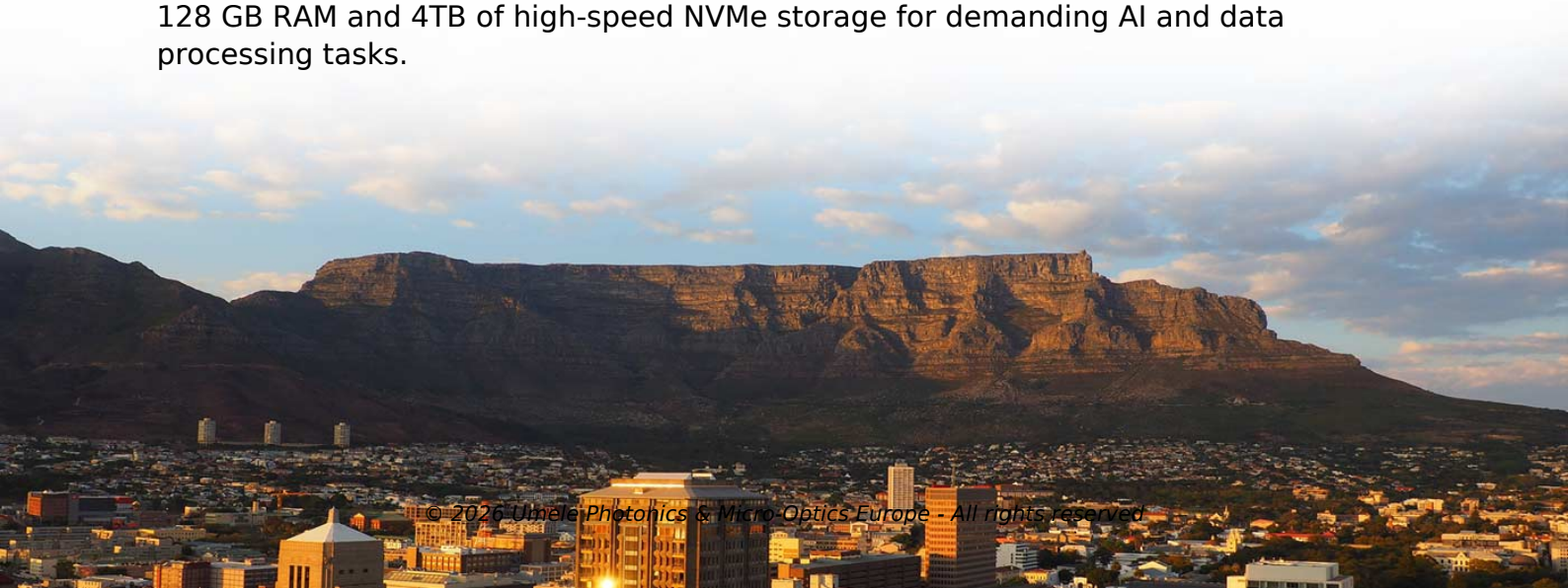
# AI server with 128GB of memory

## Overview

All About AI takes a closer look at how the AMD Ryzen AI Pro chip, paired with a staggering 128GB of RAM, is making this a reality in 2026. Imagine running advanced language models, generating Python code, or analyzing images, all locally, without relying on cloud servers. The launch lands at a precarious moment for hyperscale AI. The top five hyperscalers spent $443 billion on capital expenditures in 2025, a number expected to climb 36% to $602 billion in 2026, much of it pouring into Nvidia GPUs whose 192 GB of HBM3e per chip is increasingly the limiting factor for. Majestic Labs introduced Prometheus, an AI server engineered to address the “memory wall,” a growing constraint in scaling modern AI workloads. The system adopts a memory-first architecture that connects significantly larger pools of high-speed memory directly to processing elements, aiming to. Powered by the NVIDIA GB10 Grace Blackwell Superchip, NVIDIA DGX Spark™ delivers up to one petaFLOP1 of FP4 AI performance in a power-efficient, compact form factor. This shift isn't just. In our tests, an RTX 5090 (December 2025 market price around $2800 just for the GPU) tops out around a 32B model with a 45K context in stable configurations, mainly due to VRAM limits. A compact unified-memory system with 128GB (around $2000) can load models up to roughly 120B in 4-bit with room. Configure your server with up to 128 GB RAM and 4TB of high-speed NVMe storage for demanding AI and data processing tasks.



## Article Content

Majestic Labs Targets AI Memory Bottleneck with 128TB Server

The Prometheus platform integrates up to 128 TB of shared, contiguous memory within a single standard-sized server, enabling execution of large-scale AI models that typically require

ASUS Ascent GX10□Desktop AI supercomputer□ASUS Global

The ASUS Ascent GX10, featuring 1 petaFLOP of AI performance and 128GB of memory, empowers you to run agentic AI (such as OpenClaw 1 or other frameworks) on-demand.

As consumer RAM prices soar, Micron unveils an eye-watering 256

Case in point, I am sweating just thinking about Micron''s 256 GB DDR5 server module. The Boise, Idaho-based memory manufacturer unveiled the tech on Tuesday.

Best Unified Memory Computers for Local LLMs (2025): Bandwidth, Memory ...

Since local LLM performance is shaped by three main factors – how much memory you have for the model, how much bandwidth you can push during inference, and how much GPU

Beelink | Beelink GTR9 Pro AMD Ryzen™ AI Max+ 395

□□ □Unmatched Memory & Storage□ — Packed with 128GB LPDDR5X-8000 RAM and dual M.2 2280 PCIe 4.0 slots that support up to 16TB (each slot supports up

ASUS Ascent GX10□Desktop AI

Extreme AI Performance: Powered by NVIDIA ® GB10 Grace Blackwell Superchip delivering 1 petaFLOP of AI performance and 128GB memory for 200B model

SQL Server 2025: The Top Concerns IT Teams Should Plan For

SQL Server 2025 reached general availability on November 18, 2025. When we first wrote about common SQL Server concerns back in 2022, the list was mostly evergreen: indexes,

Huawei Launches Atlas 350 AI Accelerator with Ascend 950PR

Huawei launched the Atlas 350 AI accelerator with Ascend 950PR, 128GB memory, FP4/FP8 computing, and partners released servers for AI and industry applications.

Lemonade Frequently Asked Questions

Lemonade Server is a lightweight, open-source local LLM server that allows you to run and manage multiple AI applications on your local machine. It provides a simple CLI for managing applications

Micron First to Ship Critical Memory for AI Data Centers | Micron ...

Micron reaches industry milestone as first to validate and ship 128GB DDR5 32Gb server DRAM to address the rigorous speed and capacity demands of memory-intensive Gen AI

Majestic Labs Prometheus: 128TB AI Server

Prometheus is a single-node AI server announced on April 28, 2026, with up to 128 TB of shared memory and proprietary AIU compute, designed to run trillion-parameter language models,

Micron Raises DDR5 Server Memory with 256 GB Modules

Why is the 256 GB per module significant? Because it allows increasing memory per socket and server without occupying more slots, which is beneficial in AI, HPC, in-memory databases, and dense

NVIDIA Expands AI Portfolio to Counter ASIC Push

TrendForce reports that as CSPs like Google and Amazon increase their internal chip development, ASIC-based AI servers are forecast to represent 27.8% of all AI server shipments in

Ryzen AI Max Dedicated Servers | AI-Driven Cloud Infrastructure

Ryzen AI servers support up to 128 GB RAM for handling large models and datasets. Storage is expandable up to 4TB using dual M.2 NVMe slots, providing high-speed I/O for training data and

Local AI Laptop 2026 Setup : AMD 128GB Rig Tested Across GPT

In 2026, advanced AI models can now run locally on laptops equipped with AMD Ryzen AI Pro chips and 128GB of RAM, eliminating reliance on cloud services and enhancing privacy,

Best CPUs for AI Workflows in 2026: Ryzen AI, Core Ultra

When choosing the best CPU for AI workstation, you are presented with 4 classes of CPU: 1. AMD Threadripper PRO: High-Core Workstation CPU for AI Engineering 2. AMD EPYC: The

Dell 128GB 370-BCJN DDR5-5600MT/s ECC Reg DIMM Server Memory

Dell 128GB 370-BCJN DDR5-5600MT/s ECC Reg DIMM Server Memory Samsung OEM New | BRAND NEW Part Number: 370-BCDG Condition: BRAND NEW #### **DELL 370-bcdf 128gb (8x16gb) Ddr5

Tested: AMD''s Strix Halo vs Nvidia''s DGX Spark

In single-user scenarios, token generation is usually bottlenecked by memory bandwidth. The GB10 claims about 273 GB/s of memory bandwidth while AMD''s Strix Halo manages about 256

IBASE Releases ES1002 Edge AI Server Powered by AMD EPYC

IBASE Technology Inc., a leading manufacturer of embedded and edge computing solutions, unveils the ES1002 Edge AI Server, a high-performance platform engineered to accelerate

Personal AI Supercomputer Powered by Blackwell | NVIDIA DGX Spark

Fifth-generation Tensor Cores with support for FP4 deliver up to 1 petaFLOP of AI computing performance, combined with 128GB of system memory, accelerate inference of state-of-the-art AI

## Contact Us

For more information, pricing, or custom solutions, please contact us:

Website: https://umele-pestovani.eu
Email: info@umele-pestovani.eu
Phone: +49 152 287 6349
Address: Neue Mainzer Straße 66-68, 60311 Frankfurt am Main, Germany

This document is for informational purposes only. Specifications subject to change without notice.