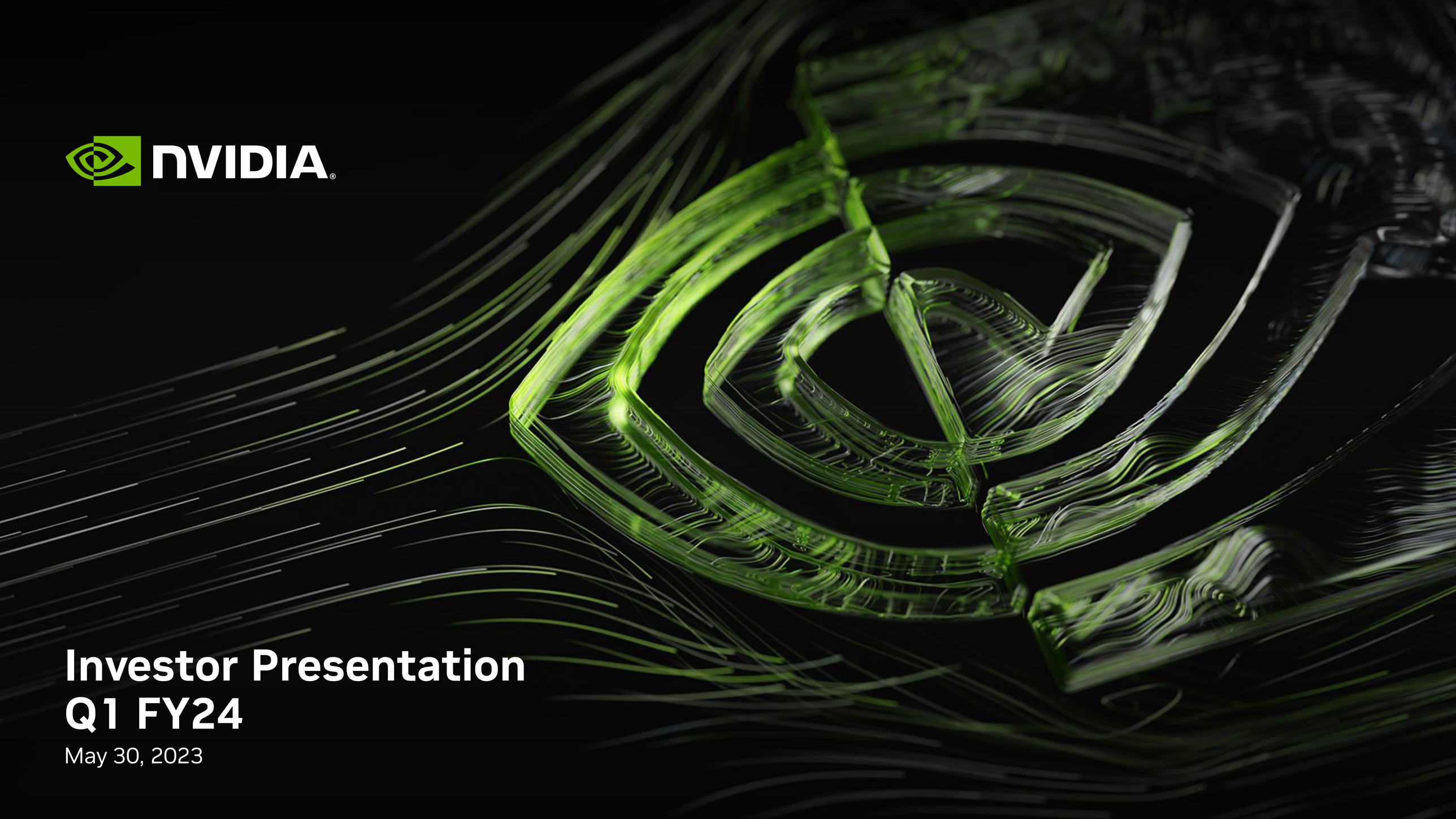


Investor Presentation
Q1 FY24

May 30, 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; Gaming and Professional Visualization platforms emerging from channel inventory corrections which we believe are behind us; CSPs around the world racing to deploy our NVIDIA Hopper- and Ampere-architecture GPUs to meet the surge in interest; consumer internet companies being at the forefront of adopting generative AI and deep-learning-based recommendation systems; GeForce RTX 40 Series GPU laptops being off to a strong start; generative AI driving a step-function increase in inference workloads; generative AI being transformative to gaming and content creation; our expectation that slower-than-expected demand growth for NEV customers in China will linger for the rest of the calendar year; the impact and timing details for our partnership with BYD; our financial outlook, Data Center drivers and visibility, our expectation to increase investments, our expected tax rates and our expected capital expenditures for the second quarter of fiscal 2024; the benefits, impact, performance, features and availability of our products and technologies; the benefits, impact, features and timing of our collaborations or partnerships; accelerated computing being needed to tackle the most impactful opportunities of our time; AI as the greatest technology force of our time; data centers across industries becoming AI factories; NVIDIA's AI expertise and scale helping to revolutionize businesses; Omniverse being essential for the next wave of AI – robotics; NVIDIA's value to every stakeholder in the ecosystem; the cost and time-to-solution savings of application speed-ups; our remaining repurchase authorization and dividend program plan; upcoming launches of our Data Center products; our Automotive design win pipeline, ramp and production expectations; NVIDIA accelerated computing being broadly recognized as the way to advance computing as Moore's law ends; generative AI unlocking new opportunities; NVIDIA's expanding accelerated computing ecosystem; the next wave of AI being robotics; NVIDIA helping bring AI to the world's largest industries; building and operating Metaverse applications being the next wave; NVIDIA software and services enabling the world's enterprises to revolutionize industries with AI; our goal of effecting supplier adoption of science-based environmental targets by 2026; and our plan for 100% of our global electricity usage for our offices and data centers to be renewable by 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.

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

- Q1 FY24 Earnings Summary
- Key Announcements This Quarter
- NVIDIA Overview
- Financials
- Reconciliation of Non-GAAP to GAAP Financial Measures

Q1 FY24
Earnings Summary



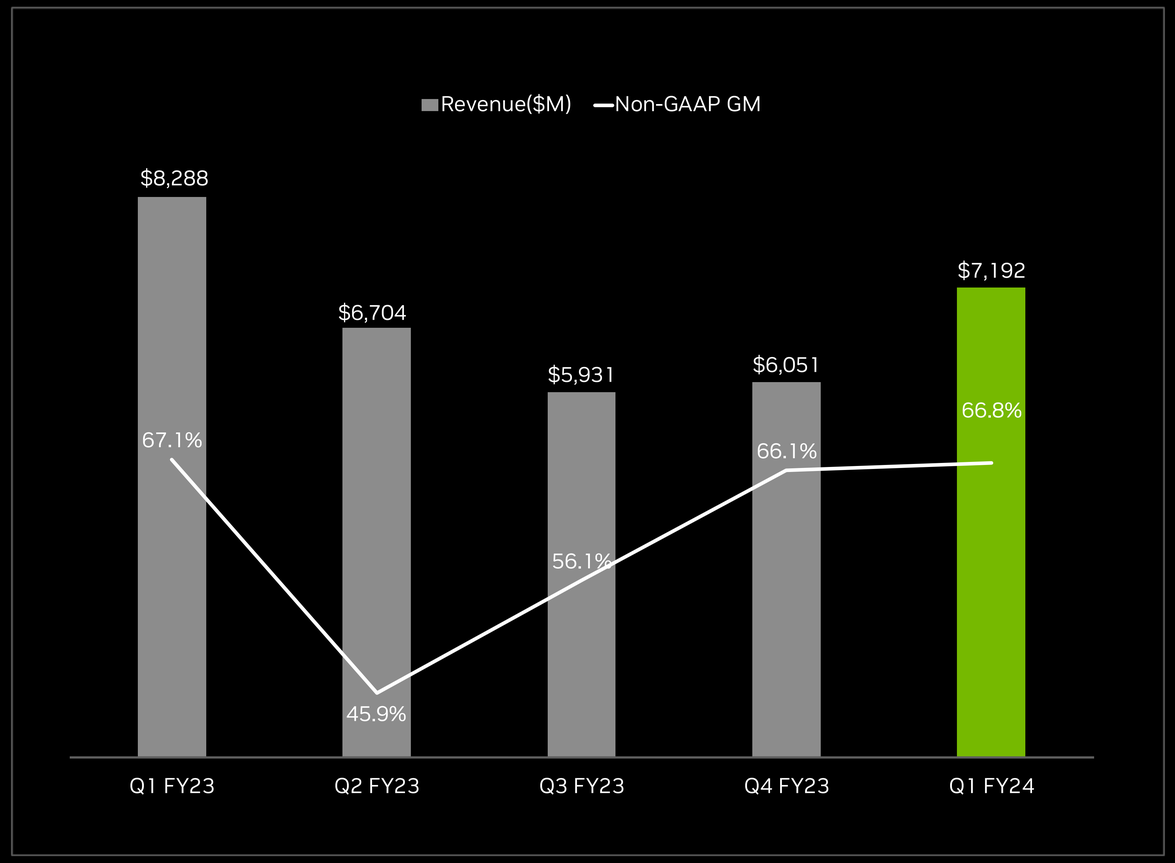
Highlights

- **Strong Q/Q growth was driven by record Data Center revenue; Gaming and Professional Visualization platforms emerging from channel inventory corrections**
  - Total revenue down 13% Y/Y to $7.19B, well above outlook of $6.50B +/- 2%
  - Data Center up 14% Y/Y to $4.28B
  - Gaming down 38% Y/Y to $2.24B

- **Record Data Center revenues on strong growth of our accelerated computing platform worldwide; Generative AI drove significant upside in demand**
  - Cloud Service Providers (CSPs) racing to deploy flagship NVIDIA Hopper- and Ampere-architecture GPUs to meet surge in interest
  - Strong growth from consumer internet companies adopting generative AI and deep-learning-based recommendation systems
  - Strong enterprise demand for AI and accelerated computing; momentum in automotive, financial services, healthcare, telecom

- **Strong Q/Q Gaming growth was driven by sales of NVIDIA GeForce RTX 40 Series GPUs for both notebooks and desktops**
  - End demand was solid and consistent with seasonality, demonstrating resilience against a challenging consumer spending backdrop
  - GeForce RTX 40 Series GPU laptops are off to a strong start
  - Ramped the RTX 4070 in desktop, joining previously launched RTX 4090, 4080 and 4070Ti GPUs

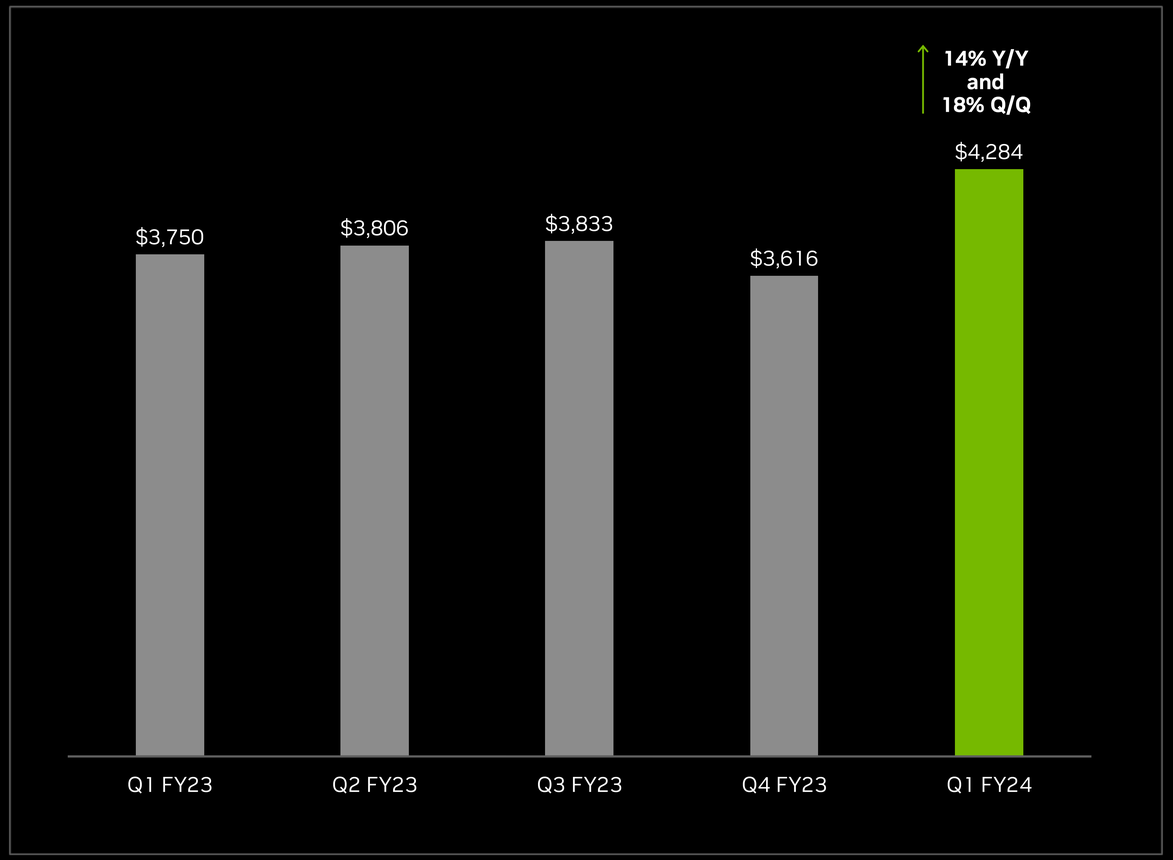
Q1 FY24 Financial Summary



	GAAP			Non-GAAP		
	Q1 FY24	Y/Y	Q/Q	Q1 FY24	Y/Y	Q/Q
Revenue	$7,192	-13%	+19%	$7,192	-13%	+19%
Gross Margin	64.6%	-0.9 pts	+1.3 pts	66.8%	-0.3 pts	+0.7 pts
Operating Income	$2,140	+15%	+70%	$3,052	-23%	+37%
Net Income	$2,043	+26%	+44%	$2,713	-21%	+25%
Diluted EPS	$0.82	+28%	+44%	$1.09	-20%	+24%
Cash Flow from Ops	$2,911	+68%	+29%	$2,911	+68%	+29%

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

Data Center

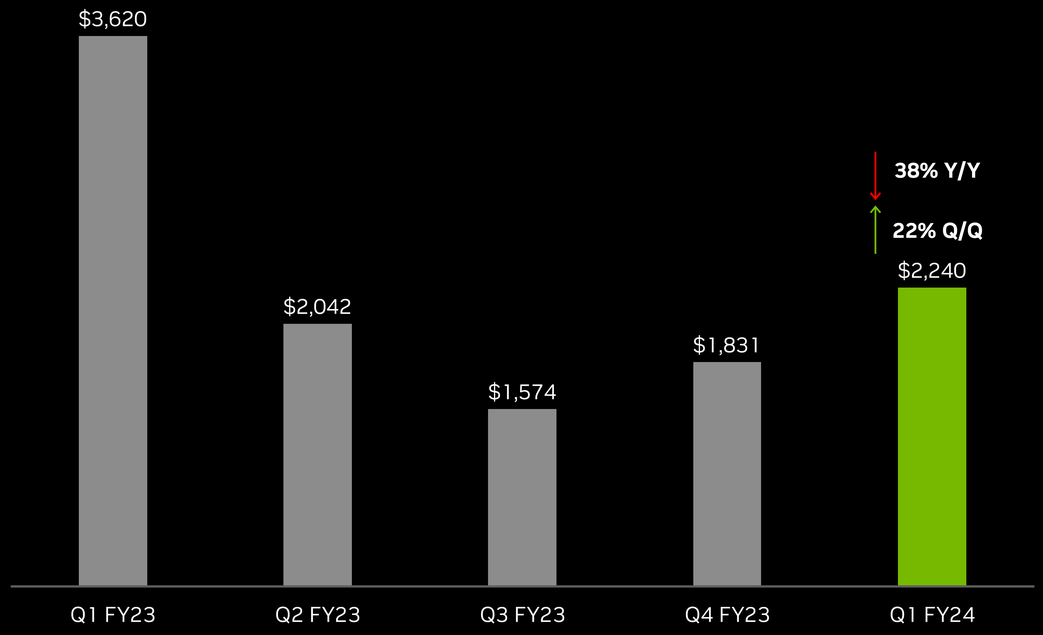


Revenue ($M)

Highlights

- Started shipping DGX H100 – our Hopper-generation AI system – which customers can deploy on-prem
- Generative AI is driving a step-function increase in training and inference workloads
- Strong demand in networking at both CSPs and enterprise customers for generative AI and accelerated computing; demand relating to general-purpose CPU infrastructure remains soft
- Bluefield-3 is in production and has been adopted by multiple hyperscale and CSP customers, including Microsoft Azure, Oracle Cloud, CoreWeave, Baidu and others
- Grace CPU and Grace-Hopper Superchips sampling now

Gaming

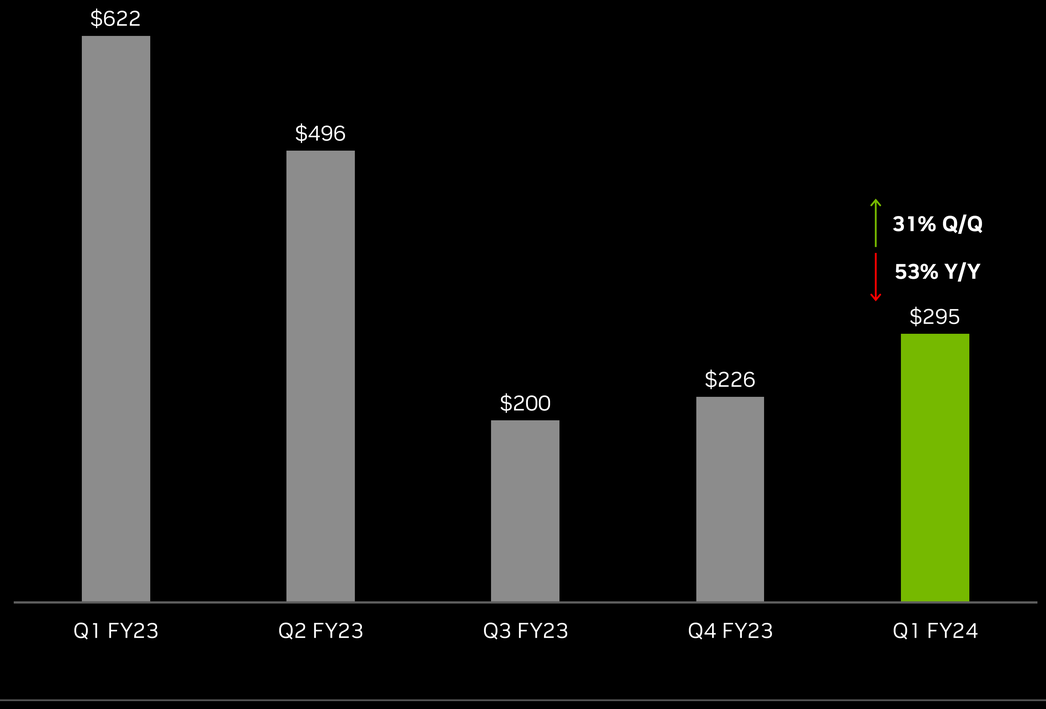


Revenue ($M)

Highlights

- Believe channel inventory corrections are behind us
- GeForce RTX 40 Series GPU laptops are off to a strong start
- Ramped the RTX 4070 in desktop, joining previously launched RTX 4090, 4080 and 4070Ti GPUs
- Generative AI will be transformative to gaming and content creation – from development to runtime
- Over 1,600 games on GeForce NOW (GFN), the richest content available on any cloud gaming service
- First game from Microsoft Xbox partnership, *Gears 5*, now available on GFN; more set to be released in the coming months

Professional Visualization

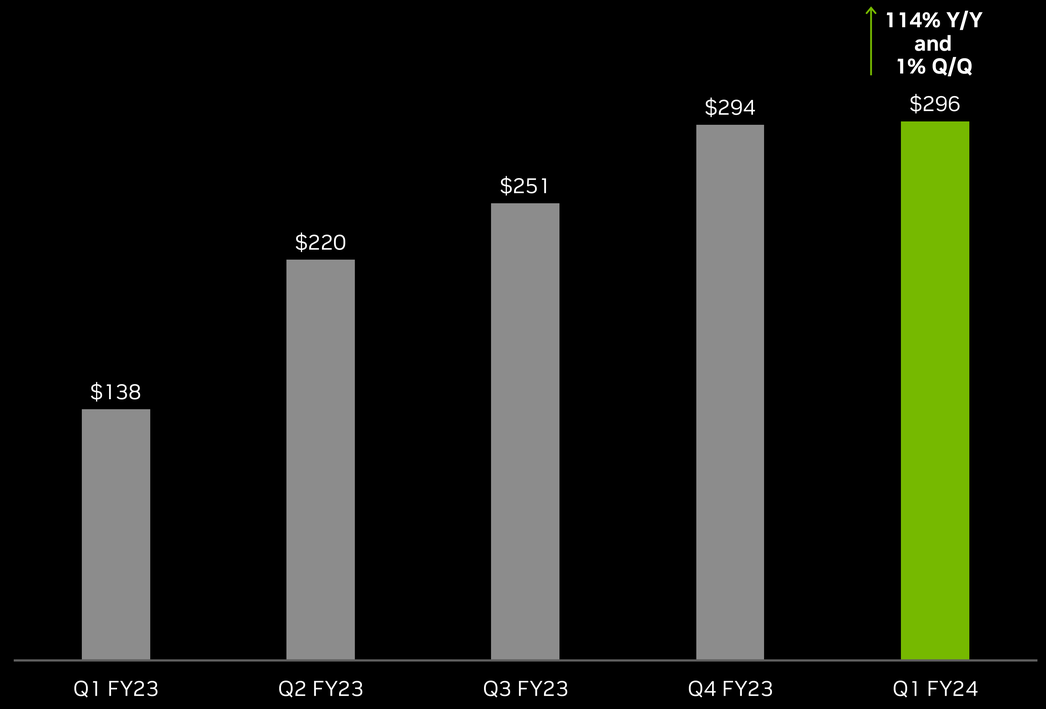


Revenue ($M)

Highlights

- Sequential growth was driven by stronger workstation demand across both mobile and desktop form factors
- Strength in key verticals such as public sector, healthcare, and automotive
- Believe the channel inventory correction is behind us
- Ramp of Ada Lovelace GPU architecture in workstations kicks off a major product cycle
- Announced six new RTX GPUs for laptop and desktop workstations; further rollouts planned in the coming quarters

Automotive

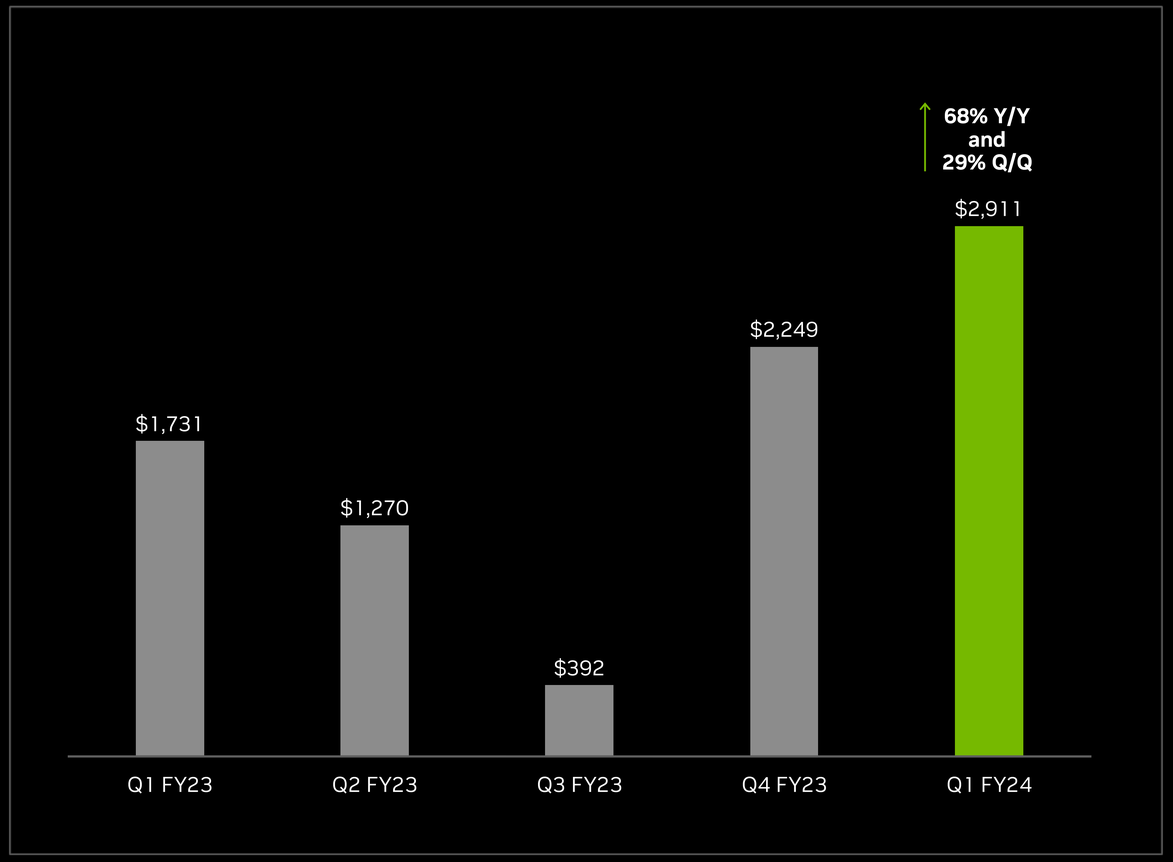


Revenue ($M)

Highlights

- Strong Y/Y growth was driven by the ramp of the NVIDIA DRIVE Orin across a number of new energy vehicles
- Q/Q growth moderated as some NEV customers in China are adjusting their production schedules to reflect slower-than-expected demand growth
  - Expect this dynamic to linger for the rest of calendar year
- Expanded partnership with BYD; new design win will extend BYD's use of DRIVE Orin to its next-generation, high-volume Dynasty and Ocean series of vehicles; production start in calendar 2024

Sources & Uses of Cash



Cash Flow from Operations ($M)

Highlights

- Y/Y increase reflects lower inventory prepayments and changes in inventory, partially offset by lower revenue
- Q/Q increase was driven by higher revenue
- Returned $99M to shareholders in the form of cash dividends
- Invested $268M in capex (includes principal payments on PP&E)
- Ended the quarter with $15.3B in gross cash and $11.0B in debt; $4.3B in net cash

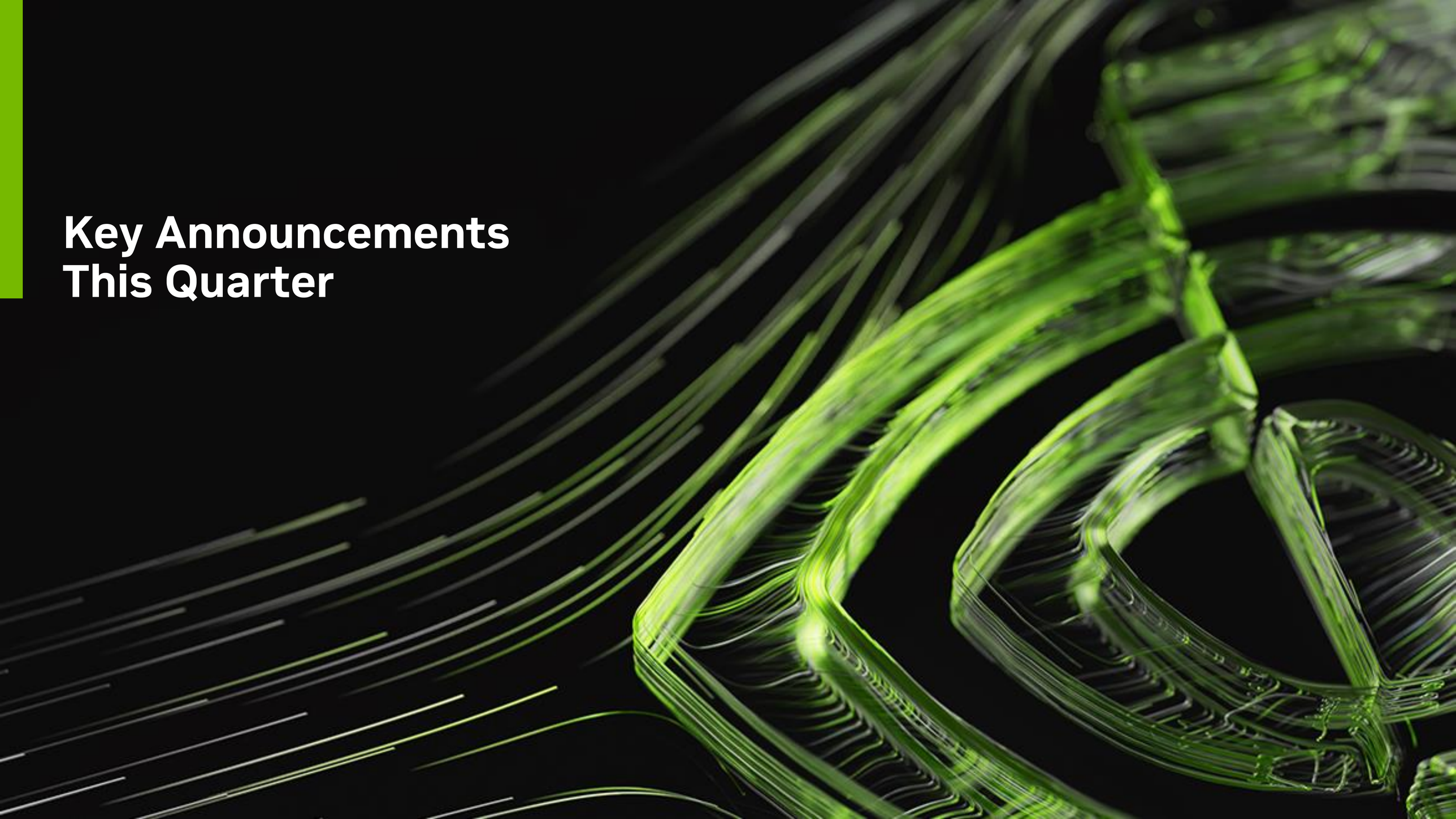
Gross cash is defined as cash/cash equivalents & marketable securities.
Debt is defined as principal value of debt.
Net cash is defined as gross cash less debt.

Q2 FY24 Outlook

Revenue	**$11.0 billion**, plus or minus 2% Expect q/q growth to largely be driven by Data Center, reflecting a steep increase in demand related to generative AI and large language models. This demand has extended our Data Center visibility out a few quarters, and we have procured substantially higher supply for the second half of the year
Gross Margins	**68.6%** GAAP and **70.0%** non-GAAP, plus or minus 50 basis points
Operating Expense	Approximately **$2.71 billion** GAAP and **$1.90 billion** non-GAAP Expect to increase investments in the business while also delivering operating leverage
Other Income & Expense	Income of approximately **$90 million** for GAAP and non-GAAP Excluding gains and losses on non-affiliated investments
Tax Rate	**14.0%** GAAP and non-GAAP, plus or minus 1%, excluding discrete items
Capital Expenditures	Approximately **$300 million** to **$350 million** For full year fiscal 2024, capital expenditures are still expected to be approximately $1.10 billion to $1.30 billion

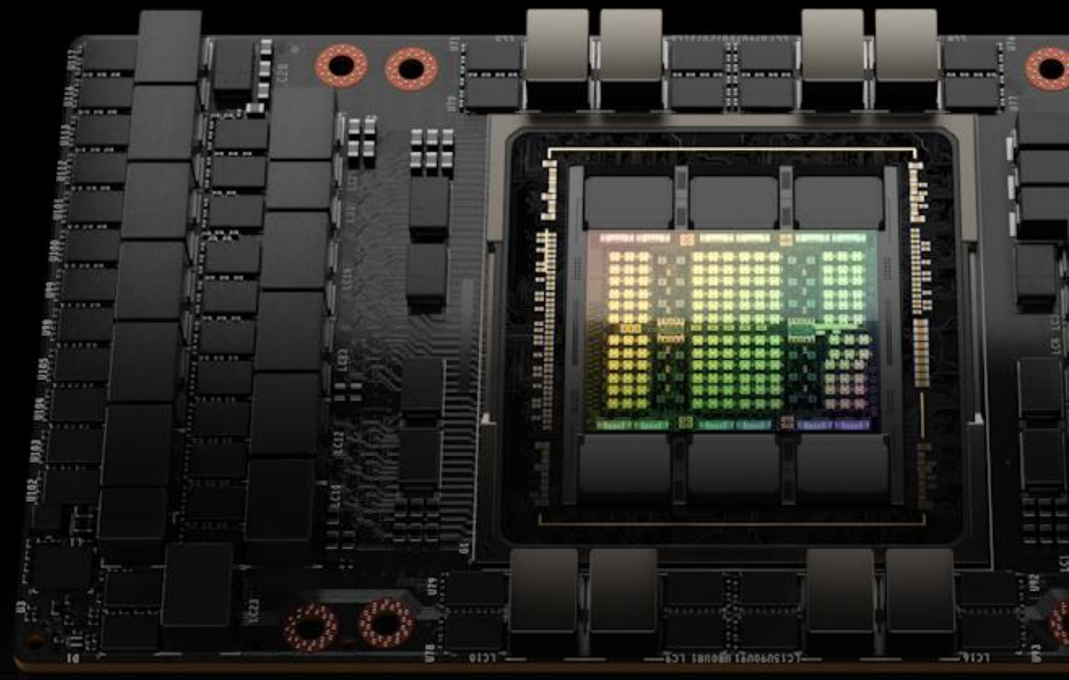
Refer to Appendix for reconciliation of Non-GAAP measures.

Key Announcements This Quarter



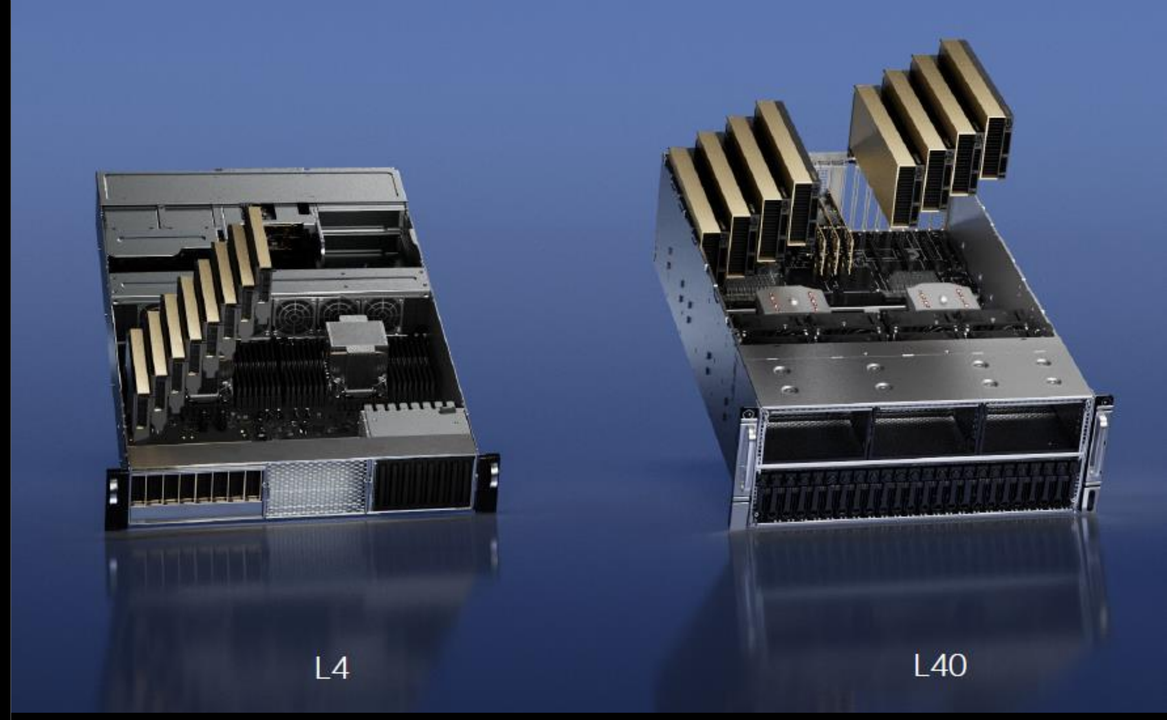
H100 Gets Broad Adoption by CSPs and AI Pioneers; DGX H100 Now Shipping

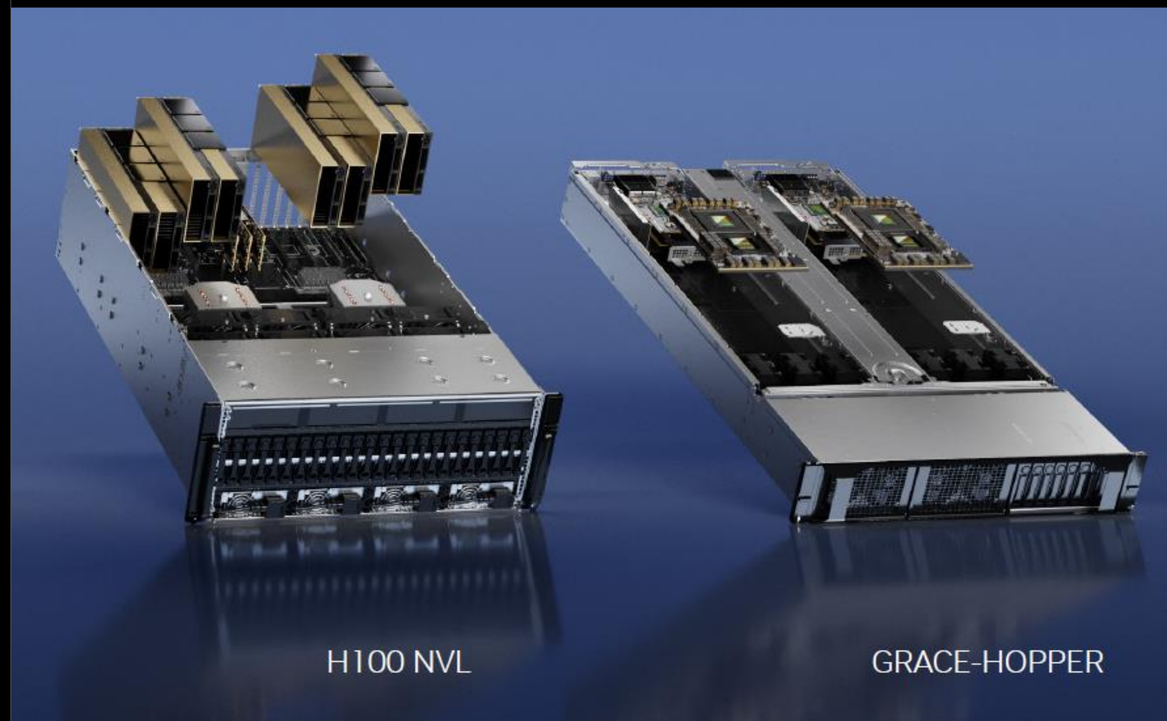
- Multiple cloud service providers (CSPs) announced the availability of H100 on their platforms
  - Private previews at Microsoft Azure, Oracle Cloud Infrastructure, Google Cloud and upcoming offering at AWS
- GPU-specialized cloud providers Cirrascale, Coreweave and Lambda have announced H100 instances; with upcoming offerings from Paperspace and Vultr
- AWS announced a multi-part collaboration with NVIDIA including its EC2 UltraClusters of P5 instances which can scale in size up to 20,000 interconnected H100 GPUs
- Meta has deployed its H100-powered Grand Teton AI supercomputer; other AI pioneers adopting H100 include OpenAI, Stability.ai and Anlatan
- DGX H100 AI supercomputers are shipping to enterprises worldwide



New NVIDIA Inference Platforms for Large Language Models and Generative AI Workloads

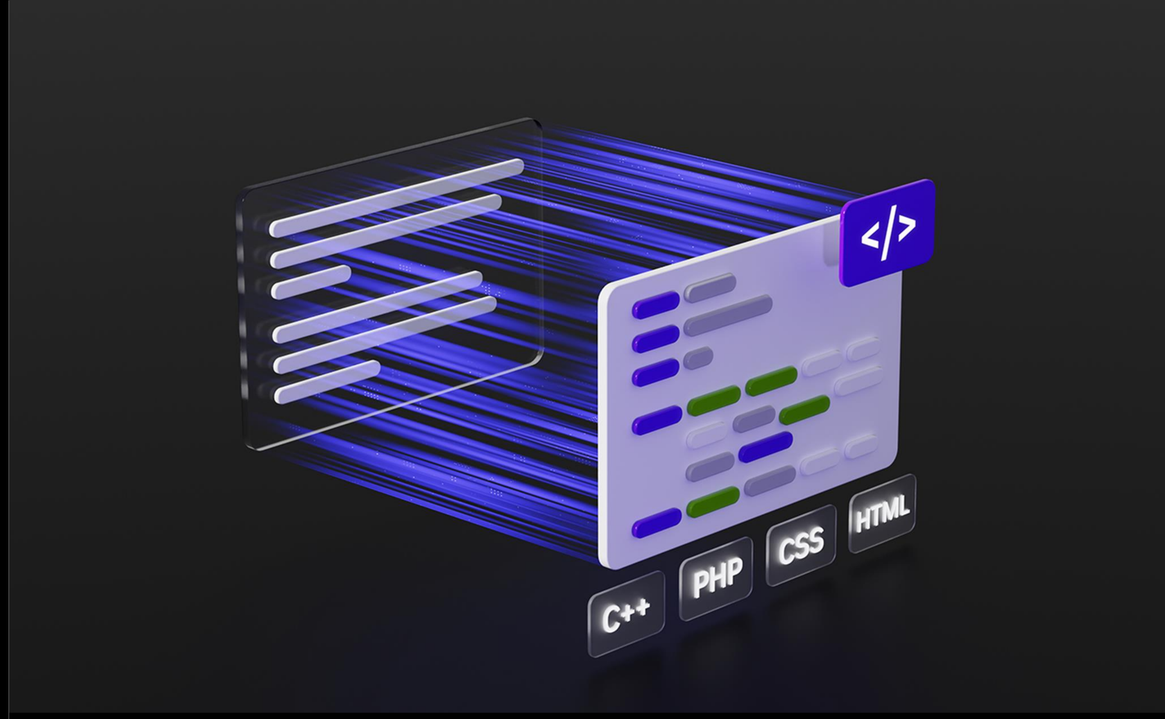
- Four new inference platforms optimized for generative AI applications, combining NVIDIA's inference software with the NVIDIA Ada, Hopper and Grace Hopper processors
  - NVIDIA L4 for AI Video
  - NVIDIA L40 for Image Generation
  - NVIDIA H100 NVL for Large Language Models
  - NVIDIA Grace Hopper for Recommendation Models
- NVIDIA L4 and L40 GPUs are available from leading system builders
- NVIDIA H100 NVL GPU and Grace Hopper Superchip are expected to be available in 2H 2023

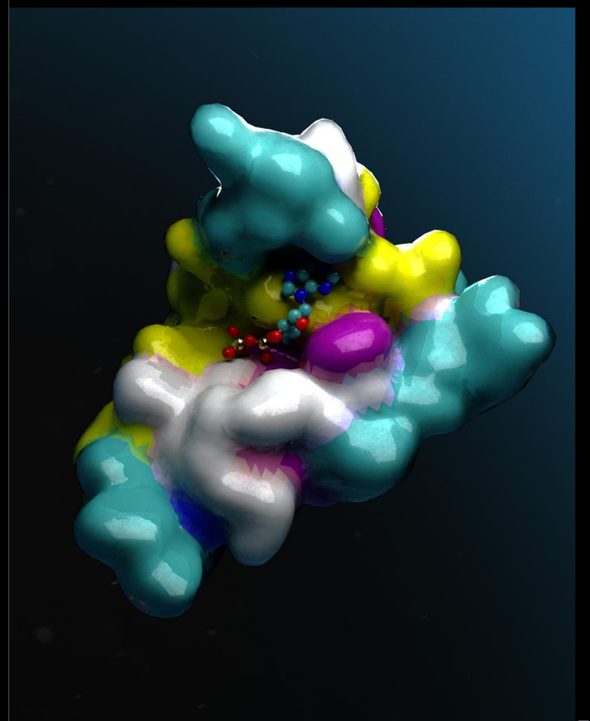




Google Cloud Adopts NVIDIA L4 Inference Platform for Generative AI and Other Major Workloads

- Google Cloud is the first CSP to adopt L4 with the launch of its new G2 virtual machines, available in public preview
- Workloads that Google Cloud will accelerate on L4 include:
  - Generative AI models for customers like Wombo and Descript
  - Integrating Triton Inference Server with Google Kubernetes Engine and Vertex AI platform
  - Google Cloud Dataproc with NVIDIA Spark-RAPIDS
  - Google's AlphaFold, and UL2 and T5 large language models
  - Google Cloud's Immersive Stream that renders 3D and AR experiences







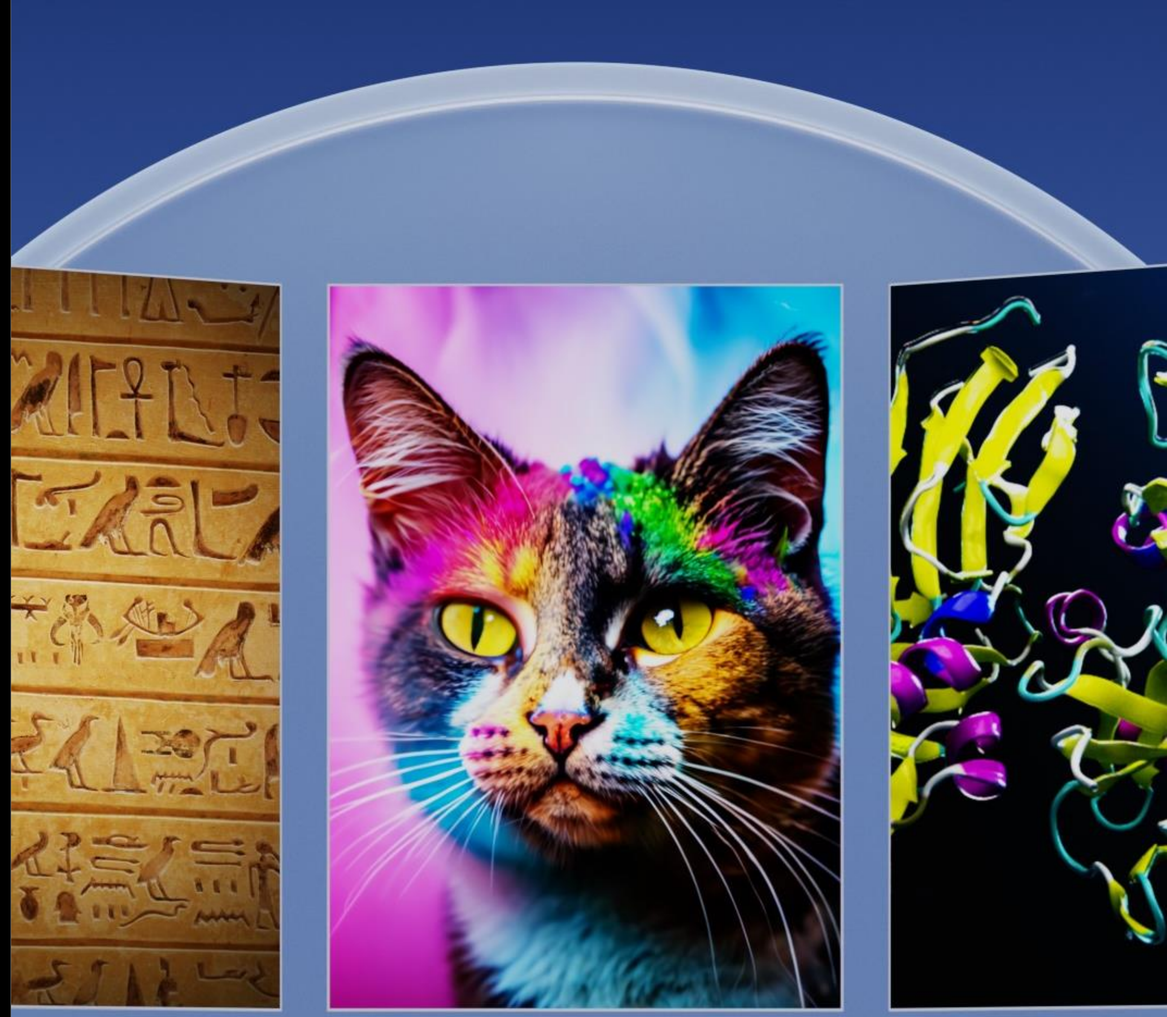
NVIDIA Launches DGX Cloud

- NVIDIA DGX Cloud is an AI supercomputing service that gives enterprises access to the infrastructure and software needed to train and deploy advanced models for generative AI and other groundbreaking applications
- Each DGX Cloud instance starts at $36,999 per month and includes:
  - Eight NVIDIA H100 or A100 80GB Tensor Core GPUs, a high-performance, low-latency fabric built with NVIDIA Networking, and high-performance storage
  - NVIDIA Base Command and NVIDIA AI Enterprise software
  - Support from NVIDIA experts
- Leading CSPs partnering with NVIDIA to host DGX Cloud infrastructure include Oracle Cloud Infrastructure, Microsoft Azure, and Google Cloud, with more on the way
- Early adopters include Amgen, CCC Intelligent Solutions, and ServiceNow



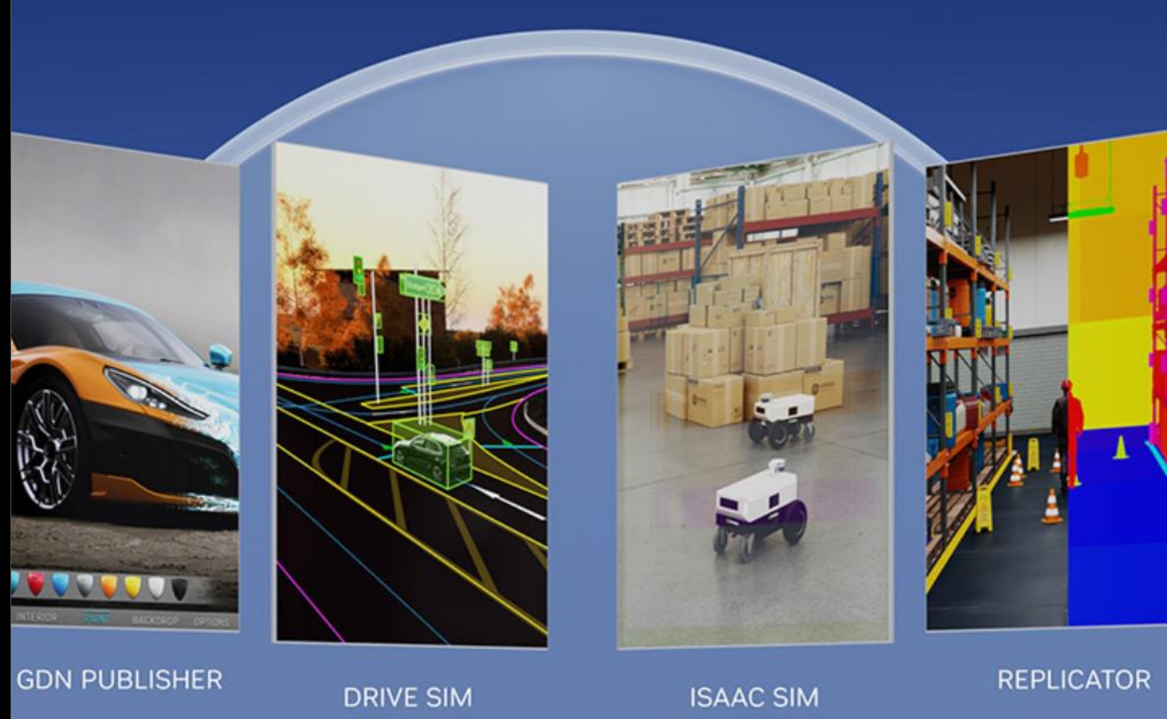
NVIDIA AI Foundations – Cloud Services For Generative AI

- NVIDIA AI Foundations are cloud services that enable businesses to build, refine and operate custom LLMs and generative AI models trained with their own proprietary data and created for their unique domain-specific tasks
- Services span language, images, video, and 3D and run on DGX Cloud. Includes:
  - NVIDIA NeMo for language
  - NVIDIA Picasso for image, video and 3D
  - NVIDIA BioNeMo for life sciences
- Getty Images, Morningstar, Quantiphi and Shutterstock are among the companies creating and using AI models built with NVIDIA AI Foundations
- Amgen, Evozyne, and Insilico Medicine are among the early adopters of BioNeMo
- Partnering with Adobe to co-develop and bring to market advanced generative AI models through Adobe's Creative Cloud and NVIDIA Picasso
- NVIDIA NeMo and BioNeMo are in early access; NVIDIA Picasso is in private preview



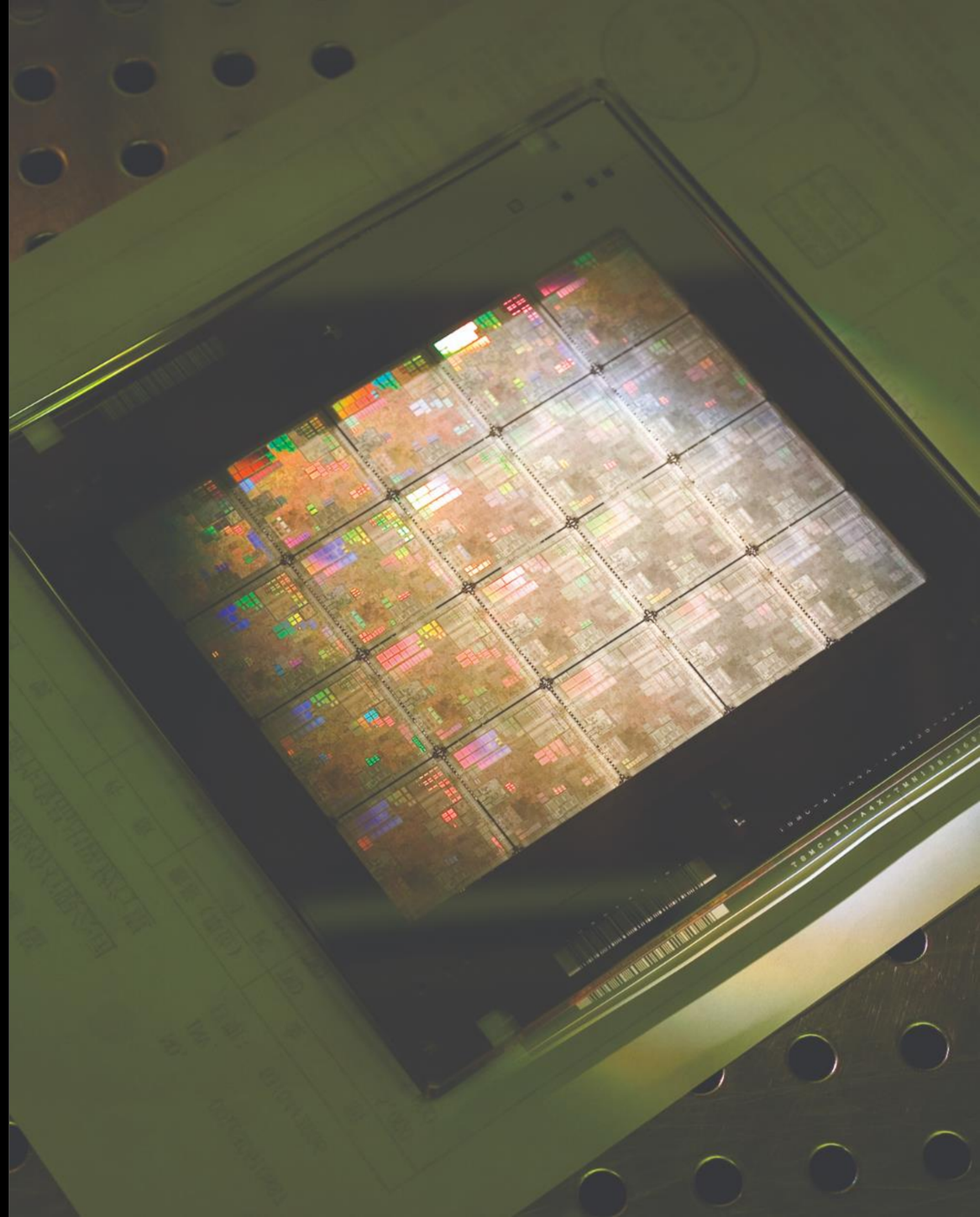
NVIDIA Omniverse Cloud to Power Industrial Digitalization

- NVIDIA Omniverse Cloud is a fully managed cloud service that enables companies to digitalize their workflows from design and engineering to smart factory to marketing
- Omniverse Cloud gives enterprises access to the full suite of Omniverse software applications and NVIDIA OVX infrastructure
- Microsoft Azure is the first CSP to host Omniverse Cloud, available now in private preview; Microsoft will also connect Microsoft 365 apps with Omniverse
- The auto industry has been an early adopter of Omniverse Enterprise; customers include BMW Group, Geely Lotus and Jaguar Land Rover
- Omniverse Cloud-based services will also be available from ecosystem partners including WPP, the world's largest marketing and communications company



NVIDIA cuLitho Software for Computational Lithography

- Computational lithography is the largest workload in chip design and manufacturing, consuming tens of billions of CPU hours annually
- The new NVIDIA cuLitho software library accelerates computational lithography by over 40X
- cuLitho would enable 500 NVIDIA DGX H100 systems to achieve the work of 40,000 CPU systems, reducing power from 35MW to just 5MW
- It is being integrated by TSMC, Synopsys and ASML into their manufacturing processes, software, and systems



NVIDIA and ServiceNow Partner to Build Generative AI Across Enterprise IT

- Partnership to develop enterprise-grade generative AI capabilities that can transform business processes with faster, more intelligent workflow automation
- Using NVIDIA software, services and accelerated infrastructure, ServiceNow is developing custom LLMs trained on data specifically for its ServiceNow Platform
- Expands ServiceNow's already extensive AI functionality with new uses for generative AI across the enterprise
- ServiceNow is also helping NVIDIA streamline its IT operations, using NVIDIA data to customize NVIDIA NeMo foundation models running on hybrid-cloud infrastructure





NVIDIA GeForce RTX 4060 and 4060 Ti

- New GeForce RTX 4060 and 4060 Ti GPUs deliver RTX technology and Ada Lovelace architecture to core gamers, starting at $299
- Include DLSS 3 and third-generation RTX technologies
- For the first time, provide 2x the performance of the latest gaming consoles at mainstream price points
- Access to the 300+ games and applications that now support DLSS
- 4060 Ti available now; 4060 available in July



NVIDIA Grace Drives Wave of New Energy-Efficient Arm Supercomputers

- Announced the Isambard 3 supercomputer at the Bristol & Bath Science Park in the U.K., the latest energy-efficient supercomputer based on Arm CPUs
- Isambard 3 will feature 384 NVIDIA Grace CPU Superchips to power medical and scientific research
- Expected to deliver 6x the performance and energy efficiency of Isambard 2, placing it among Europe's most energy-efficient systems
- ~2.7 petaflops of FP64 peak performance and consume less than 270 kilowatts of power, ranking it among the world's three greenest non-accelerated supercomputers
- Joins a growing list of NVIDIA Arm-based supercomputers. Systems at the Swiss National Supercomputing Centre and Los Alamos National Lab



Dell Technologies and NVIDIA Introduce Project Helix for Secure, On-Premises Generative AI

- Joint initiative between Dell and NVIDIA to make it easier for businesses to build and use trustworthy generative AI models on premises
- Quickly and securely deliver better customer service, market intelligence, enterprise search, and other capabilities
- Project Helix will deliver a series of full-stack solutions with technical expertise and pre-built tools based on Dell and NVIDIA infrastructure and software
- Includes complete blueprint to help enterprises use their proprietary data and more easily deploy generative AI responsibly and accurately
- Available through channel partners and APEX flexible consumption options beginning in July 2023



NVIDIA Collaborates With Microsoft to Accelerate Enterprise-Ready Generative AI

- NVIDIA is integrating its NVIDIA AI Enterprise software into Microsoft's Azure Machine Learning to help enterprises accelerate their AI initiatives
- NVIDIA AI Enterprise integration with Azure Machine Learning is available in a limited technical preview; also available on Azure Marketplace
- At the Microsoft Build developer conference, NVIDIA also showcased how Windows PCs and workstations with NVIDIA RTX GPUs will be AI-powered
  - End-to-end software engineering –from the Windows operating system to the NVIDIA graphics drivers and NeMo LLM framework – to make Windows on NVIDIA RTX Tensor Core GPUs a supercharged platform for generative AI





NVIDIA Overview



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 74% of the TOP500 supercomputers, and boasts over 4 million developers.

Headquarters: Santa Clara, CA



Generative AI Unlocks New Opportunities

How has NVIDIA contributed to the advancement of AI?

530B

NVIDIA has pioneered the art and science of visual computing and AI (Artificial Intelligence). NVIDIA GPUs—the heart of deep learning—are central to AI. They power self-driving cars, intelligent machines and robots, and extraordinary scientific discovery. NVIDIA is building the future of AI and the GPU is at its core.*

*Generated using NVIDIA NeMo service

TEXT GENERATION



Summarization

GPT-3

Marketing Copy

TRANSLATION



Translating Wikipedia

NLLB-200

Real-Time Metaverse Translation

CODING



Dynamic Code Commenting

CODEX

Function Generation

IMAGE GENERATION

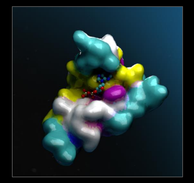


Brand Creation

eDiff-I

Gaming Characters

LIFE SCIENCE



Molecular Representations

MegaMolBART

Drug Discovery

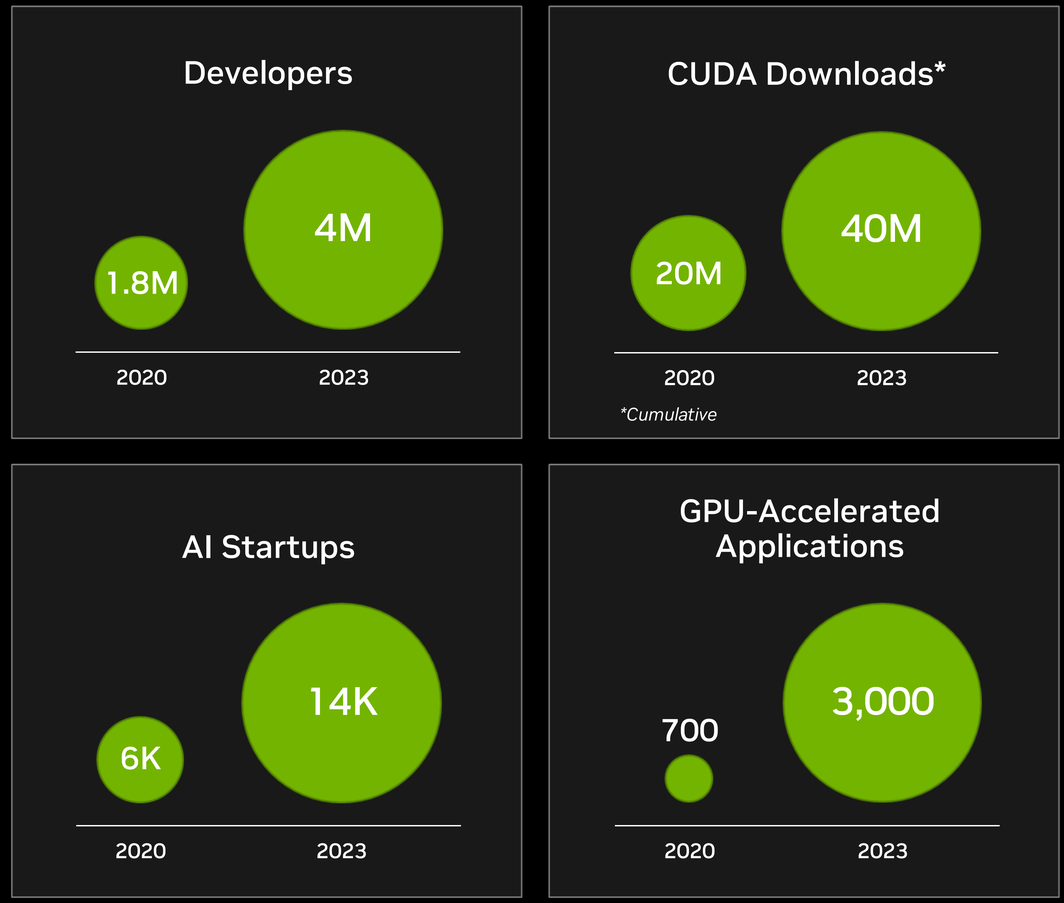
NVIDIA

NVIDIA’s Expanding Accelerated Computing Ecosystem

300 Libraries

400 AI Models

100 Updated in the Last Year



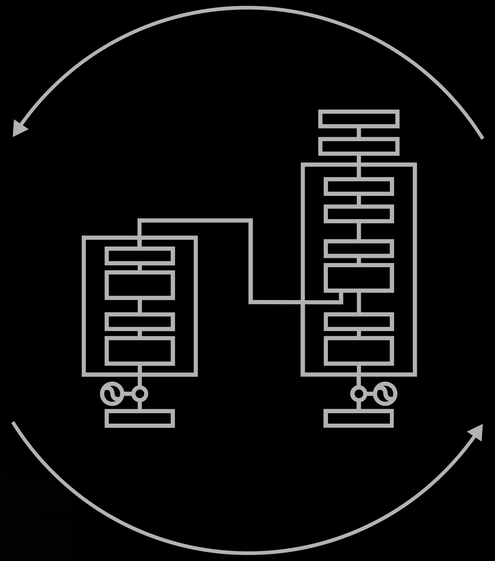
Training & Inference — One Architecture

Cloud | On-Prem | Edge

TRAINING

AI FACTORIES





INFERENCE

IN THE DATA CENTER

NVIDIA L40
Image Generation

NVIDIA L4
AI Video

NVIDIA H100 NVL
Large Language Models

NVIDIA Grace Hopper
Recommendation Models

AT THE EDGE

IGX
Industrial-Grade System
for Healthcare, Logistics,
Manufacturing

AGX
Functionally-Safe System
for Autonomous Vehicles

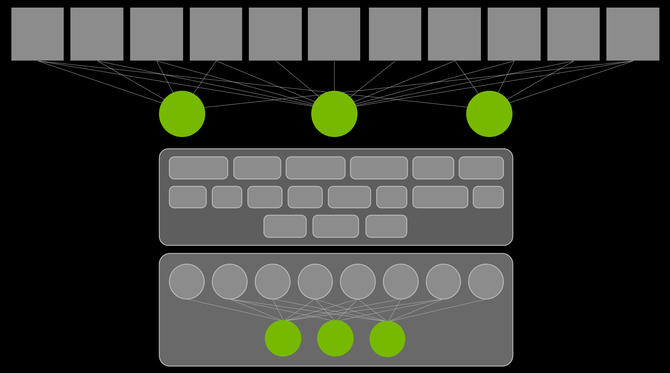
What Is Accelerated Computing?

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.

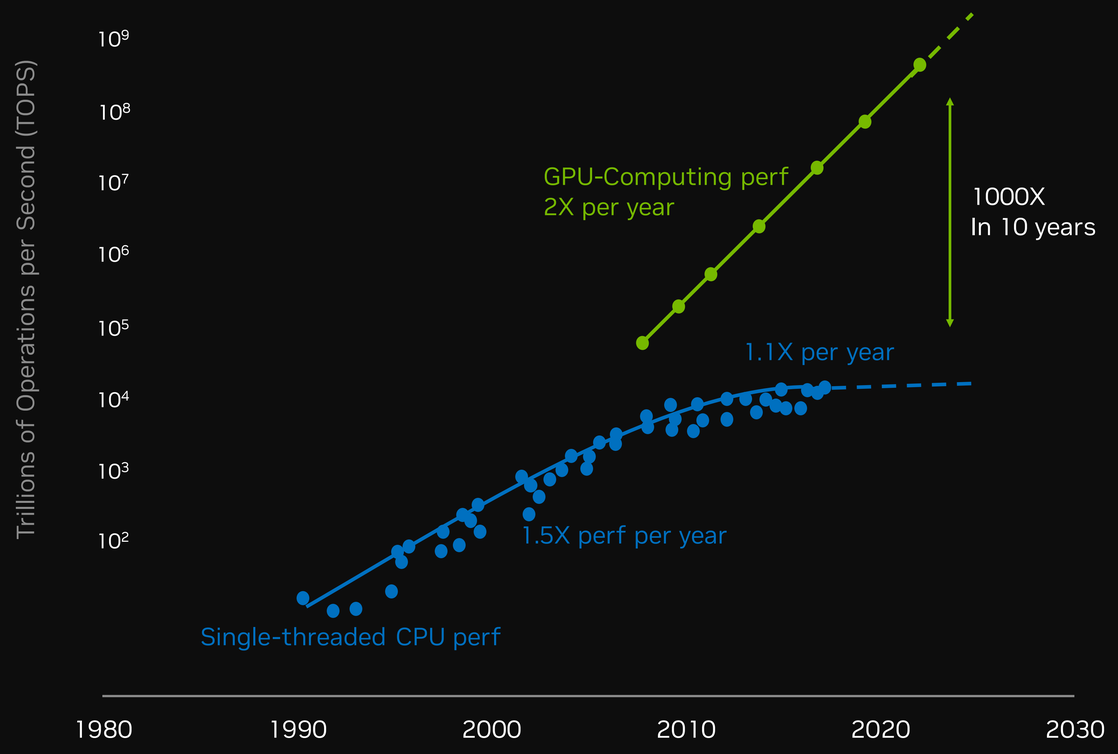
NVIDIA is uniquely dedicated to accelerated computing —working top-to-bottom—refactoring applications and creating new algorithms, and bottom-to-top—inventing new specialized processors, like RT Core and Tensor Core.

"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



AI Is the Greatest Technology Force of Our Time

Data centers across industries will become AI factories

AI has fundamentally changed what software can make and how you make software.

Companies are processing & refining their data, making AI software—becoming intelligence manufacturers. Their data centers are AI factories.

The first wave of AI is learned perception and inference, like recognizing images, understanding speech, recommending a video, or an item to buy.

The next wave of AI is robotics—AI planning actions. Digital robots, avatars, and physical robots will perceive, plan and act.

NVIDIA's acceleration stacks and ecosystems help bring AI to the world's largest industries.

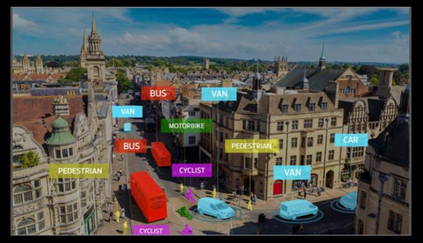
NVIDIA's world-class AI expertise and scale can help revolutionize businesses.



Contact Center AI
500M Calls / Day



Meeting Transcription
3B Meeting Minutes / Day



Public Safety
>1B Smart City Cameras Deployed



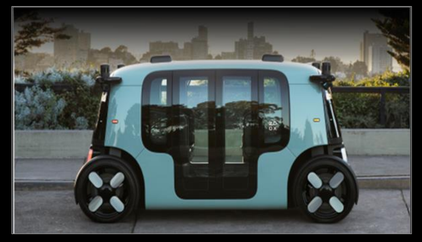
Retail Asset Protection
$94.5B Inventory Loss / Year



Medical Imaging
10M Diagnostic Scans / Day



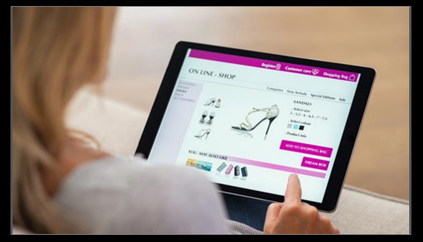
Industrial Inspection
$32M Vision Sensors Installed by 2025



Transportation
10T Miles / Year



Credit Card Fraud
1.28B Credit Transactions / Day



Product Recommendations
1B E-Commerce Visitors / Day

Source: Nilson Report, IHS Markit, Similar Web, NRF, WHO, ABI and NVIDIA internal analysis

NVIDIA

Building and Operating Metaverse Applications Is the Next Wave

NVIDIA Omniverse—Runs on NVIDIA OVX servers | RTX workstations | Enterprise software | Cloud services

NVIDIA Omniverse is a software platform for building and operating metaverse applications.

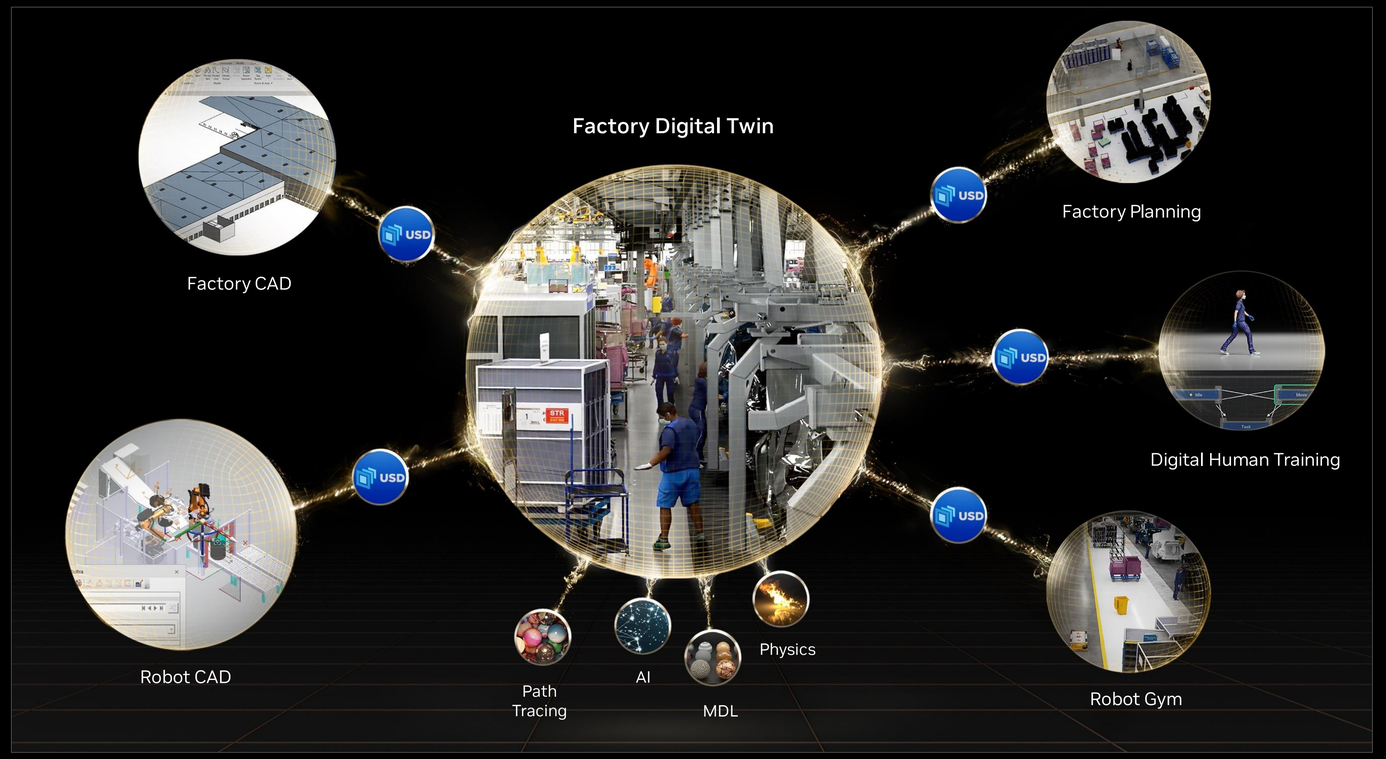
Our initial focus is on industrial metaverses, such as digital twins used to emulate the behavior of products or factories in the physical world.

Omniverse uses a real-time, large-scale 3D database that connects to 3D worlds via the USD (Universal Scene Descriptor) framework.

Just as the internet connects websites over HTML, Omniverse connects 3D worlds over USD.

Omniverse is essential for the next wave of AI—robotics—where AI interacts with the physical world.

Applications built to run on Omniverse are like portals into the Omniverse virtual world.



Product Development

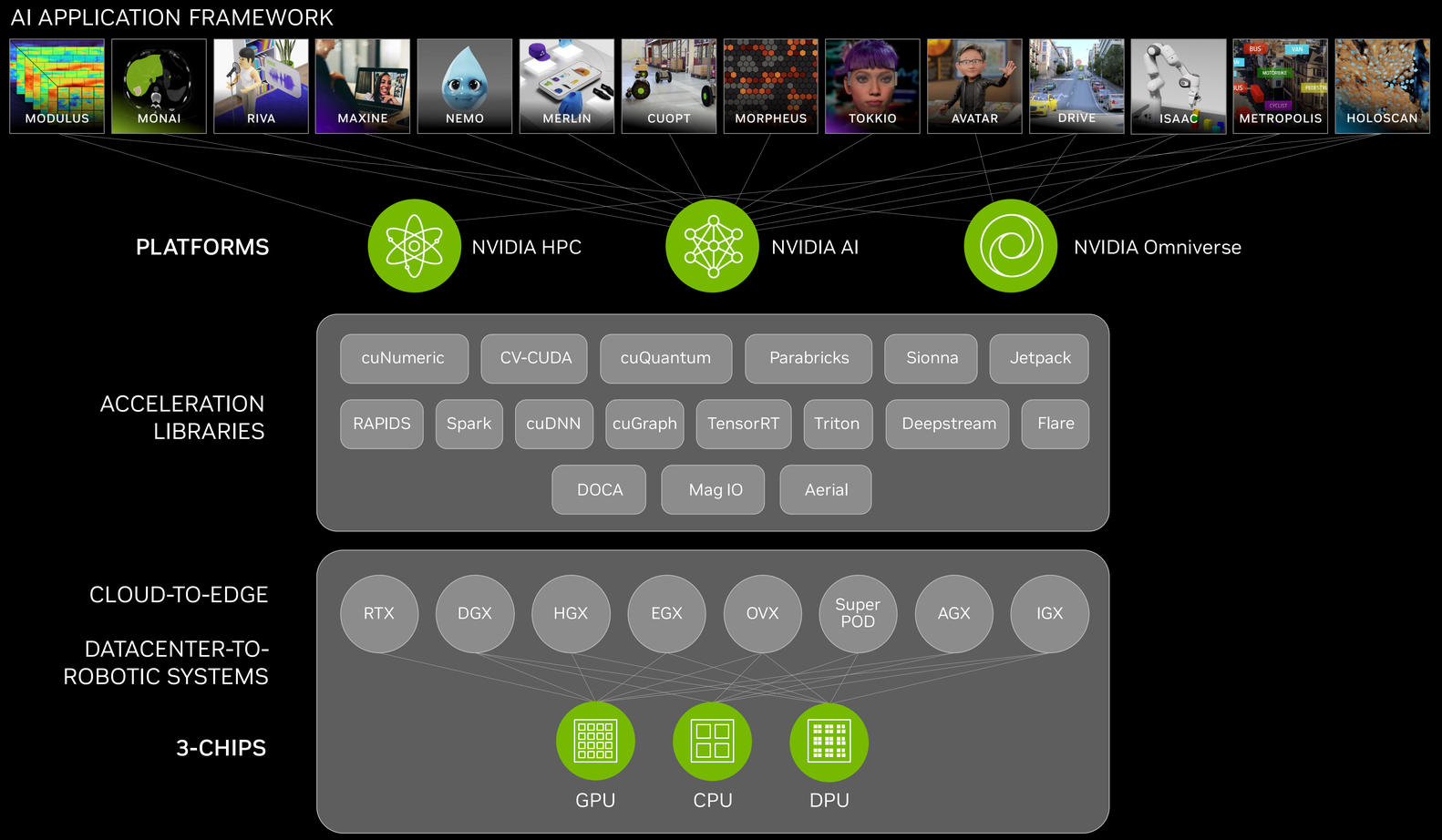
Connecting 3D Creators & AI Assistants in Virtual Worlds

Operate Digital Twin

Connecting Robots in a Virtual World

NVIDIA’s Accelerated Computing Platform

Full-stack innovation across silicon, systems and software



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

Accelerated computing is a full-stack challenge, demanding deep understanding of the problem domain, optimizing across every layer of computing, and all three chips —GPU, CPU, and DPU.

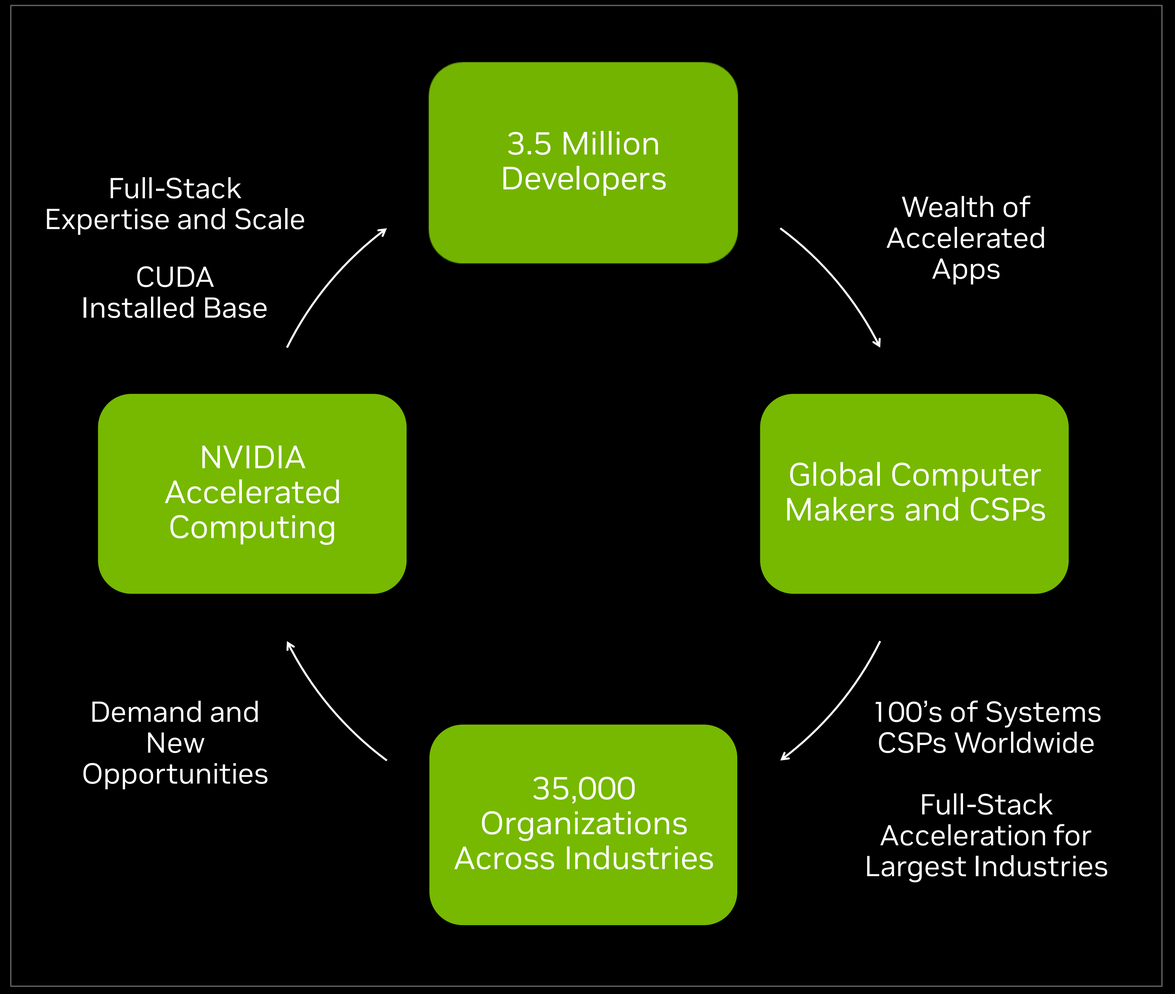
Scaling across multi-GPUs and multi-nodes is a data center-scale challenge and requires treating the network and storage as part of the computing fabric.

Our platform extends from PCs to supercomputing centers, enterprise data centers, cloud and edge environments.

NVIDIA's Multi-Sided Platform and Flywheel

NVIDIA is valued by every stakeholder in the ecosystem:

- **For developers** – NVIDIA's One Architecture and large installed base give developer's software the best performance and greatest reach
- **For computer makers and CSPs** – 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 customers** – NVIDIA is offered by virtually every computing provider and accelerates the most impactful applications from cloud to edge
- **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



Full-Stack & Data Center Scale Acceleration

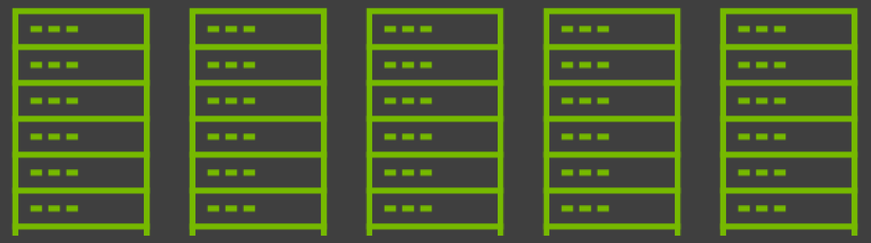
Drive Significant Cost Savings and Workload Scaling

Classical Computing—92 CPU-only servers

$3.3M (including switches, cables, racks)

Application

CPU server racks



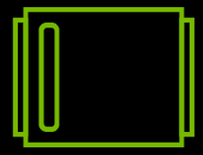
Accelerated Computing—1 NVIDIA DGX A100

$220,000 DGX and $100,000 NVIDIA AI software

Application
Re-Engineered for Acceleration

CUDA-X Acceleration Libraries

Magnum IO



DGX

10X lower cost
14X better energy-efficiency

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.

NVIDIA Software and Services

Enabling the world's enterprises to revolutionize industries with AI

NVIDIA-hosted cloud service for training Large Language Models to perform specific task s— e.g., summarize legal documents, write marketing copy, analyze market sentiment, chatbot to support customers, search documents, write and document code, paraphrase.

Nemo can help thousands of companies, train language AI's to do hundreds of tasks, in 10's of languages.

NVIDIA NeMo LLM

NVIDIA-hosted cloud service for training and deploying large biomolecular models that understand the language of chemistry, proteins, RNA, and DNA.

BioNemo can help researchers, biotech, and pharma companies to process chemical and biological datasets to accelerate drug discovery.

NVIDIA BioNeMo

NVIDIA-hosted cloud service for building generative AI–powered visual applications.

Enterprises, software creators, and service providers can run inference on their models, train NVIDIA Edify foundation models on proprietary data, or start from pretrained models to generate image, video, and 3D content from text prompts.

Picasso service is fully optimized for GPUs and streamlines training, optimization, and inference.

NVIDIA Picasso

The operating engine of AI for end-to-end data-driven software development.

One engine license accelerates end-to-end modern AI and data science.

One engine license unlocks wealth of data processing, AI, and robotics frameworks and applications — e.g., RAPIDS, Spark, Merlin, Monai, Metropolis, cuOpt, Morpheus, Tokkio.

NVIDIA AI Enterprise

A platform for designing, building, and operating 3D and virtual world simulations.

Consists of a virtual world engine, USD connectors, and portals browsing the virtual world simulation.

Omniverse is an enterprise application that connects architects, designers, hardware and software engineers, marketers, to supply-chain and factory planners.

NVIDIA Omniverse

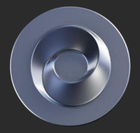
NVIDIA Cloud Services

Engaging with customers at every layer

CUSTOM AI MODEL MAKING SERVICE

NEMO | PICASSO | BIONEMO

PLATFORM-AS-A-SERVICE

NVIDIA AI   NVIDIA Omniverse

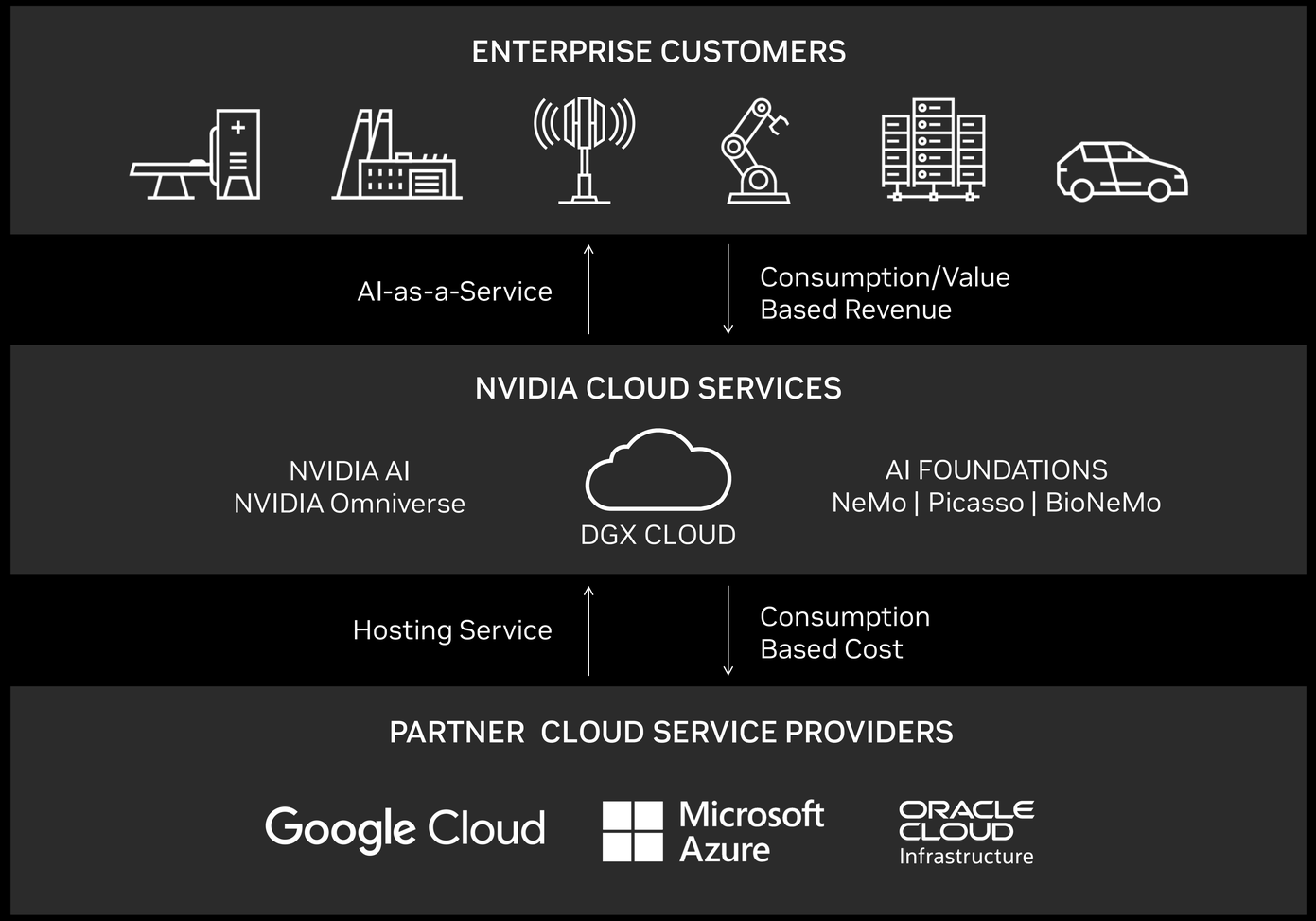
AI INFRASTRUCTURE-AS-A-SERVICE

ON PREM  DGX

HYBRID CLOUD 

MULTI CLOUD

NVIDIA’s Cloud Business Model



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

Reaching customers everywhere

	CLOUD	ON-PREM
NVIDIA	DGX Cloud (Google Cloud, Microsoft Azure, Oracle Cloud Infrastructure) NeMo | Picasso | BioNeMo	DGX
PARTNERS	aws, Google Cloud, Microsoft Azure, Oracle Cloud Infrastructure HGX, INFERENCE	Dell Technologies, Hewlett Packard Enterprise, Lenovo EGX, AGX, IGX

Giant Market Opportunity

Gaming & Metaverse | Financial Services | Healthcare | Logistics | Manufacturing | Retail | Transportation



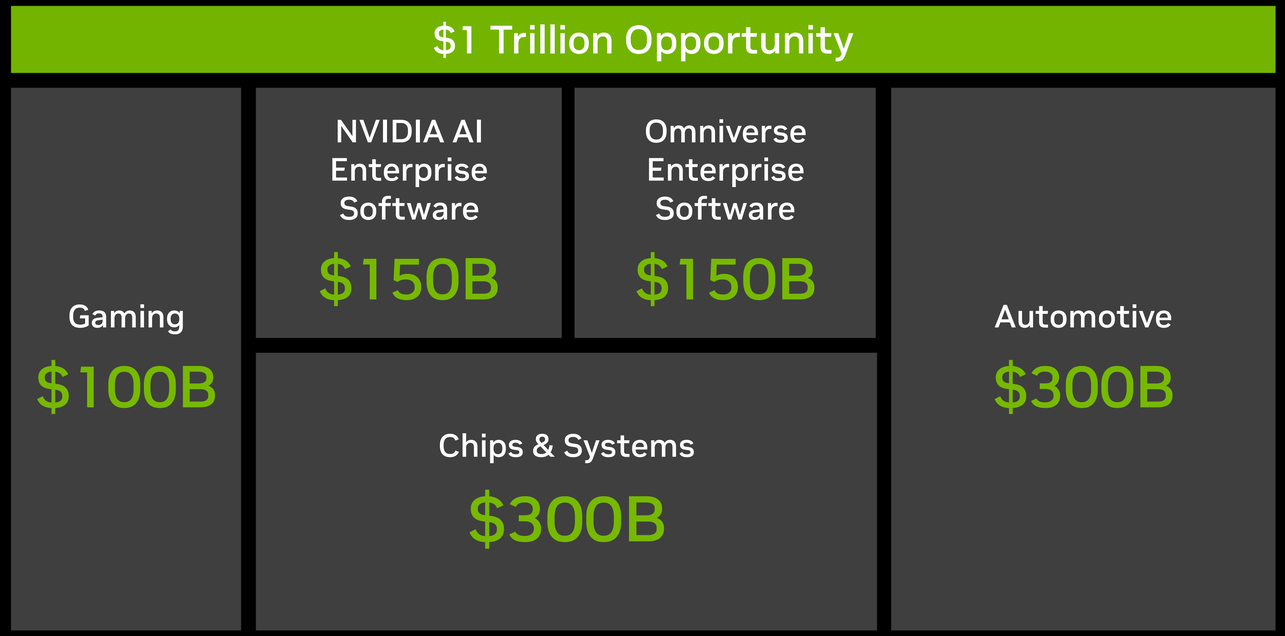
Gaming
Over 3B gamers and creators, a quarter of them spending over $100/year for GPUs in desktops, laptops, cloud or consoles

NVIDIA AI Enterprise Software
50M enterprise server installed base; per-server, per-year subscription price

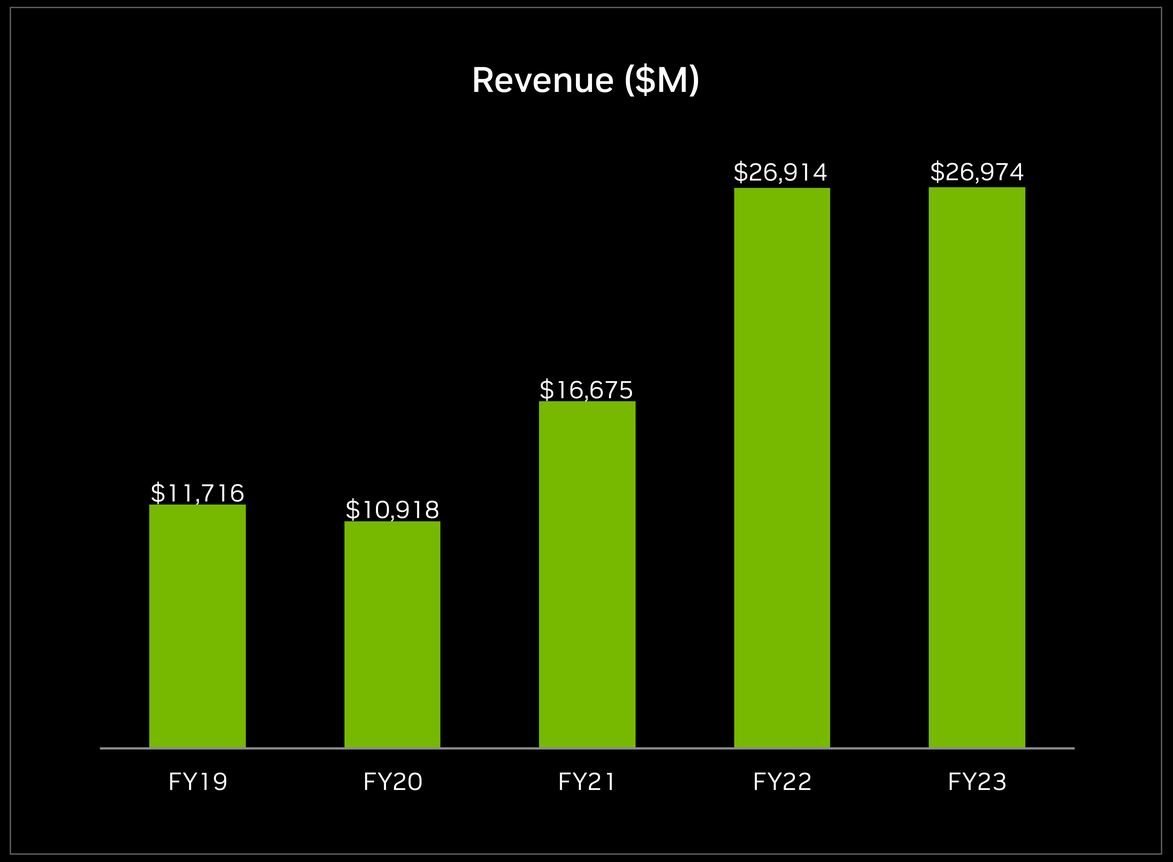
Omniverse Enterprise Software
Over 45M designers and creators; 10s of millions of digital twins —per-user/digital twin, per-year subscription price

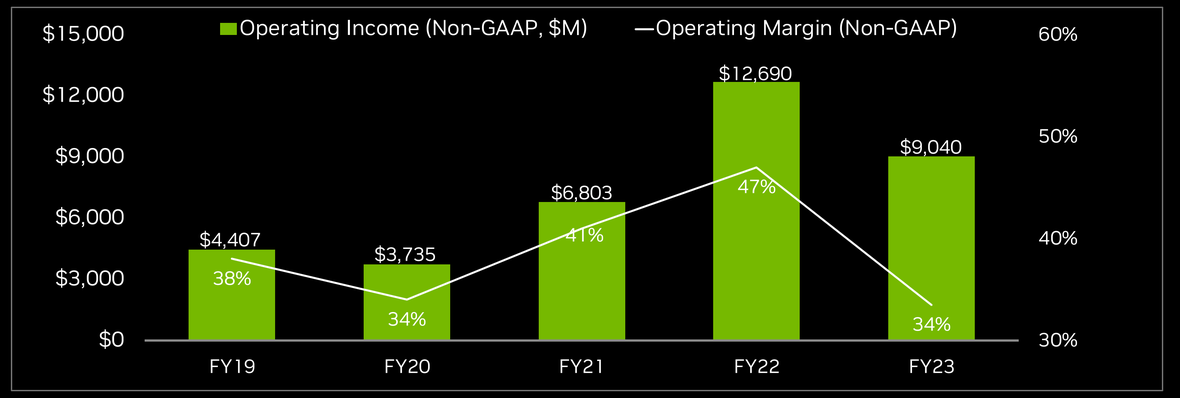
Chips and Systems
~20M servers/year—GPUs, CPUs, DPUs, NICs, switches

Automotive
100M vehicles/year hardware opportunity; 100s of millions of AV vehicles installed base software opportunity

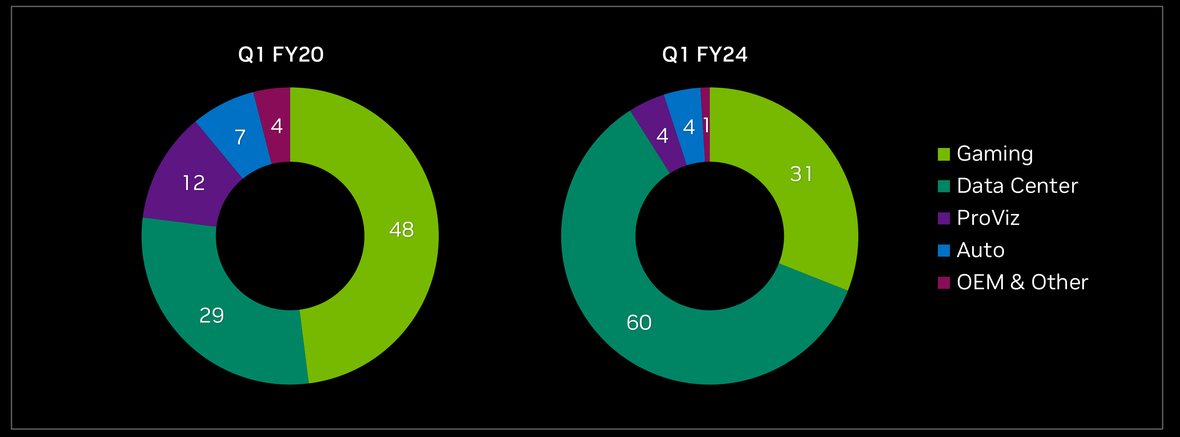


Driving Strong & Profitable Growth





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.

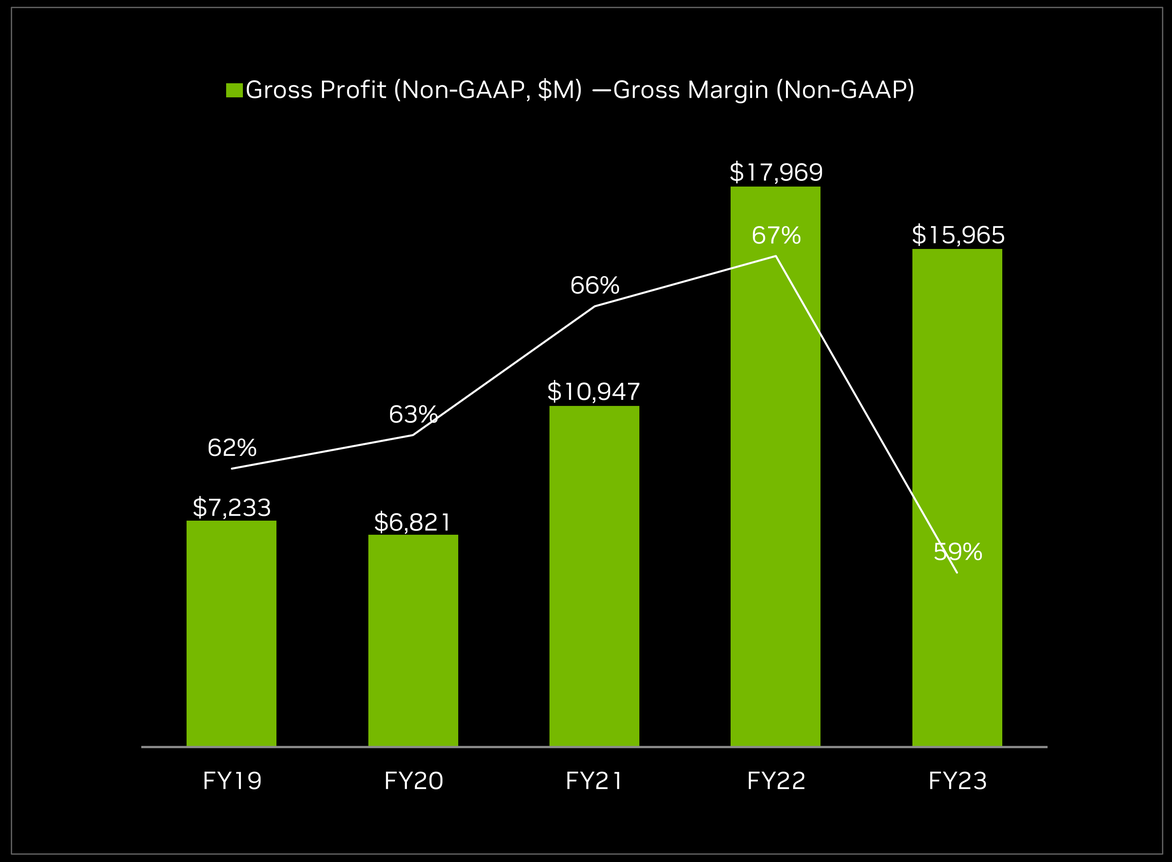
NVIDIA Gross Margins Reflect Value of Acceleration

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, 10 NVIDIA HGX nodes with 80 NVIDIA A100 GPUs that cost $4M can replace 920 nodes of CPU servers that cost over $50M for AI inference.

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)



Capital Allocation

Share Repurchase

Resumed Buybacks in Q1 FY 2023

$10.0B repurchased in FY23; $7.2B Remaining Authorization Through Dec 2023 as of Apr 30, 2023

Dividend

$398M in FY 2023

Plan to Maintain[1]

Strategic Investments

Growing Our Talent

Platform Reach & Ecosystem

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

[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/EGX/IGX systems

GPU | CPU | DPU | Networking
NVIDIA AI software



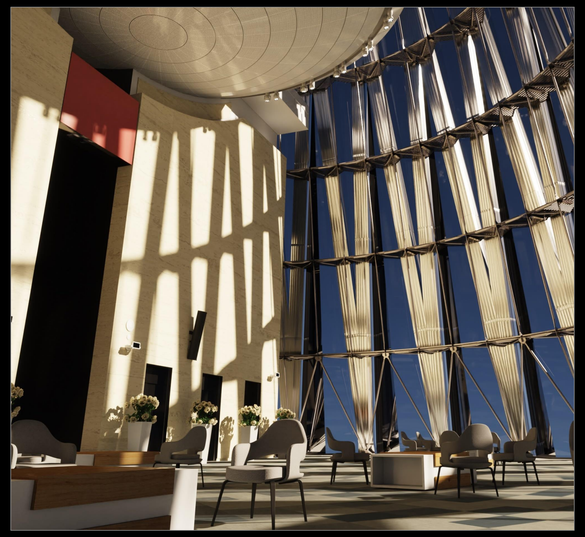
Gaming

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%

Quadro/NVIDIA RTX GPUs
for workstations

Omniverse software



Automotive

3% of FY23 revenue

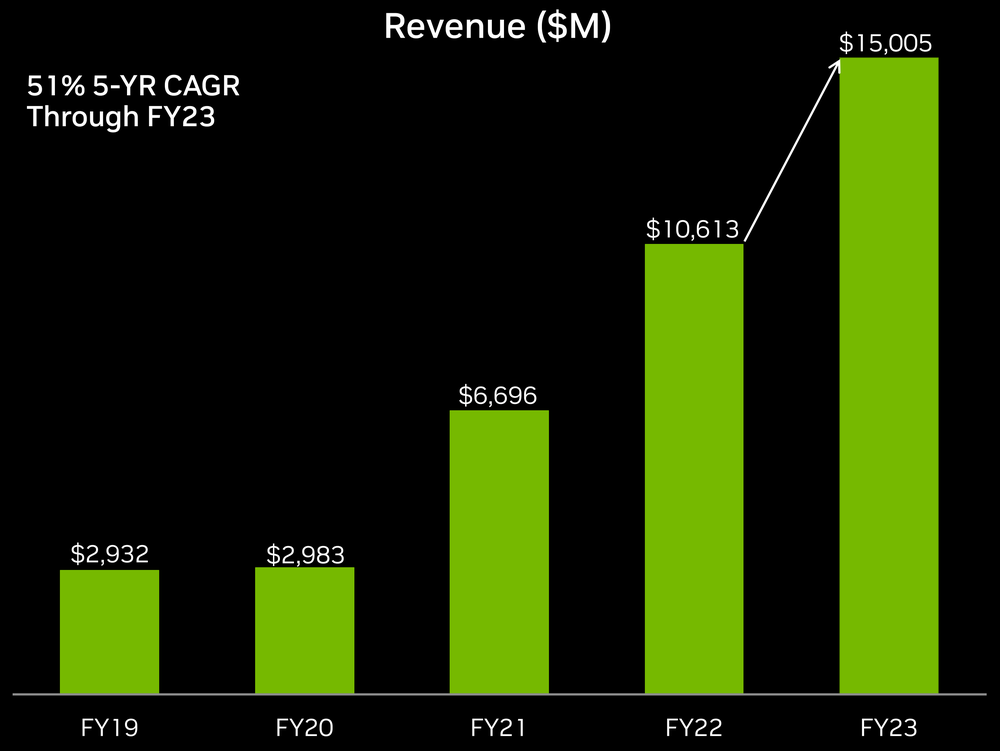
FY23 Revenue $0.9B
5-yr CAGR 10%

DRIVE Hyperion sensor architecture
with AGX compute

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

Data Center

The leading computing platform for AI, HPC & graphics



Leader in AI & HPC

#1 in AI training and inference

Used by all hyperscale & major cloud computing providers and 35,000 organizations

Powers 74% of the TOP500 supercomputers

Growth Drivers

Rapid AI adoption across industries

Full-stack AI | Software

Three chip strategy—GPU | CPU | DPU

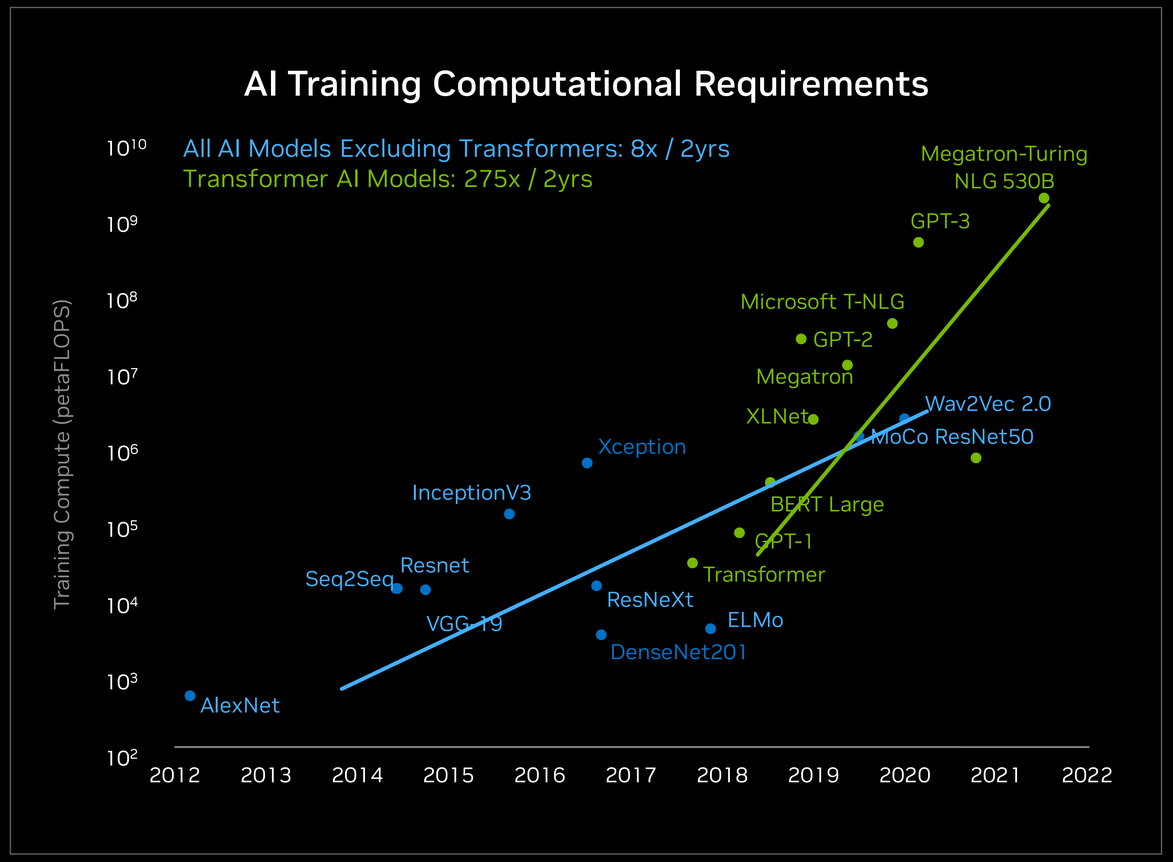
Rising computation requirements for modern AI

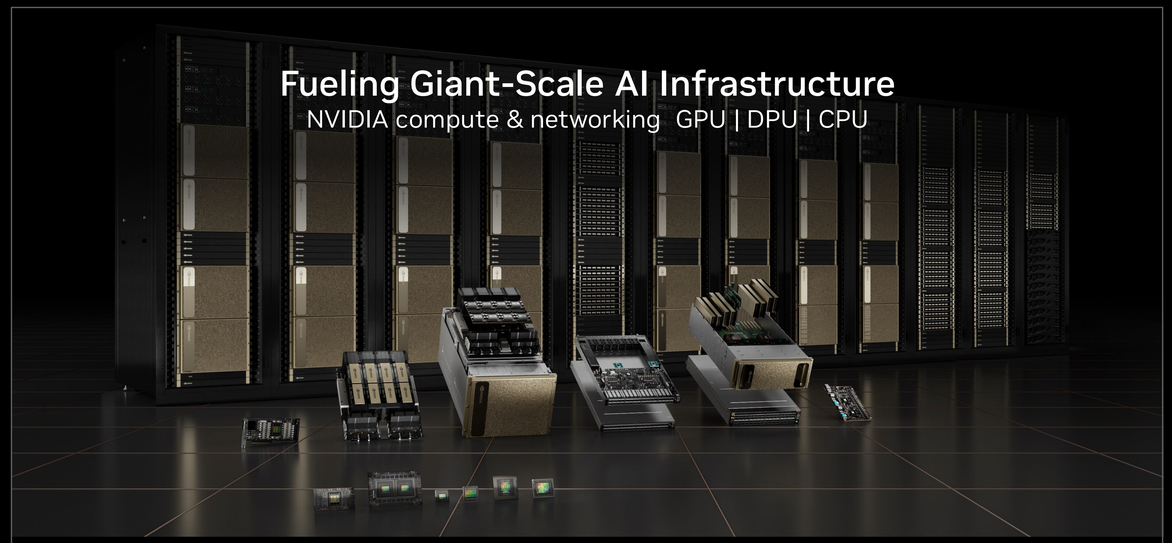
Data-center scale innovation

Omniverse

Modern AI is a Data Center Scale Computing Workload

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





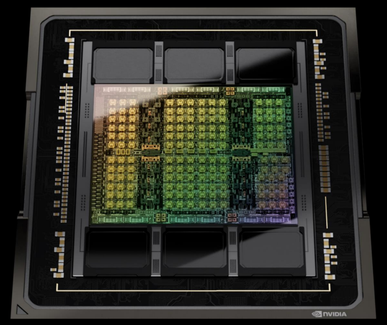
Large Language Models, based on the Transformer architecture, are one of today's most important advanced AI technologies, involving up to trillions of parameters that learn from text.

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.

Wave of New Data Center Products

Ramping new architectures for GPU, CPU and DPU

H100 GPU



World's Most Advanced Chip

80B Transistors

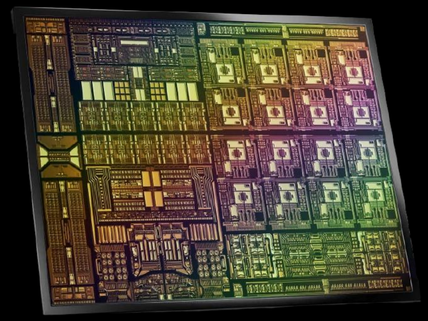
Transformer Engine – 6X Perf

Confidential Computing

4th Gen NVLink—7X PCIe Gen5

2H FY23

Bluefield-3 DPU



First 400 Gb/s DPU

Line-rate processing of software-defined networking, storage, and cybersecurity

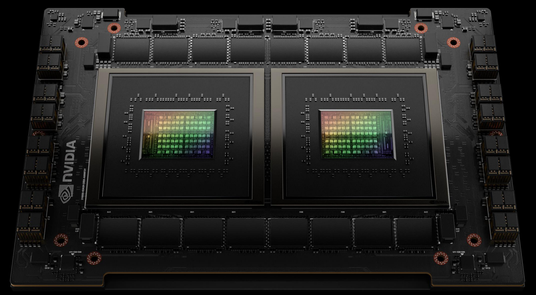
VMware vSphere 8 integration

Zero-trust security

~600 infrastructure software partners

2H FY24

Grace CPU Superchip



High Performance CPU for HPC and AI

144 Cores | 740 SPECrate®2017_int_base est.

1TB/s Memory Bandwidth

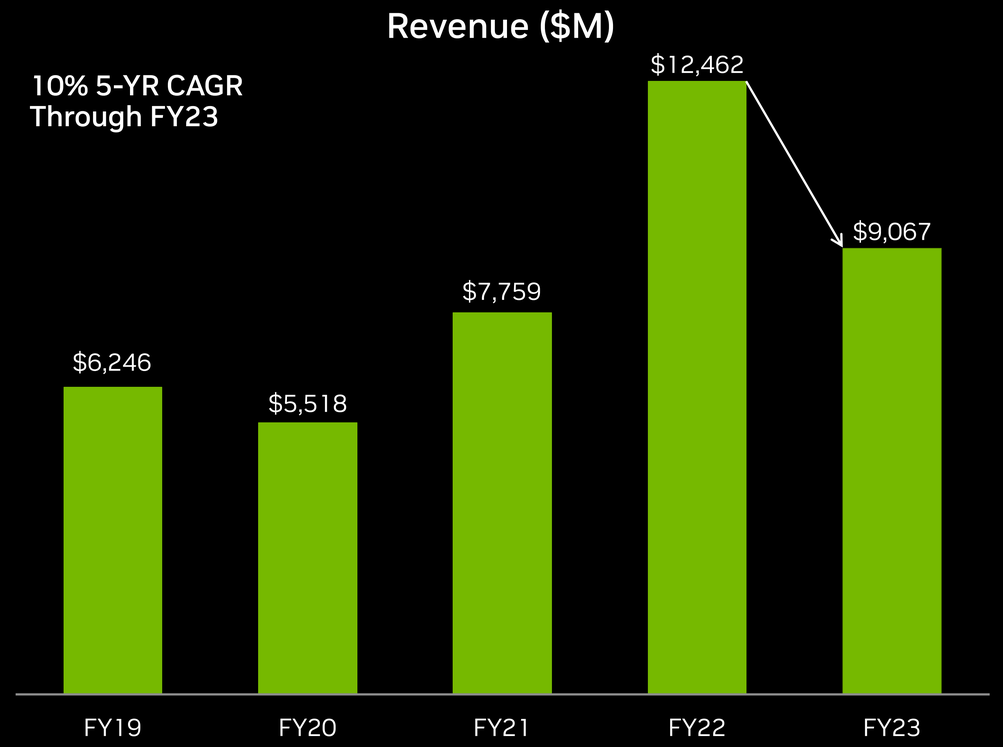
2X Perf/Watt Over Traditional Servers

Runs NVIDIA Computing Stacks

2H FY24

Gaming

GeForce—the world's largest gaming platform



Leader in PC Gaming

Strong #1 market position with over 80% share

15 of the Top 15 most popular GPUs on Steam

Leading performance & innovation

200M+ gamers on GeForce

Growth Drivers

Rising adoption of NVIDIA RTX

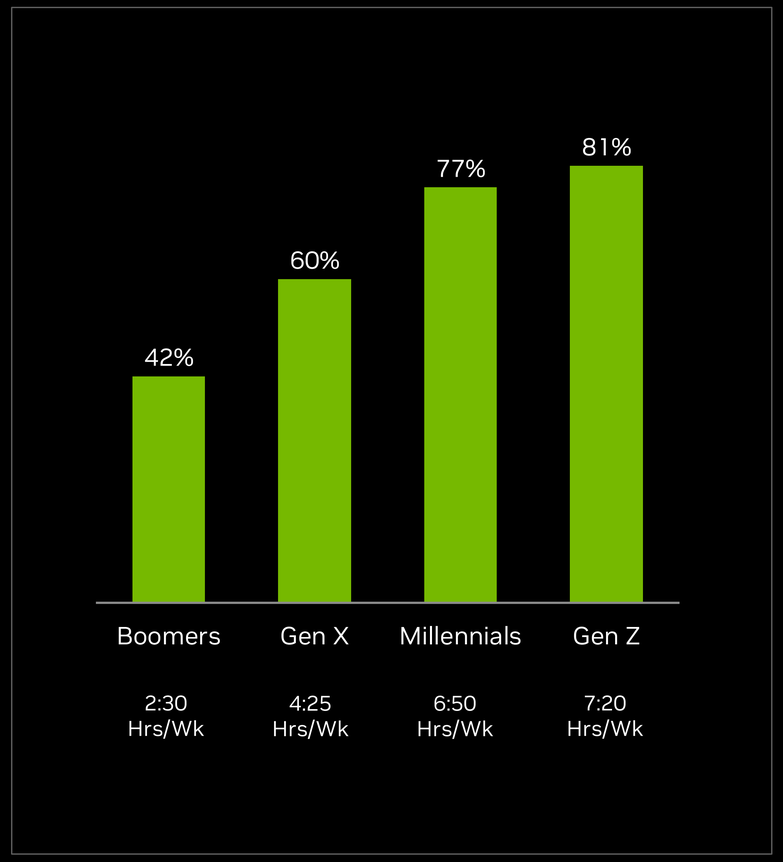
Expanding universe of gamers & creators

Gaming laptops & game consoles

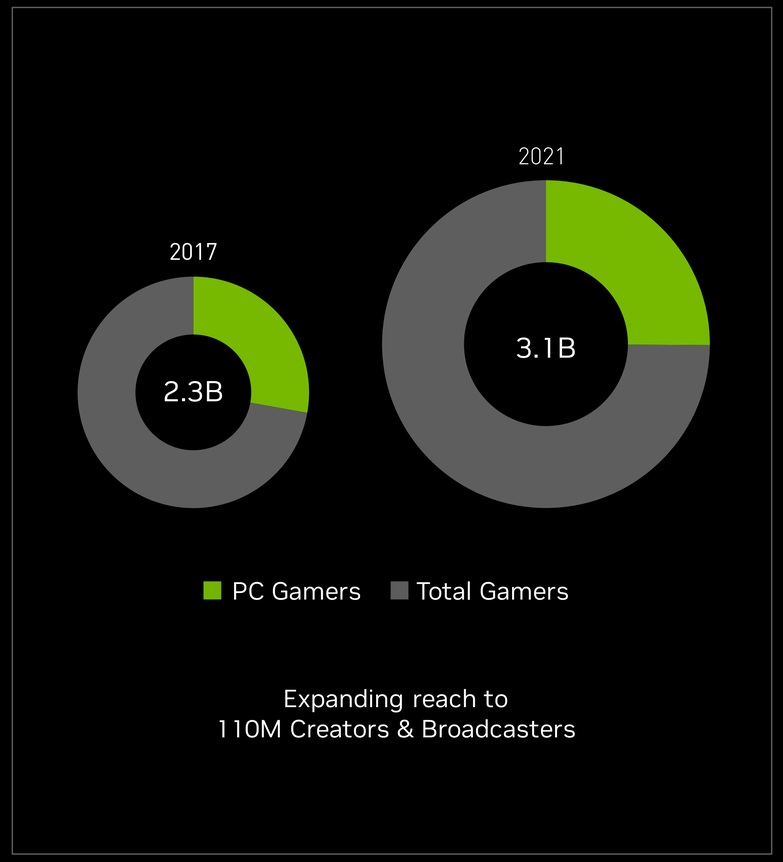
GeForce NOW Cloud gaming

Strong Gaming Fundamentals

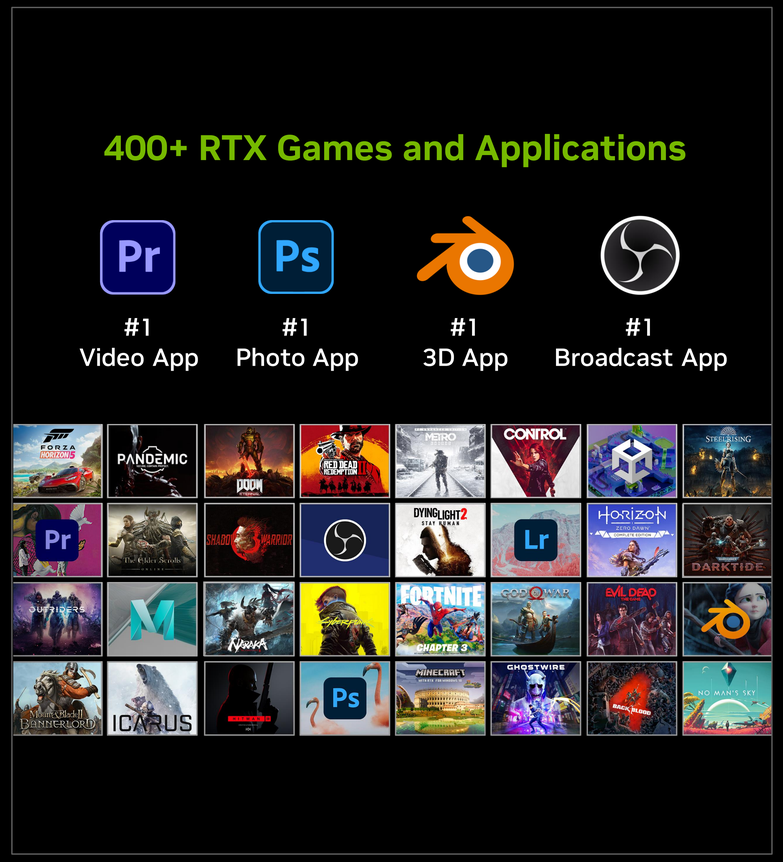
New generation, more gamers



New generation, more gamers



Expanding universe of gamers and creators

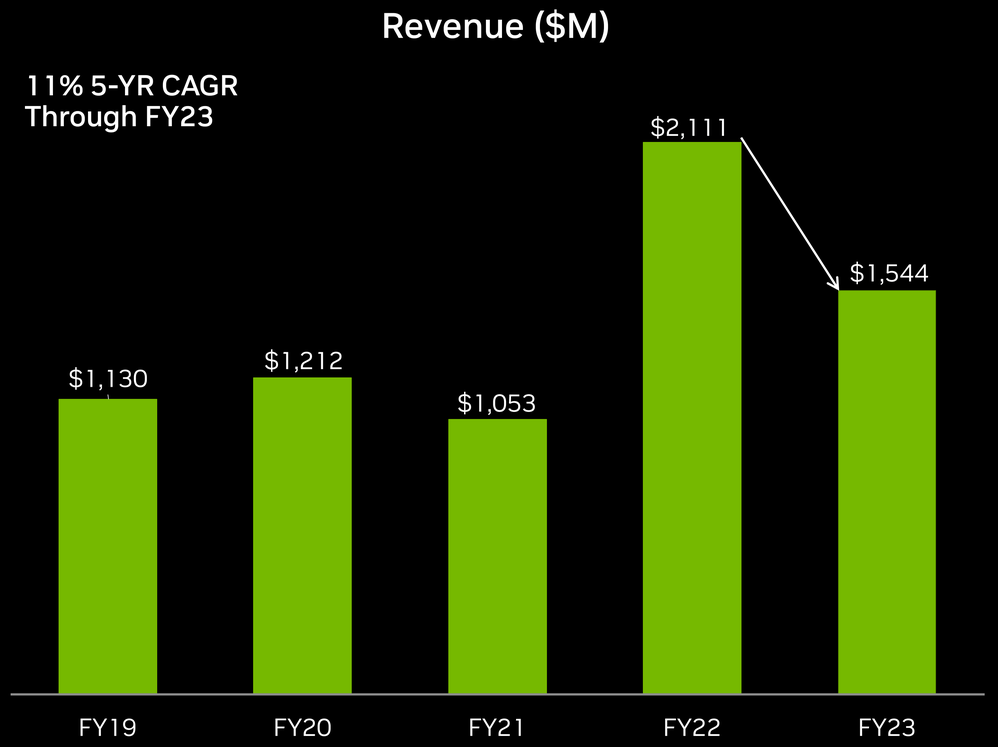


Robust NVIDIA ecosystem

Source: NewZoo and NVIDIA internal analysis

Professional Visualization

Workstation graphics



Leader in Workstation Graphics

90%+ market share in graphics for workstations

45M Designers and Creators

Strong software ecosystem with over 100 supported applications

Growth Drivers

Ray Tracing and AI revolutionizing design

Expanding universe of designers and creators

Collaborative 3D design / Omniverse

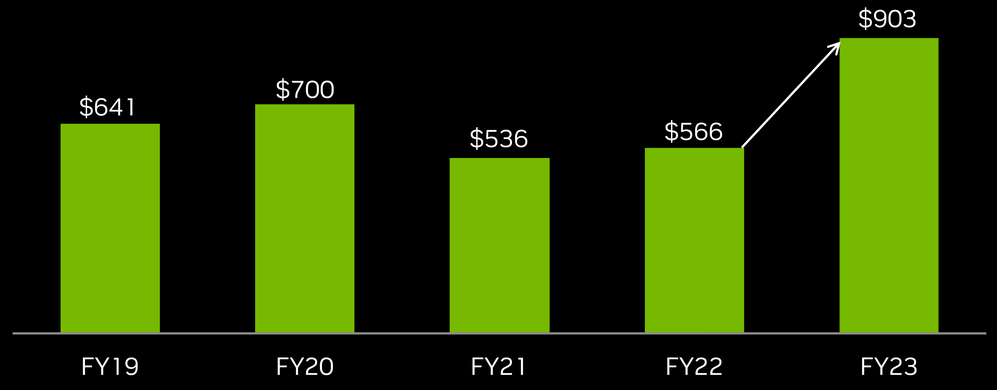
Hybrid work environments

Automotive

Autonomous Vehicles (AV) & AI Cockpit

Revenue ($M)

10% 5-YR CAGR
Through FY23



Leader in Autonomous Driving

Historical revenue driven largely by infotainment

Future growth primarily fueled by NVIDIA DRIVE, our AV and AI cockpit platform with full software stack

Over $14B design win pipeline through FY29 based on DRIVE Orin

Next-generation DRIVE Thor to ramp in FY26

Growth Drivers

Adoption of centralized car computing and software-defined vehicle architectures

AV software and services:

Mercedes Benz FY25 SOP*

Jaguar Land Rover FY26 SOP

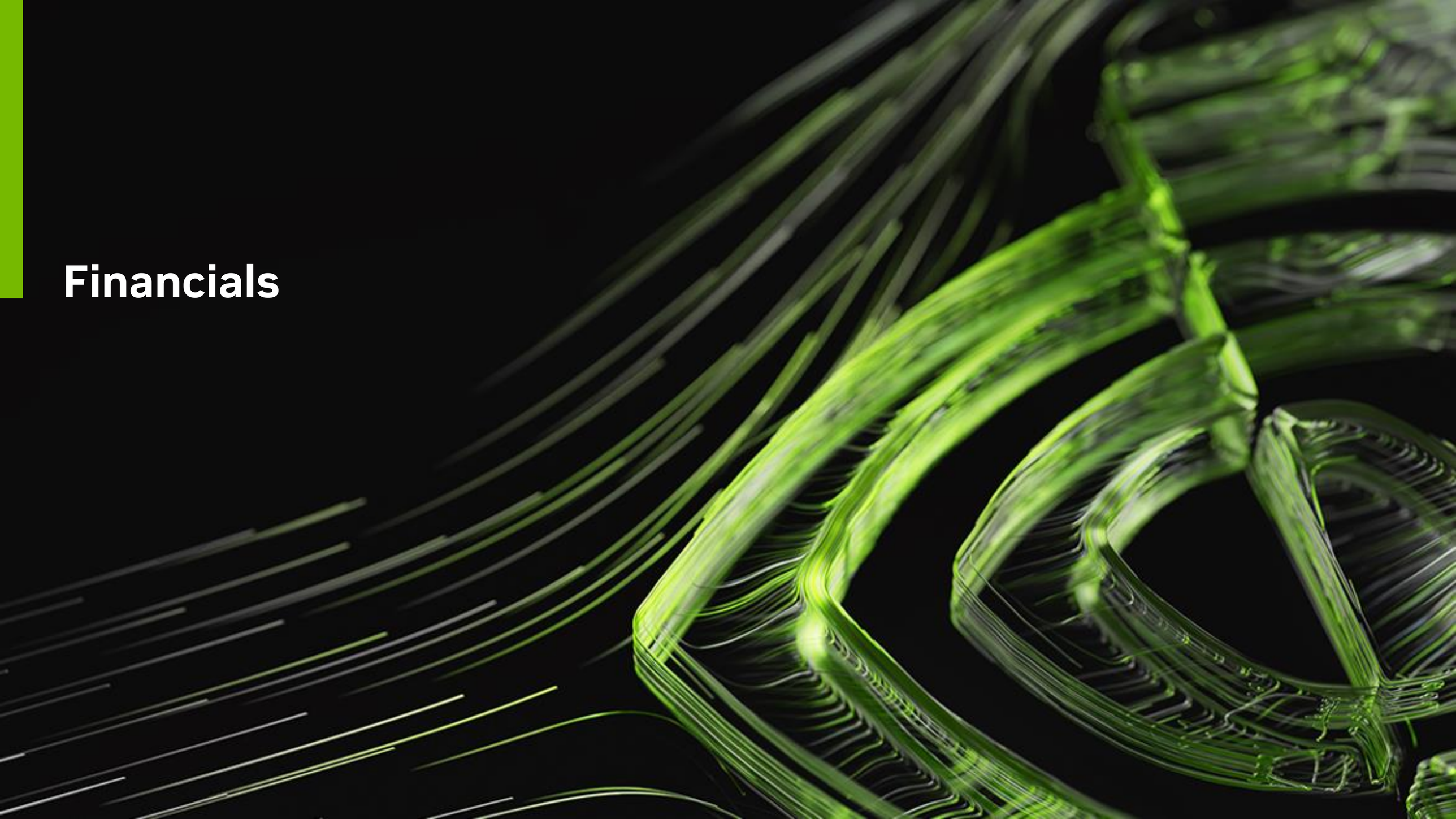
**SOP = Start of Production*

Growing NVIDIA DRIVE Pipeline

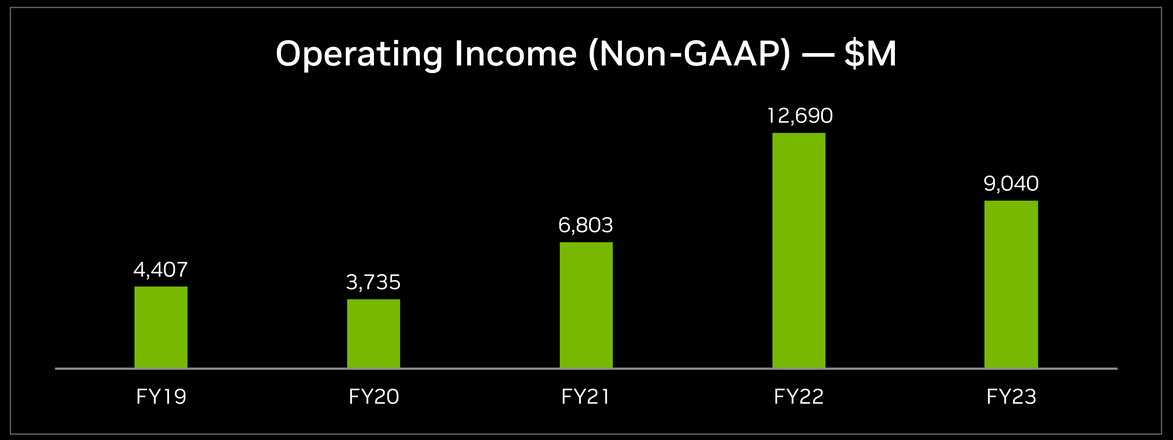
$14B design win pipeline — 6 year horizon

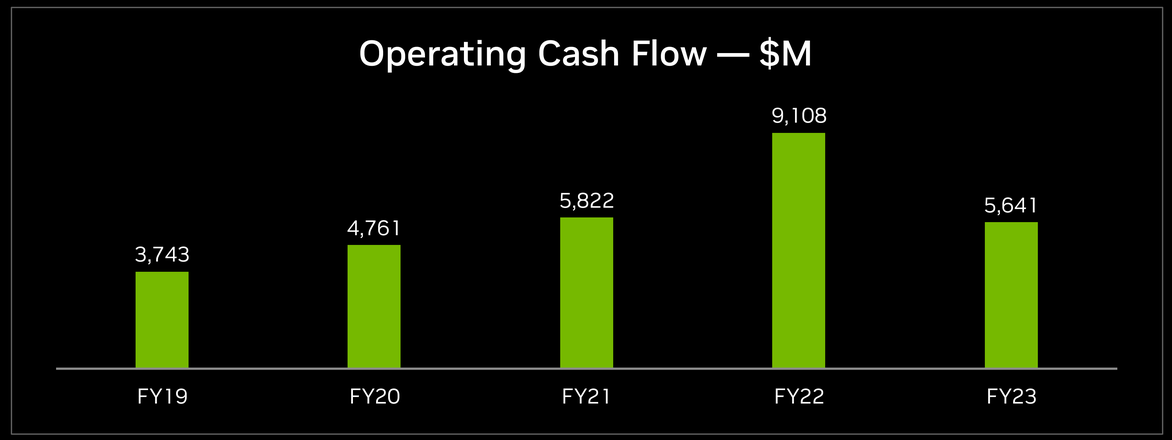


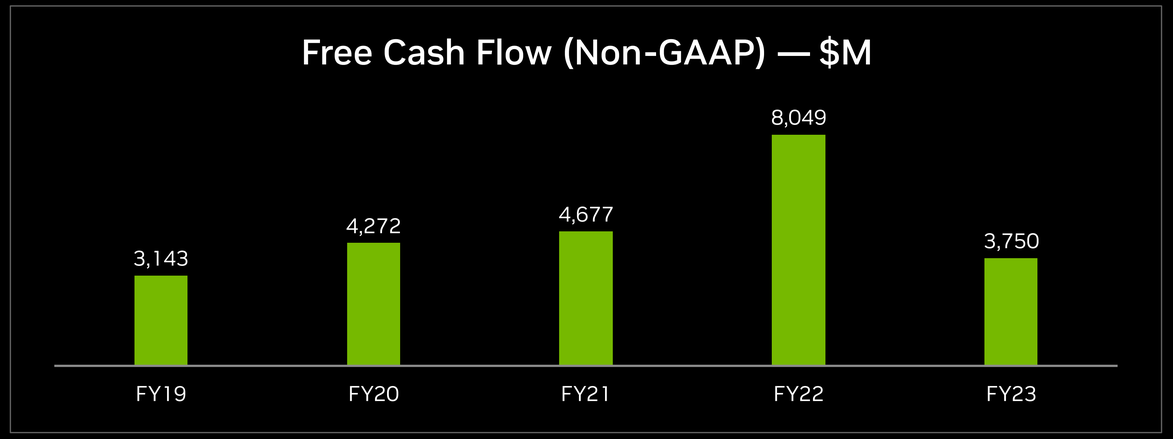
Financials

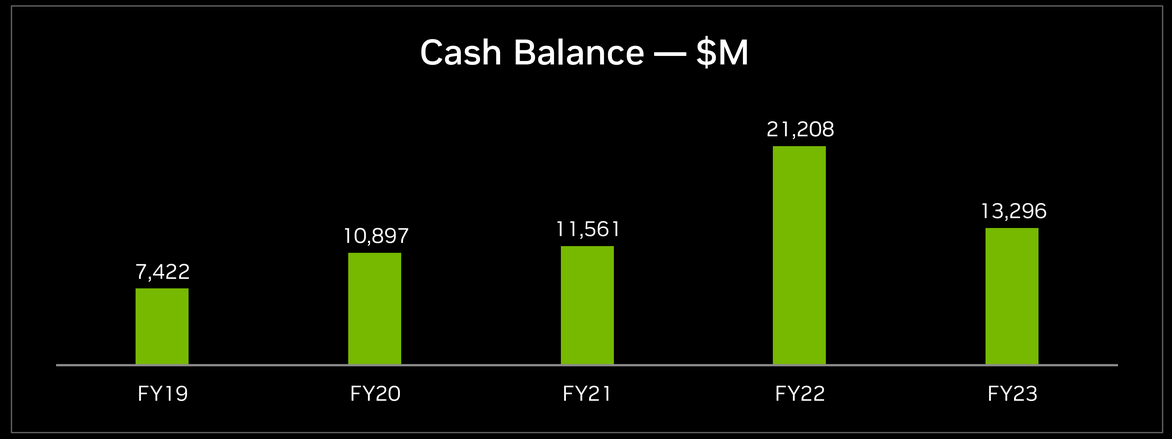


Annual Cash & Cash Flow Metrics









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

Corporate Responsibility

Environmentally Conscious



By FY26, 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



23 of Top 30 Supercomputers on the June 2023 Green500 powered by NVIDIA including the #1 system, Henri



Will achieve and 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"
FORBES

"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

38% Of Board is Gender, Racially, or Ethnically Diverse

92% 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 Cost	Other (C)	Tax Impact of Adjustments	GAAP
Q1 FY24							
Gross margin ($ in million)	$4,802	(119)	(27)	(8)	—	—	$4,648
	66.8%	(1.7)	(0.4)	(0.1)	—	—	64.6%
Operating income ($ in million)	$3,052	(173)	(735)	(8)	4	—	$2,140
Net income ($ in million)	$2,713	(173)	(735)	(8)	(11)	257	$2,043
Shares used in diluted per share calculation (millions)	2,490	—	—	—	—	—	2,490
Diluted EPS	$1.09	—	—	—	—	—	$0.82

A. Consists of amortization of intangible assets, transaction costs, and certain compensation charges.
B. Stock-based compensation charge was allocated to cost of goods sold, research and development expense, and sales, general and administrative expense.
C. Other comprises of assets held for sale related adjustments and losses from non-affiliated investments

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 Cost	GAAP
Q1 FY2023	67.1%	(1.1)	(0.5)	—	65.5%
Q2 FY2023	45.9%	(1.8)	(0.6)	—	43.5%
Q3 FY2023	56.1%	(2.0)	(0.5)	—	53.6%
Q4 FY2023	66.1%	(2.0)	(0.5)	(0.3)	63.3%

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%

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 Margin ($ in Millions & Margin Percentage)	Non-GAAP	Acquisition Termination Cost	Acquisition-Related and Other Costs (A)	Stock-Based Compensation (B)	IP-Related Cost	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%

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 costs, 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

Reconciliation of Non-GAAP to GAAP Financial Measures

($ in Millions)	Q2 FY24 Outlook
Non-GAAP gross margin	70.0%
Impact of stock-based compensation expense, acquisition-related costs, and other costs	(1.4%)
GAAP gross margin	68.6%
Non-GAAP operating expenses	$1,900
Impact of stock-based compensation expense, and acquisition-related costs, and other costs	810
GAAP operating expenses	$2,710


NVIDIA