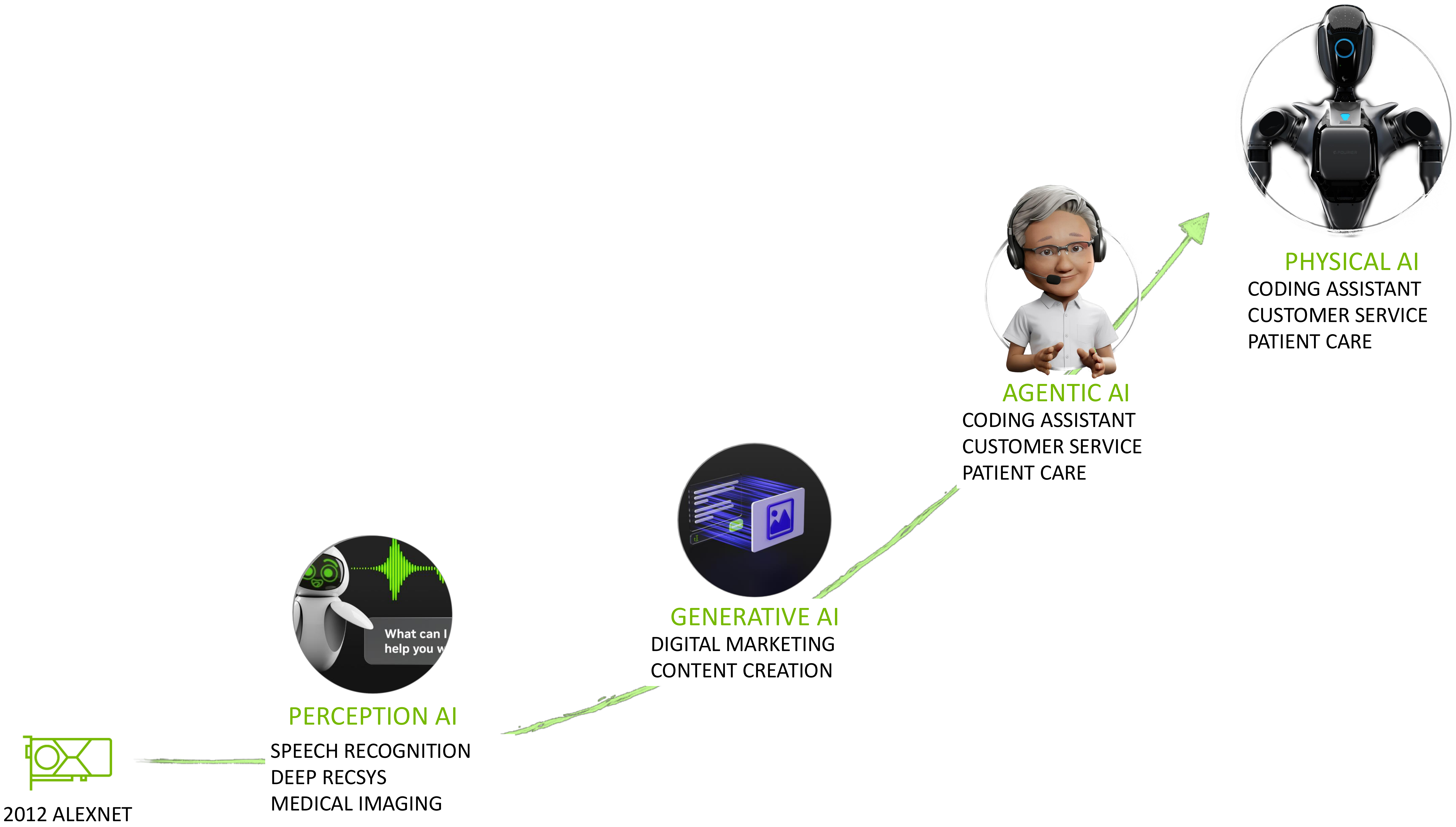
# Next Era of AI - AI Factories

Frank Lin, NVIDIA Senior Solutions Architect

# Next Wave of AI



2012 ALEXNET

PERCEPTION AI

SPEECH RECOGNITION
DEEP RECSYS
MEDICAL IMAGING

GENERATIVE AI

DIGITAL MARKETING
CONTENT CREATION

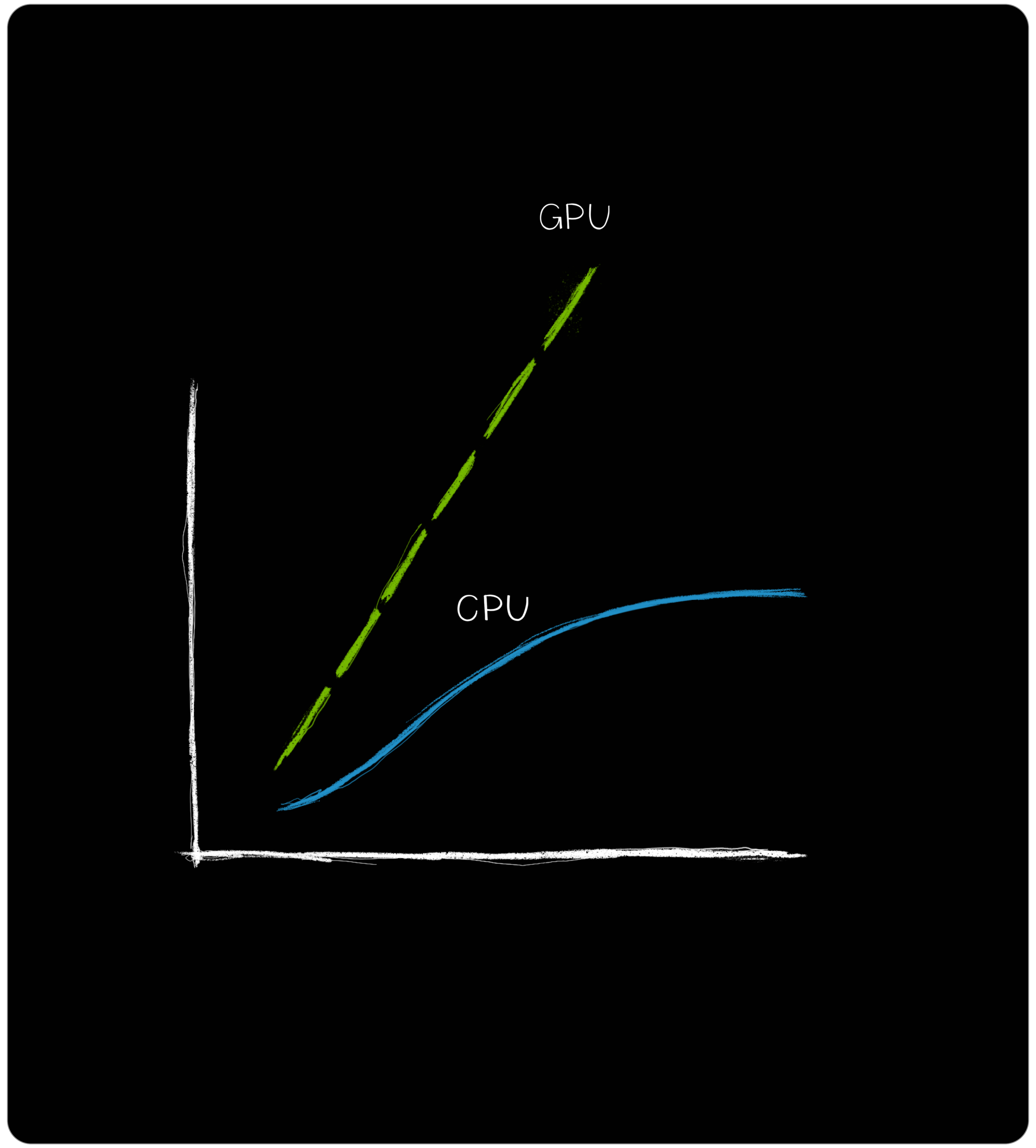
AGENTIC AI

CODING ASSISTANT
CUSTOMER SERVICE
PATIENT CARE

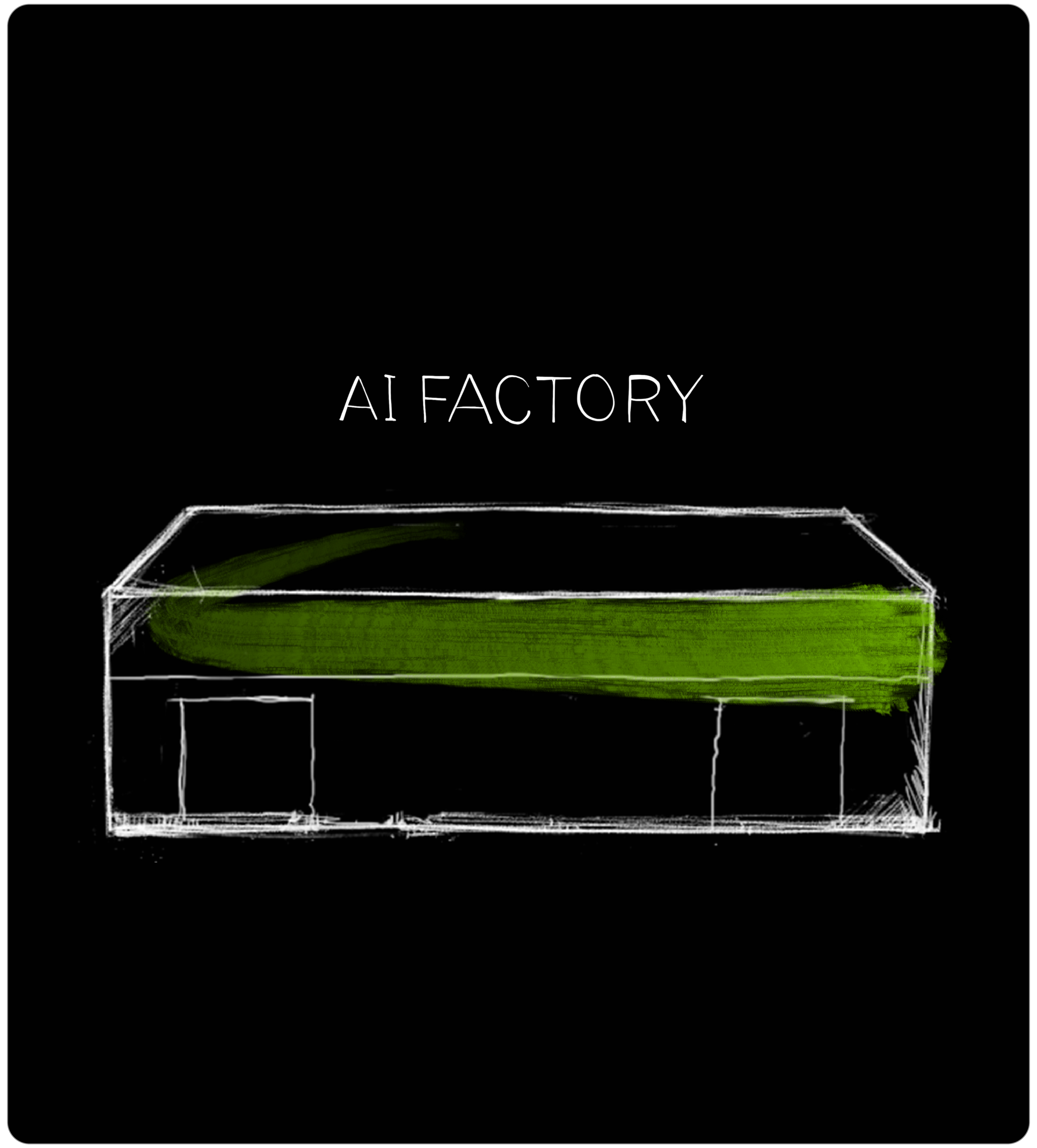
PHYSICAL AI

CODING ASSISTANT
CUSTOMER SERVICE
PATIENT CARE

# Three Major Trends in Computing



Traditional → Accelerated



Data Centers → AI Factories

Generative AI → Physical AI

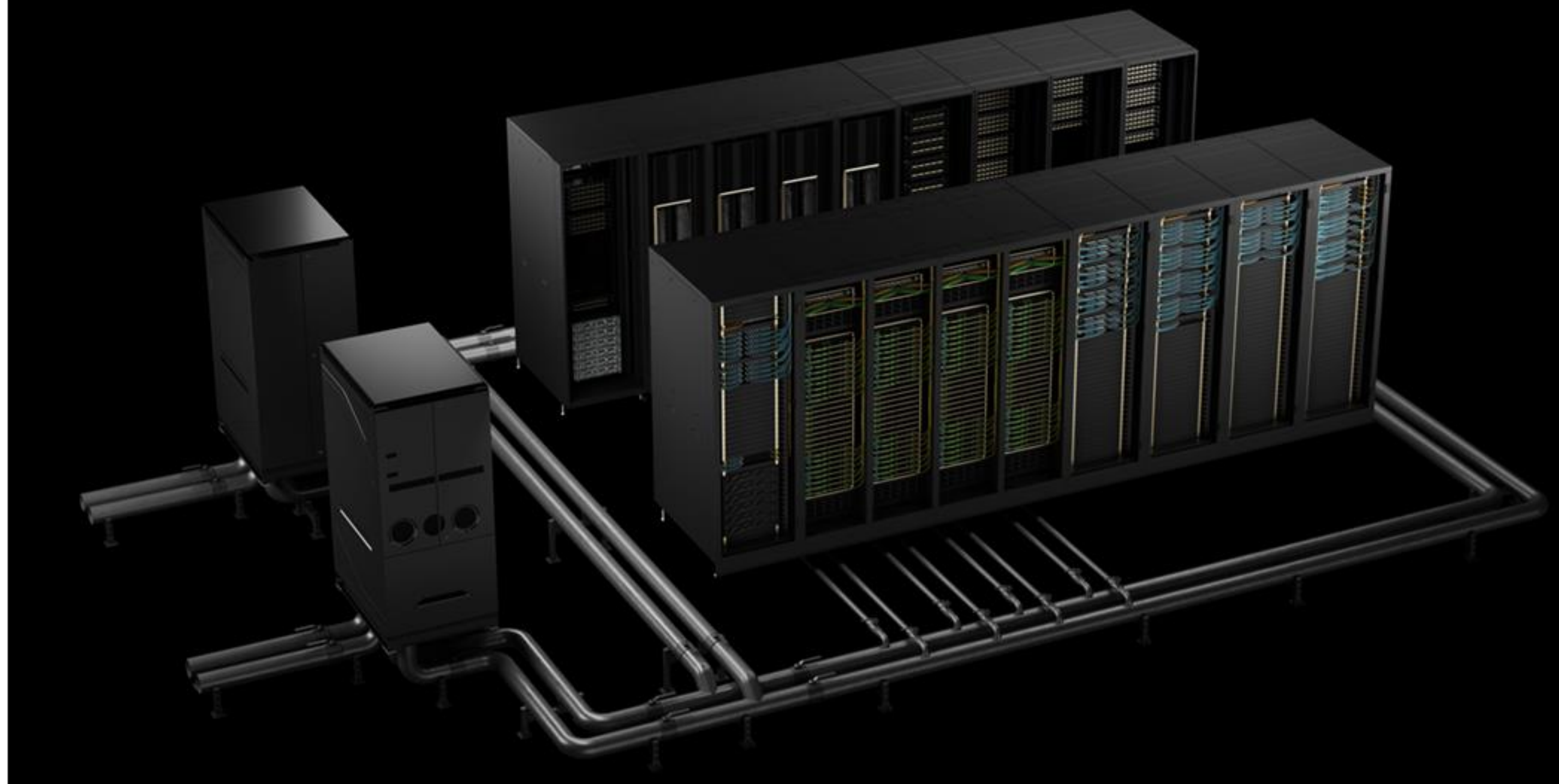
# Trillion-Dollar Global IT Investment Shifting to AI Factories

**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



# Agents Work Together to Solve Complex Problems

Employee Support Representative
Human Resources
Helpdesk Technician
Information Technology
Business Development Representative
Sales
Software Developer
Software Engineering
Hello World
Supply Chain Management
Operations
Analyst
Finance
Customer Service Manager
Operations
Content Creator
Marketing

# NVIDIA Provides the Building Blocks for Agentic AI

AI Blueprints

Research Assistant Agent

Customer Service Agent

Software Security Agent

Virtual Lab Agent

Video Analytics Agent

NVIDIA NeMo

Curator

Customizer

Evaluator

Guardrails

Retriever

NVIDIA NIM

Understanding & Reasoning

Information Retrieval

AI Safety

Digital Humans

Visual Content Generation

Digital Biology

Physical AI

**AI Infrastructure**

GPU

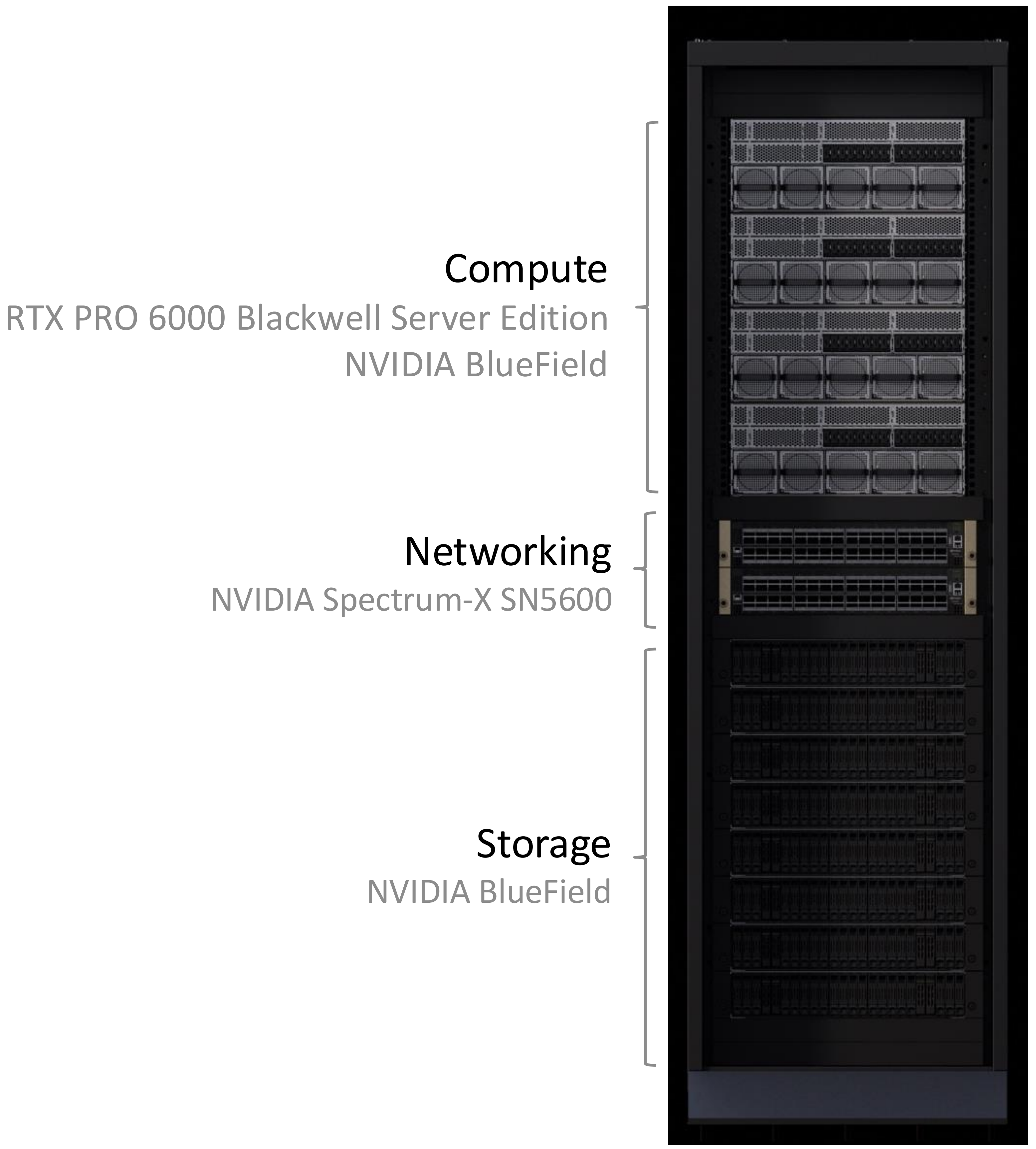
CPU

DPU

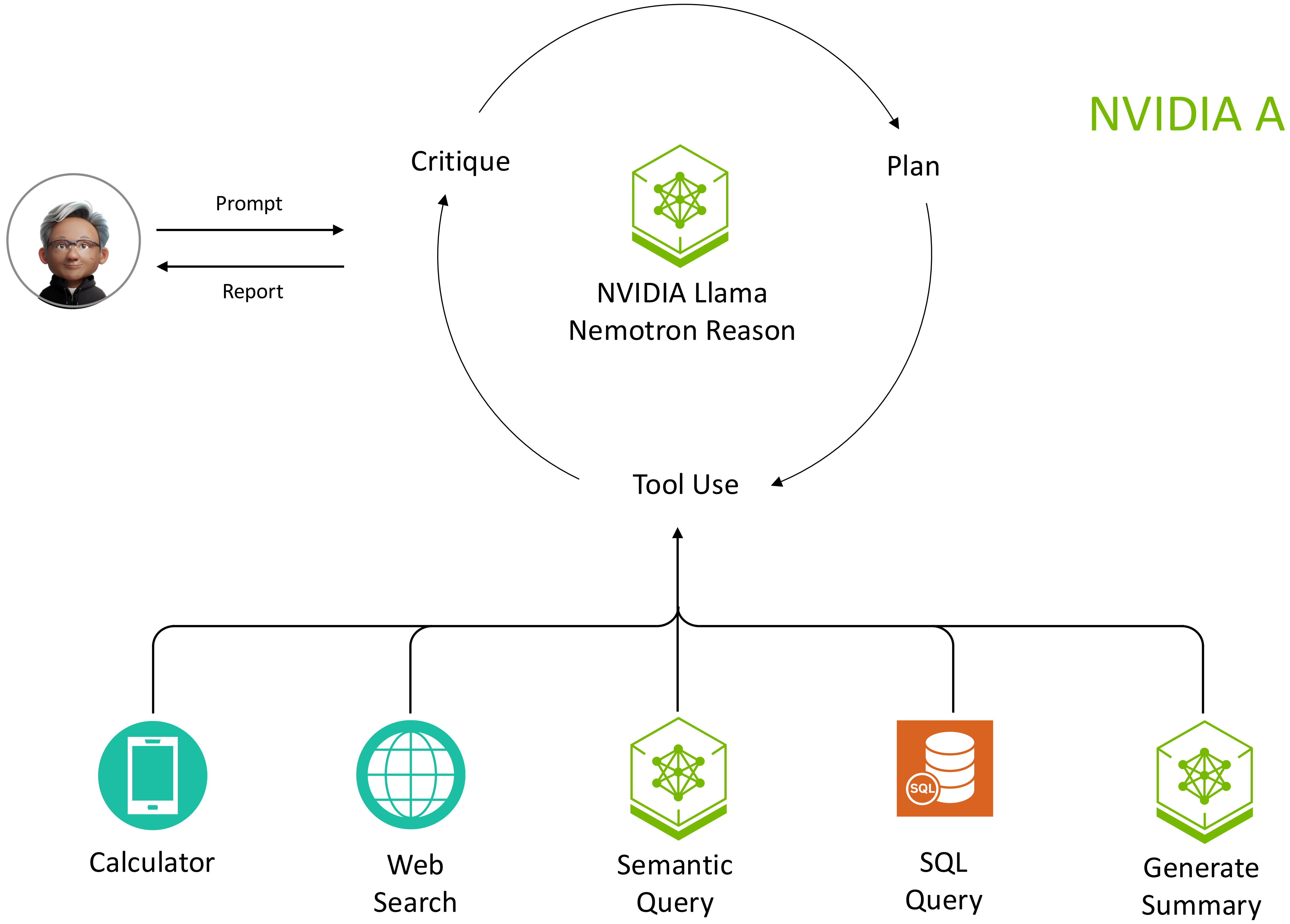
Networking

# AI-Q: AI Agent Interface to Enterprise Data Stores

## AI Data Platform Reference Design

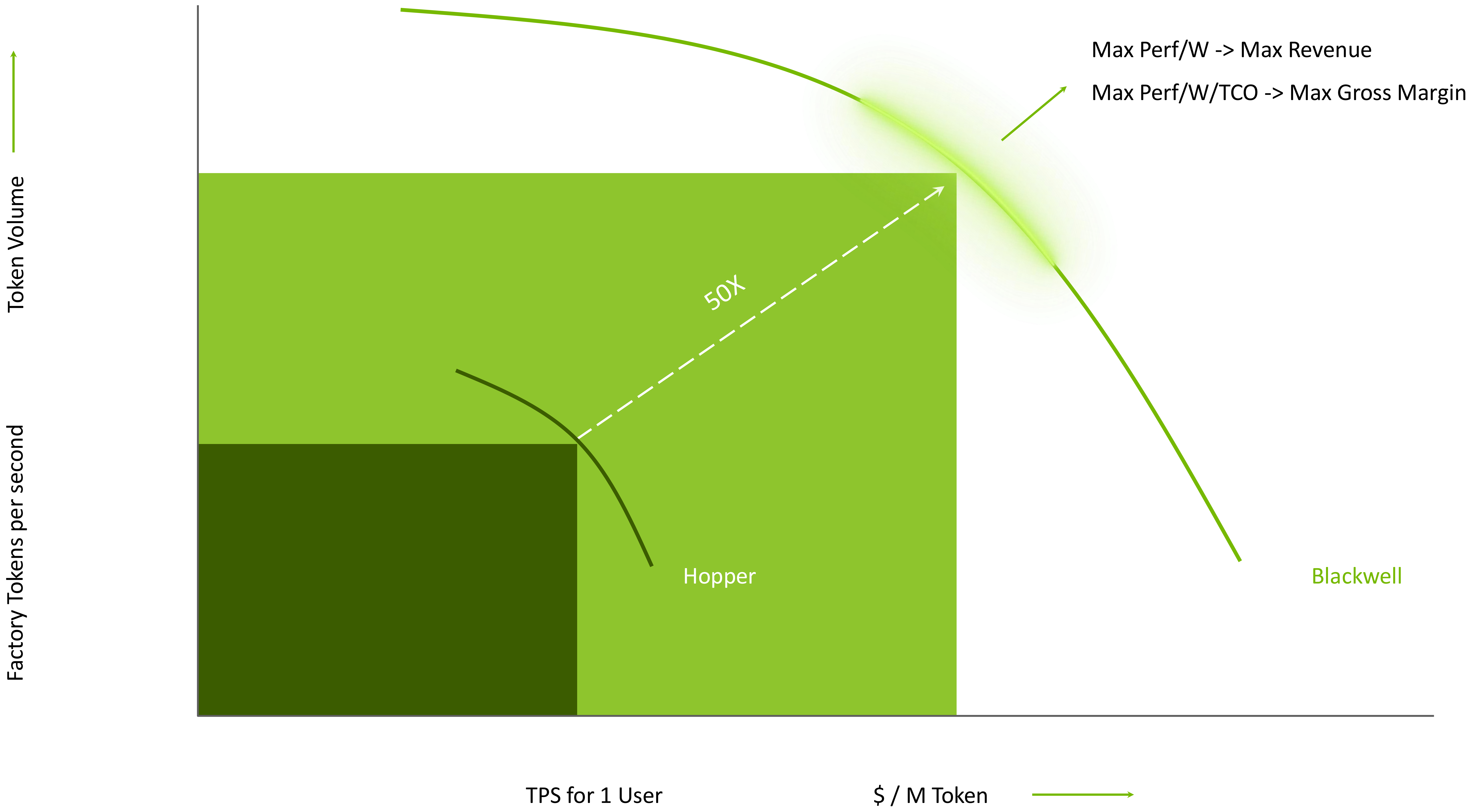


NVIDIA AI-Q



# AI Factory Output Drives Revenue

High throughput multiplied by high interactivity = total token output



# Today’s AI Challenges Require AI Factories

AI workloads require optimized full stack solutions

### Design Complexity

Spans project prioritization,
data acquisition,
and infrastructure

### Deployment and Cost

Infrastructure, security,
and customization

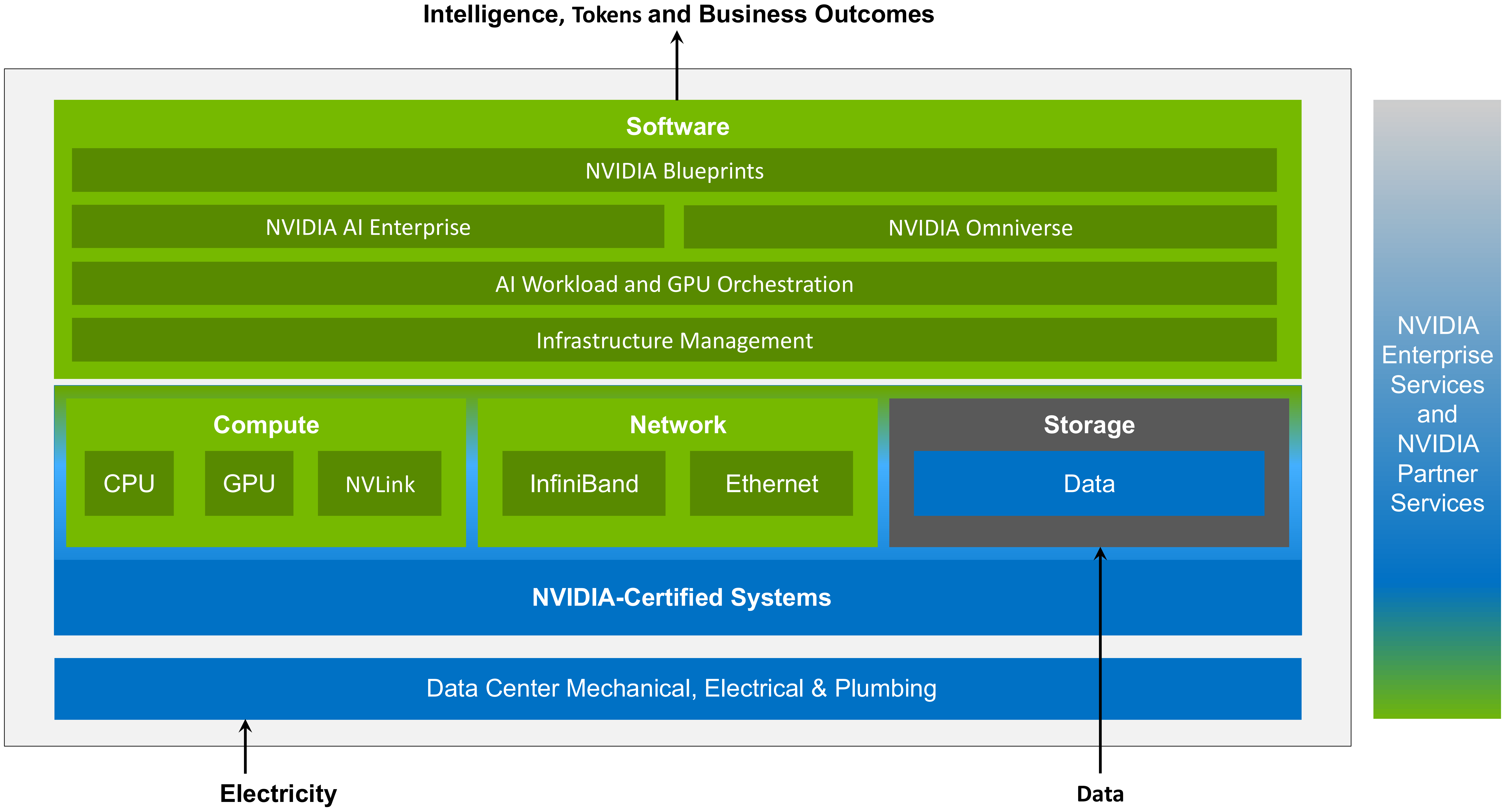
### Time to Value

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

# Enterprise Reference Architectures

# NVIDIA Provides a Full Stack for AI Factories

Built on NVIDIA customer-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
Storage
Data
NVIDIA-Certified Systems
Data Center Mechanical, Electrical & Plumbing
Electricity
Data
NVIDIA Enterprise Services and NVIDIA Partner Services

# Every Enterprise Will Have an AI Factory

## Enterprise Owned

### Enterprise AI Factory

Enterprise AI Factory Validated Design

**NVIDIA Enterprise RAs**
NVIDIA Compute
NVIDIA Networking
Partner Ecosystem

## Enterprise Rented

### NCP AI Factory

NCP Software Stack
NVIDIA AI Enterprise

NVIDIA NCP RAs
NVIDIA Compute
NVIDIA Networking
Partner Ecosystem

### CSP AI Factory

CSP Software Stack
NVIDIA AI Enterprise

CSP Designed Infrastructure
NVIDIA Compute
NVIDIA Networking
Partner Ecosystem

# AI Factory Compounds Benefits Of NVIDIA Reference Architectures

Software
- NVIDIA AI Enterprise — **5X** Throughput with NIM
- NVIDIA Run:ai — **5X** GPU utilization

Compute
- CPU — **2X** Energy efficiency with Grace Superchip
- GPU — **5X** TFLOPS with Blackwell
- NVLink — **1.5X** Throughput with NVLink

Network
- InfiniBand
- SP-X Ethernet

**2X** Performance

**1.3X** Faster Time Deploy

Time-to-value

Time-to-first-train

Time-to-inference

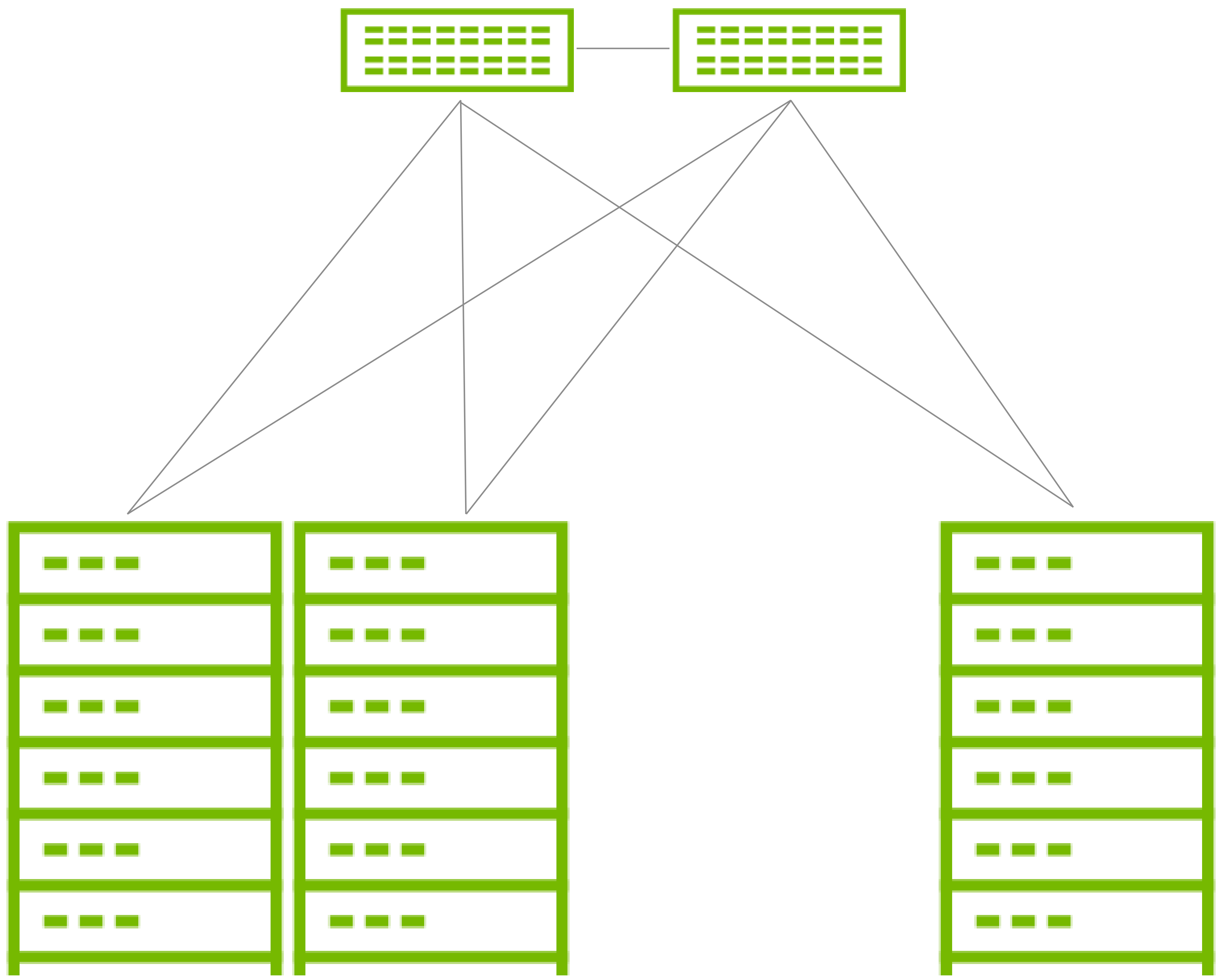
Scalability

Up to these speeds, may vary based on input/output and model

# Building on NVIDIA Enterprise Reference Architectures

**Networking**

NVIDIA Spectrum-X Ethernet Networking



**Compute**

**Storage**

NVIDIA-Certified Systems

**RTX PRO Blackwell Servers**
or HGX B200 Systems

NVIDIA-Certified Storage

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

# Introducing: Enterprise Reference Architectures

**Comprehensive full-stack design recommendations for building high-performance, scalable data center infrastructure.**

- NVIDIA-Certified Systems
  - Optimized Scale-Up & Scale-Out Configurations
- High-Performance AI Networking
  - Spectrum-X
- AI Software Stack
  - NVIDIA AI Enterprise

**Deployment Guides for Multiple Workloads**

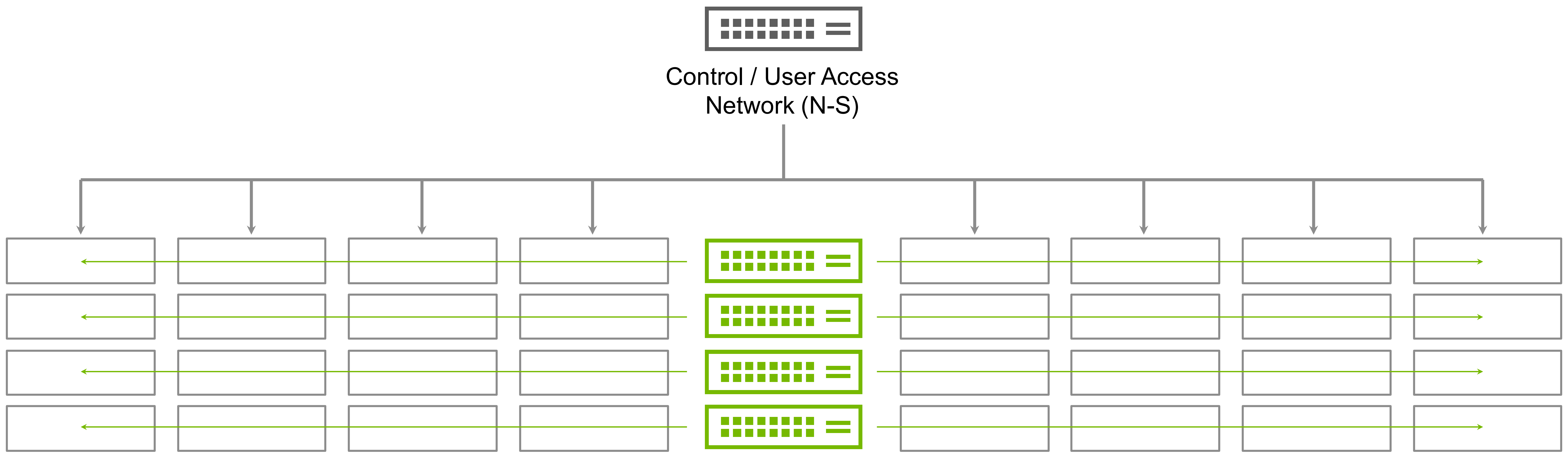
- LLM, RAG, NIM, and NIM Agent Blueprints

**Flexible Sizing for Expansion Needs**

- Multiple discrete design points for 32 - 512 GPUs
  - Optimized resource utilization

# Only NVIDIA Networking Delivers the Fabric for AI Factories

Tightly coupled | Isolated | High-bandwidth | Low jitter | Predictable performance at any scale

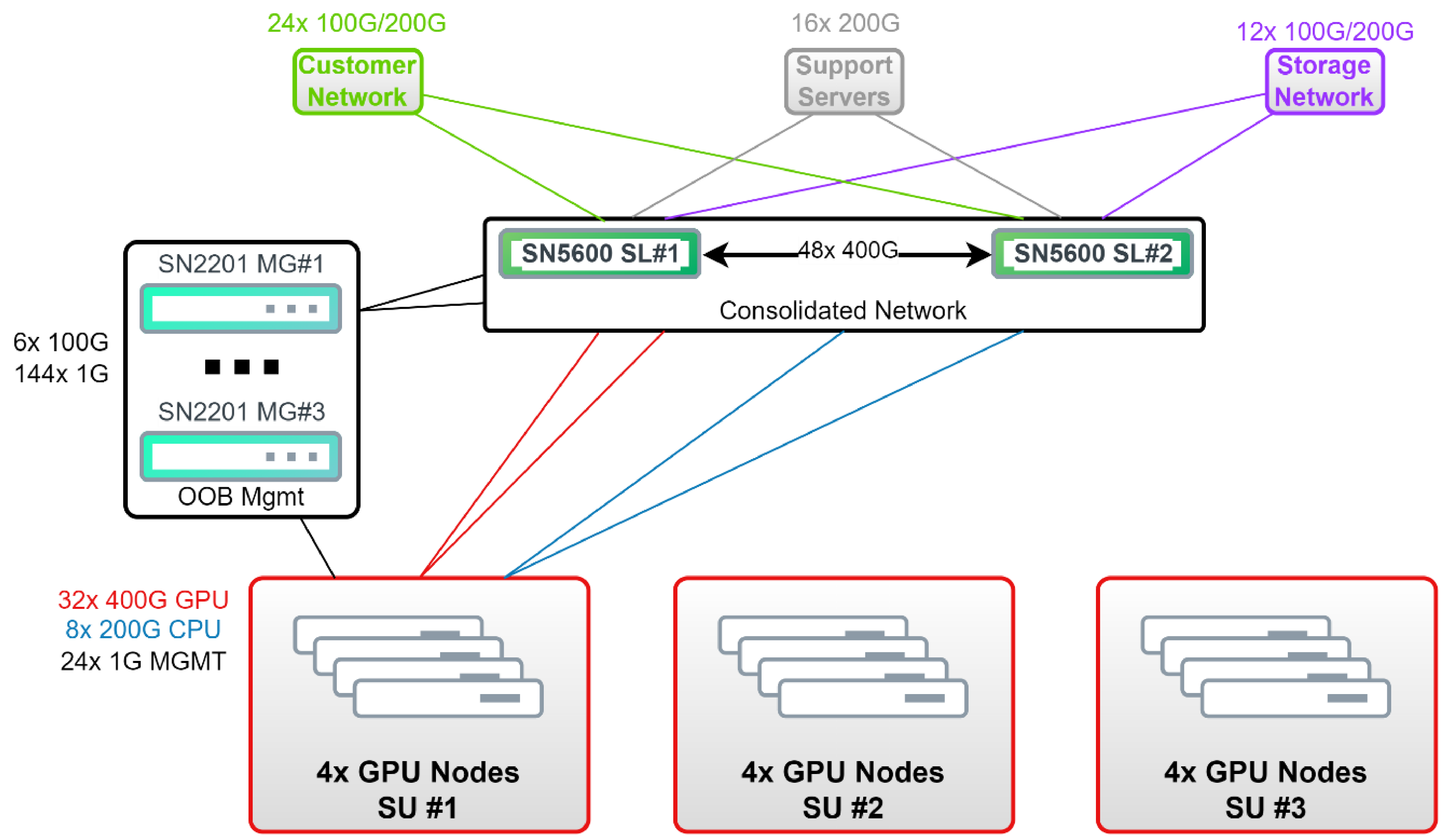


| Control / User Access Network (N-S) | AI Fabric (E-W) |
|---|---|
| Loosely-coupled applications, no isolation required | Tightly-coupled processes, tenant isolation required |
| TCP (low bandwidth flows and utilization) | RDMA (high bandwidth flows and utilization) |
| High jitter tolerance | Low jitter tolerance |
| Heterogeneous traffic, statistical multi-pathing | Bursty network capacity, predictive performance |

# Extensible Designs Enable Efficient Scaling

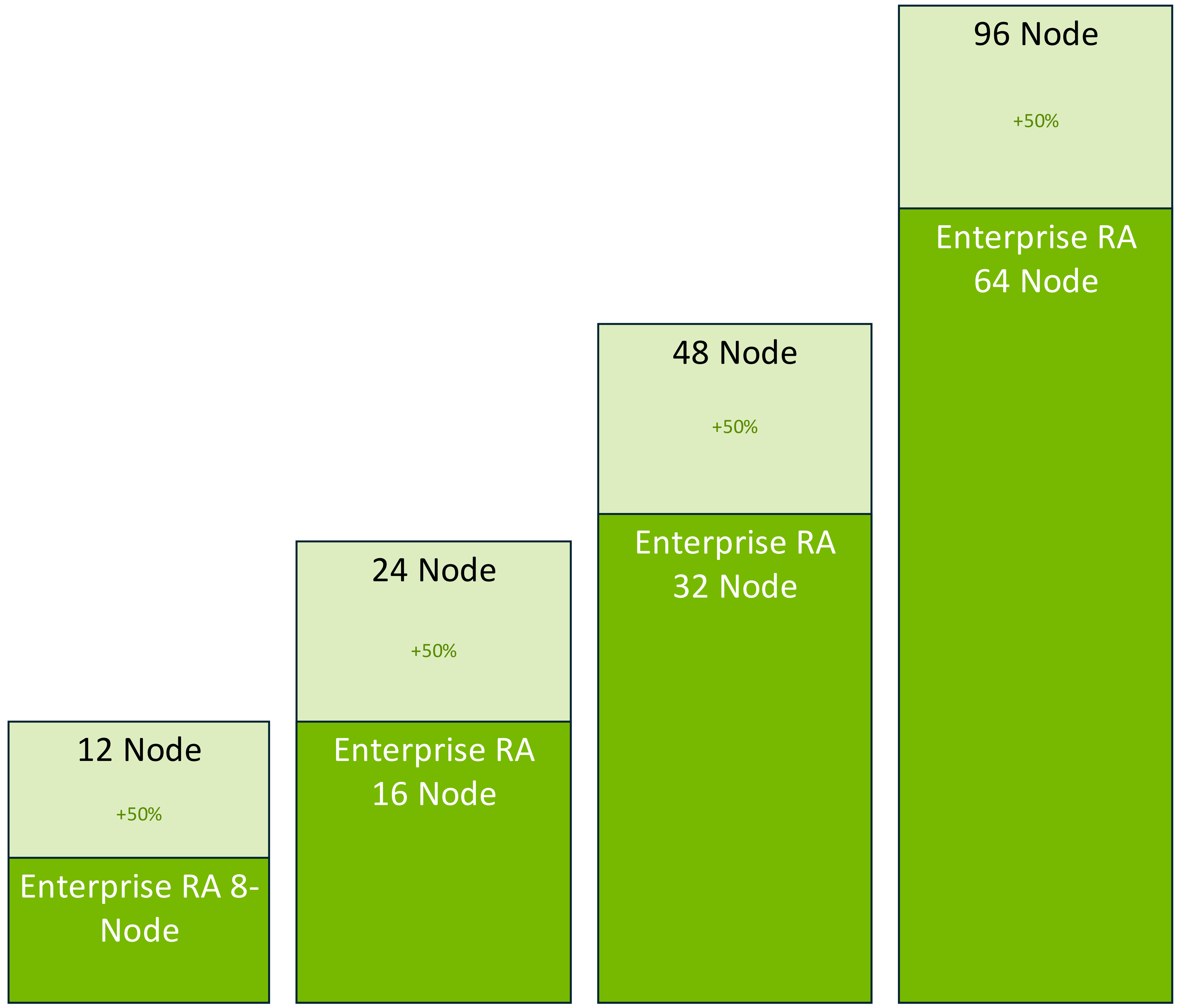
## Enterprise RA Networking w Spectrum-X

### Rail-Optimized Topology



**Example : HGX H100 3SU Cluster, 96 GPUs**
**1SU = 4 Nodes**

### Efficient scaling at multiple design points



# Resources

## **Enterprise** Reference Architecture Announcement

### Resources

**Enterprise RA**

- Announcement Blog
- Web
- Enterprise RA Whitepaper

**NVIDIA-Certified**

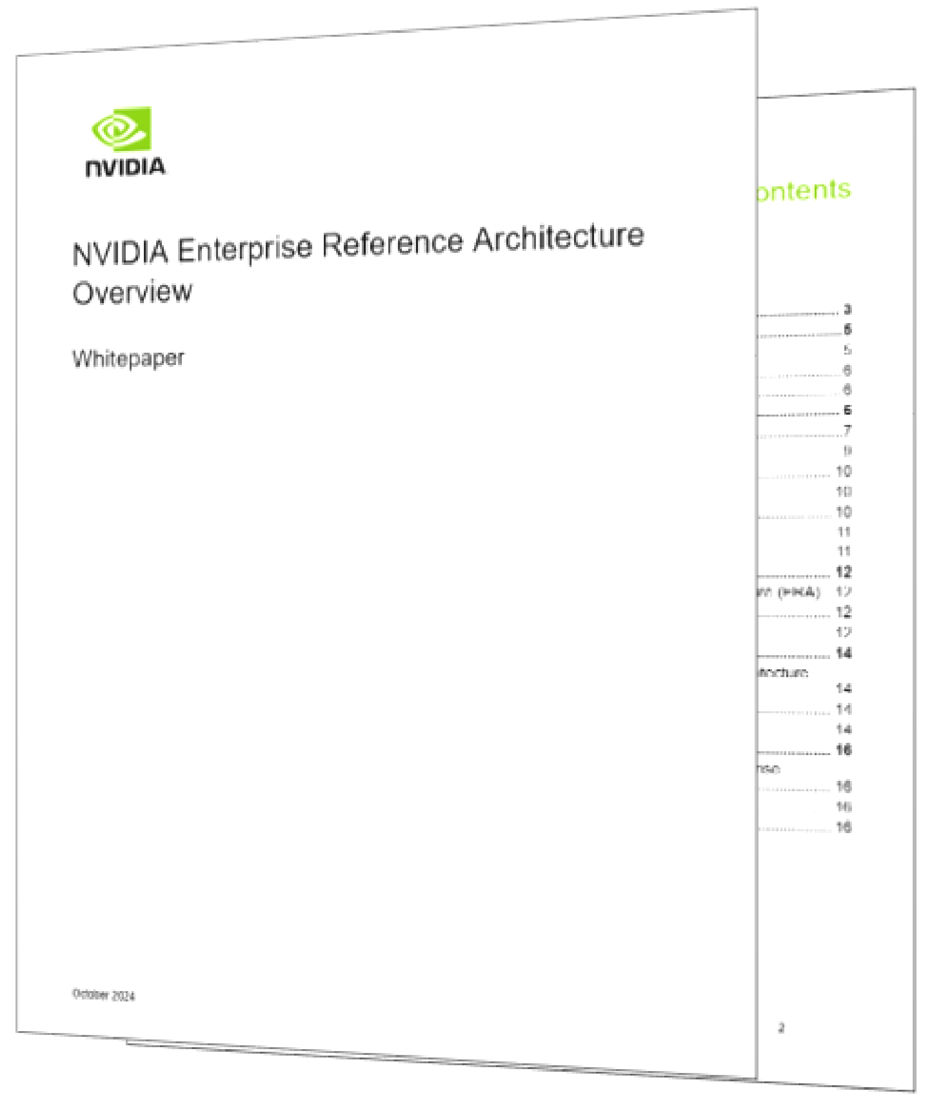
- Whitepaper
- Data Sheet

A New ERA of AI Factories: NVIDIA Unveils Enterprise Reference Architectures

Global enterprises can now tap into new reference architectures to build high-performance, scalable and secure data centers.

By Bob Pette

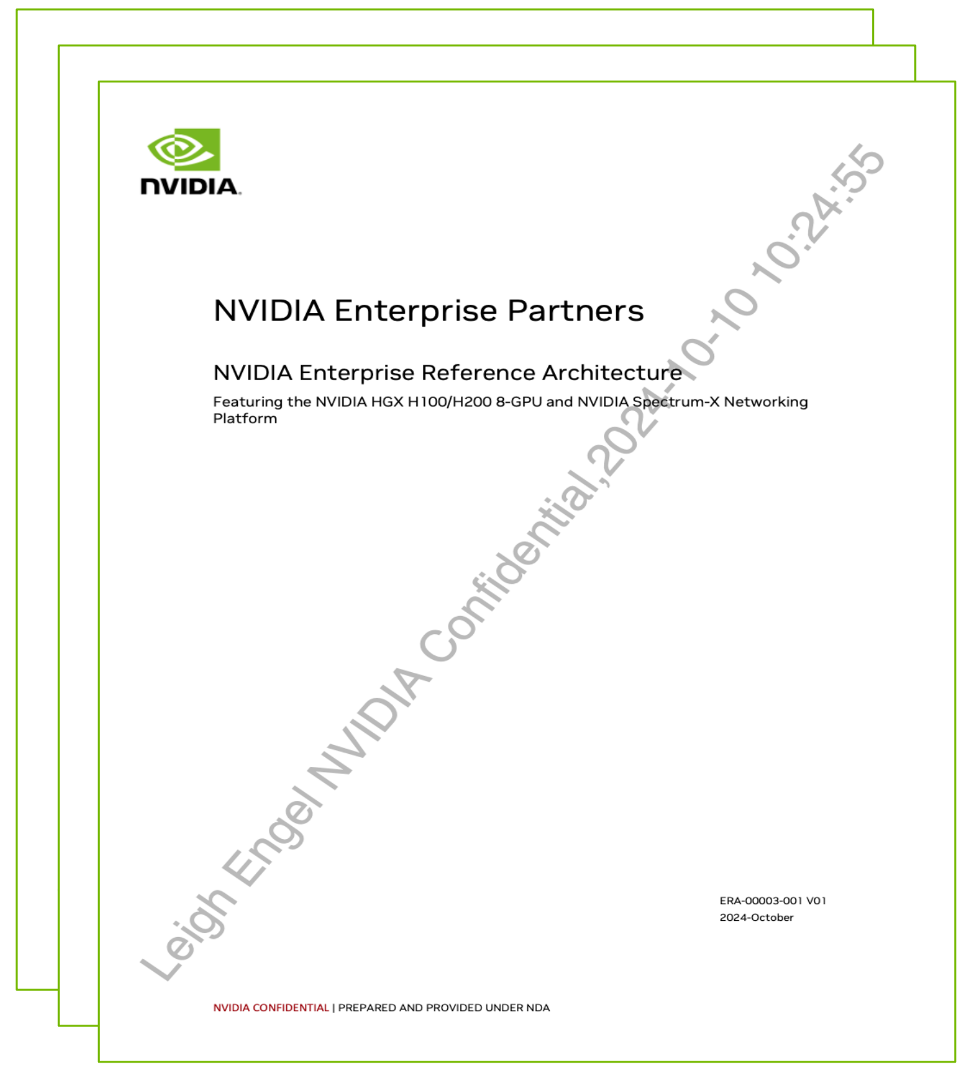
All NVIDIA News



NVIDIA Enterprise Reference Architecture Overview

Whitepaper

October 2024



NVIDIA Enterprise Partners

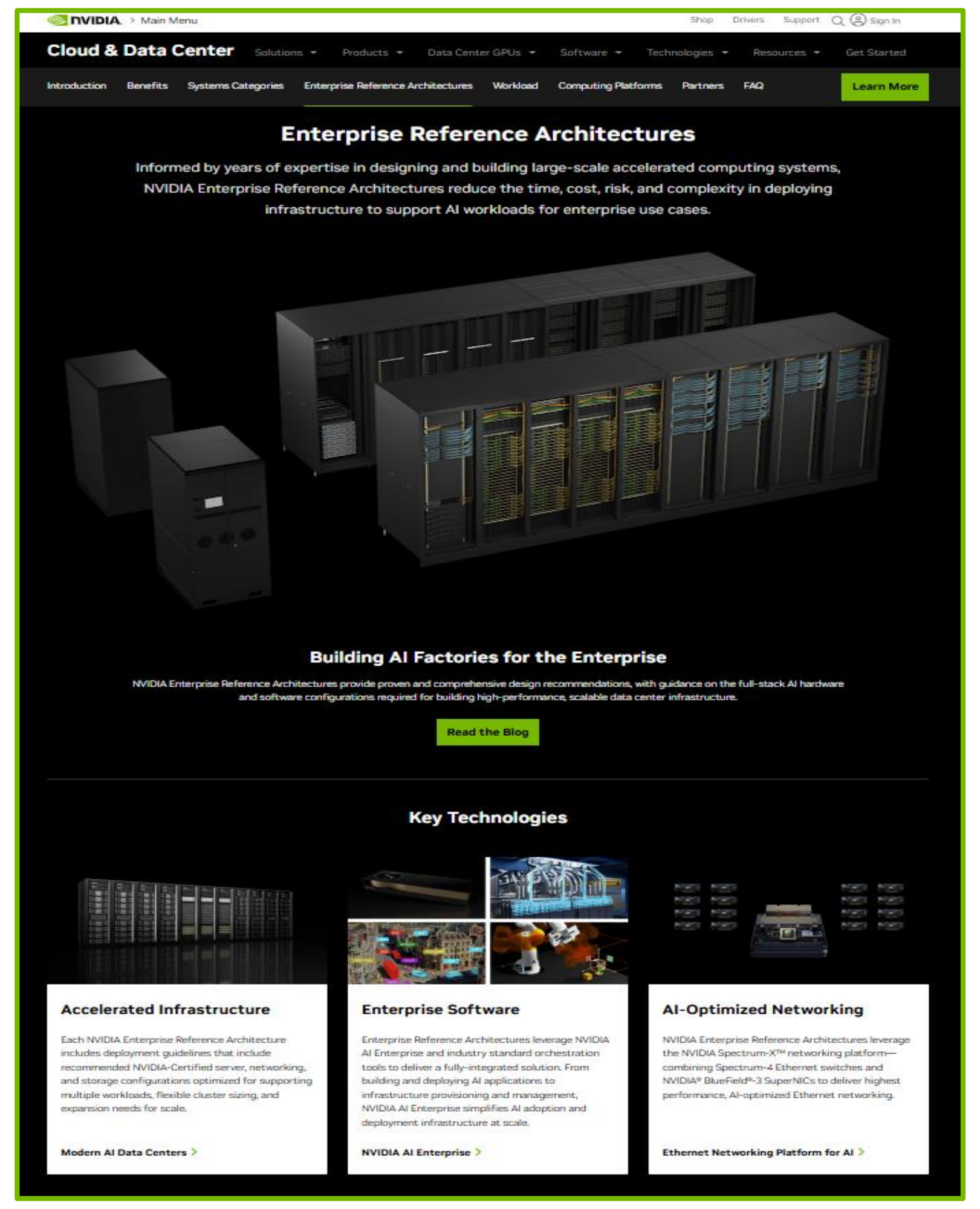
NVIDIA Enterprise Reference Architecture

Featuring the NVIDIA HGX H100/H200 8-GPU and NVIDIA Spectrum-X Networking Platform

ERA-00003-001 V01

2024-October

NVIDIA CONFIDENTIAL | PREPARED AND PROVIDED UNDER NDA



Enterprise Reference Architectures

Informed by years of expertise in designing and building large-scale accelerated computing systems, NVIDIA Enterprise Reference Architectures reduce the time, cost, risk, and complexity in deploying infrastructure to support AI workloads for enterprise use cases.

Building AI Factories for the Enterprise

NVIDIA Enterprise Reference Architectures provide proven and comprehensive design recommendations, with guidance on the full-stack AI hardware and software configurations required for building high-performance, scalable data center infrastructure.

Read the Blog

Key Technologies

Accelerated Infrastructure

Each NVIDIA Enterprise Reference Architecture includes deployment guidelines that include recommended NVIDIA-Certified server, networking, and storage configurations optimized for supporting multiple workloads, flexible cluster sizing, and expansion needs for scale.

Modern AI Data Centers

Enterprise Software

Enterprise Reference Architectures leverage NVIDIA AI Enterprise and industry standard orchestration tools to deliver a fully-integrated solution. From building and deploying AI applications to infrastructure provisioning and management, NVIDIA AI Enterprise simplifies AI adoption and deployment infrastructure at scale.

NVIDIA AI Enterprise

AI-Optimized Networking

NVIDIA Enterprise Reference Architectures leverage the NVIDIA Spectrum-X™ networking platform—combining Spectrum-4 Ethernet switches and NVIDIA® BlueField®-3 SuperNICs to deliver highest performance, AI-optimized Ethernet networking.

Ethernet Networking Platform for AI

# Helping Enterprises Build AI Factories

## Enterprise Reference Architecture

### Enterprise Reference Architectures

**Publish Proven and Comprehensive Designs**

- NVIDIA-Certified systems
- Network Topology
- Storage + Management
- Software Stack

### NVIDIA Partners

CISCO

DELL Technologies

Hewlett Packard Enterprise

Lenovo

SUPERMICRO

**Adopted by Our Partners**

- Faster TTM
- Flexible Workload Sizing
- Smoother deployment process

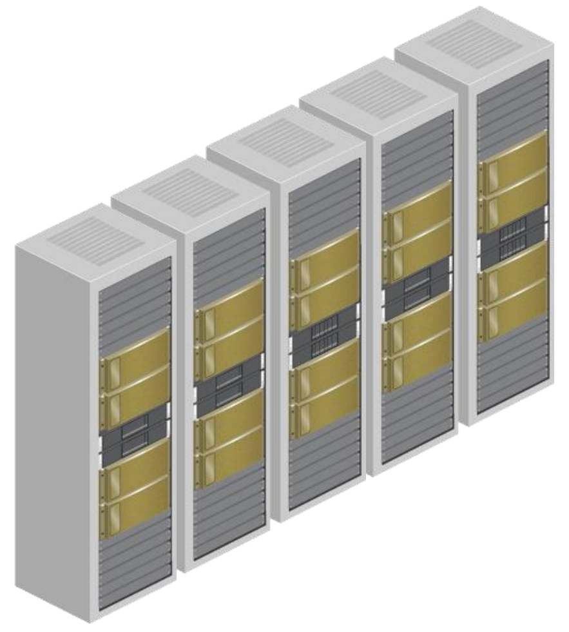
### Enterprises

**Superior Customer Experience**

- Faster Time to Solution
- Performant Infrastructure
- Reduced Complexity and Cost

# Understanding DGX BasePOD Vs DGX SuperPOD

## NVIDIA DGX BasePOD



VS

## NVIDIA DGX SuperPOD

**Scalable, foundational architecture**

**Physical twin of NVIDIA's infrastructure**

- Flexible reference architectures
- Powered by Base Command
- Validated against key **benchmarks**
- Foundation for **partner branded offerings**

- Turnkey data center **product**
- Powered by Base Command
- **Certified performance** for the most **complex workloads**
- **No customization, no partner re-branding**

### I need:

- Choice of flexible vs performance optimized designs
- Inclusion of non-SuperPOD certified storage

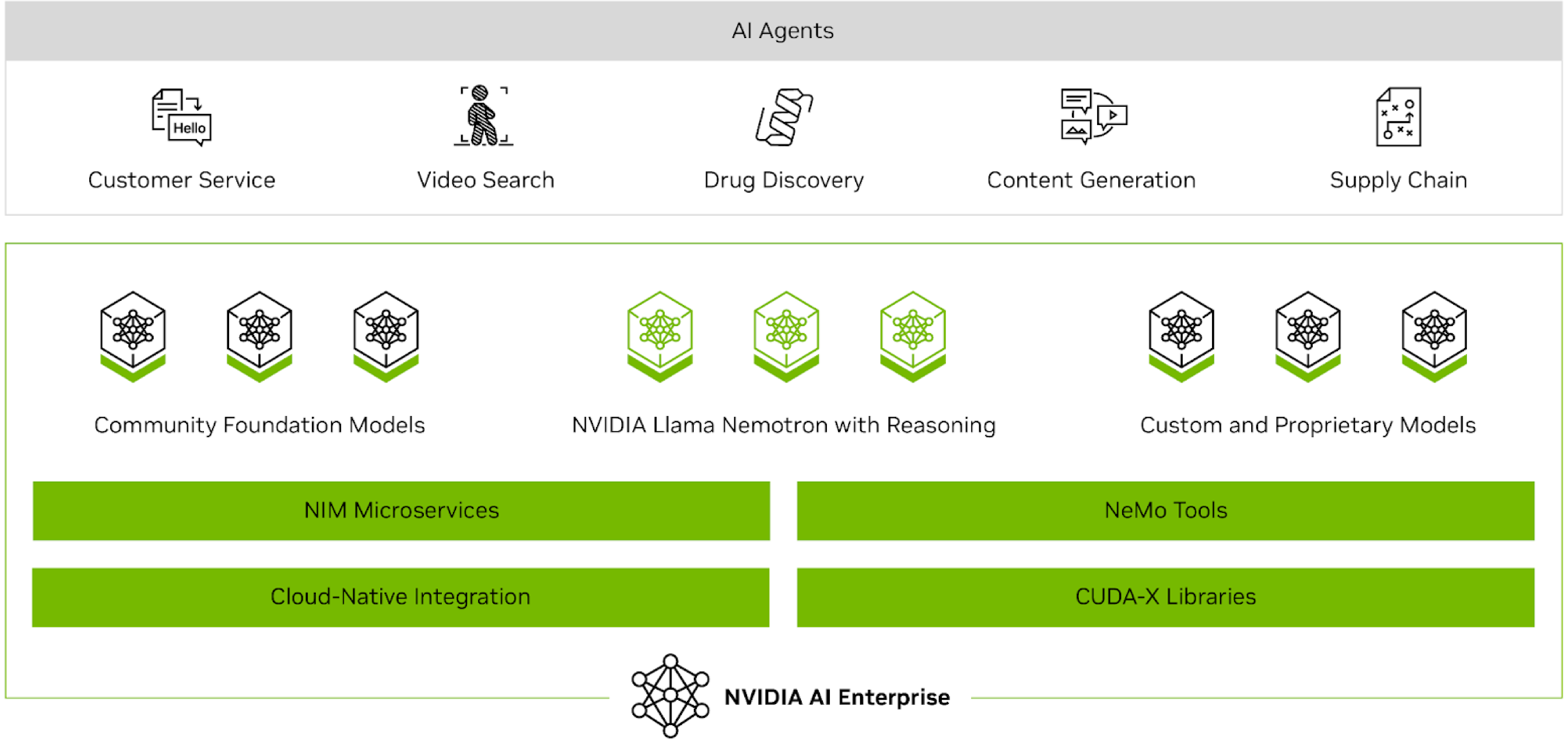
### I need:

- A replica of NVIDIA infrastructure
- A turnkey deployment
- Full-stack support of the entire deployment

# Fastest Path to Production

# NVIDIA AI Enterprise

## What's Inside?



AI Agents

Customer Service

Video Search

Drug Discovery

Content Generation

Supply Chain

Community Foundation Models

NVIDIA Llama Nemotron with Reasoning

Custom and Proprietary Models

NIM Microservices

NeMo Tools

Cloud-Native Integration

CUDA-X Libraries

NVIDIA AI Enterprise

**Comprehensive collection of preconfigured NIM microservices** for efficient inferencing of state-of-the-art foundation models for any use case.

**Powerful, ready-to-use NeMo training, evaluation, and guardrailing tools** and RAG building blocks to accelerate time to deployment.

**Infrastructure software** to help manage AI clusters at scale, across the edge and data center, both bare-metal and virtualized.

## Onboard Your AI Agents!

# NVIDIA NIM: Optimized AI Models Run Up to 5X Faster

Community models – partner models – NVIDIA models

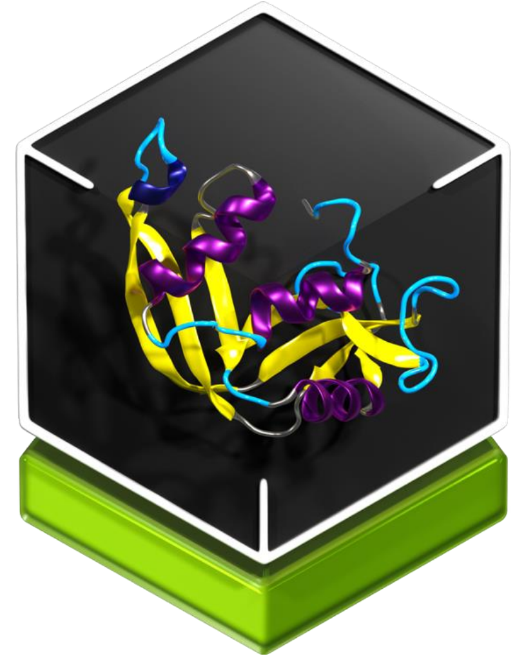
NVIDIA INFERENCE MICROSERVICE

Pre-Trained AI Models
Packaged and Optimized to Run Across
CUDA Installed Base

Speech

Digital Human

Computer Vision

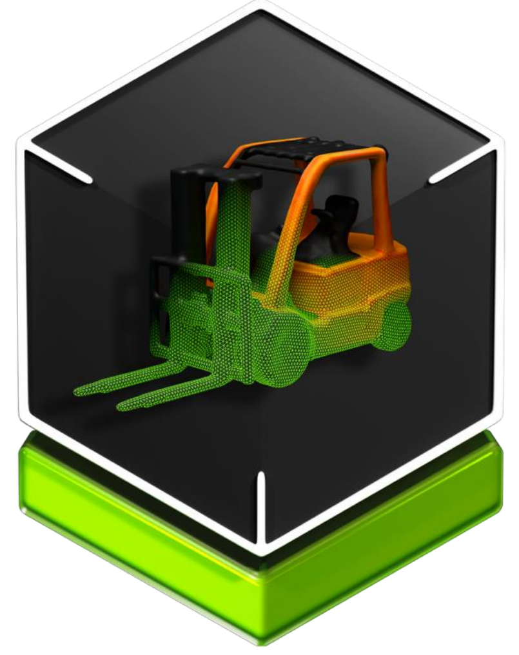


Biology

Simulation

Language

Regional Language



Vision
Language

RAG

ADEPT gettyimages Google Meta MIT MISTRAL AI_ NVIDIA shutterstock snowflake

# NVIDIA NIM Optimized Inference Microservices

Rapidly deploy reliable building blocks for accelerated generative AI anywhere

**Portable** Run cloud-native microservices anywhere, maintaining security and control of data and apps

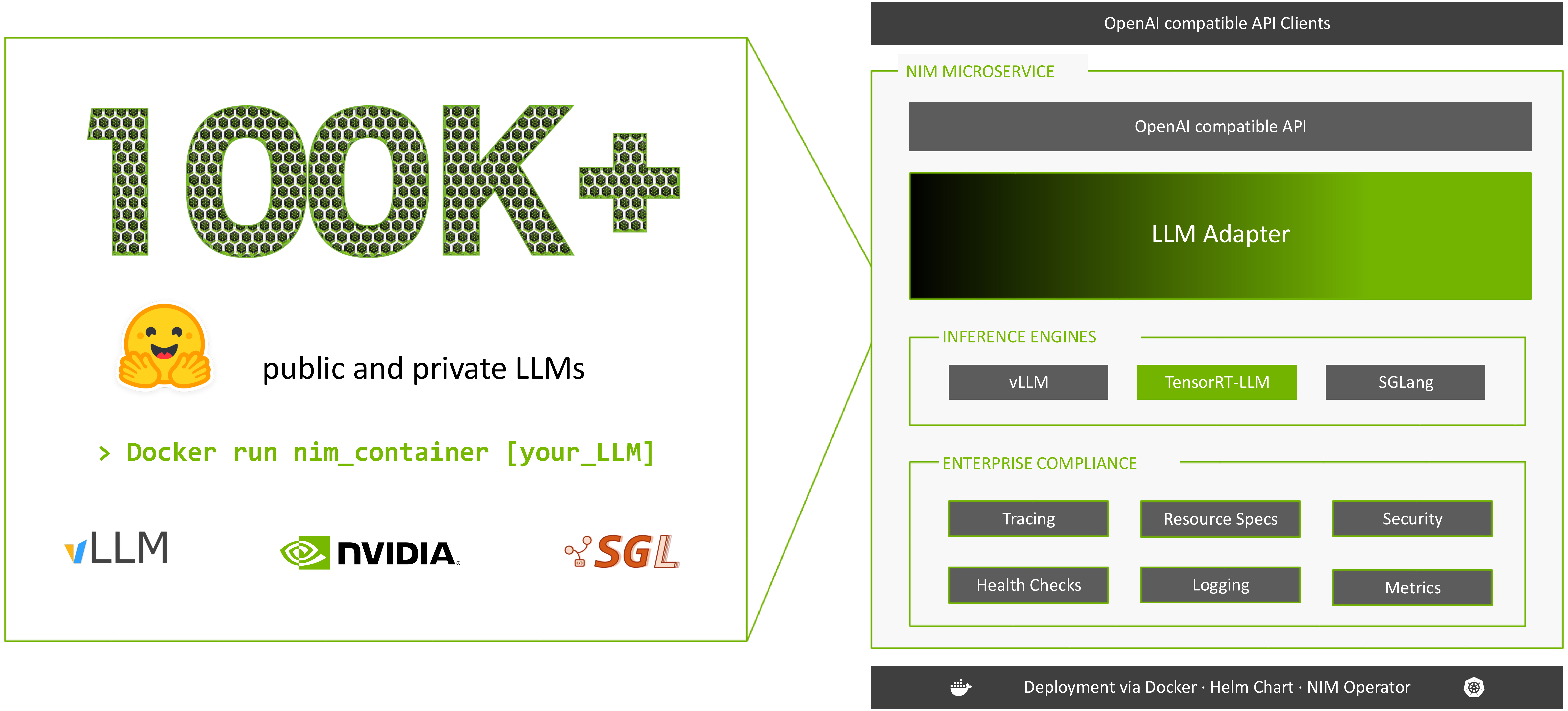
**Easy to Use** Move fast with the latest agentic AI building blocks for reasoning, retrieval, images and more, deployed in minutes with standard APIs

**Enterprise Supported** Gain confidence with stable APIs, quality assurance, continuous updates, security patching, and support

**Performance** Optimize accuracy, latency and throughput to meet requirements with lowest TCO

# Enterprise-ready Inference for a World of LLMs

Single NIM container with multiple inference backends for rapid, reliable deployment of a broad range of LLMs



100K+

public and private LLMs

```
> Docker run nim_container [your_LLM]
```

OpenAI compatible API Clients

NIM MICROSERVICE

OpenAI compatible API

LLM Adapter

INFERENCE ENGINES

vLLM · TensorRT-LLM · SGLang

ENTERPRISE COMPLIANCE

Tracing · Resource Specs · Security

Health Checks · Logging · Metrics

Deployment via Docker · Helm Chart · NIM Operator

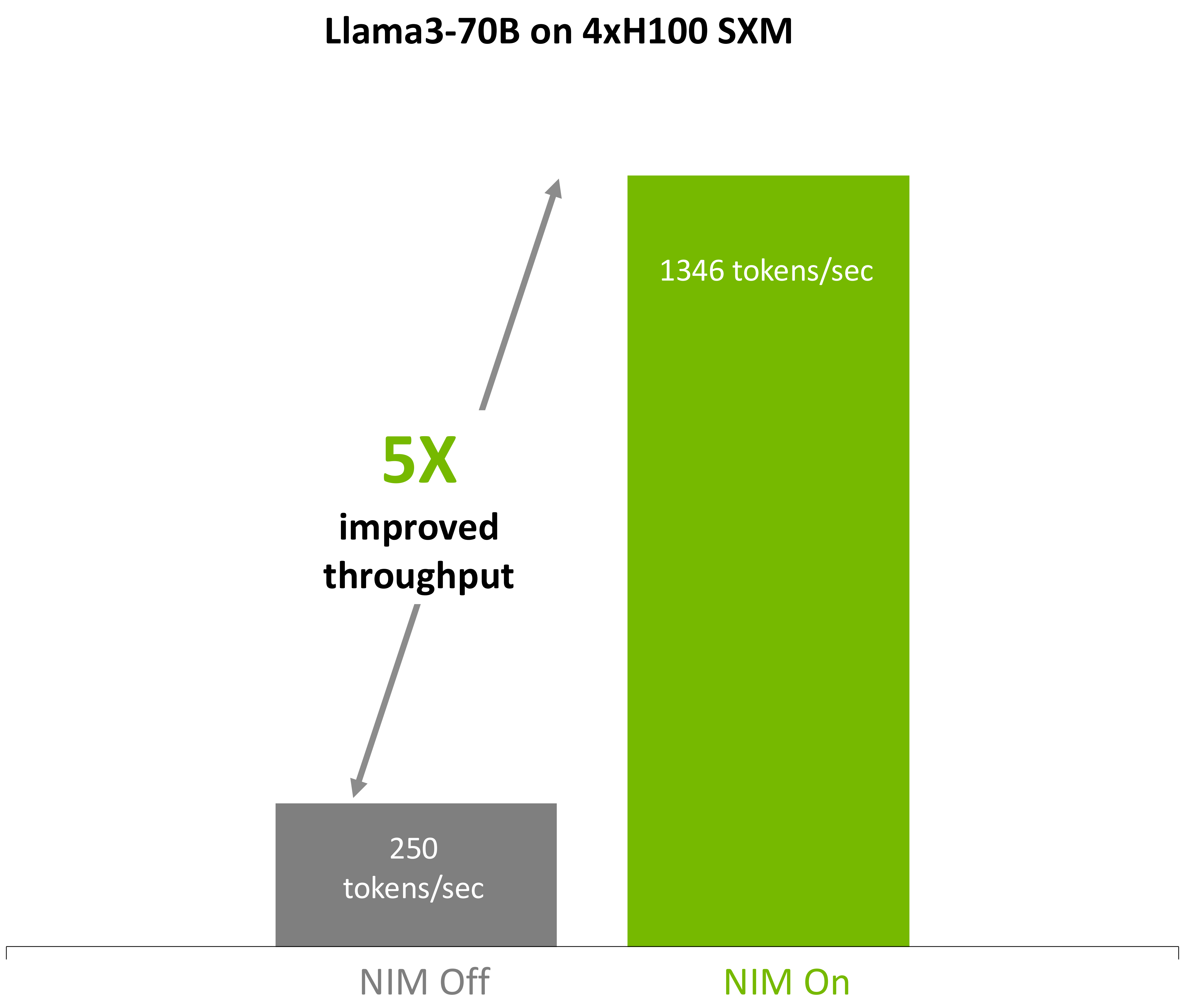
# NVIDIA NIM is the Fastest Path to AI Inference

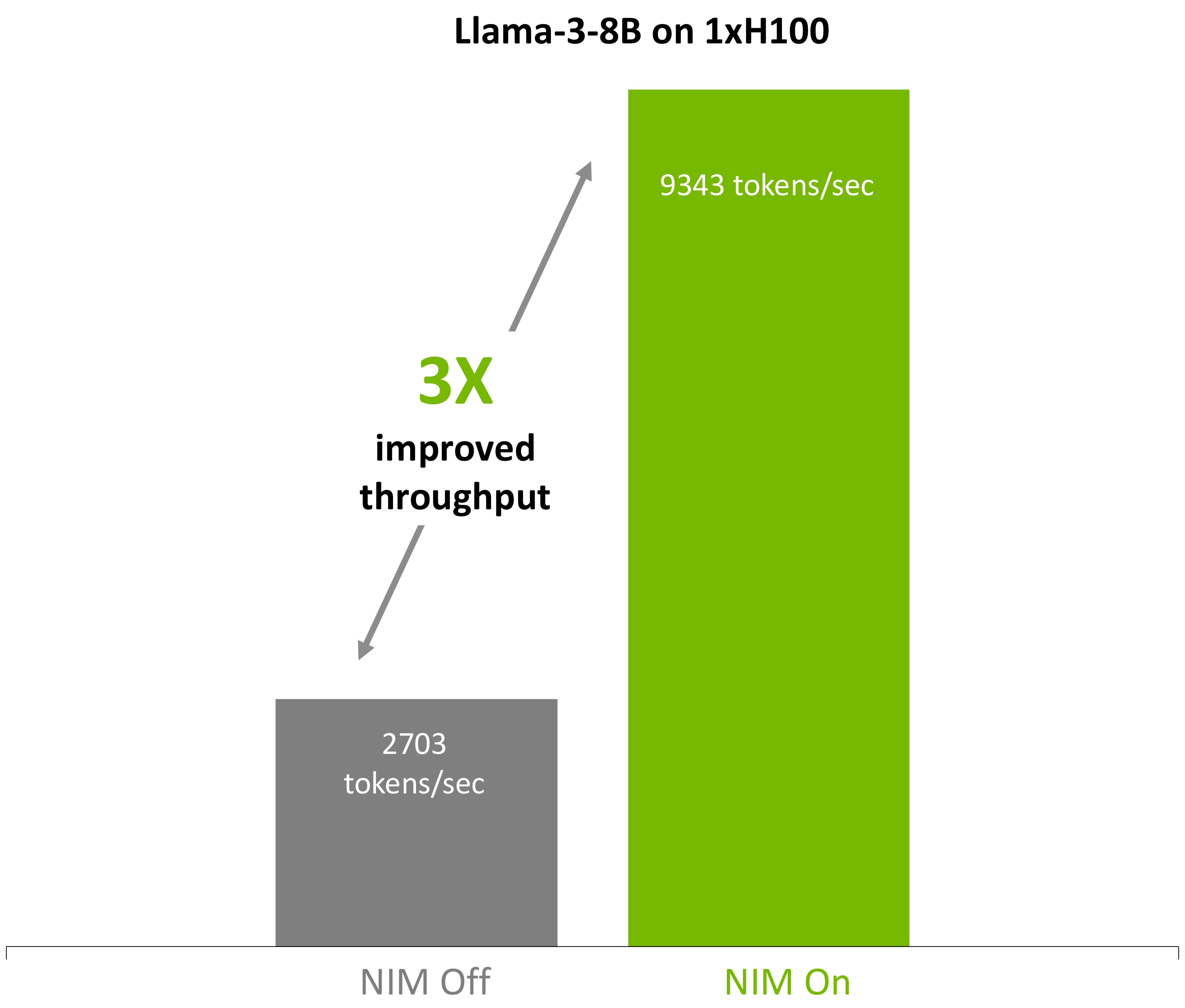
Reduces engineering resources required to deploy optimized, accelerated models

| | NVIDIA NIM | Do It Yourself |
|---|---|---|
| Deployment Time | 5 minutes | 1 week + |
| API Standardization | Industry standard protocol<br>OpenAI for LLMs, Google Translate for Speech | Implement the API layer for each domain and model family according to industry standard specifications |
| Optimized Engines | Pre-built engines for NVIDIA and community models<br>MISTRAL AI_ Meta starcoder NVIDIA Nemotron | Build your own engine and<br>manually customize for workload and hardware specific requirements |
| Pre and Post Processing Pipelines | Pre-built with optimized pipeline engines<br>to handle pre/post processing (tokenization) | Implement custom logic |
| Model Server Deployment | Automated | Manual setup and configuration |
| Customization | LoRA is supported, more planned | Create custom logic |
| Container Validation | Extensive workload specific QA support matrix validation | No validation |
| Enterprise Support | Delivered with NVIDIA AI Enterprise<br>Security and CVE scanning/patching and tech support | Self supported |

# Up to 5x Cost Savings

## Improved Efficiency Reduces Overall Cost of Solution

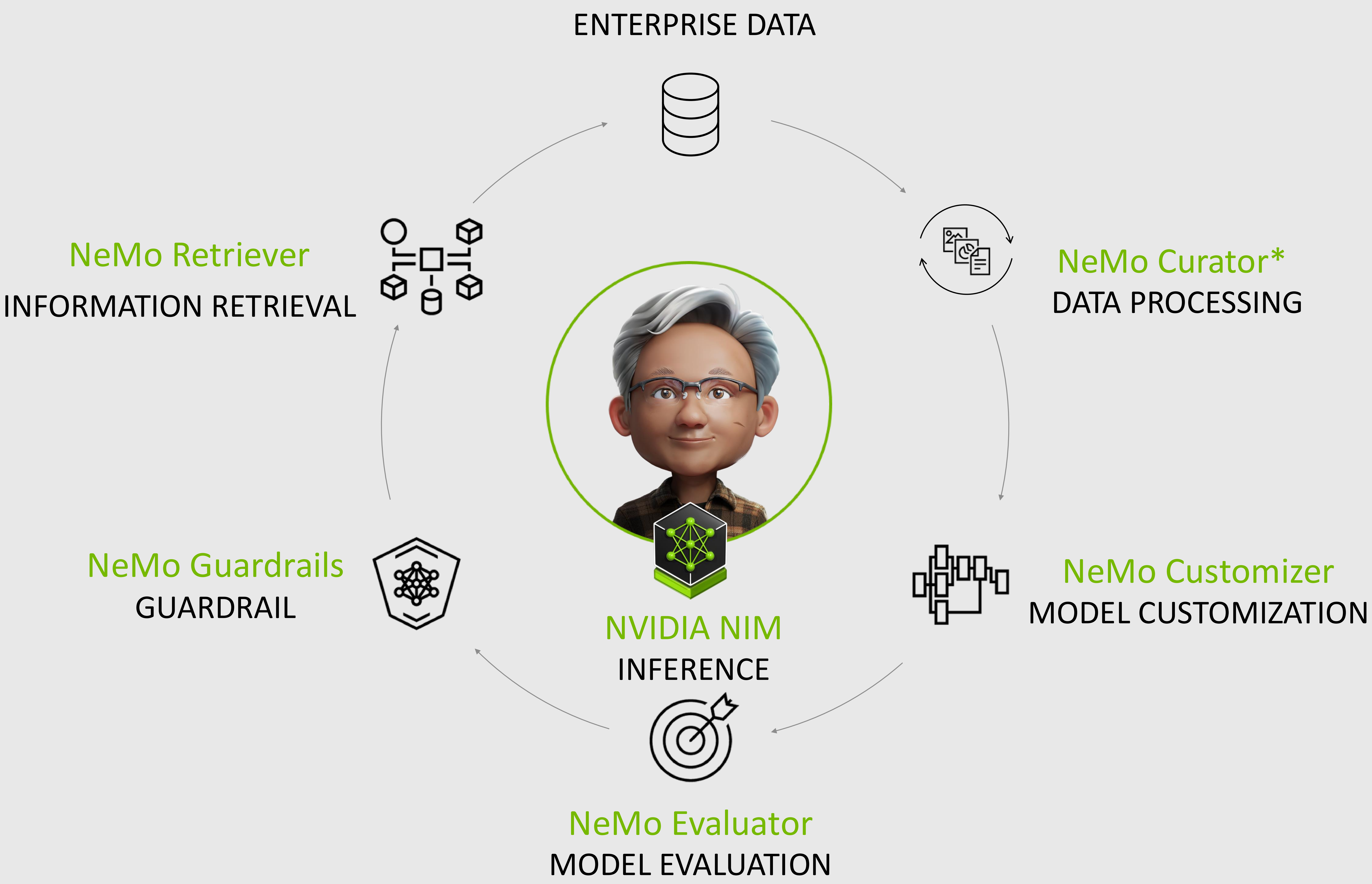




Llama 3-70b-instruct, input token length: 7,000, output token length: 1,000. Concurrent client requests: 100. 4xH100 SXM NVLink. NIM Off: FP16, TTFT: ~120s, ITL: ~180ms. NIM On: FP8. TTFT: ~4.5s, ITL: ~70ms.

# NVIDIA NeMo Microservices

Modular End-to-End Platform to easily build Data Flywheels



ENTERPRISE DATA

NeMo Curator*
DATA PROCESSING

NeMo Customizer
MODEL CUSTOMIZATION

NeMo Evaluator
MODEL EVALUATION

NeMo Guardrails
GUARDRAIL

NeMo Retriever
INFORMATION RETRIEVAL

NVIDIA NIM
INFERENCE

**Easy Setup**
Modular microservices, deployable with standard APIs

**Broad Ecosystem Support**
Integrated in popular open frameworks and AI software platforms

**Enterprise-grade**
Secure, stable, and supported software

**Run Anywhere**
Provides higher security, privacy,  and flexibility

# NVIDIA Blueprints

Available on build.nvidia.com

NVIDIA NIM & microservices

Blueprints

Reference Application

Sample Data

Reference Code

Architecture

Customization Tools

Orchestration Tools

AI Agents

# NVIDIA Provides the Building Blocks for Agentic AI

AI Blueprints

Research Assistant Agent

Customer Service Agent

Software Security Agent

Virtual Lab Agent

Video Analytics Agent

NVIDIA NeMo

Curator

Customizer

Evaluator

Guardrails

Retriever

NVIDIA NIM

Understanding & Reasoning

Information Retrieval

AI Safety

Digital Humans

Visual Content Generation

Digital Biology

Physical AI

AI Infrastructure

GPU

CPU

DPU

Networking

# NVIDIA AI Enterprise Supported Software

For additional information, see NGC Catalog

| SDKs and Frameworks | | |
|---|---|---|
| Maxine | Riva | Modulus |
| TAO Toolkit | MONAI | CUDA |
| DeepStream | Clara | CUDA Toolkit |
| Metropolis | Clara Holoscan | ACE |
| NeMo | RAPIDS | Kubernetes Device Plugin |
| TensorRT | cuOpt | PyTorch Geometric |
| Triton Inference Server | Merlin | PyTorch |
| TensorFlow | Morpheus | DGL |

| NIM Microservices | | |
|---|---|---|
| *Full list at build.nvidia.com/nim* | | |
| **Infrastructure Software Collection** | | |
| NIM Operator | GPU Operator | Network Operator |
| NVIDIA Data Center Driver | NVIDIA vGPU for AI | NVIDIA DOCA Driver for Networking |
| Base Command Manager | | |

| Extended Life Software Branches | |
|---|---|
| Production Branch (9-month) | Long-Term Support Branch (3 year) |

# Summary

# Value Proposition

Enterprise Reference Architectures for NVIDIA Partners and Customers

## Optimized Performance

Comprehensive cluster design recommendations built upon tested and validated technologies to ensure peak computing performance for generative AI workloads; NVIDIA AI Enterprise, NIM, Training, Inference

## Reduced Complexity

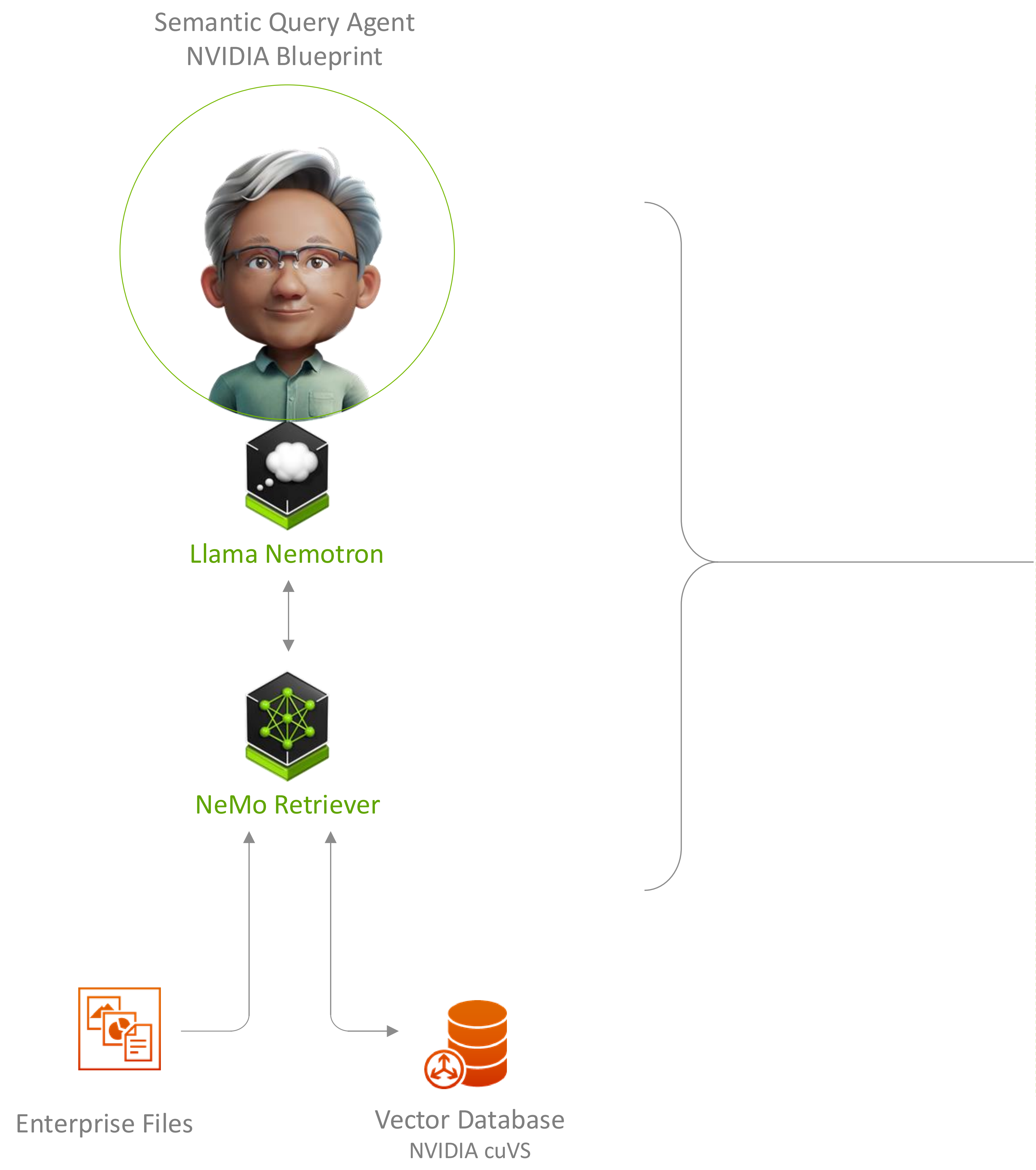
Avoid design and planning pitfalls when deploying infrastructure with detailed guidance on server, cluster, and network configuration, minimizing setup errors and accelerating deployment timelines.

## Flexibility & Scale

Improve resource utilization and eliminate over-provisioning with discrete design points for compute, network, and storage infrastructure based upon deployment scale.

# NVIDIA AI Factory for Enterprise

New class of infrastructure serving knowledge instead of data



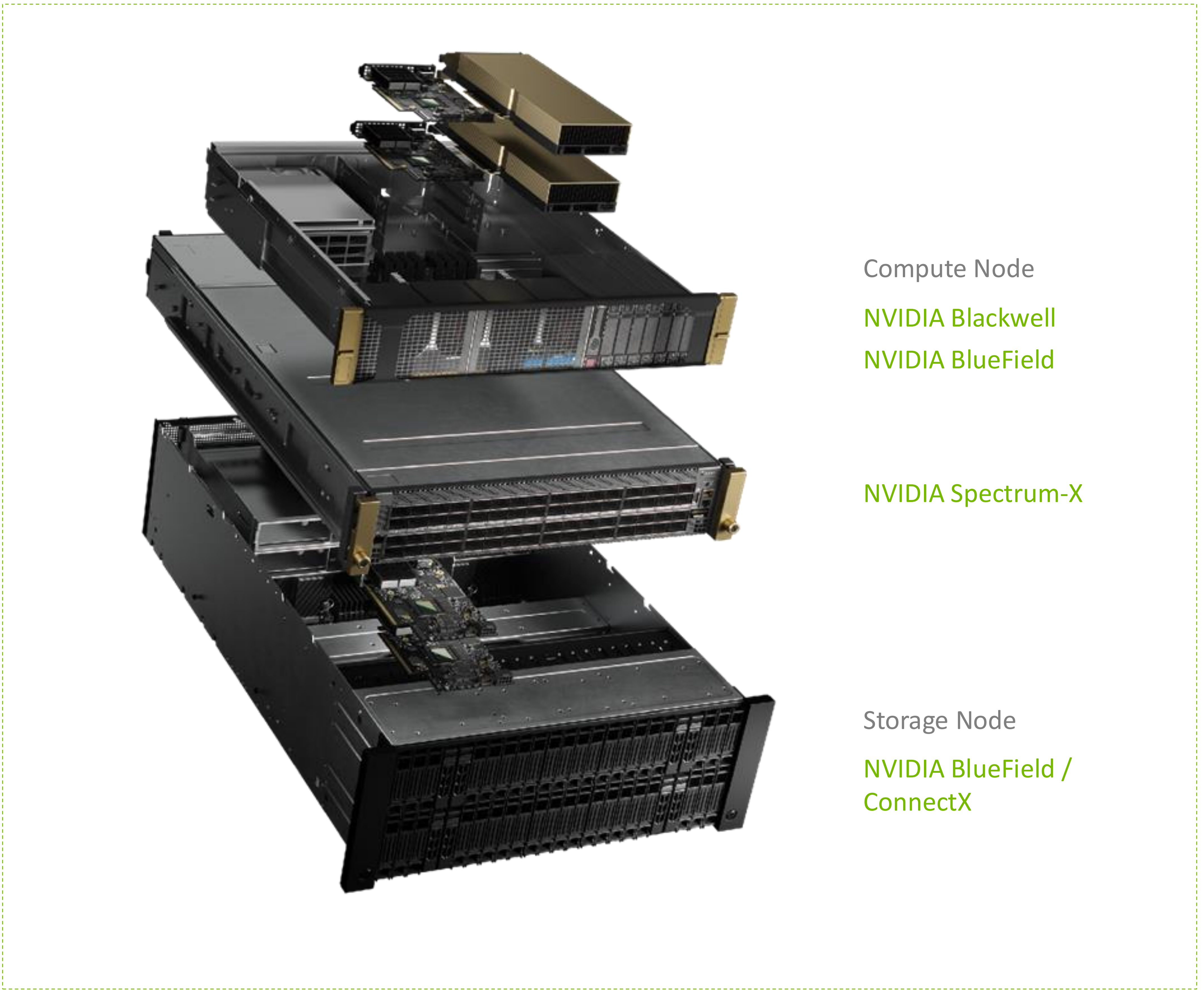
Semantic Query Agent
NVIDIA Blueprint

Llama Nemotron

NeMo Retriever

Enterprise Files

Vector Database
NVIDIA cuVS



Compute Node

NVIDIA Blackwell

NVIDIA BlueField

NVIDIA Spectrum-X

Storage Node

NVIDIA BlueField / ConnectX

PURESTORAGE®

# Enterprise AI Factory Stack

Agentic AI

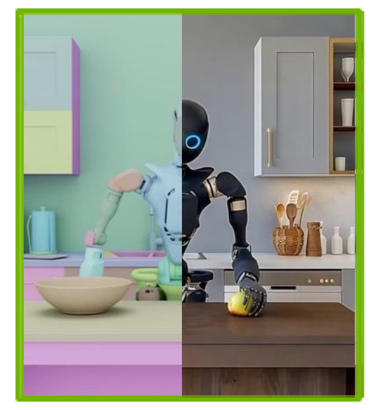
Supply Chain

Customer Service

Marketing

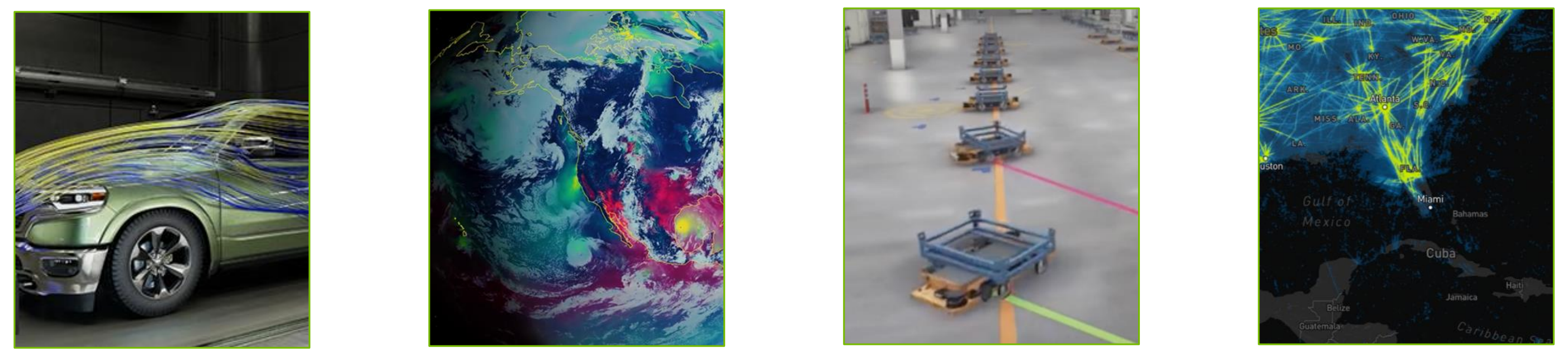
Video Analytics

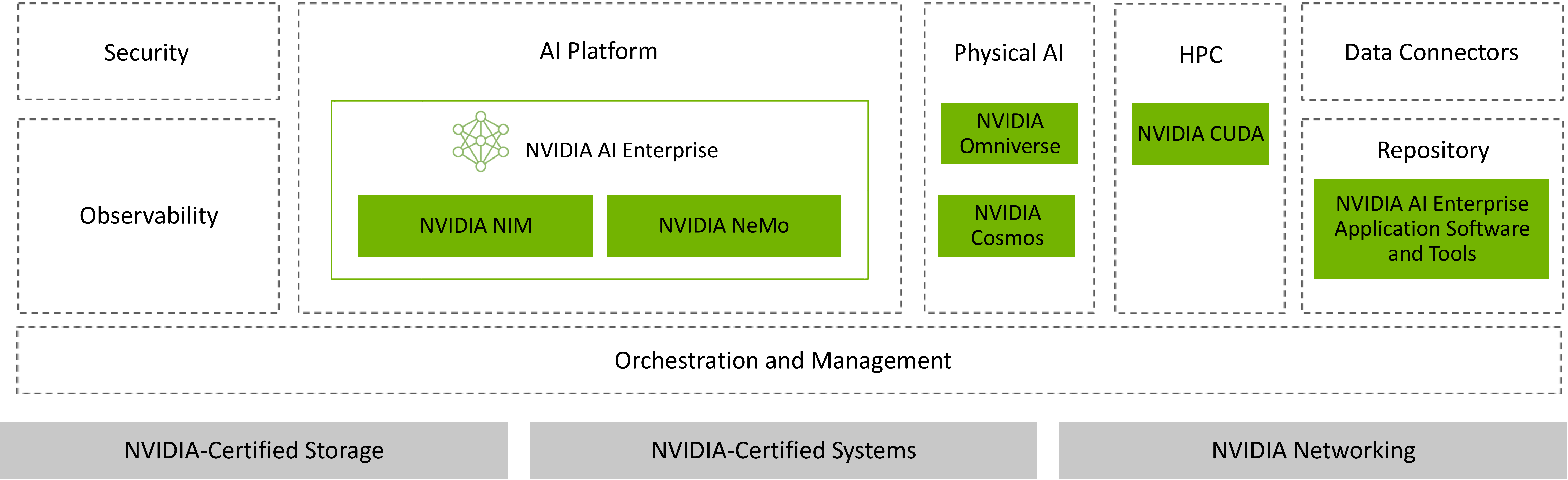
Physical AI


Robotics

Digital Twins

HPC


CUDA-X Libraries


Security

Observability

AI Platform

NVIDIA AI Enterprise

NVIDIA NIM

NVIDIA NeMo

Physical AI

NVIDIA Omniverse

NVIDIA Cosmos

HPC

NVIDIA CUDA

Data Connectors

Repository

NVIDIA AI Enterprise Application Software and Tools

Orchestration and Management

NVIDIA-Certified Storage

NVIDIA-Certified Systems

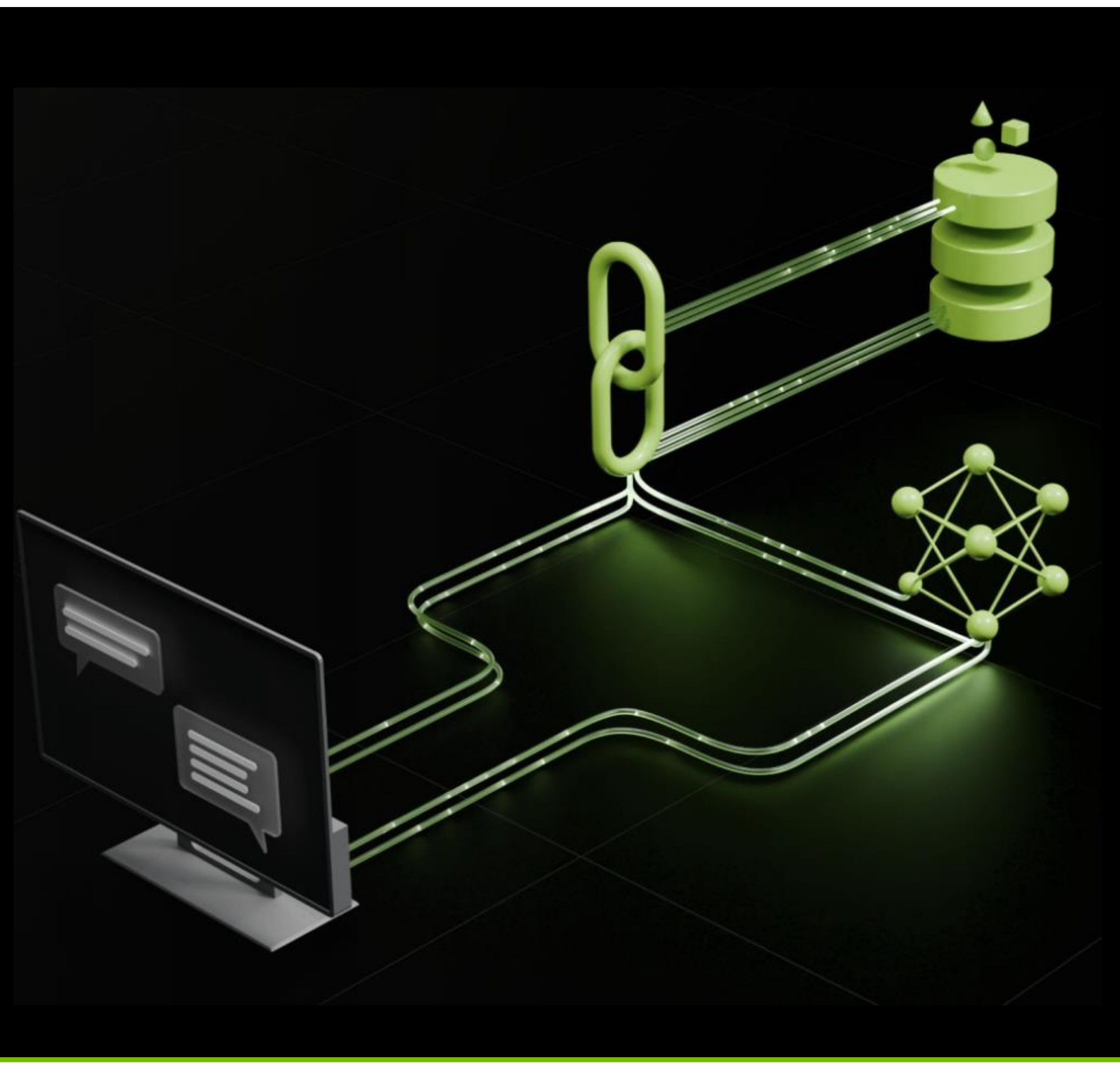
NVIDIA Networking

# Getting Started With NVIDIA AI Enterprise

Experience for free on NVIDIA-hosted infrastructure or deploy on your own infrastructure

## NVIDIA API Catalog

- Try out sample applications from your web browser
- Free API access to experiment and prototype with NVIDIA-optimized microservices



## Developer Access

NVIDIA Developer Program members can

- Download any NIM containers
- Self-host for research and testing (up to 16 GPUs)

## 90-Day Evaluation

- Free evaluation licenses for POC
- Running on compatible on-prem or cloud accelerated infrastructure
- Access to exclusive features

NVIDIA®