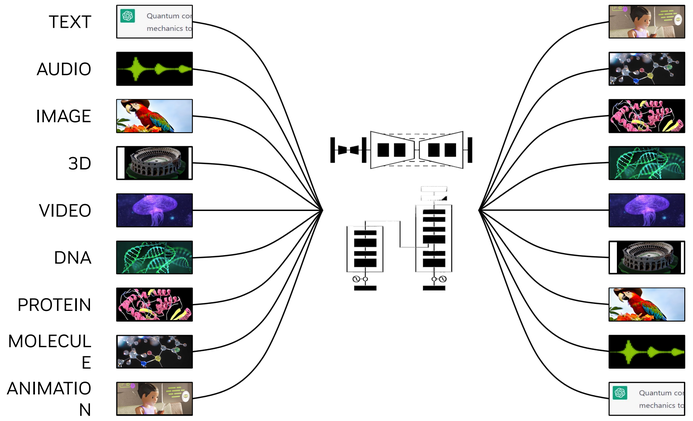
The Next Era of Generative AI

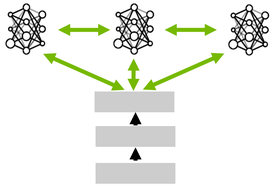
Image Classification

Transformer

Large Language Models (Transformer)

Large Language Models (Transformer)



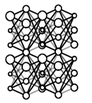


Google
Gemini

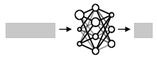
Meta
NLLB

MISTRAL AI_
Mixtral

Realtime
<50ms latency



Parameters
>10T



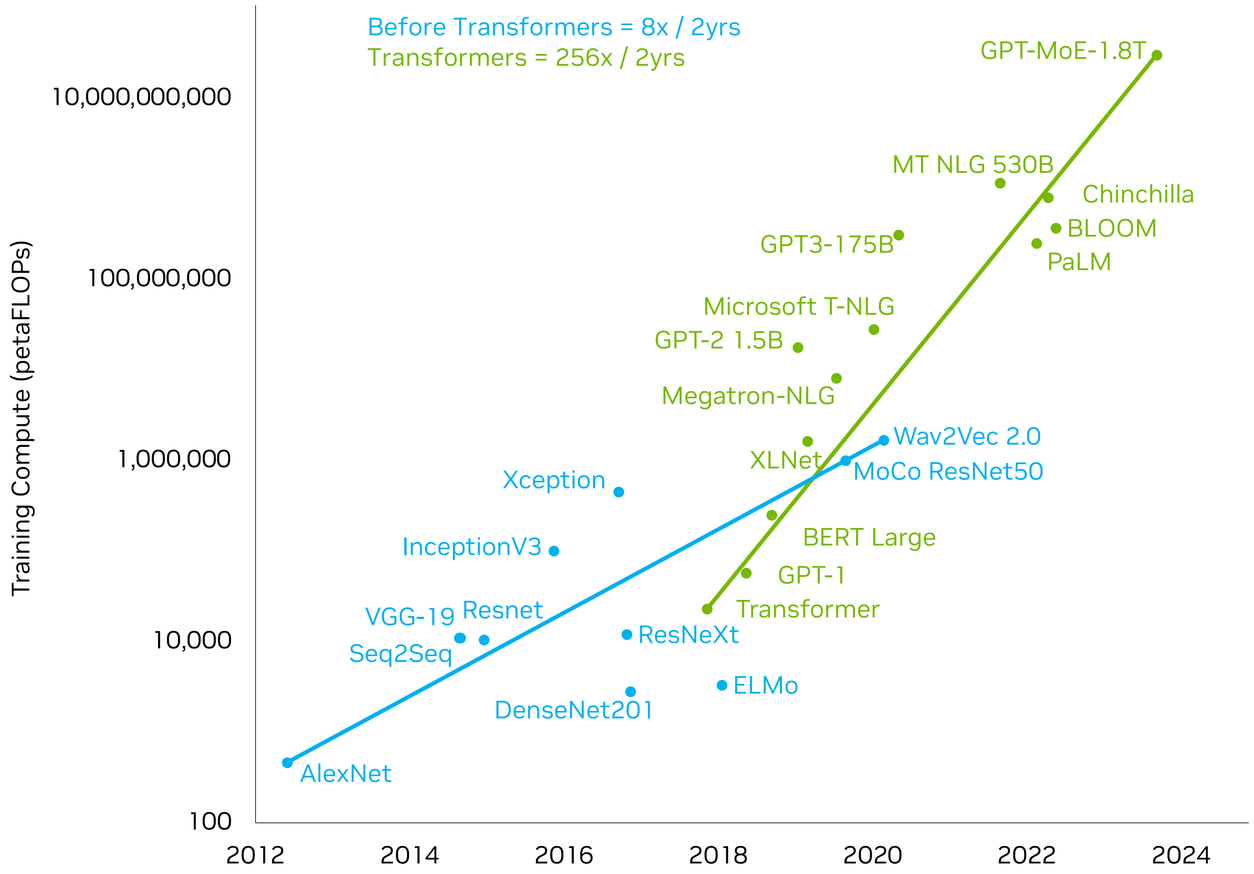
Sequence Length
>32K input

Labeled Datasets → Unlabeled Datasets → Generative AI → Multimodal Generative AI → Production GenAI Inference

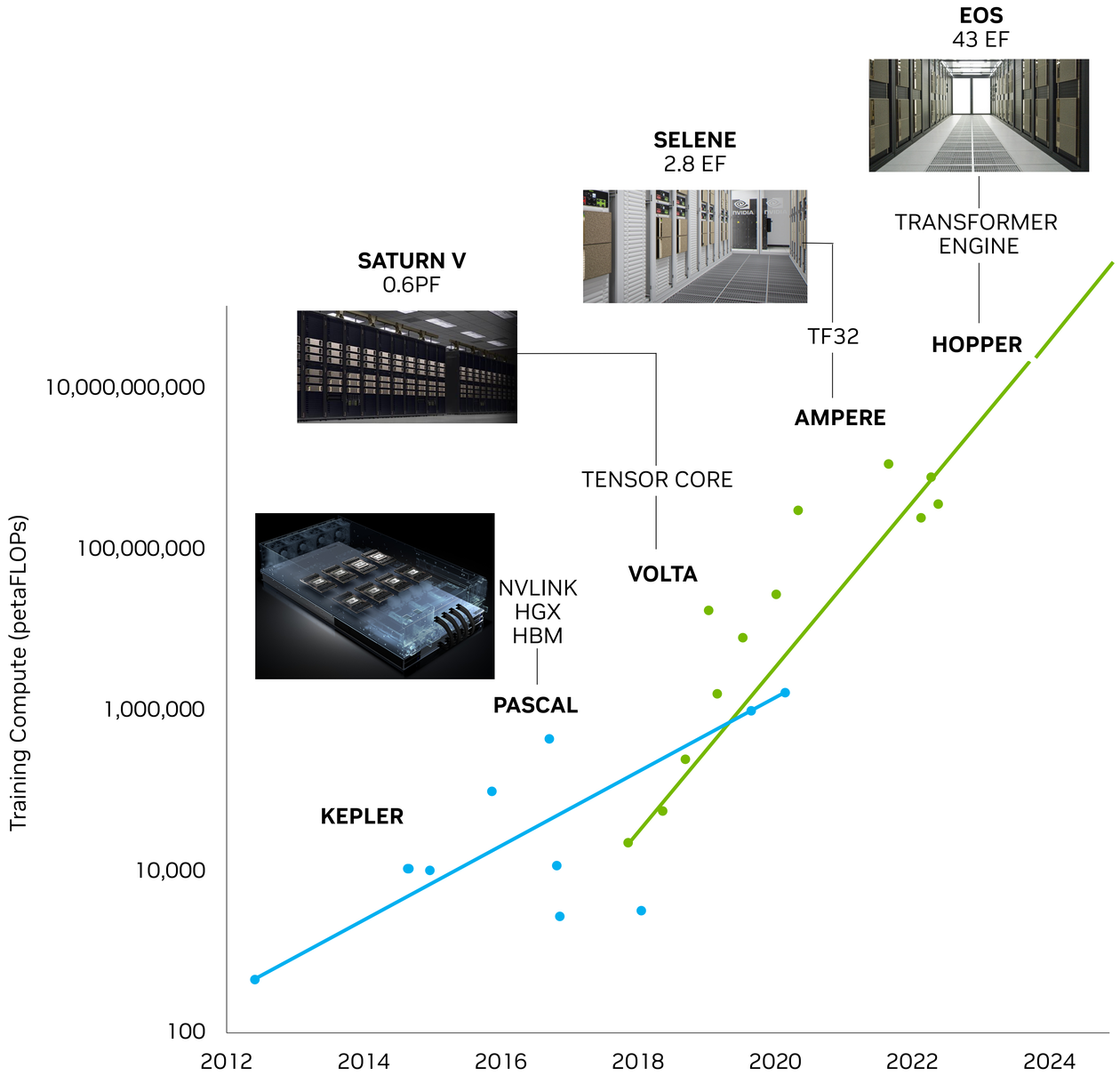
Mixture of Experts (MoE)

NVIDIA

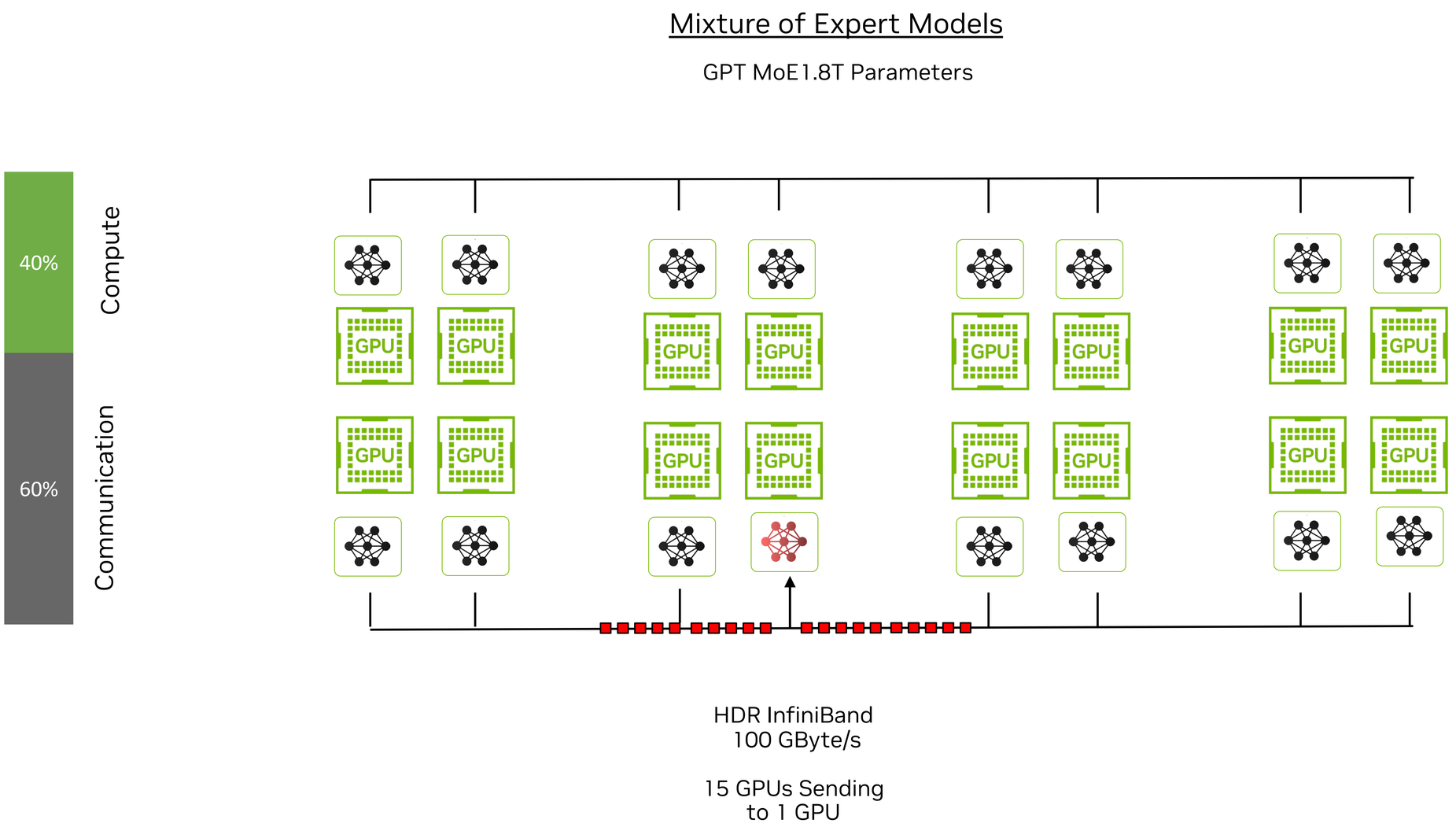
Explosive Growth in AI Computational Requirements



NVIDIA Enables Explosive Growth in AI Computational Requirements

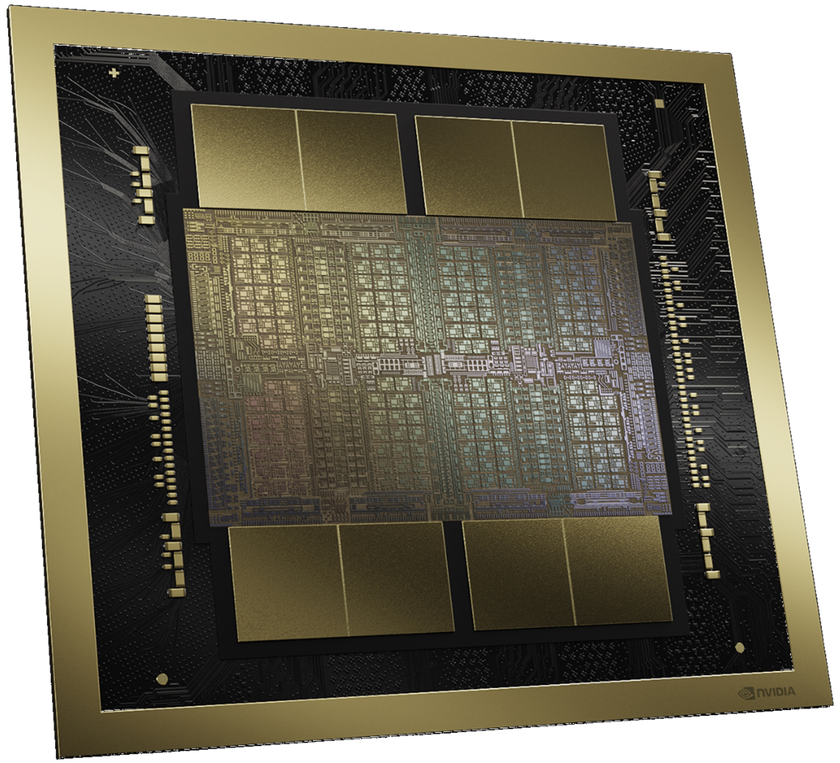


Next Generation Models Communication Bottleneck



Announcing NVIDIA Blackwell

The Engine of the New Industrial Revolution



Built to Democratize Trillion-Parameter AI

20 PetaFLOPS of AI performance on a single GPU

4X Training | 30X Inference | 25X Energy Efficiency & TCO

Expanding AI Datacenter Scale to beyond 100K GPUs

AI SUPERCHIP
208B Transistors

2nd GEN TRANSFORMER ENGINE
FP4/FP6 Tensor Core

5th GENERATION NVLINK
Scales to 576 GPUs

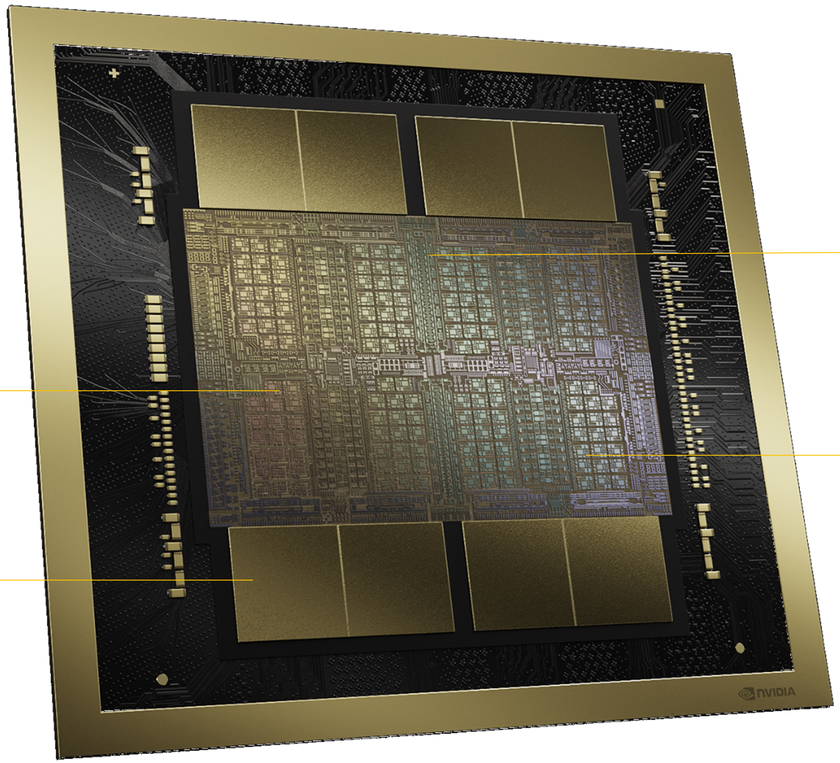
RAS ENGINE
100% In-System
Self-Test

SECURE AI
Full Performance
Encryption & TEE

DECOMPRESSION ENGINE
800 GB/s

New Class of AI Superchip

The Two Largest Dies Possible—Unified as One GPU



Reticle-sized Die 1

Fast Memory
192GB HBM3e

2 reticle-limited dies operate as One Unified CUDA GPU

NV-HBI 10TB/s High Bandwidth Interface

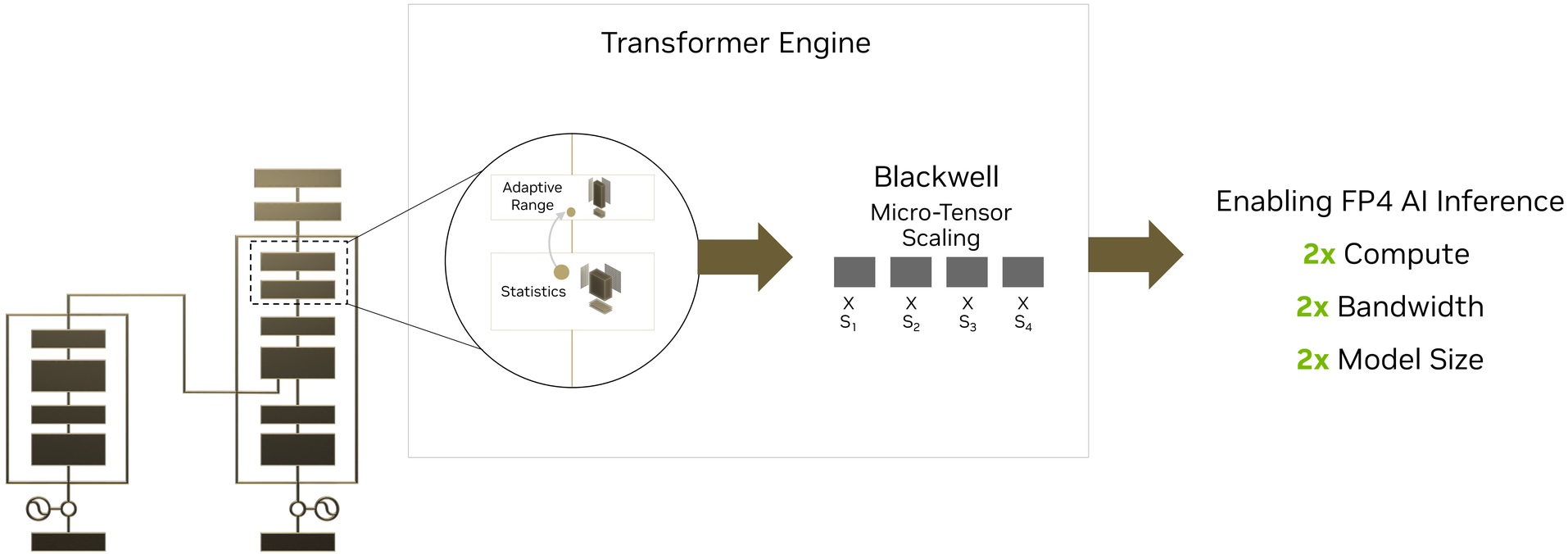
Full performance. No compromises

Reticle-sized Die 2

10 PetaFLOPS FP8 | 20 PetaFLOPS FP4
192GB HBM3e | 8 TB/sec HBM Bandwidth | 1.8TB/s NVLink

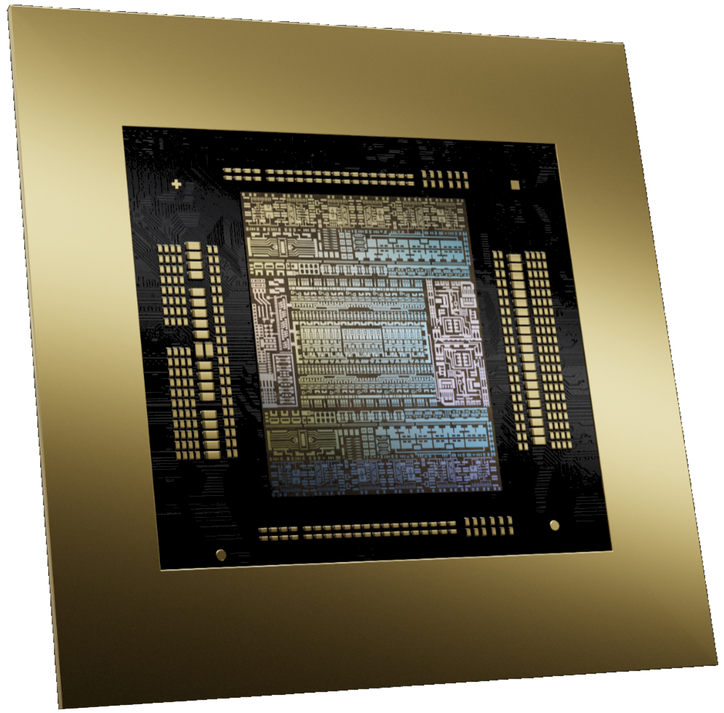
2nd Generation Transformer Engine

Accelerating Throughput with Intelligent 4-Bit Precision



Announcing Fifth Generation NVLink and NVLink Switch Chip

Efficient Scaling for Trillion Parameter Models



7.2 TB/s Full all-to-all Bidirectional Bandwidth

Sharp v4 plus FP8

3.6 TF In-Network Compute

Expanding NVLink up to 576 GPU NVLink Domain

18X Faster than Today’s Multi-Node Interconnect

Announcing GB200 NVL72

Delivers New Unit of Compute

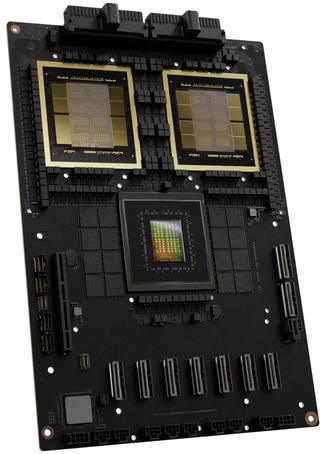
GB200 NVL72

- 36 GRACE CPUs
- 72 BLACKWELL GPUs
- Fully Connected NVLink Switch Rack

Training FP8	720 PFLOPs
Inference FP4	1,440 PFLOPs
NVL Model Size	27T params
Multi-Node All-to-All	130 TB/s
Multi-Node All-Reduce	260 TB/s

GB200 NVL72 Compute and Interconnect Nodes

Building Blocks for the GB200 NVL72 Rack



GB200 SUPERCHIP

40 PETAFLOPS FP4 AI INFERENCE
20 PETAFLOPS FP8 AI TRAINING
864GB FAST MEMORY

GB200 SUPERCHIP COMPUTE TRAY

2x GB200
80 PETAFLOPS FP4 AI INFERENCE
40 PETAFLOPS FP8 AI TRAINING
1728 GB FAST MEMORY
1U Liquid Cooled
18 Per Rack

NVLINK SWITCH TRAY

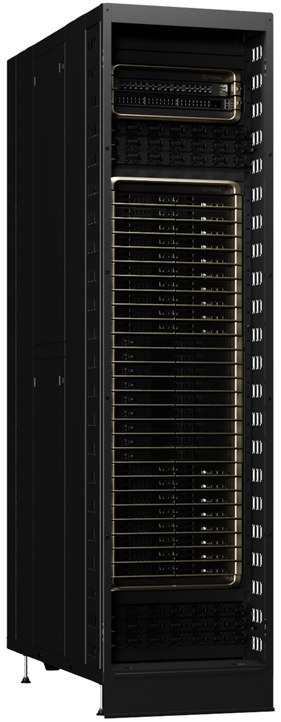
2x NVLINK SWITCH CHIP
14.4 TB/s Total Bandwidth
SHARPv4 FP64/32/16/8
1U Liquid Cooled
9 Per Rack

DGX GB200 NVL72
1 Giant GPU

Training FP8	720 PFLOPS	**22X**
Inference FP4	1.44 ExaFLOPS	**45X**
Multi-Node All-to-All	130 TB/sec	**18X**
Multi-Node All-Reduce	260 TB/sec	**36X**

Blackwell for Every Generative AI Use Case

Delivering the New Era of Performance for Every Data Center



GB200 NVL72

Compute for Trillion Parameter Scale AI
Maximum Performance and Lowest TCO

HGX B200

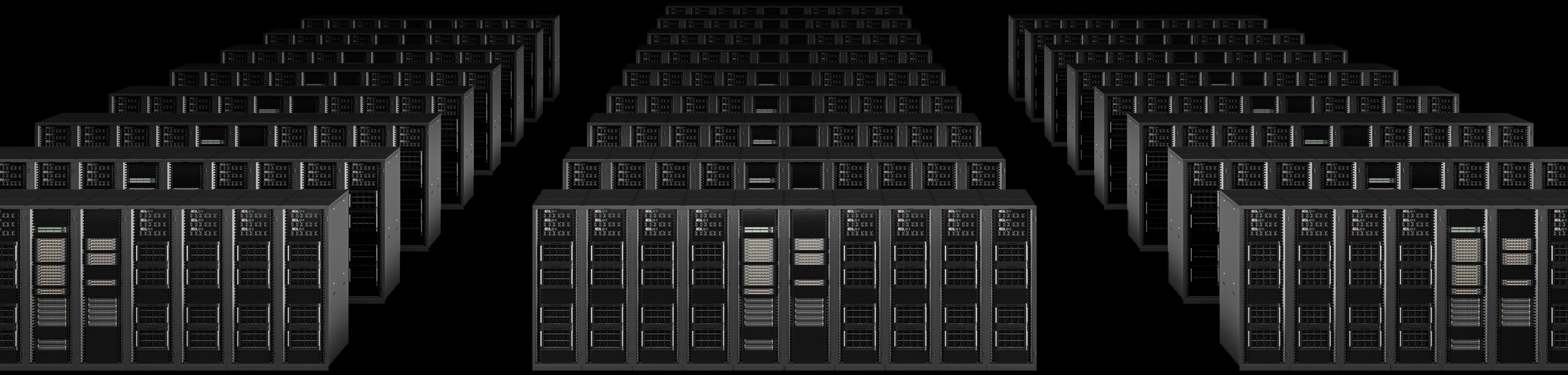
Best Performance and TCO for HGX Platform

HGX B100

Drop-in Upgrade for Existing Hopper Infrastructure

Train GPT-MoE-1.8T in 90 Days

Hopper
8000 GPUs | 15MW



Train GPT-MoE-1.8T in 90 Days

Blackwell GB200 NVL72
2000 GPUs | 4MW

1/4th the Power

LLM Training Workload: GPT-MoE-1.8T | H100 vs GB200 NVL72