LEAPSWITCH | GPU CLOUD INFRASTRUCTURE

Leapswitch H200 GPU Server

8× GPU Node | Enterprise AI Infrastructure for Large-Scale Training, Distributed AI & HPC

8 × H200 NVL	1,128 GB	38.4 TB/s	96 Cores
NVIDIA Hopper GPUs	Total HBM3e GPU Memory	Aggregate Memory Bandwidth	Dual Intel Xeon 6747P

Overview

The Leapswitch H200 8× GPU Node is a purpose-built AI powerhouse for enterprise training workloads. Built on the HPE ProLiant DL380a Gen12 NVL platform with eight NVIDIA H200 NVL GPUs, the node delivers 1,128 GB of HBM3e GPU memory — enabling the largest, most complex AI training and inference jobs, including large language models, generative AI, natural language processing, and deep learning recommendation systems. Deployed as dedicated bare metal, the node provides uncompromised, single-tenant GPU performance for enterprise AI centres of excellence.

Designed For

- **Large Language Model Training —** Train and pre-train foundation and domain-specific LLMs with over 1.1 TB of unified GPU memory in a single node.
- **Distributed AI Training —** Multi-GPU data and model parallelism for large-scale deep learning pipelines.
- **Enterprise AI Platforms —** A consolidated compute foundation for enterprise-wide AI/ML initiatives.
- **Scientific Computing & HPC —** Accelerated simulation, modelling, and research computing workloads.
- **Generative AI at Scale —** Diffusion models, multimodal AI, and large-context generative workloads.

Why Leapswitch GPU Infrastructure

Leapswitch delivers enterprise-grade GPU infrastructure for AI startups, research organisations, ISVs, and enterprises — high-performance compute, scalable NVMe storage, and flexible consumption models backed by enterprise SLAs and 24×7 infrastructure support.

- **Enterprise-Grade Hardware —** HPE ProLiant platforms with NVIDIA H200 NVL GPUs for proven reliability and serviceability.
- **Performance Without Compromise —** Bare-metal deployment eliminates virtualisation overhead for maximum GPU utilisation.
- **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.

SPECIFICATIONS	
GPU	8 × NVIDIA H200 NVL PCIe
GPU Memory	1,128 GB HBM3e total (141 GB per GPU)
Memory Bandwidth	38.4 TB/s aggregate (4.8 TB/s per GPU)
GPU Architecture	NVIDIA Hopper
CUDA Cores	16,896 per GPU
Tensor Cores	528 (4th Gen) per GPU
CPU	2 × Intel Xeon 6747P
Cores / Threads	96 Cores / 192 Threads
System Memory	1 TB DDR5-5600 ECC (16 × 64 GB)
Boot Storage	2 × 960 GB NVMe SSD
Data Storage	2 × 7.68 TB NVMe SSD
GPU Interface	PCIe Gen5
Storage Interface	PCIe NVMe
Server Platform	HPE ProLiant DL380a Gen12 NVL
Form Factor	Enterprise Rack Server
Multi-Instance GPU	Supported (MIG)
Deployment	Bare Metal (Dedicated Single-Tenant)
Supported Workloads	AI Training, LLM Training, Distributed Training, HPC, Deep Learning, Computer Vision, Generative AI

Key Platform Features

- **Hopper Architecture GPUs —** 8 × NVIDIA H200 NVL PCIe with 4th Generation Tensor Cores and Transformer Engine acceleration.
- **Massive GPU Memory —** 141 GB HBM3e per GPU at 4.8 TB/s — nearly double the memory of the previous generation for larger models and bigger batch sizes.
- **High-Performance CPU Complex —** Dual Intel Xeon 6747P processors (96 cores / 192 threads) for data pre-processing and pipeline orchestration.
- **1 TB DDR5 ECC Memory —** DDR5-5600 ECC RDIMMs for high-throughput data staging and in-memory operations.
- **Enterprise NVMe Storage —** Redundant NVMe boot drives plus 15.36 TB of high-speed NVMe data storage for datasets and checkpoints.
- **PCIe Gen5 Fabric —** High-bandwidth host connectivity across GPUs, storage, and networking.
- **Enterprise Reliability —** HPE ProLiant Gen12 platform engineering with enterprise-grade serviceability and manageability.

Deployment Options

- Dedicated GPU Server (Bare Metal)
- Private Cloud Deployments
- Enterprise GPU Clusters (multi-node scale-out)
- Custom AI Infrastructure engagements

Powered for the Full AI Lifecycle

From data preparation and distributed training to checkpointing and evaluation, the 8× GPU node consolidates the full training lifecycle on a single high-density platform. Pair multiple nodes into an enterprise GPU cluster for scale-out training, or combine with Leapswitch 4× GPU inference nodes for a complete train-to-serve AI infrastructure footprint.

Leapswitch GPU infrastructure enables organisations to accelerate AI innovation with secure, scalable, and high-performance computing resources. Specifications are subject to change without notice. For sizing, pricing, and availability, contact the Leapswitch solutions team.