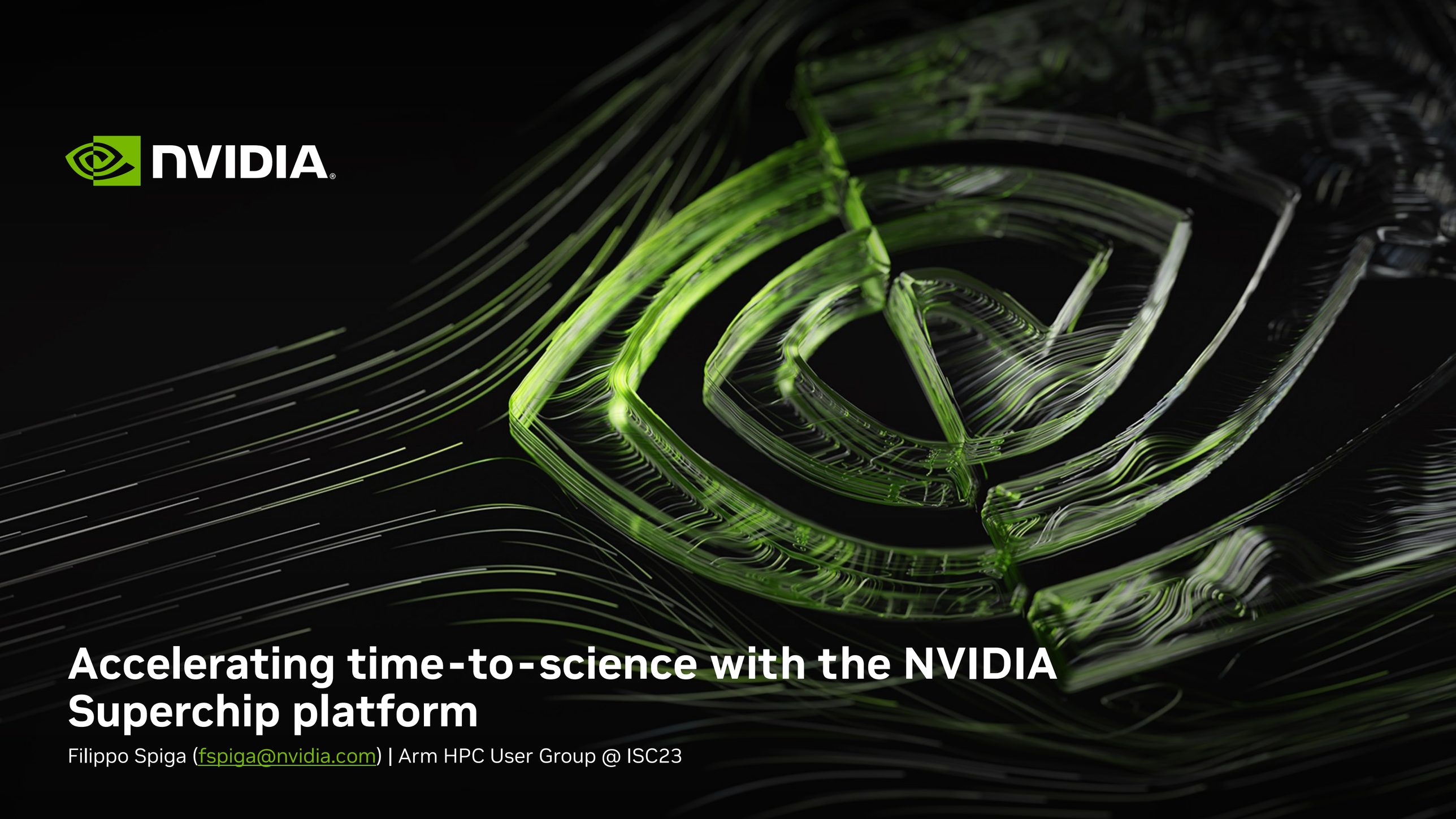


Accelerating time-to-science with the NVIDIA Superchip platform

Filippo Spiga (fspiga@nvidia.com) | Arm HPC User Group @ ISC23

NVIDIA Grace for Cloud, AI and HPC Infrastructure

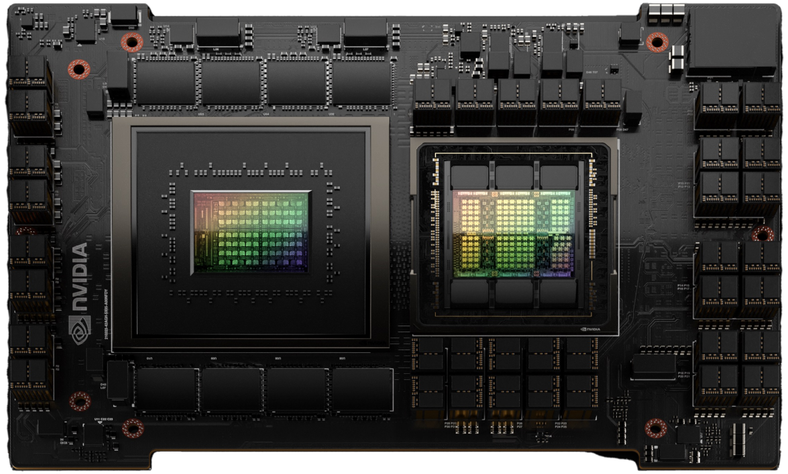
Grace CPU Superchip

CPU Computing

CPU-based applications where absolute performance, energy efficiency, and data center density matter, such as scientific computing, data analytics, enterprise and hyperscale computing applications

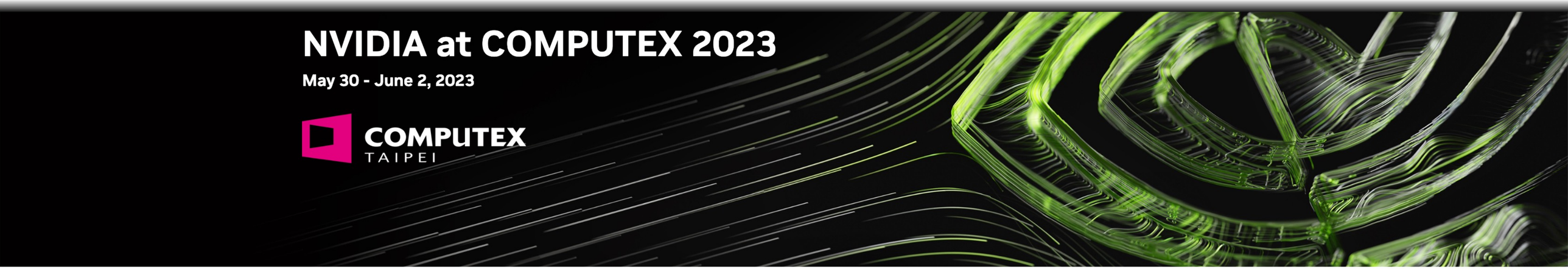
Grace Hopper Superchip

Large Scale AI & HPC



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

NEW ANNOUNCEMENTS NEXT WEEK



NVIDIA Keynote at COMPUTEX 2023

Monday, May 29, 2023 at 11:00 a.m. Taipei Time

Sunday, May 28, 2023 at 8:00 p.m. Pacific Time

Join NVIDIA founder and CEO Jensen Huang at COMPUTEX 2023 for a special keynote address streaming online.

Save the Date

SCAN ME

https://www.nvidia.com/en-us/events/computex/

NVIDIA

NVIDIA Grace CPU

Building Block of the Superchip

High Performance Power Efficient Cores

72 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

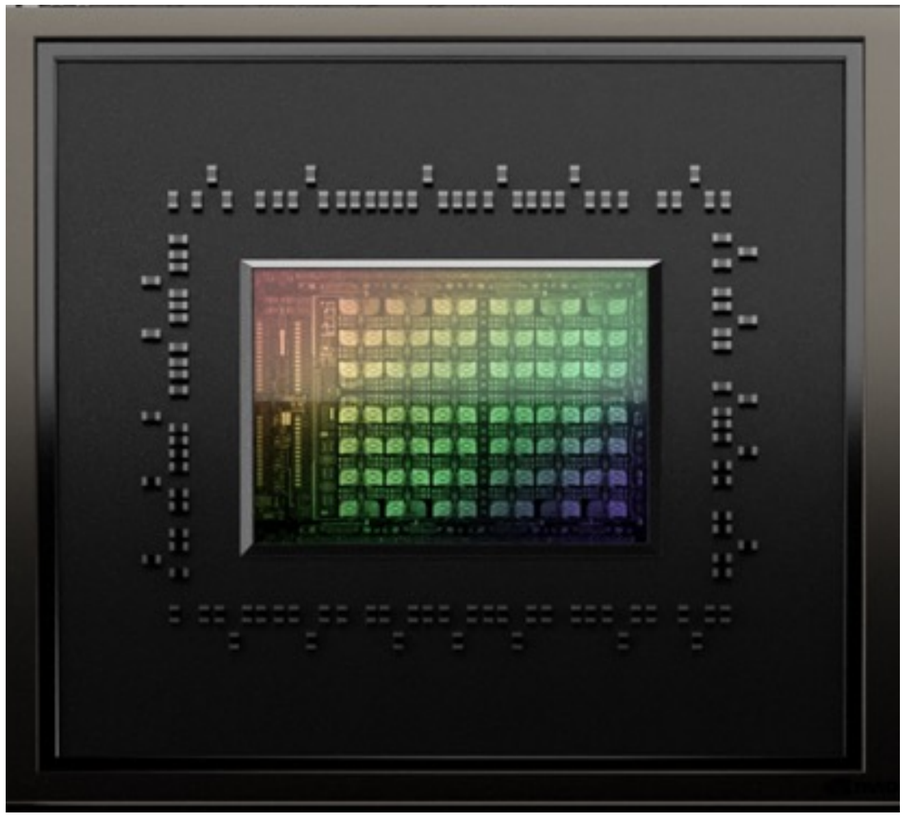
Up to 480 GB of data center enhanced LPDDR5X Memory that
delivers up to 500 GB/s of memory bandwidth

Coherent Chip-to-Chip Connections

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

Industry Leading Performance Per Watt

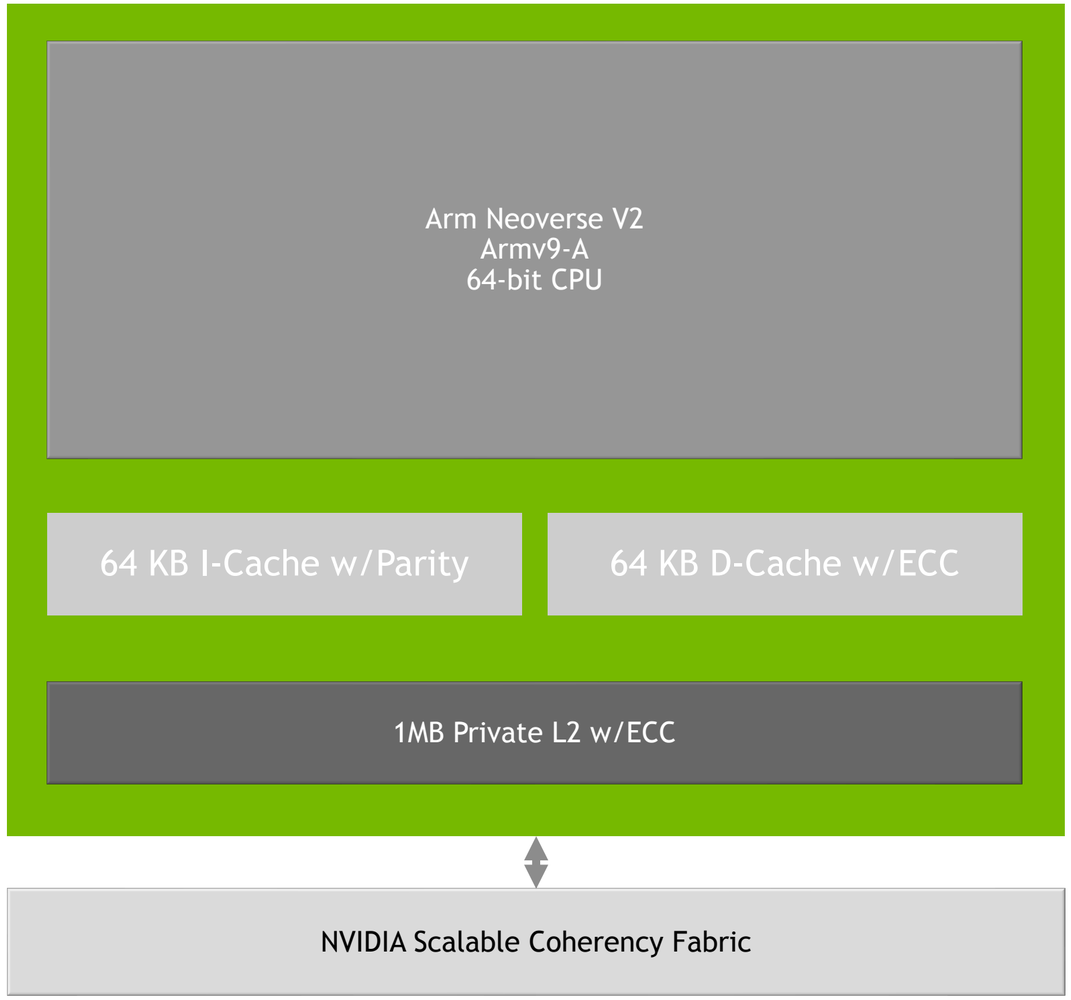
Up to 2X perf / W over today's leading servers



NVIDIA Grace core

Neoverse V2

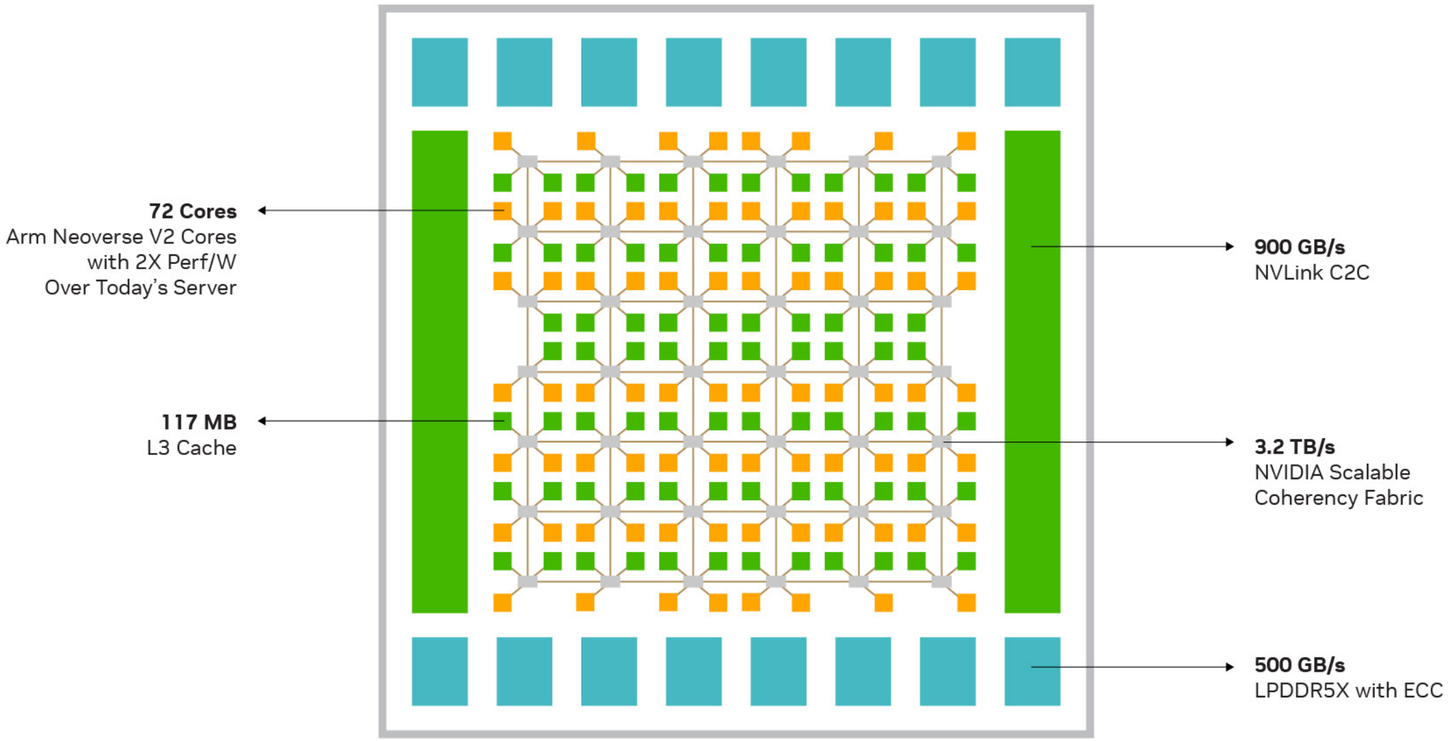
- **Arm Neoverse V2 core - Arm v9.0**
- AARCH64 at all ELs
- v9.0 scalable vector extensions
  - **Scalable Vector Extension 2 (SVE2) - 4 x 128b**
  - Scalable Vector AES (SVE_AES)
  - Scalable Vector PMULL (SVE_PMULL)
  - Scalable Vector SHA3 (SVE_SHA3)
  - Scalable Vector Pit Permutes (SVE_BitPerm)
- V9.0 debug
  - Embedded Trace Extension (ETE)
  - Trace Buffer Extension (TBE)



NVIDIA Grace is a Compute and Data Movement Architecture

NVIDIA Scalable Coherency Fabric and Distributed Cache Design

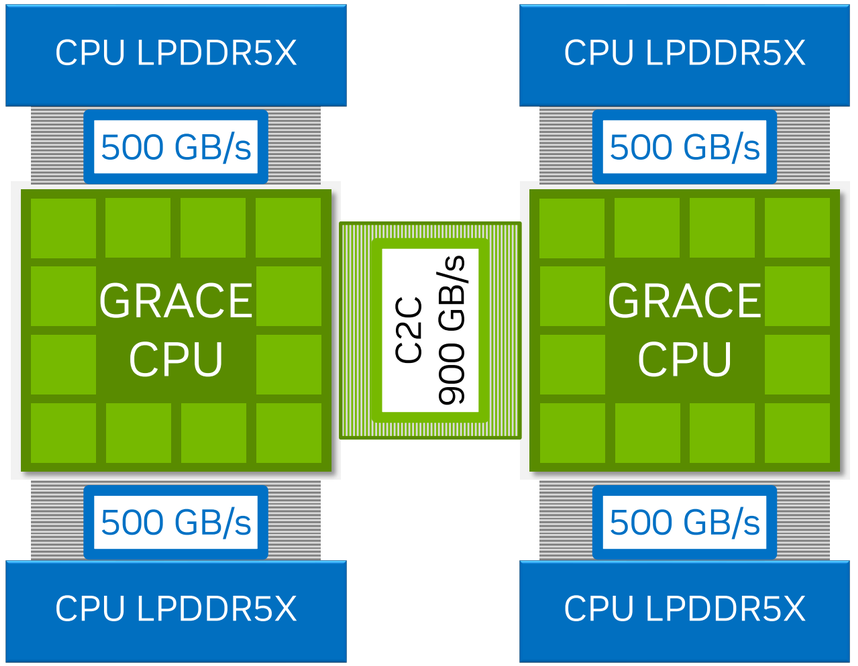
- 72 high performance Arm Neoverse V2 cores with 4x128b SVE2
- 3.2 TB/s bisection bandwidth
- 117MB of L3 cache
- Local caching of remote die memory
- Background data movement via Cache Switch Network



Low-Power High-Bandwidth Memory Subsystem

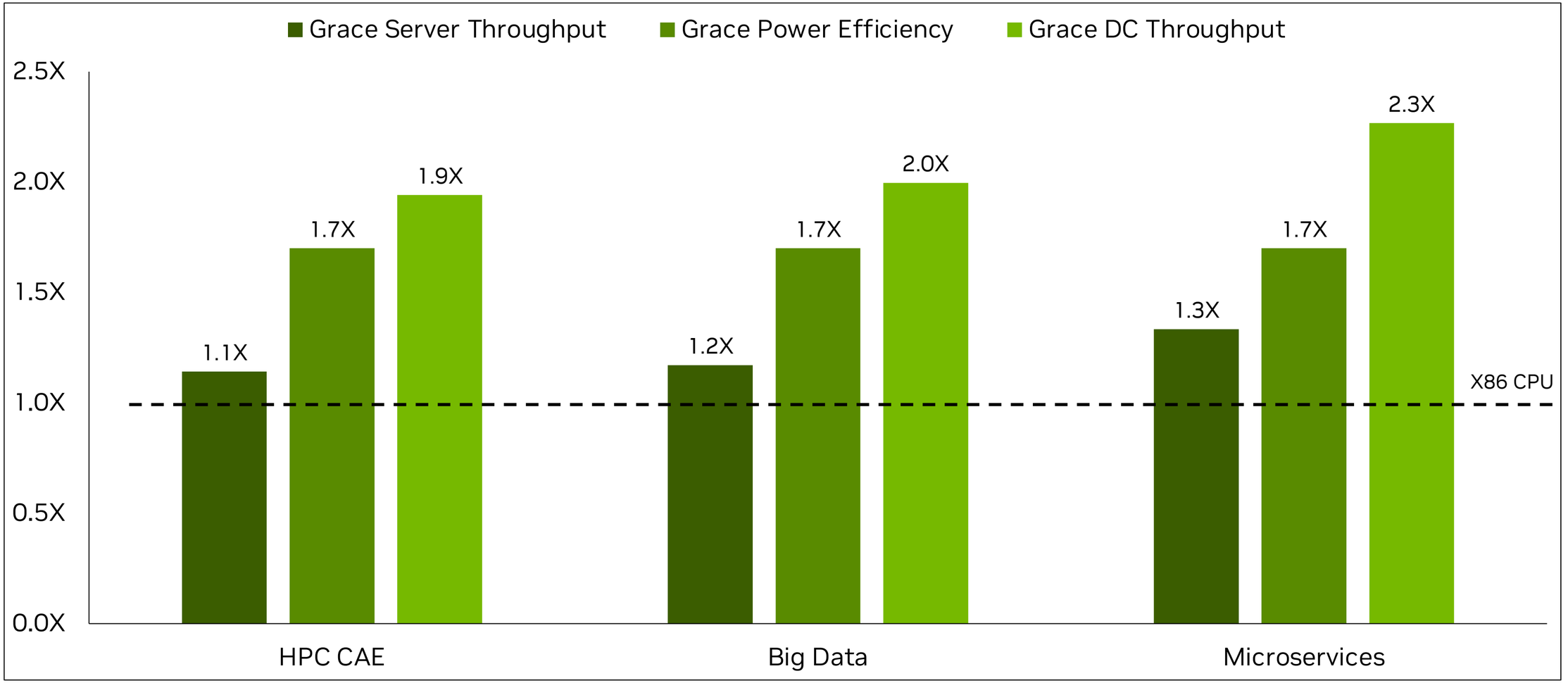
LPDDR5X Data Center Enhanced Memory

- Optimal balance between bandwidth, energy efficiency and capacity
- Up to 1TB/s of raw bidirectional BW
- 1/8th power per GB/s vs conventional DDR memory
- Similar cost / bit to conventional DDR memory
- Data Center class memory with error code correction (ECC)



Nvidia Grace CPU Delivers 2X Data Center Throughput at the Same Power

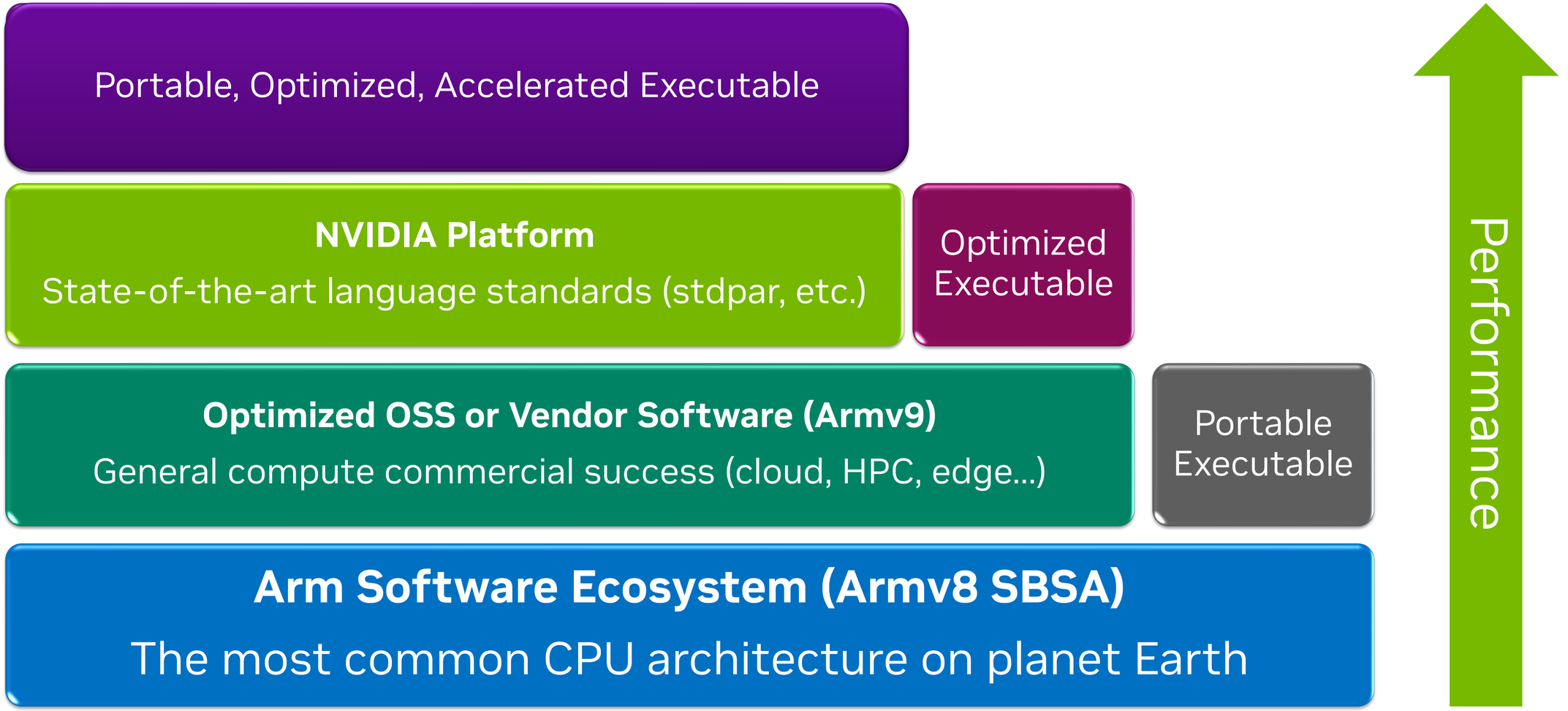
Breakthrough Performance and Efficiency



Data Center level projection of NVIDIA Grace Superchip vs x86 flagship 2-socket data center systems (112 and 192 core systems). HPC CAE: OpenFOAM (Motorbike | Small) Big Data: HiBench+K-means Spark (HiBench 7.1.1, Hadoop 3.3.3, Spark 3.3.0) and Microservices: Google Protobufs (Commit 7cd0b6fbf1643943560d8a9fe553fd206190b27f | N instances in parallel).
NVIDIA Grace Superchip performance based on engineering measurements. Results subject to change.

Grace Software Ecosystem is Built on Standards

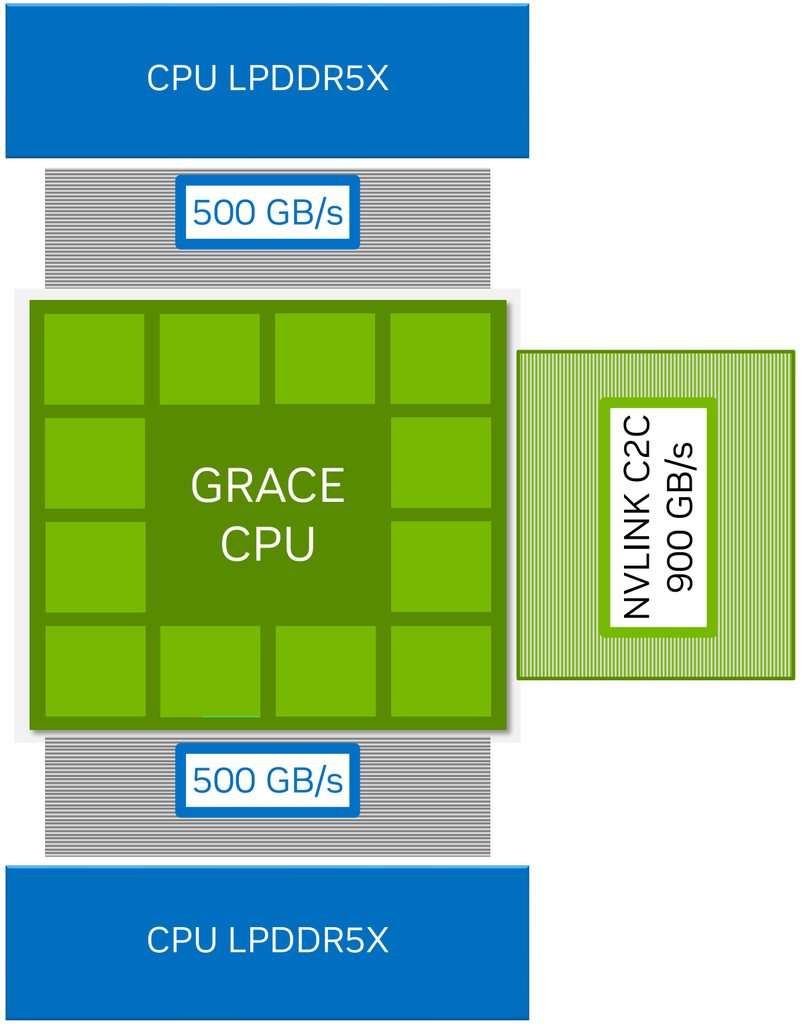
Grace brings the full NVIDIA software stack to Arm.



NVLINK-C2C

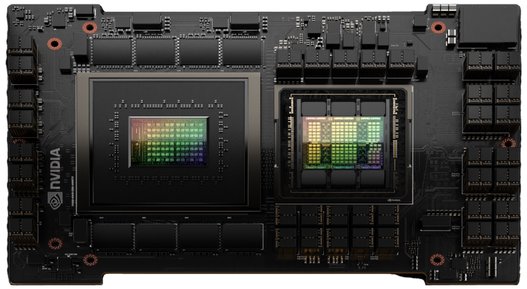
High Speed Chip to Chip Interconnect

- Creates Grace Hopper and Grace Superchips
- Removes the typical cross-socket bottlenecks
- **Up to 900GB/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
- **Enables coherency** for both Grace and Grace Hopper superchips



GRACE HOPPER SUPERCHIP

The breakthrough accelerated CPU for Large-Scale AI and HPC applications



Grace CPU + H100 GPU

72 Arm Neoverse V2 Cores with SVE2 4x128b
Transformer Engine and ~4PFLOPS of FP8

Fast NVLink-C2C Connection

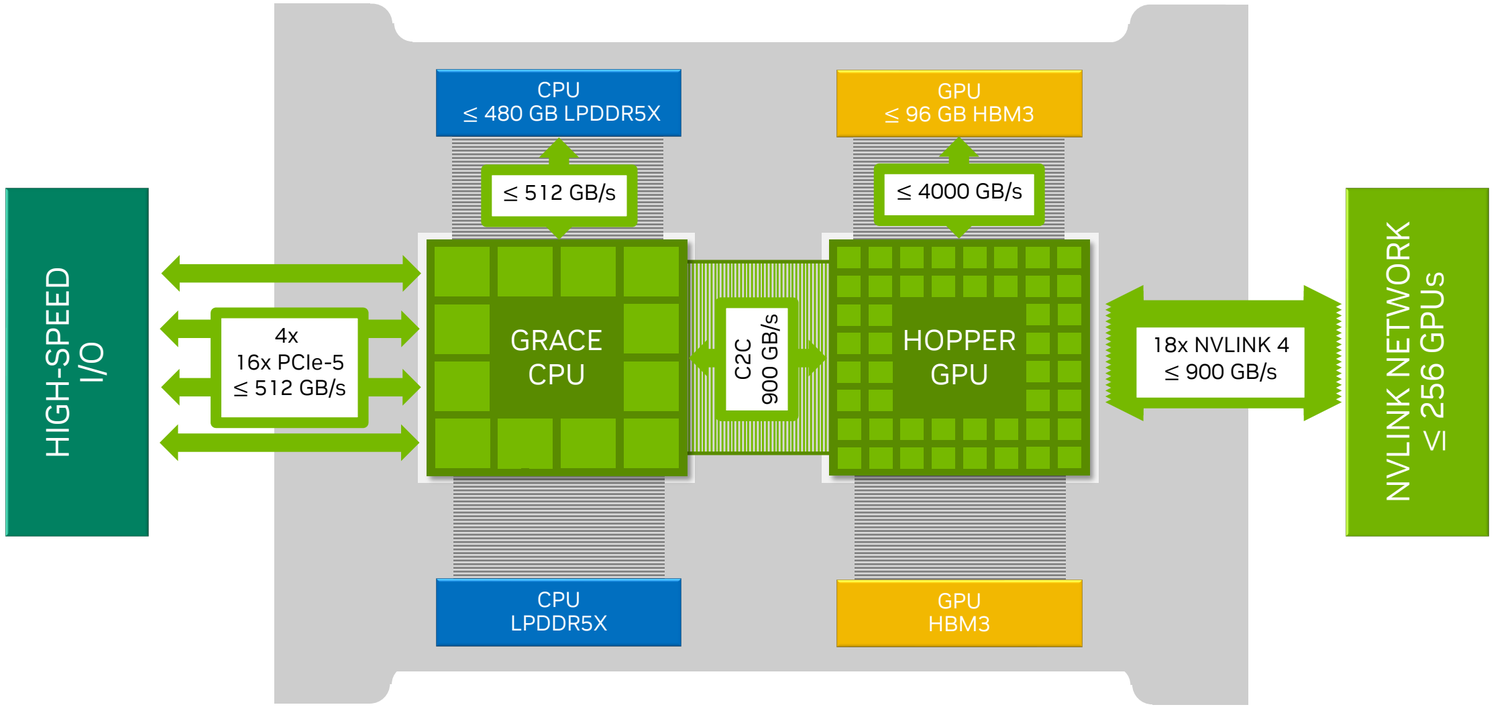
900GB/s bi-directional bandwidth CPU to GPU
7X faster than PCIe Gen 5

~600GB of Fast Access Memory

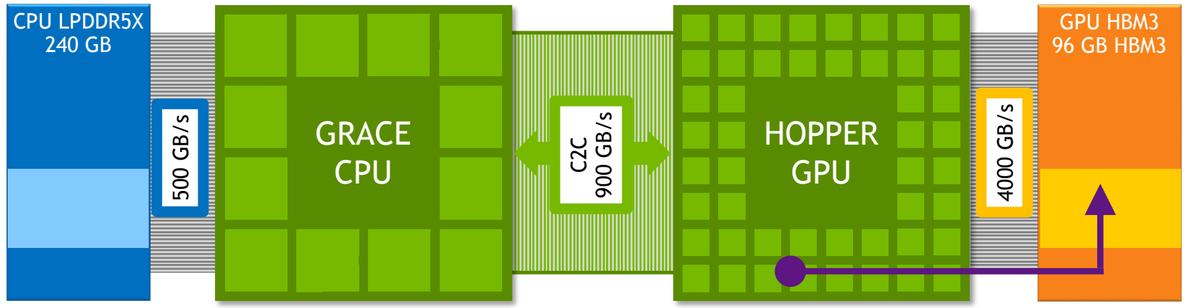
Up to 96GB HBM3, 4TB/s bandwidth
Up to 480GB LPDDR5X, 512GB/s bandwidth

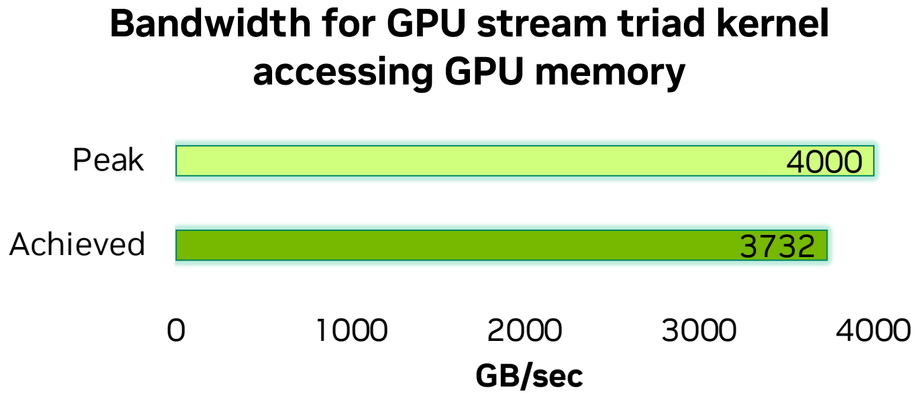
Full NVIDIA Compute Stack

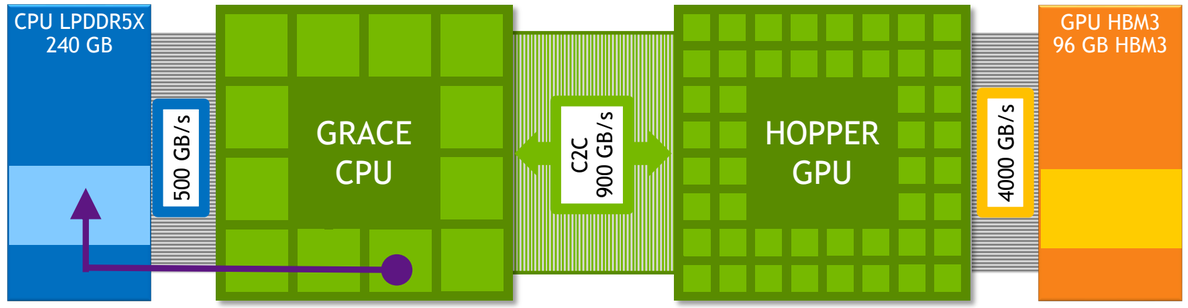
HPC, AI, Omniverse

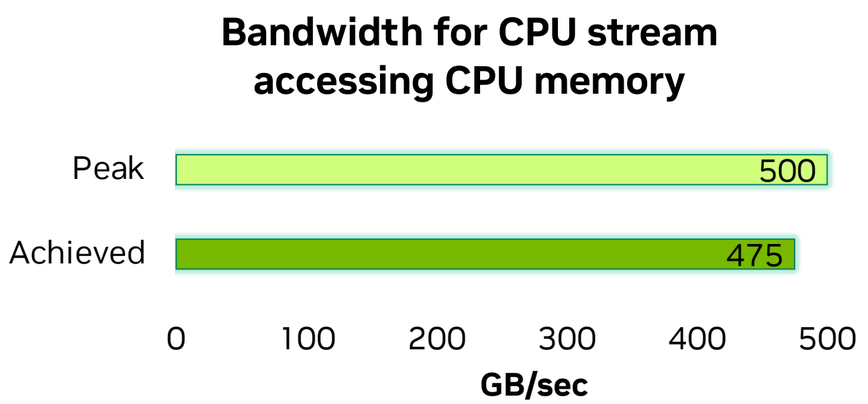


High Bandwidth Memory Access & Automatic Data Migration

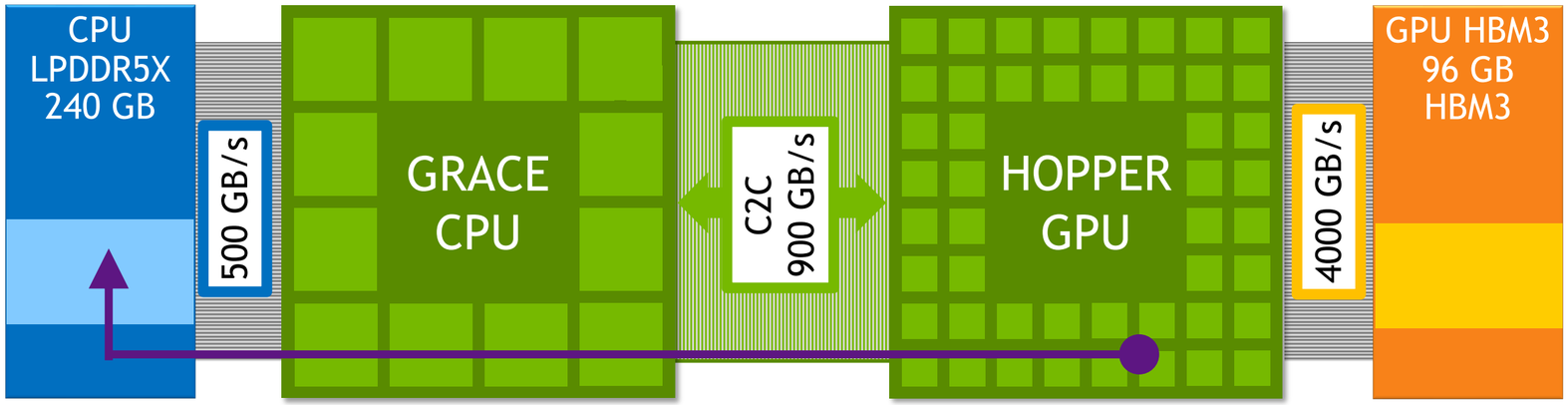


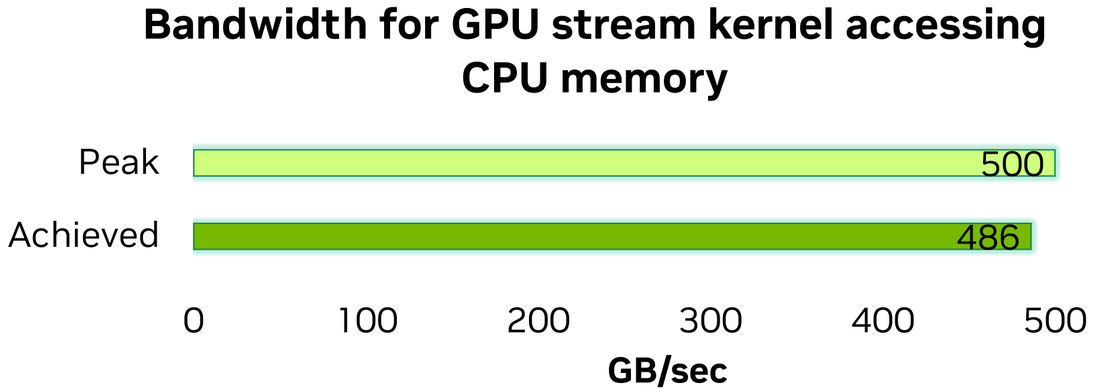






High Bandwidth Memory Access & Automatic Data Migration





These models work best with a hardware supported shared address space

GRAMMING THE NVIDIA PLATFORM

Unmatched Developer Flexibility

On PCs

On Prem

At the Edge

In the Public Cloud

Parallelism in Standard Languages

python

Directives For Existing Apps

OpenACC

OpenMP

Peak Performance

NVIDIA CUDA

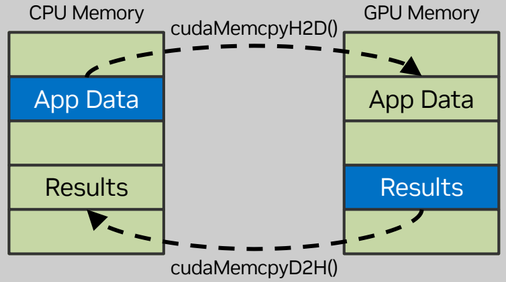
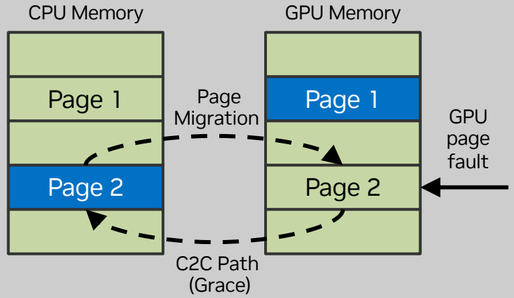
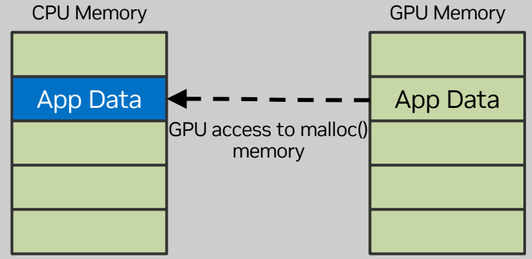
C++ | Fortran | Python

Acceleration Libraries

(AI, Data Analytics, Algebra, Quantum, Communication)

ADVANTAGES OF THE GRACE HOPPER MEMORY MODEL

Full CUDA support with additional Grace memory extensions

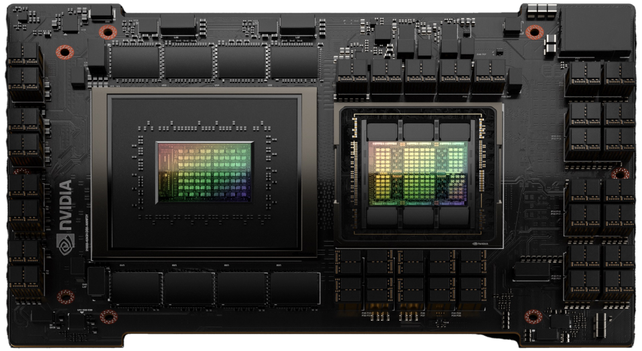
	Explicit Copy	Managed Memory	System Allocated
	Application explicitly moves data between CPU & GPU as needed 	CPU and GPU can access memory on-demand and data migrated locally for higher BW access 	GPU can access memory allocated from `malloc()`, `mmap()`, etc. 
HGX	~60 GB/s PCIe Gen5 transfers (H2D/D2H)	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

Grace Hopper HPC Platform

Unified Memory and Cache Coherence for next gen HPC performance

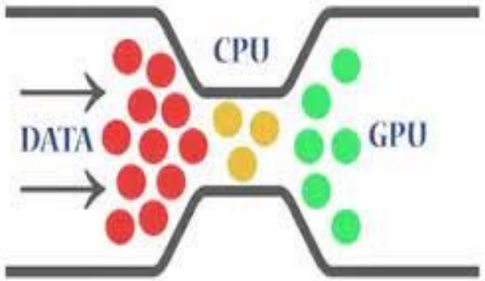
Partially GPU Accelerated Apps

Big performance gains with no code changes



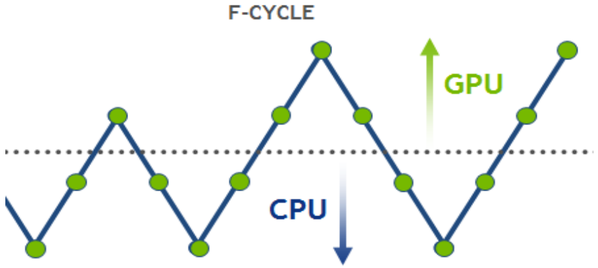
No More PCIe Bottleneck

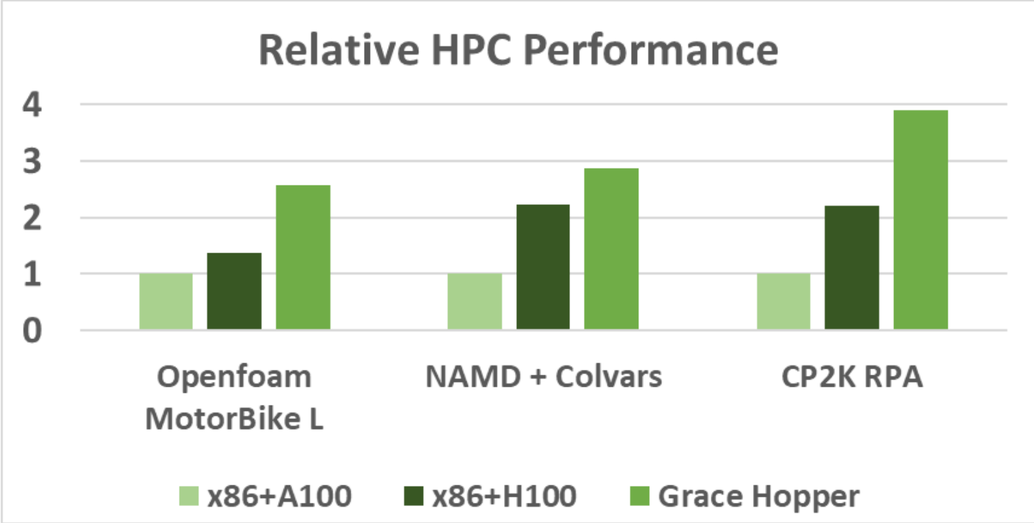
NVLink-C2C is 7X PCIe BW



CPU & GPU Cache Coherence

Incremental code changes yield big gains





Fast Access Memory

600GB

Memory Bandwidth

4TB/s

HPC: Preliminary results comparing DGX A100, DGX H100 and Grace Hopper systems.
AI: Simulation results comparing DGX H100 and Grace Hopper systems. GNN: GraphSAGE model for an augmented ogbn- products dataset of 626M nodes and 31B edges, DRLM: Deep Learning Recommendation Model (DLRM) 27TB model, NLP: fine-tuning a GPT-175B model

Application on Accelerated Systems

Partially GPU Accelerated

As GPUs become faster applications become increasingly limited by non-GPU factors

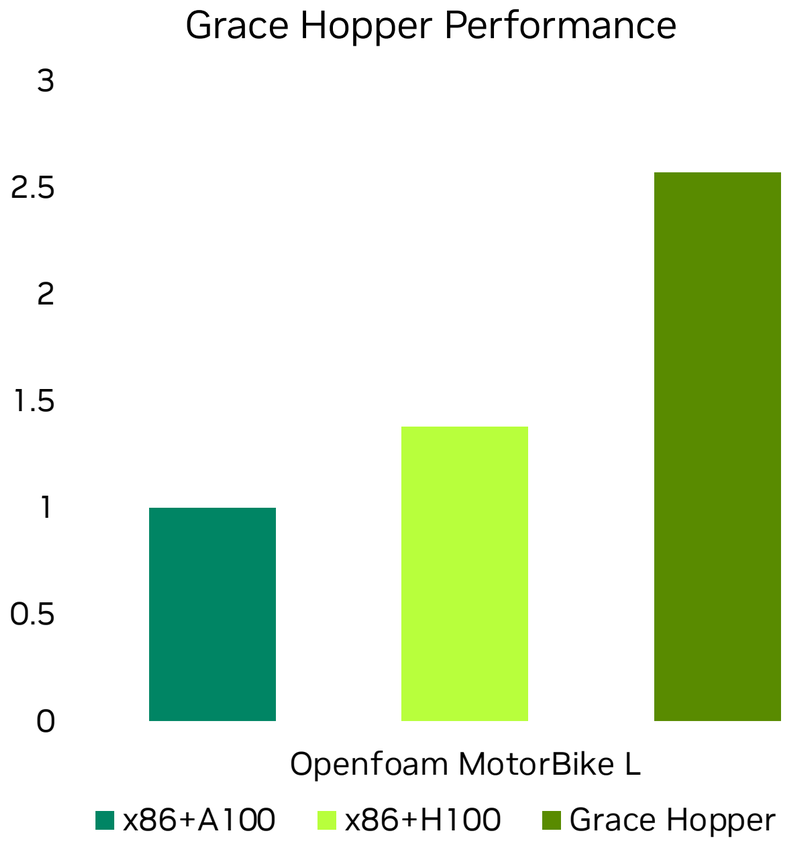
e.g. mostly data transfer (PCIe) limited

- mostly CPU limited

OpenFoam

Partially GPU Accelerated – mostly CPU limited

- Computational fluid dynamics (CFD) toolbox developed by OpenCFD
  - Popular in automotive and other engineering sectors
  - Highly configurable fluid flow solvers with turbulence / heat transfer / etc.
  - Leverage GPU accelerated AMGX linear solvers
- HPC motorbike problem (Large)
  - Around 30% of CPU-only execution is spent in linear solves
- Performance on Grace Hopper
  - High CPU and GPU memory bandwidth improve compute performance
  - C2C bandwidth minimises the cost of migrating CPU matrix data



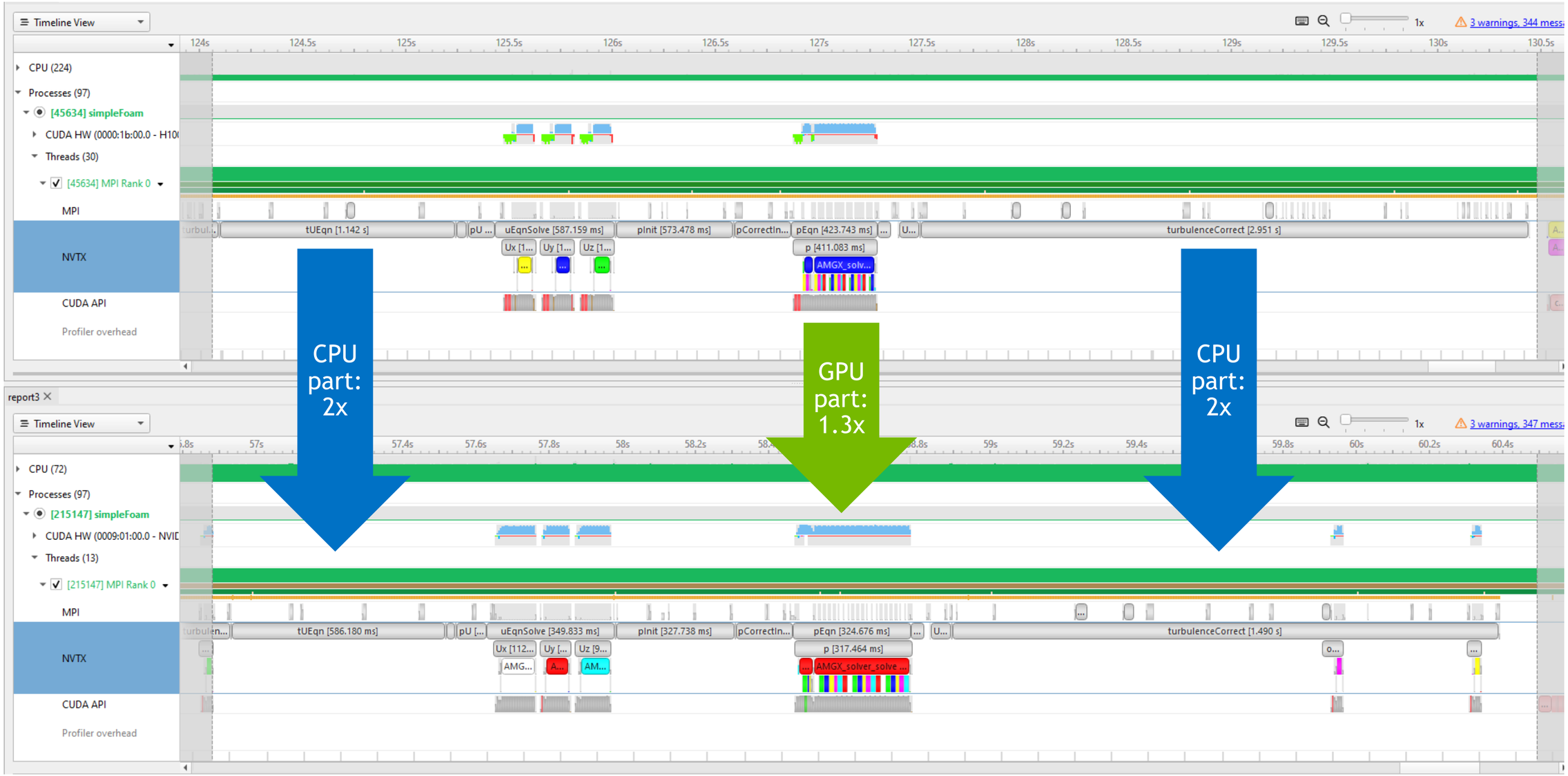
~35M cells benchmark designed by OpenFOAM HPC technical committee

OpenFoam

Nsight Systems Profile

x86 + H100 (H100 80GB HBM3)

Grace Hopper (H100 96GB HBM3)



NVIDIA

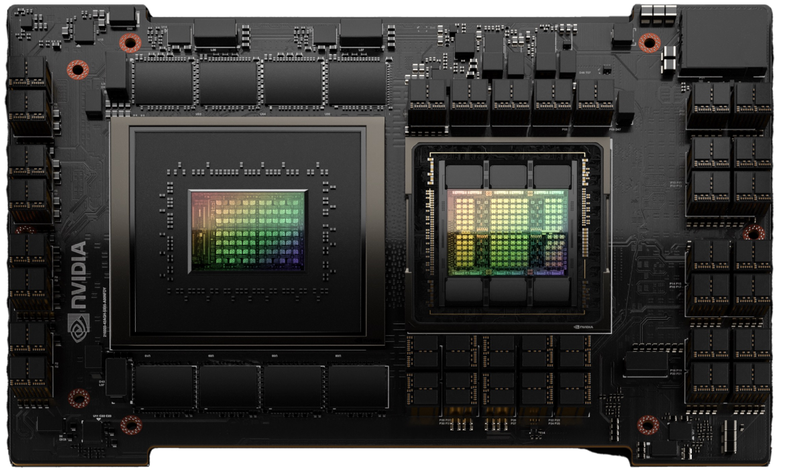
Further Resources for Grace CPU and Grace Hopper

Grace CPU Superchip

- Grace CPU Superchip Architecture Whitepaper
- Grace CPU Architecture In-Depth Blog
- Grace CPU Superchip Data Sheet
- Grace CPU Energy Efficiency Blog
- A Demonstration of AI and HPC Applications for NVIDIA Grace CPU [S51880]

Grace Hopper Superchip

- Grace Hopper Superchip Architecture Whitepaper
- Grace Hopper Architecture In-Depth Blog
- Grace Hopper Superchip Architecture Data Sheet
- Grace Hopper Recommender System Blog
- Programming Model and Applications for the Grace Hopper Superchip [S51120]




NVIDIA