AI Factories at Scale: Advancing Open Infrastructure for Agentic AI

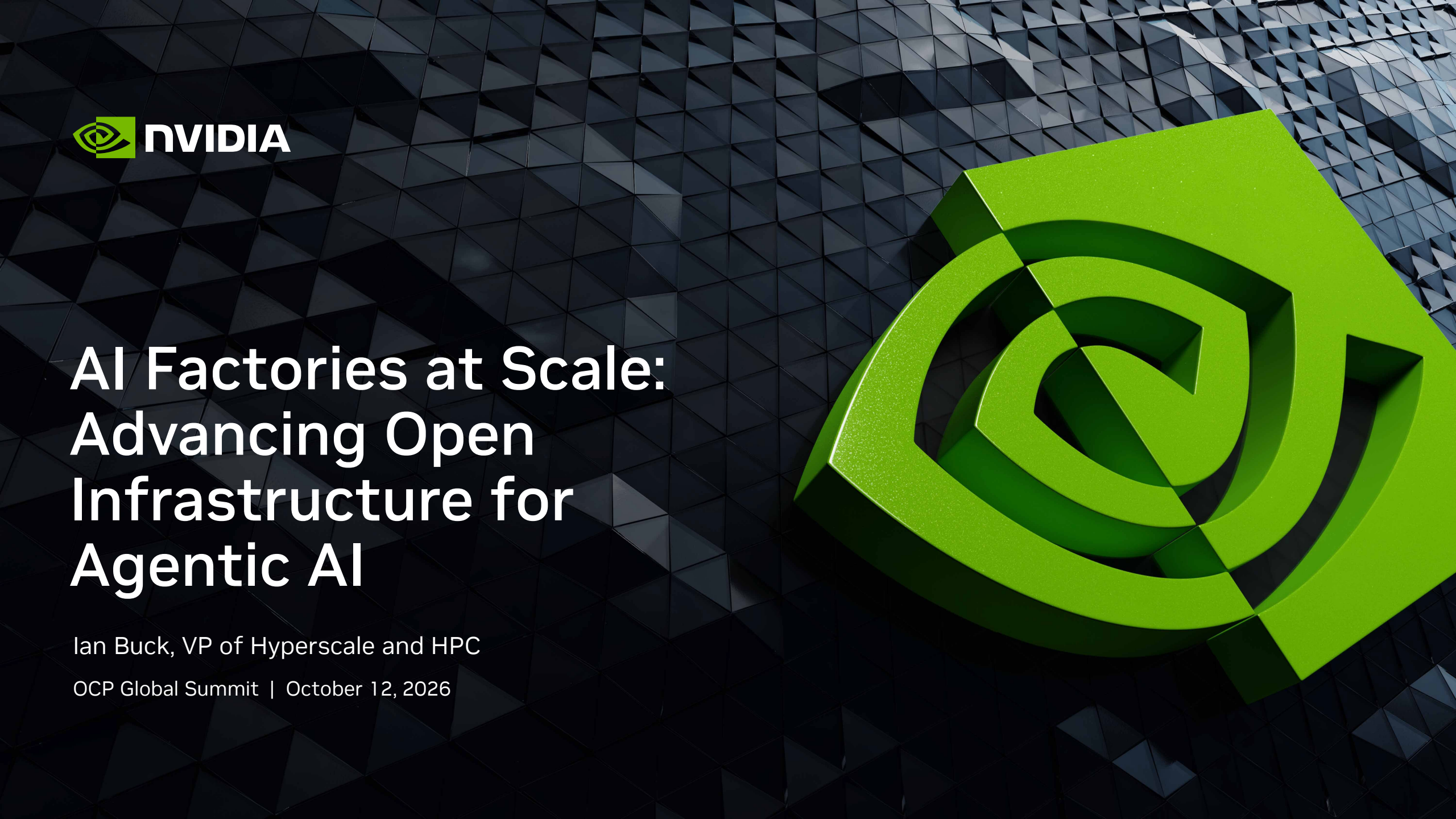
Ian Buck

Vice President of Hyperscale and HPC Computing, NVIDIA



OCP GLOBAL SUMMIT

October 12-15, 2026
San Jose, California



AI Factories at Scale: Advancing Open Infrastructure for Agentic AI

Ian Buck, VP of Hyperscale and HPC

OCP Global Summit | October 12, 2026

History of OCP Contributions

2013



SwitchX-2

2014



ConnectX-3

2015



Tesla M40



ConnectX-4

2017



Tesla P100



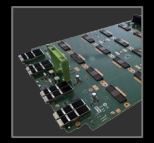
Tesla P100 HGX-1 baseboard

2019



V100 OAM form factor

2020



A100 OAM and baseboard v1.0

2022



HGX-H100 baseboard

2023



ConnectX-7 adapter

2024



MGX 1.0 Blackwell rack

2025



MGX 1.1 Blackwell Ultra

Participating Projects

- Open Platform Firmware
- Digital Twin
- Open Network Install Environment
- NIC
- Open Data Centers / Systems for AI
- Power Distribution
- Cooling Environment
- Server
- Hardware Management
- Open Accelerator
- Data Center Facility
- Security
- Switch Abstraction Interface
- IT Infrastructure
- Time Appliance Project
- MHS
- S/W for Open Network in the Cloud

NVIDIA

NVIDIA AI Factories Are Productive, Durable, and Fungible

Built on OCP open standards, every model, every workload, every partner with appreciating value

Rich Ecosystem

10M Developers | CSP, NCP, OEM, ODM, 300+ Supply Chain

Fungible

Every Model, Every Workload | Open, Closed, and Frontier

Pre-Training | Post-Training | Inference | Agentic | Accelerated Computing

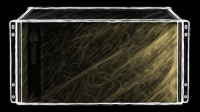
Year 1 — CUDA | Faster Every Release — Year 10

Durable



V100/DGX-1

10 yrs in service



A100

6 yrs old

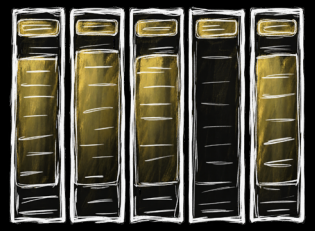
9 yrs sold



NVL72

1 yr old

7 yrs sold



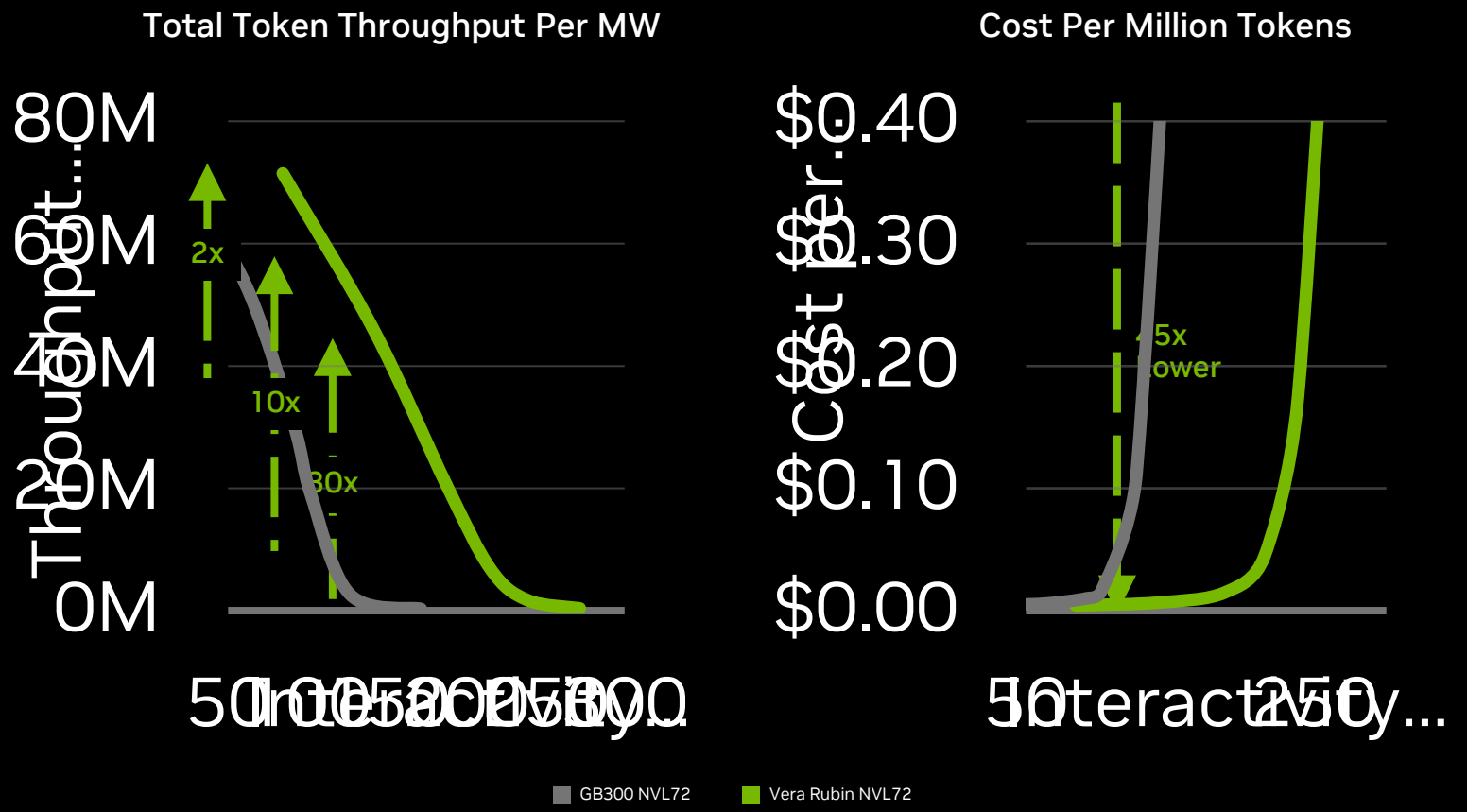
Vera Rubin

Now in full production

NVIDIA

Full-Stack Co-design Delivers up to 30x Better AI Factory Throughput on Agentic Workloads

DeepSeek-V4-Pro | SemiAnalysis AgentX





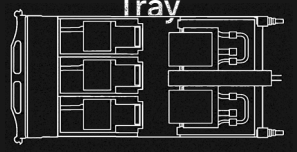
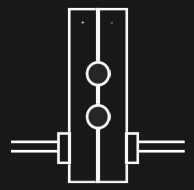
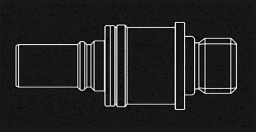
Vera Rubin NVL72 MGX Rack

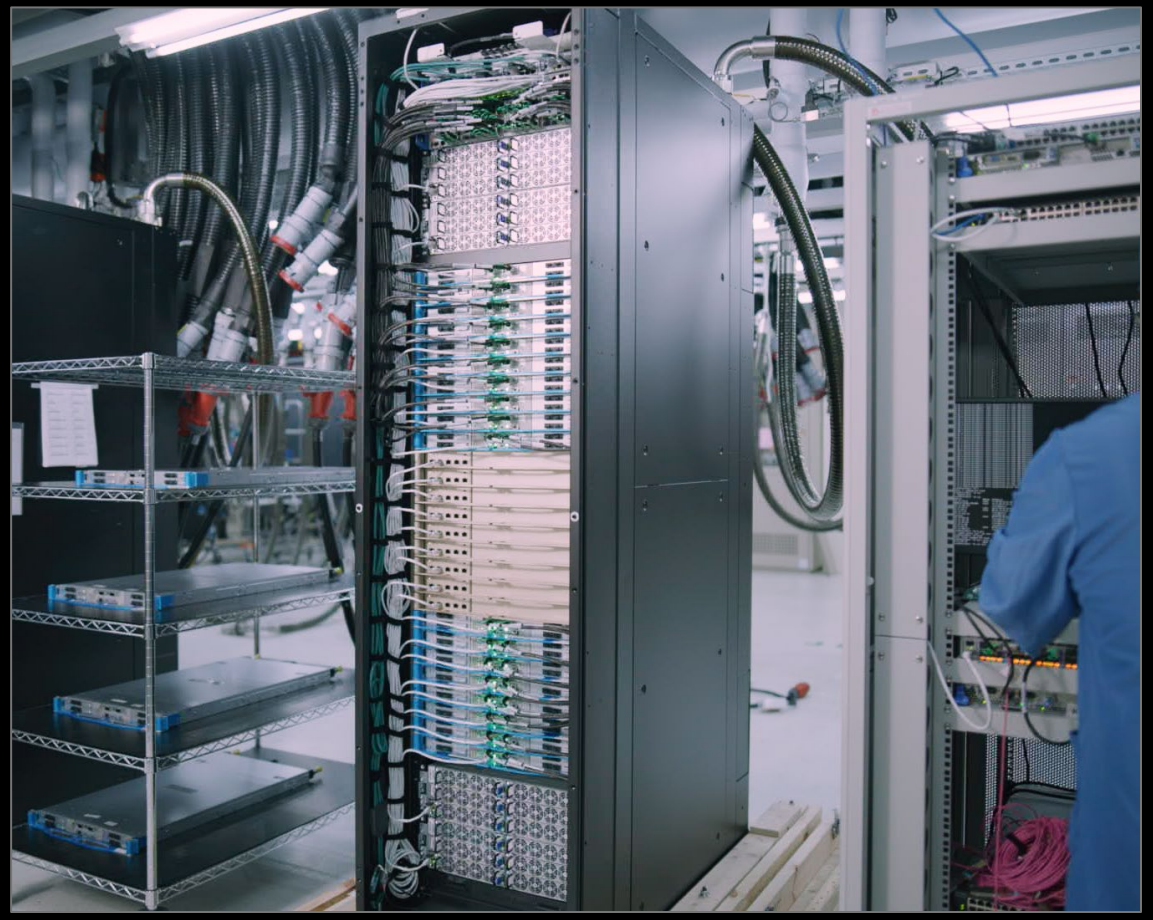
- Disaggregated serving
- Rate matching
- Large-scale expert parallel
- Distributed KV-caching
- KV-cache offloading
- KV-aware routing
- Fused CUDA kernels

NVIDIA

Announcing NVIDIA Contributes 3rd Generation MGX Specification to OCP

MGX 1.2 is a factory-scale, high-density, liquid cooled architecture designed for easy assembly and maintenance

Cableless, Hoseless Compute Tray  100% liquid cooled, faster assembly	**Liquid-Cooled Busbar**  5,000 amp capacity	**In-tray MQD/MQDB**  Easier tray servicing
Copper Spine 4 Cable Cartridges for NVLink or Spectrum-X Ethernet	**Blind Mate Manifold** 45C Liquid Cooling	**Latch Ejector Bar** Support 6,000 lb. mating force
Floating Blind Mate Reduces breakage	**Blind Mate Slide Rails** NVLink and LC connections	**Larger Rack Cable Tray** Fits higher cable density
Rack Reinforcements +100 lb. of steel support	**Rear Rack Extenders** Protects manifold fittings	**Compute Tray Bay System** Highly configurable
Switch Tray Form Factor 1U LC: NVLink or Spectrum-X	**Compute Tray DC-SCM** Fits 1RU form factor	**Liquid-Cooling Flow Rates** UQD08

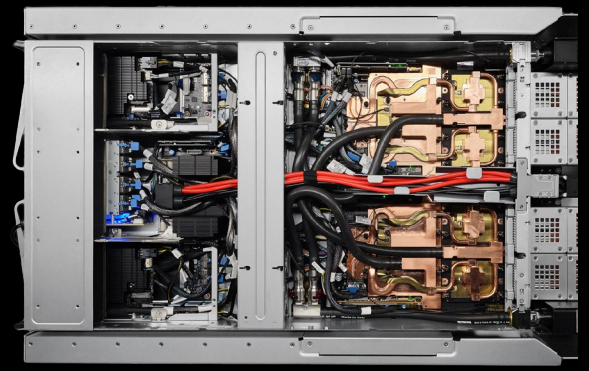


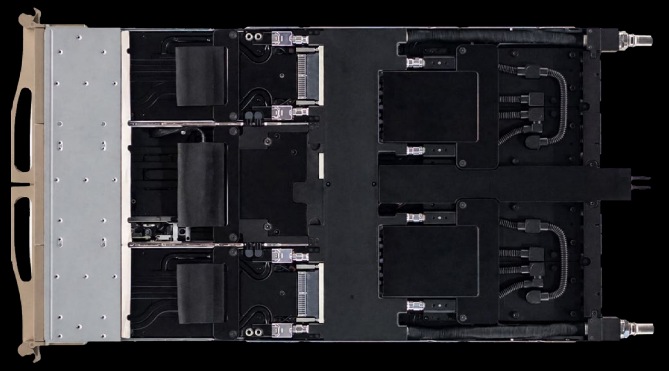
NVIDIA

3rd Gen MGX Improves Manufacturability and Maintenance

Vera Rubin: No cables, fans, or hoses

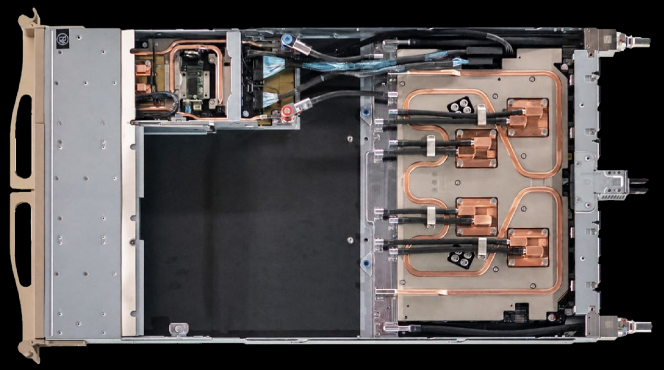
Compute Tray





Switch Tray





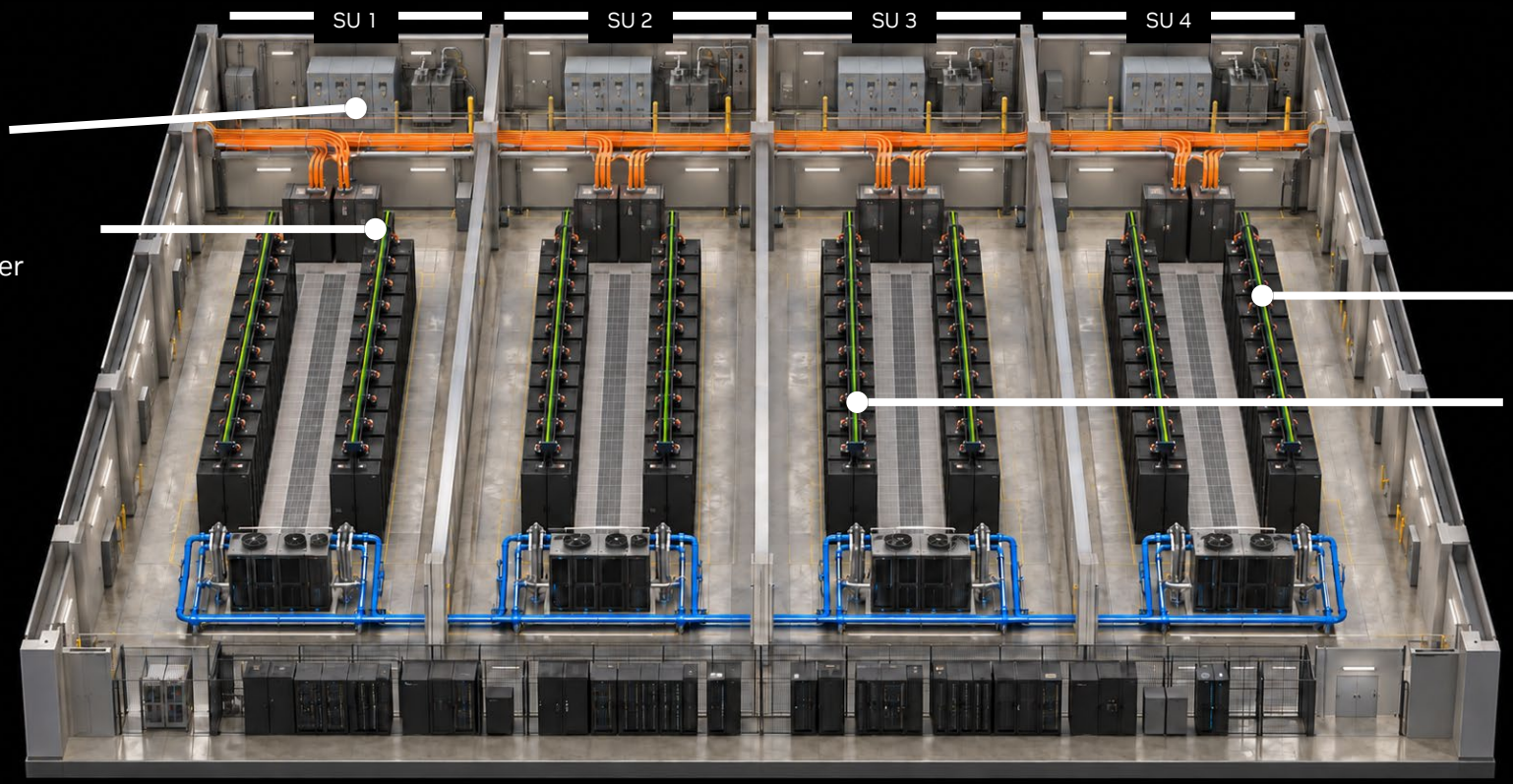
GB200 NVL72 (1st Gen OCP MGX 1.0)

Vera Rubin NVL72 (3rd Gen OCP MGX 1.2)

NVIDIA

Visionbay.ai Introduces the First 800 VDC AI Factory Engineered for the Next Decade of Accelerated Computing

Scalable power design for high-density, multi-gen AI infrastructure, aligned with OCP 800 VDC architecture

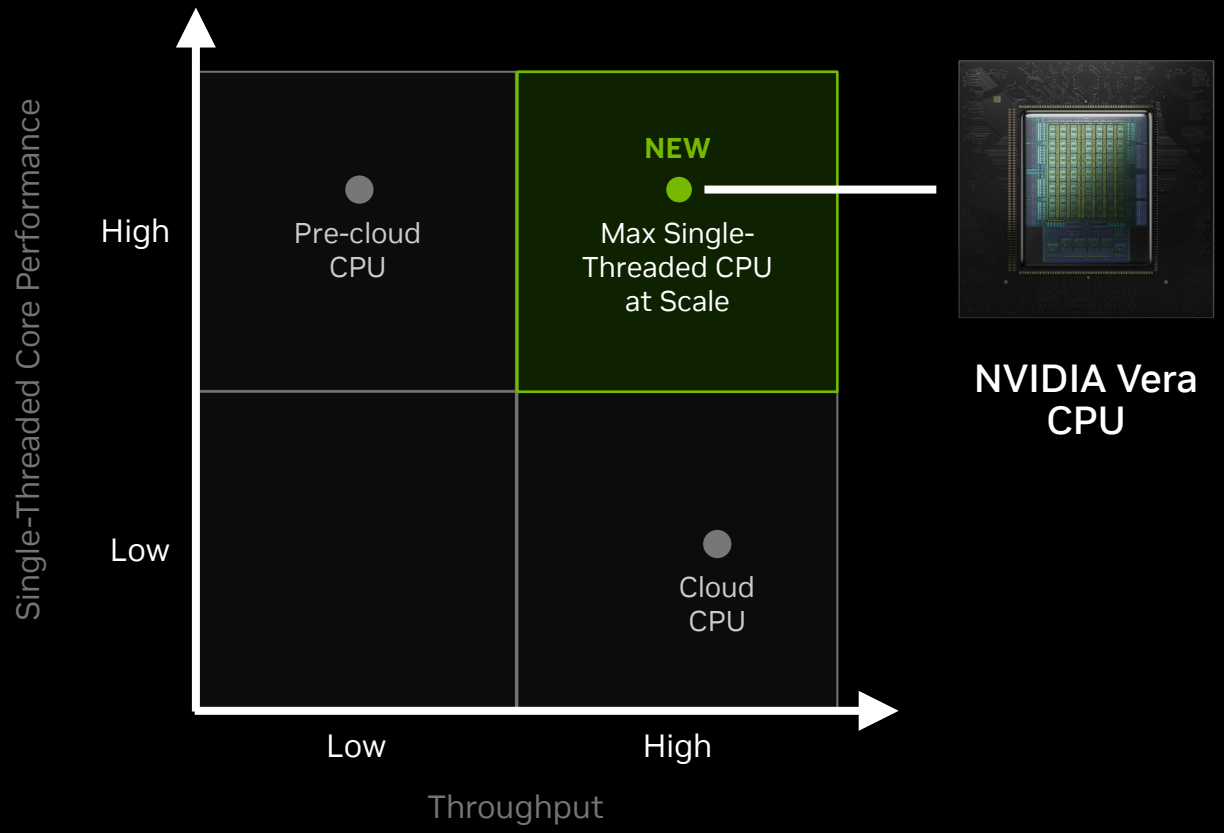


16 MW Deployment Unit

NVIDIA

NVIDIA Vera CPU for Agents

The max single-threaded CPU at scale



NVIDIA Vera CPU

2x

Performance

NVIDIA Olympus Core

3.4 TB/s

Core-to-core bandwidth

Zero Chiplet Tax
Scalable Coherent Fabric

1.2 TB/s

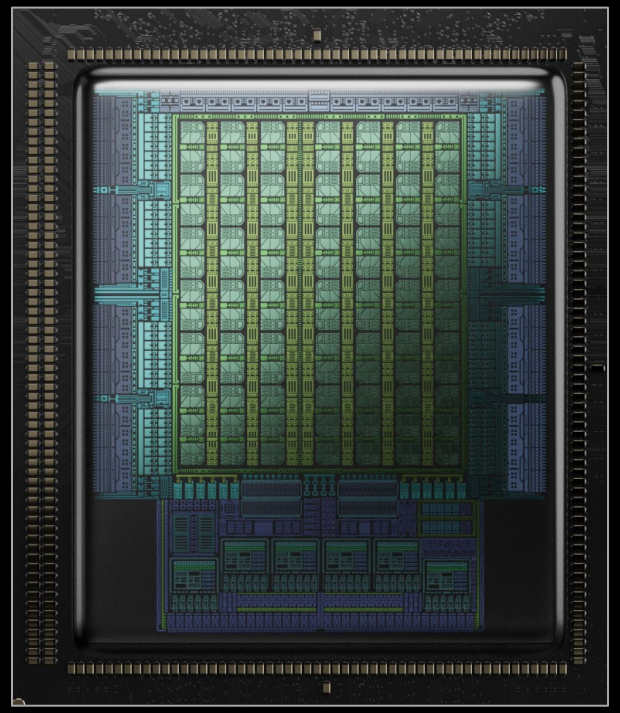
Memory bandwidth per core

LPDDR5X subsystem

NVIDIA

NVIDIA Vera Delivers 2x Performance for Agentic AI

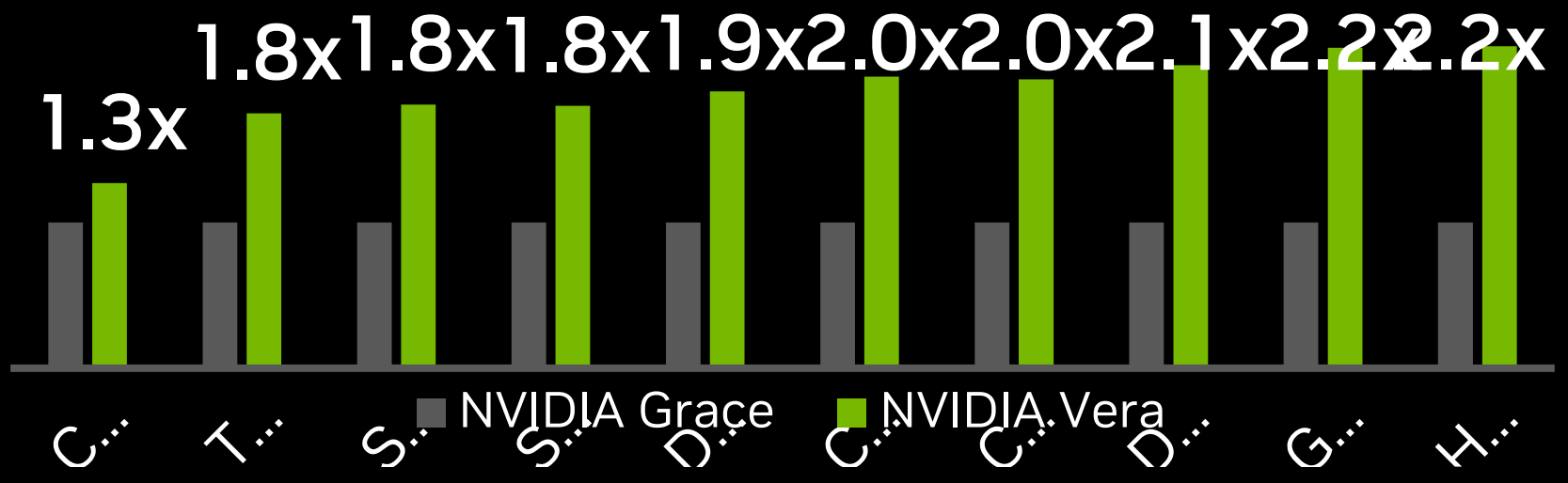
Built for sustained AI execution at scale



2x
Hyperscale Cloud
Meta DCPerf

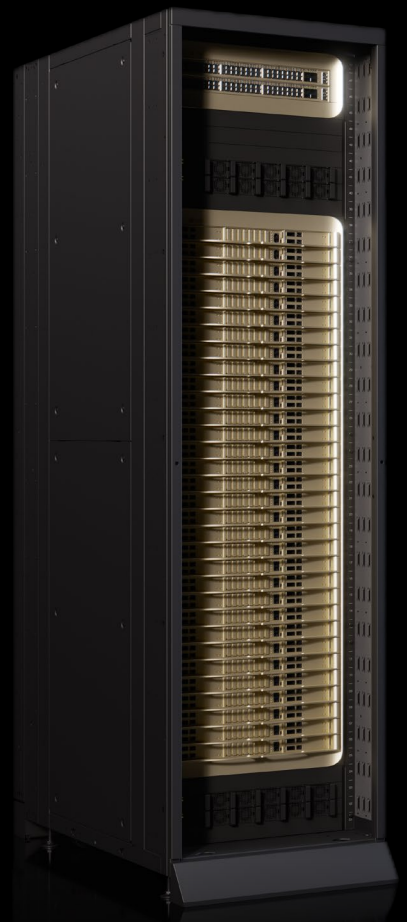
2x
HPC Golden Suite
TACC

3.5x
Apache Kafka
Redpanda



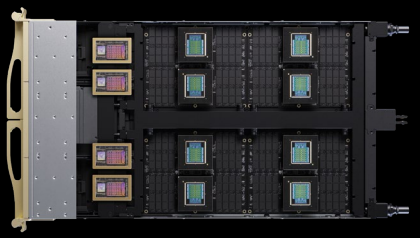
Announcing NVIDIA Vera CPU MGX Rack Now Available

Partner ecosystem building customizable Vera OCP MGX racks



Vera MGX Rack
Shipping now

Cableless, Modular, 100% Liquid Cooled



Faster assembly and easier tray servicing

Customizable



Customizable across inference, agentic, storage, and more

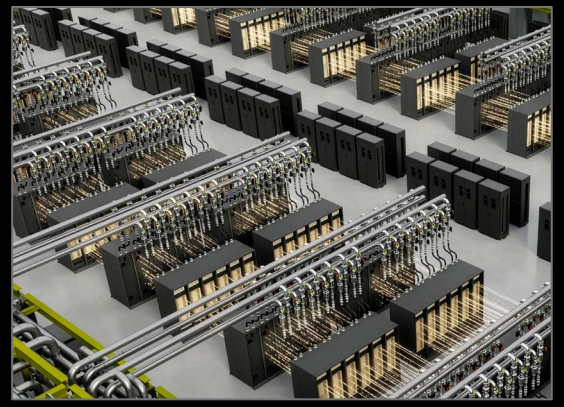
Bull · DELL Technologies · HPE · Lenovo · Supermicro
AIVRES · ASRock Rack · ASUS · COMPAL · Foxconn · GIGABYTE · hyve solutions
Inventec · MiTAC Computing · msi · PEGATRON · QCT · wistron · wiwynn

NVIDIA

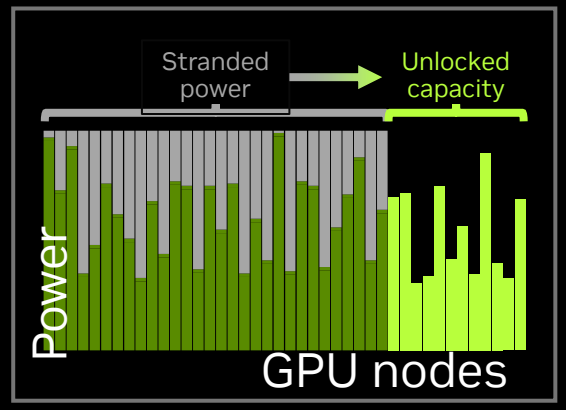
NVIDIA DSX Advances OCP-Compatible AI Factories

Grid resilience, more compute per megawatt, and simpler AI factory operations

New DSX-compatible Spec Contributions
Open data centers & grid resilience

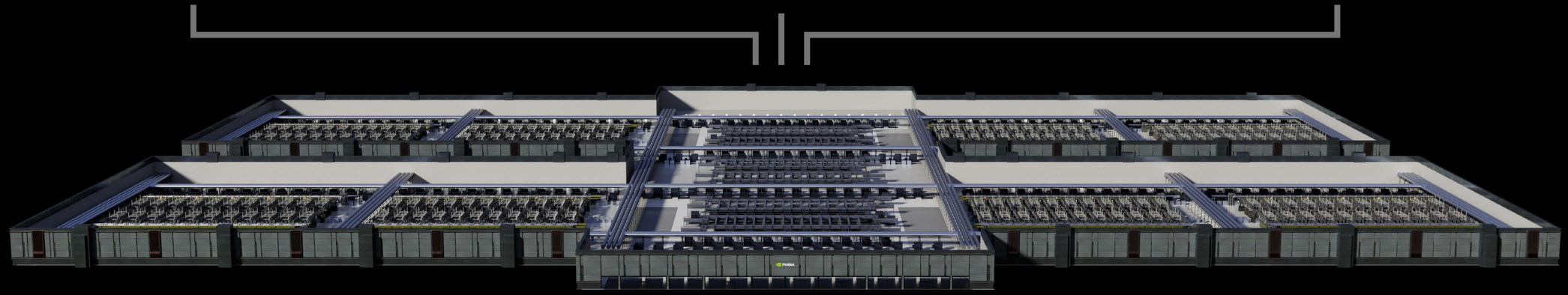


MaxLPS – 40% More GPUs in Same Power
Turns stranded power to GPU capacity



New DSX AI Factory Open Software
Easier factory configuration & operations

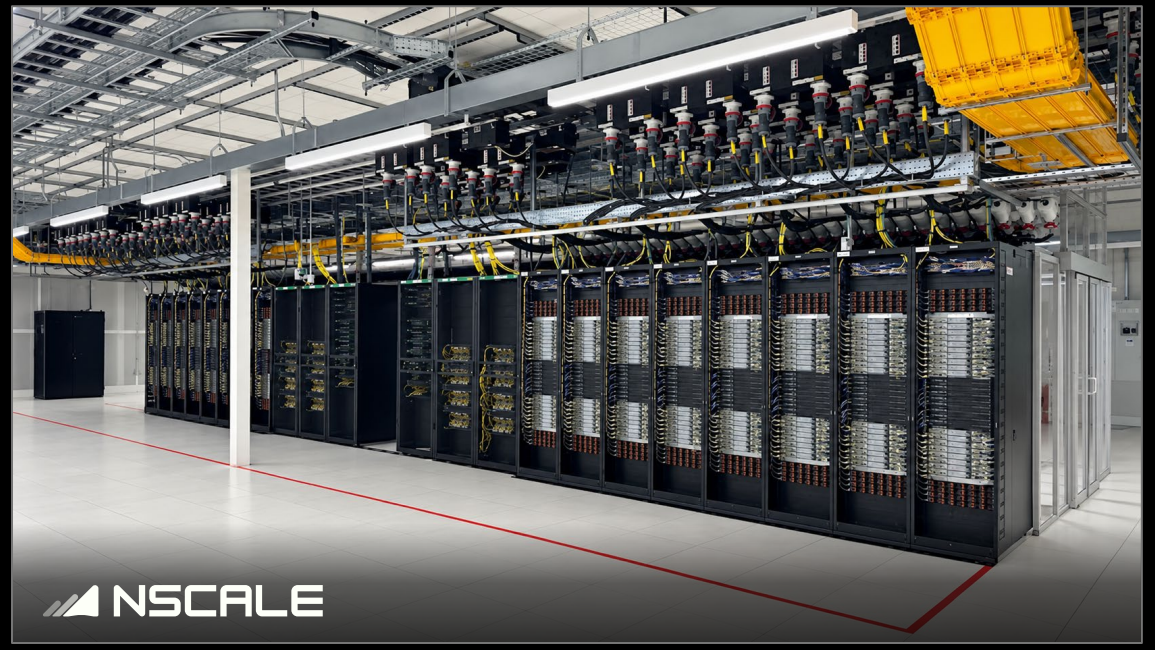
- NVIDIA Rack Management Service
- NVIDIA Cluster Readiness Engine
- NVIDIA Config Manager



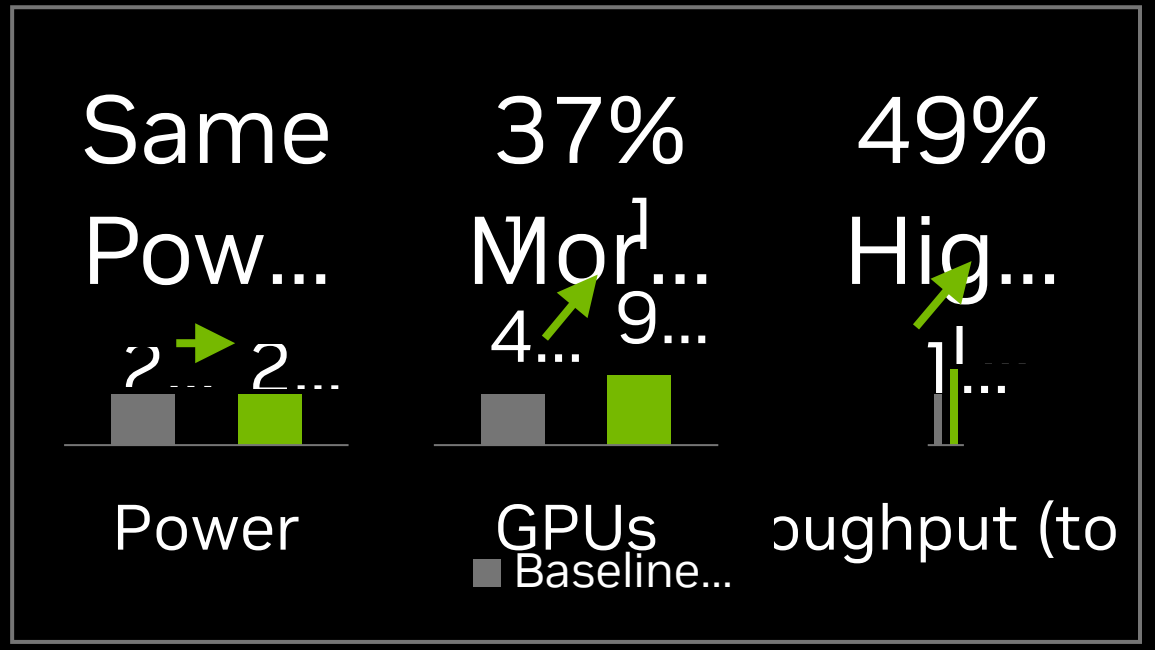
NVIDIA

Nscale Increases Throughput Nearly 50% with MaxLPS

37% more GB300 GPUs deployed within the same provisioned power budget (Kimi K2.5, 8K/1K, Iceland)



Nscale GB300 Cluster



Nearly 50% higher throughput in same power

NVIDIA Spectrum-X Photonics Multiplane Ethernet Uninterrupted AI Connectivity

Spectrum-X Multiplane Photonics enables AI factories to scale up to 512K GPUs with a flat, two-tier network





ConnectX-9 SuperNICs



BlueField-4 DPUs

NVIDIA Spectrum-6
In deployment now

2.6 ms
Fault detection

40%
Fewer switches

5x
Better power efficiency

10x
Higher MTBI

Meta Microsoft ORACLE CLOUD Infrastructure

NVIDIA

Announcing Spectrum-6 Ethernet Switch for Meta's AI Factory

Open Networking for AI with FBOSS and OCP SAI

Spectrum-X Multiplane collaboration for higher resiliency and smooth deployment at scale

World's fastest fault detection, 30x faster than traditional software-based fault detection

NVIDIA open Ethernet silicon with OCP stack

Meta



Meta 102.4T FBOSS Switch (N-IceCube)

Based on NVIDIA Spectrum-6

NVIDIA

Announcing Google Falcon Transport on NVIDIA Networking

Silicon-to-software partnership for consistent performance

Extends NVIDIA networking to Google data center transport

Supported across NVIDIA Spectrum-X SuperNICs and NVIDIA BlueField DPUs

Predictable performance at any fabric scale



Multipath Load Balancing

Spreads traffic, reduces bottlenecks



Programmable Congestion Control

Adapts traffic to reduce delays



Hardware Loss Recovery

Keeps data moving



Per-Connection Security

Protects data in transit

NVIDIA

OPEN Compute Project

NVIDIA

Spectrum-X Ethernet Extends Open Networking Leadership

One accelerated Ethernet platform supporting multiple AI transports

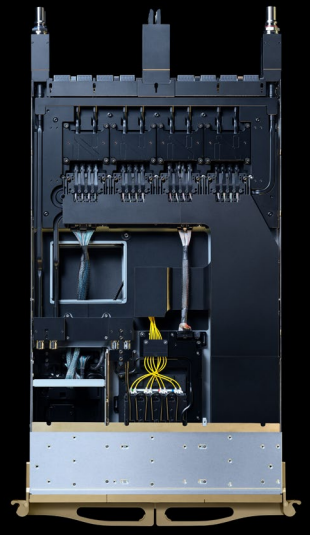
	Meta	OpenAI	NVIDIA
	Meta RoCE	MRC	Spectrum-X Adaptive RDMA
Falcon	RoCE networking for Meta's AI infrastructure	Multipath delivery for resilient AI communication	Adaptive routing for efficient AI data movement
Interoperability with Google's data center transport			

NVL576 Is Built for Breakthrough Scale and Efficiency

Designed for performance per watt for 10T+ parameter models



Vera Rubin Ultra Compute Tray



NVLink 7 CPO Switch Tray



Vera Rubin Ultra NVL576 CPO

Join Us at OCP

OCP Talks—33 Total Sessions

Session	Date, Time, and Location
AI Tokenomics Explained Shruti Koparkar and Eduardo Alvarez, NVIDIA	Tuesday, Oct 13, 2:30 p.m. | Concourse Level, 220B
800 VDC AI Factories: From Arch. Alignment to Indust. Deployment Mike Tu and Jared Huntington, NVIDIA	Tuesday, Oct 13, 4:10 p.m. | Concourse Level, 210CG
Flexibility the Grid Can Count On: Turn Curtailable Load to Capacity Sean James, NVIDIA	Tuesday, Oct 13, 4:35 p.m. | Concourse Level, 210CG
Revolutionizing Networking for Gigawatt AI Factories Gilad Shainer, NVIDIA	Wednesday, Oct 14, 9:15 a.m. | Concourse Level, 210CG
Next-Gen ML Infrastructure: The Open DC Architecture Ayan Majumdar and Heather Huang, Google | Mike Sabotta, NVIDIA | Paul Artman, AMD | Mike Edie, Meta	Thursday, Oct 15, 12:30 p.m. | Concourse Level, 210CG
The AI Factory Digital Twin: Full Stack Tokens/Watt Optimization Kourosh Nemati, Tony Marfil and Usha Kodali, NVIDIA | Sherman Ikemoto and Matthew Käufeler, Cadence	Thursday, Oct 15, 3:35 p.m. | Concourse Level, 210CG



MGX Ecosystem

Nearly 100 partners on display
800 VDC power technology
Next generation innovations


NVIDIA