Global TMT

NVIDIA: A leader amid a transformational mega trend; Initiating Coverage with Outperform

Guilherme Bellizzi Motta, CFA
+55 21 99989 1133
guilherme.bellizzi@safra.com.br

Silvio Dória
+55 11 3175 7929
silvio.v@safra.com.br

We are initiating coverage of NVIDIA Corporation (NASDAQ: NVDA) with an Outperform rating and YE'26 TP of USD 300/share (~40% upside). NVDA is a fabless semiconductor company and the global leader in accelerated computing, designing the full-stack of processors, networking, and software that form the compute backbone of the AI buildout. AI has emerged as the most consequential technology megatrend since the internet, and NVDA's GPUs are the natural hardware substrate upon which the entire ecosystem is built. Founded in 1993 and led by its visionary CEO Jensen Huang since inception, NVDA has evolved from a graphics chip company into an AI infrastructure platform. We believe its business model is progressively shifting from cyclical to recurring, with compelling upgrade dynamics and deeper customer lock-in with each generation — a shift that, in our view, remains underappreciated by the market.

Our Outperform rating is supported by four pillars: (i) a platform moat spanning chip architecture, proprietary scale-up and scale-out networking, privileged access to scarce manufacturing capacity, and the CUDA software ecosystem, reinforced by extreme co-design across the full-stack that competitors have been unable to replicate; **(ii)** an AI factory framework and annual generation cadence, in which each new architecture materially improves the operator's unit economics (tokens per watt, revenue per GW), translating NVDA's annual hardware cadence into recurring upgrade cycles that progressively reduce the cyclicality of the business; **(iii)** the agentic AI inflection as a new growth vector, in which inference workloads decouple compute demand from the initial training-driven investment cycle, as agentic workflows compound across the pre-training, post-training, and test-time compute scaling laws — chaining multiple reasoning model calls per task and materially multiplying tokens per interaction; and **(iv)** demand durability more resilient than the market currently prices, with TAM expanding across hyperscalers' AI buildout, sovereign AI programs, and enterprise CPU-to-GPU migration.

Forecasts and valuation. We forecast revenue to compound at a ~29.5% CAGR over FY'26A - FY'31E, driven predominantly by the Data Center segment, with NVDA's asset-light operating model sustaining an average gross margin of ~74%, an average EBIT margin of ~64%, and an average net margin of ~56% over the forecast horizon. This combination of top-line growth and sustained margins translates into a cumulative operating cash flow of ~USD 1.67trn over the next 5 fiscal years. **Given the massive cash flow generation and little reinvestment needs, we expect NVDA to return ~20% of the company's current market cap to shareholders through stock buybacks and dividends over the period**. On our estimates, NVDA currently trades at a still attractive ~24x FY'27E (~CY'26E) P/E and ~18x FY'28E (~CY'27E) P/E, below mega-cap tech and semiconductor peer medians. Our YE'26 target price of USD 300/share derives from a FCFF DCF.

Mind the risks. The principal risks to our thesis are customer concentration among hyperscaler companies, a potential deceleration in AI capex, competitive pressure from ASICs, and supply chain and geopolitical exposure. Secondary risks include power availability and grid constraints affecting AI deployment timing, inventory risk during architecture transitions, and key-person risk.

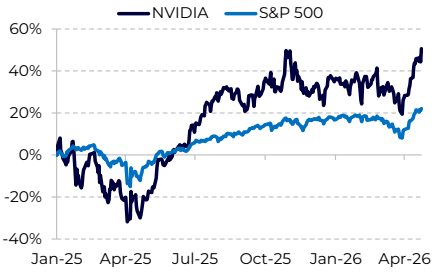
New to semiconductors? If you are new to the world of semiconductors and want to understand the complex supply chain of the industry (and where NVDA and its ecosystem participants sit across it), we recommend starting by reading **Appendix A** ("Semiconductors industry: a supply chain primer"). Furthermore, given the extensive technical terminology used throughout this report, we also provide a Glossary in **Appendix B**.

NVIDIA Corporation (NASDAQ: NVDA)	
Rating	**Outperform**
Local price	USD 213.13
Target price	USD 300.00
Upside	~40%

Multiples	FY27E	FY28E	FY29E
EV/EBIT	19.8x	14.9x	12x
EV/EBITDA	19.6x	14.7x	11.8x
P/E	23.8x	18.1x	14.6x
FCF Yield	3.4%	4.8%	6.1%

Market data	
Market cap	USD 5.2trn
52 week low/high	USD 103.11/216.61
ADTV (3m)	~USD 35bn
Fully-diluted shares outstanding	24,432mn
YTD performance	+14.3%

NVDA vs. S&P 500



Source: Bloomberg.

Comps table

Comps table	Price/GAAP Earnings			EV/GAAP EBIT			Next 3-years CAGR			PEG Ratio	ROIC, TTM
	CY'26E	CY'27E	CY'28E	CY'26E	CY'27E	CY'28E	Revenue	GAAP EBIT	GAAP EPS		
NVIDIA	**23.7x**	**18.0x**	**14.6x**	**19.9x**	**14.9x**	**12.0x**	**40.6%**	**43.4%**	**43.5%**	**0.55x**	**106%**
Mega-cap technology peers											
Apple	31.5x	28.9x	26.4x	26.5x	24.9x	23.3x	7.9%	8.1%	10.4%	3.0x	438%
Alphabet	29.2x	25.2x	21.8x	25.2x	21.5x	18.4x	15.6%	18.1%	11.2%	2.6x	51%
Microsoft	22.9x	20.1x	17.1x	18.7x	16.2x	13.8x	16.4%	16.5%	16.7%	1.4x	38%
Amazon	33.1x	26.7x	21.6x	27.9x	22.2x	17.7x	12.2%	20.0%	11.6%	2.8x	22%
Meta	22.2x	19.2x	16.5x	19.3x	16.5x	14.0x	19.7%	12.6%	20.8%	1.1x	45%
Broadcom	39.2x	25.3x	19.5x	33.9x	21.6x	16.8x	43.3%	43.7%	54.5%	0.7x	32%
TSMC	22.4x	17.1x	14.0x	18.7x	14.9x	12.4x	27.0%	31.0%	30.6%	0.7x	55%
ASML	38.1x	28.9x	25.2x	32.4x	24.7x	21.7x	16.9%	23.6%	25.1%	1.5x	88%
Median	**30.4x**	**25.3x**	**20.6x**	**25.8x**	**21.5x**	**17.2x**	**16.6%**	**19.0%**	**18.8%**	**1.44x**	**48%**
Semiconductors peers											
Broadcom	39.2x	25.3x	19.5x	33.9x	21.6x	16.8x	43.3%	43.7%	54.5%	0.7x	32%
TSMC	22.4x	17.1x	14.0x	18.7x	14.9x	12.4x	27.0%	31.0%	30.6%	0.7x	55%
ASML	38.1x	28.9x	25.2x	32.4x	24.7x	21.7x	16.9%	23.6%	25.1%	1.5x	88%
AMD	64.2x	35.0x	24.5x	54.4x	29.3x	21.2x	35.5%	51.6%	52.5%	1.2x	48%
Marvell Technology	85.5x	48.1x	30.5x	70.0x	42.4x	26.7x	33.0%	37.1%	37.2%	2.3x	55%
Applied Materials	34.0x	27.9x	25.0x	31.6x	24.7x	22.4x	13.2%	16.7%	17.4%	1.9x	32%
LAM Research	38.0x	28.9x	24.3x	32.9x	25.1x	21.4x	21.8%	27.9%	28.9%	1.3x	88%
KLA Corporation	43.7x	35.6x	31.6x	37.0x	30.1x	27.3x	14.6%	17.0%	18.6%	2.4x	48%
Median	**38.1x**	**28.9x**	**25.0x**	**32.9x**	**25.1x**	**21.7x**	**21.8%**	**27.9%**	**28.9%**	**1.52x**	**55%**

Source: Safra, Visible Alpha.

Safra vs. consensus

	Safra Estimates			Consensus Estimates			Safra vs. Consensus		
	FY'27E	FY'28E	FY'29E	FY'27E	FY'28E	FY'29E	FY'27E	FY'28E	FY'29E
Revenue	**386,815**	**497,877**	**600,537**	**363,279**	**481,869**	**565,226**	**6.5%**	**3.3%**	**6.2%**
Gross Profit (GAAP)	**290,111**	**370,919**	**442,896**	**269,824**	**359,882**	**419,482**	**7.5%**	**3.1%**	**5.6%**
Gross Margin %	*75.0%*	*74.5%*	*73.8%*	*74.3%*	*74.7%*	*74.2%*	*73bps*	*-18bps*	*-46bps*
Operating Income (GAAP)	**252,448**	**323,533**	**386,512**	**234,908**	**314,782**	**367,681**	**7.5%**	**2.8%**	**5.1%**
Operating Margin %	*65.3%*	*65.0%*	*64.4%*	*64.7%*	*65.3%*	*65.1%*	*60bps*	*-34bps*	*-69bps*
EBITDA	**256,268**	**328,519**	**392,943**	**238,942**	**320,137**	**373,901**	**7.3%**	**2.6%**	**5.1%**
EBITDA Margin %	*66.3%*	*66.0%*	*65.4%*	*65.8%*	*66.4%*	*66.2%*	*48bps*	*-45bps*	*-72bps*
Net Income (GAAP)	**216,822**	**280,275**	**337,080**	**199,917**	**267,615**	**318,060**	**8.5%**	**4.7%**	**6.0%**
Net Margin %	*56.1%*	*56.3%*	*56.1%*	*55.0%*	*55.5%*	*56.3%*	*102bps*	*76bps*	*-14bps*

Source: Safra, Visible Alpha.

Table of contents

Investment thesis

We are initiating coverage of NVIDIA (NASDAQ: NVDA) with an Outperform rating and YE'26 TP of USD 300/share (~40% upside). We believe NVDA is transitioning from a cyclical semiconductor designer to a platform business, generating upgrade economics that are increasingly recurring in nature and deepening customer lock-in with each architecture generation. The semiconductor cycle has not been repealed, but we believe that NVDA's business model is progressively reducing its cyclicality — a shift that, in our view, remains underappreciated by the market.

Our Outperform rating is supported by four pillars: (i) a platform moat spanning chip architecture, proprietary scale-up and scale-out networking, privileged access to scarce manufacturing capacity, and the CUDA software ecosystem, reinforced by extreme co-design across the full-stack that competitors have been unable to replicate; **(ii)** an AI factory framework and annual generation cadence, in which each new architecture materially improves the operator's unit economics (tokens per watt, revenue per GW), translating NVDA's annual hardware cadence into recurring upgrade cycles that progressively reduce the cyclicality of the business; **(iii)** the agentic AI inflection as a new growth vector, in which inference workloads decouple compute demand from the initial training-driven investment cycle, as agentic workflows compound across the pre-training, post-training, and test-time compute scaling laws — chaining multiple reasoning model calls per task and materially multiplying tokens per interaction; and **(iv)** demand durability more resilient than the market currently prices, with TAM expanding across hyperscalers' AI buildout, sovereign AI programs, and enterprise CPU-to-GPU migration.

I. Platform moat and extreme co-design

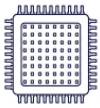
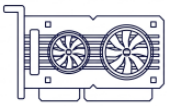
The semiconductor industry is inherently cyclical, and AI infrastructure is not an exception. However, NVDA's ecosystem lock-in and economically compelling upgrades across generations increasingly produce demand characteristics that resemble platform economics rather than the cyclical pattern the market applies.

A primer on GPU architecture and AI workloads

The GPU's dominance in AI is architectural. A **CPU** is built around a small number of powerful cores optimized for executing complex and sequential instructions with low latency, like running an operating system, managing a database or processing a spreadsheet.

A **GPU** is a fundamentally different design: it contains tens of thousands of smaller cores organized for massively parallel throughput. Originally built for rendering 3D graphics, GPU architecture maps directly onto linear algebra (matrix multiplications and vector operations) that underpins **neural networks** (the layered mathematical process by which AI models learn patterns from data). Where a CPU optimizes for low latency on each individual thread (a single sequence of instructions), a GPU maximizes **FLOPS** (floating point operations per second) across massively parallel threads.

Figure 1 – CPUs vs. GPUs: architecture

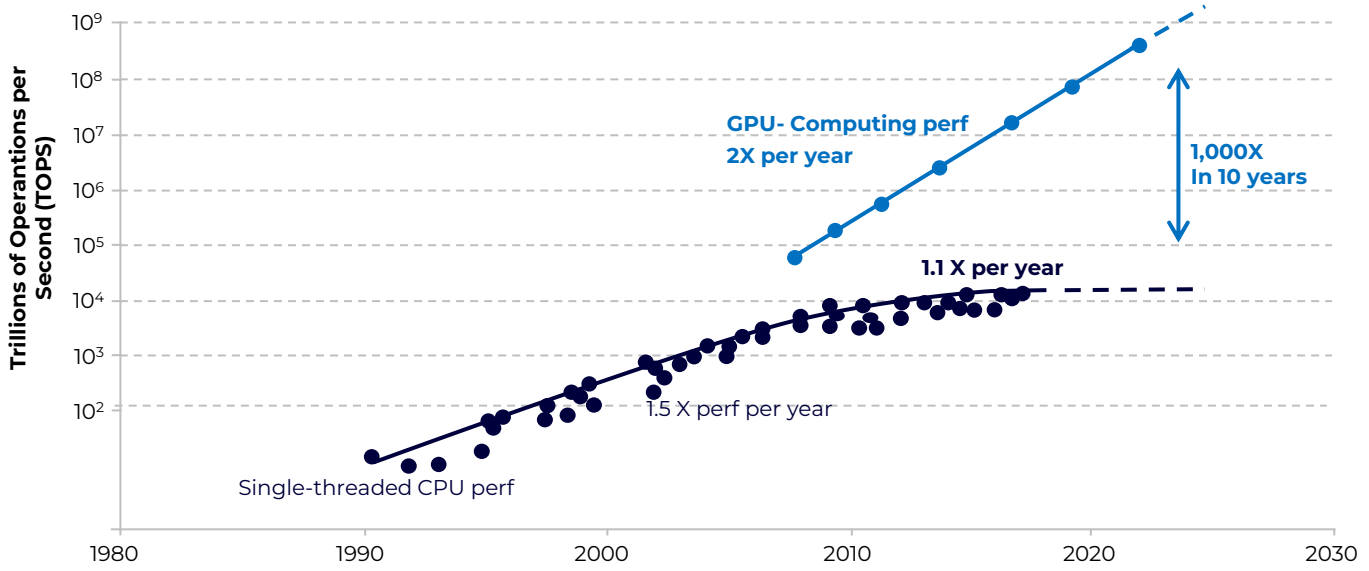
	Central Processing Unit (CPU)	Graphics Processing Unit (GPU)
		
Core count	Tens to low hundreds of cores	Tens of thousands of cores
Design philosophy	Low latency per thread	High aggregate throughput
Peak AI throughput	Single-digit TFLOPS (10^{12} operations/sec)	Tens of PFLOPS (10^{15} ops/sec) via Tensor Cores
Memory interface/bandwidth	DDR5 / Sub 1-TB/s	HBM / Multi-TB/s
Suited for	Sequential, branching, control-flow workloads	Parallel, uniform, compute-dense workloads
AI role	System orchestration, data management	Matrix math, model training and inference

Source: Safra.

At its core, an **LLM** (Large Language Model) processes input data through its numerical parameters, layer by layer, across a neural network. Within each layer, the operations are independent and identical in structure — matrix multiplications that can be distributed across the GPU cores simultaneously. This is precisely the workload GPUs were architected to execute, and the alignment between the mathematics of graphics rendering and the mathematics of neural networks is what positioned NVDA's GPUs as the foundation of modern AI.

The performance differential between GPUs and CPUs is not marginal. GPU performance has doubled roughly every year over the past decade, compounding to a ~1,000x improvement, while conventional CPU performance has grown at ~1.1x per year. More importantly, according to NVDA's Chief Scientist, only a ~2.5x multiplier within that ~1,000x improvement is attributable to semiconductor process advances, meaning that most of the performance differential is attributed to architecture, numerical precision, and software-level optimizations specific to GPU-accelerated workloads.

Figure 2 – CPUs vs. GPUs: computing performance over time (trillions of operations/second, log scale)

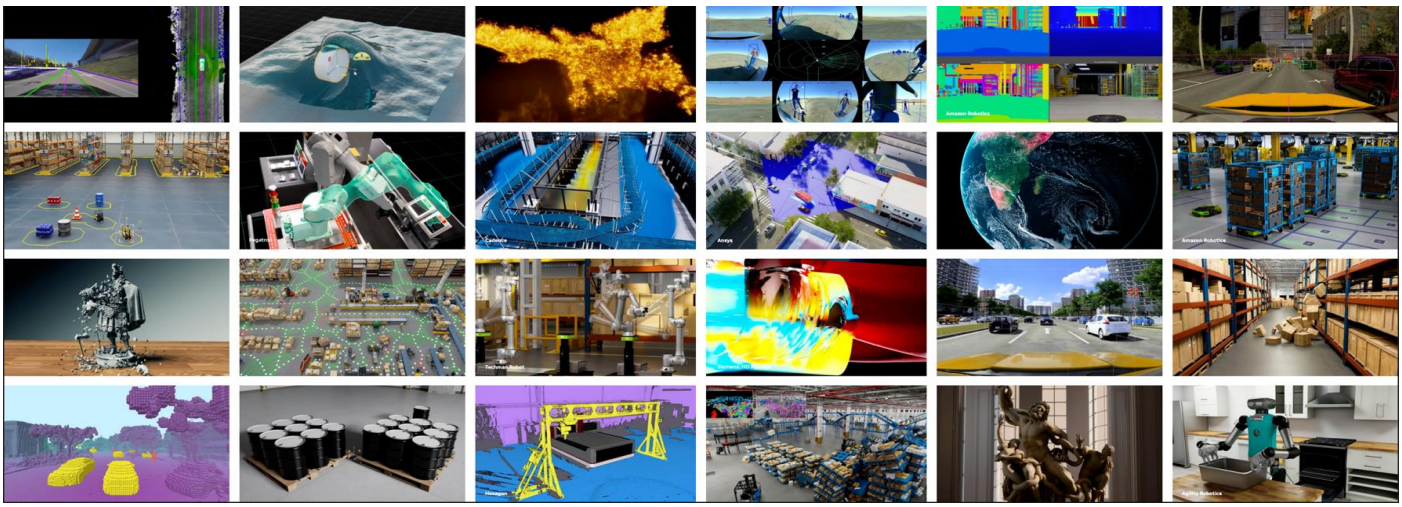


Source: Safra, NVIDIA.

The combination of parallel architecture, high computational throughput, and superior memory bandwidth (the speed at which data moves between memory and processing units, a bottleneck when handling the massive parameter sets of frontier models) is why virtually every state-of-the-art AI model today is primarily trained and served on NVDA's GPUs.

Beyond AI, the same parallel architecture accelerates workloads across computational geophysics, molecular dynamics and drug discovery, climate modeling, financial risk simulation, real-time rendering, and much more.

Figure 3 – Accelerated computing applications: autonomous driving, robotics, climate modeling, and much more



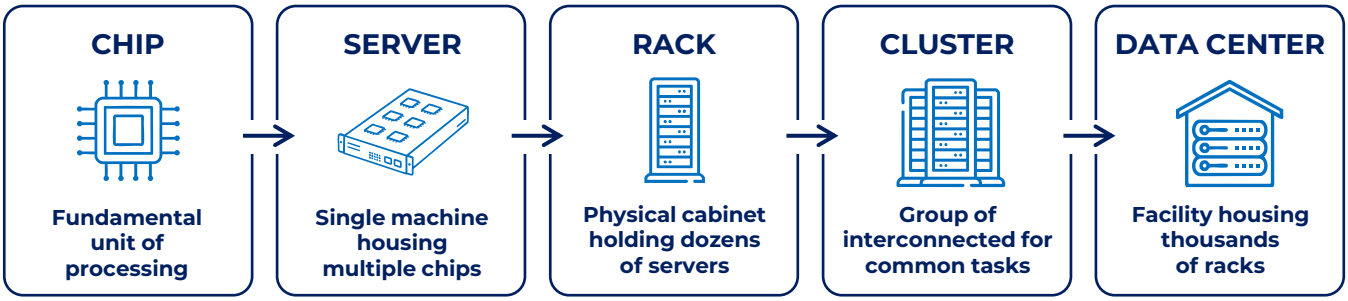
Source: Safra.

The data center as the unit of AI infrastructure

The relevant unit of AI infrastructure is not the individual chip but the data center. Modern AI workloads run on chips operating as a coordinated system, housed in purpose-built facilities.

A **server**, a single machine containing GPUs, CPUs, and memory, is mounted vertically inside a **rack**, a standardized physical cabinet that stacks dozens of servers with associated power and cooling infrastructure. Racks are organized into **clusters**, groups of interconnected systems assigned to work on a common task, that can span dozens or hundreds of racks. A single **data center** may house thousands of racks across multiple clusters.

Figure 4 – AI infrastructure hierarchy – from chip to data center



Source: Safra.

One of the binding engineering constraints in scaling these systems is communication: GPUs training an AI model must exchange data continuously, and any latency between chips — whether within a server, across a rack, or between clusters — idles the most expensive hardware in the system. **Networking**, the combination of specialized chips, cables, switches, and protocols that move data between processors at high speed, is the infrastructure that solves this problem.

NVDA's product scope today extends well beyond GPUs into a full-stack infrastructure platform. The Data Center segment (~90% of revenue as of FY'26A) is built on three processors type — **GPUs** (Blackwell/Rubin) for parallel AI computation, **CPUs** (Grace/Vera) for host processing, and, from 2H26 onward, **LPUs** (Groq 3 LPU) for low-latency decode acceleration — together with the networking silicon that connects them: **NVSwitch** chips that materialize the **NVLink** scale-up fabric within a rack, **ConnectX SuperNICs** and **BlueField DPUs** embedded in each server, and **Quantum** (InfiniBand) and **Spectrum-X** (Ethernet) switches that scale out across racks within a data center.

On top of hardware, NVDA's platform extends into software through **CUDA** (Compute Unified Device Architecture). Released in 2006, CUDA made GPUs programmable for general-purpose computation using familiar languages (C, C++, and Python), replacing the graphics-API approach that GPU computing had previously required and opening the architecture to a far broader developer base.

The corollary is that NVDA captures value at every layer of the AI data center — compute, scale-up and scale-out networking, and the software stack that binds them — through components designed and co-optimized as part of a single integrated architecture.

Amdahl's Law and the case for extreme co-design

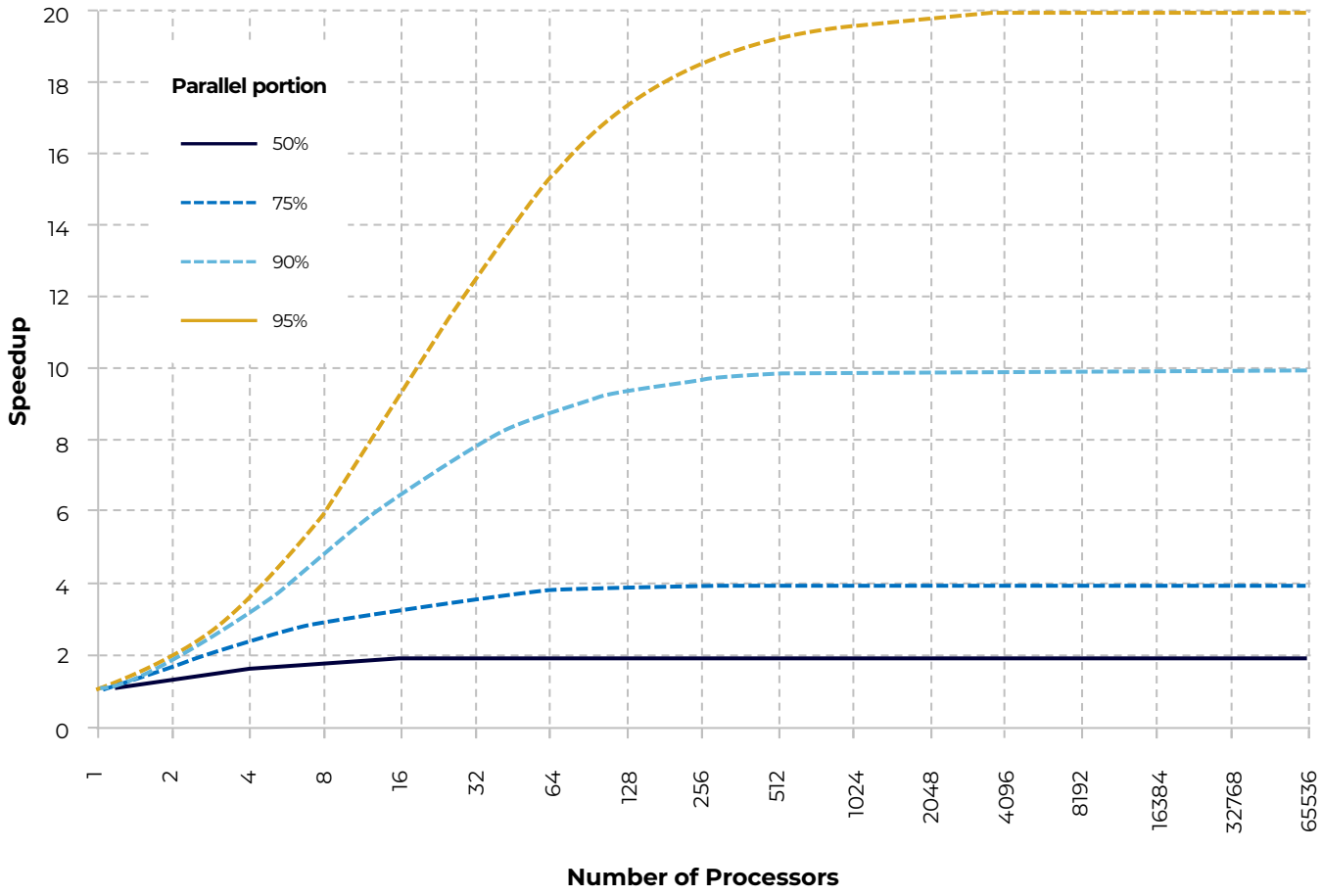
The justification for this integrated approach is a principle known as Amdahl's Law, which states that the performance of a system is limited by its slowest component.

Consider a simplified example: if a GPU completes its computation in 1 millisecond but then waits 1 millisecond for data to arrive from neighboring GPUs across a slow interconnect, the total time per cycle is 2 milliseconds. Even if the GPU's computation is infinitely accelerated, the system would still take 1 millisecond per cycle, limited entirely by the interconnect. The maximum possible speedup is 2x, regardless of how much the GPU performance improves. In other words, the bottleneck dominates the outcome.

In an AI data center, potential bottlenecks are numerous, and any single one of them can idle the rest of the system, from memory bandwidth limiting how quickly data reaches the GPU, communication latency between GPUs slowing distributed training, network congestion between racks idling entire clusters, to power and cooling constraints capping capacity.

This is the reason why NVDA has structured itself around what it calls "extreme co-design": the simultaneous optimization of every component so that no single layer constrains the others.

Figure 5 – Amdahl's Law: system speedup vs. number of processors by parallelizable share



Source: Safra.

Note: Adding processors leads to diminishing returns as speedup asymptotes toward a ceiling set by the non-parallelizable fraction of the workload. If 50% of the work can be parallelized, maximum speedup is 2x regardless of processor count; at 95%, the ceiling rises to 20x.

The organization of the company itself mirrors this philosophy — CEO Jensen Huang maintains direct reports spanning every engineering domain and holds no one-on-one meetings, a structure designed to ensure that decisions reflect full-system optimization rather than siloed thinking.

The four-layer moat: why NVDIA's competitive position is self-reinforcing

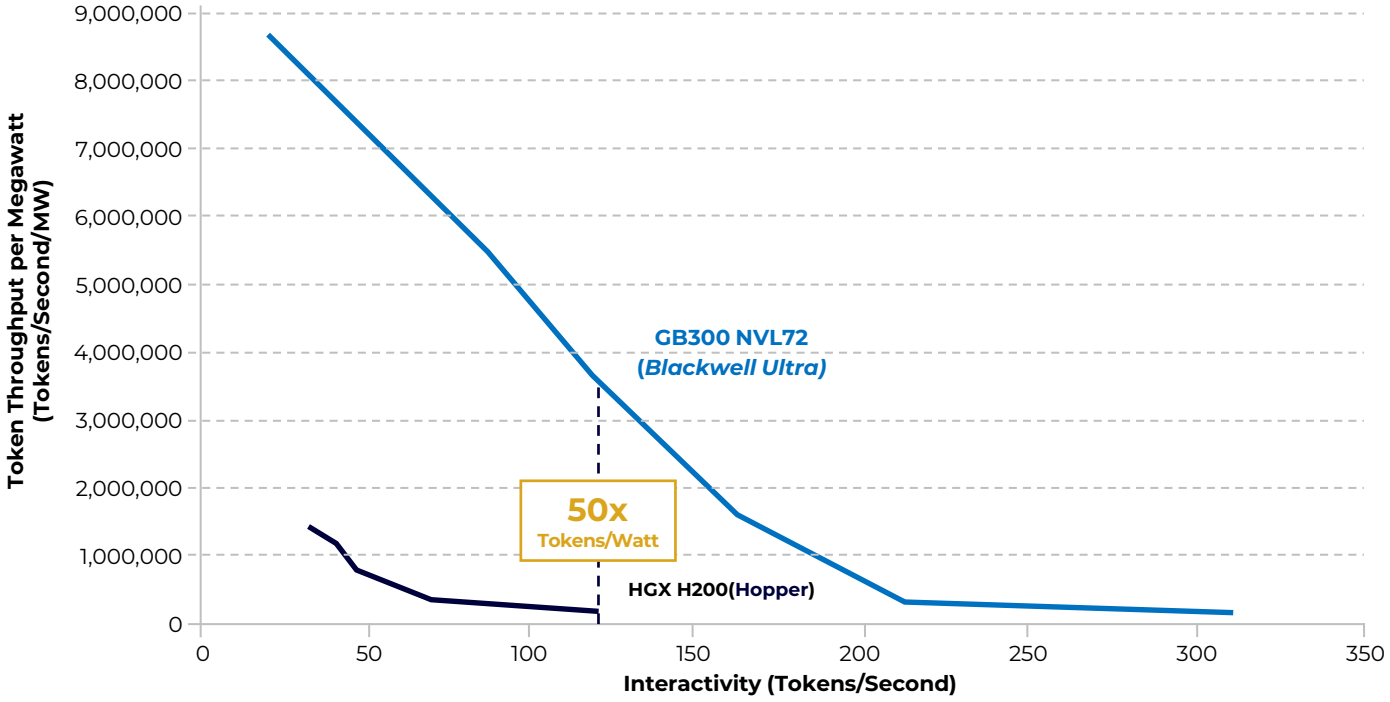
With this foundation established, we turn to the durability of NVDA's competitive position. In our view, the company's competitive position rests on four layers that together are progressively reducing the cyclicality of its business model.

The first layer is leadership in chip level performance on an annual cadence. Each GPU architecture delivers step-function gains over its predecessor on AI workloads, and the metric that best captures this progress is **throughput per watt**, defined as the number of tokens a system produces per second for each watt of power consumed.

A **token** is a sub-word unit of text, typically three to four characters, that AI models use as their atomic unit of input and output. The entire AI infrastructure stack is economically organized around it: models are priced per token, operators measure cost per token, and hardware generations are benchmarked on tokens per watt. Throughput per watt has become the binding metric for infrastructure economics — as it rises, cost per token falls.

To illustrate, the **GB300 NVL72** (Blackwell Ultra), NVDA's current flagship rack system, delivers up to ~50x more throughput per megawatt than the **HGX H200** (Hopper) on inference workloads, per SemiAnalysis InferenceMAX benchmarks. **The efficiency gains are cumulative and accelerating: the energy required to generate a single token has fallen by ~100,000x over the past decade**.

Figure 6 – Inference throughput per megawatt: GB300 NVL72 (Blackwell Ultra) vs. HGX H200 (Hopper) – DeepSeek R1 benchmark



Source: Safra, NVIDIA.

Note: The chart plots aggregate token throughput per megawatt (Y-axis) against interactivity, measured as tokens delivered per second to each individual user (X-axis). Higher interactivity corresponds to faster, real-time responses. As interactivity increases, aggregate throughput declines because the system dedicates more resources to each individual request rather than processing them in batch. Relative to HGX H200 (Hopper), the GB300 NVL72 (Blackwell Ultra) delivers up to ~50x more throughput per megawatt at comparable interactivity levels. Notably, the GB300 also serves interactivity levels that are economically unviable on Hopper, enabling a new class of AI workloads

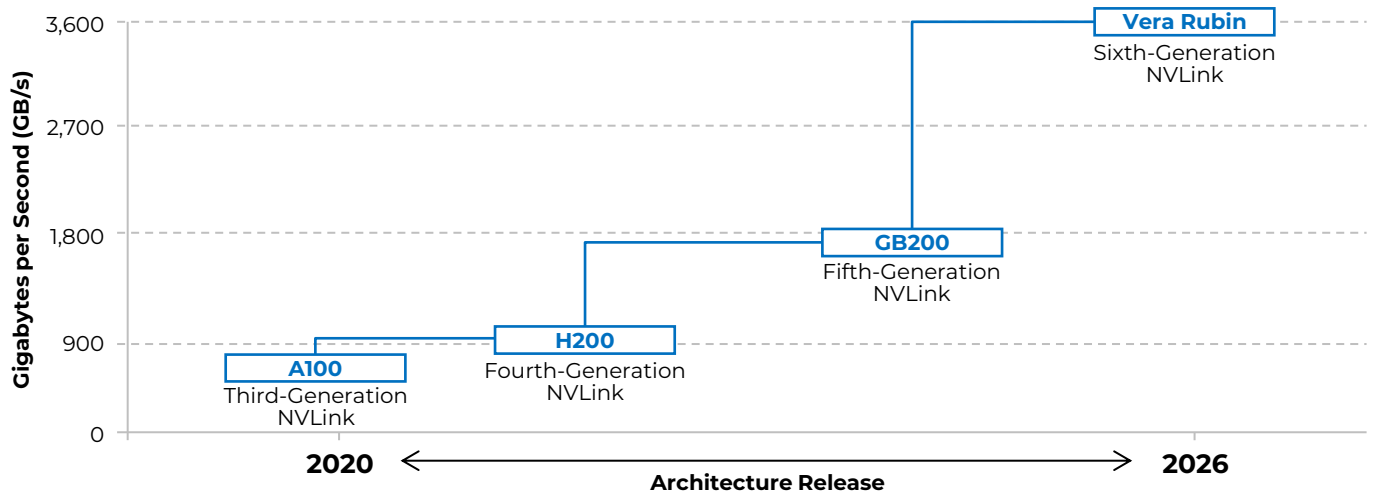
NVDA's annual architecture cadence compounds its competitive advantage. A would-be challenger must match not only the current generation but the next one, which will typically reach volume production before the challenger's first competitive alternative has.

The second layer of the moat is proprietary end-to-end networking. Modern AI workloads run across two communication layers: **scale-up**, which ties GPUs together into a single tightly coupled compute system whose size has grown from a single server to multiple racks, and **scale-out**, which connects those systems across the data center into clusters of tens of thousands of GPUs. NVDA is the only merchant vendor that owns both layers with proprietary, co-designed technology: **NVLink** and **NVSwitch** for scale-up, and **Quantum** (InfiniBand) and **Spectrum-X** (Ethernet) switches for scale-out.

Scale-up creates a single logical compute unit, and it has expanded generation by generation. NVLink is the high-speed interconnect between GPUs, and NVSwitch is the fabric chip that ties every link together so that all connected GPUs pool memory and exchange data simultaneously at full bandwidth, operating as one unified machine rather than as independent processors. The size of this unified pool directly shapes how efficiently large AI models can be trained and served: interconnect bandwidth translates into workload performance, and the larger the pool, the greater the fraction of a model that can be distributed within the unified system without incurring latency penalties.

With Hopper, NVSwitch connected 8 GPUs at 900 GB/s per GPU within a single server (HGX H100/H200). Blackwell expanded this to 72 GPUs at 1,800 GB/s across an entire rack (GB200 NVL72 and GB300 NVL2). The VR200 NVL72 (Vera Rubin), expected 2H26, maintains the 72-GPU rack-scale pool while doubling per-GPU bandwidth to 3,600 GB/s, ~10x the bandwidth of mainstream server interconnects. Rubin Ultra (2H27) will further expand the unified pool with the NVL576 Kyber architecture.

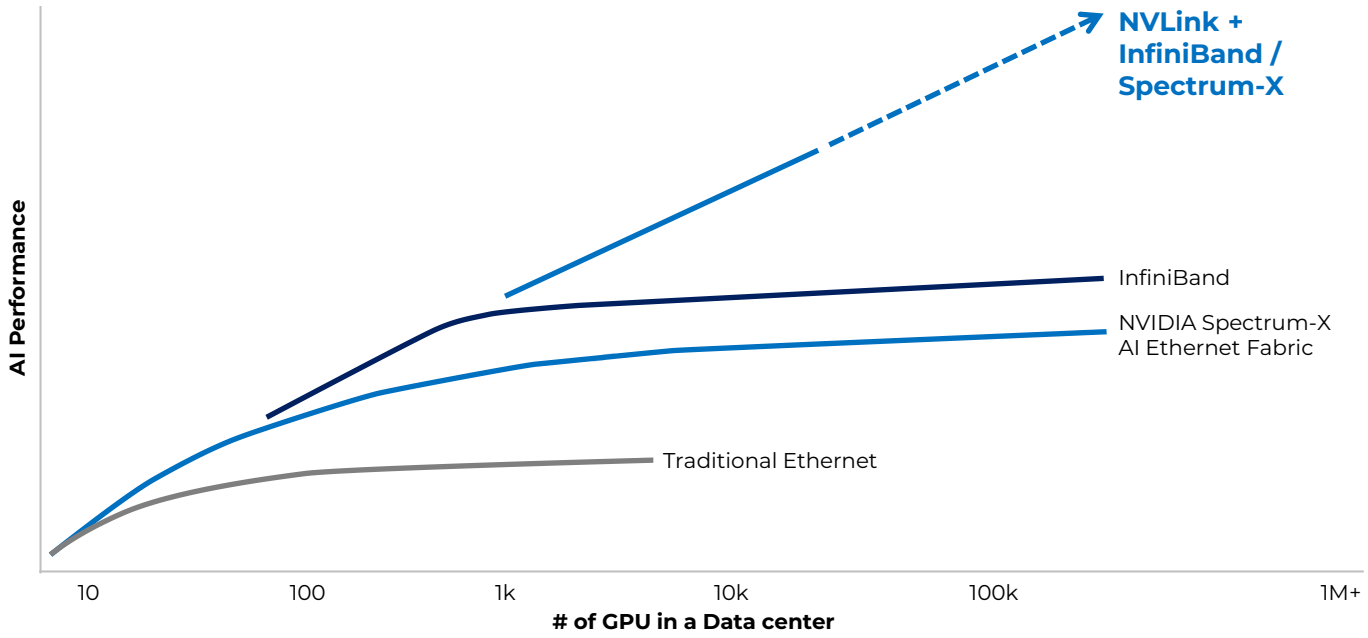
Figure 7 – NVLink bandwidth per GPU by architecture generation (GB/s)



Source: Safra, NVIDIA.

Scale-out is a data center-wide fabric, built on two industry-standard protocols. InfiniBand, originally designed for high-performance computing, is the protocol of choice for the largest training workloads, where latency is the binding constraint, while Ethernet is the standard protocol in cloud and enterprise deployments. NVDA covers both: Quantum switches for InfiniBand and Spectrum-X switches for Ethernet, alongside the network adapters inside each server (ConnectX and BlueField) and the software that coordinates communication between thousands of GPUs during training and inference. The acquisition of **Mellanox Technologies** in 2020 for ~USD 7bn gave NVDA ownership of this fabric, and networking revenue has since grown from ~USD 1.3bn at the time of the deal to ~USD 31bn in FY'26A.

Figure 8 – AI as a data center-scale workload: scale-up + scale-out networking



Source: Safra, NVIDIA.

Taken together, scale-up and scale-out form a networking stack that no merchant competitor currently matches end to end. A would-be challenger may build a competitive chip in isolation, but without comparable networking that chip cannot scale to the demands of frontier AI workloads.

The third layer of the moat is privileged access to scarce manufacturing capacity. NVDA's chips depend on constrained inputs, each sourced from a narrow set of suppliers: **leading-edge logic fabrication**, **CoWoS** (Chip-on-Wafer-on-Substrate) advanced packaging, and **HBM** (high-bandwidth memory).

Leading-edge logic refers to the most advanced semiconductor manufacturing processes, measured in **nanometers** (nm). The smaller the number, the more transistors per unit of area, and the higher the performance and energy

efficiency, at scales approaching the limits of physics. NVDA's current Blackwell architecture is fabricated on **TSMC's** custom 4NP process (a custom 4nm-class node), and the next-generation Rubin moves to TSMC's 3nm. TSMC (Not Covered) is the only foundry in the world capable of manufacturing at these nodes at scale, and leading-edge capacity is running near fully utilized, with 3nm allocation reportedly committed through 2027.

CoWoS is the advanced packaging technology that integrates the GPU and HBM stacks onto a single silicon interposer, a thin layer of silicon that acts as a high-speed wiring board, connecting the processor and its memory into a unified package. This physical proximity is what enables the memory bandwidth that modern AI accelerators require. CoWoS is the binding supply bottleneck for AI accelerators, as the industry estimates that demand exceeds supply by ~40%–50%, with lead times exceeding 40 weeks.

HBM is a specialized form of **DRAM** (dynamic random-access memory, the working memory that temporarily stores data while a processor operates on it) that stacks multiple memory layers vertically to deliver far higher data throughput at lower energy cost than standard memory. It is now the single largest component of each GPU's manufacturing cost. HBM is produced by an oligopoly of three suppliers: **SK Hynix** (Not Covered) and **Samsung** (Not Covered), both based in South Korea, and US-based **Micron** (Not Covered). All three are fully allocated. NVDA has secured multi-year supply agreements that lock in priority access, and we estimate the company consumes ~50%+ of global HBM output.

Figure 9 – Estimated leading-edge logic foundry market share

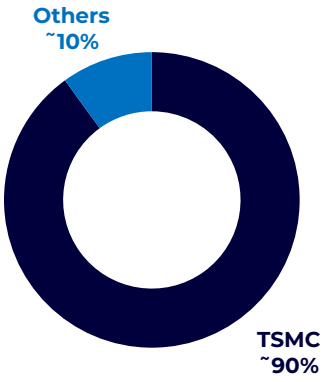
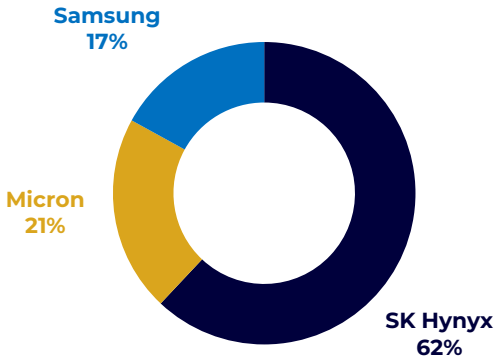


Figure 10 – Estimated high-bandwidth memory market share



Source: Safra.

NVDA is among the largest customers of each of these suppliers, and its ability to secure VIP access across all inputs simultaneously is a competitive advantage that compounds with scale.

The fourth layer of the moat is the CUDA software ecosystem. Released in 2006, CUDA made GPUs programmable for general-purpose computation using familiar languages (C, C++, and Python), replacing the graphics-API approach that GPU computing had previously required and opening the architecture to a far broader developer base.

The platform has accumulated over 7.5 million developers over two decades, thousands of optimized libraries, and native integration with every major AI framework, including PyTorch, TensorFlow, and JAX, on which virtually all AI development occurs. Backward compatibility across GPU generations ensures that code written for older hardware continues to run on current architecture without modification. Each generation of CUDA-native software therefore becomes an incremental layer of switching cost.

The switching costs are not only technical but operational. Migrating away from CUDA requires performance optimizations to be rebuilt from scratch, production systems to undergo months of revalidation, and engineering teams to be retrained on an unfamiliar stack. The academic literature, production codebases, and engineering hiring pipelines of the AI industry all assume CUDA as baseline. Alternatives exist (**AMD's** (Not Covered) ROCm, **Intel's** (Not Covered) oneAPI, and open-source standards such as **OpenCL** and **SYCL**), but none approaches CUDA's breadth of library support, depth of tooling, or installed base of trained engineers.

Figure 11 – The CUDA flywheel

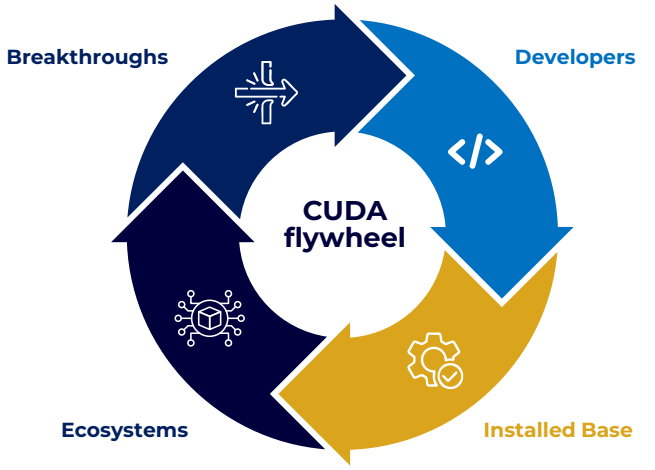
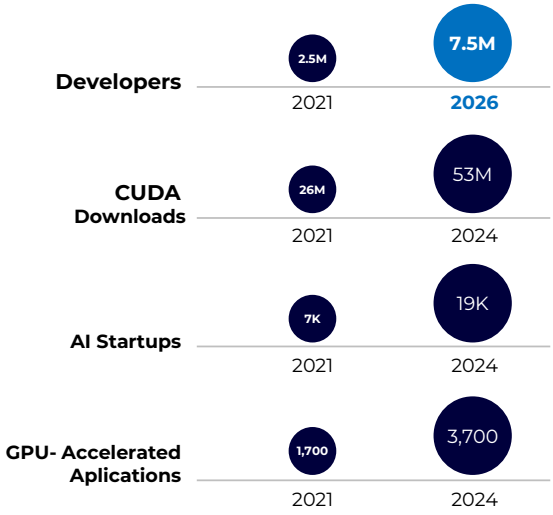


Figure 12 – The CUDA ecosystem



Source: Safra, NVIDIA.

Between the low-level CUDA programming model and the user-facing AI frameworks sits **CUDA-X**, a collection of more than 400 domain-specific acceleration libraries that translate raw GPU compute into optimized performance for deep learning and data science. The ecosystem's accumulated switching costs (technical, operational, and institutional) form the most durable layer of the moat precisely because they are the hardest to replicate through capital investments alone.

While CUDA itself generates no direct revenue, NVDA has begun monetizing the ecosystem commercially through **NVIDIA AI Enterprise**, a software platform sold as a per-GPU annual subscription (~USD 4,500/GPU/year) that bundles CUDA-X acceleration libraries, **NVIDIA Inference Microservices** (NIM, which contains pre-trained models packaged for immediate deployment), optimized AI frameworks, and enterprise support.

A would-be challenger would need to match all four layers at once. Matching one, even two or three, is not enough. A competitor faces a moving target: by the time its product reaches volume production, NVDA's next generation will have already shifted the performance baseline upwards. Without comparable networking, competitors cannot scale beyond a single rack. Without priority allocation at TSMC and across HBM suppliers, competitors cannot ship at volume. Without CUDA, competitors cannot tap the developer flywheel.

We also believe the moat is reinforced by how NVDA responds to competitive threats. The pattern resembles **Microsoft's** (Not Covered) historical playbook: when a standalone competitor gains traction in an adjacent category, Microsoft has neutralized the threat by integrating a competing product into its existing platform and leveraging its installed base so the standalone product cannot survive against the bundle. Internet Explorer displaced Netscape by shipping inside Windows, Teams eroded Slack's enterprise position by bundling into Microsoft 365, and OneDrive marginalized Dropbox through the same mechanism.

While NVDA does not "bundle" (customers can purchase compute and networking separately), the strategy is similar — identify a potential adjacent threat, absorb it into the platform, and make the integrated offering harder to replicate than the standalone offering.

The moat, ultimately, is not any single layer but the co-design across all of them: silicon, interconnect, networking, and software function as an integrated system whose performance exceeds the sum of its parts.

II. The AI factory framework and a primer on tokenomics

The AI factory and its building blocks

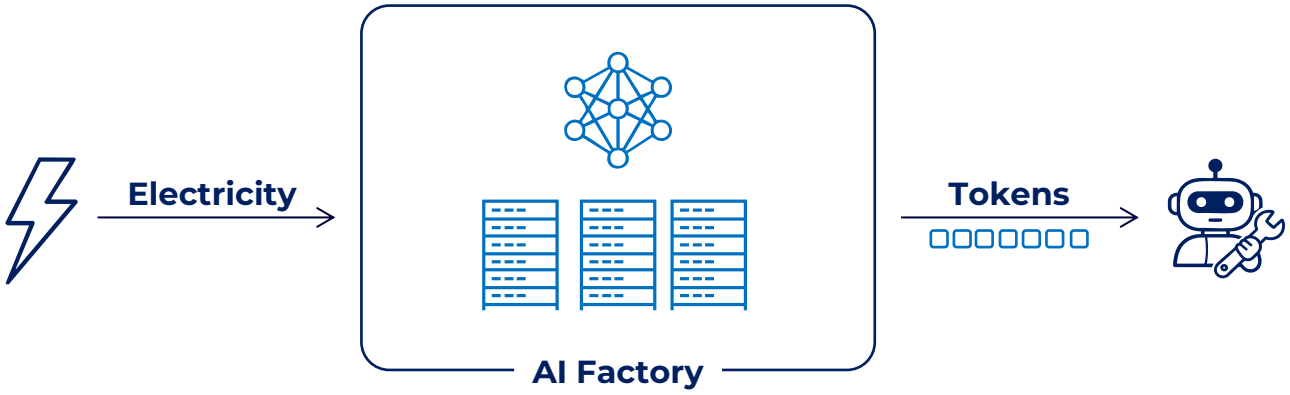
The cost structure of an AI data center creates a powerful upgrade incentive that translates NVDA's annual architecture cadence into what we view as structurally recurring demand. To understand why, it helps to start with what a modern AI data center does and how it makes money.

For decades, computing was predominantly retrieval-based: a user submitted a query, the system searched a database, and it returned a stored result. The computational cost per interaction was minimal, a lookup rather than a creation. AI has inverted this model. When a user prompts an LLM, the system does not retrieve a stored answer but instead generates one, performing billions of mathematical operations across the model's parameters to produce each token of output.

Management frames this through its "AI factory" framework: just as a traditional factory converts raw materials into finished goods, an AI data center converts electricity into tokens through computation. Operators monetize AI factories by selling tokens, typically through cloud API access, managed inference services, or proprietary applications. Token pricing varies by model size and delivery speed, but the revenue model is the same across operator types: volume of tokens produced, multiplied by the price per token. The AI factory's economic return is therefore determined by how efficiently it converts energy into useful, monetizable output.

Once an AI factory is operational, the dominant costs are largely fixed and sunk — land, contracted grid capacity, shell construction, and cooling infrastructure. The operator's return on those assets is driven primarily by one variable, the throughput of the installed GPUs. Any hardware upgrade that increases tokens per watt within the same power envelope directly improves the facility's ROIC, creating a powerful economic incentive to upgrade with each architecture generation.

Figure 13 – The AI factory framework: electricity in, tokens out



Source: Safra.

AI factory operators span several categories — hyperscalers, neoclouds renting pure compute capacity, and AI labs building proprietary infrastructure — but the underlying economics are the same across all of them.

The building block of the AI factory is the rack-scale system. Today, that means the NVL72: 36 superchips (integrated modules pairing 2 GPUs with 1 CPU) for a total of 72 GPUs, 36 CPUs, and ~1.3+ million co-designed components. Each GPU package contains 2 physical GPU dies (the individual chips cut from a fabricated wafer) connected by a high-speed bridge, bringing the total to 144 GPU dies per rack. It is the GPU die that consumes TSMC's scarcest capacity: leading-edge wafer output and CoWoS packaging.

There are currently 3 generations of the NVL72 either shipping or in development. Each generation shares the same rack architecture but differs in chip design, memory configuration, and power consumption. While NVDA does not disclose list prices for its rack systems, we estimate pricing based on industry data:

- **GB200 NVL72 (Blackwell)**, shipping since early 2025, is the current baseline at ~120 kW per rack. We estimate NVDA's compute revenue at ~USD 3.0mn per rack and networking revenue at ~USD 0.6mn, for a total of ~USD 3.6mn;
- **GB300 NVL72 (Blackwell Ultra)**, ramping up since late 2025, is a midcycle refresh built on the same Blackwell architecture, with revised silicon optimized for inference and reasoning workloads. Power consumption is ~140 kW per rack. We estimate NVDA's compute revenue at ~USD 3.8mn and networking revenue at ~USD 0.8mn, for a total of ~USD 4.6mn; and
- **VR200 NVL72 (Vera Rubin)**, expected for 2H26, is a full architectural generation change: new Rubin GPU on TSMC's 3nm node, HBM4 memory, new Vera CPU, and a new NVLink generation. Power consumption is projected at ~200 kW per rack. We estimate NVDA's compute revenue at ~USD 6.3mn and networking revenue at ~USD 1.4mn, for a total of ~USD 7.7mn, reflecting both the architectural step-up and NVDA's expanding content share per rack.

Unit economics: from power to racks to dollars

The industry sizes AI data center capacity in watts rather than floor space, a reflection of how power-dense modern AI racks have become. For reference, a 1 GW (1,000 MW or 1,000,000 kW) data center facility consumes roughly the same electricity as a mid-sized city of ~800k residents.

However, not all that power reaches the IT racks. A portion is consumed by power conversion and distribution losses (transformers, UPS systems, switchgear) and by cooling at the facility level (coolant distribution units, heat rejection systems, and pumps).

The industry measures this overhead through a metric called **Power Usage Effectiveness** (PUE), the ratio of total power entering the facility to the power consumed by IT equipment. For instance, a PUE of 1.15x means ~87% of total power reaches IT equipment, with the remainder lost to conversion and cooling overhead.

For purposes of this exercise, we assume a PUE of 1.15x across NVL72 generations, reflective of the industry's transition to liquid-cooled AI facilities with increasingly efficient power distribution. Therefore, a 1 GW data center facility delivers ~870 MW of effective IT capacity (1,000 MW / 1.15). Dividing the effective IT capacity by each rack's power consumption yields the total number of racks per GW, and multiplying the total racks count by our estimated cost per rack yields NVDA's revenue opportunity per GW of installed data center capacity.

Figure 14 – Unit economics assumptions: NVDA revenue opportunity per GW of data center installed capacity

Unit economics assumptions: NVDA revenue opportunity per GW of data center installed capacity	Unit	GB200 NVL72 (Blackwell)	GB300 NVL72 (Blackwell Ultra)	VR200 NVL72 (Vera Rubin)
(a) Data center capacity (1GW = 1mn kW)	kW	1,000,000	1,000,000	1,000,000
(b) Data center efficiency	PUE	1.15x	1.15x	1.15x
(c)= (a)/(b) Effective data center capacity	kW	869,565	869,565	869,565
(d) NVL72 rack power density	kW	120	140	200
(e) = (c)/(d) NVL72 racks per GW of DC capacity	Units	7,246	6,211	4,348
(f) NVDA compute revenue per NVL72 rack	US$mn	3.0	3.8	6.3
(g) NVDA networking revenue per NVL72 rack	US$mn	0.6	0.8	1.4
(h) = (f) + (g) NVDA revenue per NVL72 rack	**US$mn**	**3.6**	**4.6**	**7.7**
(i) = (e)*(h) NVDA revenue per GW of DC capacity	**US$mn**	**26,087**	**28,571**	**33,478**

Source: Safra.

Under our unit economics assumptions, NVDA's revenue per rack more than doubles from GB200 to VR200 (~USD 3.6mn to ~USD 7.7mn), driven by higher-performance silicon and expanding NVDA content within each rack. However, NVDA's revenue opportunity per GW grows more modestly (from ~USD 26bn to ~USD 33.5bn, or ~29% across two architecture generations). In other words, as each generation consumes more power per rack, fewer racks fit within the same fixed data center power budget.

Furthermore, we estimate that NVDA's direct hardware content accounts for ~90% of the rack's build cost. The remaining ~10% is split between 3P components (storage, cabling, rack-level ancillaries) and OEM integration markup, and that share has been declining with each generation as NVDA pulls more system-level components into its own reference design.

Beyond IT equipment (silicon and networking), the AI factory operator **total cost of ownership (TCO) per GW** includes physical infrastructure (land, grid connections, shell construction, and facility-level cooling), which is largely fixed at ~USD 15bn per GW according to our estimates.

Figure 15 – Total cost of ownership per GW of data center capacity

Total cost of ownership per GW of data center capacity	Unit	GB200 NVL72 (Blackwell)	GB300 NVL72 (Blackwell Ultra)	VR200 NVL72 (Vera Rubin)
(a) NVDA content per GW of DC capacity	US$mn	26,087	28,571	33,478
(b) Other IT equipment per GW of DC capacity	US$mn	2,899	3,175	3,720
(c) Physical infrastructure per GW of DC capacity	US$mn	15,000	15,000	15,000
(d) = (a)+(b)+(c) All-in cost	**US$mn**	**43,986**	**46,746**	**52,198**

Source: Safra.

At a more granular level, we decompose revenue per NVL72 into revenue per GPU package and per GPU die. We note that revenue per GPU package roughly doubles from GB200 to VR200 (~USD 50,000 to ~USD 107,000), and revenue per GPU die follows the same trajectory (~USD 25,000 to ~USD 53,000).

However, the gap between per-chip value growth across the generations (~2x) and per-GW revenue growth (~29%) is precisely what creates the upgrade incentive for the AI factory operator. While NVDA captures more value per chip, the AI factory operator captures even more economic surplus, considering that each architecture produces substantially more throughput per watt (meaning the cost per token falls faster than the TCO per GW rises).

Figure 16 – Unit economics assumptions: NVDA revenue per GPU package and per GPU die

Unit economics assumptions: NVDA revenue per GPU package and per GPU die	Unit	GB200 NVL72 (Blackwell)	GB300 NVL72 (Blackwell Ultra)	VR200 NVL72 (Vera Rubin)
(a) NVDA revenue per NVL72 rack	US$mn	3.6	4.6	7.7
(b) GPU packages per NVL72 rack	Units	72	72	72
(c) GPU dies per NVL72 rack	Units	144	144	144
(d) = (a)/(b) NVDA revenue per GPU package	**US$**	**50,000**	**63,889**	**106,944**
(e) = (a)/(c) NVDA revenue per GPU die	**US$**	**25,000**	**31,944**	**53,472**

Source: Safra.

The consequence is twofold. **AI factory operators have a strong incentive to deploy the newest architecture in each incremental GW of capacity. And since each generation improves the economics of every new deployed GW, they also have a strong incentive to build more capacity than they would have at the prior generation's economics**. This is what we view as a structurally recurring upgrade cycle tied to NVDA's annual architecture cadence.

A final note on management's per-GW framework. NVDA has cited a revenue opportunity of ~USD 40bn–50bn per GW of AI data center capacity. Our bottom-up estimates yield ~USD 26bn–28bn for Blackwell and Blackwell Ultra and ~USD 33.5bn for Vera Rubin, well below management's range. We read the higher management's figure as forward-looking, implicitly pricing in next-generation platforms (Rubin Ultra and Feynman) expected in the 2027–2028 timeframe, which will carry higher system ASPs and expanded NVDA content per rack.

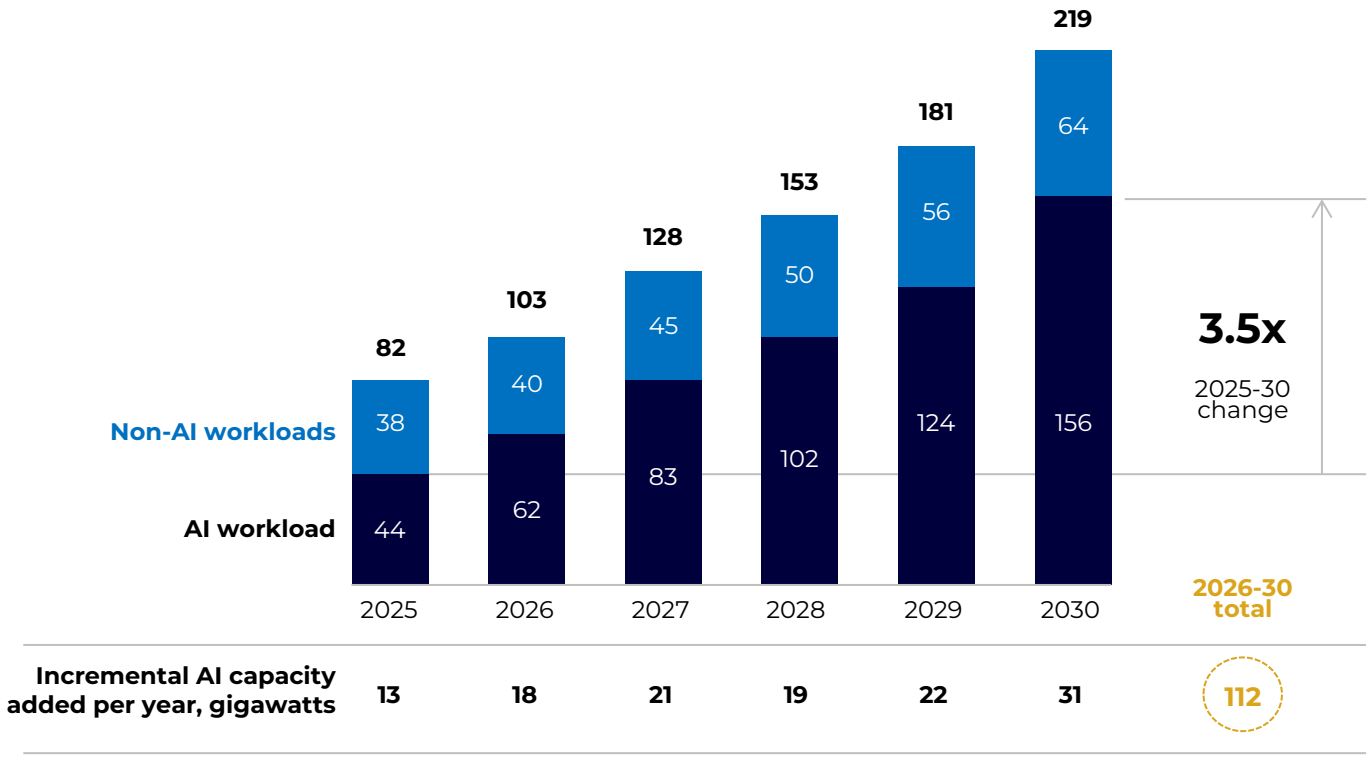
Top-down cross-check: sizing the AI infrastructure buildout

The previous analysis establishes the TCO per GW of accelerated computing capacity and how much of that spend would potentially accrue to NVDA. The next natural question is how many GWs will actually be built.

McKinsey projects the global installed base of data center capacity at ~219 GW by 2030E, up from ~82 GW in 2025, in its base case, a figure corroborated by the IEA through a bottom-up methodology (~226 GW by 2030E). Accelerated

computing accounts for ~70% of the 2030E total, or ~156 GW, implying ~112 GW of incremental accelerated computing capacity added between 2026E and 2030E.

Figure 17 – Estimated global data center capacity demand, in GW



Source: McKinsey.

We estimate NVDA's incremental addressable TAM over this period as a function of two variables: **(i)** incremental GW of accelerated computing capacity deployed, and **(ii)** NVDA's blended revenue per GW, which rises over time as the architecture mix shifts toward higher ASP generations.

Adopting McKinsey's base case of ~112 GW of incremental capacity and a blended NVDA addressable revenue of ~USD 40bn per GW (reflecting operators deploying successive generations with progressively higher NVDA content per rack) yields an **incremental TAM of ~USD 4.4trn through CY'30E**. Figure 18 provides a sensitivity analysis of the key assumptions.

Figure 18 – Estimated incremental TAM for NVDA over the next 5 years (CY'26E – CY'30E)

NVIDIA's TAM per GW of data center capacity (in USD bn)	Estimated incremental AI capacity, next 5 years (2026E-2030E)										
	60GW	70GW	80GW	90GW	100GW	110GW	120GW	130GW	140GW	150GW	160GW
25.0	1,500	1,750	2,000	2,250	2,500	2,750	3,000	3,250	3,500	3,750	4,000
27.5	1,650	1,925	2,200	2,475	2,750	3,025	3,300	3,575	3,850	4,125	4,400
30.0	1,800	2,100	2,400	2,700	3,000	3,300	3,600	3,900	4,200	4,500	4,800
32.5	1,950	2,275	2,600	2,925	3,250	3,575	3,900	4,225	4,550	4,875	5,200
35.0	2,100	2,450	2,800	3,150	3,500	3,850	4,200	4,550	4,900	5,250	5,600
37.5	2,250	2,625	3,000	3,375	3,750	4,125	4,500	4,875	5,250	5,625	6,000
40.0	2,400	2,800	3,200	3,600	4,000	4,400	4,800	5,200	5,600	6,000	6,400
42.5	2,550	2,975	3,400	3,825	4,250	4,675	5,100	5,525	5,950	6,375	6,800
45.0	2,700	3,150	3,600	4,050	4,500	4,950	5,400	5,850	6,300	6,750	7,200
47.5	2,850	3,325	3,800	4,275	4,750	5,225	5,700	6,175	6,650	7,125	7,600
50.0	3,000	3,500	4,000	4,500	5,000	5,500	6,000	6,500	7,000	7,500	8,000
52.5	3,150	3,675	4,200	4,725	5,250	5,775	6,300	6,825	7,350	7,875	8,400
55.0	3,300	3,850	4,400	4,950	5,500	6,050	6,600	7,150	7,700	8,250	8,800

Source: Safra.

Management has framed the AI infrastructure opportunity at USD 3trn – USD 4trn over this period. Our base case lands at the upper end of that range and we view the range itself as feasible, particularly as sovereign AI initiatives introduce a layer of price-inelastic, strategically motivated demand.

We also note that this framework captures only net new accelerated computing capacity. The separate opportunity from migrating the installed base of legacy x86 CPU workloads to accelerated computing, a large and largely untapped market, is referenced later in the report.

Tokenomics: why not all tokens are created equal

The unit economics above express the AI factory's return in tokens per watt. Tokens, however, are not a homogeneous commodity, and the AI factory operator's revenue opportunity depends not only on volume but on the value of each token produced.

Token pricing is determined primarily by two variables: the architecture of the model generating the token (a frontier reasoning model versus a small chat model) and the interactivity level at which it is delivered (the number of tokens streamed per second to each user). AI factory operators monetize this heterogeneity through a spectrum of service tiers, from low-cost batch processing for non-interactive workloads at one end, to premium tiers at the other, where frontier reasoning models serve latency-sensitive agentic applications at substantially higher prices per million tokens.

Larger models consume more compute per token and therefore reduce aggregate throughput within a fixed power budget, but the pricing premium they command more than offsets the higher cost to produce them. What matters for the operator is therefore not aggregate token throughput but the ability to produce high-tier tokens per watt.

Each successive NVDA architecture unlocks model sizes and interactivity levels that prior generations cannot serve at viable unit economics, expanding the AI factory operator's addressable revenue within the same fixed power envelope.

III. Scaling laws and inference as a new growth vector: the agentic AI inflection

Why compute demand keeps accelerating

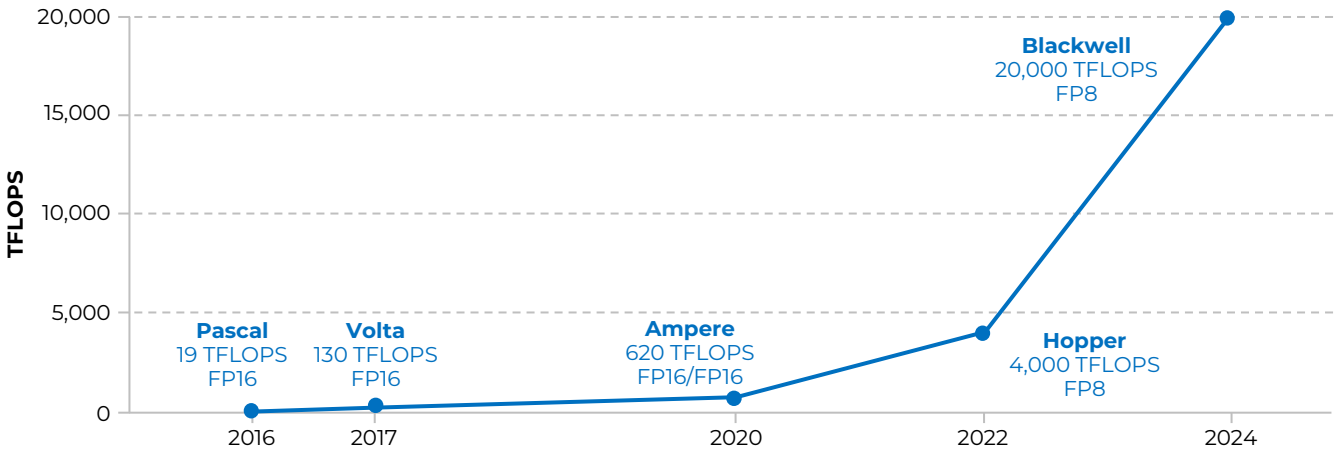
For over five decades, the semiconductor industry operated under **Moore's Law**: transistor density doubled roughly every two years, enabling predictable performance gains that shaped the entire investment cycle. A companion principle known as **Dennard scaling** reinforced the trend: shrinking transistors consumed proportionally less power, allowing performance to scale at constant power density and making each new chip generation simultaneously faster and more energy efficient.

Dennard scaling ceased to hold in the mid-2000s: below ~65 nanometers, voltage could no longer fall in proportion to transistor size, and performance-per-watt gains slowed sharply. Moore’s Law itself continued, but began to slow materially in the mid-2010s as nodes began approaching atomic scale and the limits of physics. The compounding engine that had governed the semiconductors industry for fifty years was gone.

GPU-accelerated AI compute filled the void, but with a different compounding mechanism. GPU performance has improved by ~1,000x over the past decade, a trajectory that was popularly termed “**Huang's Law**”.

Where Moore's Law depended on a single input (transistor density), Huang's Law is steeper because it stacks multiple independent improvement curves: chip architecture, lower-precision arithmetic (successive GPU generations process AI operations in fewer bits per calculation, multiplying throughput), software optimization, and memory bandwidth. Each of these levers has its own improvement trajectory, and their compounding is what drives step-change generational gains.

Figure 19 – Huang's Law: GPU performance has improved ~1,000x in the past decade, as measured by TFLOPs (trillions of floating-point operations per second)



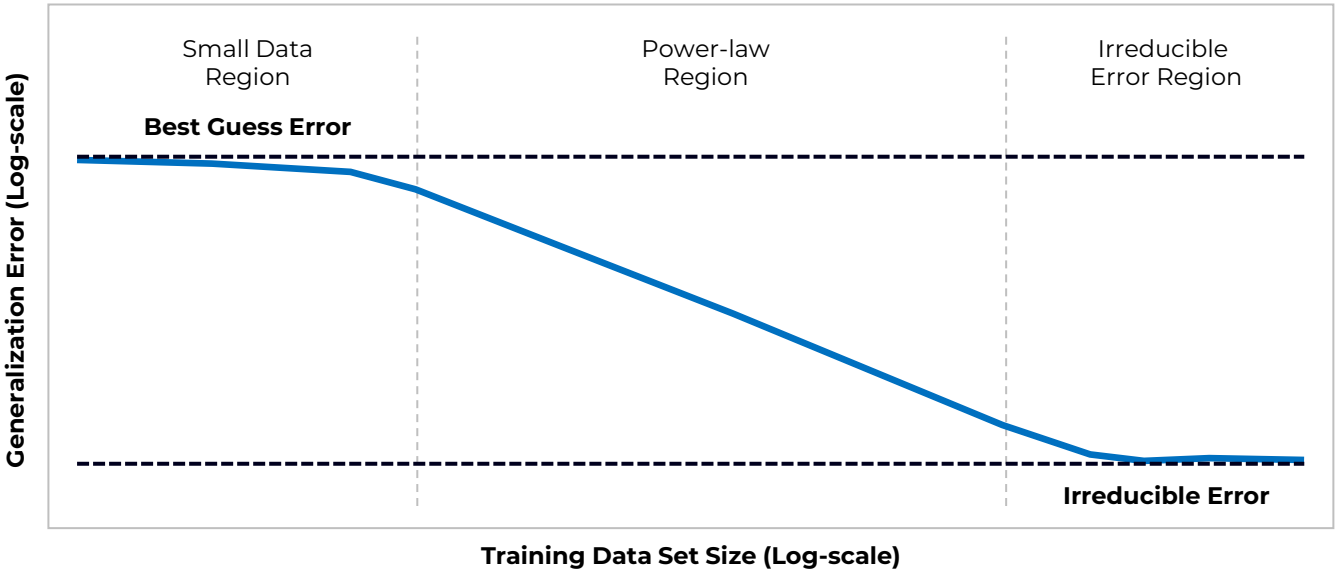
Source: Safra, NVIDIA.

If Huang's Law explains why NVDA is able to supply more compute at declining per token cost, "Scaling Laws" explain why the market keeps demanding it.

Scaling Laws describe the empirical finding that AI models improve predictably when given more compute. As the compute budget used to train a model increases, its error rate decreases in a log-linear pattern, with percentage increases in compute yielding roughly constant percentage decreases in error.

This relationship has held from the earliest LLMs through the most capable systems in production today. Put simply, more compute reliably produces better models, which has justified every successive wave of AI buildout investments.

Figure 20 – Canonical shape of a neural scaling law



Source: Hestness et all (2017), Baidu Research.

Note: Generalization error declines predictably as training data grows, transitioning through different regimes: a small-data region where models lack sufficient signal to learn meaningful structure; a power-law region where error declines log-linearly with data; and an irreducible error region where achievable performance is capped. The bull case for continued AI buildout rests on frontier models remaining in the power-law region.

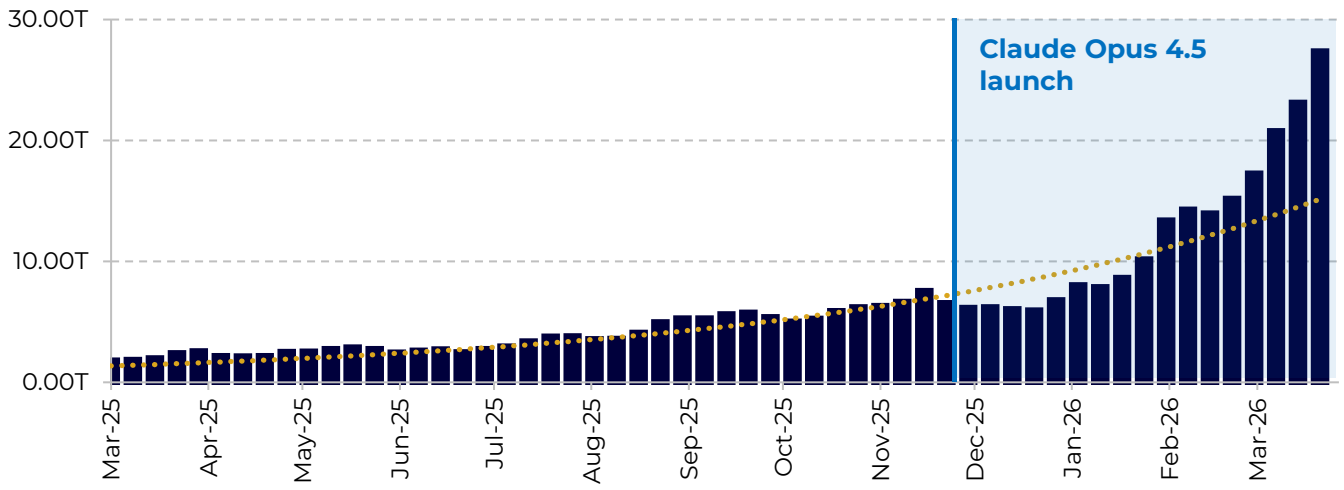
The combination of Huang's Law and Scaling Laws produces a self-reinforcing cycle: hardware gets predictably cheaper per unit of compute, models trained on that hardware get predictably better, better models unlock new applications, and new applications generate demand for next generations of hardware.

The natural objection is that efficiency gains eventually satisfy demand — if each new GPU architecture reduces cost per unit of compute, cumulative compute consumption should eventually plateau.

Historical precedent suggests the opposite. In 1865, the economist William Stanley Jevons observed that improvements in the efficiency of coal-powered steam engines did not reduce total coal consumption. Instead, the opposite occurred: efficiency improvements resulted in a substantial increase in coal consumption, as cheaper mechanical power made railways, steamships, and industrial production viable at scales that were previously out of reach.

We believe that this phenomenon, since known as the "Jevons Paradox", appears to be playing out in AI compute as well. As each new architecture generation reduces the cost per token, previously uneconomical applications become viable and total token consumption has expanded by multiples of the cost reduction. Efficiency gains are therefore expanding NVDA's TAM, not cannibalizing it.

Figure 21 – Weekly token consumption across OpenRouter



Source: OpenRouter.

Note: OpenRouter is an API aggregation platform routing developer traffic across 400+ LLMs from 60+ providers through a single interface, publishing weekly token throughput in near real time. We use it as a proxy for AI token consumption as it is model—agnostic and skews toward the developer and agentic segment driving incremental compute demand. Notably, token consumption has inflected since the November 2025 release of Anthropic's Claude Opus 4.5 and the broader shift from conversational chat to agentic AI workflows.

Where compute demand is going: the agentic AI inflection

We believe that compute demand is now compounding across several vectors. Each vector scales on its own logic, each consumes GPU compute, and each reinforces the others.

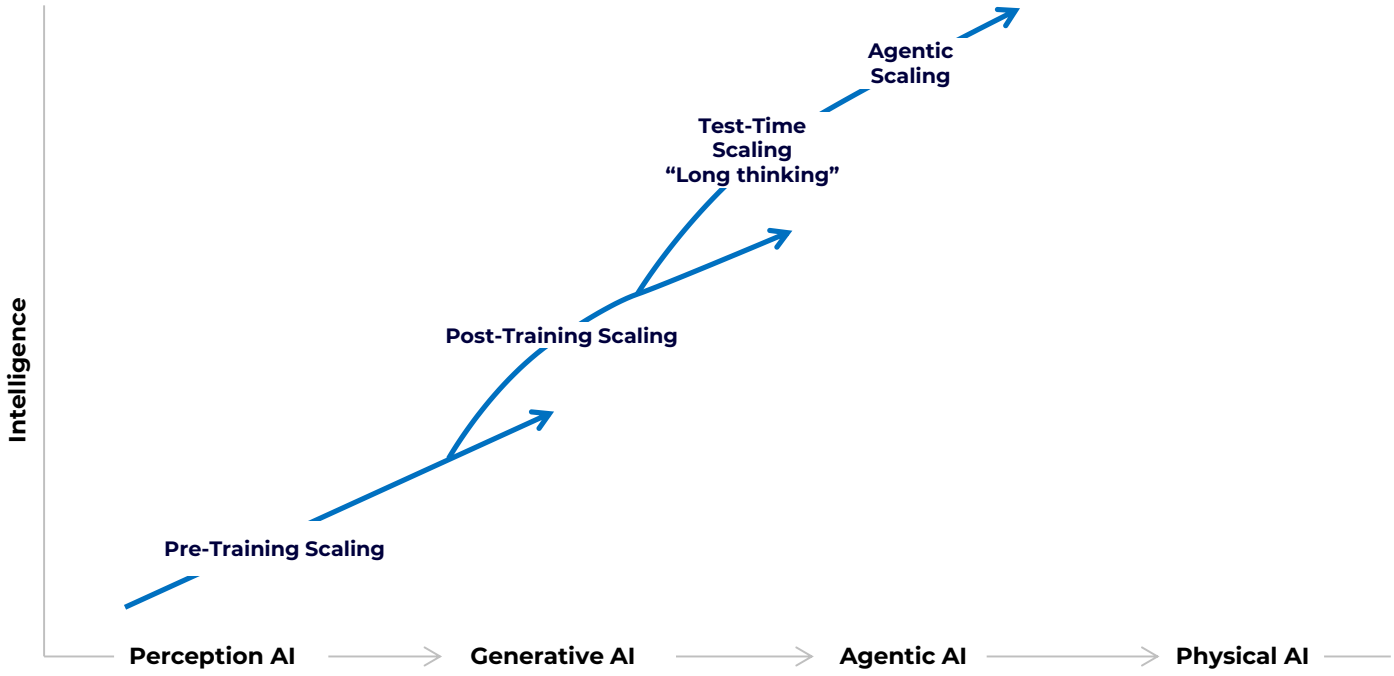
The first vector is pre-training, the large-scale training run that produces a foundation model. The system is exposed to enormous volumes of data (text, code, images) and iteratively adjusts billions of internal parameters over months of continuous GPU computation until it learns to generate useful outputs. Each frontier model requires dozens of experimental runs before a final production run, with total costs (experiments, infrastructure, personnel) running to multiples of the final run alone — the all-in cost of developing a single frontier model now reaches into the billions of dollars. Pre-training defined the first wave of AI compute demand, from the launch of ChatGPT in November 2022 through successive generations of frontier models.

The second vector is post-training, a set of techniques applied after pre-training to improve a model's capabilities on specific tasks. The most important post-training technique is **RLVR** (reinforcement learning with verifiable rewards), in which the model is trained on tasks with objectively measurable outcomes (a math problem has a correct answer, code either compiles or does not) and iteratively rewarded for correct outputs. Post-training drives GPU demand between pre-training runs, extending the utilization cycle of training clusters, and has emerged as an important source of capability gains in recent model generations.

The third vector is test-time compute, also known as inference-time scaling. Rather than improving the model during the training phase, additional compute is applied at deployment to let the model reason through harder problems. The model generates multiple chains of intermediate "thinking tokens", evaluates them, and selects the best answer, with compute demand scaling directly with task difficulty. This vector became commercially visible with the launch of OpenAI's o1 in September 2024, the first production reasoning model. Where a conventional chat model produces a few hundred tokens per response, a reasoning model can generate thousands to tens of thousands of thinking tokens for the same query.

The fourth vector is agentic scaling, which reinforces the entire loop. Agents are software systems in which a harness (code built around the model) manages an entire workflow on the user's behalf. The user provides a high-level instruction, and the harness directs the model to invoke external tools (databases, APIs, code execution environments), verify whether the output actually works, and iterate autonomously until the task is complete.

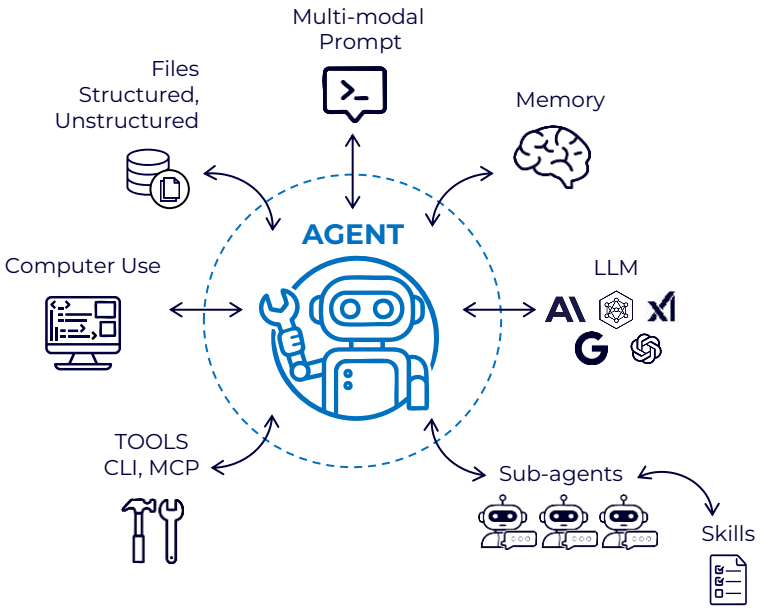
Figure 22 – Scaling laws: vectors of compute demand growth



Source: Safra.

Agents executing real-world tasks produce a stream of objectively verifiable outcomes, which is precisely the type of signal RLVR is designed to consume. Deployed agents generate outcomes, those outcomes train the next model generation, that generation is redeployed as more capable agents, and the loop restarts. Compute demand compounds at every turn: agents consume test-time compute at deployment, and the data they generate feeds ever-larger training runs.

Figure 23 – The anatomy of an agentic AI system: the model coordinates memory, tools, files, sub-agents, and other external sources to plan, execute, and verify multi-step workflows

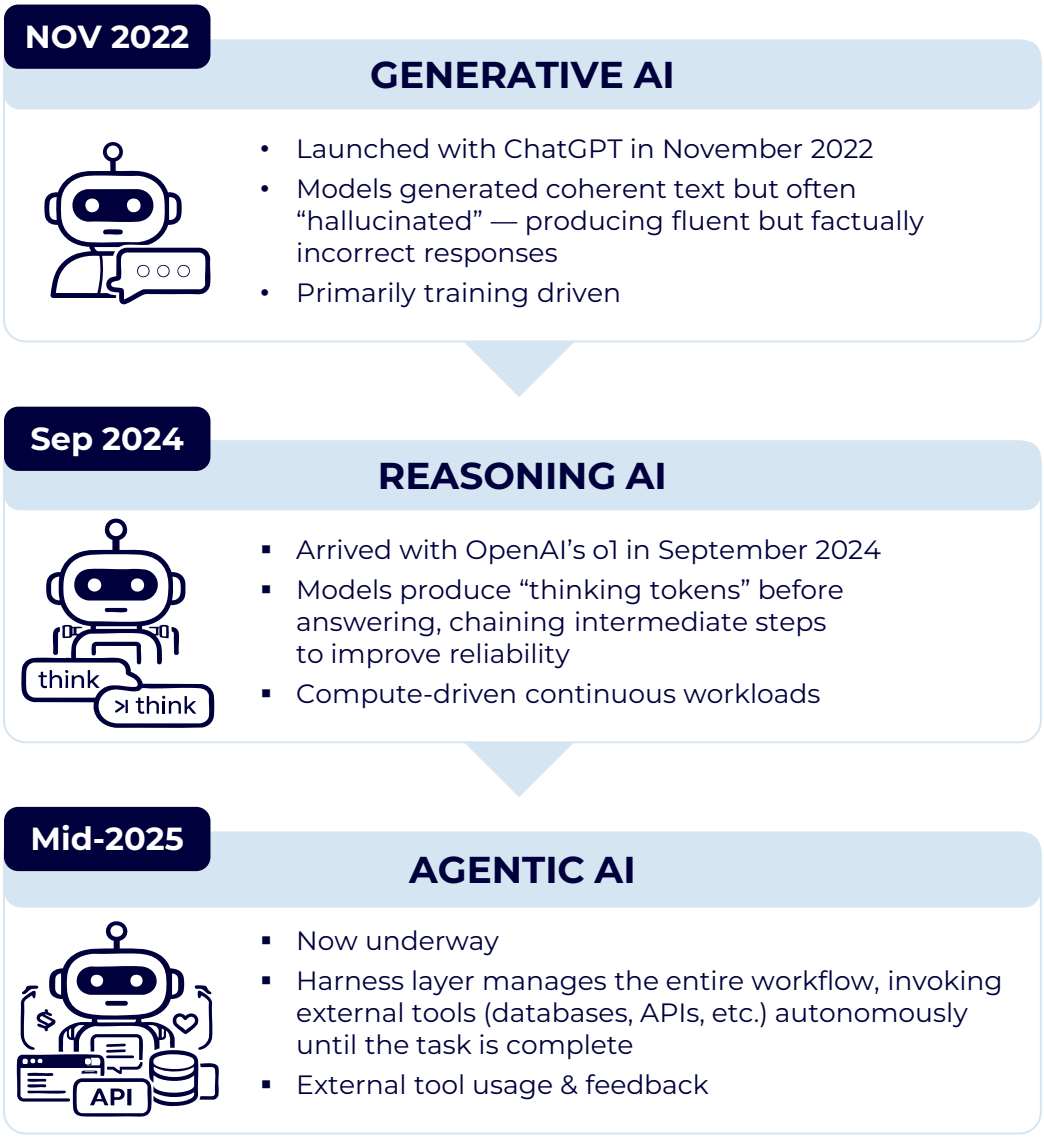


Source: Safra.

To illustrate how the vectors compound in a single interaction: in 2022, a user asking "*who is the best supplier for this component?*" would receive a text answer that might or might not be accurate. By 2024, a reasoning model would

work through pricing data, lead times, and quality metrics before recommending one. By 2026, the user instead instructs "*find the best supplier, negotiate terms, and place the order*", and an agent executes the entire workflow autonomously, calling reasoning models at each decision point, spawning sub-agents in parallel, and invoking external tools.

Figure 24 – Paradigms of AI: generative, reasoning, and agentic



Source: Safra.

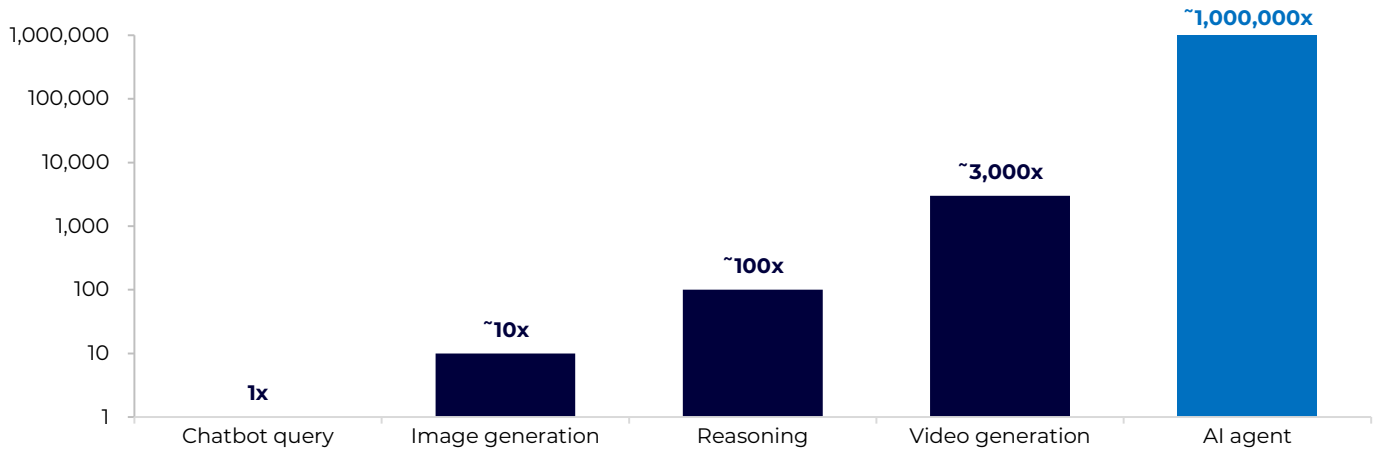
The direct consequence of this paradigm is a structural shift in the composition of AI compute demand from training towards inference.

Training, which encompasses both pre-training and post-training, is an upfront investment: a hyperscaler or frontier AI lab builds a cluster, trains a model, and reallocates capacity until the next training run. NVDA has been the principal beneficiary of this cycle since the widespread adoption of LLMs.

Inference, by contrast, is what happens after the model is built, and where the training investment is amortized. Every user prompt or agentic workflow activates the trained model, generating each token of output by processing the input through dozens to over a hundred layers of computation. Inference demand scales with every user, every application, and every query across the economy. As AI embeds into search, advertising, enterprise software, and autonomous workflows, the number of inference calls grows without natural ceiling.

A reasoning-model query consumes ~100x more inference compute than a simple generative-model query, and an agentic interaction multiplies this by another ~10,000x, since a single instruction can trigger thousands of inference calls across multiple models and tool invocations — implying a cumulative ~1,000,000x increase in compute per interaction.

Figure 25 – Token consumption relative to single chatbot query (log scale)



Source: Safra, OpenAI.

As a result, we believe inference will surpass training as the largest consumer of AI compute capacity by 2027E and reach ~60% of total AI compute by 2030E, up from ~50% today. In absolute terms, inference capacity is projected to grow from ~21 GW in 2025 to ~93 GW by 2030E (~35% CAGR), while training is expected to grow from ~23 GW to ~62 GW over the same period (~22% CAGR).

Figure 26 – Projected growth in AI training and inference, in GW

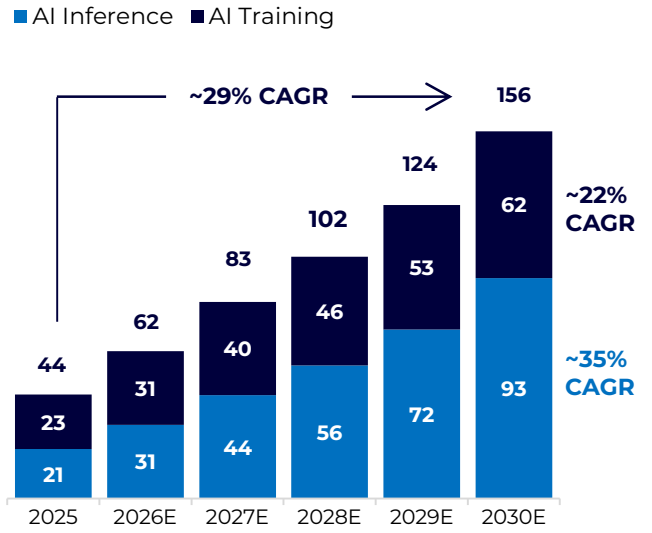
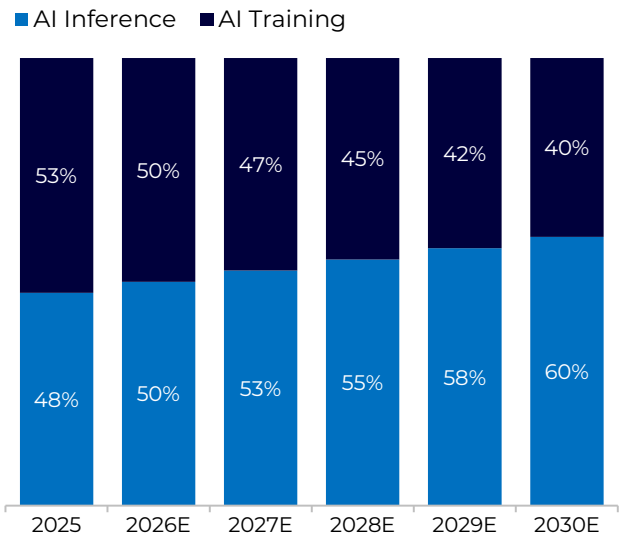


Figure 27 – Projected share of AI training and inference, in %

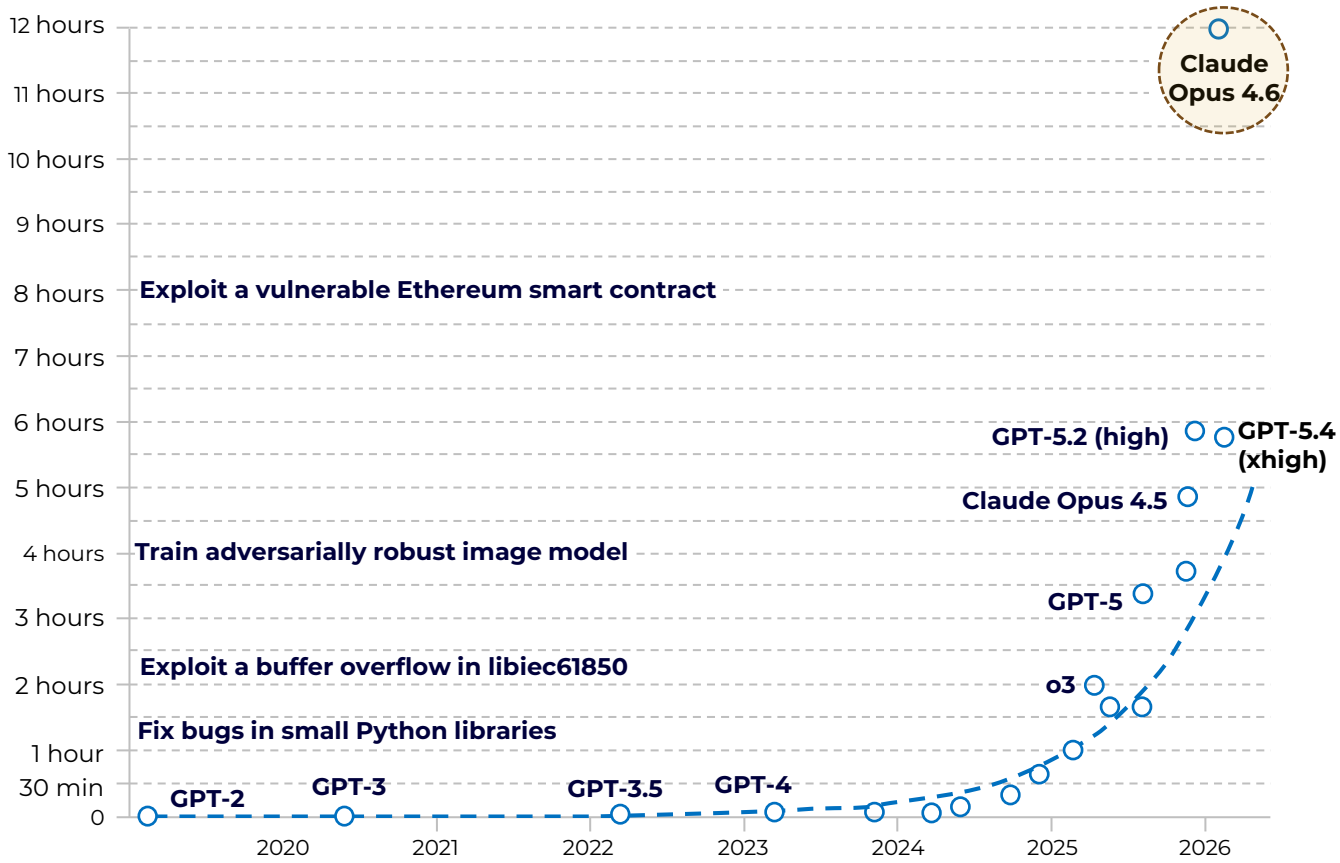


Source: Safra, McKinsey.

The agentic AI paradigm entered production deployment through 2025, as frontier models from multiple labs gained the capability to sustain autonomous work over progressively longer time horizons.

Anthropic's trajectory is illustrative: successive Opus model releases (Opus 4 in May 2025, Opus 4.5 in November 2025, and Opus 4.6 in February 2026) delivered step-function improvements in agentic capability and powered two products that drove Anthropic's ARR from ~USD 1bn to ~USD 30bn in ~15 months: Claude Code, an autonomous coding agent, and Claude Cowork, a system-level agent for non-technical users that connects to enterprise tools and executes multi-step workflows.According to the Ramp AI Index, ~31% of US businesses now pay for Anthropic products, with adoption highest among knowledge-intensive sectors such as software engineering, finance, and legal.

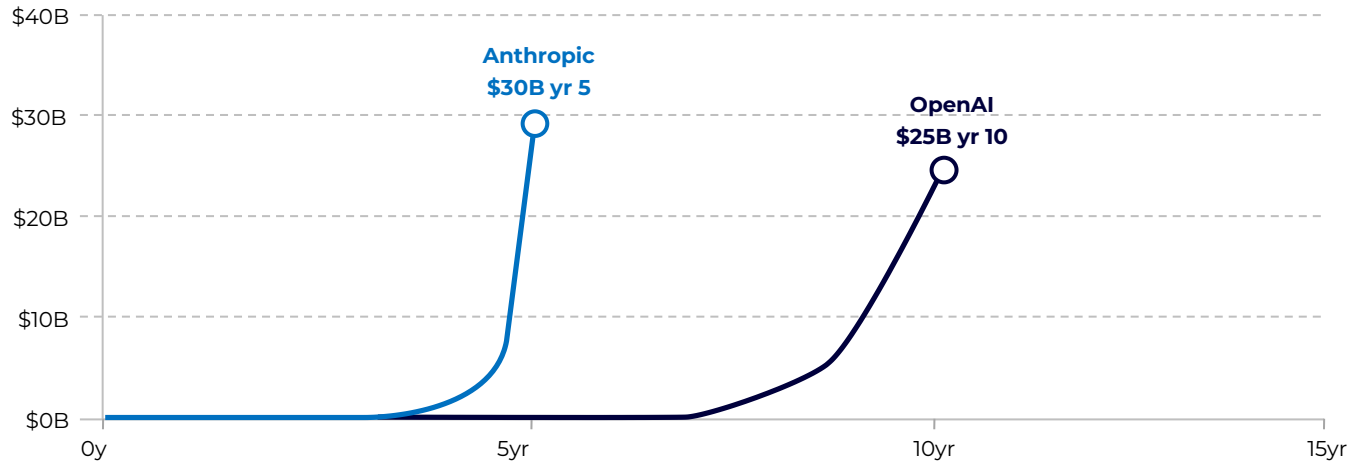
Figure 28 – METR time horizon: task length that AI models can complete with a 50% success rate, linear scale



Source: METR.

Note: METR's (Model Evaluation & Threat Research) time horizon benchmarks measure task difficulty in human-expert-time units (i.e., how long a skilled human would take to complete) and reports the threshold at which an AI succeeds on 50% of tasks. Claude Opus 4.6 currently sits at ~12 hours, meaning it completes tasks that would take a human nearly two full workdays.

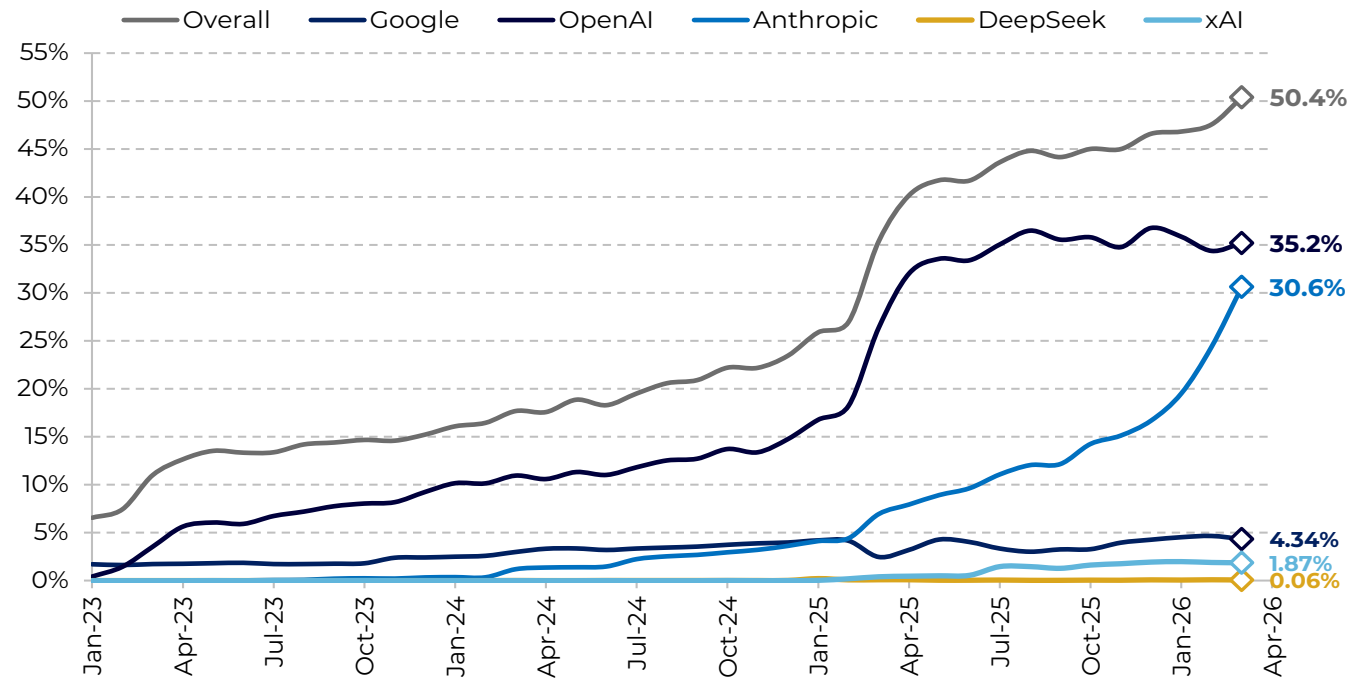
Figure 29 – Anthropic and OpenAI ARR



Source: Safra.

Note: Anthropic revenue run rate has surpassed USD 30 bn, up from USD 9 bn at the end of 2025, driven primarily by enterprise deployments of agentic AI systems across software engineering, legal, finance, and other knowledge-intensive sectors. Claude services have accelerated meaningfully YTD, with more than 1,000 enterprise customers each spending over USD 1 million annually.

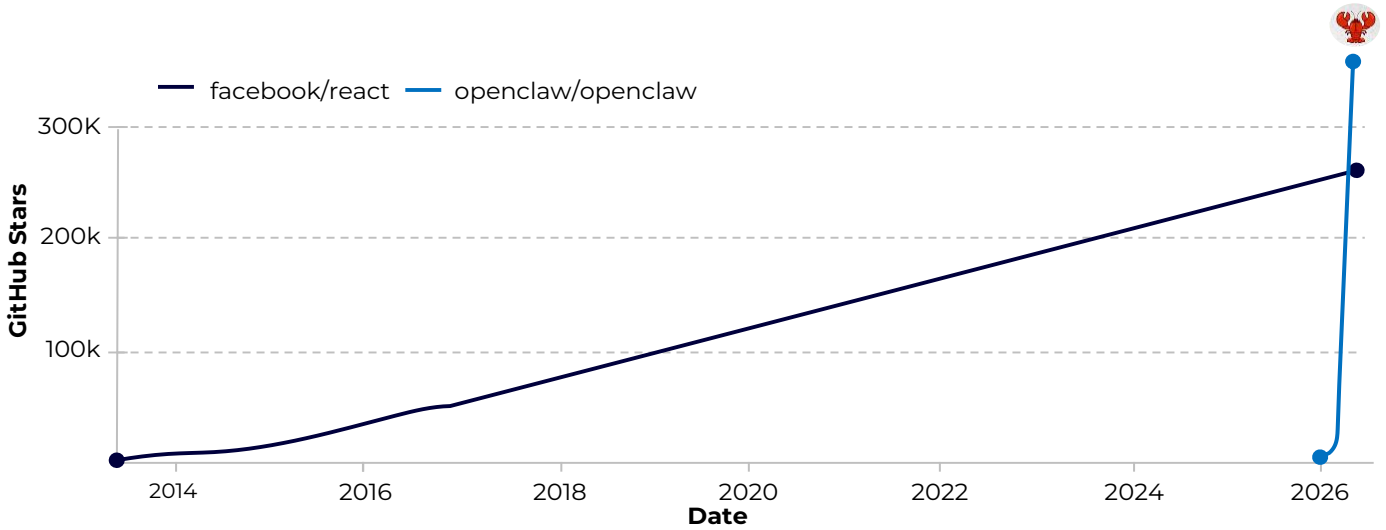
Figure 30 – Ramp AI index: share of US businesses with paid subscriptions to AI models, platforms, and tools



Source: Ramp.

A second reference point is OpenClaw, an open-source agentic framework first published in late 2025 that enables agents to connect to operating systems, messaging platforms, and applications to execute real-world tasks autonomously. OpenClaw surpassed 300,000 GitHub stars shortly after release, with NVDA management describing it as the most significant open-source release since Linux. NVDA has since launched NemoClaw, an enterprise-ready version of the OpenClaw framework that adds sandboxing, privacy, and network security layers to serve enterprise environments.

Figure 31 – OpenClaw has become a widespread phenomenon, surpassing 300,000+ GitHub stars shortly after release



Source: Star History.

The corollary is a material shift in NVDA's demand profile. Training compute is discretionary — a cluster can be reallocated across research priorities or idled between model releases without an immediate revenue consequence at the operator level. Inference compute is not: once deployed to serve live enterprise agents, consumer AI products, or autonomous coding workflows, throttling capacity translates directly into lost revenue. **Therefore, as the mix shifts from training toward inference, GPU deployments transition from more speculative training capacity to revenue-generating production assets. NVDA's demand becomes anchored to operators' P&L through inference utilization, producing a stickier, less cyclical demand profile that is harder to scale back once in place**.

The ASIC threat and NVDA's response

The continuous shift towards inference has created competitive openings for alternatives to NVDA's GPU architecture. To understand why, it helps to look at how inference actually runs on the hardware.

The inference process works in two phases, each with different hardware requirements.

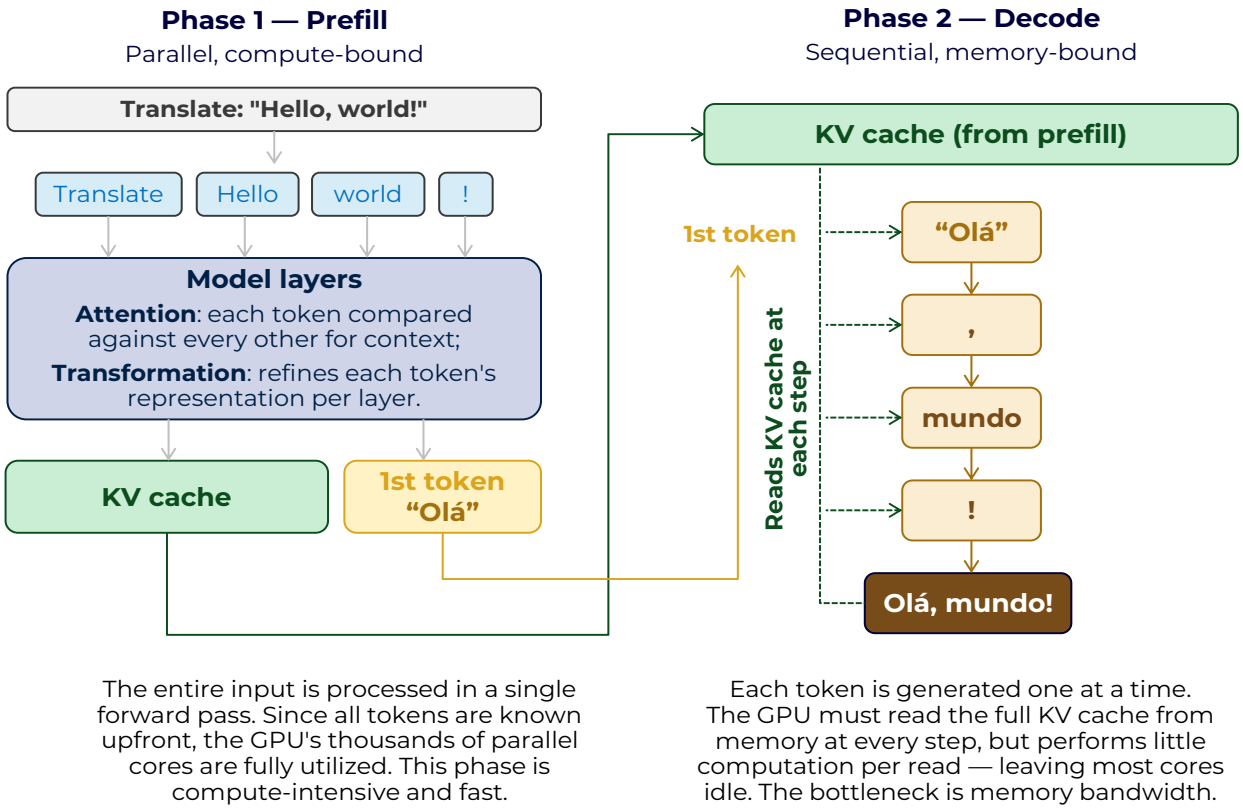
The first phase, prefill, is where the model reads the entire input prompt at once. Each token is converted into a numerical vector (a list of numbers that captures its meaning in a format the model can process) and run through dozens to over a hundred layers of computation, each performing two core operations: **attention**, where every token is compared against every other token to determine which words are most relevant to each other in context (for example, when processing "*Apple reported strong iPhone sales*", the attention mechanism determines that "*Apple*" refers to the technology company rather than the fruit by examining its relationship to "*reported*", "*iPhone*", and "*sales*"), and **transformation**, where the model refines its numerical representation of each token based on those contextual relationships.

Since the entire prompt is known upfront, all these operations run in parallel across the GPU's thousands of cores, which is why GPUs excel at prefill.

The second phase, decode, is where the model generates the rest of its response one token at a time. Each new token depends on everything that came before it (the original prompt and all previously generated tokens), so the model must read the full accumulated context from memory at every step.

Unlike prefill, this phase cannot be parallelized: each token must be generated before the next one can begin. The bottleneck shifts from raw processing power to memory bandwidth, specifically how fast the GPU can read context from its high-bandwidth memory. This is why API providers typically charge significantly more for output tokens than for input tokens: input tokens are processed efficiently in parallel during prefill, while output tokens are generated sequentially during decode, consuming far more compute time per token.

Figure 32 – The two phases of LLM inference: prefill (compute—bound) and decode (memory—bound)



Source: Safra.

Note: LLM inference runs in two phases. Prefill processes the full input prompt in one parallel pass – a workload that GPU handles efficiently, using all its cores at once. It also builds the KV cache, a memory store of intermediate context that avoids recomputing the same information at every subsequent step. Decode then generates output tokens one at a time, reading the full KV cache at each step but performing little computation per read – leaving most of the GPU's cores idle and making the phase constrained by memory rather than raw compute power. Since the KV cache grows with context length, decode becomes the dominant cost as agentic workflows consume far more tokens per request than conventional chat.

Decode is the part of inference where most time and cost are spent, and it is also where the GPU is least efficient. The GPU cores are largely idle because decode processes only one token at a time, leaving parallel capacity unused while waiting for data from memory.

This mismatch has opened space for **ASICs** (Application-Specific Integrated Circuits) designed for sequential token generation. The principal efforts are Google's (Not Covered) TPU and Meta's (Not Covered) MTIA, both co-designed with Broadcom (Not Covered), and Amazon's (Not Covered) Trainium and Inferentia, developed in-house by AWS's Annapurna Labs and by Marvell Technology (Not Covered).

In our view, however, the ASIC threat is bounded by the breadth of AI workloads. An ASIC is optimized for a narrow profile: a specific model architecture, precision format, context window, or latency target. Covering the full diversity of tasks that a GPU addresses would require a portfolio of distinct ASIC designs, compounding design and validation costs with every new workload class. A GPU, by contrast, absorbs new workloads through software updates to an existing architecture. The wider the range of AI applications grows, the stronger the economic case for a general-purpose accelerator becomes.

NVDA's response has been to extend the platform into a more heterogeneous compute system. Each successive GPU generation continues to widen the cost-per-token advantage over prior architectures, raising the bar that any alternative must clear to justify switching costs.

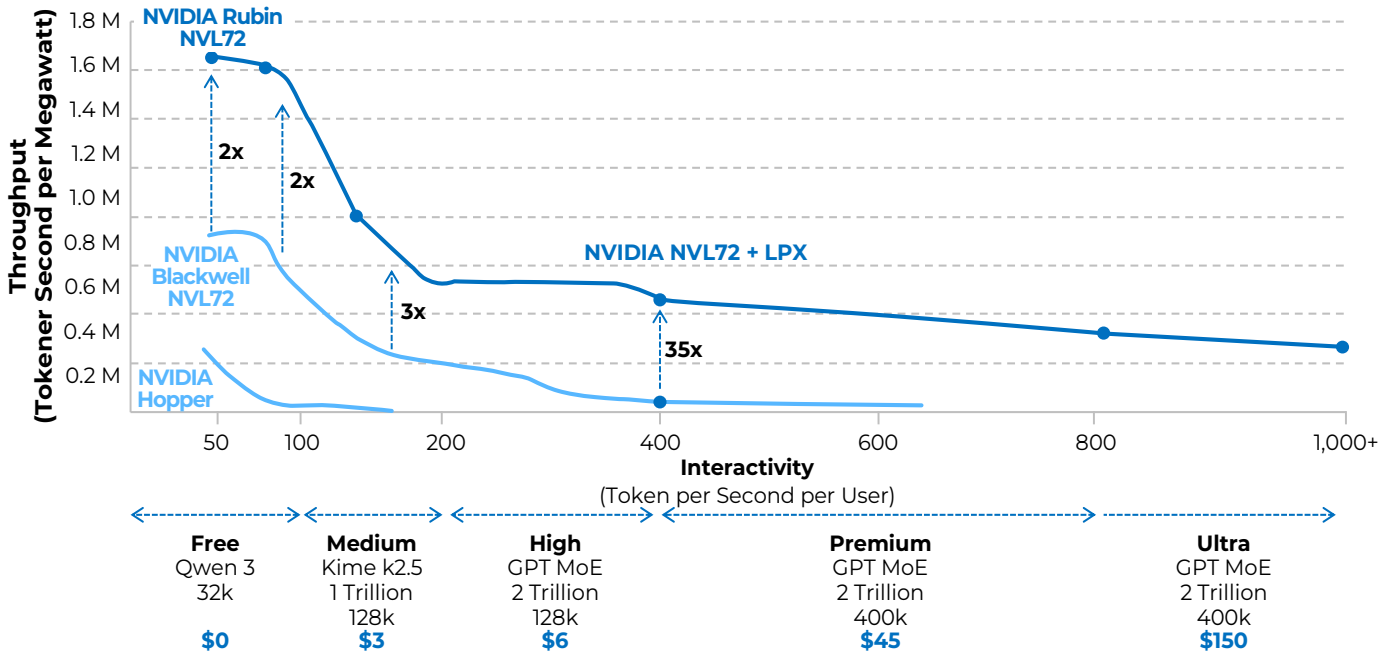
However, inference at scale demands high-throughput parallel processing for prefill (reading and understanding the input) and ultra-low-latency sequential generation for decode (producing the output one token at a time).

Addressing this concern, NVDA's first response came at the software layer. **Dynamo**, its open-source inference serving framework, disaggregates prefill and decode across separate GPU pools within a cluster and routes requests to the GPU that already holds the relevant KV cache in memory, reducing redundant computation in long-context and agentic workloads. This raises throughput and utilization on pure-GPU infrastructure without requiring new silicon.

At the hardware layer, NVDA licensed Groq's LPU technology in December 2025 and, at GTC 2026 (March 2026), unveiled the **Groq 3 LPX**, a standalone 256-LPU rack co-designed to pair with Vera Rubin NVL72 within a single AI supercomputer (the **Vera Rubin POD**). The LPU uses a fundamentally different architecture than GPUs, storing the model's parameters in on-chip SRAM, a faster memory type built directly into the processor rather than in off-chip HBM. This dramatically reduces the memory-access latency that constrains GPU decode speed.

In deployment, Rubin GPUs handle the compute-intensive prefill phase while Groq LPUs handle decode, with Dynamo orchestrating the prefill/decode split across the heterogeneous hardware. SemiAnalysis benchmarks indicate that the Vera Rubin POD (Vera Rubin NVL72 + Groq 3 LPX) delivers up to ~35x more throughput per megawatt relative to Blackwell NVL72.

Figure 33 – Vera Rubin NVL72 + Groq 3 LPX vs. Blackwell and Hopper: inference throughput per megawatt across user interactivity levels



Source: Safra, NVIDIA, SemiAnalysis.

Note: The chart plots inference deployments along two dimensions: y-axis is the data center economics (tokens/second per megawatt) and the x-axis is user experience (tokens/second per user, a measure of response speed). The two trade off: serving users faster leaves less aggregate throughput from the same hardware. Each curve is the performance frontier for a given hardware, and each generation pushes it outwards – with Rubin NVL72 + LPX (NVDA's next-gen GPU+LPU rack, launching 2H26) delivering up to ~35x the throughput per megawatt of Blackwell at high interactivity levels.

While we view inference ASICs as a legitimate competitive risk, NVDA's integration of a dedicated decode accelerator into the rack effectively absorbs the threat into its own platform, further deepening its ecosystem lock-in.

IV. Demand durability and TAM expansion: hyperscalers' capex, sovereign AI, and CPU-to-accelerated computing migration

The compute demand described in the preceding sections, driven by scaling laws, expanding inference workloads, and the agentic multiplier on token volume, materializes as physical orders for GPU racks, networking equipment, and the data center capacity to house them.

The AI infrastructure buildout

The AI buildout demand flows through NVDA's Data Center segment, which accounts for ~90% of the company's total revenue. Hyperscalers (Google, Amazon, Microsoft, Meta, and Oracle, all Not Covered) account for ~60% of Data Center revenue. The remaining ~40% is diversified across NVIDIA Cloud Partners (NCPs, GPU-focused cloud providers commonly referred to as "neoclouds"), frontier AI labs with direct purchasing relationships (Anthropic, OpenAI, xAI, and others procuring dedicated capacity outside their hyperscaler compute allocations), sovereign AI programs, supercomputing centers, and industrial and enterprise customers.

As NVDA's largest client group, hyperscalers and major NCPs are projected by consensus to spend ~USD 743bn on capex in FY'27E (NVDA fiscal year ending January 2027, roughly equivalent to CY'26E), with the vast majority directed toward AI-related infrastructure per management commentary across the group. On a cumulative basis, consensus estimates embed ~USD 4.3trn of capex over the next 5 fiscal years (FY'27E - FY'31E).

Figure 34 – NVDA's demand composition

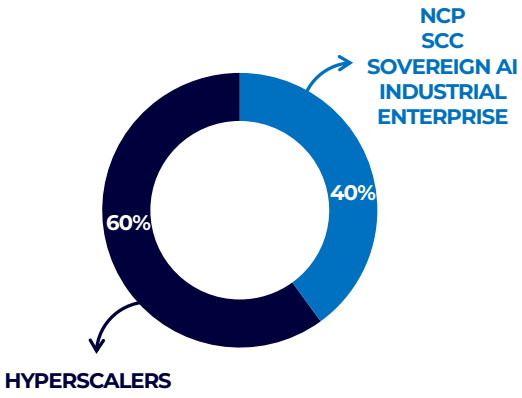
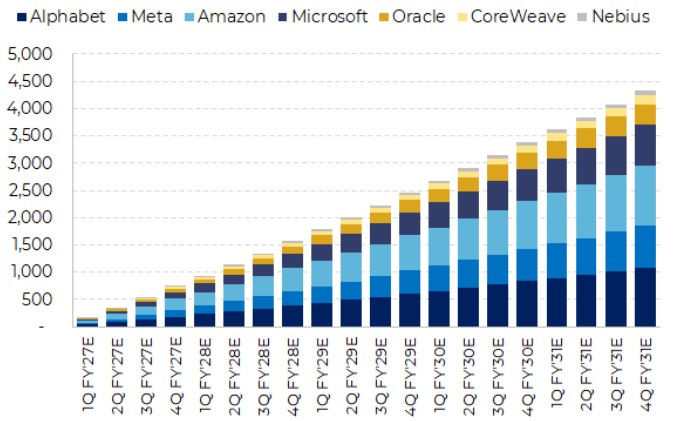


Figure 35 – Hyperscalers' cumulative capex, in USD bn (FY'27E – FY'31E)



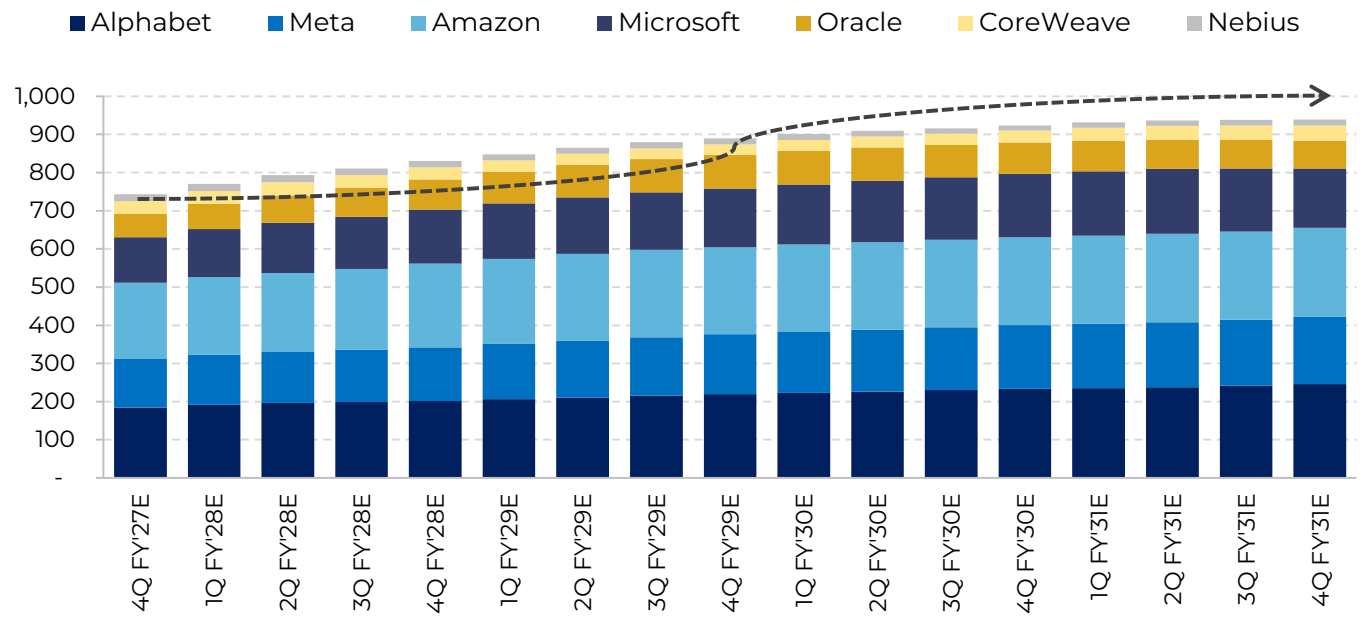
Source: Safra, Visible Alpha, NVIDIA.

Note: Includes capex estimates of Amazon, Alphabet, Meta, Microsoft, Oracle, CoreWeave and Nebius. All Not Covered.

A significant portion of this capex (~40%+) flows directly into NVDA's Data Center revenue segment through two channels: **compute** (GPUs, CPUs, LPUs from 2H26 onwards, and integrated systems such as DGX, HGX, and NVL72 racks) and **networking** (NVLink, NVSwitch, Spectrum-X, Quantum, ConnectX, BlueField). Networking operates as an attach rate on compute, so as rack-scale architectures grow denser, each GPU deployed pulls through a higher dollar value of networking infrastructure.

However, that cumulative consensus capex path embeds a sharp deceleration in growth across the group on a consolidated basis: from +63% YoY in FY'27E to +12% YoY in FY'28E (~USD 830bn), +7% YoY in FY'29E (~USD 889bn), +4% YoY in FY'30E (~USD 924bn), and +1.7% YoY in FY'31E (~USD 939bn), converging toward PCE inflation growth. Under consensus projections, capex-to-D&A for the group tells a similar story, declining from ~3.2x in FY'27E toward maintenance-level reinvestment.

Figure 36 – Hyperscalers and neoclouds LTM capex path, quarterly consensus estimates (in USD bn) – a material growth slowdown from FY'27E onwards



Source: Safra, Visible Alpha.

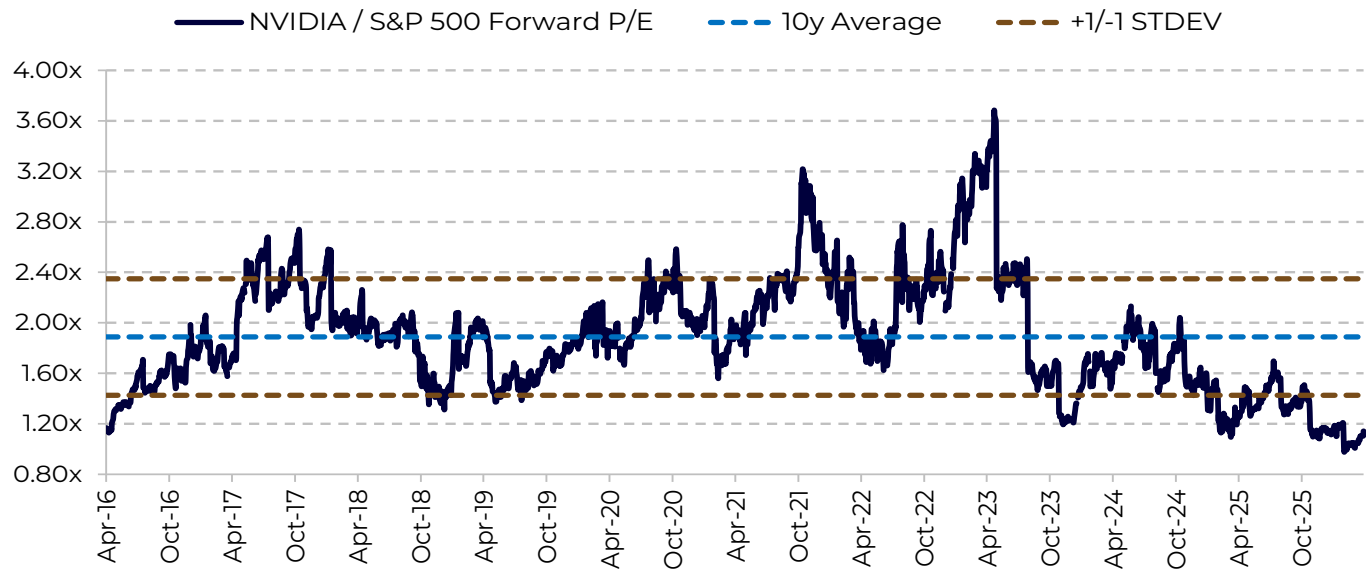
Note: Includes capex estimates of Amazon, Alphabet, Meta, Microsoft, Oracle, CoreWeave and Nebius. All Not Covered.

We believe consensus estimates are internally inconsistent with the demand trajectory established through this report. Consensus has systematically underestimated the investment cycle, with actual spending materially exceeding estimates. We believe the projected deceleration is largely attributable to two reinforcing concerns:

The first concern is an internal financing constraint. Capex as a percentage of combined OCF for the group is projected to reach ~97% in FY'27E under consensus, and estimates embed a sharp recomposition of FCF margins going forward. **The second concern is ROIC uncertainty**, specifically whether AI investments at scale will earn returns above cost of capital.

Together, these concerns have fueled "**peak capex**" fears that in our view are among the largest contributors to NVDA's valuation de-rating.

Figure 37 – NVIDIA's forward P/E relative to S&P 500 index



Source: Safra, Bloomberg.

We disagree with this framing. While we acknowledge that the ultimate dispersion of outcomes is wide, we view the concerns as more short-sighted than structural. Three factors support a more durable demand trajectory for NVDA than the market currently prices:

Factor 1: The competitive equilibrium penalizes underinvestment

The AI infrastructure buildout closely resembles a Prisoner's Dilemma, where defection (pulling back on capex) carries asymmetrically worse downside than continued spending. Each hyperscaler faces a strategic trade-off: if competitors sustain high capex while one player throttles back, the throttler risks falling behind in model capability, developer ecosystem, and enterprise platform relevance.

Critically, the costs of underinvestment are largely irrecoverable (capability gaps compound as models and ecosystems scale), while the costs of overinvestment can be absorbed through a longer cycle of capacity reallocation. The dominant strategy therefore becomes continued spending, and the persistent upward revisions to capex guidance since the start of the AI cycle are consistent with this equilibrium.

This dynamic was stress-tested in early 2025, when Chinese AI lab DeepSeek released its R1 reasoning model, trained on export-restricted NVDA GPUs. DeepSeek disclosed training costs of ~USD 6mn for its V3 base model, with an incremental ~USD 294K for the R1 reinforcement learning phase. Although independent estimates (SemiAnalysis) placed DeepSeek's total server capex closer to USD 1.6bn once infrastructure, hardware, and operating costs were included, the headline narrative was that comparable reasoning performance to OpenAI's o1 had been achieved at a fraction of the commonly assumed compute budget.

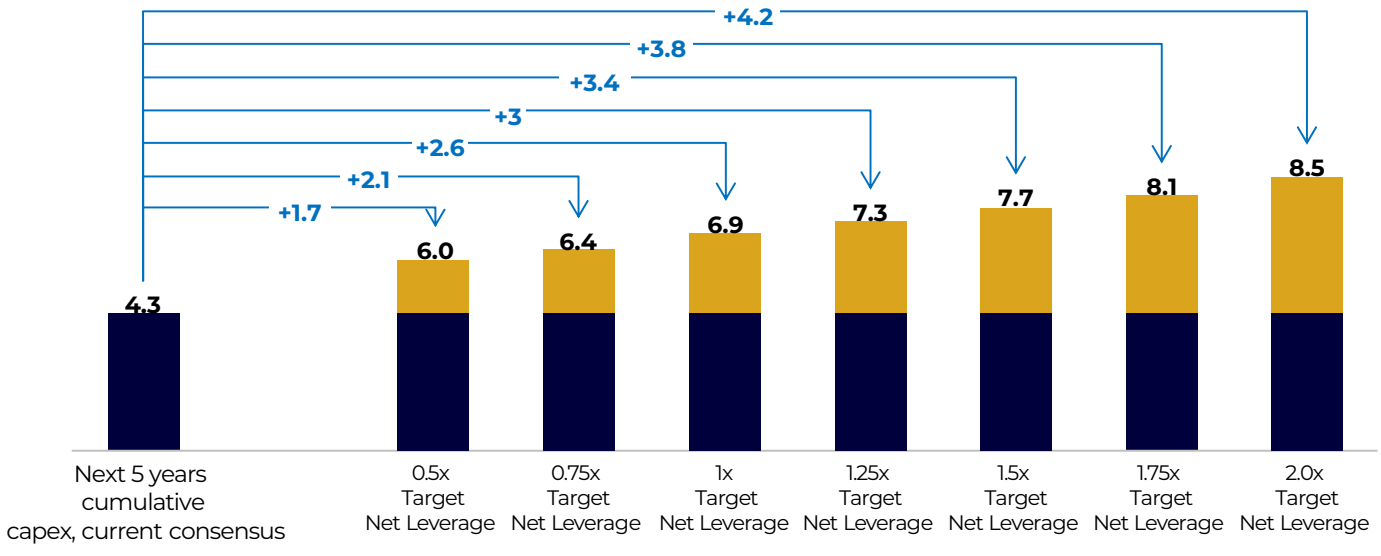
If hyperscaler capex were primarily a function of compute required per unit of AI capability, this news should have triggered deceleration. Instead, every major hyperscaler reiterated or raised capex guidance in the subsequent quarters. The DeepSeek episode demonstrated that efficiency gains reinforce rather than erode the investment cycle, consistent with the "Jevons paradox" described earlier.

Factor 2: Balance sheet capacity sets a floor under the demand curve

Beyond the competitive equilibrium, the hyperscalers' balance sheets represent an underappreciated source of additional investment capacity. As of FY'26, aggregate net debt-to-EBITDA for the group stands at ~0.1x, near net cash. Consensus implicitly assumes the AI buildout remains fully internally financed, with capex bounded by operating cash flow.

As a sensitivity exercise, we model scenarios in which aggregate net leverage rises *pro-rata* from current levels to a target net debt-to-EBITDA ratio by FY'31E, ranging from 0.5x to 2.0x. Even at a 0.50x target net leverage, well below any investment-grade corporate benchmark, the group would have ~USD 1.7trn in incremental capex capacity through FY'27E - '31E, or ~40% above the ~USD 4.3trn already embedded in consensus. At 2.00x target net leverage, still modest relative to traditional utilities (which share a comparable recurring revenue profile and routinely operate at ~3x or higher), incremental capacity rises to ~USD 4.2trn, nearly doubling the consensus path to ~USD 8.5trn.

Figure 38 – Hyperscalers and neoclouds cumulative capex for the next 5 years under different target net debt/EBITDA levels (in USD trn)



Source: Safra, Visible Alpha.

Note: Includes Amazon, Alphabet, Meta, Microsoft, Oracle, CoreWeave and Nebius. All Not Covered.

While these scenarios are not embedded in our base case, they illustrate a potential NVDA's demand lever that could be activated. We would not, however, interpret balance sheet financed capex as straightforward upside for NVDA, as the market would likely not reward profits derived from clients' debt-funded investment with sustainably higher multiples (as it is not repeatable).

For NVDA to sustain a prolonged re-rating, investors would need greater visibility on the ROI of deployed AI buildout, which leads directly to the question of downstream monetization.

Factor 3: Downstream monetization is already materializing

The most important factor, in our view, is that AI infrastructure is already paying off where it has been deployed. AI workloads have reaccelerated cloud revenue growth at scale, with every major hyperscaler describing itself as supply-constrained and inference demand exceeding available capacity across the group.

The demand signal is independently corroborated by **TSMC** (Not Covered), whose management has verified demand directly with its customers and end-customers, confirming measurable returns on deployed AI buildout. We believe that TSMC's assessment carries weight given its position as the monopoly foundry of all leading-edge AI accelerators and its incentive to be conservative given the capital intensity of fab construction.

On the application layer, we highlight **Meta's** (Not Covered) AI-driven recommendation models, which have expanded ad impressions and average price per ad simultaneously at double-digit rates, something rare in the company's operating history. Historically, when Meta launched new ad surfaces (Stories, then Reels), impression growth came at the expense of pricing as incremental inventory diluted ad quality.

The simultaneous expansion of both suggests that AI compute is creating net new advertising value rather than redistributing existing spend. We view this as particularly relevant because Meta's TAM is ultimately anchored to consumer spending, the largest share of US GDP. If AI cannot generate measurable returns there, the readthrough for cloud customers deploying AI workloads (and ultimately for the CSPs as well) would be less compelling, in our view.

While there is evidence of encouraging ROI from AI deployments across public equities, we believe the largest source of market discomfort with AI ecosystem monetization may lie in the frontier AI labs themselves. AI labs remain unprofitable and cash flow negative, requiring continuous VC funding given the unprecedented capital intensity of frontier model development. Furthermore, their growing share of hyperscalers' backlogs has led investors to discount the associated demand as commitments that may not fully materialize.

We note, however, that the leading labs are rapidly pivoting toward enterprise and API contracts with materially higher willingness to pay and multi-year visibility. For instance, Anthropic derives ~80%+ of revenue from this channel, and OpenAI appears to be following a similar trajectory, discontinuing consumer ventures to concentrate on enterprise deployments.

Taken together, we believe the balance of probabilities favors a more durable demand trajectory for NVDA than the market currently prices.

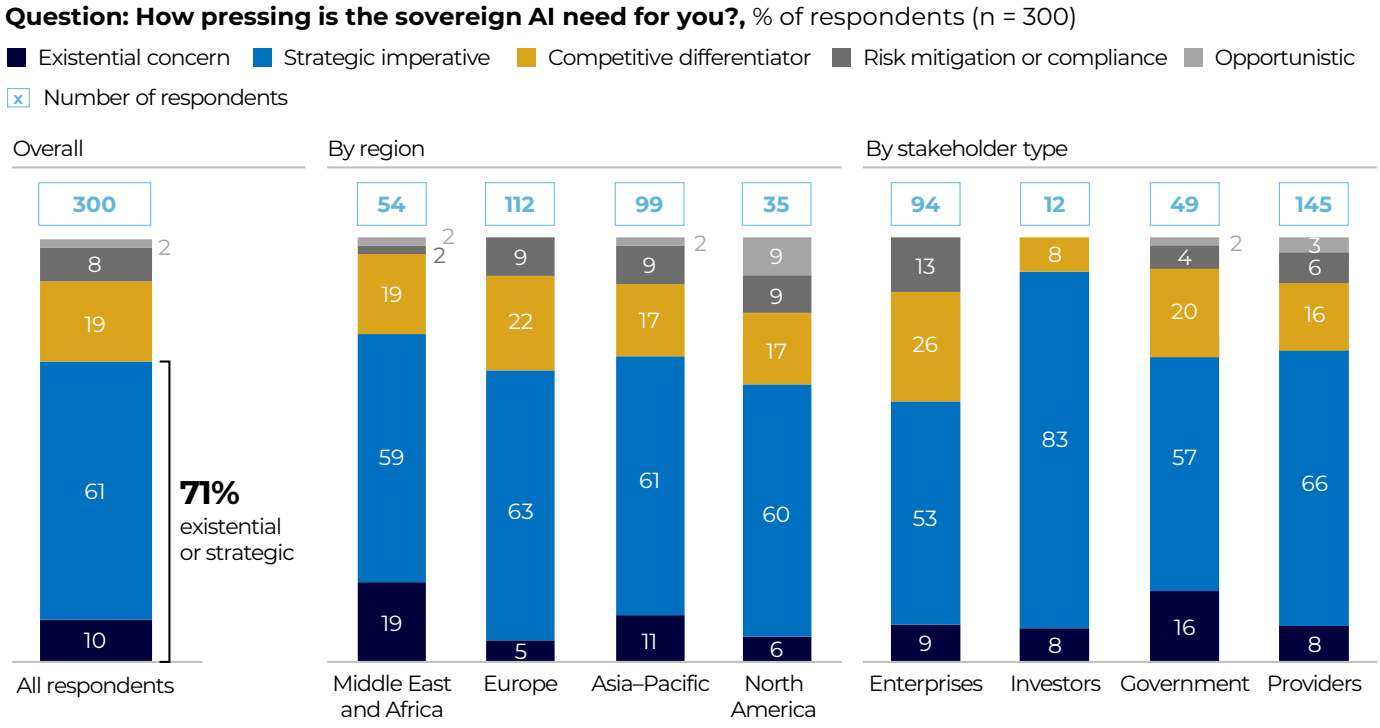
TAM expansion beyond hyperscalers: sovereign AI and CPU-to-GPU workload migration

The arguments above establish the durability of hyperscaler-driven demand. But additional vectors expand NVDA's addressable market outside the hyperscaler envelope: sovereign AI and the migration of legacy CPU workloads to accelerated computing.

Sovereign AI represents a demand vector that is geopolitically distinct from enterprise and cloud capex and, in our view, more inelastic. NVDA generated over USD 30bn in sovereign AI revenue in FY'26A, more than tripling YoY to ~15% of Data Center revenue. We estimate sovereign AI revenue could roughly double again in FY'27E to ~USD 60bn, and thereafter grow at least in line with AI infrastructure spending as a proportion of global GDP.

The underlying demand is broad-based. According to a McKinsey global survey of 300 executives, investors, and government officials, 71% characterize sovereign AI as an "existential concern" or "strategic imperative" to their organizational goals.

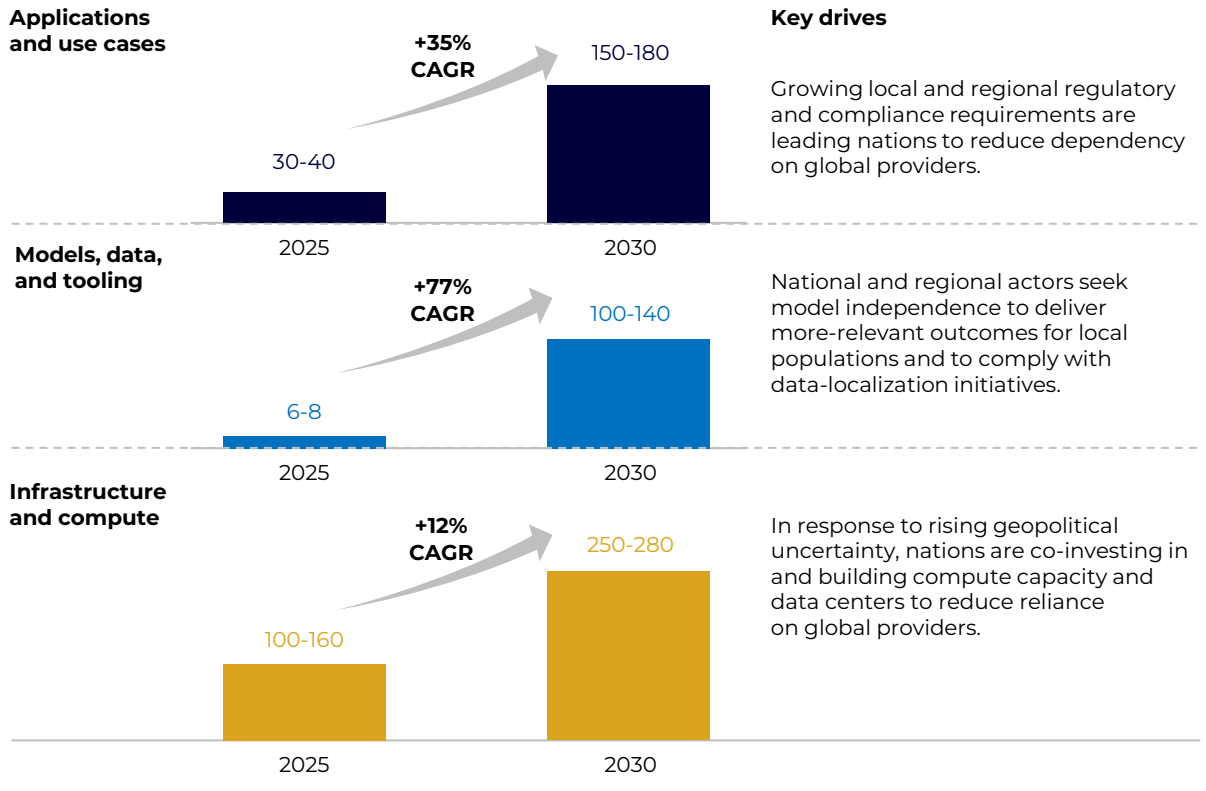
Figure 39 – McKinsey's sovereign AI survey: 71% of respondents classify AI investments as an existential or strategic imperative



Source: McKinsey

Nations are treating AI infrastructure as a competitiveness and national security asset, not a discretionary technology expenditure. This framing has two important implications for NVDA's revenue quality. First, sovereign purchasing decisions respond to strategic imperatives rather than ROIC thresholds, meaning the demand is more price-inelastic relative to enterprise or hyperscaler capex. Second, sovereign programs typically span multi-year national commitments, which further extends NVDA's demand visibility.

Figure 40 – Sovereign AI opportunity, already sizable, could expand to a USD 600 bn market by CY'30E



Source: McKinsey.

Sovereign customers purchase the same full-stack infrastructure as hyperscalers at comparable ASPs and margin profiles. Per NVDA, over 50 countries are now building AI infrastructure on the company's platforms, with pipelines spanning signed contracts to multi-year national programs.

Beyond AI, the migration of legacy computing workloads from CPUs to GPU-accelerated architectures represents an additional and largely independent vector of TAM expansion. The boundary between "AI demand" and "CPU-to-GPU migration" is blurry in practice, and NVDA does not segment the two internally. But there is a large installed base of non-AI data center workloads (data analytics, scientific simulation, database acceleration, video processing) that are economically suited to GPU acceleration regardless of the trajectory of AI spending.

McKinsey's data center capacity projections corroborate this: of the ~219 GW of global installed capacity projected by 2030, approximately 30% (~63 GW) is attributed to non-AI workloads that nonetheless would benefit from accelerated computing.

Management has framed this opportunity as a transition of the ~USD 1trn installed base of general-purpose data center infrastructure toward accelerated computing, with the global installed base projected to reach ~USD 2trn over the next four to five years as GPU-accelerated architectures replace CPU-only configurations at higher ASPs.

Finally, we highlight that NVDA's entry into the standalone CPU server market with Grace (general-purpose data center computing) and Vera (purpose-built for agentic AI workloads) extends this opportunity further, competing with **Intel** (Not Covered) and **AMD** (Not Covered) in a market they have historically dominated. Even servers that do not require a GPU may now carry NVDA silicon, expanding the company's footprint beyond accelerated computing.

Valuation and forecasts

Key assumptions. NVDA is well-positioned to capture the AI infrastructure buildout as its co-designed platform, annual generation cadence, and demand durability across hyperscalers, NCPs, AI labs, sovereigns, and enterprise CPU-to-GPU migration all compound through the forecast window.

We forecast **revenue to grow at a 29.5% CAGR from FY'26A to FY'31E**, supported by the **Data Center segment scaling from USD 194bn in FY'26A to USD 747bn by FY' 31E (31% CAGR), reaching ~95% of total revenue**. Within Data Center, we forecast **(i) Compute CAGR of ~30% and (ii) Networking CAGR of ~36% from FY'26A – FY'31E**, **with networking outgrowing compute as its attach rate rises from 19.3% to 24.5% (+520bps) over the forecast horizon**. This is driven by increasing NVDA networking content per rack across generations and the expansion of scale-out fabrics.

Figure 41 – NVDA's total revenue growth, in USD bn (FY' 2026A – FY' 2031E)

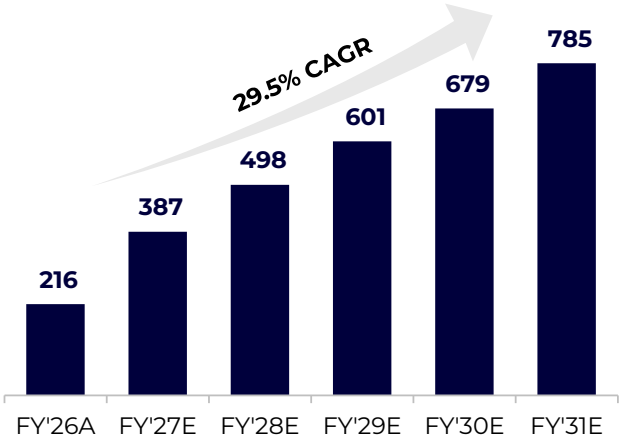
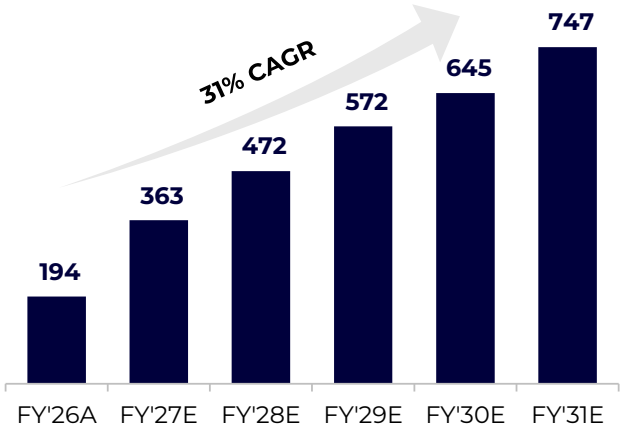


Figure 42 – NVDA's Data Center revenue growth, in USD bn (FY' 2026A – FY' 2031E)



Source: Safra.

Figure 43 – NVDA's Networking and Compute revenue growth, in USD bn (FY' 2026A – FY' 2031E)

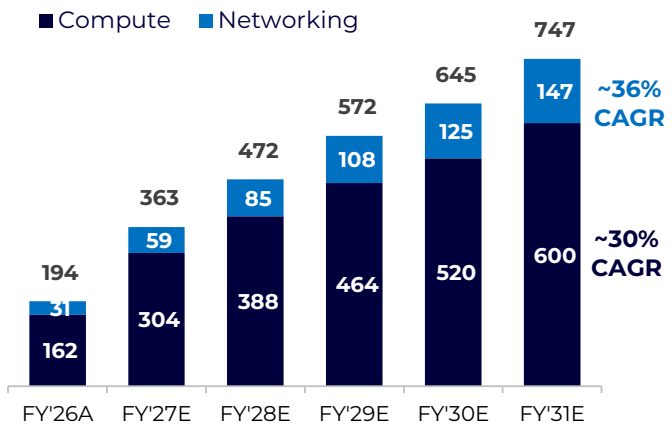
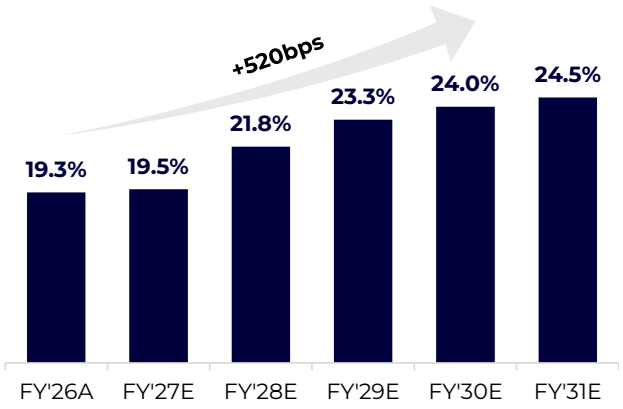


Figure 44 – Networking attach rate to Compute (FY' 2026A – FY' 2031E)



Source: Safra.

NVDA operates an asset-light, fabless model in which the bulk of the supply chain's capital intensity is absorbed upstream — for instance, PP&E and intangible assets (ex-goodwill) represented only ~6.6% of total assets as of FY'26A, with this ratio projected to decline over time.

The combination of little reinvestment needs alongside pricing power (reinforced by the favorable customer's upgrade-cycle economics) produces an **average gross margin of ~74%**, an **average EBIT margin of ~64%**, and an **average net margin of ~56%** over the forecast horizon. From FY'26A through FY'31E, we project **NOPAT and net income CAGR of 29.2% and 28.9%, respectively**, with an average **free cash flow conversion of ~98%**.

Figure 45 – NVDA's PP&E and intangibles (ex-goodwill, in USD bn (FY' 2026A – FY' 2031E)

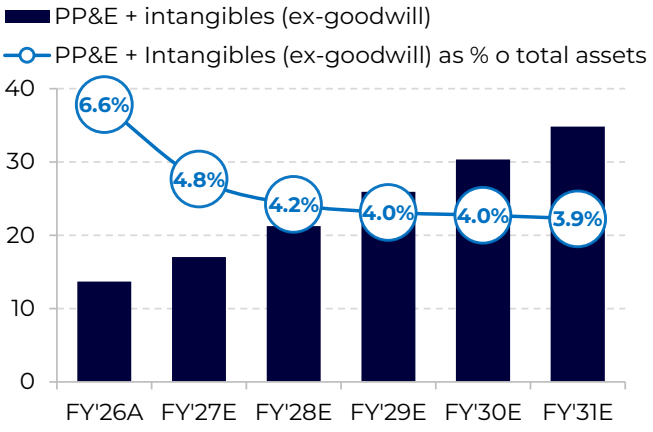
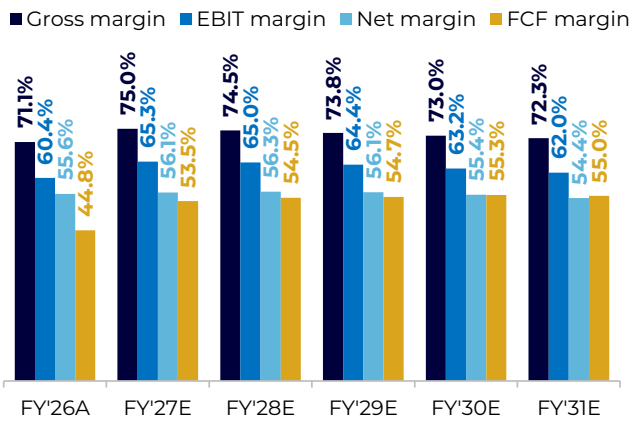


Figure 46 – NVDA's margin profile evolution (FY' 2026A – FY' 2031E)



Source: Safra.

Figure 47 – NVDA's NOPAT, in USD bn (FY' 2026A – FY' 2031E)

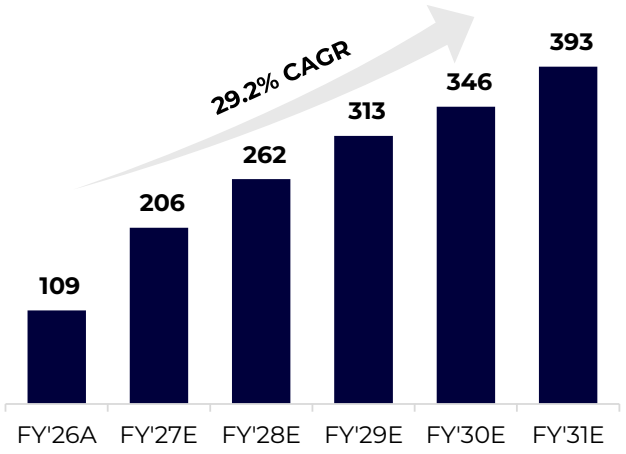
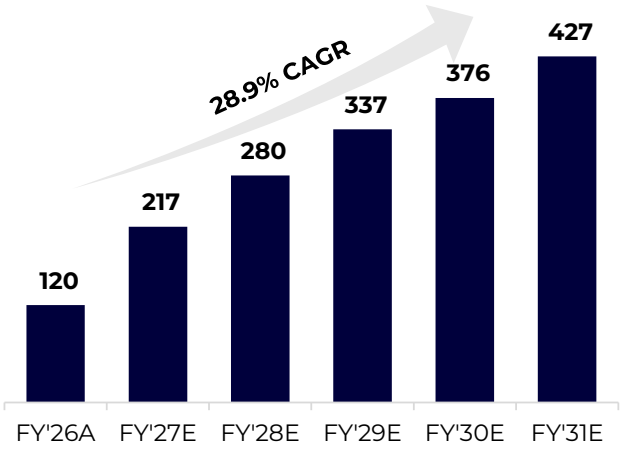


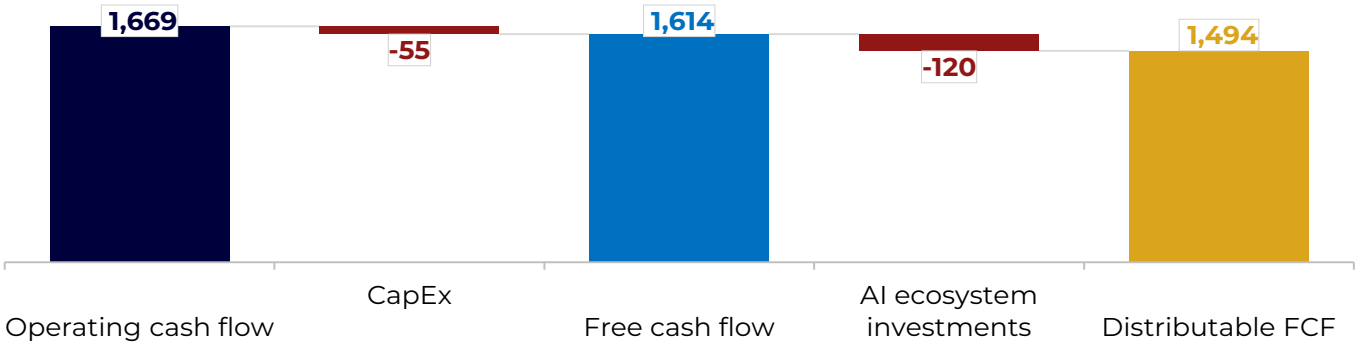
Figure 48 – NVDA's net income, in USD bn (FY' 2026A – FY' 2031E)



Source: Safra.

Capital allocation. We forecast **cumulative operating cash flow of ~USD 1.67trn over the next 5 fiscal years (FY'27E – '31E), of which ~USD 54.5bn is deployed into capex and ~USD 120bn into AI ecosystem investments** across OpenAI, Anthropic, CoreWeave, xAI, and a long tail of frontier labs and AI infrastructure companies. The anchor commitment is the ~USD 100bn OpenAI partnership announced in late 2025, under which NVIDIA commits to deploy at least 10 GW of systems to OpenAI progressively in exchange for equity stake. We view the broader portfolio as a venture-style optionality layer on top of the operating business, as positions may mark up in value through successive private funding rounds before potential IPOs. **This leaves ~USD 1.5trn of free cash flow available for distribution over the next 5 fiscal years.**

Figure 49 – NVDA's cumulative distributable cash flow, in USD bn (FY'2026A – FY' 2031E)



Source: Safra.

Shareholders' return. Although management has a stated policy of returning ~50% of free cash flow to shareholders, we read this as a floor rather than a ceiling, particularly in the outer years of the forecast. NVDA operates a low capital intensity business model, and continuing to deploy ~50% of FCF would lead to material and inefficient cash pooling on the balance sheet.

We believe the company will return the majority of its free cash flow available for distribution to shareholders, deploying ~USD 1,020bn in cumulative buybacks and ~USD 15bn in cumulative dividends through FY'31E, for a total shareholder return of ~USD 1,035bn over the next 5 years — equivalent to ~20% of the company's current market cap.

Figure 50 – NVDA's distributions to shareholders, in USD bn (FY' 2026A – FY' 2031E)

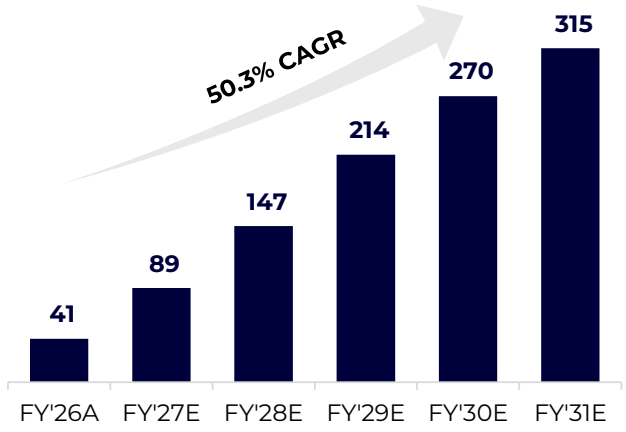
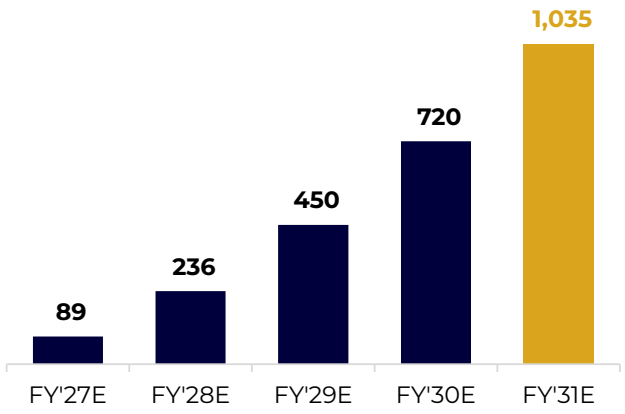


Figure 51 – NVDA's cumulative distributions to shareholders, in USD bn (FY' 2027E – FY' 2031E)



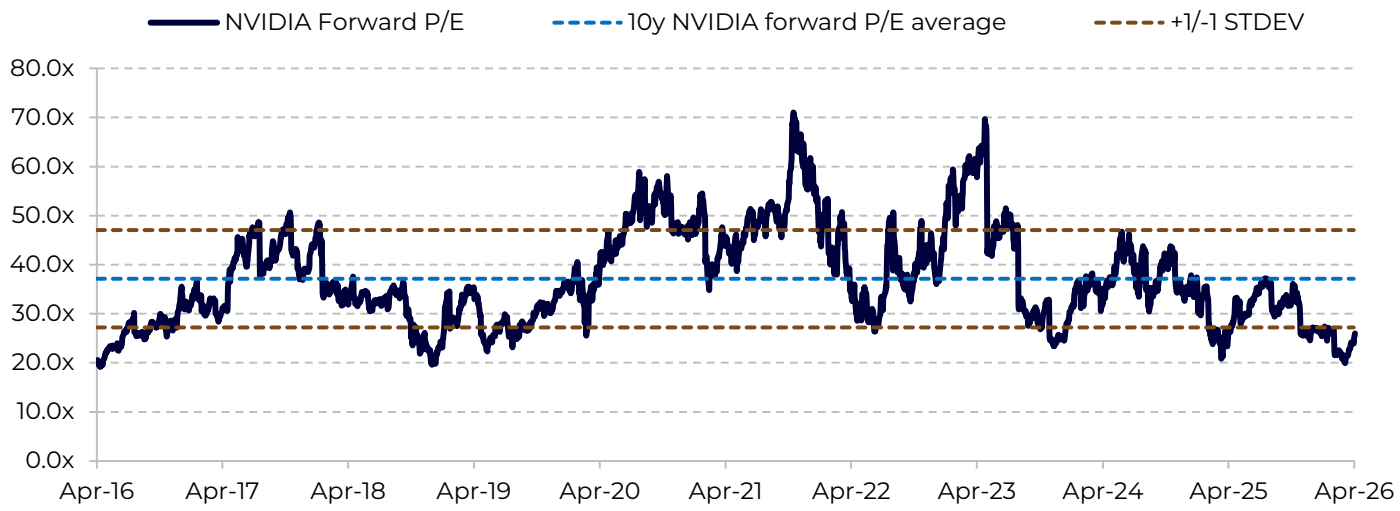
Source: Safra.

Even under our projections, NVDA would still end the forecast horizon with **~USD 480bn of cash and short-term investments in the balance sheet, equivalent to ~54% of total asset** (above its historical average of ~35% from FY'23A – FY'26A). Therefore, we see upside to our estimates, potentially positioning NVDA as one of the major capital return stories across global equities.

NVDA's massive cash flow generation and the resulting capital distributions should also act as downside protection during broad market valuation drawdowns. This creates a time-arbitrage opportunity for long-term shareholders: in periods of de-rating, the company can accelerate repurchases at lower prices, which increases per-share accretion and raises forward shareholder IRR.

Valuation. On our estimates, NVDA currently trades at a still attractive ~24x FY'27E (~CY'26E) P/E and ~18x FY'28E (~CY'27E) P/E, compared to a long-term average of ~37x, a mega-cap tech peer median of ~30x and semiconductor peer median of ~38x.

Figure 52 – NVDA's forward P/E



Source: Safra, Bloomberg.

Figure 53 – Comps table

Comps table	Price/GAAP Earnings			EV/GAAP EBIT			Next 3-years CAGR			PEG Ratio	ROIC, TTM
	CY'26E	CY'27E	CY'28E	CY'26E	CY'27E	CY'28E	Revenue	GAAP EBIT	GAAP EPS		
NVIDIA	**23.7x**	**18.0x**	**14.6x**	**19.9x**	**14.9x**	**12.0x**	**40.6%**	**43.4%**	**43.5%**	**0.55x**	**106%**
Mega-cap technology peers											
Apple	31.5x	28.9x	26.4x	26.5x	24.9x	23.3x	7.9%	8.1%	10.4%	3.0x	438%
Alphabet	29.2x	25.2x	21.8x	25.2x	21.5x	18.4x	15.6%	18.1%	11.2%	2.6x	51%
Microsoft	22.9x	20.1x	17.1x	18.7x	16.2x	13.8x	16.4%	16.5%	16.7%	1.4x	38%
Amazon	33.1x	26.7x	21.6x	27.9x	22.2x	17.7x	12.2%	20.0%	11.6%	2.8x	22%
Meta	22.2x	19.2x	16.5x	19.3x	16.5x	14.0x	19.7%	12.6%	20.8%	1.1x	45%
Broadcom	39.2x	25.3x	19.5x	33.9x	21.6x	16.8x	43.3%	43.7%	54.5%	0.7x	32%
TSMC	22.4x	17.1x	14.0x	18.7x	14.9x	12.4x	27.0%	31.0%	30.6%	0.7x	55%
ASML	38.1x	28.9x	25.2x	32.4x	24.7x	21.7x	16.9%	23.6%	25.1%	1.5x	88%
Median	**30.4x**	**25.3x**	**20.6x**	**25.8x**	**21.5x**	**17.2x**	**16.6%**	**19.0%**	**18.8%**	**1.44x**	**48%**
Semiconductors peers											
Broadcom	39.2x	25.3x	19.5x	33.9x	21.6x	16.8x	43.3%	43.7%	54.5%	0.7x	32%
TSMC	22.4x	17.1x	14.0x	18.7x	14.9x	12.4x	27.0%	31.0%	30.6%	0.7x	55%
ASML	38.1x	28.9x	25.2x	32.4x	24.7x	21.7x	16.9%	23.6%	25.1%	1.5x	88%
AMD	64.2x	35.0x	24.5x	54.4x	29.3x	21.2x	35.5%	51.6%	52.5%	1.2x	48%
Marvell Technology	85.5x	48.1x	30.5x	70.0x	42.4x	26.7x	33.0%	37.1%	37.2%	2.3x	55%
Applied Materials	34.0x	27.9x	25.0x	31.6x	24.7x	22.4x	13.2%	16.7%	17.4%	1.9x	32%
LAM Research	38.0x	28.9x	24.3x	32.9x	25.1x	21.4x	21.8%	27.9%	28.9%	1.3x	88%
KLA Corporation	43.7x	35.6x	31.6x	37.0x	30.1x	27.3x	14.6%	17.0%	18.6%	2.4x	48%
Median	**38.1x**	**28.9x**	**25.0x**	**32.9x**	**25.1x**	**21.7x**	**21.8%**	**27.9%**	**28.9%**	**1.52x**	**55%**

Source: Safra, Visible Alpha.

Figure 54 – Safra vs. consensus

	Safra Estimates			Consensus Estimates			Safra vs. Consensus		
	FY'27E	FY'28E	FY'29E	FY'27E	FY'28E	FY'29E	FY'27E	FY'28E	FY'29E
Revenue	**386,815**	**497,877**	**600,537**	**363,279**	**481,869**	**565,226**	**6.5%**	**3.3%**	**6.2%**
Gross Profit (GAAP)	**290,111**	**370,919**	**442,896**	**269,824**	**359,882**	**419,482**	**7.5%**	**3.1%**	**5.6%**
Gross Margin %	*75.0%*	*74.5%*	*73.8%*	*74.3%*	*74.7%*	*74.2%*	*73bps*	*-18bps*	*-46bps*
Operating Income (GAAP)	**252,448**	**323,533**	**386,512**	**234,908**	**314,782**	**367,681**	**7.5%**	**2.8%**	**5.1%**
Operating Margin %	*65.3%*	*65.0%*	*64.4%*	*64.7%*	*65.3%*	*65.1%*	*60bps*	*-34bps*	*-69bps*
EBITDA	**256,268**	**328,519**	**392,943**	**238,942**	**320,137**	**373,901**	**7.3%**	**2.6%**	**5.1%**
EBITDA Margin %	*66.3%*	*66.0%*	*65.4%*	*65.8%*	*66.4%*	*66.2%*	*48bps*	*-45bps*	*-72bps*
Net Income (GAAP)	**216,822**	**280,275**	**337,080**	**199,917**	**267,615**	**318,060**	**8.5%**	**4.7%**	**6.0%**
Net Margin %	*56.1%*	*56.3%*	*56.1%*	*55.0%*	*55.5%*	*56.3%*	*102bps*	*76bps*	*-14bps*

Source: Safra, Visible Alpha.

Figure 55 – Safra Analysis and Valuation Scorecard (SAVS)

P&L. in USD bn	FY'25A	FY'26A	FY'27E	FY'28E	FY'29E
Revenue	**130.5**	**215.9**	**386.8**	**497.9**	**600.5**
Data center	115.2	193.7	363.2	472.2	571.5
Gaming	11.4	16.0	15.0	15.6	17.1
Professional visualization	1.9	3.2	5.1	5.8	6.2
Automotive	1.7	2.3	2.8	3.6	4.9
OEM and others	0.4	0.6	0.7	0.7	0.8
Cost of revenue	(32.6)	(62.5)	(96.7)	(127.0)	(157.6)
Gross profit	**97.9**	**153.5**	**290.1**	**370.9**	**442.9**
Operating income	**81.5**	**130.4**	**252.4**	**323.5**	**386.5**
EBITDA	**83.3**	**133.2**	**256.3**	**328.5**	**392.9**
EBT	84.0	141.4	263.7	341.8	411.1
Income tax expense	(11.1)	(21.4)	(46.9)	(61.5)	(74.0)
Net income	**72.9**	**120.1**	**216.8**	**280.3**	**337.1**
Balance Sheet, in USD bn)	**FY'25A**	**FY'26A**	**FY'27E**	**FY'28E**	**FY'29E**
Cash and cash equivalents	43.2	62.6	141.7	233.1	315.7
Accounts receivable, net	23.1	38.5	60.2	76.3	92.6
Inventories	10.1	21.4	29.4	37.9	47.4
Other current assets	3.8	3.2	5.0	6.4	7.7
Current assets	80.1	125.6	236.4	353.6	463.4
Property and equipment, net	6.3	10.4	14.3	19.2	24.4
Goodwill	5.2	20.8	20.8	20.8	20.8
Others non current assets	20.0	50.0	79.9	107.7	132.2
Total assets	**111.6**	**206.8**	**351.4**	**501.4**	**640.8**
Accounts payable	6.3	9.8	14.5	18.7	23.4
Accrued and other current liabilities	11.7	21.4	33.7	42.7	51.9
Short-term debt	0.0	1.0	0.0	1.3	0.0
Current liabilities	18.0	32.2	48.2	62.7	75.2
Long-term debt	8.5	7.5	7.5	6.2	6.2
Other long-term liabilities	5.8	9.9	9.9	9.9	9.9
Total liabilities	**32.3**	**49.5**	**65.6**	**78.8**	**91.3**
Shareholder's Equity	**79.3**	**157.3**	**285.8**	**422.6**	**549.5**

Others, in USD bn	FY'25A	FY'26A	FY'27E	FY'28E	FY'29E
NOPAT	70.3	109.0	205.5	262.0	312.5
Operating Cash Flow	**64.1**	**102.7**	**214.0**	**280.6**	**339.7**
CapEx	(3.2)	(6.0)	(7.2)	(9.2)	(11.1)
Free Cash Flow	**60.9**	**96.7**	**206.9**	**271.4**	**328.5**
Buybacks & Dividends	34.5	41.1	88.9	147.2	214.4
Shareholders' Yield (%)	0.7%	0.8%	1.8%	3.1%	4.6%
FCF conversion (%)	83.5%	80.5%	95.4%	96.8%	97.5%

Margins	FY'25A	FY'26A	FY'27E	FY'28E	FY'29E
Gross margin	75.0%	71.1%	75.0%	74.5%	73.8%
EBITDA margin	63.8%	61.7%	66.3%	66.0%	65.4%
EBIT margin	62.4%	60.4%	65.3%	65.0%	64.4%
Pre-tax margin	64.4%	65.5%	68.2%	68.7%	68.5%
Net Income margin	55.8%	55.6%	56.1%	56.3%	56.1%
Free Cash Flow margin	46.6%	44.8%	53.5%	54.5%	54.7%

Profitability Analysis	FY'25A	FY'26A	FY'27E	FY'28E	FY'29E
ROE	91.9%	101.5%	99.4%	101.5%	66.3%
ROA	91.0%	116.7%	121.4%	96.6%	78.1%
ROIC	157.7%	105.6%	136.8%	137.3%	130.7%

Valuation	FY'25A	FY'26A	FY'27E	FY'28E	FY'29E
EV/Sales	37.6x	22.4x	12.1x	9.1x	7.2x
EV/Gross Profit	50.1x	31.5x	16.2x	12.2x	9.7x
EV/EBIT	60.2x	37.1x	18.6x	13.9x	11.1x
EV/EBITDA	58.9x	36.3x	18.3x	13.7x	11.0x
P/E	67.8x	40.7x	22.2x	16.9x	13.7x
P/FCF	81.2x	50.5x	23.3x	17.4x	14.1x
FCF Yield	1.2%	2.0%	4.3%	5.7%	7.1%
WACC (%)					10.0%
Perpetual growth (%)					6.0%

Source: Safra.

Figure 56 – Safra estimates – P&L, in USD mn

P&L, in USD mn	FY2025A	FY2026A	FY2027E	FY2028E	FY2029E	FY2030E	FY2031E
Total Revenue	**130,497**	**215,938**	**386,815**	**497,877**	**600,537**	**678,764**	**785,144**
Data Center	115,186	193,737	363,186	472,167	571,525	645,253	747,374
Gaming	11,350	16,042	15,046	15,616	17,084	18,464	20,473
Professional Visualization	1,878	3,191	5,087	5,764	6,205	6,503	6,802
Auto	1,694	2,349	2,837	3,645	4,944	7,745	9,799
OEM & IP	389	619	660	685	778	799	696
Cost of Revenues	(32,638)	(62,475)	(96,704)	(126,959)	(157,641)	(183,266)	(217,878)
Cost of revenue SBC	(179)	(261)	(334)	(343)	(399)	(569)	(841)
Cost of revenues ex-SBC	(32,459)	(62,214)	(96,370)	(126,616)	(157,242)	(182,697)	(217,036)
Gross Profit (GAAP)	**97,859**	**153,463**	**290,111**	**370,919**	**442,896**	**495,497**	**567,267**
Total operating expenses (GAAP)	**(16,405)**	**(23,076)**	**(37,663)**	**(47,385)**	**(56,384)**	**(66,502)**	**(80,512)**
R&D ex-SBC	(9,490)	(13,820)	(24,753)	(32,388)	(38,618)	(43,986)	(50,773)
R&D SBC	(3,424)	(4,677)	(5,882)	(6,050)	(7,045)	(10,035)	(14,834)
SG&A ex-SBC	(2,356)	(3,131)	(5,351)	(7,222)	(8,713)	(9,620)	(10,676)
SG&A SBC	(1,135)	(1,448)	(1,677)	(1,725)	(2,009)	(2,861)	(4,230)
Other charges (Non-GAAP)	-	-	-	-	-	-	-
Total stock-based compensation	(4,738)	(6,386)	(7,893)	(8,118)	(9,453)	(13,465)	(19,905)
Operating income (GAAP)	**81,454**	**130,387**	**252,448**	**323,533**	**386,512**	**428,995**	**486,754**
Total other income and expense	2,572	11,062	11,282	18,265	24,561	29,575	34,409
Interest income	1,786	2,299	3,221	5,956	8,360	10,115	12,274
Interest expense	(247)	(260)	(231)	(214)	(199)	(194)	(150)
Other, net	1,033	9,023	8,292	12,524	16,400	19,653	22,285
Depreciation & amortization	1,864	2,842	3,820	4,986	6,431	8,119	10,061
EBITDA	**83,318**	**133,229**	**256,268**	**328,519**	**392,943**	**437,114**	**496,815**
Pre-tax income (GAAP)	84,026	141,449	263,730	341,799	411,073	458,570	521,163
Income taxes	(11,146)	(21,382)	(46,909)	(61,524)	(73,993)	(82,543)	(93,809)
Net income	**72,880**	**120,067**	**216,822**	**280,275**	**337,080**	**376,027**	**427,354**

Source: Safra.

Figure 57 – Safra estimates – balance sheet, in USD mn

Balance Sheet, in USD mn	FY2025A	FY2026A	FY2027E	FY2028E	FY2029E	FY2030E	FY2031E
Total Assets	**111,601**	**206,805**	**351,396**	**501,377**	**640,841**	**763,282**	**894,298**
Current assets	**80,126**	**125,607**	**236,362**	**353,618**	**463,403**	**561,406**	**671,958**
Cash and cash equivalents	43,210	62,558	141,746	233,056	315,746	393,791	482,443
Other current assets	36,916	63,049	94,616	120,562	147,658	167,615	189,515
Non-current assets	**31,475**	**81,198**	**115,034**	**147,759**	**177,438**	**201,876**	**222,341**
Property and equipment, net	6,283	10,383	14,346	19,190	24,411	29,274	34,241
Goodwill	5,188	20,832	20,832	20,832	20,832	20,832	20,832
Intangible assets, net	807	3,306	2,679	2,060	1,518	1,094	591
Other non-current assets	19,197	46,677	77,177	105,677	130,677	150,677	166,677
Total Liabilities and Shareholders' Equity	**111,601**	**206,805**	**351,396**	**501,377**	**640,841**	**763,282**	**894,298**
Current liabilities	**18,047**	**32,163**	**48,230**	**62,693**	**75,230**	**86,874**	**97,751**
Accounts payable	6,310	9,812	14,489	18,715	23,379	27,049	31,150
Accrued and other current liabilities	11,737	21,352	33,741	42,728	51,851	58,324	65,351
Short-term debt	-	999	-	1,250	-	1,500	1,250
Non-current liabilities	**14,227**	**17,347**	**17,346**	**16,096**	**16,096**	**14,596**	**13,346**
Long-term debt	8,463	7,469	7,468	6,218	6,218	4,718	3,468
Other long-term liabilities	5,764	9,878	9,878	9,878	9,878	9,878	9,878
Shareholder's Equity	**79,327**	**157,295**	**285,821**	**422,588**	**549,515**	**661,812**	**783,201**

Source: Safra.

Figure 58 – Safra estimates – cash flow statements, in USD mn

Cash Flow Statements, in USD mn	FY2025A	FY2026A	FY2027E	FY2028E	FY2029E	FY2030E	FY2031E
Net income	72,880	120,067	216,822	280,275	337,080	376,027	427,354
Depreciation & amortization	1,864	2,842	3,820	4,986	6,431	8,119	10,061
Stock based compensation	4,738	6,386	7,893	8,118	9,453	13,465	19,905
Working Capital	(9,383)	(15,948)	(14,502)	(12,732)	(13,308)	(9,814)	(10,773)
Decreases / (Increases) in working capital assets	(18,241)	(26,145)	(31,567)	(25,946)	(27,095)	(19,957)	(21,900)
Increases / (Decreases) in working capital liabilities	7,637	8,353	17,066	13,214	13,787	10,143	11,128
Others	(6,008)	(10,628)	-	-	-	-	-
Deferred income taxes	(4,476)	(1,424)	-	-	-	-	-
Other cash flow from operations items	(1,532)	(9,204)	-	-	-	-	-
Cash flow from operations	**64,091**	**102,719**	**214,033**	**280,647**	**339,655**	**387,797**	**446,546**
Capital expenditures	(3,236)	(6,042)	(7,156)	(9,211)	(11,110)	(12,557)	(14,525)
Other investing activities	(2,299)	(28,249)	(30,500)	(28,500)	(25,000)	(20,000)	(16,000)
(Purchase)/Sales of marketable securities	(14,886)	(17,937)	-	-	-	-	-
Cash flow from investing	**(20,421)**	**(52,228)**	**(37,656)**	**(37,711)**	**(36,110)**	**(32,557)**	**(30,525)**
Dividends	(834)	(974)	(1,166)	(1,400)	(2,799)	(3,359)	(4,031)
Capital increases and debt issuance/(payment)	(7,691)	(7,304)	(8,334)	(4,465)	(6,449)	(7,406)	(12,448)
Stock buybacks	(33,706)	(40,087)	(87,689)	(145,762)	(211,607)	(266,430)	(310,891)
Cash flow from financing	**(42,231)**	**(48,365)**	**(97,189)**	**(151,626)**	**(220,855)**	**(277,195)**	**(327,370)**
Free cash flow	**60,855**	**96,677**	**206,877**	**271,436**	**328,545**	**375,240**	**432,021**
Free cash flow available for distribution	**58,556**	**68,428**	**176,377**	**242,936**	**303,545**	**355,240**	**416,021**

Source: Safra.

Our YE'26 target price of USD 300/share derives from our FCFF DCF, for which we assume a **cost of equity of 10.2%**, an **after-tax cost of debt of 3.8%**, and a target capital structure of **97.5% equity / 2.5% debt**, leading to a **WACC of 10%.** We assume a **perpetuity growth (g) of 6%.**

Figure 59 – Target price sensitivity to WACC and terminal growth rate (g)

WACC \ g	4.5%	5.0%	5.5%	6.0%	6.5%	7.0%	7.5%
12.5%	$ 165	$ 171	$ 178	$ 185	$ 195	$ 206	$ 219
12.0%	$ 176	$ 183	$ 191	$ 201	$ 212	$ 226	$ 242
11.5%	$ 189	$ 197	$ 207	$ 219	$ 233	$ 250	$ 272
11.0%	$ 203	$ 214	$ 226	$ 241	$ 259	$ 281	$ 310
10.5%	$ 220	$ 233	**$ 248**	**$ 267**	**$ 291**	$ 321	$ 361
10.0%	$ 241	$ 257	**$ 276**	**$ 300**	**$ 332**	$ 374	$ 433

WACC \ g	4.5%	5.0%	5.5%	6.0%	6.5%	7.0%	7.5%
12.5%	-22%	-19%	-16%	-12%	-8%	-2%	4%
12.0%	-17%	-13%	-9%	-5%	1%	7%	15%
11.5%	-10%	-7%	-2%	4%	10%	19%	29%
11.0%	-4%	1%	7%	14%	23%	33%	47%
10.5%	4%	10%	**18%**	**27%**	**38%**	52%	71%
10.0%	14%	22%	**31%**	**42%**	**57%**	77%	105%

Source: Safra.

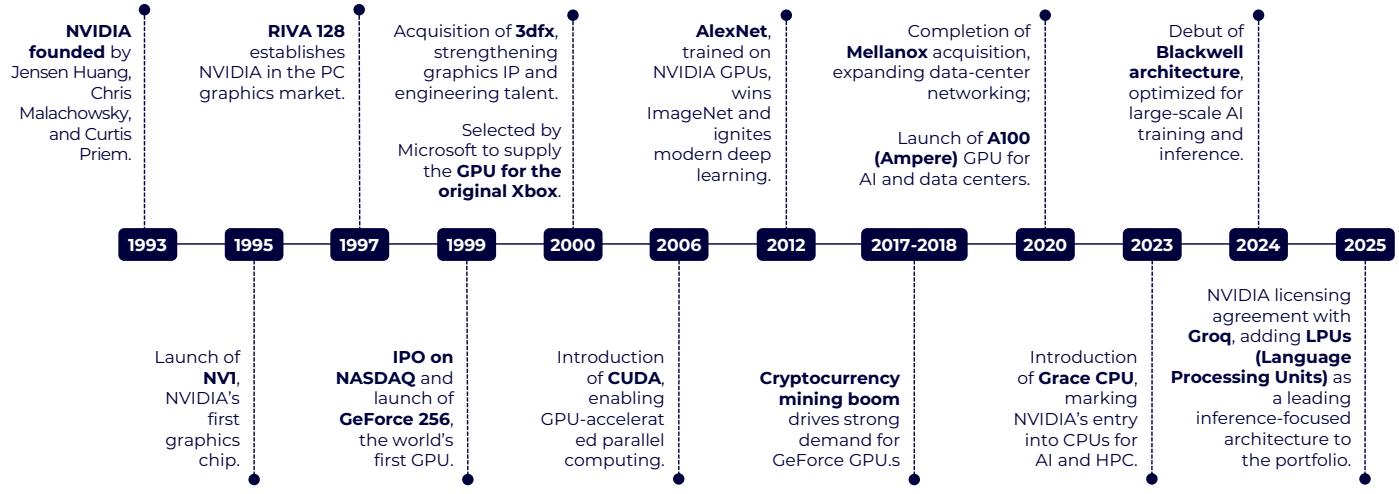
Company overview

NVDA is a semiconductor company founded in 1993 by visionary CEO Jensen Huang that designs graphics processing units (GPUs), central processing units (CPUs), networking chips, and the software that orchestrates them into full-stack AI infrastructure. The company operates a fabless model, outsourcing all fabrication to foundry and packaging partners and concentrating its full R&D investment (and ~3/4 of its ~42k workforce) on design and architecture, sustaining an annual cadence of generation iterations that no competitor has ever matched.

AI represents the most consequential platform shift in computing since the internet. Virtually every frontier AI model is built on the Transformer, a neural network architecture whose core operations (matrix multiplications across tens to hundreds of billions of parameters) are inherently parallel. GPUs, originally designed for rendering 3D graphics, pack thousands of cores capable of executing trillions of FLOPS, making them architecturally suited to exactly this type of workload. NVDA's GPUs are the natural hardware substrate for the AI ecosystem, and the company's platform is the default infrastructure on which that ecosystem runs.

Company history

Figure 60 – Timeline of major events



Source: Safra.

NVDA was founded on April 5, 1993, by Jensen Huang, Chris Malachowsky, and Curtis Priem. The three engineers conceived the company over coffee at a Denny's restaurant in East San Jose, where Huang had worked as a

dishwasher and busboy as a teenager. Their thesis was that the pace of computing would outstrip CPU capacity, creating demand for specialized processors to handle increasingly complex 3D graphics. The company was incorporated with USD 200 from each founder and subsequently raised USD 20mn in venture capital from Sequoia Capital, Sutter Hill Ventures, and others.

NVDA's first product, NV1 (1995), was a commercial failure. The chip used a proprietary rendering approach at a time when the industry was standardizing on a different method, and a parallel partnership with video game company Sega collapsed.

By late 1996, NVDA was within months of bankruptcy, and Huang cut the workforce from roughly 100 to 40. A pivotal negotiation with Sega's president yielded a USD 5mn contract buyout that provided a six-month lifeline.

That lifeline produced the RIVA 128 (1997), NVDA's first industry-standard chip. Despite supporting only a fraction of the specification, it outperformed competitors on raw speed and sold over 1mn units in 4 months, saving the company. The accelerated development process (nine months rather than the industry-standard two years) established a product cadence roughly twice as fast that of competitors, a structural advantage maintained ever since.

In January 1999, the company completed its IPO on NASDAQ, raising ~USD 42mn at a valuation of ~USD 625mn. That same year, NVDA introduced the GeForce 256. Graphics accelerator chips had existed since the early 1990s, but NVDA was the first to brand the category as a "GPU", a term the company coined to describe a processor that handled the entire graphics rendering pipeline on a single chip, rather than splitting the work with the CPU.

In 2000, NVDA acquired 3dfx, acquired key assets of 3dfx, a former pioneer in consumer graphics accelerators that had been losing share, consolidating its position as the dominant designer of discrete consumer GPUs.

In 2001, Microsoft selected NVDA to supply the GPU for the original Xbox, validating the company's position in the gaming ecosystem.

In 2006, NVDA released CUDA, a parallel computing platform that made the GPU programmable for general-purpose workloads beyond graphics. The company's insight was that the hundreds of millions of GeForce GPUs already installed in gaming PCs contained massively parallel processors that could be repurposed for scientific and industrial computation. By distributing CUDA for free, NVDA turned its gaming installed base into a general-purpose computing platform, deliberately at the expense of near-term profitability.

The scientific computing community adopted it first. Then, in 2012, a research team at the University of Toronto used NVDA GPUs to build AlexNet, a neural network that won the ImageNet image—recognition competition by a wide margin, demonstrating that GPUs could train AI models orders of magnitude faster than conventional processors. The result catalyzed the deep learning research community's adoption of CUDA as its default computing platform.

Over the next decade, CUDA evolved into a deep software stack encompassing CUDA—X acceleration libraries, AI models, APIs, SDKs, and domain-specific frameworks, accumulating over 7.5 million developers and native integration with every major AI framework.

The GPU's parallel processing capabilities also made it the hardware of choice for cryptocurrency mining during the proof-of-work booms that peaked in 2017-2018 and again in 2020-2021. Crypto miners purchased GeForce GPUs at scale, creating persistent retail shortages that priced out gamers and drew public scrutiny.

In 2020, NVDA acquired Mellanox Technologies for ~USD 7bn, adding high-speed data center networking and enabling the company to architect systems at data center scale, a move that proved important as AI models grew to require thousands of interconnected GPUs. The A100 GPU (Ampere), also launched in 2020, became the workhorse of the initial AI training wave that followed the release of ChatGPT in late 2022.

In the years since, NVDA has extended the platform across multiple dimensions. The Grace CPU (2023) moved the company beyond GPUs and into the market for the chips that manage data flow and orchestrate workloads within AI servers.

In 2024, Blackwell architecture marked the shift from selling individual servers to selling entire racks, complete liquid-cooled cabinets of 72 GPUs functioning as a single computing unit.

In 2025, a licensing agreement with Groq added LPUs to the portfolio. NVDA packaged this technology into the standalone Groq 3 LPX rack — a 256-LPU system co-deployed with Vera Rubin NVL72 to accelerate latency-sensitive inference workloads, with shipments starting 2H26.

Management now describes the data center itself as an "AI factory," a framework detailed through this report.

Revenue segmentation

NVDA's revenue is disclosed across four market platforms — Data Center (~90% of revenue in FY'26), Gaming, Professional Visualization, and Automotive — plus a residual OEM & Other category. For financial reporting, these are consolidated into two segments: Compute & Networking, encompassing Data Center, Automotive, and AI Enterprise software; and Graphics, encompassing Gaming and Professional Visualization

Data Center revenue segment

The Data Center platform accelerates compute-intensive workloads (AI training and inference, data analytics, scientific computing) in cloud, hyperscale, on-premises, and edge data centers. Revenue is booked across two categories: **compute** (GPUs, CPUs, LPUs from 2H26 onwards, and integrated systems such as DGX, HGX, and NVL72 racks) and **networking** (NVLink, NVSwitch, Spectrum-X, Quantum, ConnectX, BlueField). The offering has four layers (compute, systems, networking, and software) that together constitute a vertically integrated infrastructure stack, with NVDA designing every component and optimizing them as a co-designed architecture.

The GPU is the foundation of the Data Center platform. NVDA has released a new GPU generation roughly every year, each generation enabling AI workloads that the prior one could not serve at the required speed or scale.

The **A100 (Ampere, 2020)** was the GPU on which the generative AI era was built. It was also the first data center accelerator to offer MIG (Multi-Instance GPU), a feature that allows a single physical chip to be partitioned into up to seven smaller virtual GPUs, each running independent workloads.

The **H100 (Hopper, 2022**) was designed for the Transformer architecture, the type of neural network that underpins all modern LLMs. It introduced the Transformer Engine, a hardware feature that automatically adjusts numerical precision during computation to maximize speed without sacrificing accuracy, delivering up to 9x faster training and 30x faster inference versus the A100. The **H200** (2024) upgraded Hopper's memory without changing the architecture.

The current generation is **Blackwell**, which introduced two innovations. The **B200 GPU** uses a two-die design (two silicon chips connected by a high-speed bridge on a single package) that doubles compute density. Blackwell also added native support for 4-bit numerical precision (FP4), meaning each number the chip processes takes up half the memory of the prior generation's 8-bit format, allowing larger AI models to fit on fewer chips and run faster.

Networking has evolved from a peripheral component to a co-equal pillar of the Data Center platform. AI workloads at scale are as constrained by communication speed as by computation. Scale-up networking connects GPUs within a server or rack. **NVLink** is NVDA's proprietary high-speed interconnect, and **NVSwitch** is the chip that implements it. Together, they allow GPUs to share memory and operate as a unified system. Scale-out networking connects racks across the data center. **ConnectX** is a network interface card installed on each server. **Spectrum-X** is an Ethernet switching platform optimized for AI traffic, while **Quantum** (InfiniBand) serves the highest-performance training clusters.

The software layer converts NVDA hardware into a programmable platform. CUDA, released in 2006, is the programming model that makes NVDA GPUs programmable in standard languages. The ecosystem includes over 7.5 million developers, over 400 acceleration libraries, and native integration with PyTorch, TensorFlow, and JAX.

NVIDIA AI Enterprise is the commercial software offering, sold as a per-GPU annual subscription. It bundles **NVIDIA Inference Microservices** (NIM, optimized inference microservices), **NeMo** (tools for data curation, fine-tuning, and reinforcement learning), and **AI Blueprints** (pre-built templates for enterprise AI agent workflows).

Nemotron is NVDA's family of open-weight AI models (models whose parameters are publicly available for anyone to use and modify) designed to run on NVDA hardware. They serve a strategic function: seeding the ecosystem with NVDA-optimized workloads, driving hardware adoption without requiring dependence on a single proprietary model provider. **NemoClaw**, announced at GTC 2026, is NVDA's open-source enterprise stack for OpenClaw, the autonomous agent framework that has emerged as a default platform for deploying AI agents. NemoClaw bundles Nemotron models with a secure runtime and privacy controls, positioning NVDA as the infrastructure layer beneath autonomous agents.

Data Center go-to-market strategy. NVDA sells its Data Center compute and networking at distinct levels of integration, each targeting a different customer segment and capturing progressively more of the system economics.

At the component level, **HGX** is an 8-GPU baseboard sold to **OEMs** (i.e., companies like **Dell**, **Supermicro**, **HPE**, **Lenovo**, all Not Covered) who integrate it into their own server chassis — the highest-volume channel, dominant in enterprise and mid-market deployments.

At the system level, **DGX** is NVDA's own 8-GPU server, pre-integrated with CPUs, networking, cooling, and software – capturing integration margin that would otherwise flow to OEMs, and targeting AI-first enterprises and research labs.

At the rack level, **NVL72** is NVDA's top-tier, liquid-cooled rack system that links 72 GPUs and 36 CPUs so they operate as a single logical system — enabling training and inference of trillion-parameter LLMs that cannot fit on smaller configurations. Distribution is hybrid: NVDA sells directly as DGX NVL72, OEMs ship their branded versions, and hyperscalers source directly from Taiwanese manufacturers. However, across all channels, NVDA captures the vast majority of the rack's component cost. Current generation is **GB200**/**GB300** NVL72.

Figure 61 – Live deployment of NVIDIA's GB200 NVL72 systems

Source: CoreWeave.

Gaming revenue segment

GeForce is NVDA's consumer GPU brand, serving over 200 million gamers and creators worldwide. The product line includes GeForce RTX GPUs for desktop and laptop PCs, and GeForce NOW, a cloud gaming subscription service that streams PC games from NVDA's data centers to low-powered devices. NVDA holds a dominant position in discrete PC GPUs (graphics chips sold as a separate component rather than integrated into the main processor).

The current generation, the GeForce RTX 50 Series (Blackwell, launched January 2025), introduced neural graphics and the latest generation of DLSS, an AI technology that uses a neural network to generate high-quality frames at lower computational cost, boosting performance without sacrificing visual quality.

Gaming is NVDA's original franchise and the business that funded the company's expansion into AI. Beyond gaming, Tensor Cores embedded in every RTX chip increasingly enable generative AI applications that run locally on the user's PC (chatbots, image generators, code assistants).

Figure 62 – Example of NVIDIA's GeForce RTX 5090 Ti, designed for high performance in games with ray tracing and AI applications

Source: NVIDIA.

Figure 63 – DLSS 5 enables real-time AI enhancements for photorealistic gaming graphics

Source: Safra, NVIDIA.

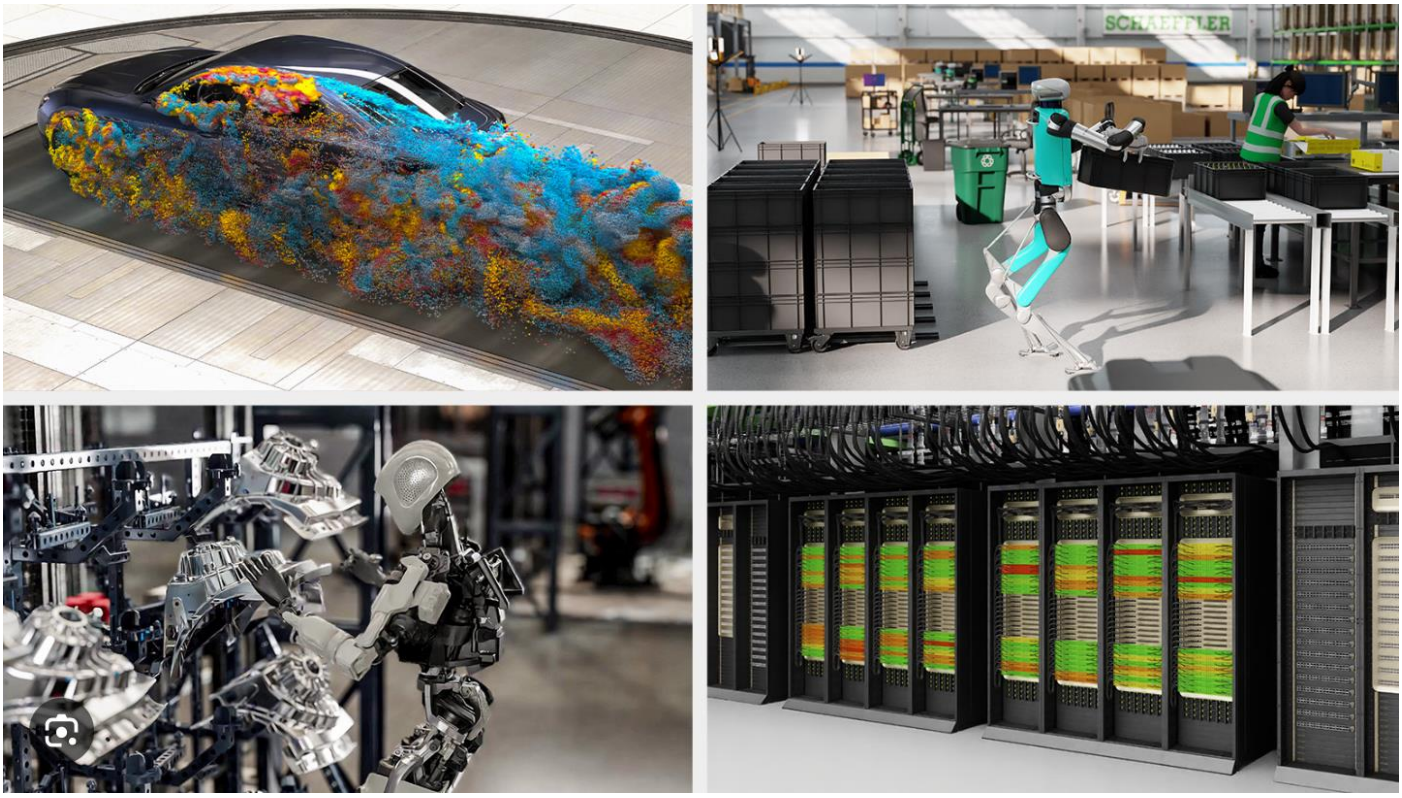
Professional Visualization revenue segment

NVIDIA's RTX PRO GPUs (formerly Quadro) serve designers, engineers, architects, and digital content creators with enterprise-grade reliability and features optimized for visual accuracy.

Omniverse is a simulation and digital twin platform that enables enterprises to build photorealistic virtual replicas of physical assets such as factories, warehouses, and road networks, and run AI models inside these virtual environments to test outcomes before committing resources in the real world.

The segment is small today but carries an optionality. Omniverse was initially designed for designers and engineers, but the same simulation capabilities are being increasingly adopted by manufacturers testing factory layouts virtually, logistics companies optimizing warehouse operations digitally, and autonomous vehicle developers running virtual driving miles.

Figure 64 – NVIDIA Omniverse: simulation and digital twin platform to model, test, and optimize real-world assets with AI



Source: Safra, NVIDIA.

Automotive revenue segment

NVIDIA DRIVE is a platform for self-driving vehicles and intelligent cockpits, spanning the chip inside the car, the software that runs on it, and the cloud infrastructure where driving algorithms are developed and tested in simulation.

Robotaxi fleets (taxis that operate without a human driver) represent one of the first large-scale commercial deployments of physical AI, a domain in which NVDA supplies the full-stack from training infrastructure to in-vehicle compute. NVDA's DRIVE platform is positioned to power most other OEM self-driving programs. Over 40 customers are building on NVDA's platform, including 20 of the top 30 electric vehicle makers.

Figure 65 – End-to-end platform for developing, simulating, and deploying autonomous driving systems.

Figure 66 – A global partner ecosystem of automakers, suppliers, and AI companies building on NVIDIA DRIVE platform

Source: NVIDIA.

OEM & Other revenue segment

A residual category that includes entry-level discrete GPUs for notebooks and legacy products for game consoles. In 2022, the segment also included revenue from dedicated cryptocurrency mining processors (CMPs), a product line NVDA created to separate crypto mining demand from the Gaming segment after miners co-opted GeForce GPUs and created persistent retail shortages. CMP revenue became negligible after Ethereum's transition from proof-of-work to proof-of-stake in late 2022 eliminated the primary use case.

Management

Figure 67 – Management

Founders

Jensen Huang
Founder, President and CEO

Chris A. Malachowsky
Founder and NVIDIA Fellow

Company Officers

Colette Kress
EVP and Chief Financial Officer

Jay Puri
EVP, Worldwide Field Operations

Debora Shoquist
EVP, Operations

Tim Teter
EVP, General Counsel and Secretary

Source: Safra.

NVDA has been led by co-founder Jensen Huang since its inception in 1993, making it one of the longest continuously founder-led companies in the technology industry.

Jensen Huang studied electrical engineering at Oregon State and Stanford, worked at AMD and LSI Logic, and co-founded NVDA at 30. He maintains a flat organizational structure with no COO and dozens of direct reports spanning every engineering domain. Co-founder **Chris Malachowsky** remains on the executive staff as an NVIDIA Fellow advising on long-term technology direction. Third co-founder **Curtis Priem** departed in 2003 but remains recognized for his role in the company's early chip design.

Colette Kress has served as EVP and CFO since 2013, bringing prior experience from Microsoft and Cisco, and has overseen capital allocation through the company's most transformative period. **Jay Puri**, EVP of Worldwide Field Operations, leads global sales across hyperscalers, enterprise, sovereign, and OEM channels. **Debora Shoquist**, EVP of Operations, manages the global supply chain and procurement. **Tim Teter**, EVP and General Counsel, oversees legal, IP, and export control compliance.

The **Board of Directors** comprises 13 members elected annually: 12 independent directors and Huang as the sole management representative.

Compensation and alignment. The compensation structure is designed to align management with long-term shareholders. The largest component of executive pay is performance-based equity vesting over four years, tied to EBIT targets and relative total shareholder return versus S&P 500.

100% of Huang's equity grants are performance-based, and when the company has missed targets, cash bonuses have been zero and performance equity has not vested. Huang kept his base salary unchanged for a full decade before a modest increase in FY'25A, and his personal shareholding of ~3.5% dwarfs his annual compensation by several orders of magnitude, meaning his wealth compounds or declines alongside every other shareholder. The company has cultivated a broad-based equity culture, with stock-based compensation forming a foundational element of pay at all levels.

Shareholder structure

NVDA employs a single class share structure of one share/one vote, meaning no individual (including the founder) holds outsized voting control relative to their economic stake. This distinguishes NVDA from technology peers, where dual-class structures grant founders majority voting power with minority economic ownership.

Institutional investors hold ~70% of shares, predominantly through passive index funds (reflecting NVDA's weight in the S&P 500 and NASDAQ 100). The largest holders are Vanguard (~9.3%), BlackRock (~8%), State Street (~4%), and Fidelity (~4%), alongside Jensen Huang's personal stake (~3.5%).

In practice, Huang exercises strategic influence well beyond his voting interest. He has led NVDA without interruption, maintains a flat organization that concentrates decision-making, and no succession plan has been publicly disclosed. The institutional base, being overwhelmingly passive, has limited history of activist engagement at this scale. The governance structure is best understood as founder-led with passive institutional oversight, with Huang's authority deriving from tenure and institutional knowledge.

Risks

Hyperscaler customer concentration and capex cycle deceleration. NVDA's Data Center segment (~90% of revenue) derives ~60% of revenue from five hyperscalers (Google, Amazon, Microsoft, Meta, Oracle, all Not Covered). While we believe the AI buildout is going to be durable, a material deceleration in aggregate capex given potential macro headwinds or slower downstream AI monetization would compress the forecast. Given customer concentration, even a single hyperscaler recalibrating capex would have non-trivial impact on our revenue outlook. The same concentration is reflected on the balance sheet, with accounts receivable scaling alongside ticket size of racl-scale orders, potentially exposing NVDA to quarterly variance in cash conversion should a top customer stretch payment terms.

Competition from merchant GPU alternatives. NVDA currently holds ~90% of the merchant AI GPU market, with **AMD** (Not Covered) as the principal alternative. While we believe the company will maintain its industry-leading position, the scaling of AMD's MI300X/MI350 platforms and the launch of the MI450-based Helios rack-scale system represent competitive pressure. We view outright displacement as low probability given NVDA's integrated platform moat and the CUDA ecosystem, but share loss through second-sourcing is a genuine risk: hyperscalers managing supplier concentration may allocate a share of next-cycle compute to AMD as a hedge.

Competition from custom silicon (ASICs). Hyperscalers have scaled internal ASIC programs (Google's TPU, AWS's Trainium and Inferentia, Microsoft's Maia, Meta's MTIA), with Broadcom (Not Covered) capturing adjacent share through custom-silicon partnerships. Extending beyond GPUs and including ASICs, we estimate that NVDA's share of the broader AI accelerator market sits at ~78% today, and we assume in our forecast that this share drifts toward ~72% by FY'31E as ASICs scale. Unlike the merchant-GPU risk, where competitive pressure comes from a third-party supplier, ASIC competition is an insourcing dynamic: the customer becomes the competitor.

We view a meaningful ASIC transition as improbable, requiring multi-year execution and carrying significant opportunity cost versus continuing to scale on NVDA's platform, but if any single hyperscaler transitioned a meaningful portion of workloads to internal silicon in outer years, the share impact could compound with the merchant-GPU risk above.

Supply chain concentration and geopolitical exposure to Taiwan. NVDA depends on **TSMC** (Not Covered) as the sole foundry for leading-edge wafers and CoWoS advanced packaging, a binding constraint across the AI buildout. HBM is concentrated in **SK Hynix** (Not Covered), with **Samsung** and **Micron** (both Not Covered) in secondary positions. Disruption at either (natural disasters, yield or ramp delays) would materially affect shipment capacity in the near term, as alternate sources cannot replicate leading-edge or packaging capacity at scale. Furthermore, the concentration of leading-edge fab capacity in Taiwan represents an ongoing geopolitical tail risk that does not have a short-term remedy, even as TSMC's US fab expansion (Arizona, first 3nm output expected in 2028E) gradually diversifies the footprint.

Export controls and the "chip war". NVDA no longer embeds China Data Center compute revenue in its forward guidance, following US export controls that reduced the company's China AI GPU share (China previously represented ~20%+ of Data Center revenue), and our forecast similarly assumes no material China contribution. The broader regulatory tail risk, however, remains meaningful: a "chip war" expansion into other sovereign AI destinations (tighter licensing requirements or restrictions) could compress the sovereign AI pipeline. In parallel, China's domestic stack could advance faster than consensus appreciated, with Huawei's Ascend series and open-weight Chinese models potentially diffusing into emerging markets as a default alternative to the NVDA/CUDA stack.

Power availability and grid constraints. The AI buildout is increasingly bottlenecked by power delivery rather than silicon. Hyperscaler capacity plans now routinely cite multi-year grid interconnection queues, constrained substation capacity, and long-dated power purchase agreements as the binding constraints on deployment timing. If meaningful GW of planned capacity slips because grid interconnection cannot be secured on the stated timeline, NVDA's revenue would be delayed (rather than lost), even if the underlying demand signal remains intact.

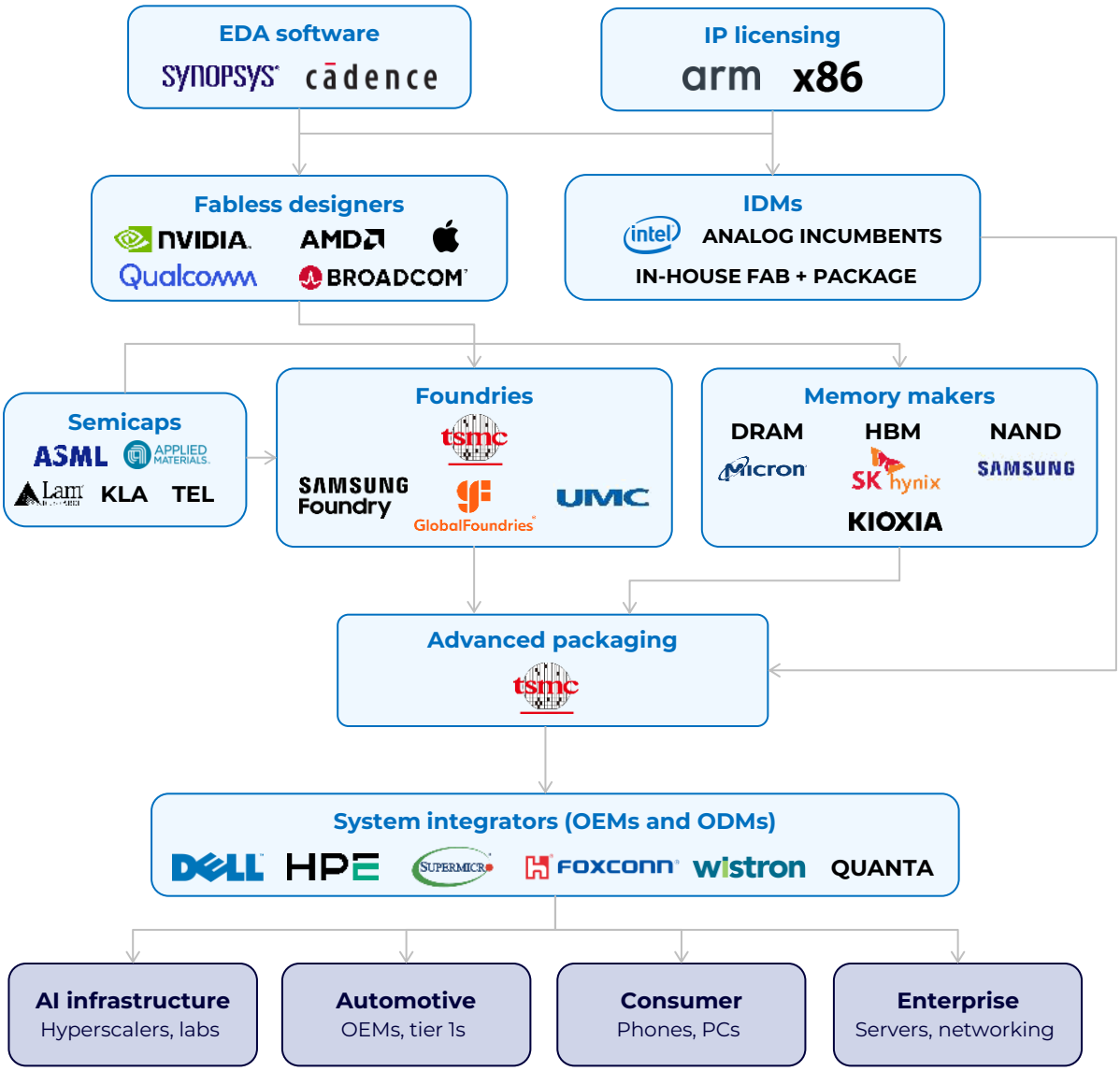
Inventory and generation-transition risk. NVDA's annual generation cadence creates natural pressure points at each transition, as customers may opt to defer purchases of the outgoing architecture ahead of the incoming one. If this dynamic plays out more sharply than expected at the Blackwell-to-Rubin handoff in 2H26, NVDA would face inventory exposure and short-term gross margin pressures.

Key-person risk. NVDA has been led by co-founder Jensen Huang since 1993. Huang operates a flat organizational structure with dozens of direct reports, with no publicly disclosed succession plan. Any unplanned departure or health event would be materially destabilizing to the company's strategic direction and investor confidence, particularly given the extent to which NVDA's platform strategy, customer relationships, and execution cadence have been shaped by Huang's personal leadership.

Appendix A: Semiconductors industry — a supply chain primer

The semiconductor supply chain is the most intricate industrial ecosystem ever constructed. Understanding where value is created within it is the foundation of any investment thesis in the sector.

Figure 68 – Overview of the semiconductor supply chain



Source: Safra.

Overview

A single leading-edge chip may cross international borders more than seventy times before reaching the end customer, passing through design software developed in California, intellectual property licensed from Cambridge, lithography equipment built in the Netherlands, silicon wafers refined in Japan, fabrication in Taiwan, memory

stacked in South Korea, advanced packaging back in Taiwan, and final assembly in Southeast Asia. **No other industry combines this degree of capital intensity and geographic dispersion.**

Given its strategic significance, the semiconductor supply chain's geopolitical implications have escalated materially since COVID. US export controls have restricted China's access to the most advanced equipment and chips, while governments globally have committed resources to reshore or de-risk semiconductor capacity. These programs are simultaneously subsidizing and fragmenting the chain along geopolitical lines, and the companies positioned at the chokepoints of this fragmented chain have been among the most strategically important industrial enterprises in the world.

What also makes the chain particularly unusual is that its extreme specialization produces a self-regulating pricing dynamic. Each segment depends on the health of the segments above and below it, and this creates a *de facto* equilibrium in which the dominant companies earn extraordinary returns but refrain from predatory extraction that would destabilize the ecosystem.

TSMC (Not Covered) could raise wafer prices far beyond current levels given its monopoly at leading-edge nodes, but doing so would impair its largest customers, whose volume growth is what funds TSMC's reinvestment cycles. **ASML** (Not Covered) could price its lithography equipment more aggressively given its monopolistic position, but its installed base economics and service revenue depend on fabs remaining fully utilized. **SK Hynix** (Not Covered) could extract scarcity rents on the memory that accompanies every AI accelerator, but depends on sustained volumes from its largest buyer to justify the capacity expansion that locks in its position.

The result is a multi-dependency cascade, where each player earns returns commensurate with its value added, while preserving its counterparties' incentives to keep buying.

Where the chips come from: the design and manufacturing model

The semiconductor industry is divided into two fundamentally different business models.

Historically, chip companies have done everything in-house: they designed the products and operated the fabrication facilities that manufactured them. This is the **Integrated Device Manufacturer (IDM)** model, and it dominated the industry from the 1960s through the 1990s. **Intel** (Not Covered) was the archetype, and for decades its manufacturing capability was the source of its competitive advantage.

The problem with the IDM model is that it is extraordinarily capital-intensive. A single leading-edge fab today can cost USD 40bn+ to build, and the rate of obsolescence is such that the equipment inside must be replaced or upgraded every few years to remain competitive. Very few companies can sustain investments at that scale, and those that fail to keep pace fall permanently behind.

The alternative model, pioneered in the late 1980s, is the fabless model. Fabless companies design chips but own no manufacturing facilities. They outsource fabrication, packaging, and testing to third parties, which allows them to concentrate their resources on architecture, software, and the ecosystems that make their chips valuable to customers.

This model was enabled by the emergence of dedicated foundries, companies whose sole business is to manufacture chips for others. The foundry model was pioneered by **TSMC**, founded in Taiwan in 1987 by Morris Chang. Chang's insight was that a manufacturing partner that never competed with its own customers by designing chips could unlock an entire generation of design-focused companies. The structural alignment between TSMC and its customers, rooted in the fact that TSMC has no proprietary chip business that might compete with theirs, is what allowed fabless designers to trust TSMC with their most sensitive next-generation designs.

NVIDIA sits at the center of the fabless model, alongside virtually every other company that has defined the modern semiconductor industry: **AMD** (Not Covered), **Apple** (Not Covered), **Broadcom** (Not Covered), and **Qualcomm** (Not Covered) all design chips and route them to TSMC for fabrication. The hyperscalers that have begun designing custom AI silicon all follow the same model, typically partnering with **Broadcom** or **Marvell Technology** (Not Covered) on design and fabricating at TSMC.

The migration from IDM to fabless is so comprehensive that Intel, the last meaningful IDM outside of memory, has spent the last five years attempting to convert itself into a foundry competitor to TSMC, with so far uneven results.

The products: logic, analog, and memory

The industry produces three fundamentally distinct categories of chip, and each sits in a different place within the investment universe.

Logic chips process information. They are the processors that execute software, perform calculations, and run the computations that underpin everything from smartphones to data centers. The most important logic chips in the current cycle are the GPUs and custom AI accelerators that power AI training and inference workloads.

Logic is where the leading edge of process technology matters most, because every shrink in transistor dimensions delivers more computational performance per watt of energy consumed. This is why logic chipmakers compete fiercely for access to the most advanced manufacturing capacity, and why the chokepoint at TSMC's leading-edge nodes has become the defining supply constraint of the AI era.

Analog chips interface with the physical world. They convert continuous real-world signals such as voltage, current, temperature, light, sound, and motion into digital data that logic chips can process, and they translate digital outputs back into the physical actions that drive motors, displays, and communications equipment.

Analog is a structurally different business from logic. It is less dependent on leading-edge manufacturing because the physics of voltage handling and signal fidelity are actually better behaved at older, more mature manufacturing nodes.

The analog market is dominated by **Texas Instruments** (Not Covered), **Analog Devices** (Not Covered), and **Infineon** (Not Covered), serving automotive, industrial, and communications end markets.

Memory chips store information, and the category is divided into two technologies with meaningfully different economics.

DRAM provides the working memory that processors access in real time, and its most important AI-relevant variant is high-bandwidth memory, known as **HBM**. HBM stacks multiple DRAM layers vertically and positions them directly adjacent to the GPU on the same package, which allows for the extraordinary data throughput that AI workloads require. HBM costs roughly three times as much per gigabyte as standard DRAM and has become the single largest component of an AI accelerator's bill of materials. The DRAM market is a consolidated triopoly dominated by **Samsung** (Not Covered), **SK Hynix**, and **Micron** (Not Covered).

SK Hynix holds an estimated 50%+ share of current-generation HBM and has become the primary supplier to NVIDIA for the GPUs that underpin the AI buildout. All three suppliers are running at full capacity, and HBM allocation has been one of the defining supply constraints of the cycle.

The second memory technology, **NAND** flash, is non-volatile and provides the storage medium for solid-state drives, smartphones, and data center storage arrays. NAND follows its own cyclical pattern driven primarily by consumer electronics and enterprise storage demand rather than by AI, and the supplier base is broader than DRAM.

The enablers: design software and intellectual property

Before a chip can be manufactured, it must be designed, and modern chip design is impossible without two layers of specialized infrastructure.

The first is **Electronic Design Automation (EDA)**, which is the software on which every semiconductor is conceived, simulated, and verified before it is sent to a foundry for manufacturing. EDA tools are to chip designers what computer-aided design software is to architects, with the difference that a modern chip contains tens of billions of transistors and thousands of kilometers of interconnect wiring compressed into a few square centimeters.

The complexity is such that the market has consolidated around two companies. **Synopsys** (Not Covered) and **Cadence** (Not Covered) together account for ~80% of global EDA revenue, with **Siemens** (Not Covered) a smaller third. Switching costs are measured in years of requalification, because foundries develop their manufacturing specifications in close collaboration with EDA vendors, and a chip designed with one vendor's tools cannot be trivially ported to another's.

The second layer is **intellectual property (IP) licensing**, and the most consequential IP in the industry is the instruction set architecture, or **ISA**. The ISA is the fundamental language that defines how software communicates with a processor, and two ISAs dominate modern computing.

The first is **x86**, developed by Intel in 1978 and later extended by AMD, which has powered virtually every personal computer and data center server for four decades. The second is **ARM**, designed by **ARM Holdings** (Not Covered) of Cambridge, UK, which took the opposite approach by optimizing for power efficiency rather than raw performance.

ARM's architecture became dominant in mobile devices because every smartphone chip requires low power consumption, and it is now expanding rapidly into the data center where power constraints have become a binding limit on AI infrastructure growth. NVIDIA's Grace CPU is built on ARM, as are Amazon's Graviton, Microsoft's Cobalt, and the custom server processors increasingly deployed by hyperscalers.

ARM licenses its architecture and pre-designed processor cores to an ecosystem of licensees and earns revenue through upfront fees and per-unit royalties, which makes it one of the highest-margin businesses in the semiconductor value chain.

The equipment that makes the chips

Semiconductor fabrication involves hundreds of discrete manufacturing steps, each performed by specialized equipment. The companies that make this equipment (collectively known as semicaps) sit at one of the most concentrated chokepoints in the entire supply chain.

In most equipment sub-segments, a single vendor holds more than half of the global market, and in the single most important category, the position is effectively monopolistic. That category is lithography, the process by which the circuit patterns that define a chip's functionality are projected onto the silicon wafer.

Lithography matters because the resolution of the pattern determines how small the transistors can be, and smaller transistors mean more computational performance for the same amount of silicon and energy.

The industry used a particular type of ultraviolet light for decades, but by the mid-2010s the physics of that approach had reached its practical limits. The transition to extreme ultraviolet lithography (EUV) resolved this constraint and enabled the current generation of leading-edge chips.

ASML is the sole manufacturer of EUV equipment in the world. A single EUV machine costs between USD 350mn and USD 400mn, takes more than a year to build, and requires such precision that its components must be assembled in a vacuum. ASML's order backlog and delivery schedules function as a leading indicator of global semiconductor supply 6 to 18-months forward, because every new leading-edge fab depends on receiving ASML tools on schedule to ramp up production.

Beyond lithography, the rest of the equipment ecosystem is dominated by a handful of specialized suppliers. **Applied Materials** (Not Covered) leads in the equipment that deposits thin films of material onto silicon wafers layer by layer. **Lam Research** (Not Covered) specializes in the equipment that carves patterns into those deposited layers through chemical etching, with a particularly strong position in memory manufacturing. **KLA** (Not Covered) dominates inspection equipment, the quality—control layer that detects microscopic defects at each step of the production process, and has accumulated a proprietary database of defect signatures built over decades that creates high switching costs. **Tokyo Electron** (Not Covered) of Japan and **ASM International** (Not Covered) of the Netherlands complete the major semicaps ecosystem.

Collectively, these five companies plus ASML form what is effectively an oligopoly that sits upstream of every chip made in the world, and their quarterly earnings and order books are among the most reliable independent signals of semiconductor industry activity.

Foundries: where the chips are actually made

Once a chip has been designed and the manufacturing equipment has been procured, the chip must actually be fabricated, and this is where TSMC's dominance becomes the defining feature of the modern industry.

TSMC commands ~70%+ of global foundry revenue and an estimated 90%+ share of leading-edge production at the most advanced manufacturing nodes. Every major AI accelerator, every Apple mobile processor, every AMD CPU and GPU, and every major hyperscaler custom chip is fabricated at TSMC.

TSMC's dominance rests on reinforcing advantages:

The first is its process technology leadership. TSMC is currently manufacturing at the 3nm node in volume and ramping a more advanced 2nm process for production in 2026. Each generation of process technology enables chip designers to either pack more computational capability into the same area or deliver the same capability at lower power consumption, and TSMC has consistently achieved better manufacturing yields at each new node than any competitor.

Higher yield means more functional chips per wafer, which translates directly into lower effective cost per chip for the customer, which reinforces TSMC's share at each subsequent generation.

The second advantage is advanced packaging. Modern AI accelerators are not single chips but integrated modules that combine a GPU die, several stacks of HBM memory, and specialized interconnects, all assembled onto a silicon platform that provides the high—speed communication pathways between them.

TSMC's proprietary packaging technology, known as CoWoS, has become the binding supply constraint for AI accelerator production globally, so tightly constrained that even when wafer fabrication capacity is available, a shortage of CoWoS capacity independently bottlenecks system shipments to customers.

The third advantage is TSMC's concentrated exposure to high-performance computing, which is the fastest-growing end market in the industry. Revenue from AI accelerators funds the next generation of TSMC's process and packaging investment, which enables the next generation of AI chips, creating a virtuous cycle that continuously widens TSMC's lead over competitors.

Samsung Foundry is TSMC's closest competitor and the only alternative that currently offers leading-edge manufacturing to external customers. Samsung has made technical progress, but its structural challenge is that it is a division of the same company that operates competing chip businesses in mobile processors, memory, and image sensors.

When a fabless designer sends its most sensitive next-generation design to a foundry, it trusts that the foundry will not use that knowledge to benefit competing products, and this trust is harder to establish when the foundry shares corporate walls with potential competitors. Qualcomm has shifted production away from Samsung toward TSMC on multiple occasions, citing yield and delivery concerns, but the structural misalignment compounds the technical disadvantage and has limited Samsung's ability to capture leading-edge customers.

Intel represents the most consequential open question in the foundry landscape. For decades, Intel was the uncontested leader in both chip design and manufacturing, and its fabs were consistently first to each new transistor node. That leadership broke during the 10nm manufacturing transition in the mid—2010s, when repeated delays allowed TSMC to pull ahead on process technology and AMD, by fabricating at TSMC, closed the performance gap in data center CPUs. Intel's server CPU market share, once above 95%, has declined materially.

Under CEO Pat Gelsinger from 2021 to 2024, Intel separated its design and manufacturing organizations, established Intel Foundry as a standalone business unit to compete for external customers, and committed to an aggressive process roadmap to regain parity with TSMC. Gelsinger was replaced by Lip-Bu Tan in March 2025, who has continued the foundry strategy alongside additional operational restructuring.

Execution has been uneven, external foundry customer wins have been limited, and Intel's most advanced process remains unproven at volume. If Intel's foundry revival succeeds, the global manufacturing landscape may shift from a TSMC near-monopoly to a more competitive market with geopolitically diversified supply. If it does not, TSMC's dominance at the leading edge will further consolidate.

Below the leading edge, the foundry market opens up considerably. GlobalFoundries (Not Covered) and UMC (Not Covered) specialize in mature manufacturing nodes serving automotive, industrial, and communications customers where leading—edge density is unnecessary and cost per wafer is the primary consideration.

The buyers: where demand comes from

The final layer of the chain is the end demand that pulls chips through the supply ecosystem. The composition of that demand is what determines which segments of the industry enjoy structural tailwinds and which face cyclical pressures.

AI infrastructure is the single most important incremental demand driver in the current cycle, flowing predominantly into leading-edge logic chips and HBM memory.

Hyperscale cloud providers purchase GPUs, custom ASICs, and networking for the training and inference workloads that underpin their AI services. Frontier AI labs procure capacity both through their hyperscaler relationships and through direct agreements with specialized cloud providers (known as neoclouds). Sovereign AI programs also add government-funded compute demand that is driven by strategic rather than commercial imperatives. Server OEMs such as **Dell** (Not Covered), **HPE** (Not Covered), **Lenovo** (Not Covered), **Supermicro** (Not Covered), and **Foxconn** (Not Covered) assemble the finished physical systems that house all of this silicon.

Automotive and industrial demand is the primary driver of analog and power semiconductors, and its magnitude has grown substantially as vehicles have become more electrified. This demand flows predominantly through the analog IDMs and more mature—node foundries. However, as autonomous driving technology progresses, leading-edge logic nodes are becoming an increasingly material channel of silicon demand within the automotive end-market.

Consumer electronics drives demand for mobile application processors, discrete GPUs, and commodity DRAM and NAND, and sits on a cycle largely independent of AI capex. Smartphones, PCs, and gaming consoles collectively represent the largest volume segment of the industry by unit count, and their demand cycle is driven by cyclical replacement rates and device upgrade cycles.

Finally, enterprise demand taps both the traditional and AI-accelerated worlds, with CPU-based compute flowing through **Intel** and **AMD** and AI-accelerated workloads flowing through NVIDIA and hyperscalers' ASICs.

Appendix B: Glossary

Accelerated computing. Computing paradigm in which specialized processors (GPUs, ASICs, LPUs, or other accelerators) handle workloads that general-purpose CPUs cannot efficiently execute — particularly parallel workloads such as AI training and inference, data analytics, scientific simulation, and graphics rendering.

Accelerator. Any specialized processor designed to execute a specific class of workload more efficiently than a general-purpose CPU. In the AI context, includes GPUs, TPUs, Trainium, Inferentia, Maia, MTIA, LPUs, and other ASICs.

Agentic AI. AI systems designed to autonomously plan, reason, and execute multi-step tasks by invoking external tools, verifying outputs, and iterating until a goal is achieved. Consumes substantially more compute per interaction than single-prompt chatbot models.

AI Enterprise (NVIDIA AI Enterprise). NVIDIA's enterprise software suite, sold as a per-GPU annual subscription, bundling CUDA-X acceleration libraries, NVIDIA Inference Microservices (NIM), optimized AI frameworks, and enterprise support.

AI factory. Framework used by NVIDIA management describing AI data centers as facilities that convert electricity into monetizable tokens through computation, analogous to how a traditional factory converts raw materials into finished goods.

AI lab. Frontier research organizations developing state-of-the-art AI models (e.g., OpenAI, Anthropic, Google DeepMind, xAI). Typically purchase compute capacity through hyperscaler relationships and direct agreements with NCPs.

AlexNet. Neural network built by a University of Toronto research team in 2012, trained on NVIDIA GPUs, that won the ImageNet image-recognition competition by a wide margin. Catalyzed the AI research community's adoption of GPUs and CUDA as the default compute platform for deep learning.

Amdahl's Law. Principle stating that the maximum speedup of a system is limited by its slowest component. In AI data centers, even large improvements in GPU compute yield little benefit if interconnect, memory, or networking are bottlenecks — which underpins NVIDIA's "extreme co-design" strategy.

Ampere. NVIDIA's 2020 GPU architecture (A100), the data center accelerator on which the generative AI era was built. First architecture to offer MIG (Multi-Instance GPU).

Analog chips. Semiconductors that interface with the physical world by converting continuous real-world signals (voltage, current, temperature, light, sound, motion) into digital data, and vice versa. Dominated by Texas Instruments, Analog Devices, and Infineon; typically manufactured at mature process nodes.

API (Application Programming Interface). Programmatic interface through which software applications access services, including AI inference endpoints exposed by model providers.

ARM. Instruction set architecture designed by ARM Holdings of Cambridge, UK, optimized for power efficiency. Dominant in mobile devices and increasingly deployed in data centers (NVIDIA Grace, Amazon Graviton, Microsoft Cobalt). ARM licenses its architecture and pre-designed cores to an ecosystem of licensees.

ASIC (Application-Specific Integrated Circuit). Chip designed for a specific application or workload rather than general-purpose computing. In the AI context, typically refers to hyperscaler-internal accelerators such as Google's TPU, AWS's Trainium and Inferentia, Microsoft's Maia, and Meta's MTIA.

ASML. Dutch company and sole global manufacturer of extreme ultraviolet lithography (EUV) equipment, the tools required to manufacture chips at leading-edge nodes. ASML's order backlog functions as a leading indicator of global semiconductor supply 6 to 18 months forward.

Attention. Core computational mechanism of transformer neural networks. Each token's processing depends on learned weights applied to all other tokens in the context window, enabling the model to determine which tokens are most relevant to each other. Highly parallel and compute-intensive.

Backward compatibility. Property of CUDA and NVIDIA's software stack by which code written for older GPU generations continues to run on current architectures without modification, turning each generation of CUDA-native software into an incremental layer of switching cost.

Bandwidth. Rate of data transfer, measured in bytes per second. In AI systems, memory bandwidth (GPU-to-HBM) and interconnect bandwidth (GPU-to-GPU via NVLink, rack-to-rack via InfiniBand or Ethernet) are both performance-critical.

Blackwell. NVIDIA's 2024–2025 GPU architecture (B200, GB200), which introduced rack-scale NVL72 systems. Introduced a two-die design and native FP4 numerical precision support. Successor to Hopper.

Blackwell Ultra. Mid-cycle refresh of Blackwell (GB300, late 2025), built on the same architecture with revised silicon optimized for inference and reasoning workloads. Power consumption of ~140 kW per rack.

BlueField. NVIDIA's family of DPUs (Data Processing Units), inherited through the Mellanox acquisition, used to offload networking, storage, and security workloads from host CPUs.

Chip war. Colloquial term for the geopolitical competition over semiconductor supply chains and access to leading-edge AI compute, characterized by US export controls on advanced chips and equipment and by parallel industrial policy programs globally.

Claude. Family of frontier AI models developed by Anthropic, available via API and through consumer and enterprise products. Claude Code is an autonomous coding agent; Claude Cowork is a system-level agent for non-technical users that connects to enterprise tools and executes multi-step workflows.

Cluster. Group of interconnected server and rack systems assigned to work on a common AI training or inference task, typically spanning dozens to tens of thousands of GPUs.

Co-design (extreme co-design). NVIDIA's engineering approach in which silicon, interconnect, networking, and software are simultaneously optimized as a single integrated architecture, rather than as independent components. Organizational correlate: a flat reporting structure that concentrates full-system decision-making.

Cobalt. Microsoft's internal ARM-based server CPU, deployed in Azure data centers as part of the hyperscaler trend toward custom silicon for general-purpose compute.

ConnectX. NVIDIA's family of network interface cards (NICs), inherited from Mellanox, providing Ethernet and InfiniBand connectivity at the server level.

Context window. Maximum number of tokens an AI model can process in a single request, including both input prompt and generated output. Larger context windows enable more complex reasoning and longer agentic workflows but increase memory bandwidth requirements at inference.

CoWoS (Chip-on-Wafer-on-Substrate). TSMC's advanced packaging technology that integrates GPU silicon and HBM memory stacks onto a single silicon interposer. The binding supply bottleneck for AI accelerators, with industry demand estimated to exceed supply by ~40%–50% and lead times exceeding 40 weeks.

CPU (Central Processing Unit). General-purpose processor built around a small number of powerful cores optimized for executing complex sequential instructions at low latency. Historically the dominant chip in data centers; increasingly paired with or displaced by GPUs for AI workloads.

CSP (Cloud Service Provider). Companies offering cloud infrastructure services. In the AI context, often used interchangeably with hyperscalers.

CUDA (Compute Unified Device Architecture). NVIDIA's proprietary parallel computing platform and programming model, released in 2006. Makes GPUs programmable for general-purpose computation using familiar languages (C, C++, Python), binding AI workloads to NVIDIA hardware and anchoring the company's software moat.

CUDA-X. Collection of more than 400 domain-specific CUDA-based libraries for deep learning, data science, networking, physics simulation, and other workload classes, translating raw GPU compute into optimized performance.

Data Center (NVIDIA segment). NVIDIA's largest revenue segment (~90% of total revenue as of FY'26A). Encompasses compute (GPUs, CPUs, LPUs from 2H26 onward, and integrated systems such as DGX, HGX, and NVL72 racks) and networking (NVLink, NVSwitch, Spectrum-X, Quantum, ConnectX, BlueField).

Decode. Second phase of LLM inference, in which the model generates output tokens one at a time. Each new token depends on all previously generated tokens, and the model must read the full accumulated context (KV cache) from memory at every step — making the phase memory-bandwidth-bound rather than compute-bound.

DeepSeek. Chinese AI lab that released the R1 reasoning model in early 2025, disclosing training costs well below those commonly assumed for comparable capabilities. The episode stress-tested the AI capex cycle and, rather than triggering deceleration, was followed by upward revisions to hyperscaler capex guidance — consistent with the Jevons Paradox dynamic.

Dennard scaling. Companion principle to Moore's Law, stating that as transistors shrank, their power consumption fell proportionally, holding power density constant so each chip generation was simultaneously faster and more energy-efficient. Dennard scaling broke down around 2025 as transistor leakage current became dominant at sub-90nm nodes.

DGX. NVIDIA's own-brand pre-integrated AI server, sold directly to end customers. Captures integration margin that would otherwise flow to OEMs and targets AI-first enterprises and research labs.

Die. Single rectangular piece of silicon cut from a wafer, containing one or more chips. Modern GPU packages (Blackwell onward) contain multiple dies connected by high-speed bridges.

Digital twin. Photorealistic virtual replica of a physical asset (factory, warehouse, road network, vehicle) in which AI models can be trained, simulated, and tested before deployment in the real world. NVIDIA Omniverse is a leading platform.

DLSS (Deep Learning Super Sampling). NVIDIA's AI-based image upscaling technology for gaming, using Tensor Cores to generate high-quality frames at lower computational cost.

DPU (Data Processing Unit). Specialized processor that offloads networking, storage, and security workloads from the CPU, allowing CPUs and GPUs to focus on application logic. NVIDIA's DPU family is BlueField.

DRAM (Dynamic Random Access Memory). Volatile working memory used in computing systems to temporarily store data while a processor operates on it. Distinct from HBM, which is a specialized stacked variant used alongside GPUs. DRAM market is a consolidated triopoly (Samsung, SK Hynix, Micron).

DRIVE (NVIDIA DRIVE). NVIDIA's platform for self-driving vehicles and intelligent cockpits, spanning in-vehicle chips, driving software, simulation environments, and cloud training infrastructure.

Dynamo. NVIDIA's open-source inference serving framework, which disaggregates prefill and decode across separate GPU pools within a cluster and routes requests to the GPU already holding the relevant KV cache in memory — raising throughput and utilization in long-context and agentic workloads.

EDA (Electronic Design Automation). Software used to design, simulate, and verify semiconductors before manufacturing. Market is consolidated around Synopsys and Cadence (~70% combined share), with Siemens EDA as the smaller third competitor.

Ethernet. General-purpose networking protocol, the standard in cloud and enterprise deployments. In the AI context, NVIDIA's Spectrum-X is an Ethernet switching platform optimized for AI traffic, positioned alongside InfiniBand depending on customer preference.

EUV (Extreme Ultraviolet lithography). Advanced chip manufacturing technology required for leading-edge semiconductor nodes. ASML is the sole global supplier of EUV equipment.

Export controls. US government restrictions on the export of advanced semiconductors and equipment, principally targeting China. Have reduced NVIDIA's China AI GPU share to effectively zero (from ~20%+ of Data Center revenue) since 2025, and represent an ongoing regulatory risk for sovereign AI markets more broadly.

Fabless. Business model in which a company designs semiconductors without owning manufacturing facilities, outsourcing fabrication to third-party foundries (typically TSMC). NVIDIA is fabless, as are AMD, Apple, Broadcom, and Qualcomm.

Feynman. NVIDIA's next-generation GPU architecture expected post-Rubin, referenced in management's architecture roadmap and in the company's forward-looking per-GW economics.

Fine-tuning. Post-training step in which a pre-trained foundation model is further trained on domain-specific data to specialize it for particular applications (enterprise knowledge, coding, regulated industries).

FLOP / FLOPS (Floating-Point Operation / Operations per Second). Standard unit of chip compute throughput. TFLOPS (trillions), PFLOPS (quadrillions), and exaflops (quintillions) measure performance at chip, rack, and data center scale respectively.

Foundation model. Large AI model pre-trained on broad data, serving as the base for multiple downstream applications through fine-tuning or prompting. Examples include GPT, Claude, Gemini, and Llama.

Foundry. Semiconductor manufacturing company that fabricates chips designed by fabless customers. TSMC is the dominant leading-edge foundry; Samsung Foundry and Intel Foundry operate secondary leading-edge fabs; GlobalFoundries and UMC specialize in mature nodes.

FP4 (4-bit Floating Point). Low-precision numerical format first natively supported by Blackwell, enabling larger models to fit on fewer chips and higher inference throughput per watt at the cost of representational precision.

Frontier model. State-of-the-art AI model representing the current performance frontier, typically developed by an AI lab and requiring multi-billion-dollar training runs including experiments, infrastructure, and personnel.

Gaming (NVIDIA segment). NVIDIA's original franchise, centered on GeForce RTX GPUs for consumer desktops and laptops, plus GeForce NOW cloud gaming. The GeForce RTX 50 Series (Blackwell, launched January 2025) is the current generation.

GB200 NVL72. NVIDIA's rack-scale Blackwell system with 72 GPUs and 36 CPUs unified under NVLink, shipping since early 2025 at ~120 kW per rack.

GB300 NVL72. Blackwell Ultra rack-scale system ramping since late 2025, built on the same architecture with silicon optimized for inference and reasoning at ~140 kW per rack.

GeForce. NVIDIA's consumer GPU brand for gaming and creative workloads. The term "GPU" was coined by NVIDIA with the 1999 launch of the GeForce 256.

GPU (Graphics Processing Unit). Processor containing tens of thousands of smaller cores organized for massively parallel throughput. Originally designed for rendering 3D graphics; now the dominant accelerator for AI training and inference due to the architectural alignment between neural network mathematics and parallel linear algebra.

Grace. NVIDIA's ARM-based CPU for general-purpose data center computing, paired with GPUs in Grace Blackwell superchips.

Graviton. Amazon's internal ARM-based server CPU, deployed across AWS as part of the hyperscaler trend toward custom silicon.

Groq. Semiconductor startup that developed the LPU (Language Processing Unit), an inference-optimized architecture that stores model parameters in on-chip SRAM rather than off-chip HBM. NVIDIA licensed Groq's LPU technology in December 2025 and unveiled the Groq 3 LPX rack at GTC 2026 to pair with Vera Rubin.

GW (Gigawatt). Unit of power used to size AI data center capacity. 1 GW ≈ 1,000 MW ≈ electricity consumption of a mid-sized city (~800k residents).

HBM (High-Bandwidth Memory). Specialized DRAM that stacks multiple memory layers vertically and sits adjacent to GPU silicon via CoWoS packaging, delivering far higher data throughput at lower energy cost than standard DRAM. Now the single largest component of an AI accelerator's bill of materials. Produced by an oligopoly of SK Hynix, Samsung, and Micron.

HBM3E. Current-generation HBM, shipping with Blackwell and Blackwell Ultra.

HBM4. Next-generation HBM, shipping with Vera Rubin from 2H26 at higher capacity and bandwidth per stack.

Helios. AMD's rack-scale AI platform based on the MI450 series, developed through the Open Compute Project. Competes with NVIDIA's NVL72 architecture; Oracle has been disclosed as lead customer.

HGX. NVIDIA's 8-GPU reference baseboard, sold to OEMs (Dell, HPE, Supermicro, Lenovo) that integrate it into their own server chassis. The highest-volume Data Center channel, dominant in enterprise and mid-market deployments.

Hopper. NVIDIA's 2022 GPU architecture (H100, H200), designed around the transformer neural network. Introduced the Transformer Engine and 8-GPU server configuration (HGX H100/H200). Predecessor to Blackwell.

Huang's Law. Empirical observation that NVIDIA GPU performance has improved by ~1,000x over the past decade — far steeper than Moore's Law — by stacking multiple independent improvement curves: chip architecture, lower-precision arithmetic, software optimization, and memory bandwidth.

Huawei Ascend. Huawei's domestically developed family of AI accelerators, positioned as China's principal alternative to NVIDIA GPUs under US export controls. A potential vector for diffusion into emerging markets as an alternative to the NVIDIA/CUDA stack.

Hyperscaler. Largest cloud and internet platform operators — Google, Amazon, Microsoft, Meta, and Oracle — characterized by multi-billion-dollar annual capex and proprietary global infrastructure footprints. Account for ~60% of NVIDIA Data Center revenue.

IDM (Integrated Device Manufacturer). Business model in which a company both designs and manufactures its own chips in-house. Dominant from the 1960s through the 1990s; Intel was the archetype. Largely displaced by the fabless/foundry model outside of memory.

ImageNet. Large-scale image classification dataset and competition that catalyzed the deep learning era. AlexNet's 2012 victory using NVIDIA GPUs demonstrated that GPU-accelerated neural networks could train orders of magnitude faster than conventional processors.

Inference. Deployment phase of an AI model, in which the trained model generates outputs (tokens) in response to user queries. Scales with every user, application, and query across the economy; expected to surpass training as the largest consumer of AI compute capacity by 2027E.

Inferentia. AWS's internal inference-focused accelerator ASIC, complementing Trainium for training workloads.

InfiniBand. High-performance scale-out networking protocol originally designed for high-performance computing, acquired by NVIDIA via Mellanox in 2020 and dominant in the largest AI training clusters where latency is the binding constraint. NVIDIA Quantum is the InfiniBand switch family.

Instruction Set Architecture (ISA). Fundamental language that defines how software communicates with a processor. The two dominant ISAs in modern computing are x86 (Intel, AMD) and ARM (licensed from ARM Holdings).

Interconnect. Physical and logical fabric that moves data between processors at high speed. Within a server or rack, NVLink and NVSwitch form the scale-up interconnect; across a data center, InfiniBand or Ethernet (Spectrum-X) form the scale-out interconnect.

Interposer (silicon interposer). Thin layer of silicon that acts as a high-speed wiring board inside an advanced package, connecting GPU dies and HBM memory stacks into a unified package with short, high-bandwidth electrical pathways. CoWoS is TSMC's interposer-based packaging technology.

Jevons Paradox. Economic principle, formulated by William Stanley Jevons in 1865, stating that efficiency improvements in the use of a resource can lead to higher, not lower, aggregate consumption — as lower per-unit cost expands the set of viable applications. Applied in this report to AI compute demand.

KV cache. Memory structure that stores the Key-Value representations of processed tokens in a transformer model, allowing the model to attend to prior tokens without recomputing them at every decode step. Memory-intensive and grows linearly with context length, making it a dominant cost driver in long-context and agentic workloads.

Latency. Time delay between a request and a response. For AI inference, low latency is critical for interactive applications (chat, agentic workflows). For GPU-to-GPU communication, interconnect latency directly translates into idle time on the most expensive hardware in the system.

Leading-edge. Most advanced semiconductor manufacturing processes, measured in nanometers (nm). Smaller nodes mean more transistors per unit area, higher performance, and greater energy efficiency. NVIDIA's Blackwell is fabricated on TSMC's custom 4NP process (a custom 4nm-class node); Rubin moves to TSMC's 3nm.

Linear algebra. Branch of mathematics centered on matrix and vector operations. The computations underlying neural networks (matrix multiplications and vector transformations) map directly onto GPU architecture, which was originally designed for the linear algebra of 3D graphics rendering.

Liquid cooling. Data center thermal management method in which a liquid coolant is circulated directly over hot components. Required for high-density AI racks (NVL72 onward), which exceed the capabilities of traditional air cooling.

LLM (Large Language Model). Class of AI models (GPT, Claude, Gemini, Llama) that generate text by predicting probability-weighted tokens based on massive training datasets. Processes input data through its numerical parameters, layer by layer, across a neural network.

Logic chips. Semiconductors that process information — the processors that execute software, perform calculations, and run the computations that underpin everything from smartphones to data centers. Includes CPUs, GPUs, and custom AI accelerators. Most leading-edge-dependent segment of the semiconductor industry.

LPU (Language Processing Unit). Inference-optimized accelerator developed by Groq, storing model parameters in on-chip SRAM rather than off-chip HBM to reduce memory-access latency in the decode phase. Integrated into NVIDIA's platform via the Groq 3 LPX rack.

Maia. Microsoft's internal AI accelerator ASIC, deployed in Azure as part of the hyperscaler trend toward custom silicon for AI workloads.

Matrix multiplication. Core mathematical operation underlying neural networks, in which large grids of numbers are multiplied together. Inherently parallel — which is why GPUs, designed for the matrix math of 3D graphics, are architecturally suited for AI.

Mellanox. Networking company acquired by NVIDIA in 2020 for ~USD 7bn, bringing InfiniBand, Ethernet switching, ConnectX NICs, and BlueField DPUs into NVIDIA's portfolio. Foundational to NVIDIA's ownership of the full scale-up and scale-out networking stack.

Memory bandwidth. Rate at which data moves between memory and processing units, measured in GB/s or TB/s. A primary bottleneck when handling the massive parameter sets of frontier AI models, particularly in the decode phase of inference.

METR (Model Evaluation & Threat Research). Independent AI evaluation organization whose time-horizon benchmarks measure task difficulty in human-expert-time units (how long a skilled human would take to complete a task) and report the threshold at which an AI succeeds on 50% of tasks. Referenced as a proxy for agentic AI capability progression.

MI300X / MI350 / MI450. AMD's Instinct series of data center GPUs. MI300X is the prior-generation platform; MI350 is the current platform; MI450 is the architecture underlying the Helios rack-scale system expected in 2H26.

MIG (Multi-Instance GPU). Feature introduced in Ampere that allows a single physical GPU to be partitioned into up to seven smaller virtual GPUs, each running independent workloads. Supports multi-tenancy in cloud deployments.

Moore's Law. Empirical observation, formalized by Gordon Moore in 1965, that transistor density doubles roughly every two years — enabling predictable performance gains that shaped five decades of semiconductor industry investment. Density scaling has slowed materially since the mid-2010s as node approach physical limits.

MTIA. Meta's internal AI training and inference accelerator ASIC.

NAND flash. Non-volatile memory technology used as the storage medium in solid-state drives, smartphones, and data center storage arrays. Follows a cycle driven primarily by consumer electronics and enterprise storage demand, largely independent of AI capex.

Nanometer (nm). Unit in which semiconductor process nodes are measured. Smaller numbers correspond to greater transistor density, higher performance, and better energy efficiency. Current leading-edge nodes are 3nm; TSMC's 2nm is in ramp for 2026.

NCP (NVIDIA Cloud Partner). GPU-focused cloud providers ("neoclouds") that purchase NVIDIA hardware at scale and resell compute capacity to AI labs and enterprises. Examples include CoreWeave, Nebius, and Nscale.

NemoClaw. NVIDIA's enterprise-ready version of OpenClaw, adding sandboxing, privacy, and network security layers to serve regulated environments.

Nemotron. NVIDIA's family of open-weight AI models designed to run on NVIDIA hardware, seeding the ecosystem with NVIDIA-optimized workloads without requiring dependence on a single proprietary model provider.

Neocloud. Informal term for NVIDIA Cloud Partners (see NCP).

Neural network. Layered mathematical structure by which AI models learn patterns from data. Within each layer, operations are independent and identical in structure — matrix multiplications that can be distributed across GPU cores simultaneously.

NIC (Network Interface Card). Hardware component installed on each server that provides network connectivity. NVIDIA's ConnectX family supports Ethernet and InfiniBand.

NIM (NVIDIA Inference Microservice). Pre-packaged, optimized AI inference endpoints running on NVIDIA hardware, distributed as part of NVIDIA AI Enterprise.

Node (process node). Generation of chip manufacturing process, typically denominated in nanometers (3nm, 5nm, 7nm). Smaller nodes offer higher transistor density and better performance per watt but require more advanced equipment (EUV) and carry higher capital intensity.

NVL72. Rack architecture with 72 GPUs unified under a single NVLink domain, introduced with GB200 and extended through GB300 and Vera Rubin.

NVL576 Kyber. Next-generation NVLink domain architecture, extending the unified GPU pool to 576 GPUs. First deployed with Rubin Ultra, expected in 2H27.

NVLink. NVIDIA's proprietary high-speed interconnect protocol for GPU-to-GPU communication, operating as scale-up networking within a rack. Bandwidth per GPU has scaled from 900 GB/s (Hopper) to 1,800 GB/s (Blackwell) to 3,600 GB/s (Rubin).

NVSwitch. The physical switch chip that implements NVLink routing, enabling all-to-all GPU communication within a rack so that all connected GPUs pool memory and exchange data simultaneously at full bandwidth.

OEM (Original Equipment Manufacturer). Server manufacturers (Dell, HPE, Supermicro, Lenovo, Foxconn) that integrate NVIDIA HGX reference designs into finished servers sold to end customers.

Omniverse. NVIDIA's simulation and digital twin platform, enabling photorealistic virtual replicas of physical assets for running AI models before physical deployment. Used in manufacturing, logistics, and autonomous vehicle development.

Open-weight model. AI model whose parameters (weights) are publicly released, enabling users to run, fine-tune, and modify the model on their own infrastructure. Contrast with closed models accessed only via API.

OpenClaw. Open-source agentic framework, first published in late 2025, that enables agents to connect to operating systems, messaging platforms, and applications to execute real-world tasks autonomously. Surpassed 300,000 GitHub stars shortly after release.

OpenRouter. API aggregation platform routing developer traffic across 400+ LLMs from 60+ providers through a single interface, publishing weekly token throughput in near real time. Used in this report as a model-agnostic proxy for AI token consumption.

Packaging (semiconductor packaging). Process of enclosing one or more silicon dies into a finished chip package that can be mounted on a board. Advanced packaging (CoWoS) integrates multiple dies and memory stacks in a single package via a silicon interposer.

Parallelism. Property of a workload that can be distributed across many processors executing simultaneously, rather than a single processor executing sequentially. Neural network computations are massively parallel by nature, which is why GPUs dominate AI.

Parameters (weights). Learnable numerical values inside a neural network that are adjusted during training. Model size is typically measured in billions of parameters; frontier models exceed one trillion.

Post-training. Training steps applied after pre-training, including fine-tuning, alignment, and reasoning-oriented reinforcement learning (RLVR). A growing source of capability gains in recent model generations and a driver of GPU utilization between pre-training runs.

Power Usage Effectiveness (PUE). Ratio of total power entering a data center facility to the power reaching IT equipment. Lower is better; ~1.15x is typical for modern liquid-cooled AI facilities, meaning ~87% of incoming power reaches IT equipment.

Prefill. First phase of LLM inference, in which the model reads the entire input prompt at once, converts each token into a numerical vector, and runs all operations in parallel across the GPU's cores. Compute-bound; efficiently handled by GPUs.

Pre-training. Initial training phase in which a foundation model learns general patterns from massive datasets (often trillions of tokens). Compute-intensive and typically performed on dedicated GPU clusters over months of continuous computation.

Prisoner's Dilemma. Game-theoretic framework in which defection (unilateral withdrawal) carries asymmetrically worse downside than continued cooperation, making continued spending the dominant strategy. Applied in this report to the hyperscaler AI capex cycle, where underinvestment risks irrecoverable capability gaps.

Professional Visualization (NVIDIA segment). NVIDIA's enterprise graphics segment, serving designers, engineers, architects, and content creators via RTX PRO GPUs (formerly Quadro) and the Omniverse simulation platform.

Quantum. NVIDIA's InfiniBand switch family, part of the scale-out networking stack inherited from the Mellanox acquisition.

Rack. Standardized physical cabinet that stacks dozens of servers with associated power and cooling infrastructure. In rack-scale architectures, the entire rack operates as a single coherent computing system.

Rack-scale. Architecture in which multiple GPUs, CPUs, and networking components across a full rack operate as a single logical system, rather than as independent servers. Introduced at scale with GB200 NVL72.

Ramp AI Index. Index published by Ramp tracking the share of US businesses with paid subscriptions to AI models, platforms, and tools. Used as a proxy for enterprise adoption of AI services.

Ray tracing. Rendering technique that simulates the physical behavior of light to produce photorealistic graphics. Hardware-accelerated by dedicated RT cores in NVIDIA RTX GPUs.

Reasoning model. AI model trained to perform multi-step reasoning through chains of intermediate "thinking tokens" before producing a final answer, typically consuming thousands to tens of thousands of tokens per query versus a few hundred for conventional chat models. Commercially introduced with OpenAI's o1 in September 2024.

Reinforcement Learning (RL). Training methodology in which a model learns through a reward signal rather than through supervised examples. Applied across model training (RLHF, RLVR) and agentic workloads.

RLVR (Reinforcement Learning with Verifiable Rewards). Post-training technique in which a model is trained on tasks with objectively measurable outcomes (math problems with correct answers, code that either compiles or does not) and iteratively rewarded for correct outputs. A key driver of recent capability gains in reasoning and agentic workflows.

Robotaxi. Taxi service operated with vehicles that drive autonomously without a human driver. One of the first large-scale commercial deployments of physical AI; NVIDIA supplies the full stack from training infrastructure to in-vehicle compute.

ROCm (Radeon Open Compute). AMD's open-source software stack for GPU computing, positioned as an alternative to NVIDIA's CUDA. Currently less mature in AI workloads and without comparable breadth of library support or installed base of trained engineers.

RTX / RTX PRO. NVIDIA's current GeForce branding, introduced with real-time ray tracing support, and the professional workstation GPU line (formerly Quadro) serving designers, engineers, and content creators.

Rubin. NVIDIA's next-generation GPU architecture, launching in the Vera Rubin platform (VR200 NVL72) in 2H26. Built on TSMC's 3nm node with HBM4 memory.

Rubin Ultra. Mid-cycle refresh of Rubin, expected in 2H27, deployed in the NVL576 Kyber architecture.

Scale-out networking. Networking that connects multiple racks into larger AI clusters, typically spanning hundreds to tens of thousands of GPUs. NVIDIA's options are InfiniBand (Quantum switches) and Ethernet (Spectrum-X switches).

Scale-up networking. High-bandwidth interconnect that binds GPUs within a single rack into a coherent computing domain, allowing them to function as a unified system that pools memory and exchanges data at full bandwidth. NVIDIA uses NVLink and NVSwitch for scale-up.

Scaling laws. Empirical principles describing how AI model capability improves predictably when given more compute. As the compute budget used to train a model increases, its error rate decreases in a log-linear pattern. Multiple scaling law vectors now coexist: pre-training, post-training (RLVR), test-time compute, and agentic compute.

SemiAnalysis. Independent semiconductor research firm whose InferenceMAX benchmarks and supply-chain analyses are referenced throughout the report as third-party sources for hardware performance comparisons and AI infrastructure economics.

Semicaps. Semiconductor capital equipment companies that manufacture the tools used to fabricate chips. Dominated by ASML (lithography), Applied Materials (thin-film deposition), Lam Research (etching), KLA (inspection), Tokyo Electron, and ASM International. Sit at one of the most concentrated chokepoints in the supply chain.

Server. Single machine containing GPUs, CPUs, memory, and networking, mounted vertically inside a rack. Typically the smallest discrete building block of AI infrastructure, though increasingly subsumed into rack-scale architectures.

Sovereign AI. National-level AI infrastructure programs, purchased directly by governments and state-linked entities rather than through hyperscaler channels. Anchored around strategic competitiveness rather than ROIC thresholds, and therefore more price-inelastic than enterprise or hyperscaler capex.

Spectrum-X. NVIDIA's AI-optimized Ethernet switching platform for scale-out networking, positioned alongside InfiniBand depending on customer preference.

SRAM (Static Random Access Memory). On-chip memory used for caches and register files. Much faster than DRAM/HBM but smaller and more expensive per bit; the basis of Groq's LPU architecture, which stores model parameters in on-chip SRAM rather than off-chip HBM.

Substrate. Physical layer on which semiconductor dies and interposers are mounted within an advanced package, providing electrical connections to the circuit board.

Superchip. NVIDIA-specific term for an integrated module pairing GPUs with a CPU. The Grace Blackwell Superchip pairs 2 Blackwell GPUs with 1 Grace CPU; an NVL72 rack contains 36 superchips.

TCO (Total Cost of Ownership). Sum of all direct and indirect costs of operating a system over its useful life, including hardware (compute, networking), physical infrastructure (land, shell, power, cooling), and ongoing operating expense.

Tensor Cores. Specialized processing units within NVIDIA GPUs designed to accelerate the matrix operations that underpin neural network computation. Embedded in every RTX chip, enabling generative AI applications that run locally on consumer PCs.

Test-time compute (inference-time scaling). Compute allocated at inference time to improve model response quality, typically through extended reasoning, multiple sampling, or tree search. A distinct scaling law vector from pre-training; became commercially visible with OpenAI's o1 in September 2024.

Thinking tokens. Intermediate reasoning tokens generated by a reasoning model between the input prompt and the final answer, representing the model's chain of thought. A reasoning model can generate thousands to tens of thousands of thinking tokens for a single query.

Thread. Single sequence of instructions executed by a processor. CPUs optimize for low latency on each individual thread; GPUs maximize throughput across massively parallel threads.

Throughput. Rate at which a system processes work, typically measured in tokens per second for AI inference or FLOPs per second for training.

Token. Sub-word unit of text, typically three to four characters, that AI models use as their atomic unit of input and output. Models are priced per token, operators measure cost per token, and hardware generations are benchmarked on tokens per watt.

Tokens per watt. Binding metric for AI factory economics, defined as the number of tokens a system produces per second for each watt of power consumed. As it rises, cost per token falls — which is why each new NVIDIA architecture generation that improves tokens per watt directly improves operator ROIC.

TPU (Tensor Processing Unit). Google's family of internal AI accelerator ASICs, originally designed for TensorFlow workloads and now deployed across Google Cloud.

Training. Process of developing an AI model by exposing it to massive datasets and iteratively adjusting model parameters. Encompasses both pre-training and post-training. Contrast with inference.

Trainium. AWS's internal training accelerator ASIC, complementing Inferentia for inference workloads.

Transformer. Neural network architecture introduced in 2017 that uses attention mechanisms to process sequences. Foundation of essentially all modern LLMs.

Transformer Engine. Hardware feature introduced in Hopper that automatically adjusts numerical precision during computation to maximize speed without sacrificing accuracy, delivering up to 9x faster training and 30x faster inference versus the A100.

TSMC. Taiwan Semiconductor Manufacturing Company, the monopoly foundry of all leading-edge AI accelerators. Commands ~70%+ of global foundry revenue and ~90%+ share of leading-edge production, with 3nm allocation reportedly committed through 2027.

Vector. Ordered list of numbers representing data (such as a token's meaning) in a format the model can process. The basic data structure of neural network computation, alongside matrices.

Vera. NVIDIA's next-generation CPU, designed for agentic AI workloads and paired with Rubin GPUs in the Vera Rubin platform.

Vera Rubin (VR200 NVL72). NVIDIA's 2H26 rack-scale platform combining Vera CPU and Rubin GPU, projected at ~200 kW per rack with HBM4 memory.

Vera Rubin POD. Heterogeneous AI supercomputer combining Vera Rubin NVL72 (GPUs) and Groq 3 LPX (LPUs), with Dynamo orchestrating the prefill/decode split across the two architectures. Delivers up to ~35x more throughput per megawatt than Blackwell NVL72 per SemiAnalysis benchmarks.

Wafer. Thin disc of silicon (typically 300mm diameter) on which semiconductor chips are fabricated. Each wafer yields multiple dies depending on chip size; leading-edge wafer capacity at TSMC is the industry's most constrained manufacturing input.

Watt. Unit of power. AI systems are rated in kilowatts (kW, per rack), megawatts (MW, per cluster), and gigawatts (GW, per data center facility). Power has replaced floor space as the primary unit in which AI data center capacity is sized.

x86. CPU instruction set architecture developed by Intel in 1978 and extended by AMD, historically dominant in data center computing. The incumbent workload base that NVIDIA's Grace and Vera CPUs target for migration to accelerated computing.

Yield. Proportion of functional chips produced per wafer in semiconductor manufacturing. Higher yield means lower effective cost per chip for the customer — a core driver of TSMC's share gains at each new process node.

Analyst disclosures

1. The analyst responsible for preparing this document, highlighted in bold, hereby certifies that all opinions expressed in this report accurately, solely and exclusively reflect his/her personal views and opinions regarding all of the issuers and securities analyzed herein and were provided in this document independently and autonomously. Whereas the personal opinions of investment analysts may diverge, Safra Corretora and/or Banco Safra and/or any of their affiliated companies may have published or may publish other reports that are inconsistent with and/or reach different conclusions than those provided herein.

2. The analyst responsible for preparing this report is not registered and/or not qualified as a research analyst at the NYSE or FINRA and such analyst is not in any way associated with Safra Securities LLC ("SSL") and is, therefore, not subject to the provisions of Rule 2242 on communications with researched companies, public appearances and transactions involving securities held in a research analyst account.

3. An analyst's compensation is based upon the total revenues of Safra Corretora, a portion of which is generated through investment banking activities. Like all employees of Safra Corretora, its subsidiaries and affiliates, analysts receive compensation that is impacted by their overall profitability. For this reason, analysts' compensation can be considered to be indirectly related to this report. However, the analyst responsible for the content of this report hereby represents that no part of his or her compensation was, is, or will be directly or indirectly related to any specific recommendation or views contained herein or linked to the pricing of any of the securities discussed herein. The analyst declares that (s)he does not maintain any relationship with any individual affiliated with the companies or governments mentioned herein and does not receive any compensation for services rendered to or have any commercial relationship with the company or any individual or entity representing the interests of the company. Neither the analyst(s) nor any member of their household holds, directly or indirectly, more than 5% of their personal net worth in any securities issued by the companies or governments analyzed in this report in his/her personal investment portfolio, nor is (s)he personally involved in the acquisition, sale or trading of such securities in the market. Neither the analyst(s) nor any member of their household serves as an officer, director or member of the advisory board of the companies analyzed in this report.

Disclosure items

Analysts	1	2	3	4

1. The securities analyst(s) involved in preparing this report are associated with individuals who work for the issuers addressed herein.
2. The securities analyst's(s') spouse(s) or partner(s) hold, either directly or indirectly, on their own behalf or on behalf of third parties, the stocks and/or other securities discussed in this report.
3. The securities analyst(s) or their spouse(s) or partner(s) are directly or indirectly involved in the purchase, sale or intermediation of the securities discussed in this report.
4. The securities analyst(s), their respective spouse(s) or partner(s) hold, either directly or indirectly, any financial interest in the issuers of the securities analyzed in this report.

IMPORTANT INFORMATION ABOUT SAFRA

Safra Corretora and/or its affiliates declare that they (i) have significant financial and commercial relationships with and/or (ii) receive compensation for services rendered to the following company(ies) and investment fund(s):

Agropecuária Maggi Ltda., Alfa Holdings S.A., Alianza Trust Renda Imobiliária FII — 7ª Emissão, Amaggi Exportação e Importação Ltda., Ambiental Metrosul Concessionária de Saneamento SPE S.A., Antônio Venâncio da Silva Empreendimentos Imobiliários Ltda., Armarinhos Fernando Ltda., Ártemis FII — 2ª Emissão, ARX Dover Recebíveis FII — 3ª Emissão, Atacadão S.A., AZ Quest Panorama Log FII — 2ª Emissão, B3 S.A. Brasil, Bolsa, Balcão, Banco Alfa de Investimento S.A., Banco CNH Industrial Capital S.A., Banco GM S.A., Bocaina Infra FIC FI Infra RF CP — 5ª Emissão, BPG Av Mofarrej Empreendimentos e Participações S.A., BRF S.A., BRZ Infra FIC FI — 1ª Emissão, Banco BTG Pactual, Caixa Seguridade Participações S.A., Cantu Store S.A., Carrefour Comércio e Indústria Ltda. , CashMe S.A., CCR AutoBan, Cemig Distribuição S.A., Centrais Elétricas Brasileiras S.A. — Eletrobras, Centrais Elétricas do Norte do Brasil S.A., Cereal Comércio Exportação e Representação Agropecuária S.A., Cerradinho Bioenergia S.A., Cimed & CO. S.A.,

Cloudwalk Instituição de Pagamento e Serviços Ltda., Companhia Catarinense de Águas e Saneamento – CASAN, Companhia de Saneamento Básico de Sao Paulo — SABESP, Companhia de Saneamento de Minas Gerais, Companhia de Transmissão de Energia Elétrica Paulista — CTEEP, Companhia Hidro Elétrica São Francisco, Companhia Pernambucana de Gás — COPERGÁS, Companhia Riograndense de Saneamento, Concessionária do VLT Carioca S.A., Consórcio Alfa de Administração S.A., Construtora Baggio Ltda., Cooperativa Regional de Cafeicultores em Guaxupé Ltda. Cooxupé, Copel Distribuição S.A., Copel Geração e Transmissão S.A., Copérnico Comercializadora de Energia Ltda., Copérnico Energias Renováveis S.A., Cordeiro Fios e Cabos Elétricos Ltda., Companhia Piratininga de Força e Luz S.A., CPV Energia FII Responsabilidade Limitada — 2ª Emissão, Cruzeiro do Sul Educacional S.A., Cury Construtora e Incorporadora S.A., Cyrela Brazil Realty S.A. , Cyrela Crédito Fundo de Investimento Imobiliário, Diagnósticos da América S.A., Direcional Engenharia S.A., EDP São Paulo Distribuição de Energia S.A., Eletronorte, Eletrozema S.A., Empresa Brasileira de Loteamentos Ltda. — EMBRALOT, Energisa S.A., Engie Brasil Energia S.A., Equatorial Pará Distribuidora de Energia S.A., Equatorial Participações e Investimentos IV S.A., ETF Buena Vista Neos Bitcoin High Income — 1ª Emissão, ETF II Buena Vista, ETF Investo Bitcoin, ETF Investo Renda Fixa, ETF QR CME CF Solana Dollar Reference Rate, Eucatex Indústria e Comércio Ltda., Exes FII — 4ª Emissão, Fazenda Pioneira Empreendimentos Agrícolas S.A., FII Capitânia Shoppings — 4ª Emissão, FII Invista Brazilian Business Park — 2ª Emissão, FII REC Fundo de CRI Cotas Amortizáveis — 1ª Emissão, Financeira Alfa S.A., Forma Certa Soluções Gráficas Ltda., Frigol S.A., GDM Genética S.A., Gestora de Inteligência de Crédito S.A., GLP Capital Partners Gestão de Recursos e Administração Imobiliária Ltda., Grupo Cereal S.A., Guardian Real Estate FII — 6ª Emissão, Hashdex Momentum ETF, Hedge Brasil Logístico Industrial — 5ª Emissão, Hedge Recebiveis — 6ª Emissão, HSI Malls FII — 4ª Emissão, Huma Capital Ltda., Icatu Vanguarda GRU Logístico FII — 1ª emissão, Iguatemi Empresa de Shopping Centers S.A., In—Haus Servicos Industriais e Logística S.A., Inter Amerra Fiagro Imobiliário — 2ª Emissão, Inter Infra FIC Renda Fixa — 3ª Emissão, Inter Oportunidade Imobiliária FII — 1ª Emissão, JBS S.A., JHSF Participações S.A., JiveMauá Bossanova FIC FI—Infra — 1ª Emissão, JS Crédito Estruturado, Kinea Agro Income USD FIAGRO — 1ª Emissão, Kinea Rendimentos Imobiliários — 11ª Emissão, Kinea Securities FII — 5ª Emissão, Lar Cooperativa Agroindustrial, Lavvi Empreendimentos Imobiliarios S.A., Life Capital Partners FII — 6ª Emissão, Localiza Rent a Car S.A., Log Commercial Properties e Participações S.A., LOGCP Inter FII — 4ª Emissão, Lojas Belian Moda Ltda., Lucca Incorporações e Participações S.A., Maha Energy Finance SARL, Manati Capital Hedge Fund FII — 5ª Emissão, Marfrig Global Foods S.A., Maxi Renda FII — 10° Emissão, Melnick Desenvolvimento Imobiliario S.A., Minas Mineração Ltda. , Moura Dubeux Engenharia e Empreendimentos Ltda., MRS Logística S.A., MRV Engenharia e Participações S.A., Multiplan Empreendimentos Imobiliários S.A., Navi Infra FIP IE — 1ª Emissão, Nex Crédito Fiagro Imobiliário — 2ª Emissão, Nortis Incorporadora e Construtora S.A., Open K Ativos e Recebíveis Imobiliários FII — 7ª Emissão, Oryx Bonds Conversíveis EUA ETF, Parsan S.A., Parshop Participações Ltda., Patria Crédito Infra Renda FIC FI— Infra — 1ª Emissão, Patria Recebíveis Imobiliários FII — 10ª Emissão, Patria Renda Urbana FII RL Unica — 5ª Emissão, Paulista Praia Hotel S.A., Petróleo Brasileiro S.A., Plano & Plano Desenvolvimento Imbiliário S.A., Prati, Donaduzzi & Cia Ltda., Quartzo Real Estate Development Mult FII — 1ª Emissão, RB Capital Infraestrutura FIC FI—Infra — 1ª Emissão, RB Investimentos Multiestratégia — 1ª Emissão, Rec Master CRI FII — 1ª Emissão, REC Multiestratégia Fundo de Investimento Imobiliário — 1ª Emissão, RIFF FIC FI — Infra — 1ª emissão, Rio Bravo ESG IS FIC FI Infra RF CP — 3ª Emissão, Rojemac Importação e Exportação Ltda., Santos Brasil Participações S.A., São Martinho S.A., Seara Alimentos LTDA, Sendas Distribuidora S.A., SLC Agrícola S.A., SLC Máquinas Ltda., Smart Real Estate FII — 2ª Emissão, Sociedade Beneficiente Israelita Brasileira Hospital Albert Einstein, Sparta FIAGRO — 3ª Emissão, Sparta Fiagro — 4ª Emissão, Sparta Infra CDI FI FIC Infra — 5ª Emissão, Suno Energias Limpas — 3ª Emissão, Transmissora Aliança de Energia Elétrica S.A. — TAESA, Tanac S.A., Tenax RFA Incentivado — 1ª Emissão, TG Renda Urbana FII — 1ª Emissão, Tigre S.A. Participações, TJK Renda Imobiliária FII — 2ª Emissão, TRX Hedge Fund FII — 1ª Emissão, TRX Real Estate FII — 11ª Emissão, Union Agro S.A., Usina Vale do Tijuco, V.Tal — Rede Neutra de Telecomunicacoes S.A., Vale S.A., Valora CRI CDI FII — 8ª Emissão, Valora CRI CDI FII — 9ª Emissão, Valora Debêntures INC FIC FI—Infra — 1ª Emissão, Vectis Gestão de Recursos Ltda., Vectis Securities FII — 1ª Emissão, Vera Cruz Agropecuária Ltda., Vera Cruz CRI Residencial High Grade — 1ª Emissão, Viação Piracicabana S.A., WHG Real Estate — 3ª Emissão, YVY Capital FIC Infra RF — 1ª Emissão, Zagros Multiestratégia FII — 2ª Emissão.

Important global disclosures

1. This report was prepared by Safra Distribuidora de Títulos e Valores Mobiliários Ltda. (“Safra Corretora”), a subsidiary of Banco Safra S.A., a company regulated by the Brazilian Securities and Exchange Commission (CVM). Safra Corretora is responsible for the distribution of this report in Brazil.

2. This report is for distribution only under such circumstances, as may be permitted by applicable law. This report is not directed at you if Safra Corretora and/or Banco Safra is prohibited or restricted by any legislation or regulation in any jurisdiction from making it available to you. You should satisfy yourself before reading it that Safra Corretora and/or Banco Safra are permitted to provide research material concerning investments to you under relevant legislation and regulations.

3. This report is provided for informational purposes only and does not constitute or should not be construed as an invitation, solicitation, offer or inducement to buy or sell any financial instrument or to participate in any particular trading strategy in any jurisdiction.

4. The information herein was deemed reasonable on the date of its publication and was obtained from reliable public sources. Safra Corretora does not ensure or guarantee, either expressly or implicitly, that the information contained herein is accurate or complete. Safra Corretora has no obligation to update, modify or amend this report and informs the reader accordingly, except when terminating coverage of the companies discussed in the report. The opinions, estimates, information and/or forecasts expressed in this report constitute the current judgment of the analyst responsible for the content of this report as of the date in which it was issued and are therefore subject to change without notice. The prices and availability of the financial instruments are merely indicative and subject to change beyond the control of Safra Corretora. Safra Corretora is not obligated to update, amend or otherwise alter this report, or to inform readers of any changes in its content, except upon termination of coverage of the issuers of the securities discussed herein.

5. Due to international regulations, not all financial services/instruments may be available to all clients. You should be aware of and observe any such restrictions when considering a potential investment decision.

6. The financial instruments discussed in this document may not be available to or suitable for all investors. This report does not take under consideration the specific investment objectives, financial situation or needs of any particular investor. Investors who intend to purchase or trade the securities covered in this report must seek out the applicable information and documents in order to decide whether to invest or not in such securities. Investors must independently seek out financial, accounting, legal, economic and market guidance, based on their personal profile, before making any investment decision regarding the securities of the issuers analyzed in this report. Each investor must make the final investment decision based on a range of related risks, fees and commissions. In all cases, investors should conduct their own investigation and analysis of such information before taking or omitting to take any action in relation to the securities or markets that are analyzed in this report.

7. The report should not be regarded by recipients as a substitute for the exercise of their own judgment. The opinions, estimates and projections expressed herein constitute the current judgment of the analyst responsible for the substance of this report as of the date on which the report was issued and are therefore subject to change without notice and may differ or be contrary to the opinions expressed by other business areas of Banco Safra as a result of their use of different assumptions and criteria.

8. If a financial instrument is expressed in currencies other than the one used by the investor, exchange rate fluctuations may adversely affect the price, value or profitability. Yields of financial instruments may vary, ultimately increasing or decreasing the price or value of financial instruments, either directly or indirectly. Past performance is not necessarily indicative of future results, and this report does not ensure or guarantee, either expressly or implicitly, any possible future performance or any other aspect thereof. Safra Corretora and its affiliated companies may not be held liable for any losses, either direct or indirect, arising from the use of this report or its content. Upon using the content herein, investors undertake to irrevocably and irreparably hold Safra Corretora and/or any of its affiliated companies harmless from and against any claims, complaints and/or losses.

9. Any opinions, estimates and projections must not be construed as a representation that the matters referred to therein will occur. The prices and availability of financial instruments are indicative only and subject to change without notice. The Research department will initiate, update and cease coverage solely at the discretion of Banco Safra.

10. This report may not be reproduced or redistributed to any other person, wholly or in part, for any purpose, without the prior written consent of Safra Corretora. Additional information relative to the financial instruments discussed in this report is available upon request.

Additional disclaimer for reports distributed to:

USA:

Safra Securities LLC (“SSL”), a FINRA/SIPC member firm, is distributing this report in the United States and accepts responsibility for the content of this report. SSL assumes responsibility for this research for purposes of U.S. law. Any U.S. investor who receives this report and intends to trade any of the securities addressed herein must do so through Safra Securities LLC at 546 5th Ave, 2nd Floor, New York, NY.

UK AND EUROPEAN ECONOMIC AREA (EEA):

The communication of this report is not being made and has not been approved by an authorized person for the purposes of Section 21 of the United Kingdom Financial Services and Markets Act 2000 ("FSMA 2000"). Accordingly, the report is not being distributed to, and must not be passed on to, the general public in the United Kingdom and is to be circulated only to persons outside the United Kingdom or to persons within the United Kingdom falling within the definition of investment professionals (as defined in Article 19(5) of the FSMA 2000 (Financial Promotion) Order 2005 ("Order")) or to other persons to whom it may be lawfully communicated in accordance with the Order.

While all reasonable efforts have been made to ensure that the information contained herein is not untrue or misleading at the time of its publication, no representation is made as to its accuracy or completeness, and it should not be relied upon as such. Past performances are not a guarantee of future performances. All opinions expressed in the present document reflect the current context and are subject to change without notice.

OTHER COUNTRIES:

This report and the securities discussed herein may not be eligible for distribution or sale in all countries or to certain categories of investors. In general, this report may be distributed only to professional and institutional investors. By accessing this report, you confirm that you are aware of the laws in your jurisdiction relating to the provision and sale of financial service products and acknowledge that this material contains proprietary information and that you are to keep this information confidential. Additionally, you confirm that you understand the risks related to the financial instruments discussed herein. Due to international regulations, not all financial services/instruments may be available to all clients.