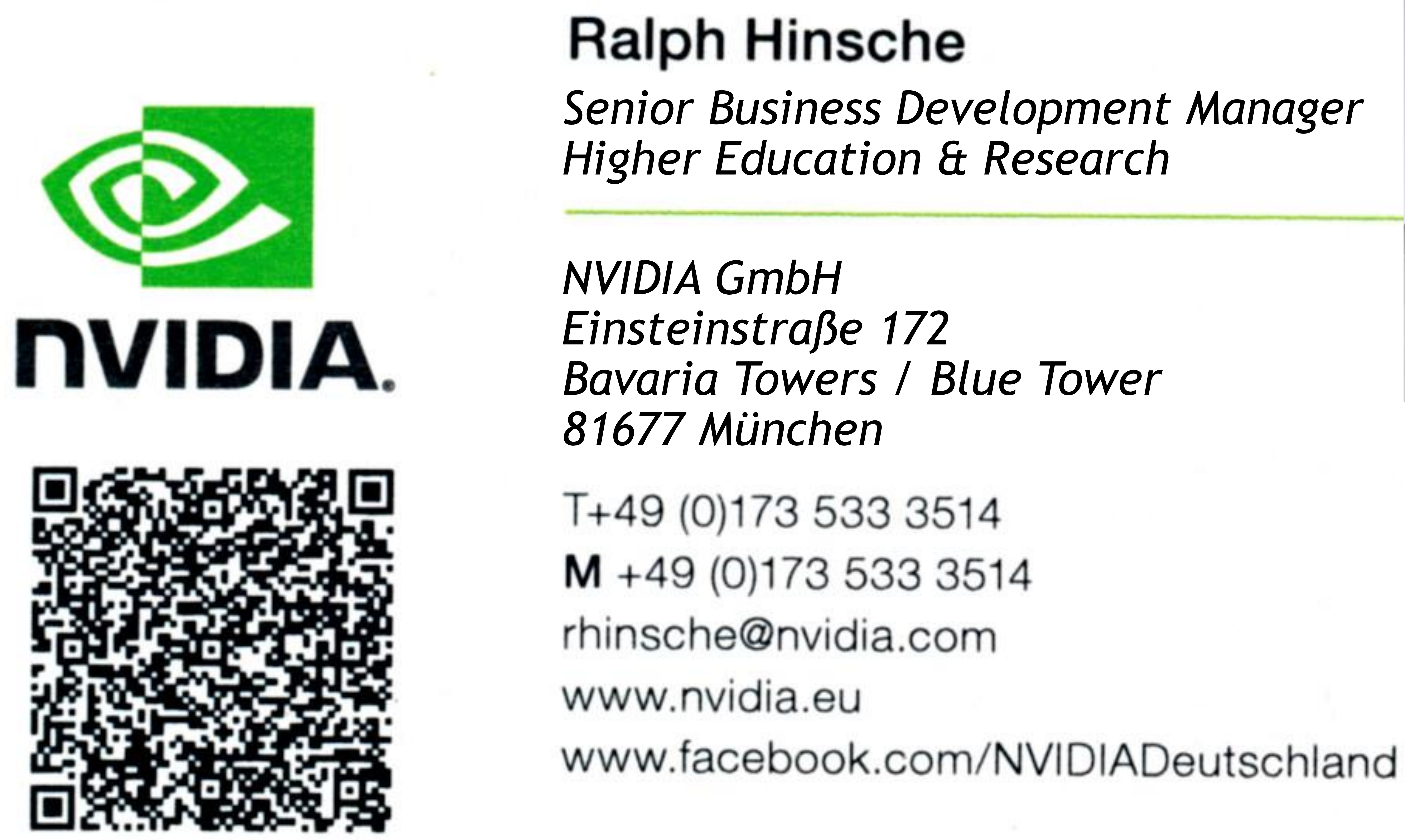
Blackwell Customer Deck

March 2024

CONTACT DETAILS



Ralph Hinsche

Senior Business Development Manager
Higher Education & Research

NVIDIA GmbH
Einsteinstraße 172
Bavaria Towers / Blue Tower
81677 München

T+49 (0)173 533 3514
M +49 (0)173 533 3514
rhinsche@nvidia.com
www.nvidia.eu
www.facebook.com/NVIDIADeutschland

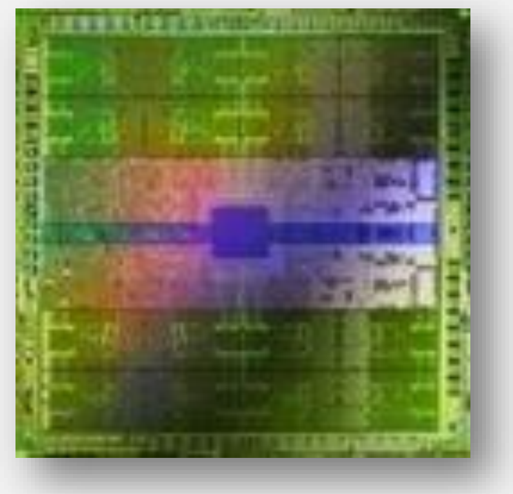
NOW 18 YEARS OF GPU COMPUTING / 12 YEARS OF AI-ACCELERATION

CUDA Launched

NVIDIA
CUDA

World's First GPU
Top500 System

Fermi: World's
First HPC GPU



Oak Ridge Deploys World's Fastest
Supercomputer w/ GPUs

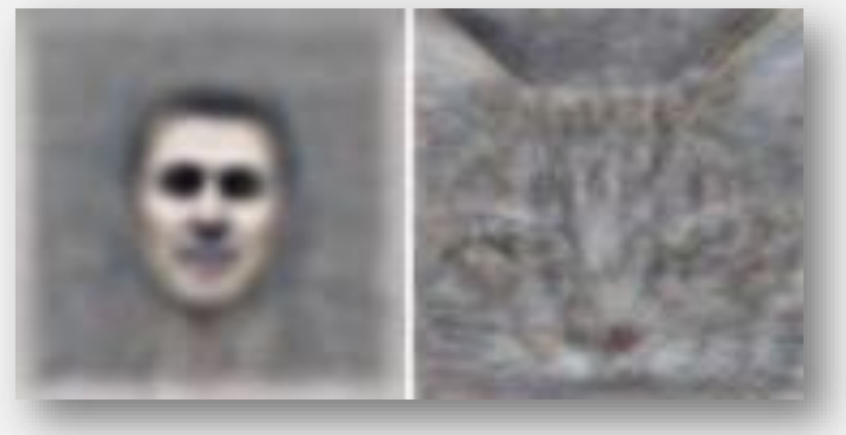
Discovered How H1N1 Mutates
to Resist Drugs

AlexNet beats expert code by
huge margin using GPUs

IMAGENET

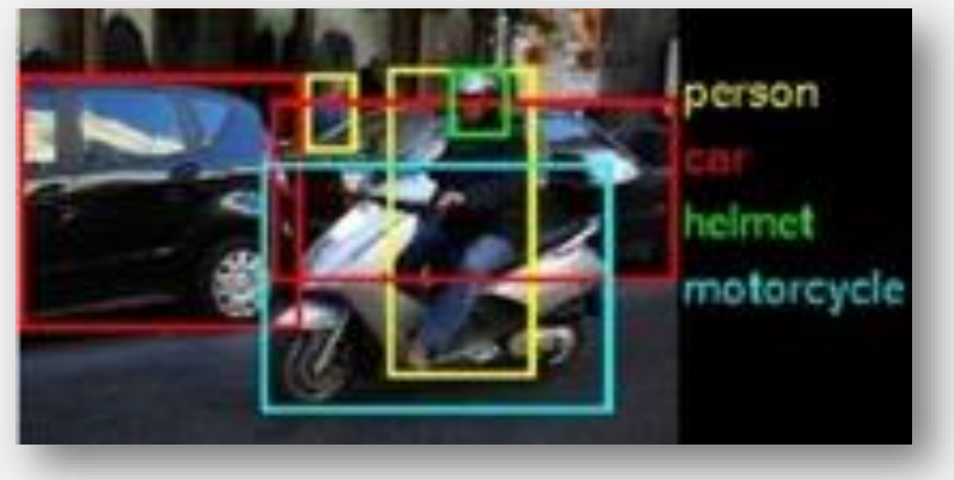
World's First Atomic Model
of HIV Capsid

Stanford Builds AI Machine
using GPUs



World's First 3-D Mapping of
Human Genome

Google Outperforms
Humans in ImageNet



GPU-Trained AI Machine Beats
World Champion in Go

nature
At last — a computer program that can beat a champion Go player
ALL SYSTEMS GO

Top 13 Greenest
Supercomputers Powered
by NVIDIA GPUs

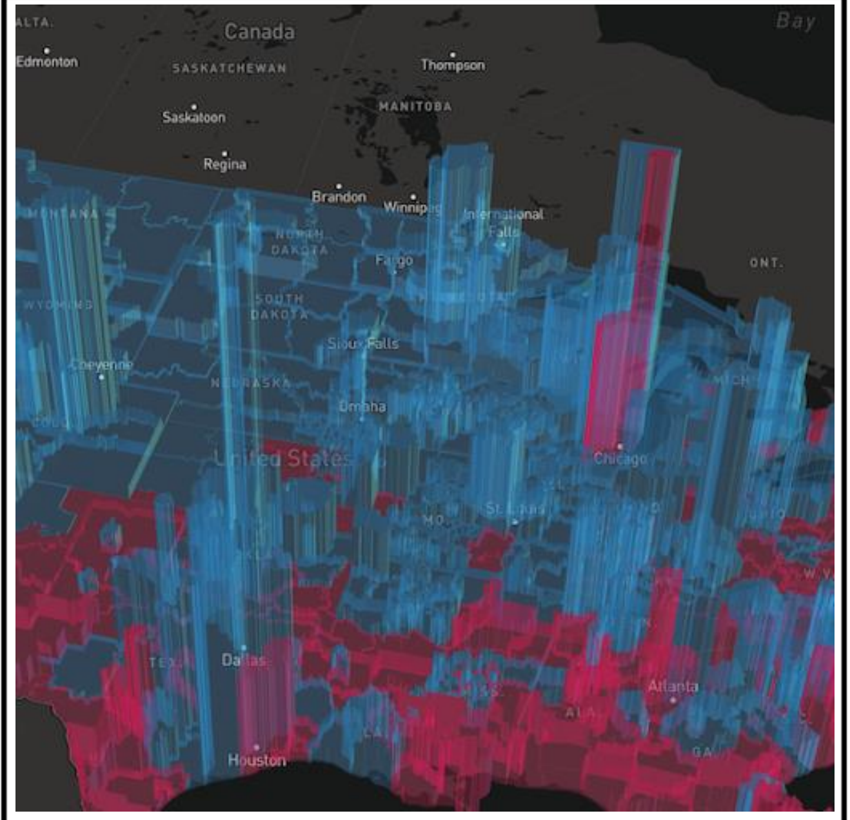
2006 2008 2010 2012 2014 2017

OUR BODY OF WORK

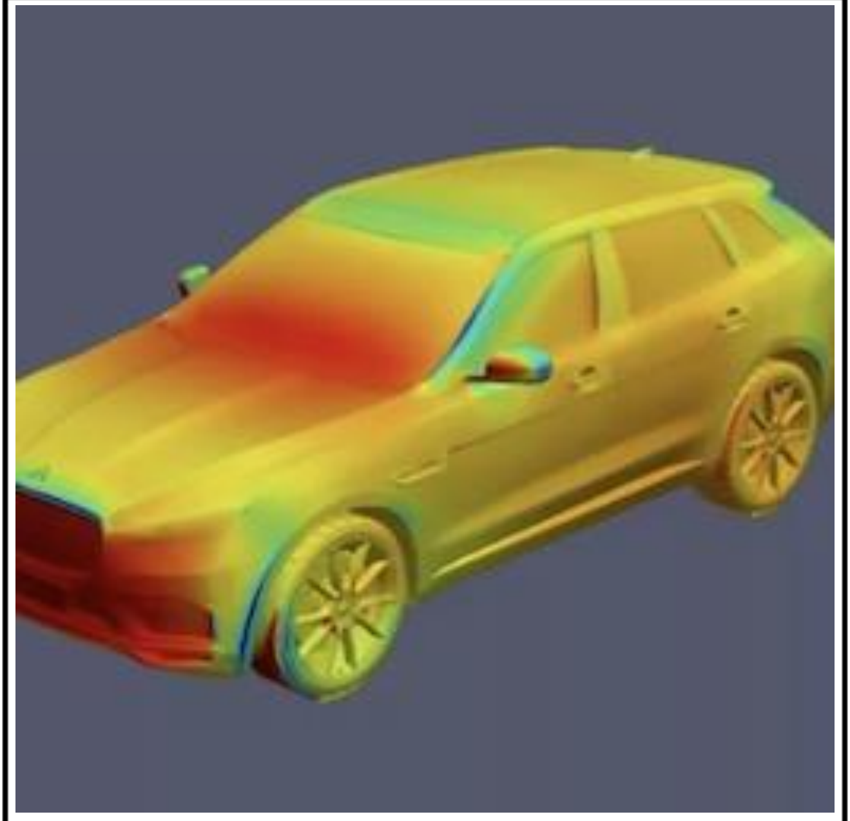
NVIDIA pioneered accelerated computing to tackle challenges no one else can solve. We engineer technology for the da Vincis and Einsteins of our time. Our work in AI is transforming 100 trillion dollars worth of industries, from gaming to healthcare to transportation, and profoundly impacting society.

NVIDIA AI Accelerated Computing Platform

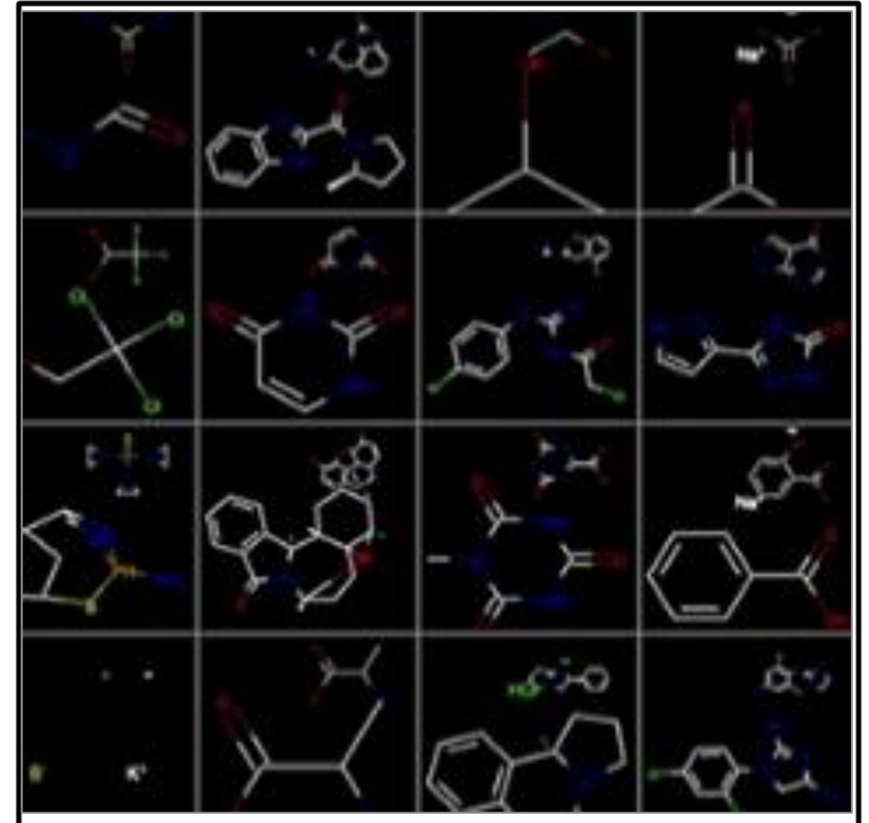
Hardware and Software Acceleration Across Every Workload and Vertical



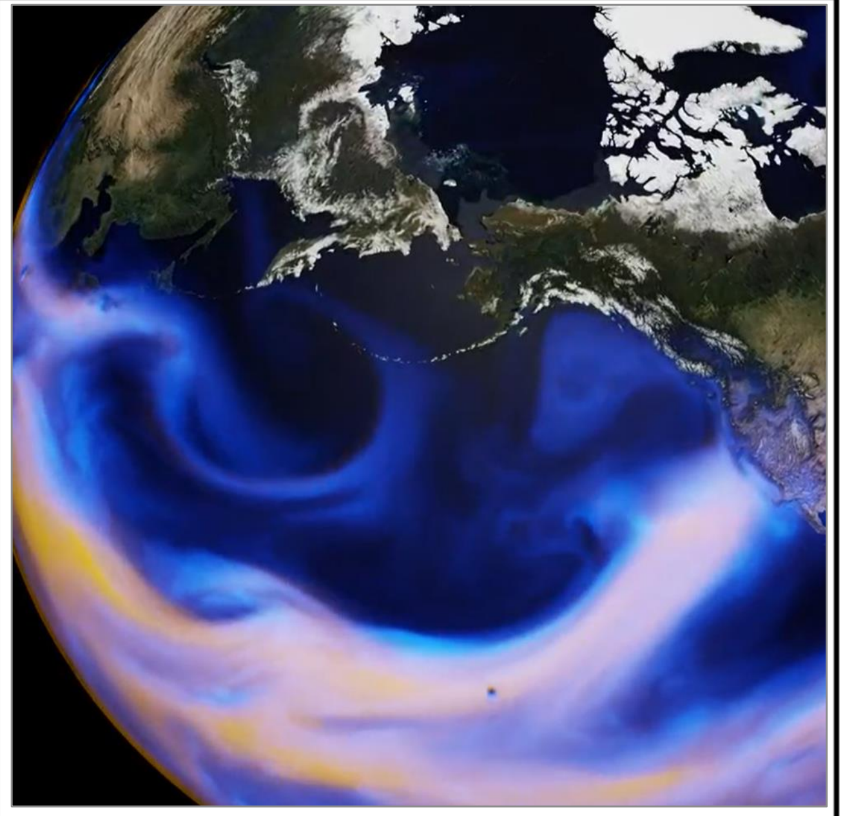
DATA PROCESSING



CAD, CAE, SDA

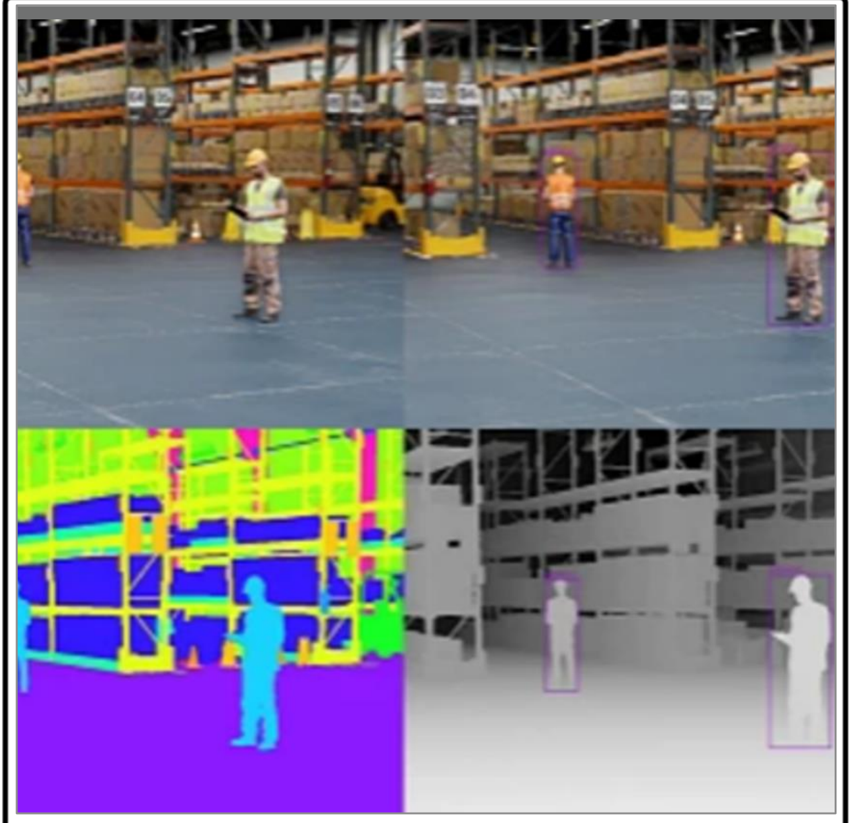


COMPUTER-AIDED DRUG DESIGN

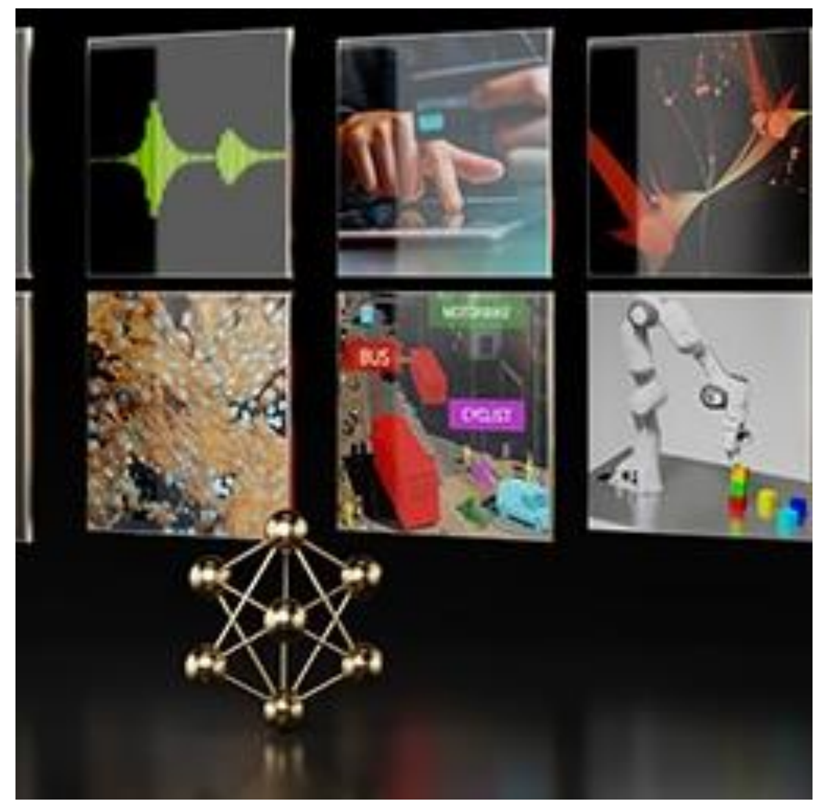


CLIMATE SIMULATION

QUANTUM

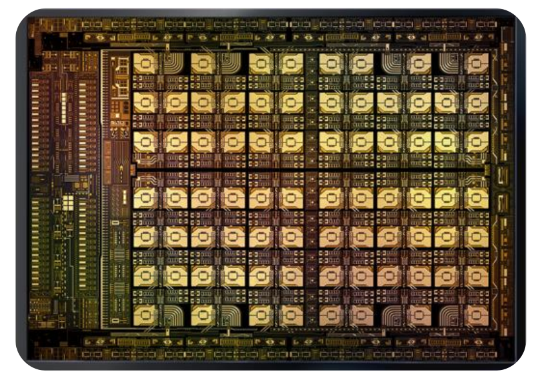


ROBOTICS
INDUSTRIAL
DIGITAL
TWINS

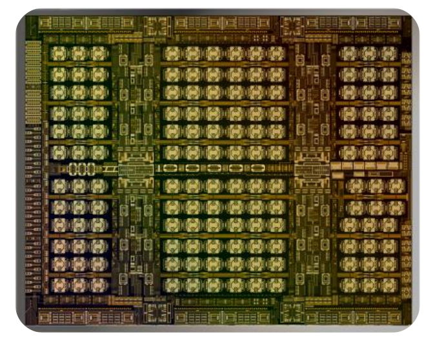


ENTERPRISE AI

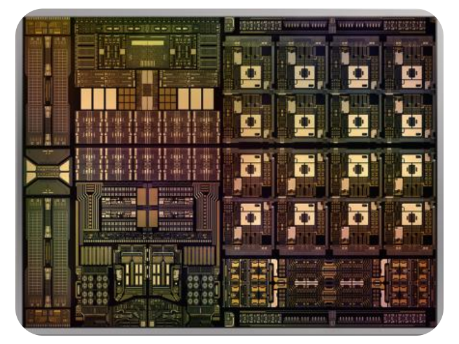
CUDA-X LIBRARIES



CPU

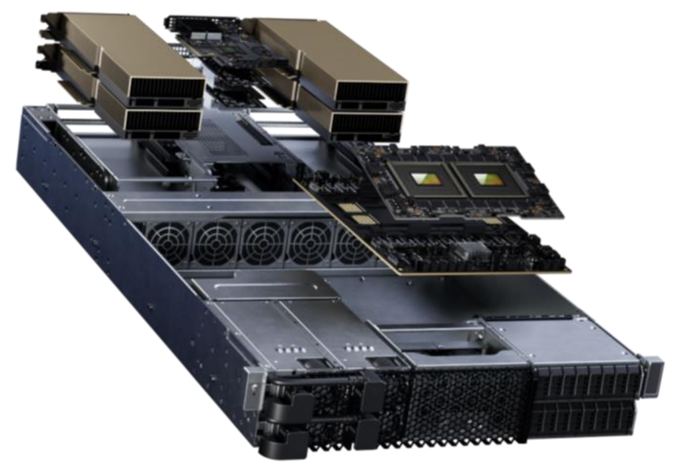


GPU



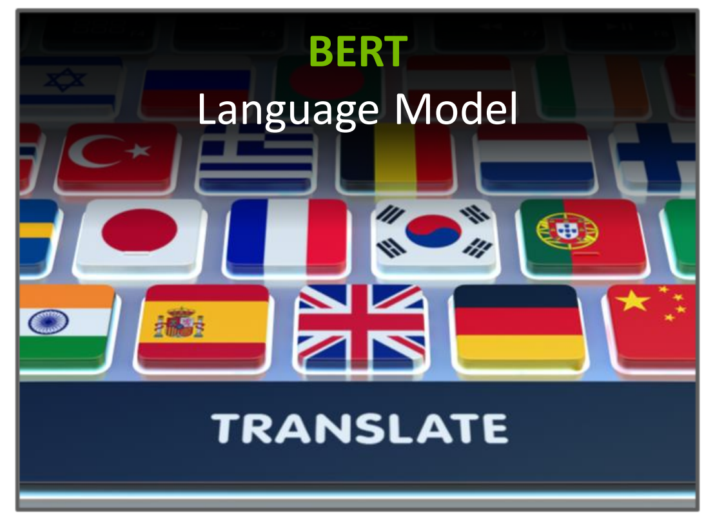
DPU

ACCELERATED COMPUTING



The Next Era of Generative AI

Image Classification



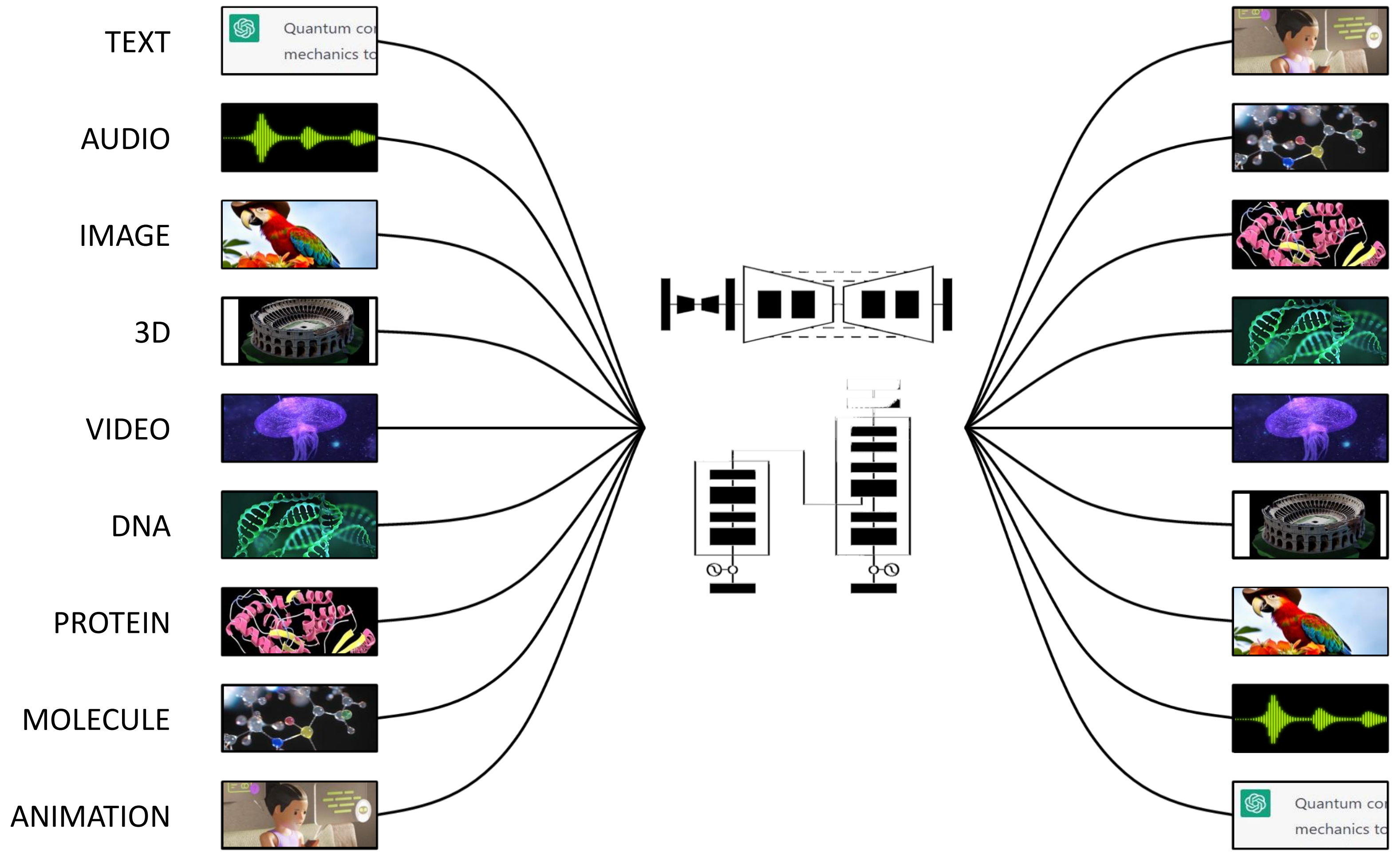
Transformer

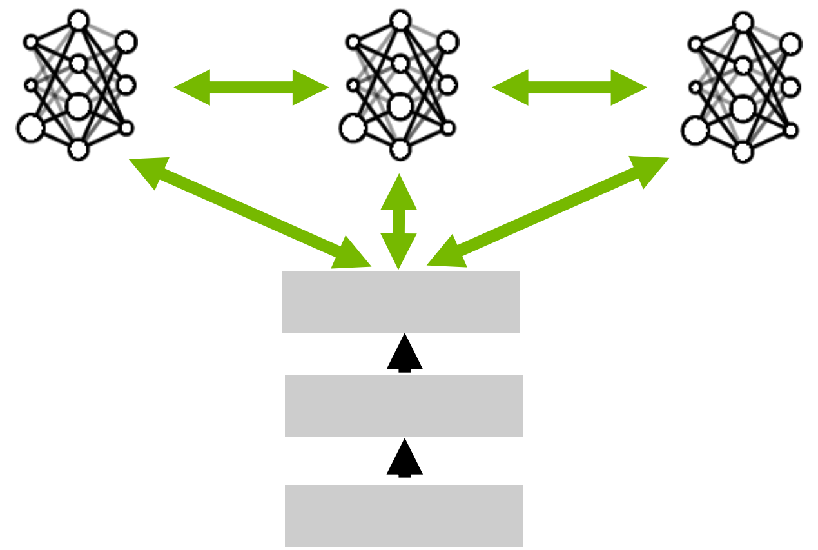
Large Language Models
(Transformer)

Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way.

Large Language Models
(Transformer)





Google
Gemini

Meta
NLLB

MISTRAL AI_
Mixtral

Realtime
<50ms latency

Parameters
>10T

Sequence Length
>32K input

Labeled Datasets

Unlabeled Datasets

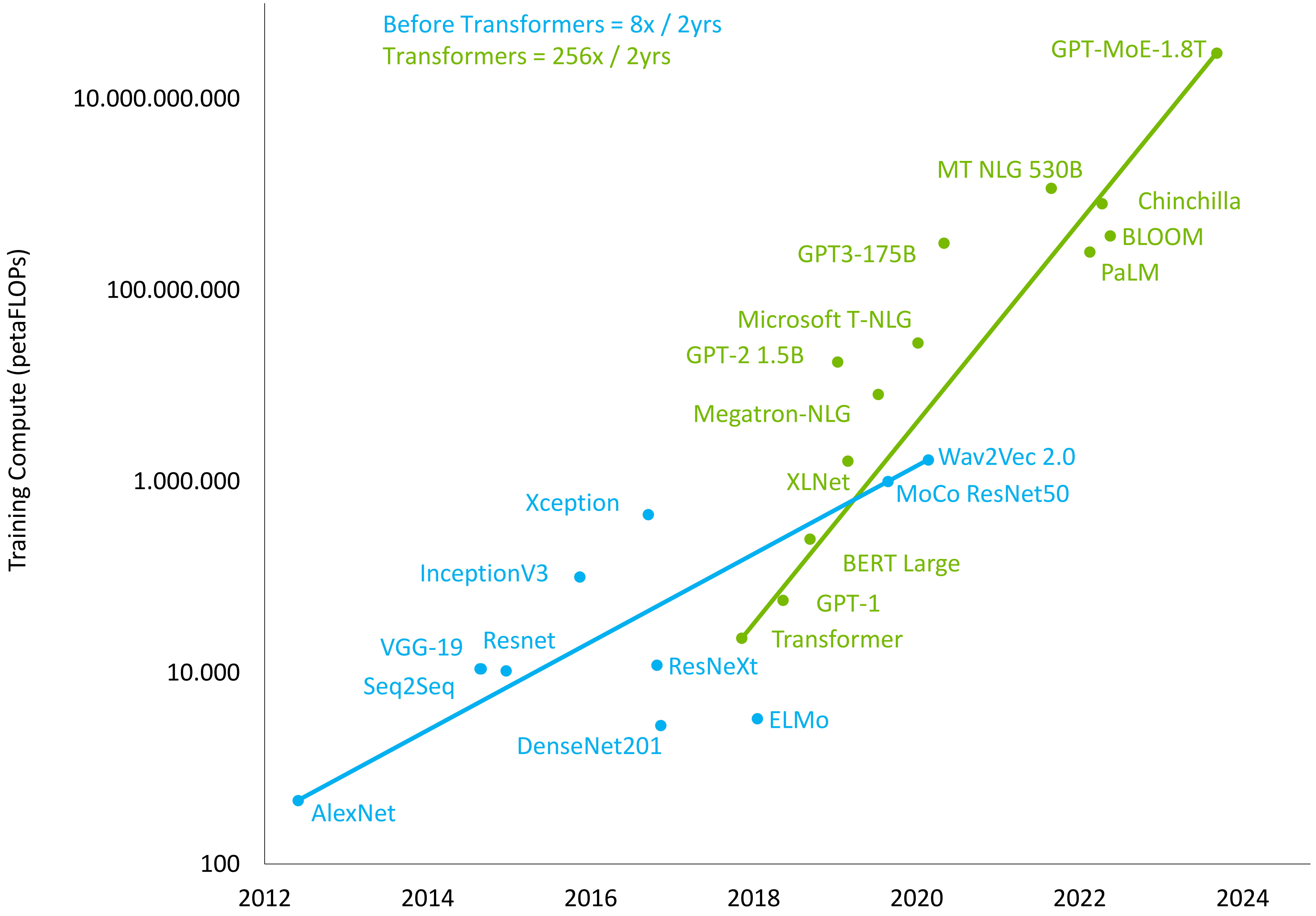
Generative AI

Multimodal Generative AI

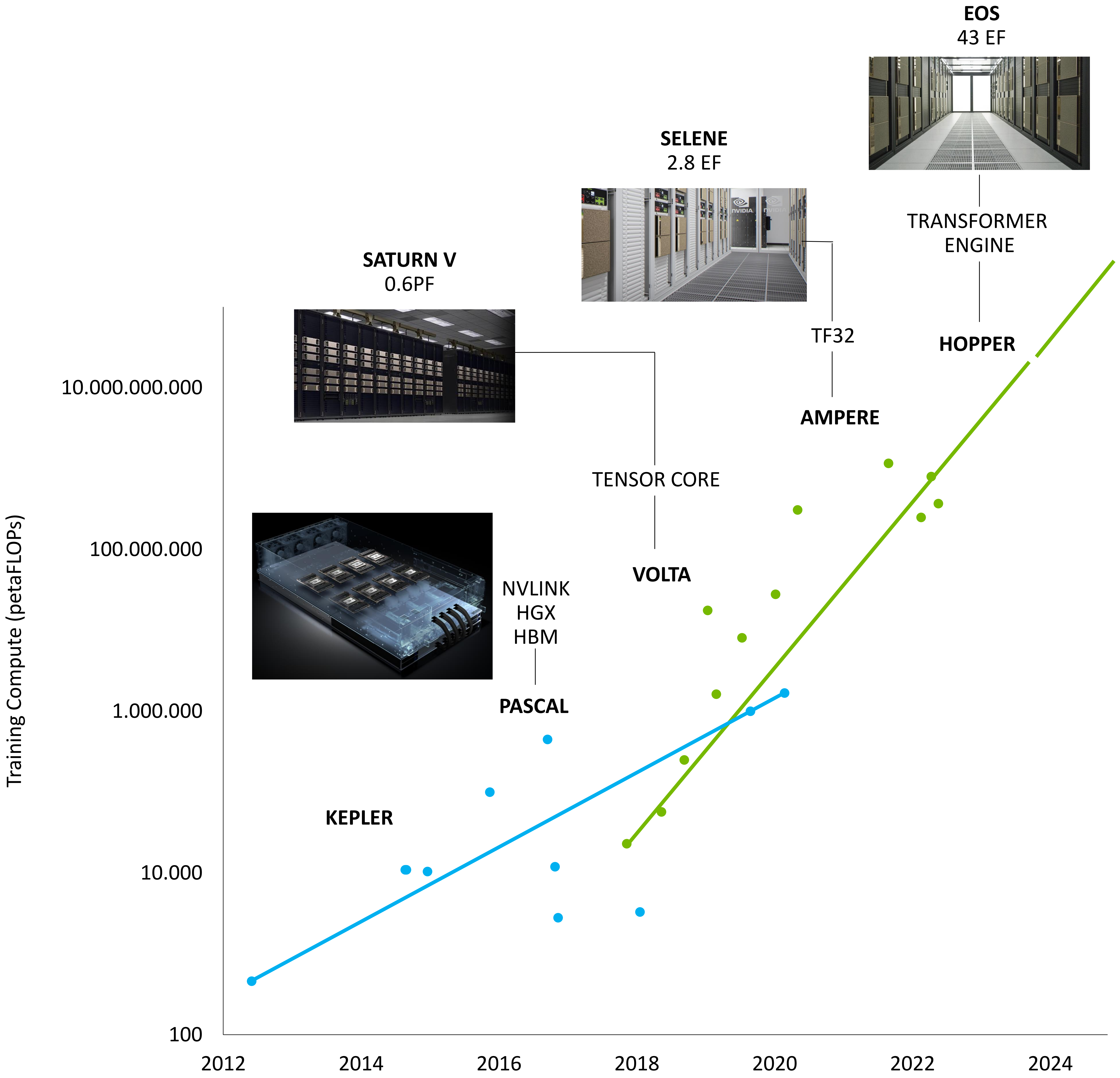
Mixture of Experts (MoE)

Production GenAI Inference

Explosive Growth in AI Computational Requirements

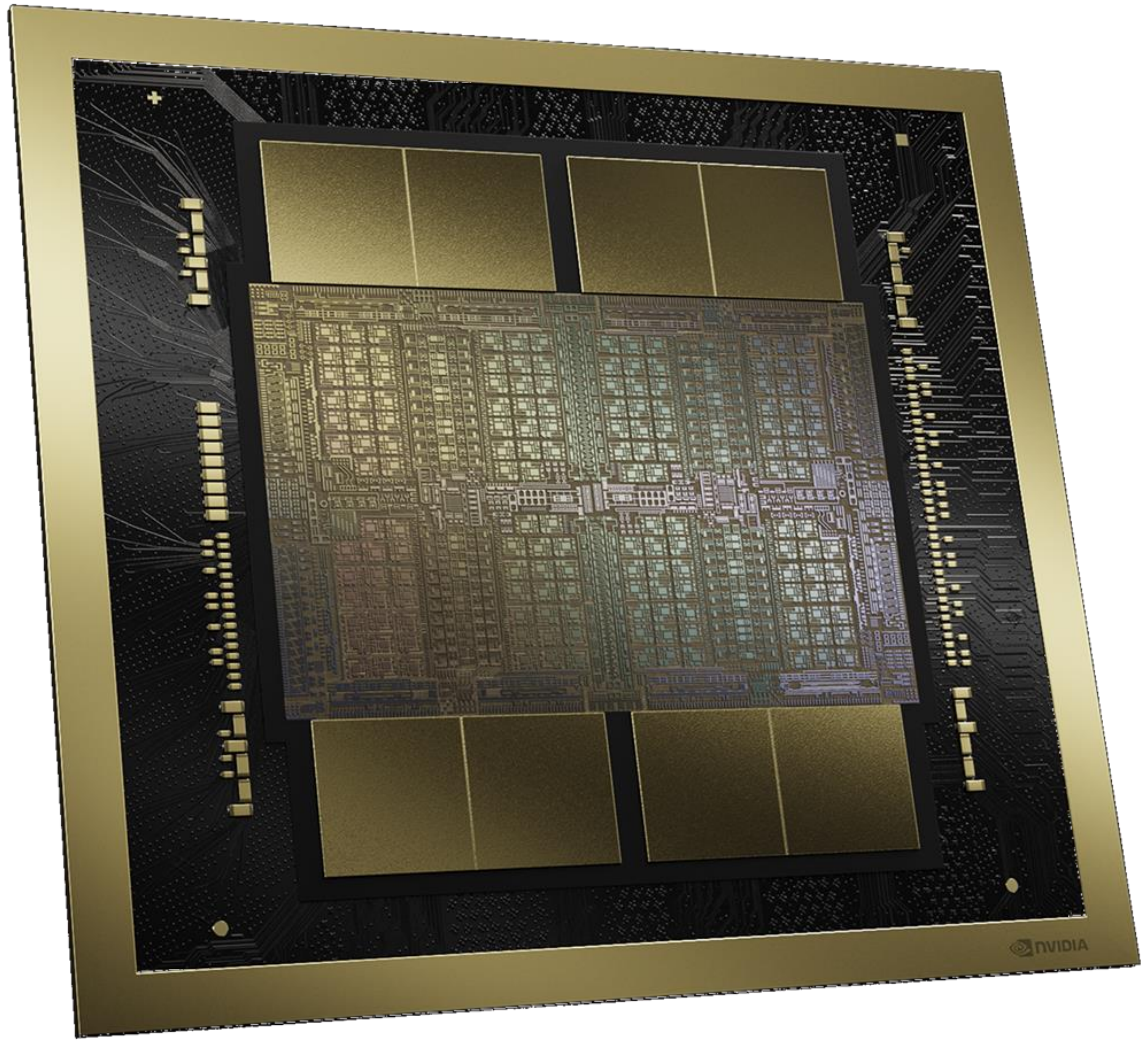


NVIDIA Enables Explosive Growth in AI Computational Requirements



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 beyond100K GPUs

AI SUPERCHIP
208B Transistors

2nd GEN TRANSFORMER ENGINE
FP4/FP6 Tensor Core

5th GENERATION NVLINK
Scales to 576 GPUs

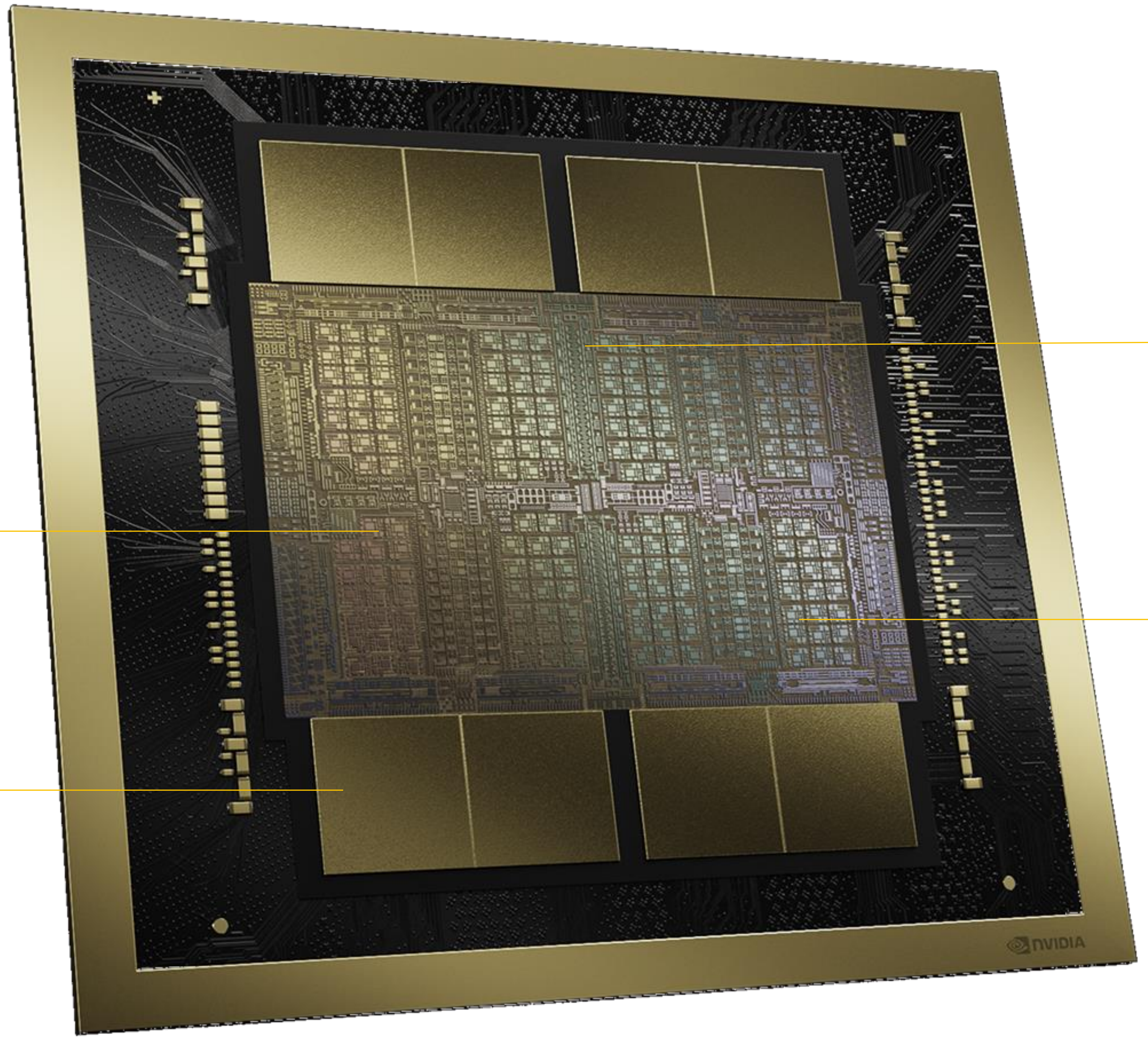
RAS ENGINE
100% In-System
Self-Test

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

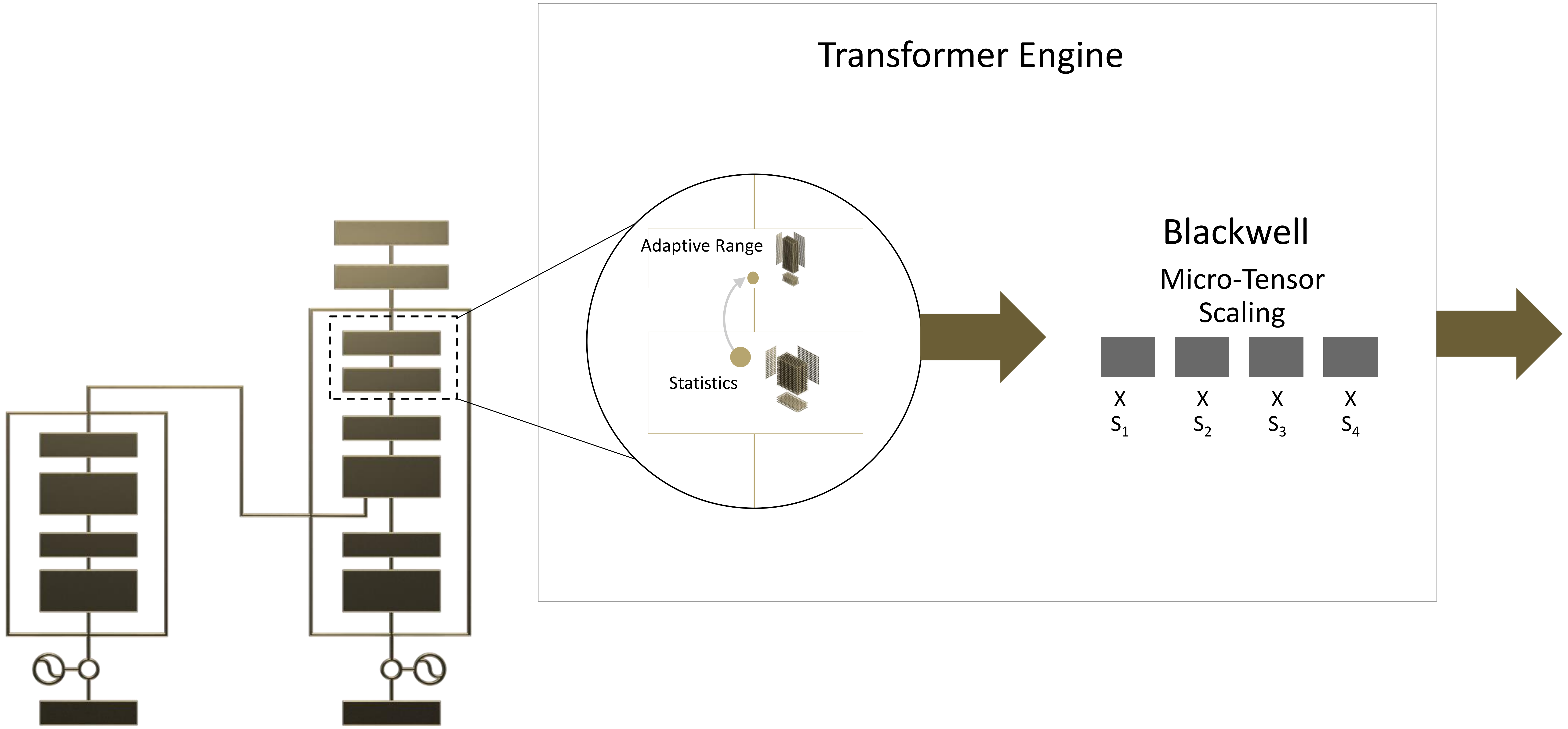
Full performance. No compromises

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



Transformer Engine

Adaptive Range

Statistics

Blackwell

Micro-Tensor Scaling

X S_1 X S_2 X S_3 X S_4

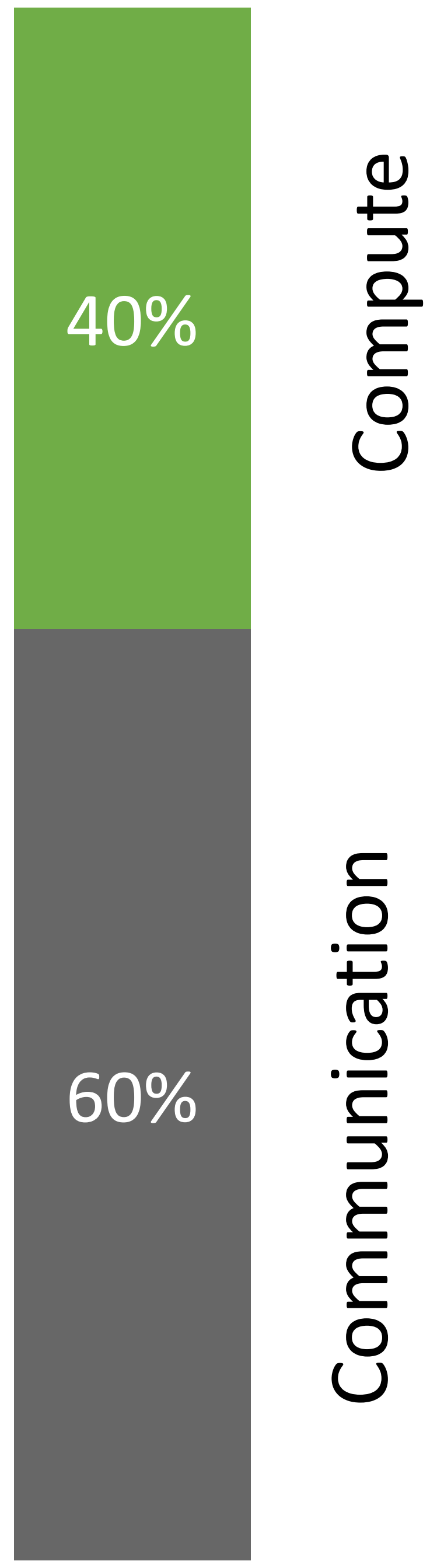
Enabling FP4 AI Inference

2x Compute

2x Bandwidth

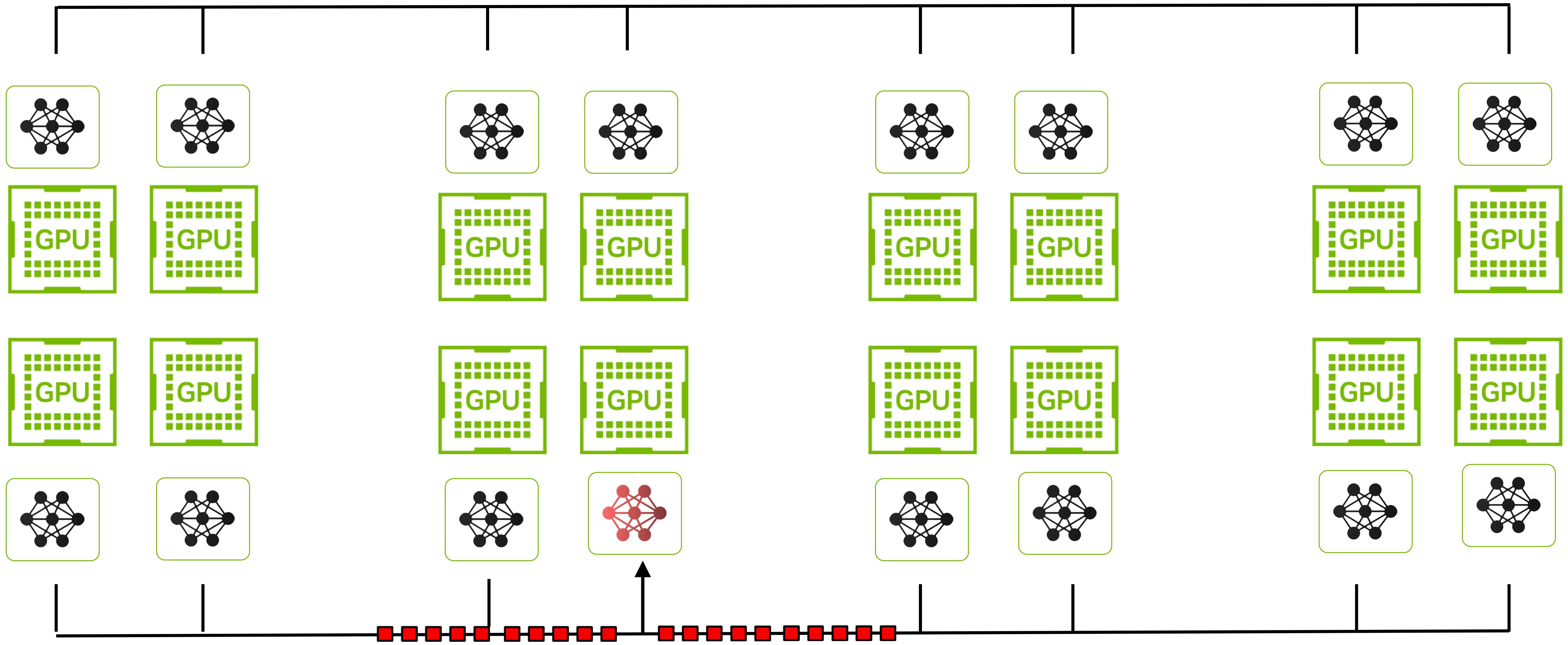
2x Model Size

Next Generation Models Communication Bottleneck



Mixture of Expert Models

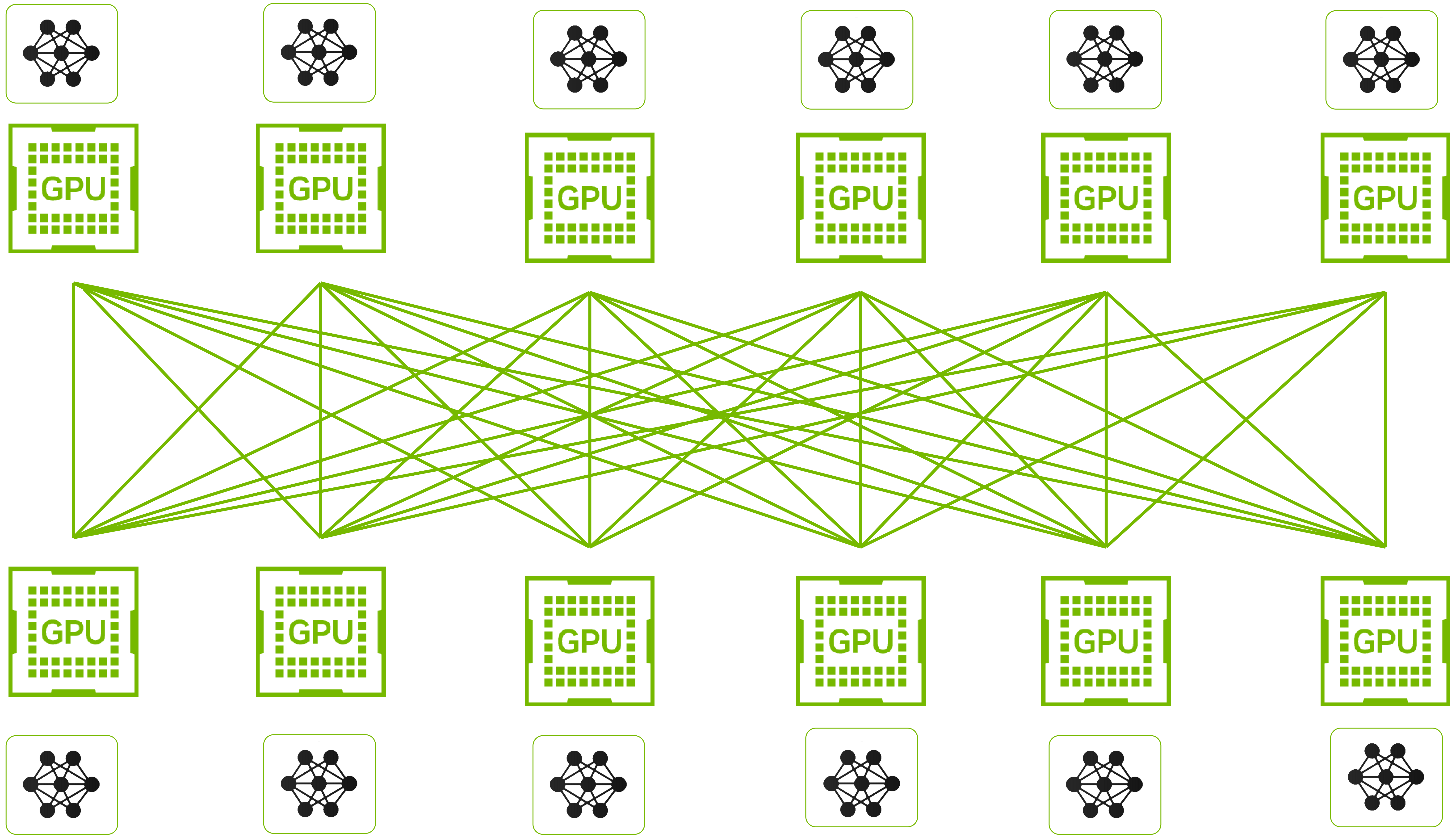
GPT MoE1.8T Parameters



HDR InfiniBand
100 GByte/s

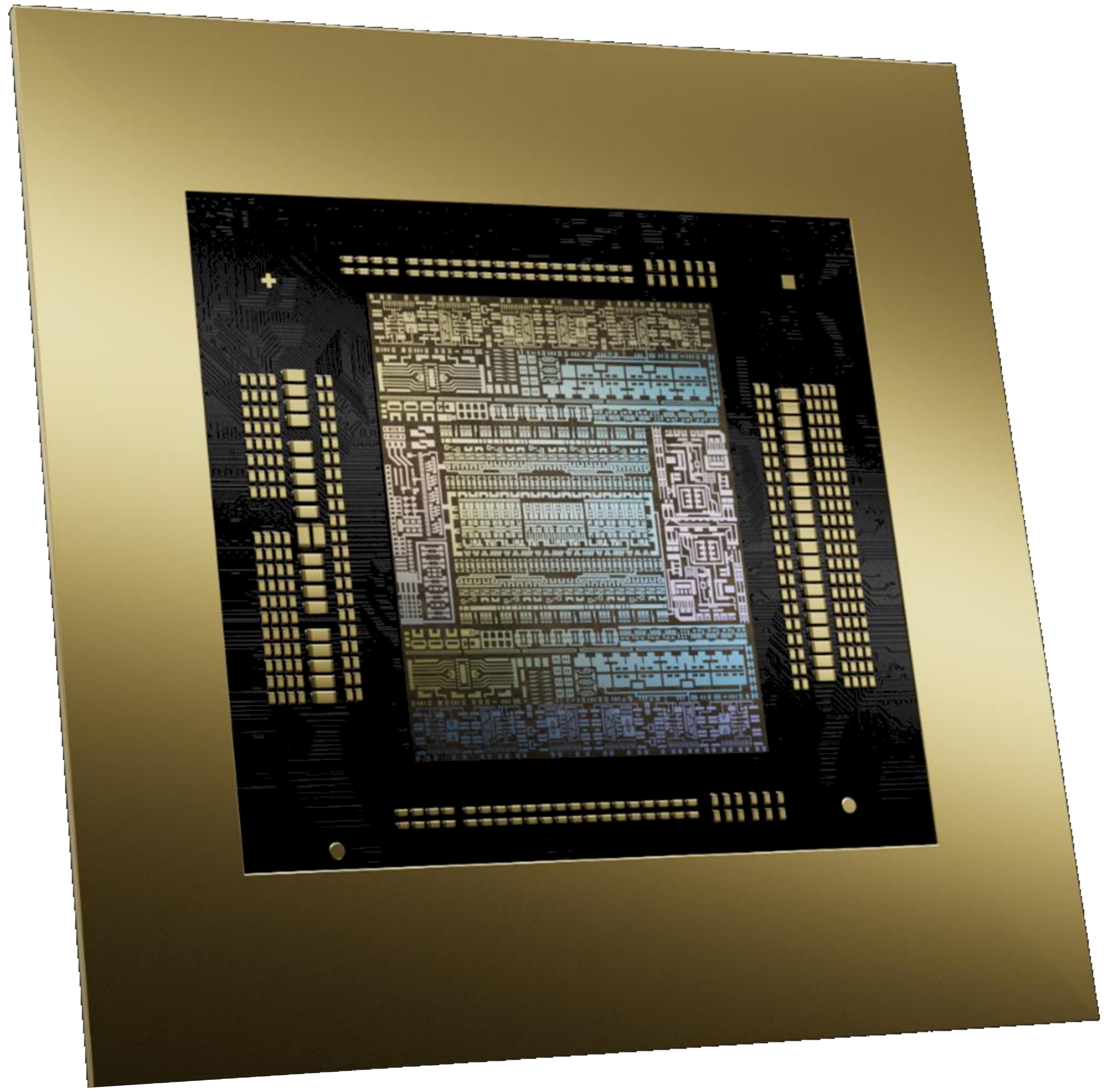
15 GPUs Sending
to 1 GPU

The World Needs a New AI Compute Fabric



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

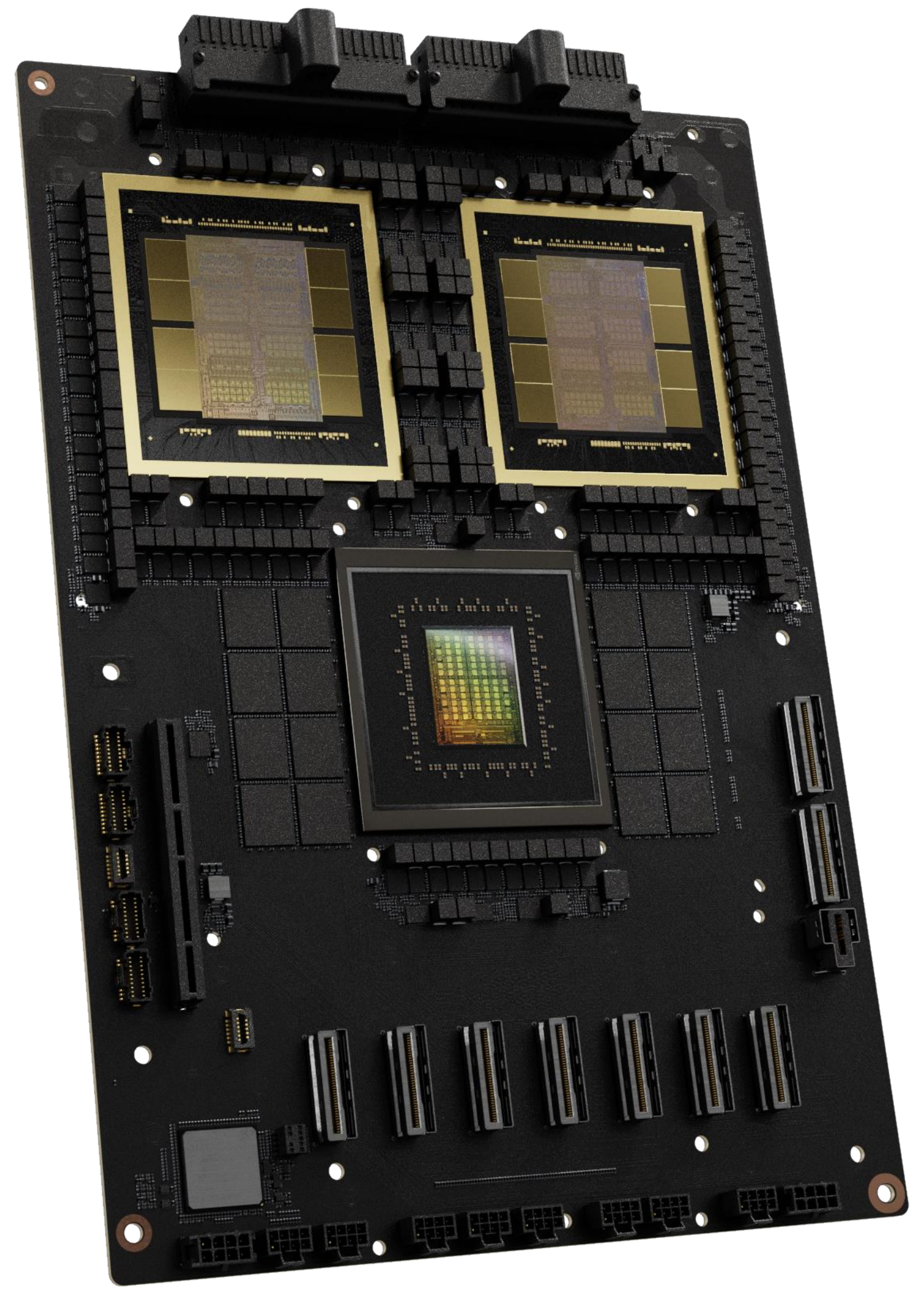
GB200 NVL72

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

Training	720 PFLOPs
Inference	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

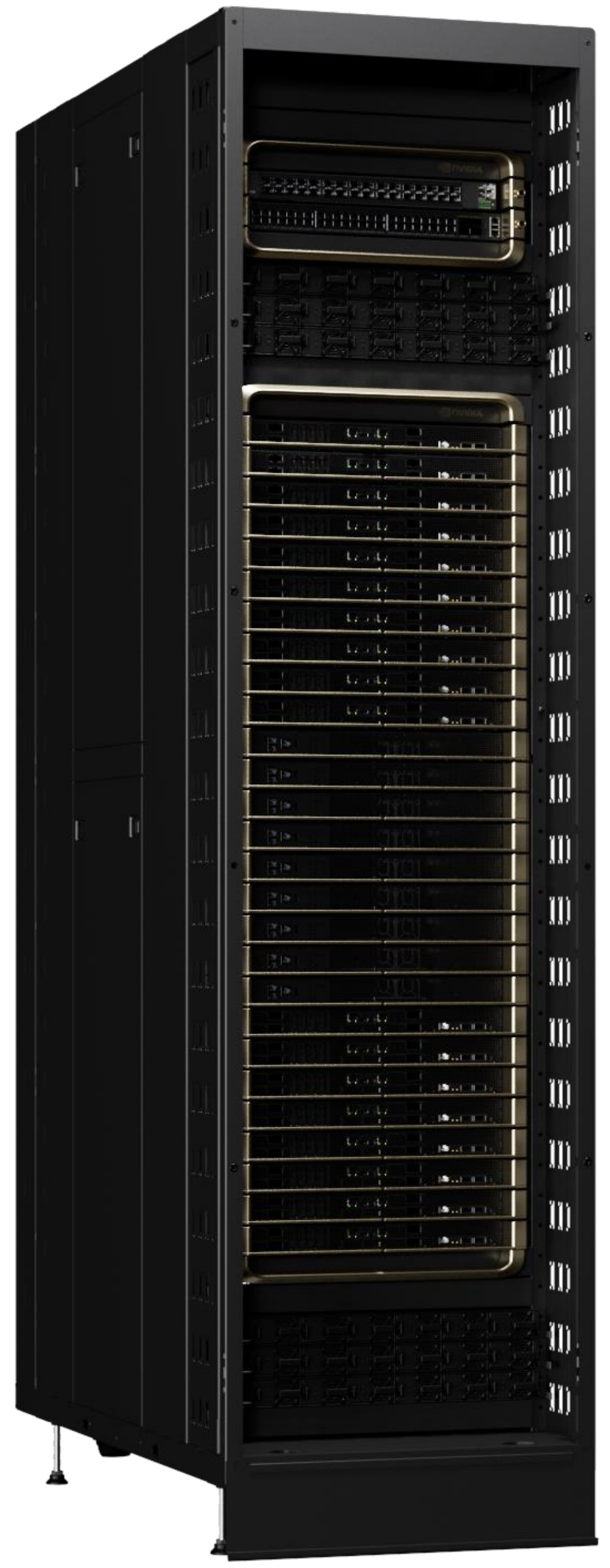
2x GB200
80 PETAFLOPS FP4 AI INFERENCE
40 PETAFLOPS FP8 AI TRAINING
1728 GB FAST MEMORY
1U Liquid Cooled
18 Per Rack

NVLINK SWITCH TRAY

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

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

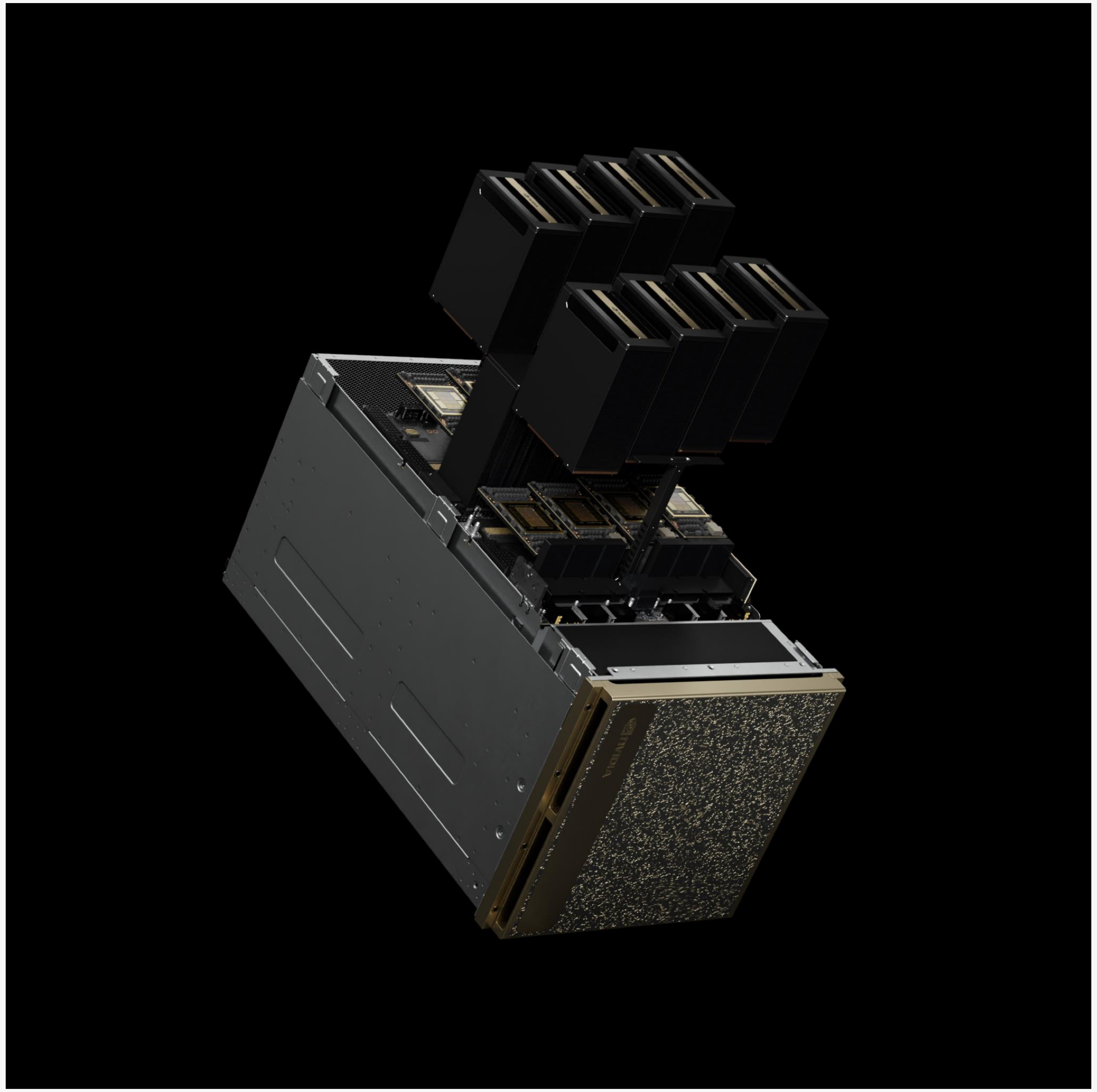
Maximum Performance and Lowest TCO

HGX B200

Best Performance and TCO for HGX Platform

HGX B100

Drop-in Upgrade for Existing Hopper Infrastructure



DGX B200
The Foundation for Your AI Center of Excellence

DGX B200 System

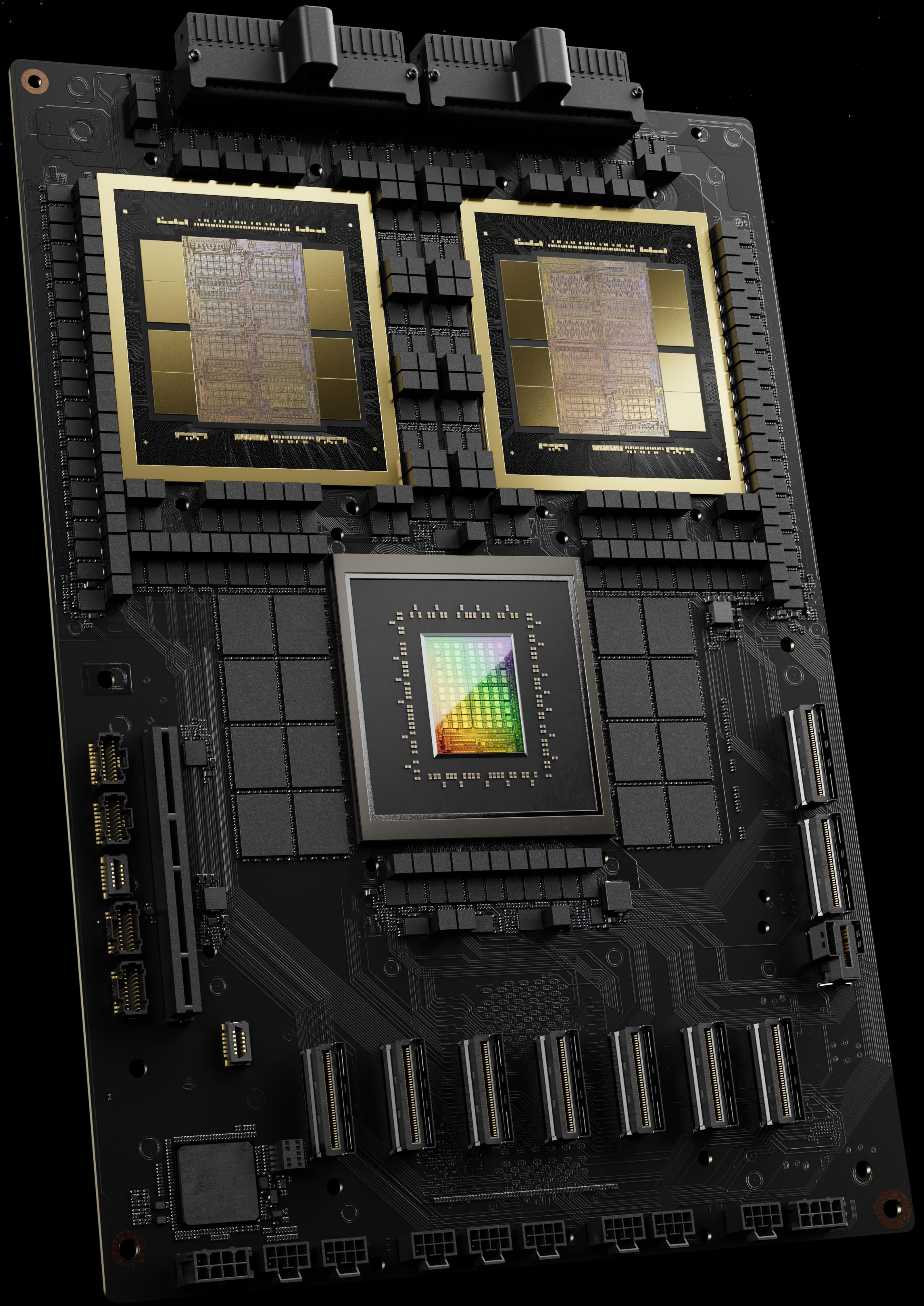
- 6th generation of air-cooled DGX system
- Unified platform for every workload from training, to fine-tuning, to inference
- **8x** NVIDIA Blackwell GPUs
- **1.4TB** of GPU memory, enabling training of large generative AI models
- 15X inference, 3X training, and 12X energy savings
- NVIDIA Blackwell architecture in rack mount design
- Scalable with **DGX SuperPOD**

Blackwell System Specifications

	GB200 NVL72	HGX B200	HGX B100
Blackwell GPUs	72	8	8
FP4 Tensor Core	1,440 petaFLOPS	144 petaFLOPS	112 petaFLOPS
FP8/FP6/INT8	720 petaFLOPS	72 petaFLOPS	56 petaFLOPS
Fast Memory	Up to 30 TB	up to 1.5 TB	Up to 1.5TB
Aggregate Memory Bandwidth	Up to 600 TB/s	Up to 64 TB/s	Up to 64 TB/s
Aggregate NVLink Bandwidth	130 TB/s	14.4 TB/s	14.4 TB/s
CPU Cores	2592 Arm Neoverse V2 cores	-	-
Per GPU Specifications			
FP4 Tensor Core	20 petaFLOPS	18 petaFLOPS	14 petaFLOPS
FP8/FP6 Tensor Core	10 petaFLOPS	9 petaFLOPS	7 petaFLOPS
INT8 Tensor Core	10 petaOPS	9 petaOPS	7 petaOPs
FP16/BF16 Tensor Core	5 petaFLOPS	4.5 petaFLOPS	3.5 petaFLOPS
TF32 Tensor Core	2.5 petaFLOPS	2.2 petaFLOPS	1.8 petaFLOPS
FP64 Tensor Core	45 teraFLOPS	40 teraFLOPS	30 teraFLOPS
GPU memory | Bandwidth	Up to 192 GB HBM3e | Up to 8 TB/s		
Multi-Instance GPU (MIG)	7		
Decompression Engine	Yes		
Decoders	2x 7 NVDEC 2x 7 NVJPEG		
Power	Configurable up to 1,200W	Configurable up to 1,000W	Configurable up to 700W
Interconnect	5th Generation NVLink: 1.8TB/s PCIe Gen6: 256GB/s		
Server options	NVIDIA GB200 NVL72 partner and NVIDIA-Certified Systems with 72 GPUs	NVIDIA HGX B200 partner and NVIDIA-Certified Systems with 8 GPUs	NVIDIA HGX B100 partner and NVIDIA-Certified Systems with 8 GPUs

1.Preliminary specifications subject to change. All Tensor Core numbers with sparsity.

2.GB200 Superchip configuration includes 2 high performance B200 GPUs and one Grace CPU



ANNOUNCING NVIDIA BLACKWELL PLATFORM FOR TRILLION-PARAMETER SCALE GENERATIVE AI

AI SUPERCHIP
208B Transistors

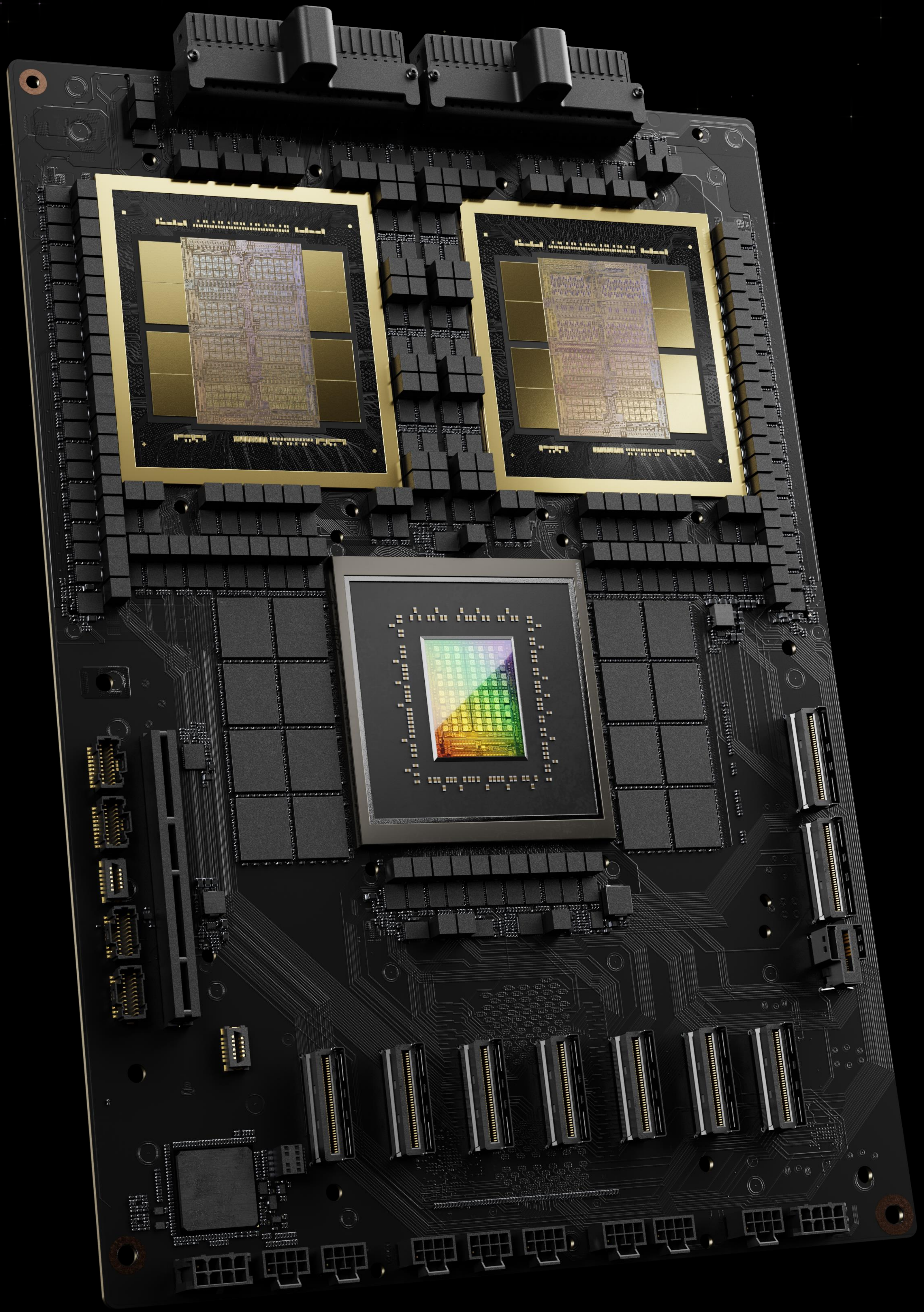
2nd GEN TRANSFORMER ENGINE
FP4/FP6 Tensor Core

5th GENERATION NVLINK
Scales to 576 GPUs

RAS ENGINE
100% In-System Self-Test

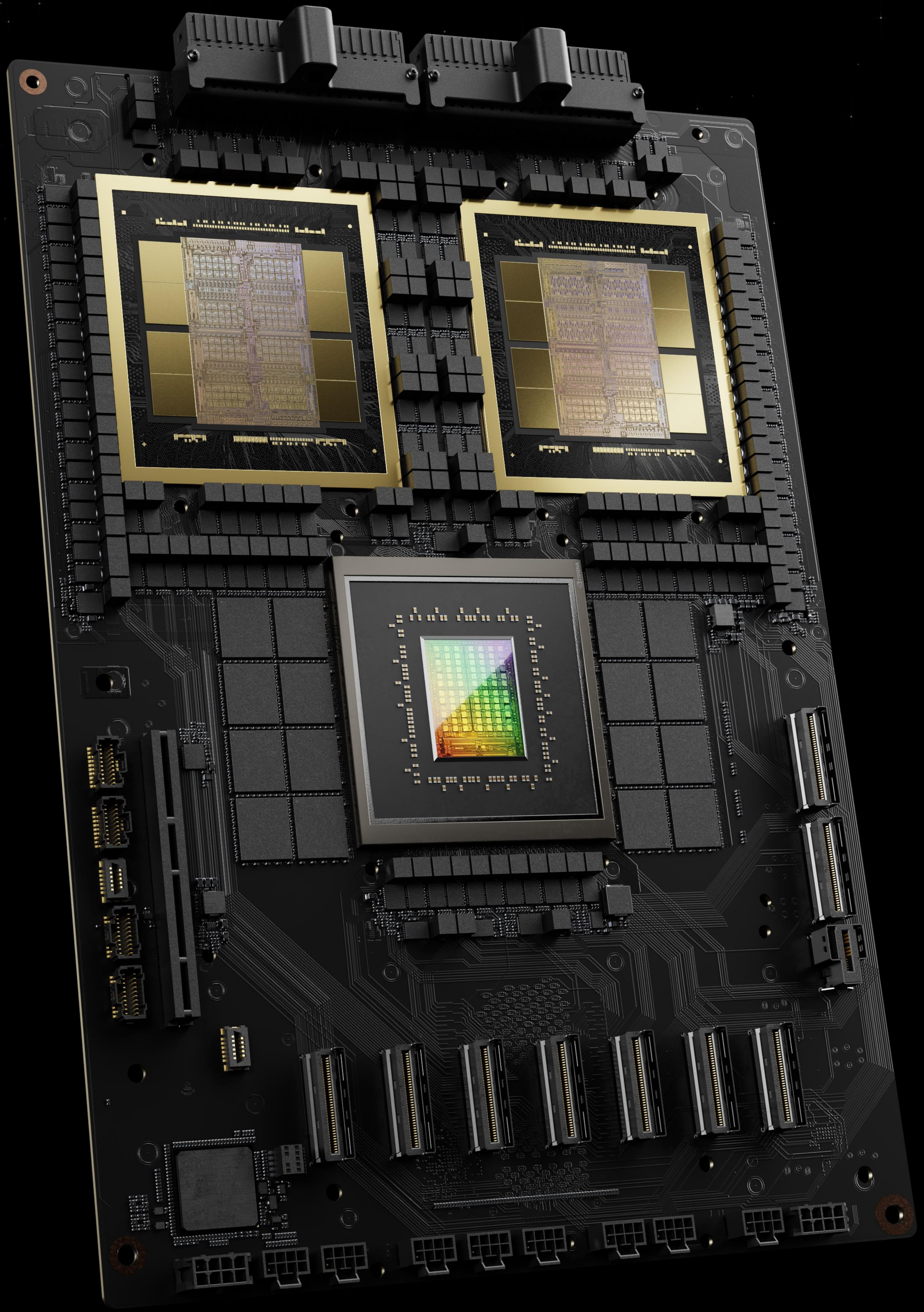
SECURE AI
Full Performance
Encryption & TEE

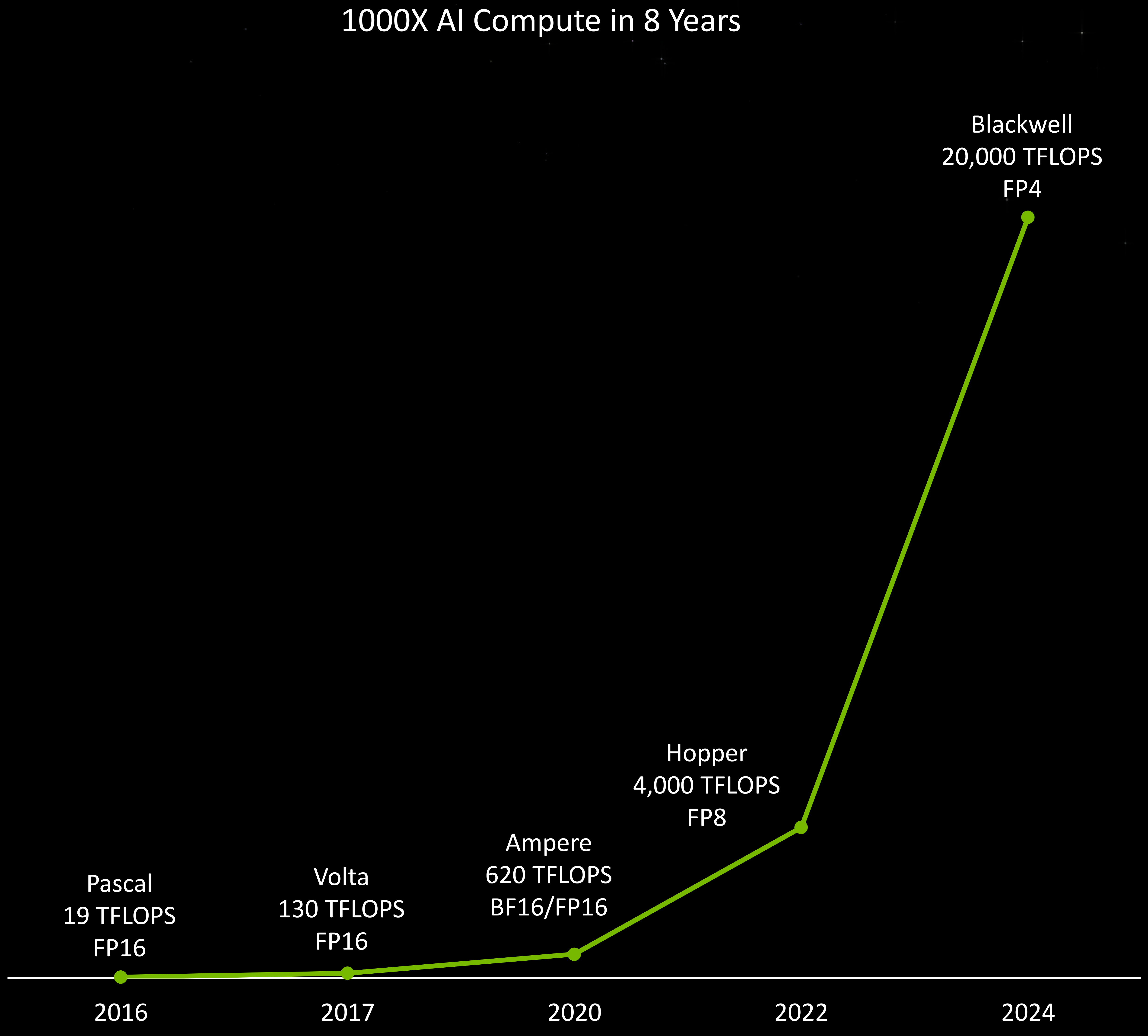
DECOMPRESSION ENGINE
800 GB/sec

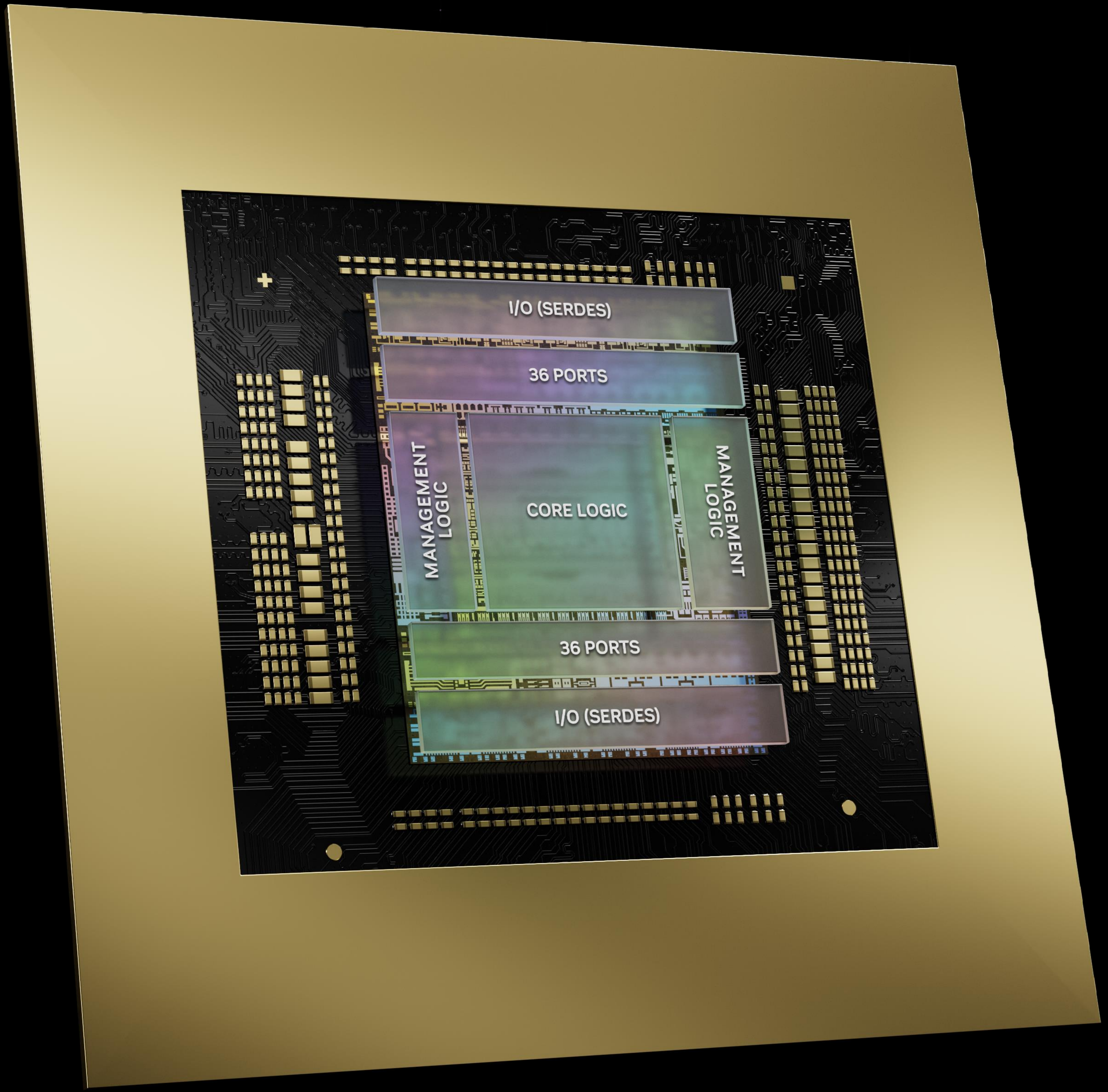


Blackwell GPU

FP8	20 PFLOPS	2.5X Hopper
NEW FP6	20 PFLOPS	2.5X
NEW FP4	**40 PFLOPS**	**5X**
HBM Model Size	740B param	6X
HBM Bandwidth	34T param/sec	5X
NVLINK All-Reduce with SHARP	7.2 TB/s	4X




1000X AI Compute in 8 Years
Blackwell
20,000 TFLOPS
FP4
Hopper
4,000 TFLOPS
FP8
Ampere
620 TFLOPS
BF16/FP16
Volta
130 TFLOPS
FP16
Pascal
19 TFLOPS
FP16
2016
2017
2020
2022
2024



NVLink Switch Chip

50B Transistors in TSMC 4NP

72-Ports Dual 200 Gb/sec SerDes

4 NVLinks at 1.8TB/sec

7.2TB/sec Full-Duplex Bandwidth

SHARP In-Network Compute – 3.6 TFLOPS FP8

NVIDIA Blackwell Platform

HGX B100

NVLINK Switch

GB200 Superchip
Compute Node

Quantum X800 Switch
ConnectX-8 SuperNIC

Spectrum X800 Switch
BlueField-3 SuperNIC

Train GPT-MoE-1.8T in 90 Days

Hopper
8000 GPUs | 15MW

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

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

Foto: G. Otto, GSI Helmholtzzentrum für Schwerionenforschung

Foto: G. Otto, GSI Helmholtzzentrum für Schwerionenforschung

GSI Helmholtzzentrum für Schwerionenforschung GmbH

Green IT Cube / GSI Darmstadt

(Almost looks like an NVidia-Product / at least it has the right color and is illuminated GREEN at night …)

Final configuration consists of :

- 79x HPE Apollo 6500 Gen10 Plus 8-way AI-servers (with 8x A100/80GB each -> HGX-8), with a total of 632x A100 this is the largest academic AI-Cluster in Germany
- € 6.27 Mill net revenue generated by HGX-8 A100 boards
- € 1.14 Mill net revenue on InfiniBand network products (IB-switches, IB-adapters, cables)

Green IT Cube specifics :

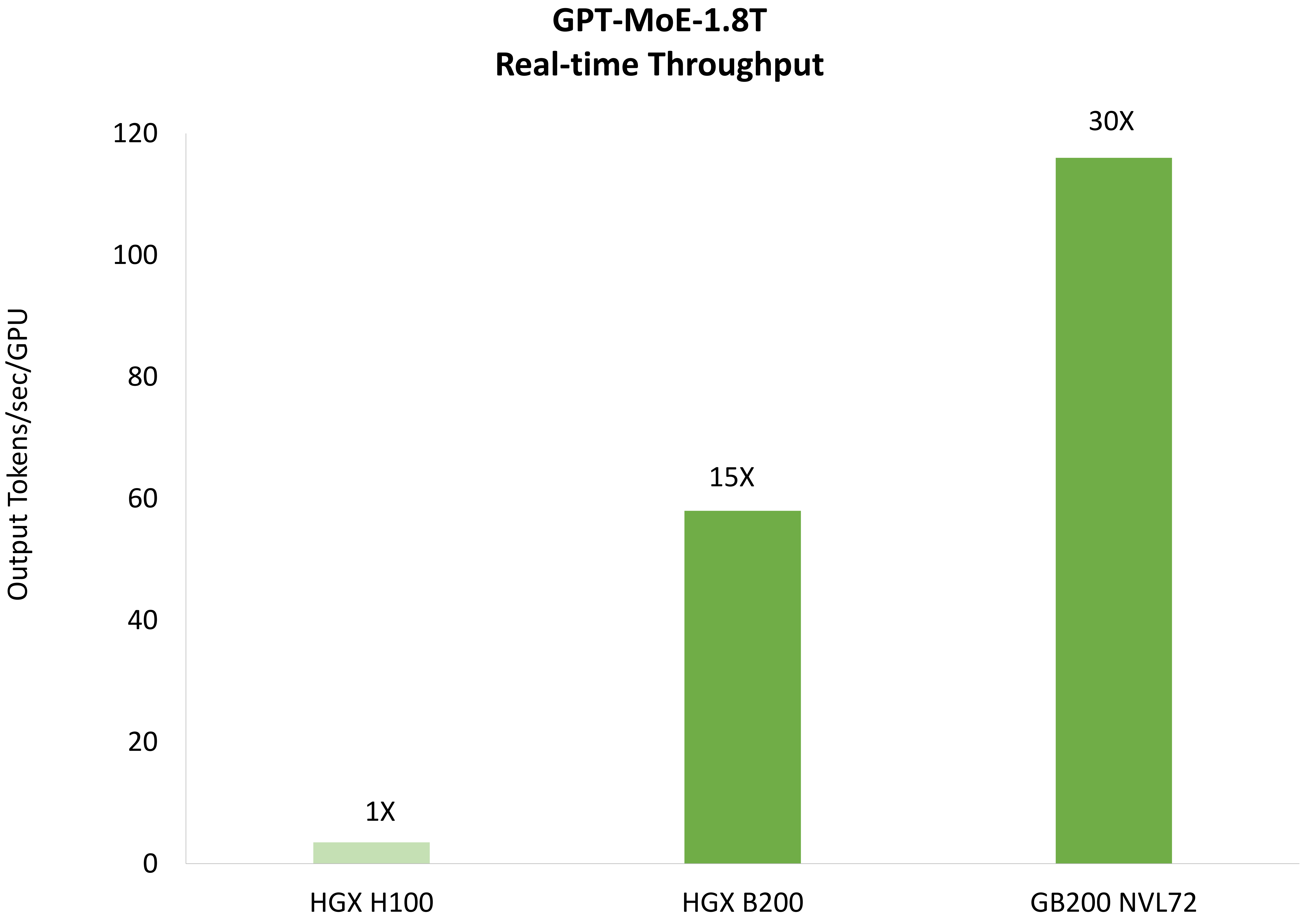
- 6 stories high / 12 MegaWatt cooling capacity (final phase)
- Only €28 Mill total costs / initial phase (33%) in only 1 year construction time
- PUE-factor < 1.05 / Blue Angel eco-label (2020) / European Patent 2020

Future plans at "hessian.AI"

hessian.AI intends to invest a further ca. €10 Mill to directly expand this new AI-Cluster by additional HGX H100 8-way systems. As soon as the preferred contender (HPE) is able to deliver their HGX-based H100 systems the procurement phase will start; presumably mid-2023. With this addition the hessian.AI cluster is poised to become the largest AI-Cluster in Europe.

Performance and Blackwell Miracles

Real-Time Inference for Next Generation Models



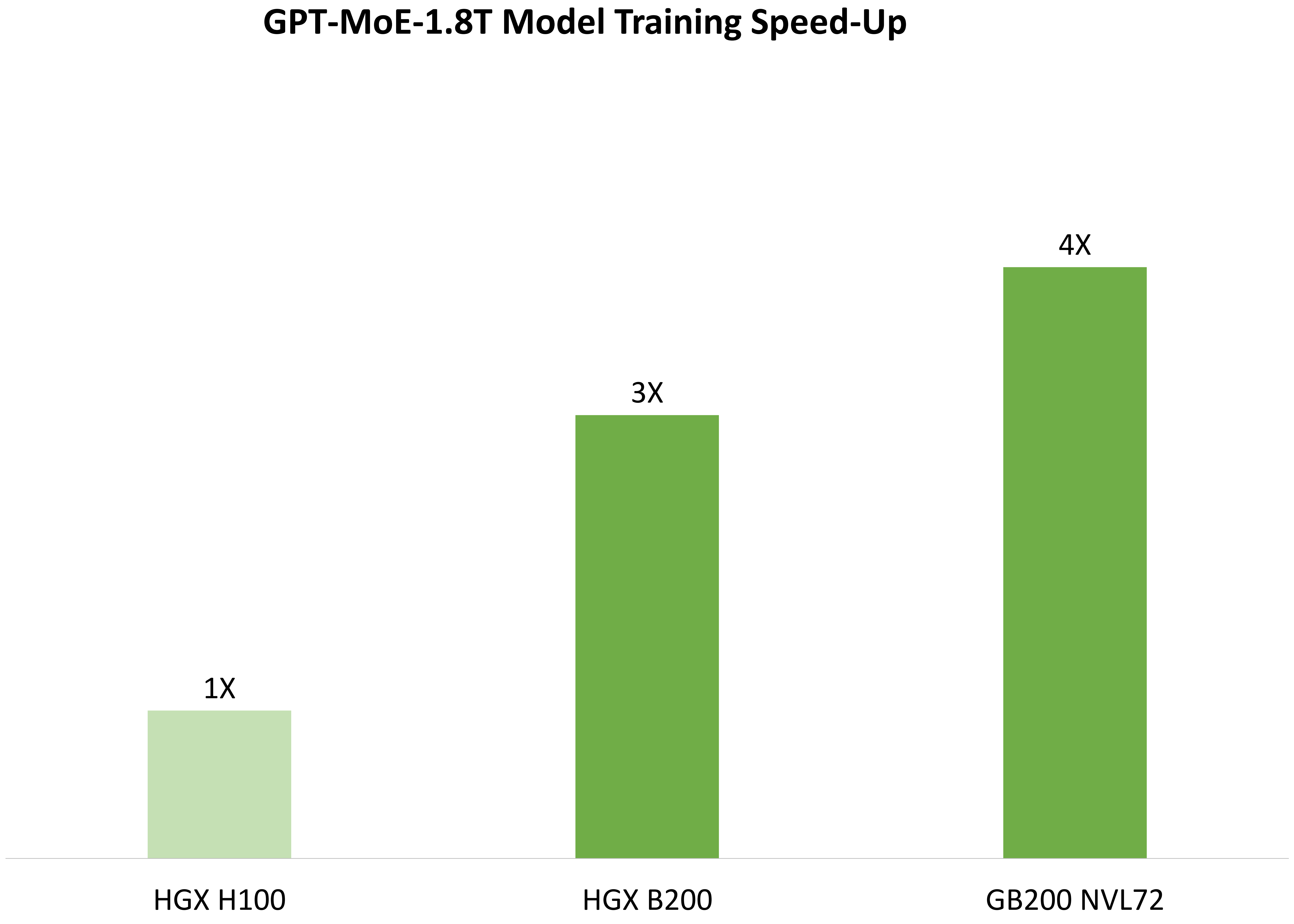
30X
Higher
Throughput

25X
Energy
Efficiency

25X
Lower
TCO

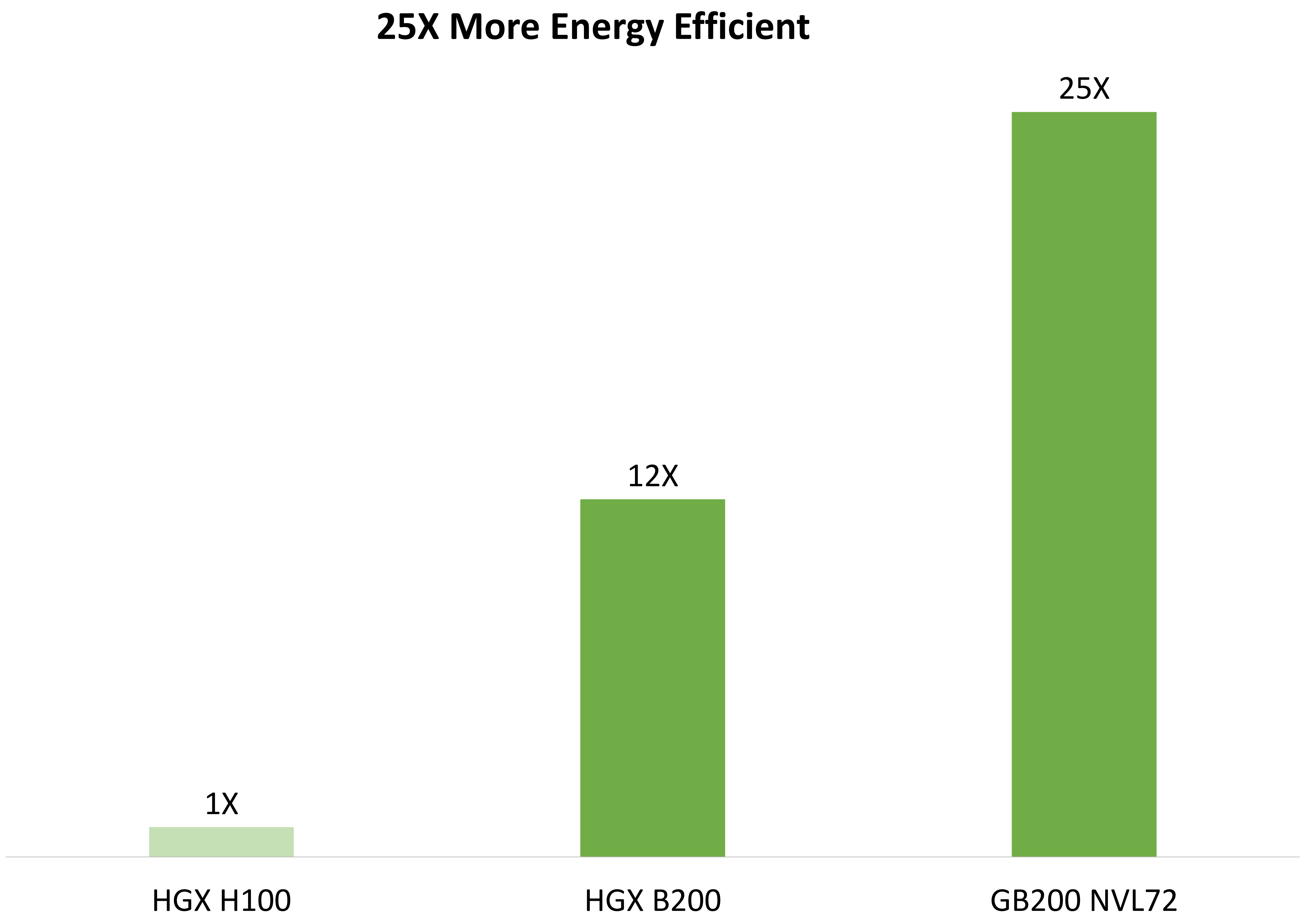
Projected performance subject to change. Token-to-token latency (TTL) = 50 milliseconds (ms) real time, first token latency (FTL) = 5s, input sequence length = 32,768, output sequence length = 1,024 output, 8 HGX H100 air-cooled : 400GB IB Network vs 18 GB200 Superchip liquid-cooled : NVL36, 8x eight-way HGX H100 GPUs air-cooled vs. 1x eight-way HGX B200 air-cooled, per GPU performance comparison.

Supercharged AI Training Performance



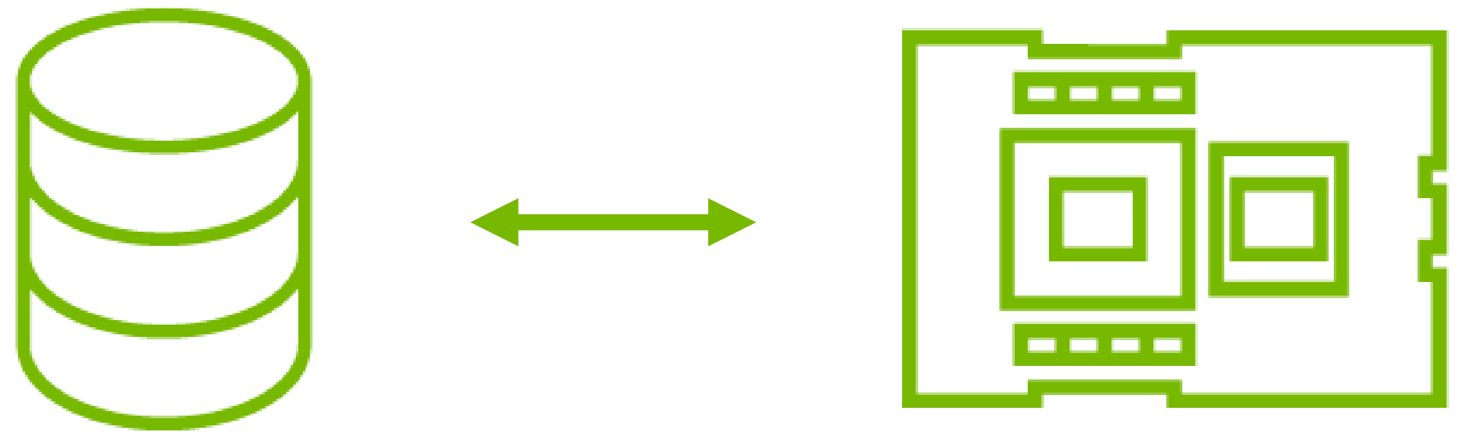
Projected performance subject to change. 32,768 GPU scale, 4,096 HGX H100 air-cooled cluster: 400G IB network, 456 GB200 NVL72 liquid-cooled cluster: 800G IB network

Reduced Energy Use and Lower Cost of Ownership

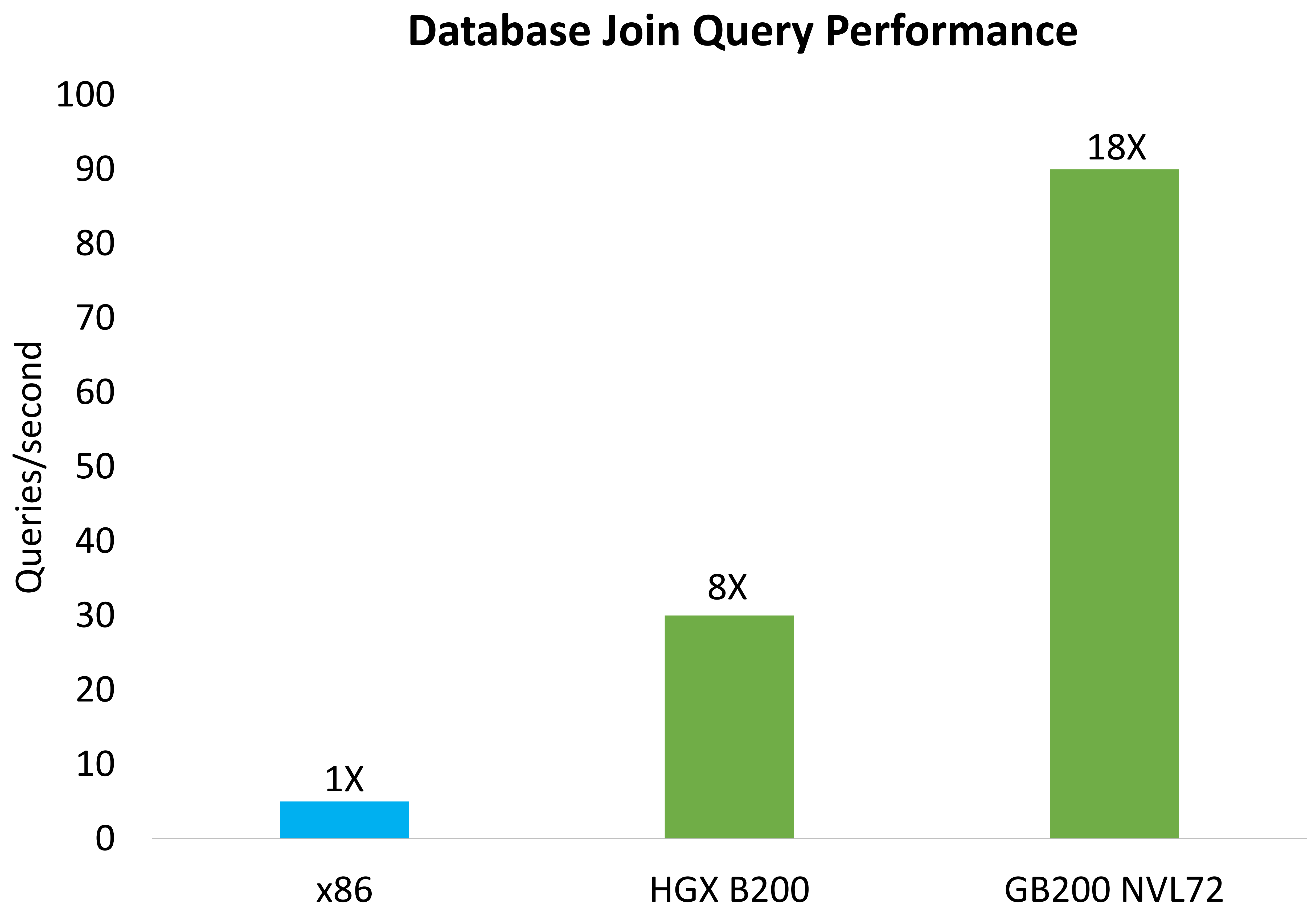


Projected performance subject to change. Token-to-token latency (TTL) = 50ms real time, first token latency (FTL) = 5s, input sequence length = 32,768, output sequence length = 1,028, 8x eight-way HGX H100 GPUs air-cooled vs. 1x eight-way HGX B200 air-cooled, per GPU performance comparison.. TCO and energy savings for 100 racks eight-way HGX H100 air-cooled versus 8 racks eight-way HGX B200 air-cooled with equivalent performance

Accelerating Data Processing with Decompression Engine



- Supported Formats: Deflate, Snappy and LZ4
- Grace CPU memory and high speed C2C link - rapid access to Blackwell for large databases



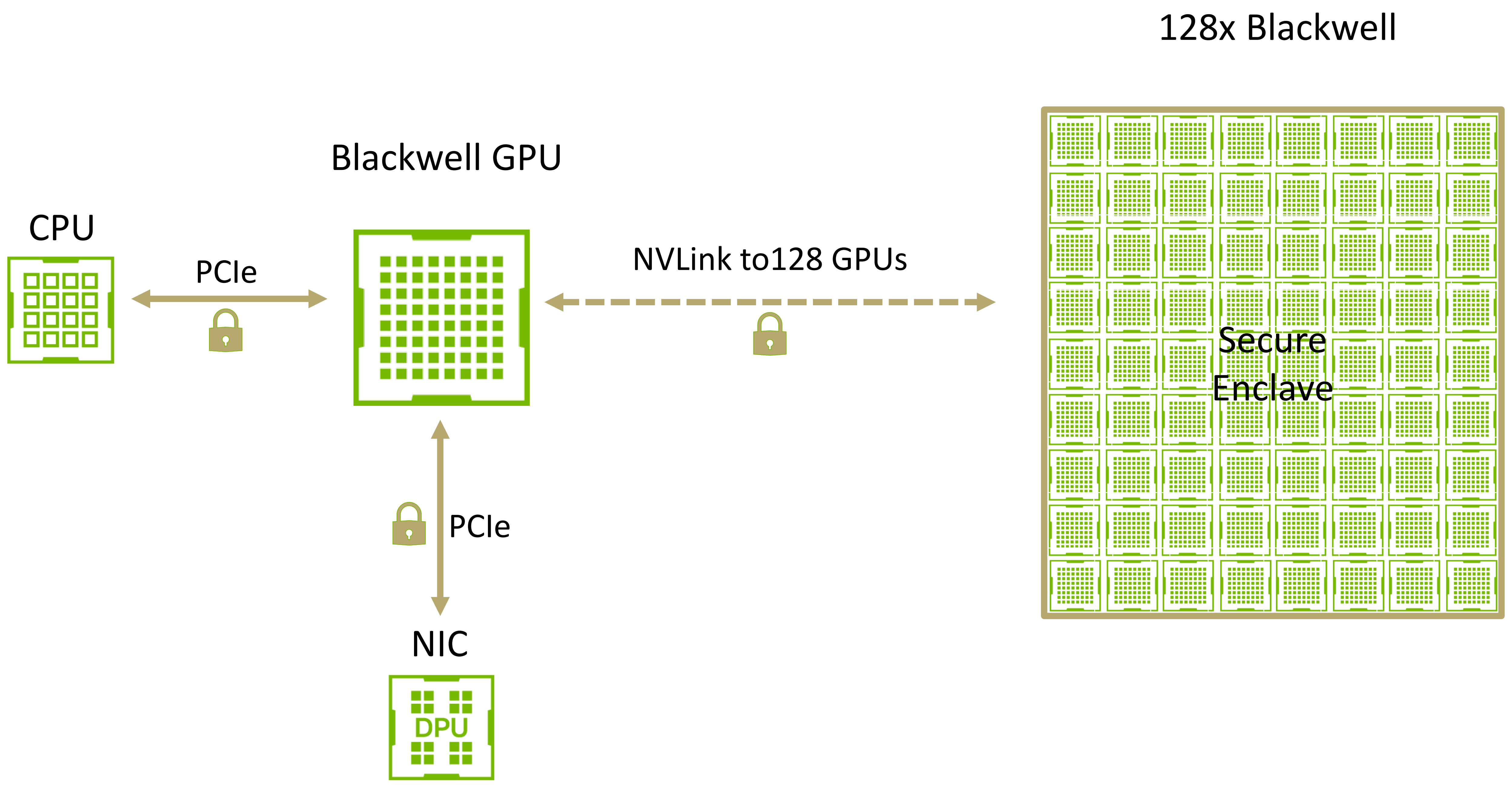
Projected performance subject to change. Database join query with Snappy / Deflate compression derived from TPC-H Q4 query. 1x x86, 1x H100 GPU, and 1x Blackwell GPU from Grace Blackwell GB200 Superchip.

New Era of Secure AI

Confidential Computing for Performant Massive LLMs

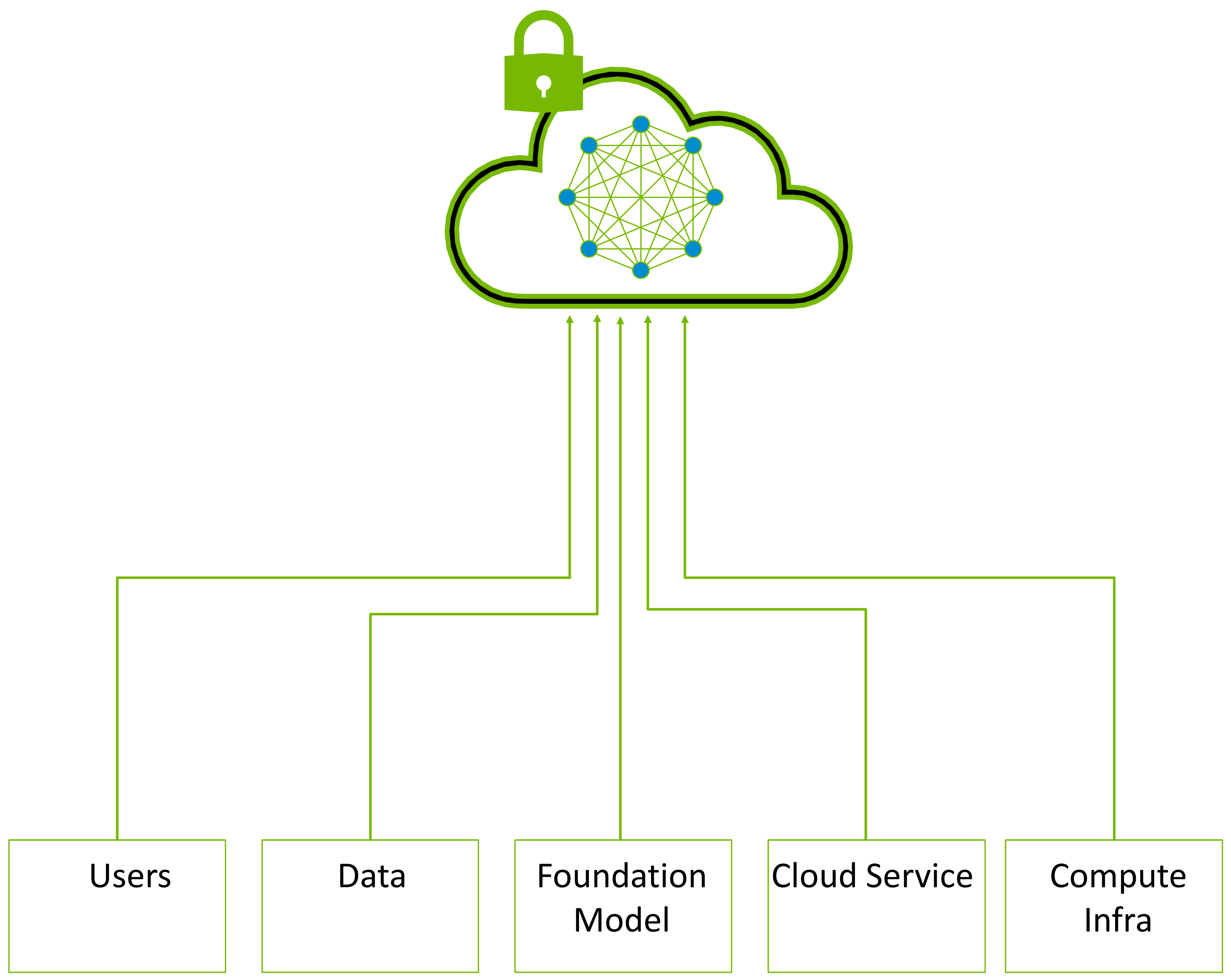
PERFORMANT END-TO-END AI SECURITY

Encrypted on Every Channel
Same Performance



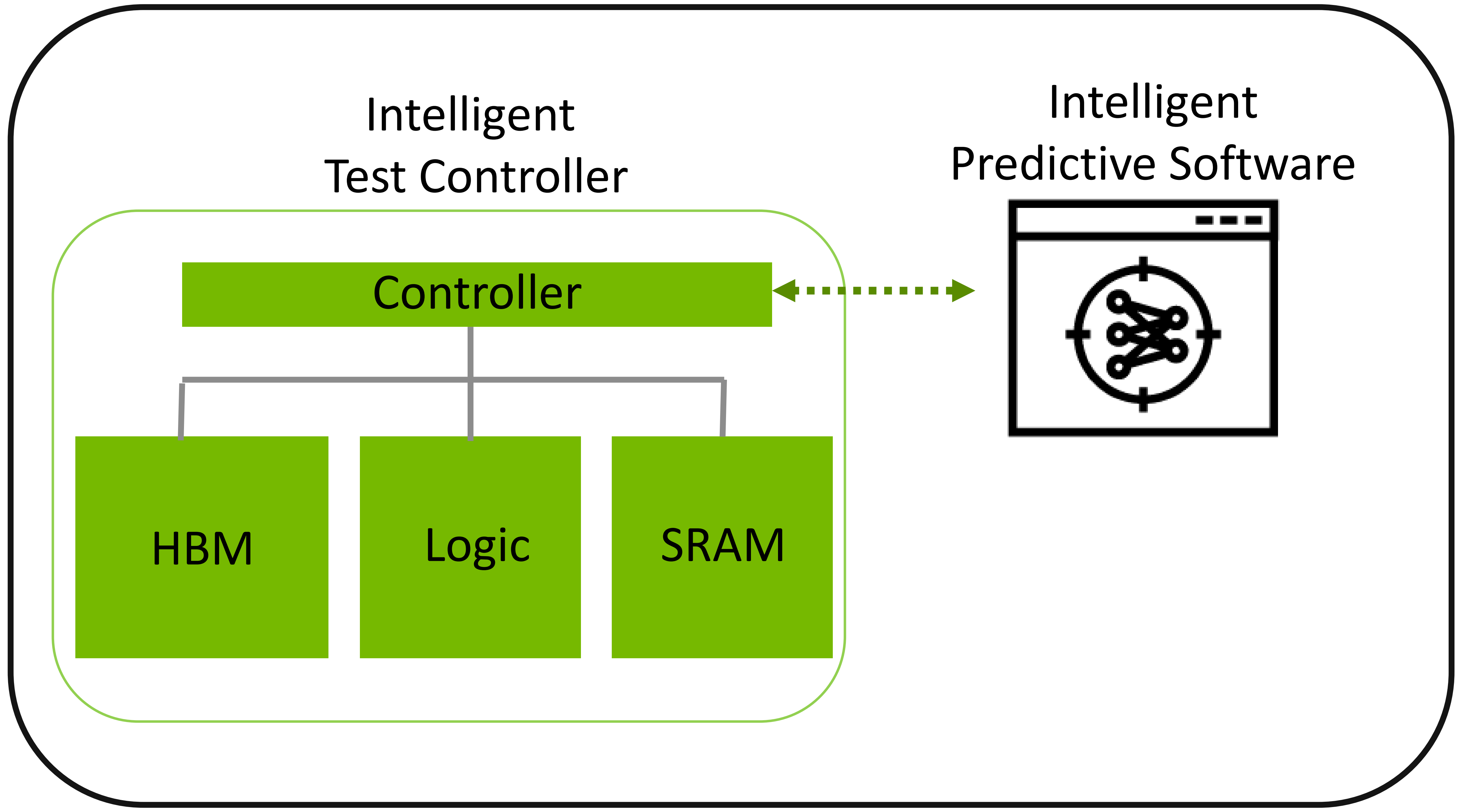
ENABLING DISTRIBUTED AI ECOSYSTEM

Allowing Every Contributor to Protect IP



RAS Engine

Minimize Downtime on Massive at Scale Workloads



2.5M

Test chains

5 Billion

SRAM bits tested

Predict Failures

Minimize unplanned outages

NVIDIA