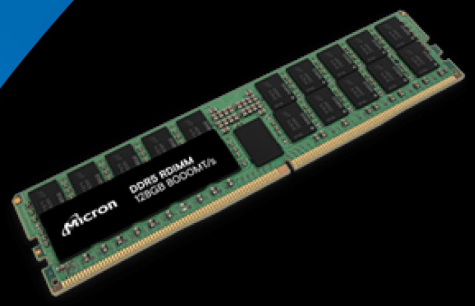
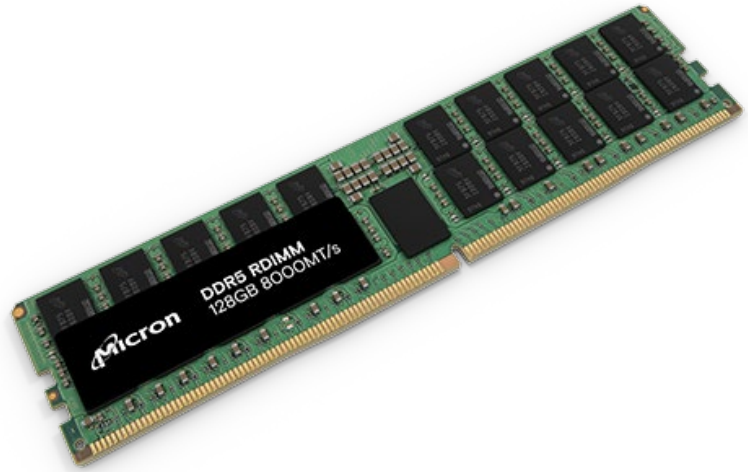
High-capacity DDR5 solution from Micron's leading-edge 1β (1-beta) technology



Micron is first to enable the market with a monolithic 32Gb-based DDR5 128GB DRAM module that delivers the fastest speed and lowest latency for enabling generations of server platforms to come.

Micron's DDR5 128GB RDIMM module, based on our 1β (1-beta) 32Gb technology, delivers high performance and better energy efficiency for memory-intensive applications including generative AI training and inference, real-time data analytics, and in-memory databases.

Built using Micron's industry-leading 1β process node technology, which uses advanced CMOS device technology, Micron DDR5 128GB RDIMM enables up to 6TB of system memory capacity and data rates up to 8,000MT/s.[1]



Key benefits

Fast Performance for AI in the data center

Micron high-capacity DDR5 delivers up to 28% faster performance for AI training[2].

Up to 16% improved latency[3]

Important for memory-bound workloads such as generative AI, in-memory databases, and real-time data analytics, where high-capacity is needed, and prompt response times are critical for real-time inference.

Highest Bandwidth DDR5 with capability up to 8000 MT/s

Low power and energy efficiency for datacenter workloads

>24% improved energy efficiency (pj/bit)[3]

Innovative 1β technology

- >45% improvement in wafer bit density using Micron's leading 1-beta technology based on 32Gb die, enabling the best bit density in the industry.[3]
- Micron's 1β 128GB RDIMM helps to balance CPU core counts with memory capacity, bandwidth, and power for optimized system performance, enabling future datacenter infrastructure.

1. Based on x86 systems with dual-socket, 12 channels per socket, and 2 DIMMs per channel.
2. 28% faster training time is a projected value at 8000MT/s based on empirical measurements of AI/ML model runs at different memory frequencies.
3. As compared with commercially available competitive 3DS modules.

Micron 1β enables generations of server platforms to come

Industry's fastest speed, lowest latency DDR5 based on a 32Gb monolithic die, Micron 1β will be available in 5600MT/s, 6400MT/s, 7200MT/s, and 8000MT/s, enabling future generations of server platforms. Micron's 1β 16Gb-based modules are available in 16GB, 32GB, and 64GB module densities for general compute applications not requiring larger capacities.

Micron DDR5 Modules

MT C 40 F 2 O 4 7 S 1 R C 80B B 1

Module Capacity

CODE	CAPACITY
5	32GB
6	64GB
7	128GB

Module Type

CODE	DESCRIPTION
R	288-pin RDIMM X80

DDR5 Module Speed

Module Speed Bin (Part Mark)	Component JEDEC Speed Bin
56B	DDR5-5600B
64B	DDR5-6400B
72B	DDR5-7200B
80B	DDR5-8000B

Contact Micron Field Sales Network for more information.

Product Specifications

Specification	Value
Data Rate	8000MT/s
Nominal Voltage (V_{DD}/V_{DDQ}/V_{PP})	1.1V/1.1V/1.8V
Burst Length	BL16, BC8
Bank Count (x4)	32
Module Configuration/Form Factor	2Rx4 RDIMM
DRAM Component Density	32Gb
Module Capacity	128GB
Power Consumption	10W @ 4800
Operating Temperature	0-95C
DRAM Component Package Size	7.5x11.5x1.0mm
32Gb tRFC1/tRFC2/tRFCsb	410/220/190ns
JEDEC Specification Compliance	JESD79-5B
JEDEC-Optional Feature Support	MBIST/mPPR
	ARFM/DRFM
	4-phase clocking
	Clock Sync
	Rx DQS CTLE

Improved AI inference and training

Micron's high-capacity, high-speed DDR5 128GB RDIMM enables and accelerates memory-bound workloads such as training for Large Language Models (LLMs). Micron DDR5 128GB RDIMM shows up to 28% improvement in AI training performance[4]. Llama 2 was chosen as a benchmark to evaluate the training and inference times because it best reflects modern LLMs (for example, ChatGPT).

Training for Llama 2

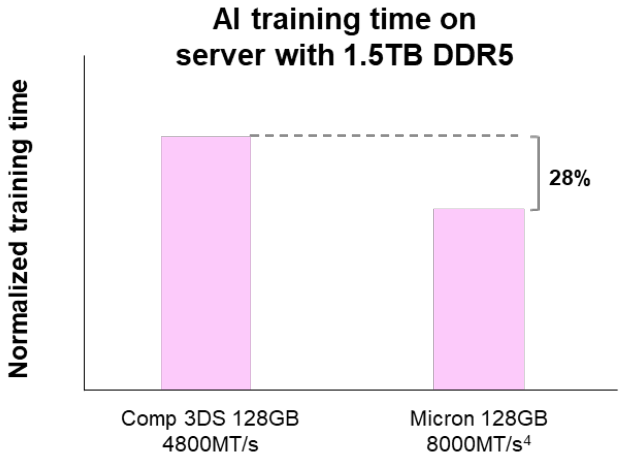


Figure 1. Llama 2 Training Time Projection

Inference for Llama 2

For inference and power consumption results using Llama 2, 3DS TSV RDIMM modules show up to 48% higher power when compared to Micron's DDR5 128GB RDIMM.

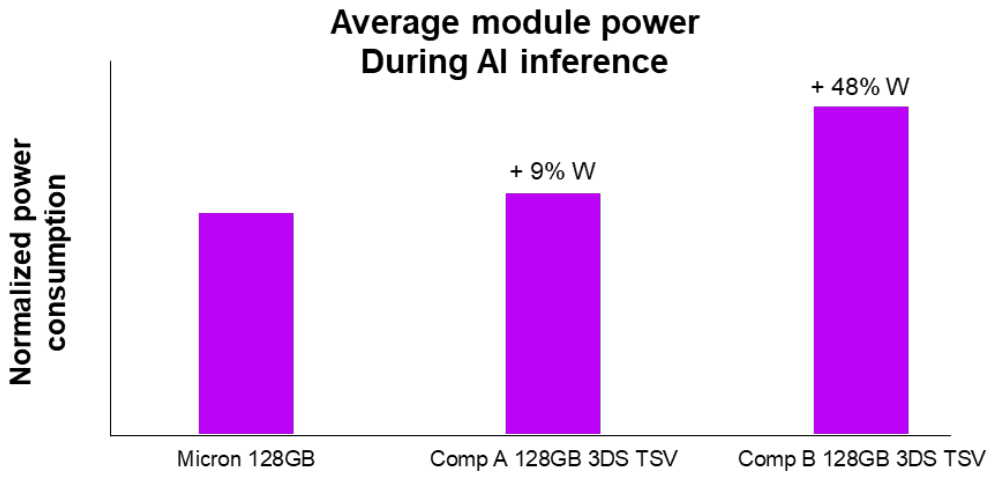


Figure 2. Llama 2 Inference and Average Module Power

4. The 28% training time projection at 8000MT/s is based on empirical measurements of AI/ML model runs at different memory frequencies.

micron.com/ddr5