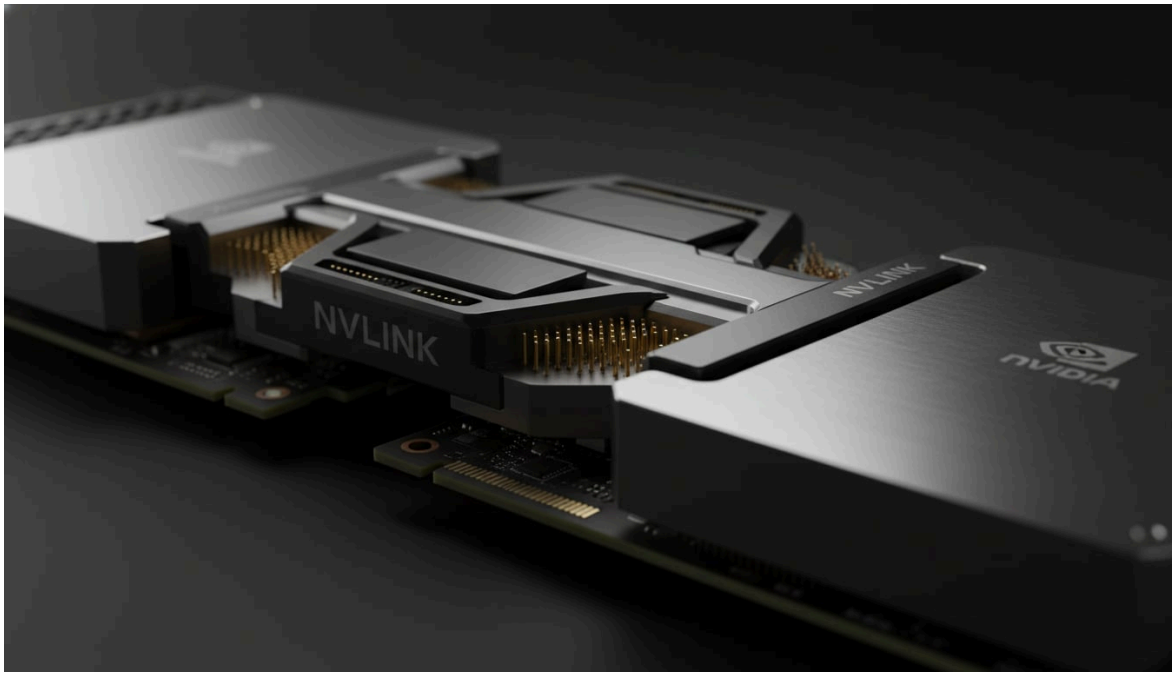
NVIDIA NVLink Explained: A Guide to the GPU Interconnect

By Adrien Laurent, CEO at IntuitionLabs • 10/22/2025 • 30 min read

nvidia nvlink | gpu interconnect | nvswitch | pcie vs nvlink | multi-gpu systems | gpu bandwidth | high performance computing | ai hardware



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 link ([1] nvidianews.nvidia.com) ([2] developer.nvidia.com). By scaling to multiple NVLink lanes per GPU, NVLink delivers **order-of-magnitude higher data rates** than traditional PCIe (e.g. an NVIDIA Tesla P100 achieves 160 GB/s bidirectional vs ~32 GB/s over PCIe Gen3×16 ([1] 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 ([2] 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).

In short, NVLink is *revolutionary* because it breaks the former PCIe bottleneck for GPU interconnects. It offers orders of magnitude greater 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 ([3] www.olcf.ornl.gov) ([2] developer.nvidia.com). More recently, in 2025 Microsoft deployed a **4,608-GPU NVLink-connected cluster** on Azure (72 GPUs per rack with IBM Grace CPUs) achieving **92.1 exaFLOPS** of FP4 inference by tying GPUs with NVLink 5 fabric ([4] 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 ([5] www.nextplatform.com) ([6] 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 of bidirectional throughput 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 ([1] nvidianews.nvidia.com) ([7] 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 ([2] developer.nvidia.com)

([8] 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"* ([5] 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 ([9] en.wikichip.org) ([2] 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 GHz (GT/s), 25 GB/s per direction (50 GB/s bidirectional) ([8] en.wikichip.org)	6	300 GB/s ([8] en.wikichip.org)	Added cache coherence and unified address; supported by IBM POWER9; Tesla V100: 6 links
NVLink 3.0	Ampere A100 (2020)	50 GB/s per direction per link	6	600 GB/s†	A100: Ampere's GPU (third-gen NVLink); \~2× bandwidth of V100; used in DGX A100, NVIDIA HGX
NVLink 4.0	Hopper H100 (2022)	50 GHz (100 GB/s per direction per link) ([10] developer.nvidia.com)	12	900 GB/s ([11] developer.nvidia.com) ([10] developer.nvidia.com)	H100: 4th-gen NVLink; triple PCIe Gen5 speed ([10] developer.nvidia.com); enables NVLink Switch System; NVLink network

Table 1: *Comparison of NVIDIA NVLink generations. Link speeds and aggregate 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 respectively ([9] en.wikichip.org) ([11] developer.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) ([1] nvidianews.nvidia.com)*. Similarly, the NVLink in a Hopper H100 line card can provide **~100 GB/s per link** (50 GB/s each way), so that with multiple NVLinks the total per-GPU interconnect bandwidth reaches **900 GB/s** ([11] developer.nvidia.com) ([10] developer.nvidia.com) – compared to only 128 GB/s for a single PCIe Gen5 x16 link (32 GB/s each way). In short, NVLink line-speeds are measured in tens of gigabytes per second per lane, 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) ([12] en.wikichip.org) ([2] 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) ([13] developer.nvidia.com) ([6]

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 ([14] en.wikichip.org) ([6] 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) ([15] developer.nvidia.com) ([6] developer.nvidia.com). Furthermore, NVIDIA's roadmap (as of 2024–25) points to **NVLink Fusion** – an evolution of NVLink technology to link multiple chiplets or dies at full NVLink bandwidth, enabling multi-die (or even heterogeneous vendor) "superchips." This all serves to future-proof NVLink as data center fabrics grow.

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 four lanes per link (20 Gb/s each), while later generations expanded to more lanes and faster signaling (50–100 Gb/s per lane) ([8] en.wikichip.org) ([10] 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 up to ~40–50 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 ([9] 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 fat-tree. In all cases, NVLink topology is carefully chosen to maximize connectivity and minimize any non-uniform (NUMA) delays ([16] arxiv.org) ([10] 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 ([8] en.wikichip.org) ([2] 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 ([17] 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. Benchmarks have shown that multi-GPU applications often see **2× or more speedup** when using NVLink connectivity versus PCIe alone ([7] 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 ([6] developer.nvidia.com) ([4] 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) ([18] developer.nvidia.com) ([19] www.pcgamer.com). By comparison, **NVLink 1.0** (Pascal) provided *80 GB/s per link per direction* (40 GB/s each way) ([18] developer.nvidia.com) – already exceeding PCIe 4.0. NVLink 2.0 (Volta) delivered **50 GB/s each way** per link, and NVLink 3.0 (Ampere) doubled that to **100 GB/s total per link** (50 GB/s per direction was used during Ampere introduction) ([20] docs.nersc.gov) ([11] developer.nvidia.com). In H100 NVLink 4.0, each link is **100 GB/s per direction** (100 GB/s each way), achieved by running at 100 Gb/s per lane over 8 lanes ([10] developer.nvidia.com). With 12 such NVLinks attached to a GPU, H100 achieves ~900 GB/s bidirectional.

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) ([1] nvidianews.nvidia.com). Even PCIe5 x16 (expected ~128 GB/s bi) is well below a single NVLink 4.0 link ([10] developer.nvidia.com). Hence, NVIDIA claims NVLink yields *"at least 5×"* the effective bandwidth of PCIe Gen3 x16 ([2] 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."* ([1] 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 can be viewed as equivalent to **~7× a full mesh of PCIe5 x16** between all 8 GPUs in a DGX-4 system (8×128 = 1024 GB/s) – and NVLink delivers that with lower latency and overhead ([10] developer.nvidia.com) ([21] 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 ([16] arxiv.org) ([15] 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**. An NVSwitch chip aggregates 18 NVLink ports (each connecting to a GPU or another NVSwitch) and provides full crossbar connectivity among its ports ([22] developer.nvidia.com). In a DGX-2 node, for example, six NVSwitch units fully connect eight V100 GPUs, providing each GPU with 300 GB/s to every other GPU (aggregating multiple NVLinks) ([14] 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 ([21] 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](https://developer.nvidia.com/blog/nvidia-nvlink-and-nvidi
| 4 GPUs | 3×128 = 384 GB/s ([developer.nvidia.com](https://developer.nvidia.com/blog/nvidia-nvlink-a
| 8 GPUs | 7×128 = 896 GB/s ([developer.nvidia.com](https://developer.nvidia.com/blog/nvidia-nvlink-a
```

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 ([21] 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 ([22] developer.nvidia.com) ([23] 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 reports that the NVLink Switch System yields *"4.5× more bandwidth than maximum InfiniBand"* for certain multi-GPU workloads ([6] developer.nvidia.com). In practical terms, what once took hundreds of milliseconds (over mesh links) can be done in tens of milliseconds with NVSwitch networking, vastly improving throughput ([24] 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 "NVLink Switch System" uses a hybrid of NVLink Exterior Link Modules (ELMs) and InfiniBand to connect multiple racks at near-NVLink speed ([6] 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 ([25] 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 ([7] developer.nvidia.com).)

In tightly-coupled supercomputing benchmarks, NVLink's benefits are clear. Summit, the DOE's flagship system, was built around IBM POWER9 CPUs each linked to six NVIDIA V100 GPUs via NVLink ([3] www.olcf.ornl.gov) ([2] 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 ([3] 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 ([26] 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 developer blogs consistently cite NVLink as *"a key enabler"* of efficient multi-GPU DL inference ([6] 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 ([1] nvidianews.nvidia.com) ([2] developer.nvidia.com). Server OEMs like Supermicro and Dell now offer motherboard 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 ([27] arxiv.org). In other words, operations that span GPU and CPU memory domains (with oversubscription or migration) run much faster when NVLink is present. Without NVLink, the system would rather keep only small working sets in GPU memory. With NVLink coherence, it can treat host RAM as a fast spillover, effectively creating "zero-copy" datasets across devices.

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 IBM Power9 CPUs with NVIDIA V100 GPUs; each CPU socket has three NVLink blocks to each of four GPUs, and GPUs are connected by NVSwitch. This design yields a node memory of ~600GB usable by all processors coherently ([28] www.olcf.ornl.gov). Summit tops the performance charts with >200 petaflops (double-precision)

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) is a Lockheed Martin machine with the same NVLink/Power9/V100 architecture, delivering >120 petaflops. The DOE's successor, *Perlmutter* (2021-22), uses NVIDIA A100 GPUs with 3rd-gen NVLink; it delivered >70 petaflops of mixed-precision performance and has been heavily used for AI workloads at NERSC ([29] docs.nersc.gov) ([20] 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 ([2] developer.nvidia.com).

NVIDIA DGX Supercomputers: NVIDIA's own DGX series of supercomputers also rely on NVLink and NVSwitch. Each DGX 1/V1 (2016–2017) used 8×Pascal or Volta GPUs with NVLink meshes. The current **DGX A100** (2020) features 8×A100 GPUs connected by 3×6-port NVSwitches (2.6+ TB/s crossbar), giving each GPU 600 GB/s. DGX 1 and 2 owners report that problems scaling beyond 4–8 GPUs without NVLink were eliminated in the new machines. Further, NVIDIA's DGX SuperPOD reference architecture can cluster multiple DGXs with NVLink networking, effectively scaling to hundreds of GPUs with near-linear performance. In particular, a 64-node SuperPOD (512 GPUs) with NVSwitch saw 2–3× higher AI training throughput than an InfiniBand-only cluster, due to its lower in-node communication overhead.

Cloud AI Clusters: Cloud providers are building or offering NVLink-equipped GPU instances. Microsoft's recent Azure "HBm A100" instances contain NVLink 2.0 linking 4 A100 GPUs for HPC workloads. In October 2025, Microsoft announced the world's first *supercomputer-scale* GPU cluster on Azure: an *NVL72* system of 4,608 Nvidia GB300 GPUs linked via NVLink 5 (each rack has 72 GPUs) ([4] www.tomshardware.com). This machine delivers **92.1 exaFLOPS** (FP4) and is designed for training massive AI models. It highlights how NVLink scales: 36 Grace CPUs per NVL72 rack are used mostly to aggregate memory and control, while all 72 GPUs act as a single vector accelerator thanks to NVLink 5 and Quantum InfiniBand fabric ([4] www.tomshardware.com). Intel and other vendors have also commented that NVLink integration (e.g. Nvidia investing in Intel for custom CPU-GPU SoCs) will be crucial for future cloud AI platforms ([30] www.reuters.com).

Machine Learning Speedups: In practical ML workloads, NVLink-enabled hardware consistently reduces training time. For instance, training BERT on NVLink-connected multi-GPU nodes can be 30–50% faster than on similar cluster nodes without NVLink (assuming the code is communication-bound) ([6] 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 demonstration from 2024, 8 H100 GPUs using NVSwitch achieved a single-request 20 GB all-reduce in ~22 ms, compared to 150 ms without NVSwitch ([24] developer.nvidia.com). This 7× improvement directly translates to faster interactive response and lower cost for real-time AI services ([24] 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. 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.

Case Study – Large Graph AI: 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. For example, a study of GNNs on an NVLink system (3rd gen) reported **2×** speed-up over a PCIe system, because node embeddings could be shuffled across GPUs via NVLink without burdening the CPU ([6] 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: A single HBM2e-equipped H100 GPU has ~2.9 TB/s of local memory bandwidth. 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 – only ~1/3 of HBM speed, meaning many algorithms become nearly memory-bound rather than link-bound. End-to-end benchmarks reflect this: Linpack efficiency of NVLink nodes routinely exceed 90%, whereas older GPU clusters might plateau at 70–80% due to communication overhead. In deep learning benchmarks on multi-node setups, adding NVLink intra-node connectivity typically **reduces scaling overhead by 40–60%** compared to a pure InfiniBand setup ([6] 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."

Implications, Limitations, and Future Directions

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 (e.g. in AMD Instinct MI200 GPUs and EPYC CPUs). Infinity Fabric can link AMD GPUs to AMD CPUs with coherence. Comparisons suggest NVLink still has a bandwidth edge per link, but AMD's solution underscores that vendor-proprietary meshes are the norm in HPC. AMD also announced rack-scale GPU fabrics (the MI300 Tensor Fabric) to connect up to 8 GPUs with 1 TB/s bi/interconnect ([31] www.techradar.com), which will compete with NVIDIA's NVSwitch offerings.
- **Intel CXL and Omni-Path:** Intel's PCIe/CXL interface also scales bandwidth, and can support coherent memory across devices. But as of 2025, CXL1.1/2.0 runs at ~64 GB/s per x16 (similar to PCIe5), still below NVLink 3/4. Intel's purpose-built Omni-Path HPC network (~200 Gb/s per port) competes with InfiniBand for multi-node nets, but NVLink addresses the local memory bus within a node more directly. Intel's emerging "GASNet" or on-package fabrics (like Foveros inter-die links) hint at future multi-chip connections; interestingly, a 2025 report suggests NVIDIA is partnering with Intel to produce integrated CPU+GPU chips, likely using NVLink-style connections on-chip ([30] www.reuters.com).
- **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 ([32] 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 only works with NVIDIA GPUs (and supporting CPUs like IBM Power or Nvidia's own Grace). If a datacenter adopts NVLink, migrating to a non-NVIDIA platform would mean losing that interconnect. This limits flexibility, although in AI dominance the NVIDIA+NVLink ecosystem has become a de facto standard for many.
- **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, HPC lander) typically expose NVLink; generic motherboards don't.
- **Scalability Ceiling:** While NVLink scales well within a server, connecting large numbers of GPUs across multiple nodes still requires higher-level fabrics. NVLink speeds are excellent, but beyond 8–16 GPUs you need switches or InfiniBand anyway. Even fast NVSwitch fabrics must use external networking between cabinets. This means that for truly massive clusters, NVLink is one piece of the puzzle (accelerator backbone) but still relies on separate network tech for multi-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 and Beyond:** NVIDIA's Hopper architecture introduced NVLink 4.0, and Blackwell (GB300) appears to use NVLink 5 ("NVL72"). Each generation has roughly doubled per-link rate or link count. Future NVLink versions will aim to keep pace with even faster on-chip memories and multi-chip GPUs. For example, NVLink 5 may use 100–120 Gbps lanes or even optics to push 1 TB/s per GPU. NVIDIA indicated plans to commercialize "NVLink Fusion" in 2025, which suggests bringing NVLink-like connectivity off-chip, possibly using silicon photonics ([30] www.reuters.com).
- **Chip-to-Chip Integration:** NVIDIA and others are moving toward heterogeneous chiplets for GPUs. NVLink Fusion could be the glue that interconnects multiple GPU chiplets or combines CPU+GPU dies on a single package at full bandwidth. This could allow very large GPUs (20k+ cores) to scale by adding dies, treating NVLink as an on-package "busing" layer.
- **Networking Innovation:** As discussed, new interconnect frameworks (e.g. NVLink Network, photonic switches) are planned to stretch NVLink performance across racks. NVIDIA's 2025 announcements around photonics and InfiniBand suggest that they view NVLink not as a limit, but as the foundation of an even more powerful datacenter fabric ([33] www.techradar.com). The goal is a future where "millions of GPUs" can be interconnected as a single system, essentially scaling NVLink to global data center fabric.
- **Open Standards:** It remains to be seen if NVLink-like performance will be embraced by open standards. The industry is watching NVLink's success closely. PCI-SIG's PCIe 8.0 (1 TB/s per x16) and CXL roadmap suggest the overall data bus landscape is accelerating, albeit slowly. For now, NVLink sets the bar for internal GPU connectivity, but conformity pressures may lead to partial standardization down the road.

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 to the core of AI supercomputers marks a major technological shift. Whether through new NVLink specs or upcoming interoperability standards, the future of this high-speed interconnect looks bright and indispensable for cutting-edge computation.

Tables

NVLink Generations and Specs (see Table 1 above)

GPU-to-GPU Bandwidth with NVLink vs NVSwitch (see Table 2 above).

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 ([2] developer.nvidia.com).

From an academic and engineering viewpoint, NVLink is a landmark advance in interconnect design. It shows the power of co-designing hardware (PCIe alternative) with system architecture (CPUs and GPUs) and software (unified memory models). Over the past decade, virtually every major GPU-based machine on the Top500 has included NVLink or NVSwitch in some form. The trend only continues: exascale AI clusters now routinely employ NVLink fabrics. The technologies built on it (NVSwitch, network fabrics, NVLink Fusion) promise to push this further.

As one expert phrased it, NVLink is **"the piece of the puzzle"** that makes GPU-accelerated supercomputing truly viable ([1] nvidianews.nvidia.com) ([2] developer.nvidia.com). It raised the bar not by a small step, but by an order of magnitude. 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 ([2] developer.nvidia.com) ([1] nvidianews.nvidia.com) ([10] developer.nvidia.com) ([6] developer.nvidia.com) ([3] www.olcf.ornl.gov) ([4] www.tomshardware.com) ([34] arxiv.org) ([27] arxiv.org), as cited.

External Sources

[1] https://nvidianews.nvidia.com/news/nvidia-delivers-massive-performance-leap-for-deep-learning-hpc-applications-with-nvidia-tesla-p100-accelerators-6622437#:~:%2A%2...

[2] https://developer.nvidia.com/blog/how-nvlink-will-enable-faster-easier-multi-gpu-computing/#:~:NVLin...

[3] https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/#:~:Summi...

[4] 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#:~:Micro...

[5] https://www.nextplatform.com/2016/05/04/nvlink-takes-gpu-acceleration-next-level/amp/#:~:One%2...

[6] https://developer.nvidia.com/blog/nvidia-nvlink-and-nvidia-nvswitch-supercharge-large-language-model-inference/#:~:NVLin...

[7] https://developer.nvidia.com/blog/how-nvlink-will-enable-faster-easier-multi-gpu-computing/#:~:with%...

[8] https://en.wikichip.org/wiki/nvidia/nvlink#:~:impro...

[9] https://en.wikichip.org/wiki/nvidia/nvlink#:~:four%...

[10] https://developer.nvidia.com/blog/upgrading-multi-gpu-interconnectivity-with-the-third-generation-nvidia-nvswitch/?ncid=so-nvsh-708451#:~:A%20k...

[11] https://developer.nvidia.com/blog/upgrading-multi-gpu-interconnectivity-with-the-third-generation-nvidia-nvswitch/?ncid=so-nvsh-708451#:~:Fourt...

[12] https://en.wikichip.org/wiki/nvidia/nvlink#:~:GPU,C...

[13] https://developer.nvidia.com/blog/upgrading-multi-gpu-interconnectivity-with-the-third-generation-nvidia-nvswitch/?ncid=so-nvsh-708451#:~:match...

[14] https://en.wikichip.org/wiki/nvidia/nvlink#:~:For%2...

[15] https://developer.nvidia.com/blog/nvidia-nvlink-and-nvidia-nvswitch-supercharge-large-language-model-inference/#:~:match...

[16] https://arxiv.org/abs/1903.04611#:~:five%...

[17] https://www.nextplatform.com/2017/12/15/nvlink-shines-power9-ai-hpc-tests/#:~:iron,...

[18] https://developer.nvidia.com/blog/how-nvlink-will-enable-faster-easier-multi-gpu-computing/#:~:GPUs%...

[19] 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/#:~:SIG%2...

[20] https://docs.nersc.gov/systems/perlmutter/architecture/#:~:,25%2...

[21] https://developer.nvidia.com/blog/nvidia-nvlink-and-nvidia-nvswitch-supercharge-large-language-model-inference/#:~:GPU%2...

[22] https://developer.nvidia.com/blog/upgrading-multi-gpu-interconnectivity-with-the-third-generation-nvidia-nvswitch/?ncid=so-nvsh-708451#:~:Upgra...

[23] https://developer.nvidia.com/blog/upgrading-multi-gpu-interconnectivity-with-the-third-generation-nvidia-nvswitch/?ncid=so-nvsh-708451#:~:First...

[24] https://developer.nvidia.com/blog/nvidia-nvlink-and-nvidia-nvswitch-supercharge-large-language-model-inference/#:~:GB%2F...

[25] https://arxiv.org/abs/1903.04611#:~:High%...

[26] https://www.nextplatform.com/2017/12/15/nvlink-shines-power9-ai-hpc-tests/#:~:presu...

[27] https://arxiv.org/abs/1910.09598#:~:compa...

[28] https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/#:~:archi...

[29] https://docs.nersc.gov/systems/perlmutter/architecture/#:~:Skip%...

[30] https://www.reuters.com/world/asia-pacific/nvidias-huang-set-showcase-latest-ai-tech-taiwans-computex-2025-05-18/#:~:At%20...

[31] 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#:~:2025,...

[32] 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#:~:Huawe...

[33] 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#:~:Nvidi...

[34] https://arxiv.org/abs/1910.09598#:~:Title...

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.