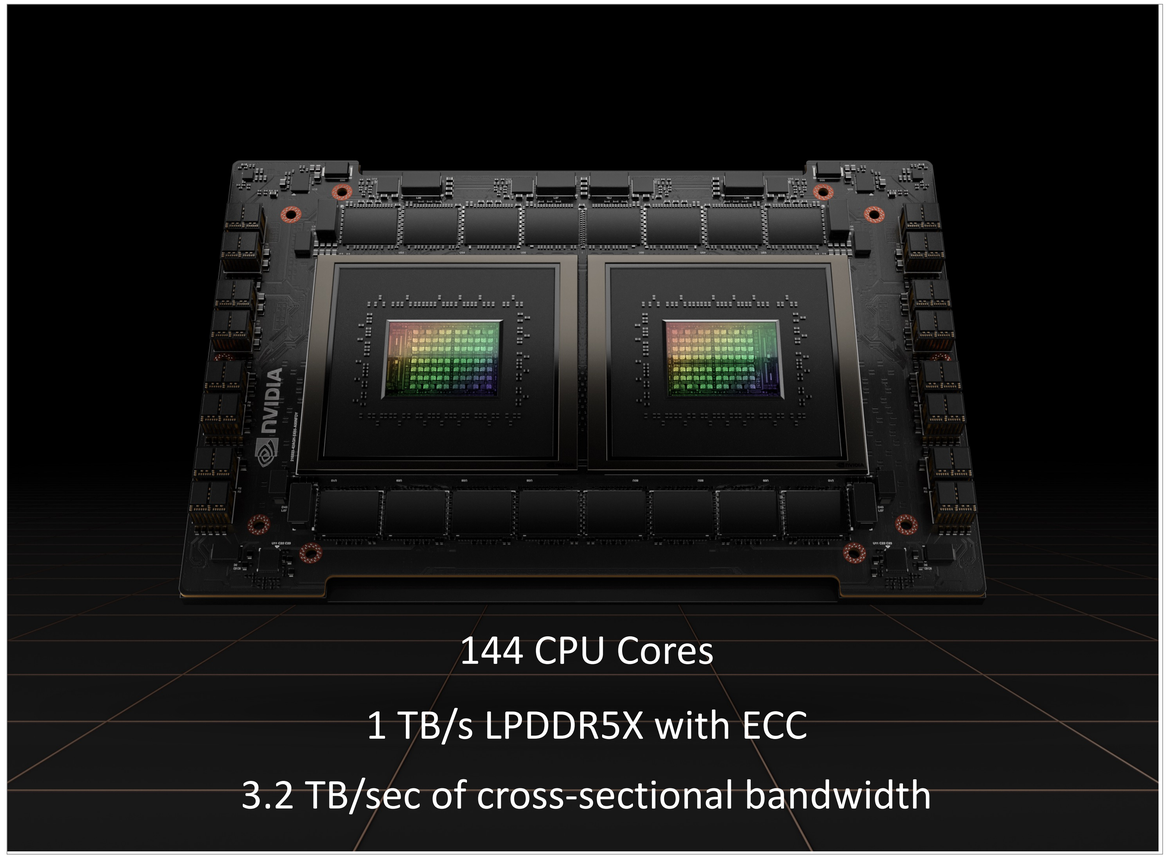
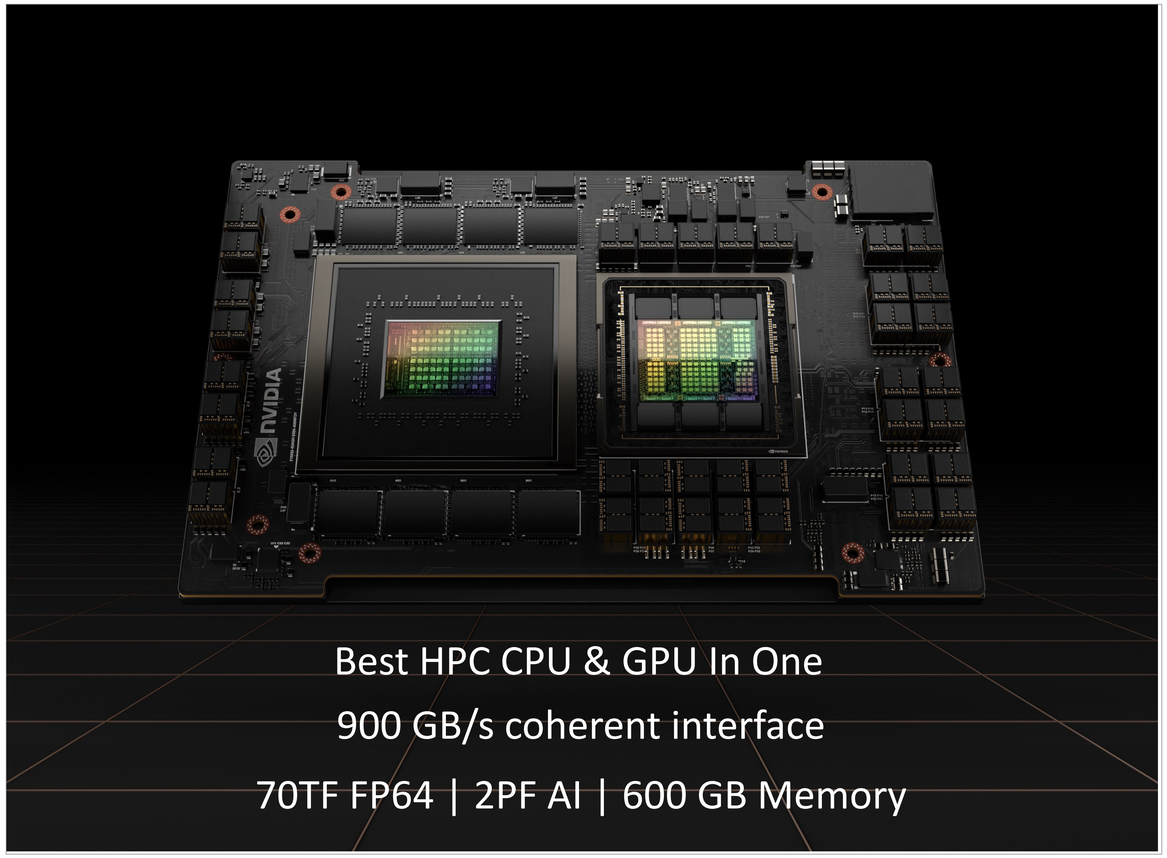
GH200 (Grace Hopper)

NVIDIA Grace and Grace Hopper

High Performance for an Energy Constrained World



Grace CPU Superchip
High-performance CPU for HPC and cloud computing



Grace Hopper Superchip
CPU+GPU designed for giant-scale AI and HPC

Grace CPU Superchip

Delivers Superior Performance and Efficiency for HPC

NVIDIA Grace CPU Superchip

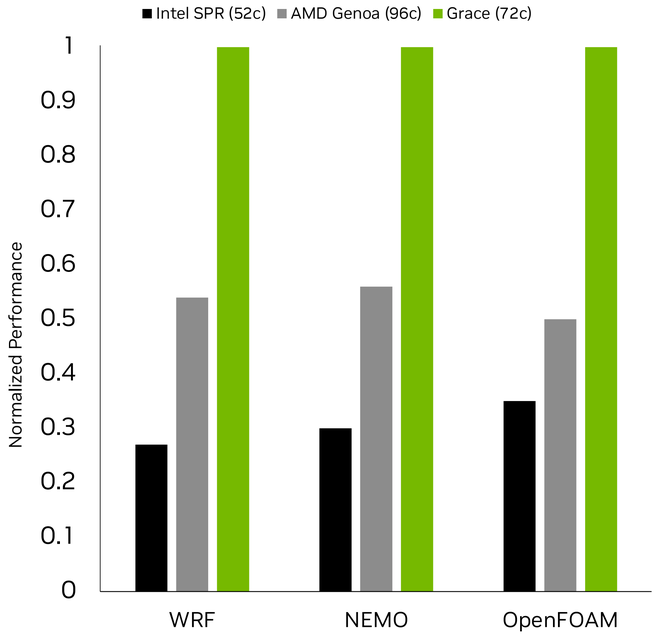
High Performance Power Efficient Cores
144 Arm Neoverse V2 Cores with SVE2 4x128b

High-Bandwidth Low-Power Memory
Up to 960GB of data center class LPDDR5X Memory that
delivers up to1TB/s of memory bandwidth

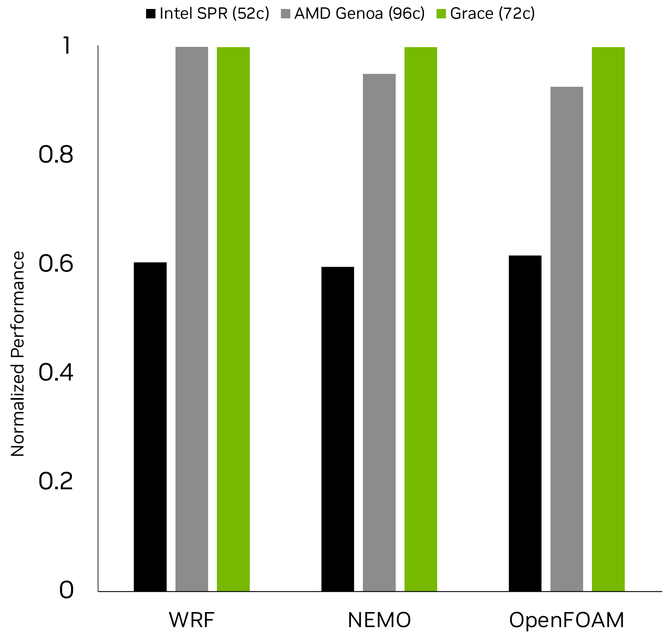
Fast On-Chip Fabric
3.2 TB/s of bi-sectional bandwidth connects
CPU cores, NVLink-C2C, memory, and system IO

Full NVIDIA Compute Stack
HPC, AI, Omniverse

Performance per Watt



Performance

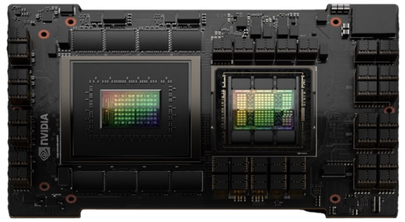


WRF 4.4.2 ; dataset: CONUS12, 24hr simulation; NVIDIA Grace Superchip: 4 ranks/18 threads, Intel SPR 13 ranks/4 threads, AMD Genoa 12 ranks/8 threads | NEMO 4.2.0; dataset: GYRE_PICES(25) 50 steps | OpenFOAM v2212 dataset: Motorbike Large
NVIDIA Grace measurement on preproduction hardware | AMD EPYC 9654 (Genoa) / 96 cores / 768 GB RAM / 400W cTDP | Intel(R) Xeon(R) Platinum 8470Q (Sapphire Rapids) / 52 cores / 512 GB RAM / 350 TDP | Energy comparisons are based on CPU socket power + DRAM power
Intel and AMD Socket power were captured using Intel PCM and AMD uprof telemetry tools | Grace system was measured by total CPU & Memory power before module voltage regulator losses

Grace Hopper Superchip

The breakthrough accelerated CPU delivering performance and efficiency

NVIDIA Grace Hopper Superchip



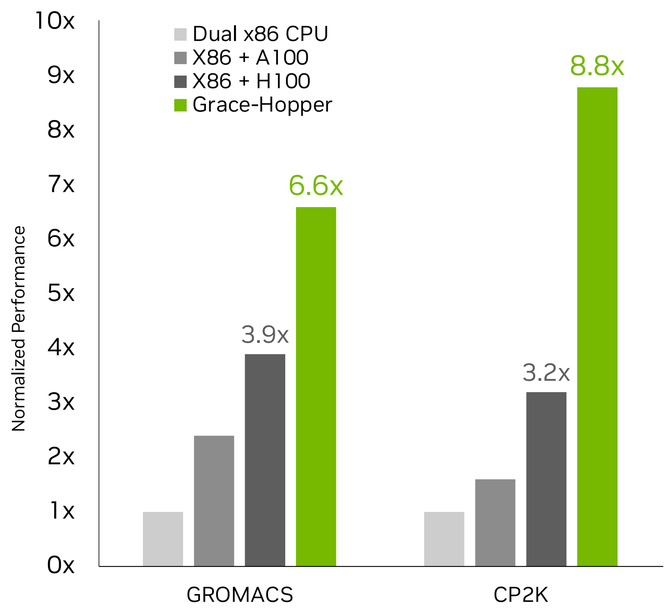
Grace CPU + H100 GPU
72 Arm Neoverse V2 Cores with SVE2 4x128b
Transformer Engine and ~4PFLOPS of FP8

Fast NVLink-C2C Connection
900GB/s bi-directional bandwidth CPU to GPU
7X faster than PCIe Gen 5

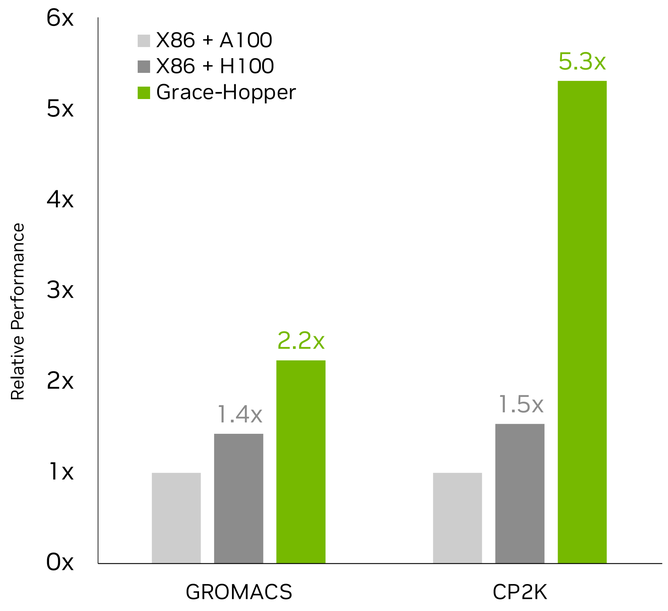
~600GB of Fast Access Memory
Up to 96GB HBM3, 4TB/s bandwidth
Up to 480GB LPDDR5X, 512GB/s bandwidth

Full NVIDIA Compute Stack
HPC, AI, Omniverse

Performance per Watt



Performance



GROMACS version 2023.1 dataset: STMV | CPK version RPA dataset: H2O-32-RI-dRPA-96points.
Baseline [Center Panel] CPU Intel SPR 8480C 52c / socket, dual socket server. Baseline [Right Panel] NVIDIA HGX-A100 80 GB SXM 4 GPU server. NVIDIA HGX-H100 80 GB SXM 4 GPU server. Single server Grace-Hopper Superchip
NVIDIA Grace-Hopper Superchip performance based on engineering measurements on production hardware. CPU data extrapolated from Dual Xeon Gold 6420 at 2.6 GHz performance measurements. Results subject to change.

Announcing Grace Hopper Now in Full Production

Grace and Grace Hopper transforming HPC and AI

LANL (Venado)
Grace Hopper
10 EFLOPS AI Perf

Univ. of Bristol (Isambard 3)
Grace CPU
2 PFLOPS HPC Perf

BSC (MareNostrum 5)
Grace CPU
2 PFLOPS HPC Perf

CSCS (ALPS)
Grace Hopper
20 EFLOPS AI Perf

KAUST (Shaheen-III)
Grace Hopper
7 EFLOPS AI Perf

NCHC (Taiwania 4)
Grace CPU
300 TFLOPS HPC Perf

GH200 GRACE HOPPER SUPERCHIP

The breakthrough accelerated CPU for Large-Scale AI and HPC applications

Grace CPU + H100 GPU

72 Arm Neoverse V2 Cores with SVE2 4x128b
Transformer Engine and ~4PFLOPS of FP8

Fast NVLink-C2C Connection

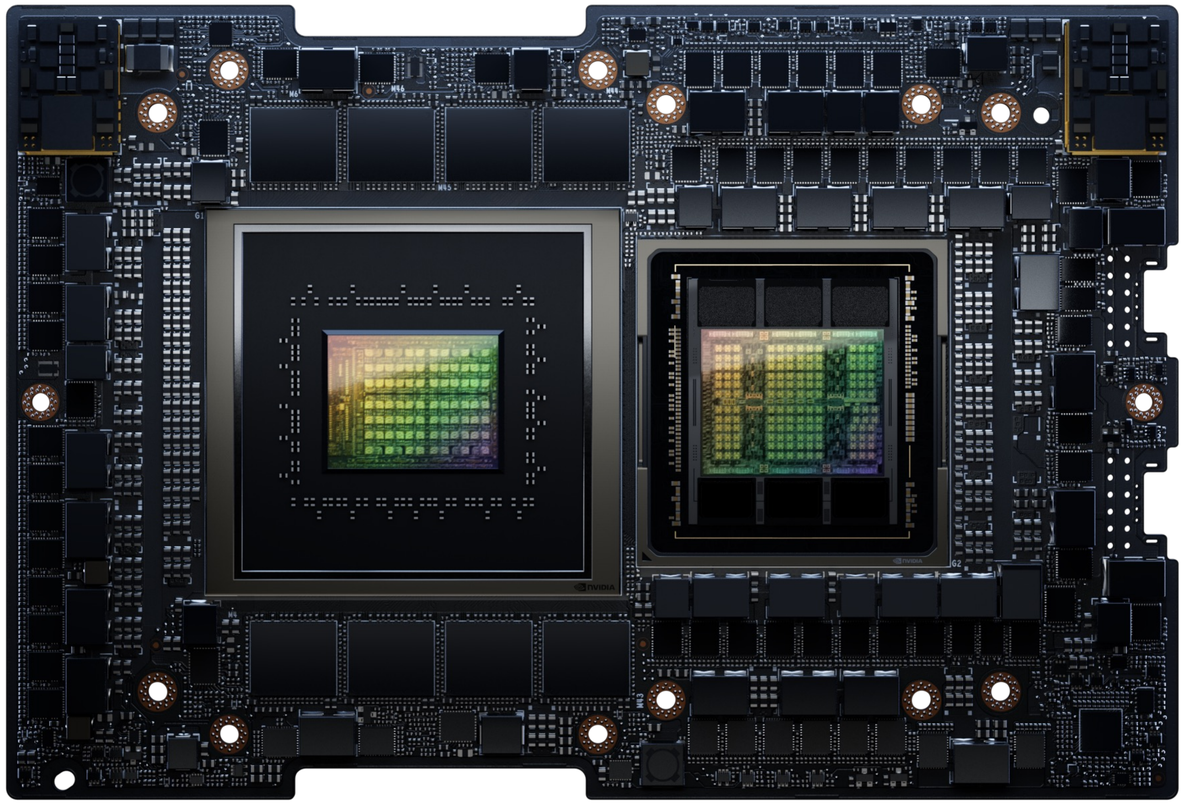
900GB/s bi-directional bandwidth CPU to GPU
7X faster than PCIe Gen 5

~600GB of Fast Access Memory

Up to 96GB HBM3, 4TB/s bandwidth
Up to 480GB LPDDR5X, 512GB/s bandwidth

Full NVIDIA Compute Stack

AI, Omniverse

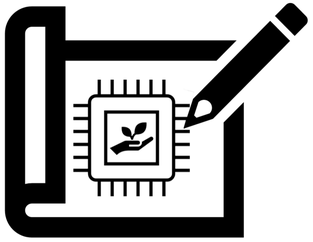


Optimizing Power and Performance at Data Center Scale

Higher Performance for Less Power

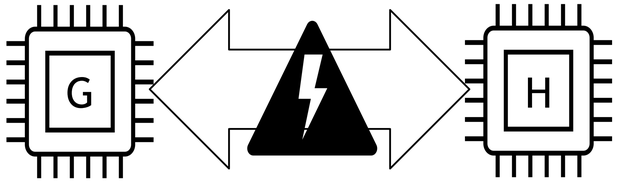
Energy Efficient Design

Energy Efficient CPU, GPU, Memory, IO



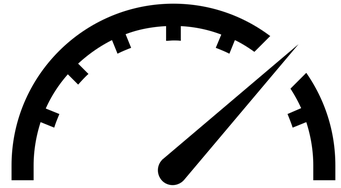
Automatic Power Shifting

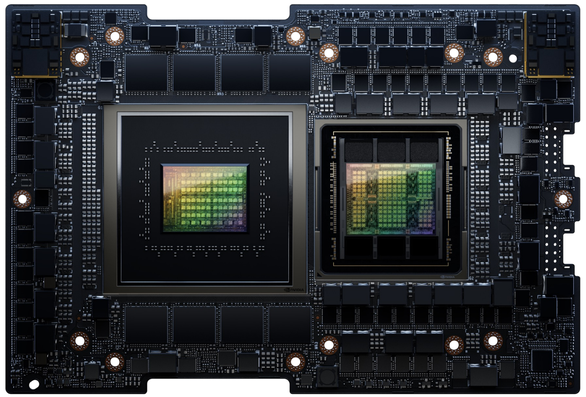
Automatically shifts power between CPU & GPU



Application Power Tuning

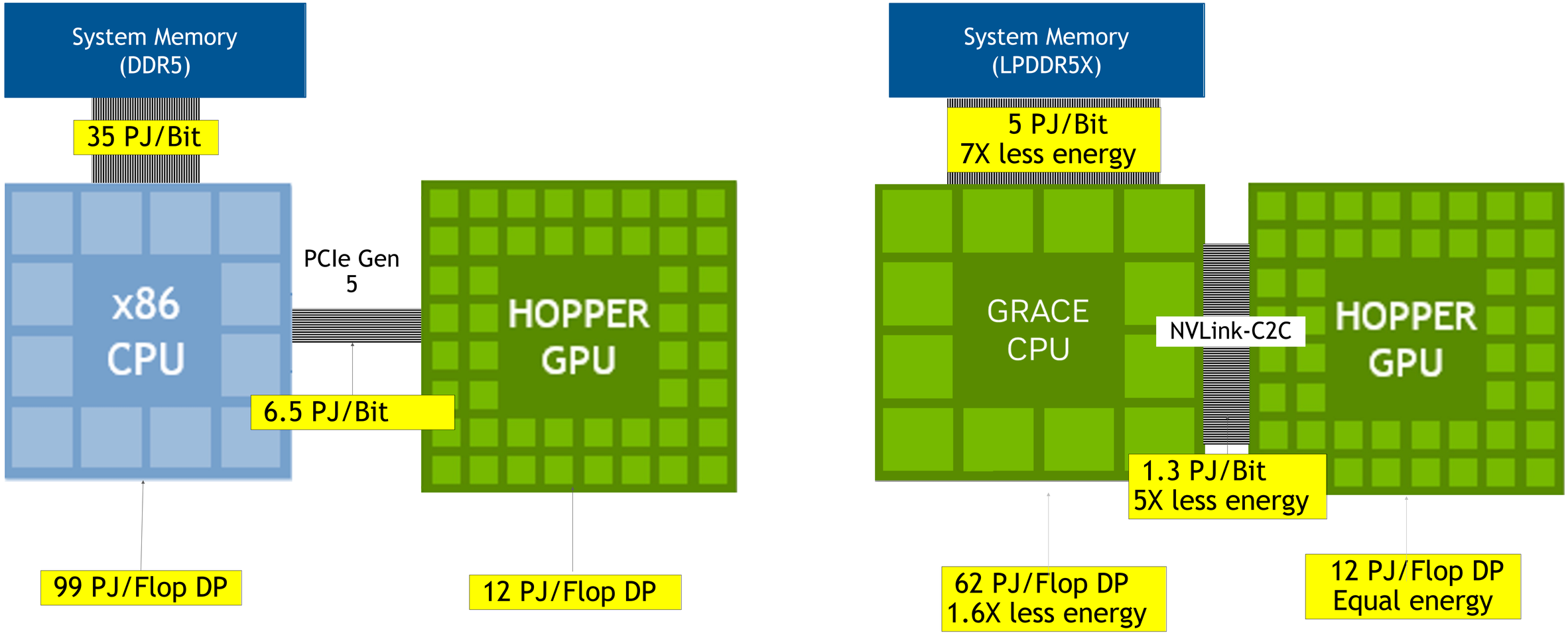
Adjustable clocks for improved energy efficiency





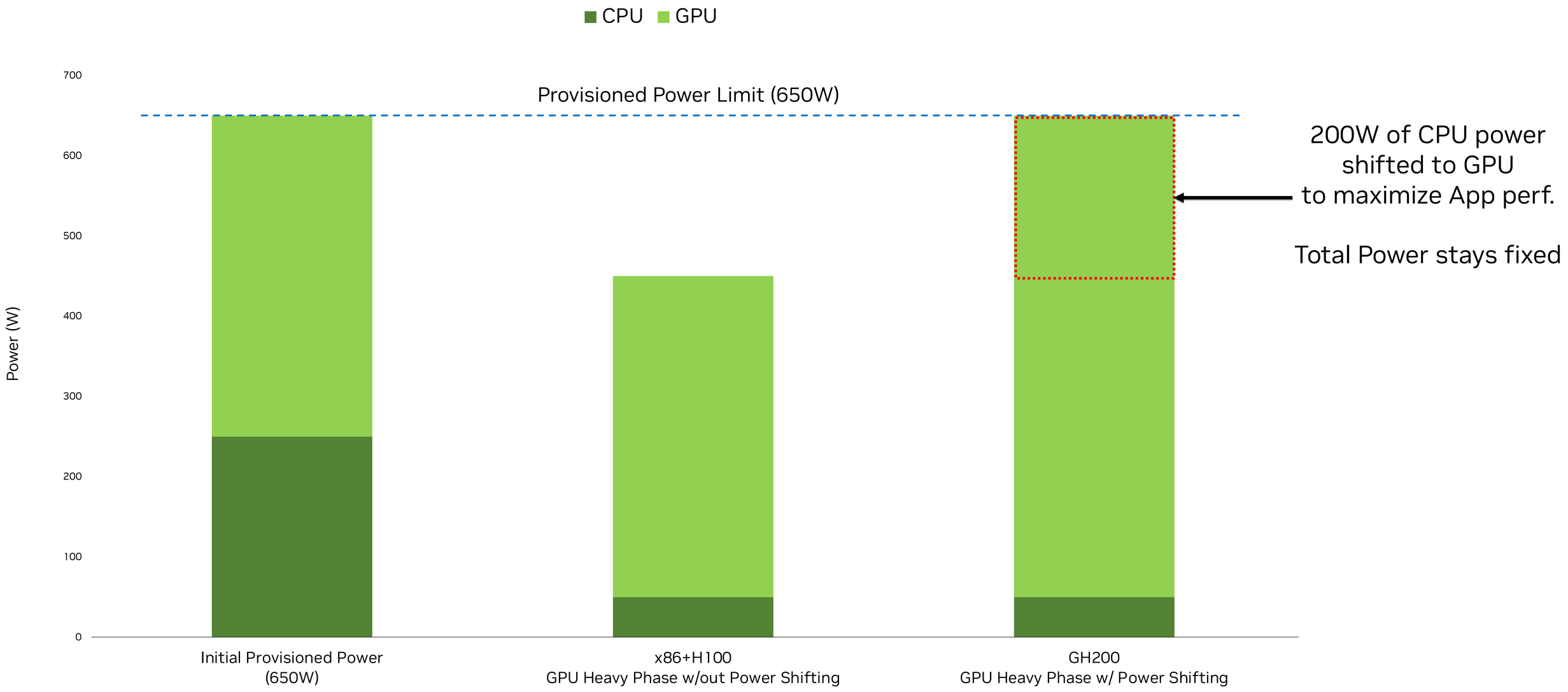
Energy Efficient Design

More Efficient Computation and Data Movement



Optimizing Performance Through Power Shifting

Getting the Most Out of provisioned power

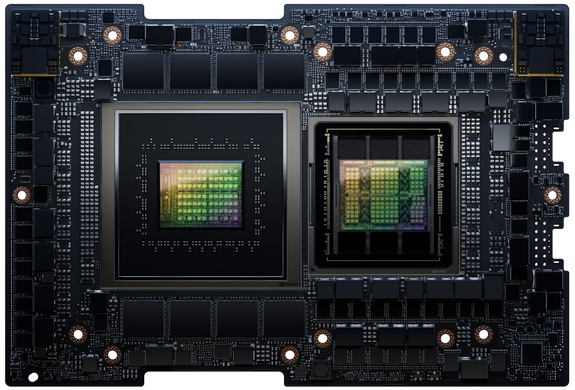


GH200 Grace Hopper HPC Platform

Unified Memory and Cache Coherence for Next Gen HPC Performance

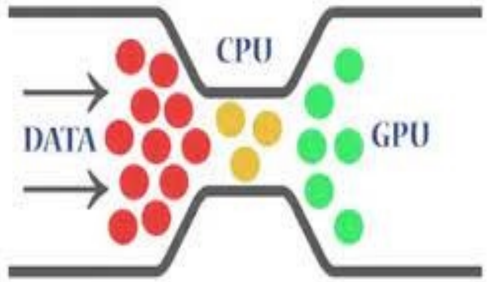
Partially GPU Accelerated Apps

Big performance gains with no code changes



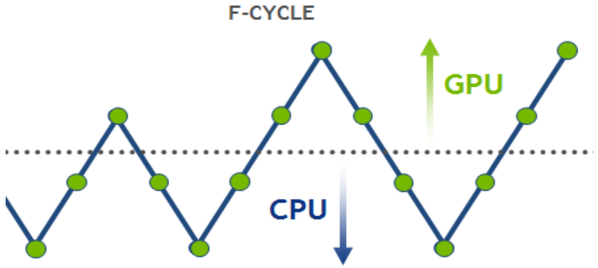
No More PCIe Bottleneck

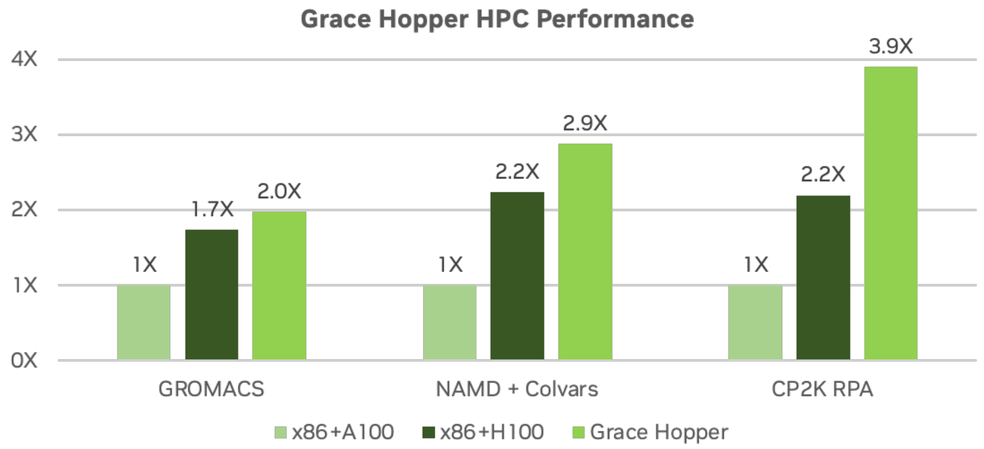
NVLink-C2C is 7X PCIe BW



CPU & GPU Cache Coherence

Incremental code changes yield big gains





Fast Access Memory

576GB

Memory Bandwidth

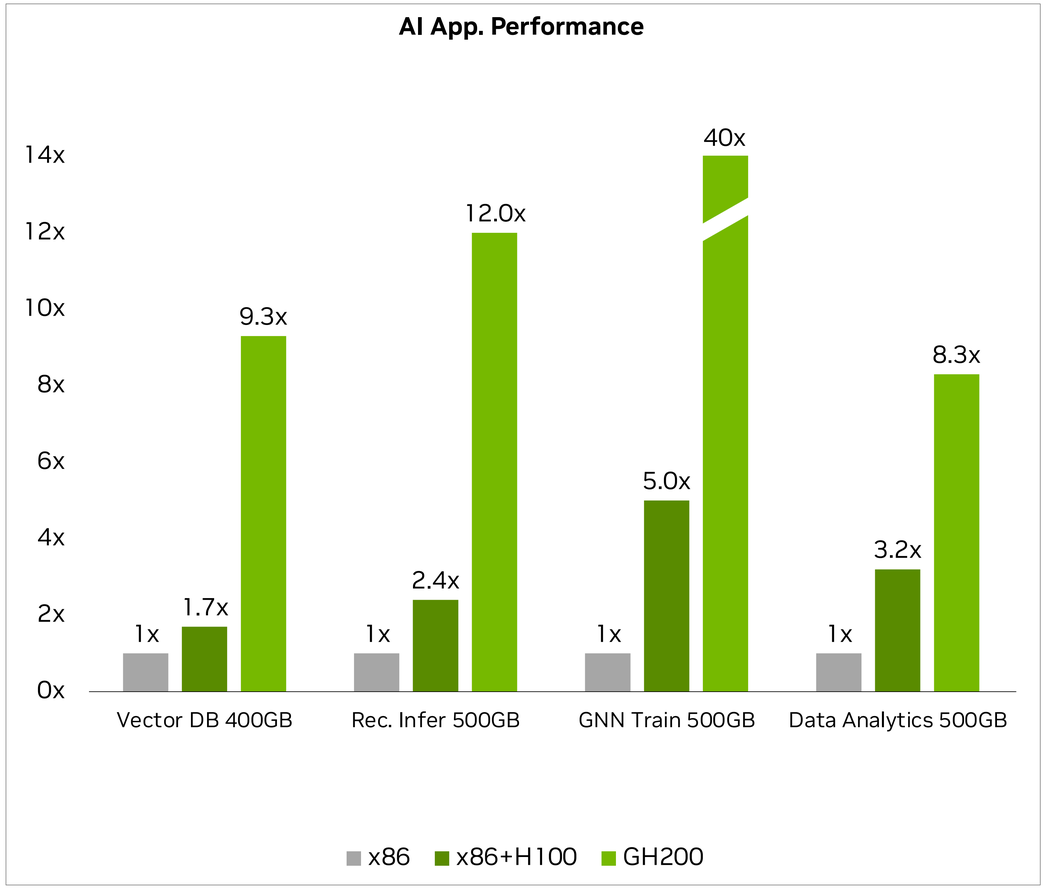
4TB/s

HPC: Preliminary results comparing DGX A100, DGX H100 and Grace Hopper systems.
AI: Simulation results comparing DGX H100 and Grace Hopper systems. GNN: GraphSAGE model for an augmented ogbn- products dataset of 626M nodes and 31B edges, DRLM: Deep Learning Recommendation Model (DLRM) 27TB model, NLP: fine-tuning a GPT-175B model

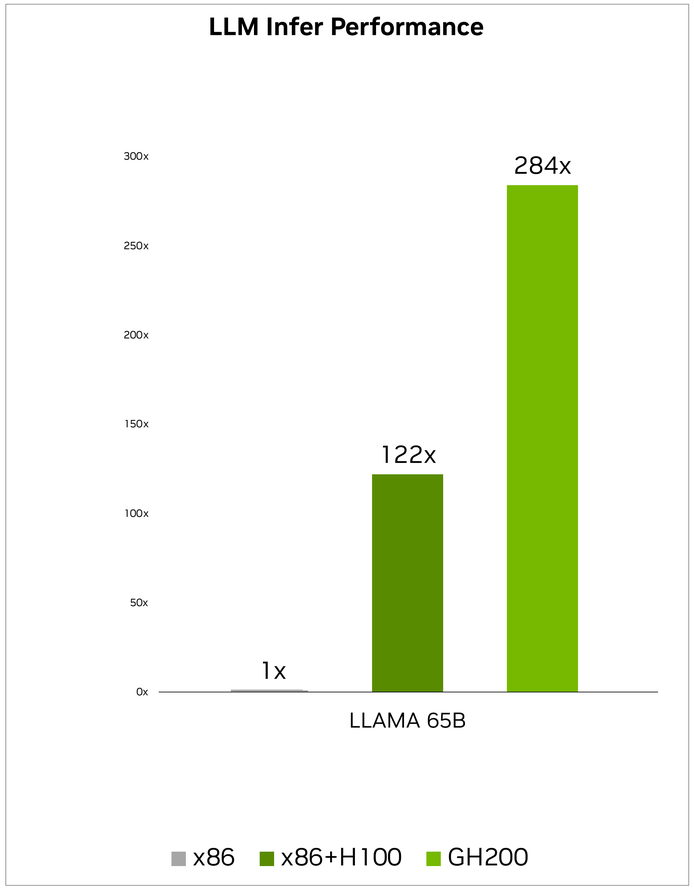
GH200 Grace Hopper AI Inference Platform

Versatile Scale Out with Unmatched Performance

Memory Intensive

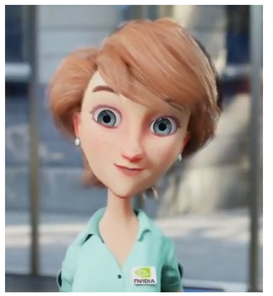


GPU Intensive



Use Cases

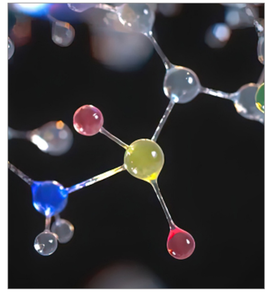
LLM
Conversational AI
Domain Knowledge



Recommender Systems
eCommerce
Personalized Content

Vector Database
Fraud Detection
Drug Discovery

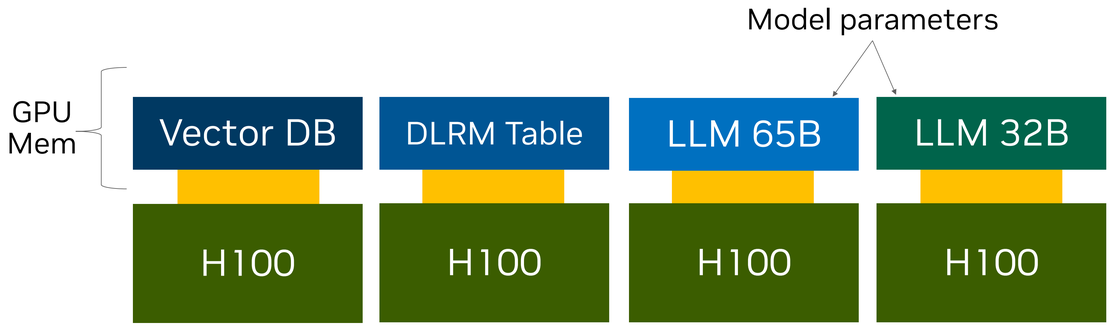
GNN
Computer Vison
Recommenders



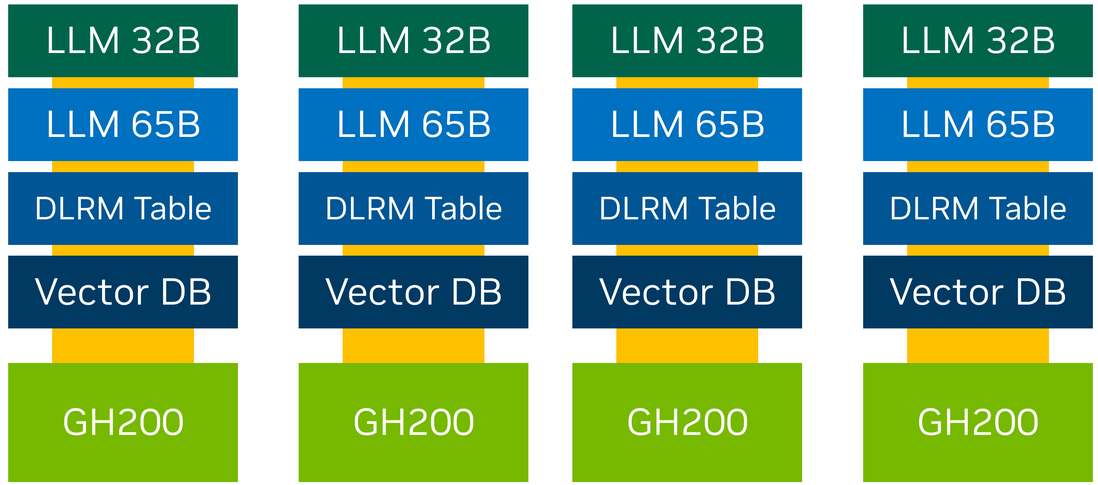
x86 CPU is SPR, A100/H100 are PCIe cards, except A100/H100 SXM used for HPC benchmarking
Vector DB search (IVF-Flat-MM used), DLRM Rec infer, DA are projection based on scaling kernel measurements
Graph Neural Network, LLM and HPC based on full app measurements
OpenFOAM based on MotorBike, NAMD with Colvars, CP2K with RPA, DA with TPC-H query4

GH200 Maximizes Utilization in Diverse Workload Environment

Typical GPU Data Center



GH200 Data Center



- Typical data center supports diverse workload using many models of varying sizes
- PCIe GPU has limited GPU memory and different models need to be stored by different GPUs
- Usage patterns cause some GPU to stay idle, while others are over-subscribed

- GH200 with 576GB GPU memory can store all models on every node
- Every GH200 runs everything - scheduling is easier, scale out more efficient
- **Maximizing data center utilization**

DGX GH200

The Need for Massive Memory AI Systems

Common challenges associated with giant workloads

AI Training

Recommender Systems, Giant Models

SOCIAL MEDIA

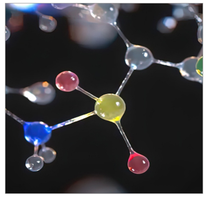
E-COMMERCE

Data Analytics

Graph Neural Networks

FRAUD DETECTION

DRUG DISCOVERY



Generative AI

Emerging AI Models

MULTI-MODAL GENERATIVE MODELS

CONTENT GENERATION

CHALLENGE

1. Slower output, models could take months to train instead of weeks, limiting progress
2. Reduced model features due to memory limitations with larger datasets
3. Impossible to address certain problems due to CPU-GPU memory access bottleneck

Introducing NVIDIA DGX GH200

AI supercomputer for the generative AI era

Giant Memory for Giant Models

- New class of AI supercomputer powered by 256 Grace Hopper Superchips
- 1 giant GPU and giant pool of memory for simplified programming model

Super Power-Efficient Computing

- Eliminates CPU to GPU PCIe bottleneck
- Reduces interconnect power consumption, increases bandwidth

Turnkey, Integrated, and Ready-to-Run

- White glove experience from NVIDIA, giant models in weeks not months
- Powered by NVIDIA Base Command and NVIDIA AI Enterprise software

The Trillion Parameter Instrument of AI

Massive memory supercomputing for emerging AI

World’s first system built with NVIDIA NVLink Switch System

- Nearly **500X** more system memory
- **48X** GPU-to-GPU bandwidth
- **7X** CPU-to-GPU bandwidth
- **5X** interconnect power efficiency

Available year-end 2023

256 Grace Hopper Superchips | **144TB** unified fast memory
36 L2 NVLink switches | **900 GB/s** GPU-to-GPU bandwidth | **128 TB/s** bisection bandwidth

How DGX GH200 Enables Giant Models

Solving the previously unsolvable challenges

AI Training

Recommender Systems, Giant Models

SOCIAL MEDIA

E-COMMERCE

Data Analytics

Graph Neural Networks

FRAUD DETECTION

DRUG DISCOVERY

Generative AI

Emerging AI Models

MULTI-MODAL GENERATIVE MODELS

CONTENT GENERATION

CHALLENGE

1. Slower output, models could take months to train instead of weeks, limiting progress

2. Reduced model features due to memory limitations with larger datasets

3. Simply impossible to tackle certain problems due to CPU-GPU memory access bottleneck

SOLUTION

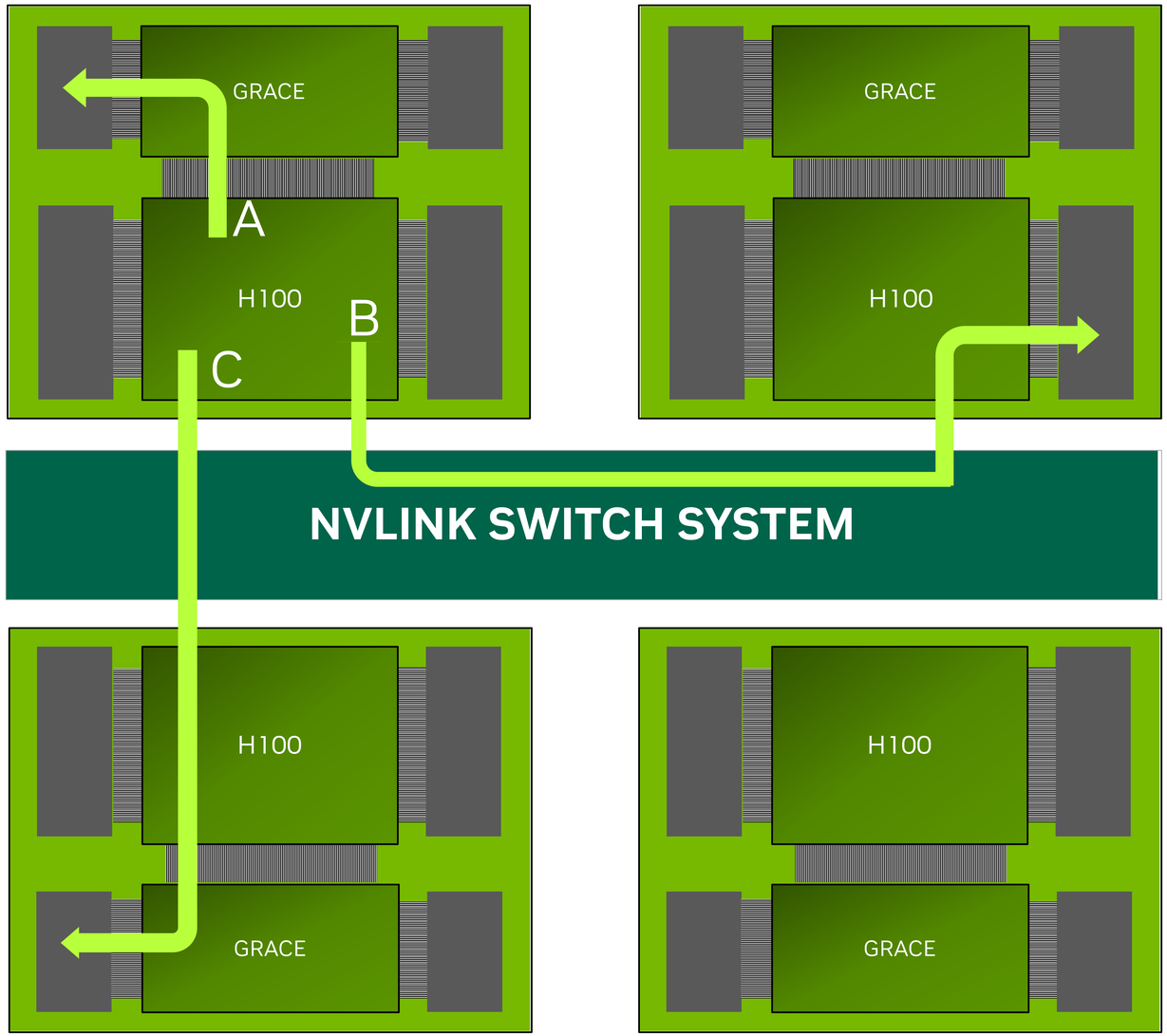
1. 128 TB/s bisection bandwidth offers massive speedup in 1 data-center-size GPU

2. 144 TB of fully-connected GPU memory, allowing larger than ever datasets

3. NVLink-C2C speeds up communication between CPU and GPU with 7X more bandwidth than PCIe Gen 5

Fast NVLink Bandwidth for All 256 GPUs

Up to 7X faster than latest generation PCIe Gen 5 architecture



Memory Access	Bandwidth GBPs/direction
	Full
A. Local CPU memory	450
B. Remote GPU memory	450
C. Remote CPU memory	450

DGX GH200 NVLink Architecture

Up to 256 GPU NVLink full bandwidth fat-tree, enabling 144TB Fast Memory

115.2 TBps bisection bandwidth

GPU Chassis 1

L1 NVLink Switches

450GB/s/dir. NVLink copper cables

H100 Grace — Superchip 1

H100 Grace — Superchip 2

H100 Grace — Superchip 3

H100 Grace — Superchip 8

GPU Chassis 2

L1 NVLink Switches

450GB/s/dir. NVLink copper cables

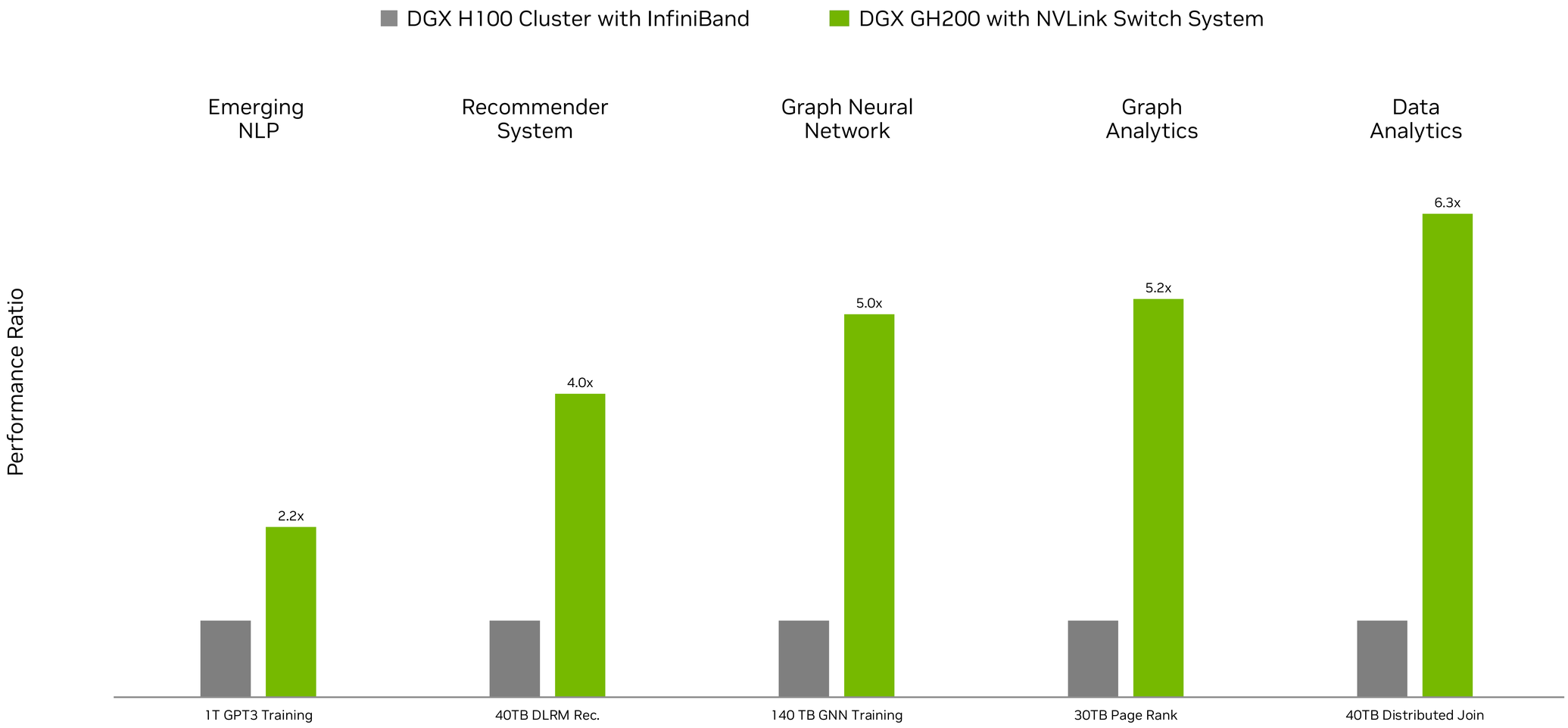
H100 Grace — Superchip 1

H100 Grace — Superchip 2

H100 Grace — Superchip 3

H100 Grace — Superchip 8

DGX GH200 Fastest for Giant Memory Models



Source: NVIDIA internal projections
1T GPT3 Training: 32 GPU; 40TB DLRM Rec: 128 GPU; 140 TB GNN Training: 256 GPU; 30TB Page Rank: 128 GPU; 40TB Distributed Join: 128 GPU

GH200 for Quantum

In memory perspective

量子電腦	CPU 傳統電腦	x86+GPU 高速電腦	New GH200 高速電腦
1 qubit	$2*4*2^{1}$ bytes (16 Bytes)	Not Necessary	Not Necessary
2 qubits	$2*4*2^{2}$ bytes (32 Bytes)	Not Necessary	Not Necessary
20 qubits	$2*4*2^{20}$ bytes (8 MBs)	1*H100-40GB	1*GH200-576GB
32 qubits	$2*4*2^{32}$ bytes (32 GBs)	1*H100-40GB	1*GH200-576GB
36 qubits	$2*4*2^{36}$ bytes (512 GBs)	8*H100-80GB (1*DGX H100)	1*GH200-576GB
45 qubits	$2*4*2^{45}$ bytes (256 TBs)	3200*H100-80GB (400 DGX H100) (~< 1*Selene)	445*GH200-576GB

NVIDIA DGX GH200

A new epoch in AI

256x Grace Hopper Superchips with **144 TB** of total GPU memory

18,432 Arm Neoverse V2 CPU Cores

900GB/s GPU to GPU bandwidth

128 TBps bisection bandwidth

NVLink Switch System

256x NVIDIA Connect-X 7 400Gb/s InfiniBand Network Interface and **256x NVIDIA BlueField-3 DPUs**

Green Computing

Saving Carbon and Money Running Common HPC Apps

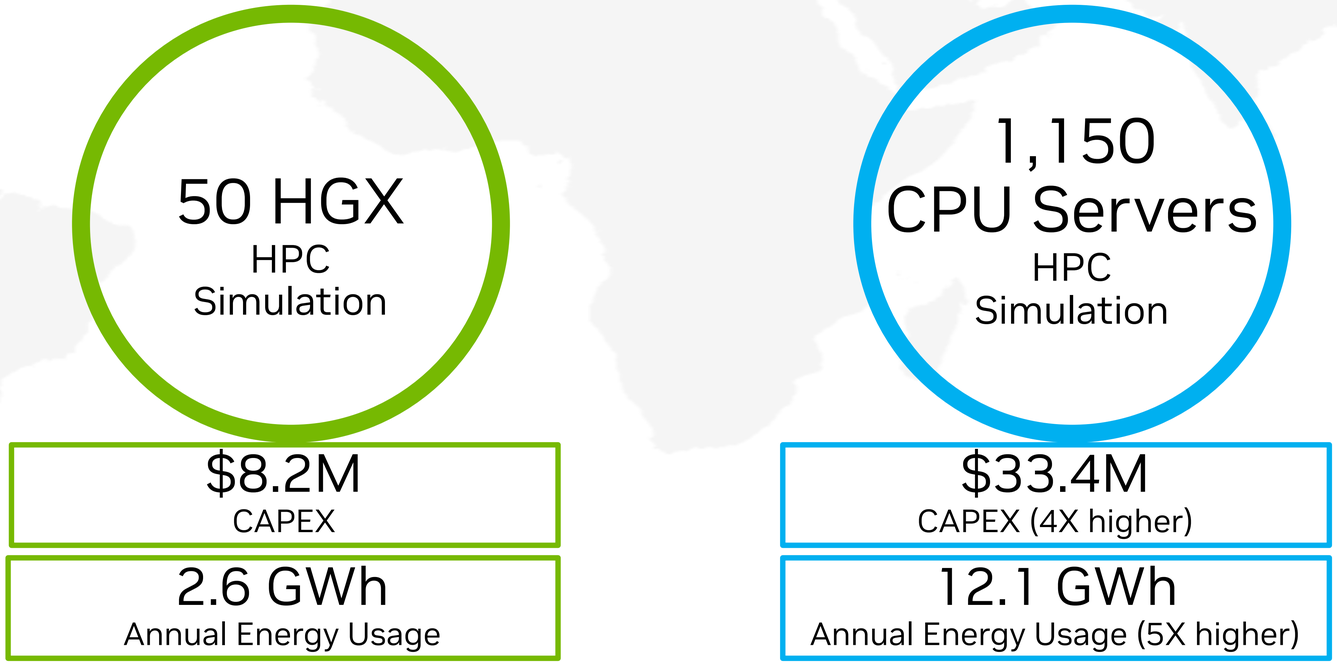
Benefits of Transitioning to NVIDIA GPUs

Case Study: HPC Datacenter running common simulators

HPC Applications: AMBER, Chroma, FUN3D, GROMACS, ICON, MILC, NAMD, Quantum Espresso, and VASP

Equal runtime for each application

NVIDIA H100 vs. Intel Sapphire Rapids



Grace and Hopper: Transforming HPC and AI

Delivers 4.4X More Performance at the Same Power

X86 CPU AND X86 + H100 DATA CENTER

180 dual x86 CPU nodes
168 dual x86 + 4xH100 nodes
44,544 Cores
1 MW total power

GRACE HOPPER DATA CENTER

712 Grace Hopper Superchip nodes

51,300 Cores
1 MW total power

1 MW Data Center in both cases. X86 Data Center consists of 2S Milan CPU and Dual Milan and 4xH100 GPU nodes. Grace Hopper Superchip Nodes. GPU workloads consisting of a mix of HPC applications (including LAMMPS, GROMACS, and ICON)
Projected data center workload performance, subject to change.

Platforms for Discovery

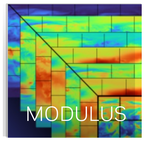
AI APPLICATION FRAMEWORK

Edge Processing

Speech AI

Customer Service

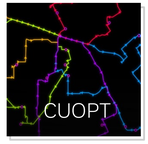
Recommenders



Physics ML

Video

Video Analytics



Logistics

Conversational AI

Robotics

Autonomous Vehicles

Cyber-security

Healthcare

Customer Service

PLATFORMS


NVIDIA HPC


NVIDIA AI


NVIDIA Omniverse

ACCELERATION LIBRARIES

cuNumeric | CV-CUDA | cuQuantum | Parabricks | Sionna | JetPack

RAPIDS | Spark | cuDNN | cuGraph | TensorRT | Triton | DeepStream | Flare

CUDA Quantum | DOCA | Magnum IO | Aerial | cuFFTMp

4M+ Developers

450+ SDKs/Models

CLOUD-TO-EDGE DATACENTER-TO-ROBOTIC SYSTEMS

RTX | DGX | HGX | EGX | OVX | Super POD | IGX

3 CHIPS


GPU


CPU


DPU

NVIDIA


NVIDIA
Thank You