GTC 2025 直送！AIコンピューティング最新情報

2025/4/9
佐々木 邦暢 (@_ksasaki)
エヌビディア合同会社

NVIDIA
GTC
What's Next in AI Starts Here
SAN JOSE McENERY CONVENTION CENTER

NVIDIA AI & HPC プラットフォーム

NIM
CUDA-Accelerated
Agentic AI Libraries

Omniverse
CUDA-Accelerated
Physical AI Libraries

CUDA-X Libraries

CUDA • DOCA • NCCL
Cluster-Scale Software
System Software
Chip Software

Accelerated
Software Stack

GB200 NVL72 SuperPOD

Grace Blackwell
MGX Node

NVLink Switch

Quantum Switch

Spectrum-X Switch

Chips Purpose-Built for AI Supercomputing
GPU | CPU | DPU | NIC | NVLink Switch | IB Switch | Enet Switch

Blackwell があらゆる場所に

aws CoreWeave Google Cloud Meta Microsoft Azure NEBIUS OpenAI ORACLE CLOUD Infrastructure xAI

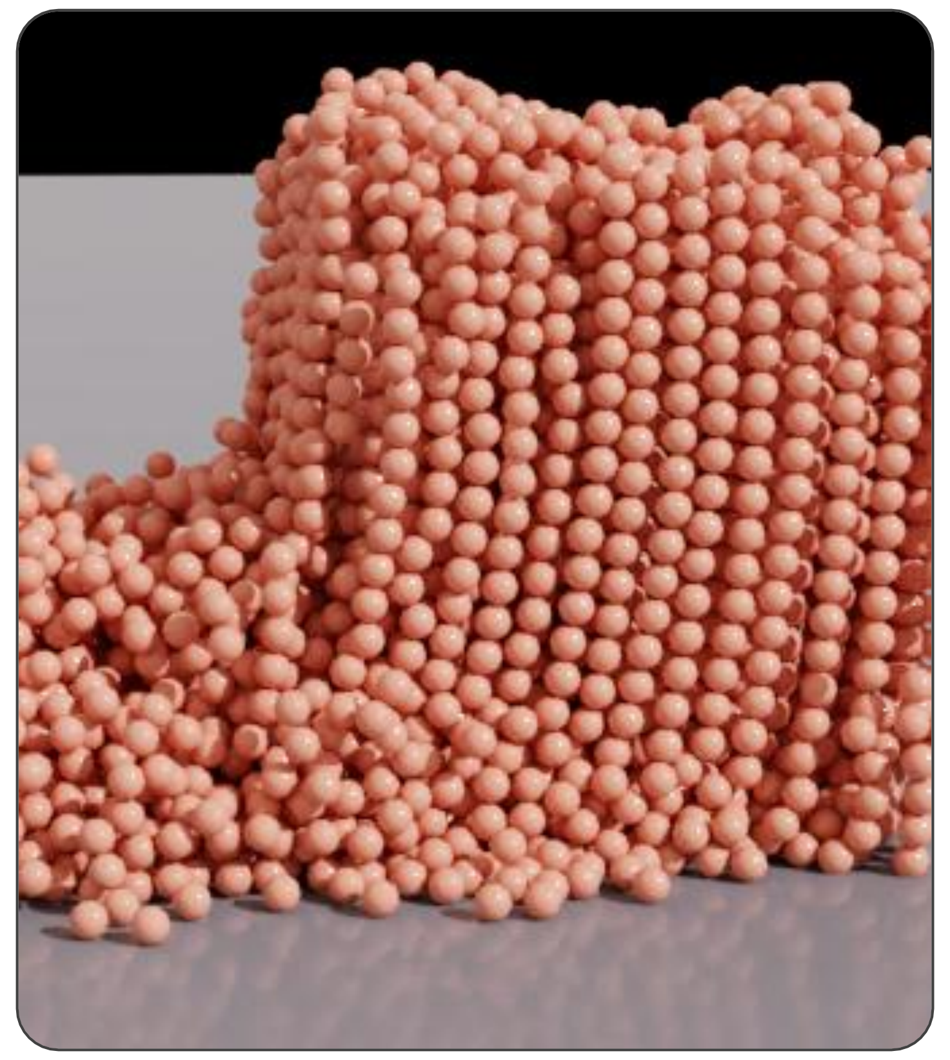
DELL Technologies Hewlett Packard Enterprise Lenovo SUPERMICRO AIVRES ASUS FOXCONN HON HAI TECHNOLOGY GROUP GIGABYTE Inventec PEGATRON QCT wistron wiwynn

Fastest Ramping Product in NVIDIA History

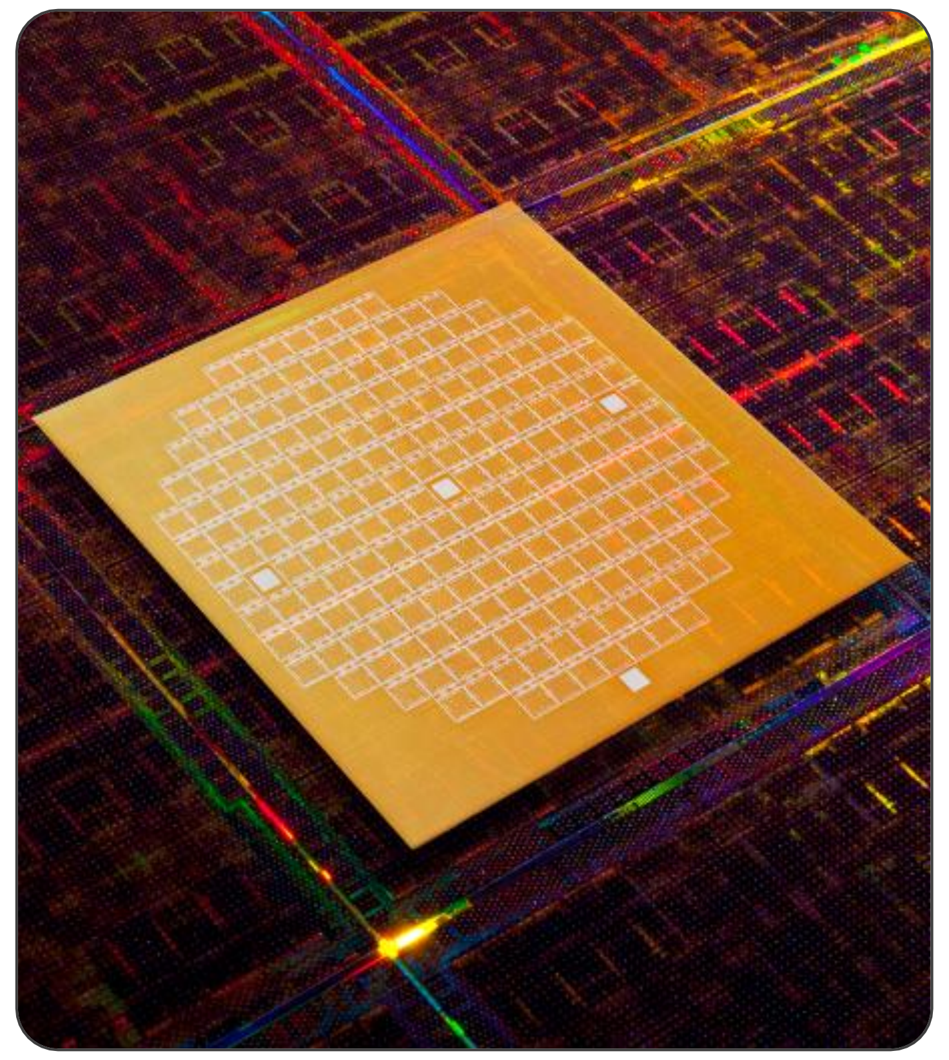
CUDA-X が多様なアプリケーションを高速化

cuDSS
CAE

PhysicsNeMo
AI Physics

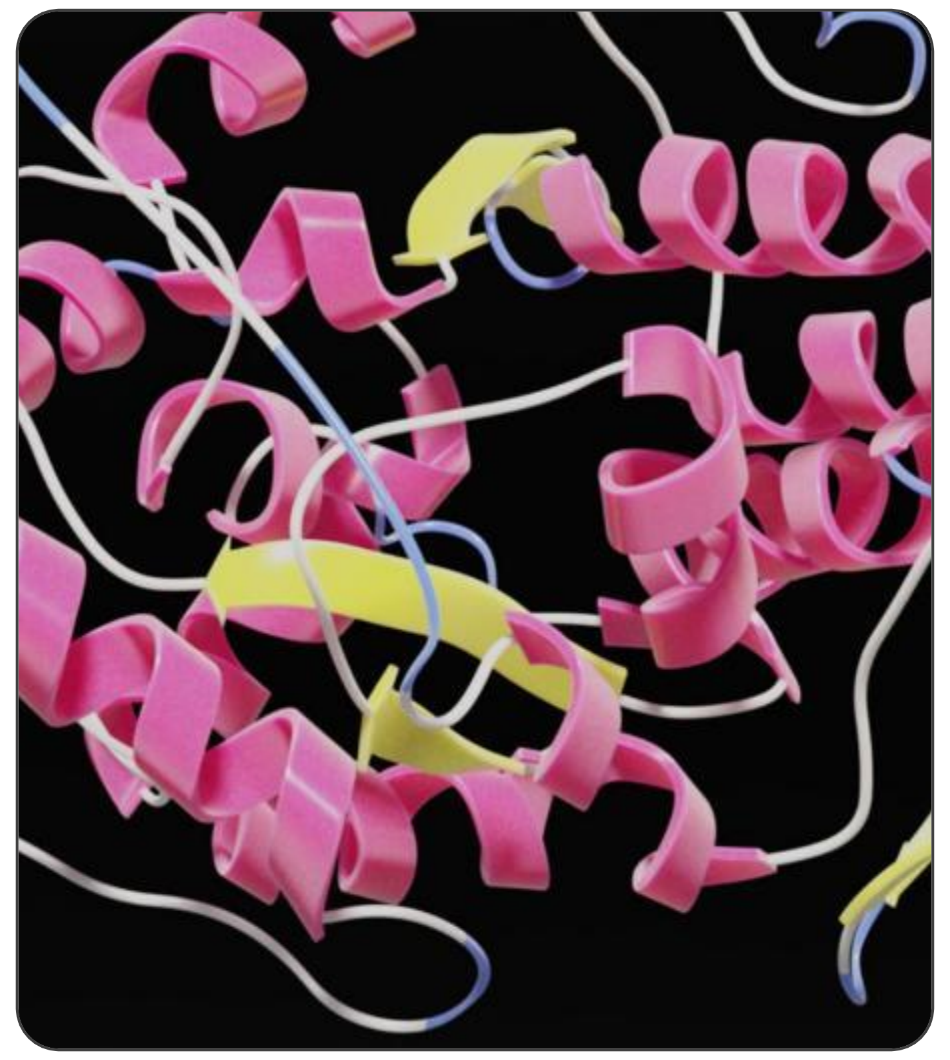


Warp
Physical Simulation

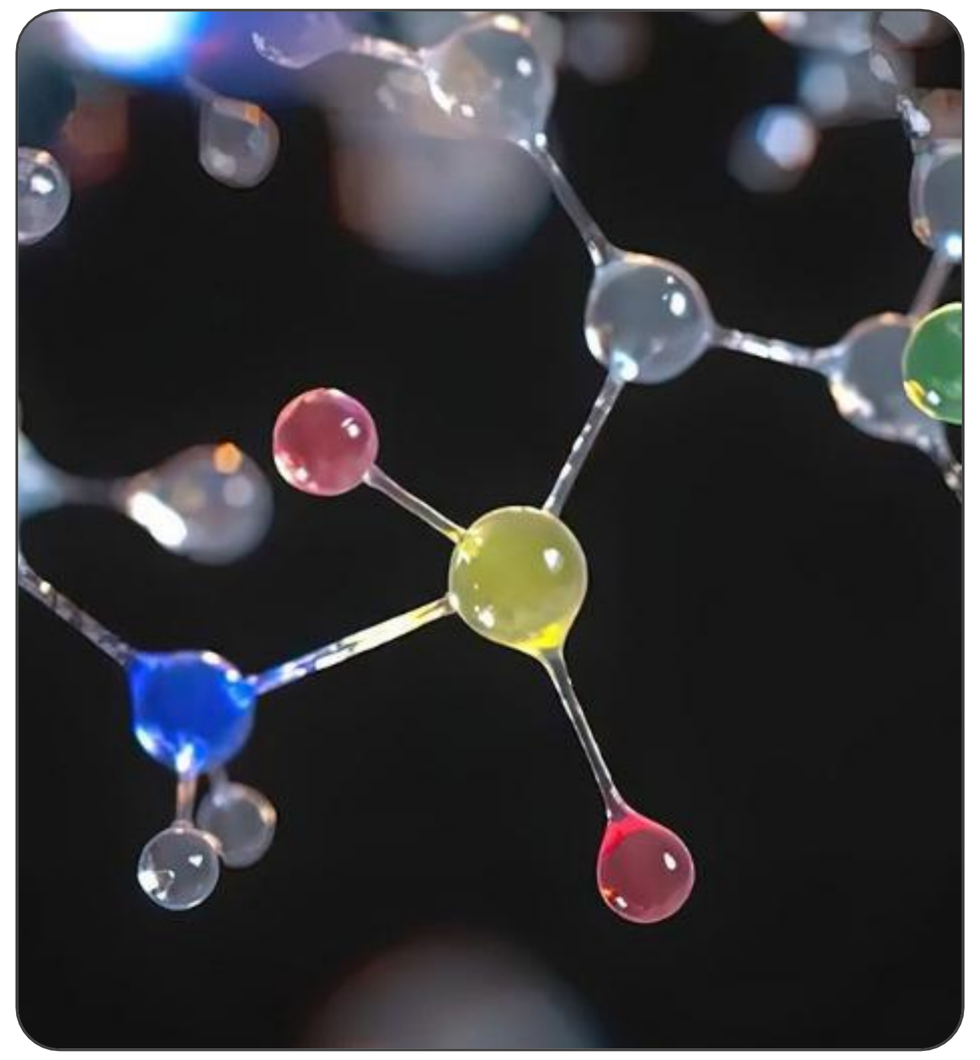


cuLitho
Computational Lithography

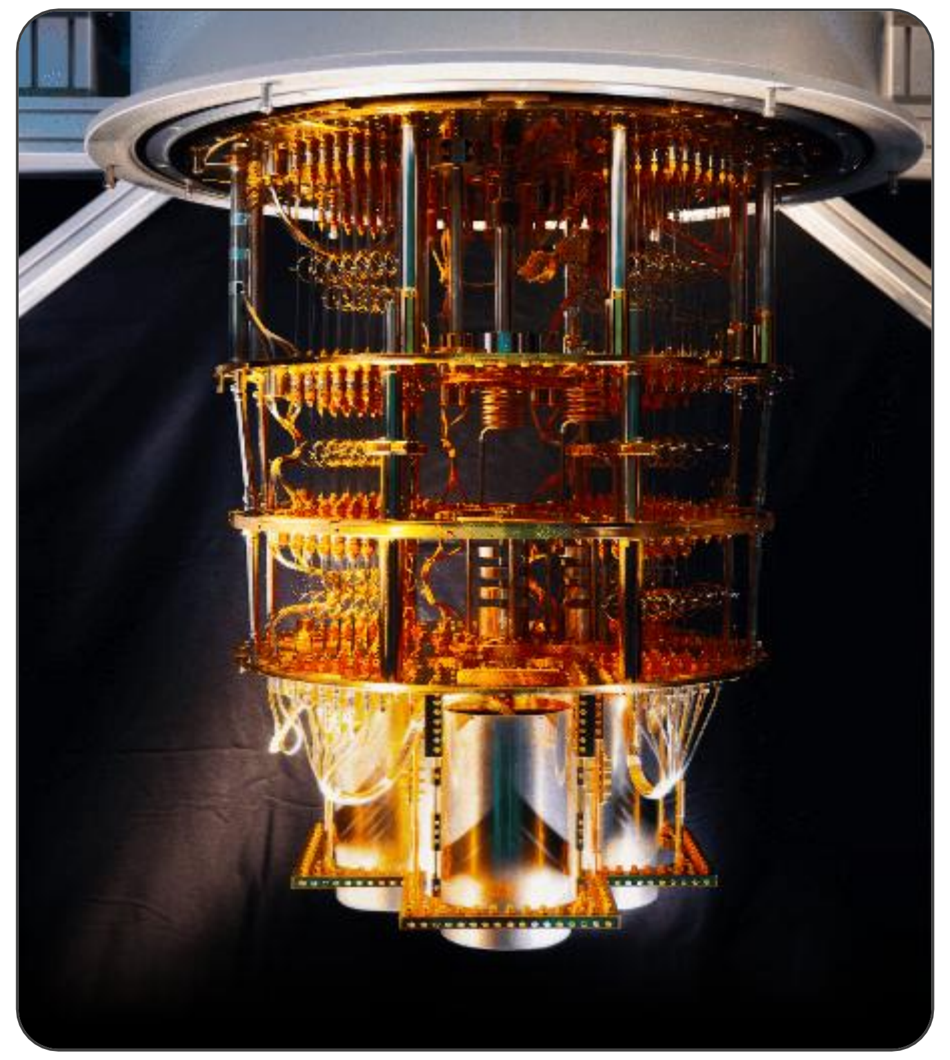
ALCHEMI
AI Materials Science



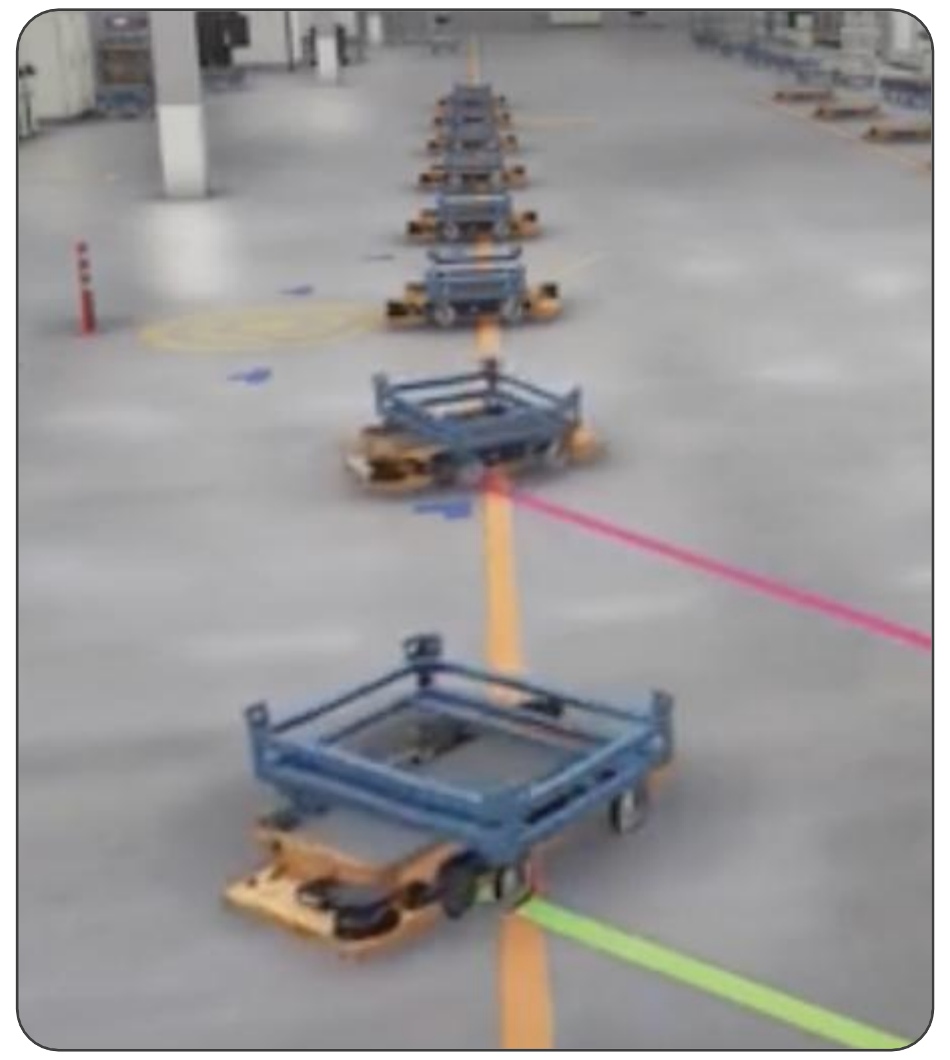
cuEquivariance
Drug & Materials Discovery



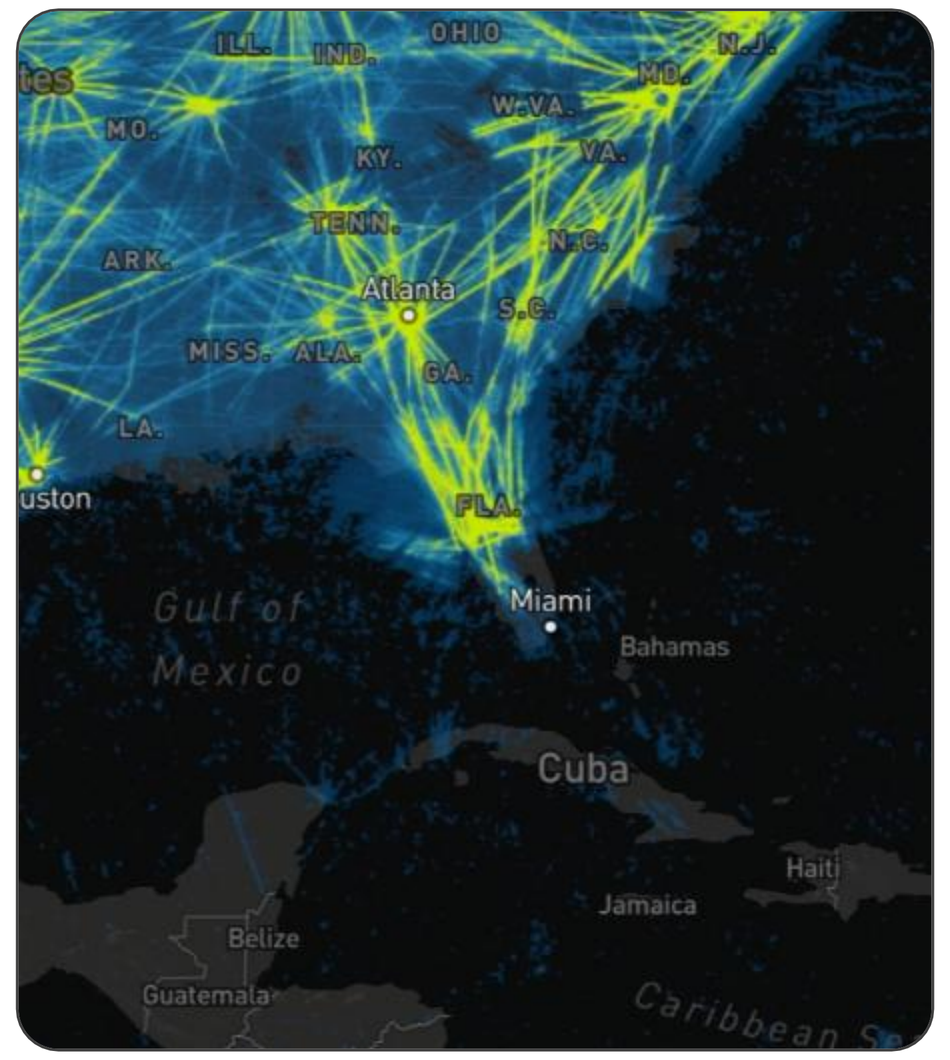
Parabricks
Gene Sequencing



CUDA-Q
Quantum Computing



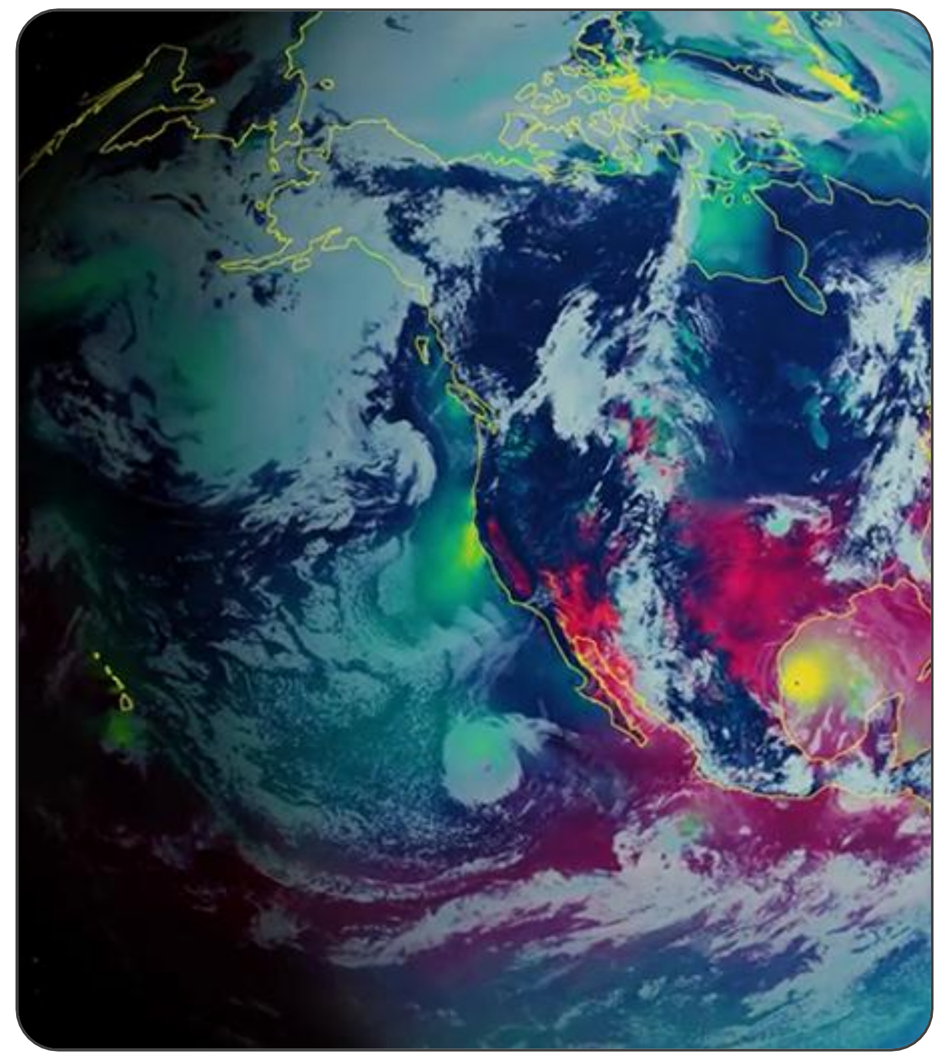
cuOpt
Decision Optimization



cuDF
Data Processing

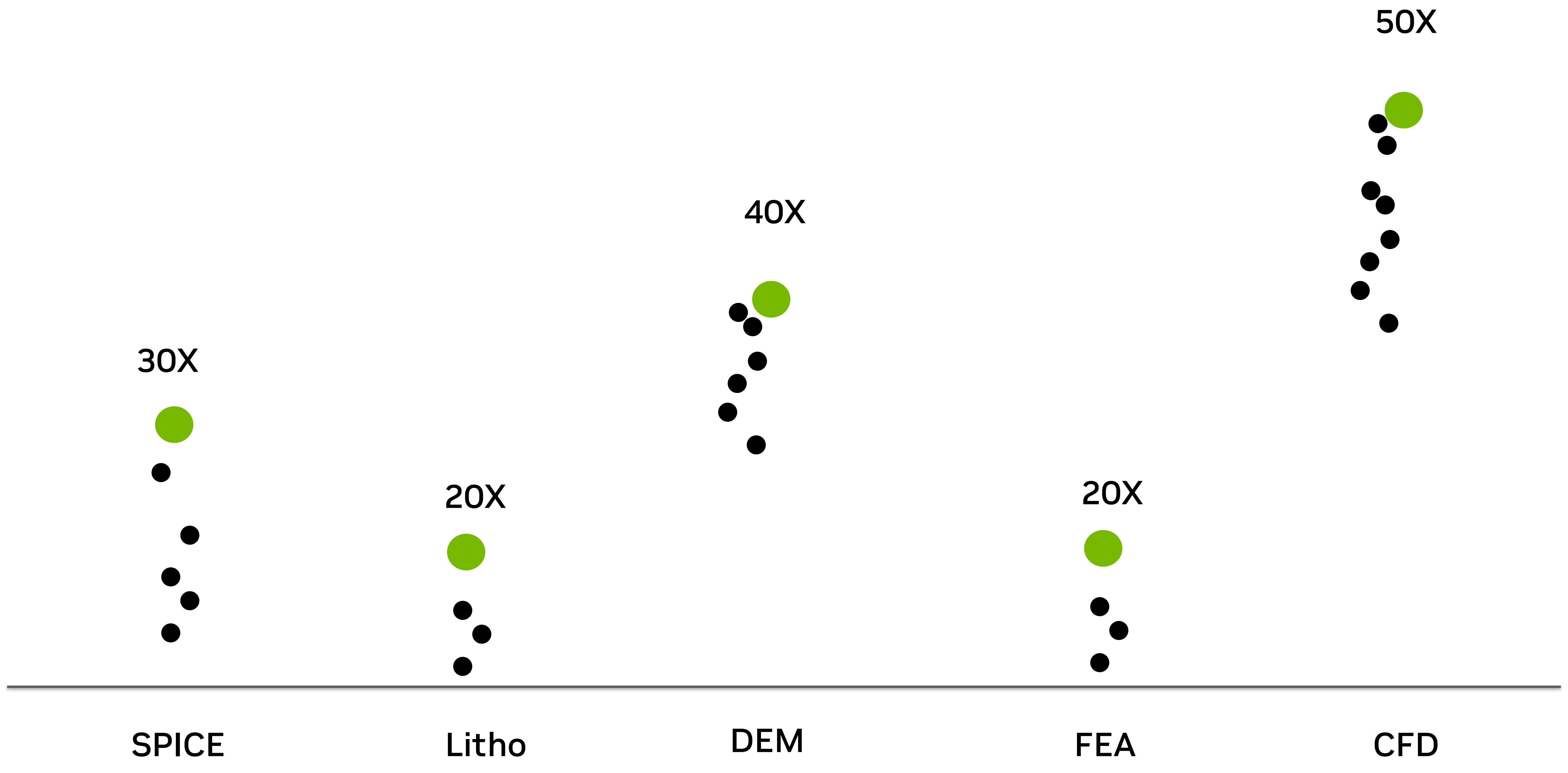
cuPyNumeric
Numerical Computing

Holoscan
Edge HPC



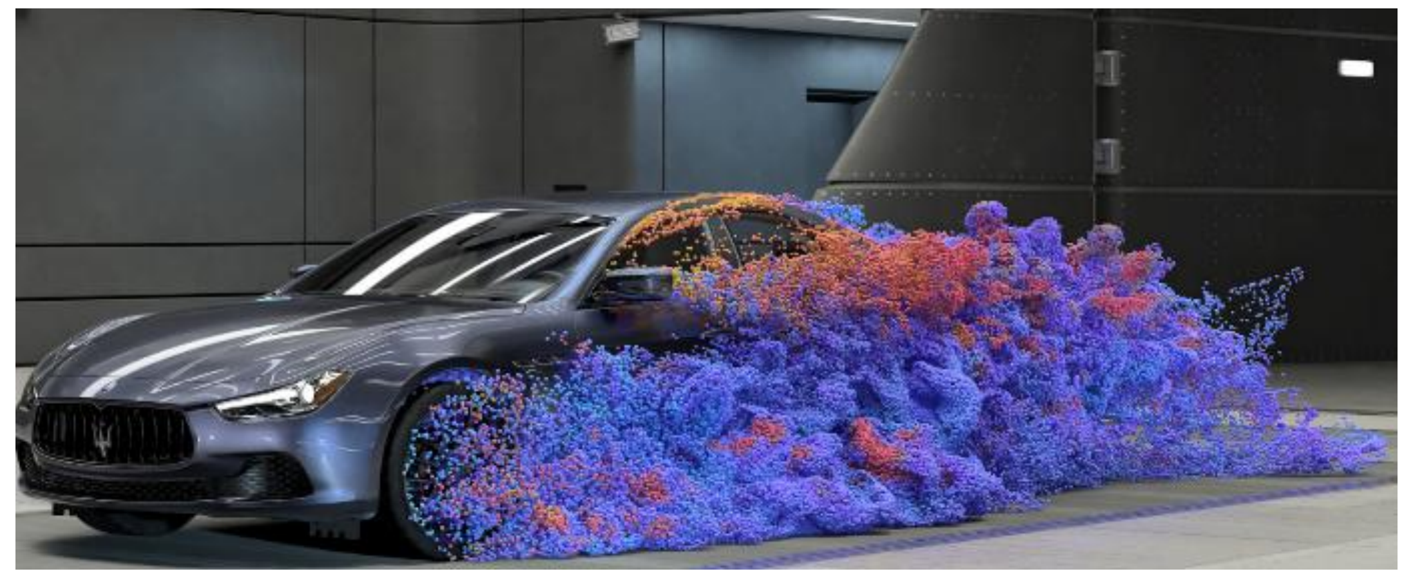
Earth-2
Weather Analytics

Blackwell で加速する計算工学のエコシステム



rescale

Boom Supersonic, Rescale, and NASA Fun3D



SIEMENS

Maserati with Siemens Simcenter STAR-CCM+

cādence

NVIDIA data-center digital twins with Cadence Fidelity

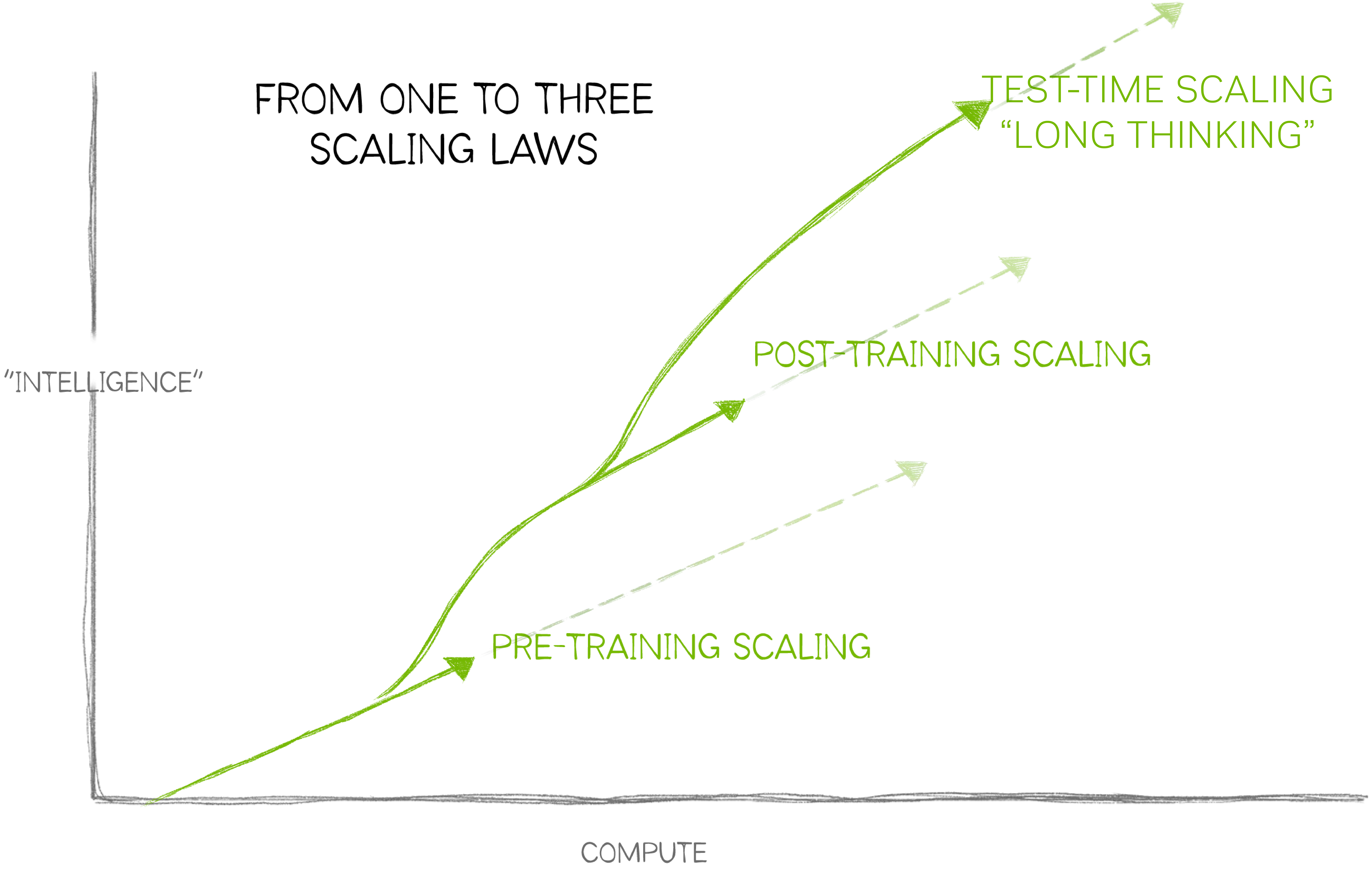
Ansys

Volvo accelerating with Ansys Fluent

ALTAIR
Ansys
AUTODESK
BeyondMath
cādence
COMSOL
engys
Open-source CFD for Enterprise
HEXAGON
luminary
M-Star
NAVASTO
An AUTODESK Company
NEURAL CONCEPT
nTop
PHYSICSX
rescale
SIEMENS
SIMSCALE
VOLCANO Platforms

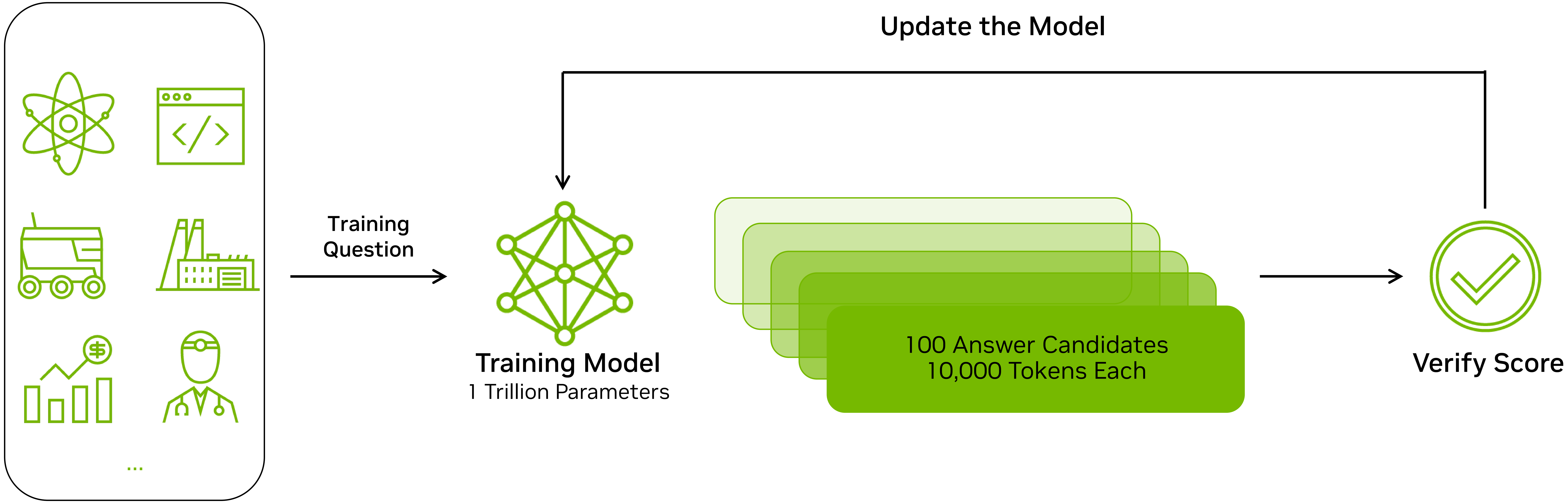
AI スケーリング則による演算需要の指数関数的な高まり

“長考” する AI が推論をスケーリング



リーズニング時代の Post-Training スケーリング

Teaching models to think requires a lot of compute



100
Topic Areas

1 Million
Questions Per topic

100 Topics x 1 Million Questions x 100 Answers x 10,000 Tokens

100 Trillion Tokens

7X More Than What Llama 3 Was Pre-Trained On

リーズニングによるトークン数が急激に増大

Cost increasing with compute – requires full stack innovation to lower cost/token

KNOWLEDGE

Pre-trained models simulate internet knowledge
100 tokens | $0.50 / M Token

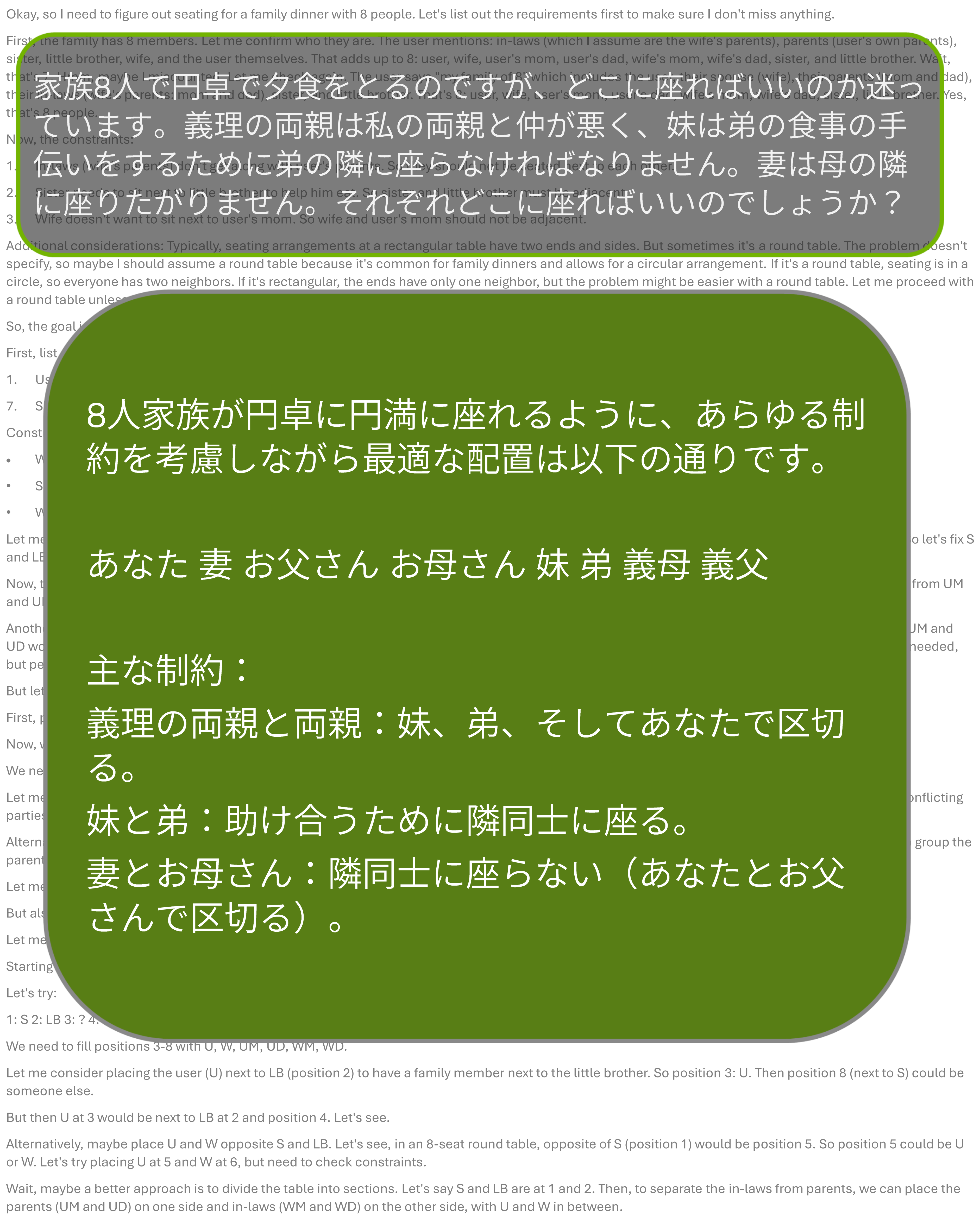
NVIDIA とは？

NVIDIAは、グラフィックス・プロセッシング・ユニット（GPU）、高性能コンピューティング・ハードウェア、人工知能（AI）技術の設計・製造を専門とする多国籍テクノロジー企業です。1993年にジェンスン・フアン、クリス・マラコウスキー、カーティス・プライムによって設立されました。

NVIDIAはカリフォルニア州サンタクララに本社を置き、コンピューターグラフィックス、ゲーム、AIの分野におけるリーディングカンパニーの一つとなっています。

REASONING

Reasoning models think through problems
~10K tokens → 100X cost



家族8人で円卓で夕食をとるのですが、どこに座ればいいのか迷っています。義理の両親は私の両親と仲が悪く、妹は弟の食事の手伝いをするために弟の隣に座らなければなりません。妻は母の隣に座りたがりません。それぞれどこに座ればいいのでしょうか？

Okay, so I need to figure out seating for a family dinner with 8 people. Let's list out the requirements first to make sure I don't miss anything.

First, the family has 8 members. Let me confirm who they are. The user mentions: in-laws (which I assume are the wife's parents), parents (user's own parents), sister, little brother, wife, and the user themselves. That adds up to 8: user, wife, user's mom, user's dad, wife's mom, wife's dad, sister, and little brother. Wait, that's 7. Hmm, maybe I miscounted. Let me check again. The user says "my family of 8" which includes the user, their spouse (wife), their parents (mom and dad), their in-laws (wife's parents: mom and dad), sister, and little brother. That's 8: user, wife, user's mom, user's dad, wife's mom, wife's dad, sister, little brother. Yes, that's 8 people.

Now, the constraints:

1. In-laws (wife's parents) don't get along with user's parents. So they should not be seated next to each other.
2. Sister needs to sit next to little brother to help him eat. So sister and little brother must be adjacent.
3. Wife doesn't want to sit next to user's mom. So wife and user's mom should not be adjacent.

Additional considerations: Typically, seating arrangements at a rectangular table have two ends and sides. But sometimes it's a round table. The problem doesn't specify, so maybe I should assume a round table because it's common for family dinners and allows for a circular arrangement. If it's a round table, seating is in a circle, so everyone has two neighbors. If it's rectangular, the ends have only one neighbor, but the problem might be easier with a round table. Let me proceed with a round table unles

8人家族が円卓に円満に座れるように、あらゆる制約を考慮しながら最適な配置は以下の通りです。

あなた 妻 お父さん お母さん 妹 弟 義母 義父

主な制約：
義理の両親と両親：妹、弟、そしてあなたで区切る。
妹と弟：助け合うために隣同士に座る。
妻とお母さん：隣同士に座らない（あなたとお父さんで区切る）。

We need to fill positions 3-8 with U, W, UM, UD, WM, WD.

Let me consider placing the user (U) next to LB (position 2) to have a family member next to the little brother. So position 3: U. Then position 8 (next to S) could be someone else.

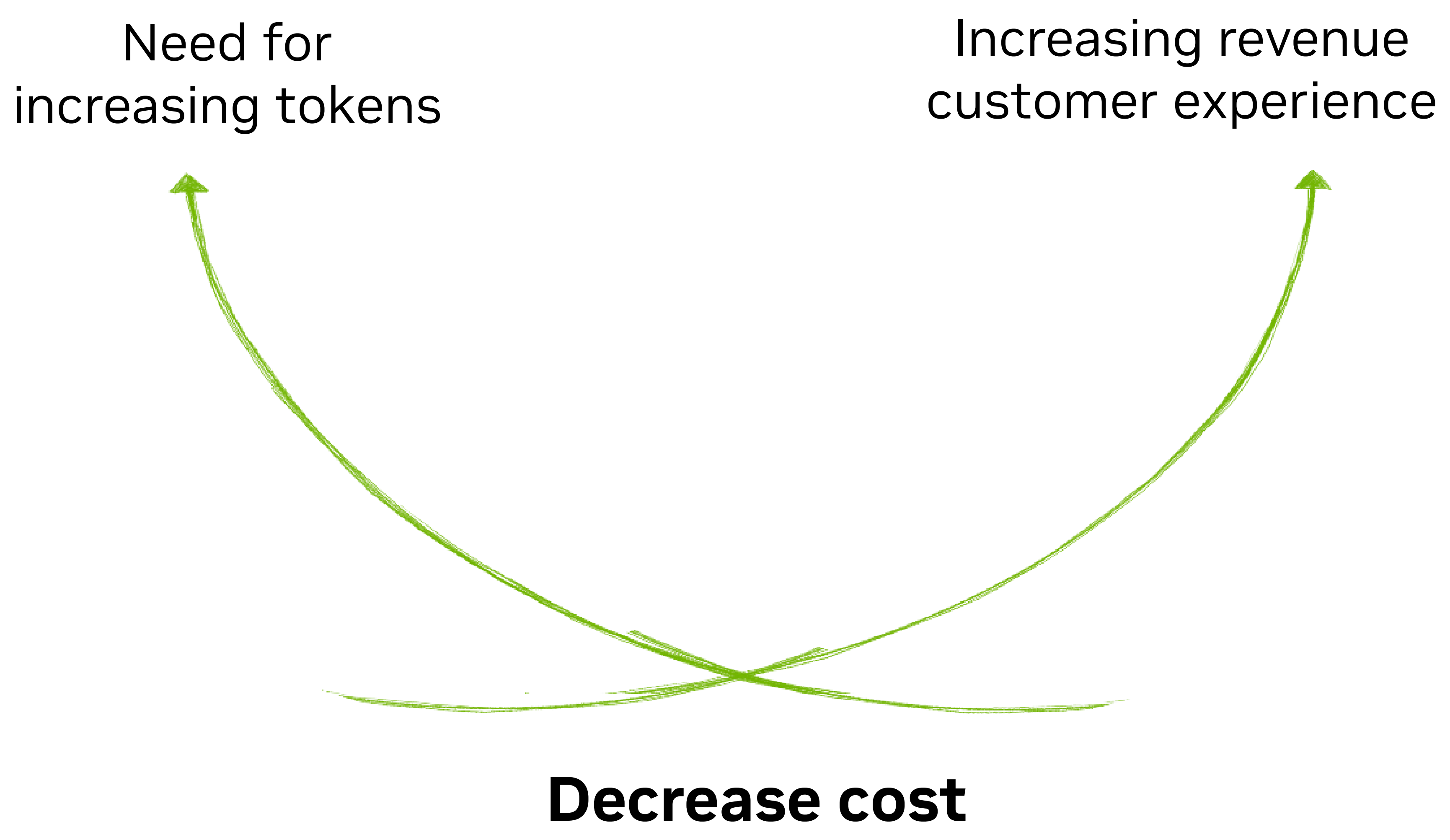
But then U at 3 would be next to LB at 2 and position 4. Let's see.

Alternatively, maybe place U and W opposite S and LB. Let's see, in an 8-seat round table, opposite of S (position 1) would be position 5. So position 5 could be U or W. Let's try placing U at 5 and W at 6, but need to check constraints.

Wait, maybe a better approach is to divide the table into sections. Let's say S and LB are at 1 and 2. Then, to separate the in-laws from parents, we can place the parents (UM and UD) on one side and in-laws (WM and WD) on the other side, with U and W in between.

ACCELERATING AI FACTORY VALUE

>$1/M Token | >1T Param | >300 TPS



NVIDIA Dynamo

NVIDIA Dynamo

リーズニング AI を支える推論ソフトウェア

分散型および非集約型の推論サービス

30X

AI Factory
Throughput
& Revenue

Deepseek R1
Based models

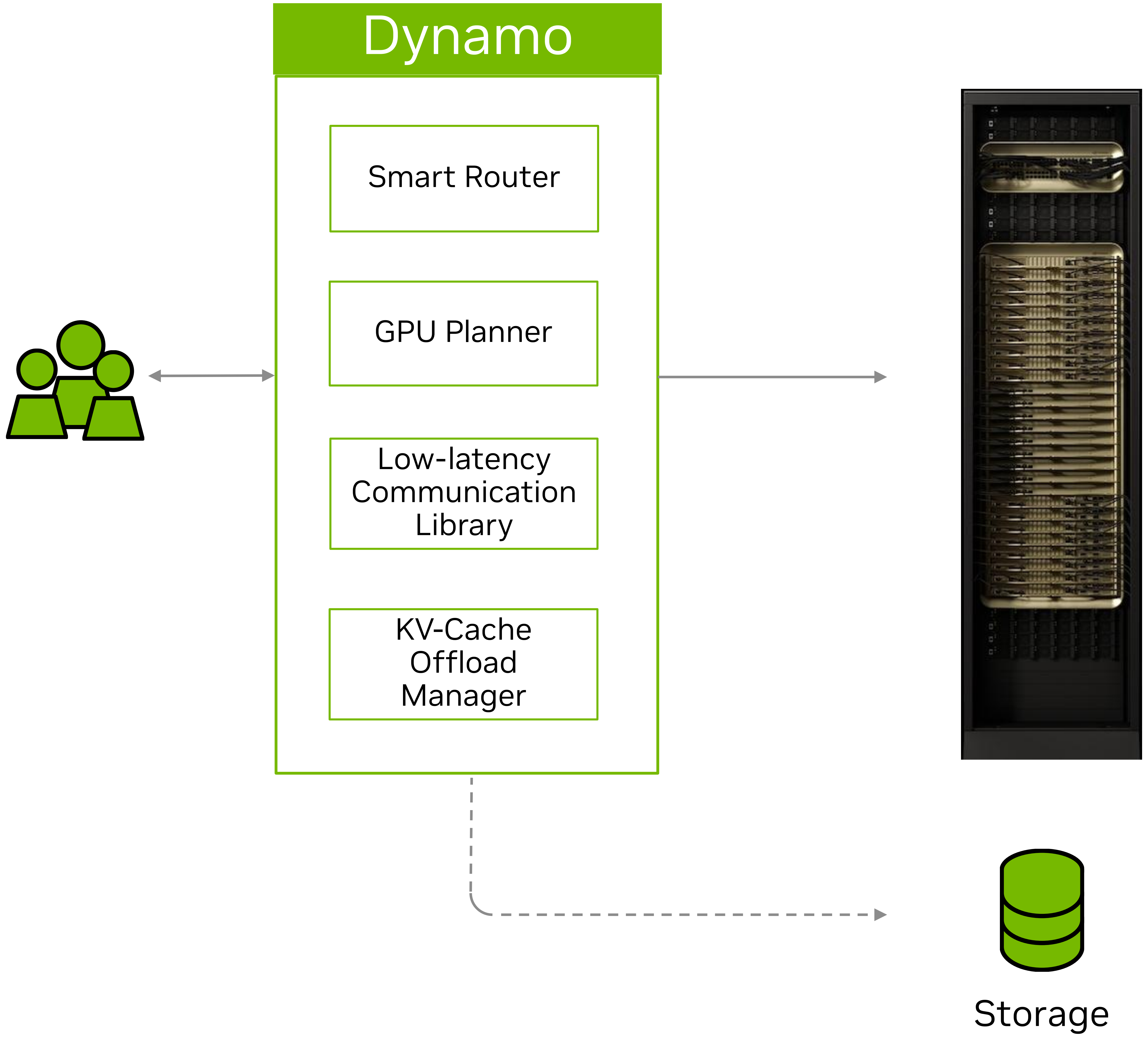
2.5X

AI Factory
Throughput
& Revenue

Llama
Based Models

1000+

GPU Scale for
a single query



Fully Open Source and Open Backend

PyTorch vLLM SGL TensorRT

30X increase: GB200 NVL72, DeepSeek R1 671B. Without Dynamo: Inflight Batching, With Dynamo: Disaggregated Serving. Projected performance subject to change.
2.5X increase: Hopper, Llama 3.1 70B. Without Dynamo: Inflight Batching. With Dynamo: Disaggregated Serving.

Architecture and Components

User Requests

API Server

Compatible with OpenAI API, LlamaStack API and more

Planner

Realtime performance tunning

Smart Router

KV Cache-Aware routing with specialized KV cache insertion and eviction algorithms

Disaggregated Serving

Prefill Worker

Prefill Engine

Distributed KV cache

Decode Worker

Decode Engine

Distributed KV cache

Event Plane

Metric transfer across DYNEMO components

KV Cache Manager

KV cache offloading

Data Transfer Engine

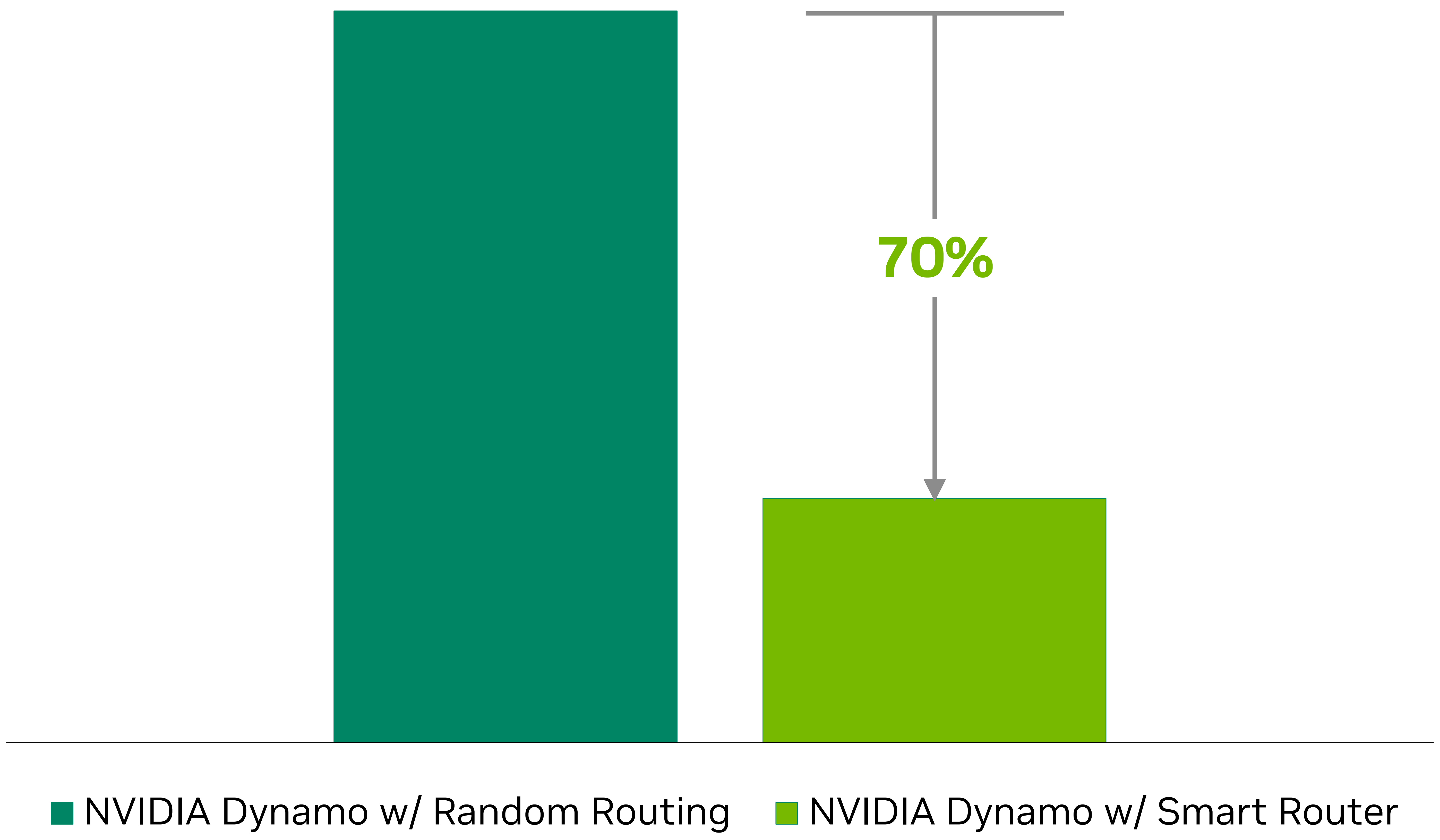
Low latency, interconnect-agnostic, multi-node data transfer

NVIDIA Dynamo: Smart Router

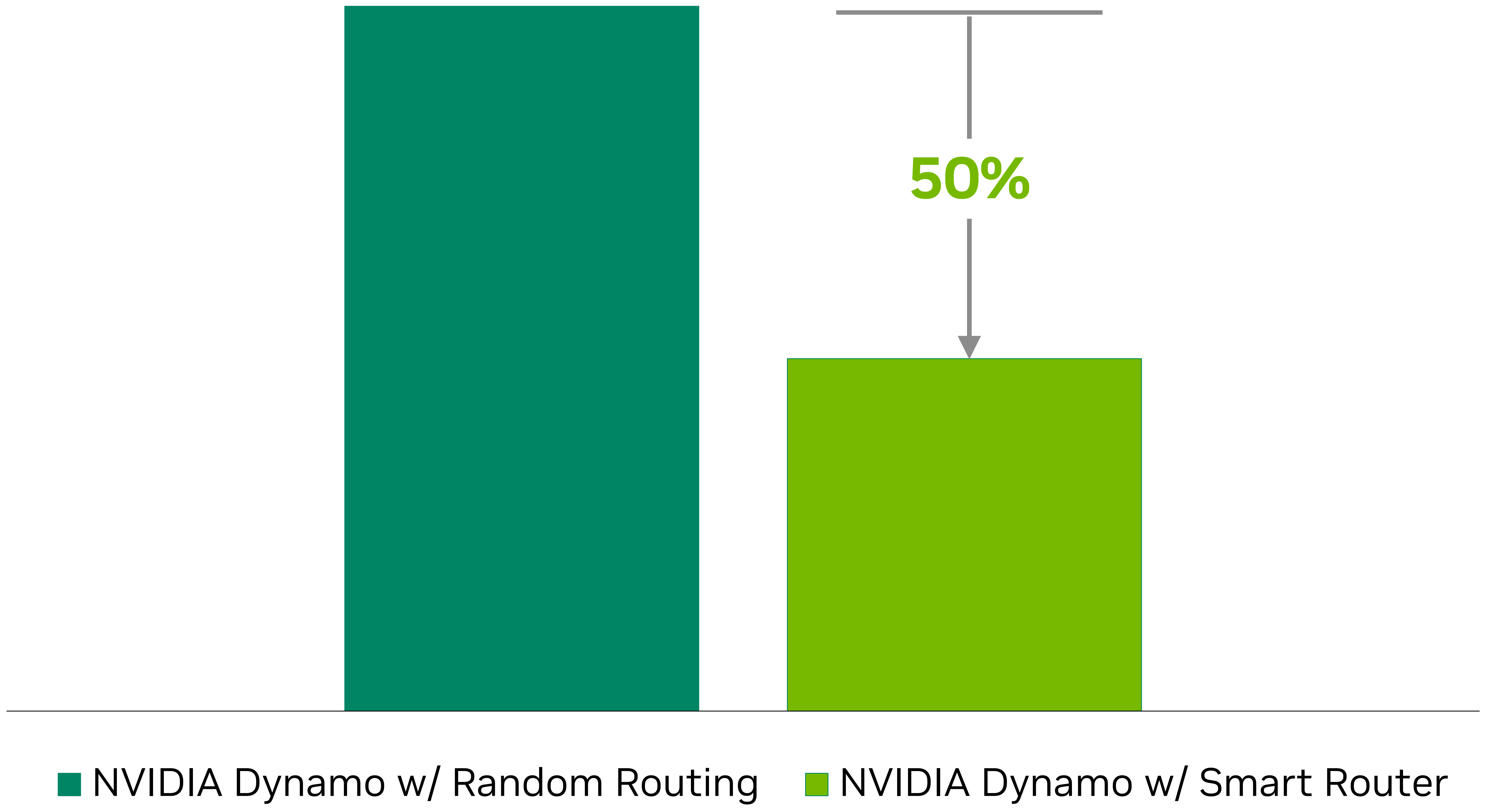
Reducing costly re-computation of KV cache

DeepSeek-R1 Distill Llama 70B | NVIDIA HGX-H100
(Lower is Better)

Time to First Token



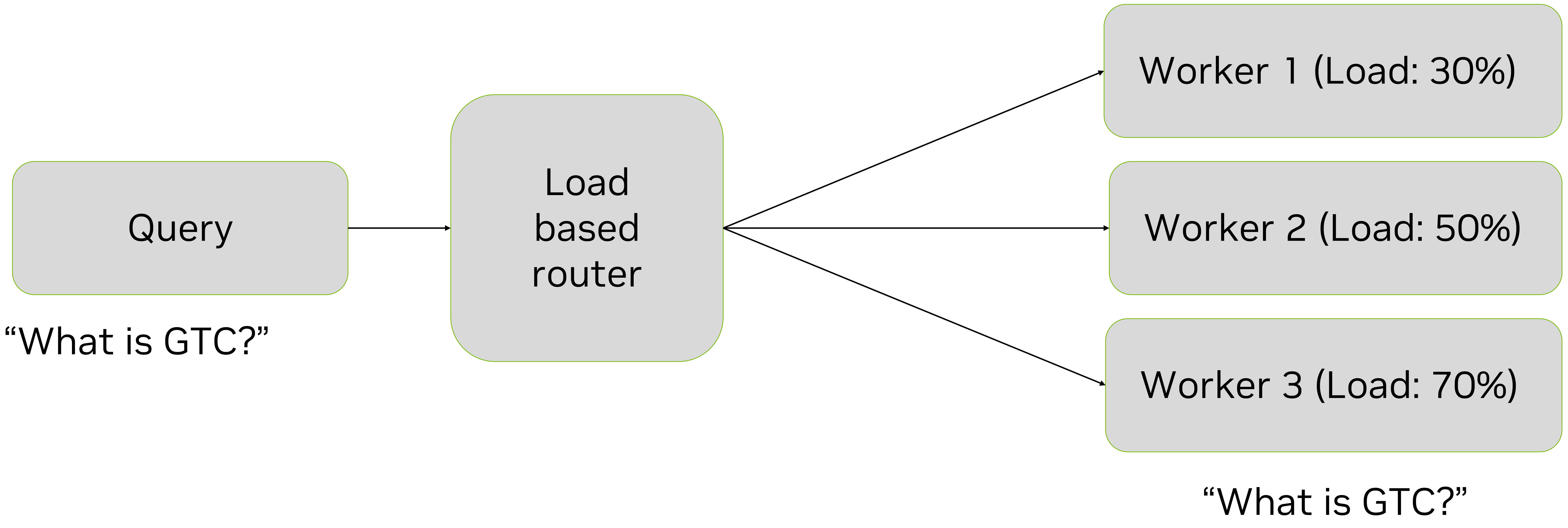
Avg. Request Latency

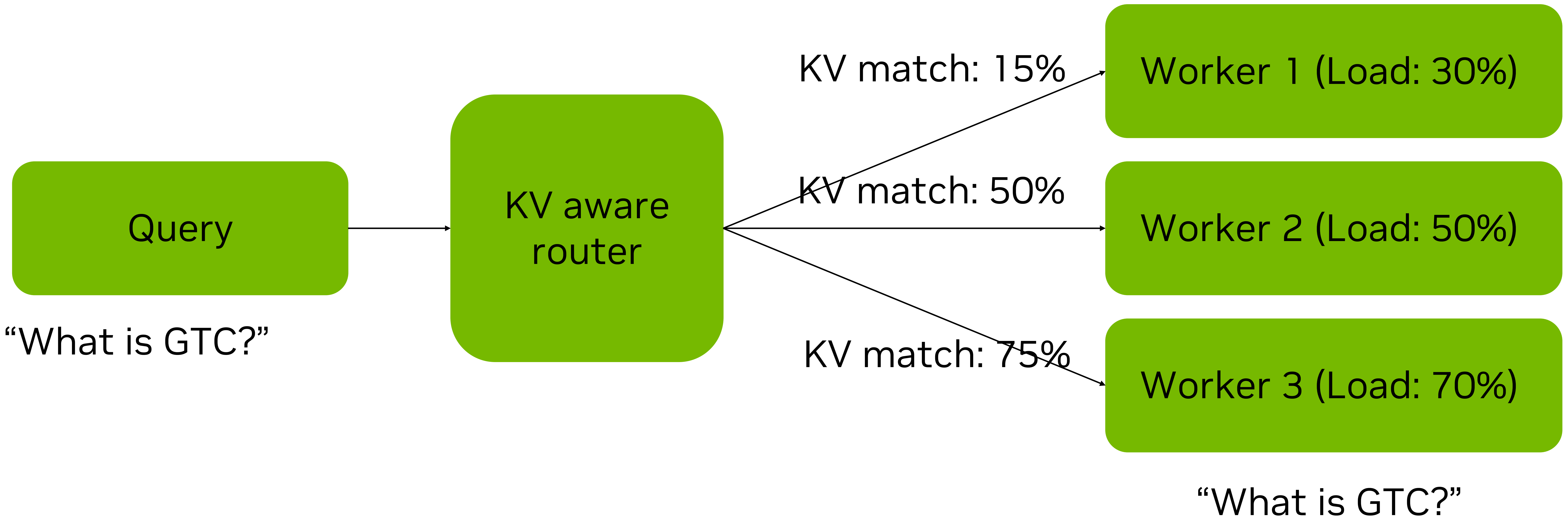


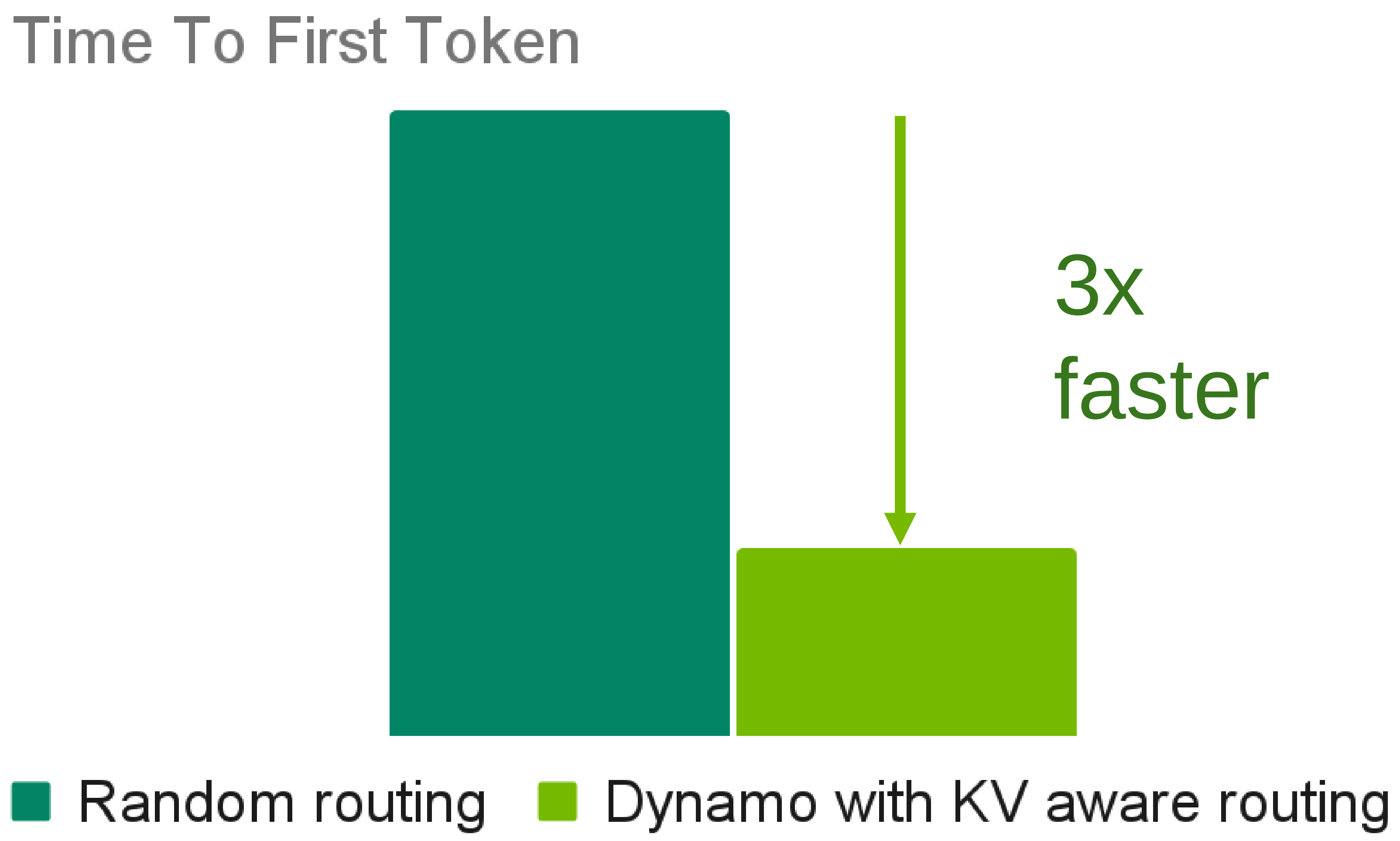
2x HGX-H100 nodes
8x DeepSeek-R1- Distill-Llama-70B. vLLM, FP8, Tensor Parallel: 2
Data Source: 100K real R1 requests, Avg ISL/OSL: 4k/800

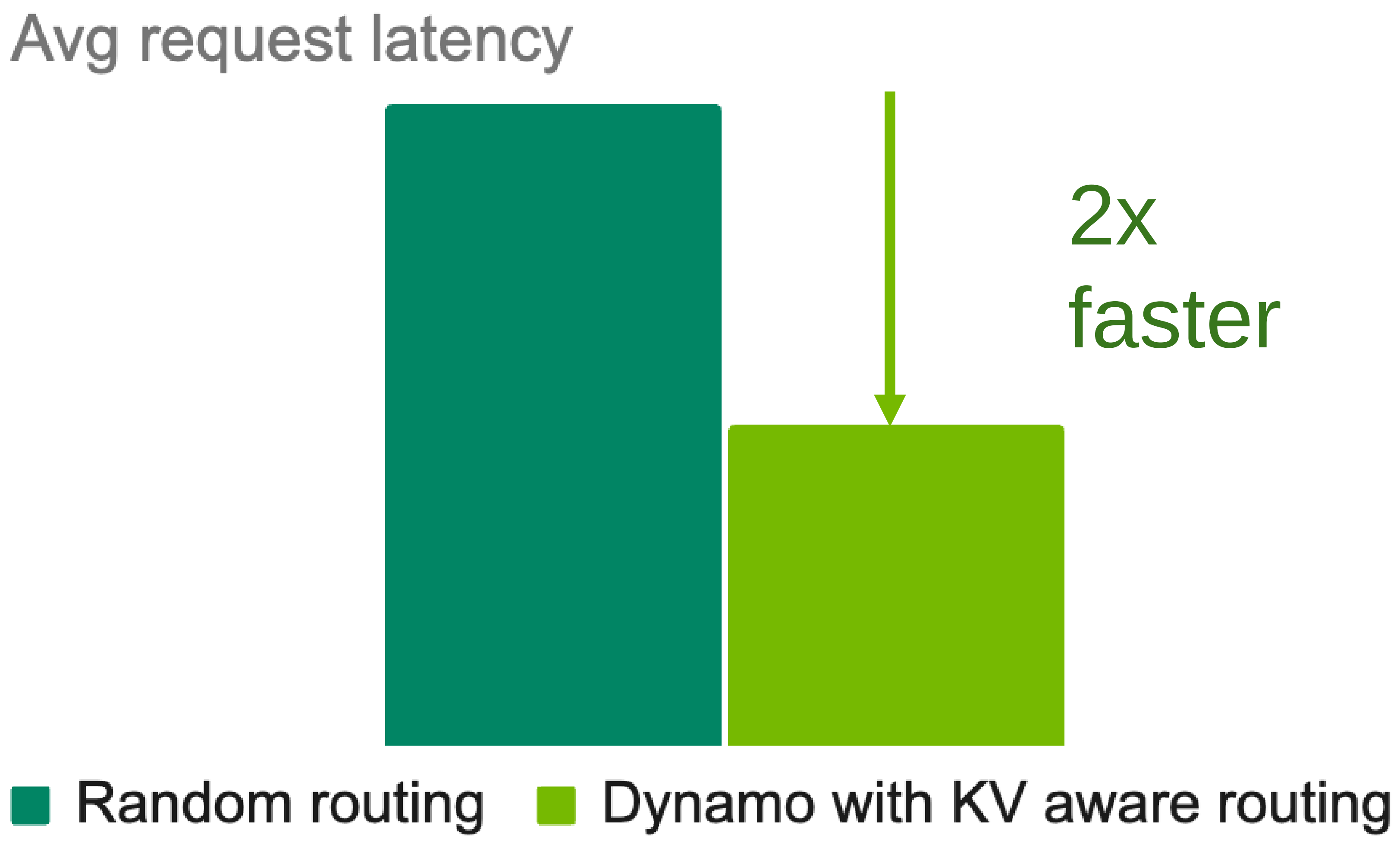
KV Cache Aware Routing

Significant boost in TTFT and End to End Latency with real data (100K requests with R1)



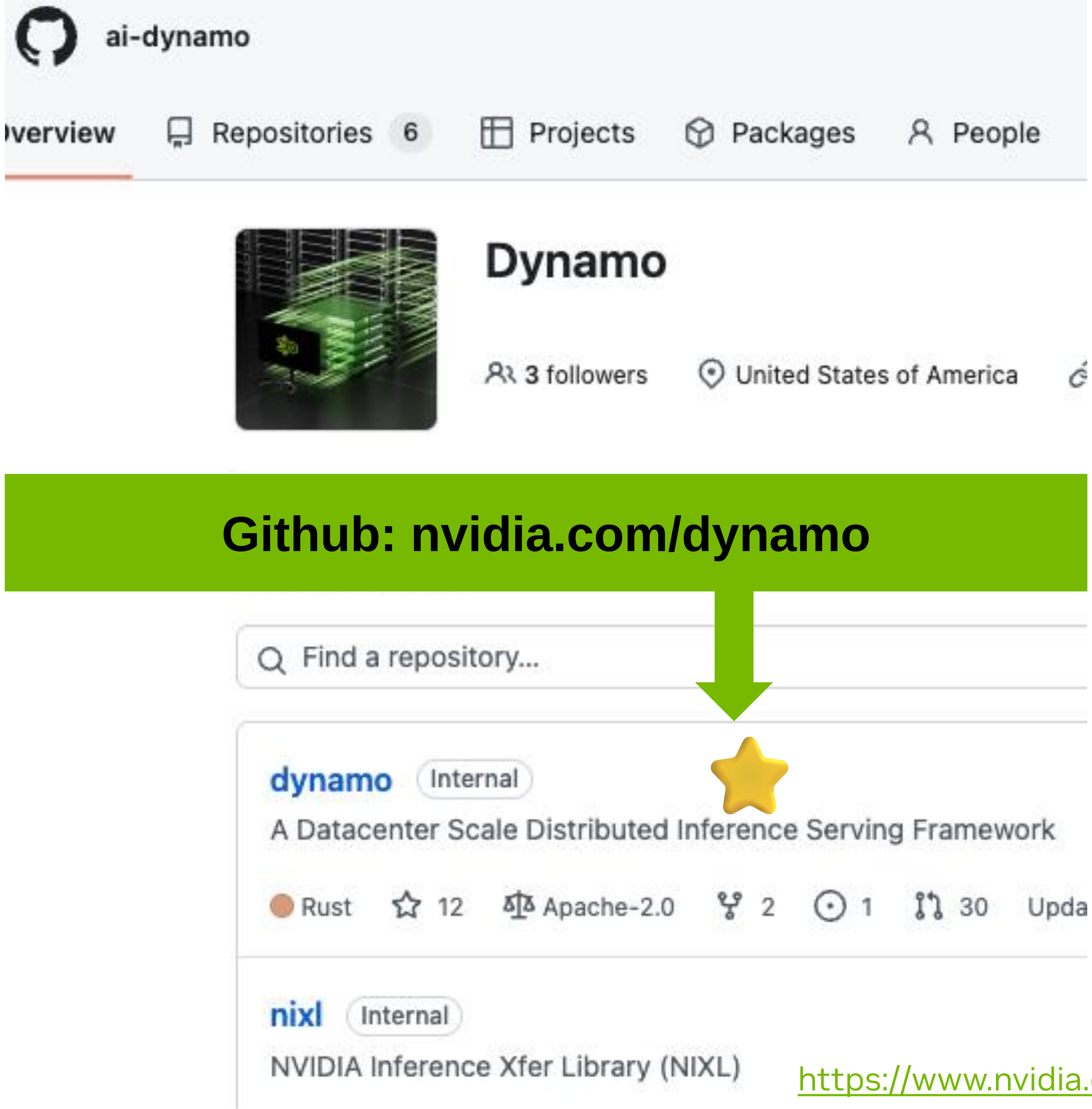






Tested with R1 Distilled Llama 70B over 2 nodes of 8 x H100s with vLLM 0.7.3

NVIDIA Dynamo as of Today (=GTC week)



Dynamo

3 followers United States of America

Github: nvidia.com/dynamo

dynamo Internal

A Datacenter Scale Distributed Inference Serving Framework

Rust 12 Apache-2.0 2 1 30 Upda

nixl Internal

NVIDIA Inference Xfer Library (NIXL)

- Apache 2 license and public CI
- Pip wheels on PyPi
- Rust for perf and Python for extensibility
- Discord for developer community

- Dynamo CLI
  - dynamo run: Quick start with model, input and output
  - dynamo serve: Construct graph of workers and serve
  - (EA) dynamo build: Containerize
  - (EA) dynamo deploy: Deploy to K8

- Three backends: TRT-LLM, vLLM, & SGLang
- Disaggregated serving with TRT-LLM and vLLM
- KV aware routing with TRT-LLM and vLLM
- (EA) KV manager with vLLM
- NIXL for RDMA and TCP (fallback for AWS EFA)

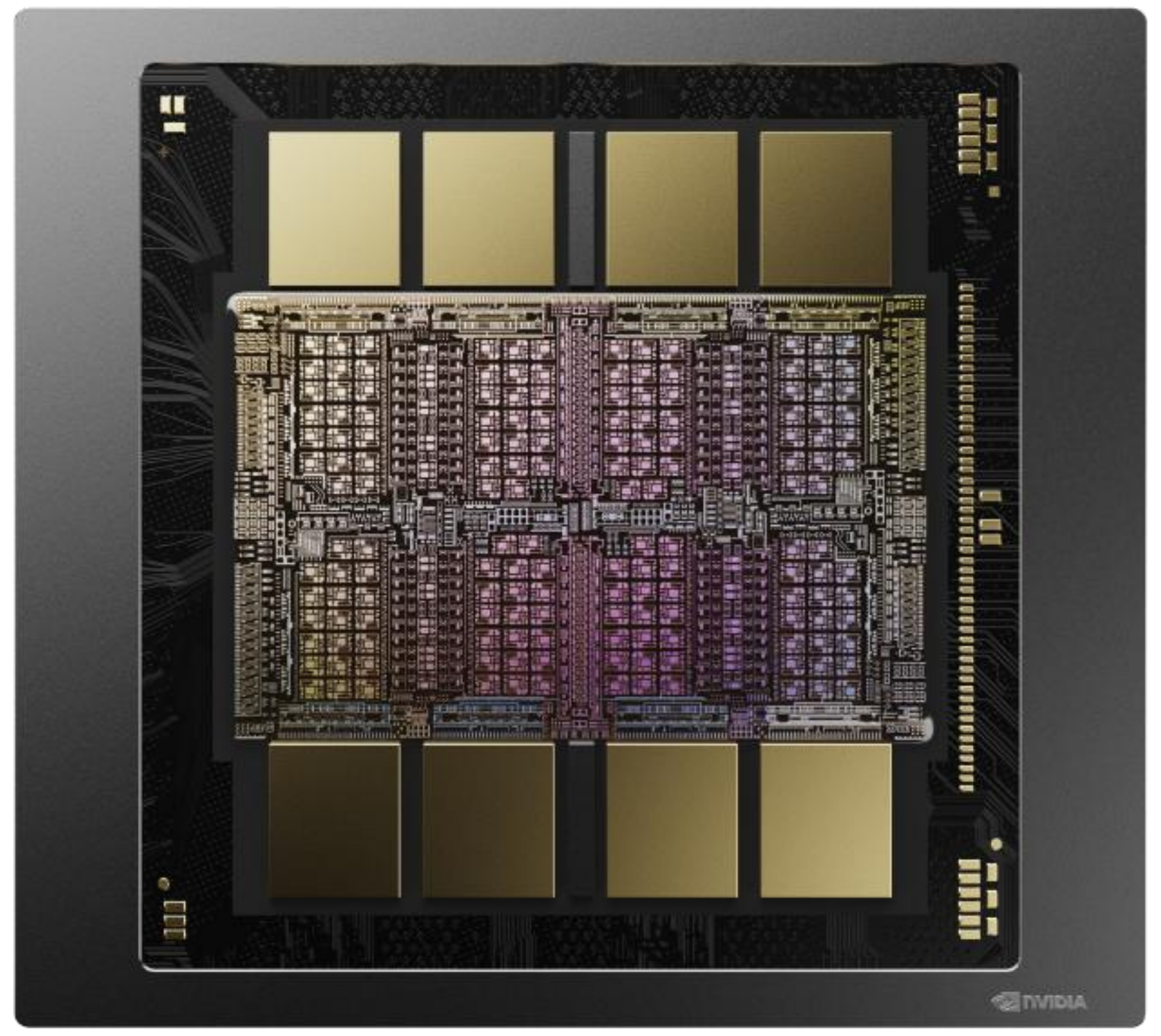
https://www.nvidia.com/en-us/on-demand/session/gtc25-S73042/

Blackwell Ultra

Announcing Blackwell Ultra

Built for the Age of AI Reasoning

BLACKWELL ULTRA GPU



Upgraded FP4	15PF Dense
HBM Memory	Up to 288GB HBM3e
Attention	2.5x Hopper

Blackwell 288GB GPU | 1.5x more FP4 Inference

50X AI Factory Output

Upgraded NVL72 Design for Improved Energy Efficiency

NVIDIA GB300 NVL72

FP4 (Dense)	1.1 ExaFLOPS
HBM Memory	20 TB
Fast Memory	40 TB
Networking	14.4 TB/s

Compared to HGX H100 | AI Reasoning on Deepseek R1 | FP4 – Dense indicated. Projected performance subject to change

NVIDIA GB300 NVL72

Built for the Age of AI Reasoning

Increasing AI Reasoning
Throughput and Responsiveness

50X

AI Factory Output

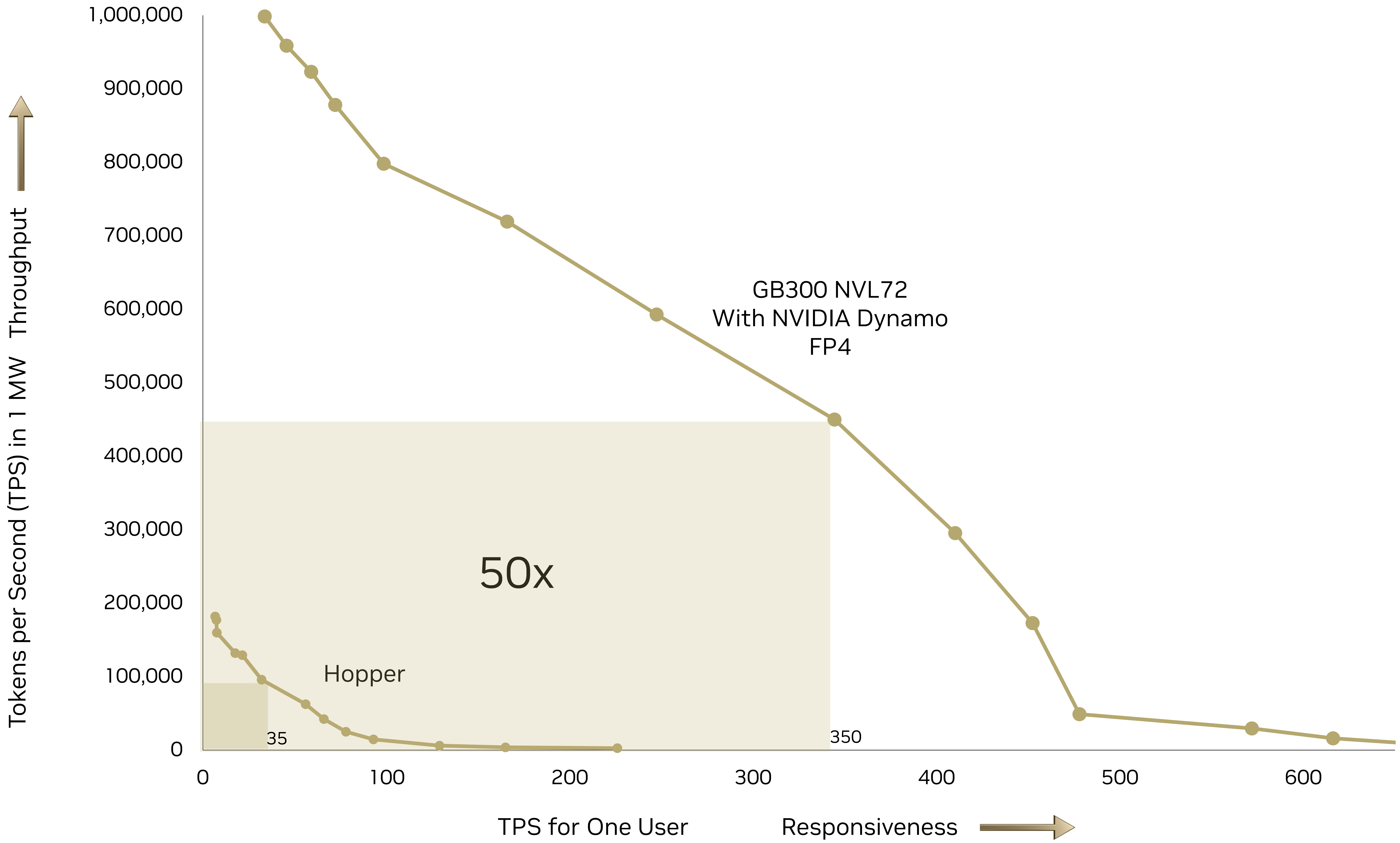
Interactive DeepSeek-R1 671B

H100		GB300 NVL72
35 TPS for one user	→	350 TPS for one user
1.5 Mins	→	10 Seconds

Projected performance subject to change

Compared to HGX H100 | AI Reasoning on Deepseek R1

NVIDIA Blackwell Ultra AI Factory Output



DeepSeek R1 ISL = 32K, OSL = 8K, GB300 NVL72 with FP4 Dynamo disaggregation. H100 with FP8 In-flight batching. Projected performance subject to change.

DGX GB300

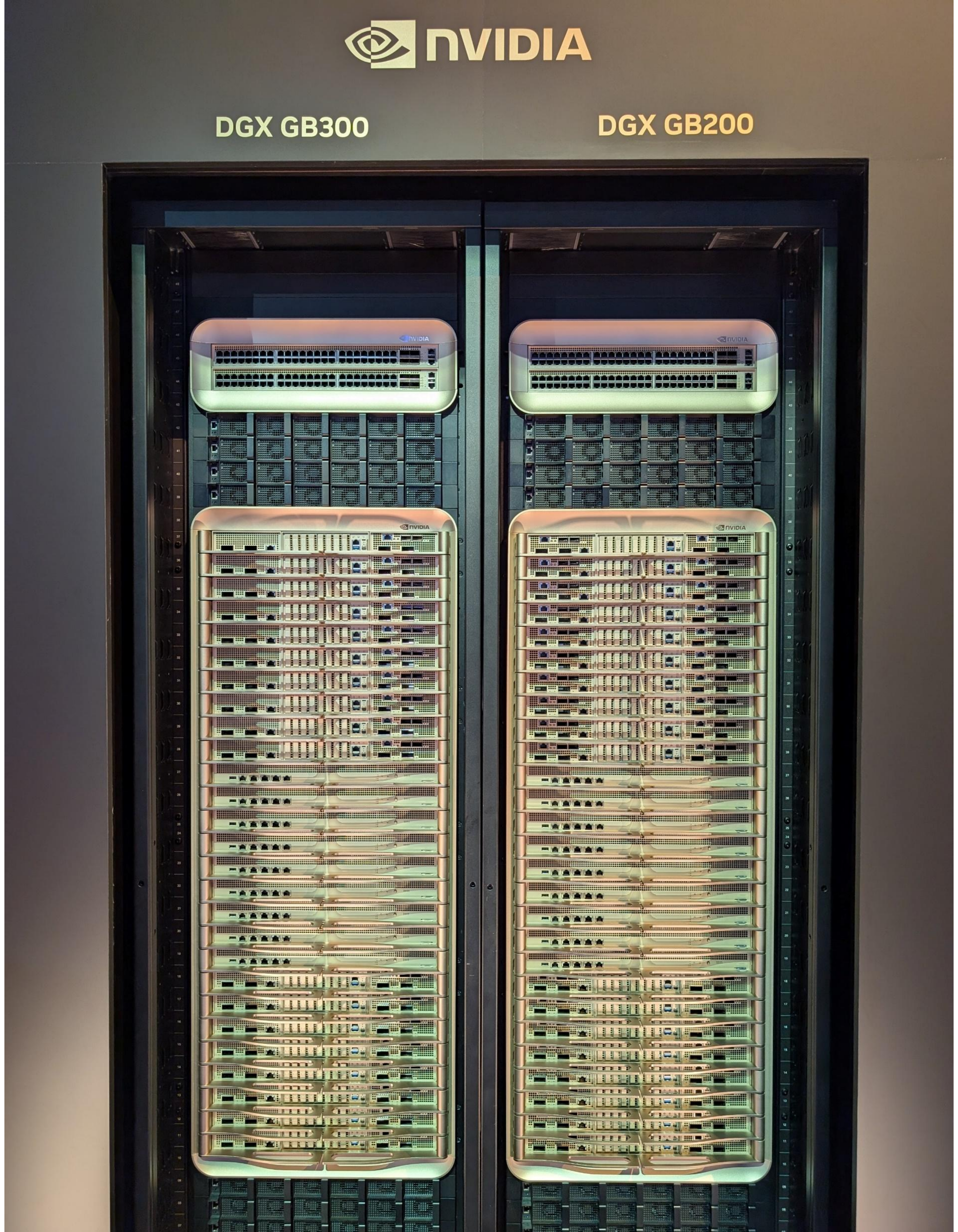
Accelerate Real-Time State of the Art Inference Models

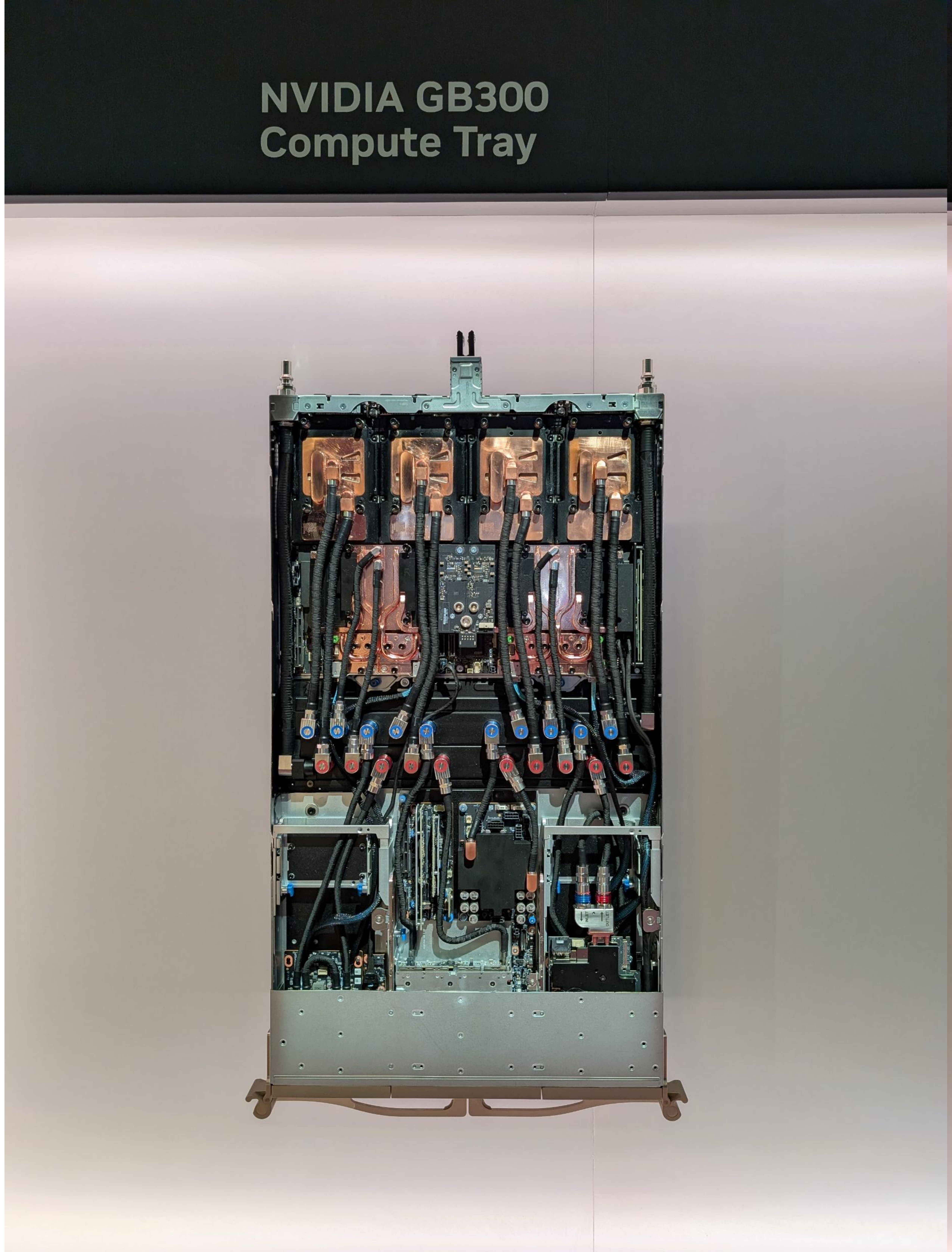
- Based on the NVIDIA GB300 NVL72 rack architecture
- Provides the foundation for NVIDIA DGX SuperPOD with DGX GB300
- Powered by Grace Blackwell Ultra Superchips connected with fifth-generation NVIDIA NVLink
- 36 Grace CPUs and 72 Blackwell Ultra GPUs
- 1.4 exaFLOPS of AI performance and 38TB of fast memory
- Massive shared memory space to accelerate the most data-intensive workloads

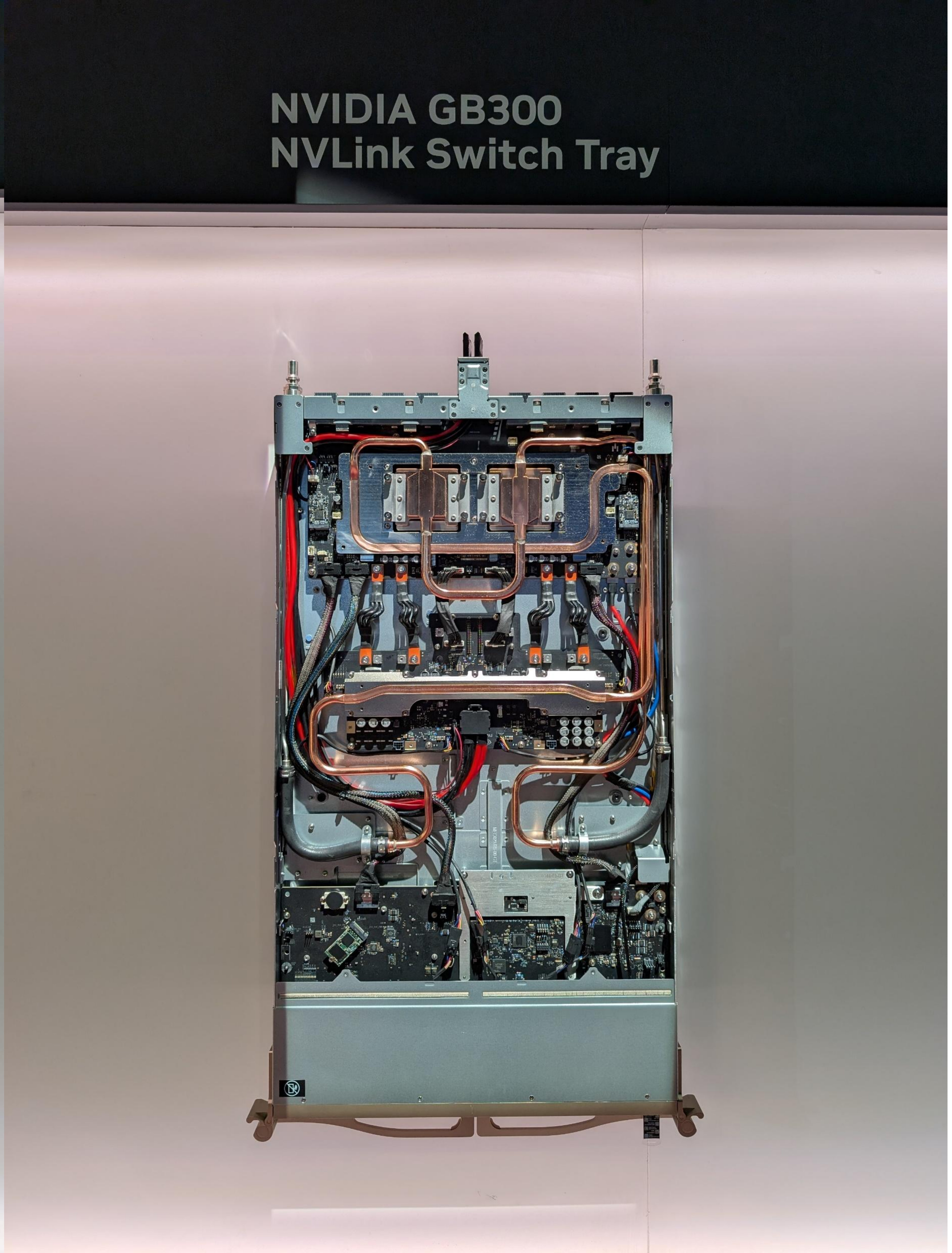
NVIDIA DGX GB200

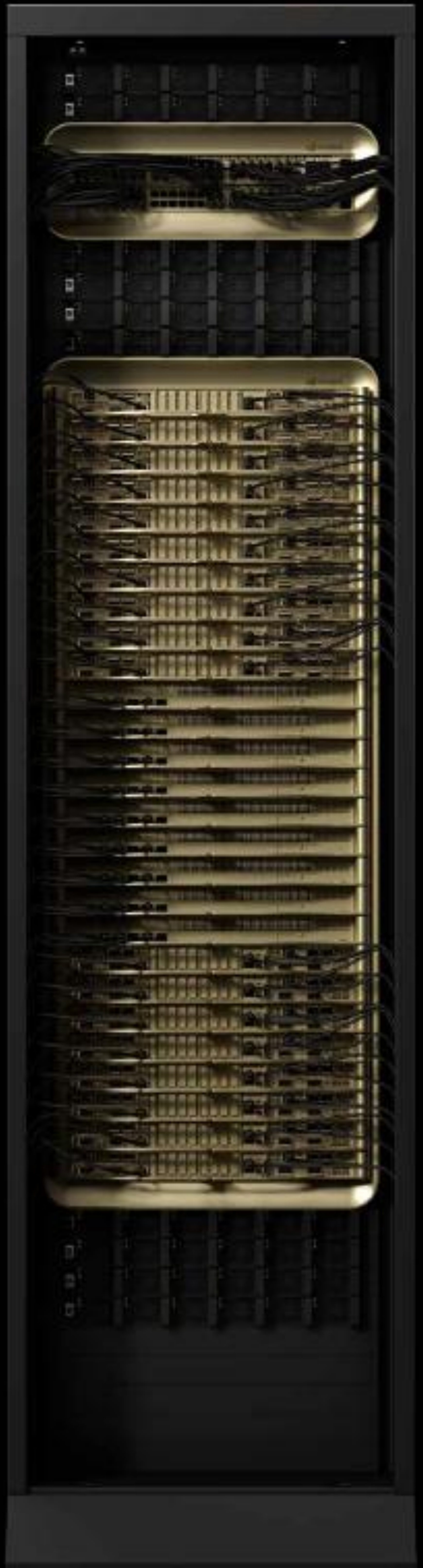
Always-available enterprise infrastructure for mission-critical AI

- The building block of DGX SuperPOD with DGX GB200 systems
- Based on the NVIDIA GB200 NVL72
- Provides a fully-integrated, ready-to-scale infrastructure solution for generative AI
- Built with 36 GB200 Superchips and fifth-gen NVLink
- Connects 36 Grace CPUs and 72 Blackwell GPUs for compute intensive workloads
- 1.4 exaFLOPS of AI performance and 30TB of fast memory
- Handles the most complex generative AI workloads


NVIDIA
DGX GB300
DGX GB200


NVIDIA GB300
Compute Tray


NVIDIA GB300
NVLink Switch Tray



Oberon Rack
Liquid Cooled

Vera Rubin NVL144

Second Half 2026

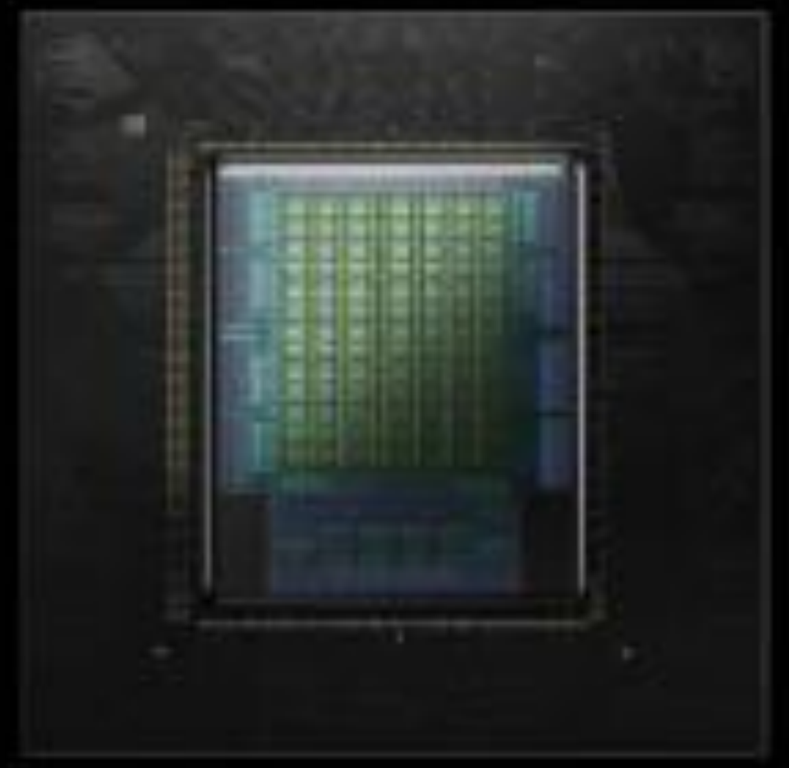
3.6 EF FP4 Inference
1.2 EF FP8 Training
3.3X GB300 NVL72

13 TB/s HBM4
75 TB Fast Memory
1.6X

260 TB/s NVLink6
2X

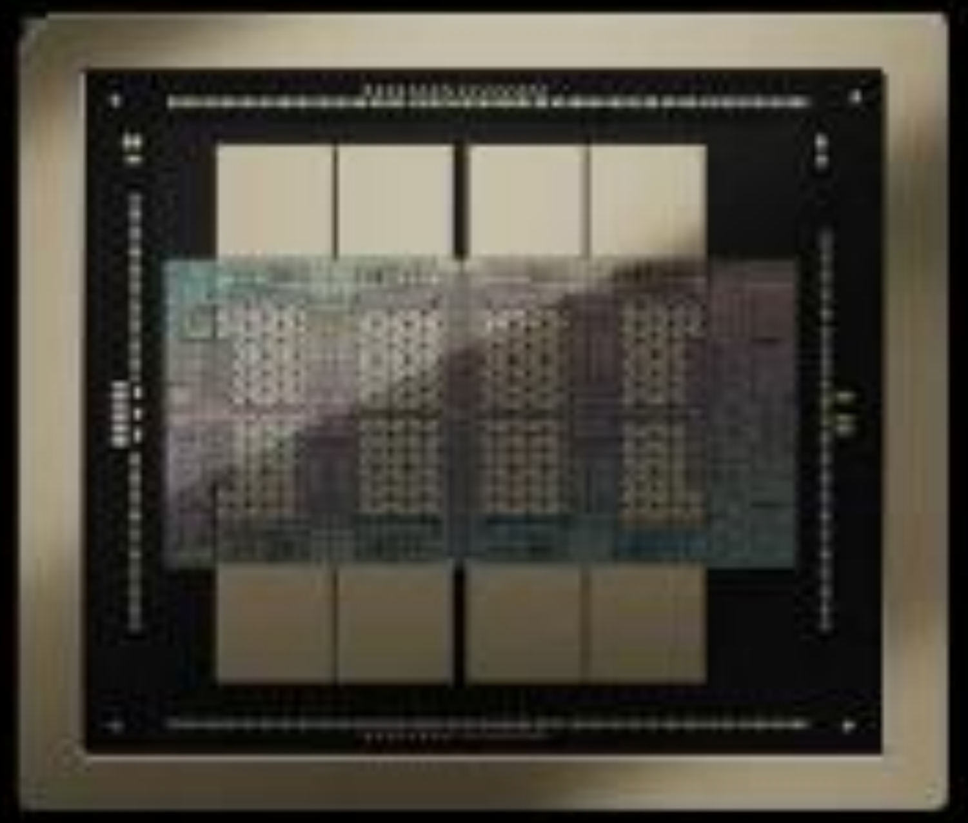
28.8 TB/s CX9
2X

Vera

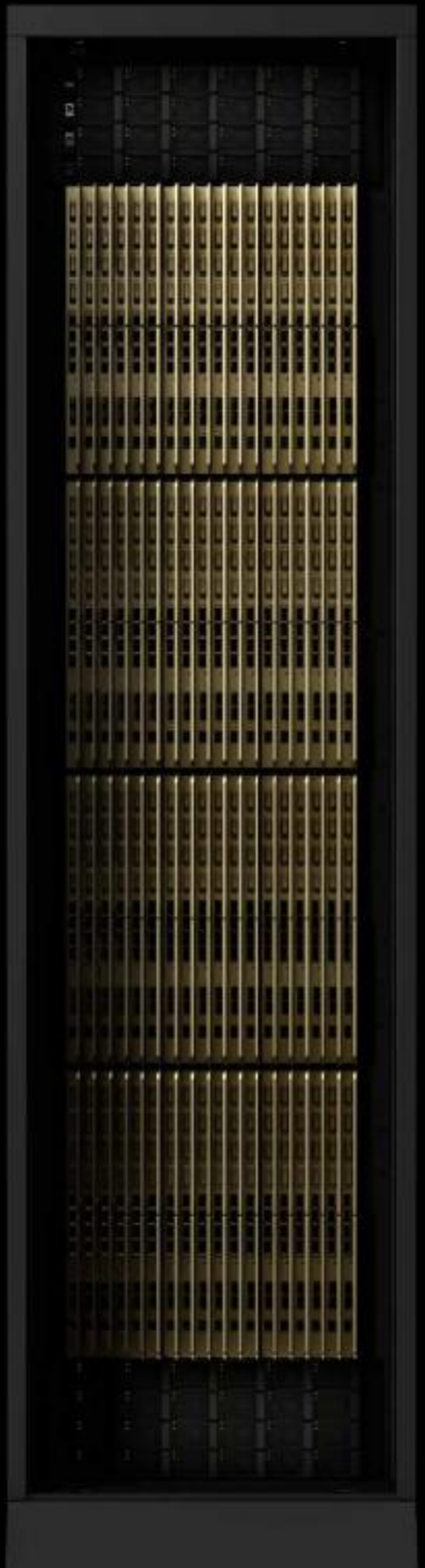


88 Custom Arm Cores
176 Threads
1.8 TB/s NVLink-C2C

Rubin



2 Reticle-Sized GPUs
50PF FP4 | 288GB HBM4



Kyber Rack
Liquid Cooled

Rubin Ultra NVL576

Second Half 2027

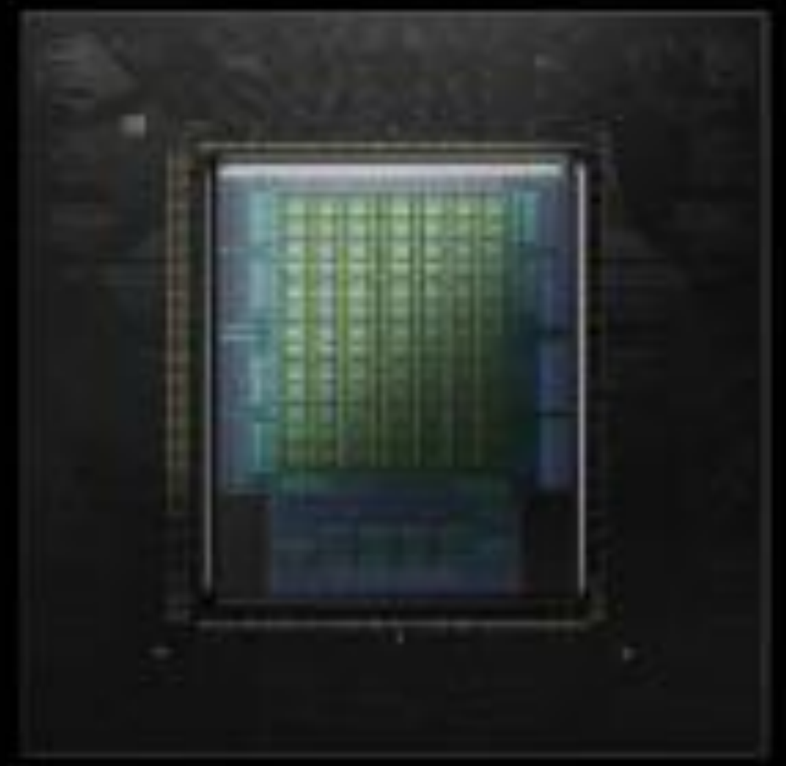
15 EF FP4 Inference
5 EF FP8 Training
14X GB300 NVL72

4.6 PB/s HBM4e
365 TB Fast Memory
8X

1.5 PBs NVLink7
12X

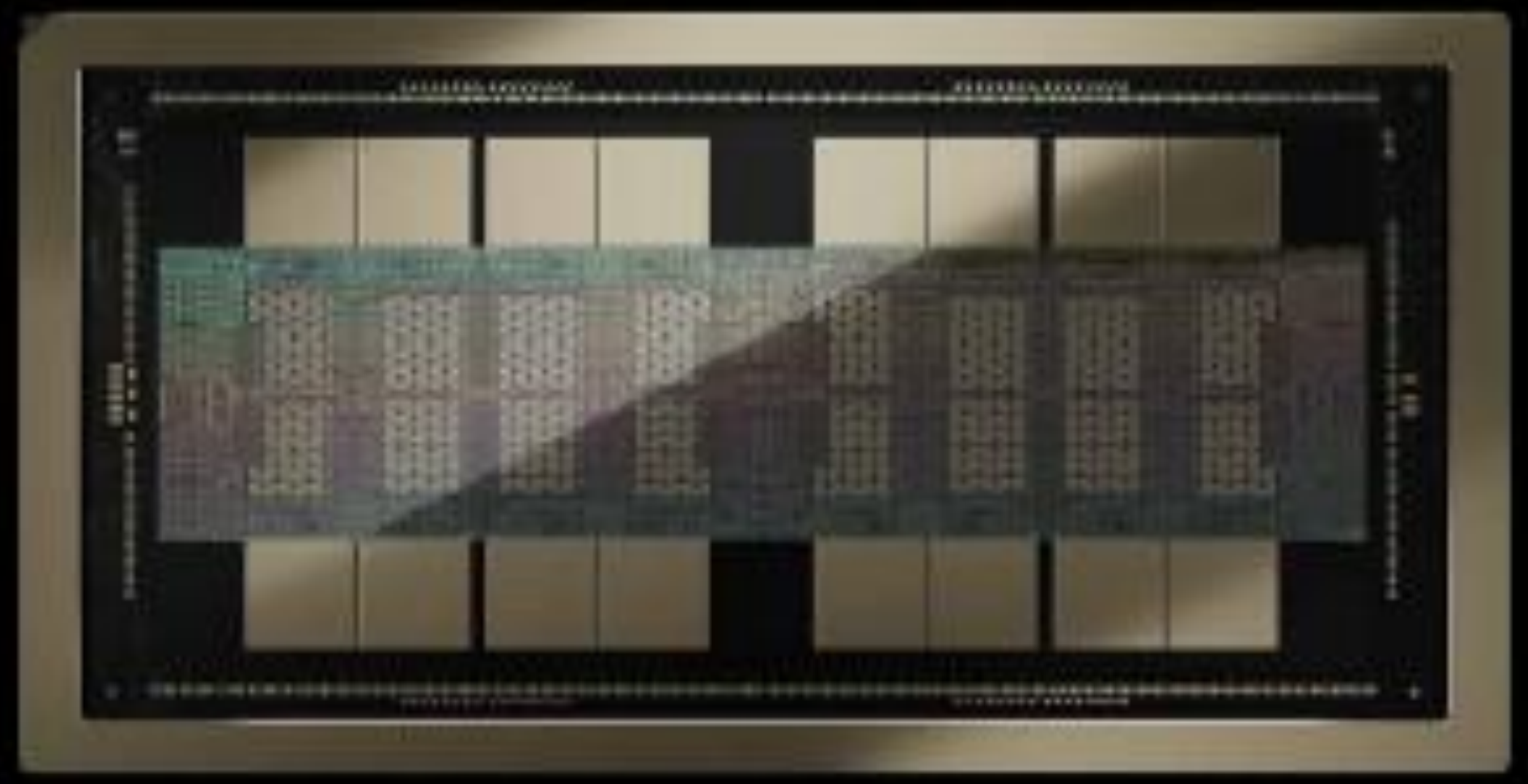
115.2 TB/s CX9
8X

Vera

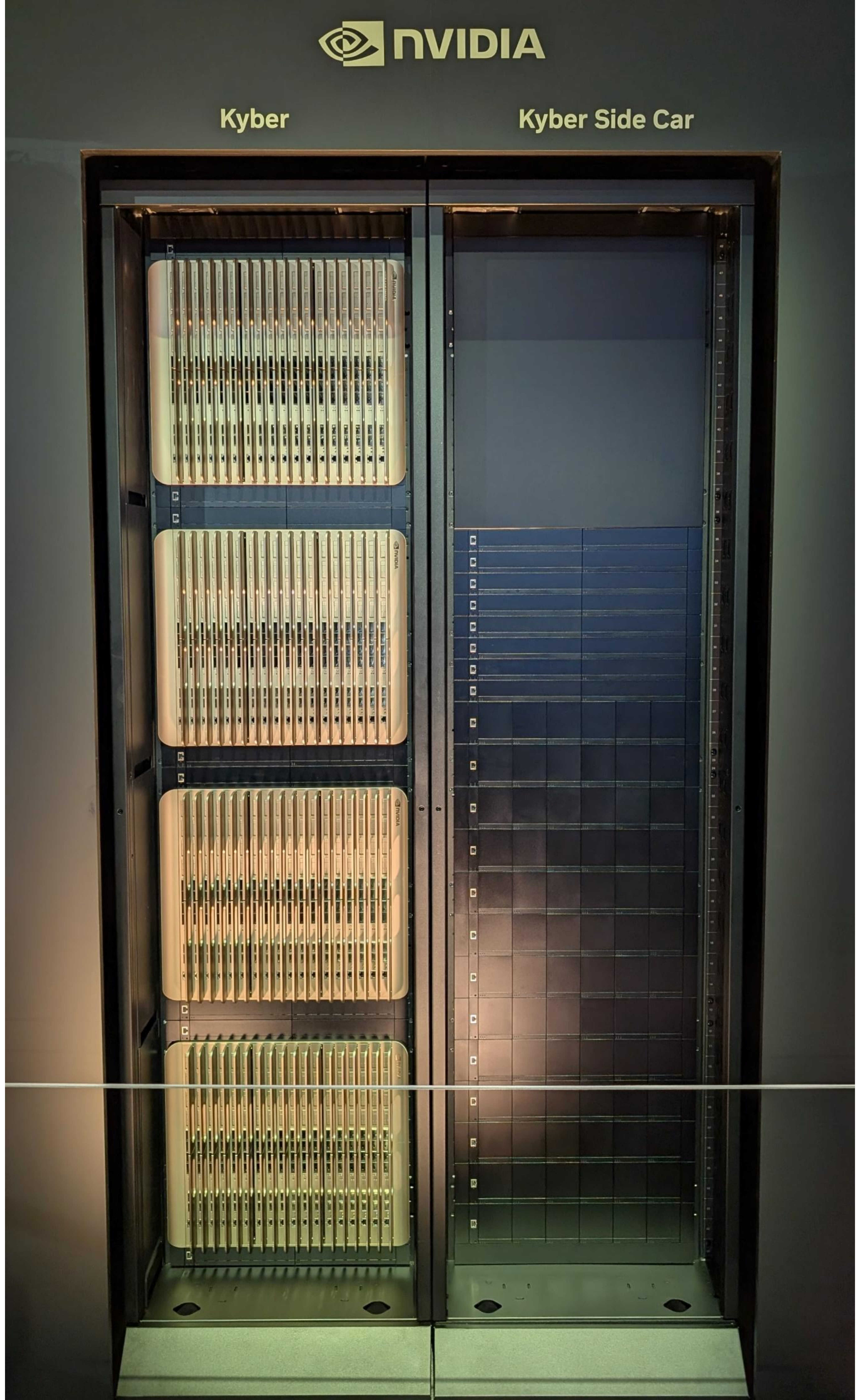


88 Custom Arm Cores
176 Threads
1.8 TB/s NVLink-C2C

Rubin Ultra



4 Reticle-Sized GPUs
100PF FP4 | 1TB HBM4e

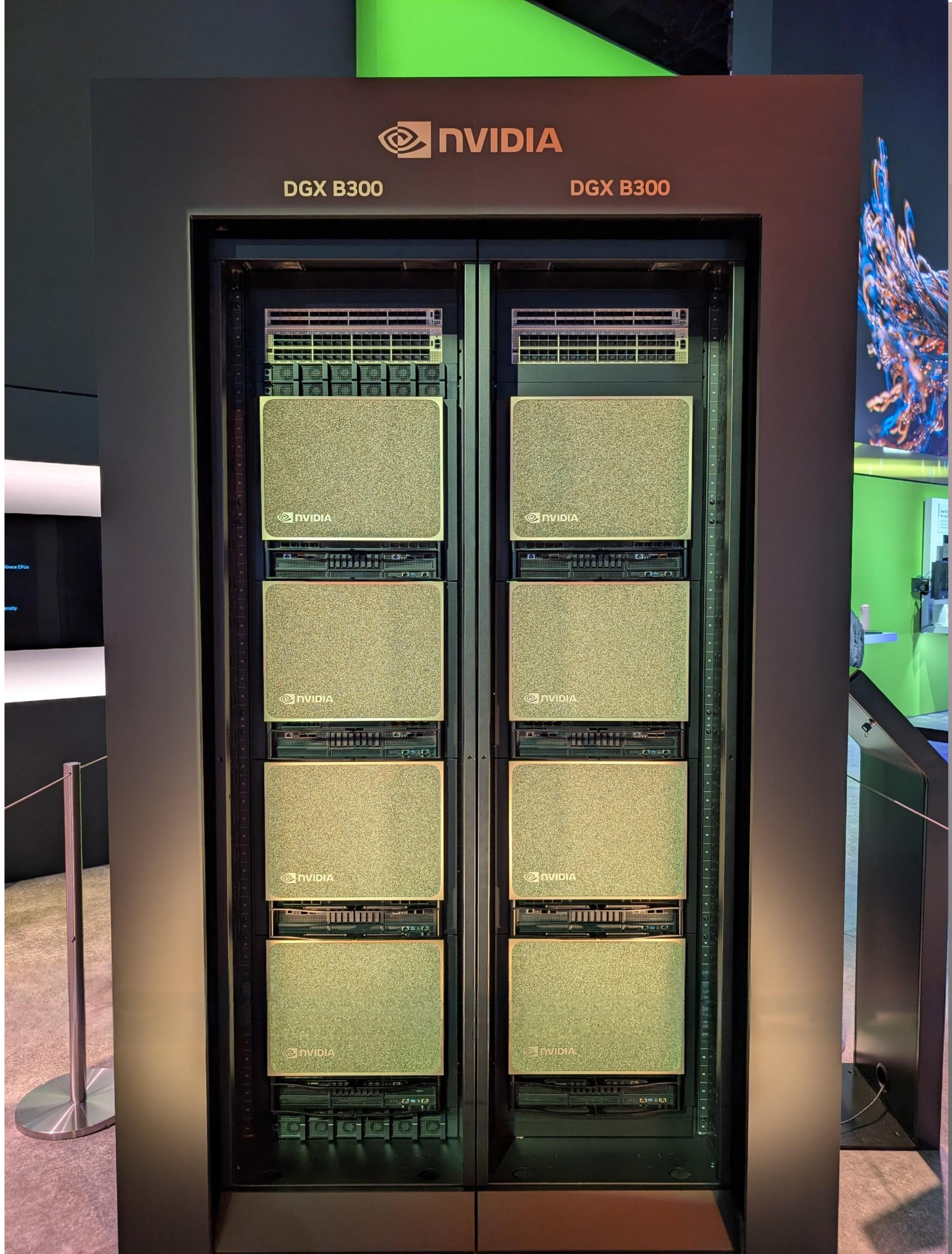

NVIDIA
Kyber
Kyber Side Car

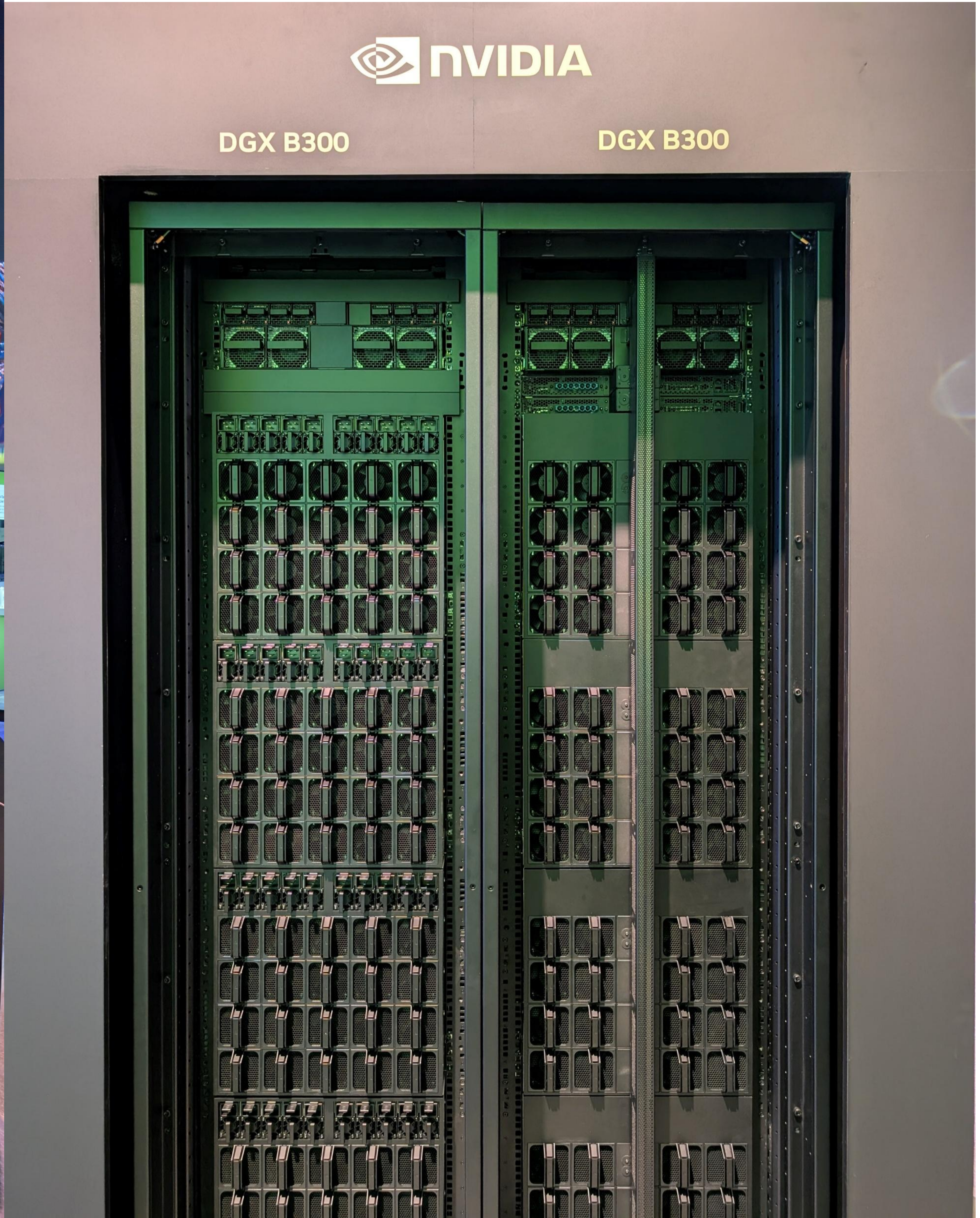
DGX B300

Accelerated Infrastructure for the Era of AI Reasoning

10U Chassis | ~14 kW system
Designed for the modern data center

- Newest air-cooled DGX system with NVIDIA Blackwell Ultra GPUs
- All new system design seamlessly integrates into NVIDIA MGX or traditional enterprise racks
- **2.3TB of GPU memory**, enabling training and inference of complex models
- Equipped with NVIDIA ConnectX-8 high speed networking at **800Gb/s**
- Delivers **72 PFLOPS AI training** and **144 PFLOPS AI inference** performance
- Purpose-built platform for the era of AI reasoning, setting a new bar for LLM inference


NVIDIA
DGX B300
DGX B300


NVIDIA
DGX B300
DGX B300

NVIDIA DGX B200

The foundation of the modern AI data center

10U Chassis | ~14.3 kW system
Deployable in today's data centers

- Air-cooled DGX system with 8X NVIDIA Blackwell GPUs
- **1.4TB of GPU memory**, enabling training of large generative AI models , **64 TB/s Bandwidth**
- **1.8 TB/s** NVLink GPU-to-GPU Bandwidth
- Purpose-built, unified platform for every workload from training, to fine-tuning, to inference
- Delivers **3X AI training** and **15X AI inference** performance as previous generation (DGX H100)

Blackwell Ecosystem

Blackwell Ultra Coming Later 2025

aws Google Cloud Microsoft Azure ORACLE CLOUD Infrastructure SoftBank

Scaleway CoreWeave indosat OOREDOO HUTCHISON Lambda YOTTA YTL COMMUNICATIONS NORTHERN DATA GROUP IBM Cloud APPLIED DIGITAL CRUSOE

DELL Technologies Hewlett Packard Enterprise Lenovo CISCO SUPERMICRO FUJITSU GIGABYTE ASUS Singtel

FOXCONN HON HAI TECHNOLOGY GROUP PEGATRON Inventec AIVRES wiwynn wistron ASRock Rack QCT

GB200 NVL72

HGX B200

HGX B300 NVL16

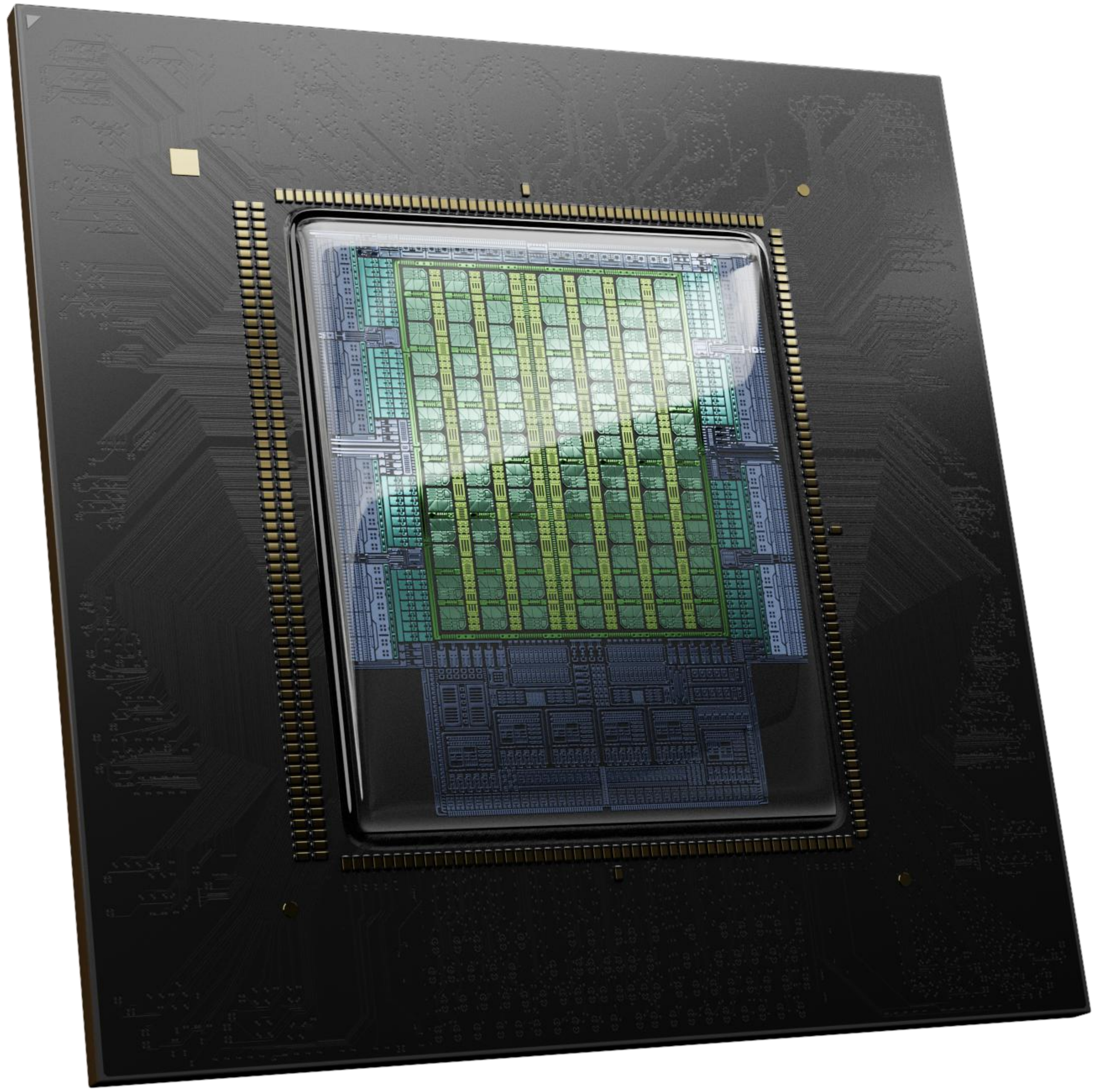
GB300 NVL72

Vera CPU / Networking

Introducing NVIDIA Vera: A Next-Generation CPU

AI Factories, compute and memory intensive CPU workloads

- **>2X CPU Compute Capability, 2.4x threads vs. Grace,**
  - 88 Cores with Spatial Multi-Threading
- **5X Memory BW per Watt**
  - Memory power per socket under 50W vs 280W MRDIMM
- **>3x Memory Capacity**
  - 1.5 TB of coherent LPDDR5X in Vera Rubin platforms
  - 2 TB of DDR5
- **>2x Bisection Bandwidth vs. x86**
  - Single NUMA design for optimal tuning out-of-box
- **7x Faster GPU connectivity vs. traditional CPU**
  - 1.8 TB/s NVLink-C2C CPU:GPU bandwidth vs PCIe Gen 6



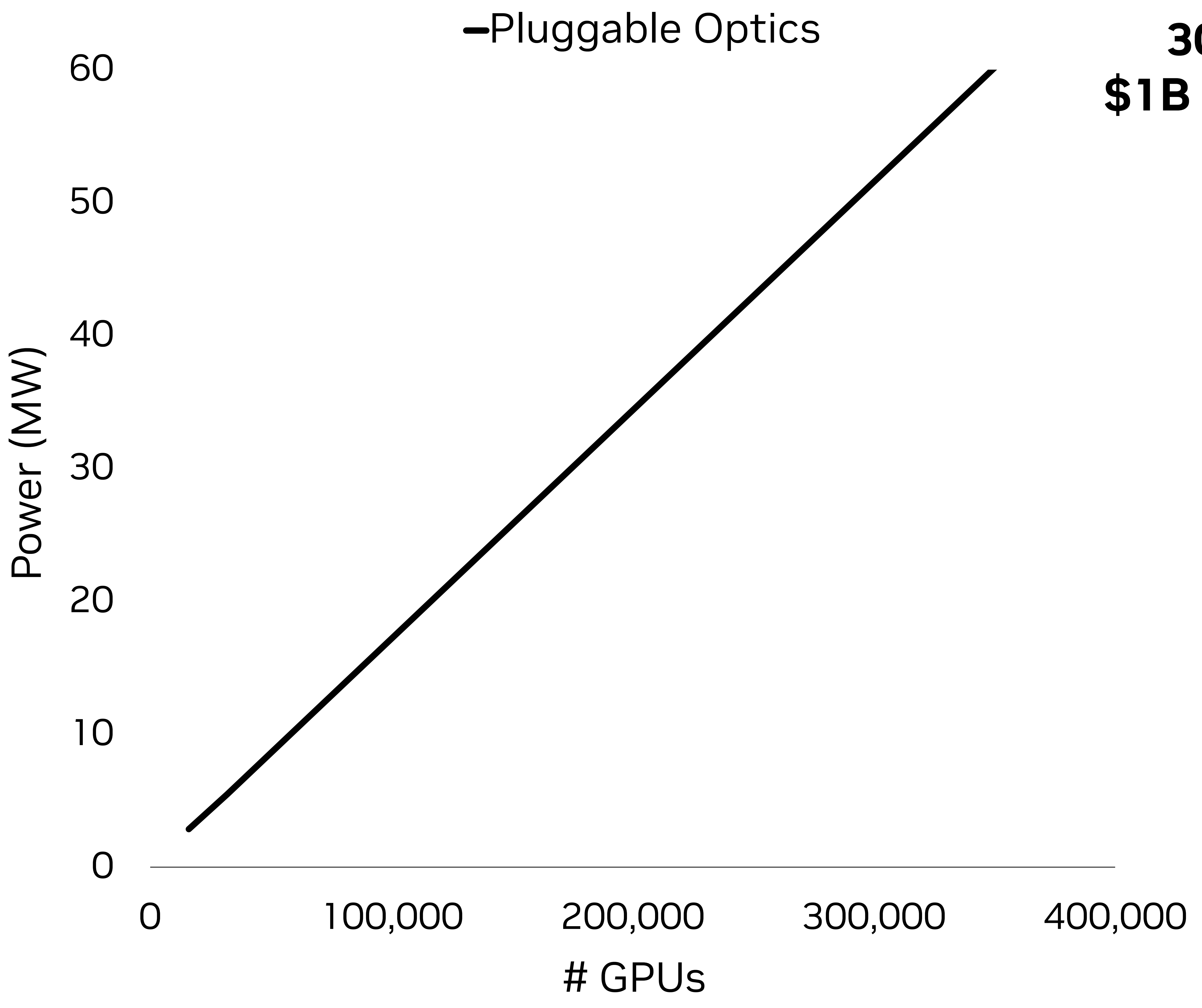
Projected performance subject to change

Spectrum-X Ethernet Accelerates the World's Largest AI Supercomputers

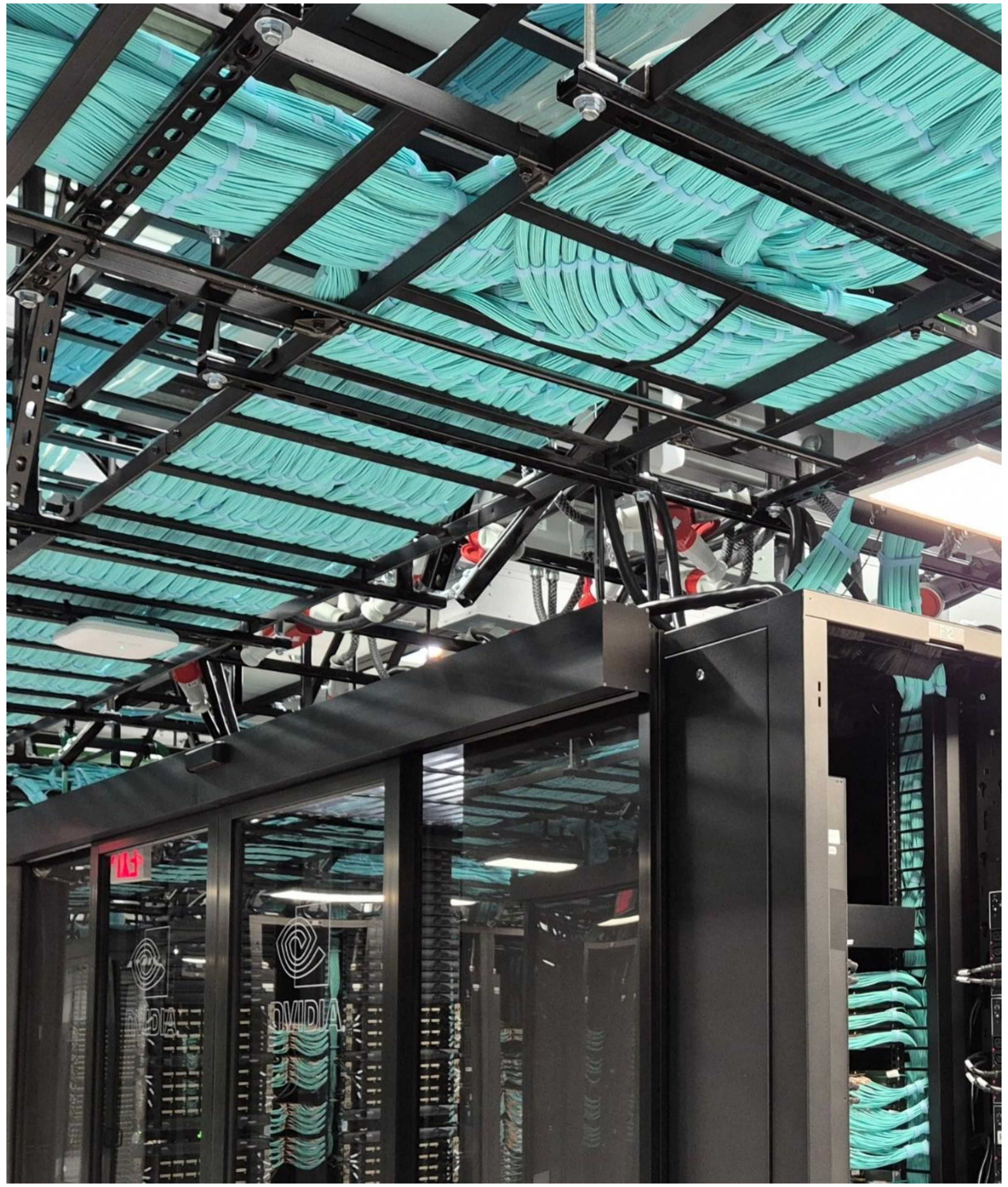
Scaling Compute to 400K GPU AI

100K	400K	2.4M	40 MW
Servers	GPUs	Optical Transceivers	Transceiver Power

Power and Reliability Challenges of AI Scale-Out and Density



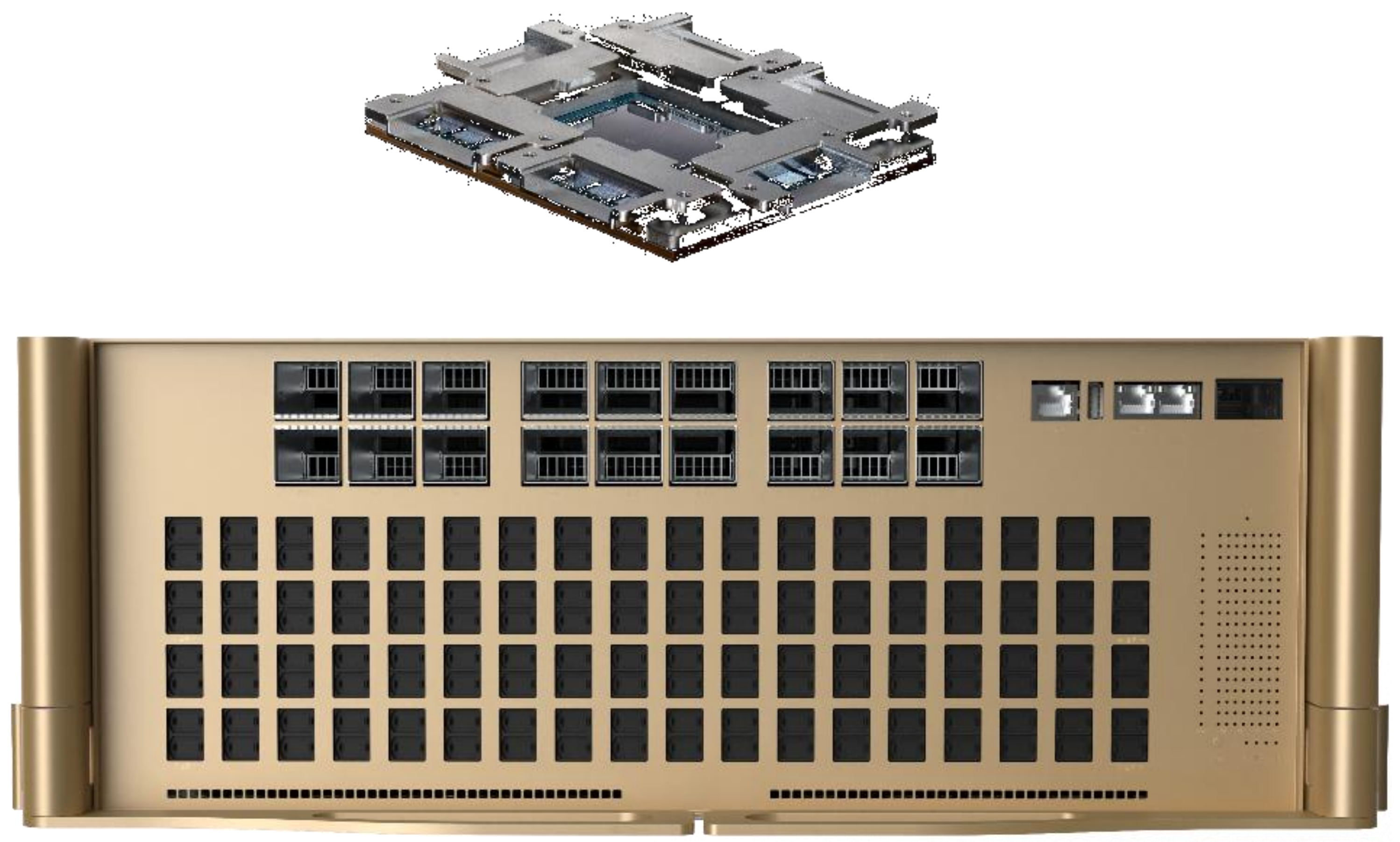
60MW of Switch Optics Power
=
30K GPUs
$1B GPU Oppty



Announcing NVIDIA Photonics Switch Systems

Co-packaged optics networking switches to scale AI factories to millions of GPUs

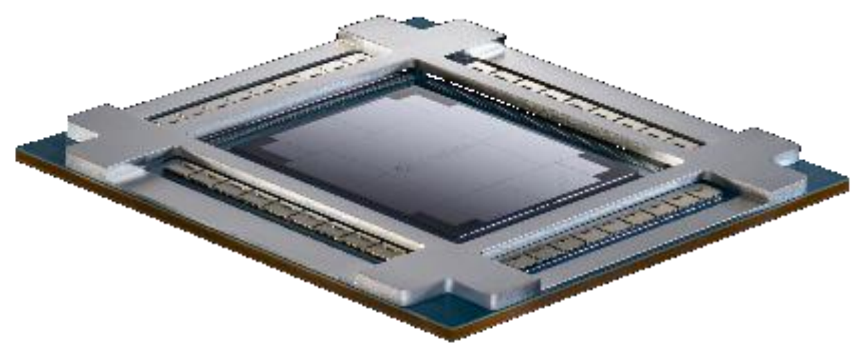
Quantum-X Photonics



Quantum 3450-LD

115Tb/s
144 ports of 800G
(576 ports x 200G)
Liquid cooled

Spectrum-X Photonics



Spectrum SN6810

102.4Tb/s
128 ports of 800G
(512 x 200G)
Liquid cooled

Spectrum SN6800

409.6Tb/s
512 ports of 800G
(2048 x 200G)
Liquid cooled

3.5X
Power efficiency

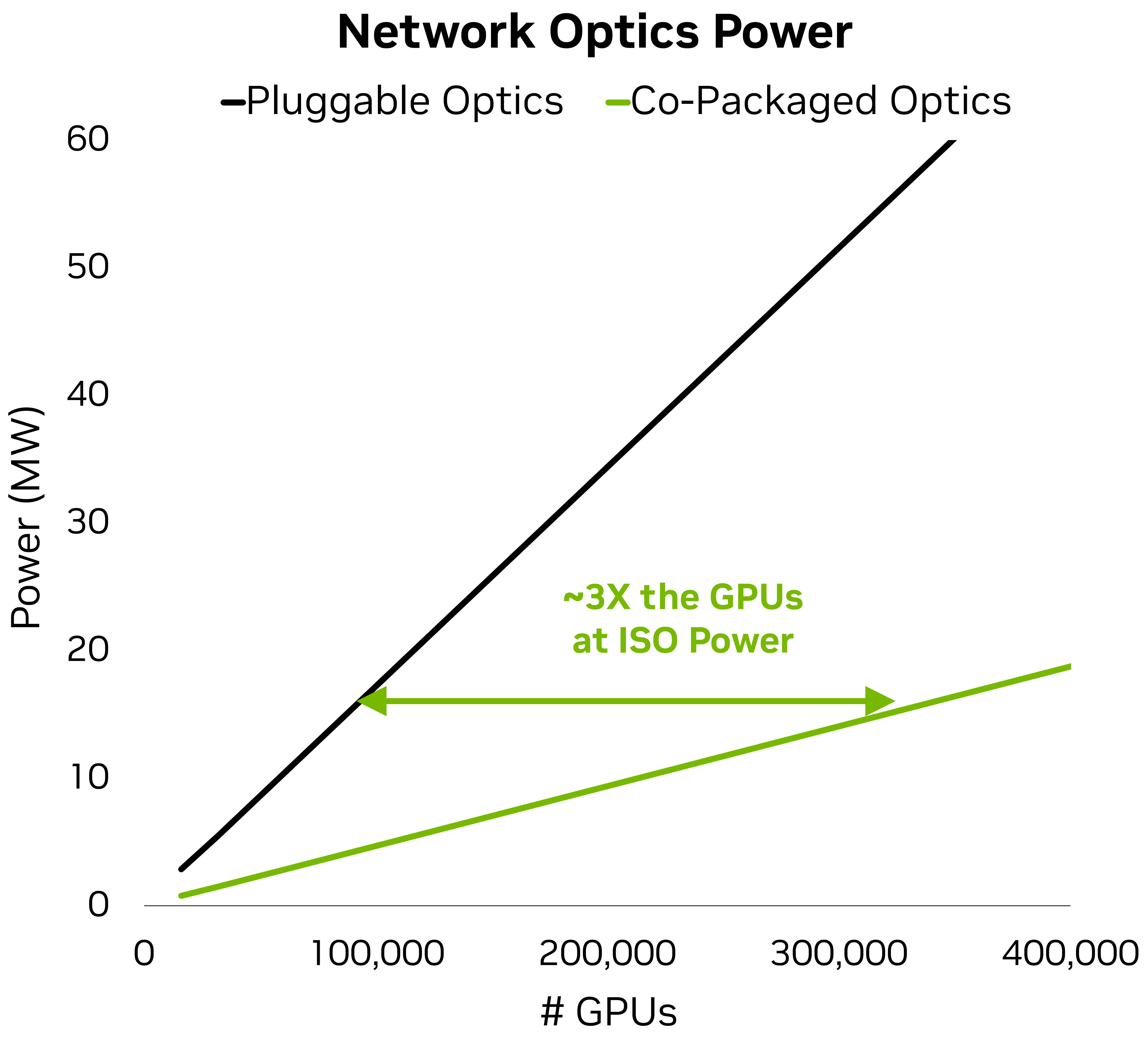
10X
Higher resiliency

1.3X
Faster time to deploy

Projected performance subject to change

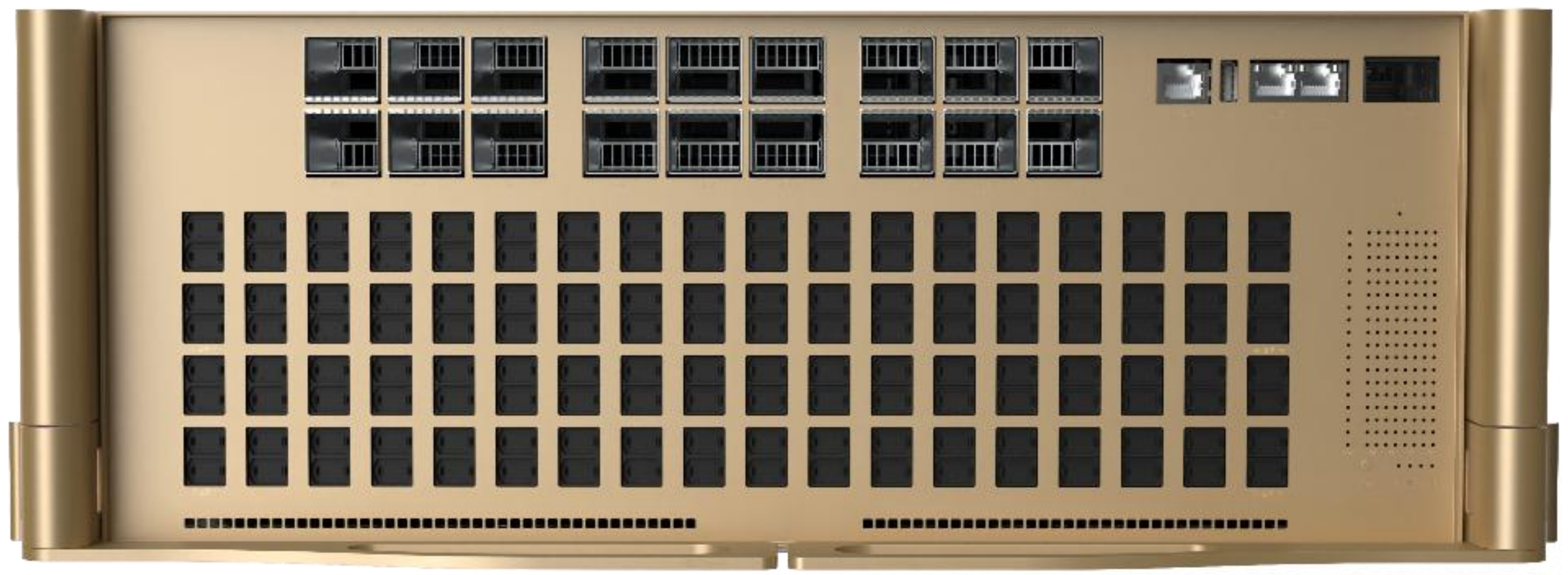
NVIDIA Photonics Solves Power and Reliability Challenges of AI Scale-Out

Co-packaged silicon photonics networking switches to scale AI factories to millions of GPUs



72 Transceivers Replaced

432 Fewer Lasers



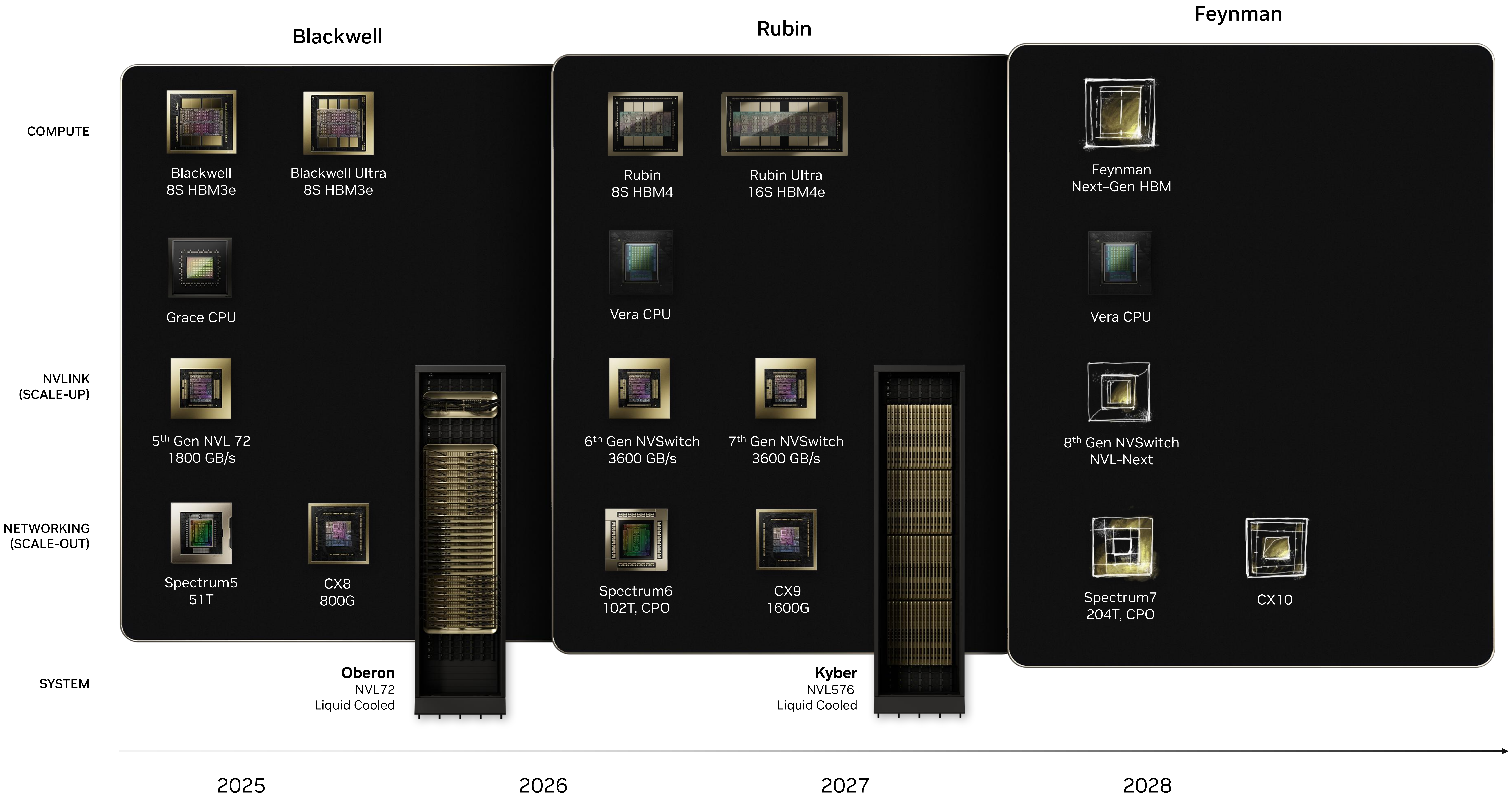
3.5X Power efficiency

10X Higher resiliency

1.3X Time to operation

NVIDIA Paves Road to Gigawatt AI Factories

One-Year Rhythm | Full-Stack | One Architecture | CUDA Everywhere



COMPUTE

NVLINK (SCALE-UP)

NETWORKING (SCALE-OUT)

SYSTEM

Blackwell

Blackwell 8S HBM3e

Blackwell Ultra 8S HBM3e

Grace CPU

5th Gen NVL 72 1800 GB/s

Spectrum5 51T

CX8 800G

Oberon
NVL72
Liquid Cooled

Rubin

Rubin 8S HBM4

Rubin Ultra 16S HBM4e

Vera CPU

6th Gen NVSwitch 3600 GB/s

7th Gen NVSwitch 3600 GB/s

Spectrum6 102T, CPO

CX9 1600G

Kyber
NVL576
Liquid Cooled

Feynman

Feynman Next-Gen HBM

Vera CPU

8th Gen NVSwitch NVL-Next

Spectrum7 204T, CPO

CX10

2025

2026

2027

2028

DGX Spark

NVIDIA

Fine-tuning DeepSeek-R1-Distill-Qwen-32B
On DGX Spark using the NV-Code dataset
Installation and Setup
Model Setup
Welcome to NVIDIA DGX Spark

NVIDIA®