H200 platform brief

HGX SXM and NVL architectures for AI and HPC

Choose the GPU fabric for the workload before choosing the chassis. H200 SXM is an HGX platform component; H200 NVL is a PCIe accelerator. NexTec scopes the complete supported system, including host compute, storage, external networking and facility requirements.



Independent NexTec server concept illustration. This is not the specified Supermicro chassis, an HGX baseboard view, or a photograph of inventory. See linked manufacturer references for exact hardware.

Per GPU specification	H200 NVL	H200 SXM
Memory	141 GB HBM3e	141 GB HBM3e
Local memory bandwidth	4.8 TB/s	4.8 TB/s
Maximum GPU power	Up to 600 W	Up to 700 W
Physical integration	PCIe 5.0 x16 add-in card	SXM module on HGX baseboard
GPU interconnect	Supported 2- or 4-GPU NVLink bridge groups	Eight-GPU HGX NVLink fabric with four NVSwitch devices

Both support up to 900 GB/s NVLink bandwidth per GPU in the appropriate supported topology. GPU memory totals and GPU counts do not by themselves define communication behavior or application scaling.

HGX REFERENCE SYSTEM

Eight GPU HGX H200 platform

The HGX reference is the Supermicro SYS-821GE-TNHR, an 8U dual-Xeon system with eight H200 SXM GPUs. NVL cards cannot replace SXM modules on the same baseboard.

System element	Reference configuration
GPU population	8 x H200 SXM, 141 GB HBM3e each; 1,128 GB aggregate device memory
Host processors	Dual 4th or 5th Gen Intel Xeon Scalable; exact models, TDP and firmware require OEM qualification
ECC host memory	32 DDR5 DIMM slots; up to 4 TB at 1DPC or 8 TB at 2DPC. CPU, DIMMs and population determine speed
Boot storage	Two onboard M.2 NVMe slots; drives and mirror method specified in the proposal
Front storage	12 default hot-swap 2.5-inch NVMe bays; four additional NVMe bays optional. Confirm the ordered SATA configuration and controllers
PCIe expansion	Eight PCIe 5.0 x16 low-profile slots, two FHHL slots and two optional FHHL slots
GPU power subtotal	Up to 5.6 kW for eight 700 W SXM GPUs, before host and other components
Fabric planning	Internal NVLink and NVSwitch connectivity is separate from the external Ethernet or InfiniBand cluster fabric

Select the whole system topology

Define the GPU-affiliated NIC arrangement, external network, optics and cables, shared-storage throughput and management network. Validate the intended training, inference or HPC software on the selected internal and inter-node fabrics.

Treat memory capacity as distributed

The eight-GPU total of 1,128 GB is aggregate memory across devices, not automatically one shared pool. The application or framework determines partitioning, communication and usable capacity for the workload.

This reference uses eight SXM GPUs. It is not a two- or four-GPU SXM offering. Confirm the exact supported H200 cooling implementation, rack requirements and service scope for the ordered build.

ARCHITECTURE AND PROCUREMENT

Qualification before deployment

H200 projects require a complete-system proposal. GPU specifications alone do not establish facility readiness, application performance or the supported service model.

When to evaluate each path

Requirement	Evaluation focus
Communication-intensive distributed workloads	Evaluate the HGX eight-GPU fabric, collective performance and inter-node scaling on a representative workload
PCIe deployment with smaller NVLink groups	Evaluate H200 NVL on a qualified PCIe host; specify the supported 2- or 4-GPU bridge kit, card spacing and GPU count
Large-model serving	Record model, precision, context, concurrency, latency and serving software before sizing capacity

Acceptance and facility checks

Agree GPU and fabric health checks, PCIe and NVLink enumeration, ECC memory and storage tests, sustained-load power and thermal behavior, and representative application acceptance. Document firmware and software versions. Test duration and acceptance thresholds require agreement.

Confirm complete-system electrical load, PSU redundancy, input-voltage requirements, cooling implementation, rack loading and service clearances. SXM and NVL have different thermal and host requirements; do not assume every H200 configuration is air-only.

Commercial conditions

Quotation and availability. Reference specifications and options are not a stock list. Exact parts, condition, compatibility, price, lead time and service scope are confirmed in the written quotation. An inquiry does not place an order or reserve inventory. Sales and delivery are within the United States only.

Purchase approval. Buyer vetting and NexTec approval are required before purchase. Manufacturer or distributor approval is required where applicable, including PNY review for applicable orders. Buyer, end-user, intended-use and installation-location details may be requested.

Manufacturer warranty. The applicable component manufacturer provides the warranty. Duration, eligibility and claim terms vary by part. Customers contact the manufacturer for claims; NexTec can assist with order and warranty information.

Technical references and next step

NVIDIA H200 specifications

Supermicro SYS-821GE-TNHR platform

Request an H200 system proposal