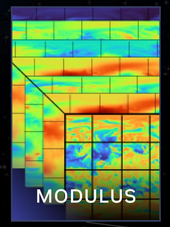
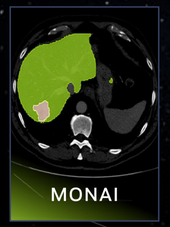
NVIDIA’s Blackwell Architecture, Powering the AI Factory

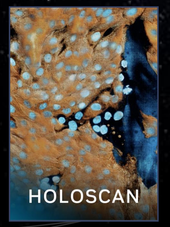
Richard Wright | AI Day, Prague - October 23rd, 2024


NVIDIA

APPLICATION FRAMEWORKS







PLATFORM

NVIDIA AI

NVIDIA OMNIVERSE

ACCELERATION LIBRARIES

RTX

CUDA-X

CUDA

SYSTEM SOFTWARE

Magnum IO

DOCA

Base Command

Forge

HARDWARE

RTX DGX HGX EGX OVX AGX MLNX

GPU CPU DPU NIC SWITCH SOC

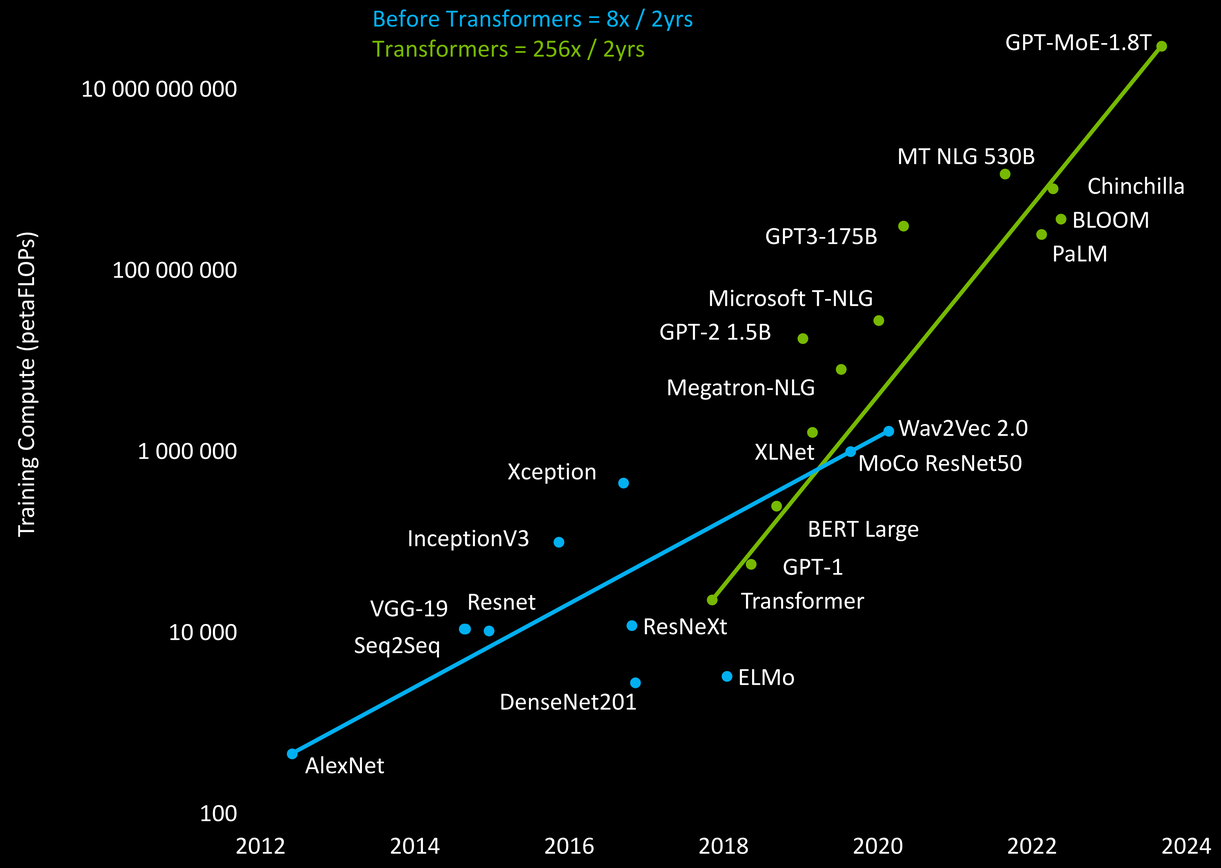
"A NEW INDUSTRIAL REVOLUTION"

AI FACTORY

$100T

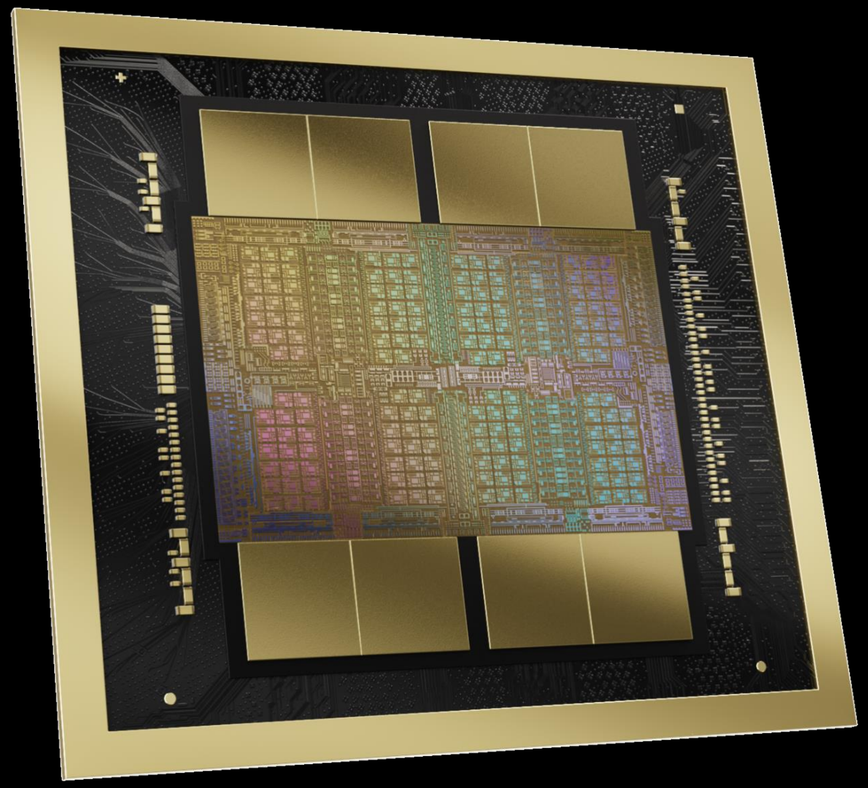
MANUFACTURING
TRANSPORTATION
HEALTHCARE
COMPUTING

Explosive Growth in AI Computational Requirements



Announcing NVIDIA Blackwell

The Engine of the New Industrial Revolution



Built to Democratize Trillion-Parameter AI

20 PetaFLOPS of AI performance on a single GPU

4X Training | 30X Inference | 25X Energy Efficiency & TCO

Expanding AI Datacenter Scale to beyond100K GPUs

AI SUPERCHIP
208B Transistors

2nd GEN TRANSFORMER ENGINE
FP4/FP6 Tensor Core

5th GENERATION NVLINK
Scales to 576 GPUs

RAS ENGINE
100% In-System
Self-Test

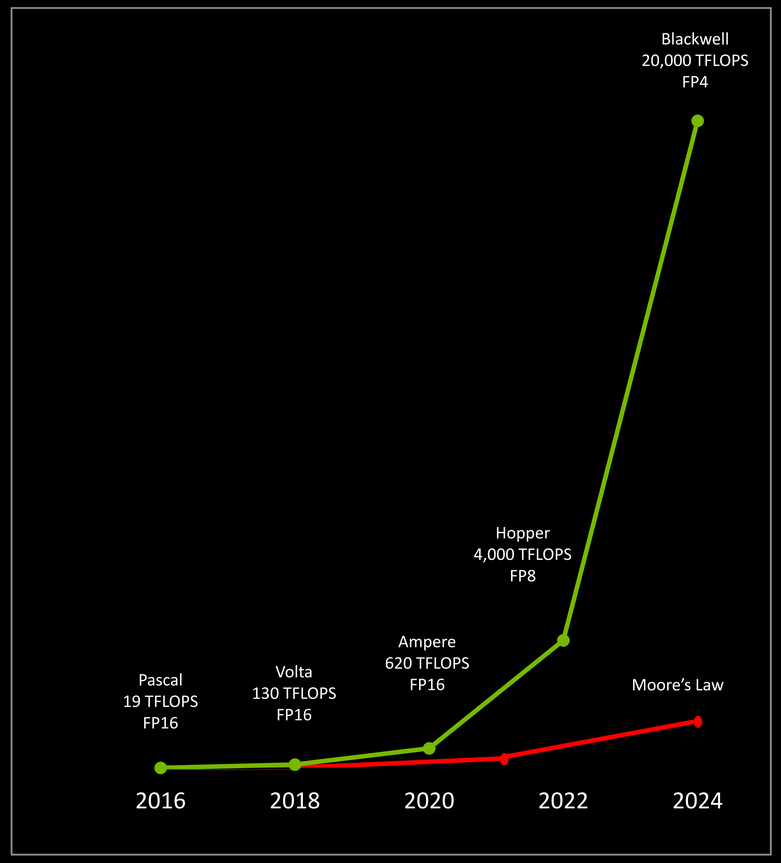
SECURE AI
Full Performance
Encryption & TEE

DECOMPRESSION ENGINE
800 GB/s

NVIDIA

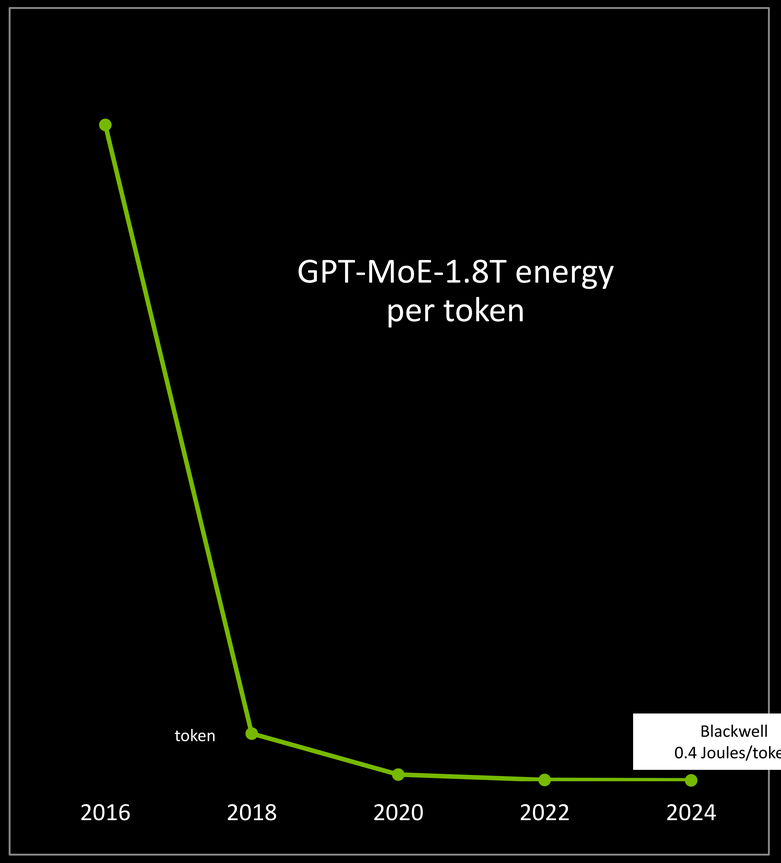
NVIDIA Blackwell Drives Exponential Compute and Efficiency

1,000X Compute and 45,000X Efficiency inference gains over eight years



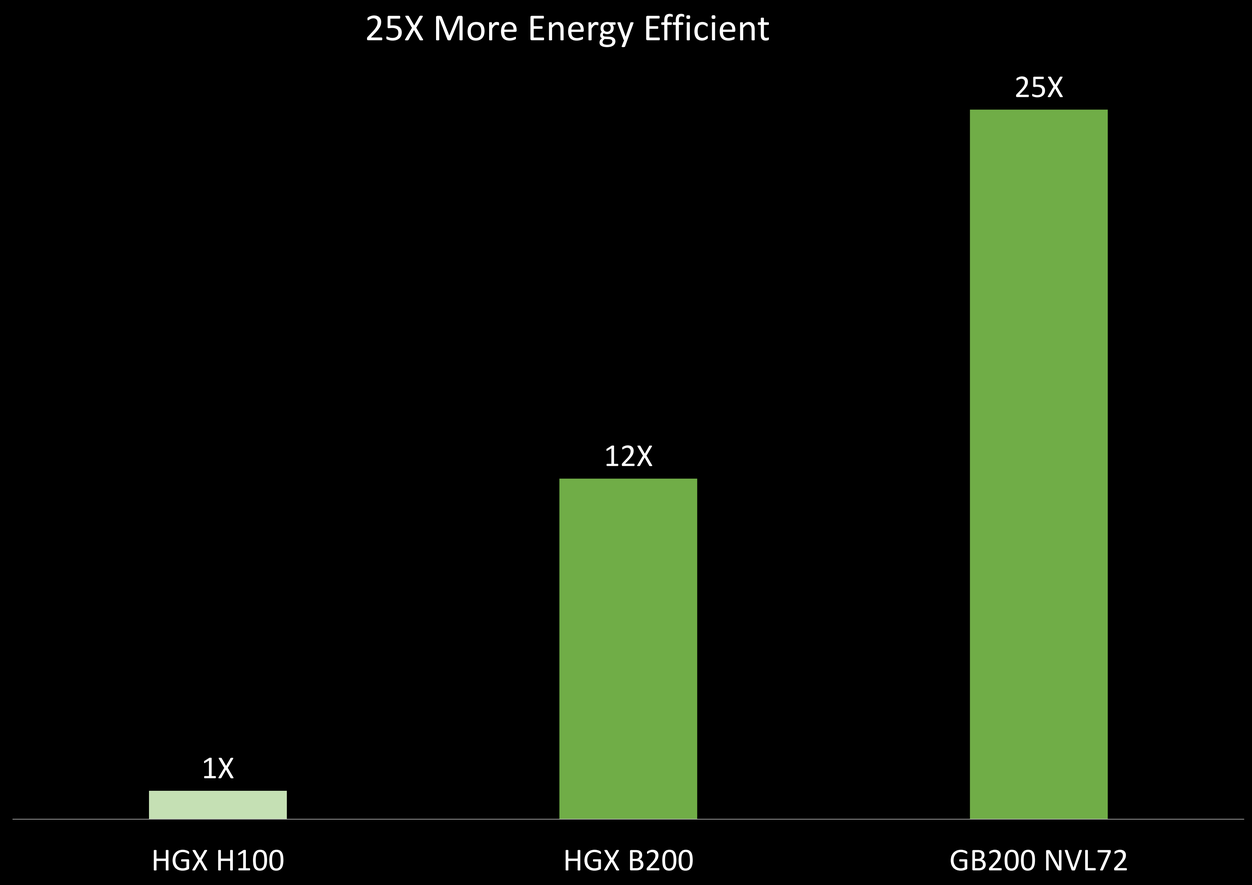
1,000X AI compute in eight years

Democratizing Trillion Parameter AI



Energy/token drops 45,000X in eight years

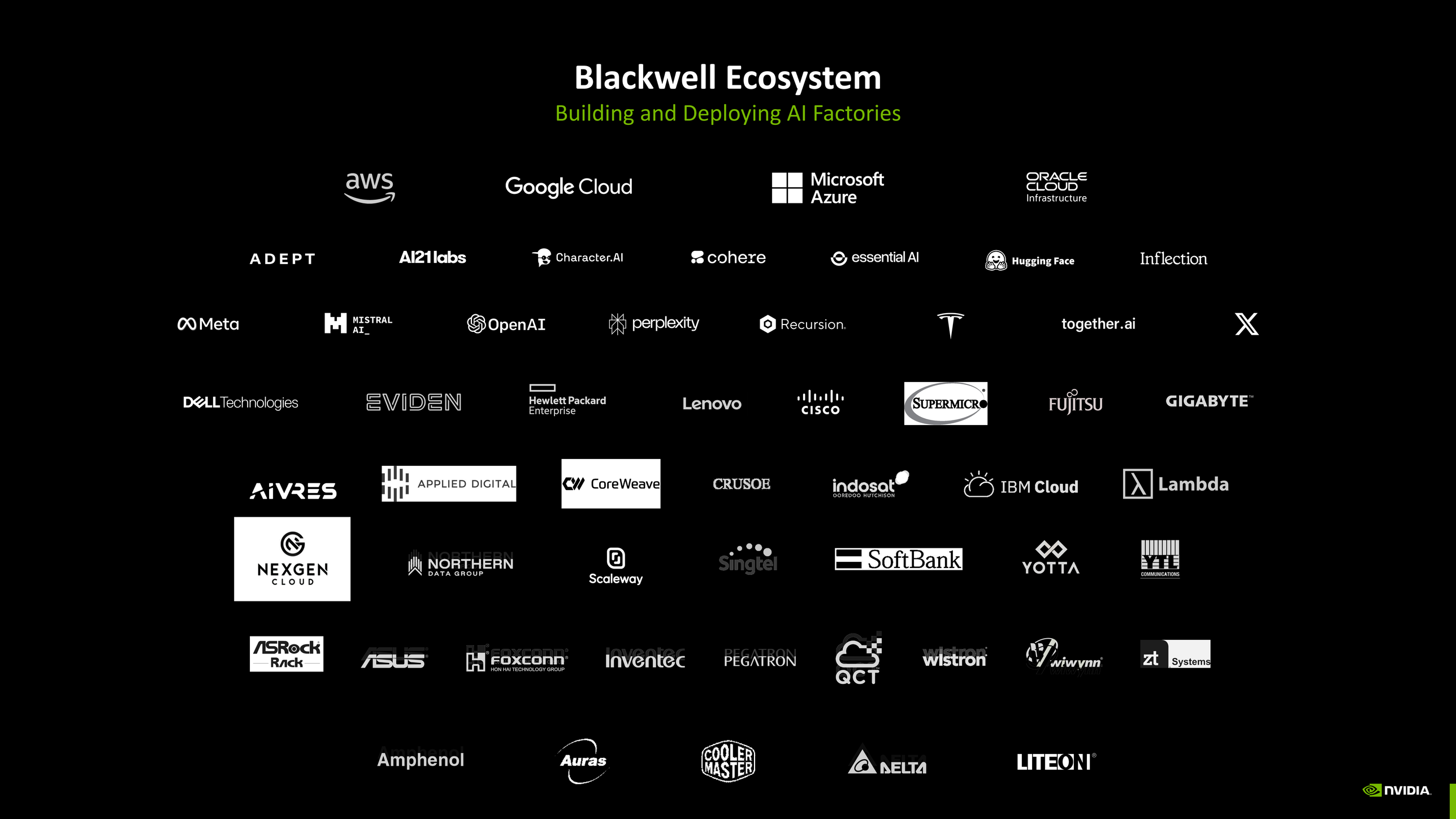
Reduced Energy Use and Lower Cost of Ownership



Projected performance subject to change. Token-to-token latency (TTL) = 50ms real time, first token latency (FTL) = 5s, input sequence length = 32,768, output sequence length = 1,028, 8x eight-way HGX H100 GPUs air-cooled vs. 1x eight-way HGX B200 air-cooled, per GPU performance comparison.. TCO and energy savings for 100 racks eight-way HGX H100 air-cooled versus 8 racks eight-way HGX B200 air-cooled with equivalent performance

Blackwell Ecosystem

Building and Deploying AI Factories



NVIDIA AI Enterprise AI Software Stack

MLOps

AI Applications

NVIDIA AI Enterprise

Infrastructure Management

Cloud Native Management and Orchestration
GPU Operator, Network Operator

Cluster Management
Base Command Manager Essentials

Infra Acceleration Libraries
Magnum IO, vGPU, CUDA

Application Frameworks

LLM
NeMo

Speech AI
Riva

Cybersecurity
Morpheus

Medical Imaging
Clara

More

AI Development

Data Science / Prep
RAPIDS, RAPIDS Accelerator for Apache Spark

Model Training and Customization
NeMo, TAO, PyTorch, TensorFlow

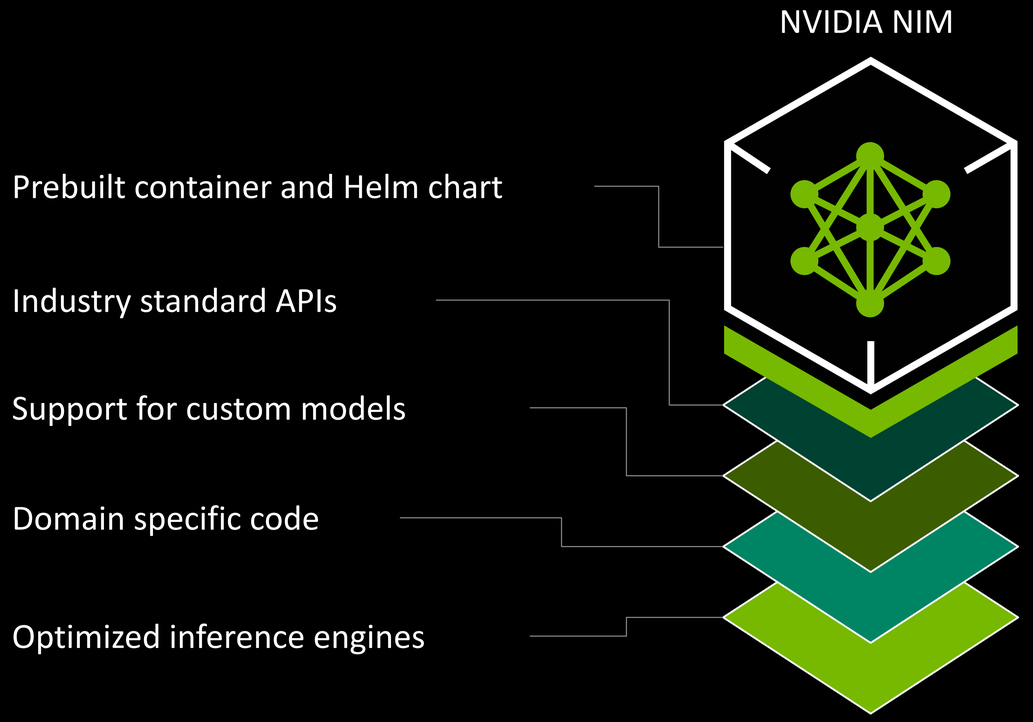
Deploy at Scale
Triton Inference Server

Optimize for Inference
TensorRT, TensorRT-LLM

Cloud | Data Center | Workstations | Edge

NVIDIA NIM: Inference Microservices for Generative AI

Accelerated runtime for generative AI



Deploy anywhere and maintain control of generative AI applications and data

Simplified development of AI application that can run in enterprise environments

Day 0 support for all generative AI models providing choice across the ecosystem

Improved TCO with best latency and throughput running on accelerated infrastructure

Best accuracy for enterprise by enabling tuning with proprietary data sources

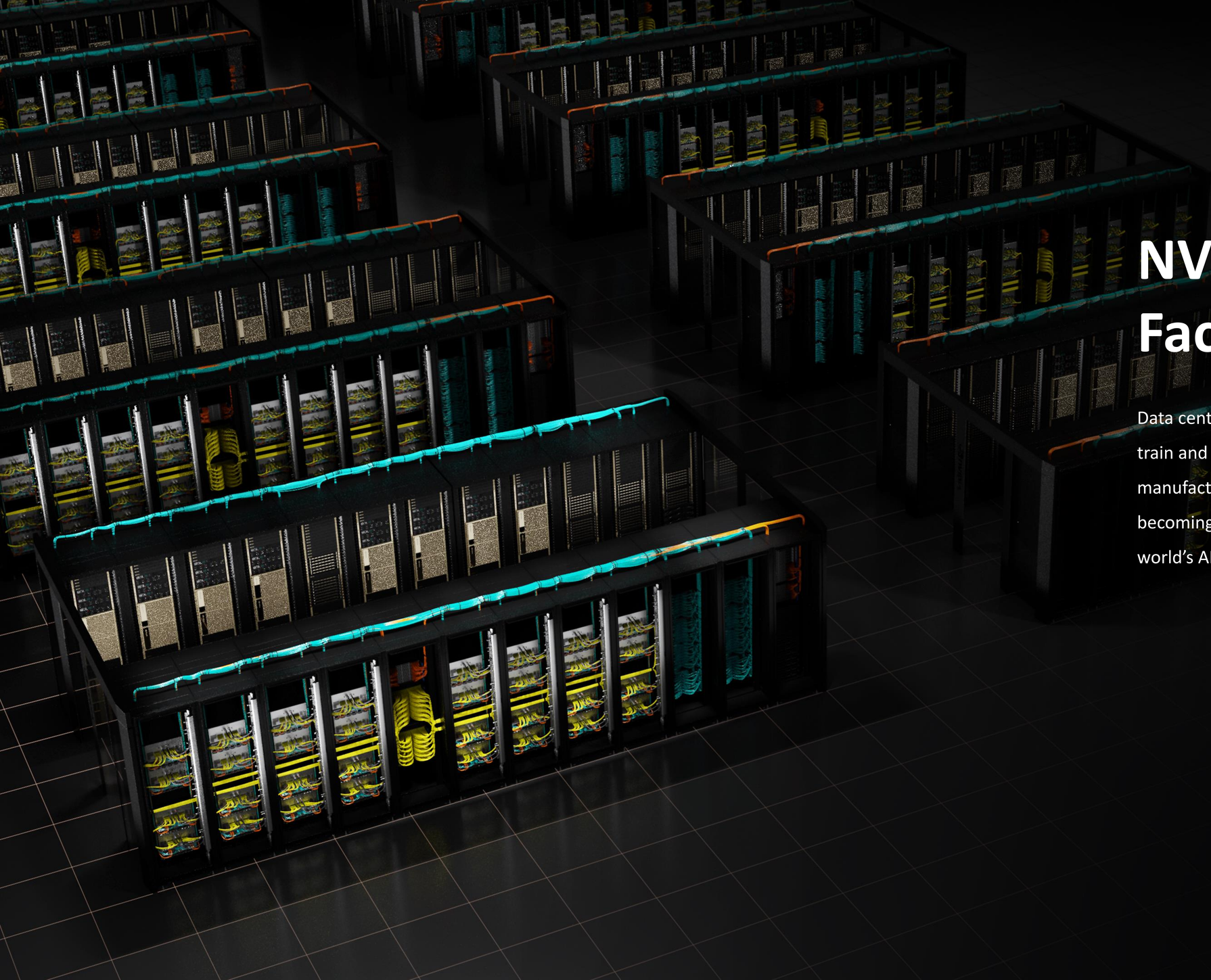
Enterprise software with feature branches, validation and support

Microsoft Azure · aws · Google Cloud · ORACLE · DGX & DGX Cloud · DELL Technologies · Hewlett Packard Enterprise · Lenovo · SUPERMICRO

NVIDIA

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.



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.

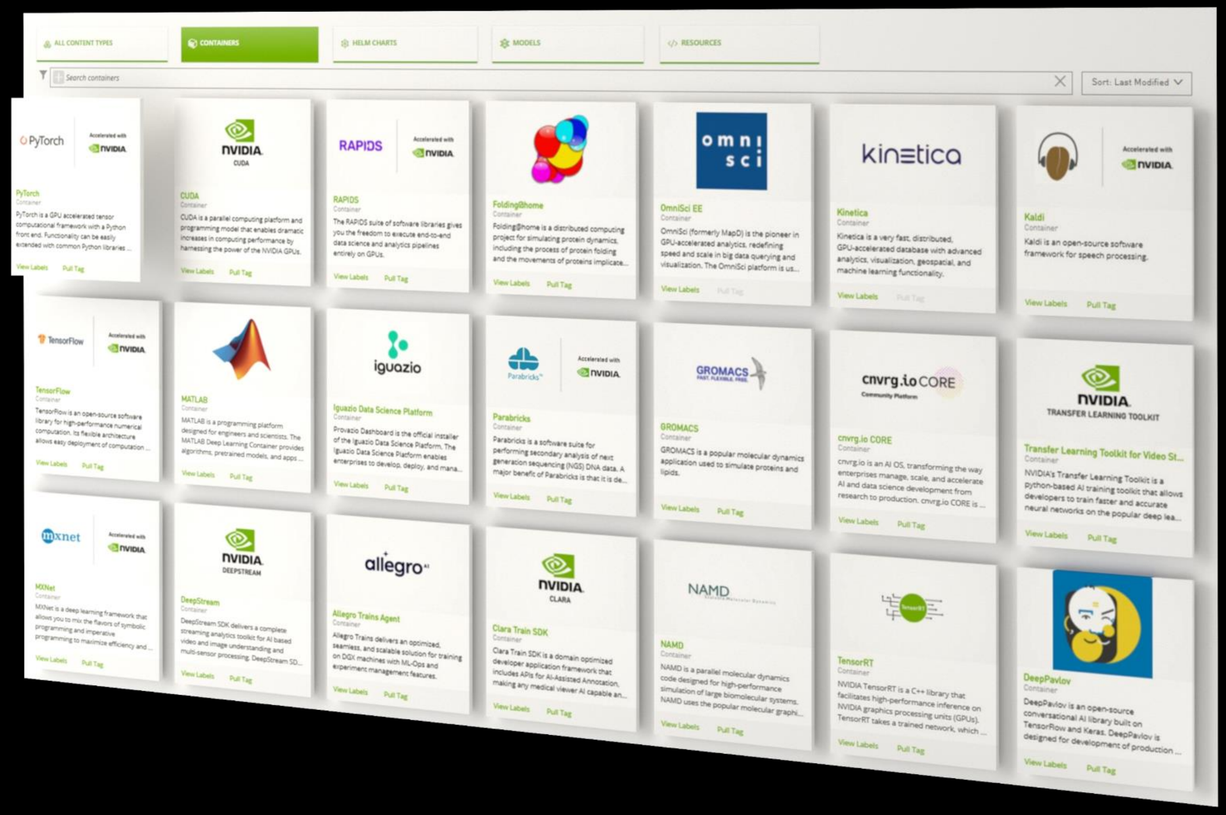
Start Accelerating Your AI Journey with NGC Resources

GPU-optimized AI, Machine Learning, & HPC Software | NVIDIA NGC

Resources

- Try NVIDIA AI Enterprise with a 90-day evaluation license
- Technical blogs
- Webinars
- YouTube Playlist

- Learn more at www.nvidia.com/ngc




The #1 AI Conference | NVIDIA GTC 2025
17th – 20th March
www.nvidia.com/gtc

The Journey Ahead