AI Factories – Transforming Data Centers with NVIDIA and HPE

Eric Kang 康勝閔, 資深解決方案架構協理, NVIDIA

October 16, 2025

Trillion-Dollar Global IT Investment Shifting to AI Factories

92% of enterprises investing in AI

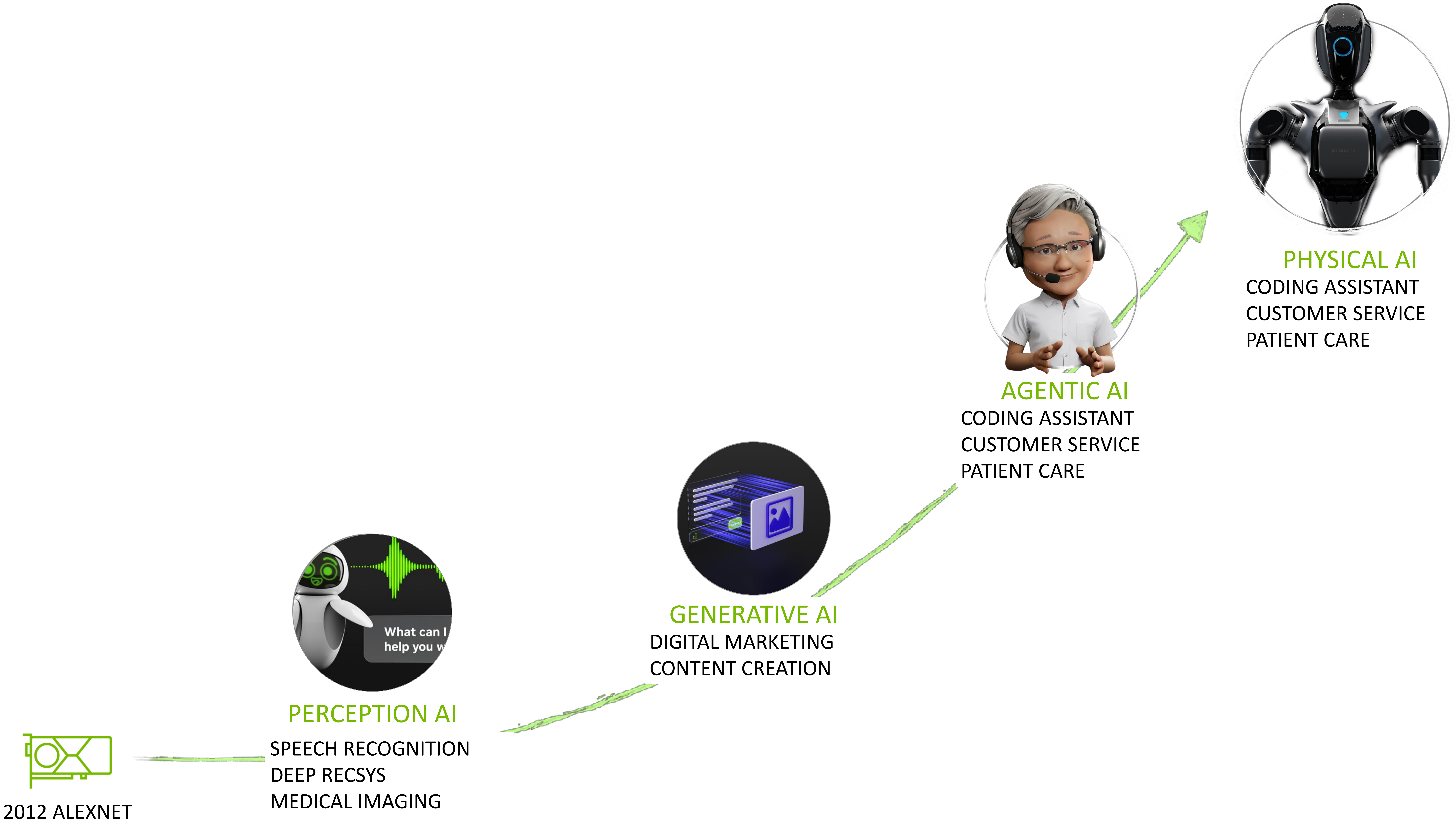
50% will use AI agents to achieve business value

33% find complexity top barrier for adoption

1% have mature AI deployments

Evolution of AI

Agentic AI Enables More Powerful AI Applications



2012 ALEXNET

PERCEPTION AI
SPEECH RECOGNITION
DEEP RECSYS
MEDICAL IMAGING

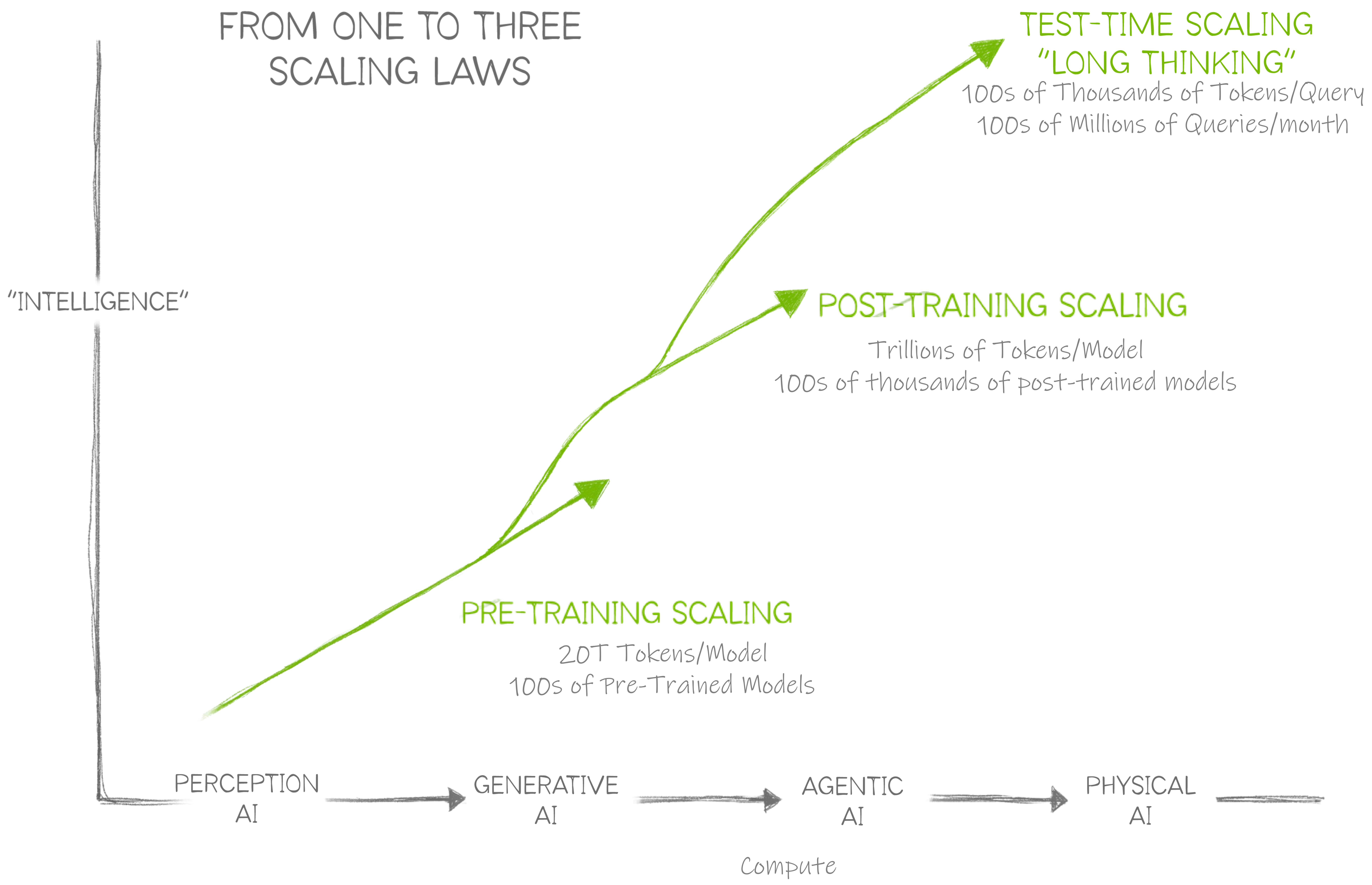
GENERATIVE AI
DIGITAL MARKETING
CONTENT CREATION

AGENTIC AI
CODING ASSISTANT
CUSTOMER SERVICE
PATIENT CARE

PHYSICAL AI
CODING ASSISTANT
CUSTOMER SERVICE
PATIENT CARE

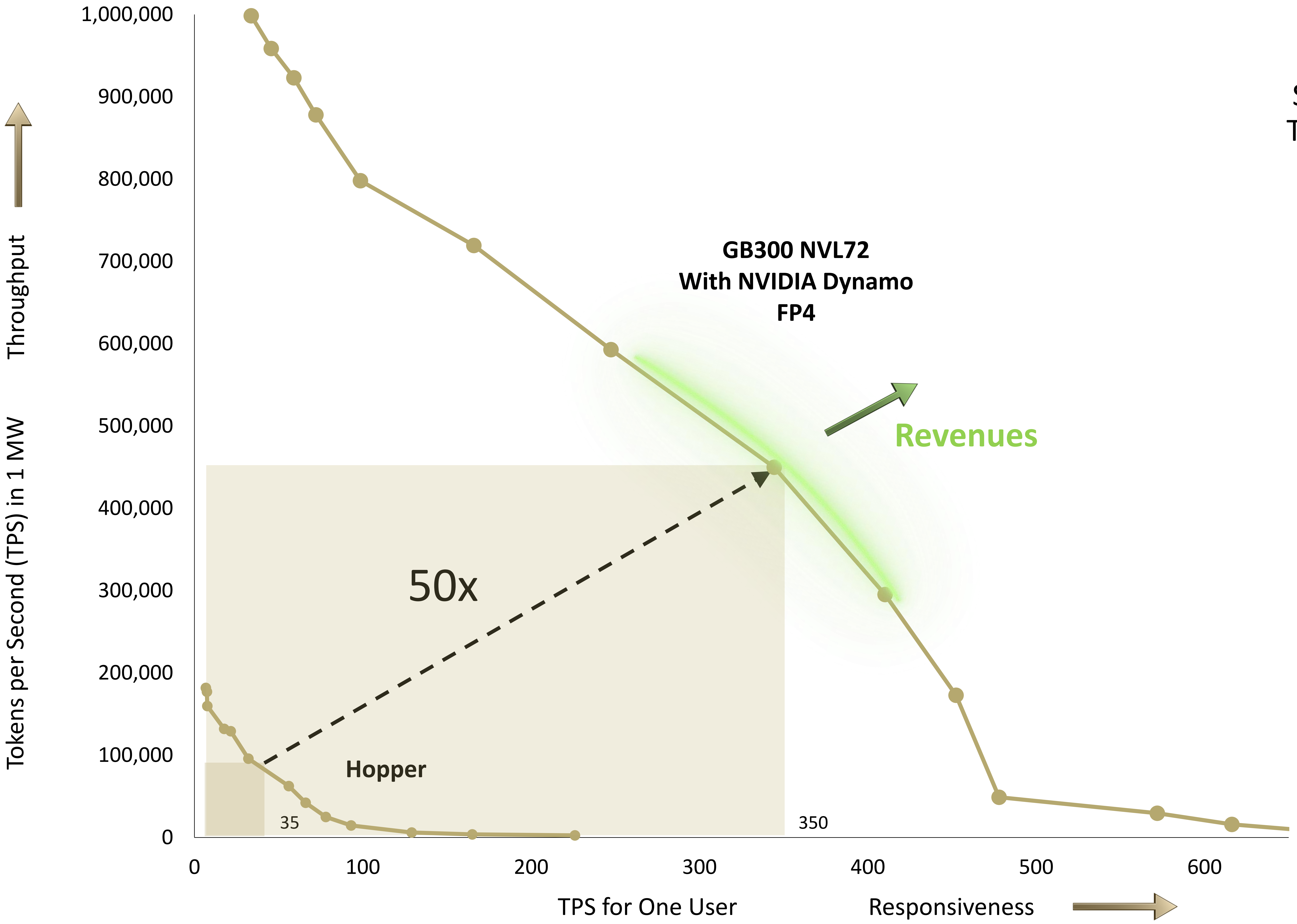
AI Scaling Laws Drive Exponential Demand for Compute

New “long thinking” required for agentic and physical AI


FROM ONE TO THREE SCALING LAWS
“INTELLIGENCE”
TEST-TIME SCALING “LONG THINKING”
100s of Thousands of Tokens/Query
100s of Millions of Queries/month
POST-TRAINING SCALING
Trillions of Tokens/Model
100s of thousands of post-trained models
PRE-TRAINING SCALING
20T Tokens/Model
100s of Pre-Trained Models
PERCEPTION AI
GENERATIVE AI
AGENTIC AI
PHYSICAL AI
Compute

AI Factory Output Drives Revenue

High throughput multiplied by high interactivity = total token output



Supercharging AI Reasoning Throughput, Responsiveness and Revenue

50X

AI Factory Output

10X

Responsiveness

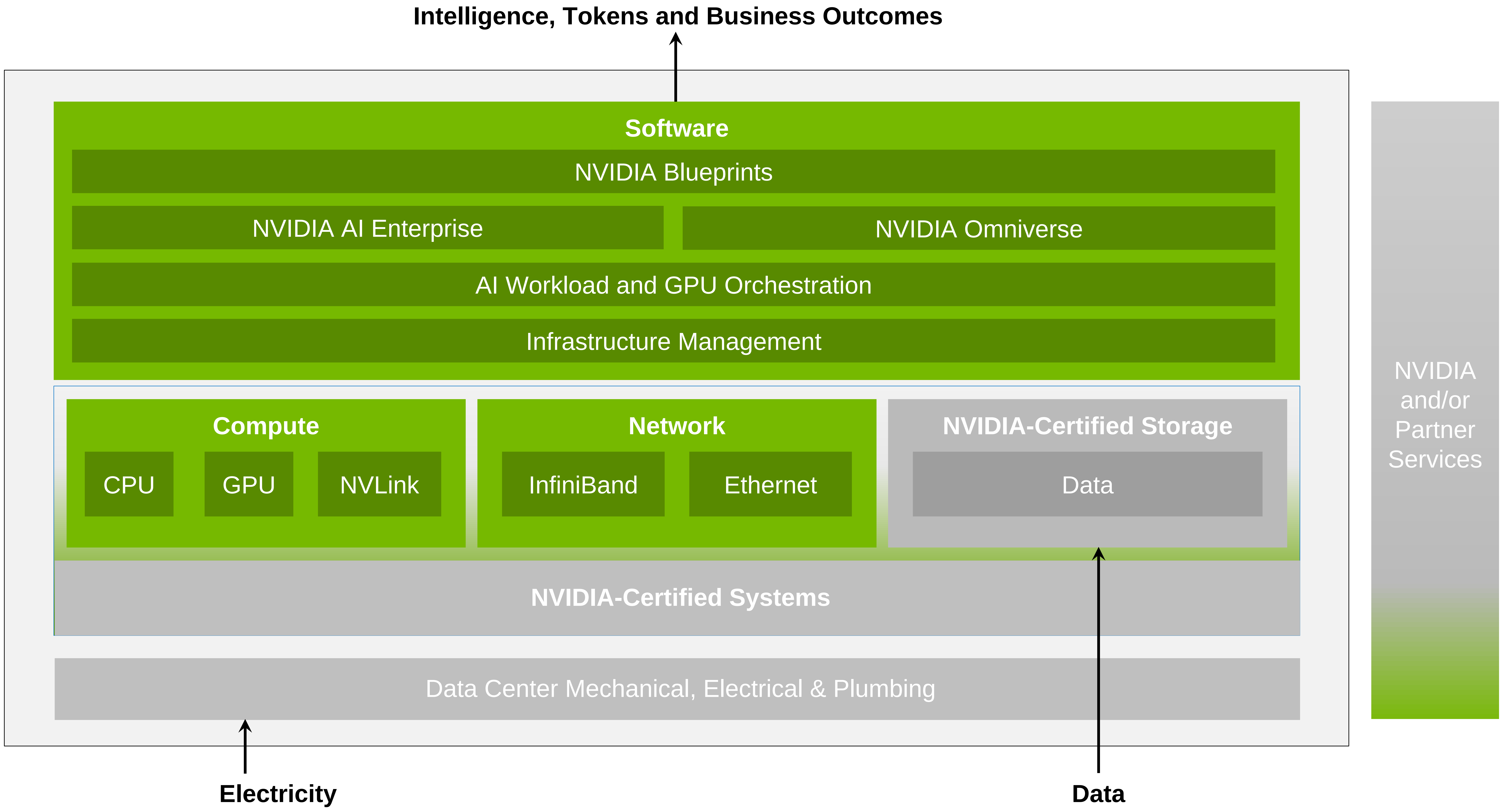
5X

ISO Power Throughput

DeepSeek R1 ISL = 32K, OSL = 8K, GB300 NVL72 with FP4 Dynamo disaggregation. H100 with FP8 In-flight batching. Projected performance subject to change.

NVIDIA Provides a Full Stack for AI Factories

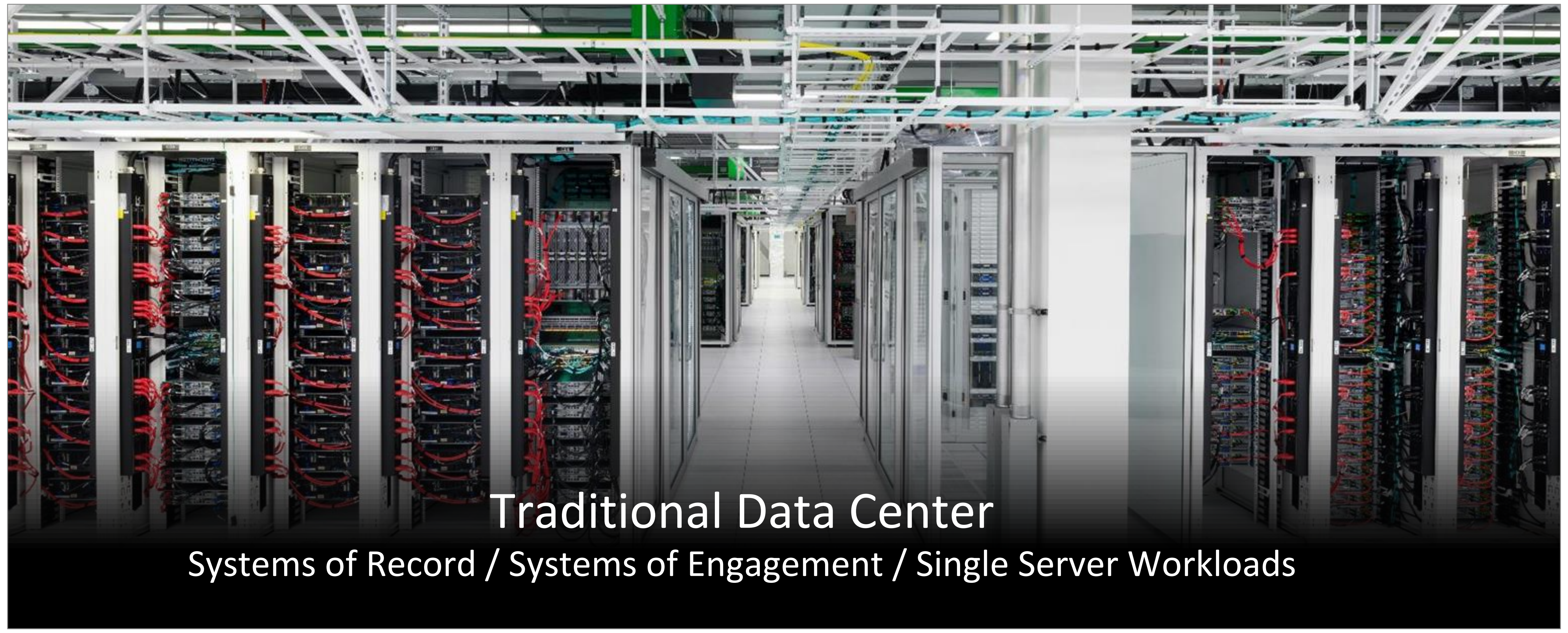
Built on NVIDIA Validated Data Center Reference Architectures



Intelligence, Tokens and Business Outcomes
Software
NVIDIA Blueprints
NVIDIA AI Enterprise
NVIDIA Omniverse
AI Workload and GPU Orchestration
Infrastructure Management
Compute
CPU
GPU
NVLink
Network
InfiniBand
Ethernet
NVIDIA-Certified Storage
Data
NVIDIA-Certified Systems
Data Center Mechanical, Electrical & Plumbing
NVIDIA and/or Partner Services
Electricity
Data

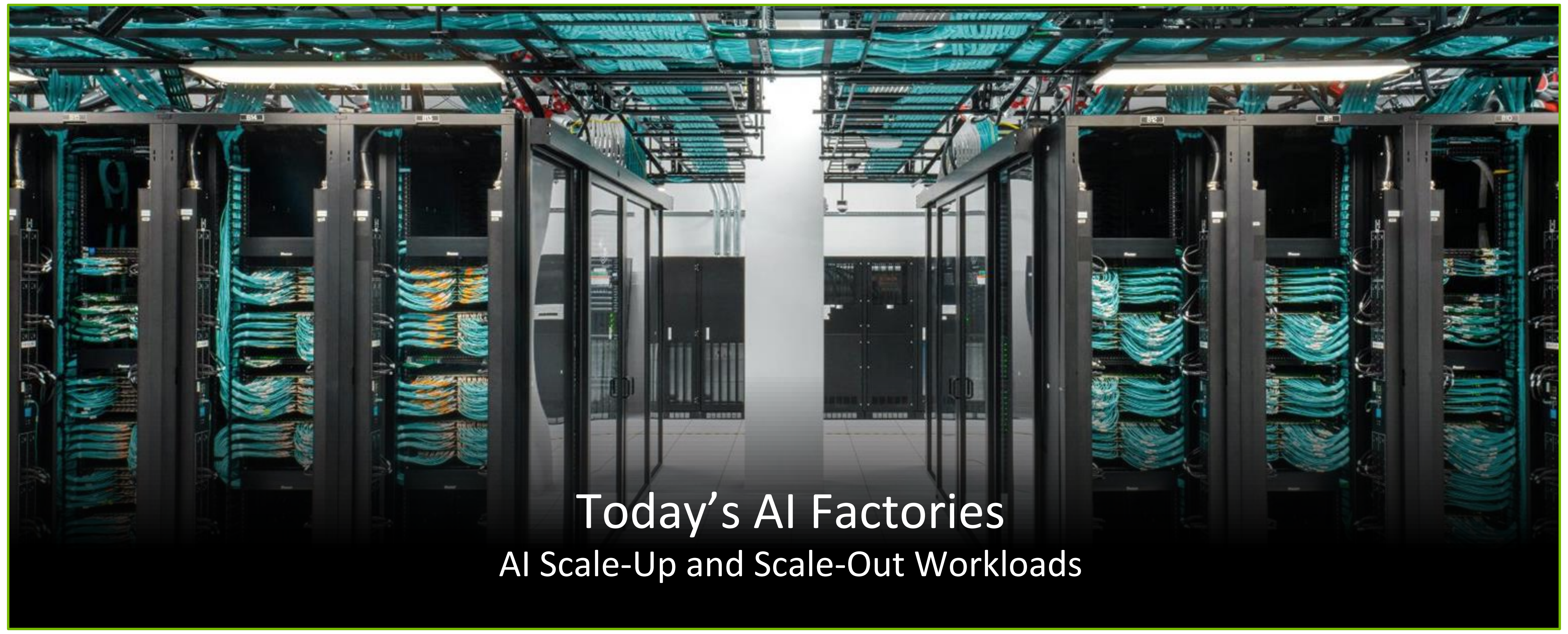
Transforming Data Centers into AI Powerhouses

Enterprises Need for Purpose-Built AI Factories for the Age of AI Reasoning



Traditional Data Center

Systems of Record / Systems of Engagement / Single Server Workloads



Today’s AI Factories

AI Scale-Up and Scale-Out Workloads

Cost Center	Profit Engine
Retrieval	Generative
X86 CPU	Accelerated
20 kW/Rack	>130 kW/Rack
Air Cooled	Liquid and Air Cooled
100k NICs	400k Network Fabric

Today’s AI Challenges Require AI Factories

AI Workloads Require Optimized Full Stack Solutions

Design Complexity

Spans project prioritization, data acquisition, infrastructure, and sizing

Deployment and Cost

Infrastructure, security, and customization

Time to Value

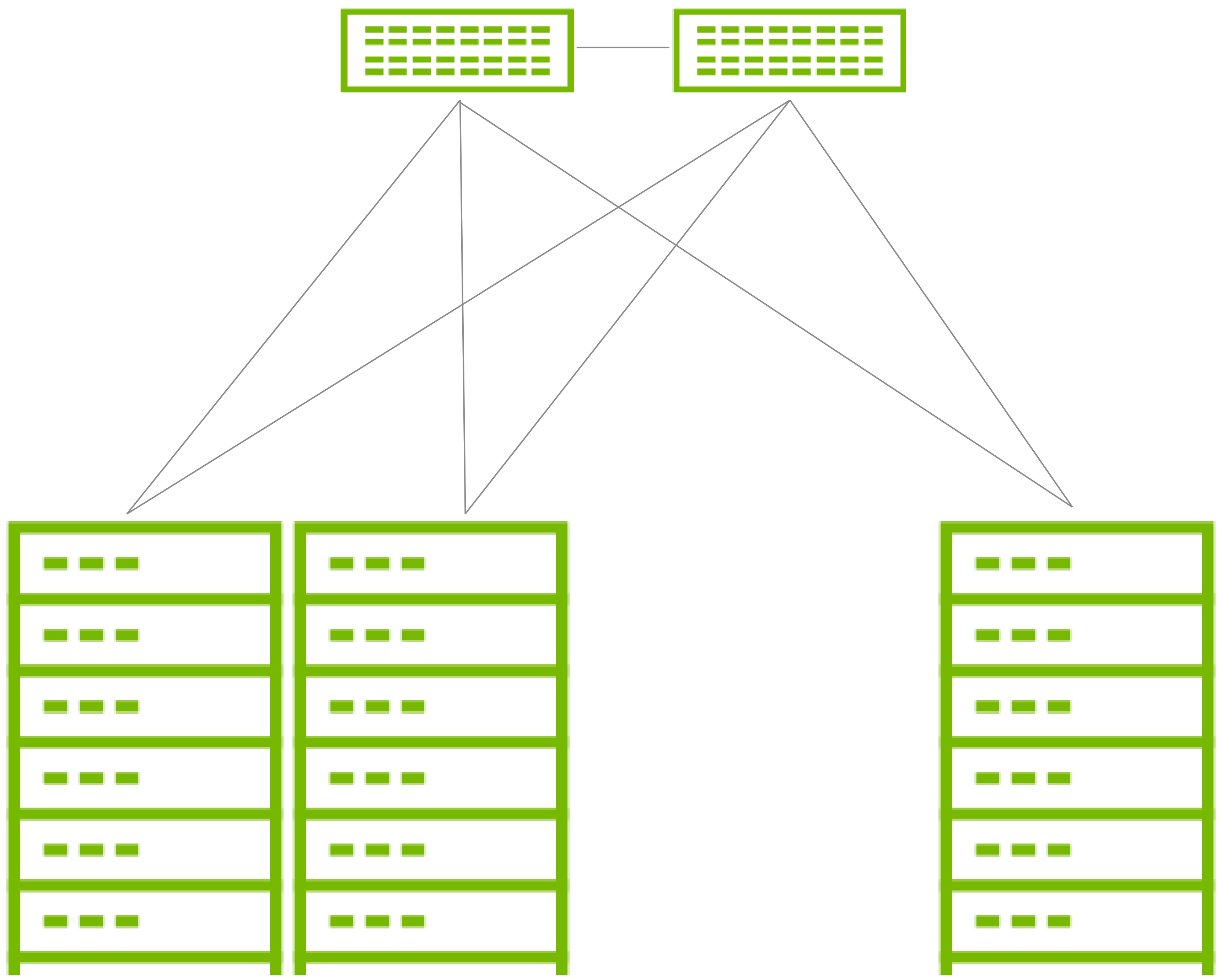
Resource management, time-to-first-train, time-to-inference

NVIDIA Enterprise AI Factory Validated Design

Building on NVIDIA Enterprise Reference Architectures

Networking

NVIDIA Spectrum-X Ethernet Networking



Compute

Storage

NVIDIA-Certified Systems

RTX PRO Servers

or HGX Blackwell Systems

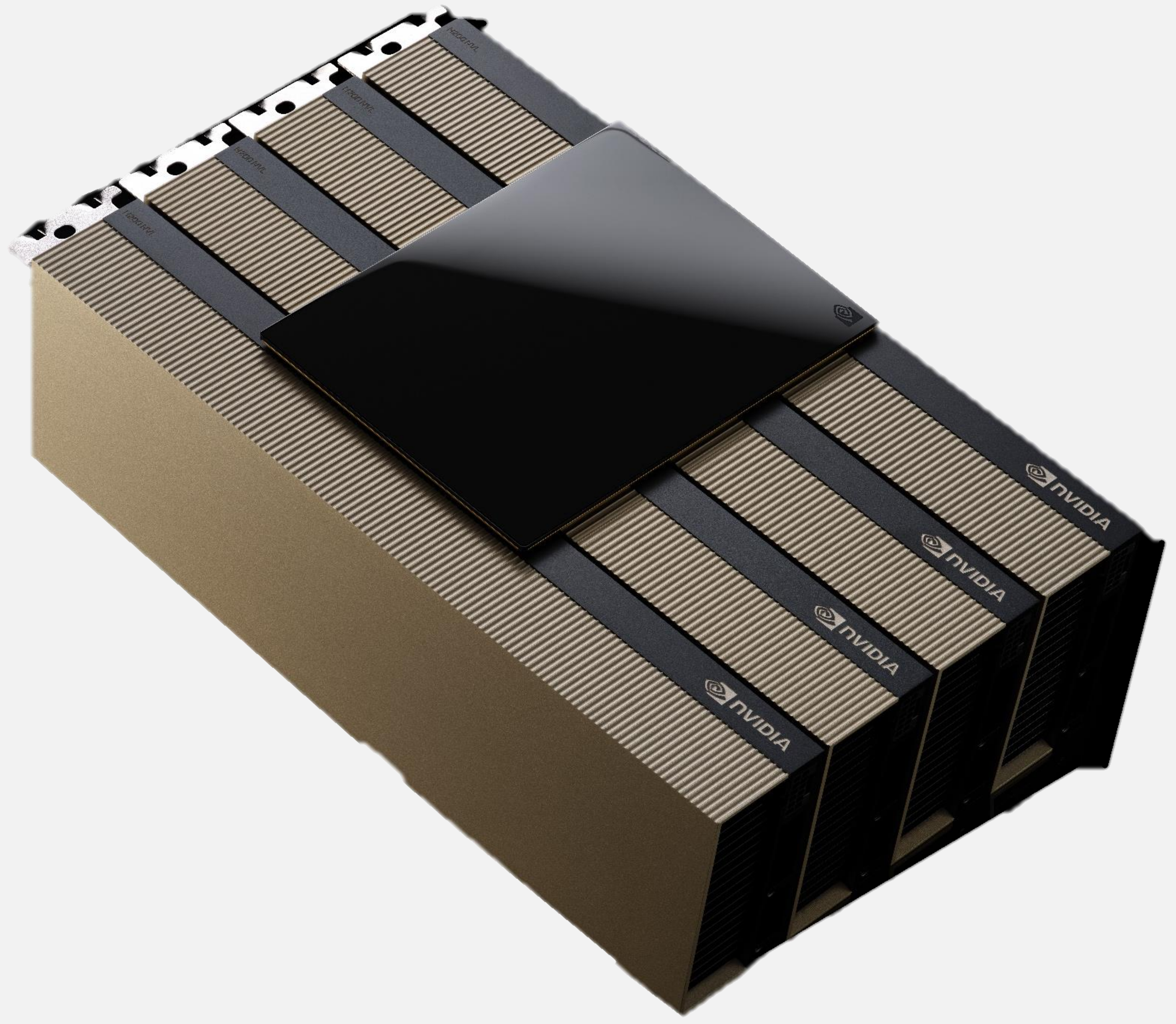
NVIDIA-Certified Storage

- ✓ **Time to value**
- ✓ **Scalability**
- ✓ **Manageability**
- ✓ **Security**

NVIDIA Accelerated Computing for Enterprise

From Hopper to Blackwell, Enterprise AI Factory Building Blocks

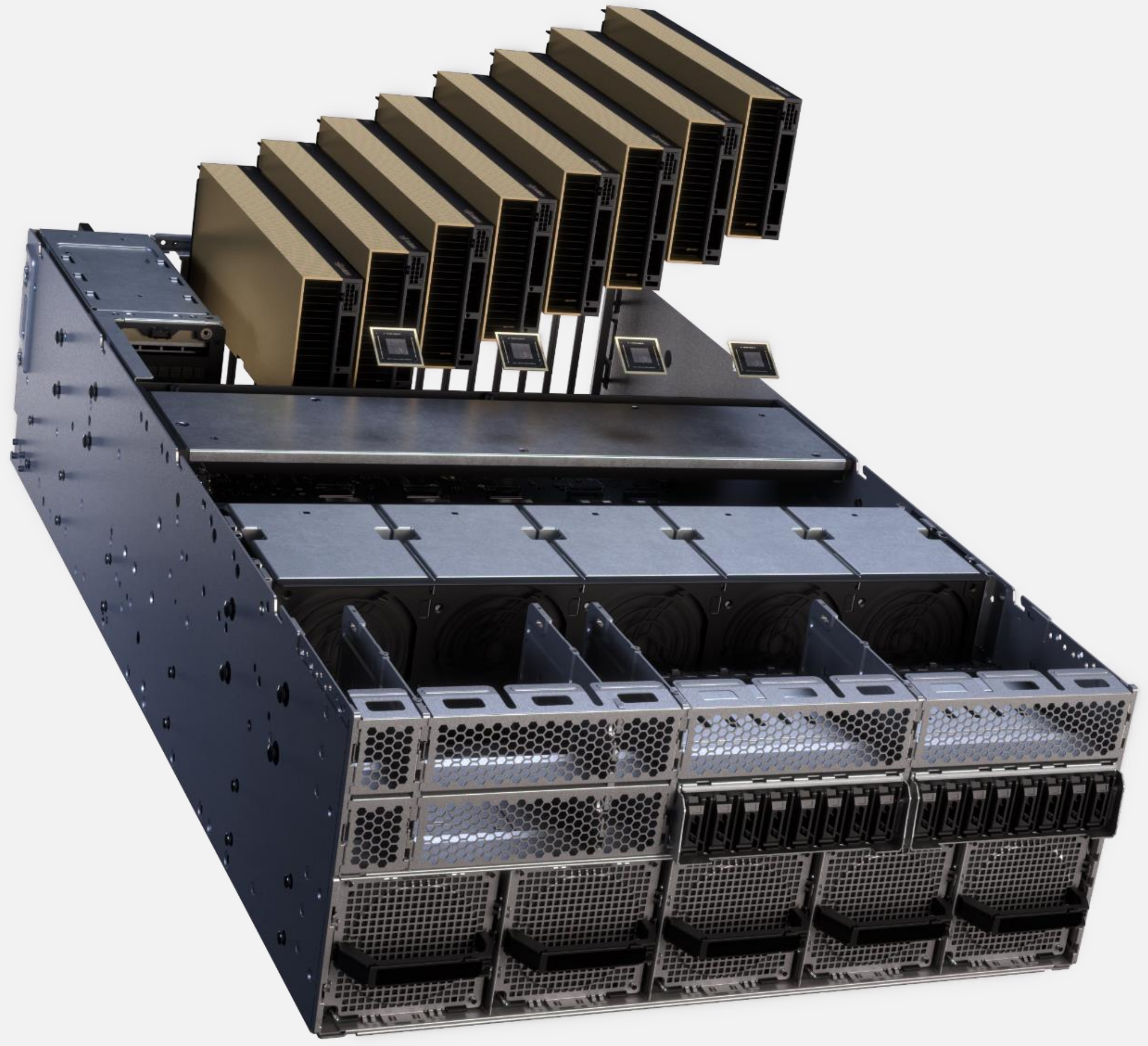
LLM Inference and Enterprise AI
Available for Deployments Now



H200 NVL

Hopper, incl 5 yr NVIDIA AI Enterprise license
8kW, Air-Cooled PCIe, x86, with NVL4 and FP64
Enterprise RA (2-8-5)

Flexibility and Performance
for Enterprise AI and Industrial AI



RTX PRO Server

Blackwell, Enterprise DC Compatibility, Multi-Workload
7kW, Air-Cooled PCIe, x86, RTX Graphics, vGPU
Enterprise RA (2-8-5)

Best Air-Cooled Platform
for LLM Training and Inference

HGX B200

Blackwell, Optimized Performance and TCO
14kW, Air-Cooled HGX, x86
Enterprise RA (2-8-9)

* 2-8-5 config (PCIe-Optimized) = 2 CPUs, 8 GPUs, 5 network adapters; 2-8-9 (HGX) = 2 CPUs, 8 GPUs, 9 network adapters

RTX PRO 6000 Blackwell Server Edition

The Most Powerful Blackwell Data Center Platform for AI and Visual Computing

Breakthrough Multimodal AI Inference

- 5th-Gen Tensor, 2nd-Gen Transformer Engine, FP4
- Full Media Pipeline: 4 NVENC/ NVDEC/ NVJPEG

Powerful Graphics and Visual Computing

- 4th-Gen RTX, Neural Shaders, DLSS 4

Data Center Ready

- 96GB GDDR7,1.6 TB/s Memory BW, 128MB L2 Cache
- Multi-Instance GPU (MIG), TEE Confidential Compute

Performance Specs

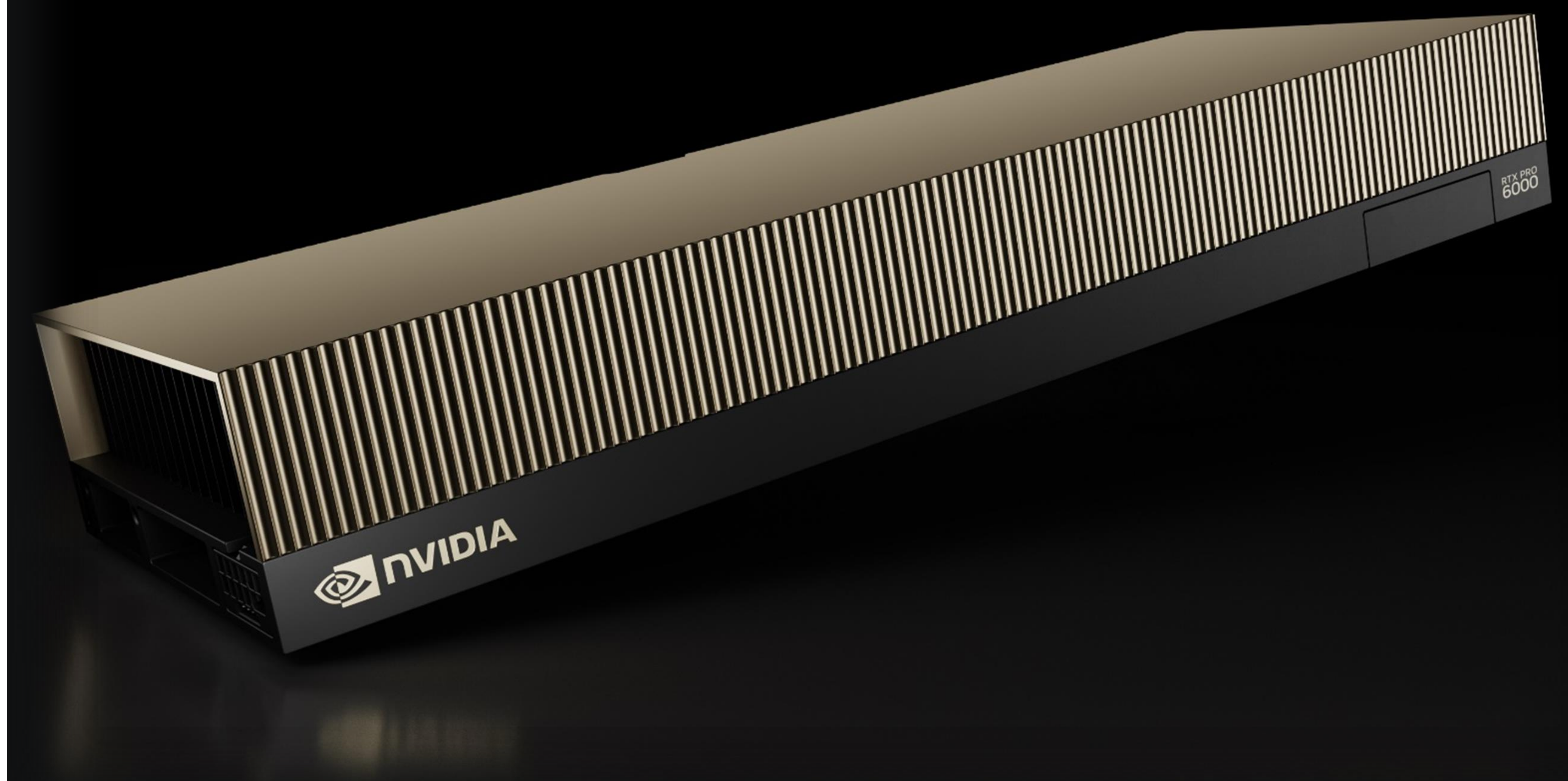
✓ 188 Ray Tracing Cores

✓ 752 Tensor Cores

✓ 24,064 Cuda Cores

✓ Peak FP4 AI Performance: 3.7 PFLOPS

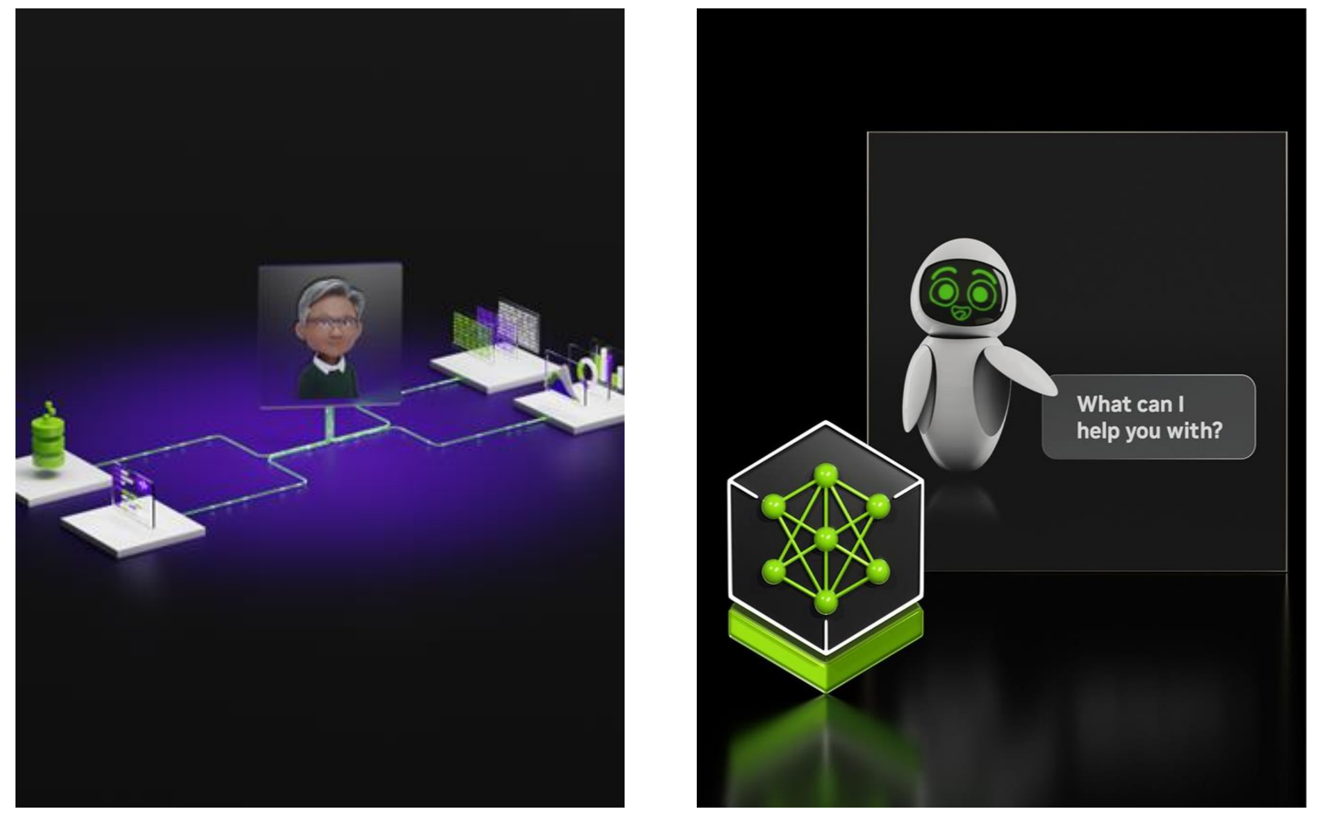
✓ Peak RT Core Performance: 354.5 TFLOPS



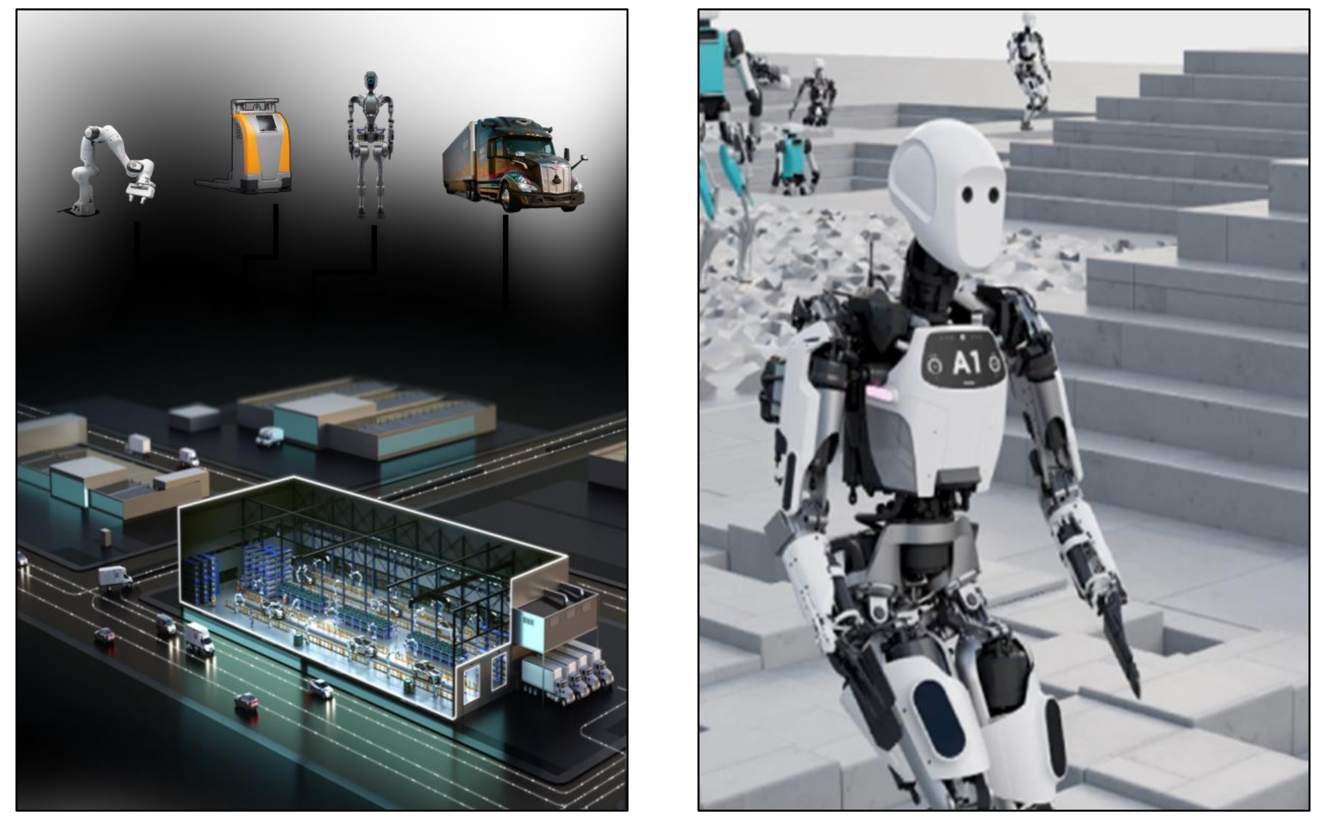
Dual Slot, FHFL I Up to 600W

Modern Enterprises Have Diverse Accelerated Workloads

From Agentic AI and Physical AI to AI-Enabled Applications

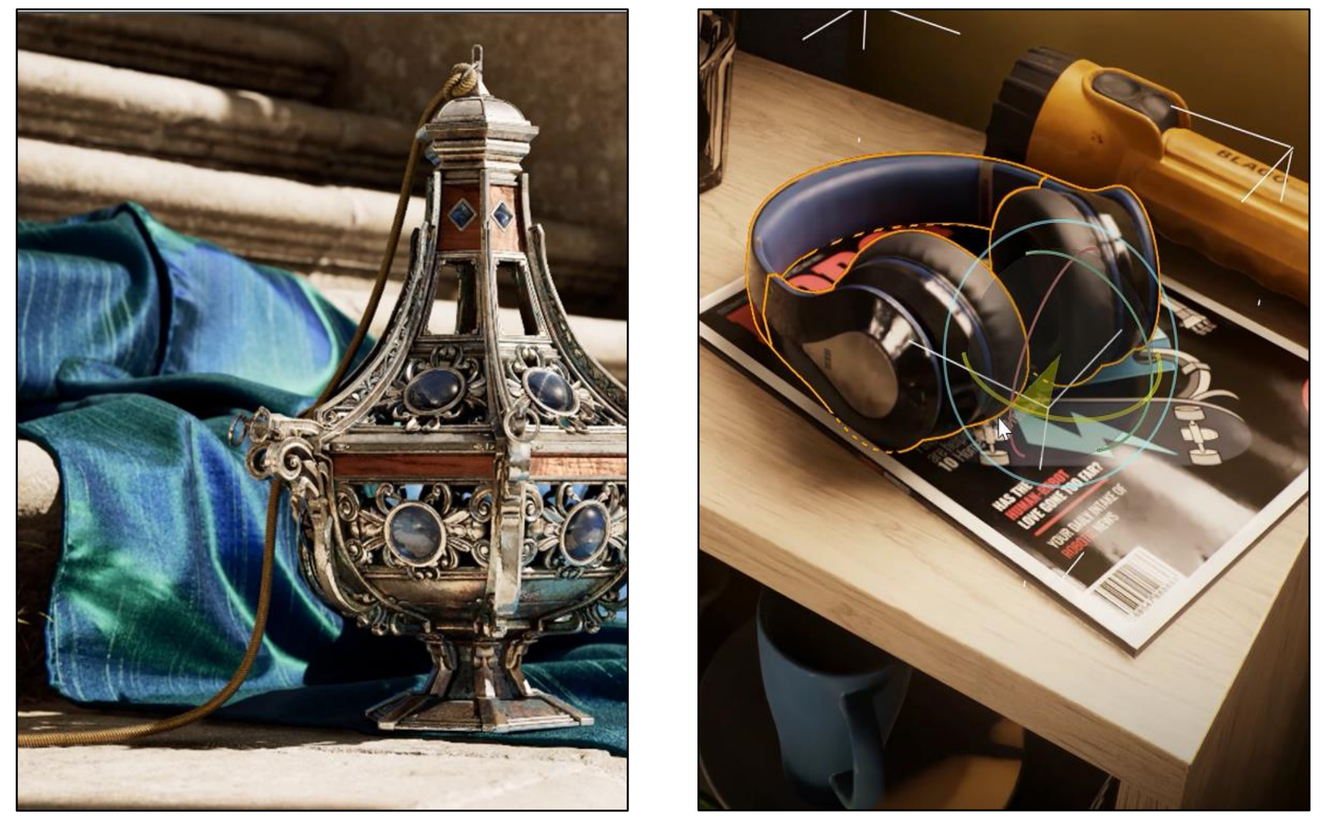


AGENTIC AI



INDUSTRIAL & PHYSICAL AI

SCIENTIFIC COMPUTING, DATA ANALYTICS, & SIMULATION



VISUAL COMPUTING

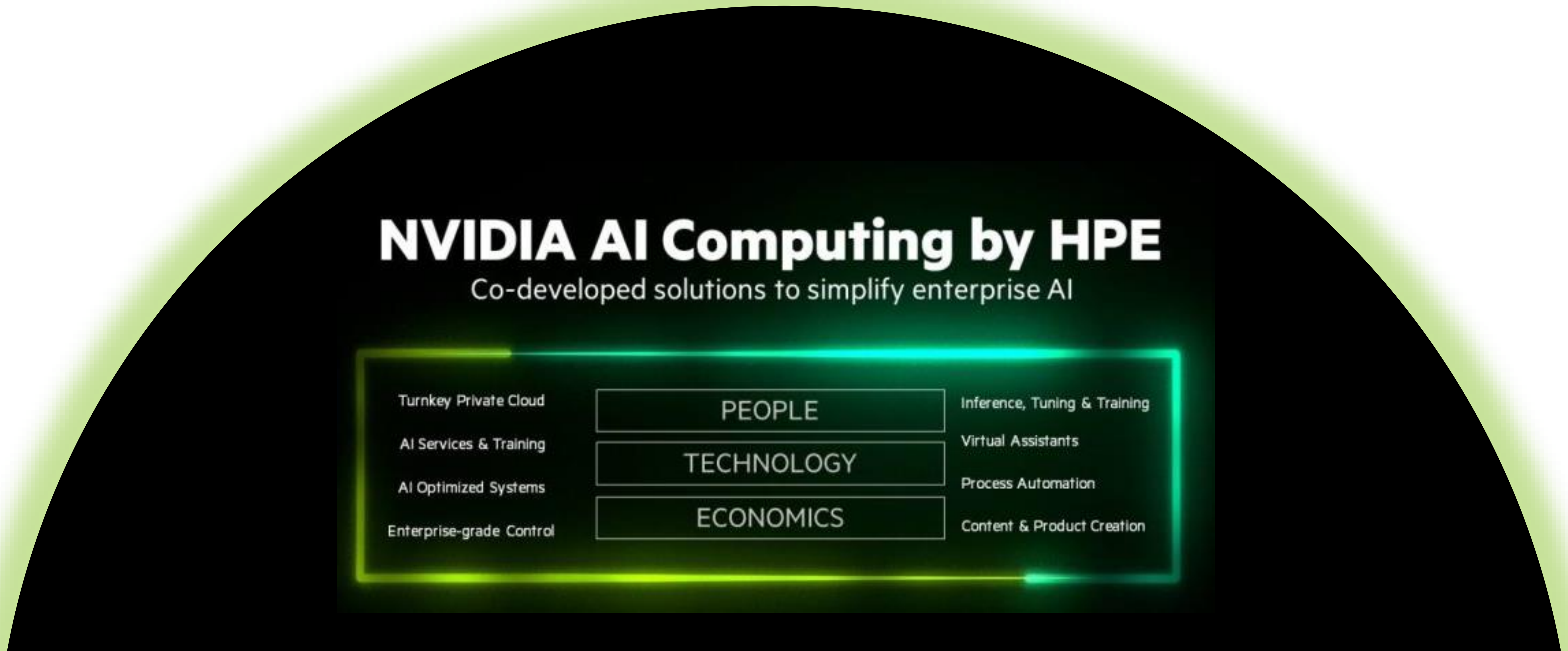
ENTERPRISE APPLICATIONS

NVIDIA AI Enterprise

NVIDIA Omniverse

NVIDIA CUDA-X Microservices

HPE ProLiant 380a Gen12



NVIDIA AI Computing by HPE

Co-developed solutions to simplify enterprise AI

- Turnkey Private Cloud
- AI Services & Training
- AI Optimized Systems
- Enterprise-grade Control

PEOPLE

TECHNOLOGY

ECONOMICS

- Inference, Tuning & Training
- Virtual Assistants
- Process Automation
- Content & Product Creation

NVIDIA RTX PRO 6000 Blackwell Server Edition

Best Server for Enterprise AI

RTX PRO Server Up to 5X Better Price Performance

AI Inference Performance/$
Normalized to L40S
6.0X
5.0X
4.0X
3.0X
2.0X
1.0X
0.0X
2.0X
3.4X
1.4X
5.0X
2.5X
Flux 12B
Cosmos
Llama 3 8B
Llama 3 70B
Mixtral 8x7B
Image Generation
Video Generation
LLM Inference
L40S
HGX H100
RTX PRO 6000

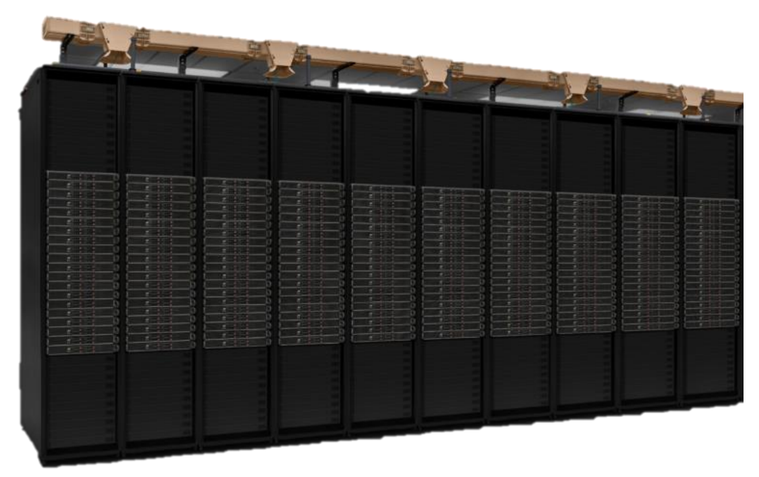
Performance/$ = Performance / TCO for a Single Node (Server + Power Costs) for L40S and HGX H100 compared to RTX PRO 6000 Blackwell Server Edition

Building Blocks For Enterprise AI Factories

NVIDIA Certified

Servers

Storage



High performance, scalable, and secure accelerated computing platform for AI & HPC

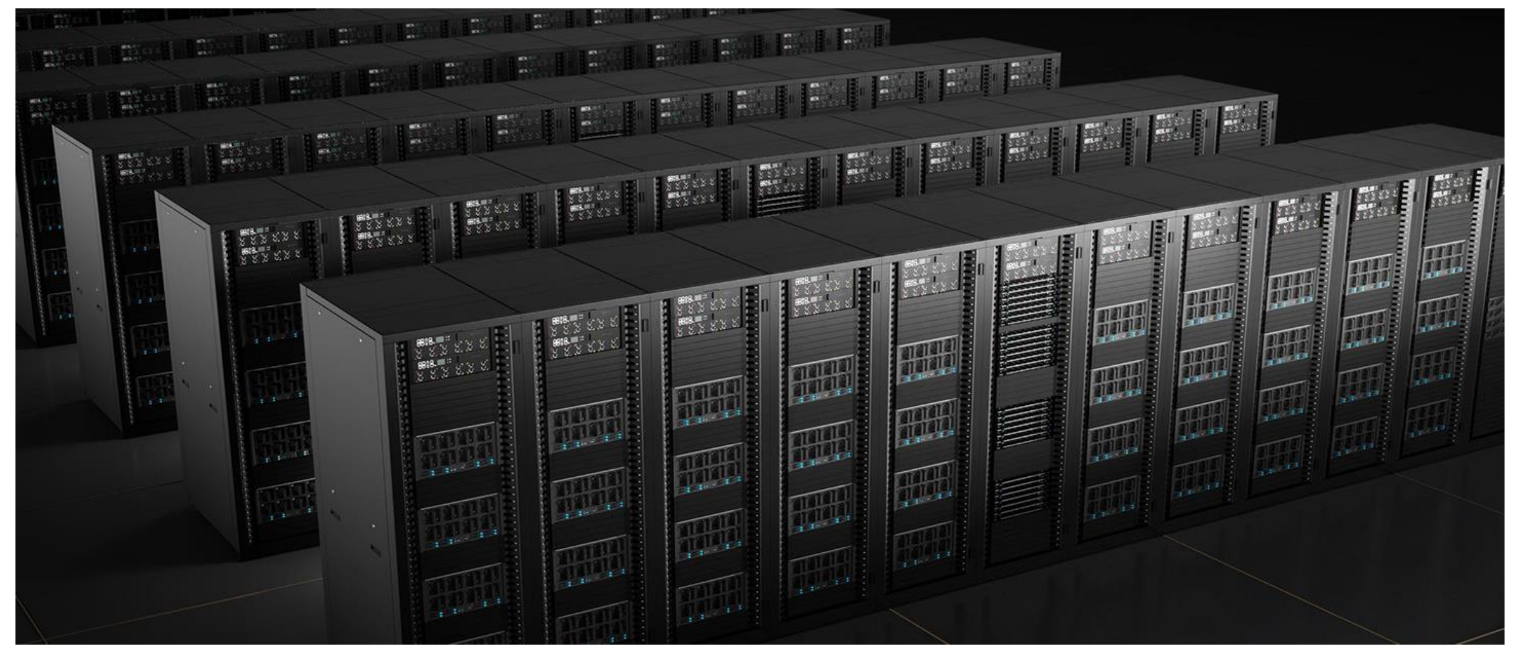
PERFORMANCE

MANAGEABILITY

SECURITY

SCALABILITY

NVIDIA Enterprise Reference Architectures



Proven and Comprehensive Recommendations for Scaling Enterprise Deployment

NVIDIA-Certified Servers | Network Topology Management | Software Stack

- Faster Time to Solution
- Performant & Scalable Infrastructure
- Reduced Complexity and Cost
- Supportable design patterns and scalable units

Adopted by NVIDIA and HPE

HPE | NVIDIA

Informing designs to speed TTM with an enhanced deployment for AI Factories of various sizes

Superior Customer Experience

- Faster Time to Deploy
- Optimized Performance
- Reduced Complexity and Cost
- Faster Time to Resolution

State of the Art Infrastructure Management

Empower model builders and enterprise IT with full stack intelligence

AI Practitioners & MLOps

IT & DevOps

Cloud Native AI Software

NVIDIA AI Enterprise

NVIDIA Mission Control

AI Workload Management — Seamless workload orchestration with **NVIDIA Run:ai** technology

Autonomous Recovery Engine — Problem identification, isolation, and recovery

Advanced Cluster Provisioning — Next-generation cluster management for AI's next frontiers

Telemetry and Observability — Customizable dashboards

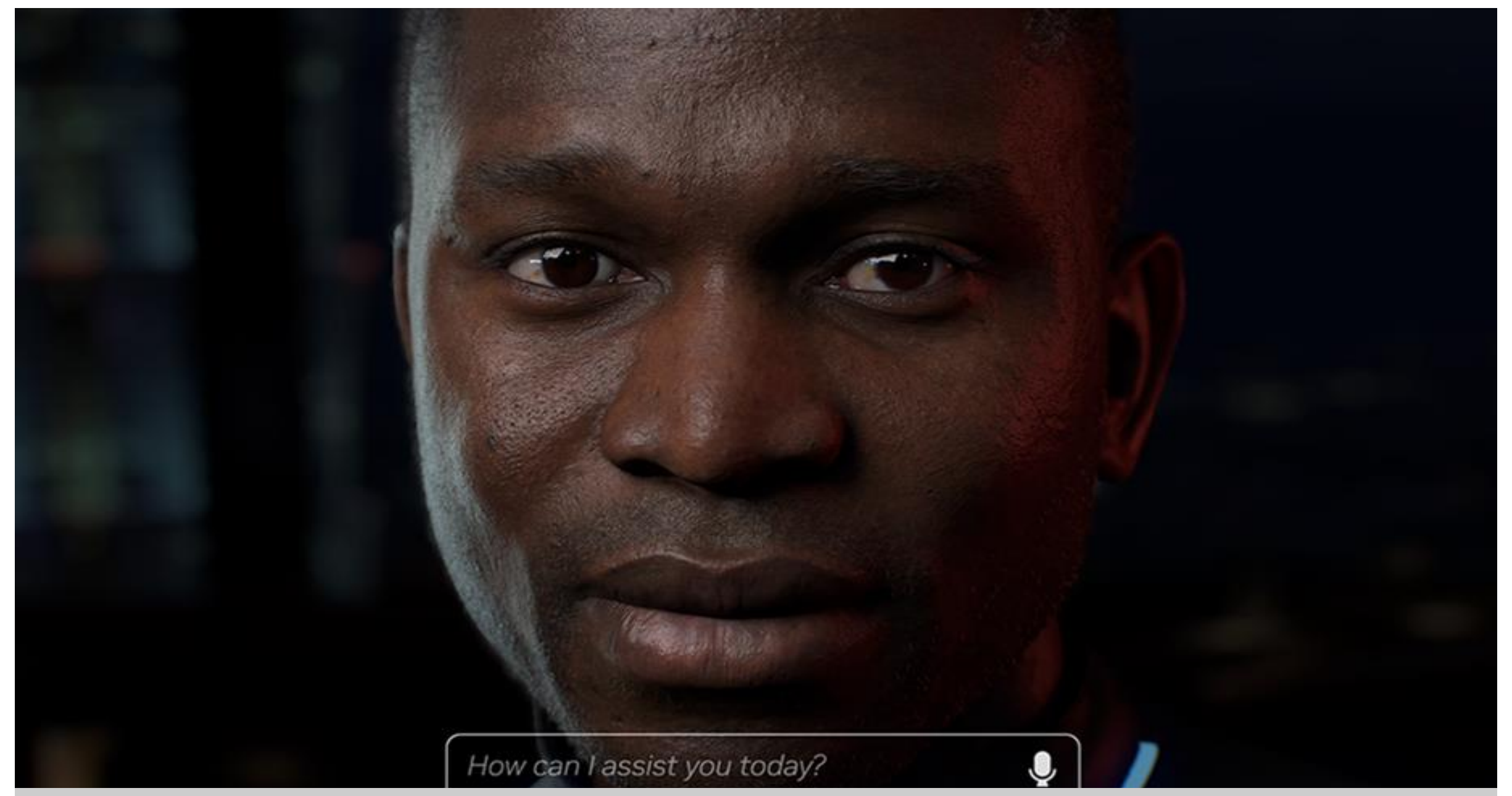
Validation and Diagnostics — On-demand health checks

AI Factory Infrastructure

NVIDIA Blueprints

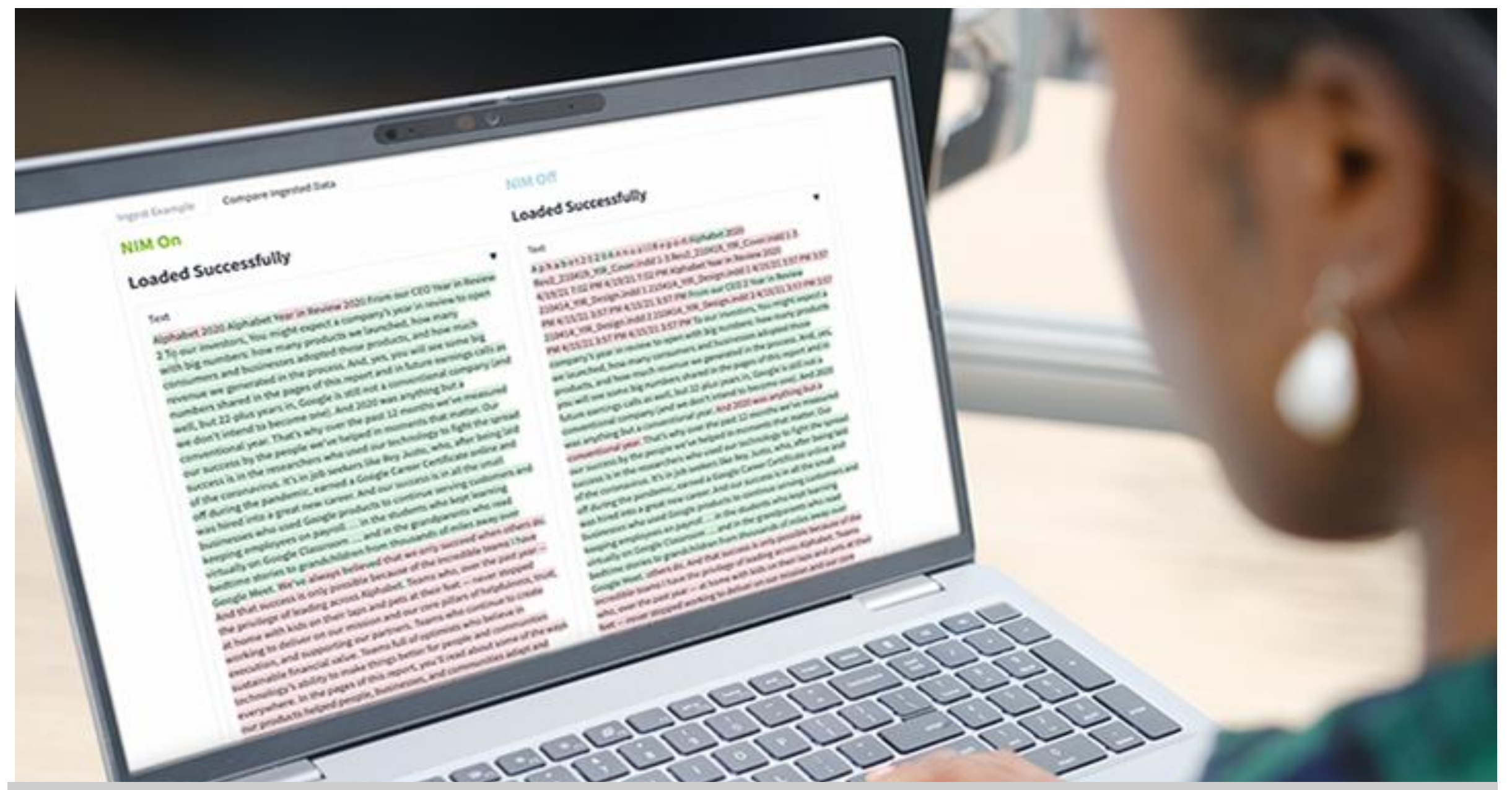
Available on https://build.nvidia.com/

Digital Humans
for Customer Service



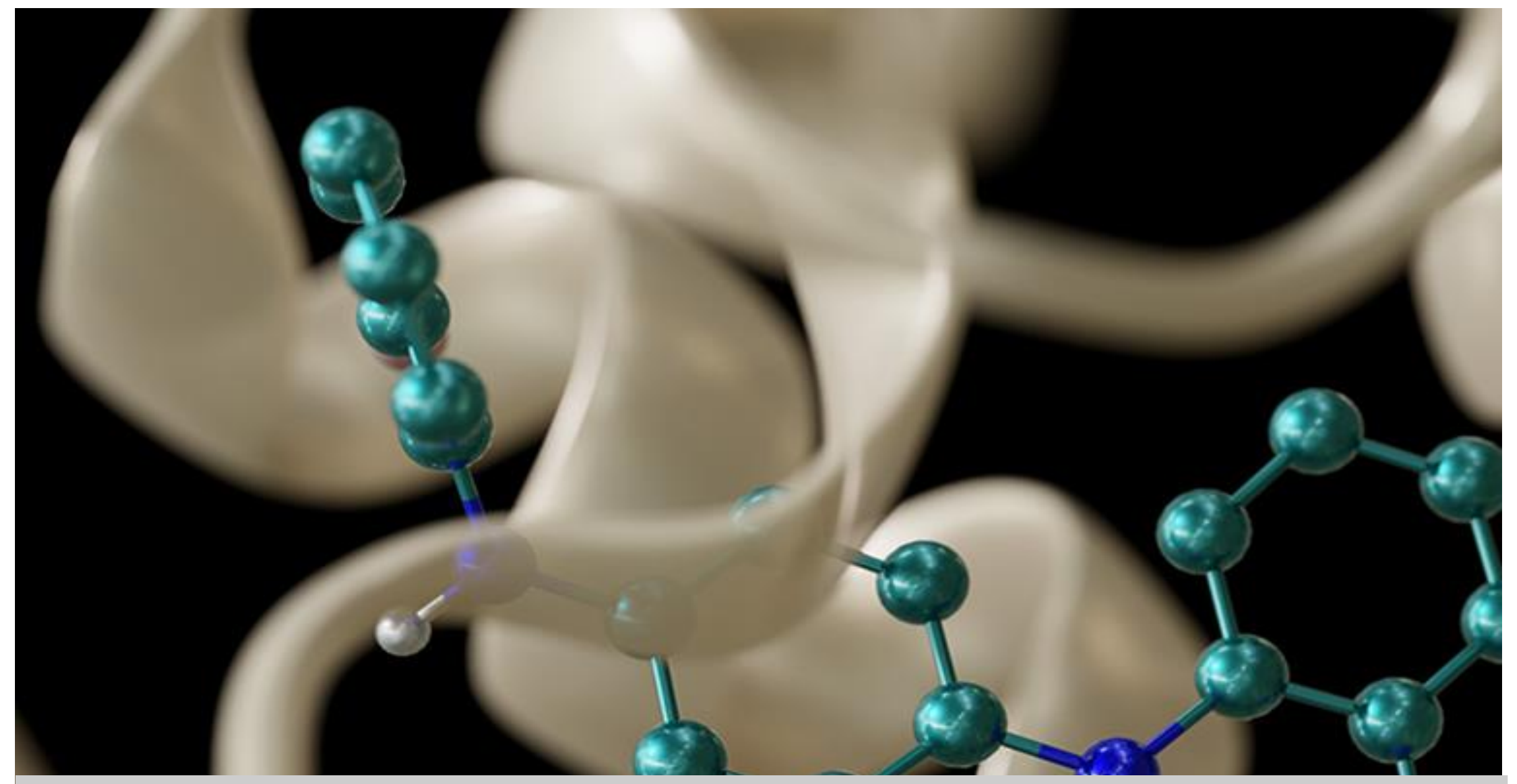
NVIDIA AI Blueprint

Multimodal PDF Data Extraction
for Enterprise RAG



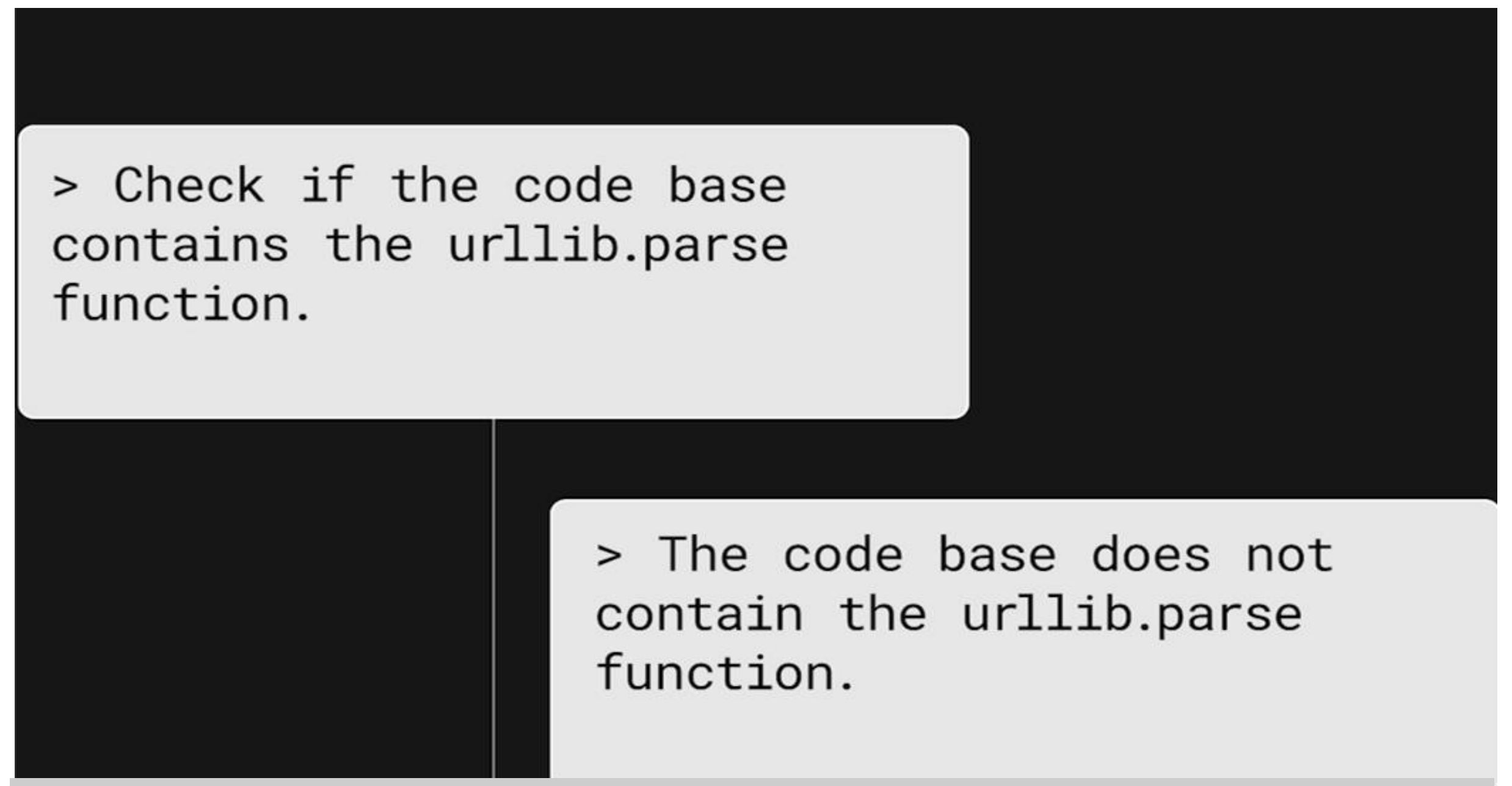
NVIDIA AI Blueprint

Generative Virtual Screening
for Drug Discovery



NVIDIA BioNeMo Blueprint

Vulnerability Analysis
for Container Security

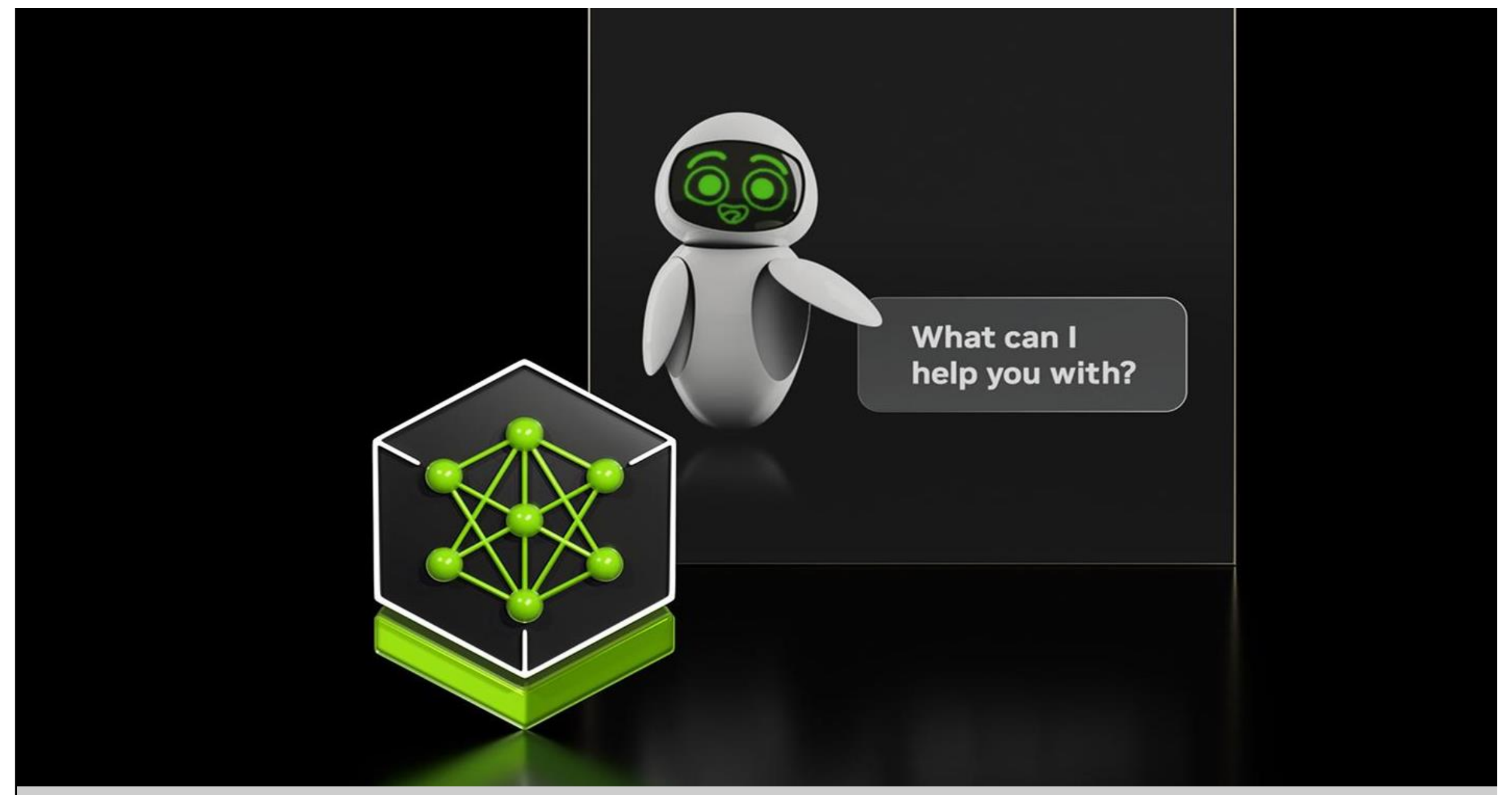


```
> Check if the code base
contains the urllib.parse
function.
```

```
> The code base does not
contain the urllib.parse
function.
```

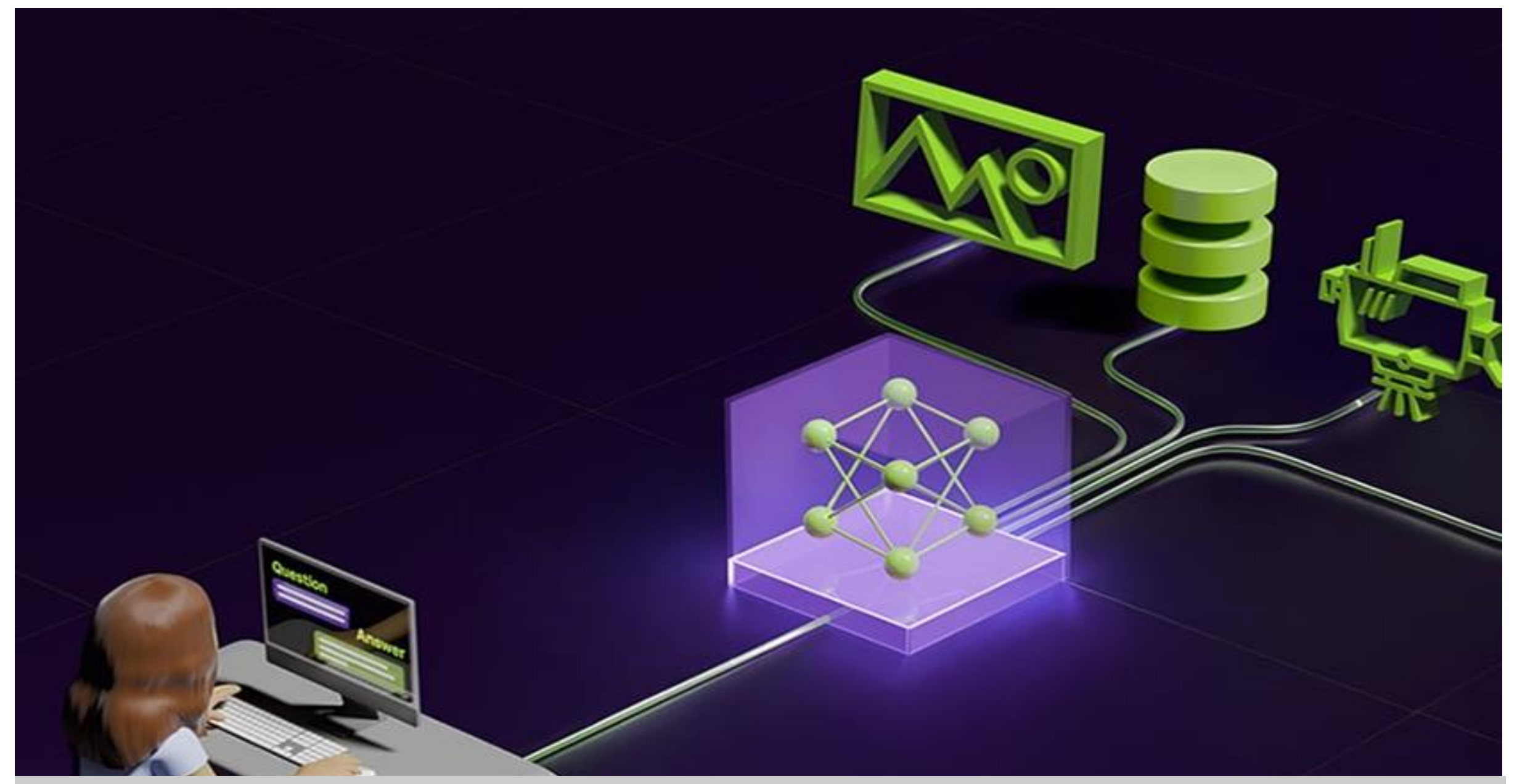
NVIDIA AI Blueprint

AI Virtual Assistants
for Customer Service



NVIDIA AI Blueprint

Visual AI Agent
for Video Search and Summarization



NVIDIA AI Blueprint

3D Conditioning for
Precise Visual Generative AI

NVIDIA Omniverse Blueprint

Build a Digital Twin for
Interactive Fluid Simulation

NVIDIA Omniverse Blueprint

Expanded AI factory portfolio from HPE

for every AI ambition, across clouds, cores and countries

Turnkey AI factory
Enterprises

AI factory at scale
Model builders & SP’s

Sovereign AI factory
Governments, public sector

Common control plane: HPE Morpheus and HPE OpsRamp

Turnkey, engineered systems

Customized, validated solutions

Infrastructure | Software | Services | Ecosystem | Sustainability

The AI factory portfolio

For every AI ambition

Size & Complexity

At scale
Model Builders | Service Providers

and

Sovereign
Sovereign environments | defined borders

Turnkey
(HPE Private Cloud AI)
Traditional Enterprises

Data ingestion → Model development → Training → Fine tuning → Inferencing → Monitoring

Thank You

© 2025 Hewlett Packard Enterprise Development LP