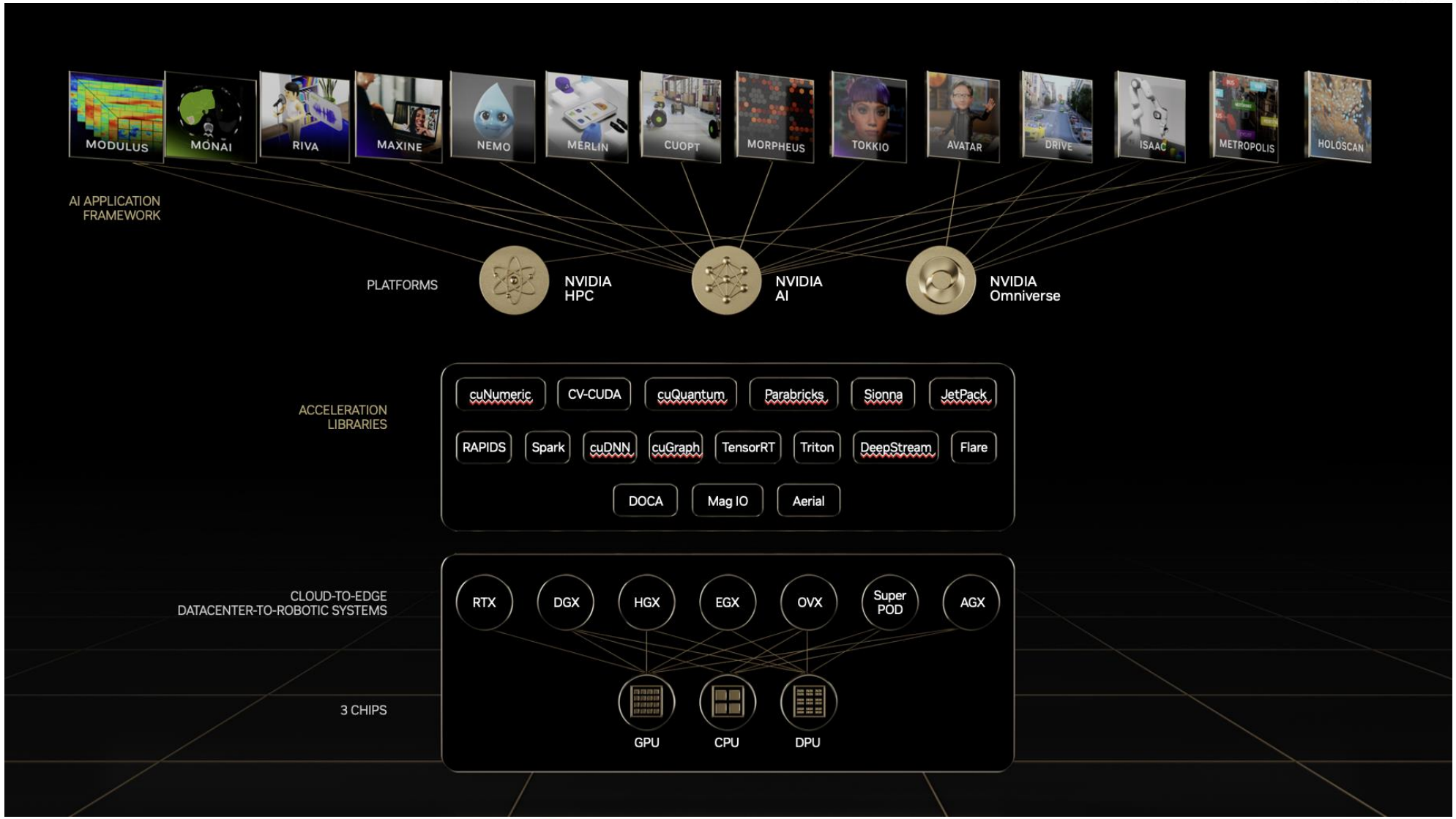
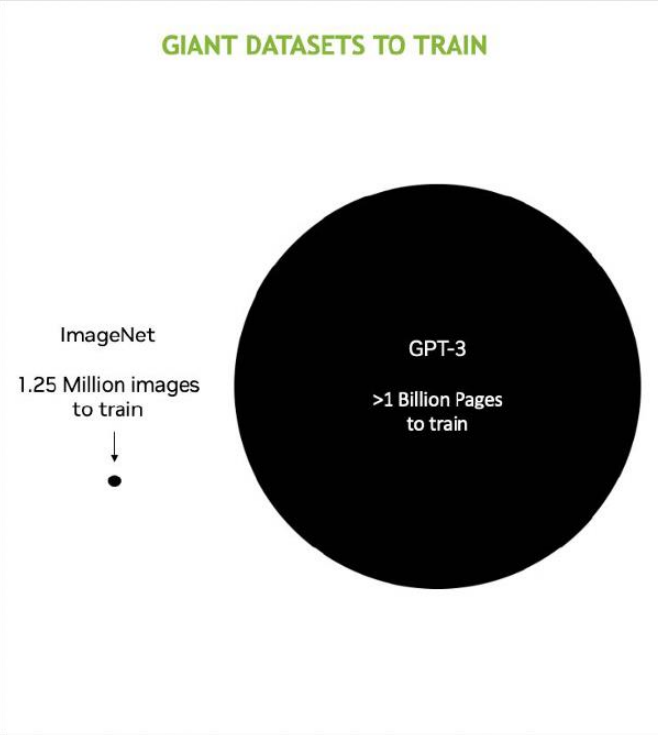

Architect
SACC
2022 中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2022
激发架构性能 点亮业务活力
云上会议 网络直播 | 2022年10月27-29日
ITPUB

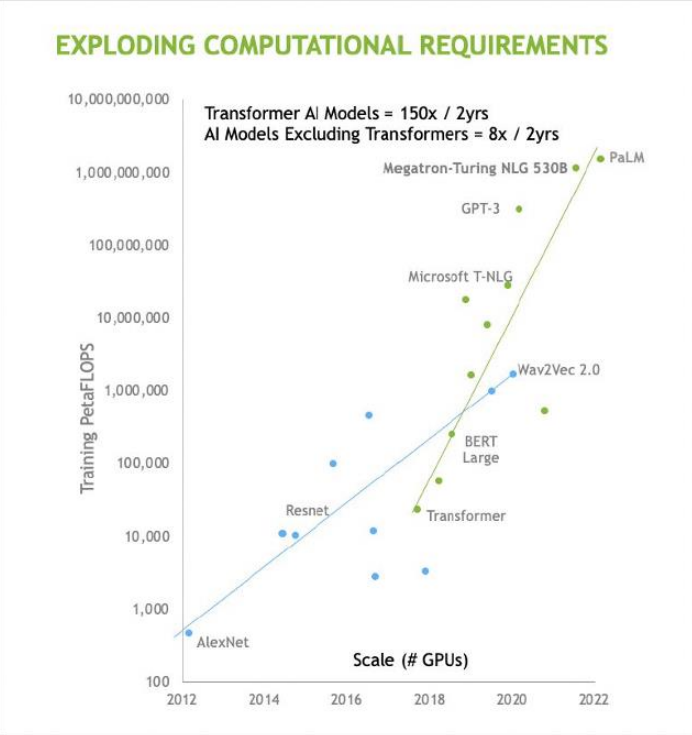
Designing Your Compute Data Center Architecture of Excellence in 2023

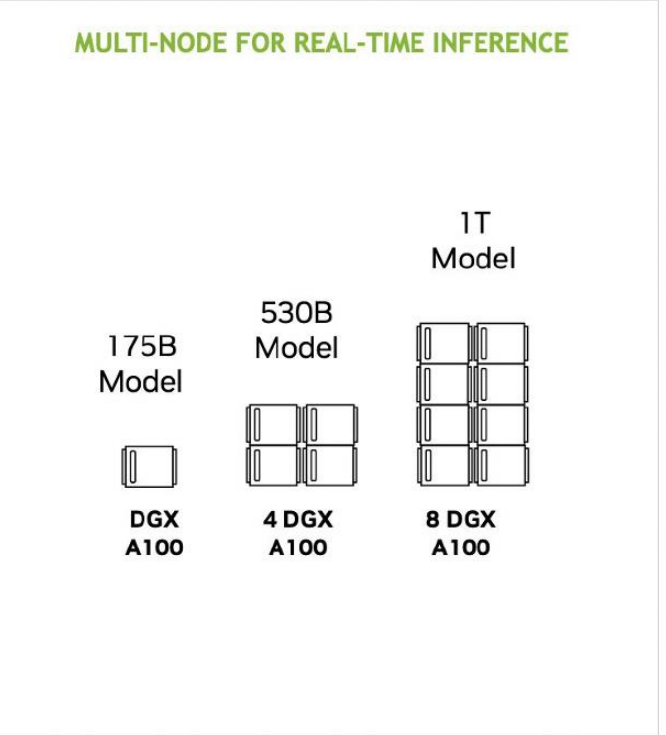
NVIDIA Senior Solution Architecture manager
Lu Chuan


MODULUS
MONAI
RIVA
MAXINE
NEMO
MERLIN
CUOPT
MORPHEUS
TOKKIO
AVATAR
DRIVE
ISAAC
METROPOLIS
HOLOSCAN
AI APPLICATION FRAMEWORK
PLATFORMS
NVIDIA HPC
NVIDIA AI
NVIDIA Omniverse
ACCELERATION LIBRARIES
cuNumeric
CV-CUDA
cuQuantum
Parabricks
Sionna
JetPack
RAPIDS
Spark
cuDNN
cuGraph
TensorRT
Triton
DeepStream
Flare
DOCA
Mag IO
Aerial
CLOUD-TO-EDGE DATACENTER-TO-ROBOTIC SYSTEMS
RTX
DGX
HGX
EGX
OVX
Super POD
AGX
3 CHIPS
GPU
CPU
DPU

Mounting challenges of computing at scale







The next wave of AI fueled by transformers

SuperGLUE: https://super.gluebenchmark.com/leaderboard | Google ViT: https://arxiv.org/abs/2010.11929 | OpenAI CLIP: https://openai.com/blog/clip/ | AlphaFold2: https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

NVIDIA NeMo Megatron with DGX SuperPOD

Train what was once impossible

Efficiency at Extreme Scale
Training GPT-3 175B takes 355 years on a V100, 14.8 years on 1 DGX A100 and about 1 month on a 140-node DGX SuperPOD

Tools to Build Your Own Custom Language Models
Train the world's largest transformer-based language models using Megatron's advanced optimizations and parallelization algorithms.

Optimized Topology for Multi-Node Training
Train the largest models using model parallelism, with NVLINK and InfiniBand for fast cross-node communication.

Turnkey Experience for Rapid Deployment
A full-stack data center platform that includes industry-leading computing, storage, networking, software, and management tools.

Direct access to world-class NLP experts
Access dedicated expertise from install to infrastructure management to scaling workloads to streamlined production AI.

NVIDIA Hopper GPU

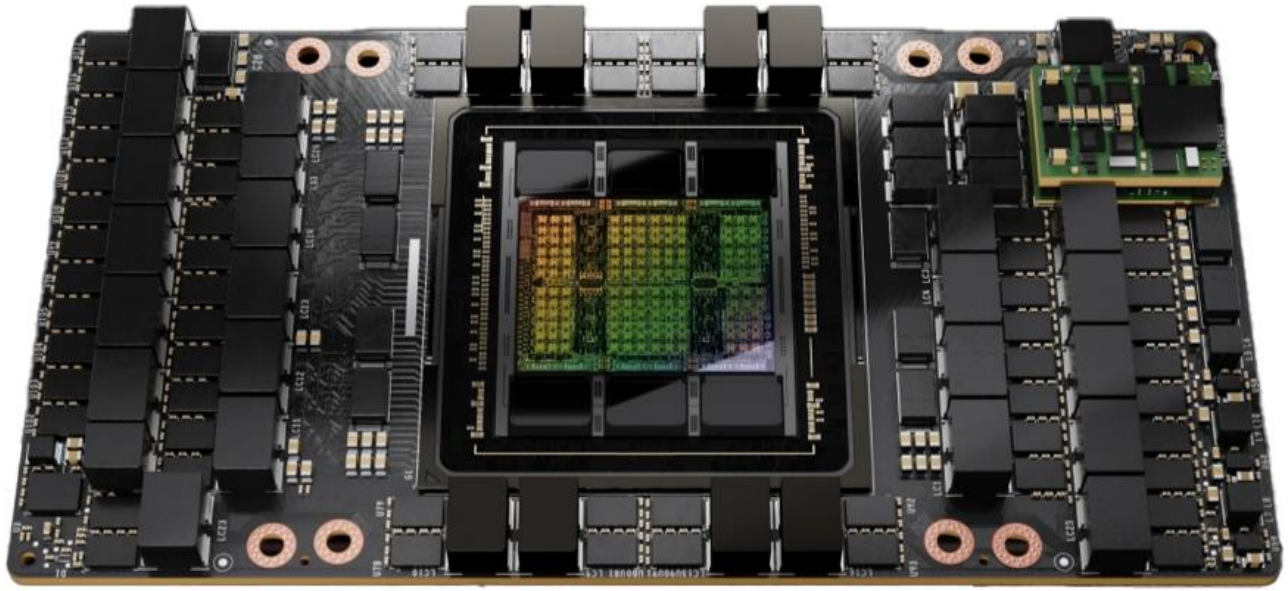
Unprecedented performance, scalability, and security for every data center

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

TRANFORMER MODEL OPTIMIZATIONS
6X faster on largest transformer models

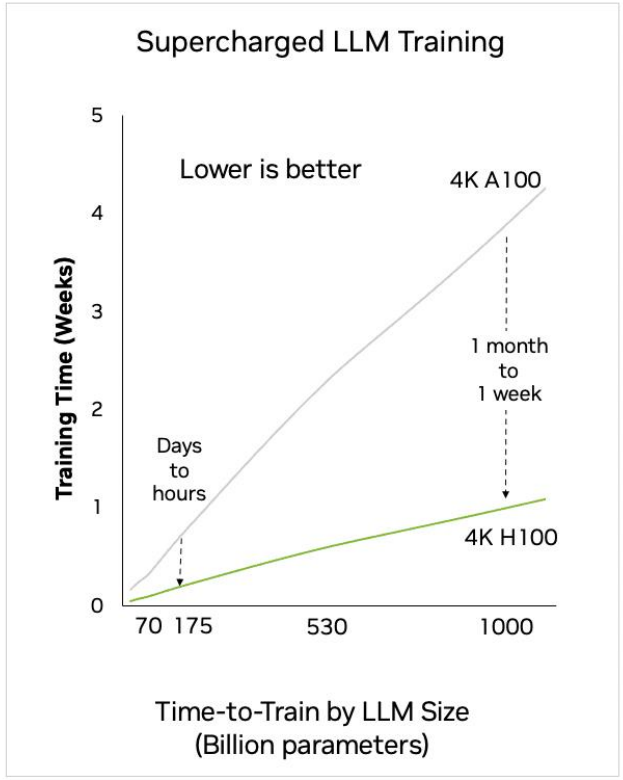
HIGHEST UTILIZATION EFFICIENCY AND SECURITY
7 Fully isolated & secured instances, guaranteed QoS
2nd Gen MIG | Confidential Computing

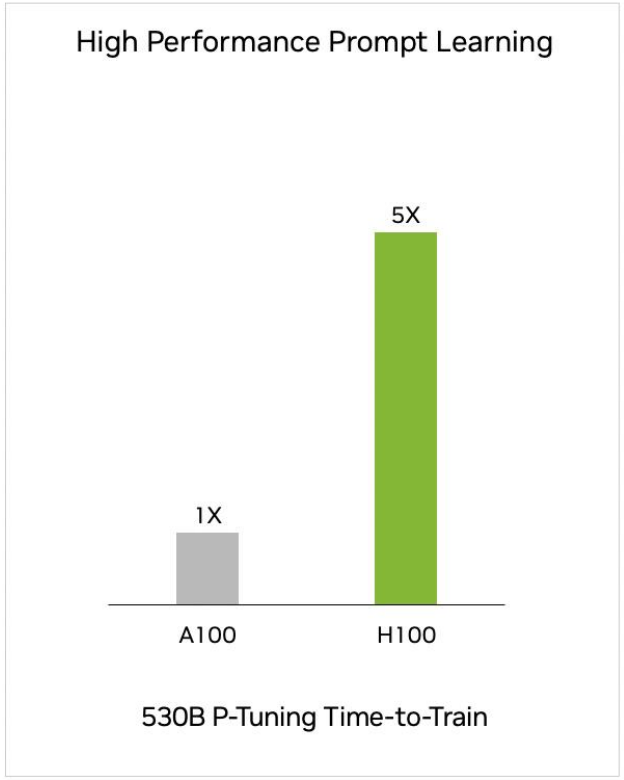
FASTEST, SCALABLE INTERCONNECT
900 GB/s GPU-2-GPU connectivity (1.5X)
up to 256 GPUs with NVLink Switch | 128GB/s PCI Gen5

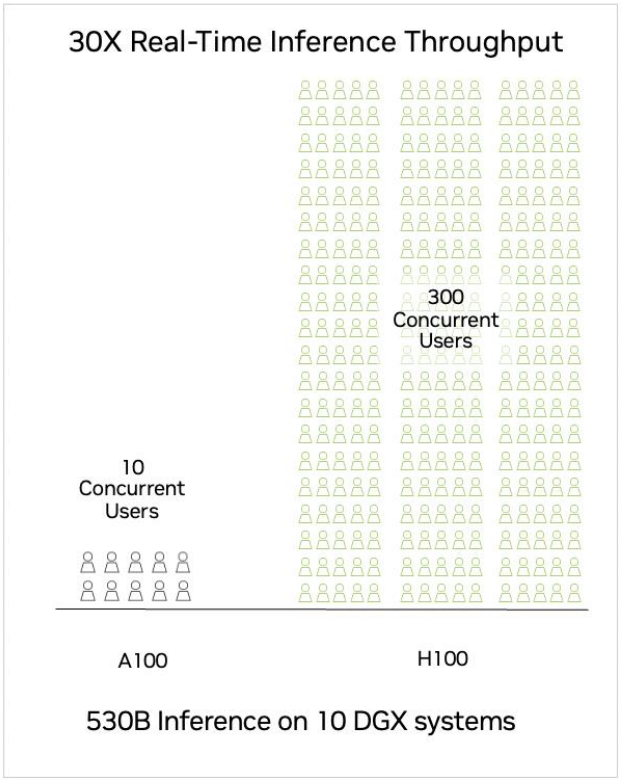


NVIDIA Hopper Supercharges LLMs

Hopper architecture addresses LLM needs at scale



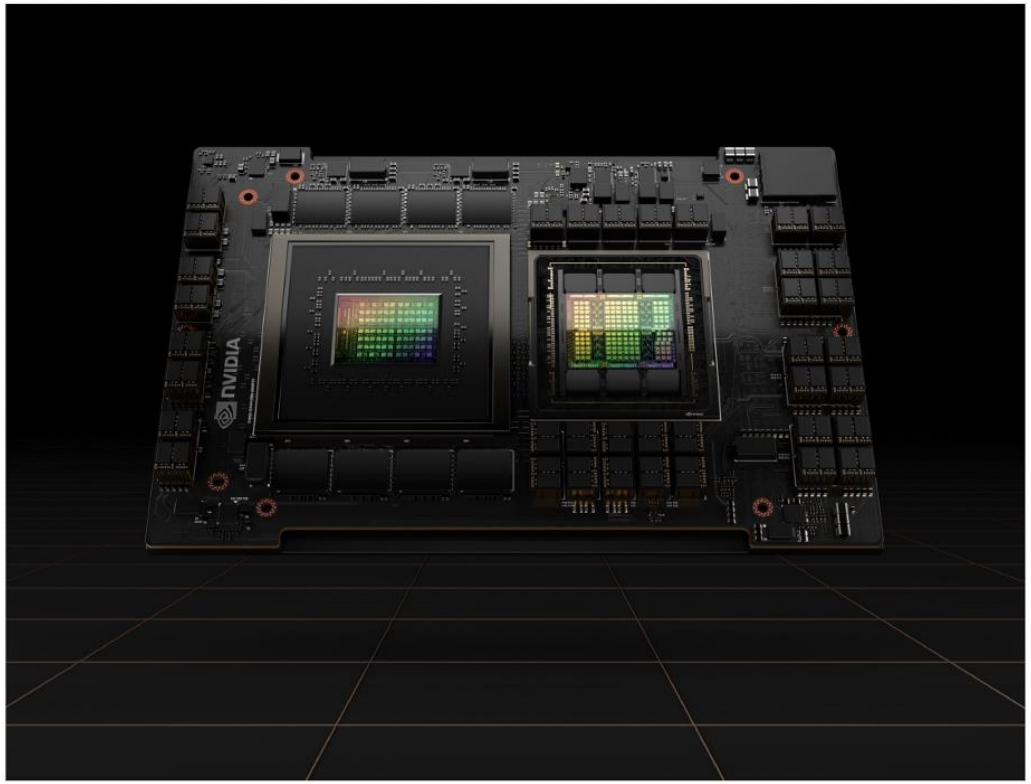


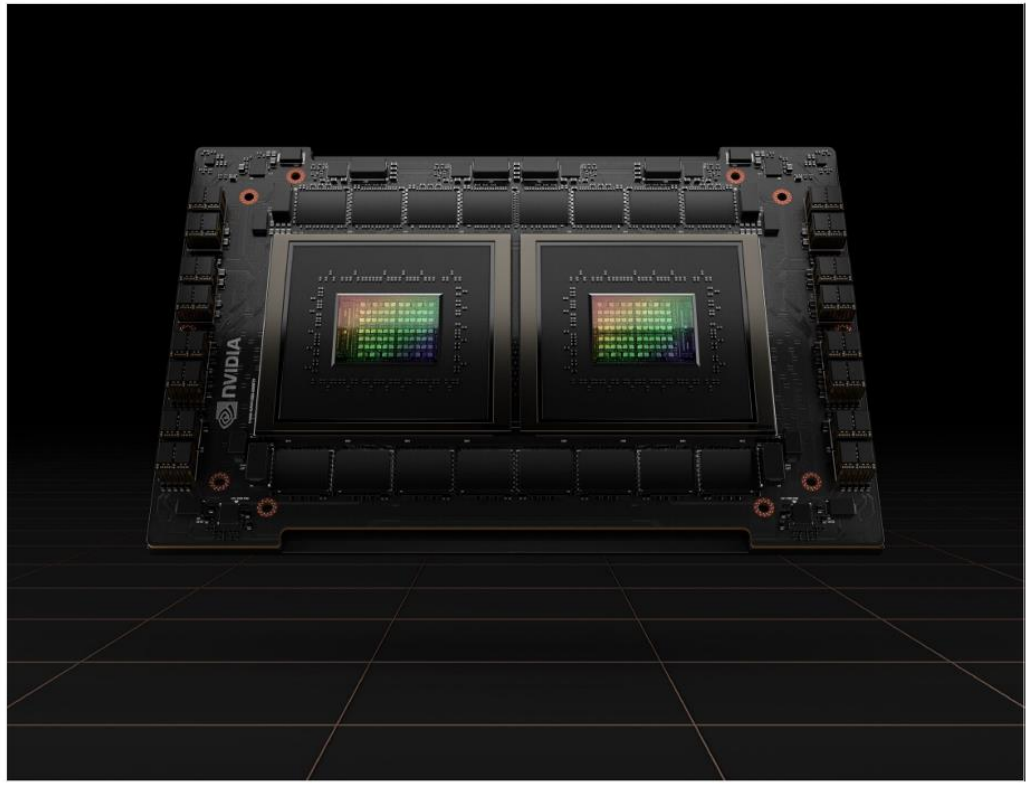


LLM Training | 4096 GPUs | H100 NDR IB | A100 HDR IB | 300 Billion tokens.
P-Tuning | DGX H100 | DGX A100 | 530B Q&A tuning using SQuAD dataset
Inference | chatbot| 10 DGX H100 NDR IB | 10 DGX A100 HDR IB | <1 sec latency | 1 inference/second/user.
H100 data center projected workload performance, subject to change

NVIDIA GRACE

Designed from the Ground-UP to be a Superchip

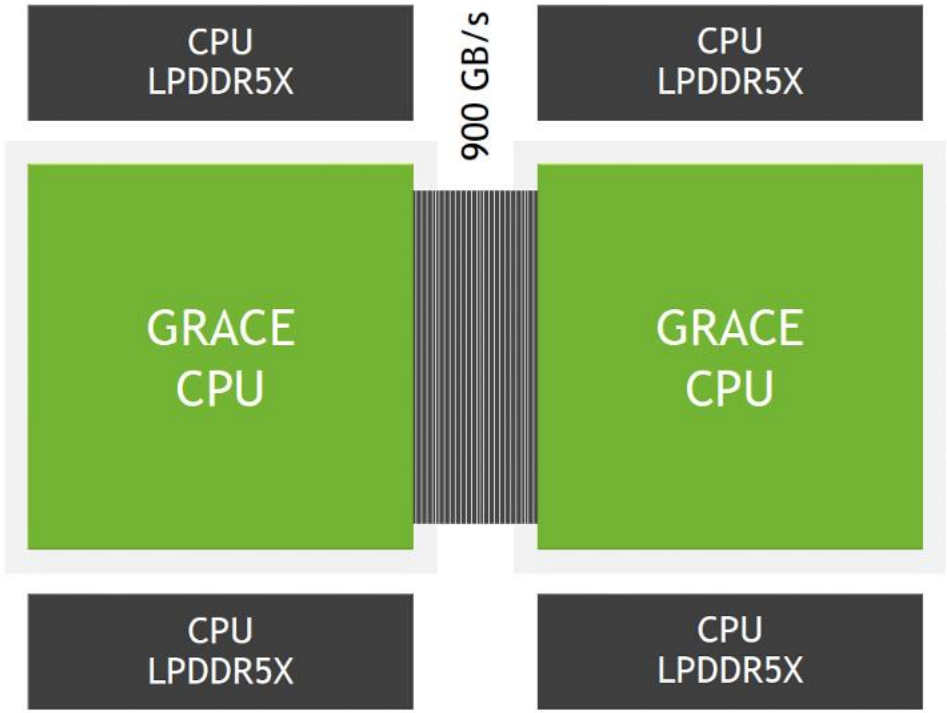




NVLINK-C2C

High speed chip to chip interconnect

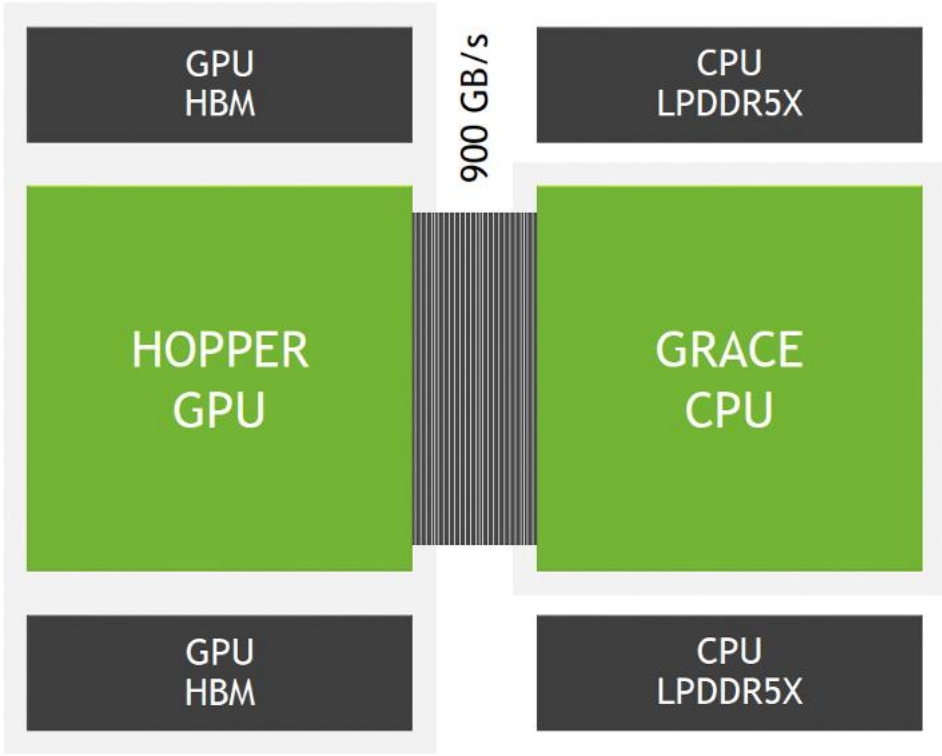
- Used to create the Grace Hopper, and Grace Superchips
- Removes the typical cross-socket bottlenecks
- Up to 900GB/s of raw bidirectional BW
  - Same BW as GPU to GPU NVLINK on Hopper
- Low power interface - 1.3 pJ/bit
  - More than 5x more power efficient than PCIe
- Enables coherency for both Grace and Grace Hopper superchips



GRACE HOPPER

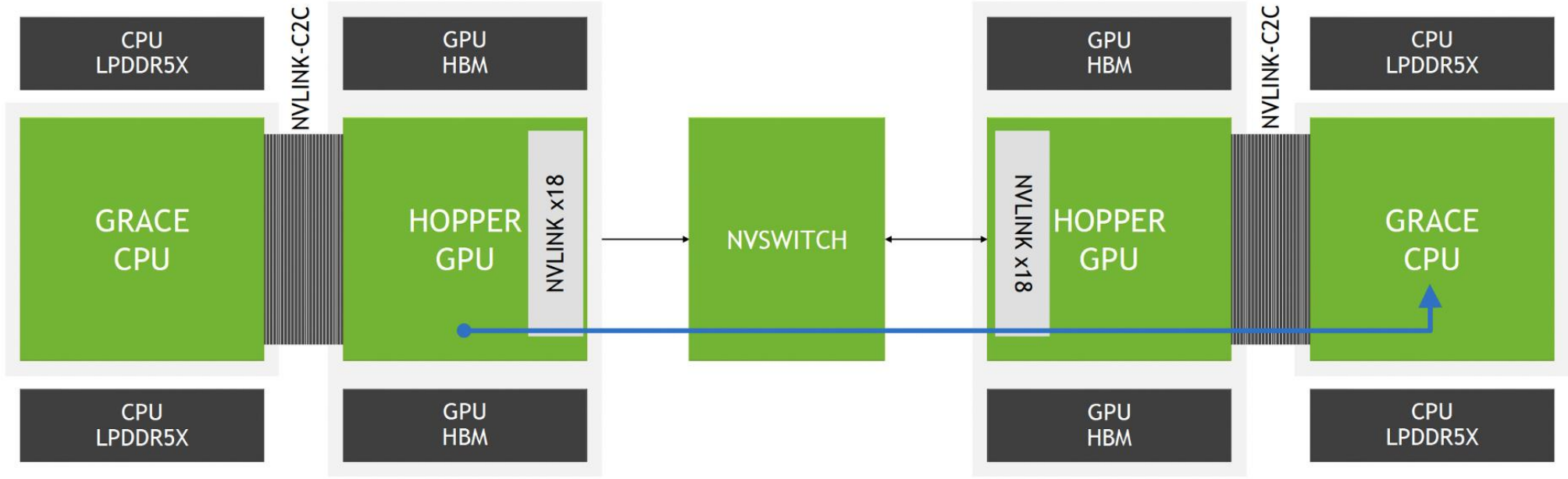
Heterogenous Coherency

- Unified Memory with shared page tables
  - Shared CPU and GPU virtual address space
  - Transparent GPU access to pageable memory
  - System allocator support for GPU memory
    - Yes, malloced and mmaped pointers!
- Native atomics, including standard C++ atomic support



NVLINK- SCALING

Superchip Scaling | CPU/GPU | Extended GPU Memory

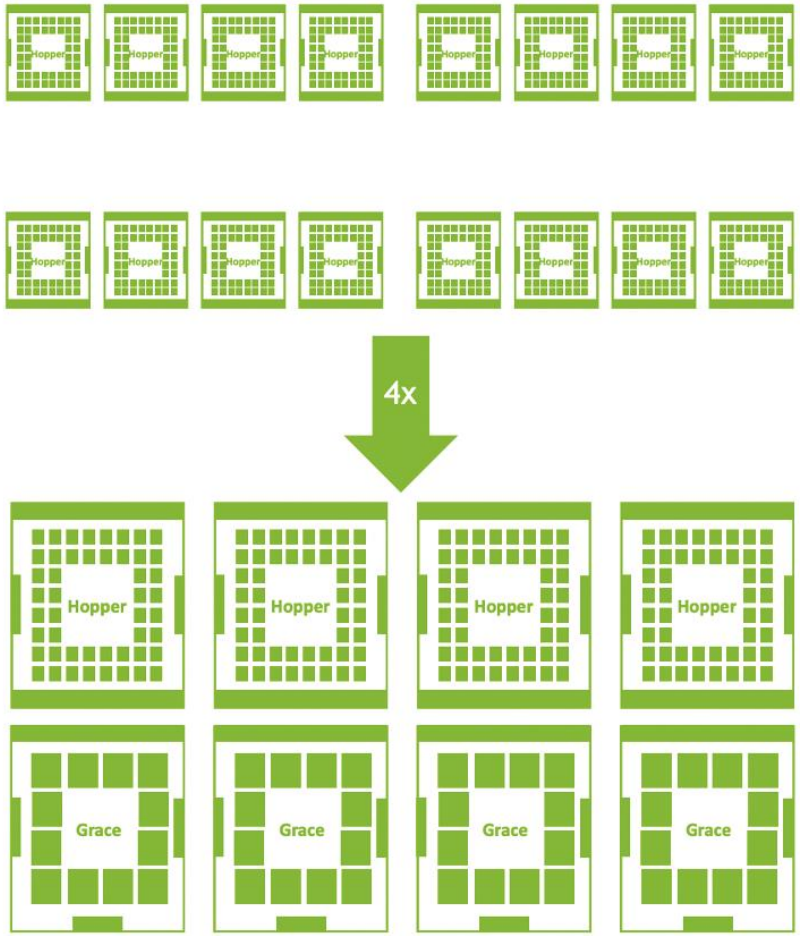


Enables remote NVLINK connected GPUs, to access Grace's memory at native NVLINK speeds

NVIDIA GRACE

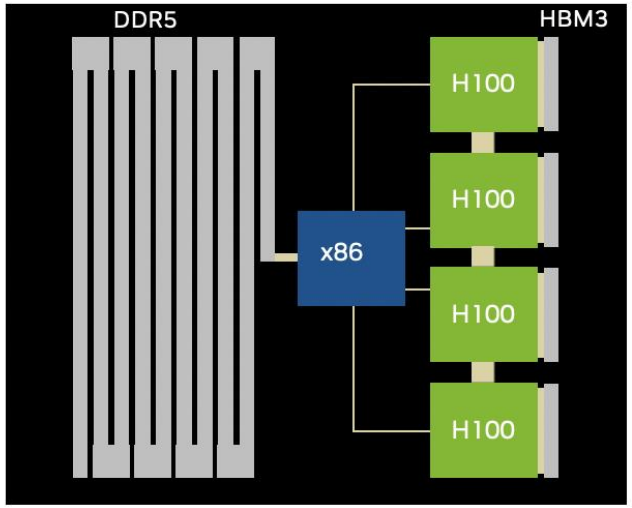
- Natural Language Processing
- GPT-3 inference – fp8 – 175GB of memory
- GPT-3 training – over 2.5TB of memory

- Extended GPU Memory to the rescue!
- 4x decrease in the number of GPUs needed to fit the training set in memory



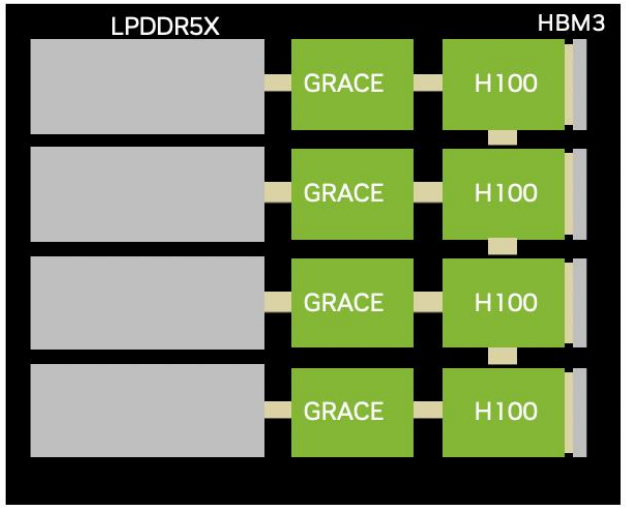
Grace Hopper to Supercharge Recommender Systems

CURRENT x86 ARCHITECTURE



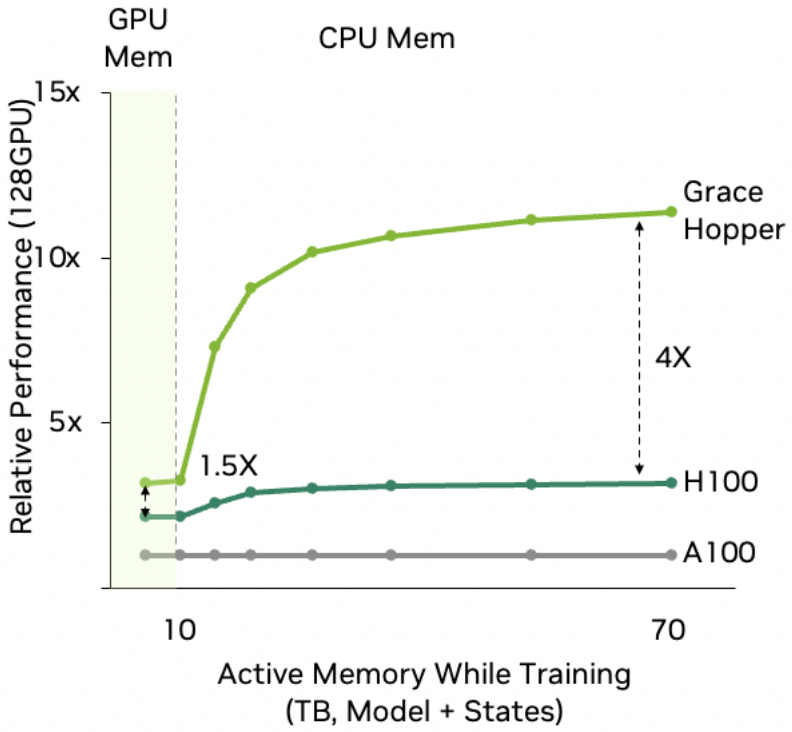
GPU	12,000	GB/sec
CPU	350	GB/sec
PCIE Gen5 (Effective Per GPU)	**128**	**GB/sec**
Fast Memory	**320**	**GB**

GRACE HOPPER ARCHITECTURE

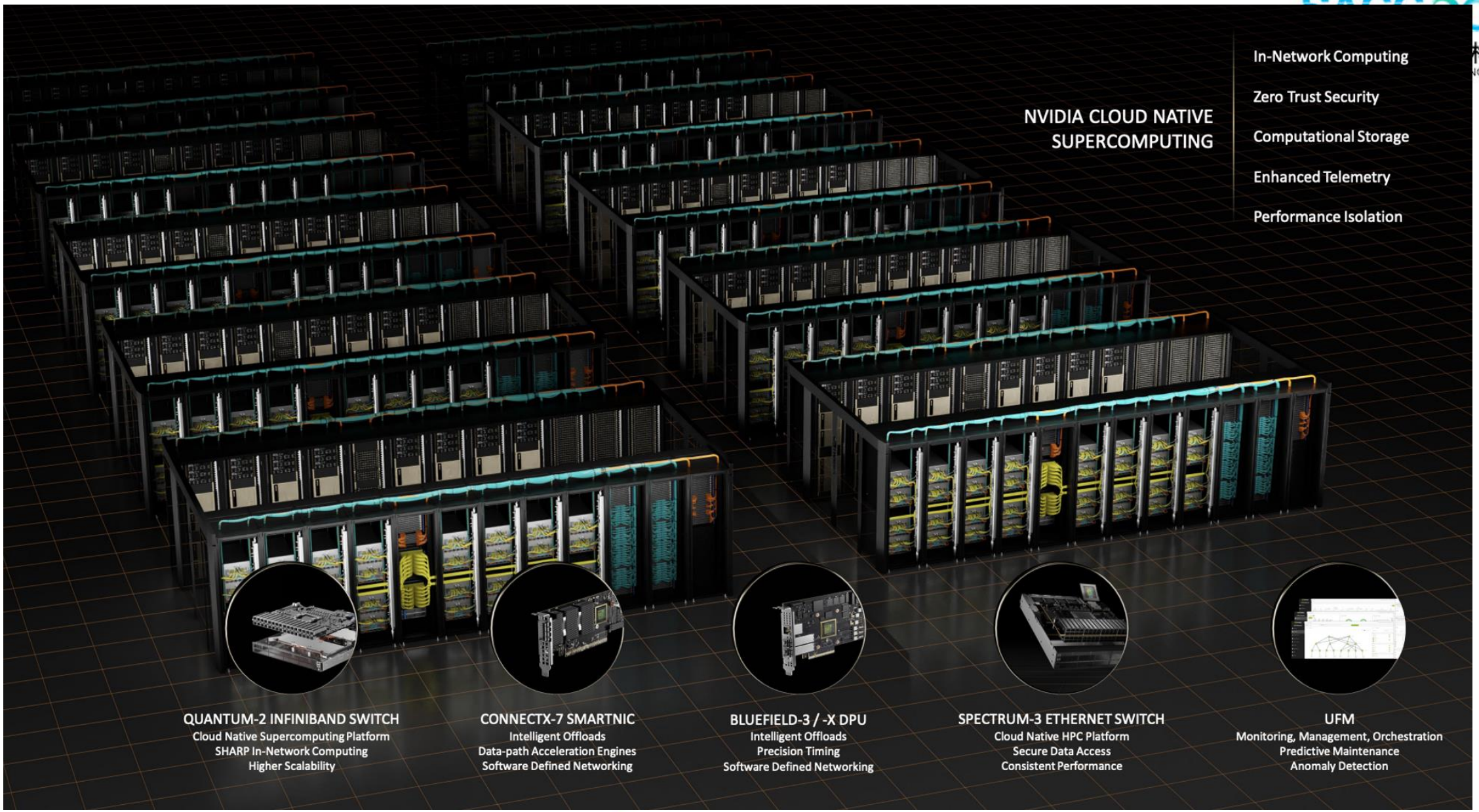


GPU	12,000	GB/sec	
CPU	500	GB/sec	
NVLink C2C	**900**	**GB/sec**	**7X**
Fast Memory (LPDDR5X + HBM)	**2300**	**GB**	**7X**

Grace Hopper 4X Faster Large Recommender Training



Bandwidth claims rounded for illustration.
Performance results based on projections on these configurations Grace : 128xGrace Hopper Superchip with MNNVL, 128x DGX H100 with IB and 128x DGX A100 with IB
Synthetic recommender system model


NVIDIA CLOUD NATIVE
SUPERCOMPUTING
In-Network Computing
Zero Trust Security
Computational Storage
Enhanced Telemetry
Performance Isolation
QUANTUM-2 INFINIBAND SWITCH
Cloud Native Supercomputing Platform
SHARP In-Network Computing
Higher Scalability
CONNECTX-7 SMARTNIC
Intelligent Offloads
Data-path Acceleration Engines
Software Defined Networking
BLUEFIELD-3 / -X DPU
Intelligent Offloads
Precision Timing
Software Defined Networking
SPECTRUM-3 ETHERNET SWITCH
Cloud Native HPC Platform
Secure Data Access
Consistent Performance
UFM
Monitoring, Management, Orchestration
Predictive Maintenance
Anomaly Detection

NVIDIA SPECTRUM PLATFORM

Accelerated Ethernet Technologies

ACCELERATED
Best-in-class hardware performance with cloud-scale software efficiency

INNOVATIVE
5th generation in-house ASIC design optimizes Cloud, AI, & storage workloads

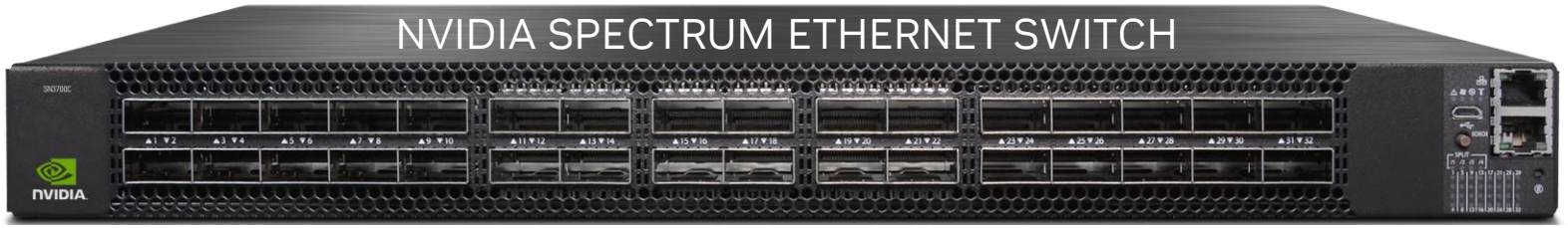
NVIDIA NETQ | NVIDIA AIR
NETWORK SOLUTION TOOLS

NVIDIA CUMULUS | PURE SONIC
NETWORK OPERATING SYSTEM

NVIDIA SPECTRUM ASIC
SWITCH ASIC

OPTIMIZED
Faster network deployments with lowest TCO and highest ROI

RELIABLE
Exclusive features enabling fairness, predictability and actionable visibility



BlueField Data Processing Unit

SOFTWARE DEFINED NETWORKING

SOFTWARE DEFINED SECURITY

SOFTWARE DEFINED STORAGE

Infrastructure Services

Data Center on a Chip

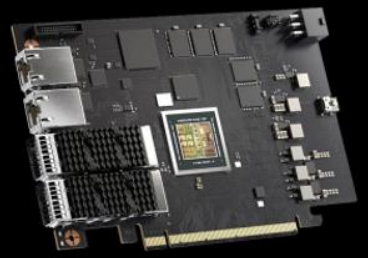
16 Arm 64-Bit Cores

16 Core / 256 Threads Datapath Accelerator

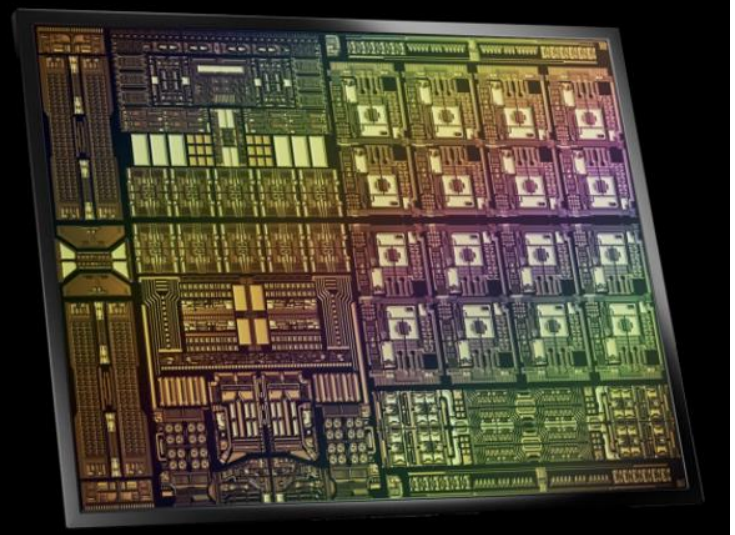
ConnectX InfiniBand / Ethernet

DDR memory interface

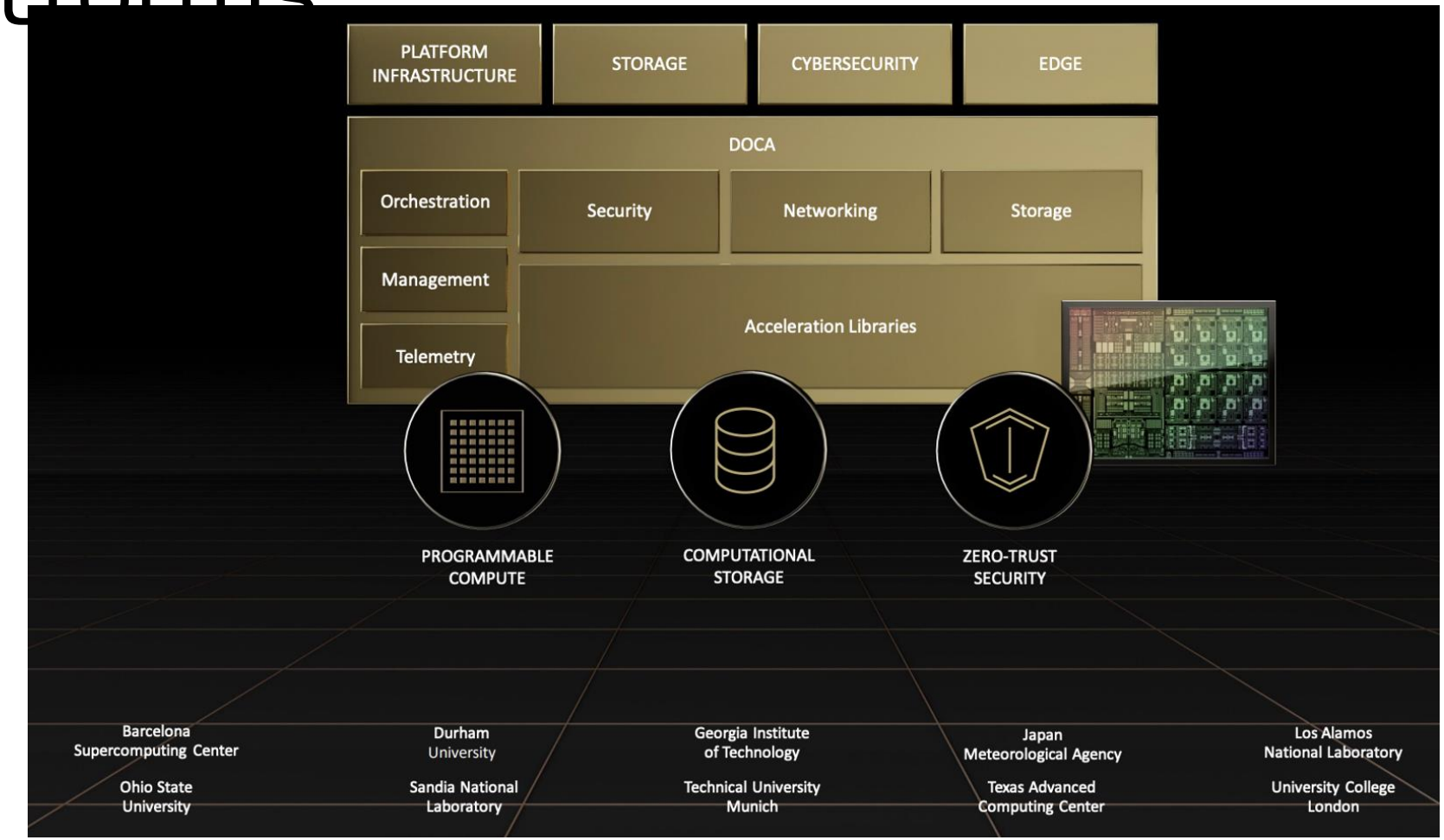
PCIe switch

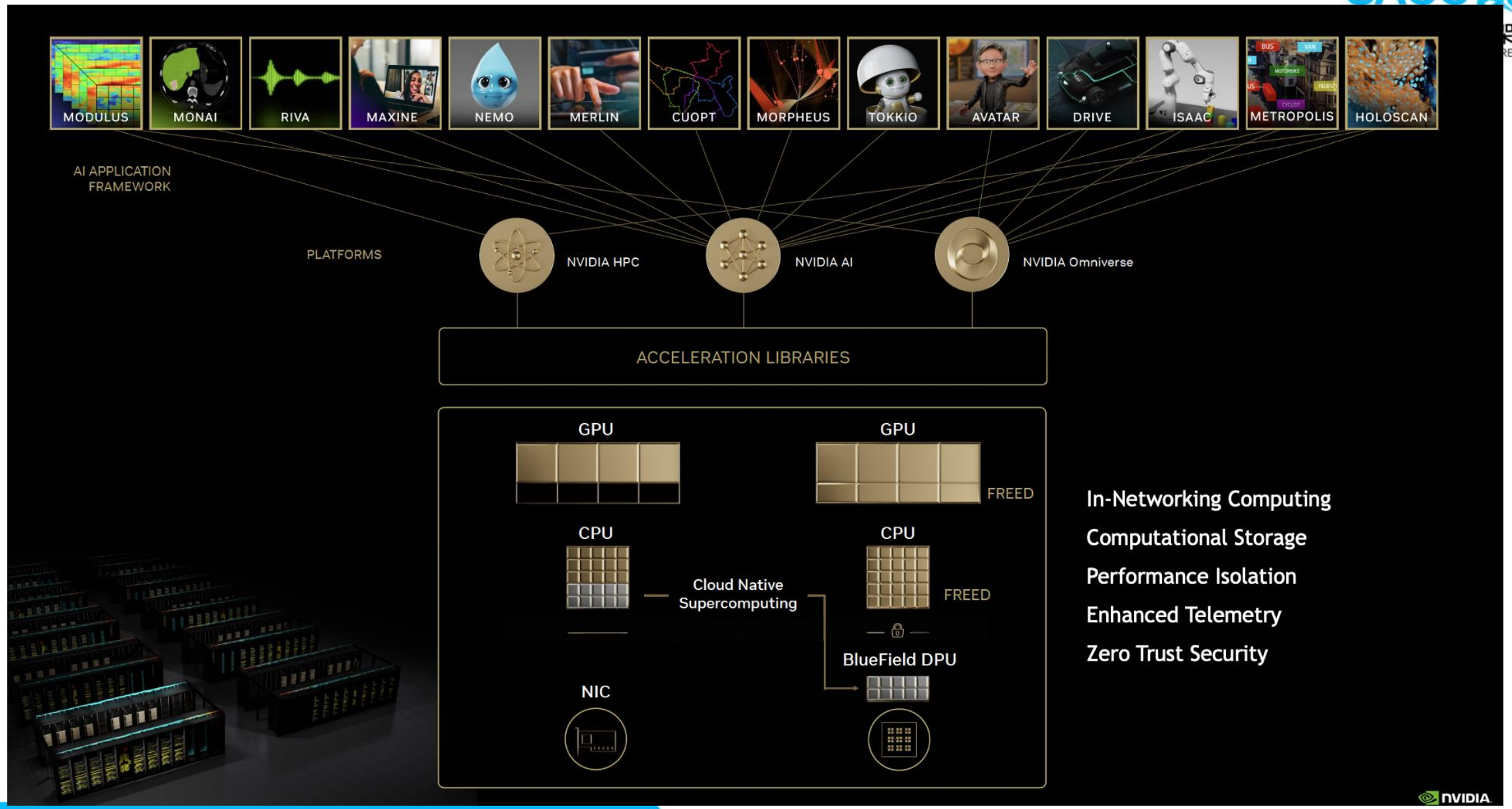


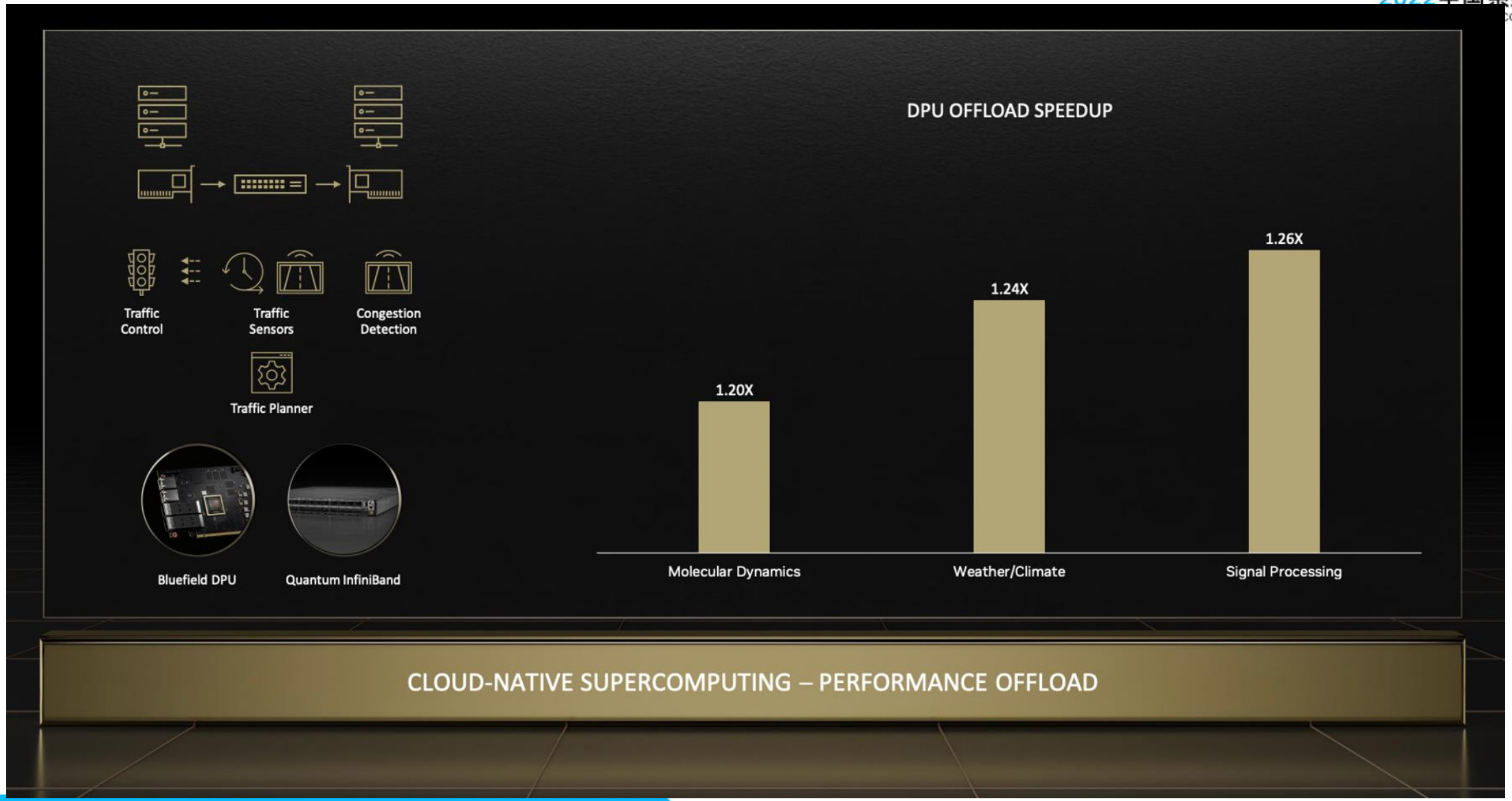
BlueField Infrastructure
Compute Platform

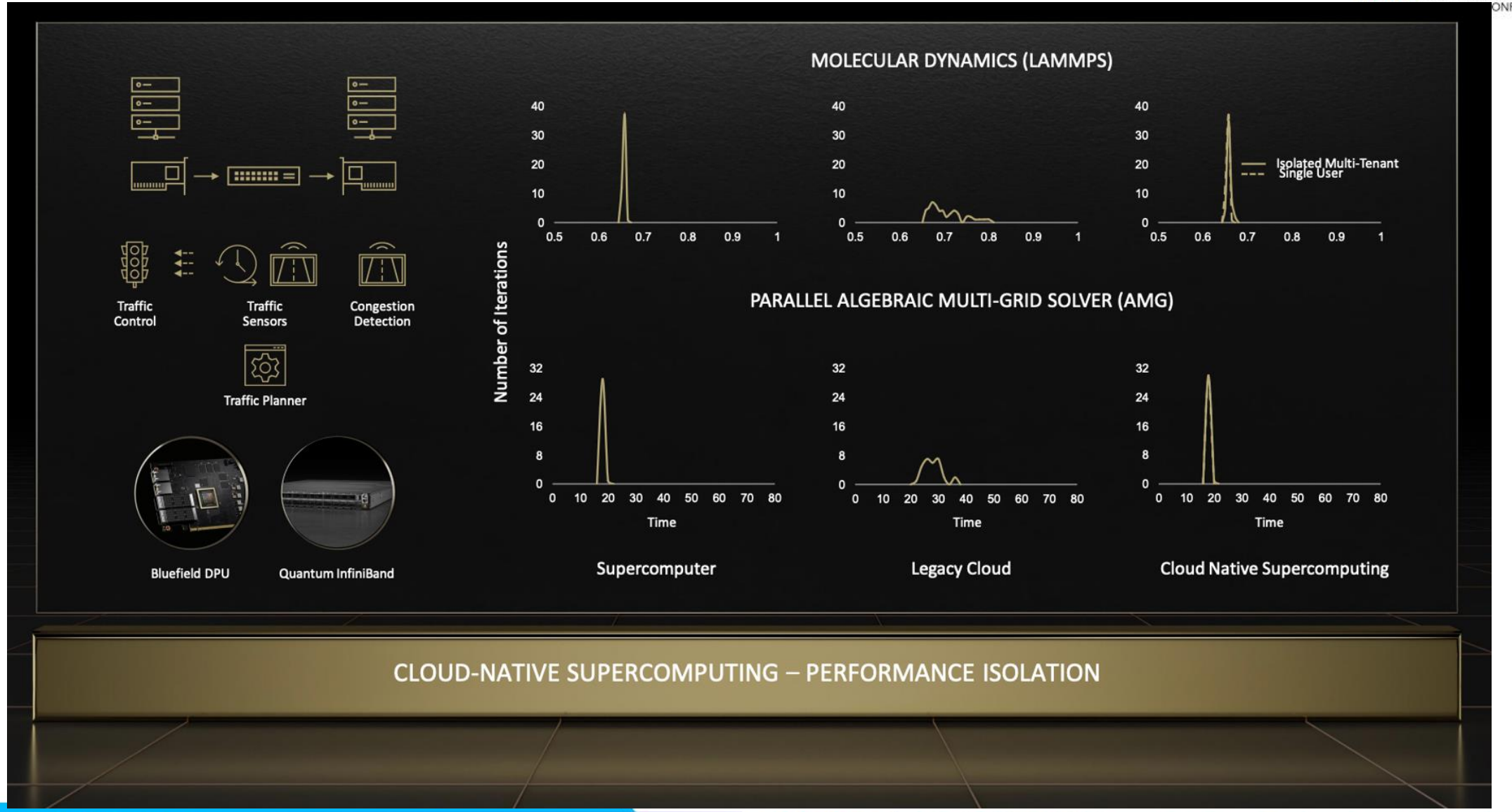


Cloud Native Supercomputing Platforms



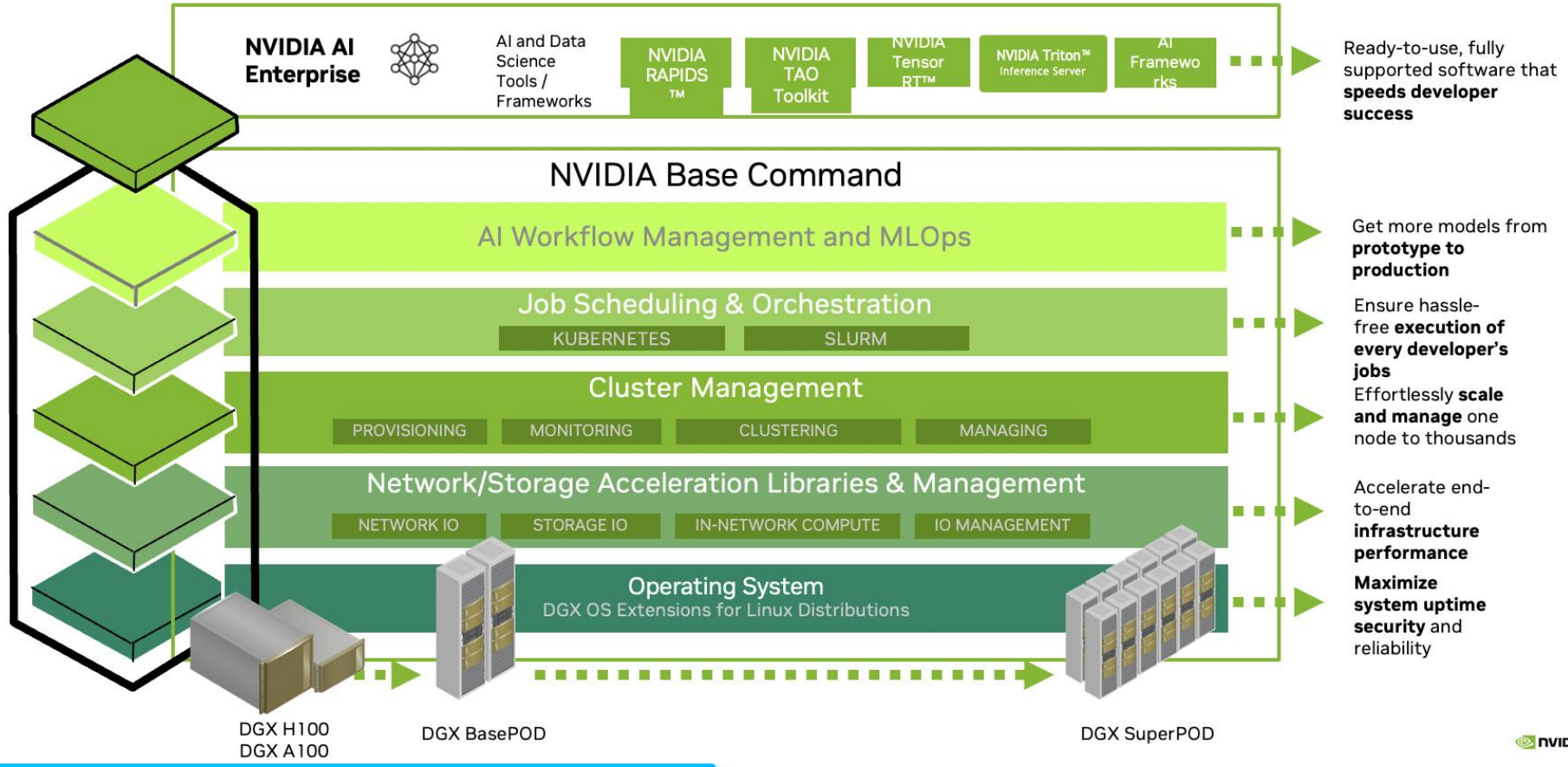

MODULUS
MONAI
RIVA
MAXINE
NEMO
MERLIN
CUOPT
MORPHEUS
TOKKIO
AVATAR
DRIVE
ISAAC
METROPOLIS
HOLOSCAN
AI APPLICATION FRAMEWORK
PLATFORMS
NVIDIA HPC
NVIDIA AI
NVIDIA Omniverse
ACCELERATION LIBRARIES
GPU
GPU
FREED
CPU
CPU
Cloud Native Supercomputing
FREED
BlueField DPU
NIC
In-Networking Computing
Computational Storage
Performance Isolation
Enhanced Telemetry
Zero Trust Security
NVIDIA


Traffic Control
Traffic Sensors
Congestion Detection
Traffic Planner
Bluefield DPU
Quantum InfiniBand
DPU OFFLOAD SPEEDUP
1.20X
1.24X
1.26X
Molecular Dynamics
Weather/Climate
Signal Processing
CLOUD-NATIVE SUPERCOMPUTING – PERFORMANCE OFFLOAD


MOLECULAR DYNAMICS (LAMMPS)
Isolated Multi-Tenant
Single User
Traffic Control
Traffic Sensors
Congestion Detection
Traffic Planner
PARALLEL ALGEBRAIC MULTI-GRID SOLVER (AMG)
Number of Iterations
Time
Bluefield DPU
Quantum InfiniBand
Supercomputer
Legacy Cloud
Cloud Native Supercomputing
CLOUD-NATIVE SUPERCOMPUTING – PERFORMANCE ISOLATION

Cluster Management

Enterprises tools that drive the value of AI investment



NGC

Portal to AI services, software, support

NGC Catalog

Cloud Services
End-to-End AI development

AI Services for NLP, biology, speech

AI Workflow Management & Support

Performance Optimized
Tested across GPU-accelerated platforms

Monthly sw container updates

SOTA models

Fully Transparent
Quickly find and deploy the right sw

Detailed security scan reports

Model resumes

Accelerates Development
Focus on building, not setup

One click deploy from NGC

Develop once. Deploy anywhere w/ NVIDIA VMI

ngc.nvidia.com

THANKS