Unleashing the Power of RTX: Accelerating Professional Workflows with NVIDIA RTX 6000 Pro Server Edition and HPE

Mason Wu, NVIDIA Senior Solutions Architect

October 16, 2025

Unleashing the Power of RTX: Accelerating Professional Workflows with NVIDIA RTX 6000 Pro Server Edition and HPE

Mason Wu, NVIDIA Senior Solutions Architect | 2025

Rise of the AI Factory

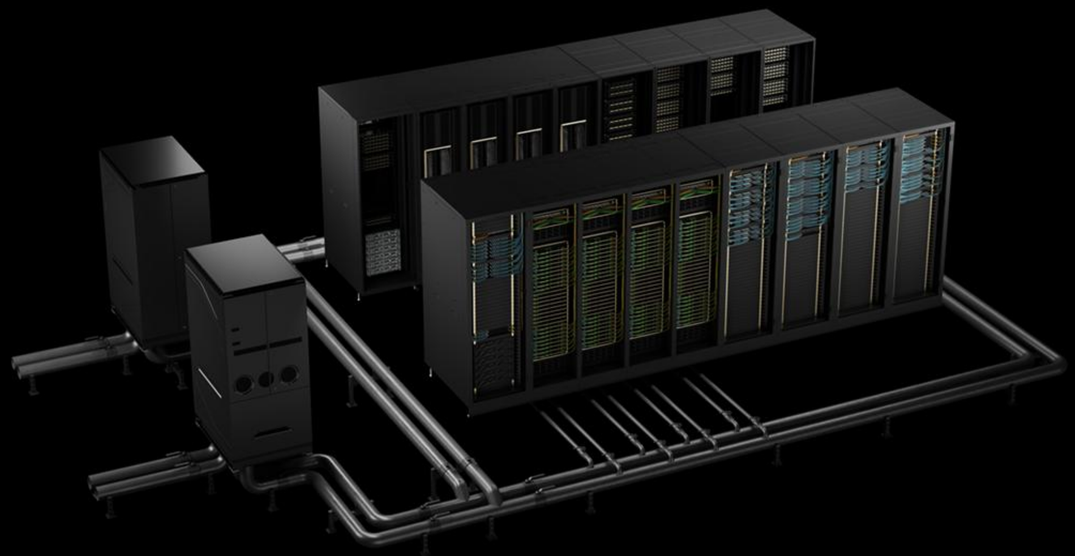
Trillion-Dollar Global IT Investment Shifting to AI Factories

92% of enterprises investing in AI

50% will use AI agents to achieve business value

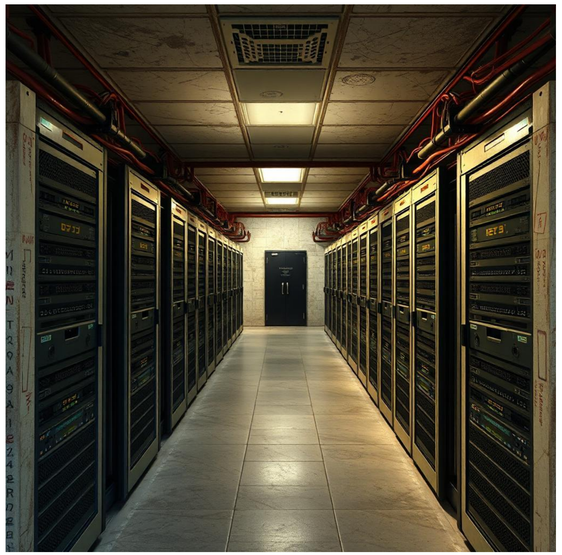
33% find complexity top barrier for adoption

1% have mature AI deployments



Every Enterprise Needs an AI Factory

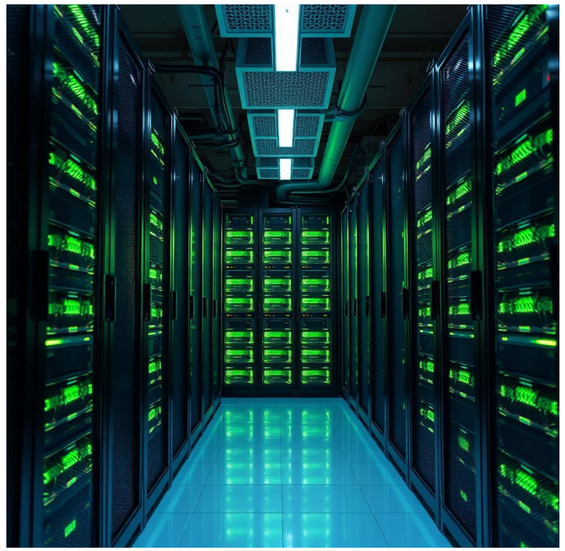
Manufacturing Intelligence at Scale



Traditional Data Center

Transactional Workflows

Electricity & Data →



→ Tokens*, Intelligence & Outcomes

AI Factory

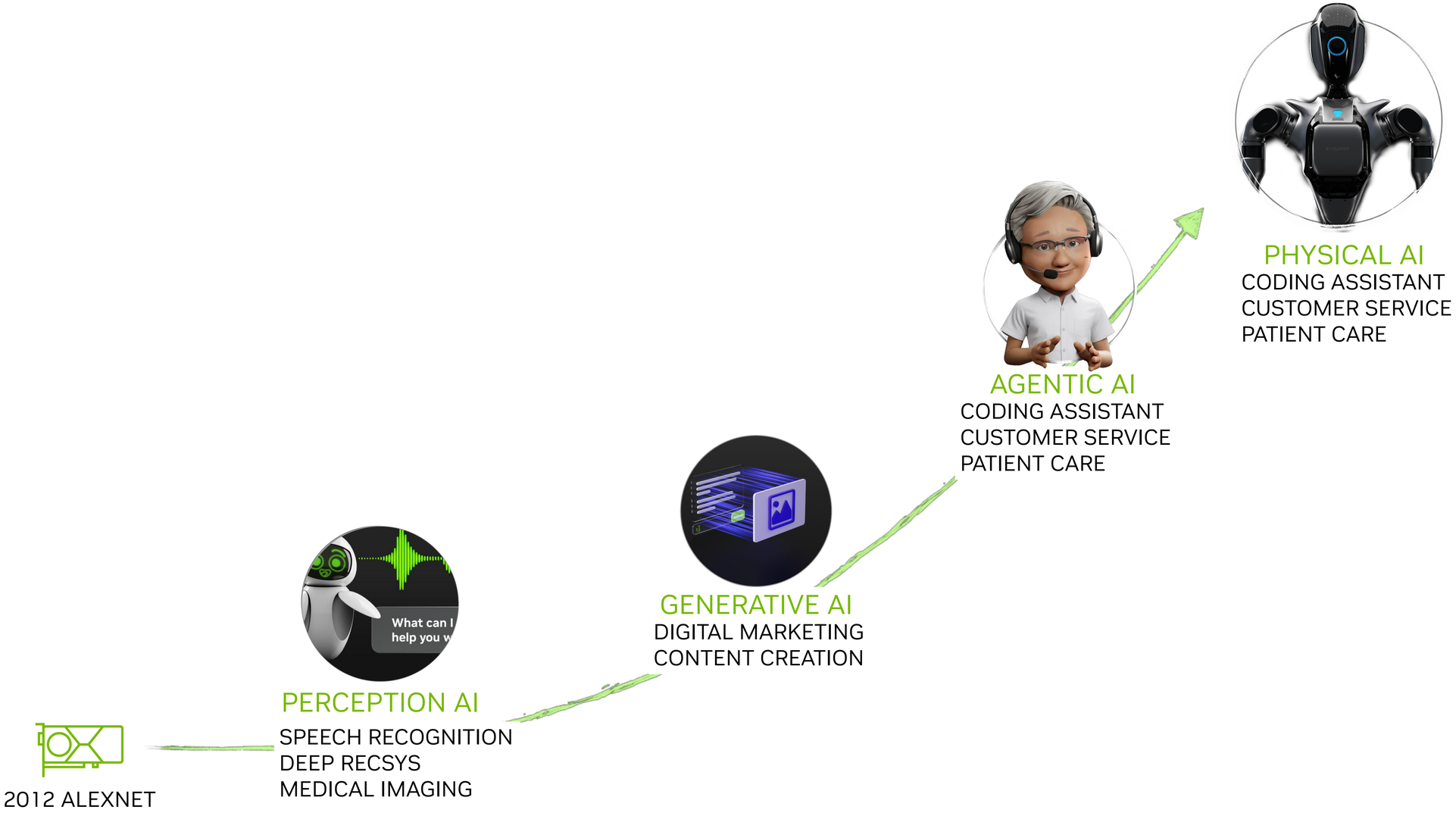
Accelerated Workloads

*What is a token? It's about half a word, so a 10-word ChatGPT query is 20-30 tokens.

NVIDIA

Evolution of AI

Agentic AI Enables More Powerful AI Applications



Agentic AI will Transform the Enterprise

AI Agents will drive performance gains, better problem solving, and faster time to action

Organizations will use agents to achieve faster business value from AI by 2025[1]

Improvement in cycle times, with improved quality, with AI agents[2]

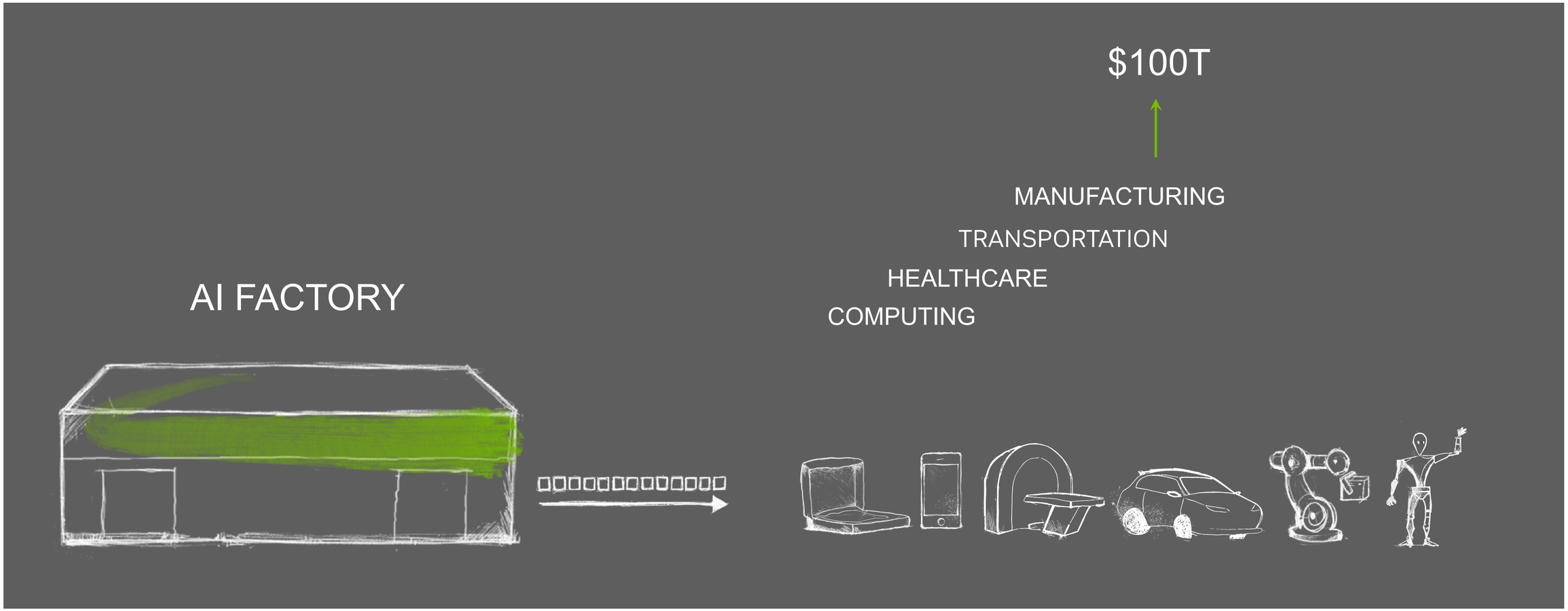
Of enterprise software applications will include agentic AI by 2028[3]

[1] IDC FutureScape: Worldwide Artificial Intelligence and Automation 2025 Predictions US51666724, October 30, 2024
[2] IDC, IDC FutureScape: Worldwide Artificial Intelligence and Automation 2025 Predictions US51666724, October 30, 2024
[3] Gartner®, Innovation Insight: No-Code Agent Builders by Jason Wong, Sohail Majumdar, etc., March 26, 2025

NVIDIA

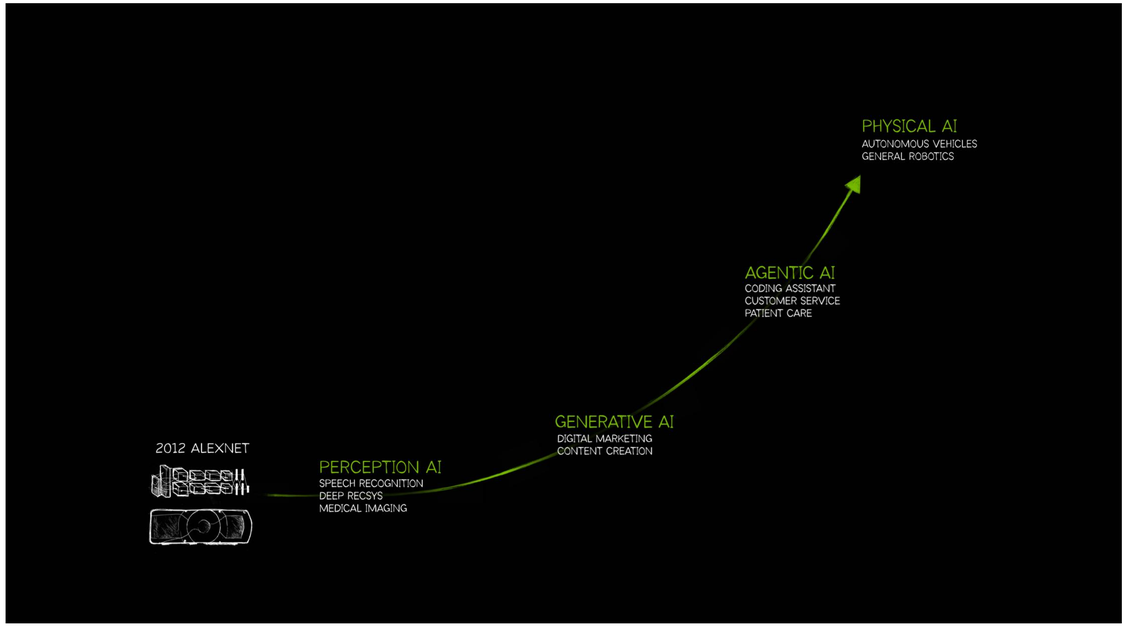
A New Industrial Revolution - Tokens are the New Currency

AI Factories: Converting Data and Electricity into Profit-Generating Assets



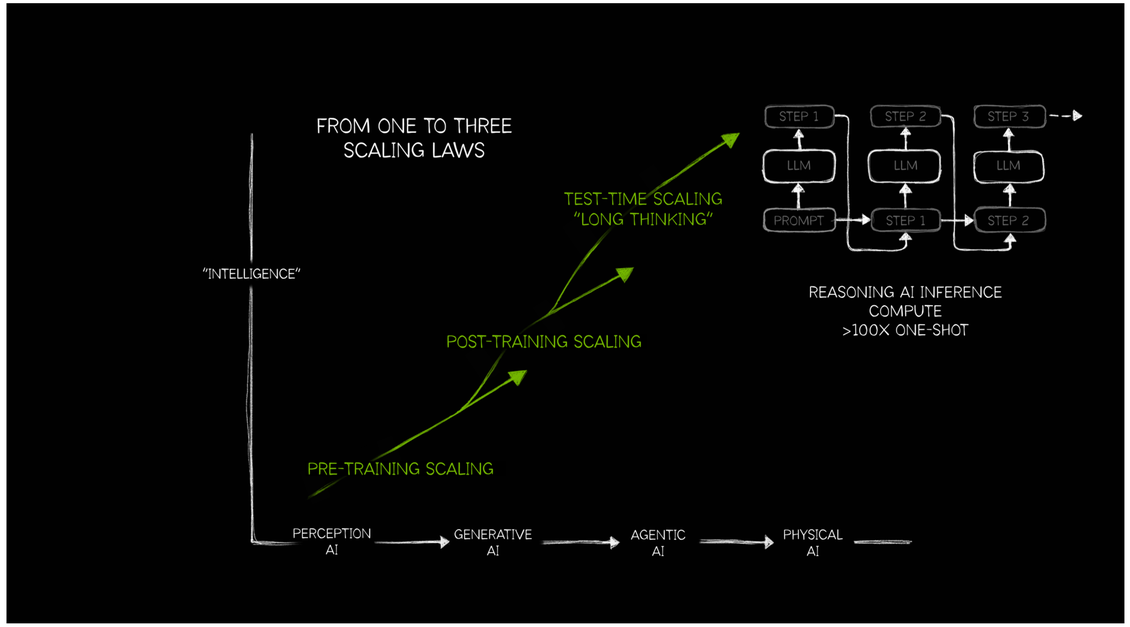
AI is Changing the World, and the World of Business

And it’s changing fast!



AI progress opens new market opportunities

50% of organizations will use AI agents by the end of 2025, and 33% of enterprise SW applications will use agentic AI by 2028.

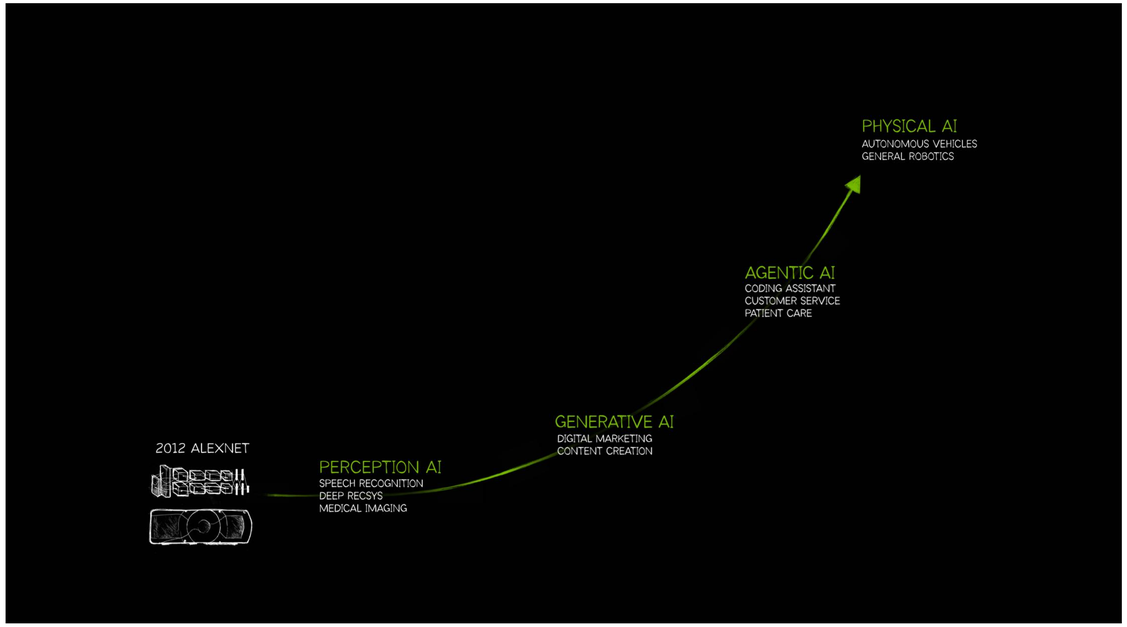


Scaling laws of AI are creating smarter answers

“Long Thinking” focus on deep analysis and thoughtful decision making (Reasoning) vs. a one-shot answer. Reasoning requires >100X compute vs. one-shot.

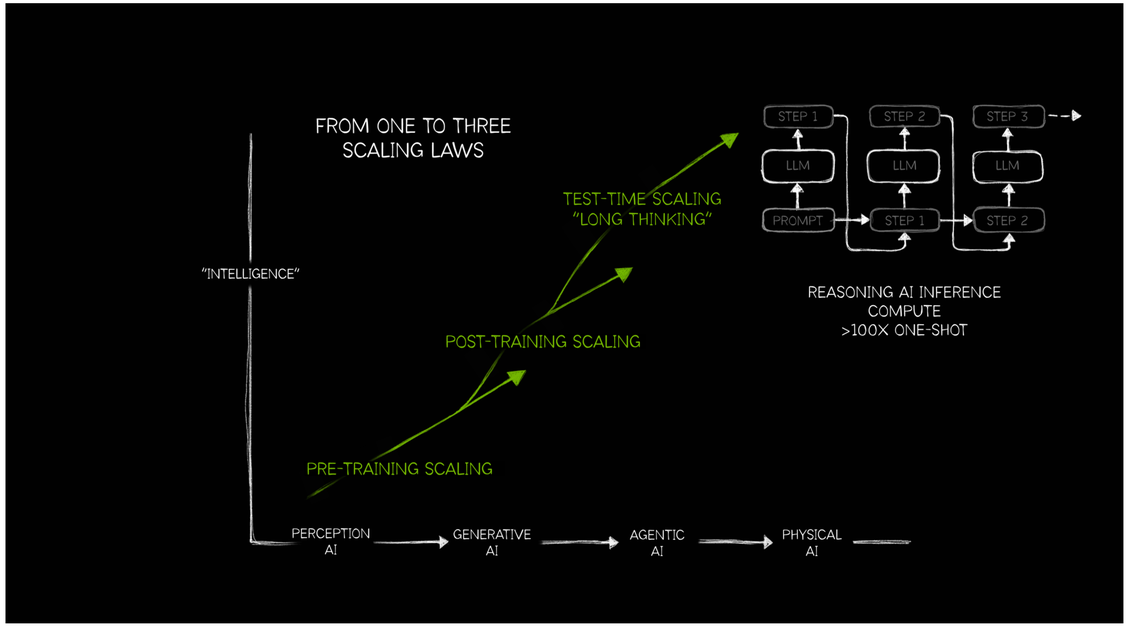
AI is Changing the World, and the World of Business

And it's changing fast!



AI progress opens new market opportunities

50% of organizations will use AI agents by the end of 2025, and 33% of enterprise SW applications will use agentic AI by 2028.

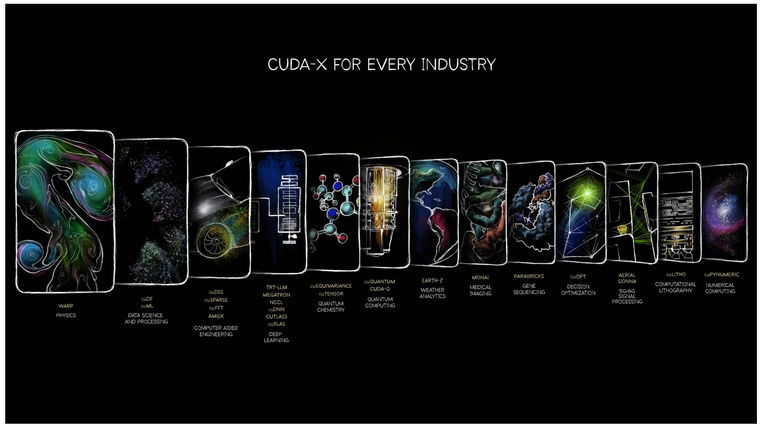


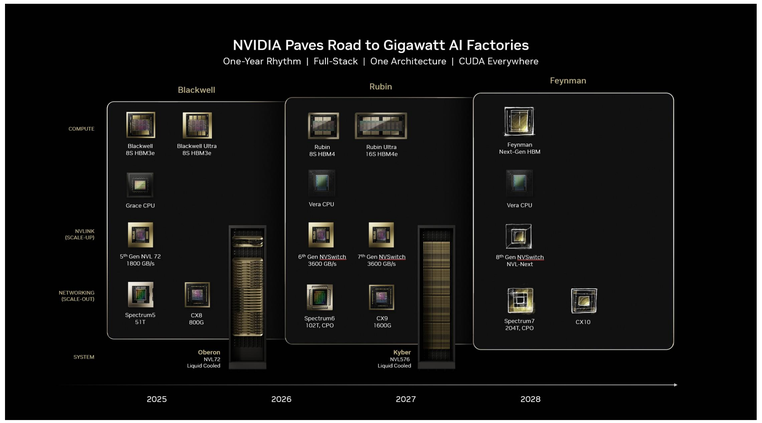
Scaling laws of AI are creating smarter answers

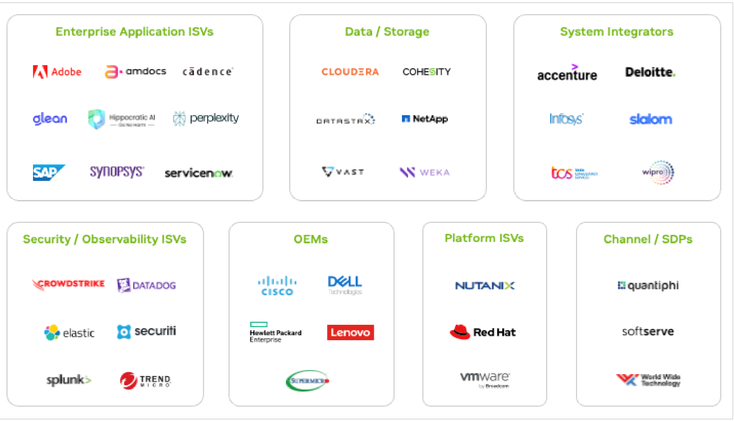
"Long Thinking" focus on deep analysis and thoughtful decision making (Reasoning) vs. a one-shot answer. Reasoning requires >100X compute vs. one-shot.

AI is a Datacenter-Level Problem, and a Full-Stack Problem

And it's very difficult!







You Need Developers

NVIDIA invests in CUDA-X libraries across a wide range of use cases, domains, and industries to enable the world's developers and ISVs.

You Need Datacenter Solutions

AI is a full stack problem, requiring end-to-end accelerated compute and networking hardware, with a huge set of software tools, libraries, etc.

You Need An Ecosystem

AI solutions require consultants, ISVs, data storage, deployment services, management tools, & other services to ensure customer success.

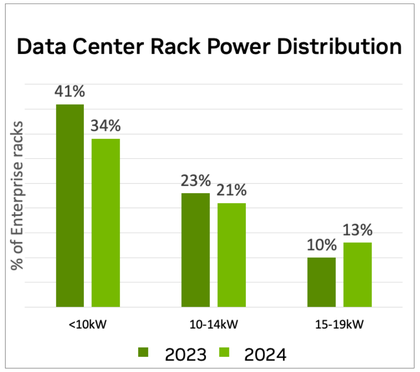
NVIDIA RTX PRO Server

Enterprise IT Constraints

Meeting Enterprises Where They Are

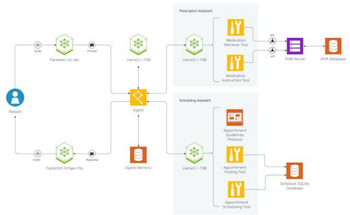
Data Center

~70%[1] of data centers are below 20kW/rack



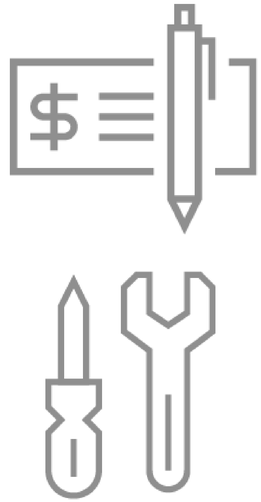
62% of data centers will *NOT* have liquid cooling by 2026[2]

Workload



Most enterprises have mixed and evolving workloads

Financial & Tooling



Varied budget and software tooling requirements

1 Source: Uptime Institue Global Data Center survey 2023 & 2024 (n=850)
2. Source: The Register survey 2024 (n=812)

From Desktop to AI Factory Blackwell is Everywhere

DGX Spark

DGX Station

GB200 / GB300 NVL72

Arm Platforms

RTX PRO Laptop

RTX PRO Workstation

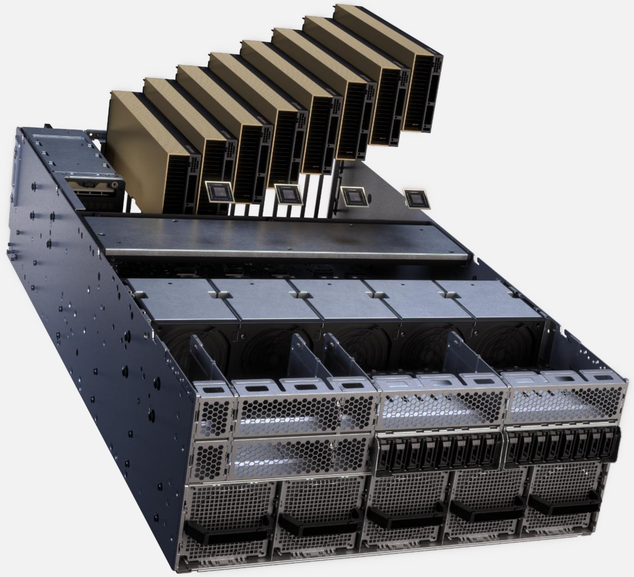
RTX PRO Blackwell Server

DGX B200 / B300

x86 Platforms

Not All Factories for Manufacturing are the Same

Not all Factories for AI are either



RTX PRO Server

Multi-Workload Acceleration,
Enterprise DC Compatibility

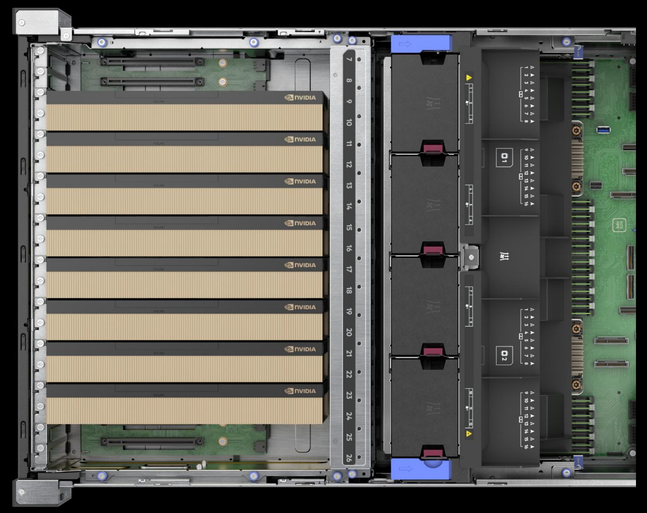
HGX B200

AI-Optimized Performance for Training and
Large Model Inference

GB200 NVL72

Exascale Computer in a Rack for Massive
LLM Training and Inference

NVIDIA Certified



NVIDIA RTX PRO 6000 Blackwell Server Edition

HPE ProLiant 380A Gen12

Canonical

NUTANIX

Red Hat

vmware by Broadcom

HPE

Enterprise Server for NVIDIA Accelerated Computing

HPE ProLiant Compute DL380a Gen12

NVIDIA Blackwell for the Enterprise

RTX PRO Server

Designed for enterprise IT requirements

Multi-user AI with MIG and virtualization

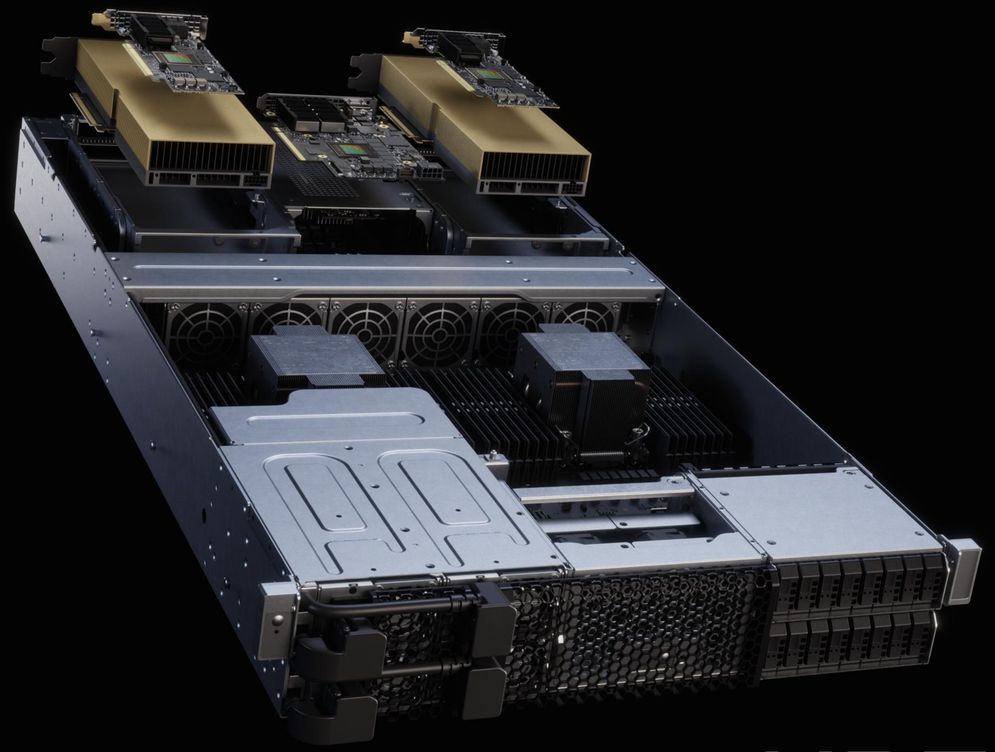
Ease of Orderability / Install / Deploy / Support

Air-Cooled / PCIe / x86 / 7kW per Node

RTX PRO Server 7kW power requirement based on 8x RTX PRO 6000 Blackwell Server Edition GPUs

NVIDIA

NVIDIA Certified



HPE

Announcing New Mainstream NVIDIA RTX PRO Servers

High-volume servers from global system partners

Introducing New 2U Configurations

Most Popular Form Factor (~57% of Rack Servers)

2x RTX PRO 6000 Blackwell Server Edition GPUs

Multi-Workload Acceleration

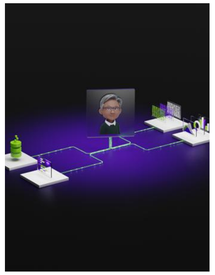
15X price-performance and energy efficiency vs CPU-only server

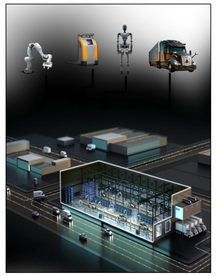
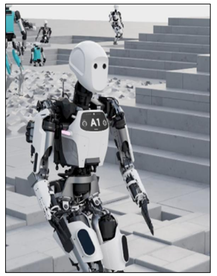
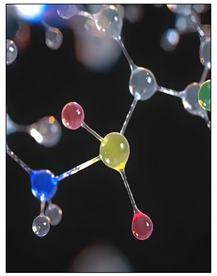
NVIDIA RTX PRO Server (2U Chassis, 2x RTX PRO 6000 Blackwell Server Edition GPU, 2x x86 CPU, 3kW) vs. 2x x86 CPU (2U Chassis, 1.1kW); Multi-Workload Relative Performance; Geomean of measured performance speedups for HPC Applications (FP32), Rendering (VRed), Spark RAPIDS, Encode Streams (1080p30, HEVC). Shown for representation only. Please contact partners for pricing.

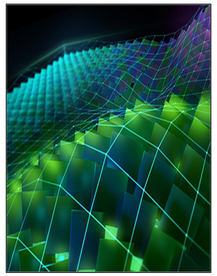
NVIDIA

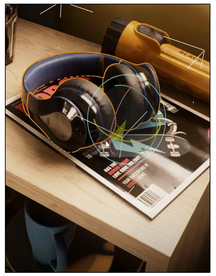
Modern Enterprises Have Diverse Accelerated Workloads

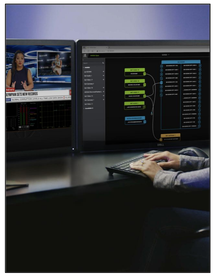
From Agentic AI and Physical AI to AI-Enabled Applications









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 DL385 Gen11

HPE ProLiant 380a Gen12

NVIDIA AI Computing by HPE

Co-developed solutions to simplify enterprise AI

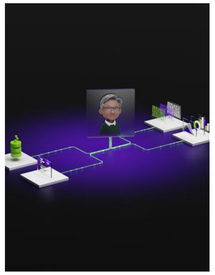
Turnkey Private Cloud	PEOPLE	Inference, Tuning & Training
AI Services & Training	TECHNOLOGY	Virtual Assistants
AI Optimized Systems	ECONOMICS	Process Automation
Enterprise-grade Control		Content & Product Creation

NVIDIA RTX PRO 6000 Blackwell Server Edition

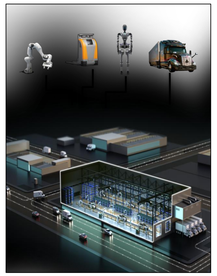
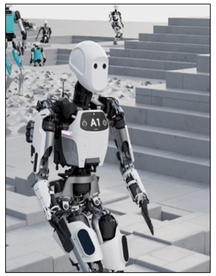
NVIDIA

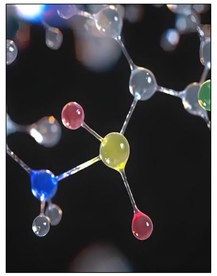
NVIDIA RTX PRO Server

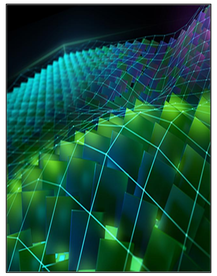
Exceptional Performance for RTX Graphics, Industrial AI, and Generative AI

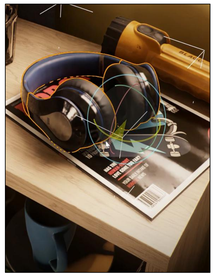


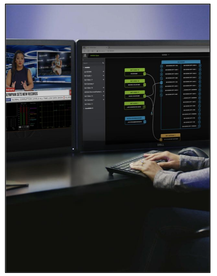

AGENTIC AI



INDUSTRIAL & PHYSICAL AI




SCIENTIFIC COMPUTING, DATA ANALYTICS, & SIMULATION



VISUAL COMPUTING



ENTERPRISE APPLICATIONS

6X
Throughput
LLM Inference

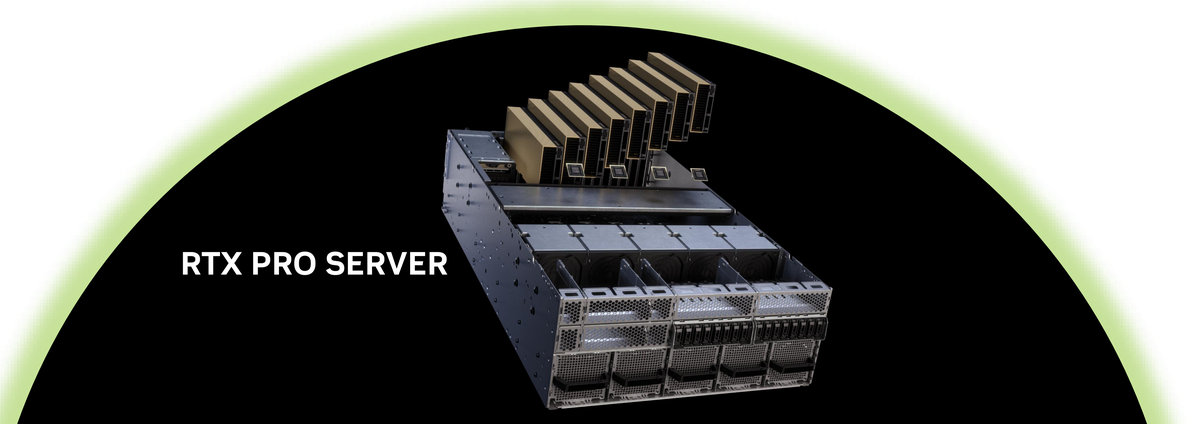
4X
Faster
Synthetic Data Generation

7X
Faster
Genome Sequence Alignment

2X
Throughput
Engineering Simulation

6X
Higher FPS
Real-Time Rendering

4X
User Density
Virtual Workstations



NVIDIA RTX PRO 6000 Blackwell Server Edition vs. NVIDIA L40S

Projected Performance. Subject to change.
1. LLama3 70B Inference; 8K/256, 20 t/s/usr
2. NVIDIA Cosmos 7B; Text-Video Generation, 2.5s 720p Video

Measured Performance
3. Smith & Waterman (TCUPs); 8-bits; RTX PRO 6000 vs L40S
4. RTM, Isotropic Radius; Mcells/s; FP32
5. Omniverse; Debrecan; Real Time Rendering- FPS. RT2+DLSS4 On vs DLSS3
6. # of concurrent vGPU users supported

NVIDIA

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

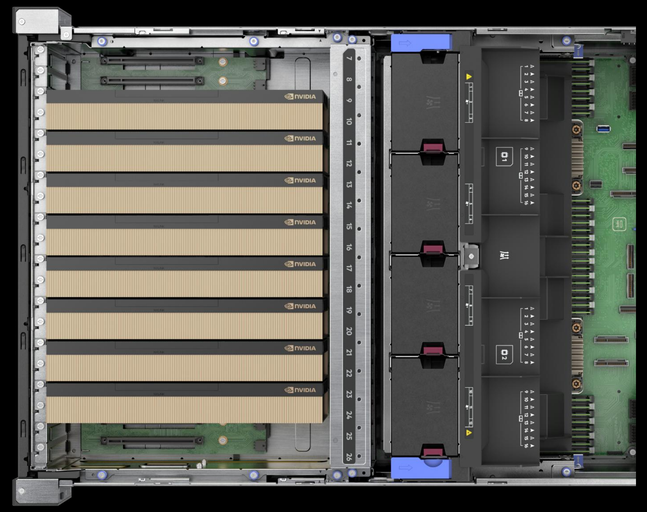
- 96GB GDDR7,1.6 TB/s Memory BW, 128MB L2 Cache
- Multi-Instance GPU (MIG), TEE Confidential Compute

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

NVIDIA Certified



NVIDIA RTX PRO 6000 Blackwell Server Edition

Canonical

NUTANIX

Red Hat

vmware by Broadcom

HPE

HPE ProLiant 380A Gen12

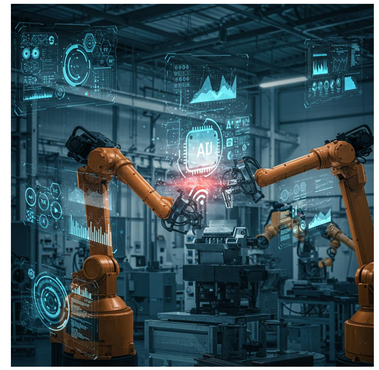
Enterprise Server for NVIDIA Accelerated Computing

HPE ProLiant Compute DL380a Gen12

NVIDIA Blackwell for the Enterprise

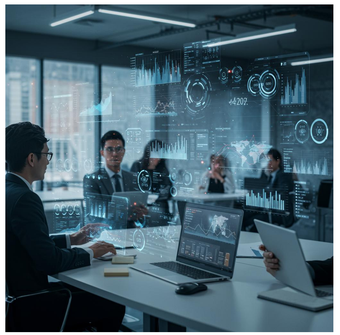
RTX Graphics & Visual Computing

Industrial and Physical AI



Enterprise HPC

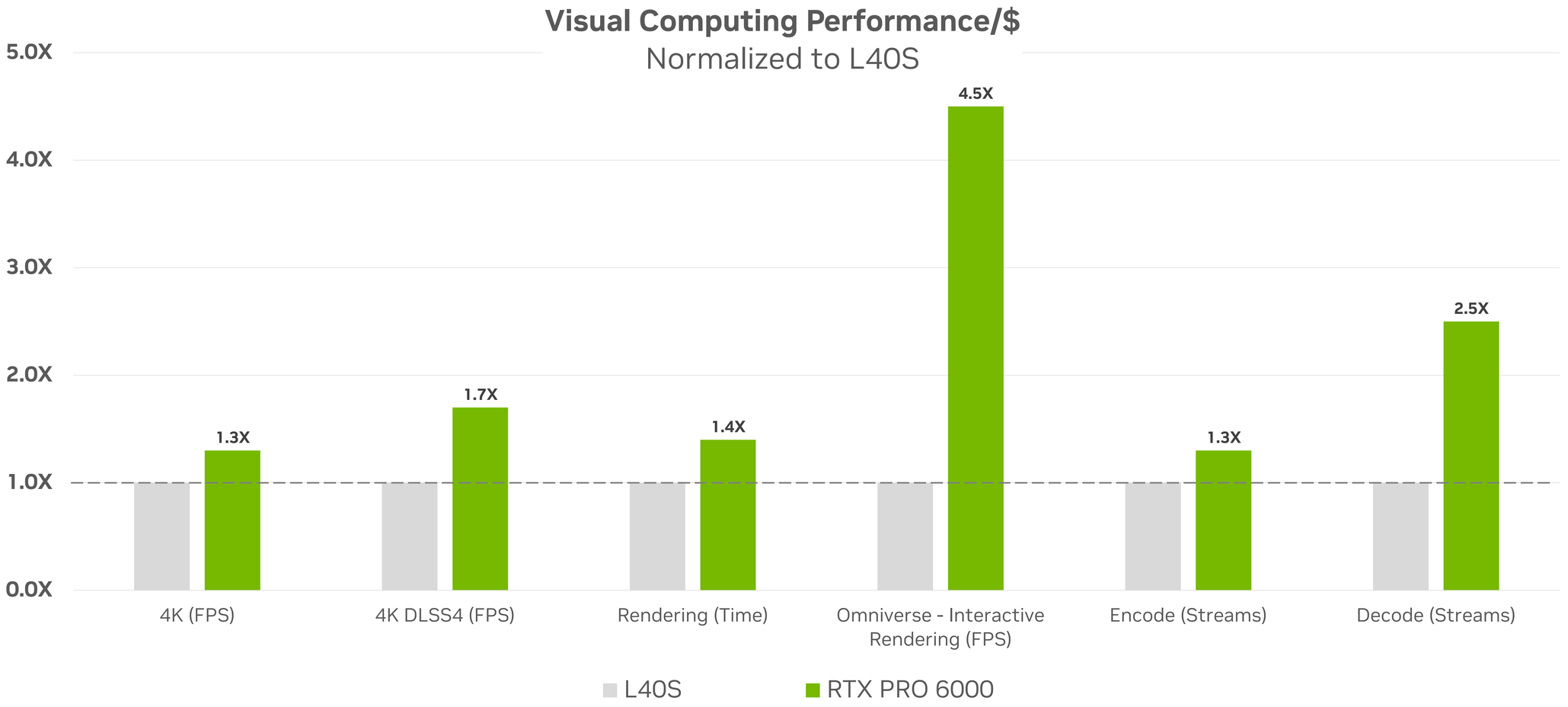
Enterprise AI



NVIDIA

Best Server for RTX Graphics and Visual Computing

RTX PRO Server Up to 4.5X Better Price Performance



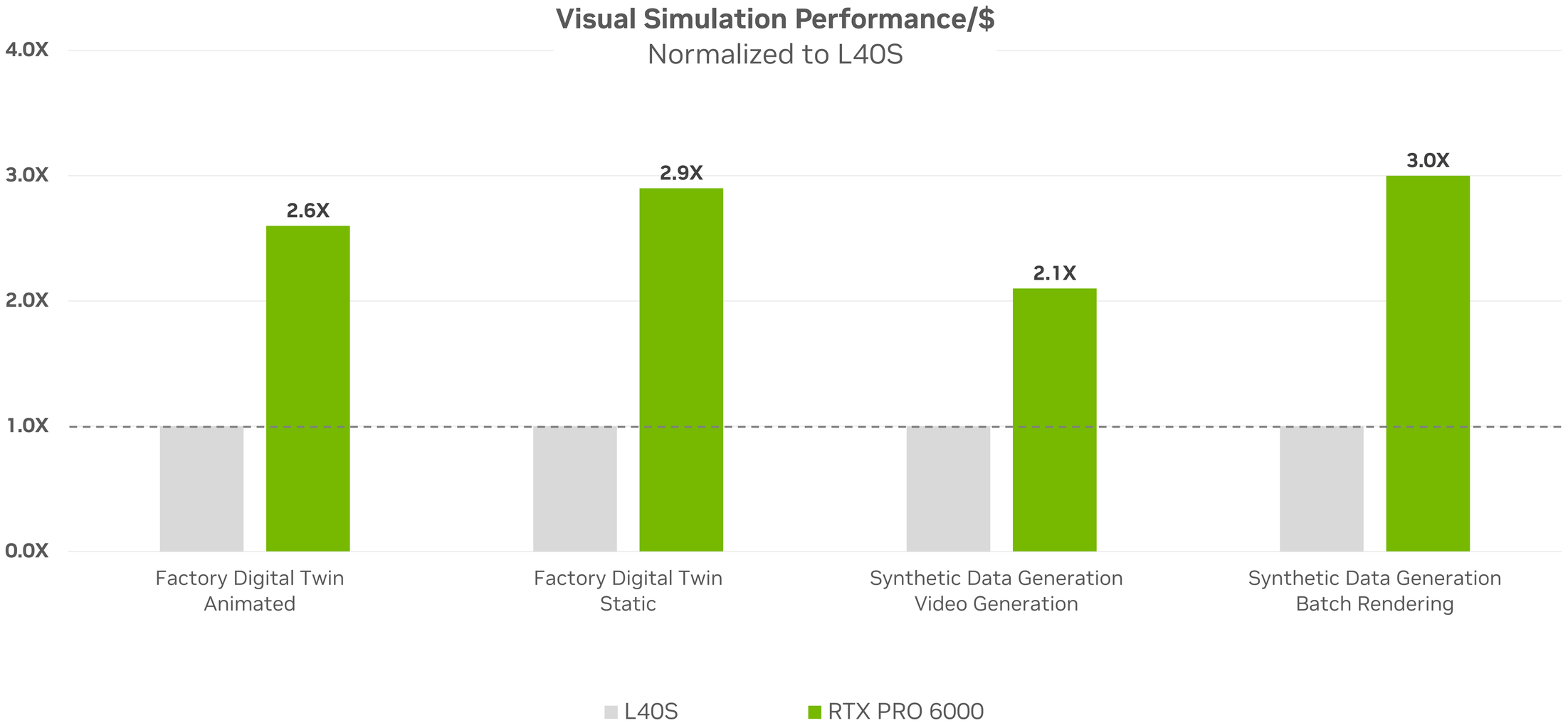
Performance/$ = Performance / TCO for a Single Node (Server + Power Costs) for L40S compared to RTX PRO 6000 Blackwell Server Edition

Visual Computing Ecosystem

Applications for Creative and Technical Professionals

Best Server for Industrial AI and Physical AI

RTX PRO Server Up to 3X Better Price Performance



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

NVIDIA RTX PRO Server

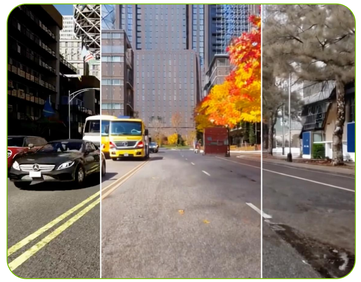
World-class performance and scalability for the era of industrial & physical AI

Accelerate Complex Industrial & Physical AI Workloads

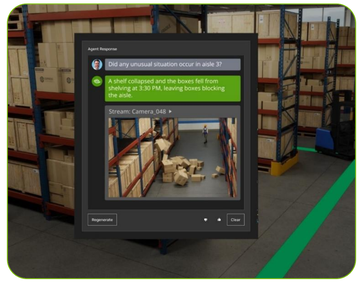
CAE Digital Twins

Factory Digital Twins

Robotics Simulation

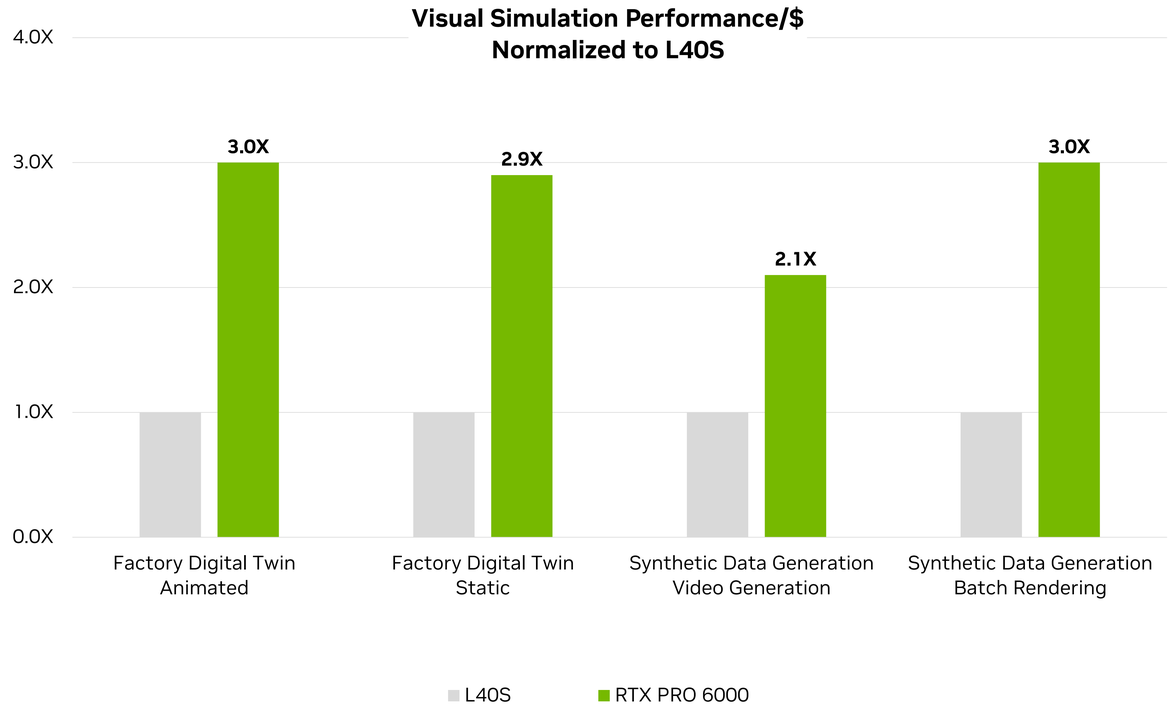


Synthetic Data Generation



Vision AI Agents

Up to **3X** Better Price-Performance



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

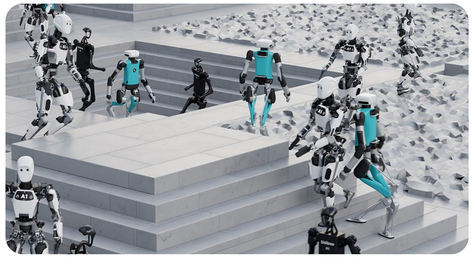
NVIDIA

Industrial and Physical AI Ecosystem

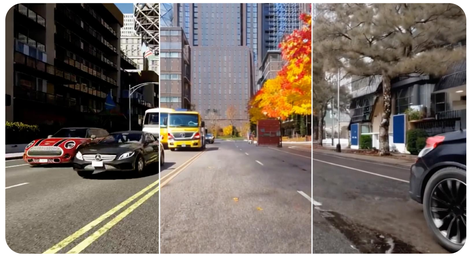
Example RTX PRO Server Applications and Workloads

CAE Digital Twins
NVIDIA CUDA-X and Omniverse

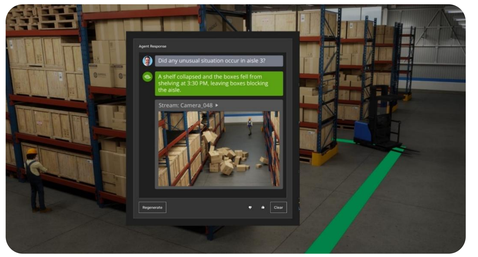
Factory Digital Twins
NVIDIA Omniverse and Metropolis



Robotics Simulation
NVIDIA Omniverse and Isaac Sim



Synthetic Data Generation
NVIDIA Omniverse and Cosmos



Vision AI Agents
NVIDIA Metropolis and Cosmos

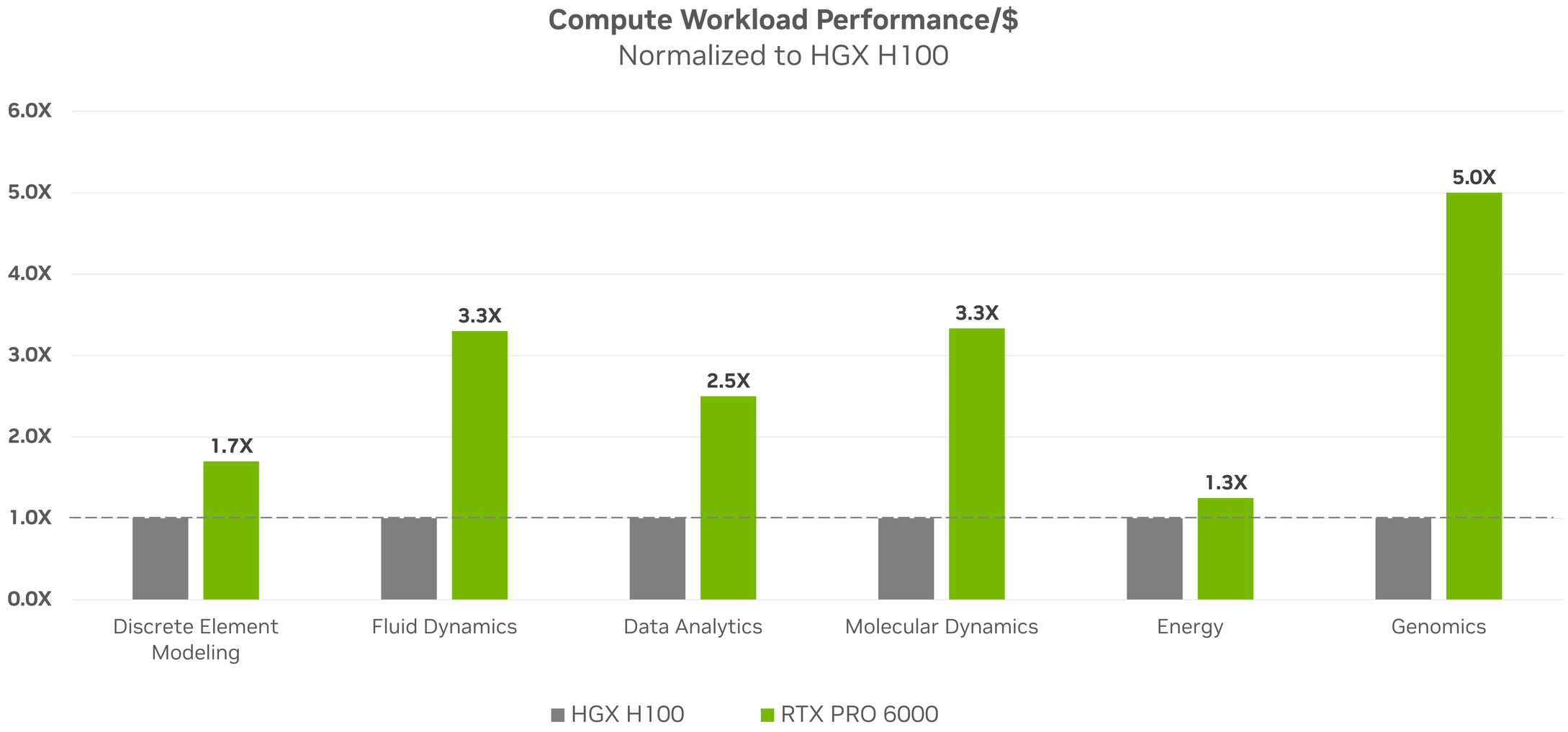
ALTAIR · Ansys · AUTODESK · cadence · DASSAULT SYSTEMES

HEXAGON · idealworks · MuJoCo · ptc · roboflow

Rockwell Automation · SIEMENS · SIMSCALE · VISUAL COMPONENTS

Best Server for Enterprise HPC

RTX PRO Server Up to 5X Better Price Performance



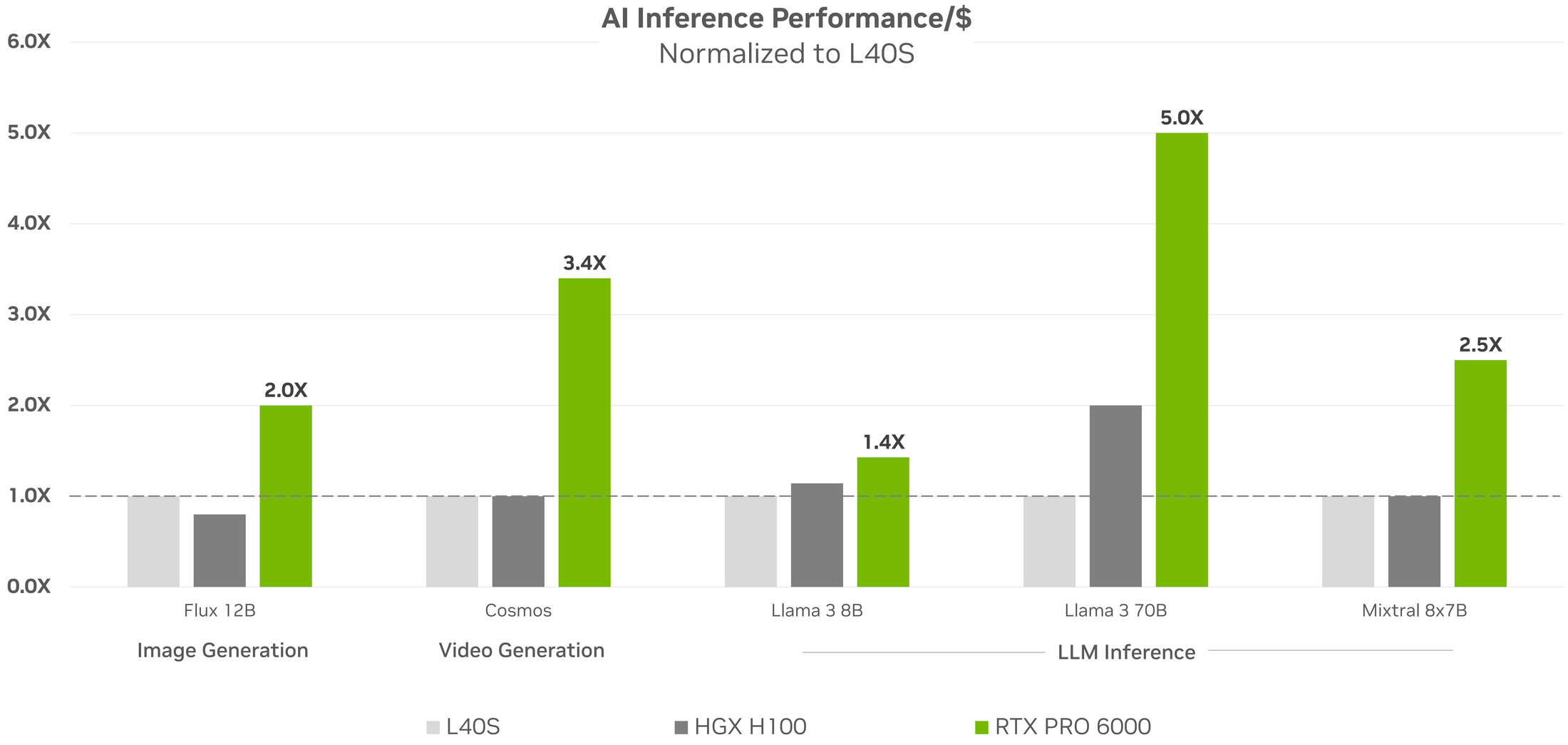
Performance/$ = Performance / TCO for a Single Node (Server + Power Costs) for HGX H100 compared to RTX PRO 6000 Blackwell Server Edition

Enterprise HPC Ecosystem

Accelerated Applications for Industries

Best Server for Enterprise AI

RTX PRO Server Up to 5X Better Price Performance



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

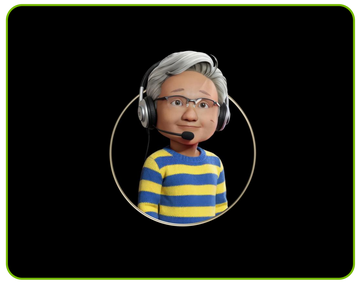
NVIDIA RTX PRO Server

Breakthrough performance for agentic and generative AI

Accelerate Agentic AI and Generative AI Applications

Employee Support

Content Creation

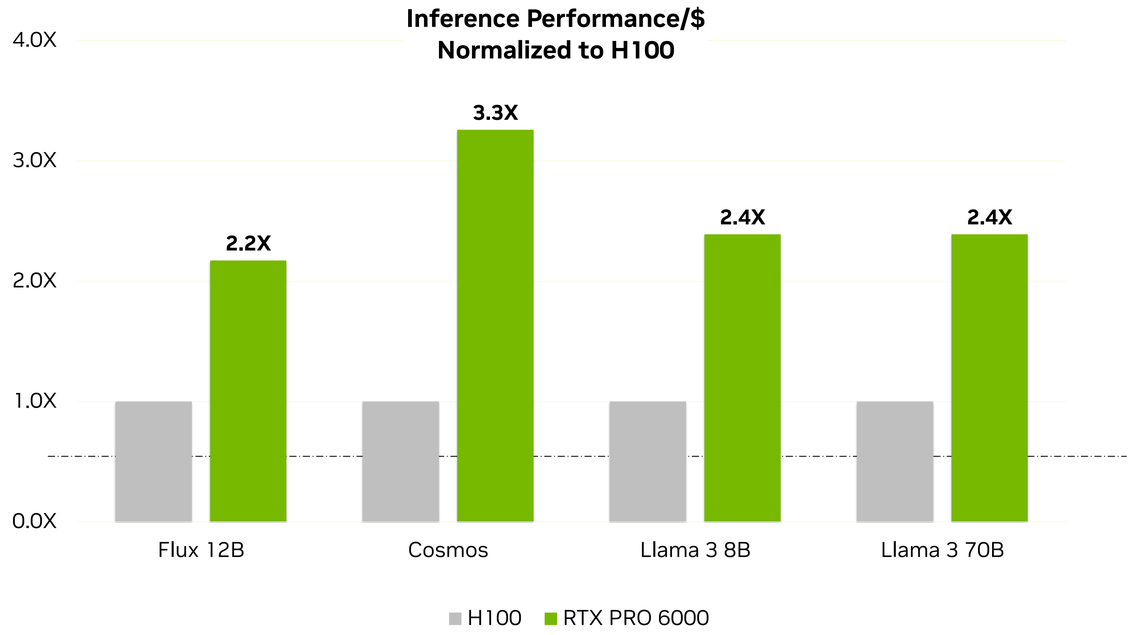


Customer Service

Coding

Deep Research

Over 3X Better Price-Performance



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

Image Generation
Flux 12B – 1024x1024 Images Per Min- FP4
Video Generation/Synthetic Data Generation
COSMOS VFM 7B – Text-Video Generation- 720p -RTX PRO FP4, HGX H100 (FP8)
LLM Inference
LLAMA3 8B – 2k/256, TPS/GPU; RTX PRO FP4, HGX H100 (FP8)
LLAMA3 70B – 8k/256, TPS/GPU; RTX PRO FP4, HGX H100 (FP8)

NVIDIA

Enterprise AI Ecosystem

Example RTX Server PRO Server AI Agents

Supply Chain Agent

Security Agent

Customer Service Agent

Coding Agent

Deep Research Agent

Bio-Medical Agent

Agent Applications

amdocs · cisco · IQVIA · tabnine · UiPath · UNIPHORE

AIOps

CROWDSTRIKE · dataiku · DataRobot · DATASTAX (an IBM Company) · dynatrace

elastic · H2O.ai · JFrog · TREND MICRO · Weights & Biases by CoreWeave

Orchestration

Canonical · NUTANIX · Red Hat · vmware by Broadcom

Blueprints

NVIDIA Data Flywheel Blueprint	NVIDIA AI-Q Blueprint
Refine AI Agents through Continuous Model Distillation with Data Flywheels	Build AI Agents for Enterprise Research

Business case for consolidation – simplify, accelerate and save

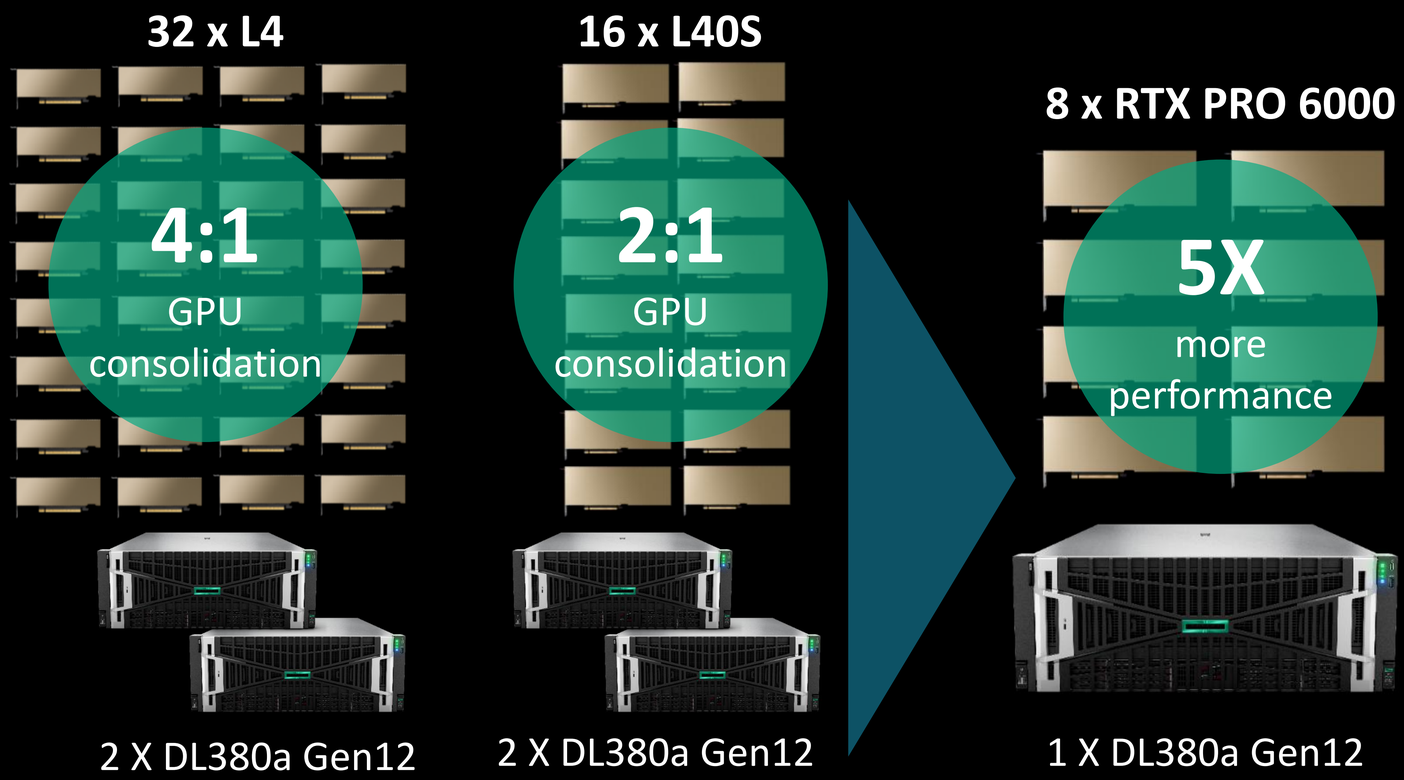
HPE ProLiant Compute DL380a Gen12 + NVIDIA RTX PRO 6000 Blackwell Server Edition GPU

Up to:

5X more performance

48% lower cost

25% power savings



Costs savings	**41%**	**48%**
Power savings		**25%**

Hewlett Packard Enterprise | NVIDIA

*NVIDIA RTX PRO 6000 performance 2X versus L40S (Actual 1.7X-6.8X), 4X versus L4 (Actual 4.4X - 7.6X) - blogs.nvidia.com/blog/rtx-pro-6000-blackwell-server-edition/
2:1 Consolidation, cost and power savings are based on based on comparison 1 x DL380a Gen12 with eight NVIDIA RTX PRO 6000 versus 2 X DL380a Gen12 each with sixteen NVIDIA L4 or 8 L40S GPU. Costs savings based on list price and subject to change without notice.

NVIDIA Enterprise AI Factory Options

NVIDIA Blackwell Accelerated Computing Comparison

	RTX PRO Server	HGX B200
Industrial AI / Physical AI / Digital Twin	☆☆☆☆☆	---
Graphics / Rendering / VDI	☆☆☆☆☆	---
Video Agents (VSS) / Video Analytics	☆☆☆☆☆	☆☆
Inference (Multimodal AI, Generative AI, Agentic AI) – Perf/$	☆☆☆☆☆	☆☆☆☆☆
Fine Tuning	☆☆☆	☆☆☆☆
Single/Mixed Precision HPC[1] Applications- Perf/$	☆☆☆☆	☆☆☆
# GPUs per node	2-8	8
Power Requirement[2]	Up to 7 kW*	14 kW
Cooling	Air Cooled	Air or Water Cooled

1. FP64 not supported on RTX PRO
2. RTX PRO Server 7kW power requirement based on 8x RTX PRO 6000 Blackwell Server Edition GPUs

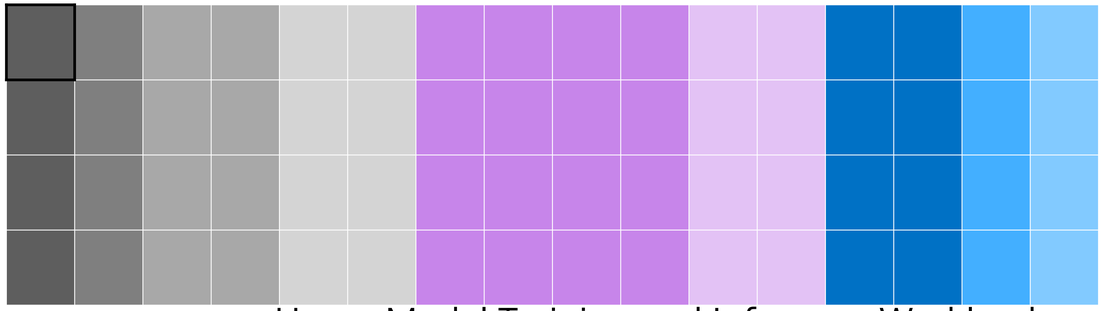
Example AI Factory Workload Projection

World-Class AI Factories for Multiple Enterprise Workloads

HGX B200

AI Factory Workloads

1 Node

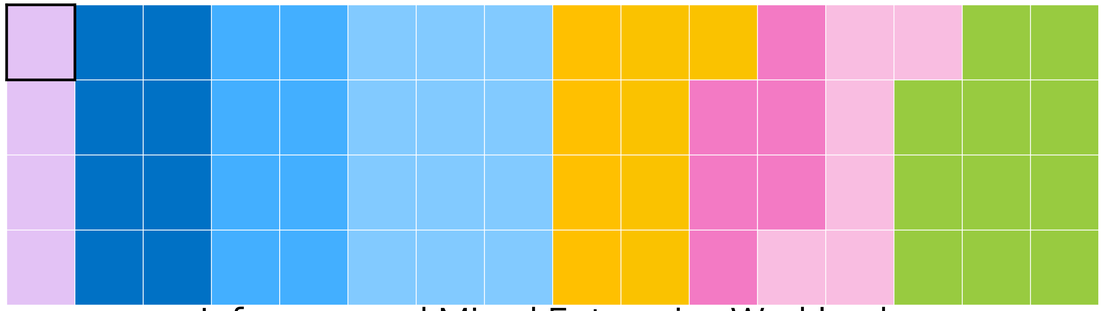


Heavy Model Training and Inference Workloads

RTX PRO Server

AI Factory Workloads

1 Node



Inference and Mixed Enterprise Workloads

Training	Fine Tuning	Inference	RAG/Embedding	Data Analytics	Simulation	Visual Computing
Model Size: Medium-→ Small	Model Size: Medium-→ Small	AI Agents, SDG, VSS, CV, ReccSys	Multimodal Retrieval Pipelines	Spark RAPIDS, Data Science	CAE, CFD, Scientific Simulation	Digital Twins, CAD, Rendering, VDI

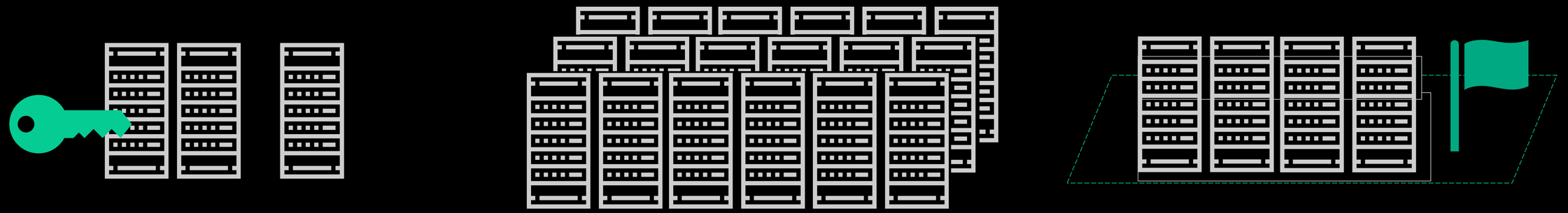
Illustrative chart for mapping workloads within an AI Factory for a representative Enterprise customer

Expanded AI factory portfolio from HPE

for every AI ambition, across clouds, cores and countries

Turnkey AI factory	**AI factory at scale**	**Sovereign AI factory**
Enterprises	Model builders & SP's	Governments, public sector

Common control plane: HPE Morpheus and HPE OpsRamp



Turnkey, engineered systems

Customized, validated solutions

Infrastructure | Software | Services | Ecosystem | Sustainability

Thank You

Thank You