Revolutionizing Memory for AI/ML's Future: MRDIMM

Pankaj Goel
Associate Director, Siemens EDA
Questa One Avery VIP

Agenda

Evolving Memory Demands in AI

Bandwidth Needs	AI/ML models require moving large volumes of data between memory and accelerators
	Standard memory architectures often can't supply data fast enough, causing the processor to wait idly for memory fetches causing bandwidth bottleneck.
Latency Issues	Latency is the time it takes to fetch data from memory.
	In AI training repeated access to large data causes long fetch cycle slowing down computation
Capacity Constraints	Large models need high memory capacity to avoid constant paging.
	Limited DRAM causes frequent data movement to slower storage, reducing performance.
Poor Memory Access Patterns	AI workloads often involve irregular memory access which traditional memory hierarchies aren't optimized for.
	This leads to cache misses, inefficient data prefetching, and increased memory access overhead.

Evolving Memory Demands in AI

Parallelism in Memory Access	AI accelerators process operations in parallel, however traditional memory subsystems are not designed for massive parallel data access, creating a mismatch.
	This limits multi-threaded throughput and stalls processing pipelines.
Power and Thermal needs	High memory bandwidth and capacity requirements increase power consumption and heat
	Systems may throttle performance to stay within thermal limits for AI applications.
Memory Scaling	As models grow (billions of parameters) however memory technologies haven't scaled proportionally in speed or efficiency.
	This widening gap limits how fast and efficiently models can be trained or deployed.

MR-DIMM Technology

MRDIMM	• Multiplexed Rank Dual In-line Memory Module (MR-DIMM) is next evolution in memory modules • High-Performance Design for AI and HPC Workloads
Key Components	• Multiplexing Registered Clock Driver(MRCD) Manages clock signals for precise timing • Multiplexing Data Buffer (MDB): Enables high-speed data muxing/demuxing and reduces CPU load
Key Features:	• Supports up to 4 Ranks per DIMM • Timing Modes Supported: 1N and 2N • Operation Modes: Mux Mode and Rank Mode
Data Rates	• Up to 12,800 MT/s today • Future-ready for higher rate (17,600 MT/s)
Capacity	• Up to 256 GB per DIMM

MR-DIMM Technology

General Layout

MR-DIMM Technology : Mux Mode

- Simultaneously accessing ranks within a sub-channel
- Each sub-channel is divided into two pseudo-channels which operate somewhat independently from each other
- Resulting in total bandwidth of the Host bus can be at double the data rate of a standard DIMM with the DRAMs operating at the same data rate

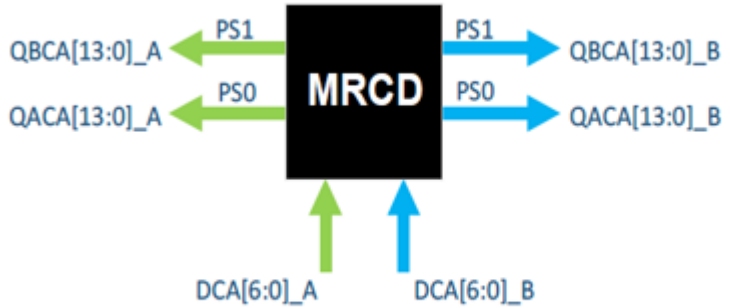


Figure 2 — MRCD DCA and QCA Ports in Mux mode

Table 9 — Mode A - 2Rx4 DCS to QCS Mapping

Clock	DCS Signal Active	QCS Signal Active
Even	DCS0_n	Q[A]CS[1:0]_n (PS0 Rank0)
Even	DCS1_n	This case is illegal
Odd	DCS0_n	Q[B]CS[1:0]_n (PS1 Rank0)
Odd	DCS1_n	This case is illegal

Table 10 — Mode B - 4Rx4 DCS to QCS Mapping

Clock	DCS Signal Active	QCS Signal Active
Even	DCS0_n	QACS0 (PS0 Rank0)
Even	DCS1_n	QACS1 (PS0 Rank1)
Odd	DCS0_n	QBCS0 (PS1 Rank0)
Odd	DCS1_n	QBCS1 (PS1 Rank1)

MR-DIMM Technology : Mux Mode

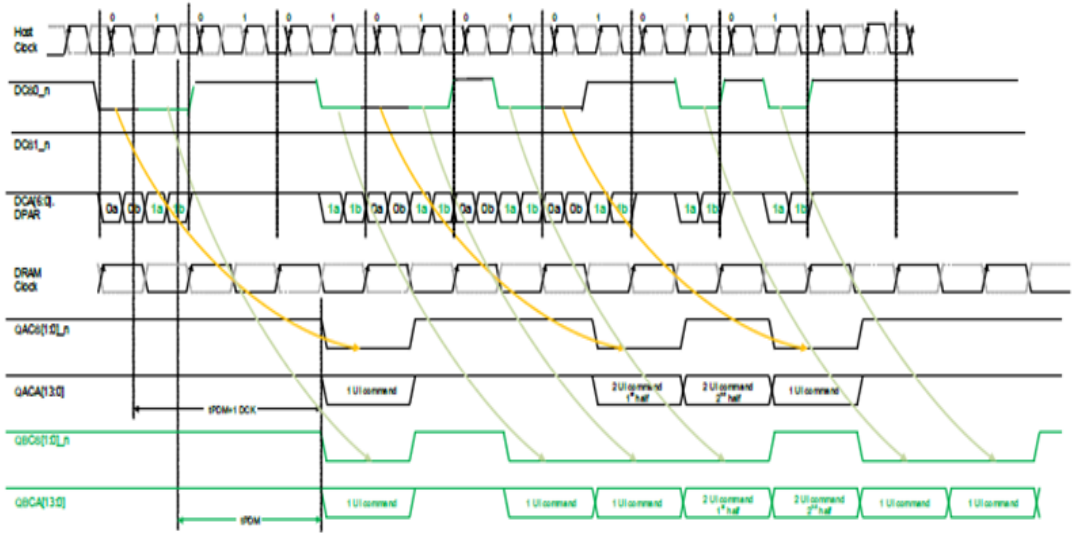


Figure 5 — Mode A - 2Rx4 Command Address Bus

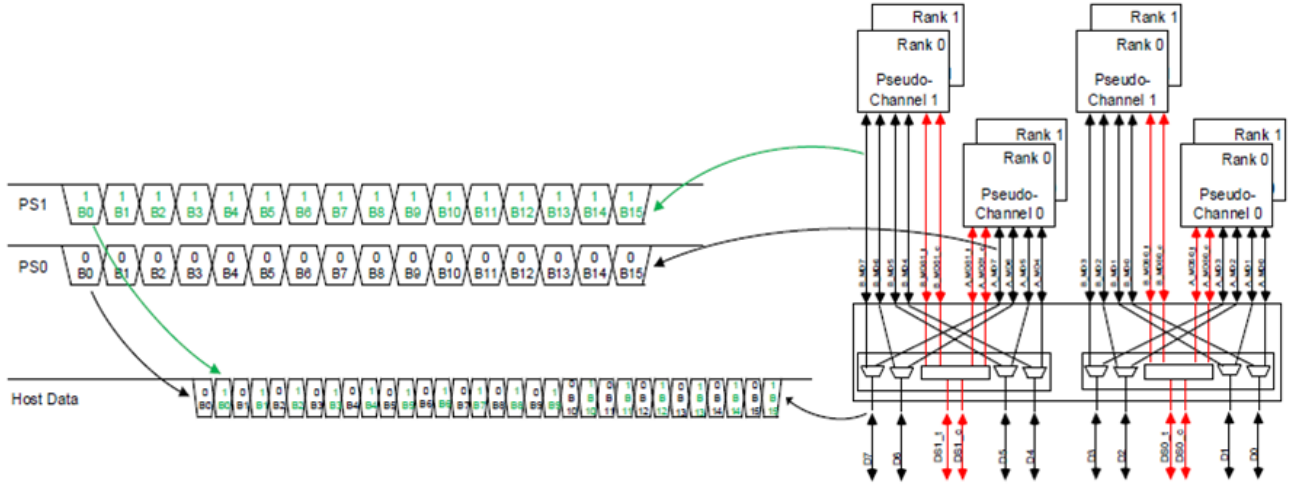


Figure 3 — MDB x4 Mux Mode

MR-DIMM Technology : Rank Mode

- Similar to DDR5 LRDIMM
- Host bus operates at the same data rate as the DRAM
- Can support up to 4 ranks
- Comprises Rank 1N and Rank 2N modes

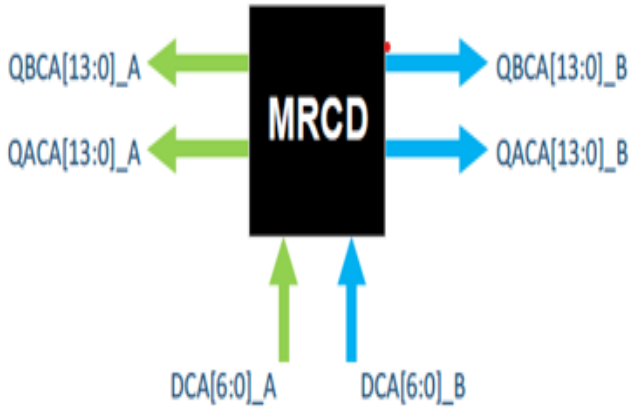


Figure 3 — MRCD DCA and QCA Ports in Rank mode

Table 6 — DCS to QCS mapping in Rank mode[1,2]

Input CA13 in the first UI	Input CS	Output CS	Rank #
0	DCS0_n	QACS0_n	Rank 0
0	DCS1_n	QACS1_n	Rank 1
1	DCS0_n	QBCS0_n	Rank 2
1	DCS1_n	QBCS1_n	Rank 3

MR-DIMM Technology : Rank Mode

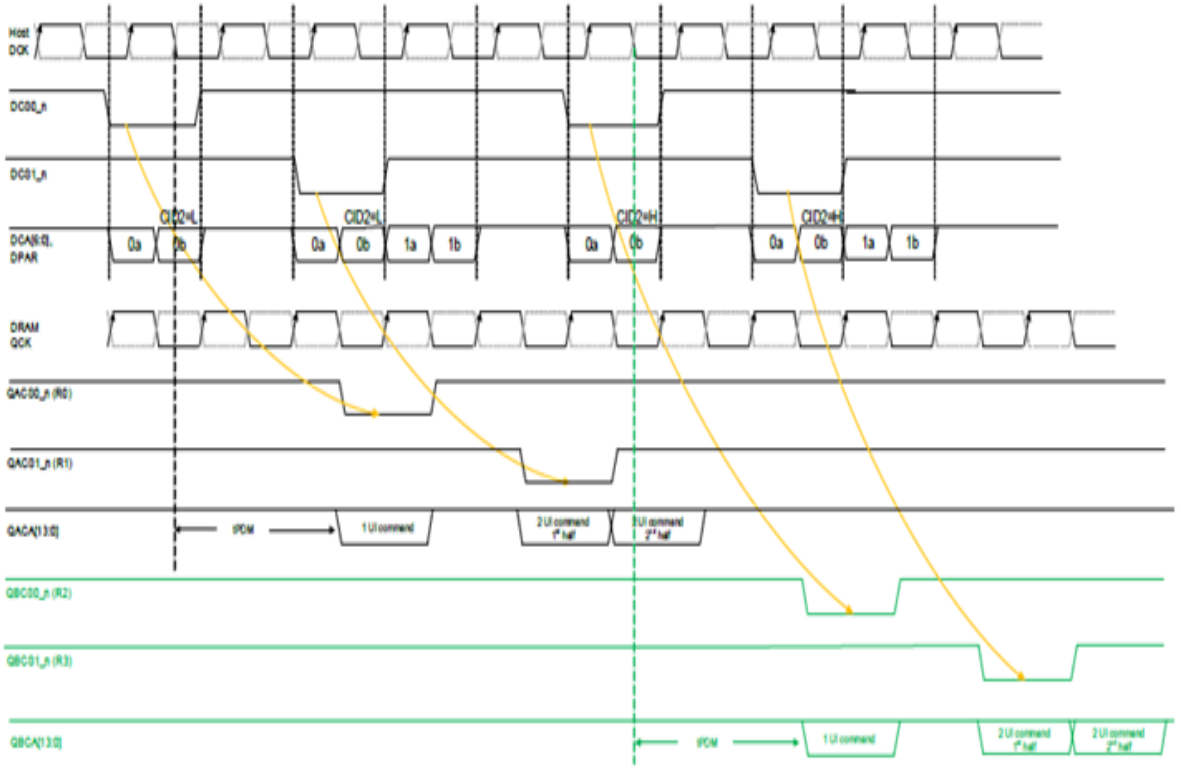


Figure 10 — 1N Rank mode

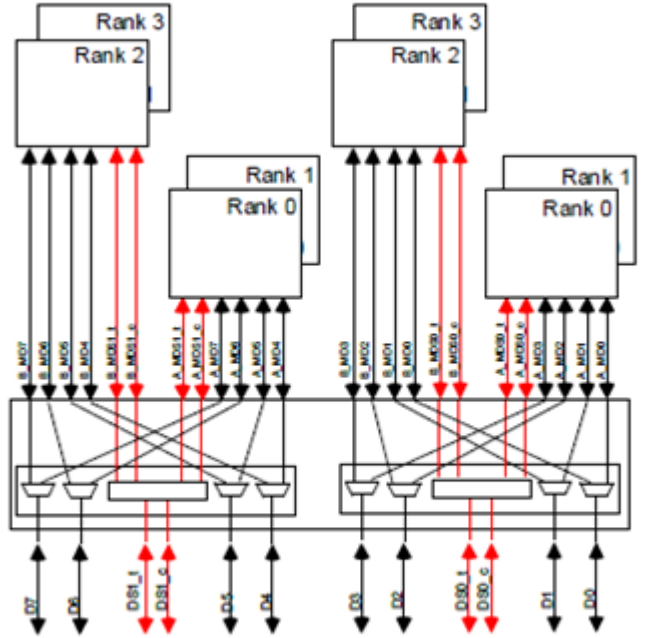
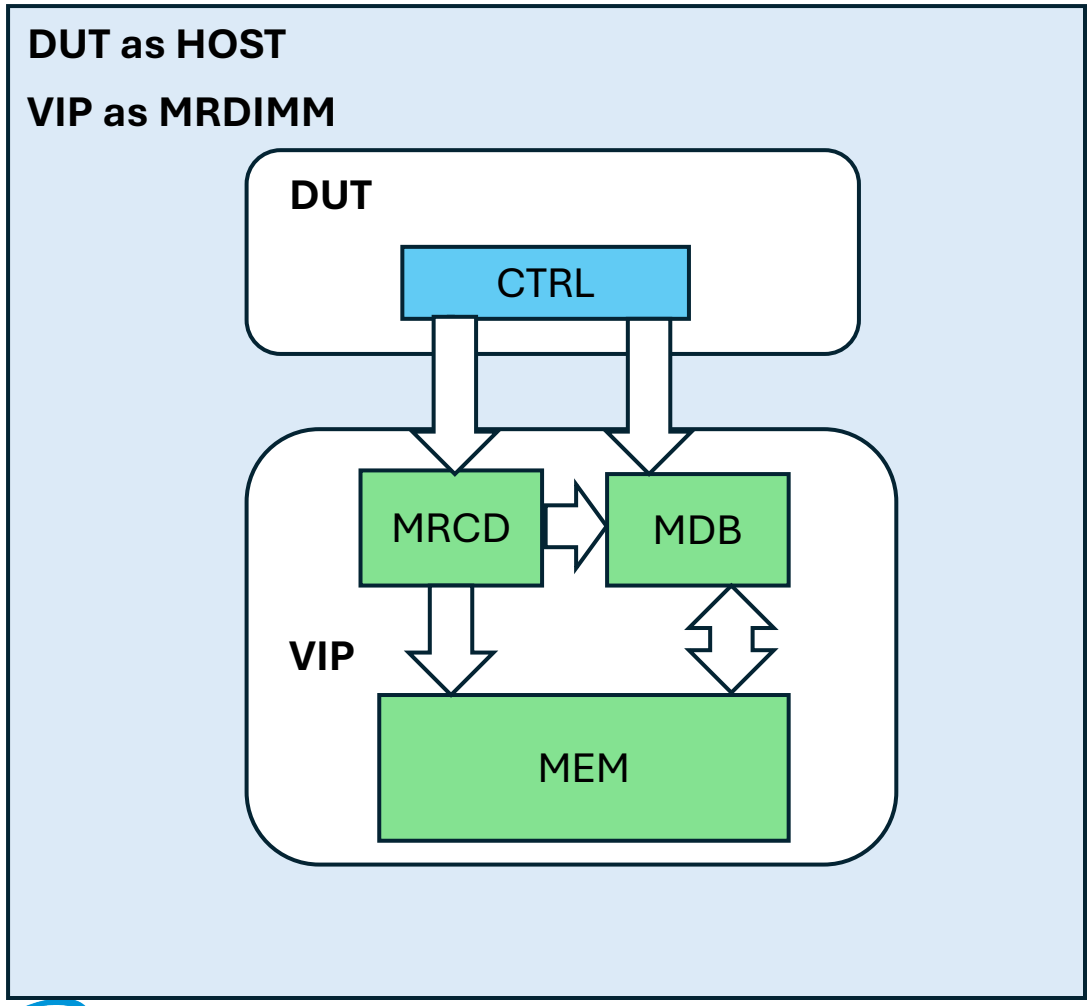


Figure 5 — MDB x4 Rank Mode

MR-DIMM Technology: Verification Use Case



- VIP Vendor (Avery VIP) provides modular components
- Any component can be replaced by VIP or DUT
- Dedicated and clean connection
- In built Tests and Checks for compliance
- Coverage for spec completeness

MR-DIMM Edge

Higher Memory Bandwidth	Uses signal multiplexing to double the number of ranks per memory channel without doubling the physical interface width, allowing for much higher bandwidth per channel.
	Up to ~39% higher effective memory bandwidth MRDIMM boosts bandwidth from ~6,400 MT/s (DDR5 RDIMM) to ~12,800 MT/s, delivering more data to AI cores faster
Better Latency Optimization	Latency reduced by ~40% in real-world workloads Lower round-trip delay compared to 128 GB @ 6,400 MT/s RDIMM
	Although the multiplexing introduces a slight increase in rank-to-rank latency, increased parallelism across ranks and internal buffering helps reduce effective memory access delays in AI workloads.
Capacity & Form Factor	Supports 32–256 GB per module, including Tall Form Factor (TFF) for 2U/4U servers
	Enables 256 GB TFF MRDIMM using 32-Gb dies at same power as 128 GB @ 16-Gb

MR-DIMM Edge

Thermal & Energy Efficiency	TFF design reduces DRAM temperature by ~20 °C under same airflow and power
	On-module PMIC buffers deliver lower power-per-task, improving data-center performance-per-watt
Ecosystem Compatibility	Drop-in compatible with DDR5 RDIMM slots and Xeon-6/Granite Rapids CPUs—no motherboard or BIOS changes needed
	Maintains full RDIMM RAS features (ECC, buffering, error correction)
AI-Workload Impact	Real-world tests on Xeon-6 show ~33% faster completion of AI-style tasks
	Ideal for LLMs, AI retraining, and inference workloads using CPUs and accelerators

Ecosystem Integration and Industry Adoption

Standardization through JEDEC	JEDEC released in early 2024	Defines the DDR5 MR-DIMM
		Pinout and form-factor alignment with DDR5 RDIMM
		On-DIMM PMIC and data buffers (MDBs)
Memory Vendor Adoption	Micron: MR-DIMM samples at 8800 MT/s launched Q1 2025.	
	Samsung: Developed 128GB MR-DIMM modules with DDR5-8000 speed targeting AI training.	
	SK hynix: Announced roadmap for MR-DIMM targeting AI inference in HPC clusters.	
VIP Vendors Support	Enables SoC and memory controller developers to begin validation ahead of mass adoption.	
	Helps accelerate time-to-market for platforms adopting MR-DIMM.	
	Avery, a key player in VIP solutions, has already added MR-DIMM to its offerings.	
Server OEMs & Systems	HPE, Dell, Lenovo: Prototype MR-DIMM-enabled systems	Generative AI workloads
		Large-scale training clusters
		In-memory analytics (SAP HANA, Spark)

MRDIMM vs the Rest: Taking the Lead

Feature	RDIMM	LRDIMM	MRDIMM
Components	DRAM, RCD	DRAM, RCD, DB	**DRAM, MRCD, MDB** – advanced and modular for higher flexibility
Modes	Not Applicable	Not Applicable	**Supports Rank Mode and Mux Mode** – enabling performance and scalability options
Pseudo Channel	Not supported	Not supported	**Supported** – improves parallelism and memory access efficiency
Data Multiplexing (Host)	Not supported	Not supported	**Supported** – reduces I/O & increases bandwidth
DQ Rate	Same on Host & DRAM	Same on Host & DRAM	**Mux Mode**: Host DQ rate is **2× DRAM DQ rate** – faster data handling
Speed Bins	Up to DDR5-8000	Up to DDR5-8000	**Mux Mode**: Up to **DDR5-12800** – unmatched speed
Ranks	2	2	**Rank Mode**: 4 **(higher capacity)**

Conclusion

MR-DIMM is a major leap in memory subsystem design

Solves bandwidth and capacity bottlenecks without platform changes

Forward-looking, scalable solution compatible with:

- DDR5 ecosystems
- CXL memory expansion
- Future DDR6 support

Enables:

- Next-gen AI and memory-bound workloads
- Hybrid topologies: RDIMM + MR-DIMM + CXL
- Modular and heterogeneous memory systems

References

- JEDEC Memory Modules
- Micron MRDIMM Innovations
- Intel MRDIMM announcement
- TechPowerUp article on MRDIMM
- Rambus DDR5 MRDIMM technical
- Tom’s Hardware Intel MRDIMM coverage