Nvidia Business Model

Idea In Short

Nvidia designs graphics processing units (GPUs), chips originally built to render video game worlds that turned out to be exceptionally good at the parallel math behind artificial intelligence (AI). Founded in 1993 by Jensen Huang, Chris Malachowsky and Curtis Priem, the company spent its first two decades competing for gamers' wallets before a strategic bet on general-purpose computing repositioned it as the primary supplier of chips that train and run AI models. That bet paid off spectacularly: Nvidia's data center revenue now dwarfs its gaming business, and the company became the first to cross a $4 trillion market valuation in 2025. This article traces how Nvidia built that position, why its CUDA software platform functions as a moat few competitors can cross, and why U.S.-China export restrictions have become one of the biggest risks to its business.

From gaming chips to a $4 trillion company

Jensen Huang, Chris Malachowsky and Curtis Priem founded Nvidia in 1993 with a focus on graphics chips for personal computers, launching the GeForce 256 in 1999 as the product that established the company's reputation in the gaming market. For more than a decade, Nvidia competed primarily against ATI, later acquired by AMD, for the business of powering ever more realistic video games. That business remained profitable and grew steadily, but it was not what eventually made Nvidia one of the most valuable companies in the world.

The turning point came in the mid-2000s, when Nvidia developed CUDA, a platform that let developers use GPUs for computing tasks well beyond graphics rendering. GPUs process many calculations in parallel, which happens to match the mathematical structure of training neural networks. As AI research accelerated through the 2010s, researchers increasingly turned to Nvidia hardware, and by 2021 the company's data center revenue had surpassed its gaming revenue for the first time. Nvidia crossed a $1 trillion market capitalization in mid-2023, then $3 trillion in mid-2024, and became the first public company to exceed $4 trillion in July 2025.1

CUDA as the moat competitors can't cross

CUDA is not a single product but an accumulated software ecosystem built over roughly two decades, now used by an estimated 5 million developers. Every major AI framework, from PyTorch to TensorFlow, runs most efficiently on Nvidia hardware because CUDA has become the default layer developers build against. Switching a large AI infrastructure to a competitor's chips means rewriting and re-optimizing code that took years to tune for Nvidia's architecture, a cost few companies are willing to absorb even when rival chips are cheaper.

> Moving away from CUDA is a multiyear, multibillion-dollar hurdle for any customer.

This software lock-in compounds with Nvidia's supply chain relationships, particularly its manufacturing partnership with Taiwan Semiconductor Manufacturing Company (TSMC), which took more than a decade to build to its current scale. Combined, these factors give Nvidia an estimated 80% to 90% share of the AI accelerator market, a level of dominance rare in any hardware category.2

Data centers now drive the business

Nvidia's data center segment generated about $193.7 billion in fiscal 2026, roughly 90% of the company's total revenue of $215.9 billion for the year. That revenue comes from selling GPUs and full server systems to cloud providers, including Amazon Web Services, Microsoft Azure and Google Cloud, along with enterprises building their own AI infrastructure. Gaming, once Nvidia's entire identity, generated $16 billion in the same period, still a record for that segment but now a small fraction of the overall business.3

China export restrictions cut into growth

Nvidia's dependence on advanced chip exports has run directly into U.S. national security policy toward China. In April 2025, the Commerce Department ruled that Nvidia's H20 chip, a version specifically downgraded to comply with earlier export limits, still required an export license to reach Chinese customers. Nvidia took a $4.5 billion charge in the first

quarter of fiscal 2026 related to H20 inventory and purchase commitments it could no longer fulfill. Policy reversed again in mid-2025, allowing H20 shipments to resume, and by early 2026 the administration had cleared even more advanced H200 chips for export to China, subject to a 25% tax.

The back-and-forth illustrates a structural risk in Nvidia's business: a large and previously growing market can be closed off or reopened by policy decisions outside the company's control. China's own AI champions, including Huawei, continue developing domestic alternatives, meaning even a full policy reversal might not fully restore Nvidia's prior market share there.4

Diversifying into automotive and professional visualization

Beyond gaming and data centers, Nvidia supplies chips and software platforms for autonomous vehicles through its DRIVE platform, working with automakers on advanced driver assistance systems and in-car computing. Volvo and other manufacturers use Nvidia hardware to run perception and decision-making software for self-driving features. The professional visualization segment, built on Nvidia's Quadro and RTX workstation chips, serves engineers and designers using computer-aided design and rendering software, a smaller but steady revenue line that predates the AI boom.5

Hyperscalers building their own chips

Nvidia's largest customers are also its most capable potential competitors. Google has developed its own AI chips, called Tensor Processing Units (TPUs), since 2015 and released its seventh generation in November 2025, while Amazon and Microsoft have pursued similar custom silicon programs through chips like Trainium. Midjourney, an AI image-generation company, reported cutting its monthly compute costs from $2.1 million to $700,000 by shifting workloads from Nvidia GPUs to Google's TPUs, an example rivals point to as evidence that custom chips can undercut Nvidia's pricing for specific workloads.

Even so, Nvidia remains TSMC's largest customer, and the foundry's capacity allocation has favored Nvidia over newer entrants competing for the same advanced manufacturing slots. Custom AI chip sales are projected to grow faster than general-purpose GPU sales in 2026, but hyperscalers building their own silicon still rely on Nvidia GPUs for the bulk of their AI infrastructure, since CUDA compatibility and software maturity remain harder to replicate

than the chips themselves.6

Key Partners

Nvidia relies on TSMC as its principal chip manufacturer, a relationship built over more than a decade that gives Nvidia priority access to advanced manufacturing capacity. Cloud providers including Microsoft, Amazon and Oracle serve as both customers and infrastructure partners, deploying Nvidia GPUs at massive scale. Automakers like Volvo integrate Nvidia's DRIVE platform into their vehicles, while companies such as Foxconn assemble Nvidia's server systems for data center deployment.

Key Activities

Research and development sits at the center of Nvidia's operations, spanning GPU architecture design, AI software frameworks and chip packaging innovations released on an annual cadence. Manufacturing coordination with TSMC and other foundry partners ensures Nvidia can scale production of its most advanced chips despite persistent capacity constraints. Collaboration with software developers, cloud providers and automakers keeps CUDA and Nvidia's broader software stack embedded across industries.

Key Resources

Nvidia's most valuable resource is arguably its CUDA software ecosystem, which locks in developers far more durably than any single chip generation. Its intellectual property portfolio, covering GPU architecture, AI acceleration techniques and networking technology gained through its acquisition of Mellanox, protects its technical lead. A highly skilled engineering workforce and a global brand reputation built first in gaming and now in AI computing round out the company's core resources.

Value Propositions

For AI developers and cloud providers, Nvidia offers the highest-performing chips available for training and running large models, paired with a mature software stack that reduces engineering overhead. For gamers, Nvidia provides GPUs capable of real-time ray tracing and high frame rates that competitors have struggled to match consistently. For automakers, the company offers an end-to-end platform spanning in-vehicle computing hardware,

autonomous driving software and the data center infrastructure needed to train those systems.

Customer Relationships

Nvidia maintains direct relationships with its largest customers, the hyperscale cloud providers and major AI labs, through dedicated account teams and co-engineering arrangements on future chip generations. Enterprise and automotive customers work with Nvidia through longer sales cycles involving technical integration support. Individual gaming consumers interact with the company primarily through retail partners, online support channels and its developer community forums.

Channels

Direct sales to large cloud providers and enterprise customers account for the bulk of Nvidia's data center revenue, often involving multiyear supply agreements. Retail and e-commerce partners, including major electronics retailers, distribute gaming GPUs to individual consumers. Nvidia's own website and developer portal serve as channels for software downloads, technical documentation and the CUDA toolkit that keeps developers engaged with its ecosystem.

Customer Segments

Cloud providers and AI labs building large-scale training and inference infrastructure represent Nvidia's largest and fastest-growing customer segment. Gamers and content creators remain a substantial base, purchasing GPUs for high-performance PCs and streaming setups. Automakers developing autonomous and driver-assistance systems, along with professionals in engineering and design who rely on visualization workstations, round out Nvidia's customer base.

Cost Structure

Research and development represents one of Nvidia's largest expense categories, reflecting the pace at which the company must iterate on chip architecture to stay ahead of competitors. Manufacturing costs paid to TSMC and other foundry partners scale with the volume and complexity of chips produced each generation. Stock-based compensation and

selling, general and administrative expenses complete the company's primary cost lines, alongside export-related charges tied to shifting trade policy.

Revenue Streams

Data center sales of GPUs and full server systems generate the large majority of Nvidia's revenue, sold to cloud providers and enterprises building AI infrastructure. Gaming GPU sales remain a substantial, if now secondary, revenue stream tied to consumer hardware upgrade cycles.

- 1Nvidia becomes first company to reach $4 trillion market cap
- 2Nvidia's CUDA lock-in and supply scarcity
- 3Nvidia fourth-quarter and fiscal 2026 financial results
- 4Nvidia H200 chips cleared for China export in 2026
- 5Nvidia automotive and DRIVE platform partnerships
- 6Hyperscalers scale custom AI chips alongside Nvidia GPUs

Summary

Nvidia's transformation from a graphics card maker into the infrastructure layer of the AI economy did not happen by accident. The company invested in CUDA for nearly two decades before AI demand made that investment pay off, and it has kept pouring resources into the software ecosystem that keeps customers locked into its hardware. That positioning now generates the overwhelming majority of its revenue from data centers rather than the gaming market that built its brand. But concentration brings risk: a handful of cloud providers account for an outsized share of purchases, rivals like AMD and custom chip programs at Google, Amazon and Microsoft are chipping at the edges, and shifting U.S. export policy toward China has already cost Nvidia billions. Whether Nvidia keeps its lead depends less on any single chip generation than on whether CUDA stays the default language developers reach for.