



04 · HIGH-MEMORY INFERENCE

X8 H200 NVL

High-memory inference
with node-local NVLink.

A 4U dual-socket AMD EPYC platform with
eight NVIDIA H200 NVL GPUs arranged as
two four-GPU NVLink bridge domains.



PRODUCT VIEW · X8 H200 NVL

4U

AIR-COOLED PLATFORM

8 ×

NVIDIA H200 NVL

141 GB

HBM3E / GPU

2 × 4

GPU NVLINK DOMAINS

DESIGNED FOR

long-context inference · governed RAG · reasoning models · private model hosting · selected fine-tuning and HPC

01

Increase memory per accelerator

Each H200 NVL provides 141 GB nominal HBM3e, supporting long-context inference, reasoning models and memory-intensive serving profiles.

02

Use node-local NVLink scale-up

Two supported four-GPU bridge domains provide defined intra-node GPU groups without presenting the eight GPUs as one NVLink or NVSwitch domain.

03

Separate the network paths

Four 400GbE ConnectX-7 links distribute two endpoints to each A/B rail; BlueField-3 supplies dual 200GbE service paths.

Product imagery is illustrative. Final chassis details and installed options are defined by the order configuration.

ENTERPRISE OPERATIONS

Manage the node without merging the data planes.

Independent A/B workload paths, separate service connectivity and dedicated out-of-band endpoints support controlled operations and fault isolation.

A/B FABRIC

DPU OOB

3+1 POWER

High-memory inference. Two defined NVLink domains.

X8 H200 NVL pairs 141 GB HBM3e GPUs with dual-socket host compute, two four-GPU NVLink bridge domains and independent A/B 400GbE paths.

CUSTOMER BENEFITS

- 01 Increase memory per accelerator**
Each H200 NVL provides 141 GB nominal HBM3e, supporting long-context inference, reasoning models and memory-intensive serving profiles.
- 02 Use node-local NVLink scale-up**
Two supported four-GPU bridge domains provide defined intra-node GPU groups without presenting the eight GPUs as one NVLink or NVSwitch domain.
- 03 Separate the network paths**
Four 400GbE ConnectX-7 links distribute two endpoints to each A/B rail; BlueField-3 supplies dual 200GbE service paths.
- 04 Balance compute and local data**
Two EPYC 9455 processors, 2 TB nominal host memory, mirrored boot and local E1.S data drives provide a balanced node profile.

IDEAL WORKLOADS

- 01 Long-context inference**
- 02 Retrieval-augmented generation**
- 03 Reasoning models**
- 04 Private model hosting**
- 05 Selected fine-tuning and HPC**

CONFIGURATION FLEXIBILITY

- EPYC 9455 or order-defined 9555 option
- 2 TB standard / 4 TB capacity profile
- Independent A/B 400GbE rails
- Two four-GPU NVLink bridge domains

TECHNICAL HIGHLIGHTS

FORM FACTOR

4U air-cooled eight-GPU rack server

ACCELERATORS

8 × NVIDIA H200 NVL; 141 GB nominal HBM3e each; PCIe Gen5 x16; configurable up to 600 W/GPU

GPU FABRIC

Two four-GPU NVLink bridge domains; bridge capability and arrangement follow the ordered GPU/platform combination

PROCESSORS

2 × AMD EPYC 9455; 48 cores / 96 threads and 300 W default TDP each

HOST MEMORY

16 × 128 GB DDR5 RDIMMs = 2 TB nominal; supported speed and population follow the exact motherboard and DIMM part

BOOT / LOCAL DATA

2 × 1.92 TB-class mirrored boot devices; 4 × 7.68 TB E1.S local data drives in the reference configuration

EAST-WEST NETWORK

4 × 400GbE ConnectX-7 links; two endpoints to each independent A/B rail

SERVICE / MANAGEMENT

BlueField-3 with dual 200GbE service paths; separate server BMC and DPU out-of-band access

PCIe VALIDATION

GPU and adapter link width, root complex, peer access, NUMA locality and GPU-to-NIC affinity are confirmed against the ordered platform

POWER / COOLING

Reference 3+1 arrangement of 3.2 kW power supplies; front-to-rear air cooling; exact limits require order confirmation

CONFIGURATION NOTE

The node contains two four-GPU NVLink bridge domains, not one eight-GPU NVLink domain and not an NVSwitch system. GPU/platform compatibility, bridge support, PCIe topology, power, cooling, firmware, operating system and NVIDIA support scope require qualification of the exact ordered configuration.

DEFINE THE RIGHT SYSTEM

Align compute, memory, storage, network and acceleration around the workload.

DISCUSS A CONFIGURATION

www.sidera-is.com