NVIDIA

Revolutionizing Artificial Intelligence and Financial Services:
A dive into NVIDIA's Technological Impact

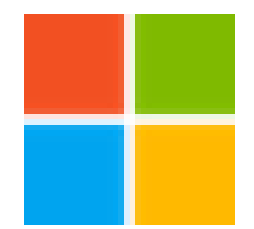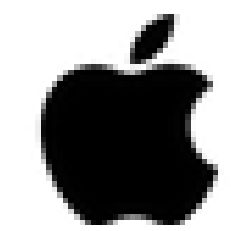Marc Stampfli, May 2024

E-Mail: mstampfli@nvidia.com
Linkedin: https://www.linkedin.com/in/marcstampfli

Our Body of Work

NVIDIA pioneered **accelerated computing to tackle challenges no one else can solve**.

Our work in AI and digital twins is transforming the world's largest industries and profoundly impacting society.

Rank	Name	Market Cap	Price	Today
1	Microsoft MSFT	$3.197 T	$430.16	▲ 0.74%
2	Apple AAPL	$2.913 T	$189.98	▲ 1.66%
3	NVIDIA NVDA	$2.618 T	$1,065	▲ 2.57%
4	Alphabet (Google) GOOG	$2.170 T	$176.33	▲ 0.73%
5	Amazon AMZN	$1.880 T	$180.75	▼ 0.17%
6	Meta Platforms (Facebook) META	$1.213 T	$478.22	▲ 2.67%

Sparking the iPhone Moment of AI

The acceleration of deep learning ignited the big bang of AI. ChatGPT, a large language model powered by an NVIDIA DGX™ AI supercomputer, reached 100 million users in just two months. Its magical capabilities have captured the world's imagination. Generative AI is a new computing platform, like the PC, internet, and mobile-cloud. Accelerated computing and AI have fully arrived.

What's the definition of a large language model?

A large language model is a type of artificial intelligence system that has been trained on massive amounts of text data and can generate human-like language responses to input.
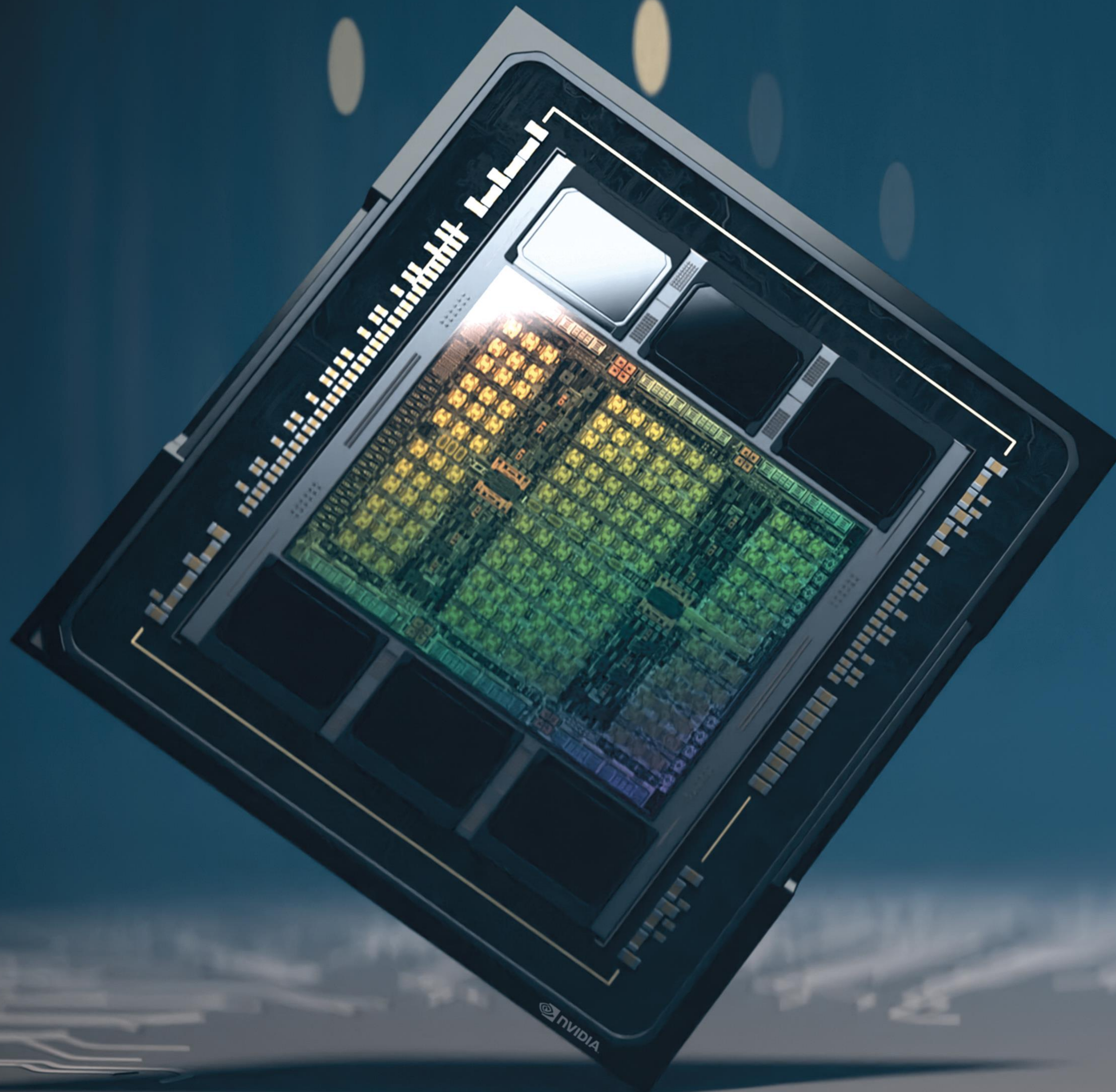
Enterprise are on the Generative AI Journey

2022

Explosion

ChatGPT gets announced late in 2022, gaining over 100 million users in just two months. Users of all levels can experience AI and feel the benefits firsthand.

2023

Experimentation

Enterprise application developers kick off POCs for generative AI applications with API services and open models including Llama 2, Mistral, NVIDIA, and others.
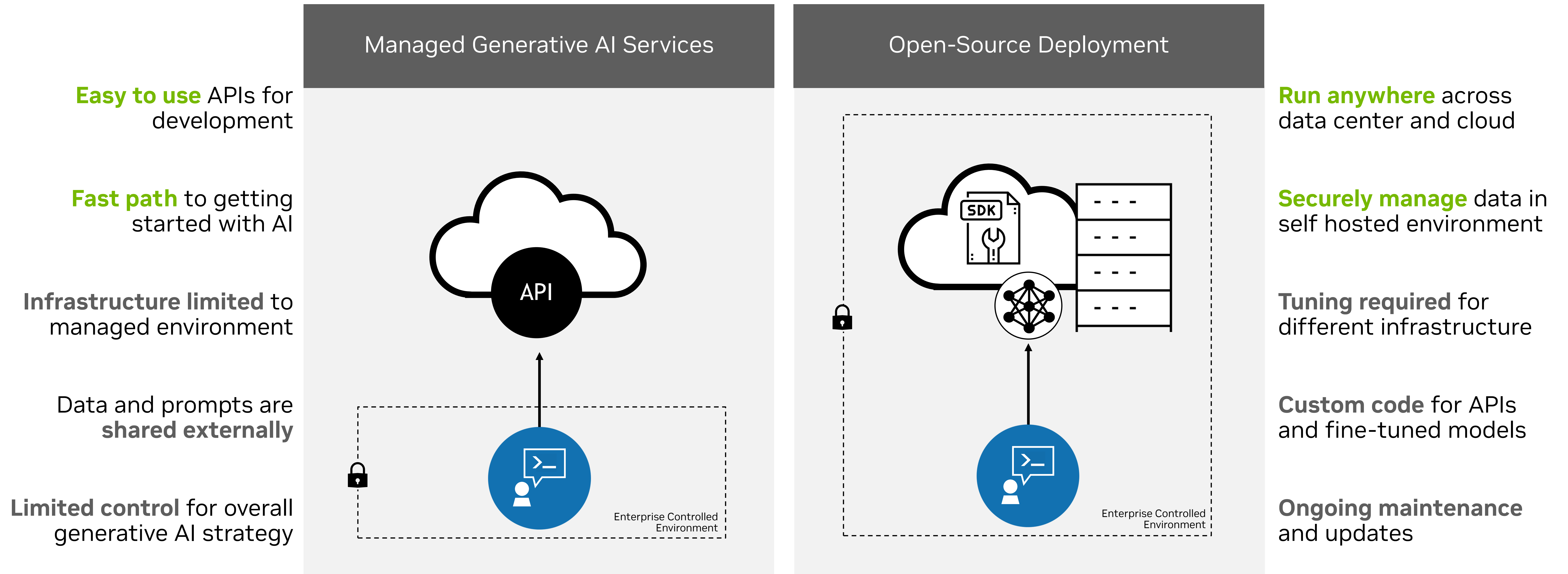
2024

Production

Organizations have set aside budget and are ramping up efforts to build accelerated infrastructure to support generative AI in production.

NVIDIA

Enterprises Face Challenges Experimenting with Generative AI

Organizations must choose between ease of use and control

Managed Generative AI Services

Open-Source Deployment

Easy to use APIs for development

Fast path to getting started with AI

Infrastructure limited to managed environment

Data and prompts are **shared externally**

Limited control for overall generative AI strategy

API

Enterprise Controlled Environment

Run anywhere across data center and cloud

Securely manage data in self hosted environment

Tuning required for different infrastructure

Custom code for APIs and fine-tuned models

Ongoing maintenance and updates

SDK

Enterprise Controlled Environment

NVIDIA AI Enterprise

High Performance and Efficient Runtime for Generative AI

Reasoning	Speech & Translation	BioMedical	Content Generation	Route Planning

Community Models	NVIDIA Models	Custom Models

NIM Microservices	CUDA-X Microservices

Security Advisory	Enterprise Support	Cluster Management	Infra Optimization	Production Branch

NVIDIA AI Enterprise
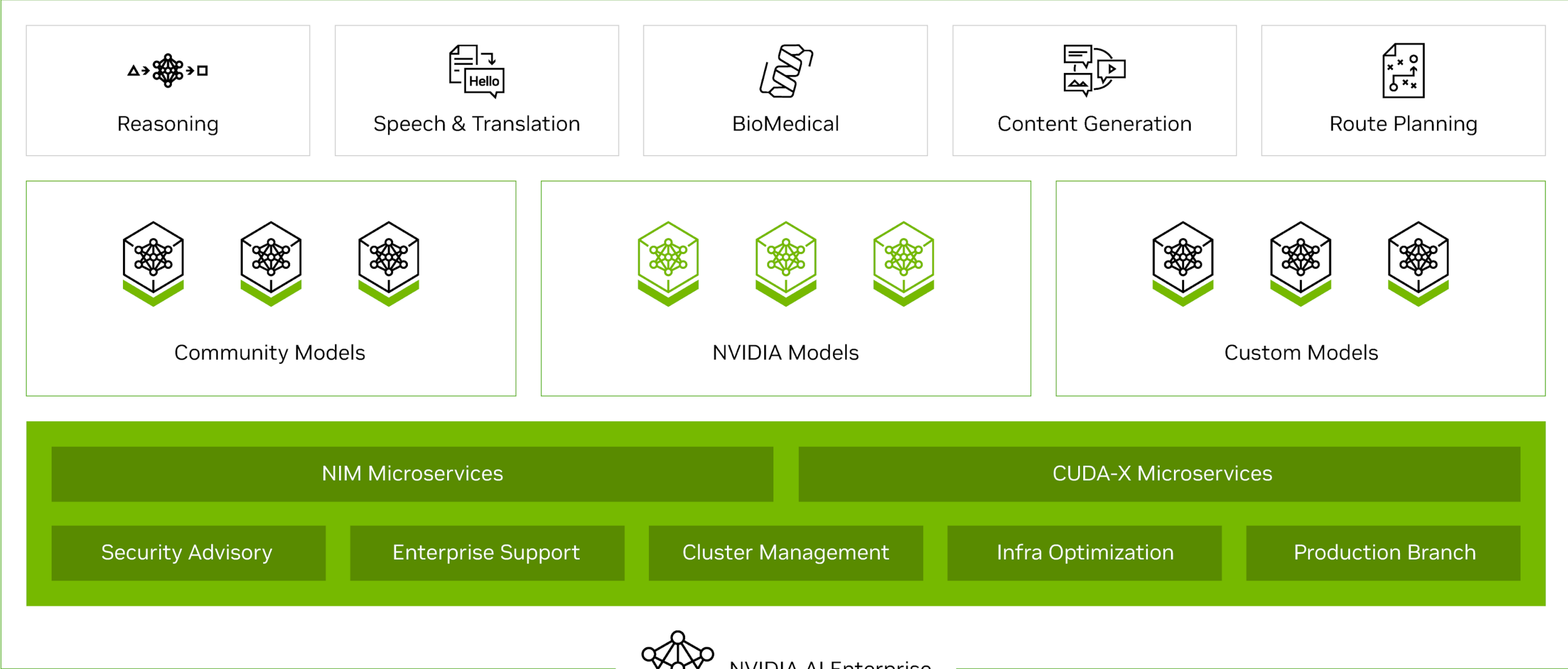
Cloud | Data Center | Workstations | Edge

Try NVIDIA AI: https://build.nvidia.com/explore/discover

NVIDIA Powers AI Factories

Data centers process mountains of continuous data to train and refine AI software. Companies are manufacturing intelligence, and their data centers are becoming giant AI factories. NVIDIA is the engine of the world's AI infrastructure.
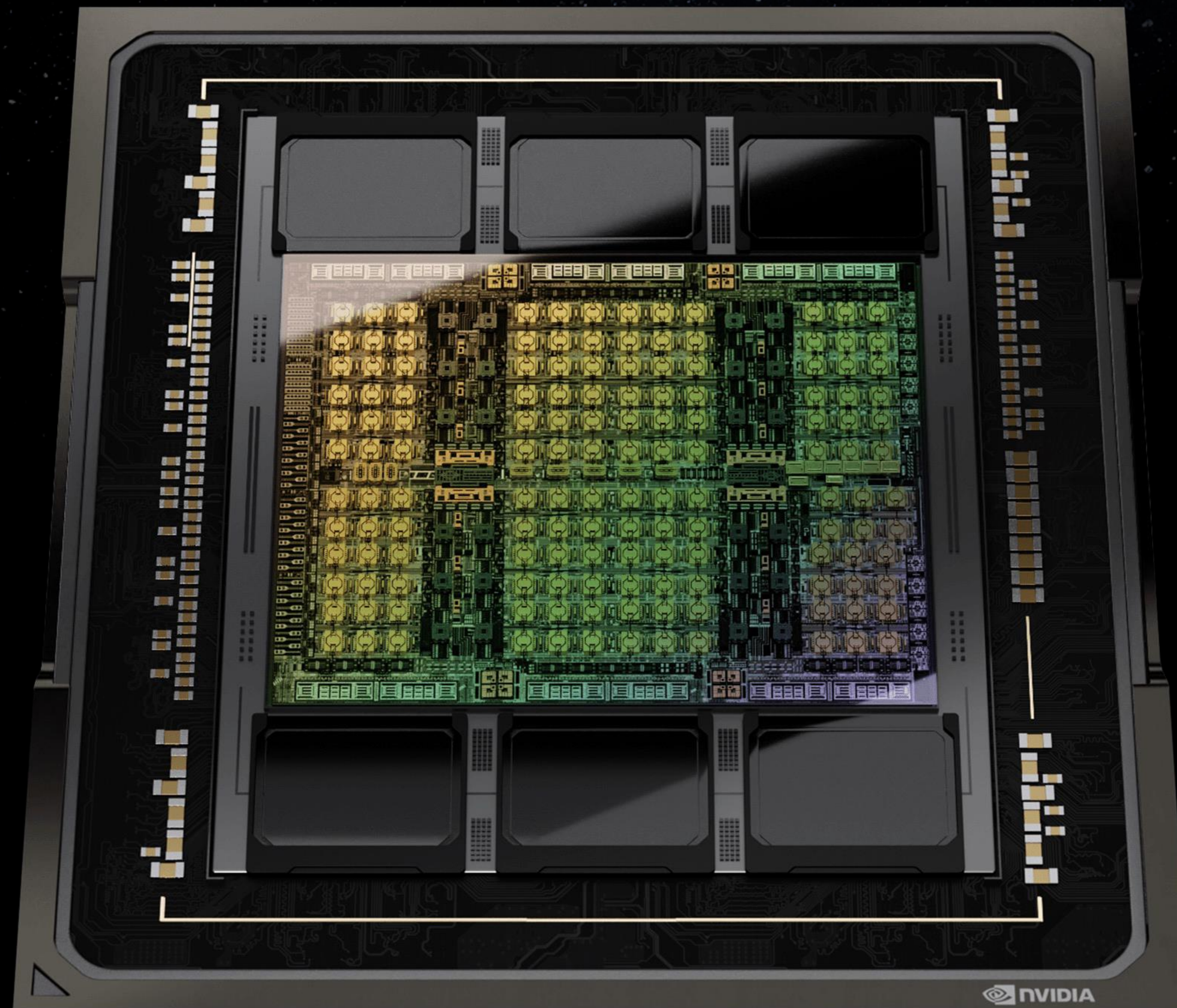
Hopper —
The Engine for the
World's AI Infrastructure

The NVIDIA Hopper™ architecture will power
the next wave of AI data centers. The first
Hopper-based GPU, the NVIDIA H100, comes
packed with 80 billion transistors and delivers
an order-of-magnitude performance leap over
its predecessor.

NVIDIA DGX—Purpose-Built for the Unique Demands of AI

Our fourth-generation NVIDIA DGX system is the world's first AI platform to be built with the new H100 GPUs. Each DGX H100 provides 32 petaflops of AI performance at FP8 precision—6X more than the prior generation. The next-generation DGX SuperPOD™ will expand the frontiers of AI with the ability to run massive workloads with trillions of parameters.
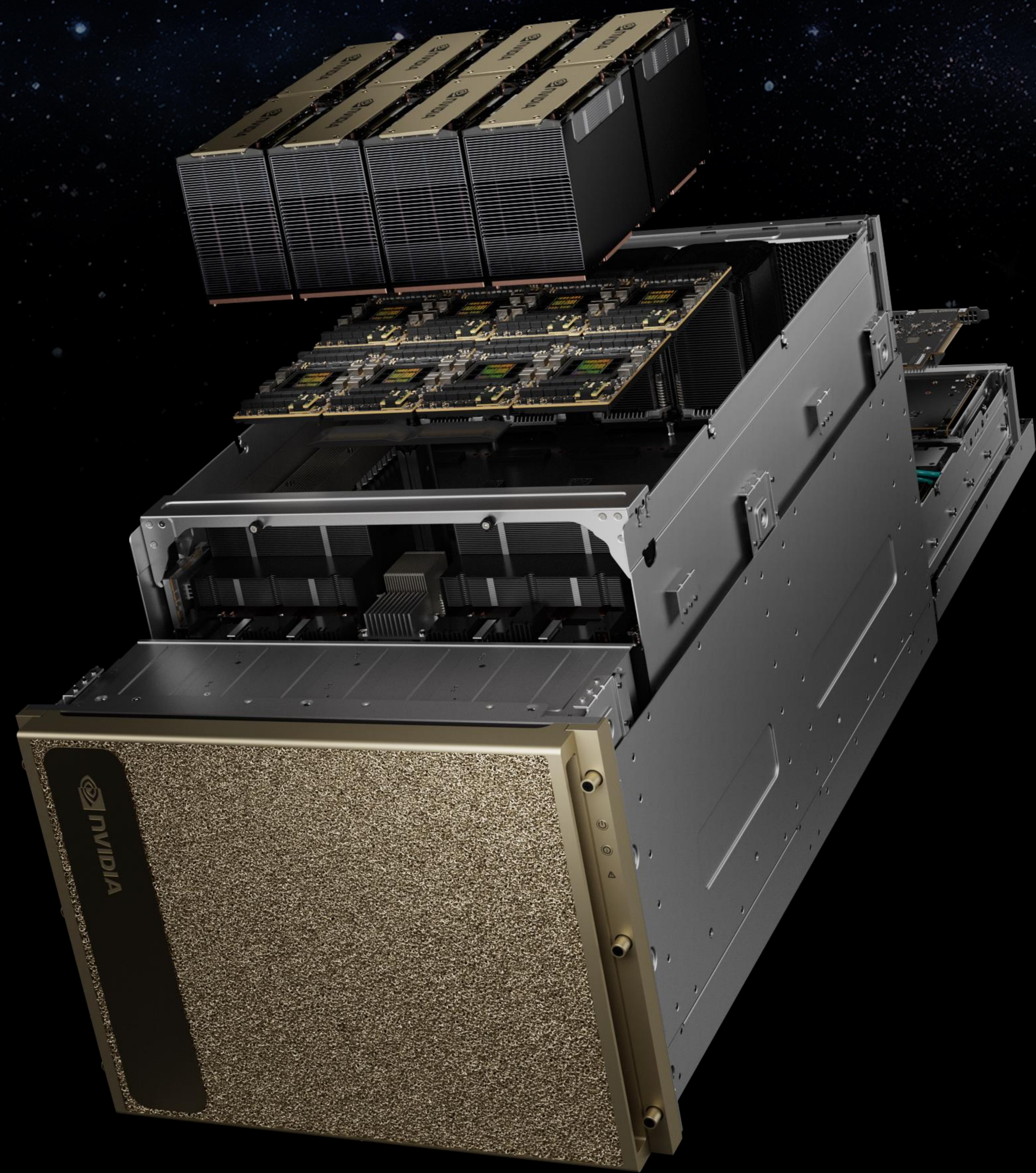
Revolutionizing Financial Services with Artificial Intelligence

"Best Places to Work in 2023"

Glassdoor

"Most Innovative Companies"

Fast Company

"World's Best CEOS"

Barron's

"100 Best Companies to Work For"

Fortune

"World's Best Performing CEO"

Harvard Business Review

"50 Smartest Companies"

MIT Tech Review

Founded in 1993   Jensen Huang Founder & CEO   27,000 Employees $27B in FY23

Marc Stampfli, May 2024

E-Mail: mstampfli@nvidia.com
Linkedin: https://www.linkedin.com/in/marcstampfli