HOPPER: NVIDIA'S NEW GPU ARCHITECTURE & WHY IT MATTERS!

TIM WOOD SR. SOLUTION ARCHITECT, FSI-EMEA

NVIDIA DATACENTER PLATFORM

Scalable Platform for HPC and Inovation in AI

BUSINESS APPLICATIONS

DL for Securities Valuation | Synthetic Data Generation | Predictive Analytics | Accelerated Risk Analytics | Portfolio Optimization | Pattern and Anomaly Detection | Advanced NLP/NLU | ++

NGC

SOFTWARE HUB

HELM

Certified Containers

Pre-trained Models

SDKs

APPLICATION FRAMEWORKS

DATA SCIENCE	CONVERSATIONAL AI	IVA	RECOMMENDATION SYSTEMS	SIMULATION	CYBERSECURITY	++
RAPIDS	*Riva*	*Metropolis*	*Merlin*	*Omniverse*	*Morpheus*	…

DEVELOPER TOOLKITS

ML & DATA ANALYTICS	AI TRAINING & INFERENCE	HIGH PERFORMANCE COMPUTING	RENDERING & VISUALIZATION
RAPIDS, dmlc XGBoost	TensorFlow, TensorRT, PYTORCH, mxnet	*NVIDIA HPC SDK*	*IndeX, OptiX, CloudXR, MDL*

ACCELERATION LIBRARIES

COMPUTE	NETWORKING, STORAGE & SECURITY
CUDA-X	*DOCA, MAGNUM IO*

OPERATIONS

VIRTUAL GPU SW

TRITON INFERENCE SERVER

FLEET COMMAND

NVIDIA GPU Operator

Red Hat

vmware

NVIDIA CERTIFIED

VALIDATED SOLUTIONS

SERVERS & CLOUD

DGX HGX

Purpose Built

EGX

Mainstream & Edge

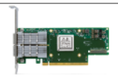
CSP Instances

HARDWARE TECHNOLOGIES

GPU

NVSwitch

BF DPU



SMART NIC

NVIDIA Switch

MONITORING

DCGM

Grafana

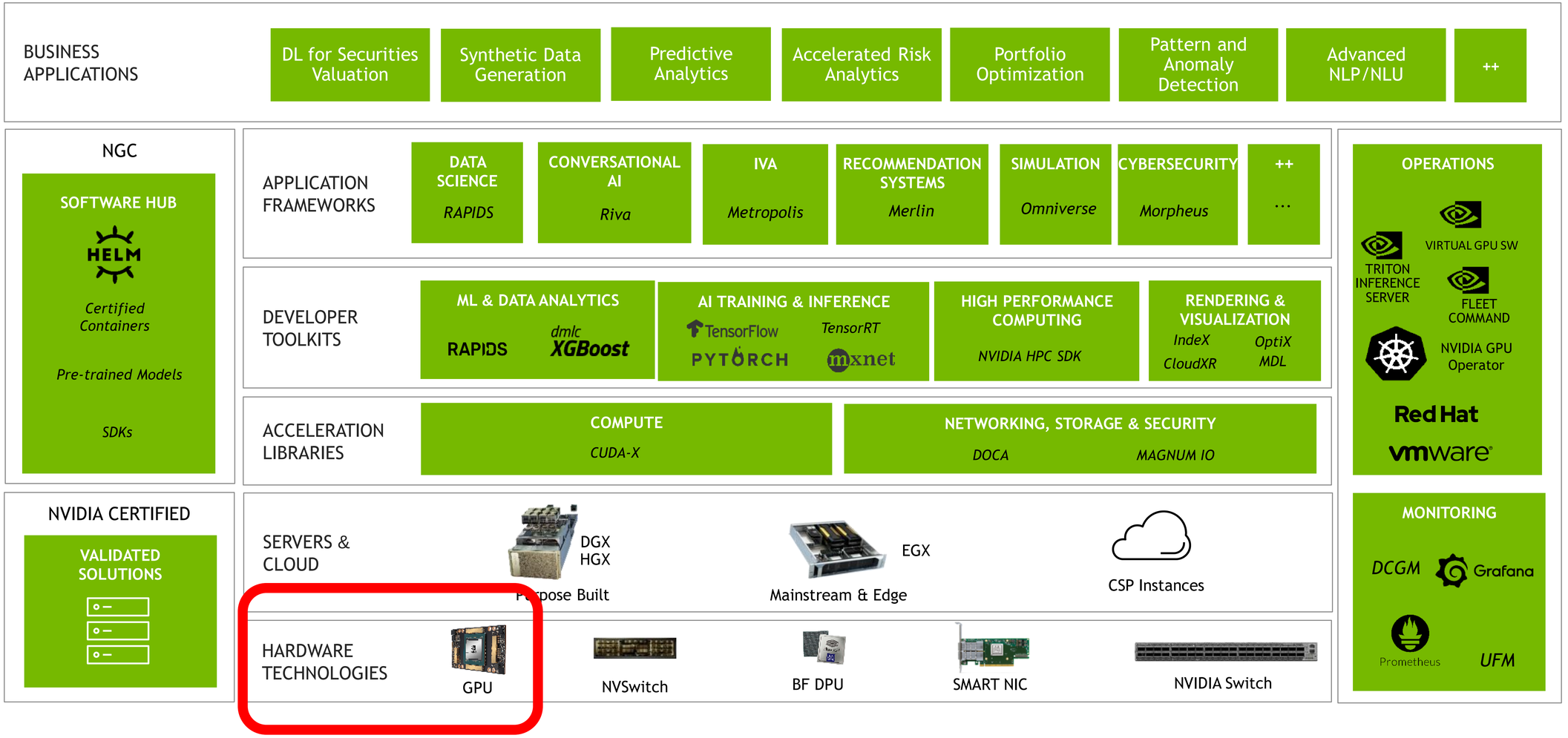
Prometheus

UFM

NVIDIA

NVIDIA DATACENTER PLATFORM

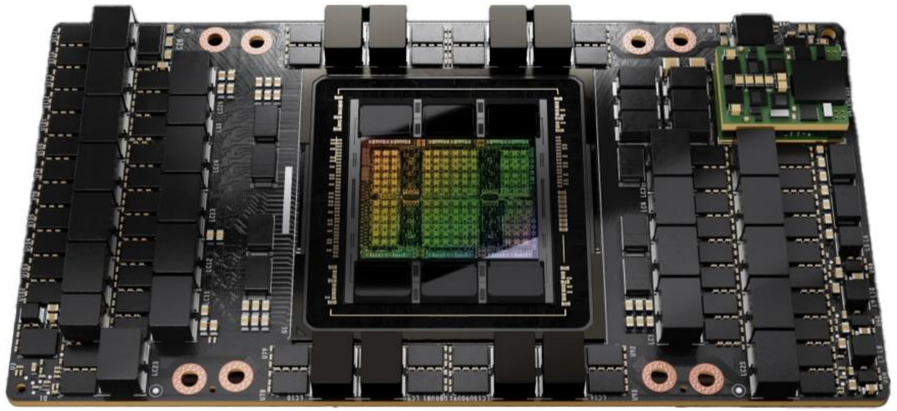
Scalable Platform for HPC and Inovation in AI

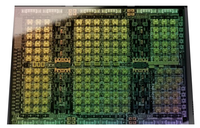


NVIDIA H100

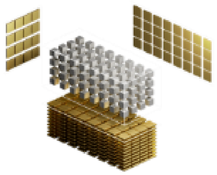
Unprecedented Performance, Scalability, and Security for Every Data Center

- **Highest 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 Model Optimizations**
  6X faster on largest transformer models
- **Highest Utilization Efficiency and Security**
  7 Fully isolated & secured instances, guaranteed QoS
  2nd Gen MIG | Confidential Computing
- **Fastest, Scalable Interconnect**
  900 GB/s GPU-2-GPU connectivity (1.5X)
  up to 256 GPUs with NVLink Switch | 128GB/s PCI Gen5

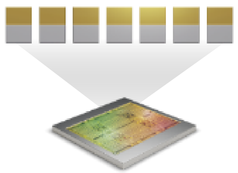




World's Most Advanced Chip



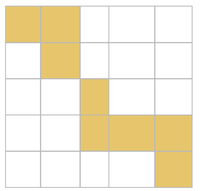
Transformer Engine



2nd Gen MIG

Confidential Computing

4th Gen NVLink

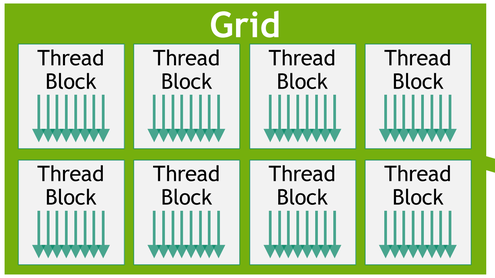


DPX Instructions

Not STAC Benchmarks

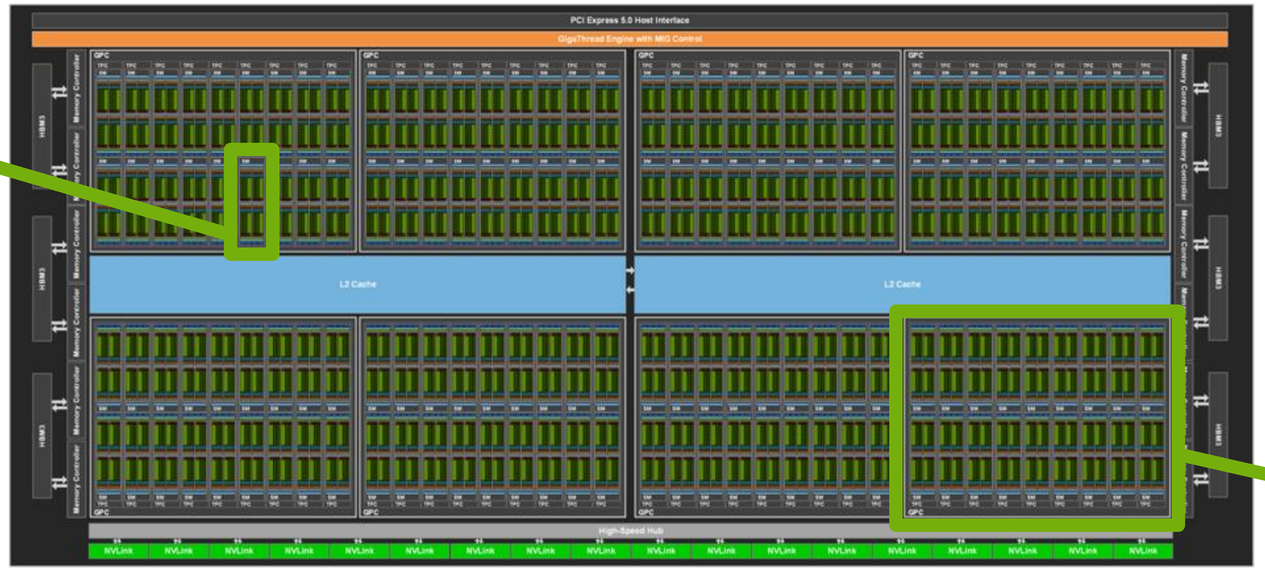
HOPPER & CUDA 12 ONWARDS

NEW: Thread Block Clusters and the Tensor Memory Accelerator



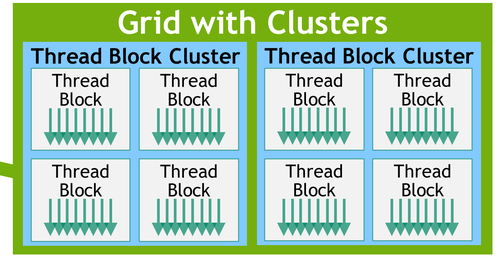
CUDA Till Version 11.x

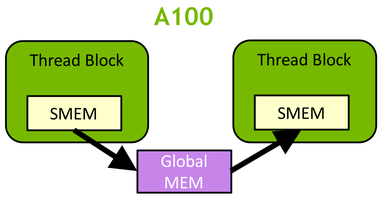
- Execution Grids composed of *thread blocks*
- Thread blocks map to *streaming multiprocessors*
- Threads may communicate via limited *shared memory*



CUDA 12 and onwards

- New level in thread hierarchy called *thread block clusters*
- Thread block clusters map to *GPCs*
- Threads comprising a thread block cluster communicate via *distributed shared memory*



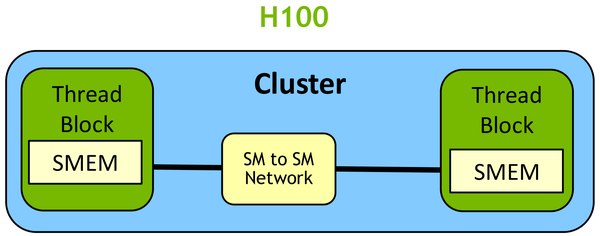


Thread Block Clusters

- Map naturally to the GPU architecture
- Are guaranteed to be scheduled together and have fast synchronization
- With fast asynchronous communication enabled by the...

Tensor Memory Accelerator

- Manages distributed shared memory within the GPC
- Enables thread direct block-to-block communication with local barrier synchronization



NVIDIA DATACENTER PLATFORM

Scalable Platform for HPC and Inovation in AI

BUSINESS APPLICATIONS

DL for Securities Valuation | Synthetic Data Generation | Predictive Analytics | Accelerated Risk Analytics | Portfolio Optimization | Pattern and Anomaly Detection | Advanced NLP/NLU | ++

NGC

SOFTWARE HUB

HELM

Certified Containers

Pre-trained Models

SDKs

APPLICATION FRAMEWORKS

DATA SCIENCE	CONVERSATIONAL AI	IVA	RECOMMENDATION SYSTEMS	SIMULATION	CYBERSECURITY	++
RAPIDS	*Riva*	*Metropolis*	*Merlin*	*Omniverse*	*Morpheus*	…

DEVELOPER TOOLKITS

ML & DATA ANALYTICS	AI TRAINING & INFERENCE	HIGH PERFORMANCE COMPUTING	RENDERING & VISUALIZATION
RAPIDS, dmlc XGBoost	TensorFlow, TensorRT, PYTORCH, mxnet	*NVIDIA HPC SDK*	*IndeX, OptiX, CloudXR, MDL*

ACCELERATION LIBRARIES

COMPUTE	NETWORKING, STORAGE & SECURITY
CUDA-X	*DOCA, MAGNUM IO*

OPERATIONS

VIRTUAL GPU SW

TRITON INFERENCE SERVER

FLEET COMMAND

NVIDIA GPU Operator

Red Hat

vmware

NVIDIA CERTIFIED

VALIDATED SOLUTIONS

SERVERS & CLOUD

DGX HGX

Purpose Built

EGX

Mainstream & Edge

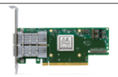
CSP Instances

HARDWARE TECHNOLOGIES

GPU

NVSwitch

BF DPU



SMART NIC

NVIDIA Switch

MONITORING

DCGM

Grafana

Prometheus

UFM

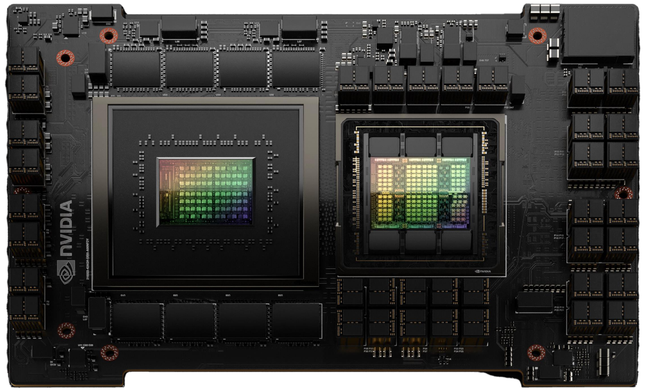
NVIDIA

GRACE HOPPER & GRACE SUPERCHIP

GPU+CPU and CPU+CPU Modules Designed for Giant Scale AI and HPC

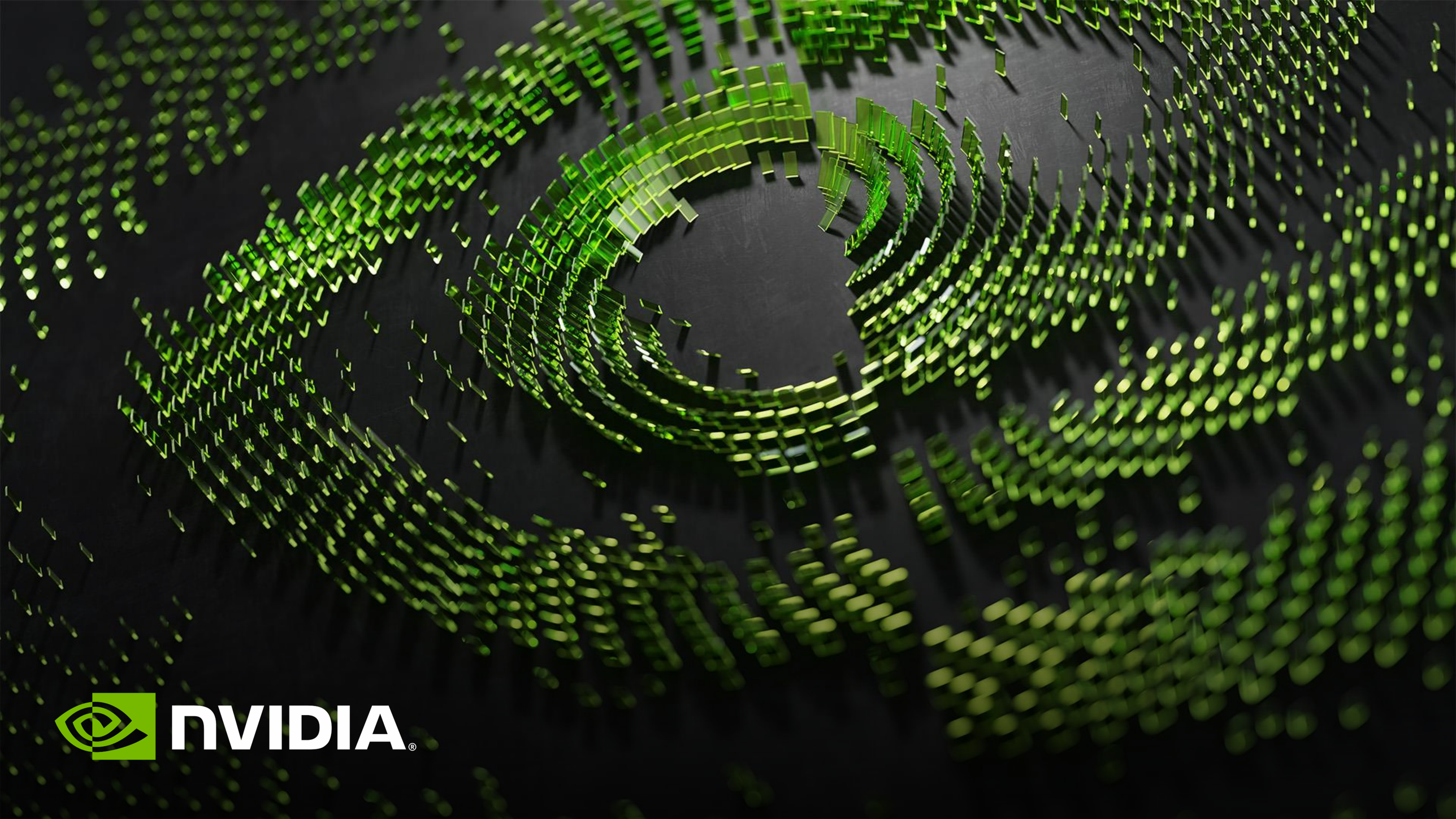
Grace Hopper

- 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



Grace Superchip

- Grace Superchip
- Highest CPU Performance, with 144 high-performance Armv9 Cores
- Highest memory bandwidth with world's first LPDDR5x memory with ECC, 1TB/s Memory Bandwidth
- HIGHEST ENERGY EFFICIENCY
- 2X Packing Density compared to DIMM based design
- Runs all NVIDIA AI and HPC computing stacks


NVIDIA