What Is NVLink?

NVLink is a high-speed interconnect for GPU and CPU processors in accelerated systems, propelling data and calculations to actionable results.

Author: Rick Merritt

Accelerated computing — a capability once confined to high-performance computers in government research labs — has gone mainstream.

Banks, car makers, factories, hospitals, retailers and others are adopting AI supercomputers to tackle the growing mountains of data they need to process and understand.

These powerful, efficient systems are superhighways of computing. They carry data and calculations over parallel paths on a lightning journey to actionable results.

GPU and CPU processors are the resources along the way, and their onramps are fast interconnects. The gold standard in interconnects for accelerated computing is NVLink .

NVLink is a high-speed connection for GPUs and CPUs formed by a robust software protocol, typically riding on multiple pairs of wires printed on a computer board. It lets processors send and receive data from shared pools of memory at lightning speed.

Now in its fourth generation, NVLink connects host and accelerated processors at rates up to 900 gigabytes per second (GB/s).

That's more than 7x the bandwidth of PCIe Gen 5, the interconnect used in conventional x86 servers. And NVLink sports 5x the energy efficiency of PCIe Gen 5, thanks to data transfers that consume just 1.3 picojoules per bit.

First introduced as a GPU interconnect with the NVIDIA P100 GPU, NVLink has advanced in lockstep with each new NVIDIA GPU architecture.

In 2018, NVLink hit the spotlight in high performance computing when it debuted connecting GPUs and CPUs in two of the world's most powerful supercomputers, Summit and Sierra .

The systems, installed at Oak Ridge and Lawrence Livermore National Laboratories, are pushing the boundaries of science in fields such as drug discovery , natural disaster prediction and more.

In 2020, the third-generation NVLink doubled its max bandwidth per GPU to 600GB/s, packing a dozen interconnects in every NVIDIA A100 Tensor Core GPU .

The A100 powers AI supercomputers in enterprise data centers, cloud computing services and HPC labs across the globe.

Today, 18 fourth-generation NVLink interconnects are embedded in a single NVIDIA H100 Tensor Core GPU . And the technology has taken on a new, strategic role that will enable the most advanced CPUs and accelerators on the planet.

NVIDIA NVLink-C2C is a version of the board-level interconnect to join two processors inside a single package, creating a superchip. For example, it connects two CPU chips to deliver 144 Arm Neoverse V2 cores in the NVIDIA Grace CPU Superchip, a processor built to deliver energy-efficient performance for cloud, enterprise and HPC users.

NVIDIA NVLink-C2C also joins a Grace CPU and a Hopper GPU to create the Grace Hopper Superchip . It packs accelerated computing for the world's toughest HPC and AI jobs into a single chip.

Alps , an AI supercomputer planned for the Swiss National Computing Center, will be among the first to use Grace Hopper. When it comes online later this year, the high-performance system will work on big science problems in fields from astrophysics to quantum chemistry.

Grace and Grace Hopper are also great for bringing energy efficiency to demanding cloud computing workloads.

For example, Grace Hopper is an ideal processor for recommender systems . These economic engines of the internet need fast, efficient access to lots of data to serve trillions of results to billions of users daily.

In addition, NVLink is used in a powerful system-on-chip for automakers that includes NVIDIA Hopper, Grace and Ada Lovelace processors. NVIDIA DRIVE Thor is a car computer that unifies intelligent functions such as digital instrument cluster, infotainment, automated driving, parking and more into a single architecture.

NVLink also acts like the socket stamped into a LEGO piece. It’s the basis for building supersystems to tackle the biggest HPC and AI jobs.

For example, NVLinks on all eight GPUs in an NVIDIA DGX system share fast, direct connections via NVSwitch chips. Together, they enable an NVLink network where every GPU in the server is part of a single system.

To get even more performance, DGX systems can themselves be stacked into modular units of 32 servers, creating a powerful, efficient computing cluster .

Users can connect a modular block of 32 DGX systems into a single AI supercomputer using a combination of an NVLink network inside the DGX and NVIDIA Quantum-2 switched Infiniband fabric between them. For example, an NVIDIA DGX H100 SuperPOD packs 256 H100 GPUs to deliver up to an exaflop of peak AI performance.

To get even more performance, users can tap into the AI supercomputers in the cloud such as the one Microsoft Azure is building with tens of thousands of A100 and H100 GPUs . It’s a service used by groups like OpenAI to train some of the world’s largest generative AI models.

And it’s one more example of the power of accelerated computing.

Original URL: https://blogs.nvidia.com/blog/2023/03/06/what-is-nvidia-nvlink/