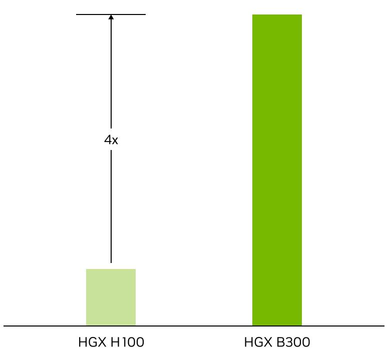
AI / LLM RackScale Solutions

Purpose-built GPU solutions for AI training and inference

AMAX ENGINEERING

With 40 years of engineering expertise, our team specializes in transforming standard IT components into high-performance computing solutions with optimized thermal, electrical, mechanical, and networking design.



Llama 3.1 405B Model Training Speedup

AI Infrastructure and Validated Design

AMAX accelerates AI and LLM workloads that demand extreme GPU density, massive memory capacity, and high-speed I/O. AMAX RackScale GPU solutions designed for AI, combine multi-GPU servers with high bandwidth in node fabric, 800G class networking between nodes, efficient cooling, and monitored power distribution. This approach supports long training runs and high throughput inference while keeping performance stable.

Engineered for Modern AI/LLM Workloads

- **Tensor/Transformer Cores:** Boost LLM-specific operations like attention layers for faster compute.
- **Large HBM Memory Capacity:** Handle massive models and long context windows without bottlenecks.
- **In-node Bandwidth:** NVLink/NVSwitch ensures GPUs stay fully utilized in large-scale AI/LLM training.
- **800G Networking:** Rapid multi-node synchronization for distributed AI/LLM workloads.
- **Energy Efficient Cooling:** Maintain stable AI/LLM performance while reducing energy costs.

Faster Real-Time Throughput with NVIDIA Blackwell

AMAX designs infrastructure for training and serving large language models, built to process massive datasets and keep GPUs synchronized across nodes. RackScale GPU solutions with NVIDIA HGX™ B300 deliver up to 4× faster training for Llama 3.1 405B models compared to HGX H100, accelerating development cycles and reducing time to results. The widely available B200 remains a strong choice for high-throughput inference, while the next-generation B300 sets the standard for the most demanding dense training workloads.

AI/LLM RackScale Solutions

AMAX RackScale 32 with NVIDIA HGX™ B300 (Air-Cooled)

Air-cooled, rack-scale solution built on the NVIDIA reference architecture, delivering high-performance AI computing with NVIDIA HGX B300 GPUs, high-speed interconnects, and offering up to 9.2TB HBM3e memory per rack.

RackScale 32 with NVIDIA HGX™ B300 GPU	
CPU	Dual Socket Intel® Xeon® Scalable processors
GPU	32x NVIDIA Blackwell Ultra GPUs
Cooling	High-efficiency air cooling
GPU Memory	Up to 9.2TB total HBM3e GPU memory per rack
Networking	NVIDIA NDR 800Gbps InfiniBand switches
Storage	High Performance Storage Appliance
Total FP4 Tensor Core	576 PFLOPS
Total FP8 Tensor Core	288 PFLOPS

AMAX LiquidMax® RackScale 64 (Liquid-Cooled)

High-density, fully liquid-cooled rack solution designed to support advanced AI training and inference. With up to 64x NVIDIA HGX B300 GPUs, this system delivers maximum performance and efficient thermal management in a compact footprint.

LiquidMax® RackScale 64 with NVIDIA HGX™ B300 GPU	
CPU	Dual Socket Intel® Xeon® Scalable processors
GPU	64x NVIDIA Blackwell Ultra GPUs
Cooling	Direct liquid cooling, CDU and manifold included
GPU Memory	18.4TB total HBM3e memory
Networking	Integrated power distribution, CDU with dual hot-swap pumps
Storage	High Performance Storage Appliance
Total FP4 Tensor Core	1152 PFLOPS
Total FP8 Tensor Core	576 PFLOPS

End-to-End Deployment Services

AMAX delivers both liquid-cooled and air-cooled rack-scale solutions engineered, assembled, and validated before shipment. Each rack undergoes a site survey, full burn-in, performance benchmarking, and environmental testing to ensure readiness for AI/LLM training and inference. Remote monitoring and ongoing support options help maintain peak performance and uptime. For customers awaiting permanent data center space, HostMax™ provides temporary hosting so systems can go live immediately after build.

Visit www.amax.com/contact to get started today

AMAX | 1565 Reliance Way, Fremont, CA 94539 | 1 (800) 800-6328 | www.amax.com | info@amax.com |