LLM-2U-AM5 & LLM-3U-AM5 Series

2U & 3U SHORT-DEPTH EDGE AI RACKMOUNT SERVERS



SCALE PRIVATE AI WITH SHORT-DEPTH EDGE AI SERVERS

The LLM-2U-AM5 and LLM-3U-AM5 are high-performance, short-depth Edge AI rackmount servers engineered for single-node, on-premises AI deployments. Powered by AMD EPYC™ and Ryzen™ processors, these servers combine high-capacity DDR5 memory, PCIe Gen5 expansion, and professional GPU acceleration to support private LLM inference, retrieval-augmented generation (RAG), and Agentic AI workloads.

Available in compact 2U and expansion-oriented 3U configurations, the LLM-2U-AM5 & LLM-3U-AM5 Series Series delivers flexible computing performance and high-speed networking for deploying dedicated AI infrastructure closer to enterprise and industrial data sources.

- Short-Depth Rackmount Design
- Single-Node AI Performance
- GPU-Accelerated AI Inference
- High-Speed Edge Networking

Product Brief
Edge AI Rackmount Servers
LLM-3U-AM5, LLM-2U-AM5



POWERING AI AT THE ON-PREM DATA CENTER EDGE

As AI applications move beyond centralized cloud infrastructure, enterprises and industrial organizations increasingly require dedicated computing resources closer to their operational data.

The LLM-2U-AM5 and LLM-3U-AM5 are purpose-built for the on-premises data center edge, delivering GPU-accelerated single-node computing for private LLMs, RAG, and intelligent AI applications. Positioned between distributed edge devices and centralized cloud resources, these short-depth rackmount servers enable local AI processing while reducing cloud dependency, network bandwidth requirements, and sensitive data exposure.

Their compact 2U and 3U designs make them suitable for micro data centers, enterprise server rooms, and distributed industrial facilities requiring dedicated Edge AI infrastructure.

Key Features

- Supports AMD EPYC 4005/4004 & Ryzen™ 9000 Series Processors
- AMD B650 Chipset
- 4x DDR5 UDIMM, up to 5200 MT/s, Max. 192GB (Non-ECC)
- 2x 10GbE (SFP+), 2x GbE LAN
- Dual 4K Display: 1x DP, 1x HDMI
- 2x PCIe x16 Gen 5, 1x PCIe x1
- 2x M.2 M-Key, 1x M.2 E-Key
- 2U: 4x 2.5" SATA SSD Bays; 3U: 4x 2.5" + 1x 3.5" SATA Drive Bays (support RAID 0/1/5/10)
- 6x USB 3.2 Gen 2 (Rear), 2x USB 3.2 (Front)
- Optional Out-of-Band Management
- 850W Single Power Supply
- Support TPM 2.0
- UL 62368 Ed. 4, CE, FCC

Private LLM & Generative AI

Retrieval-Augmented Generation (RAG)

Agentic AI

Physical AI & Industrial Automation

Smart Infrastructure & Security

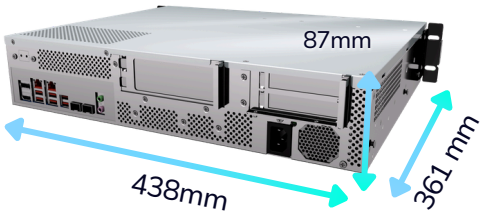
Short-Depth Rackmount Design for Space-Efficient AI

The LLM-2U-AM5 and LLM-3U-AM5 feature short-depth rackmount designs that bring dedicated Edge AI computing into space-constrained on-premises environments. With a chassis depth of just 361 mm, both servers provide GPU-accelerated computing in compact, standard 19-inch rackmount configurations.

The 2U model prioritizes space efficiency for balanced single-node AI deployments, while the 3U model offers additional chassis height for greater physical expansion flexibility.

Both models are suited for micro data centers, industrial control rooms, and enterprise server installations where rack space is limited.

LLM-2U-AM5



LLM-3U-AM5

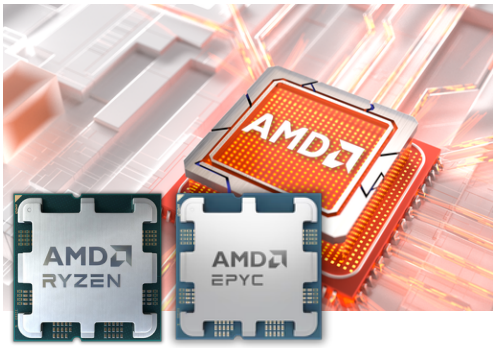


Key Features

- LLM-2U-AM5: 438 (W) x 361 (D) x 87 (H) mm
- LLM-3U-AM5: 438 (W) x 361 (D) x 132 (H) mm
- Standard 19-inch rackmount installation
- Sliding-rail installation support

Scalable AMD EPYC™ & Ryzen™ Performance for Edge AI

Supporting AMD EPYC™ 4005/4004 and Ryzen™ 9000 Series processors, the LLM-2U-AM5 and LLM-3U-AM5 deliver flexible multicore performance for single-node Edge AI rackmount servers. Built on the AMD B650 chipset, these servers support compute-intensive workloads, from private LLM inference and generative AI to enterprise RAG and multi-workload processing in on-premises environments.



The AMD AM5 architecture delivers:

- Up to 16 cores / 32 threads, processor dependent
- EPYC™ options for server-oriented and multi-workload deployments
- Ryzen™ options for high-performance Edge AI and inference workloads
- PCIe Gen 5 and DDR5 support for scalable system performance

High-Capacity DDR5 Memory for Data-Intensive AI

Equipped with four DDR5 5600 UDIMM slots (up to 5600 MT/s), the LLM-2U-AM5 and LLM-3U-AM5 support up to 192GB of system memory, providing greater headroom for multitasking, virtualization, AI inference, and other memory-intensive workloads.



GenAI Acceleration with Dedicated GPU Support

For on-premises LLM inference and generative AI workloads, dedicated GPU acceleration is essential. The LLM-2U-AM5 and LLM-3U-AM5 support a validated range of NVIDIA® RTX PRO™ Blackwell workstation GPUs, from the compact RTX PRO 2000 Blackwell to the high-performance RTX PRO 6000 Blackwell Max-Q Workstation Edition. Combined with PCIe Gen 5 expansion, these GPUs provide scalable acceleration and GPU memory capacity for private LLMs, RAG, and multimodal AI workloads in single-node Edge AI rackmount servers.

GPU Model	GPU Memory	Power
NVIDIA RTX PRO 2000 Blackwell	16GB	70W
NVIDIA RTX PRO 4000 Blackwell SFF	24GB	70W
NVIDIA RTX PRO 4000 Blackwell	24GB	140W
NVIDIA RTX PRO 4500 Blackwell	32GB	200W
NVIDIA RTX PRO 5000 Blackwell	48GB	300W
NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition	96GB	300W

Flexible PCIe Gen5 Expansion for AI Workloads

The LLM-2U-AM5 and LLM-3U-AM5 feature four PCIe expansion slots to support high-performance GPUs, AI accelerators, network adapters, and storage controllers. With configurable PCIe Gen5 lane allocation, these single-node Edge AI rackmount servers provide flexible expansion for evolving AI inference and data-intensive workloads.

PCIe Expansion Specifications

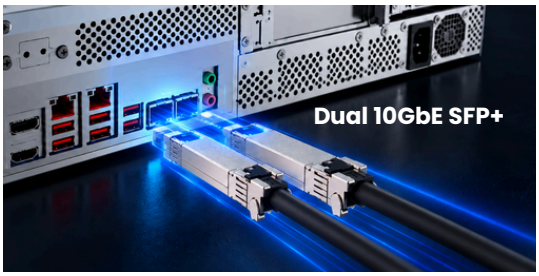
- Slot #1: 1x PCIe x16 (16-lane or 8-lane,)
- Slot #2: 1x PCIe x16 (8-lane or No Signal)
- Slot #3: 1x PCIe x4 (1-lane, Open End)



High-Bandwidth Networking with Dual 10GbE SFP+

Designed for high-speed data transmission between distributed edge devices, networked storage, and on-premises AI infrastructure, the LLM-2U-AM5 and LLM-3U-AM5 integrate dual 10GbE SFP+ ports and dual one-Gigabit Ethernet ports.

Powered by the Broadcom BCM57412 controller, the dual 10GbE interfaces support fiber-optic or direct-attach networking, delivering high-throughput connectivity for private LLM inference, RAG applications, and data-intensive Edge AI workloads.



Key Specifications

- 2x 10GbE SFP+ for fiber or direct-attach networking
- 2x 1GbE RJ45 for standard Ethernet connectivity
- Broadcom BCM57412 10GbE controller

Flexible Local Storage for AI Data & Models

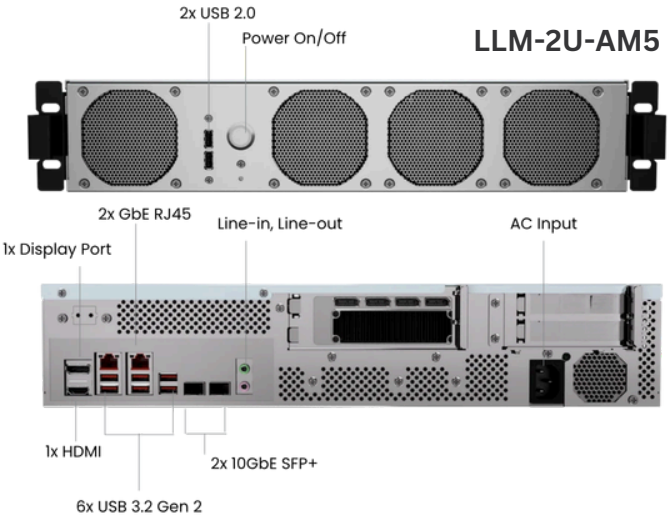
Local storage supports on-premises AI operations by keeping model files, vector databases, application data, and retrieval resources close to the inference server.

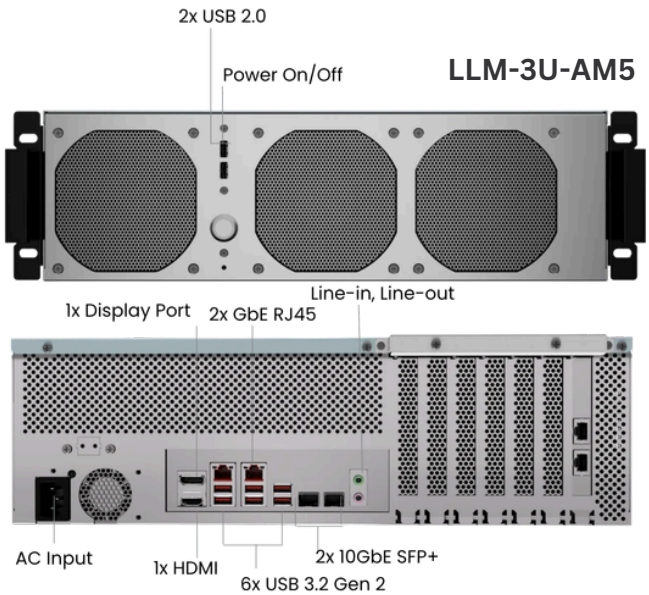
The LLM-2U-AM5 and LLM-3U-AM5 combine high-speed M.2 NVMe storage with SATA drive bays, allowing organizations to configure storage according to their AI application requirements.

The 2U model provides dual internal 2.5-inch SATA bays,
while the 3U model accommodates both 2.5-inch and 3.5-inch drives for added storage configuration flexibility.

Key Specifications

- 2x M.2 M-Key 2280 for PCIe NVMe SSDs
- 1x M.2 E-Key 2230 for wireless expansion
- SATA RAID 0/1/5/10 support







LLM-2U-AM5

- 2x 2.5" or 4x 2.5" SATA SSD Drive Bays



LLM-3U-AM5

- 2x 2.5" + 1x 3.5" or 4x 2.5" or 1x 3.5" SATA Drive Bays



Remote Management for Distributed Edge AI Servers

Managing Edge AI servers across multiple locations can increase maintenance costs and operational complexity. The LLM-2U-AM5 and LLM-3U-AM5 offer optional out-of-band (OOB) management, enabling IT administrators to remotely monitor system health, diagnose issues, and perform maintenance even when the operating system is unresponsive. This helps minimize on-site service visits, reduce downtime, and maintain business continuity across distributed Edge AI deployments.

Active Cooling for sustained Edge AI Workloads

The LLM-2U-AM5 and LLM-3U-AM5 feature dedicated CPU cooling and internal system fans to efficiently manage heat during continuous AI inference and data processing. Combined with a heavy-duty metal chassis, the active cooling design supports stable performance for sustained GPU-accelerated workloads.





Integrated 850W Power Supply

The LLM-2U-AM5 and LLM-3U-AM5 feature an integrated 850W single power supply to support compatible processor, GPU, storage, and PCIe expansion configurations.

This integrated power design simplifies deployment of dedicated single-node AI computing within standard rackmount environments.

Industrial Longevity, Security & Safety Compliance

Built for long-term Edge AI deployments, the LLM-2U-AM5 and LLM-3U-AM5 combine embedded lifecycle support with TPM 2.0 security to help protect sensitive data and support dependable on-premises operations. Designed to meet UL 62368 Ed. 4 safety standards, CE requirements, and FCC Class B, these rackmount servers support reliable integration into industrial and enterprise AI infrastructure.

UL CERTIFIED
SAFETY US·CA
E357184

CE FC

We Design, Manufacture, and Service Customers Around the World

CALIFORNIA (HQ)
CHICAGO
TAIWAN (HQ)

LLM-2U-AM5 & LLM-3U-AM5 Series

Short-Depth Edge AI Rackmount Servers





Model	LLM-2U-AM5	LLM-3U-AM5
Form Factor	2U Edge AI Rackmount Server	3U Edge AI Rackmount Server
Processor	AMD EPYC™ 4005/4004 & Ryzen™ 9000 Series Processors (Max 170W)	
Chipset	AMD B650	
Memory	4x DDR5 UDIMM, up to 5200 MT/s, Max. 192GB (Non-ECC)	
PCIe Expansion	Slot #1: 1x PCIe x16 (16-lane or 8-lane,) Slot #2: 1x PCIe x16 (8-lane or No Signal) Slot #3: 1x PCIe x4 (1-lane, Open End) (2 Full-Height, 2 Low-Profile Expansion Slots)	Slot #1: 1x PCIe x16 (16-lane or 8-lane,) Slot #2: 1x PCIe x16 (8-lane or No Signal) Slot #3: 1x PCIe x4 (1-lane, Open End) (7 Full-Height Expansion Slots)
Expansion	2x M.2 M-Key, 1x M.2 E-Key	
SATA Storage	2x or 4x 2.5" SATA SSD Bays	2x or 4x 2.5" SATA SSD Bays + 1x 3.5" SATA Bay
Networking	2x 10GbE SFP+, 2x GbE RJ45	
Display	Dual Display: 1x DP, 1x HDMI	
USB	6x USB 3.2 Gen 2 (Rear), 2x USB 2.0 (Front)	
Remote Management	Optional Out-of-Band Management	
Security	Support TPM 2.0	
Power Supply	850W Single Power Supply	
System Cooling	4x System Fan, 1x CPU Fan, 1x PSU Fan	3x System Fan, 1x CPU Fan, 1x PSU Fan
Operating Temperature	0°C to 35°C (with 0.6 m/s airflow)	
Vibration	1 Grms (5–500 Hz, 0.5 hr/axis), with SSD	
Certifications	UL 62368 Ed. 4, CE, FCC	
Dimensions (W x D x H)	438 x 361 x 87 mm	438 x 361 x 132 mm
Mounting	Standard 19" Rackmount	

USA 918 Radecki Ct., City of Industry, CA 91748
TW 176 Jianyi Rd., Zhonghe Dist., New Taipei City
800.977.3646
886-2-7731-7888
sales@premioinc.com
premioinc.com