Investor Presentation
Q3 FY24

November 27, 2023

Except for the historical information contained herein, certain matters in this presentation including, but not limited to, statements as to: our financial position; our markets, market opportunity, demand and growth drivers; a broadening set of GPU-specialized CSPs; entering the holidays with our best-ever line-up for gamers and creators; generative AI emerging as the new "killer app" for high-performance PCs; being on track to exit the year at an annualized revenue run rate of $1 billion for our recurring software, support, and services offerings; AI emerging as a powerful demand driver for Professional Visualization; Foxconn incorporating Omniverse into its manufacturing process; our financial outlook, and expected tax rates for the fourth quarter of fiscal 2024; our expectations of sequential growth to be driven by Data Center, continued strong demand for compute and networking, and Gaming likely declining sequentially; the U.K. government building one of the world's fastest AI supercomputers; Jülich building its next-gen AI supercomputer; the combined AI compute capacity of all the supercomputers built on Grace Hopper across the U.S., EMEA and Japan next year; the benefits, impact, performance, features and availability of our products and technologies; the benefits, impact, features and timing of our collaborations or partnerships; NVIDIA accelerated computing being broadly recognized as the way to advance computing as Moore's law ends and AI lifts off; accelerated computing being needed to tackle the most impactful opportunities of our time; AI driving a platform shift from general purpose to accelerated computing, and enabling new, never-before-possible applications; trillion dollars of installed global data center infrastructure transitioning to accelerated computing; broader enterprise adoption of AI and accelerated computing under way; AI and accelerated computing making possible the next big waves of autonomous machines and industrial digitalization; a rapidly growing universe of applications and industry innovation; AI's ability to augment creativity and productivity; generative AI as the most important computing platform of our generation; data centers becoming AI factories; full-stack and data center scale acceleration driving significant cost savings and workload scaling; the high ROI of high compute performance; our belief that every important company will run its own AI factories; our dividend program plan; AI factories expanding our market opportunity; our Automotive design win pipeline, ramp and production expectations; our aim to engage manufacturing suppliers and goal of effecting supplier adoption of science-based environmental targets by fiscal 2026; and our plan for 100% renewable electricity for our operations and data centers by fiscal 2025 and annually thereafter 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 will occur. Except as required by law, NVIDIA disclaims any obligation to update these 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 time to time with the SEC, including our most recent Annual Report on Form 10-K, Quarterly Reports on Form 10-Q, and Current Reports on Form 8-K. Copies of reports we file with the SEC are posted on our website and are available from NVIDIA without charge.

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 gross profit, non-GAAP gross margin, non-GAAP operating expenses, non-GAAP operating income, non-GAAP operating margin, non-GAAP net income, non-GAAP diluted earnings per share, 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".

Content

- Q3 FY24 Earnings Summary
- Key Announcements This Quarter
- NVIDIA Overview
- Financials
- Reconciliation of Non-GAAP to GAAP Financial Measures

Q3 FY24
Earnings Summary

Highlights

Record quarter driven by strong Data Center growth

- Total revenue up 206% Y/Y to $18.12B, well above outlook of $16.00B +/- 2%
- Data Center up 279% Y/Y to $14.51B
- Gaming up 81% Y/Y to $2.86B

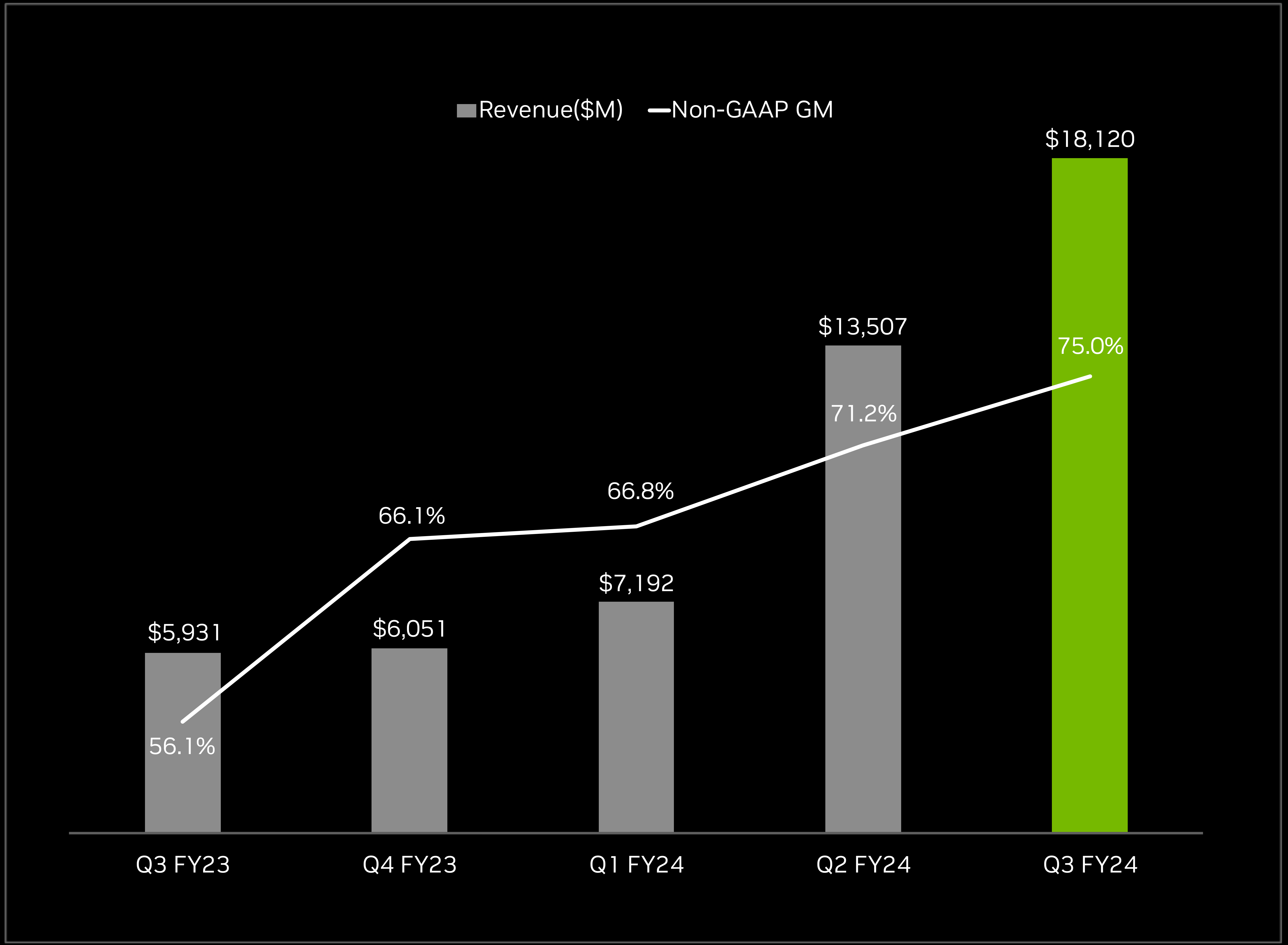
Record Data Center revenue driven by continued ramp of NVIDIA HGX platform and InfiniBand networking

- Consumer internet and enterprise companies drove exceptional sequential growth, outpacing total growth
- Strong demand from all hyperscale cloud service providers (CSPs), and a broadening set of GPU-specialized CSPs
- Inference is contributing significantly to NVIDIA Data Center demand as AI is now in full production

Gaming growth reflects strong demand for GeForce RTX 40 series GPUs for back-to-school and the holidays

- GeForce RTX available at price points as low as $299 — entering the holidays with best-ever line-up for gamers and creators
- Gaming has doubled relative to pre-COVID levels even against the backdrop of lackluster PC market performance
- Gen AI emerging as new “killer app” for high-performance PCs — NVIDIA RTX is the natural platform for AI-application developers

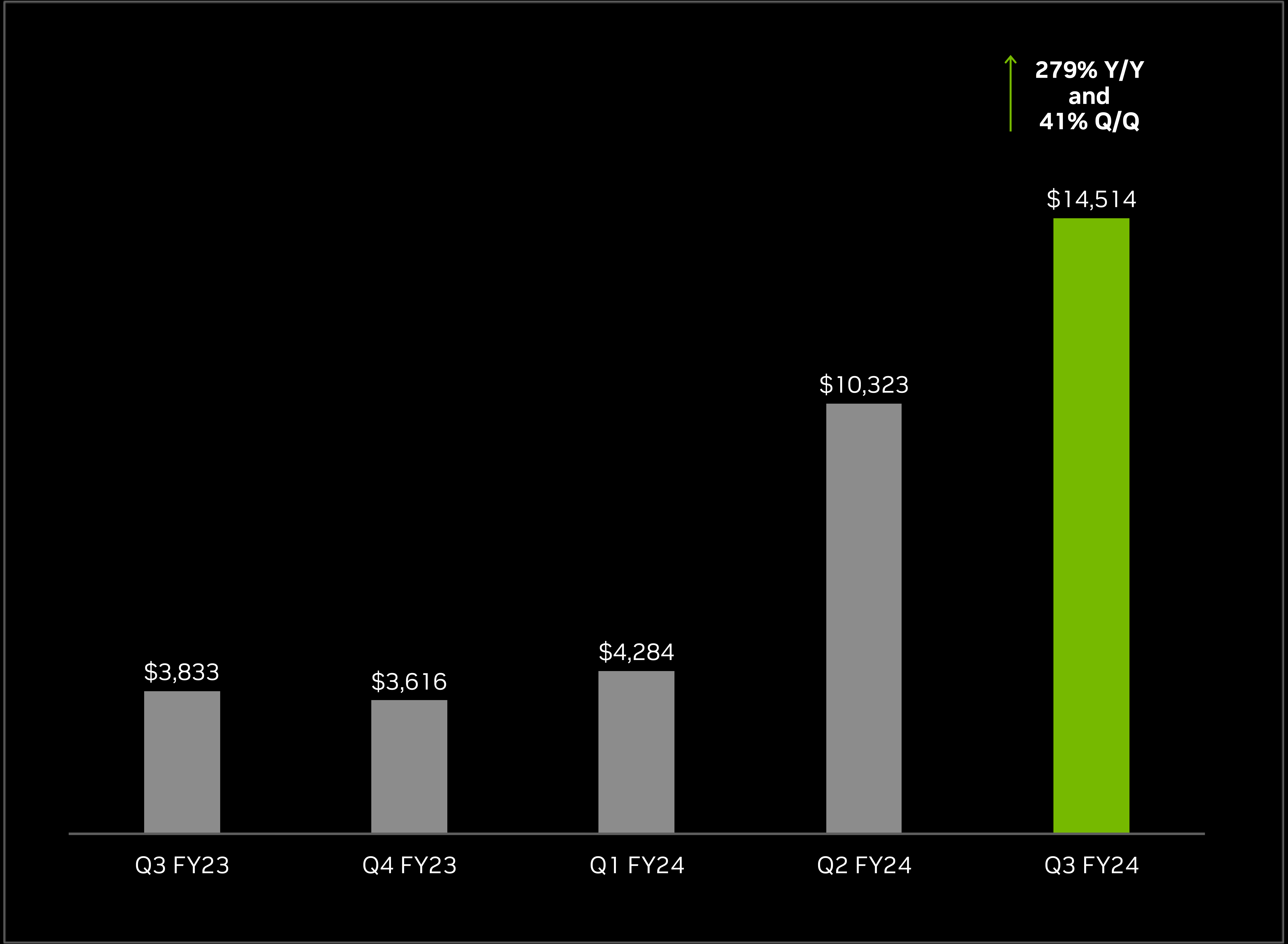
Q3 FY24 Financial Summary



	GAAP			Non-GAAP		
	Q3 FY24	Y/Y	Q/Q	Q3 FY24	Y/Y	Q/Q
Revenue	$18,120	+206%	+34%	$18,120	+206%	+34%
Gross Margin	74.0%	+20.4 pts	+3.9 pts	75.0%	+18.9 pts	+3.8 pts
Operating Income	$10,417	+1,633%	+53%	$11,557	+652%	+49%
Net Income	$9,243	+1,259%	+49%	$10,020	+588%	+49%
Diluted EPS	$3.71	+1,274%	+50%	$4.02	+593%	+49%
Cash Flow from Ops	$7,333	+1,771%	+16%	$7,333	+1,771%	+16%

All dollar figures are in millions other than EPS. Refer to Appendix for reconciliation of Non-GAAP measures.

Data Center

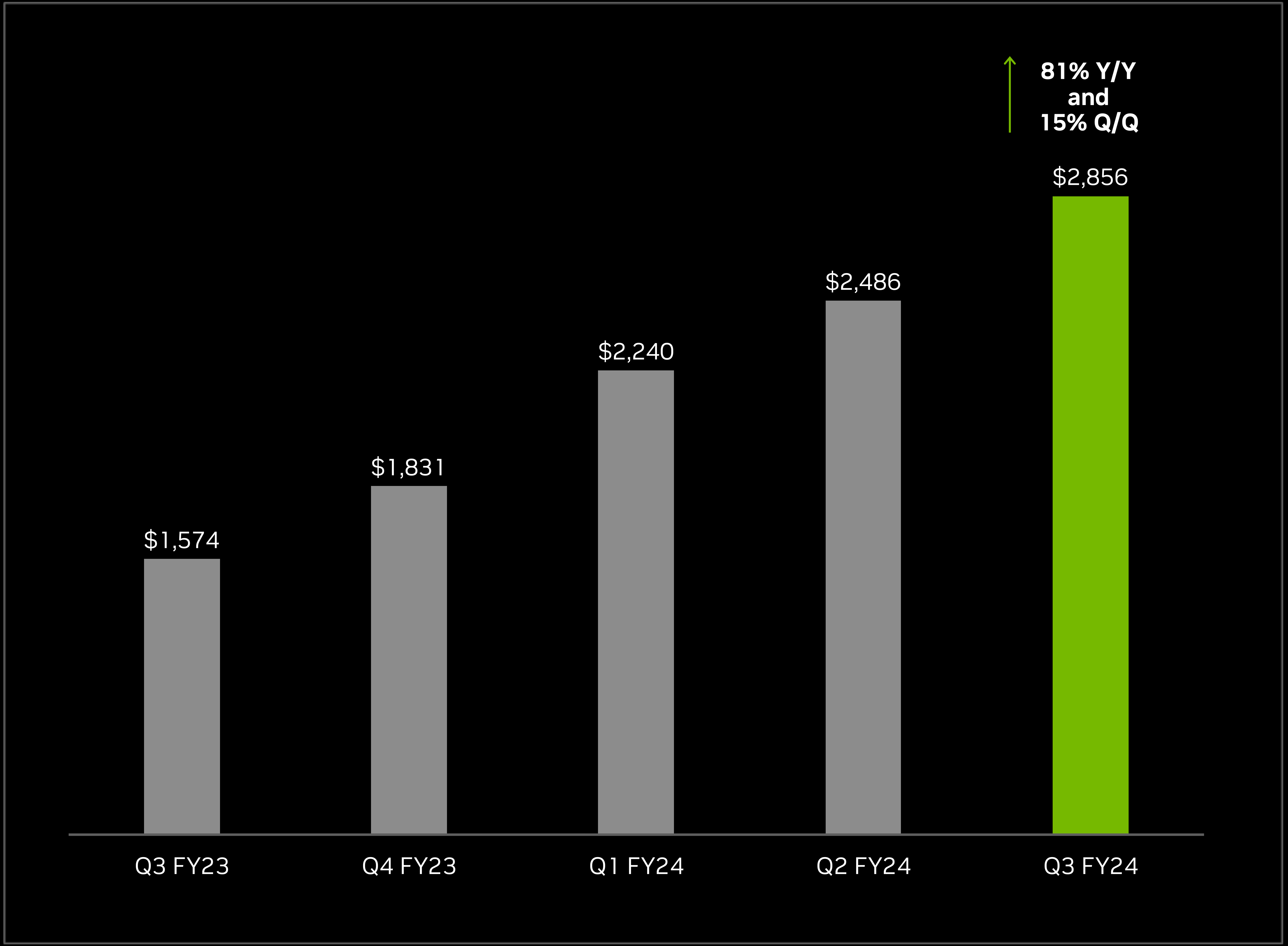


Revenue ($M)

Highlights

- Data Center compute revenue quadrupled from last year, Networking revenue nearly tripled
- Strong, broad-based demand for NVIDIA accelerated computing fueled by investment in the buildout of infrastructure for LLMs, recommendation engines, and gen AI applications
- Networking business now exceeds a $10 billion annualized revenue run rate
- NVIDIA H100 Tensor Core GPU instances are now generally available in virtually every cloud, and are in high demand
- Vast majority of revenue driven by NVIDIA Hopper HGX, with a lower contribution from the prior-gen Ampere GPU architecture
- New L40S GPU began to ship; first revenue quarter for GH200
- On track to exit the year at an annualized revenue run rate of $1 billion for our recurring software, support, and services offerings

Gaming

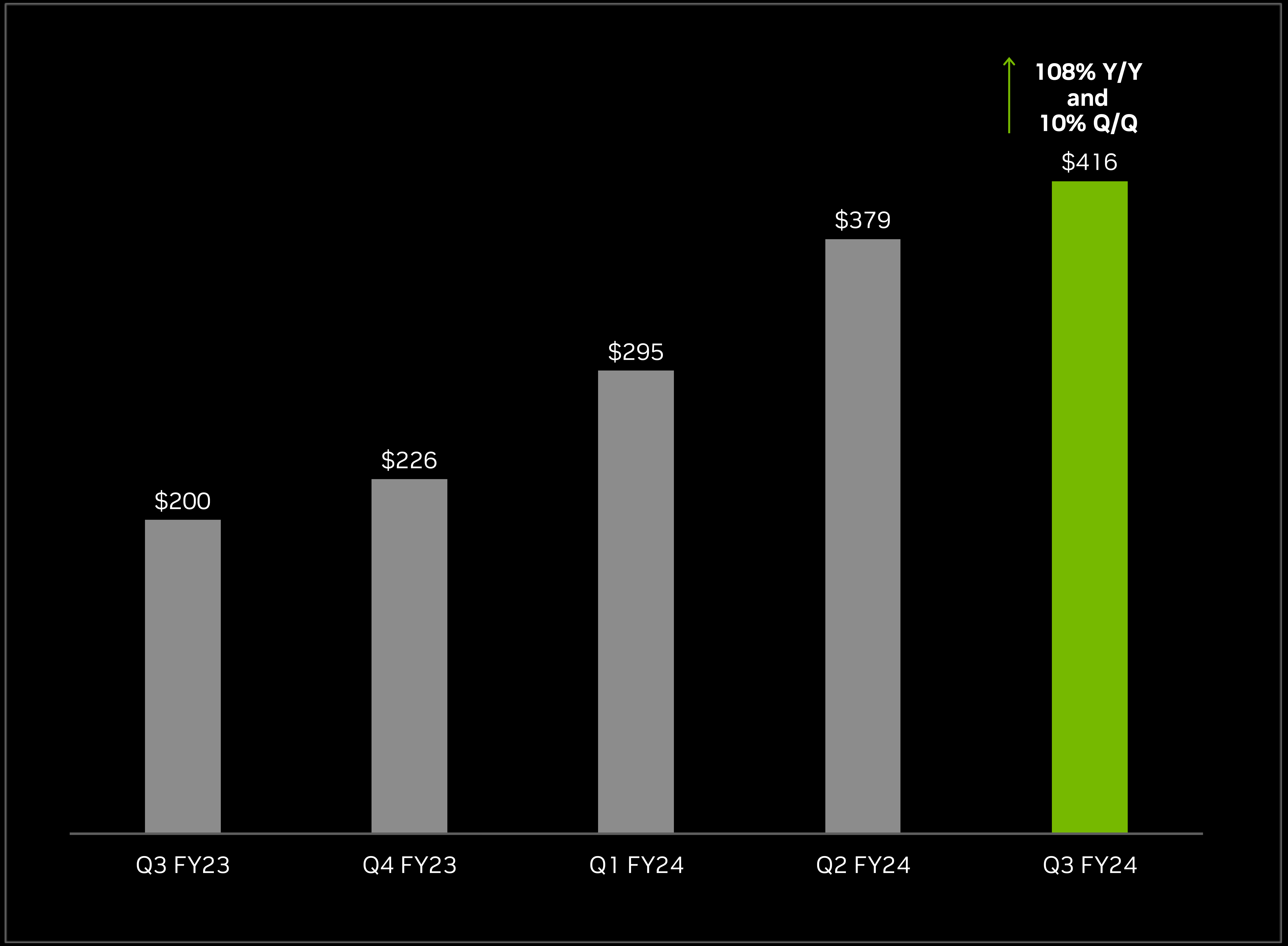


Revenue ($M)

Highlights

- Strong demand in the important back-to-school shopping season
- The RTX ecosystem continues to grow; there are now over 475 RTX enabled games and applications
- Released TensorRT-LLM for Windows, which speeds on-device LLM inference by up to 4X
- GeForce NOW surpassed 1,700 PC titles including Alan Wake II, Baldur's Gate 3, Cyberpunk 2077: Phantom Liberty, and Starfield

Professional Visualization

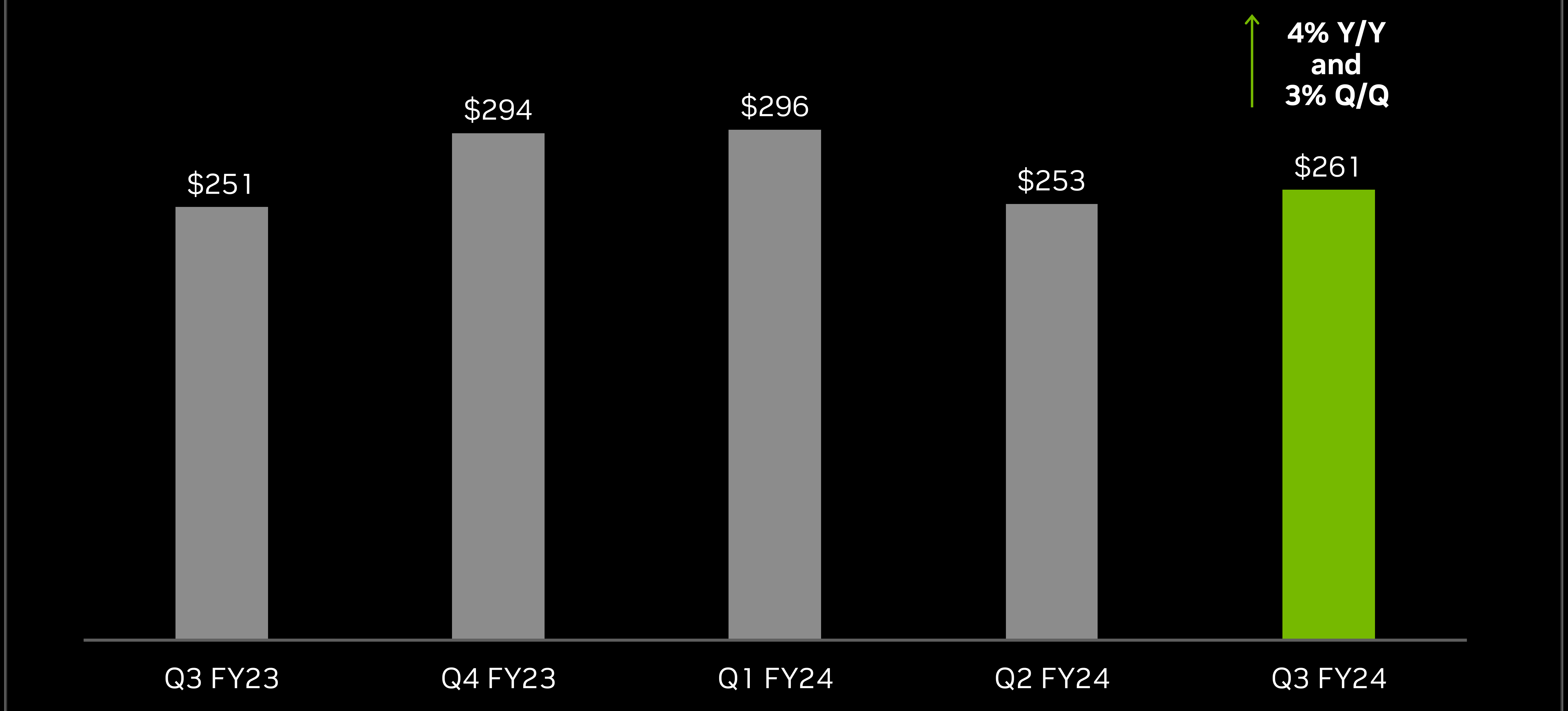


Revenue ($M)

Highlights

- AI emerging as a powerful demand driver, including inference for AI imaging in healthcare, edge AI in smart spaces and the public sector
- Launched a new line of desktop workstations based on NVIDIA RTX Ada Lovelace generation GPUs and ConnectX SmartNICs
- Mercedes-Benz is using Omniverse-powered digital twins to plan, design, build and operate its manufacturing and assembly facilities
- Foxconn will incorporate Omniverse into its manufacturing process
- Announced two new Omniverse Cloud services on Microsoft Azure — for virtual factory simulation and autonomous vehicle simulation

Automotive

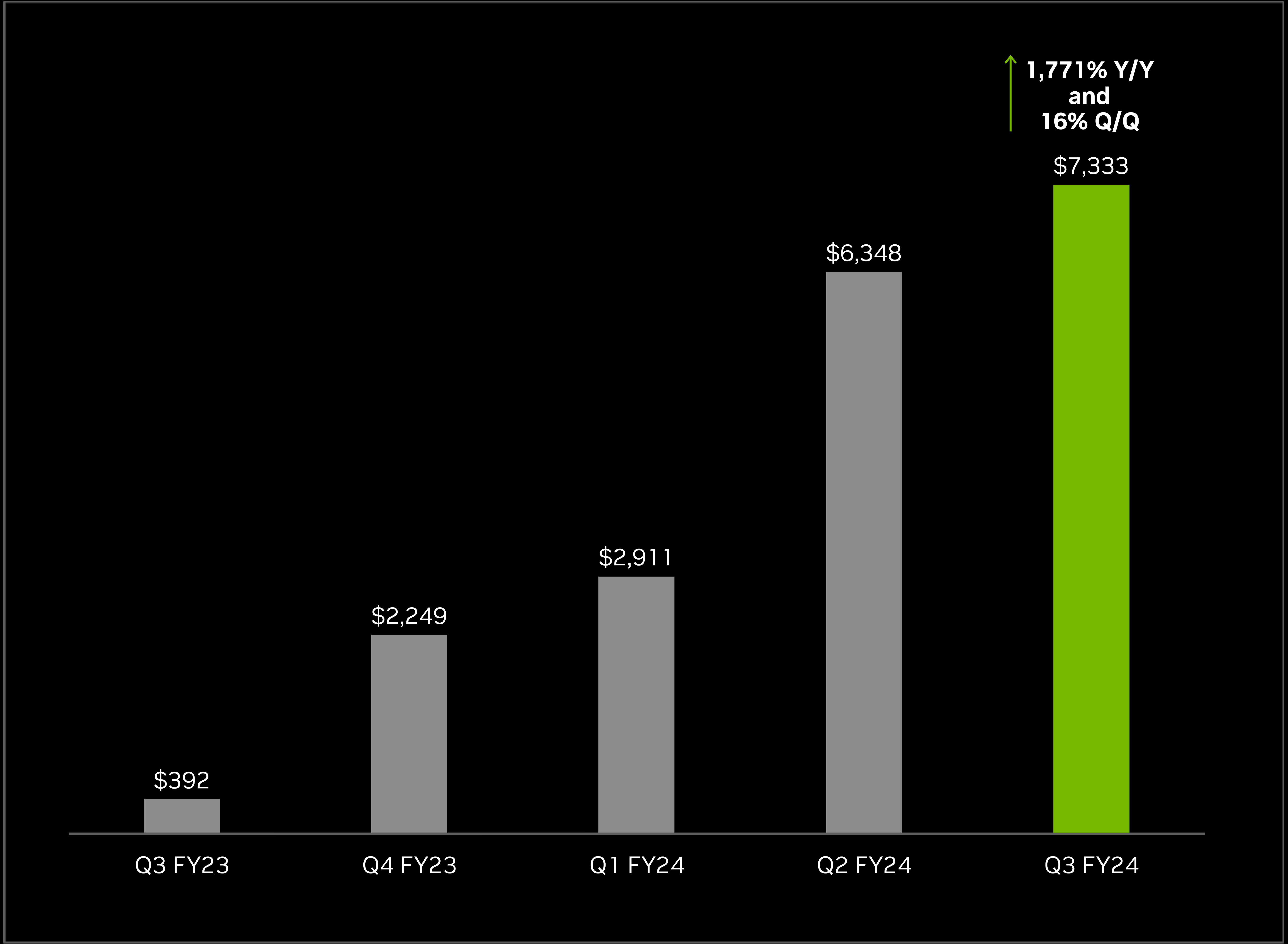


Revenue ($M)

Highlights

- Growth primarily driven by continued growth in self-driving platforms based on NVIDIA DRIVE Orin SoC, and the ramp of AI cockpit solutions with global OEM customers
- Extended automotive partnership with Foxconn to include NVIDIA DRIVE Thor, next-generation automotive SoC

Sources & Uses of Cash



Cash Flow from Operations ($M)

Highlights

- Y/Y and Q/Q growth primarily driven by higher revenue partially offset by higher cash tax payments
- Utilized cash of $3.9 billion towards shareholder returns, including $3.8 billion in share repurchases and $99 million in cash dividends
- Invested $291M in capex (includes principal payments on PP&E)
- Ended the quarter with $18.3B in gross cash and $9.8B in debt; $8.5B in net cash

Gross cash is defined as cash/cash equivalents & marketable securities.
Debt is defined as principal value of debt.
Net cash is defined as gross cash less debt.

Q4 FY24 Outlook

Revenue	**$20.0 billion**, plus or minus 2% Expect strong Q/Q growth to be driven by Data Center, with continued strong demand for both compute and networking. Gaming will likely decline Q/Q, as it is now more aligned with notebook seasonality
Gross Margins	**74.5%** GAAP and **75.5%** non-GAAP, plus or minus 50 basis points
Operating Expense	Approximately **$3.17 billion** GAAP and **$2.20 billion** non-GAAP
Other Income & Expense	Income of approximately **$200 million** for GAAP and non-GAAP Excluding gains and losses on non-affiliated investments
Tax Rate	**15.0%** GAAP and non-GAAP, plus or minus 1%, excluding discrete items

Refer to Appendix for reconciliation of Non-GAAP measures.

Key Announcements This Quarter

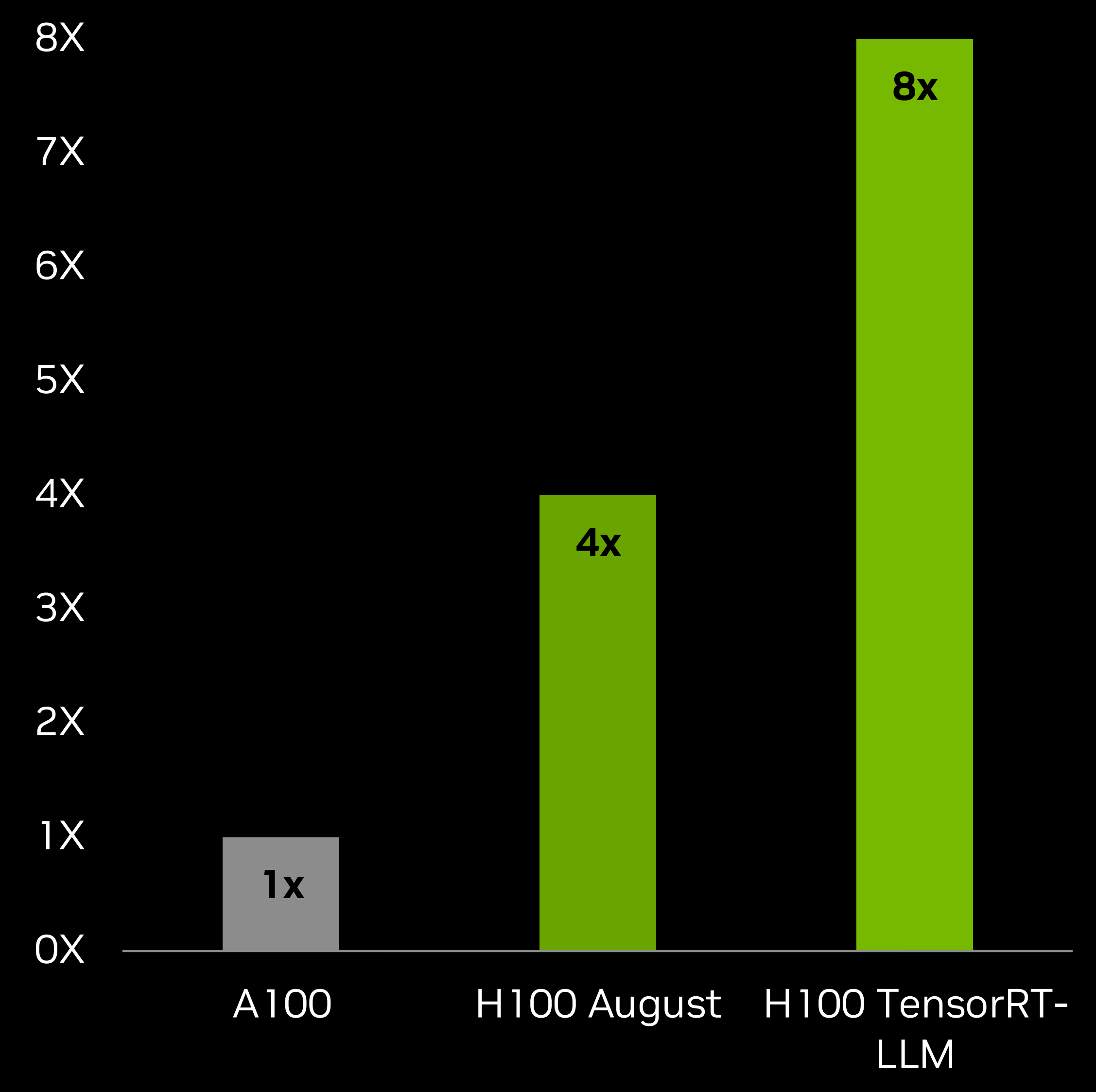
New TensorRT-LLM Software More Than Doubles Inference Performance

- NVIDIA developed TensorRT-LLM, an open-source software library that enables customers to more than double the inference performance of their GPUs
- TensorRT-LLM on H100 GPUs provides up to an 8X performance speedup compared to prior generation A100 GPUs running GPT-J 6B without the software
  - 5.3X reduction in TCO and 5.6X reduction in energy costs
- With TensorRT-LLM for Windows, LLMs and generative AI applications can run up to 4x faster locally on PCs and Workstations powered by NVIDIA GeForce RTX and NVIDIA RTX GPUs
- TensorRT-LLM for data centers now publicly available; TensorRT-LLM for Windows in beta

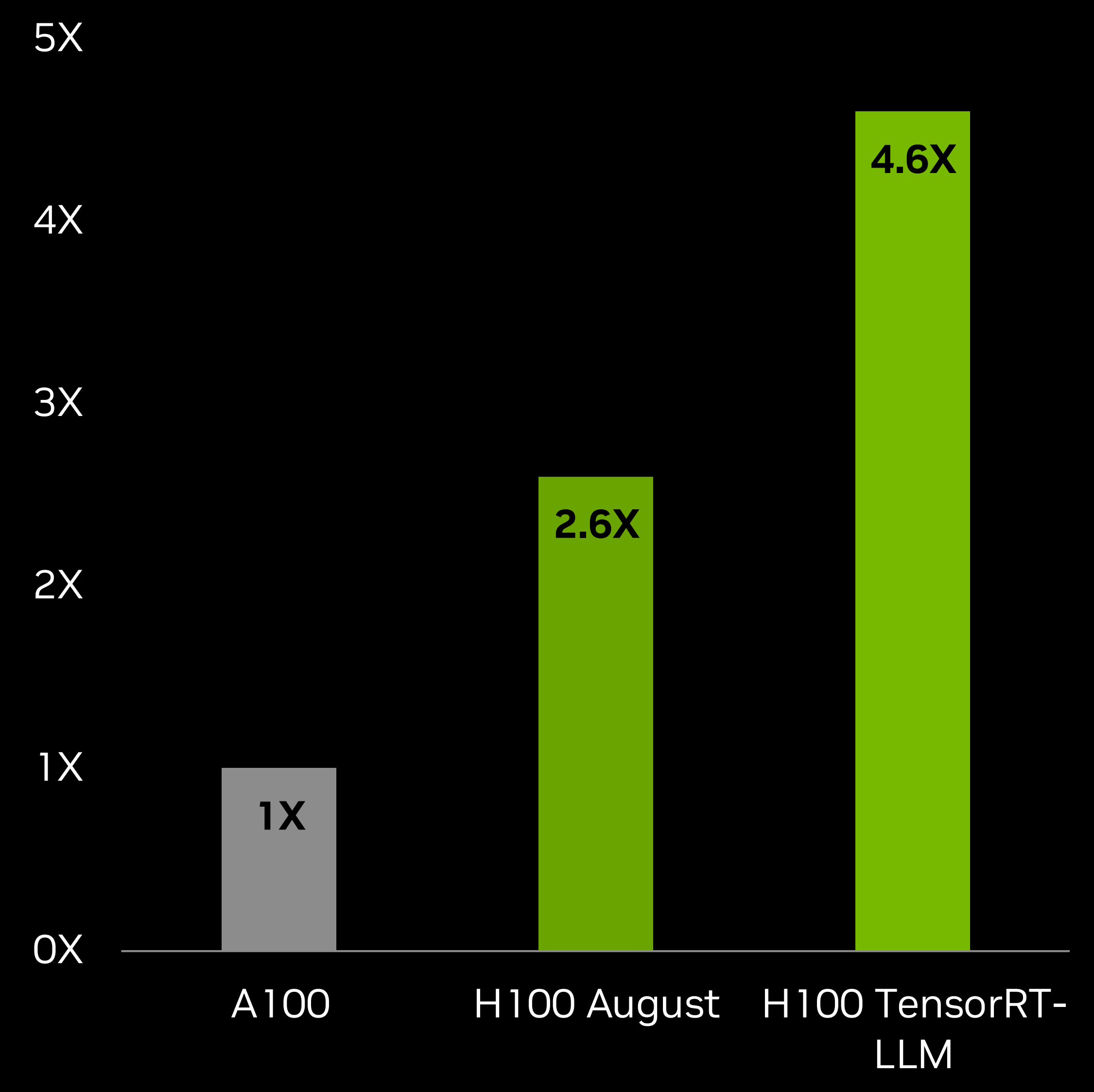
TensorRT-LLM Supercharges Hopper Performance

Software optimizations double leading performance

8X Increase in GPT-J 6B Inference Performance



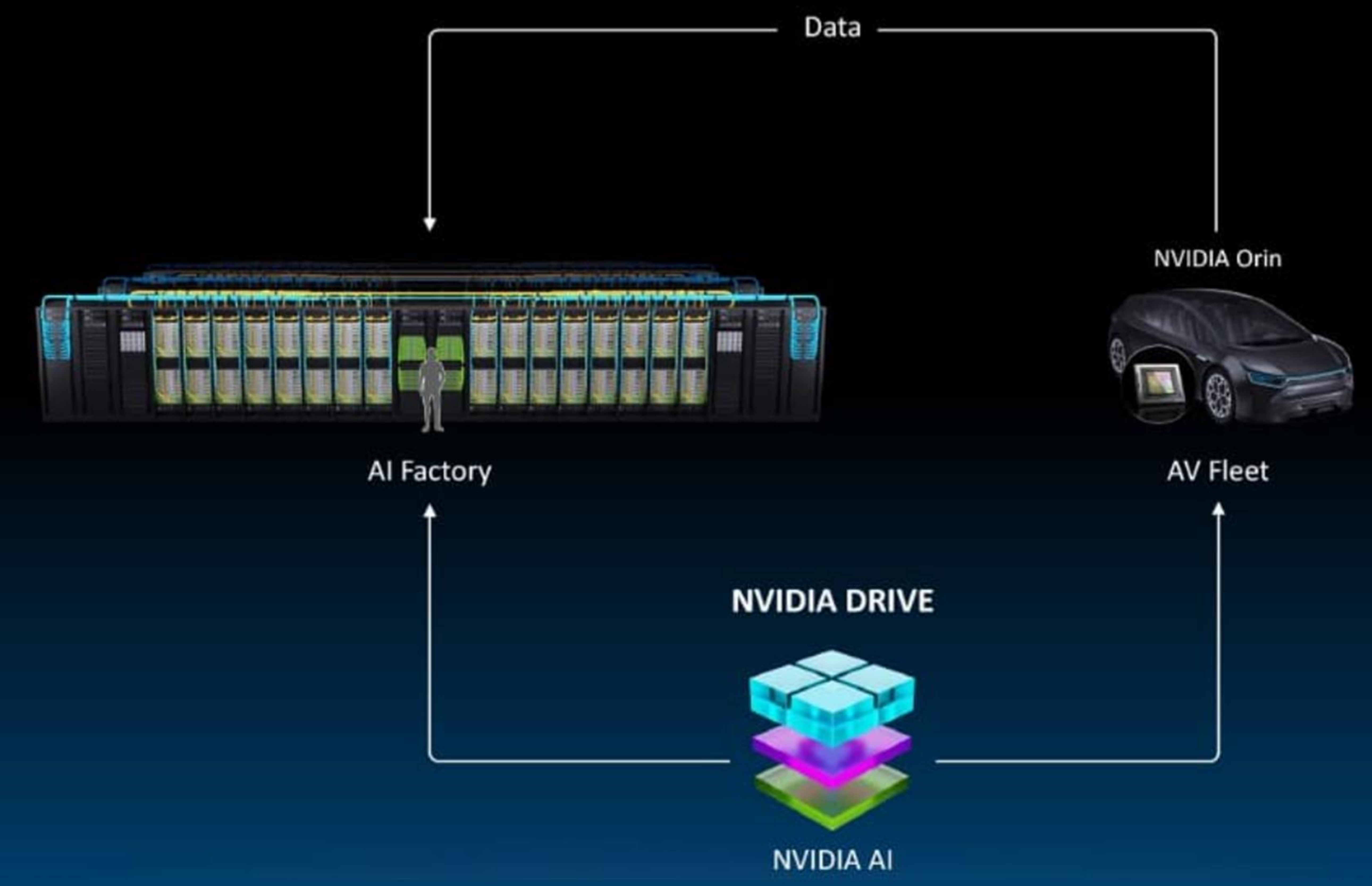
4.6X Higher Llama2 Inference Performance



Text summarization, variable input/output length, CNN/ DailyMail dataset | A100 FP 16 PyTorch eager mode / H100 FP8 | H100 FP8, TensorRT-LLM, in-flight batching

NVIDIA Partners With Foxconn to Build Factories and Systems for the AI Industrial Revolution

- Foxconn, the world’s largest manufacturer, will integrate NVIDIA technology to develop “AI factories”, a new class of data centers
- Based on the NVIDIA accelerated computing platform, including NVIDIA GH200 and NVIDIA AI Enterprise software, these AI factories will power a wide range of applications, including:
  - Digitalization of manufacturing and inspection workflows
  - Development of AI-powered EVs and robotics platforms
  - A growing number of language-based generative AI services
- In addition:
  - Foxconn Smart EV will be built on NVIDIA DRIVE Hyperion 9, next-gen platform for autonomous automotive fleets, powered by NVIDIA DRIVE Thor, our future automotive SoC
  - Foxconn Smart Manufacturing robotic systems will be built on the NVIDIA Isaac autonomous mobile robot platform.
  - Foxconn Smart City will incorporate the NVIDIA Metropolis intelligent video analytics platform



AI factories are a new class of data centers, optimized for refining data and training, inferencing, and generating AI

NVIDIA Partners With India Tech Giants to Advance AI Across World's Most Populous Nation

NVIDIA announced collaborations with Reliance Industries, Tata Group and Infosys to bring AI technology and skills to India

- With **Reliance**, the companies will work together to develop India's own foundation LLM trained on India's diverse languages and tailored for generative AI applications; build supercomputing infrastructure to support the exponential computational demands of AI
- With **Tata**, the collaboration will bring a state-of-the-art AI supercomputer to provide infrastructure-as-a-service and platform for AI services in India
- With **Infosys**, the partnership will bring the NVIDIA AI Enterprise ecosystem of models, tools, runtimes and GPU systems to drive productivity gains with generative AI applications and solutions
  - Infosys plans to set up an NVIDIA Center of Excellence where it will train and certify 50,000 of its employees on NVIDIA AI technology

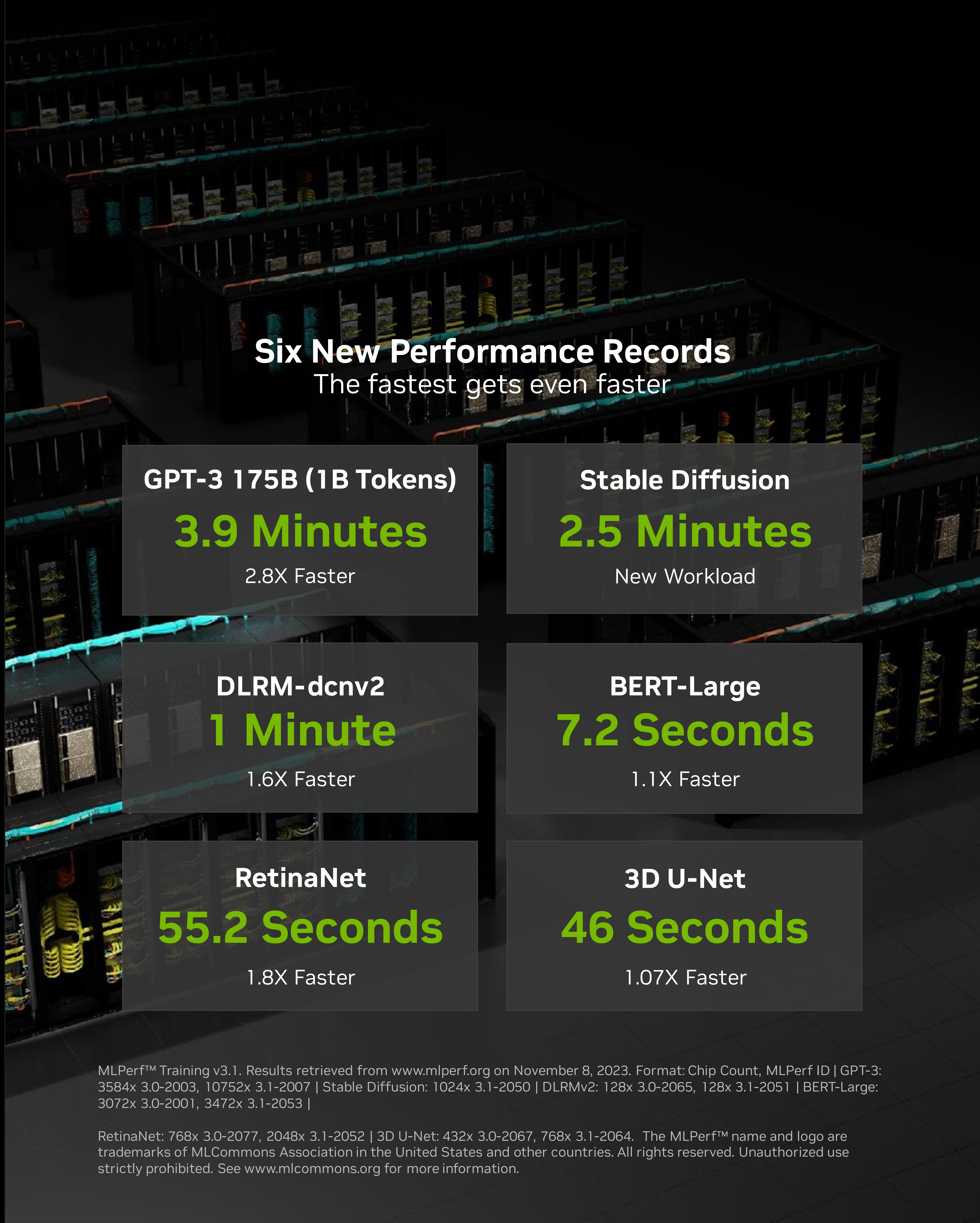
Infosys®

Reliance
Industries Limited

TATA

NVIDIA Sets New LLM Training Record With Largest MLPerf Submission Ever

- NVIDIA set six new performance records in this round, with the performance increase stemming from a combination of advances in software and scaled-up hardware
  - 2.8x faster on generative AI – completing a training benchmark based on a GPT-3 model with 175 billion parameters trained on 1 billion tokens in just 3.9 minutes
  - 1.6x faster on training recommender models
  - 1.8x faster on training computer vision models
- The GPT-3 benchmark ran on NVIDIA Eos – a new AI supercomputer powered by 10,752 H100 GPUs and NVIDIA Quantum-2 InfiniBand networking
- The 10,752 H100 GPUs far surpassed the scaling in AI training in June, when NVIDIA used 3,584 Hopper GPUs
  - The 3x scaling in GPU numbers delivered a 2.8x scaling in performance, a 93% efficiency rate thanks in part to software optimizations
- Microsoft Azure achieved similar results on a nearly identical cluster, demonstrating the efficiency of NVIDIA AI in public cloud deployments



Six New Performance Records

The fastest gets even faster

GPT-3 175B (1B Tokens) **3.9 Minutes** 2.8X Faster	**Stable Diffusion** **2.5 Minutes** New Workload
DLRM-dcnv2 **1 Minute** 1.6X Faster	**BERT-Large** **7.2 Seconds** 1.1X Faster
RetinaNet **55.2 Seconds** 1.8X Faster	**3D U-Net** **46 Seconds** 1.07X Faster

MLPerf™ Training v3.1. Results retrieved from www.mlperf.org on November 8, 2023. Format: Chip Count, MLPerf ID | GPT-3: 3584x 3.0-2003, 10752x 3.1-2007 | Stable Diffusion: 1024x 3.1-2050 | DLRMv2: 128x 3.0-2065, 128x 3.1-2051 | BERT-Large: 3072x 3.0-2001, 3472x 3.1-2053 |

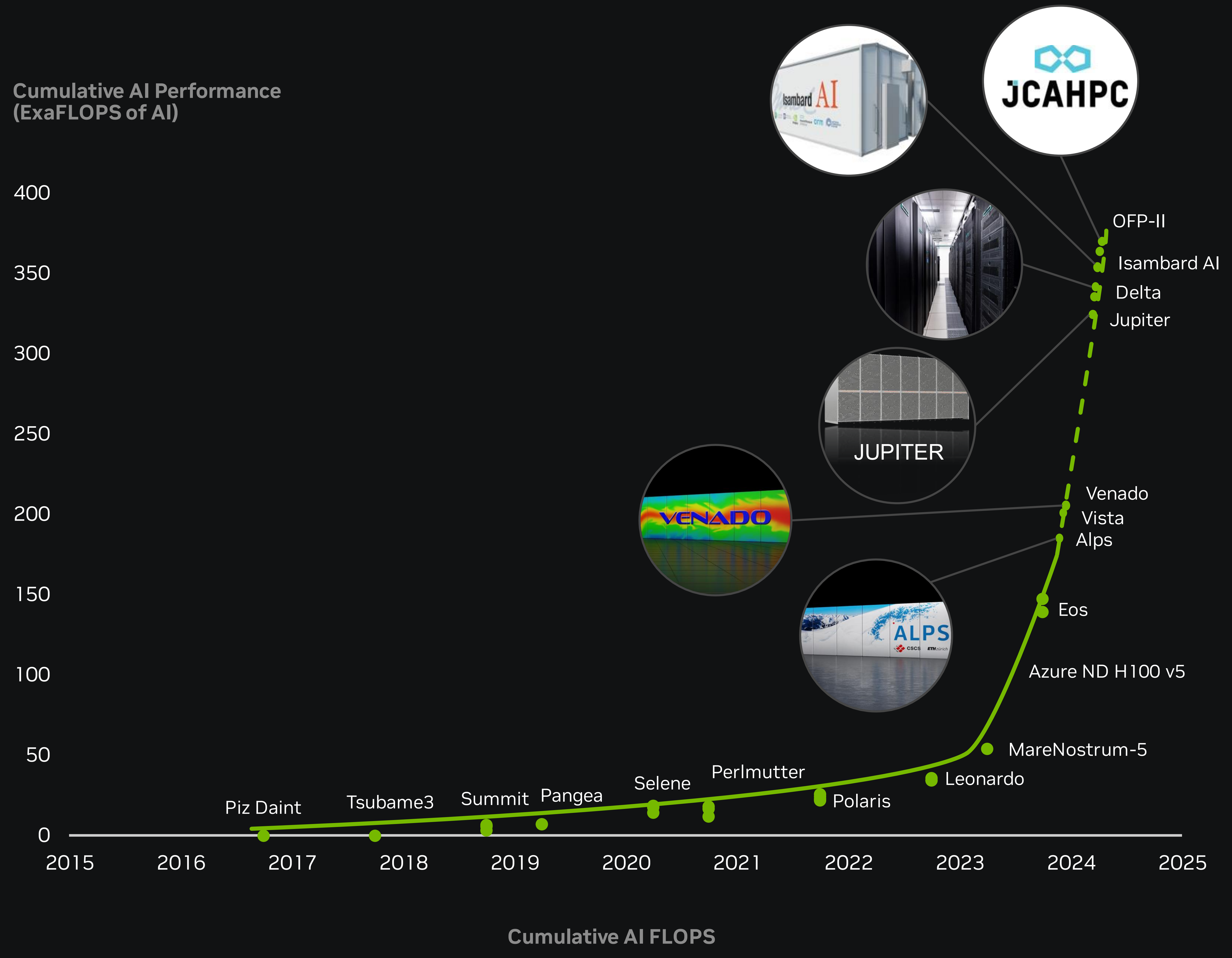
RetinaNet: 768x 3.0-2077, 2048x 3.1-2052 | 3D U-Net: 432x 3.0-2067, 768x 3.1-2064. The MLPerf™ name and logo are trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

New NVIDIA HGX H200 Supercharges Hopper

- NVIDIA H200 is the first GPU to offer HBM3e — faster, larger memory to fuel the acceleration of generative AI and large language models, while advancing scientific computing for HPC workloads
- H200 delivers 141GB of memory at 4.8 terabytes per second, nearly double the capacity and 2.4X more bandwidth compared with its predecessor, NVIDIA A100
- Boosts inference speed by up to 2X compared to H100 GPUs when handling LLMs such as Llama2
- Microsoft announced plans to add the H200 to Azure next year for larger model inference with no increase in latency
- H200-powered systems from the world's leading server manufacturers and cloud service providers are expected to begin shipping in the second quarter of 2024

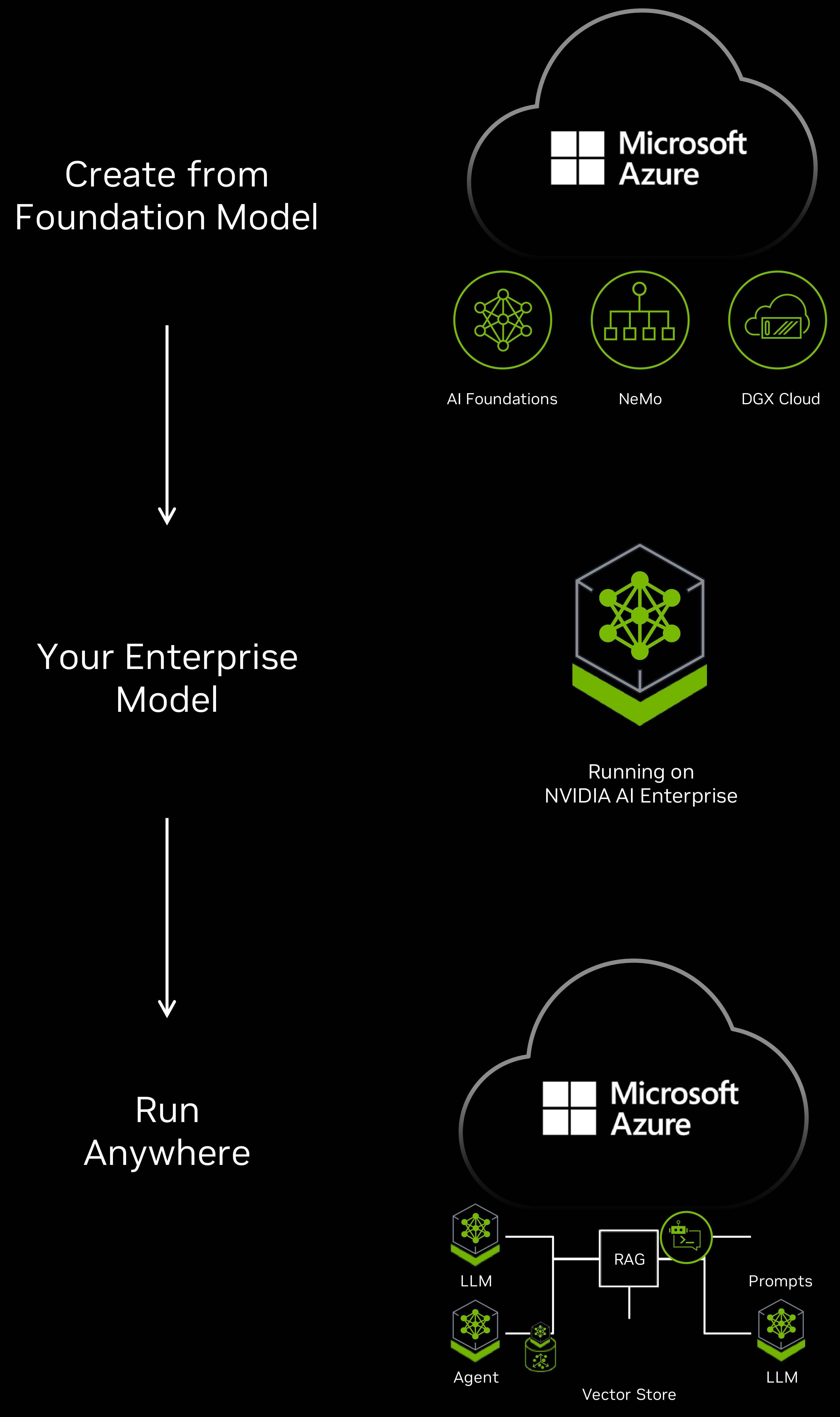
Grace Hopper Gains Significant Traction with Supercomputing Customers

- Initial shipments to Los Alamos National Lab and the Swiss National Supercomputing Centre took place in the third quarter
- The U.K. government announced it will build one of the world's fastest AI supercomputers with almost 5.5K Grace Hopper Superchips
- German supercomputing center Jülich will build its next-gen AI supercomputer, with close to 24K Grace Hopper Superchips and Quantum-2 InfiniBand
  - Will be the world's most powerful AI system with over 90 exaflops of AI performance
  - Marks the debut of a quad NVIDIA GH200 Grace Hopper Superchip node configuration
- Combined AI compute capacity of all the supercomputers built on Grace Hopper across the U.S., EMEA and Japan next year estimated to exceed 200 exaflops



NVIDIA AI Foundry Service for Enterprises on Microsoft Azure

- Introduced new NVIDIA AI foundry service for the development and tuning of custom generative AI enterprise applications, running on Microsoft Azure
- Customers can bring their domain knowledge and proprietary data, and we help them build their AI models using our AI expertise and software stack in DGX Cloud AI factory – all with enterprise-grade security and support
- Businesses can deploy their customized models with the NVIDIA AI Enterprise software runtime to power generative AI applications such as intelligent search, summarization, and content generation
- Industry leaders SAP SE, Amdocs and Getty Images are among the first customers of NVIDIA AI foundry service



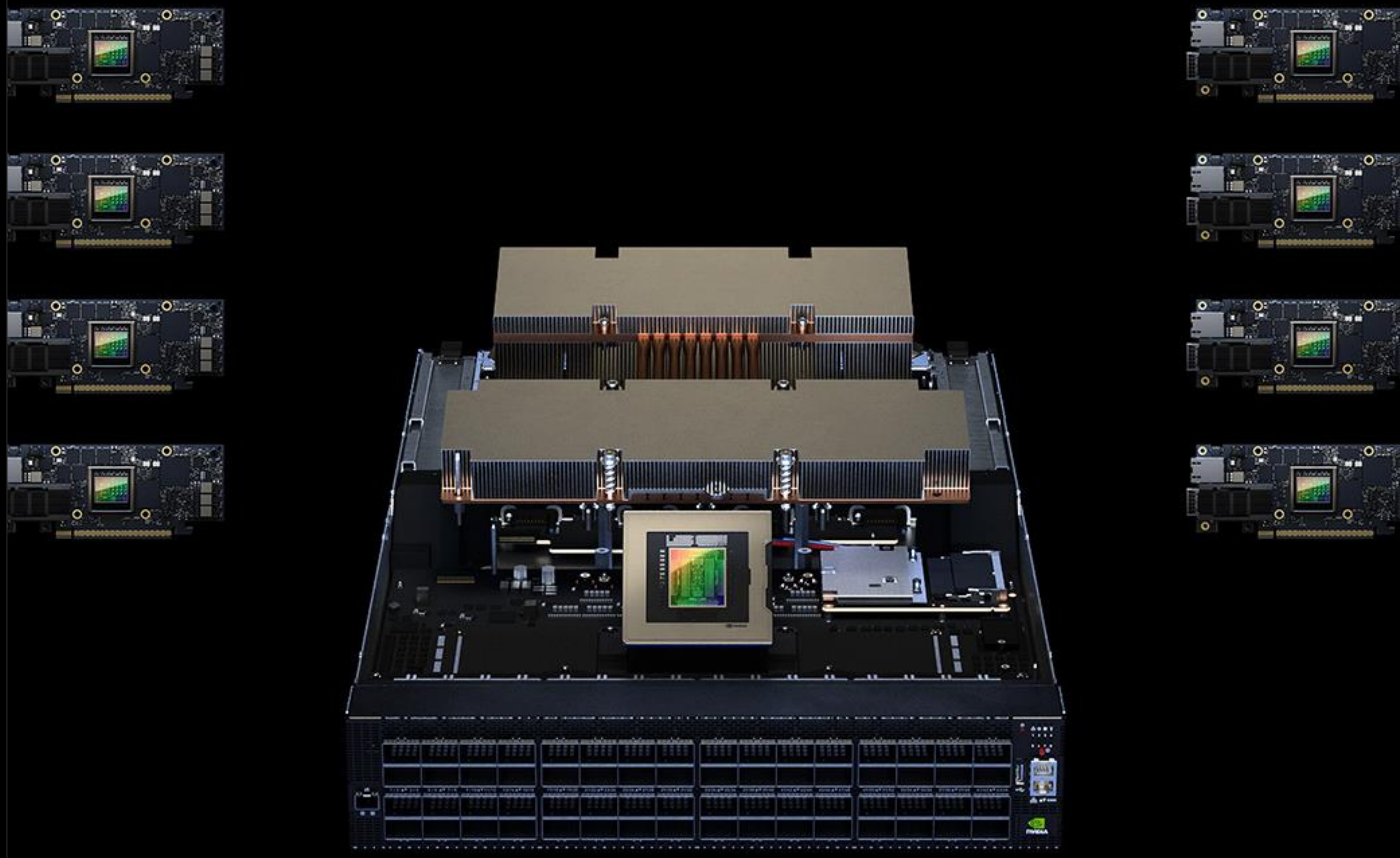
Create from Foundation Model

Your Enterprise Model

Run Anywhere

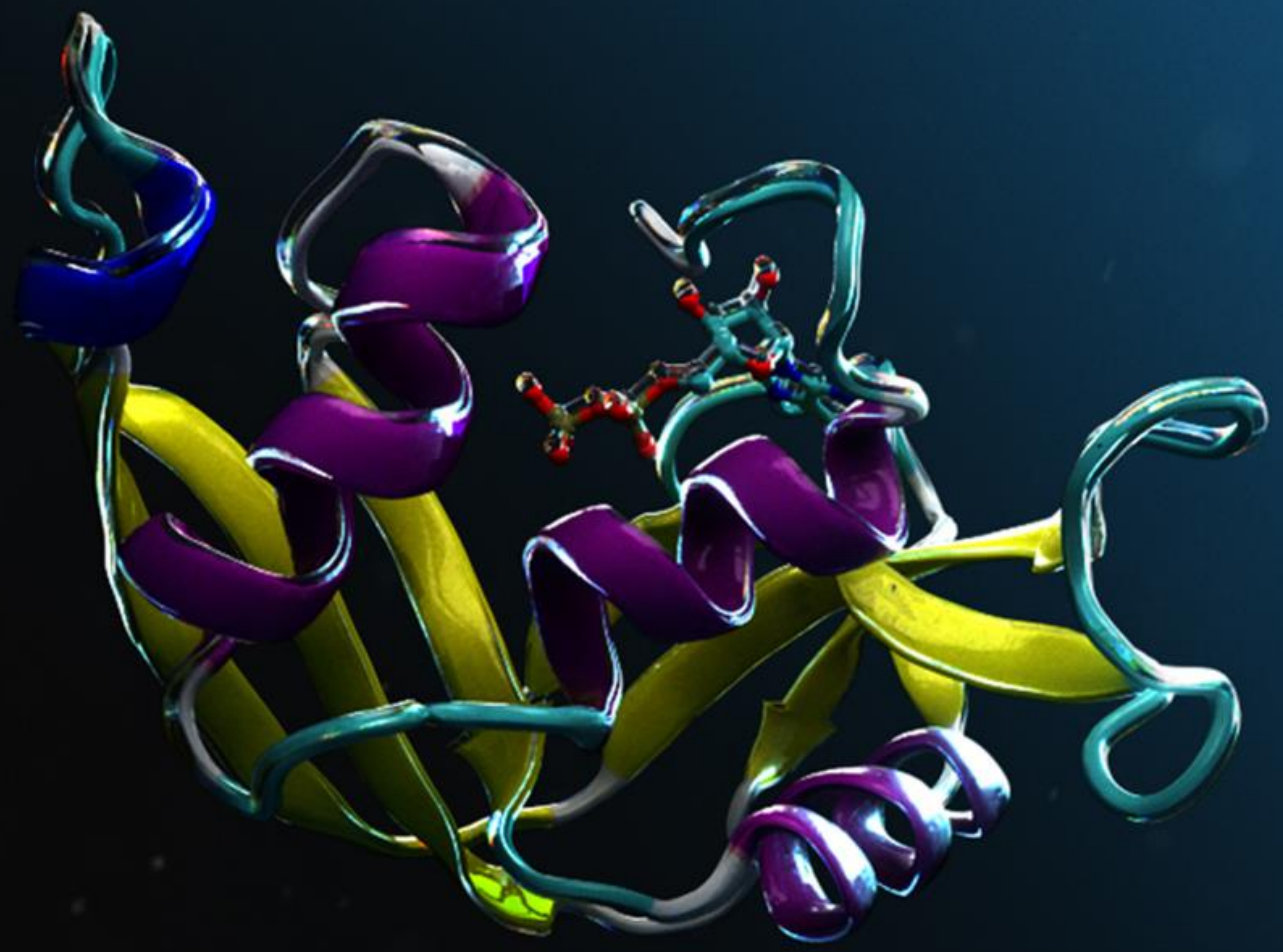
NVIDIA Spectrum-X Ethernet networking platform for AI Available Soon from Dell, HPE and Lenovo

- Purpose-built for gen AI, Spectrum-X offers enterprises a new class of Ethernet networking that can achieve 1.6x higher networking performance for AI communication versus traditional Ethernet offerings
- Dell, Hewlett Packard Enterprise and Lenovo will be the first to integrate NVIDIA Spectrum-X Ethernet networking technologies for AI into their server lineups
- New systems bring together Spectrum-X with NVIDIA GPUs, NVIDIA AI Enterprise software and NVIDIA AI Workbench software to provide enterprises the building blocks to transform their businesses with generative AI
- Available in the first quarter of next year



NVIDIA Collaborates With Genentech to Accelerate Drug Discovery Using Generative AI

- Genentech is pioneering the use of generative AI to discover and develop new therapeutics and deliver treatments to patients more efficiently
- NVIDIA will work with Genentech to accelerate Genentech's proprietary algorithms on NVIDIA DGX Cloud
- Genentech plans to use NVIDIA BioNeMo to help accelerate and optimize their AI drug discovery platform
- NVIDIA plans to use insights learned from this collaboration to improve its BioNeMo platform
- BioNeMo is now generally available as a training service



NVIDIA Overview

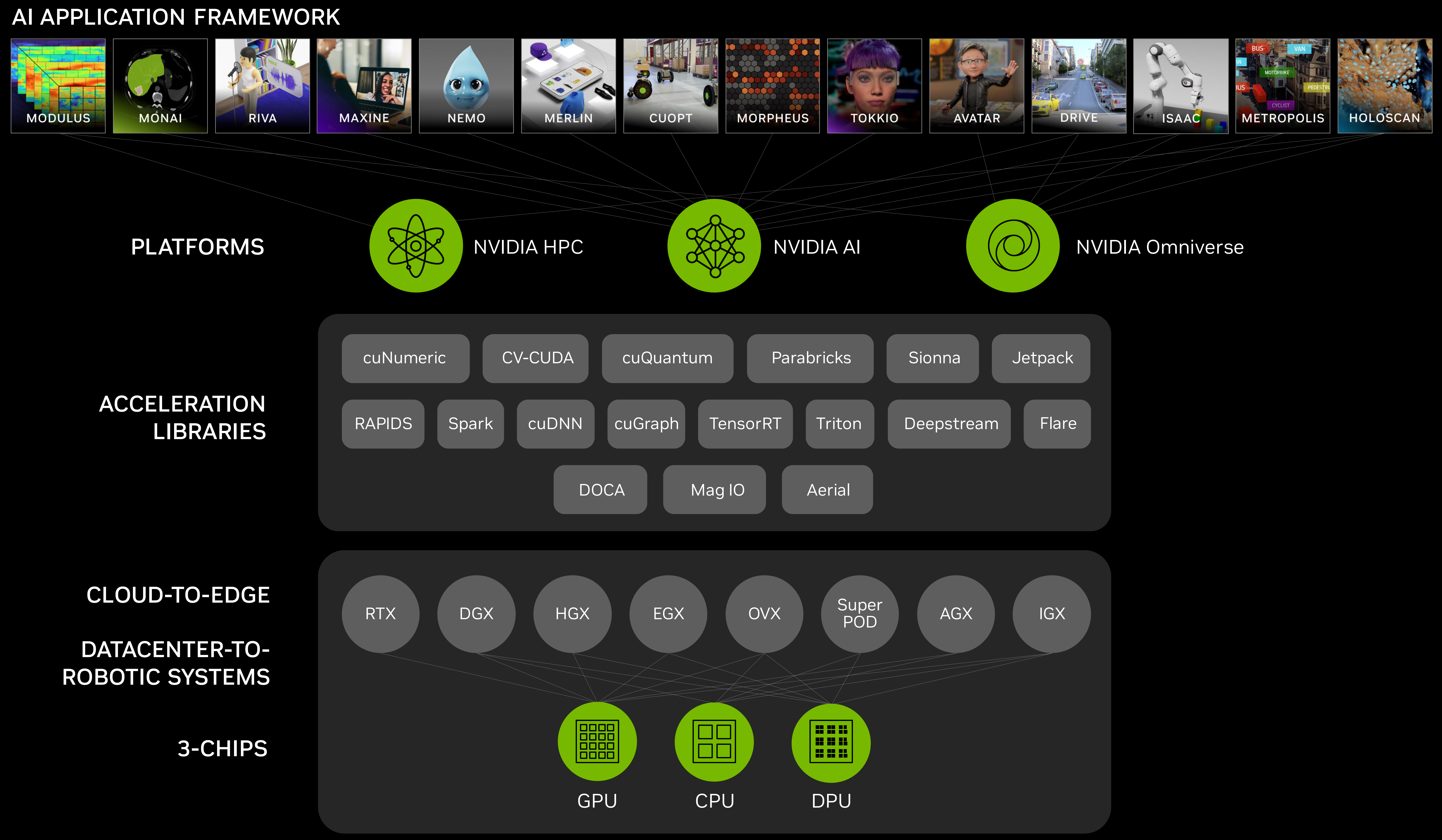
Headquarters: Santa Clara, CA

NVIDIA pioneered accelerated computing to help solve impactful challenges classical computers cannot. A quarter of a century in the making, NVIDIA accelerated computing is broadly recognized as the way to advance computing as Moore's law ends and AI lifts off.

NVIDIA's platform is installed in several hundred million computers, is available in every cloud and from every server maker, powers 76% of the TOP500 supercomputers, and boasts 4.5 million developers.

NVIDIA's Accelerated Computing Platform

Full-stack innovation across silicon, systems and software



AI APPLICATION FRAMEWORK

MODULUS · MONAI · RIVA · MAXINE · NEMO · MERLIN · CUOPT · MORPHEUS · TOKKIO · AVATAR · DRIVE · ISAAC · METROPOLIS · HOLOSCAN

PLATFORMS

NVIDIA HPC · NVIDIA AI · NVIDIA Omniverse

ACCELERATION LIBRARIES

cuNumeric · CV-CUDA · cuQuantum · Parabricks · Sionna · Jetpack

RAPIDS · Spark · cuDNN · cuGraph · TensorRT · Triton · Deepstream · Flare

DOCA · Mag IO · Aerial

CLOUD-TO-EDGE

DATACENTER-TO-ROBOTIC SYSTEMS

RTX · DGX · HGX · EGX · OVX · Super POD · AGX · IGX

3-CHIPS

GPU · CPU · DPU

With nearly three decades of singular focus, NVIDIA is expert at accelerating software and scaling compute by a **Million-X**, going well beyond Moore's law

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

Our full-stack platforms — NVIDIA HPC, NVIDIA AI, and NVIDIA Omniverse — accelerate high performance computing, AI and industrial digitalization workloads

We accelerate workloads at **data center scale**, across thousands of compute nodes, treating the network and storage as part of the computing fabric

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

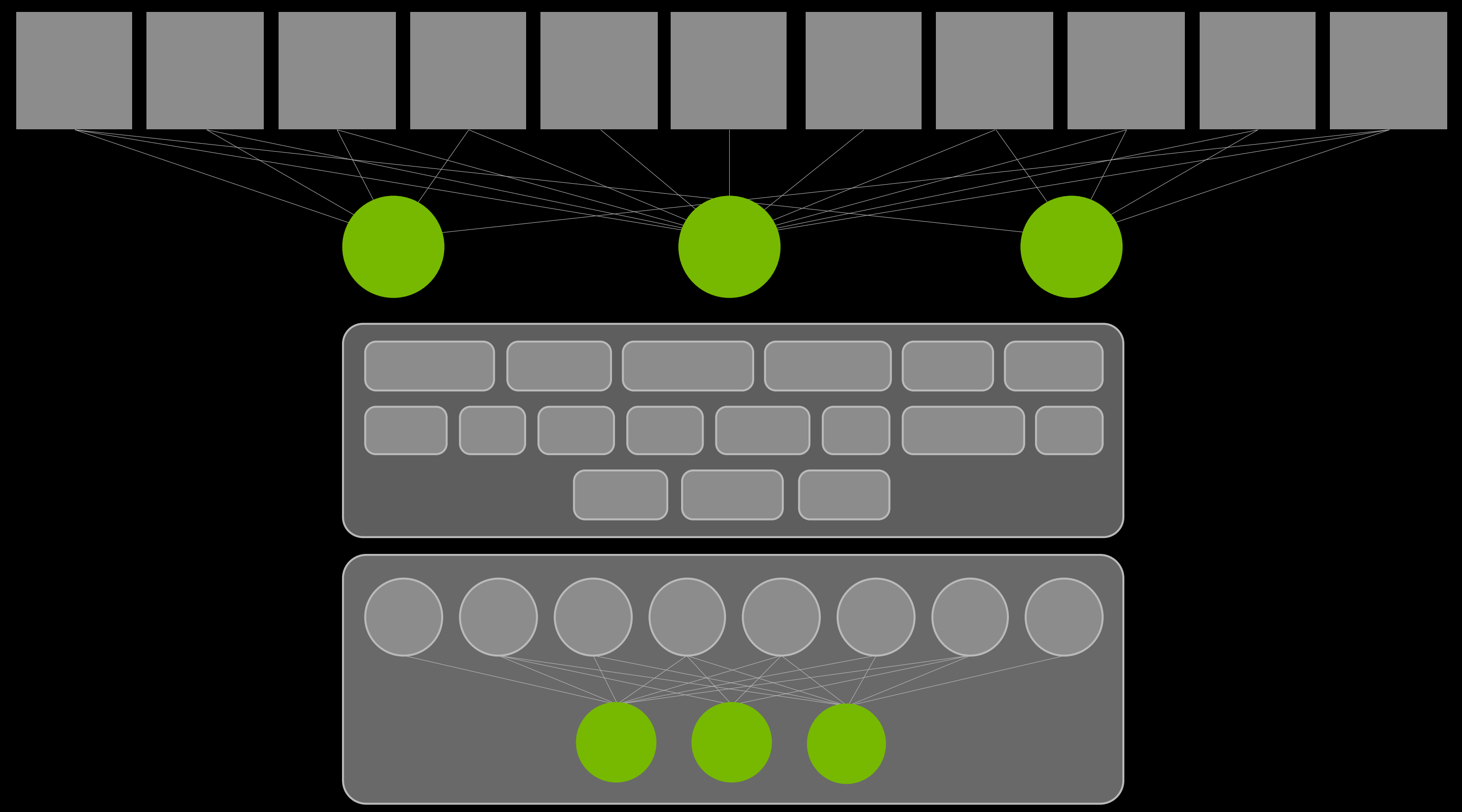
What Is Accelerated Computing?

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 1000X, 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

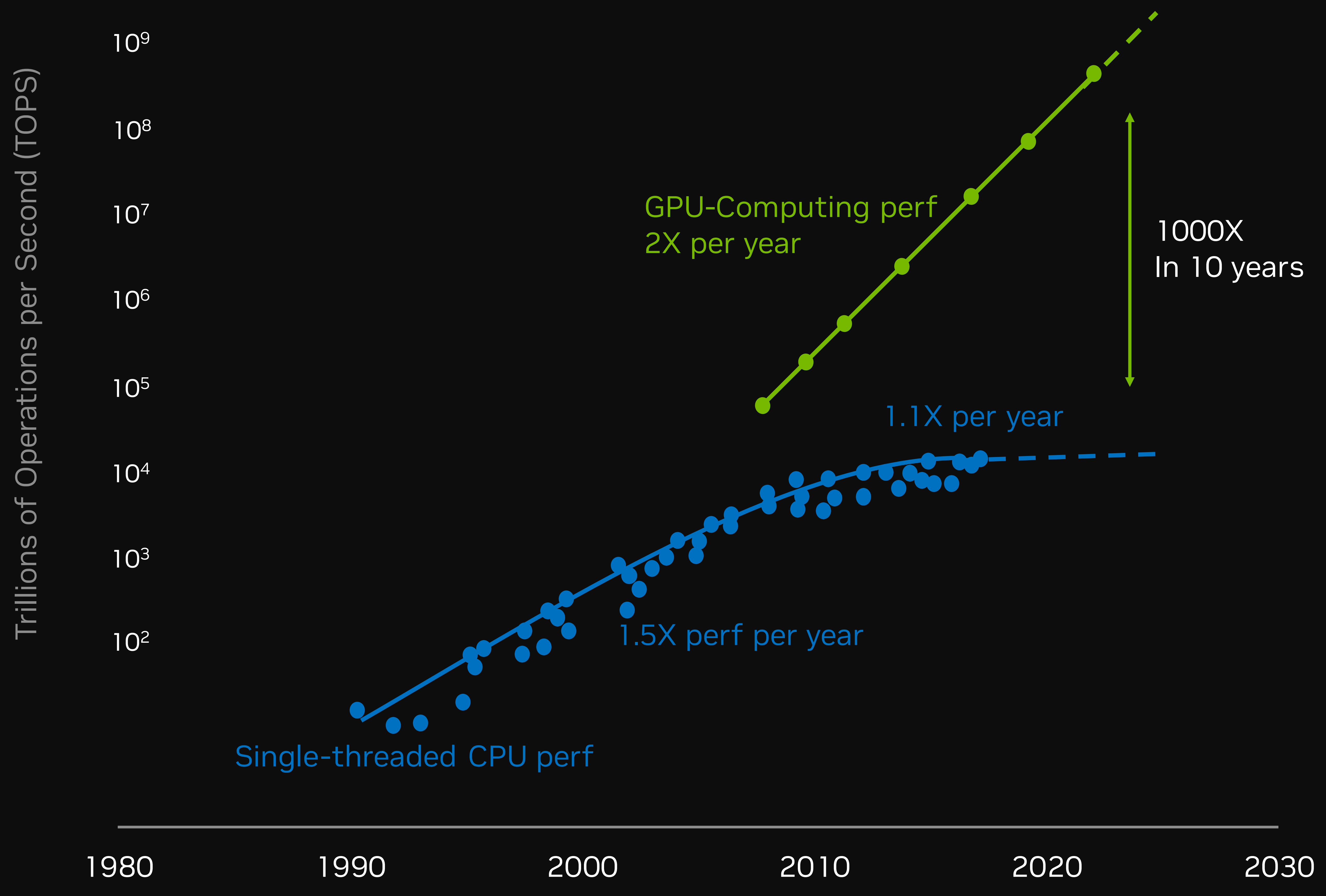
NVIDIA is uniquely dedicated to accelerated computing—working top-to-bottom, refactoring applications and creating new algorithms, and bottom-to-top—inventing new specialized processors, like RT Core and Tensor Core

"It's the end of Moore's Law as we know it."

- John Hennessy Oct 23, 2018

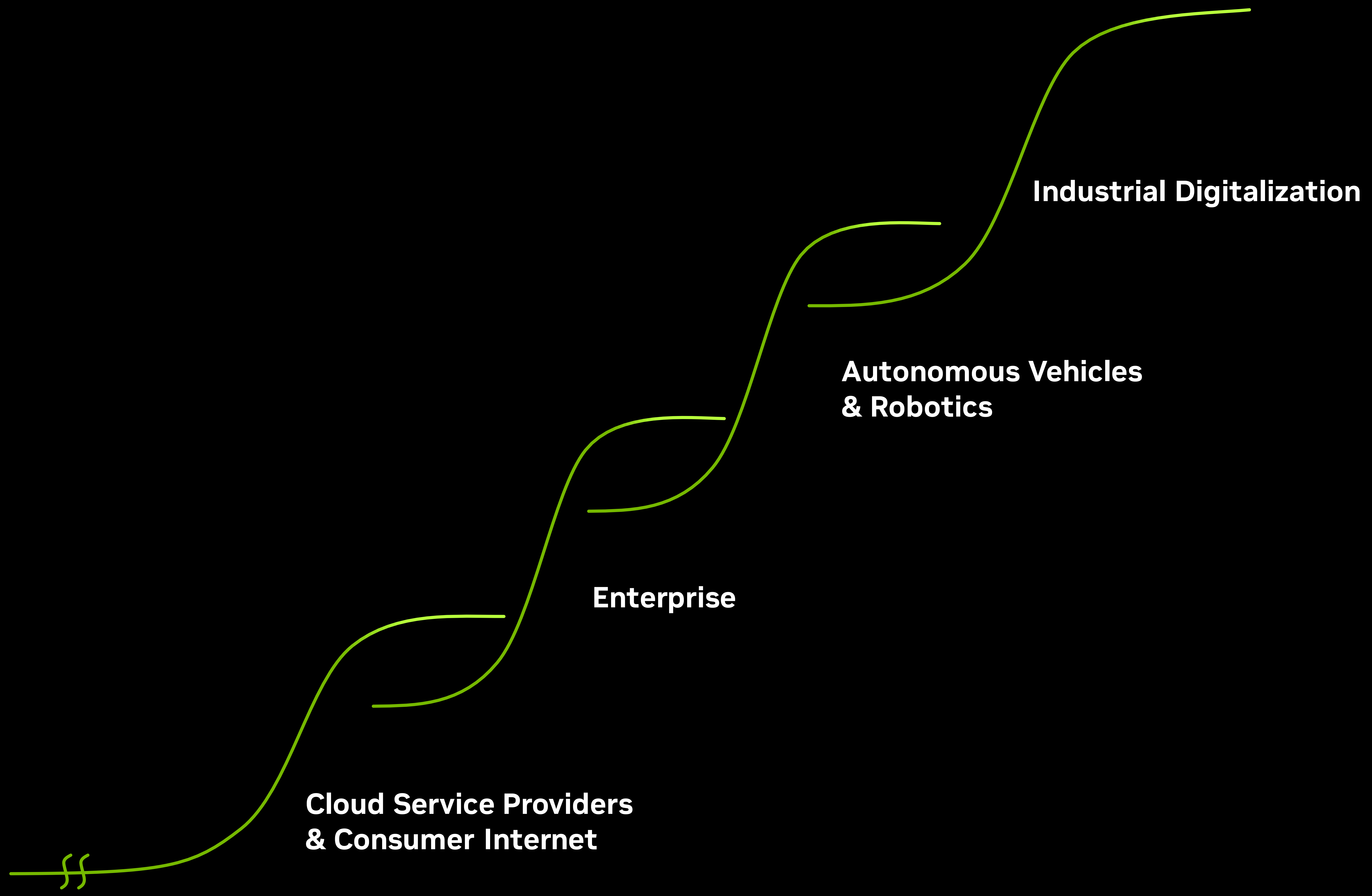
"Moore's Law is dead."

- Jensen Huang, GTC 2013



Waves of Adoption of Accelerated Computing

A generational computing platform shift


Industrial Digitalization
Autonomous Vehicles & Robotics
Enterprise
Cloud Service Providers & Consumer Internet

A new computing era has begun

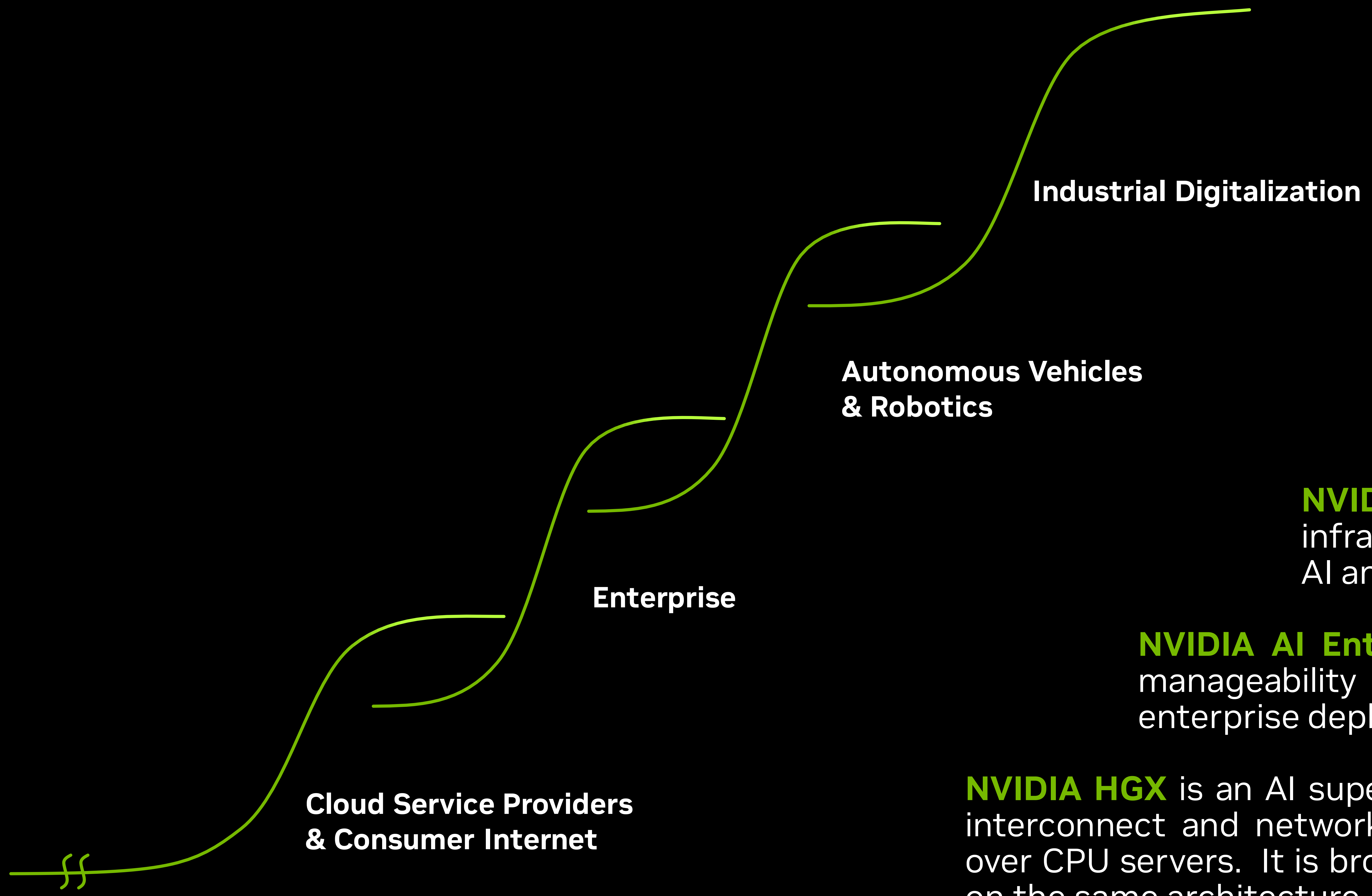
Accelerated computing enabled the rise of AI, which is driving a platform shift from general purpose to accelerated computing, and enabling new, never-before-possible applications

The trillion dollars of installed global data center infrastructure will transition to accelerated computing to achieve better performance, energy-efficiency and cost by an order of magnitude

Hyperscale cloud service providers and consumer internet companies have been the early adopters of AI and accelerated computing, with broader enterprise adoption now under way

AI and accelerated computing will also make possible the next big waves — autonomous machines and industrial digitalization

NVIDIA Accelerated Computing for Every Wave


Industrial Digitalization
Autonomous Vehicles & Robotics
Enterprise
Cloud Service Providers & Consumer Internet

NVIDIA Omniverse is a software platform for designing, building, and operating 3D and virtual world simulations. It harnesses the power of NVIDIA graphics and AI technologies and runs on NVIDIA-powered data centers and workstations

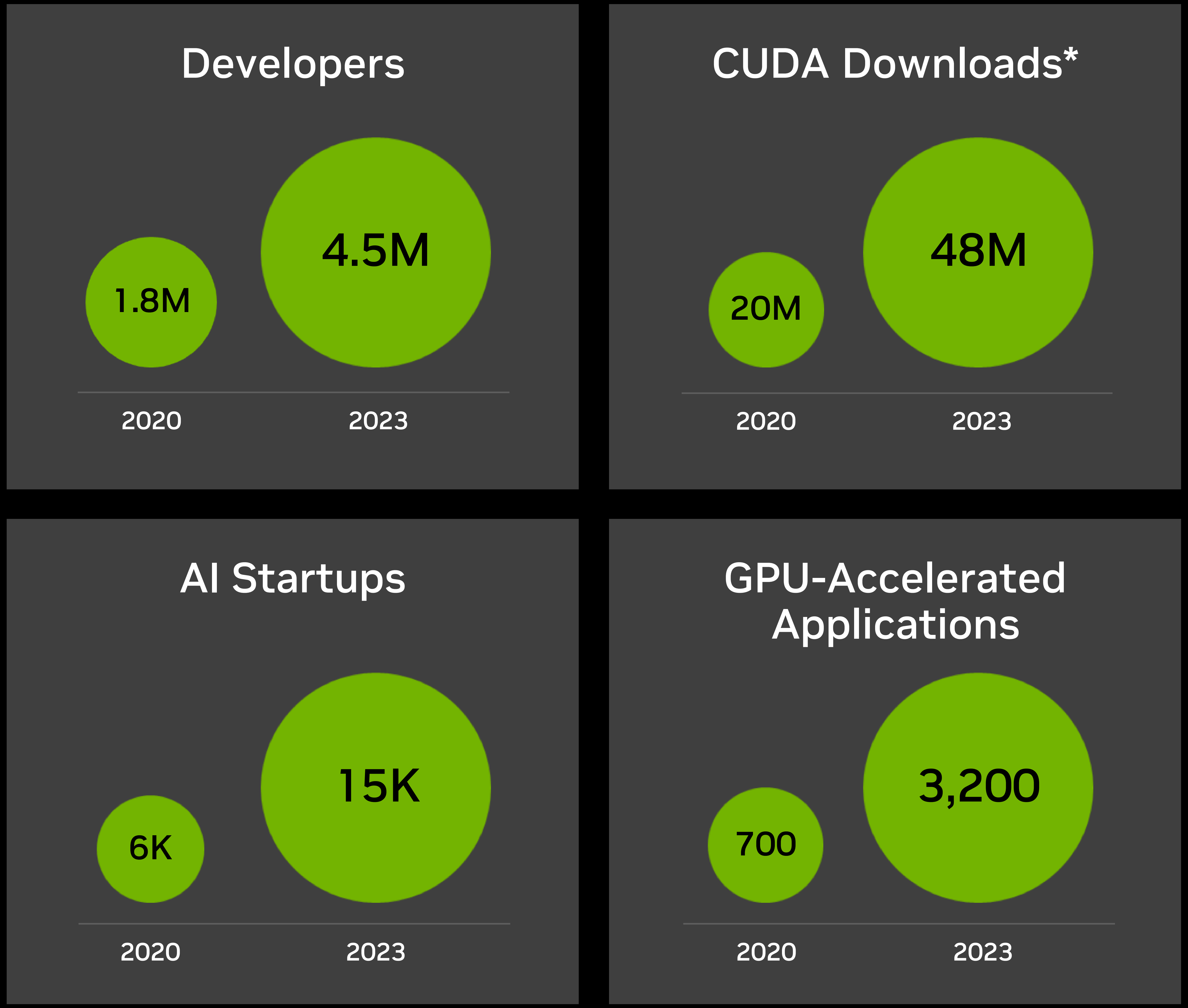
NVIDIA DRIVE is a full-stack platform for autonomous vehicles (AV) that includes hardware for in-car compute, such as the Orin system-on-chip, and the full AV and AI cockpit software stack

NVIDIA DGX Cloud is a cloud service that allows enterprises immediate access to the infrastructure and software needed to train advanced models for generative AI and other groundbreaking applications

NVIDIA AI Enterprise is the operating system of AI, with enterprise-grade security, stability, manageability and support. It is available on all major CSPs and server OEMs and supports enterprise deployment of AI in production

NVIDIA HGX is an AI supercomputing platform purpose-built for AI. It includes 8 NVIDIA GPUs, as well as interconnect and networking technologies, delivering order-of-magnitude performance speed-ups for AI over CPU servers. It is broadly available from all major server OEMs/ODMs. **NVIDIA DGX**, an AI server based on the same architecture, along with NVIDIA AI software and support, is also available

NVIDIA’s Accelerated Computing Ecosystem



Developers

CUDA Downloads*

AI Startups

GPU-Accelerated Applications

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, pre-trained AI models, SDKs and other development tools

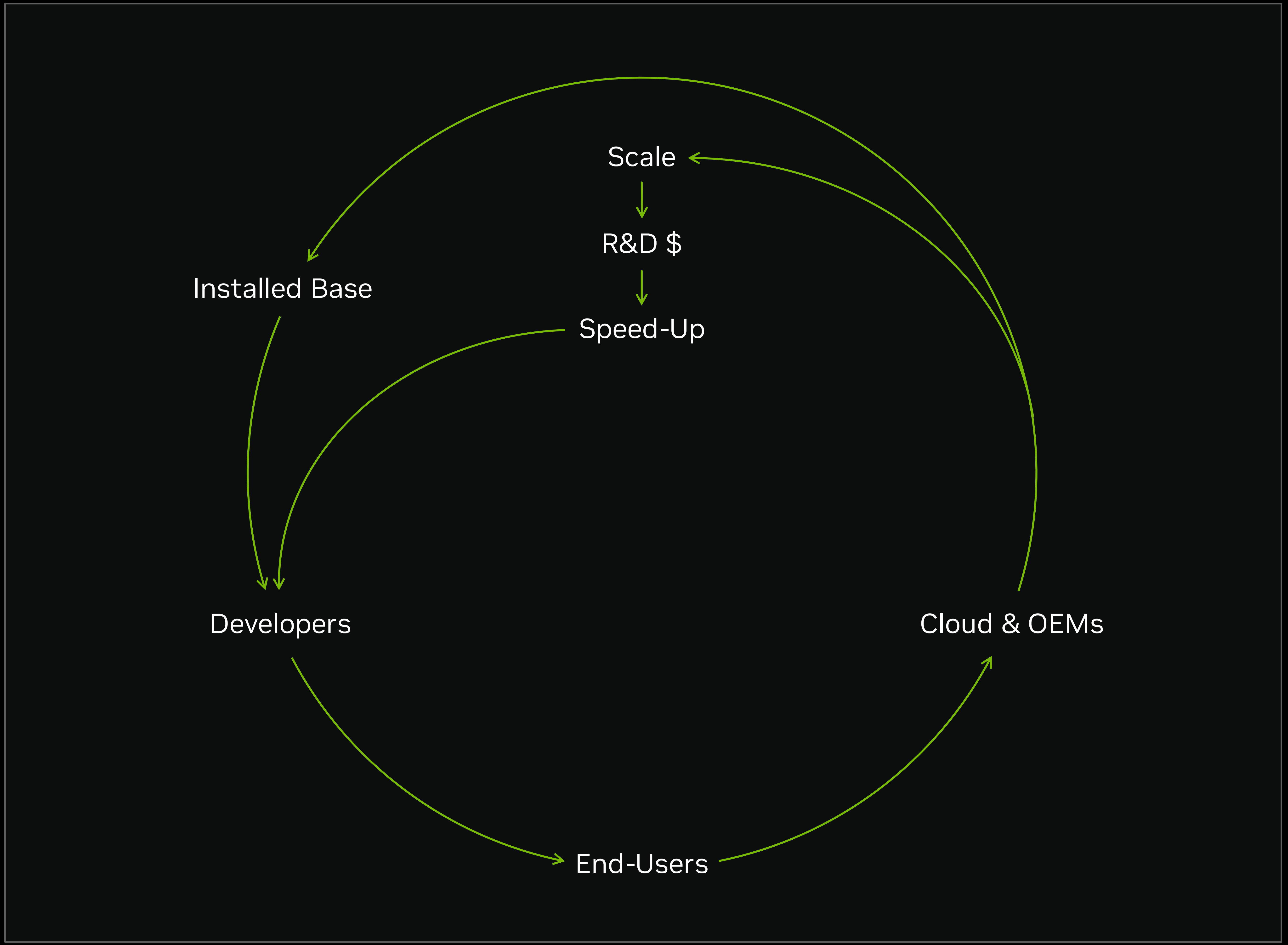
300 Libraries

600 AI Models

100 Updated in the Last Year

*Cumulative

NVIDIA's Multi-Sided Platform and Flywheel



NVIDIA Accelerated Computing Virtuous Cycle

The virtuous cycle of NVIDIA's accelerated computing starts with an installed base of several hundred million GPUs, all compatible with the CUDA programming model

- **For developers** — NVIDIA's one architecture and large installed base give developer's software the best performance and greatest reach
- **For end users** — NVIDIA is offered by virtually every computing provider and accelerates the most impactful applications from cloud to edge
- **For cloud providers and OEMs** — NVIDIA's rich suite of Acceleration Platforms lets partners build one offering to address large markets including media & entertainment, healthcare, transportation, energy, financial services, manufacturing, retail, and more
- **For NVIDIA** — Deep engagement with developers, computing providers, and customers in diverse industries enables unmatched expertise, scale, and speed of innovation across the entire accelerated computing stack — propelling the flywheel

Huge ROI from AI Driving a Powerful New Investment Cycle

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

Knowledge workers will use copilots based on large language models to generate documents, answer questions, or summarize missed meetings, emails and chats — adding hours of productivity per week

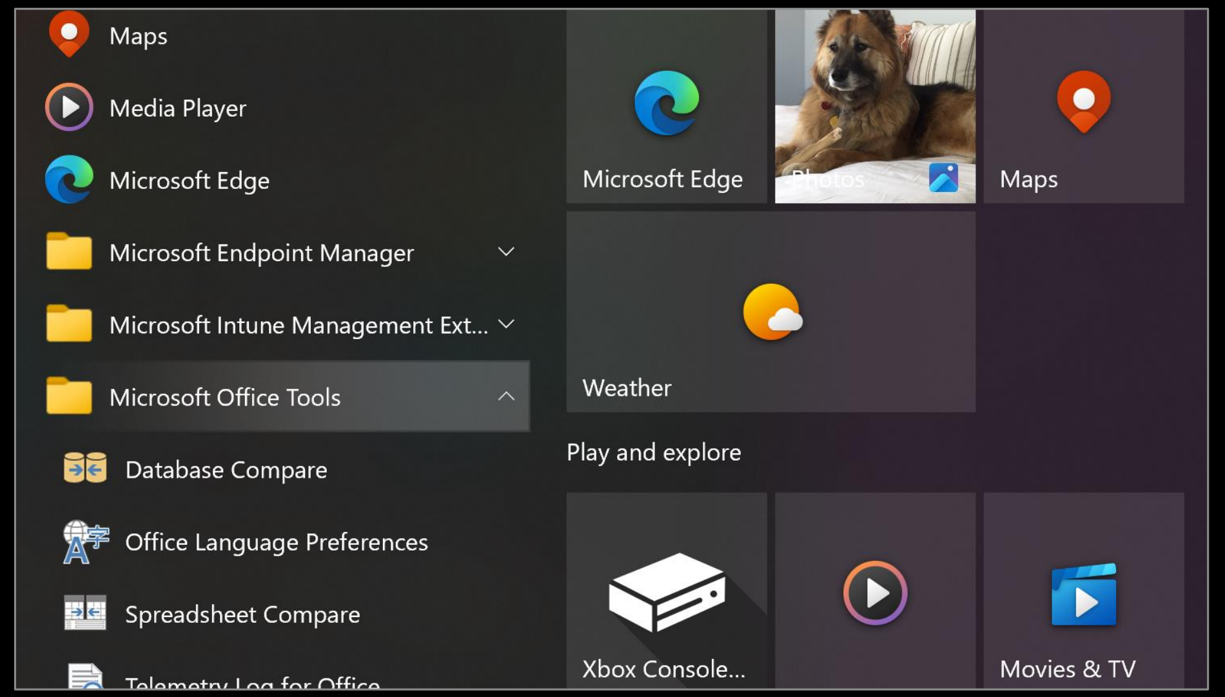
Copilots specialized for fields such as software development, legal services or education can boost productivity by as much as 50%

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

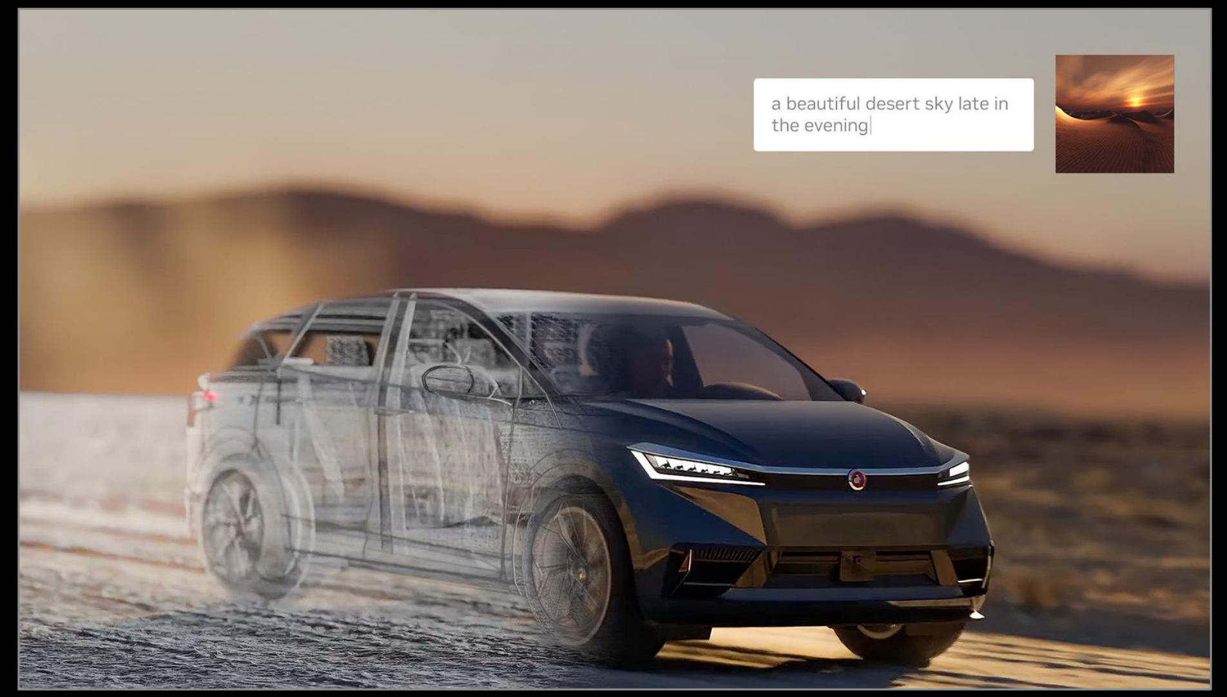
Call center agents augmented with AI chatbots can dramatically increase productivity and customer satisfaction

Drug discovery, financial services, agriculture and food services and climate forecasting are seeing order-of-magnitude workflow acceleration from AI



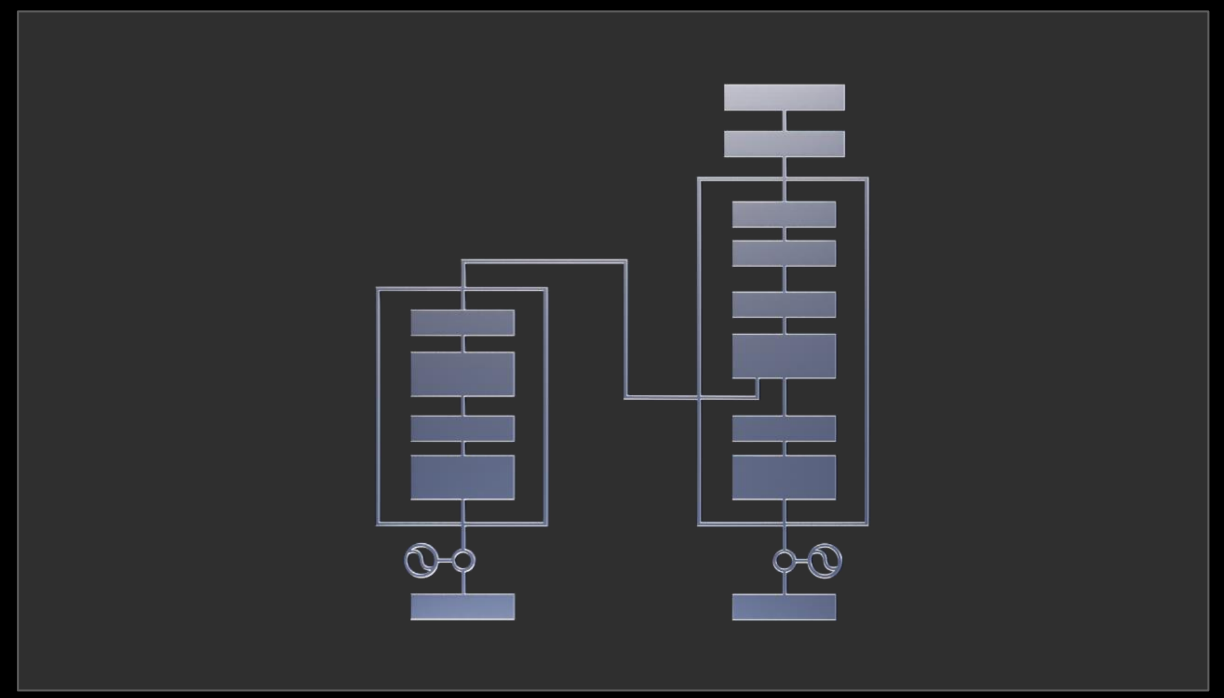
Office AI Copilots
Over 1B knowledge workers

Search & Social Media
$700B in digital advertising annually



AI Content Creation
50M creators globally

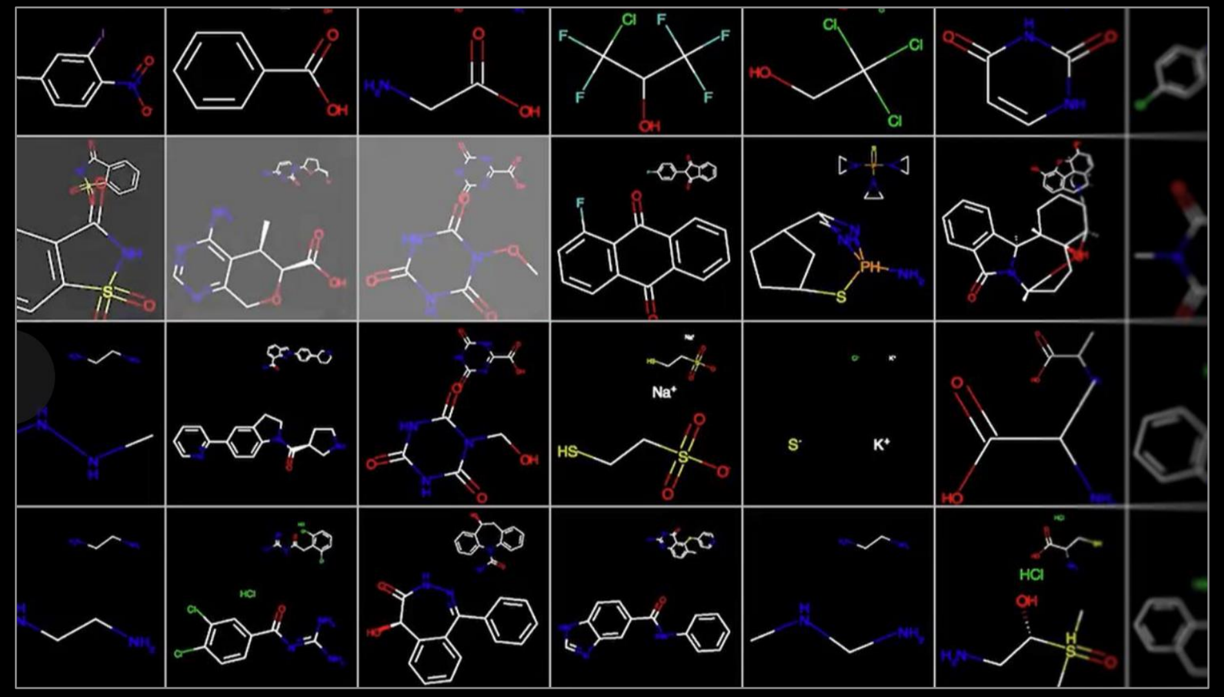
Legal Services, Education
1M legal professionals in the US
9M educators in the US



AI Software Development
30M software developers globally

Financial Services
678B annual credit card transactions

Customer Service with AI
15M call center agents globally



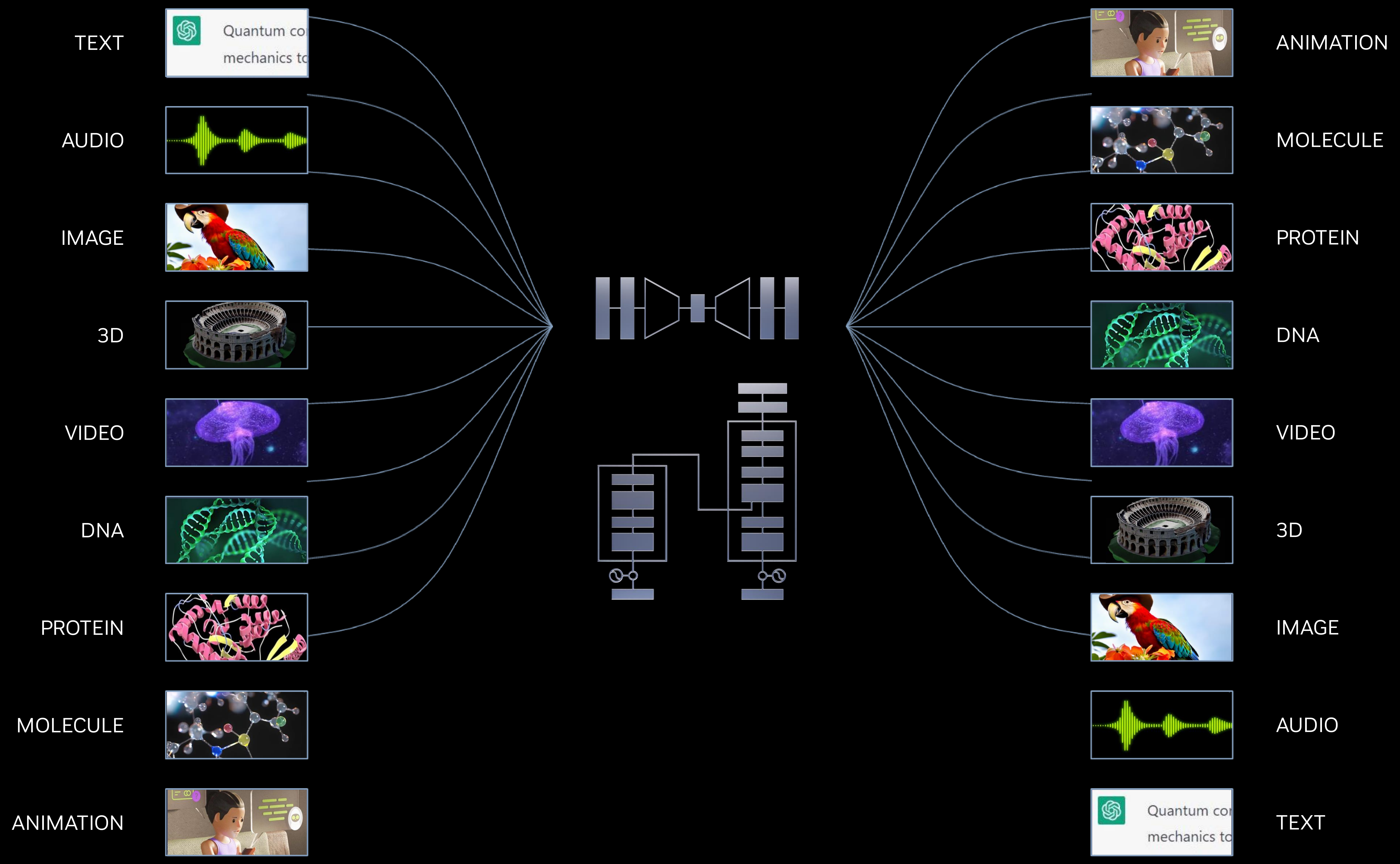
Drug Discovery
10^{18} molecules in chemical space
40 exabytes of genome data

Agri-Food | Climate
1B people in agri-food worldwide
Earth-2 for km-scale simulation

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

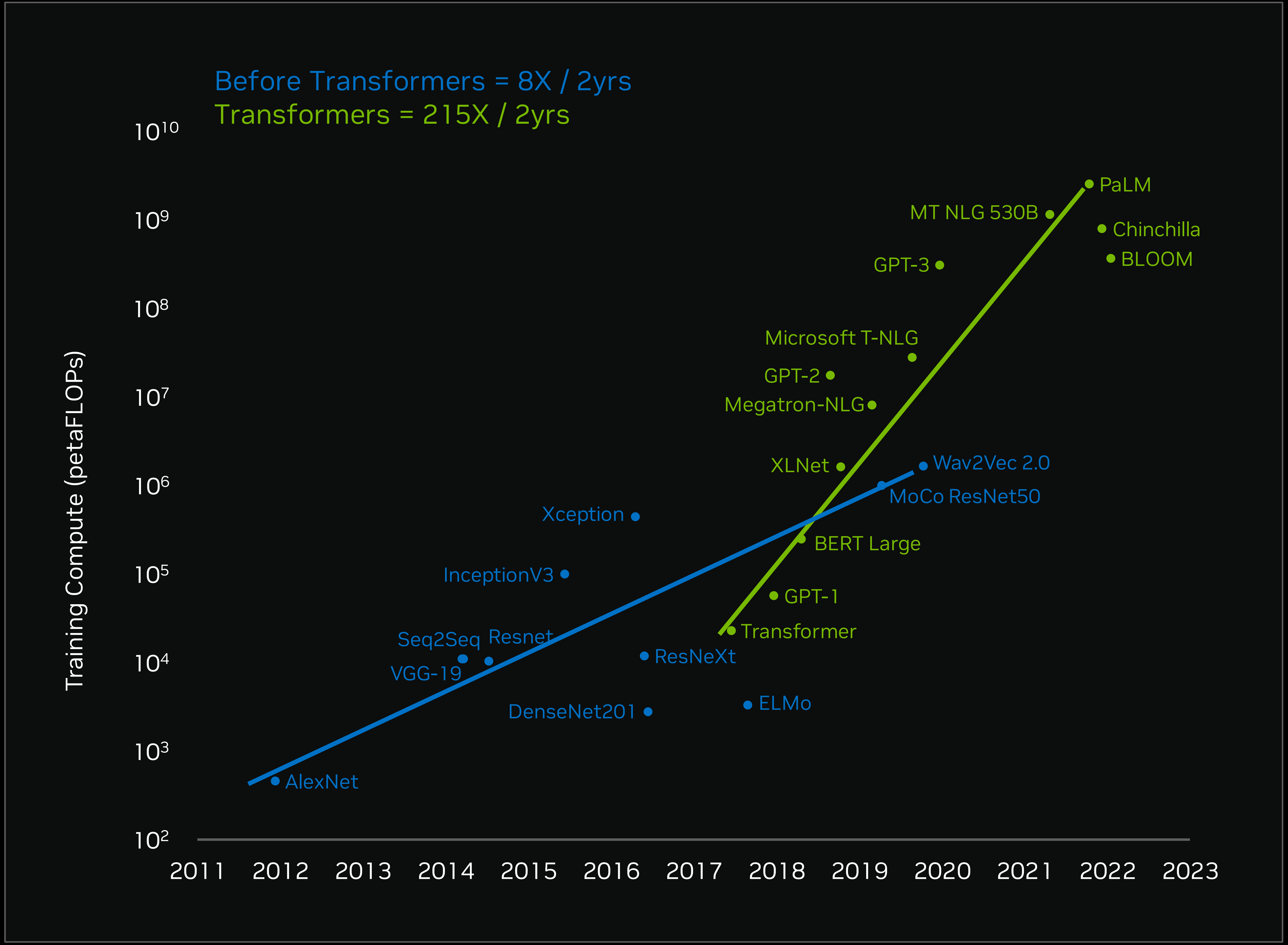
It can then be prompted to generate text, images, video, code, or even proteins

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

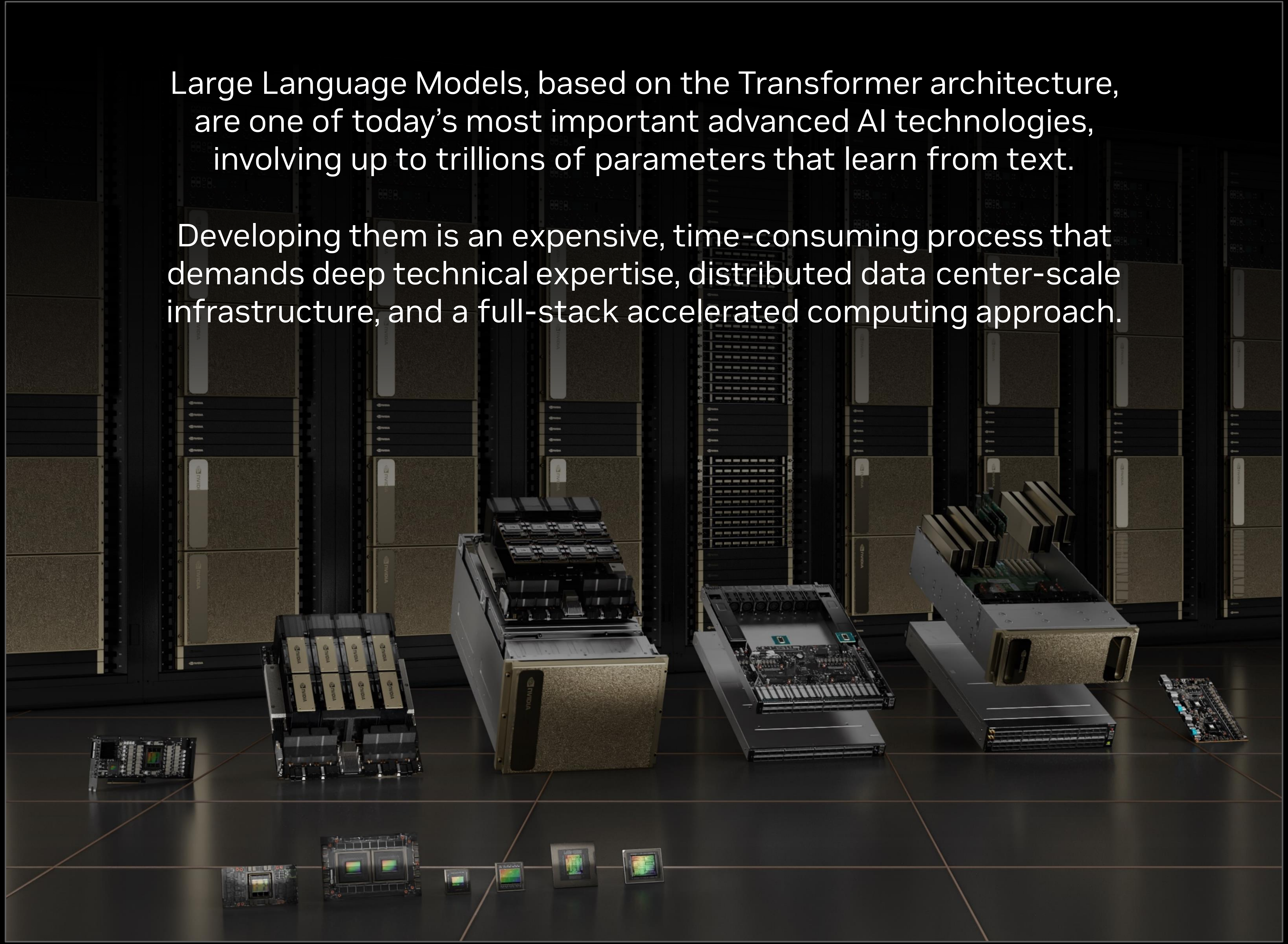
1,600+ Generative AI companies are building on NVIDIA

Modern AI is a Data Center Scale Computing Workload

Data centers are becoming AI factories: Data as input, intelligence as output



AI Training Computational Requirements



Large Language Models, based on the Transformer architecture, are one of today's most important advanced AI technologies, involving up to trillions of parameters that learn from text.

Developing them is an expensive, time-consuming process that demands deep technical expertise, distributed data center-scale infrastructure, and a full-stack accelerated computing approach.

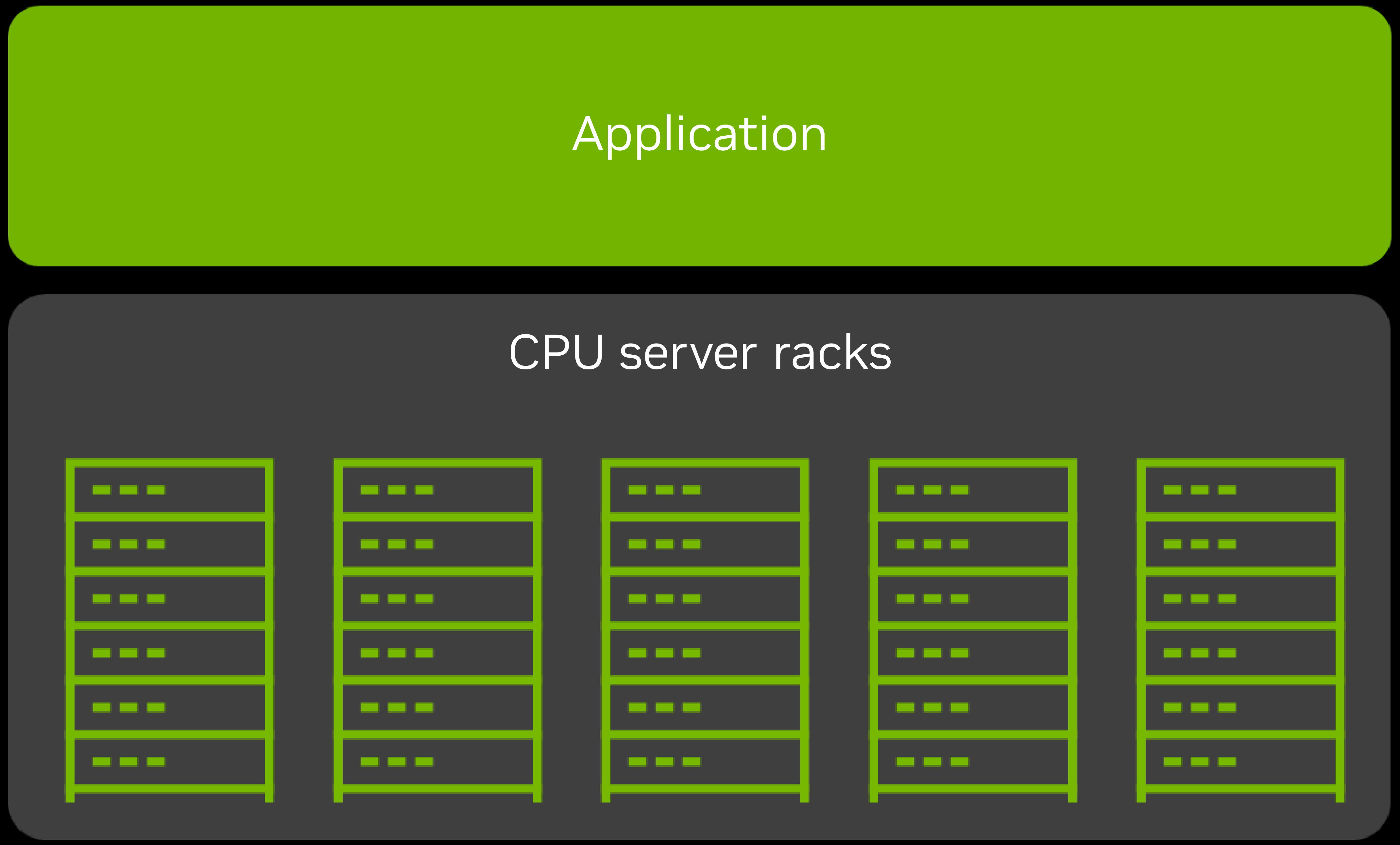
Fueling Giant-Scale AI Infrastructure

NVIDIA compute & networking GPU | DPU | CPU

Full-Stack & Data Center Scale Acceleration

Drive significant cost savings and workload scaling

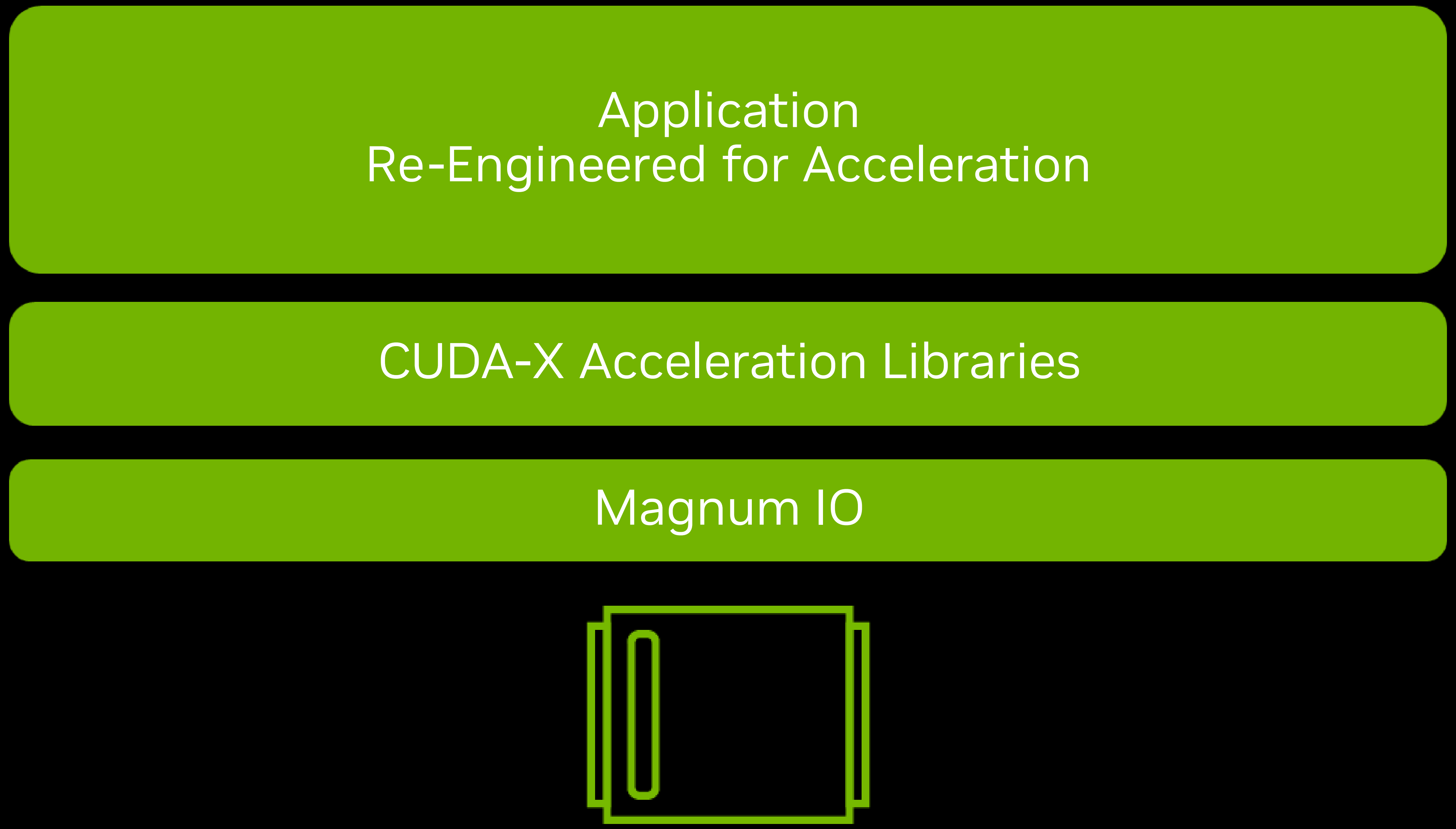
Classical Computing—960 CPU-only servers



Application

CPU server racks

Accelerated Computing—2 GPU servers



Application
Re-Engineered for Acceleration

CUDA-X Acceleration Libraries

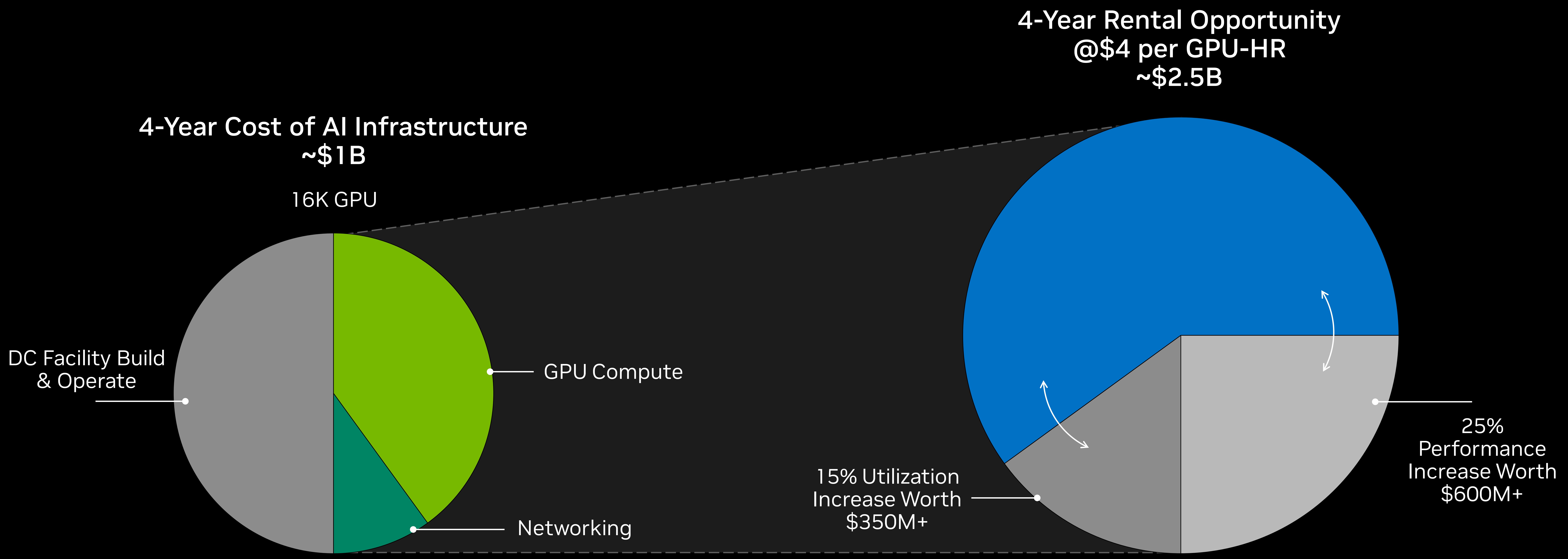
Magnum IO

25X lower cost
84X better energy-efficiency

LLM Workload: Bert-Large Training and Inference | CPU Server: Dual-EYPC 7763 | GPU Server: Dual-EPYC 7763 + 8X H100 PCIe GPUs

The High ROI of High Compute Performance

$1 upfront investment in NVIDIA compute and networking can translate to $5 in CSP revenue over 4 years



Illustrative example of NVIDIA GPU cost vs AI infrastructure total cost of ownership (TCO)

Training & Inference — One Architecture

Cloud | On-Prem | Edge

NVIDIA DGX | HGX H100
NVIDIA L40S

Training

IN THE DATA CENTER

NVIDIA L40
Image Generation

NVIDIA L4
AI Video

NVIDIA H100 | L40S
Universal GPUs

NVIDIA Grace Hopper
RecSys, Gen AI

AT THE EDGE

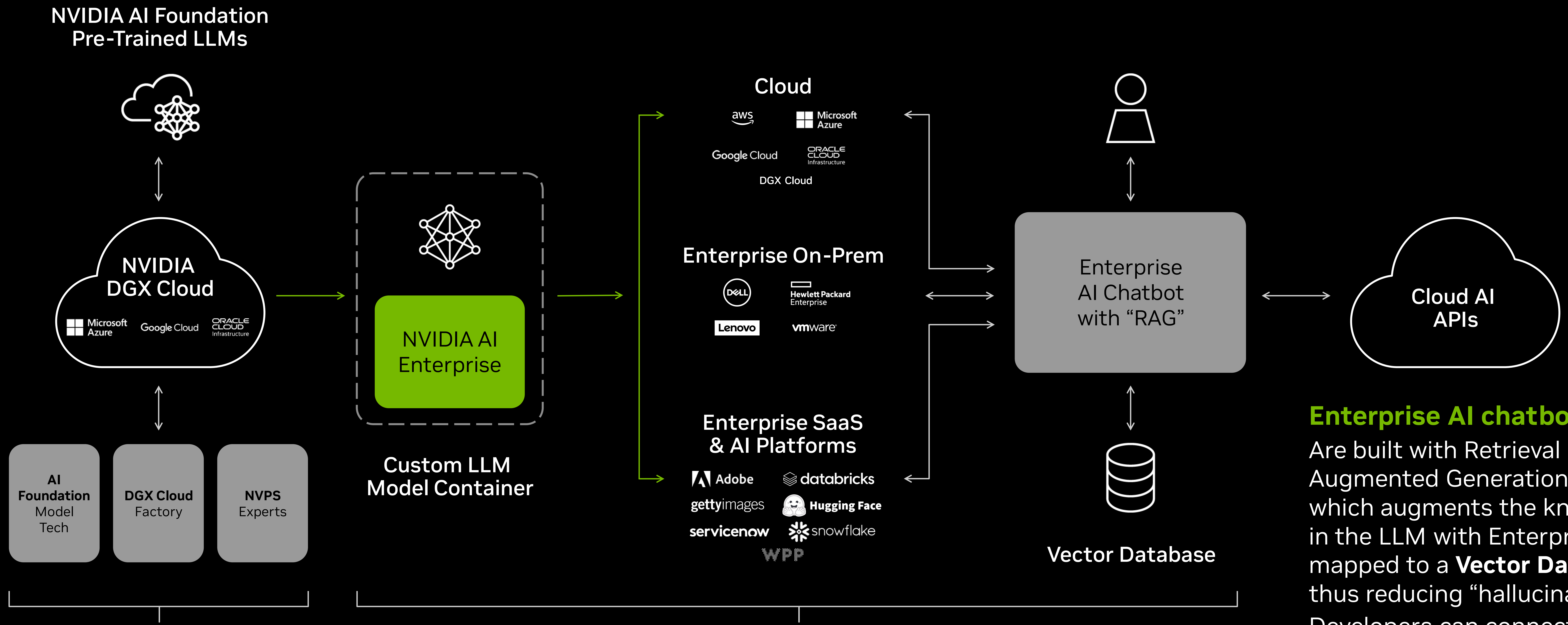
IGX
Industrial-Grade System for Healthcare, Logistics, Manufacturing

AGX
Functionally-Safe System for Autonomous Vehicles

Inference

Powering the AI Industrial Revolution

Building and Running Enterprise Gen AI Applications



NVIDIA AI Foundation
Pre-Trained LLMs

NVIDIA
DGX Cloud
Microsoft Azure
Google Cloud
ORACLE CLOUD Infrastructure

AI Foundation Model Tech

DGX Cloud Factory

NVPS Experts

NVIDIA AI Enterprise

Custom LLM
Model Container

Cloud
aws
Microsoft Azure
Google Cloud
ORACLE CLOUD Infrastructure
DGX Cloud

Enterprise On-Prem
DELL
Hewlett Packard Enterprise
Lenovo
vmware

Enterprise SaaS
& AI Platforms
Adobe
databricks
gettyimages
Hugging Face
servicenow
snowflake
WPP

Enterprise
AI Chatbot
with "RAG"

Vector Database

Cloud AI
APIs

NVIDIA AI foundry service
for *building* Enterprise AI applications

NVIDIA AI enterprise ecosystem
for *running* Enterprise AI applications

Enterprise AI chatbots

Are built with Retrieval Augmented Generation (**RAG**), which augments the knowledge in the LLM with Enterprise data mapped to a **Vector Database**, thus reducing "hallucinations".

Developers can connect additional or 3rd party services to the AI chatbot via **cloud AI APIs**.

The NVIDIA AI Foundry Model on DGX Cloud

For building enterprise AI applications

NVIDIA's "AI foundry" service leverages our AI infrastructure and expertise to build custom AI models for enterprise customers — analogous to a semiconductor foundry that uses its infrastructure and expertise to build custom chips for fabless customers.

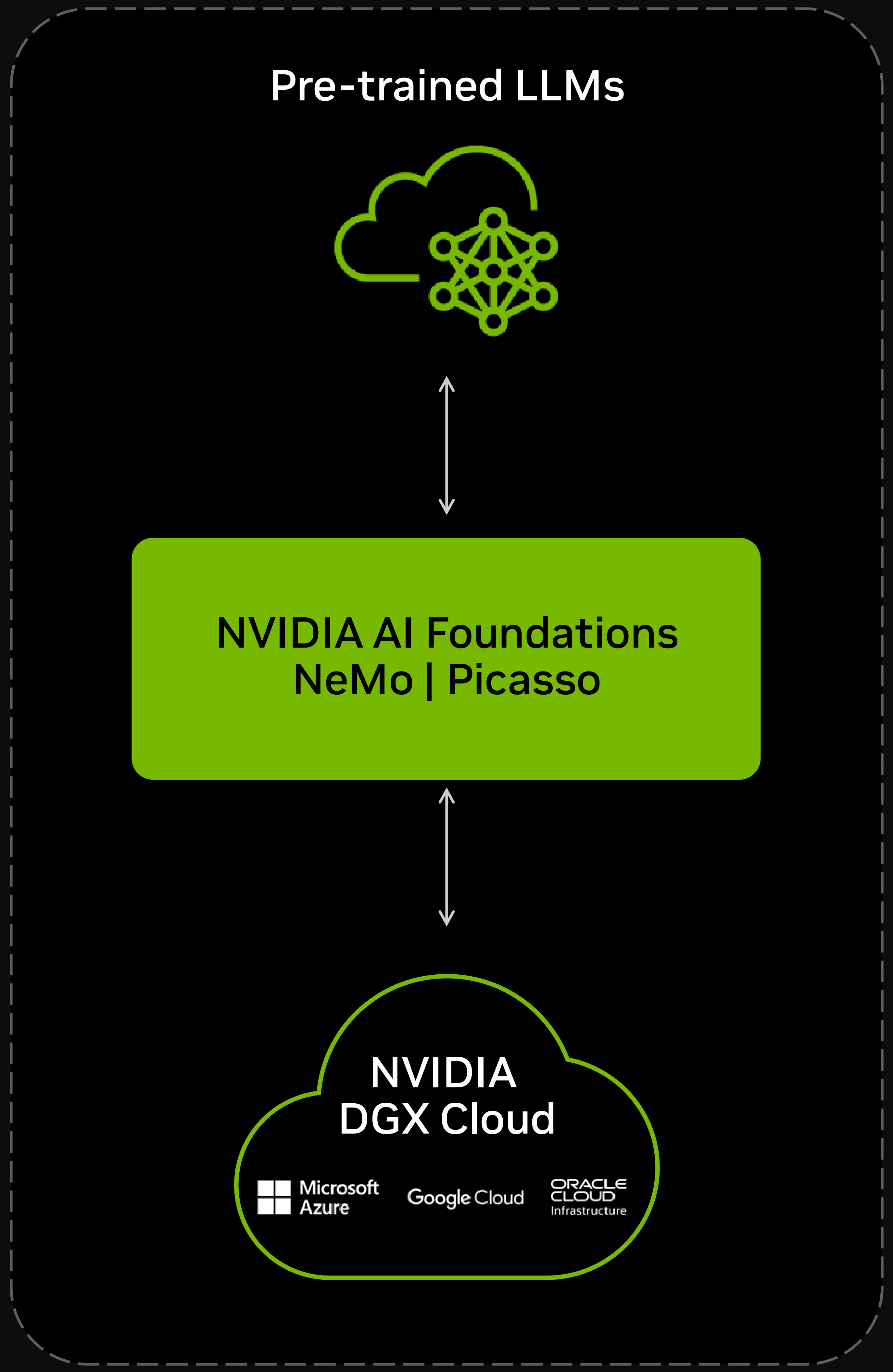
An enterprise customer starts with an NVIDIA or 3rd party pre-trained AI model, available in **NVIDIA AI Foundations**. This model making service includes frameworks such as **NVIDIA NeMo** for custom LLMs and **NVIDIA Picasso** for custom generative AI for visual design.

With help from NVIDIA experts, the enterprise customer fine-tunes the model on their proprietary enterprise data and adds guardrails, using tools available in NVIDIA AI Foundations.

The fine-tuning and optimization is done on **NVIDIA DGX Cloud**, a cloud service that allows enterprises immediate access to NVIDIA AI infrastructure and software, hosted at partner cloud providers.

The enterprise customer ends up with a fully-trained and optimized AI model, fine-tuned on their proprietary enterprise data, that can be deployed anywhere — in the cloud or on-prem.

The NVIDIA AI Foundry model generates revenue based on per-node, per-month consumption of NVIDIA DGX Cloud.



NVIDIA AI foundry

AI Factories — A New Class of Data Centers

For running enterprise AI applications

“AI factories” are a new class of data centers specially built for processing, refining and transforming vast amounts of data into valuable AI models and tokens.

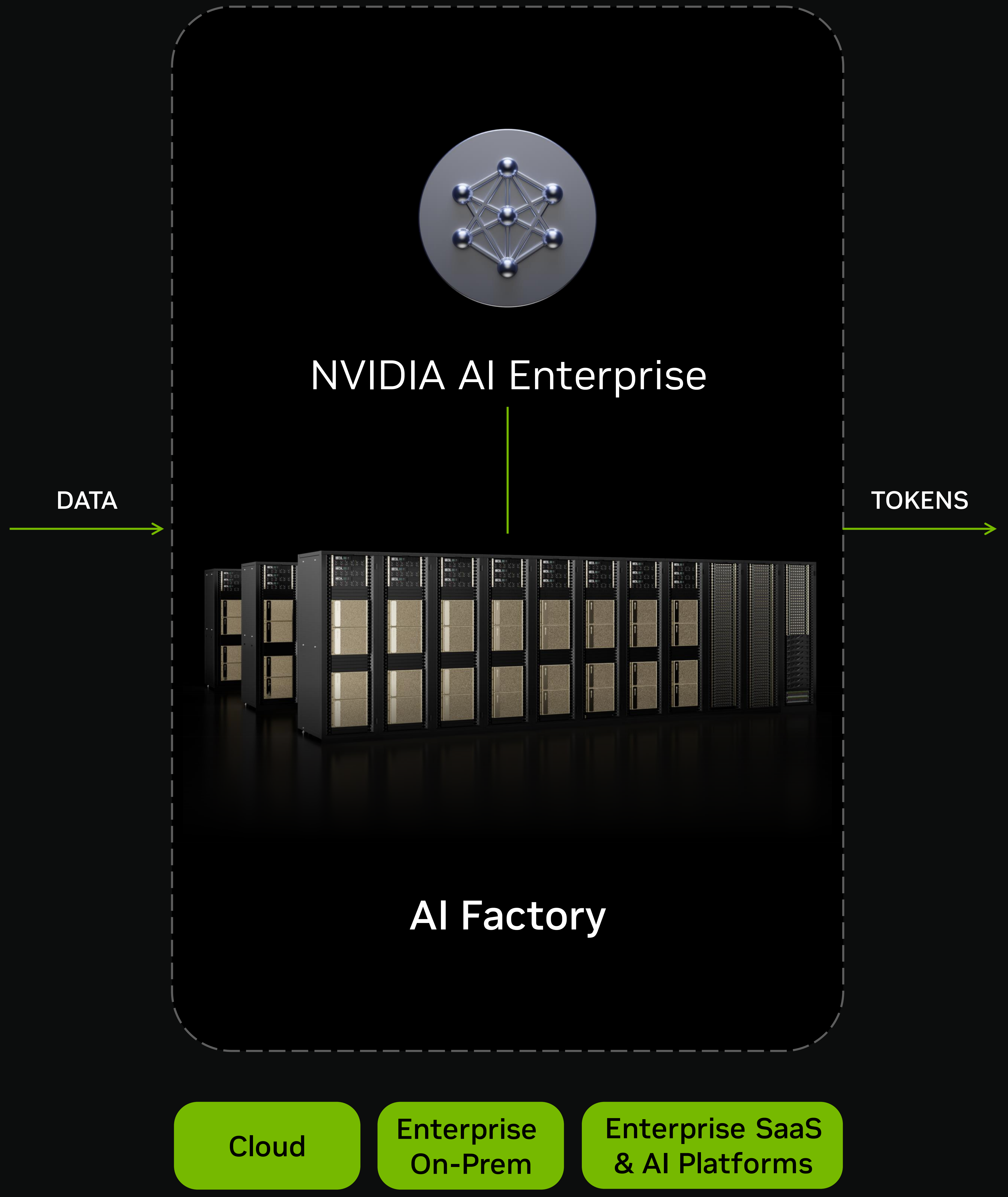
Unlike traditional data centers built for IT workloads, AI factories are built to deliver automated, professional skills.

AI factories are not multi-workload or multi-tenant. They run one workload – an AI model – and have just one customer or owner — analogous to a traditional factory.

AI factories can be built on-prem, in the cloud, or in the data centers of SaaS and AI platform vendors.

We believe that in the future, every important company will run its own AI factories in order to securely process its valuable proprietary data and turn it into monetizable tokens, encapsulating its knowledge, intelligence, and creativity.

In addition to the up-front revenue opportunity from data center systems, **NVIDIA can generate recurring revenue from AI factories for their use of NVIDIA AI Enterprise**, the operating system for enterprise AI.



NVIDIA AI Enterprise

The operating system for enterprise AI

NVIDIA AI Enterprise

NVIDIA AI Enterprise is software for deploying and running AI with enterprise-grade security, API stability, manageability and support.

Cloud-native and available in every major cloud marketplace.

Certified to run on servers and workstations from all major OEMs.

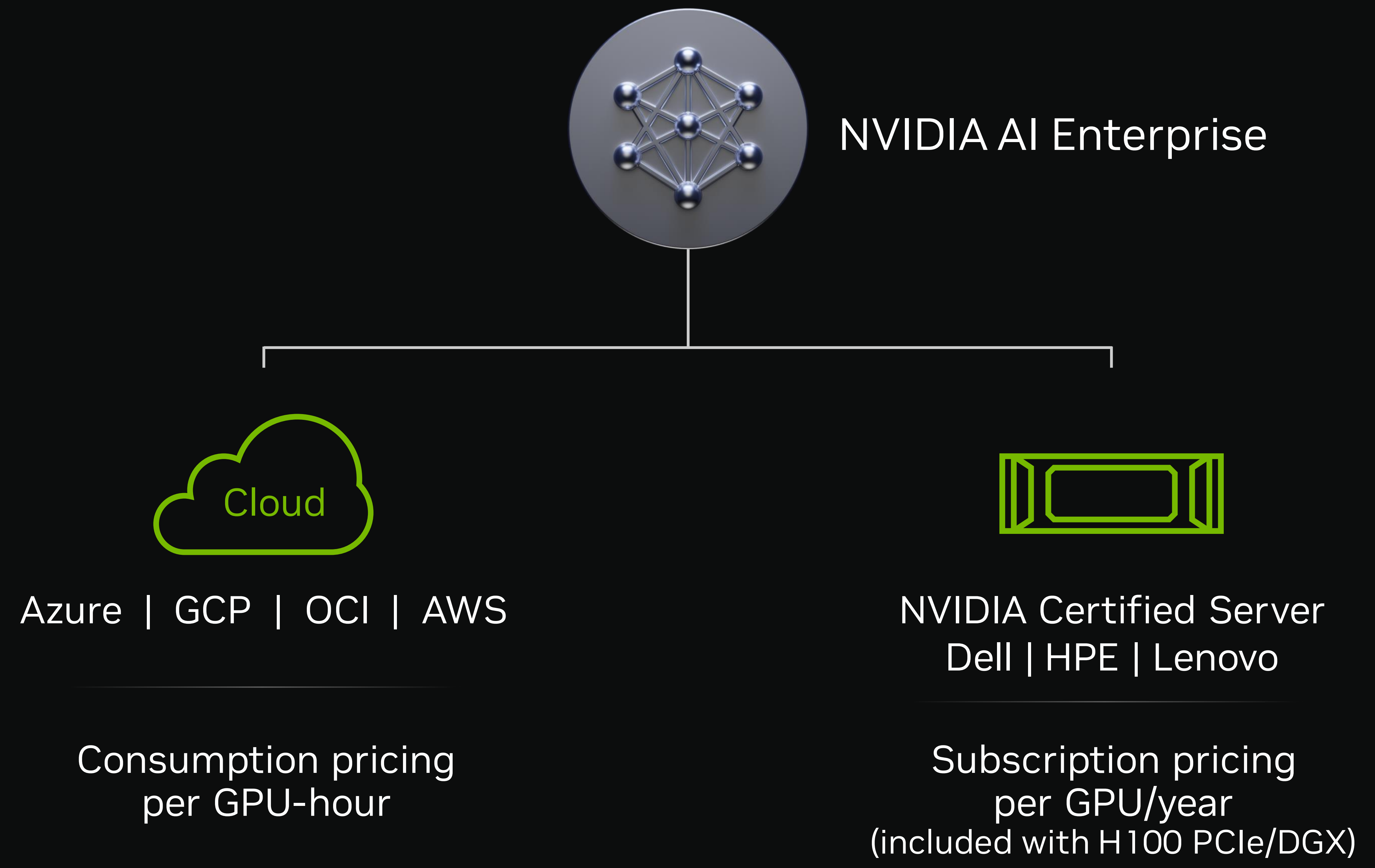
Supported by all major global system integrators.

Integrated with and distributed by VMware.

AI Use Cases and Workflows

- LLM
- Speech AI
- Recommenders
- Cybersecurity
- Medical Imaging
- Video Analytics
- Route Optimization
- More

Run Anywhere



NVIDIA AI Enterprise

Azure | GCP | OCI | AWS

Consumption pricing
per GPU-hour

NVIDIA Certified Server
Dell | HPE | Lenovo

Subscription pricing
per GPU/year
(included with H100 PCIe/DGX)

NVIDIA AI Enterprise

Broad and deep ecosystem and distribution to reach every enterprise

GSI & Service Delivery

accenture

Booz | Allen | Hamilton

Capgemini

Deloitte.

Infosys

tcs TATA CONSULTANCY SERVICES

wipro

AI Platforms

databricks

Hugging Face

snowflake

Software Platforms

gettyimages

servicenow

shutterstock

Adobe

WPP

Public Cloud Marketplaces

aws

Google Cloud

Microsoft Azure

ORACLE Cloud Infrastructure

Private Cloud

vmware

Server OEMs

BOXX

cisco

DELL Technologies

HPE GreenLake

hp

Lenovo

SUPERMICRO

NVIDIA Go-to-Market Across Cloud and On-Premises

Reaching customers everywhere

NVIDIA

DGX Cloud

Google Cloud · Microsoft Azure · ORACLE CLOUD Infrastructure

NVIDIA AI Foundations - Cloud services for customizing and operating generative AI models

DGX

Partners

aws · Google Cloud · Microsoft Azure · ORACLE CLOUD Infrastructure

HGX · INFERENCE

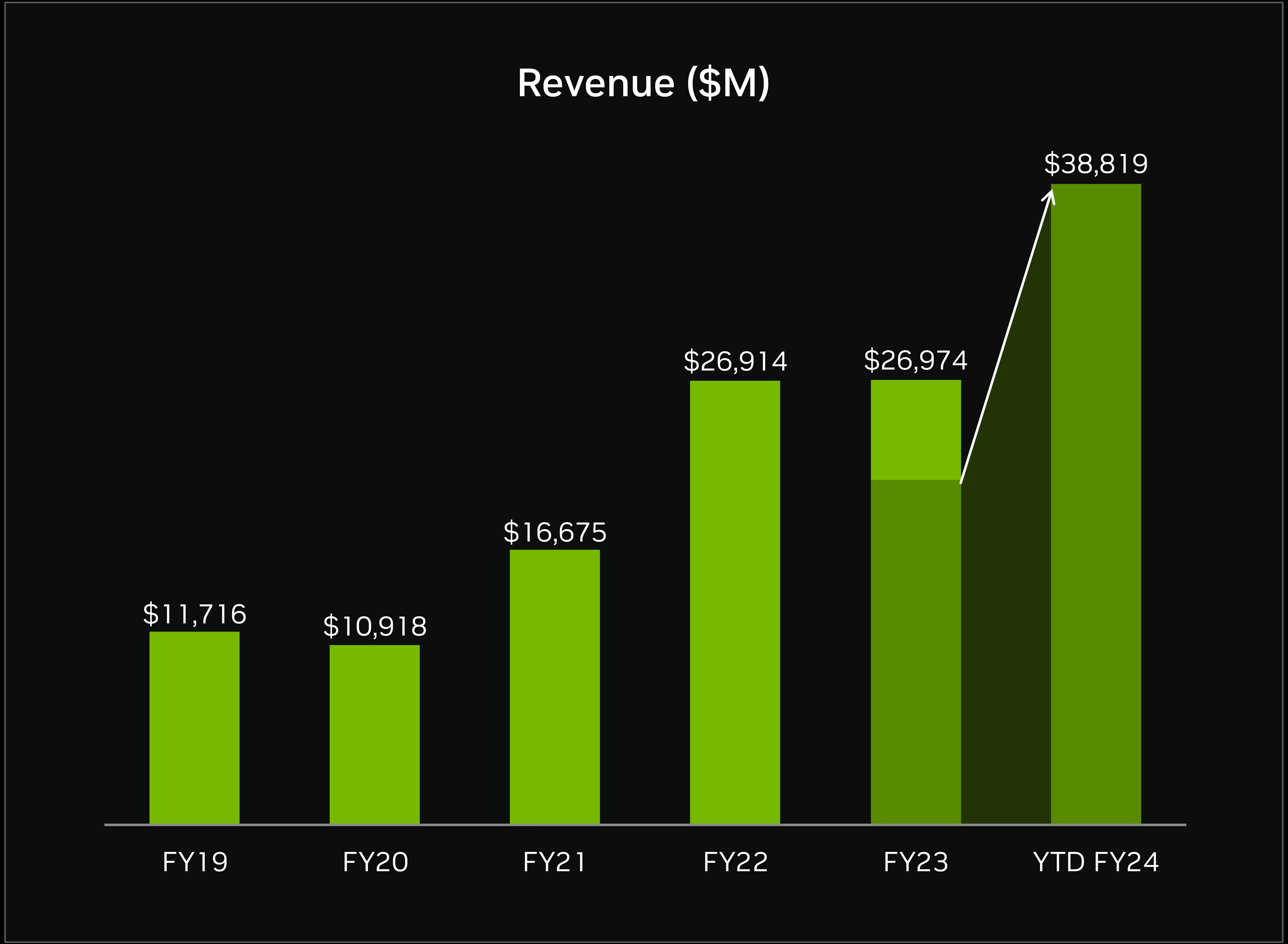
DELL Technologies · Hewlett Packard Enterprise · Lenovo

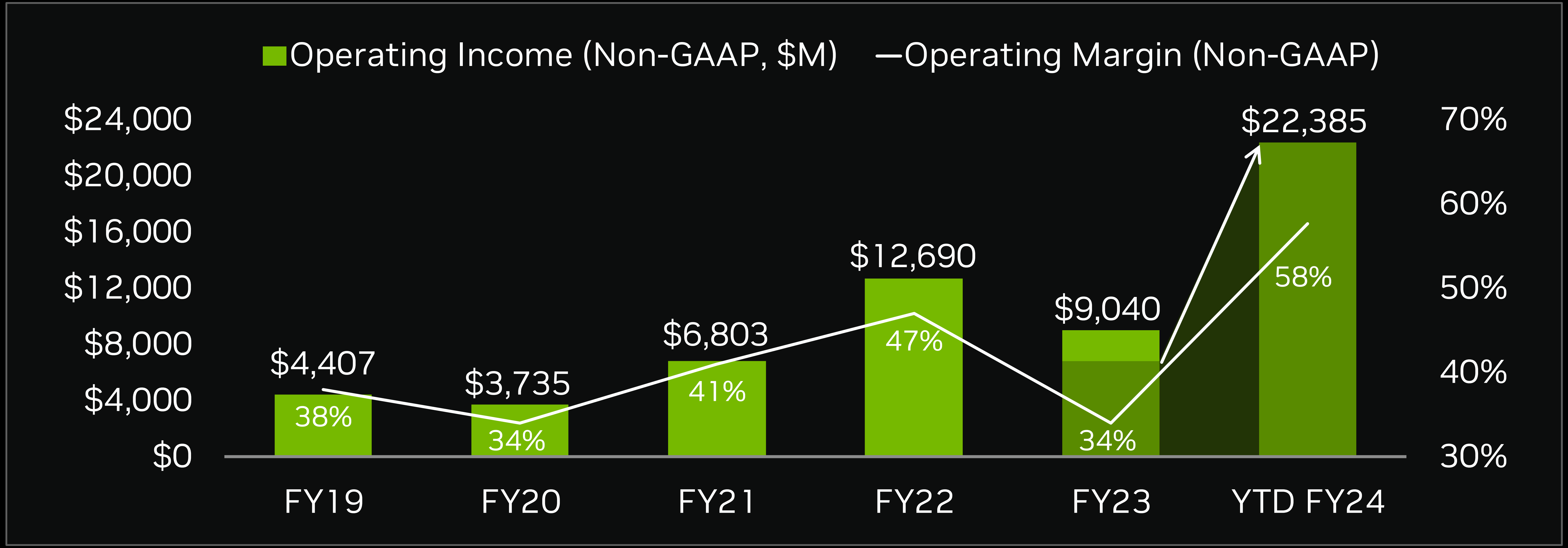
MGX · AGX · IGX

Cloud

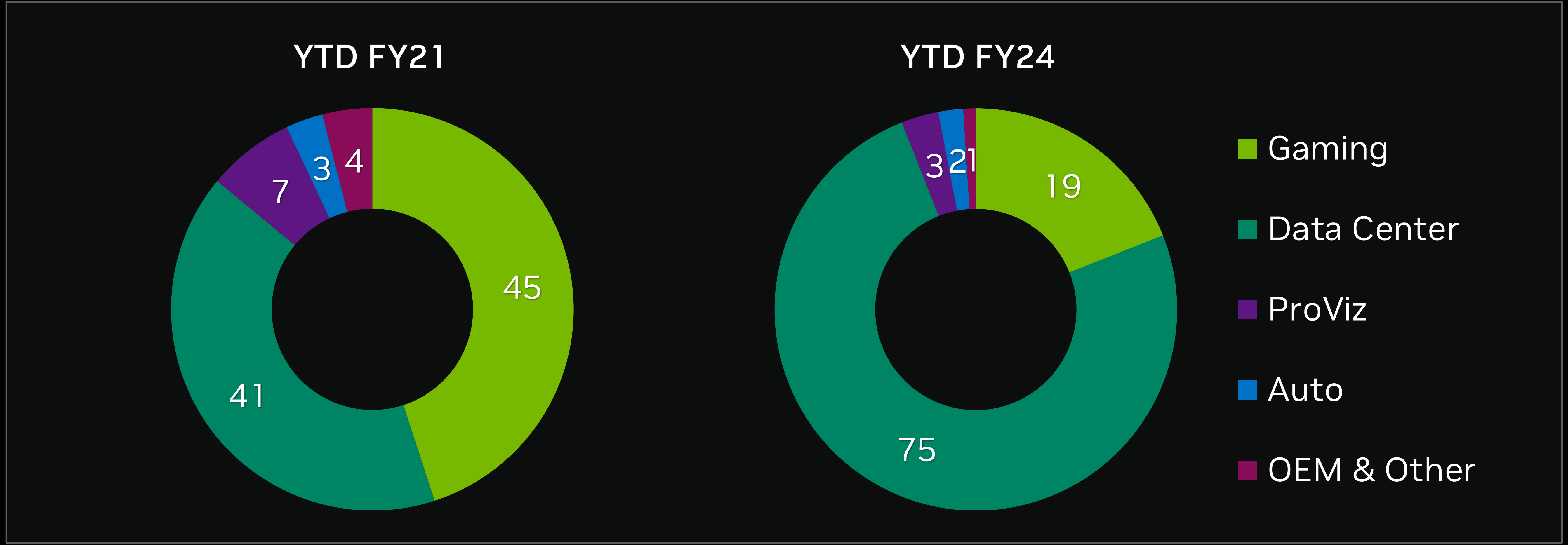
On-Prem

Driving Strong & Profitable Growth





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



FY23 financial metrics reflect a $2.2B charge for inventory and related reserves primarily related to Data Center and Gaming.

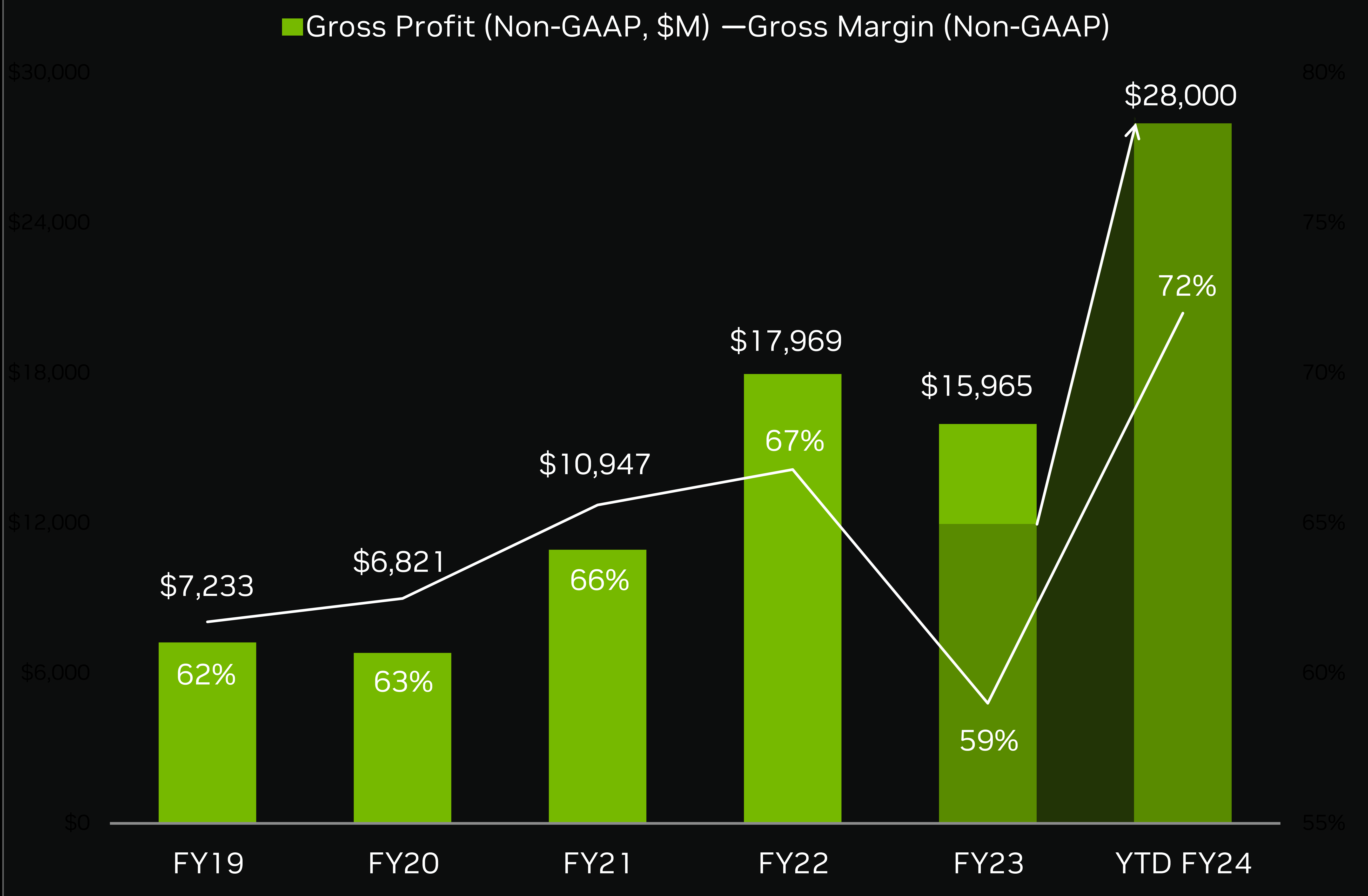
NVIDIA Gross Margins Reflect Value of Acceleration

Accelerated computing requires full-stack and data center-scale innovation across silicon, systems, algorithms and applications.

Significant expertise and effort are required, but application speed-ups can be incredible, resulting in dramatic cost and time-to-solution savings.

> For example, 2 NVIDIA HGX nodes with 16 NVIDIA H100 GPUs that cost $400K can replace 960 nodes of CPU servers that cost $10M for the same LLM workload.

NVIDIA chips carry the value of the full-stack, not just the chip.

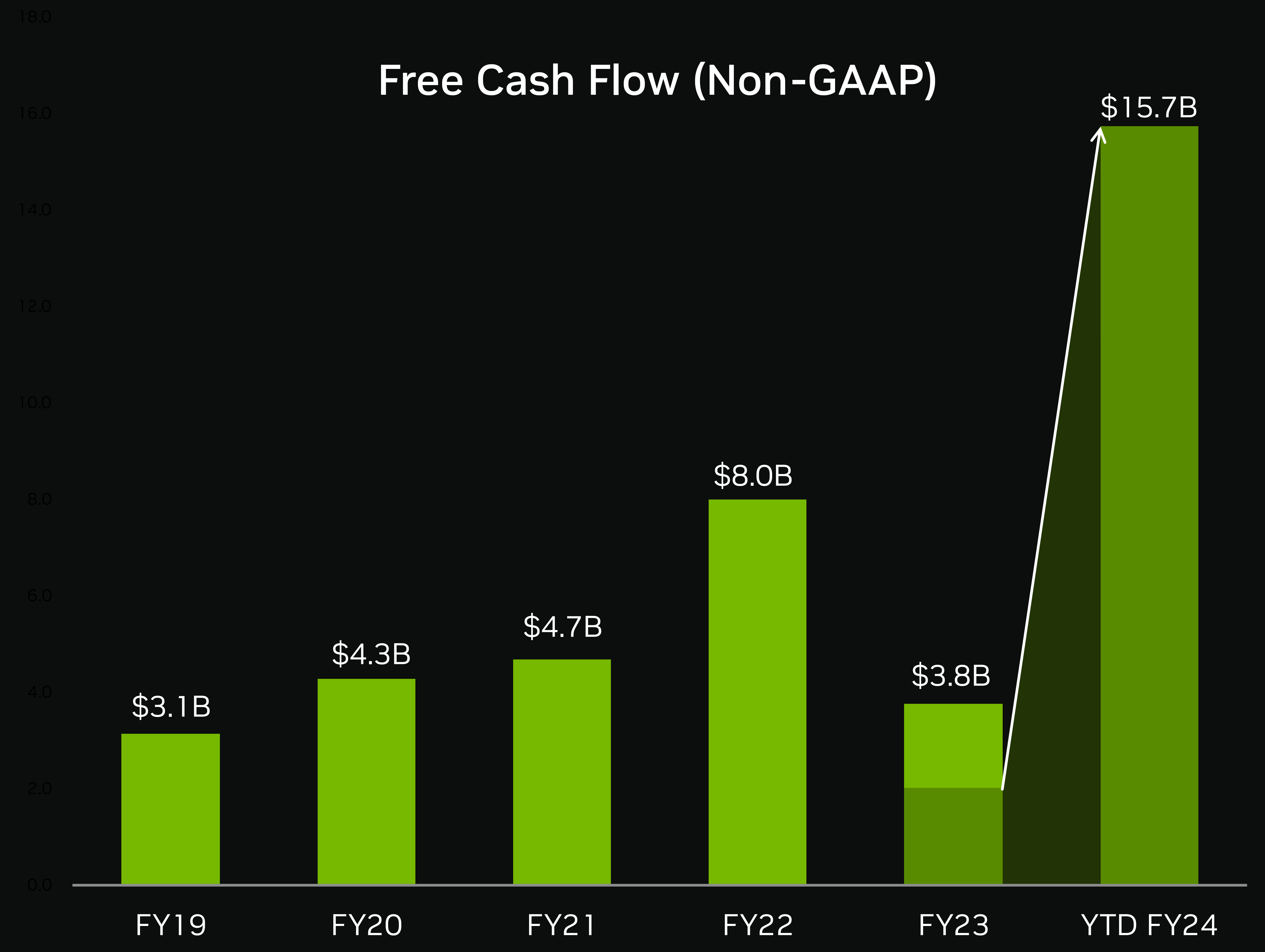


Cost comparison example based on latest available NVIDIA A100 GPU and Intel CPU inference results in the commercially available category of the MLPerf industry benchmark; includes related infrastructure costs such as networking.

FY23 financial metrics reflect a $2.2B charge for inventory and related reserves primarily related to Data Center and Gaming. Fiscal year ends in January. Refer to Appendix for reconciliation of Non-GAAP measures. Gross margins are rounded to the nearest percent.

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

$10B repurchased in FY23

$25.2B Remaining Authorization as of end of Q3

Dividend

$398M in FY 2023

Plan to Maintain[1]

Strategic Investments

Growing Our Talent

Platform Reach & Ecosystem

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

Our Market Platforms at a Glance

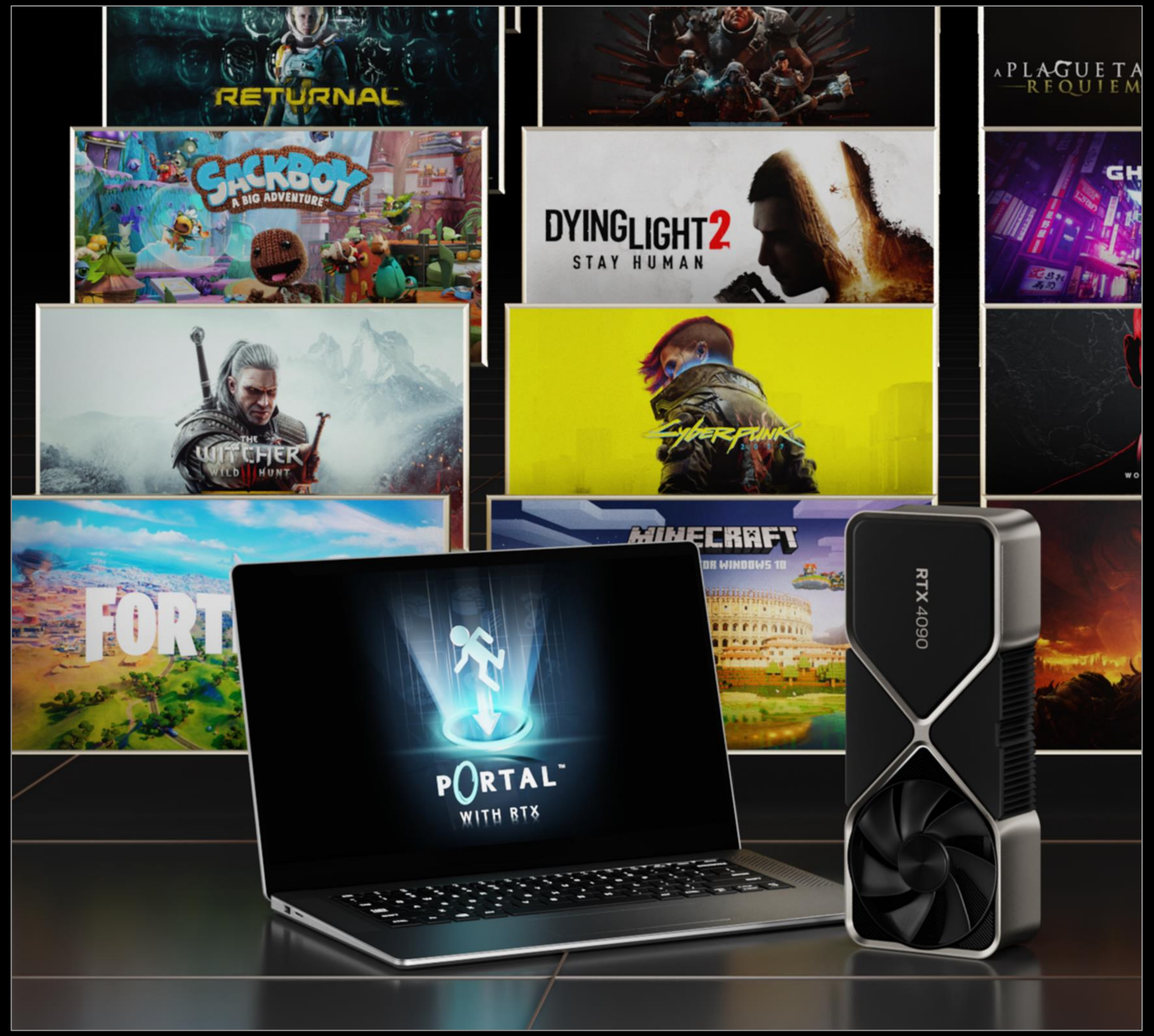
Data Center

56% of FY23 Revenue

FY23 Revenue $15.0B
5-YR CAGR 51%

DGX/HGX/MGX/IGX systems

GPU | CPU | DPU | Networking
NVIDIA AI software



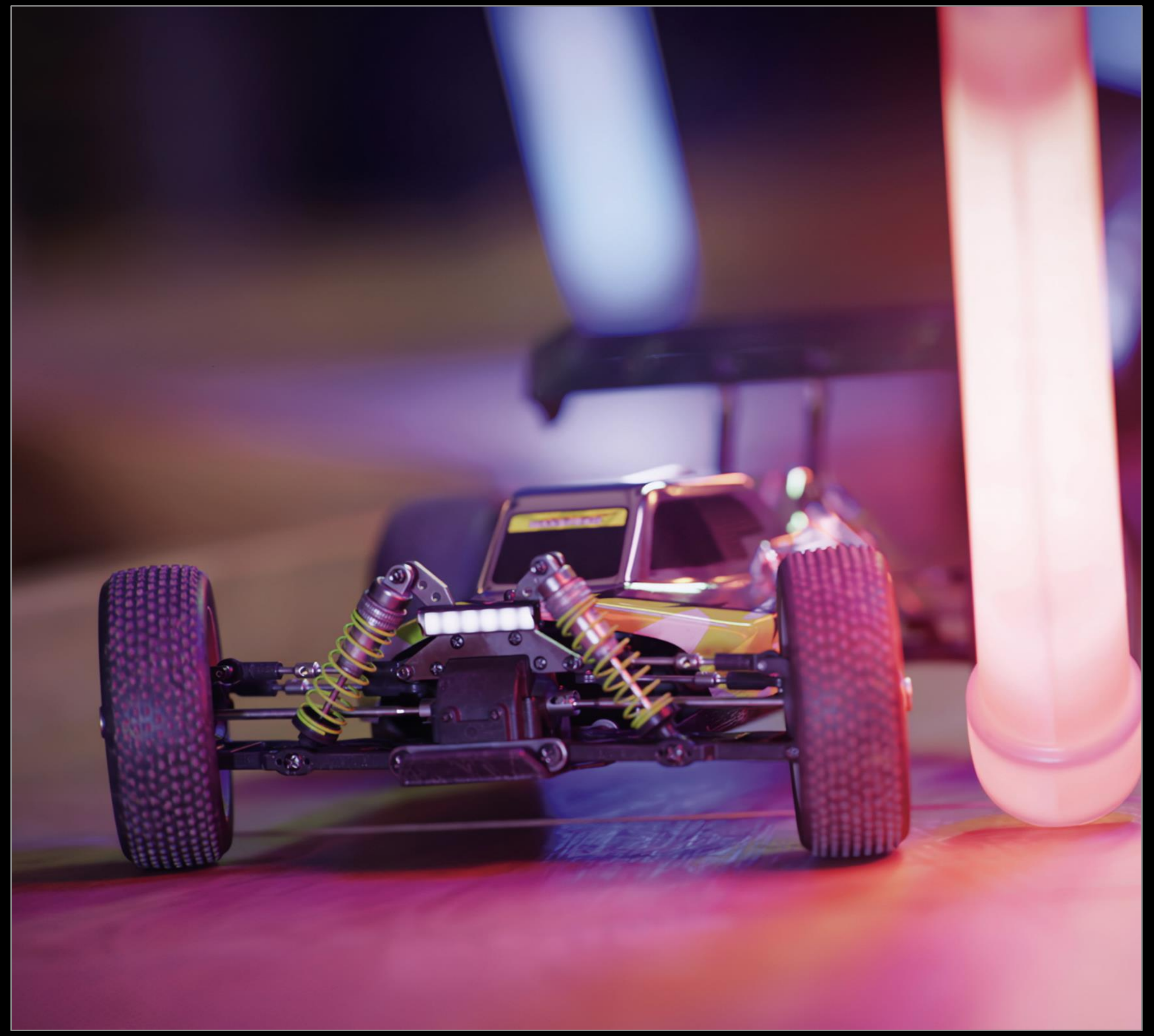
Gaming

33% of FY23 Revenue

FY23 Revenue $9.1B
5-YR CAGR 10%

GeForce GPUs for PC gaming

GeForce NOW cloud gaming



Professional Visualization

6% of FY23 Revenue

FY23 Revenue $1.5B
5-YR CAGR 11%

NVIDIA RTX GPUs
for workstations

Omniverse software

Automotive

3% of FY23 Revenue

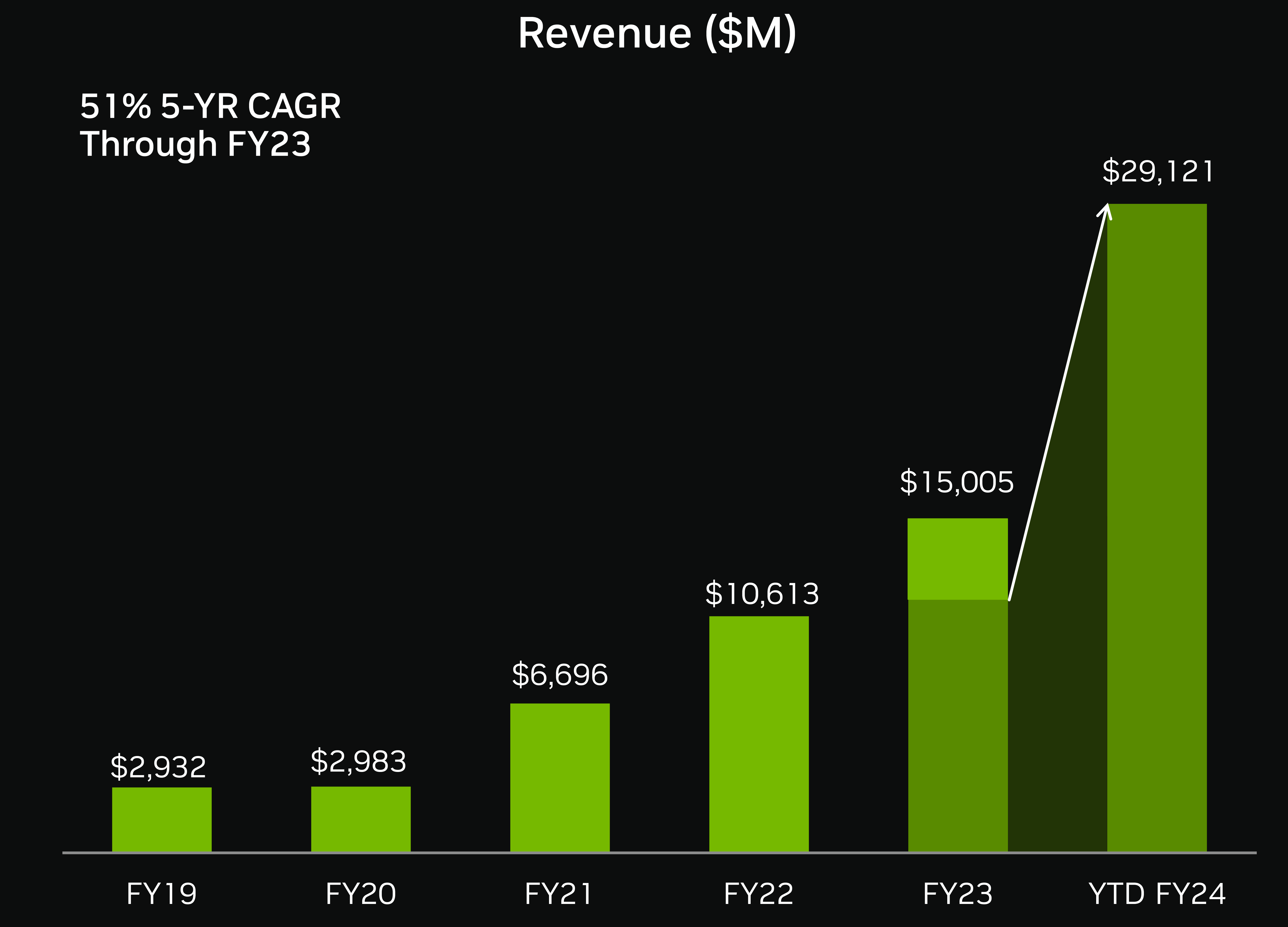
FY23 Revenue $0.9B
5-YR CAGR 10%

DRIVE Hyperion sensor architecture
with AGX compute

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

Data Center

The leading accelerated computing platform



Leader in AI & HPC

#1 in AI training and inference

Used by all hyperscale and major cloud computing providers and 40,000 enterprises

Powers 76% of the TOP500 supercomputers

Growth Drivers

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

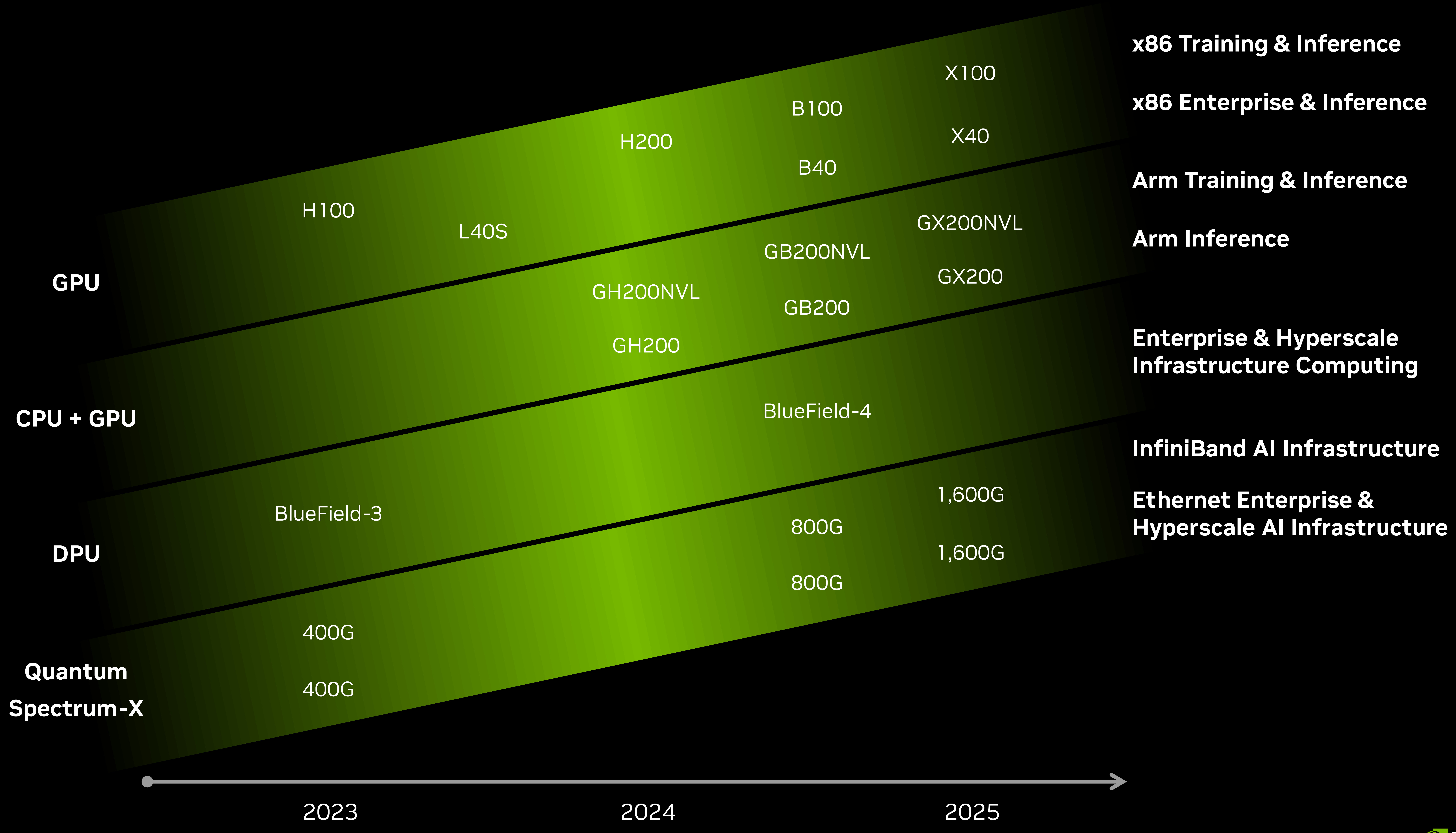
Emergence of “AI factory” — optimized for refining data and training, inferencing, and generating AI

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

DGX Cloud services and NVIDIA AI Enterprise software for building and running enterprise AI applications

NVIDIA AI — One Architecture | Train and Deploy Everywhere

One-Year Rhythm



GPU

H100

L40S

H200

B100

B40

X100

X40

x86 Training & Inference

x86 Enterprise & Inference

CPU + GPU

GH200NVL

GH200

GB200NVL

GB200

GX200NVL

GX200

Arm Training & Inference

Arm Inference

DPU

BlueField-3

BlueField-4

Enterprise & Hyperscale Infrastructure Computing

Quantum

Spectrum-X

400G

400G

800G

800G

1,600G

1,600G

InfiniBand AI Infrastructure

Ethernet Enterprise & Hyperscale AI Infrastructure

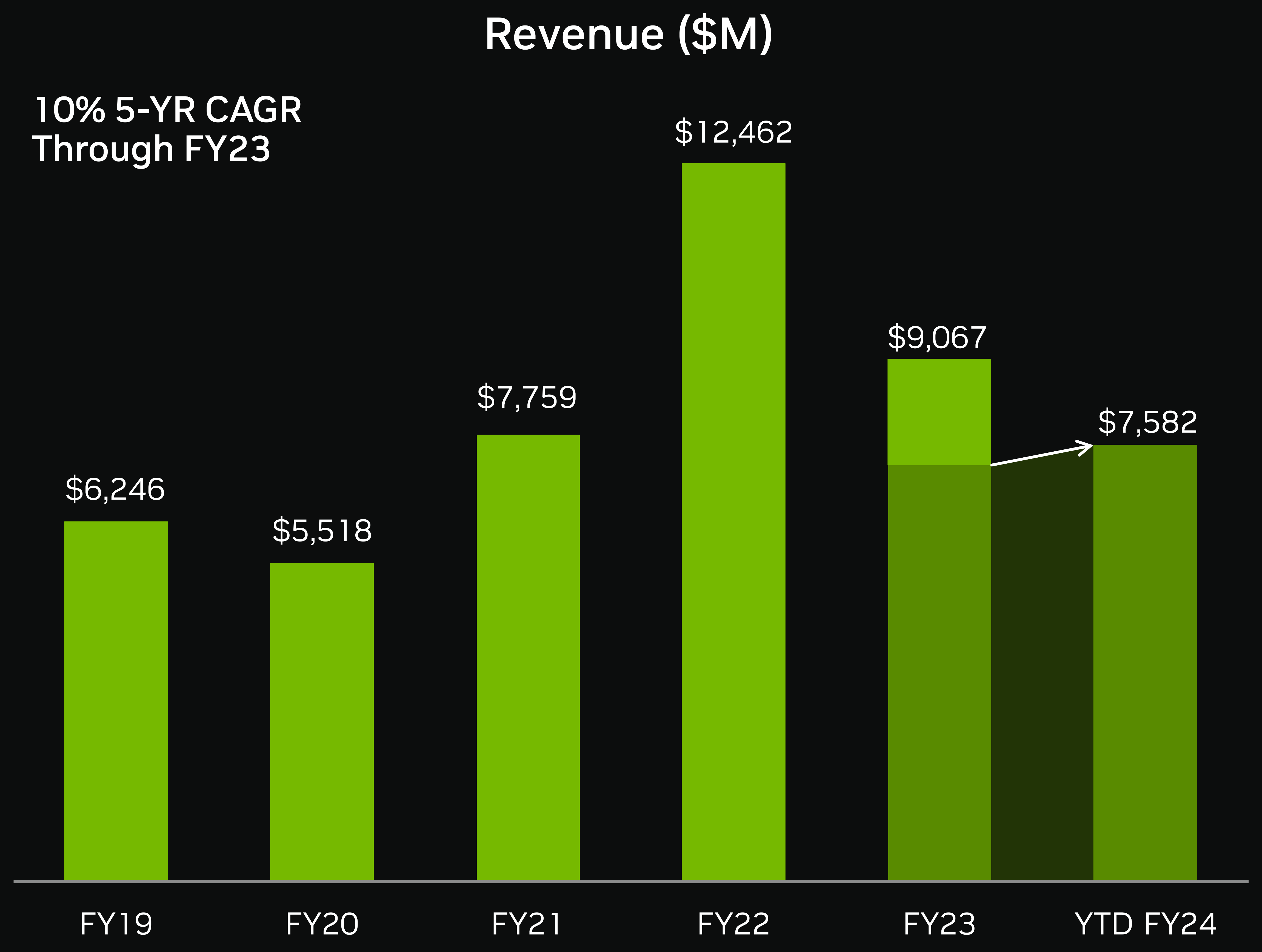
2023

2024

2025

Gaming

GeForce — the world's largest gaming platform



Leader in PC Gaming

Strong #1 market position

15 of the top 15 most popular GPUs on Steam

Leading performance & innovation

200M+ gamers on GeForce

Growth Drivers

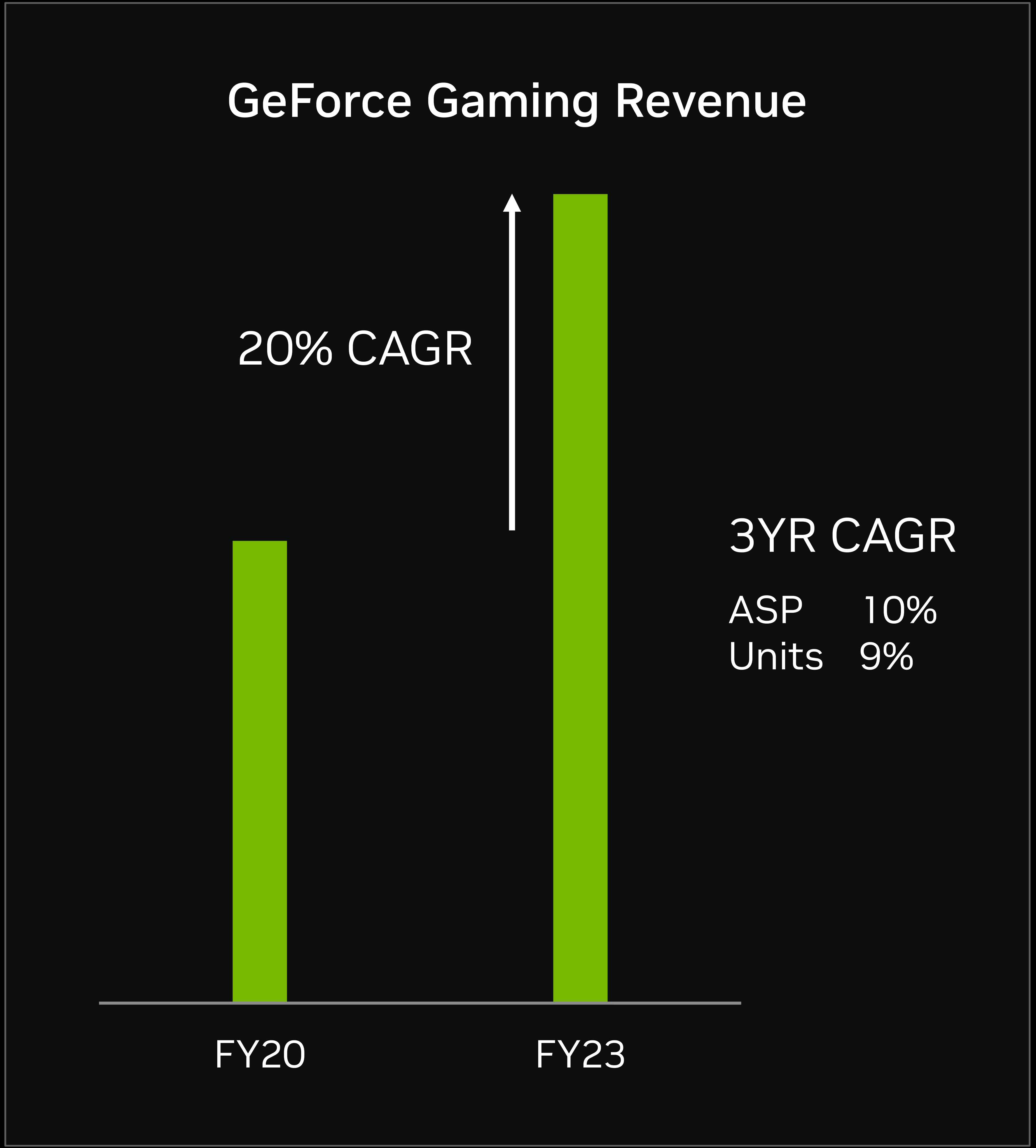
Rising adoption of NVIDIA RTX in games

Expanding universe of gamers & creators

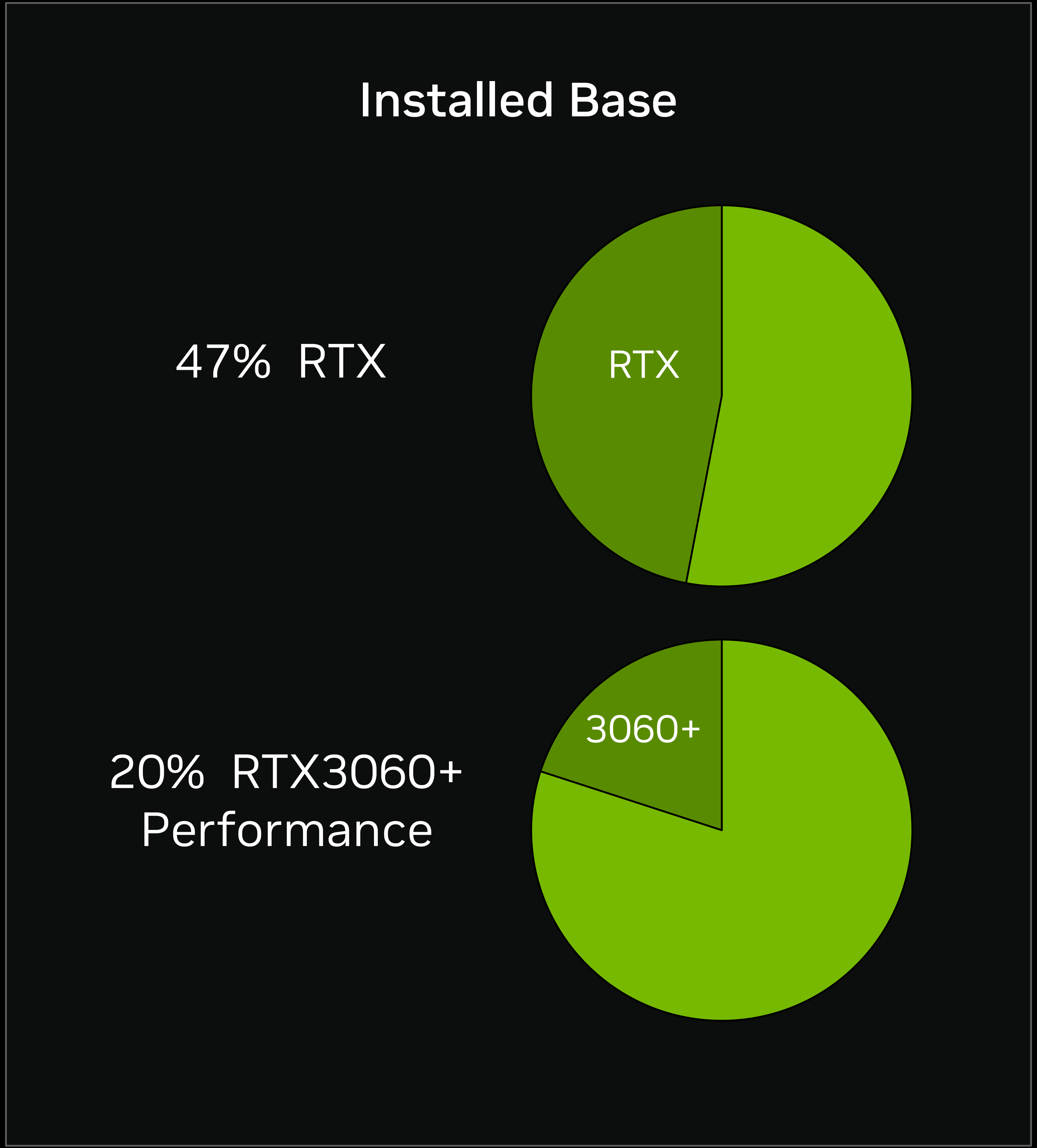
Gaming laptops & Gen AI on PCs

GeForce NOW Cloud gaming

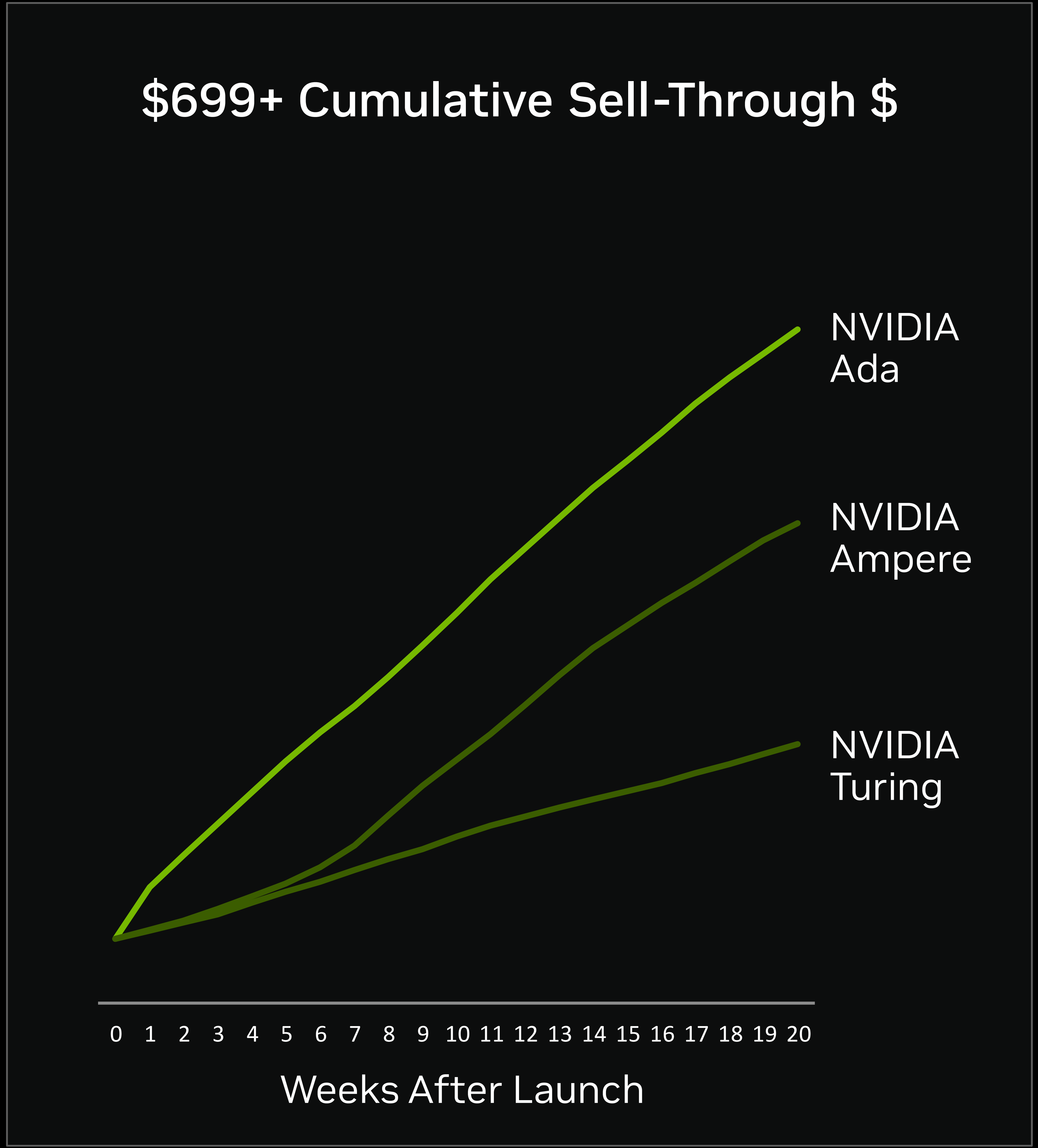
GeForce Extends Growth, Large Upgrade Opportunity



More Gamers, Richer Mix



Installed Base Needs Upgrade



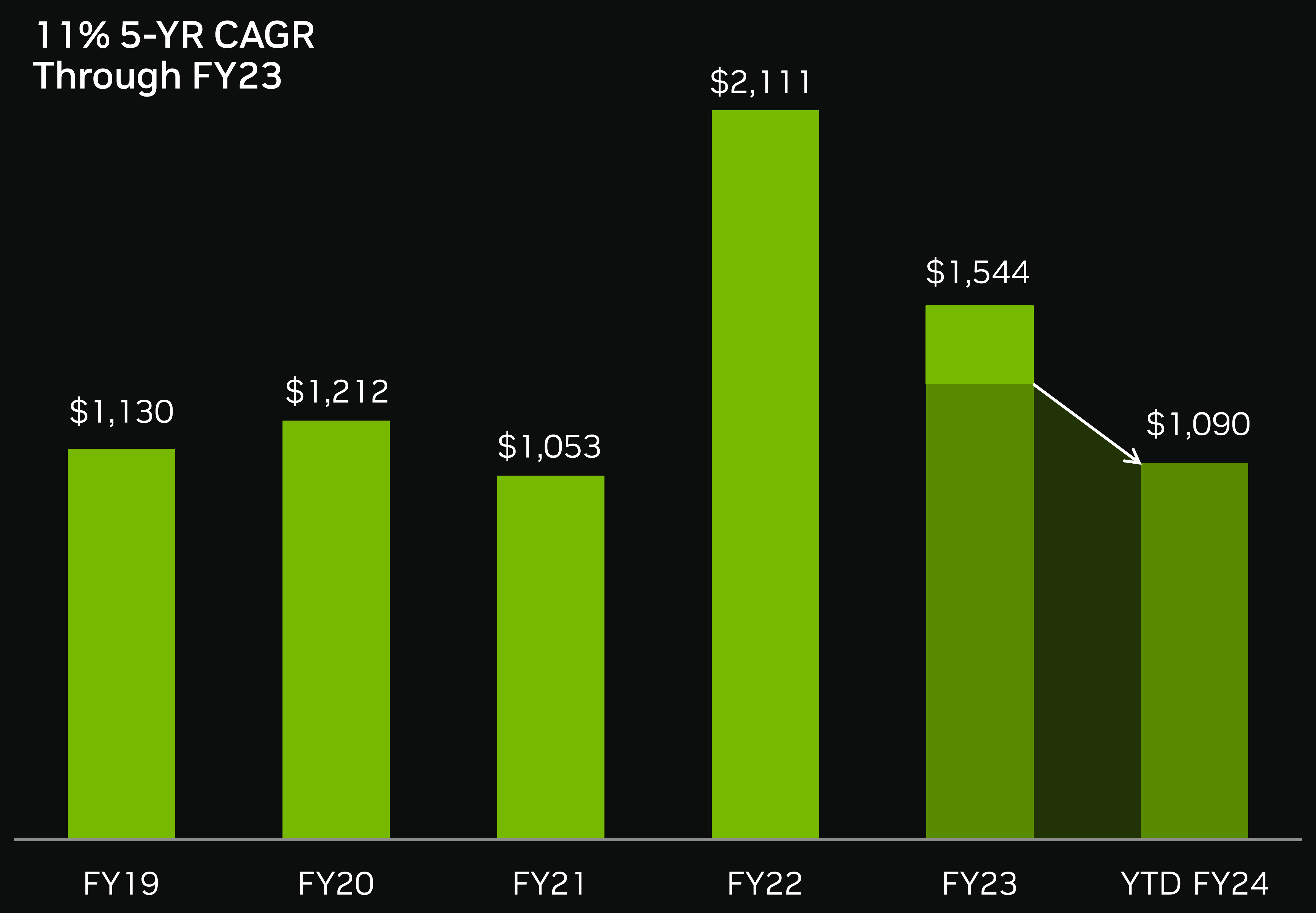
Ada: 3X Turing Ramp at $699+

Source: NVIDIA estimates

Professional Visualization

Workstation graphics

Revenue ($M)



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

Ray Tracing and generative AI revolutionizing design and content creation

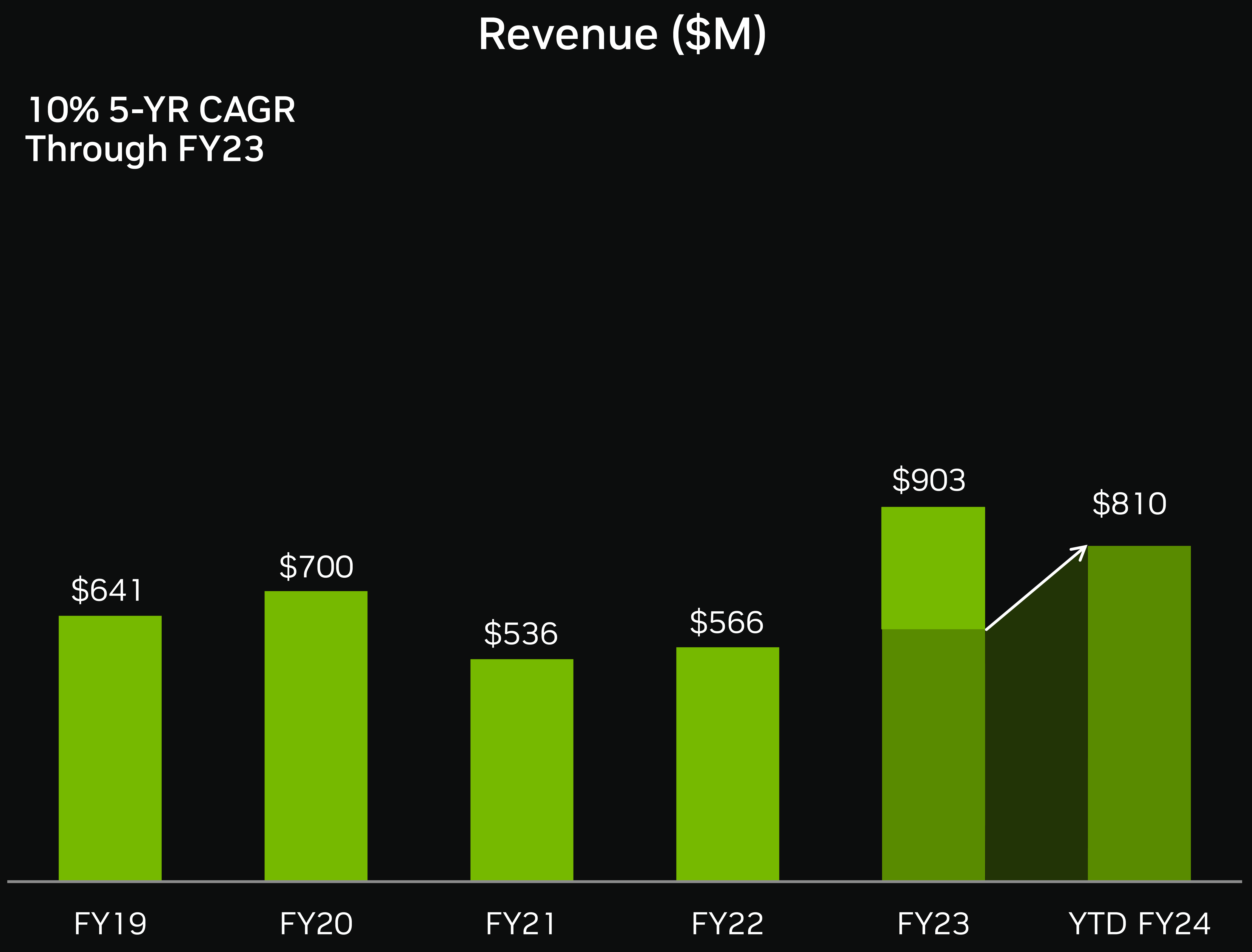
Expanding universe of designers and creators

Collaborative 3D design / Omniverse

Hybrid work environments

Automotive

Autonomous Vehicles (AV) and AI Cockpit



Leader in Autonomous Driving

NVIDIA DRIVE is our end-to-end Autonomous Vehicle (AV) and AI Cockpit platform featuring a full software stack and is powered by NVIDIA (systems-on-a-chip) SoCs in the vehicle

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

$1 Trillion Long-Term Available Market Opportunity

Cloud Service Providers & Consumer Internet

Enterprise

Autonomous Vehicles & Robotics

Industrial Digitalization

Omniverse Enterprise $150B

Autonomous Machines $300B

NVIDIA AI Enterprise & DGX Cloud $150B

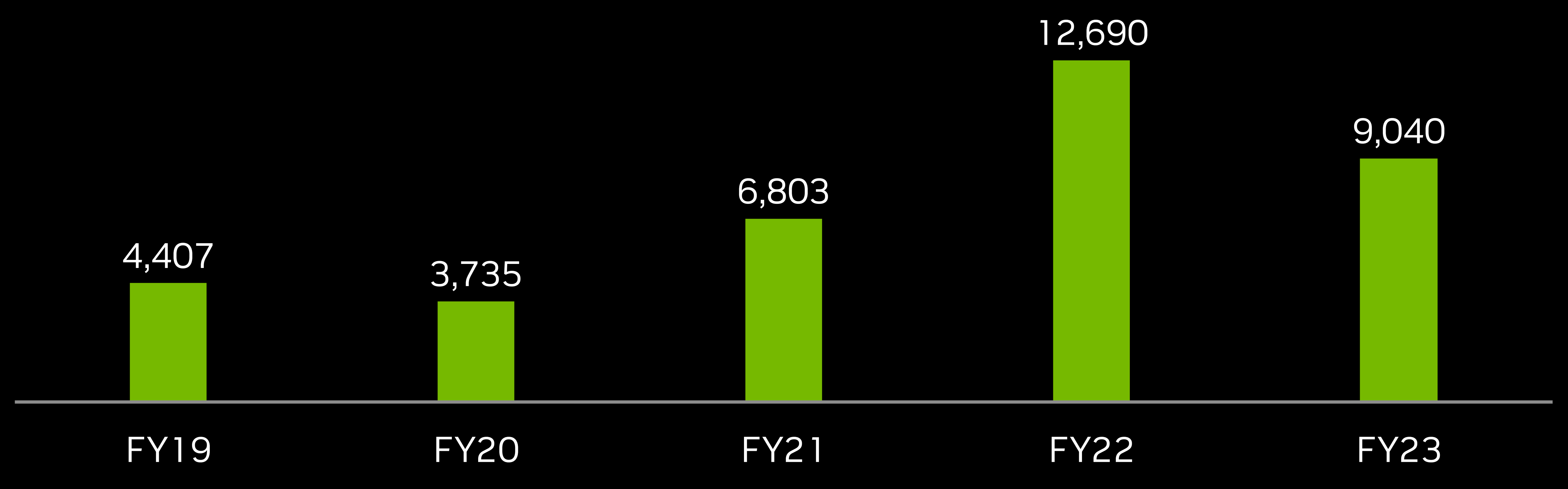
Gaming $100B

Data Center Systems $300B

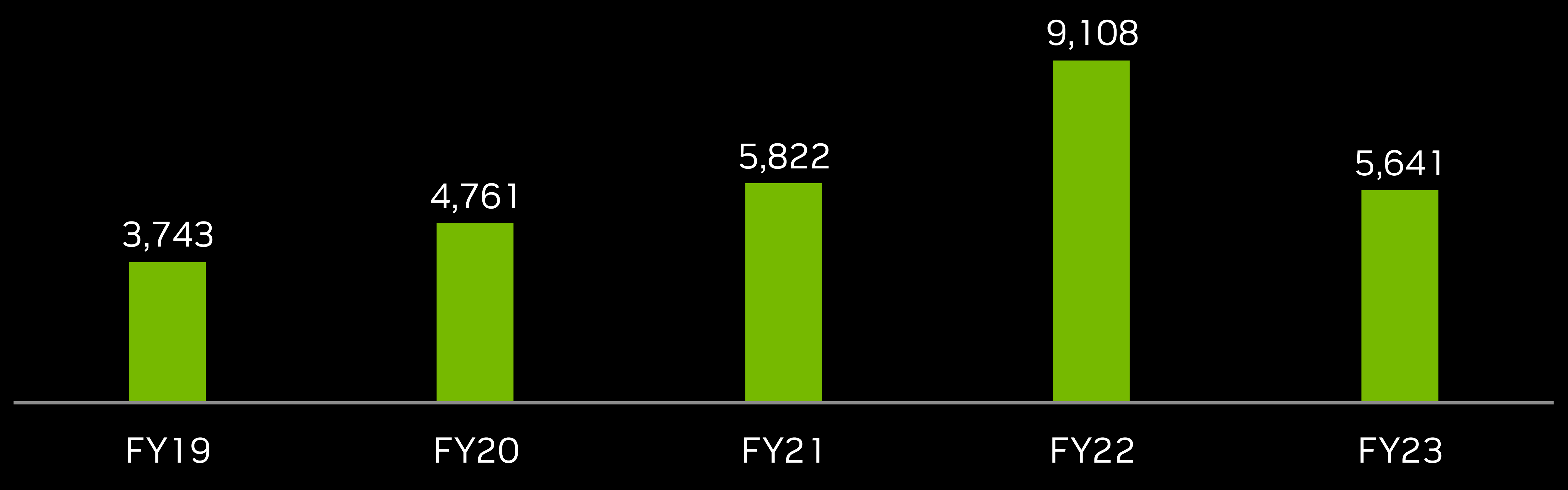
Financials

Annual Cash & Cash Flow Metrics

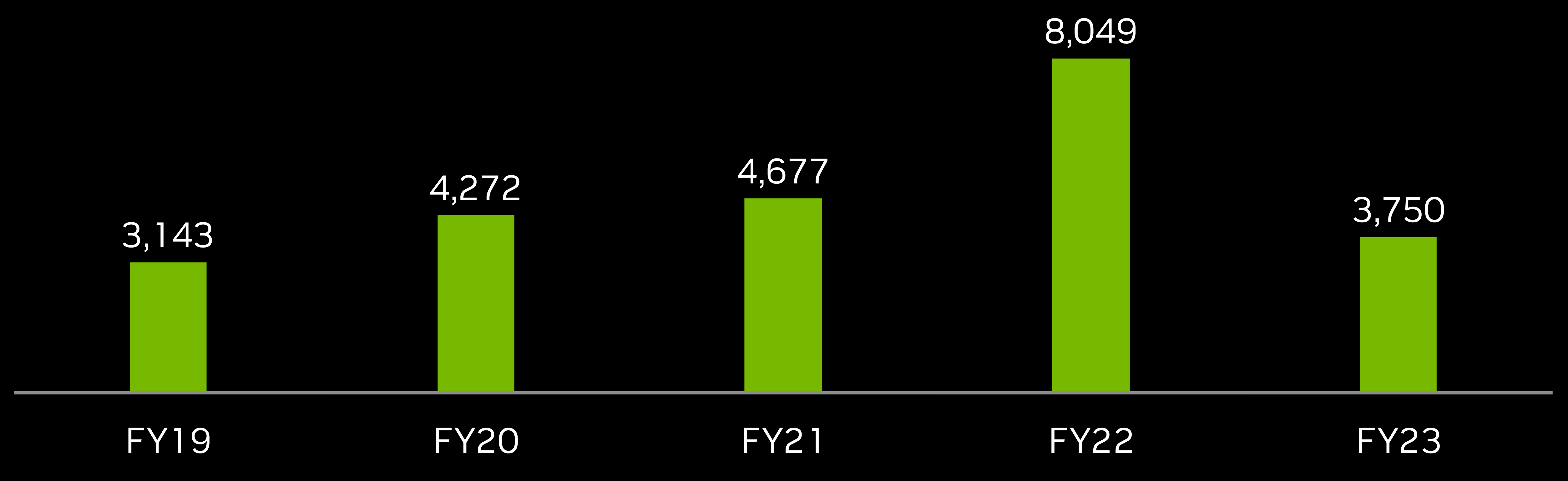
Operating Income (Non-GAAP) — $M



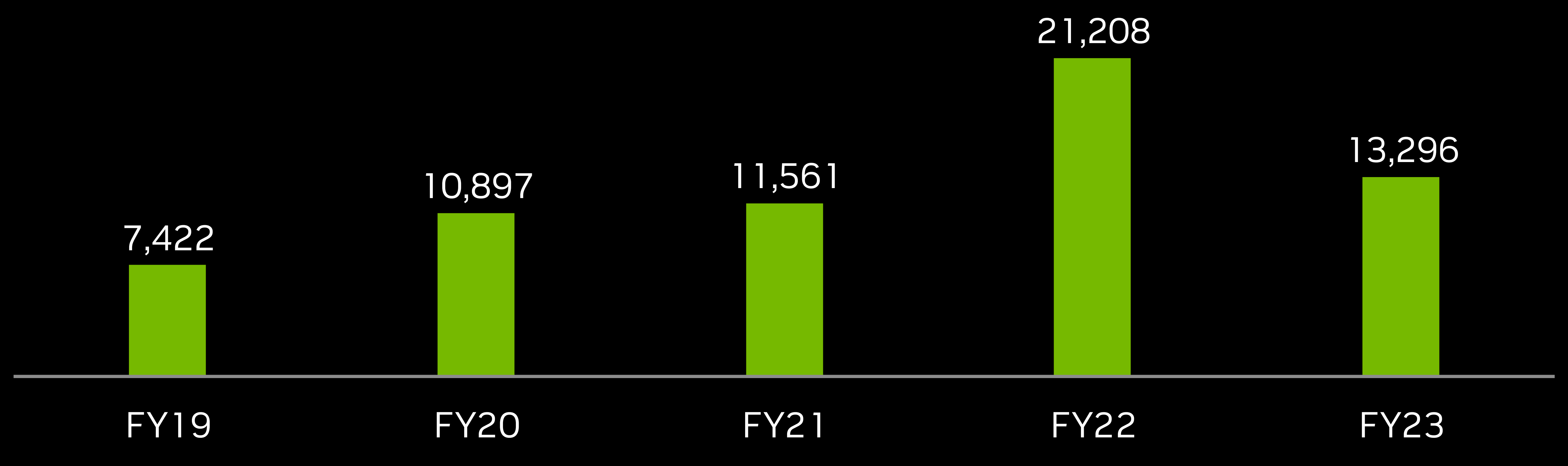
Operating Cash Flow — $M



Free Cash Flow (Non-GAAP) — $M



Cash Balance — $M



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

Corporate Responsibility

Environmentally Conscious

By FY26, aim to engage manufacturing suppliers comprising at least 67% of NVIDIA's scope 3 category 1 GHG emissions with goal of effecting supplier adoption of science-based targets

NVIDIA GPUs are typically 20X more energy efficient for certain AI and HPC workloads than traditional CPUs

Plan to achieve & maintain 100% renewable electricity for our operations and data centers by FY25 and annually thereafter

A Place For People To Do Their Life's Work

glassdoor
BEST PLACES TO WORK 2023

"100 Best Companies to Work For"
FORTUNE

"America's Most Just Companies"
CNBC

"Most Responsible Companies"
NEWSWEEK

"Best Places to Work for LGBT Equality"
HUMAN RIGHTS CAMPAIGN

Management

Time Magazine's 100 Most Influential Companies

Fast Company's Best Workplaces for Innovators

Fortune's World's Most Admired Companies

Wall Street Journal's Management Top 250 All-Stars

Corporate Governance

43% of Board is Gender, Racially, or Ethnically Diverse

93% of Directors are independent

Reconciliation of Non-GAAP to GAAP Financial Measures

Reconciliation of Non-GAAP to GAAP Financial Measures

	Non-GAAP	Acquisition-Related and Other Costs (A)	Stock-Based Compensation (B)	IP-Related Costs	Other (C)	Tax Impact of Adjustments	GAAP
Q3 FY24							
Gross margin ($ in million)	$13,583	(119)	(38)	(26)	—	—	$13,400
	75.0%	(0.7)	(0.2)	(0.1)	—	—	74.0%
Operating income ($ in million)	$11,557	(135)	(979)	(26)	—	—	$10,417
Net income ($ in million)	$10,020	(135)	(979)	(26)	(70)	433	$9,243
Shares used in diluted per share calculation (millions)	2,494	—	—	—	—	—	2,494
Diluted EPS	$4.02	—	—	—	—	—	$3.71

A. Consists of amortization of intangible assets and transaction costs.

B. Stock-based compensation charge was allocated to cost of goods sold, research and development expense, and sales, general and administrative expense.

C. Other represents net losses from non-affiliated investments and interest expense related to amortization of debt discount

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

Gross Margin	Non-GAAP	Acquisition-Related and Other Costs (A)	Stock-Based Compensation (B)	IP-Related Costs	GAAP
Q3 FY 2023	56.1%	(2.0)	(0.5)	—	53.6%
Q4 FY 2023	66.1%	(2.0)	(0.5)	(0.3)	63.3%
Q1 FY 2024	66.8%	(1.7)	(0.4)	(0.1)	64.6%
Q2 FY 2024	71.2%	(0.9)	(0.2)	—	70.1%

A. *Consists of amortization of intangible assets*
B. *Stock-based compensation charge was allocated to cost of goods sold*

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

Gross Margin ($ in Millions & Margin Percentage)	Non-GAAP	Acquisition-Related and Other Costs (A)	Stock-Based Compensation (B)	IP-Related Costs	GAAP
FY 2019	$7,233	—	(27)	(35)	$7,171
	61.7%	—	(0.2)	(0.3)	61.2%
FY 2020	$6,821	—	(39)	(14)	$6,768
	62.5%	—	(0.4)	(0.1)	62.0%
FY 2021	$10,947	(425)	(88)	(38)	$10,396
	65.6%	(2.6)	(0.5)	(0.2)	62.3%
FY 2022	$17,969	(344)	(141)	(9)	$17,475
	66.8%	(1.4)	(0.5)	—	64.9%
FY 2023	$15,965	(455)	(138)	(16)	$15,356
	59.2%	(1.7)	(0.5)	(0.1)	56.9%
YTD Q3 2023	$11,966	(335)	(108)	—	$11,523
	57.2%	(1.6)	(0.5)	—	55.1%
YTD Q3 2024	$28,000	(358)	(96)	(36)	$27,510
	72.1%	(0.9)	(0.2)	(0.1)	70.9%

A. Consists of amortization of intangible assets and inventory step-up
B. Stock-based compensation charge was allocated to cost of goods sold

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

Operating Income and Margin ($ in Millions & Margin Percentage)	Non-GAAP	Acquisition Termination Cost	Acquisition-Related and Other Costs (A)	Stock-Based Compensation (B)	IP-Related Costs	Other (C)	GAAP
FY 2019	$4,407	—	(2)	(557)	(35)	(9)	$3,804
	37.6%	—	—	(4.7)	(0.3)	(0.1)	32.5%
FY 2020	$3,735	—	(31)	(844)	(14)	—	$2,846
	34.2%	—	(0.3)	(7.7)	(0.1)	—	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)	(16)	(63)	$4,224
	33.5%	(5.0)	(2.5)	(10.0)	(0.1)	(0.2)	15.7%
YTD Q3 2023	$6,816	(1,353)	(499)	(1,971)	—	(25)	$2,968
	32.6%	(6.5)	(2.4)	(9.4)	—	(0.1)	14.2%
YTD Q3 2024	$22,385	—	(446)	(2,555)	(36)	10	$19,358
	57.7%	—	(1.1)	(6.6)	(0.1)	—	49.9%

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 2019	$3,143	600	—	$3,743
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
YTD Q3 2023	$2,015	1,324	54	$3,393
YTD Q3 2024	$15,732	815	44	$16,591

Reconciliation of Non-GAAP to GAAP Financial Measures

($ in Millions)	Q4 FY24 Outlook
Non-GAAP gross margin	75.5%
Impact of stock-based compensation expense, acquisition-related costs, and other costs	(1.0%)
GAAP gross margin	74.5%
Non-GAAP operating expenses	$2,200
Impact of stock-based compensation expense, acquisition-related costs, and other costs	965
GAAP operating expenses	$3,165

NVIDIA