NVIDIA Corporation

The compute platform the entire AI buildout runs on

Explainer · June 1, 2026

Designs the GPUs and the CUDA software platform that train and run most of the world's AI workloads.

COMPANY	NVIDIA Corporation	HEADQUARTERS	Santa Clara, California, USA
OWNERSHIP	Public	TICKER	NASDAQ: NVDA
ENTERPRISE VALUE	~$5.2T (approx., June 2026)	EV / EBITDA	~32x (approx., TTM)
CEO	Jensen Huang		

ABSTRACT

NVIDIA designs the graphics processing units (GPUs) and the surrounding software platform that train and serve the bulk of modern artificial-intelligence workloads. What once was a gaming-chip company is now overwhelmingly a data-center business: roughly nine of every ten revenue dollars in fiscal 2026 came from selling AI compute to cloud providers, model labs, and enterprises. The distinctive asset is not the silicon alone but the pairing of a fast-moving GPU architecture roadmap with CUDA, a proprietary parallel-computing software layer that developers have built on for nearly two decades. That combination produces both raw performance and high switching costs. The implications are concentrated: NVIDIA's fortunes are now tied tightly to the durability of AI capital spending, to its ability to ship each new architecture on schedule against capacity constraints, and to whether its software lock-in holds as well-funded rivals and customers' own chips attack the margin.

Keywords: NVIDIA; GPU; CUDA; AI data center; Blackwell; accelerated computing; AI capex; semiconductors

1. Snapshot

NVIDIA designs GPUs and the accelerated-computing platform on which most AI is built. NVIDIA was founded in 1993 by Jensen Huang, Chris Malachowsky, and Curtis Priem, and is headquartered in Santa Clara, California. Huang, who still leads the company as CEO, co-founded it after engineering and management roles in the chip industry. NVIDIA is public (NASDAQ: NVDA) and now sits above a $5 trillion valuation, though the precise figure varies across data sources and no filing-based number anchors it. It is one of the largest companies in the world by market value, with a workforce in the tens of thousands. As a public company most headline financials are disclosed; what remains murkier are the quantified effects of China export restrictions, true end-customer concentration, and the cleanest current operating-segment breakdown, all of which are reported inconsistently.

2. Thesis: Why This Company, Why Now

The bet is that NVIDIA has become the indispensable platform layer of AI computing, and that the structural shift toward training and serving large models is durable rather than a spike. The evidence is in the revenue mix. NVIDIA reported record full-year FY2026 revenue of $215.9 billion (up roughly 65% year over year), with Data Center revenue of $193.7 billion (up about 68%), versus Gaming ($16.04B), Professional Visualization ($3.19B), and Automotive ($2.35B). Data Center is therefore close to 90% of the company.

This is the most direct AI-demand linkage in the market: NVIDIA does not merely benefit from the AI compute wave, it is the product the wave is buying. In Q1 FY2027, the quarter ending April 26, 2026, NVIDIA posted record quarterly revenue of $81.6 billion, up 85% year over year, with Data Center at a record $75.2 billion, up 92%, and GAAP diluted EPS of $2.39. The reachable market is large but cyclical, and that concentration is the thesis and the risk in one figure.

3. The Core Idea in Plain English

NVIDIA sells the engines and the operating manual for parallel computation. A CPU is like a few brilliant generalists solving problems one after another; a GPU is like thousands of specialized workers doing the same arithmetic at once, which is exactly the shape of training and running a neural network. That hardware advantage is only half the story. The other half is CUDA, the software framework that lets developers actually program those thousands of cores. Old world: graphics chips drew video-game frames. New world: the same architecture, plus a deep software stack, became the substrate for the entire AI industry, so buying an NVIDIA system means buying into a programming environment, not just a piece of silicon.

4. The Technical Space

The underlying problem is throughput: AI training and inference are enormous volumes of matrix math that must be executed in parallel, cheaply, and reliably, often across thousands of chips wired together. Accelerated computing solves this by offloading that math from general-purpose CPUs to specialized processors.

The standard approaches fall into three camps. General-purpose GPUs are the dominant tool, flexible enough to handle research and production. Custom accelerators, the ASICs that hyperscalers design for their own workloads, trade flexibility for efficiency on a known task. And a long tail of startup silicon chases specific bottlenecks like memory bandwidth or inference latency.

What "good" looks like turns on a few dimensions that actually matter. Raw performance per chip is the headline, but performance at the cluster level, how well thousands of chips communicate without stalling, increasingly decides real-world training speed. Performance per dollar and per watt governs total cost of ownership at data-center scale. And the software ecosystem, the libraries, compilers, and developer familiarity, determines how quickly a chip can be put to productive use. A faster chip nobody can easily program is not actually good.

5. How Their Technology Works (and What's Proprietary)

NVIDIA's stack decomposes into two reinforcing layers.

- **The architecture roadmap.** NVIDIA ships successive GPU generations on an aggressive cadence. Its competitive position rests on the proprietary CUDA parallel-computing platform, first released in 2007, plus a successive GPU architecture roadmap: Hopper (the H100, 80 billion transistors, built on TSMC's 4N process) advancing to Blackwell (B100/B200, and the GB200 Grace Blackwell Superchip). The 4N node is a custom NVIDIA process variant at TSMC, not the generic N4. The Grace Blackwell design pairs GPUs with NVIDIA's own CPU and high-speed interconnect, attacking the cluster-level communication bottleneck that increasingly gates training.

› **The CUDA software layer.** CUDA is the programming model and the library ecosystem built on top of it over nearly two decades. It is what turns thousands of cores into something developers can actually target, and it is the harder thing to replicate.

On defensibility, the two pieces differ. The architecture itself is genuinely hard but not unassailable: TSMC fabricates for everyone, and well-funded rivals and hyperscalers can and do design competitive silicon. CUDA is the deeper technical moat, because two decades of accumulated libraries and developer muscle memory cannot be cloned by taping out a faster chip. That said, treating CUDA as absolute lock-in overstates it. NVIDIA's recent collaboration with Intel on co-designed x86 CPUs and integrated SoCs suggests the company is extending its position through partner hardware, which complicates a pure software-exclusivity story. The accurate read is a significant, layered moat rather than an impregnable one.

6. Business and Go-to-Market

NVIDIA sells hardware systems and platforms, increasingly as integrated rack-scale machines rather than loose chips, to cloud providers, AI labs, enterprises, and OEM partners who build servers around its silicon. The motion is a mix of direct relationships with the largest buyers and a broad partner channel. Gross margins are high by hardware standards, a function of the software-plus-silicon bundle that lets NVIDIA price on delivered AI performance rather than component cost.

Capital returns underline the cash generation. NVIDIA returned approximately $20.0 billion to shareholders in Q1 FY27 via repurchases and dividends and announced an additional $80.0 billion share repurchase authorization.

Historically the company organized itself around four markets. NVIDIA's processors and platforms have historically been organized around four markets, Gaming, Professional Visualization, Data Center, and Automotive, a focus the company set in 2014. That framing is now dated: reporting has shifted, and in Q1 FY27 the non-data-center lines were grouped into an "Edge Computing" category of $6.4 billion, a taxonomy change worth noting when comparing across periods. The commercial reality is simpler than the old segmentation suggests, with the data-center business carrying the company.

7. Competitive Landscape and Moats

The comp set spans direct GPU rivals, custom-silicon efforts, and software challengers: AMD, Intel, the hyperscalers' in-house accelerators (Google's TPU, Amazon's Trainium/Inferentia), and a long tail of inference-focused startups.

AMD is the closest direct competitor. Where NVIDIA wins is the software ecosystem and the integrated rack-scale system: CUDA's maturity and NVIDIA's interconnect make full clusters easier to stand up and program. Where it can lose is on price and supply, where AMD's roadmap traction and reports of customer migration give buyers a credible second source, especially for inference, precisely where switching costs are lowest. The other rivals press at the edges: hyperscaler ASICs erode NVIDIA's share inside the largest buyers' own fleets, while Intel partners on co-designed hardware as much as it competes.

Three moats matter.

› **Software ecosystem lock-in.** CUDA's installed base of developers and libraries is the durable, hardest-to-copy advantage, on top of the technical edge above.

› **Scale and supply position.** NVIDIA's volume commands priority access to constrained manufacturing and advanced packaging, which is itself a barrier when capacity is the bottleneck.

› **System-level integration.** Owning GPU, CPU, and interconnect lets NVIDIA optimize the whole rack, which is harder to match piecemeal.

The platform risk is real: every major customer is also a potential competitor building its own chips, and one independent claim of roughly 92% discrete-GPU share traces only to vendor materials, not a third-party tracker, so treat the dominance as large but unquantified.

8. Risks and Open Questions

The picture would change most on AI-capex cyclicality, supply, and the durability of the software moat. The questions I would put to management:

- How exposed is revenue to a slowdown in hyperscaler AI capital spending, and what is the visibility into multi-year purchase commitments versus one-time buildout?
- What is the quantified revenue impact of U.S. export restrictions on China, given that China revenue is currently excluded from guidance with no disclosed figure?
- How much of the largest customers' future demand is shifting to their own in-house accelerators, and over what timeframe?
- Can the architecture roadmap ship on schedule against manufacturing and advanced-packaging capacity constraints, and what is the realistic supply ceiling?
- How sticky is CUDA for inference specifically, where rivals and open frameworks are most credible and switching costs lowest?

A diligence flag on the financials: one source reports trailing-twelve-month revenue of $253.49 billion, above the FY2026 annual figure, likely because the TTM window absorbs the surging Q1 FY27 quarter. Reconcile the windows before relying on either.

9. Bottom Line

The core read: NVIDIA is the dominant platform of the AI computing era, with a data-center business that now defines the company and a CUDA-plus-architecture stack that is a genuine, layered moat rather than an absolute one. The single biggest reason it works is that the hardware advantage is wrapped in nearly two decades of software lock-in that rivals cannot tape out their way past. The one thing to watch next is the durability of AI capital spending and how quickly the largest customers' in-house silicon erodes NVIDIA's share from inside its biggest accounts.

10. For the Nerds

The architectural bet worth tracking is at the cluster, not the chip. The Grace Blackwell design is a statement that the binding constraint on large-model training has shifted from per-GPU FLOPS to interconnect bandwidth and memory coherence across thousands of accelerators. By co-packaging CPU and GPU and owning the high-speed fabric between nodes, NVIDIA is competing on system topology, which is far harder for a single faster ASIC to displace than a chip-versus-chip benchmark would suggest.

The open question is whether the CUDA moat holds at the framework layer. Most training already runs through higher-level abstractions, and compiler stacks that target multiple hardware backends are maturing. If portability becomes good enough that switching backends is a recompile rather than a rewrite, the lock-in weakens fastest in inference, where workloads are more standardized. NVIDIA's counter is to keep the whole-rack performance gap wide enough that portability is not worth the performance tax, a moving target that depends on staying a node and an architecture ahead.

WHERE SOURCES DISAGREE

Founding city

One source states NVIDIA was founded April 5, 1993 in Sunnyvale, California; other sources state 1993 with Santa Clara as the location

Unresolved: the sources do not reconcile whether "Santa Clara" refers to the founding site or only the current headquarters

Current market capitalization

One source (a stock-data page) puts it at roughly $5.47 trillion as of June 1, 2026; another (a broker page) shows $5.11 trillion, and an older snapshot shows ~$4.23 trillion

Unresolved: the figures derive from different real-time scrapes at different moments and no authoritative filing-based figure is in the evidence; "above $5 trillion" is the only safe framing

"First company to reach a $4 trillion market cap" (July 2025)

One account asserts NVIDIA was the world's first to cross $4T, and earlier crossed $3.3T in June 2024. The available evidence contains no cross-company market-cap history to substantiate the "first" superlative

Unresolved: no independent market-data record in the corpus supports the comparative claim

~92% discrete GPU market share

One reading cites NVIDIA holding roughly 92% of the discrete GPU market as of 2025. The figure traces only to a vendor press bundle and an encyclopedia entry, not to a third-party tracker (e.g., Jon Peddie/Mercury Research), while separate evidence shows AMD claiming roadmap traction and customer migration

Unresolved: no independent market-share source is present in the evidence

Trailing-twelve-month vs. annual financials

One source reports TTM revenue of $253.49 billion and net income of $159.61 billion, which exceeds the FY2026 annual revenue of $215.9 billion from the primary results

Unresolved: the TTM window likely incorporates the surging Q1 FY27 quarter, but the corpus does not explicitly reconcile the two figures

Whether CUDA alone explains NVIDIA's durability

One reading treats the CUDA-plus-architecture stack as a self-reinforcing lock-in. Competing evidence describes Intel and NVIDIA co-designing x86 CPUs and RTX SoCs, suggesting NVIDIA's position is being extended through partner hardware rather than CUDA exclusivity alone

Unresolved: the evidence shows ecosystem expansion that complicates, but does not falsify, the moat claim