

NVIDIA RTX PRO 6000 Blackwell

8× GPU Server

8× GPU Node | Workstation-Class Blackwell Power for AI Fine-Tuning, Rendering & Simulation at Scale

**8× RTX PRO
6000**

NVIDIA BLACKWELL GPUS

768 GB

TOTAL GDDR7 GPU MEMORY

14.3 TB/s

AGGREGATE MEMORY
BANDWIDTH

96 Cores

DUAL INTEL XEON 6747P

Overview

The Leapswitch RTX PRO 6000 8× GPU Node brings workstation-class NVIDIA Blackwell performance to enterprise scale. Built on the HPE ProLiant DL380a Gen12 NVL platform with eight NVIDIA RTX PRO 6000 Blackwell GPUs, the node delivers 768 GB of aggregate GDDR7 GPU memory — enabling large-scale LLM fine-tuning, high-throughput rendering pipelines, and multi-GPU generative AI and simulation workloads that outgrow a single card.

Deployed as dedicated bare metal, the node provides uncompromised, single-tenant GPU performance for enterprise creative, engineering, and AI teams that need multiple workstation-class GPUs consolidated on one high-density platform.

Designed For

- **Multi-GPU LLM Fine-Tuning** — Fine-tune and serve large open reasoning models across up to 8 GPUs with 768 GB of aggregate GDDR7 memory.
- **Render Farm Consolidation** — Photorealistic, real-time ray-traced rendering at scale for architecture, automotive design, and VFX pipelines.
- **Agentic AI Platforms** — Run multiple always-on, local AI agents in parallel with NVIDIA NemoClaw privacy and security guardrails.
- **Digital Twins & Simulation at Scale** — Multi-GPU physics and computational fluid dynamics simulation for manufacturing and engineering teams.
- **Generative AI Content Pipelines** — Parallel image, video, and 3D asset generation workloads across multiple isolated GPUs.

Why Leapswitch GPU Infrastructure

- **Enterprise-Grade Hardware** — HPE ProLiant platforms with genuine NVIDIA RTX PRO 6000 Blackwell GPUs, rigorously tested and driver-optimized for production workloads.
- **Performance Without Compromise** — Bare-metal deployment eliminates virtualisation overhead for maximum GPU utilisation.
- **India Data Residency** — Deploy from Tier-4 Indian data centres, supporting RBI, DPDP, and SEBI data-localization requirements.
- **Scalable by Design** — Grow from a single node to multi-node enterprise GPU clusters.
- **Security & Isolation** — Single-tenant dedicated infrastructure with data centre-grade physical and network security.

QUICK SPECIFICATIONS

GPU	8× NVIDIA RTX PRO 6000 Blackwell
GPU Architecture	NVIDIA Blackwell
Total GPU Memory	768 GB GDDR7 (96 GB per GPU)
Aggregate Bandwidth	14.3 TB/s (1.8 TB/s per GPU)
CUDA Cores	24,064 per GPU
Tensor / RT Cores	5th Gen / 4th Gen per GPU
CPU	2× Intel Xeon 6747P
System Memory	1 TB DDR5-5600 ECC
Total GPU Power	4,800 W (600 W per GPU)
Server Platform	HPE ProLiant DL380a Gen12 NVL
Deployment	Bare Metal, Dedicated Single-Tenant

Genuine NVIDIA hardware, deployed on
● Leapswitch enterprise infrastructure with 24×7 support.

Key Platform Features

- **Blackwell Streaming Multiprocessors** — Enhanced SMs with neural shaders that integrate neural networks directly inside programmable shaders for AI-augmented graphics.
- **5th-Generation Tensor Cores** — Up to 3× the performance of the previous generation, with FP4 precision support for faster, lower-memory AI model processing and local LLM fine-tuning.
- **4th-Generation RT Cores** — Double the ray-triangle intersection rate of the previous generation, with support for NVIDIA RTX Mega Geometry for up to 100× more ray-traced triangles.
- **96 GB GDDR7 Memory** — 1,792 GB/s of bandwidth to handle massive 3D datasets, multi-billion-parameter models, and large-scale VR environments.
- **PCIe Gen 5** — Double the host bandwidth of PCIe Gen 4 for faster data transfer in AI, data science, and 3D modeling pipelines.
- **Universal MIG Support** — Partition a single GPU into isolated instances (up to 4× 24 GB, 2× 48 GB, or 1× 96 GB) for secure multi-workload, multi-tenant utilization.
- **Double-Flow-Through Thermal Design** — Dual axial cooling pathways sustain 600 W of continuous compute without throttling.
- **Next-Gen Video Engines** — 4× 9th-gen NVENC and 4× 6th-gen NVDEC engines with 4:2:2 support for high-resolution video and streaming workflows.

Per-GPU Specifications

GPU Architecture	NVIDIA Blackwell	Display Connectors	4× DisplayPort 2.1b
NVIDIA CUDA Cores	24,064	Max Simultaneous Displays	4× 4096×2160@120Hz / 4× 5120×2880@60Hz / 2× 7680×4320@60Hz
Tensor Cores	Fifth generation	Video Engines	4× NVENC (9th gen), 4× NVDEC (6th gen)
Ray Tracing Cores	Fourth generation	MIG Support	Up to 4× 24 GB / 2× 48 GB / 1× 96 GB
AI TOPS	4,000 AI TOPS ^{1,2}	Power Consumption	Total board power: 600 W
Single-Precision Perf.	126 TFLOPS ¹	Power Connector	1× PCIe CEM5 16-pin
RT Core Performance	382 TFLOPS ¹	Thermal Solution	Double-flow-through
GPU Memory	96 GB GDDR7 with ECC	Form Factor	5.4" H × 12" L, dual slot, extended height
Memory Interface	512-bit	Graphics APIs	DirectX 12, Shader Model 6.6, OpenGL 4.6 ³ , Vulkan 1.4 ³
Memory Bandwidth	1,792 GB/s	Compute APIs	CUDA 12.8, OpenCL 3.0, DirectCompute
System Interface	PCIe 5.0 x16		

Server & Node Specifications

GPU Count	8× NVIDIA RTX PRO 6000 Blackwell	System Memory	1 TB DDR5-5600 ECC (16 × 64 GB)
Total GPU Memory	768 GB GDDR7 (96 GB per GPU)	Boot Storage	2 × 960 GB NVMe SSD
Aggregate Memory Bandwidth	14.3 TB/s (1.8 TB/s per GPU)	Data Storage	2 × 7.68 TB NVMe SSD
Total GPU Board Power	4,800 W (600 W per GPU)	GPU Interface	PCIe Gen5
CPU	2× Intel Xeon 6747P	Storage Interface	PCIe NVMe
Cores / Threads	96 Cores / 192 Threads	Server Platform	HPE ProLiant DL380a Gen12 NVL

Note: Aggregate GPU board power for this configuration is 4,800 W (600 W × 8 GPUs). Confirm current power, cooling, and rack availability for this configuration with the Leapswitch solutions team when sizing a deployment.

Built For These Workloads

3D Rendering & Visualization

Photorealistic scenes in real time with ray tracing — architecture, automotive design, and VFX studios.

Real-Time RT

8K Viewport

Multi-GPU

AI-Accelerated Content Creation

Generative AI for image, video, and 3D asset creation with 96 GB of VRAM headroom.

Generative AI

Video Production

96 GB VRAM

Digital Twin & Simulation

Complex digital twins for manufacturing, logistics, and urban planning at interactive frame rates.

Omniverse Ready

Physics Sim

Real-Time

Deployment Options

- **Dedicated GPU Instance** — Single-tenant RTX PRO 6000 access on Leapswitch cloud infrastructure.
- **Workstation-as-a-Service** — Remote, on-demand workstation-class GPU for AI and creative teams.
- **Multi-GPU Configurations** — Scale to multiple RTX PRO 6000 GPUs for high-performance rendering and AI clusters.
- **Custom Enterprise Engagements** — Tailored deployment for studios, ISVs, and enterprise AI teams.

Ready to deploy the RTX PRO 6000?

Contact the Leapswitch solutions team for sizing, pricing, and availability.

cloudpe.com/gpu/rtx-pro-6000

1. Peak rates based on GPU Boost Clock. 2. Effective FP4 TOPS with sparsity. 3. Per Khronos specification, pending conformance testing. NVIDIA, RTX, RTX PRO, CUDA, NemoClaw are NVIDIA trademarks. Specifications subject to change without notice.