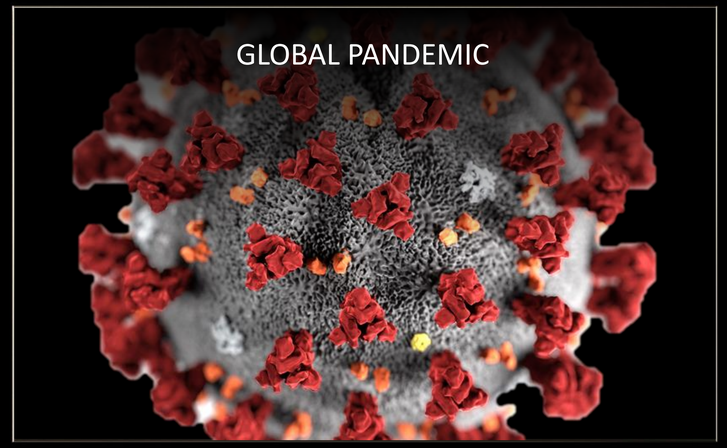
EFFICIENT AT SCALE PERFORMANCE USING NVIDIA ACCELERATED SOLUTIONS FOR HPC AND AI

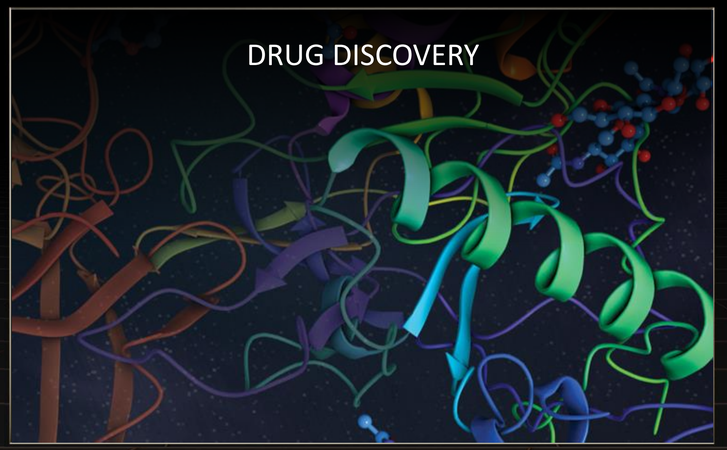
DR. GABRIEL NOAJE

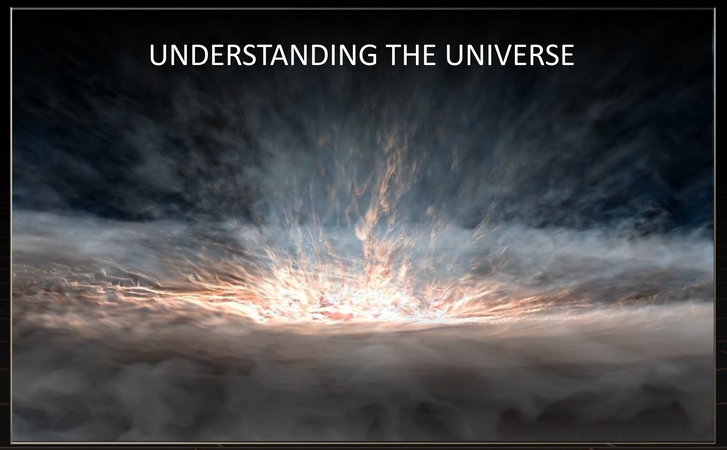
PRINCIPAL SOLUTIONS ARCHITECT, APAC SOUTH, NVIDIA

GNOAJE@NVIDIA.COM

JULY 26TH, 2022







GRAND CHALLENGES OF THE 21ST CENTURY

Full Stack. Data Center Scale

2,700 Accelerated Applications

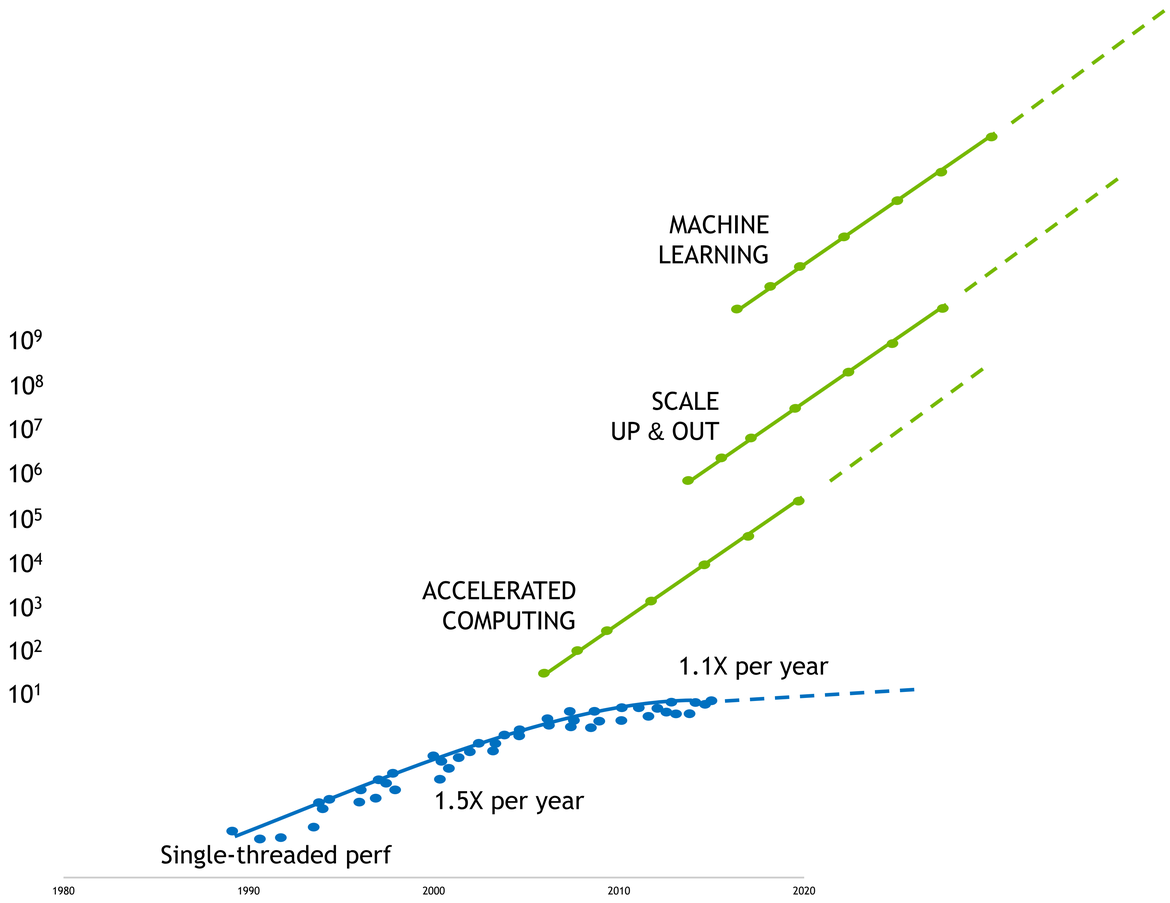
450 SDKs, AI Models

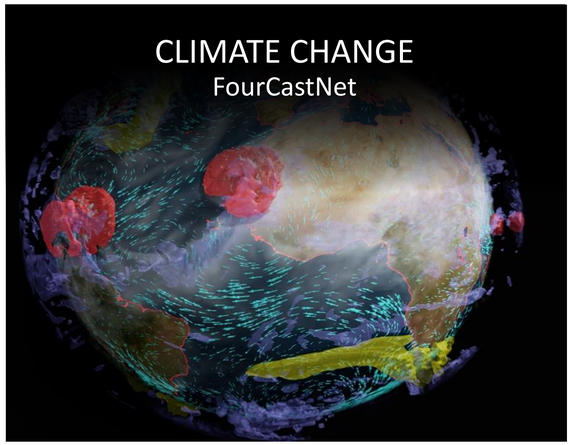
30 Million CUDA Downloads

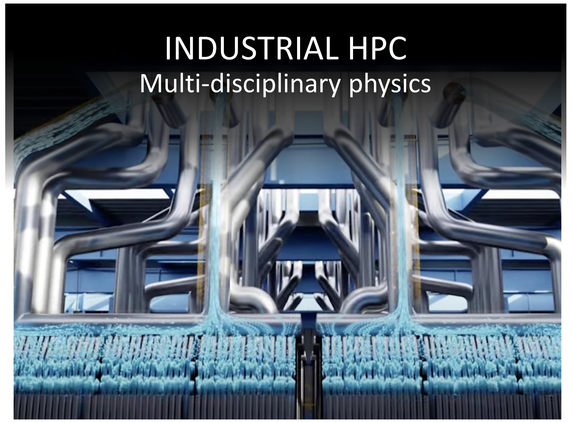
3 Million Developers

MILLION-X SPEEDUP FOR INNOVATION AND DISCOVERY

Combination of Accelerated Computing, Data Center Scale and AI







SIMULATION + AI: MILLION-X SPEEDUP FOR INNOVATION AND DISCOVERY

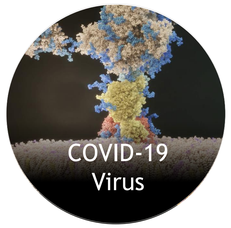
Combination of Accelerated Computing, Data Center Scale and AI

WORKLOADS OF THE MODERN SUPERCOMPUTER

EDGE

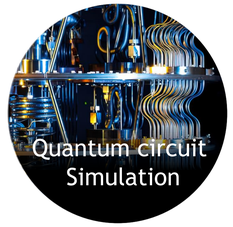
HPC + AI

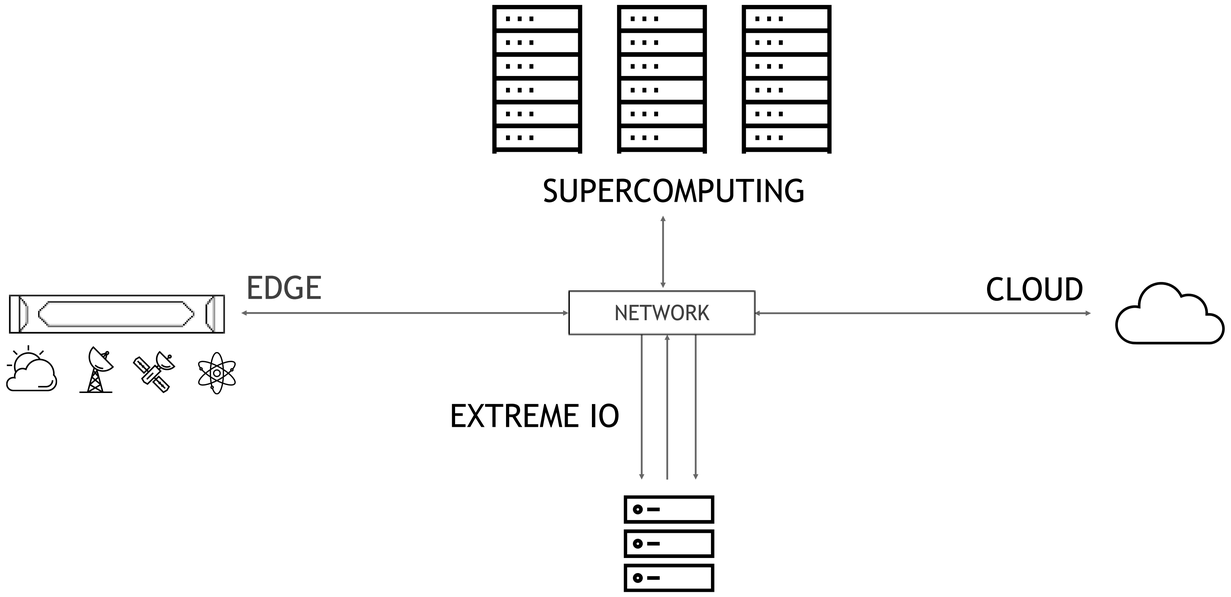
SIMULATION



DIGITAL TWIN

QUANTUM COMPUTING

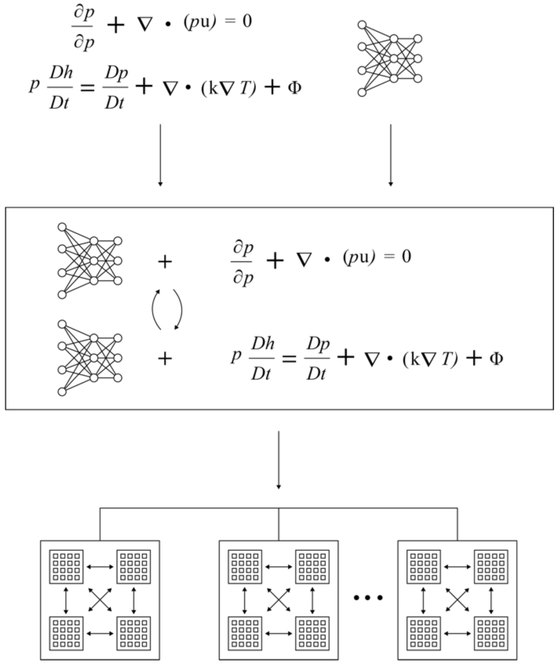




NVIDIA MODULUS

Physics Machine Learning Platform

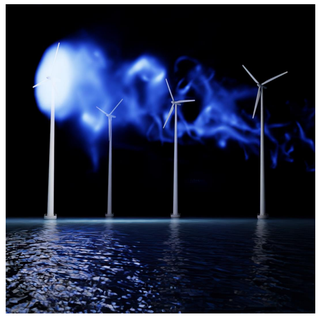
TRAINING NEURAL NETWORKS USING BOTH DATA AND THE GOVERNING EQUATIONS



ADVANCING SCIENTIFIC DISCOVERY WITH MODULUS

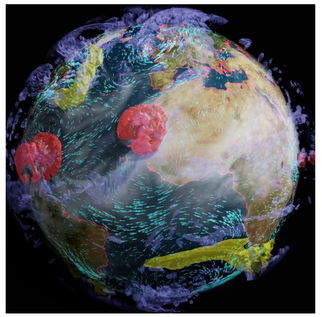
RENEWABLE ENERGY

Siemens Gamesa: Up to 4000X Speedup of Wind Turbine Wake Optimization



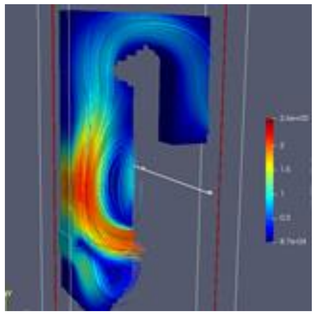
CLIMATE CHANGE

45,000X Speedup of Extreme weather Prediction with FourCastNet



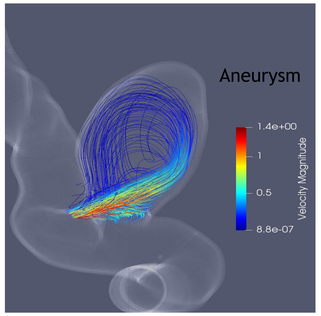
INDUSTRIAL HPC

NETL: 10,000X Faster Build Of high-fidelity surrogate models



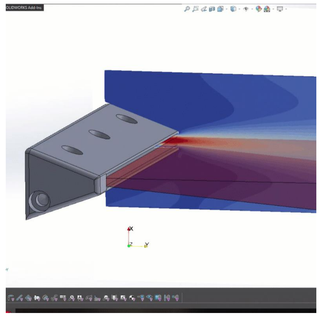
HEALTHCARE

Achieve high-fidelity results faster for blood flow in inter-cranial aneurysm



DIGITAL TWINS

Kinetic Vision: Design Optimization Using parameterized models

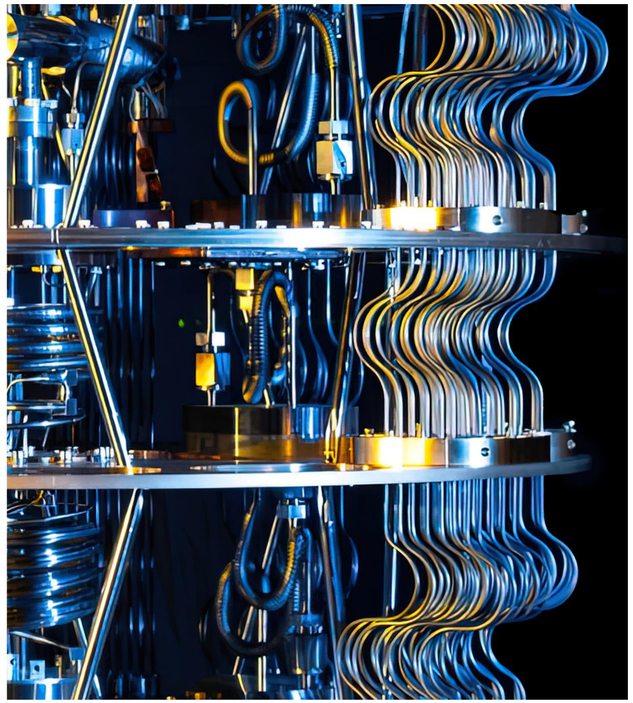


Get started today with NVIDIA Modulus

ACCELERATE QUANTUM COMPUTING RESEARCH WITH NVIDIA CUQUANTUM

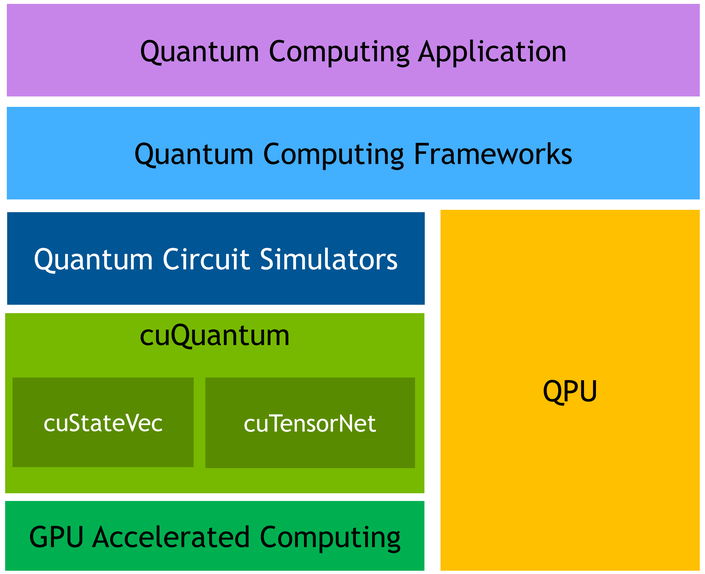
Research the Computer of Tomorrow with the Most Powerful Computer Today

COMPUTING OF TOMORROW

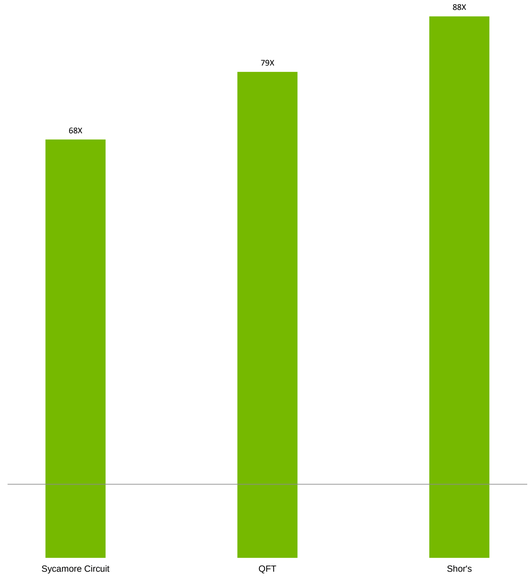


PLATFORM TO ACCELERATE QUANTUM COMPUTNG RESEARCH

Identify Quantum Candidates | Experiment Algorithms | Simulate Large Designs



UP TO 88X SPEEDUP* ON KEY QUANTUM ALGORITHMS



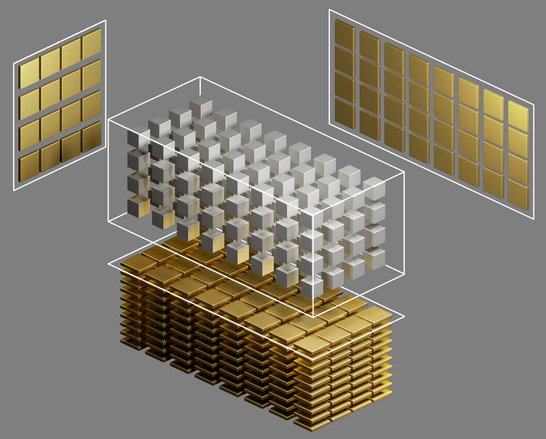
*over Dual Core CPU

NVIDIA

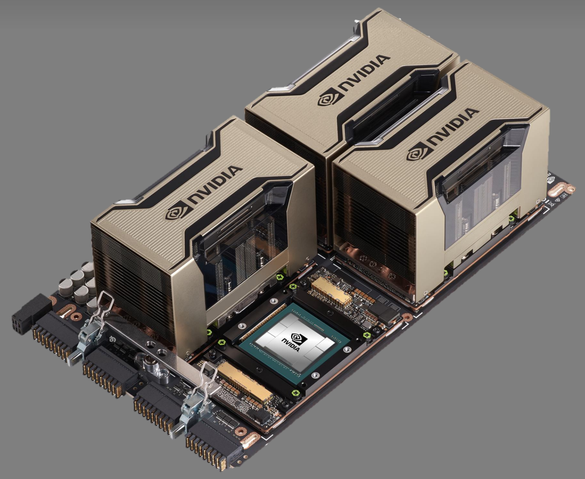
NVIDIA PERFORMANCE LIBRARIES

Major Directions

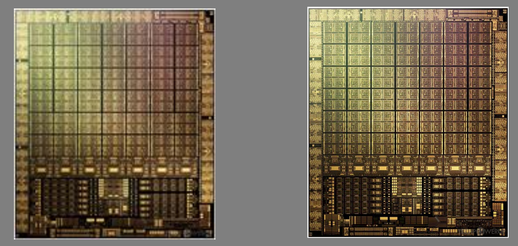
Seamless Acceleration
Tensor Cores, Enhanced L2$ & SMEM



Scaling Up
Multi-GPU and Multi-Node Libraries

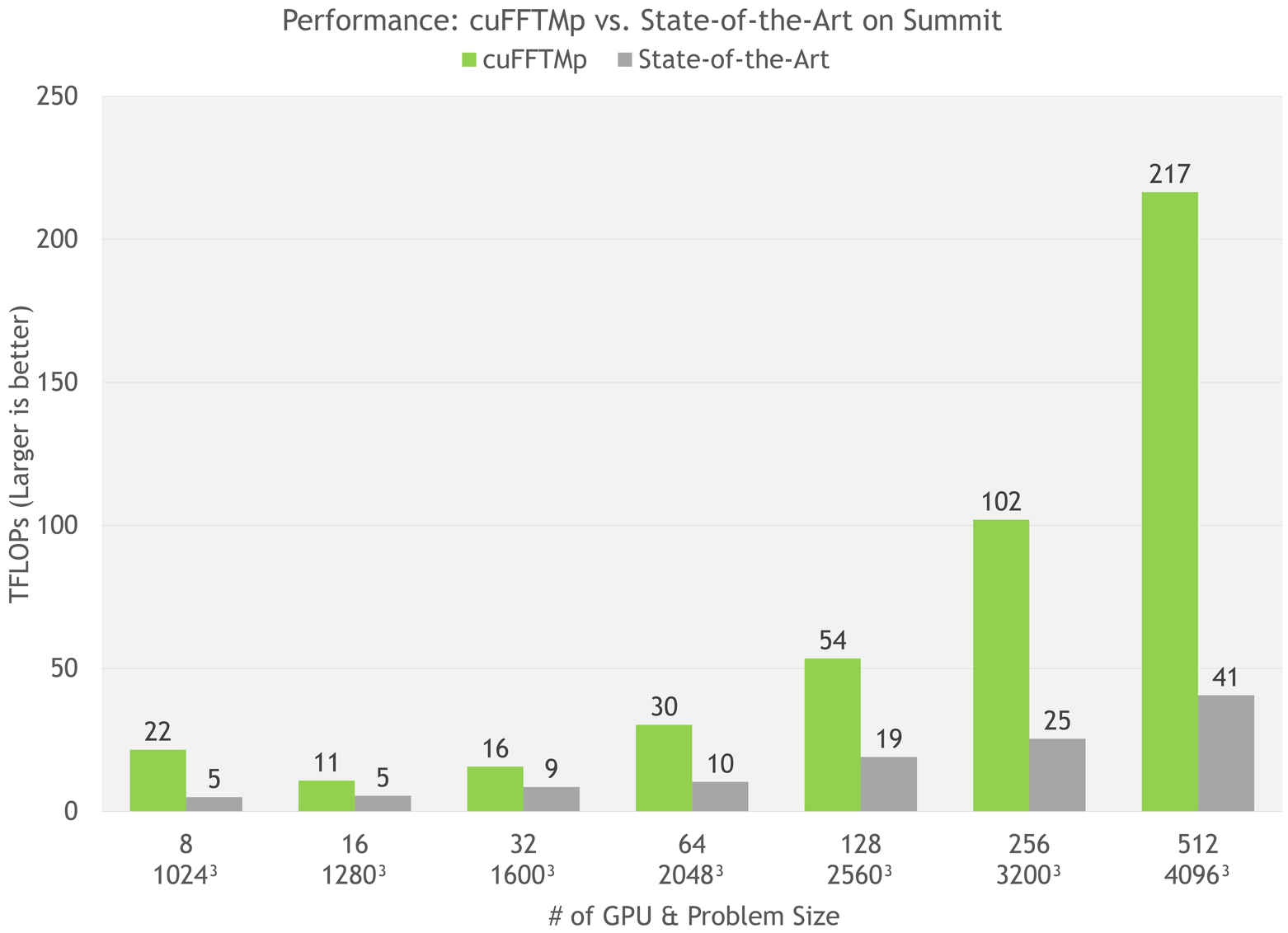


Composability
Device Functions



MULTI-NODE MATH LIBRARIES

cuFFTMp: Fast Fourier Transforms at scale



cuFFTMp

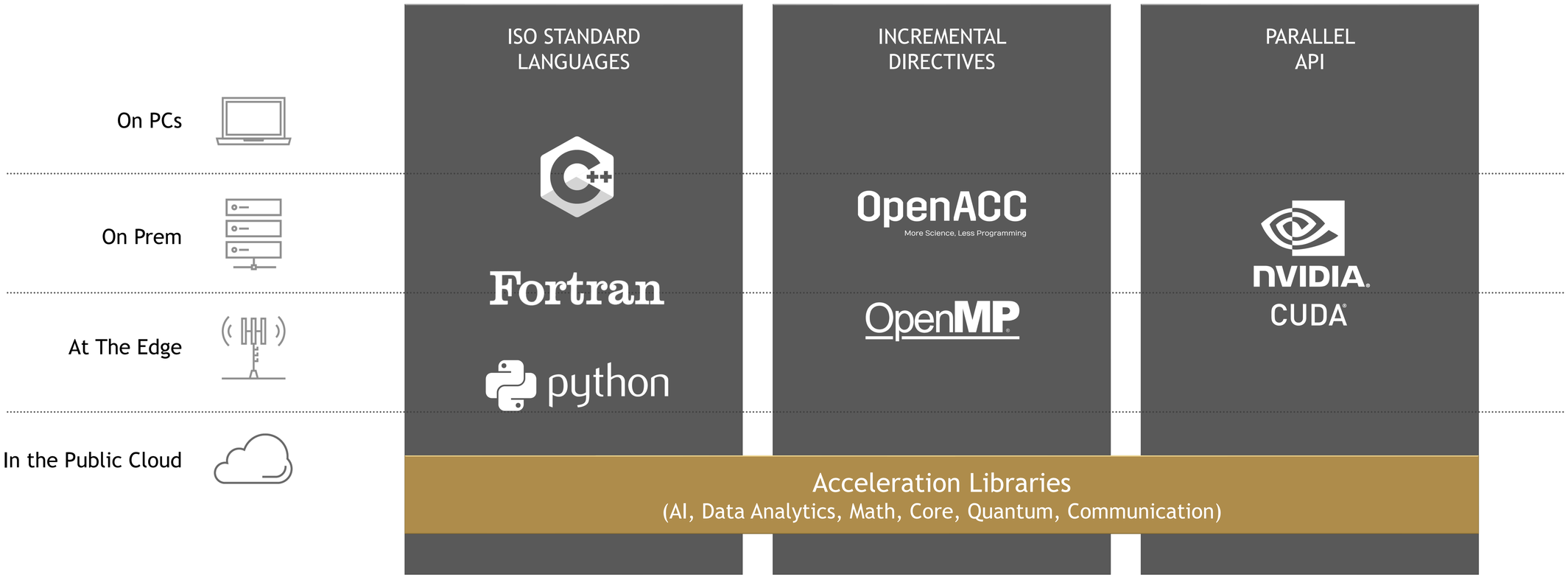
A distributed-memory multi-node and multiGPU solution for solving FFTs at scale.

EA release available in Fall ’21
https://developer.nvidia.com/cudamathlibraryea

Initial release to 2D & 3D with Slab composition

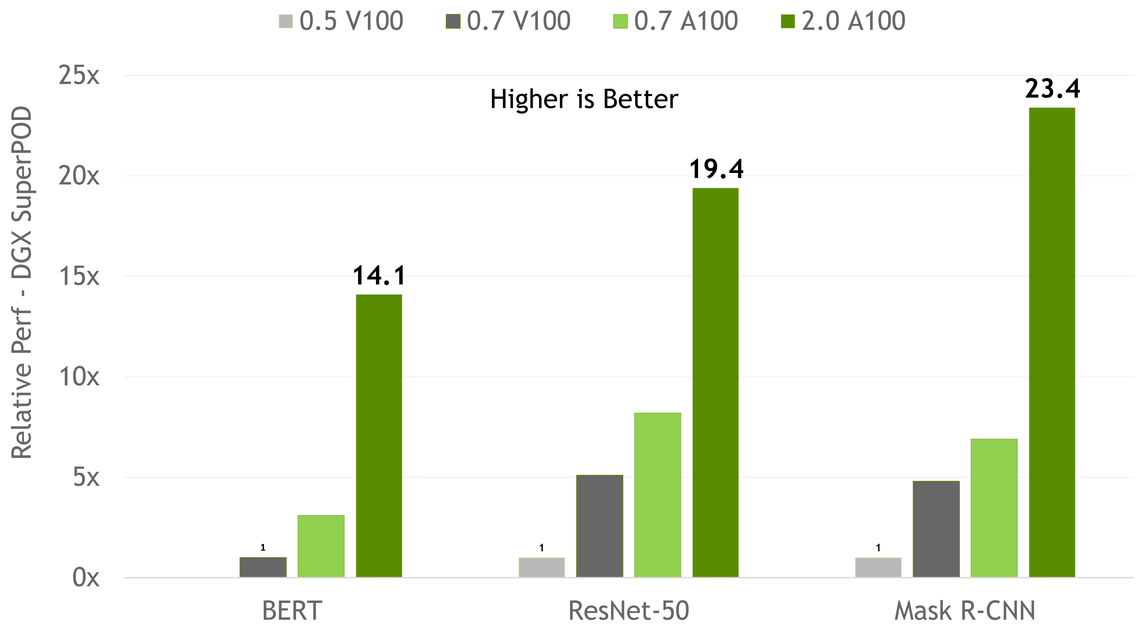
FLEXIBLE HPC PROGRAMMING OPTIONS

Develop How You Want, Where You Want



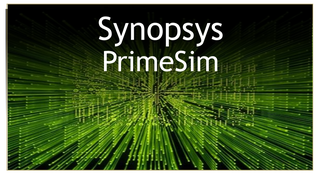
23X MORE PERFORMANCE IN 3.5 YEARS

NVIDIA AI Delivers Continuous Gains With SW and At-Scale Improvements

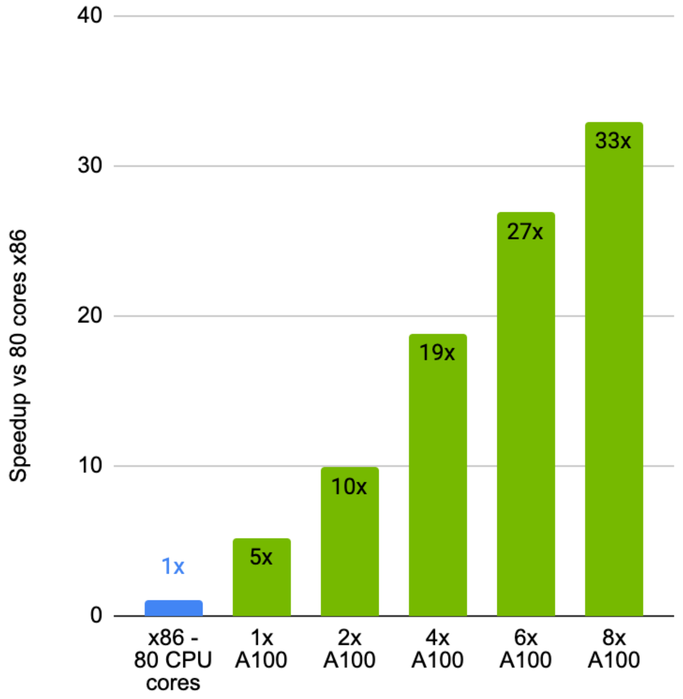


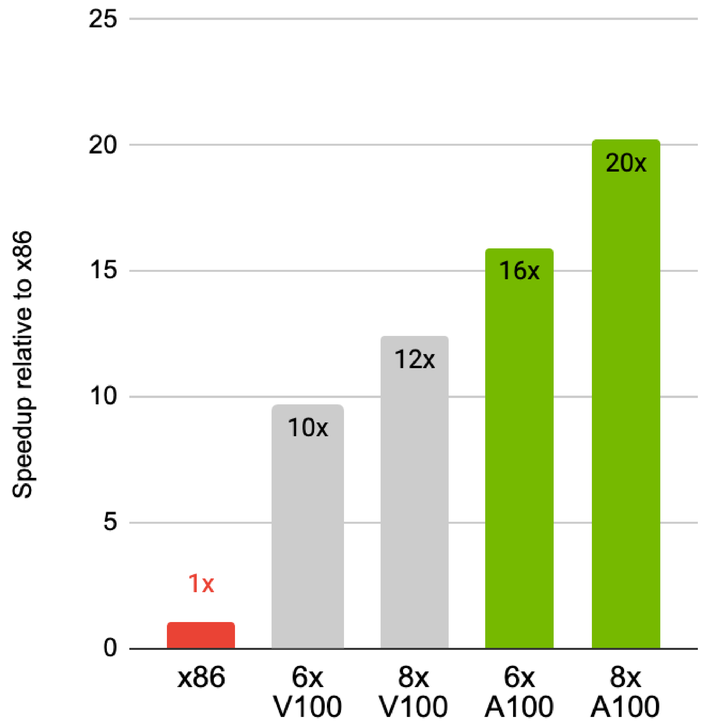
Results normalized for throughput due to higher accuracy requirements on latest round of MLPerf Training 2.0. MLPerf ID 0.5/0.7/2.0 comparison:
BERT: 0.7-56/0.7-38/2.0-2106 | ResNet50v1.5: 0.5-17/0.7-55/0.7-37/2.0-2107 | Mask R-CNN: 0.5-22/0.7-48/0.7-28/2.0-2099
MLPerf name and logo are trademarks. See www. .org for more information.

ACCELERATING INDUSTRIAL HPC SIMULATIONS



Altair
Ansys
cadence
CASCADE TECHNOLOGIES
DASSAULT SYSTEMES
HEXAGON
MathWorks
ptc
SIEMENS
SYNOPSYS
VASP





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

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

TRANSFORMER MODEL OPTIMIZATIONS

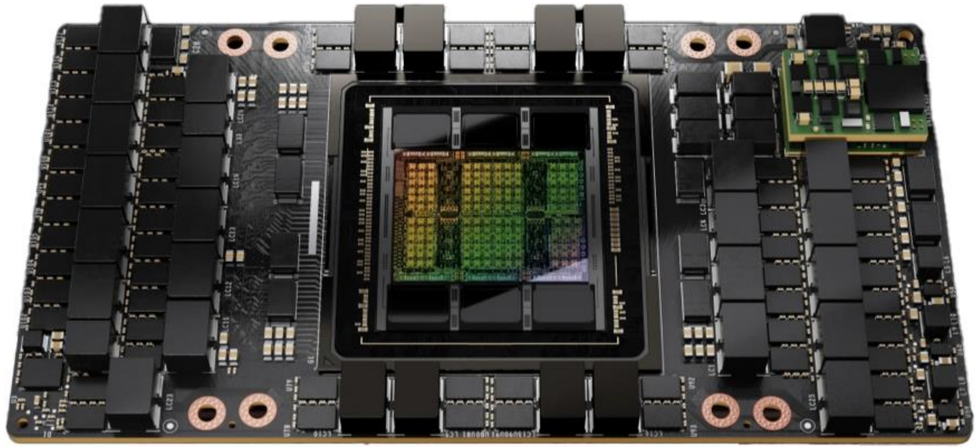
6X faster on largest transformer models

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



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

NVIDIA H100 CNX

Unprecedented performance for GPU-powered, IO-intensive workloads

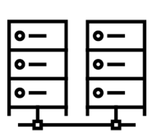
Unique Versatile Architecture
Combines H100 GPU and ConnectX-7 SmartNIC on a single board

Better I/O Performance
High-speed, contention-free GPU-Network data transfer

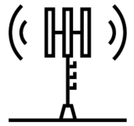
Optimized Design
Inherently balanced scale-out architecture

Cost Savings
High performance with mainstream servers

Application Ready
Acceleration benefits without application modification

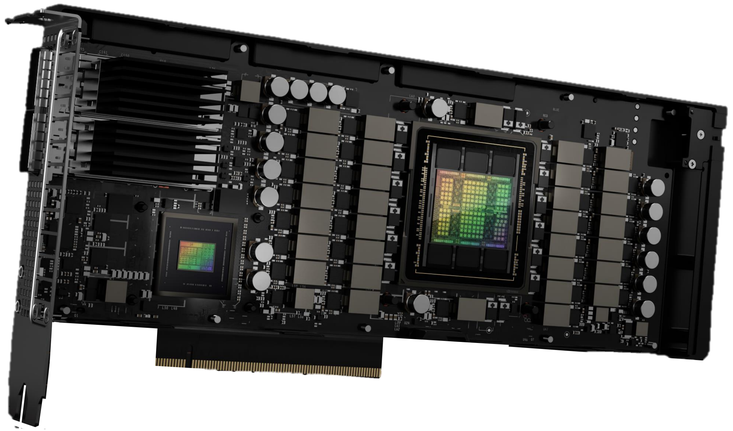


Multi-node Training



5G vRAN

AI on 5G



350W | 80GB | 400 Gb/s Eth or IB
PCIe Gen 5 within board and to host
2-Slot FHFL | NVLink

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

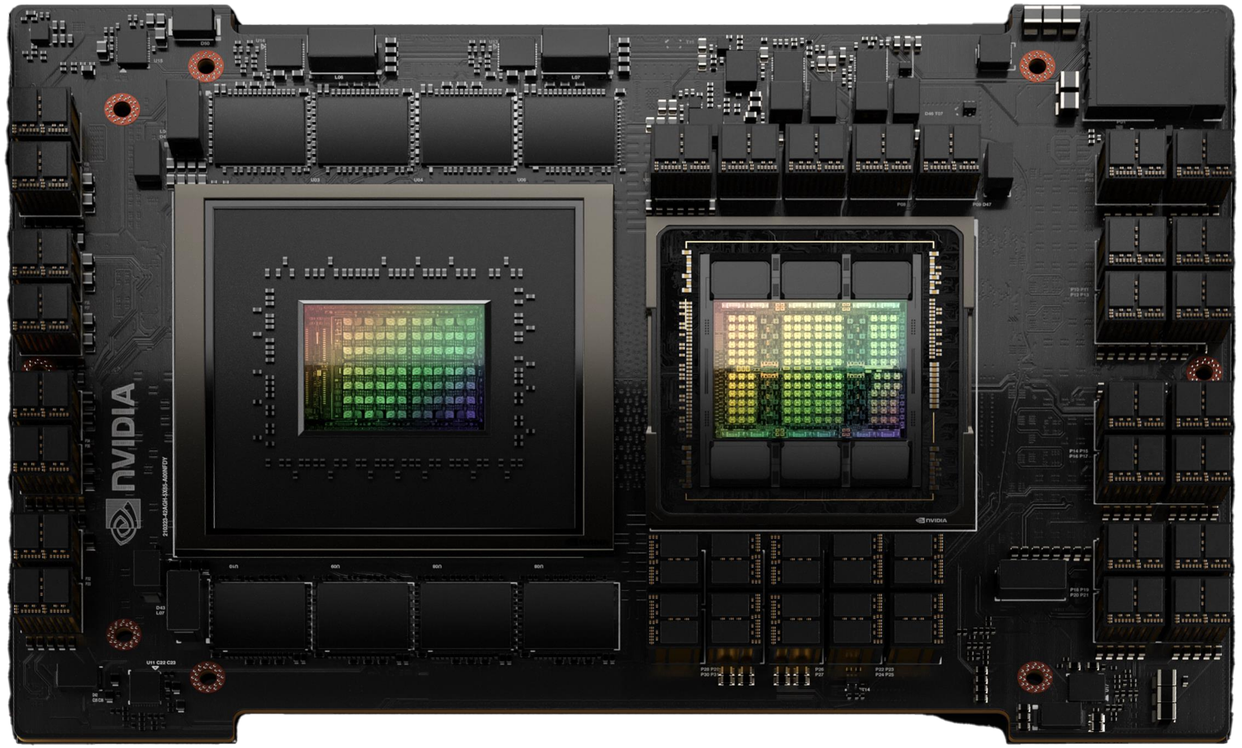
HIGHEST MEMORY BANDWIDTH 3.5GB/s
LPDDR5x and HBM3

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

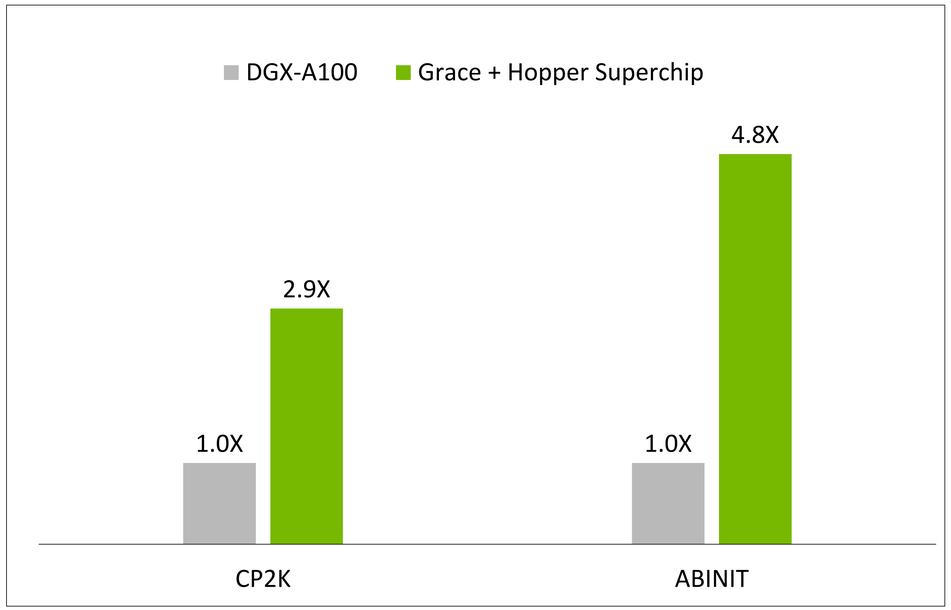
RTX, HPC, AI, Omniverse

AVAILABLE 1H 2023

DELIVERING UP TO 5X HPC HIGHER PERFORMANCE

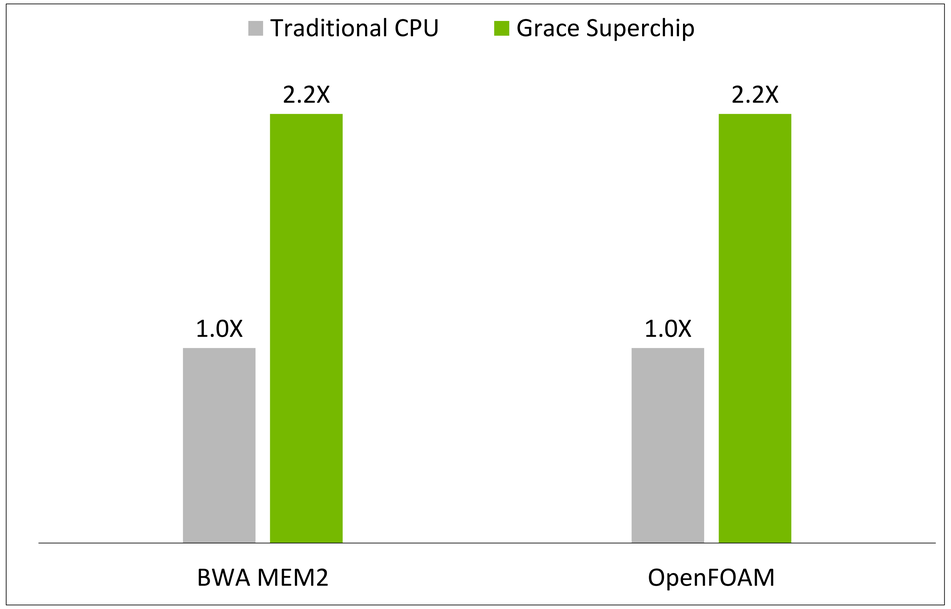
GRACE HOPPER SUPERCHIP

Accelerated HPC Application Performance



GRACE CPU SUPERCHIP

Nonaccelerated HPC Application Performance



Grace Hopper Superchip and Grace CPU Superchip Pre-silicon ESTIMATES ONLY: | Performance comparisons: (left) DGX-A100 (4x A100 SXM4 80 GB) | CP2K version 9.1 Random Phase Approximation dataset H2O-128-RI-dRPA-TZ.inp | ABINIT version 8.10.3 dataset Titanium 255 Atoms, LOBPCG algorithm | (Right) Traditional CPU Intel Ice Lake 8380 | BWA MEM2 v2.2.1 from the git repo running Full Human Genome (human_g1k_v37) | OpenFOAM version 2112 for model Large Motorbike v1912

GRACE REFERENCE DESIGNS FOR MODERN DATA CENTER WORKLOADS

CGX	OVX	HGX	HGX
Cloud Graphics & Gaming	Digital Twins & Omniverse	HPC	AI Training, Inference & HPC
Grace CPU Superchip	Grace CPU Superchip	Grace CPU Superchip	Grace Hopper Superchip
A16	NVIDIA GPUs	BlueField-3	BlueField-3
BlueField-3	BlueField-3	OEM Defined IO	OEM Defined IO / 4th Gen NVLink

DGX SUPERPOD WITH NVLINK SWITCH SYSTEM

Purpose Built for Giant NLP, DLRM, Scientific Computing (3DFFT), MoE Models

Availability

2023
Only sold as part of DGX SuperPOD solution
for enterprises and CSPs.

Scale Up AI and HPC

Delivers massive collective communication performance
9X bandwidth compared to HDR InfiniBand

H100 CLUSTER (1 SCALABLE UNIT)

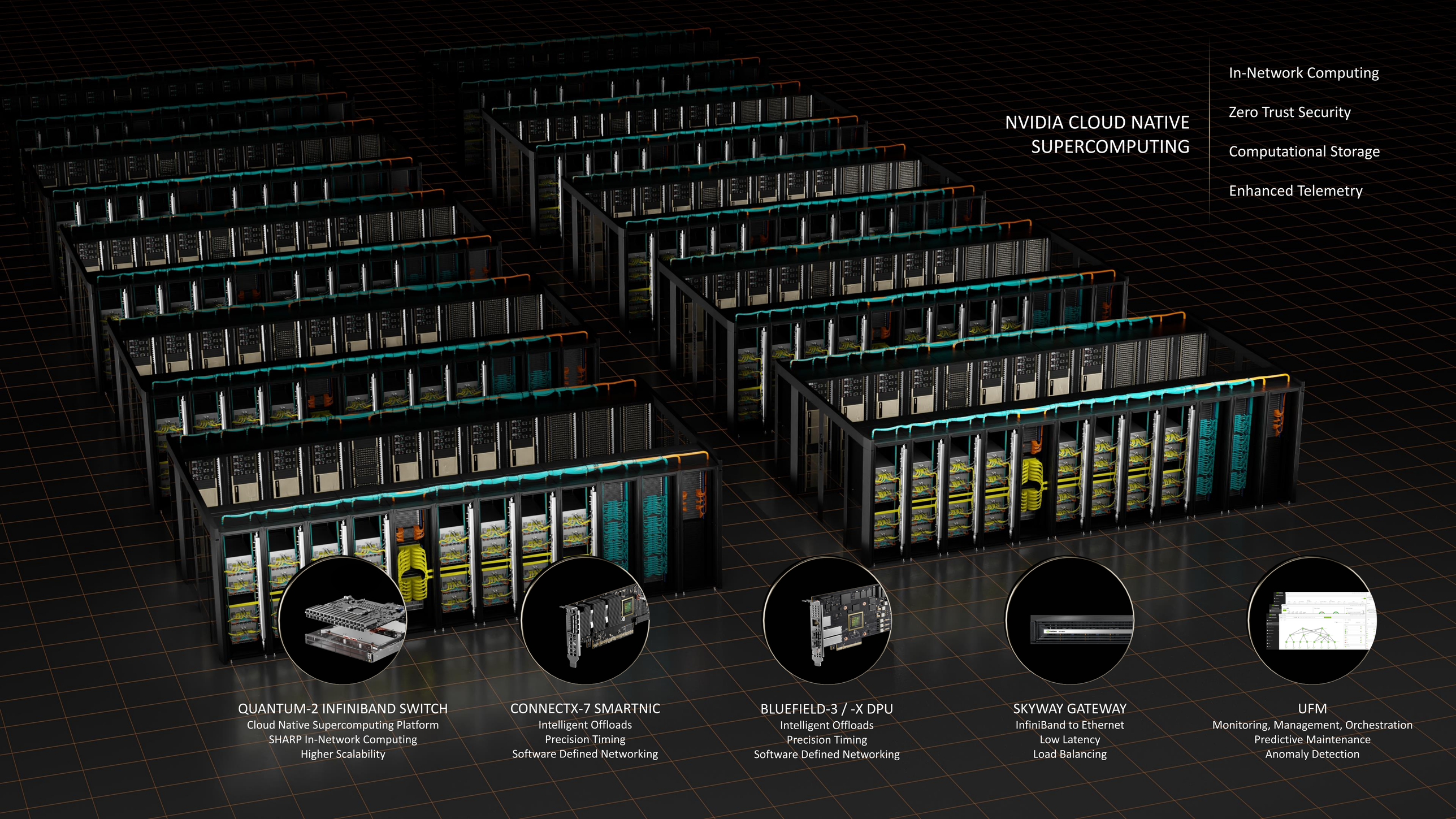
57.6 TB/s Bi-section Bandwidth
32 servers | 18 NVLink switches | 1,152 NVLink optical cables

Scale Out Compute and Storage Communications

uses Quantum-2 NDR InfiniBand

NVIDIA CLOUD NATIVE SUPERCOMPUTING

- In-Network Computing
- Zero Trust Security
- Computational Storage
- Enhanced Telemetry



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

RADICAL IMPROVEMENTS IN CPU EFFICIENCY

Grace Delivers On Both Compute And Memory Bandwidth

Node Configuration	2x 96c Genoa *Estimated*	Single Grace Superchip *Projected*
General-Purpose FP64	7.4 Gflops	7.4 Gflops
Memory Bandwidth	1 TB/s	1 TB/s
CPU + Memory Power	900W	**500W**
Bandwidth per Core	5.3 GB/s	**7.0 GB/s**

For 5 Petaflops of HPC capacity (~2 GB/core):

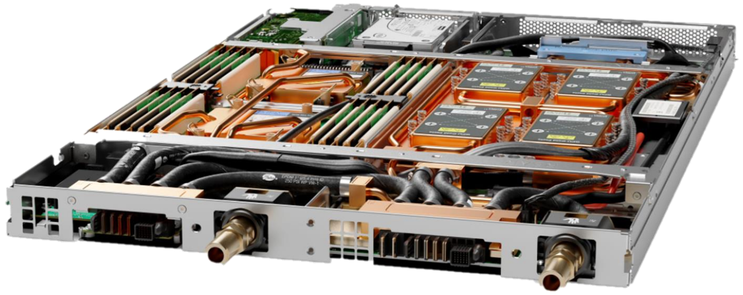
Datacenter Power Needed	1.6x	**1x**
5-Year TCO Savings ($.25/KWH, Estimated: Power, Cooling, CapEx)	-	**$2.3 Million**
Number of Cabinets (300 KW LC or similar: 2.5 of Genoa, 1.5 of Grace)	3	**2**
Lifetime CO_2 Savings (Assuming 5-Year Deployment)	-	**10,000 Metric Tons**

Grace Superchip vs. 2S X86-64 + Memory
1.8x Perf / Watt.

Big Results
35% Lower OpEx
50% Improved Density
Save Millions in TCO
...and CO_2 emissions equivalent to **87,000 trees.**

LIQUID-COOLED A100 TENSOR CORE GPU PORTFOLIO

A100 Offered with Liquid-Cooled Options



A100 HGX
Shipping from OEMs

A100 PCIe
Q3 2022

ENERGY & SPACE EFFICIENCY
A100 PCIE LIQUID COOLED VS AIR COOLED

66%

LESS RACK SPACE

30%

LOWER POWER CONSUMPTION

<1.2
PUE

EASILY MEET EFFICIENCY TARGETS

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

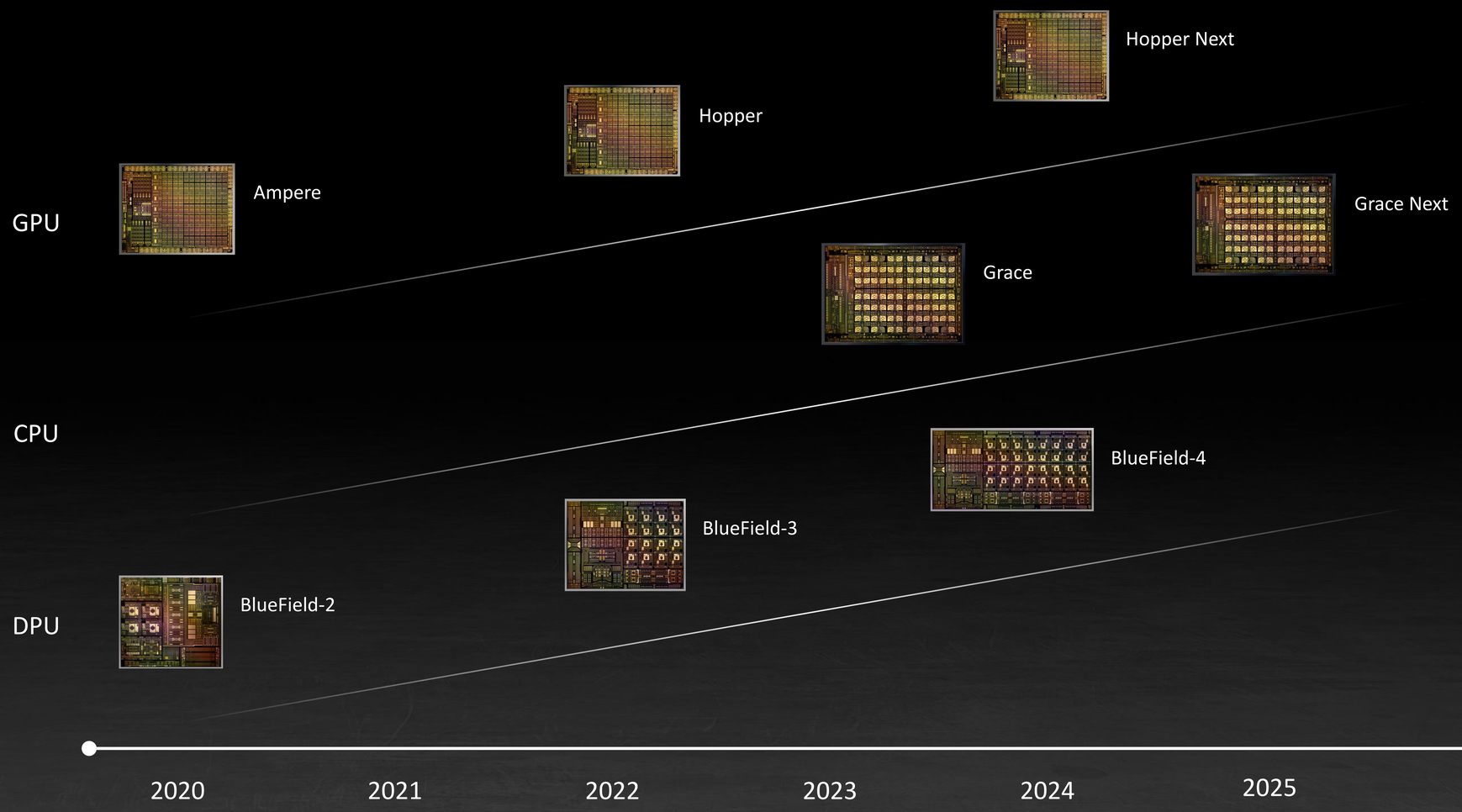
CINECA (Leonardo)

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

3 CHIPS. YEARLY LEAPS. ONE ARCHITECTURE.



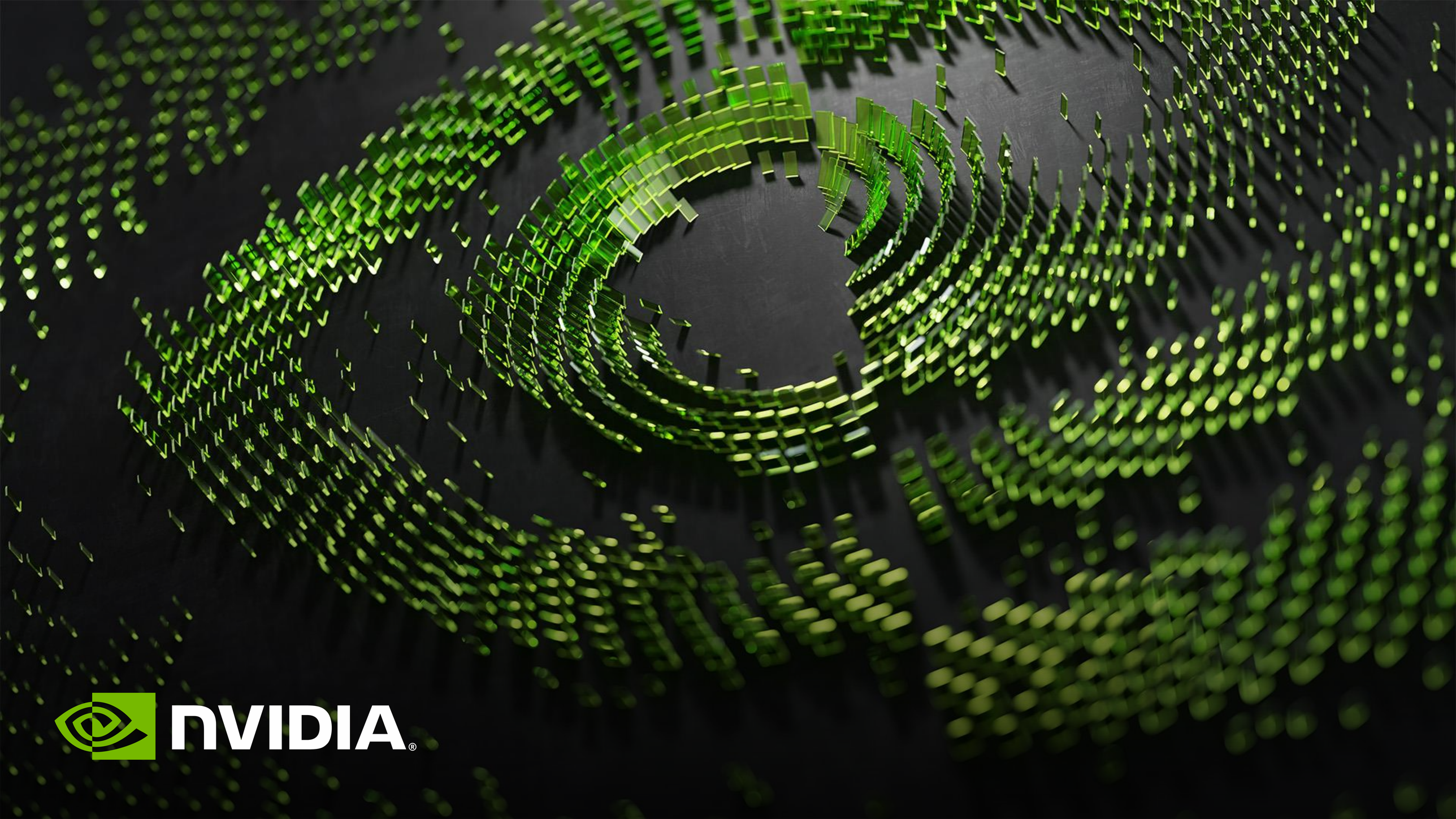
Full Stack. Data Center Scale

2,700 Accelerated Applications

450 SDKs, AI Models

30 Million CUDA Downloads

3 Million Developers


NVIDIA