Kick-off Meeting (Nov 16) Virtual	• 02:00PM - 02:05PM: Welcome and event overview (Jay, CK) • 02:05PM – 02:10PM: NCHC opening (王順泰組長@NCHC) • 02:10PM – 02:15PM: Hackathon team opening (Bharat) • 02:15PM - 03:00PM: Round table self-introduction (Team & Mentor). • 3 mins for each team lead • 1 mins for two mentors per team • 03:00PM - 03:05PM: 5 mins break • 03:05PM - 03:15PM: Introduction to computing resources (Kuan-Ting) • 03:15PM - 04:00PM: Introduction to Nsight Analysis Tools (Leo Chen) • 04:00PM - 04:30PM: breakout rooms (Team & Mentor)
Day 1 (November 23) Virtual	• 02:00PM - 03:00PM: Scrum #1 (5 mins presentation per team)
Day 2 (November 30) Virtual	• 02:00PM - 03:00PM: Scrum #2 (5 mins presentation per team)
Day 3 (Dec 07) In-Person	• 10:00 AM - 10:30 PM: Welcome and event description • 10:30 AM - 12:00 PM: Final presentation (12 mins presentation +3 minutes QA per team) • 12:00 PM - 01:30 PM: Lunch time • 01:30 PM - 03:00 PM: Final presentation (12 mins presentation +3 minutes QA per team) • 03:00 PM - 04:00 PM: Wrap-up session

NARLabs 國家實驗研究院
國家高速網路與計算中心
National Center for High-performance Computing

OPEN HACKATHONS
OpenACC

Total presentation time is 5 minutes = 4 minutes presentation + 1 minutes QA or any inputs.

Team Name

NVIDIA GH200 Grace Hopper Superchip

Built for the New Era of AI Supercomputing

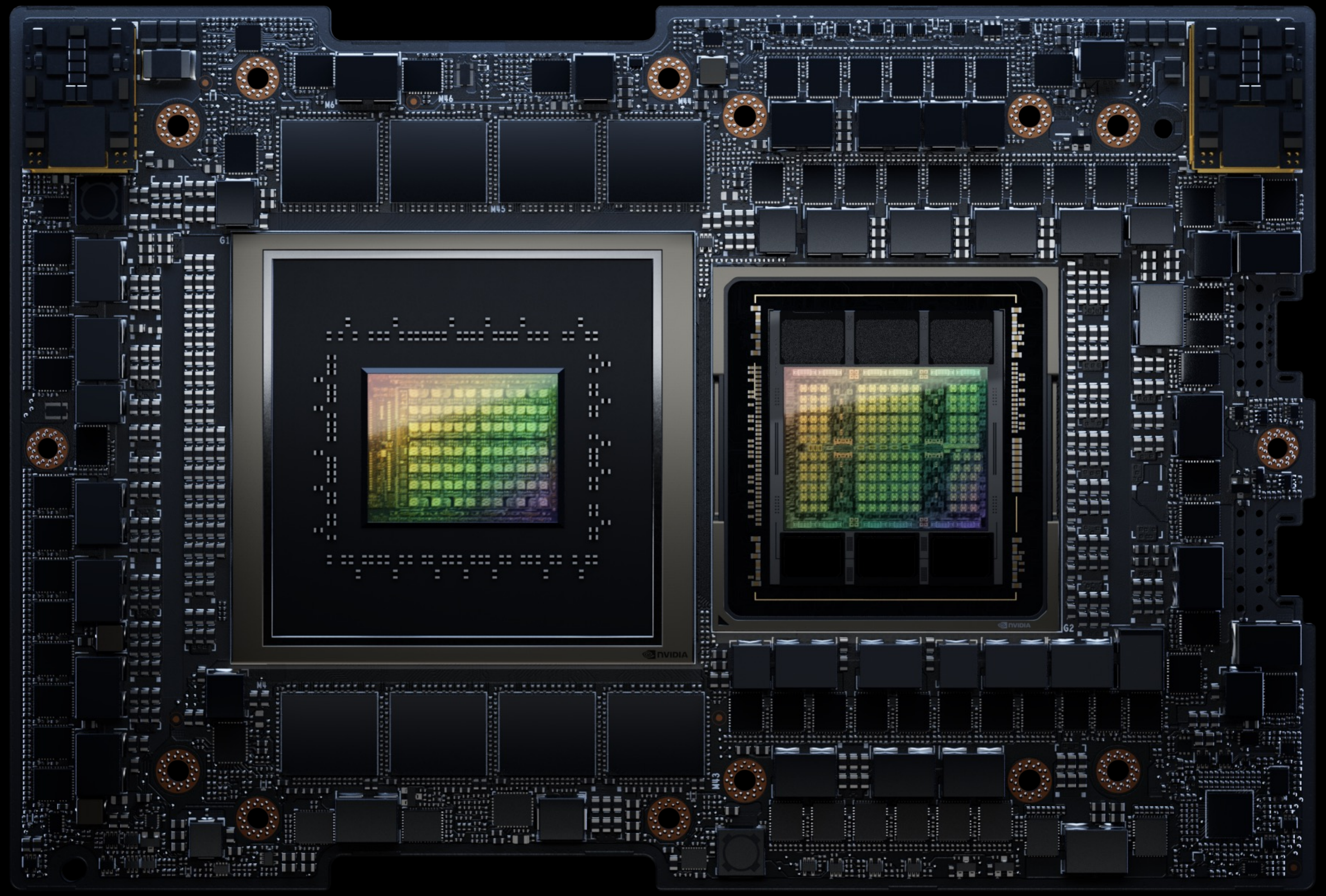
CPU to GPU Bandwidth
900GB/s
NVLink-C2C

GPU Memory Bandwidth
4.9TB/s
HBM3e

Energy Efficiency
52X
MILC Efficiency vs 2S x86 CPUs

QFT Quantum Simulation
90X
Performance vs 2S x86 CPUs

Llama 2 70B Inference
100X
Performance vs 2S x86 CPUs



624GB High-Speed Memory | 4 PF AI Perf | 72 Arm Cores

Preliminary measured performance, subject to change
Energy Efficiency: GH200 144GB vs 2S Xeon 8480+ CPU for MILC running dataset: Apex Medium
QFT Quantum Simulation: QFT 2S Xeon 8480+ vs GH200 144GB
Llama2 70B: ISL 2K, OSL 512 | Throughput | GH200 144GB BS 128 | 2S Xeon 8480+

Now in Full Production – NVIDIA GH200 Grace Hopper Superchip

Built for the New Era of AI Supercomputing

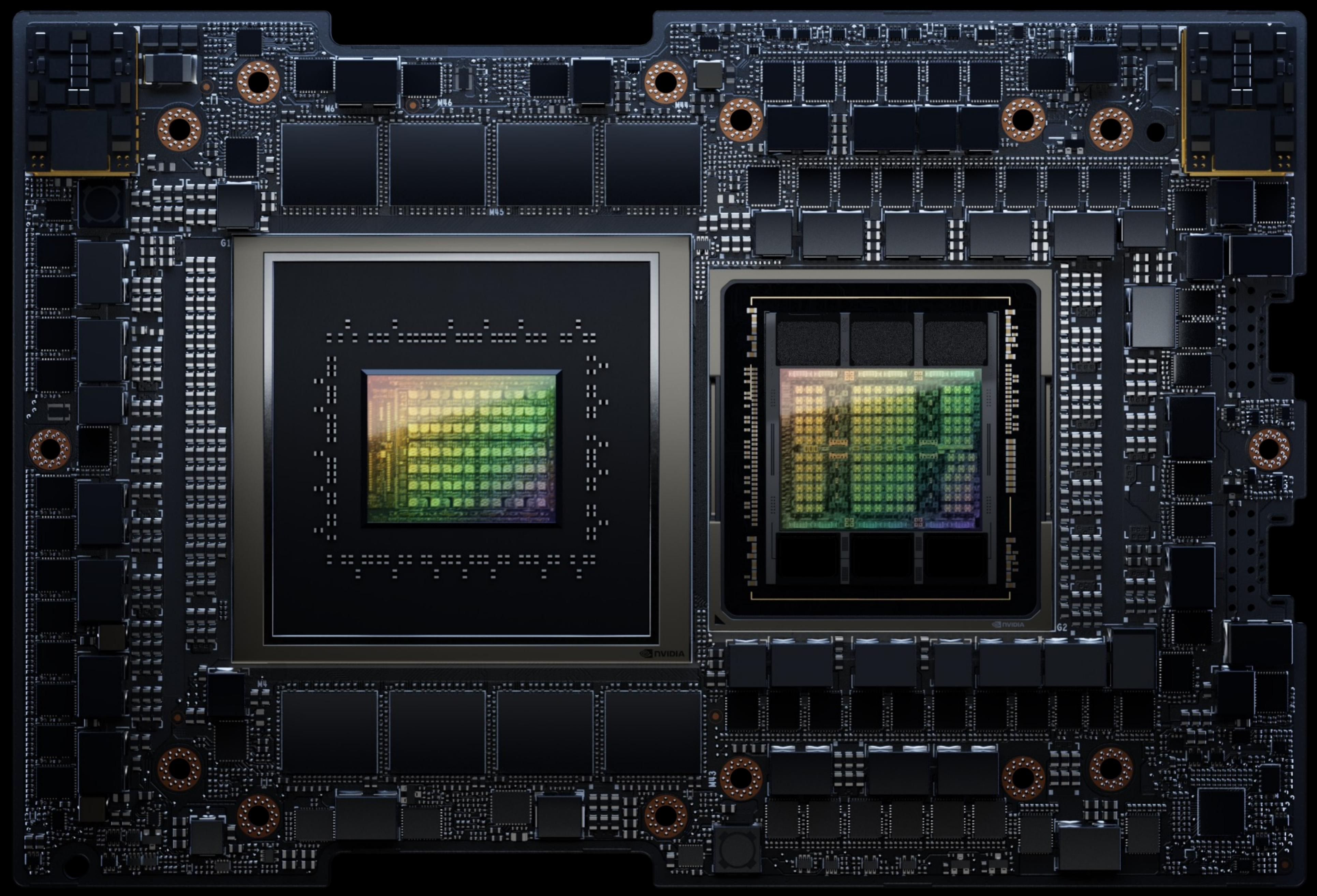
Most versatile compute
Best performance across CPU, GPU or memory intensive applications

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

Now Available in Launchpad
Available Starting Early December; Sign-up Today

https://www.nvidia.com/en-us/launchpad/grace-hopper/

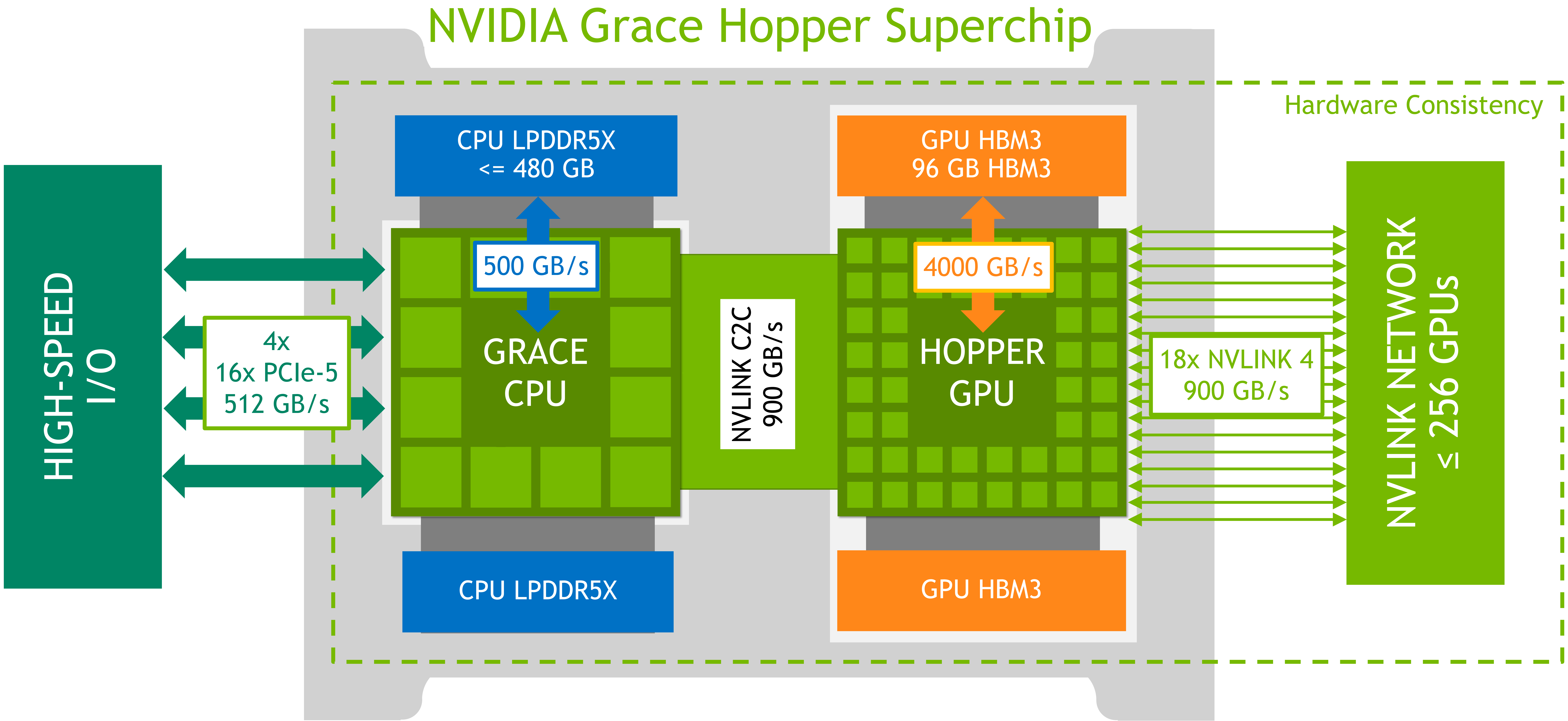


Shipping Globally with Early Access from CSPs

ASRock Rack
ASUS
DELL Technologies
CoreWeave
EVIDEN an atos business
Foxconn Industrial Internet
GIGABYTE
Hewlett Packard Enterprise
Inventec
Lambda
Lenovo
NVIDIA
ORACLE CLOUD Infrastructure
PEGATRON
QCT
SUPERMICRO
VULTR
wistron
wiwynn
zt Systems

Grace Hopper Superchip

GPU can access CPU memory at CPU memory speeds



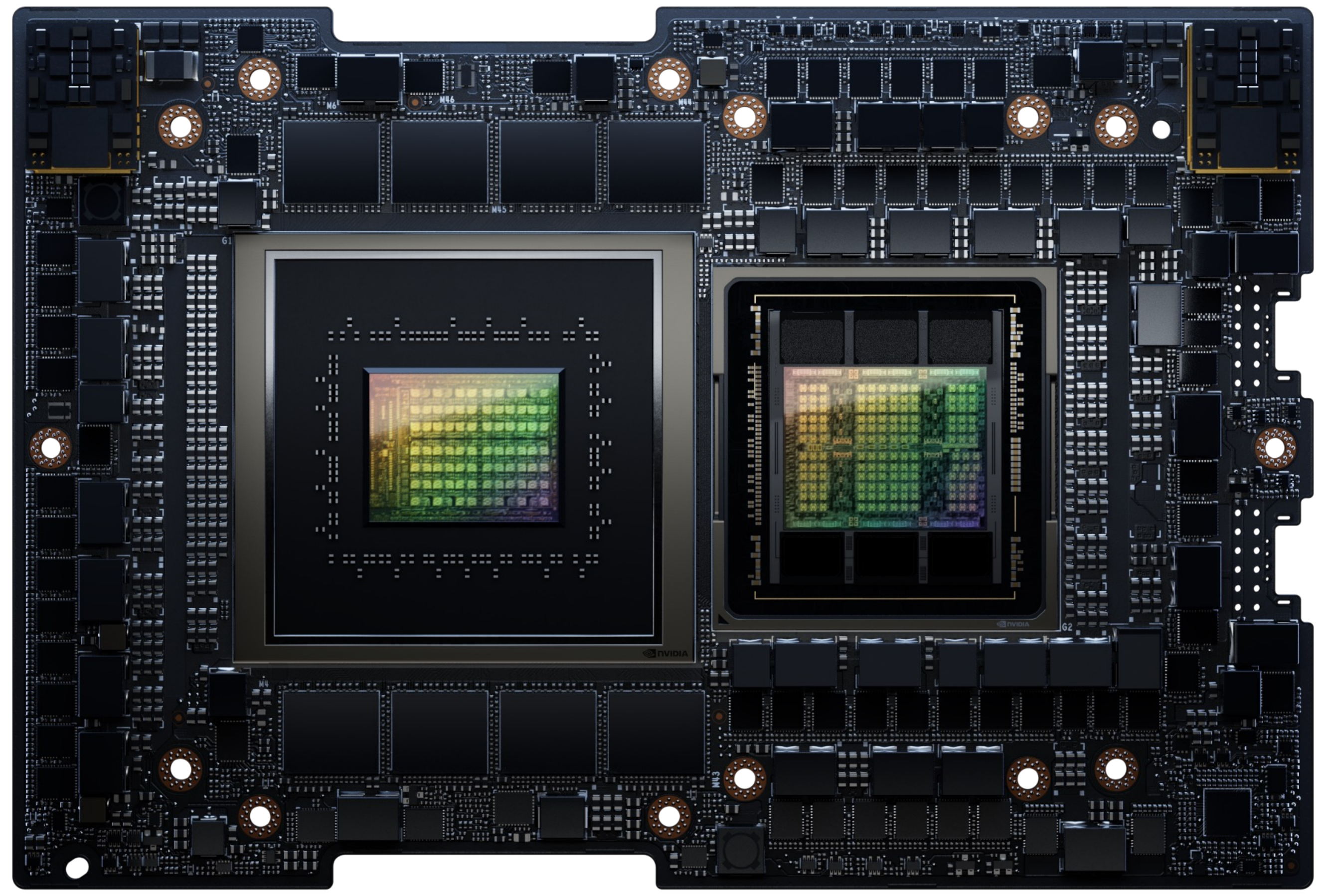
https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper

GH200 Grace Hopper HPC Platform

Unified Memory and Cache Coherence for Next Gen HPC Performance

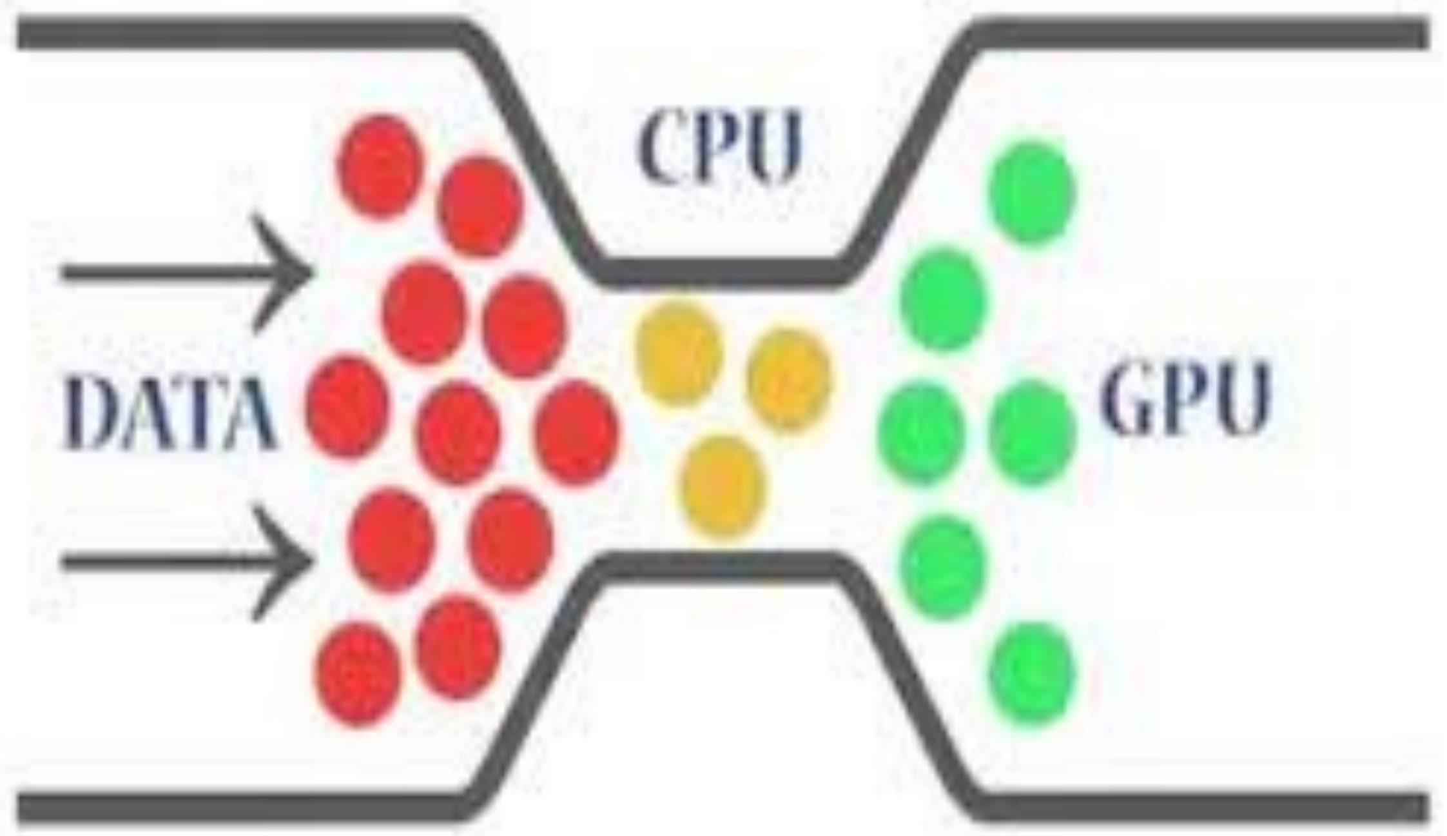
Partially GPU Accelerated Apps

Big performance gains with no code changes



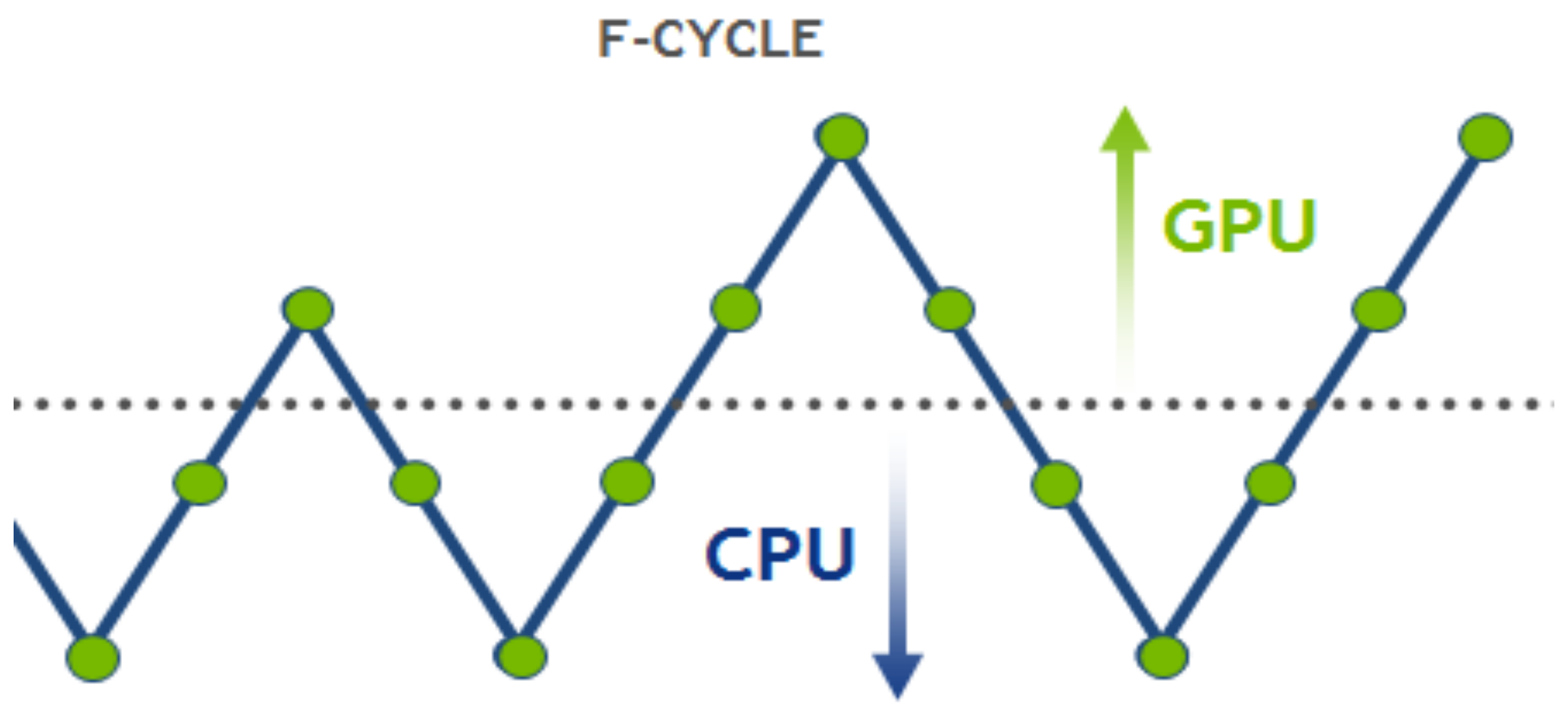
No More PCIe Bottleneck

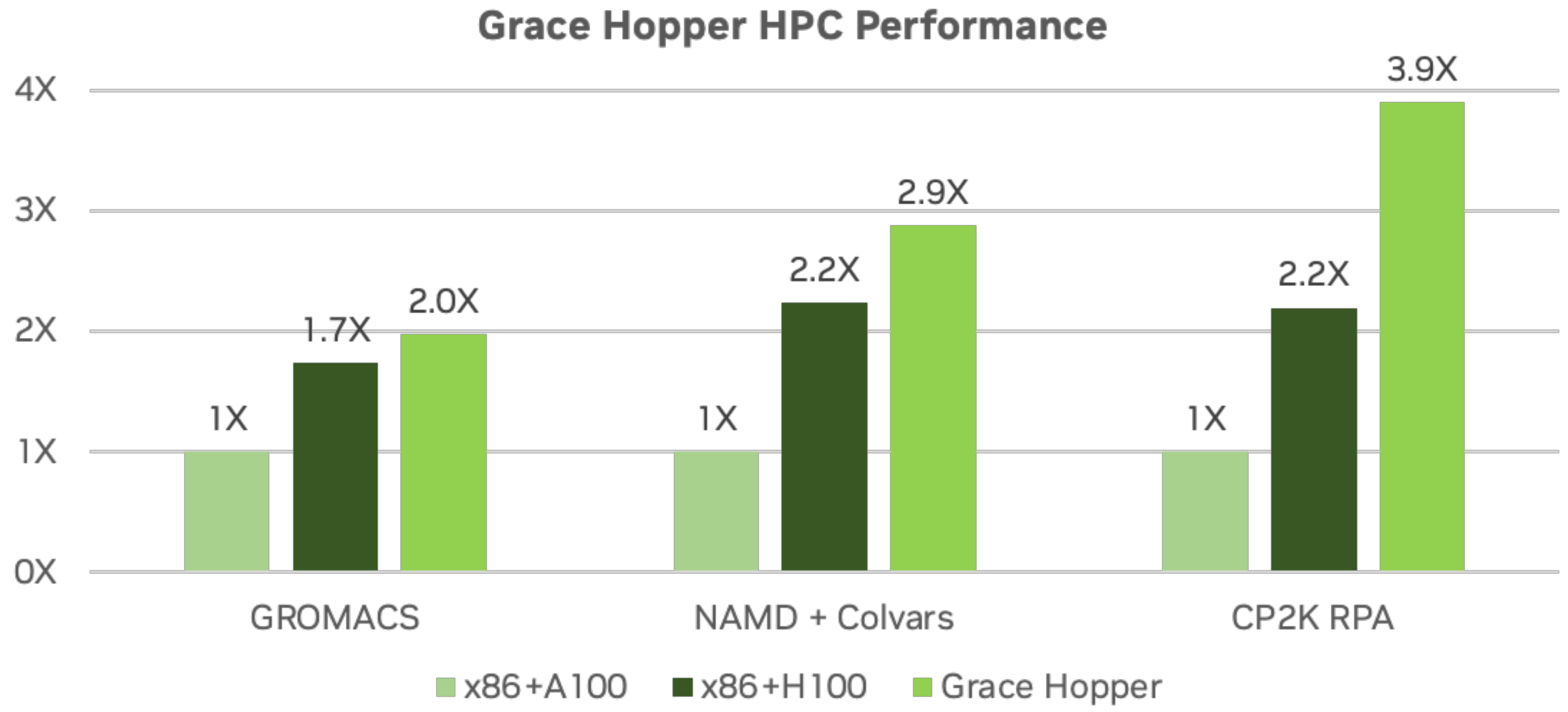
NVLink-C2C is 7X PCIe BW



CPU & GPU Cache Coherence

Incremental code changes yield big gains





Fast Access Memory

624GB

Memory Bandwidth

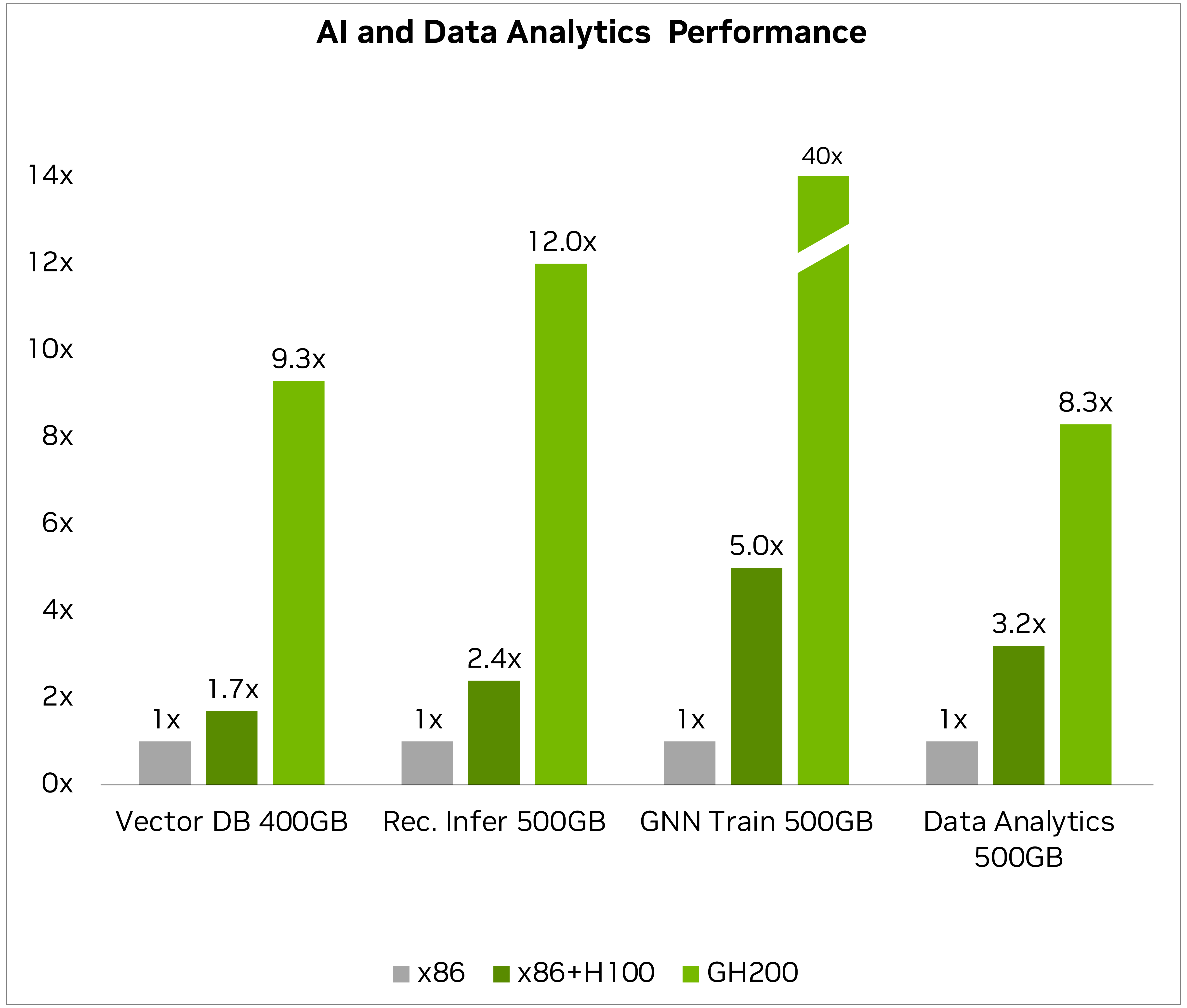
5TB/s

HPC: Preliminary results comparing DGX A100, DGX H100 and GH200 with HBM3 systems.
OpenFOAM based on MotorBike, NAMD with Colvars, CP2K with RPA,

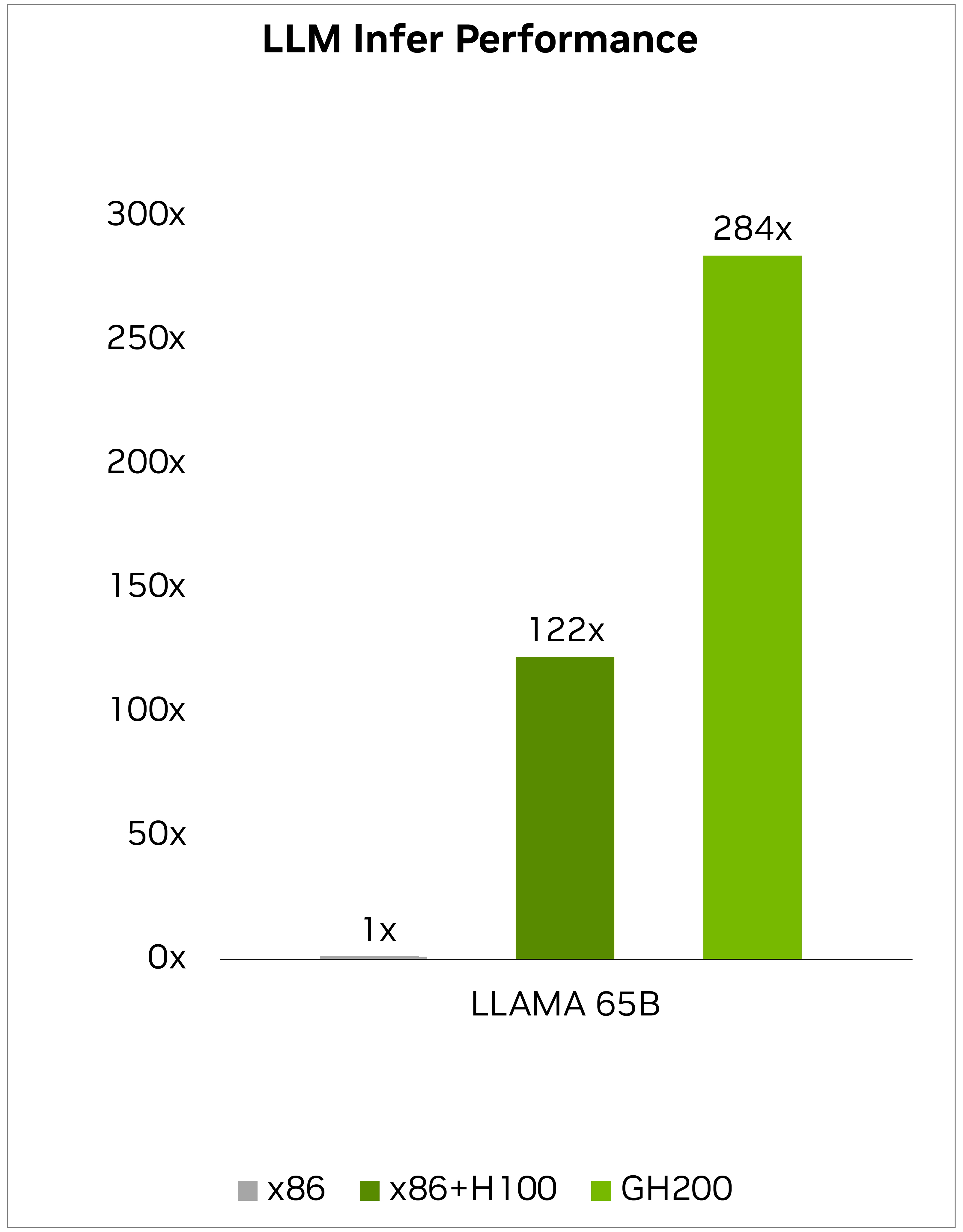
GH200 Grace Hopper AI Inference Platform

Versatile Scale Out with Unmatched Performance

Memory Intensive



GPU Intensive



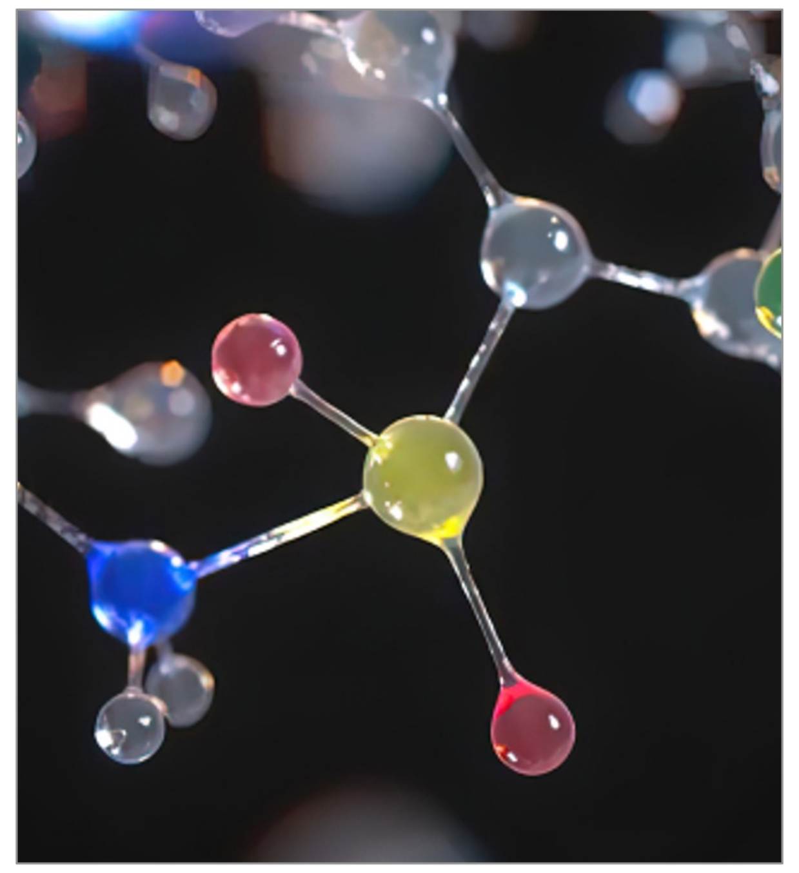
Use Cases

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, GH200 with HBM3
Vector DB search (IVF-Flat-MM used), DLRM Rec infer, DA (TPC-H query4) are projection based on scaling kernel measurements
Graph Neural Network, based on full GraphSAGE model measurements

Simplifying GPU Programming for HPC with NVIDIA Grace Hopper Superchip

https://developer.nvidia.com/blog/simplifying-gpu-programming-for-hpc-with-the-nvidia-grace-hopper-superchip/