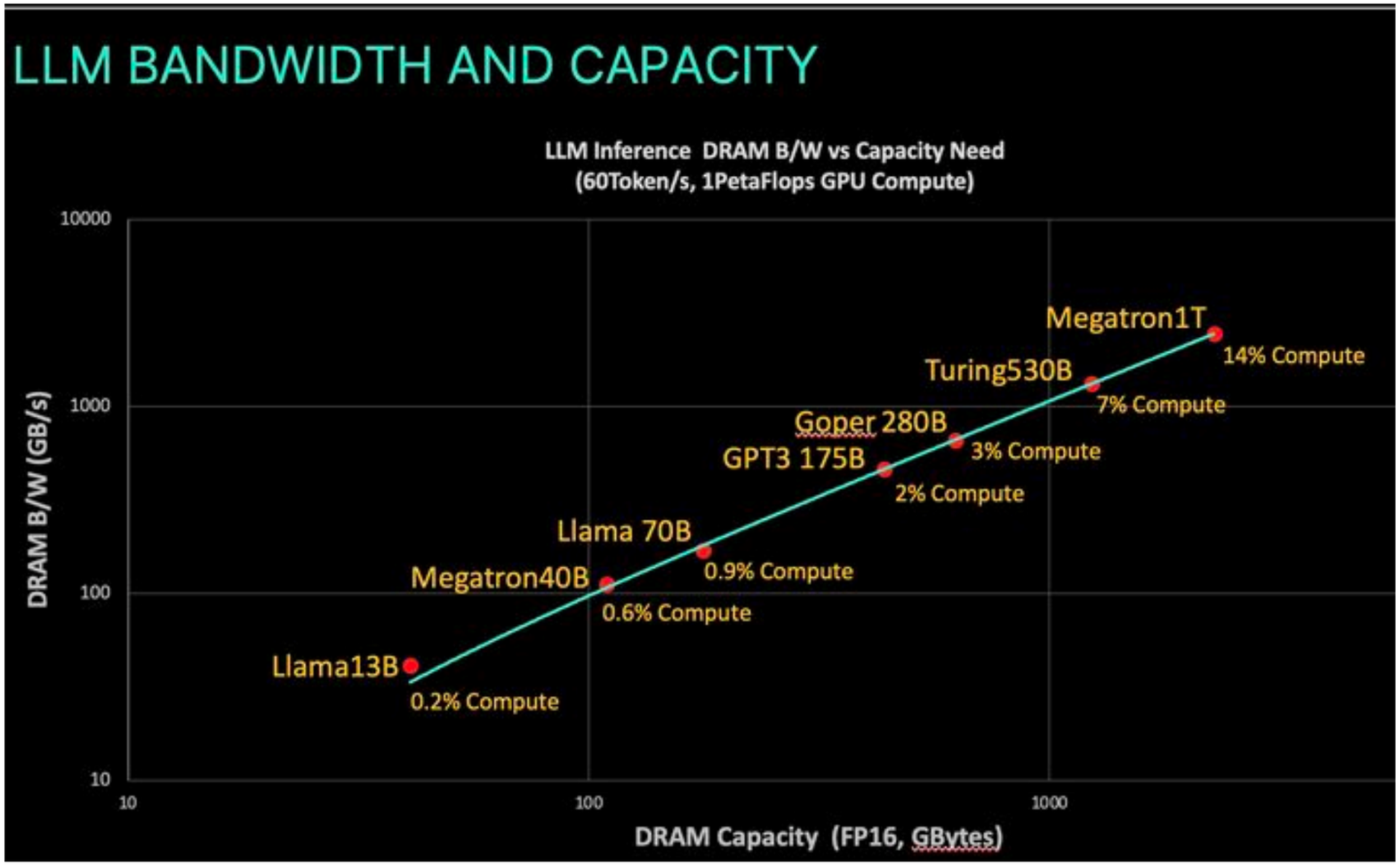
A Deep Dive Into Next-Gen MRDIMM Server Memory

Steven Woo
Fellow and Distinguished Inventor

Matt Jones
Vice President, Strategic Marketing

Loop Capital Expert Call
February 18, 2025

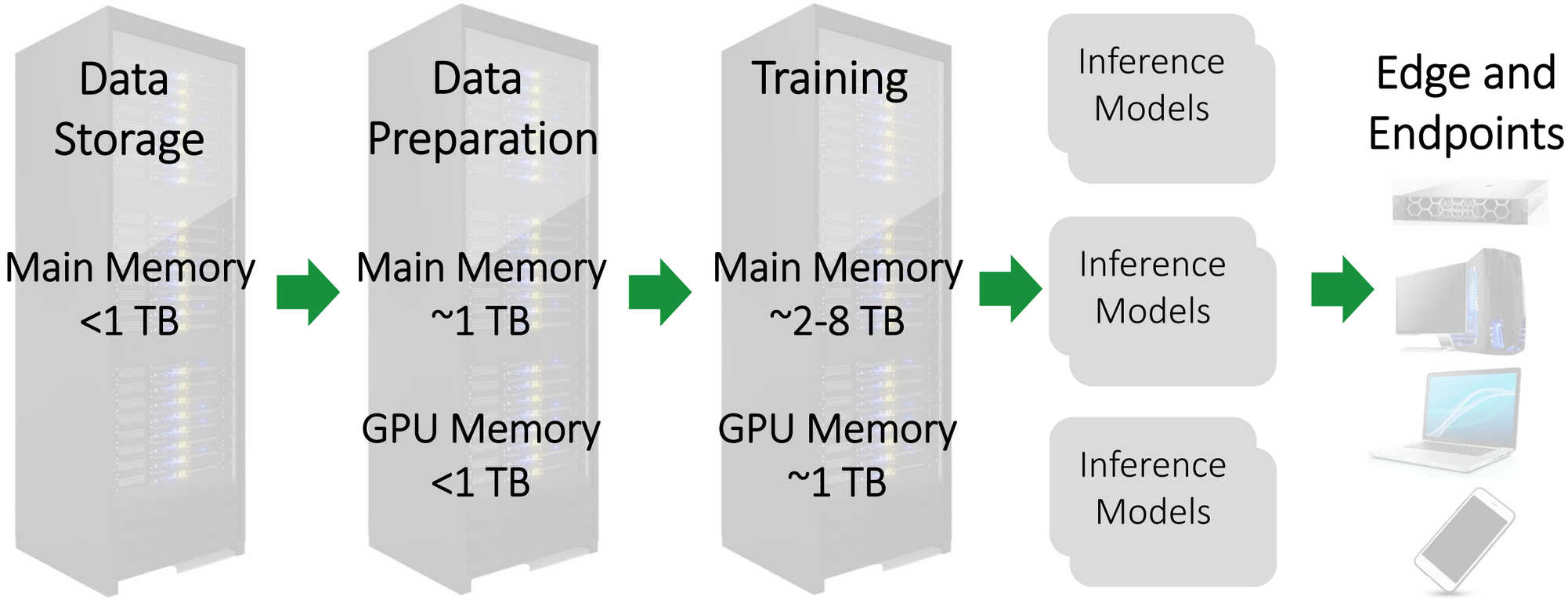
AI Depends on Memory Bandwidth and Capacity



Raja Koduri, https://lnkd.in/gZHHrFCZ

- AI becoming more capable through larger, more sophisticated models
- Dramatic advances in AI fueled by memory bandwidth and capacity
- Next-gen AI driving demand for even higher memory bandwidths and capacities

AI Training Pipeline: From Data Center to Edge and End Points



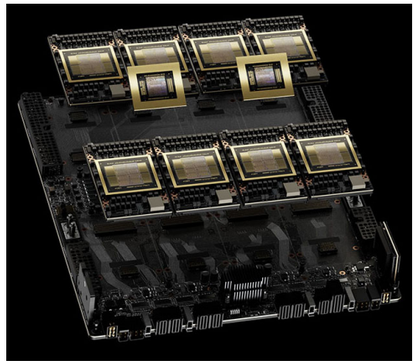
Increased memory bandwidth and capacity → larger models → better results more quickly

DDR plays a critical role in modern AI processing pipelines

Leading Edge AI Servers and Racks: HBM and DDR Critical

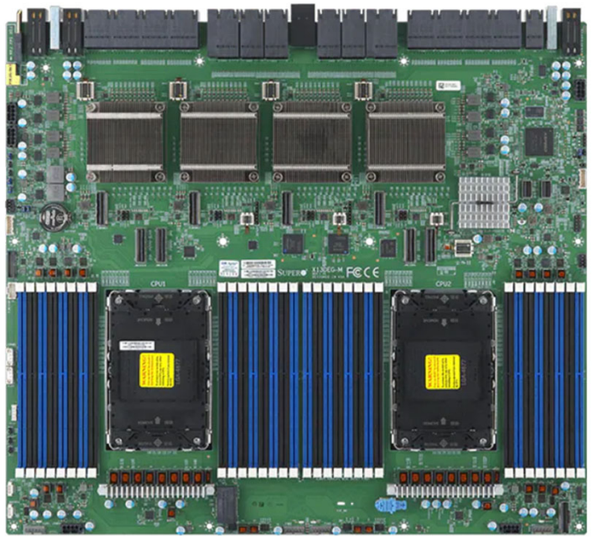
SuperMicro Generative AI Supercluster

8 - 4U Compute Nodes per Rack



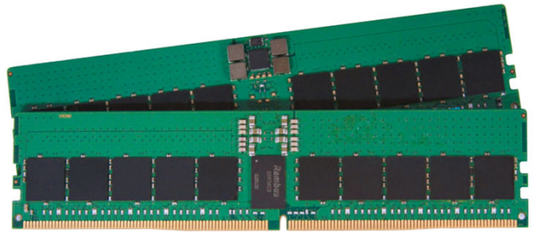
GPU Subsystem

- 8 NVIDIA H200 GPUs
- 1.1TB HBM3e
- 38.4TB/s bandwidth



CPU Subsystem today

- 2 Dual Socket Intel Xeon or AMD EPYC CPUs
- Up to 8TB DDR5



- AI demanding more memory bandwidth and capacity
- Physical limits on DIMM count, placement of DIMMs
- Need each DIMM to provide more bandwidth and capacity, but how?
- Strong desire to leverage DDR5 infrastructure to reduce cost, time to market

Modern Cloud Workloads and the New “Memory Wall”

Memory is Key For Cloud Infrastructure (Google MemCon 2023)

Key to new era is **memory**

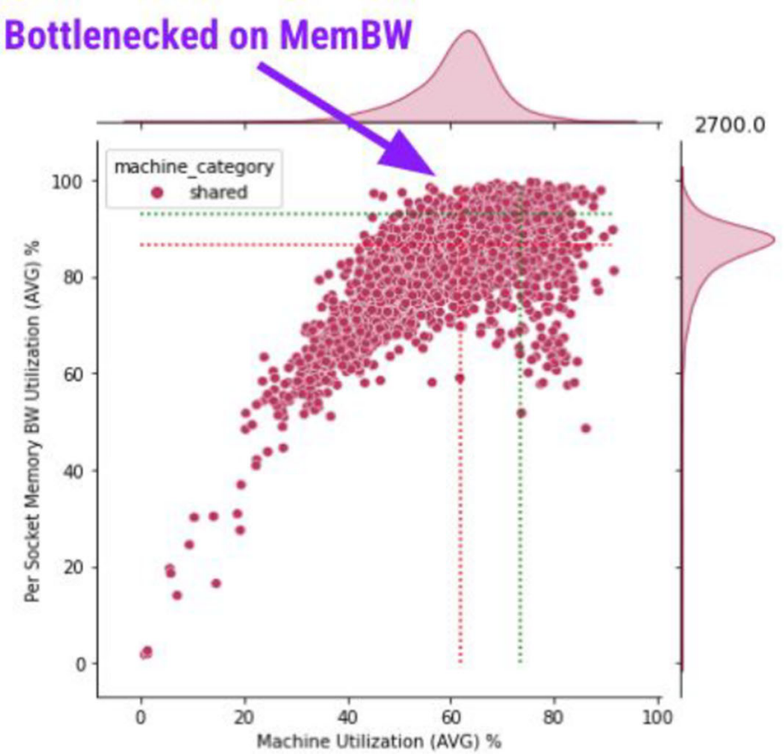
Foundational metric is Perf/TCO

Core to Memory Dilema: Memory BW not keeping up with core count increases.

CPU memory controllers lagging DDR frequency increases - latency penalties.

DDR4 -> DDR5, more BW is coming but cost overheads and latency are eroding value.

Superior Perf/TCO critical to drive mainstream adoption.



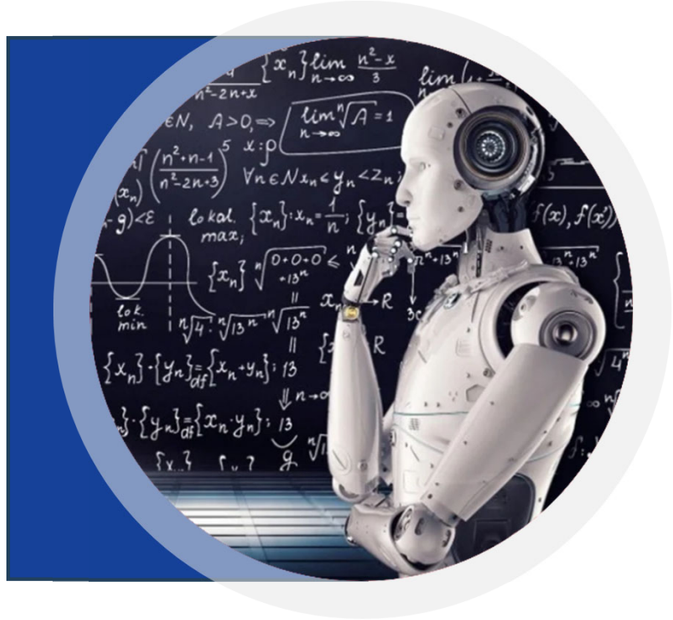
Memory BW Utilization at 85% when CPU utilization only 60%

- Core counts continuing to scale at fast pace
- CPUs bottlenecked on memory bandwidth
- More memory bandwidth and capacity needed for rising CPU core counts

Source: “The Renaissance in Datacenter Design: Delivering Modern and Scalable Solutions,” Tom Garvens, MemCon 2023

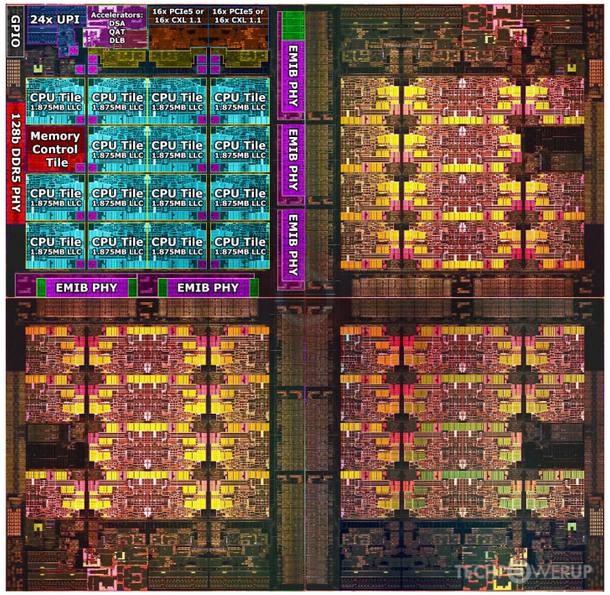
Key Drivers for Memory Bandwidth and Capacity

Advanced AI



- More human-like reasoning
- Intelligent Agents
- Robotics

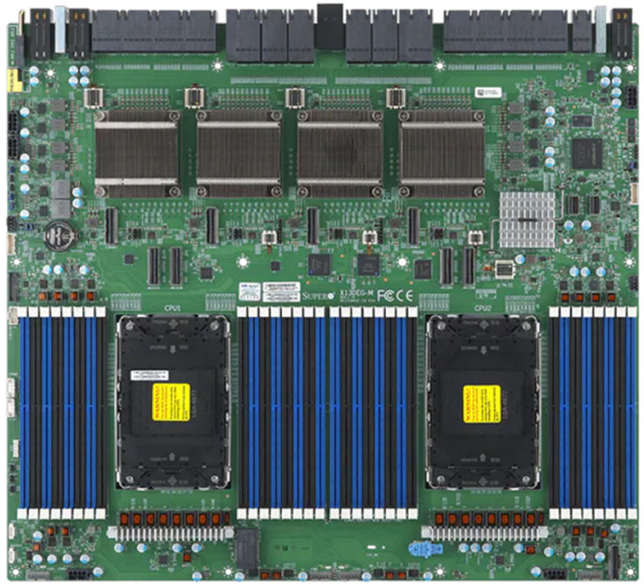
Growing CPU Core Counts



https://www.techpowerup.com/292204/intel-sapphire-rapids-xeon-4-tile-mcm-annotated

- Data Center workloads driving demand for more bandwidth

Physical Constraints



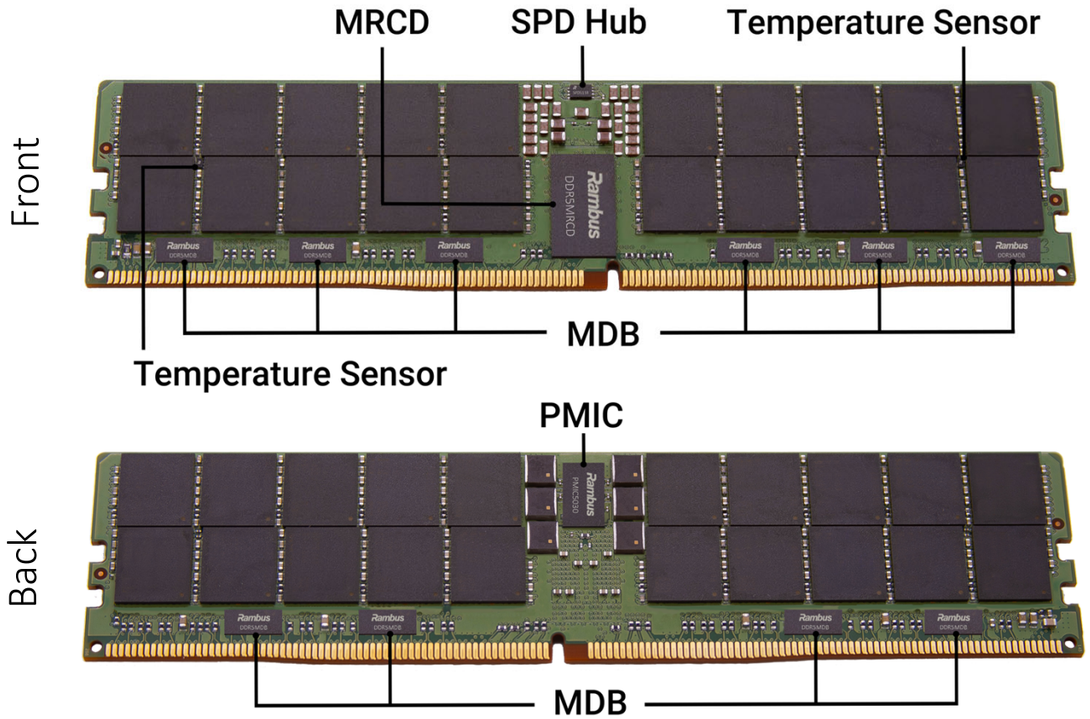
https://www.supermicro.com/en/products/motherboard/x14dbg-gd

- DIMMs must provide more bandwidth and capacity
- Leverage DDR5 infrastructure

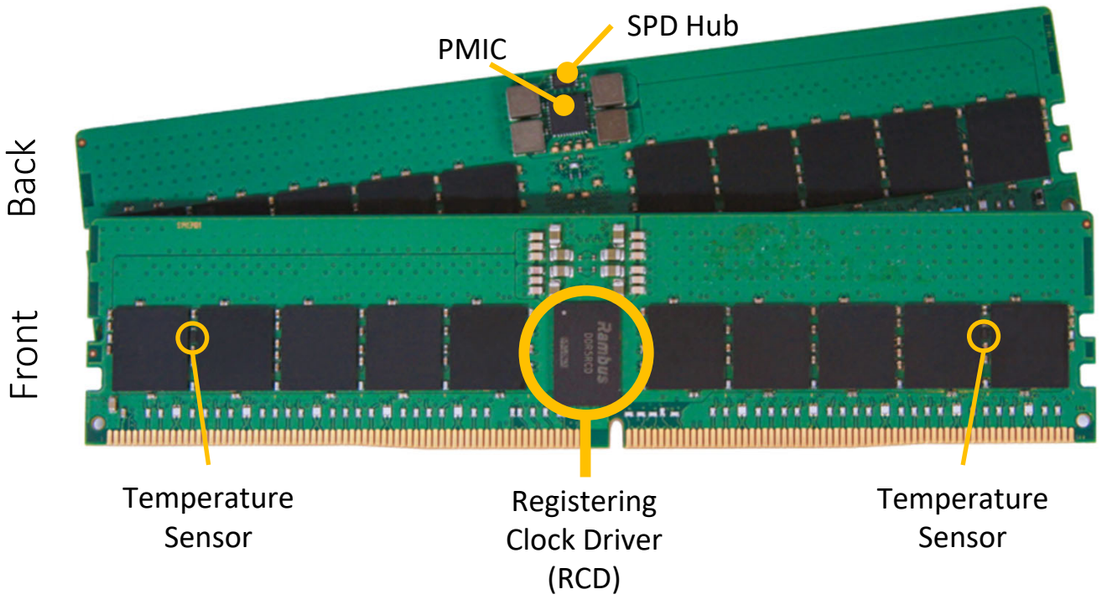
MRDIMMs provide greater memory bandwidth and capacity, leverage existing DDR5 DRAMs and infrastructure

MRDIMMs vs RDIMMs

MRDIMM



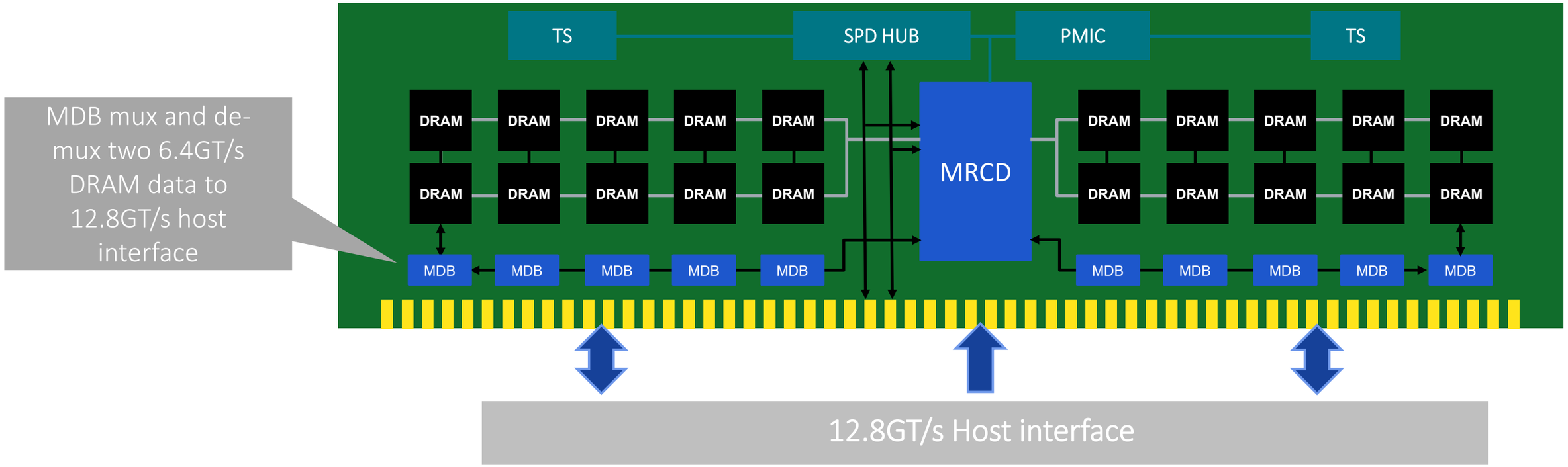
RDIMM



- MRDIMMs include a new component: 10 MDBs (data buffers)
- MRDIMMs also use a Registering Clock Driver (MRCD), a PMIC, an SPD Hub, and Temperature Sensors
- Same SPD Hub and Temperature Sensors used in MRDIMM and RDIMM
- Rambus provides a complete MRDIMM chipset with an enhanced PMIC for MRDIMM 12800

High Level Architecture of MRDIMM (Multiplexed Ranked DIMM)

with one MRCD and ten MDB components



Benefits of MRDIMM

Greater bandwidth per channel

- The DDR5 DRAM roadmap maxes out at 9200 MT/s
- MRDIMM extends the data rate to 12800 MT/s, with follow-on generations running 18000+ MT/s
- Host bus at 2X native speed of the DDR5 DRAMs

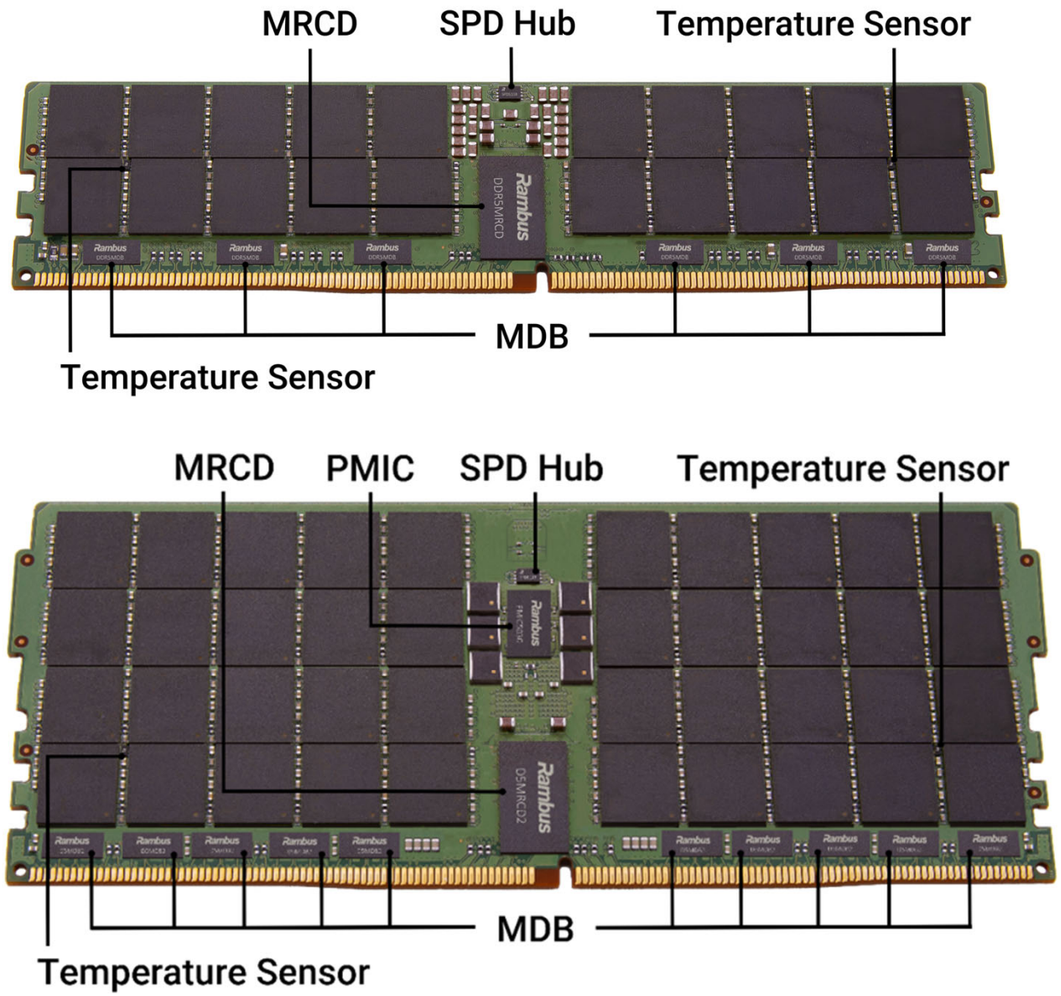
Greater capacity per channel

- MRDIMM supports 4 ranks in a tall DIMM form factor using standard DRAMs and 4 ranks in a standard DIMM form factor using DDP DRAMs

Leverages existing infrastructure

- Extends the DDR5 lifetime and leverages investment, seamless upgrade in memory bandwidth and capacity

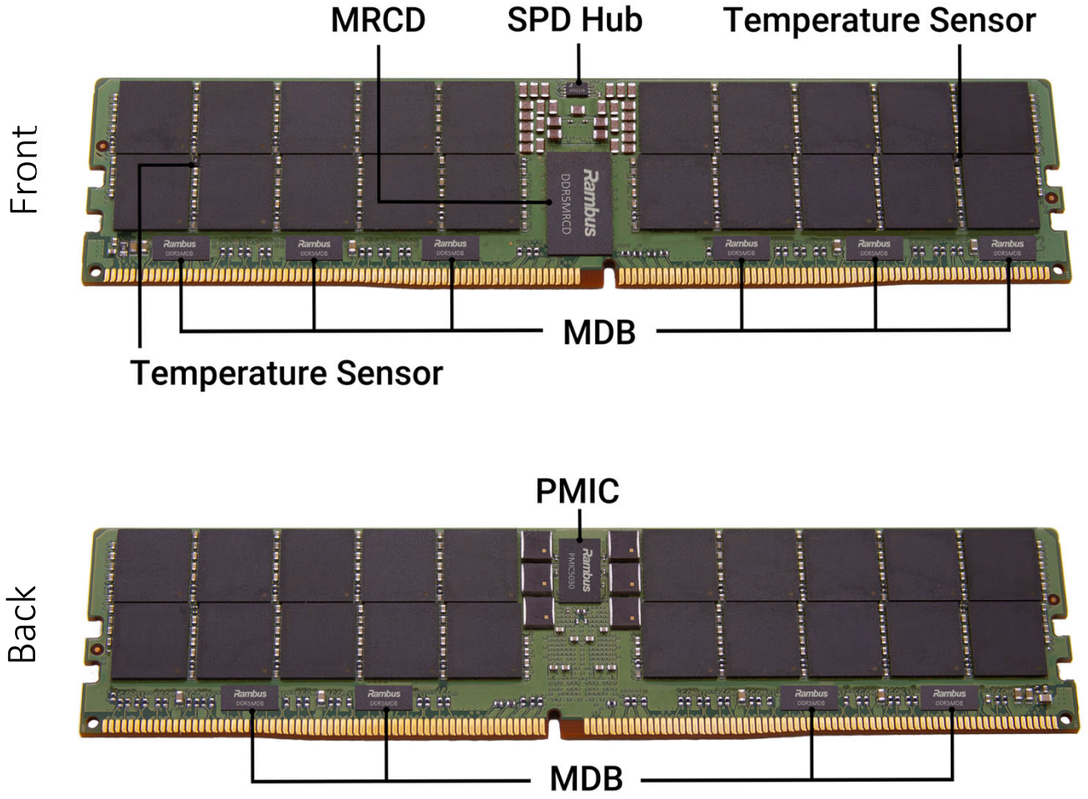
Standard and Tall DDR5 MRDIMM 12800



Rambus Offers Industry-First Chipset for Next-Generation DDR5 MRDIMMs to Deliver Breakthrough Performance for Data Center & AI

- Industry's first MRCD and MDB chips for next-generation MRDIMMs at 12,800 MT/s
- Next-generation server PMIC (PMIC5030) for MRDIMM and RDIMM 8000+
- Incorporates advanced clocking, control, and power management features needed for higher capacity and bandwidth modules
- Enables flexible and scalable end-user server configuration with compatibility across server platforms
- Feeds insatiable demand for higher memory performance in advanced data center and AI workloads

Rambus MRDIMM 12800 Chipset




Thank you