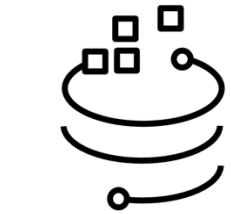
NVIDIA Technology for AI and HPC Overview

IBM Storage Scale Days 2025 DE
March 19th – 20th, 2025 | Heidelberg, Germany

Dr. Thomas Schoenemeyer (NVIDIA)
mailto:thomass@nvidia.com

Frank Kraemer (IBM)
mailto: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.

Dev Teams are faced with **data challenges** to scale AI

1 - There's more data

Exploding data growth

The aggregate volume of data stored is set to grow over 250% in the next 5 years.

2 - In more locations

Multiple locations, clouds, applications and silos

82% of enterprises are inhibited by data silos.

3 - In more formats

Documents, images, video

80% of time is spent on data cleaning, integration and preparation.

4 - With less quality

Stale and inconsistent

82% of enterprises say data quality is a barrier on their data integration projects.

IBM **G**lobal **D**ata **P**latform helps to unlock the full potential of AI

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.

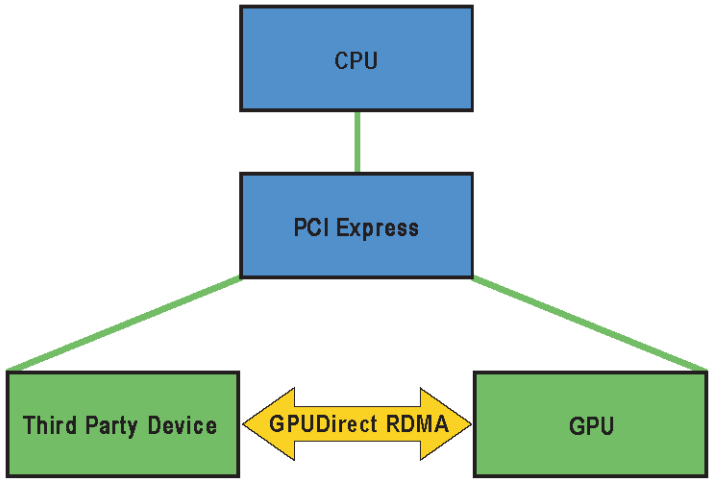
GPUDirect Storage (GDS) enables a direct data path between local or remote storage. GDS leverages direct memory access (DMA) transfers between GPU memory and storage, which avoids a bounce buffer through the CPU. This direct path increases system bandwidth and decreases the latency and utilization load 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/

NVIDIA CERTIFIED

NVIDIA-Certified Systems

IBM 6000 GPUDirect Storage



NVIDIA

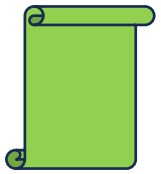
A100 / H100 / GH200 / GB200

Nvidia DGX **SuperPOD** & DGX **BasePOD**

Enhancing Data Movement and Access for GPUs

NVIDIA DGX SuperPOD is AI data center infrastructure that enables IT to deliver performance—without compromise—for every user and workload. As part of the NVIDIA DGX platform, **DGX SuperPOD** offers leadership-class accelerated infrastructure and scalable performance for the most challenging AI workloads, with industry-proven results.

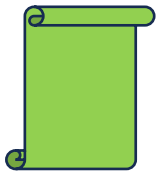
https://www.nvidia.com/en-us/data-center/dgx-superpod/



NVIDIA-Certified Systems

NVIDIA DGX BasePOD AI is powering mission-critical use cases in every industry—from healthcare to manufacturing to financial services. As part of the NVIDIA DGX platform, **NVIDIA DGX BasePOD** provides the critical foundation on which business transformation is realized and AI applications are born.

https://www.nvidia.com/en-us/data-center/dgx-basepod/

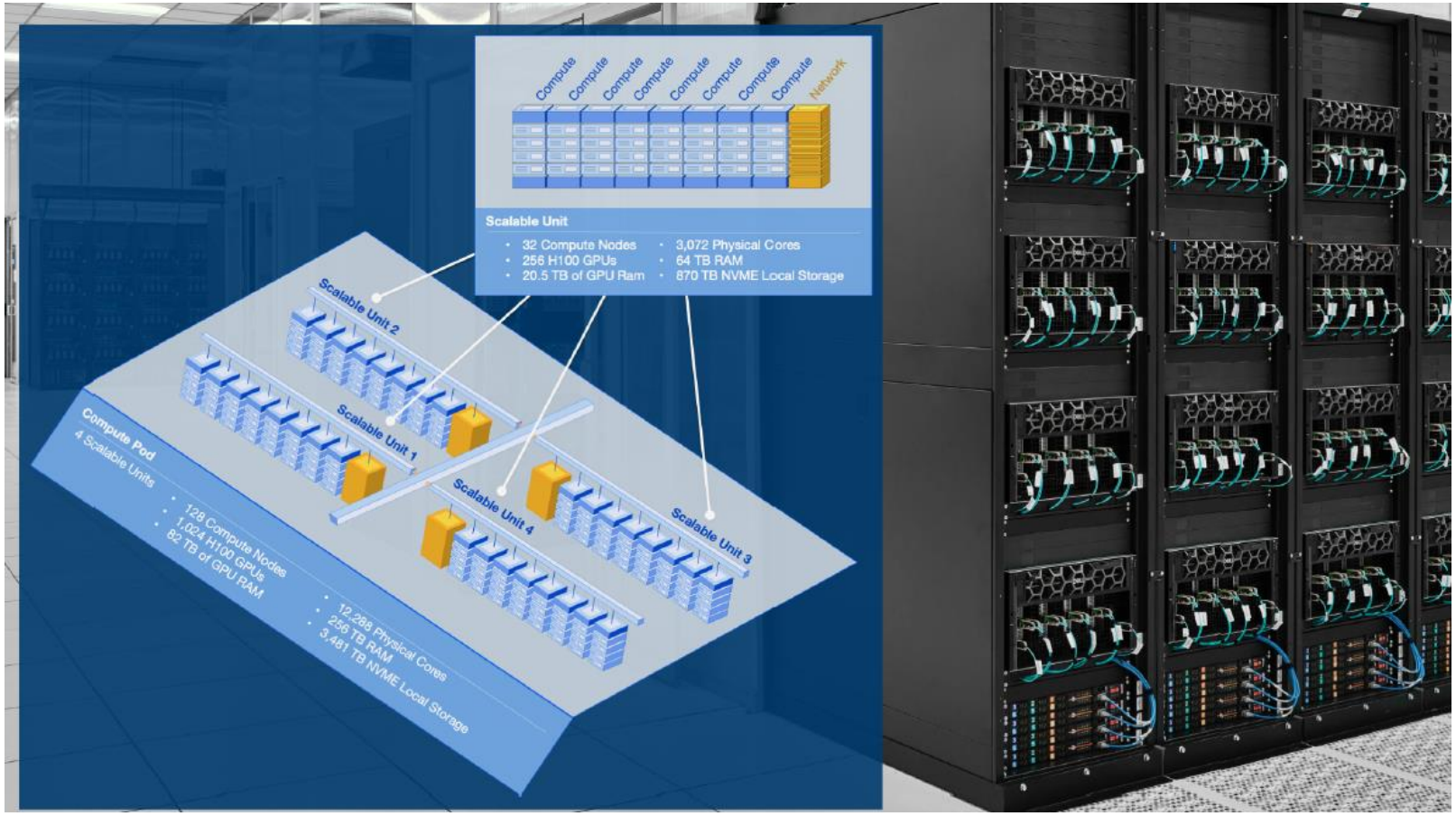


NVIDIA-Certified Systems

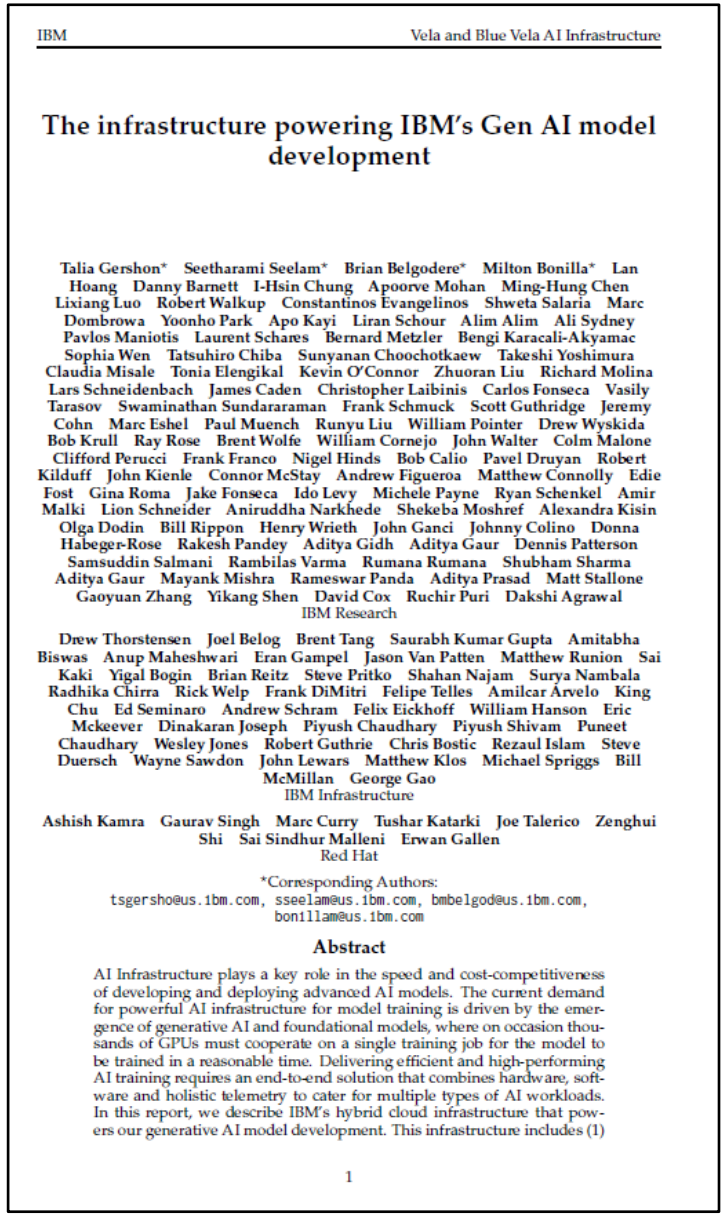
Jensen @ IBM booth Nvidia GTC 2024 with IBM 6000

IBM Blue Vela AI Supercomputer (Compute)

Blue Vela AI supercomputer is based on NVIDIA reference HGX Platform guidelines.
Dell PowerEdge XE9680 nodes each with 8x NVIDIA H100 GPUs (80 GB HBM)



IBM's AI supercomputer:
https://research.ibm.com/blog/vela-ai-supercomputer-updates



IBM Vela and Blue Vela AI Infrastructure

The infrastructure powering IBM's Gen AI model development

Talia Gershon* Seetharami Seelam* Brian Belgodere* Milton Bonilla* Lan Hoang Danny Barnett I-Hsin Chung Apoorve Mohan Ming-Hung Chen Lixiang Luo Robert Walkup Constantinos Evangelinos Shweta Salaria Marc Dombrowa Yoonho Park Apo Kayi Liran Schour Alim Alim Ali Sydney Pavlos Maniotis Laurent Schares Bernard Metzler Bengi Karacali-Akyamac Sophia Wen Tatsuhiro Chiba Sunyanan Choochotkaew Takeshi Yoshimura Claudia Misale Tonia Elengikal Kevin O'Connor Zhuoran Liu Richard Molina Lars Schneidenbach James Caden Christopher Laibinis Carlos Fonseca Vasily Tarasov Swaminathan Sundararaman Frank Schmuck Scott Guthridge Jeremy Cohn Marc Eshel Paul Muench Runyu Liu William Pointer Drew Wyskida Bob Krull Ray Rose Brent Wolfe William Cornejo John Walter Colm Malone Clifford Perucci Frank Franco Nigel Hinds Bob Calio Pavel Druyan Robert Kilduff John Kienle Connor McStay Andrew Figueroa Matthew Connolly Edie Fost Gina Roma Jake Fonseca Ido Levy Michele Payne Ryan Schenkel Amir Malki Lion Schneider Aniruddha Narkhede Shekeba Moshref Alexandra Kisin Olga Dodin Bill Rippon Henry Wrieth John Ganci Johnny Colino Donna Habeger-Rose Rakesh Pandey Aditya Gidh Aditya Gaur Dennis Patterson Samsuddin Salmani Rambilas Varma Rumana Rumana Shubham Sharma Aditya Gaur Mayank Mishra Rameswar Panda Aditya Prasad Matt Stallone Gaoyuan Zhang Yikang Shen David Cox Ruchir Puri Dakshi Agrawal
IBM Research

Drew Thorstensen Joel Belog Brent Tang Saurabh Kumar Gupta Amitabha Biswas Anup Maheshwari Eran Gampel Jason Van Patten Matthew Runion Sai Kaki Yigal Bogin Brian Reitz Steve Pritko Shahan Najam Surya Nambala Radhika Chirra Rick Welp Frank DiMitri Felipe Telles Amilcar Arvelo King Chu Ed Seminaro Andrew Schram Felix Eickhoff William Hanson Eric Mckeever Dinakaran Joseph Piyush Chaudhary Piyush Shivam Puneet Chaudhary Wesley Jones Robert Guthrie Chris Bostic Rezaul Islam Steve Duersch Wayne Sawdon John Lewars Matthew Klos Michael Spriggs Bill McMillan George Gao
IBM Infrastructure

Ashish Kamra Gaurav Singh Marc Curry Tushar Katarki Joe Talerico Zenghui Shi Sai Sindhur Malleni Erwan Gallen
Red Hat

*Corresponding Authors:
tsgersho@us.ibm.com, sseelam@us.ibm.com, bmbelgod@us.ibm.com, bonillam@us.ibm.com

Abstract

AI Infrastructure plays a key role in the speed and cost-competitiveness of developing and deploying advanced AI models. The current demand for powerful AI infrastructure for model training is driven by the emergence of generative AI and foundational models, where on occasion thousands of GPUs must cooperate on a single training job for the model to be trained in a reasonable time. Delivering efficient and high-performing AI training requires an end-to-end solution that combines hardware, software and holistic telemetry to cater for multiple types of AI workloads. In this report, we describe IBM's hybrid cloud infrastructure that powers our generative AI model development. This infrastructure includes (1)

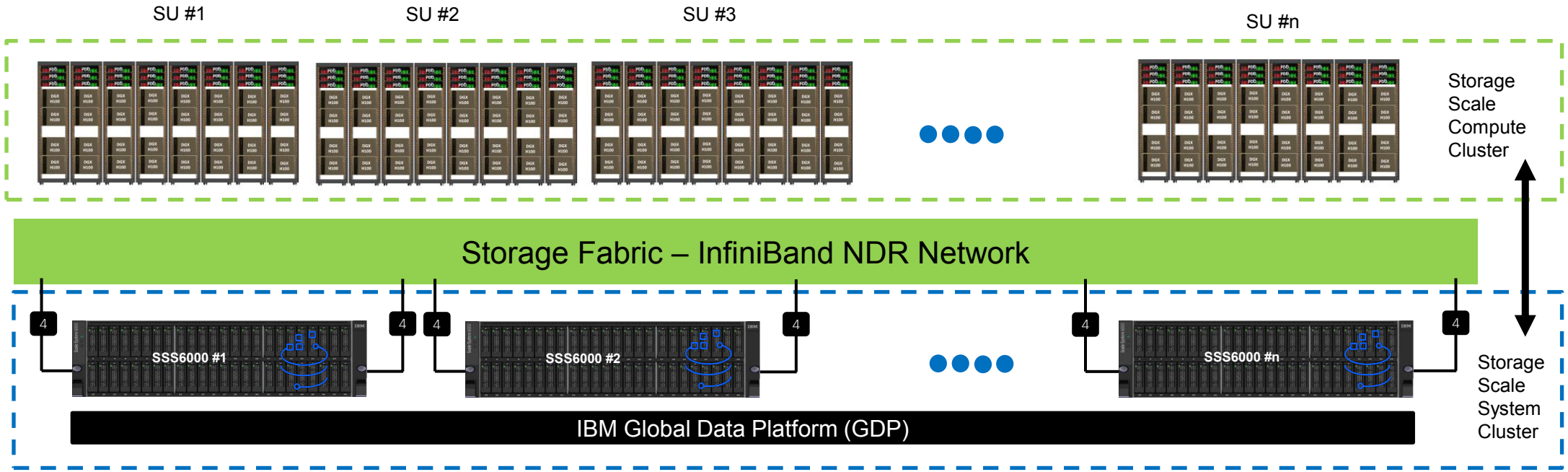
https://arxiv.org/abs/2407.05467

IBM Blue Vela AI Supercomputer (Storage)

- AI Supercomputer scalable up to 5000 H100 HGX Systems
- Phase #2 will have 20 Scalable Units (SU)
- 32x SSS6000 planned for Phase #2
- NVIDIA NDR Network Fabric for both Compute and Storage
- Enterprise AI Service
- Training LLM Models with 100B+ Parameters
- Linear scaling of performance and capacity

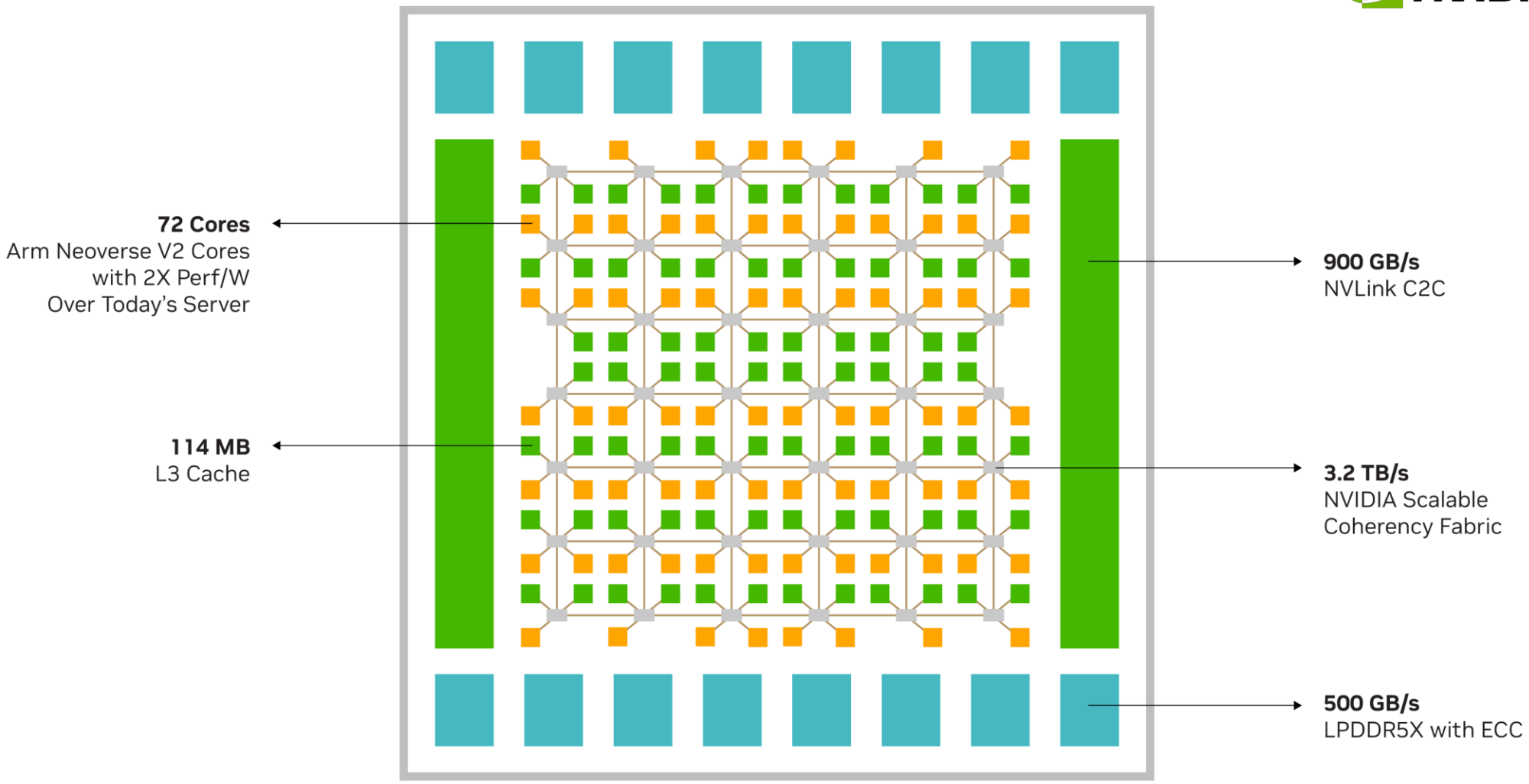
The key I/O operation in AI training is **re-read**. It is not just that data is read, but it must be reused again and again due to the iterative nature of AI training.

Write performance is important as AI models grow and time-to-train, writing checkpoints is necessary for fault tolerance. The size of checkpoint files can be terabytes in size.



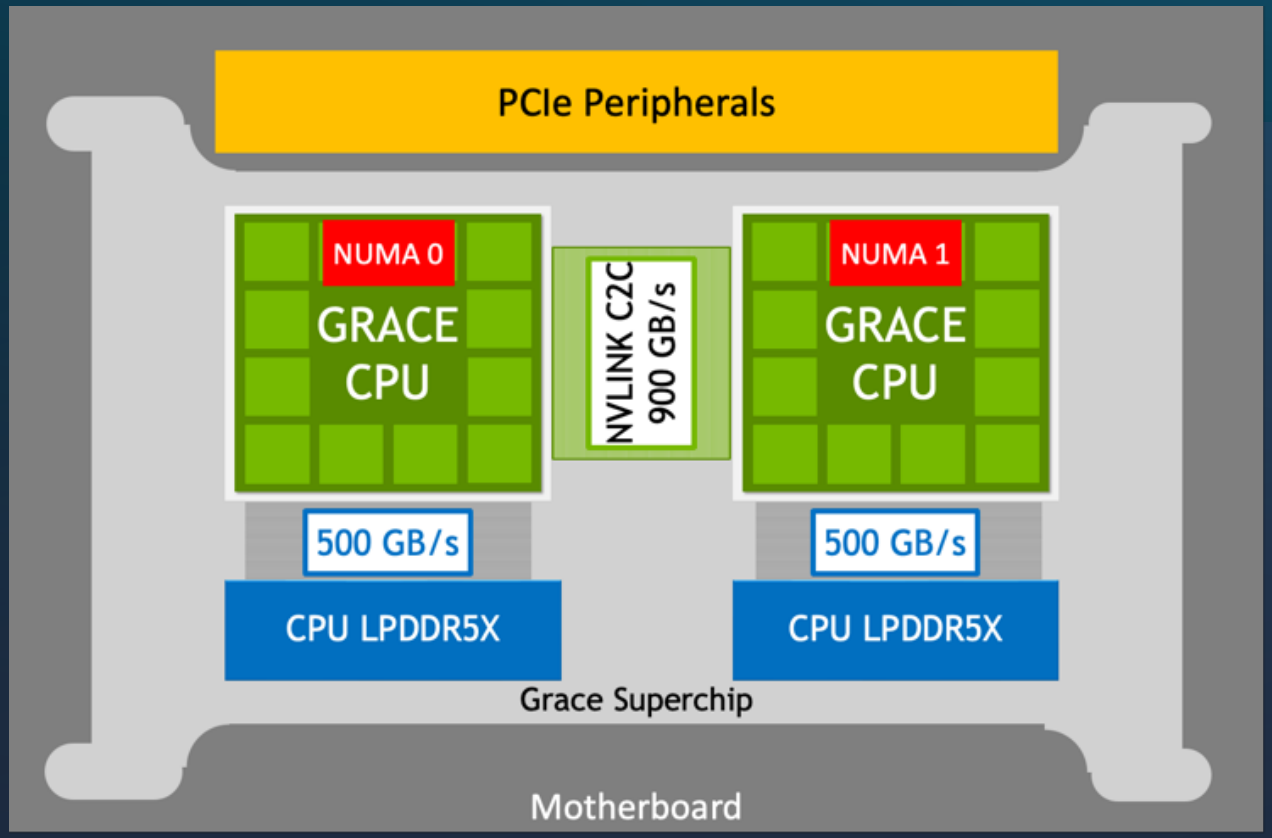
https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastructure-h100/latest/storage-architecture.html

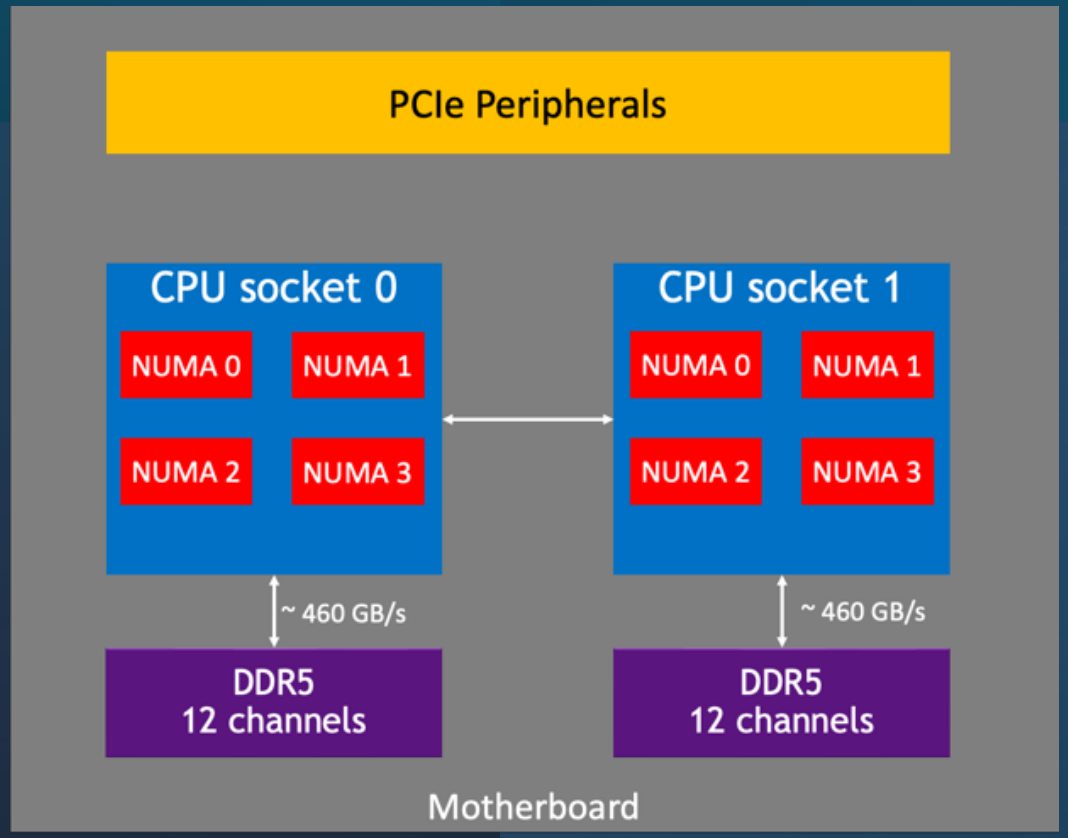
NVIDIA GRACE CPU Architecture



https://developer.nvidia.com/blog/nvidia-grace-cpu-integrates-with-the-arm-software-ecosystem/

NVIDIA GRACE SuperChip





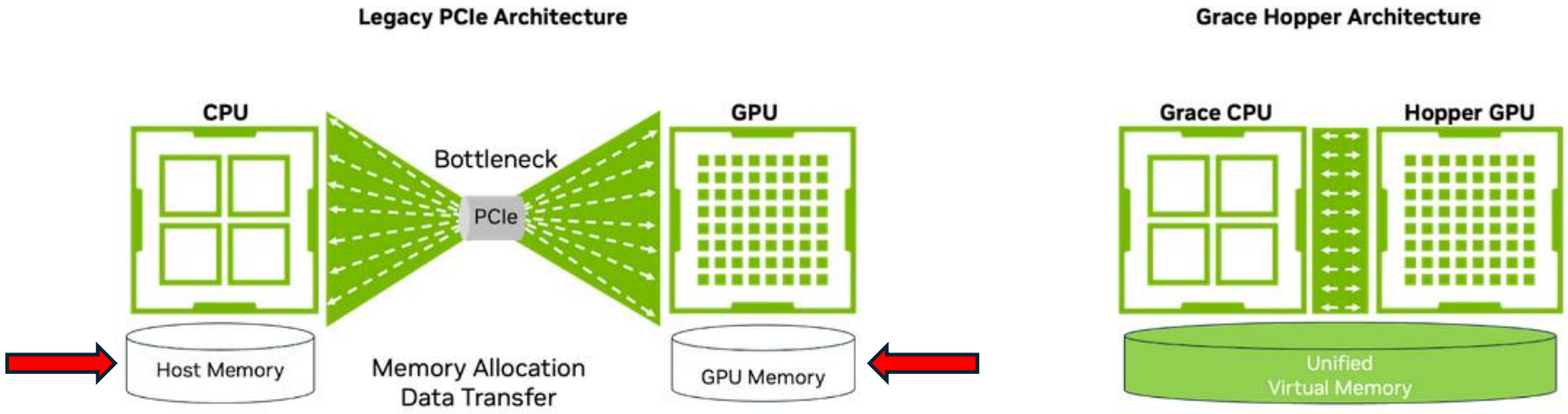
https://developer.nvidia.com/blog/nvidia-grace-cpu-superchip-architecture-in-depth/

NVIDIA GH200 Grace Hopper (GH) Architecture

Grace CPU optimized for massive memory bandwidth and efficiency. It's based on **ARM** architecture and designed to work closely with the Hopper GPU.

Hopper GPU is designed for AI and HPC, featuring advanced capabilities like support for the new Transformer Engine, which is crucial for training large AI models.

The architecture uses NVLink-C2C technology to connect the Grace CPU and Hopper GPU directly, allowing them to **share memory and data** seamlessly. This tight integration leads to significant improvements in performance, efficiency, and scalability, especially for data-intensive applications like AI model training and scientific simulations.



NVIDIA's **NVLink-C2C** (Chip-to-Chip) is a high-speed, low-latency interconnect technology designed to link multiple chips directly, such as GPUs, CPUs, or custom accelerators, on the same board. It enables faster communication between these components, reducing bottlenecks and improving overall system performance, particularly in high-performance computing and AI workloads.

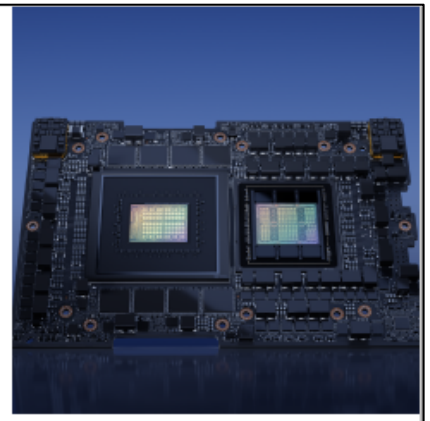
https://www.nvidia.com/en-us/data-center/grace-hopper-superchip/

Datasheet

NVIDIA

NVIDIA GH200 Grace Hopper Superchip

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



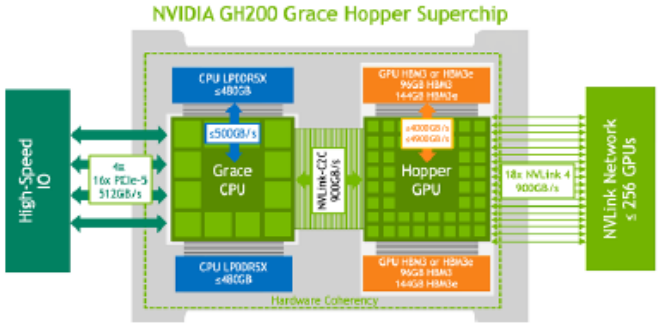
The World's Most Versatile Computing Platform

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



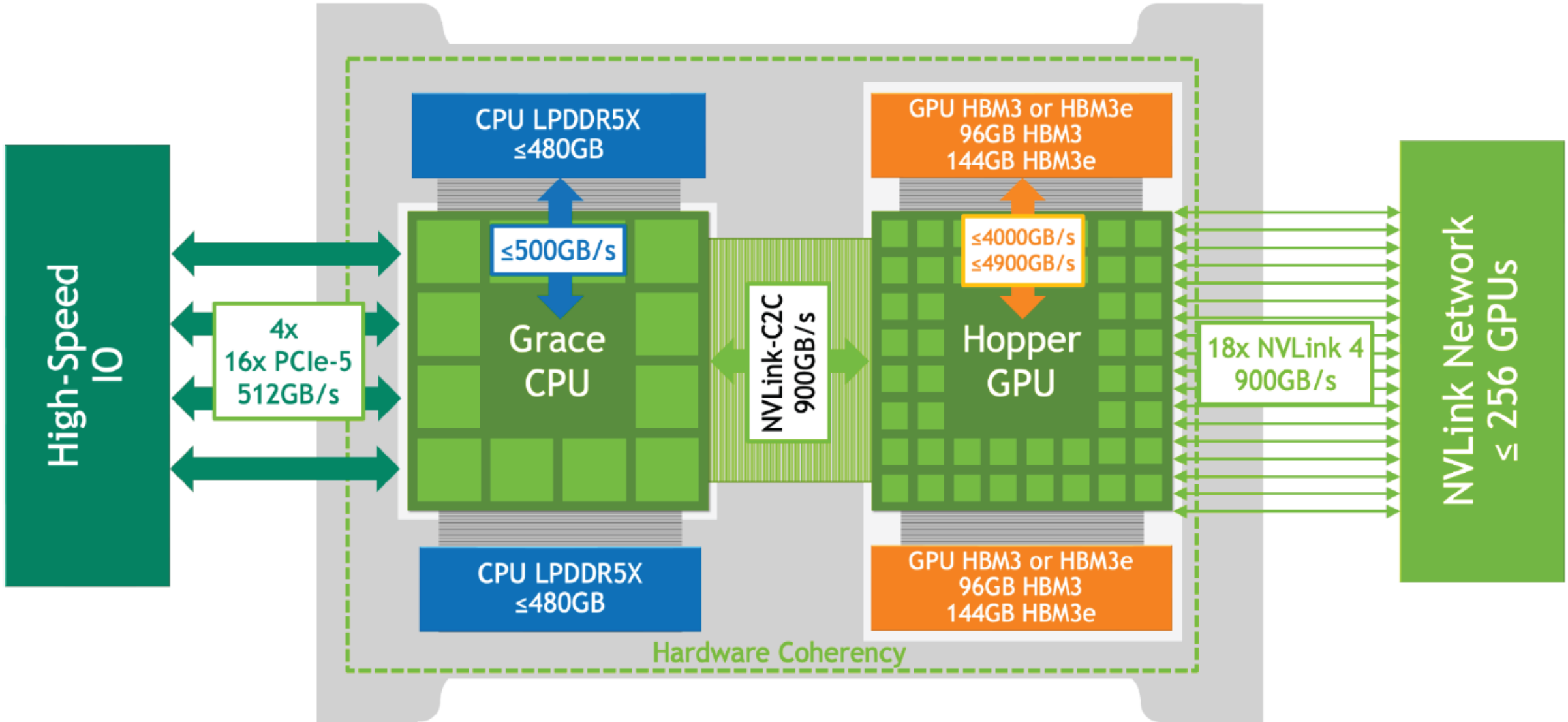
NVIDIA Grace Hopper GH200 Superchip Architecture

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.

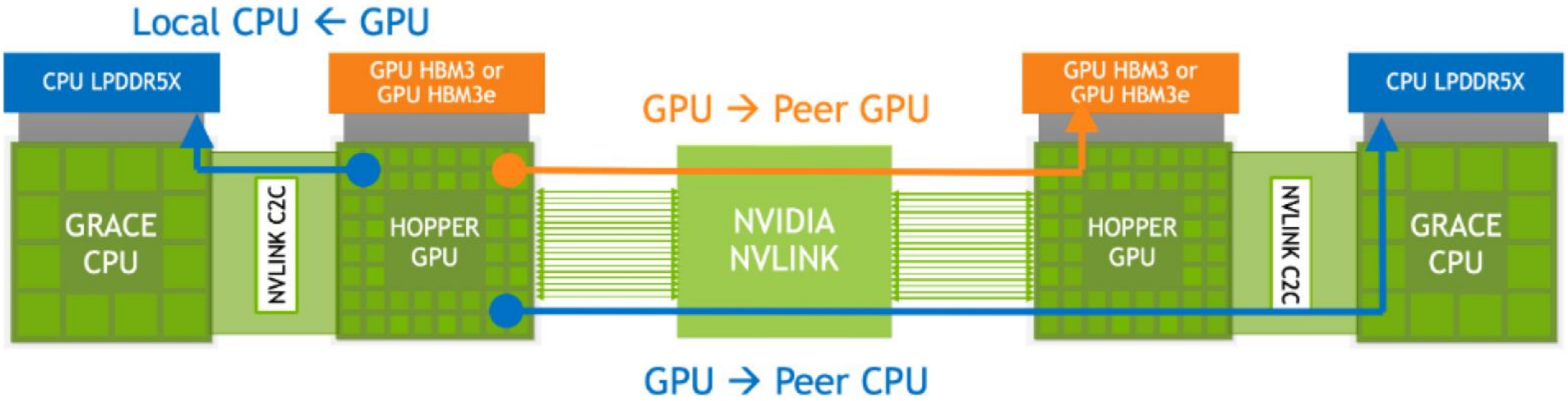
https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper-2

NVIDIA GH200 Grace Hopper Superchip

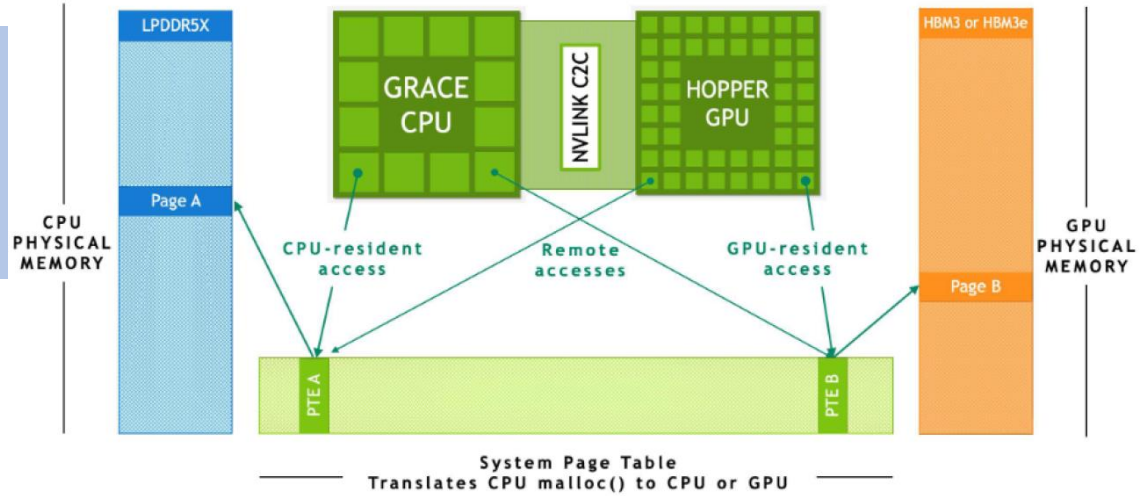


Source: https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper

Accelerating AI Applications with Extended GPU Memory

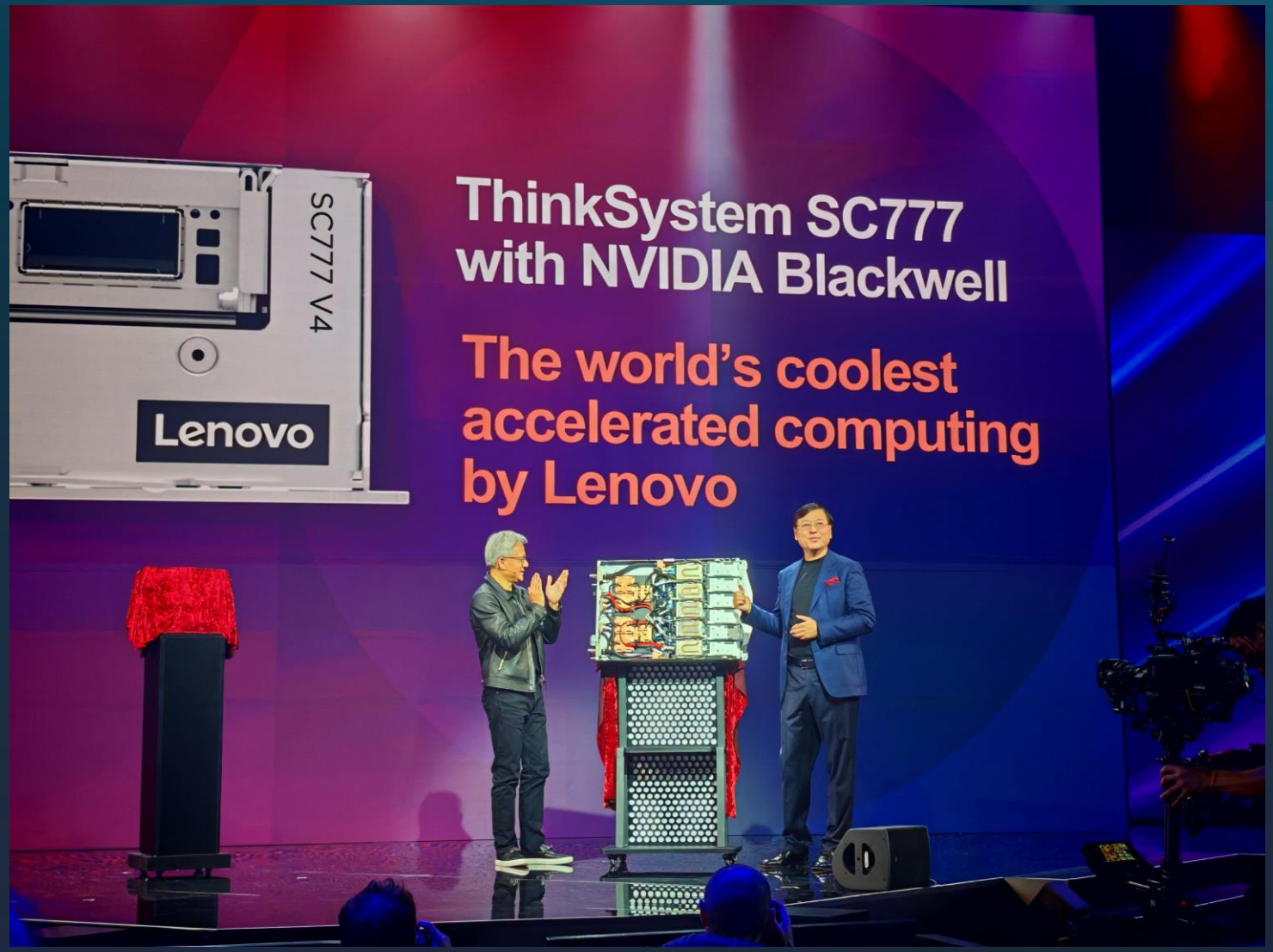


The Extended GPU Memory (EGM) feature over the high-bandwidth NVLink-C2C enables GPUs to access all the system memory efficiently. EGM provides up to 19.5 TBs system memory in a multi-node NVSwitch-connected system.



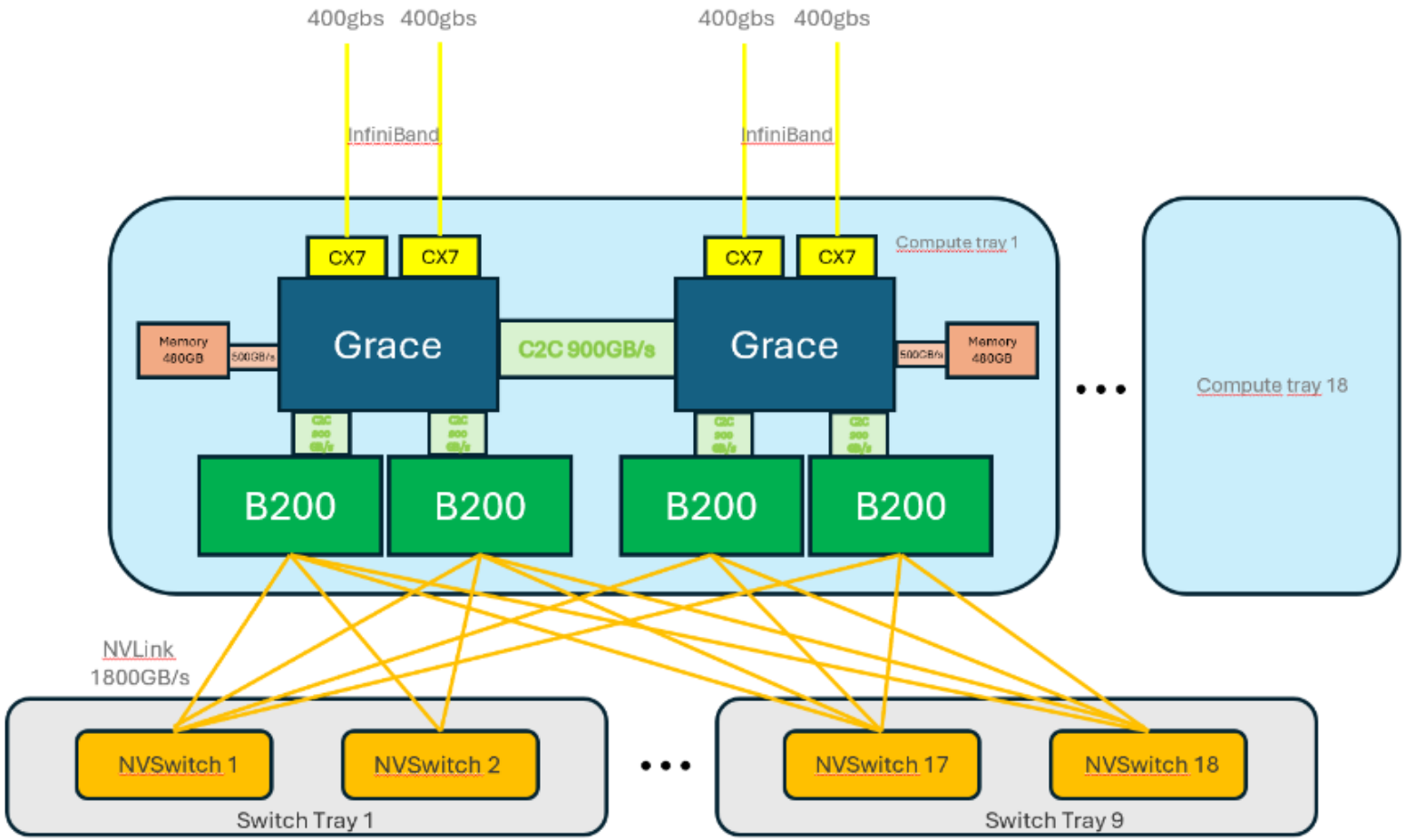
Source: https://developer.nvidia.com/blog/nvidia-grace-hopper-superchip-architecture-in-depth/

NVIDIA GB200 Grace Blackwell Superchip



https://www.nvidia.com/en-us/data-center/gb200-nvl72/

NVIDIA GB200 NVL72 - Grace Blackwell Superchip



Top of Rack 1G Switches

4x Power Shelves

10x Compute Trays

9x NVLink Switches

8x Compute Trays

Drip Pan

4x Power Shelves

DGX GB200 NVL72 System

NVIDIA GB200 NVL72 connects 36 NVIDIA Grace CPUs and 72 NVIDIA Blackwell GPUs in a rack-scale design, supercharging generative AI, data processing, and high-performance computing.

Supermicro NVIDIA GB200 NVL72 SuperCluster

https://www.supermicro.com/manuals/brochure/Brochure-AI-SuperCluster-NVIDIA-GB200-NVL72.pdf
https://www.supermicro.com/datasheet/datasheet_SuperCluster_GB200_NVL72.pdf

An Exascale of Compute in a Rack

End-to-end Liquid-cooling Solution for NVIDIA GB200 NVL72

Supermicro's GB200 NVL72 solution represents a breakthrough in AI computing infrastructure, combining Supermicro's end-to-end liquid-cooling technology. It enables up to 25x performance improvement at the same power envelope compared to previous NVIDIA Hopper generations, while reducing data center electricity costs by up to 40%. The system integrates 72 NVIDIA Blackwell GPUs and 36 NVIDIA Grace CPUs in a single rack, delivering exascale computing capabilities through NVIDIA's most extensive NVLink™ network to date, achieving 130 TB/s of GPU communications.
The 48U solution's versatility supports both liquid-to-air and liquid-to-liquid cooling configurations, accommodating various data center environments.

- 36 NVIDIA Grace CPUs and 72 NVIDIA Blackwell GPUs
- 1.8TB/s GPU-GPU Interconnect across all 72 GPUs and CPUs
- 36x NVIDIA 72-core Grace Arm Neoverse V2 CPUs
- Max 250kW CDU Liquid-to-liquid or Liquid-to-air Option

NVIDIA GB200 NVL72 SuperCluster

for NVIDIA GB200 Grace™ Blackwell Superchip

Liquid-to-Air Solution for NVIDIA GB200 NVL72

Datasheet | Learn More

Datasheet

NVIDIA

NVIDIA Blackwell

The engine of the new industrial revolution.

Breaking Barriers in Accelerated Computing

The NVIDIA Blackwell architecture introduces groundbreaking advancements for generative AI and accelerated computing. The incorporation of the second-generation Transformer Engine, alongside the faster and wider NVIDIA NVLink™ interconnect, propels the data center into a new era, with orders of magnitude more performance compared to the previous architecture generation. Further advances in NVIDIA Confidential Computing technology raise the level of security for real-time LLM inference at scale without performance compromise. And Blackwell's new decompression engine combined with Spark RAPIDS™ libraries deliver unparalleled database performance to fuel data analytics applications. Blackwell's multiple advancements build upon generations of accelerated computing technologies to define the next chapter of generative AI with unparalleled performance, efficiency, and scale.

Key Offerings

- NVIDIA GB200 NVL72
- NVIDIA HGX B200

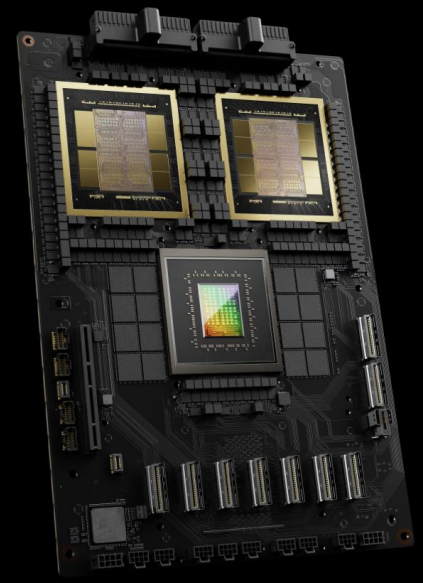
NVIDIA GB200 NVL72

Powering the New Era of Computing

NVIDIA

NVIDIA Grace Blackwell GB200 Superchip Architecture

The NVIDIA GB200 NVL72 connects 36 GB200 Grace Blackwell Superchips with 36 Grace CPUs and 72 Blackwell GPUs in a rack-scale design. The GB200 NVL72 is a liquid-cooled solution with a 72-GPU NVLink domain that acts as a single massive GPU—delivering 30X faster real-time inference for trillion-parameter large language models.



Redpaper:

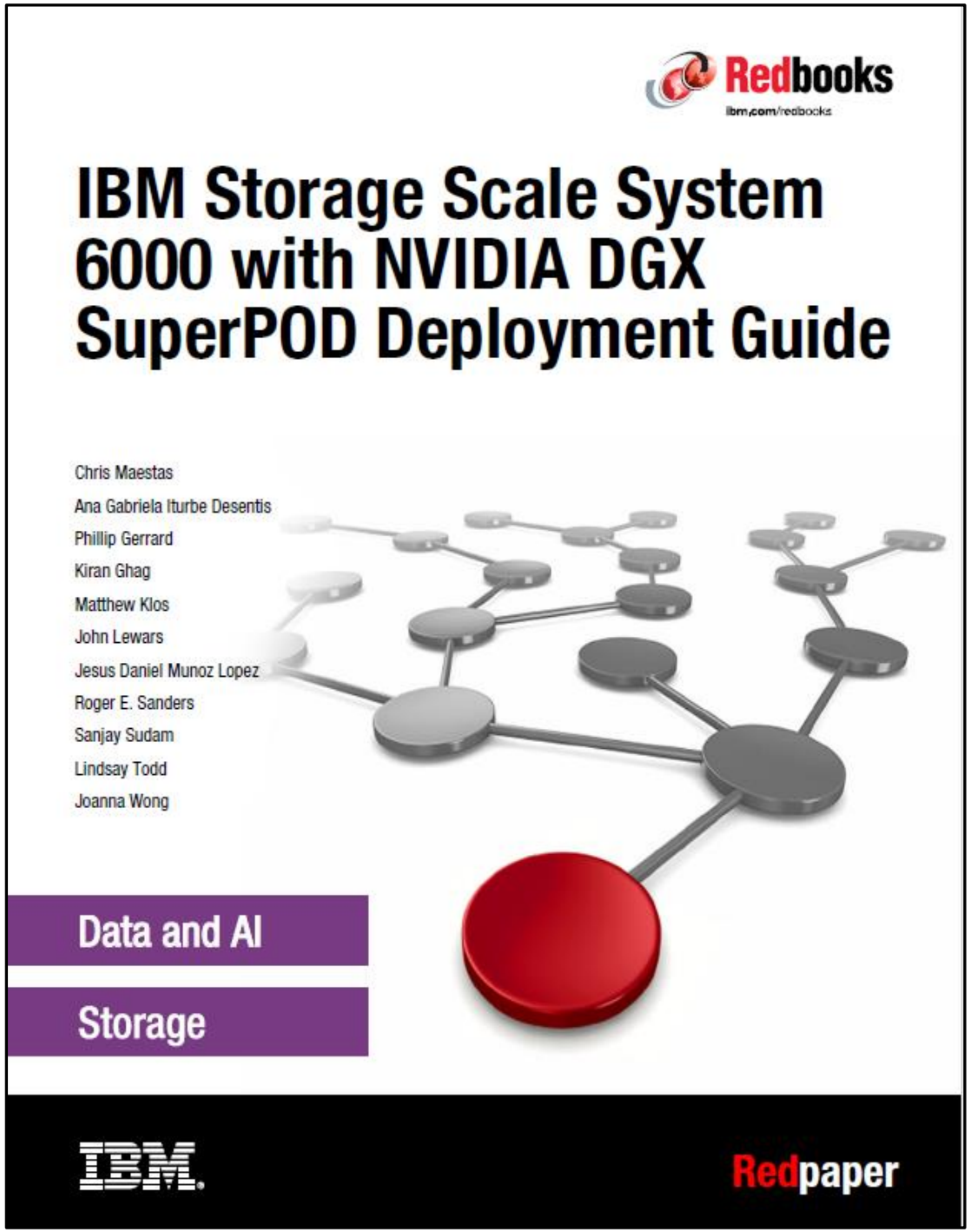
IBM Storage Scale System 6000 with NVIDIA DGX SuperPOD **Deployment Guide**

IBM Redbook REDP-5746-00

Authors: Chris Maestas, Ana Gabriela Iturbe Desentis, Phillip Gerrard, Kiran Ghag, Nikhil Khandelwal, Matthew Klos, John Lewars, Jesus Daniel Munoz Lopez, Roger E. Sanders, Sanjay Sudam, Lindsay Todd and Joanna Wong

https://www.redbooks.ibm.com/redpapers/pdfs/redp5746.pdf

Table of Contents:



CoreWeave Partners with **IBM** to Deliver New AI Supercomputer for IBM **Granite Models**

YORKTOWN HEIGHTS, N.Y. – **January 15, 2025** – CoreWeave, the AI Hyperscaler, today announced its plans to deliver one of the first NVIDIA GB200 Grace Blackwell Superchip-enabled AI supercomputers to IBM, equipped with **NVIDIA GB200 NVL72** systems, interconnected with NVIDIA Quantum-2 InfiniBand networking. IBM will use CoreWeave's highly performant, reliable, and resilient cloud platform to train the next generations of its Granite models, IBM's open source, enterprise-ready series of AI models that deliver state-of-the-art performance relative to model size while maximizing safety, speed, and cost-efficiency for enterprise use cases.

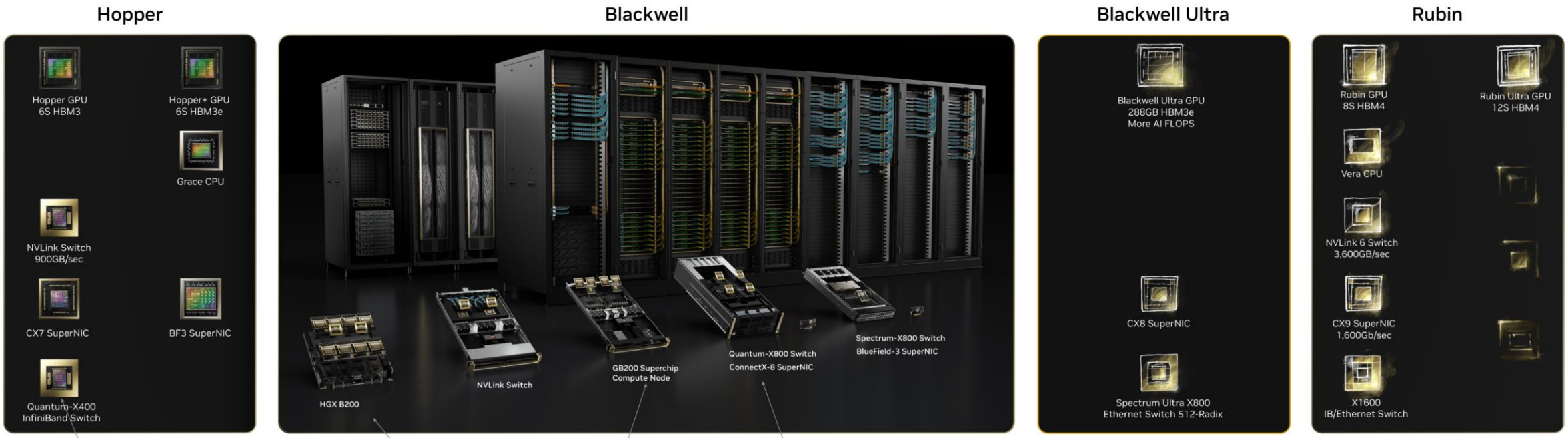
The supercomputer will leverage **IBM Storage Scale System 6000**, which is combined with NVMe flash technology to deliver high-performance storage for AI, data analytics, and other demanding workloads. As part of this agreement, CoreWeave customers can access the IBM Storage platform within CoreWeave's dedicated environments and AI cloud platform.

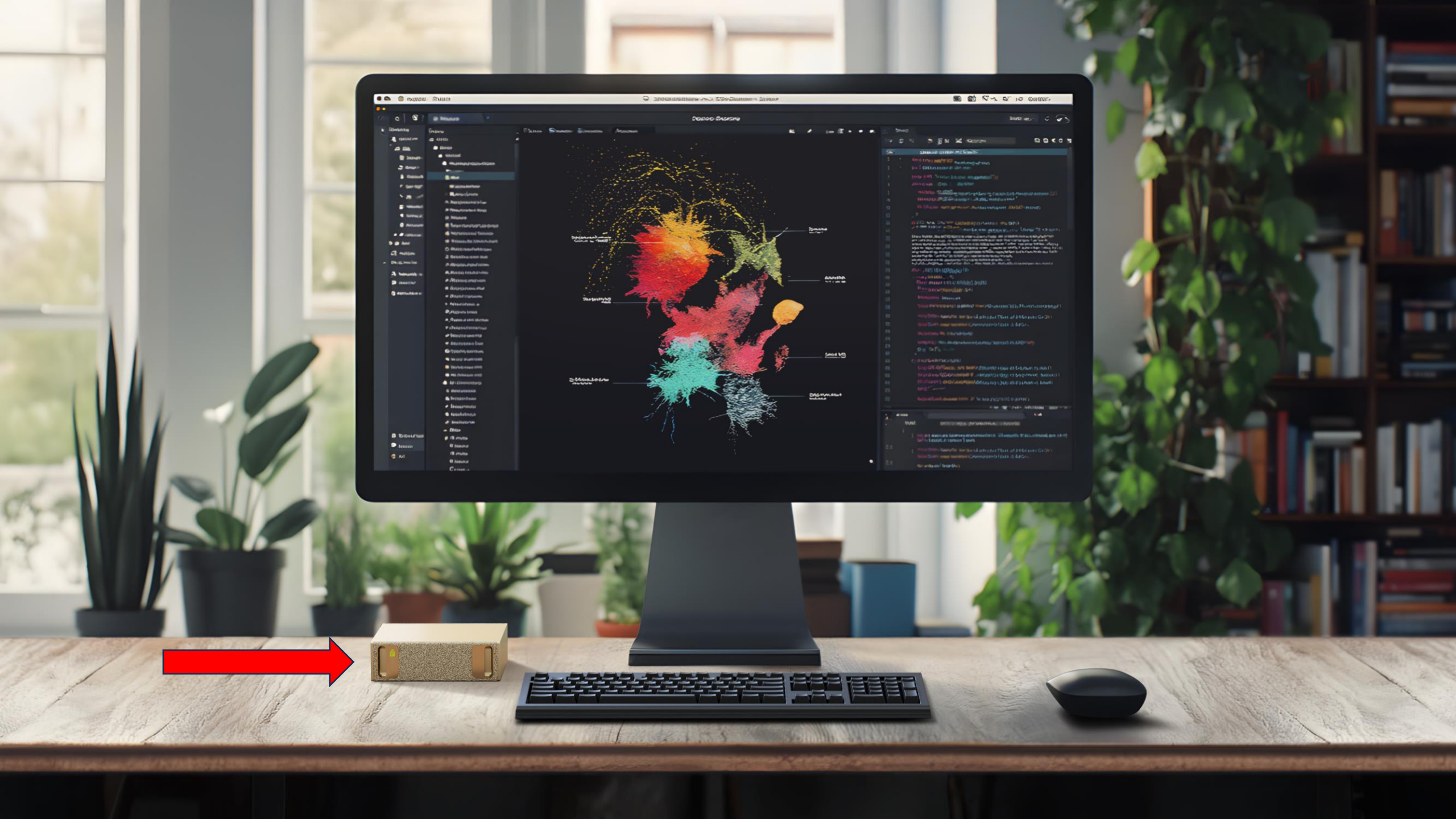
https://newsroom.ibm.com/2025-01-15-coreweave-partners-with-ibm-to-deliver-new-ai-supercomputer-for-ibm-granite-models
https://www.coreweave.com/blog/coreweave-becomes-the-first-cloud-provider-with-generally-available-nvidia-gb200-nvl72-instances

IBM Storage Scale System 6000
GPUDirect enabled NVMe Storage

One-Year Rhythm | Supercluster Scale | Full Stack | CUDA Everywhere

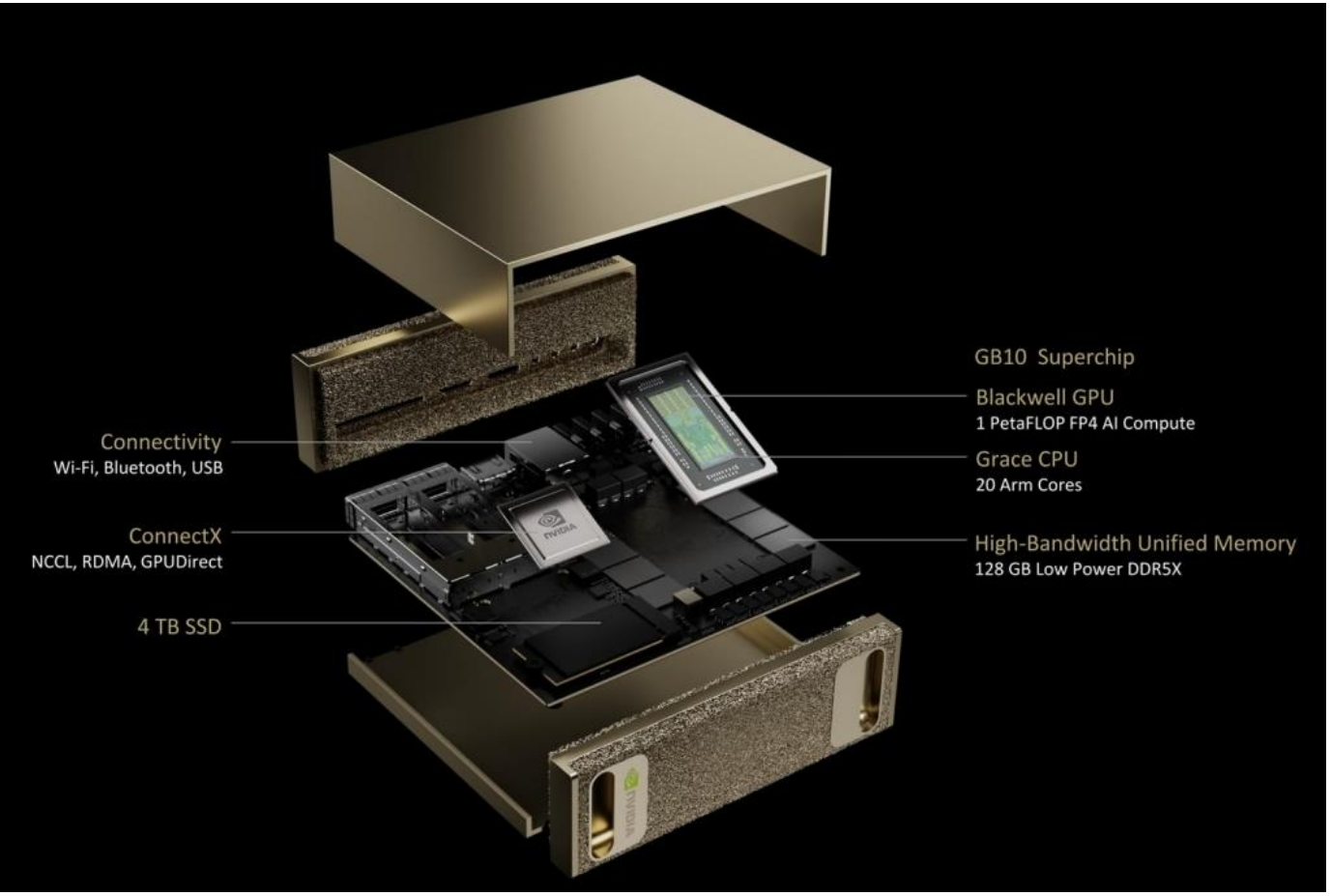
GPU Generations 2022-2027





Project DIGITS - NVIDIA GB10 Grace Blackwell Superchip

[Project DIGITS will be available in May 2025 from NVIDIA and top partners, starting at $3,000.]

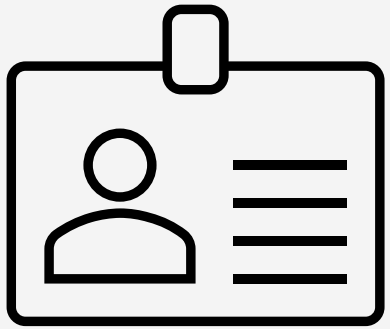


System on Chip	NVIDIA GB10 Grace Blackwell Superchip
GPU	NVIDIA Blackwell
CPU	20-core Arm®
AI Performance	1 PFLOP FP4
Memory	128 GB LPDDR5x coherent unified system memory
Connectivity	4x USB4 Type C, Bluetooth
Networking	ConnectX, Wi-Fi, Ethernet
Storage	Up to 4 TB NVMe M.2 with self-encryption
Display Output	1x HDMI 2.1a
Power	Standard wall outlet
OS	NVIDIA DGX™ Base OS, Ubuntu Linux

*Preliminary specifications, subject to change

https://nvidianews.nvidia.com/news/nvidia-puts-grace-blackwell-on-every-desk-and-at-every-ai-developers-fingertips

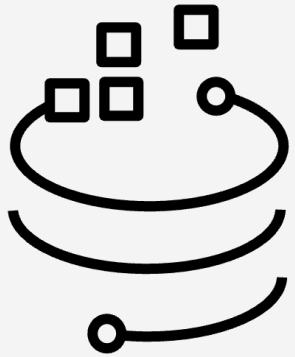
Thank you



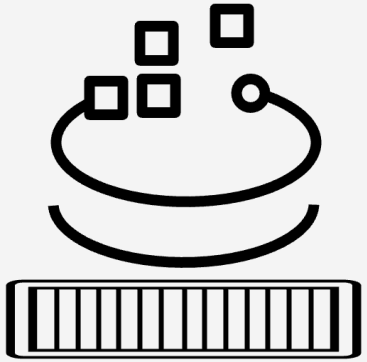
Dr. Thomas Schoenemeyer (NVIDIA)
mailto:thomass@nvidia.com

Frank Kraemer (IBM)
mailto:kraemerf@de.ibm.com

Thank you for using



Storage Scale



Storage Scale System