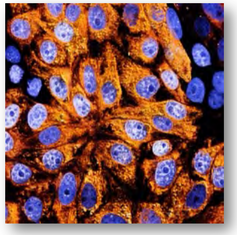
Grace and Grace Hopper Update

Dion Harris, Director of Accelerated HPC, AI, Quantum Solutions

September 7, 2023

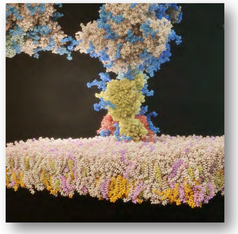
Workloads of the Modern Supercomputer

EDGE

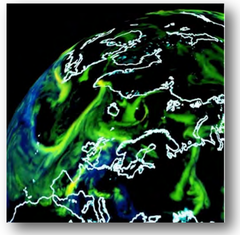


SIM + AI

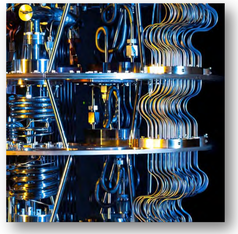
SIMULATION

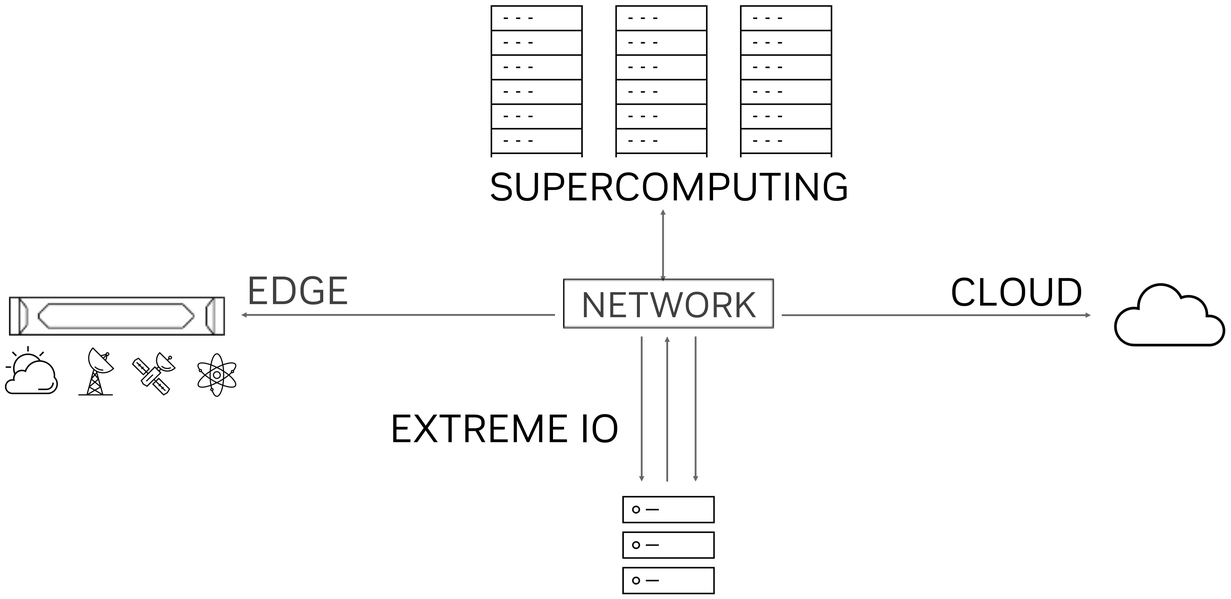


DIGITAL TWIN



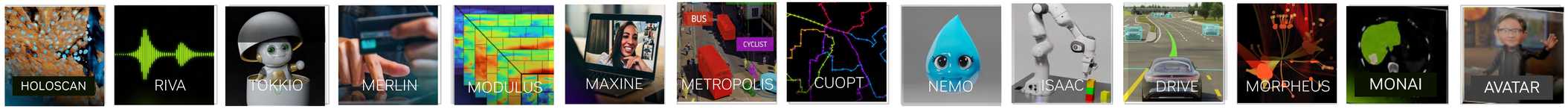
QUANTUM COMPUTING





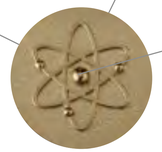
NVIDIA Platform

AI APPLICATION FRAMEWORK



Edge Processing | Speech AI | Customer Service | RecSys | 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

CLOUD-TO-EDGE DATACENTER-TO-ROBOTIC SYSTEMS

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

3 Processors

GPU | CPU | DPU

NVIDIA Grace and Grace Hopper SuperChips

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

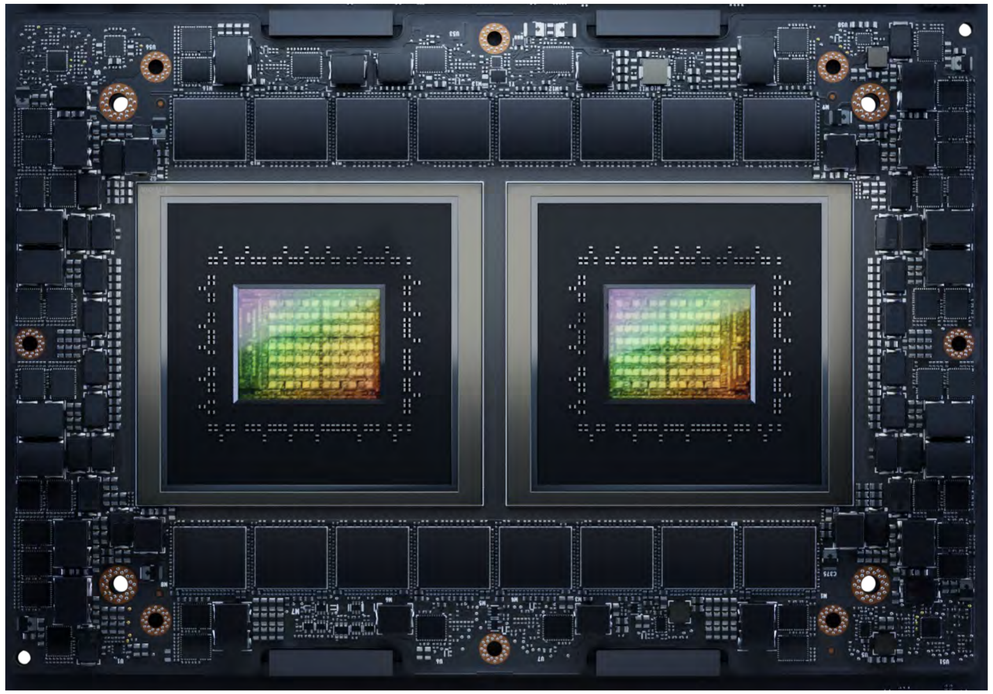
CSCS (ALPS)
Grace Hopper
20 EFLOPS AI Perf

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

NVIDIA Grace for Cloud, AI and HPC Infrastructure

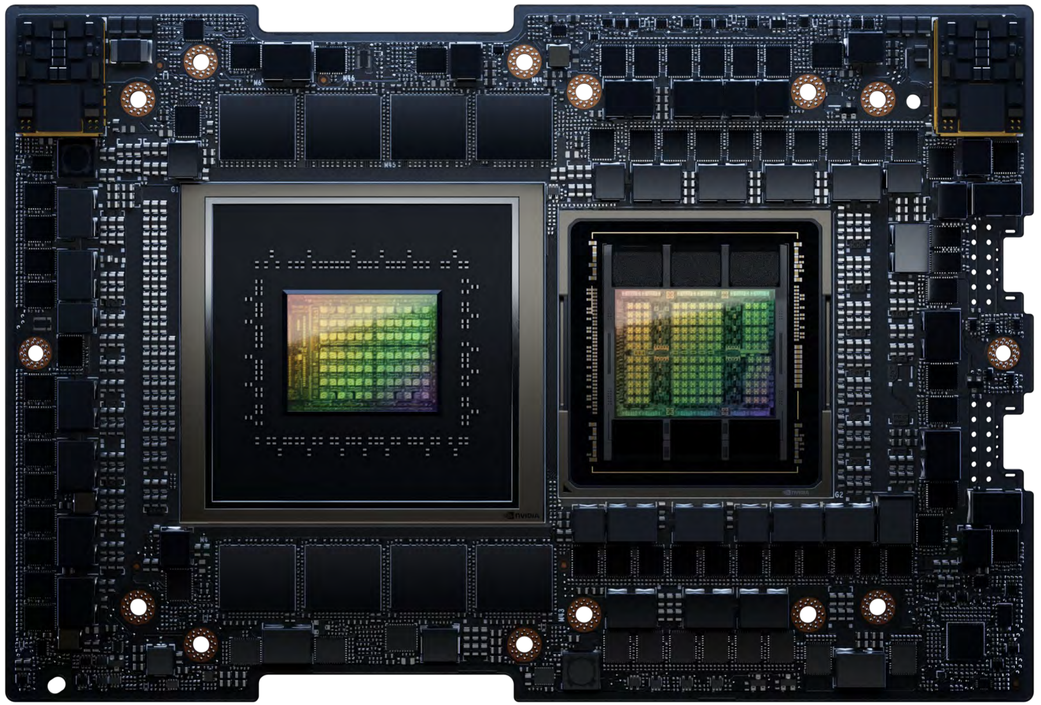
Grace CPU Superchip

CPU Computing



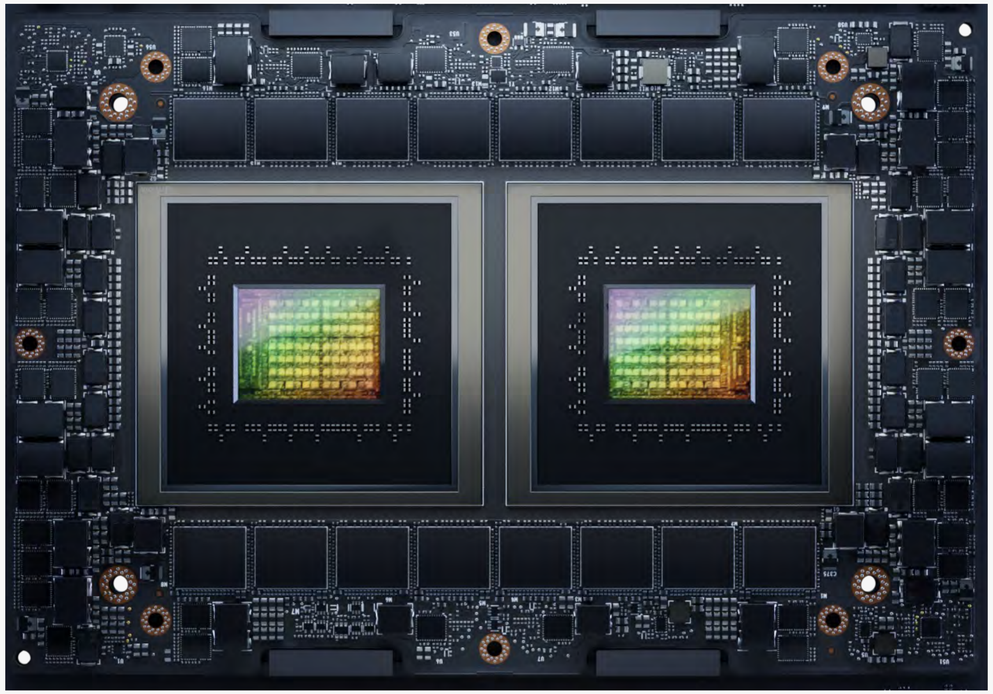
GH200 Grace Hopper Superchip

Large Scale AI & HPC



NVIDIA Grace CPU Superchip

Energy efficient CPU computing



High Performance Power Efficient Cores

144 flagship Arm Neoverse V2 Cores with
SVE2 4x128b SIMD per core

Fast On-Chip Fabric

3.2 TB/s of bisection bandwidth connects
CPU cores, NVLink-C2C, memory, and system IO

High-Bandwidth Low-Power Memory

Up to 480 GB of data center enhanced LPDDR5X Memory that
delivers up to 500 GB/s of memory bandwidth

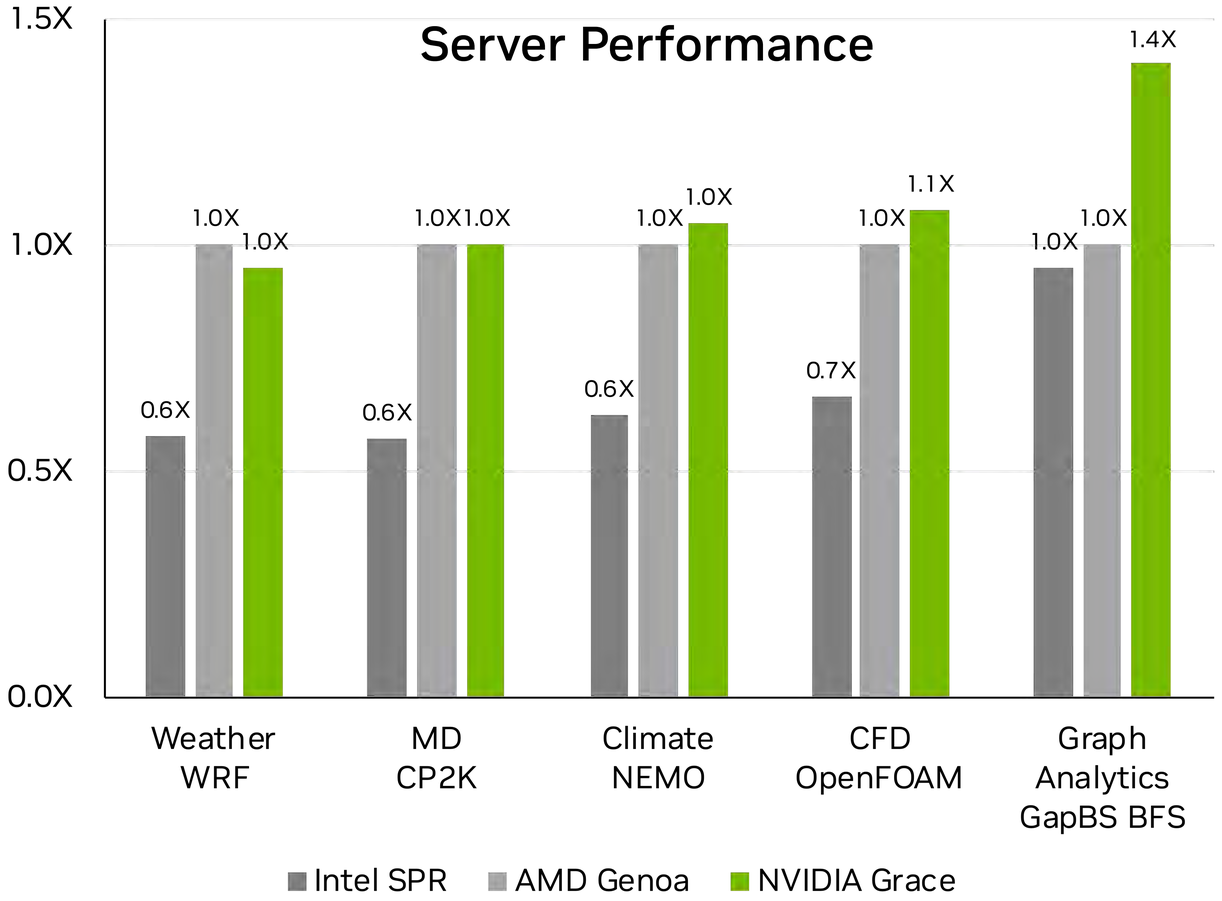
Coherent Chip-to-Chip Connections

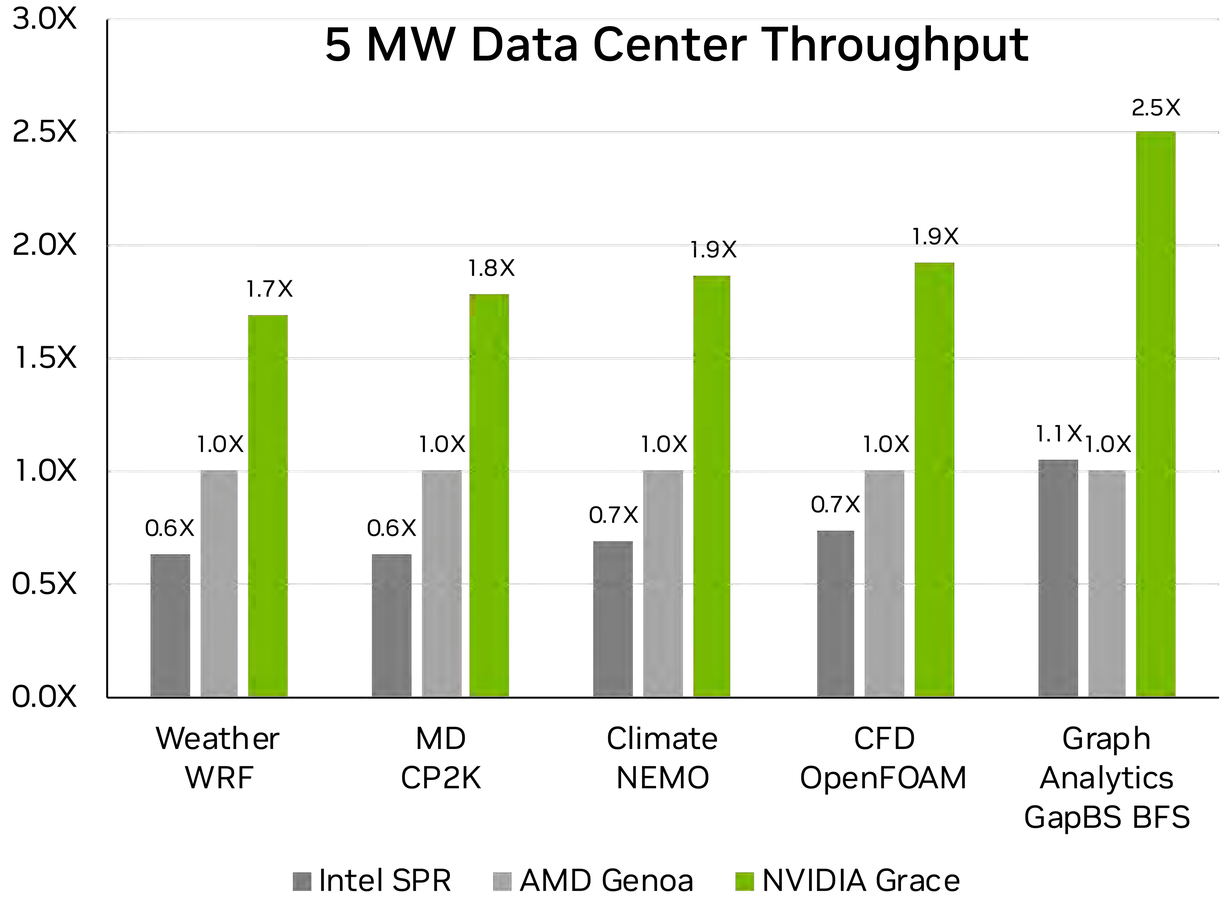
NVLink-C2C with 900 GB/s bandwidth for coherent
connection to CPU or GPU

Industry Leading Performance Per Watt

Up to 2X perf / W over today's leading servers

NVIDIA Grace CPU Delivers 2X Throughput at the Same Power





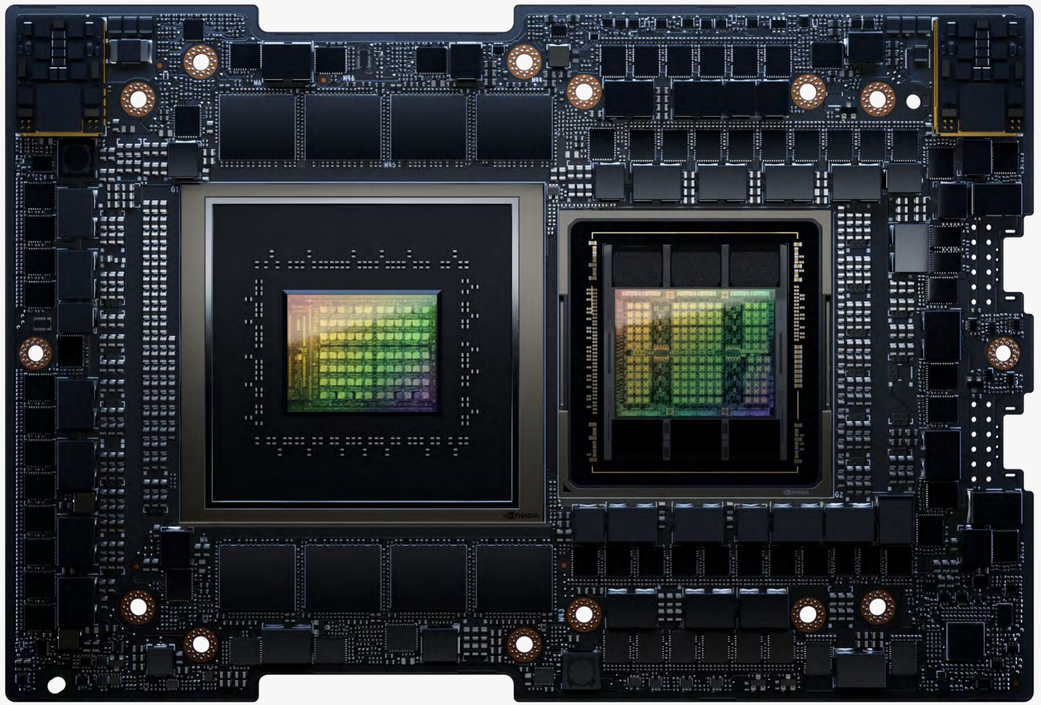
5 MW Data Center level projection based on server measurement of NVIDIA Grace Superchip vs x86 2-socket data center systems (AMD Epyc 9654 and Intel Xeon 8480+). Weather: WRF CONUS12, 24 hr simulation 4.4.2 MD: CP2K RPA 2023.1 Climate: NEMO Gyre_Pisces v4.2.0 CFD: OpenFOAM Motorbike | Large v2212 Graph Analytics: The Gap Benchmarks Suite BFS NVIDIA Grace Superchip performance based on engineering measurements. Results subject to change.

NVIDIA

NVIDIA Grace Hopper Superchip

"Super" → more than a "chip"

NVIDIA CPU + NVIDIA GPU w/o compromises

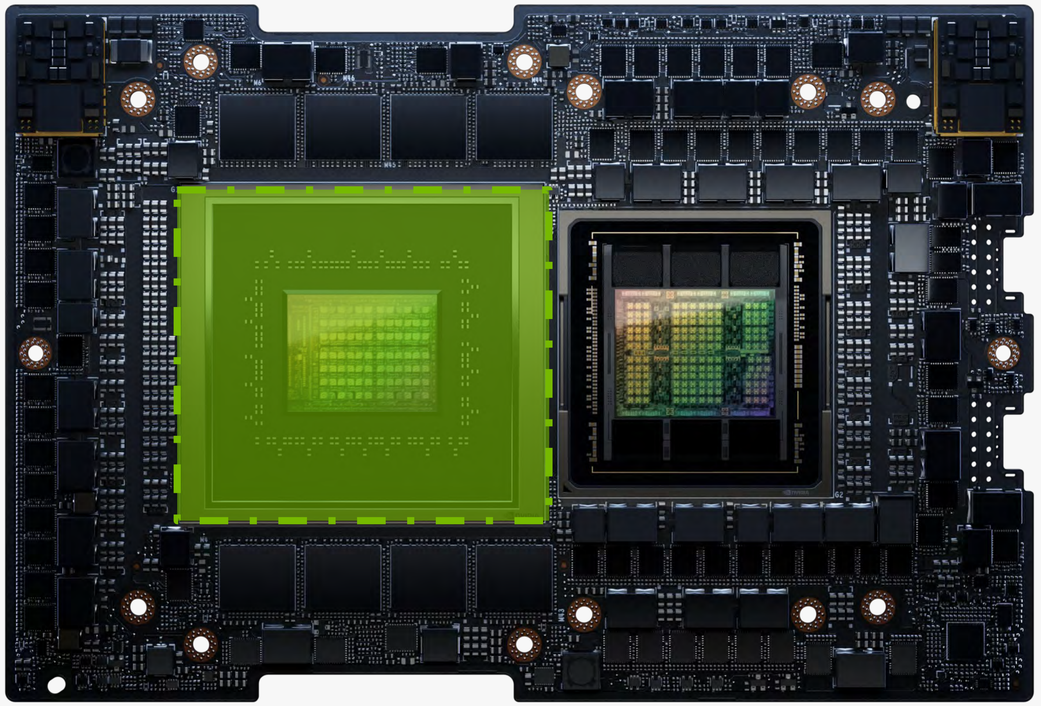


NVIDIA Grace Hopper Superchip

"Super" → more than a "chip"

NVIDIA CPU + NVIDIA GPU w/o compromises

- **NVIDIA Grace CPU + LPDDR5 Memory**
  - 72 Arm-v9 Neoverse V2 CPU cores with SVE2.
    - →Efficiency: 62pJ/DFMA (x86: ~99); 1.6x more efficient
    - →Performance: 3.6 FP64 TFLOP/s
  - Memory:
    - → High capacity: ≤ **480 GB** LPDDR5X (5pJ/bit vs 36 DDR)
    - → High bandwidth: ≤ **500 GB/s**
    - → Low latency: less than competitors at peak bandwidth

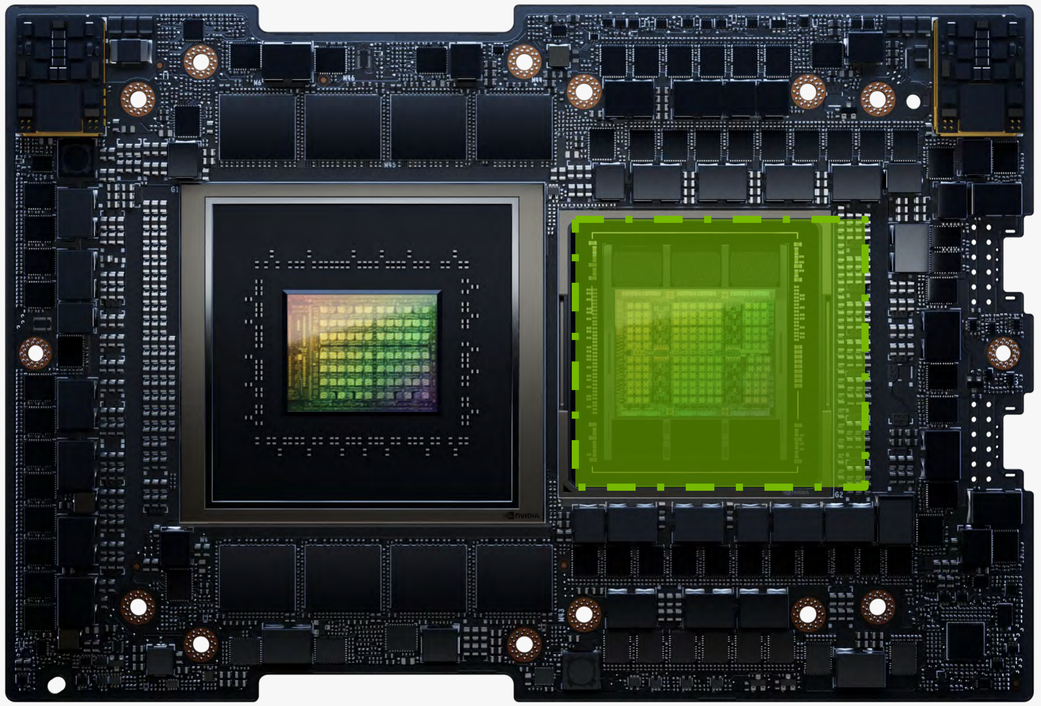


NVIDIA Grace Hopper Superchip

"Super" → more than a "chip"

NVIDIA CPU + NVIDIA GPU w/o compromises

- **NVIDIA Grace CPU + LPDDR5 Memory**
  - 72 Arm-v9 Neoverse V2 CPU cores with SVE2.
    - →Efficiency: 62pJ/DFMA (x86: ~99); 1.6x more efficient
    - →Performance: 3.6 FP64 TFLOP/s
  - Memory:
    - → High capacity: ≤ **480 GB** LPDDR5X (5pJ/bit vs 36 DDR)
    - → High bandwidth: ≤ **500 GB/s**
    - → Low latency: less than competitors at peak bandwidth
- **NVIDIA Hopper GPU**
  - →High performance: 60 FP64 TC TFLOP/s
  - Memory:
    - → High capacity: 96 GB HBM3
    - → Extreme bandwidth ≤ 4000 GB/s
  - → Threads are threads (not SIMD lanes)

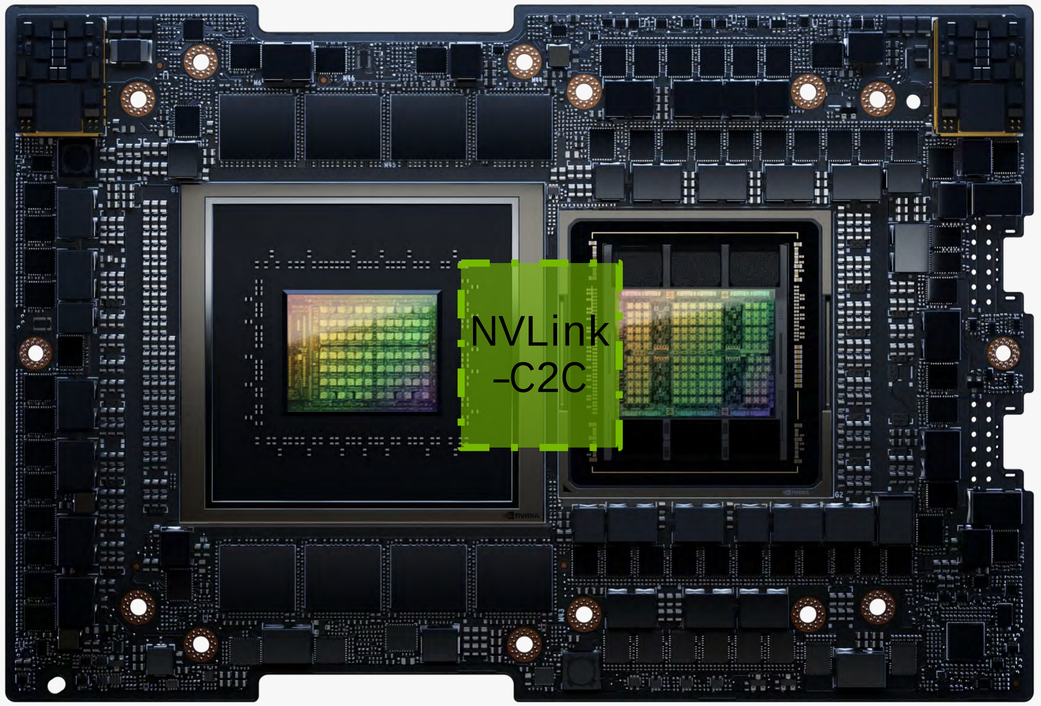


NVIDIA Grace Hopper Superchip

Soul is the new **NVLink-C2C**
CPU ←→ GPU interconnect

- **Memory coherency**: ease of use
  - →*All* threads – GPU and CPU – access system memory: C++ `new`, `malloc`, `mmap`'ed files, atomics, …
  - → Fast automatic page migrations HBM3←→LPDDR5X.
  - → Threads cache peer memory → Less migrations.
- **High-bandwidth**: 900 GB/s (same as peer NVLink 4)
  - → GPU reads or writes local/peer LPDDR5X at ~peak BW
- **Low-latency**: GPU→HBM latency
  - →GPU reads or writes LPDDR5X at ~HBM3 latency

For all threads in the system
memory is memory
expected behavior + latency + bandwidth.



NVIDIA Grace Hopper Superchip

Made 💚 for any programming model

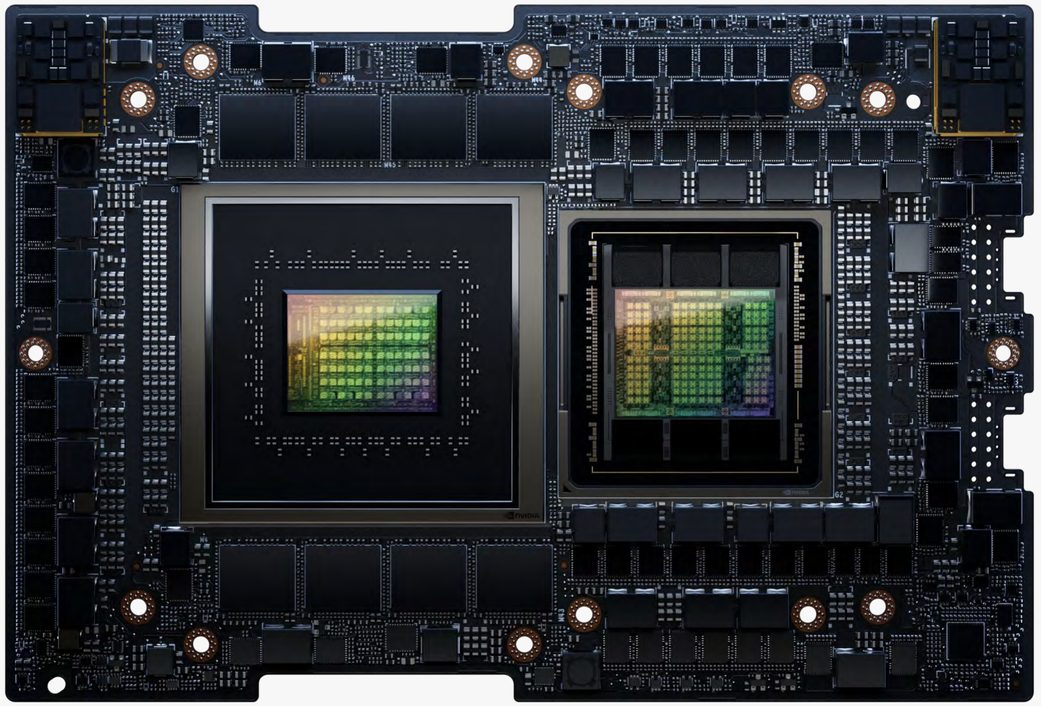
Portable ISO C++, ISO Fortran, Python

- **Simplifies parallelization:** less SW changes

  → **ISO C++, ISO Fortran, Python**: Threads are “threads” (!SIMD), memory consistency, automatic memory management, …

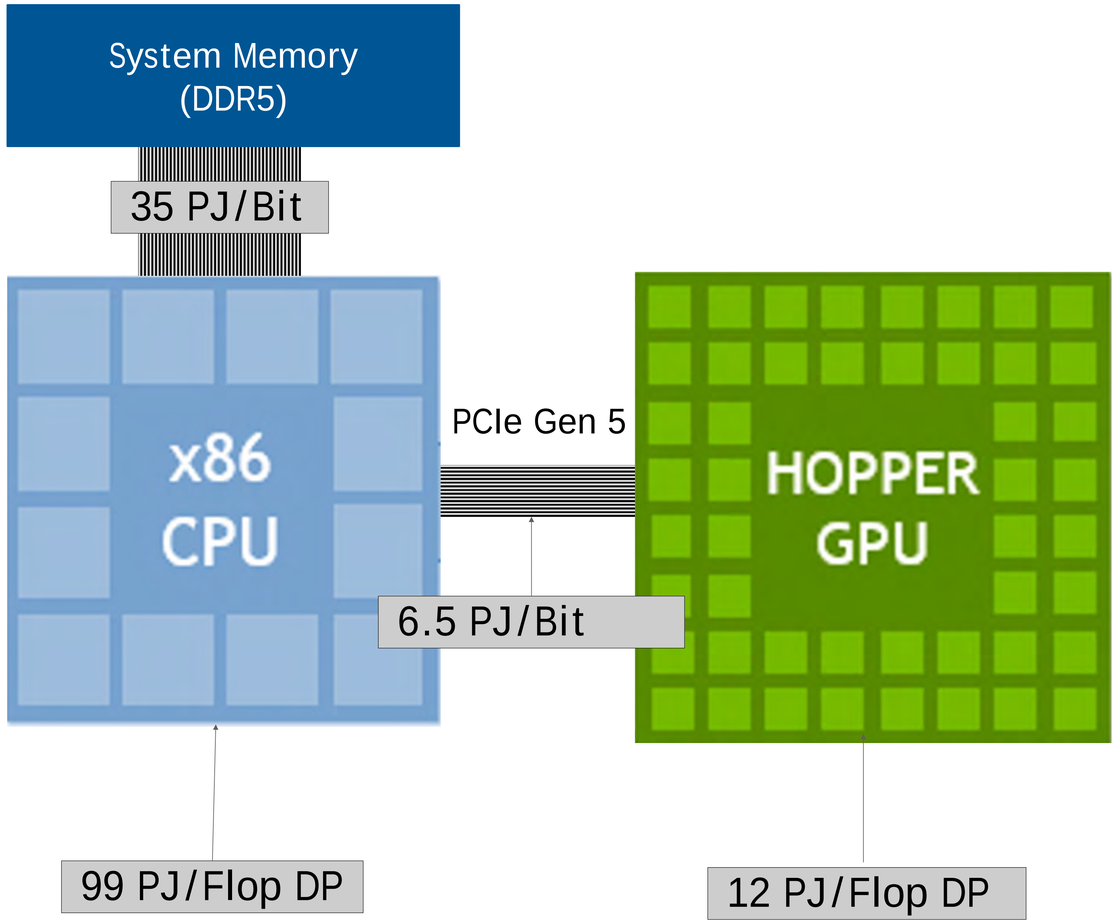
  → **Applications**: complex code stays on CPU, infrequently used memory stays on DDR, large GPU memory capacity (600 GB).

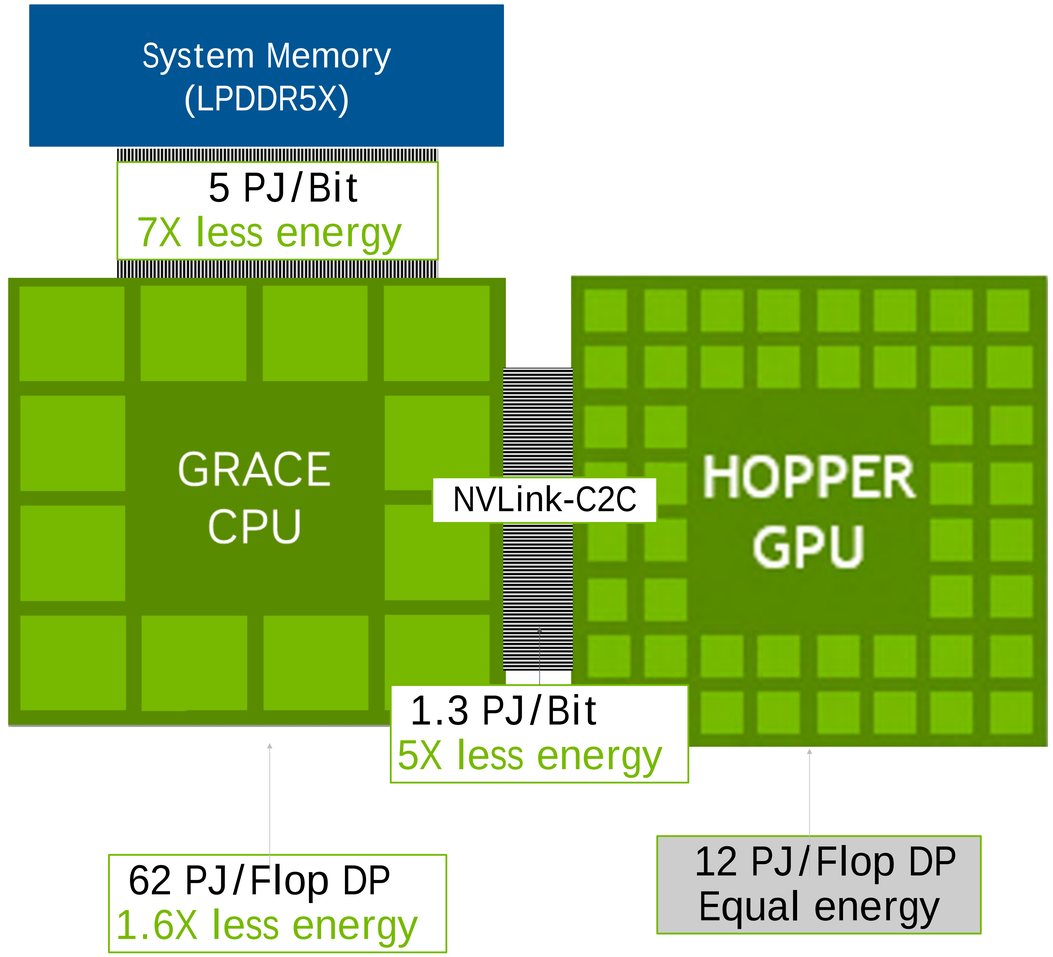
- **Easiest system to:**
  → teach & learn heterogeneous programming
  → parallelize applications
  → use the right HW for each algorithm



Energy Efficient Design

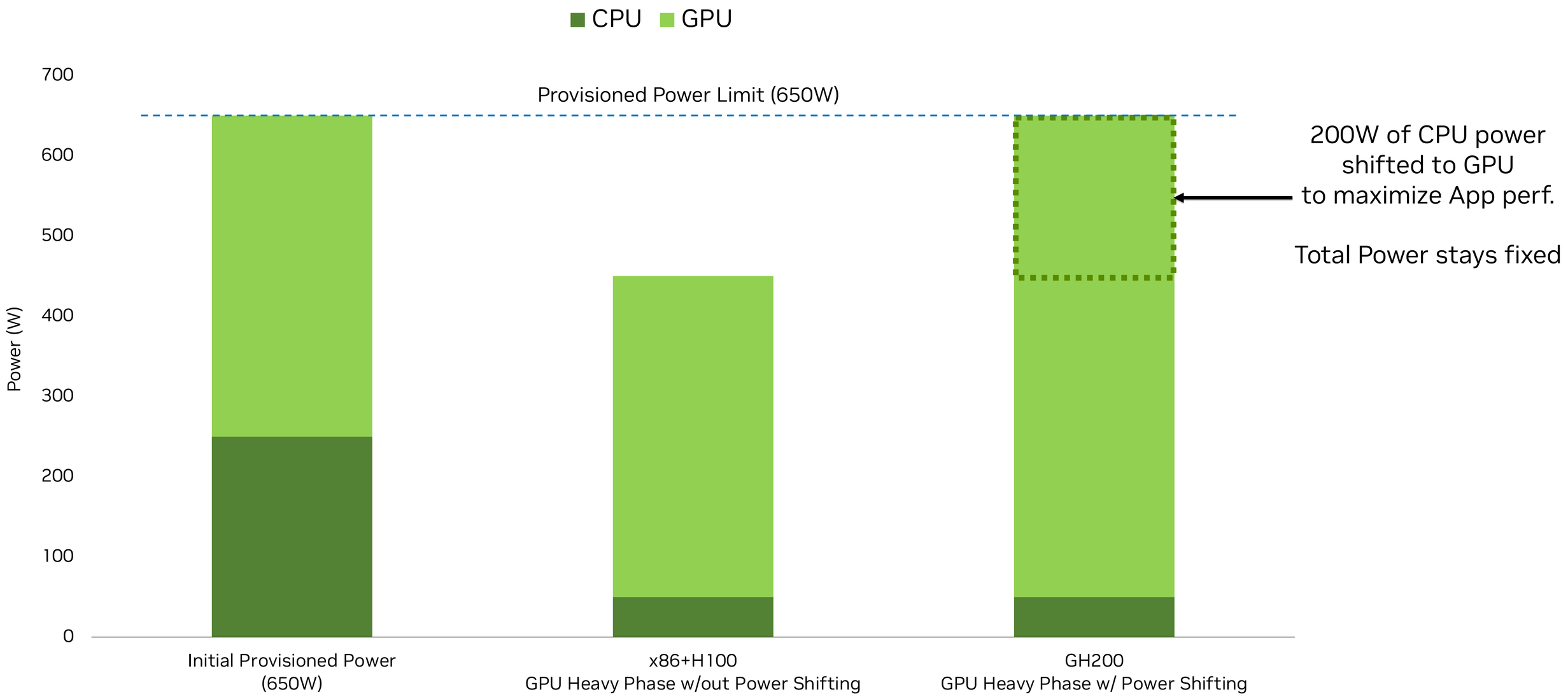
More Efficient Computation and Data Movement



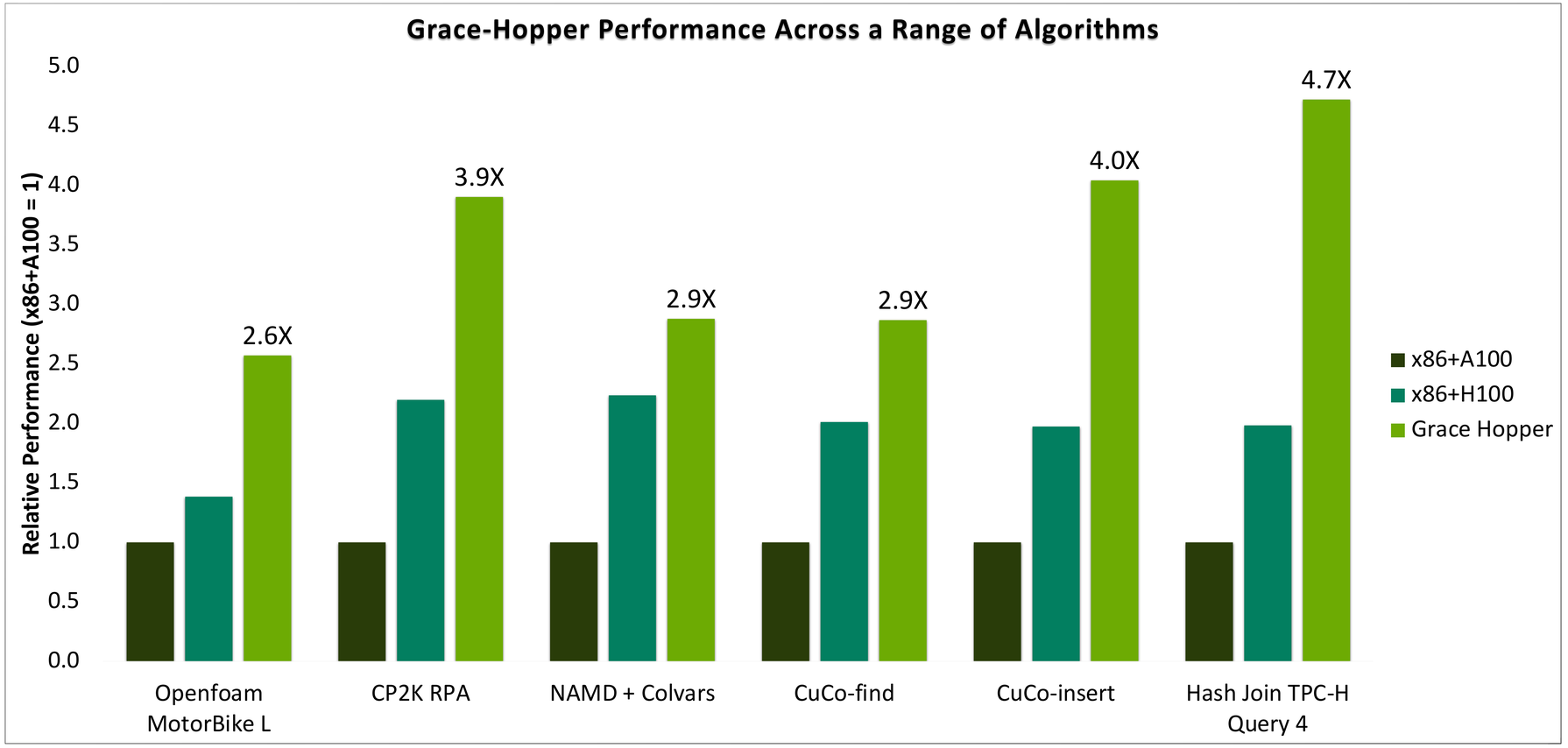


Optimizing Performance Through Power Shifting

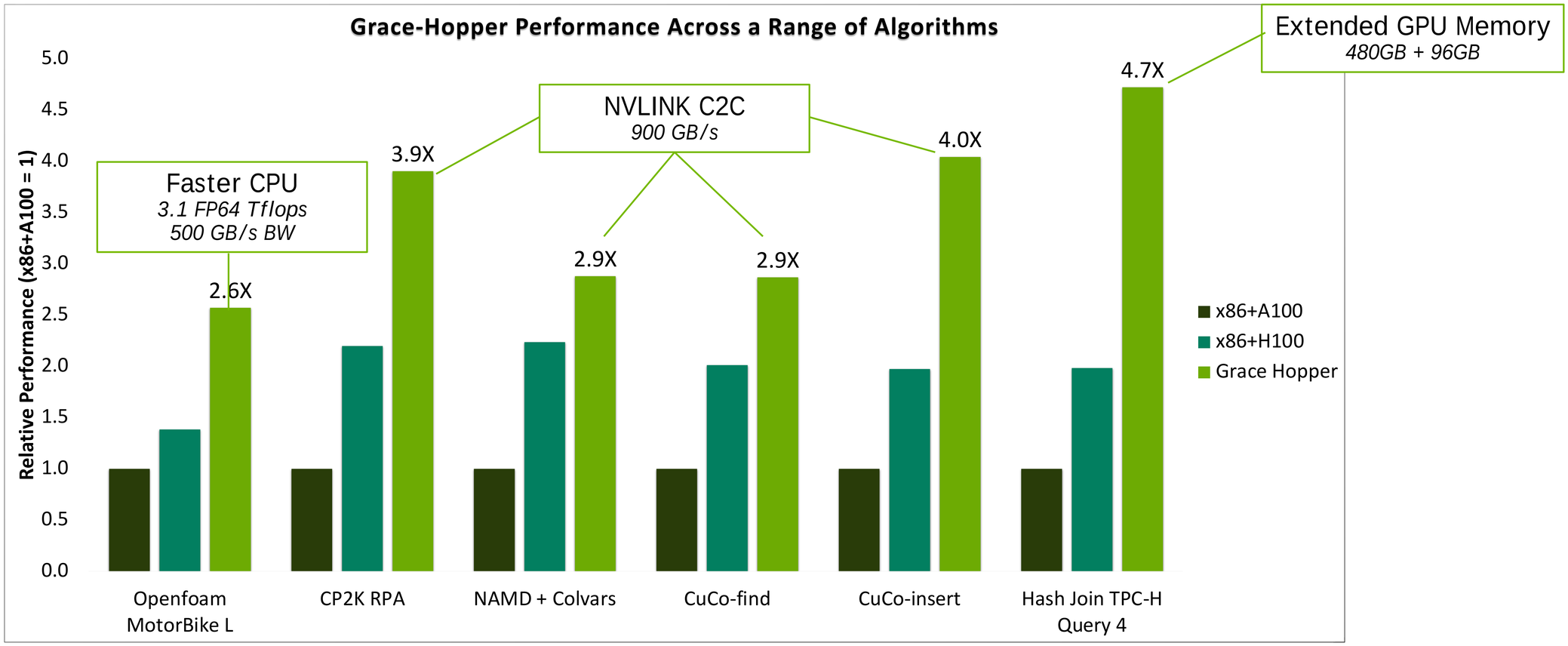
Getting the most out of provisioned power



Grace-Hopper Superchip Workload Performance



Grace-Hopper Superchip Workload Performance



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

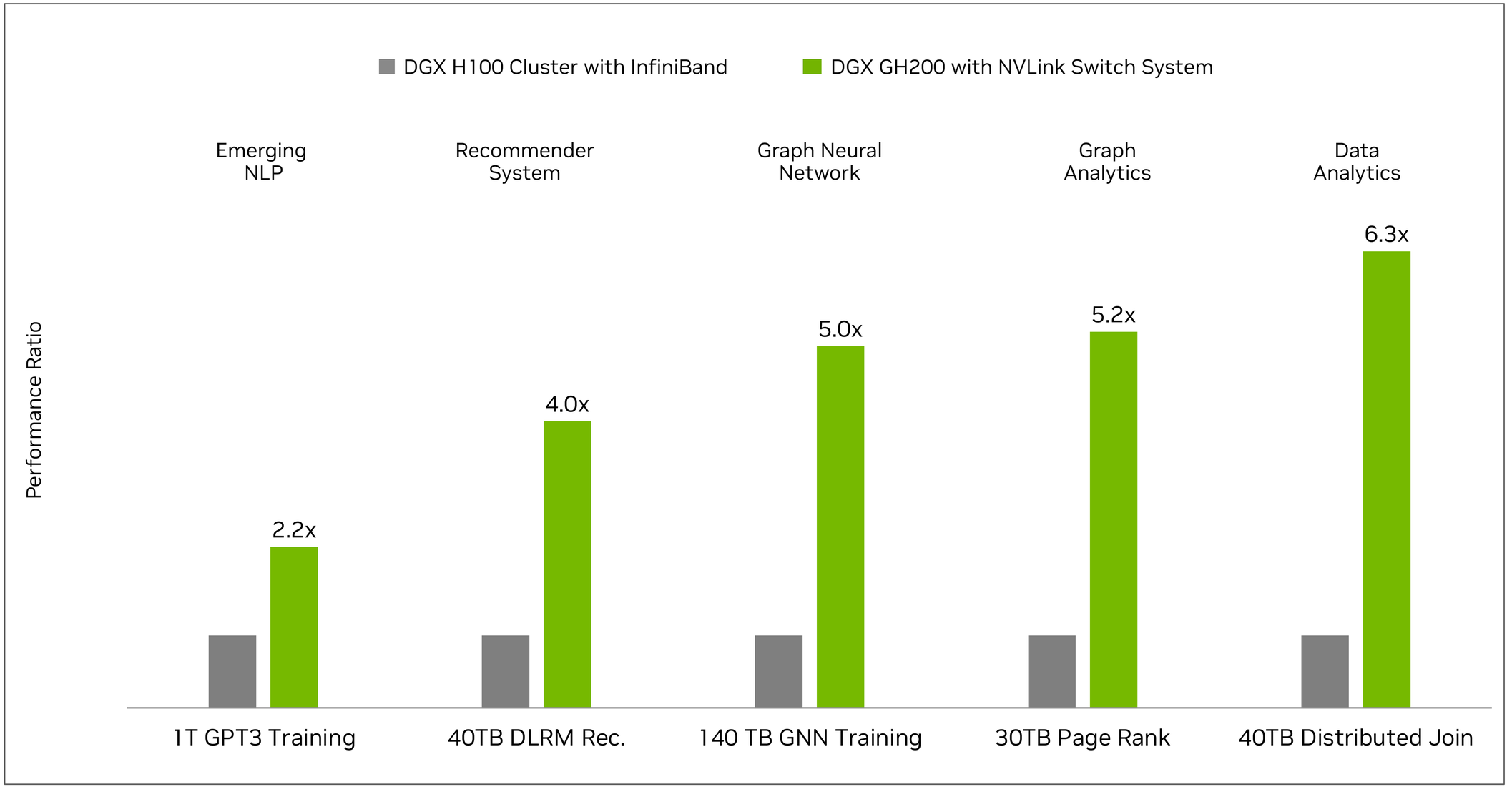
1 MW Data Center in both cases. X86 Data Center consists of 2S Milan CPU and Dual Milan and 4xH100 GPU nodes. Grace Hopper Superchip Nodes. GPU workloads consisting of a mix of HPC applications (including LAMMPS, GROMACS, and ICON)
Projected data center workload performance, subject to change.

NVIDIA DGX GH200

World’s first system built with NVIDIA NVLink Switch System

256 GH200 Grace Hopper Superchips | **1 EFLOPS** AI Performance | **144TB** unified fast memory
36 L2 NVLink switches | **900GB/s** GPU-to-GPU bandwidth | **128TB/s** bisection bandwidth

DGX GH200 Fastest for Giant Memory Models



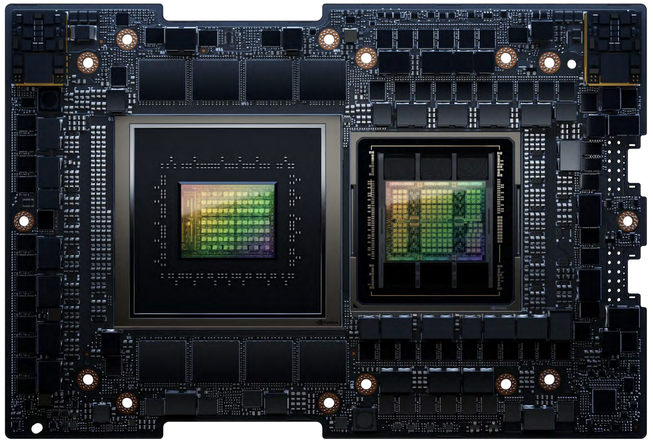
Source: NVIDIA internal projections
1T GPT3 Training: 32 GPU; 40TB DLRM Rec: 128 GPU; 140 TB GNN Training: 256 GPU; 30TB Page Rank: 128 GPU; 40TB Distributed Join: 128 GPU

Next-Generation NVIDIA GH200 Grace Hopper Superchip

Processor for the era of accelerated computing and generative AI

Next-Gen NVIDIA GH200 Superchip

- World's first HBM3e GPU
- 144 GB HBM3e memory with 5 TB/sec bandwidth
- 1.7X capacity | 1.5x bandwidth vs H100



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

New NVLink dual-GH200 system

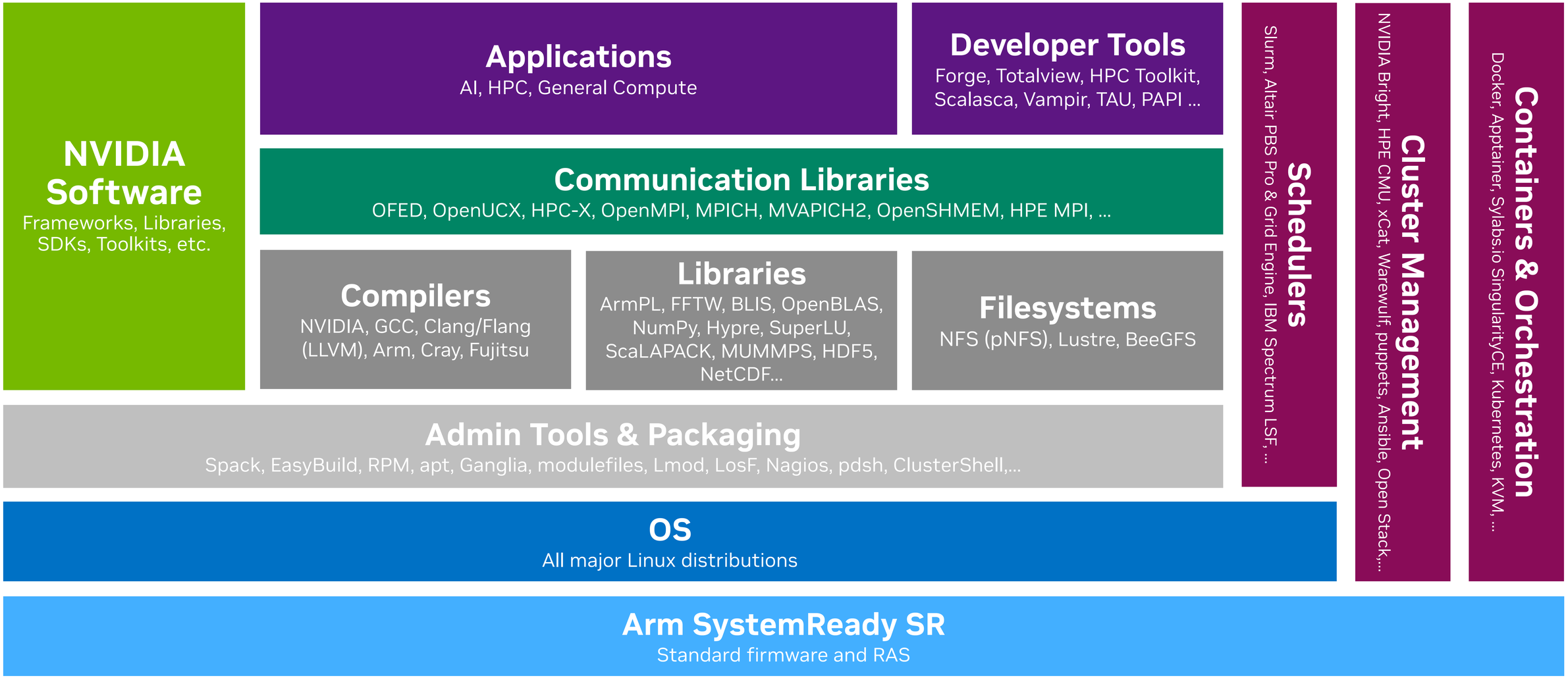
- Combined 1.2 TB fast memory
- 3.5x capacity | 3x bandwidth
- Simple to deploy MGX-compatible design

144 Core Grace CPU | 8 PFLOPS Hopper GPU | 282 GB HBM3e | 10 TB/s

Available Q2 2024

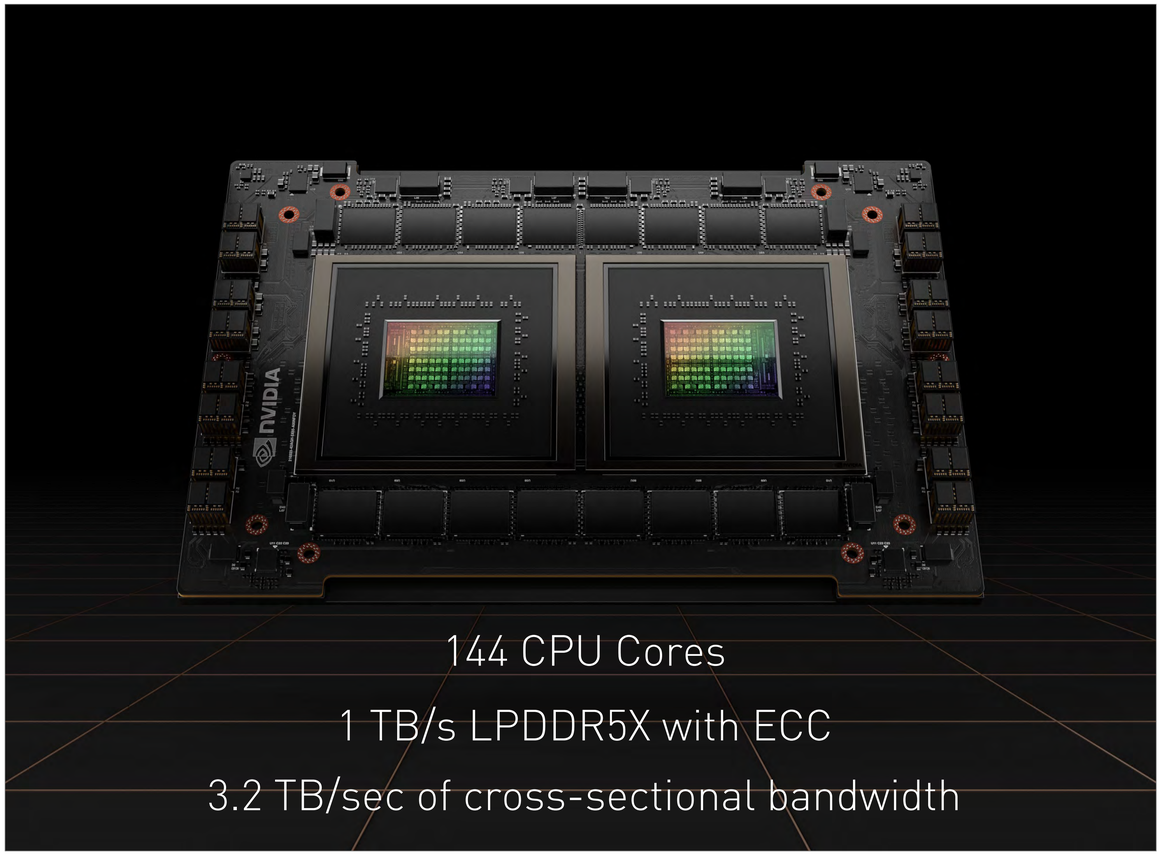
NVIDIA Grace HPC/AI Software Ecosystem

Full support for the broad Arm software ecosystem, both open source and commercial

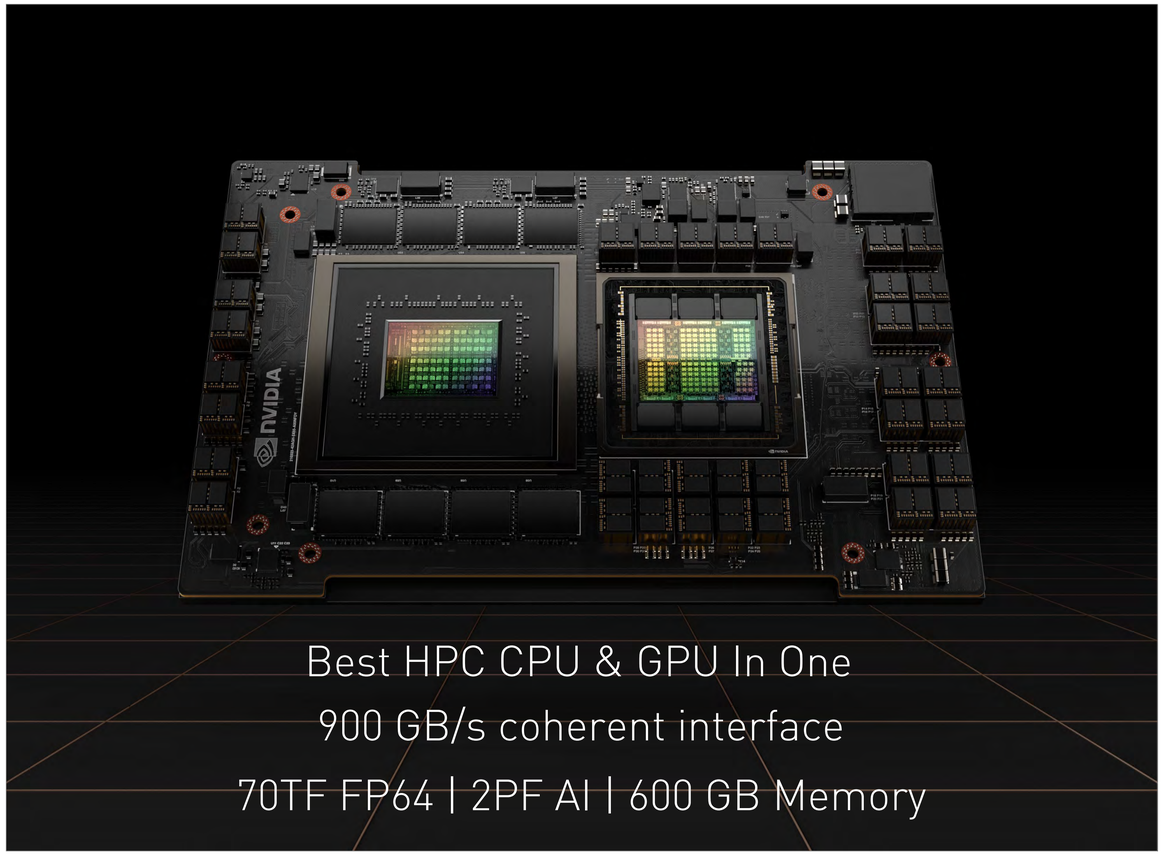


NVIDIA Grace and Grace Hopper

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

Questions?