SXM vs PCIe GPUs: Form Factor, Cooling and Use Cases

Published July 21, 2026 39 min read

Executive Summary

SXM (Server PCI Express Module) and **PCIe** (Peripheral Component Interconnect Express) are the two form factors NVIDIA uses to package its data center graphics processing units (GPUs), and the choice between them shapes bandwidth, power, cooling, and total cost of ownership more than the underlying silicon does. NVIDIA sells the same **H100** die in both packages: the SXM version ships on a proprietary mezzanine module that plugs into an **HGX** baseboard and draws up to 700 watts (W) of thermal design power (TDP), while the PCIe version is a standard dual-slot, air-cooled card (Source: nvidia.com). The interconnect gap is the decisive variable: SXM GPUs communicate over **NVLink** at up to 900 gigabytes per second (GB/s) per GPU on Hopper-generation hardware, versus the roughly 128GB/s ceiling of a PCIe Gen5 x16 link that PCI-SIG's own specification confirms provides "an effective raw bandwidth of 128.0 Gigabytes/second in each direction simultaneously" (Source: nvidia.com) (Source: pcisig.com). That bandwidth advantage is why every major hyperscale cloud, including Amazon Web Services (AWS) EC2 P5, Google Cloud A3, and Microsoft Azure ND H100 v5, provisions its flagship multi-GPU training instances exclusively on SXM hardware wired through NVIDIA's NVSwitch fabric (Source: cloud.google.com) (Source: learn.microsoft.com).

Pricing data gathered directly from provider pages as of July 2026 confirms a consistent, if modest, premium for SXM within the same provider: Lambda charges $3.29 per GPU-hour for **H100 PCIe** versus $4.29 for **H100 SXM**, and CoreWeave charges $4.25 versus $4.76 for the equivalent pair (Source: lambda.ai) (Source: coreweave.com). AWS prices its 8-GPU, SXM-only p5.48xlarge Capacity Block at an effective $5.191 per H100 per hour, while Google Cloud's a3-highgpu-8g instance works out to roughly $11.06 per H100 per hour on demand (Source: aws.amazon.com) (Source: cloud.google.com). On the power and cooling side, Uptime Institute's most recent global survey found that average server rack densities remain below 6 kilowatts (kW), with most operators reporting no racks beyond 20kW, even as 700W-class SXM accelerators push individual 8-GPU nodes well past that envelope (Source: uptimeinstitute.com).

The practical answer to "SXM versus PCIe GPU" depends on whether a workload is bound by inter-GPU communication. Large-scale distributed training of large language models (LLMs), the kind Meta ran across 24,576 **NVIDIA Tensor Core H100 GPUs** per cluster to train Llama 3, and the kind NVIDIA used to set MLPerf training records on 3,584 H100 GPUs with CoreWeave, requires SXM's NVLink and NVSwitch fabric because gradient synchronization across GPUs becomes the bottleneck at scale (Source: developer.nvidia.com). Conversely, single-GPU inference, fine-tuning

workloads that fit on one or two accelerators, and cost-sensitive deployments favor PCIe cards, which cost less to buy, fit standard air-cooled chassis, and do not require a specialized HGX baseboard. This report walks through the architecture, cooling requirements, benchmark evidence, cloud pricing, and five named deployments that illustrate when each form factor makes financial and technical sense, closing with a decision framework organizations can apply directly to procurement.

Five real-world cases anchor the analysis: Meta's 24,576-GPU Grand Teton clusters, NVIDIA's own Eos supercomputer built from 4,608 SXM-based H100 GPUs, the Department of Energy's NERSC Perlmutter system (which shows SXM and PCIe coexisting inside a single node), AWS's single-GPU p5.4xlarge instance used by insurer AON for actuarial modeling, and CoreWeave's record-setting MLPerf training submissions on both HGX H100 and the newer GB300 NVL72 (Source: docs.nersc.gov) (Source: coreweave.com). Taken together, the evidence points to a durable rule of thumb rather than a universal winner: SXM is structurally necessary once a workload must synchronize state across more than roughly two GPUs, and PCIe remains the more economical, operationally simpler choice everywhere else.

Introduction and Background

Every NVIDIA data center **GPU** (graphics processing unit) released since the **Volta** generation has shipped in two distinct physical packages that plug into a server in fundamentally different ways. **PCIe** is a decades-old, universal expansion standard that PCI-SIG describes as having "served as the de facto interconnect of choice for nearly two decades," and the same slot design that carries a network card, a storage controller, or a gaming graphics card also accepts a data center GPU shaped as a dual-slot card (Source: pcisig.com). **SXM**, by contrast, is NVIDIA's own high-performance mezzanine socket, a bare circuit module without an enclosure that bolts face-down onto a specialized system board called **HGX**, and NVIDIA's HGX platform page describes it as bringing "together the full power of NVIDIA GPUs" with NVLink and networking into a single certified design that server manufacturers build around (Source: nvidia.com). NVIDIA's own product literature does not publicly spell out what the "SXM" initialism stands for; trade publications that cover the hardware most closely describe it as **Server PCI Express Module**, and this report adopts that convention while noting the ambiguity (Source: amcompute.com).

The distinction matters because AI training and high-performance computing (HPC) workloads increasingly split a single job across many GPUs simultaneously, and how fast those GPUs can exchange intermediate results, gradients, or activations directly determines whether the cluster spends its time computing or waiting on data transfer. NVIDIA's H100, announced in 2022 and still widely deployed as of July 2026, is available in both packages built from identical silicon: the same 80 billion-transistor Hopper die, manufactured on the same Taiwan Semiconductor Manufacturing Company (TSMC) 4N process, is simply mounted differently depending on the target form factor (Source: megware.com). The same pattern holds for the older **A100** (Ampere generation), available in both SXM4 and PCIe form factors, and the newer **H200**, which pairs Hopper compute with larger high-bandwidth memory (HBM) (Source: megware.com). It is worth flagging one small but genuine discrepancy in the public record: NVIDIA's original 2022 H100 datasheet listed PCIe TDP as "300-350W (configurable)," while NVIDIA's current live product page lists the same card at "350-400W (configurable)," a revision this report attributes to a later specification update rather than an error, since the SXM figure of "Up to 700W (configurable)" has remained consistent across both versions (Source: megware.com) (Source: nvidia.com).

This report answers the question **"SXM vs PCIe GPU"** comprehensively: what each form factor is, how their interconnects differ, what cooling and power infrastructure each requires, how they perform on published benchmarks, what they cost to rent from major clouds, and which real-world deployments illustrate the tradeoffs.It also addresses the related queries practitioners search for, including what an SXM GPU is, how SXM and PCIe compare specifically for the H100 and A100, SXM cooling requirements, NVLink behavior across the two form factors, and concrete guidance for when to choose one over the other. All figures are anchored to publicly available vendor documentation, cloud pricing pages, standards-body specifications, and named deployment case studies gathered and verified as of July 2026; where sources disagree or data has changed across product generations, this report flags the discrepancy rather than presenting a false consensus.

SXM GPUs: Architecture and Interconnect

Capabilities

An SXM GPU is a bare module, exposing its die and memory stack for direct contact cooling, that mounts onto NVIDIA's proprietary HGX baseboard rather than plugging into a conventional PCIe slot (Source: amcompute.com). The HGX board wires every GPU position to a dedicated switching chip called **NVSwitch**, which gives each GPU a direct, single-hop path to every other GPU on the same board rather than routing traffic through the host central processing unit (CPU); AWS describes this design as letting "each GPU communicate with every other GPU in the same instance with single-hop latency" (Source: aws.amazon.com). On the H100 generation, this NVLink 4.0 fabric delivers 900GB/s of bidirectional bandwidth per GPU, and NVIDIA's datasheet describes the design as delivering "over 7X the bandwidth of PCIe Gen5" for GPU-to-GPU traffic, while also stating the NVLink Switch System "delivers 9X higher bandwidth than InfiniBand HDR on the NVIDIA Ampere architecture" for scale-out clusters of up to 256 H100 GPUs

(Source: megware.com). Beyond the NVLink fabric itself, NVIDIA describes its broader HGX networking stack as "combining NVIDIA NVLink scale-up, NVIDIA Quantum InfiniBand and Spectrum-X Ethernet scale-out, Spectrum-XGS Ethernet multi-data-center scale-across" into a single full-stack fabric purpose-built for GPU clusters, a layered design PCIe-only servers do not replicate (Source: nvidia.com). Because the socket is not power-constrained by a shared motherboard bus, SXM modules also support materially higher TDP than their PCIe siblings, and that extra power headroom is what lets NVIDIA clock SXM parts higher and pair them with faster memory: H100 SXM ships with 80 gigabytes (GB) of HBM3 at 3.35 terabytes per second (TB/s) of bandwidth, versus the original H100 PCIe's HBM2e running at roughly 2TB/s (Source: nvidia.com).

HGX boards hold four or eight GPUs and form the core of NVIDIA's own **DGX** systems as well as third-party servers certified under the NVIDIA-Certified Systems program. NVLink bandwidth has scaled with every architecture generation since: fourth-generation NVLink on Hopper delivers 900GB/s per GPU, fifth-generation NVLink on Blackwell (used in the **B200** and the rack-scale **GB200 NVL72**) delivers 1,800GB/s, and NVIDIA's sixth-generation NVLink, introduced with the Rubin platform, targets 3,600GB/s of bandwidth per GPU, more than fourteen times a PCIe Gen6 link (Source: nvidia.com). NVIDIA's newest HGX Rubin NVL8 board pushes this further still, integrating "eight NVIDIA Rubin GPUs with sixth-generation high-speed NVLink interconnects, delivering up to 10x more token factory throughput versus HGX B200" (Source: nvidia.com).

SXM also supports NVIDIA's **Multi-Instance GPU (MIG)** partitioning technology, which lets administrators split a single physical GPU into as many as seven fully isolated instances, each with dedicated memory, cache, and compute cores; NVIDIA says this "gives administrators the ability to support every workload, from the smallest to the largest, with guaranteed quality of service (QoS)," and separately notes the feature can unlock "up to 7x more GPU resources on a single GPU" for teams that would otherwise need seven separate cards (Source: nvidia.com) (Source: nvidia.com). NVIDIA notes that without MIG, "different jobs running on the same GPU...compete for the same resources," while MIG guarantees quality of service for each tenant (Source: nvidia.com). MIG is available on both SXM and PCIe GPUs across "NVIDIA Rubin, NVIDIA Blackwell, and NVIDIA Hopper GPUs," so partitioning flexibility is not by itself a reason to prefer one form factor over the other (Source: nvidia.com). On the older A100 generation, NVIDIA describes the same capability more simply: "An A100 GPU can be partitioned into as many as seven GPU instances, fully isolated at the hardware level with their own high-bandwidth memory, cache, and compute cores," a feature available identically on A100 SXM and A100 PCIe cards (Source: nvidia.com).

Adoption

SXM is the default configuration for every hyperscale cloud's flagship AI training instance as of mid-2026. AWS's **EC2 P5** family provides up to eight H100 GPUs per instance with up to 900GB/s of NVSwitch interconnect, for a total of 3.6TB/s of bisectional bandwidth across the instance, and up to 3,200 gigabits per second (Gbps) of cross-node networking through its second-generation Elastic Fabric Adapter (EFA) (Source: aws.amazon.com). Google Cloud's **A3** supercomputer instances pair eight H100 SXM GPUs with 3.6TB/s of bisectional bandwidth through NVSwitch and NVLink 4.0, plus custom 200Gbps networking chips that Google says provide "up to 10x more network bandwidth compared to our A2 VMs," its prior A100 generation (Source: cloud.google.com). Microsoft's **Azure ND H100 v5** series gives each of its eight GPUs a dedicated, topology-agnostic 400Gbps NVIDIA Quantum-2 InfiniBand connection alongside NVLink 4.0 for intra-VM GPU communication, and pairs that with "96 physical fourth Gen Intel Xeon Scalable processor cores" per virtual machine so the host CPU does not become a secondary bottleneck; Microsoft adds that these instances "provide excellent performance for many AI, ML, and analytics tools that support GPU acceleration" such as "TensorFlow, Pytorch, Caffe, RAPIDS, and other frameworks" (Source: learn.microsoft.com) (Source: learn.microsoft.com). Microsoft states that "ND H100 v5-based deployments can scale up to thousands of GPUs with 3.2 Tbps of interconnect bandwidth per VM" (Source: learn.microsoft.com). NVIDIA's own announcement of the Azure instance highlights that H100's "latest NVLink technology...lets GPUs talk to each other at 900GB/sec," and adds that "the inclusion of NVIDIA Quantum-2 CX7 InfiniBand with 3,200 Gbps cross-node bandwidth ensures seamless performance across the GPUs at massive scale, matching the capabilities of top-performing supercomputers globally," crediting the combination with helping the ND H100 v5 VMs "achieve up to 2x speedup in LLMs like the BLOOM 175B model for inference versus previous generation instances" (Source: blogs.nvidia.com) (Source: blogs.nvidia.com) (Source: blogs.nvidia.com).

Adoption is not limited to public cloud. Meta built its 24,576-GPU Llama 3 training clusters on its own **Grand Teton** SXM platform, an open GPU hardware design it contributed to the Open Compute Project, choosing the form factor specifically because it "integrate[s] power, control, compute, and fabric interfaces into a single chassis for better overall performance, signal integrity, and thermal performance" (Source: engineering.fb.com). NVIDIA's own DGX SuperPOD reference architecture, which underpins its Eos research supercomputer, is likewise built entirely from eight-GPU SXM DGX H100 nodes, each including "two NVIDIA BlueField-3 DPUs" for networking, storage, and security offload plus "Eight NVIDIA ConnectX-7" Quantum-2 InfiniBand networking adapters providing 400Gbps of throughput per adapter, connected by "a fourth-generation NVLink, combined with NVSwitch, [that] provides 900 gigabytes per second connectivity between every GPU in each DGX H100 system" (Source: nvidianews.nvidia.com) (Source: nvidianews.nvidia.com) (Source: nvidianews.nvidia.com).

Strengths and Limitations

SXM's central strength is bandwidth density: NVSwitch gives every GPU on a board a full-mesh, single-hop connection to every other GPU, which keeps multi-GPU training compute-bound rather than communication-bound as job sizes grow. Its second strength is thermal and power headroom, since the HGX baseboard delivers dedicated high-wattage power to each GPU position rather than sharing a general-purpose motherboard bus, making it practical to sustain 700W-class parts with a cold plate mounted directly on the exposed die (Source: amcompute.com).

The limitations are structural. SXM modules require a purpose-built HGX baseboard rather than a general-purpose motherboard, which most organizations buy pre-integrated inside an HGX-certified or DGX chassis rather than assembling themselves. On the newest architectures, that requirement is now absolute for the flagship part: NVIDIA's HGX platform page states plainly that "HGX is available in a single baseboard with eight NVIDIA Rubin, NVIDIA Blackwell, or NVIDIA Blackwell Ultra SXMs," with no PCIe alternative offered for those generations (Source: nvidia.com). That baseboard, plus the higher per-GPU power draw, generally pushes SXM systems toward liquid cooling once density climbs, which raises facility requirements relative to a conventional air-cooled rack, a tradeoff examined in more detail in the Data Analysis section below. SXM configurations are also typically sold and priced as fixed four- or eight-GPU units rather than the single-card increments PCIe supports, which reduces flexibility for organizations that only need one or two accelerators.

PCIe GPUs: Architecture and Deployment Model

Capabilities

A PCIe data center GPU is a standard dual-slot, air-cooled card that slots into the same x16 expansion connector used by consumer graphics cards, network interface cards, and storage controllers, requiring no proprietary baseboard (Source: nvidia.com). The PCI Special Interest Group (PCI-SIG), which governs the standard, has doubled per-lane bandwidth with each generation: PCIe 1.0 provided 2.5 gigabits per second per lane per direction, PCIe 2.0 raised that to 5.0Gb/s, PCIe 3.0 to 8.0Gb/s, and "PCIe 4.0 added support for 16.0 Gigabits/second/Lane/direction of raw bandwidth," with PCIe 5.0 reaching 32.0 gigatransfers per second (GT/s), which across a 16-lane link provides "an effective raw bandwidth of 128.0 Gigabytes/second in each direction simultaneously" (Source: pcisig.com) (Source: pcisig.com). The newer PCIe 6.0 specification doubles bandwidth again, to "64.0 GT/s raw data rate and up to 256.0 GB/s via x16 configuration" using Pulse Amplitude Modulation with four levels (PAM4) signaling, with "Lightweight Forward Error Correct (FEC) and Cyclic Redundancy Check (CRC)" added specifically to "mitigate the bit error rate increase associated with PAM4 signaling," and PCI-SIG notes the new specification "maintains backwards compatibility with all previous generations of PCIe technology," so data center operators can adopt PCIe 6.0 GPUs without replacing existing PCIe 5.0 or 4.0 infrastructure wholesale, though as of July 2026 most deployed H100 and A100 PCIe cards still use Gen5 or Gen4 connections (Source: pcisig.com) (Source: pcisig.com) (Source: pcisig.com). That ceiling, which applies to both a GPU's connection back to the host CPU and, absent a bridge, between two PCIe GPUs in the same server, is the fundamental constraint distinguishing PCIe from SXM's NVLink fabric.

Because PCIe GPUs share a conventional motherboard's power delivery, their TDP is capped well below SXM's. To recover some multi-GPU bandwidth without a full HGX baseboard, NVIDIA sells **NVLink Bridge** kits that connect up to two PCIe GPUs directly: on the A100 generation, NVIDIA states that "NVLink is available in A100 SXM GPUs via HGX A100 server boards and in PCIe GPUs via an NVLink Bridge for up to 2 GPUs," delivering 600GB/s between the pair (Source: pny.com). A comparable bridged configuration exists for the specialized H100 NVL card, well short of the eight-way, full-mesh 900GB/s fabric that NVSwitch provides but far faster than routing through the PCIe bus and host CPU. Notably, NVIDIA has not released an NVL or bridged-PCIe variant of its Blackwell-generation GPUs, consistent with Blackwell shipping only as SXM modules as described in the SXM Limitations section above, so the bridged, two-GPU PCIe path currently remains specific to Hopper-generation hardware.

Adoption

PCIe GPUs remain the default choice for single- and dual-GPU deployments across independent GPU clouds and on-premises servers. Lambda sells H100 PCIe on-demand at $3.29 per GPU-hour, alongside A100 PCIe at $1.99 per GPU-hour in its smaller instance tiers, while also offering multi-GPU 1-Click Clusters that "spin up 1-Click Clusters featuring 16-2,000+ interconnected HGX B200 and H100 GPUs for large-scale AI workloads" for customers that outgrow single cards (Source: lambda.ai) (Source: lambda.ai). CoreWeave lists NVIDIA H100 PCIe at $4.25 per hour separately from its NVLink-connected A100 and HGX H100 offerings on its classic pricing page, and interestingly prices A100 80GB PCIe and A100 80GB NVLink identically at $2.21 per hour, showing that pricing premiums for SXM are not universal across every GPU generation or provider; the same catalog also itemizes vCPUs and memory separately from the GPU line item, billed by the hour per core and per gigabyte, a granular model independent GPU clouds commonly use instead of the bundled instance pricing hyperscalers favor (Source: coreweave.com). PCIe form factors are also favored in HPC clusters that pair a modest number of GPUs with a CPU-heavy node design: the National Energy Research Scientific Computing Center's (NERSC)

Perlmutter supercomputer, for example, equips each GPU node with four NVIDIA A100 GPUs connected to the host over a "PCIe 4.0 GPU-CPU connection," even though the GPUs themselves communicate with each other over "4 third generation NVLink connections between each pair of GPUs," illustrating that PCIe and NVLink are not mutually exclusive within a single node design (Source: docs.nersc.gov) (Source: docs.nersc.gov).

Cloud providers also use PCIe-class single-GPU instances for workloads that do not need multi-GPU scaling, a pattern illustrated by AWS's smallest P5 configuration and detailed in the AWS EC2 P5 and AON case study below.

Strengths and Limitations

PCIe's core strength is compatibility and cost. Because it uses the same slot as every other server expansion card, a PCIe GPU can go into an existing, general-purpose chassis without a proprietary baseboard, keeping capital costs and supply chain flexibility higher than an SXM deployment requires (Source: amcompute.com). Lower TDP per GPU also means PCIe systems more often stay within the reach of conventional air cooling, avoiding the facility upgrades that high-density liquid-cooled racks require.

The limitation is the same bandwidth ceiling described above: once a workload must shard across more than two GPUs, PCIe's approximately 128GB/s Gen5 link (or a single 600GB/s NVLink Bridge pair) cannot match an eight-way NVSwitch fabric, and independent MLPerf submissions show the practical effect. Chip maker Qualcomm's MLPerf Inference results noted that its CloudAI100 accelerators, "each limited to 75W TDP," beat "eight Nvidia H100 (PCIe)" on throughput per watt for image classification and object detection, while NVIDIA's SXM-based DGX-H100 system topped the same round's language-model benchmarks, underscoring that the PCIe H100 configuration used in that comparison was not NVIDIA's highest-throughput entry (Source: eetimes.com). PCIe deployments also generally require purchasing an NVLink Bridge separately, and only in pairs, to recover any of the multi-GPU bandwidth SXM provides natively across all eight positions on an HGX board (Source: pny.com).

Feature Comparison

Table 1 below consolidates the official NVIDIA specifications for the H100 and A100 generations across both form factors, drawing on NVIDIA's datasheets. Figures reflect the vendor's published configurable maximums as of July 2026 and should be read as ceilings rather than typical sustained values.

SPECIFICATION	H100 SXM	H100 PCIE	A100 SXM (80GB)	A100 PCIE (80GB)
Form factor	Proprietary mezzanine module on HGX baseboard	Standard dual-slot, air-cooled card	Proprietary mezzanine module on HGX baseboard	Standard PCIe card
Max TDP	Up to 700W (configurable)	350 to 400W (configurable)	400W	300W
GPU memory	80GB HBM3	80GB HBM2e	80GB HBM2e	80GB HBM2e
Memory bandwidth	3.35TB/s	~2TB/s	2,039GB/s	1,935GB/s
GPU-to-GPU interconnect	NVLink: 900GB/s	PCIe Gen5: 128GB/s (NVLink Bridge optional, 2 GPUs)	NVLink: 600GB/s (up to 16 GPUs via NVSwitch)	NVLink Bridge: 600GB/s (2 GPUs)
Server options	NVIDIA HGX H100 or DGX H100 with 4 or 8 GPUs	Partner and NVIDIA-Certified Systems, 1 to 8 GPUs	NVIDIA HGX A100 with 4, 8, or 16 GPUs	Partner systems, 1 to 8 GPUs

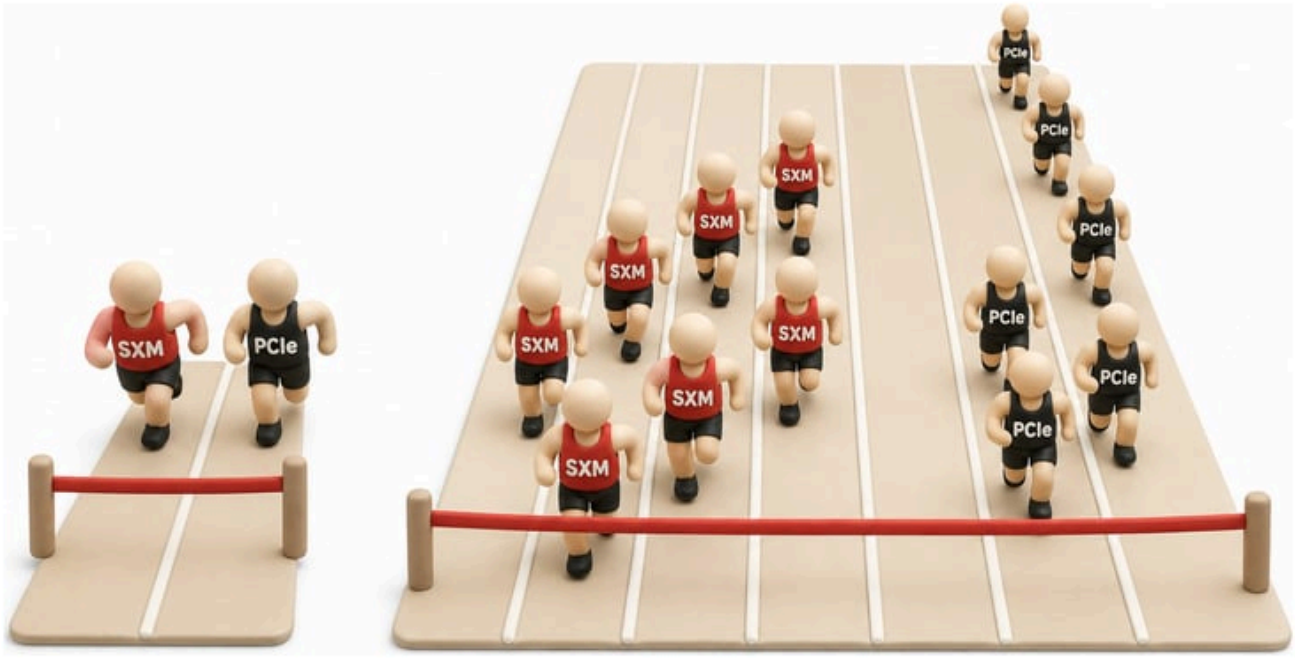
Sources: (Source: megware.com) (Source: nvidia.com) (Source: pny.com)

The table makes two points clear. First, the memory and compute silicon is close to identical within a generation, so the SXM premium is entirely a function of power delivery and interconnect, not a faster chip. Second, the interconnect gap compounds with GPU count: on an 8-GPU HGX H100 board, every GPU reaches every other GPU at the full 900GB/s NVLink rate simultaneously, whereas eight PCIe cards without bridges are limited to routing traffic through the host CPU's PCIe fabric, and even with bridges, NVLink Bridge pairs only connect two cards at a time rather than forming a full mesh. It is also worth noting that NVIDIA's rival AMD ships an analogous but non-interchangeable mezzanine format for its Instinct accelerators:

the MI300X uses the Open Compute Project's **OCP Accelerator Module (OAM)** format, itself connected to the host over a "PCIe 5.0 x16" bus but relying on AMD's own Infinity Fabric links, eight of them, rated at up to "128 GB/s" of peak per-link bandwidth for GPU-to-GPU traffic, on a card whose own performance footnotes describe "192 GB HBM3 memory capacity and 5.325 TFLOPS peak theoretical memory bandwidth" (Source: amd.com) (Source: amd.com) (Source: amd.com). This confirms the SXM-versus-PCIe tradeoff described here is an industry pattern rather than an NVIDIA-only design choice.

Performance and Benchmarks

Independent and vendor-reported benchmarks consistently show that the performance gap between SXM and PCIe widens as GPU count increases, while remaining small or negligible for single-GPU workloads. In NVIDIA's own MLPerf Training v3.0 submissions, the company's SXM-based H100 platform set new per-accelerator performance records across every workload category and delivered up to 3.1 times more performance per accelerator than the prior A100 generation's MLPerf Training v2.1 results (Source: developer.nvidia.com). At scale, NVIDIA and CoreWeave's joint submission trained a 175-billion-parameter GPT-3 model to convergence in just 10.9 minutes using 3,584 H100 SXM GPUs on CoreWeave's HGX H100 infrastructure, a result the companies describe as demonstrating "89% performance scaling efficiency even when moving from hundreds to thousands of H100 GPUs" (Source: developer.nvidia.com). NVIDIA's smaller-scale submissions reinforce the same scaling story: a 512-GPU H100 run trained the same GPT-3 benchmark in 64.3 minutes, and "scaling to 768 GPUs on the same Pre-Eos system reduced time-to-train to 44.8 minutes for near-linear scaling efficiency" (Source: developer.nvidia.com). That kind of scaling across thousands of GPUs is only achievable because NVLink and NVSwitch keep GPU-to-GPU communication off the comparatively slow PCIe bus.



Inference benchmarks tell a more nuanced story. MLPerf Inference results reported by EE Times show NVIDIA's SXM-based **DGX-H100** system "topping the table for Bert-Large results" in most scenarios, with one exception where a Dell PowerEdge server using eight H100s (with a different host CPU) edged ahead in a single offline-mode category, underscoring that host CPU and system integration, not just the GPU form factor, affect inference throughput (Source: eetimes.com). The same report found that scaling from one to eight GPUs within a DGX-H100 system stayed close to a perfect 8.0 times multiplier for most workloads, with only modest outliers in specific scenarios, again reflecting NVSwitch's near-linear scaling characteristics (Source: eetimes.com). At the same time, third-party inference accelerators demonstrated that PCIe-connected H100 systems are not automatically the most power-efficient option for every workload: Qualcomm's CloudAI100 chips outperformed eight PCIe H100 GPUs on queries per second per watt for image classification and object detection by a factor of 1.5 to 2.1 times, a reminder that raw interconnect bandwidth is only one input to inference efficiency (Source: eetimes.com). On the cloud side, Google reports its H100-based A3 instances deliver "up to a 30x inference performance boost when compared to our A2 VM's that are powered by NVIDIA A100 Tensor Core GPU," a comparison that mixes generational GPU

improvements with the interconnect upgrade from A2 to A3, and separately credits the platform's scale for "up to 26 exaFlops of AI performance," a figure that depends entirely on stitching thousands of NVSwitch-connected GPUs together rather than any single accelerator's raw throughput (Source: cloud.google.com) (Source: cloud.google.com).

Generational comparisons reinforce the pattern. NVIDIA's H200 datasheet, which specifies both an SXM and a PCIe-form-factor "NVL" variant, reports that H200 doubles inference performance relative to H100 on large models such as Llama 2 70B, and delivers up to 110 times faster time-to-results on select HPC applications when measured on the SXM configuration against an H100 SXM baseline, while stating that H200's added performance arrives "all within the same power profile as the H100 Tensor Core GPU" (Source: megware.com) (Source: megware.com). Because both H200 variants share the same 141GB of HBM3e memory and 4.8TB/s of memory bandwidth, and differ mainly in TDP (up to 700W for SXM versus up to 600W for the PCIe-based NVL card) and interconnect topology, the H200 comparison illustrates that even as memory capacity converges across form factors, interconnect remains the primary performance differentiator at scale.

Data Analysis and Evidence

Cloud pricing is one of the clearest quantitative signals of how the market values SXM's interconnect advantage. *Table 2* below compiles on-demand, per-GPU-hour pricing gathered directly from provider pricing pages in July 2026.

PROVIDER	INSTANCE OR PLAN	GPU	FORM FACTOR	PRICE PER GPU-HOUR
Lambda	1x H100 PCIe	H100	PCIe	$3.29
Lambda	1x H100 SXM	H100	SXM	$4.29
Lambda	1x A100 SXM (80GB)	A100	SXM	$2.79
Lambda	1x B200 SXM6	B200	SXM	$6.69
CoreWeave	H100 PCIe (Classic)	H100	PCIe	$4.25
CoreWeave	HGX H100 (Classic)	H100	SXM	$4.76
CoreWeave	A100 80GB PCIe (Classic)	A100	PCIe	$2.21
AWS	p5.48xlarge (8x H100, effective rate)	H100	SXM	$5.191
AWS	p4d.24xlarge (8x A100, effective rate)	A100	SXM	$1.475
AWS	p6-b200.48xlarge (8x B200, effective rate)	B200	SXM	$12.355
AWS	p6-b300.48xlarge (8x B300, effective rate)	B300	SXM	$14.04
AWS	u-p6e-gb200x72 UltraServer (72x B200, effective rate)	B200	SXM (NVL72)	$10.582
Google Cloud	a3-highgpu-8g (8x H100)	H100	SXM	~$11.06
Google Cloud	a2-ultragpu-8g (8x A100 80GB)	A100	SXM	~$5.07
Google Cloud	a4-highgpu-8g (8x B200)	B200	SXM	~$8.06

Sources: (Source: lambda.ai) (Source: coreweave.com) (Source: aws.amazon.com) (Source: cloud.google.com)

Table 2 shows a consistent pattern across the two providers that sell both form factors side by side: Lambda charges roughly 30 percent more for H100 SXM than H100 PCIe ($4.29 versus $3.29), and CoreWeave charges roughly 12 percent more ($4.76 versus $4.25), which is a modest premium relative to the sevenfold interconnect bandwidth difference the hardware provides. It also shows that the largest hyperscalers, AWS and

Google Cloud, do not offer PCIe-based multi-GPU instances at flagship tiers at all; their high-end training products are SXM-only, which is consistent with the architecture and adoption evidence discussed above. Google Cloud's A3 pricing, at roughly $11.06 per H100 per hour, sits notably above AWS and the independent clouds, a discrepancy this report flags rather than resolves, since provider pricing reflects networking, support, and regional cost structures beyond the GPU itself. The newest Blackwell-generation pricing shown in the table, from $6.69 per GPU-hour at Lambda for B200 up to an effective $14.04 per GPU-hour for AWS's newer p6-b300.48xlarge Capacity Block, is SXM-only across every provider surveyed, since NVIDIA has not shipped a PCIe or NVL variant of the B200 or B300 as of July 2026; notably, AWS prices rack-scale GB200 NVL72 UltraServer capacity at an effective $10.582 per GPU-hour across a 72-GPU domain, close to its standalone 8-GPU B200 instance rate, suggesting the NVL72 rack-scale premium is smaller than the jump from PCIe to SXM within a single generation (Source: aws.amazon.com). AWS's Capacity Block pricing model itself is worth noting as a structural difference from ordinary on-demand billing: it requires reserving GPU capacity for a fixed future time window, and AWS explains that "when you use an EC2 Capacity Block, you pay for the operating system you use when your instances are running," a mechanism the company introduced specifically to guarantee availability for large, time-boxed training runs on scarce SXM-based instances (Source: aws.amazon.com). Google Cloud's own accelerator pricing page similarly confirms its lowest-cost A100 tier, the single-GPU A2 Standard configuration, remains an SXM part rather than a PCIe one, underscoring how thoroughly SXM has displaced PCIe across every hyperscaler's catalog regardless of instance size (Source: cloud.google.com).

Power and cooling economics are the other major quantitative dimension. Uptime Institute's most recent global data center survey, its 13th annual edition, found that "server rack densities are climbing steadily, but slowly," with average densities remaining below 6kW per rack, and the survey separately notes that "most operators do not have any racks beyond 20 kW," a finding whose authors say "suggests the widespread use of direct liquid cooling is not imminent" across the industry as a whole (Source: uptimeinstitute.com) (Source: uptimeinstitute.com). That baseline is a meaningful shift from just a few years earlier: an Uptime Institute survey of the industry found "the mean average density in our 2020 survey sample was 8.4 kW/rack," with "the most common density" then sitting at "5-9 kW/rack," and separately reported that "racks with densities of 20 kW and higher are becoming a reality for many data centers," even though such high densities remained concentrated among a small number of hyperscale operators rather than the industry at large (Source: journal.uptimeinstitute.com) (Source: journal.uptimeinstitute.com) (Source: journal.uptimeinstitute.com). That baseline is important context for SXM deployments: an 8-GPU HGX H100 board alone can draw up to 5.6kW from the GPUs before accounting for host CPUs, memory, storage, and networking, meaning a fully populated SXM rack can exceed typical facility provisioning even before reaching the 20 to 25kW threshold that Uptime Institute's research separately identifies as the point where "direct liquid cooling and precision air cooling becomes more economical and efficient" (Source: journal.uptimeinstitute.com). At the extreme end of the roadmap, NVIDIA's GB200 NVL72 rack-scale system, which uses SXM-format Blackwell GPUs exclusively, is described by NVIDIA as "a rack-scale, liquid-cooled design" that concentrates 72 GPUs and 36 Grace CPUs in a single rack, and NVIDIA states that compared with air-cooled H100 infrastructure, "GB200 delivers 25x more performance at the same power, while reducing water consumption" through that liquid-cooling architecture (Source: nvidia.com) (Source: nvidia.com). By contrast, the H100 PCIe card remains a conventional air-cooled design, fitting within the density envelope most data centers already support without facility upgrades.

Interconnect standards data supplies the third quantitative pillar. PCI-SIG's own specification confirms the underlying PCIe ceiling that both form factors' host connections share, and describes PCIe 6.0 as "the cost-effective and scalable interconnect solution for data-intensive markets like Data Center, Artificial Intelligence/Machine Learning, HPC," language that positions even the fastest PCIe generation as complementary to, rather than a substitute for, dedicated GPU interconnects like NVLink (Source: pcisig.com). A 16-lane PCIe 5.0 link, the generation underlying most deployed H100 and A100 PCIe cards, tops out at 128.0GB/s bidirectional per the same PCI-SIG specification, doubling to as much as 256.0GB/s on PCIe 6.0 x16 configurations, but both figures remain well below the 900GB/s per-GPU NVLink bandwidth SXM systems achieve today (Source: pcisig.com). NVIDIA's own generational NVLink table, published alongside its GB200 NVL72 materials, confirms the sixth-generation NVLink figure of 3.6TB/s per GPU cited throughout this report, reinforcing that the interconnect gap versus PCIe widens rather than narrows with each new architecture (Source: nvidia.com).

Case Studies and Real-World Examples

Meta's Grand Teton Clusters: SXM at 24,576-GPU Scale

Meta disclosed in March 2024 that it had built two 24,576-GPU clusters on its in-house **Grand Teton** SXM platform specifically to train Llama 3, describing the combination of "the efficiency of the high-performance network fabrics within these clusters" and "the 24,576 NVIDIA Tensor Core H100 GPUs in each" as what allowed both cluster versions to support larger and more complex models than its prior AI Research SuperCluster, which had used "16,000 NVIDIA A100 GPUs, in 2022" (Source: engineering.fb.com) (Source: engineering.fb.com). Meta built one cluster on a Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) fabric and the other on NVIDIA Quantum2 InfiniBand specifically to compare scale-out interconnect options, and reported that "both of these solutions interconnect 400 Gbps endpoints," while both clusters shared the same SXM-based

Grand Teton compute platform for intra-node GPU communication (Source: engineering.fb.com). By the end of 2024, Meta said it planned to grow its total infrastructure to 350,000 NVIDIA H100 GPUs, part of a portfolio delivering compute power equivalent to nearly 600,000 H100s (Source: engineering.fb.com). The scale of that build-out is only tractable with SXM's NVSwitch fabric; routing that many GPUs' gradient synchronization traffic over PCIe alone would leave the CPUs and PCIe fabric as the bottleneck long before GPU compute became the limiting factor.

NVIDIA Eos and the DGX SuperPOD: SXM for National-Scale HPC

NVIDIA's own **Eos** supercomputer, announced alongside the DGX H100 platform, is built from 576 DGX H100 systems for a total of 4,608 H100 GPUs, all connected via SXM modules on HGX baseboards, NVLink Switch System, and Quantum-2 InfiniBand (Source: nvidianews.nvidia.com). NVIDIA said the resulting DGX SuperPOD architecture delivers a total of "70 terabytes/sec of bandwidth" across the pod, and framed the entire system as designed to "run massive LLM workloads with trillions of parameters" (Source: nvidianews.nvidia.com). This case illustrates SXM adoption at the vendor's own reference-architecture level: NVIDIA does not offer an equivalent PCIe-based SuperPOD design for large-scale distributed training, reflecting the same bandwidth logic that drove Meta's and the hyperscalers' choices.

NERSC Perlmutter: A Hybrid Deployment That Clarifies the Boundary

The Department of Energy's NERSC facility operates **Perlmutter**, a system based on the HPE Cray Shasta platform that is "a heterogeneous system comprised of 3,072 CPU-only and 1,792 GPU-accelerated nodes," each pairing a single AMD EPYC CPU with four NVIDIA A100 GPUs (Source: docs.nersc.gov). Within a node, the four A100 GPUs communicate with each other over four third-generation NVLink connections, each rated at "25 GB/s/direction for each link," a capability the PCIe Adoption section above shows works alongside, not instead of, the node's separate PCIe 4.0 host connection (Source: docs.nersc.gov). This is a valuable real-world clarification of a common misconception: SXM and PCIe are not mutually exclusive labels for an entire system, since a node can use SXM GPUs with NVLink for intra-node GPU-to-GPU traffic while still relying on PCIe for the separate GPU-to-CPU and GPU-to-network-interface-card paths. Perlmutter's design also demonstrates that four-GPU SXM nodes, rather than the eight-GPU configuration common in hyperscale cloud offerings, remain a common HPC procurement pattern.

AWS EC2 P5 and AON: When Single-GPU PCIe-Class Economics Win

Not every production deployment needs multi-GPU interconnect at all. AWS's own EC2 P5 marketing highlights insurer AON's use of the single-GPU p5.4xlarge instance, quoting the company's Global Head of Life Solutions describing how the instance let AON "run machine learning models and economic forecasts that used to take days in just a matter of hours," and noting that "the ability to use a single H100 GPU instance (p5.4xlarge) means we're not only saving time but also optimizing our computational resources" (Source: aws.amazon.com) (Source: aws.amazon.com). AWS's own instance specification confirms that p5.4xlarge does not support GPU peer-to-peer connectivity and states plainly that "GPUDirect RDMA is not supported in P5.4xlarge," the features that matter only once a workload spans multiple GPUs (Source: aws.amazon.com). This case underscores the report's central decision principle: for actuarial modeling and similar single-accelerator analytical workloads, the multi-GPU interconnect debate is simply irrelevant, and the cheapest single-GPU instance that fits the model in memory is the economically correct choice, regardless of whether the underlying chip happens to be packaged for a multi-GPU board.

CoreWeave and NVIDIA: SXM at the Frontier of MLPerf Records

CoreWeave's joint MLPerf Training v3.0 submission with NVIDIA used HGX H100 (SXM) infrastructure to train a 175-billion-parameter GPT-3 model at 3,584-GPU scale in 10.9 minutes, a result the companies say demonstrates that "the NVIDIA platform delivers great performance in both on-premises and commercially available cloud instances" (Source: developer.nvidia.com). More recently, in MLPerf Training v6.0, CoreWeave reported training the DeepSeek-V3 671-billion-parameter mixture-of-experts model in 2.02 minutes on 8,192 NVIDIA GB300 GPUs, calling it "the fastest time recorded in the round across all available-cloud submissions on this benchmark" and noting the run used "the largest GB300 NVL72 cluster ever benchmarked on this workload," a cluster CoreWeave says is "16x the next-largest GB300 submission in the round" (Source: coreweave.com) (Source: coreweave.com) (Source: coreweave.com). CoreWeave also reported a 2.8 times wall-clock speedup on its Llama 3.1 405B training benchmark year over year, crediting the combination of "the latest NVIDIA GB300 NVL72 systems with NVIDIA Spectrum-X networking, the topology-aware SUNK scheduler, and CoreWeave Mission Control" (Source: coreweave.com). Both flagship records depend on SXM-format GPUs (Hopper-generation H100 in 2023, Blackwell Ultra-generation GB300 in 2026) mounted on NVIDIA's rack-scale NVLink fabric, reinforcing that frontier-scale training performance records remain an SXM-exclusive domain even as the underlying chip generation changes.

Implications and Future Directions

The trajectory of NVIDIA's roadmap suggests the SXM-versus-PCIe divide will widen rather than narrow at the high end, even as it becomes less relevant for smaller deployments. NVIDIA's newest Blackwell and Rubin-generation flagship parts are available only as SXM modules integrated into liquid-cooled HGX or NVL72 designs, unlike the H100 and H200 generations that retained PCIe options. That shift pushes frontier-scale training and reasoning workloads further toward the rack-scale, liquid-cooled model that GB200 NVL72 pioneers, where NVIDIA reports a "130 terabytes per second (TB/s) of low-latency GPU communications" domain spanning all 72 GPUs in a single rack, a bandwidth figure with no PCIe analog at any generation (Source: nvidia.com). NVIDIA also offers a smaller-scale bridged option for organizations that need more than one GPU acting together without committing to a full rack: the "NVIDIA GB200 NVL4...deliver[s] revolutionary performance through a bridge connecting four NVIDIA NVLink Blackwell GPUs unified with two Grace CPUs," compatible with liquid-cooled modular servers rather than the full NVL72 rack (Source: nvidia.com). Organizations planning multi-year AI infrastructure investments should assume that any workload requiring more than roughly two GPUs of tightly coupled compute will need to budget for SXM-class power and liquid-cooling infrastructure, not just for the current generation but for the generations likely to follow it.

At the same time, PCIe is not disappearing from the inference and edge tiers. MIG partitioning works on both SXM and PCIe-mounted GPUs and, on newer NVIDIA architectures, an administrator "could create two instances with 93GB of memory each, four instances with 46GB each, or seven instances with 23GB each" from a single GB200-class GPU, letting organizations right-size one physical accelerator to many smaller inference workloads regardless of form factor (Source: nvidia.com). This reduces the practical urgency of choosing SXM purely for utilization reasons. NVIDIA has also continued releasing PCIe-native "NVL" variants for memory-bound inference use cases through the Hopper generation, and the H200 datasheet explicitly frames the PCIe-form-factor H200 NVL as "the ideal choice for customers with space constraints within the data center," targeting large language model serving where a single GPU lacks sufficient memory but an eight-GPU SXM system would be over-provisioned (Source: megware.com).

Cooling infrastructure investment is the practical constraint most likely to determine adoption speed. Uptime Institute's finding that most data centers still operate well below the 20 to 25kW threshold where liquid cooling becomes economically favorable implies that a large share of the installed base cannot host SXM-based, multi-GPU systems without capital investment in power distribution and cooling upgrades, even if the compute economics otherwise favor SXM (Source: uptimeinstitute.com). This is one reason independent GPU clouds like Lambda and CoreWeave continue to sell PCIe capacity alongside SXM rather than migrating entirely to liquid-cooled fleets, as reflected in both providers' published rate cards: a meaningful share of demand, particularly fine-tuning, batch inference, and research prototyping, does not need SXM's bandwidth and would rather avoid its facility requirements and price premium (Source: coreweave.com).

The following decision matrix, *Table 3*, synthesizes the evidence above into a practical procurement guide organized by workload type.

WORKLOAD TYPE	RECOMMENDED FORM FACTOR	RATIONALE
Single-GPU inference or prototyping	PCIe	No inter-GPU traffic to accelerate; lowest cost and simplest air-cooled deployment (Source: amcompute.com)
Fine-tuning on 1 to 2 GPUs	PCIe (with optional NVLink Bridge)	NVLink Bridge recovers most of the multi-GPU benefit for a two-GPU pair without a full HGX baseboard (Source: pny.com)
Distributed training, 4 to 8+ GPUs, single node	SXM	NVSwitch full-mesh interconnect keeps training compute-bound as GPU count grows (Source: nvidia.com)
Multi-node distributed training or foundation model pretraining	SXM plus InfiniBand or comparable scale-out fabric	Required to reach the thousands-of-GPU scale used by Meta, NVIDIA Eos, and CoreWeave's MLPerf records (Source: coreweave.com)
Memory-bound LLM inference serving on 2 GPUs	PCIe-based NVL variant	Purpose-built for space-constrained data centers per NVIDIA's own H200 NVL positioning (Source: megware.com)
Facility limited to air cooling, below 20kW per rack	PCIe	Matches the density envelope most data centers already support (Source: uptimeinstitute.com)

Frequently Asked Questions (FAQs)

What is an SXM GPU? An SXM GPU is a NVIDIA data center graphics processing unit packaged as a bare mezzanine module rather than an enclosed card, designed to mount onto a proprietary HGX baseboard that wires every GPU on the board to a dedicated NVSwitch chip for direct, high-bandwidth GPU-to-GPU communication over NVLink. It cannot be installed in a conventional server motherboard slot and instead ships pre-integrated inside an HGX-certified or DGX chassis.

How does SXM compare to PCIe for the H100 specifically? H100 SXM offers up to 700W of TDP, 3.35TB/s of memory bandwidth, and 900GB/s of NVLink interconnect per GPU, while H100 PCIe is capped at 350 to 400W, roughly 2TB/s of memory bandwidth, and approximately 128GB/s of PCIe Gen5 interconnect unless paired with an NVLink Bridge (Source: megware.com).

How does SXM compare to PCIe for the A100? A100 SXM (80GB) supports up to 400W of TDP and 2,039GB/s of memory bandwidth, with NVLink providing 600GB/s of interconnect across up to 16 GPUs via NVSwitch; A100 PCIe (80GB) is capped at 300W and 1,935GB/s of memory bandwidth, with NVLink Bridge kits limited to connecting two GPUs at 600GB/s (Source: nvidia.com) (Source: pny.com).

What are the cooling requirements for SXM GPUs? SXM's exposed die design allows a cold plate to sit directly on the chip, which is what makes it practical to sustain 700W-class TDP; NVIDIA's own reference designs increasingly pair high-wattage SXM parts, like the Blackwell-generation GB200 NVL72, with liquid cooling rather than air cooling, while H100 PCIe remains explicitly specified as an air-cooled, dual-slot card, as detailed in the Data Analysis section above.

Does PCIe GPU performance lag SXM in every workload? No. For single-GPU inference or fine-tuning that never requires GPU-to-GPU communication, PCIe and SXM versions of the same chip perform comparably, since they share the same compute cores and similar memory bandwidth; the gap only opens for workloads that must synchronize state across multiple GPUs, where SXM's NVLink and NVSwitch fabric provides substantially higher throughput than PCIe's shared bus (Source: eetimes.com).

Do all cloud providers offer both SXM and PCIe options? No. AWS, Google Cloud, and Microsoft Azure provision their flagship multi-GPU training instances exclusively on SXM hardware, while independent clouds such as Lambda and CoreWeave sell both PCIe and SXM configurations of the same GPU generation side by side (Source: cloud.google.com) (Source: cloud.google.com).

When should an organization choose PCIe over SXM? PCIe is the more cost-effective and operationally simpler choice for single- or dual-GPU inference, model fine-tuning that fits in the memory of one or two accelerators, batch analytical workloads like the actuarial modeling AON runs on AWS's single-GPU p5.4xlarge instance, and any deployment where the facility cannot support liquid cooling or 700W-class rack density.

Conclusion

SXM and PCIe are not competing GPU products but two different ways of packaging the same silicon, and the choice between them should be driven by a single technical question: does the workload need multiple GPUs to synchronize state with each other at high bandwidth, or does it not. When the answer is yes, as with large-scale LLM training at Meta's 24,576-GPU scale, NVIDIA's own DGX SuperPOD and Eos supercomputer, and CoreWeave's record-setting MLPerf submissions, SXM's NVLink and NVSwitch fabric is not merely faster but structurally necessary, since PCIe's roughly 128GB/s ceiling would leave thousands of GPUs waiting on data transfer rather than computing. When the answer is no, as with AON's single-GPU actuarial modeling on AWS or the many fine-tuning and inference workloads that independent clouds like Lambda and CoreWeave still price at a meaningful discount for PCIe, the lower cost, simpler air cooling, and standard chassis compatibility of PCIe make it the more rational purchase.

The data gathered for this report, spanning official NVIDIA specifications, cloud provider pricing pages, PCI-SIG's own interconnect standards (Source: pcisig.com), Uptime Institute's facility survey data, AMD's competing OAM format, and five named deployments, points to a durable rather than temporary divide. As NVIDIA's Blackwell and Rubin generations drop PCIe options for flagship parts entirely and push rack-scale liquid cooling further into the mainstream, the practical decision for most organizations will increasingly be framed not as SXM versus PCIe in the abstract, but as a question of how many GPUs a given workload genuinely needs to act as one, and whether the organization's facilities, budget, and workload size justify building or renting the infrastructure that answer requires. Procurement teams should treat this as a workload-sizing exercise first and a hardware-shopping exercise second: benchmark the actual multi-GPU communication pattern of the target workload before committing to either form factor, since the wrong choice in either direction, over-provisioning SXM for a single-GPU inference service or under-provisioning PCIe for a distributed training job, carries a real and measurable cost.

Tags: SXM, PCIe, GPU, NVIDIA H100, NVIDIA A100, NVLink, HGX, data center GPUs, cloud computing, AI infrastructure

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 GPUSmith 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.

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