EINPRESSWIRE

Premio Expands its Edge Computing Hardware Portfolio with New LLM Series Edge Servers

1U Edge Server LLM-1U-RPL designed to enable real-Time GenAI and LLM Workloads for on-premises data centers at the edge

LLM-1U-RPL Series, 1U Short-Depth Edge AI Server

CITY OF INDUSTRY, CA, UNITED STATES, July 7, 2025 /EINPresswire.com/ -- Premio Inc., a global leader in rugged edge AI computing and industrial display technology, announces the release of the LLM-1U-RPL Series, the first model in its new LLM Series line of edge servers. This compact, short-depth 1U edge server is designed to bring real-time Generative AI (GenAI) and Large Language Model (LLM) workloads directly to the on-premises data center edge. This new series is engineered for demanding IT/OT enterprise deployments that require better performance through lower latency inferencing and data processing closer to its source of data generation. The LLM-1U-RPL Series addresses the growing demand for more on-premises AI capabilities, moving beyond traditional cloud reliance to provide reduced bandwidth strain, safeguarded data sovereignty, and support for real-time decisions in hybrid cloud environments now at the edge.

> The LLM-1U-RPL is purpose-built for on-premise data centers to deliver high-performance, low-latency AI inferencing for LLM workloads—without the need for traditional centralized cloud resources.”
>
> *Dustin Seetoo, VP of Product Marketing*

“Designed for the demands of edge deployments, this new edge server integrates 13th Gen Intel® Core™ processors with performance-hybrid architecture, dedicated NVIDIA GPUs for accelerated computing, and industrial-grade power redundancy—key capabilities that enable real-time intelligence, reduce latency, and give organizations greater control over their data." says Dustin Seetoo, VP of Product Marketing at Premio.

LLM-1U-RPL Key Features:
- Short-Depth 1U Rackmount Design (483 (W) x 480 (D) x 44 (H) mm)
- 13th Gen Intel® Core™ Processors
- Supports up to an NVIDIA RTX 5000 Ada GPU for accelerated computing
- PCIe Gen 4 Expansion for GPU AI accelerators or high-throughput network cards
- Flexible and High-Speed Storage Option in m.2 NVME and dual hot-swappable 2.5” SATA bays
- Optimized I/O Connectivity for On-Premises Edge AI: 3x 2.5GbE LAN ports, 6x USB 3.2 Gen2 ports, and COM ports
- 600W (1+1) redundant power supply
- Hot-swappable redundant smart fans
- Enhanced Cybersecurity and Physical Security
- World-Class Certifications (UL, FCC, CE)

The LLM-1U-RPL is engineered to bring low-latency inferencing directly to the edge, where time-sensitive decisions must be made in real-time. The edge server is powered by 13th Gen Intel® Core™ processors (up to i9, 65W TDP), leveraging a performance hybrid architecture with P-cores for low-latency inferencing like LLM prompt response and token generation and E-cores for general-purpose background applications. It also supports up to 64GB of dual-channel DDR4 3200MT/s SODIMM memory for streamlining multi-modal data streams without performance bottlenecks.

For local storage options it includes high-speed NVMe via an M.2 M-Key slot and front-accessible dual hot-swappable 2.5” SATA bays. The local storage capability reduces reliance on cloud resources, eases last mile backhaul bandwidth usage, and accelerates response times, while also enhancing data privacy and sovereignty

Designed with performance flexibility in mind, the server also supports PCIe Gen 4 expansion slots for high-throughput network interface cards (NIC) or a dedicated AI GPU accelerator. With compatibility for up to an NVIDIA RTX™ 5000 Ada, the system enables high-performance inferencing for private, on-prem LLM deployments, such as digital twins and generative AI inferencing, to minimize cloud dependency and preserve data sovereignty.

In addition to performance, the LLM-1U-RPL is designed for long-term reliability and secure operation in a compact, short-depth 1U form factor. Redundant power supplies and hot-swappable fans enable continuous 24/7 uptime and simplified maintenance. Security features such as a tamper-resistant front bezel, chassis intrusion detection, and TPM 2.0 help safeguard sensitive data in regulated or privatized environments.

The LLM-1U-RPL is ideal for a range of key markets and Industry 4.0 applications that demand local AI processing, from manufacturing automation and robotics to smart infrastructure and security. Its ability to bring generative AI workloads closer to the data source helps reduce cloud exposure and ensures compliance with evolving data governance standards. Overall, the LLM-1U-RPL serves as a scalable, on-premises edge computing node, bridging real-time AI processing with in-the-field devices across Industry 4.0, mobility, and intelligent infrastructure deployments.

For more information about Premio's edge AI server, LLM-1U-RPL Series, contact our embedded and edge computing experts at sales@premioinc.com or visit www.premioinc.com.

Dustin Seetoo
Premio Inc.
+1 626-839-3100
email us here
Visit us on social media:
LinkedIn
Facebook
YouTube