




NVIDIA Updates

May 2024



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.



Contents

- AI and Science

- NVIDIA Hopper

- NVIDIA Grace & Grace Hopper

- Programming the NVIDIA Platform

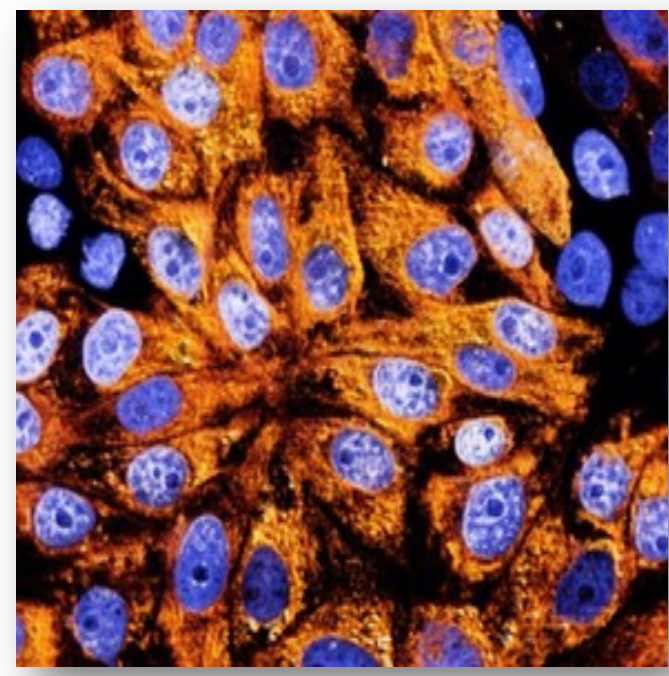
- Next Generation: NVIDIA Blackwell

The background features a series of overlapping, wavy, light green bands that create a sense of depth and movement. On the far left, there is a solid, vertical green bar. The overall aesthetic is clean, modern, and futuristic.

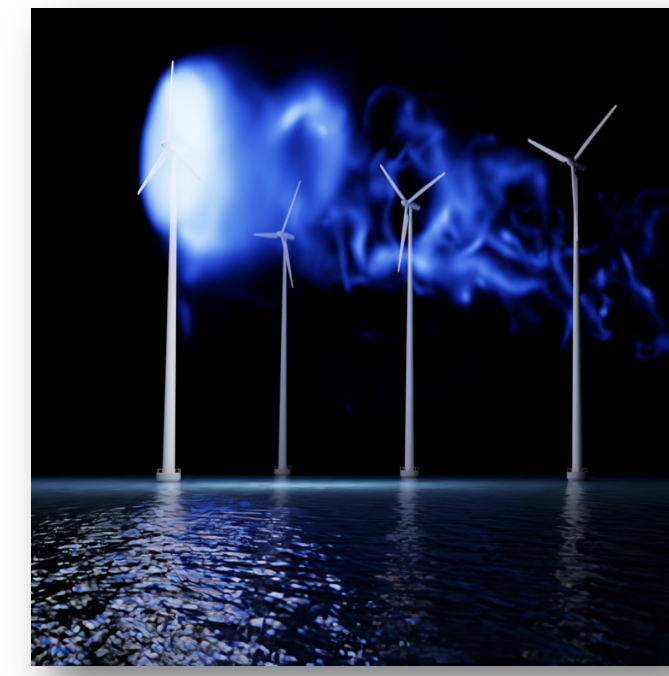
AI and Science

Workloads of the Modern Supercomputer

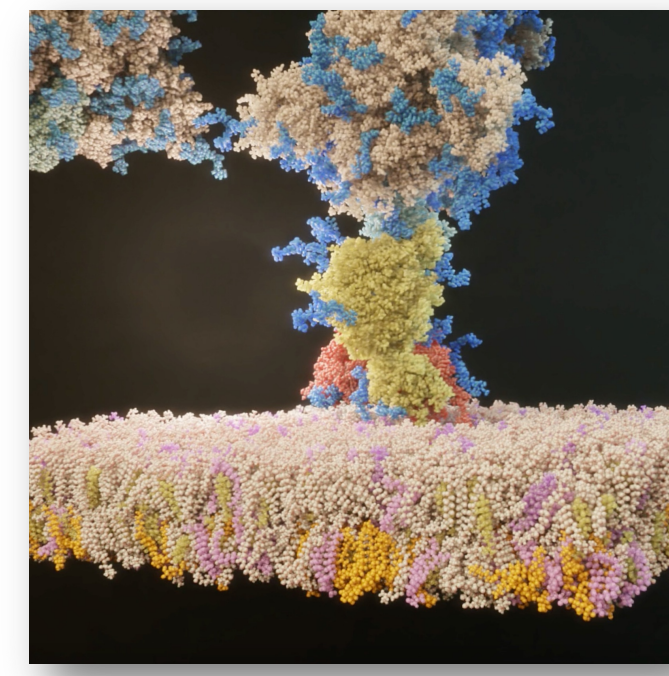
EDGE



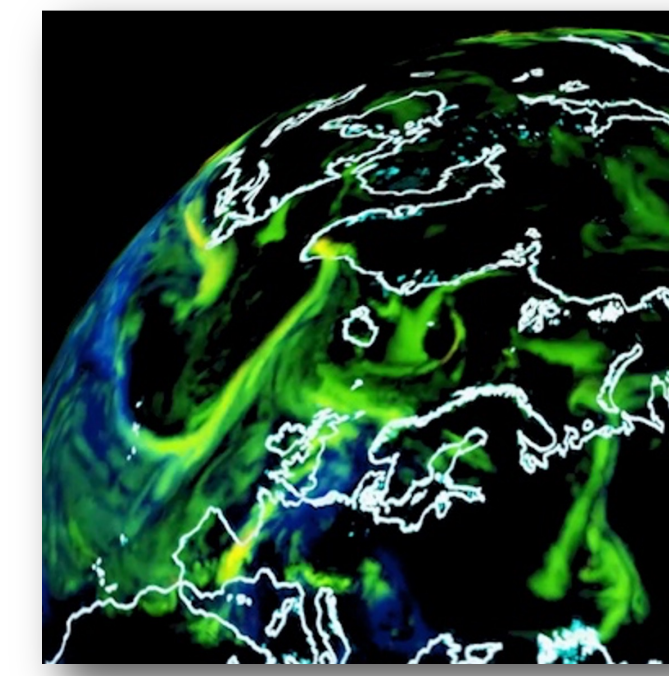
SIM + AI



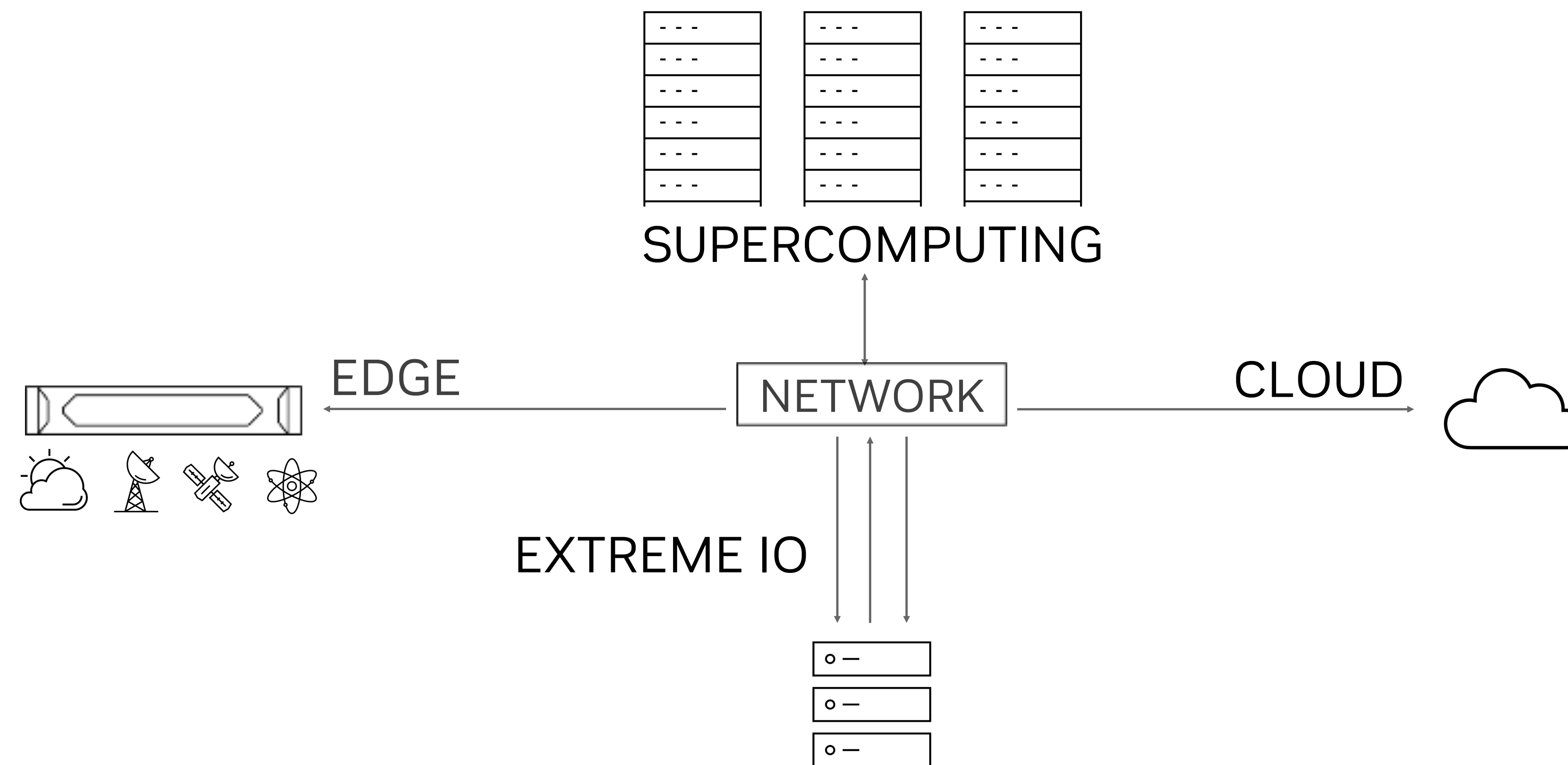
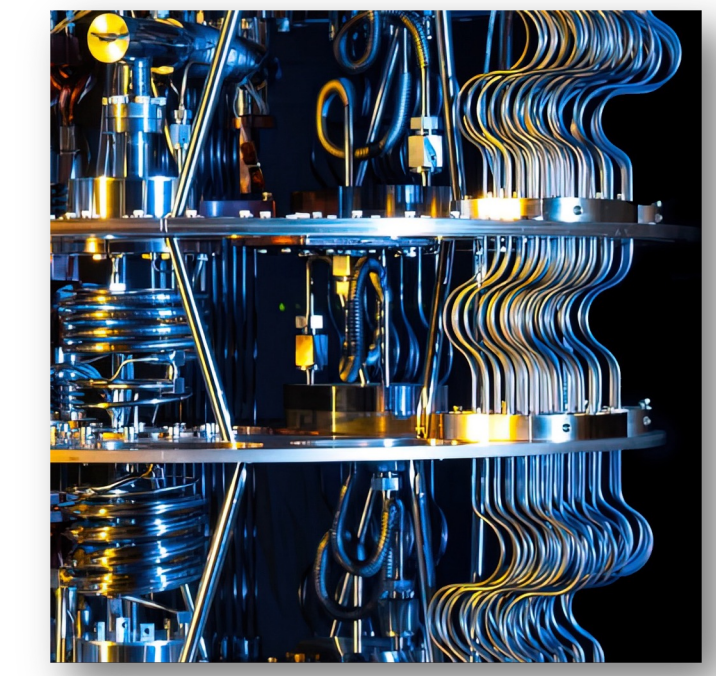
SIMULATION



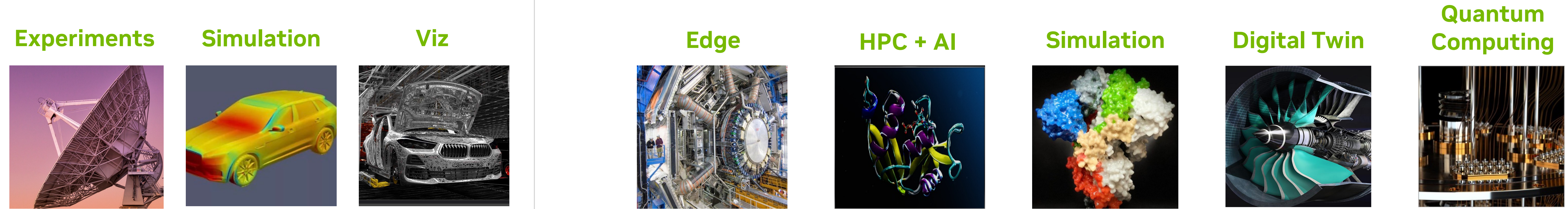
DIGITAL TWIN



QUANTUM COMPUTING



HPC Reinvented with AI

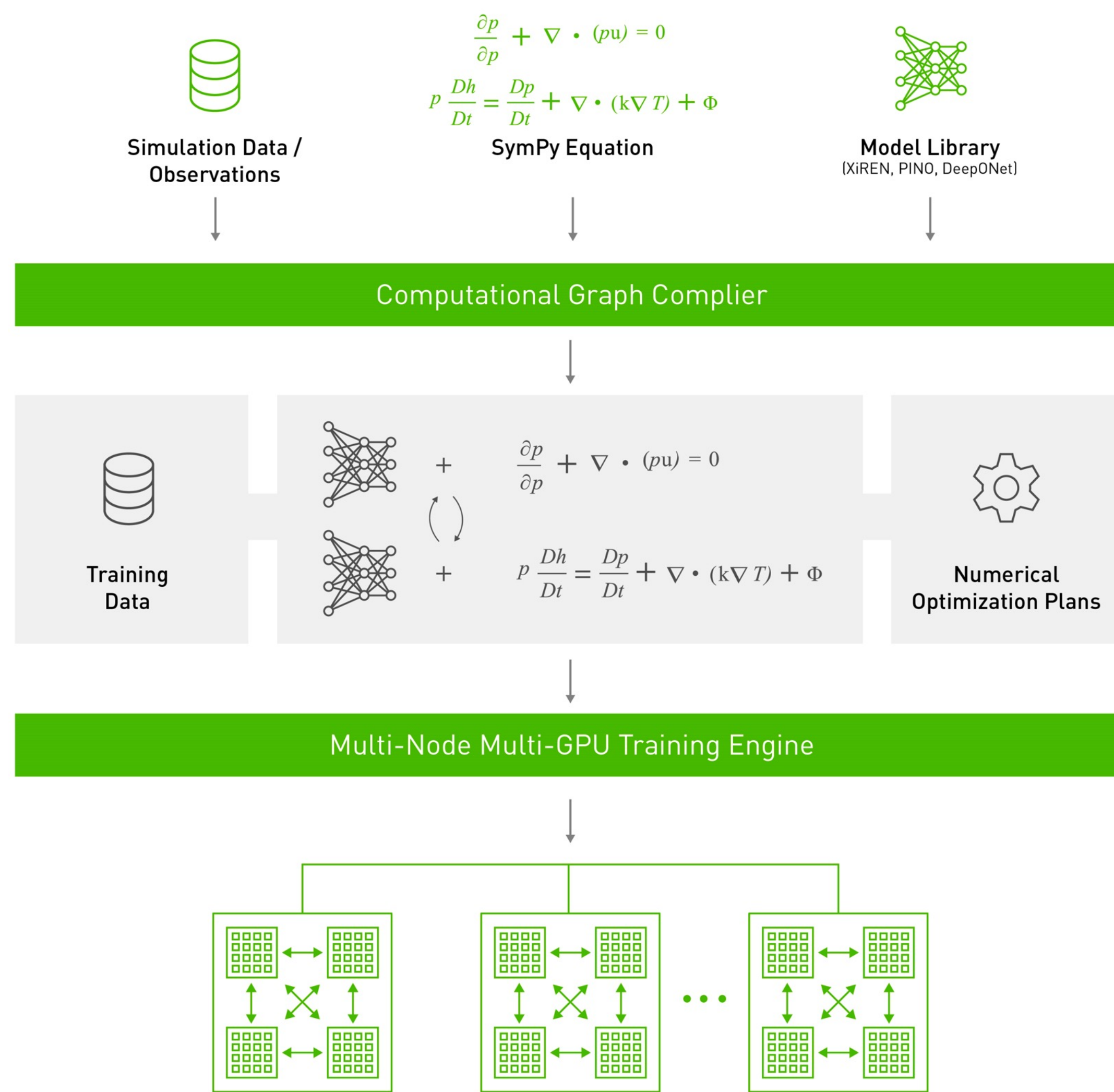


| FEATURE | PRE-EXASCALE | EMERGING POST EXA-SCALE |
|--------------------|---------------------------------------|---|
| USAGE | BATCH | INTERACTIVE & DISTRIBUTED |
| WORKLOAD | SINGLE SIMULATION/ENSEMBLES | SIMULATION/ENSEMBLES, AI TRAINING AND INFERENCE |
| EXPERIMENTS | OFFLINE DATA ANALYSIS FOR EXPERIMENTS | MIX OF REAL-TIME ANALYSIS, STEERING AND OFFLINE |
| DIGITAL TWINS | IN-SITU VISUALIZATION | <u>INTERACTIVE COMBINATION</u> OF SIMULATION AND OBSERVATIONAL DATA |
| QUANTUM COMPUTING | SIMULATION | PREPARING FOR A HYBRID MODEL |
| PROGRAMMING MODELS | FORTRAN, C++, MPI, OPENMP | STANDARD PARALLELISM SUPPORT IN FORTRAN, C++, MPI, OPENMP, OPENACC, PYTHON, JULIA, PYTORCH, JAX, TENSORFLOW |
| CLOUD | GRID | BURST CAPABILITIES, FASTER REFRESH CYCLE, ACCESS TO LATEST TECHNOLOGY AT SCALE |

NVIDIA Modulus

Open-Source Platform for Developing Physics-Based Machine Learning

Training Neural Networks Using Both Data And The Governing Equations



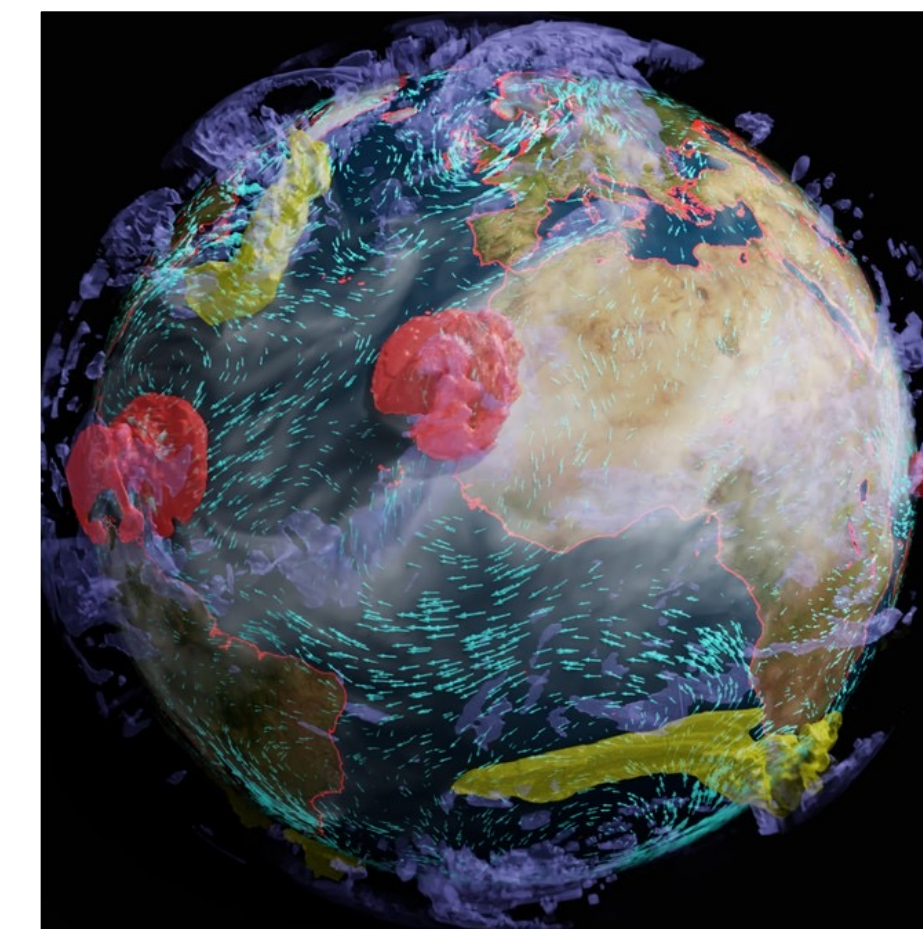
With generative AI using diffusion models, you can enhance engineering simulations and generate higher-fidelity data for scalable, responsive designs.

Advancing Scientific Discovery With Modulus

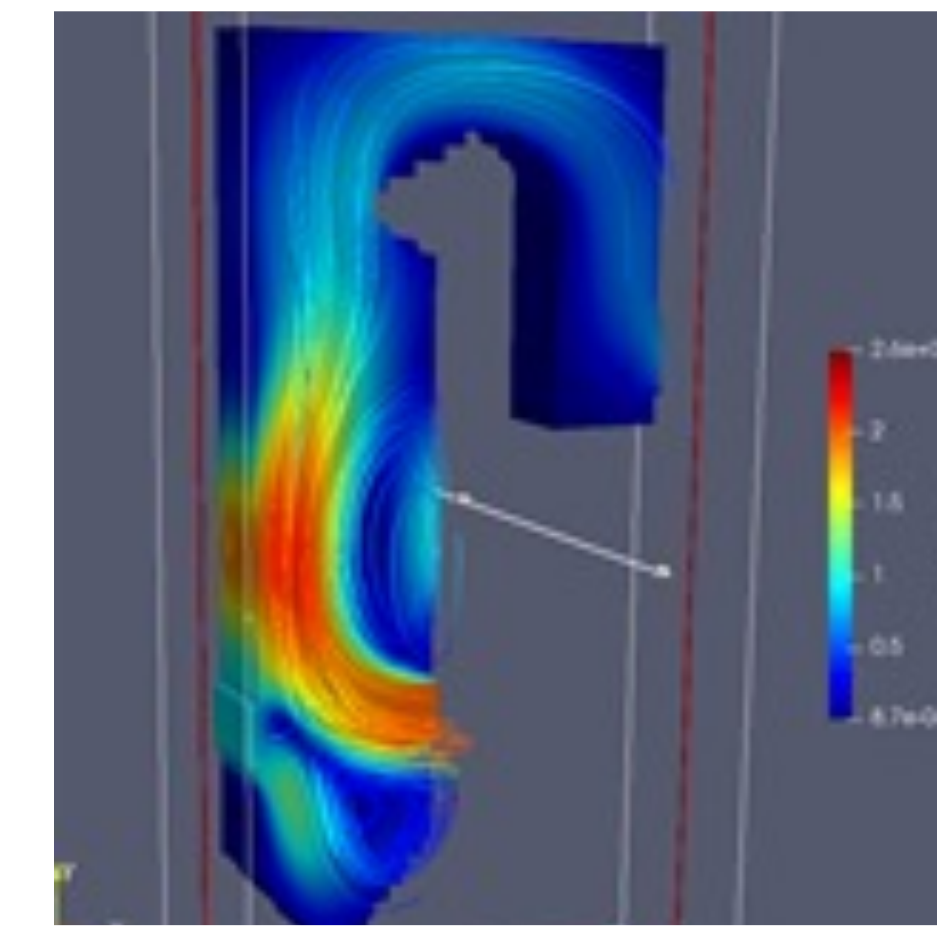
Renewable Energy
Siemens Gamesa: 4000X Faster wind turbine wake optimization



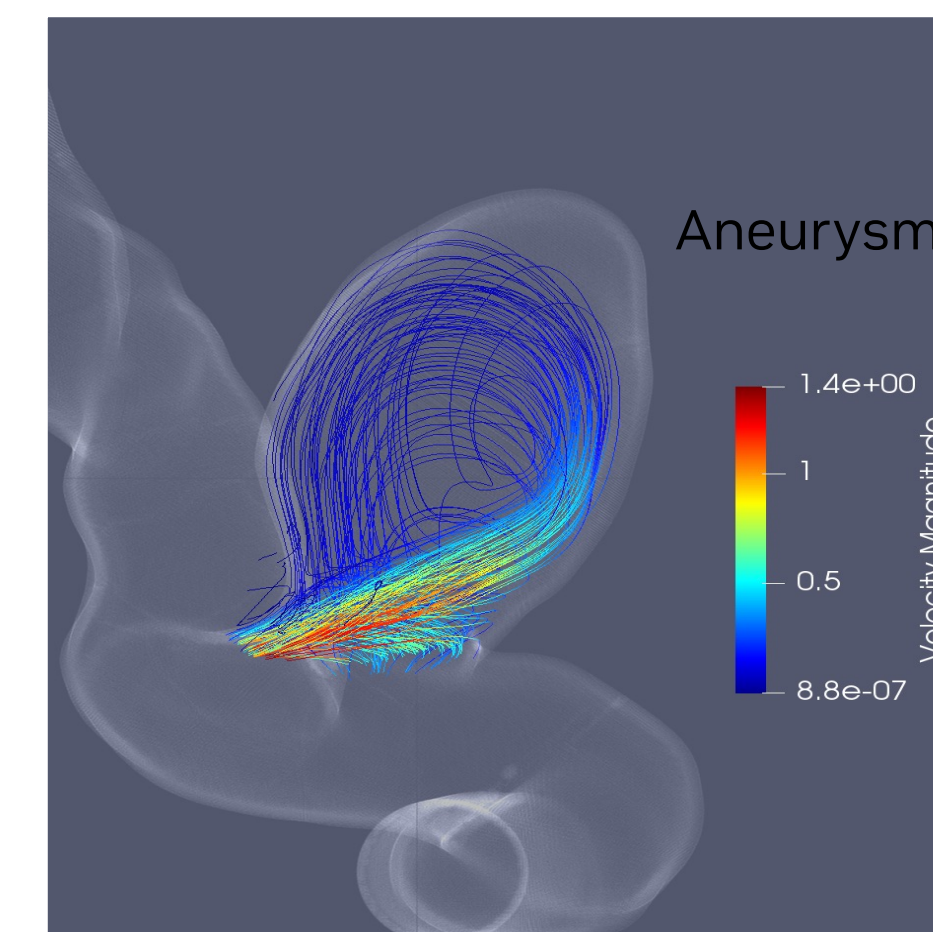
Climate Change
45,000X Faster extreme weather prediction with FourCastNet



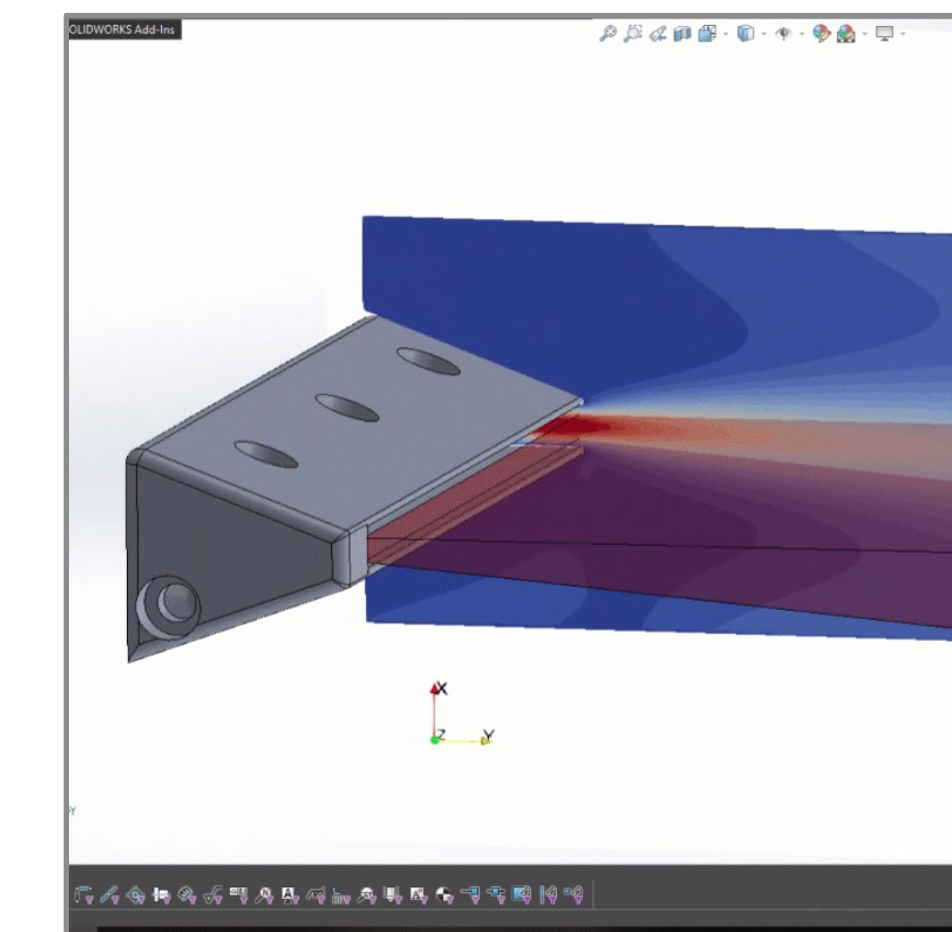
Industrial HPC
NETL: 10,000X Faster build of high-fidelity surrogate models



Healthcare
High-fidelity results faster for blood flow in inter-cranial aneurysm



Digital Twins
Kinetic Vision: Design optimization using parameterized models



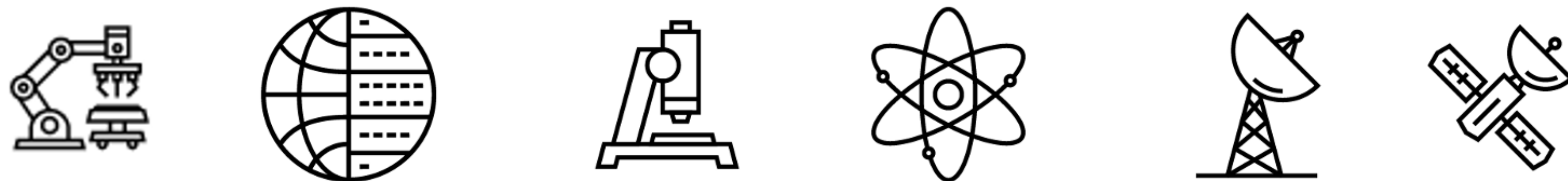
Science and Engineering Teaching Kit available now.

NVIDIA Holoscan

SDK for Building AI-Enabled Sensor Processing Applications

Holoscan SDK

From surgery to satellites



[Holohub](#)
Operators | Reference Apps

Model Zoo
NGC | MONAI

Holoscan SDK

Python (JAX | CuPy | RAPIDS)

C++

Graph Composer

Operators

IO

AI Inference

Viz

Custom

I/O Library

DPDK

[Rivermax](#)

AI Library

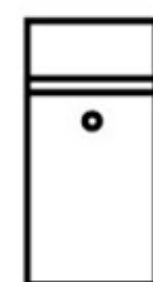
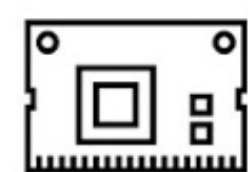
[TensorRT](#)

Triton

Visualization

Vulkan

Accelerated Libraries



Features

- C++ and Python APIs for **domain agnostic** sensor data processing workflows
- Scalable from IGX (ARM + GPU) to DGX (x86 + A100)
- Sample applications to jump-start ML/AI-enabled and Accelerated Computing sensor pipelines with [Holohub](#)
- AI Inference with pluggable backends such as TensorRT
- Apache 2 Licensed and Available on [GitHub](#)

Benefits

- Simplifies sensor I/O to GPU
- Simplifies the deployment of an AI model in a streaming pipeline
- Provides customizable, reusable, and flexible components to build and deploy GPU-accelerated algorithms
- Scale workloads with Holoscan Cloud Native

Building Digital Twin With NVIDIA Omniverse

Foundational platform components

NUCLEUS



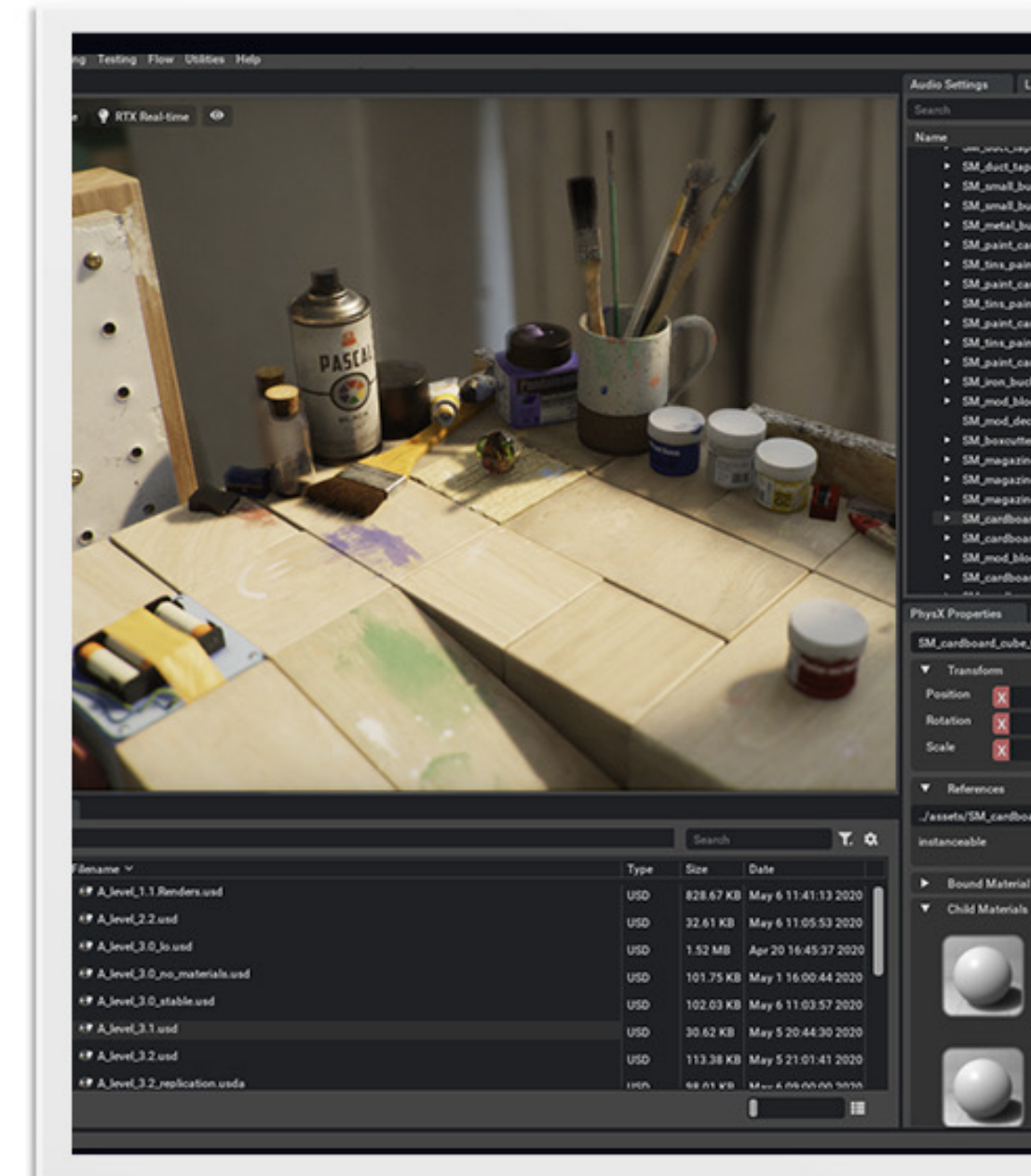
Source of truth
Database &
Collaboration
Engine

CONNECT



Coupling
Connectors

KIT



Application API
Python-USD Toolkit

SIMULATION



Virtual Actor
Rigid body dynamics
Real-time CFD, FEM

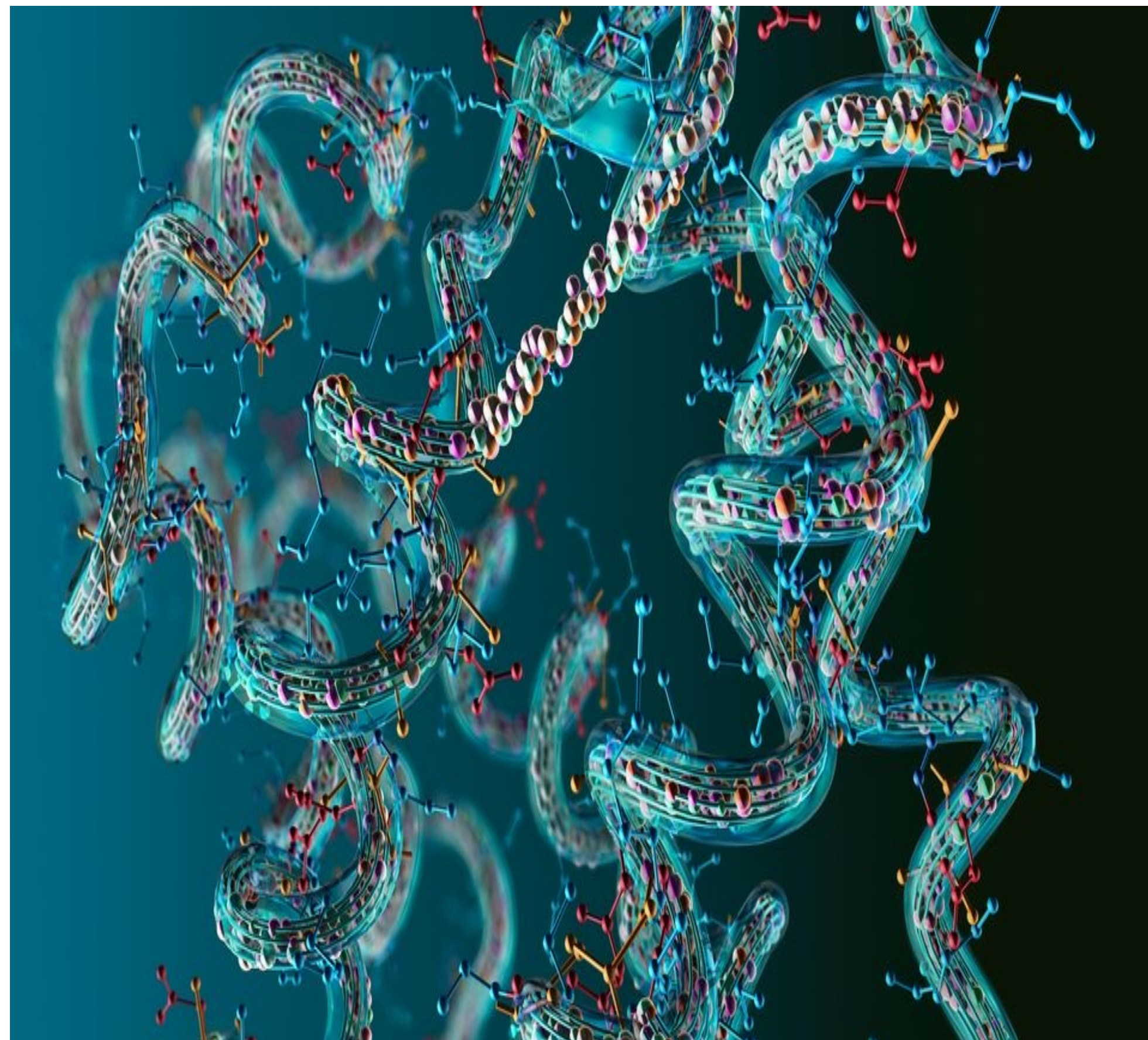
RTX RENDERER



Virtual Sensor
Visible spectrum
Experimental: RF, IR, ..

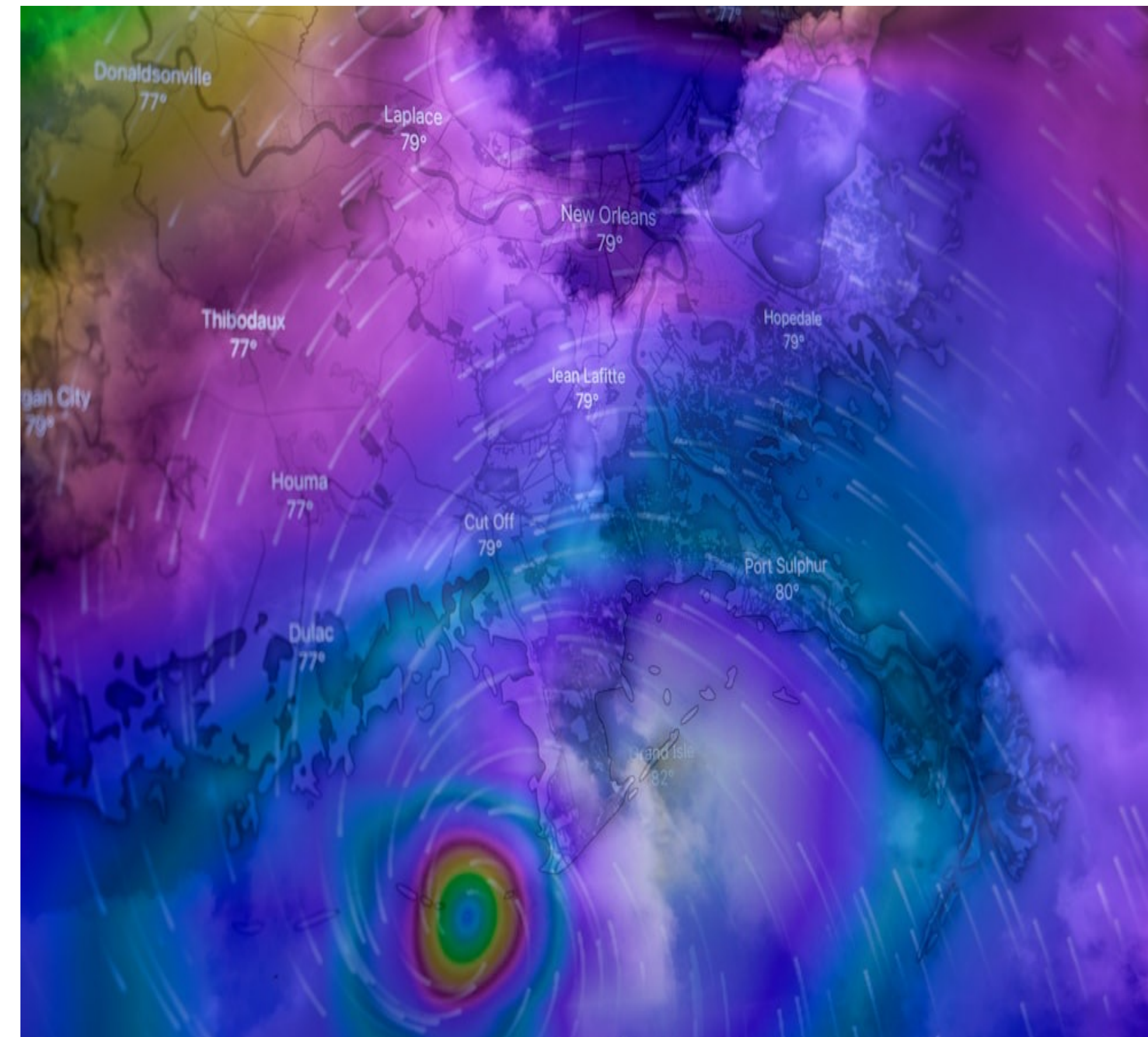
GenAI For Science Research and Discoveries

The Race for Foundation Models for Science is on



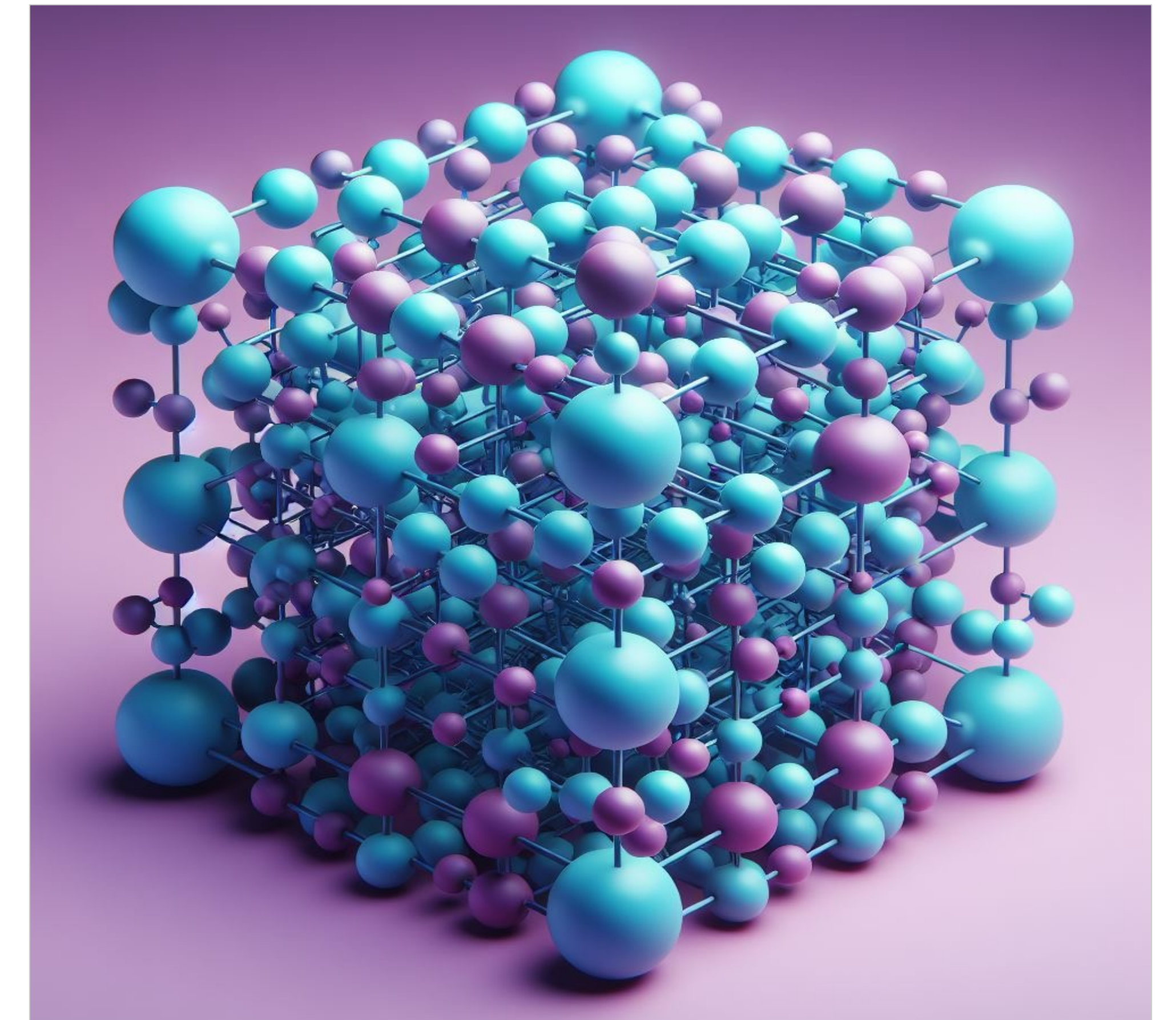
Biology: AlphaFold

2021



Climate : ClimaX/Stormer

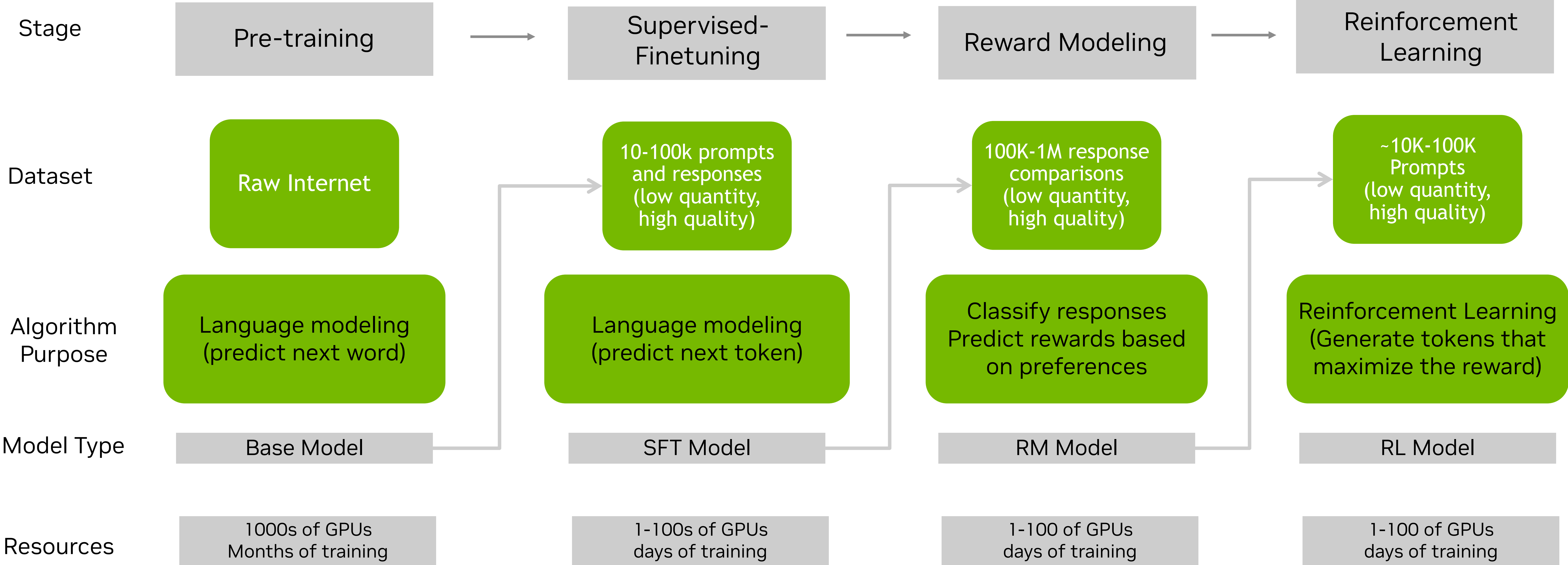
Jan 2023



Materials : MatterGen

Dec 2023

Building a Domain Specific Gen AI model is a Multistage Process

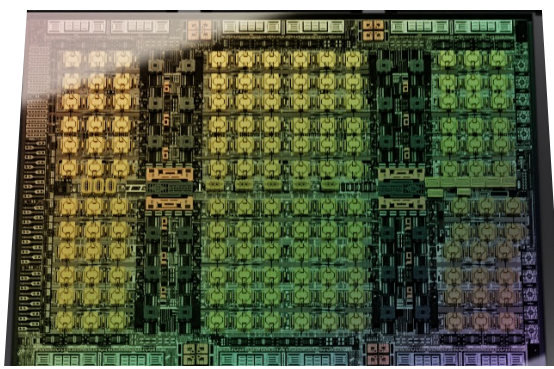


Reference : <https://www.youtube.com/watch?v=bZQun8Y4L2A>

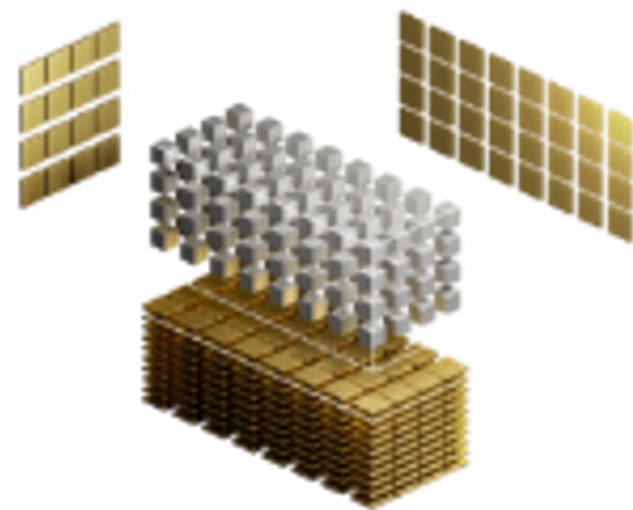
NVIDIA Hopper

NVIDIA Hopper

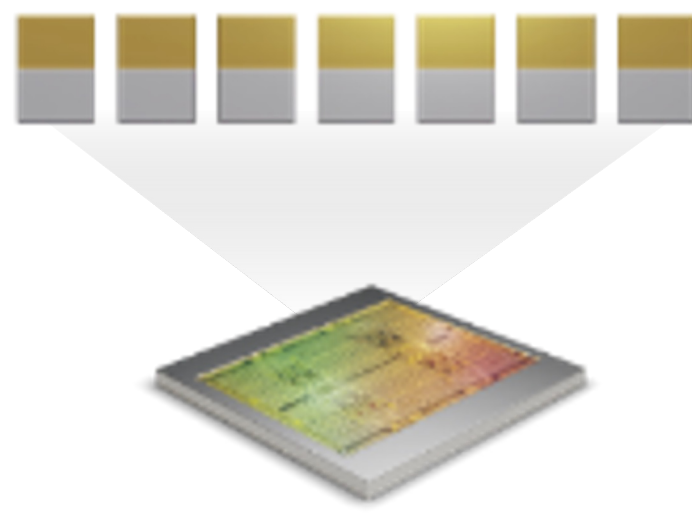
The new engine for the world's AI infrastructure



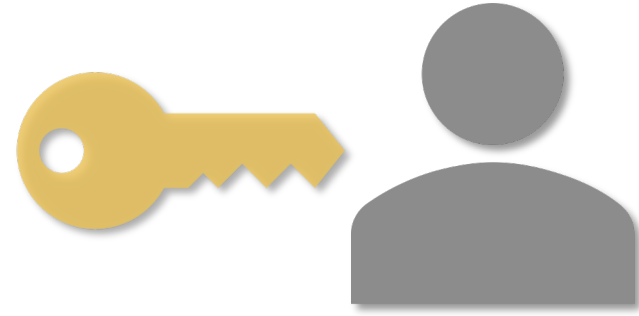
World's Most Advanced Chip



Transformer Engine



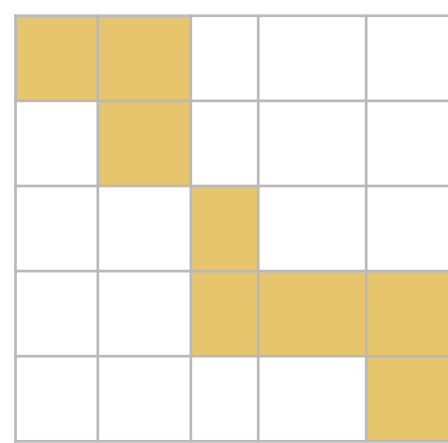
2nd Gen MIG



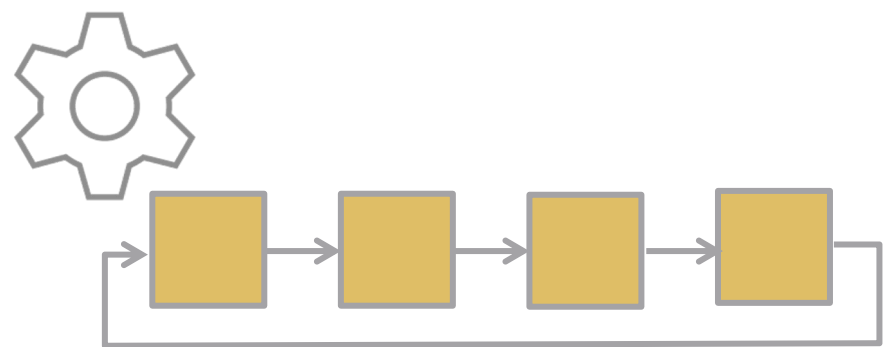
Confidential Computing



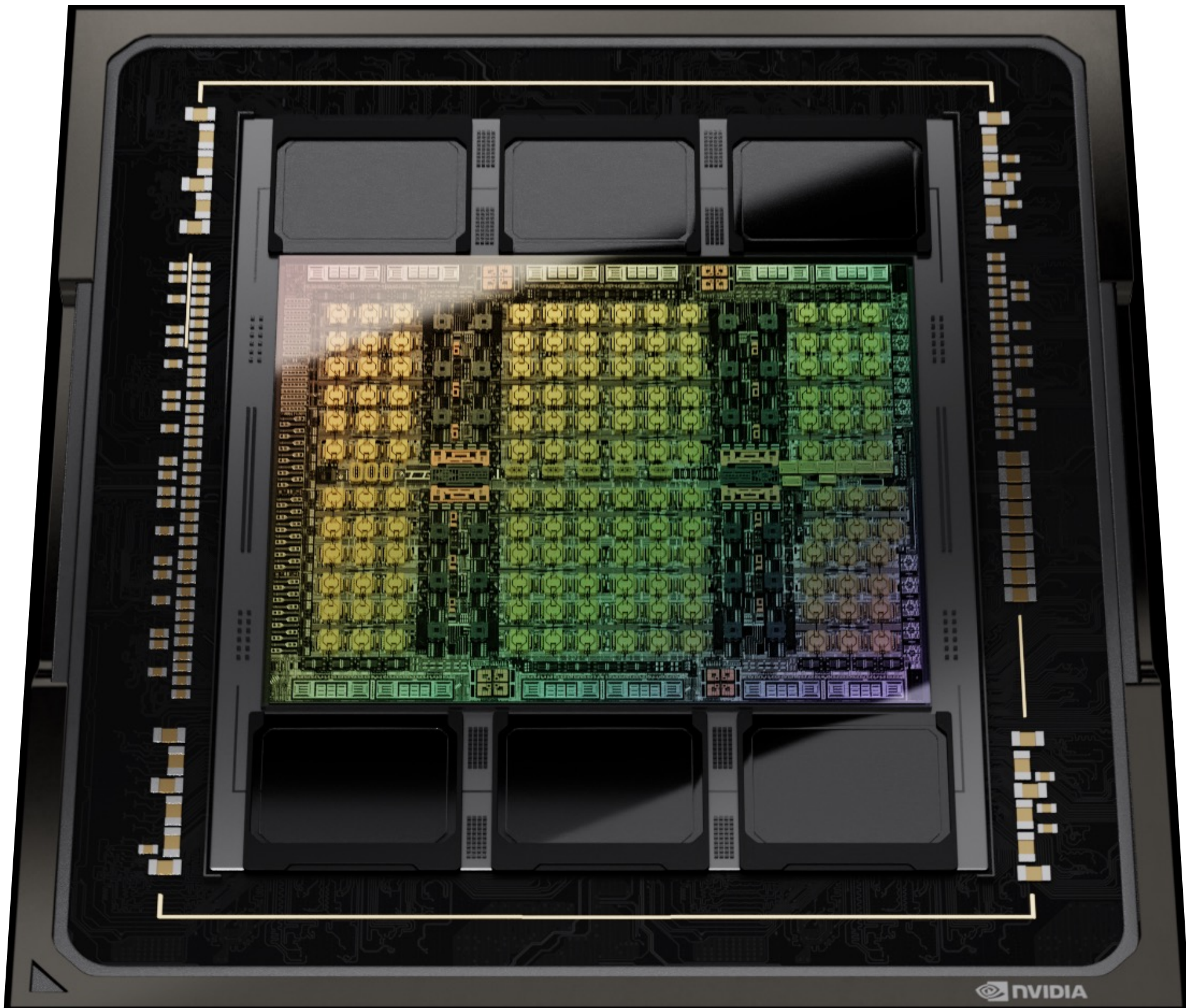
4th Gen NVLink



DPX Instructions



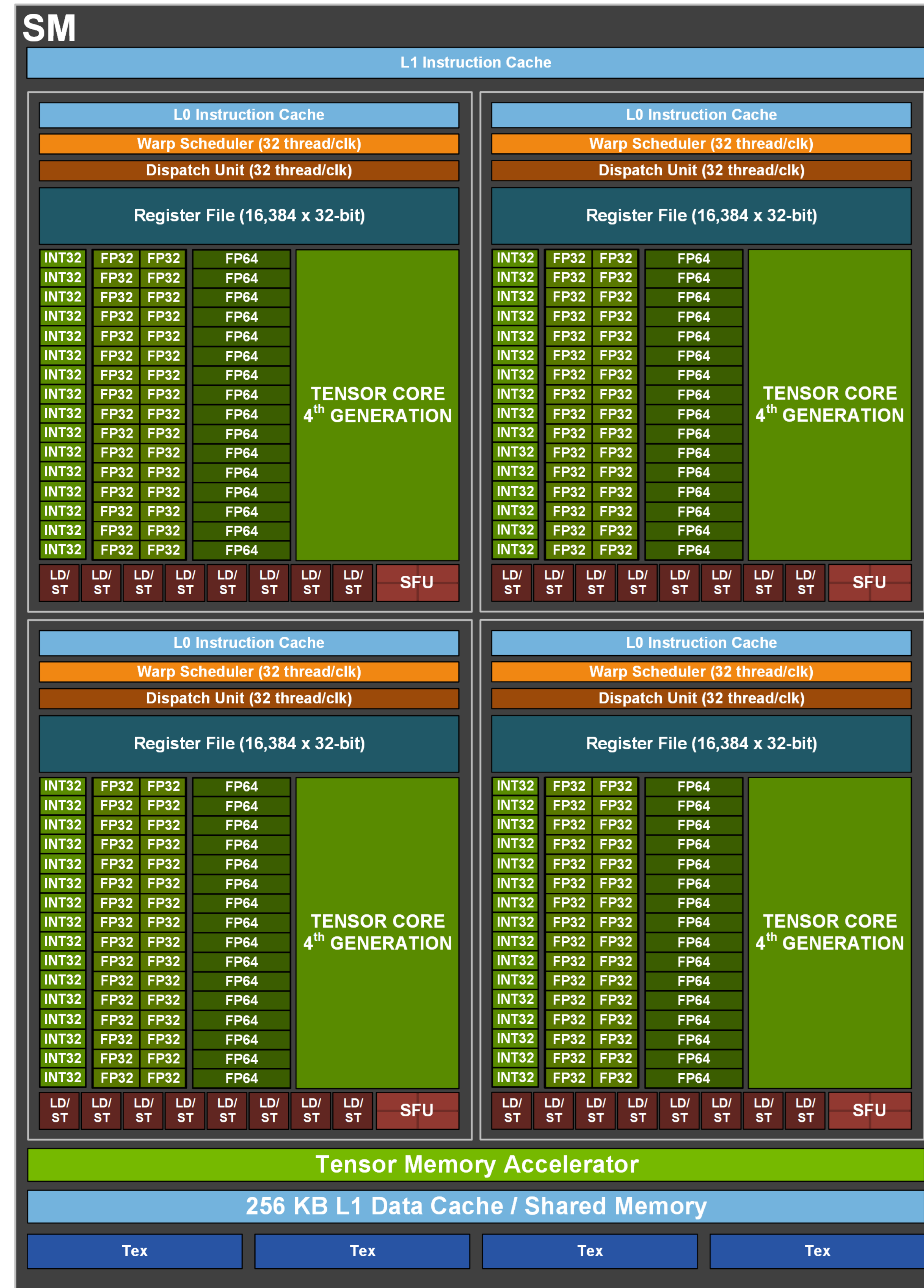
NVIDIA AI Enterprise Software Suite
*Redeemable NVIDIA AI Enterprise 5 Year Subscription**



Custom 4N TSMC Process | 80 billion transistors

**Included for H100 PCIe in mainstream systems*

Inside H100 SM Architecture



- New **4th Gen Tensor Core**
- **2x faster** clock-for-clock
- Supports **wide range** of storage and math formats
- New **FP8 format support**
- Accelerates **sparse** tensor arithmetic
- New **DPX** instruction set
- Improves programmer productivity:
 - New **Thread Block Clusters**
 - Turn locality into efficiency
 - New **Tensor Memory Accelerator**
 - Fully asynchronous data movement
- **256 KB** L1 cache / Shared Memory
- More **efficient data management** saves up to 30% operand delivery power

10 years of evolution in GPU hardware

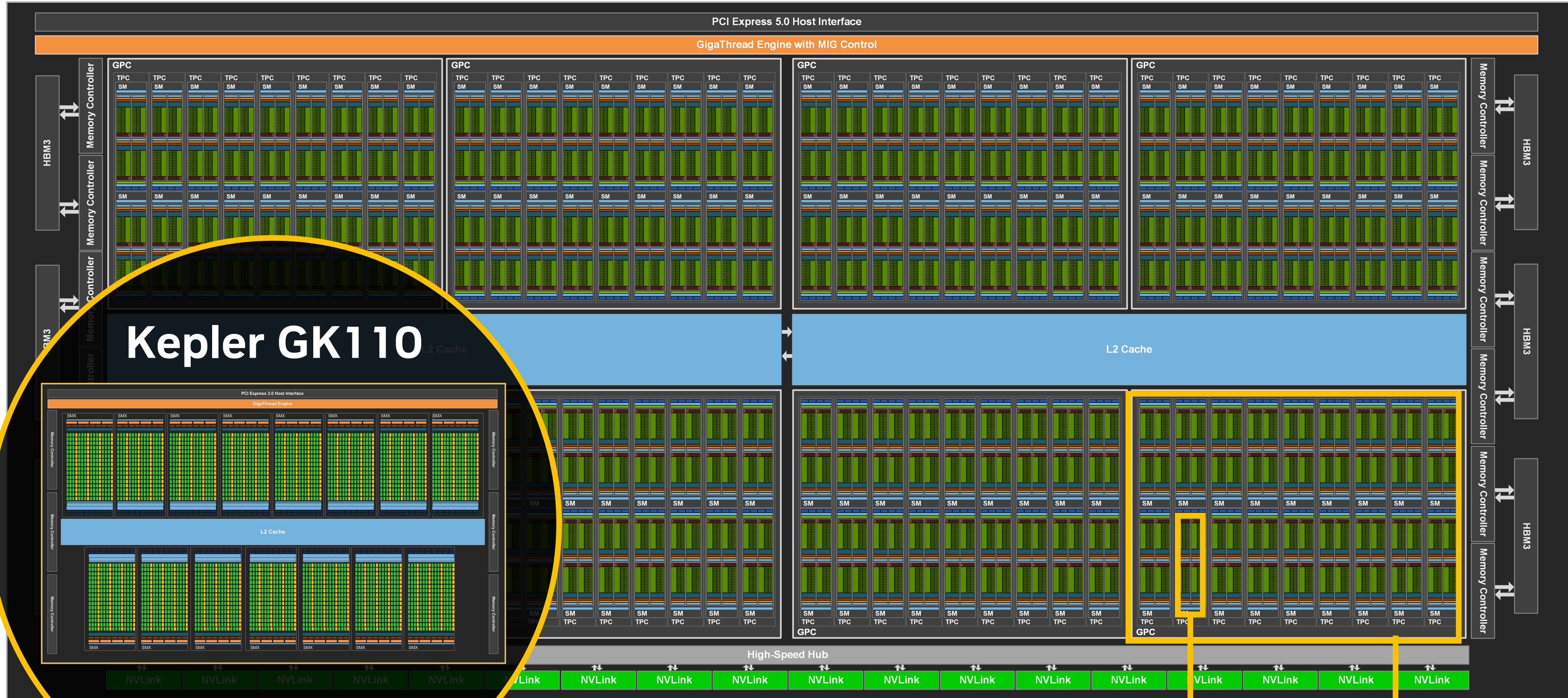
Kepler GK110 GPU (2012)

Hopper H100 GPU (2022)



SM
15 SMs

3.52 TFLOPS single
1.17 TFLOPS double



Kepler GK110

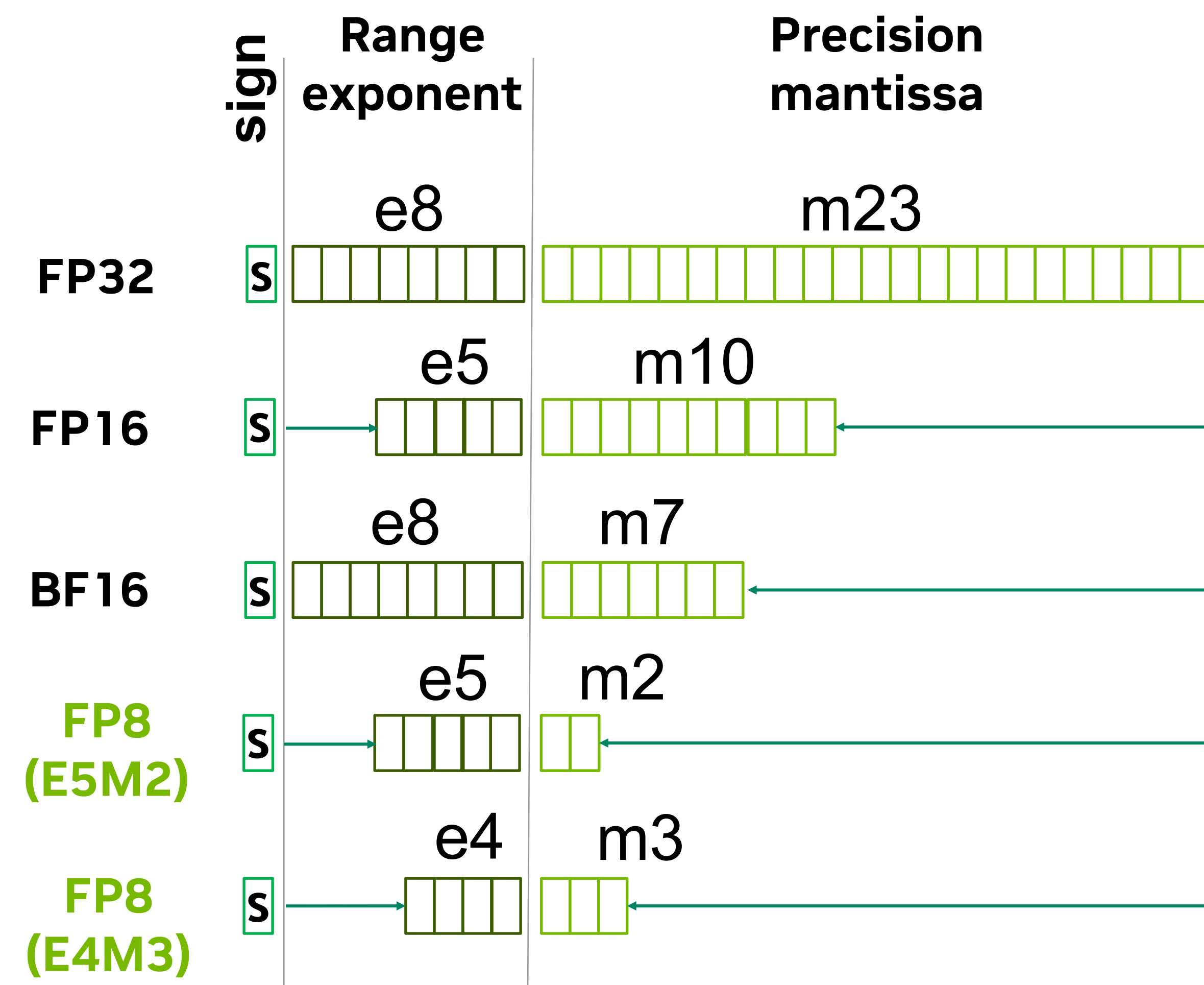
132 SMs

SM GPC

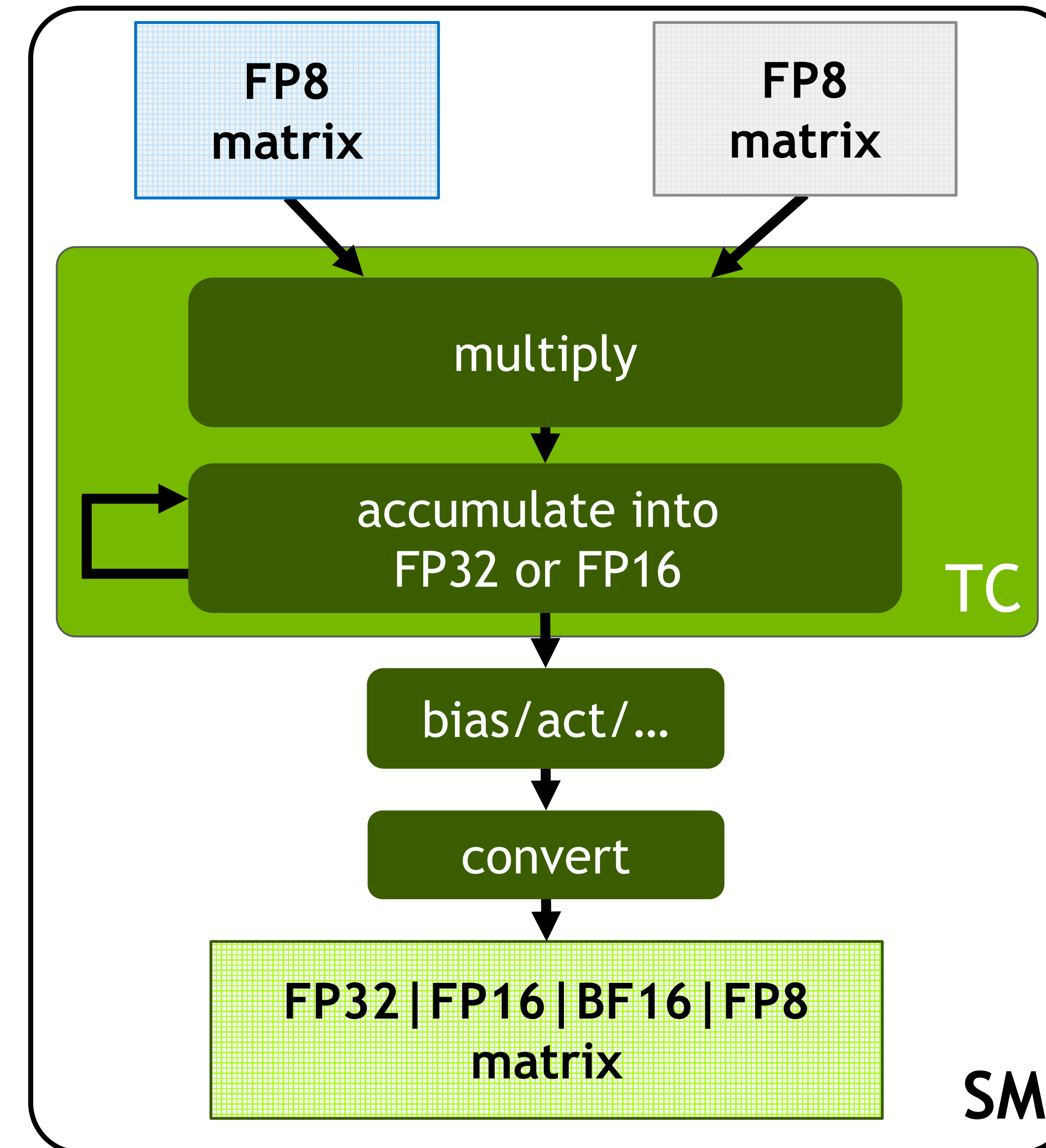
67 TFLOPS single [19x]
34 TFLOPS double [29x]
67 TFLOPS double with TC [57x]

Inside 8-bit Floating Point (FP8)

2x throughput & half footprint of FP16/BF16

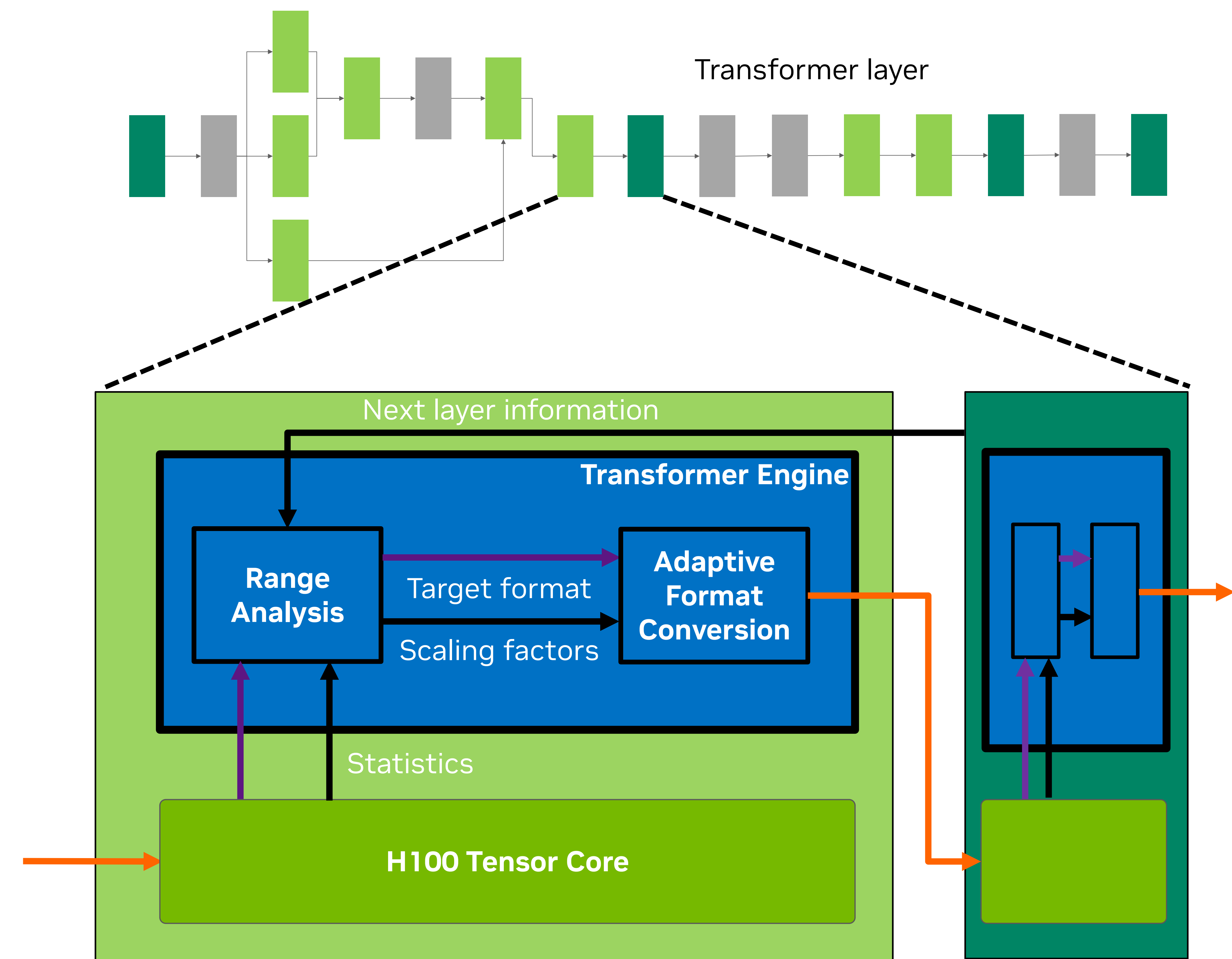


Allocate 1 bit to either range or precision



Support for multiple accumulator and output types

Transformed Engine



Optimal Transformer acceleration with Hopper Tensor Core

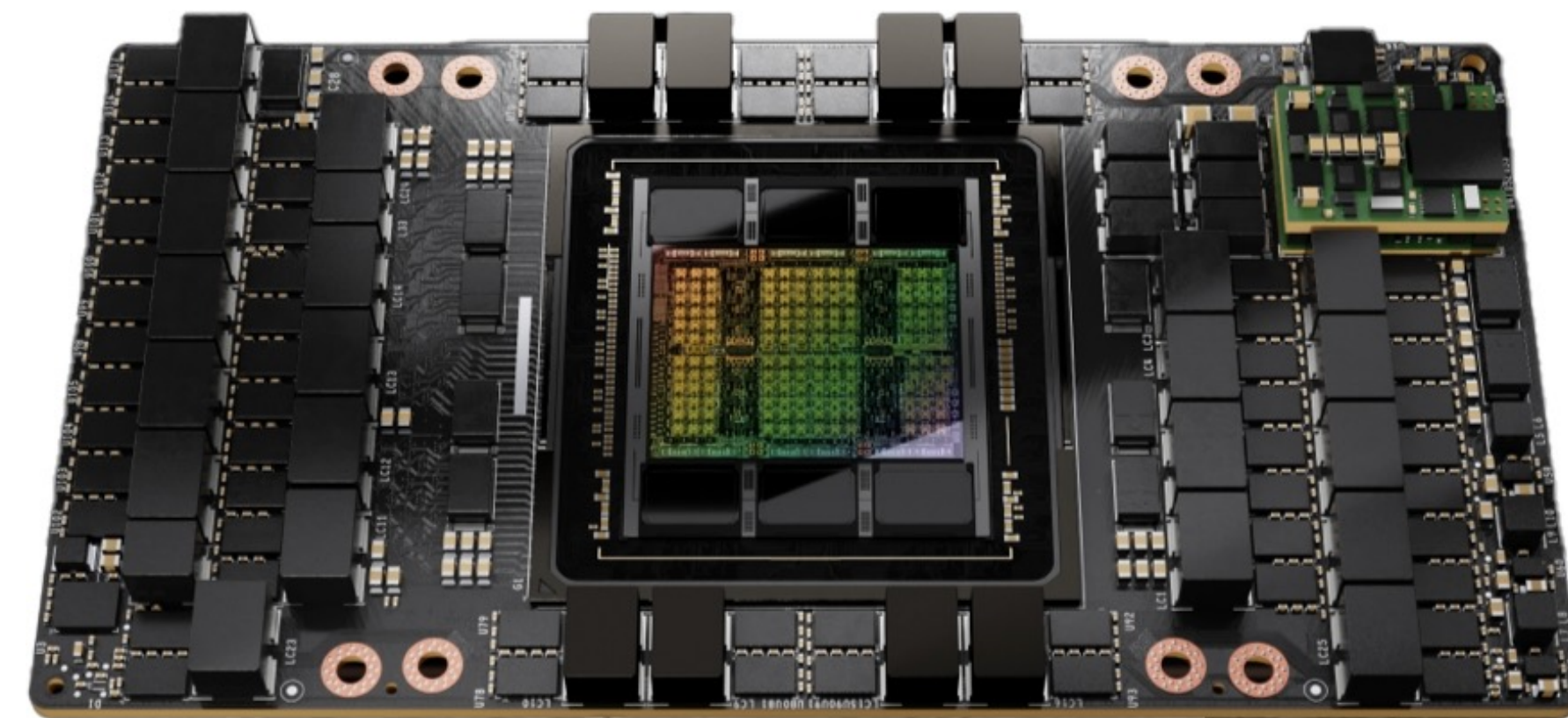
- **Transparent** to DL frameworks
- User can **enable/disable**
- Selectively **applies new FP8 format** for highest throughput
- **Monitors tensor statistics** and **dynamically adjusts range** to maintain accuracy

→ Adaptive precision → High precision → Auxiliary data

NVIDIA H100

Unprecedented Performance, Scalability, and Security for Every Data Center

H100



AI and HPC Performance

4PF FP8 (6X) | 2PF FP16 (3X) | 1PF TF32 (3X) | 67TF FP64 (3.4X)
3.35TB/s (1.5X), 80GB HBM3 memory

Transformer Engine

6X faster on largest transformer models

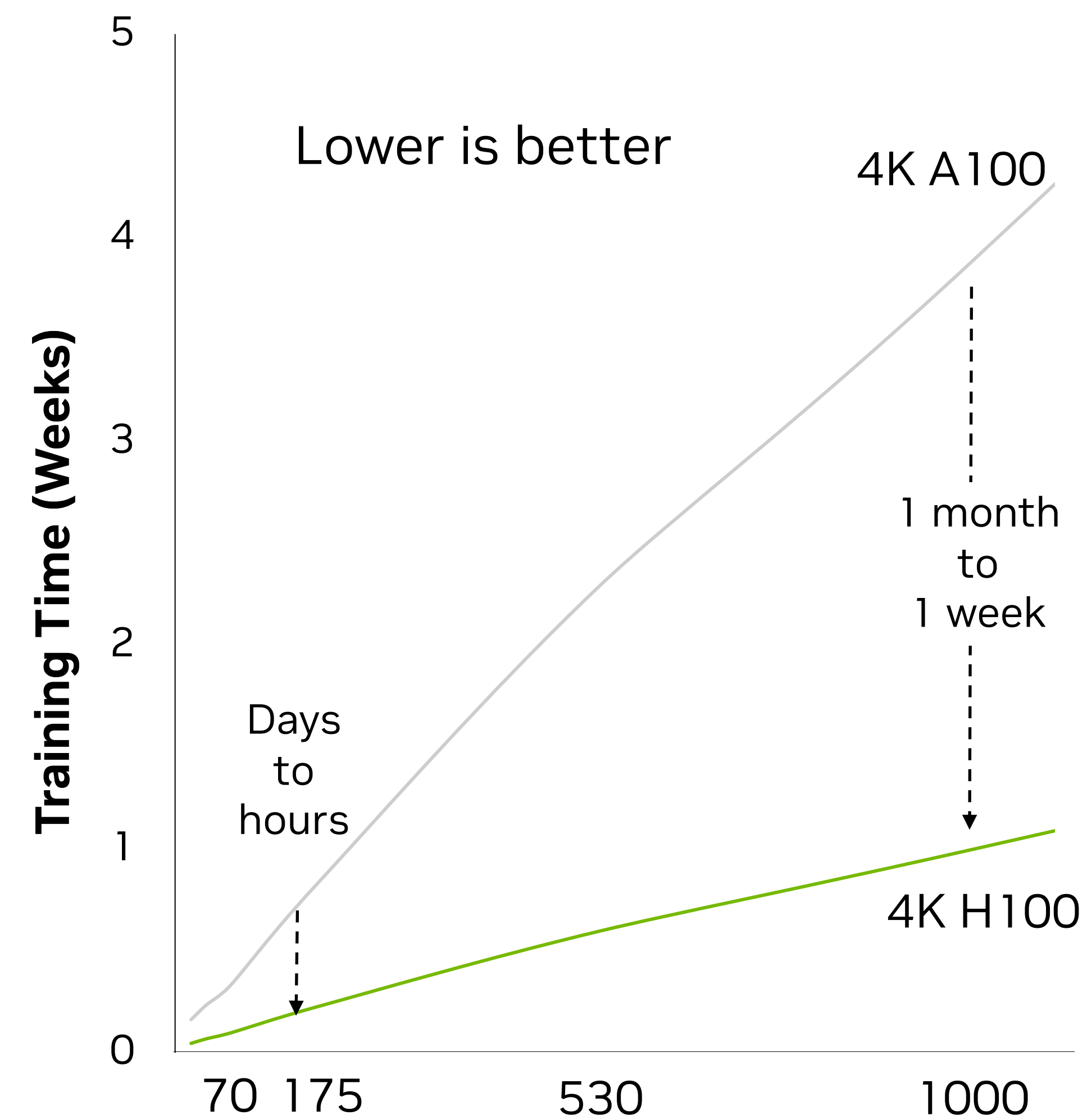
High Utilization Efficiency and Security

7 Fully isolated & secured instances, guaranteed QoS
2nd Gen MIG | Confidential Computing

Fast, Scalable Interconnect

900 GB/s GPU-2-GPU connectivity (1.5X)
up to 256 GPUs with NVLink Switch | 128GB/s PCI Gen5

Supercharged LLM Training



Time-to-Train by LLM Size
(Billion parameters)

Available Everywhere

All Major OEMs



All Major CSPs



NVIDIA L40S

Unparalleled AI and Graphics Performance for the Data Center

NVIDIA L40S GPU



Ada Lovelace Architecture Features

New Streaming Multiprocessor
4th-Gen Tensor Cores
3rd-Gen RT Cores

Gen-AI, LLM Training, & Inference

Transformer Engine - FP8
1.5 petaFLOPS Tensor Performance

OVX Reference Architecture

Powerful AI and Graphics Performance at Scale
NVIDIA AI Enterprise | Omniverse Enterprise
Powered by L40S GPUS

Most Powerful Universal Accelerator

Generative AI

Small Model Training | Fine-Tuning | Inference

Visual Computing

Omniverse | Rendering | 3D Graphics | Video

LLM Fine Tuning

8 hours

Llama 2-70B 1 Billion Tokens¹

LLM 1st Token Latency

<30 ms

Llama 2-13B Inference 225/20²

Small LLM Training

3 days

Llama 2-7B 100 Billion Tokens³

Image Gen AI

>82

Images per Minute⁴

LLM E2E Latency

<0.5 s

Llama 2-13B Inference 225/20⁵

Full Video Pipeline

184

AV1 Encode Streams⁶

L40S Performance Highlights

Broad Availability

Data Center

ADVANTECH ASUS

Atos CISCO

DELL Technologies FUJITSU

GIGABYTE H3C

Hewlett Packard Enterprise inspur



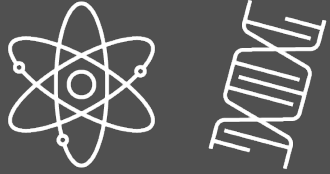





Lenovo QCT

SUPERMICR X FUSION

Cloud

aws ORACLE CoreWeave VULTR

GPU Portfolio: NVIDIA Hopper™ and Ada Lovelace Architectures

| | GPU | Networking |  DL Training & DA |  DL Inference |  HPC / AI |  Omniverse / Render Farms |  Virtual Workstation |  Virtual Desktop (VDI) |  AI Video |  Far Edge Acceleration |
|------------------------------------|--------------|------------|--|--|--|--|---|---|--|---|
| Compute | GH200 | QTM2 SPTM4 | ● | ● | ● | | | | | |
| | H100 | QTM2 SPTM4 | ● | ● | ● | | | | | |
| Graphics / Compute | L40S | QTM1 SPTM3 | ● | ● | ● | ● | ● | | ● | |
| | L40 | SPTM3 | | | | ● | ● | | | |
| Small Form Factor Compute/Graphics | L4 | SPTM3 | | ● | | ● | ● | ● | ● | ● |

 Best
 Better
 Good

Price-performance comparison relative across each entire workload column. This chart should be used in conjunction with measured data for targeted workloads.

QTM1 Quantum-1 IB switch plus BlueField2 DPUs or ConnectX-6/6 DX SmartNICs

QTM2 Quantum-2 IB switch plus BlueField3 DPUs or ConnectX-7 SmartNICs

SPTM3 Spectrum-3 ethernet switch plus Bluefield2 DPUs or ConnectX-6 /6 Dx SmartNICs

SPTM4 Spectrum-4 ethernet switch plus Bluefield3 DPUs or ConnectX-7 SmartNICs

NVIDIA Grace & Grace Hopper

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 to 1TB/s of memory bandwidth

Fast and Flexible CPU IO

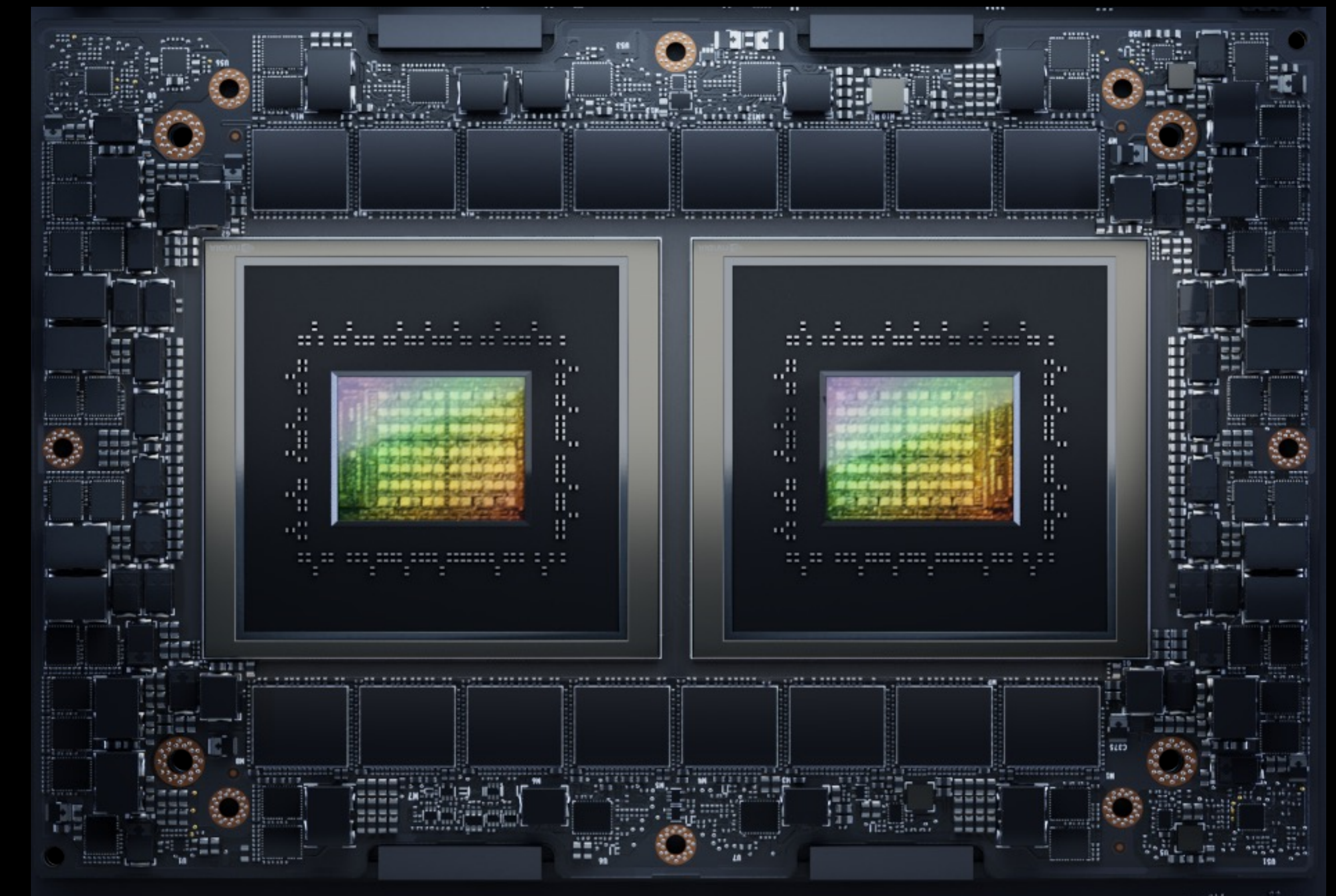
Up to 8x PCIe Gen5 x16 interface. PCIe Gen 5 up to 128GB/s
2X more bandwidth compared to PCIe Gen 4

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

Data Center CPU Landscape

The Future is Data Center Power Limited

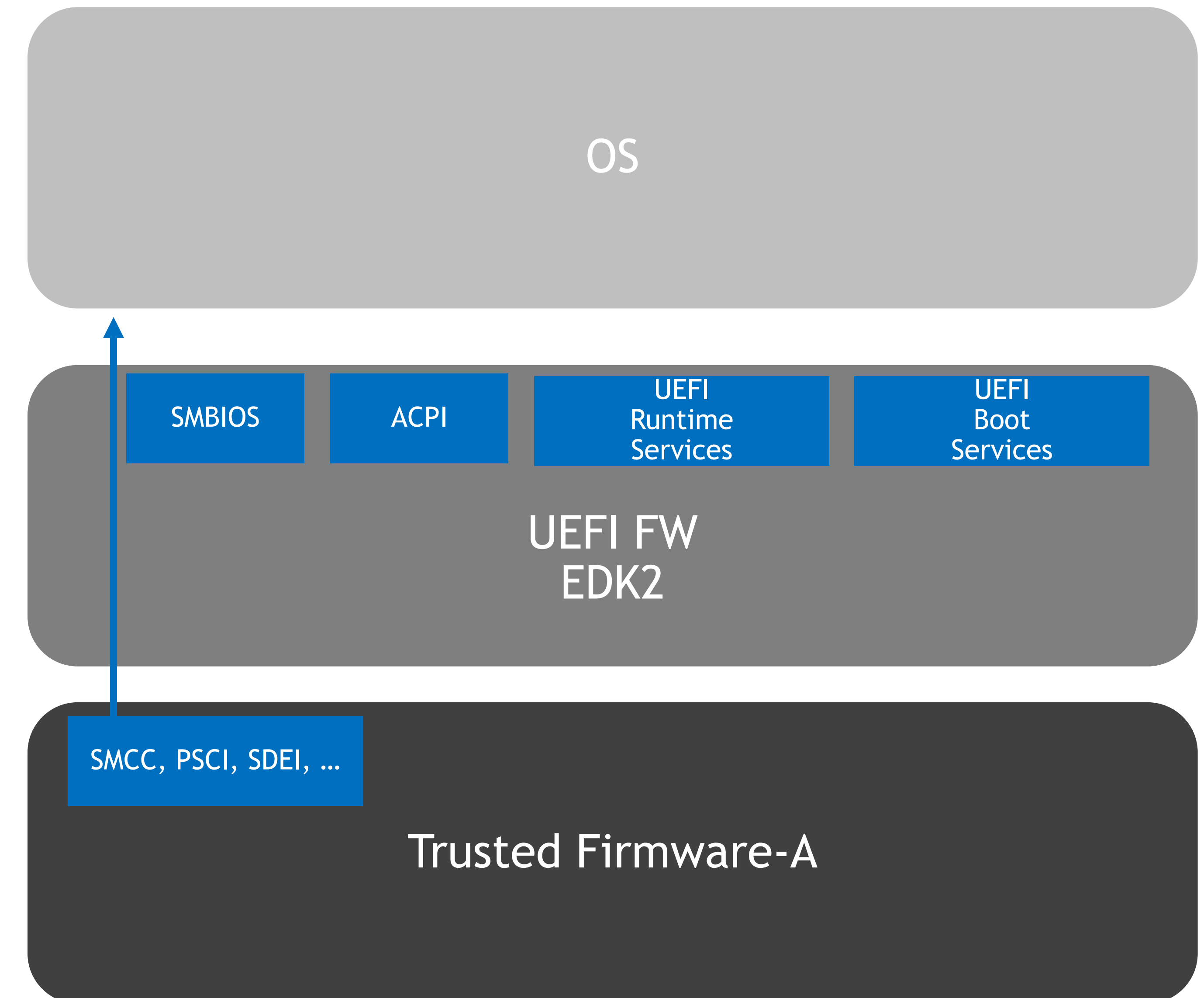


| Goal | Low-Cost Cores (Availability drives TCO) | Balance TCO and Performance | High-Performance (Perf. is the numerator on TCO) |
|--------------------------|--|---|---|
| Example Workloads | Proxies, Load Balancers, Web Frontend, Service Endpoints | DB Servers, Analytics, Video & Image Processing, Application Servers, CI/CD, game servers | Simulation, high-end analytics, traditional ML. |
| Design Point | Cores & vCPUs per watt | Balanced fabric, integer, memory BW, and FP performance and perf. per watt. | High memory BW, high FP performance |
| Criteria | Cost of Persistence: rent cores for IaaS 24x7 at low cost | Cost of Peak compute: best TCO to achieve a defined maximum goal. | Cost of Job: best overall compute throughput per \$. |

Grace

Server Base System Architecture (SBSA) Base Boot Requirements (BBR)

- Standard set of platform requirements and recommendations to enable off-the-shelf OS support
- SBBR recipe support from BBR
- Allows OS and system SW to expect consistency across different SOCs
 - Standard Private Peripheral Interrupt (PPI) assignments
 - Standard UART
 - PCIe – ECAM, ITS for MSI(-X)

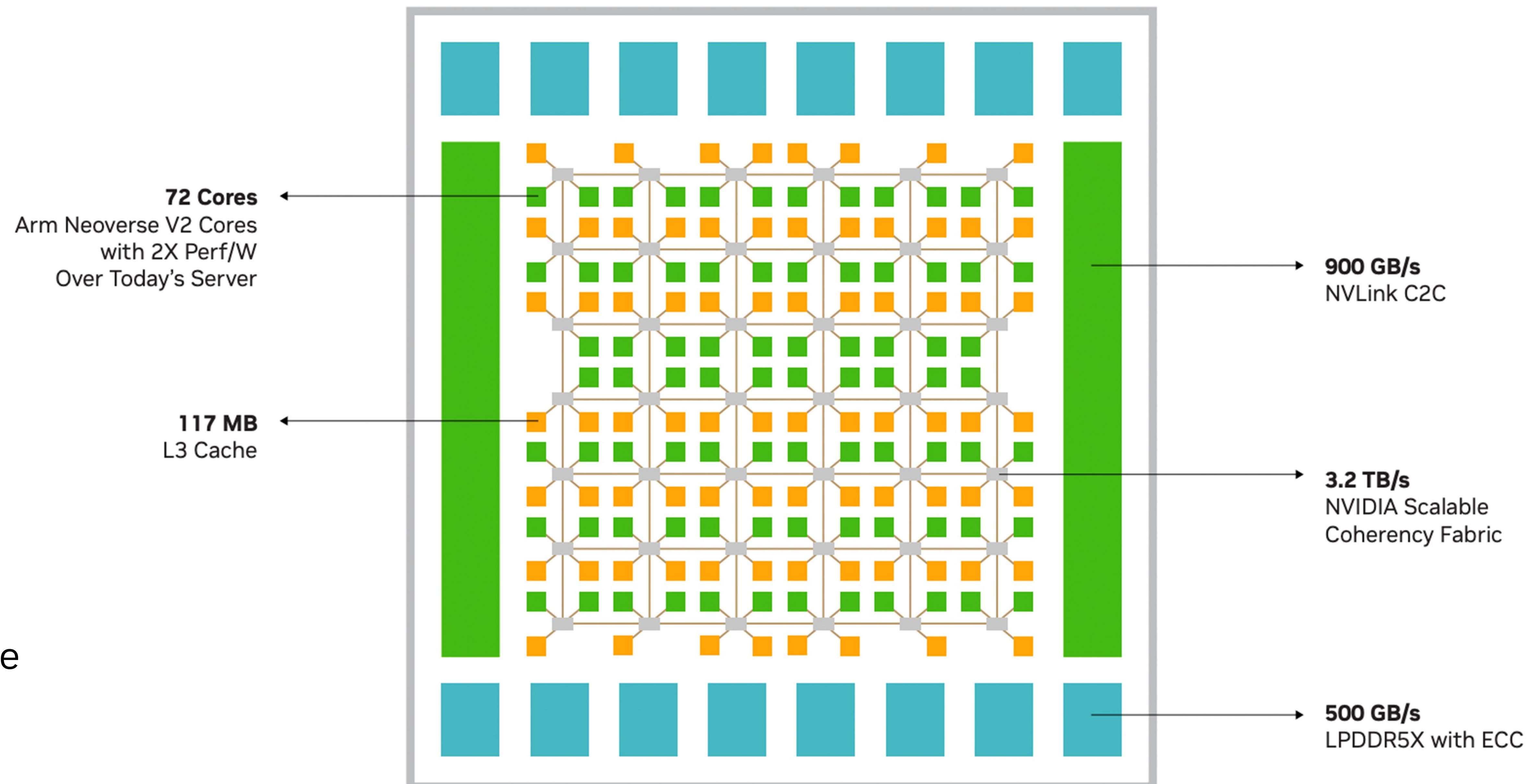


arm
SystemReady

GRACE IS A COMPUTE & DATA MOVEMENT ARCHITECTURE

NVIDIA Scalable Coherency Fabric (SCF) and distributed cache design

- Up to 512GB of LPDDR5X memory
 - **32 channels**
 - **Up to 546 GB/s of memory BW**
 - **Competitive power/perf**
- NVIDIA Scalable Coherency Fabric
 - 3,225.6 GB/s bi-section BW
 - **117MB of distributed L3 cache**
 - Scalable to 72+ cores per die
 - Background data movement via Cache Switch Network
- Supports up to 4-die coherency over Coherent NVLINK

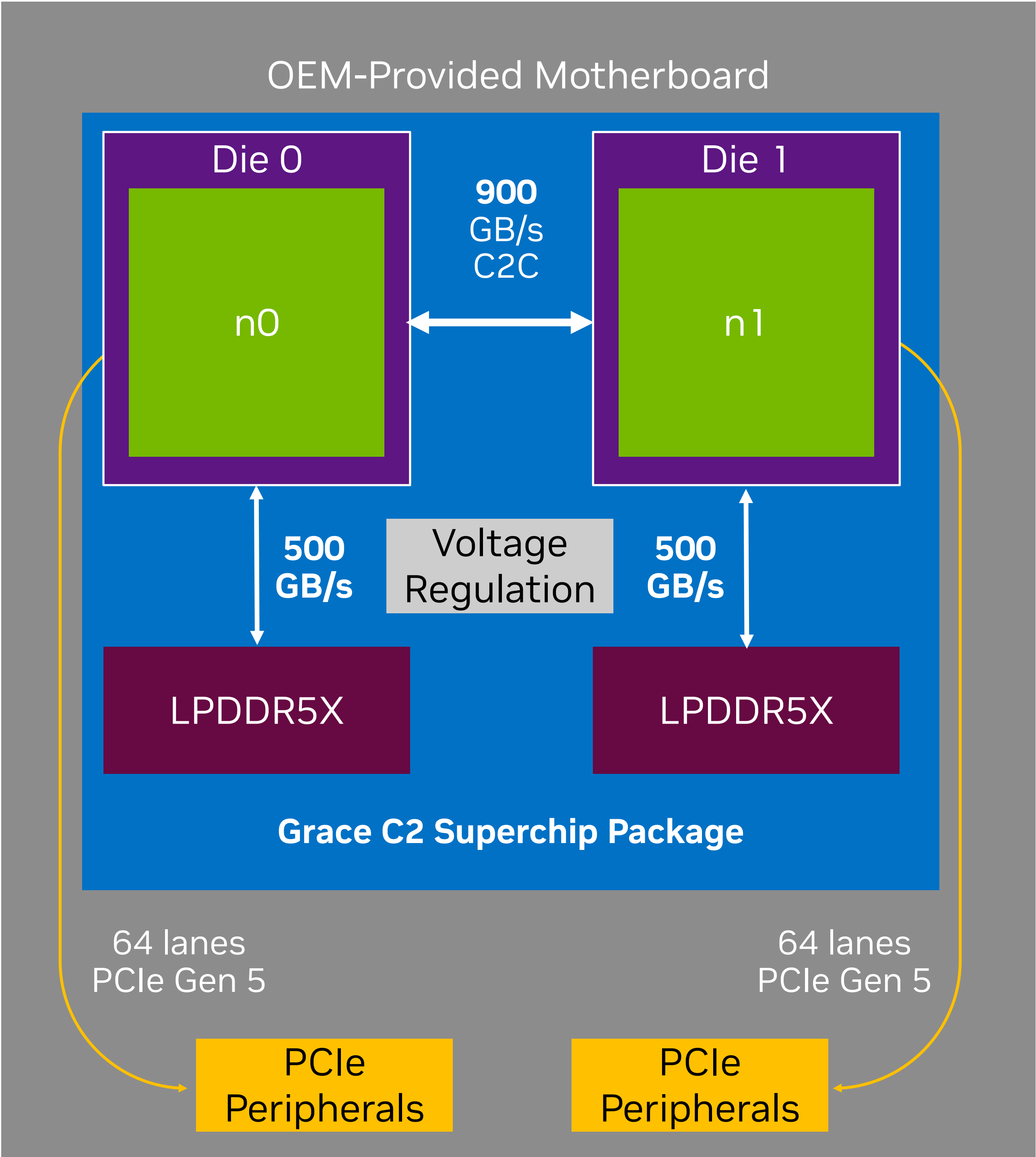


Example possible fabric topology for illustrative purposes

Grace Simplifies System Design and Workload Optimization

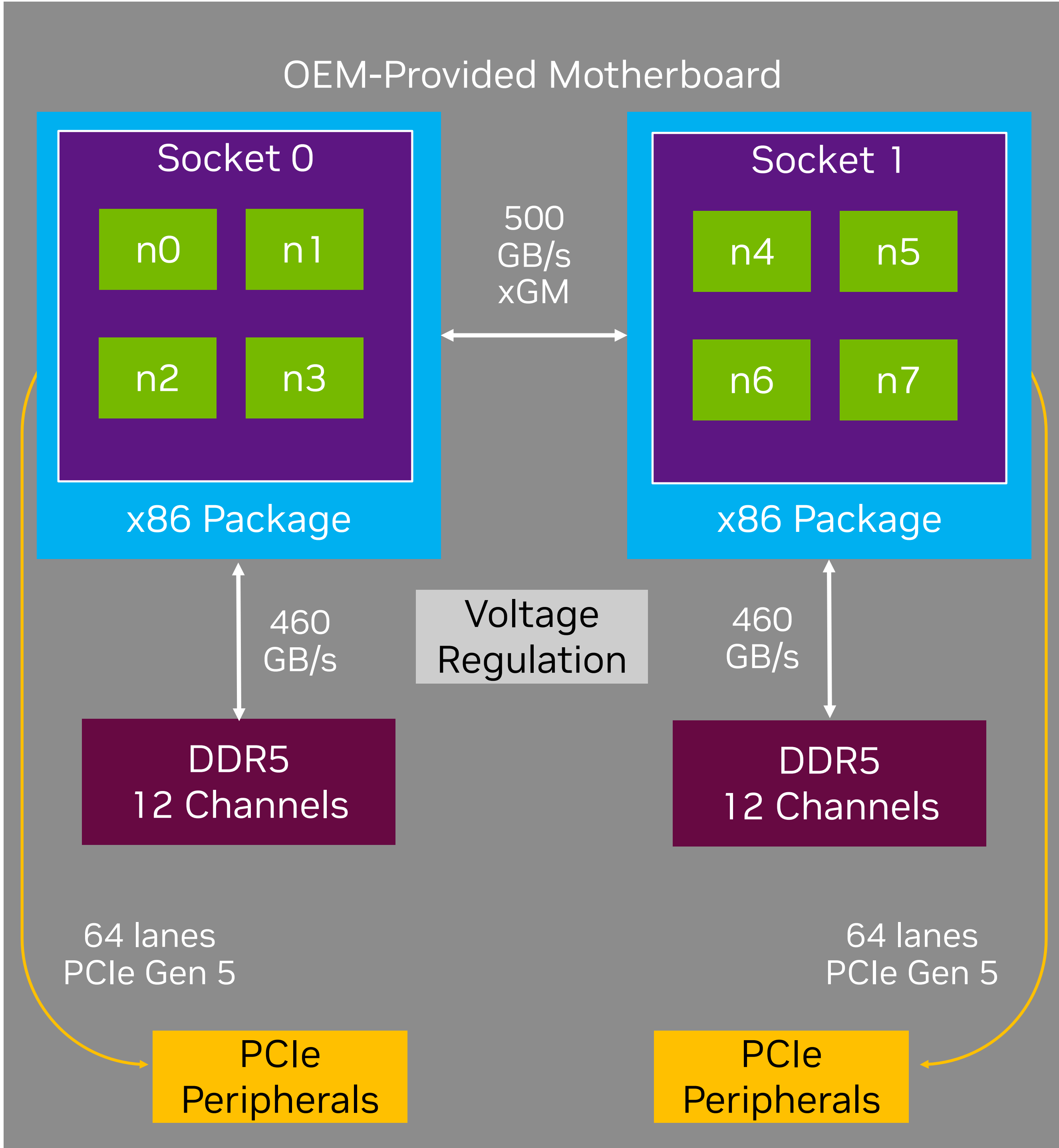
Reduces NUMA Bottlenecks

Grace Server
Grace C2 Superchip



- 2 NUMA Nodes
- 2 Compute Dies
- 500 Watts (CPU + MEM)
- 900 GB/s n-to-n

Conventional 2-Socket Server
Example: 2x AMD Genoa, Native NPS=4



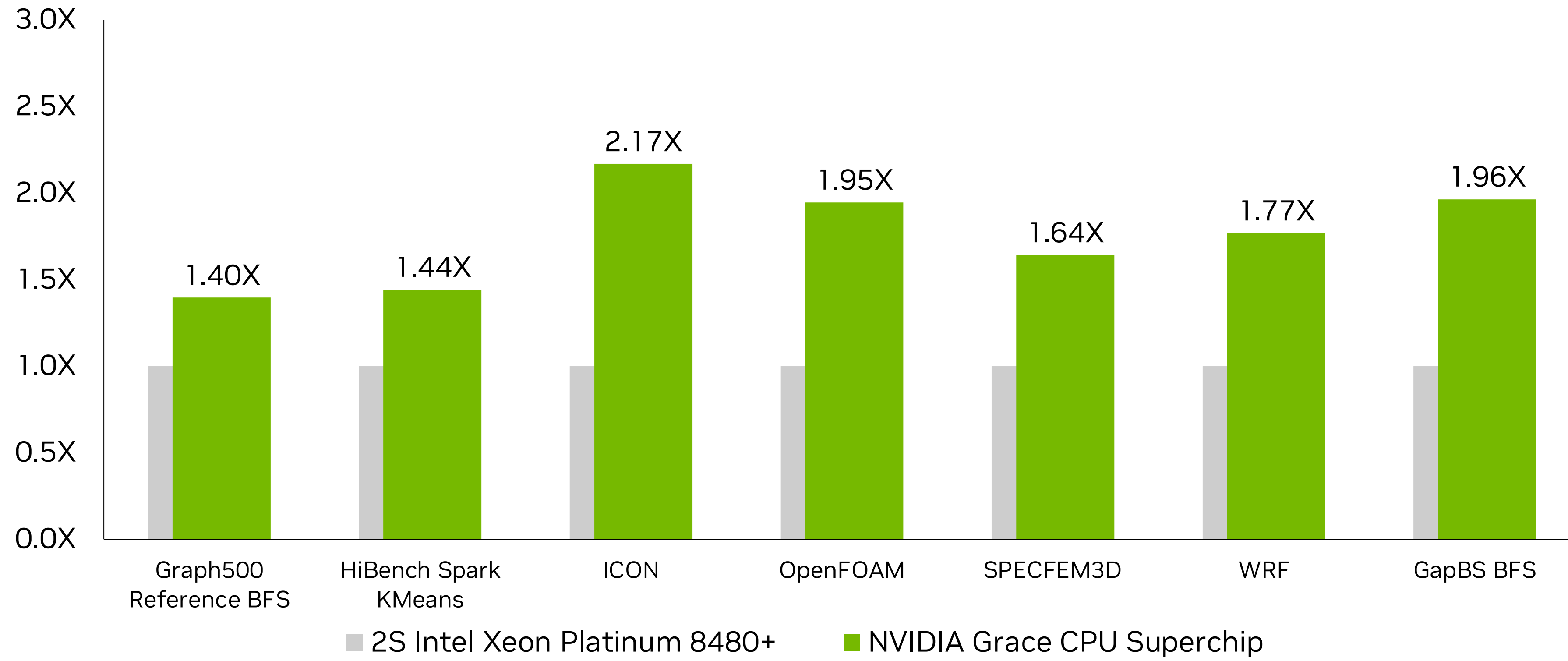
- 8 NUMA Nodes
- 24 Compute Chiplets
- 900+ Watts (CPU + MEM)
- 500 GB/s n-to-n

NVIDIA Grace CPU Doubles HPC Data Center Throughput

Breakthrough Performance and Efficiency

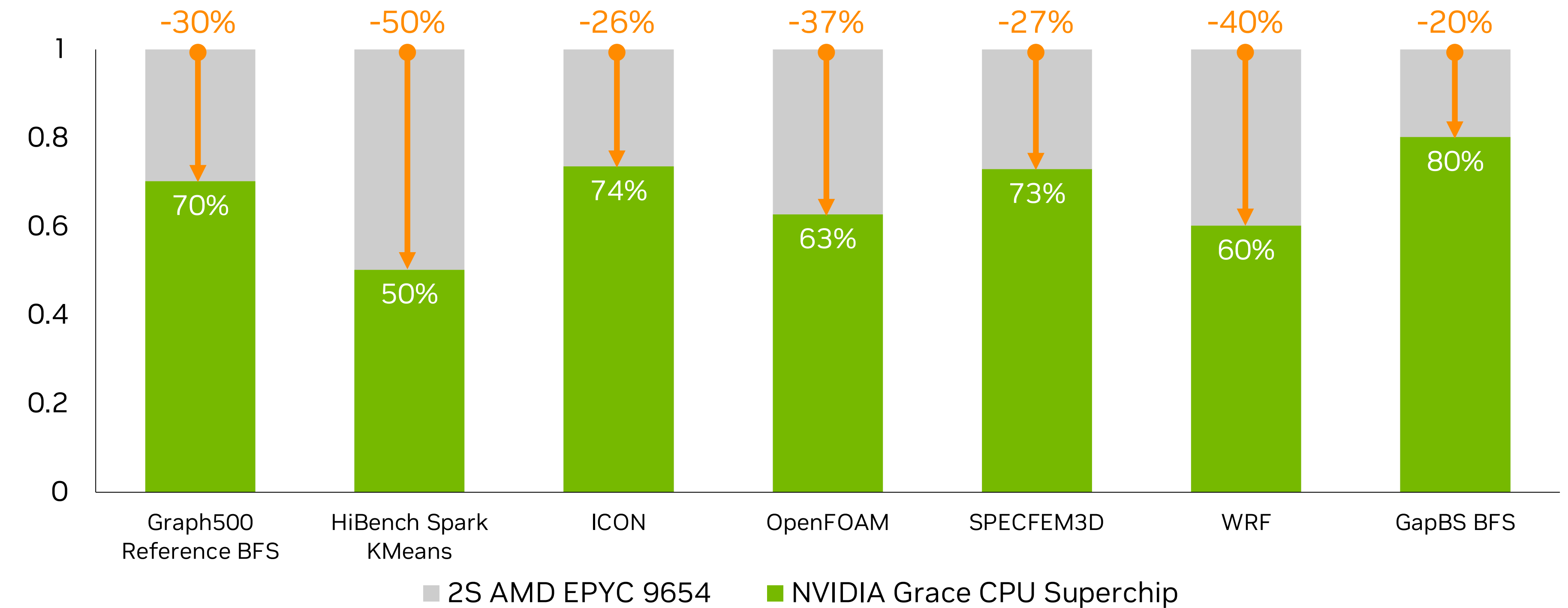
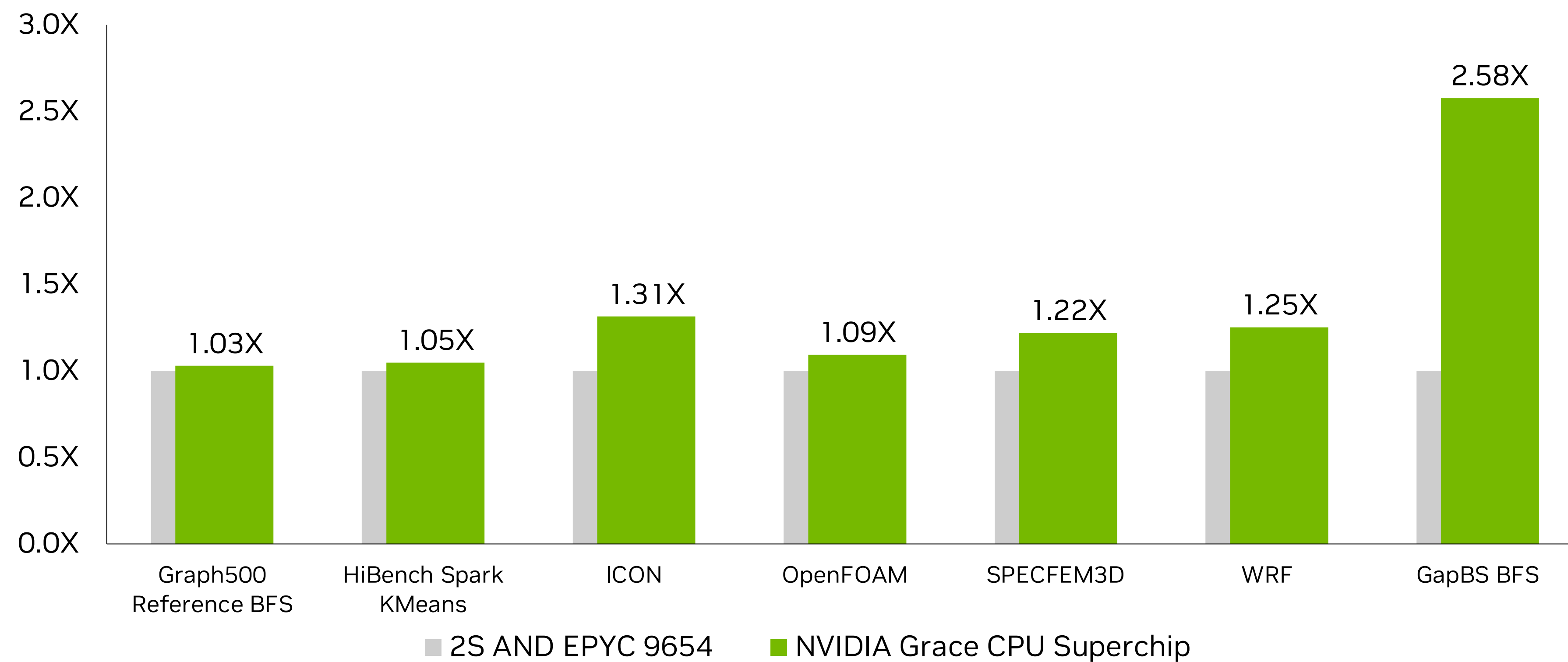
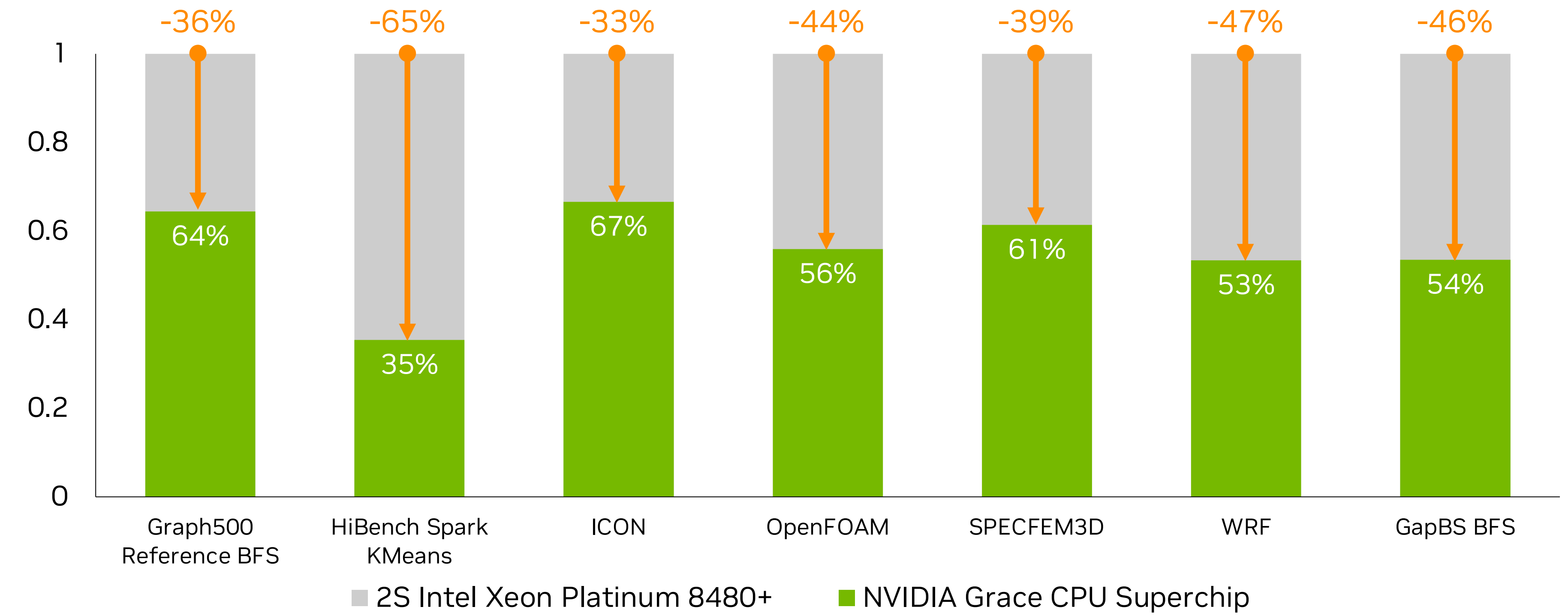
Time-to-Solution

(HIGHER is better)



Energy-To-Solution

(LOWER is better)



Data Center levsingle-node lab measured Grace Superchip vs. x86 flagship 2-socket data center systems (AMD EPYC 9654 and Intel Xeon 8480+). Seismic Data Proc: SPECFEM3D four_material_simple_model CFD: OpenFOAM Motorbike | Large v2212 Climate: NEMO Gyre_Pisces v4.2.0 Weather: ICON QUBICC 80 km resolution Weather: WRF CONUS12km NVIDIA Grace Superchip performance based on measurements. Results subject to change.

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

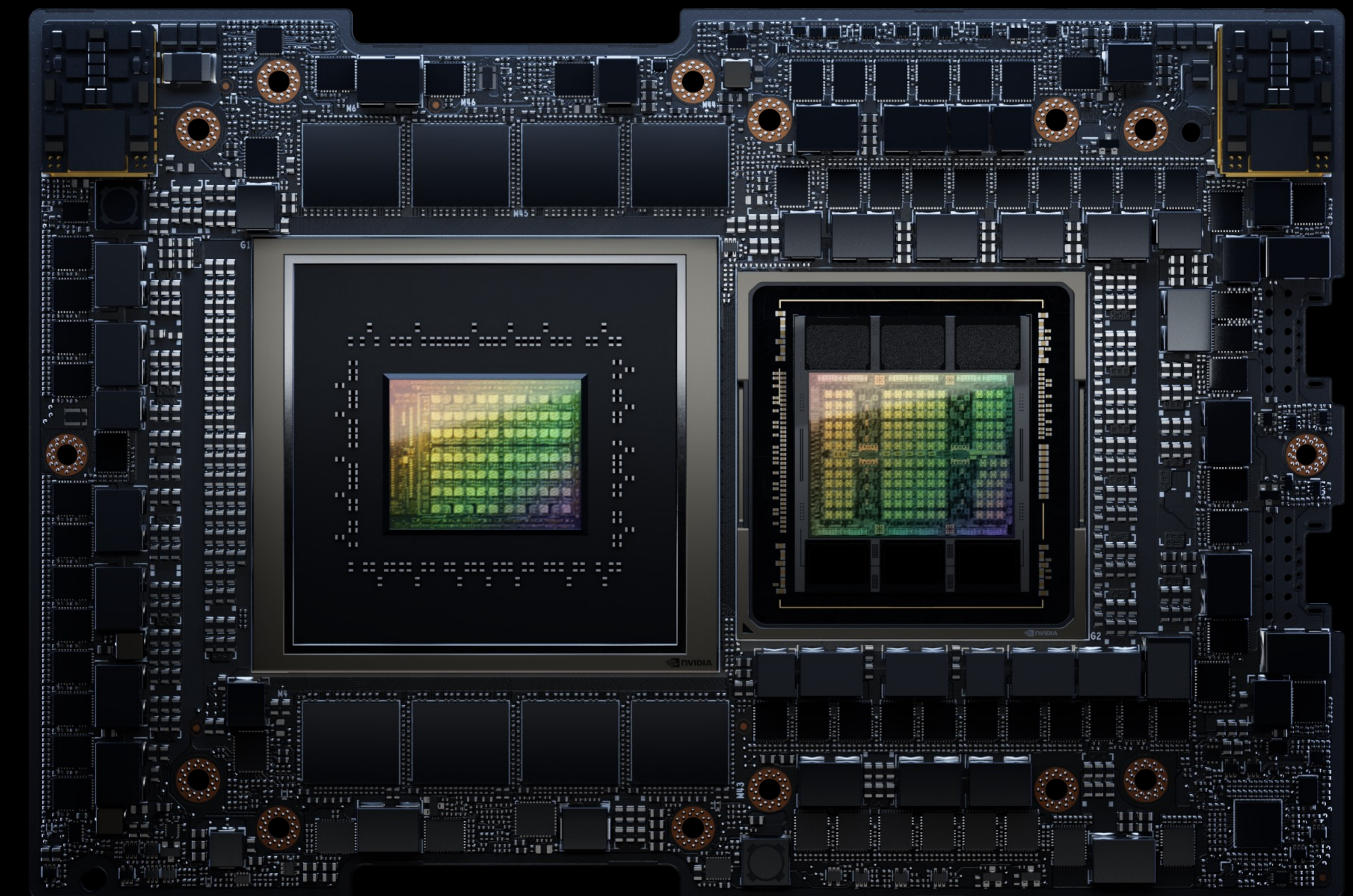
1 CPU:1 GPU node simple to manage and schedule for for HPC, enterprise, and cloud

Best Perf/TCO for diverse workloads

Maximize data center utilization and power efficiency

Continued Innovation

Grace and Hopper-Next in 2024

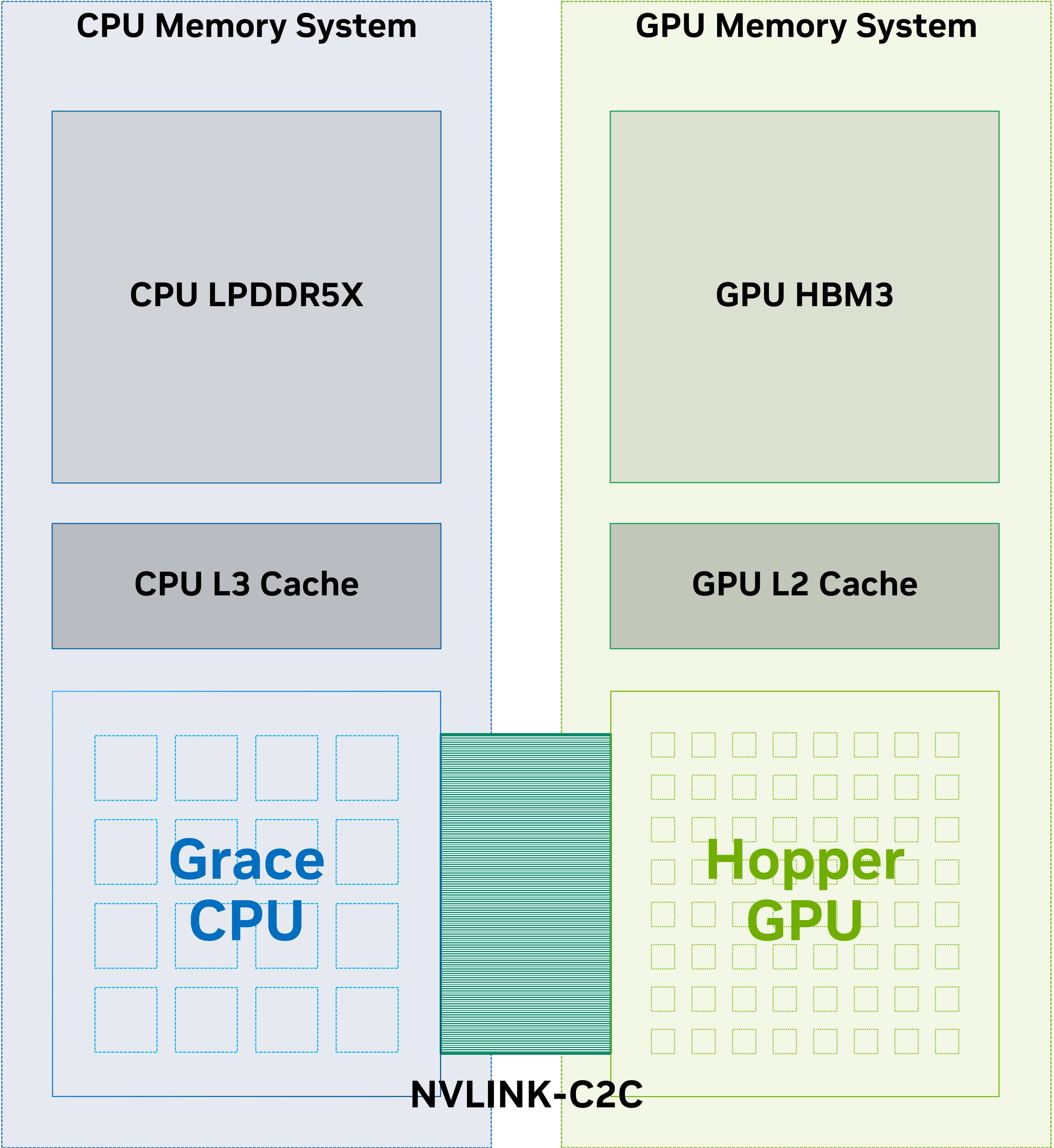
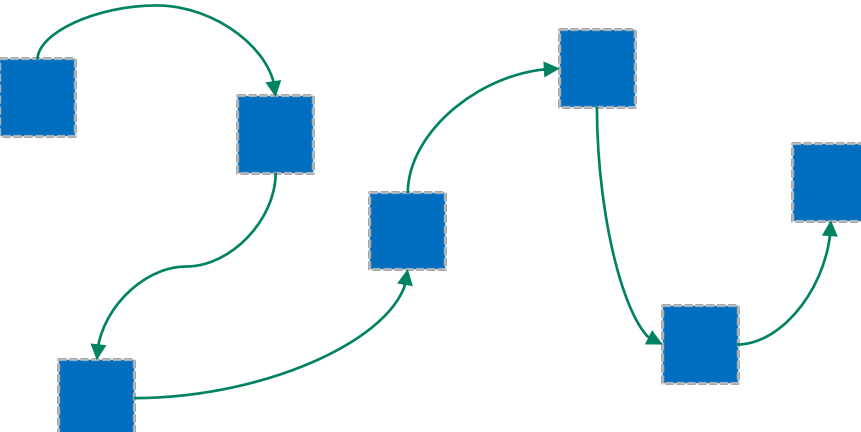


900GB/s NVLink-C2C | 576GB High-Speed Memory
4 PF AI Perf | 72 Arm Cores

Two Memory Systems, Each Optimized For Its Processor

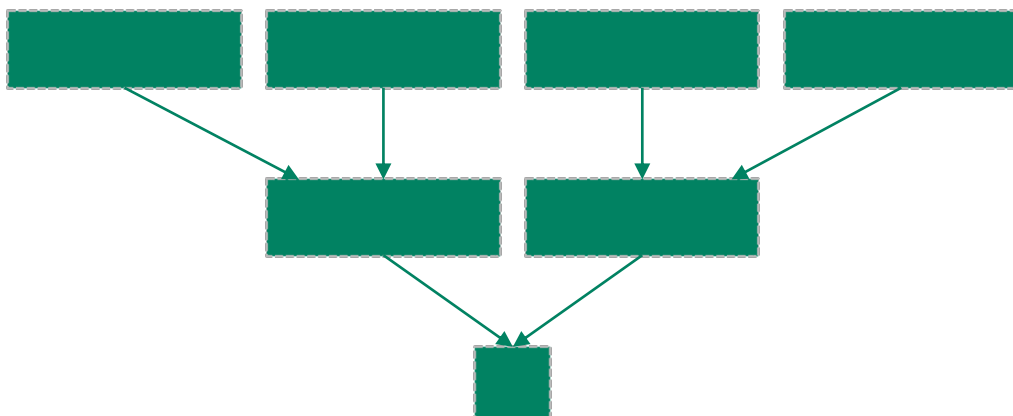
CPU memory system is optimized for **low latency** and **deep cache hierarchy**

Run **latency-sensitive** code on the CPU, e.g. a linked list



GPU memory system is optimized for **high throughput** and **high bandwidth cache**

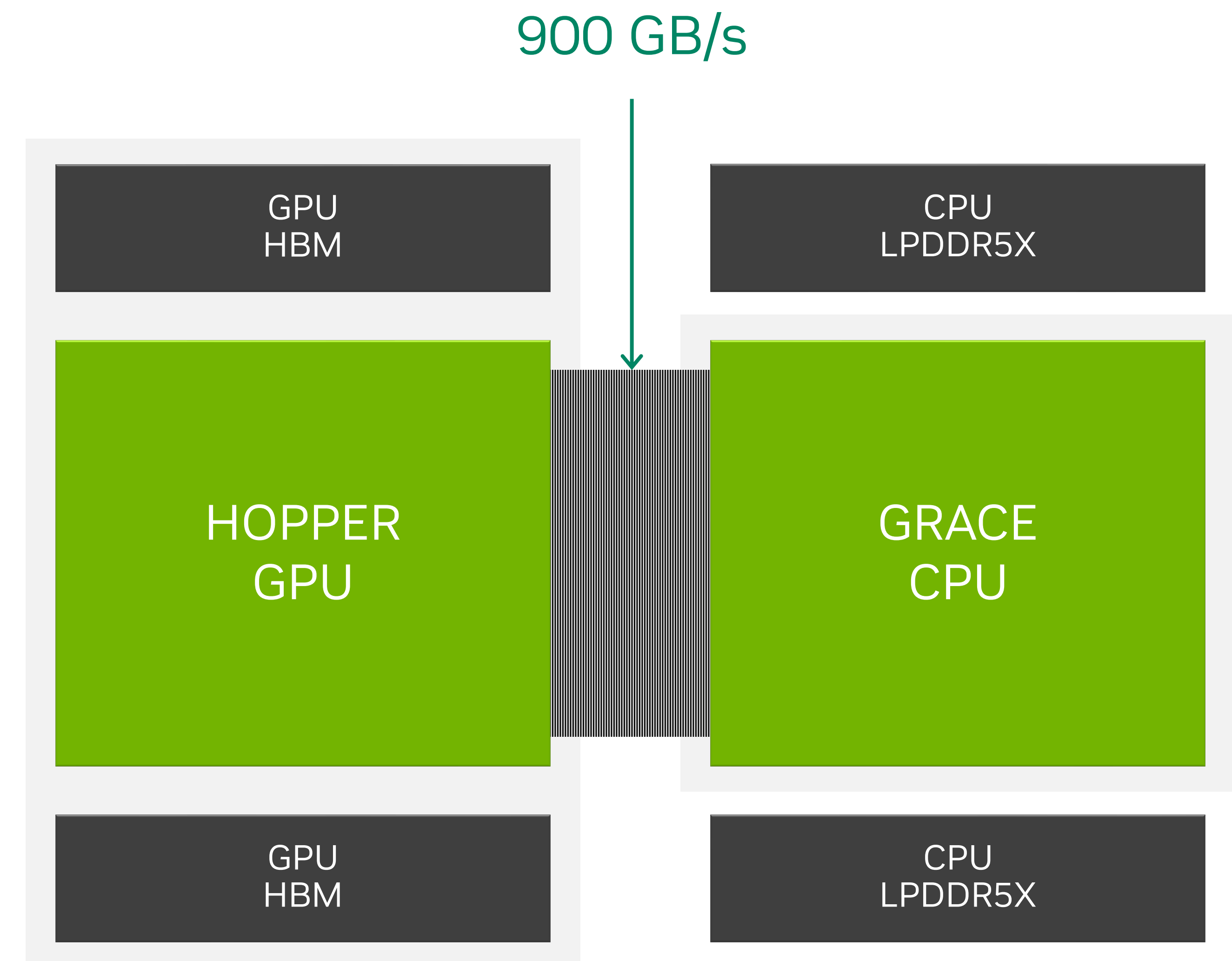
Data- and maths-intensive code on the GPU, e.g. vector reduction



NVLINK-C2C

High Speed Chip to Chip Interconnect

- Used to create the Grace Hopper, and Grace Superchips
 - Native atomics, including standard C++ atomic support
 - **Enables coherency**
- Up to 900 GB/s of raw bidirectional BW
 - Same BW as GPU to GPU NVLINK on Hopper
- Low power interface - 1.3 pJ/bit
 - **More than 5x more power efficient than PCIe**
- Unified Memory with shared page tables
 - **Shared CPU and GPU virtual address space (AST)**

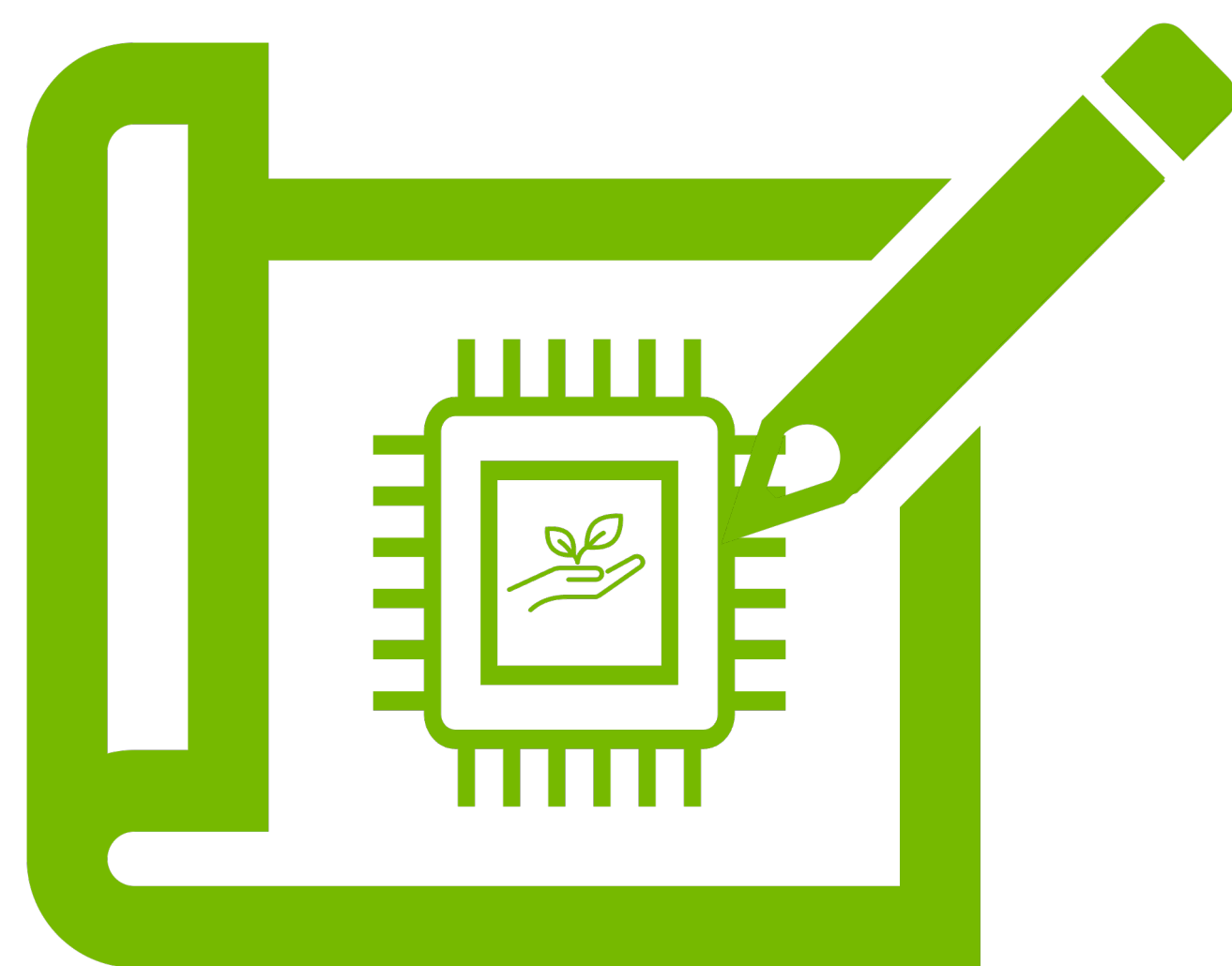


NVIDIA GH200 Optimizing Power and Performance at Data Center Scale

Higher Performance for Less Power

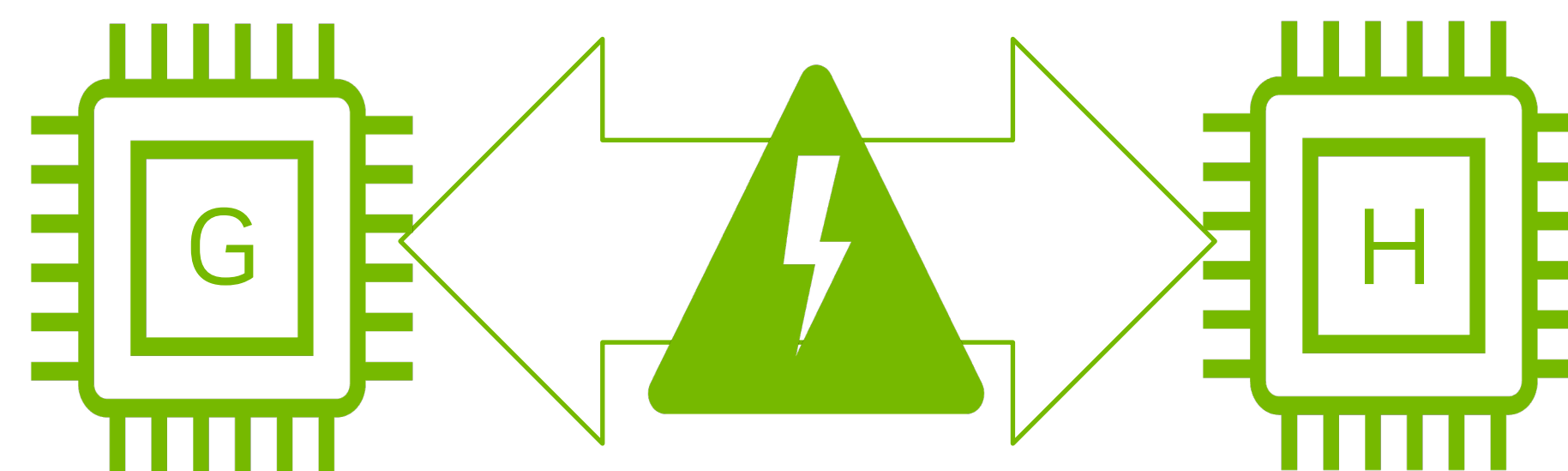
Energy Efficient Design

Energy efficient CPU, GPU, memory and IO



Automatic Power Steering

Automatically shifts power between CPU and GPU

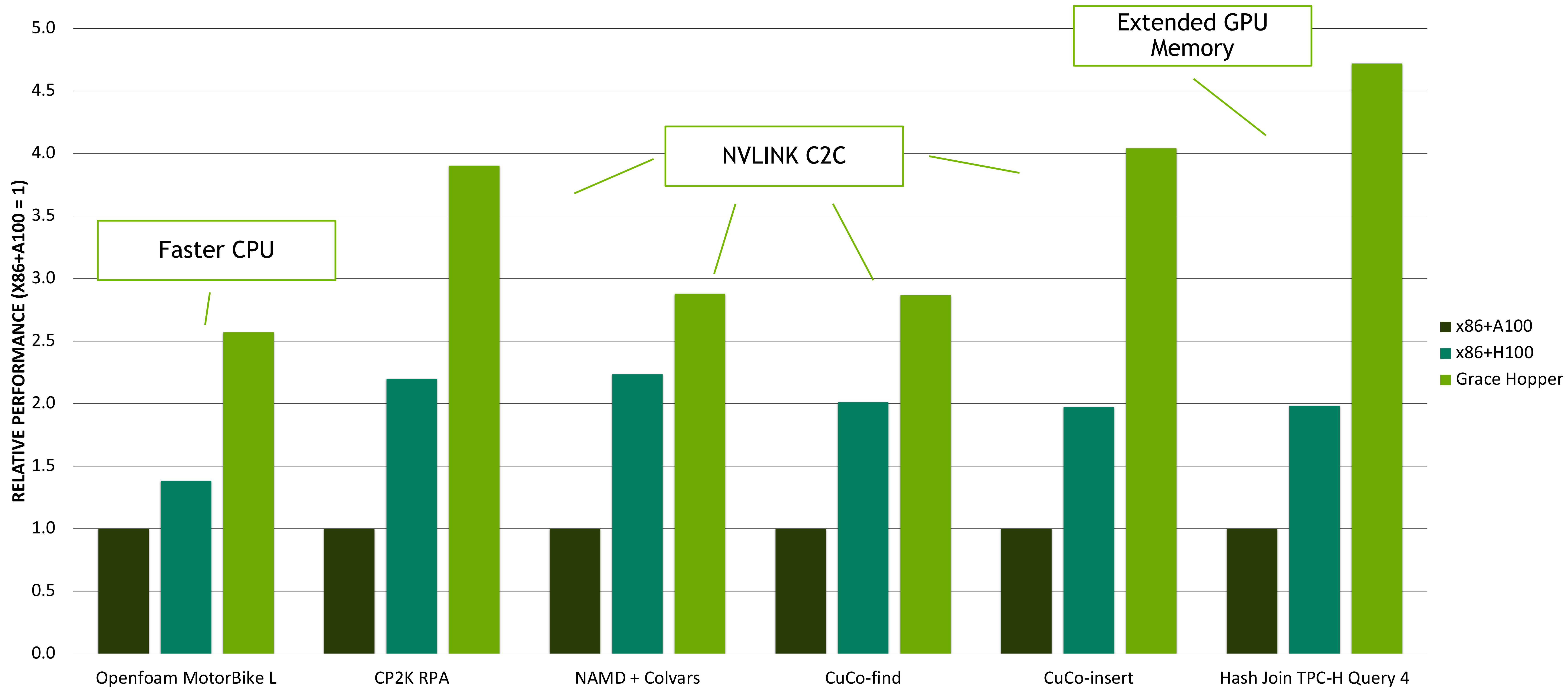


Application Power-Tuning

Adjustable clocks for improved energy efficiency

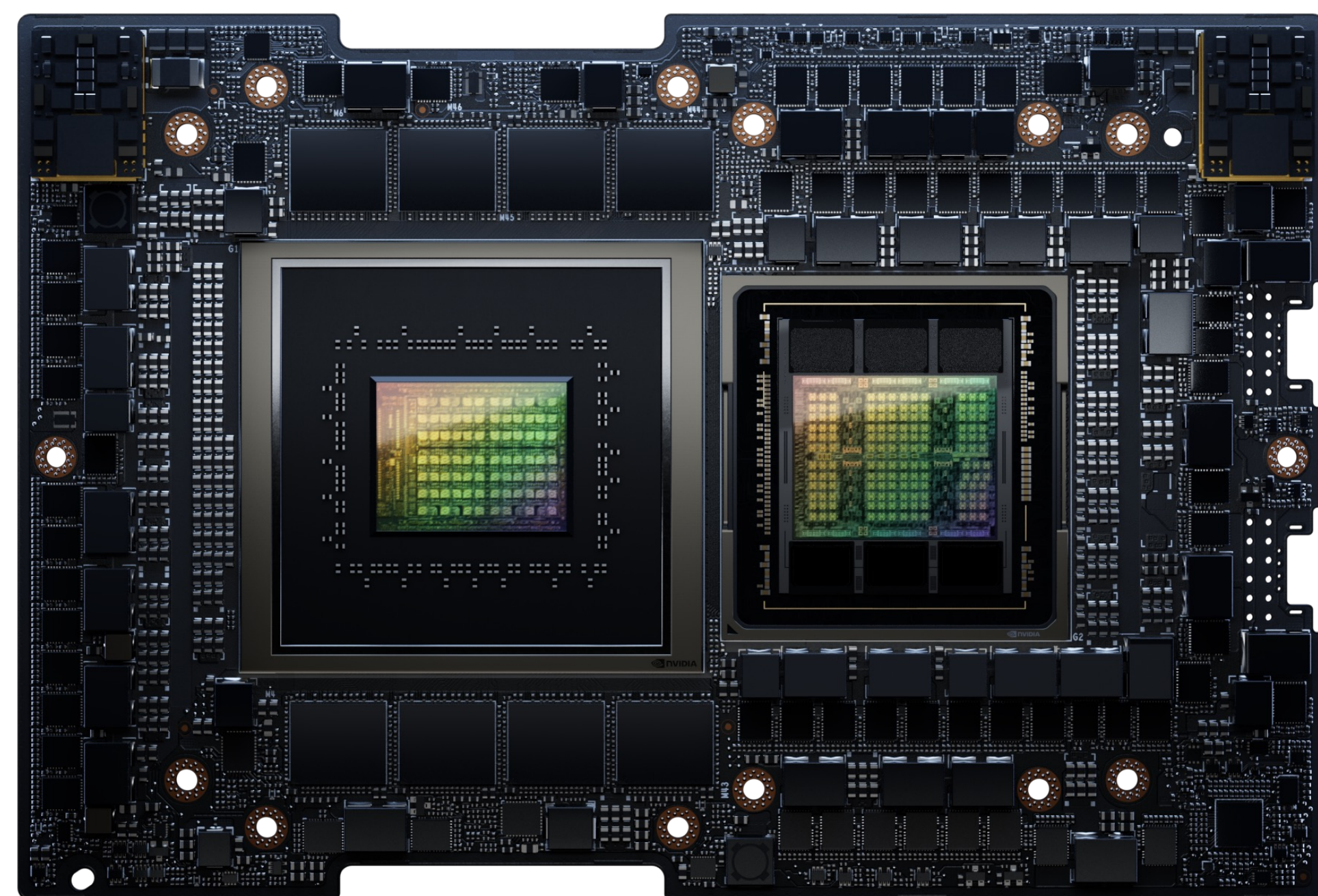


GH200 Performance On Different Workloads



NVIDIA GH200 Grace Hopper Superchip

Processor For The Era of Accelerated Computing And Generative AI

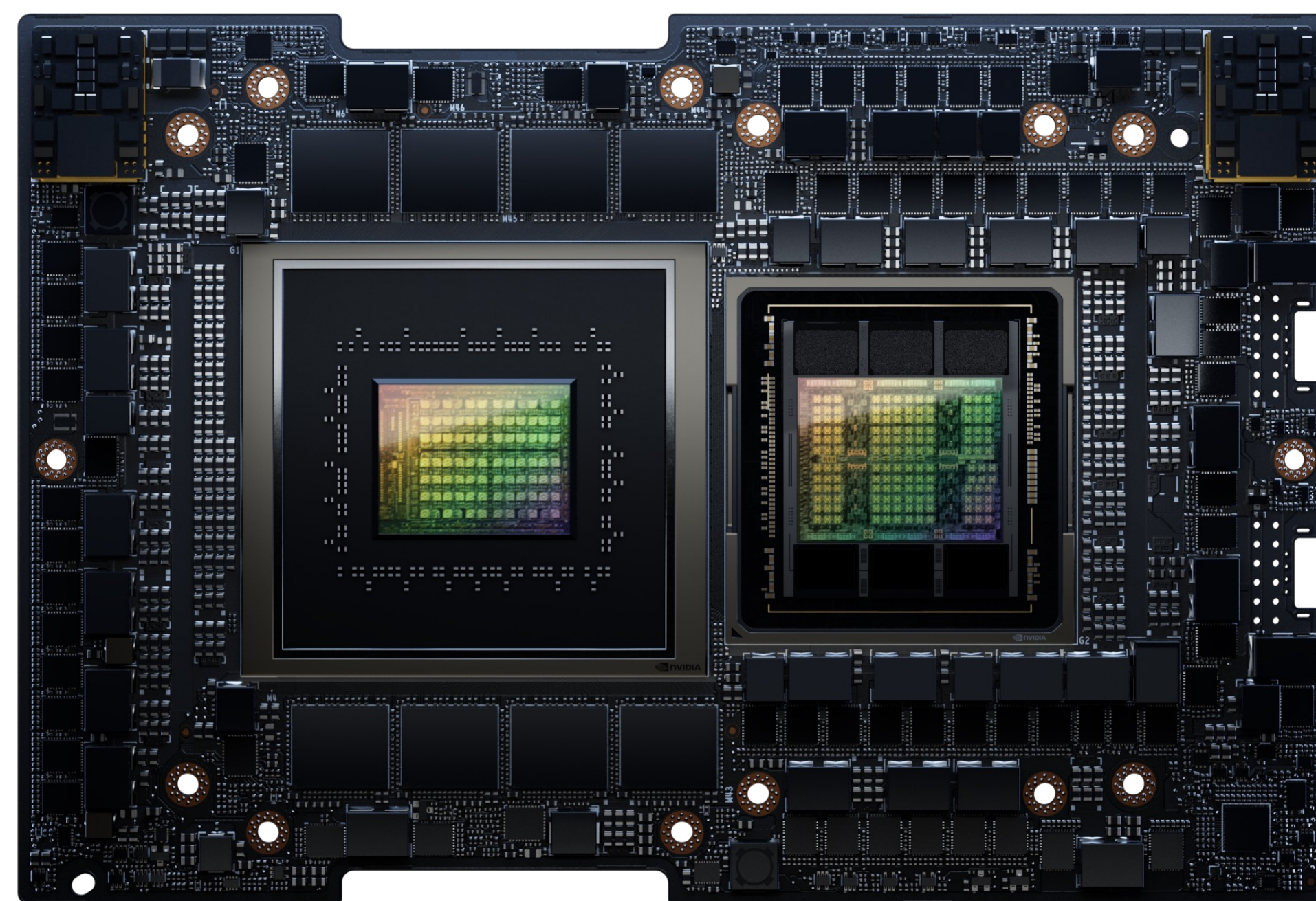


72 Core Grace CPU | 4 PFLOPS Hopper GPU
96 GB HBM3 | 4 TB/s | 900 GB/s NVLink-C2C

- 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 now

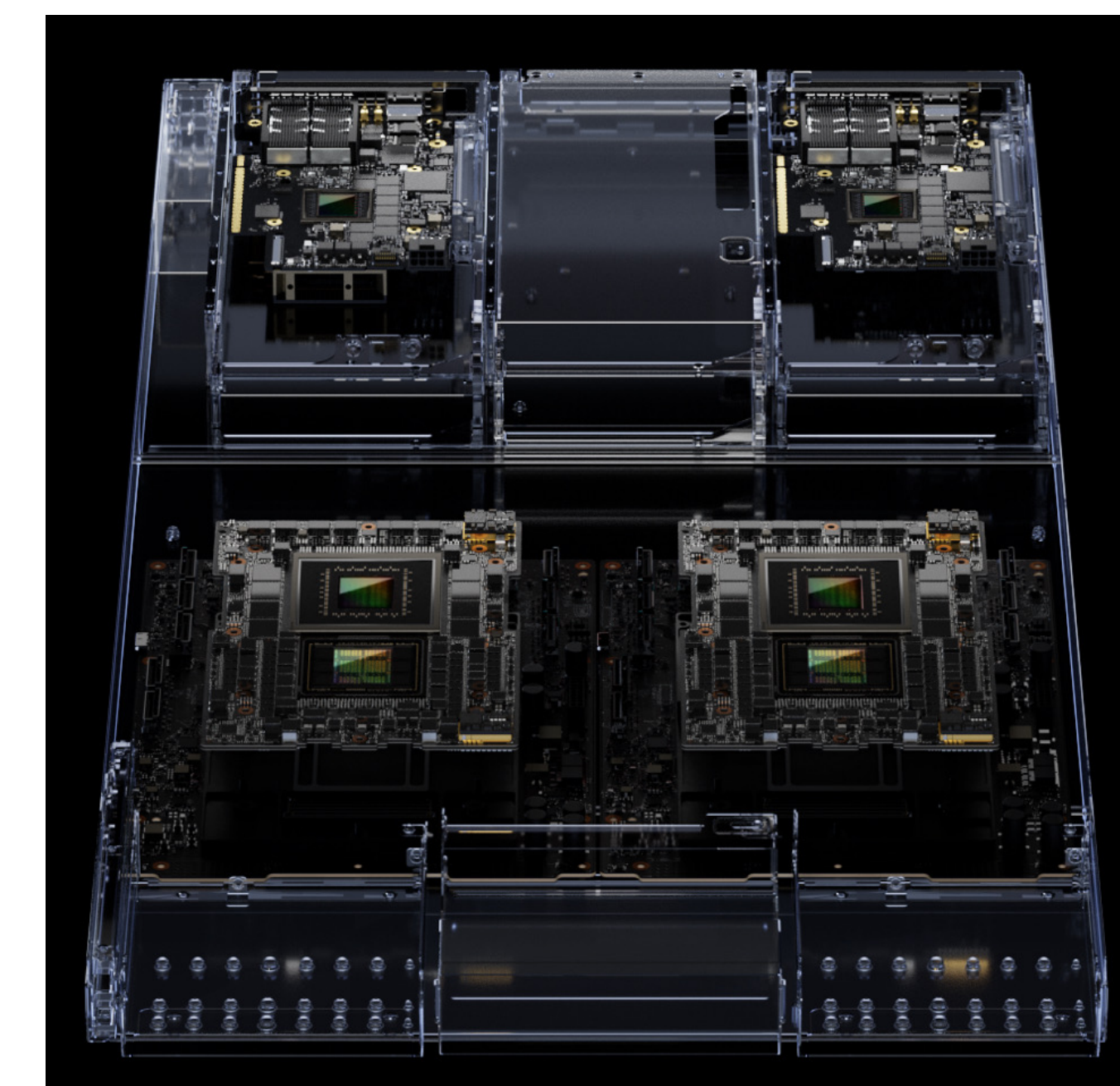


72 Core Grace CPU | 4 PFLOPS Hopper GPU
144 GB HBM3e | 5 TB/s | 900 GB/s NVLink-C2C

- 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

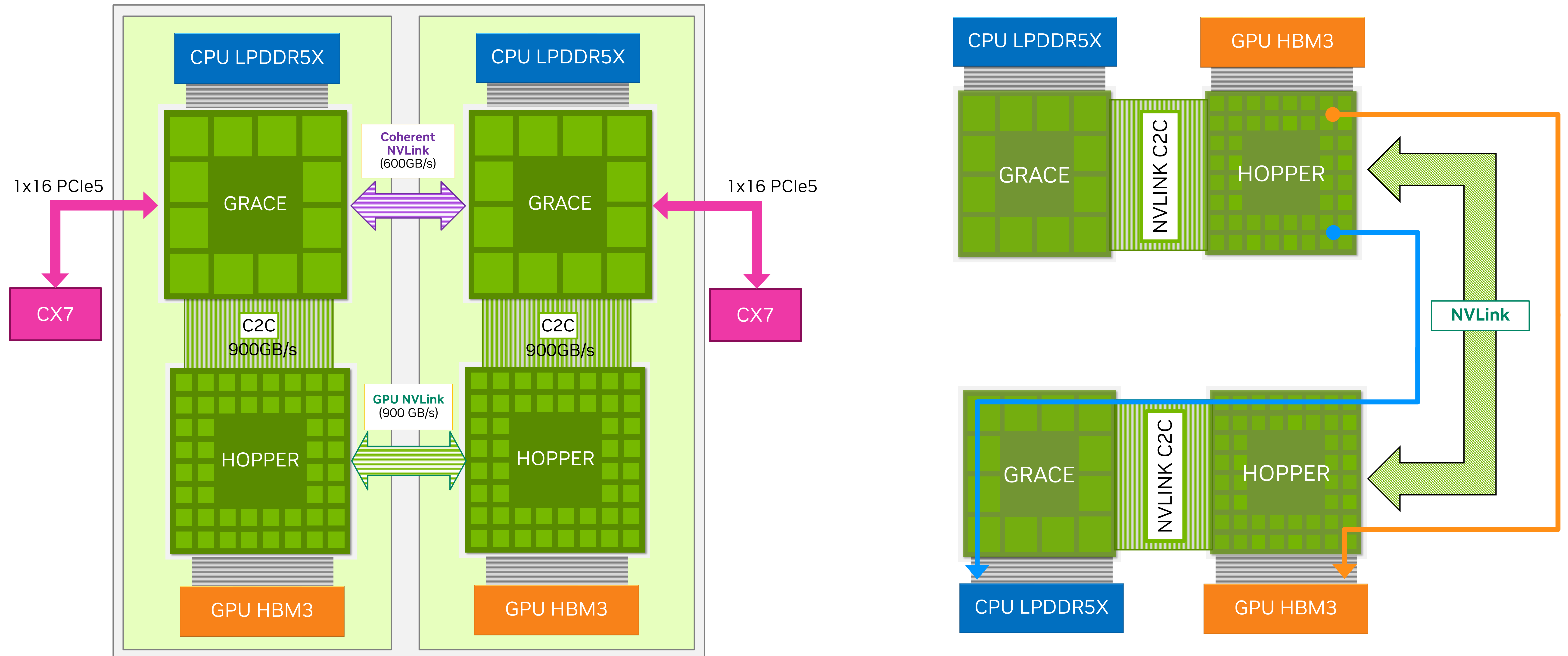
- Simple to deploy MGX-compatible design
- Combined 1.2 TB fast memory
- 3.5x capacity and 3x bandwidth vs H100
- Full NVIDIA Compute Stack

NVLink Dual GH200 System

Available late Q2 2024

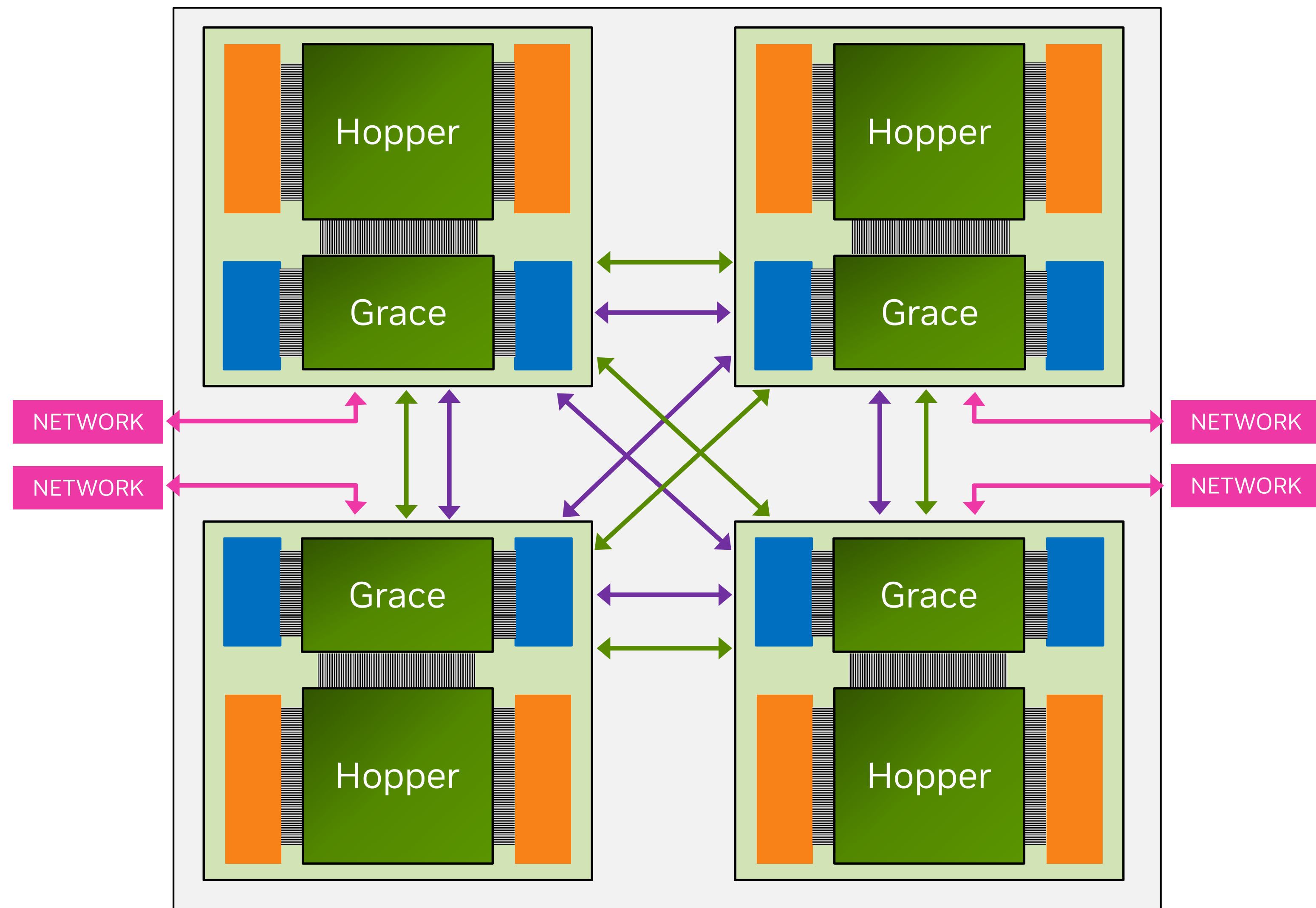
NVLink Dual GH200 System

One OS image, double CPU & GPU performance

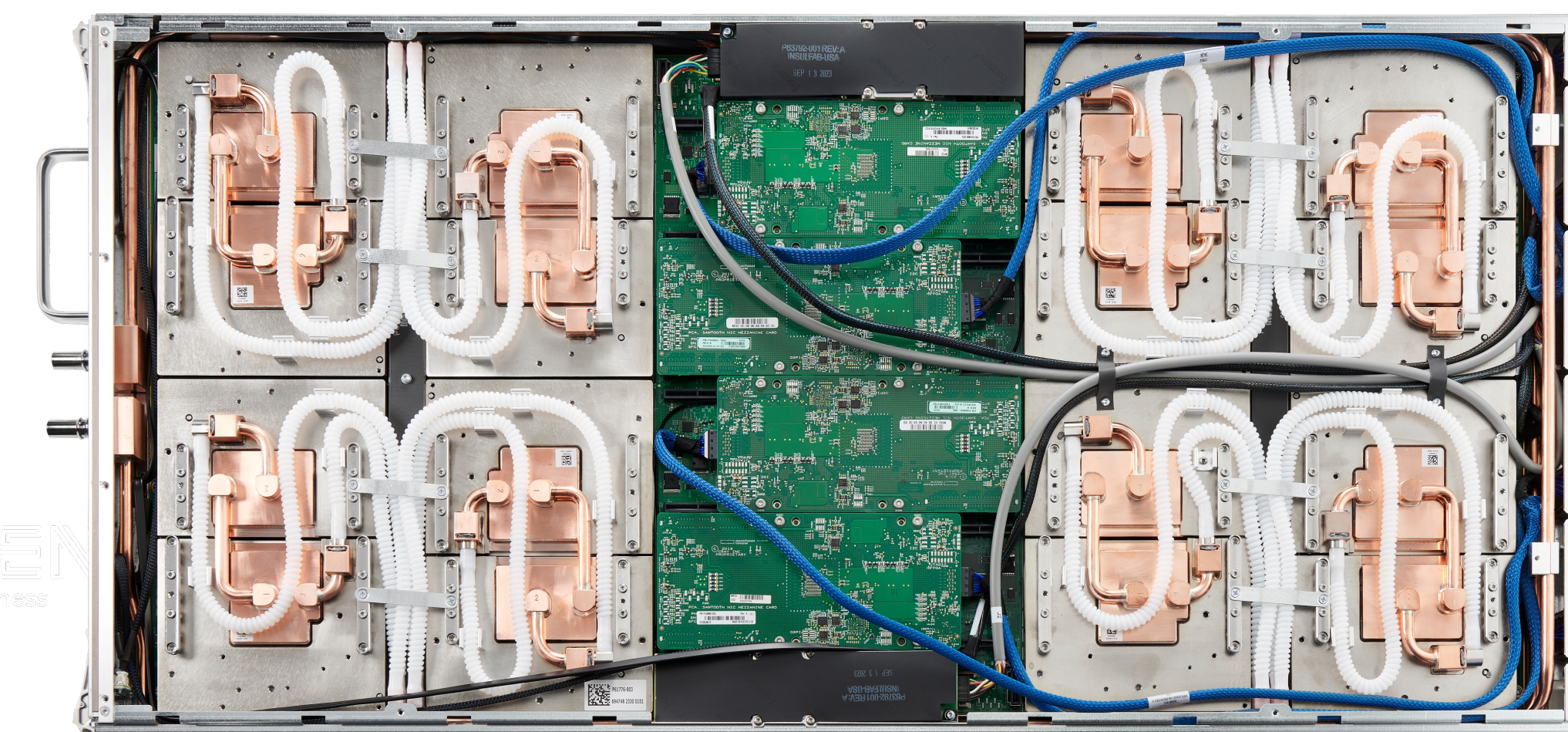


NVIDIA Quad GH200

Node architecture for scalable dense supercomputing



EVIDEN



Hewlett Packard
Enterprise

GPU NVLINK
(300 GB)



COHERENT cNVLINK
(200 GB)



Programming the NVIDIA Platform

Programming the NVIDIA platform

CPU, GPU, and Network

ACCELERATED STANDARD LANGUAGES

ISO C++, ISO Fortran

```
std::transform(par, x, x+n, y, y,  
              [=] (float x, float y) { return y +  
a*x; }  
);
```

```
do concurrent (i = 1:n)  
  y(i) = y(i) + a*x(i)  
enddo
```

```
import cunumeric as np  
...  
def saxpy(a, x, y):  
  y[:] += a*x
```

INCREMENTAL PORTABLE OPTIMIZATION

OpenACC, OpenMP

```
#pragma acc data copy(x,y) {  
...  
std::transform(par, x, x+n, y, y,  
              [=] (float x, float y) {  
                return y + a*x;  
              });  
...  
}  
  
#pragma omp target data map(x,y) {  
...  
std::transform(par, x, x+n, y, y,  
              [=] (float x, float y) {  
                return y + a*x;  
              });  
...  
}
```

PLATFORM SPECIALIZATION

CUDA

```
__global__  
void saxpy(int n, float a,  
           float *x, float *y) {  
  int i = blockIdx.x*blockDim.x +  
          threadIdx.x;  
  if (i < n) y[i] += a*x[i];  
}  
  
int main(void) {  
  ...  
  cudaMemcpy(d_x, x, ...);  
  cudaMemcpy(d_y, y, ...);  
  
  saxpy<<<(N+255)/256,256>>>(...);  
  
  cudaMemcpy(y, d_y, ...);  
}
```

ACCELERATION LIBRARIES

Core

Math

Communication

Data Analytics

AI

Quantum

Choosing A Programming Model

There can be ~~only~~ *more than* one.

| Libraries | Standard Languages | Compiler Directives | CUDA Languages |
|---|---|--|--|
| <ul style="list-style-type: none">• Accelerate common operations with little/no code changes.• Expert-tuned performance.• Forward support guarantees. | <ul style="list-style-type: none">• Strong cross-platform support.• Single source code for multiple platforms.• Reduced learning curve. | <ul style="list-style-type: none">• High cross-platform support.• Single source code for multiple platforms.• Reduced learning curve.• Additional programmer control. | <ul style="list-style-type: none">• Exposes full GPU capabilities.• Trades portability for performance.• Distinct GPU/CPU code paths.• Full programmer control. |
| Programmer Productivity | | | Programmer Control |

By design these approaches are interoperable so developers can choose the right balance for their needs.

CUDA Toolkit

Develop, Optimize and Deploy GPU-Accelerated Applications

CUDA Toolkit 12

NVIDIA® CUDA® Toolkit (CTK) provides what you need to create high performance GPU-accelerated applications.

Develop, optimize, and deploy GPU-accelerated applications on embedded systems to Cloud-based platforms and HPC supercomputers.

Included in the toolkit:

- GPU-accelerated libraries
- Debugging and optimization tools
- C/C++ compilers
- Runtime library

Also, supports Fortran and Python parallel language constructs.

Learn more about the [CUDA Toolkit](#) and [Nsight Developer Tools](#) on our DevZone.

Accelerated Computing Software Engine

CUDA 12 introduces support for:

- Hopper and Ada Lovelace architectures
- Arm server processors
- Lazy Module and Kernel Loading
- New developer tools for Python and multi-GPU and multi-node (MGMN) clusters

NVIDIA **Hopper** architecture support includes:

- Next gen Tensor Cores and Transformer Engine
- Next gen Multi-Instance GPU (MIG)
- Mixed precision modes
- Advanced memory management
- Hi-speed NVLink Switch system

CUDA is an Enterprise Scalable Platform

Scale from single GPU laptops and workstations to cloud and supercomputer installations with thousands of GPUs.



HPC SDK Updates

Grace Hopper, unified memory, and more

• HPC SDK 23.11:

- Unified memory support for stdpar, OpenACC, and CUDA C++/Fortran
- NVTX improvements for stdpar codes
 - Now you can see your stdpar in NSight: improved tools support, developer experience, performance optimizations
- C-Fortran Interface
 - Better multi-paradigm interoperability for mixed C, C++, and Fortran codes
 - F2008 MPI bindings for nvfortran
- C++20 Coroutines for CPU
 - Future GPU support will enable alternative async models for stdpar
- Support for Grace Hopper in all bundled components
 - Compilers, Math Libraries, Networking, Tools.
- HPC-X is the default MPI implementation optimized for NV platform
- Grace(/Arm) performance (-tp=neoverse-v2)
 - Re-engineered vectorizer, intrinsics, system math library functions

• HPC SDK 24.3:

- Improved compile speed for nvc++
 - Up to 1.15x - 2x faster for some workloads
- Unified memory support for OpenMP Target Offload
- Integrated NVIDIA Performance Library (NVPL) for Grace CPUs
- CUDA Fortran `unified` attribute

• HPC SDK 24.5:

- New NVPL integrations
- Ubuntu 24.04 support
- Improved memory model CLI for HPC Compilers

Unified Memory

- C++ stdpar improvements
- Fortran stdpar improvements
- OpenACC improvements
- CUDA Fortran
- OpenMP Target Offload
- Unified Functions

cuNumeric and Legate

Accelerating Python Applications at Scale

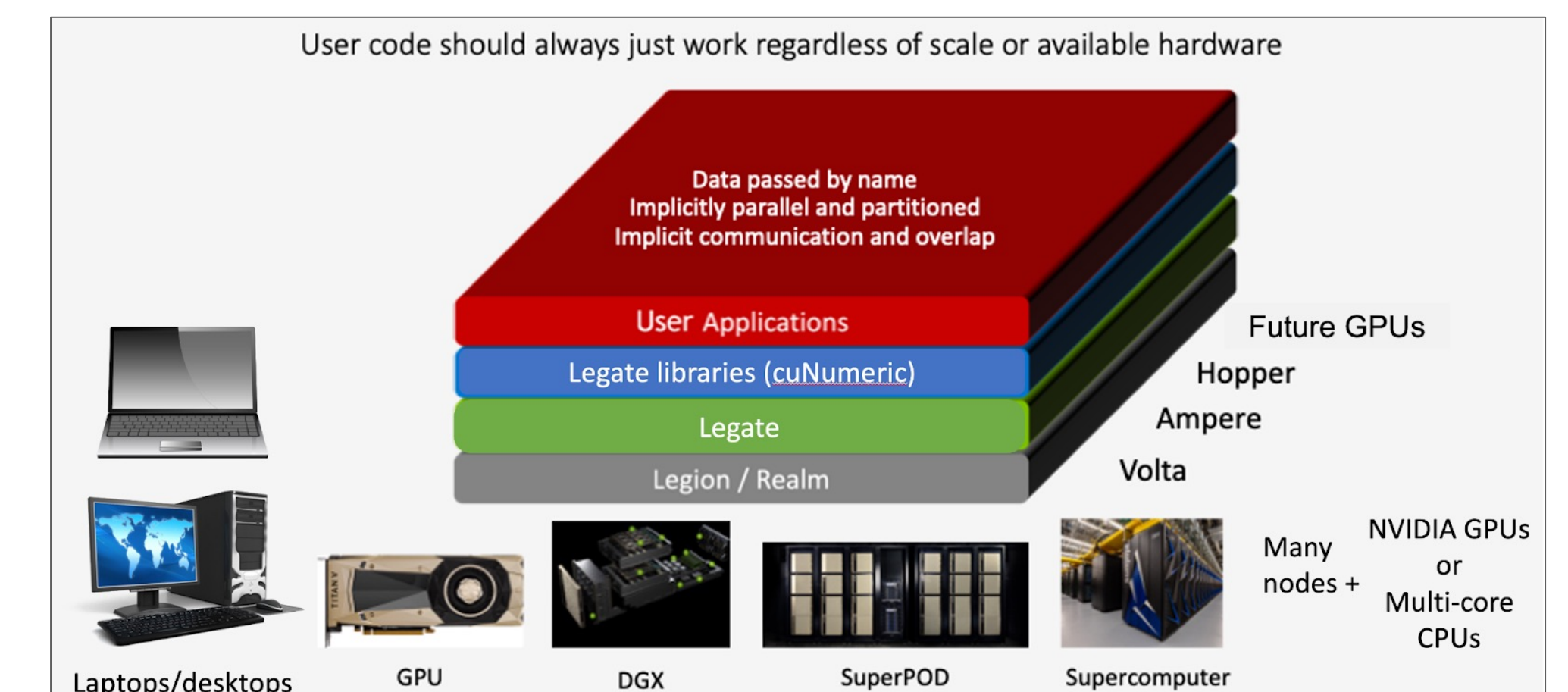
Introducing cuNumeric and Legate

- NVIDIA **cuNumeric** is a **Legate** library aspiring to be a drop-in replacement for *NumPy*
- **Legate** is an abstraction layer running over a runtime system providing multi-GPU and multi-node (MGMN) computing
- Helps developers leverage power of large CPU and GPU clusters by running the same code that runs on laptops
- Learn more about cuNumeric and Legate from the [Accelerating Python Applications with cuNumeric and Legate](#) TechBlog post.

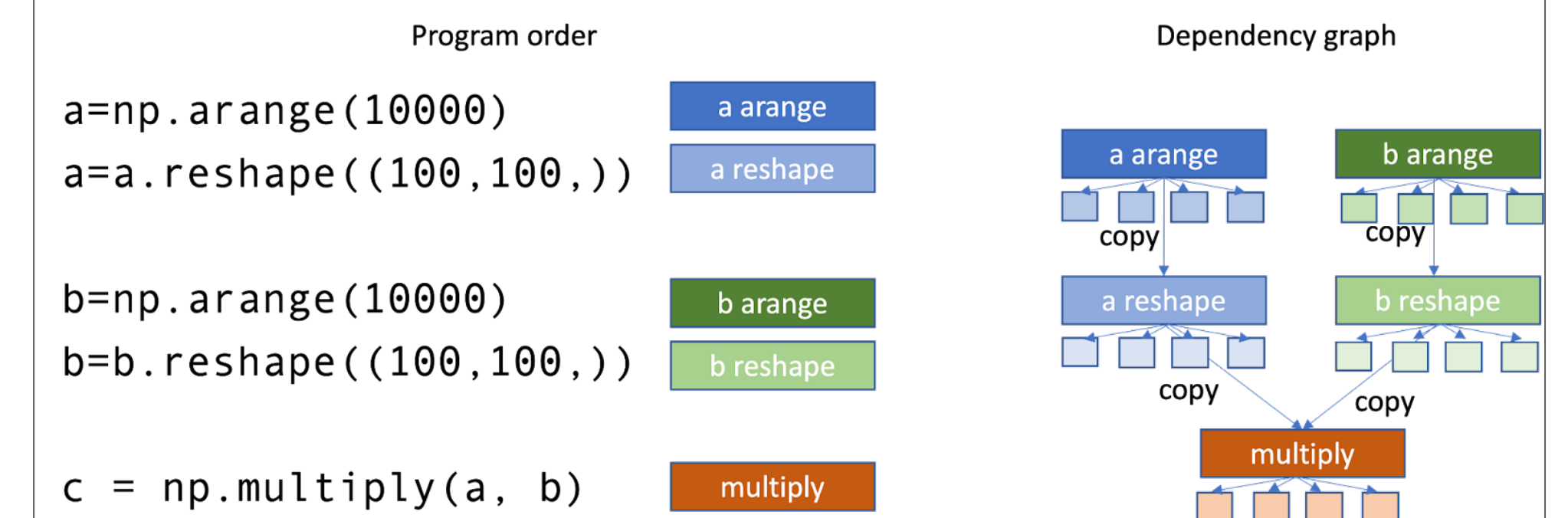
Democratizing Scientific Computing for Python

- Develop and test programs on small datasets on a laptop or workstation
- Scale up to larger datasets deployed on 1000's of GPUs in the cloud, or on a supercomputer -- without code changes
- Key benefits of the cuNumeric library on Legate:
 - Transparently accelerates and scales existing NumPy workflows
 - Scales to up to 1000's of GPUs optimally
 - Requires zero code changes
 - Is freely available. Get started on [GitHub](#) or [Conda](#)

Software Stack and Performance



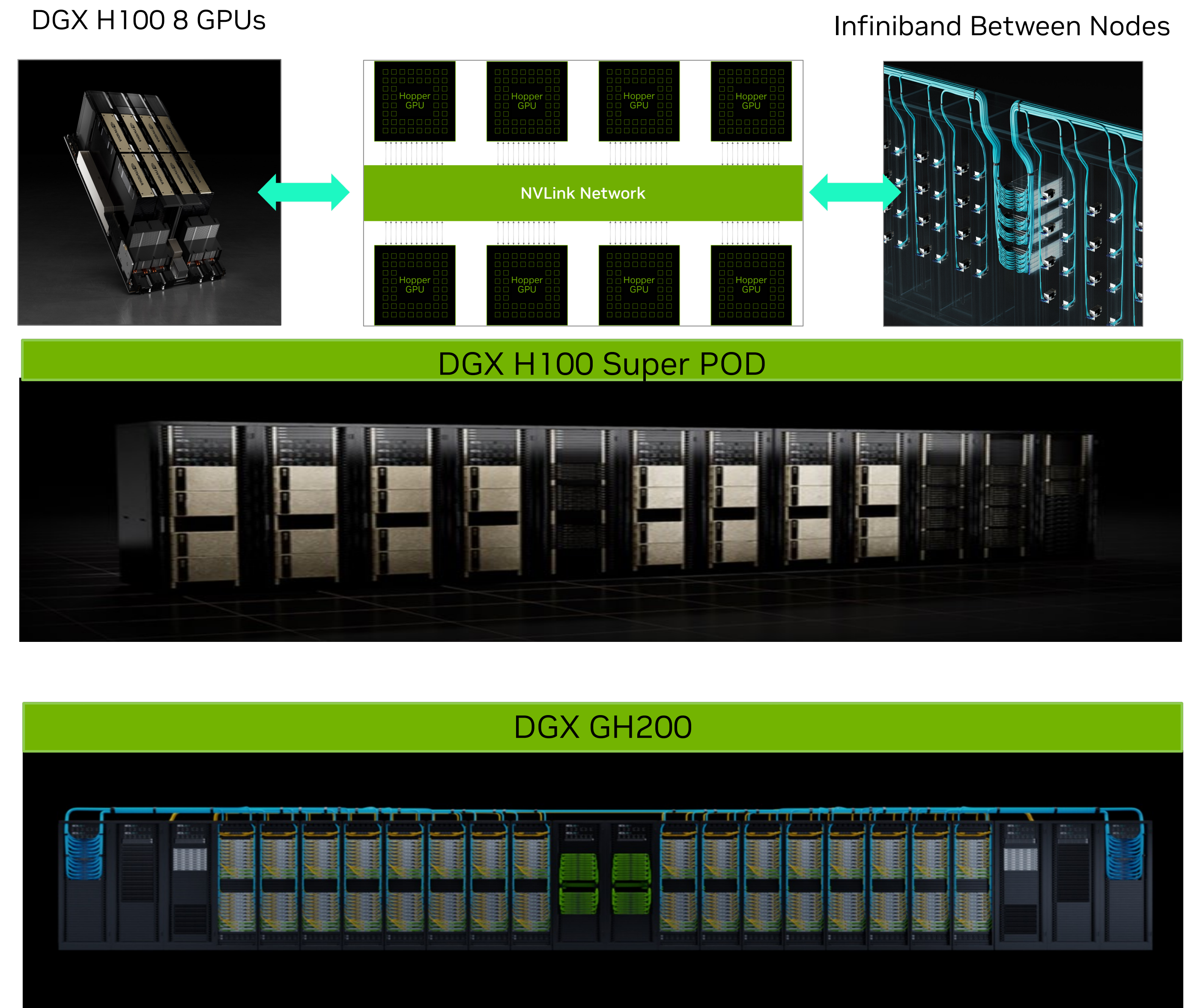
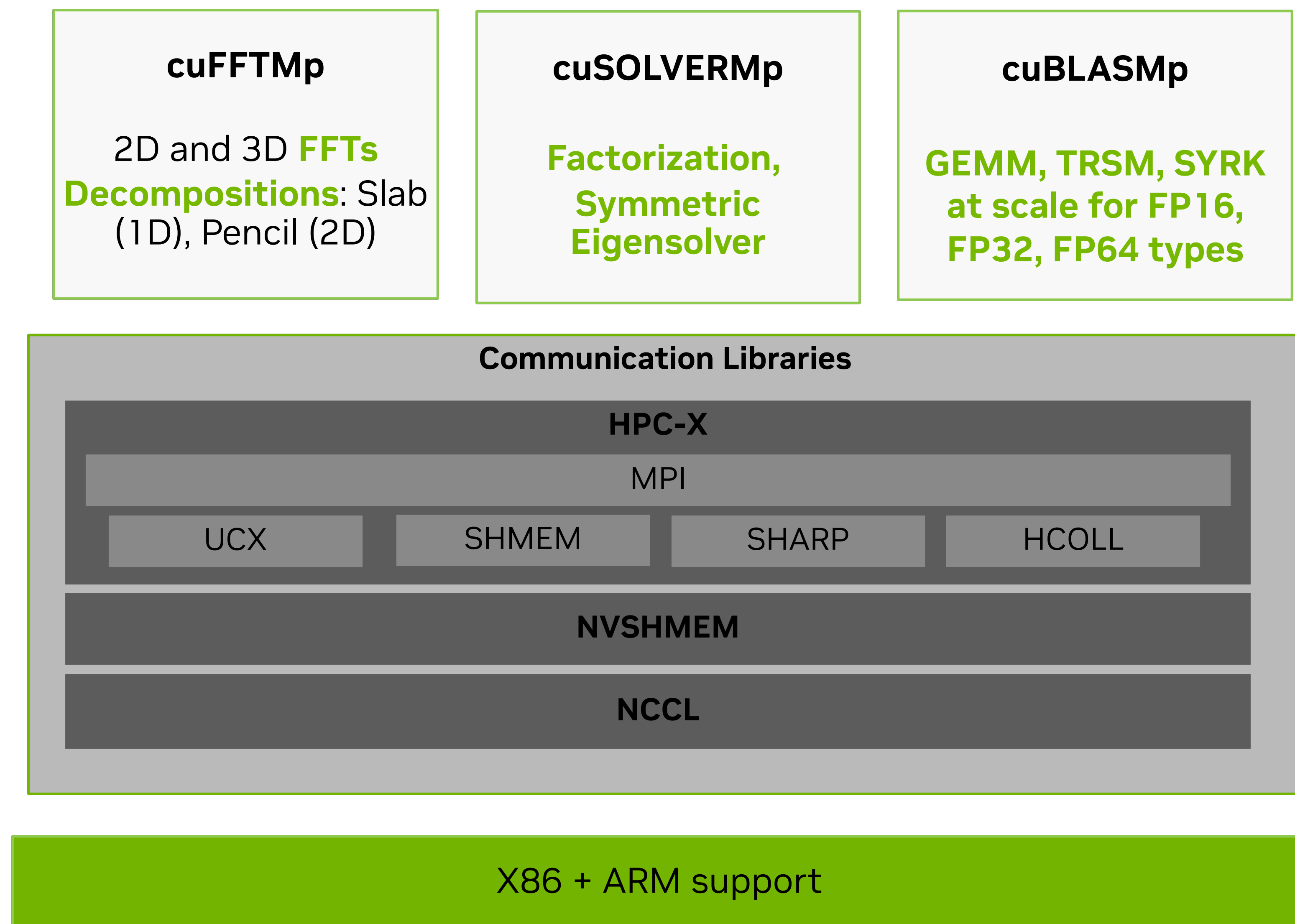
Asynchronous execution in cuNumeric



This figure visualizes a dependency graph executed on four GPUs (single node). Here, *arange*, *reshape* tasks, and *copy* operations for array 'a' can be performed in parallel with those for array 'b'. Note that each of the array-wide operations are also split into four suboperations.

Multi GPU Multi Node APIs

Scalable and Grace Hopper Support

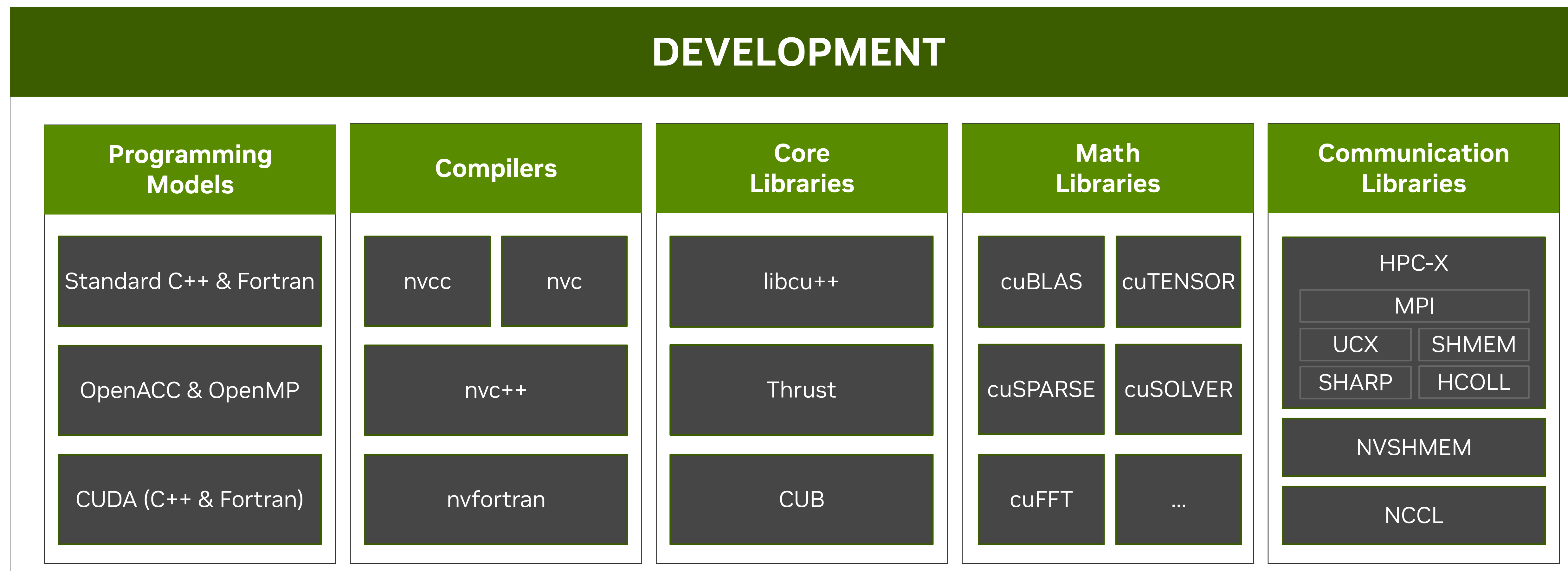


256 Grace Hopper Superchips | **1EFLOPS** AI Performance
| **144TB** unified fast memory
36 L2 NVLink switches | **900 GB/s** GPU-to-GPU
bandwidth | **128 TB/s** bisection bandwidth

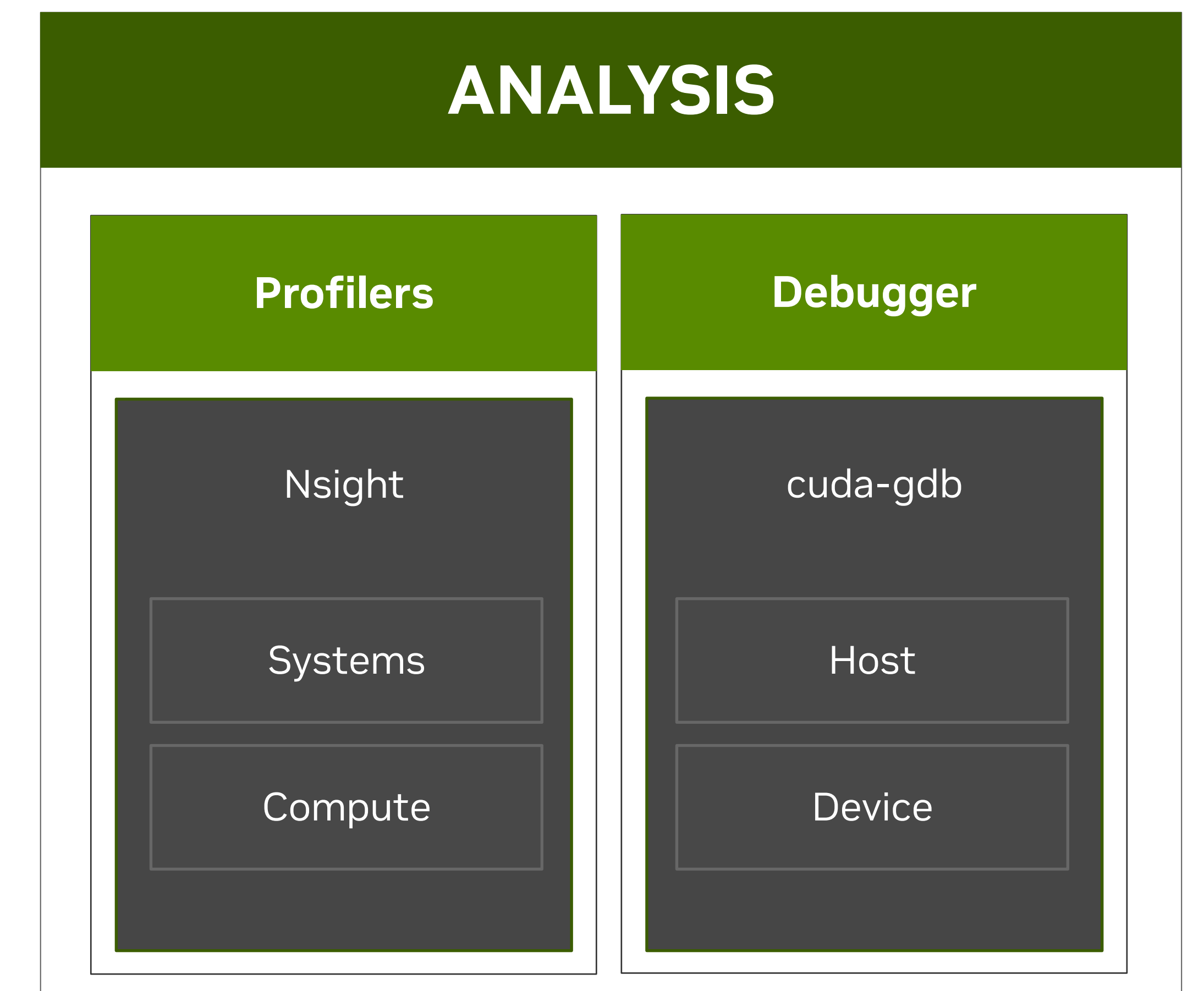
NVIDIA HPC SDK

Available at developer.nvidia.com/hpc-sdk, on NGC, via Spack, and in the Cloud

DEVELOPMENT



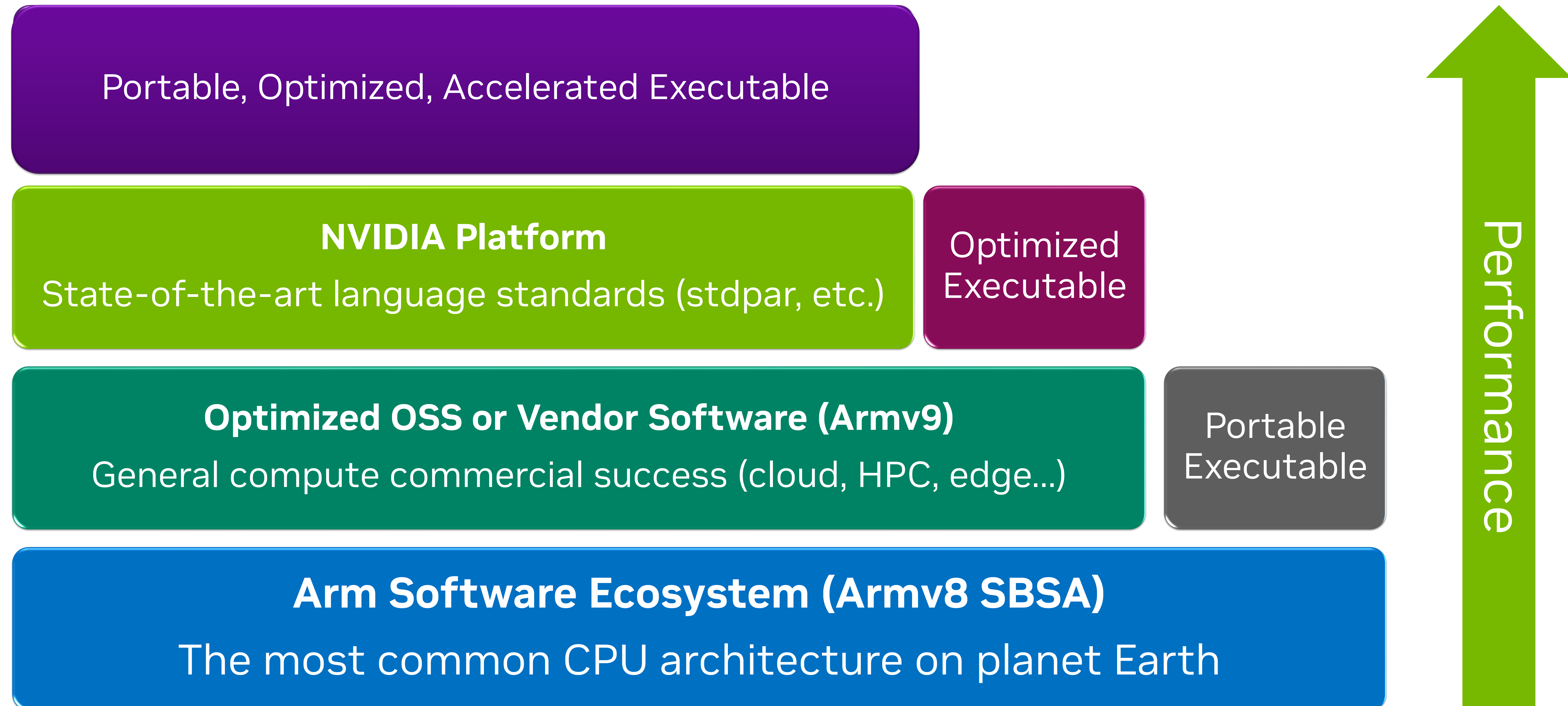
ANALYSIS



Develop for the NVIDIA Platform: GPU, CPU and Interconnect
Libraries | Accelerated C++ and Fortran | Directives | CUDA
x86_64 | Arm
6 Releases Per Year | Freely Available

Grace Software Ecosystem is Built on Standards + NVIDIA's Ecosystem

Grace brings the full NVIDIA software stack to Arm



Advancing the State-of-the-Art in Compilers

NVIDIA invests in open source and commercial compilers for NVIDIA Grace

- **NVIDIA HPC Compilers**

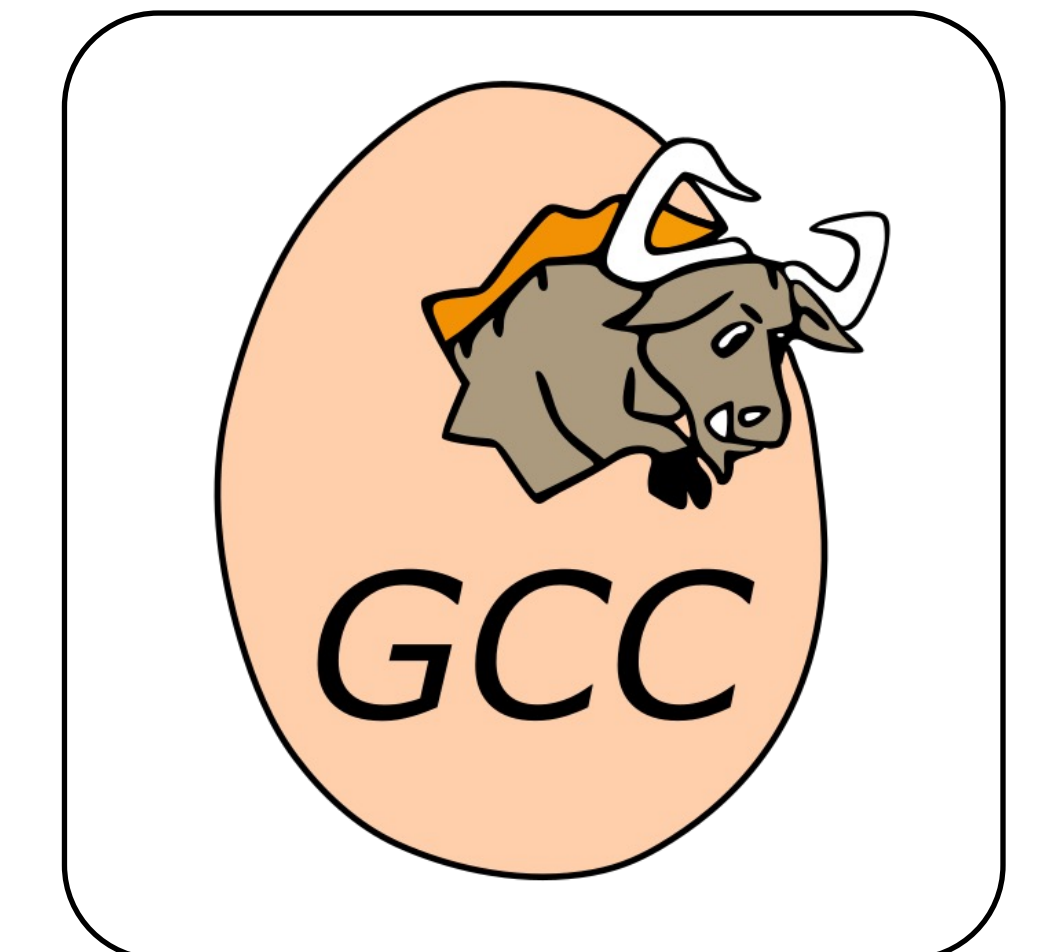
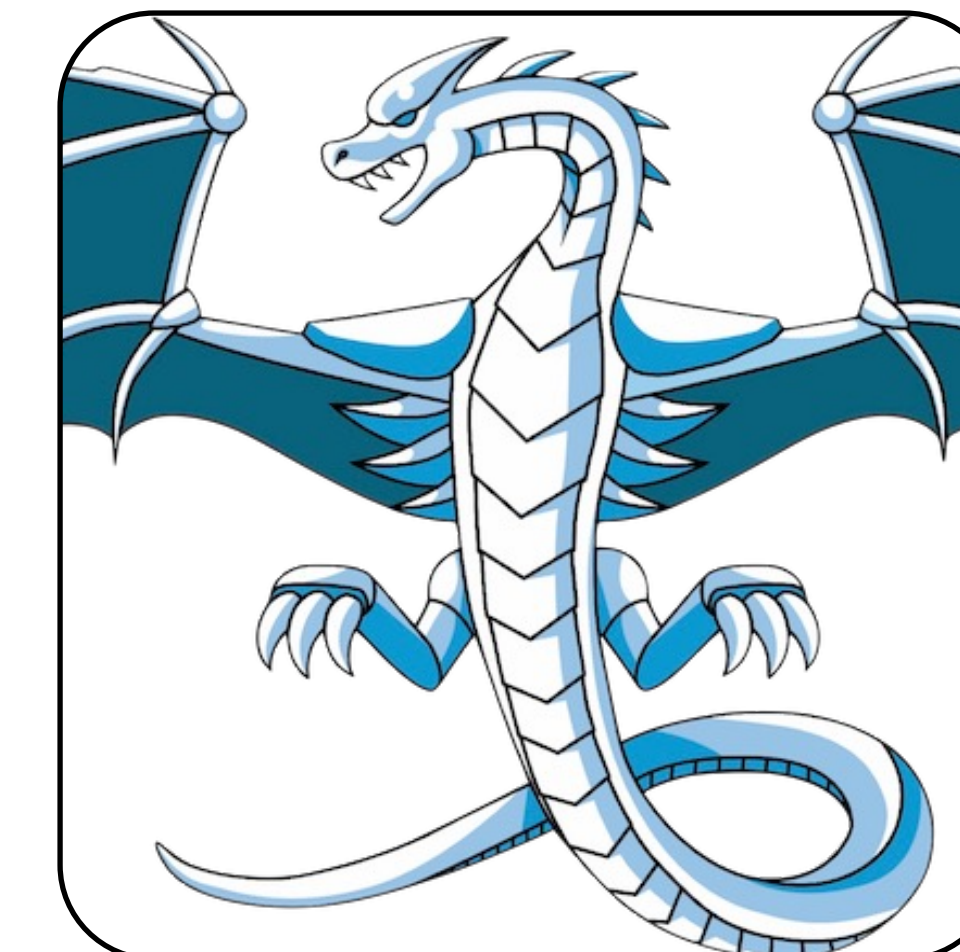
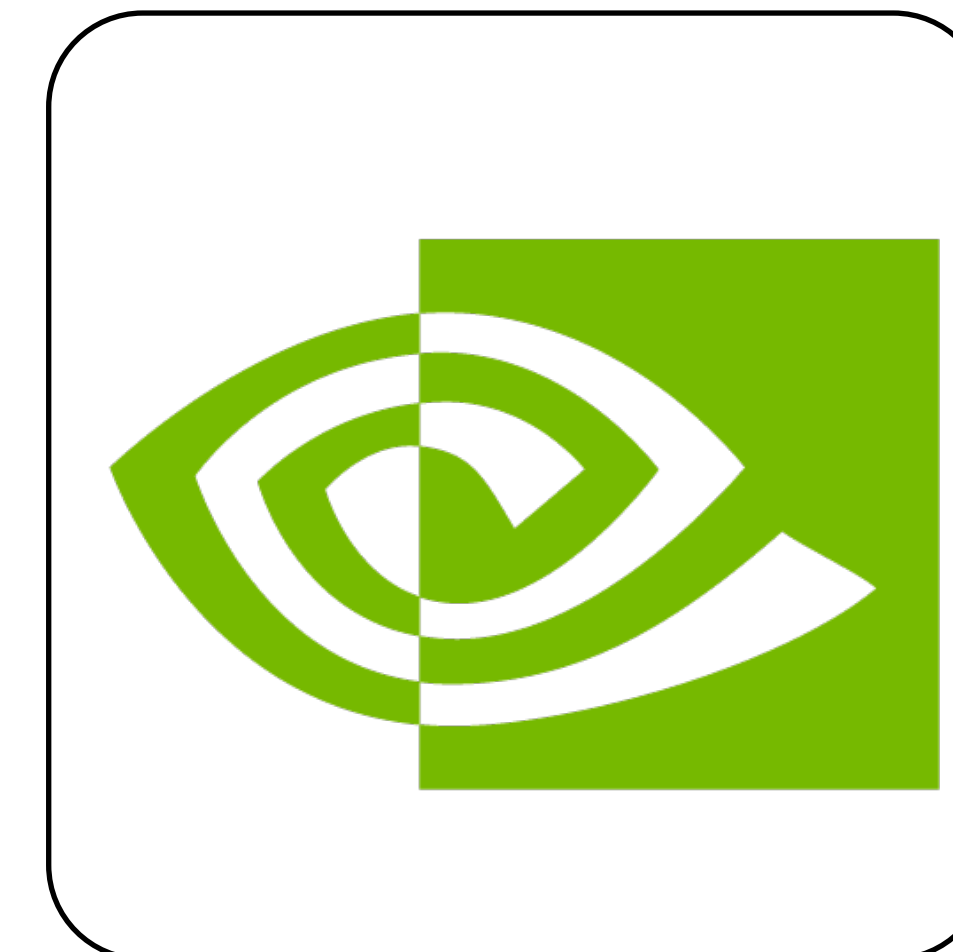
- Focused on application performance and programmer productivity
- High velocity, constant innovation
- Freely available with commercial support option

- **LLVM and Clang**

- NVIDIA provides builds of Clang for Grace
 - <https://developer.nvidia.com/grace/clang>
- Drop-in replacement for mainline Clang
- 100% of Clang enhancements for Grace are contributed to mainline LLVM

- **GCC**

- NVIDIA contributes to mainline GCC to support Grace
- Working with all major Linux distros to improve availability of Grace optimizations in GCC



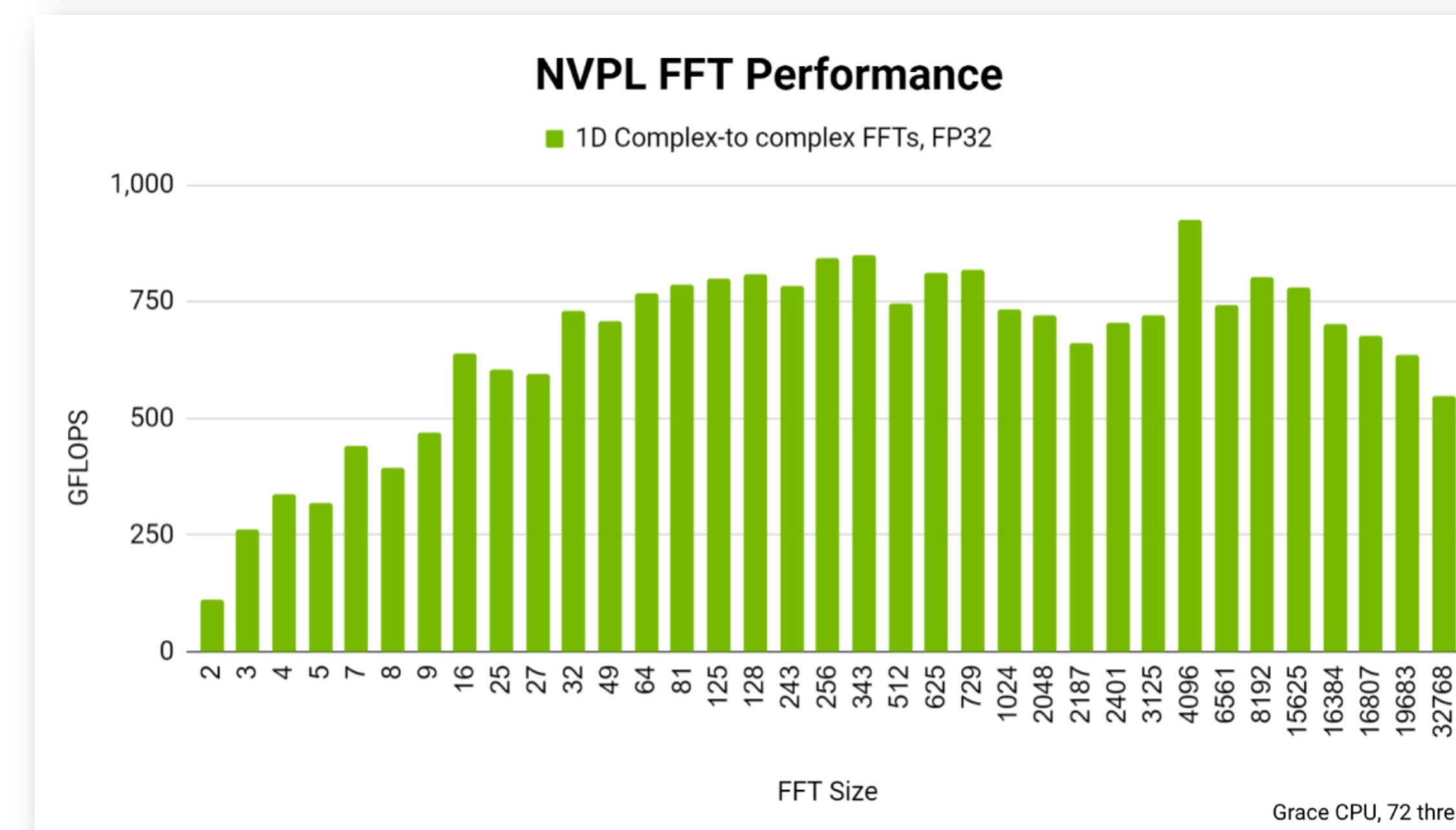
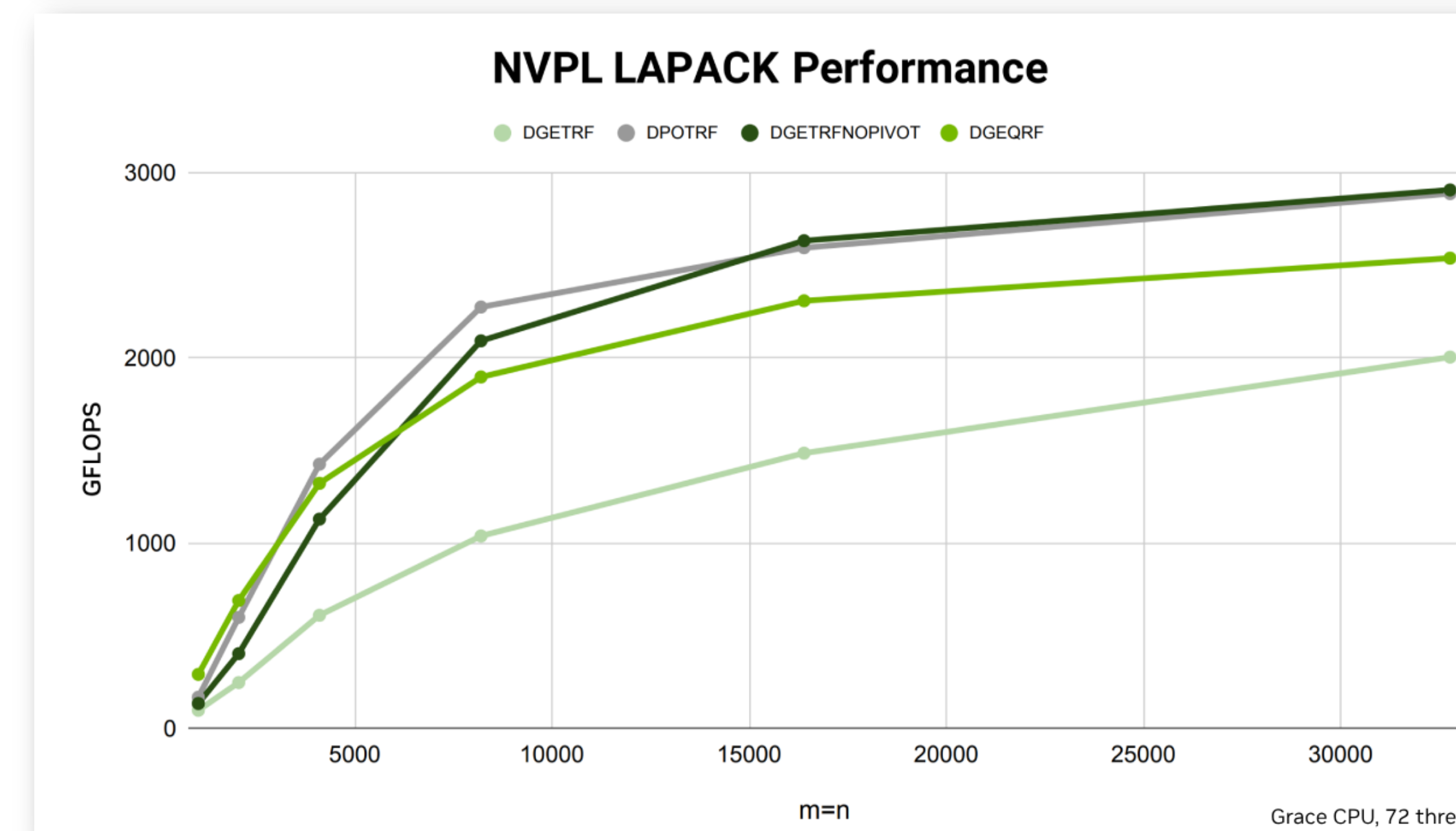
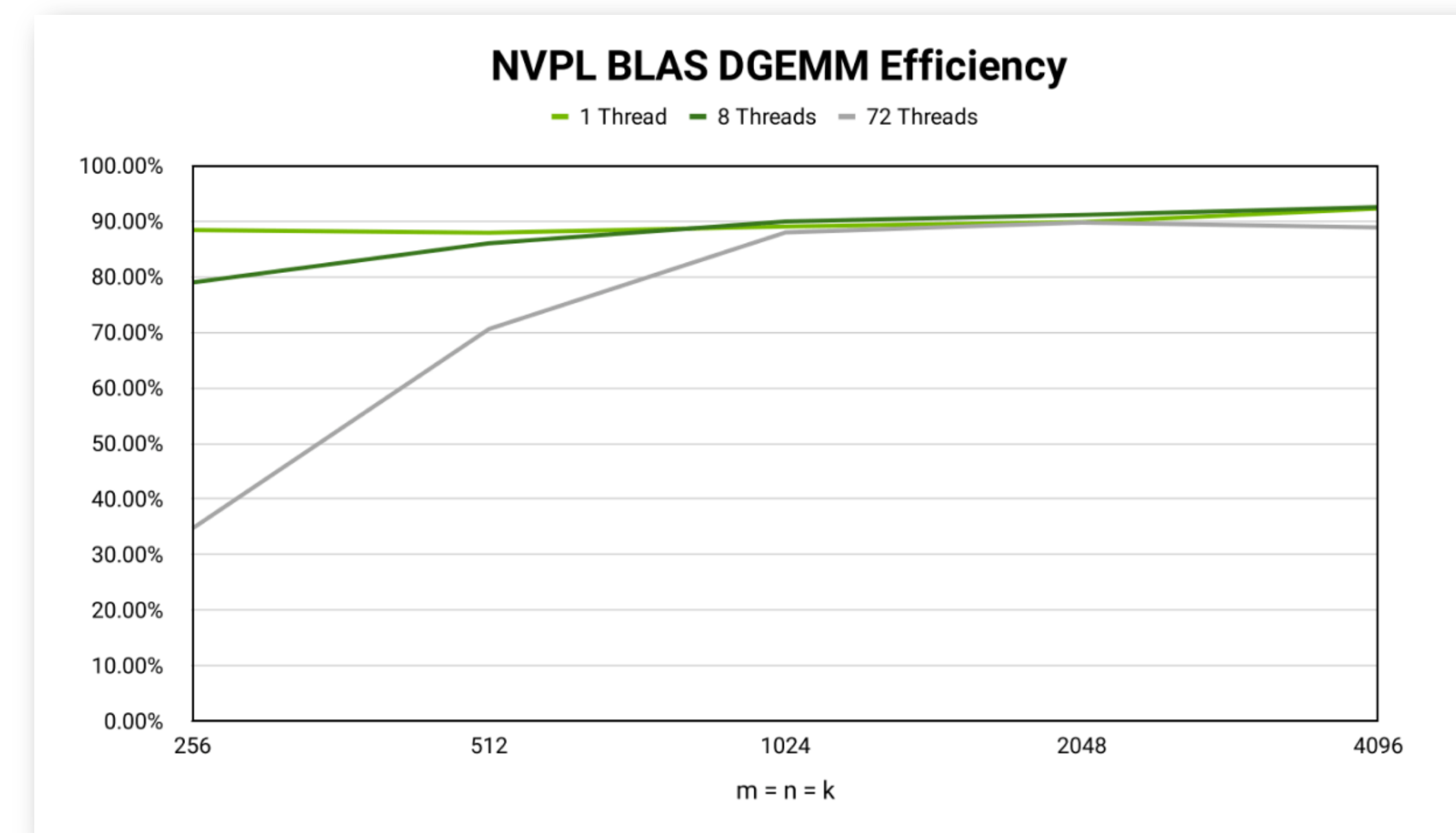
NVIDIA Performance Libraries (NVPL)

Optimized math libraries for NVIDIA CPUs

- Easily port applications to NVIDIA's Arm CPUs
- Drop-in replacement for any math library implementing standard interfaces (e.g. Netlib, FFTW)
- New interfaces for high-performance libraries

| | | | |
|--------|--------|-------|-----------|
| BLAS | LAPACK | PBLAS | SCALAPACK |
| TENSOR | SPARSE | RAND | FFT |

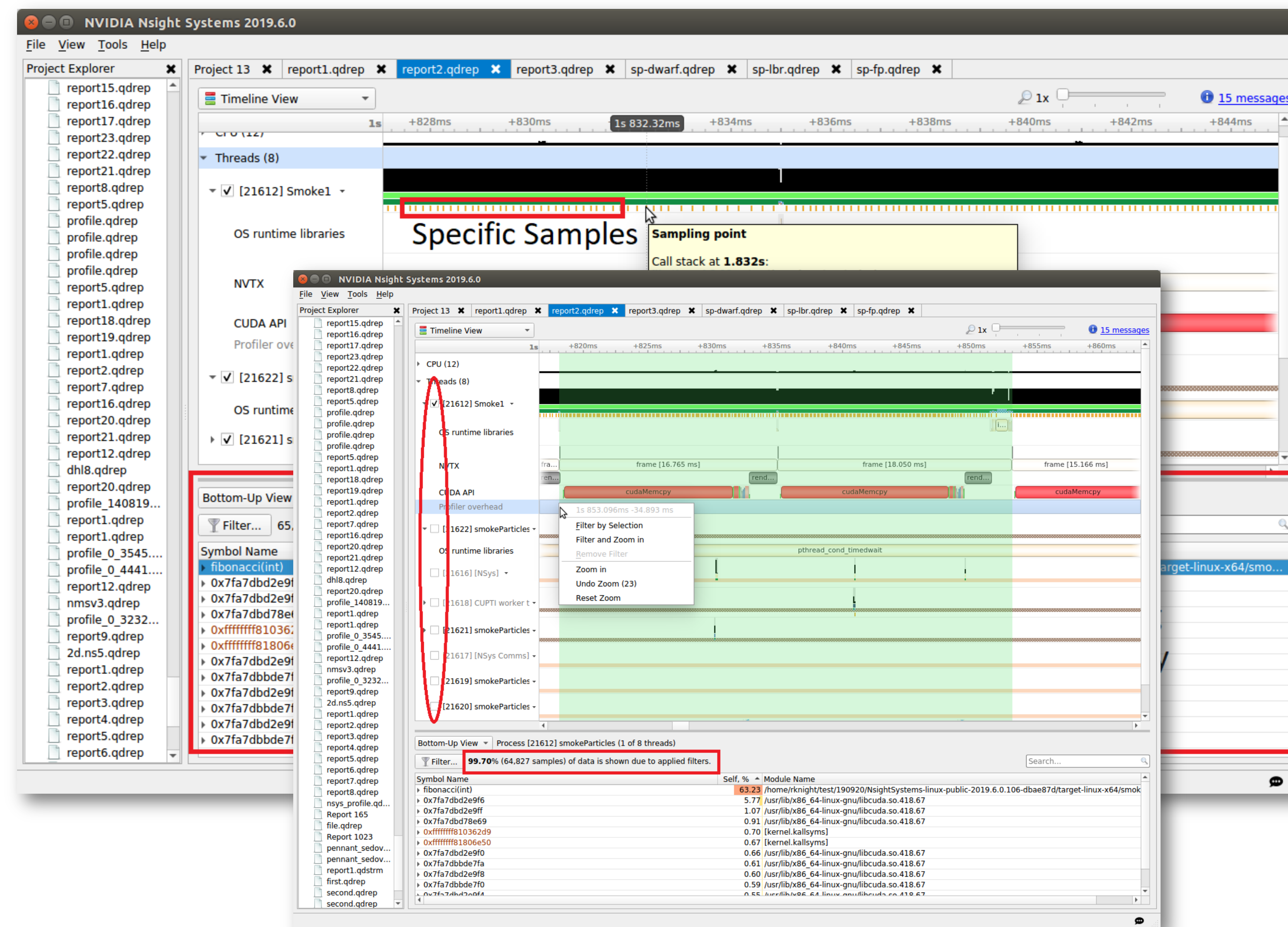
Download Now
www.developer.nvidia.com/nvpl



Debuggers and Profilers for GH200 and Grace CPU Superchip

Full capability on Grace-Hopper

- **NVIDIA Nsight has full feature-parity on GH200**
 - Anything you can do with Nsight tools on x86+Hopper, you can do on GH200 with the same workflow
- **GH200 has hundreds of performance counters (PMUs)**
 - **Computational intensity, bandwidth, instruction mix...**
- **Generally, all major debugging and profiling tools for x86+Hopper are available on GH200**
 - Similar capabilities are provided by other tools on Grace

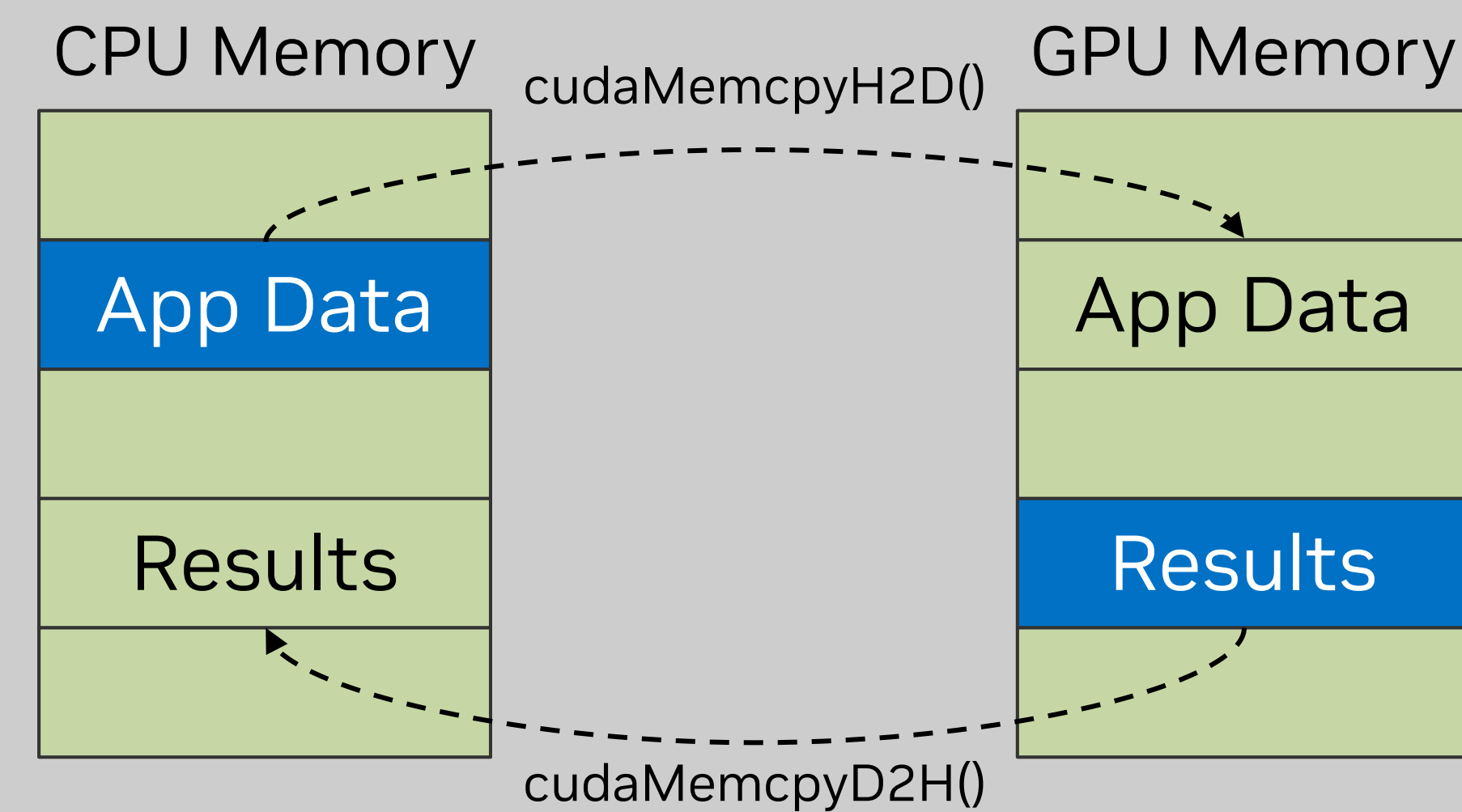


The Grace Hopper Advantage

Full CUDA support with additional Grace memory extensions

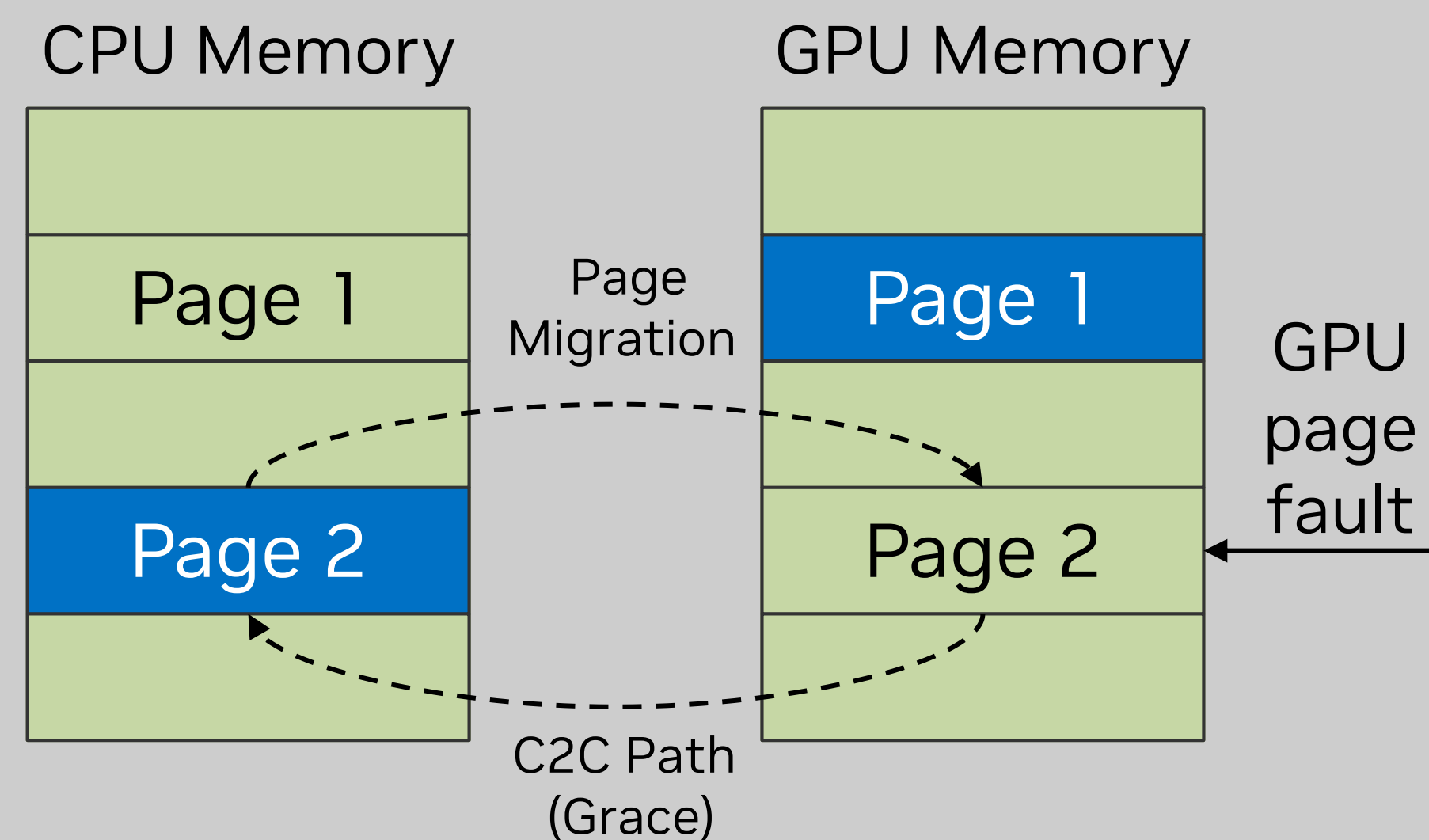
Explicit Copy

Application explicitly moves data between CPU & GPU as needed



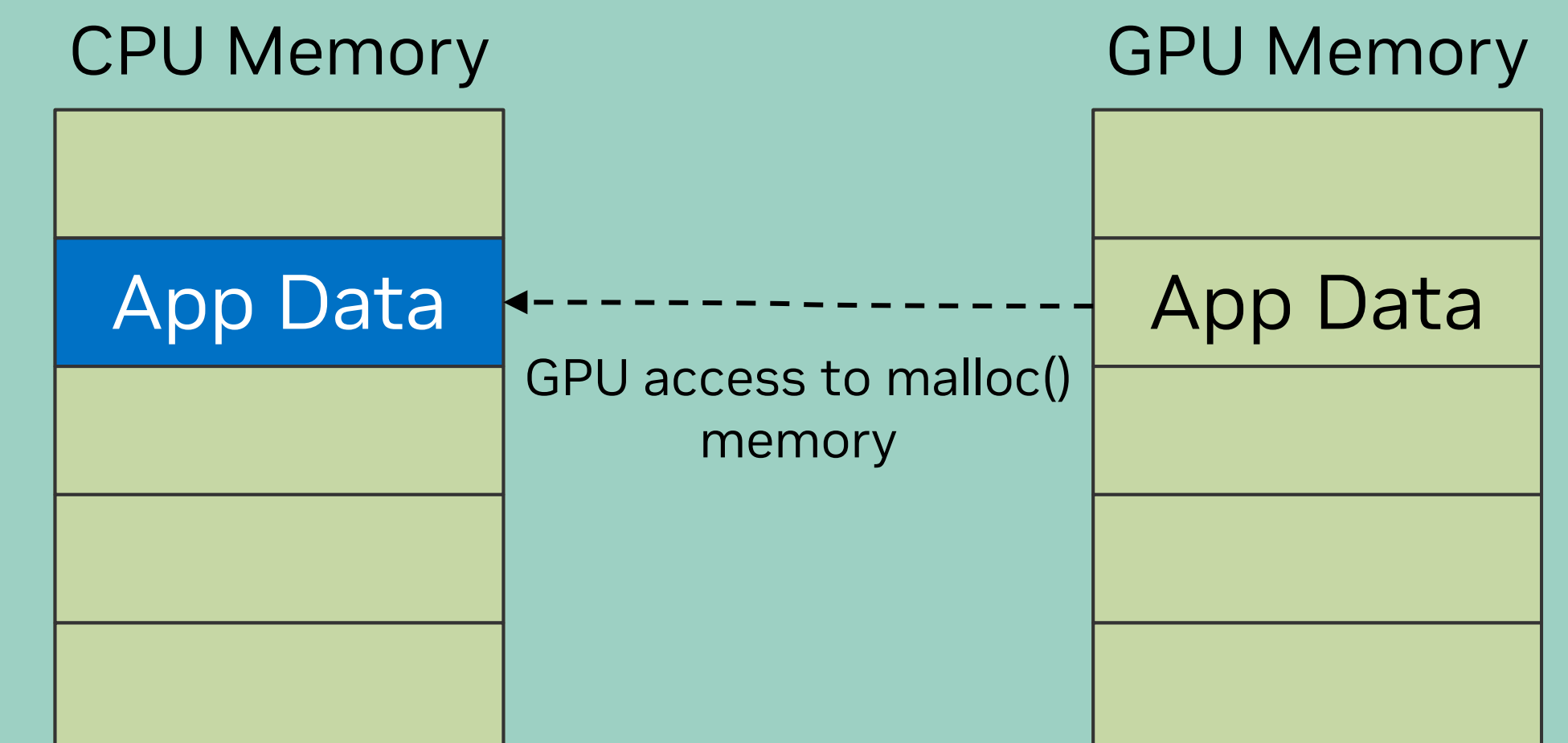
Managed Memory

CPU and GPU can access memory on-demand and data migrated locally for higher BW access



System Allocated

GPU can access memory allocated from `malloc()`, `mmap()`, etc.



HGX

~60 GB/s PCIe Gen5 transfers (H2D/D2H)

Requires migration to GPU

Access possible with explicit call to `cudaHostRegister()` at PCIe speeds
Requires HMM patch in Linux Kernel

G+H

7x faster transfers, up to 450 GB/s (NVLink C2C)

Migrations not required and faster migrations when they happen at NVLink C2C speed

`cudaHostRegister()` not needed; access at NVLink C2C speeds

CUDA Explicit Memory Allocators

Maximum **portable** performance Out-of-the-box

- **No programming model changes!**
 - No new APIs
 - No changes to existing APIs
 - No source code changes
- **Unified Memory**
 - Available on *most* platforms supported by CUDA 12.x: GH, P9+V100, PCIe x86 & Arm, etc.
 - Same Unified Memory Programming Model for all platforms
 - "memory accesses just work" + "hints".
- Unified Memory **Hints**
 - *Hints* only impact performance, not results.
 - **cudaMemAdvise** hints: PreferredLocation, AccessedBy.
 - **cudaMemPrefetch** hints: prefetch to NUMA node.
 - Work with all memory, e.g., including malloc.

| Memory | Placement | Access-based Migration | Accessible From | |
|---|----------------------------|------------------------|-----------------|------|
| | | | CPU | GPUs |
| System-allocated (malloc, mmap) | First-touch (GPU CPU) | ✓ | ✓ | ✓ |
| CUDA managed (cudaMallocManaged) | | ✓ | ✓ | ✓ |
| CUDA device memory (cudaMalloc) | GPU | ✗ | ✗ | ✓ |
| CUDA host memory (cudaMallocHost) | CPU | ✗ | ✓ | ✓ |
| <i>...and many others: interprocess, virtual, fabric, ...</i> | | | | |

CUDA Unified Memory Hints

```
cudaMemAdvise(ptr, nbytes, advice, device);
```

| | | | |
|----------------|-------------------|------------|------------|
| AdVICES | PreferredLocation | AccessedBy | ReadMostly |
|----------------|-------------------|------------|------------|

| | | |
|----------------|--------|-----|
| DEVICES | GPU id | CPU |
|----------------|--------|-----|

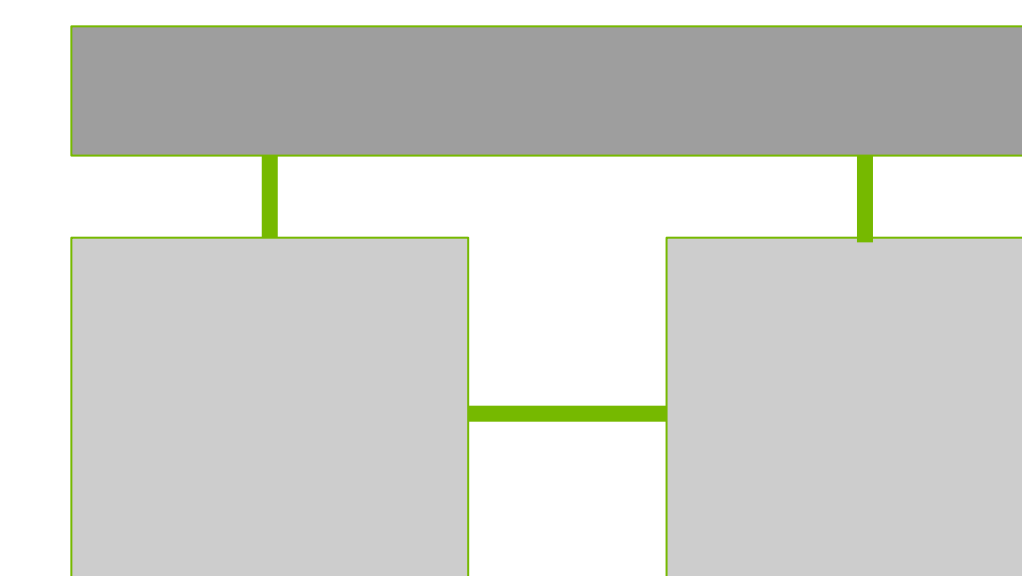
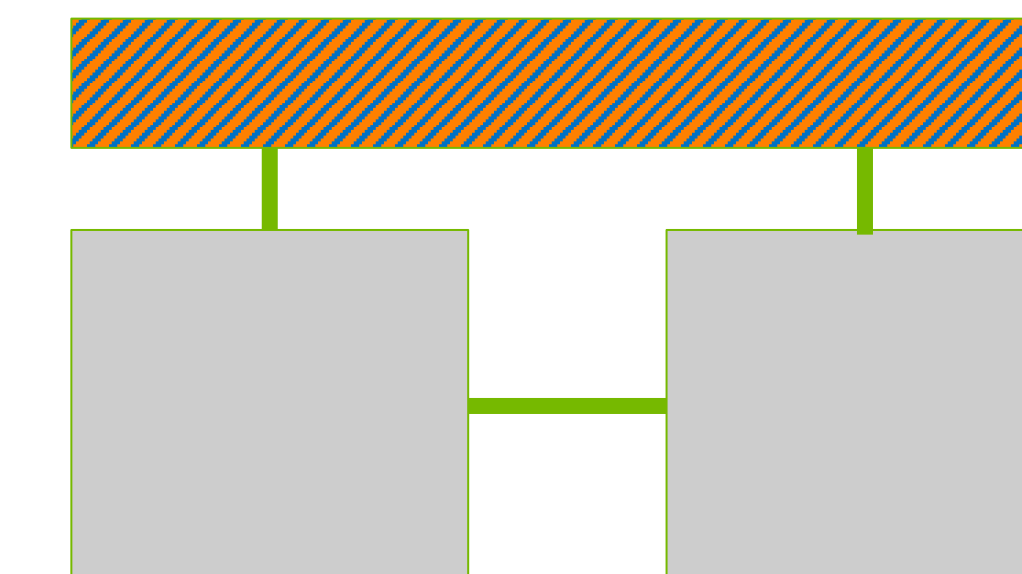
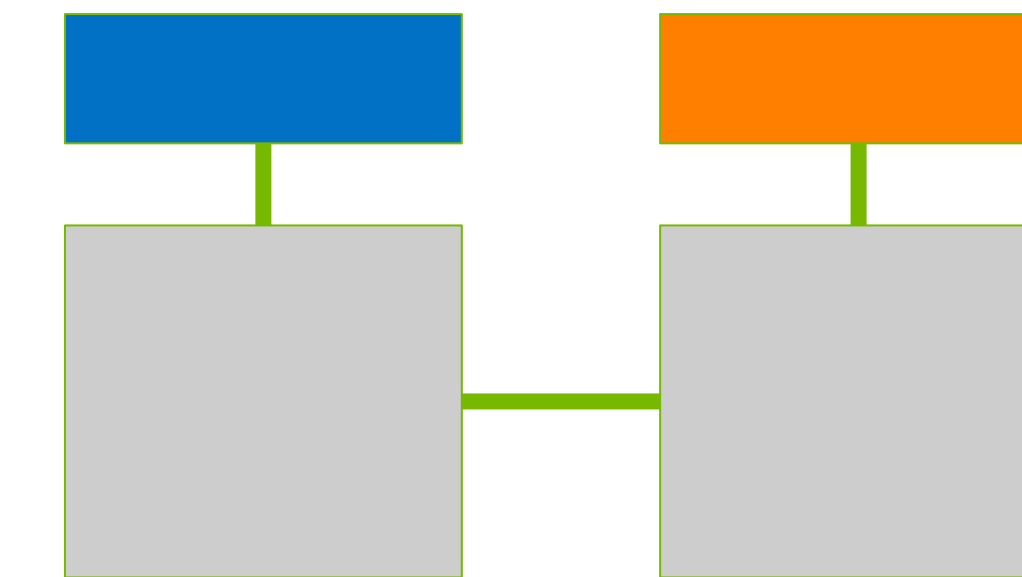
```
cudaMemPrefetchAsync(ptr, nbytes, destination);
```

| | | |
|---------------------|--------|-----|
| DESTINATIONS | GPU id | CPU |
|---------------------|--------|-----|

NVHPC Compilers GPU Memory Model

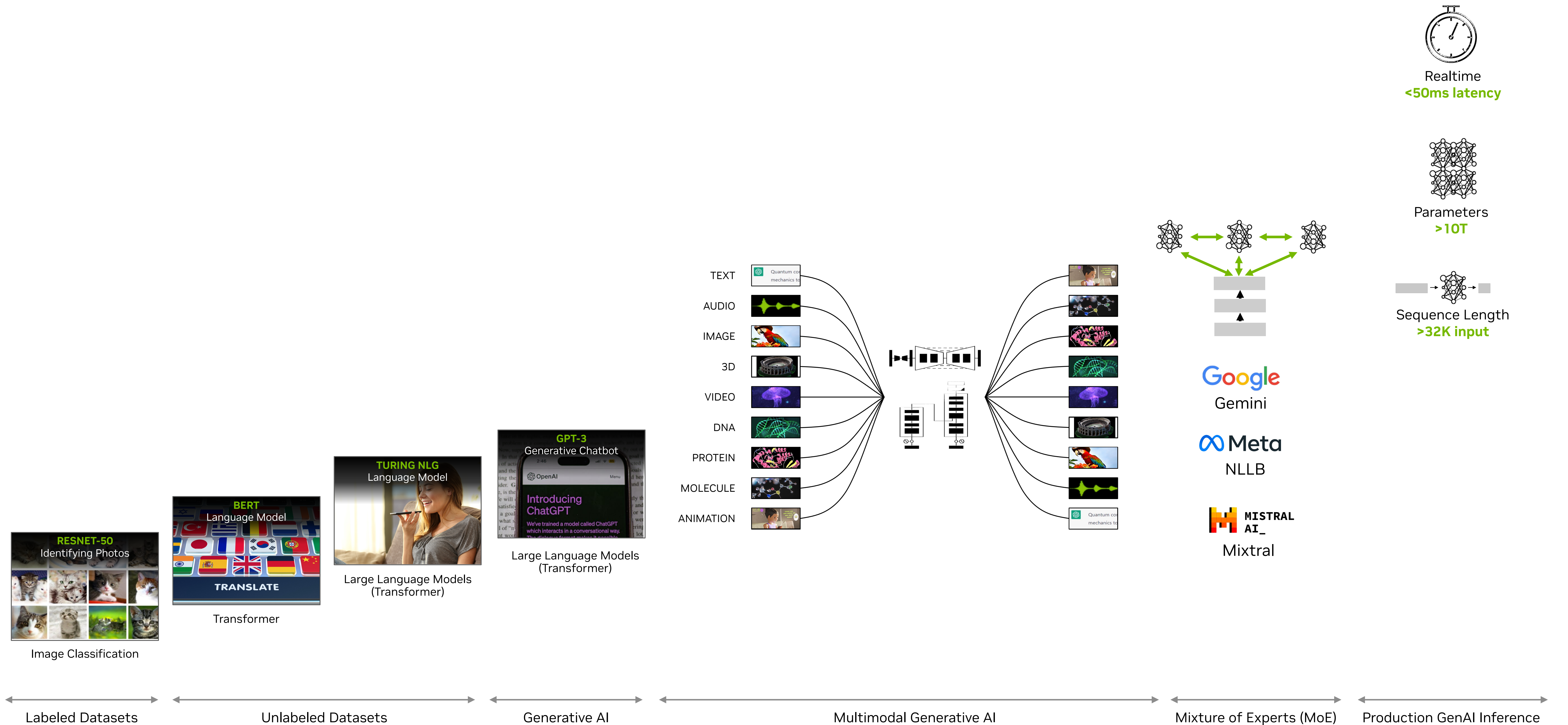
Abstraction over HW simplifying GPU programming

- **Separate** - CPU and GPU have distinct memories when data are shared between the two explicit copy is needed
- **Managed** (-gpu=managed) - CPU and GPU have a single address space for **dynamically-allocated** data, data is migrated automatically on-demand. Other memory remains separate.
 - Only dynamically allocated data are shared between CPU and GPU
 - Stack and global variables outside of parallel algorithm code Can't be accessed in parallel algorithm
- **Unified** (-gpu=unified) - CPU and GPU have single address space which allows accessing **all data** locations from both processors, data may be migrated or accessed in-place.
 - All CPU data are accessible from the GPU utilizing full CUDA Unified Memory (with ATS/HMM)
 - This mode may or may not also utilise CUDA Managed Memory



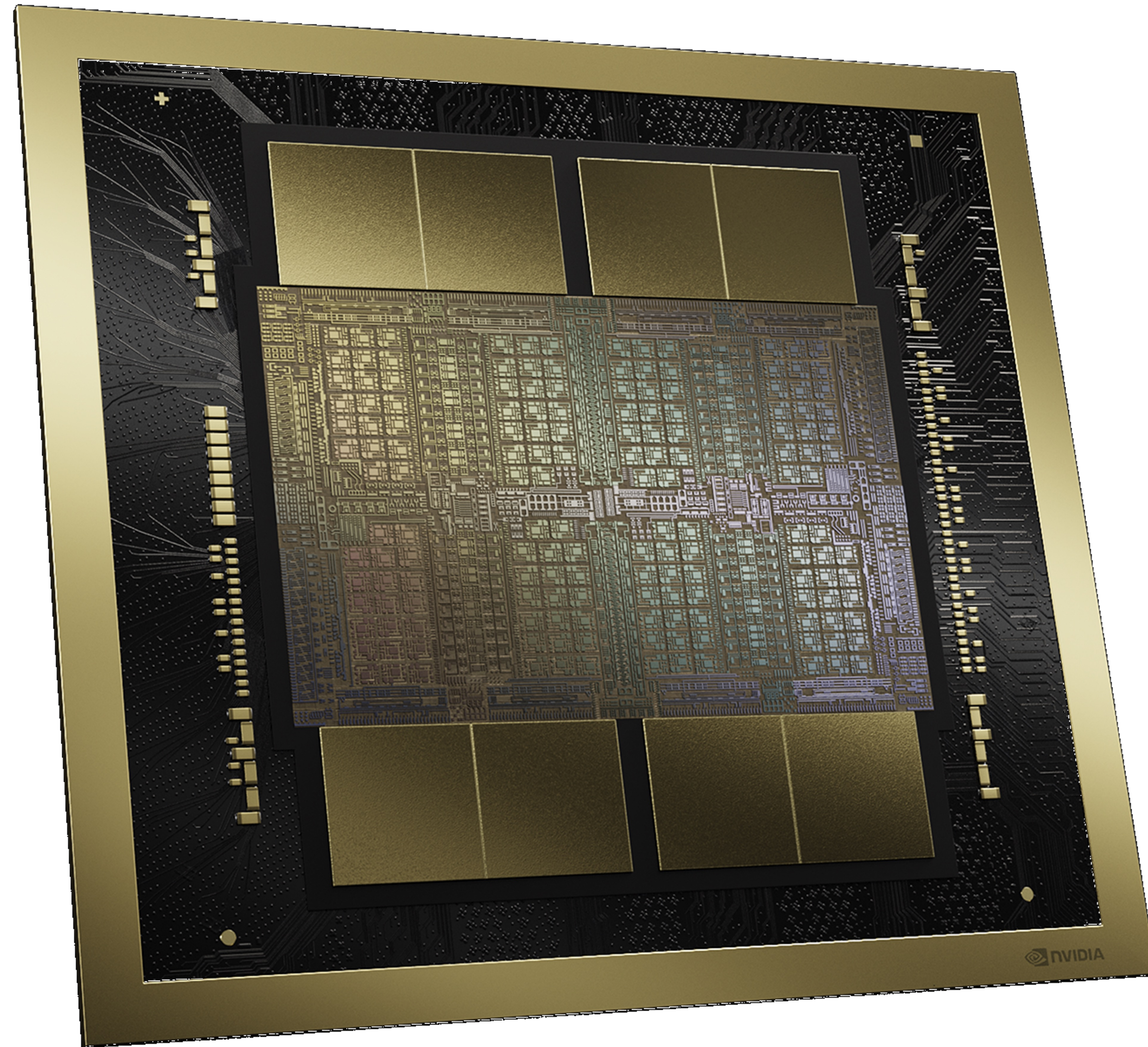
Next gen: NVIDIA Blackwell

The Next Era of Generative AI



Announcing NVIDIA Blackwell

The Engine of the New Industrial Revolution

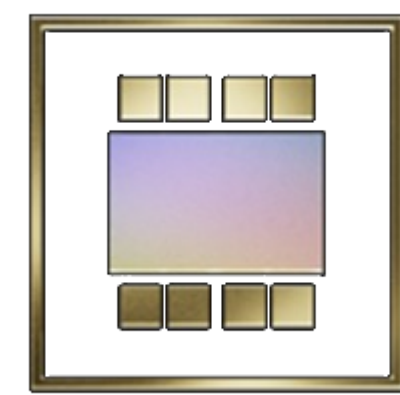


Built to Democratize Trillion-Parameter AI

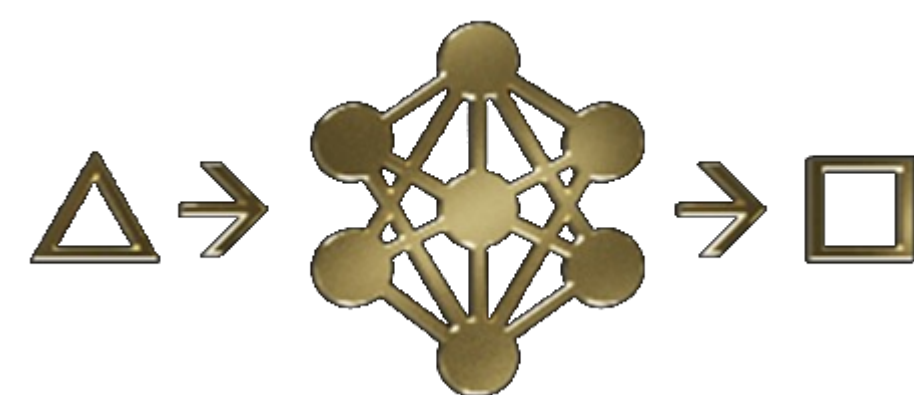
20 PetaFLOPS of AI performance on a single GPU

4X Training | 30X Inference | 25X Energy Efficiency & TCO

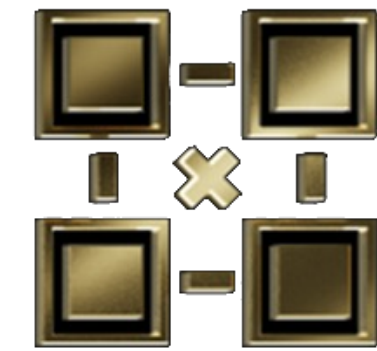
Expanding AI Datacenter Scale to beyond 100K GPUs



AI SUPERCHIP
208B Transistors



2nd GEN TRANSFORMER ENGINE
FP4/FP6 Tensor Core



5th GENERATION NVLINK
Scales to 576 GPUs



RAS ENGINE
100% In-System
Self-Test



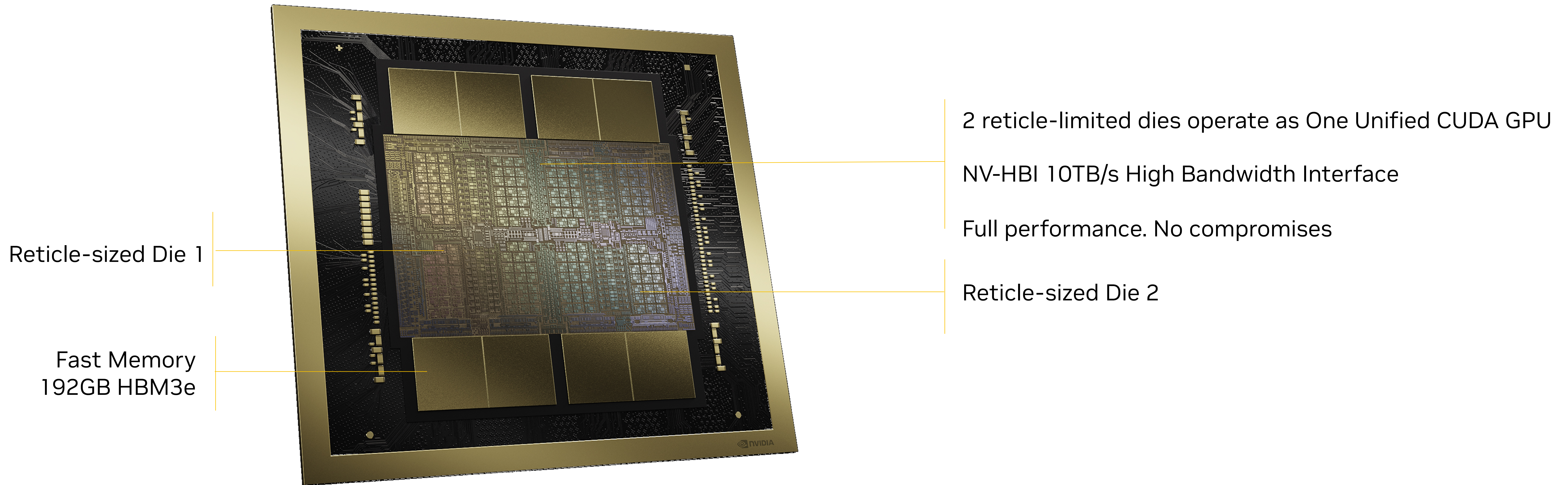
SECURE AI
Full Performance
Encryption & TEE



DECOMPRESSION ENGINE
800 GB/s

New Class of AI Superchip

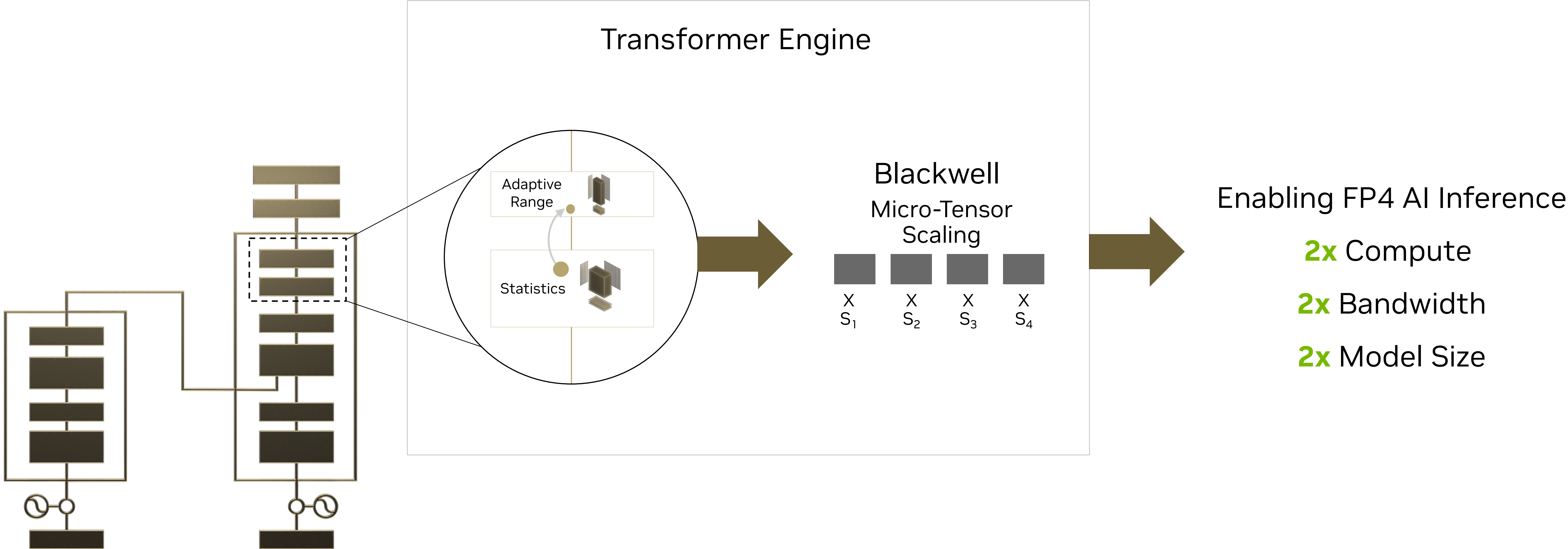
The Two Largest Dies Possible—Unified as One GPU



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

2nd Generation Transformer Engine

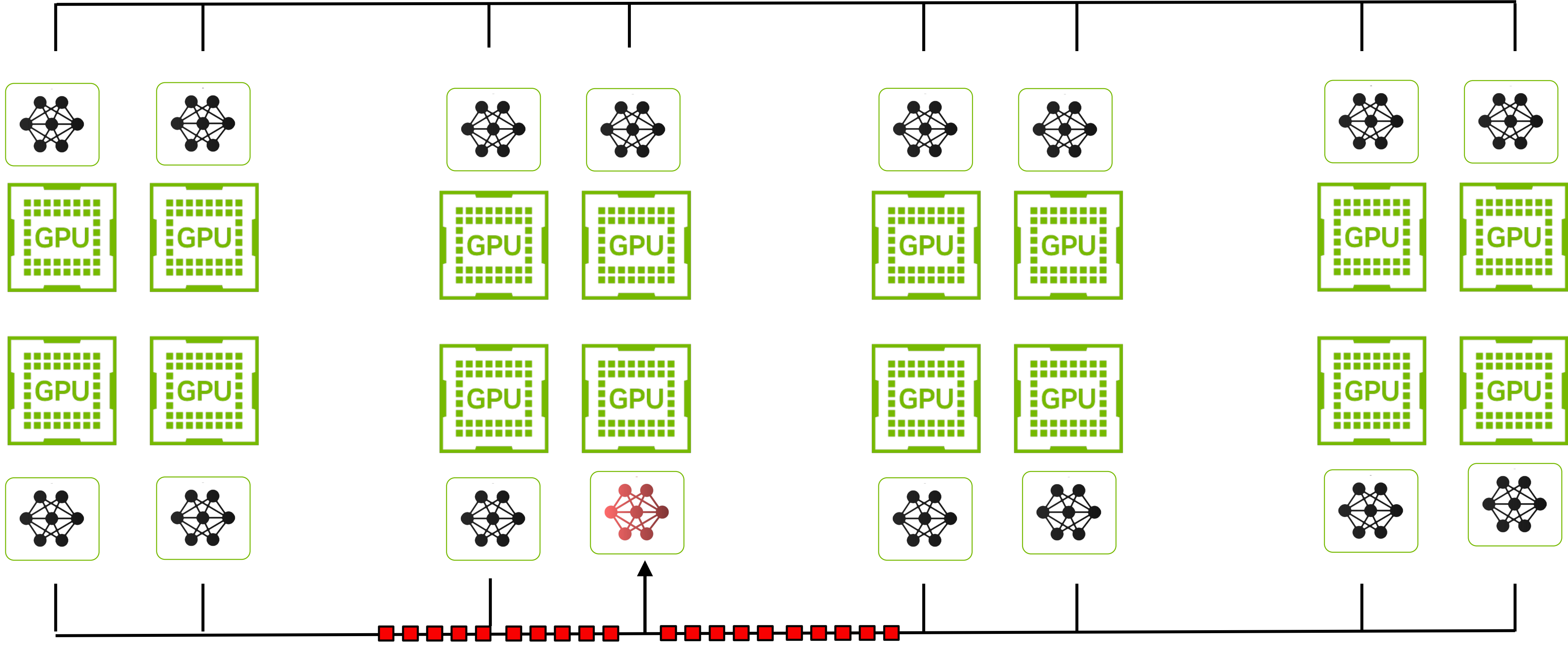
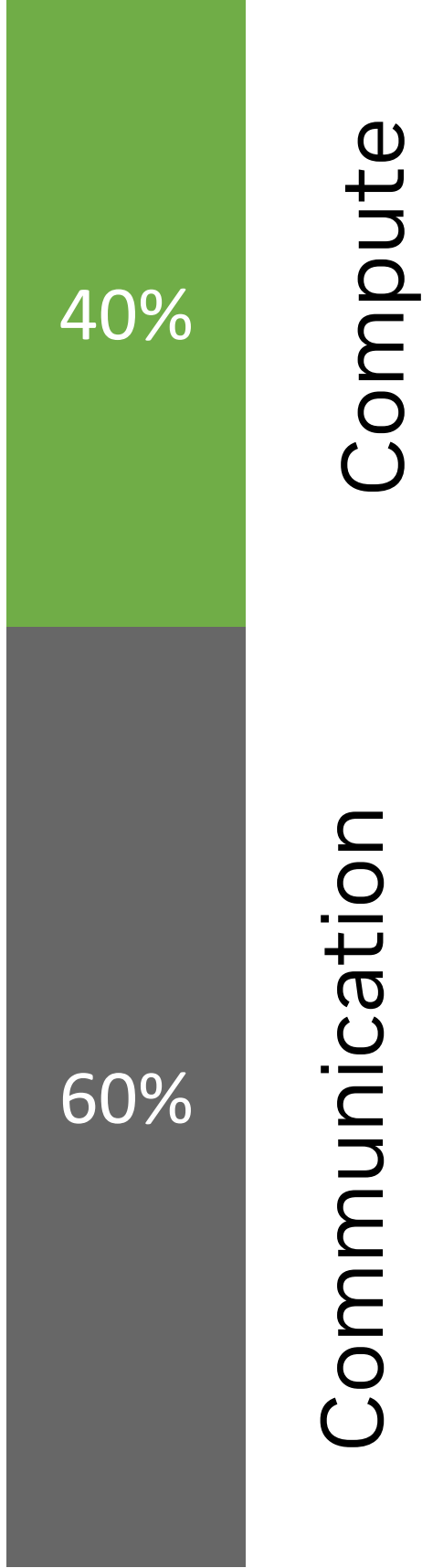
Accelerating Throughput with Intelligent 4-Bit Precision



Next Generation Models Communication Bottleneck

Mixture of Expert Models

GPT MoE 1.8T Parameters

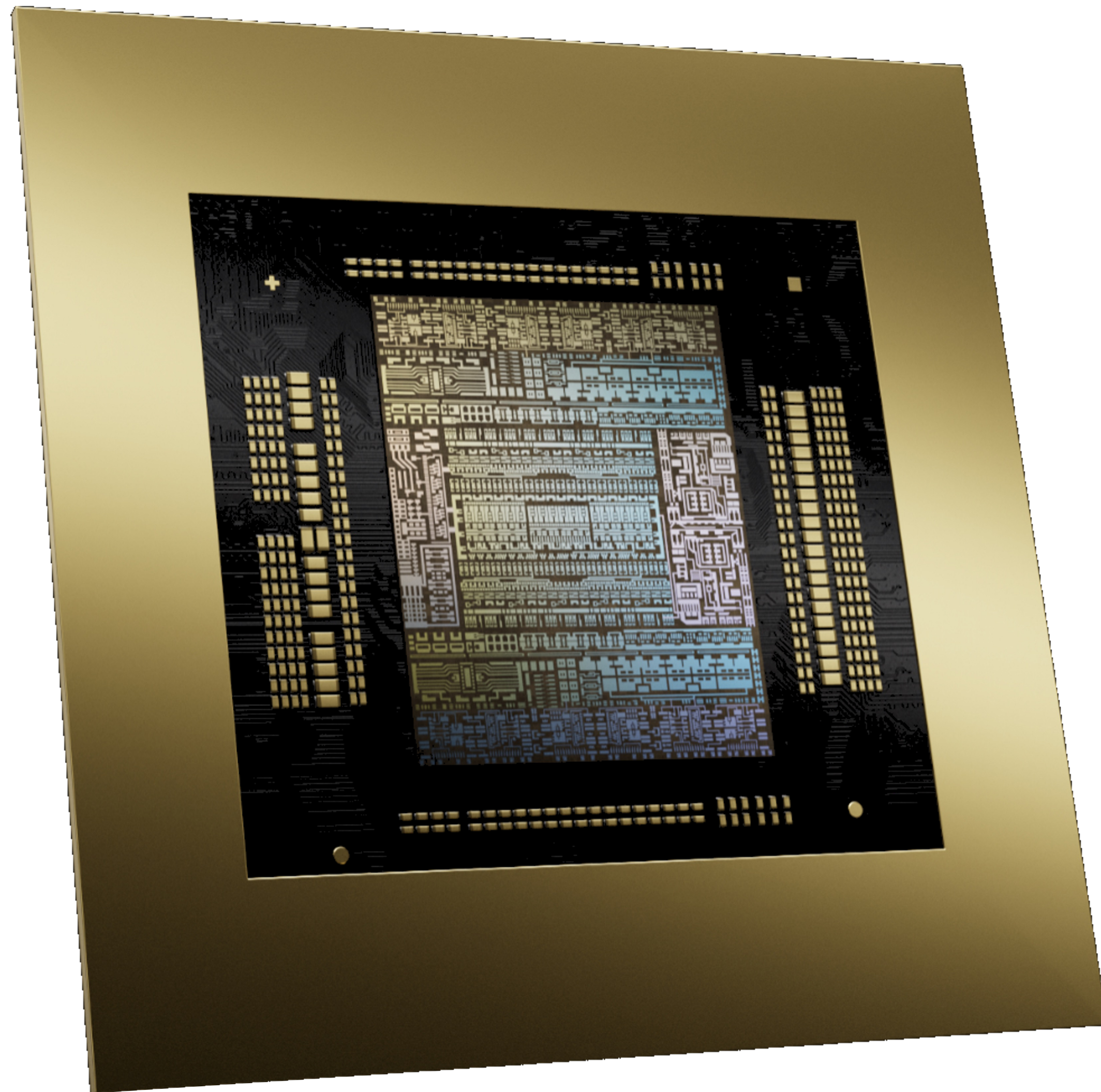


HDR InfiniBand
100 GByte/s

15 GPUs Sending
to 1 GPU

Announcing Fifth Generation NVLink and NVLink Switch Chip

Efficient Scaling for Trillion Parameter Models



7.2 TB/s Full all-to-all Bidirectional Bandwidth

Sharp v4 plus FP8

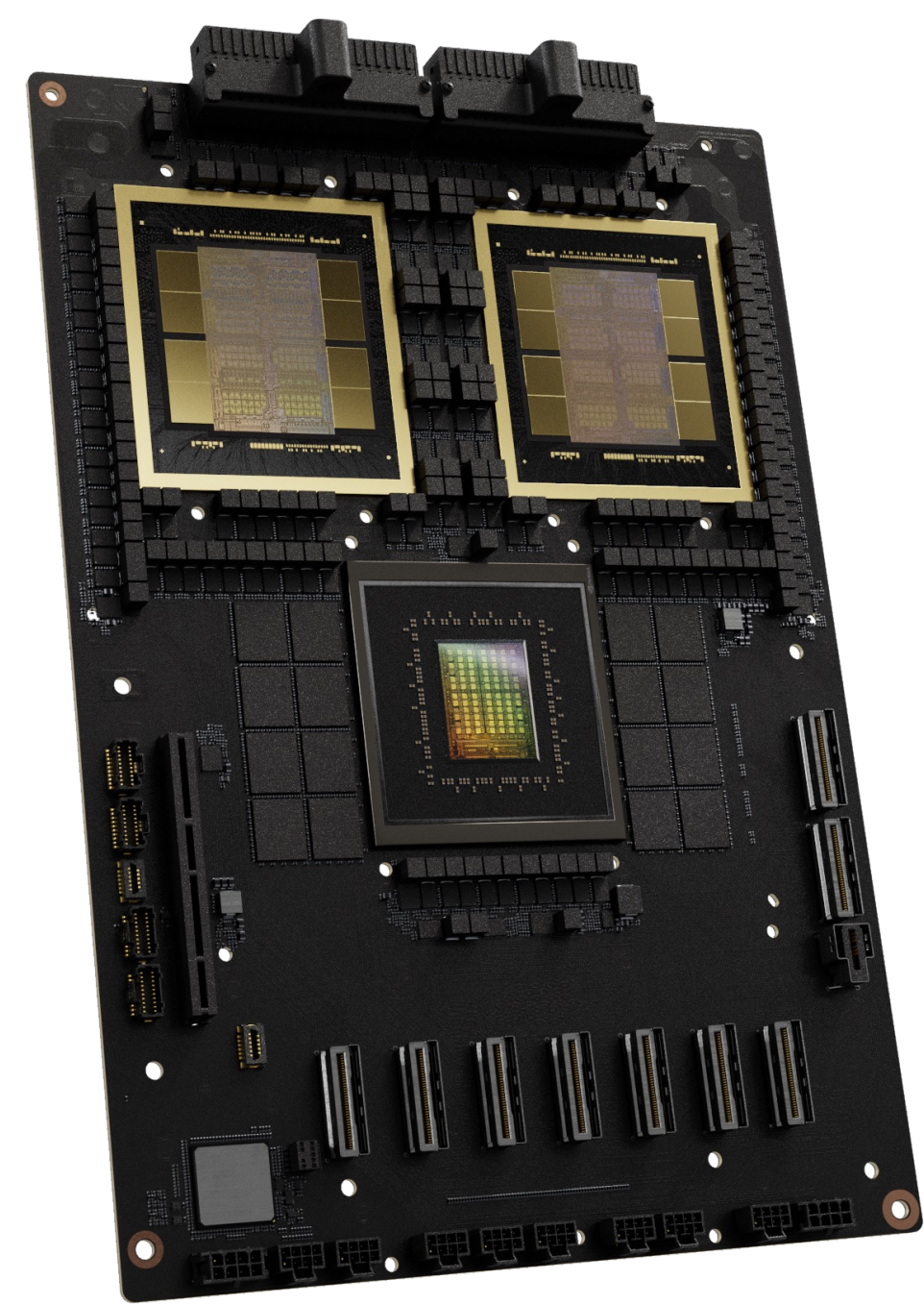
3.6 TF In-Network Compute

Expanding NVLink up to 576 GPU NVLink Domain

18X Faster than Today's Multi-Node Interconnect

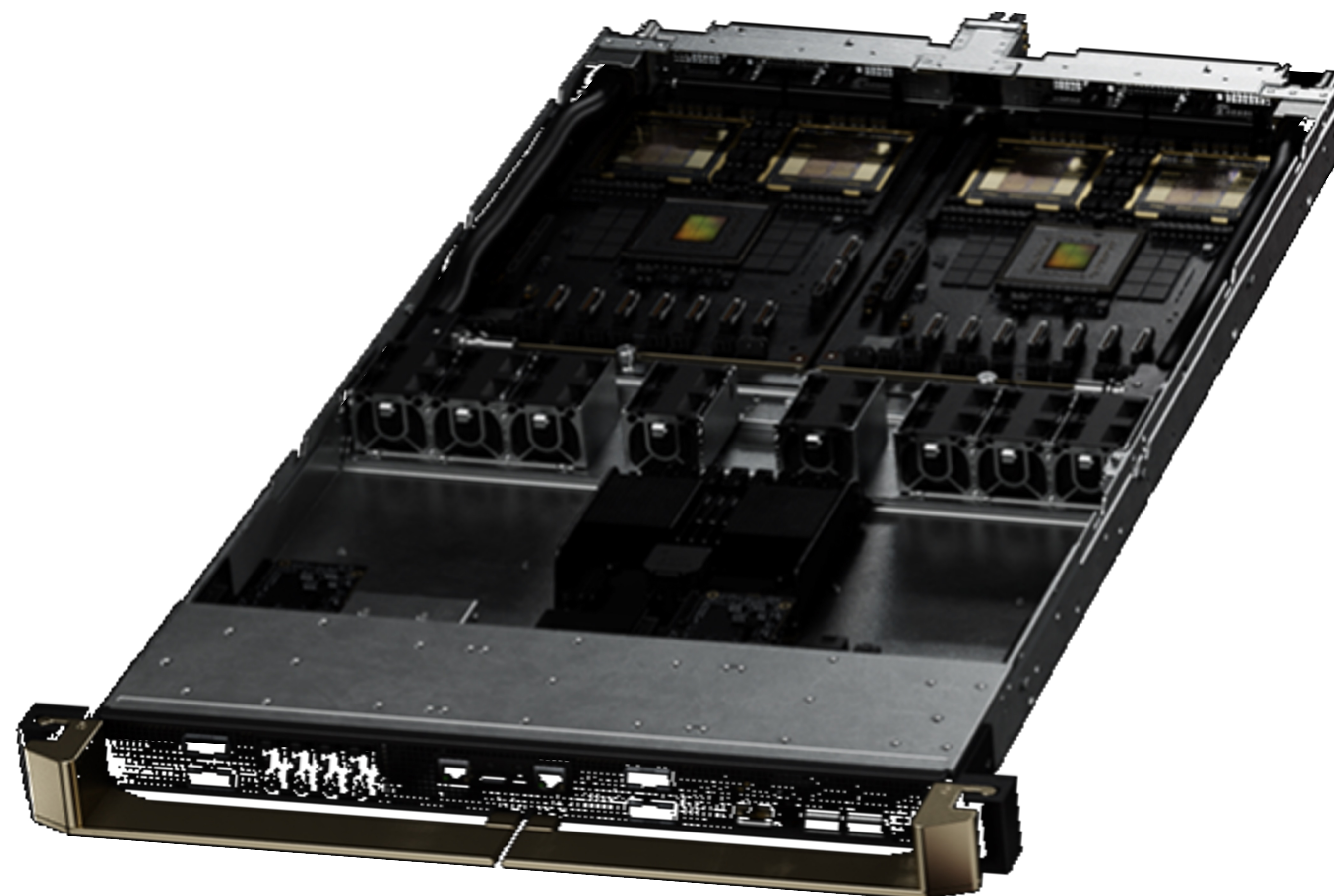
GB200 NVL72 Compute and Interconnect Nodes

Building Blocks for the GB200 NVL72 Rack



GB200 SUPERCHIP

40 PETAFLUPS FP4 AI INFERENCE
20 PETAFLUPS FP8 AI TRAINING
864GB FAST MEMORY



GB200 SUPERCHIP COMPUTE TRAY

2x GB200
80 PETAFLUPS FP4 AI INFERENCE
40 PETAFLUPS FP8 AI TRAINING
1728 GB FAST MEMORY
1U Liquid Cooled
18 Per Rack



NVLINK SWITCH TRAY

2x NVLINK SWITCH CHIP
14.4 TB/s Total Bandwidth
SHARV4 FP64/32/16/8
1U Liquid Cooled
9 Per Rack

Announcing 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 for Every Generative AI Use Case

Delivering the New Era of Performance for Every Data Center



GB200 NVL72

Compute for Trillion Parameter Scale AI
Maximum Performance and Lowest TCO



HGX B200

Best Performance and TCO for HGX Platform

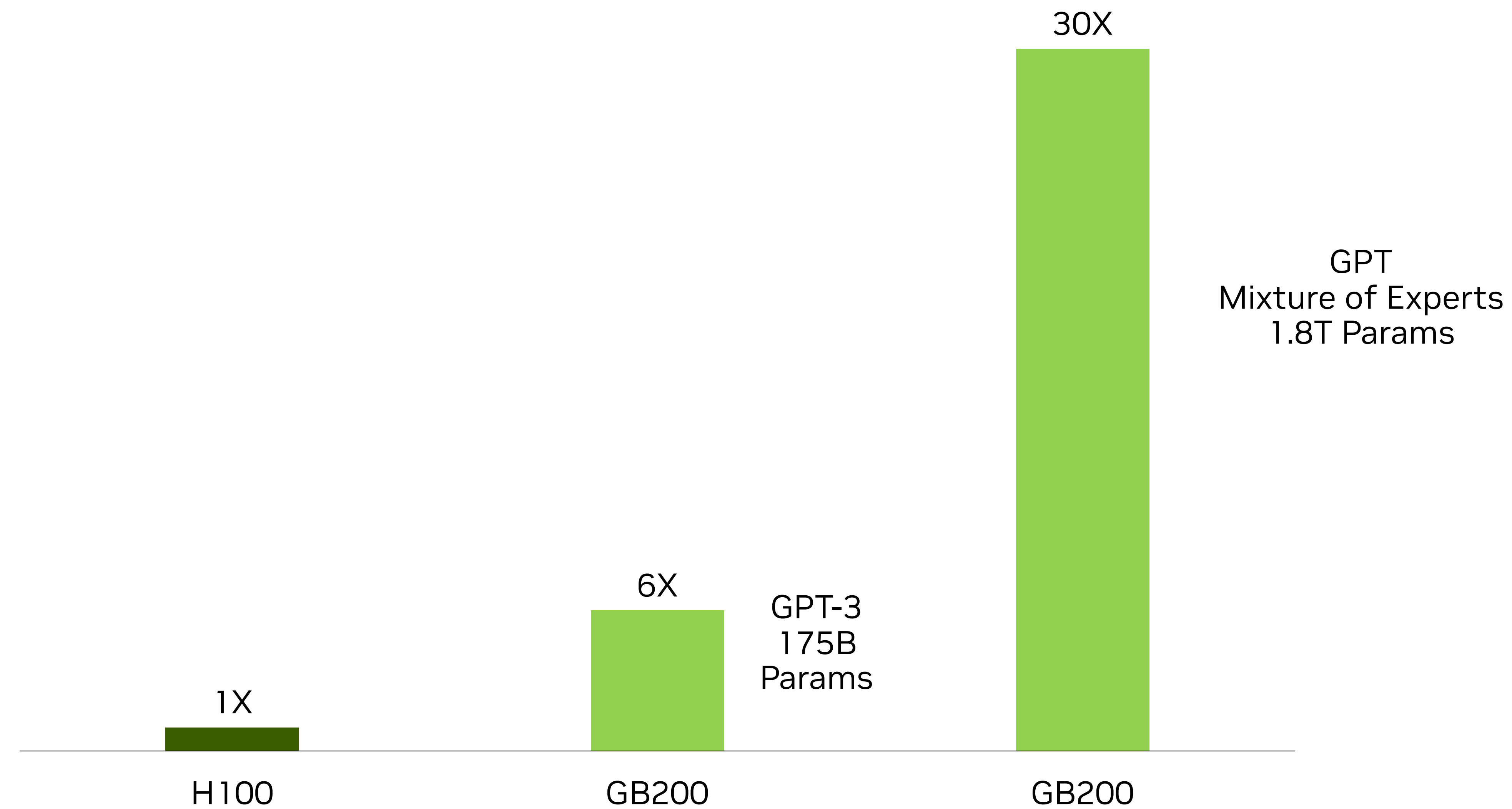


HGX B100

Drop-in Upgrade for Existing Hopper Infrastructure

GB200 NVL72 Enabling Trillion Parameter AI

30x Realtime Mixture of Experts Inference, 25X Improved Energy Efficiency



Projected performance subject to change

Token-to-token latency (TTL) = 50 milliseconds (ms) real time

GPT-3 175B: First token latency (FTL) 2s; input sequence length = 2,048, output sequence length = 128, 4 HGX H100 air-cooled 400GB IB Network vs 2 GB200 Superchips liquid-cooled NVLink; per GPU performance comparison,

GPT-MoE-1.8T: FTL = 5s; input sequence length = 32,768, output sequence length = 1,024, 8 HGX H100 air-cooled 400GB IB Network vs 18 GB200 Superchips liquid-cooled NVL36; per GPU performance comparison

Blackwell Ecosystem

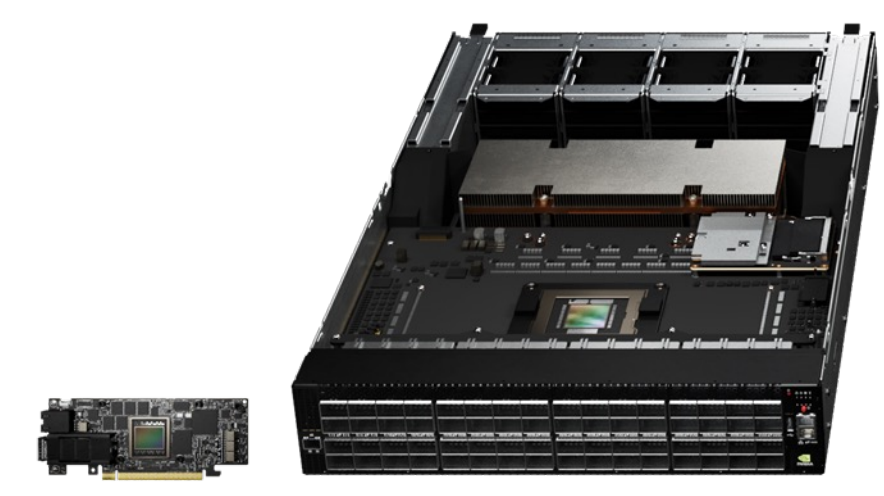
Coming Later 2024



Google Cloud



ORACLE CLOUD Infrastructure



Spectrum-X800



Quantum-X800



GB200 NVL72



HGX B200



HGX B100

