Optimized Solutions For Generative AI Inference & RAG

Enterprise generative AI inference and RAG

Enterprise AI is transforming organizations by embedding advanced artificial intelligence into core business processes and workflows. Unlike consumer AI, which supports general individual tasks, enterprise AI addresses specific workloads at scale—spanning operations, customer service, supply chain management, and decision-making.

Generative AI inference and Retrieval-Augmented Generation (RAG) are central to this shift, helping organizations to streamline business processes and realize value from their data. Organizations can leverage generative AI to create new content—text, images, and videos—using pre-trained models, helping automate content creation, enhance customer engagement, and streamline internal processes. Building on this, RAG can further enhance the effectiveness of generative AI outputs by drawing on an organization's own data to deliver more accurate, contextually relevant responses. Together, these workloads drive efficiency, strengthen decision-making, and open new paths for innovation and growth.

Supermicro and NVIDIA accelerating enterprise AI

Supermicro and NVIDIA are working together to deliver turnkey solutions built to simplify the process of deploying enterprise generative AI and RAG. Based on flexible architectures and latest-generation GPUs, these solutions provide enterprises with the scalability and efficiency needed to accelerate AI adoption and innovation.

Key technical advantages

- **Broad portfolio:** Form factors designed to be deployed right throughout the enterprise AI pipeline simplify the process of integrating AI-optimized hardware into existing environments where thermal, power, and space constraints may limit the use of one-size-fits-all solutions.
- **Modular design:** Supermicro's Building Block Solutions® provide a flexible, modular approach enabling the integration of new GPU platforms without needing to completely redesign system architectures, reducing costs and shortening development times.
- **Latest-Generation GPUs:** NVIDIA GPUs—including the NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU and NVIDIA H200 NVL GPU—provide AI compute performance and memory capacity needed to support the models required for today's GenAI and RAG workloads.
- **Optimized performance:** Supermicro's NVIDIA-Certified Systems™ are tested and validated to work with NVIDIA software. Leverage NVIDIA's full-stack AI inference platform including NVIDIA NIM™ and NVIDIA NeMo™, as well as NVIDIA Blueprints to accelerate the creation of complete AI solutions.

Business outcomes

- Right-sized compute power that integrates easily into existing infrastructure and can handle multiple, evolving AI workloads.
- Simplify the deployment of AI, allowing your organization to maintain a competitive edge and shorten time to revenue and ROI.
- Streamline operations and boost productivity through AI solutions tailored to specific workloads and environments.
- Improve the relevance of generative AI content using your own data.

Supermicro Systems Optimized For Generative AI & RAG

Key points of value

- Fast time-to-online through modular, pre-validated designs that simplify deployment and scaling
- Flexible GPU capacity per system, enabling right-sized AI performance and the ability to scale workloads across a range of form factors and GPU densities
- Air-cooled architectures designed for sustained efficiency and reliability without major modifications to existing data center power and HVAC
- NVIDIA-Certified Systems™ validated and guaranteed to work with NVIDIA networking and software for full-stack AI solutions

Relevent verticals/industries

- Healthcare
- Federal
- Fintech
- AI development
- AI startup

4U NVIDIA RTX PRO Server

NVIDIA RTX PRO Server reference architecture

Up to 8 GPUs in 4U

E1.S drive support

SYS-422GL-NR

- **CPU** - Dual Socket Intel® Xeon® 6 6900 series processors with P-cores
- **Maximum GPU Quantity** - Up to 8 double-width
- **Memory** - 24 DIMM slots; up to DDR5-6400
- **Storage** - Up to 8 front hot-swap E1.S NVMe drive bays
- **Power** - 4x 3200W Redundant (3 + 1) Titanium Level power supplies

NVIDIA H200 NVL (141GB) – 8 GPUs (NVIDIA-Qualified)

NVIDIA RTX PRO 6000 BSE – 8 GPUs (NVIDIA-Qualified)

5U 8-GPU

13 PCIe 5.0 slots for GPU and NICs

Thermally-optimized 5U design

Up to 24 NVMe storage drives

SYS-522GA-NRT/AS -5126GS-TNRT(2)

- **CPU** - Dual Intel® Xeon® 6900 series processors with P-cores or AMD EPYC™ 9005 Series
- **Maximum GPU Quantity** - Up to 8 double-width
- **Memory** - 24 DIMM slots; up to DDR5-6400
- **Storage** - Up to 24 front hot-swap 2.5" PCIe 5.0 NVMe drive bays
- **Power** - 6x 2700W Redundant Titanium Level power supplies

SYS-522GA-NRT

NVIDIA H200 NVL (141GB) – 8 GPUs

NVIDIA RTX PRO 6000 BSE – 8 GPUs (NVIDIA-Qualified)

AS -5126GS-TNRT2

NVIDIA H200 NVL (141GB) – 8 GPUs

NVIDIA RTX PRO 6000 BSE – 8 GPUs (NVIDIA-Qualified)

AS -5126GS-TNRT

NVIDIA H200 NVL (141GB) – 4 GPUs

NVIDIA RTX PRO 6000 BSE – 4 GPUs (NVIDIA-Qualified)

2U NVIDIA RTX PRO Server

Scalable AI platform

Up to 2 GPUs in 2U

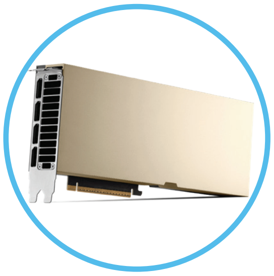
Supports DC-MHS standard

SYS-222C-TN

- **CPU** - Dual Intel® Xeon® 6700/6500 series processors with P-cores or 6700 series processors with E-cores
- **Maximum GPU Quantity** – Up to 2 double-width
- **Memory** - 32 DIMM slots; up to DDR5-6400
- **Storage** - 8 front hot-swap 2.5" NVMe/SAS/SATA drive bays
- **Power** - 2x 2000W Redundant Titanium Level power supplies

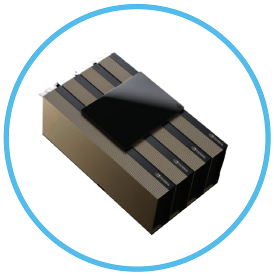
NVIDIA RTX PRO 6000 BSE – 2 GPUs (NVIDIA-Qualified)

NVIDIA GPUs optimized for generative AI and RAG



NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU

- 5th Gen Tensor Cores + 4th Gen Ray Tracing Cores
- 96GB GDDR7 to support models up to 70b parameters
- Up to 600W
- Peak FP4 AI Performance: 3.7 PFLOPS
- Best for: Graphics & visual computing, Industrial & Physical AI, Enterprise AI, Enterprise HPC



NVIDIA H200 NVL (141GB) GPU

- 141GB HBM3e
- Up to 600W
- NVIDIA NVLink™ Bridge for up to 4 GPUs
- Best for: Full-scale LLM training, high-batch GenAI inference, HPC, scientific computing

Oct 2025