NVIDIA

AI Acceleration with NVIDIA

Dr. Arts Yang, Sr. Product Architect DGX & Cloud

arts@nvidia.com

AI DAYS in Prague

25. Jan. 2024

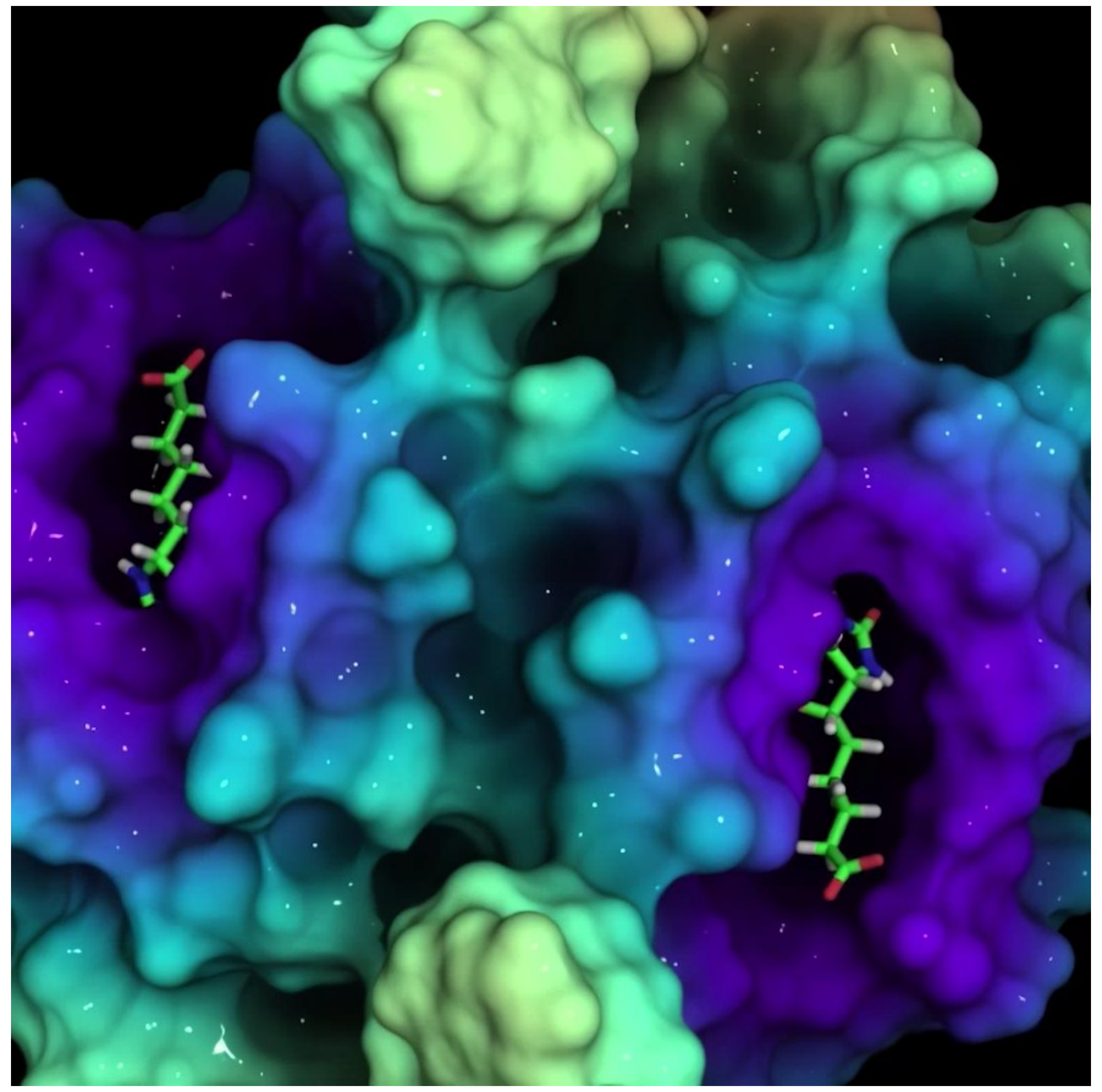
The In-Person GTC Experience Is Back

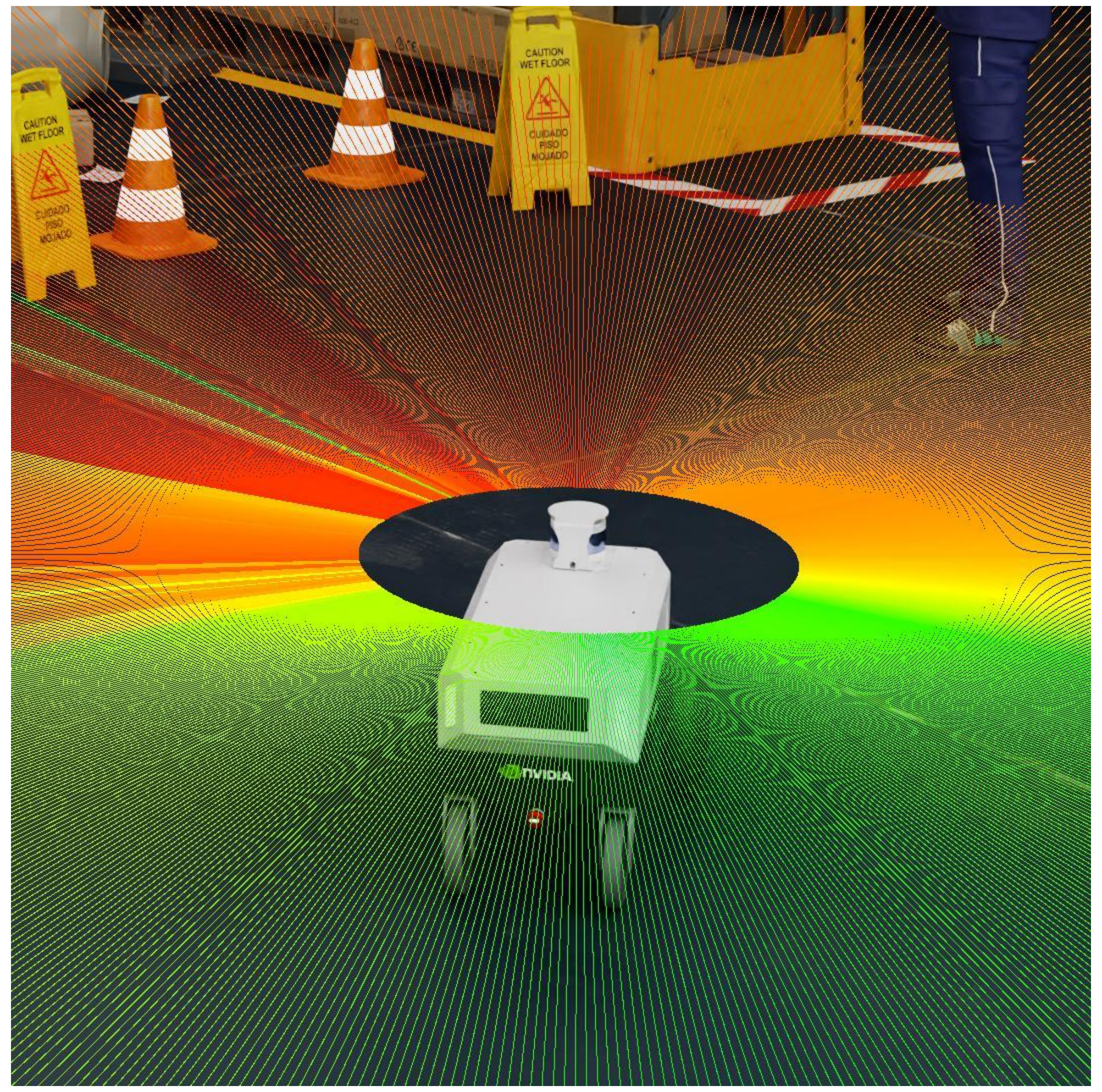
Come to GTC—the conference for the era of AI—to connect with a dream team of industry luminaries, developers, researchers, and business experts shaping what’s next in AI and accelerated computing.

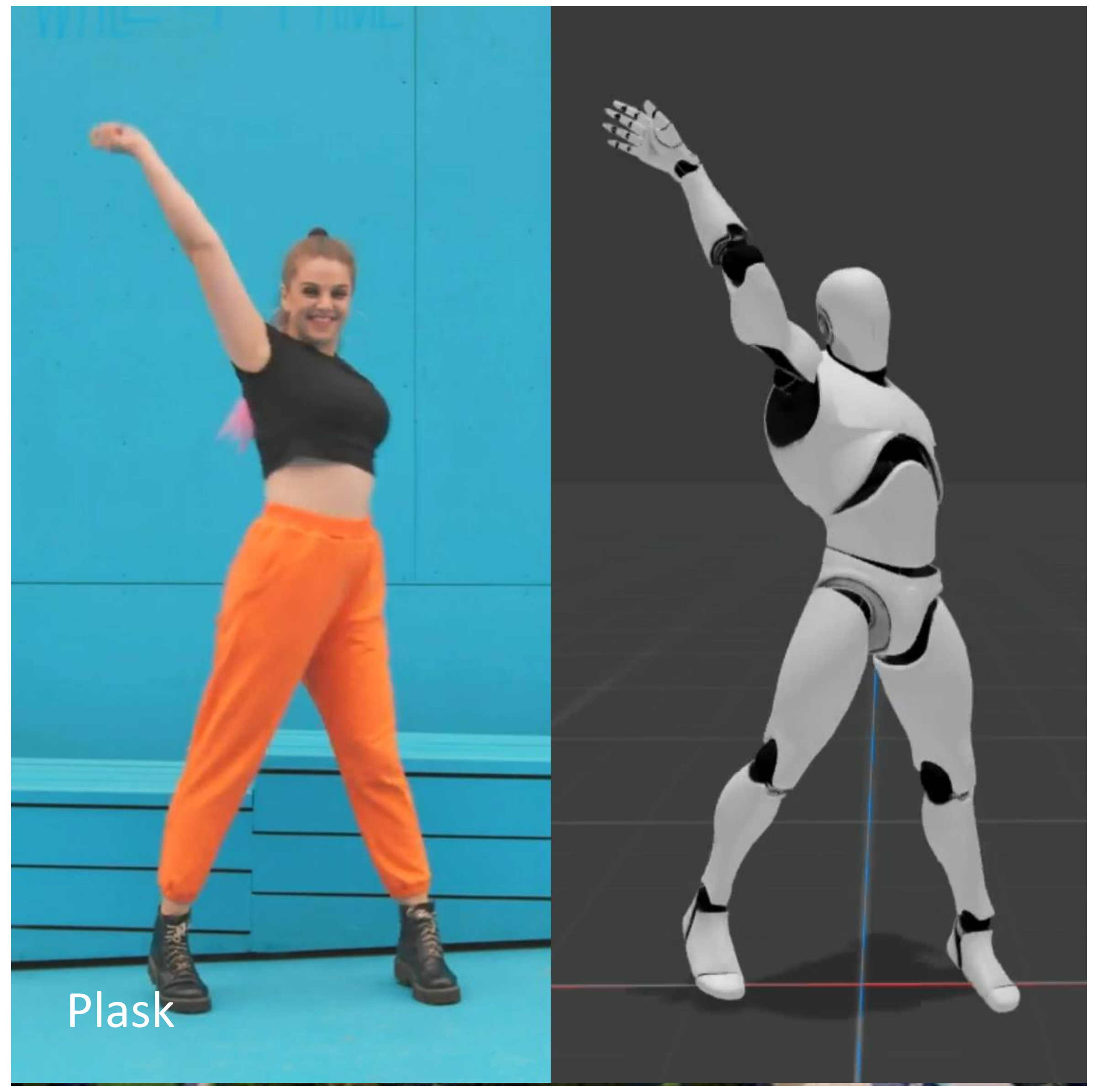
From the highly anticipated keynote by NVIDIA CEO Jensen Huang to over 600 inspiring sessions, 200+ exhibits, and tons of networking events, GTC delivers something for every technical level and interest area.

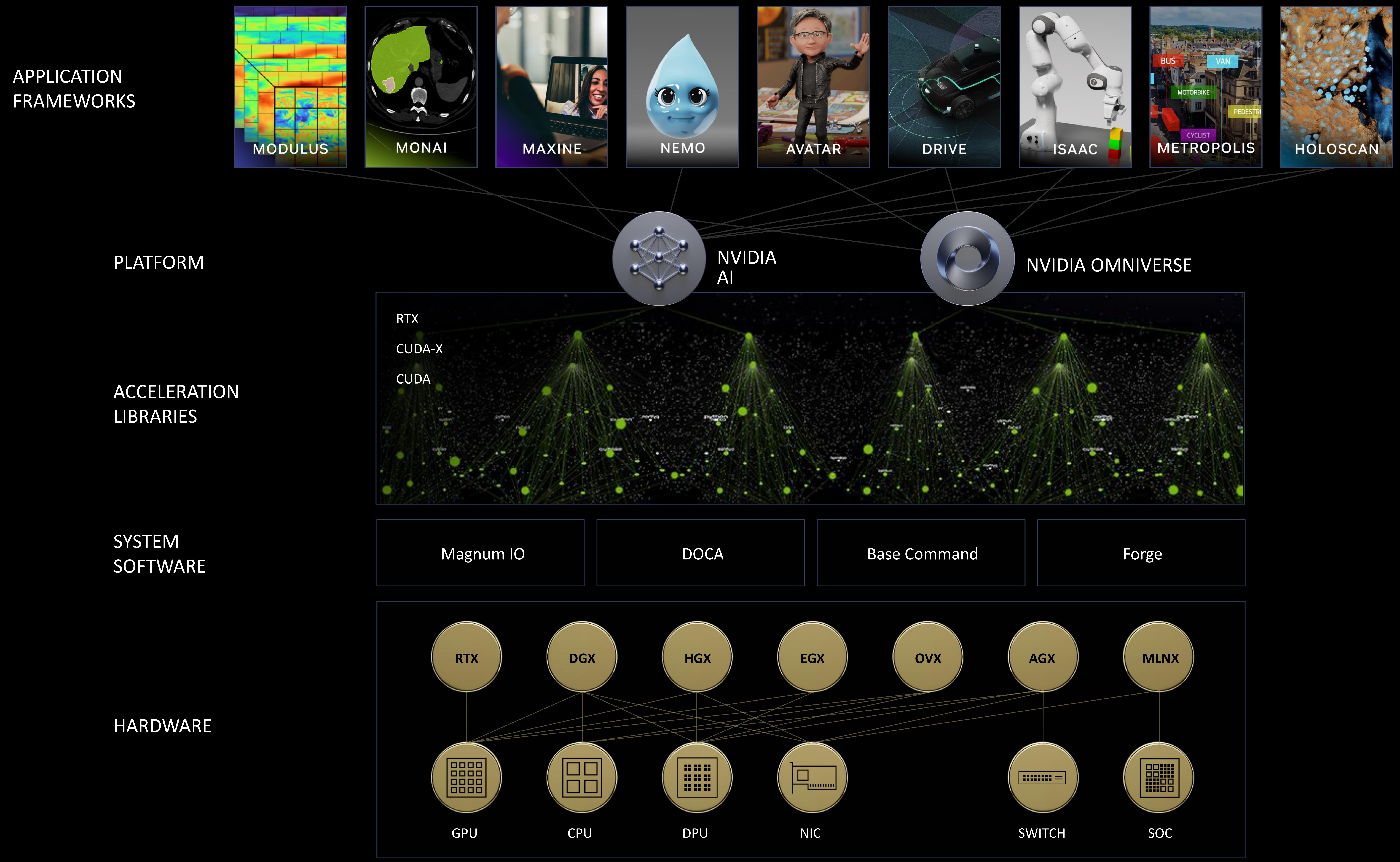
Be sure to save your spot for this transformative event. You can even take advantage of early-bird pricing when you register by February 7.

March 18-21, 2024 | www.nvidia.com/gtc









APPLICATION FRAMEWORKS

MODULUS
MONAI
MAXINE
NEMO
AVATAR
DRIVE
ISAAC
METROPOLIS
HOLOSCAN

PLATFORM

NVIDIA AI
NVIDIA OMNIVERSE

ACCELERATION LIBRARIES

RTX
CUDA-X
CUDA

SYSTEM SOFTWARE

Magnum IO
DOCA
Base Command
Forge

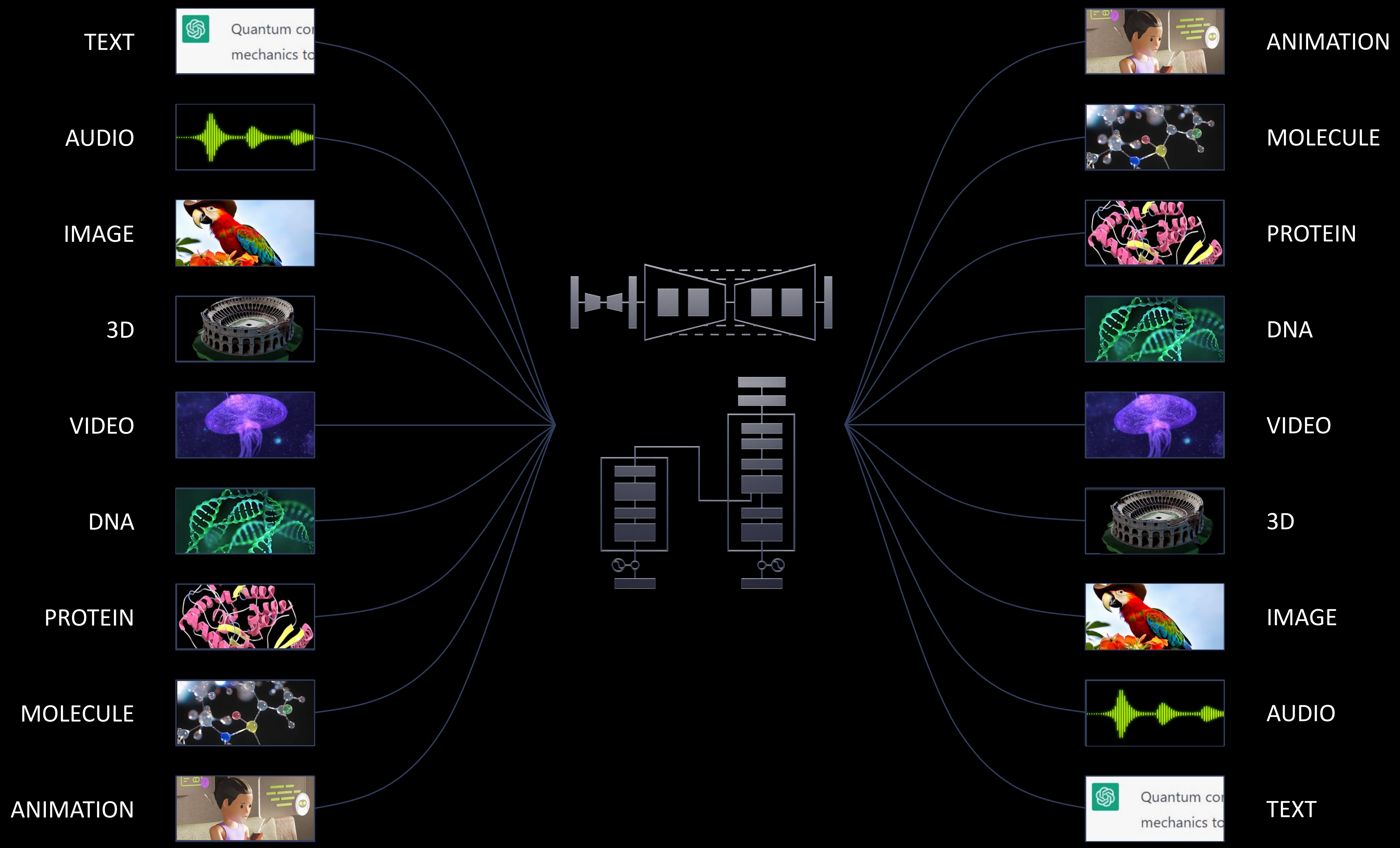
HARDWARE

RTX
DGX
HGX
EGX
OVX
AGX
MLNX

GPU
CPU
DPU
NIC
SWITCH
SOC

Generative AI

The iPhone moment of AI



Generative AI Market and Use Cases

What's the Hype!

Generative AI is here

Key Players

Microsoft

Reinvent search with Generative AI

OpenAI ChatGPT

110 million users in 2 months

NVIDIA

Generative AI computing power & platform of the world

Meta

Business group focused on Generative AI

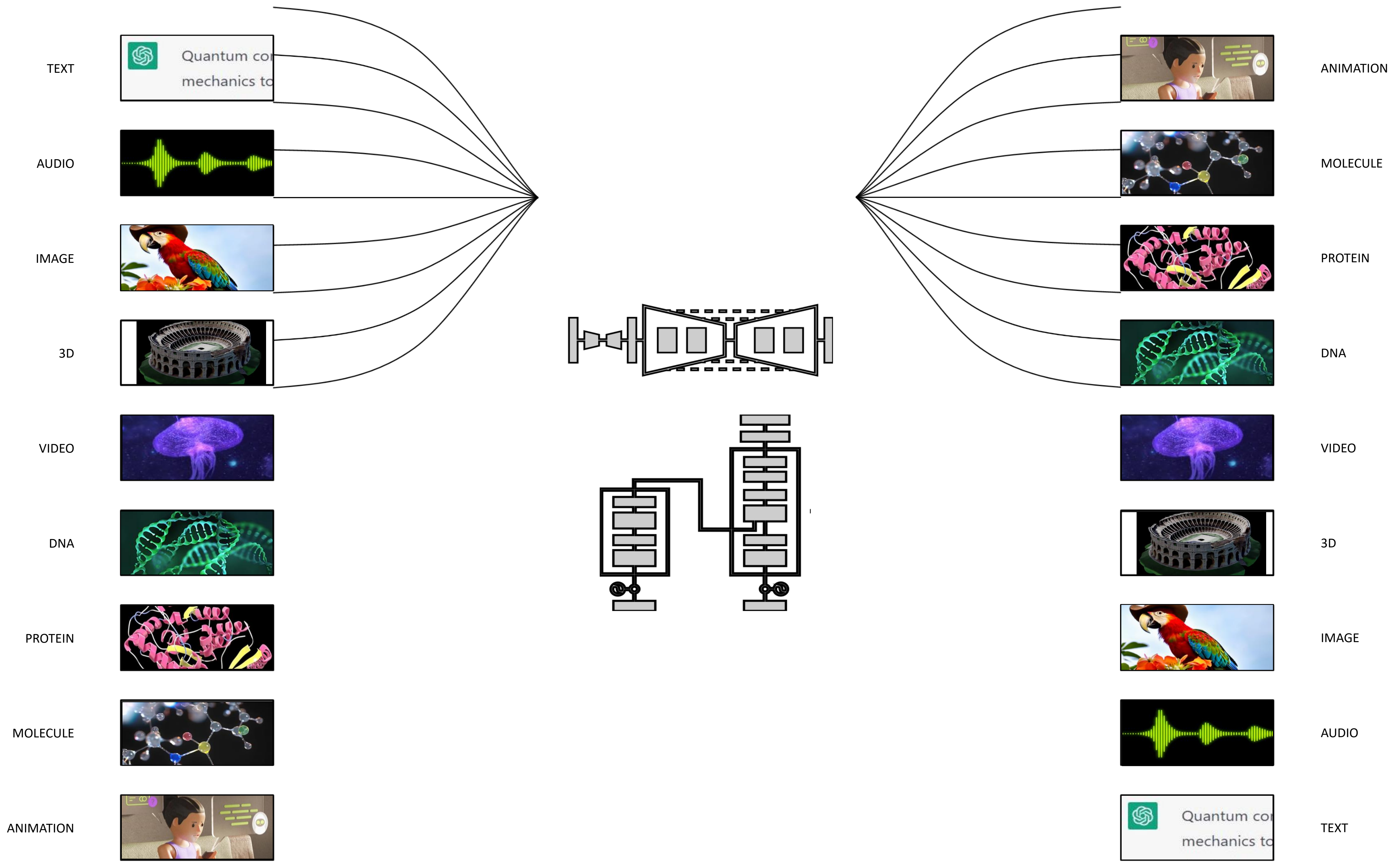
Enterprises & Startups

1600+

Google

Generative AI changing the main products

Generative AI Possibilities



Generative AI is Having a Moment!

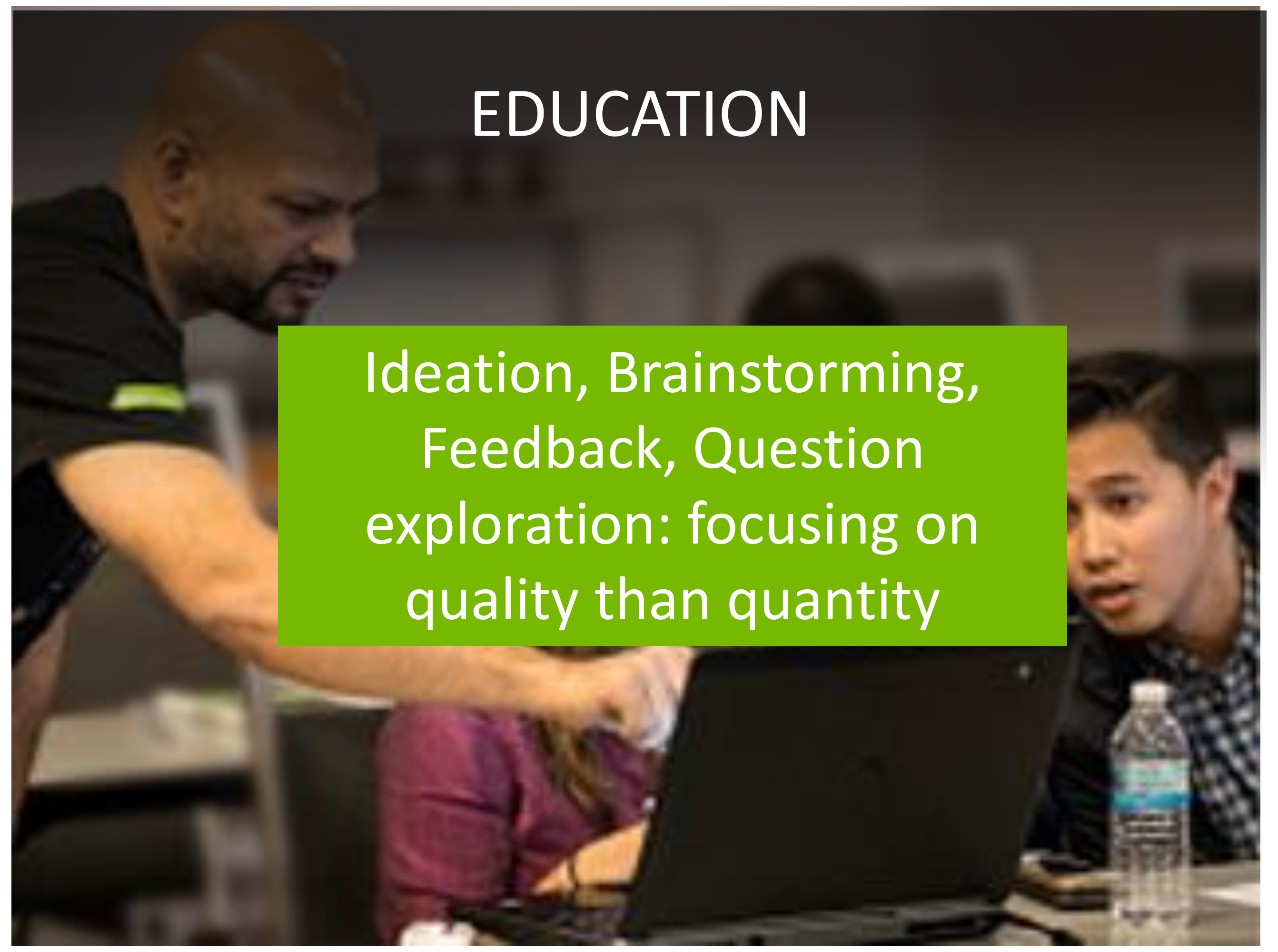
Captivating World's Attention | Happening Fast | Solving Problems in a Way Never Done Before

Benefits of Generative AI on Various Business

Generative AI is transforming every industries

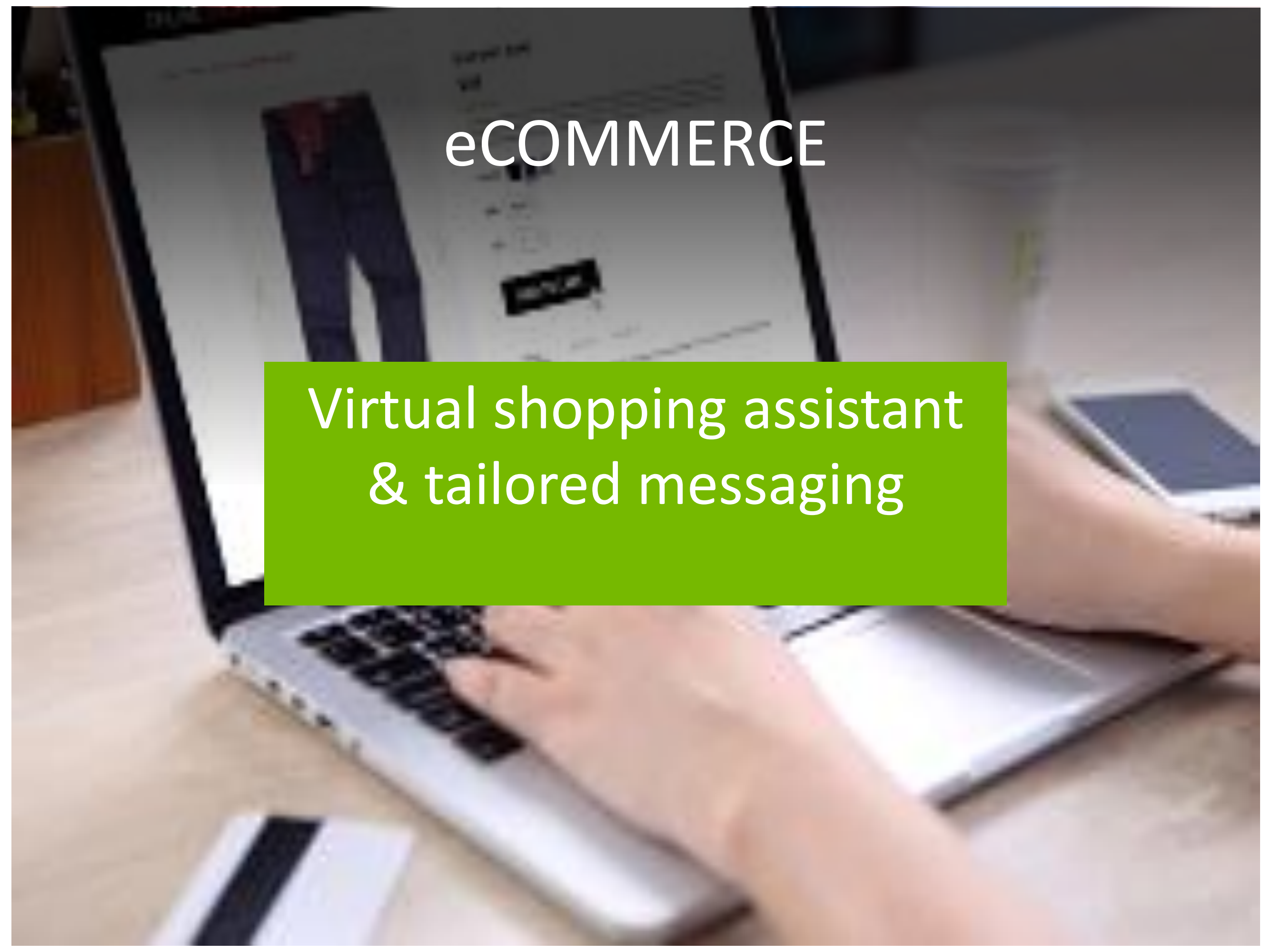
HEALTHCARE & PHARMA

Drug Discovery,
Predicting patient outcomes, &
summarizing Doctor's notes



EDUCATION

Ideation, Brainstorming, Feedback, Question exploration: focusing on quality than quantity

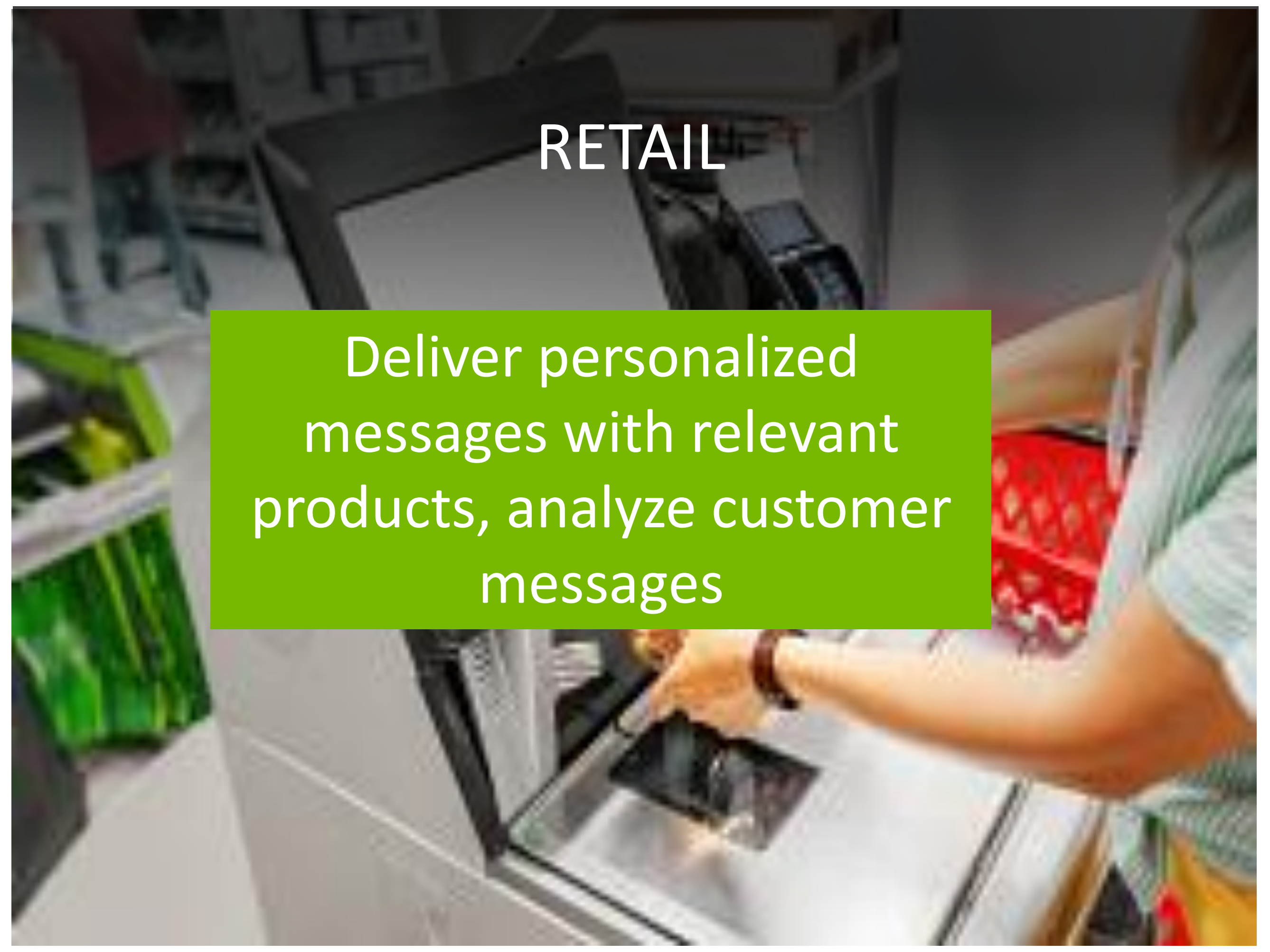


eCOMMERCE

Virtual shopping assistant & tailored messaging

MARKETING

Copy generation, New age advertising, Headline generation, Personalized images and text



RETAIL

Deliver personalized messages with relevant products, analyze customer messages

MEDIA & ENTERTAINMENT

Improve efficiency of running new ideas

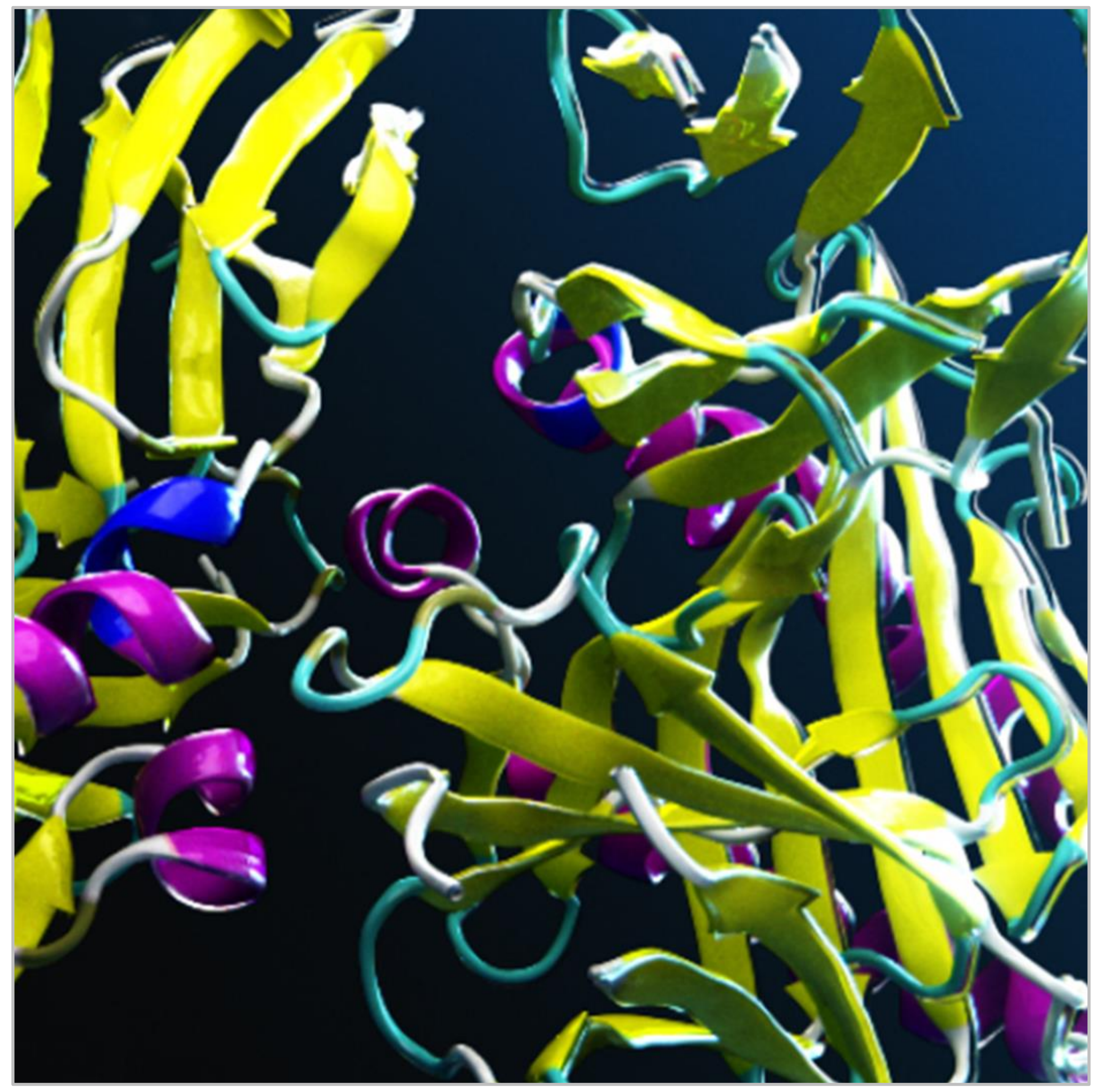
Generative AI Risks

Challenges/Limitations of Generative AI

- Data Privacy & Security
- IP Rights & Copyright
- Biases, Errors, & Limitations
- Ethical Implication
- Malevolent Activities

NVIDIA AI Foundations

NEMO



BIONEMO

PICASSO

NVIDIA AI ENTERPRISE

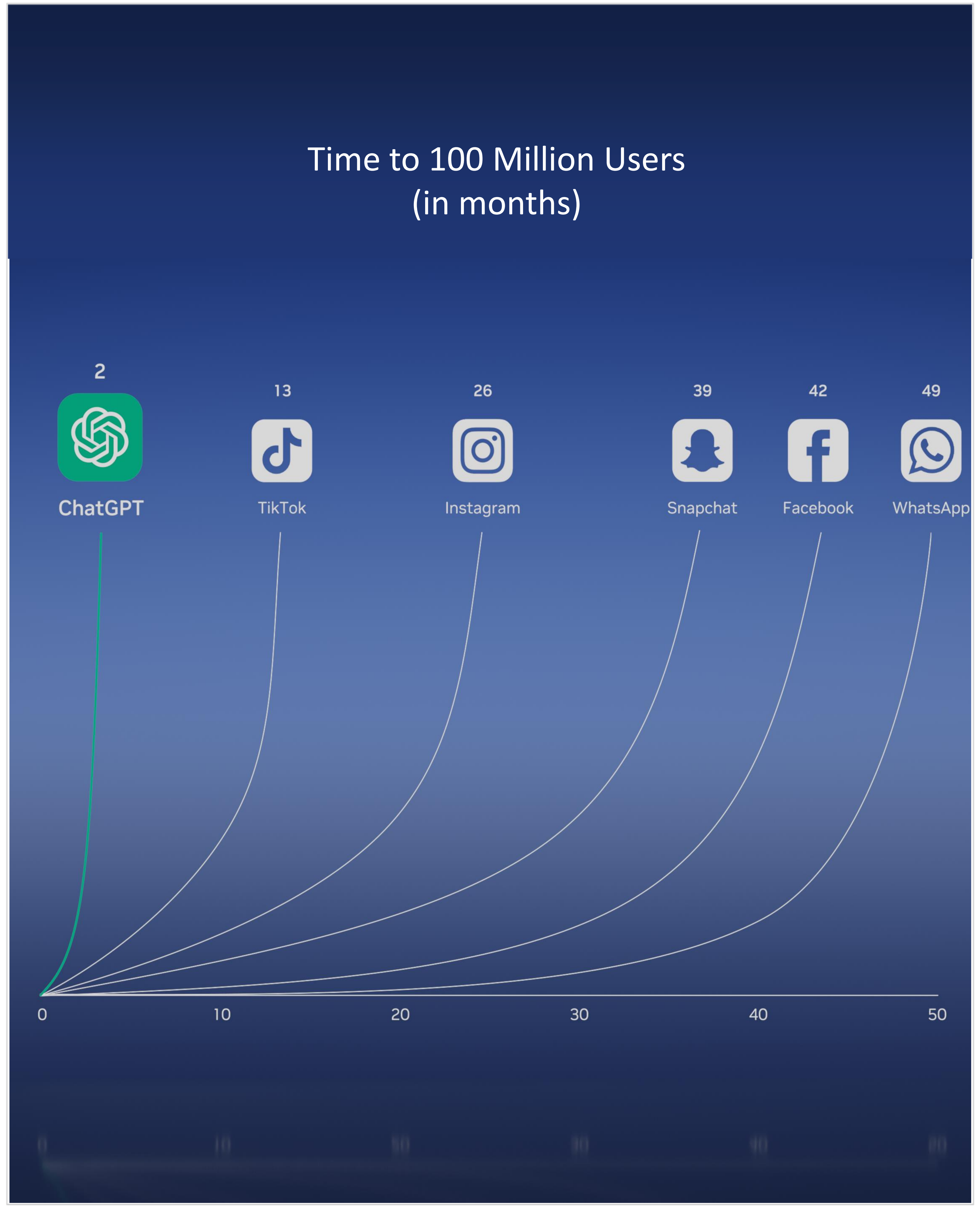
NVIDIA DGX Cloud

The iPhone Moment of AI is Here

Every major application and workflow is going to include AI

CHATBOTS

Fastest Growing Application Ever

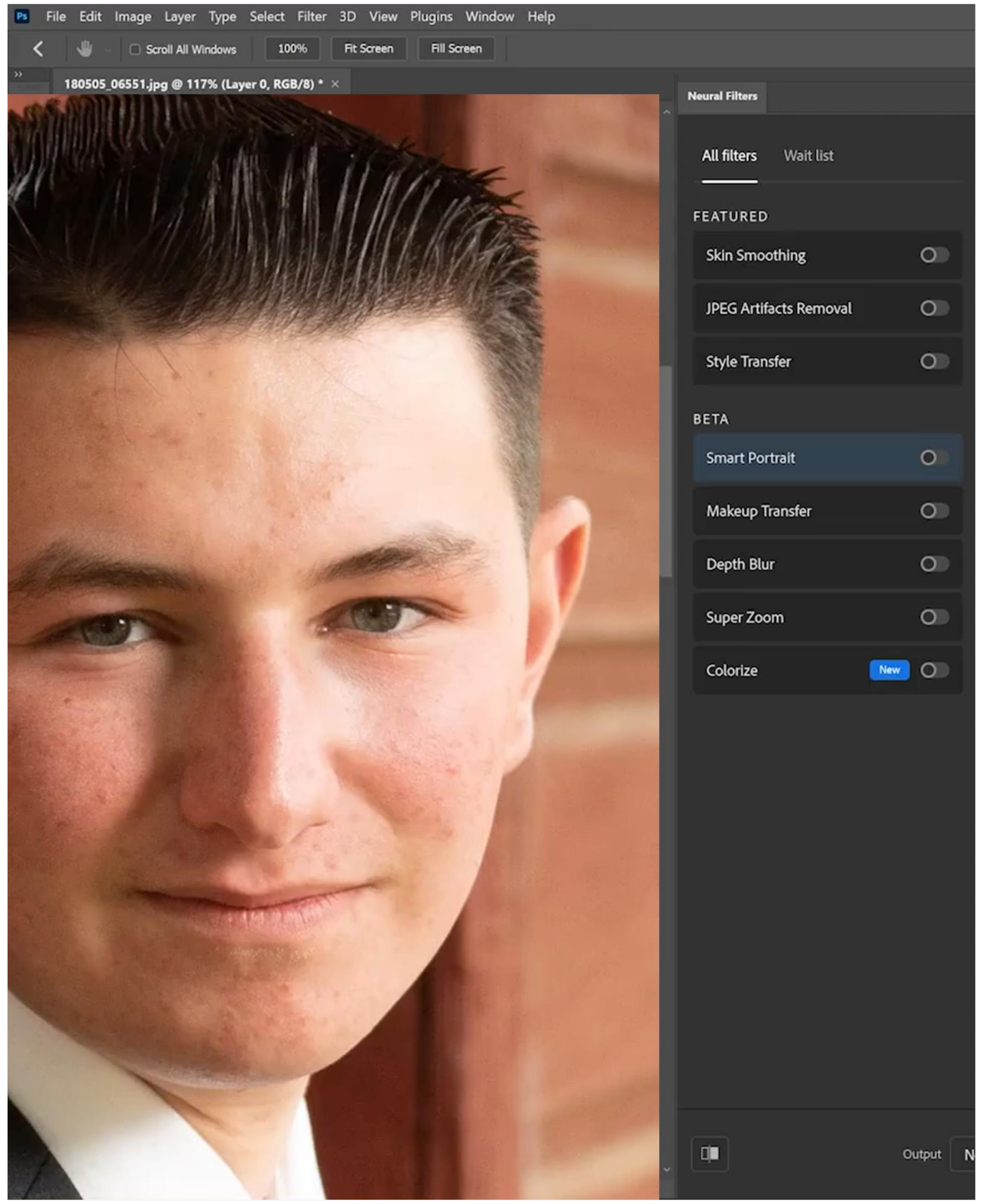


GENERATIVE ART

Over 200M+ Users

AI-AUGMENTED APPLICATIONS

ISVs Accelerating AI Integration



Generative AI Technology Preview

AI Playground on NGC

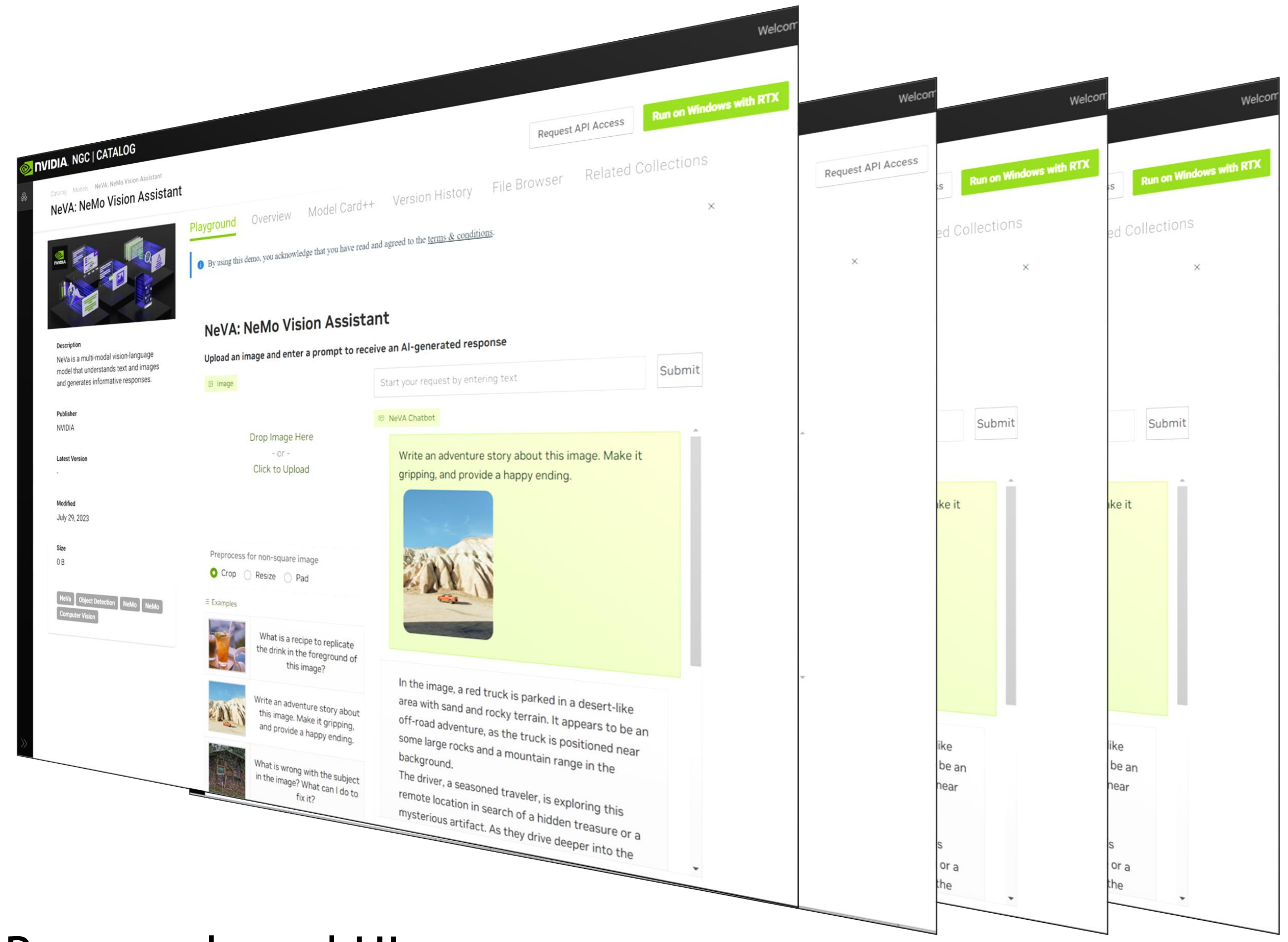
Experience Top Generative AI Models via Simple Web UI

Explore top NVIDIA & OSS AI models on NGC

Experience on DGX Cloud via AI Playground

One-click install to RTX PC

Fine tune with AI Workbench



NeVA: NeMo Vision Assistant

By using this demo, you acknowledge that you have read and agreed to the terms & conditions.

Description

NeVa is a multi-modal vision-language model that understands text and images and generates informative responses.

Publisher: NVIDIA

Latest Version: -

Modified: July 29, 2023

Size: 0 B

NeVA: NeMo Vision Assistant

Upload an image and enter a prompt to receive an AI-generated response

Start your request by entering text

Preprocess for non-square image: ● Crop ○ Resize ○ Pad

What is a recipe to replicate the drink in the foreground of this image?

Write an adventure story about this image. Make it gripping, and provide a happy ending.

What is wrong with the subject in the image? What can I do to fix it?

Write an adventure story about this image. Make it gripping, and provide a happy ending.

In the image, a red truck is parked in a desert-like area with sand and rocky terrain. It appears to be an off-road adventure, as the truck is positioned near some large rocks and a mountain range in the background.

The driver, a seasoned traveler, is exploring this remote location in search of a hidden treasure or a mysterious artifact. As they drive deeper into the

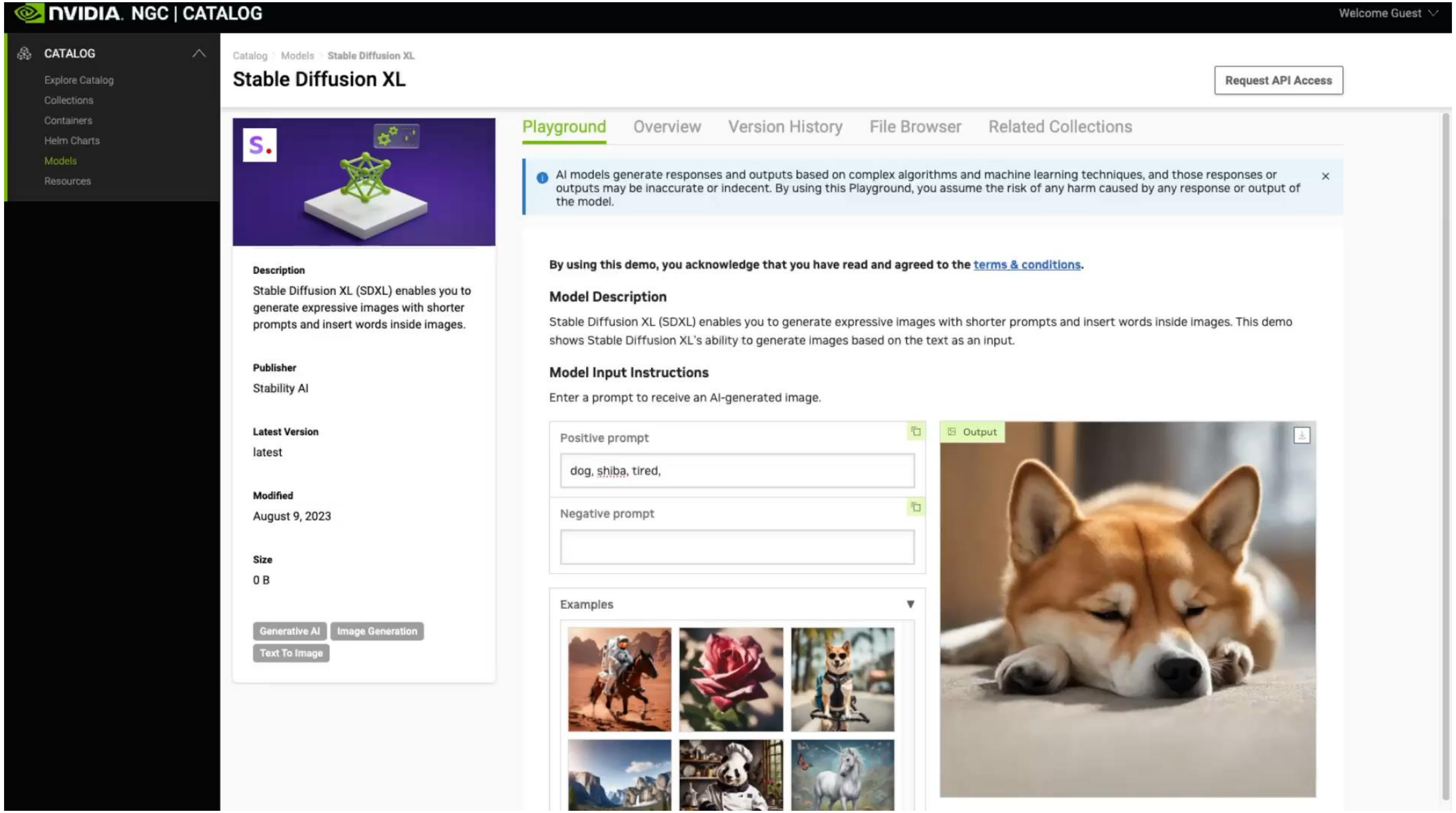
Browser-based UI

DGX Cloud

Windows RTX PC/WS

AI Playground on NGC

Demo



Stable Diffusion XL

Description

Stable Diffusion XL (SDXL) enables you to generate expressive images with shorter prompts and insert words inside images.

Publisher

Stability AI

Latest Version

latest

Modified

August 9, 2023

Size

0 B

AI models generate responses and outputs based on complex algorithms and machine learning techniques, and those responses or outputs may be inaccurate or indecent. By using this Playground, you assume the risk of any harm caused by any response or output of the model.

By using this demo, you acknowledge that you have read and agreed to the terms & conditions.

Model Description

Stable Diffusion XL (SDXL) enables you to generate expressive images with shorter prompts and insert words inside images. This demo shows Stable Diffusion XL's ability to generate images based on the text as an input.

Model Input Instructions

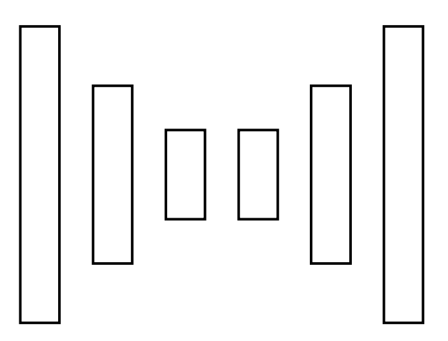
Enter a prompt to receive an AI-generated image.

Positive prompt: dog, shiba, tired,

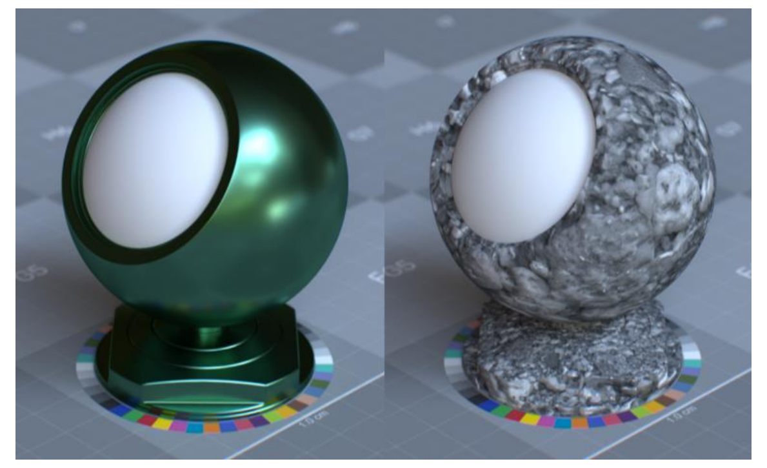
Negative prompt:

Populating 3D Worlds With Generative AI

GenAI Model

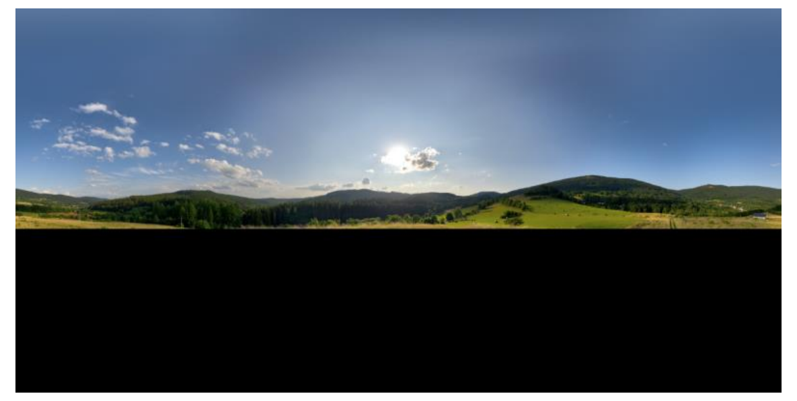


Background 3D

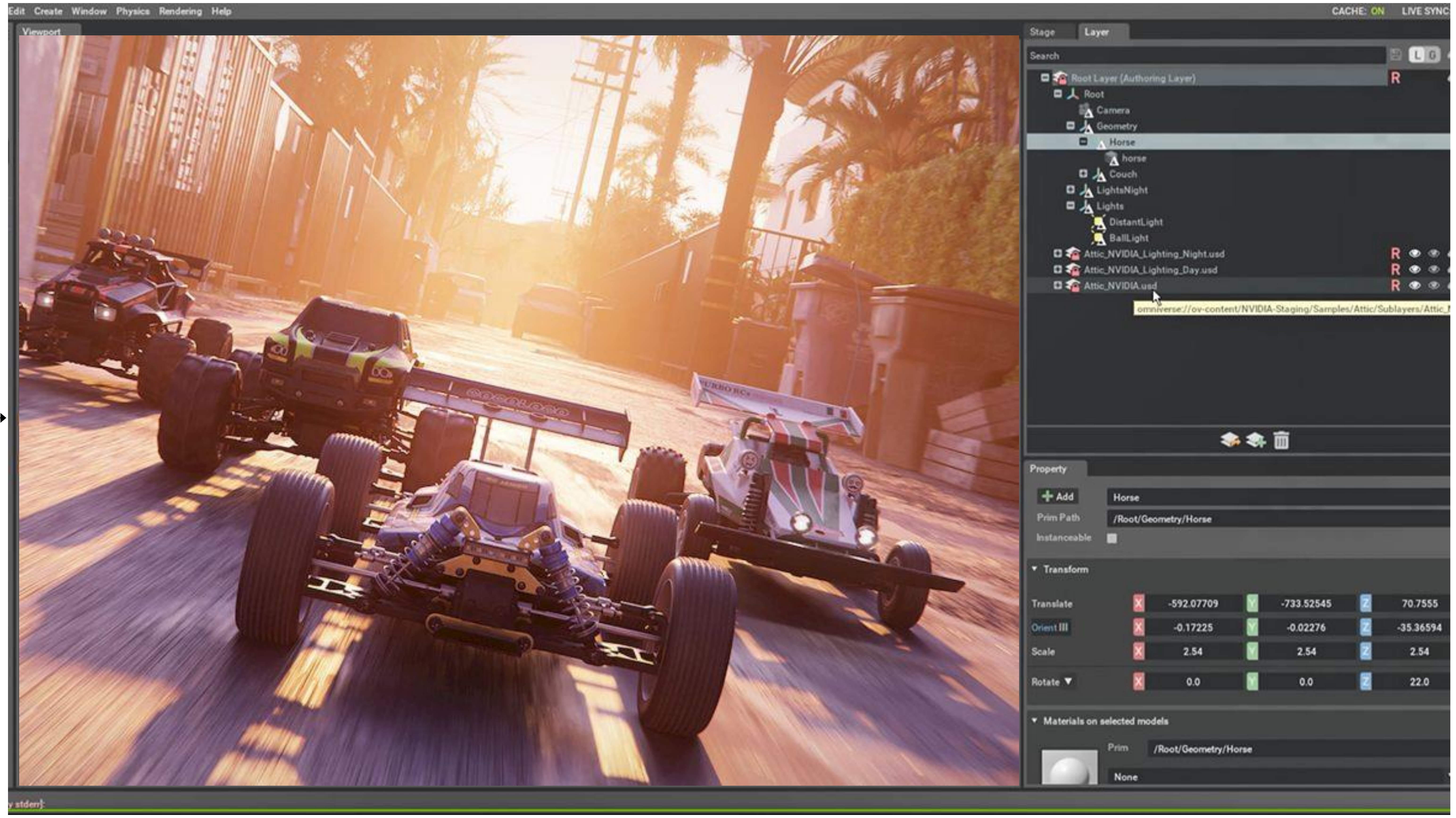


Materials

2D Posters & Backplates



360 HDRi


Edit Create Window Physics Rendering Help
CACHE: ON
LIVE SYNC
Viewport
Stage
Layer
Search
Root Layer (Authoring Layer)
Root
Camera
Geometry
Horse
horse
Couch
LightsNight
Lights
DistantLight
BallLight
Attic_NVIDIA_Lighting_Night.usd
Attic_NVIDIA_Lighting_Day.usd
Attic_NVIDIA.usd
omniverse://ov-content/NVIDIA-Staging/Samples/Attic/Sublayers/Attic_
Property
Add
y stderr]:

Horse

Prim Path /Root/Geometry/Horse

Instanceable □

▾ Transform

Translate	X	-592.07709	Y	-733.52545	Z	70.7555
Orient	X	-0.17225	Y	-0.02276	Z	-35.36594
Scale	X	2.54	Y	2.54	Z	2.54
Rotate ▼	X	0.0	Y	0.0	Z	22.0

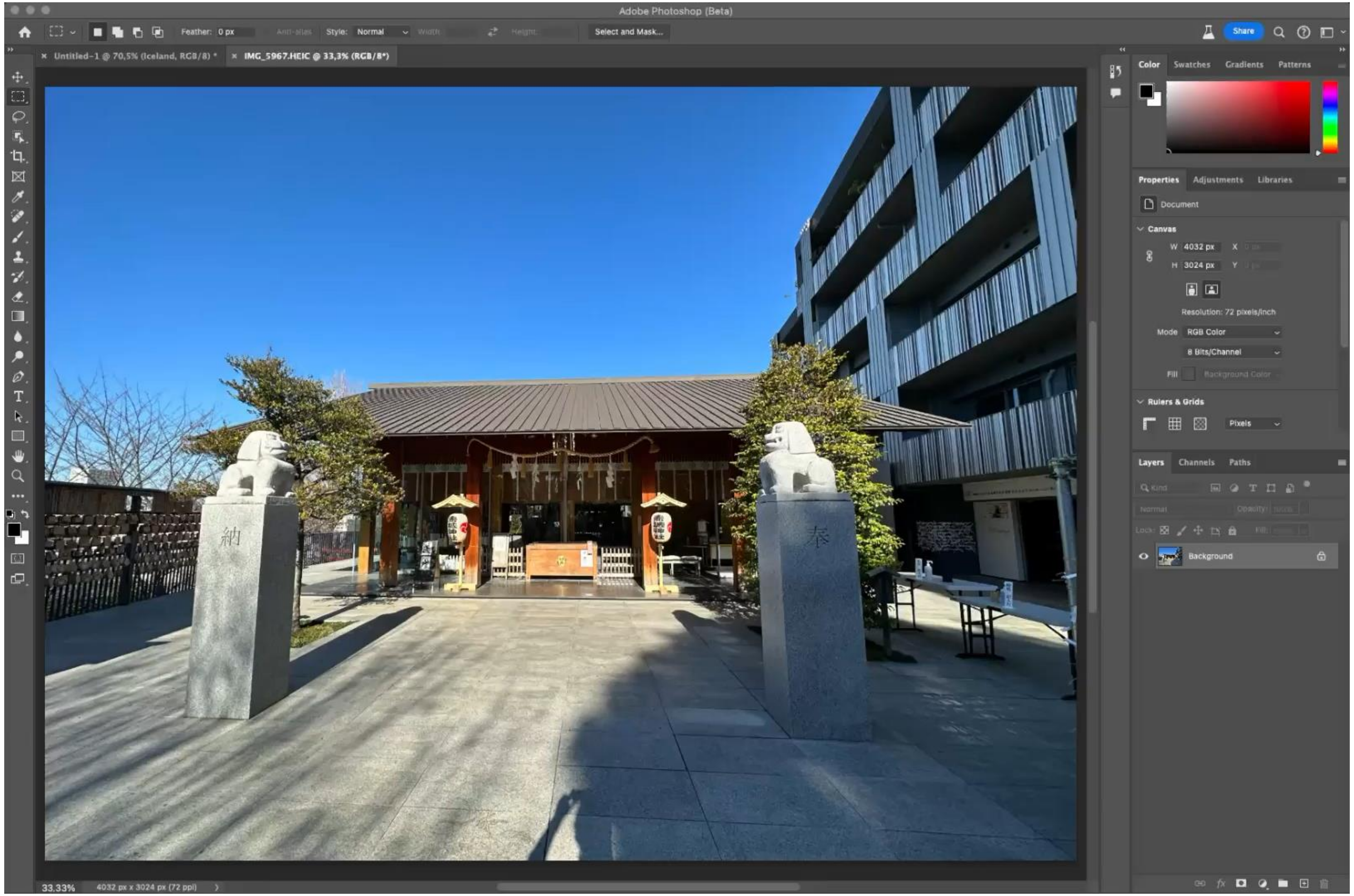
▾ Materials on selected models

Prim /Root/Geometry/Horse

None

Adobe Photoshop Example

Adding Content with Generative Fill



Adobe Photoshop (Beta)

Feather: 0 px

Style: Normal

Untitled-1 @ 70,5% (Iceland, RGB/8) *

IMG_5967.HEIC @ 33,3% (RGB/8*)

Canvas

W 4032 px

H 3024 px

Resolution: 72 pixels/inch

Mode RGB Color

8 Bits/Channel

Rulers & Grids

Pixels

33.33%

4032 px x 3024 px (72 ppi)

Generative AI Transforming Workflows Across Industries

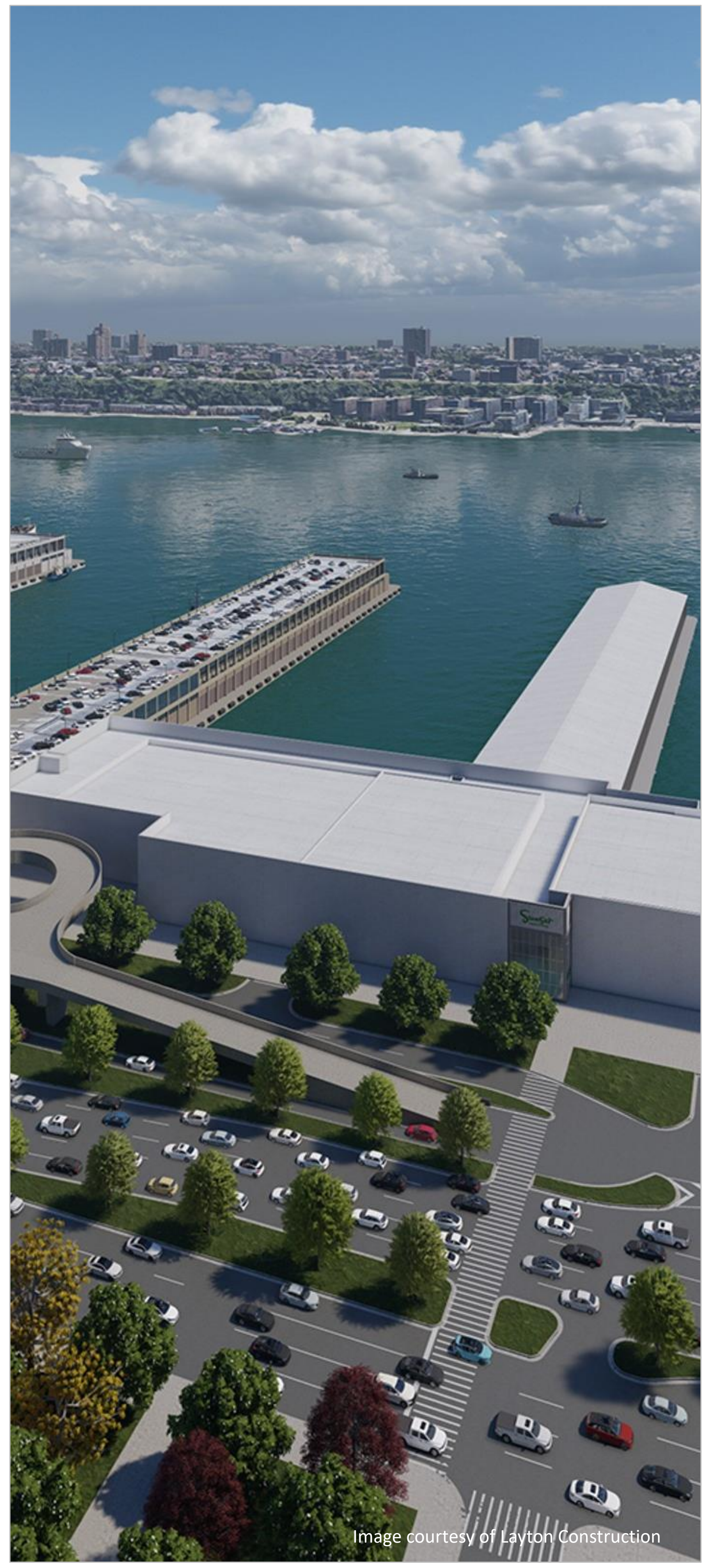
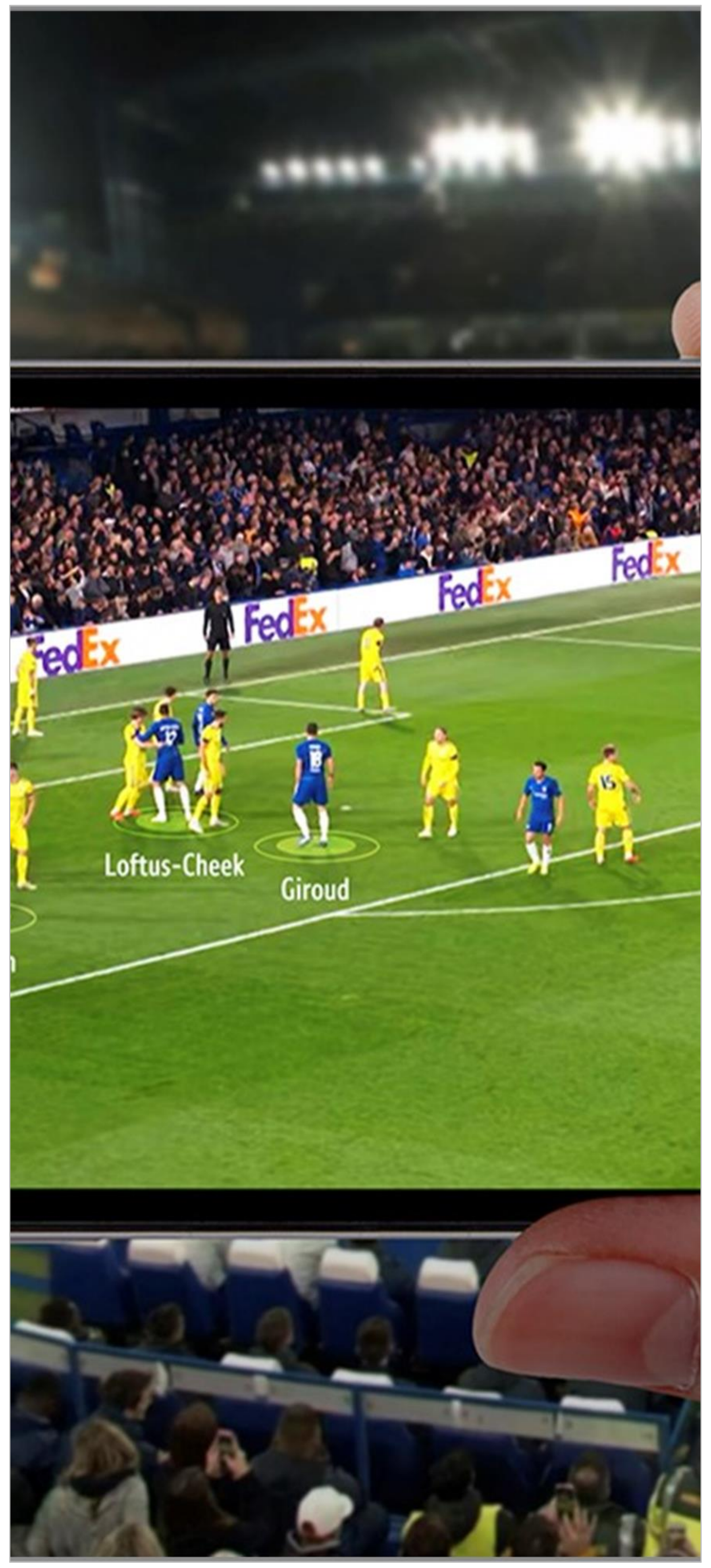


Image courtesy of Layton Construction

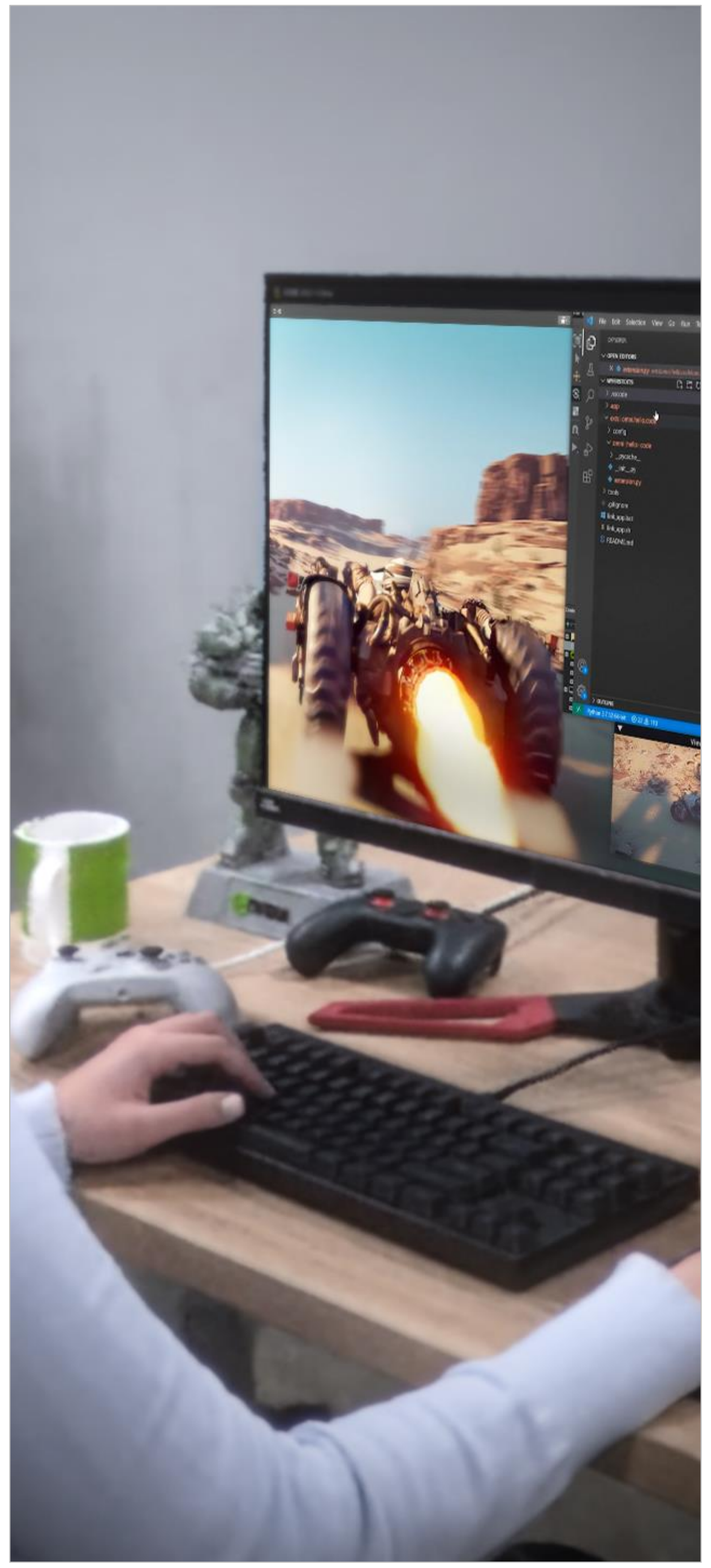
Architecture

Courtesy of Debarti Studio

Product Design



Film / Video



3D FX / Game Dev

Marketing

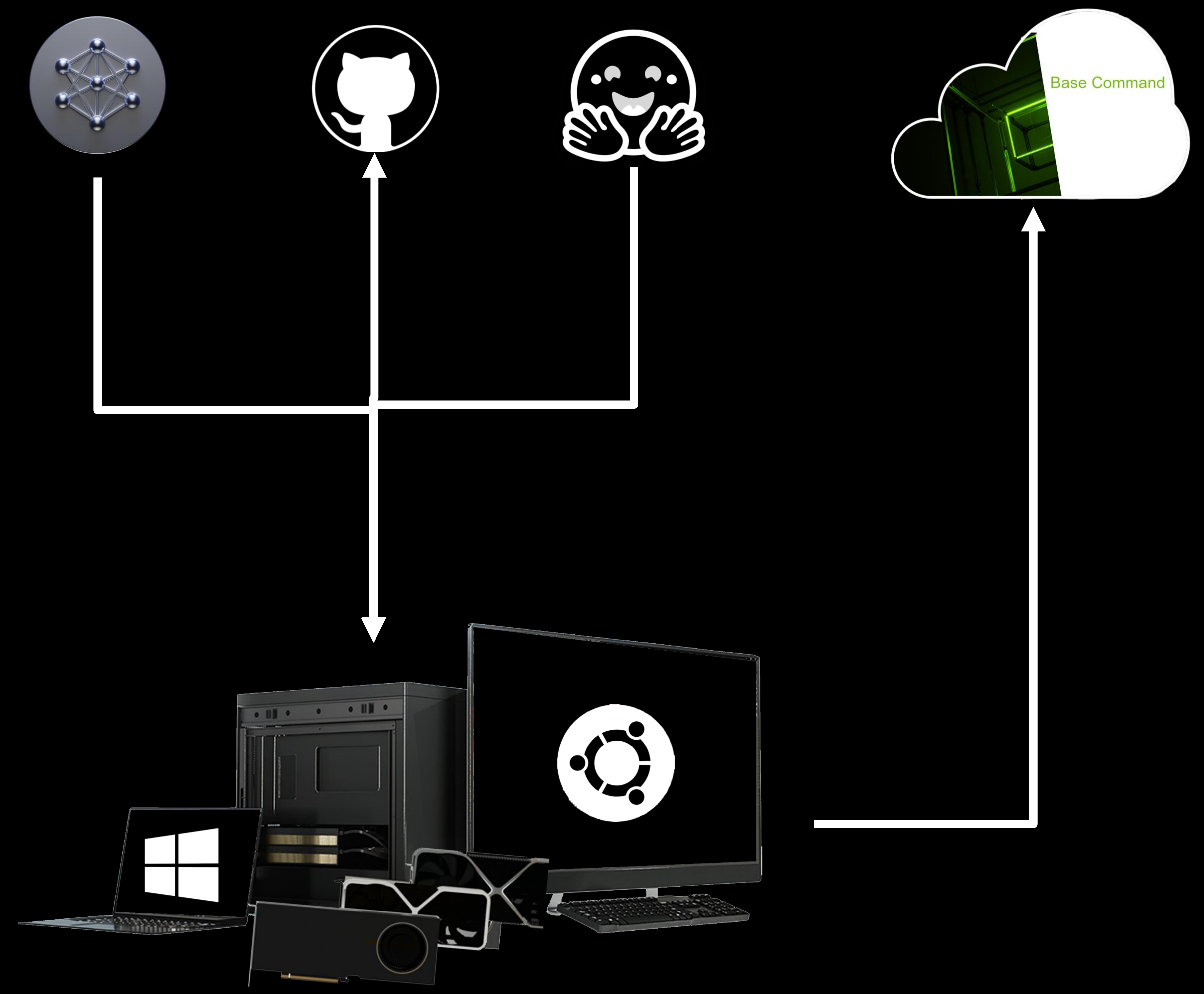
Courtesy of Cinema4D

Photography

September Early Access for NVIDIA AI Workbench

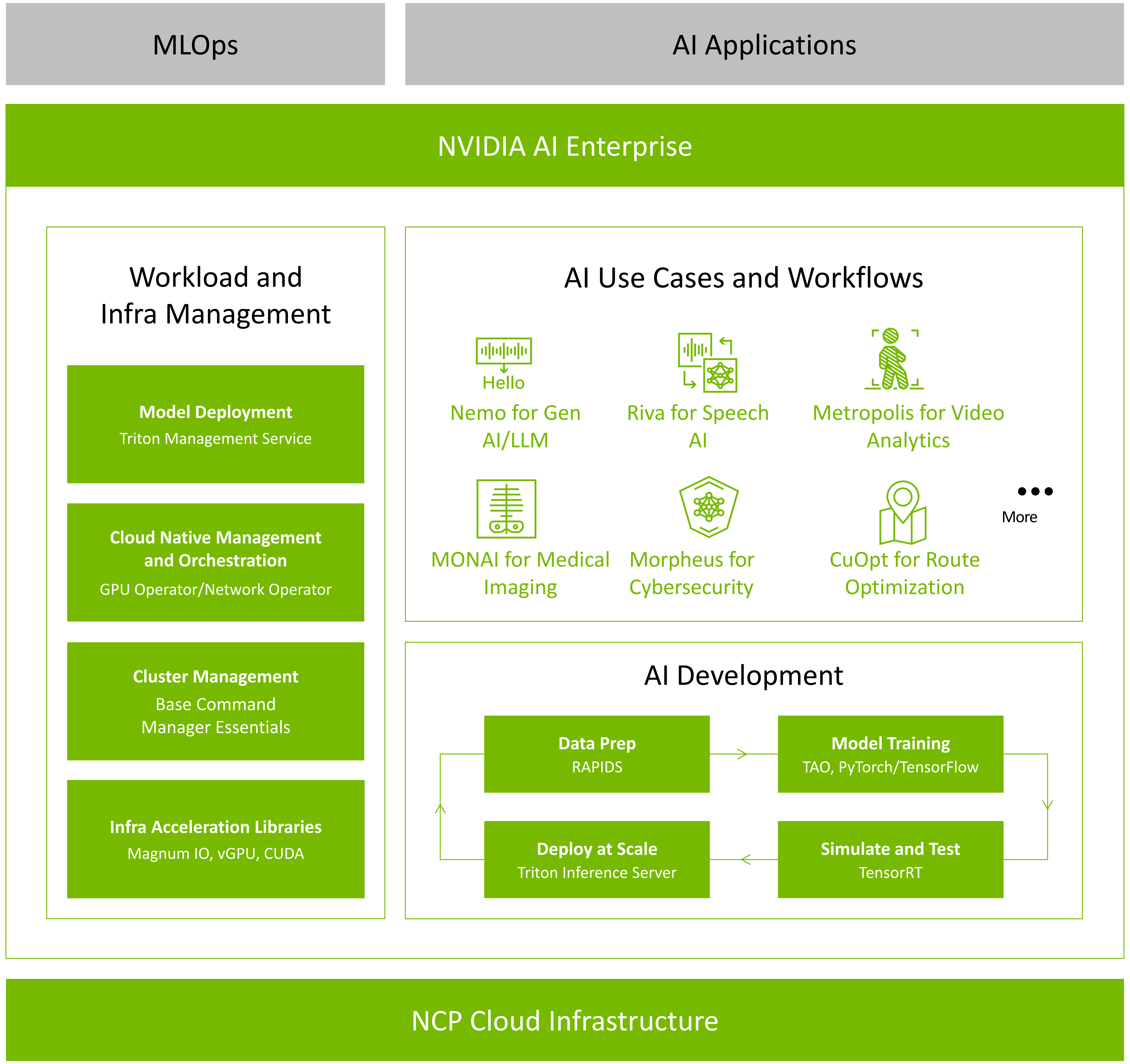
Announcing at SIGGRAPH

- New development platform for data scientists & AI/ML engineers
- Brings Gen AI development to RTX GPUs on PCs & Linux
- Streamlines NVIDIA HW & SW for a broad audience
- Integrates with NGC, GitHub & Hugging Face
- One-click push of local work into DGX Cloud
- Easy hybrid deployment on local machines & clouds
- Free and available for self-service install from NGC
- EA starting in September



NVIDIA AI Enterprise

End to end AI software



MLOps

AI Applications

NVIDIA AI Enterprise

Workload and Infra Management

Model Deployment
Triton Management Service

Cloud Native Management and Orchestration
GPU Operator/Network Operator

Cluster Management
Base Command Manager Essentials

Infra Acceleration Libraries
Magnum IO, vGPU, CUDA

AI Use Cases and Workflows

Hello

Nemo for Gen AI/LLM

Riva for Speech AI

Metropolis for Video Analytics

MONAI for Medical Imaging

Morpheus for Cybersecurity

CuOpt for Route Optimization

••• More

AI Development

Data Prep
RAPIDS

Model Training
TAO, PyTorch/TensorFlow

Simulate and Test
TensorRT

Deploy at Scale
Triton Inference Server

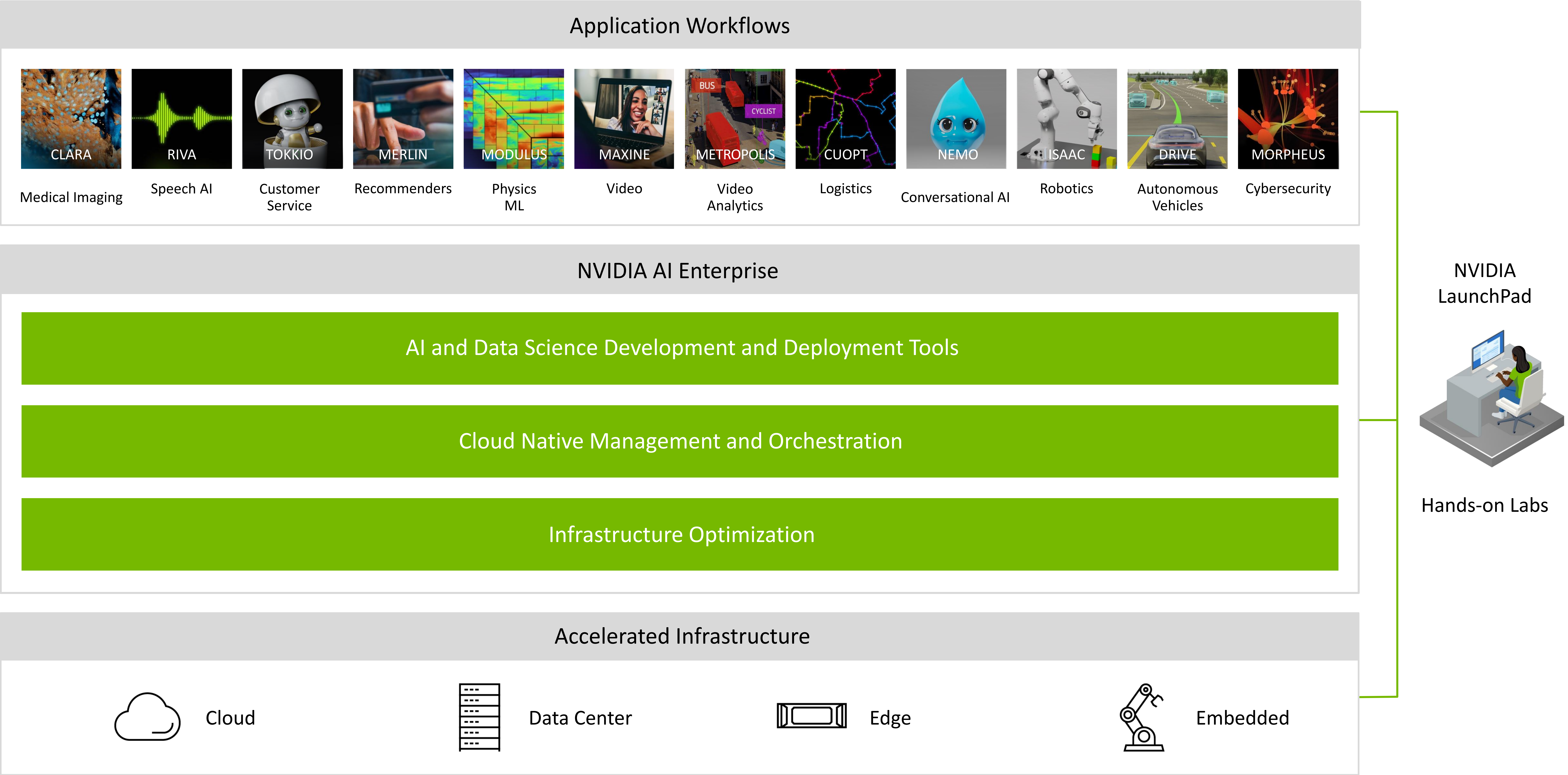
NCP Cloud Infrastructure

- Cloud Native, Hybrid Optimized
- Enterprise Grade OSS Components
- Secure And Scalable
- Technical Standard Support 9 X 5, Premium 24x7

NVIDIA Software

NVIDIA AI

End-to-end open platform for production AI



Application Workflows

CLARA	RIVA	TOKKIO	MERLIN	MODULUS	MAXINE	METROPOLIS	CUOPT	NEMO	ISAAC	DRIVE	MORPHEUS
Medical Imaging	Speech AI	Customer Service	Recommenders	Physics ML	Video	Video Analytics	Logistics	Conversational AI	Robotics	Autonomous Vehicles	Cybersecurity

NVIDIA AI Enterprise

AI and Data Science Development and Deployment Tools

Cloud Native Management and Orchestration

Infrastructure Optimization

Accelerated Infrastructure

Cloud

Data Center

Edge

Embedded

NVIDIA LaunchPad

Hands-on Labs

Productivity with Base Command

Enterprises tools that drive the value of AI investment

NVIDIA AI Enterprise

AI and Data Science Tools / Frameworks

NVIDIA RAPIDS™

NVIDIA TAO Toolkit

NVIDIA Tensor RT™

AI Frameworks

NeMo Megatron

- **NVIDIA AI Enterprise** (included in DGX SW Bundle)

Ready-to-use, fully supported software that speeds **developer success**

NVIDIA Base Command

AI Workflow Management and MLOps

- **DGX-Ready Software Partners**

Get more models from **prototype to production**

Job Scheduling & Orchestration

KUBERNETES

SLURM

- **Kubernetes**
- **Slurm**

Ensure hassle-free **execution of every developer's jobs**

Cluster Management

PROVISIONING

MONITORING

CLUSTERING

MANAGING

- **Bright Cluster Manager**
- **Base Command Manager** (DGX SuperPOD)

Effortlessly **scale and manage** one node to thousands

Network/Storage Acceleration Libraries & Management

NETWORK IO

STORAGE IO

IN-NETWORK COMPUTE

IO MANAGEMENT

- **Magnum IO, UFM, NetQ**

Accelerate end-to-end **infrastructure performance**

Operating System

DGX OS Extensions for Linux Distributions

- **DGX OS**

Maximize system uptime security and reliability

Single Node

Multi-Node

DGX H100 SuperPOD

NVIDIA Reference Architectures

Introducing Base Command Manager Essentials

Purpose-built for Enterprise AI Infrastructure Management

Infrastructure Provisioning

- Maintain a secure, up-to-date, and reliable AI infrastructure

Workload Management

- Easily provide data scientists with all the tools and resources they need

Resource Monitoring

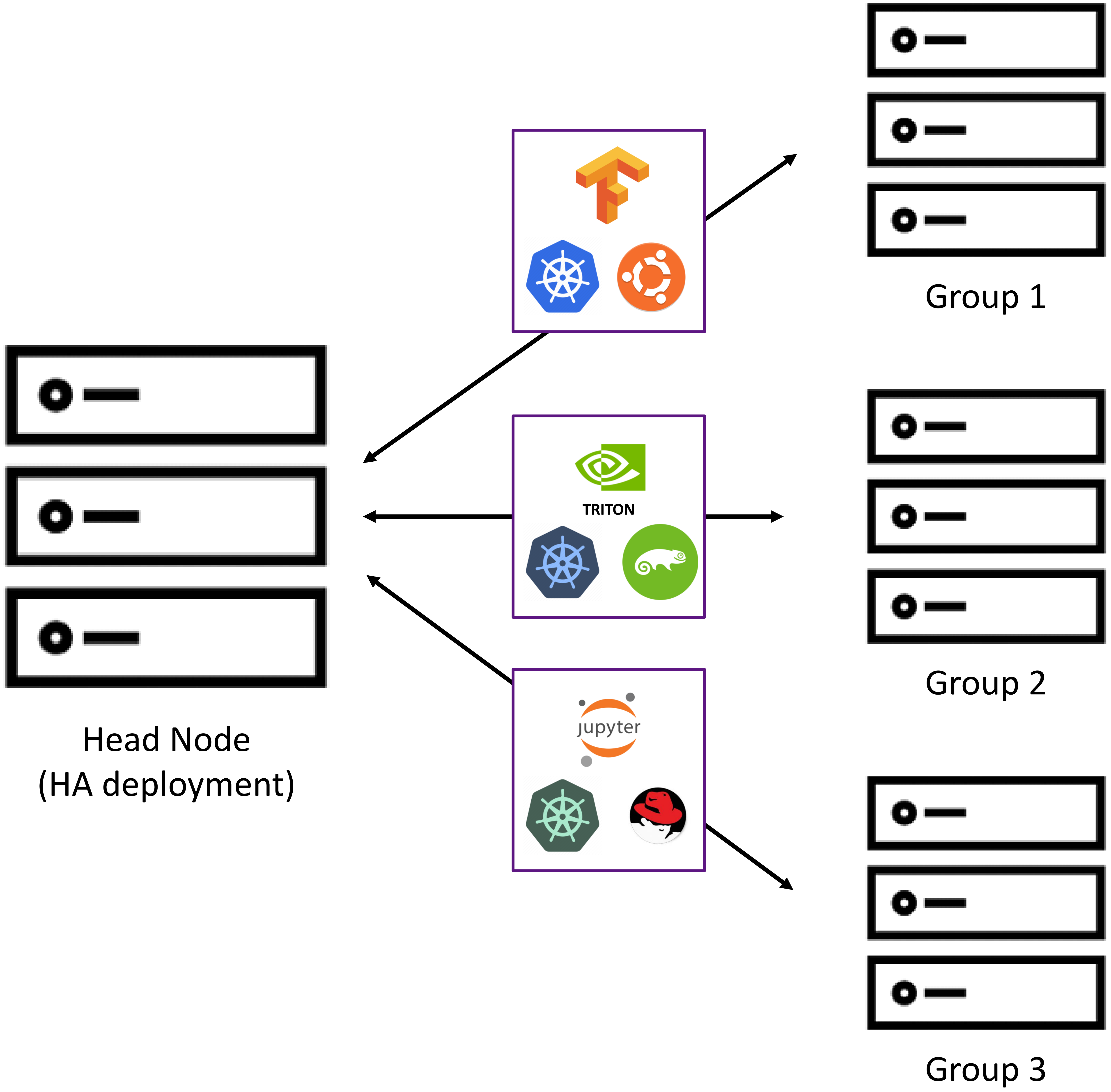
- Get detailed data for informed decision-making

Data Center

Cloud

Powerful Image Management and Cluster Configuration

Greatly increases admin productivity and prevents configuration drift



Provision systems by categories

- Share common software stack

Software stack synchronization

- Occurs at each reboot

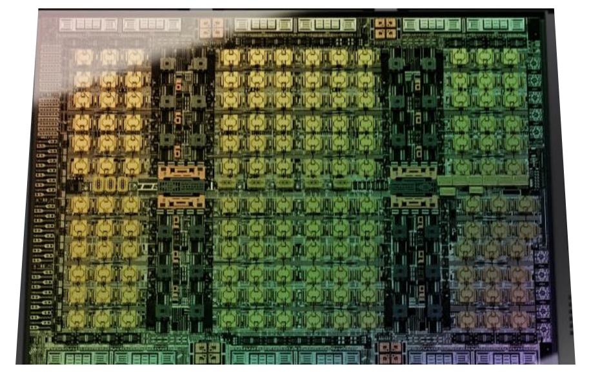
Post-install configuration

- Networking & Security
- User authentication
- GPU settings
- BMC & Power settings
- *and more*

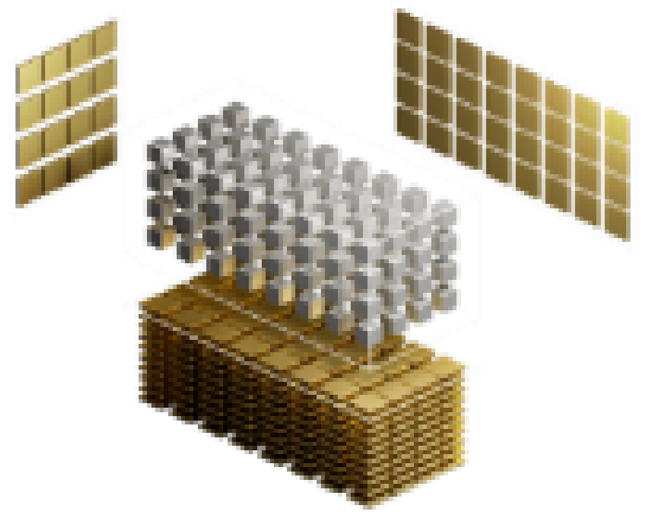
Selected NVIDIA GPUs

NVIDIA Hopper

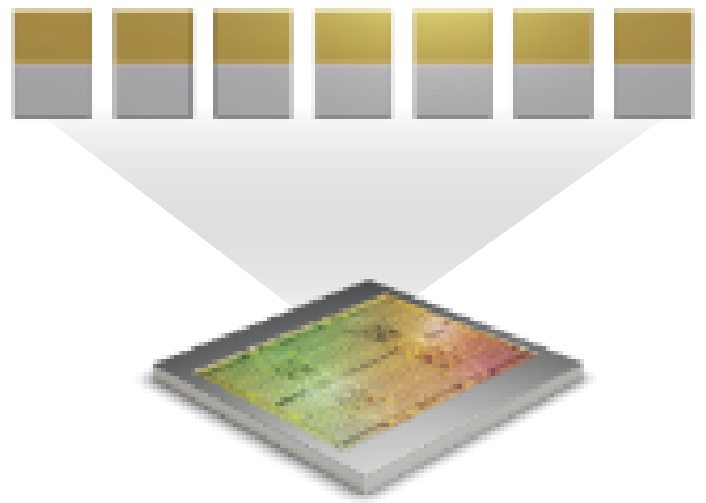
The Engine for the World's AI Infrastructure



World's Most Advanced Chip



Transformer Engine

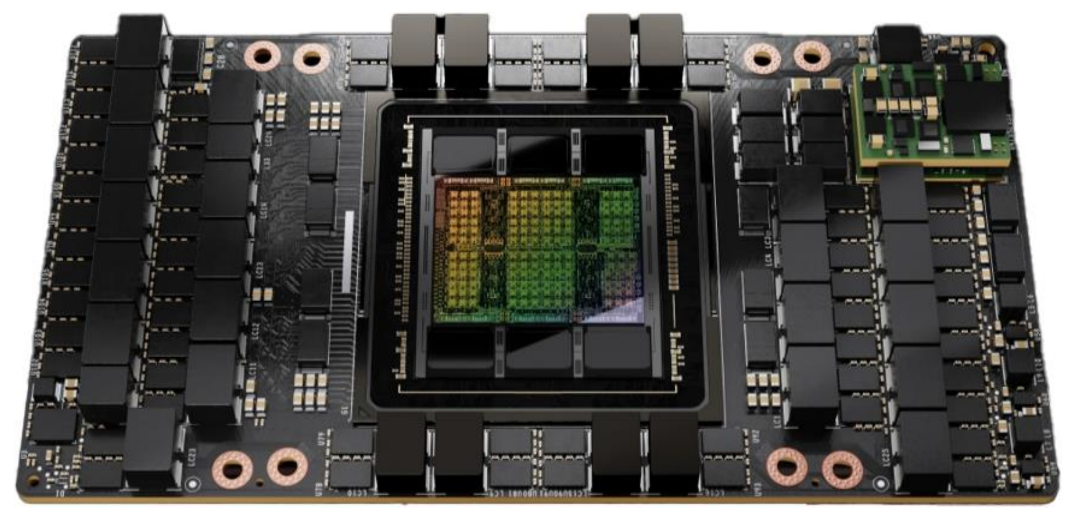


2nd Gen MIG

Confidential Computing

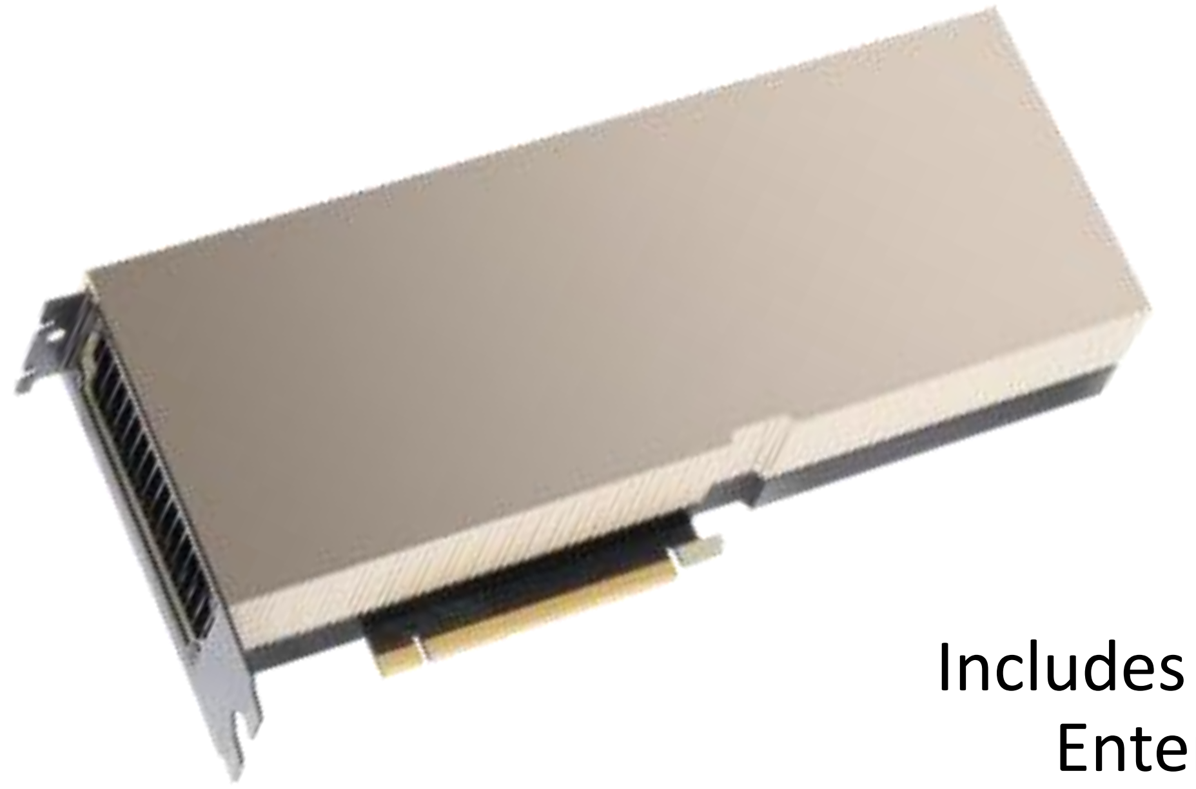
4th Gen NVLink

DPX Instructions



H100 SXM

H100 NVL



H100 PCIE

Includes NVIDIA AI Enterprise

NVIDIA AI Enterprise is a 5-Year Subscription

NVIDIA H100

Unprecedented Performance, Scalability, and Security for Every Data Center

Highest AI and HPC Performance

4PF FP8 (6X)| 2PF FP16 (3X)| 1PF TF32 (3X)| 60TF FP64 (3.4X)
3.35TB/s (1.5X), 80GB HBM3 memory

Transformer Model Optimizations

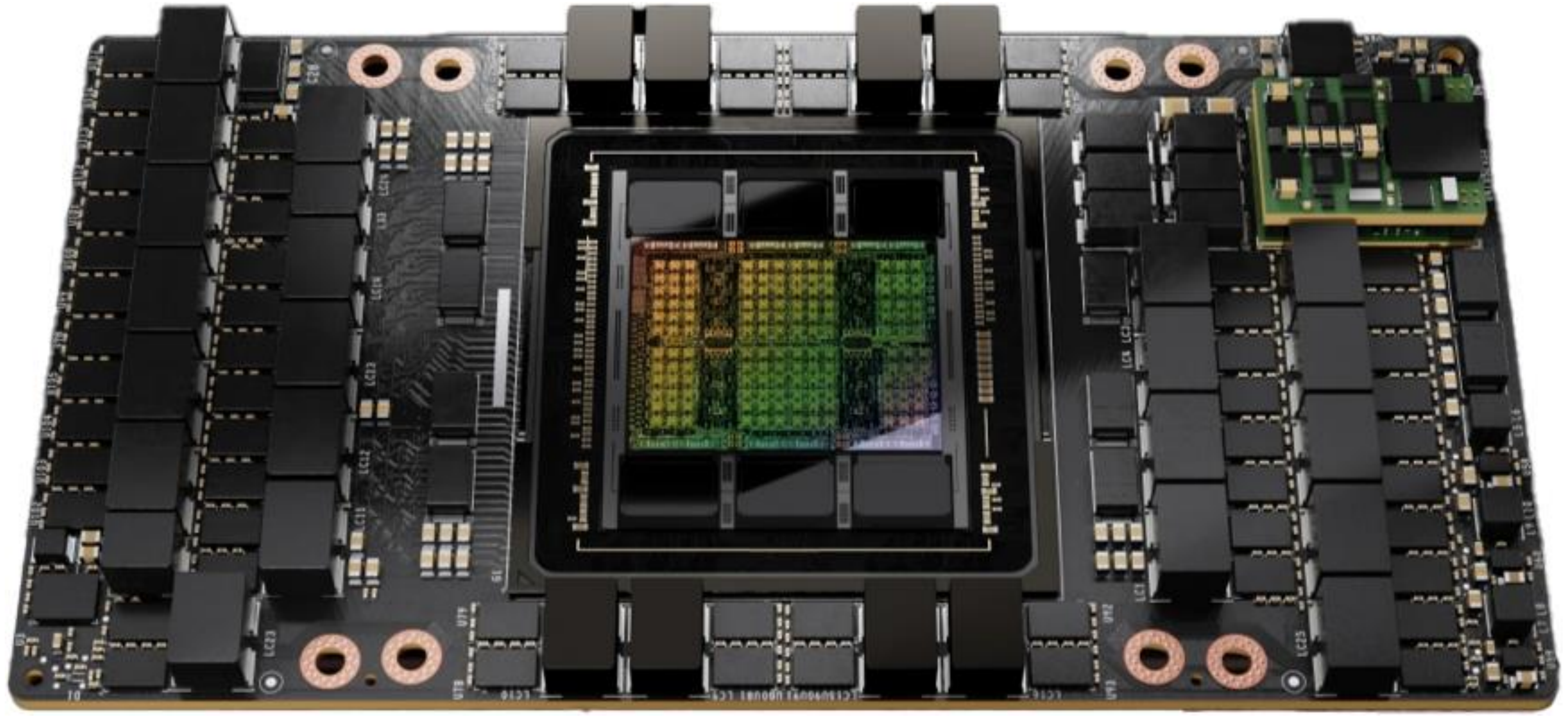
6X faster on largest transformer models

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)
128GB/s PCI Gen5



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

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 Proven Choice for Enterprise AI

8x NVIDIA H100 GPUs | 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

DGX SuperPOD WITH DGX H100

32 DGX H100 | 1 EFLOPS AI
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

NVIDIA L40S

The highest performance universal GPU for AI, graphics, and video

LLM Fine Tuning

8 hours

Llama 2-70B 1 Billion Tokens[1]

LLM 1st Token Latency

<30 ms

Llama 2-13B Inference 225/20[2]

Small LLM Training

3 days

Llama 2-7B 100 Billion Tokens[3]

Image Gen AI

>82

Images per Minute[4]

LLM E2E Latency

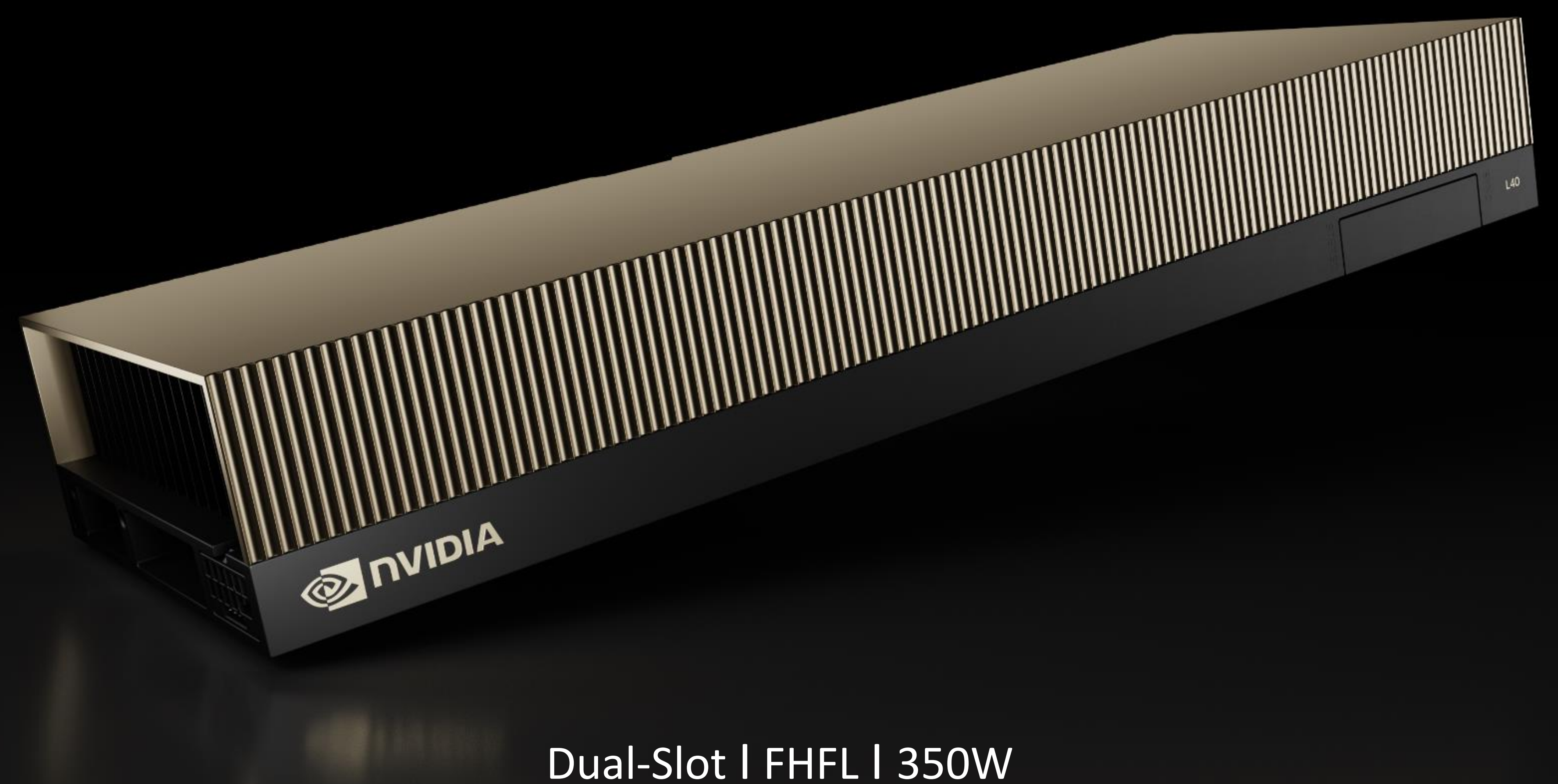
<0.5 s

Llama 2-13B Inference 225/20[5]

Full Video Pipeline

184

AV1 Encode Streams[6]

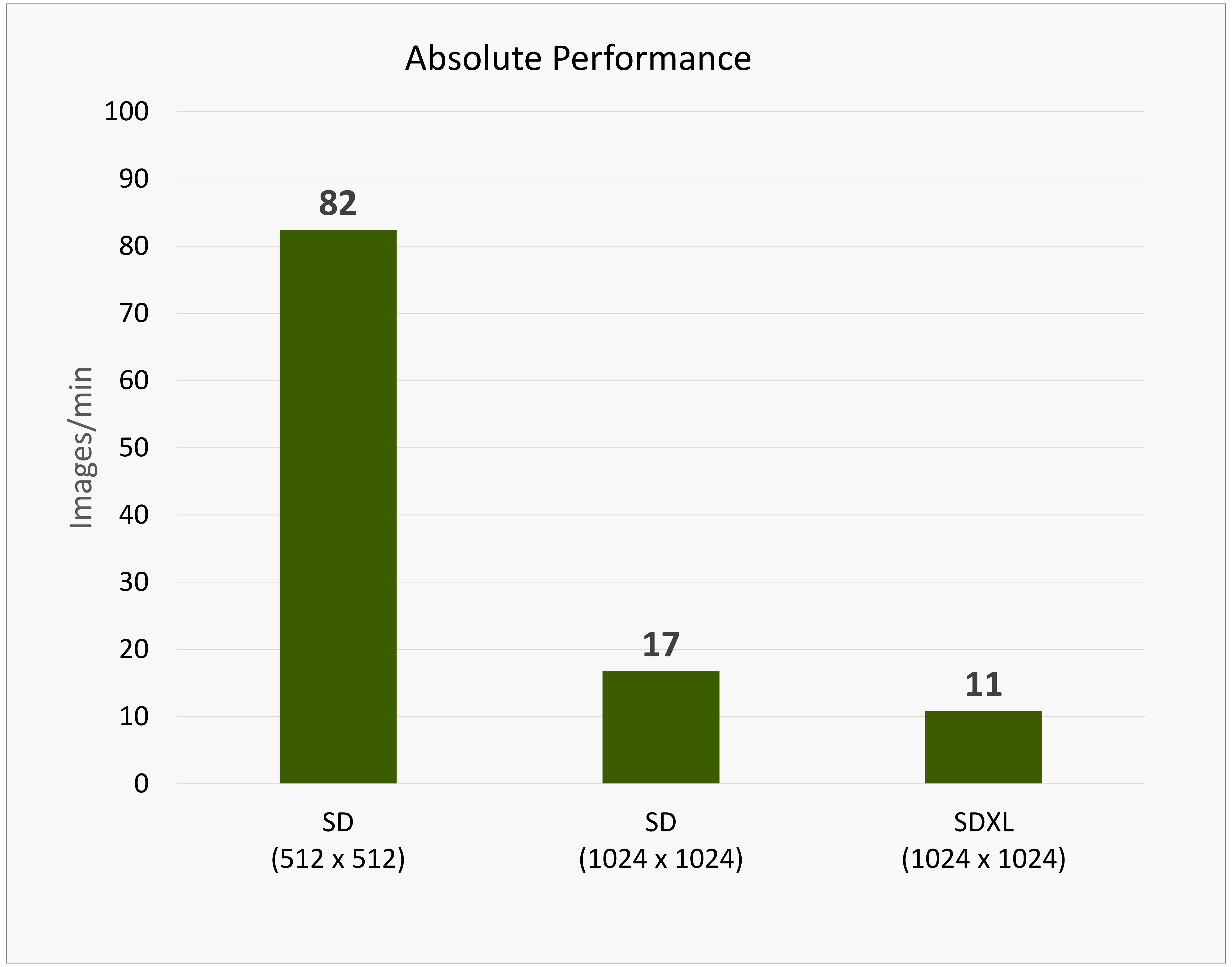


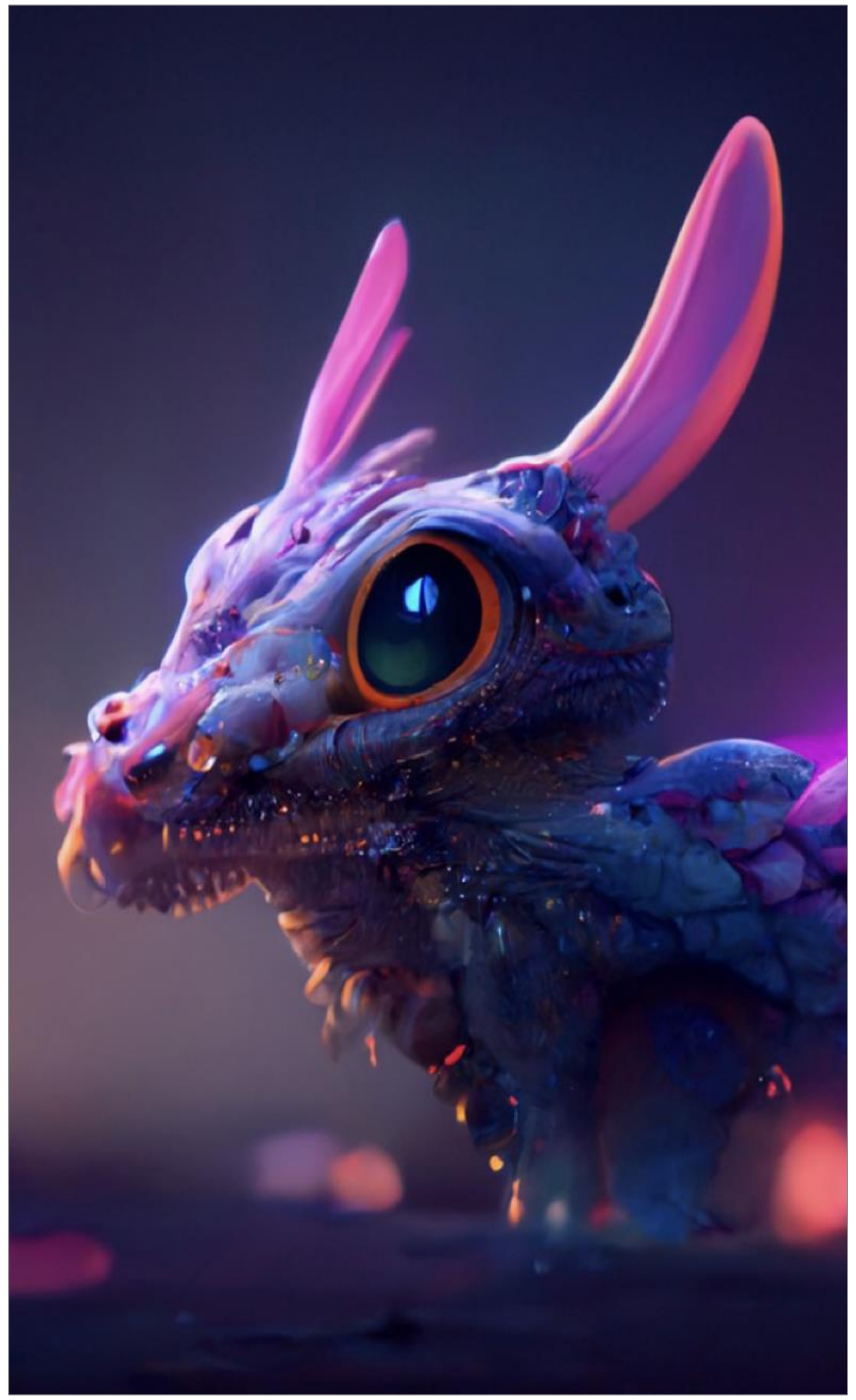
Dual-Slot I FHFL I 350W

Preliminary performance, subject to change
1. Fine-Tuning Llama 2-70B SFT, 1 Billion Tokens; 64x L40S (simulated).
2. Llama 2-13B, ISL=225, OSL=20, BS=1, FP8, 1x L40S. 1st Token Latency.
3. Llama 2-7B, 100 Billion Tokens; 64x L40S (simulated).
4. Image Generation, Stable Diffusion v2.1, 50 iterations, 512 x 512 resolution; 1x L40S.
5. Llama 2-13B, ISL=225, OSL=20, BS=1, FP8, 1x L40S. E2E Latency.
6. Concurrent Encoding Streams; 720p30; 1x L40S.

L40S Generates > 82 Images/min for Image Gen AI Inference

Incredible Performance Across Different Image Sizes and Resolutions



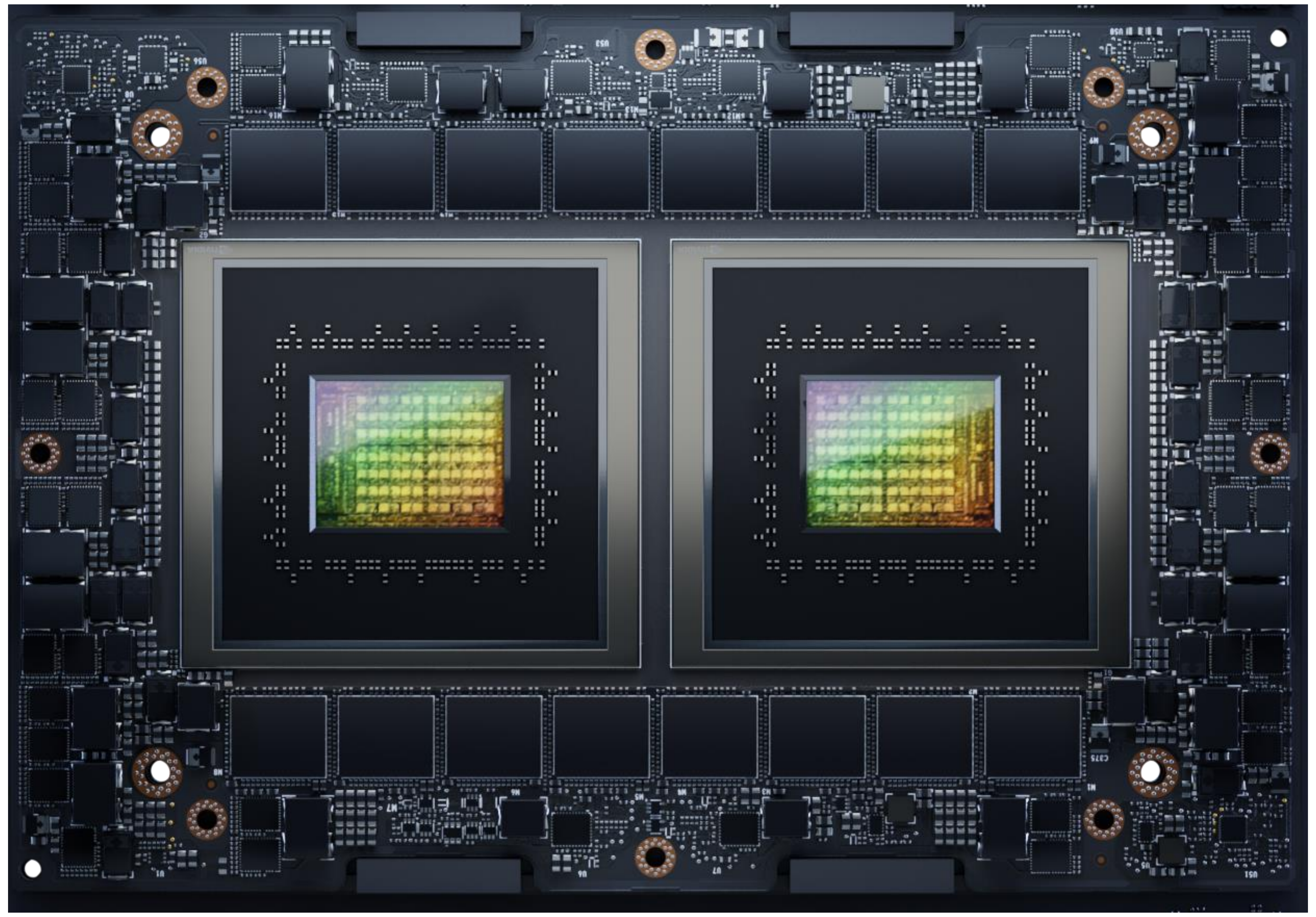


Measured performance
Stable Diffusion v2.1, 50 iterations, TRT 8.6.1, BS:1, FP16
Stable Diffusion XL, 50 iterations, TRT 8.6.1, BS:1, FP16

NVIDIA Grace for Cloud, AI and HPC Infrastructure

Grace CPU Superchip

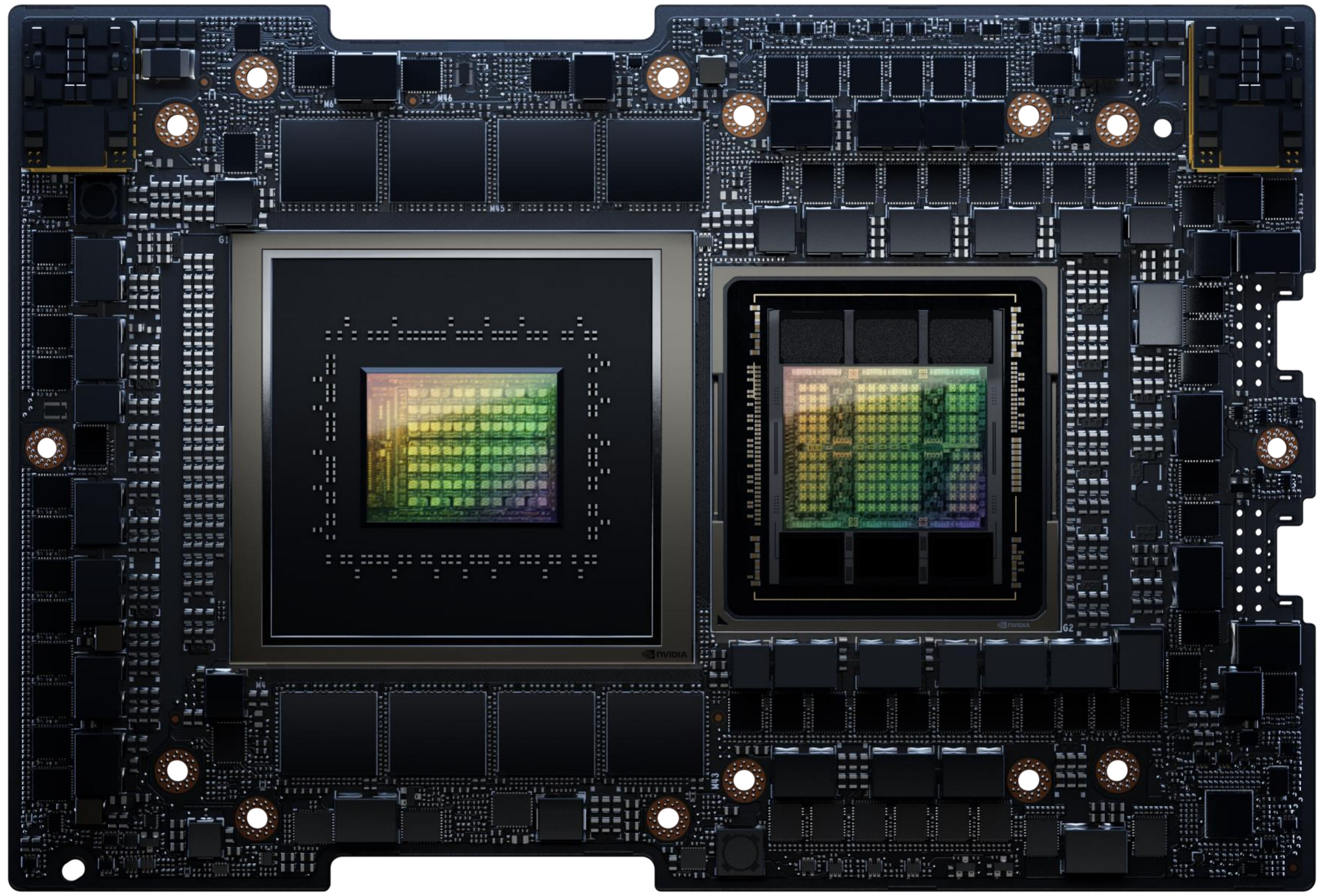
CPU Computing



CPU-based applications where absolute performance, energy efficiency, and data center density matter, such as scientific computing, data analytics, enterprise and hyperscale computing applications

GH200 Grace Hopper Superchip

Large Scale AI & HPC



Accelerated applications where CPU performance and system memory size and bandwidth are critical; tightly coupled CPU & GPU for flagship AI & HPC. Most versatile compute platform for scale out.

NVIDIA GH200 Grace Hopper Superchip

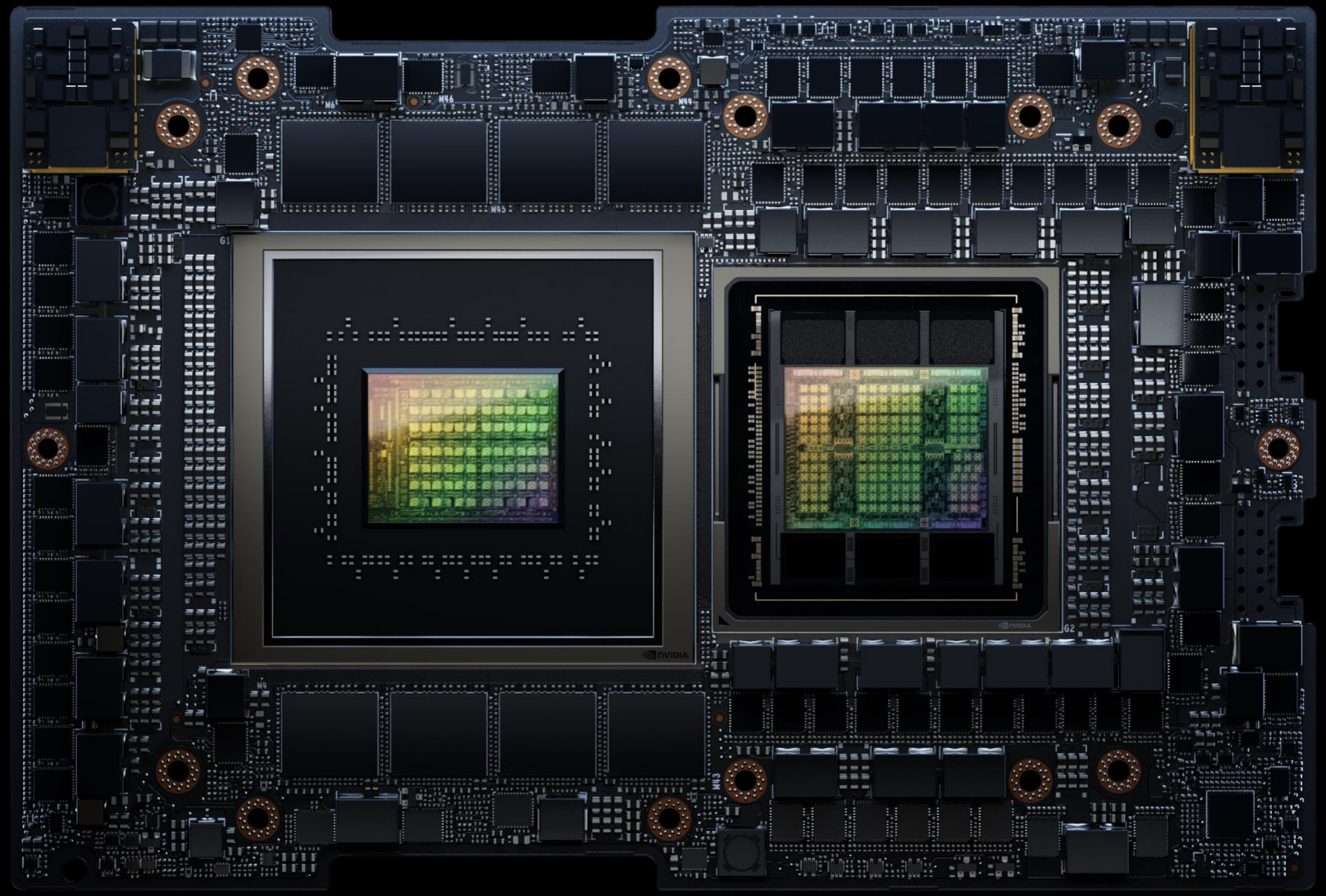
Built for the New Era of Accelerated Computing and Generative AI

Most versatile compute
Best performance across CPU, GPU or memory intensive applications

Easy to deploy and scale out
1 CPU:1 GPU node simple to manage and schedule for for HPC, enterprise, and cloud

Best Perf/TCO for diverse workloads
Maximize data center utilization and power efficiency

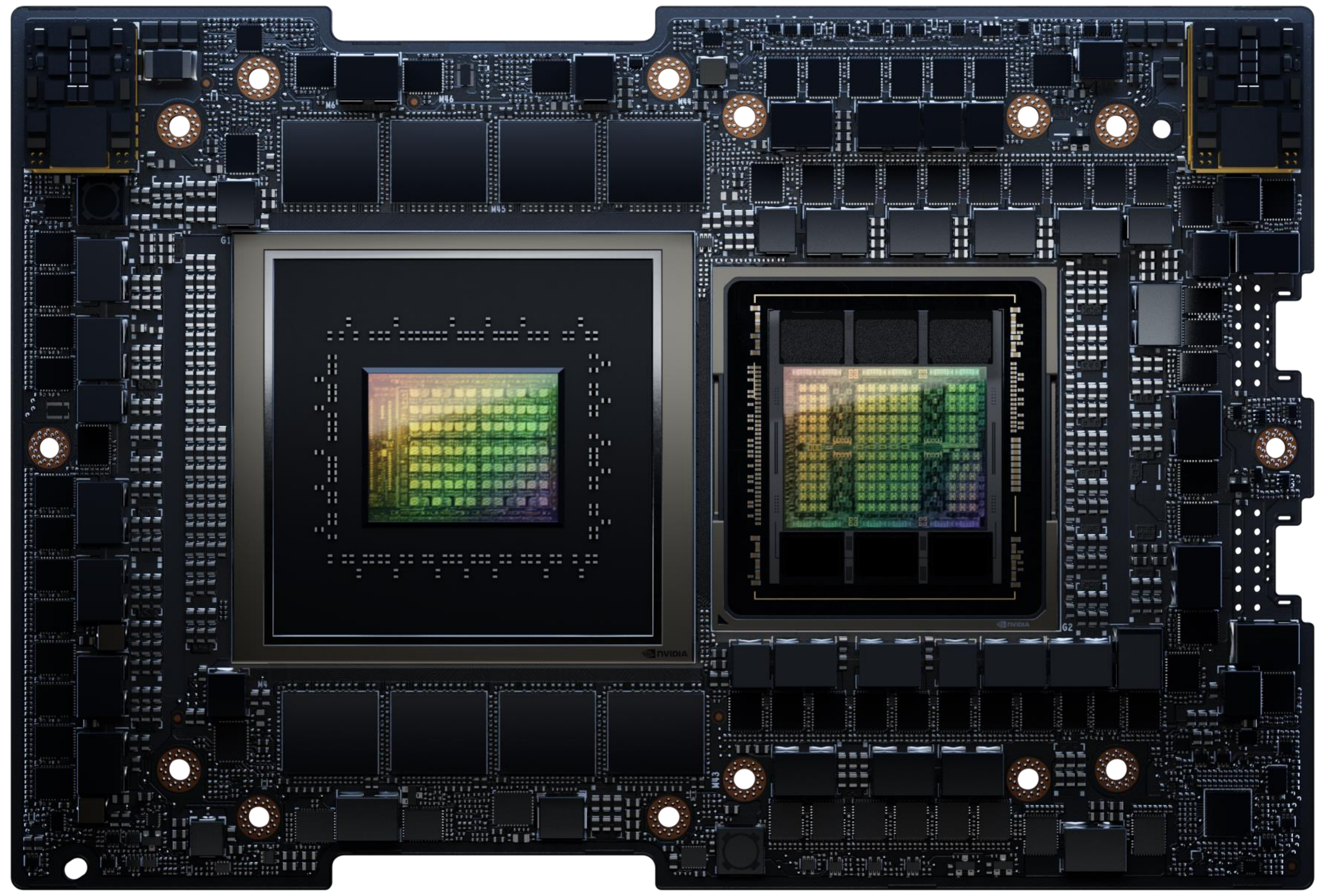
Continued Innovation
Grace and Hopper-Next in 2024



900GB/s NVLink-C2C | 624GB High-Speed Memory
4 PF AI Perf | 72 Arm Cores

NVIDIA GH200 Grace Hopper Superchip

Processor For The Era of Accelerated Computing And Generative AI

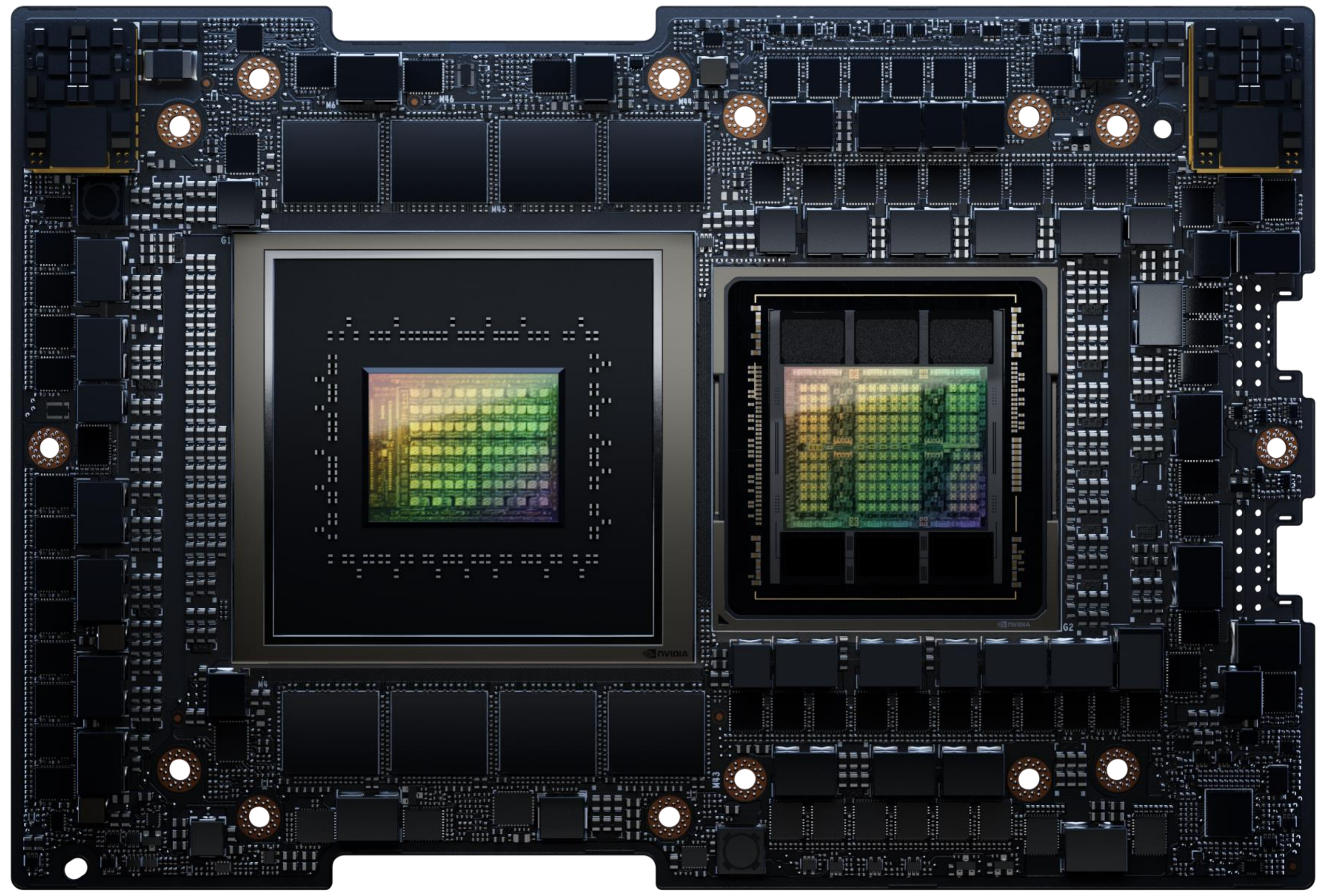


72 Core Grace CPU | 4 PFLOPS Hopper GPU
96 GB HBM3 | 4 TB/s | 900 GB/s NVLink-C2C

- 7X bandwidth to GPU vs PCIe Gen 5
- Combined 576 GB of fast memory
- 1.2x capacity and bandwidth vs H100
- Full NVIDIA Compute Stack

GH200 with HBM3

Available for order

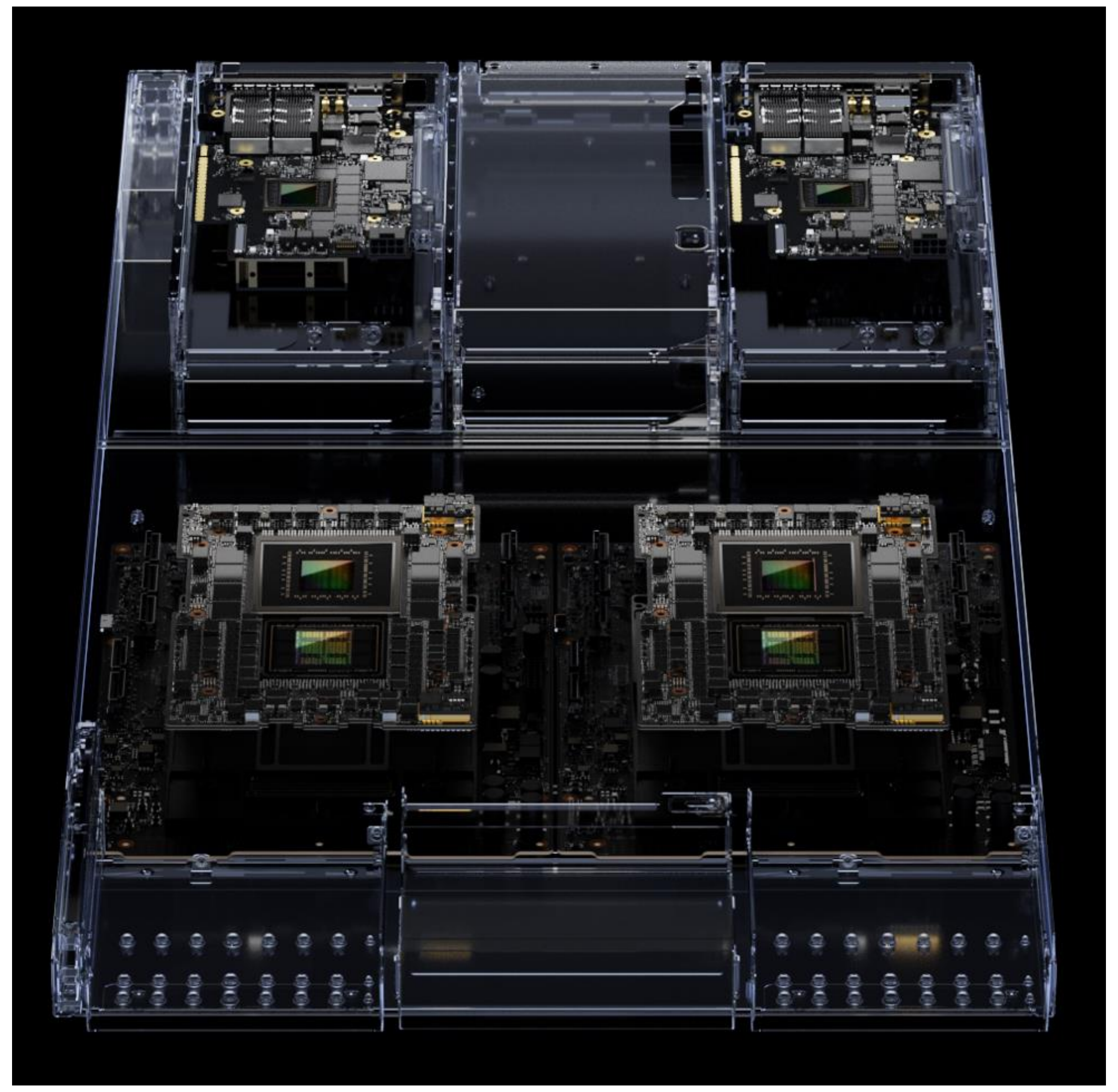


72 Core Grace CPU | 4 PFLOPS Hopper GPU
144 GB HBM3e | 5 TB/s | 900 GB/s NVLink-C2C

- World's first HBM3e GPU
- Combined 624 GB of fast memory
- 1.7x capacity and 1.5x bandwidth vs H100
- Full NVIDIA Compute Stack

GH200 with HBM3e

Available Late Q2 2024



144 Core Grace CPU | 8 PFLOPS Hopper GPU
288 GB HBM3e | 10 TB/s | 900 GB/s NVLink-C2C

- Simple to deploy MGX-compatible design
- Combined 1.2 TB fast memory
- 3.5x capacity and 3x bandwidth vs H100
- Full NVIDIA Compute Stack

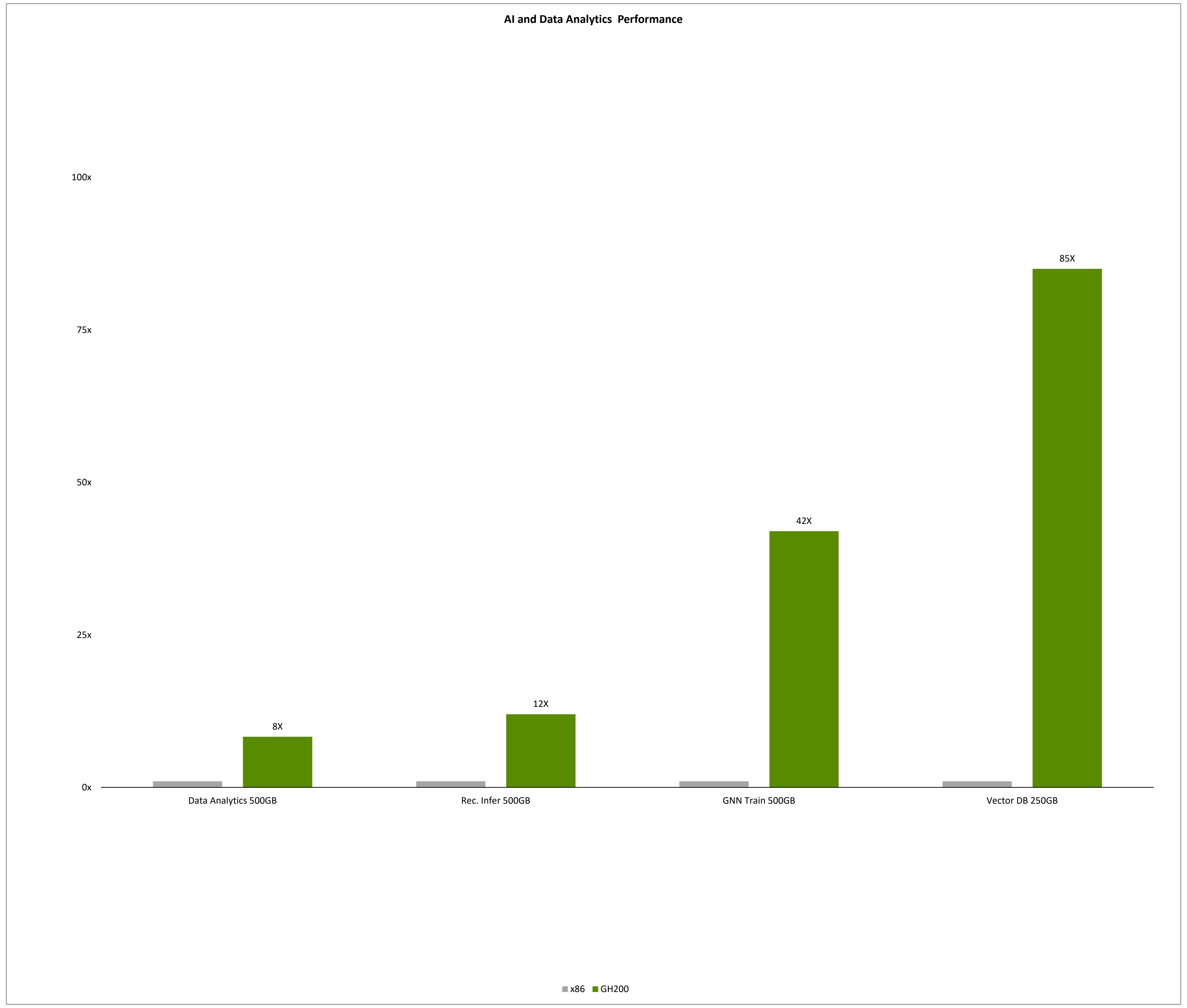
NVLink Dual GH200 System

Available Q2 2024

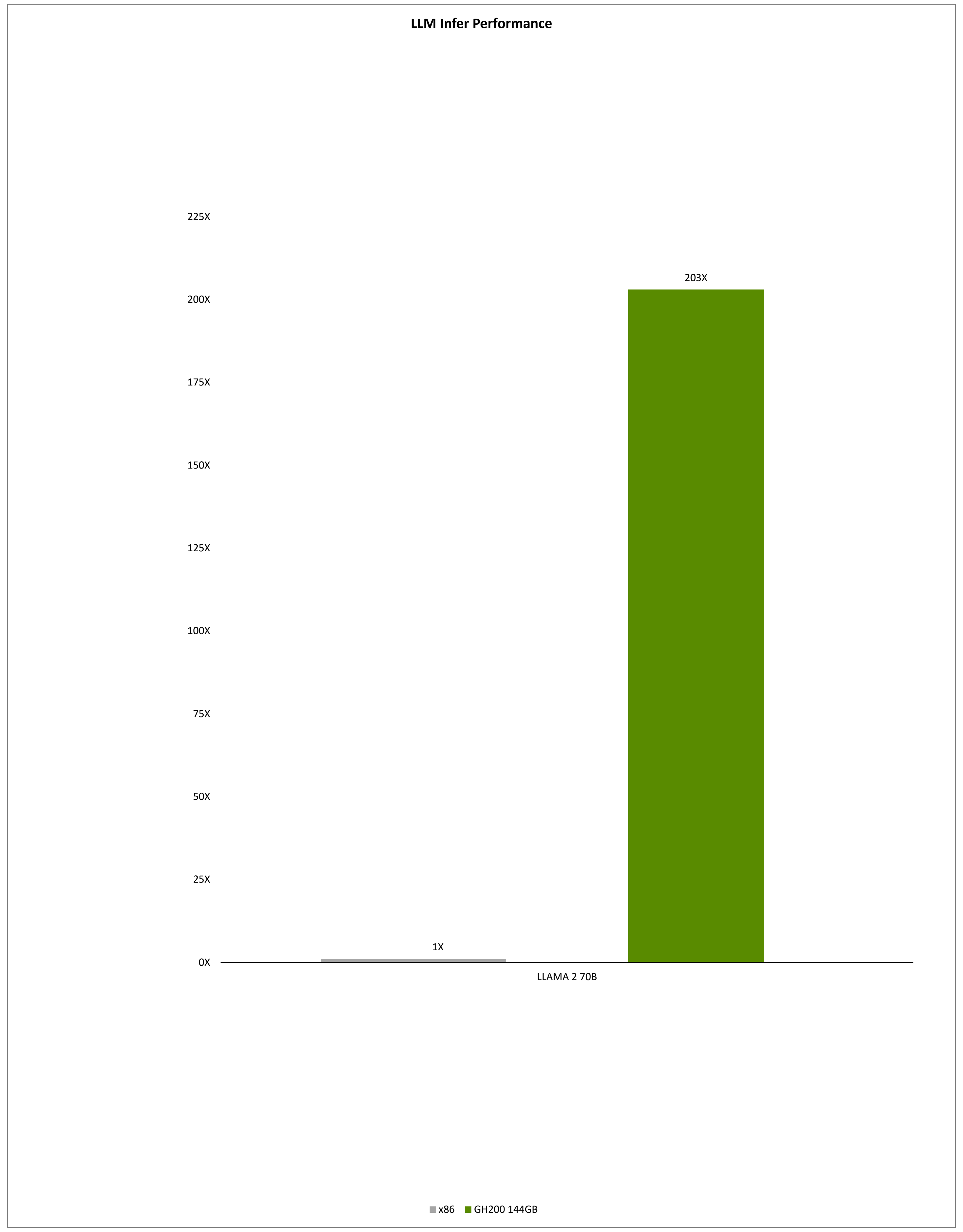
GH200 Grace Hopper AI Inference Platform

Versatile Scale Out with Unmatched Performance

Memory Intensive



GPU Intensive



Use Cases

LLM

Conversational AI
Domain Knowledge

Recommender Systems

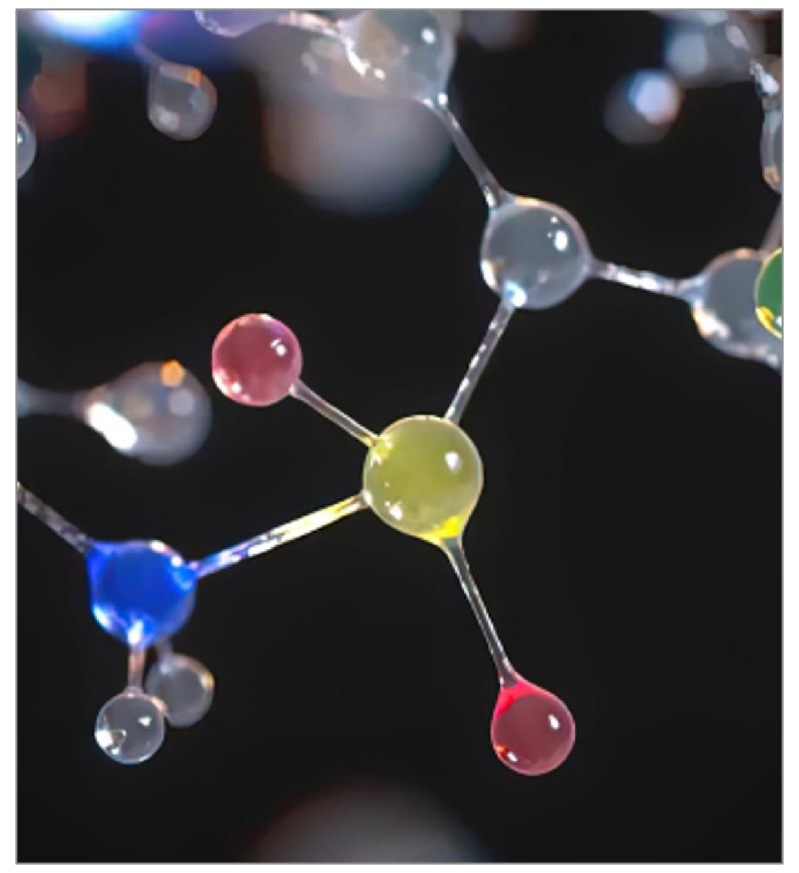
eCommerce
Personalized Content

Vector Database

Fraud Detection
Drug Discovery

GNN

Computer Vison
Recommenders



x86 CPU is Intel Xeon 8480+, GH200 with HBM3 unless noted
DLRM Rec infer, DA (TPC-H query4) are projection based on scaling kernel measurements
Vector DB search (IVF-PQ) with GH200 144GB, Graph Neural Network (GraphSAGE model) Llama 2 70B (1024 input / 512 output seq length, BS = 128) based on measurements measurements

NVIDIA Networking

NIC

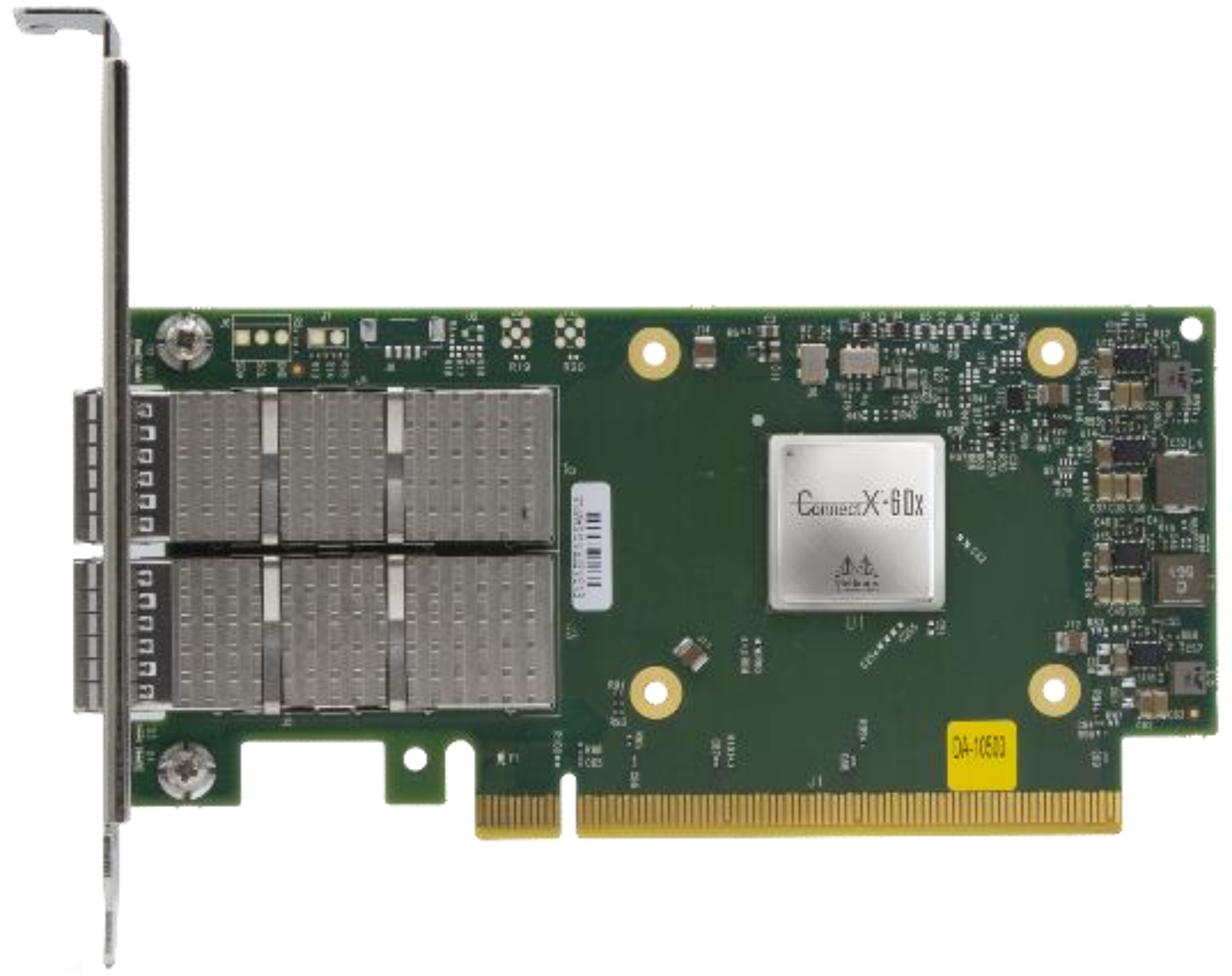
Network Interface Card

A network interface controller (a.k.a. network adapter) is a computer HW component that connects a computer to a computer network.

Early network interface controllers were commonly implemented on expansion cards that plugged into a computer bus…

Modern network interface controllers offer advanced features such as interrupt and DMA interfaces to the host processors, support for multiple receive and transmit queues, …

https://en.wikipedia.org/wiki/Network_interface_controller



2019 ConnectX-6 200gE NIC

2020 ConnectX-7 400gIB/E NIC

2007 ConnectX-EN 10gE NIC – First Ethernet NIC

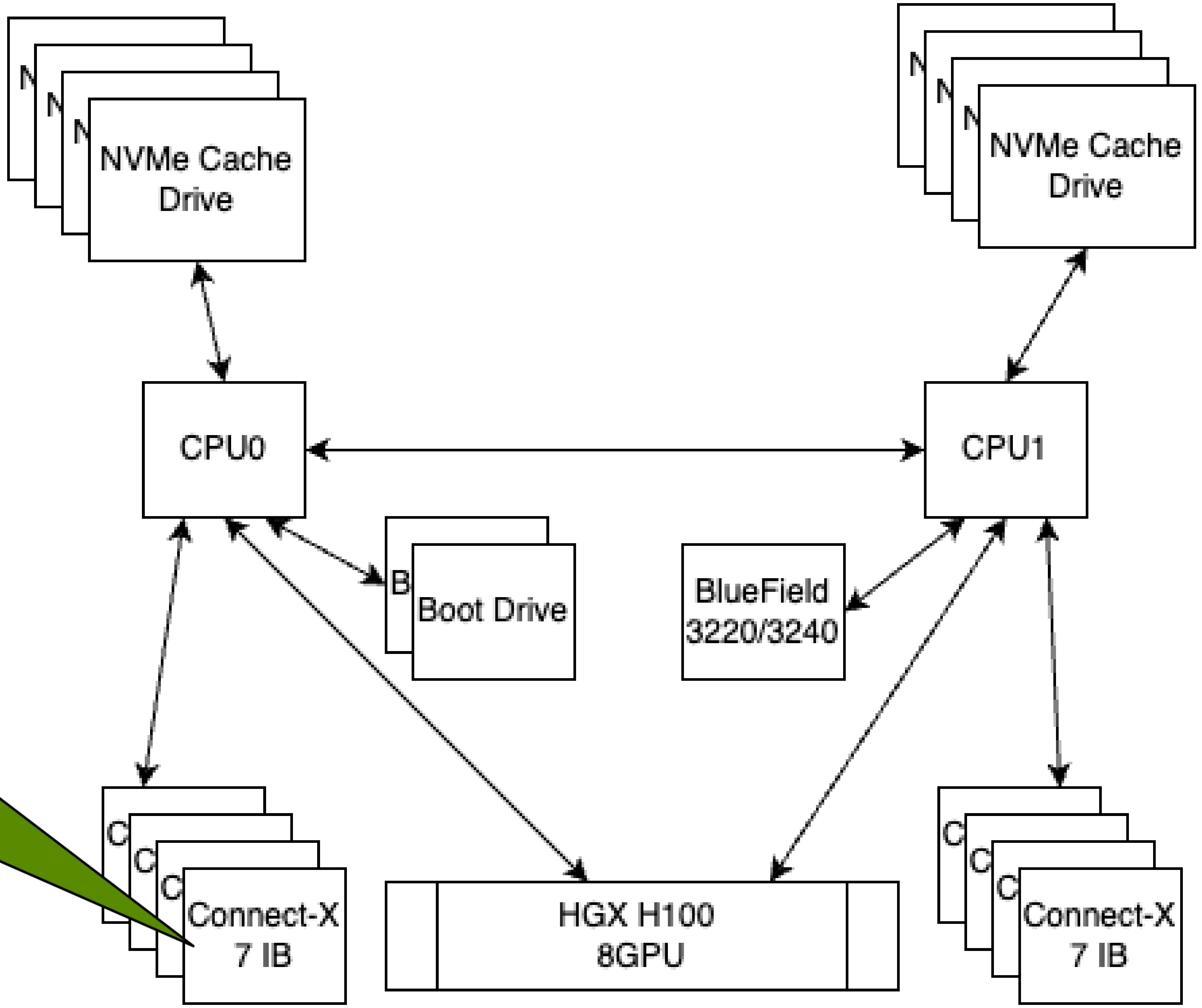
2001 InfiniBridge 10g IB HCA – Mellanox First
Host Channel Adapter (HCA)

Recommended Compute Nodes

NVIDIA DGX Systems & Qualified and Certified HGX H100 Servers

NVIDIA ConnectX-7 Networking Adapters

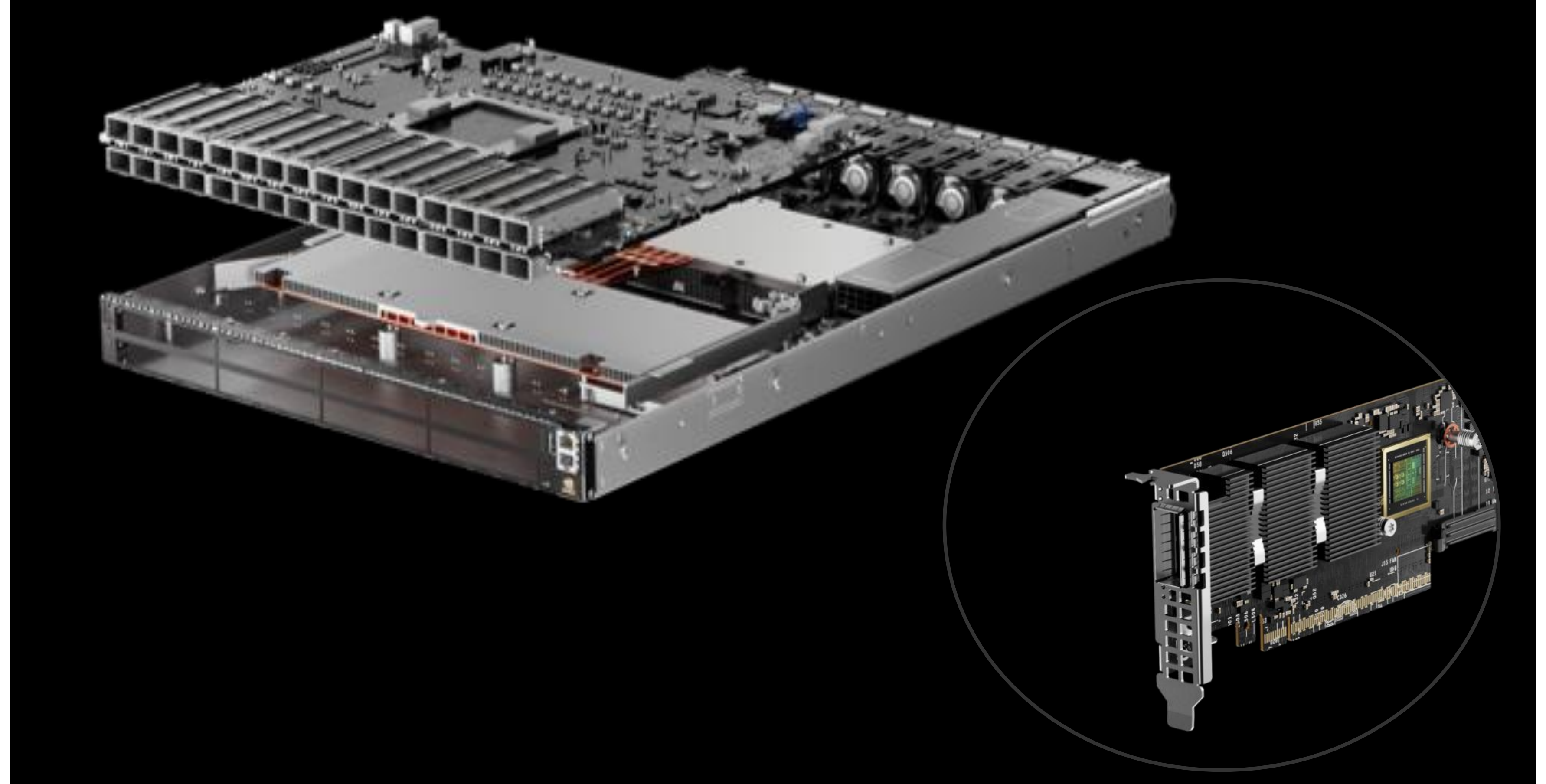
The ConnectX®-7 networking adapter is the latest ConnectX® line. It can provide 25/50/100/200/400Gb/s of throughput. NVIDIA recommended systems use the single-port ConnectX-7® to provide best performance in deployments with NDR. Specifications are available here.



NVIDIA Networking Platforms

Accelerated Networking Solutions for the Era of AI

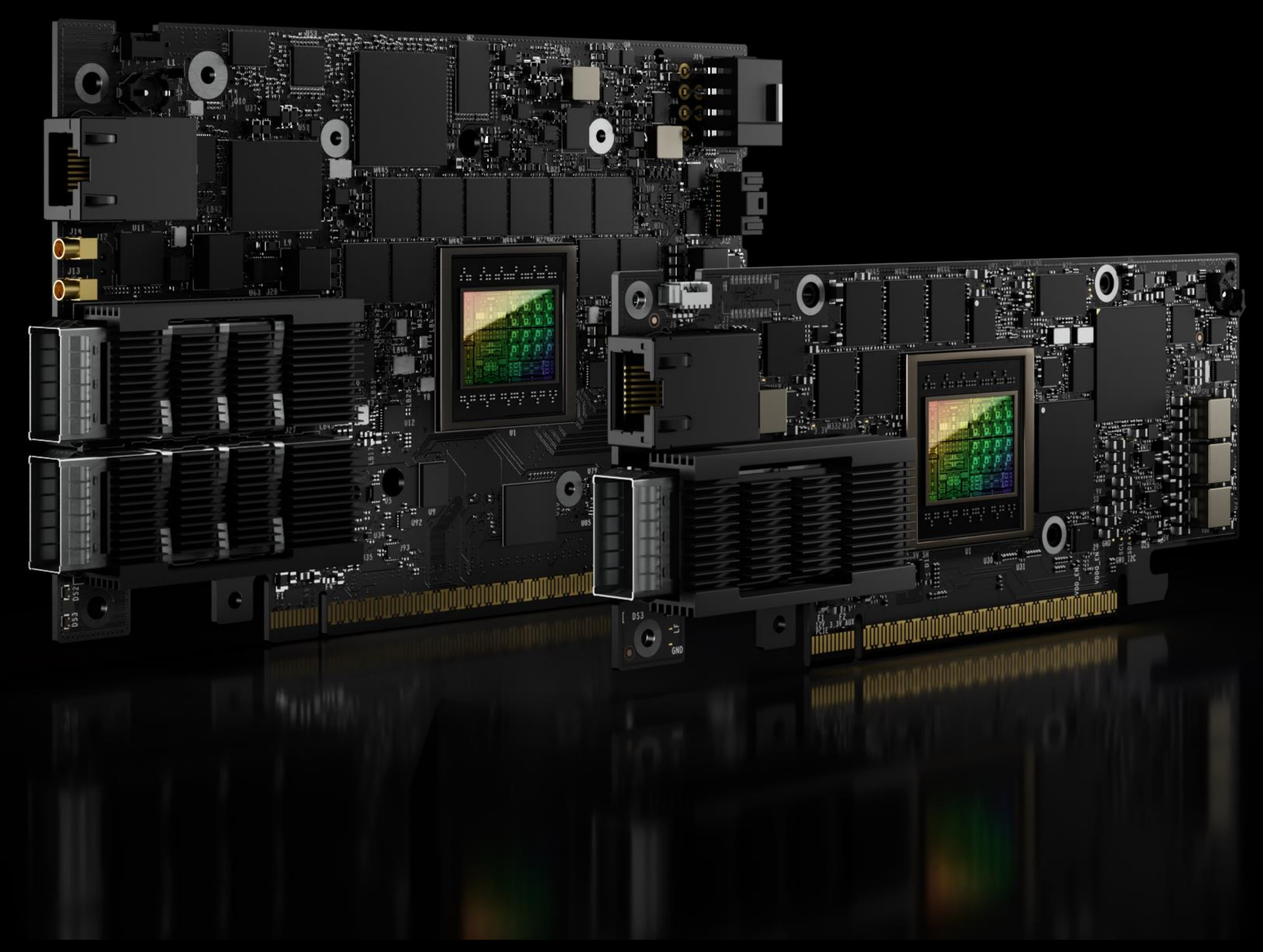
Quantum-2 InfiniBand



Supercomputing Networking Platform

AI Factories and Cloud-Native Supercomputing

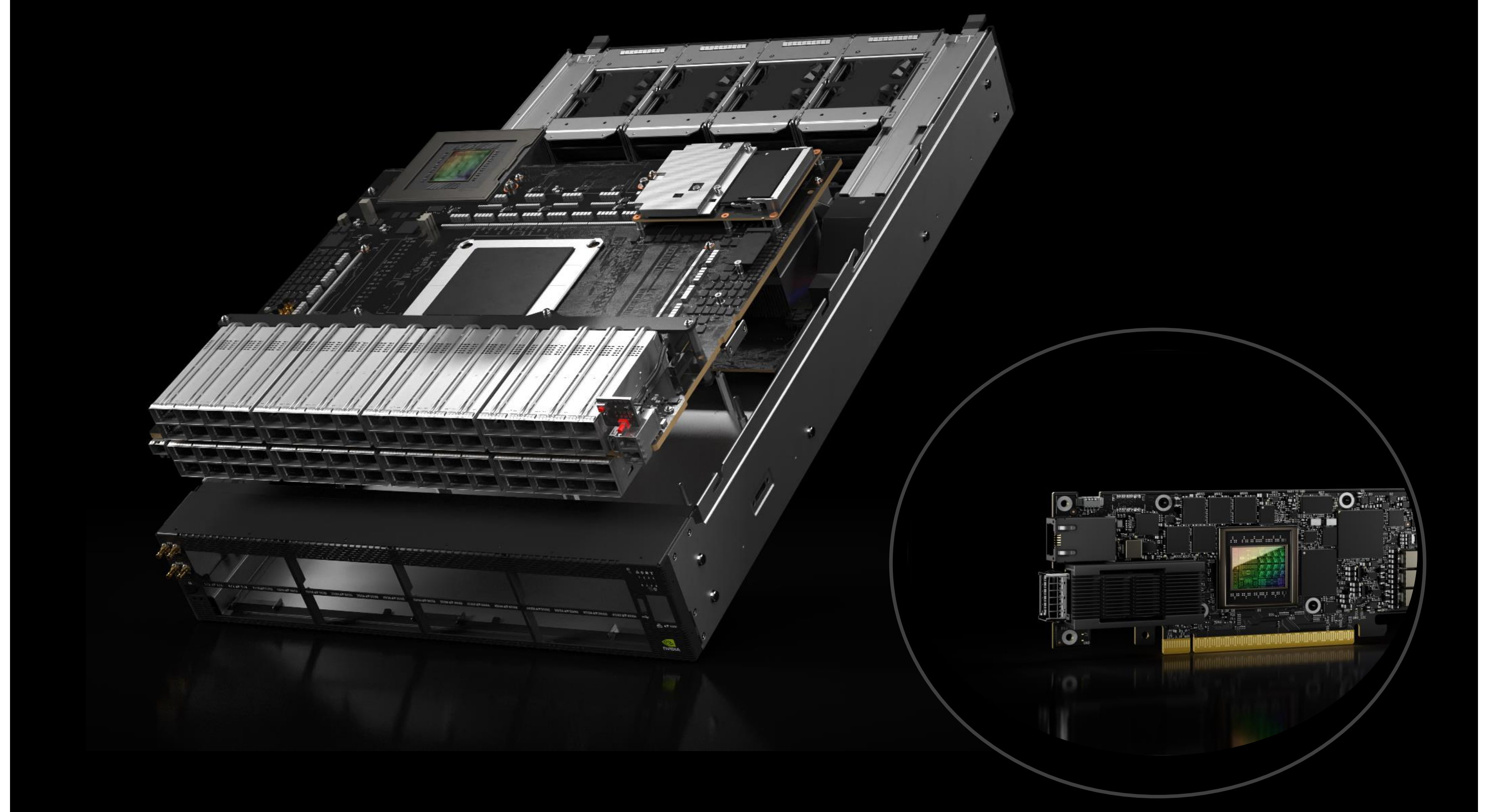
BlueField-3 Networking



Infrastructure Compute Platform

Offload, Accelerate, and Isolate Data Center Infrastructure

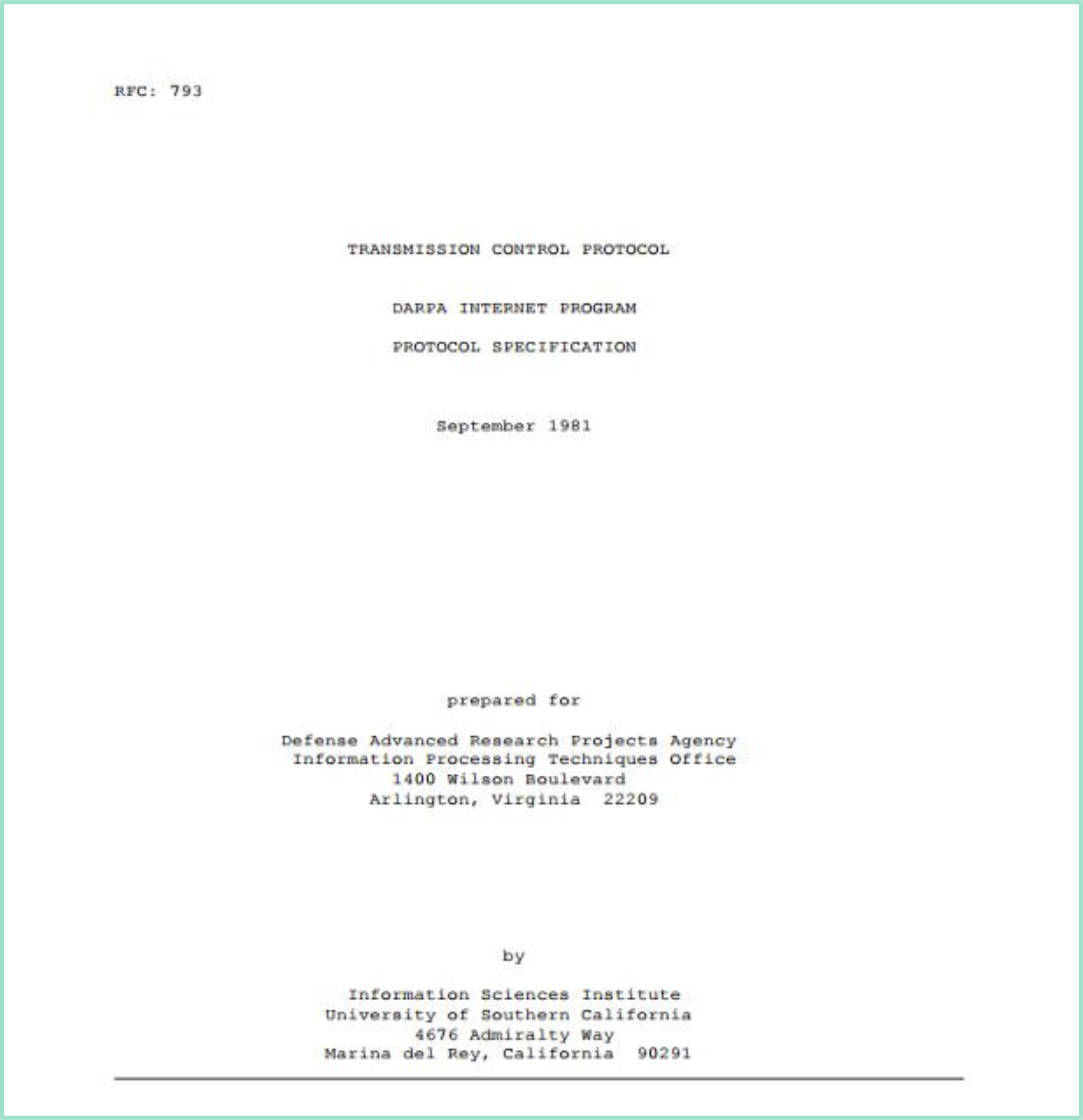
Spectrum-X Ethernet



Hyperscale Networking Platform

Purpose-built Ethernet Networking for AI Clouds

TCP and Sockets



RFC: 793

TRANSMISSION CONTROL PROTOCOL

DARPA INTERNET PROGRAM

PROTOCOL SPECIFICATION

September 1981

prepared for

Defense Advanced Research Projects Agency
Information Processing Techniques Office
1400 Wilson Boulevard
Arlington, Virginia 22209

by

Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, California 90291

RFC 793 – Transmission Control Protocol (TCP)

```
SocketFD = socket(PF_INET, SOCK_STREAM, IPPROTO_TCP);
 if (SocketFD == -1) {
   perror("cannot create socket");
   exit(EXIT_FAILURE);
 }

 memset(&sa, 0, sizeof sa);

 sa.sin_family = AF_INET;
 sa.sin_port = htons(1100);
 res = inet_pton(AF_INET, "192.168.1.3", &sa.sin_addr);

 if (connect(SocketFD, (struct sockaddr *)&sa, sizeof sa) == -1) {
   perror("connect failed");
   close(SocketFD);
   exit(EXIT_FAILURE);
 }

 n = write(SocketFD,buffer,strlen(buffer));
 if (n < 0)
      error("ERROR writing to socket");

 n = read(SocketFD,buffer,MAX_BUFFER_SIZE);
 if (n < 0)
      error("ERROR reading from socket");

 shutdown(SocketFD, SHUT_RDWR);

 close(SocketFD);
```

BSD/POSIX Sockets

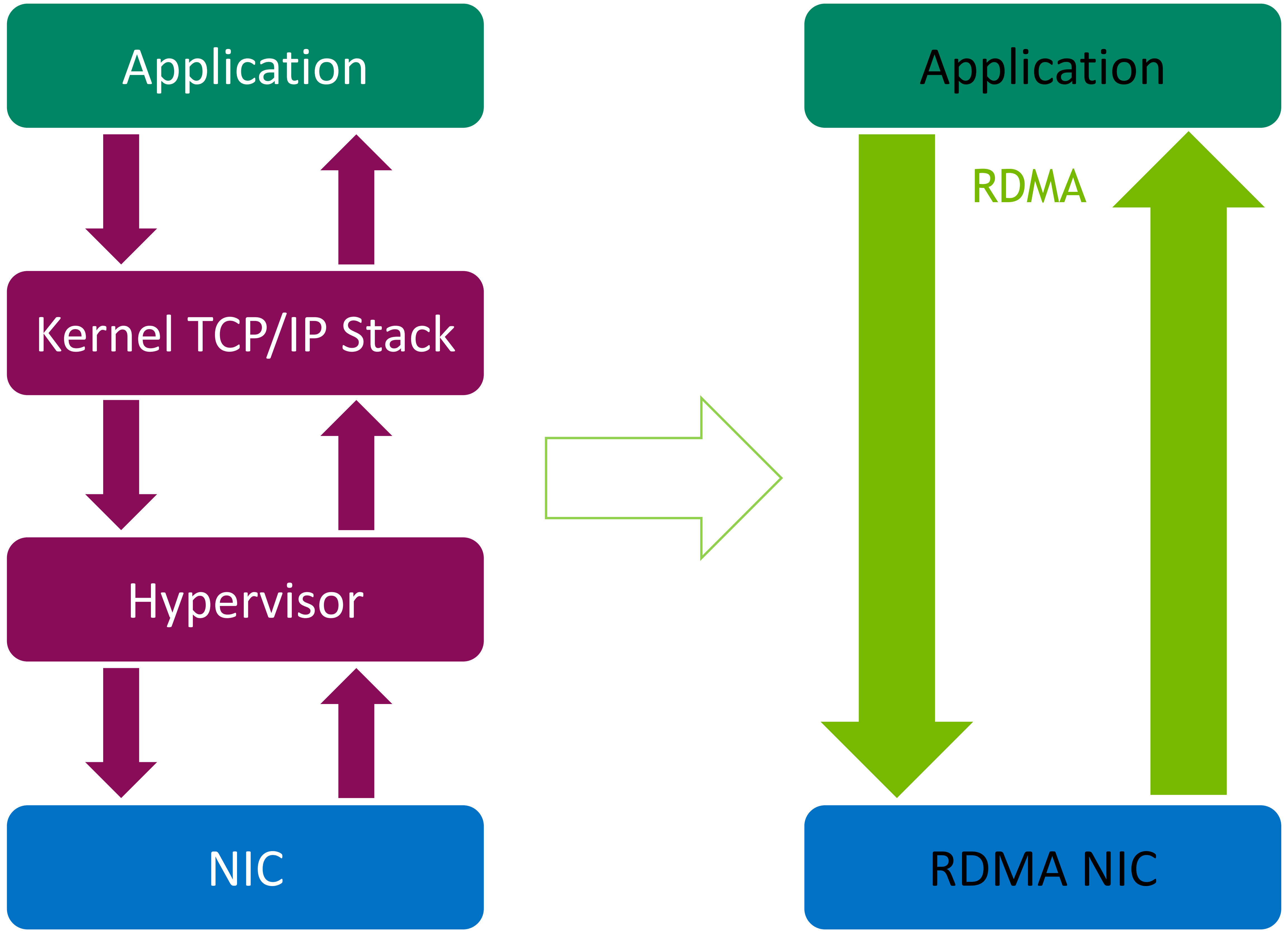
RDMA – Remote DMA

InfiniBand and RoCE (RDMA over Converged Ethernet)

- Messages send/receive, remote DMA and remote atomics
- Hardware transport
- Kernel and hypervisor bypass

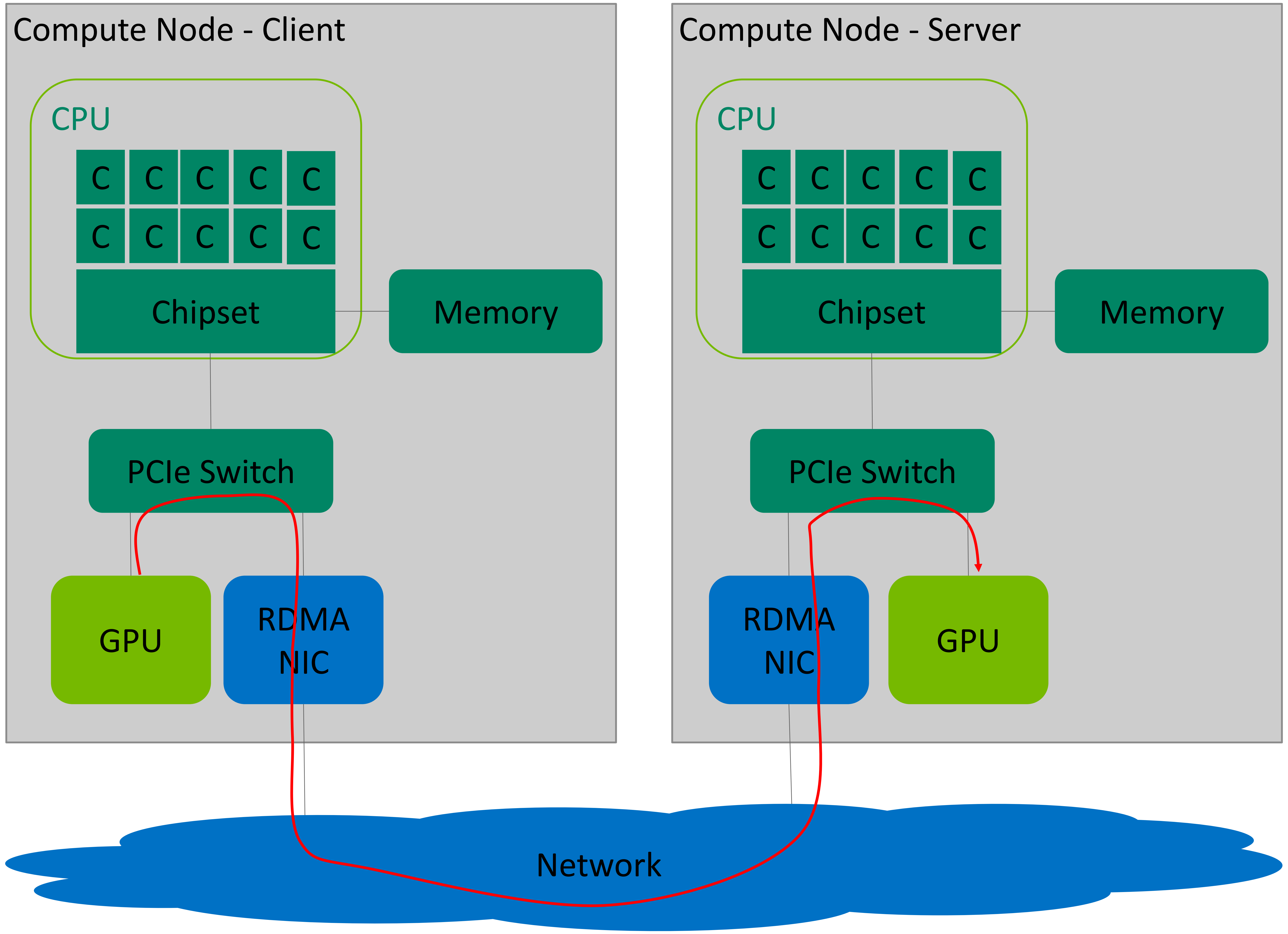
Advantages

- Lower latency – 10us → 700ns
- Higher message rate – 215M messages/s
- Lower CPU utilization – 0%



RDMA enables Peer to Peer

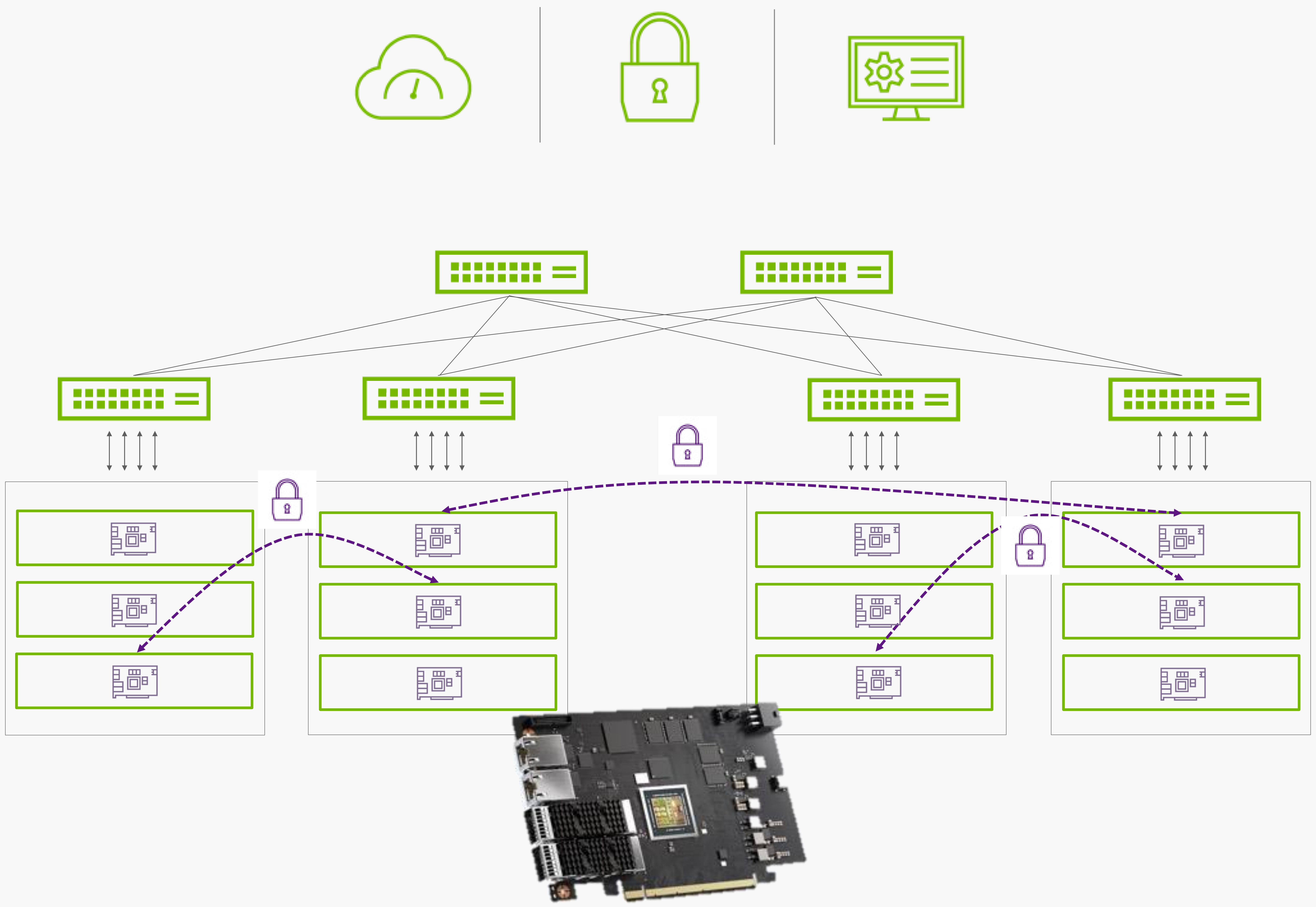
GPU Direct



DOCA Host-based networking

Focus areas

- Controller-less VPC networking for BMaaS
- Accelerated Routing and EVPN on Host
- Simplifying underlay network fabric
  - End-to-end IP Fabric from the host
  - L3 ECMP to replace proprietary LAG/MLAG
- Advanced ToR switch features on DPU
  - Reduce risk with Whitebox deployments (SONiC, Switchdev)
  - Save on 3rd party switch feature licenses
- NVIDIA E2E value-add with Cumulus and HBN

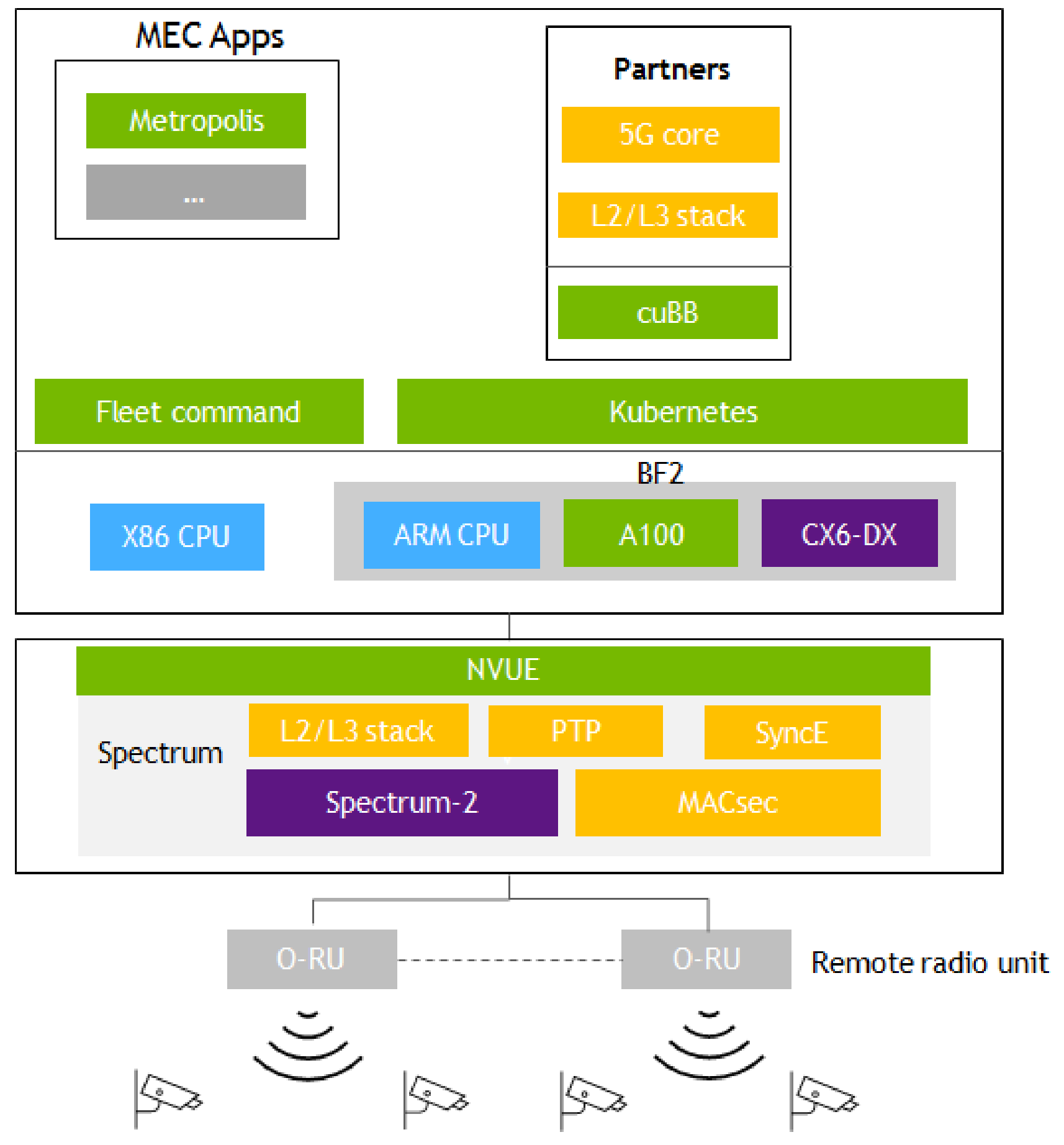


NVIDIA Cumulus Apps on DPU

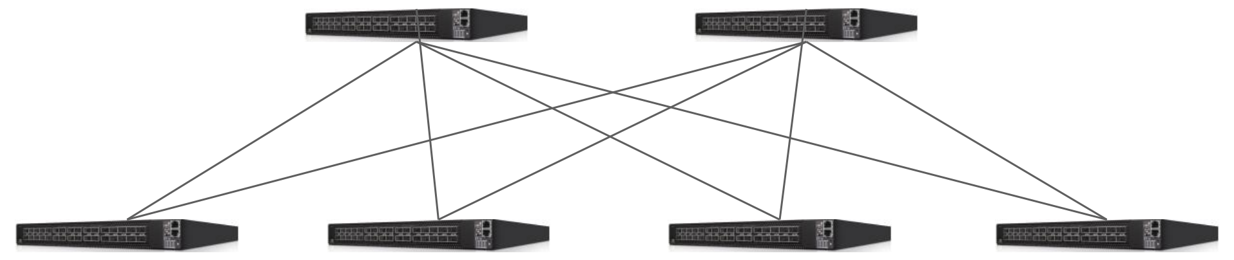
NetQ in Full Stack NVIDIA Solutions

Telemetry Data Collection, Network Monitoring and Troubleshooting

Aerial (EGX POD)

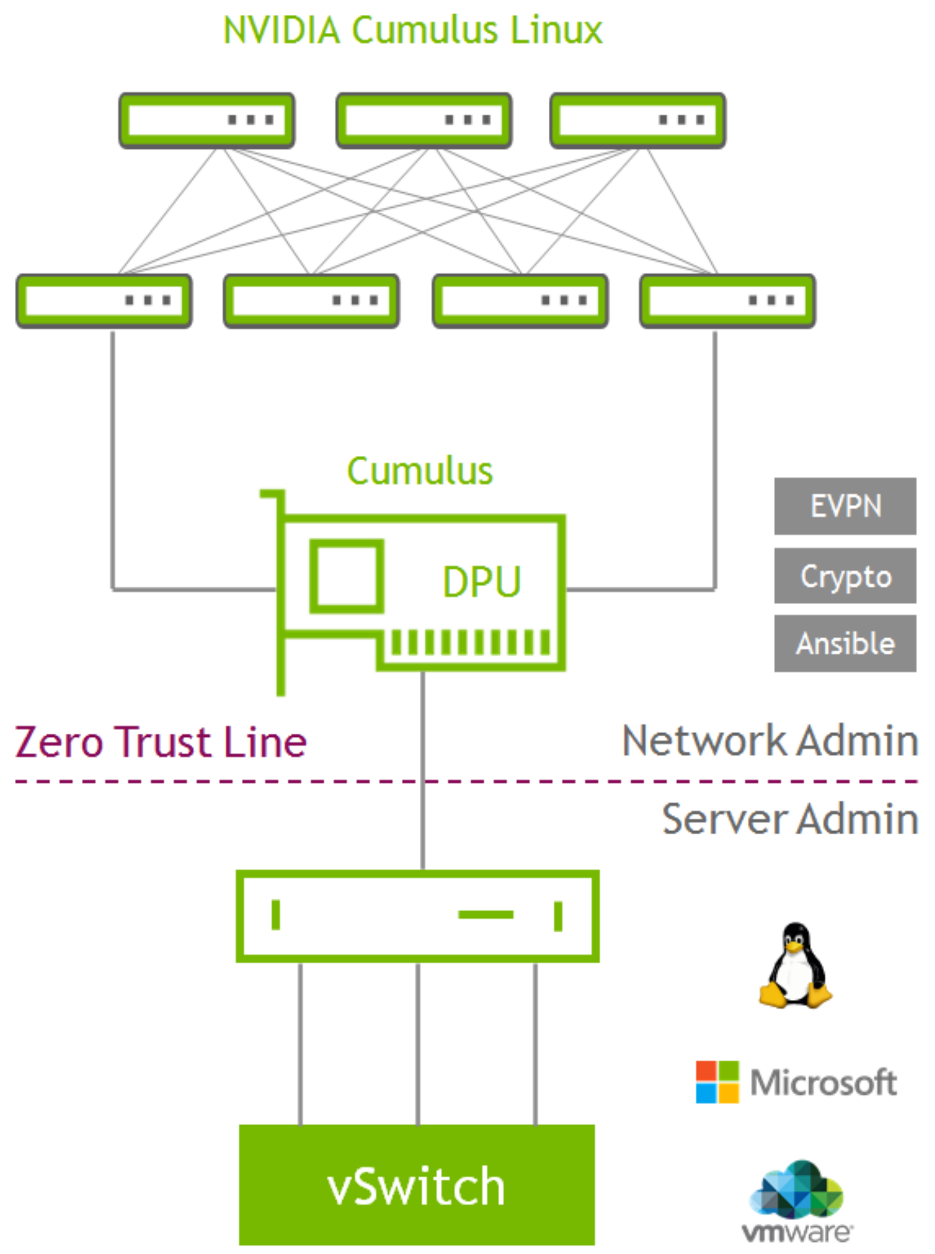


DGX/HGX

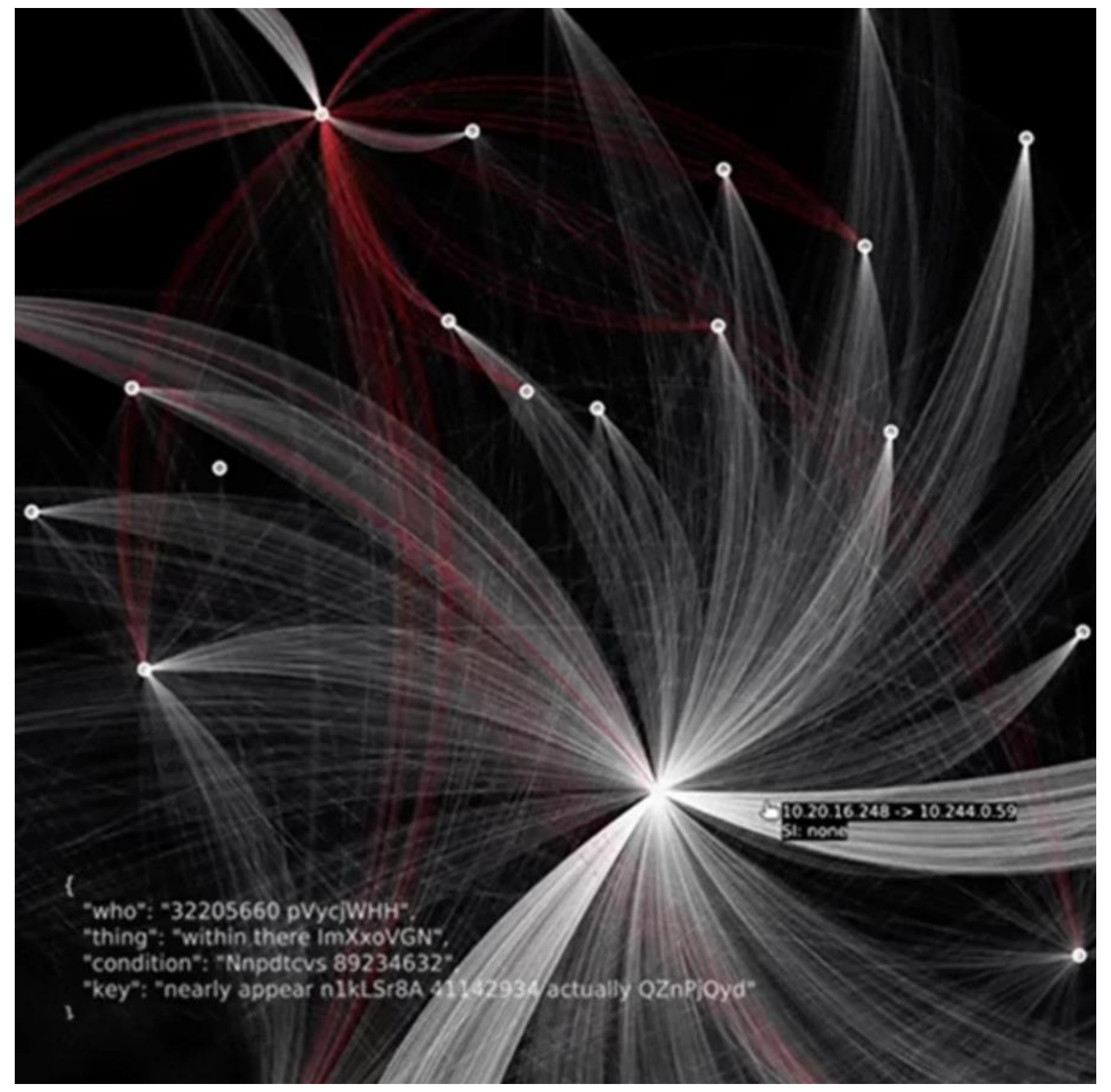


OVX POD / Super POD

Host Based Networking (HBN)



AI Powered by Morpheus Pipeline



```
{
  "who": "32205660 pVycjWHH",
  "thing": "within there ImXxoVGN",
  "condition": "Nnpdtcvs 89234632",
  "key": "nearly appear n1kLSr8A 41142934 actually QZnPjOyd"
}
```

- Switch & DPU inventory
- Validations

- Metrics monitoring
- RoCE monitoring

- WJH and other events
- Trace & flow telemetry

- Network data collection
- APIs for integration

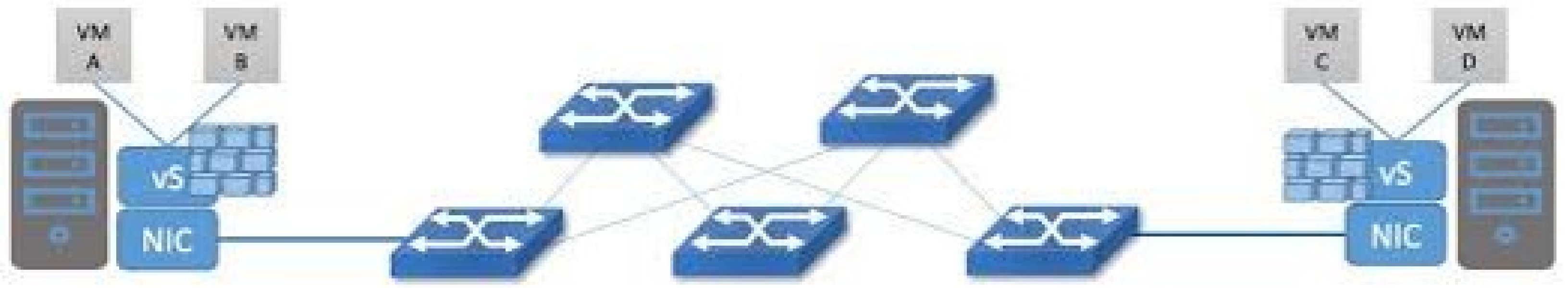
Getting Started with NVIDIA Reference Architectures

Reference Architecture Components

Compute Node

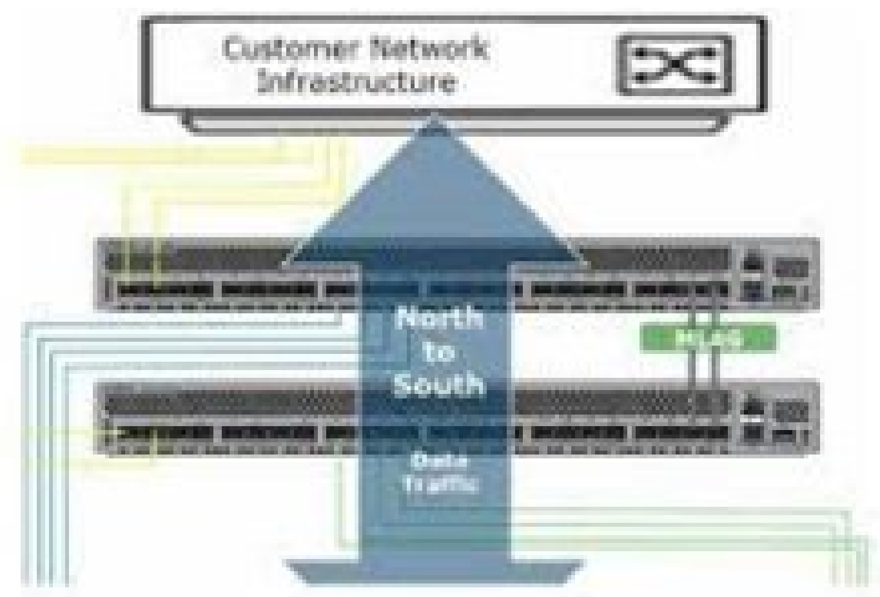
- HGX H100 8 GPUs Compute node design
- Ensures performance within the compute node

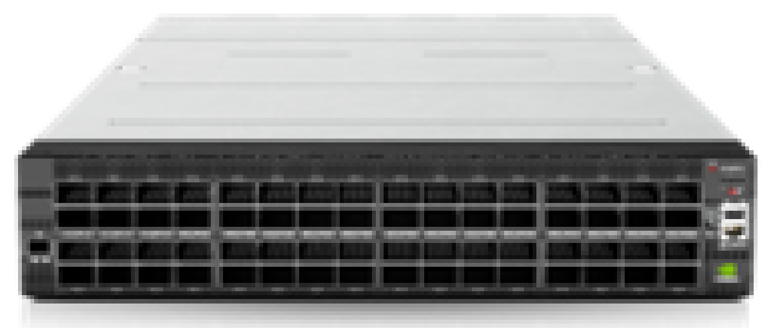
E-W Network



- 8-Rails-optimized, non-blocking fat tree IB topology
- Validated with 3 production generations of SuperPOD
- Highly scalable with large locality groups
- SharpV3 for multi tenancy
- Managed by UFM & BCM
- Easy to validate and operate

Tenant Access Network: N-S



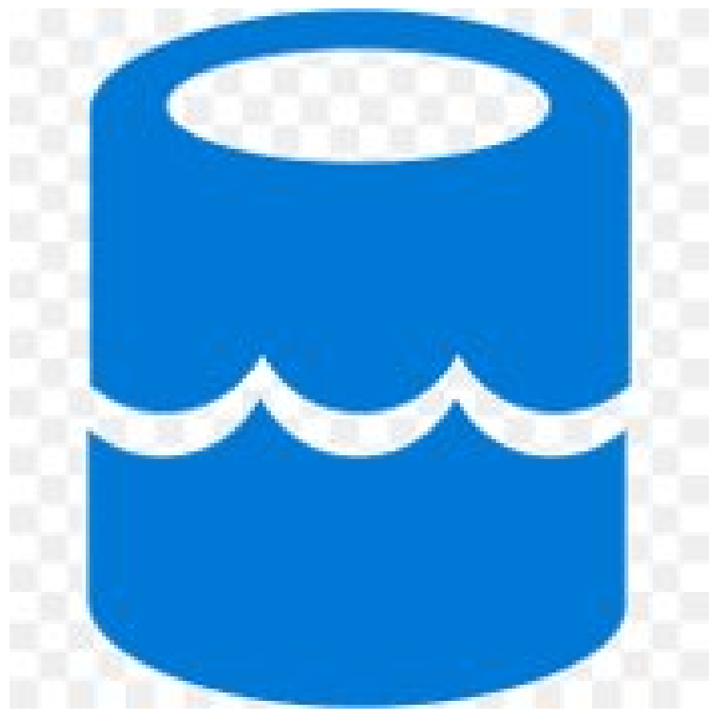


- Each HGX has 2 x 200GbE BF3 connections to Leaf switches
- BF3 enables a true, zero-trust BMaaS
- Infrastructure isolation from the compute workload
- Enhanced security
- Storage offloaded and accelerated by BF3

High Speed Storage (HSS)

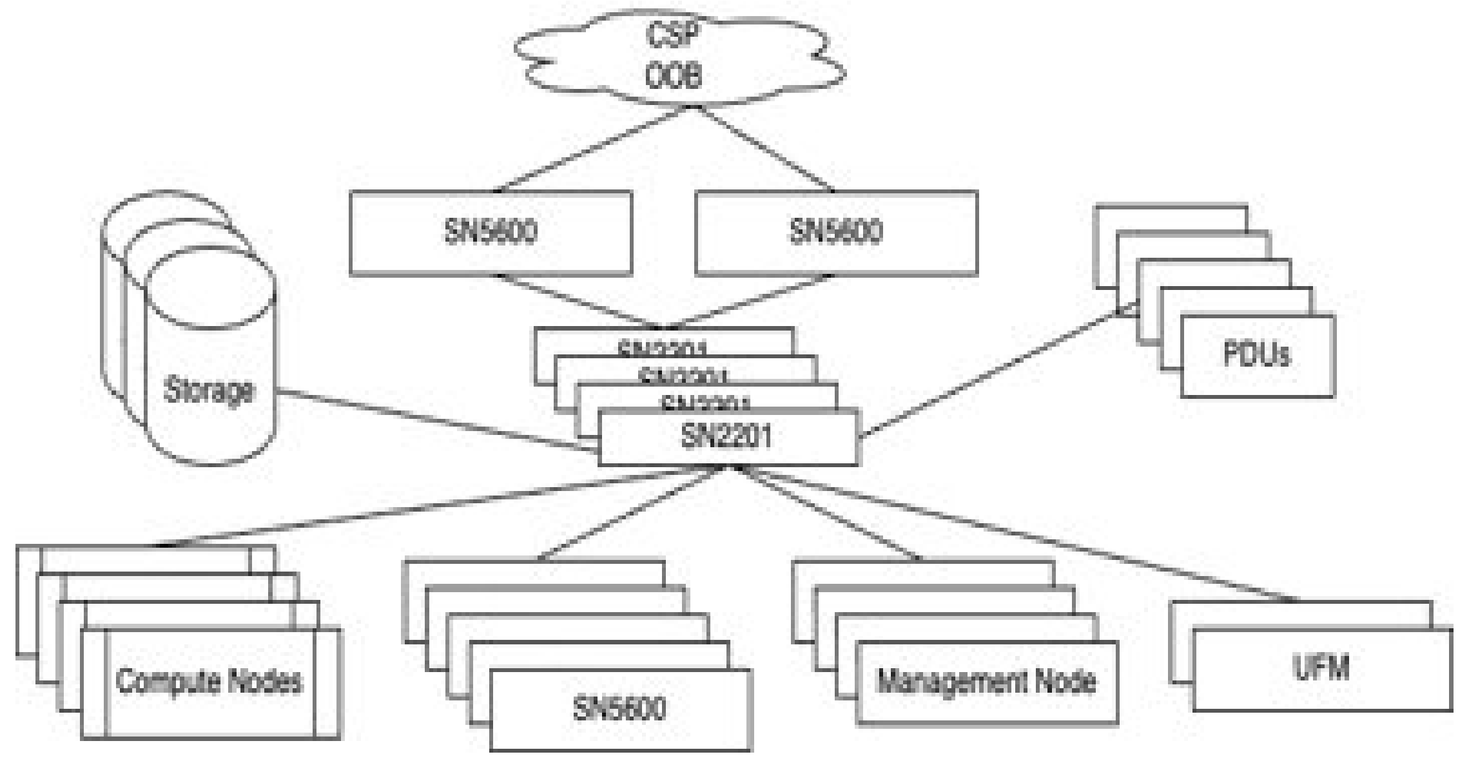
- Clear minimum recommended bandwidth for storage
- Optimized for best performance of LLM Training
- Enabling a choice of storage solutions
- Per POD or centralized HSS placement

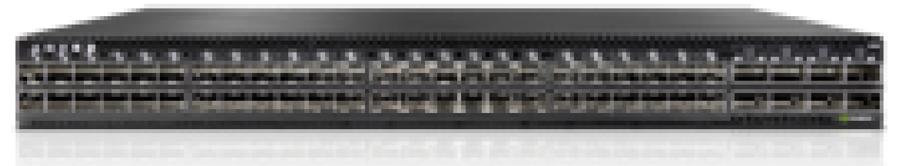
Outer Ring Storage (Data Lake)



- Raw-data storage
- Minimal requirements for a Datalike storage
- Can be a unified solution with HSS

Out Of Band Management





- 1GbE management network
- Providing monitoring and management of all DC devices
- Enabling integration of different APIs between all devices

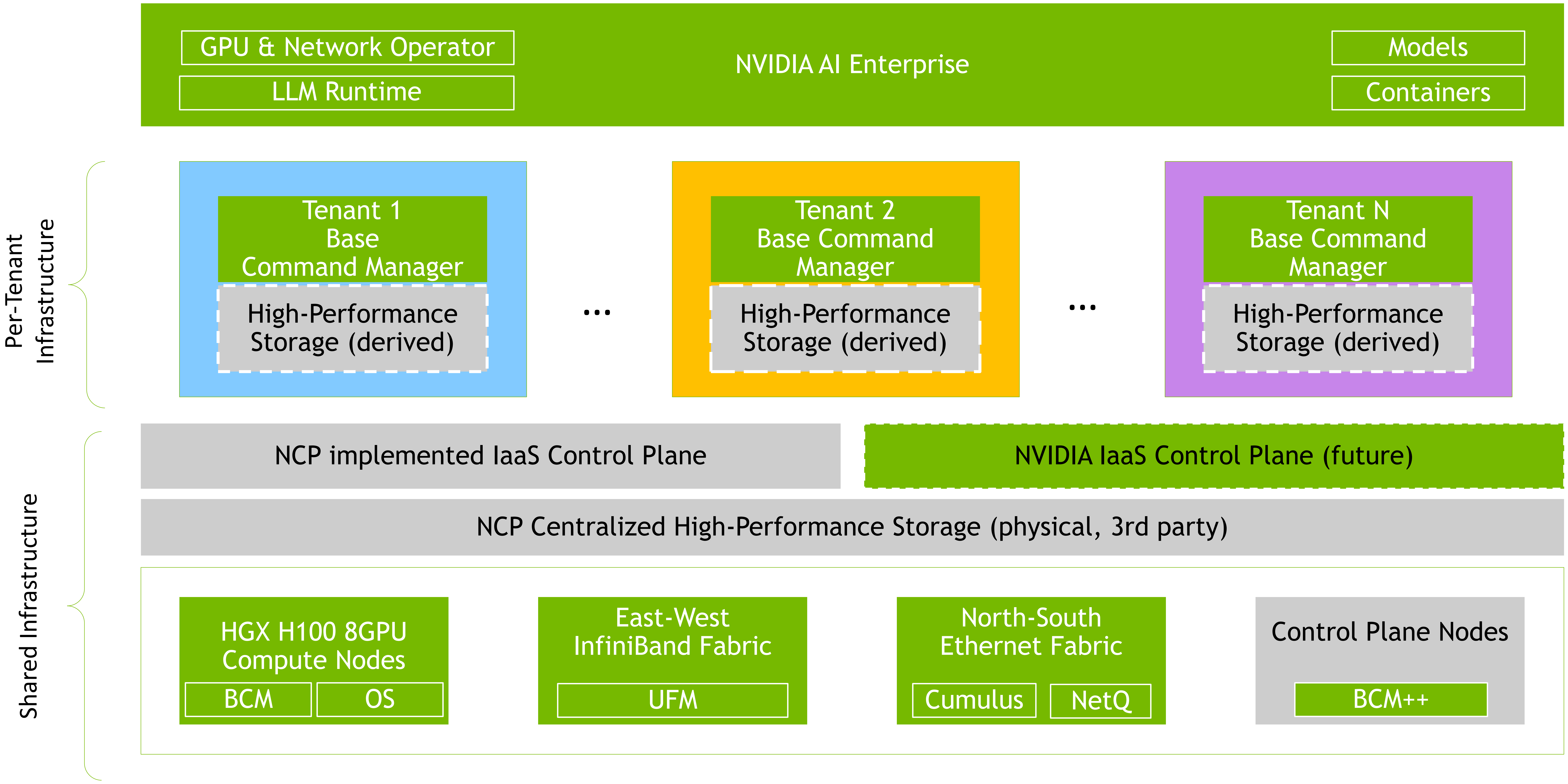
Flagship for AI Training

NVIDIA DGX SuperPOD with H100

Scalable Unit (SU)
Utility Racks
DGX H100 #4
DGX H100 #3
DGX H100 #2
DGX H100 #1
DGX H100 #8
DGX H100 #7
DGX H100 #6
DGX H100 #5
DGX H100 #12
DGX H100 #11
DGX H100 #10
DGX H100 #9
DGX H100 #16
DGX H100 #15
DGX H100 #14
DGX H100 #13
IB Compute Leaf #8
IB Compute Leaf #7
IB Compute Leaf #6
IB Compute Leaf #5
IB Compute Leaf #4
IB Compute Leaf #3
IB Compute Leaf #2
IB Compute Leaf #1
OOB Leaf #1
DGX H100 #20
DGX H100 #19
DGX H100 #18
DGX H100 #17
DGX H100 #24
DGX H100 #23
DGX H100 #22
DGX H100 #21
DGX H100 #28
DGX H100 #27
DGX H100 #26
DGX H100 #25
DGX H100 #31
DGX H100 #30
DGX H100 #29
IB Compute Spine #4
IB Compute Spine #3
IB Compute Spine #2
IB Compute Spine #1
In-band Switch #1
In-band Switch #2
OOB Leaf #2
Compute UFM #2
Compute UFM #1
NFS Home Directory
Management Node #5
Management Node #4
Management Node #3
Management Node #2
Management Node #1
IB Storage Leaf #4
IB Storage Leaf #3
IB Storage Leaf #2
IB Storage Leaf #1
IB Storage Spine #2
IB Storage Spine #1
Storage UFM #2
Storage UFM #1
Storage Mgmt Node
Parallel Filesystem

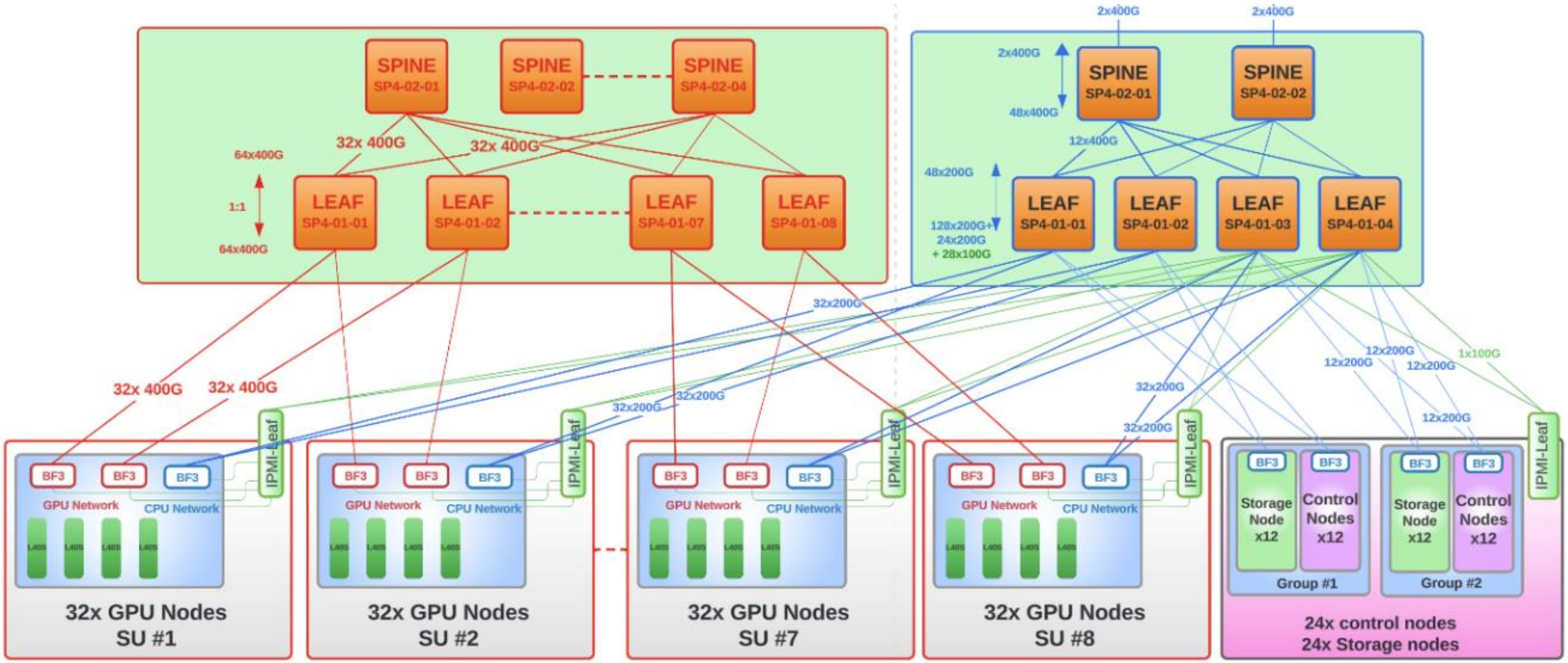
Taking the ”SuperPOD” to Cloud

NCP Reference Architecture


GPU & Network Operator
LLM Runtime
NVIDIA AI Enterprise
Models
Containers
Per-Tenant Infrastructure
Tenant 1 Base Command Manager
High-Performance Storage (derived)
…
Tenant 2 Base Command Manager
High-Performance Storage (derived)
…
Tenant N Base Command Manager
High-Performance Storage (derived)
Shared Infrastructure
NCP implemented IaaS Control Plane
NVIDIA IaaS Control Plane (future)
NCP Centralized High-Performance Storage (physical, 3rd party)
HGX H100 8GPU Compute Nodes
BCM
OS
East-West InfiniBand Fabric
UFM
North-South Ethernet Fabric
Cumulus
NetQ
Control Plane Nodes
BCM++

The Versatile Answer for Inference and Graphics

NVIDIA L40s Reference Architecture



New Option for Fine-Tuning and Inference

NVIDIA Reference Architecture with GraceHopper Superchip GH200

SPINE
SP3-02-01
SPINE
SP3-02-02
SPINE
SP3-02-08
4x 200G
4x 200G
4x 200G
32x200G
1:1
32x200G
LEAF
SP3-01-01
LEAF
SP3-01-02
LEAF
SP3-01-15
LEAF
SP3-01-16
16x200G
16x200G
16x200G
16x200G
16x200G
16x200G
4x200G
4x200G
4x200G
SPINE
SP3-02-01
SPINE
SP3-02-02
SPINE
SP3-02-08
48x200G
8x200G
3x200G
3x200G
3x200G
2x200G
24x200G
3:4
32x200G
LEAF
SP3-01-01
LEAF
SP3-01-02
LEAF
SP3-01-15
LEAF
SP3-01-16
M&S
Leaf#1
M&S
Leaf #2
M&S
Leaf #3
M&S
Leaf #4
16x200G
4:3
12x200G
IPMI-Spine
IPMI-Spine
100G
16x200G
16x200G
16x200G
16x200G
IPMI-Leaf
CX7
GPU
Network
BF3
CPU
Network
GH200
16x
MGX GH200
Nodes #1
16x
MGX GH200
Nodes #2
16x
MGX GH200
Nodes #3
16x
MGX GH200
Nodes #4
16x
MGX GH200
Nodes #29
16x
MGX GH200
Nodes #30
16x
MGX GH200
Nodes #31
16x
MGX GH200
Nodes #32
6x200G
6x200G
6x200G
6x200G
BF3
Storage
Node
x6
Control
Nodes
x6
Storge
Node
x6
Control
Nodes
x6

NVIDIA®