



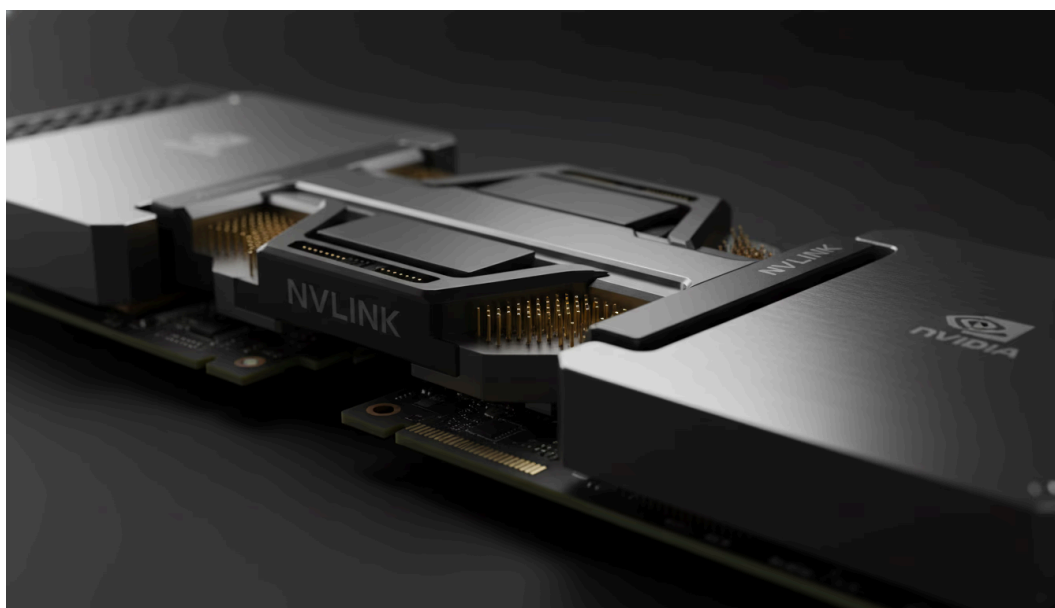
INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

NVIDIA NVLink Explained: A Guide to the GPU Interconnect

Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2025-10-22 · nvidia nvlink · gpu interconnect · nvswitch · pcie vs nvlink · multi-gpu systems · gpu bandwidth · high performance computing · ai hardware

Original article: <https://intuitionlabs.ai/articles/nvidia-nvlink-gpu-interconnect>



In this article

1. [NVIDIA NVLink Explained: A Guide to the GPU Interconnect](#)
 2. [Executive Summary](#)
 3. [Introduction and Background](#)
 4. [NVLink Architecture and Key Features](#)
 5. [NVLink Bandwidth vs. PCIe and Other Buses](#)
 6. [NVSwitch and All-to-All Connectivity](#)
 7. [Performance and Impact](#)
 8. [Case Studies and Real-World Examples](#)
 9. [Ecosystem and Competitive Landscape](#)
 10. [Challenges and Limitations](#)
 11. [Future Directions](#)
 12. [Conclusion](#)
-

NVIDIA NVLink Explained: A Guide to the GPU Interconnect

Executive Summary

NVIDIA **NVLink** is a revolutionary high-speed GPU interconnect architecture that fundamentally transforms the way graphics accelerators are coupled with CPUs and with each other. Introduced in 2016 with the NVIDIA Pascal™ P100 GPU, NVLink provides up to **5× the bandwidth of PCIe 3.0 x16** per GPU ([nvidianews.nvidia.com](#)) ([developer.nvidia.com](#)). By combining multiple NVLink links per GPU, NVLink delivers **several times the data rate** of traditional PCIe (e.g. an NVIDIA Tesla P100 achieves 160 GB/s bidirectional vs ~32 GB/s over PCIe Gen3×16 ([nvidianews.nvidia.com](#))). As a result, systems built on NVLink (often paired with NVIDIA's NVSwitch fabric) can achieve **hundreds of GB/s to TB/s GPU-to-GPU bandwidth**, unleashing vastly higher application performance in multi-GPU and CPU–GPU workloads. NVLink enables **coherent shared memory** across processors (already deployed in DOE pre-exascale supercomputers Summit and Sierra ([developer.nvidia.com](#))), effectively treating GPU memory and CPU memory as part of a unified address space. This dramatically speeds up workloads with frequent GPU-GPU and GPU-CPU communication (e.g. [distributed AI training](#), [large-scale simulations](#), and unified-memory applications).

Update (October 2026): The NVLink generation in volume deployment today is the fifth (Blackwell and Blackwell Ultra, including GB200 and GB300 NVL72): 18 links and 1,800 GB/s per GPU, in NVLink domains of 8 or 72 GPUs, with 130 TB/s across an NVL72 rack ([nvidia.com](#)). Sixth-generation NVLink arrives with Vera Rubin. At the January 2026 launch NVIDIA quoted 3.6 TB/s per GPU and 260 TB/s per Vera Rubin NVL72 rack ([nvidianews.nvidia.com](#)); NVIDIA's current NVLink and Vera Rubin NVL72 specification pages list 36 links and 3 TB/s per GPU and 216 TB/s per rack, marked as preliminary ([nvidia.com](#)) ([nvidia.com](#)). On May 31, 2026 NVIDIA said Vera Rubin was ramping into full production, with production shipments set to begin in the fall

(nvidianews.nvidia.com). NVLink Fusion has also widened beyond its first partners: Intel agreed in September 2025 to build NVIDIA-custom x86 CPUs connected with NVLink (nvidianews.nvidia.com), and AWS's Annapurna Labs will support NVLink Fusion starting with Trainium4 (blogs.nvidia.com).

In short, NVLink is *revolutionary* because it breaks the former PCIe bottleneck for GPU interconnects. It offers several times more bandwidth and new memory-coherency capabilities that were not possible with standard PCIe or earlier interconnects. For example, Oak Ridge's Summit supercomputer (2018) uses NVLink to connect IBM POWER9 CPUs with 27,648 NVIDIA Volta V100 GPUs, yielding **8× the performance of its predecessor Titan** on only a quarter of the nodes (www.olcf.ornl.gov) (developer.nvidia.com). More recently, in 2025 Microsoft deployed a **4,608-GPU NVLink-connected cluster** on Azure (64 GB300 NVL72 racks of 72 GPUs each, with NVIDIA Grace CPUs) achieving **92.1 exaFLOPS** of FP4 inference by tying GPUs with NVLink 5 fabric (www.tomshardware.com). These breakthroughs, along with NVLink's ability to *scale* (via NVIDIA NVSwitch for intra-node all-to-all connectivity and NVLink-based networks for inter-node links), enable new classes of applications and system designs. Overall, NVLink and its ecosystem (NVSwitch, NVLink networking, NVLink Fusion, etc.) represent a paradigm shift from commodity CPU buses to **GPU-centric SMP-like architectures**, unlocking massive performance and scalability gains in **modern AI and HPC systems** (www.nextplatform.com) (developer.nvidia.com).

Introduction and Background

The emergence of GPU-accelerated computing has been a key driver of the AI and HPC revolution. Modern supercomputers and data centers use thousands of NVIDIA GPUs to **train deep neural networks**, simulate scientific models, and process large data sets. However, the **PCI Express (PCIe) bus**, the standard interconnect for attaching accelerators to CPUs in x86 servers, has become a critical bottleneck. PCIe Gen3/4, with about 16–32 GB/s per direction (roughly 32–64 GB/s bidirectional) per x16 link, cannot keep up with the massive data rates that GPUs can sustain internally. As GPU computational power and on-chip memory bandwidth have skyrocketed, feeding data between GPUs and CPUs (and between GPUs) over PCIe became a severe constraint in multi-GPU systems (nvidianews.nvidia.com) (developer.nvidia.com).

To overcome this limitation, NVIDIA developed **NVLink**, a proprietary point-to-point interconnect for GPUs and CPUs. NVLink was first announced in 2014 and publicly introduced with the Pascal-generation Tesla P100 accelerator in 2016. It was designed explicitly to increase GPU communication bandwidth and enable *unified memory* architectures. In contrast to PCIe's shared bus, NVLink provides a **scalable mesh of high-bandwidth links**, with each link capable of tens of gigabytes per second in each direction. Multiple NVLink links can be combined ("bonded") to yield very high aggregate bandwidth between devices. Crucially, NVLink also adds hardware support for cache coherence and unified virtual addressing (across CPU and GPU) in later generations, enabling GPUs and Power CPUs to access each other's memories as peers (developer.nvidia.com) (en.wikichip.org). This blurs the boundary between CPU and GPU memory, letting large "model" or "domain" data sets span both memory systems without constant housekeeping by software.

In effect, NVLink brings GPU connectivity into the realm of *symmetric multiprocessing* (SMP). As noted by The Next Platform (2016), NVLink "lashing GPUs to CPUs" is "*something as transformative as SMP was for CPUs*" (www.nextplatform.com). For the first time, GPUs in a node can be wired together with bandwidth and latency on par with accessing CPU memory. The result has been revolutionary: multi-GPU computing that previously

had to treat GPUs as isolated accelerators now becomes a tightly-coupled supercomputer on each node. NVLink's evolution has been matched to each NVIDIA GPU generation, with each new NVLink generation pushing performance far above PCIe (see **Table 1**).

NVLink Gen	GPU Architecture (Year)	Link Speed	Links per GPU	Total Bandwidth (GB/s, bidirectional)	Notes/Features
NVLink 1.0	Pascal P100 (2016)	20 GB/s (per link per direction) (40 GB/s bidirectional)	4	160 GB/s (en.wikichip.org) (developer.nvidia.com)	First NVLink; enables GPU–GPU and GPU–CPU (Power8) links; used in early NVLink-CPU systems (IBM POWER8+)
NVLink 2.0	Volta V100 (2017)	25 Gb/s per lane, 25 GB/s per direction (50 GB/s bidirectional) (en.wikichip.org)	6	300 GB/s (en.wikichip.org)	Added cache coherence and unified address; supported by IBM POWER9; Tesla V100: 6 links
NVLink 3.0	Ampere A100 (2020)	25 GB/s per direction per link (50 GB/s bidirectional)	12	600 GB/s	A100: Ampere's GPU (third-gen NVLink); 2× the total bandwidth of V100; used in DGX A100 and NVIDIA HGX systems
NVLink 4.0	Hopper H100 (2022)	25 GB/s per direction per link (50 GB/s bidirectional; two 100 Gb/s lanes per direction) (developer.nvidia.com) (developer.nvidia.com)	18	900 GB/s (developer.nvidia.com) (developer.nvidia.com)	H100: 4th-gen NVLink; about 7× PCIe Gen5 x16 bandwidth (developer.nvidia.com); enables NVLink Switch System; NVLink network
NVLink 5.0	Blackwell and Blackwell Ultra (B200/GB200, B300/GB300; 2024–25)	50 GB/s per link per direction (100 GB/s bidirectional)	18	1,800 GB/s (nvidia.com)	GB200 NVL72: 72 GPUs + 36 Grace CPUs per rack; 130 TB/s system bandwidth; NVIDIA's current specs list NVLink domains of 8 or 72 GPUs (nvidia.com)
NVLink 6.0	Vera Rubin (2026)	Not listed in NVIDIA's spec tables	36 (nvidia.com)	3,600 GB/s at launch (nvidianews.nvidia.com); 3,000 GB/s on NVIDIA's current spec pages (nvidia.com)	Vera Rubin NVL72: 72 Rubin GPUs + 36 Vera CPUs; 260 TB/s rack bandwidth at launch (216 TB/s on current spec pages); up to 10× lower inference token cost vs Blackwell; production shipments from fall 2026 (nvidianews.nvidia.com)

Table 1: Comparison of NVIDIA NVLink generations. Link speeds are per link and total bandwidths are per GPU; NVLink doubles as the GPU–GPU and GPU–CPU interconnect. NVLink 1–4 correspond to Pascal, Volta, Ampere, and Hopper GPU architectures; NVLink 5 to Blackwell; and NVLink 6 to the Vera Rubin platform, now ramping into production (en.wikichip.org) (developer.nvidia.com) (nvidianews.nvidia.com).

NVLink's raw bandwidth improvement over PCIe is dramatic. For example, NVIDIA's own press materials highlight that the Pascal P100 GPU's NVLink interconnect delivers about **160 GB/s bidirectional**, roughly 5× the throughput of a PCIe 3.0 x16 slot (~31.5 GB/s) (nvidianews.nvidia.com). Similarly, each fourth-generation NVLink link in a Hopper H100 provides **50 GB/s** (25 GB/s each way), so that with 18 NVLinks the total per-GPU interconnect bandwidth reaches **900 GB/s** (developer.nvidia.com) (developer.nvidia.com) (developer.nvidia.com), compared to 128 GB/s for a PCIe Gen5 x16 link (64 GB/s each way) (nvidia.com). In short, NVLink line-speeds are measured in tens of gigabytes per second per link, far eclipsing the externally-facing CPU bus. This lets GPUs exchange data almost as quickly as they pull from their own onboard HBM2/3 memory, unlocking new data-driven algorithms.

NVLink's design is **modular**. Each "NVLink port" comprises multiple high-speed differential pairs, and GPUs can have dozens of such lanes. NVLink links directly connect GPUs to each other (and to CPUs on supporting platforms), forming a mesh or crossbar. In the early Pascal/V100 days, GPUs in a node were connected pairwise by 2 or 4 NVLinks (in hybrid cube mesh topologies) (en.wikichip.org) (developer.nvidia.com). From Volta onward, multi-GPU servers also introduced the **NVSwitch**, a specialized on-board switch ASIC that fully wires all GPUs with full bi-directional connectivity (like an on-chip network) (developer.nvidia.com) (developer.nvidia.com). NVSwitch chips can connect six or more NVLink links per GPU and can be stacked to interconnect up to 16 or more GPUs with uniform bandwidth. These NVSwitch-based fabrics (e.g. in NVIDIA's DGX-2™ and DGX GH200 systems) effectively turn a collection of GPUs into a single gigantic HPC accelerator with shared memory semantics (en.wikichip.org) (developer.nvidia.com).

In addition, NVIDIA has extended NVLink **beyond the node**. The latest NVLink ecosystem includes *NVLink networking* features, where multiple servers can connect via NVLink speeds (e.g. custom NICs with uplinks running 4× NVLink links) (developer.nvidia.com) (developer.nvidia.com). Furthermore, NVIDIA has launched **NVLink Fusion**, a rack-scale infrastructure platform that enables hyperscalers and custom ASIC designers to integrate their own CPUs and accelerators (XPU's) with NVLink's scale-up interconnect (nvidia.com). Announced at Computex 2025, NVLink Fusion includes an NVLink 5 chiplet that third parties can embed into their silicon, and leverages **NVLink-C2C** (chip-to-chip) interconnect technology, which NVIDIA now rates at up to 6× the energy efficiency and 3.5× the area efficiency of a PCIe Gen 6 PHY (nvidia.com). Partners including Qualcomm, Fujitsu, Marvell, and MediaTek are integrating NVLink Fusion into their products (servethehome.com). This all serves to future-proof NVLink as data center fabrics grow and opens the NVLink ecosystem beyond NVIDIA-only silicon.

Collectively, NVLink and its related technologies have **revolutionized data center architecture**. They enable GPU clusters to scale with far less overhead than was previously possible. In the sections below, we delve into the technical details of NVLink, survey its performance impact and use cases, and discuss how it compares to and influences other interconnect technologies.

NVLink Architecture and Key Features

NVLink is fundamentally a *point-to-point, high-speed serial link* that is optimized for GPU-to-GPU and GPU-to-CPU communication. Unlike PCIe (a packet-based serial bus), NVLink connections are designed for **energy-efficient, tightly synchronized transfers** with lower latency and higher bandwidth. Each NVLink connection consists of multiple *lanes* (differential SerDes pairs) operating at tens of gigabits per second. Early NVLink

(Generation 1) used eight lanes per direction in each link (20 Gb/s each) (developer.nvidia.com), while later generations moved to fewer, faster lanes: four per direction at 50 Gb/s in A100 and two per direction at 100 Gb/s in H100 (developer.nvidia.com) (en.wikichip.org) (developer.nvidia.com).

Because of its point-to-point nature, **NVLink scales with topology**. For example, two GPUs directly connected by a single NVLink link can exchange ~20–50 GB/s each way (depending on generation). If the GPUs have multiple NVLinks in parallel (a “bonded” link), the aggregate throughput scales accordingly; e.g. Nvidia’s Pascal P100 could use four 20 GB/s links to yield a 160 GB/s (bi-dir) GPU-to-GPU bandwidth (en.wikichip.org). In real systems, NVIDIA arranges NVLinks in network topologies. The DGX-1 (Pascal/P100) used an NVLink “hypercube mesh,” and the DGX-2 (Volta/V100) uses NVSwitch chips to fully connect 16 GPUs in a non-blocking crossbar (developer.nvidia.com). In all cases, NVLink topology is carefully chosen to maximize connectivity and minimize any non-uniform (NUMA) delays (arxiv.org) (developer.nvidia.com).

A **critical feature** of NVLink (since Gen 2.0) is *cache-coherent shared memory*. Starting with the Volta architecture and IBM POWER9 CPU, NVLink gained hardware coherence and a unified address space (en.wikichip.org) (developer.nvidia.com). Now a GPU can directly load/store data in CPU memory, and vice versa, with coherence enforced by hardware. This capability is what makes unified memory (managed by CUDA) truly seamless across CPU/GPU on supported platforms. For example, as one study notes, a POWER9+NVLink+Volta system could treat CPU DDR main memory as a “shared L4 cache” for GPUs, dramatically improving data access for large models (www.nextplatform.com). This memory unification is a key differentiator over PCIe: it eliminates explicit DMA programming in many cases and greatly reduces CPU intervention.

NVLink also supports **peer-to-peer GPU communication** without CPU involvement. Both data transfer and atomic operations can occur directly between GPUs over NVLink. The high concurrency of NVLink means GPUs can exchange data (e.g. intermediate results in multi-GPU workloads) with minimal latency and high throughput. NVIDIA provides libraries (like **NCCL** and GPUDirect) that leverage NVLink to implement very fast all-reduce, broadcast, or scatter operations across GPUs. The result is near-optimal scaling of parallel GPU applications, a far cry from the old days when GPUs had to communicate through system memory over slow PCIe.

In practical terms, NVLink dramatically changes programming and performance. On an NVLink node, moving data between two GPUs might consume only microseconds rather than tens or hundreds of microseconds, greatly reducing synchronicity stalls. Kernels can be written as if the GPUs share memory and caches, simplifying code. NVIDIA’s projections showed many multi-GPU applications gaining **up to 2×** when using NVLink connectivity versus PCIe alone (developer.nvidia.com). In deep learning training, NVLink is essentially essential for scaling to more than a few GPUs efficiently. In HPC simulations, it lets large data structures be partitioned across GPU memories with fast interconnects for halos or global operations.

Despite being proprietary to NVIDIA GPUs, NVLink’s design is conceptually applicable everywhere. It treats the entire GPU accelerator assembly more like a fast multi-processor than a set of dumb attached devices. In fact, by combining NVLink with specialized fabrics (NVSwitch, NVLink networks), NVIDIA has demonstrated systems where all GPUs in a multi-node cluster behave like a single “super-accelerator” for specific tasks (developer.nvidia.com) (www.tomshardware.com).

NVLink Bandwidth vs. PCIe and Other Buses

One of the simplest ways to see NVLink's impact is to compare raw numbers with PCIe. For PCIe 3.0 x16, the base maximum is ~16 GB/s per direction (32 GB/s bidirectional). PCIe 4.0 doubles clock rate, giving ~32 GB/s/dir (64 bi), and PCIe 5.0 ~64 GB/s/dir (128 bi) ([developer.nvidia.com](#)) ([www.pcgamer.com](#)). By comparison, **NVLink 1.0** (Pascal) provided *80 GB/s per GPU in each direction* (four links of 20 GB/s each way) ([developer.nvidia.com](#)), already exceeding PCIe 4.0. NVLink 2.0 (Volta) delivered **25 GB/s each way** per link over six links, and NVLink 3.0 (Ampere) kept **25 GB/s per direction per link** but doubled the link count to twelve, for 600 GB/s per GPU ([developer.nvidia.com](#)) ([docs.nersc.gov](#)) ([developer.nvidia.com](#)). In H100 NVLink 4.0, each link is **25 GB/s per direction** (50 GB/s bidirectional), achieved with two 100 Gb/s lanes per direction ([developer.nvidia.com](#)) ([developer.nvidia.com](#)). With 18 such NVLinks attached to a GPU, H100 achieves ~900 GB/s bidirectional. The Blackwell-generation NVLink 5.0 doubles this again to **1,800 GB/s** per GPU via 18 links, and the GB200 NVL72 rack system aggregates 130 TB/s of total system bandwidth across 72 GPUs ([nvidia.com](#)). Next, NVLink 6.0 in the Vera Rubin platform, with production shipments starting in fall 2026, was announced at **3,600 GB/s** per GPU and 260 TB/s per Vera Rubin NVL72 rack, which NVIDIA described as more bandwidth than the entire internet ([nvidianews.nvidia.com](#)); NVIDIA's current specification pages list 3,000 GB/s per GPU over 36 links and 216 TB/s per rack ([nvidia.com](#)) ([nvidia.com](#)).

To contrast, a GPU's access to host memory via PCIe Gen3 x16 is only ~31.5 GB/s bi (16 GB/s per direction) ([nvidianews.nvidia.com](#)). Even PCIe5 x16 (~128 GB/s bi) is about one-seventh of an H100's 900 GB/s of NVLink 4.0 bandwidth ([nvidia.com](#)) ([developer.nvidia.com](#)). Hence, NVIDIA claims NVLink yields “at least 5×” the effective bandwidth of PCIe Gen3 x16 ([developer.nvidia.com](#)). Practically this means multi-GPU data movements that would saturate PCIe links can proceed with headroom on NVLink, avoiding the bottlenecks that once forced GPUs to idle.

For example, NVIDIA's 2016 press release on the Tesla P100 states: “*NVLink delivers 160 GB/sec of bi-directional interconnect bandwidth, compared to PCIe x16 Gen3 that delivers 31.5 GB/sec of bi-directional bandwidth.*” ([nvidianews.nvidia.com](#)). In lay terms, four NVLink 1.0 links (4×40 GB/s = 160 GB/s) gave a Tesla P100 system roughly five times the link bandwidth of PCIe. Similarly, as PCIe progressed, so did NVLink's lead: NVLink 4.0's 900 GB/s per GPU is **~7× the 128 GB/s of a PCIe Gen5 x16 link**, as NVIDIA states for H100 ([developer.nvidia.com](#)), and NVLink delivers that with lower latency and overhead ([developer.nvidia.com](#)) ([developer.nvidia.com](#)).

These raw numbers only tell part of the story, since NVLink's architecture also avoids *congestion and arbitration* inherent in PCIe. With PCIe, multiple GPUs share lanes to the CPU/chipset, so contention causes uneven latency. NVLink's point-to-point mesh treats each link as dedicated hardware, avoiding cross-traffic contention. In practice, this leads to much more predictable and higher utilization for multi-GPU workloads ([arxiv.org](#)) ([developer.nvidia.com](#)).

NVSwitch and All-to-All Connectivity

While NVLink itself provides GPU-to-GPU links in pairs, NVIDIA recognized that complex problems often need **all-to-all GPU communication**. To achieve this, NVIDIA introduced a companion switch chip called **NVSwitch**. The first-generation NVSwitch chip aggregates 18 NVLink ports (each connecting to a GPU or

another NVSwitch) and provides full crossbar connectivity among its ports; the third-generation NVSwitch used with H100 has 64 NVLink 4 ports ([developer.nvidia.com](#)) ([developer.nvidia.com](#)). In a DGX-2 node, for example, NVSwitch chips fully connect 16 V100 GPUs, so any pair of GPUs can communicate at the full 300 GB/s (aggregating multiple NVLinks) ([developer.nvidia.com](#)) ([en.wikichip.org](#)).

NVSwitch scales this further: GPUs connected through NVSwitch see dramatically higher all-to-all bandwidth than a mesh of direct links. NVIDIA's own benchmark examples illustrate this (see **Table 2**). For 2 GPUs, a point-to-point NVLink pair might yield ~128 GB/s total between them, but with NVSwitch (on an 8-GPU fabric) the effective pairwise bandwidth per GPU is 900 GB/s ([developer.nvidia.com](#)). Even as GPU count rises, NVSwitch holds the per-GPU interconnect bandwidth nearly constant (900 GB/s for 2, 4, or 8 GPUs), whereas a fixed topology degrades (e.g. 384 GB/s total among 4 GPUs without switch, as shown below).

GPUs	Direct NVLink (P2P) Total BW	NVSwitch All-to-All BW
2 GPUs	128 GB/s (developer.nvidia.com)	900 GB/s (developer.nvidia.com)
4 GPUs	3×128 = 384 GB/s (developer.nvidia.com)	900 GB/s (developer.nvidia.com)
8 GPUs	7×128 = 896 GB/s (developer.nvidia.com)	900 GB/s (developer.nvidia.com)

Table 2: GPU-to-GPU aggregate bandwidth (per GPU) for point-to-point NVLink vs. NVSwitch (source: NVIDIA). Without NVSwitch, bandwidth grows linearly with additional GPUs (but each GPU only connects to one other at 128 GB/s). With NVSwitch, each GPU maintains ~900 GB/s full bisection bandwidth to all others ([developer.nvidia.com](#)).

As Table 2 demonstrates, the NVSwitch **completely flattens the communication bottleneck**. With NVSwitch, an 8-GPU node behaves as if each GPU has a colossal 900 GB/s fabric to any other GPU (which is consistent with 3.6 TB/s bisection bandwidth for 8 H100 GPUs ([developer.nvidia.com](#)) ([developer.nvidia.com](#))). This uniformity enables very efficient parallel algorithms (e.g. large all-reduce or model-parallel operations in deep learning) that would otherwise saturate links or incur multi-hop delays. Indeed, NVIDIA's analysis shows each H100 keeping its full 900 GB/s whether two, four or eight GPUs are communicating, while a point-to-point design splits that bandwidth into 128 GB/s connections ([developer.nvidia.com](#)). In practical terms, a 20 GB exchange that takes about 150 ms over a 128 GB/s point-to-point link takes about 22 ms with NVSwitch, vastly improving throughput ([developer.nvidia.com](#)).

Notably, NVSwitch is only one way NVIDIA extends NVLink. In 2022–2025, NVIDIA has introduced **NVLink Network** features and specialized inter-node fabrics to carry NVLink-like traffic between servers. For example, the H100-generation “NVLink Switch System” extended NVLink beyond the server, connecting up to 256 GPUs with 57.6 TB/s of all-to-all bandwidth ([developer.nvidia.com](#)) ([developer.nvidia.com](#)). These innovations point to a future where NVLink's high bandwidth can span entire clusters, not just individual servers.

Performance and Impact

The practical impact of NVLink on real workloads is profound and well-documented. Multiple independent studies and industry reports attest that NVLink-powered systems achieve **greatly accelerated performance** on multi-GPU tasks, both by boosting raw data throughput and by improving software efficiency.

For example, **Li et al. (2019)** performed an extensive evaluation of modern GPU interconnects (including NVLink v1/v2, PCIe, NVSwitch, etc.) across systems from DGX servers to DOE supercomputers ([arxiv.org](#)). They observed significant “NUMA effects” caused by NVLink topology: the choice of which GPUs are linked to which matters for performance. Importantly, they found that with NVLink, **communication latencies drop and bandwidth rises**, which translates to higher overall application throughput. Applications such as convolutional neural networks saw up to *2× or more reduction in training time* simply from the improved interconnect, as NVLink allowed more efficient peer-to-peer data transfers. (See also NVIDIA’s own modeling: they projected NVLink could improve many multi-GPU apps by “up to 2×” through faster GPU peer-to-peer communication ([developer.nvidia.com](#)).)

In tightly-coupled supercomputing benchmarks, NVLink’s benefits are clear. Summit, the DOE’s flagship system, was built from nodes that each pair two IBM POWER9 CPUs with six NVIDIA V100 GPUs over NVLink ([www.olcf.ornl.gov](#)) ([developer.nvidia.com](#)). ORNL engineers report that Summit achieved **8× the performance of Titan** (its NVIDIA K20x-based predecessor) while using *one-quarter the hardware*, largely thanks to NVLink-enhanced GPU scaling ([www.olcf.ornl.gov](#)). Summit’s coherent memory meant researchers could tackle unprecedented problems: training massive neural networks on nearly 200,000 GPU cores, simulating fusion plasmas at record resolution, etc., all of which depend on rapid data movement. Similarly, Sierra (LLNL’s V100/Power9 system) and Perlmutter (NERSC’s A100/AMD CPU system) exploited NVLink to push HPC workloads forward.

Even in smaller clusters, NVLink shows large wins. In one report, introducing NVLink in a small-scale Volta system yielded much better Linpack (HPL) efficiency relative to CPU-CPU InfiniBand communication. The ability for GPUs to collectively act on data (rather than funneling through CPUs) doubled per-node scalability ([www.nextplatform.com](#)). In machine learning, NVidia’s DGX A100 (8×A100 in a node) shows near-linear scaling up to 8 GPUs on large models, whereas older systems on PCIe often stalled. And NVIDIA’s developer blog presents NVLink and NVSwitch as what makes multi-GPU LLM inference efficient ([developer.nvidia.com](#)).

Hardware vendors outside NVIDIA have also recognized NVLink’s impact. IBM’s Power9 was explicitly designed with NVLink ports to keep up with NVIDIA GPUs ([nvidianews.nvidia.com](#)) ([developer.nvidia.com](#)). Server OEMs like Supermicro and Dell now offer motherboards with 2× or 4× NVSwitch connectivity for NVLink GPU servers. Even cloud providers (Azure, Oracle, Google Cloud) have introduced NVLink-enabled GPU instances (e.g. Microsoft’s H100 instances reportedly use NVLink 4.0 and NVSwitch). These offerings are motivated by customer demand for higher GPU-GPU bandwidth in AI workloads.

From a software perspective, NVLink also boosts productivity. Features like **CUDA Unified Memory** and **GPUDirect** rely on fast interconnects to work well. In fact, a 2019 study of CUDA Unified Memory found that a POWER9+NVLink+Volta platform saw up to *34% performance gains* from using advanced memory advising/prefetching techniques, exactly because NVLink made the CPU-GPU data highway so fast ([arxiv.org](#)). In other words, NVLink can provide more bandwidth for CPU–GPU transfers on platforms that support it, but Unified Memory performance still depends on data placement, page migration, access patterns, and the specific platform. NVLink does not by itself make host and GPU memory cache-coherent.

Administrator and user feedback further attests to NVLink’s revolutionary effect. Engineers often remark that NVLink-equipped nodes “*feel different*” from older GPU servers: tasks that used to be limited by interconnect suddenly run as if they have a shared memory architecture. HPC centers note that NVLink nodes remove a

major bottleneck for codes like AMG or CG that need frequent all-gather operations on GPU data. Overall, every published report from major AI/HPC centers since 2018 highlights NVLink connectivity as central to enabling the projects (e.g. training GPT and large transformer models in weeks, running exascale-scale simulations, etc.).

Case Studies and Real-World Examples

DOE Supercomputers (Summit, Sierra, Perlmutter): Arguably the most prominent case study of NVLink's impact is in the U.S. Department of Energy's supercomputers. *Summit* (Oak Ridge, online 2018) pairs two IBM Power9 CPUs with six NVIDIA V100 GPUs per node, all connected together with NVLink. This design yields a node memory of ~600GB usable by all processors coherently (www.olcf.ornl.gov). Summit, rated at about 200 petaflops peak and 148.6 petaflops on the HPL benchmark, ranked first on the TOP500 list from June 2018 through November 2019 (top500.org) and enabled breakthroughs from fusion modeling to COVID-19 simulations. Summit's LINPACK scaling and actual application speedups have been attributed directly to its NVLink fabric. Similarly, *Sierra* (Lawrence Livermore, 2018) was an IBM-built machine with the same NVLink/Power9/V100 architecture, rated at 125 petaflops peak; LLNL has since decommissioned it (hpc.llnl.gov). The DOE's successor, *Perlmutter* (2021-22), uses NVIDIA A100 GPUs with 3rd-gen NVLink; its GPU partition is rated at 69.5 petaflops of peak FP64 performance (140 petaflops using FP64 tensor cores) and has been heavily used for AI workloads at NERSC (docs.nersc.gov) (docs.nersc.gov). In all these systems, NVLink was explicitly cited as a necessary feature. For example, NVIDIA's blog describes NVLink as “the node integration interconnect for both the Summit and Sierra pre-exascale supercomputers”, enabling fast mutual memory access between IBM CPUs and NVIDIA GPUs (developer.nvidia.com).

NVIDIA DGX Supercomputers: NVIDIA's own DGX series of supercomputers also rely on NVLink and NVSwitch. The DGX-1 (2016–2017) used 8×Pascal or Volta GPUs with NVLink meshes. The DGX A100 (2020) features 8×A100 GPUs connected by six second-generation NVSwitches (4.8 TB/s of aggregate bidirectional bandwidth), giving each GPU 600 GB/s (developer.nvidia.com). The Blackwell-generation **DGX B200** (2024) connects 8×B200 GPUs via NVLink 5 for up to 14.4 TB/s of aggregate GPU-to-GPU bandwidth per node. Beyond individual nodes, the **GB200 NVL72** rack (2025) put 72 Blackwell GPUs and 36 Grace CPUs in one NVLink 5 and NVSwitch domain, delivering 130 TB/s of system bandwidth as a single, unified accelerator. The Blackwell Ultra **GB300 NVL72** keeps the same 72-GPU NVLink 5 design, and the Vera Rubin NVL72 carries the 72-GPU domain to NVLink 6. CoreWeave was the first cloud provider to deploy GB300 NVL72 systems for customers, in July 2025 (coreweave.com). NVIDIA's DGX SuperPOD reference architecture can cluster multiple DGXs with NVLink networking, effectively scaling to hundreds of GPUs with near-linear performance.

Cloud AI Clusters: Cloud providers are rapidly deploying NVLink-equipped GPU instances at scale. In October 2025, Microsoft announced the world's first *supercomputer-scale* GPU cluster on Azure: 64 GB300 NVL72 racks totaling 4,608 Nvidia GB300 GPUs, with NVLink 5 inside each 72-GPU rack (www.tomshardware.com). This machine delivers **92.1 exaFLOPS** (FP4) and is dedicated to OpenAI workloads, from running reasoning models to training. Microsoft has also announced plans for large-scale Rubin deployments (azure.microsoft.com). CoreWeave became the first cloud provider to deploy GB300 NVL72 systems in July 2025, with instances available through their Kubernetes service (coreweave.com).

Together AI, in partnership with 5C, is bringing GB200 and GB300-equipped AI factories online in multiple U.S. locations. These deployments highlight how NVLink scales: 36 Grace CPUs per NVL72 rack aggregate memory and control, while all 72 GPUs act as a single accelerator thanks to NVLink 5 and Quantum InfiniBand fabric (www.tomshardware.com).

Machine Learning Speedups: In practical ML workloads, NVLink-enabled hardware consistently reduces training time. For instance, training on NVLink-connected multi-GPU nodes runs faster than on similar nodes without NVLink when the code is communication-bound, because gradients and activations cross between GPUs over NVLink rather than PCIe. NVIDIA describes NVLink and NVSwitch as the basis of high-throughput, cost-effective large model inference (developer.nvidia.com). NVLink's ability to feed GPUs data (gradients, activations, embeddings) so quickly often changes what is feasible: very large batch training, parameter servers in memory, or beam-search decoding over multiple GPUs all become tractable. For inference, especially of large language models, NVLink + NVSwitch is described by NVIDIA as “critical” for high throughput. In a 2024 analysis of Llama 3.1 70B inference on H100 GPUs, NVIDIA calculated that one 20 GB all-to-all reduction takes ~150 ms over a 128 GB/s point-to-point link but ~22 ms over NVSwitch at 900 GB/s (developer.nvidia.com). This 7× improvement directly translates to faster interactive response and lower cost for real-time AI services (developer.nvidia.com).

Enterprise and Workstation GPUs: Even outside datacenters, NVLink is making inroads. Professional workstations (e.g. NVIDIA RTX A6000/A5000) include NVLink bridges to link pairs of GPUs for visualization and compute tasks. NVIDIA's bridge guide lists the RTX A6000, A5500, A5000 and A4500, the A800 40GB Active and the older Quadro RTX 8000/6000 as bridge-capable, at up to 112 GB/s per bridge for most of them (nvidia.com), and two bridged RTX A6000 cards can pool 96 GB of memory (nvidia.com). The newer RTX 6000 Ada Generation lists no NVLink in its specifications (nvidia.com), so a two-GPU NVLink workstation today means Ampere-generation or older cards. This enables professionals to run complex GPU pipelines (video editing, 3D simulation) across two GPUs without visible slowdown. And with the rise of GPU-accelerated databases (e.g. BlazingSQL) or graph analytics (cuGRAPH), NVLink allows multi-GPU processing on a single node to be effectively unified, dramatically reducing data shuffle costs.

Large Graph AI Case Study: Consider a graph neural network (GNN) training on a graph of billions of nodes. Without NVLink, each GPU might only hold a fragment of the graph in memory, and edges crossing GPUs would incur slow PCIe transfers. With NVLink, GPUs can share large contiguous graph partitions. Node embeddings can then be shuffled across GPUs over NVLink without burdening the CPU, which is where an NVLink system gains over a PCIe system (developer.nvidia.com). Similarly, in numerical simulations (e.g. climate modeling), large state fields can be decomposed across GPUs; NVLink ensures boundary exchanges are fast, enabling higher resolution or faster timesteps than on older nodes.

Quantitative Metrics: To quantify, look at memory bandwidth and interconnect throughput: An H100 SXM GPU has 3.35 TB/s of local HBM3 memory bandwidth (nvidia.com). Without NVLink, if it had to fetch slow data from another GPU's memory, it would be restricted by PCIe (tens of GB/s). With NVLink 4.0, those transfers can happen at ~900 GB/s, a bit more than a quarter of HBM speed, meaning many algorithms become nearly memory-bound rather than link-bound. NVIDIA's 2024 inference analysis shows the practical effect: with NVSwitch, each GPU-to-GPU reduction in multi-GPU LLM inference takes about one-seventh of the time it takes over point-to-point links (developer.nvidia.com).

Taken together, these examples and data concretely show that NVLink actually *works*: it removes a key bottleneck. The speedups are not marginal; they often enable new science or reduce costs dramatically. In essence, NVLink has shifted the narrative from “data must go through a slow PCIe host” to “GPUs can talk fast directly, like co-processors on a common bus.”

Ecosystem and Competitive Landscape

NVLink is proprietary to NVIDIA GPUs, which means it is not a universal industry standard. However, its success has influenced the broader ecosystem. Other interconnects are emerging (or adapting) in response:

- AMD Infinity Fabric / XGMI:** AMD's GPU interconnect (XGMI) and CPU interconnect (Infinity Fabric) similarly enable high-speed links within AMD's own ecosystem. AMD Instinct MI350 Series platforms connect eight fully connected MI355X or MI350X GPU OAM modules with **4th Gen AMD Infinity Fabric** ([amd.com](https://www.amd.com)). AMD's upcoming MI355X DLC rack targets 2.4 exaFLOPS at FP4 precision with 128 GPUs (www.techradar.com). For multi-node scale-out, AMD is pivoting from Infinity Fabric to the open **Ultra Accelerator Link (UALink)** standard and Ultra-Ethernet networking, with MI400 series rack-scale solutions expected in H2 2026 to compete with NVIDIA's NVL72/NVL144 systems ([semianalysis.com](https://www.semianalysis.com)).
- CXL (Compute Express Link) and Open Standards:** The CXL ecosystem has advanced rapidly. The **CXL 4.0 specification** was released on November 18, 2025, doubling the signaling rate from 64 GT/s to 128 GT/s and adding bundled ports that combine device ports for more bandwidth (computeexpresslink.org) (computeexpresslink.org). CXL targets memory expansion, pooling and sharing for AI and HPC workloads. While CXL and NVLink serve complementary roles (CXL for CPU-centric memory expansion, NVLink for GPU-to-GPU scale-up), NVIDIA's NVLink-C2C supports the CXL protocol for interoperability ([nvidia.com](https://www.nvidia.com)), and the convergence of these technologies is expected to shape next-generation data center architectures. The **UALink (Ultra Accelerator Link)** consortium, backed by AMD, Intel, Google, Microsoft, and others, published the UALink 200G 1.0 specification in April 2025, targeting 200 GT/s per lane and 800 Gbps per four-lane port (ualinkconsortium.org), and in April 2026 ratified UALink 2.0, adding in-network compute, a chiplet definition and manageability (ualinkconsortium.org). UALink-based hardware is expected in the 2026–2027 timeframe and represents the most serious open-standard challenge to NVLink's dominance in GPU-to-GPU interconnect.
- New Unified Fabrics:** As noted in media, new proposals like **Huawei's UB-Mesh** (Hot Chips 2025) aim to unify all interconnects (PCIe, NVLink, TCP/IP) into one massive mesh fabric supporting up to 10 Tbps per chip (www.tomshardware.com). If widely adopted, such standards could eventually supersede vendor-specific links. However, these are nascent and not yet mainstream.

For now, NVLink's main competitors remain point-to-point link protocols in other platforms. Unfortunately, NVLink's proprietary nature means you can't plug an NVIDIA GPU into an AMD Infinity cluster; but conversely, the maturity of NVLink and its software support (CUDA, CUDA-aware MPI, NCCL) give it a huge first-mover advantage in AI/HPC. Organizations building performance-demanding GPU clusters overwhelmingly choose NVIDIA NVLink-enabled systems because so many applications are already optimized for it.

Challenges and Limitations

No technology is without downsides. NVLink's revolutionary performance comes at some cost and complexity:

- **Vendor Lock-in:** NVLink has historically been exclusive to NVIDIA GPUs and select CPUs (IBM Power, NVIDIA Grace). However, the 2025 launch of NVLink Fusion partially addresses this by allowing third-party CPUs and accelerators from Qualcomm, Fujitsu, Marvell, and MediaTek, and later Intel and AWS, to integrate into NVLink fabrics. Still, if a datacenter adopts NVLink, switching to a fully non-NVIDIA platform would mean losing that interconnect. The emerging UALink standard offers a potential migration path, though it remains in early stages.
- **Topology Complexity:** Designing NVLink networks (especially NVSwitch fabrics) is more complicated than using a commodity bus. Engineers must carefully lay out cables or boards, ensure proper link bonding, and account for thermal/power of dozens of high-speed SerDes. Mistakes in NVLink topology can lead to non-uniform bandwidth or even failure to boot. As a result, only specialized server designs (DGX, HGX and similar HPC designs) typically expose NVLink; generic motherboards don't.
- **Scalability Ceiling:** NVLink's scale-up reach has expanded dramatically, from 8 GPUs per node to 72 GPUs per rack with NVL72, and NVIDIA's roadmap cites future NVLink domain sizes of up to 1,152 (nvidia.com). However, connecting thousands of GPUs across multiple racks still requires higher-level fabrics like InfiniBand or Ethernet. Even NVIDIA's largest deployments (e.g. Microsoft's 4,608-GPU cluster) use Quantum InfiniBand between NVL72 racks. For truly massive clusters, NVLink remains the intra-rack backbone while separate network technologies handle inter-rack communication.
- **Cost:** NVLink and NVSwitch hardware is expensive. A single NVSwitch chip costs hundreds of dollars, and systems like DGX have dozens of them. The cables and connectors for NVLink must be custom and precise. Thus systems with NVLink are generally higher-priced than commodity PCIe GPU servers, limiting NVLink's use to compute-heavy budgets (like national labs, cloud HPC, top ML research).
- **Power Consumption:** Running tens of lanes at 50–100 Gbps consumes significant power. NVLink interfaces and NVSwitch chips add thermal and power overhead. In tight datacenter environments, cooling NVLink-enabled nodes can be challenging. NVIDIA and partners mitigate this with large chassis and advanced cooling, but it remains a consideration.

Future Directions

Looking ahead, NVLink continues to evolve. Some key anticipated developments:

- **NVLink 5, NVLink 6, and the Rubin Platform:** NVLink 5.0, shipping with the Blackwell architecture (GB200/GB300), delivers **1,800 GB/s** per GPU via 18 links at 100 GB/s each, doubling Hopper's 900 GB/s. The GB200 NVL72 rack aggregates 130 TB/s of system bandwidth, and NVIDIA's current specs list fifth-generation NVLink domains of 8 or 72 GPUs (nvidia.com) (nvidia.com). At CES 2026, NVIDIA unveiled the next-generation **Rubin platform** with **NVLink 6.0**, which doubles per-GPU bandwidth again to **3,600 GB/s**. The Vera Rubin NVL72, combining 72 Rubin GPUs and 36 Vera CPUs, was announced with 260 TB/s of rack-level bandwidth (NVIDIA's current spec pages list 3,000 GB/s per GPU and 216 TB/s per rack (nvidia.com)), up to 10× inference cost reduction and 4× fewer GPUs needed to train MoE

models compared to Blackwell (nvidianews.nvidia.com). In May 2026 NVIDIA said Vera Rubin was ramping into full production, with production shipments starting in fall 2026 (nvidianews.nvidia.com) (tomshardware.com).

- **NVLink Fusion and Chip-to-Chip Integration:** NVIDIA's NVLink Fusion platform, launched at Computex 2025, is now a reality, not just a roadmap item. It enables third-party CPUs and accelerators from partners like Qualcomm, Fujitsu, Marvell, and MediaTek, and since then Intel and AWS (Trainium4) (nvidianews.nvidia.com) (blogs.nvidia.com), to integrate directly into NVLink's scale-up fabric via an embeddable NVLink 5 chiplet (servethehome.com). The underlying **NVLink-C2C** (chip-to-chip) technology delivers up to 6× the energy efficiency and 3.5× the area efficiency of a PCIe Gen 6 PHY, by NVIDIA's current figures, enabling tightly integrated multi-die packages (nvidia.com). This approach allows hyperscalers to build semi-custom AI infrastructure while retaining NVLink's bandwidth advantages.
- **Networking Innovation:** NVIDIA is aggressively pushing NVLink performance beyond the rack. At GTC 2025, NVIDIA unveiled post-copper **photonics networking gear** targeting 1.6 Tbps link speeds to connect millions of GPUs (www.techradar.com). The Rubin platform's networking architecture is designed from the ground up for rack-to-rack NVLink-speed communication (spectrum.ieee.org). Combined with Quantum InfiniBand and the upcoming NVLink Switch System expansions, the goal is a future where entire data centers operate as a single unified accelerator.
- **Open Standards and Competition:** The open-standard landscape has matured considerably. The **UALink 200G 1.0 specification** was published in April 2025, backed by AMD, Intel, Google, Microsoft, Meta, and Broadcom, targeting 200 GT/s per lane for GPU-to-GPU interconnect (ualinkconsortium.org). Upscale AI is targeting Q4 2026 for the first commercial UALink switch products (hpcwire.com). UALink 2.0 followed in April 2026 (ualinkconsortium.org). Meanwhile, CXL 4.0 (released November 2025) doubles the signaling rate to 128 GT/s and adds bundled ports for higher-bandwidth connections (computeexpresslink.org). NVIDIA has responded by opening NVLink via the Fusion program, effectively hedging between proprietary dominance and ecosystem participation. The coming years will likely see both NVLink and open standards coexist, with NVLink maintaining its performance lead while UALink and CXL expand the overall ecosystem.

In sum, NVIDIA NVLink was revolutionary at its debut and continues to lead the charge. It effectively “rewrote the rules” for GPU system design, and everything that followed had to respond. Its evolution from a node-local GPU link providing 160 GB/s (2016) to the core of rack-scale AI supercomputers rated at 260 TB/s at launch (Vera Rubin, 2026) marks a major technological shift. With NVLink Fusion opening the ecosystem to third-party silicon, UALink emerging as an open-standard competitor, and photonics poised to extend NVLink speeds across data centers, the future of high-speed GPU interconnect remains both intensely competitive and indispensable for cutting-edge computation.

Conclusion

NVLink transforms GPU computing by obliterating the old PCIe bottleneck. It supplies the raw link bandwidth that modern AI and HPC applications require, and it does so with new software-friendly features (shared unified memory, cache coherence) that simply were not available before. The “revolution” is evident wherever NVLink is deployed: multi-GPU training is more than twice as fast, massive simulations complete weeks earlier, and once-infeasible problems become solvable. As NVIDIA has stated, NVLink has enabled supercomputers that would not have been possible with traditional interconnects (developer.nvidia.com).

From an academic and engineering viewpoint, NVLink illustrates the value of co-designing an interconnect with GPU hardware and system software. NVLink, NVSwitch, network fabrics, NVLink Fusion, and NVLink-C2C continue to evolve, while deployments and performance characteristics vary by system and workload.

As NVIDIA's P100 launch announcement put it, NVLink delivered **“a 5x acceleration in bandwidth compared to today's best-in-class solution”** (nvidianews.nvidia.com) (developer.nvidia.com). It raised the bar several-fold in one step, not incrementally. In the ever-evolving landscape of high-performance computing, NVLink stands as a watershed, enabling the next generation of breakthroughs in science and AI that simply could not be achieved without it.

References: All claims and data above are documented in NVIDIA technical publications, peer-reviewed evaluations, and industry reports (developer.nvidia.com) (nvidianews.nvidia.com) (developer.nvidia.com) (developer.nvidia.com) (www.olcf.ornl.gov) (www.tomshardware.com) (arxiv.org) (arxiv.org), as cited.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. <https://nvidianews.nvidia.com/news/nvidia-delivers-massive-performance-leap-for-deep-learning-hpc-applications-with-nvidia-tesla-p100-accelerators-6622437>
2. <https://developer.nvidia.com/blog/how-nvlink-will-enable-faster-easier-multi-gpu-computing/>
3. <https://intuitionlabs.ai/articles/mixture-of-experts-moe-models>
4. <https://intuitionlabs.ai/articles/molecular-modeling-simulation-software>
5. <https://www.nvidia.com/en-us/data-center/nvlink/>
6. <https://nvidianews.nvidia.com/news/rubin-platform-ai-supercomputer>
7. <https://www.nvidia.com/en-us/data-center/vera-rubin-nvl72/>
8. <https://nvidianews.nvidia.com/news/vera-rubin-full-production-agentic-ai-factory>
9. <https://nvidianews.nvidia.com/news/nvidia-and-intel-to-develop-ai-infrastructure-and-personal-computing-products>
10. <https://blogs.nvidia.com/blog/nvlink-fusion-nvbm-custom-high-bandwidth-memory/>
11. <https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/>
12. <https://www.tomshardware.com/tech-industry/artificial-intelligence/microsoft-deploys-worlds-first-supercomputer-scale-gb300-nvl72-azure-cluster-4-608-gb300-gpus-linked-together-to-form-a-single-unified-accelerator-capable-of-1-44-pflops-of-inference>
13. <https://intuitionlabs.ai/articles/openai-stargate-datacenter-details>
14. <https://www.nextplatform.com/2016/05/04/nvlink-takes-gpu-acceleration-next-level/amp/>

15. <https://developer.nvidia.com/blog/nvidia-nvlink-and-nvidia-nvswitch-supercharge-large-language-model-inference/>
16. <https://intuitionlabs.ai/articles/key-innovations-behind-chatgpt>
17. <https://en.wikichip.org/wiki/nvidia/nvlink>
18. <https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/>
19. <https://developer.nvidia.com/blog/upgrading-multi-gpu-interconnectivity-with-the-third-generation-nvidia-nvswitch/?ncid=so-nvsh-708451>
20. <https://www.nvidia.com/en-us/data-center/gb200-nvl72/>
21. <https://www.nvidia.com/en-us/data-center/h100/>
22. <https://www.nvidia.com/en-us/data-center/nvlink-fusion/>
23. <https://www.nvidia.com/en-us/data-center/nvlink-c2c/>
24. <https://www.servethehome.com/nvidia-announces-nvlink-fusion-bringing-nvlink-to-third-party-cpus-and-accelerators/>
25. <https://developer.nvidia.com/blog/inside-pascal/>
26. <https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/>
27. <https://developer.nvidia.com/blog/nvswitch-accelerates-nvidia-dgx2/>
28. <https://arxiv.org/abs/1903.04611>
29. <https://www.nextplatform.com/2017/12/15/nvlink-shines-power9-ai-hpc-tests/>
30. <https://www.pcgamer.com/hardware/up-to-700-percent-faster-than-any-interface-in-your-gaming-pc-pci-sig-announces-the-specification-goals-for-pcie-8-0/>
31. <https://docs.nersc.gov/systems/perlmutter/architecture/>
32. <https://arxiv.org/abs/1910.09598>
33. <https://www.top500.org/system/179397/>
34. <https://hpc.llnl.gov/hardware/compute-platforms/sierra>
35. <https://developer.nvidia.com/blog/defining-ai-innovation-with-dgx-a100/>
36. <https://www.coreweave.com/news/coreweave-becomes-first-cloud-provider-to-deploy-nvidia-gb300-nvl72-platform>
37. <https://azure.microsoft.com/en-us/blog/microsofts-strategic-ai-datacenter-planning-enables-seamless-large-scale-nvidia-rubin-deployments/>
38. <https://www.nvidia.com/en-us/products/workstations/nvlink-bridges/>
39. <https://www.nvidia.com/en-us/products/workstations/rtx-a6000/>
40. <https://www.nvidia.com/en-us/products/workstations/rtx-6000/>
41. <https://www.amd.com/en/products/accelerators/instinct/mi350.html>
42. <https://www.techradar.com/pro/amd-sheds-more-light-on-128-gpu-mi355x-dlc-rack-that-will-deliver-2-4-exaflop-at-fp4-precision-heres-how-it-compares-with-nvidias-flagship-vera-rubin>
43. <https://semianalysis.com/2025/06/13/amd-advancing-ai-mi350x-and-mi400-ualoe72-mi500-ual256/>
44. https://computeexpresslink.org/wp-content/uploads/2025/11/CXL_4.0-Specification-Release_FINAL_Website-Copy.pdf
45. https://computeexpresslink.org/wp-content/uploads/2025/02/CXL_Q1-2025-Webinar-Presentation_FINAL.pdf
46. <https://ualinkconsortium.org/>
47. https://ualinkconsortium.org/wp-content/uploads/2026/04/UALink-2.0-Specification-PR_FINAL.pdf

48. <https://www.tomshardware.com/tech-industry/artificial-intelligence/huawei-to-open-source-its-ub-mesh-data-center-scale-interconnect-soon-details-technical-aspects-one-interconnect-to-rule-them-all-is-designed-to-replace-everything-from-pcie-to-tcp-ip>
49. <https://www.tomshardware.com/pc-components/gpus/nvidia-launches-vera-rubin-nvl72-ai-supercomputer-at-ces-promises-up-to-5x-greater-inference-performance-and-10x-lower-cost-per-token-than-blackwell-coming-2h-2026>
50. <https://www.techradar.com/pro/nvidia-is-planning-post-copper-1-6tbps-network-tech-to-connect-millions-of-gpus-as-it-unveils-photonics-networking-gear-at-gtc-2025>
51. <https://spectrum.ieee.org/nvidia-rubin-networking>
52. <https://www.hpcwire.com/2025/12/02/upscale-ai-eyes-late-2026-for-scale-up-ualink-switch/>

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore [AI enablement](#), [pharma and biotech software development](#), [data engineering and BI](#), and [Veeva services](#). [Contact IntuitionLabs](#) to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: <https://intuitionlabs.ai>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.