NCHC Grace Workshop

Platforms and Products

NCHC 2025

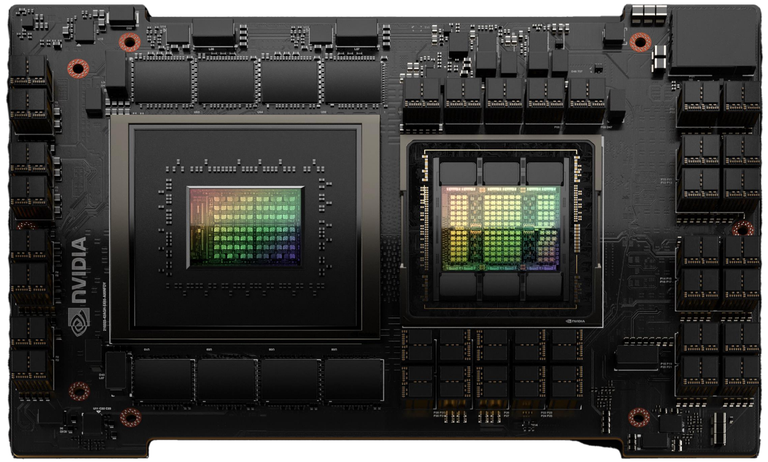
THIS INFORMATION IS INTENDED TO OUTLINE OUR GENERAL PRODUCT DIRECTION. MANY OF THE PRODUCTS AND FEATURES DESCRIBED HEREIN REMAIN IN VARIOUS STAGES AND WILL BE OFFERED ON A WHEN-AND-IF-AVAILABLE BASIS. THIS ROADMAP DOES NOT CONSTITUTE A COMMITMENT, PROMISE, OR LEGAL OBLIGATION AND IS SUBJECT TO CHANGE AT THE SOLE DISCRETION OF NVIDIA. THE DEVELOPMENT, RELEASE, AND TIMING OF ANY FEATURES OR FUNCTIONALITIES DESCRIBED FOR OUR PRODUCTS REMAINS AT THE SOLE DISCRETION OF NVIDIA. NVIDIA WILL HAVE NO LIABILITY FOR FAILURE TO DELIVER OR DELAY IN THE DELIVERY OF ANY OF THE PRODUCTS, FEATURES, OR FUNCTIONS SET FORTH IN THIS DOCUMENT.

NVIDIA Grace and Hopper Platform

NVIDIA Grace for HPC & AI Infrastructure

Grace Hopper Superchip

Giant Scale AI & HPC



Accelerated applications where CPU performance and system memory BW are critical; extreme and highly atomic collaboration between CPU & GPU contexts for flagship AI & HPC

Grace CPU Superchip

CPU Computing

Applications that run on CPU but where absolute performance, energy efficiency, and datacenter density matter, such as in scientific computing, data analytics, and hyperscale computing applications

NVIDIA Grace CPU Superchip

Breakthrough Performance and Efficiency for the Modern Data Center

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 960GB of data center enhanced LPDDR5X Memory that
delivers up to1TB/s of memory bandwidth

Fast and Flexible CPU IO

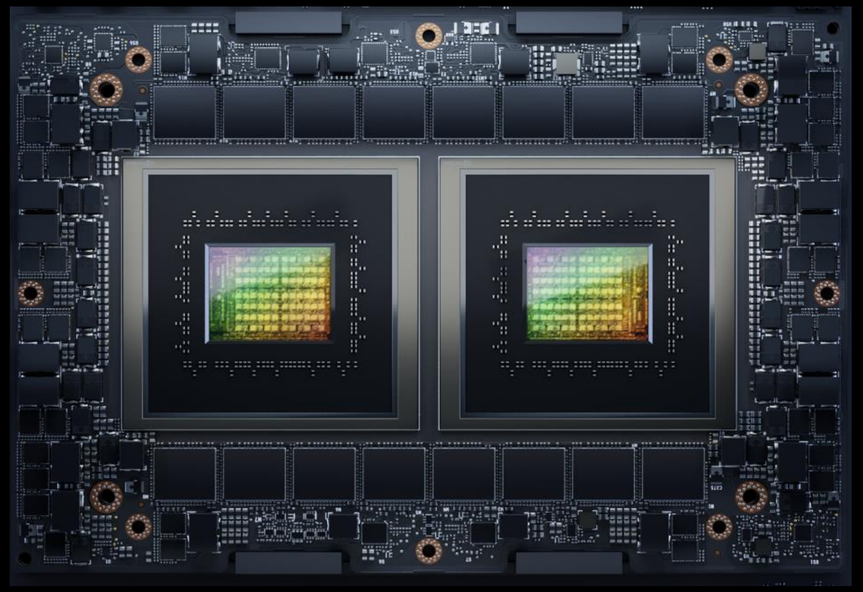
Up to 8x PCIe Gen5 x16 interface. PCIe Gen 5 up to 128GB/s
2X more bandwidth compared to PCIe Gen 4

Full NVIDIA Software Stack

AI, Omniverse

Continued Innovation

Grace-Next



144 Arm Neoverse V2 Cores | 228MB L3 Cache
3.2 TB/s NVIDIA Scalable Coherency Fabric | 960GB LPDDR5X

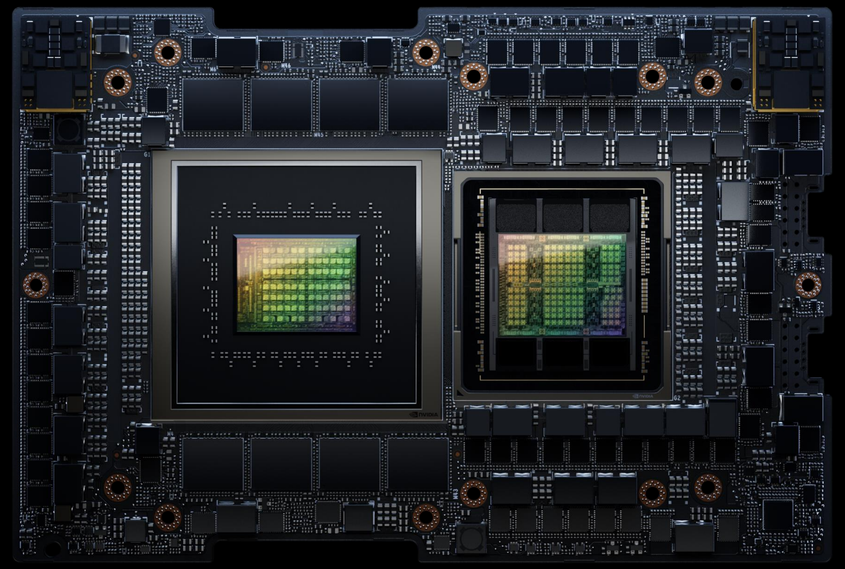
NVIDIA GH200 Grace Hopper Superchip

Built for the New Era of Accelerated Computing and Generative AI

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 HPC, enterprise, and cloud

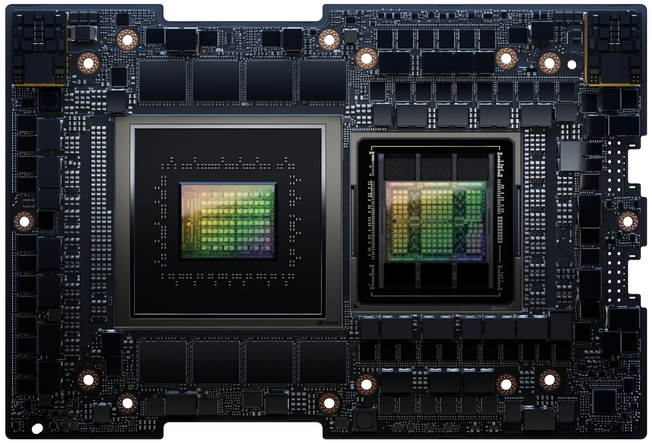
Best Perf/TCO for diverse workloads
Maximize data center utilization and power efficiency



900GB/s NVLink-C2C | 624GB High-Speed Memory
4 PF AI Perf | 72 Arm Cores

NVIDIA GH200 Grace Hopper Superchip

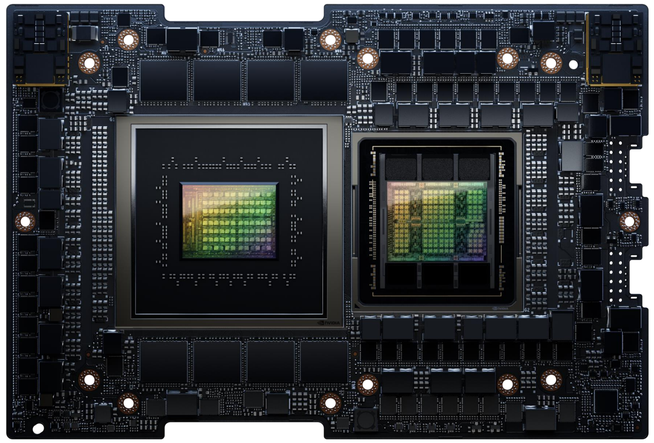
Processor For The Era of Accelerated Computing And Generative AI



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

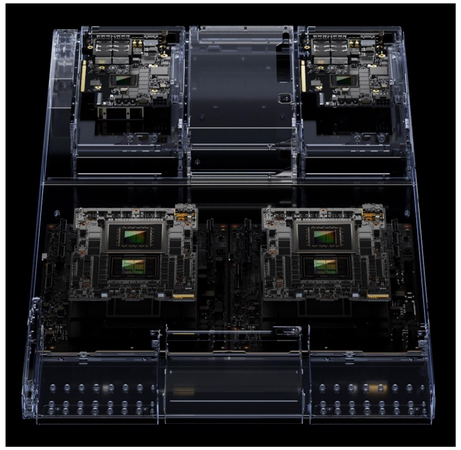
GH200 with HBM3



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



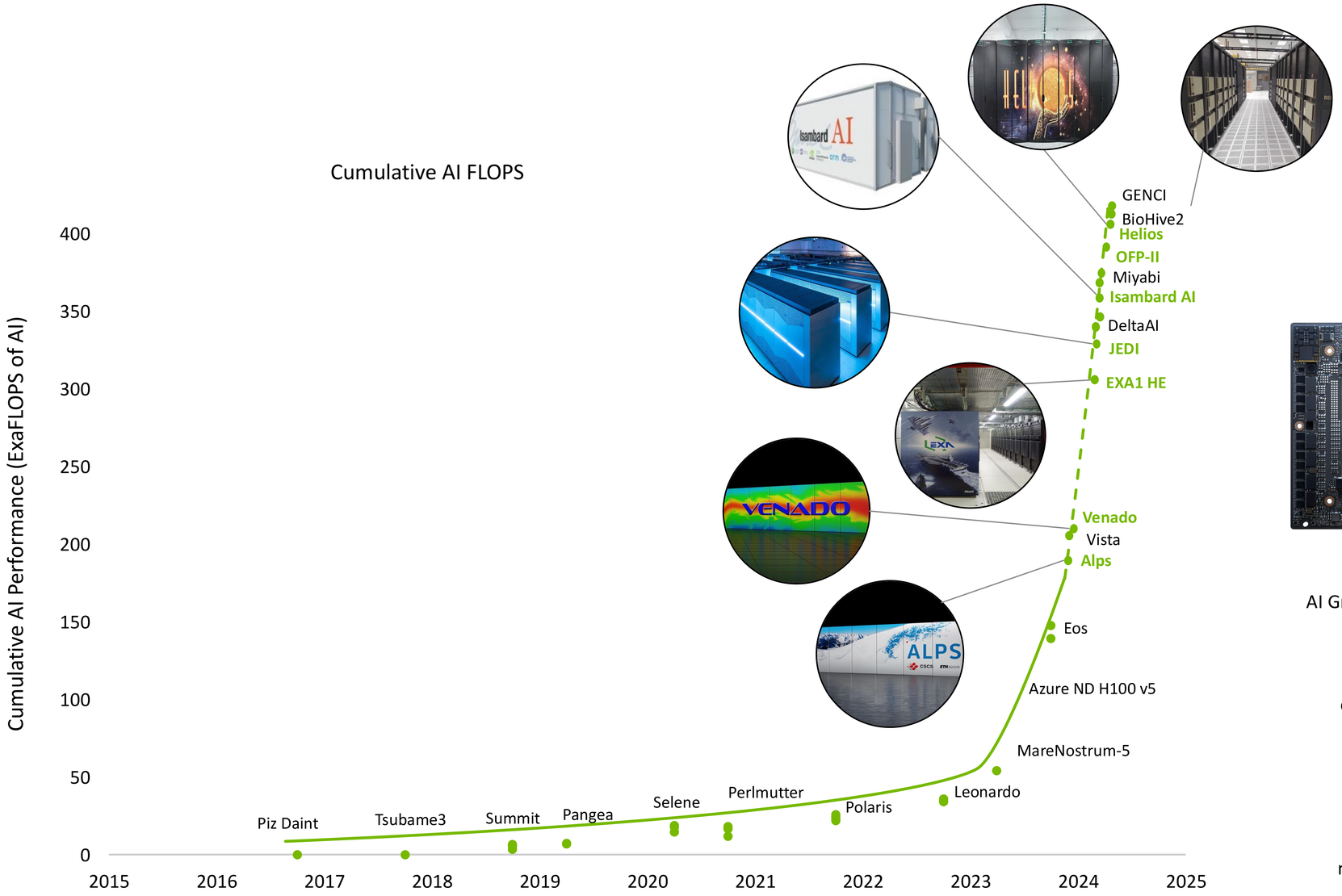
144 Core Grace CPU | 8 PFLOPS Hopper GPU
288 GB HBM3e | 10 TB/s | 900 GB/s NVLink-C2C

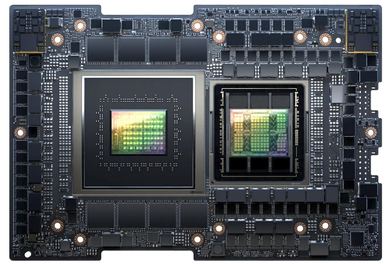
- Simple to deploy MGX-compatible design
- Combined 1.2 TB fast memory
- 3.5x capacity and 3x bandwidth vs H100
- Full NVIDIA Compute Stack

GH200 NVL2

Grace Hopper Powers AI Supercomputing Datacenters

Grace Hopper Will Deliver 200 Exaflops of AI performance for Groundbreaking Research





200 ExaFLOPS
AI Grace Hopper coming online 2024

80%
of Hopper are Grace Hopper

2X
More energy efficient

7
new GH200 supercomputers

First European Grace Hopper Supercomputer Online

ALPS, at CSCS

- Fastest AI Supercomputer in Europe
- 20 Exaflops of AI
- 10X more energy efficient than Piz Daint
- Powered by 10,000 Grace Hopper Superchips
- HPC and AI to Advance Weather, Climate (1km global models), and Material Science

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Hewlett Packard Enterprise

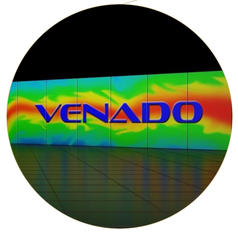
NVIDIA

Grace and Grace Hopper Now in Full Production

Grace and Grace Hopper transforming HPC and AI

Cyfronet
(Helios)

Univ. of Bristol
(Isambard AI)



LANL
(Venado)

TACC
(Vista)

NVIDIA

Univ. of Bristol
(Isambard 3)

BSC
(MareNostrum 5)

CSCS
(ALPS)

KAUST
(Shaheen-III)

NCHC
(Taiwania 4)

FZJ
(JUPITER)

JCAHPC
(Miyabi)

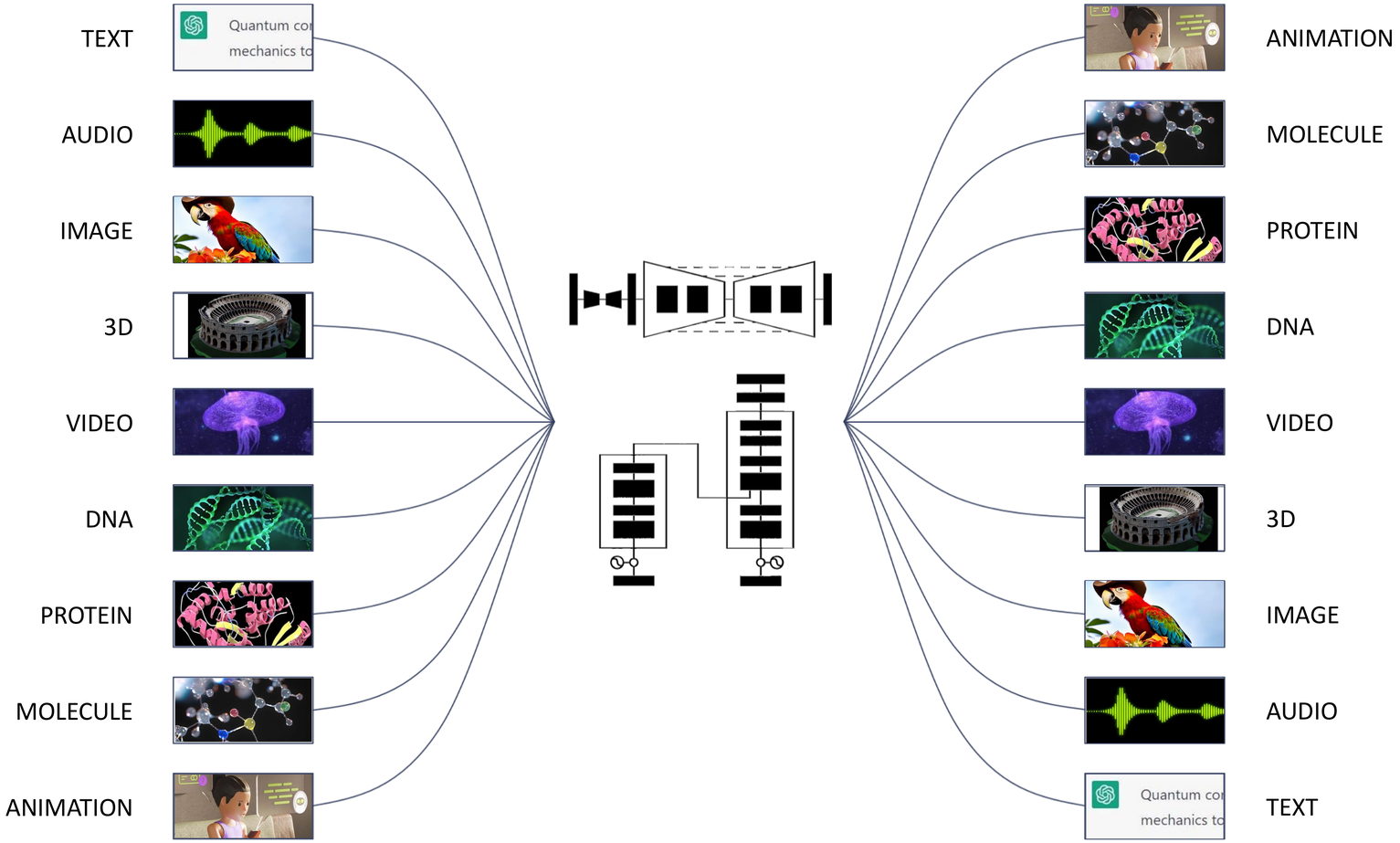
The “Raise of the AI Machines”

NVIDIA Blackwell Platform

Generative AI

Multi-modal intelligence

Generative AI refers to machine learning algorithms that enable computers to **use existing or past content** like text, audio and video files, images, and even code…

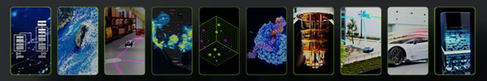


… to **generate completely original artifacts** that would look like the real deal.

NVIDIA Blackwell Platform

NIM
CUDA-Accelerated
Agentic AI Libraries

Omniverse
CUDA-Accelerated
Physical AI Libraries

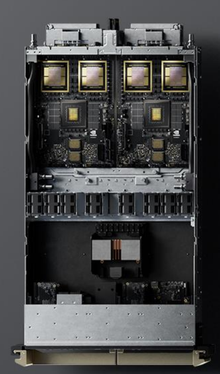


CUDA-X Libraries

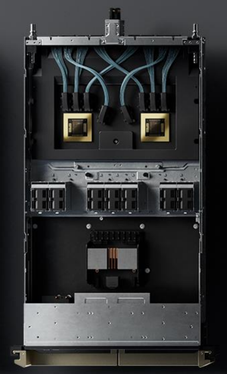
CUDA • DOCA • NCCL
Cluster-Scale Software
System Software
Chip Software

Accelerated
Software Stack

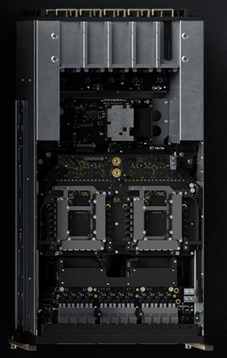
GB200 NVL72 SuperPOD



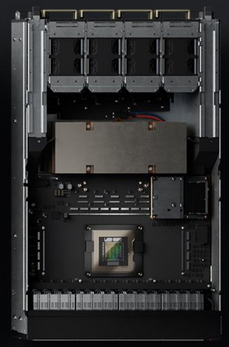
Grace Blackwell
MGX Node



NVLink Switch



Quantum Switch

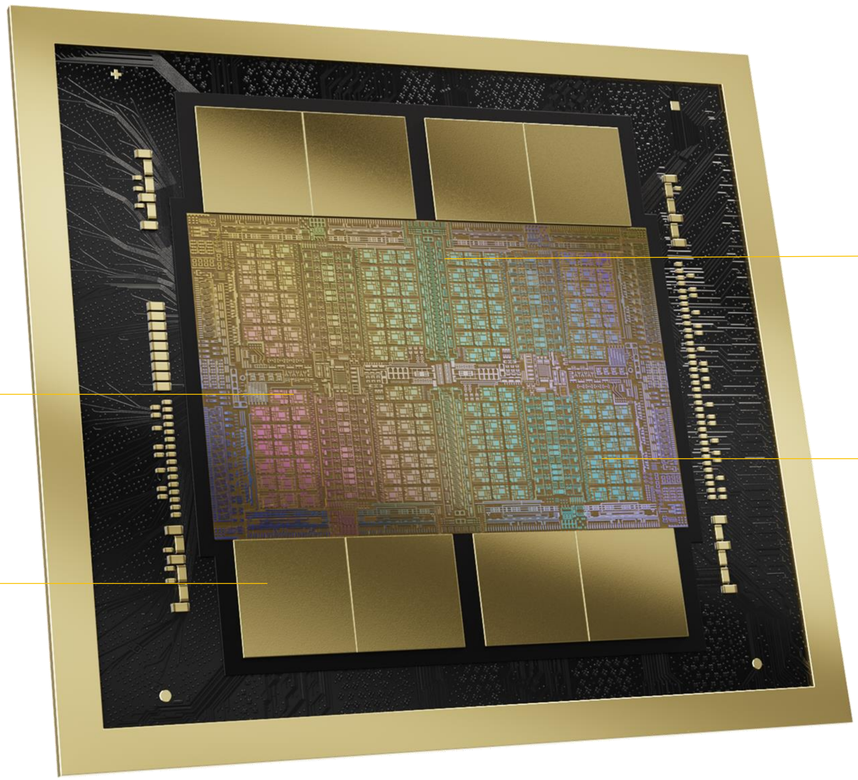


Spectrum-X Switch

Chips Purpose-Built for AI Supercomputing
GPU | CPU | DPU | NIC | NVLink Switch | IB Switch | Enet Switch

NVIDIA Blackwell

The Two Largest Dies Possible—Unified as One GPU



Reticle-sized Die 1

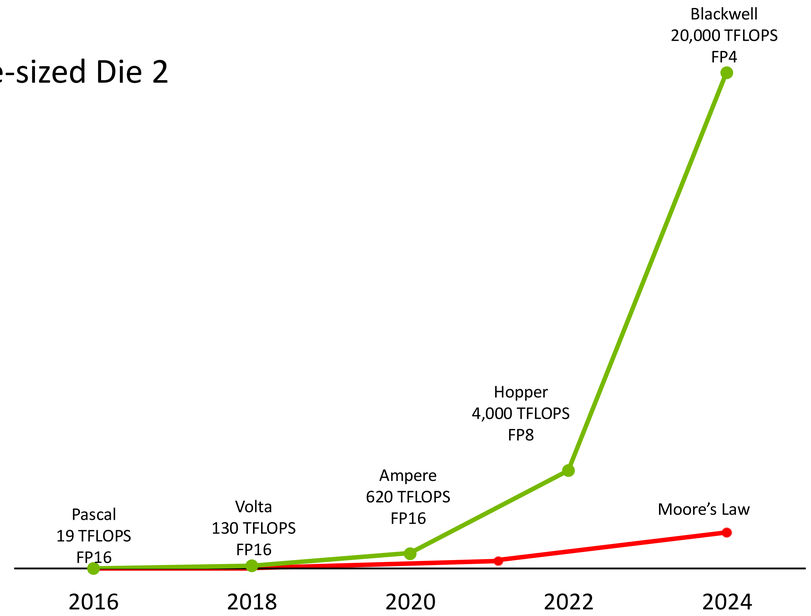
Fast Memory
192GB HBM3e

2 reticle-limited dies operate as One Unified CUDA GPU

NV-HBI 10TB/s High Bandwidth Interface

Full performance. No compromises

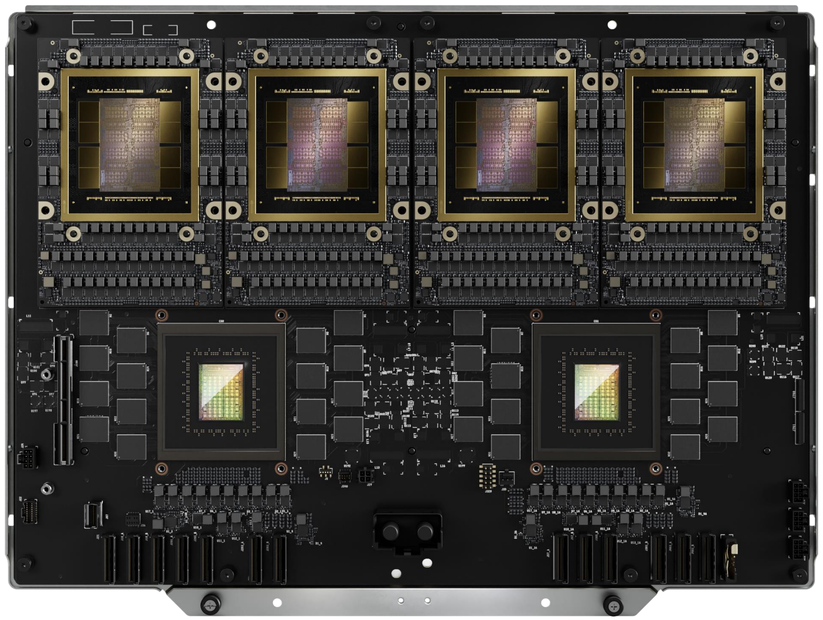
Reticle-sized Die 2



10 PetaFLOPS FP8 | 20 PetaFLOPS FP4
192GB HBM3e | 8 TB/sec HBM Bandwidth | 1.8TB/s NVLink

NVIDIA GB200 Grace Blackwell NVL4 Superchip

Flexible module designed for Blackwell-accelerated HPC and AI



NVIDIA GB200 Grace Blackwell NVL4 Superchip

Single-Server Solution

HPC and AI Hybrid Workloads

4-GPU NVLink Domain With 1.3T Coherent Memory

2.2X
Simulation

1.8X
Training

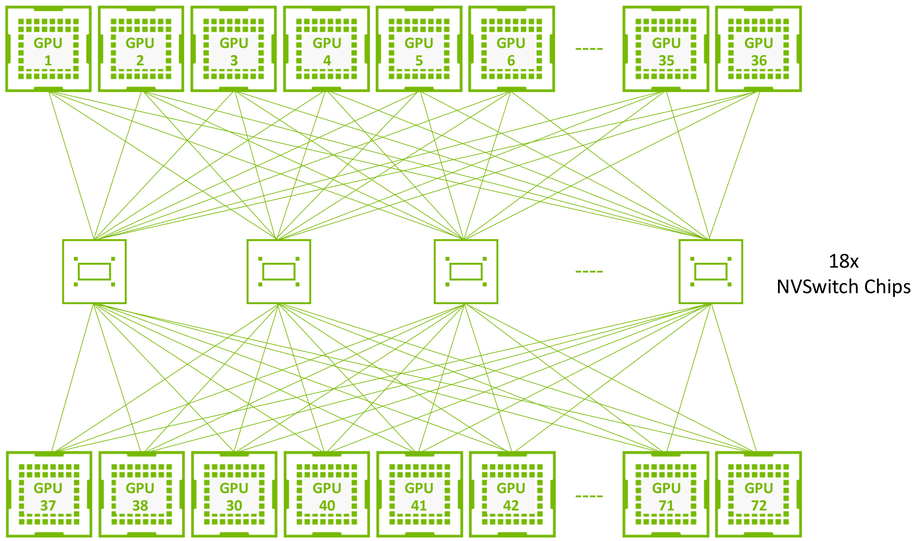
1.8X
Inference

NVIDIA GB200 NVL4 versus NVIDIA GH200 NVL4 | Scientific Computing – MILC | GNN Training – GraphCast 37M parameters, BF16 | Inference – OpenFold evoformer block, FP16, GenSLM Llama2-7B, FP16

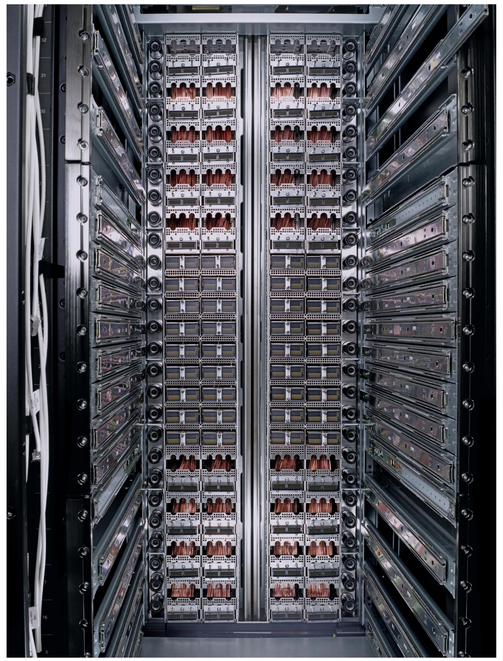
NVIDIA

Minimizing Multi-GPU Communication Latency

Expanding GPU domains with NVIDIA NVSwitch and NVIDIA NVLink

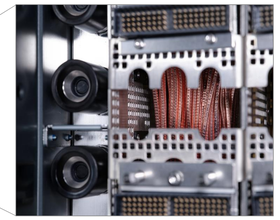


GB200 NVL72 Rack

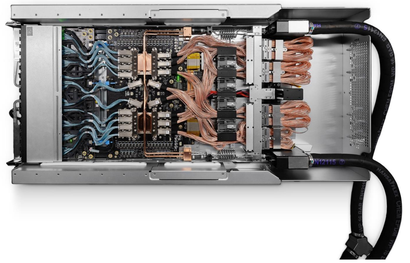


NVLink Spine

NVLink Cables



NVLink Switch Tray



NVLink	NVL Domain	All-to-All	All-Reduce
1.8 TB/s	72 GPUs	130 TB/s	260 TB/s

NVIDIA GB200 NVL72

Delivers New Unit of Compute

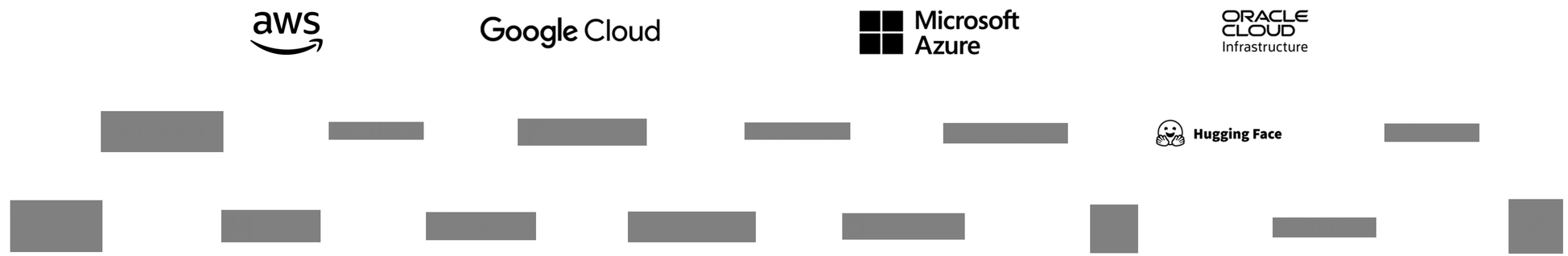
GB200 NVL72

- 36 GRACE CPUs
- 72 BLACKWELL GPUs
- Fully Connected NVLink Switch Rack

Training FP8	720 PFLOPs
Inference FP4	1,440 PFLOPs
NVL Model Size	27T params
Multi-Node All-to-All	130 TB/s
Multi-Node All-Reduce	260 TB/s

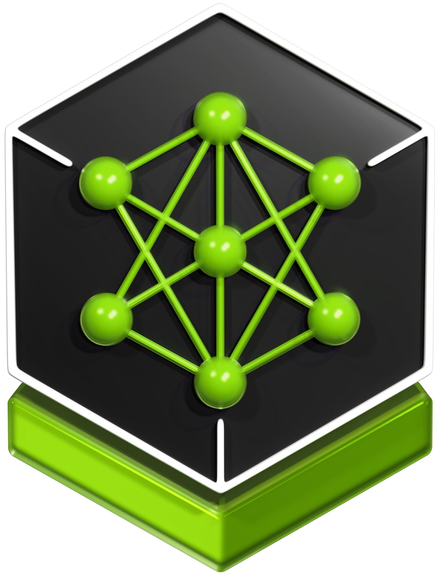
Blackwell Ecosystem

Building and Deploying AI Factories



NVIDIA NIM: Optimized AI Models Run Up to 5X Faster

Community models – partner models – NVIDIA models



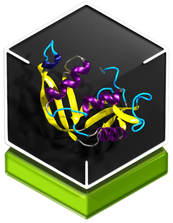
NVIDIA INFERENCE MICROSERVICE

Pre-Trained AI Models
Packaged and Optimized to Run Across
CUDA Installed Base

Speech

Digital Human

Computer Vision



Biology

Simulation

Language

Regional Language

Vision Language

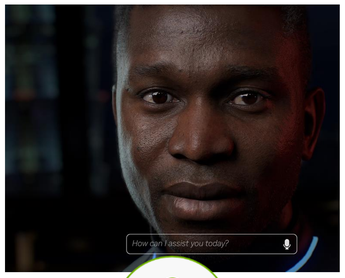
RAG

ADEPT gettyimages Google Meta MIT MISTRAL AI_ NVIDIA shutterstock snowflake

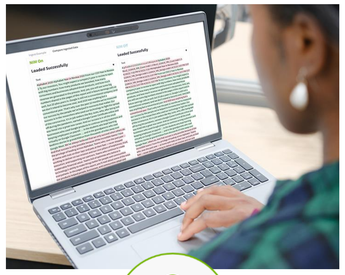
NVIDIA Blueprints

Available on build.nvidia.com

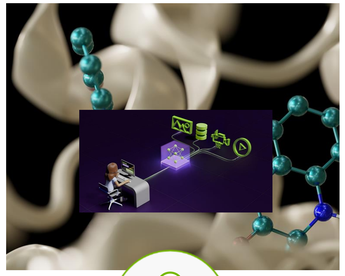
Digital Humans
for Customer Service



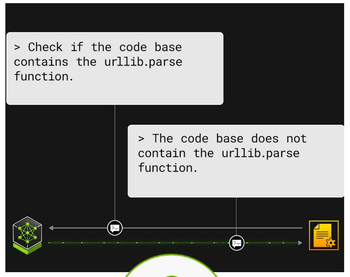
Multimodal PDF Data Extraction for
Enterprise RAG



Generative Virtual Screening for
Drug Discovery

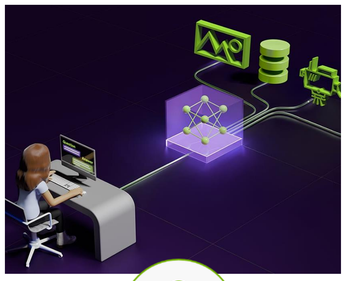


Vulnerability Analysis
for Container Security



AI Virtual Assistants
for Customer Service

Visual AI Agent for Video Search and
Summarization



Reference Application

Sample Data

Reference Code

Architecture

Customization Tools

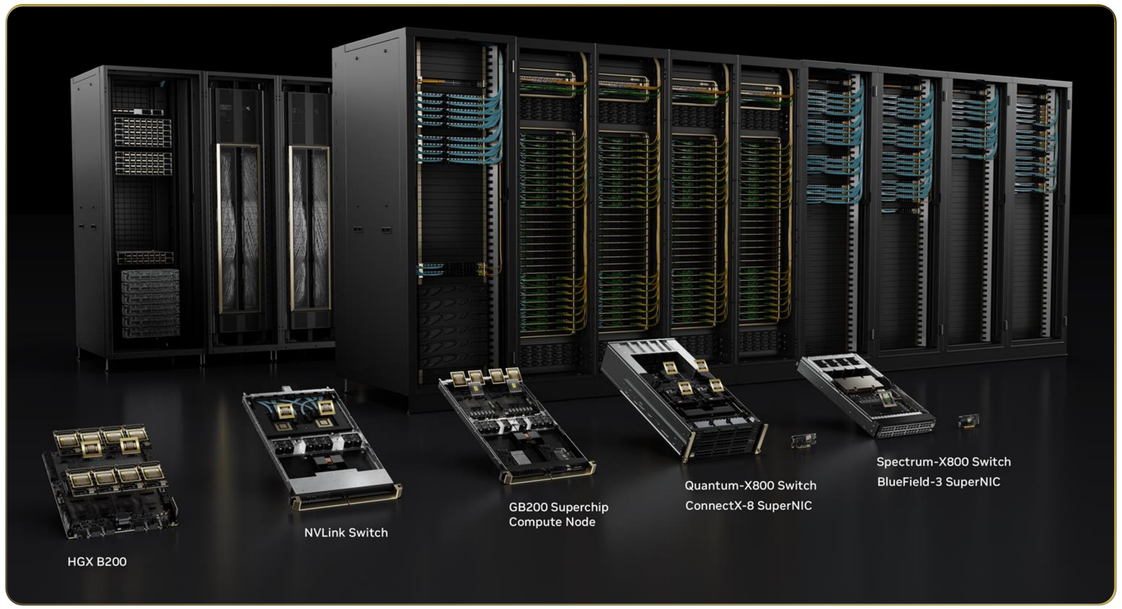
Orchestration Tools

Future Technologies

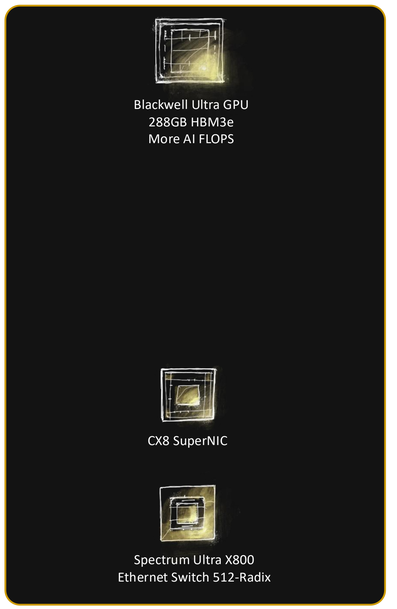
One-Year Rhythm | Supercluster Scale | Full Stack | CUDA Everywhere

Hopper

Blackwell



Blackwell Ultra



Rubin

2022 2023 2024 2025 2026 2027

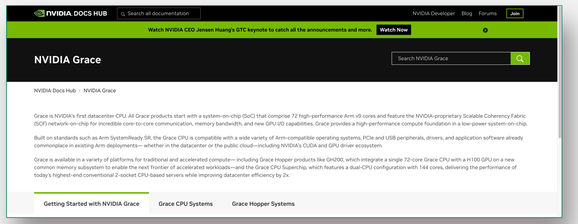
Resources and Assets

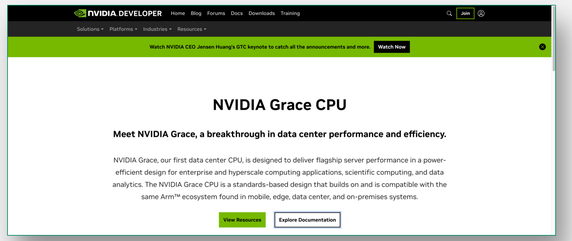
NVIDIA Grace CPU

Assets and Blogs

- Grace CPU Customer Deck
- Grace Hopper Customer Deck
- Revolutionizing Data Center Efficiency with the NVIDIA Grace Family
- Grace CPU Power Efficiency Video
- Grace CPU Bristol PR
- Grace CPU Architecture In-Depth Blog
- Grace CPU Energy Efficiency Blog
- Grace Hopper Architecture In-Depth Blog
- Grace Hopper Recommender System Blog
- Simplifying GPU Programming for HPC with NVIDIA Grace Hopper Superchip
- Building High-Performance Applications in the Era of Accelerated Computing
- NVIDIA GH200 and Grace CPU Accelerates Murex MX.3 Analytics and Reduces Power Consumption
- Boosting Mathematical Optimization Performance and Energy Efficiency on the NVIDIA Grace CPU

Gettering Started with NVIDIA Grace CPU and GH200





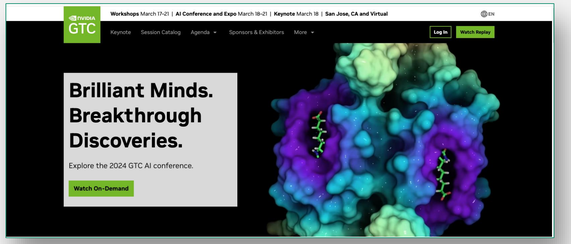
Documentation

- NVIDIA Grace Documents Hub
- NVIDIA Grace Developers Page

Data Sheets and Whitepapers

- NVIDIA Grace Hopper Datasheet
- NVIDIA Grace Hopper Whitepaper
- NVIDIA Grace CPU Datasheet
- NVIDIA Grace CPU Whitepaper

NVIDIA Grace and Grace Hopper at GTC



GTC 2024 Sessions

- Grace Hopper Superchip Architecture and Performance Optimizations for Deep Learning Applications [S61159]
- Performance Optimization for Grace CPU Superchip [S62275]
- Harnessing Grace Hopper's Capabilities to Accelerate Vector Database Search [S62339]
- A Deep Dive into the Latest HPC Software [S61203]
- Customer use Cases for NVIDIA Grace Hopper Superchip and Grace CPU Superchip [S62247]
- Accelerating Scientific Workflows with the NVIDIA Grace Hopper Platform [S62337]
- Accelerating Linear Solvers on NVIDIA Grace [S62529]
- Accelerate Gurobi Optimization with Energy-Efficient NVIDIA Grace CPU [S62555]
- NERSC-10 Benchmarks on Grace Hopper and Milan-A100 Systems: A Performance and Energy Case Study [S61402]
- Scientific Computing With NVIDIA Grace and the Arm Software Ecosystem [S61598]
- Crafting User Experience for Retrieval Augmented Generation (Presented by Lambda) [S62997]
- Early Science with Grace Hopper at Scale on Alps [S62157]
- Energy-Efficient GPU Computing With Mixed-Precision Modeling for Climate/Weather Applications [S61379]
- A New Era of AI-Driven Electronic Design Automation on Accelerated Computing [S62846]
- Magnum IO GPUDirect, NCCL, NVSHMEM, and GDA-KI on Grace Hopper and Hopper systems [S61368]
- Energy and Power Efficiency for Applications on the Latest NVIDIA Technology [S62419]
- Revolutionizing Supercomputing: Unleashing the Power of Grace [S62579]
- Unleashing Innovation With OpenShift on NVIDIA Grace Hopper Superchip (Presented by Red Hat, Inc.) [S63178]
- HPC in Quant Finance: Leveraging AI/ML Revolution in Graph-Based Computation for Risk Management of Exotics on NVIDIA GPUs, Grace Hopper and Triton Inference Server [S61856]