Company Overview

May 2025

Certain matters in this presentation including, but not limited to, statements as to: expectations with respect to growth, performance and benefits of NVIDIA's products, services, and technologies, including Blackwell, and related trends and drivers; expectations with respect to supply and demand for NVIDIA's products, services, and technologies, including Blackwell, and related matters including inventory, production and distribution; our financial position; projected market growth and trends, market opportunity, demand, and growth drivers; our financial and business outlook; our dividend program; expectations with respect to NVIDIA's third party arrangements, including with its collaborators and partners; third parties adopting our products and technologies; expectations with respect to technology developments and related trends and drivers; expectations with respect to AI and related industries; our sustainability goals; and other statements that are not historical facts are forward-looking statements within the meaning of Section 27A of the Securities Act of 1933, as amended, and Section 21E of the Securities Exchange Act of 1934, as amended, which are subject to the "safe harbor" created by those sections based on management's beliefs and assumptions and on information currently available to management and are subject to risks and uncertainties that could cause results to be materially different than expectations.

Important factors that could cause actual results to differ materially include: global economic and political conditions; NVIDIA's reliance on third parties to manufacture, assemble, package and test NVIDIA's products; the impact of technological development and competition; development of new products and technologies or enhancements to NVIDIA's existing product and technologies; market acceptance of NVIDIA's products or NVIDIA's partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of NVIDIA's products or technologies when integrated into systems; and changes in applicable laws and regulations, as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

Many of the products and features described herein remain in various stages and will be offered on a when-and-if-available basis. The statements within are not intended to be, and should not be interpreted as a commitment, promise, or legal obligation, and the development, release, and timing of any features or functionalities described for our products is subject to change and remains at the sole discretion of NVIDIA. NVIDIA will have no liability for failure to deliver or delay in the delivery of any of the products, features or functions set forth herein.

NVIDIA uses certain non-GAAP measures in this presentation including non-GAAP operating income, non-GAAP operating margin, and free cash flow. NVIDIA believes the presentation of its non-GAAP financial measures enhances investors' overall understanding of the company's historical financial performance. The presentation of the company's non-GAAP financial measures is not meant to be considered in isolation or as a substitute for the company's financial results prepared in accordance with GAAP, and the company's non-GAAP measures may be different from non-GAAP measures used by other companies. Further information relevant to the interpretation of non-GAAP financial measures, and reconciliations of these non-GAAP financial measures to the most comparable GAAP measures, may be found in the slide titled "Reconciliation of Non-GAAP to GAAP Financial Measures."

NVIDIA

NVIDIA's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, revolutionized accelerated computing, ignited the era of modern AI, and is fueling industrial digitalization across markets.

Today, two transitions are occurring simultaneously—accelerated computing and generative AI—transforming the computer industry and every other industry worldwide, and NVIDIA is enabling these transitions with our full-stack computing platform and data-center-scale offerings.

NVIDIA's platform is installed in several hundred million computers, is available in every cloud and from every server maker, powers over 75% of the TOP500 supercomputers, and has ~5.9 million developers.

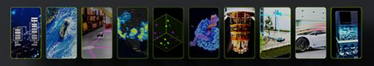
Headquarters: Santa Clara, CA | Headcount: ~36,000

NVIDIA’s Accelerated Computing Platform

Data center scale innovation across chips, networking, systems, software, and algorithms

NIM
CUDA-Accelerated
Agentic AI Libraries

Omniverse CUDA-Accelerated
Physical AI Libraries



CUDA-X Libraries

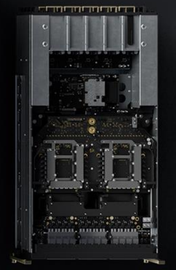
CUDA • DOCA • NCCL
Cluster-Scale Software
System Software
Chip Software

Accelerated
Software Stack

GB200 NVL72 SuperPOD

Grace Blackwell
MGX Node

NVLink Switch



Quantum Switch

Spectrum-X
Switch

Chips Purpose-Built for AI Supercomputing
GPU | CPU | DPU | NIC | NVLink Switch | IB Switch | ENET Switch

NVIDIA has accelerated software and compute by a 1,000,000X in the last decade, far surpassing Moore’s law.

Accelerated computing requires full-stack innovation—optimizing across every layer of computing—from chips and systems to software and algorithms, demanding deep understanding of the problem domain.

Our platform extends from the cloud and enterprise data centers to supercomputing, edge computing, PCs, and robotics.

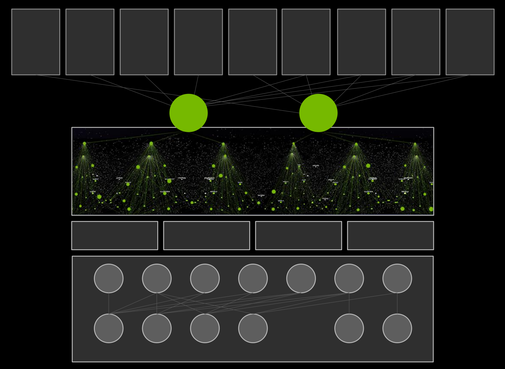
What Is Accelerated Computing?

A full-stack approach: silicon, systems, software

Not just a superfast chip—accelerated computing is a full-stack combination of:

- Chip(s) with specialized processors
- Algorithms in acceleration libraries
- Domain experts to refactor applications

To speed up compute-intensive parts of an application



Amdahl's law:

The overall system speed-up (S) gained by optimizing a single part of a system by a factor (s) is limited by the proportion of execution time of that part (p).

$$S = \frac{1}{(1-p) + \frac{p}{s}}$$

For example:

- If 90% of the runtime can be accelerated by 100X, the application is sped up 9X
- If 99% of the runtime can be accelerated by 100X, the application is sped up 50X
- If 80% of the runtime can be accelerated by 500X, or even 1,000X, the application is sped up 5X

Why Accelerated Computing?

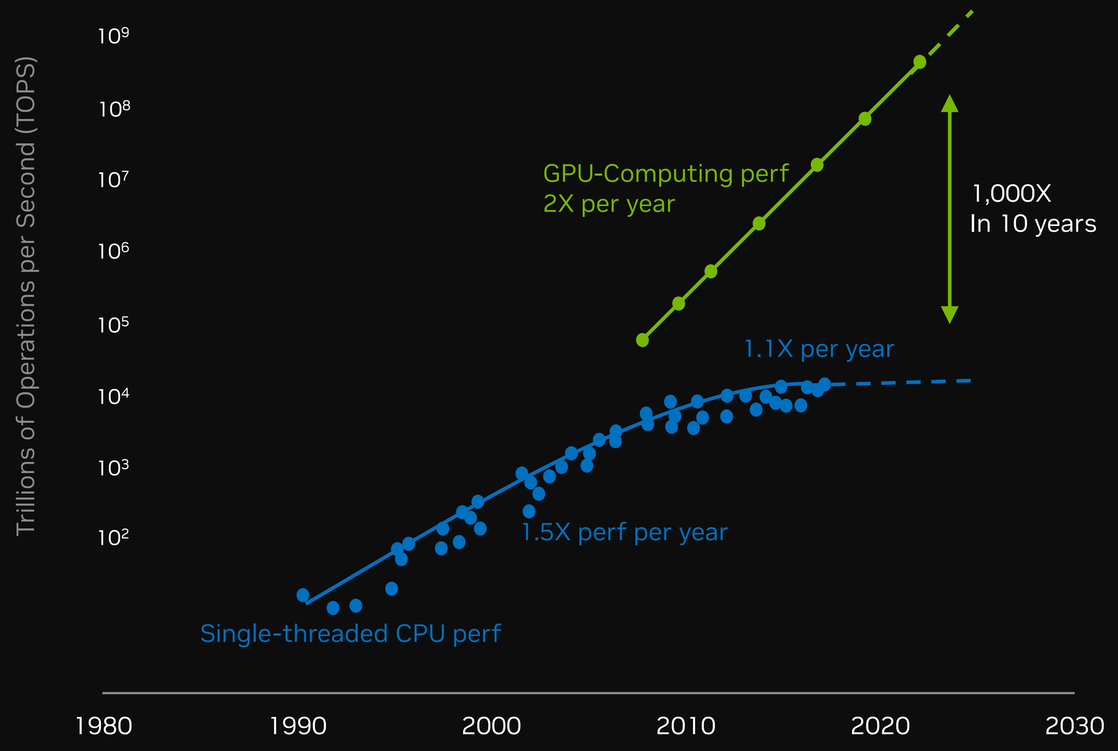
Advancing computing in the post-Moore's law era

Accelerated computing is needed to tackle the most impactful opportunities of our time—like AI, climate simulation, drug discovery, ray tracing, and robotics.

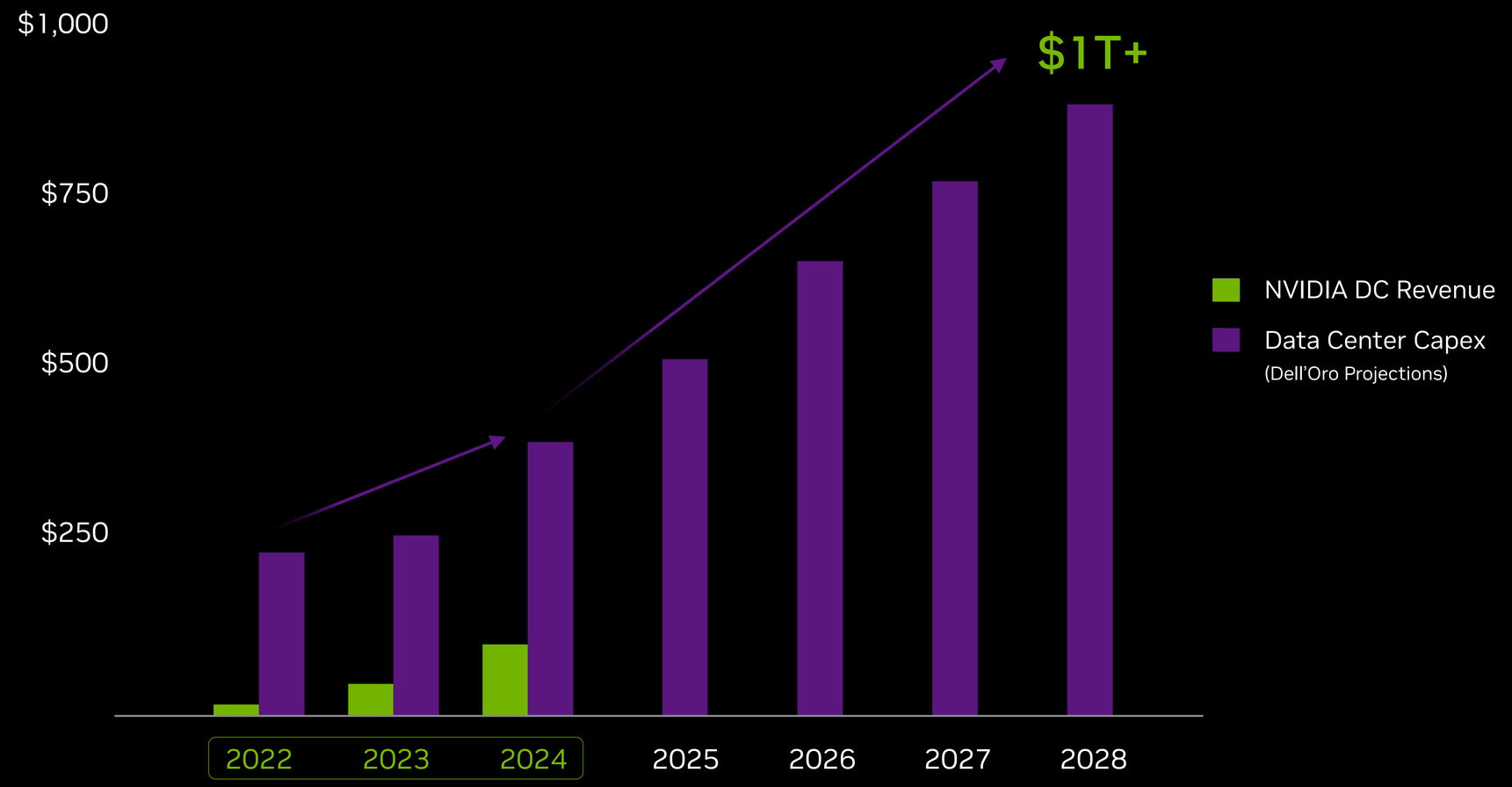
NVIDIA is uniquely dedicated to accelerated computing—working top-to-bottom, refactoring applications and creating new algorithms, and bottom-to-top inventing new specialized processors, like RT Cores and Tensor Cores.

"It's the end of Moore's law as we know it."
—John Hennessy, Oct 2018

"Moore's law is dead."
—Jensen Huang, GTC 2013

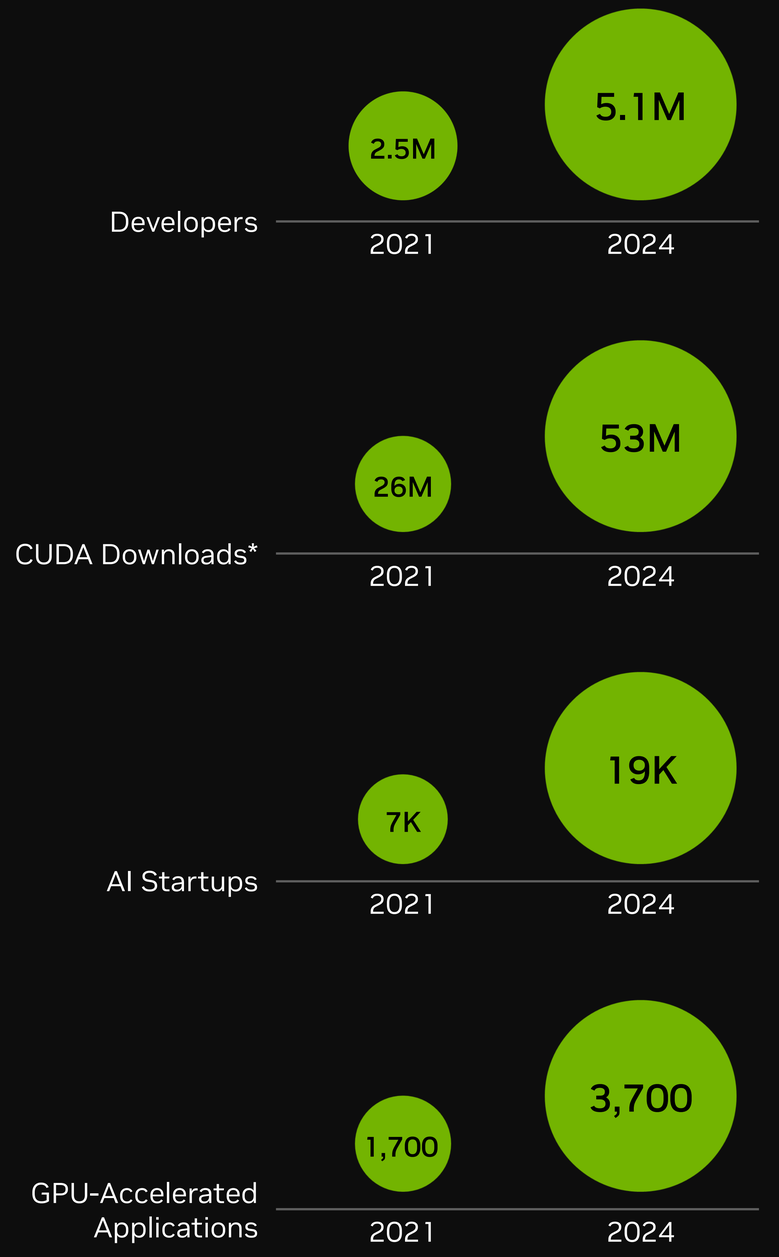


Computing at Inflection Point



NVIDIA's Accelerated Computing Ecosystem

- The NVIDIA accelerated computing platform has attracted the largest ecosystem of developers, supporting a rapidly growing universe of applications and industry innovation.
- Developers can engage with NVIDIA through CUDA—our parallel computing programming model introduced in 2006—or at higher layers of the stack, including libraries, pretrained AI models, SDKs, and other development tools.



*Cumulative

AI Driving a Powerful Investment Cycle and Significant Returns

AI can augment creativity and productivity by orders of magnitude across industries

AI Agents will take action to automate tasks at superhuman speed, transforming businesses and freeing workers to focus on other tasks.

Copilots based on LLMs will generate documents, answer questions, or summarize missed meetings, emails, and chats—adding hours of productivity per week. Specialized for fields such as **software development, legal services** or **education** and can boost productivity by as much as 50%.

Social media, search, and e-commerce apps are using deep recommenders to offer more relevant content and ads to their customers, increasing engagement and monetization.

Creators can generate stunning, photorealistic images with a single text prompt—compressing workflows that take days or weeks into minutes in industries from advertising to game development.

Call center agents augmented with AI chatbots can dramatically increase productivity and customer satisfaction.

Drug discovery and financial services are seeing order-of-magnitude workflow acceleration from AI.

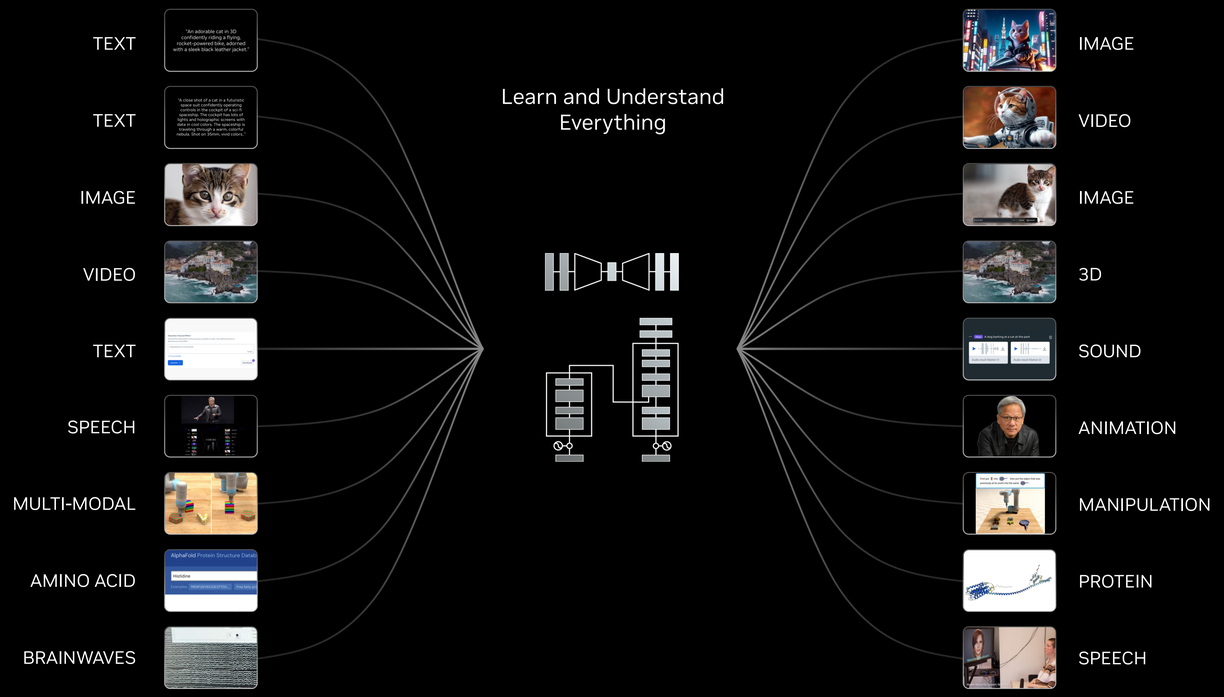
Manufacturing workflows are reinvented and automated through generative AI and robotics, boosting productivity.

NVIDIA

Source: Goldman Sachs, Cowen, Statista, Capital One, Wall Street Journal, Resource Watch, NVIDIA internal analysis

Generative AI

The most important computing platform of our generation



The era of generative AI has arrived, unlocking new opportunities for AI across many different applications.

Generative AI is trained on large amounts of data to find patterns and relationships, learning the representation of almost anything with structure.

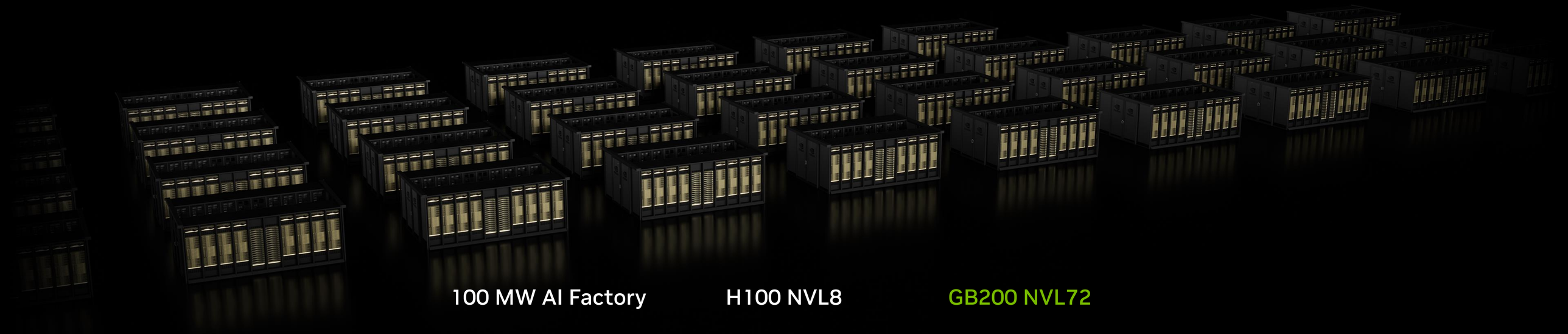
It can then be prompted to generate text, images, video, code, or even proteins.

For the very first time, computers can augment the human ability to generate information and create.

1,600+ generative AI companies are building on NVIDIA.

Blackwell 40X Hopper Inference Performance

NVLink flops per watt ~ AI factory output



100 MW AI Factory	H100 NVL8	GB200 NVL72	
GPU Dies	45K	**85K**	
Racks	1,400	**600**	
Data Center Productivity	300M	**12,000M**	40X More Token Revenue

DeepSeek R1 Context and Generation, , ISL=32K, OSL=8K @ Pareto optimal TPS/User

AI Factories—A New Class of Data Centers

Production of digital intelligence tokens

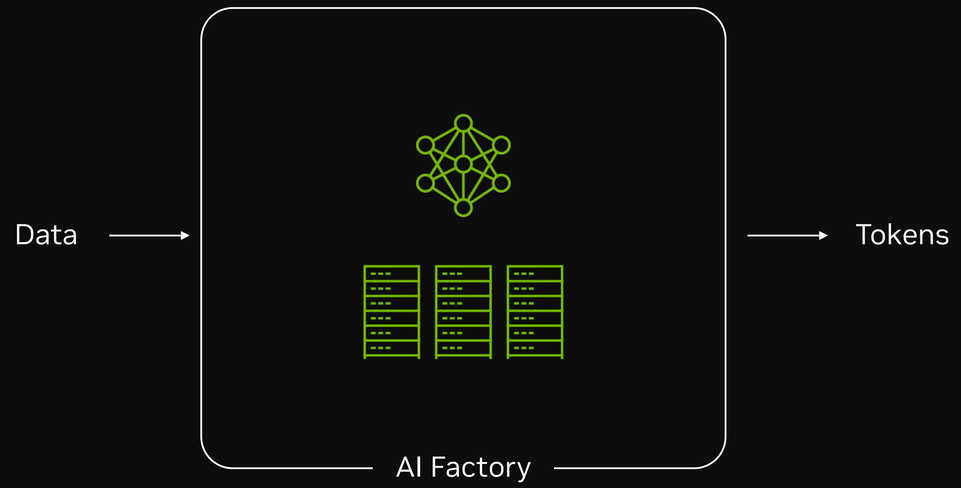
AI factories are a new form of computing infrastructure. Their purpose is not to store user and company data or run ERP and CRM applications. AI factories are highly optimized systems purpose-built to process raw data, refine it into models, and produce monetizable tokens with great scale and efficiency.

In the AI industrial revolution, data is the raw material, tokens are the new commodity, and NVIDIA is the token generator in the AI factory.

Every company will produce digital intelligence. Tokens will be transformed into intelligent responses and actions of digital nurses, tutors, customer service agents, chip designers, manufacturing robots and autonomous cars, Weather prediction agents will warn us of storms. Some companies will build and operate AI factories, while others will rent.

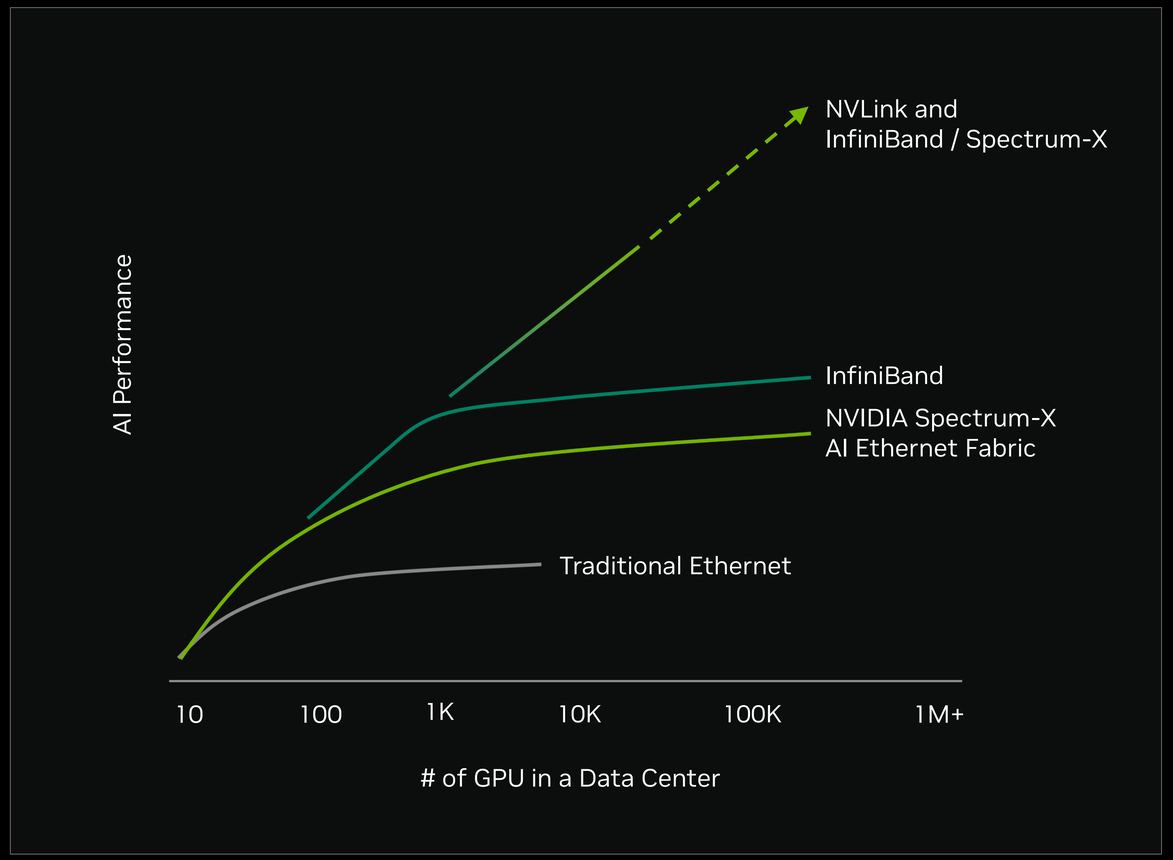
Countries are awakening to the need to treat their data as a national resource and AI factories as an essential national infrastructure. Data encodes a nation's history, knowledge, and culture, and can be transformed into the sovereign AI for its companies, startups, universities, and governments.

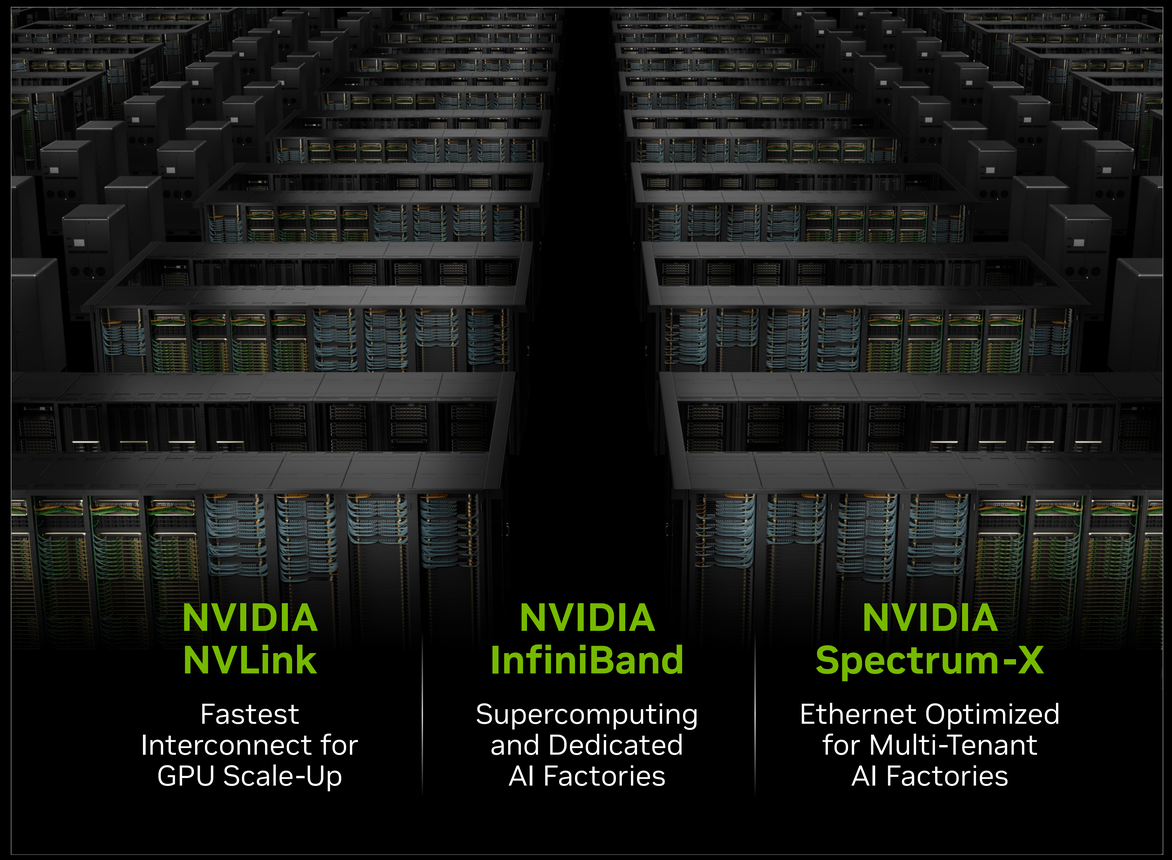
NVIDIA builds the complete AI system and licenses **NVIDIA AI Enterprise**, the AI stack and operating system for AI factories.



Extending NVIDIA Networking to Scale Up & Scale Out AI in Any Data Center

New NVLink and Spectrum-X increase networking opportunity beyond InfiniBand to every data center





Generative AI is a Data Center-Scale Computing Workload

Limitless scaling with NVLINK + InfiniBand or Spectrum-X

NVIDIA

Accelerated Computing Starts With CUDA Libraries

Delivering up to 200X speedup across major workloads

Unlike CPU general-purpose computing, GPU-accelerated computing requires software and algorithms to be redesigned. Software is not automatically accelerated in the presence of a GPU or accelerator.

NVIDIA CUDA libraries encapsulate NVIDIA-engineered algorithms that enable applications to be accelerated on NVIDIA's installed base. They deliver dramatically higher performance—compared to CPU-only alternatives—across application domains, including AI and high-performance computing, and significantly reduce runtime, cost, and energy, while increasing scale.

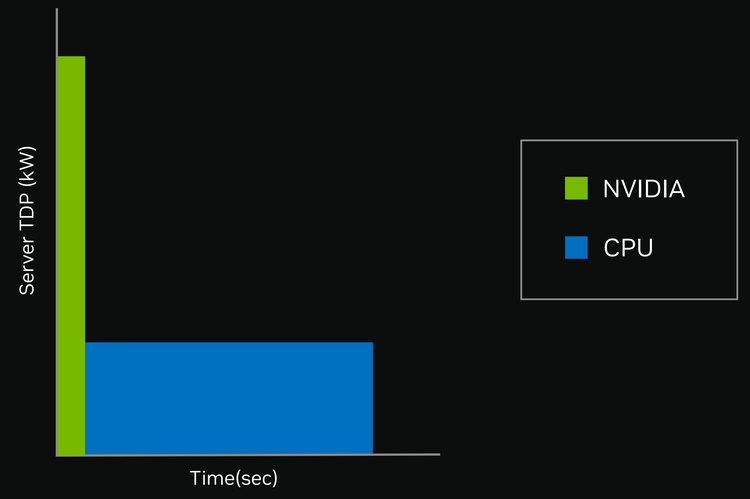
With over 400 CUDA libraries, NVIDIA can address many major workloads across a wide range of industries. As new libraries become available, they unlock new markets adding to our long-term opportunity.

~200X	**Data Processing**	cuVS, cuDF-Spark, cuDF-pandas, cuDF-Polars, cuGraph, cuML, XGBoost, RAPIDS, NeMo Curator, cuSOLVER, cuIO
~200X	**Computer Vision**	CV-CUDA, Deepstream, TAO, Holoscan, cuCIM, TensorRT, Triton Inference Server, DALI, nvImageCodec, cuDNN, nvJPEG, nvJPEG2000, nvTIFF, NPP, Video Codec SDK, Magnum IO, NCCL, cuVS, DALI
~100X	**Science**	Earth-2 CorrDiff, Holoscan, Parabricks, Monai, Modulus, Warp, cuLitho, cuQuantum, CUDA-Q, AmgX, cuDSS, cuFFT, cuSOLVER, cuBLAS, cuSPARSE, cuTENSOR, cuGraph, Magnum IO, NCCL, NVSHMEM, RAFT, cuNumeric, Sionna
~100X	**Deep Learning**	cuDNN, CUTLASS, Megatron, TensorRT, TRT LLM, NCCL, NV-Triton, CUDA-optimized PyTorch, Tensorflow, Triton, Jax
~50X	**Recommender Systems**	Merlin, HugeCTR, TensorRT, Triton Inference Server, cuBLAS, cuDNN, cuFFT, cuSPARSE, CUTLASS, Magnum IO, NCCL, cuVS
~30X	**Speech AI**	Riva, TensorRT, Triton Inference Server, NeMo, cuBLAS, cuDNN, cuFFT, CUTLASS
~100X	**Agentic and Physical AI**	ACE, Riva, Nemo, Tokkio Digital Human, Holoscan, Metropolis, Omniverse, Isaac, DRIVE, cuLitho, cuMotion, cuOpt, Aerial CUDA-accelerated RAN, Sionna, fVDB, PhysX, Warp, NVblox

Accelerated Computing Is Sustainable Computing

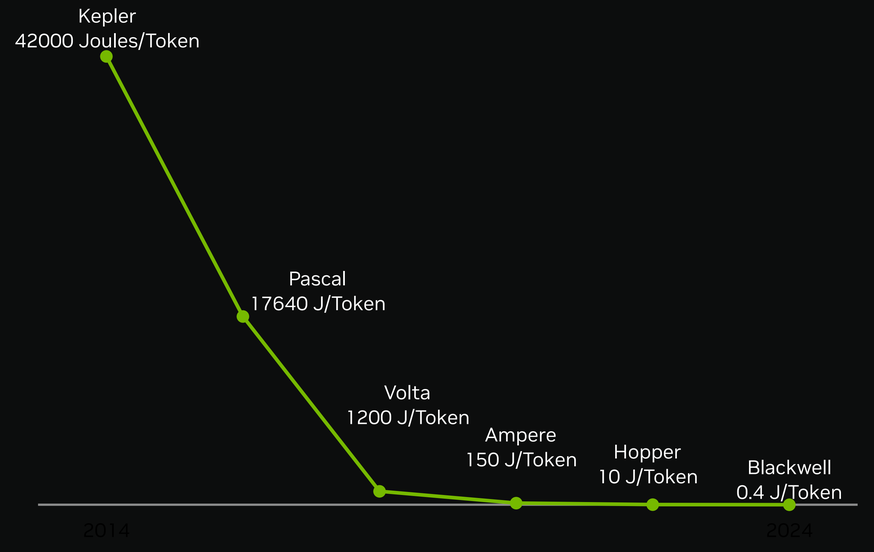
Order of magnitude more energy efficient

Accelerated computing requires **higher peak power consumption** than CPUs, however, completes workloads **significantly faster and consumes less total energy**



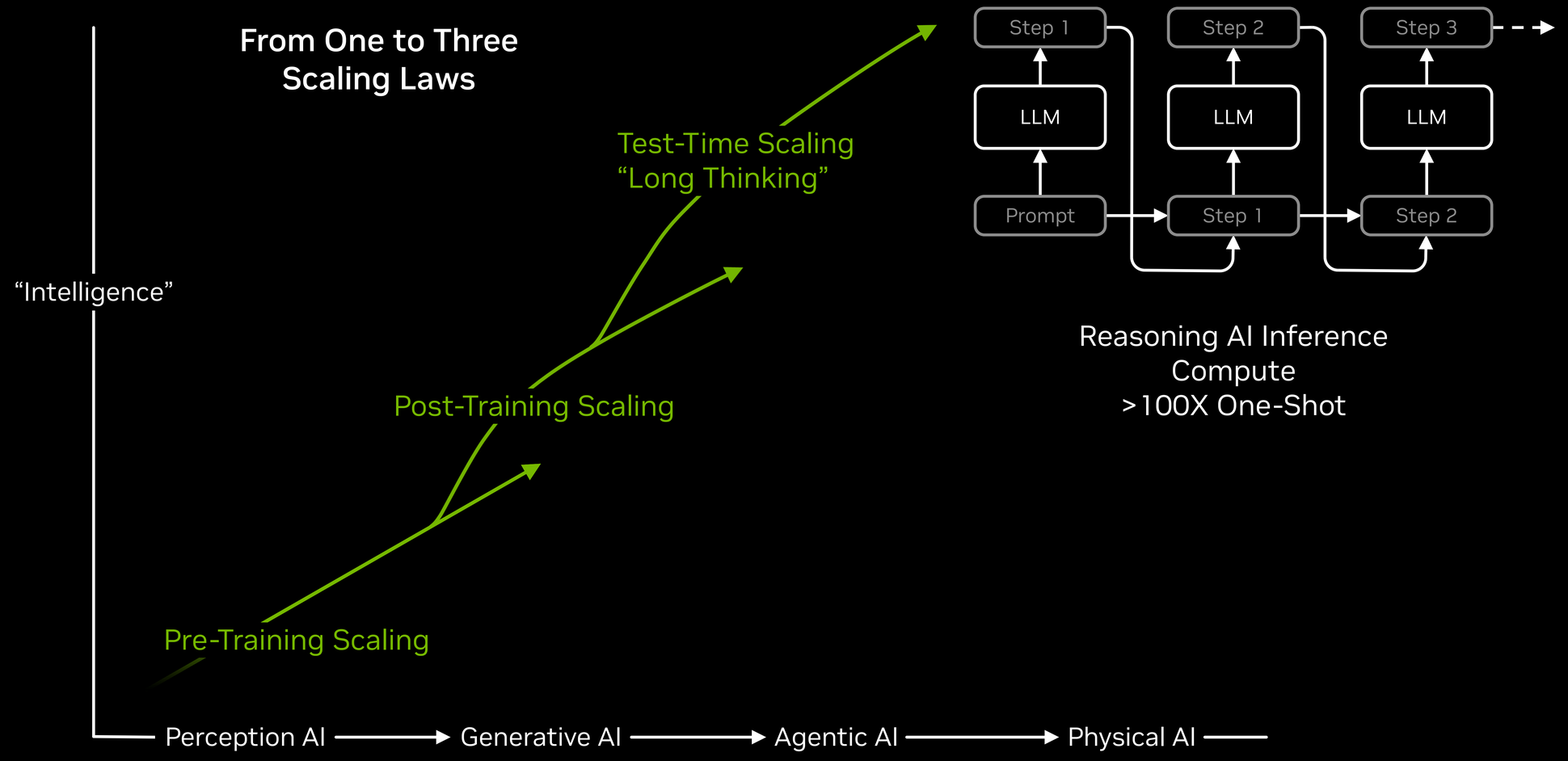
Energy Usage in AI Inference

Accelerated computing enables **full-stack optimization** from algorithm to GPU architecture, such as Tensor Core Transformer Engine; **LLM energy efficiency improved 100,000X** in the past 10 years



GPT-MoE-1.8T energy per token

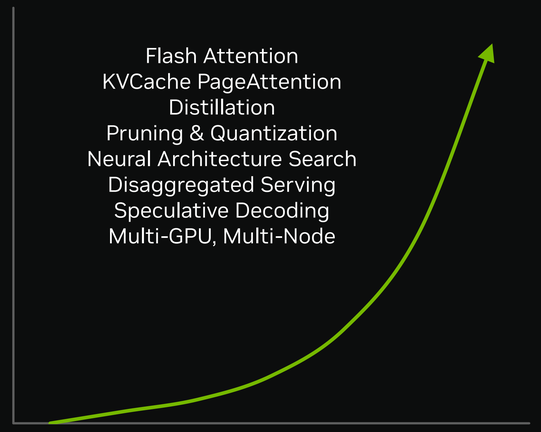
AI Scaling Laws Drive Exponential Demand for Compute



NVIDIA Is the Leading Inference Platform

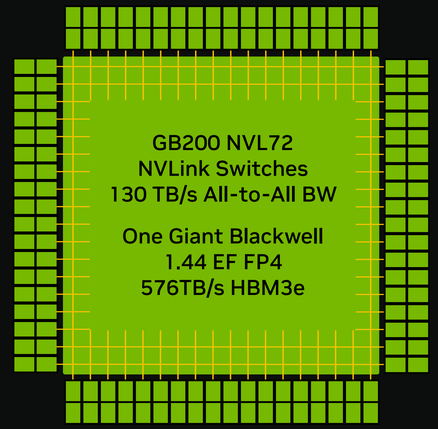
Inference compute scales exponentially with “long thinking”

Hopper inference performance increased 5X in 1 year with rapid algorithm innovations enabled by rich NVIDIA CUDA ecosystem



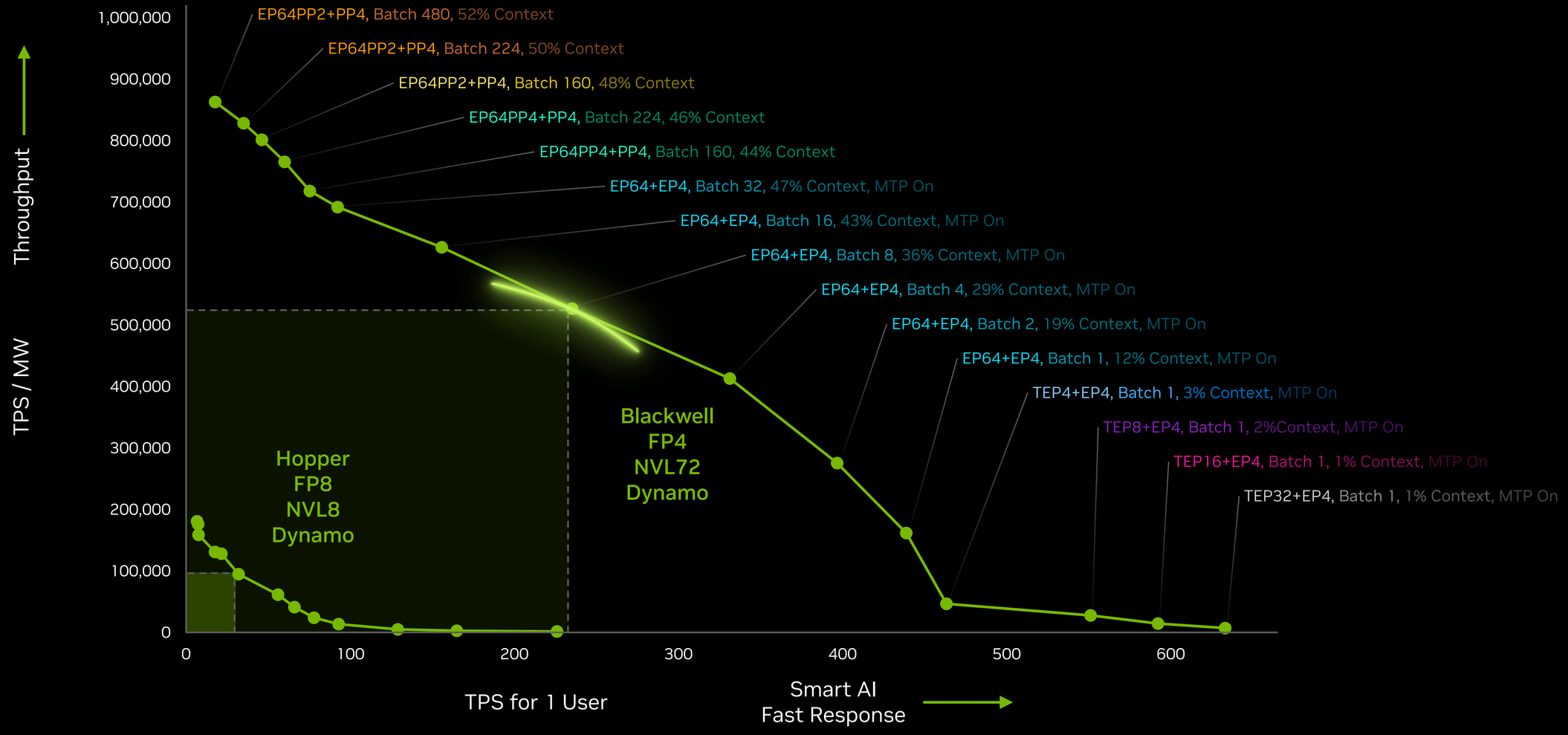
Installed base & CUDA ➔ rapid software innovation ➔ performance ➔ lower inference cost ➔ increase demand ➔ increase installed base

Inference compute scaling exponentially with large multimodal models, chain-of-thought, reasoning, agents, and low-latency responses



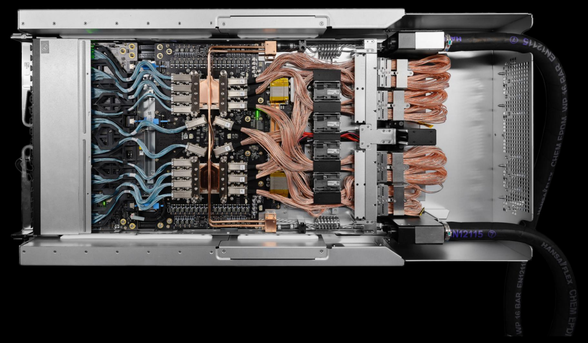
Blackwell 40X Hopper

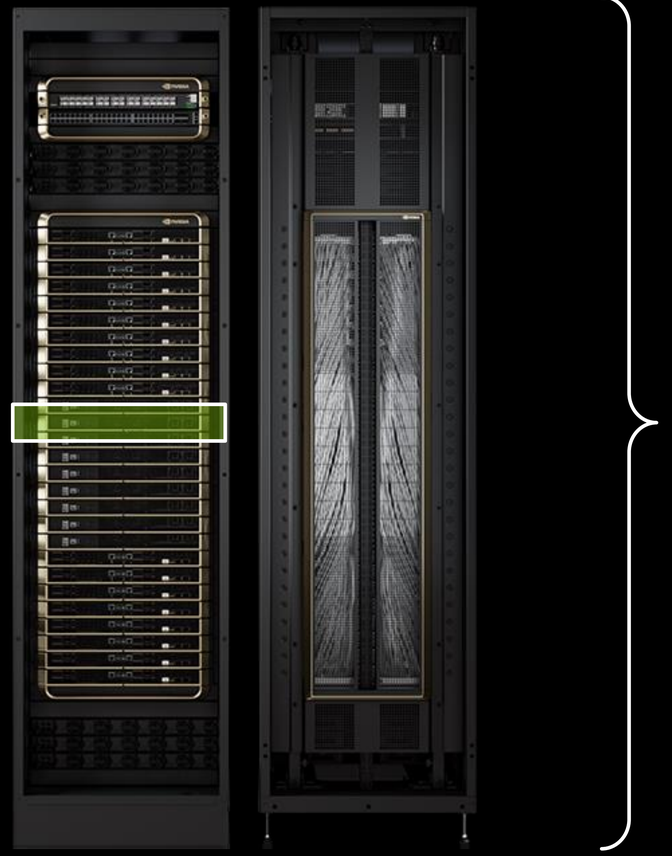
FP4 | NVL72 | Dynamo | TRT-LLM continuous optimization | 32K ISL/8K OSL



NVIDIA NVLink Enables New Level of AI Training & Inference Scaling

NVLink Switch



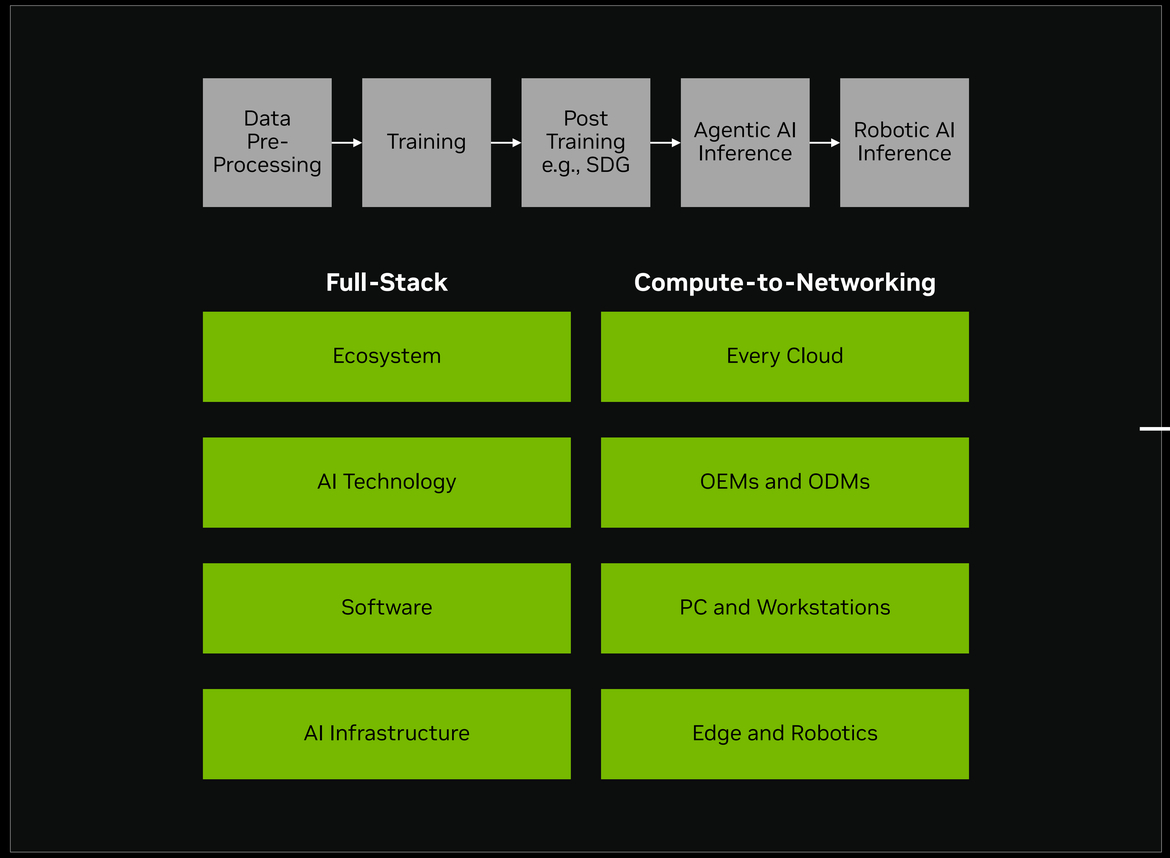


GB200 NVL72
NVLink Switches
130 TB/s All-to-All BW

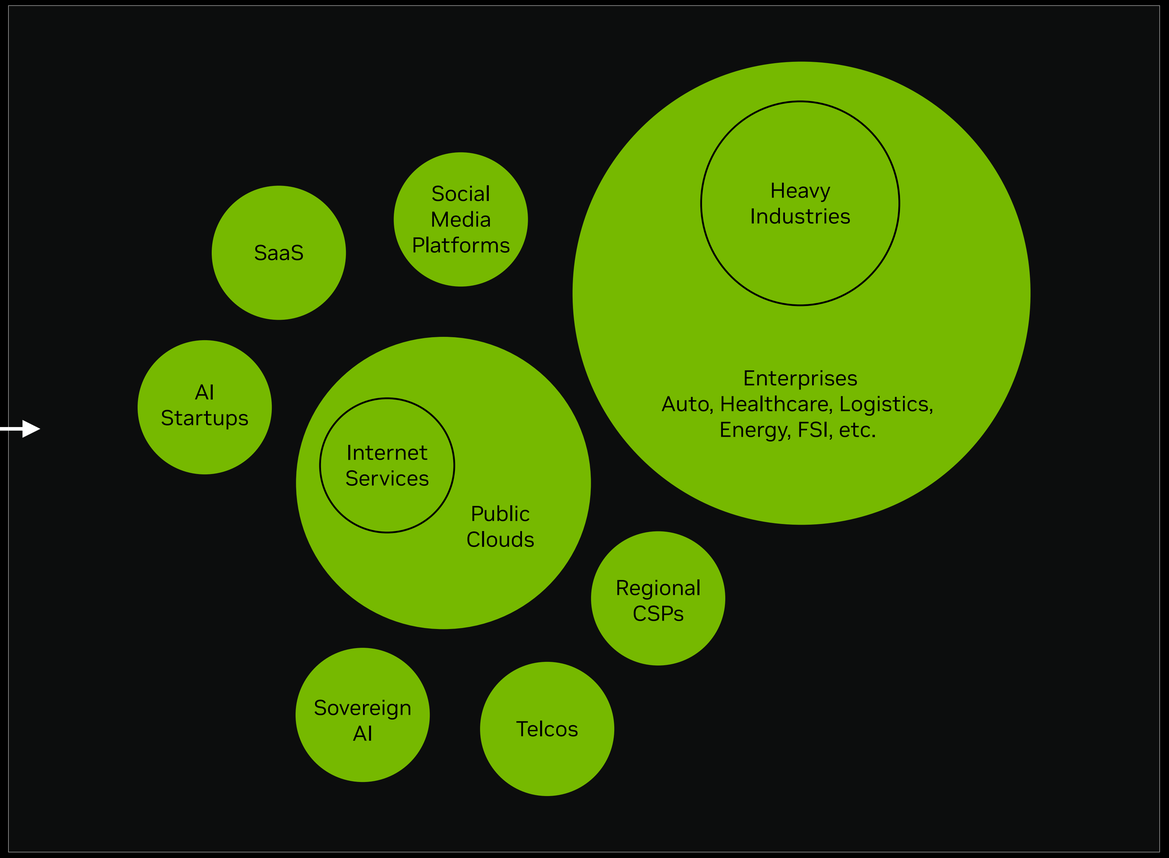
One Giant Blackwell
1.44 EF FP4
576TB/s HBM3e

NVIDIA AI Platform and Ecosystem Reaches Every Market

Every workload to address the world's industries

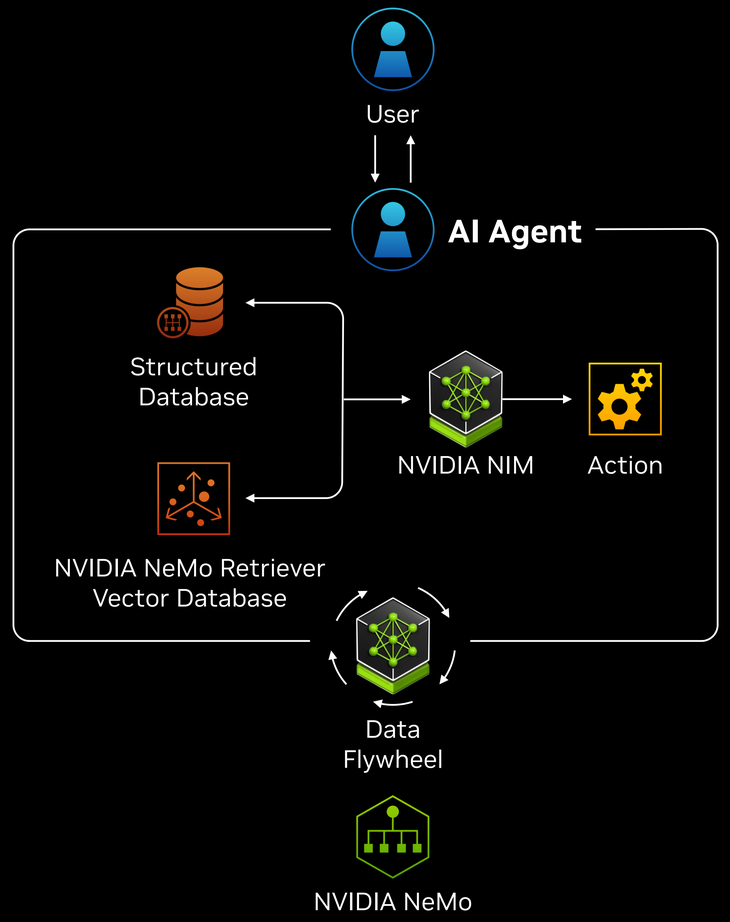


Accelerate Every Workload



Full-Stack, Entire AI Infrastructure

NVIDIA AI Enterprise Enables IT Ecosystem With State-of-the-Art AI Models and Libraries to Build Agentic AI



NVIDIA AI Enterprise Ecosystem

System Integrators

accenture Deloitte. . . .

Enterprise ISVs

amdocs cadence SAP servicenow SYNOPSYS . . .

Data Platforms

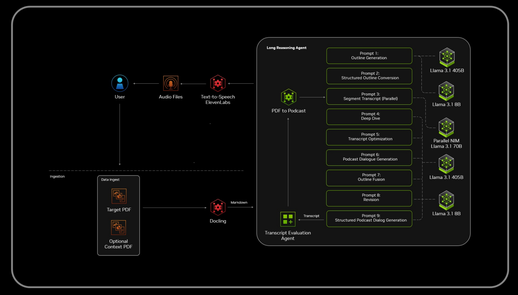
CLOUDERA Dropbox NetApp ORACLE VAST . . .

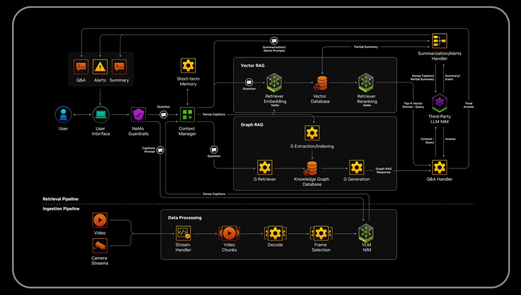
Cloud and On-Prem Infrastructure

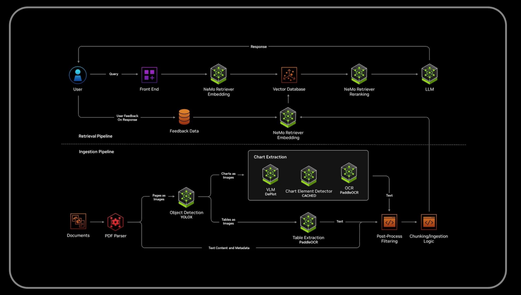
aws DELL Technologies Google Cloud Hewlett Packard Enterprise Microsoft Azure ORACLE CLOUD Infrastructure . . .

Reinventing $500B Enterprise IT for the Age of AI

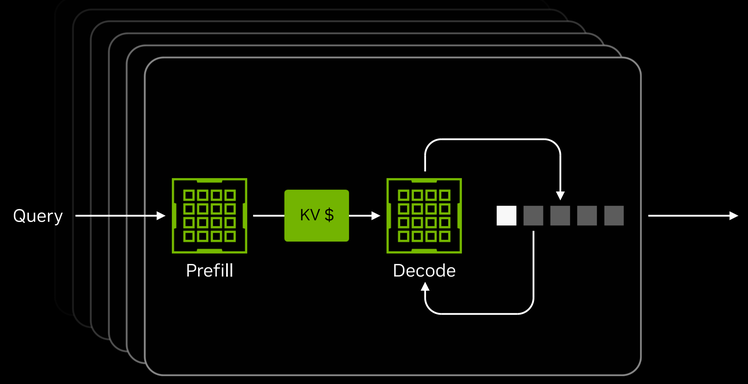
AI







OS



Infrastructure

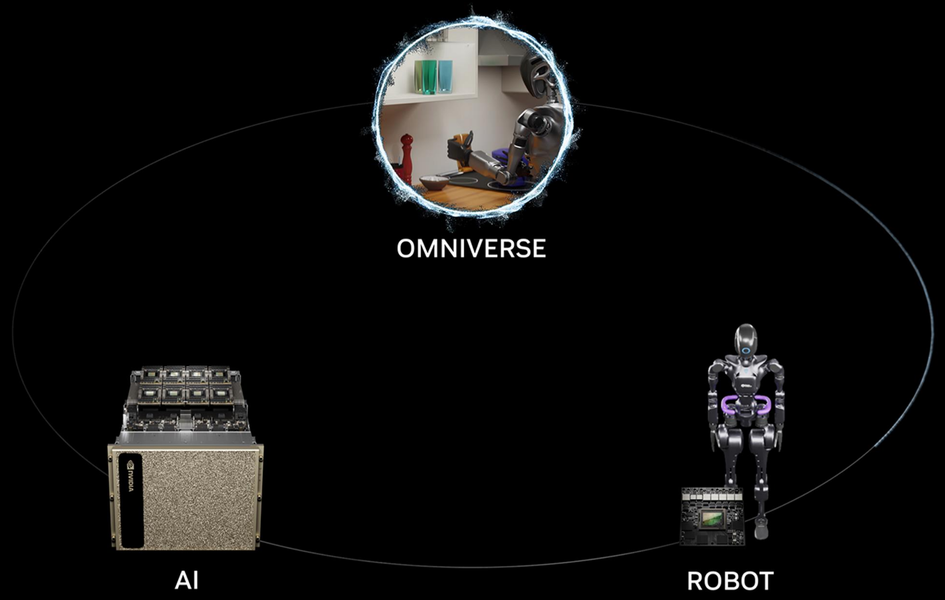
Compute

Networking

Storage

NVIDIA

NVIDIA Omniverse and AI Revolutionizing Manufacturing and Robotics



The next AI wave is physical AI—models that can perceive, understand, and interact with the physical world. Physical AI will embody robotic systems—from autonomous vehicles to industrial robots and humanoids, to warehouses and factories.

Three computers and software stacks are required to build physical AI:
NVIDIA AI on DGX to train the AI model, NVIDIA Omniverse on OVX to teach, test, and validate the AI model's skills, and NVIDIA AGX to run the AI software on the robot.

Enterprises license NVIDIA Omniverse at $4,500 per GPU per year.

100M Cars

Billions in Future

10M Factories

200K Warehouses

Sovereign AI

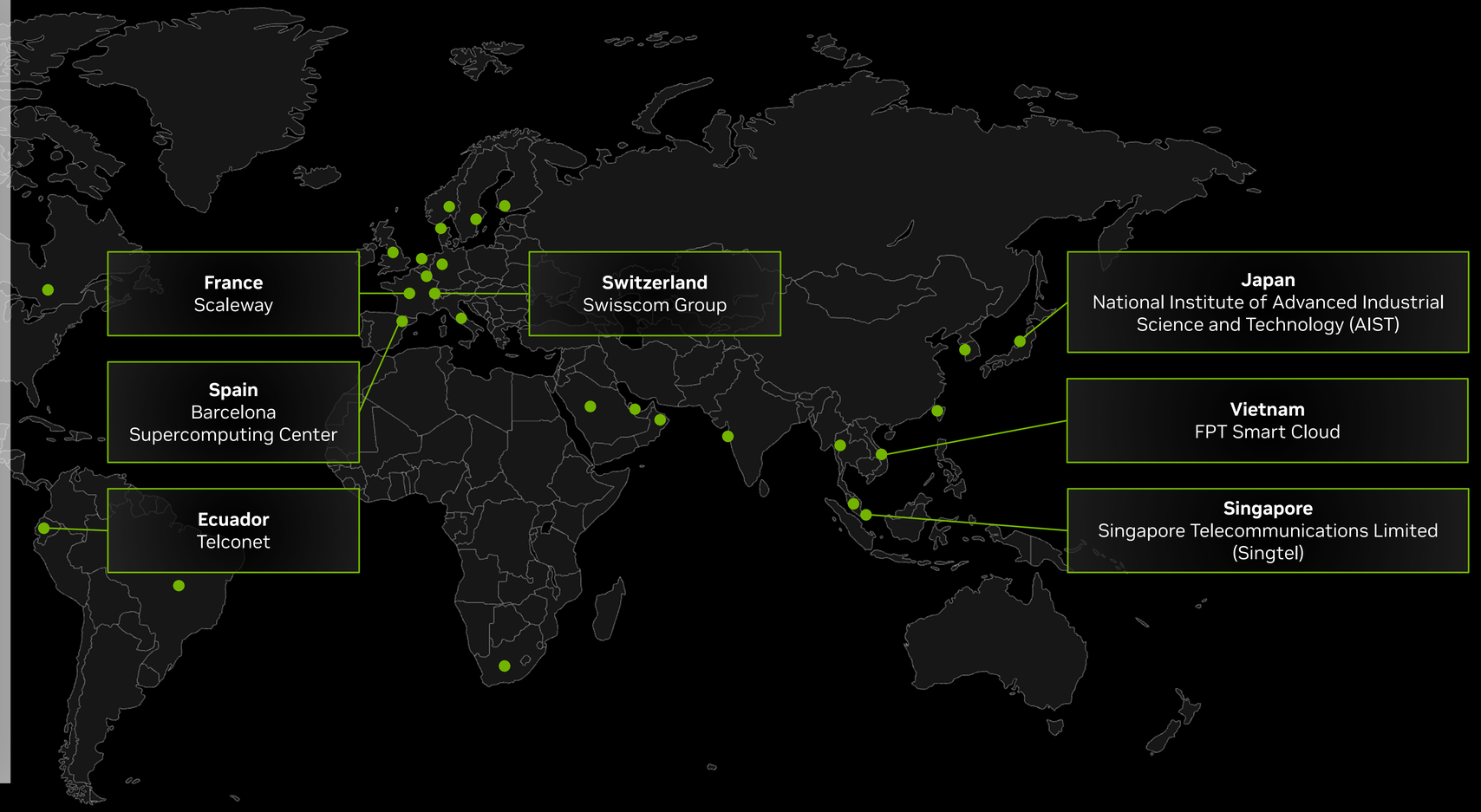
Nations produce AI using their own data, infrastructure, workforce, and business networks

Sovereign AI

Nations are awakening to the imperative to produce AI using their own infrastructure, data, workforces, and business networks. Nations are building domestic computing capacity.

Some governments operate sovereign AI clouds in collaboration with state-owned telecommunications providers or utilities. Other governments partner with local cloud providers to deliver a shared AI computing infrastructure for public and private-sector use.

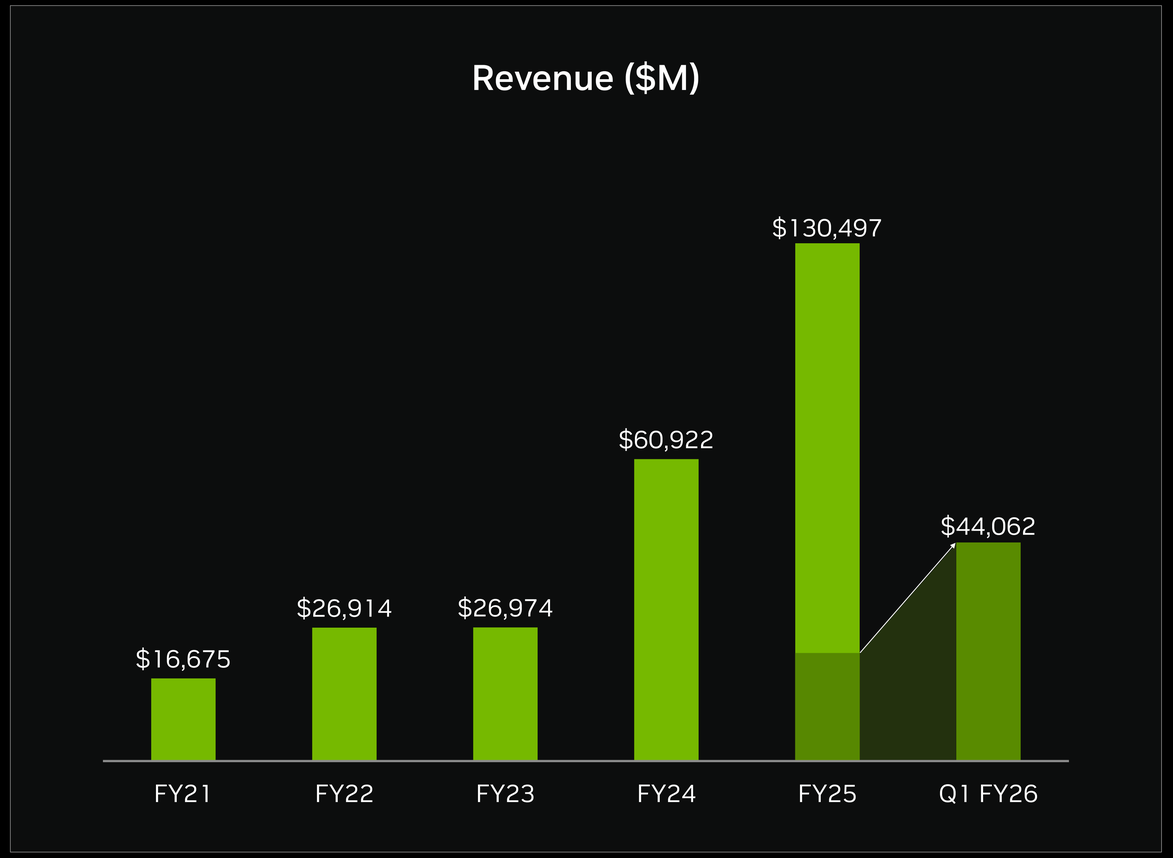
NVIDIA's ability to help build AI infrastructure with our end-to-end compute-to-networking technologies, full-stack software, AI expertise, and rich ecosystem of partners and customers allows sovereign AI and regional cloud providers to jump-start their countries' AI ambitions.

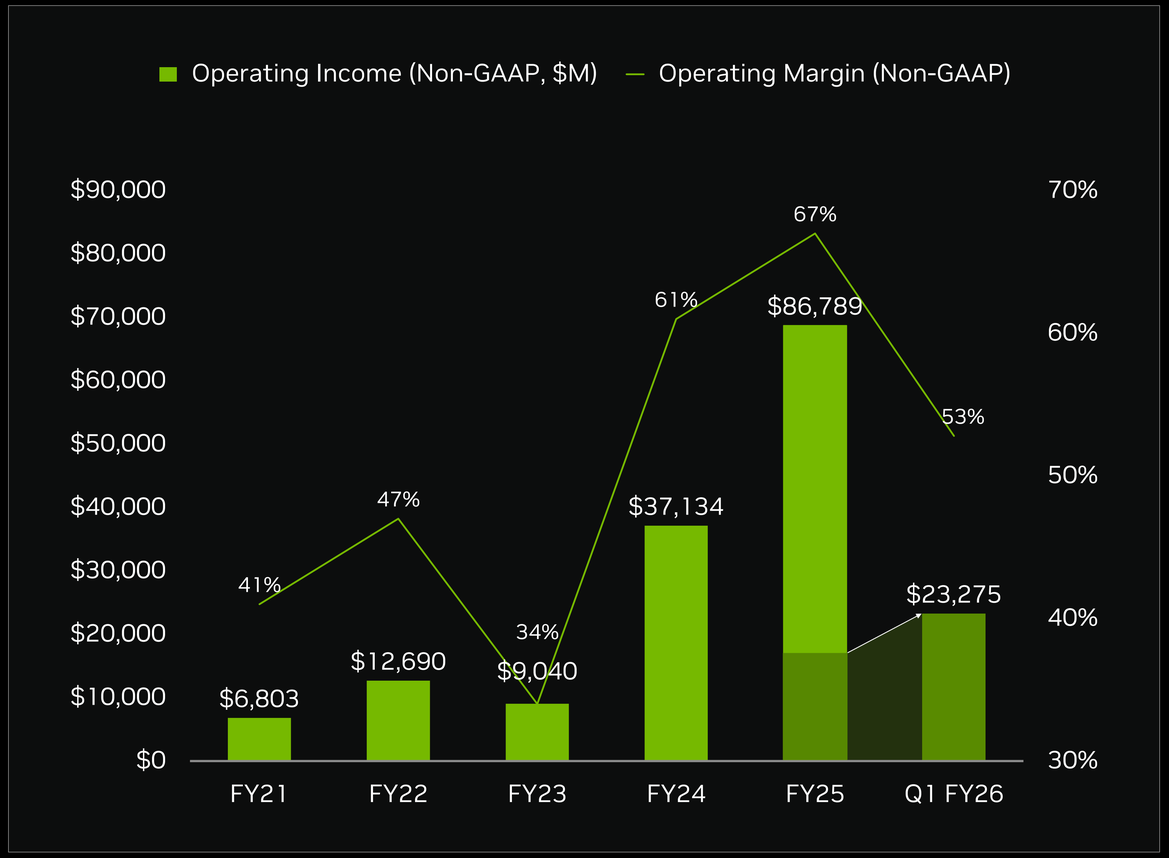


● Location of NVIDIA sovereign AI partners

NVIDIA

Driving Strong and Profitable Growth

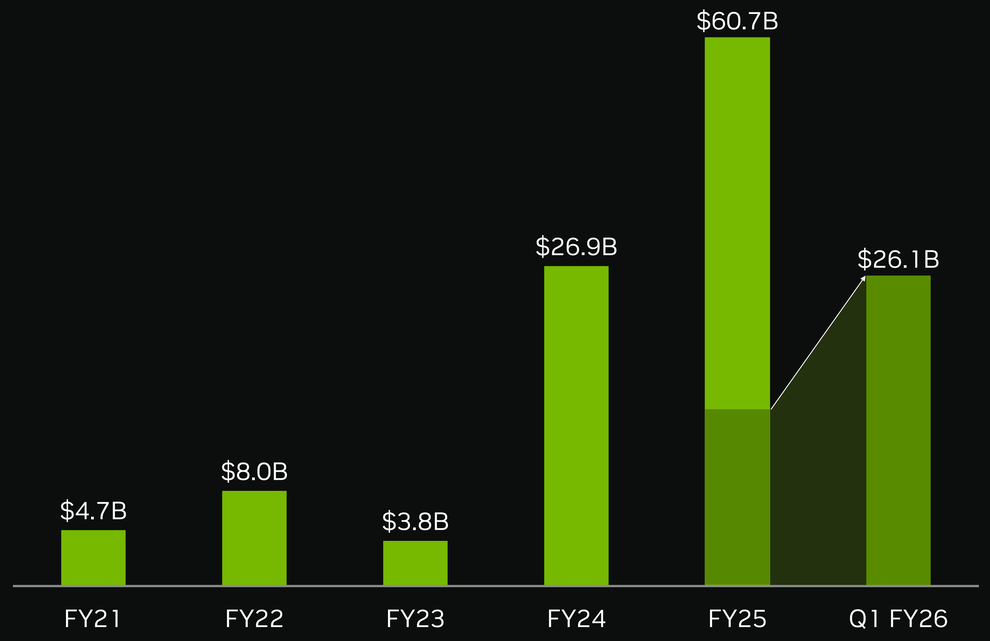




Fiscal year ends in January. Refer to Appendix for reconciliation of Non-GAAP measures. Operating margins rounded to the nearest percent. Q1 FY26 included a $4.5 billion charge associated with H20 excess inventory and purchase obligations.

Strong Cash Flow Generation

Free Cash Flow (Non-GAAP)



Fiscal year ends in January. Refer to Appendix for reconciliation of Non-GAAP measures.

Capital Allocation

Share Repurchase

Utilized $14.1B of cash for repurchases in Q1 FY26

$24.3B remaining authorization as of the end of Q1

Dividend

$244M in Q1 FY26

Plan to Maintain[1]

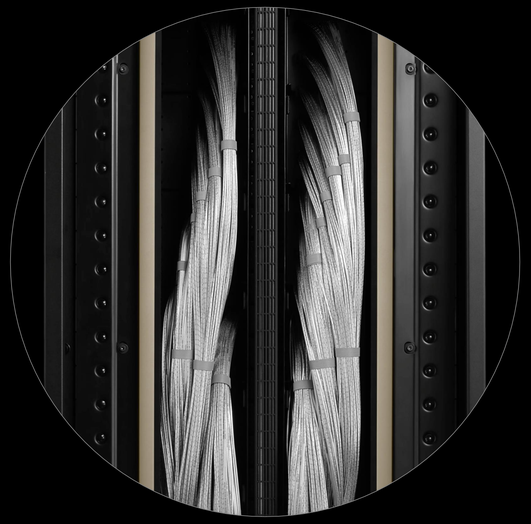
Strategic Investments

Growing Our Talent

Platform Reach and Ecosystem

[1] *Subject to continuing determination by our Board of Directors.*

Our Market Platforms at a Glance



Data Center

88% of FY25 Revenue

FY25 Revenue $115.2B
5-YR CAGR 108%

DGX/HGX/MGX/IGX systems

GPU | CPU | DPU | Networking
NVIDIA AI software

Gaming

9% of FY25 Revenue

FY25 Revenue $11.4B
5-YR CAGR 16%

GeForce GPUs for PC gaming

GeForce NOW cloud gaming

Professional Visualization

1% of FY25 Revenue

FY25 Revenue $1.9B
5-YR CAGR 9%

NVIDIA RTX GPUs for workstations

Omniverse software

Automotive

1% of FY25 Revenue

FY25 Revenue $1.7B
5-YR CAGR 19%

DRIVE Hyperion sensor architecture
with AGX compute

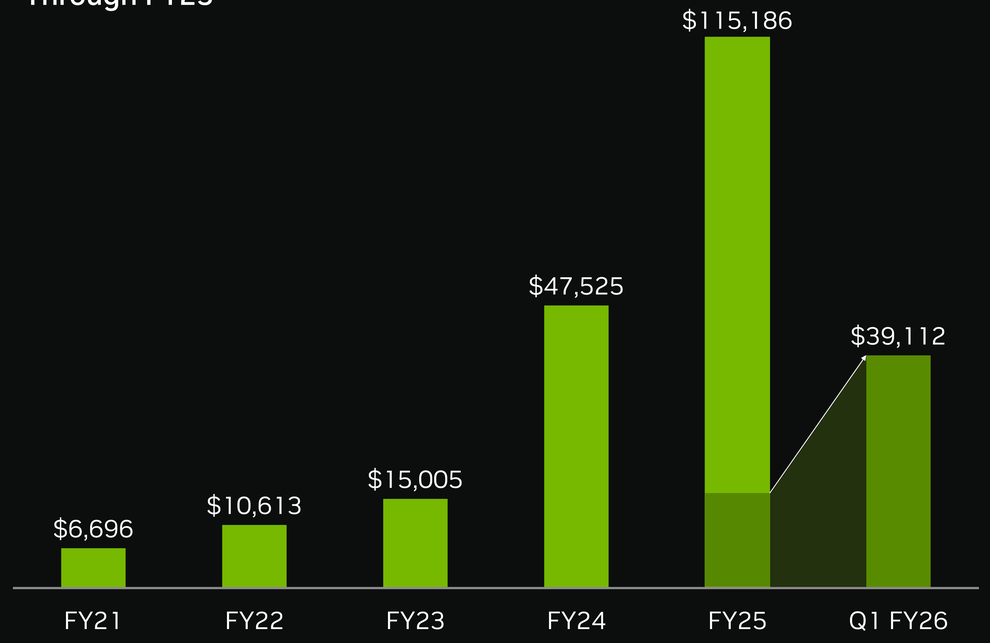
DRIVE AV & IX full-stack software
for ADAS, AV, and AI cockpit

Data Center

The leading accelerated computing platform

Revenue ($M)

108% 5-YR CAGR
Through FY25



Leader in AI and HPC

No. 1 in AI training and inference

Used by all hyperscalers, major cloud computing providers, and over 40,000 companies

Powers over 75% of the TOP500 supercomputers

Growth Drivers

Broad data center platform transition from general-purpose to accelerated computing

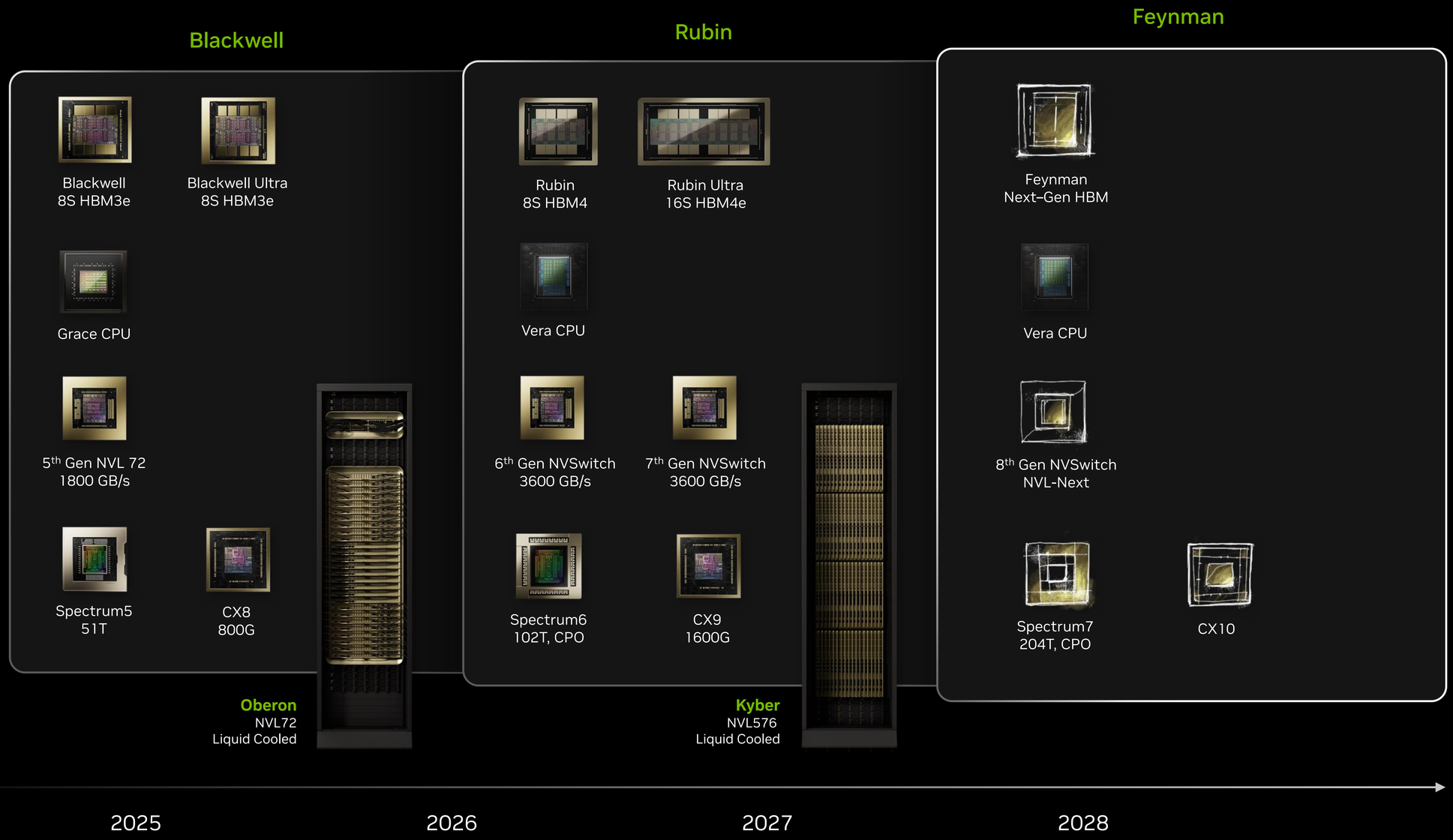
Emergence of AI factories—optimized for refining data and training, inferencing, and generating AI

Broader and faster product launch cadence to meet a growing and diverse set of AI opportunities

NVIDIA AI Enterprise/NIM for building and running enterprise AI applications

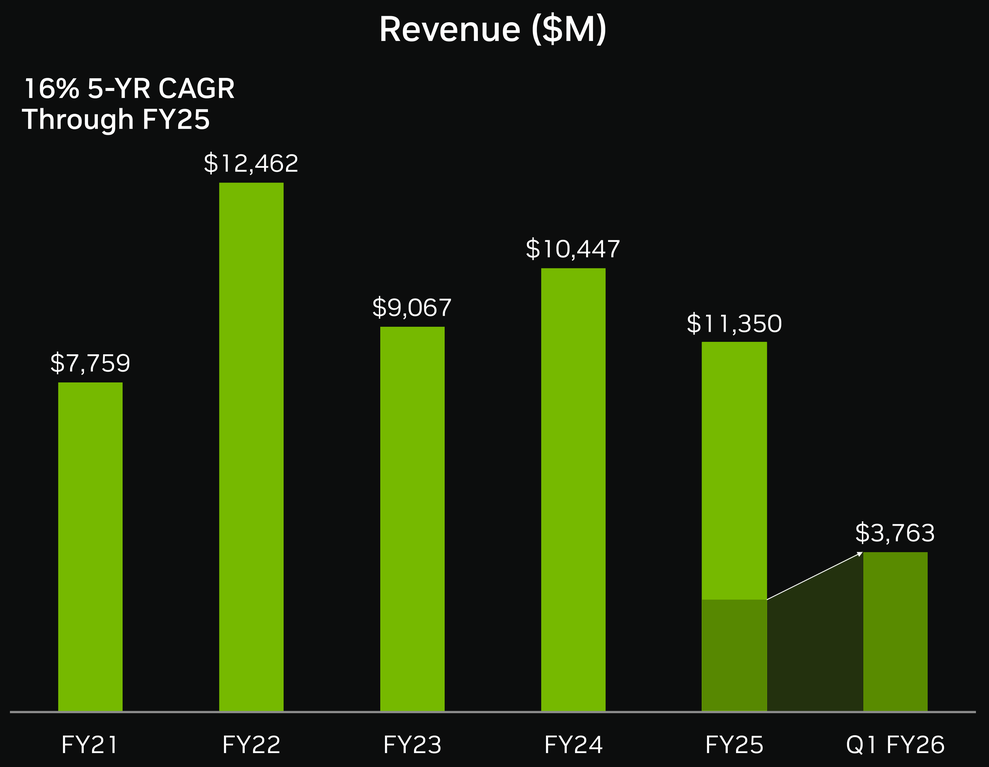
NVIDIA Paves Road to Gigawatt AI Factories

One-Year rhythm | Full-stack | One architecture | CUDA everywhere



Gaming

GeForce—world's largest gaming platform



Leader in PC Gaming

Strong No. 1 market position

15 of the top 15 most popular GPUs on Steam

Leading performance and innovation

200M+ gamers on GeForce

Growth Drivers

Rising adoption of NVIDIA RTX in games

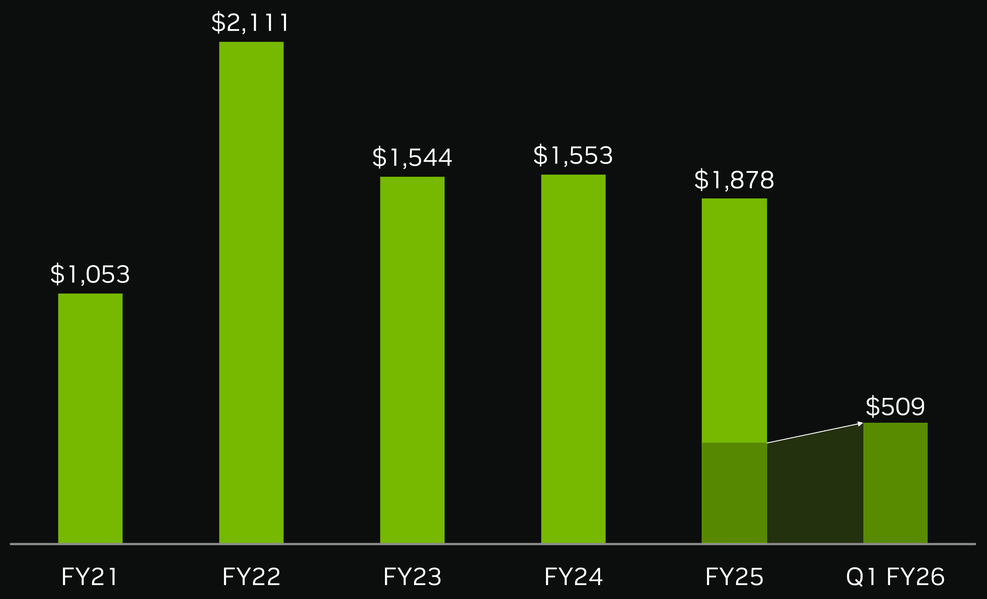
Expanding universe of gamers and creators

Gaming laptops and generative AI on PCs

GeForce NOW cloud gaming

Professional Visualization

Workstation graphics



Leader in Workstation Graphics

95%+ market share in graphics for workstations

45M designers and creators

Strong software ecosystem with over 100 RTX accelerated and supported applications

Growth Drivers

Generative AI adoption across design and creative industries

Enterprise AI development, model fine-tuning, cross-industry

Ray tracing revolutionizing design and content creation

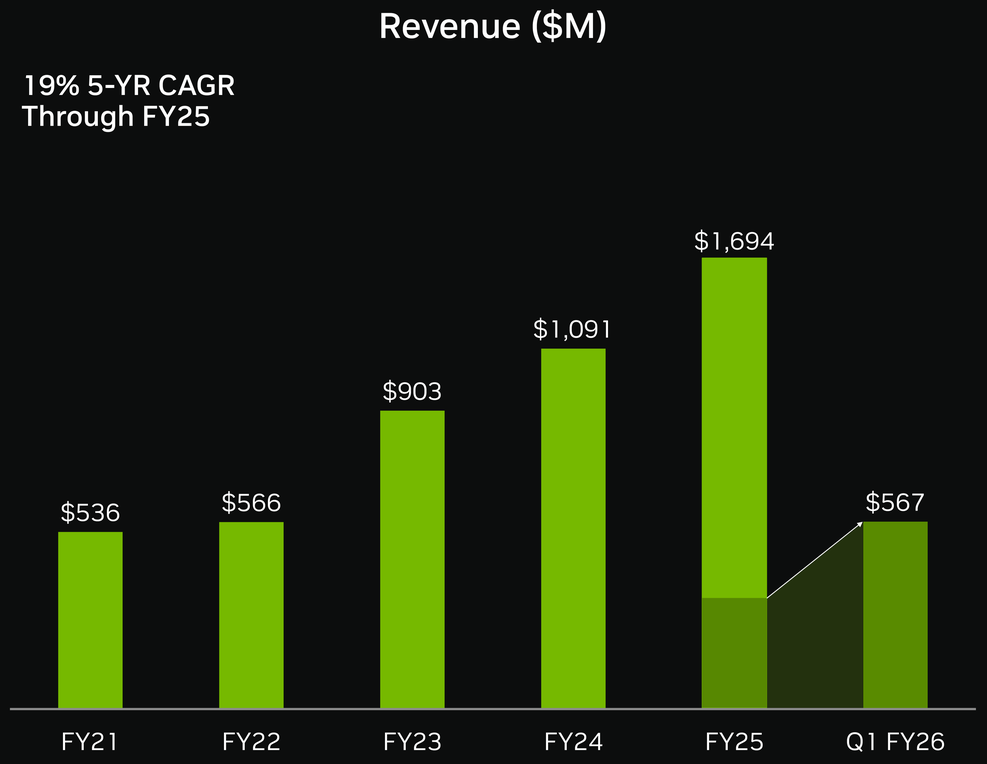
Expanding universe of designers and creators

Omniverse for digital twins and collaborative 3D design

Hybrid work environments

Automotive

Autonomous vehicles and AI cockpits



Leader in Autonomous Driving

NVIDIA DRIVE an end-to-end autonomous vehicle (AV) and AI cockpit platform featuring a full software stack and powered by NVIDIA SoCs (systems-on-a-chip) in vehicles

DRIVE Orin SoC ramp began in FY23

Next-generation DRIVE Thor SoC ramp to begin in FY26

Over 40 customers including 20 of top 30 EV makers, 7 of top 10 truck makers, 8 of top 10 robotaxi makers

Growth Drivers

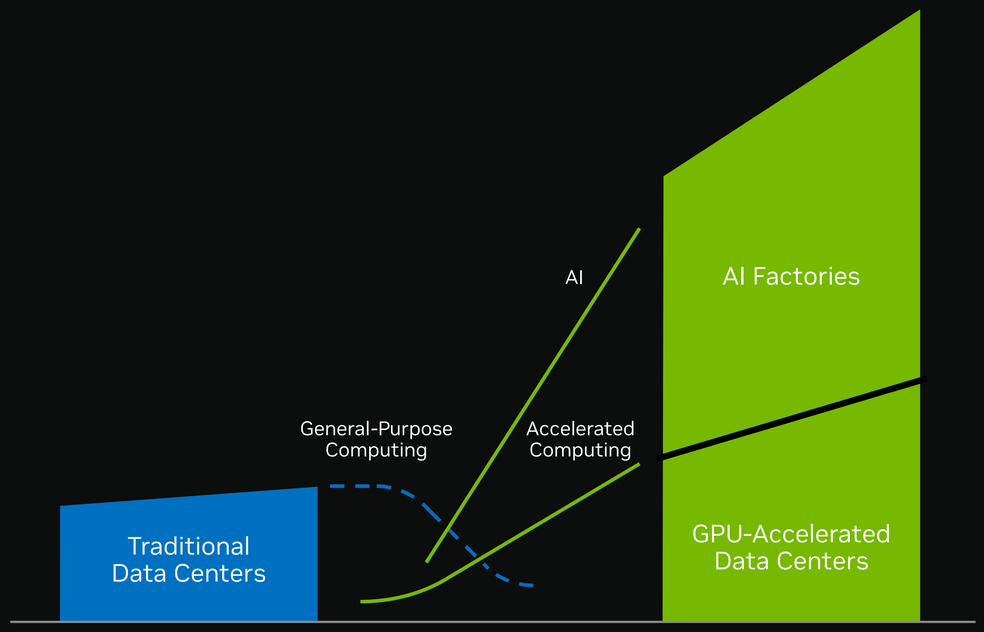
Adoption of centralized car computing and software-defined vehicle architectures

AV software and services:

Mercedes-Benz

Jaguar Land Rover

Accelerated Computing and Generative AI Create Trillion-Dollar Opportunities



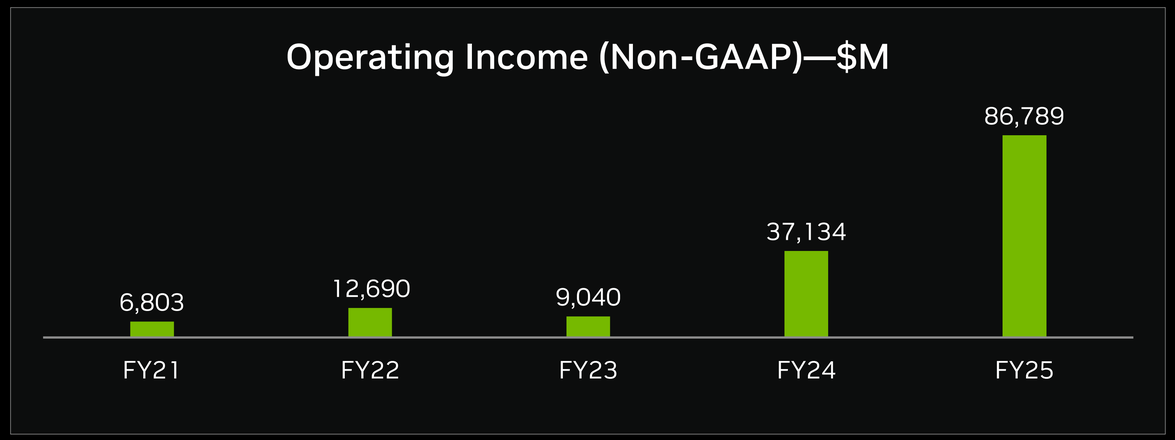
The $1T installed base of general-purpose CPU data center infrastructure is being modernized to a new GPU-accelerated computing paradigm.

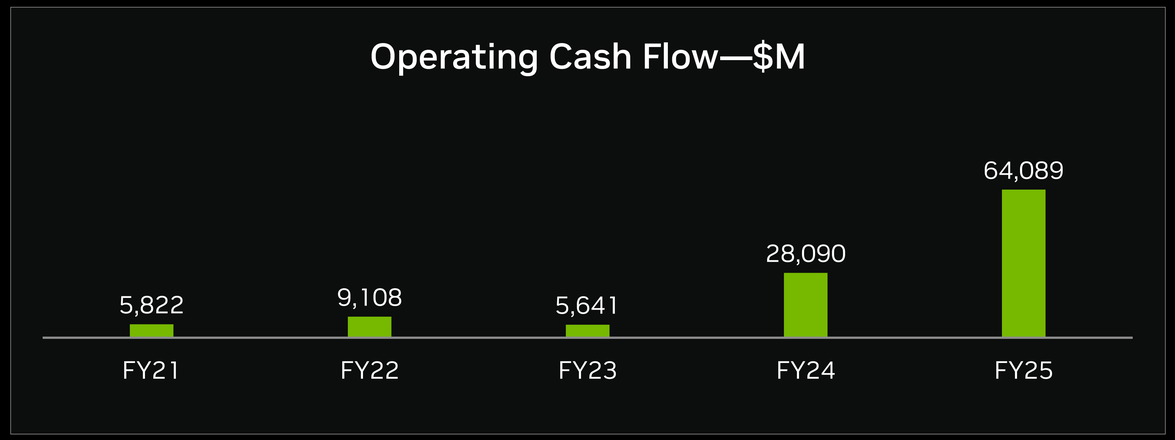
The entire computing stack has been reinvented—from CPU to GPU, from coding to machine learning, from software to generative AI. Computers generate intelligence tokens, a new commodity.

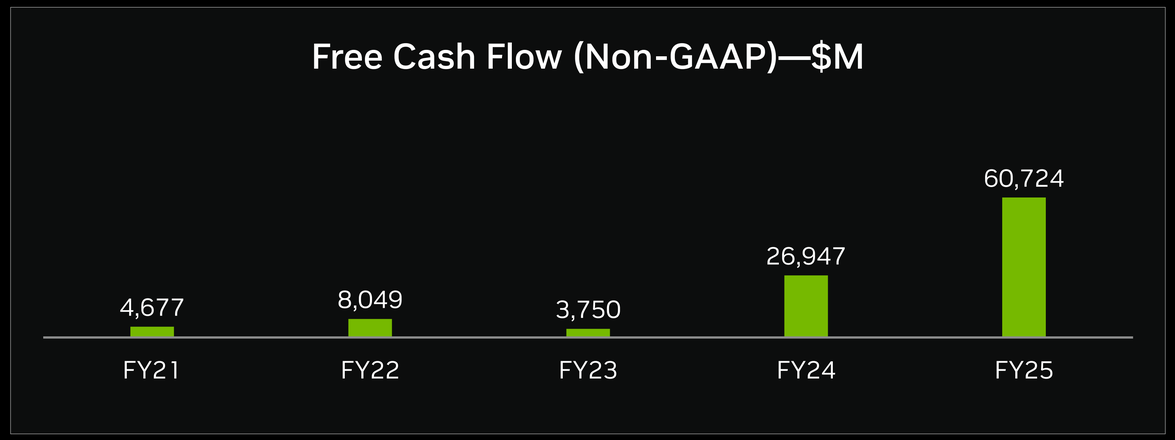
A new type of data center, AI factories, is expanding the data center footprint to $2T and beyond in the coming years. Eventually, companies in every industry will operate AI factories as the digital twin of their workforce, manufacturing plants, and products. A new industrial revolution has begun.

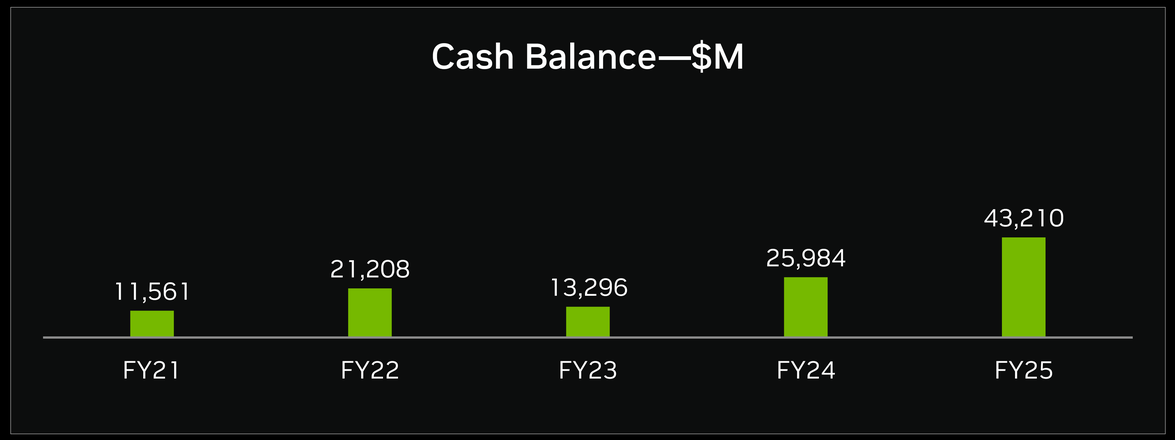
Financials

Annual Cash and Cash Flow Metrics









Cash balance is defined as cash and cash equivalents plus marketable securities | Refer to Appendix for reconciliation of non-GAAP measures

Corporate Sustainability

Environmentally Conscious

NVIDIA Blackwell platform delivers a 25X improvement in energy efficiency for LLM inference compared to the Hopper generation

We achieved our goal and will maintain 100% renewable electricity for offices and data centers under our operational control

On track to engage manufacturing suppliers comprising at least 67% of scope 3 category 1 GHG emissions with the goal of effecting supplier adoption of science-based targets by end of FY26

A Place for People to Do Their Life's Work

"Best Places to Work"
GLASSDOOR

"100 Most Sustainable U.S. Companies"
BARRON'S

"America's 100 Best Companies to Work For"
FORTUNE

"America's Most Responsible Companies"
NEWSWEEK

Management

Fast Company Magazine's World's Most Innovative Companies

Fortune's World's Most Admired Companies

Time Magazine's 100 Most Influential Companies

Wall Street Journal's Management Top 250

Corporate Governance

92% of directors are independent

Reconciliation of Non-GAAP to GAAP Financial Measures

Reconciliation of Non-GAAP to GAAP Financial Measures

Operating Income and Margin ($ in Millions and Margin Percentage)	Non-GAAP	Acquisition Termination Cost	Acquisition-Related and Other Costs (A)	Stock-Based Compensation (B)	Other (C)	GAAP
FY 2021	$6,803	—	(836)	(1,397)	(38)	$4,532
	40.8%	—	(5.0)	(8.4)	(0.2)	27.2%
FY 2022	$12,690	—	(636)	(2,004)	(9)	$10,041
	47.2%	—	(2.5)	(7.4)	—	37.3%
FY 2023	$9,040	(1,353)	(674)	(2,710)	(79)	$4,224
	33.5%	(5.0)	(2.5)	(10.0)	(0.3)	15.7%
FY 2024	$37,134	—	(583)	(3,549)	(30)	$32,972
	61.0%	—	(1.0)	(5.8)	(0.1)	54.1%
FY 2025	$86,789	—	(602)	(4,737)	3	$81,453
	66.5%	—	(0.5)	(3.6)	—	62.4%
Q1’25	$18,059	—	(140)	(1,011)	1	$16,909
	69.3%	—	(0.5)	(3.9)	—	64.9%
Q1’26	$23,275	—	(160)	(1,474)	(3)	$21,638
	52.8%	—	(0.4)	(3.3)	—	49.1%

A. Consists of amortization of acquisition-related intangible assets, inventory step-up, transaction costs, compensation charges, and other costs
B. Stock-based compensation charge was allocated to cost of goods sold, research and development expense, and sales, general and administrative expense
C. Comprises of legal settlement cost, contributions, restructuring costs and assets held for sale related adjustments

Reconciliation of Non-GAAP to GAAP Financial Measures (contd.)

($ in Millions)	Free Cash Flow	Purchases Related to Property and Equipment and Intangible Assets	Principal Payments on Property and Equipment and Intangible Assets	Net Cash Provided by Operating Activities
FY 2021	$4,677	1,128	17	$5,822
FY 2022	$8,049	976	83	$9,108
FY 2023	$3,750	1,833	58	$5,641
FY 2024	$26,947	1,069	74	$28,090
FY 2025	$60,724	3,236	129	$64,089
YTD Q1’25	$14,936	369	40	$15,345
YTD Q1’26	$26,135	1,227	52	$27,414


NVIDIA