DATASHEET

NVIDIA H100 CNX

Unified Network and Compute Acceleration

Unprecedented performance for GPU-powered, IO-intensive workloads.

Experience the unprecedented performance of converged acceleration. NVIDIA H100 CNX combines the power of the NVIDIA H100 Tensor Core GPU with the advanced networking capabilities of the NVIDIA® ConnectX®-7 smart network interface card (SmartNIC) to accelerate GPU-powered, input/output (IO)-intensive workloads, such as distributed AI training in the enterprise data center and 5G processing.

Better I/O Performance

NVIDIA H100 and ConnectX-7 are connected via an integrated PCIe Gen5 switch, which provides a dedicated high-speed path for data transfers between the GPU and network. This eliminates bottlenecks of data going through the host and provides low, predictable latency, which is important for time-sensitive applications like 5G signal processing.

Balanced, Optimized Design

The integration of a GPU and a SmartNIC into a single device results in a balanced architecture by design. In systems where multiple GPUs are desired, a converged accelerator card enforces the optimal 1:1 ratio of GPU to NIC. The design also avoids contention on the server's PCIe bus, so the performance scales linearly with additional devices.

Cost Savings

Because the GPU and SmartNIC are connected directly, customers can leverage mainstream PCIe Gen4 or even Gen3 servers to achieve a level of performance only possible with high-end or purpose-built systems. Using a single card also saves on power, space, and PCIe device slots, enabling further cost savings by allowing a greater number of accelerators per server.

Application-Ready

Core acceleration software libraries, such as the NVIDIA Collective Communications Library (NCCL) and Unified Communication X (UCX®), automatically make use of the best-performing path for data transfers to GPUs. As a result, existing GPU-accelerated multi-node applications can take advantage of the H100 CNX without any modification, delivering immediate benefits.

KEY FEATURES

- NVIDIA H100 Tensor Core GPU
- NVIDIA ConnectX-7 SmartNIC
- Integrated PCIe Gen5 switch

TOP USE CASES

- Distributed Multi-node AI Training
- Enterprise 5G

SPECIFICATIONS

GPU Memory	**80GB HBM2e**
Memory Bandwidth	**> 2.0 TB/s**
Multi-Instance GPU (MIG) instances	**7 instances @ 10GB each** **3 instances @ 20GB each** **2 instances @ 40GB each**
Interconnect	**PCIe Gen5 128 GB/s**
Networking	**Up to 400 Gb/s (NDR or 400GbE), dual-port QSFP112*, Ethernet or InfiniBand**
Form Factor	**Dual-slot full-height, full-length (FHFL)**
Max Power	**350W**

**With aggregated bandwidth of 400 GB/s*

USE CASE	BENEFITS OF UNIFIED NETWORK AND COMPUTE ACCELERATION
Distributed Multi-node AI Training	> Dedicated path from the network to the GPU enables NVIDIA GPUDirect® RDMA to operate at near line speeds > Ideal GPU-to-NIC ratio allows for balanced GPU power scale-up
Enterprise 5G	> Dedicated path from the network to the GPU paves the way for low, predictable latency > Linear scalability with additional accelerators

Ready to Get Started?

To learn more, visit **www.nvidia.com/H100CNX**