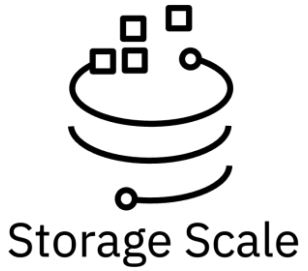
IBM Storage Scale für Daten und AI & NVIDIA GPUs

—

Frank Kraemer, IBM
kraemerf@de.ibm.com

Disclaimer



IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

IBM reserves the right to change product specifications and offerings at any time without notice. This publication could include technical inaccuracies or typographical errors. References herein to IBM products and services do not imply that IBM intends to make them available in all countries.

The AI “Data Vulcan”

Data is crucial for **training** AI models. The quality and quantity of training data sets are crucial to the accuracy and effectiveness of machine learning models. The more diverse and representative the data is, the better the model can generalize and perform on new, unseen data.

Source: Mercedes-Benz

Source: Hako

Source: TORC

Source: Digital Railways

Source: John Deere

Source: ZOOX

Autonomous Systems (AS) are everywhere impacting how we live, work, play, purchase goods and services, how we socialize, and even our commutes.

Q: How long will it take to train an LLM model like GPT-3?

A: See Paper https://arxiv.org/pdf/2005.14165.pdf

How much memory is required? GPT-3 requires **175 billion** parameters.

- Two bytes per parameter **(FP16)** means the total memory required is **350 GB**.
- Four bytes per parameter **(FP32)** means we need **700 GB** of memory.

Training LLMs is computationally intensive

GPT-3 Training Time on NVIDIA A100 GPUs

	Time to train 300B tokens in days (A100) – BF16			
	800 GPUs (5x DGX SuperPod)	480 GPUs (3x DGX SuperPod)	160 GPUs (1x DGX SuperPod)	64 GPUs (8x DGX A100)
GPT-3: 126M	0.07	0.12	0.37	0.92
GPT-3: 5B	0.8	1.3	3.9	9.8
GPT-3: 20B	3.6	6	18.1	45.3
GPT-3: 40B	6.6	10.9	32.8	82
GPT-3: 175B	28	46.7	140	349.9

x days (24h)

The Register®

Zuckerberg wants to build artificial general intelligence with 350K Nvidia H100 GPUs

Maybe the AGI can finish that Metaverse, haha – oh wait, they're serious

Katyanna Quach

Sat 20 Jan 2024 // 01:58 UTC

Facebook supremo Mark Zuckerberg is redirecting Meta-wide efforts to build artificial general intelligence and wants to secure a whopping 350,000 or more Nvidia H100 GPUs by the end of the year to make that happen.

Source: https://www.theregister.com/2024/01/20/metas_ai_plans/

Storage Scale & Storage Scale Systems

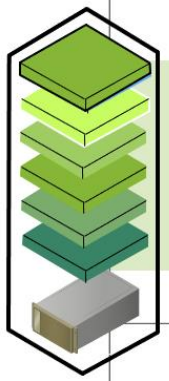
IBM

NVIDIA DGX BasePOD validated storage partner

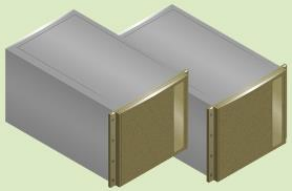
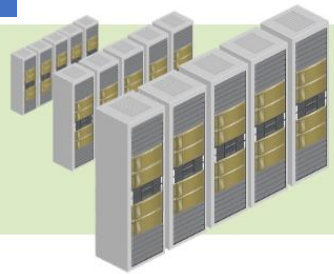
- NVIDIA Base Command supported, includes Kubernetes (K8) platform (static/dynamic storage provision).
- Simple building blocks – scalable seamless storage upgrade path as needs grow from 1st DGX to AI CoE DGX SuperPOD
- NVIDIA MagnumIO (GDS) acceleration
- Faster, simplified deployments and upgrades with Ansible automation.

NVIDIA DGX SuperPOD validated storage partner

- Leading parallel performance on enterprise storage (6-9s availability)
- Global Data Platform: automate AI workflows: ingest to inference, high-performance object, multi-site, multi-vendor & hybrid cloud
- Data Economics – Eliminate copies and transparently tier
- Trusted, global enterprise level support and services.
- Successful deployments across the globe



DGX Platform powers every step in your AI journey

	1 Day One your 1st DGX systems	2 Scaled Infrastructure DGX BasePOD	3 AI CoE DGX SuperPOD	4 AI Hybrid Cloud DGX Cloud
			 EQUINIX	 EQUINIX
Software capabilities that help you grow your AI platform	Base Command		Base Command Platform	

	• NVIDIA AI Enterprise:
	  • Pre-trained models, optimized frameworks
	  • Accelerate data prep tasks
	  • Customize/fine-tune pre-trained models
	  • Optimize/accelerate inference
	• Kubernetes or Slurm scheduling
	• Add/manage DGX within your existing compute infrastructure
	• Run Jupyter notebooks
	• Cloud-based private registry
	• Accelerate storage & network IO
	• Fully optimized OS stack
			In addition to the features on the left:
			• Base Command Platform:
			  • Simplify developer workflow
			  • Dataset management
			  • Batch processing
			  • Monitoring	

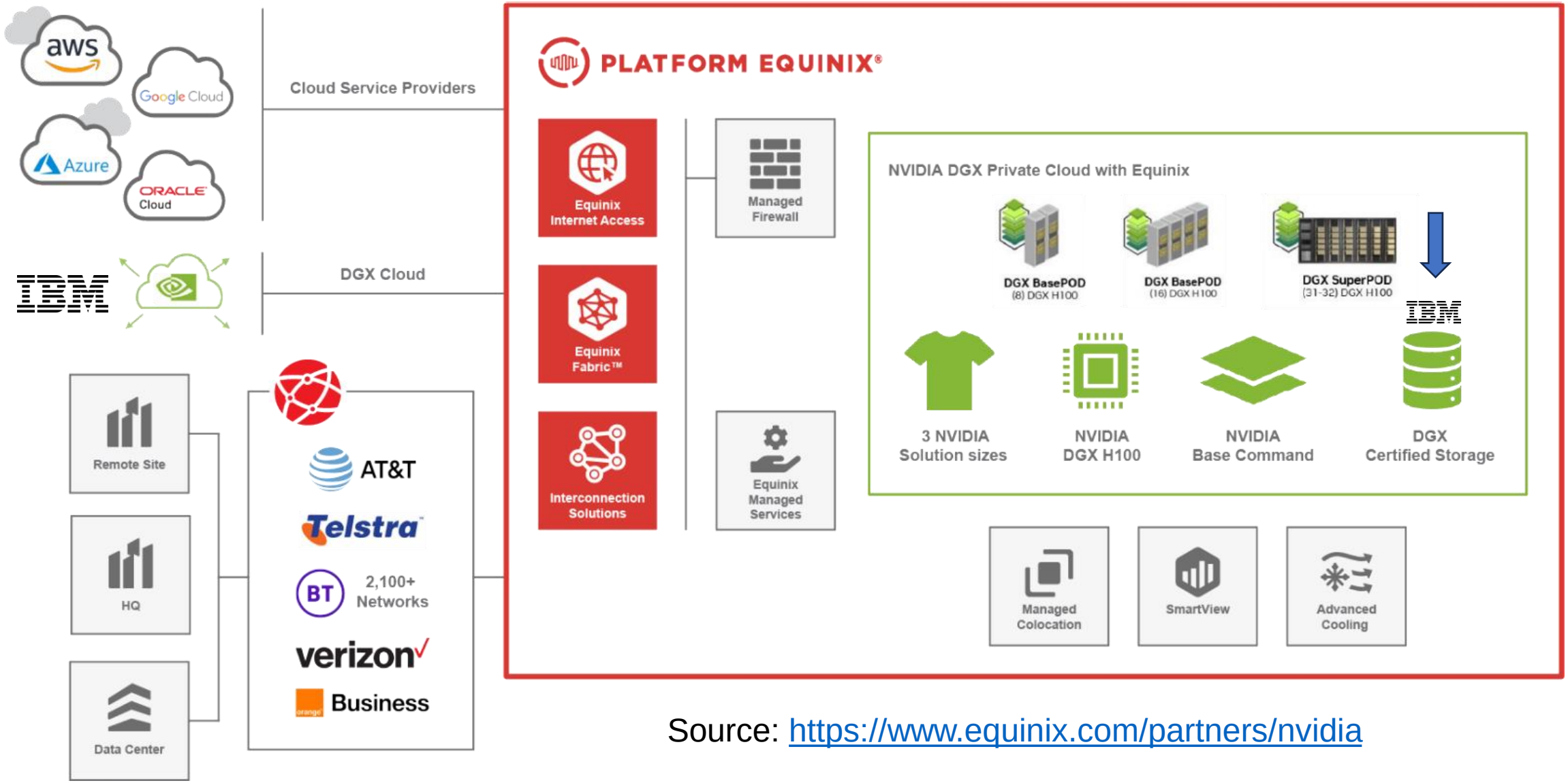
Compute Foundation of NVIDIA H100 SuperPOD Architecture

Example: 4x Nvidia **S**calable **U**nits (SU)

1024xH100 GPUs in 128-node DGX SuperPOD with 4xDGX H100 systems per rack and spine rack.

Equinix Private AI with NVIDIA DGX

Turnkey, ready-to-run AI development platform



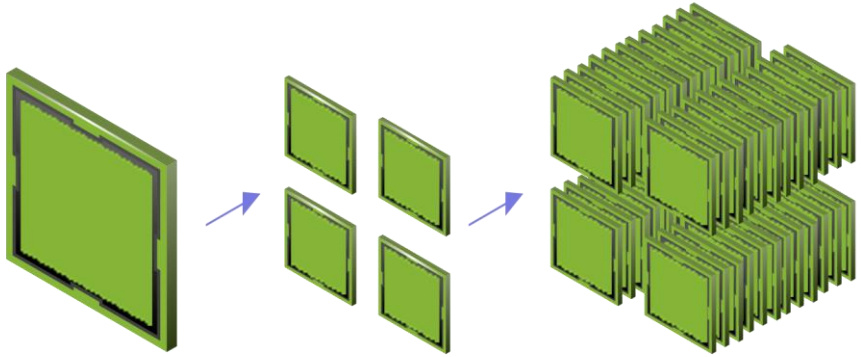
Source: https://www.equinix.com/partners/nvidia

NVIDIA Collective Communications Library (NCCL)

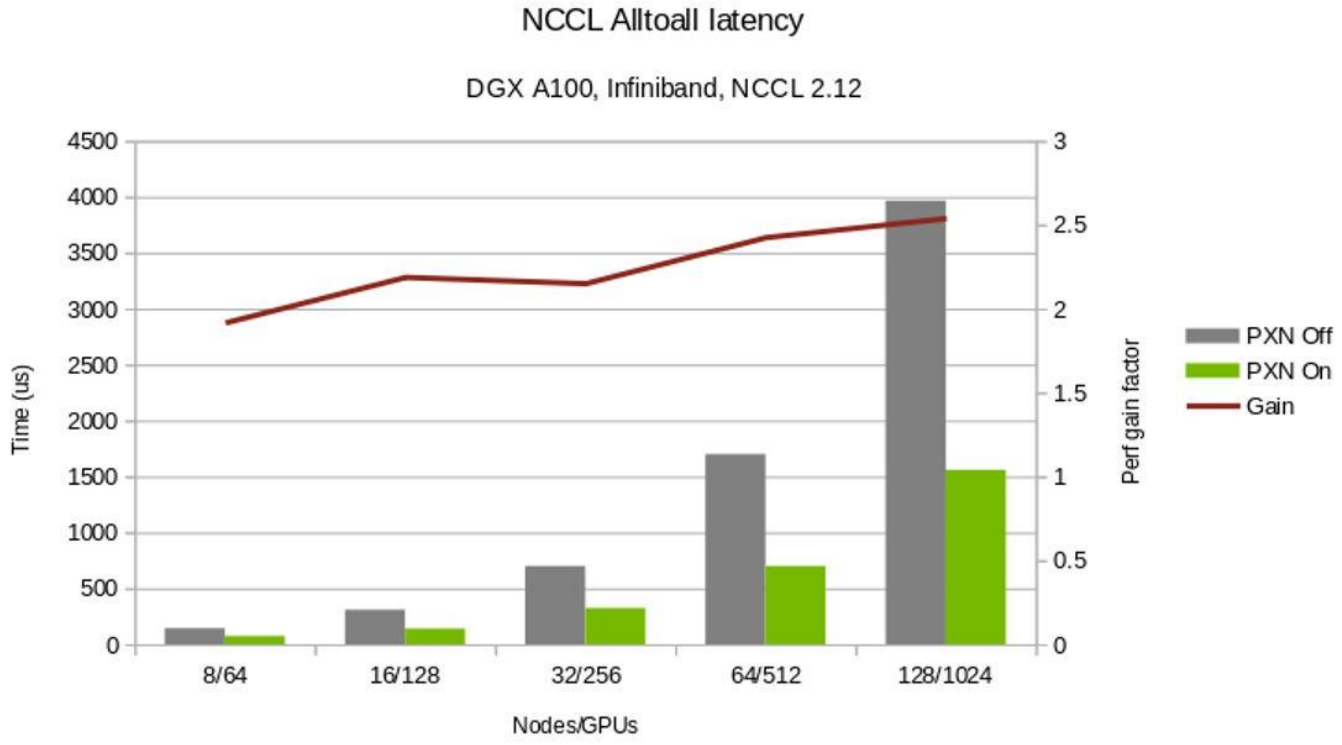
NCCL provides fast collectives over multiple GPUs

The NVIDIA Collective Communication Library (NCCL) implements multi-GPU and multi-node communication primitives optimized for NVIDIA GPUs and Networking. NCCL provides routines such as **all-gather, all-reduce, broadcast, reduce, reduce-scatter as well as point-to-point send and receive** that are optimized to achieve high bandwidth and low latency over PCIe and NVLink high-speed interconnects within a node and over NVIDIA Mellanox Network across nodes.

https://developer.nvidia.com/nccl



GPU



NVIDIA GPUDirect RDMA and GPUDirect Storage

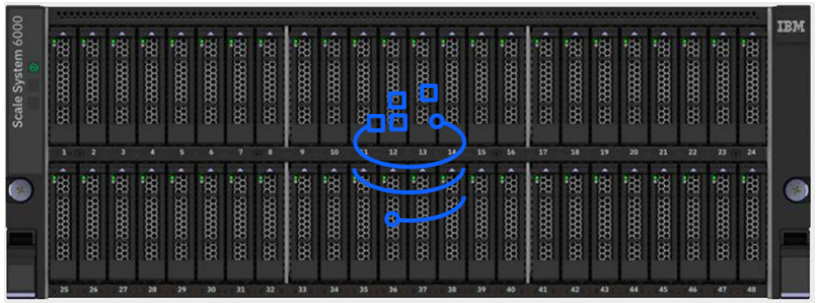
Enhancing Data Movement and Access for GPUs

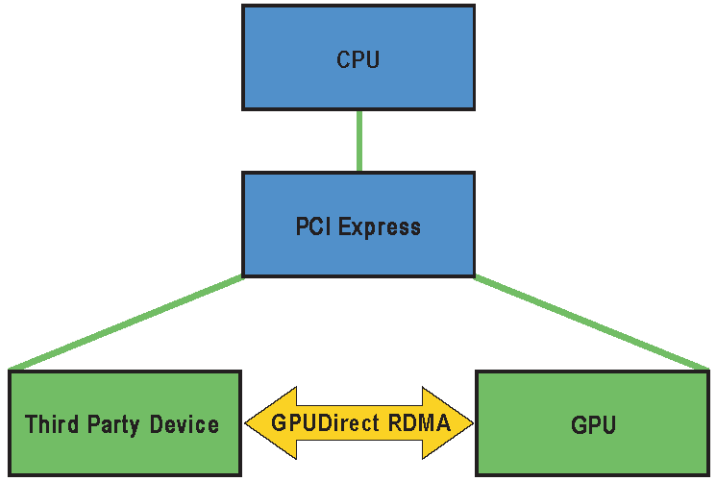
GPUDirect RDMA is a technology in NVIDIA GPUs that enables direct data exchange between GPUs and a third-party peer device using PCI Express.

GPUDirect Storage (GDS) enables a direct data path between local or remote storage. GDS leverages direct memory access (DMA) transfers between GPU memory and storage, which avoids a bounce buffer through the CPU. This direct path increases system bandwidth and decreases the latency and utilization load on the CPU. To support GPUDirect RDMA, a user space CUDA APIs and kernel mode drivers are required.

https://docs.nvidia.com/cuda/gpudirect-rdma/

IBM SSS6000 GPUDirect Storage





A100 / H100 / GH200

Storage for Data and AI & NVIDIA GPU Solutions

Start small and scale predictably in response to demand

AI Entrant

1 DGX H100
or
1x HGX H100

- Half Populated SSS 3500
- Up to 60 GB/s

AI Medium

8 x DGX H100
or
8 x HGX H100

- 1 x SSS 3500
- Up to 125 GB/s read

or

- Half populated SSS 6000
- Up to 150 GB/s read

AI Master

16 x DGX H100
or
16 x HGX H100

- 2 x SSS 3500
- Up to 250 GB/s read

or

- 1 x SSS 6000
- Up to 300 GB/s read

AI Scaler

DGX H100 SuperPOD
or
HGX H100 SuperPOD

- 2 x SSS 3500
- Up to 250 GB/s read

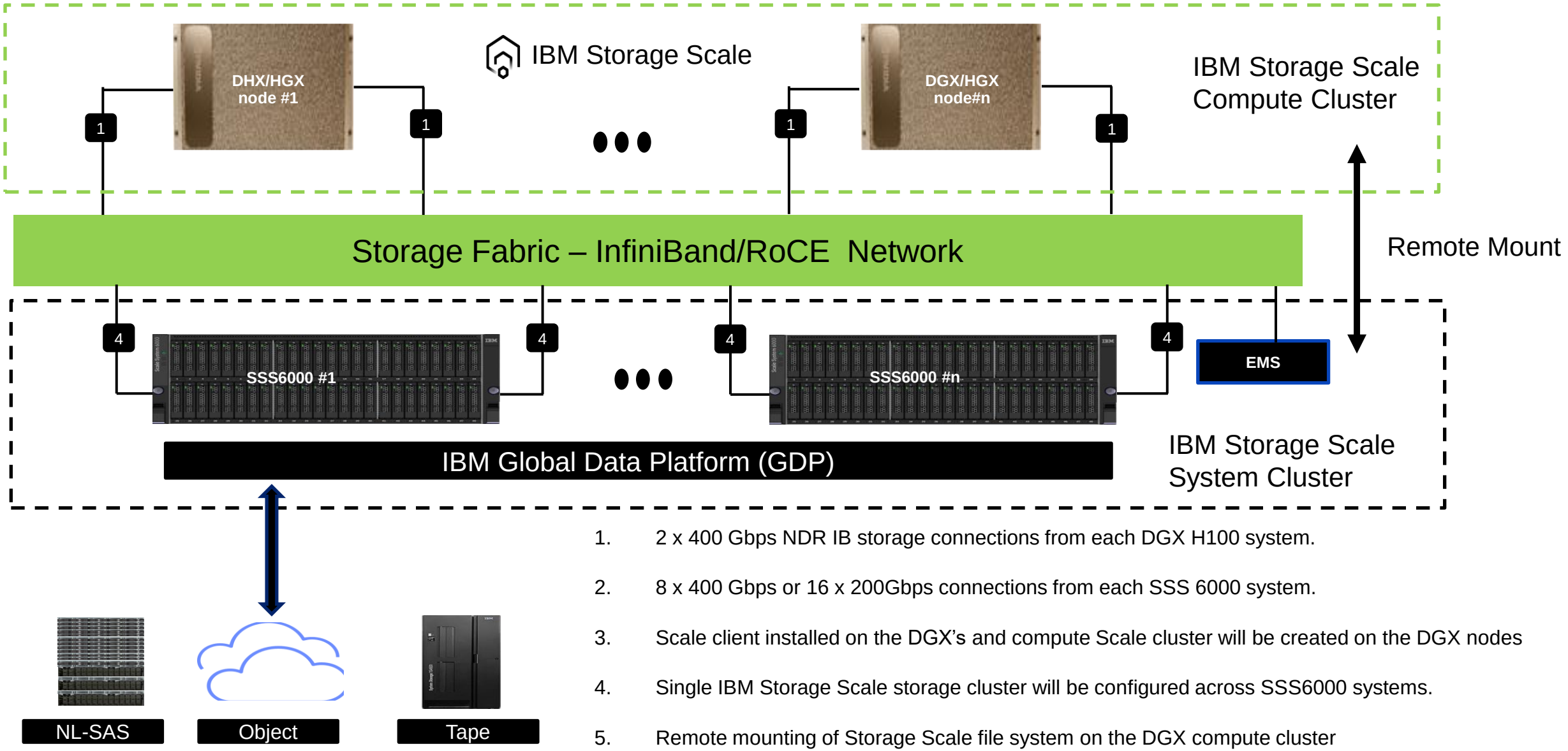
or

- 1 x SSS 6000
- Up to 300 GB/s read

- Simple building blocks – scalable seamless storage upgrade path as needs grow from 1st DGX to AI CoE DGX SuperPOD
- Global Data Platform – Data fidelity capabilities to automate AI workflows.
- Data Economics – Eliminate copies and transparently tier
- Trusted, global enterprise level support and services.
- Successful deployments across the globe

A simple, scalable upgrade path

NVIDIA H100 SuperPOD Storage Fabric – SSS 6000



1. 2 x 400 Gbps NDR IB storage connections from each DGX H100 system.
2. 8 x 400 Gbps or 16 x 200Gbps connections from each SSS 6000 system.
3. Scale client installed on the DGX's and compute Scale cluster will be created on the DGX nodes
4. Single IBM Storage Scale storage cluster will be configured across SSS6000 systems.
5. Remote mounting of Storage Scale file system on the DGX compute cluster
6. IBM Global Data Platform - Transparent external tiering to the low-cost storage.

NVIDIA H100 SuperPOD Scalable Unit storage sizing guidelines

Start small and scale predictably in response to demand

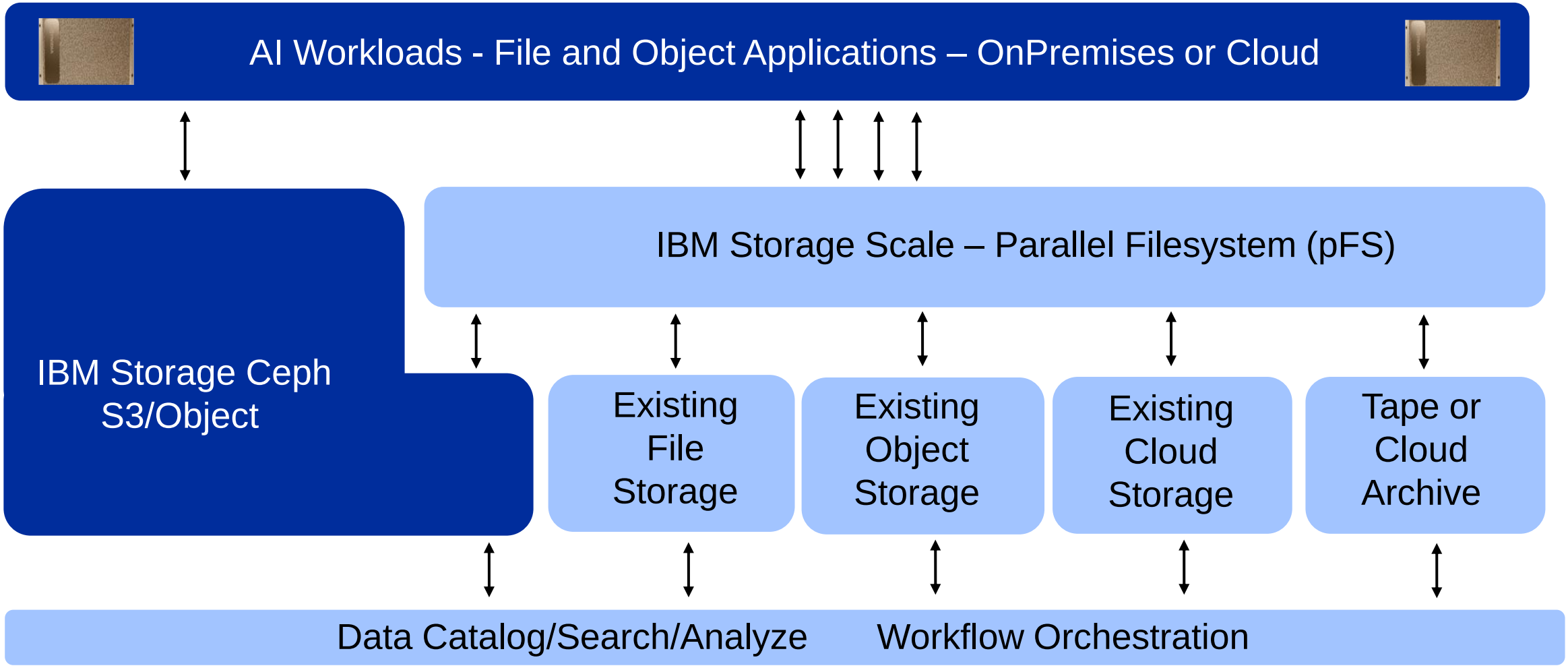
			NVIDIA Guidance		IBM SSS 6000 Performance			IBM SSS 6000 Capacity			
# of SUs	# of DGXs	# of GPUs	Read	Write	Qty	Read GB/s	Write GB/s	Qty	7.68 TB NVMe Usable	15.36 TB NVMe Usable	30.74 TB NVMe Usable
1 SU	32	256	125 GB/s	62 GB/s	1	300	150	1	0.25 PB	0.5 PB	1 PB
2 SU	64	512	250 GB/s	125 GB/s	1	300	150	1	0.25 PB	0.5 PB	1 PB
3 SU	96	786	375 GB/s	187 GB/s	2	600	300	2	0.50 PB	1 PB	2 PB
4 SU	128	1024	500 GB/s	250 GB/s	2	600	300	2	0.5 PB	1 PB	2 PB
8 SU	256	2048	1 TB/s	500 GB/s	4	1200	600	4	1 PB	2 PB	4 PB
16 SU	512	4096	2 TB/s	1 TB/s	7	2100	1050	7	1.75 PB	3.5 PB	7 PB
32 SU	1024	8192	4 TB/s	2 TB/s	**14**	**4200**	**2100**	14	3.5 PB	7 PB	14 PB

1. Lot of I/O
2. Understand the Job type & model size
   - LLMs or other GENAI workloads
3. 2:1 Read: Write
4. Most important are Read & Re-Read
5. Writes are massive with large parameters models with 175B+
6. Scalable Performance really matters
7. Parallel File Storage (pFS) is a scratch space, not long-term storage
   - PFS represents 10-20% of overall storage in modern AI scalable deployments
   - Tiering to object/NFS/Tape for 80-90% of overall Storage

Details see:
https://www.ibm.com/downloads/cas/JBVQYVXB

IBM Global Data Platform (GDP)

Unifying Unstructured Data for AI workloads



Blog: https://community.ibm.com/community/user/storage/blogs/david-wohlford1/2023/10/26/ibm-announces-the-ultimate-way-to-build-a-global-d

STORAGE for AI MATTERS

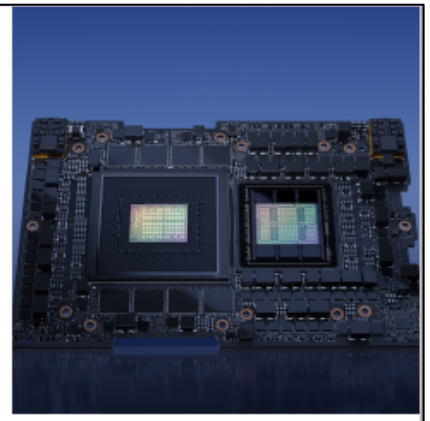
the fastest servers in the world are the world's
slowest servers if they are waiting for data
ALL GPUS & CPUS WAIT AT THE SAME SPEED

Datasheet

NVIDIA

NVIDIA GH200 Grace Hopper Superchip

The breakthrough accelerated CPU for large-scale AI and high-performance computing (HPC) applications.



The World's Most Versatile Computing Platform

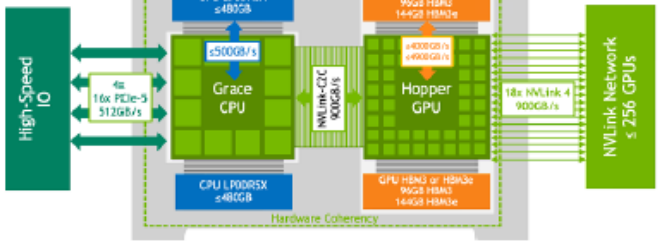
The NVIDIA Grace Hopper™ architecture brings together the groundbreaking performance of the NVIDIA Hopper™ GPU with the versatility of the NVIDIA Grace™ CPU in a single superchip, connected with the high-bandwidth, memory-coherent NVIDIA® NVLink® Chip-2-Chip (C2C) interconnect.

NVIDIA NVLink-C2C is a memory-coherent, high-bandwidth, and low-latency interconnect for superchips. The heart of the GH200 Grace Hopper Superchip, it delivers up to 900 gigabytes per second (GB/s) of total bandwidth, which is 7X higher than PCIe Gen5 lanes commonly used in accelerated systems. NVLink-C2C enables applications to oversubscribe the GPU's memory and directly utilize NVIDIA Grace CPU's memory at high bandwidth. With up to 480GB of LPDDR5X CPU memory per GH200 Grace Hopper Superchip, the GPU has direct access to 7X more fast memory than HMB3 or almost 8X more fast memory with HBM3e. GH200 can be easily deployed in standard servers to run a variety of inference, data analytics, and other compute and memory-intensive workloads. GH200 can also be combined with the NVIDIA NVLink Switch System, with all GPU threads running on up to 256 NVLink-connected GPUs and able to access up to 144 terabytes (TB) of memory at high bandwidth.

Key Features

- 72-core NVIDIA Grace CPU
- NVIDIA H100 Tensor Core GPU
- Up to 480GB of LPDDR5X memory with error-correction code (ECC)
- Supports 96GB of HBM3 or 144GB of HBM3e
- Up to 624GB of fast-access memory
- NVLink-C2C: 900GB/s of coherent memory

NVIDIA GH200 Grace Hopper Superchip



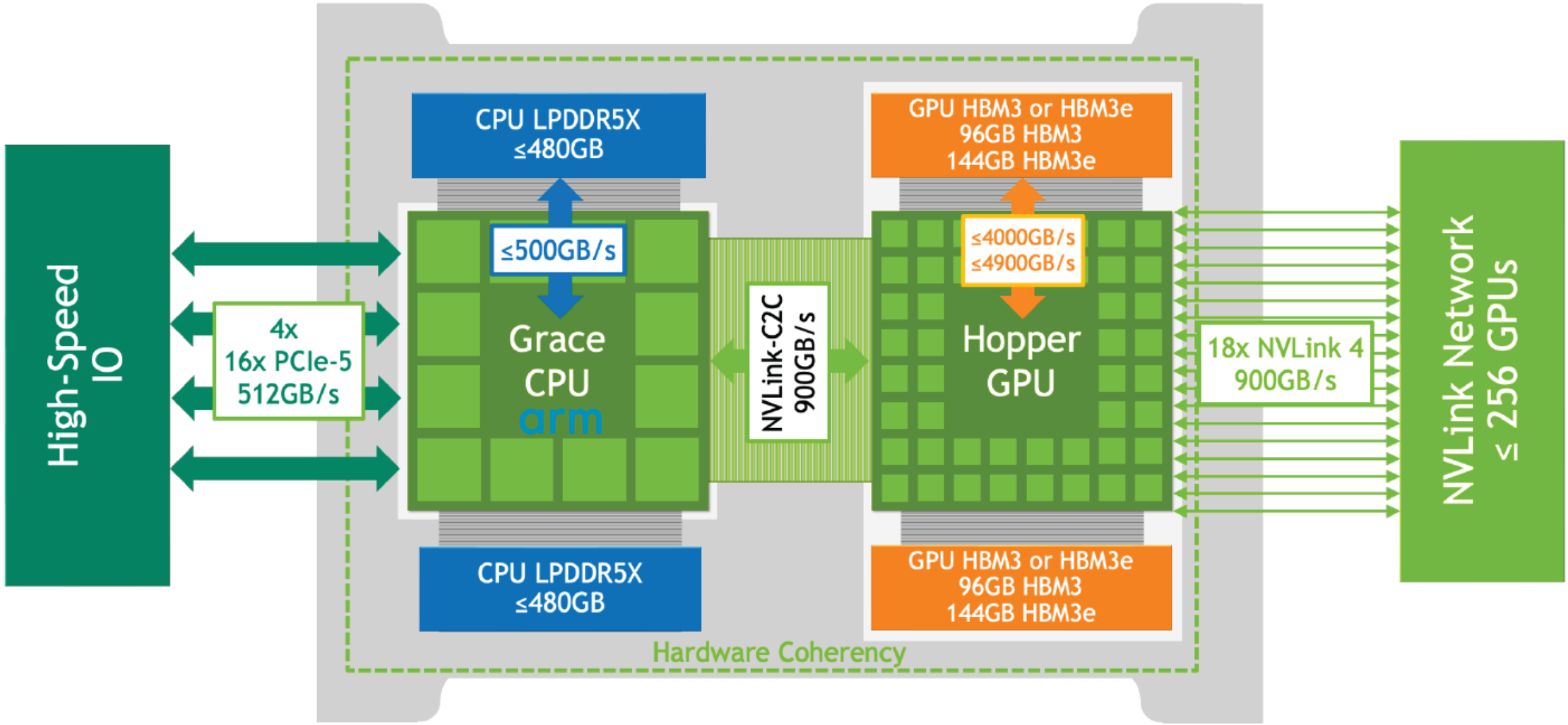
NVIDIA Grace Hopper GH200 Superchip Architecture

The NVIDIA Grace Hopper Superchip architecture brings together the performance of the **NVIDIA Hopper GPU** with the of the **NVIDIA Grace CPU**, connected with a high bandwidth and memory coherent NVIDIA NVLink Chip-2-Chip (C2C) interconnect in a single chip, and support for the new NVIDIA NVLink Switch System.

NVLink-C2C enables applications to oversubscribe the GPU's memory and directly utilize NVIDIA Grace CPU's memory at high bandwidth. With up to **512 GB of LPDDR5X CPU** memory per Grace Hopper Superchip, the GPU has direct high-bandwidth access to 4x more memory than what is available with HBM. Combined with the NVIDIA NVLink Switch System, all GPU threads running on up to 256 NVLink-connected GPUs can now access up to 150 TB of memory at high bandwidth. Fourth-generation NVLink enables accessing peer memory using direct loads, stores, and atomic operations, enabling accelerated applications to solve larger problems.

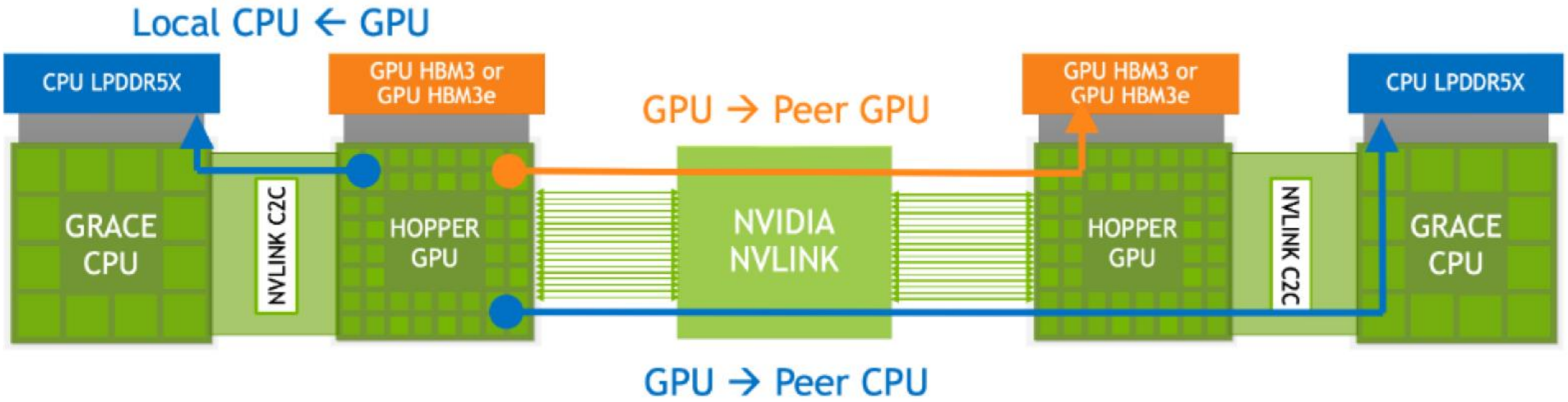
https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper-2

NVIDIA GH200 Grace Hopper Superchip

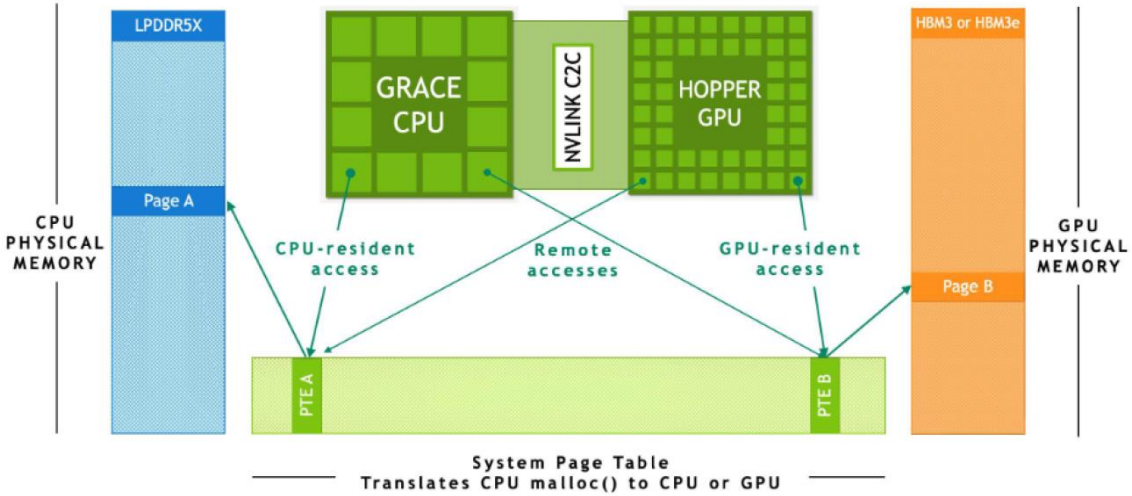


Source: https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper

Accelerating AI Applications with Extended GPU Memory (EGM)



The Extended GPU Memory (EGM) feature over the high-bandwidth NVLink-C2C enables GPUs to access all the system memory efficiently. EGM provides up to 19.5 TBs system memory in a multi-node NVSwitch-connected system.



Source: https://developer.nvidia.com/blog/nvidia-grace-hopper-superchip-architecture-in-depth/

Technology Preview - IBM Storage Scale on ARM64

Is like any other supported Linux architecture

IBM Storage Scale client for HPC compute node(s) does include:

- rpm based install
- Network Shared Disk (NSD) client
- Scale base functionality (IO, policies, remote mounts, snapshots, quotas, etc.)
- Manager roles: file system manager / token manager / cluster manager
- RDMA (IB or RoCE)
- Health Monitoring
- Target OS: RHEL 9

JUPITER, the “Joint Undertaking Pioneer for Innovative and Transformative Exascale Research”, will be the first Exascale supercomputer in Europe.

24.000 NVIDIA Grace Hopper (GH200) GPUs

- Official architecture name is aarch64 / ARM64
- There are no new configuration options specific to ARM
- All supported features work identical compared to other Linux architectures
- Reference customer is HPC Research Center Juelich
- JUPITER Systems Overview see: https://www.fz-juelich.de/en/ias/jsc/jupiter/tech
- Technical Details: https://insidehpc.com/2024/01/eviden-to-deliver-modular-data-center-for-jsc-to-host-jupiter-europes-1st-exascale-system/

IBM Global Data Platform (GDP) for AI

https://www.ibm.com/downloads/cas/NO6AGVVK

- Data Access Services (DAS) with multi-protocol performance that connects directly to your file and object applications with multiple parallel paths.
- Data Caching or Core Services (AFM) that provide global connectivity from multiple data sources and multiple locations to bring together data from IBM and non-IBM storage environments.
- Data Management Services with policy automation that transparently helps mange the flow of data and take much of the complexity out of day-to-day data management.
- Data Security Services that provide cyber secure automation to ensure your data is protected and safe and quickly recoverable when needed.

IBM Storage Scale System is designed to optimize, secure and unlock AI data

Why IBM Storage for AI?

- Reliability & Data Durability. Zero Data loss. Self-healing. Autonomic failure domains.
- Data Security & Governance. Encryption. Regulatory compliance.
- Scalability. On-demand horizontal scaling. Autonomic re-balancing. Non-disruptive. Tiering support.
- CapEx Efficiency. Runs on low-cost standard technology servers. NVMe and/or HDDs.
- OpEx Efficiency. Easy deployment, upgrades and maintenance via Ansible automation.
- Flexibility & Universal. Object, File, and Block access protocols.
- Hybrid Cloud. Having seamless access to data on-prem and on-cloud providers.