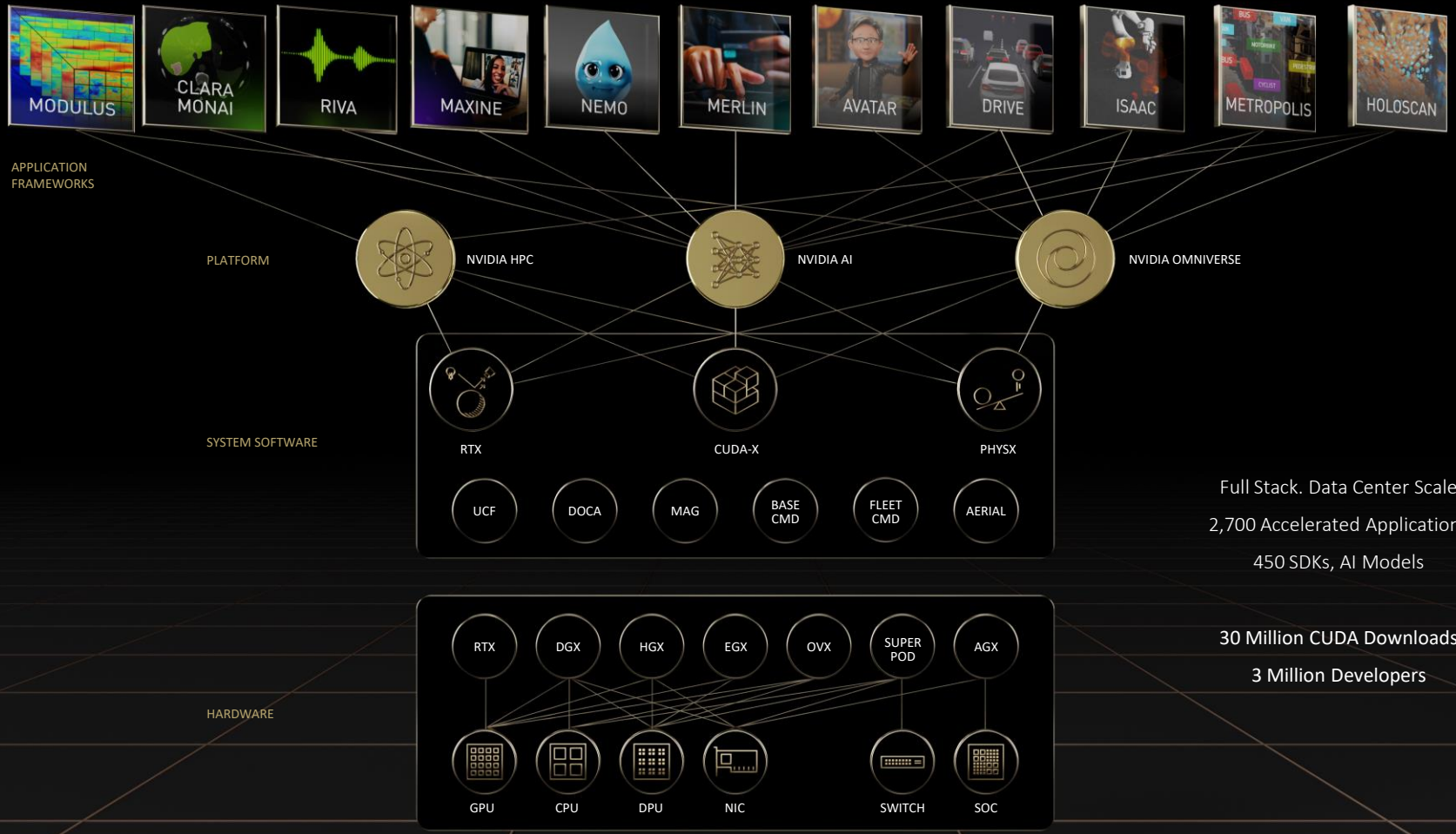
NVIDIA Hopper update

Mason Wu, NVIDIA Solutions Architect

APRIL 15, 2022



Full Stack. Data Center Scale

2,700 Accelerated Applications

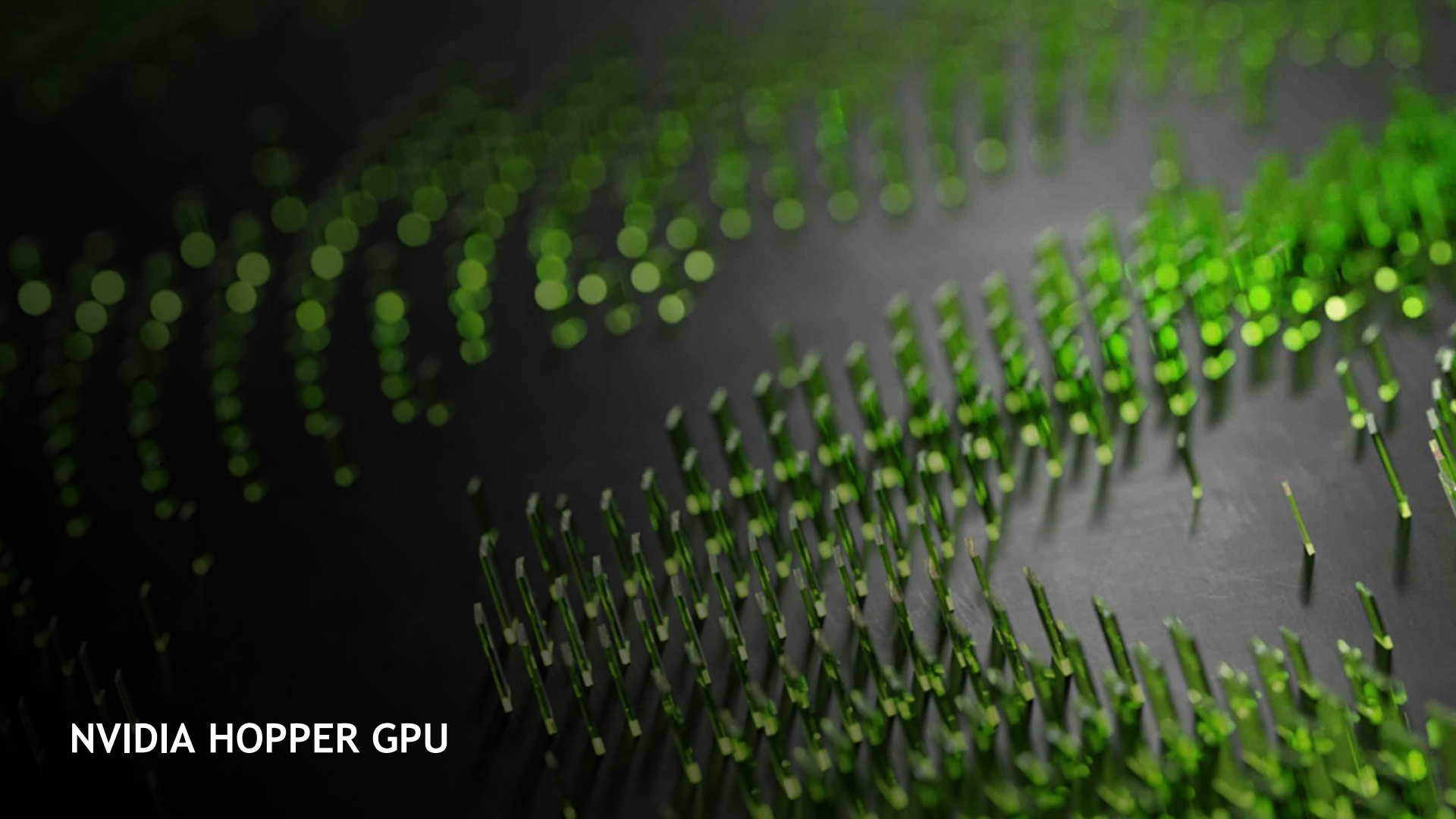
450 SDKs, AI Models

30 Million CUDA Downloads

3 Million Developers

AGENDA

- NVIDIA Hopper GPU
- NVIDIA Grace Hopper Superchip / Grace Superchip


NVIDIA HOPPER GPU

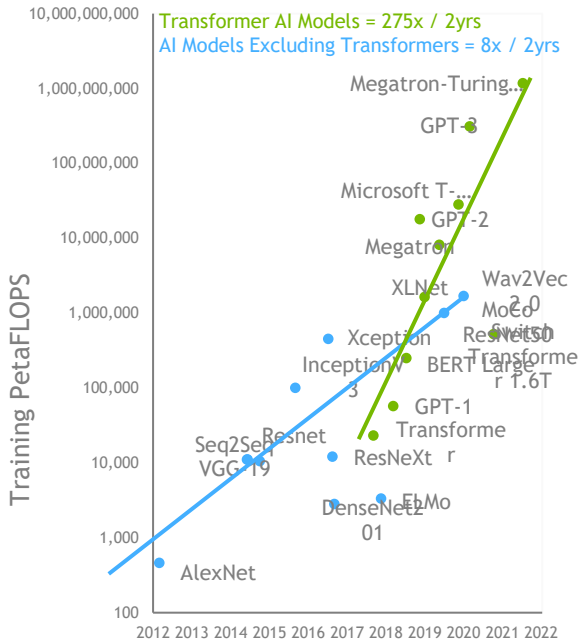
NEXT WAVE OF AI REQUIRES PERFORMANCE AND SCALABILITY

TRANSFORMERS TRANSFORMING AI

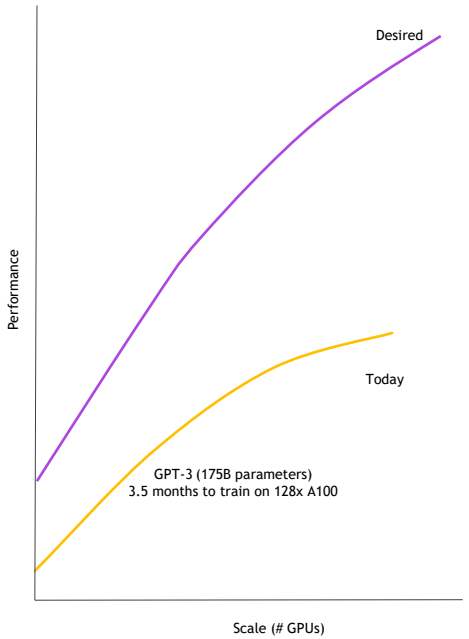
70%

AI Papers
In last 2 years discuss Transformer Models

EXPLODING COMPUTATIONAL REQUIREMENTS



HIGHER PERFORMANCE AND SCALABILITY

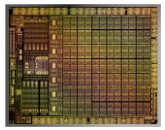


MegaMolBART: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/clara/models/megamolbart | SegFormer: https://arxiv.org/abs/2105.15203 | Decision Transformer: https://arxiv.org/pdf/2106.01345.pdf | SuperGLUE: https://super.gluebenchmark.com/leaderboard
Exploding Computational Requirements, source: NVIDIA Analysis and https://github.com/amirgholami/ai_and_memory_wall

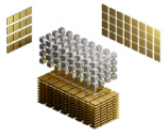
NVIDIA

ANNOUNCING NVIDIA HOPPER

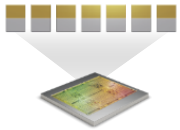
The New Engine for the World's AI Infrastructure



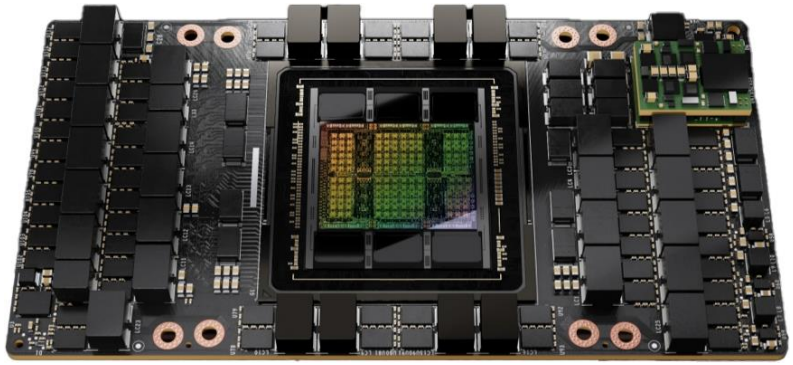
World's Most Advanced Chip



Transformer Engine



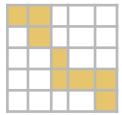
2nd Gen MIG



Custom 4N TSMC Process | 80 billion transistors

Confidential Computing

4th Gen NVLink

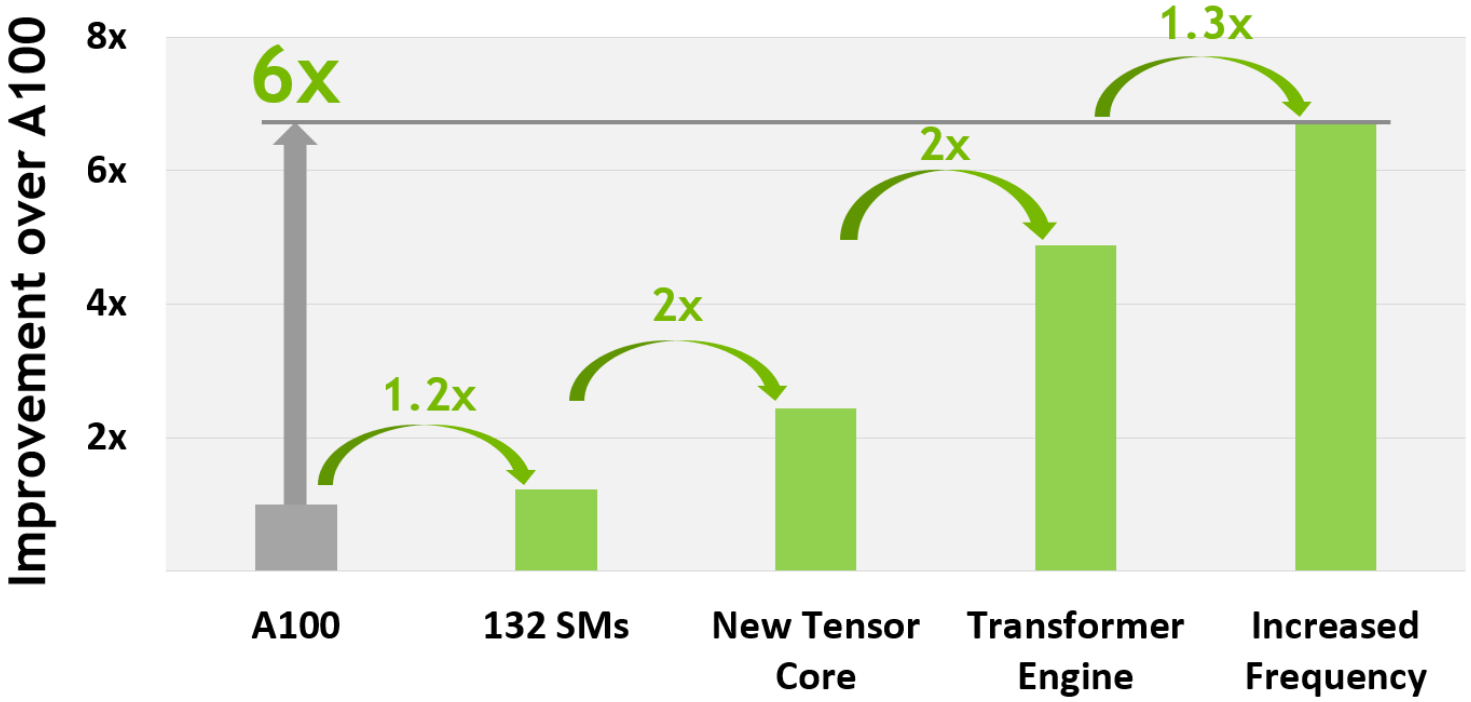


DPX Instructions

SPEEDS AND FEEDS SUMMARY

		A100-80G SXM	H100 SXM	Ratio
	# of SMs	108	132	1.2x
Scalar	FP64 TF FP32 TF	9.7 19.5	30 60	3x
Tensor	INT8 TOPS FP16 TF BF16 TF TF32 TF FP64 TF	624 312 312 156 19.5	2,000 1,000 1,000 500 60	3x
	New FP8 TF	-	2,000	**6x** vs A100 16b
	Memory Capacity	80 GB	80 GB	1x
	DRAM BW	2 TB/s	3 TB/s	1.5x
	NVLink BW	600 GB/s	900 GB/s	1.5x
	NVLink Domain	8	256	32x
	PCIe BW	64 GB/s	128 GB/s	2x

H100 COMPUTE IMPROVEMENTS SUMMARY

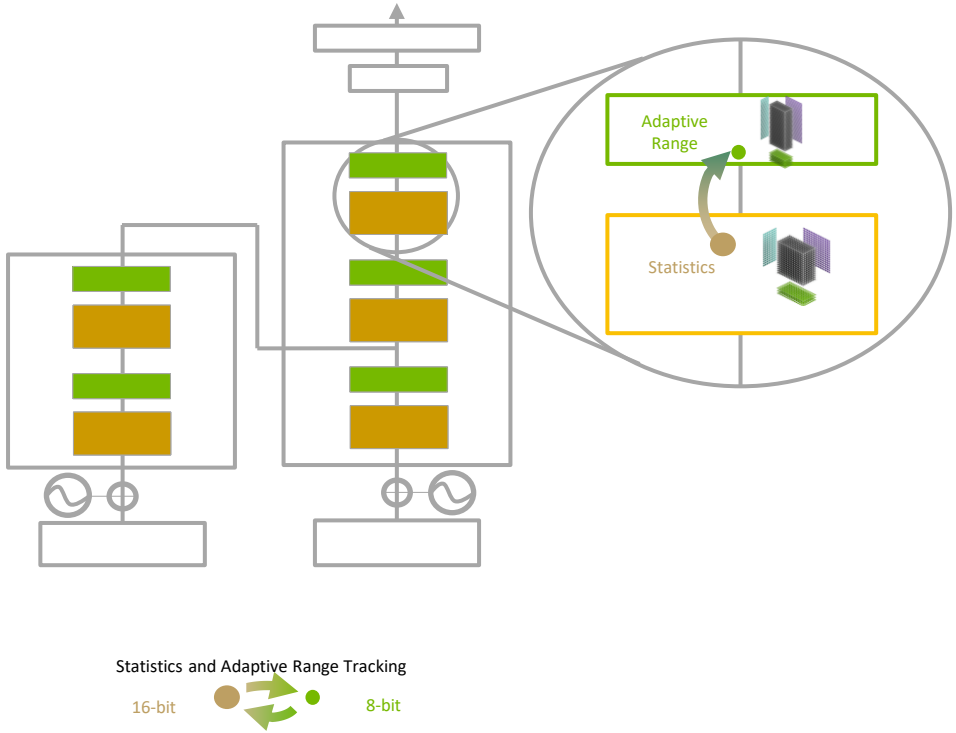


6x throughput for the world's most compute-hungry workloads

TRANSFORMER ENGINE

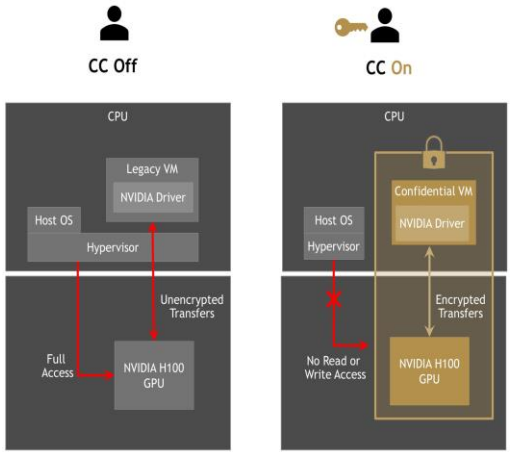
Tensor Core Optimized for Transformer Models

- 6X Faster Training and Inference of Transformer Models
- NVIDIA Tuned Adaptive Range Optimization Across 16-bit and 8-bit Math
- Configurable Macro Blocks Deliver Performance Without Accuracy Loss

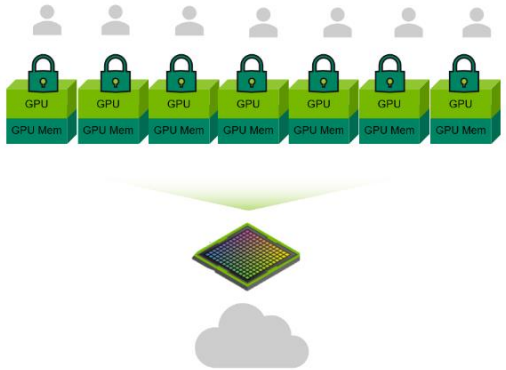


HOPPER TECHNOLOGICAL BREAKTHROUGHS

CONFIDENTIAL COMPUTING
Secure Data and AI Models In-Use

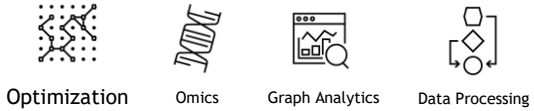


MULTI-GPU INSTANCE
7 Secure Tenants on 1 GPU

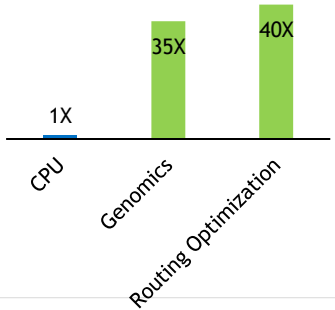


NEW DYNAMIC PROGRAMING INSTRUCTIONS
Accelerate Dynamic Programming Algorithms

A BROAD RANGE OF USE CASES



REAL-TIME PERFORMANCE



Dynamic Programming Instructions HGX H100 4-GPU vs dual socket 32 core IceLake

INSIDE 8-BIT FLOATING POINT (FP8)

E5M2

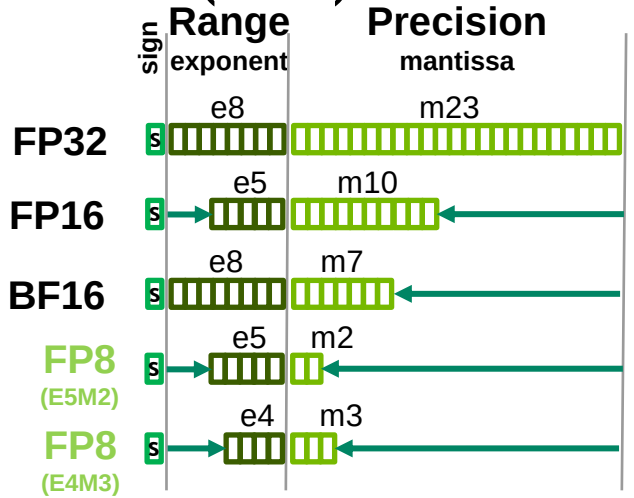
- Encoding for: Infs, NaNs, zeros
- Dynamic range: 32 powers of 2
- Precision: 4 samples between powers of two

E4M3

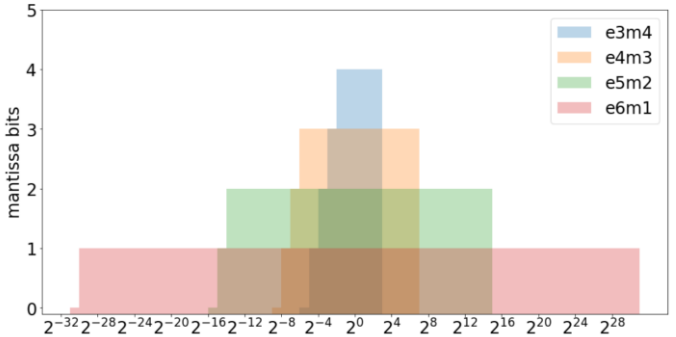
- Encoding for: NaNs, zeros
  - No Infs: instead we extend the dynamic range
- Dynamic range: 18 powers of 2
- Precision: 8 samples between powers of two

Both types support

- Denormals
- Saturating conversions from wider types
  - Values exceeding max representable fp8 value *x* get saturated to *x*



Allocate 1 bit to either range or precision



NVIDIA

NVLINK SWITCH SYSTEM

Enabling Multi-Node NVLink Up to 256 GPUs

4th GEN NVLINK

900 GB/s from 18x25GB/sec bi-directional ports
GPU-2-GPU connectivity across nodes

3rd GEN NVSWITCH

All-to-all NVLink switching for 8-256 GPUs
Accelerate collectives - multicast and SHARP

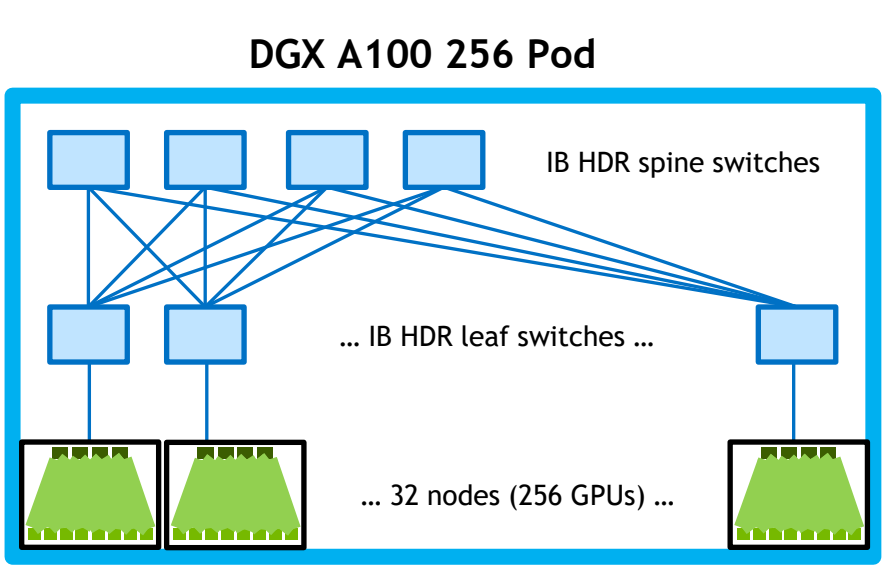
NVLINK SWITCH

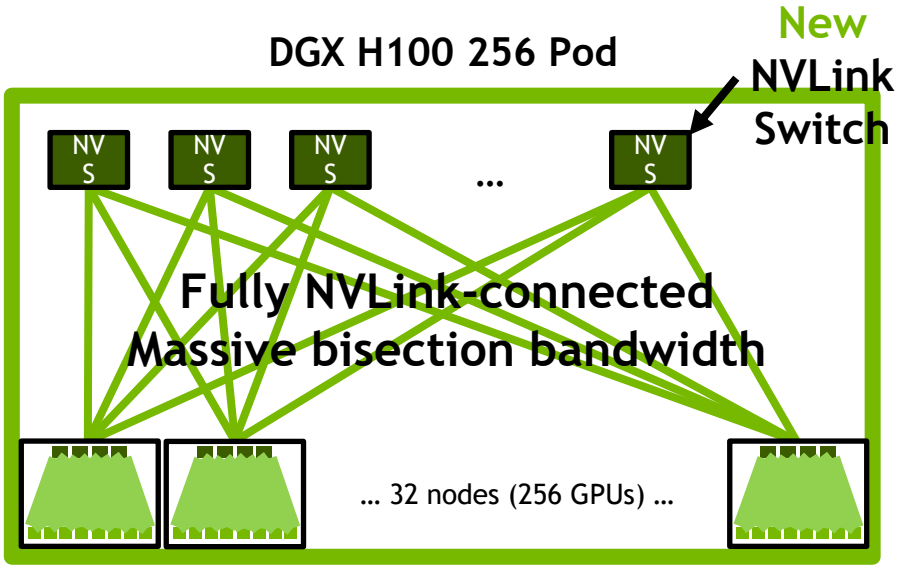
128 port cross-connect based on NVSwitch

H100 CLUSTER (1 SCALABLE UNIT)

57,600 GB/s all-to-all bandwidth
32 servers | 18 NVLink switches | 1,152 NVLink optical cables

SCALE-UP WITH NVLINK NETWORK





	A100 SuperPod			H100 SuperPod			Speedup	
	Dense PFLOP/s	Bisection [GB/s]	Reduce [GB/s]	Dense PFLOP/s	Bisection [GB/s]	Reduce [GB/s]	Bisection	Reduce
1 DGX / 8 GPUs	2.5	2,400	150	16	3,600	450	**1.5x**	**3x**
32 DGXs / 256 GPUs	80	6,400	100	512	57,600	450	**9x**	**4.5x**

ANNOUNCING NVIDIA EOS SUPERCOMPUTER

The World's Most Advanced AI Infrastructure

NVIDIA Eos

DGX SuperPOD Powered by 576 DGX H100 Systems |
500 Quantum-2 IB Switches | 360 NVLink Switches

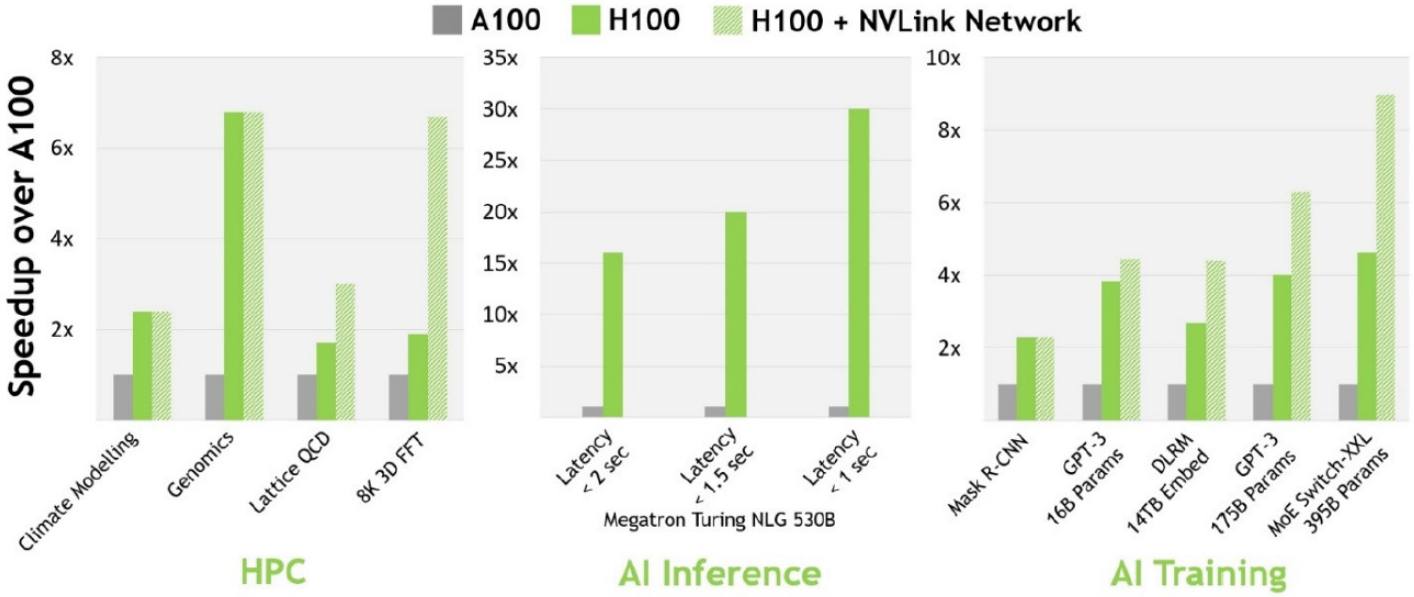
FP8	18 EFLOPS	6X
FP16	9 EFLOPS	3X
FP64	275 PFLOPS	3X
In-Network Compute	3.7 PFLOPS	36X
Bisection Bandwidth	230 TB/s	2X
NVLINK Domain	256 GPUs	32X

Blueprint for OEM and Cloud Partner Offerings

Cloud Native | Performance Isolation | Multi-Tenant

X-Factors compare performance over DGX A100 SuperPOD based supercomputer configuration with same number of Nodes

H100 ENABLES NEXT-GENERATION AI AND HPC BREAKTHROUGHS



All performance numbers are preliminary based on current expectations and subject to change in shipping products. A100 cluster: HDR IB network. H100 cluster: NDR IB network with NVLink Switch System where indicated.

GPUs: Climate Modeling 1K, LQCD 1K, Genomics 8, 3D-FFT 256, MT-NLG 32 (batch sizes: 4 for A100, 60 for H100 at 1 sec, 8 for A100 and 64 for H100 at 1.5 and 2sec),MRCNN 8 (batch 32),GPT-3 16B 512 (batch 256), DLRM 128 (batch 64K), GPT-3 16K (batch 512), MoE 8K (batch 512, one expert per GPU)

NVIDIA H100 PCIE

Unprecedented Performance, Scalability, and Security for Mainstream Servers

HIGHEST AI AND HPC MAINSTREAM PERFORMANCE

3.2PF FP8 (5X) | 1.6PF FP16 (2.5X) | 800TF TF32 (2.5X) | 48TF FP64 (2.5X)
6X faster Dynamic Programming with DPX Instructions
2TB/s , 80GB HBM2e memory

HIGHEST COMPUTE ENERGY EFFICIENCY

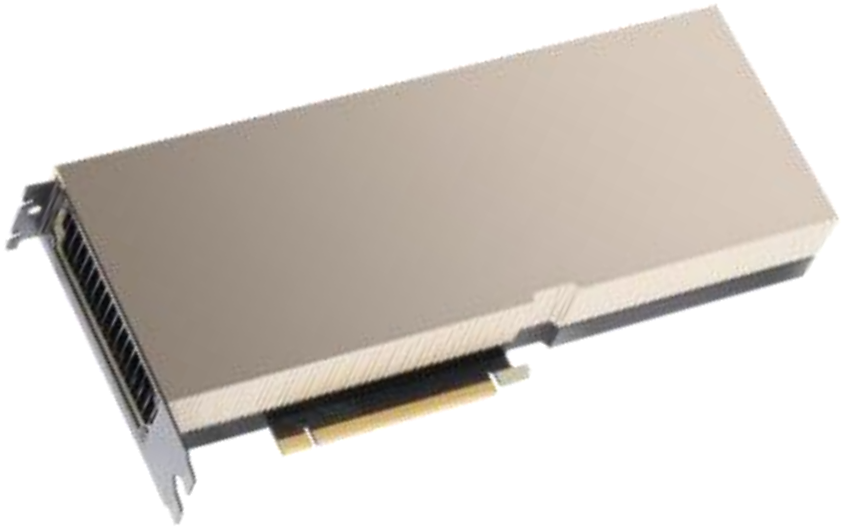
Configurable TDP - 150W to 350W
2 Slot FHFL mainstream form factor

HIGHEST UTILIZATION EFFICIENCY AND SECURITY

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

HIGHEST PERFORMING SERVER CONNECTIVITY

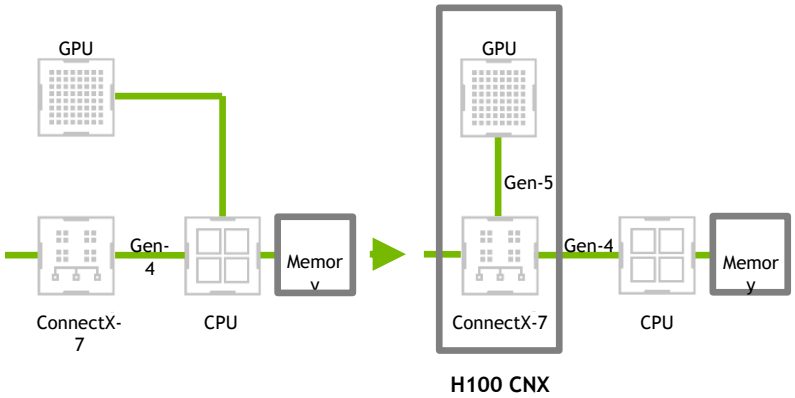
128GB/s PCI Gen5
600 GB/s GPU-2-GPU connectivity (5X PCIe Gen5)
up to 2 GPUs with NVLink Bridge



FP8, FP16, TF32 performance include sparsity. X-factor compared to A100

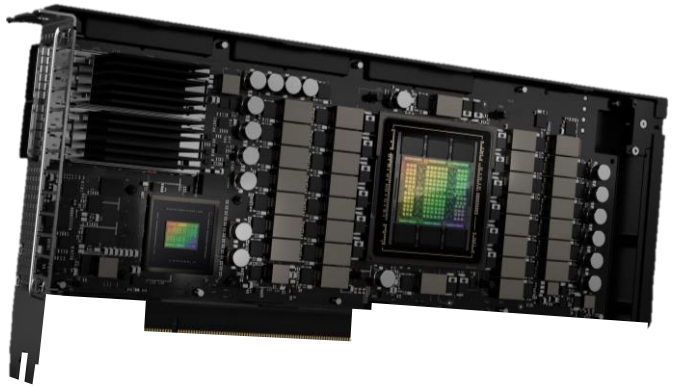
ANNOUNCING H100 CNX CONVERGED ACCELERATOR

Delivering High-Speed GPU-Network I/O to Mainstream Servers

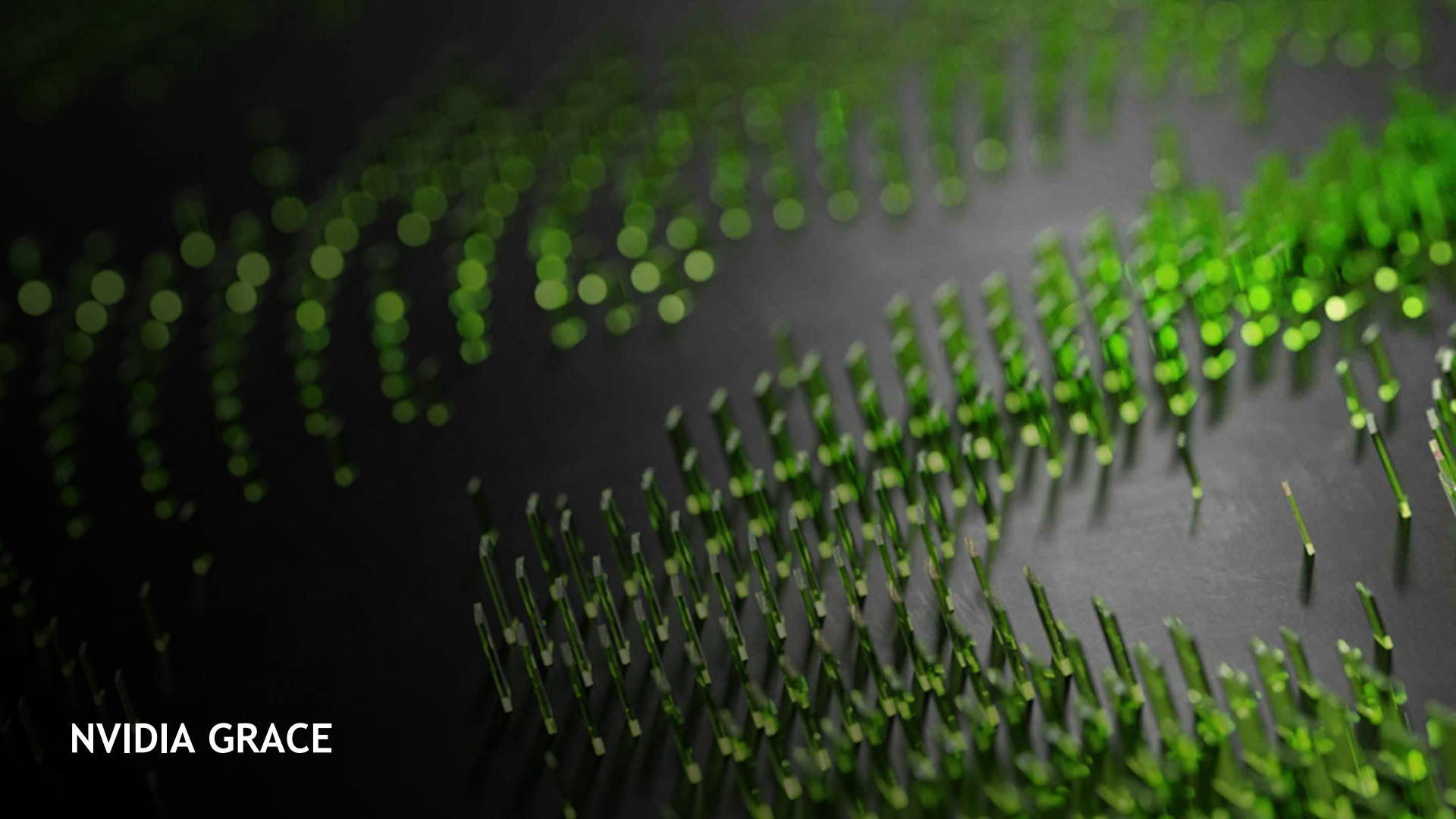


Traditional Server

Optimized for Accelerated Computing



350W | 80GB | 400 Gb/s Eth or IB
PCIe Gen 5 within board and to host
2-Slot FHFL | NVLink


NVIDIA GRACE

ANNOUNCING GRACE HOPPER

CPU+GPU Designed for Giant Scale AI and HPC

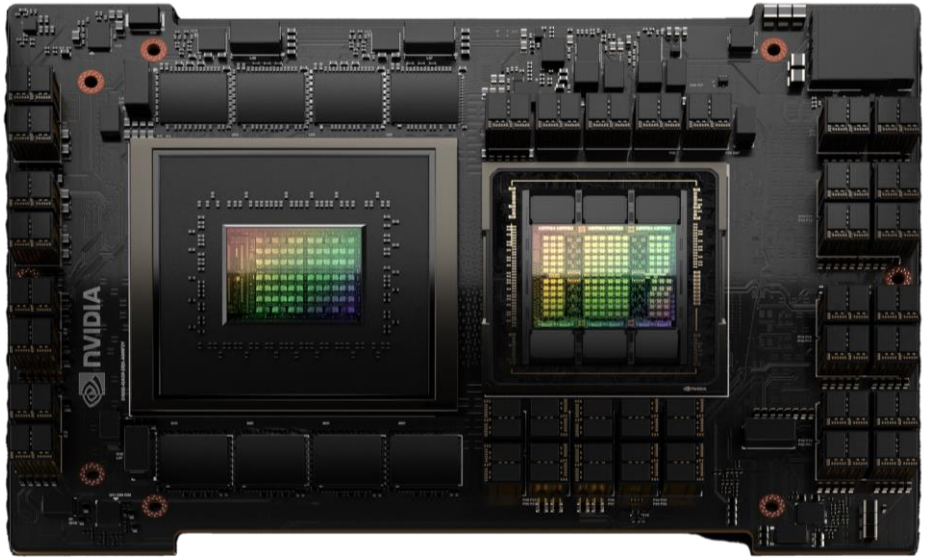
600GB Memory GPU for Giant Models

New 900 GB/s Coherent Interface

30X Higher System Memory B/W to GPU In A Server

Runs Nvidia Computing Stacks

Available 1H 2023



ANNOUNCING GRACE CPU SUPERCHIP

The Full Power of the Grace

HIGHEST CPU PERFORMANCE

Superchip Design with 144 high-performance Armv9 Cores
Estimated Specrate2017_int_base of over 740

HIGHEST MEMORY BANDWIDTH

World's first LPDDR5x memory with ECC, 1TB/s Memory Bandwidth

HIGHEST ENERGY EFFICIENCY

2X Perf/Watt, CPU Cores + Memory in 500W

2X PACKING DENSITY

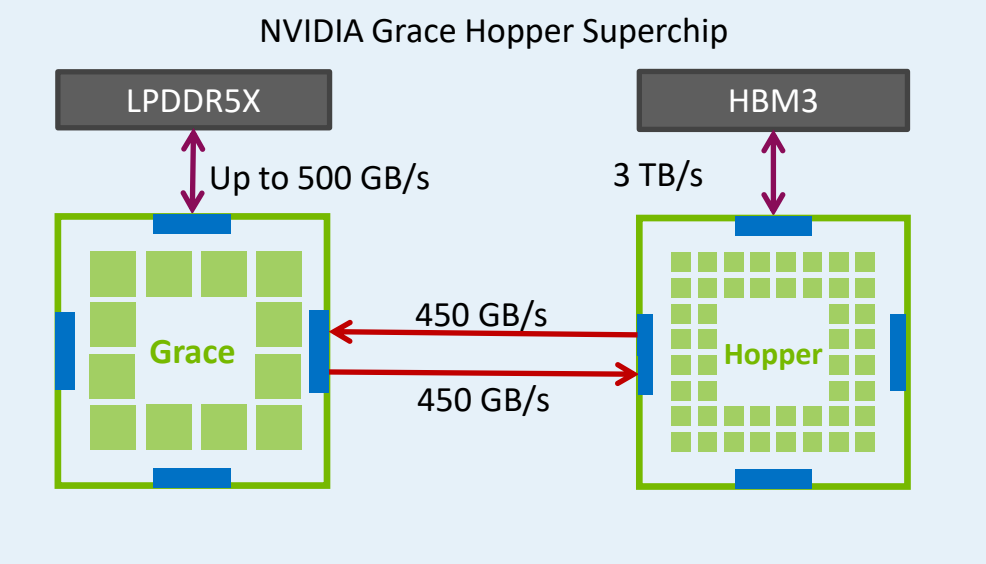
2x density of DIMM based designs

RUNS FULL NVIDIA COMPUTING STACKS

RTX, HPC, AI, Omniverse

AVAILABLE 1H 2023

NVLINK C2C


NVIDIA Grace Hopper Superchip
LPDDR5X
HBM3
Up to 500 GB/s
3 TB/s
Grace
Hopper
450 GB/s
450 GB/s


NVIDIA