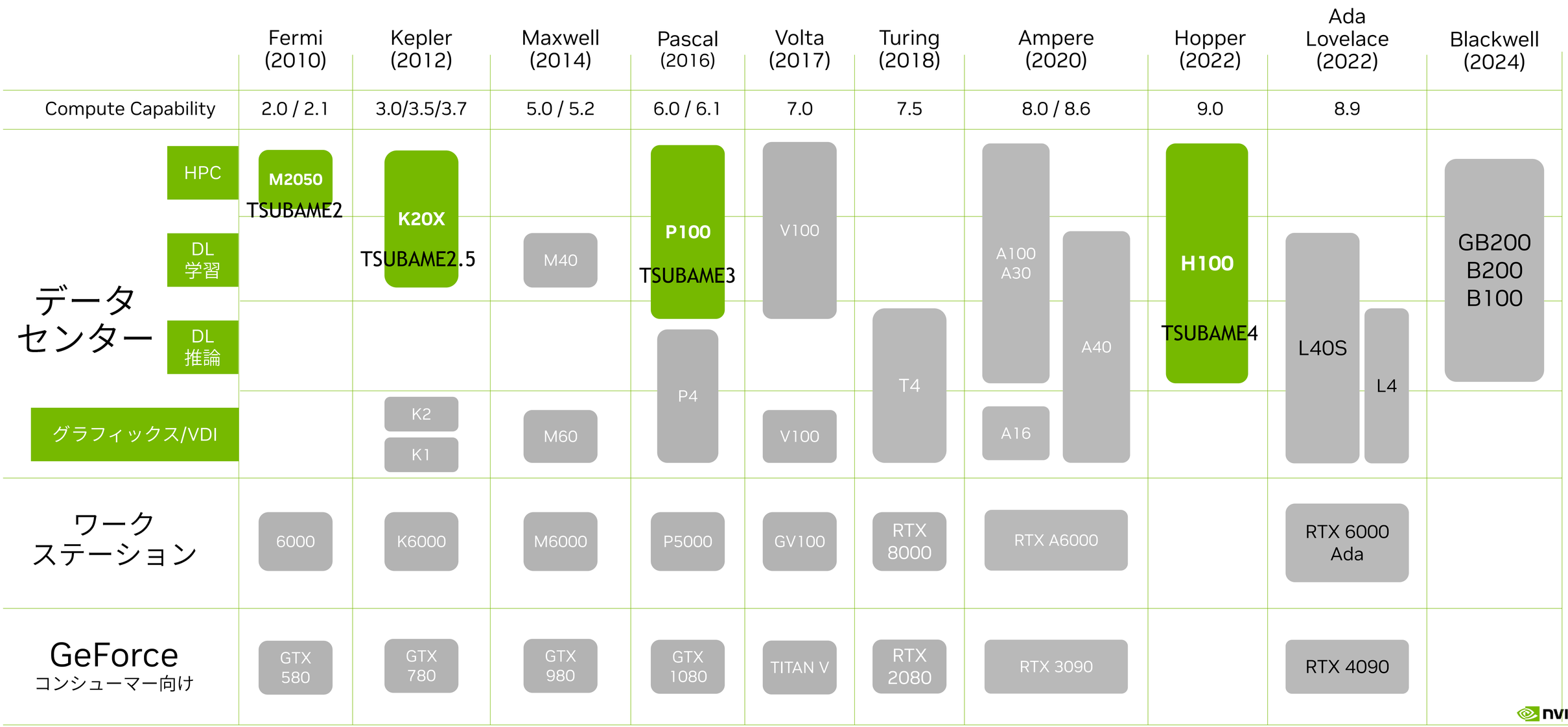
AI と HPC の進化を支えるエヌビディアのコンピューティングプラットフォーム

2024/6/27

エヌビディア合同会社
佐々木邦暢 (@_ksasaki)

GPU と HPC

NVIDIA GPU 製品のおおまかな一覧



TOP500: 私が気になる二つのシステム

3	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States https://top500.org/system/180236/	2,073,600	561.20	846.84
10	**Eos NVIDIA DGX SuperPOD** - NVIDIA DGX H100, Xeon Platinum 8480C 56C 3.8GHz, NVIDIA H100, Infiniband NDR400, Nvidia NVIDIA Corporation United States https://top500.org/system/180239/	485,888	121.40	188.65

NVIDIA H100 Tensor Core GPU

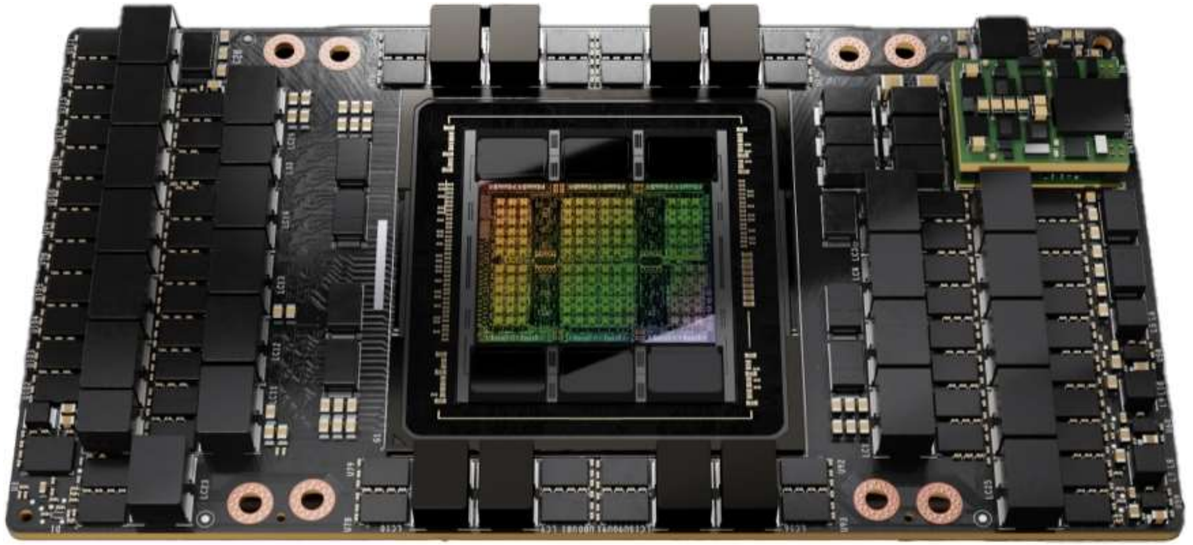
Unprecedented Performance, Scalability, and Security for Every Data Center

HIGHEST AI AND HPC PERFORMANCE
4PF FP8 (6X)| 2PF FP16 (3X)| 1PF TF32 (3X)| 60TF FP64 (3X)
3TB/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



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

DGX H100 Hardware Feature Summary

NVIDIA DGX H100

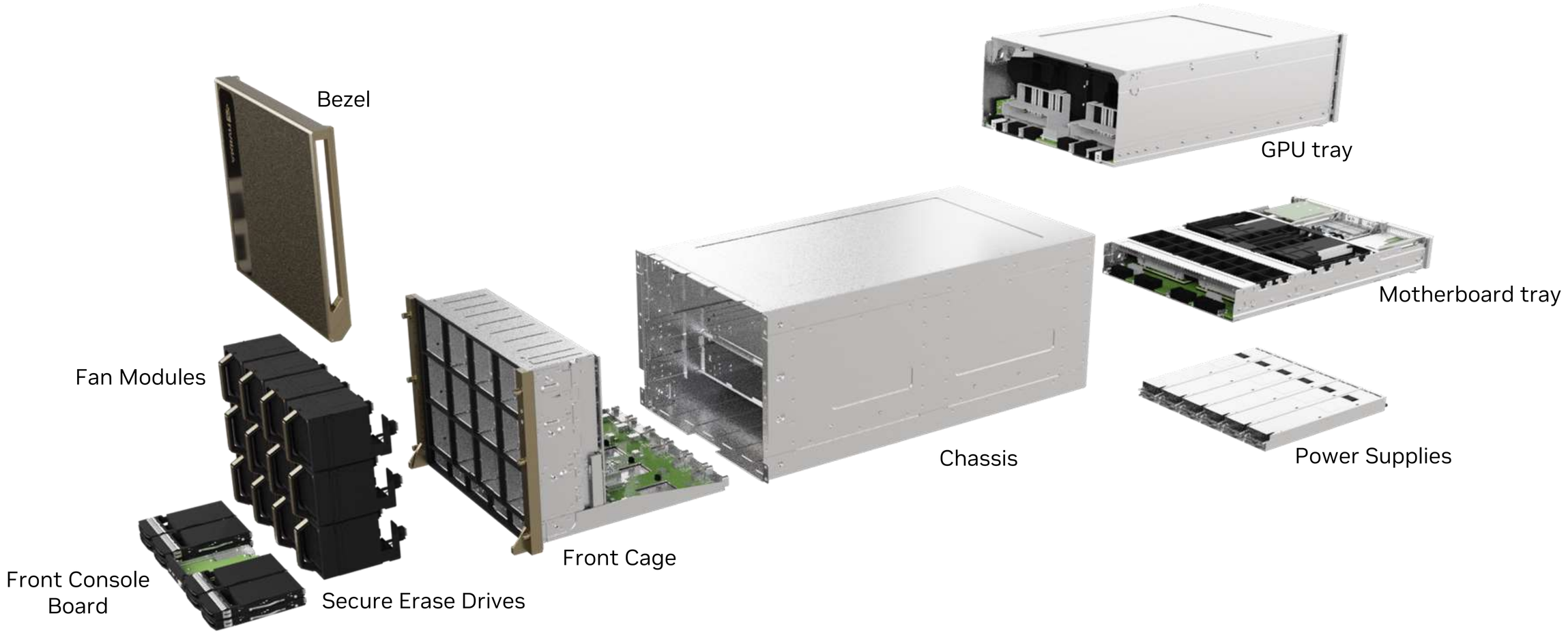
The world's first AI system with the NVIDIA H100 Tensor Core GPU

GPU	8 x NVIDIA H100 Tensor Core GPUs
GPU memory	640 GB total
Performance	32 petaFLOPs FP8
NVswitches	4
CPUs	Dual Intel Xeon 8480C, 56 cores (112 for 2 sockets), 2.0 GHz (base), 2.9 GHz (All Core Turbo), 3.8 GHz (Max Turbo)
System Memory	2 TB
Networking connectivity and speed	4 x OSFP ports provide connectivity to 8 x NVIDIA ConnectX-7s 8 x 400Gb/s InfiniBand/Ethernet* 2 x dual-port NVIDIA ConnectX-7 1 x 400Gb/s InfiniBand/Ethernet and 1 x 200Gb/s InfiniBand/Ethernet*
Cache Storage	8 x U.2 3.84TB NVMe Secure Erase Drives
Boot Storage	2 x 1.92TB M.2 NVMe (software encryptable)
Host Management	On-board 10Gb/s RJ-45 Ethernet 100Gb/s QSFP Ethernet NIC
Remote System Management	Baseboard Management Controller (BMC) 1Gb/s RJ-45 network connectivity Remote Keyboard, Video, Mouse (KVM) Remote Storage RedFish, IPMI management
Operating System	DGX OS based on Ubuntu Additional support for Ubuntu and Red Hat Enterprise Linux

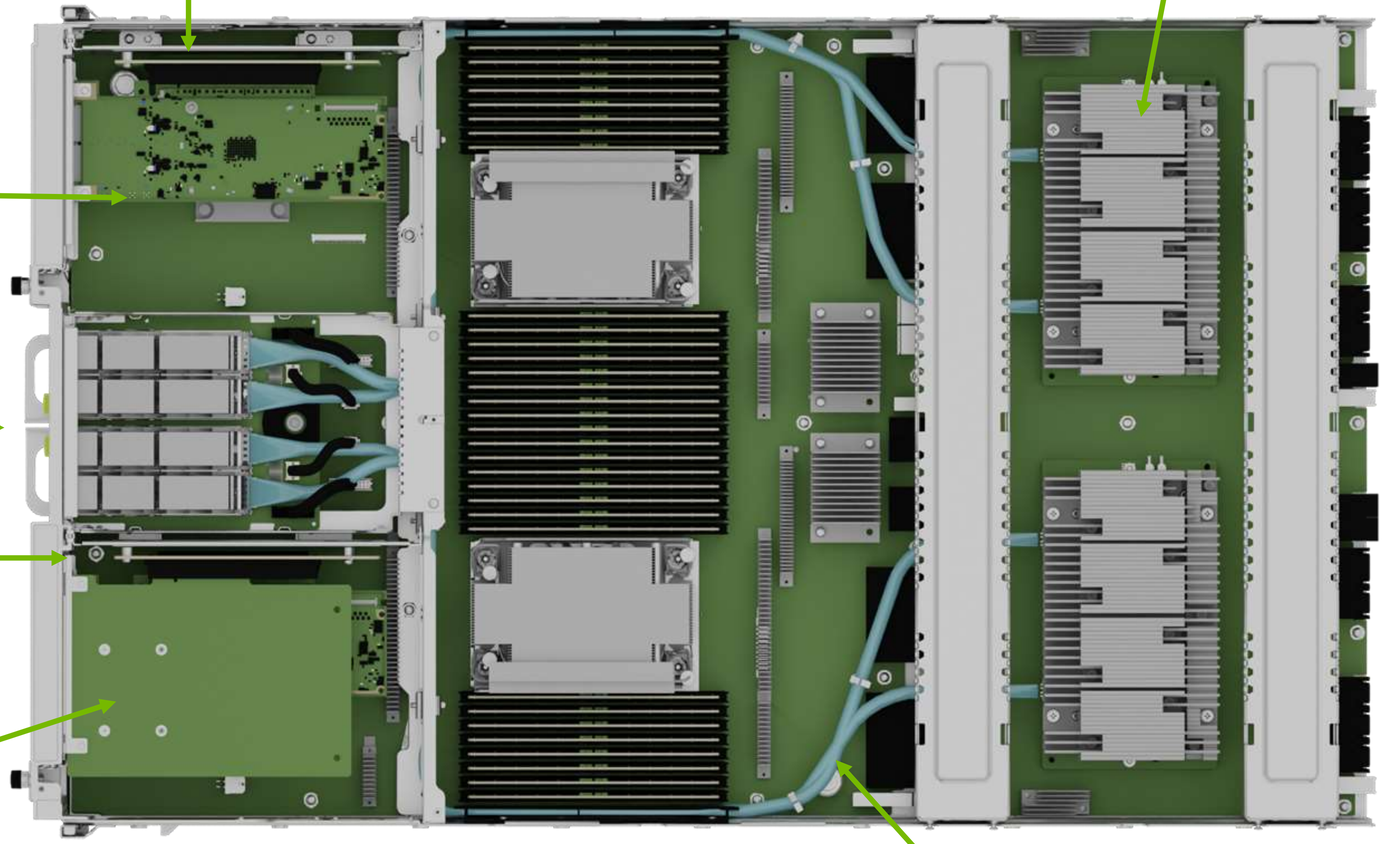
* Subject to Change

NVIDIA

DGX H100 Exploded View



DGX H100 Motherboard Tray



PCIe card riser for slots 1 and 3

32 x 64GB DIMMs for a total of 2TB of system memory

Networking modules. Each includes 4 x ConnectX-7 network devices

100Gb Ethernet NIC (top) ConnectX-7 storage networking (bottom)

OSFP transceivers for external ConnectX-7 cluster networking

PCIe card riser for slots 2 and 4

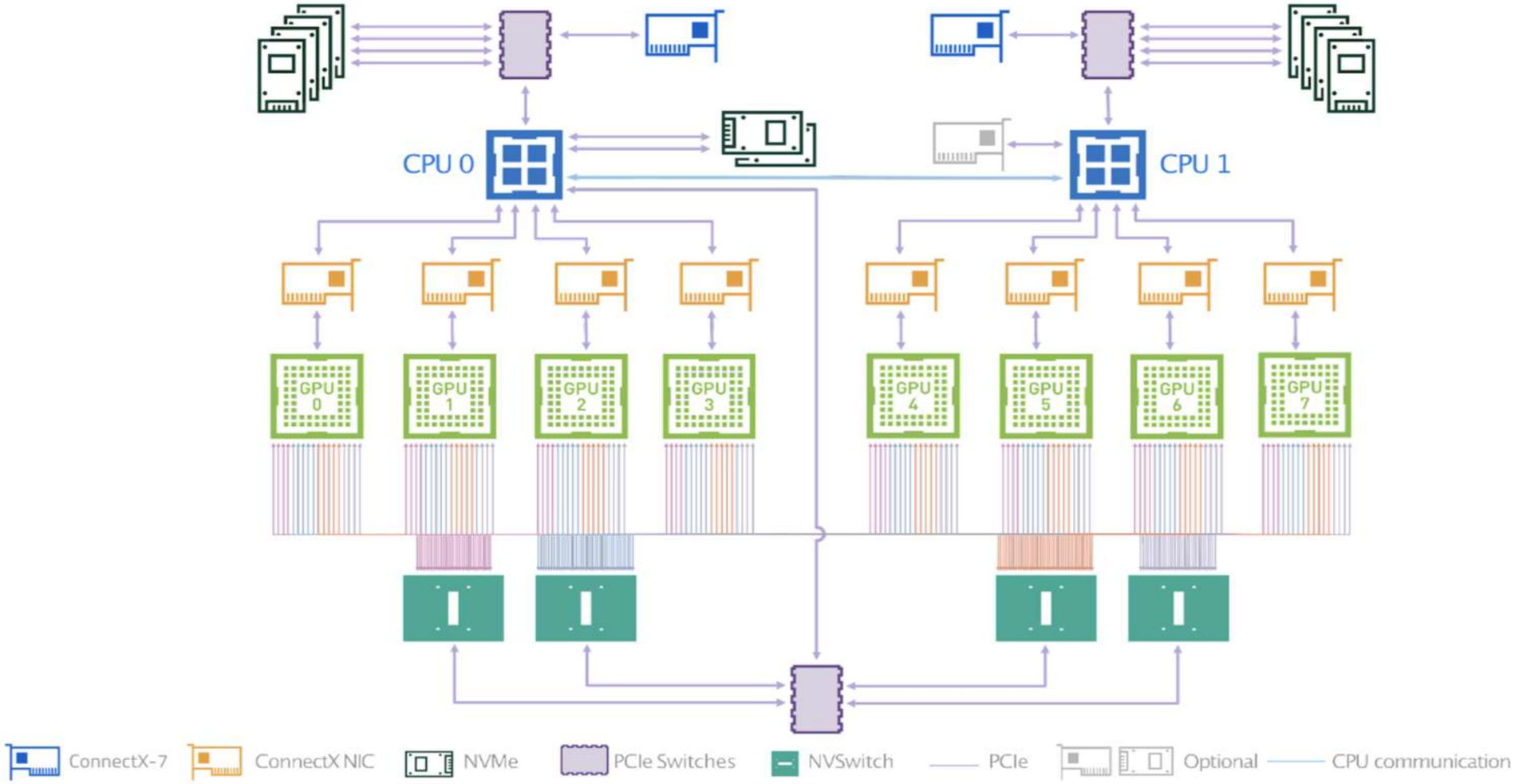
M.2 boot drive carrier (top) ConnectX-7 storage networking (bottom)

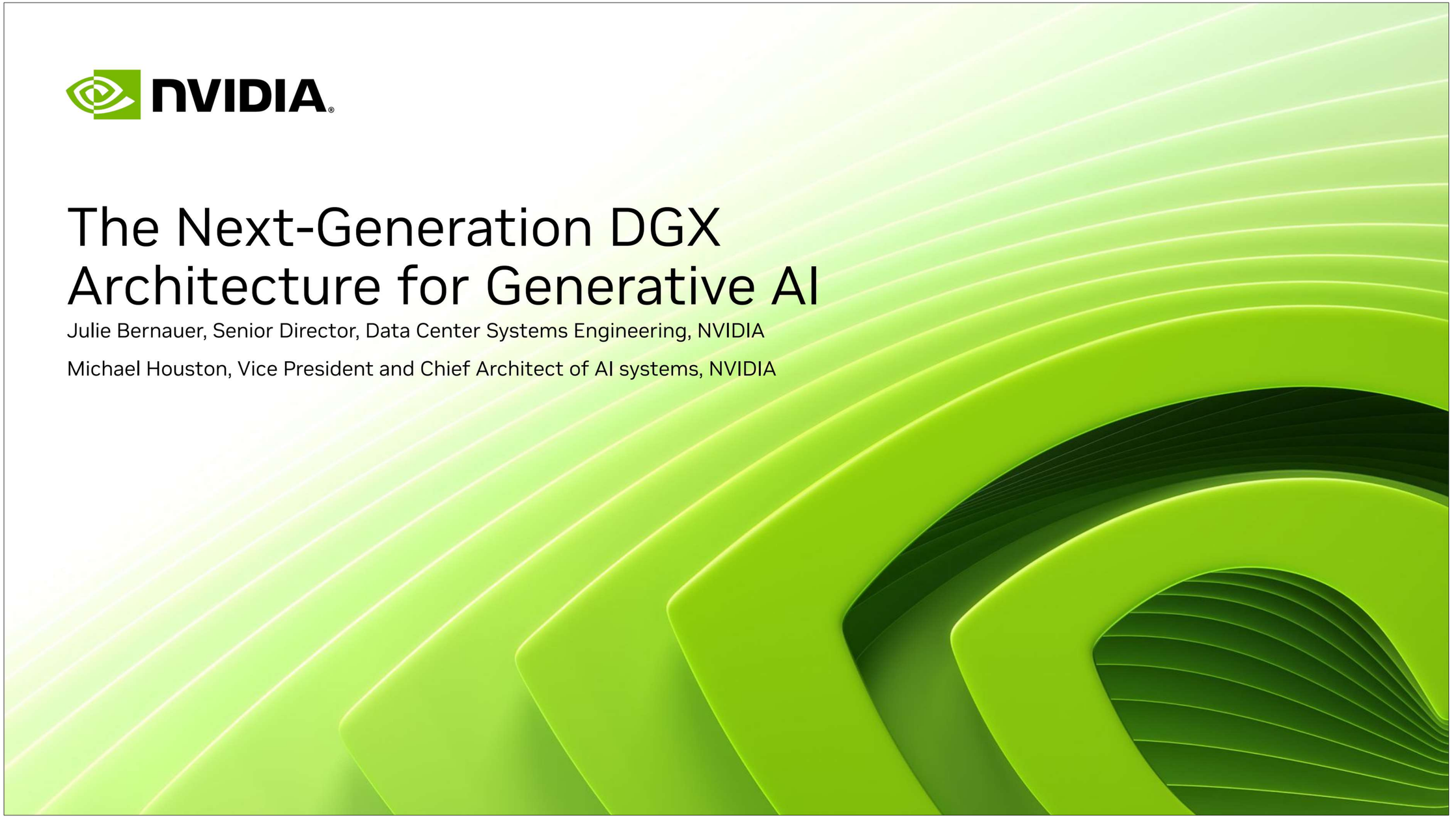
Midplane connectivity Power, PCIe, sensors and signaling communications

Dual x86 Gen5 PCIe capable CPUs

4 x Densilink cables to communicate ConnectX-7 network devices to external ports

DGX H100 Block Diagram





https://www.nvidia.com/en-us/on-demand/session/gtc24-s62421/

Applied Systems at NVIDIA

What do we do?

We bring up the next supercomputers for AI at scale

- Eos, DGXH100, 2023, #9 in Nov 2023
  - Early model for Azure Eagle, first CSP in top10 (#3)
- pre-Eos, DGXH100, 2023, #14 in May 2023
  - First H100 w/ newer Intel Gen CPUs
- Selene, DGX A100, #5 in 2020
  - First A100 w/ AMD Rome CPUs, first top5 computer to run Ubuntu
- Circe, DGX2H, #61 in 2018
  - First SuperPod deployed at customers

For our engineering and customers to get great new generation of clusters to support their product achievements

While we are at it, we get to work on new features and advances in the DL/AI world for our customers:

- MLPerf
- LLMs and other at-scale DL fun
- Other features for the at-scale computing community

top500.org/system/180153/

TOP500 The List

MEGWARE | AMD EPYC

HOME LISTS STATISTICS RESOURCES ABOUT MEDIA KIT

Home »NVIDIA Corporation »Pre-Eos 128 Node DGX SuperPOD - NVIDIA DGX H100, Xeon Platinum 8480C 56C 2GH...

PRE-EOS 128 NODE DGX SUPERPOD - NVIDIA DGX H100, XEON PLATINUM 8480C 56C 2GHZ, NVIDIA H100 TENSOR CORE GPUS, NVIDIA CONNECTX-7 NDR 400G INFINIBAND

Site:	NVIDIA Corporation
System URL:	https://www.nvidia.com/en-us/data-center/dgx-superpod/
Manufacturer:	Nvidia
Cores:	81,920
Processor:	Xeon Platinum 8480C 56C 2GHz
Interconnect:	NVIDIA Infiniband NDR
Installation Year:	2023
Performance	
Linpack Performance (Rmax)	40.66 PFlop/s
Theoretical Peak (Rpeak)	58.04 PFlop/s
Nmax	3,219,456
HPCG [TFlop/s]	340.631

List of large language models [edit]

List of large language models

Name	Release date[a]	Developer	Number of parameters[b]	Corpus size
Megatron-Turing NLG	October 2021[44]	Microsoft and Nvidia	530 billion[45]	338.6 billion tokens[45]

hpcwire.com/2023/05/22/top500-frontier-gains-92-petaflops-hpl-gets-a-little-greener/

is the second time that Microsoft has provided the highest new entry on a Top500 list; the last time was in November 2021 with a #10 ranked system.

Nvidia's new "Pre-Eos 128 Node DGX SuperPOD" is another one we've been waiting to hear word on, as Nvidia was notably quiet about Eos during their latest GTC proceedings. The apparent precursor to the full Eos system, Pre-Eos aggregates 40.66 Linpack petaflops out of a potential Rpeak of 58.05 (for a Linpack efficiency of 70.0%), using SuperPod-integrated Nvidia DGX H100 systems, equipped with Intel CPUs and Nvidia H100 GPUs, networked with Nvidia's ConnectX-7 NDR 400G InfiniBand fabric.

Eos system rendering. Courtesy of Nvidia. Click here for the full slide.

Nvidia announced the Eos system, named for the Greek goddess of the dawn, 14 months ago at its spring GTC 2022, noting at the time the system would be "online in a few months." Eos is based on the fourth-generation DGX system – the DGX H100 – which hooks together eight GPUs using Nvidia's proprietary high-speed NVLink interconnects. In its full configuration, Eos will span 18 32-DGX H100 Pods, for a total of 576 DGX H100 systems, 4,608 H100 GPUs, 500 Quantum-2 InfiniBand switches, and 360 NVLink switches. According to Nvidia, the full system will provide 18 exaflops of FP8 performance, 9 exaflops of FP16 performance or 275 petaflops of FP64 performance (all peak numbers). We did the math and came up with 156 petaflops FP64 peak when we plugged in 34 teraflops of traditional FP64 per H100 SXM (reference) and multiplied it by the machine's planned 4,608x GPUs.

Currently, between 31 and 127 DGX H100 systems can be brought together into a SuperPod (in Nvidia's parlance) using NVLink switches. We are waiting to hear back from Nvidia regarding the discrepancy between the 127-node number and the 128-node Top500 system. Nvidia's product specifications doc, dated May 2023, notes that "a pair of NVIDIA Unified Fabric Manager (UFM) appliances displaces one DGX H100 system in the DGX SuperPOD deployment pattern, resulting in a maximum of 127 DGX H100 systems per full DGX SuperPOD." Fully configured and populated, a 127-node SuperPod provides 3.2 exaflops of peak FP8 computing, 34.5 peak traditional FP64 petaflops, or 68.1 peak FP64 tensor core petaflops.

Eos is the planned successor to Selene, which was named for the Greek goddess of the moon and sister of Eos. Selene debuted in seventh place on the Top500 in June 2020 and is currently ranked ninth. The system, which comprises 280 DGX A100 systems, delivers 63.46 petaflops out of an Rpeak of 79.22 petaflops, which translates to a Linpack efficiency of 80.1%. [Note: Nvidia's Rpeak is configured differently than its marketing peak.]

NVIDIA

Eos

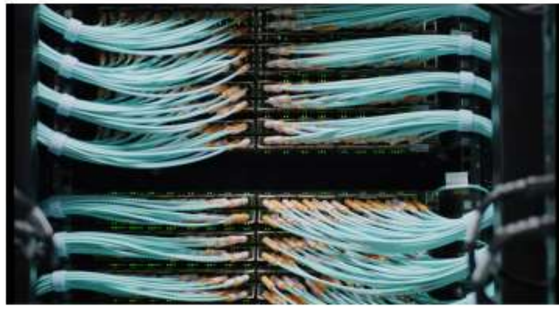
Early definition: the mandate

High level objective:

- Reference architecture for DGX H100 SuperPod that Nvidia sells and support
- Reference architecture for hyperscale designs

Consists of:

- Cluster consisting of 576 DGX H100 with Scalable Units of 32 nodes
- Networking: Quantum2 NDR IB switches, SHARPv3 optimized layout and Cumulus for ethernet management
- CPU nodes for mgmt and compute
- Thermal: Liquid Cooled Systems
- Storage: 4 TB/sec aggregated performance



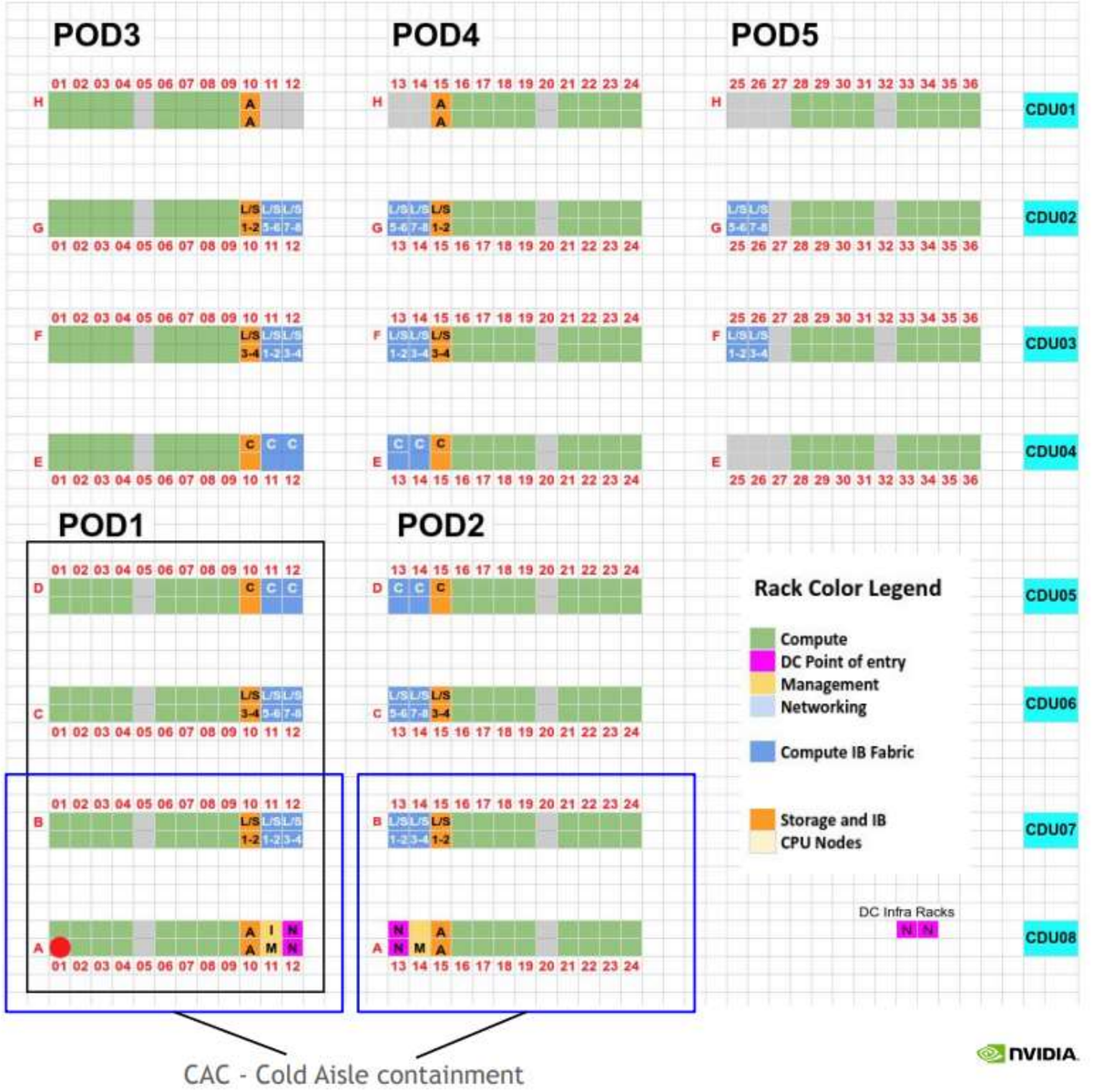
https://www.youtube.com/watch?v=J8-CgG5ewJQ

Floorplan/physical layout

There are 5 total PODs, each capable of housing 128 DGX H100 nodes.

Each POD contains two CACs with each CACs consisting of nodes from two rows (e.g.: CAC1_1 has racks A01-12 and B01-12).

While air cooling is partitioned by CACs, pairs of CDUs provide liquid cooling to pairs of row. Pairs are used to provide redundancy.

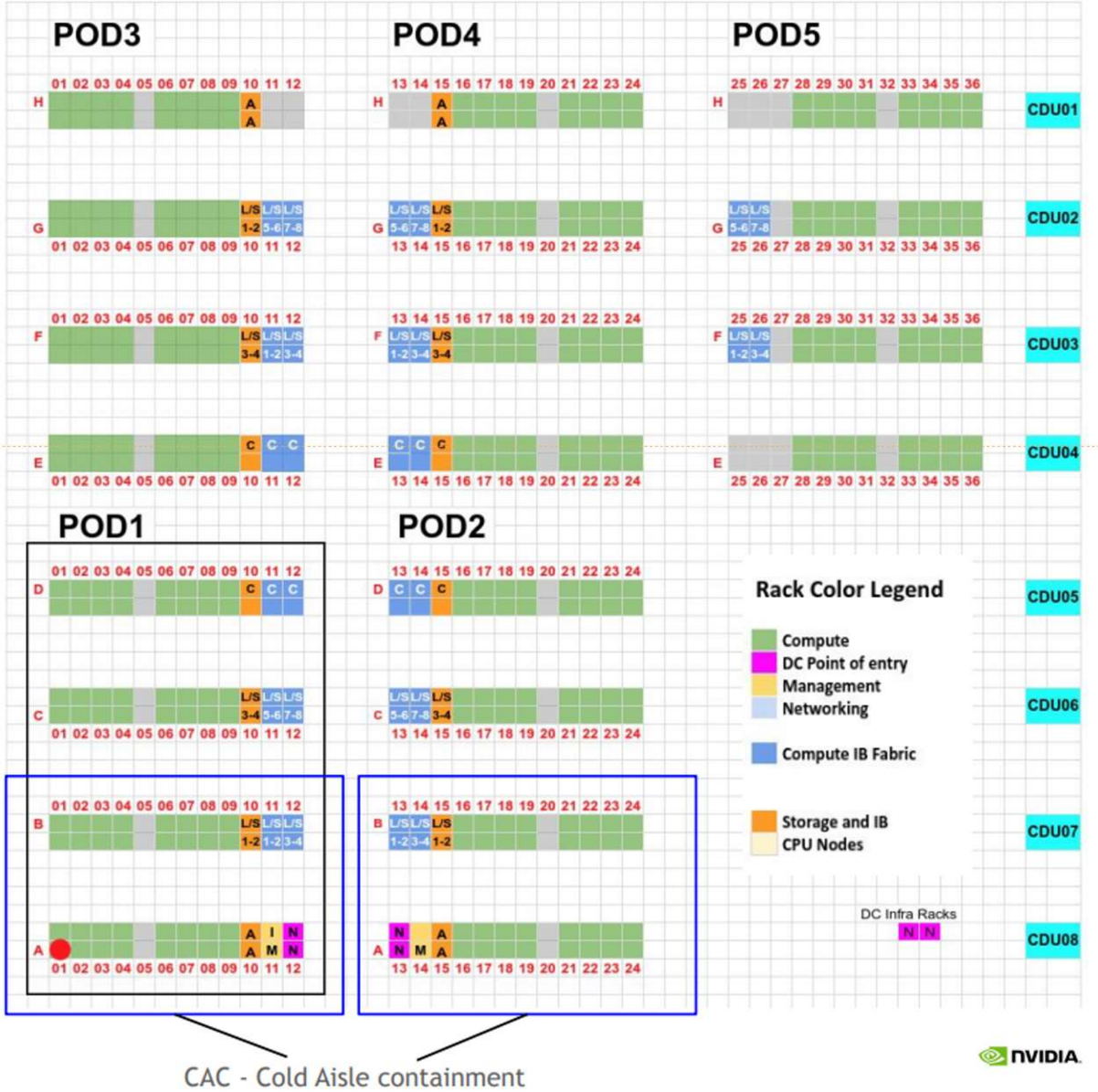


Floorplan/physical layout

There are 5 total PODs, each capable of housing 128 DGX H100 nodes.

Each POD contains two CACs with each CACs consisting of nodes from two rows (e.g.: CAC1_1 has racks A01-12 and B01-12).

While air cooling is partitioned by CACs, pairs of CDUs provide liquid cooling to pairs of row. Pairs are used to provide redundancy.

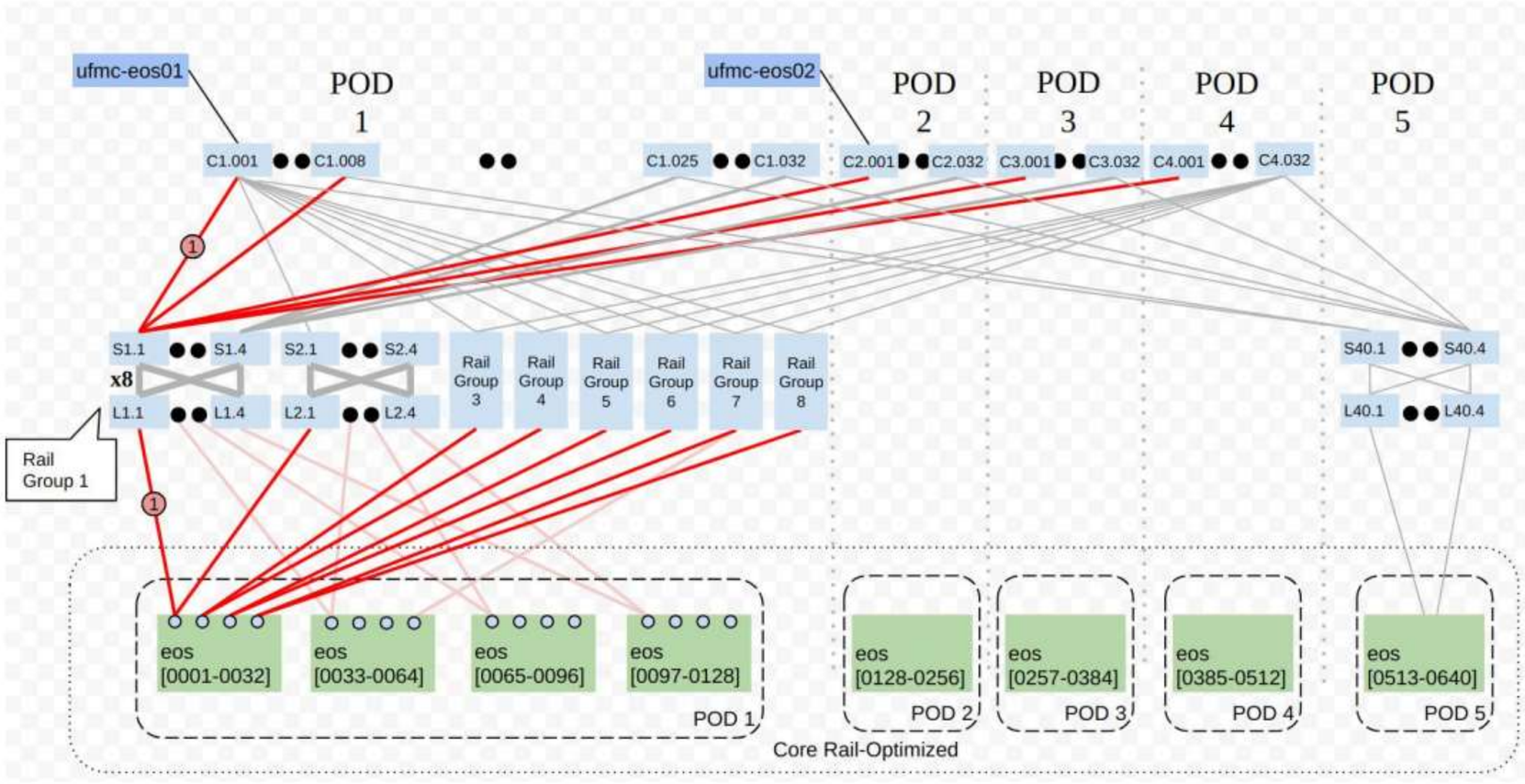


Compute infiniband architecture

Full fat tree for performance.

“Rail-optimized” to minimize latency and avoid congestion. “Rail” refers to the network link associated with a particular GPU index

- Groups of 32 nodes have each rail connected to a single switch (“Leaf Rail-Optimized”)
- Rail Groups are made of 4 leaf switches and 4 spine switches. There are 8 Rail Groups per POD, one per rail
- The core switches are installed in conjunction with the first 4 PODs only, 32 switches each. Empty ports are left on the core switches to support 3 additional PODs without recabling the existing ones



Storage architecture

Fast train with fast IO

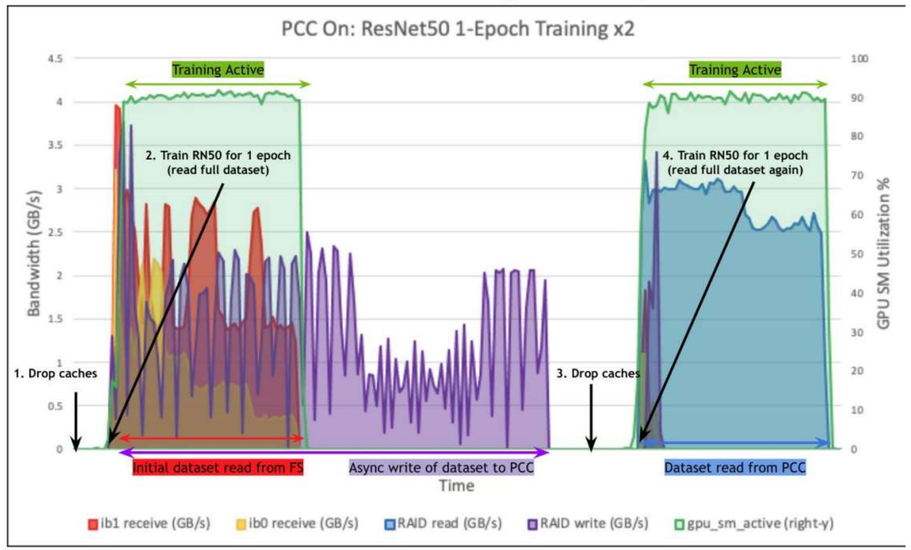
Storage as memory hierarchy:

- Memory (file) cache (aggregate): 224TB/sec - 1.1PB (2TB/node)
- NVMe cache (aggregate): 28TB/Sec - 16PB (30TB/node)
- Network filesystem (cache - Lustre): 2TB/sec - 10PB
- Object storage: 100GB/sec - 100+PB

Why we separate storage traffic:

- Maximize compute performance for RDMA based compute traffic on E/W
- Allow pre-fetching data
- Enable fast checkpoint for resiliency

1. Drop caches
2. Train RN50 for 1 epoch (read full dataset)
3. Drop caches
4. Train RN50 again for 1 epoch



Liquid cooling

2 generations of experience

Since A100, we have run with liquid cooling in our internal production systems.

Our internal systems use LC loops and CDUs (cooling distribution units).

We have been working with different options:

- In row or in rack CDUs requiring primary/secondary loops
- Using in-row heat exchangers

And have developed telemetry and systems management

NOVEMBER 2021

The system to claim the No. 1 spot for the Green500 was MN-3 from Preferred Networks in Japan. Relying on the MN-Core chip and an accelerator optimized for matrix arithmetic, this machine was able to achieve an incredible 39.38 gigaflops/watt power-efficiency. This machine provided a performance 29.7-gigaflops/watt on the last list, clearly showcasing some impressive improvement. It also enhanced its standing on the TOP500 list, moving from No. 337 to No. 302.

Green500 Release

THE LIST

GREEN500 LIST (EXCEL)

The new SSC-21 Scalable Module an HPE Apollo 6500 system installed at Samsung Electronics in South Korea achieved an impressive 33.98 gigaflops/watt. They did so by submitting an power optimized run of the HPL benchmark. It is listed at position 292 in the TOP500.

NVIDIA installed a new liquid cooled DGX A100 prototype system called Tethys. With a power optimized HPL run Tethys achieved 31.5 gigaflops/watt and garne red the No. 3 spot on the Green500. It is listed at position 296 in the TOP500.

The Wilkes-3 system improved its results but was still pushed down to the No.4 spot on the Green500. Wilkes-3, which is housed at the University of Cambridge in the U.K., had a power-efficiency of 30.8 gigaflops/watt. However, it was pushed from No. 100 to No. 281 on the TOP500 list.

The University of Florida in the USA with its HiPerGator AI system was pushed from the No. 2 spot to the No. 5 spot. This machine held steady at 29.52 gigaflops/watt. This NVIDIA system has 138,880 cores and relies on an AMD EPYC 7742 processor. Despite this impressive performance, HiPerGator AI was pushed from No. 22 to No. 31 on the TOP500.

R_{max} and R_{peak} values are in PFlop/s. For more details about other fields, check the TOP500 description.

R_{peak} values are calculated using the advertised clock rate of the CPU. For the efficiency of the systems you should take into account the Turbo CPU clock rate where it applies.

Green500 Data

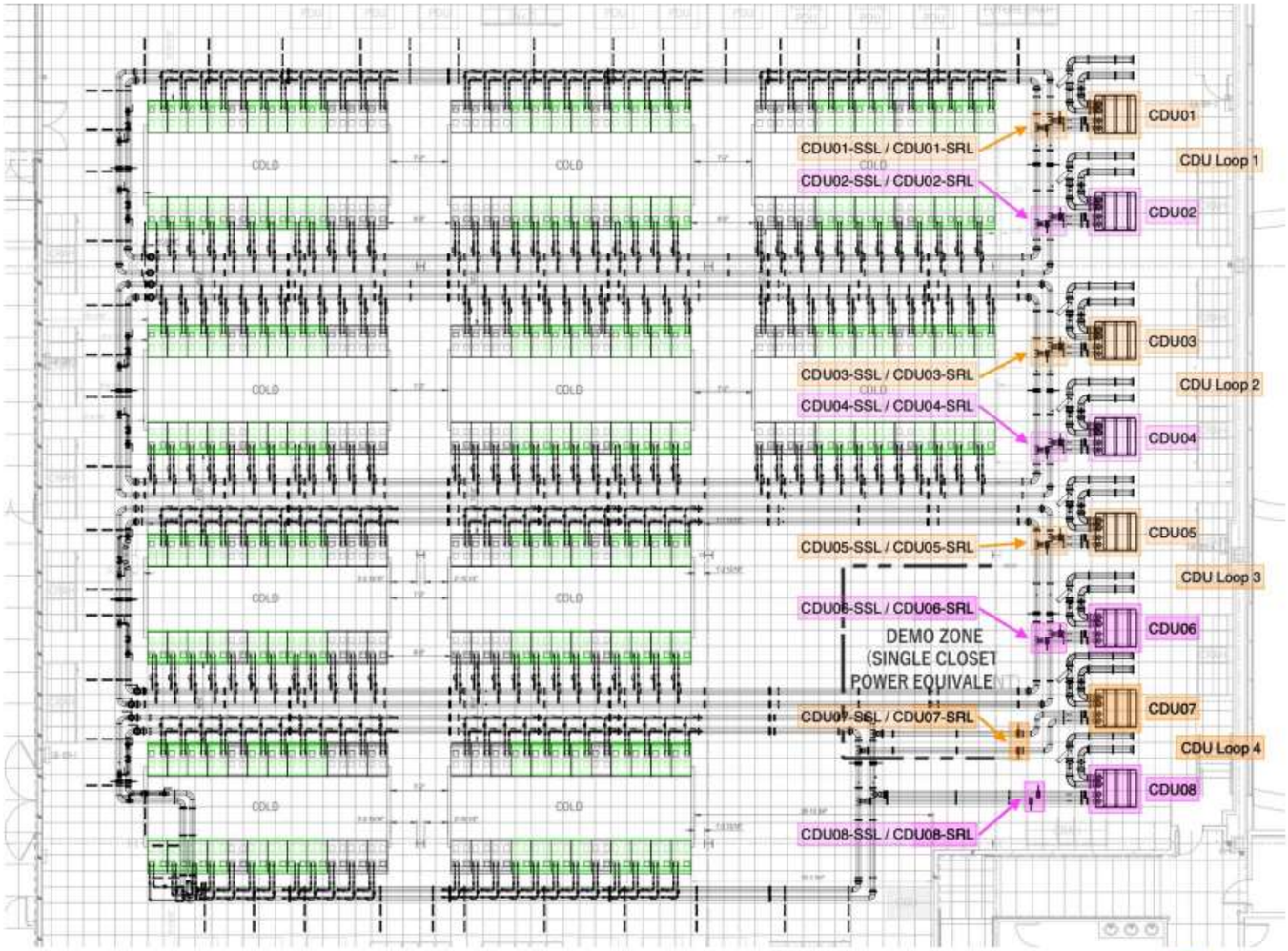
Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
3	295	**Tethys** - NVIDIA DGX A100 Liquid Cooled Prototype, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100 80GB, Infiniband HDR, **Nvidia** NVIDIA Corporation **United States**	19,840	2.25	72	31.538

Power and Thermals

Balancing phases and cooling loops

Design integrates cooling and thermals at the DC level, integrating:

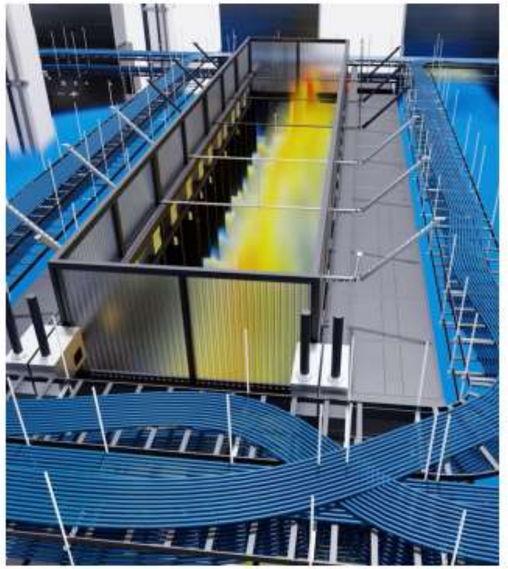
- The SuperPod layout with its split core
- Power balancing
- Liquid cooling secondary loops, CDUs and manifolds



Building at scale with GB200NVL

Building block: 576 GPU, Hot-aisle containment

- Cluster “building block”: hot-aisle containment closet (HAC)
  - Matches physical deployment
  - Group of racks with contained hot air exhaust
  - Typically also have a common liquid cooling loop
- HAC size chosen at 576 GPUs to match:
  - IB topology
  - Maximum load which can be cooled by redundant pair of CDUs
  - HAC size can be adjusted depending on specific DC capabilities
- 8 compute (OCP) racks + 8 support (EIA) racks
  - Number of support racks calculated based on network requirements: Compute IB, Storage IB, Ethernet for management
  - Management nodes, cluster storage, etc will also be spread across support racks



ISC 2024

High Performance, Energy-Efficient Systems in TOP500/GREEN500

Grace Hopper is powering the next wave of new Exaflop AI systems around the globe

Powering Worlds Fastest, Greenest Supercomputers

TOP500
The platform of choice

76%
use NVIDIA

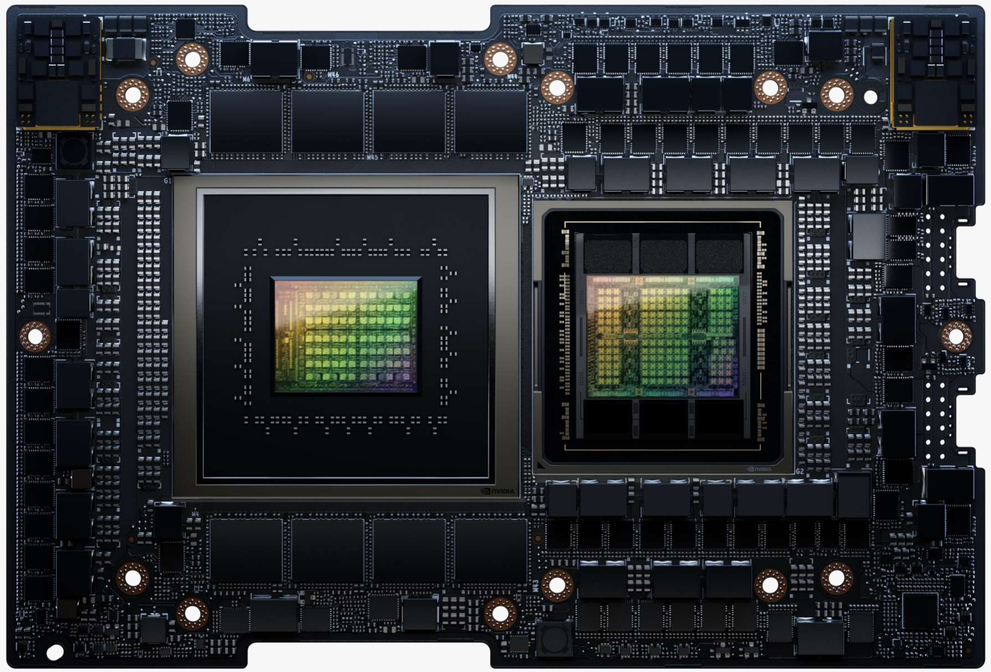
1.5 EF
added to top10

#1
powered by GH200

5 of the top10 powered by GH200

23
NVIDIA of the top 30

Green500
Energy Efficiency



NVIDIA GH200 Grace Hopper Superchip

Built for the New Era of AI Supercomputing

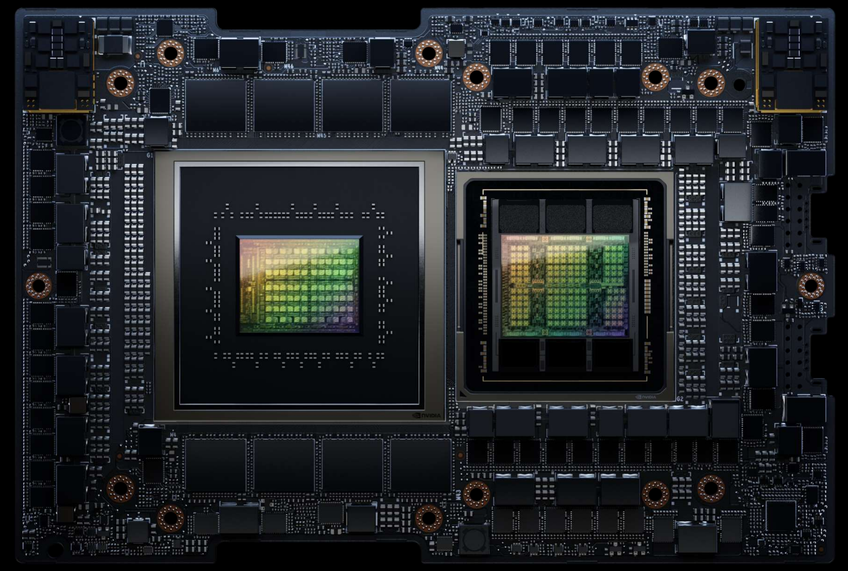
CPU to GPU Bandwidth
900GB/s
NVLink-C2C

GPU Memory Bandwidth
5TB/s
HBM3e

Energy Efficiency
50X
MILC Efficiency vs 2S x86 CPUs

QFT Quantum Simulation
90X
Performance vs 2S x86 CPUs

LLM Inference
200X
Performance vs H100 80GB



624GB High-Speed Memory | 4 PF AI Perf | 72 Arm Cores

Preliminary measured performance, subject to change
Energy Efficiency: GH200 144GB vs 2S Xeon 8480+ CPU for MILC running dataset: Apex Medium
QFT Quantum Simulation: QFT 2S Xeon 8480+ vs GH200 144GB
LLM Inference: Llama.cpp (2S 8480+) and TensorRT-LLM (GH200, H100) | Max Batch Llama2 70B | Throughput includes time to first token + token generation time

JEDI, #1 on ISC 2024 Green 500

Grace Hopper in 6 of the Top 10 Green 500

- JEDI, at EuroHPC/ FZJ, Julich, Germany
  - 48 nodes, 192 GH200 at 73 GF/Watt
- Jupiter Booster - ~6,000 nodes, ~24,000 GH200
- 36x FP64 energy efficiency vs Titan, Oakridge in 2012.
  - 18,000 K20X was 2 GF/Watt.
- GH200 over 1,000x more energy efficient on raw compute
  - Mixed precision and the AI performance,
- AI can bring time-to-solution algorithmic improvements to solving scientific compute problems faster.

Isambard-AI , #2 on ISC 2024 Green 500

University of Bristol

- Isambard-AI , University of Bristol
  - 0.7 AI EF | 42 nodes | 168 GH200 Superchips at 69 GF/Watt
  - 22 TB LPDDR | 16 TB HBM | 38TB Total Memory
- Phase 2: 12 cabinets | 660 blades
  - 21 AI EF | 1,320 nodes | 5,280 GH200 Superchips
  - 675 TB LPDDR | 506 TB HBM | 1.2 PB Total Memory
- AI-based Supercomputer to Foster Scientific Innovation

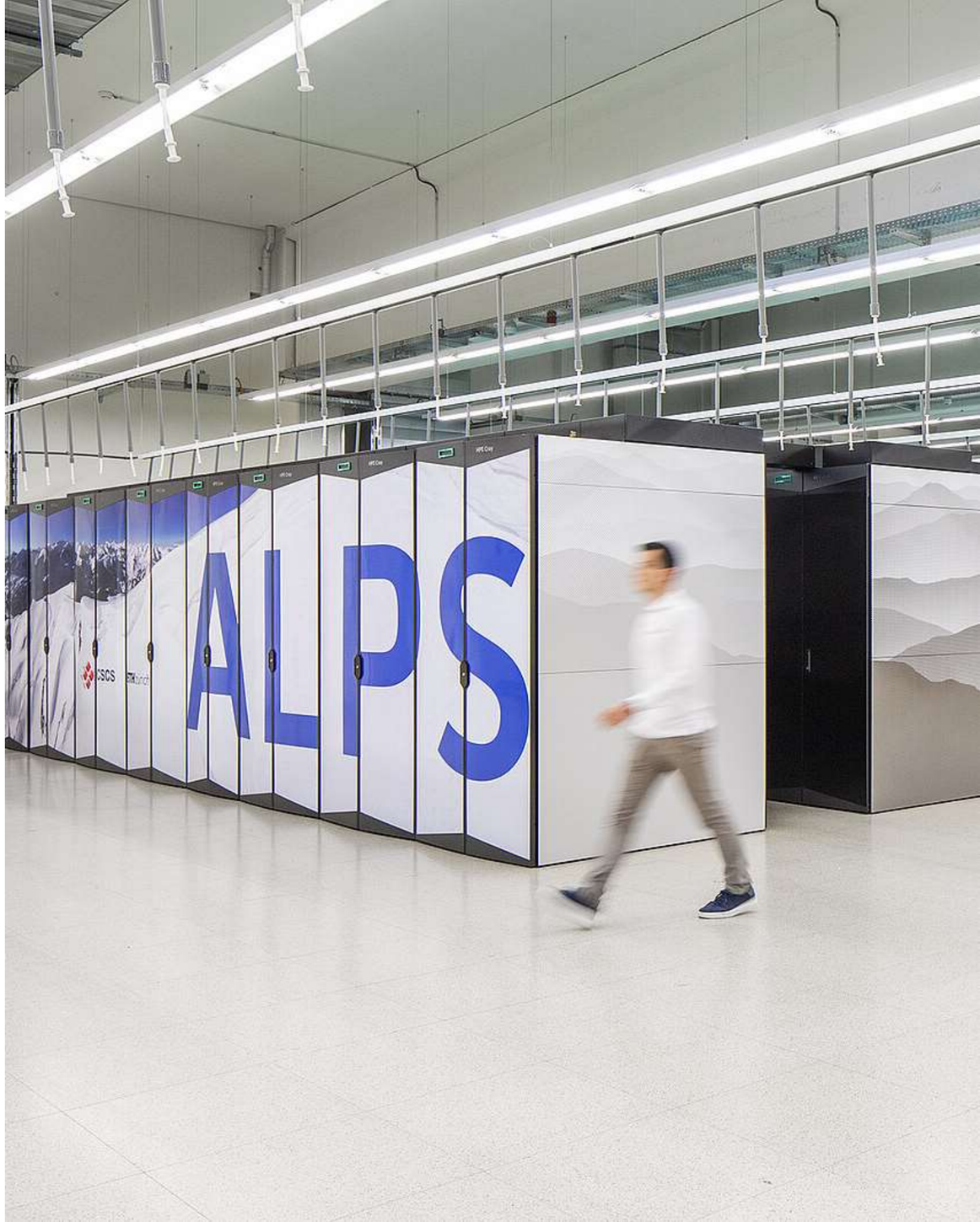
First European Grace Hopper Supercomputer Online

- Fastest AI Supercomputer in Europe
- 20 Exaflops of AI
- 10X more energy efficient than Piz Daint at 64 GF/Watt
- Powered by 10,000 Grace Hopper Superchips
- HPC and AI to Advance Weather, Climate (1km global models), and Material Science

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Hewlett Packard Enterprise

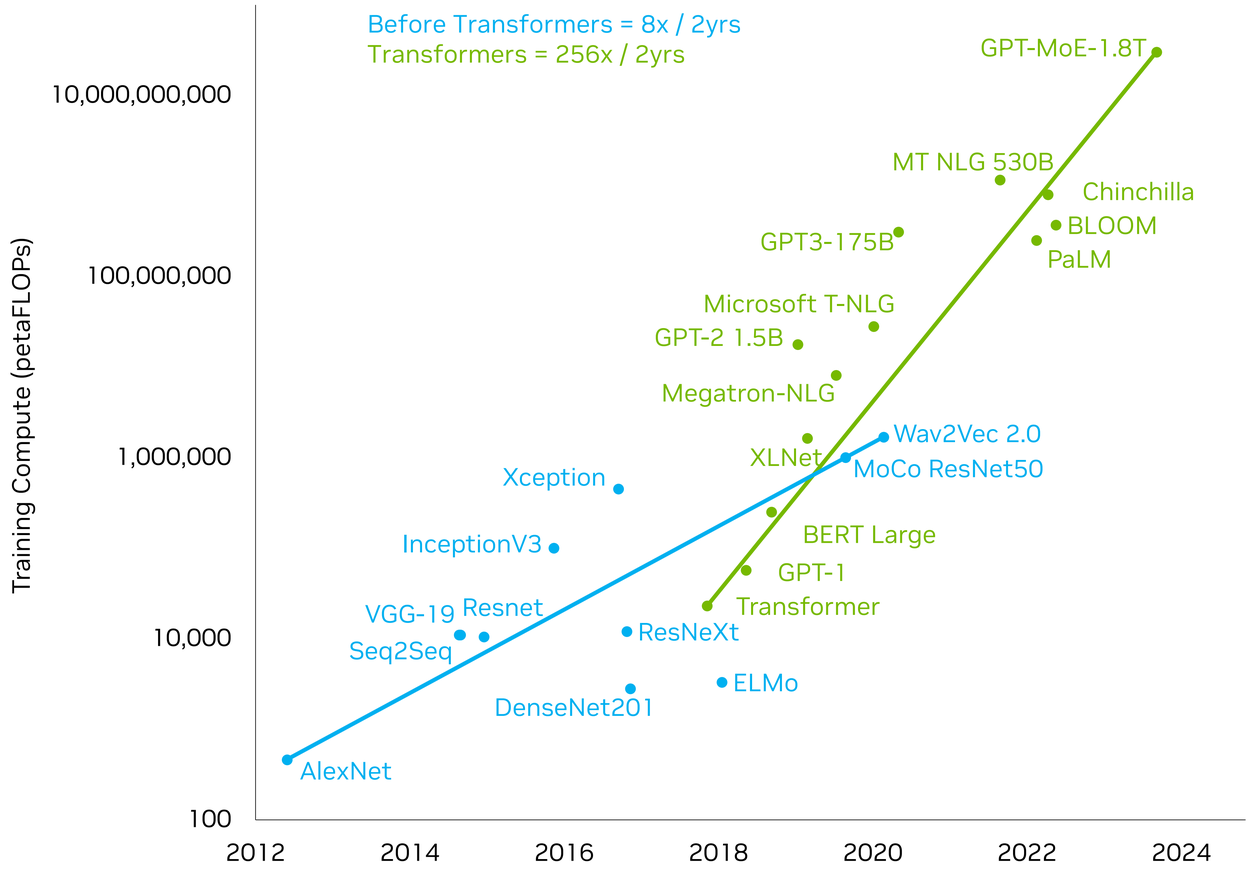
NVIDIA



Green500 (2024/6)

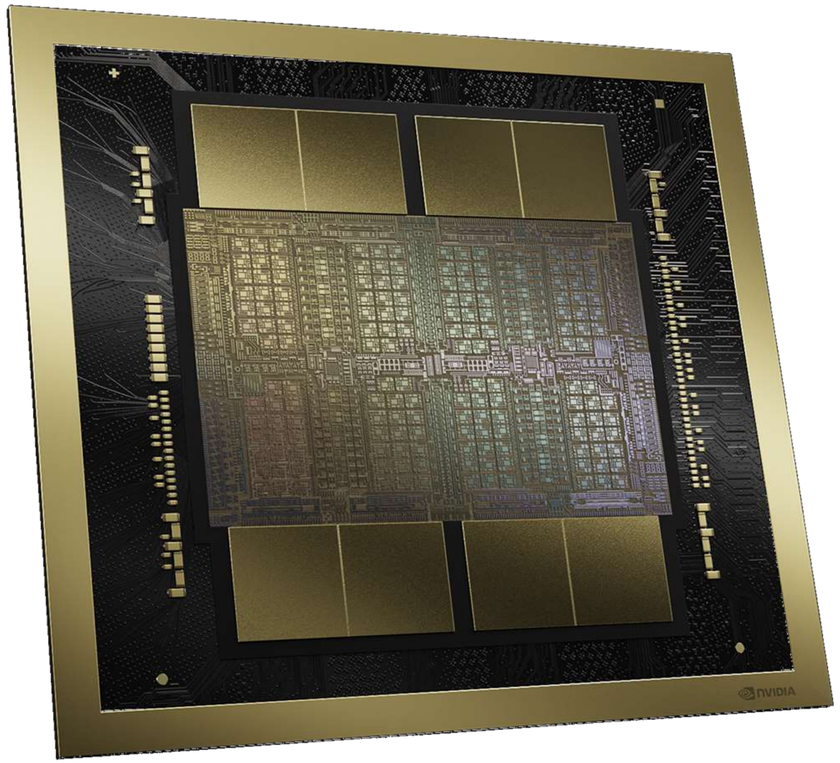
Rank	Name	Computer	Energy Efficiency [GFlops/Watts]	Processor
1	JEDI	BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200	72.73	Grace Hopper Superchip 72C 3GHz
2	Isambard-AI phase 1	HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11	68.83	NVIDIA Grace 72C 3.1GHz
3	Helios GPU	HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11	66.95	NVIDIA Grace 72C 3.1GHz
4	Henri	ThinkSystem SR670 V2, Intel Xeon Platinum 8362 32C 2.8GHz, NVIDIA H100 80GB PCIe, Infiniband HDR	65.40	Intel Xeon Platinum 8362 32C 2.8GHz
5	preAlps	HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11	64.38	NVIDIA Grace 72C 3.1GHz
6	HoreKa-Teal	ThinkSystem SD665-N V3, AMD EPYC 9354 32C 3.25GHz, Nvidia H100 94Gb SXM5, Infiniband NDR200	62.96	AMD EPYC 9354 32C 3.25GHz
7	Frontier TDS	HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11	62.68	AMD Optimized 3rd Generation EPYC 64C 2GHz
8	Venado	HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11	59.29	NVIDIA Grace 72C 3.1GHz
9	Adastra	HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11	58.02	AMD Optimized 3rd Generation EPYC 64C 2GHz
10	Setonix – GPU	HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11	56.98	AMD Optimized 3rd Generation EPYC 64C 2GHz

爆発的に伸びゆく演算需要



NVIDIA Blackwell

新たな産業革命の原動力



兆単位のパラメーターを持つ AI トレーニングのために開発

GPU 1基で 20 PetaFLOPS (FP4) の AI 性能

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

AI データセンターを 10 万 GPU 規模へ拡大

AI SUPERCHIP
208B Transistors

2nd GEN TRANSFORMER ENGINE
FP4/FP6 Tensor Core

5th GENERATION NVLINK
Scales to 576 GPUs

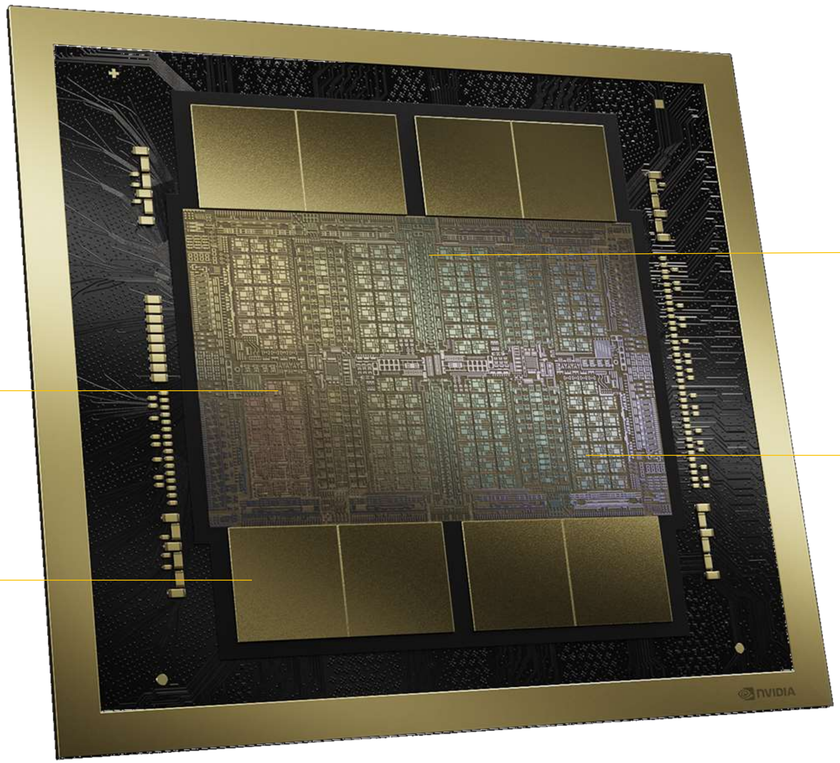
RAS ENGINE
100% In-System Self-Test

SECURE AI
Full Performance Encryption & TEE

DECOMPRESSION ENGINE
800 GB/s

AI スーパーチップの新機軸

世界最大クラスの 2 つのダイを 1 基の GPU に統合



レチクル リミットサイズのダイ 2 基を1つの GPU に結合

NV-HBI 10TB/s 広帯域インタフェース

レチクル リミット
サイズのダイ #1

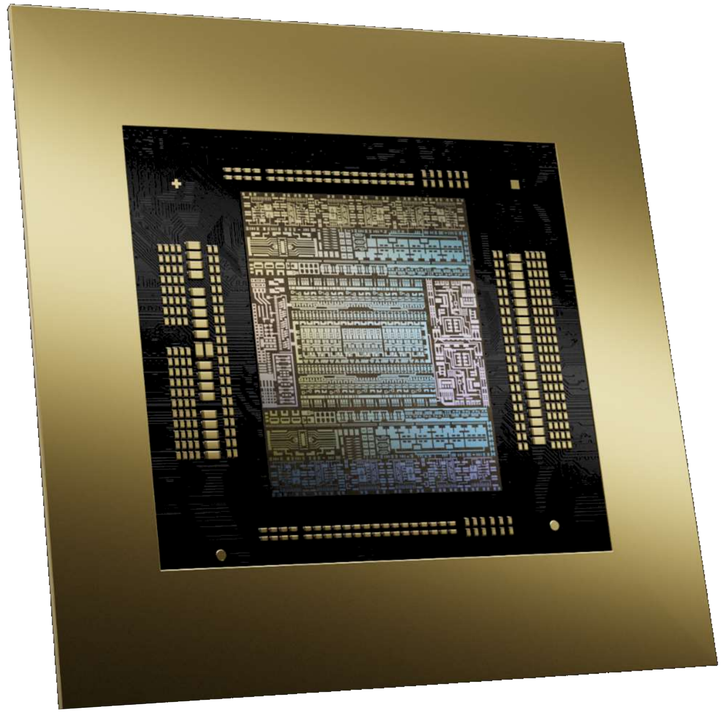
レチクル リミット
サイズのダイ #2

高速メモリ
192GB HBM3e

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

第 5 世代 NVLink と NVLink スイッチ

兆単位パラメーターモデルに対応するスケーラビリティ



7.2 TB/s のバイセクションバンド幅

Sharp v4 plus FP8

3.6 TF In-Network Compute

NVLink で最大 576 GPU を接続可能

NDR InfiniBand や 400Gbps イーサネットより 18 倍高速

GB200 NVL72

計算機の新たな形

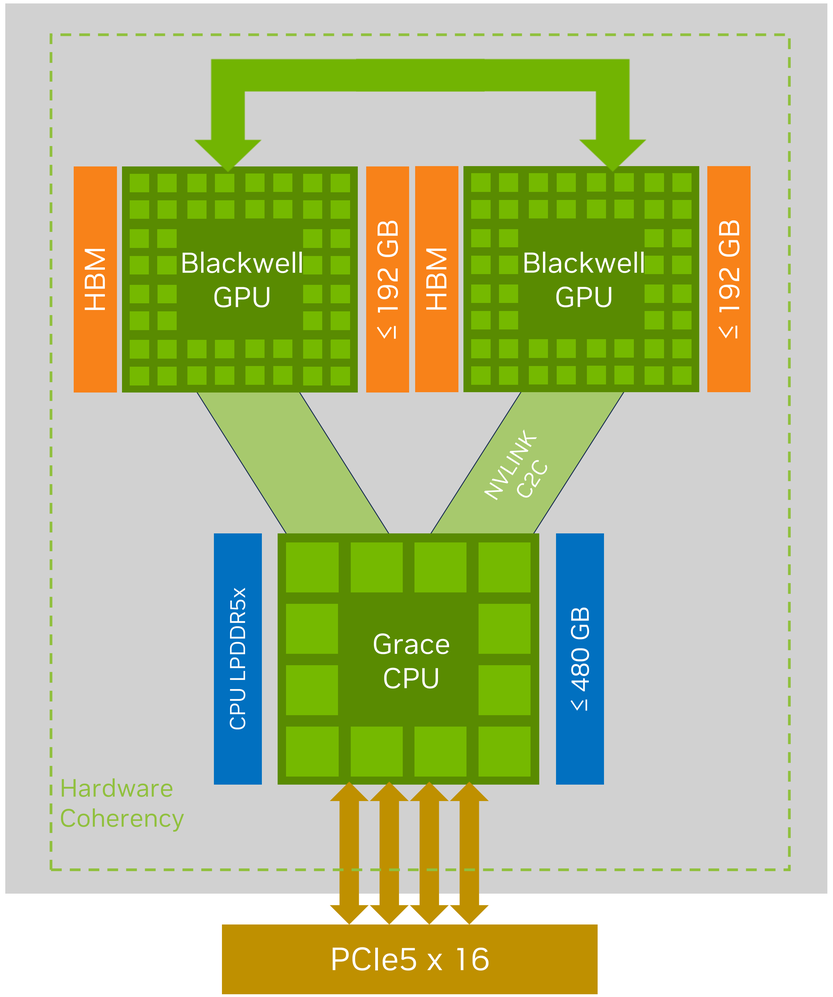
GB200 NVL72

- 36 GRACE CPUs
- 72 BLACKWELL GPUs
- Fully Connected NVLink Switch Rack

Training	720 PFLOPs
Inference	1,440 PFLOPs
NVL Model Size	27T params
Multi-Node All-to-All	130 TB/s
Multi-Node All-Reduce	260 TB/s

Grace Blackwell Superchip

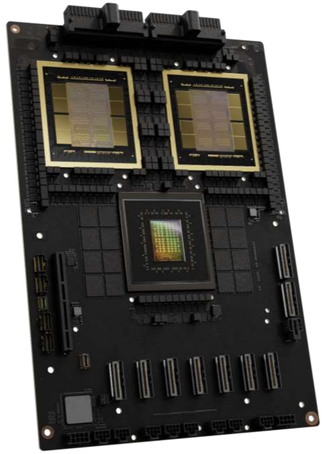
Architecture for Simulation + AI



	GH200 (1 Grace | 1 Hopper)	GB200 (1 Grace | 2 Blackwell)
Assumed Network	NDR200	XDR400
Super Chip TDP	700W	Up to 2.7kW
CPU	Grace	Grace
HBM / LPDDR (GB)	96 GB | Up to 480GB	Up to 384 GB | Up to 480 GB
HBM Bandwidth	1x	4x
FP64 FMA (Vector)	1x	2.6x
FP32 FMA (Vector)	1x	2.6x

GB200 NVL72 計算ノードとインターコネクト

GB200 NVL72 Rack の構成要素



GB200 SUPERCHIP

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

GB200 SUPERCHIP COMPUTE TRAY

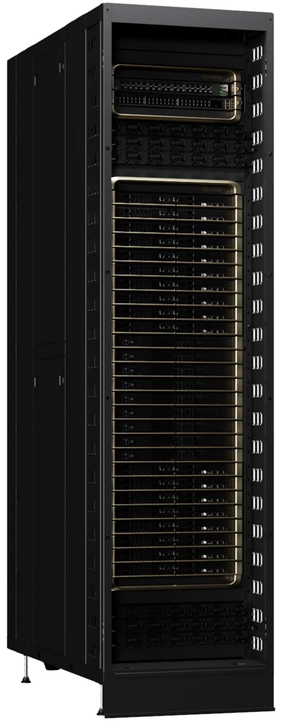
2x GB200
80 PETAFLOPS FP4 AI INFERENCE
40 PETAFLOPS FP8 AI TRAINING
1728 GB FAST MEMORY
1U Liquid Cooled
18 Per Rack

NVLINK SWITCH TRAY

2x NVLINK SWITCH CHIP
14.4 TB/s Total Bandwidth
SHARPv4 FP64/32/16/8
1U Liquid Cooled
9 Per Rack

Blackwell ラインナップ

すべてのデータセンターへ新世代の演算性能を



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

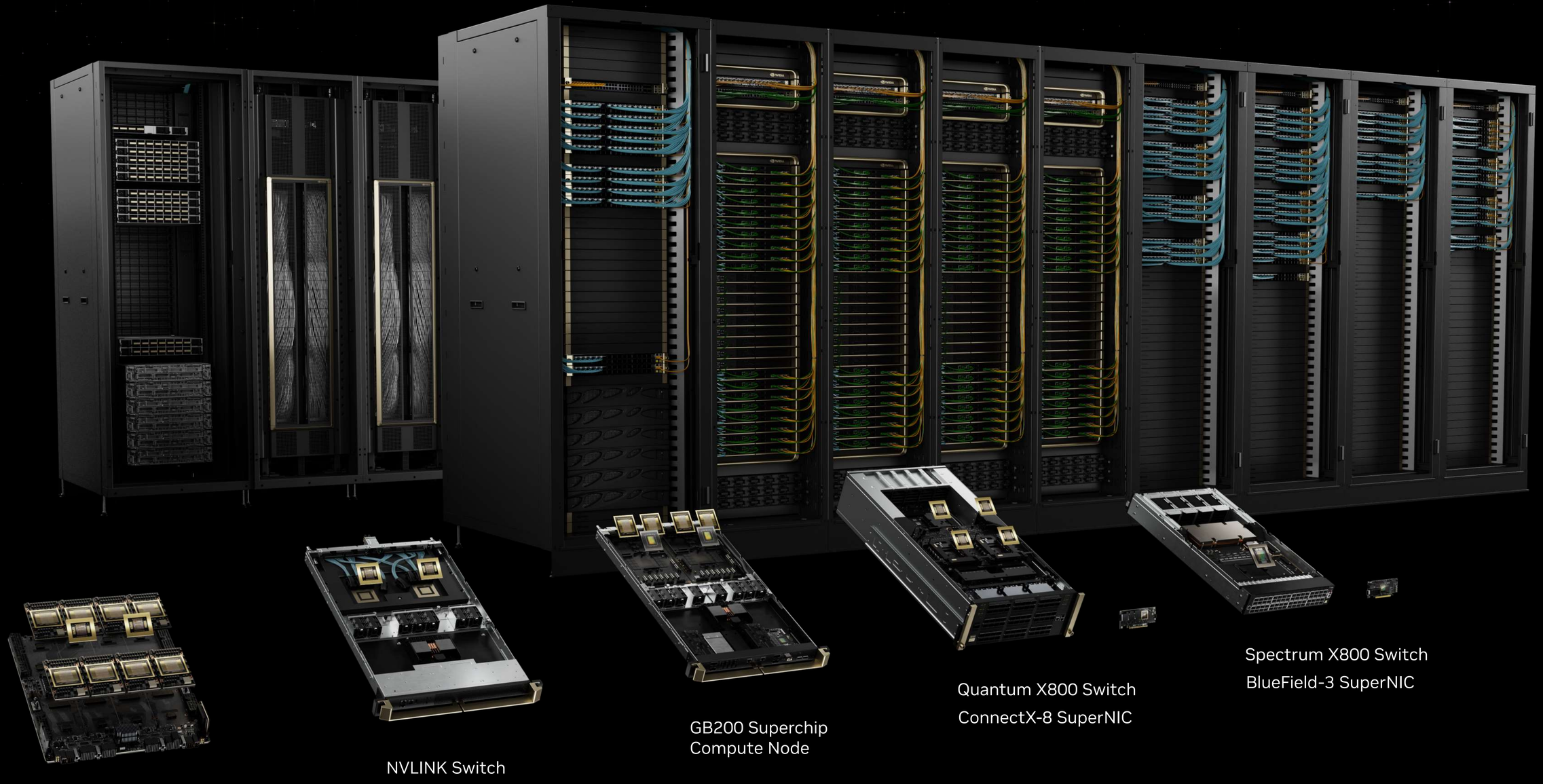
Blackwell システム仕様

	GB200 NVL72	HGX B200	HGX B100	参考までに HGX H100
GPUs	72	8	8	8 (Hopper)
FP4 Tensor Core	1,440 petaFLOPS	144 petaFLOPS	112 petaFLOPS	-
FP8/FP6/INT8	720 petaFLOPS	72 petaFLOPS	56 petaFLOPS	32 petaFLOPS
Fast Memory	Up to 30 TB	up to 1.5 TB	Up to 1.5TB	640GB
Aggregate Memory Bandwidth	Up to 600 TB/s	Up to 64 TB/s	Up to 64 TB/s	24 TB/s
Aggregate NVLink Bandwidth	130 TB/s	14.4 TB/s	14.4 TB/s	7.2 TB/s
CPU Cores	2592 Arm Neoverse V2 cores	-	-	-
Per GPU Specifications				
FP4 Tensor Core	20 petaFLOPS	18 petaFLOPS	14 petaFLOPS	-
FP8/FP6 Tensor Core	10 petaFLOPS	9 petaFLOPS	7 petaFLOPS	4 petaFLOPS
INT8 Tensor Core	10 petaOPS	9 petaOPS	7 petaOPs	4 petaOPS
FP16/BF16 Tensor Core	5 petaFLOPS	4.5 petaFLOPS	3.5 petaFLOPS	2 petaFLOPS
TF32 Tensor Core	2.5 petaFLOPS	2.2 petaFLOPS	1.8 petaFLOPS	0.98 petaFLOPS
FP64 Tensor Core	45 teraFLOPS	40 teraFLOPS	30 teraFLOPS	67 teraFLOPS
GPU memory | Bandwidth	Up to 192 GB HBM3e | Up to 8 TB/s			80GB HBM3 | 3TB/s
Multi-Instance GPU (MIG)	7			7
Decompression Engine	Yes			-
Decoders	2x 7 NVDEC 2x 7 NVJPEG			
Power	Configurable up to 1,200W	Configurable up to 1,000W	Configurable up to 700W	700W
Interconnect	5th Generation NVLink: 1.8TB/s PCIe Gen6: 256GB/s			4th Generation NVLink: 0.9TB/s PCIe Gen5: 128GB/s
Server options	NVIDIA GB200 NVL72 partner and NVIDIA-Certified Systems with 72 GPUs	NVIDIA HGX B200 partner and NVIDIA-Certified Systems with 8 GPUs	NVIDIA HGX B100 partner and NVIDIA-Certified Systems with 8 GPUs	

1.Preliminary specifications subject to change. All Tensor Core numbers with sparsity.
2.GB200 Superchip configuration includes 2 high performance B200 GPUs and one Grace CPU

NVIDIA

NVIDIA Blackwell Platform



HGX B100

NVLINK Switch

GB200 Superchip
Compute Node

Quantum X800 Switch
ConnectX-8 SuperNIC

Spectrum X800 Switch
BlueField-3 SuperNIC