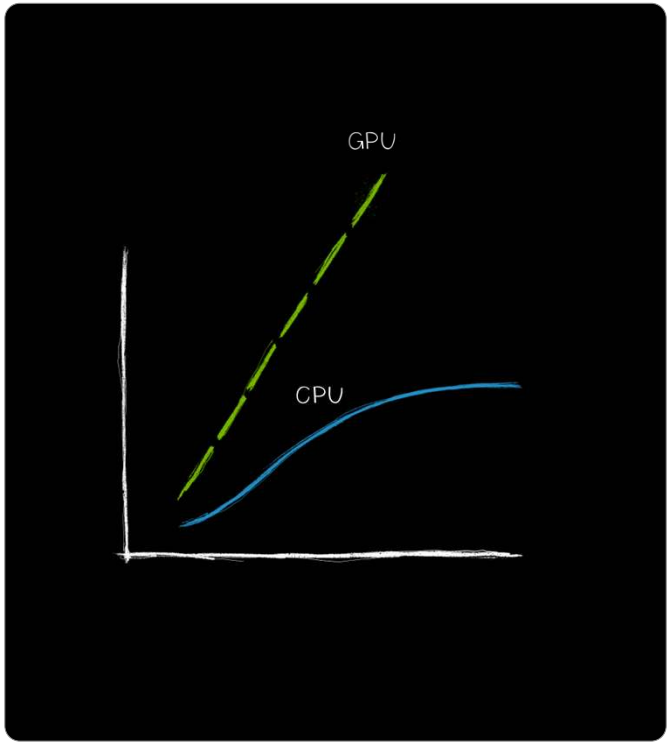
AIおよびコンピューティングの最新トレンド

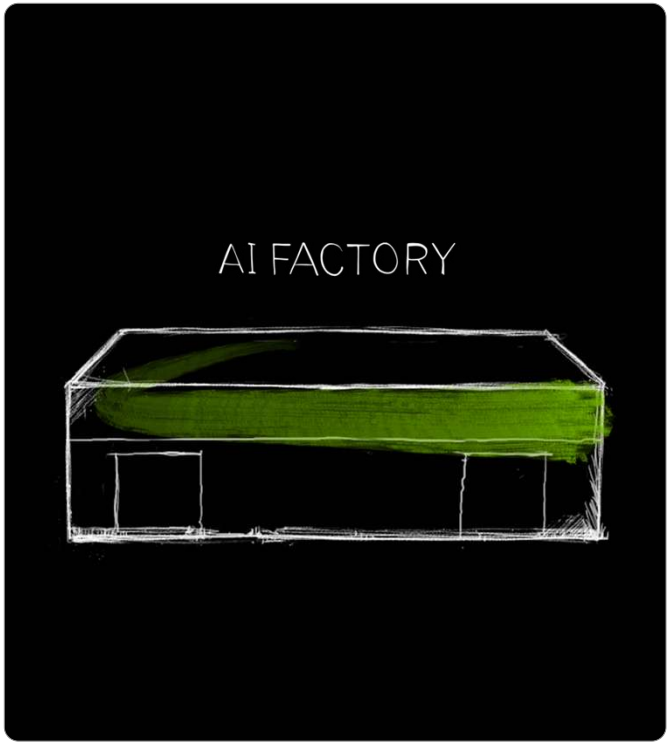
Hiroshi Aiko, Sr. Marketing Manager, NVIDIA

PCCC Workshop 柏 2025, Jun 27,2025

Three Major Trends in Computing



Traditional → Accelerated

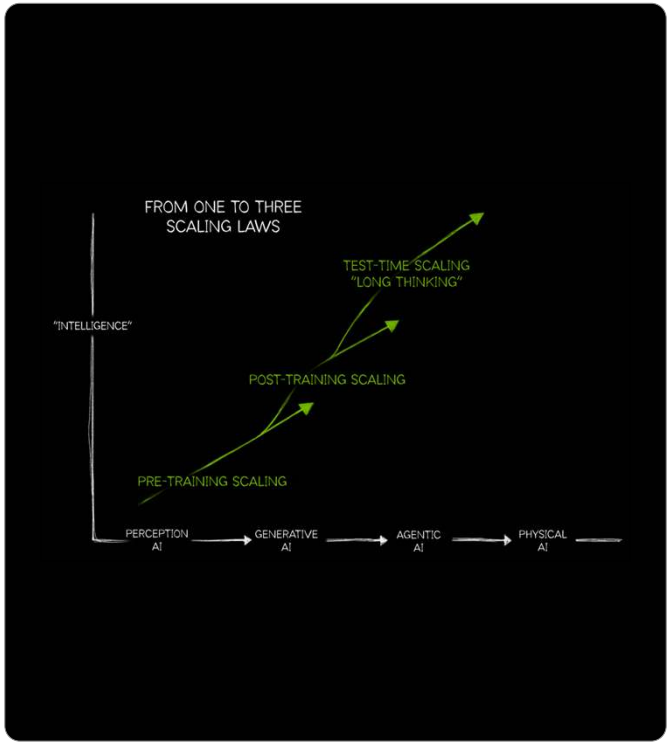


Data Centers → AI Factories

Generative AI → Physical AI

NVIDIA

AI Factory Specialized Infrastructure Challenges

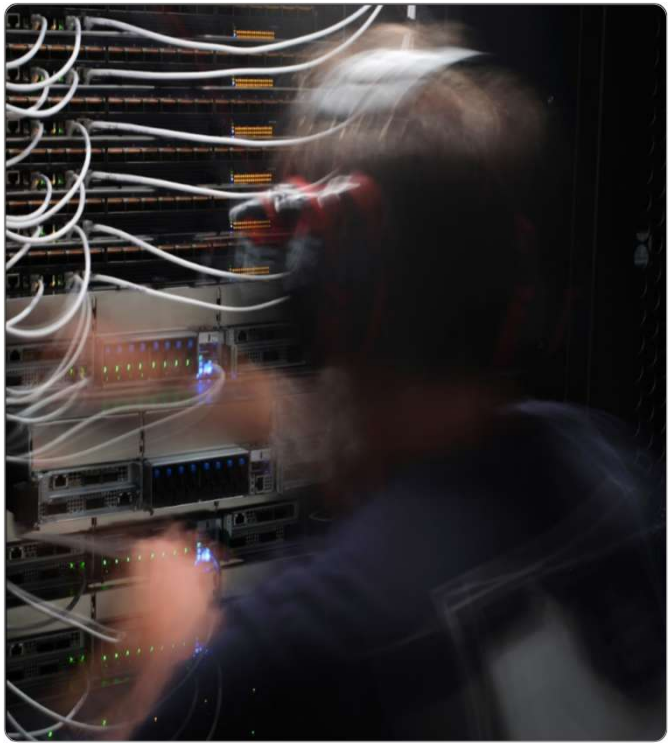


Increasing Compute Demands

Requires proven scale-up performance

Increasing Scale-up Demands

Requires fast-paced, rack-scale roadmaps



Increasing Infrastructure Complexity

Requires Rich Ecosystem and Common Infrastructure

Grace Blackwell NVL72
“A Thinking Machine”

Increasing AI Reasoning
Throughput and Responsiveness

50X
AI Factory Output

Blackwell 288GB GPU

1.5x more FP4 Inference

Upgraded NVL72 Design for Improved Energy Efficiency

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

NVIDIA

Grace Blackwell Ultra Superchip

40 PFLOPS FP4 TE

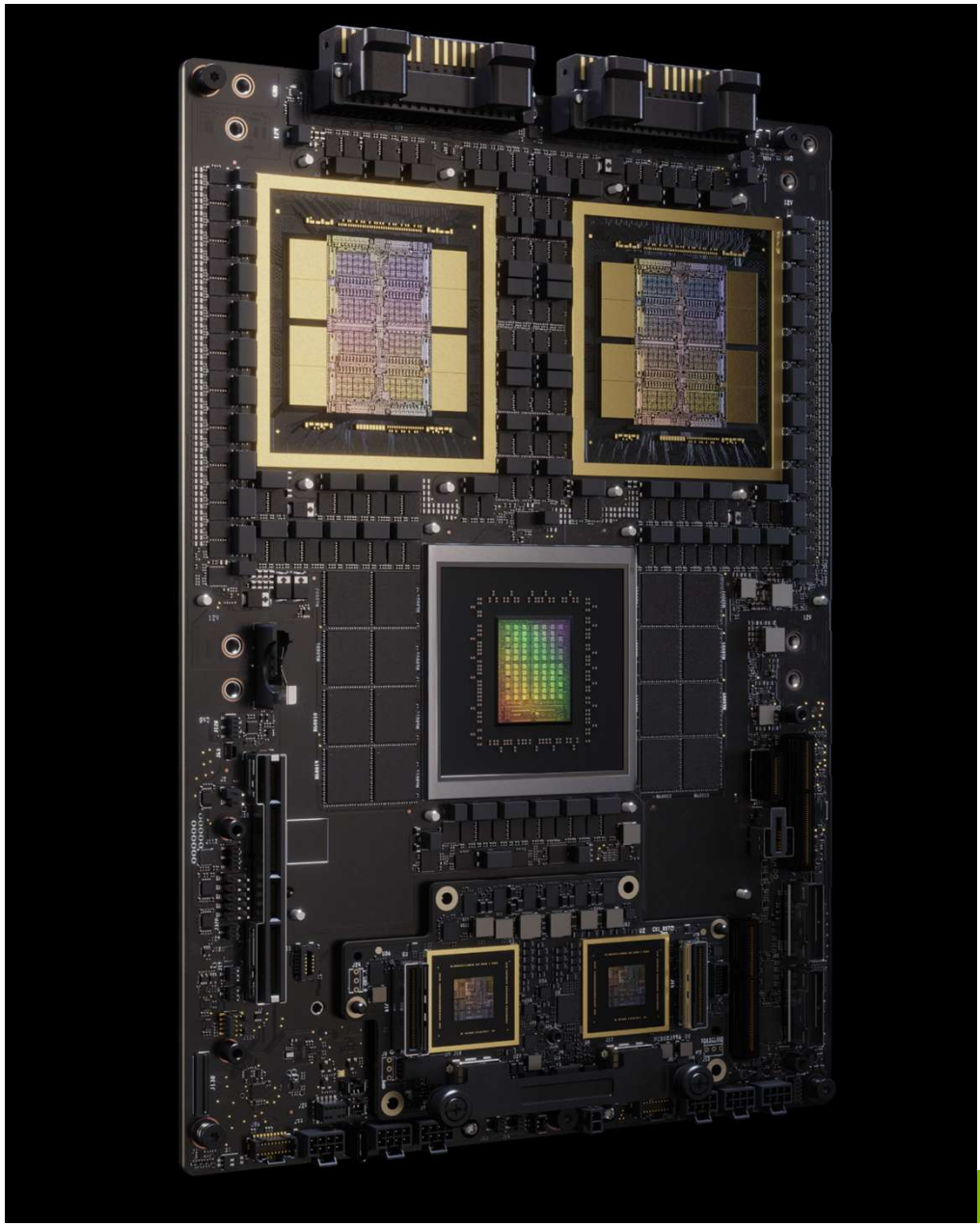
576GB at 16 TB/sec HBM3e

3.6TB/sec NVLink5

72 Core Grace CPU

480 GB LPDDR5X

800 Gb/sec CX8 x2



NVIDIA Dynamo の発表

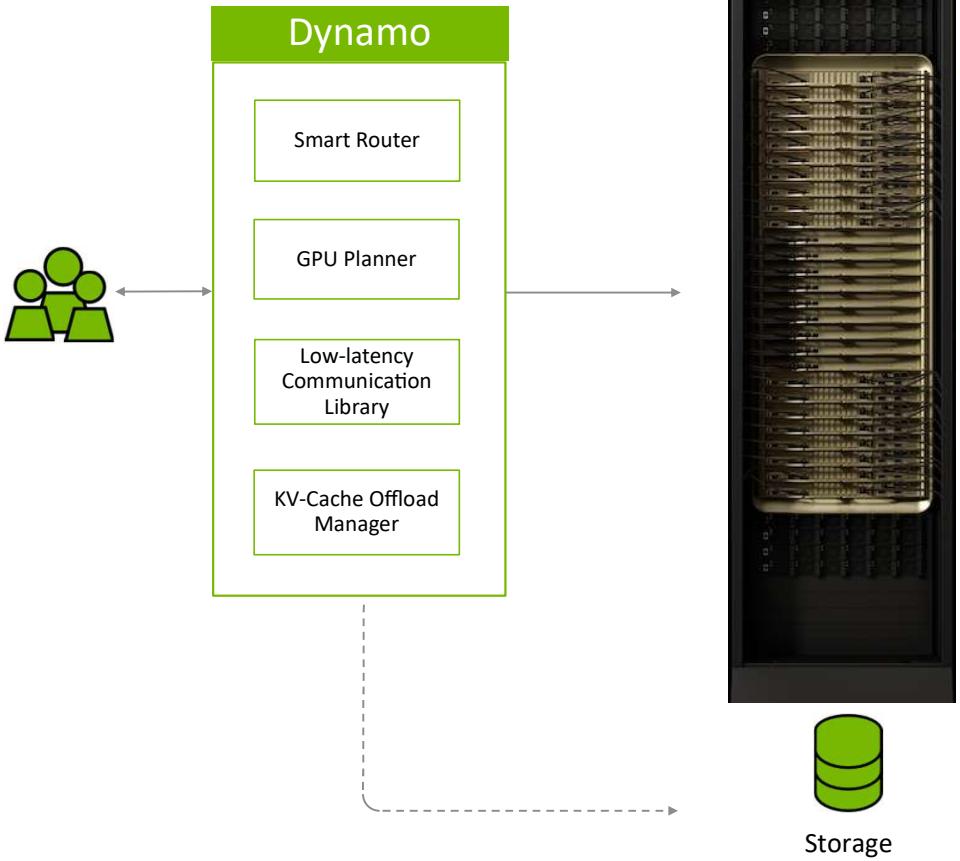
大規模なリーズニングのための推論ソフトウェア

Distributed and Disaggregated Serving

30X
AI ファクトリー
スループット &
収益
DeepSeek モデル

2.5X
AIファクトリー
スループット &
収益
Llama モデル

1000+
GPUスケール
単一のクエリ



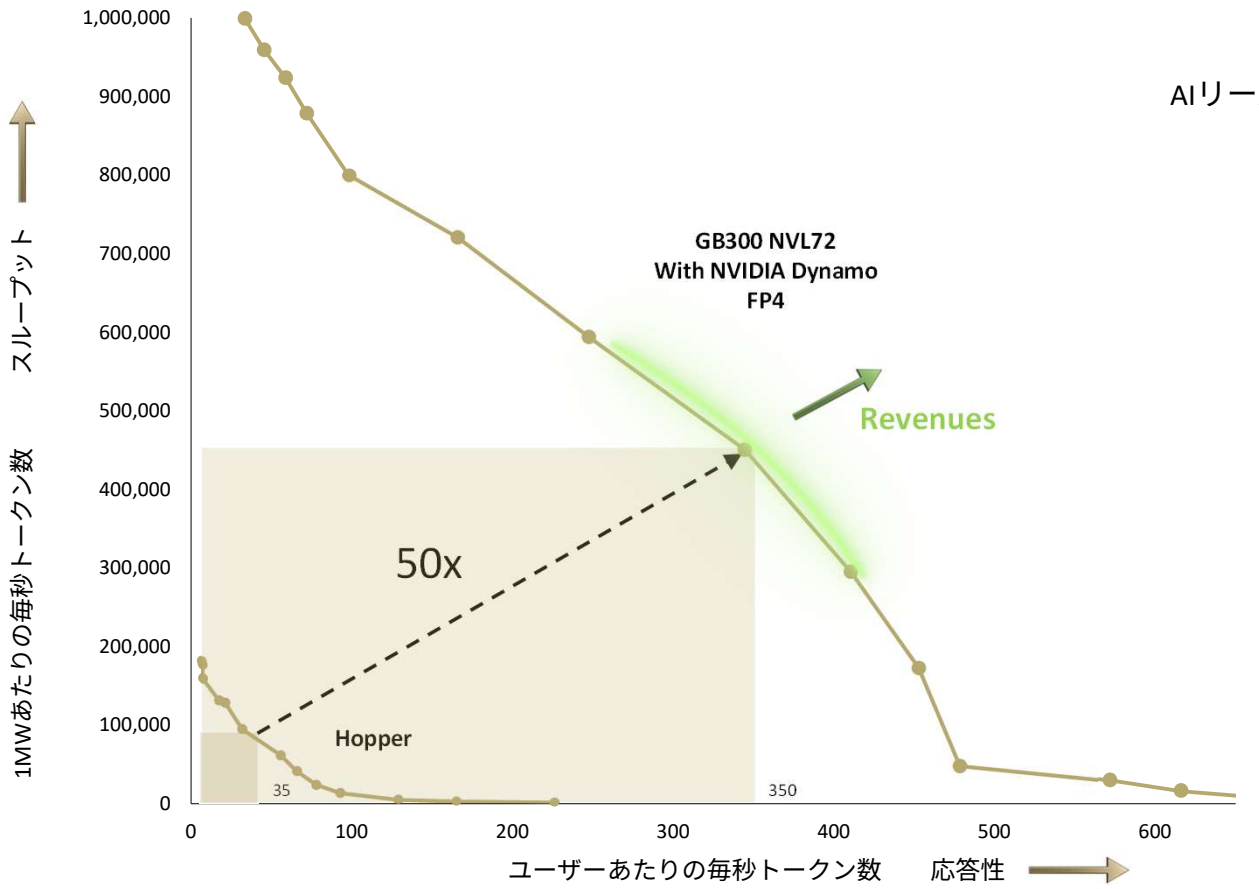
完全なオープンソースとオープンなバックエンド

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.

NVIDIA

強化されたAIファクトリは利益に貢献する



AIリーズニングのスループット強化
応答性
収益

50X
AI ファクトリ 出力

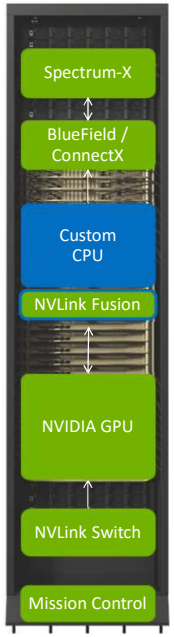
10X
応答性

5X
電力の
スループット

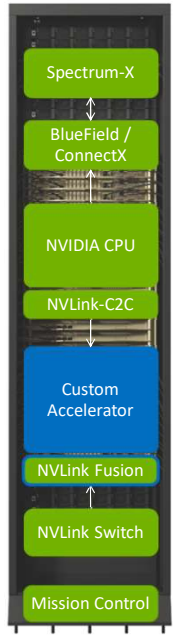
DeepSeek R1 ISL = 32K, OSL = 8K, GB300 NVL72 with FP4 Dynamo disaggregation. H100 with FP8 In-flight batching. Projected performance subject to change.

Announcing
NVIDIA NVLink Fusion

Semi-Custom AI Infrastructure



NVLink Fusion
+ Custom CPU



NVLink Fusion
+ Custom Accelerator

NVIDIA

NVIDIA NVLink Fusion

Semi-Custom AI Infrastructure

NVIDIA Plus ASIC Hybrid AI Infrastructure

Proven Scale-Up and Scale-Out Ecosystem

Single Scalable AI Factory Architecture

alchip · AsteraLabs. · MARVELL · MEDIATEK · FUJITSU · Qualcomm · cadence · SYNOPSYS

One Architecture – from Cloud AI, Enterprise AI, Personal AI, to Edge AI

Arm Platforms

x86 Platforms

Announcing NVIDIA DGX Station

One trillion parameter models on personal AI machine

DGX Spark

The world's smallest AI supercomputer

GB10 Superchip | 128GB Unified system memory | 1,000 AI TOPS
NVIDIA ConnectX-7 SmartNIC

Reserve systems on NVIDIA.com Marketplace

acer ASUS DELL Technologies GIGABYTE hp Lenovo msi

DGX Station

The ultimate desktop computer for the AI era

20 PFLOPS AI performance | 800 GB Unified system memory
| NVIDIA ConnectX-8 SuperNIC | 800 Gb/s network connection

ASUS BOXX DELL Technologies GIGABYTE hp Lambda msi SUPERMICRO

NVIDIA