

Except for the historical information contained herein, certain matters in this presentation; our markets and market opportunity; demand and supply; demand and growth drivers; the benefits, impact, performance, features and liquid cooling, and growth drivers; the benefits, impact, performance, features and liquid cooling, NVLink 8-GPU to NVLink 72-GPU, InfiniBand and Ethernet, CSP, regional AI factories, and enterprise platforms; AI training continuing scaling law; the \$1T installed base of general-purpose CPU data center infrastructure being modernized to a new GPU accelerated computing paradigm; Al Factories expanding the data center footprint to \$2T and beyond in the coming years; companies in every industry operating Al factories as the digital twin of their workforce, manufacturing plants, and products; every company producing digital intelligence; tokens being the transformed into intelligence agents, and chip designers, manufacturing robots and autonomous cars, and weather prediction agents to warn us of storms; some companies building and operating Al factories, while others renting; countries awakening to the need to treat its data as a national resource and Al factories as an essential national infrastructure; data being able to be transformed into the sovereign Al for its companies, startups, universities and governments; NVIDIA being able to address many major workloads across a wide range of industries; accelerated computing enabling full-stack optimization from algorithm to GPU architecture; Al scaling laws driving exponential demand for compute; NVIDIA being the leading inference platform; inference compute scaling exponentially; model size, multi-modality, synthetic data generation and reinforcement learning scaling computing capacity by ~4x per year; NVIDIA NVLink enabling new level of AI training & inference scaling; NVIDIA being able to address every industry from healthcare, manufacturing, and transportation to retail, CSPs, and telecommunications; NVIDIA AI platform and ecosystem reaching every market; AI driving investment cycle and significant returns; AI accelerating scientific discovery, creativity, and productivity across industries; Al agents being able to generate documents, review contracts, help code and debug software, operationalize marketing campaigns, and review medical images; social media, search and e-commerce apps using deep recommenders; creators being able to generate stunning, photorealistic images or 3D objects with a single text prompt; call center agents augmented with AI chatbots being able to scale to deliver customer care during peak times, run in-the-loop with customer support to provide better service, and capture every interaction to build institutional knowledge in high turnover industries; drug discovery seeing order-of-magnitude workflow acceleration using generative for candidate drug generation and target selection; manufacturing using generative Al to genera self-learning robots that can handle diverse tasks in changing work environments; climate tech scientists using Al to discover new materials for better batteries, emulate complex physics to predict long-range weather, estimate carbon-capture sites, optimize wind farm output; companies leveraging AI to unlock new monetizable applications, enhance employee productivity, and transform their business models; industrial AI applications serving large markets - including 10M factories, 200K warehouses, billions of humanoid robots, and 100M autos manufactured each year; specialized AI co-pilots collaborating with employees, being given access to necessary data, using platform tools like SAP, Snowflake, or ServiceNow, and collaborating with other AI agents in its workforce; NVIDIA AI Enterprise enabling IT ecosystem with state-of-the-art AI models and libraries to build agentic AI; physical AI embodying robotic systems; nations are awakening to the imperative to produce artificial intelligence using their own infrastructure, data, workforces and business networks; and nations building domestic computing capacity through various models are forward-looking statements.

These forward-looking statements and any other forward-looking statements that go beyond historical facts that are made in this presentation are subject to risks and uncertainties that may cause actual results to differ materially. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences and demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems and other factors. NVIDIA has based these forward-looking statements largely on its current expectations and projections about future events and trends that it believes may affect its financial condition, results of operations, business strategy, short-term and long-term business operations and objectives, and financial needs. These forward-looking statements are subject to a number of risks and uncertainties, and you should not rely upon the forward-looking statements as predictions of future events. The future events and trends discussed in this presentation may not occur and actual results could differ materially and adversely from those anticipated or implied in the forward-looking statements. Although NVIDIA believes that the expectations reflected in the forward-looking statements are reasonable, the company cannot guarantee that future results, levels of activity, performance, achievements or events and circumstances reflected in the forward-looking statements to reflect future events or circumstances. For a complete discussion of factors that could materially affect our financial results and operations, please refer to the reports we file from ti

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 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".





World's First GPU Turns 25

NVIDIA released GeForce 256, the world's first GPU, 25 years ago. This marked a milestone in computing by providing gamers with higher frame rates, smoother gameplay, and life-like graphics.

NVIDIA's GPU set the foundation for the big bang of modern AI, a decade later. The parallel processing power of NVIDIA GPUs and the invention of a revolutionary programming model "CUDA" attracted researchers who recognized the opportunity to accelerate computationally-demanding applications like physical and chemical simulations and deep learning.

NVIDIA released the GeForce RTX series in 2018, reinventing computer graphics once again with real-time ray tracing and neural rendering, applying AI to generate pixels at quality and speed no one thought possible.

Today, the world recognizes GPUs not only for reshaping gaming but also for powering the AI industrial revolution. NVIDIA GeForce 256 started an era where gaming, computing, and AI evolve together, fundamentally advancing entertainment, technology, and society.



NVIDIA Scaling with AI Scale-Up and Scale-Out

NVIDIA's chips and systems supply chain expanded significantly in 2024, while demand continues to exceed supply.

NVIDIA Blackwell platform in full production – expanding to support x86 and NVIDIA Grace, air and liquid cooling, NVLink 8-GPU to NVLink 72-GPU, InfiniBand and Ethernet, CSP, regional AI factories, and enterprise platforms.

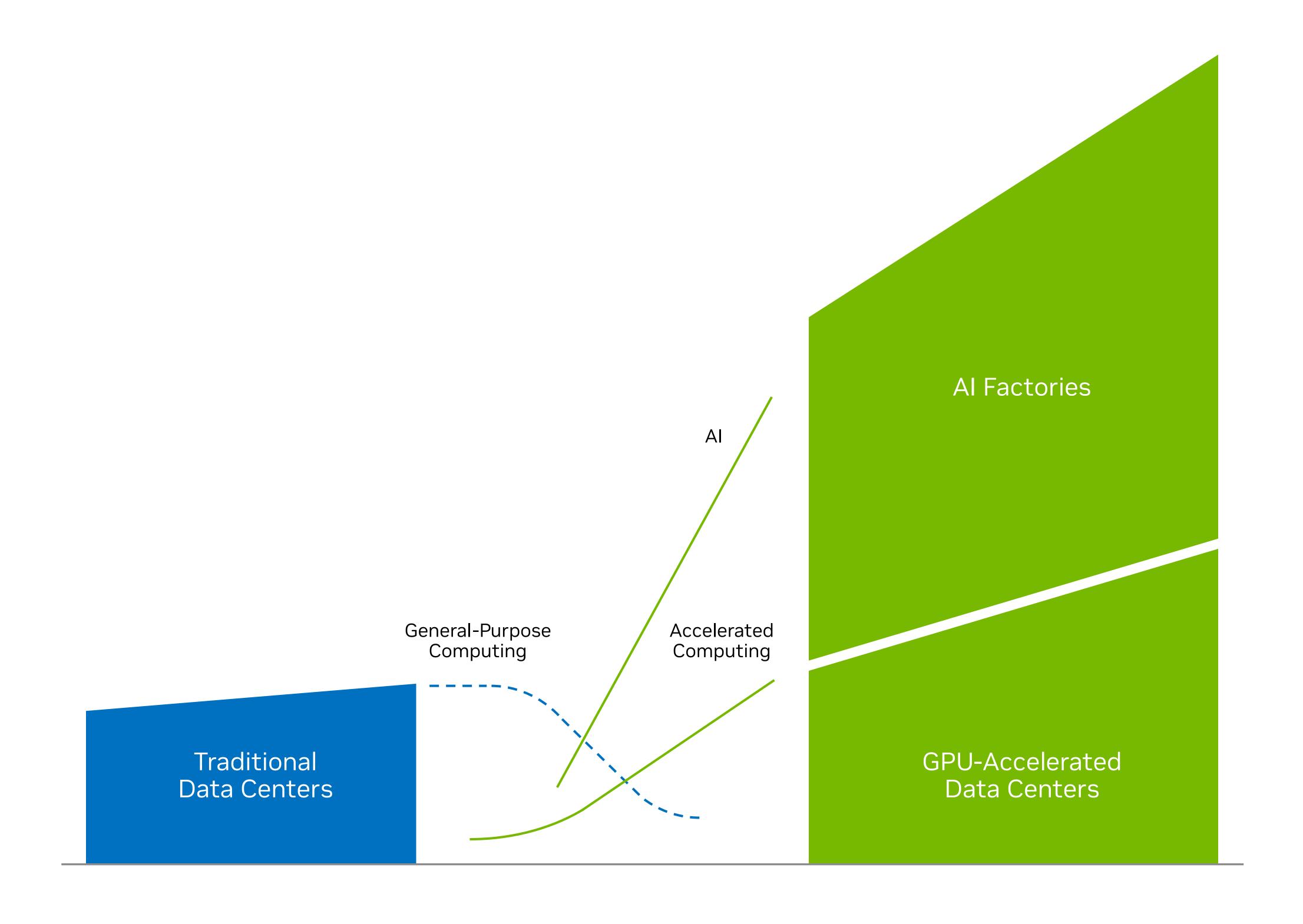
One Year Rhythm at full infrastructure scale.

Al training continues Al scaling law. *OpenAl o 1* starts inference-time scaling law. Inference is driving significant workload growth and demand.

Regional and sovereign clouds start up around the world.

NVIDIA AI Enterprise and Omniverse set the foundations for enterprise and industrial AI.





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 Al. 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.



Al Factories—A New Class of Data Centers

Production of digital intelligence tokens

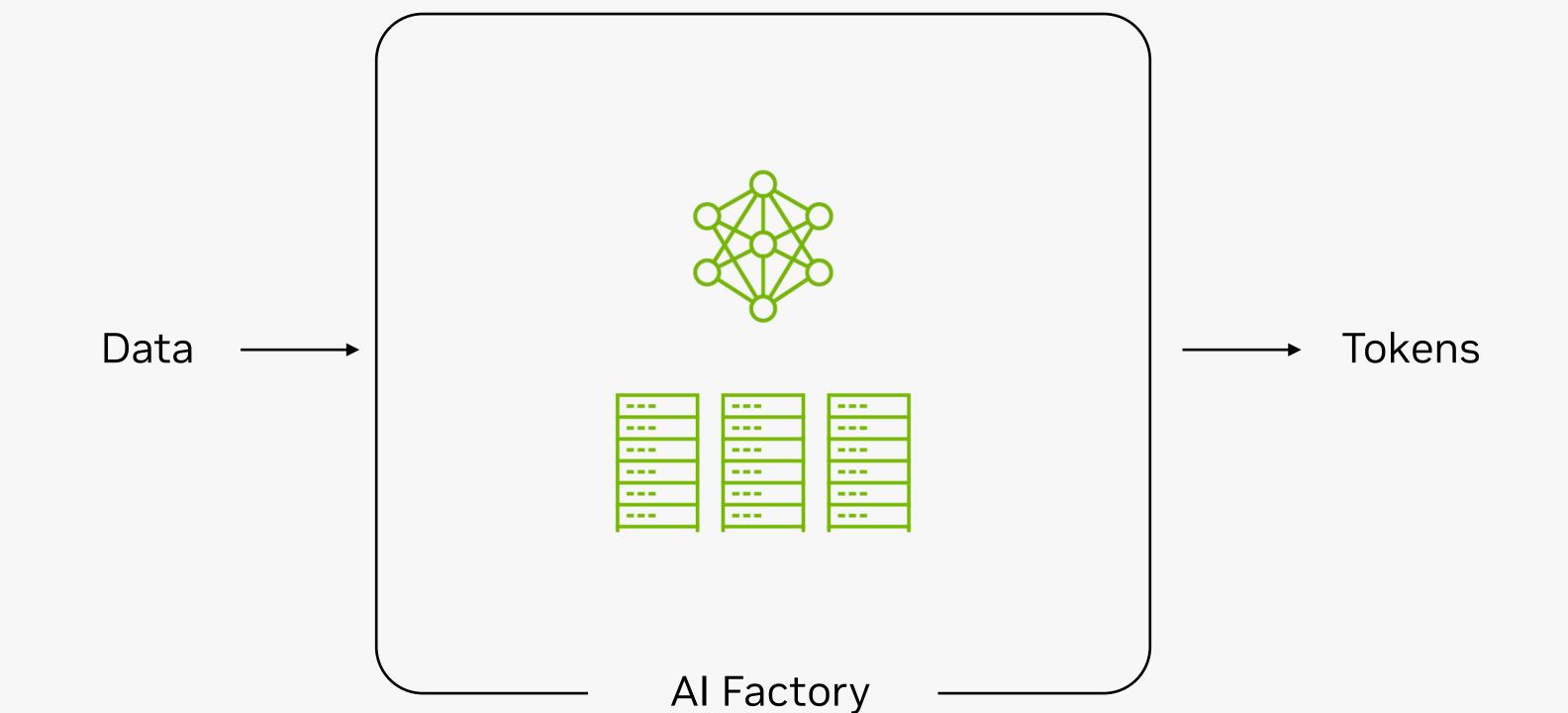
Al factories are a new form of computing infrastructure. Its purpose is not to store user and company data or run ERP and CRM applications. Al factories are highly optimized systems purposebuilt to process raw data, refine it into models, and produce monetizable tokens with great scale and efficiency.

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.

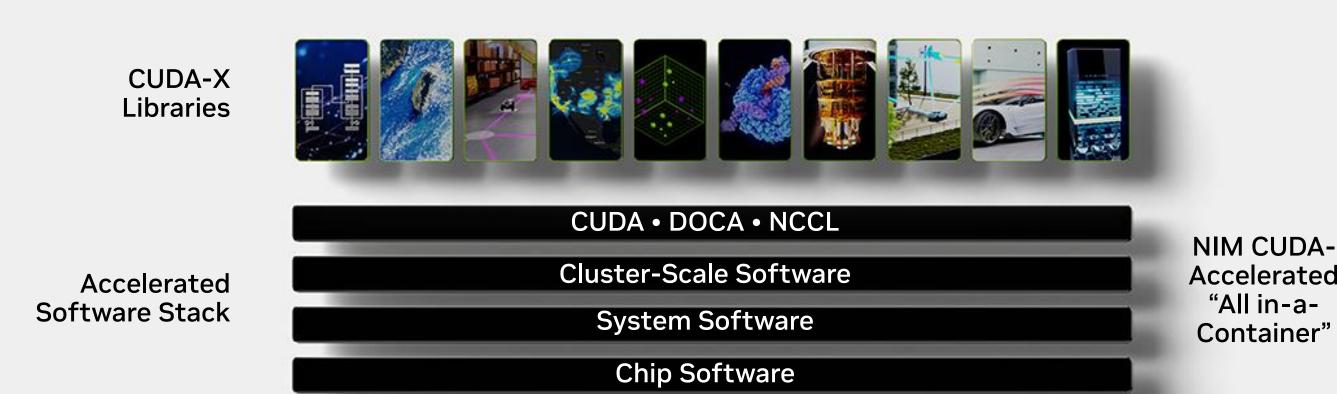
Every company will produce digital intelligence. Tokens will be transformed into intelligent responses and actions of digital nurses, tutors, customer service agents, and chip designers, manufacturing robots and autonomous cars, and weather prediction agents to warn us of storms. Some companies will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

Countries are awakening to the need to treat its 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.





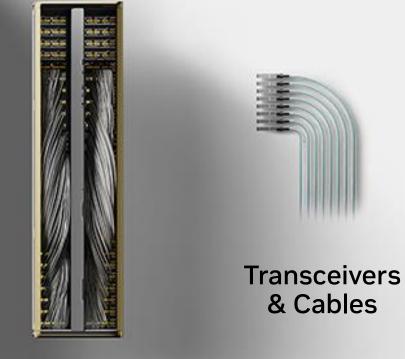




Omniverse **CUDA-Accelerated Physical AI Libraries**



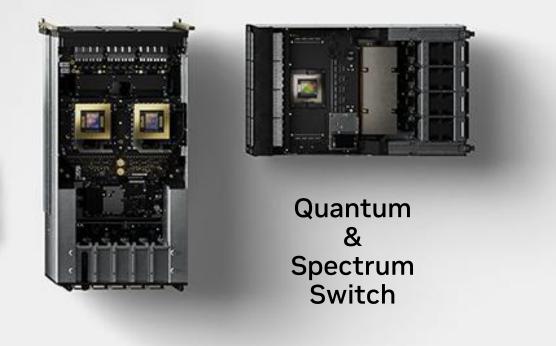






GB200 NVL72 SuperPOD











Chips Purpose-Built for Al Supercomputing GPU | CPU | DPU | NIC | NVLINK Switch | IB Switch | ENET Switch

NVIDIA's Accelerated **Computing Platform**

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

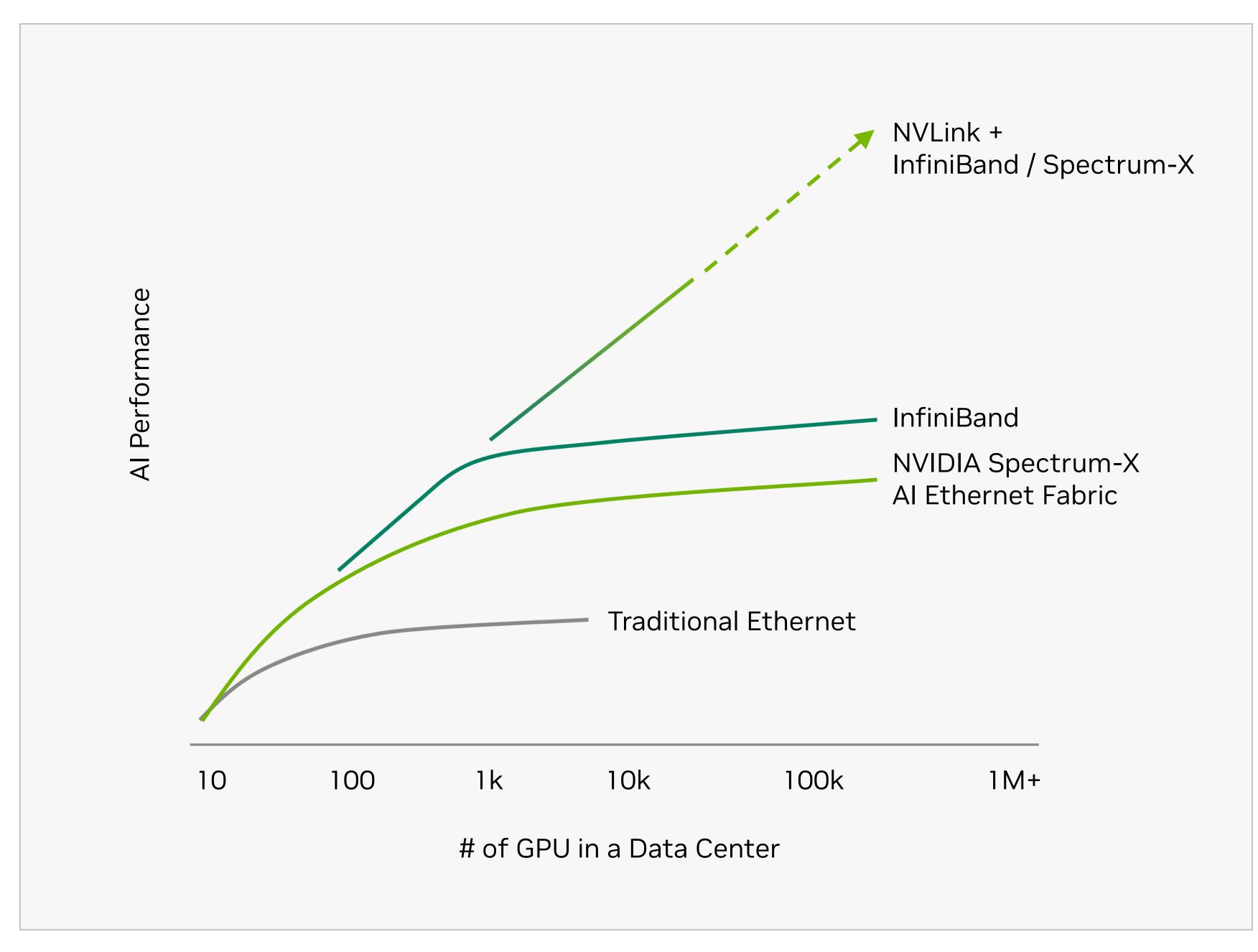
NVIDIA has accelerated software and compute by a 1,000,000X in the last decade, far surpassing Moore's law.

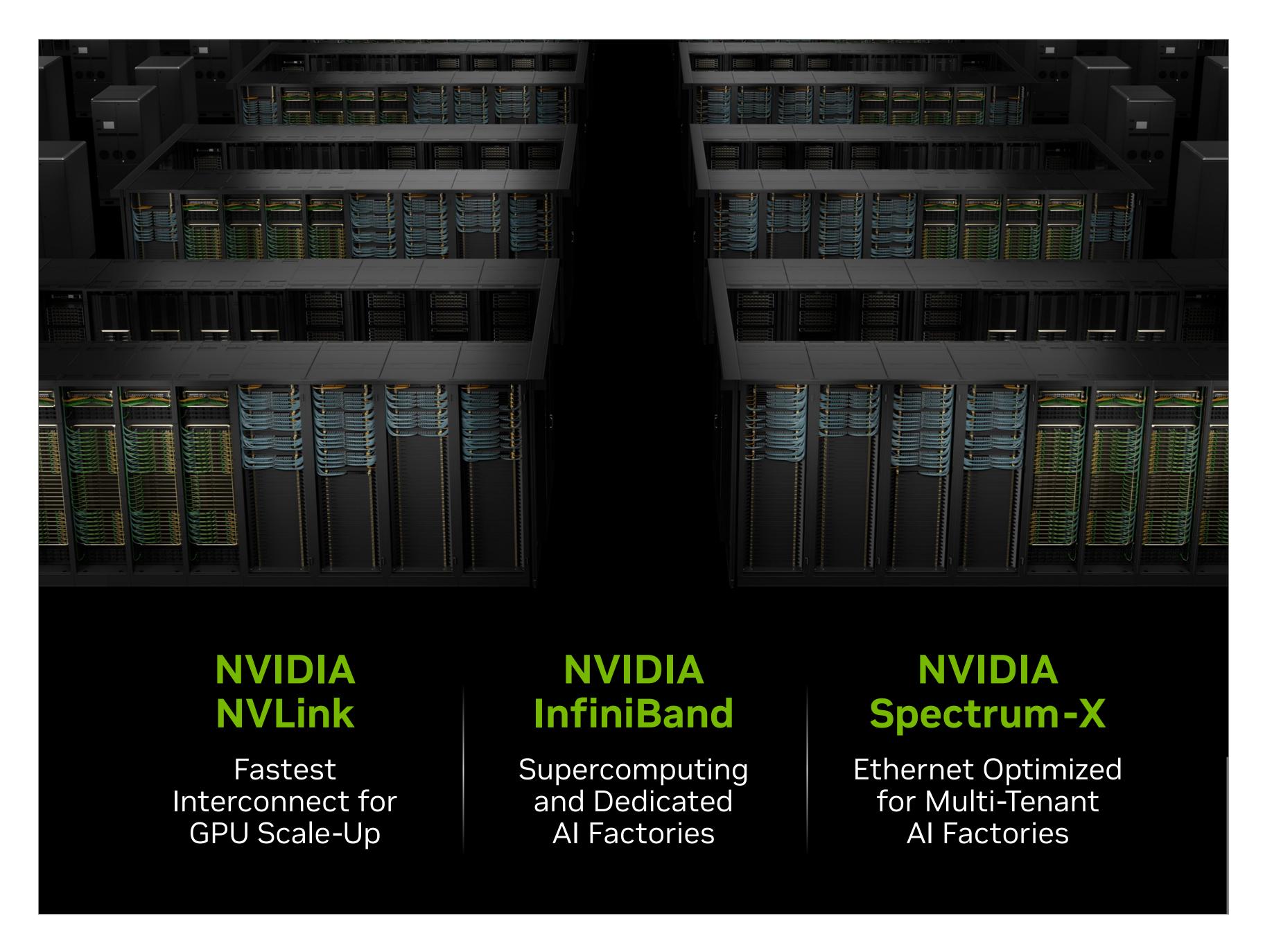
Accelerated computing requires full-stack innovation—optimizing across every layer of computing—from chips and systems to software and algorithms, demanding deep understanding of the problem domain.

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

Extending NVIDIA Networking to Scale-Up and Scale-Out AI in Any Datacenter

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

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 re-designed. 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 Al and high-performance computing significantly reducing run time, cost, energy, while increasing scale.

With over 400 CUDA libraires 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, cuClM, TensorRT, Triton Inference Server, DALI, nvlmageCodec, 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 Al

Riva, TensorRT, Triton Inference Server, NeMo, cuBLAS, cuDNN, cuFFT, CUTLASS

~100X

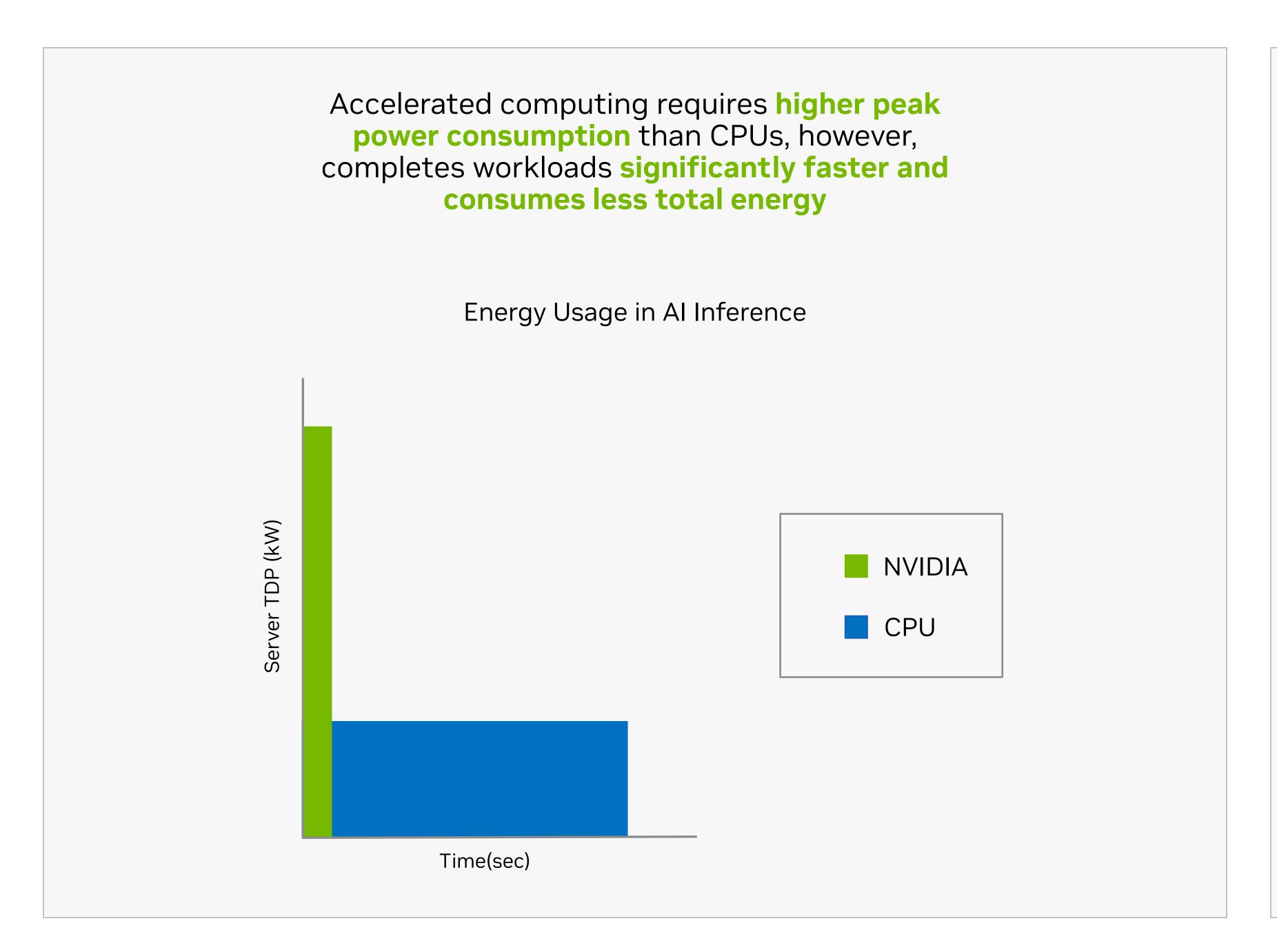
Agentic & Physical Al

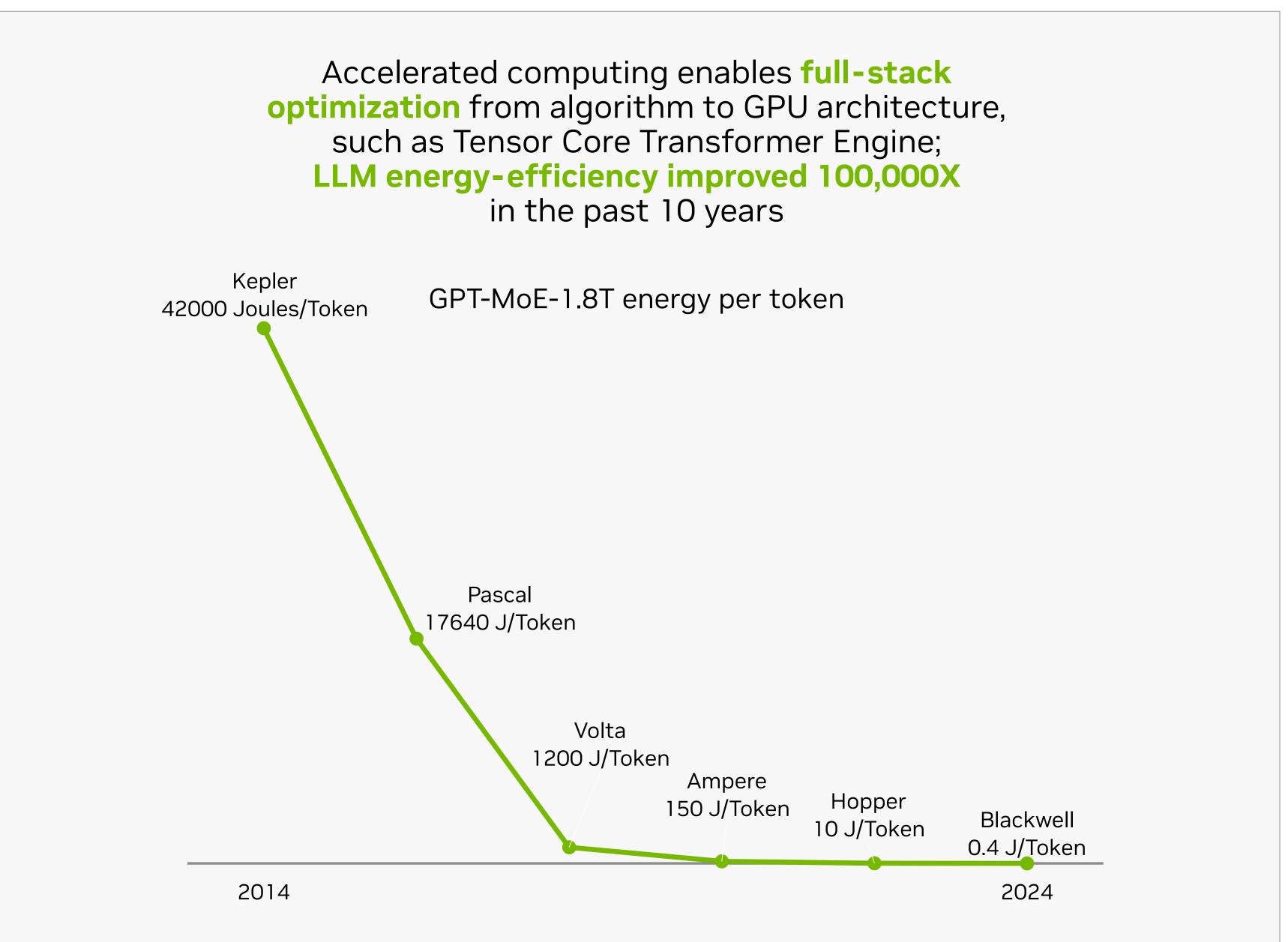
ACE, Riva, Nemo, Tokkio Digital Human, Holoscan, Metropolis, Omniverse, Isaac, DRIVE, cuLitho, cuMotion, cuOpt, Aerial CUDA-accelerated RAN, Sionna, fVDB, PhysX, Warp, NVblox



Accelerated Computing Is Sustainable Computing

Order-of-Magnitude More Energy-Efficient

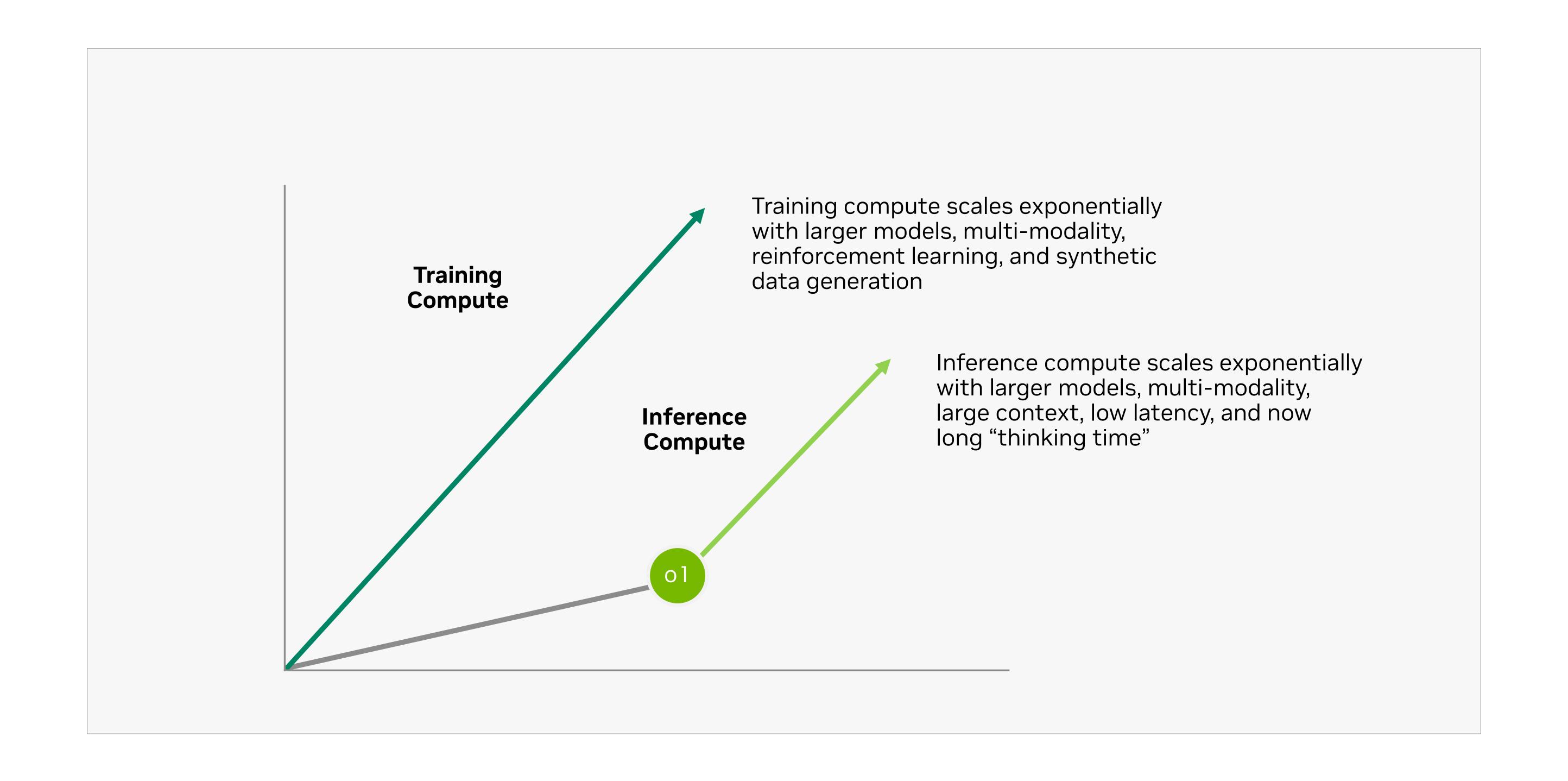






Al Scaling Laws Drive Exponential Demand for Compute

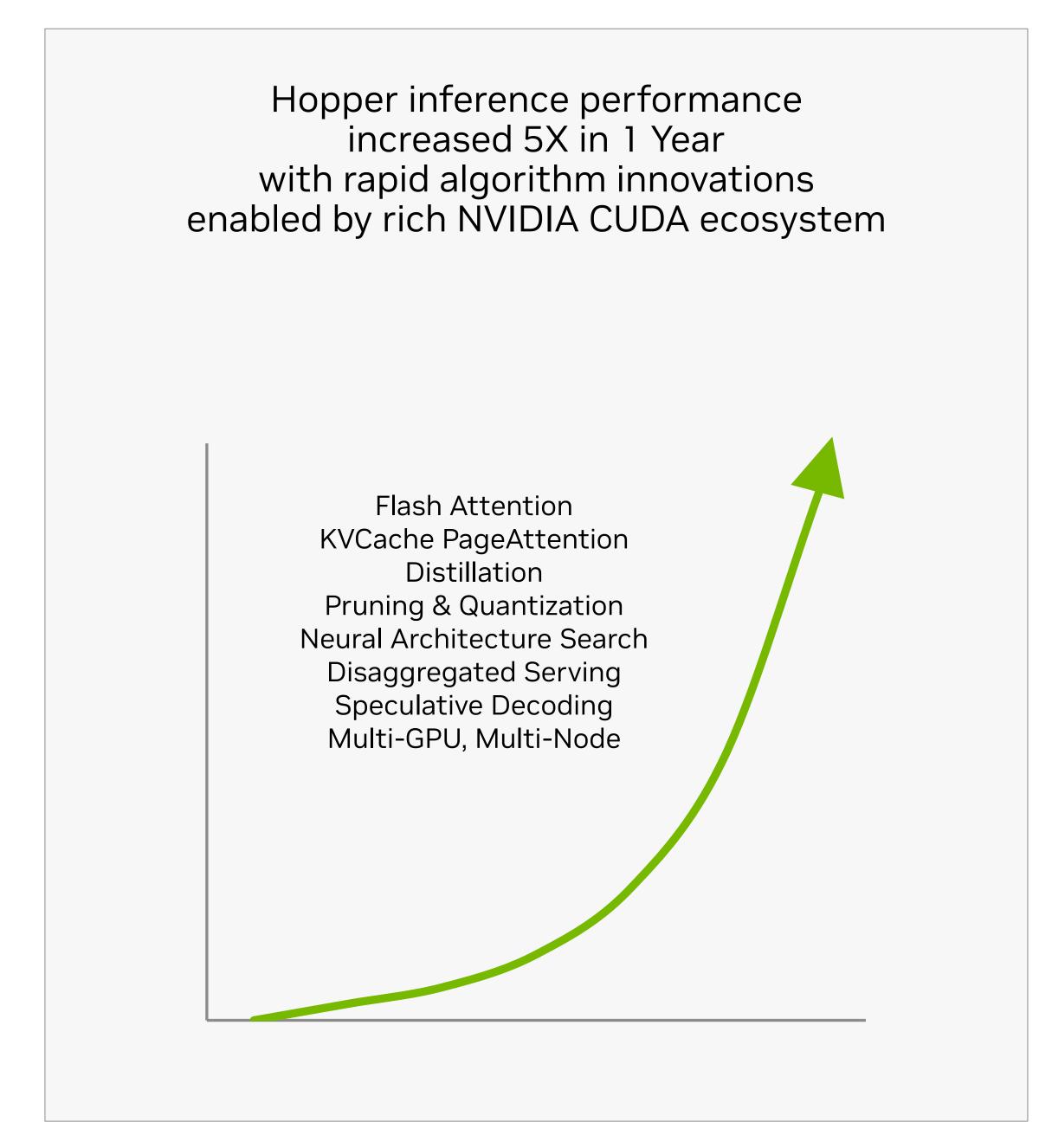
New OpenAl o 1 Long "Thinking Time" Creates a New Way to Scale

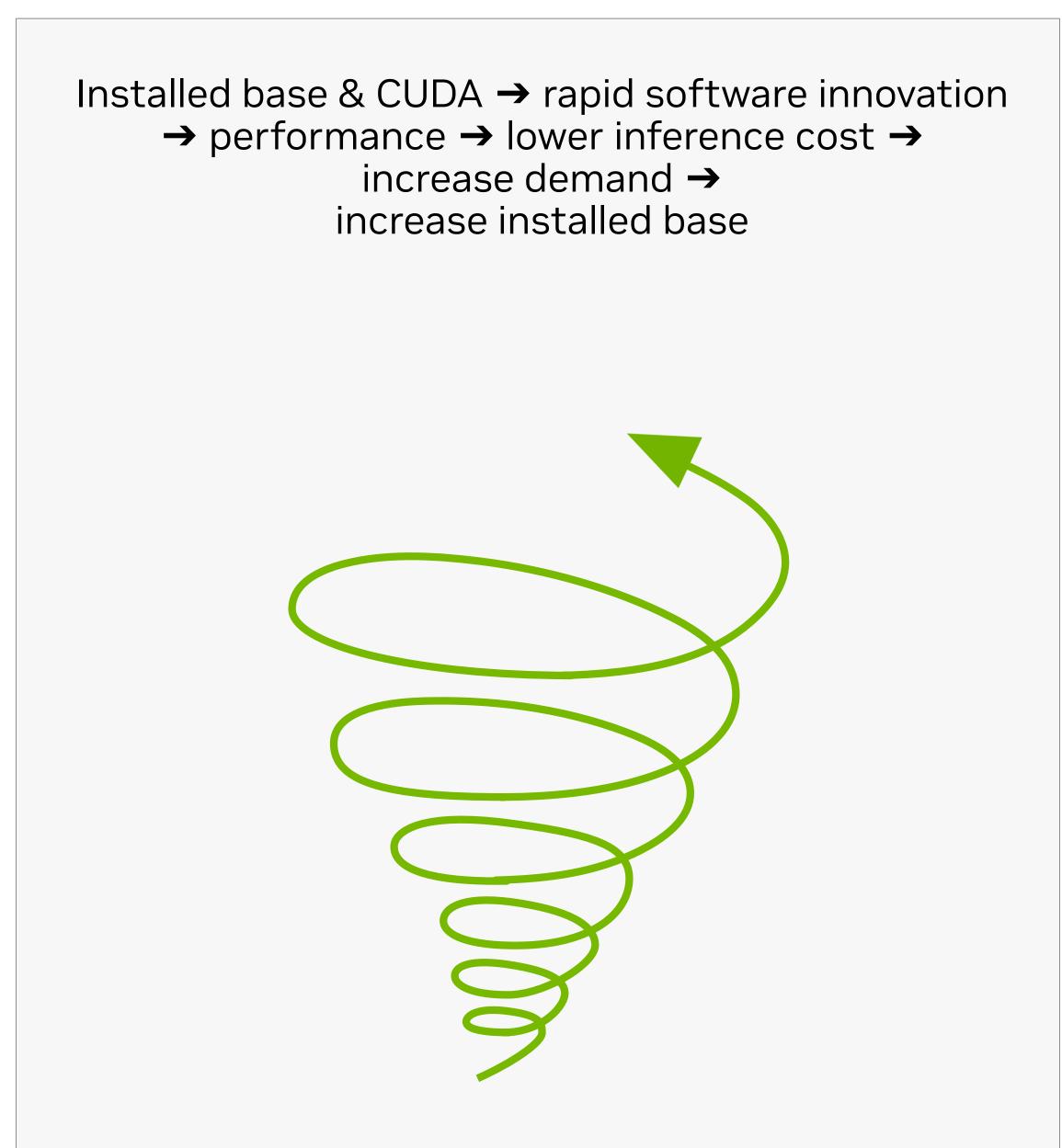


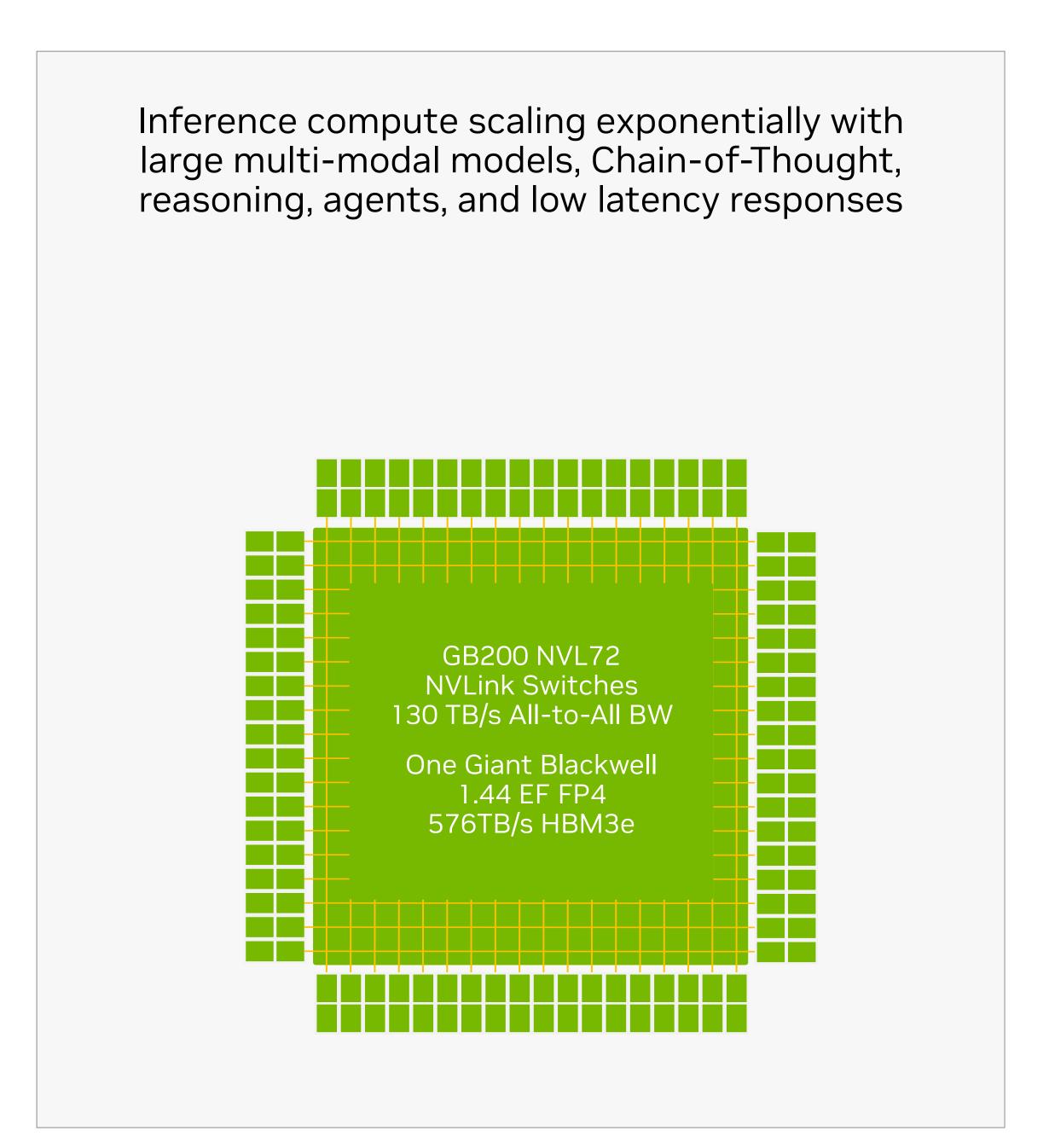


NVIDIA is the Leading Inference Platform

Inference Compute Scales Exponentially with "Long Thinking"









NVIDIA Al Infrastructure Roadmap at One Year Rhythm

NVIDIA's newest architecture – Blackwell – is an AI infrastructure platform and integrates seven chips, each contributing to performance at data center-scale.

Compute demand for AI is scaling exponentially. Model size, multi-modality, synthetic data generation and reinforcement learning is scaling computing capacity by ~4x per year. And new inference-time technologies like multi-modality, Chain-of-Thought, reasoning, and agents have introduced an exponential scaling law to inference.

NVIDIA has the scale and ability to update the entire AI infrastructure on a One Year Rhythm as we can engineer and optimize across the full stack and the entire infrastructure.

The cadence and breadth of our innovation brings more performant AI infrastructure to the market each year, delivering exceptional **TCO**, energy efficiency, and ROI to our customers.



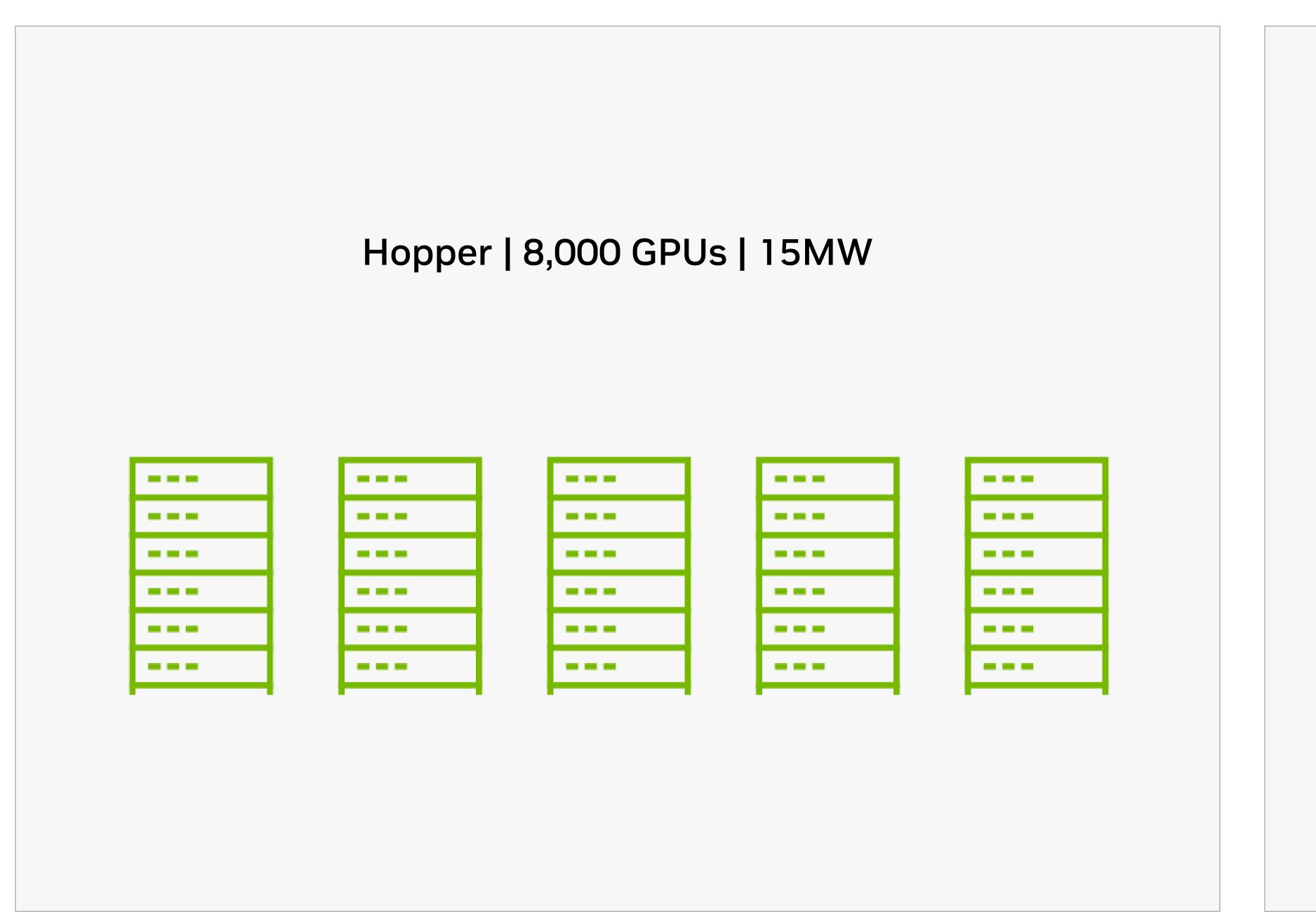
One Year Rhythm | Supercluster Scale | Full-Stack | CUDA Everywhere

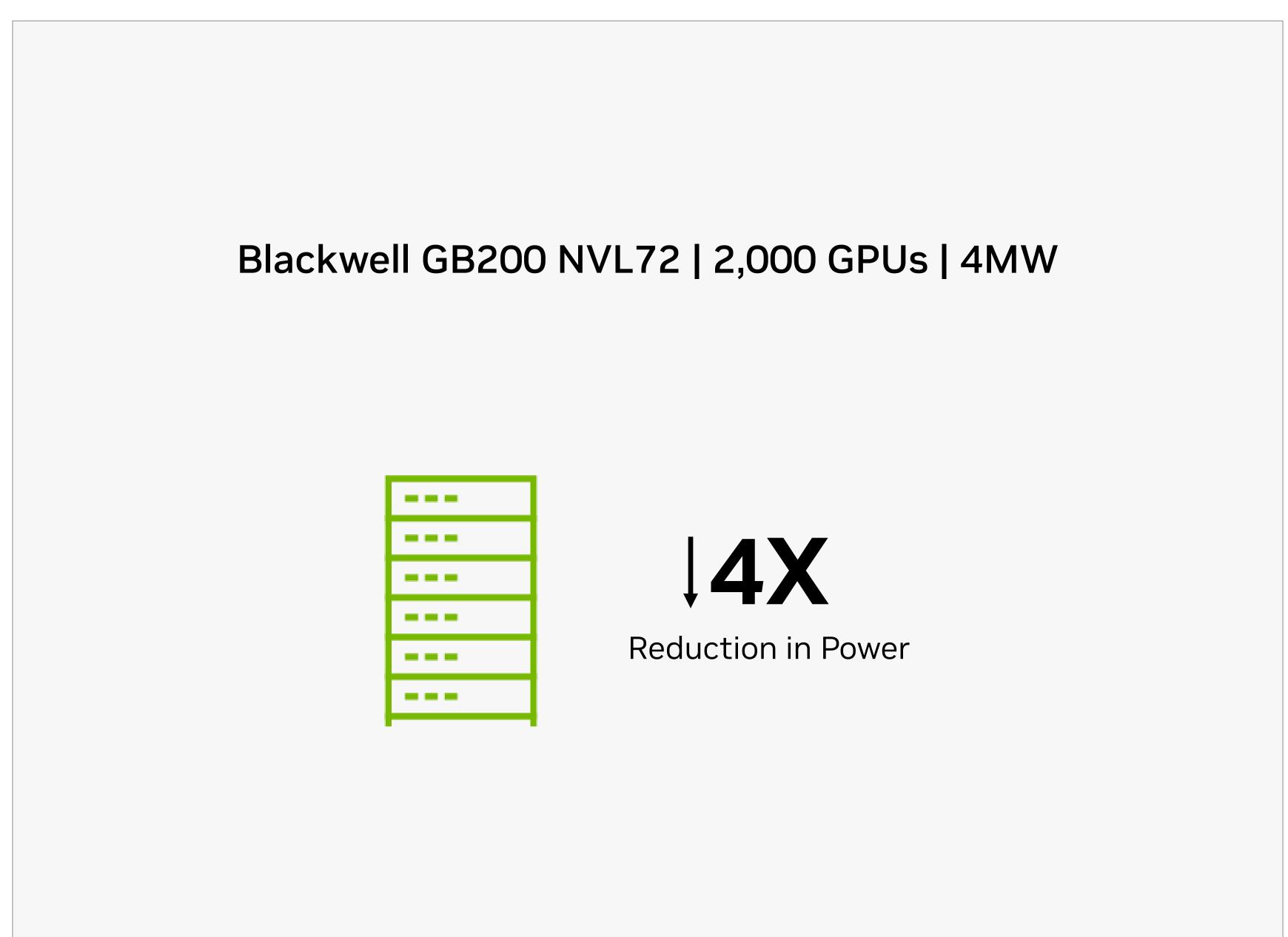
"Supercharge Al Scaling Law"



One Year Rhythm Drives Annual Cost and Energy Reduction

Significant reduction in TCO with each generation







NVIDIA NVLink Enables New Level of AI Training & Inference Scaling



NVIDIA Addresses The Entire Al Market

NVIDIA's accelerated computing platform encompasses a complete computing stack and infrastructure required for customers to build and run AI at scale.

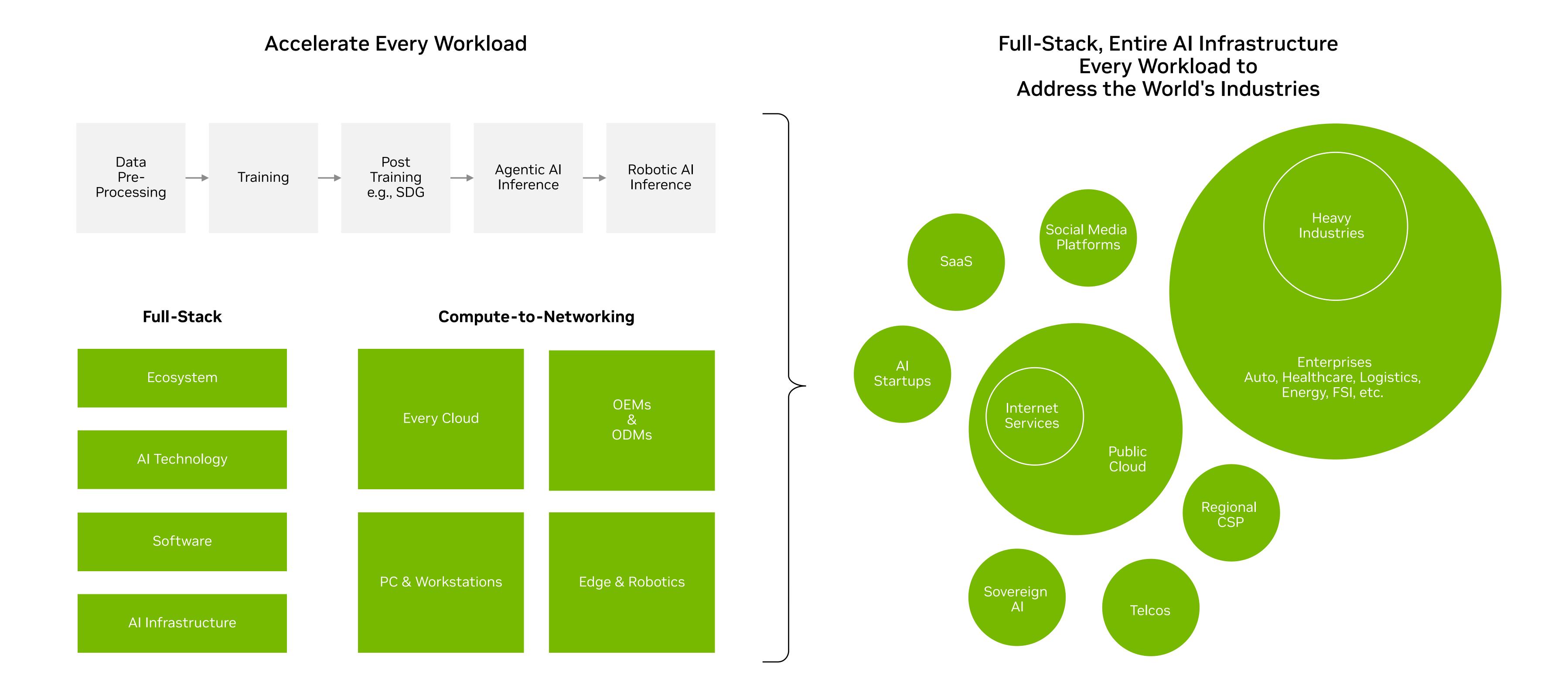
CUDA-X libraries and microservices leverage the programmability of CUDA to enable thousands of applications across a wide range of use cases to run on NVIDIA infrastructure, including Al training and inference, data processing, robotics, drug discovery, and consumer internet services.

NVIDIA has a rich ecosystem of partners who have integrated NVIDIA platform technologies and joined our go-to-market. Each industry and market segment may include different domain-specific software stacks, ODMs, OEMs, CSPs, 3rd party software platforms, system integrators, and system integrations/solution development partners.

The completeness of the NVIDIA platform for each industry and workload lets us address every industry from healthcare, manufacturing, and transportation to retail, CSPs, and telecommunications.



NVIDIA AI Platform and Ecosystem Reaches Every Market





Al Driving Investment Cycle and Significant Returns

Al is accelerating scientific discovery, creativity, and productivity across industries

Al Agents, or copilots and Al teammates, automate tasks at superhuman speed, increase employee productivity and accelerate businesses. Agents can generate documents, review contracts, help code and debug software, operationalize marketing campaigns, and review medical images

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

Creators can generate stunning, photorealistic images or 3D objects 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 scale to deliver customer care during peak times, run in-the-loop with customer support to provide better service, and capture every interaction to build institutional knowledge in high turnover industries

Drug discovery is seeing order-of-magnitude workflow acceleration using generative AI for candidate drug generation and target selection

Manufacturing is using generative AI to generate ideas to inspire designers, accelerate the planning and layout of factories and warehouses; and for self-learning robots that can handle diverse tasks in changing work environments

Climate Tech scientists are using AI to discover new materials for better batteries, emulate complex physics to predict long-range weather, estimate carbon-capture sites, optimize wind farm output



Al Agents & Copilots
Over 1B
knowledge workers



Search & Social Media \$700B in digital advertising annually

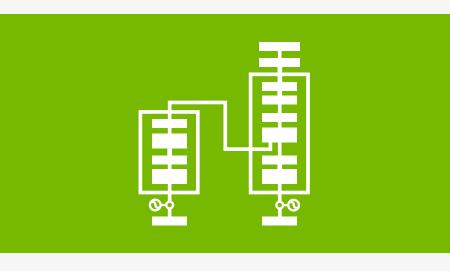


Al Content Creation
50M creators
globally



Legal Services,
Education

1M legal professionals in the US
9M educators in the US



Development
30M software developers
globally

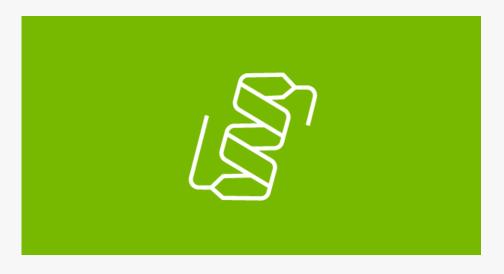
Al Software



Financial
Services
678B annual
credit card transactions



Customer Service
15M call center agents
globally



Drug Discovery
10¹⁸ molecules in chemical space
40 exabytes of genome data



Manufacturing \$50T of Heavy Industry



The Enterprise and Industrial AI Wave of Adoption Has Started

Al brings computing into the core of the \$100T global industries. Companies will leverage Al to unlock new monetizable applications, enhance employee productivity, and transform their business models.

NVIDIA AI Enterprise platform, which includes NVIDIA NeMo agent on-boarding and operating software platform, NVIDIA NIM "model-in-a-container" microservices, and AI Foundry custom-AI service, enables enterprises to build and run custom agentic AI applications.

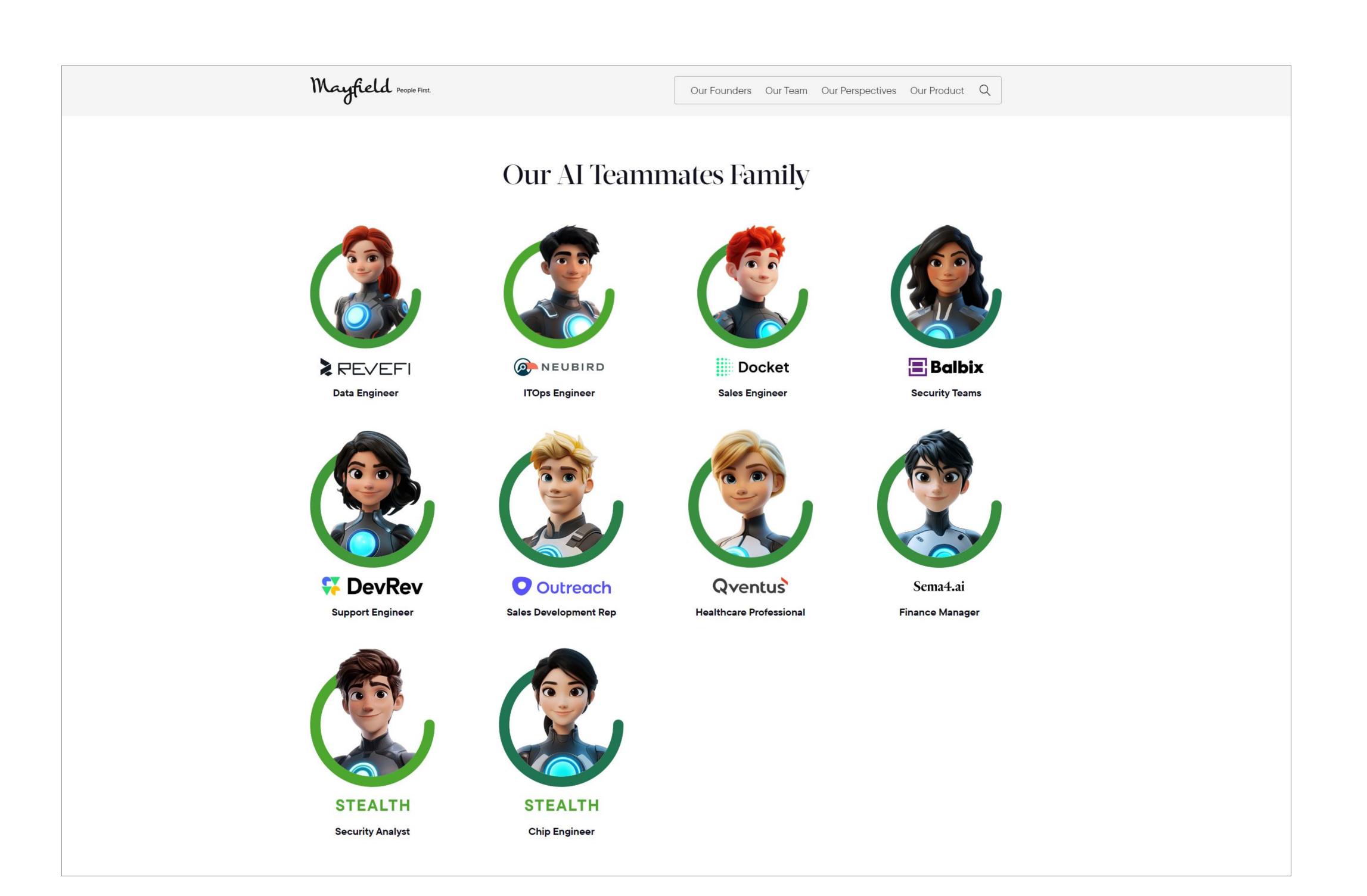
Industrial AI applications will serve large markets – including 10M factories, 200K warehouses, billions of humanoid robots, and 100M autos manufactured each year.

NVIDIA Omniverse provides a platform for building and deploying autonomous machine applications. Omniverse connects to leading design, simulation, and factory control software and enables virtual integration of factories and warehouses into operational digital twins.



Enterprise Al Wave is Here

Co-Pilots, Al Teammates, and Agents are the New Digital Workforce to Revolution Enterprise Productivity



This is Mayfield's webpage of AI companies in their portfolio. The illustration and description provide a framework of enterprise AI – agents are a new digital workforce that will collaborate and accelerate the work of employees.

Specialized Al co-pilots, teammates, or agents will collaborate with employees, be given access to necessary data, use platform tools like SAP, Snowflake, or ServiceNow, and even collaborate with other Al agents.

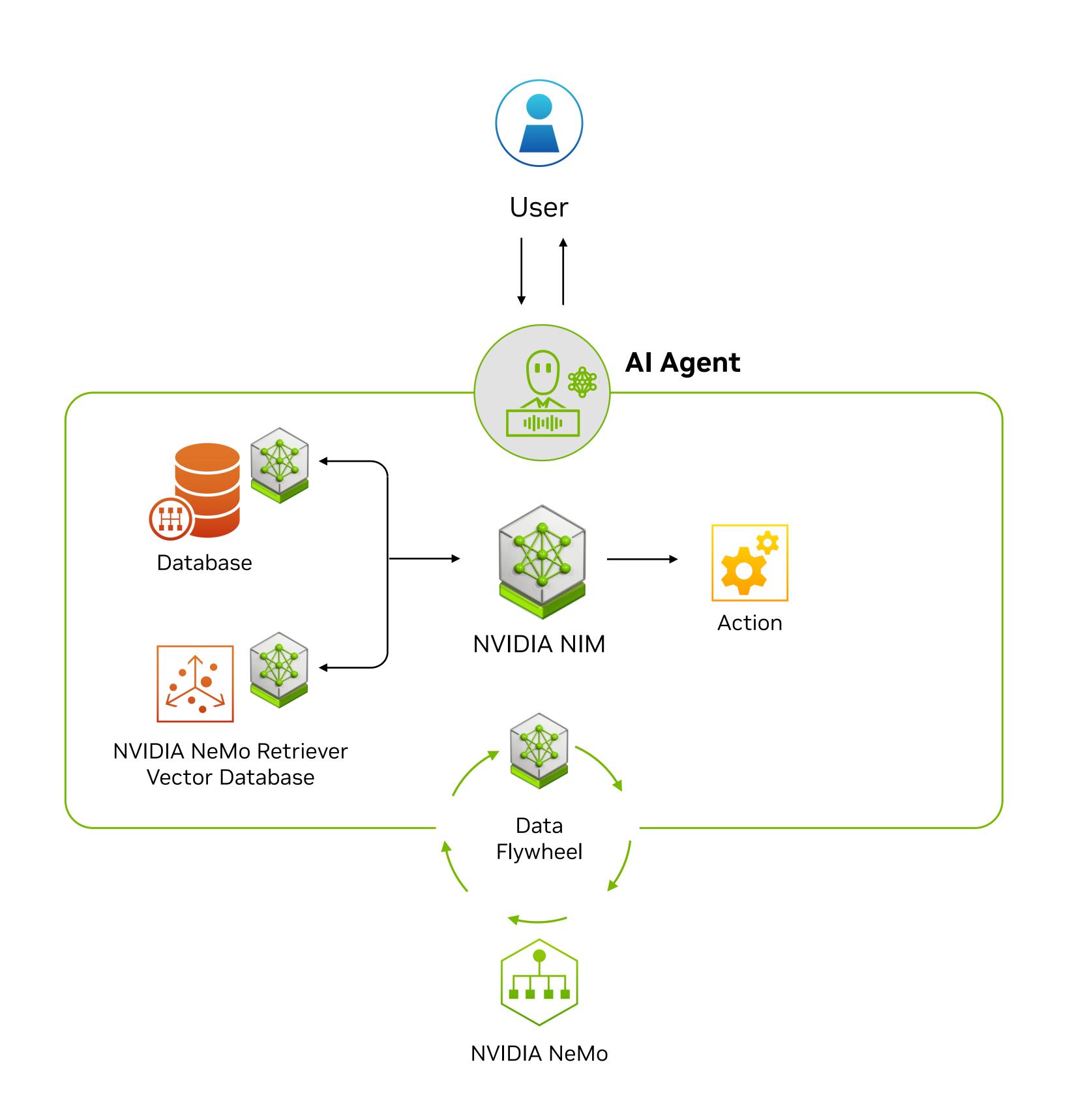
In time, the world's companies will hire billions of Al agents in its workforce.

NVIDIA AI Enterprise offers the IT and services ecosystem, NVIDIA-optimized NeMo agent lifecycle libraries, NIM pre-trained models, and DGX Cloud infrastructure to build and deploy enterprise AI.

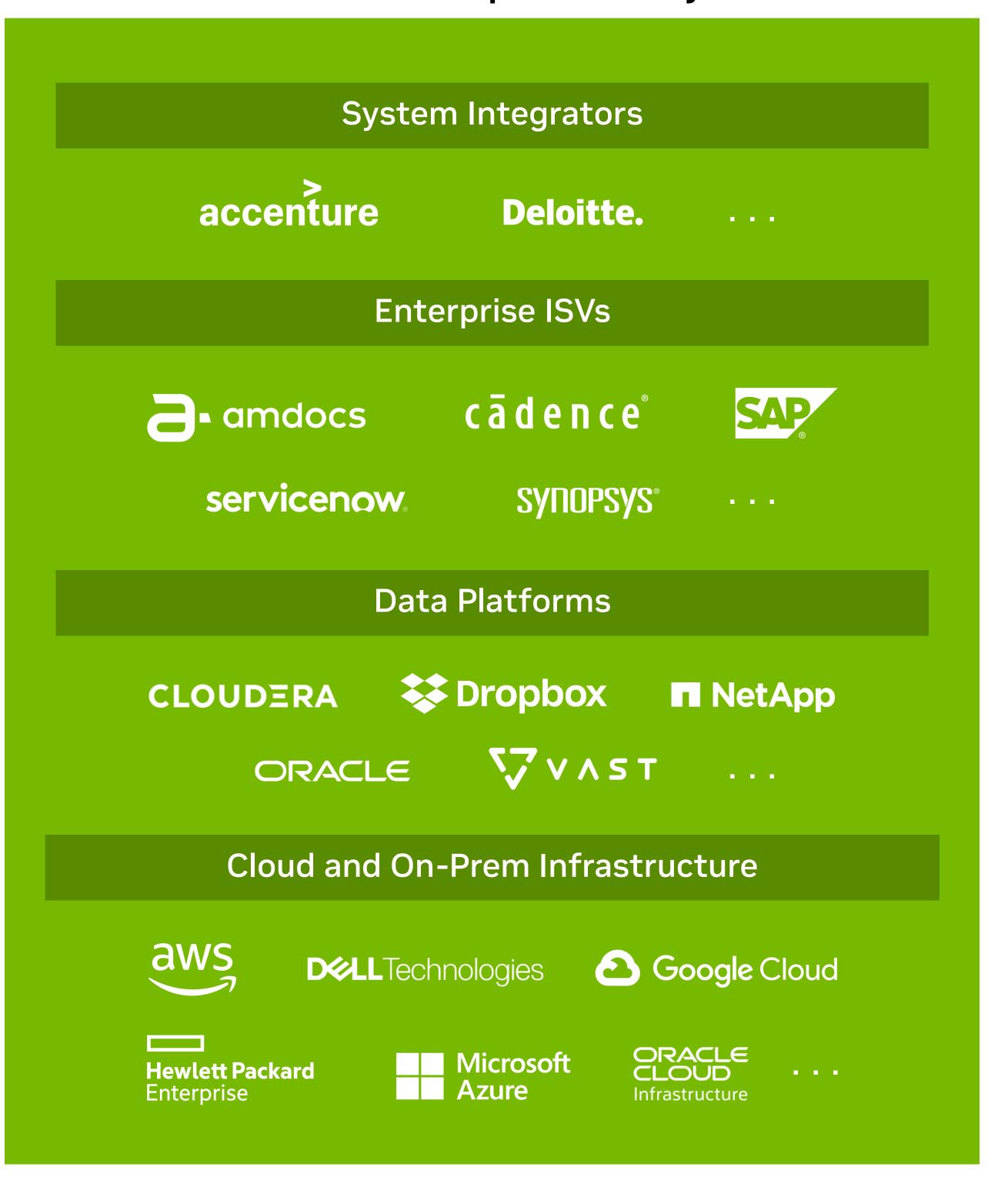
Enterprises license NVIDIA AI Enterprise at \$4500 per-GPU per-Year to run NIM and NeMo libraries.



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

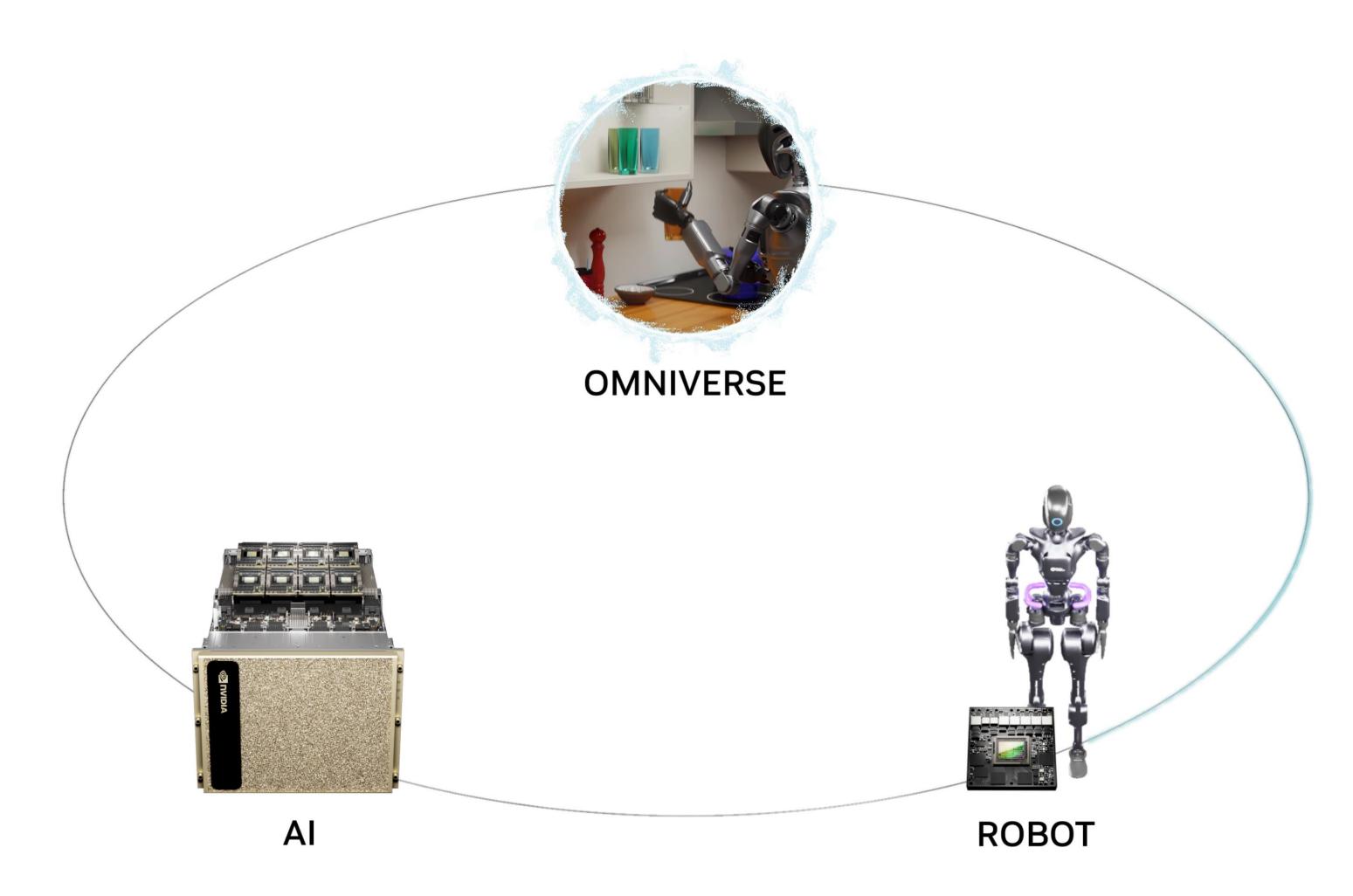


NVIDIA AI Enterprise Ecosystem





NVIDIA Omniverse and AI Revolutionizing Manufacturing & Robotics



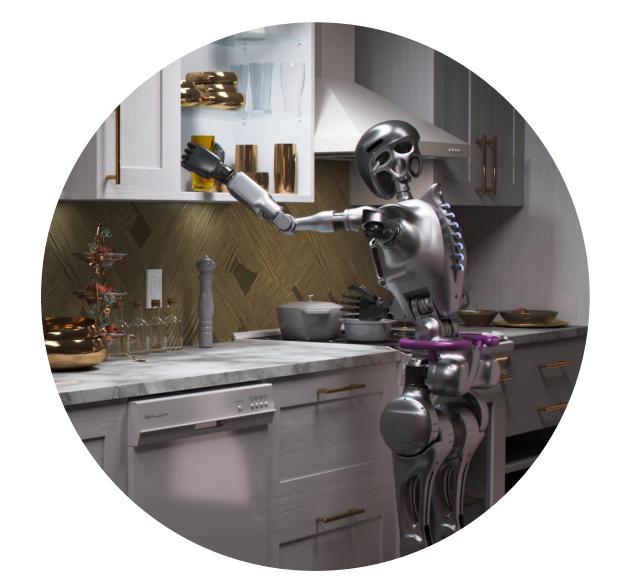
The next AI wave is physical AI - models that can perceive, understand, and interact with the physical world. Physical AI will embody robotic systems – from autonomous vehicles to industrial robots and humanoids, to warehouses and factories.

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 \$4500 per-GPU per-Year.



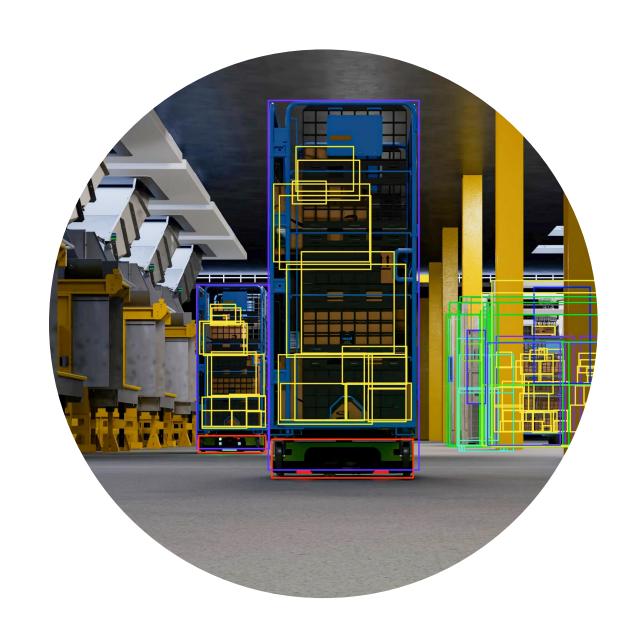
100M Cars



Billions in Future



10M Factories



200K Warehouses



Sovereign Al

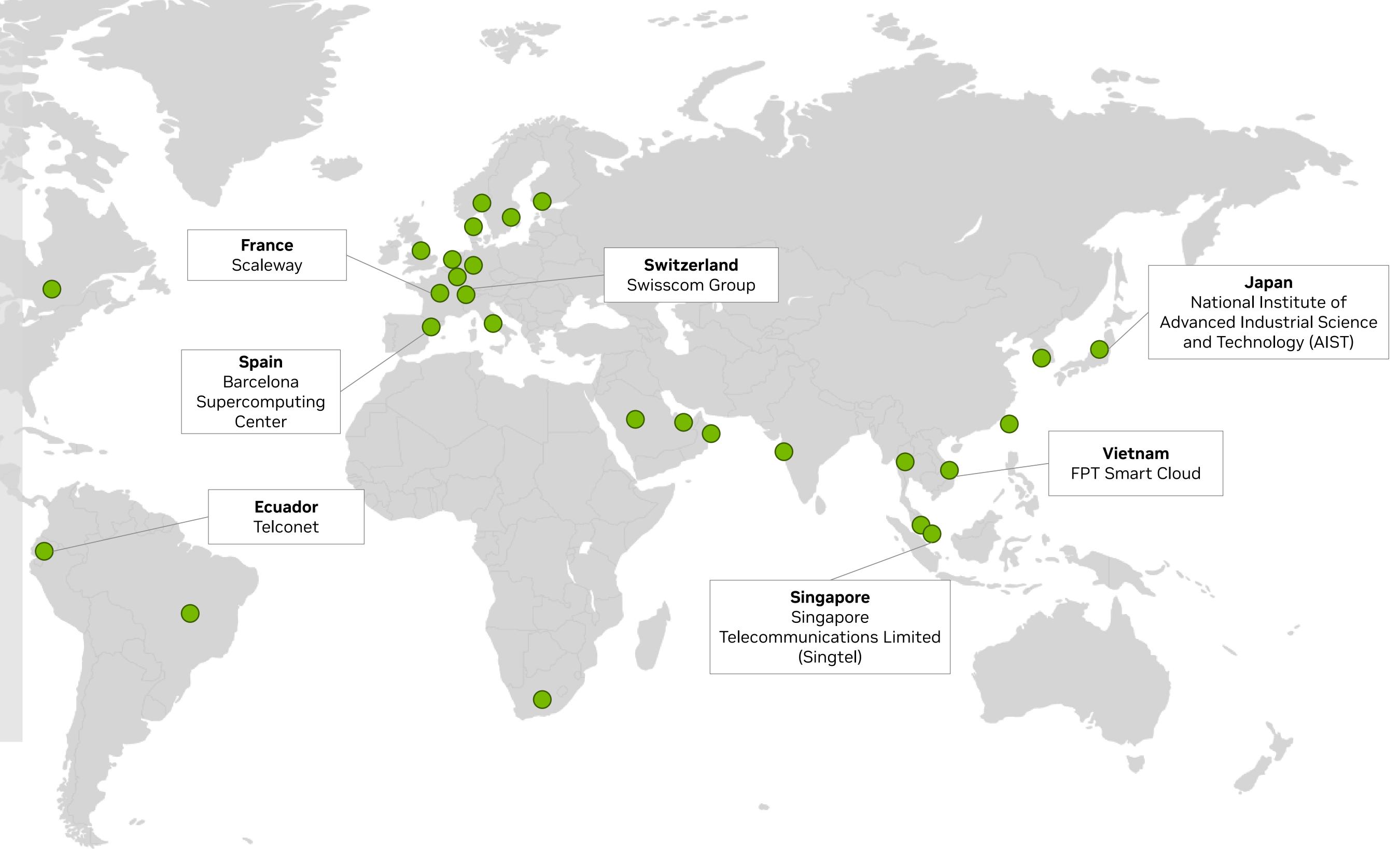
Nations produce Al using their own data, infrastructure, workforce, and business networks

Sovereign Al

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

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

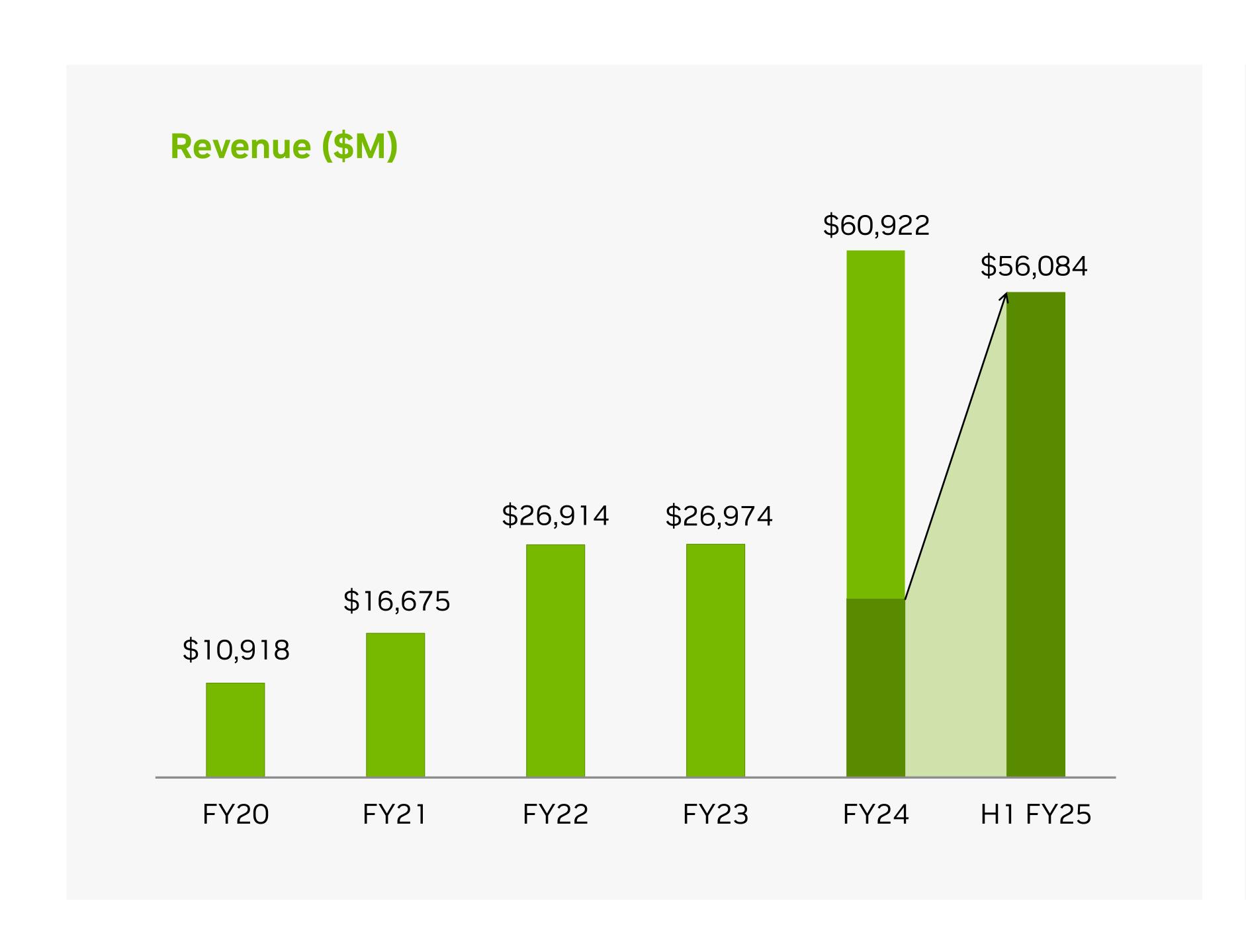
NVIDIA's ability to help build AI infrastructure with our end-to-end compute-to-networking technologies, full-stack software, AI expertise, and rich ecosystem of partners and customers allows sovereign AI and regional cloud providers to jumpstart their countries' AI ambitions.

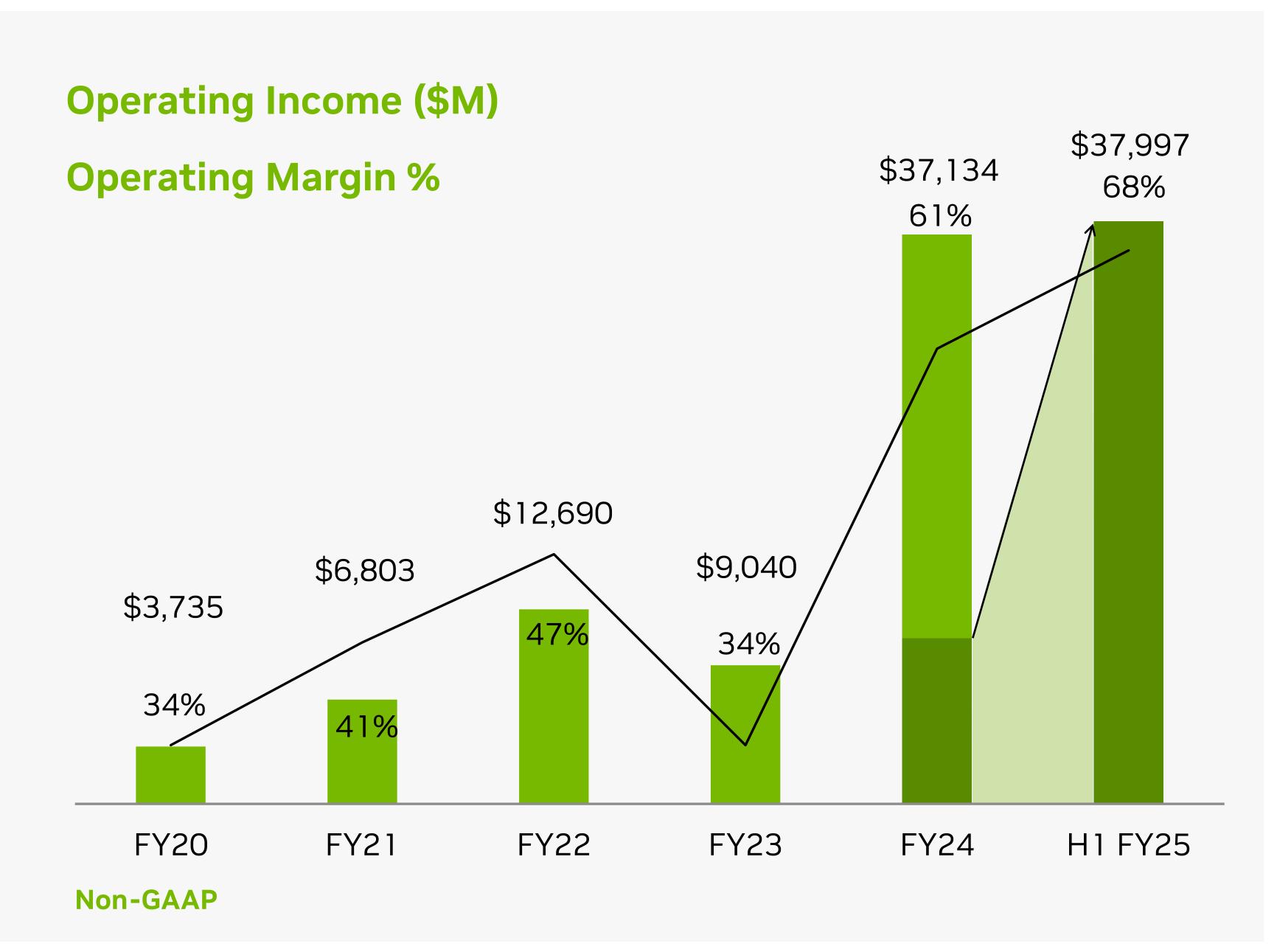






Driving Strong and Profitable Growth

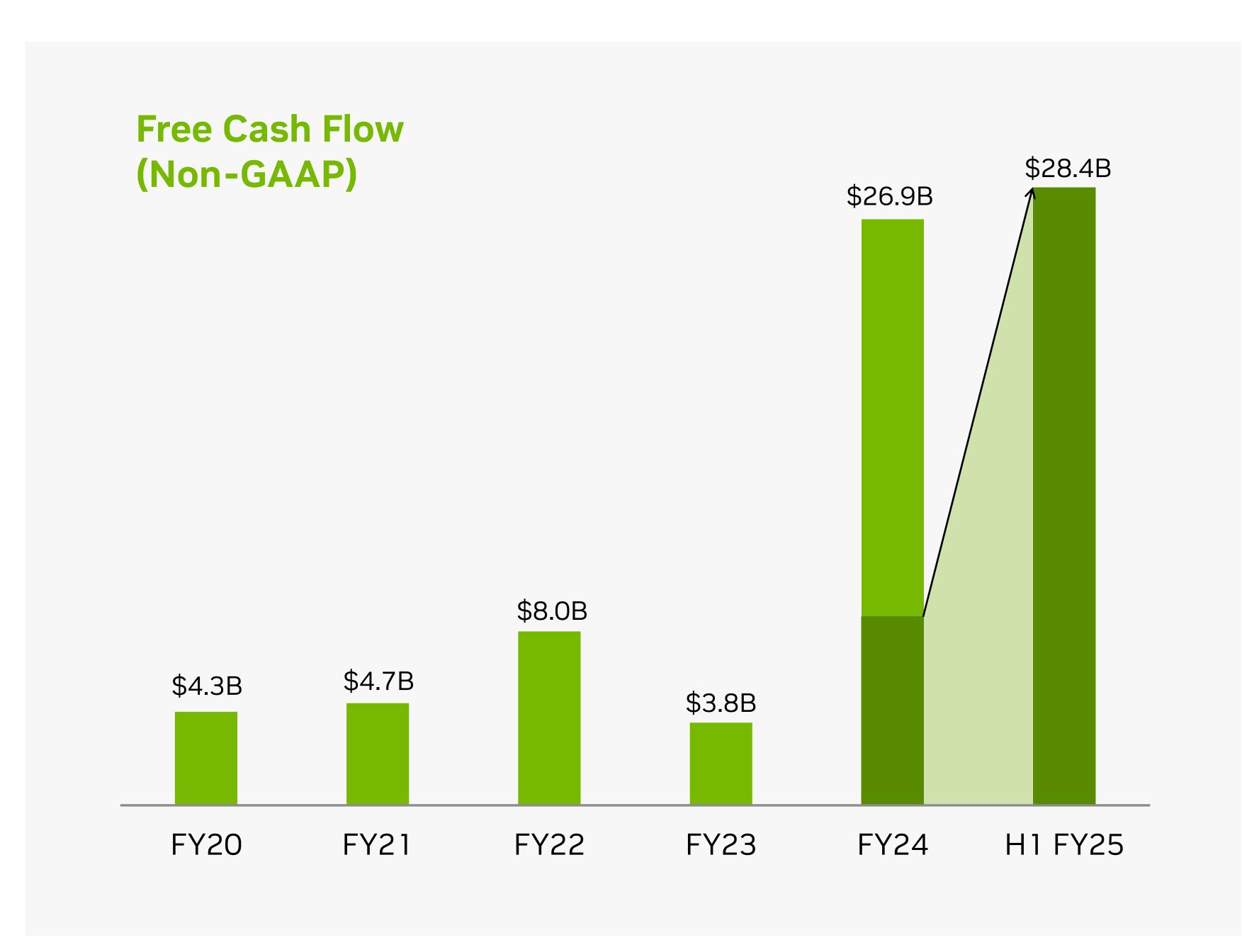




Fiscal year ends in January. Refer to Appendix for reconciliation of Non-GAAP measures. Operating margins rounded to the nearest percent.



Strong Cash Flow Generation



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

Capital Allocation

Share Repurchase

- Utilized \$14.9B of cash for repurchases in H1 FY25
- Additional \$50B in stock repurchase authorization, adding to \$7.5B which remained as of the end of Q2

Dividend

- \$344M in H1 FY25
- Dividend increased by 150% in Q2 FY25
- Plan to Maintain¹

Strategic Investments

- Growing Our Talent
- Platform Reach & Ecosystem



¹ Subject to continuing determination by our Board of Directors.

Our Market Platforms at a Glance



Data Center
78% of FY24 Revenue

FY24 Revenue \$47.5B 5-YR CAGR 75%

DGX/HGX/MGX/IGX systems

GPU | CPU | DPU | Networking

NVIDIA AI software



Gaming 17% of FY24 Revenue

FY24 Revenue \$10.4B 5-YR CAGR 11%

GeForce GPUs for PC gaming
GeForce NOW cloud gaming



Professional Visualization

3% of FY24 Revenue

FY24 Revenue \$1.6B 5-YR CAGR 7%

NVIDIA RTX GPUs for workstations

Omniverse software



Automotive

2% of FY24 Revenue

FY24 Revenue \$1.1B 5-YR CAGR 11%

DRIVE Hyperion sensor architecture with AGX compute

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



Reconciliation of Non-GAAP to GAAP Financial Measures

Reconciliation of Non-GAAP to GAAP Financial Measures

Operating Income and Margin (\$ in Millions & Margin Percentage)	Non-GAAP	Acquisition Termination Cost	Acquisition-Related and Other Costs (A)	Stock-Based Compensation (B)	Other (C)	GAAP
FY 2020	\$3,735		(31)	(844)	(14)	\$2,846
	34.2%		(0.3)	(7.7)	(O.1)	26.1%
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%
EV 2024	\$37,134		(583)	(3,549)	(30)	\$32,972
FY 2024	61.0%		(1.0)	(5.8)	(O.1)	54.1%
H1 FY 2024	\$10,828		(311)	(1,576)		\$8,941
	52.3%		(1.5)	(7.6)		43.2%
	\$37,997		(286)	(2,164)	4	\$35,551
H1 FY 2025	67.8%		(0.5)	(3.9)		63.4%

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

(\$ 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 2020	\$4,272	489		\$4,761
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
H1 FY 2024	\$8,691	537	31	\$9,259
H1 FY 2025	\$28,418	1,346	69	\$29,833



