



04 · HGX SCALE-UP COMPUTE

X8 HGX B300

Eight-GPU scale-up for training and distributed AI.

A 4U liquid-cooled platform built around an eight-GPU NVIDIA HGX B300 baseboard, fifth-generation NVLink/NVSwitch and per-GPU ConnectX-8 connectivity.



PRODUCT VIEW · X8 HGX B300

4U

LIQUID-COOLED PLATFORM

8 ×

NVIDIA B300

288 GB

HBM3E / GPU

14.4 TB/s

AGGREGATE NVLINK

DESIGNED FOR

large-model training · fine-tuning · distributed inference · scientific computing · multi-node scale-out

01

Train inside one scale-up domain

Eight B300 GPUs operate within one fifth-generation NVLink/NVSwitch domain, providing 14.4 TB/s aggregate node-local NVLink bandwidth.

02

Expand accelerator memory

Each B300 provides 288 GB nominal HBM3e; the node contains 2.304 TB nominal memory across eight separate GPU devices.

03

Give every GPU a fabric interface

Eight integrated ConnectX-8 interfaces provide up to 800 Gb/s per GPU interface for rail-aligned multi-node scale-out.

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

ENTERPRISE OPERATIONS

Control a dense liquid-cooled compute node.

Dedicated management, separate out-of-band paths and a redundant CRPS reference support controlled lifecycle operations for the HGX platform.

SERVER BMC
SEPARATE OOB
5+5 POWER

One eight-GPU HGX domain. Built to scale out.

X8 HGX B300 combines an eight-GPU NVLink/NVSwitch scale-up domain with 800Gb/s-class per-GPU networking and direct-to-chip liquid cooling.

CUSTOMER BENEFITS

- 01 Train inside one scale-up domain**
Eight B300 GPUs operate within one fifth-generation NVLink/NVSwitch domain, providing 14.4 TB/s aggregate node-local NVLink bandwidth.
- 02 Expand accelerator memory**
Each B300 provides 288 GB nominal HBM3e; the node contains 2.304 TB nominal memory across eight separate GPU devices.
- 03 Give every GPU a fabric interface**
Eight integrated ConnectX-8 interfaces provide up to 800 Gb/s per GPU interface for rail-aligned multi-node scale-out.
- 04 Engineer for sustained power density**
Direct-to-chip liquid cooling and a 5+5 × 3 kW CRPS reference support the thermal and power envelope of the HGX platform.

IDEAL WORKLOADS

- 01 Large-model training**
- 02 Model fine-tuning**
- 03 Distributed inference**
- 04 Scientific workloads**
- 05 Rail-aligned scale-out**

CONFIGURATION FLEXIBILITY

- Dual-plane 800G/GPU reference
- Single-plane 400G/GPU option
- 2 TB standard / 4 TB capacity profile
- Selected Xeon 6760P / 6767P / 6740P profiles

TECHNICAL HIGHLIGHTS

FORM FACTOR

4U liquid-cooled eight-GPU HGX server

ACCELERATORS

8 × NVIDIA B300 Blackwell Ultra SXM; 288 GB nominal HBM3e each; configurable up to 1,100 W/GPU

SCALE-UP FABRIC

One eight-GPU fifth-generation NVLink/NVSwitch domain; 14.4 TB/s aggregate NVLink; 1.8 TB/s GPU-to-GPU bandwidth class

PROCESSORS

2 × Intel Xeon 6760P; 64 cores / 128 threads and 330 W TDP each

HOST MEMORY

16 × 128 GB DDR5 RDIMMs = 2 TB nominal; exact board population and speed are order-defined

BOOT / CACHE

2 × 1.92 TB-class mirrored boot devices; 4 × 7.68 TB U.2 local cache in the reference configuration

COMPUTE NETWORK

8 × integrated ConnectX-8 interfaces, up to 800 Gb/s per GPU interface

NORTH / SOUTH

1 × BlueField-3 B3240 DPU with a dual-port 400GbE profile; one port per converged plane

MANAGEMENT

Dedicated server BMC plus separate NIC/DPU and infrastructure out-of-band paths in the ordered deployment

POWER / COOLING

5+5 × 3 kW CRPS reference and direct-to-chip liquid cooling; facility and site limits follow the exact platform

CONFIGURATION NOTE

The eight GPUs form one node-local NVLink/NVSwitch domain; Ethernet provides node-to-node scale-out and does not extend that coherent domain. Final board revision, CPU/DIMM profile, NIC and DPU modes, power delivery, liquid-cooling interface, firmware and support scope are defined by the ordered platform.

DEFINE THE RIGHT SYSTEM

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

DISCUSS A CONFIGURATION

www.sidera-is.com