NVLink and NVSwitch

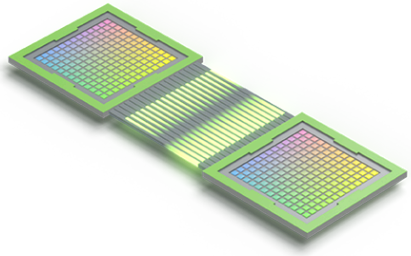
The building blocks of advanced multi-GPU communication—within and between servers.

A Need for Faster, More Scalable Interconnects

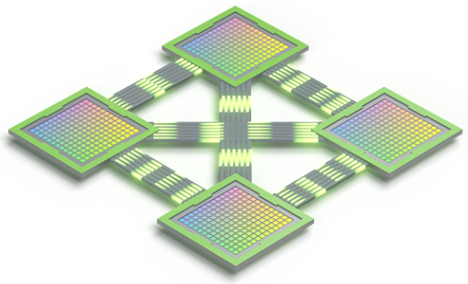
Increasing compute demands in AI and high-performance computing (HPC)—including an emerging class of trillion-parameter models—are driving a need for multi-node, multi-GPU systems with seamless, high-speed communication between every GPU. To build the most powerful, end-to-end computing platform that can meet the speed of business, a fast, scalable interconnect is needed.

improved scalability for multi-GPU system configurations. A single NVIDIA H100 Tensor

Servers like the NVIDIA DGX™ H100 take advantage of this technology to deliver greater scalability for ultrafast deep learning training.

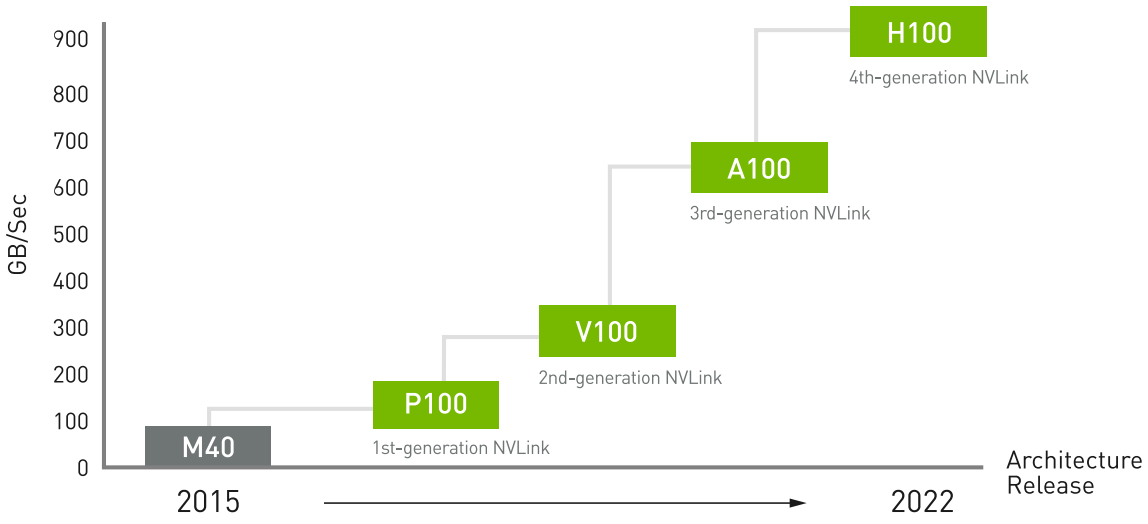


NVIDIA H100 PCIe with NVLink GPU-to-GPU connection



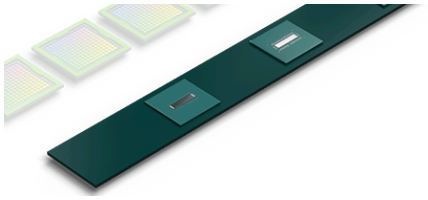
NVIDIA H100 with NVLink GPU-to-GPU connections

NVLink Performance

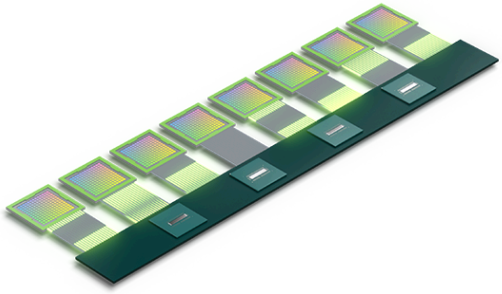


NVLink in NVIDIA H100 increases inter-GPU communication bandwidth 1.5X compared to the previous generation, so researchers can use larger, more sophisticated applications to solve more complex problems.

The third generation of NVIDIA NVSwitch™ builds on the advanced communication capability of NVLink to deliver higher bandwidth and reduced latency for compute-intensive workloads. To enable high-speed, collective operations, each NVSwitch has 64 NVLink ports equipped with engines for NVIDIA Scalable Hierarchical Aggregation Reduction Protocol (SHARP)™ for in-network reductions and multicast acceleration.



NVSwitch enables eight GPUs in an NVIDIA DGX H100 system to cooperate in a cluster with full-bandwidth connectivity.



How NVLink and NVSwitch Work Together

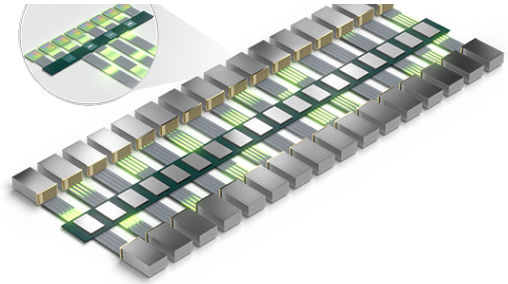
NVLink is a direct GPU-to-GPU interconnect that scales multi-GPU input/output (IO) within the server. NVSwitch connects multiple NVLinks to provide all-to-all GPU communication at full NVLink speed within a single node and between nodes.

With the combination of NVLink and NVSwitch, NVIDIA won MLPerf 1.1, the first industry-wide AI benchmark.

Parameter Models with NVLink Switch System

With NVSwitch, NVLink connections can be extended across nodes to create a seamless, high-bandwidth, multi-node GPU cluster—effectively forming a data center-sized GPU. By adding a second tier of NVLink Switches externally to the servers, future NVLink Switch Systems can connect up to 256 GPUs and deliver a staggering 57.6 terabytes per second (TB/s) of all-to-all bandwidth, making it possible to rapidly solve even the largest AI jobs.

Learn More About NVIDIA H100 ›



Scaling from Enterprise to Exascale

Full Connection for Unparalleled Performance

NVSwitch is the first on-node switch architecture to support eight to 16 fully connected GPUs in a single server node. The third-generation NVSwitch interconnects every GPU pair at an incredible 900GB/s. It supports full all-to-

The Most Powerful AI and HPC Platform

NVLink and NVSwitch are essential building blocks of the complete NVIDIA data center solution that incorporates hardware, networking, software, libraries, and optimized AI models and applications from the NVIDIA AI Enterprise software

Specifications

NVLink NVSwitch

	Second Generation	Third Generation	Fourth Generation
NVLink bandwidth per GPU	300GB/s	600GB/s	900GB/s
Maximum Number of Links per GPU	6	12	18
Supported NVIDIA Architectures	NVIDIA Volta™ architecture	NVIDIA Ampere Architecture	NVIDIA Hopper™ Architecture

Preliminary specifications, may be subject to change

Take a Deep Dive into the NVIDIA Hopper Architecture

Read Whitepaper

NVLink NVSwitch Specifications Shop Drivers Support

NVIDIA EGX Platform

NVIDIA HGX Platform

Networking Products

Virtual GPUs

Confidential Computing

NVLink-C2C

NVLink/NVSwitch

Tensor Cores

Multi-Instance GPU

IndeX ParaView Plugin

NVIDIA Morpheus AI framework

NVIDIA

Data Center Blogs

Company Overview

NVLink NVSwitch Specifications Shop Drivers Support

DGX Product Literature

Documentation

Energy Efficiency Calculator

Glossary

GPU Apps Catalog

GPU Test Drive

GTC AI Conference

NVIDIA GRID Community Advisors

Qualified System Catalog

Technical Training

Training for IT Professionals

Where to Buy

Virtual GPU Forum

Virtual GPU Product Literature

NVIDIA Foundation

Research

Social Responsibility

Technologies

Careers

Follow Data Center