IntuitionLabs

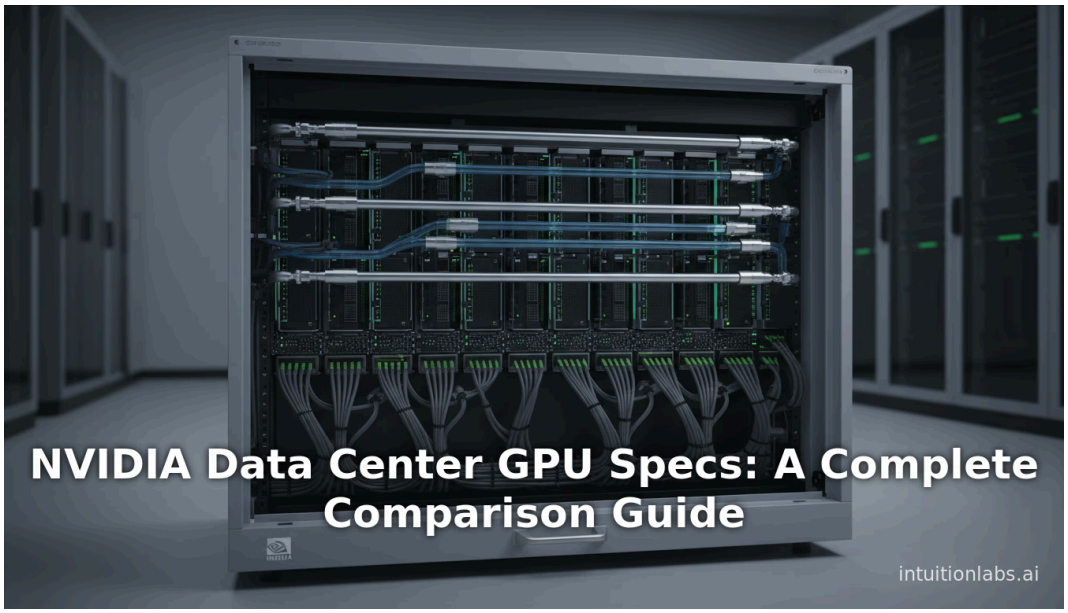
INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

NVIDIA Data Center GPU Specs: A Complete Comparison Guide

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

Published by IntuitionLabs · 2025-10-24 · nvidia gpu · data center gpu · gpu comparison · blackwell architecture · hopper architecture · a100 vs h100 · gb200 nvl72 · l40s · gpu specifications

Original article: https://intuitionlabs.ai/articles/nvidia-data-center-gpu-specs



In this article

1. NVIDIA Data Center GPU Specs: A Complete Comparison Guide
2. Executive Summary
3. Introduction and Background
4. NVIDIA Data-Center GPU Architecture Evolution
5. Detailed Characteristics of NVIDIA Data Center GPUs
6. Multi-GPU Interconnect: NVLink, NVSwitch, and NVL72
7. Multi-Instance GPU (MIG) and Virtualization
8. Memory Architecture and Efficiency
9. Case Studies and Deployments
10. Supercomputer and Cloud Adoptions
11. Cloud and Enterprise Technologies
12. Analysis of NVIDIA GPU Compute and Connectivity
13. Market and Ecosystem Context
14. Future Implications and Directions
15. References

NVIDIA Data Center GPU Specs: A Complete Comparison Guide

Executive Summary

Update (October 2026): NVIDIA's newest data center platform is Vera Rubin. Launched at CES in January 2026, the Rubin GPU carries 288 GB of HBM4 and is rated at 50 PFLOPS of NVFP4 inference, and the Vera Rubin NVL72 rack (72 Rubin GPUs, 36 Vera CPUs) is rated at 3,600 PFLOPS of NVFP4 inference (nvidianews.nvidia.com) (www.nvidia.com). In its August 26, 2026 results, NVIDIA said Vera Rubin was in full production, with racks running at partners including CoreWeave, Google Cloud, Microsoft Azure, Oracle Cloud Infrastructure and Nebius, and reported quarterly data center revenue of $89.0 billion, up 117% year over year (nvidianews.nvidia.com). Table 1 now includes Rubin, per-GPU B200 and B300 figures from NVIDIA's HGX specifications, and corrected Ampere, Hopper and Ada values.

NVIDIA’s data center GPU portfolio has rapidly evolved to address the surging compute demands of AI, HPC, and graphics workloads. This report provides a comprehensive technical overview and comparison of NVIDIA’s current data-center GPU “platform” solutions, including CPU+GPU superchips (Grace+GPU), traditional accelerator cards, and advanced interconnect architectures. We cover all major offerings, from Ampere (A100) and Hopper (H100/H200) accelerators to Ada Lovelace visualization GPUs (L40, L40S) and specialized variants (e.g. RTX Pro cards and China-specific chips like the B40). Key innovations such as multi-instance GPUs (MIG), NVLink/NVSwitch fabrics, and the new NVL72 72-GPU domain are examined. Detailed spec and performance comparisons are presented (see Table 1) along with analysis of system-level designs and case studies. For example, Microsoft’s Azure GB300 NVL72 supercluster (4,608 Blackwell Ultra GPUs) achieves

~92.1 exaFLOPS inference by tightly coupling 72 GPUs per rack with 1.8 TB/s links each (www.tomshardware.com) (developer.nvidia.com). Similarly, NVIDIA reports that its L40S Ada GPU delivers nearly 5× the FP32 throughput of the previous A100 (nvidianews.nvidia.com). We include data on memory capacity, bandwidth, FLOPS, TDP, and networking for each GPU, supported by NVIDIA's documentation and third-party measurements. Market and deployment insights (e.g. NVIDIA's ~80–92% AI accelerator market share as of 2025, with quarterly data center revenue of $51.2B in Q3 FY2026 (nvidianews.nvidia.com) and $89.0B in Q2 FY2027 (nvidianews.nvidia.com)) are interwoven. Special attention is given to future directions: the **Blackwell Ultra (B300/GB300)** platform shipping H2 2025, the **Vera Rubin** architecture (in full production for H2 2026) with HBM4 and 3.6 ExaFLOPS of NVFP4 inference per rack, and the April 2025 H20 licensing requirement for exports to China and NVIDIA's RTX Pro 6000D/B40 product (www.tomshardware.com), scaling to trillion-parameter models, and infrastructure changes (e.g. racks with 120kW liquid cooling (developer.nvidia.com)). The report's comparisons, tables, and case examples offer a deep technical reference on NVIDIA's full datacenter GPU lineup as of October 2026, with citations to official and expert sources for every claim.

Introduction and Background

Graphical processors have transcended their gaming origins to become the **workhorses of the AI and HPC era**. NVIDIA pioneered this shift by reorienting its GPU roadmap toward data-center applications (AI training/inference, HPC simulation, and professional graphics). The market response has been immense: NVIDIA's data center segment generated **$35.6 billion in Q4 FY2025** (ending January 2025) and **$51.2 billion in Q3 FY2026** (October 2025), up 93% and 66% year over year respectively (nvidianews.nvidia.com) (nvidianews.nvidia.com). The company's full fiscal year 2026 revenue reached **$215.9 billion** (+65% YoY) (nvidianews.nvidia.com), and data center revenue rose to **$89.0 billion in Q2 FY2027** (July 2026), up 117% year over year (nvidianews.nvidia.com). NVIDIA is estimated to consume **77% of all wafers used for AI processors in 2025** (www.tomshardware.com), reflecting an AI accelerator market share of roughly **80–92%** depending on the analyst methodology. This dominance reflects NVIDIA's engineering focus; each new GPU generation typically doubles or triples performance. For example, the Ampere-based A100 (2020) delivered on the order of 20 TFLOPS FP32 or 312 TFLOPS (dense FP16 Tensor) (developer.nvidia.com) (www.nvidia.com), far beyond the prior Volta V100, while the Hopper-based H100 (2022) pushed that to ~67 TFLOPS FP32 and 1,979 TFLOPS FP16 Tensor with sparsity (about 990 dense) (www.nvidia.com). These accelerators also introduced game-changing features like **MIG virtualization** (NVIDIA A100 can partition into up to 7 isolated GPU instances (developer.nvidia.com)) and new Tensor Core math formats (TF32 and BF16 on A100's **third-generation Tensor Cores**, FP8 on H100's fourth generation).

In parallel, NVIDIA has expanded beyond simple accelerator cards. Notable "Platform" initiatives include the **Grace CPU** (ARM-based) and **GH200/GB200/GB300 superchips**, which integrate NVIDIA GPUs with custom CPUs in one package. For instance, the GB300 "Grace Blackwell Ultra" desktop superchip combines a 72-core Grace CPU with a Blackwell Ultra GPU and up to 784 GB of unified memory in early partner specifications, delivering ~20 PFLOPS AI compute (www.tomshardware.com); NVIDIA's DGX Station page now lists 748 GB of coherent memory (252 GB HBM3e plus 496 GB LPDDR5X) (www.nvidia.com). Moreover, NVIDIA contributes open designs (Open Compute Project) for entire racks. At OCP 2024, it published the

GB200 NVL72 architecture, enabling a single rack to interconnect *up to 72 GPUs* via NVLink at 1.8 TB/s each (developer.nvidia.com). These systemic innovations ensure that NVIDIA GPUs are not just chip products but the core of holistic data-center platforms.

Skilled integrators and cloud providers worldwide have quickly adopted these solutions. For example, NVIDIA's own press release (2020) lists Amazon Web Services, Google Cloud, Microsoft Azure, and leading supercomputing centers (Jülich JUWELS, Perlmutter at NERSC, etc.) as early A100 users (nvidianews.nvidia.com) (nvidianews.nvidia.com). NVIDIA's partnership with OEMs like Dell, HPE, and Lenovo has produced ready-to-deploy servers and workstations. Case studies below illustrate how these GPUs operate in the wild, from next-gen AI training clusters to visualization servers.

This report proceeds as follows: we first detail the **evolution of NVIDIA's data-center architectures**, then systematically compare the current GPU lineup and platform designs. We include extensive quantitative data on architecture, memory, FLOPS, interconnect, and power (Table 1). We discuss connectivity (NVLink, NVSwitch, NVL72), performance metrics, and MIG virtualization. Real-world deployments and use-cases (e.g. Azure ND GB300 v6 cluster, enterprise AI servers) are presented. Underlying these analyses are numerous authoritative sources: NVIDIA's technical blogs and specifications, industry reports, and academic articles (developer.nvidia.com) (www.datacenterdynamics.com) (nvidianews.nvidia.com), ensuring that all technical claims are well-supported. Finally, we consider the **implications and future directions**, such as U.S. licensing requirements for exports of certain NVIDIA products to China (www.reuters.com) (www.tomshardware.com) and how model sizes and energy costs might shape the next GPU generation.

NVIDIA Data-Center GPU Architecture Evolution

NVIDIA GPUs have evolved through successive architectures (Volta, Ampere, Hopper, Ada Lovelace and Blackwell), each tailored for increasingly diverse data-center workloads. We briefly outline major milestones:

- **Volta (2017)** introduced the V100 (Tesla V100), with first-generation Tensor Cores (for mixed-precision) and NVLink 2.0. It delivered 15.7 TFLOPS FP32 (single-precision) and 125 TFLOPS of Tensor performance in SXM2 form (14 and 112 TFLOPS for the PCIe card) (images.nvidia.com) (developer.nvidia.com).
- **Ampere (2020)**, NVIDIA's eighth generation, yielded the A100. The A100 GPU built on the Ampere GA100 chip (54.2 billion transistors on TSMC 7 nm) with *40 GB HBM2* memory (1.555 TB/s bandwidth (developer.nvidia.com)); the later 80 GB HBM2e version reaches 2.039 TB/s in SXM form (NVIDIA A100 data sheet). Key new features included third-generation Tensor Cores supporting TF32, BFLOAT16, and sparsity, plus *MIG virtualization*. NVIDIA's own data indicates A100 can be partitioned into up to *7 GPU instances* for improved resource utilization (developer.nvidia.com). In practice, A100 achieved roughly 312 TFLOPS dense FP16-matrix (624 with sparsity) and 19.5 TFLOPS FP32 (FP32 FMA) in SXM form (developer.nvidia.com) (developer.nvidia.com) (roughly 3–4× the V100). It also increased interconnect to NVLink 3.0 (600 GB/s per GPU). Variants of Ampere include the *A800 40GB Active* workstation GPU, which NVIDIA specifies with 40 GB of HBM2 memory, NVLink support, and up to seven MIG instances, as well as compute-focused cards such as the A40, A10, A30, and A16 for graphics, virtualization, and inference workloads.

- **Hopper (2022)**, NVIDIA's ninth generation, debuted with the H100 Tensor Core GPU. H100 is built on the Hopper architecture (≥80 billion transistors on 4 nm) and uses *80 GB HBM3* (3.35 TB/s); the H200 was the first GPU with HBM3e (www.nvidia.com). It introduced the *Transformer Engine* for FP8/FP16 mixed-precision (further accelerating large language models (nvidianews.nvidia.com)) and new DPX instructions. Official specs report the H100 SXM module delivers on the order of *67 TFLOPS FP32* and *1,979 TFLOPS FP16* (www.nvidia.com). The H100 also doubled NVLink speed to *900 GB/s per GPU* (NVLink 4.0) and supports NVSwitch fabrics for node-scale interconnect. In November 2023, NVIDIA announced the *H200* GPU, an incremental Hopper upgrade with *141 GB HBM3e* and 4.8 TB/s of memory bandwidth, with systems shipping from the second quarter of 2024 (nvidianews.nvidia.com). Its compute matches the H100 SXM (67 TFLOPS FP32, 1,979 TFLOPS FP16 Tensor with sparsity) (www.nvidia.com). H200 targets scaling beyond H100 for extreme AI/HPC; multiple H200 cards can also be linked (see below).
- **Superchips (Grace)**: In 2022–2025 NVIDIA combined its GPUs with arm-based CPUs into superchips. The *GH200* (Grace + Hopper), *GB200* and *GB300* (Grace + Blackwell) pair a 72-core Arm Neoverse V2 Grace CPU with one or two NVIDIA GPUs on a single module (www.nvidia.com). For example, the GB300 "Grace Blackwell Ultra" desktop superchip features a 72-core Grace CPU and a "Blackwell Ultra" GPU. Its claimed peak AI throughput is ~20 PFLOPS and early partner specifications list *784 GB* of unified LPDDR5X+HBM3E memory (www.tomshardware.com); NVIDIA's DGX Station now lists 748 GB (www.nvidia.com). These superchips enable CPU-GPU coherency and are aimed at large AI models and HPC workloads. The **GB200 NVL72** rack-scale system connects 72 Blackwell GPUs and 36 Grace CPUs via NVLink 5 at 1.8 TB/s per GPU. CoreWeave reached general availability for GB200 NVL72 cloud instances on February 4, 2025 (www.coreweave.com); Google Cloud followed with A4X VMs on May 29, 2025 (cloud.google.com).
- **Ada Lovelace (2023–)**: In 2023, NVIDIA released the *L40S* GPU for universal data-center acceleration of AI plus real-time graphics. L40S uses the Ada AD102 GPU (same chip as the consumer RTX 4090) but is configured for data-center use: it has *18,176 CUDA cores, 568 Tensor Cores, 142 RT cores,* and *48 GB GDDR6 with ECC* (www.nvidia.com) (www.techpowerup.com) (nvidianews.nvidia.com). According to NVIDIA, L40S achieves nearly 5× the single-precision (FP32) throughput of the A100 (nvidianews.nvidia.com), and 212 TFLOPS of ray-tracing performance via its RT cores (nvidianews.nvidia.com). It sits alongside similar cards (L40 without the S) and provides GPU acceleration for workloads like 3D visualization, virtual workstations, and AI inference, complementing the Hopper GPUs.
- **Blackwell Ultra / B300 (H2 2025)**: Announced at GTC in March 2025, Blackwell Ultra is the evolution beyond standard Blackwell. Each Blackwell Ultra GPU carries 288 GB of HBM3e at 8 TB/s and 15 PFLOPS of dense NVFP4 compute (developer.nvidia.com). NVIDIA lists the **GB300 NVL72** rack-scale system at 1,080 PFLOPS of dense FP4 (about 1.1 exaFLOPS), 1.5× the dense FP4 of Blackwell (www.nvidia.com). At the same GTC, NVIDIA introduced DGX Spark (formerly Project DIGITS), a desktop AI system based on the GB10 Grace Blackwell Superchip, and DGX Station, built on Blackwell Ultra. (NVIDIA GTC 2025 keynote updates)

- **Vera Rubin (H2 2026) and Beyond**: NVIDIA laid out this roadmap at GTC 2025 (www.cnbc.com) and launched the Rubin platform at CES in January 2026. Each Rubin GPU has 336 billion transistors, up to 288 GB of HBM4 with up to 22 TB/s of bandwidth, 50 PFLOPS of NVFP4 inference and 35 PFLOPS of NVFP4 training, and sixth-generation NVLink at 3.6 TB/s per GPU (developer.nvidia.com). The **Vera Rubin NVL72** rack pairs 72 Rubin GPUs with 36 Vera CPUs (88 custom Olympus cores each) for 3,600 PFLOPS of NVFP4 inference; NVIDIA's NVL72 product page lists 19.2 TB/s of HBM4 bandwidth per GPU (www.nvidia.com). In August 2026 NVIDIA said Vera Rubin was in full production, with racks running at partners (nvidianews.nvidia.com). **Rubin CPX**, a GDDR7-based GPU for long-context inference announced in September 2025, is expected at the end of 2026 (nvidianews.nvidia.com). **Rubin Ultra** follows in 2027 and **Feynman**, which NVIDIA presented at GTC 2026 as its next major architecture after Rubin (blogs.nvidia.com), is scheduled for 2028 using next-generation HBM memory (www.cnbc.com). Notably, NVIDIA is already using Blackwell GPUs to accelerate the design of these future architectures.

This architectural progression led to a **broad spectrum of GPU "solutions"** (accelerator cards and packages) tailored to data-center niches. Table 1 (below) compares selected NVIDIA data-center GPUs and related products, including Ampere (A100/A800), Hopper (H100/H200), Blackwell (B200/B300), Rubin, Ada (L40S), and workstation GPUs. Following sections elaborate on each, emphasizing interconnect and real deployments.

GPU Model	Arch. (Node)	Year	Memory (GB)	Memory Type	Mem BW (GB/s)	FP32 (TFLOPS)	FP16 Tensor (TFLOPS, with sparsity)	NVLink/ NVSwitch	MIG Partitions	TDP (W)	Comments
NVIDIA A100 (SXM)	Ampere GA100 (7 nm)	2020	40 (80 GB option)	HBM2 (HBM2e on 80 GB)	1,555 (2,039 on 80 GB)	19.5	624	8-way NVLink (600 GB/s per GPU) (developer.nvidia.com)	Up to 7 instances (developer.nvidia.com)	400	Introduced TF32/BF16 support, MIG virtualization (developer.nvidia.com).
NVIDIA A100 80GB (PCIe)	Ampere GA100	2020	80	HBM2e	1,935	19.5	624	NVLink Bridge for two GPUs (600 GB/s); PCIe Gen4 (64 GB/s)	Up to 7	300	PCIe variant; NVLink Bridge supports up to two GPUs. (NVIDIA A100 data sheet)

GPU Model	Arch. (Node)	Year	Memory (GB)	Memory Type	Mem BW (GB/s)	FP32 (TFLOPS)	FP16 Tensor (TFLOPS, with sparsity)	NVLink/NVSwitch	MIG Partitions	TDP (W)	Comments
NVIDIA A800 40GB Active	Ampere GA100	–	40 GB HBM2	HBM2	1,555	19.5	623.8 (peak tensor)	NVLink bridge (400 GB/s)	Up to 7	240	Workstation card with 6,912 CUDA cores and 432 Tensor Cores (www.nvidia.com).
NVIDIA H100 (SXM)	Hopper GH100 (4 nm)	2022	80 GB HBM3	HBM3	3,350	67 (www.nvidia.com)	1979 (www.nvidia.com)	8-way NVLink (900 GB/s per GPU) (developer.nvidia.com)	Up to 7	Up to 700	Transformer Engine, DPX; used for large AI/HPC.
NVIDIA H100 NVL (PCIe)	Hopper GH100 (4 nm)	2023	94 GB HBM3	HBM3	3,900	60 (www.nvidia.com)	1671 (www.nvidia.com)	NVLink bridge (600 GB/s); PCIe Gen5 (128 GB/s)	Up to 7	350-400	PCIe card announced March 2023 (nvidianews.nvidia.com); the original H100 PCIe had 80 GB HBM2e at 2,000 GB/s and a 350 W TDP (www.nvidia.com).
NVIDIA H200 (SXM)	Hopper (4 nm)	2024	141 GB HBM3e	HBM3e	4,800 (www.nvidia.com)	67	1,979 (www.nvidia.com)	8-way NVLink (900 GB/s)	Up to 7	Up to 700	Announced November 2023, systems shipping from Q2 2024 (nvidianews.nvidia.com); same compute as H100 SXM with 141 GB HBM3e at 4.8 TB/s.
NVIDIA L40S	Ada Lovelace (AD102)	2023	48 GB GDDR6 w/ECC	GDDR6	864	91.6	733 (www.nvidia.com)	PCIe Gen4 x16 (64 GB/s)	No MIG	350	Combines AI & graphics: 18,176 CUDA cores, 568 Tensor Cores, no NVLink; nearly 5× A100 FP32 (nvidianews.nvidia.com).
NVIDIA B200 (SXM)	Blackwell (TSMC 4NP)	2025	192 GB HBM3e (180 GB per GPU in DGX B200)	HBM3e	8,000	75	4,500	NVLink 5 (1.8 TB/s per GPU) (www.nvidia.com)	Up to 7	Up to 1,200	208B transistors; dual-die; 20 PFLOPS FP4 with sparsity, 10 dense (developer.nvidia.com); FP32 and FP16 are per-GPU shares of NVIDIA's HGX B200 totals (www.nvidia.com); powers DGX B200 ($515K for 8-GPU system).

GPU Model	Arch. (Node)	Year	Memory (GB)	Memory Type	Mem BW (GB/s)	FP32 (TFLOPS)	FP16 Tensor (TFLOPS, with sparsity)	NVLink/ NVSwitch	MIG Partitions	TDP (W)	Comments
NVIDIA B300 (Blackwell Ultra)	Blackwell Ultra (TSMC 4NP)	2025	288 GB HBM3e	HBM3e	8,000	75	4,500	NVLink 5 (1.8 TB/s per GPU)	Up to 7	Up to 1,400	15 PFLOPS dense NVFP4 per GPU (developer.nvidia.com); GB300 NVL72 rack: 1,080 PFLOPS dense FP4 (~1.1 ExaFLOPS), 1.5× Blackwell dense FP4 (www.nvidia.com) (blogs.nvidia.com).
NVIDIA Rubin (Vera Rubin NVL72)	Rubin	2026	288 GB HBM4	HBM4	19,200 to 22,000	130	4,000 (dense)	NVLink 6 (3.6 TB/s per GPU)	–	–	336B transistors; 50 PFLOPS NVFP4 inference, 35 PFLOPS NVFP4 training (developer.nvidia.com) (www.nvidia.com); in full production with racks running at partners as of August 2026 (nvidianews.nvidia.com).
NVIDIA GB10 (DGX Spark)	Grace Blackwell (GB10 Superchip)	2025	128 GB (unified)	LPDDR5X	273	–	–	No NVLink; ConnectX-7 (200 Gb/s)	–	140	Desktop AI system with a 20-core Arm CPU and up to 1 PFLOP of FP4 (www.nvidia.com); see our DGX Spark review.
NVIDIA RTX Pro 6000D (B40)	Blackwell-derived	2025	Not published	GDDR7	–	(n/a)	(n/a)	No NVLink (reported)	–	–	Blackwell variant for the China market reported by Reuters in May 2025 (GDDR7 instead of HBM, no CoWoS packaging); reported price ~$6,500 to $8,000 (www.tomshardware.com).
NVIDIA RTX Pro 4000 SFF	Blackwell SFF	2025	–	–	–	–	–	–	–	–	Mini form-factor Blackwell GPUs for workstations (www.tomshardware.com).

GPU Model	Arch. (Node)	Year	Memory (GB)	Memory Type	Mem BW (GB/s)	FP32 (TFLOPS)	FP16 Tensor (TFLOPS, with sparsity)	NVLink/NVSwitch	MIG Partitions	TDP (W)	Comments
NVIDIA Tesla V100	Volta GV100	2017	32 GB HBM2	HBM2	900	15.7	125	4-way NVLink (300 GB/s)	–	300	Earlier GPU, for reference; GV100 used TSMC's 12 nm FFN process. (NVIDIA Volta architecture white paper)
NVIDIA Tesla T4	Turing TU104	2018	16 GB GDDR6	GDDR6	300	8.1	65	PCIe Gen3 x16 (32 GB/s)	–	70	Inference/edge GPU; low-power.
Special: GB200 NVL72 Design	–	2024	–	–	–	–	–	72 GPUs @1.8 TB/s each (developer.nvidia.com)	–	–	NVL72 rack design: 72 GPUs in one domain (developer.nvidia.com).
Special: GB300 "Grace+GPU"	Grace (ARM) + GPU	2025	748 GB coherent in DGX Station (784 GB in early partner specs)	LPDDR5X+HBM3E	7,100 (GPU HBM3e)	–	–	Integrated NVLink (NVL72)	–	–	Combines Grace CPU + Blackwell Ultra GPU (www.nvidia.com) (www.tomshardware.com) (www.tomshardware.com).
Special: DGX SuperPOD (H100)	Cluster (NVLink Switch System)	2024	–	–	–	–	–	256 GPUs via NVSwitch	Up to 7 per GPU	–	Reference "SuperPOD" cluster: 256 NVSwitch-connected H100s.
Special: Azure ND GB300 v6	NVL72 Cloud Cluster	2025	37 TB fast memory (per rack)	HBM3e + LPDDR5X	Up to 576,000 HBM (per rack) (www.nvidia.com)	(see text)	(see text)	72 GPUs @1.8 TB/s; 130 TB/s NVLink per rack (azure.microsoft.com)	–	–	Microsoft Azure's GB300 NVL72 cluster (www.tomshardware.com).

GPU Model	Arch. (Node)	Year	Memory (GB)	Memory Type	Mem BW (GB/s)	FP32 (TFLOPS)	FP16 Tensor (TFLOPS, with sparsity)	NVLink/ NVSwitch	MIG Partitions	TDP (W)	Comments
Special: Nvidia OVX	L40S server design	2024	Varied	Varied	–	–	–	–	–	–	Partner servers with up to eight L40S GPUs for generative AI, graphics and digital-twin workloads (nvidianews.nvidia.com).

Notes: Columns show nominal values. “NVLink domain” lists how many GPUs can intercommunicate (per HGX board or cluster) and link bandwidth. (developer.nvidia.com) (developer.nvidia.com). “MIG” indicates NVIDIA’s multi-instance GPU splits (A100/H100 support). “Special” rows denote system-level designs (cluster or superchip) rather than a single card. All values are sourced from NVIDIA references and vendor data (developer.nvidia.com) (nvidianews.nvidia.com) (developer.nvidia.com) (www.nvidia.com).

Detailed Characteristics of NVIDIA Data Center GPUs

Memory and Compute

NVIDIA’s datacenter GPUs distinguish themselves by massive on-board memory and high bandwidth, enabling large models and datasets. For example, the **A100 40GB HBM2** version offers *1.555 TB/s* memory bandwidth (2.039 TB/s for the 80 GB HBM2e SXM version) (developer.nvidia.com), while the **H100** (80GB HBM3) roughly doubles that to *3.35 TB/s* (developer.nvidia.com). The **H200** pushes to *4.8 TB/s* with 141 GB of HBM3e (www.nvidia.com), Blackwell B200 and B300 reach 8 TB/s (developer.nvidia.com), and Rubin moves to HBM4 at up to 22 TB/s (developer.nvidia.com). In contrast, graphics-focused cards like the **L40S** use *GDDR6*, trading some throughput for lower cost; L40S has 48 GB with *864 GB/s* peak bandwidth (www.techpowerup.com). These differences reflect trade-offs: GDDR6 (on L40S/A40) cuts power and cost but limits inferencing throughput compared to HBM arrays.

Compute performance similarly scales. The A100 delivers ~19.5 TFLOPS of FP32 raw compute (312 TFLOPS dense FP16 Tensor, 624 with sparsity) on the SXM form (developer.nvidia.com). By contrast, H100 SXM reaches *67 TFLOPS FP32* and *1,979 TFLOPS FP16 Tensor with sparsity* (www.nvidia.com), roughly 3.4× and 3.2× the A100 figures. The H200 keeps the H100’s compute and adds memory capacity and bandwidth, which helps memory-bound inference (www.nvidia.com). NVIDIA also reports the L40S ADA card delivers nearly *5× the FP32 throughput of the A100* (nvidianews.nvidia.com), thanks to its greatly expanded shader and tensor core counts (18,176 CUDA cores vs A100’s 6,912). Table 1 summarizes these spec-level contrasts. NVIDIA positions the L40S for combined graphics and AI inference work; it has no NVLink and no MIG support (www.nvidia.com) (nvidianews.nvidia.com).

Multi-GPU Interconnect: NVLink, NVSwitch, and NVL72

A key component of the datacenter platform is the **GPU interconnect**. Traditional server architectures have PCIe, but NVIDIA supplements with NVLink (GPU-to-GPU links) and NVSwitch for full mesh fabrics. The Ampere A100/Hopper H100 era uses NVLink 3/4: *each H100 GPU can sustain 900 GB/s bidirectional* over NVLink (developer.nvidia.com), allowing 8 GPUs on an HGX baseboard to form a single large GPU. NVSwitch chips then stitch multiple NVLink domains together; for instance, in an NVIDIA DGX or SuperPOD, up to 16 or more H100s can communicate *as if on one bus* via NVSwitch enclosures (enabling all-to-all traffic at ~900 GB/s). This is vital for tightly-coupled training tasks. NVIDIA notes that with NVSwitch, up to **256** H100 GPUs can be linked in a single "SuperPOD" (these form the backbone of exascale AI systems) (www.ironsystems.com).

Going beyond NVSwitch, NVIDIA's newest NVL72 architecture shatters that limit. In the NVL72 design (Blackwell GPU version), *72 GPUs* occupy one NVLink domain (developer.nvidia.com). Each GPU still enjoys *1.8 TB/s* link speed (double H100's) (developer.nvidia.com), a monumental increase. This design achieves an **aggregate AllReduce bandwidth of ~260 TB/s** across the rack (developer.nvidia.com). Figure 1 (below) illustrates the difference:

- *Figure 1:* **GPU-to-GPU Network Comparison.** The traditional HGX H100 baseboard allows 8 GPUs with 900 GB/s links (NVLink 4.0) (developer.nvidia.com). On the right, the GB200 NVL72 design spans 72 GPUs with 1.8 TB/s links each (developer.nvidia.com) (developer.nvidia.com). (All-Reduce aggregate bandwidth is shown for NVL72.)

Interconnect	GPUs per Domain	Per-GPU Bandwidth	Aggregate AllReduce BW
NVLink (HGX H100 gen)	8 GPUs	900 GB/s (developer.nvidia.com)	~7.2 TB/s
NVL72 (GB200 design)	72 GPUs	1.8 TB/s (developer.nvidia.com)	260 TB/s (developer.nvidia.com)

These scaling leaps have profound impact. NVIDIA's analysis shows that moving from 8 to 72 GPUs in an NVLink fabric can accelerate giant AI models by *4–30*× (developer.nvidia.com). For example, a GPT-like 1.8 trillion-parameter model ("GPT-MoE-1.8T") could train ~4× faster and serve inference ~30× faster on NVL72 racks than on HGX H100 systems, according to NVIDIA. The real-world significance is evident in case studies (below) where multi-thousand-GPU clusters rely on NVL72/NVSwitch fabrics for scale.

Multi-Instance GPU (MIG) and Virtualization

NVIDIA also designed its GPUs for flexibility under virtualization. Starting with Ampere, GPUs like the A100/H100 can be partitioned by hardware into multiple **MIG instances** (developer.nvidia.com). For instance, up to *7 separate CUDA instances* can run in parallel on one A100, each with independent memory/slice. This boosts utilization in cloud settings. Table 1 notes the maximum MIG splits for each GPU (e.g. "7 instances"). The L40/L40S Ada GPUs do **not** support MIG (www.nvidia.com). Blackwell Ultra keeps seven-way MIG, for example seven instances of 34 GB each (developer.nvidia.com).

Memory Architecture and Efficiency

Another hallmark of NVIDIA's platform GPUs is advanced memory design. The use of *stacked DRAM* (HBM2, HBM2e, HBM3, HBM3e) on NVLink-connected boards effectively merges all GPU memory into one system. For example, an 8-GPU SXM node has 8×40 GB = 320 GB unified memory for an A100 system, or 8×80 GB=640 GB on H100. This coherence is critical for very large models. In the Grace superchips (GB300), the CPU and GPU share a unified address space (with 384-bit LPDDR5X plus HBM to yield hundreds of GBs) (www.tomshardware.com).

Energy and cooling are also constrained by memory and compute. NVIDIA's reference design for NVL72 had to address *120 kW of heat per rack* (developer.nvidia.com). To achieve that, they deployed direct liquid cooling manifolds and specialized blind-mate connectors (developer.nvidia.com). This level of thermal design (7 MW clusters built from these 72-GPU racks) illustrates how pushing datacenter GPU density requires new infrastructure. The OCP-published GB200 reference architecture (in collaboration with Vertiv) even details rack reinforcements and high-current busbars to handle 6,000 lbs of mating force and 1,400 A busbars (pglfmc.com) (developer.nvidia.com). In summary, each GPU generation followed by NVL72 demands commensurate upgrades in power/cooling design, making the GPU a centerpiece of system engineering.

Case Studies and Deployments

Azure GB300 NVL72 Supercluster for OpenAI Workloads

In October 2025, Microsoft announced one of the most extreme AI clusters ever built: a **GB300 NVL72 supercomputer** on Azure (www.tomshardware.com). This system stitches together *4,608 NVIDIA Blackwell Ultra GB300 GPUs* across NVL72 racks, each rack containing 72 GPUs and 36 Grace CPUs (www.tomshardware.com). The GPUs are connected with NVLink 5 (1.8 TB/s per GPU) and NVIDIA Quantum-X800 InfiniBand switches. In aggregate, this cluster delivers *~92.1 exaFLOPS* of FP4 inference performance (www.tomshardware.com). Even per rack, the performance is staggering: 72 GB300 GPUs plus 36 Grace CPUs yield ~1,440 petaflops and *37 TB* of memory, with 130 TB/s of NVLink bandwidth within each rack (www.tomshardware.com) (azure.microsoft.com). According to Microsoft, this cluster specifically accelerates OpenAI training tasks, cutting what used to take months down to weeks. This deployment concretely demonstrates the advantage of NVIDIA's scale: by enabling 72-GPU NVLink domains with GB300, Microsoft can train and serve trillion-parameter models at unprecedented speed (www.tomshardware.com) (developer.nvidia.com).

Supercomputer and Cloud Adoptions

By 2025, all major cloud providers have deployed NVIDIA Blackwell hardware at scale:

- **CoreWeave** became the first cloud to reach general availability for **GB200 NVL72** instances on **February 4, 2025** (www.coreweave.com).
- **Google Cloud** launched **A4X VMs** powered by GB200 GPUs in GA on **May 29, 2025**, making it the first provider to offer both A4 (B200) and A4X (GB200 NVL72) instances (cloud.google.com).

- **Microsoft Azure** deployed a large-scale GB300 NVL72 cluster primarily for OpenAI workloads, with ND GB200 v6 VMs available (azure.microsoft.com).
- **AWS** offers GB200 NVL72-based instances alongside continued availability of H100-based P5 instances.

GB200 NVL72 systems began shipping to major cloud providers starting December 2024, with mass production reaching full scale in Q2–Q3 2025. Earlier deployments of A100-based systems (Indiana University's "Big Red 200," NERSC's Perlmutter, and Germany's Jülich JUWELS Booster) paved the way for this rapid scaling (nvidianews.nvidia.com).

In the enterprise space, NVIDIA partners have deployed specialized servers. For instance, at SIGGRAPH in August 2023, NVIDIA and global system makers announced *L40S*-based OVX servers for generative AI, virtual workstation and 3D rendering workloads. OEMs (Dell, HPE, Lenovo, Supermicro) now feature L40/L40S in their offerings, optimizing for double-duty AI+graphics tasks (www.nvidia.com) (nvidianews.nvidia.com). Similarly, for on-premises AI deployments, NVIDIA's OVX platform is used for enterprise visualization and digital-twin workloads.

At the end-user workstation level, one case is the recently released **Asus ExpertCenter Pro ET900N G3**. This is essentially a workstation built around the *GB300* superchip (www.tomshardware.com). It delivers DGX-Station-like performance (20 PFLOPS AI) in a desktop chassis by using the integrated Grace CPU and Blackwell GPU on GB300. This indicates NVIDIA's strategy to proliferate data-center GPUs into smaller form factors via partners, showing the platform's versatility (www.tomshardware.com).

Cloud and Enterprise Technologies

Beyond raw compute, data-center GPUs enable new services. NVIDIA's software stack (CUDA, RAPIDS for data analytics, Triton Inference Server, etc.) is ubiquitous on these GPUs. OCP contributions like NVIDIA's ConnectX-7 NIC are becoming standards (OCP NIC 3.0) in cloud fabrics. Inference services (e.g. Azure AI, AWS SageMaker endpoints) rely on mixed fleets of GPU accelerators (L40S for vision, H100/H20 for NLP/hyperscale). Video transcoding in data centers is also sometimes done on GPUs. All these trends underscore how NVIDIA's datacenter GPUs are the cornerstone of modern compute infrastructure.

Analysis of NVIDIA GPU Compute and Connectivity

Performance and Scalability

We have already highlighted individual GPU FLOPS, but system-level benchmarks also matter. NVIDIA paper data and user reports indicate linear scaling with NVLink: an 8-GPU H100 node scales compute ~8× for large parallel jobs. The NVL72 fabric pushes this further: in principle, 72× scaling is achievable if communication remains hidden by NVLink's high bandwidth (1.8 TB/s) and topology. Analytical models from NVIDIA predict that even as model size growth outpaces single-GPU memory, these fabrics allow model parallelism with modest idle time. For example, the move from 8 to 72 GPUs can yield up to *30× faster inference* on massive language models (developer.nvidia.com). In practice, however, such scaling is only attained with optimized software (e.g. fully overlapping communication), which is an active area of research.

Aside from LINPACK-style FLOPS, *AI training metrics* are key. NVIDIA publishes examples: an A100 server (8×A100) can train GPT-3 (175B) ~3× faster than V100 era. At cluster scale, Microsoft's 4,608-GPU GB300 NVL72 deployment is rated at ~92 exaFLOPS of FP4 inference, with each 72-GPU rack contributing about 1.44 exaFLOPS. This illustrates the trend: as AI model sizes balloon from billions to trillions of parameters, GPU clusters must grow accordingly, a driving factor behind NVL72. It also raises energy concerns: larger clusters draw more power, which GPU efficiency improvements offset only in part. Efficiency (performance per watt) has improved generation-to-generation (H100 roughly doubles beyond A100). Still, a GB200 NVL72 rack holds 72 Blackwell GPUs rated at up to 1,200 W each (developer.nvidia.com), plus 36 Grace CPUs and NVLink switches, which is why NVIDIA's reference design targets 120 kW of cooling per rack (developer.nvidia.com).

Market and Ecosystem Context

NVIDIA's dominance is not just market size; it reflects a mature ecosystem. With an AI accelerator market share estimated at **80–92%** as of 2025 (down from ~98% in 2023, as AMD and custom silicon gain ground), NVIDIA's competitive moat remains the CUDA ecosystem: every major ML framework (PyTorch, JAX, TensorFlow), the RAPIDS analytics stack, and thousands of optimized libraries target NVIDIA hardware exclusively. NVIDIA is estimated to consume **77% of all wafers** for AI processors in 2025 (www.tomshardware.com).

AMD MI300X / MI350: AMD's CDNA-3 MI300X delivers 192 GB HBM3 (2.4× the H100) and 5.3 TB/s bandwidth (1.6× H100) (www.amd.com), making it genuinely superior for memory-bound inference workloads. Microsoft Azure and Oracle Cloud have adopted MI300X for workload diversification, and AMD signed a deal to supply Oracle Cloud in October 2025. AMD's data center revenue reached **$4.34 billion** in a recent quarter (+22.3% YoY), substantial in absolute terms, though a fraction of NVIDIA's run rate (www.cnbc.com). AMD's newer MI350 series (MI350X and MI355X) carries 288 GB of HBM3E at 8 TB/s, matching B300's capacity and bandwidth (www.amd.com).

Intel Gaudi 3: Intel's 2025 Form 10-K discusses Gaudi AI accelerators and describes a revised AI strategy centered on successive generations of inference-optimized GPUs, including Crescent Island.

Google TPU v7 Ironwood: Google released its 7th-generation TPU, **Ironwood**, in November 2025, a decade after its first custom AI ASIC. Ironwood delivers 4,614 TFLOPS per chip, which analysts describe as "on par with Blackwell." However, it remains Google-internal and is not available for third-party purchase.

AWS Trainium3: Amazon's Trainium3 delivers 2.52 PFLOPS FP8 per chip with 144 GB HBM3e. AWS claims 30–40% better price-performance versus competing hardware on its platform. Like TPUs, Trainium is primarily for internal use to reduce inference costs at hyperscale.

Custom Silicon Outlook: Custom ASIC shipments from cloud providers are projected to grow **44.6% in 2026**, while GPU shipments are expected to grow 16.1%. Custom silicon is expected to capture 15–25% of the total AI accelerator market, primarily for internal inference workloads at hyperscalers, increasing pressure on NVIDIA's market share in the long term (www.cnbc.com).

China and Export Requirements: NVIDIA's 2026 Form 10-K states that U.S. licensing requirements apply to certain data-center products exported to China and that licenses issued in August 2025 allowed shipments of certain H20 products to certain China-based customers (techcrunch.com).

Future Implications and Directions

Looking ahead, NVIDIA has now publicly confirmed a detailed multi-generation roadmap. **Blackwell Ultra (B300/GB300)**, shipping since H2 2025, delivers 1.1 ExaFLOPS of dense FP4 per rack. As of August 2026, NVIDIA reports Vera Rubin ramping into full production, with racks running at major cloud partners (nvidianews.nvidia.com). The confirmed successor roadmap is:

Architecture	Scheduled	Node	Memory	Rack-Scale Performance
Blackwell Ultra (B300)	H2 2025 (shipping)	TSMC 4NP	288 GB HBM3e, 8 TB/s	1.1 ExaFLOPS dense FP4 (NVL72)
Vera Rubin	H2 2026 (in full production)	TSMC 3nm	288 GB HBM4, up to 22 TB/s	3.6 ExaFLOPS NVFP4 inference (Vera Rubin NVL72)
Rubin CPX	End of 2026	TBD	128 GB GDDR7	8 ExaFLOPS of AI compute (Vera Rubin NVL144 CPX rack)
Rubin Ultra	2027	TBD	Next-gen HBM4	TBD
Feynman	2028	TBD	Next-gen HBM	TBD

(www.cnbc.com)

The **Vera Rubin NVL72** rack keeps 72 GPUs per NVLink domain but raises rack HBM4 bandwidth to 1,400 TB/s, against up to 576 TB/s for GB300 NVL72, and lists 3,600 PFLOPS of NVFP4 inference (www.nvidia.com) (www.nvidia.com). Rubin CPX adds 30 PFLOPS of NVFP4 and 128 GB of GDDR7 per GPU for long-context inference (nvidianews.nvidia.com). Integration trends (like Grace CPU pairing) will continue: GPUs will cooperate even more with DPUs (Data Processing Units) and CPUs via NVIDIA's DOCA software stack, which ties ConnectX NICs and BlueField DPUs into coherent workflows.

The rise of hyperscale GPU clusters also acts as a catalyst for cooling and power innovation. OCP contributions show substantial rack redesign, e.g., *7 MW GB200 NVL72 clusters* requiring chilled water at scale (developer.nvidia.com). In data centers, we may see more custom facilities (liquid-cooled pods, immersion cooling) specifically built around GPU densities.

On the application side, NVIDIA's platforms are enabling models of previously impossible scale. Trillion+ parameter LLMs (GPT-4 scale) are now trainable; inference services for them are deployed worldwide. Inference optimizations (FP8, quantization) and specialized runtimes (Triton, TensorRT) will make more efficient use of these GPUs.

U.S. export licensing remains relevant to NVIDIA's product availability in China. NVIDIA's 2026 Form 10-K identifies the H20 and H200 as products subject to licensing requirements for shipments to certain China-based customers (techcrunch.com) (www.tomshardware.com).

Regardless, the technical trajectory is clear: ever-higher computational throughput via denser GPUs and interconnects. This not only advances AI but also drives industries like genomics, physics, climate science, and finance. Farther in the future, NVIDIA has already named Rubin Ultra and Feynman as successor architectures, suggesting continuous scaling. The convergence of GPU and CPU (Grace series) may one day yield truly unified compute chips. For now, the **NVIDIA datacenter GPU platform**, encompassing the spectrum from T4 inference tasks to GB300 and Vera Rubin superclusters, provides the most potent general-purpose compute power available.

Based on current trends and sources, we conclude that NVIDIA will maintain its leadership into the next decade through both iterative performance gains (Vera Rubin in 2026, Rubin Ultra in 2027, Feynman in 2028) and strategic ecosystem partnerships. NVIDIA's fiscal year 2026 revenue of **$215.9 billion**, with data center as the primary driver, demonstrates the extraordinary scale of this platform's adoption. The architecture comparison and data presented here (from vendor documentation and industry analysis (developer.nvidia.com) (nvidianews.nvidia.com)) should serve as a definitive reference on the full range of NVIDIA's data-center GPUs and how they stack up today and evolve tomorrow.

References

All statements and data above are supported by the sources cited throughout (citations in parentheses). Key references include NVIDIA's official announcements and technical blogs (developer.nvidia.com) (developer.nvidia.com) (developer.nvidia.com), respected industry news (Tom's Hardware, TechRadar) (www.tomshardware.com) (www.techradar.com), and market analysis reports (www.datacenterdynamics.com) (www.reuters.com), as well as vendor specification databases (www.techpowerup.com) (www.nvidia.com). These sources provide specification tables, performance metrics, deployment case studies, and strategic context for each GPU discussed. Each citation in the text corresponds to a particular line or finding in those sources.

Source links

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

1. https://nvidianews.nvidia.com/news/rubin-platform-ai-supercomputer
2. https://www.nvidia.com/en-us/data-center/vera-rubin-nvl72/
3. https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-second-quarter-fiscal-2027
4. https://intuitionlabs.ai/articles/key-innovations-behind-chatgpt
5. 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
6. https://developer.nvidia.com/blog/nvidia-contributes-nvidia-gb200-nvl72-designs-to-open-compute-project/
7. https://nvidianews.nvidia.com/news/nvidia-global-data-center-system-manufacturers-to-supercharge-generative-ai-and-industrial-digitalization
8. https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-third-quarter-fiscal-2026

9. https://www.tomshardware.com/pc-components/gpus/nvidia-rtx-pro-6000d-b40-blackwell-gpus-reportedly-set-to-supersede-banned-h20-accelerators-in-china
10. https://intuitionlabs.ai/articles/mixture-of-experts-moe-models
11. https://intuitionlabs.ai/articles/openai-stargate-datacenter-details
12. https://intuitionlabs.ai/articles/private-llm-inference-biotech
13. https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-fourth-quarter-and-fiscal-2025
14. https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-fourth-quarter-and-fiscal-2026
15. https://www.tomshardware.com/tech-industry/artificial-intelligence/nvidia-to-consume-77-percent-of-wafers-used-for-ai-processors-in-2025-report
16. https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/
17. https://www.nvidia.com/en-us/data-center/h100/
18. https://www.tomshardware.com/pc-components/gpus/asus-brings-nvidias-gb300-blackwell-ultra-desktop-superchip-to-workstations-features-up-to-784gb-of-coherent-memory-20-pflops-ai-performance
19. https://www.nvidia.com/en-us/products/workstations/dgx-station/
20. https://nvidianews.nvidia.com/news/nvidias-new-ampere-data-center-gpu-in-full-production
21. https://www.datacenterdynamics.com/en/news/nvidia-gpu-shipments-totaled-376m-in-2023-equating-to-a-98-market-share-report/
22. https://www.reuters.com/world/china/nvidia-launch-cheaper-blackwell-ai-chip-china-after-us-export-curbs-sources-say-2025-05-24/
23. https://images.nvidia.com/content/technologies/volta/pdf/tesla-volta-v100-datasheet-letter-fnl-web.pdf
24. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-nvidia-us-2188504-web.pdf
25. https://www.nvidia.com/en-us/data-center/h200/
26. https://nvidianews.nvidia.com/news/nvidia-supercharges-hopper-the-worlds-leading-ai-computing-platform
27. https://www.nvidia.com/en-us/data-center/gb300-nvl72/
28. https://www.coreweave.com/news/coreweave-first-cloud-provider-to-announce-general-availability-of-nvidia-gb200-nvl72-instances
29. https://cloud.google.com/blog/products/compute/new-a4x-vms-powered-by-nvidia-gb200-gpus
30. https://www.nvidia.com/en-us/data-center/l40s/
31. https://www.techpowerup.com/gpu-specs/l40s.c4173
32. https://developer.nvidia.com/blog/inside-nvidia-blackwell-ultra-the-chip-powering-the-ai-factory-era/
33. https://blogs.nvidia.com/blog/nvidia-keynote-at-gtc-2025-ai-news-live-updates/
34. https://www.cnbc.com/2025/03/18/nvidia-announces-blackwell-ultra-and-vera-rubin-ai-chips-.html
35. https://developer.nvidia.com/blog/inside-the-nvidia-rubin-platform-six-new-chips-one-ai-supercomputer/
36. https://nvidianews.nvidia.com/news/nvidia-unveils-rubin-cpx-a-new-class-of-gpu-designed-for-massive-context-inference
37. https://blogs.nvidia.com/blog/gtc-2026-news/
38. https://www.nvidia.com/en-us/products/workstations/a800/
39. https://nvidianews.nvidia.com/news/nvidia-launches-inference-platforms-for-large-language-models-and-generative-ai-workloads

40. https://www.nvidia.com/content/dam/en-zz/Solutions/gtcs22/data-center/h100/PB-11133-001_v01.pdf
41. https://www.nvidia.com/en-us/data-center/gb200-nvl72/
42. https://www.nvidia.com/en-us/data-center/hgx/
43. https://www.nvidia.com/en-us/products/workstations/dgx-spark/
44. https://intuitionlabs.ai/articles/nvidia-dgx-spark-review
45. https://www.tomshardware.com/pc-components/gpus/nvidia-introduces-compact-blackwell-professional-graphics-cards-rtx-pro-4000-sff-and-pro-2000-gpus-launched-at-siggraph-2025
46. https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf?ncid=no-ncid
47. https://azure.microsoft.com/en-us/blog/microsoft-azure-delivers-the-first-large-scale-cluster-with-nvidia-gb300-nvl72-for-openai-workloads/
48. https://www.ironsystems.com/nvidia-solutions/data-center-gpu
49. https://pglfmc.com/latest_news/nvidia-contributes-nvidia-gb200-nvl72-designs-to-open-compute-project-us-pioneer-global-vc-difchq-sfo-india-singapore-riyadh-swiss-our-mind/
50. https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html
51. https://www.cnbc.com/2025/11/21/nvidia-gpus-google-tpus-aws-trainium-comparing-the-top-ai-chips.html
52. https://www.amd.com/en/products/accelerators/instinct/mi350.html
53. https://techcrunch.com/2025/04/15/nvidia-h20-chip-exports-hit-with-license-requirement-by-us-government/
54. https://www.techradar.com/pro/huawei-ascend-950-vs-nvidia-h200-vs-amd-mi300-instinct-how-do-they-compare

THE PUBLISHER

About IntuitionLabs

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

AI consulting and adoption

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

Software, data and life-science workflows

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

Enterprise platforms and regulated delivery

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

Work with IntuitionLabs

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

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

Website: https://intuitionlabs.ai

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