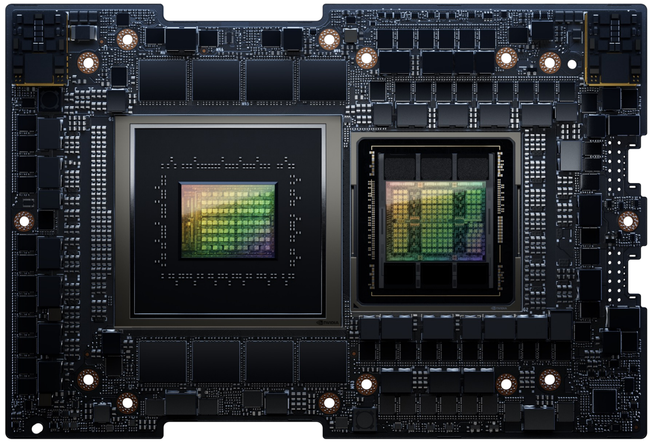
NVIDIA GH200 Grace Hopper Superchip

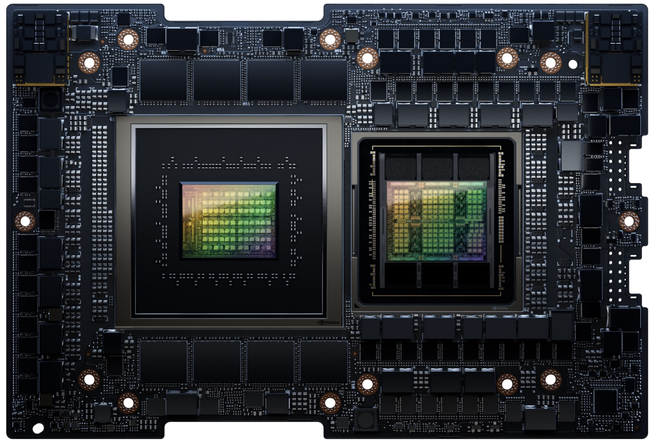
Processor For The Era of Accelerated Computing And Generative AI



72 Core Grace CPU | 4 PFLOPS Hopper GPU
96 GB HBM3 | 4 TB/s | 900 GB/s NVLink-C2C

- 7X bandwidth to GPU vs PCIe Gen 5
- Combined 576 GB of fast memory
- 1.2x capacity and bandwidth vs H100
- Full NVIDIA Compute Stack

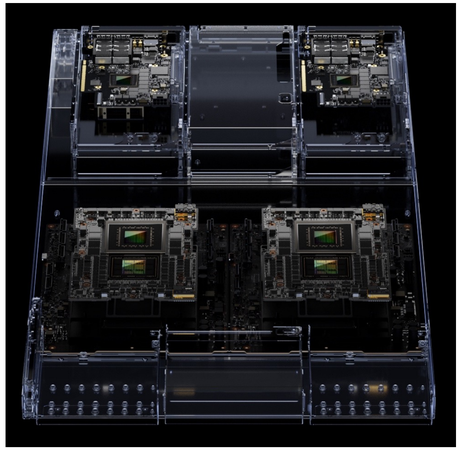
GH200 with HBM3
Available for order



72 Core Grace CPU | 4 PFLOPS Hopper GPU
144 GB HBM3e | 5 TB/s | 900 GB/s NVLink-C2C

- World's first HBM3e GPU
- Combined 624 GB of fast memory
- 1.7x capacity and 1.5x bandwidth vs H100
- Full NVIDIA Compute Stack

GH200 with HBM3e
Available late Q2 2024



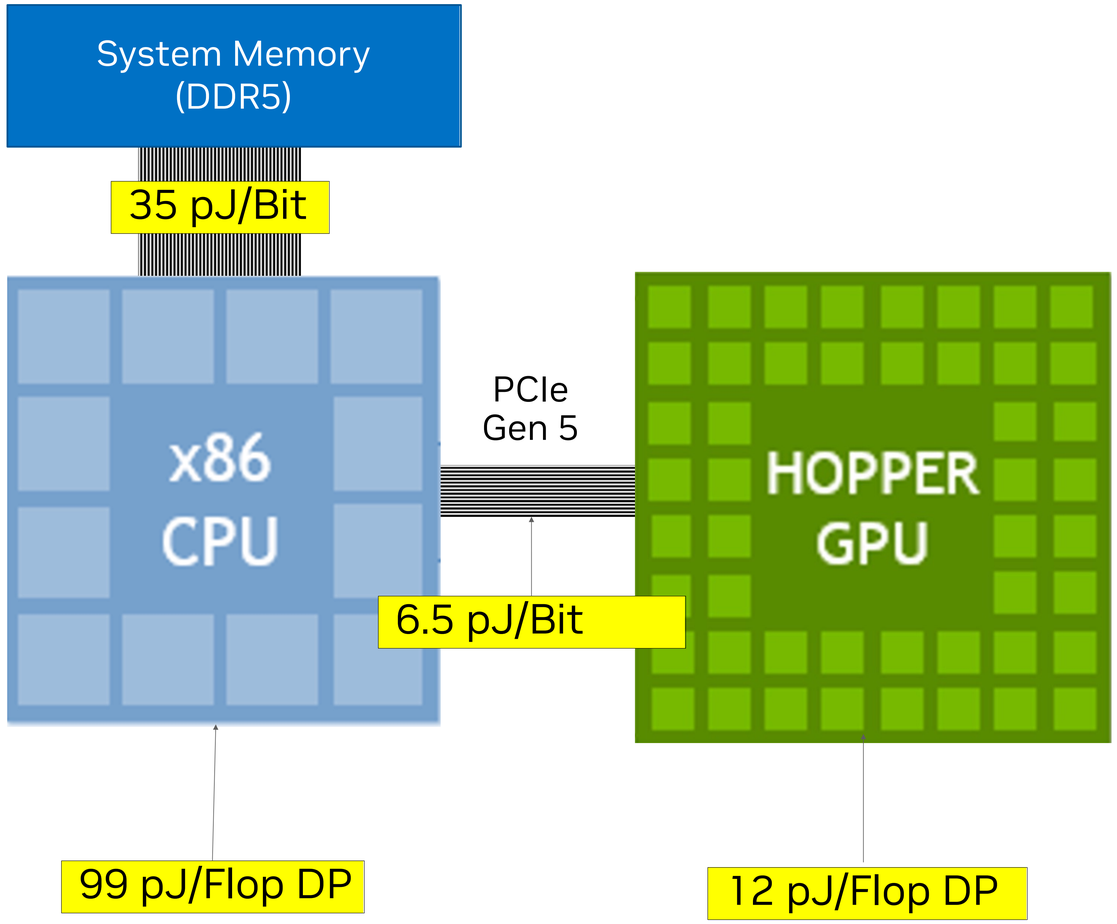
144 Core Grace CPU | 8 PFLOPS Hopper GPU
288 GB HBM3e | 10 TB/s | 900 GB/s NVLink-C2C

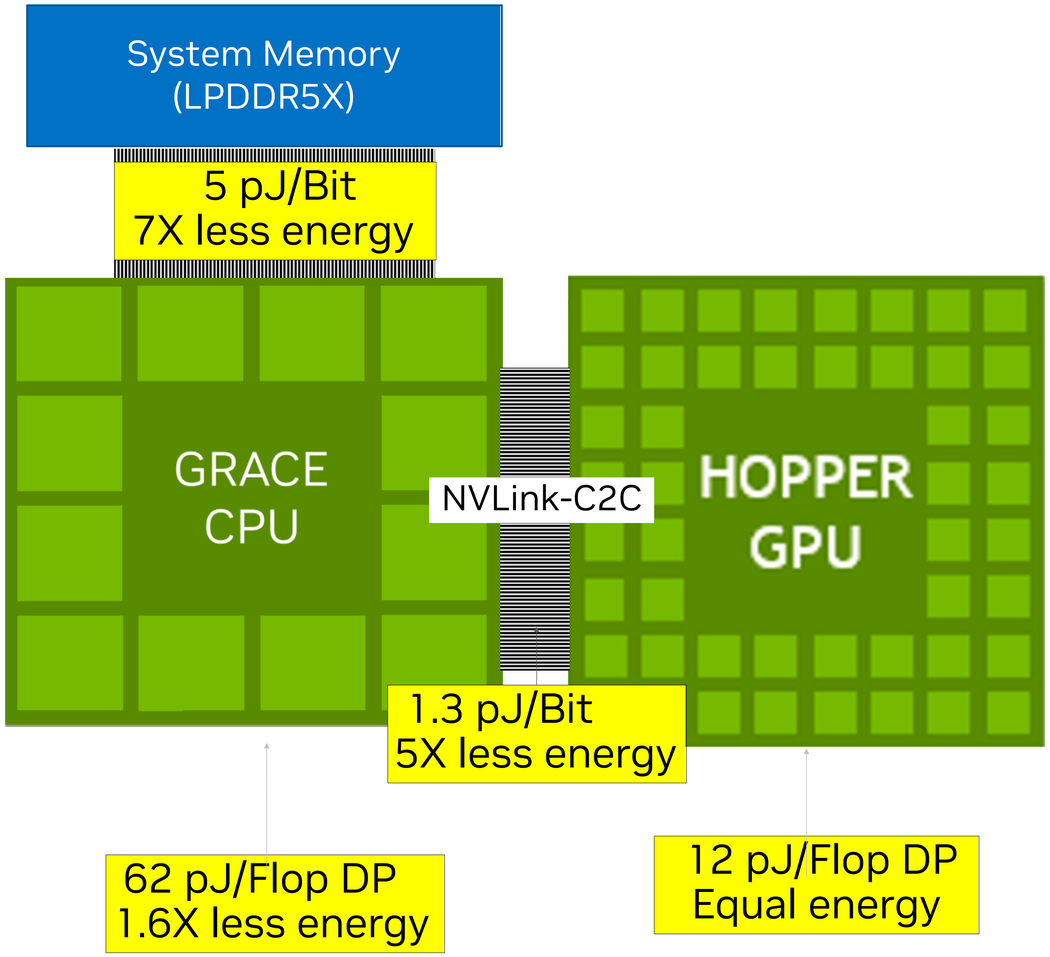
- Simple to deploy MGX-compatible design
- Combined 1.2 TB fast memory
- 3.5x capacity and 3x bandwidth vs H100
- Full NVIDIA Compute Stack

NVLink Dual GH200 System
Available late Q2 2024

Energy Efficient Design

More Efficient Computation and Data Movement



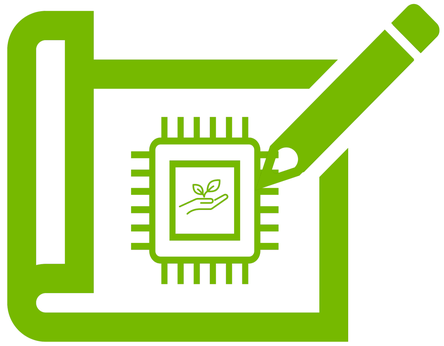


NVIDIA GH200 Optimizing Power and Performance at Data Center Scale

Higher Performance for Less Power

Energy Efficient Design

Energy efficient CPU, GPU, memory and IO



Automatic Power Steering

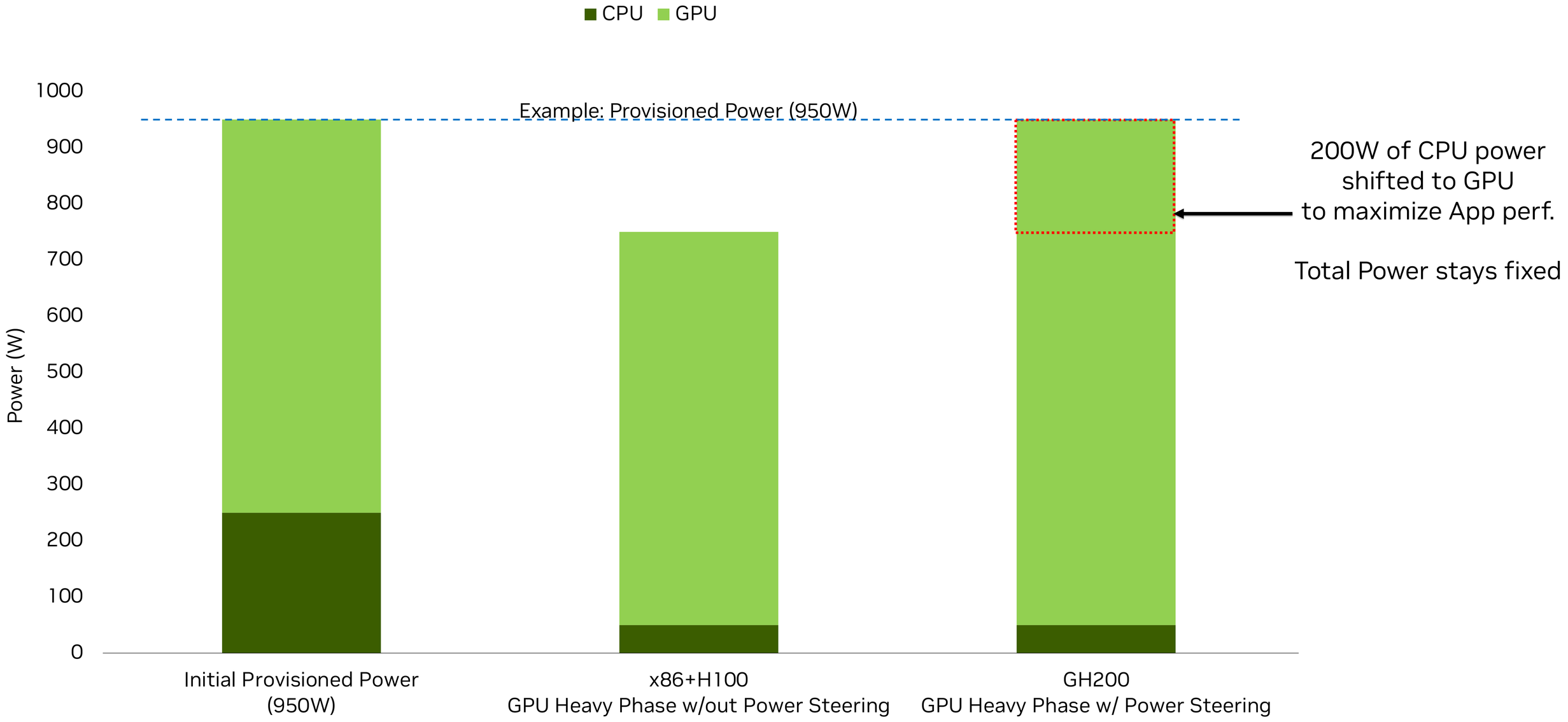
Automatically shifts power between CPU and GPU

Application Power-Tuning

Adjustable clocks for improved energy efficiency

Optimizing Performance Through Power Steering

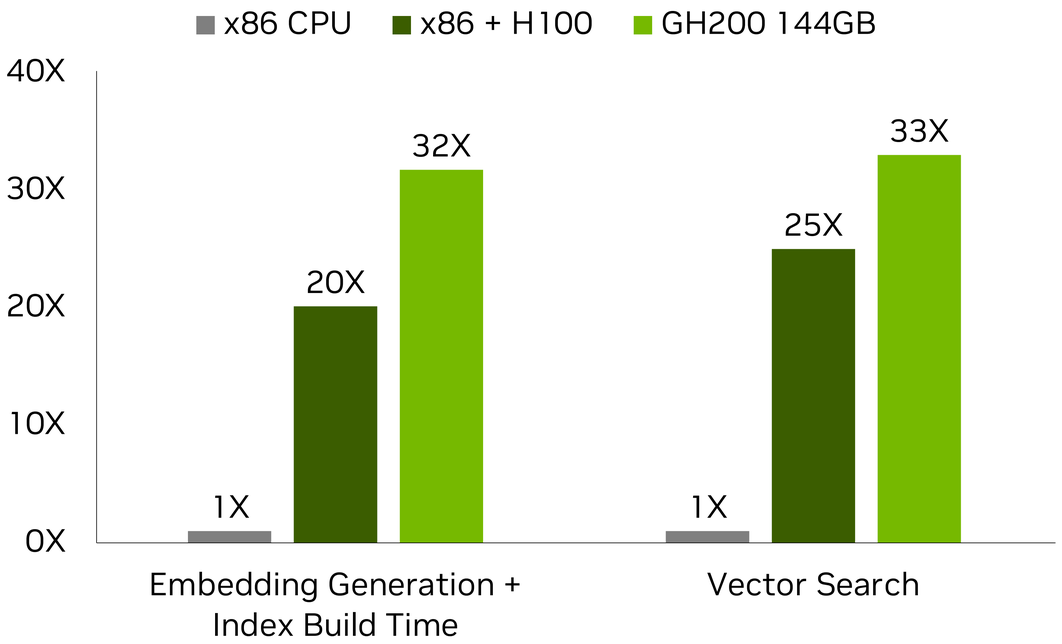
Getting the Most Out of provisioned power

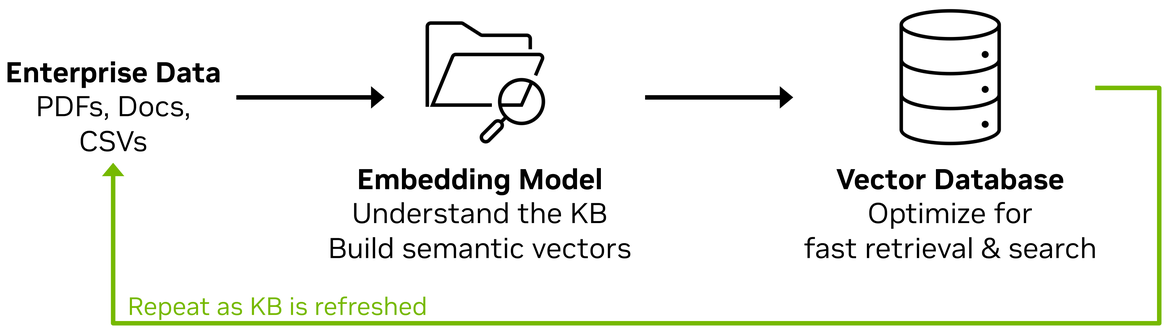


Deploying RAG Applications on GH200 Accelerates Performance

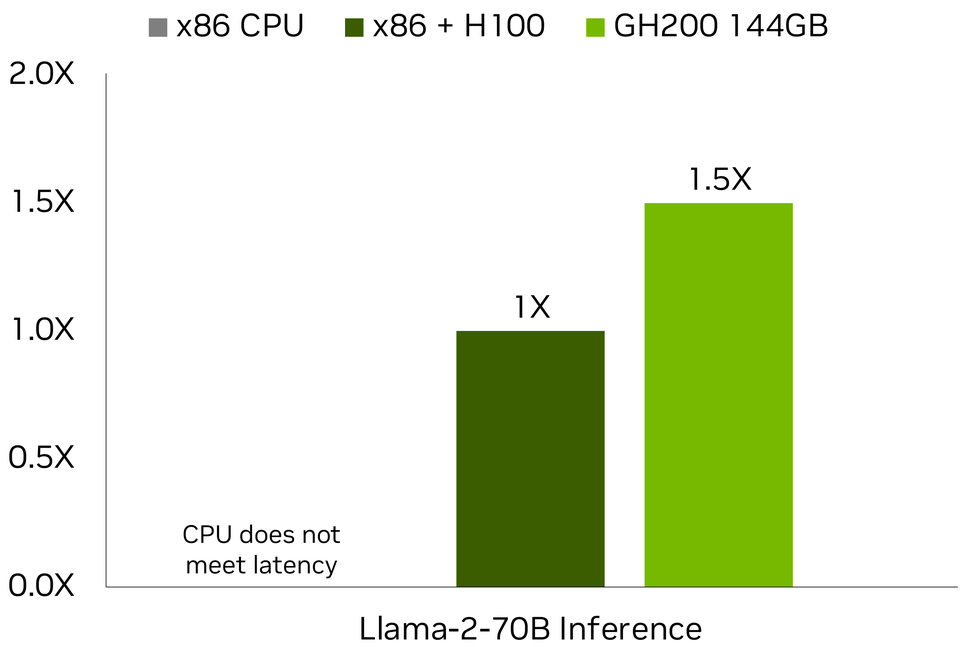
More than 30X faster Index Build and Vector Search vs x86 CPU

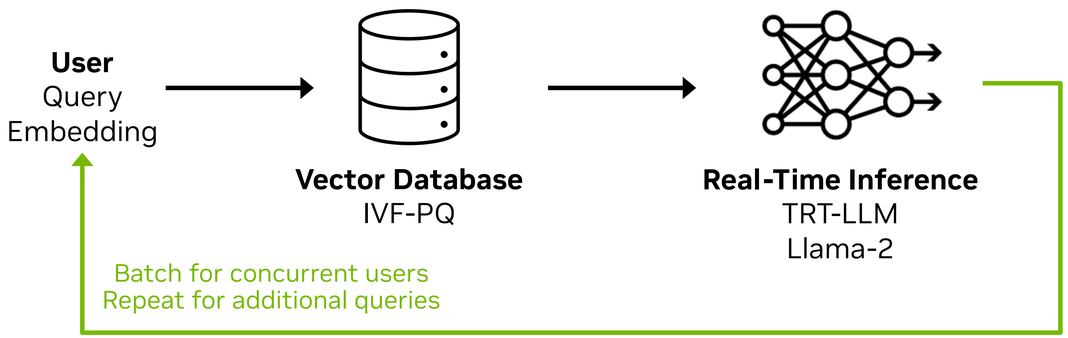
Document Retrieval, Ingestion, and Search





RAG LLM Inference with Llama-2



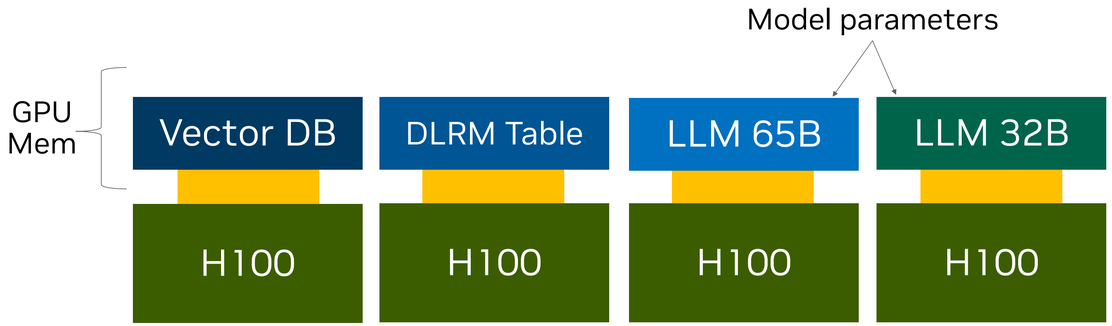


1 x GH200 (144GB) | 1 x H100 (80GB) | 2S Xeon Platinum 8480+
Embedding: Model MPNET Base v2 – Hugging Face | Batch = 1,024 | Output Vectors = 85M of size 768
Vector Search: Batch = 10,000 | Vector Search Queries = 10,000 over 85M vectors
LLM Inference: Llama.cpp (2S 8480+) and TensorRT-LLM (GH200, H100) | Max Batch | Throughput includes time to first token + token generation time
Further Performance Details: Deploying Retrieval-Augmented Generation Applications on NVIDIA GH200 Delivers Accelerated Performance | NVIDIA Technical Blog

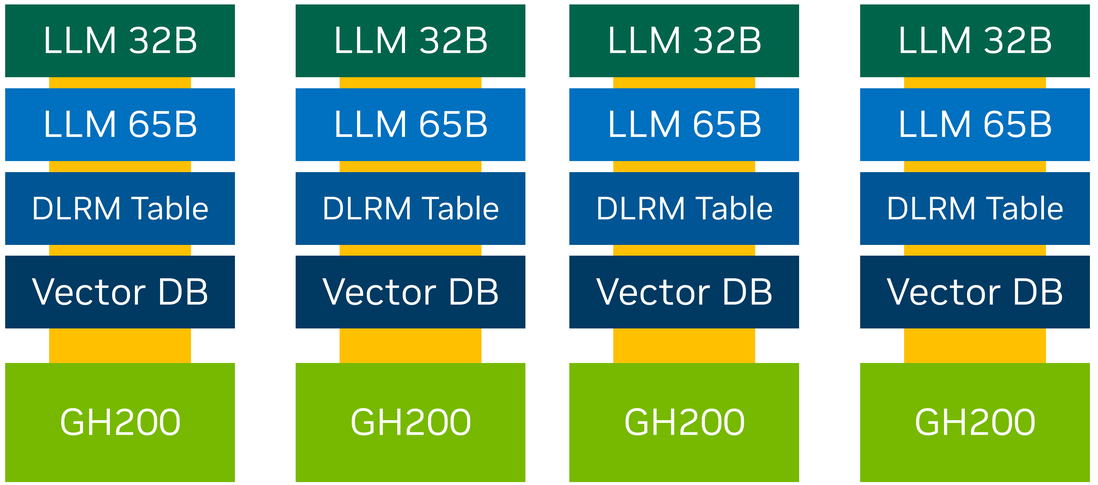
NVIDIA

GH200 Maximizes Utilization in Diverse Workload Environment

Typical GPU Data Center



GH200 Data Center

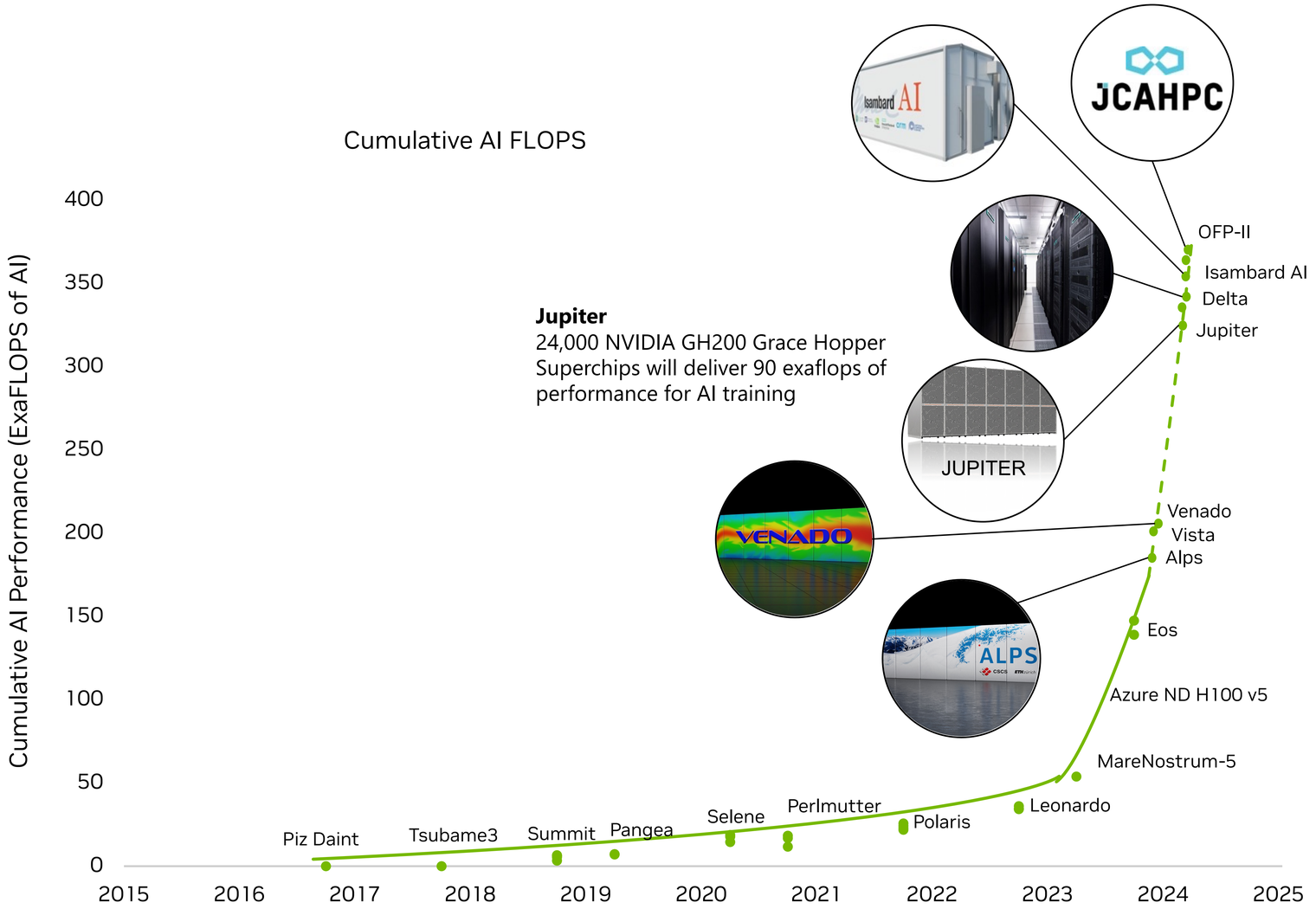


- Typical data center supports diverse workload using many models of varying sizes
- PCIe GPU has limited GPU memory and different models need to be stored by different GPUs
- Usage patterns cause some GPU to stay idle, while others are over-subscribed

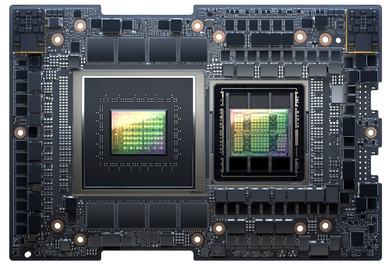
- GH200 with 624GB GPU memory can store all models on every node
- Every GH200 runs everything - scheduling is easier, scale out more efficient
- **Maximizing data center utilization**

Next-Gen AI Supercomputing Datacenter

350 ExaFLOPs of AI Performance to Drive Scientific Innovation



Isambard-AI
For advance robotics, data analytics, drug discovery, climate research, and more.
5,448 NVIDIA GH200 Grace Hopper Superchips to deliver a whopping 21 exaflops of AI

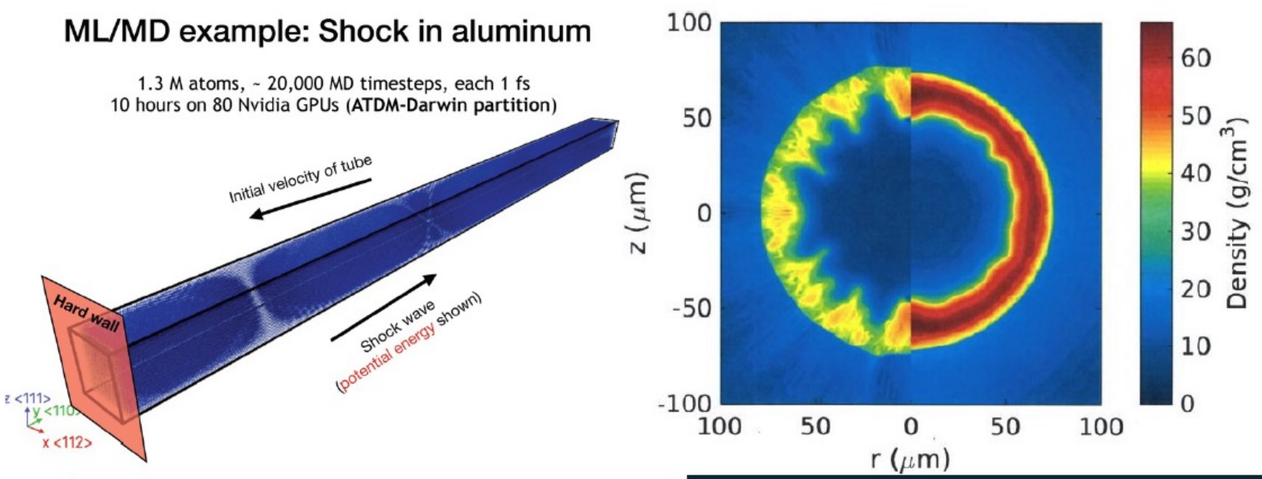


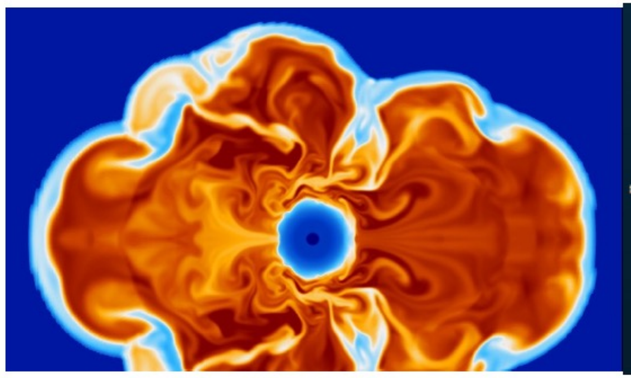
200 Exaflops AI Grace Hopper
Coming online 2024

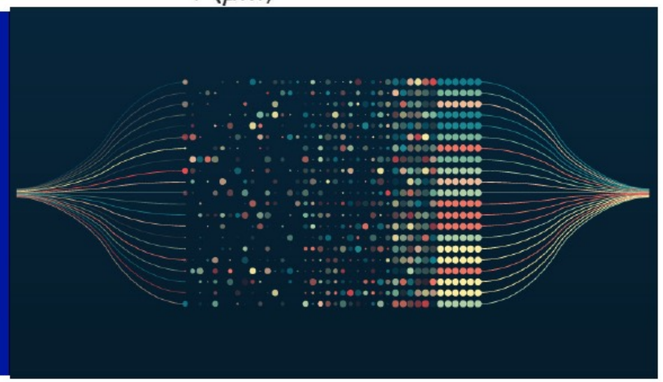
Venado
For computing improvements requires deeper collaboration with HPC/AI/ML as a driving force
AI emerged as a new foundation for progress in sciencecomputing
2560 NVIDIA GH200 Grace Hopper Superchips / 1840 Grace Superchips

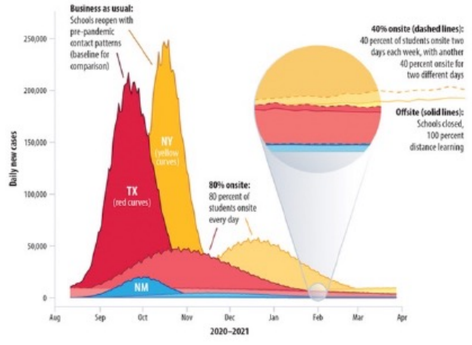
NVIDIA

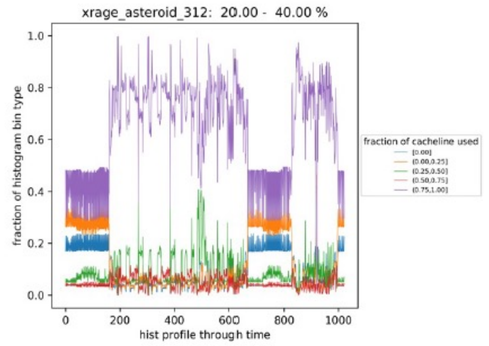
Venado: Efforts spanning applications to microarchitecture are underway











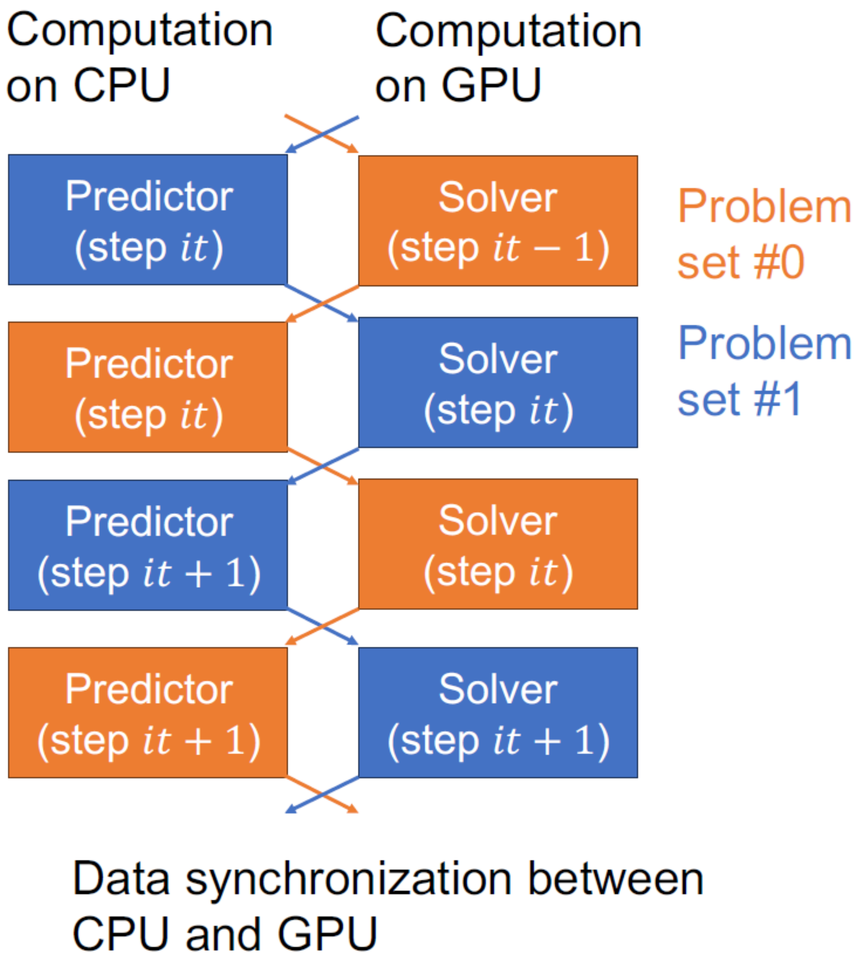
A LANL resource for codesign & discovery

Los Alamos NATIONAL LABORATORY

NVIDIA

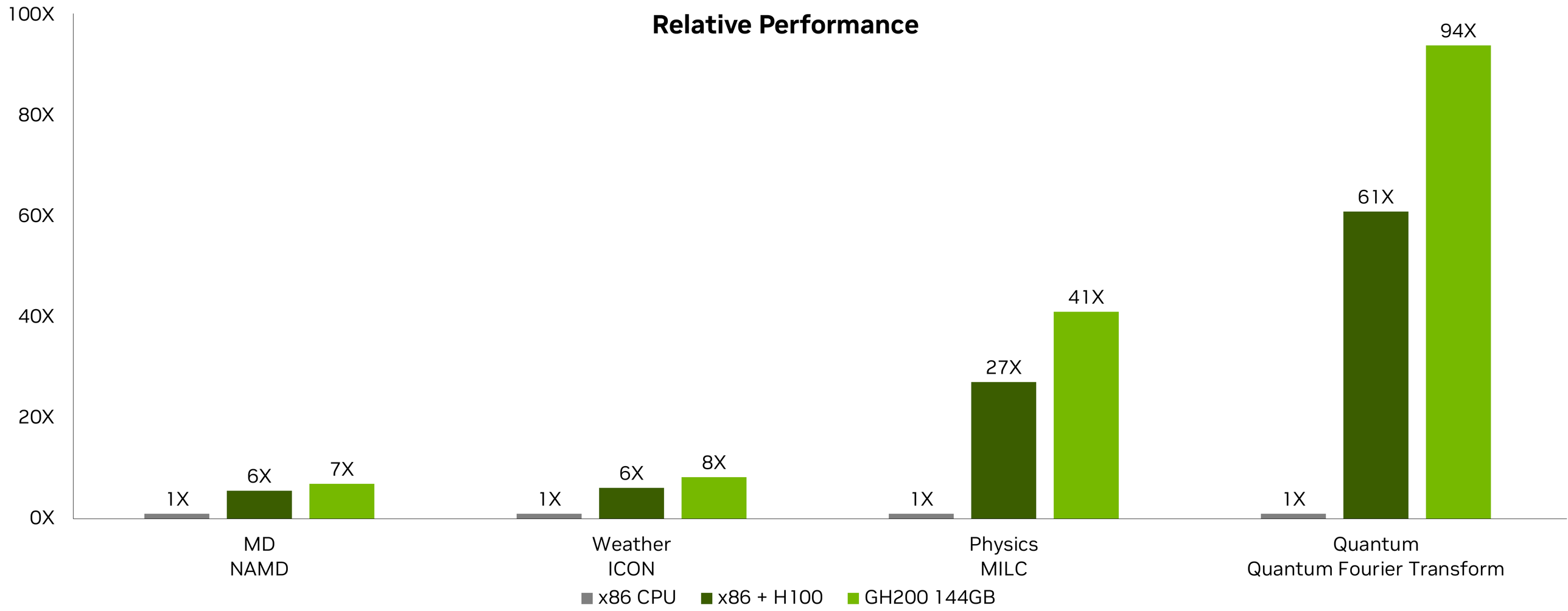
Use of data-driven predictor on a strongly-connected CPU-GPU environment

- Proposed method: concurrently use CPU and GPU by solving two sets of problems
  - CPU used for data-driven predictor for one set of problem
  - GPU used for iterative solver for the other set of problem
  - Synchronize data before and after predictor/iterative solver using fast CPU-GPU interconnect
- Additional GPU memory required is small
  - Matrix $\boldsymbol{A}$ can be shared for the two sets of problems on GPU memory
  - The largest problem size that can be solved is almost the same as that of a standard GPU-based method
- Both CPU and GPU can be fully utilized throughout the computation
  - Elapsed time of predictor@CPU and solver@GPU is about the same and the synchronization time is negligible



NVIDIA GH200 Delivers Breakthrough HPC Performance

Up to 94X Faster than 2S x86 CPU



Single Node GH200 vs 2S Xeon Platinum 8480+
Quantum: Quantum Fourier Transform | Physics: MILC Apex Medium | Weather: ICON QUBICC 191 Levels 80km Radiation MD: NAMD LaINDY_ColVars

NVIDIA GH200 Grace Hopper Superchip Now Available in Launchpad

Sign-up Today

- **Access NVIDIA GH200 via a web browser**

- **Test key features**
  - **Grace Hopper:** NVLink-C2C and memory management
  - **Grace CPU:** 72 Arm Cores and high-bandwidth memory subsystem and fabric
  - **Hopper GPU:** Transformer engine and high-performance compute

- **Run key workloads**
  - LLMs
  - HPC Applications
  - Data Analytics
  - Real-world applications and case studies

In This Free Hands-On Lab, You'll Experience:

Seamless integration of the NVIDIA GH200 Grace Hopper™ Superchip with NVIDIA's software stacks.

Interactive demos of memory bandwidth, CPU and GPU coherence, the NVIDIA NVLink®-C2C interconnect, oversubscription of GPU memory, and converting x86 Intrinsics to Arm® Neon.

Examples of real-world applications and case studies, including large language model (LLM) demonstrations that showcase the types of problems GH200 can solve.

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

NVIDIA