Tech & AI Systems | Jan 11, 2026 | 9 min read | By Bryan White

Why Nvidia Dominates AI: A History of CUDA and Parallel Computing

AI * Computing * Physics

From 3D Graphics to General-Purpose Parallel Computing

The trajectory of Nvidia Corporation -- from a 1993 San Jose startup to the primary hardware architect of the artificial intelligence economy -- is a singular case study in technological convergence. While the company's public identity for its first two decades was anchored in consumer video game graphics, its internal roadmap progressively steered toward general-purpose parallel computing. Nvidia's dominance in Artificial Intelligence (AI) was not merely fortuitous timing during the 2012 deep learning resurgence, but the outcome of a multi-decade engineering strategy to unify 3D rendering pipelines and scientific computation.[1][2] By tracing the architectural evolution from fixed-function polygon pipelines to the unified shaders of the G80, the CUDA software stack, and the low-precision tensor engines of Hopper and Blackwell, this report examines how the Graphics Processing Unit (GPU) became the foundational compute engine of modern science.[3][14][22]

Founded on April 5, 1993, by Jensen Huang, Chris Malachowsky, and Curtis Priem, Nvidia initially struggled in a crowded market of nearly 90 graphics accelerator vendors.[1][2] Its first chip, the NV1 (1995), relied on quadratic texture mapping rather than polygon rendering and nearly bankrupted the firm when Microsoft standardized DirectX around triangle polygons.[2] Pivoting to high-velocity six-month release cycles with the RIVA 128 (1997), Nvidia introduced the GeForce 256 in 1999 -- coining the term "GPU" by integrating hardware Transform and Lighting (T&L) directly onto the silicon die.[1][2] Offloading floating-point vector geometry from the serial Central Processing Unit (CPU) established the GPU as a dedicated mathematical co-processor.

Architectural Unification: GeForce 256, G80 Stream Multiprocessors, and CUDA

The G80 Unified Shader Architecture and SIMT Execution

Early programmable GPUs such as the GeForce 3 (2001) separated hardware into dedicated Vertex Shaders (geometry) and Pixel Shaders (fragment coloring), leaving silicon idle whenever workloads were imbalanced.[3] In November 2006, Nvidia's G80 architecture (GeForce 8800 GTX) abolished this split by introducing the Unified Shader Architecture built around general-purpose Stream Multiprocessors (SMs).[3][4][30]

Rather than relying on rigid CPU-style SIMD (Single Instruction, Multiple Data) vectors that stall on conditional branches, the G80 pioneered SIMT (Single Instruction, Multiple Threads) execution.[3][7][8] Hardware schedulers grouped individual threads into 32-thread "warps" executed in lockstep while automatically managing branch divergence and reconvergence.[3] Whether simulating fluid dynamics,

folding proteins in computational structural biology, or multiplying neural network weight matrices, SIMT transformed the GPU into a large-scale parallel floating-point supercomputer.

The CUDA Software Moat

Prior to 2006, scientists had to disguise numerical arrays as texture maps and shaders to compute on GPUs. Released alongside the G80, CUDA (Compute Unified Device Architecture) exposed the GPU instruction set directly to standard C/C++ programming.[1][4][5][6] By embedding CUDA hardware support across every consumer and workstation GPU shipped -- despite years of Wall Street skepticism over the added die area and R&D expense -- Nvidia seeded a global installed base across university laboratories long before deep learning commercialized.[2][6]

The Deep Learning Ignition: AlexNet to Volta Tensor Cores

The decisive proof of concept arrived at the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC). University of Toronto researchers Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton trained AlexNet -- an 8-layer Deep Convolutional Neural Network with 60 million parameters and 650,000 neurons -- across two consumer 3 GB GeForce GTX 580 GPUs using CUDA.[9][10][11] Employing model parallelism to split the network across both cards over five to six days of training, AlexNet slashed top-5 classification error from 26.2% to 15.3%, igniting the modern deep learning era.[10][12]

As neural networks scaled exponentially, Nvidia bifurcated its roadmap into consumer (GeForce) and data center architectures:[13][14]

* Pascal (P100, 2016): Broke the memory bandwidth bottleneck by pairing Chip-on-Wafer-on-Substrate (CoWoS) 3D-stacked High Bandwidth Memory (HBM2) (720 GB/s) with NVLink 1.0 (160 GB/s inter-GPU fabric, 5x PCIe bandwidth).[14][15]

* Volta (V100, 2017): Introduced the Tensor Core, a dedicated matrix-multiply-accumulate hardware unit executing 4x4 matrix GEMM operations (D = A x B + C) in a single clock cycle.[16][17] By adopting mixed-precision (FP16/FP32) arithmetic suited to the statistical tolerance of neural weights, Volta delivered 125 TFLOPS -- a nearly 6x leap over Pascal.[13][32]

The Generative Scaling Era: Ampere, Hopper, and Blackwell

Following the rise of Transformer architectures, Nvidia co-designed silicon and numerical formats to accelerate trillion-parameter Large Language Models (LLMs):

* Ampere (A100, 2020): Introduced TF32 math, 2:4 Structural Sparsity (doubling throughput by skipping zero-weight multiplications), and Multi-Instance GPU (MIG) partitioning into seven isolated cloud instances.[14][18][19]

* Hopper (H100, 2022): Engineered the Transformer Engine, dynamically shifting layer calculations between 16-bit and 8-bit floating point (FP8) to halve memory footprint and accelerate LLM training up to 9x over Ampere.[20][21][31]

* Blackwell (B200, 2025): Surmounted the photolithographic reticle limit (~800-900 mm) by fusing dual reticle-sized dies across a 10 TB/s chip-to-chip interconnect into a single logical GPU, while introducing block-scaled 4-bit precision (FP4) and

NVLink 5 to achieve up to 9,000 TFLOPS of dense/sparse inference throughput.[22][23][29]

| Architecture | | Year | | Process Node | | Key Architectural Innovation | | Peak Tensor Throughput | | Memory Bandwidth | | Max VRAM |

| Pascal (P100) | | 2016 | | TSMC 16nm | | CoWoS HBM2 & NVLink 1.0 | | 21.2 TFLOPS (FP16, pre-Tensor) | | 720 GB/s | | 16 GB |

| Volta (V100) | | 2017 | | TSMC 12nm | | 1st-Gen Tensor Cores (4x4 GEMM) | | 125 TFLOPS (FP16) | | 900 GB/s | | 32 GB |

| Ampere (A100) | | 2020 | | TSMC 7nm | | TF32, 2:4 Structural Sparsity & MIG | | 312 TFLOPS (FP16/BF16) | | 2,039 GB/s | | 80 GB |

| Hopper (H100) | | 2022 | | TSMC 4N | | Transformer Engine & Dynamic FP8 | | 990 TFLOPS (FP16) / 1,980 (FP8) | | 3,350 GB/s | | 80 GB |

| Blackwell (B200) | | 2025 | | TSMC 4NP | | Dual-Reticle Chiplet, FP4 & NVLink 5 | | 2,250 TFLOPS (FP16) / 9,000 (FP4) | | 8,000 GB/s | | 192 GB |

| Numerical Format | | Bit Width | | Dynamic Range (Approx.) | | Architecture Introduced | | Primary AI Workload |

| FP32 (Single) | | 32-bit | | 10-38 to 1038 | | Pre-2016 Baseline | | Scientific simulation & master weight accumulation |

| FP16 (Half) | | 16-bit | | 10-5 to 65,504 | | Volta (2017) | | Mixed-precision neural network training |

| BF16 (Bfloat16) | | 16-bit | | 10-38 to 1038 | | Ampere (2020) | | Stable LLM training with FP32 exponent range |

| TF32 (TensorFloat) | | 19-bit | | 10-38 to 1038 | | Ampere (2020) | | Default internal matrix math acceleration |

| FP8 (E4M3 / E5M2) | | 8-bit | | Format-dependent | | Hopper (2022) | | Transformer Engine training & high-speed inference |

| NVFP4 (Block-Scaled) | | 4-bit | | Micro-block scaled | | Blackwell (2025) | | Trillion-parameter LLM inference efficiency |

Strategic Software Moats, Sovereign AI, and Thermodynamic Limits

This technical trajectory inverted Nvidia's financial structure: by Q3 FY2026 (late 2025), Data Center revenue reached $51.2 billion per quarter (~90% of total revenue) compared to $4.3 billion for Gaming.[24][25] While competitors such as AMD (MI300X) and Intel (Gaudi 3) field competitive silicon, Nvidia's deep software ecosystem -- spanning CUDA, cuDNN, TensorRT, and NCCL collective communication libraries -- creates high switching costs even against hardware-agnostic compilers like OpenAI's Triton.[6][26][27]

However, as nations invest heavily in domestic "Sovereign AI" supercomputers,[22] scaling now confronts fundamental thermodynamic and grid constraints governed by the second law of thermodynamics: a single liquid-cooled Blackwell NVL72 rack draws 120 kW of electrical power, shifting the frontier of AI scaling from transistor lithography to heat dissipation and power-grid engineering.[28][29]

The Foundry of Intelligence

Nvidia's thirty-year ascent demonstrates how architectural investments in consumer 3D graphics inadvertently solved the core computational bottleneck of 21st-century artificial intelligence. By unifying programmable shaders in the G80, democratizing parallel C++ programming via CUDA, and iteratively compressing floating-point precision from 32-bit scalars down to 4-bit block-scaled tensors, Nvidia transformed the GPU from a video game peripheral into the industrial foundry of machine intelligence.[1][3][22][23]

Works Cited

* [1] NVIDIA Corporation (Nov 2, 2026). Our History: Innovations Over the Years. -- https://www.nvidia.com/en-us/about-nvidia/corporate-timeline/

* [2] Acquired Podcast (Mar 28, 2022). Nvidia Part I: The GPU Company (1993-2006). -- https://www.acquired.fm/episodes/nvidia-the-gpu-company-1993-2006

* [3] Sanglard, F. (May 2, 2020). A history of NVidia Stream Multiprocessor. -- https://fabiensanglard.net/cuda/

* [4] Nickolls, J., Buck, I., Garland, M., & Skadron, K. (Mar 2008). Scalable Parallel Programming with CUDA. ACM Queue, 6(2), 40-53. -- https://doi.org/10.1145/1365490.1365500

* [5] Garland, M., Le Grand, S., Nickolls, J., Anderson, J., et al. (Jul 2008). Parallel Computing Experiences with CUDA. IEEE Micro, 28(4), 13-27. -- https://doi.org/10.1109/MM.2008.57

* [6] Modular (Jan 9, 2026). What exactly is "CUDA"? (Democratizing AI Compute, Part 2). -- https://www.modular.com/blog/democratizing-compute-part-2-what-exactly-is-cuda

* [7] Glick, B. H. (Sep 21, 2025). SIMT vs SIMD: Second Edition. -- https://www.glick.cloud/blog/simt-vs-simd-second-edition

* [8] NVIDIA Developer Forums (Jun 28, 2009). SIMD Versus SIMT: What is the difference between SIMT vs SIMD. -- https://forums.developer.nvidia.com/t/simd-versus-simt-what-is-the-difference-between-simt-vs-simd/10459

* [9] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (May 24, 2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90. -- https://doi.org/10.1145/3065386

* [10] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (Dec 2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 25, 1097-1105. -- https://proceedings.neurips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

* [11] Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). Dive into Deep Learning: 8.1. Deep Convolutional Neural Networks (AlexNet). -- https://d2l.ai/chapter_convolutional-modern/alexnet.html

* [12] Pinecone (Jun 30, 2023). AlexNet and ImageNet: The Birth of Deep Learning. -- https://www.pinecone.io/learn/series/image-search/imagenet/

* [13] NVIDIA Blog (Jan 12, 2016). Accelerating AI with GPUs: A New Computing Model. -- https://blogs.nvidia.com/blog/accelerating-ai-artificial-intelligence-gpus/

* [14] NVIDIA Corporation (Sep 15, 2026). NVLink & NVSwitch: Fastest HPC Data Center Platform. -- https://www.nvidia.com/en-us/data-center/nvlink/

* [15] Network-Switch (Sep 17, 2025). The Evolution of NVIDIA NVLink: NVSwitch and PCIe Comparison. -- https://network-switch.com/blogs/networking/the-evolution-of-nvidia-nvlink-technology

* [16] DigitalOcean (May 6, 2022). Tensor Cores Explained in Simple Terms. -- https://www.digitalocean.com/community/tutorials/understanding-tensor-cores

* [17] Fasi, M., Higham, N. J., Mikaitis, M., & Pranesh, S. (Feb 10, 2021). Numerical behavior of NVIDIA tensor cores. PeerJ Computer Science, 7, e330. -- https://doi.org/10.7717/peerj-cs.330

* [18] Choquette, J., Gandhi, W., Giroux, O., Stam, N., & Krashinsky, R. (Mar 1, 2021). NVIDIA A100 Tensor Core GPU: Performance and Innovation. IEEE Micro, 41(2), 29-35. -- https://doi.org/10.1109/MM.2021.3061394

* [19] NVIDIA Corporation (May 2020). NVIDIA A100 Tensor Core GPU Architecture Whitepaper. -- https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf

* [20] Choquette, J. (May 2023). NVIDIA Hopper H100 GPU: Scaling Performance. IEEE Micro, 43(3), 9-17. -- https://doi.org/10.1109/MM.2023.3256796

* [21] NVIDIA Technical Blog (Mar 22, 2022). NVIDIA Hopper Architecture In-Depth. -- https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/

* [22] NVIDIA Corporation (Dec 18, 2025). The Engine Behind AI Factories: NVIDIA Blackwell Architecture. -- https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/

* [23] NVIDIA Developer Blog (Jun 24, 2025). Introducing NVFP4 for Efficient and Accurate Low-Precision Inference. -- https://developer.nvidia.com/blog/introducing-nvfp4-for-efficient-and-accurate-low-precision-inference/

* [24] Voronoi (Nov 20, 2025). Nvidia: From Gaming to Data Center Giant. -- https://www.voronoiapp.com/technology/-Nvidias-Shift-from-Gaming-to-Data-Center-Dominance-3153

* [25] NVIDIA Newsroom (Nov 19, 2025). NVIDIA Announces Financial Results for Third Quarter Fiscal 2026. -- https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-third-quarter-fiscal-2026

* [26] Teki, S. (Dec 9, 2025). Nvidia's AI Moat in 2025: A Deep Dive. -- https://www.sundeepteki.org/blog/nvidias-ai-moat-in-2025-a-deep-dive

* [27] Pak, A. (Jun 28, 2024). The CUDA Advantage: How NVIDIA Came to Dominate AI. -- https://medium.com/@aidanpak/the-cuda-advantage-how-nvidia-came-to-dominate-ai-and-the-role-of-gpu-memory-in-large-scale-m

* [28] NVIDIA Developer Blog (Aug 21, 2025). Scaling AI Inference Performance and Flexibility with NVIDIA NVLink and NVLink Fusion. -- https://developer.nvidia.com/blog/scaling-ai-inference-performance-and-flexibility-with-nvidia-nvlink-and-nvlink-fusion/

* [29] Cockcroft, A. (Sep 28, 2024). Deep dive into NVIDIA Blackwell Benchmarks. -- https://adrianco.medium.com/deep-dive-into-nvidia-blackwell-benchmarks-where-does-the-4x-training-and-30x-inference-0209f197

* [30] Lindholm, E., Nickolls, J., Oberman, S., & Montrym, J. (Mar 2008). NVIDIA Tesla: A Unified Graphics and Computing Architecture. IEEE Micro, 28(2), 39-55. -- https://doi.org/10.1109/MM.2008.31

* [31] NVIDIA Developer Blog (Jun 4, 2025). Floating-Point 8: An Introduction to Efficient, Lower-Precision AI Training. -- https://developer.nvidia.com/blog/floating-point-8-an-introduction-to-efficient-lower-precision-ai-training/

* [32] NVIDIA Docs (Feb 1, 2023). Train With Mixed Precision. -- https://docs.nvidia.com/deeplearning/performance/mixed-precision-training/index.html