Beyond bandwidth: MRDIMMs' edge in the data center revolution

FMS 2025

Henrique Pötter, PhD
Sr. System Performance Engineer

micron Intelligence Accelerated™

FMS

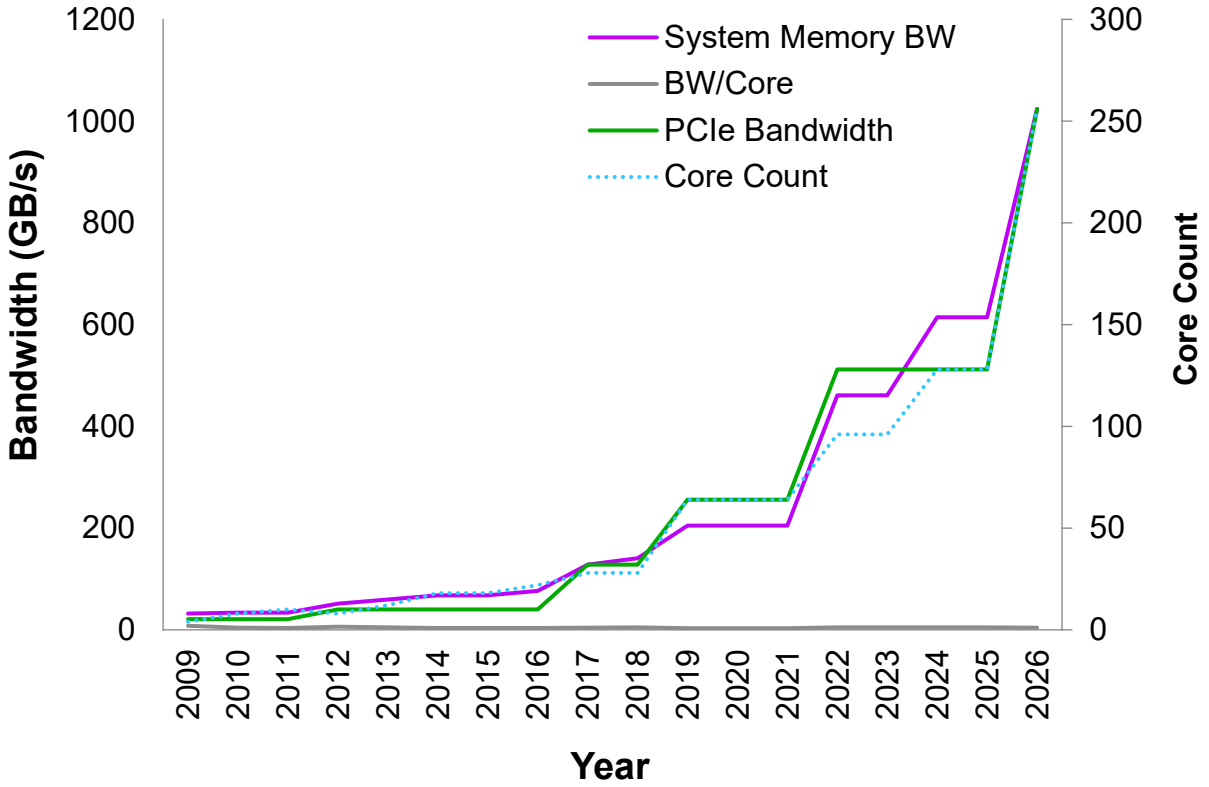
FMS25 Micron agenda

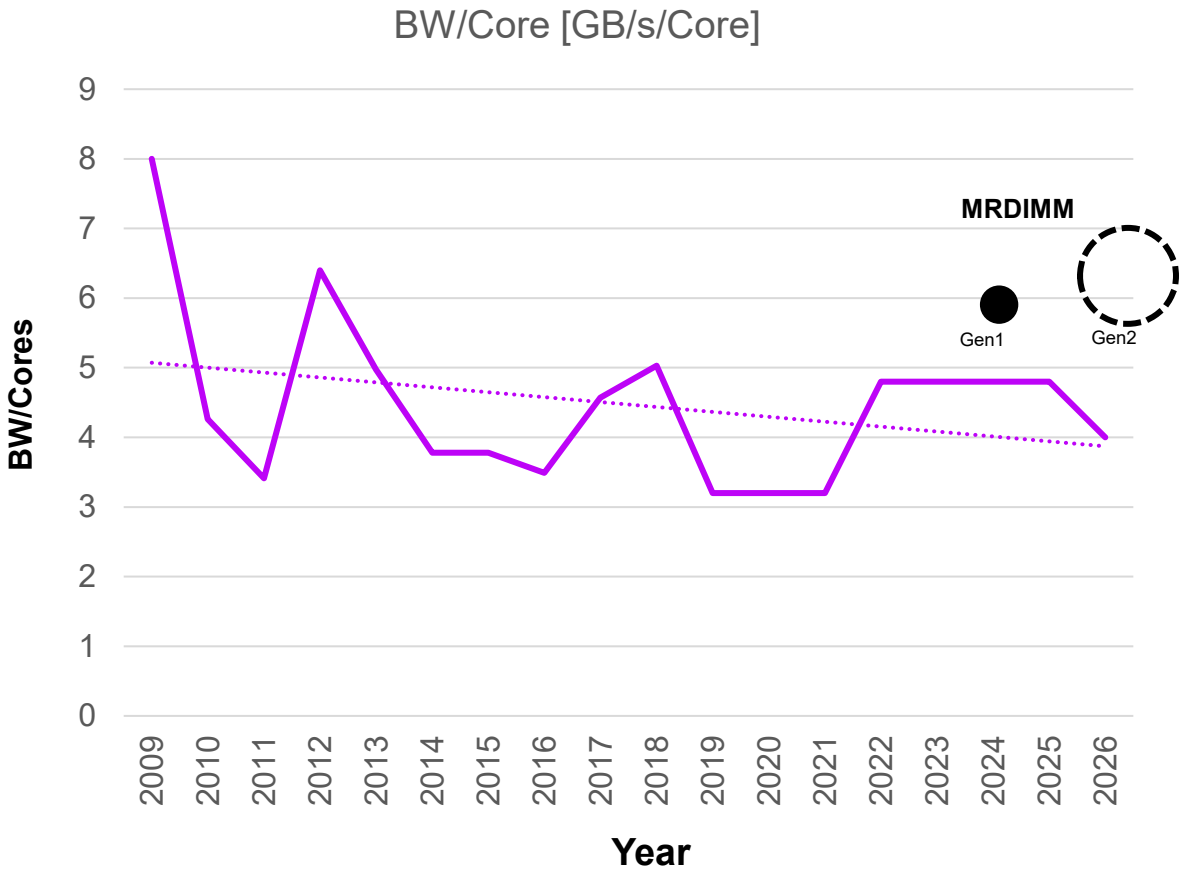
MRDIMM

- Why MRDIMM?
- How it Works?
- Workloads
  - Bandwidth
  - Capacity
  - Latency
  - Power (Task-Energy)

CPU core count, PCIe, and memory Bandwidth

Core count and Memory Bandwidth outpaces memory bandwidth increase: “Memory Wall” problem[1]



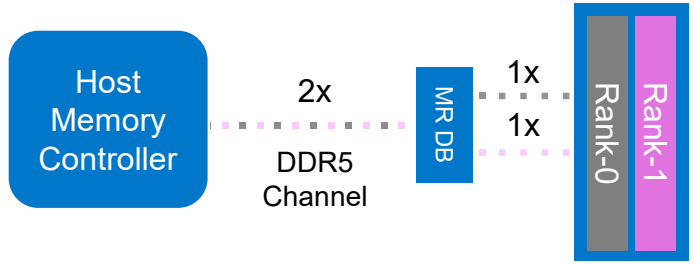


1. Based on public information gathered by Micron
2. Assumes “Performance” core counts

MRDIMM overview – multiplexed Rank DIMM

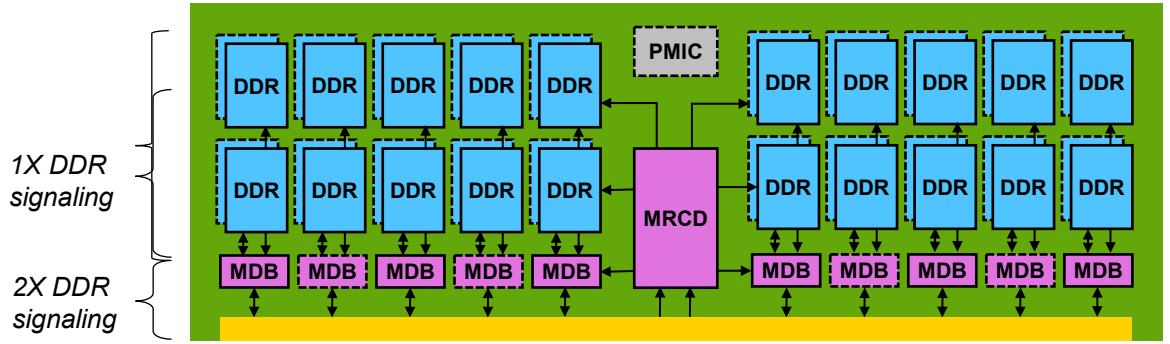
Leveraging DDR5 infrastructure while delivering increased Bandwidth to address CPU core count explosion

- Initial speeds up to 8800 MT/s (Gen1)
- Capacities from 32GB to 256GB
- Data remains aligned to 64-byte accesses from the host
  - 2 DDR ranks are multiplexed for 2X DDR speed
  - multiplexed ranks behave like pseudo channels
- RAS performance equivalent to DDR5 RDIMMs
  - Uses standard x4 or x8 DDR5 components
- Conforms to DDR5 mechanical and electrical interface
- Supports module configurations: 2Rx8, 2Rx4, 4Rx8, 4Rx4
- Broad ecosystem support



MRDIMM Concept

Similar components and floorplan as LRDIMM

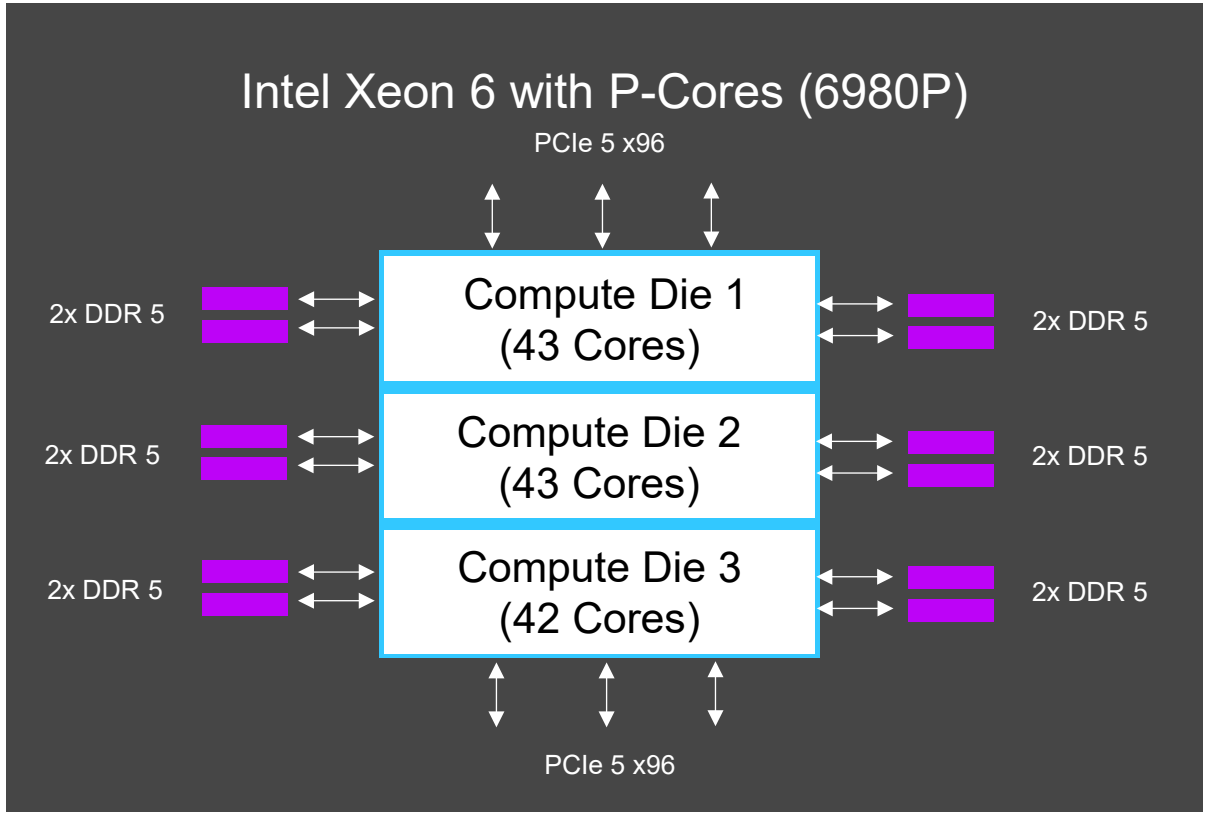


MR RCD (MRCD): simultaneously access 2 ranks, supports up to 4 total ranks
MR DB (MDB): includes 2 ports on back-side to enable muxing / de-muxing

MRDIMM workloads

GNR-AP CPU architecture

- CPU: GNR-AP, 6980P, Base 2.0 GHz, 3.9GHz 128c
  - L1d:6 MiB (128 instances)
  - L1i:8 MiB (128 instances)
  - L2: 256 MiB (128 instances)
  - L3: 504 MiB (2 instances)
- GNR-AP is composed by 3 Compute Dies
  - Each compute Die has 4 local DRAM Channels
    - Higher Bandwidth and lower Latency for local Compute die
- OS can benefit if each compute die is mapped as a numanode (Co-locate process and memory in the same numanode).
  - Enable SubNuma Cluster in BIOS (General Uncore Settings)
- Memory Latency is lower for “closer” compute dies!
- 6980P has uneven cores per compute die with 43, 43, and 42.



Bandwidth (MB/s) 64GB MRDIMM @ 8800 MT/s

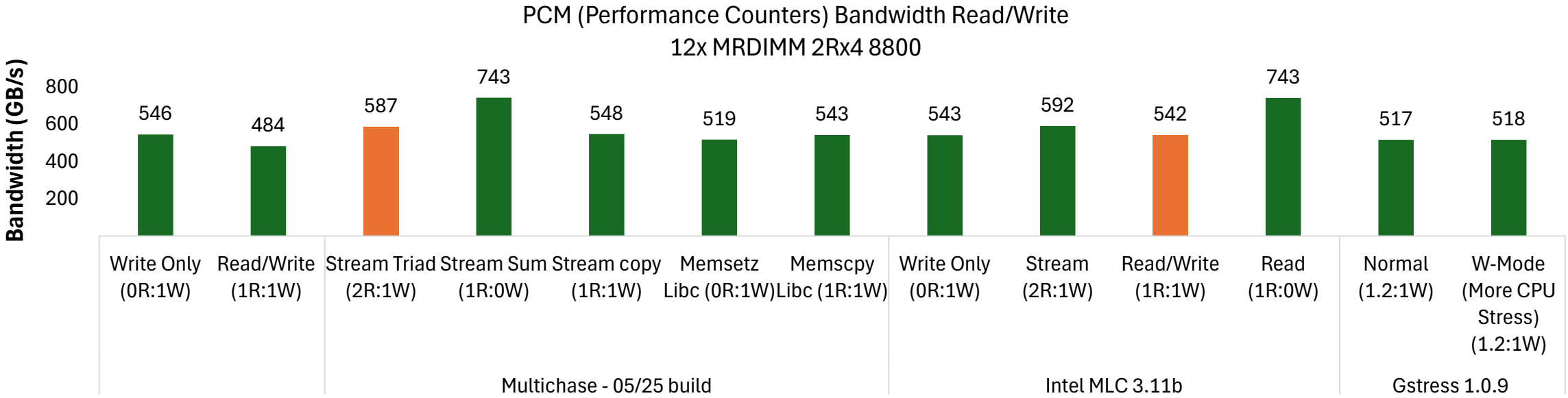
Numa Node Memory	Numa Node Process 0	1	2
0	273926	214245	214113
1	213269	274329	214199
2	212641	213342	274379

Numa Node Memory	Numa Node Process 0	1	2
0	109.3	124	149.7
1	120.4	100.6	122.3
2	140.2	120.1	102.1

Bandwidth

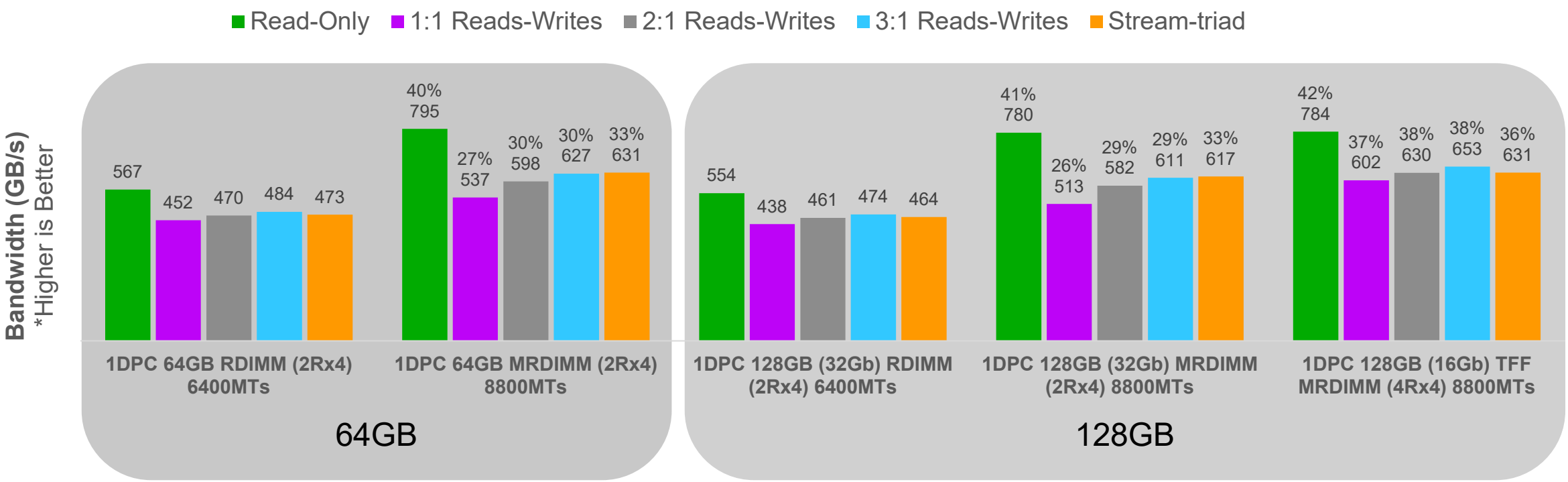
Which microbenchmark shows the highest bandwidth?

- Intel MLC Read Only (1R:0W) and Multichase Stream_sum (1R:0W) can reach the highest BW
  - Shows highest bandwidth for the different access patterns
- Intel MLC is used by Intel on any x86 from competitors
- Multichase can reach similar bandwidth as Intel MLC (Multichase shows 5% lower bandwidth than perf. counters)
- Gstress is 14% off from performance counters



Production server demonstrates dramatic performance

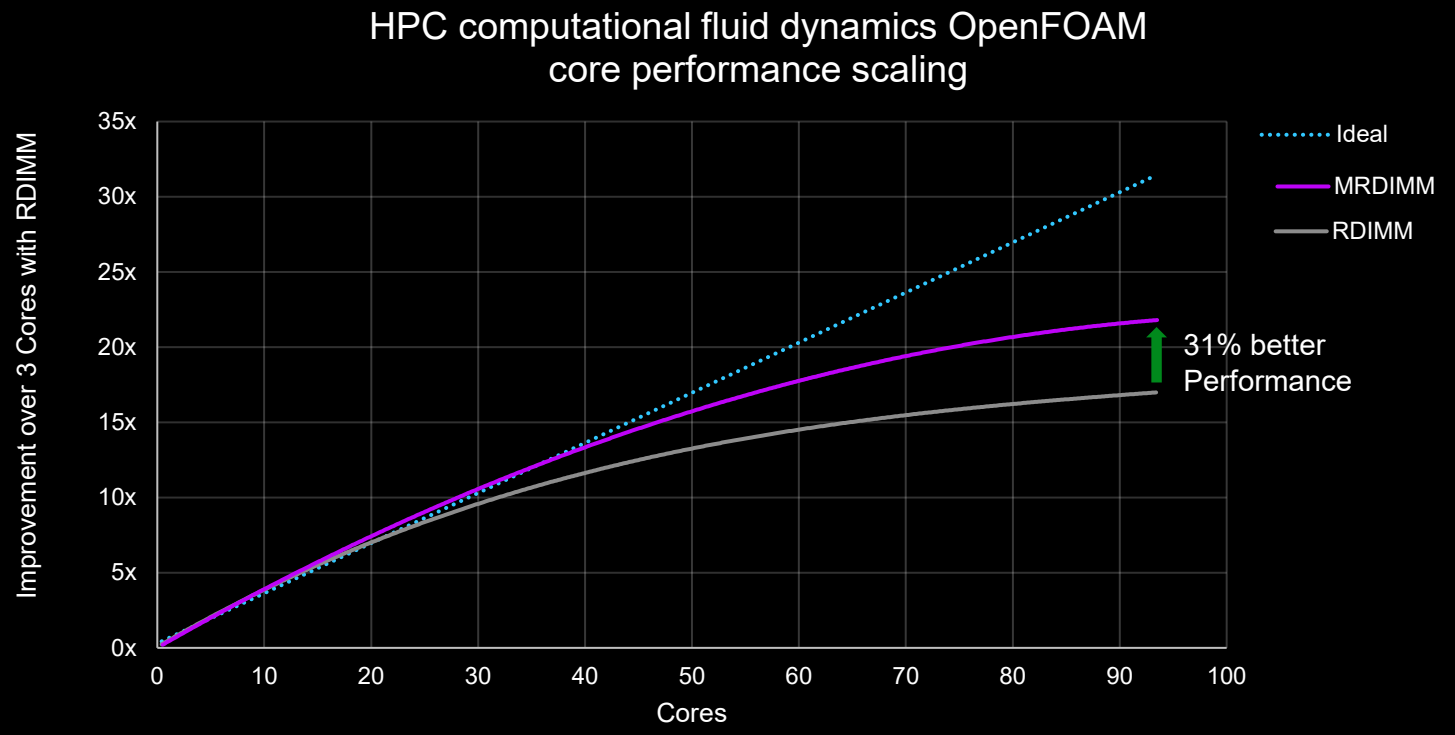
- Memory intensive Benchmark – Intel's Memory Latency Checker (MLC)
- Max bandwidth measured by the CPU generating requests at the optimal (typically the fastest) possible rate



HPC OpenFOAM

MRDIMM improves Core Performance Efficiency

- MRDIMM 8800 MT/s performance increases as more cores are used
- MRDIMM 8800 MT/s shows up to 31% Higher performance than RDIMM 6400 MT/s with 96 cores



Empirical data collected with OpenFOAM motor bike mesh 600x500x500 comparing 64GB MRDIMM 8800MT/s against 64GB RDIMM 6400MT/s

AI RAG VectorDB MRDIMM

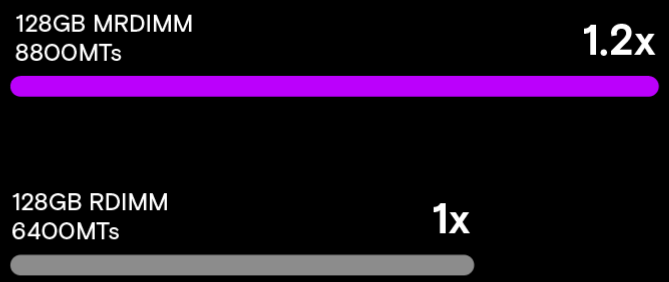
MRDIMM achieves 1.2x performance gain and 30% lower latency over RDIMM

Retrieval-Augmented Generation (RAG) is used to provide additional context for LLM models to improve the quality of response

MRDIMM 128GB vs RDIMM 128GB

- MRDIMM improves RAG performance by solving ~20% more queries per second
- MRDIMM delivers ~34% higher peak memory bandwidth utilization and 30% lower memory latency

Queries per second (QPS) improvement for 15K parallel queries
(Higher is better)



GNR-AP Avenue City | CPU 96 Cores, 2.7GHz (step 0) | BIOS BHSDCRB1.IPC.0030.D67.2404051303

AI SVM with MRDIMM

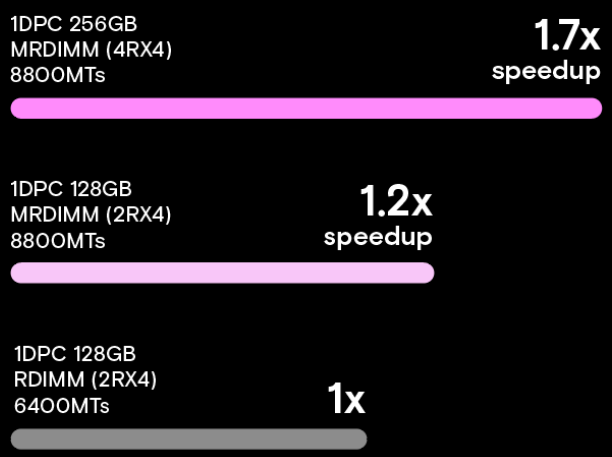
MRDIMM delivers up to 70% gain over RDIMM

Support vector machine supervised ML with 2.4TB dataset

- 128GB MRDIMM bandwidth improves SVM performance by 1.2x on GNR-AP
- 256GB MRDIMM extra capacity and bandwidth improves SVM performance by 1.7x
- SVM uses 10x less storage I/O with 256GB MRDIMM over 128GB RDIMM

Apache SPARK AI SVM improvement

Speedup (higher is better)

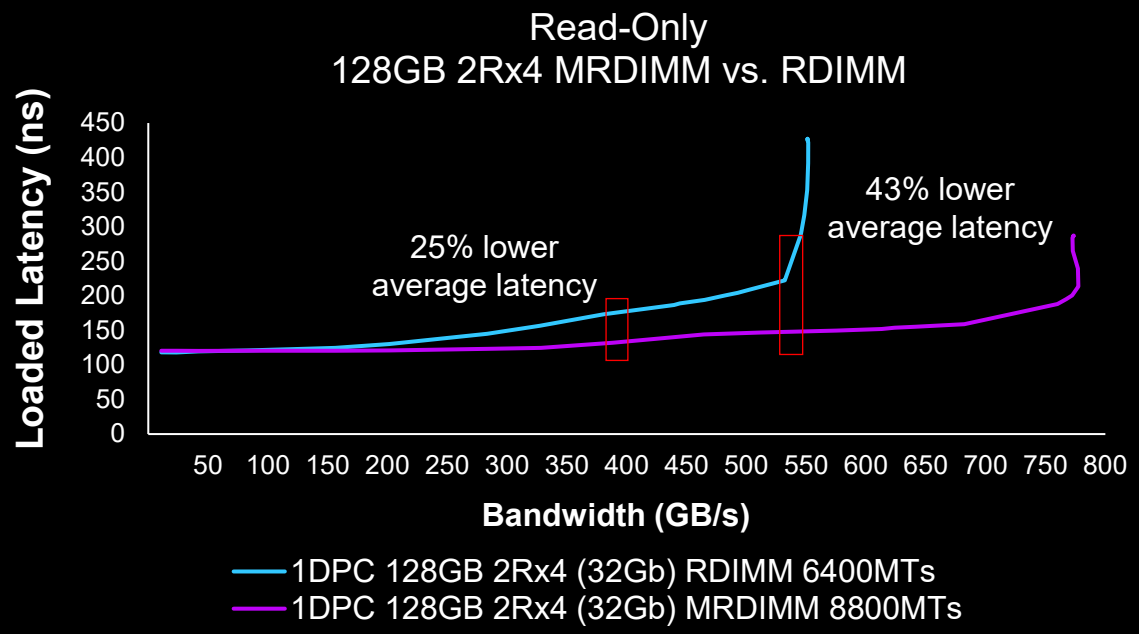


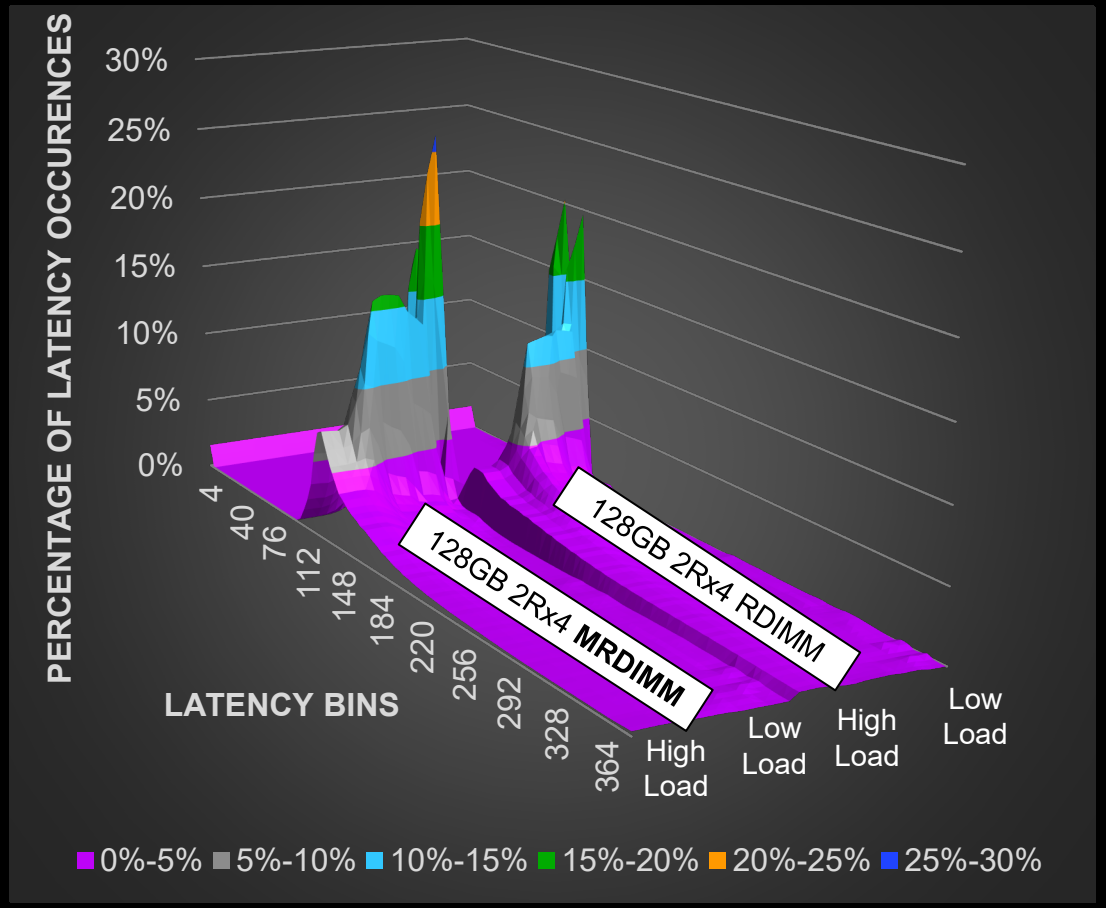
GNR-AP Avenue City | CPU 96 Cores, 2.7GHz (step 0) | BIOS BHSDCRB1.IPC.0030.D67.2404051303

Latency

MRDIMM delivers lowest consistent latency over BW spectrum

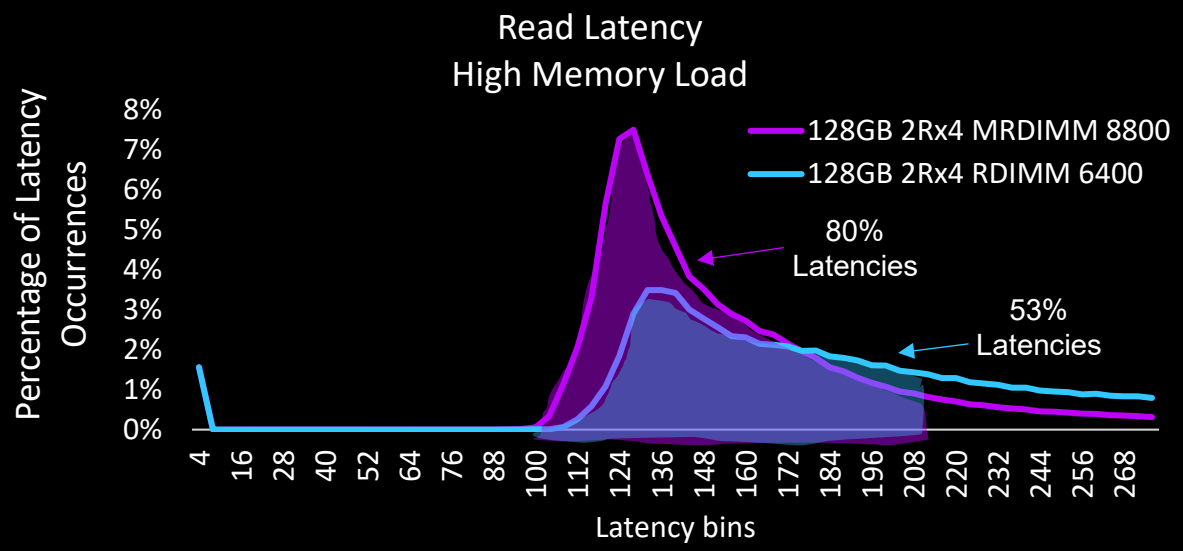
- Memory intensive Benchmark – Intel's Memory Latency Checker (MLC)
- 128GB 2Rx4 (32Gb) MRDIMM 8800 MT/s vs. 128GB 2Rx4 (32Gb) RDIMM 6400 MT/s:
  - @550GB/s: ~43% lower average latency
  - @400GB/s: ~23% lower average latency





MRDIMM delivers lowest tail latency

- MRDIMM is well behaved
- Memory intensive Benchmark – Intel's Memory Latency Checker (MLC)
- High Memory Load
  - 128GB 2Rx4 (32Gb) MRDIMM 8800 **80% of latencies** are within 208ns vs. **53% for** 128GB 2Rx4 (32Gb) **RDIMM 6400**
- More Responsive System and Improved QoS



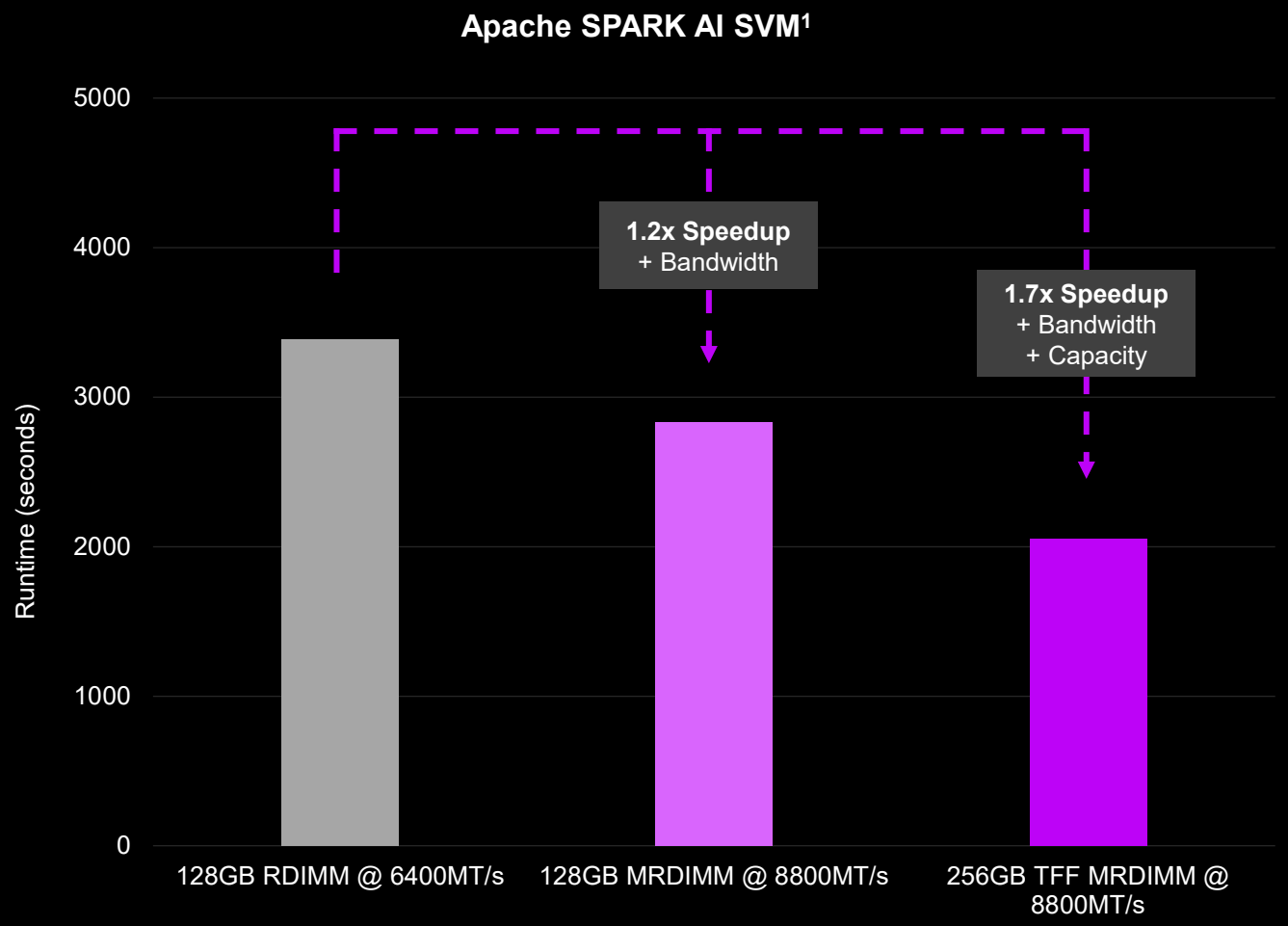
	High Load	Medium Load	Low Load
Latency Above 2µs	**-28%**	**-19%**	**-16%**

*Even under low Load MRDIMM shows lower tail latency

Capacity without limits

MRDIMM capacity advantage[1]

- ML Algorithms scaling to larger data set sizes
- Extending Data set size capability with higher MRDIMM capacity
- More than just Capacity but Capacity with Bandwidth
  - 20% increase due to MRDIMM Bandwidth
  - 70% increase adding MRDIMM bandwidth and Capacity



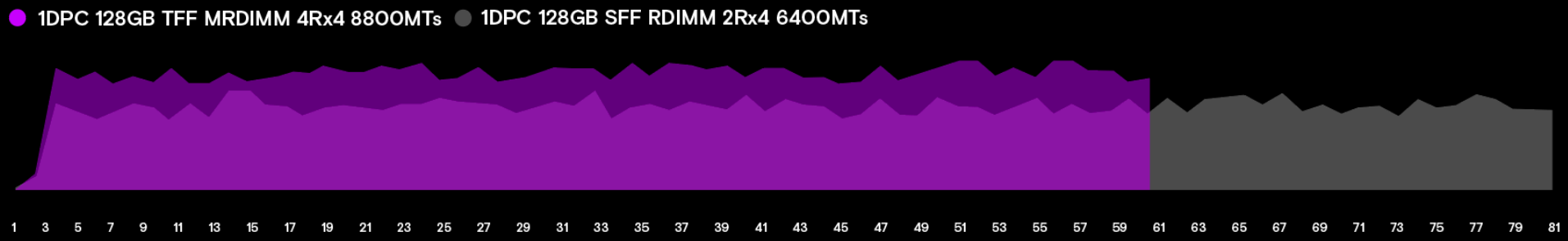
1. Spark Support Vector Machine Runtime

Power = task energy

Task energy delivers improved TCO
HPC – OpenFOAM

- OpenFOAM CFD simulation runs up to 1.34x faster with MRDIMM
- Despite higher total server power, MRDIMM uses up to 11% less energy per task (completes the task faster)

OpenFOAM server power (watts)



MRDIMM uses
11%
less energy

GNR-AP Avenue City
CPU 96 Cores, 2.7GHz (step 0)
BIOS BHSDCRB1.IPC.0030.D67.2404051303

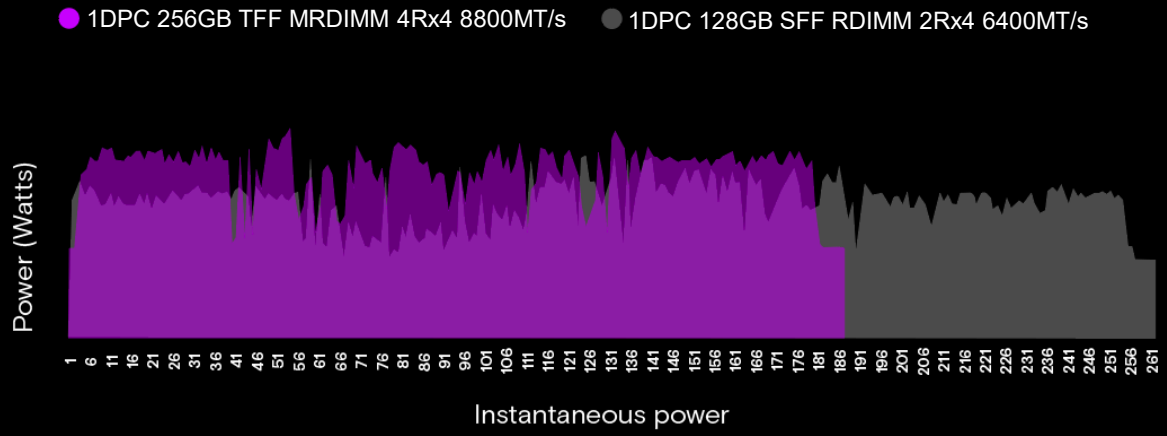
AI SVM task-energy with RDIMM

MRDIMM is >1.2x more energy efficient than RDIMM

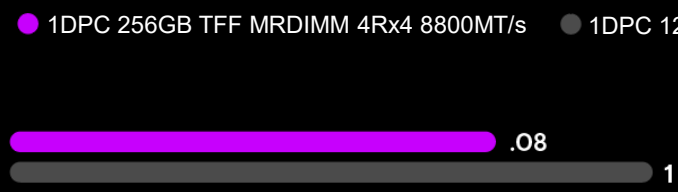
- Apache Spark SVM (Machine Learning Library)
- Performance:
  - >1.7x performance gain when processing 2.9TB dataset with 256GB 4Rx4 TFF MRDIMM over 128GB RDIMM (2Rx4) 6400MTs
- Energy efficiency:
  - > 1.2x energy efficiency when processing 2.9TB dataset with 256GB 4Rx4 TFF MRDIMM over 128GB RDIMM (2Rx4) 6400MTs

*Apache Spark SVM power measured on when processing a 2.9TB dataset. Experiment based on Intel Hibench. Intel Xeon 6 with 96 P-cores. Avenue City Prototype board. Bios 3376, Linux kernel 6.9.6-1, Alma 9

Apache Spark SVM
Server power consumption over time (Watts)



Apache Spark SVM
Relative task-energy (area under curve)



GEN1 MRDIMM Power* Map to Platform Cooling

MRDIMM Product			RDIMM Power	MRDIMM Power	
Density (GB)	Rank	DiD	6400 MT/s	8000 MT/s	8800 MT/s
256GB (TFF)	4Rx4	Y53A		20.2	21.2
128GB	2Rx4	Y53A	11.4	17.1	18.2
128GB (TFF)	4Rx4	Y52K		19.5	20.5
96GB	2Rx4	Y5CB	11.2	16.8	17.9
	4Rx8	Y5CB		13.9	15.0
64GB	2Rx4	Y52K	11.0	14.7	17.2
	4Rx8	Y52K		13.5	14.4
48GB	2Rx8	Y5CB		10.8	11.6
32GB	2Rx8	Y52K		11.4	12.2

< 18Watts	18Watts – 25Watts

* Early measured & Simulated power estimates (Feb 2025)

- Usage Conditions:
  - Nominal Voltage = 1.1V, Temperature = 60°C
  - 1.15x READs to WRITEs
  - Bus Utilization: 8000 (65%-72%), 8800 (62%-70%), 12800 (54%-61%) (Dependent on module config)
  - Google Stress (StressAppTest) Workload
  - 1X Refresh (<85C)

Micron Confidential

Gen 2 MRDIMM Power* Map to Platform Cooling

MRDIMM Product			MRDIMM Power	
Density (GB)	Rank	DID	10400 MT/s	12800 MT/s
256GB (TFF)	4Rx4	Y53A	23.5	24.8
192GB (TFF)	4Rx4	Y5CB	22.4	24.2
128GB	2Rx4	Y53A	20.3	21.5
128GB	4Rx8	Y53A		17.3
128GB (TFF)	4Rx4	Y62E	20.1	21.6
96GB	2Rx4	Y5CB	19.3	21.3
96GB	4Rx8	Y5CB		17.7
64GB	2Rx4	Y62E	16.3	17.8
	4Rx8	Y53A		15.1
48GB	2Rx8	Y5CB		12.3
32GB	2Rx8	Y62E		12.7
		< 18Watts		18Watts – 25Watts

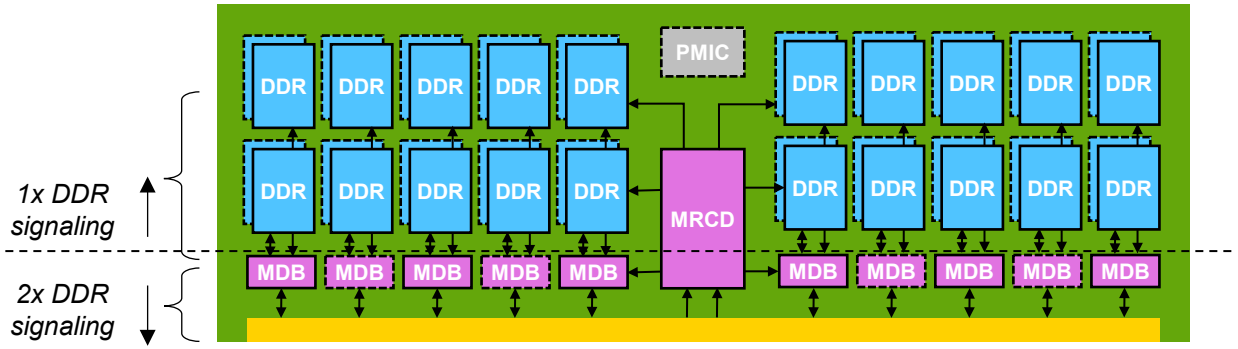
* Early measured & Simulated power estimates (Feb 2025)

- Usage Conditions:
  - Nominal Voltage = 1.1V, Temperature = 60°C
  - 1.15x READs to WRITEs
  - Bus Utilization: 8000 (65%-72%), 8800 (62%-70%), 12800 (54%-61%) (Dependent on module config)
  - Google Stress (StressAppTest) Workload
  - 1X Refresh (<85C)

Micron DDR5 Gen1 MRDIMM

Delivers the highest performance high-capacity main memory solution for data centers

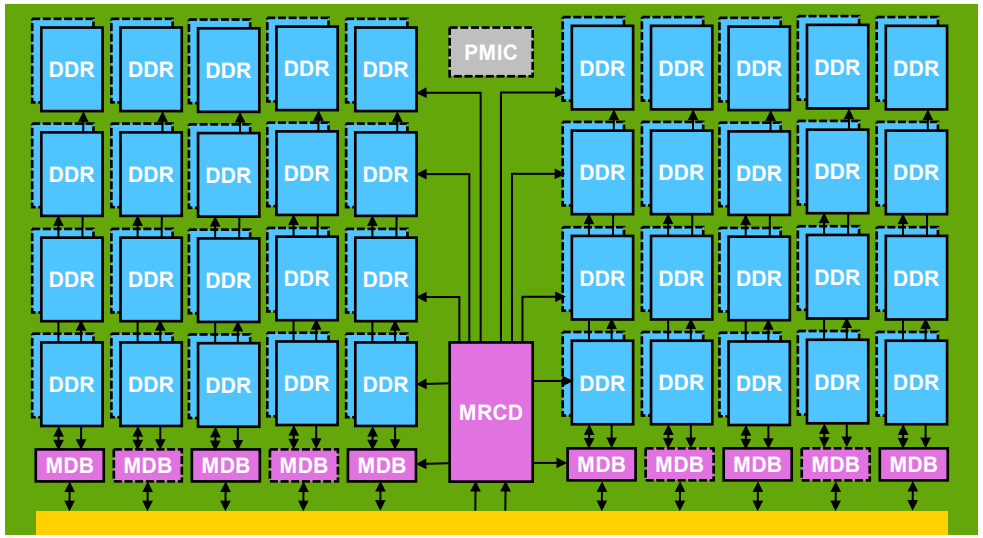
Standard Height DDR5 Form Factor MRDIMM



Server 1U fit

MRDIMM Density	MPN	Config	Speed (MT/s)	DRAM Density
32GB	MTC20F2085S1HC88XD1	2Rx8	8000, 8800	16Gb
48GB	MTC20F208XS1HC88XC1	2Rx8	8000, 8800	24Gb
64GB	MTC40F2046S1HC88XD1	2Rx4	8000, 8800	16Gb
64GB	MTC40F4086S1HC88XD1	4Rx8	8000, 8800	16Gb
96GB	MTC40F204WS1HC88XB1	2Rx4	8000, 8800	24Gb
96GB	MTC40F408WS1HC88XC1	4Rx8	8000, 8800	24Gb
128GB	MTC40F2047S1HC88XB1	2Rx4	8000, 8800	32Gb
256GB	MTC80F4048S2HC88XB1	4Rx4	8000, 8800	32Gb DDP

Tall Form Factor (TFF: H – 56.9mm) 4Rx4 MRDIMM



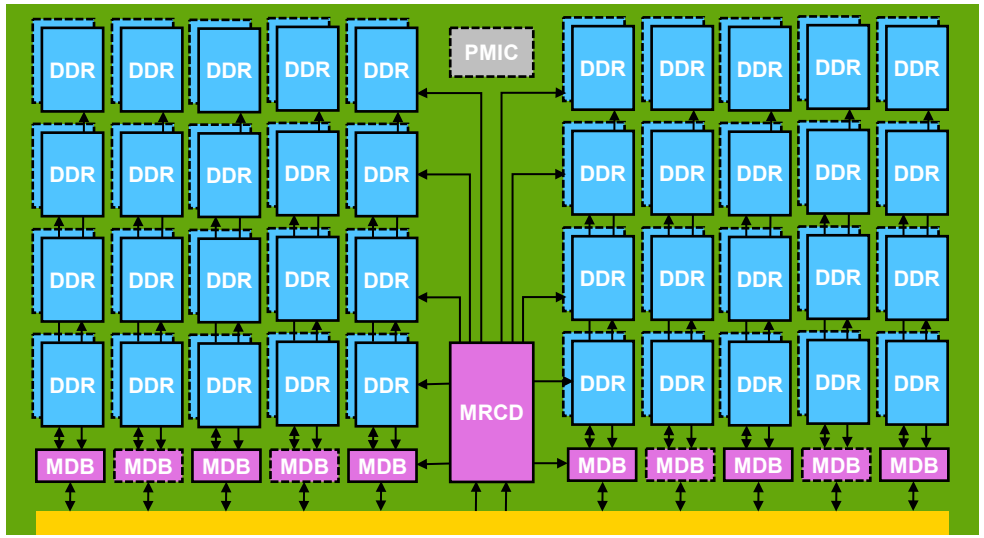
Server >1U fit

MRDIMM Density	MPN	Config	Speed (MT/s)	DRAM Density
128GB	MTC80F4047M1HC88XD1	4Rx4 (TFF)	8000, 8800	16Gb
256GB	MTC80F4048M1HC88XB1	4Rx4 (TFF)	8000, 8800	32Gb

Micron DDR5 Gen2 MRDIMM

Delivers the highest performance high-capacity main memory solution for data centers

Tall Form Factor (TFF: H – 56.9mm) 4Rx4 MRDIMM

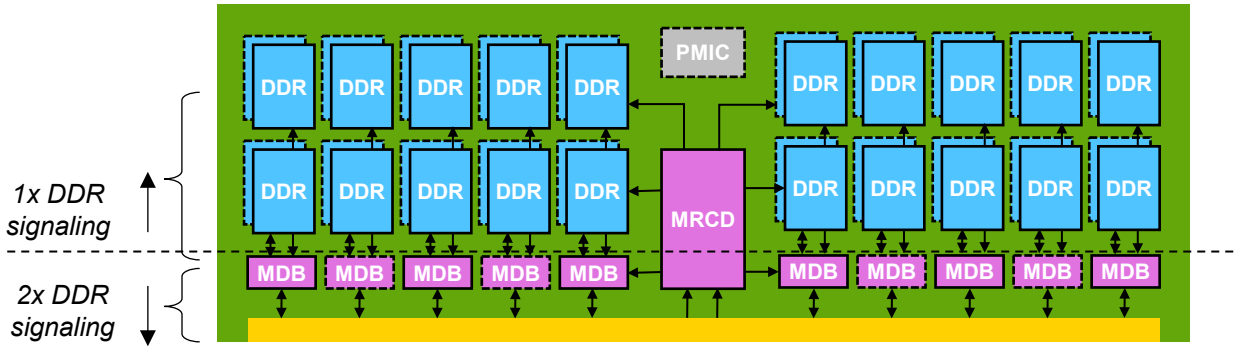


Server >1U fit

MRDIMM Density	MPN	Config	Speed (MT/s)	DRAM Density
128GB	MTC80F4047M1HC1AXH1	4Rx4 (TFF)	12800	16Gb
256GB	MTC80F4048M1HC1AXB1 MTC80F4048M1HC1AXE1 *	4Rx4 (TFF)	12800	32Gb

* 1γ 32Gb in ~2H-CY2026

Standard Height DDR5 Form Factor MRDIMM



Server 1U fit

MRDIMM Density	MPN	Config	Speed (MT/s)	DRAM Density
32GB	MTC20F2085S1HC1AXH1	2Rx8	12800	16Gb
48GB	MTC20F208XS1HC1AXC1	2Rx8	12800	24Gb
64GB	MTC40F2046S1HC1AXH1	2Rx4	12800	16Gb
64GB	MTC40F4086S1HC1AXH1	4Rx8	12800	16Gb
64GB	MTC20F2086S1HC1AXB1 MTC20F2086S1HC1AXE1 *	2Rx8	12800	32Gb
96GB	MTC40F204WS1HC1AXC1	2Rx4	12800	24Gb
96GB	MTC40F408WS1HC1AXC1	4Rx8	12800	24Gb
128GB	MTC40F2047S1HC1AXB1 MTC40F2047S1HC1AXE1 *	2Rx4	12800	32Gb
128GB	MTC40F4087S1HC1AXB1 MTC40F4087S1HC1AXE1 *	4Rx8	12800	32Gb