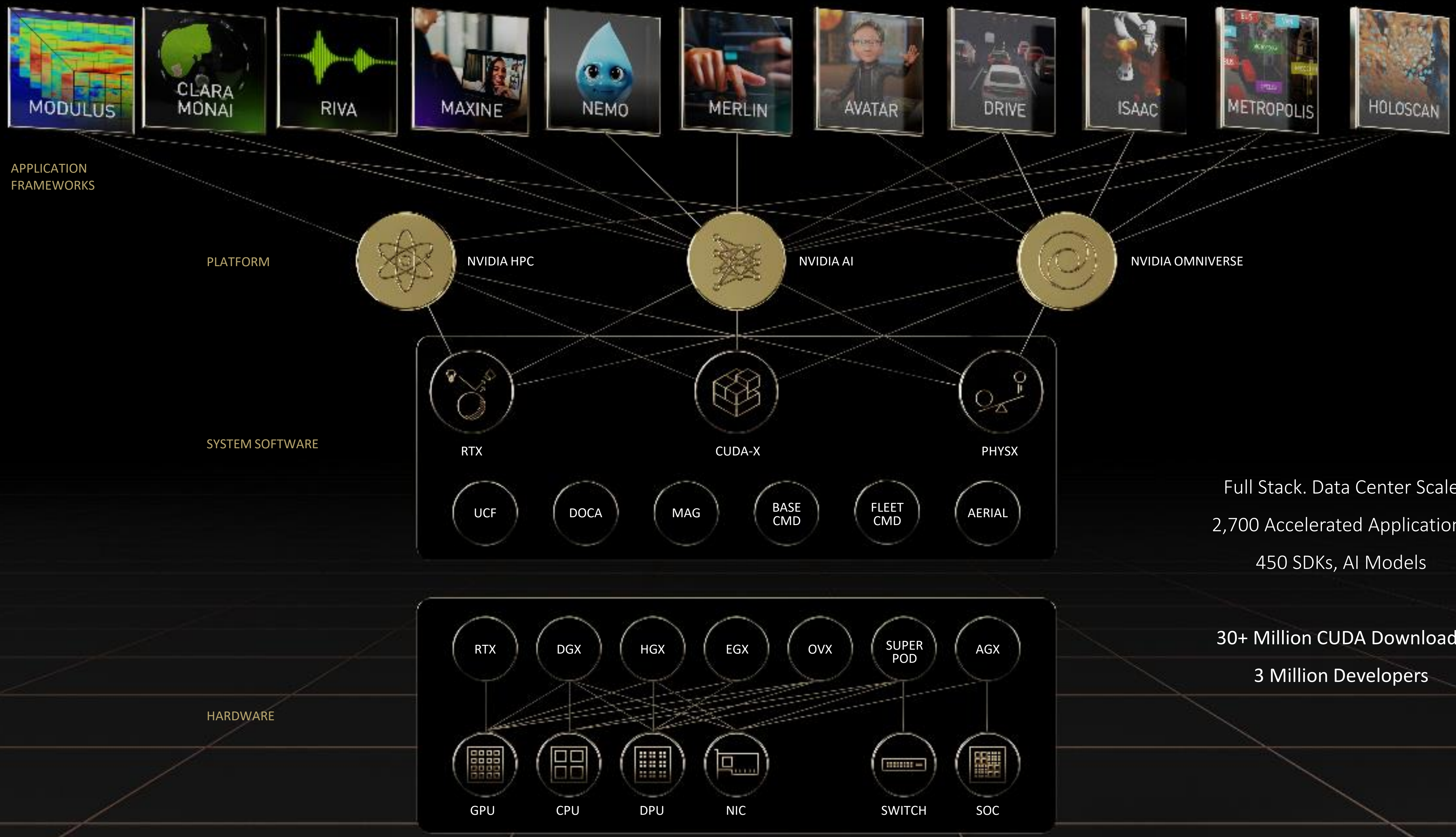
NVIDIA UPDATES FOR IUCC 6-2022

YANIV BENAMI
HIGHER EDUCATION SALES LEADER



Full Stack. Data Center Scale

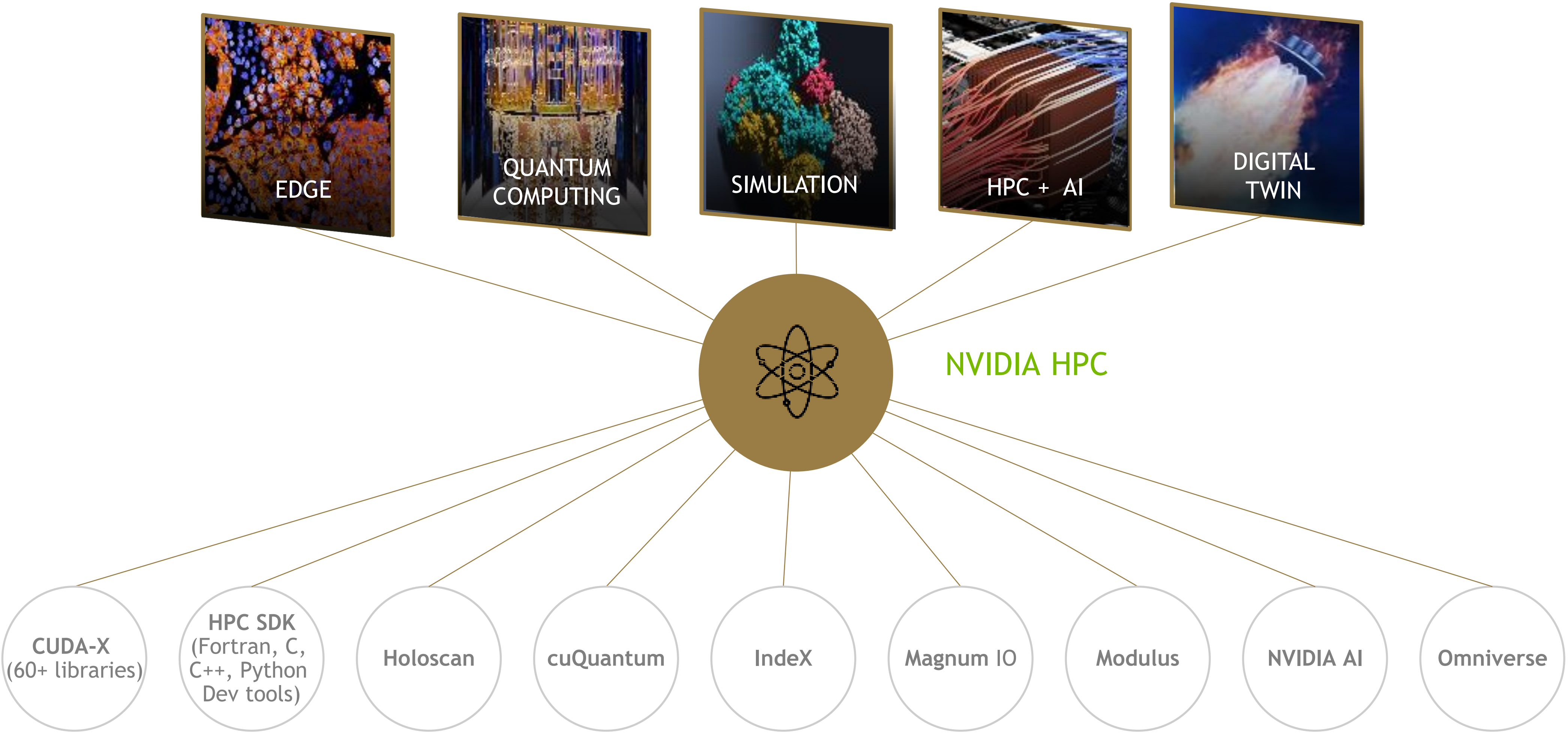
2,700 Accelerated Applications

450 SDKs, AI Models

30+ Million CUDA Downloads

3 Million Developers

NVIDIA'S HPC PLATFORM



EDGE

QUANTUM COMPUTING

SIMULATION

HPC + AI

DIGITAL TWIN

NVIDIA HPC

CUDA-X
(60+ libraries)

HPC SDK
(Fortran, C, C++, Python Dev tools)

Holoscan

cuQuantum

IndeX

Magnum IO

Modulus

NVIDIA AI

Omniverse

ACCELERATING THE WORKLOADS OF THE MODERN SUPERCOMPUTER

WORKLOADS OF THE MODERN SUPERCOMPUTER

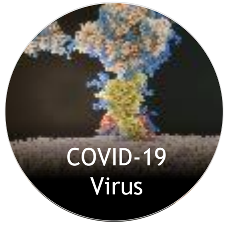
EDGE

Microscopy Image Processing

HPC + AI

Turbine Wake Optimization

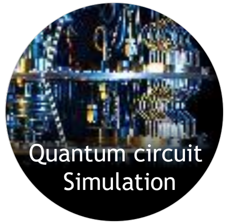
SIMULATION

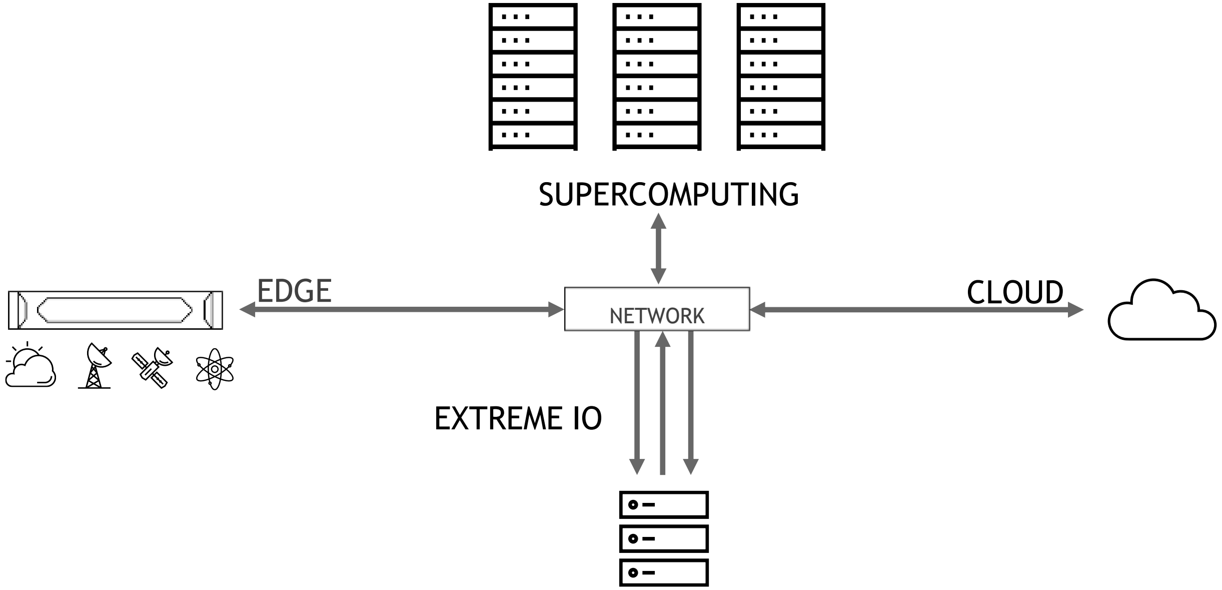

COVID-19 Virus

DIGITAL TWIN

Earth 2

QUANTUM COMPUTING


Quantum circuit Simulation

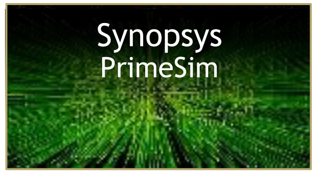


ACCELERATING INDUSTRIAL HPC SIMULATIONS

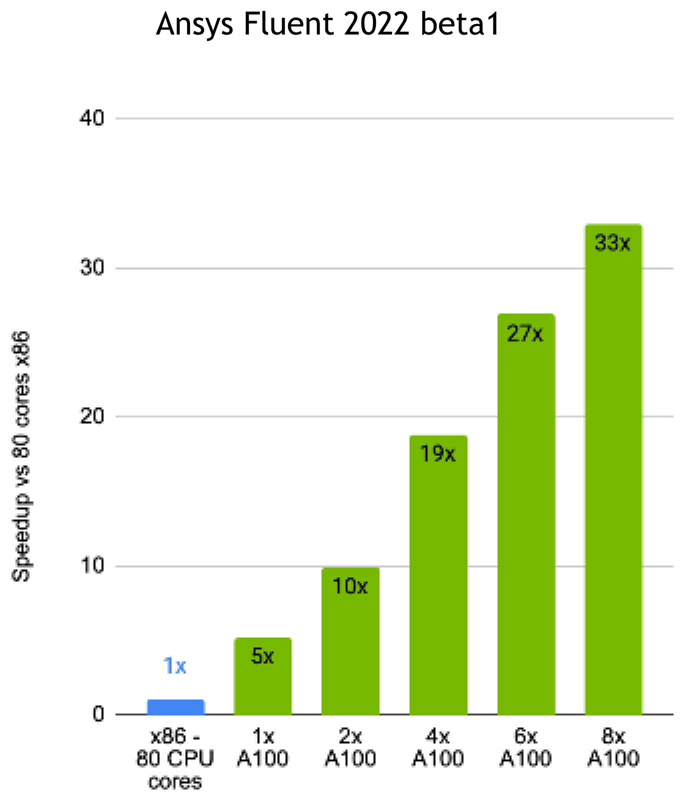
Altair
Altair CFD

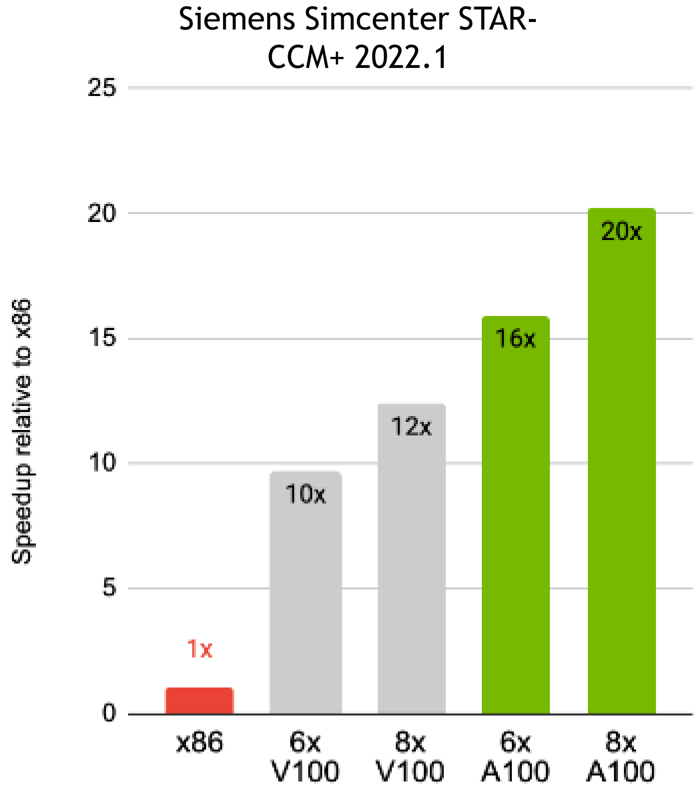
Ansys
Ansys Fluent (beta)

Siemens
Simcenter STAR-CCM+

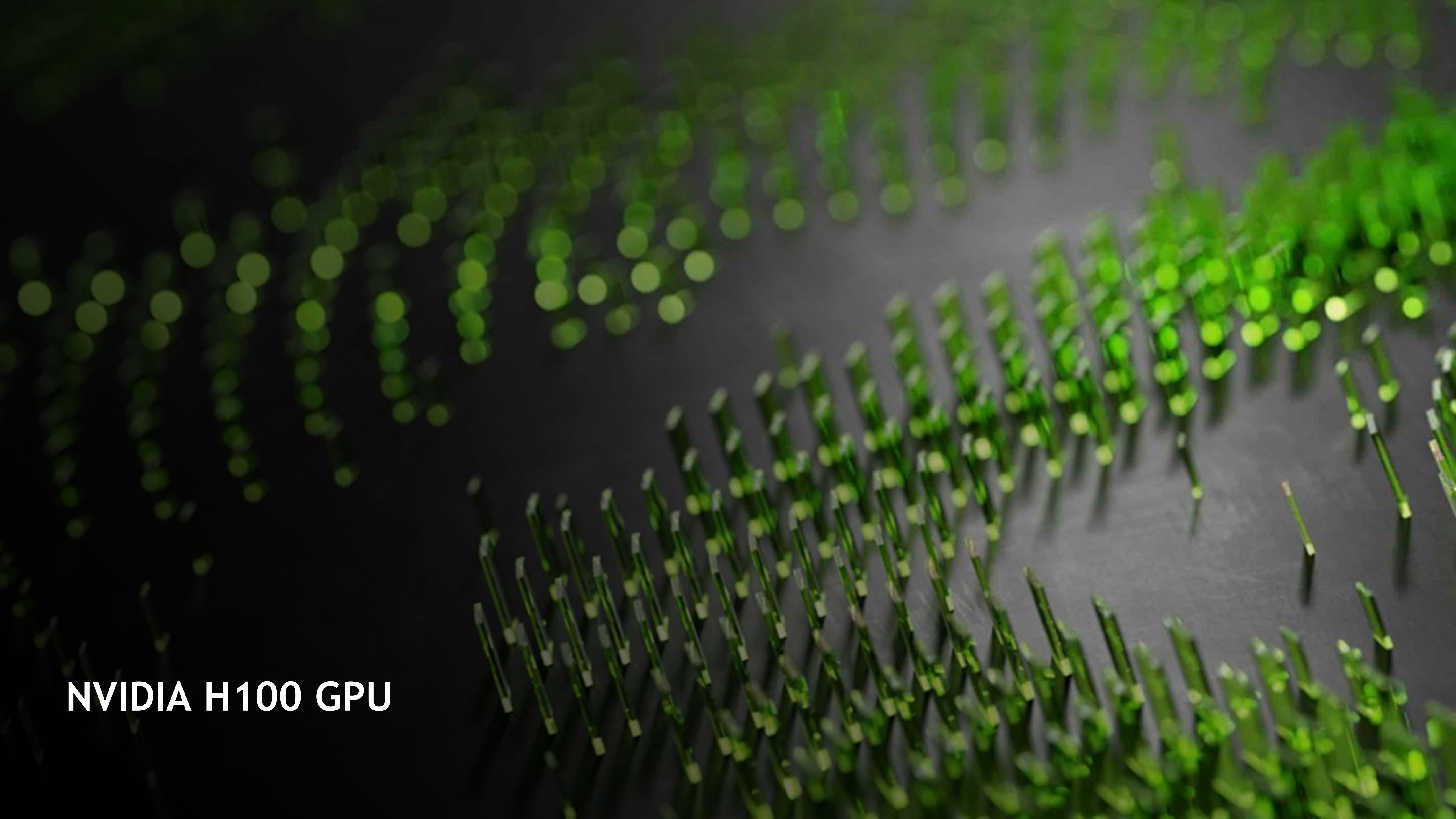

Synopsys
PrimeSim

Altair
Ansys
cādence
CASCADE TECHNOLOGIES
DASSAULT SYSTEMES
HEXAGON
MathWorks
ptc
SIEMENS
SYNOPSYS





Middle Panel: Ansys Fluent 2022 beta 1 running 105M cell car model | CPU baseline Intel Ice Lake 8380 with 80 cores | GPU A100 A100 PCIe 80GB | Right Panel: Siemens Simcenter STAR-CCM+ 2022.1 LeMans 104M model | CPU baseline AMD 7742 Rome |. GPU V100 PCIe 16 GB | A100 A100 PCIe 80GB



NVIDIA H100 GPU

NVIDIA H100

The New Engine for the World's AI Infrastructure

HIGHEST AI AND HPC PERFORMANCE

4PF FP8 (6X)| 2PF FP16 (3X)| 1PF TF32 (3X)| 60TF FP64 (3X)
3TB/s (1.5X), 80GB HBM3 memory

TRANSFORMER MODEL OPTIMIZATIONS

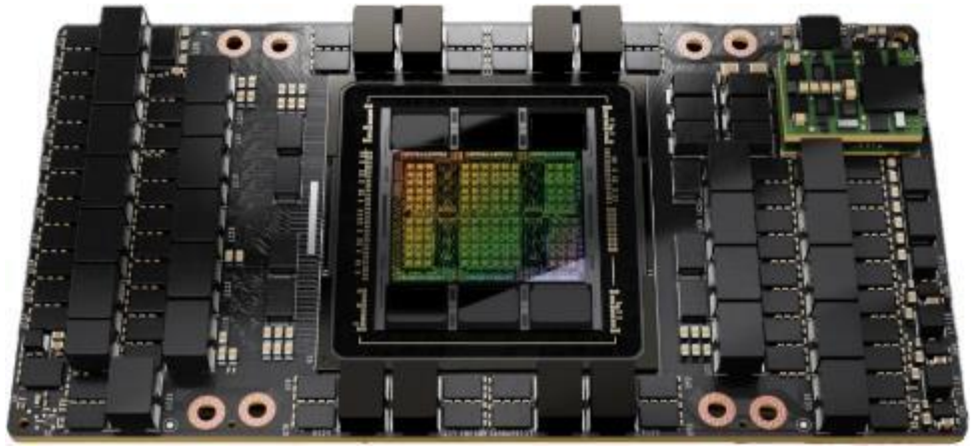
6X faster on largest transformer models

HIGHEST UTILIZATION EFFICIENCY AND SECURITY

7 Fully isolated & secured instances, guaranteed QoS
2nd Gen MIG | Confidential Computing

FASTEST, SCALABLE INTERCONNECT

900 GB/s GPU-2-GPU connectivity (1.5X)
up to 256 GPUs with NVLink Switch | 128GB/s PCI Gen5



Custom 4N TSMC Process | 80 billion transistors

World's Most Advanced Chip
80B Transistors

Transformer Engine
6X Transformer Performance

4th Gen NVLINK
7X PCIe Gen 5

Confidential Computing
Secure Data and AI Models in Use

2nd Gen MIG
7X Secure Tenants

DPX Instructions
7X Dynamic Prog Performance

DELIVERING THE AI CENTER OF EXCELLENCE FOR ENTERPRISE

Best of Breed Infrastructure for AI Development Built on NVIDIA DGX

NVIDIA DGX H100

The World's First AI System with NVIDIA H100

8x NVIDIA H100 | 32 PFLOPS FP8 (6X) | 0.5 PFLOPS FP64 (3X)
640 GB HBM3 | 3.6 TB/s (1.5X) BISECTION B/W

4th Generation of the World's Most Successful Platform Purpose-Built for Enterprise AI

COMING LATE 2022

DGX SuperPOD WITH DGX H100

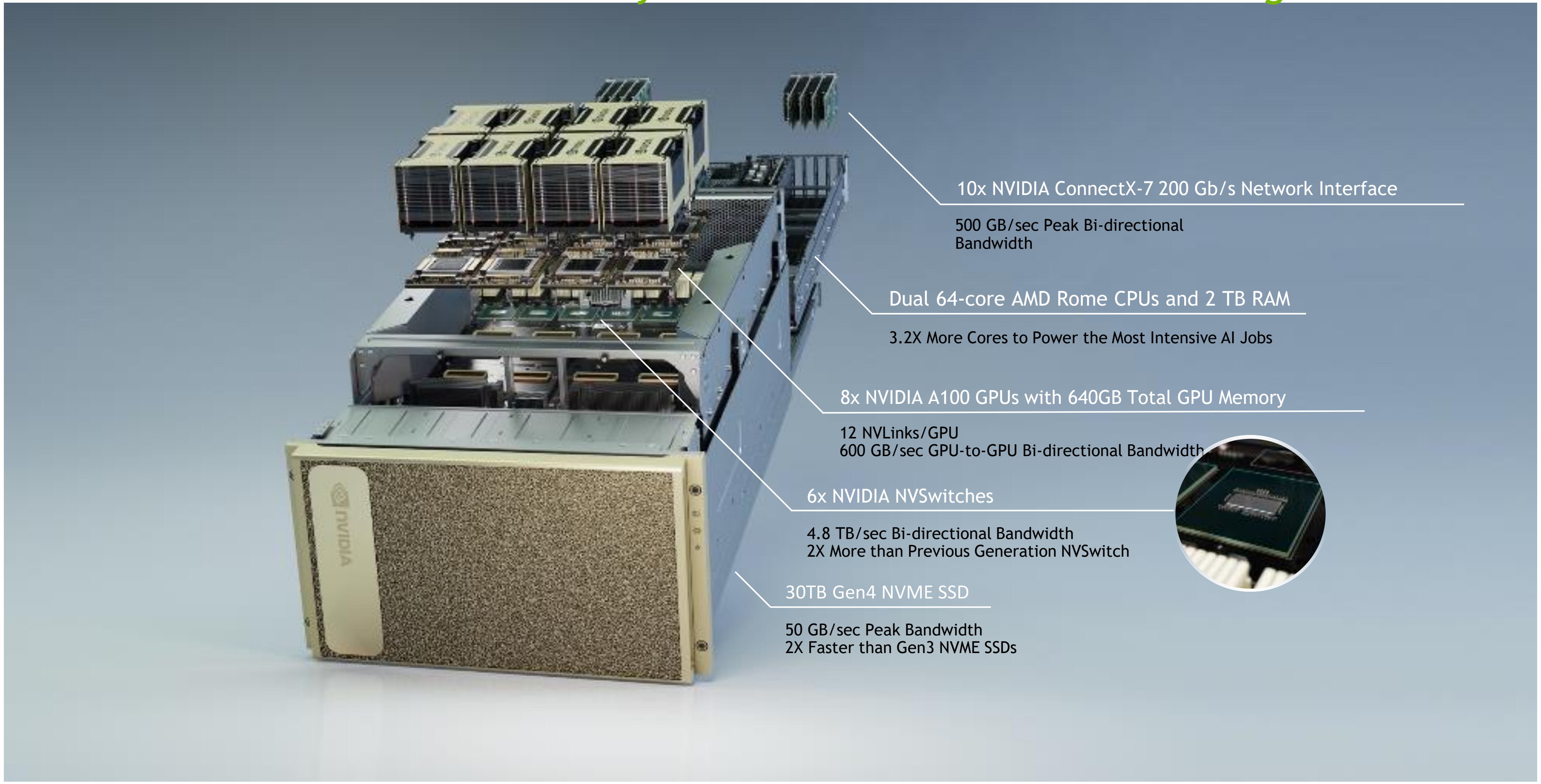
32 DGX H100 | 1 EFLOPS AI
NVLINK SWITCH SYSTEM | QUANTUM-2 IB | 20TB HBM3 |
70 TB/s BISECTION B/W (11X)

1 ExaFLOPS of AI Performance in 32 Nodes
Scale as large as needed in 32 node increments

X-Factors compare performance over DGX SuperPOD with DGX A100 supercomputer configuration with same number of nodes

GAME-CHANGING PERFORMANCE FOR INNOVATORS

NVIDIA DGX A100 640GB System – New annocuements coming soon!



10x NVIDIA ConnectX-7 200 Gb/s Network Interface

500 GB/sec Peak Bi-directional
Bandwidth

Dual 64-core AMD Rome CPUs and 2 TB RAM

3.2X More Cores to Power the Most Intensive AI Jobs

8x NVIDIA A100 GPUs with 640GB Total GPU Memory

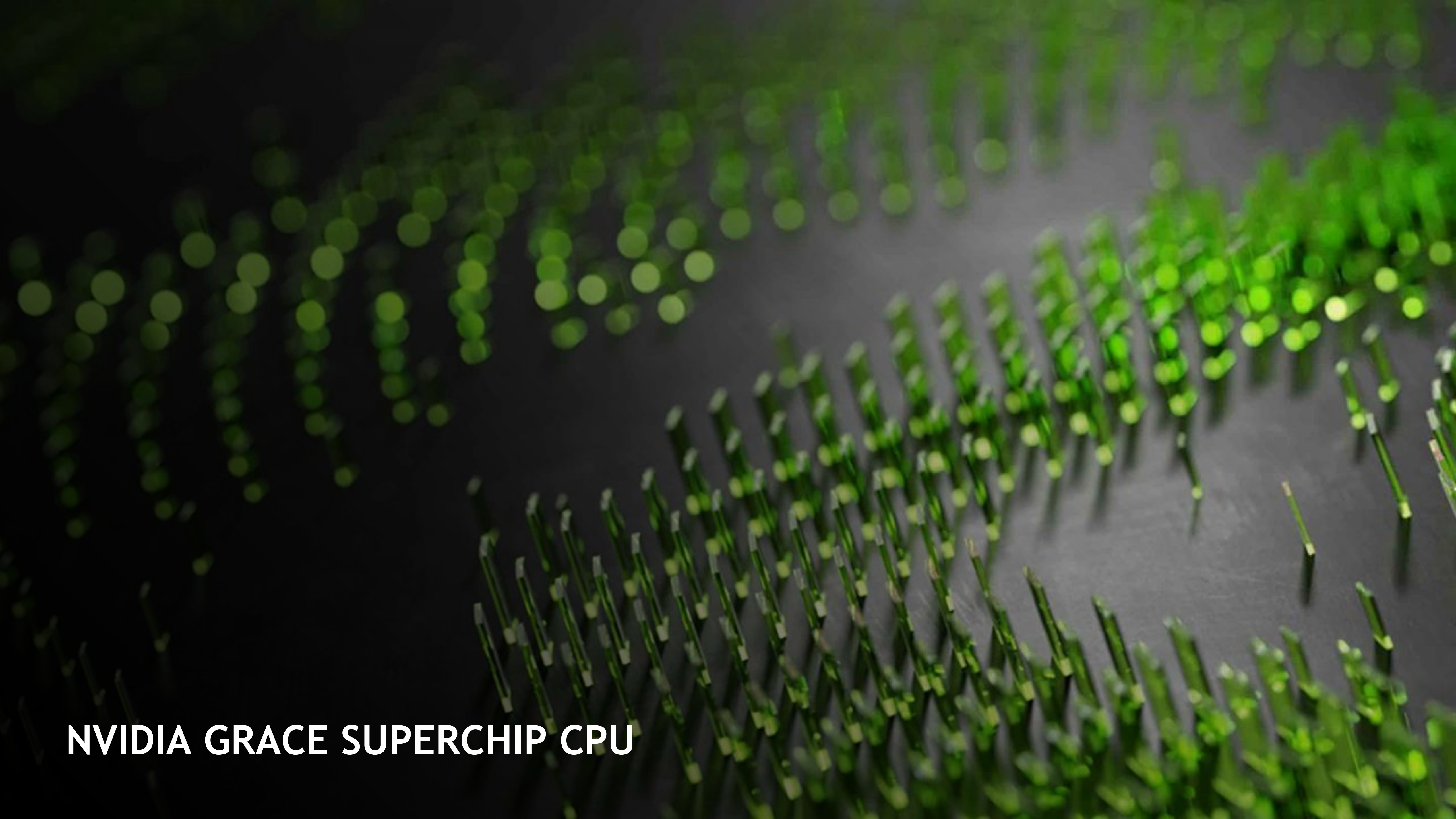
12 NVLinks/GPU
600 GB/sec GPU-to-GPU Bi-directional Bandwidth

6x NVIDIA NVSwitches

4.8 TB/sec Bi-directional Bandwidth
2X More than Previous Generation NVSwitch

30TB Gen4 NVME SSD

50 GB/sec Peak Bandwidth
2X Faster than Gen3 NVME SSDs



NVIDIA GRACE SUPERCHIP CPU

GRACE HOPPER SUPERCHIP

Built for Giant Scale AI and HPC

HIGHEST ACCELERATED PERFORMANCE

Grace CPU plus Hopper GPU Acceleration

~600GB MEMORY AVAILABLE TO GPU

Enables Giant AI Models for Training & Inference

HIGHEST MEMORY BANDWIDTH 3.5GB/s

LPDDR5x and HBM3

NEW 900GB/S COHERENT INTERFACE

NVLink-C2C connecting Grace to Hopper

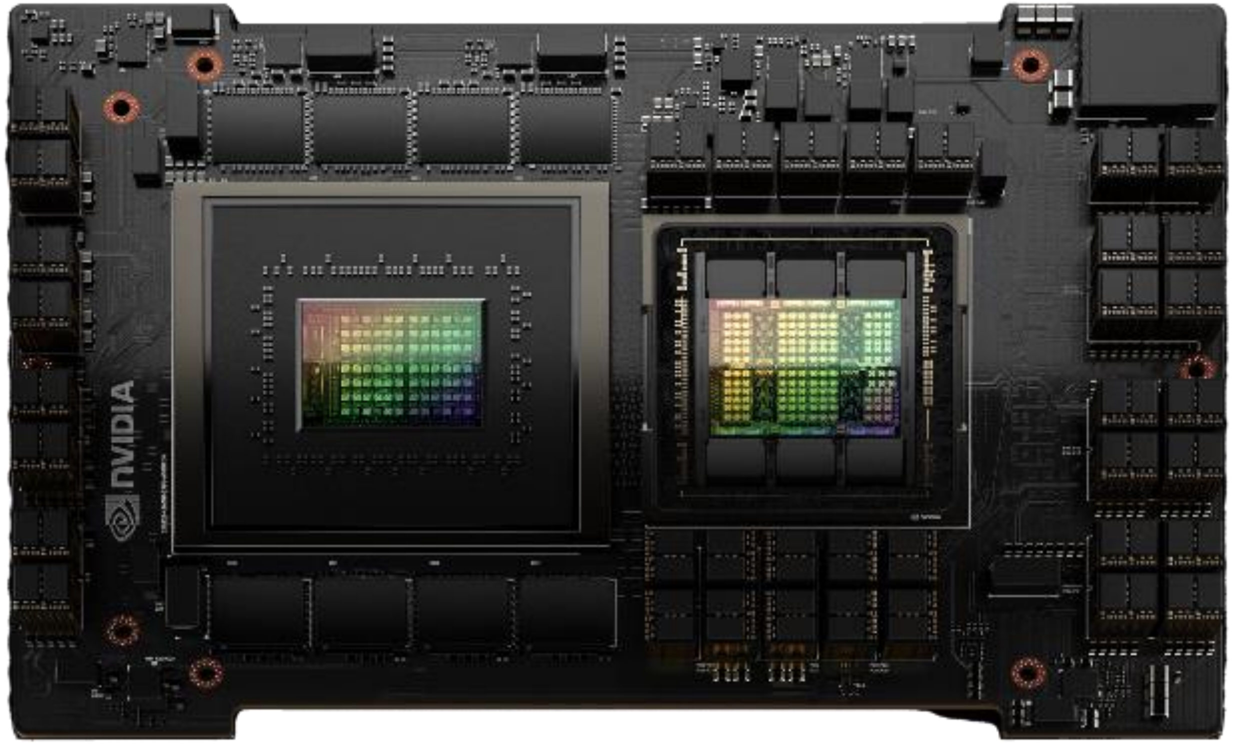
15X HIGHER SYSTEM MEMORY BANDWIDTH TO GPU

NVLink-C2C vs PCIe

RUNS FULL NVIDIA COMPUTING STACKS

RTX, HPC, AI, Omniverse

AVAILABLE 1H 2023



FP8, FP16, TF32 performance include sparsity. X-factor compared to A100

GRACE CPU SUPERCHIP

The CPU for AI and HPC Infrastructure

HIGHEST CPU PERFORMANCE

Superchip Design with 144 high-performance Armv9 Cores
Estimated Specrate2017_int_base of over 740

HIGHEST MEMORY BANDWIDTH

World’s first LPDDR5x memory with ECC, 1TB/s Memory Bandwidth

HIGHEST ENERGY EFFICIENCY

2X Perf/Watt, CPU Cores + Memory in 500W

2X PACKING DENSITY

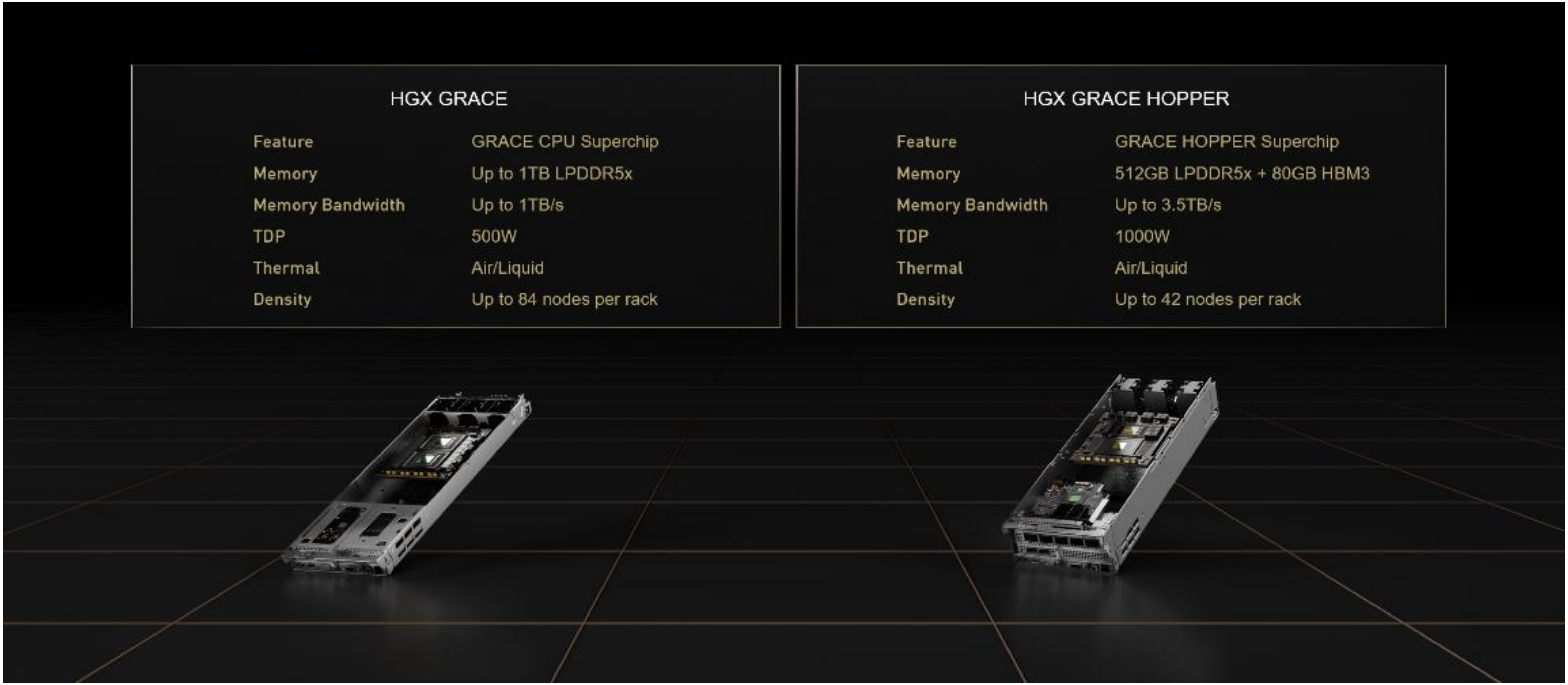
2x density of DIMM based designs

RUNS FULL NVIDIA COMPUTING STACKS

RTX, HPC, AI, Omniverse

AVAILABLE 1H 2023

2U HIGH DENSITY SERVER REFERENCE DESIGNS FOR RAPID ADOPTION



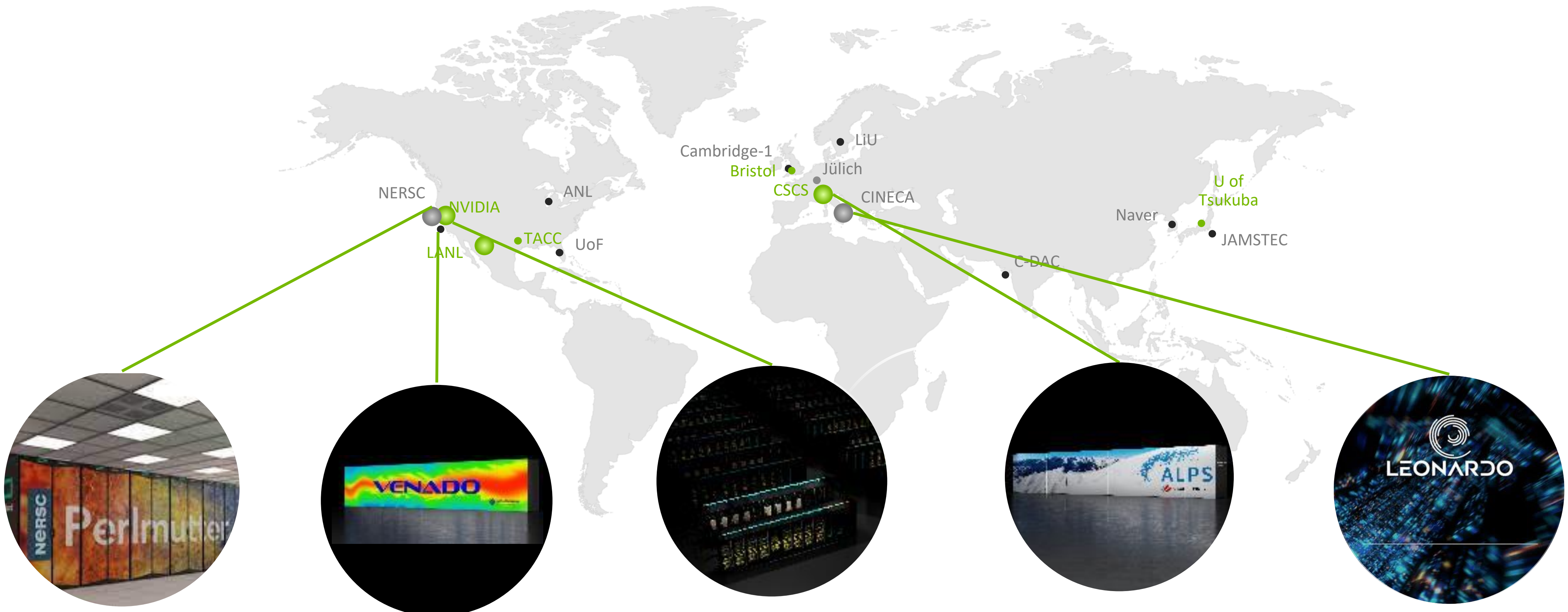
HGX GRACE

Feature	GRACE CPU Superchip
Memory	Up to 1TB LPDDR5x
Memory Bandwidth	Up to 1TB/s
TDP	500W
Thermal	Air/Liquid
Density	Up to 84 nodes per rack

HGX GRACE HOPPER

Feature	GRACE HOPPER Superchip
Memory	512GB LPDDR5x + 80GB HBM3
Memory Bandwidth	Up to 3.5TB/s
TDP	1000W
Thermal	Air/Liquid
Density	Up to 42 nodes per rack

NVIDIA NEXT-GEN COMPUTING PLATFORM
POWERING THE NEXT WAVE OF AI SUPERCOMPUTERS



NERSC (Perlmutter)
3.8 EFLOPS AI Perf

LANL (Venado)
10 EFLOPS AI Perf

NVIDIA (EOS)
18 EFLOPS AI Perf
Quantum-2 400Gb/s InfiniBand

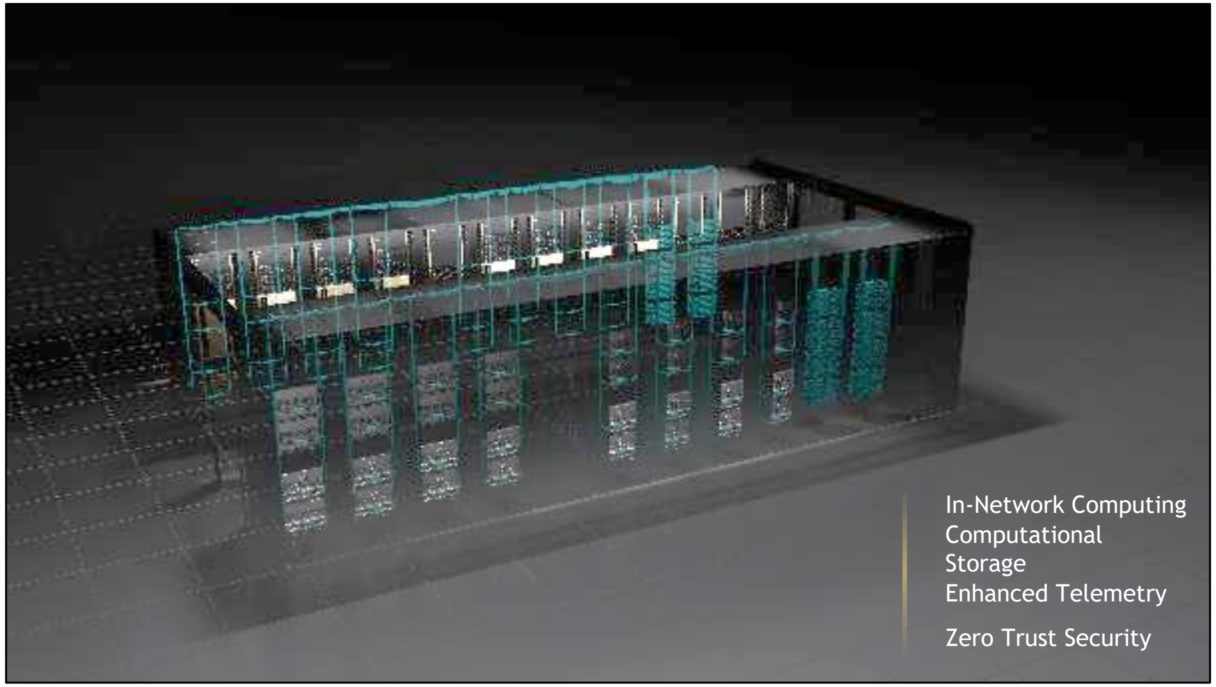
CSCS (ALPS)
20 EFLOPS AI Perf

CINECA (Leonardo)
10 EFLOPS AI Perf
Quantum 200Gb/s InfiniBand

- Hopper + X86 systems: University of Tsukuba, Bristol, and TACC
- Grace Hopper/Grace CPU Superchips systems: CSCS and LANL

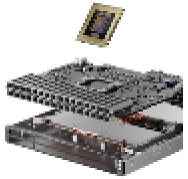
NVIDIA NETWORKING

CLOUD NATIVE SUPERCOMPUTING ENABLED BY NVIDIA QUANTUM-2 INFINIBAND PLATFORM

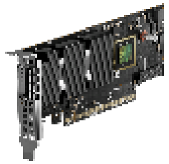


In-Network Computing
Computational
Storage
Enhanced Telemetry

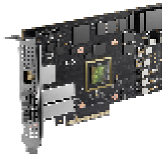
Zero Trust Security



QUANTUM-2 INFINIBAND SWITCH
Cloud Native Supercomputing Platform
SHARP In-Network Computing
Higher Scalability



CONNECTX-7 SMARTNIC
Intelligent Offloads
Precision Timing
Software Defined Networking



BLUEFIELD-3/-X DPU
Intelligent Offloads
Precision Timing
Software Defined Networking

SKYWAY GATEWAY
InfiniBand to Ethernet
Low Latency
Load Balancing

UFM
Monitoring, Management, Orchestration
Predictive Maintenance
Anomaly Detection

ANNOUNCING TACC AND LANL BLUEFIELD INFINIBAND COLLABORATIONS

RESEARCH AND DEVELOPMENT
Application Development Over BlueField/DOCA

30X PERFORMANCE SPEEDUP
Multi-Year Collaboration

NVIDIA DGX FOUNDRY – SUPERCOMPUER AS A SERVICE FOR BURST USAGE

Specifications

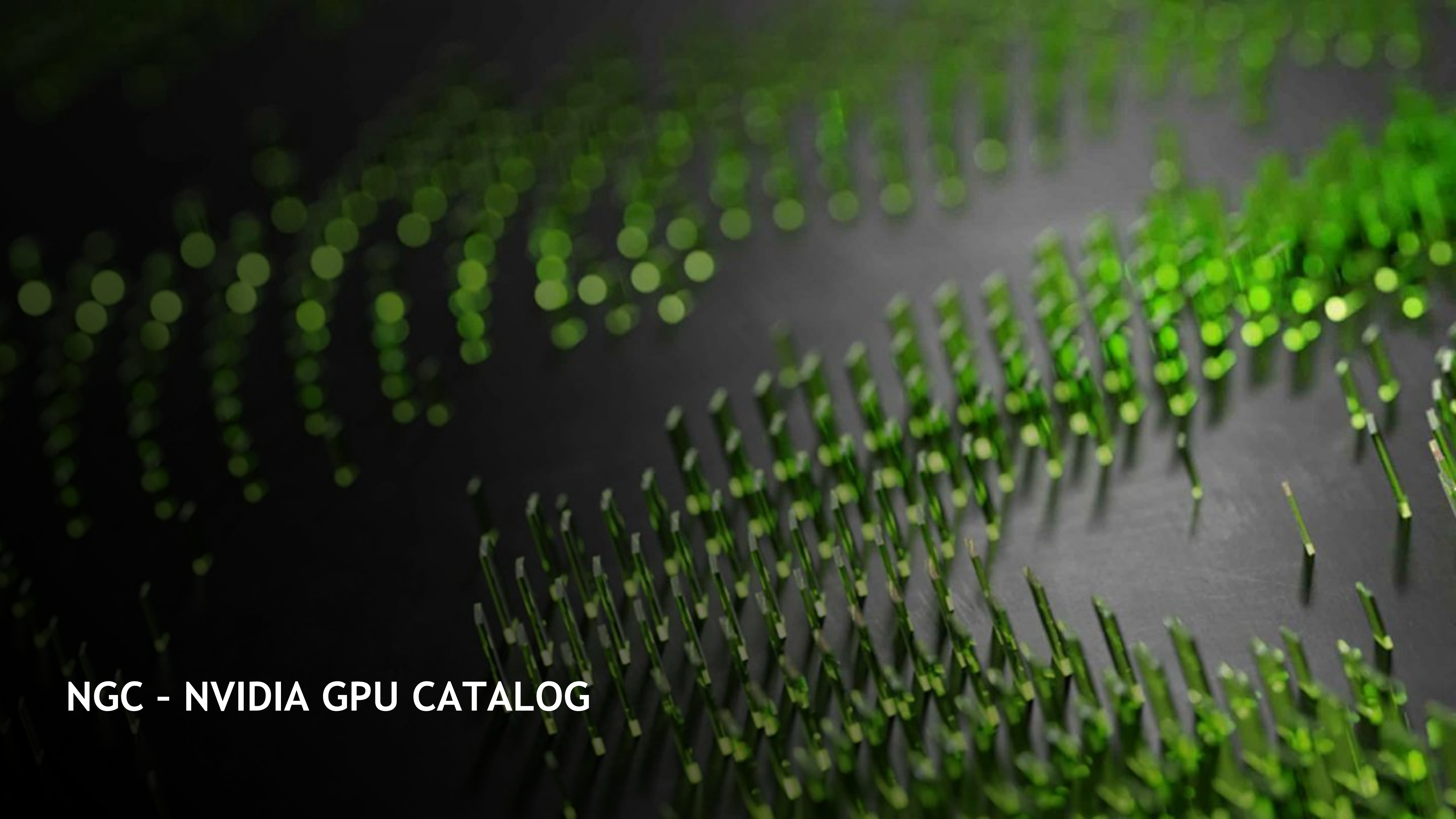
Global Program with regional deployments:

- Today: Silicon Valley, Washington DC area
- Soon: South Korea, Germany, Taiwan

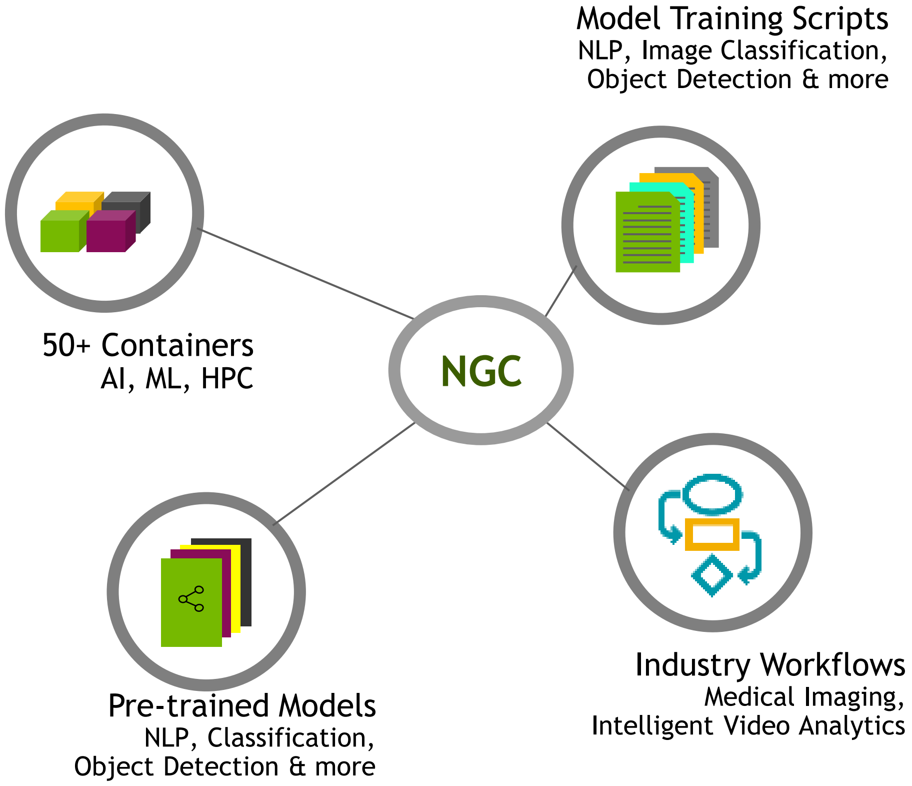
Typical deployment:

- **Compute:**
  NVIDIA DGX SuperPOD (20 or more NVIDIA DGX A100)
- **Networking:**
  Compute nodes: 8x 200Gb/s InfiniBand
  Storage: 100Gb/s Ethernet
  Internet access: 10Gb/s
- **Storage:**
  Dedicated high-availability (HA) pair of NetApp AFF A800 per customer

Refreshed with latest technology once available.



NGC – NVIDIA GPU CATALOG


Model Training Scripts
NLP, Image Classification, Object Detection & more
50+ Containers
AI, ML, HPC
NGC
Pre-trained Models
NLP, Classification, Object Detection & more
Industry Workflows
Medical Imaging, Intelligent Video Analytics

EFFORTLESS PRODUCTIVITY

NVIDIA DGX Software Stack Delivers Immediate Productivity that Saves Time and Money

Save $x00,000's on software engineering of AI frameworks

Depend on NVIDIA-optimized frameworks instead of evolving open source software

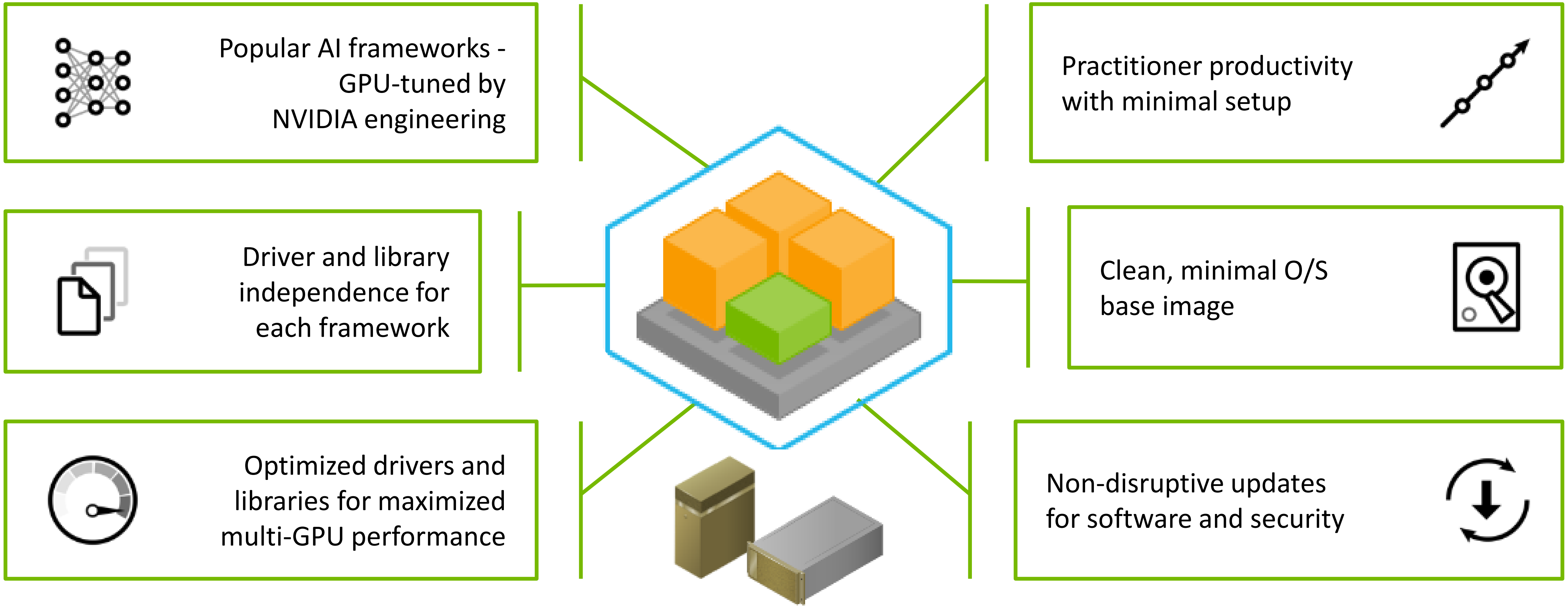
Save $100k+/yr in admin OpEx with cloud management, streamlined collaboration

Monthly framework releases ensure maximized performance for AI ROI

DGX private registry for powerful sharing and collaboration

ENTERPRISE BENEFITS OF DGX SOFTWARE

NVIDIA Investments in Deep Learning Performance and Manageability



Popular AI frameworks - GPU-tuned by NVIDIA engineering

Driver and library independence for each framework

Optimized drivers and libraries for maximized multi-GPU performance

Practitioner productivity with minimal setup

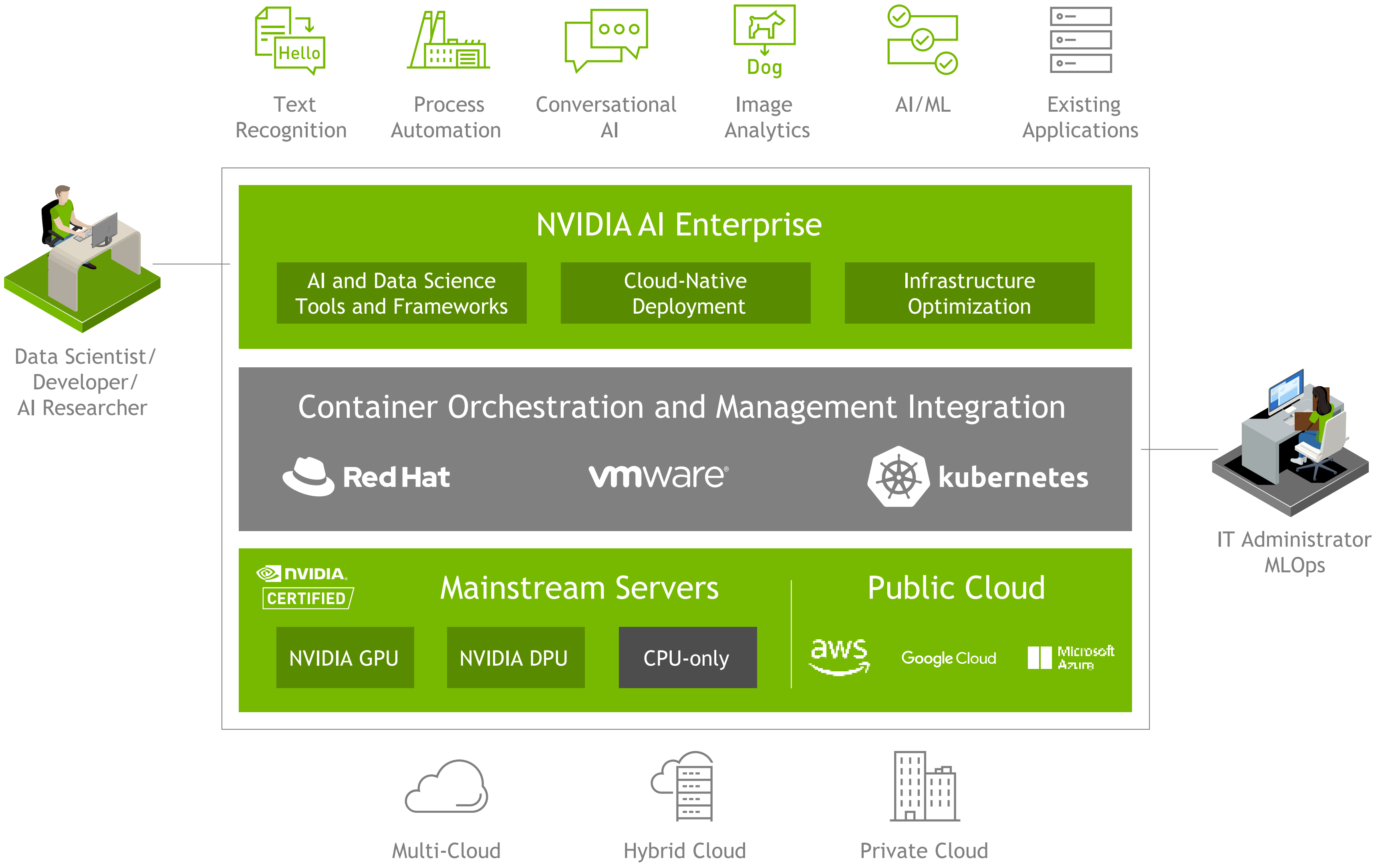
Clean, minimal O/S base image

Non-disruptive updates for software and security

NVAIE – NVIDIA AI ENTERPRISE

AI-READY ENTERPRISE PLATFORM

Enterprise AI for Everyone, Everywhere, on Every Platform



Text Recognition

Process Automation

Conversational AI

Image Analytics

AI/ML

Existing Applications

Data Scientist/ Developer/ AI Researcher

NVIDIA AI Enterprise

- AI and Data Science Tools and Frameworks
- Cloud-Native Deployment
- Infrastructure Optimization

Container Orchestration and Management Integration

Red Hat · vmware · kubernetes

IT Administrator MLOps

NVIDIA CERTIFIED

Mainstream Servers

- NVIDIA GPU
- NVIDIA DPU
- CPU-only

Public Cloud

aws · Google Cloud · Microsoft Azure

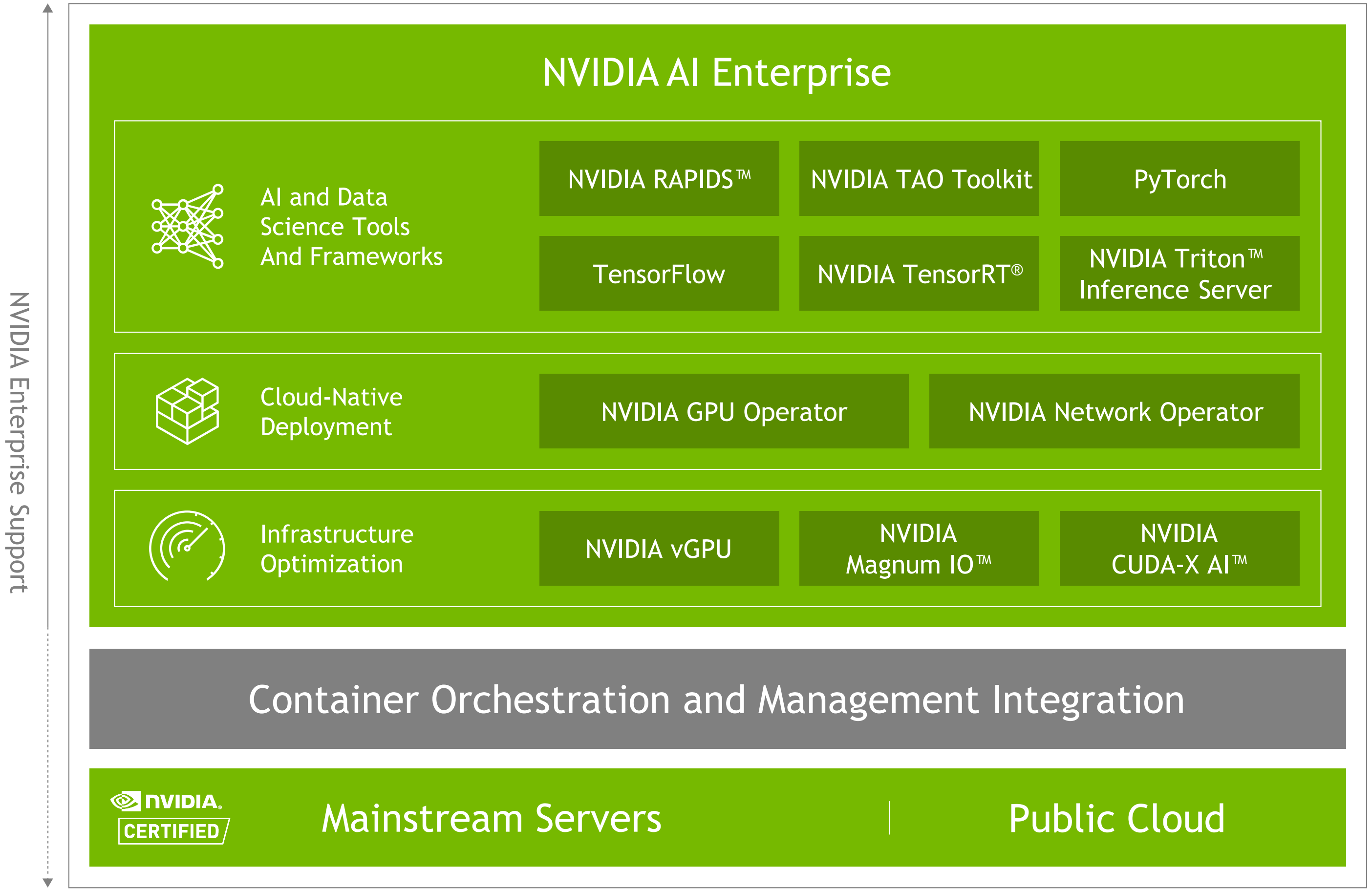
Multi-Cloud

Hybrid Cloud

Private Cloud

NVIDIA AI ENTERPRISE SOFTWARE SUITE

Enterprise AI for Everyone, Everywhere, on Every Platform



NVIDIA AI Enterprise

AI and Data Science Tools And Frameworks

NVIDIA RAPIDS™

NVIDIA TAO Toolkit

PyTorch

TensorFlow

NVIDIA TensorRT®

NVIDIA Triton™ Inference Server

Cloud-Native Deployment

NVIDIA GPU Operator

NVIDIA Network Operator

Infrastructure Optimization

NVIDIA vGPU

NVIDIA Magnum IO™

NVIDIA CUDA-X AI™

Container Orchestration and Management Integration

NVIDIA CERTIFIED

Mainstream Servers

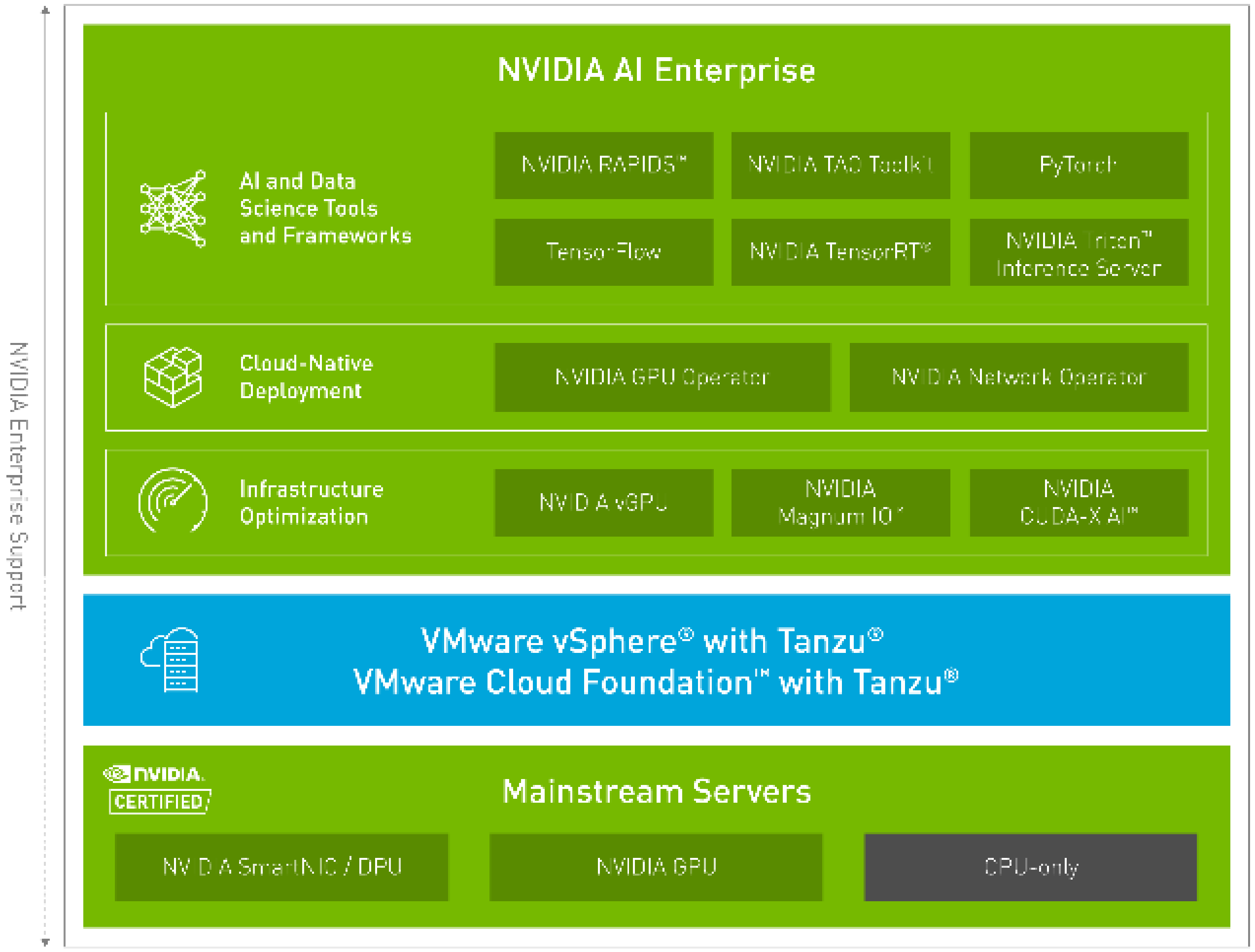
Public Cloud

NVIDIA Enterprise Support

https://www.nvidia.com/ai-enterprise-suite/

NVIDIA AI ENTERPRISE SOFTWARE SUITE

Enabling AI and Data Analytics on VMware vSphere and VMware Cloud Foundation



NVIDIA Enterprise Support

NVIDIA AI Enterprise

AI and Data Science Tools and Frameworks

NVIDIA RAPIDS™

NVIDIA TAO Toolkit

PyTorch

TensorFlow

NVIDIA TensorRT®

NVIDIA Triton™ Inference Server

Cloud-Native Deployment

NVIDIA GPU Operator

NVIDIA Network Operator

Infrastructure Optimization

NVIDIA vGPU

NVIDIA Magnum IO™

NVIDIA CUDA-X AI™

VMware vSphere® with Tanzu®
VMware Cloud Foundation™ with Tanzu®

NVIDIA CERTIFIED

Mainstream Servers

NVIDIA SmartNIC / DPU

NVIDIA GPU

CPU-only

Optimized for Performance

Comparable bare-metal performance across multiple nodes to power large, complex training and machine learning workloads virtualized

Certified for VMware vSphere

Reduce deployment risks with a complete suite of NVIDIA AI software certified for the VMware data center

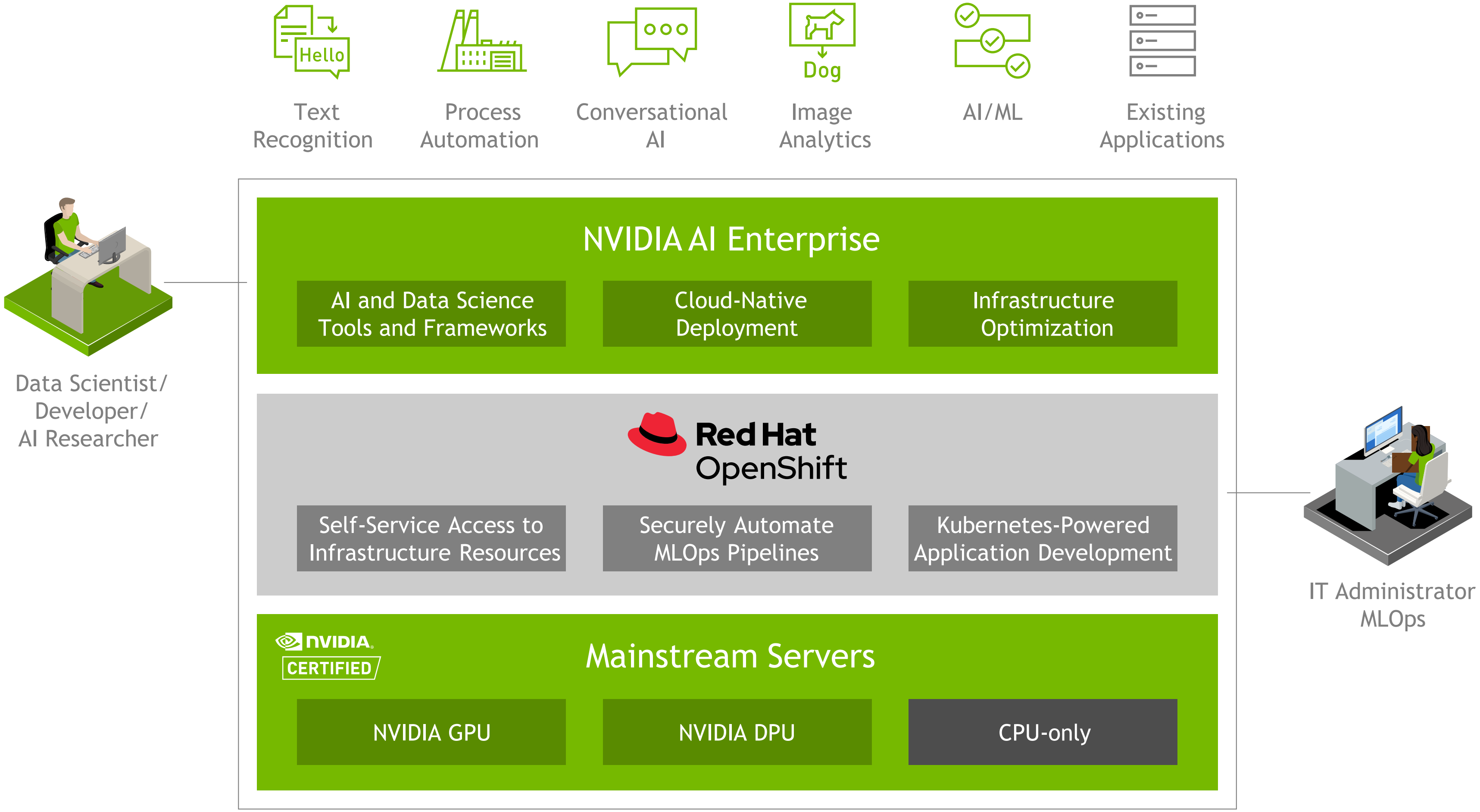
NVIDIA Enterprise Support

Ensure mission-critical AI projects stay on track with access to NVIDIA experts

*TensorFlow and PyTorch are integrated in the NVIDIA RAPIDS containers in NVIDIA AI Enterprise 1.1 and later

https://www.nvidia.com/ai-enterprise-suite/

NVIDIA AI ENTERPRISE WITH RED HAT OPENSHIFT



Text Recognition

Process Automation

Conversational AI

Image Analytics

AI/ML

Existing Applications

Data Scientist/ Developer/ AI Researcher

NVIDIA AI Enterprise

AI and Data Science Tools and Frameworks

Cloud-Native Deployment

Infrastructure Optimization

Red Hat OpenShift

Self-Service Access to Infrastructure Resources

Securely Automate MLOps Pipelines

Kubernetes-Powered Application Development

NVIDIA CERTIFIED

Mainstream Servers

NVIDIA GPU

NVIDIA DPU

CPU-only

IT Administrator MLOps

NVIDIA AI ENTERPRISE SUPPORT AND TRAINING

Open-Source Transparency with Assurance of Enterprise Grade Support

Extend Your Team

Access to NVIDIA AI Experts

8-5 local business hours
Guidance on configuration + performance
Access to engineering

Stay Up to Date

Priority Notifications

Latest security fixes, maintenance releases, coordinated support across partners

Control Upgrade and Maintenance Schedule

Long Term Support

Up to 3 years for designated SW branches

Customized Support

Mission Critical*

Designated Technical Account Manager
Business critical 24/7 live agent access

Upskill Your Workforce

Enterprise Training Services

Instructor-led workshops and self-paced trainings

* Available as upgrade options.

GETTING STARTED WITH NVIDIA AI

NVIDIA AI Enterprise Trial Programs

Test Drive Demo

- Self-directed, remote access demo
  - Predicting NYC Taxi Fares with RAPIDS
  - BERT Question Answer in TensorFlow
- Requires ~1 hr / Access for 48 hrs

NVIDIA LaunchPad

- AI development and deployment trial program
- Deep dive, hands-on labs for AI practitioners and IT staff
- Requires ~8 hrs / Access for 2 wks

Evaluation Software

- Requirements: NVIDIA-Certified System
- Free evaluation licenses for on premises POC
- 90 days to test and experience

Q&A

YANIV BENAMI
YBENAMI@NVIDIA.COM
054-5225769