Accelerating Time-To-Science using the NVIDIA platform

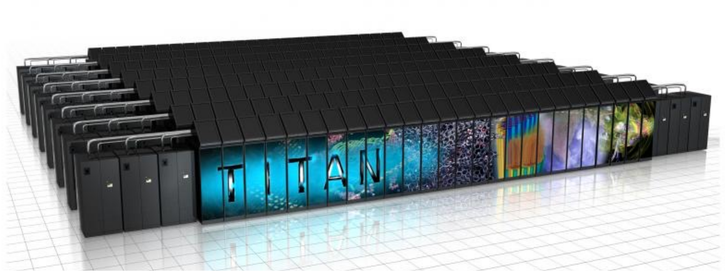
Filippo Spiga | fspiga@nvidia.com

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.

Evolution of GPU accelerated system design

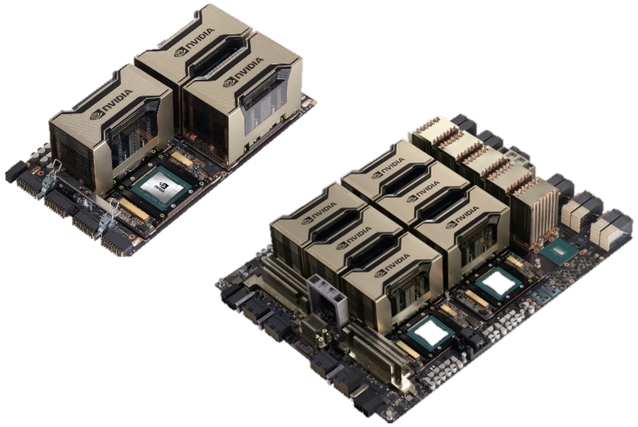
Software and Hardware innovations hand-by-hand

Single CPU - Single GPU



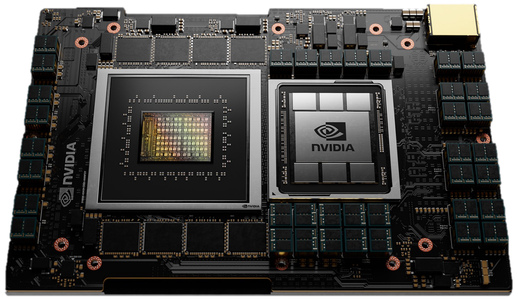
Single CPU - Multi GPU

(“fat” nodes, 2/4/8 ways)



Efficient intra-node GPU-to-GPU, PCIe bottleneck

Single CPU – Single GPU

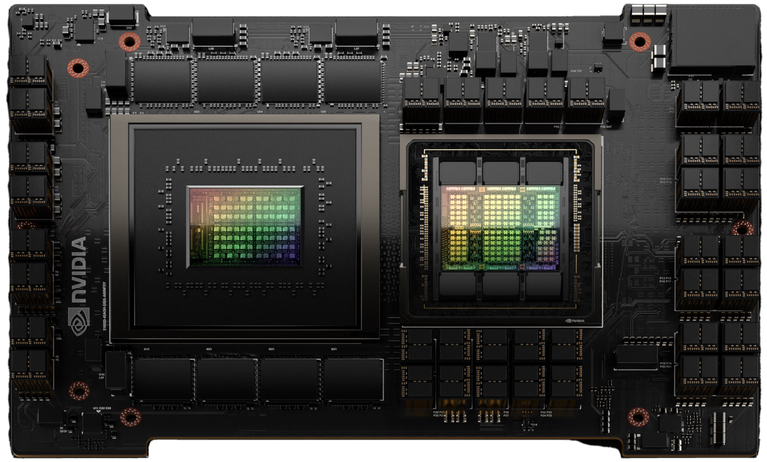


Removing the D2H/H2D bottleneck, adding HW coherency

NVIDIA Grace for HPC & AI Infrastructure

Grace Hopper Superchip

Giant Scale AI & HPC

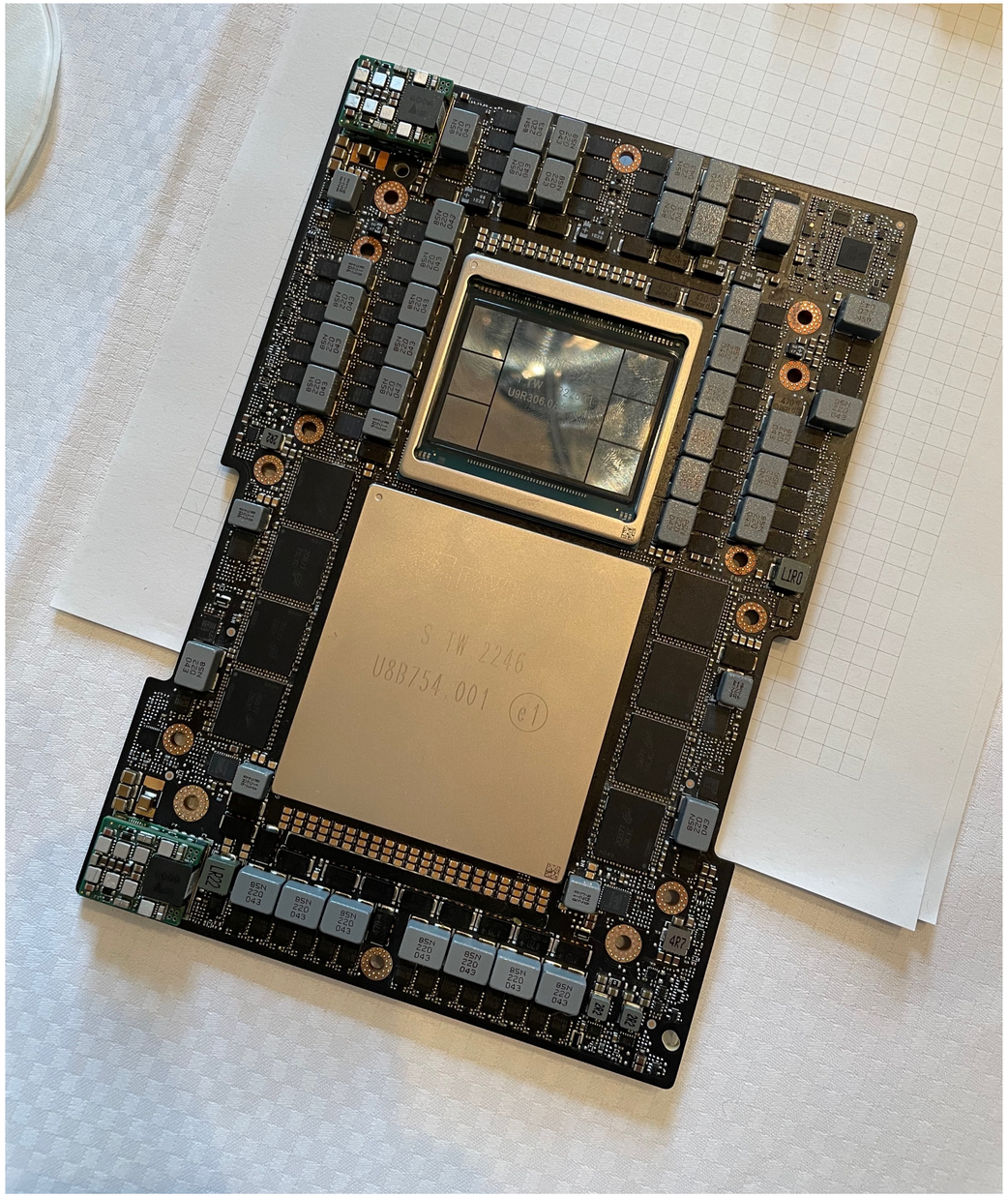


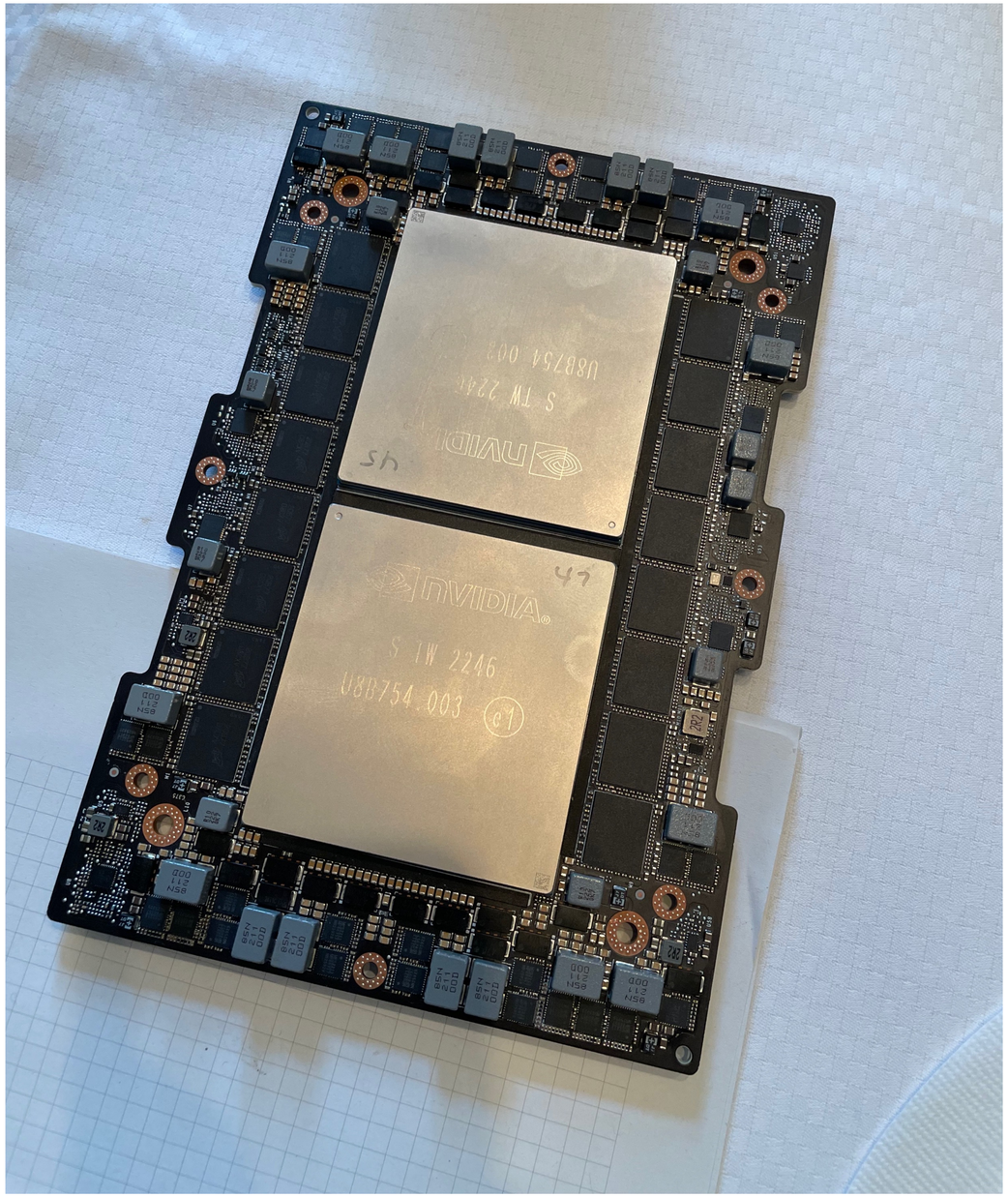
Accelerated applications where CPU performance and system memory BW are critical; extreme and highly atomic collaboration between CPU & GPU contexts for flagship AI & HPC

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


S IW 2246
U8B754.001
e1


NVIDIA
S IW 2246
U8B754.003
e1

Announcing Grace Hopper Now in Full Production

Grace and Grace Hopper transforming HPC and AI

LANL (Venado)
Grace Hopper
10 EFLOPS AI Perf

Univ. of Bristol (Isambard 3)
Grace CPU
2 PFLOPS HPC Perf

BSC (MareNostrum 5)
Grace CPU
2 PFLOPS HPC Perf

CSCS (ALPS)
Grace Hopper
20 EFLOPS AI Perf

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

NCHC (Taiwania 4)
Grace CPU
300 TFLOPS HPC Perf

Contents

Inside the NVIDIA Superchip Platform

NVIDIA Grace CPU Superchip

2X Performance at the Same Power 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 bi-section bandwidth connects
CPU cores, NVLink-C2C, memory, and system IO

High-Bandwidth Low-Power Memory

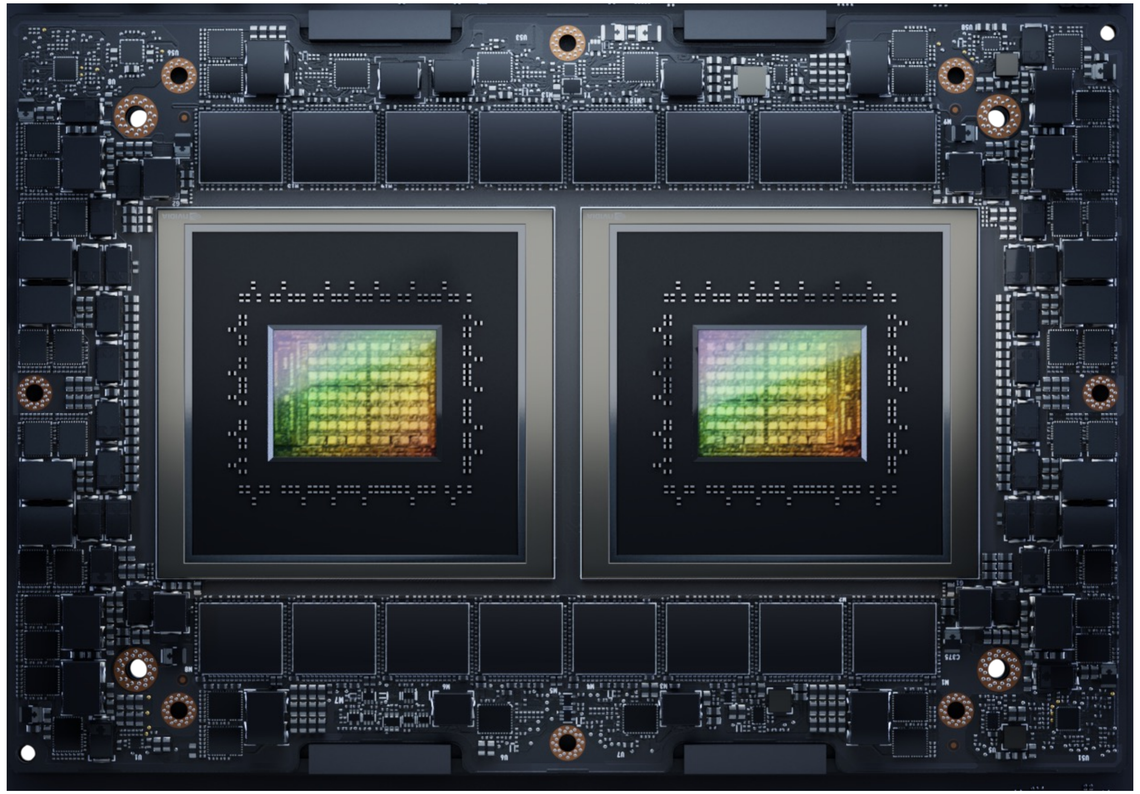
Up to 960GB of data center enhanced LPDDR5X Memory that
delivers up to1TB/s of memory bandwidth

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

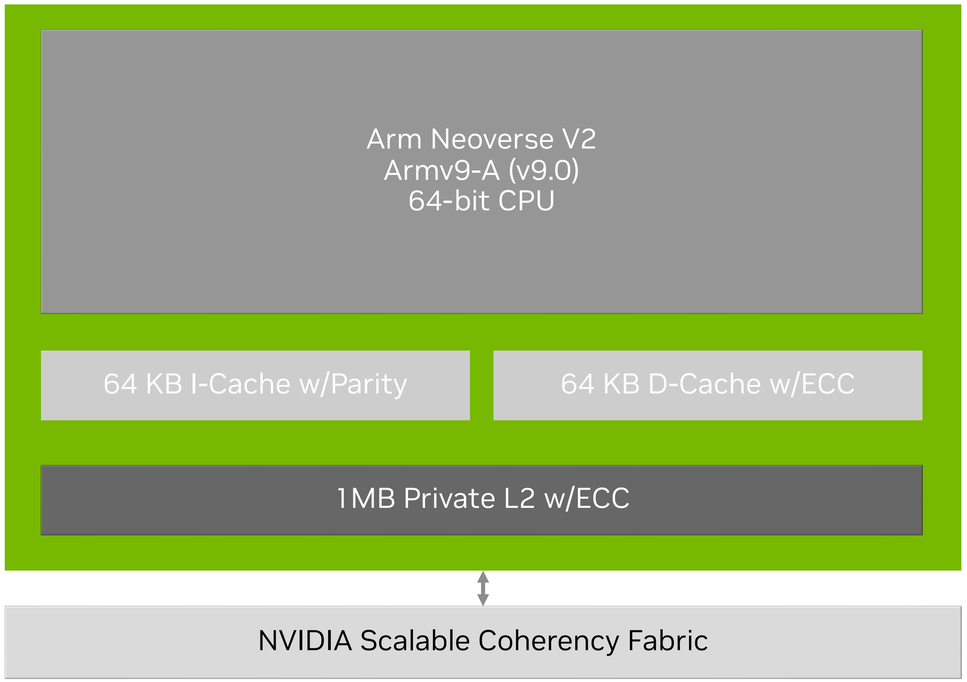
AI, Omniverse



NVIDIA Grace

Introducing Neoverse V2

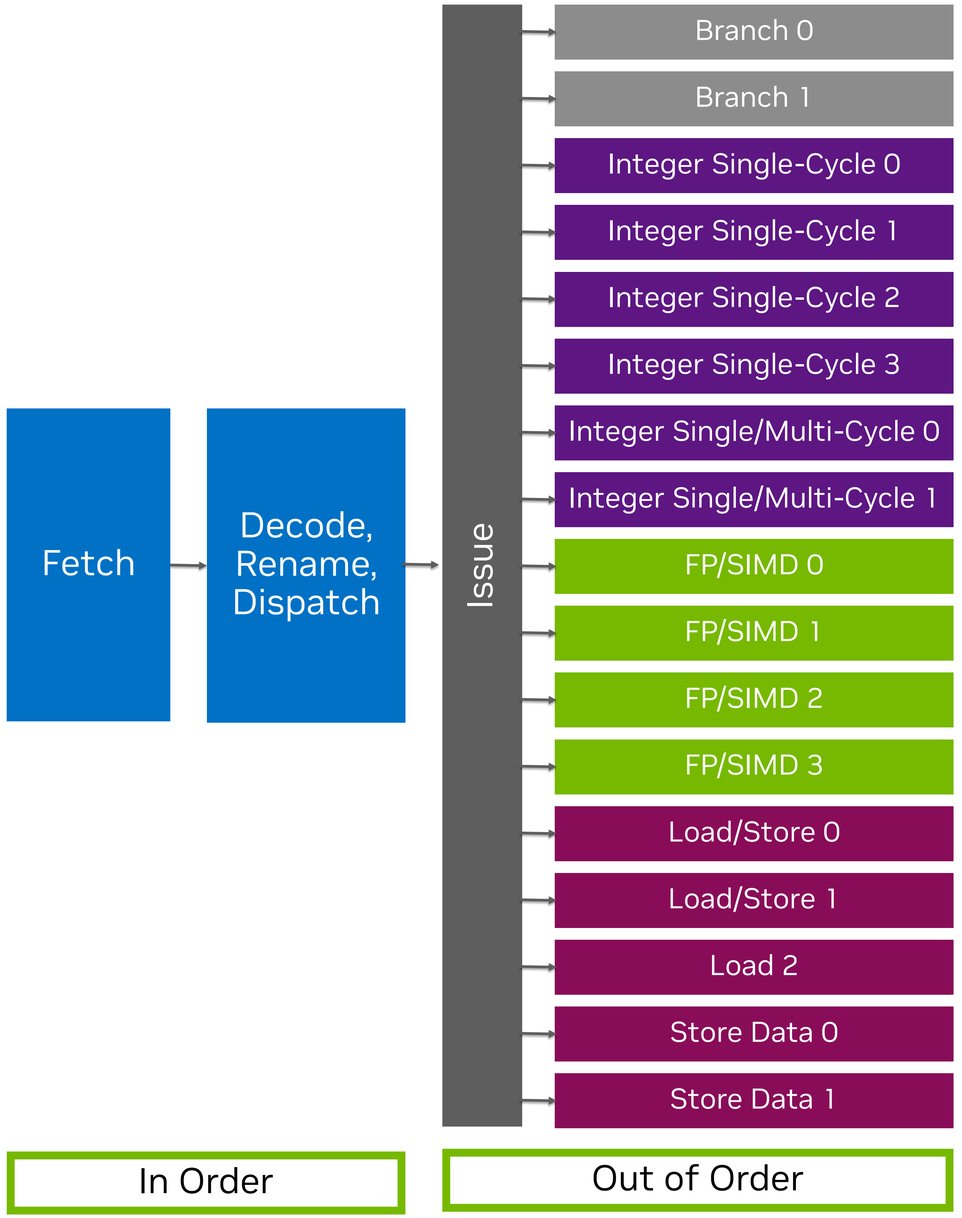
- **Arm Neoverse V2 core – Arm v9.0**
- AARCH64 at all ELs
- v9.0 scalable vector extensions
  - **Scalable Vector Extension 2 (SVE2) - 4 x 128b**
  - Scalable Vector AES (SVE_AES)
  - Scalable Vector PMULL (SVE_PMULL)
  - Scalable Vector SHA3 (SVE_SHA3)
  - Scalable Vector Pit Permutes (SVE_BitPerm)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



NVIDIA Grace SIMD Highlights

Neoverse V2 Arm IP core

- 4x128b SIMD units = 512b SIMD vector bandwidth total
- Each SIMD unit can retire NEON or SVE2 instructions
- On this architecture, SVE2 and NEON have the same peak performance
  - *Different from Fujitsu A64FX where SVE 4x faster than NEON*
- SVE2 can vectorize more complex codes and supports more data types than NEON.
- NEON doesn't require predicate calculation
  - *Neither does VLS SVE, but that's an advanced topic*



NVIDIA Grace vs. Fujitsu A64FX

A64FX is special in many way – Grace is mainstream

MAINSTREAM LEADERSHIP HPC

Familiar design	• High single-thread performance • Simple memory hierarchy
Large user community	• Runs key HPC applications out-of-the-box • Standard best practices hold true
Significant fraction of peak w/o tuning	• OSS toolchains (i.e. GNU) are tuned for u-arch • Performance curves generally follow expectation

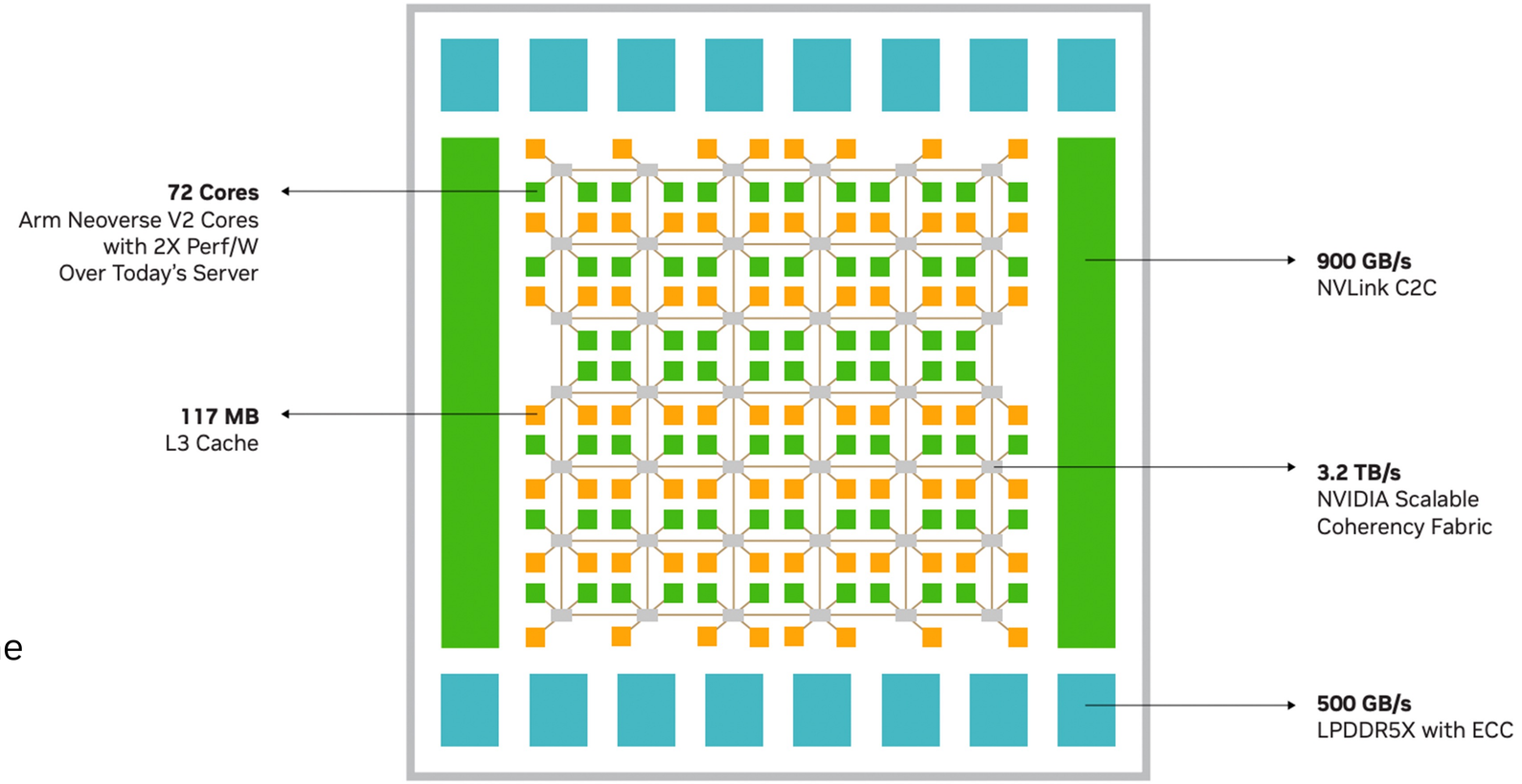
EXTREME HPC CODESIGN

Codesigned for specific application	• Custom hardware or software • Trades generality for performance
Small userbase of extreme experts	• Nonstandard software environments • Common assumptions may hurt performance
Significant tuning effort required	• OSS toolchains unlikely to be performant • Plan for man-months of optimization effort

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

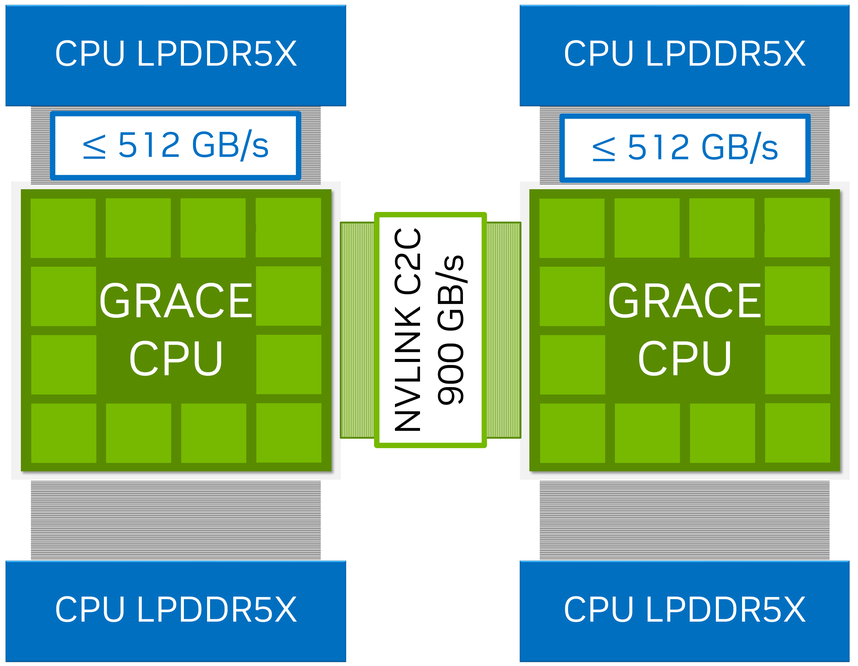


Example possible fabric topology for illustrative purposes

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)



Grace Simplifies System Design and Workload Optimization

Reduces NUMA Bottlenecks

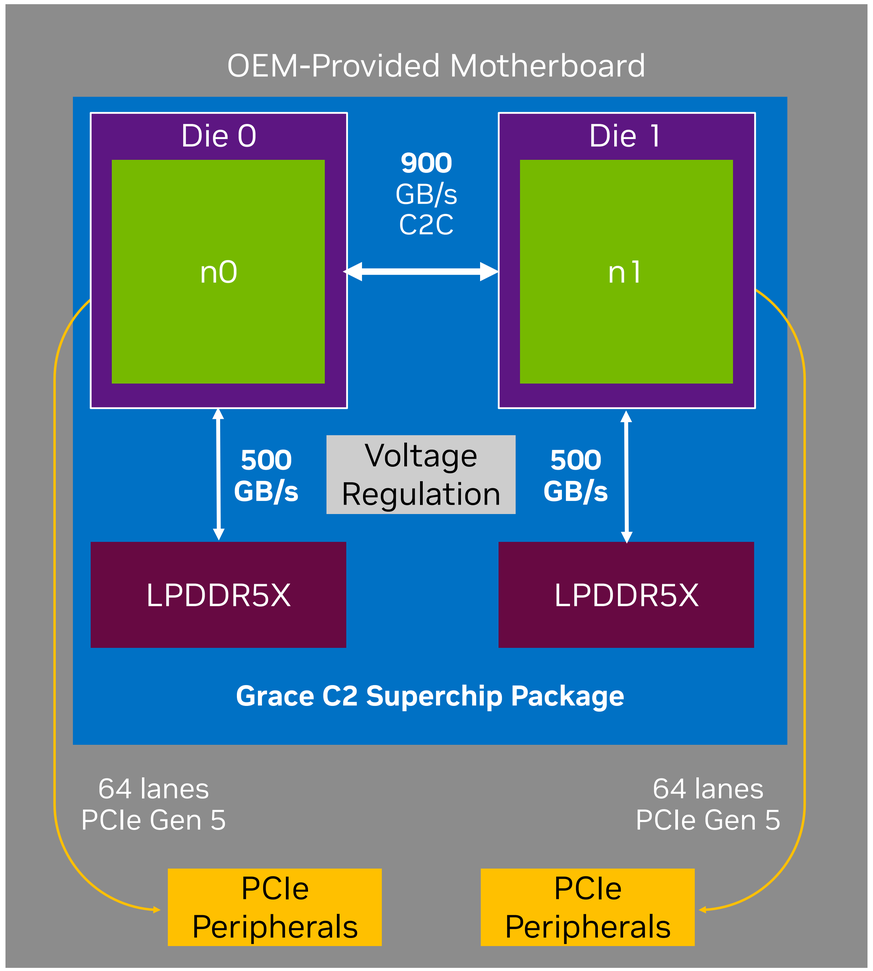
2 NUMA Nodes

2 Compute Dies

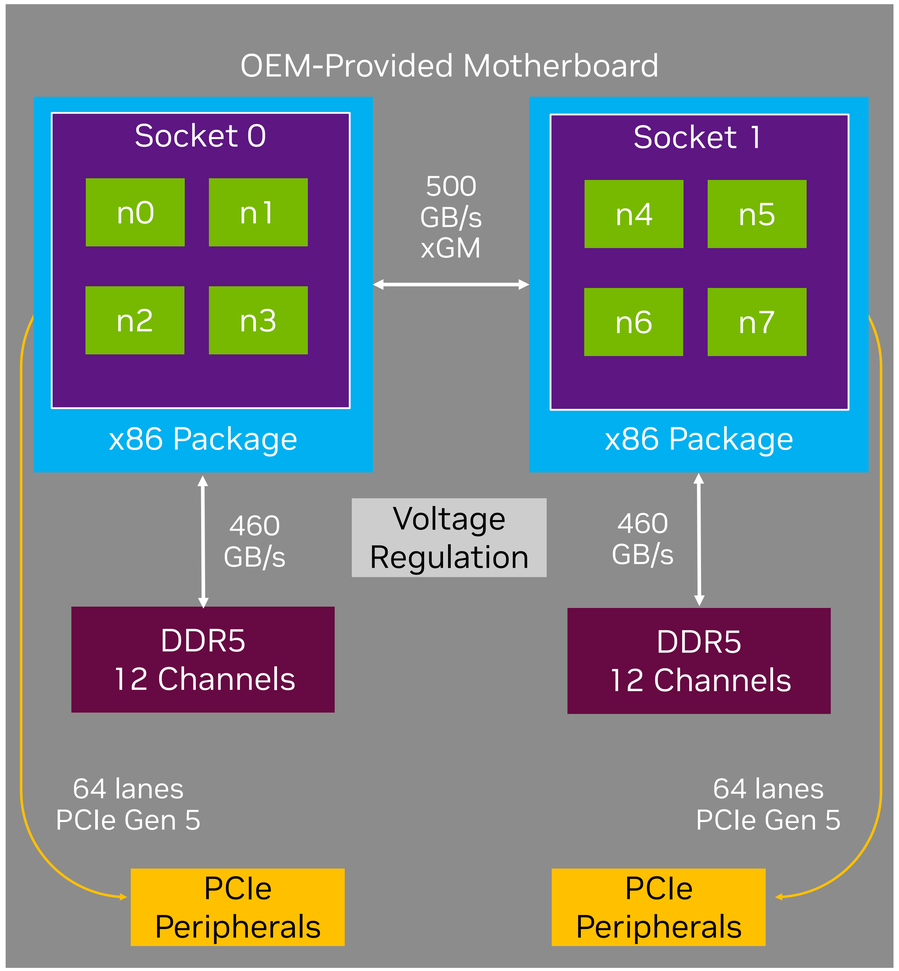
500 Watts (CPU + MEM)

900 GB/s n-to-n

Grace Server
Grace C2 Superchip



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 Serves High-Performance and Power Constrained Markets

Maximize Data Center Throughput

Cloud

Maximize CSP Revenue

Hyperscale Cloud

Consumer Internet Providers

2X Perf Per Rack vs x86

Enterprise and High-Perf / W Edge

Maximize Compute Density

Industrial

Retail

Financial

Telco

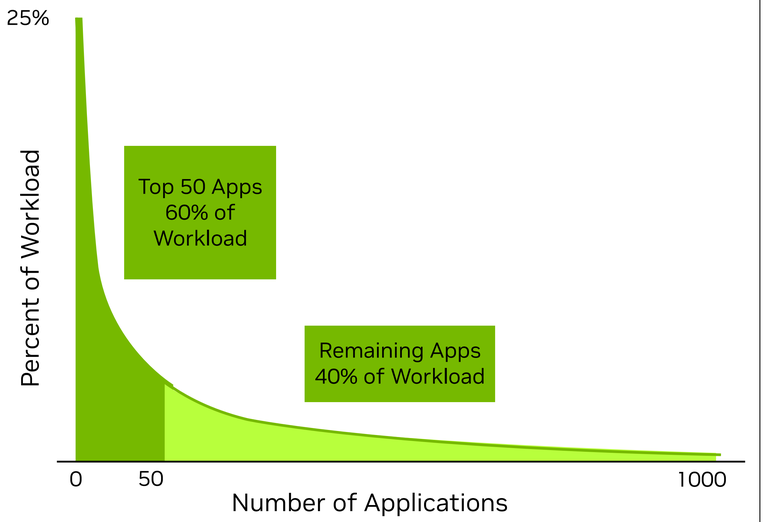
CDN

Storage

60% of x86 Power Per Server

HPC

Maximize Science



1.5X FP64 PFLOPs at Same Power vs x86

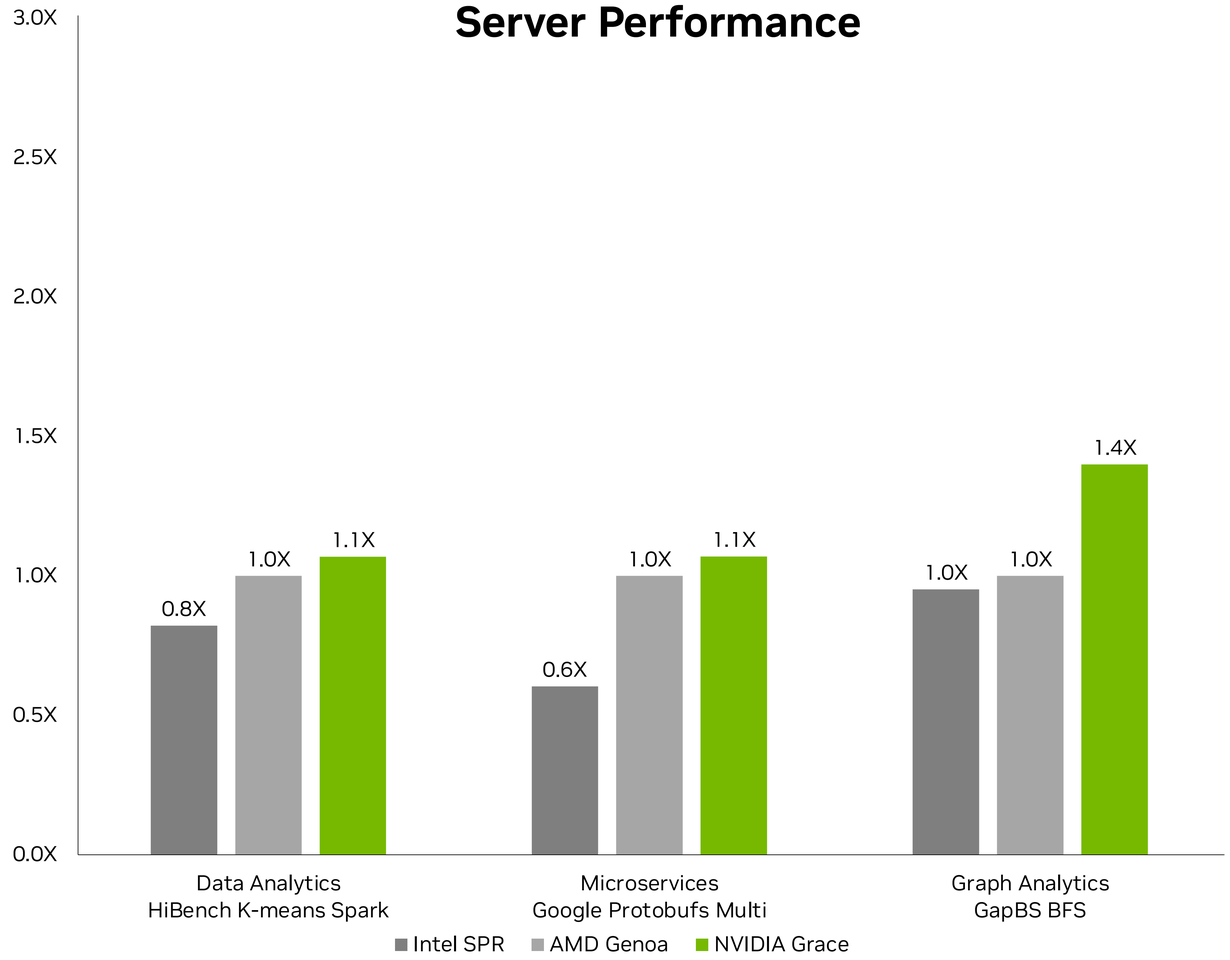
More Users, More Devices, More Intelligence, More Compute

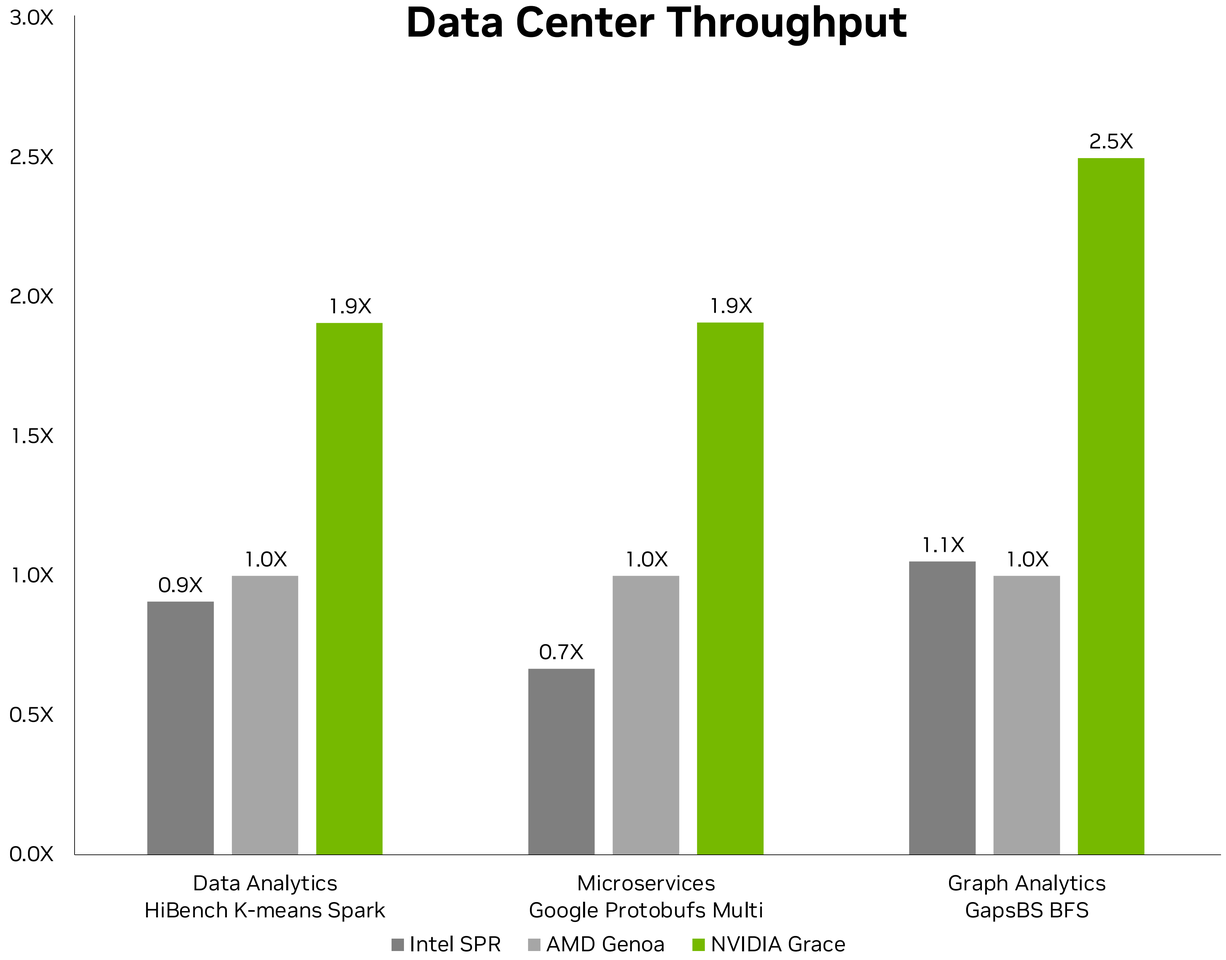
Compute Demands are Increasing Exponentially in Different Dimensions.

	Front End Surface Workloads	**AI** Accelerated Workloads	**Back End** Volume Workloads
Products	**AMD Bergamo, Ampere One**	**NVIDIA GPUs**	**NVIDIA Grace, Intel Sapphire Rapids, AMD Genoa**
Role	Broker transactions from clients and throughout the data center.	Ubiquitous AI & Acceleration	Heavy data and application processing.
CPU Qualities	Cores per socket, loosely coupled.	-	Balanced Performance (high single-thread performance, memory bandwidth, and core-to-core communication)
Applications	IaaS Cloud, Proxy, load balancing, web service endpoints.	-	Application servers, data analytics, graph databases, HPC, simulation, reinforcement learning.
Optimization	Cores & Threads per Watt, per DC	Performance per Watt, per DC	Amdahl's Law: Performance Per Core Performance per Watt, per DC

NVIDIA Grace CPU Delivers 2.5X Cloud Data Center Throughput at the Same Power

Breakthrough Performance and Efficiency



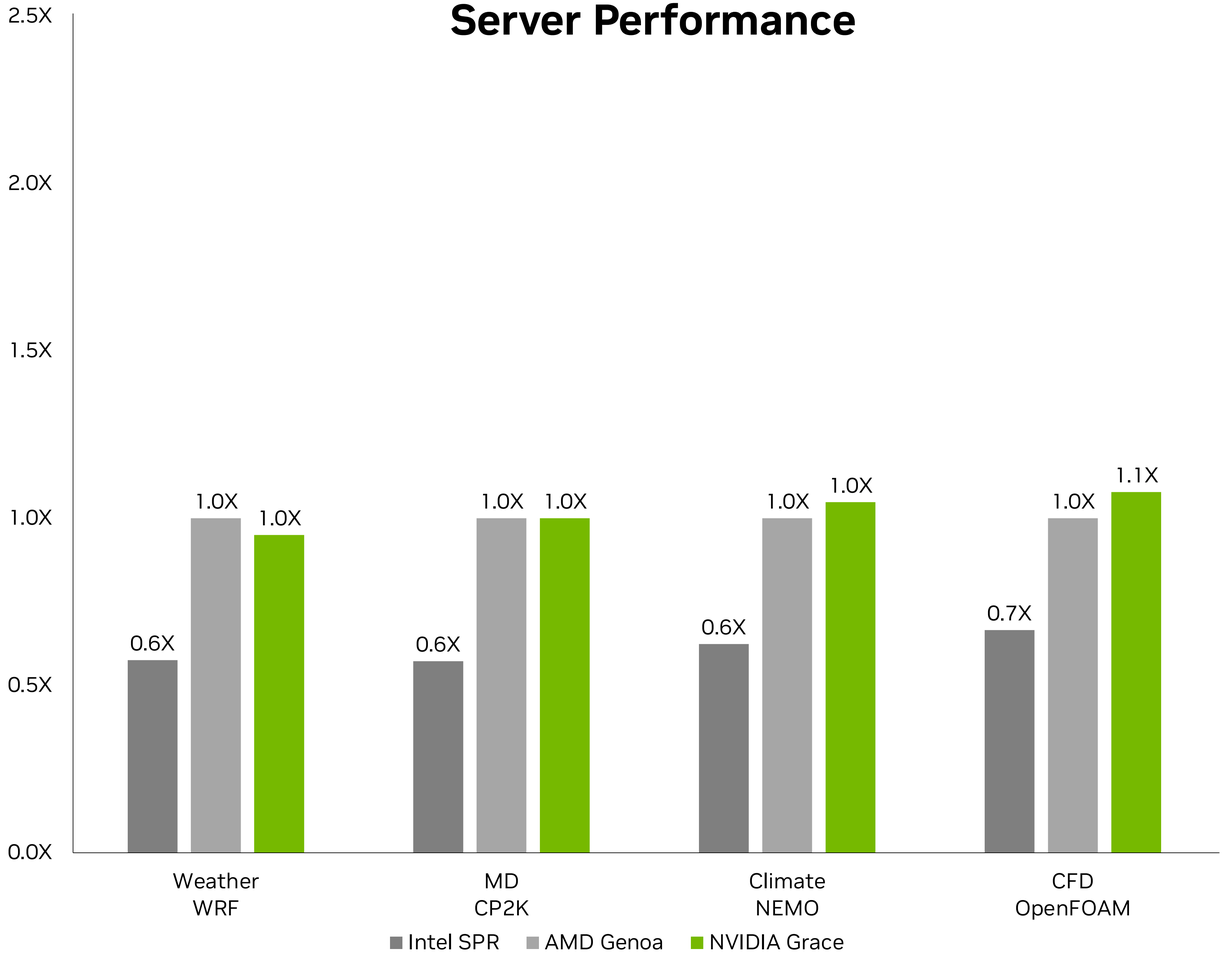


Data Center level projection of NVIDIA Grace Superchip vs x86 flagship 2-socket data center systems (AMD Epyc 9654 and Intel Xeon 8480+). Data Analytics : HiBench+K-means Spark (HiBench 7.1.1, Hadoop 3.3.3, Spark 3.3.0) and Microservices: Google Protobufs (Commit 7cd0b6fbf1643943560d8a9fe553fd206190b27f | N instances in parallel). Graph Analytics: The Gap Benchmarks Suite BFS arXiv:1508.03619 [cs.DC], 2015. NVIDIA Grace Superchip performance based on engineering measurements. Results subject to change.

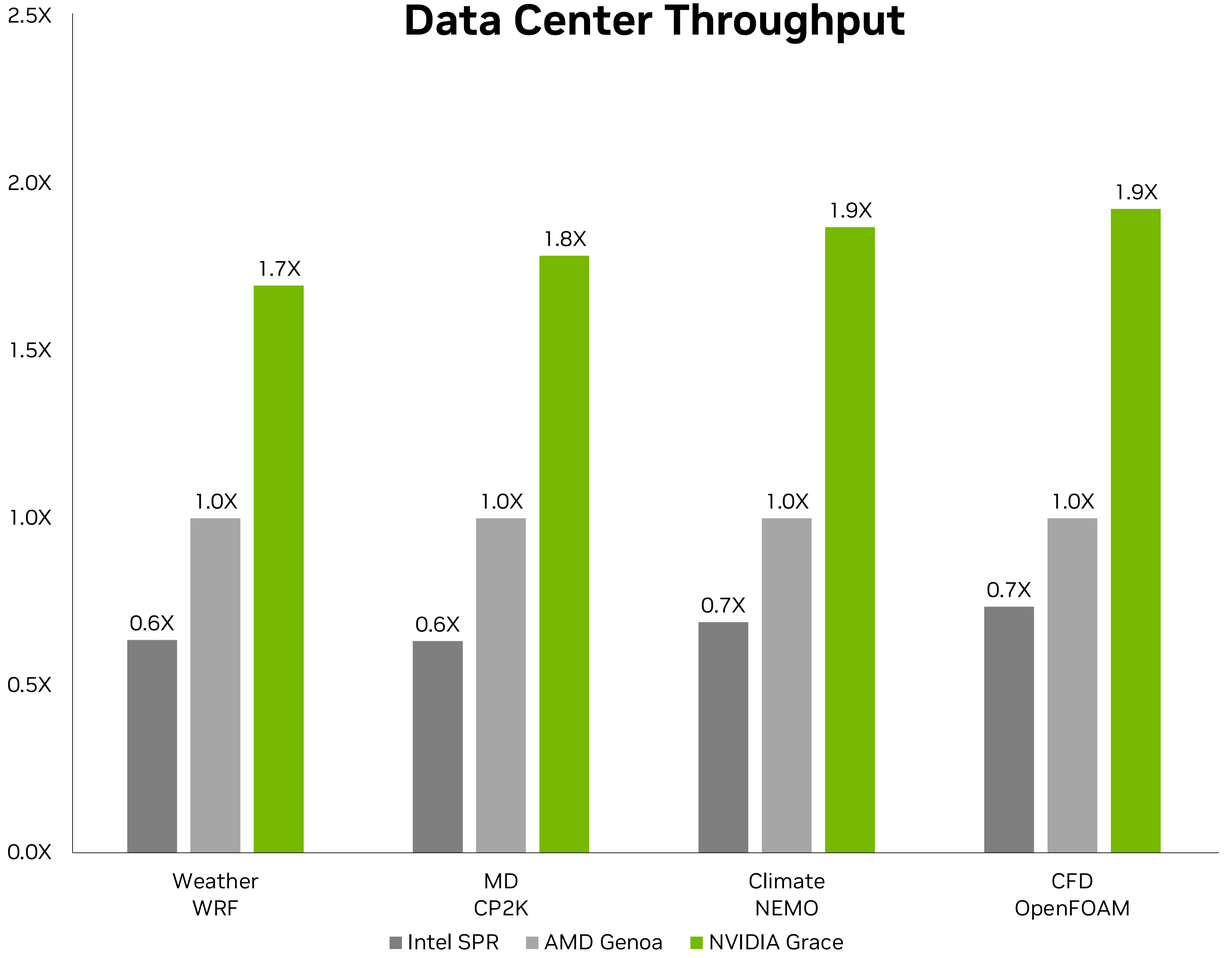
NVIDIA Grace CPU Delivers 1.9X HPC Data Center Throughput at the Same Power

Breakthrough Performance and Efficiency

Server Performance



Data Center Throughput



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

NVIDIA HOPPER H100 GPU

Breakthrough Performance and Efficiency for the Modern Data Center

Highest AI and HPC Performance

4PF FP8 (6X)| 2PF FP16 (3X)| 1PF TF32 (3X)| 67TF FP64 (3.4X)

4TB/s (2X), 96GB HBM3 memory

Transformer Engine

4th generation Tensor Core optimized for Transformer Models

6X faster on largest transformer models

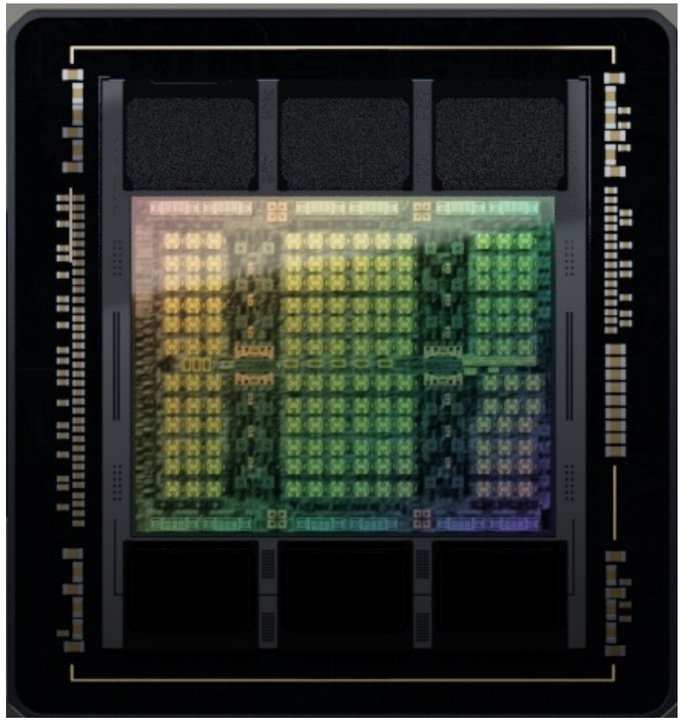
Highest Utilization Efficiency and Security

7 Fully isolated & secured instances, 2nd Gen MIG

Fastest, Scalable Interconnect

4th Gen NVLink 900 GB/s GPU-to-GPU connectivity

up to 256 linked GPUs with NVLink Switch System



10 years of evolution in GPU hardware

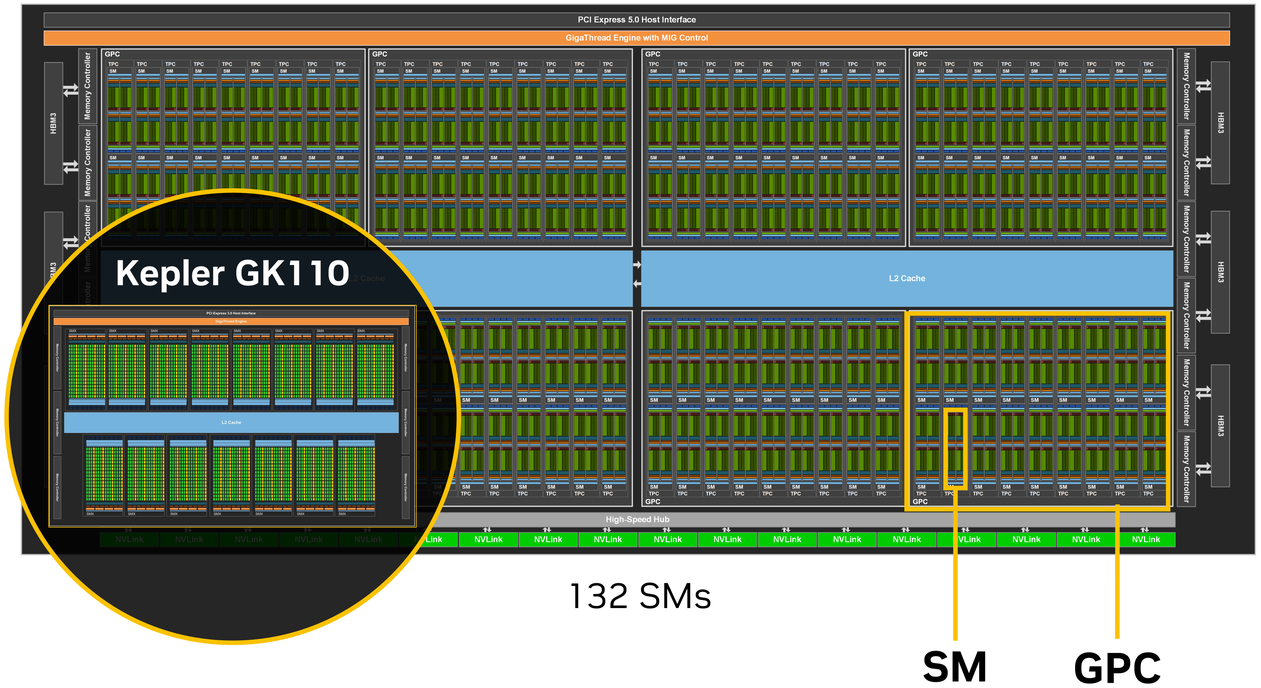
Kepler GK110 GPU (2012)

15 SMs

SM

3.52 TFLOPS single
1.17 TFLOPS double

Hopper H100 GPU (2022)



132 SMs

SM **GPC**

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

GH200 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

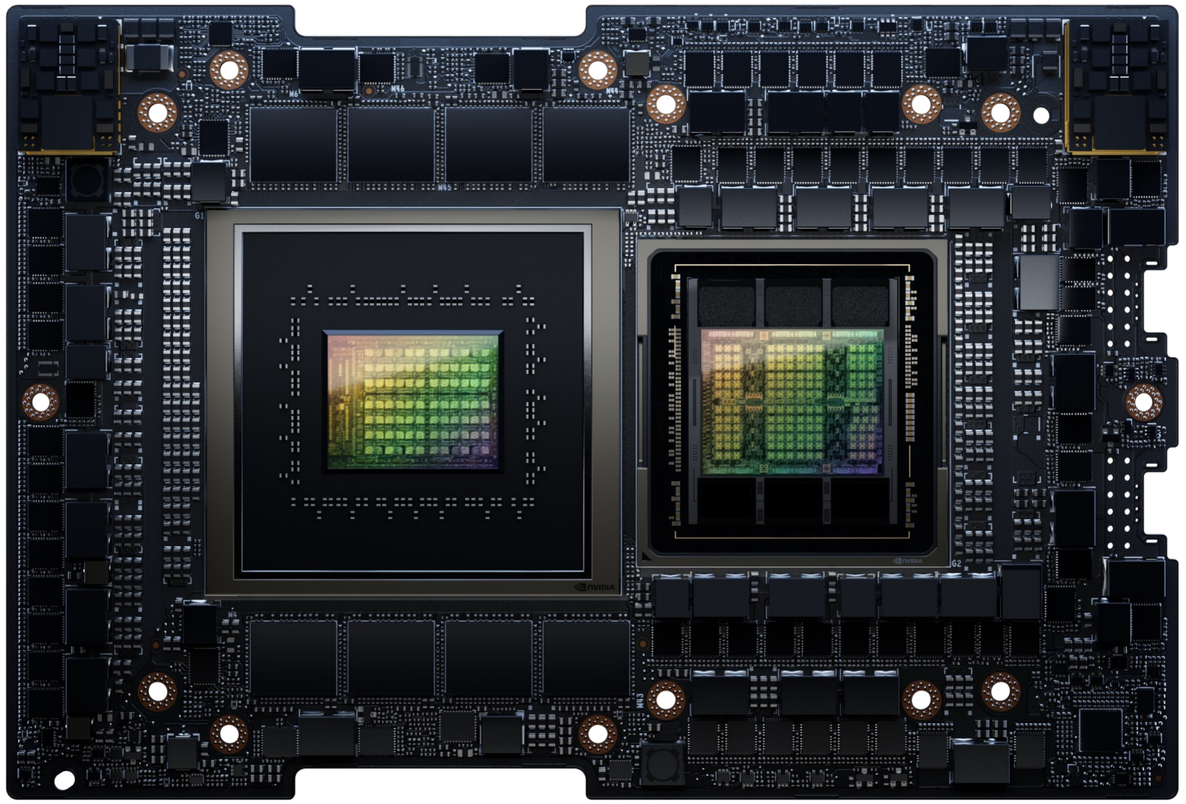
900GB/s bi-directional bandwidth CPU to GPU
7X faster than PCIe Gen 5

~600GB of Fast Access Memory

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

Full NVIDIA Compute Stack

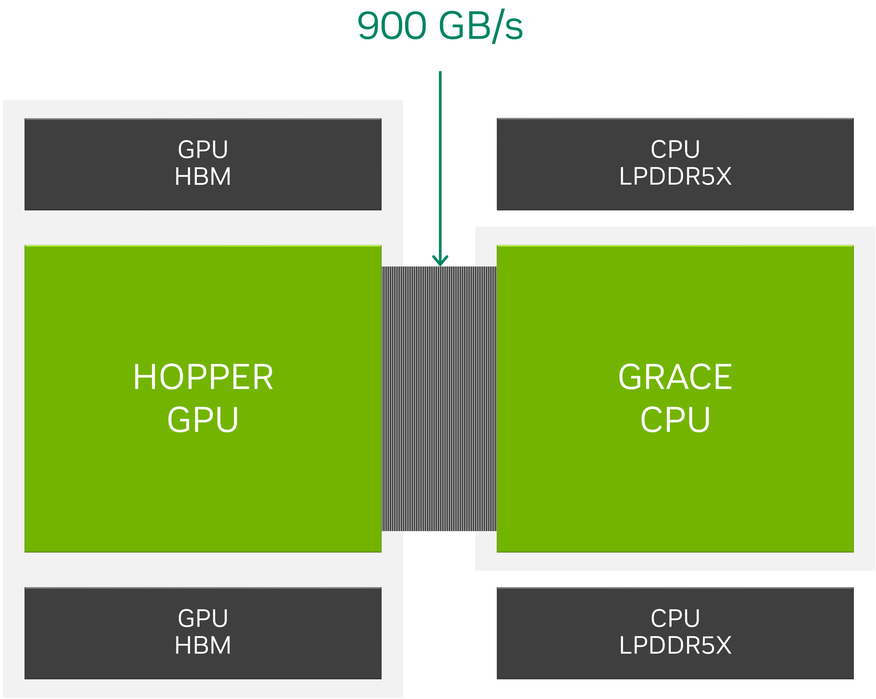
AI, Omniverse



NVLINK-C2C

High Speed Chip to Chip Interconnect

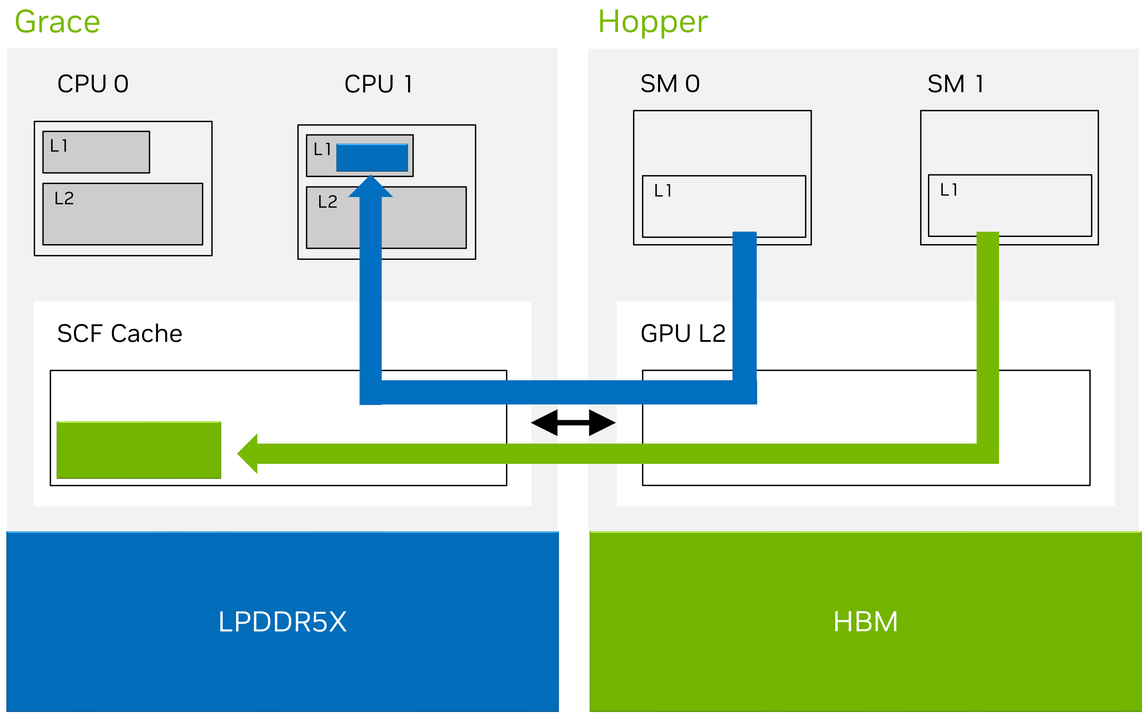
- Used to create the Grace Hopper, and Grace Superchips
  - Native atomics, including standard C++ atomic support
  - **Enables coherency**
- Up to 900 GB/s of raw bidirectional BW
  - Same BW as GPU to GPU NVLINK on Hopper
- Low power interface - 1.3 pJ/bit
  - **More than 5x more power efficient than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



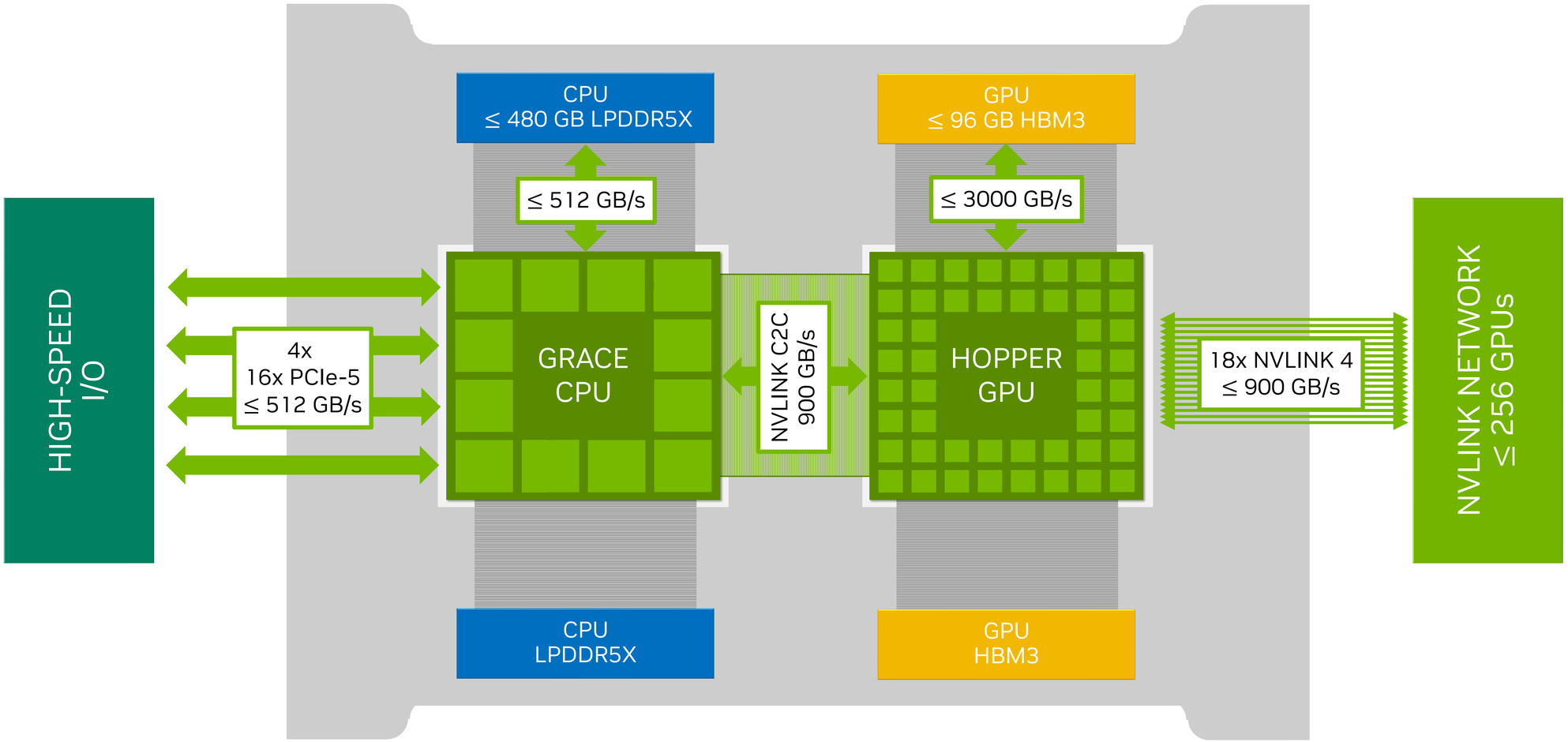
Grace Hopper Coherency

Heterogeneous Coherency

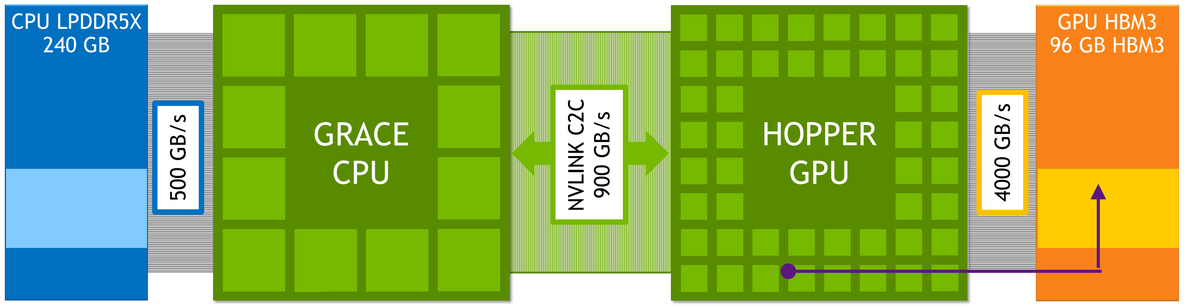
- **Grace follows the Arm memory model**
  - Coherently caches Hopper's memory
  - Data can be cached in CPU caches or SCF Cache
  - Same programming model as any other CPU memory
- **Hopper follows the CUDA memory model**
  - GPU L2 tracks if Grace has a line and will snoop Grace for data
  - GPU will snoop CPU caches for LPDDR5X memory access

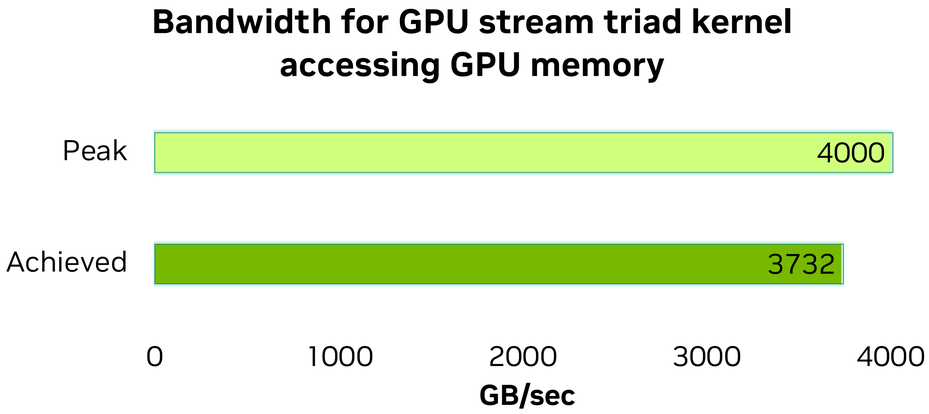


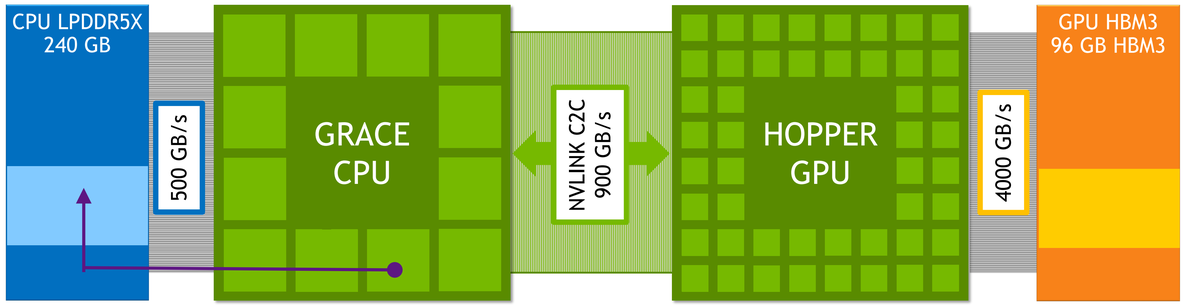
Speeds & Feeds

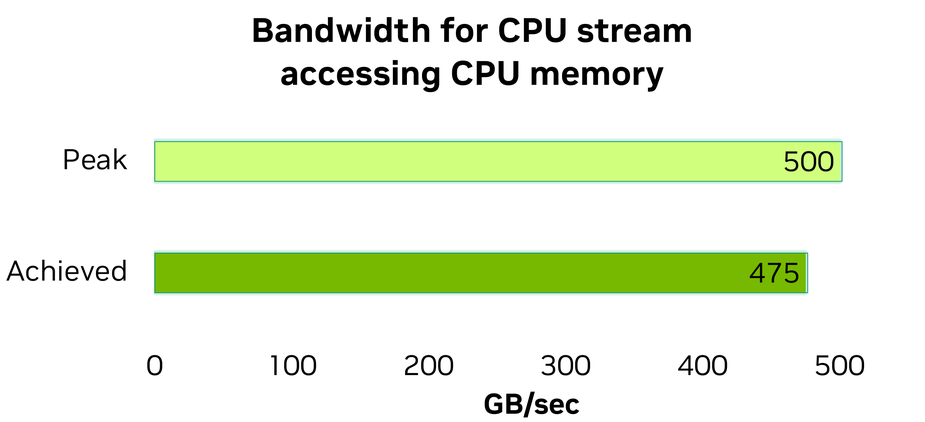


High Bandwidth Memory Access & Automatic Data Migration

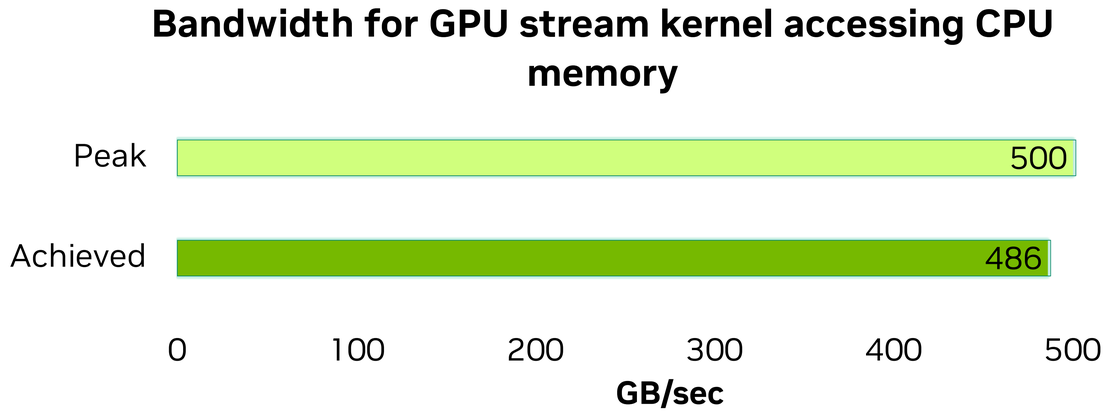






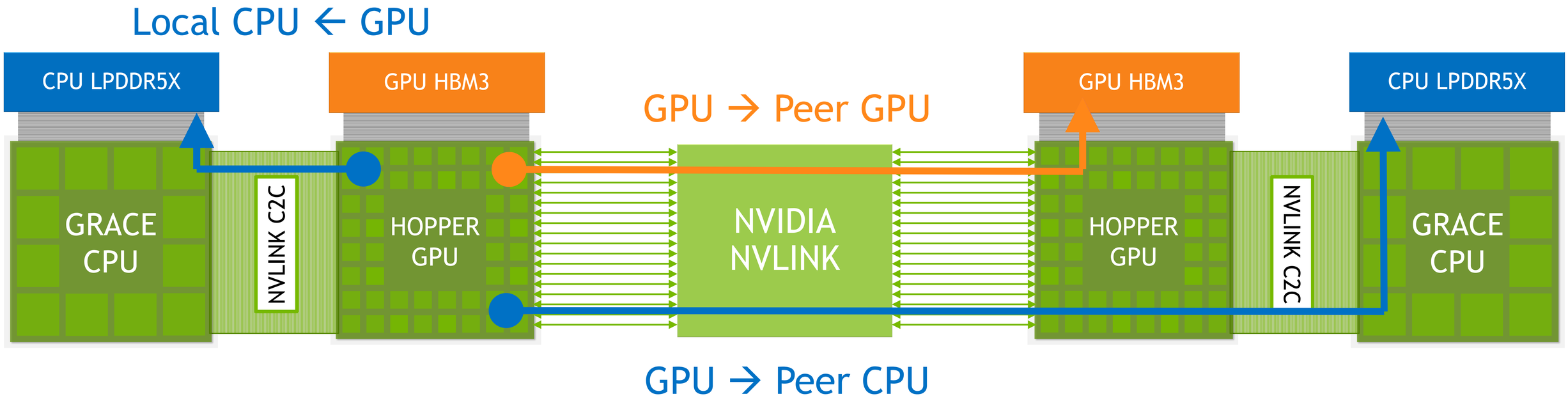


High Bandwidth Memory Access & Automatic Data Migration



NVLINK-Scaling

Superchip Scaling | CPU/GPU | Extended GPU Memory

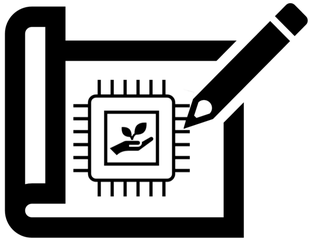


Optimizing Power and Performance at Data Center Scale

Higher Performance for Less Power

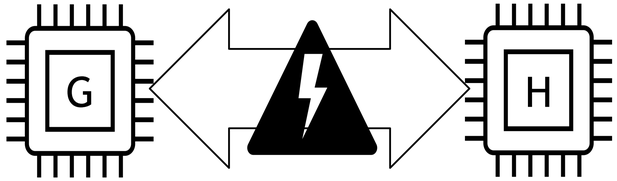
Energy Efficient Design

Energy Efficient CPU, GPU, Memory, IO



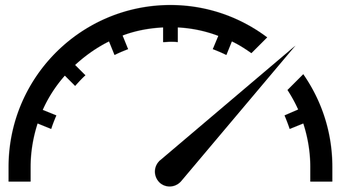
Automatic Power Shifting

Automatically shifts power between CPU & GPU



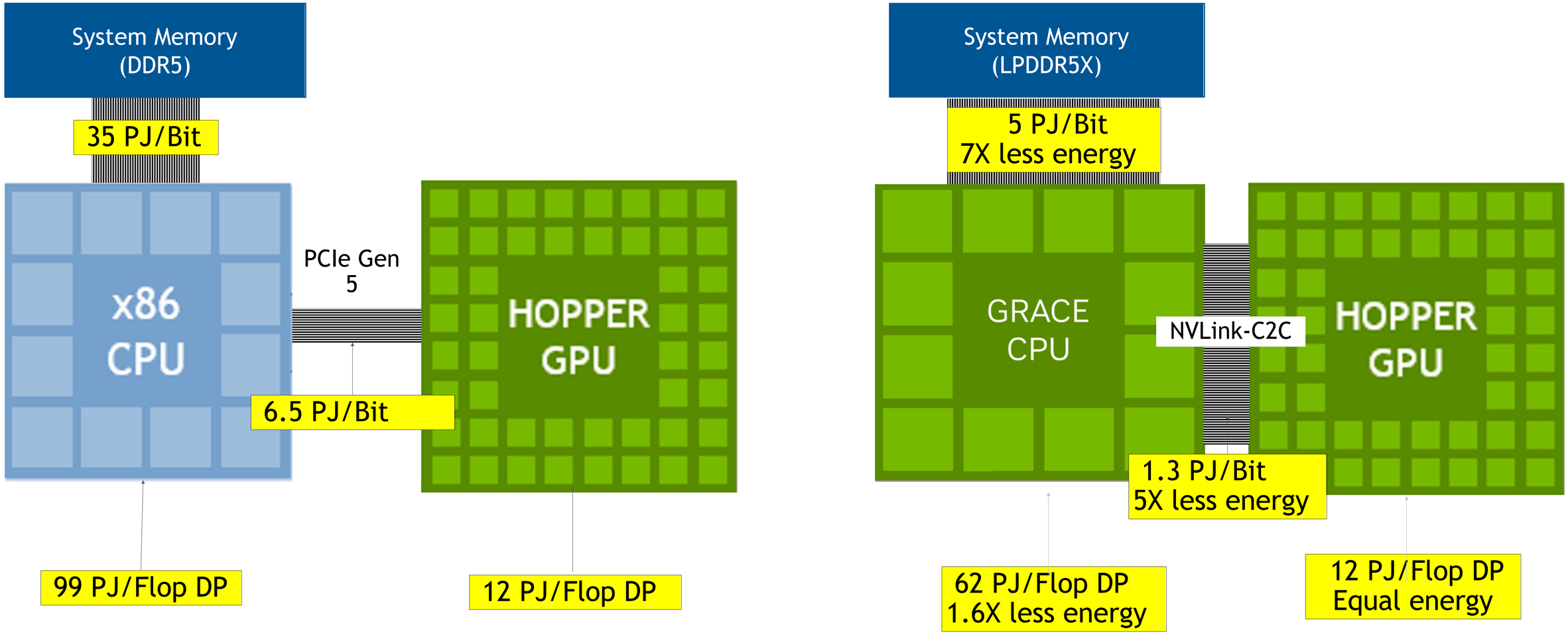
Application Power Tuning

Adjustable clocks for improved energy efficiency



Energy Efficient Design

More Efficient Computation and Data Movement



Optimizing Performance Through Power Shifting

Getting the Most Out of provisioned power