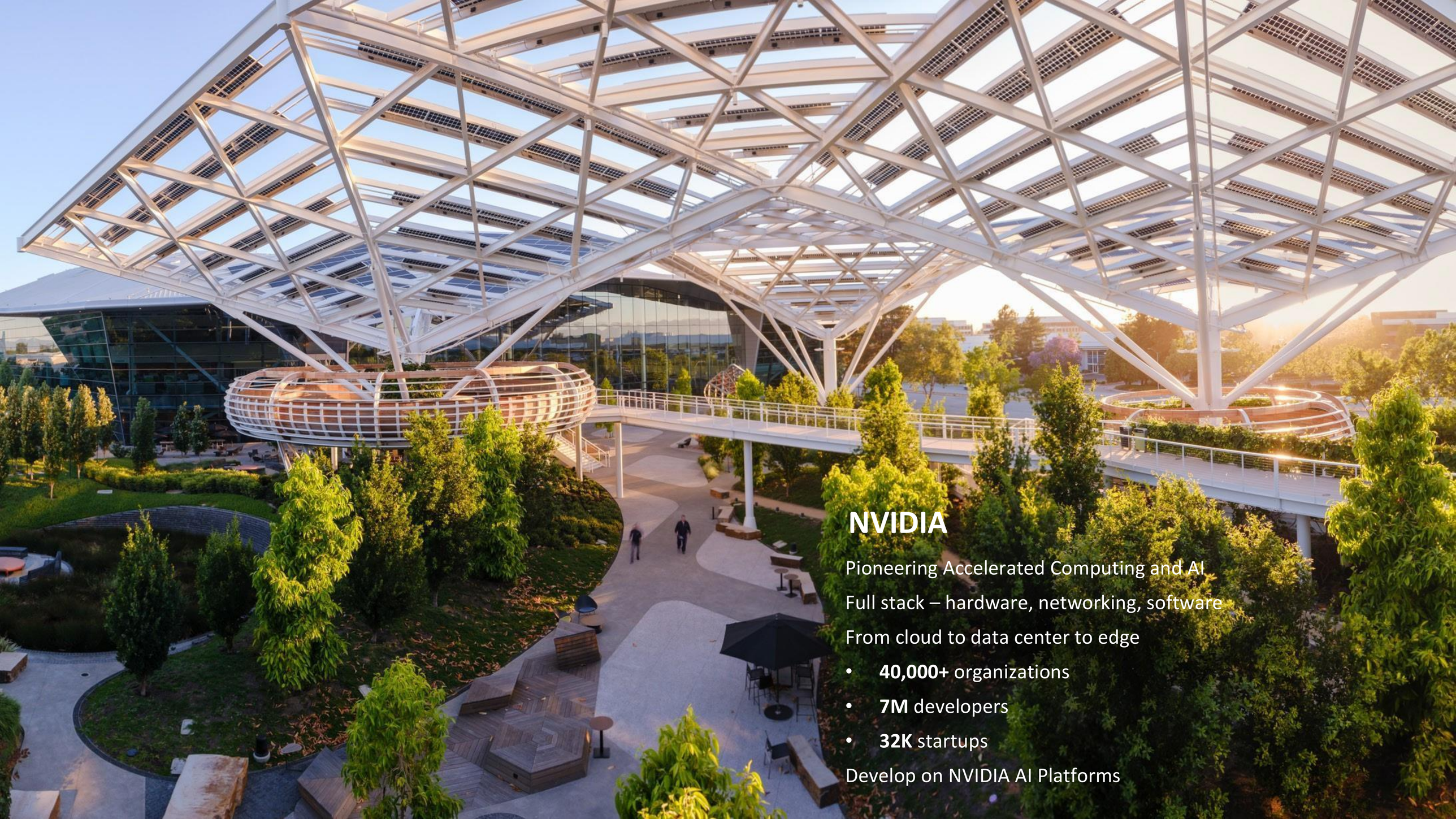
The Full-Stack Advantage: Why AI Wins Are Built Layer by Layer

Tarini Bhatnagar, Sr. Solutions Architect

October 2026



NVIDIA

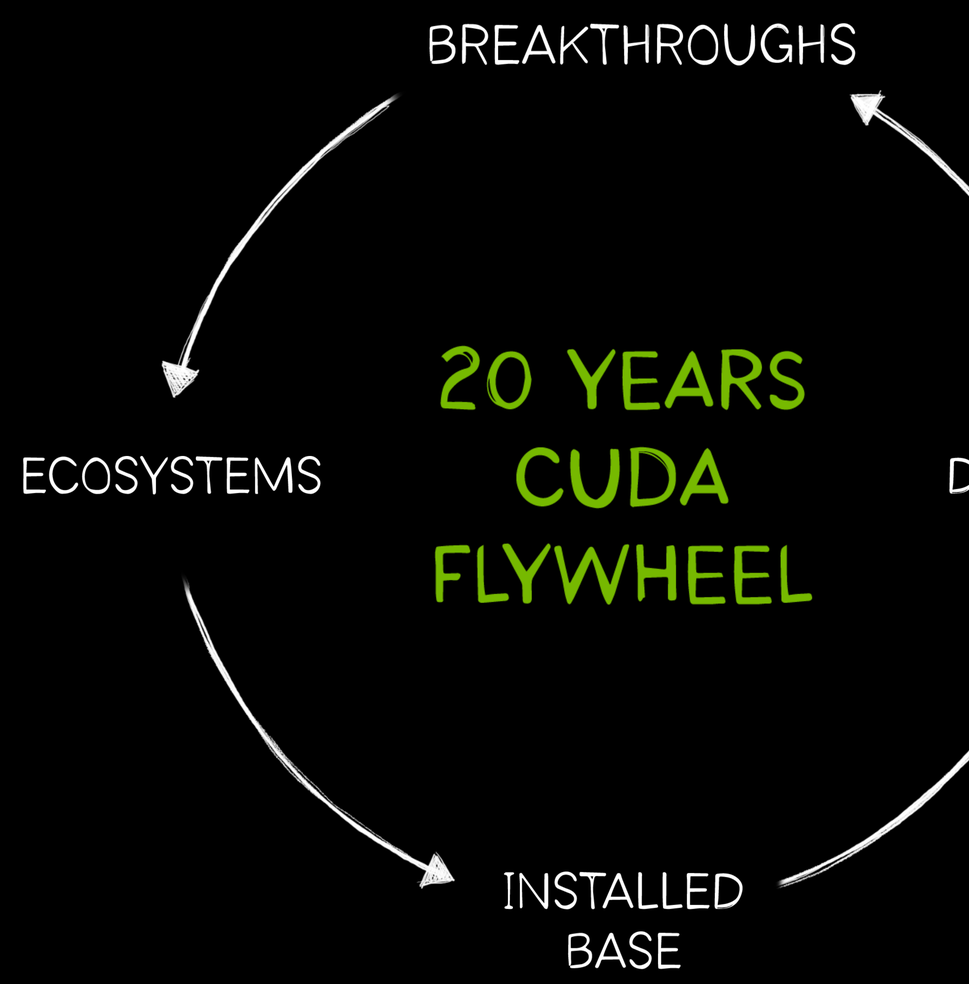
Pioneering Accelerated Computing and AI

Full stack – hardware, networking, software

From cloud to data center to edge

- **40,000+** organizations
- **7M** developers
- **32K** startups

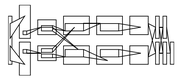
Develop on NVIDIA AI Platforms


BREAKTHROUGHS
ECOSYSTEMS
20 YEARS
CUDA
FLYWHEEL
DEVELOPERS
INSTALLED
BASE

The Evolution of AI

Enterprises require solutions that can keep up with fast-paced innovations

1950
Alan Turing's publication



2012
AlexNet

Perception AI
Speech Recognition
Recommender Sys
Medical Imaging



2017
Transformer Architecture

2022
ChatGPT

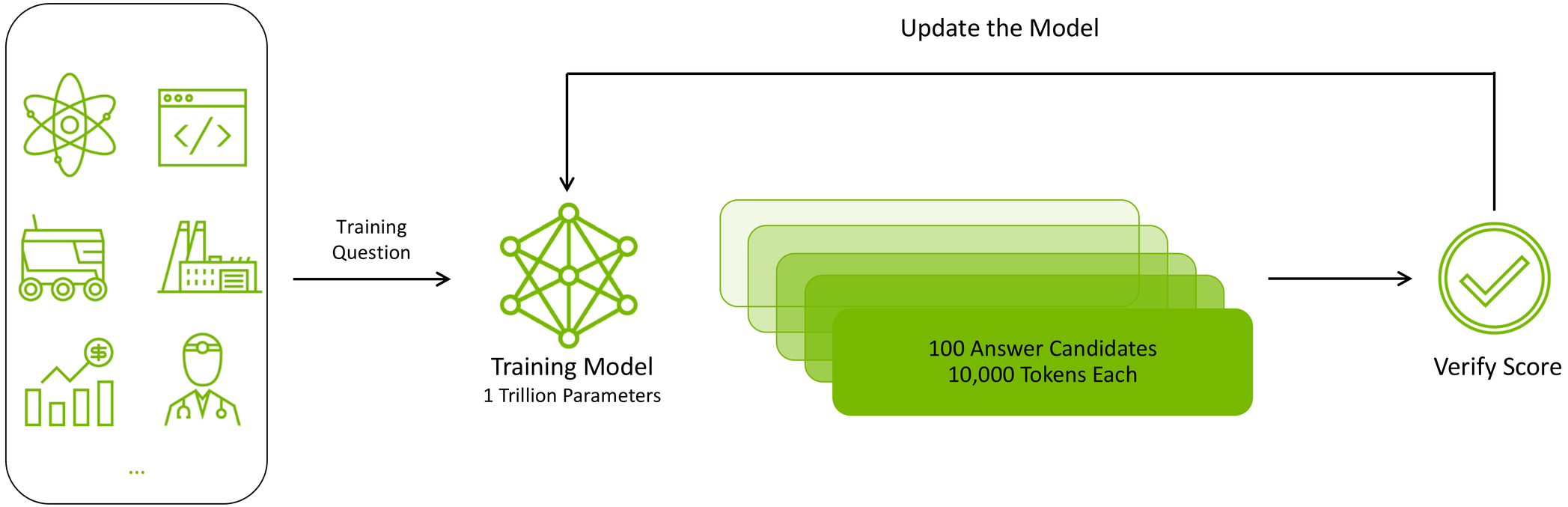
Generative AI
Digital Marketing Content Creation

Agentic AI
Coding Assistant
Customer Service
Patient Care

Physical AI
Self-driving Cars General Robotics

The Era of Reasoning

Teaching models to think requires a lot of compute



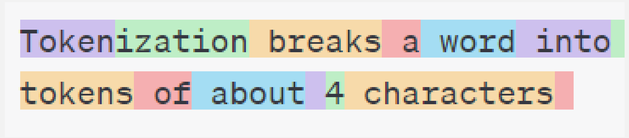
The quality of compute determines the quality of thinking → outcomes

Understanding Tokens

The Language and Currency of AI

Tokens

- Tiny units of data that come from breaking down bigger chunks of information
  - Text: ± 4 chars in length
  - Images: Group of pixels
- AI models learn by guessing the next token in the sentence (next-token prediction).
- After training, the model can perform inference to generate tokens and produce responses.



Value of Tokens

- The cost of AI inference is often measured in terms of the cost per token
- The faster tokens can be processed, the faster models can learn and respond
- Compute resources required to process and generate tokens can be significant
- Many AI services measure the value of their products based on the number of tokens consumed and generated

Explaining Tokens — the Language and Currency of AI

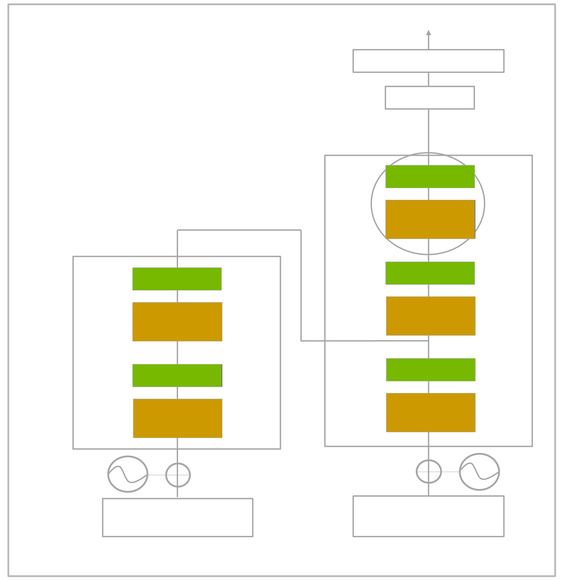
LLM = Large Language Model

Predicts the next word

A few thousand previous words for context

Predict the next word

Transformer Architecture
Deep Neural Network



Through hard work, he supported himself and his •••

Because it crossed state lines, that criminal behavior attracted the attention of the •••

My favorite professor was educated, intelligent, and •••

“family”

“*FBI*”

“understanding”

Models:
GPT-OSS-120B
LLAMA-3.1-8B
QWEN3-VL

NVIDIA.

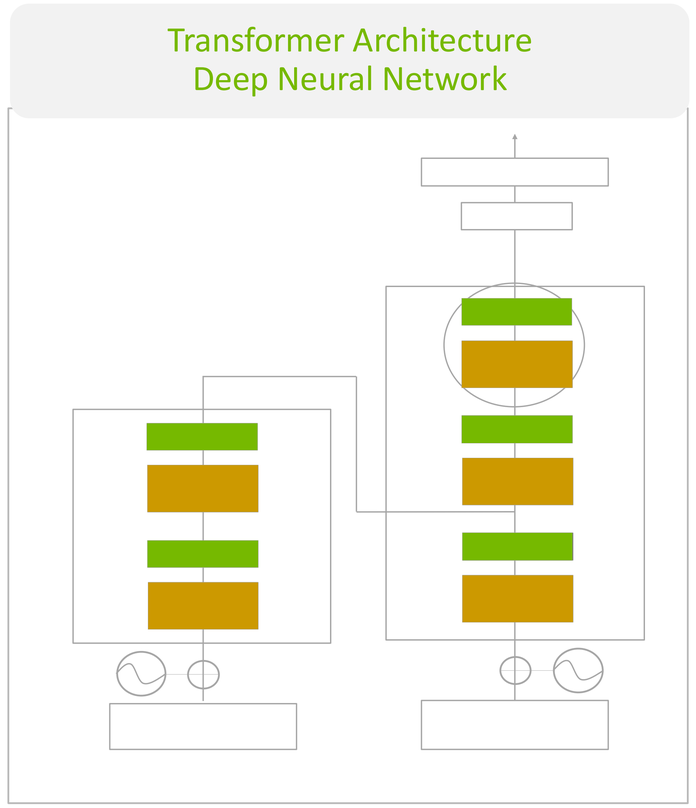
What is a token?

Unit of output for LLMs – not always a word

A few thousand previous words for context

Predict the next **token**

Transformer Architecture
Deep Neural Network



Through hard work, he supported himself and his •••

Because it crossed state lines, that criminal behavior attracted the attention of the •••

My favorite professor was educated, intelligent, and •••

“family”

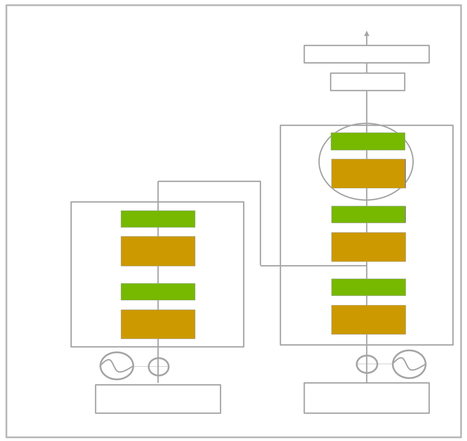
“*FBI*”

“under”, “stand”, “ing”

How do we measure LLM performance?

Input Sequence Length (ISL)

Through hard work, he supported himself and his •••



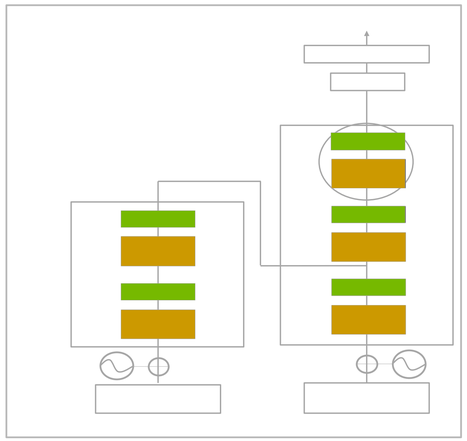
Concurrency

Output Sequence Length (OSL)

“family”

How do we measure LLM performance?

Through hard work, he supported himself and his •••

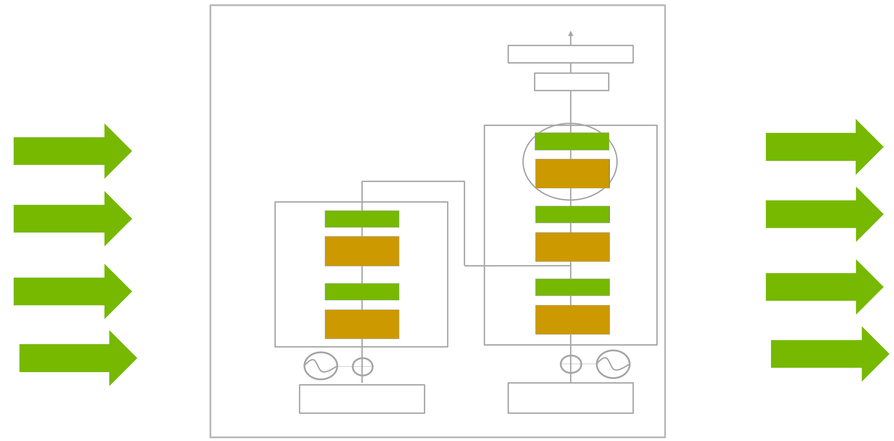


“family”

TTFT = time to first token (latency)

How do we measure LLM performance?

[many concurrent prompts]



[many concurrent responses]

TPS = tokens per second (throughput)

How do we measure LLM TCO?

1. API Costs

Total Cost = $ / token

*usually expressed per million tokens
*input and output tokens usually have different prices

2. Self-Hosted Costs

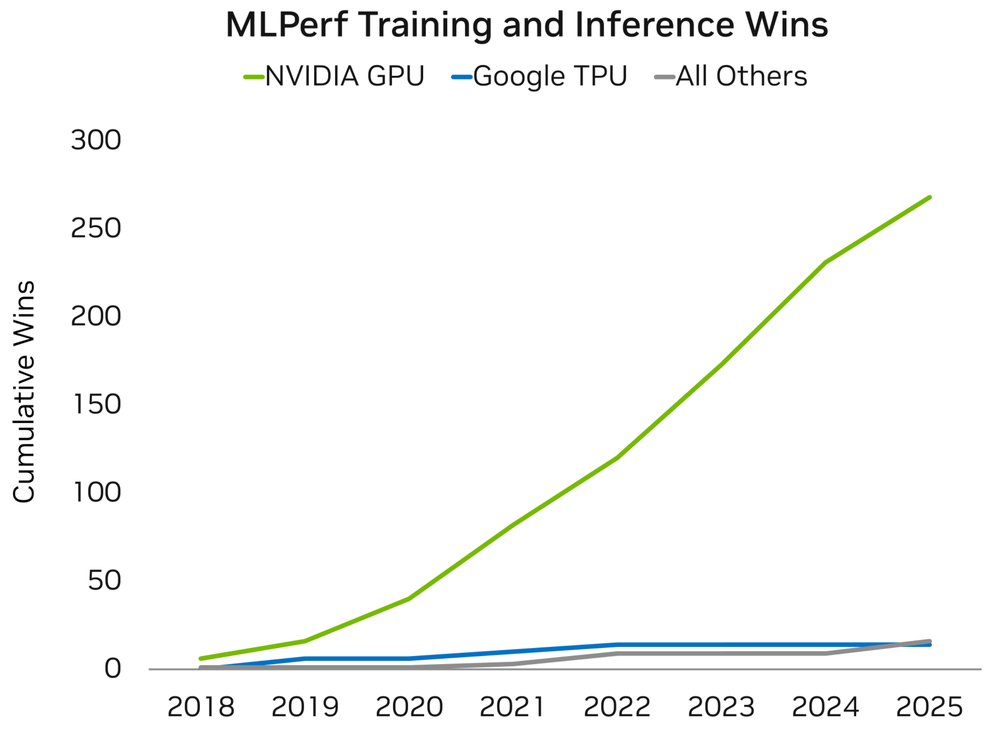
Total Cost = $ / token = $ GPU Infrastructure cost / TPS

*different GPUs have different TPS
*infrastructure cost should include operating costs (power, cooling, DC space, etc…)

NVIDIA.

Sustained Leadership Across Every MLPerf Benchmark

NVIDIA Leads MLPerf With Hundreds of Training and Inference Wins



"NVIDIA Blackwell Ultra Sets Reasoning Records in MLPerf Debut"

– GamesBeat

"NVIDIA's Blackwell Conquers Largest LLM Training Benchmark"

– IEEE Spectrum

"Blackwell GPUs Lift NVIDIA to the Top of MLPerf Training Rankings"

– HPC Wire

"NVIDIA Unveils Rubin CPX Amidst Chart-Topping Blackwell Ultra MLPerf Results"

– HotHardware

MLPerf Training reported time-to-train and MLPerf Inference: Data Center Reported Total Score

Five Layer AI Cake

AI as multi-layer infrastructure driving global job creation and economic growth

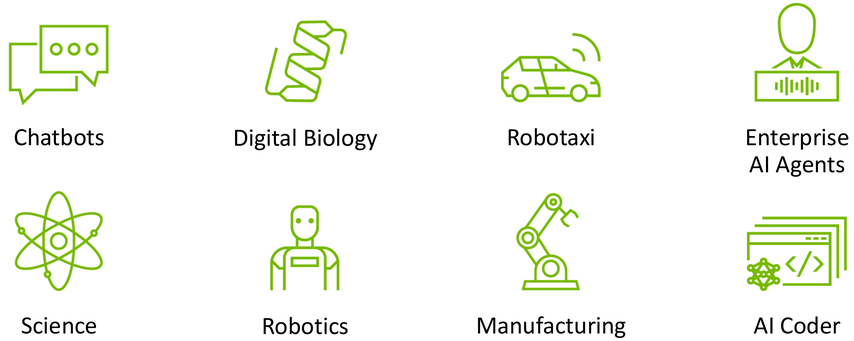
Applications

Models

Infrastructure

Chips

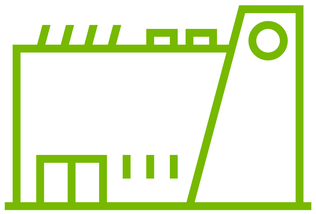
Energy



LLM VLM VLA MMLLM GPT

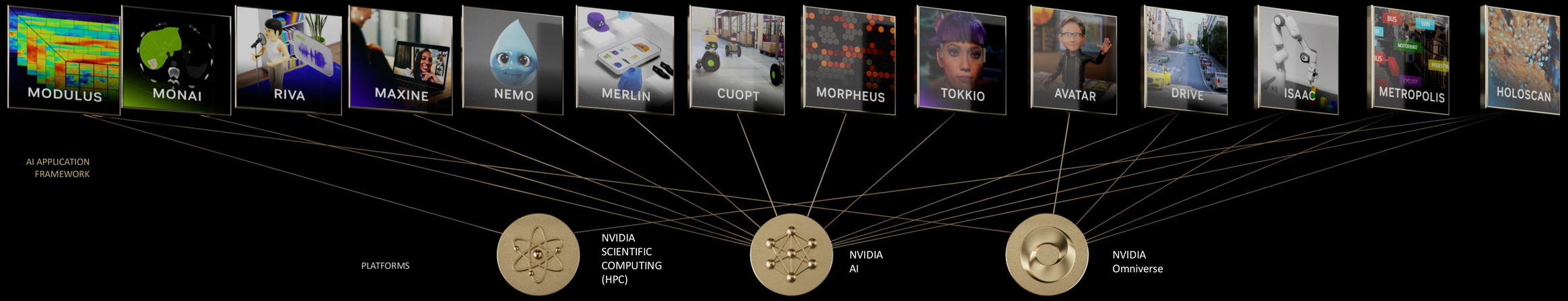
DM GNN MOE SSM LBM

AI Factories



NVIDIA

Beyond Silicon: NVIDIA Software Drives AI at Scale



NVIDIA is the accelerated computing company. Our acceleration libraries and software help developers build the highest-performance AI applications while achieving the lowest compute cost on our GPUs.

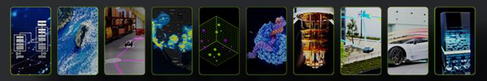
NVIDIA Computing Platform

NIM
CUDA-Accelerated
Agentic AI Libraries



Omniverse
CUDA-Accelerated
Physical AI Libraries





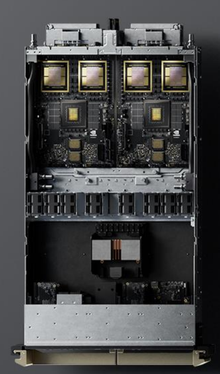
CUDA-X Libraries

CUDA • DOCA • NCCL
Cluster-Scale Software
System Software
Chip Software

Accelerated
Software Stack

GB200 NVL72 SuperPOD

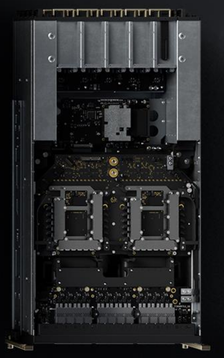




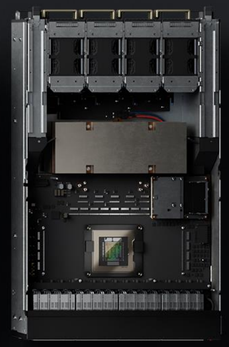
Grace Blackwell
MGX Node



NVLink Switch



Quantum Switch

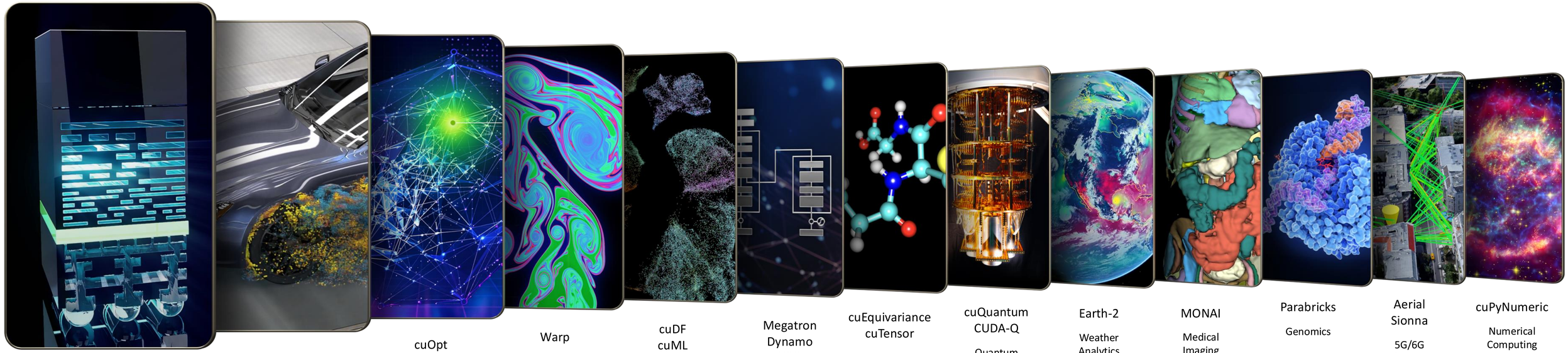


Spectrum-X Switch



Chips Purpose-Built for AI Supercomputing
GPU | CPU | DPU | NIC | NVLink Switch | IB Switch | Ethernet Switch

Accelerated Computing Begins with CUDA Libraries



cuLitho

Computational Lithography

cuDSS
cuSPARSE
cuFFT
AmgX

Computer Aided Engineering

cuOpt

Decision Optimization

Warp

Physics

cuDF
cuML

Data Science and Processing

Megatron
Dynamo
NIXL
cuDNN
CUTLASS

Deep Learning

cuEquivariance
cuTensor

Quantum Chemistry

cuQuantum
CUDA-Q

Quantum Computing

Earth-2

Weather Analytics

MONAI

Medical Imaging

Parabricks

Genomics

Aerial
Sionna

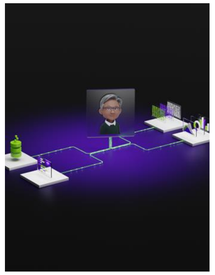
5G/6G
Signal Processing

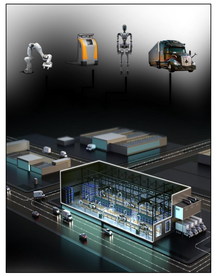
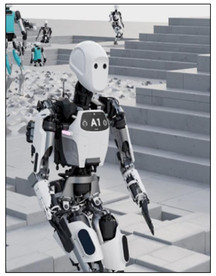
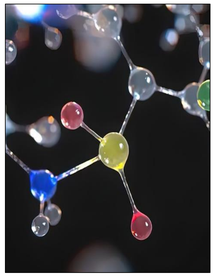
cuPyNumeric

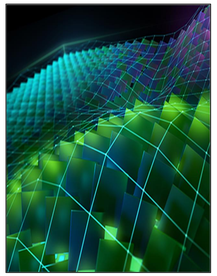
Numerical Computing

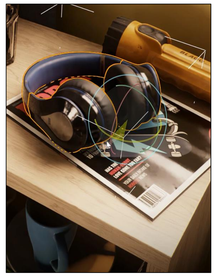
Modern Enterprises Have Diverse Accelerated Workloads

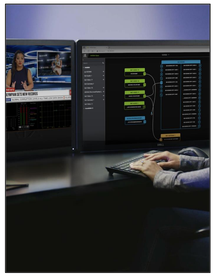
From Enterprise Applications to Visual Computing to HPC to AI









AGENTIC AI | INDUSTRIAL & PHYSICAL AI | SCIENTIFIC COMPUTING, DATA ANALYTICS, & SIMULATION | VISUAL COMPUTING | ENTERPRISE APPLICATIONS



Enterprise Data Center

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

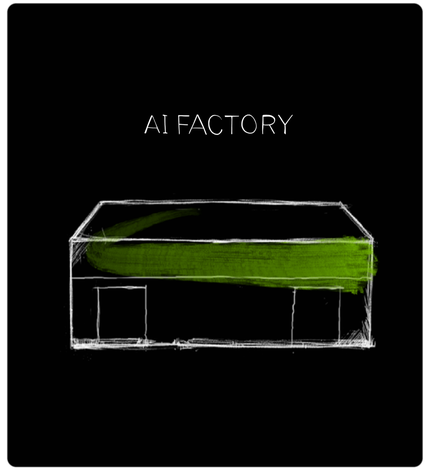
Resource management, time-to-first-train, time-to-inference

A New Industrial Revolution - Tokens are the New Currency

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

AEC | Auto | Consumer Internet | Energy | FSI | Healthcare and Life Sciences | Higher Education and Research | HPC / Supercomputing | Manufacturing and Industrials | Media and Entertainment | Public Sector (US) | Retail and CPG | Robotics | Smart Cities and Spaces | Telco

$100T



NVIDIA

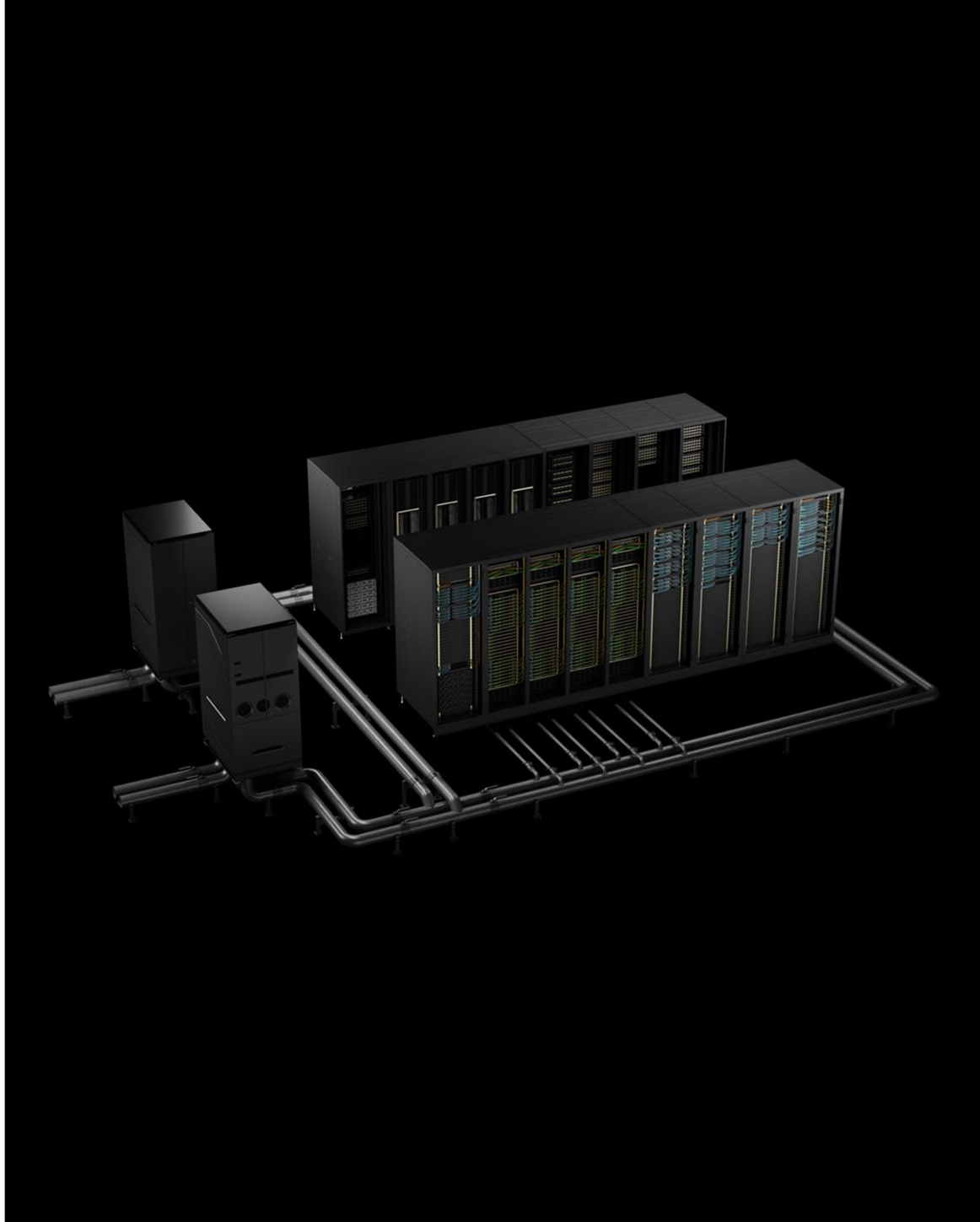
Trillion-Dollar Global IT Investment Shifting to AI Factories

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



“

By 2029, the top 25% of large-cap enterprises that master AI factories will control over 75% of their respective markets….

By 2029, 70% of large enterprises failing to effectively utilize AI factories will cease to exist.

— *Gartner®, Innovation Insight for GenAI Infrastructure – July, 2025*

”

AI Factories are the Industrial Infrastructure of the AI Era
Inference is the Workload
Tokens are the New Commodity
Compute is Revenue



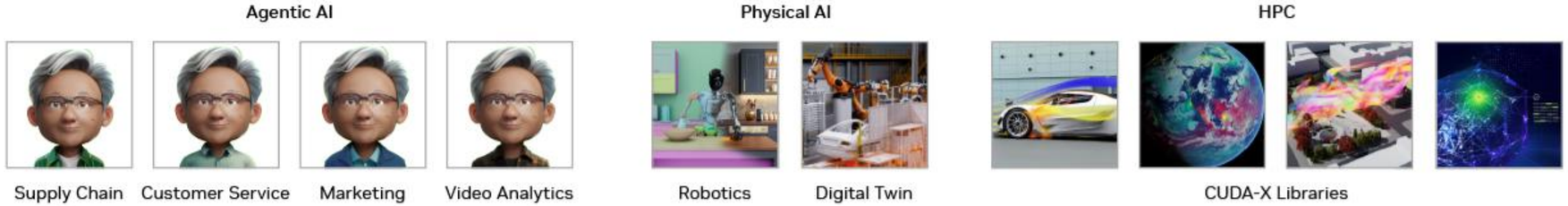
AI factories need to get to first token production as fast as fast and efficiently as possible

NVIDIA CONFIDENTIAL. DO NOT DISTRIBUTE.

NVIDIA Provides a Full Stack for AI Factories

Simplifies AI Factories with Full Stack Solutions for Agentic AI, Physical AI, and HPC

Solutions



AIOps Platform

NVIDIA Enterprise AI Factory Validated and Reference Designs

ISVs

NVIDIA Blueprints

NVIDIA AI Enterprise

NVIDIA NIM | NVIDIA NeMo | NeMo Agent Toolkit

NVIDIA Omniverse

AI Infrastructure Software

NVIDIA Enterprise Reference Architecture

NVIDIA AI Enterprise
GPU, Networking, NIM Operators

NVIDIA Run:ai | CUDA-X

Kubernetes | Operating System

AI Infrastructure

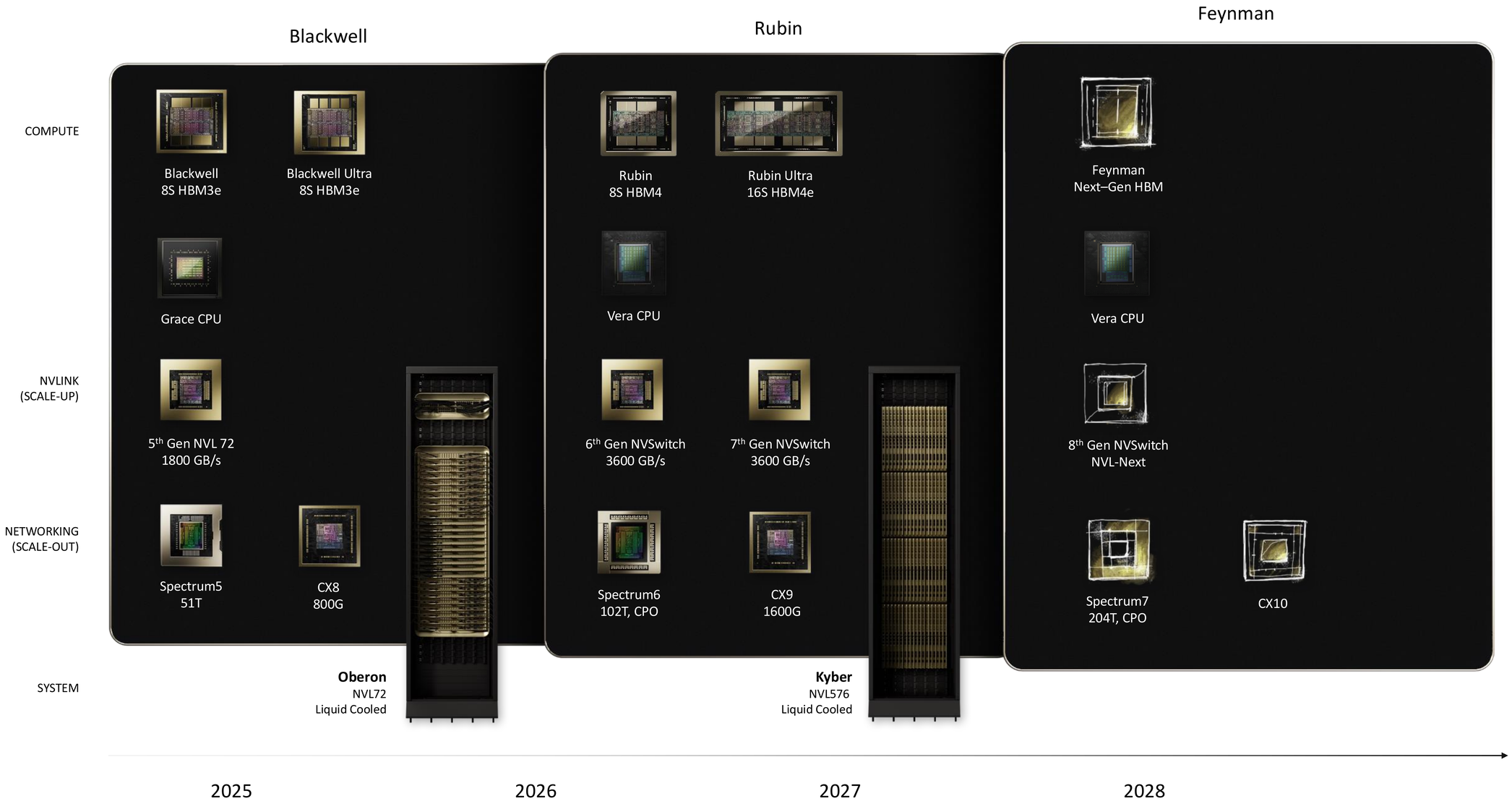
NVIDIA-Certified Compute
CPU | GPU

NVIDIA Networking
NVIDIA Spectrum-X | Bluefield | NVLINK

NVIDIA-Certified Storage

NVIDIA Paves Road to Gigawatt AI Factories

One-Year Rhythm | Full-Stack | One Architecture | CUDA Everywhere

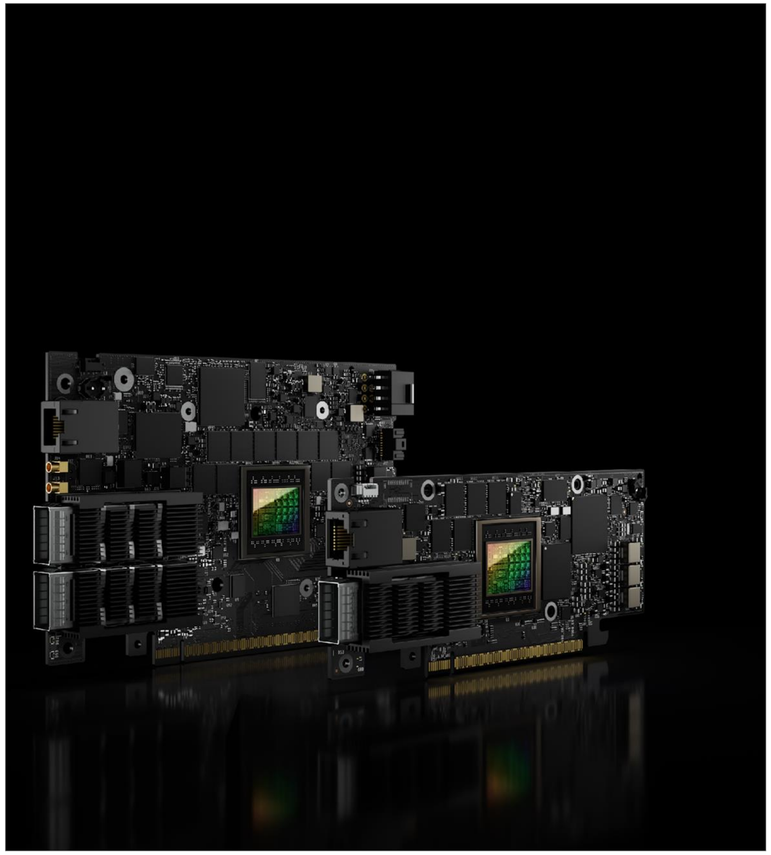


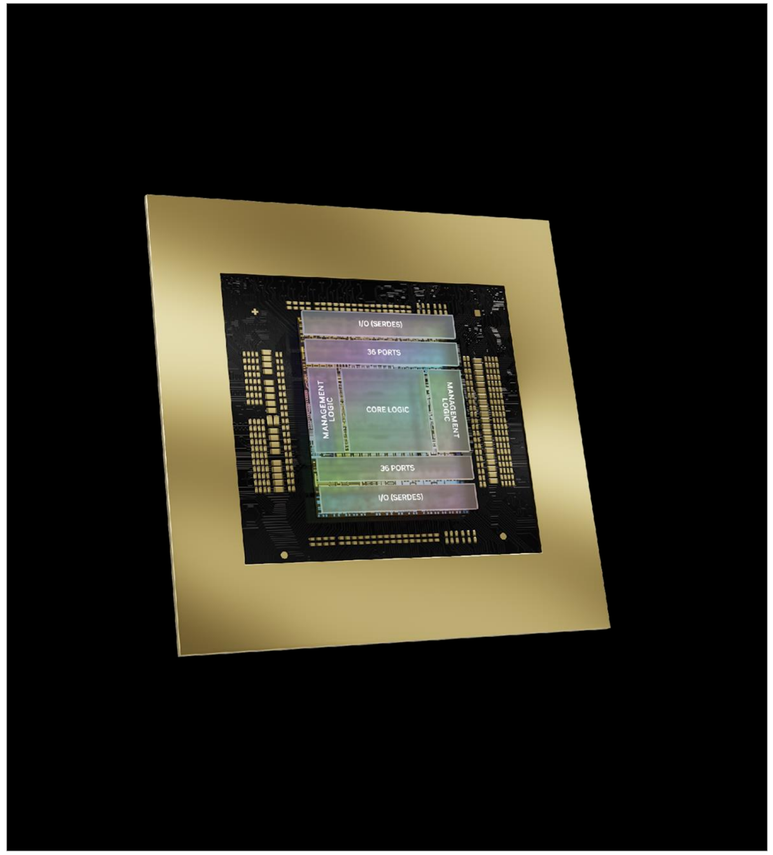
NVIDIA Networking Optimizes AI Factory Communication

Industry best networking solutions to power Enterprise AI

Storage/Access (N-S)

AI Fabric (E-W)







BlueField DPU and SuperNIC

High-performance, low-latency networking with full offload for security, storage, and telemetry

NVLink and NVSwitch

Building blocks for high-speed, multi-GPU scaleup

Spectrum-X Ethernet

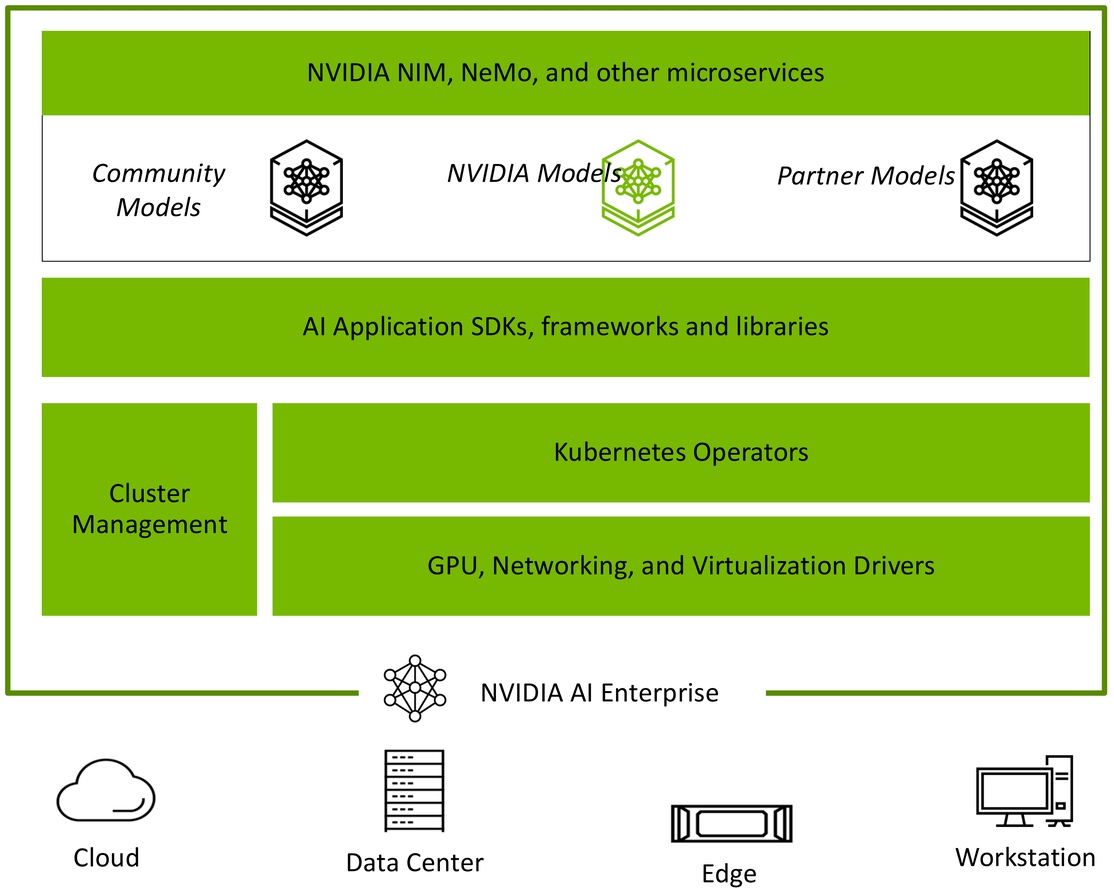
New class of AI optimized Ethernet; largest GPU cluster in the world deployed with Spectrum-X

NVIDIA

NVIDIA Software Activates AI Factory Workloads

Streamline development and deployment of production-grade, end-to-end Generative AI pipelines

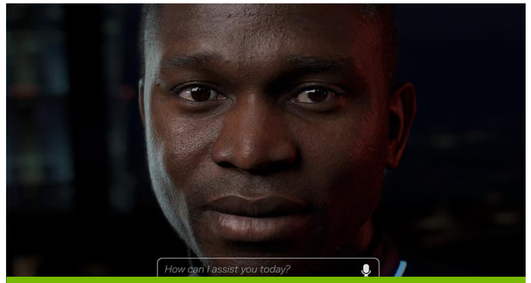
AEC | Auto | Consumer Internet | Energy | FSI | Healthcare and Life Sciences | Higher Education and Research | HPC / Supercomputing | Manufacturing and Industrials | Media and Entertainment | Public Sector (US) | Retail and CPG | Robotics | Smart Cities and Spaces | Telco



NVIDIA Blueprints

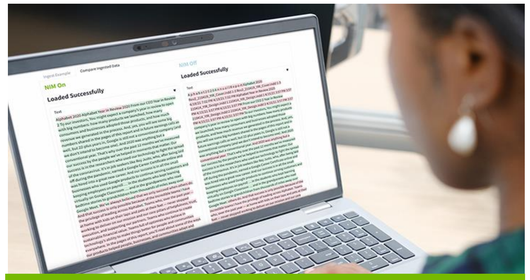
Available on build.nvidia.com

Digital Humans
for Customer Service



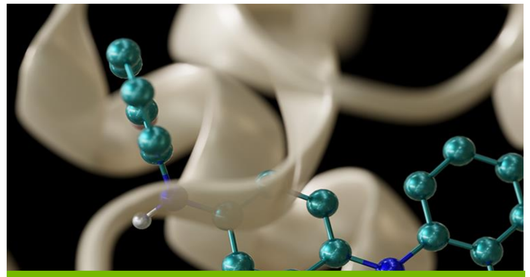
NVIDIA AI Blueprint

Multimodal PDF Data Extraction
for Enterprise RAG



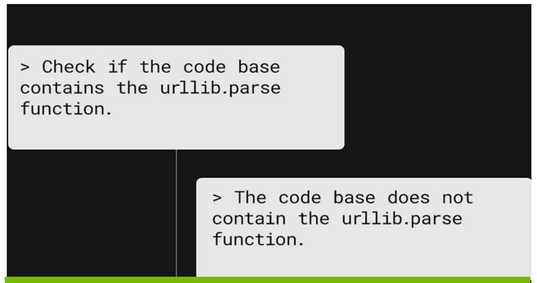
NVIDIA AI Blueprint

Generative Virtual Screening
for Drug Discovery



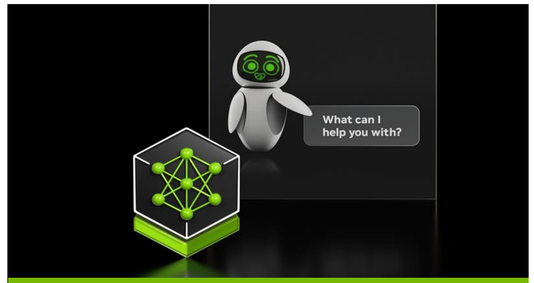
NVIDIA BioNeMo Blueprint

Vulnerability Analysis
for Container Security



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

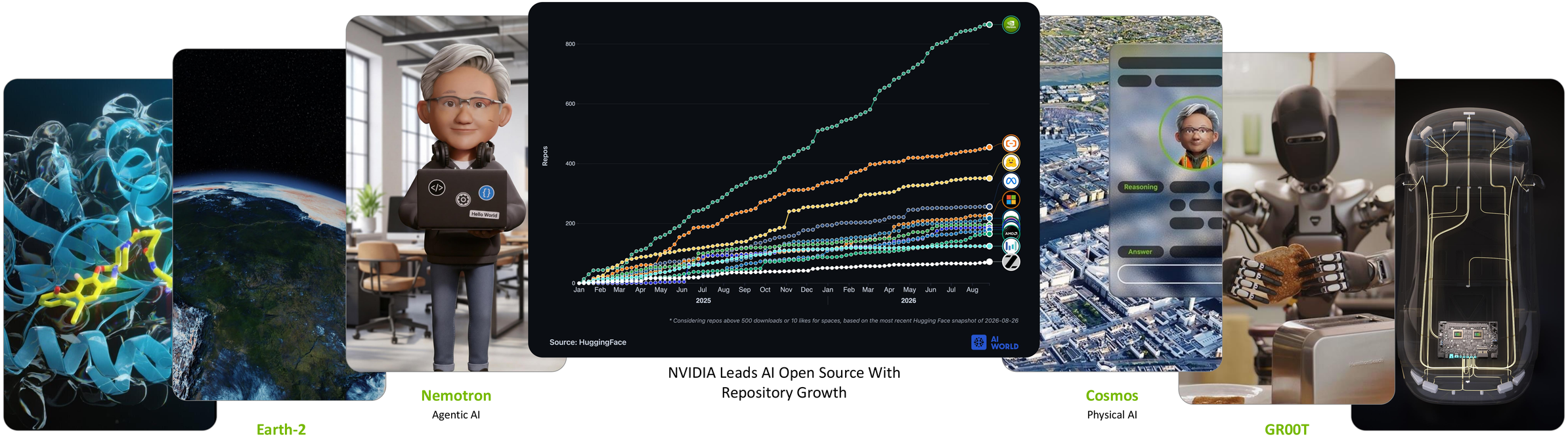
Build a Digital Twin for
Interactive Fluid Simulation



NVIDIA Omniverse Blueprint

NVIDIA

NVIDIA Leads Open-Source AI Contributions



NVIDIA Leads AI Open Source With Repository Growth

Clara
Biomedical AI

Earth-2
AI-Physics

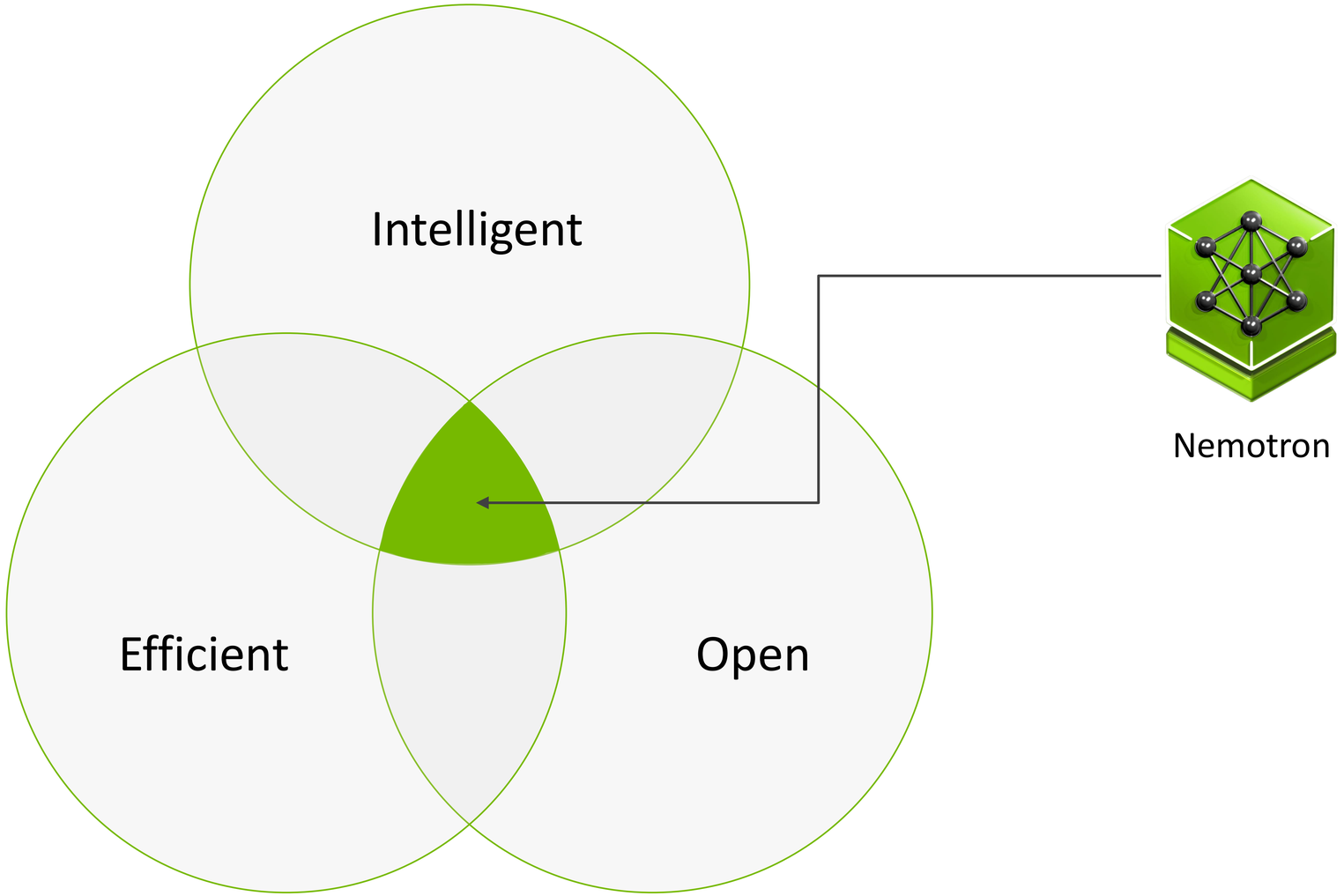
Nemotron
Agentic AI

Cosmos
Physical AI

GR00T
Robotics

Alpamayo
Autonomous Vehicle

Nemotron Designed for Peak Intelligence, Efficiency and Openness



NVIDIA Nemotron 3 Family of Open, Efficient, Intelligent Models

Bolt

74B-A9B

AVAILABLE NOW

Lightning

30B-A3B

AVAILABLE NOW

Super VL

120B-A12B

AVAILABLE NOW

Ultra

550B-A55B

AVAILABLE NOW

Nemotron 3

Open Datasets

Pre-Training Tokens, Post-Training Samples, RL Tasks

NeMo Gym

RL Environments & Skills

Open Research

Papers, Samples

NVIDIA

Driving Energy Efficiency and Cost Savings at Data Center Scale

NVIDIA compute, networking, and software are better together for sustainable computing

Classical Computing

960 CPU-only servers

Application

CPU Server Racks



X 8

Accelerated Computing

2 GPU servers

Application
Re-engineered for acceleration

CUDA-X Acceleration Libraries

Magnum IO



25X Lower cost

84X Better energy efficiency

Requires only 1.2% of the energy used by the CPU cluster

Accelerated Computing Foundation

3 elements

Blackwell GPUs are **more energy efficient** than CPUs for certain AI and HPC workloads up to **20X**

BlueField DPUs reduce server power consumption by up to **30%**

Grace CPUs deliver **better energy efficiency** than x86 CPUs up to **2X**

LLM Workload: Bert-Large Training and Inference | CPU Server: Dual-EYPC 7763 | GPU Server: Dual-EPYC 7763 + 8X H100 PCIe GPUs


NVIDIA