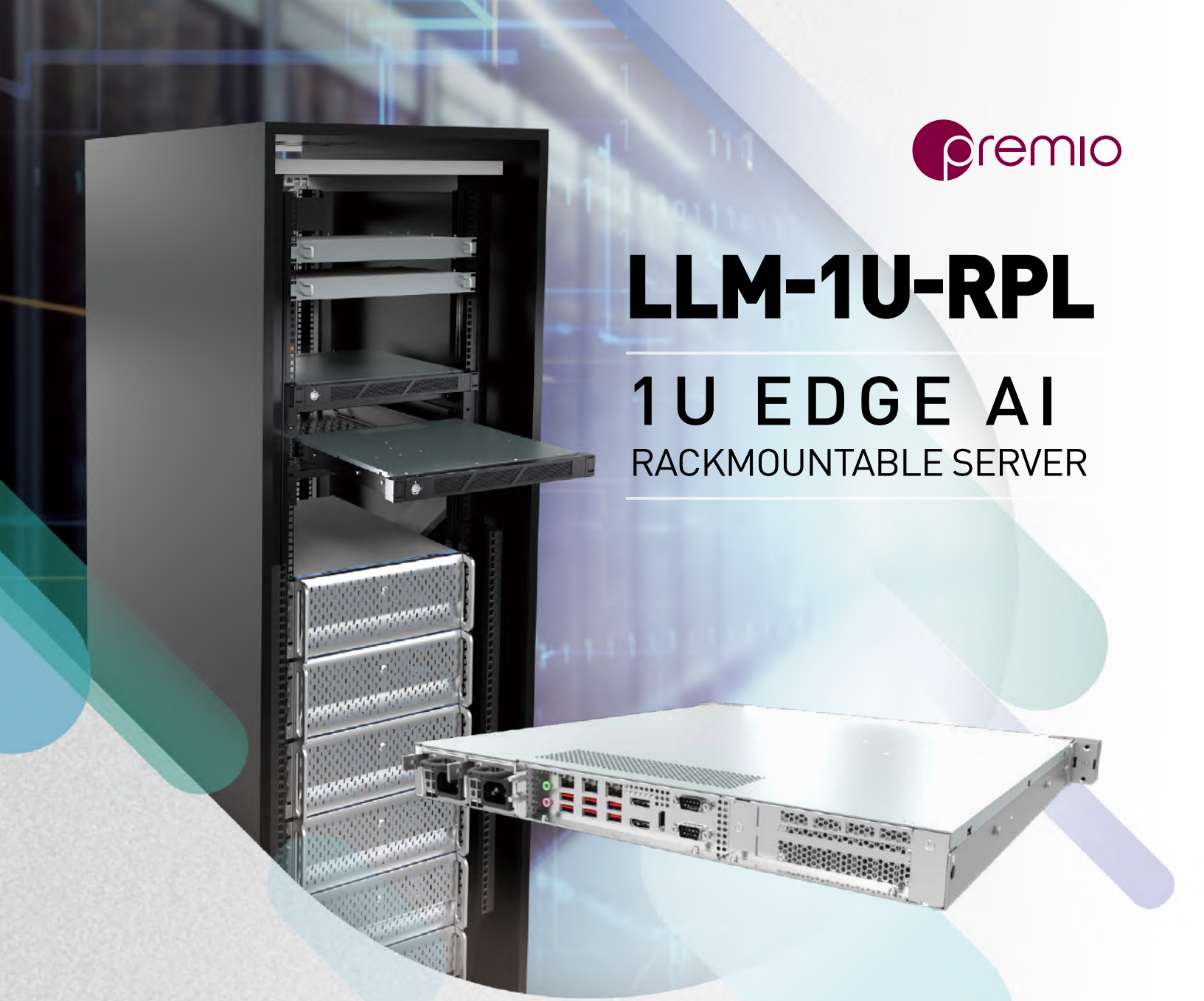
premio

LLM-1U-RPL

1U EDGE AI
RACKMOUNTABLE SERVER



SCALE GEN AI AT THE EDGE

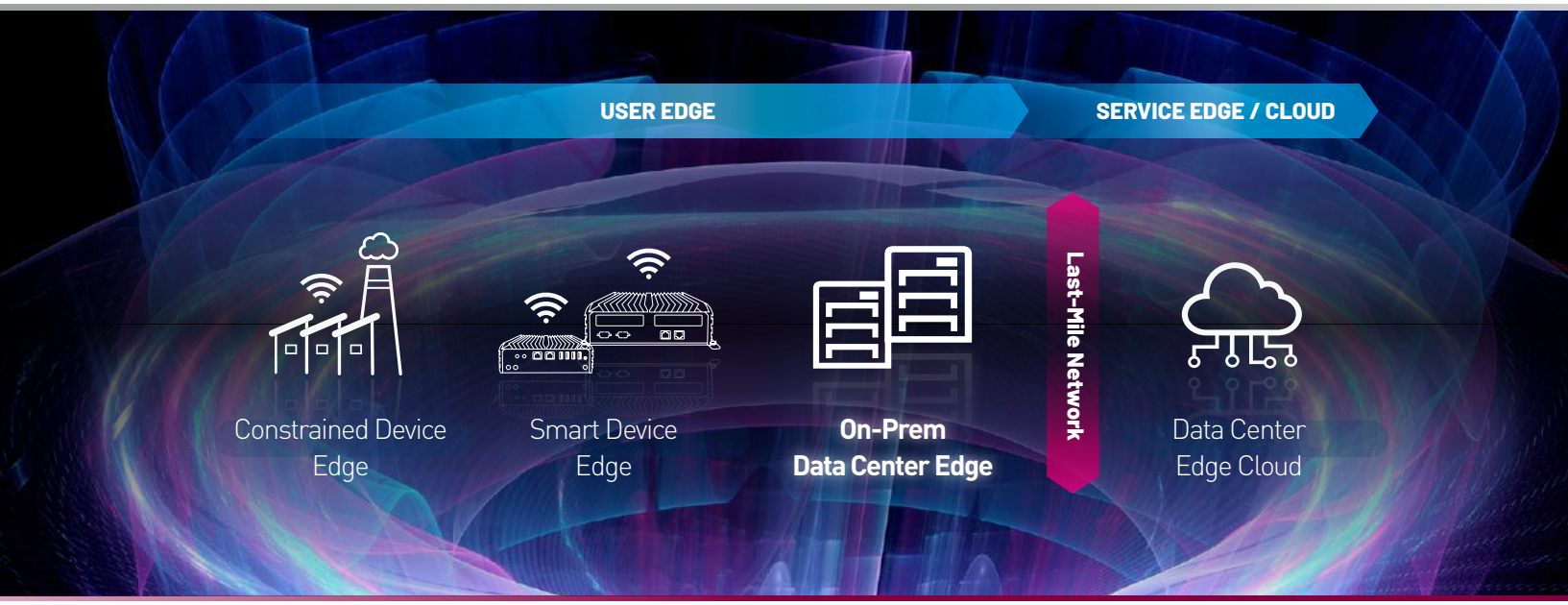
The LLM-1U-RPL Series is a high-performance, short-depth 1U edge AI server engineered for on-premises multimodal deployment in industrial environments. It delivers low latency inferencing to enable real-time genAI and LLM workloads by processing the sensor data at the source of data generation.

- Real-Time Performance
- Gen AI Inferencing
- Operational Redundancy
- Robust Physical Security

Product Brief

1U Edge AI Rackmountable Server

LLM-1U-RPL Series



Performing at the On-Prem Data Center Edge

While most of Premio's edge platforms serve the smart device edge, the LLM-1U-RPL is purpose-built for the on-premises data center edge within the User Edge tier. Positioned between localized devices and the cloud, it enables low-latency AI inferencing and LLM workloads without relying on centralized cloud resources. This reduces bandwidth strain, safeguards data sovereignty, and supports real-time decisions for generative AI tasks in hybrid cloud environments.

Key Features

- Short-Depth 1U Form Factor
- 13th Gen Intel Core E Processor
- Dedicated GPU Acceleration
- PCIe 4.0 Expansion
- Operational Redundancy & Serviceability
- Robust Physical Security
- World-Class Safety Certifications

"By 2029, at least 60% of edge computing deployments will use composite AI (both predictive and generative AI [GenAI]), compared to less than 5% in 2023." - Gartner

Key Markets and Applications

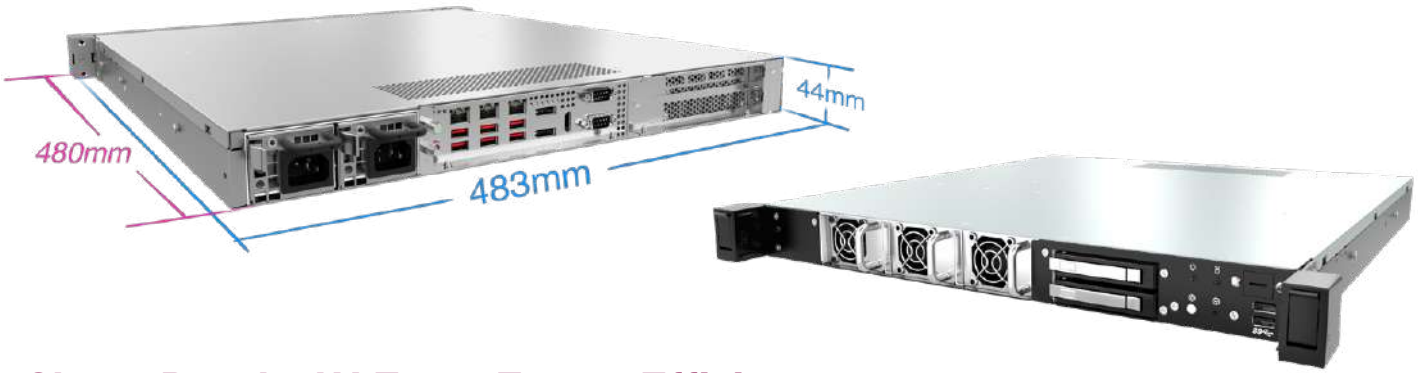
LLM & Gen AI Inference

Industrial Control & Automation

Smart City

Smart Healthcare

Retail & Hospitality



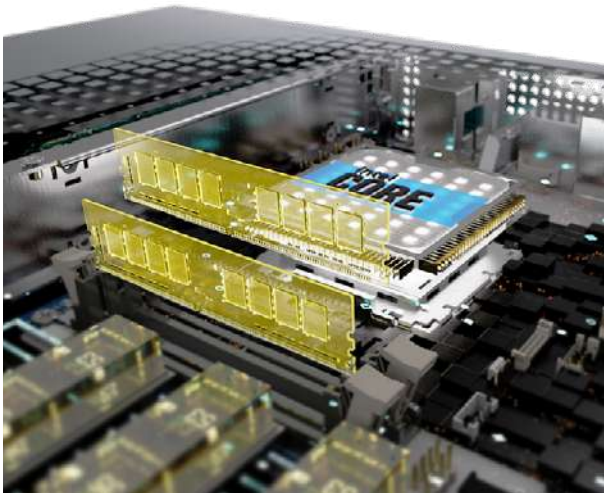
Short-Depth, 1U Form Factor Efficiency

The compact short-depth 1U design of the LLM-1U-RPL is optimized for distributed industrial sites with confined rack spaces commonly found in micro data centers, control rooms, or mobile enclosures. It maximizes overall space efficiency and ensures integration compatibility in spatially limiting edge deployments.

- 1U Short-Depth Form Factor
- 483 (W) x 480 (D) x 44 (H) mm

Optimized for Real-Time Performance

Powered by 13th Gen Intel® Core™ E processors (up to i9, 65W TDP), the LLM-1U-RPL leverages its performance hybrid architecture with performance (P) cores for low-latency inferencing such as LLM prompt response and token generation and efficiency (E) cores to manage background applications. Paired with up to 64GB of dual-channel DDR4 memory, this edge AI server can streamline multi-modal data streams and processes without encountering performance bottlenecks.



- Intel® Core™ i3/i5/i7/i9 Core TE Processors (65W TDP)
- Up to 64GB DDR4 3200MT/s SODIMM

GenAI Acceleration with Dedicated GPU Support

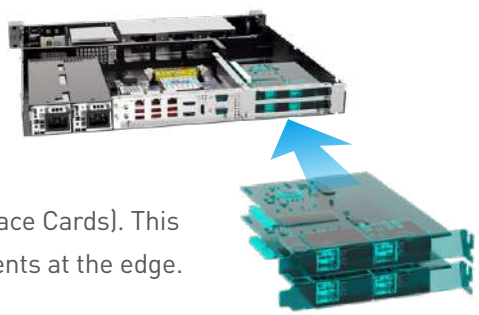
For on-premises LLMs and genAI workloads to execute effectively at the edge, AI acceleration is essential. The LLM-1U-RPL supports a validated list of workstation-class GPUs such as the NVIDIA RTX™ 5000 Ada featuring 32GB of VRAM and 1,000+ TFLOPS of tensor performance. These accelerators are specifically designed for embedded AI environments with extreme power efficiency and specifically to enable last-mile processing of LLMs directly on-site.

Supporting up to an NVIDIA RTX 5000 Ada

Validated List of Workstation-Class GPUs for LLM Workloads

GPU List	VRAM	Power	Tensor Performance
RTX A1000	8GB	50W	53.8 TFLOPS (FP8)
RTX 2000 ADA	16GB	70W	191.9 TFLOPS (FP16)
RTX 4000 SFF ADA	20GB	70W	306.8 TFLOPS (FP8)
RTX 4000 ADA	20GB	130W	327.6 TFLOPS (FP8)
RTX 4500 ADA	24GB	210W	634.0 TFLOPS (FP8)
RTX 5000 ADA	32GB	250W	1044.4 TFLOPS (FP8)

Scalable PCIe 4.0 Expansion Architecture



The edge AI server accommodates evolving AI workloads with PCIe 4.0 expansion and flexibility. Its layout is configurable as either a single PCIe 4.0 x16 or dual PCIe 4.0 x8. This allows for seamless compatibility with a dedicated AI accelerator or various frame grabbers and high-speed NICs (Network Interface Cards). This level of configurability ensures that it can meet complex deployment requirements at the edge.

Configurable between:

- 1x PCIe 4.0 x16
- 2x PCIe 4.0 x8

Flexible and High-Speed Storage Options

The LLM-1U-RPL Series features a flexible storage architecture that includes high-speed NVMe via an M.2 M-Key slot and dual hot-swappable 2.5" SATA bays, enabling rapid data access and easy drive maintenance without system downtime. This local storage capability reduces reliance on cloud resources, easing last mile backhaul bandwidth usage, and accelerating response times. Storing data locally reduces sensitive data exposure to the cloud and ensures data privacy and sovereignty.

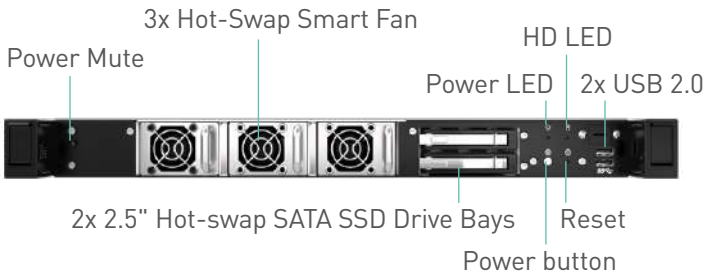
- M.2 M-Key for NVMe SSD

- 2x Hot-swappable 2.5" SATA Bays

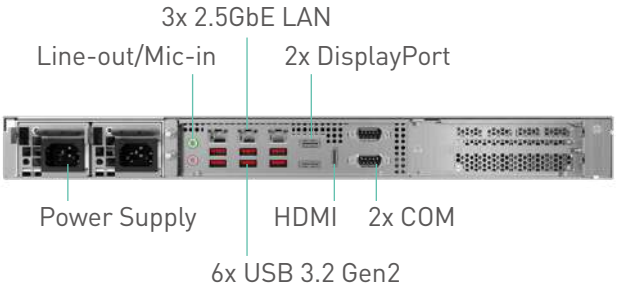
Optimized I/O Connectivity for On-Premises Edge AI

Built to anchor connected edge AI systems, the LLM-1U-RPL series offers comprehensive I/O for seamless integration with distributed networks and diverse devices. 3x 2.5GbE LAN ports enable high-throughput data exchange and network resilience. 6x USB 3.2 ports connect AI-optimized peripherals such as vision cameras and sensors. COM ports provide reliable communication with legacy control systems. This flexible architecture positions the LLM-1U-RPL as a scalable, on-premises edge node, bridging real-time AI processing with field devices across Industry 4.0, mobility, and intelligent infrastructure applications.

Rear



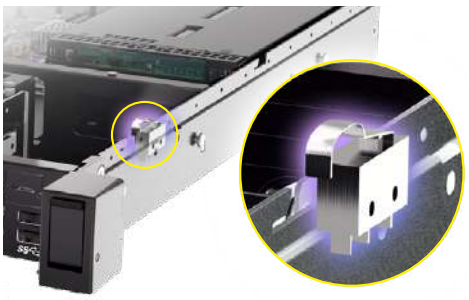
Front



Cybersecurity with Lockable Anti-Tampering and TPM

With cybersecurity becoming a top priority at the edge, the LLM-1U-RPL incorporates physical and firmware-level security features to safeguard against mission-critical infrastructure. A lockable front security bezel prevents unauthorized access, while a chassis intrusion detection switch provides tamper detection and event logging functionality. Additionally, the LLM-1U-RPL supports TPM 2.0 directly on-board to ensure secure boot and data encryption at the hardware root level.

- Security Bezel



- Chassis Intrusion Detection Switch

- TPM 2.0

Reliable 24/7 Operation

Engineered for continuous performance at the edge, the LLM-1U-RPL ensures maximum system availability in mission-critical environments. Its redundant 600W (1+1) power supply supports hot-swappable replacement, allowing failed modules to be serviced without system shutdown—minimizing downtime. Complementing this, 3x hot-swappable fans maintain optimal thermal performance during sustained AI workloads. This service-friendly, high-availability design enables the LLM-1U-RPL to deliver uninterrupted 24/7 performance in demanding edge deployments.

- Hot-swappable 600W PSU (1+1)

- Hot-swappable Fans

World-Class Safety Certifications

The LLM-1U-RPL is engineered and certified to meet global safety and compliance standards, ensuring trusted deployment across diverse industrial and commercial environments. It is UL Listed for safety, and fully compliant with CE and FCC requirements for electromagnetic compatibility and emissions. In addition, the edge AI server benefits from a secure product development process aligned with IEC 62443-4-1, as part of the certified cybersecurity product lifecycle management guidelines.

We Design, Manufacture, and Service customers around the world

LLM-1U-RPL SERIES 1U Edge AI Rackmount Server

Processor	Support 12/13th Gen Intel® ADL & RPL Processor (LGA 1700, 65W/35W TDP)
Memory	2x DDR4 3200MT/s SODIMM. Max. up to 64 GB (Default: 8GB & Non- ECC supported)
Storage	2x 2.5" Hot-swap SATA SSD Drive Bays (7mm/15mm) 1x M.2 M-Key Type (2242/2280, Support PCIe x4 Gen3 NVMe SSD)
PCIe	1x PCIe x16 Gen 4 or 2x PCIe x8 Gen 4 (Card Dimension: 267 (L) x 112 (H) mm, dual slot)
I/O	2x COM, 3x RJ45 (2.5GbE), 6x USB 3.2 Gen2, 2x USB 3.2 Gen 1 Type-A, Line-out/Mic-in
Expansion Slot	1x M.2 B key Type: 3042 • Support PCIe x1 +USB 3.2 Gen 2x1 + USB 2.0 • Support NVMe Storage/AI Module/4G/5G 1x M.2 E-Key Type: 2230 • Support PCIe x1 + USB 2.0; Support CNVi • Devices Supported: Intel® AX210 Wi-Fi 6E & BT-5.1 (vPro Supported)
Power	2x 600W Redundant (1+1) Power Supplies AC: 115 to 230 V, DC: -36 to -72 V, AT/ATX Select
System Cooling	1x CPU Fan Cooler, 3x Hot-swap Redundant Smart Fans, 1x Dust Fan Filter
Operating Temperature	0°C to 35°C with 0.6 m/s airflow
Shock & Vibration	With SSD: 1 Grms (5 - 500 Hz, 0.5 hr/axis) With SSD: NA-G half-sin 11ms
Certification	UL 62368 Ed. 3, CE, FCC Class A
Dimensions	483 (W) x 480 (D) x 44 (H) mm