NVIDIA Blackwell Platform: Advancing Generative AI and Accelerated Computing

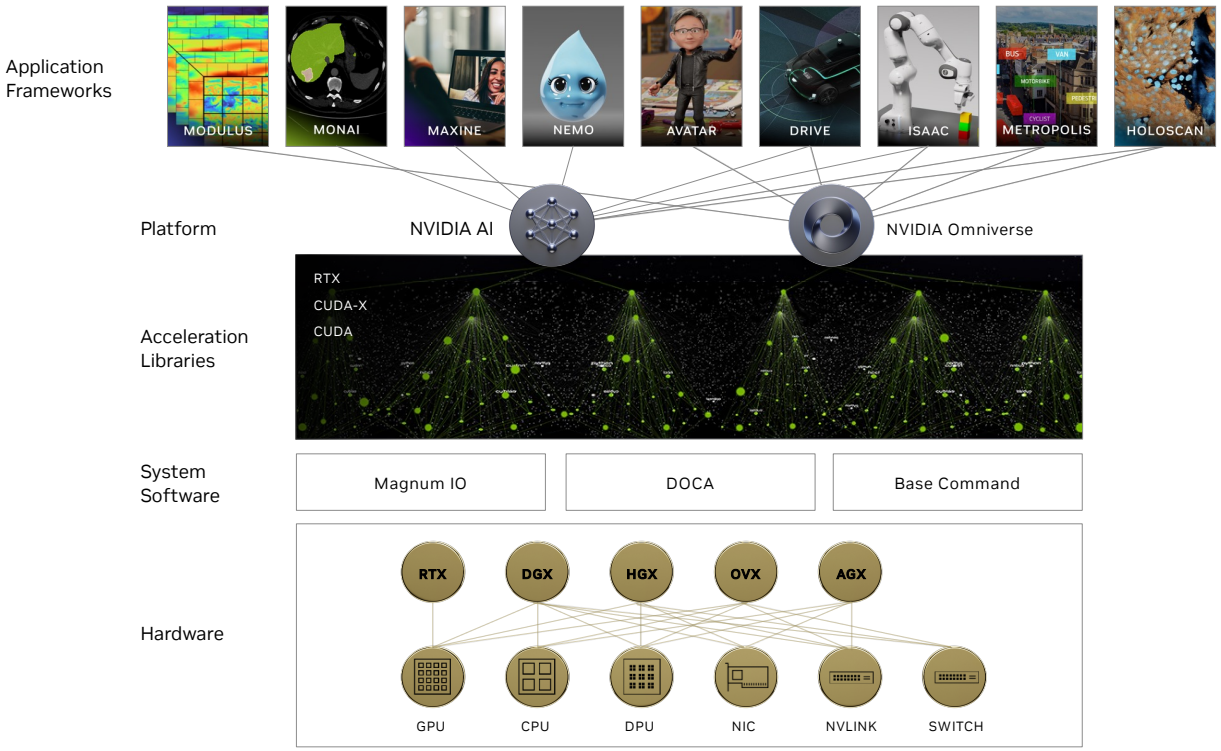
Ajay Tirumala, Raymond Wong | NVIDIA

Agenda

- NVIDIA Blackwell Platform – Data Center Scale Architecture
- Blackwell GPU
- NVIDIA Quasar Quantization System: Enabling Low Precision AI
- Networking for AI, End-to-End Performance and Power Scaling
- Conclusions

NVIDIA Blackwell Platform:
Data Center Scale Architecture

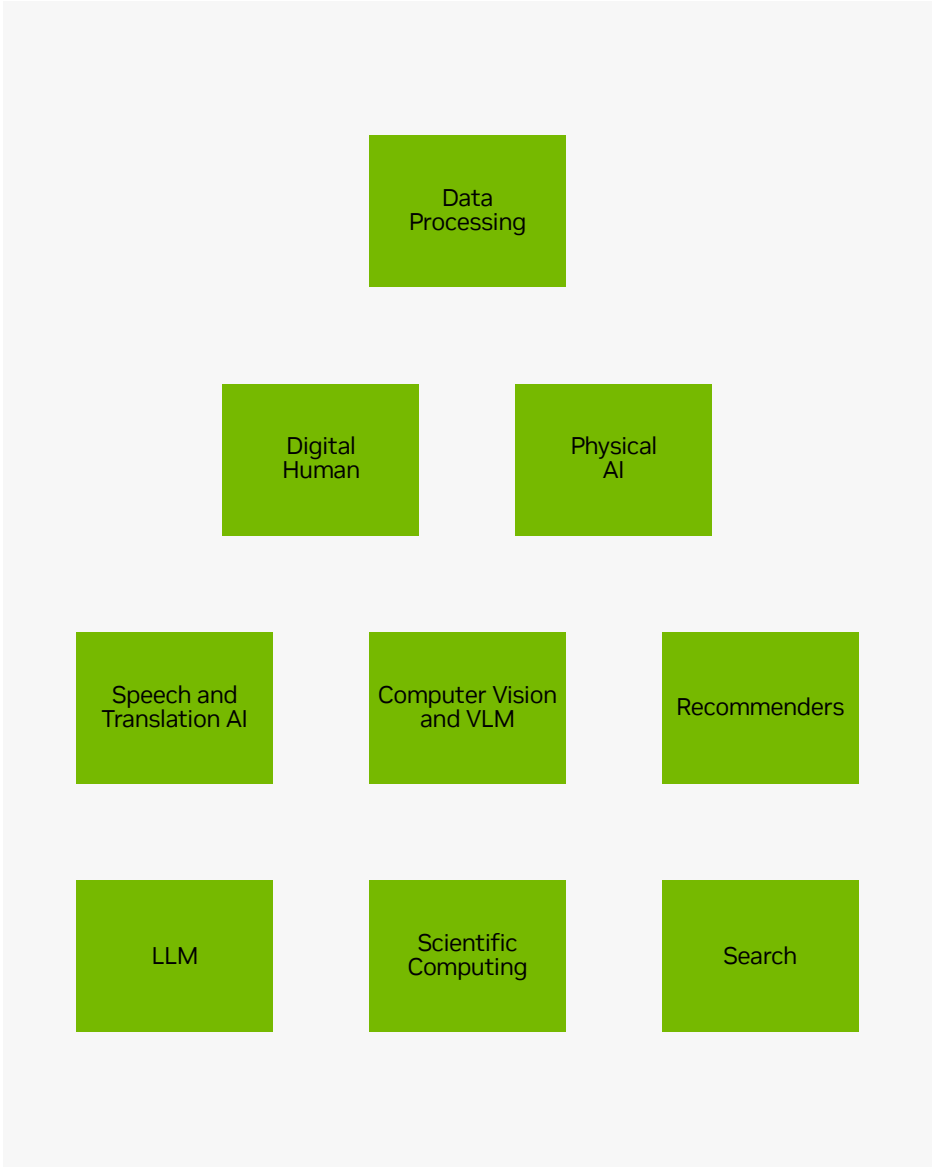
The Full Stack Challenge for AI and Accelerated Computing



Over 400 NVIDIA CUDA-X Libraries

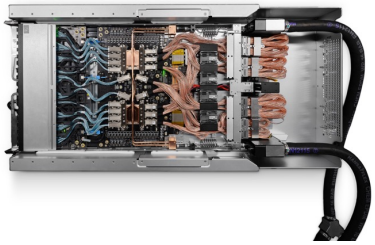
Blackwell optimized to deliver maximum performance

- Optimized libraries for each platform
- Targeting diverse application domains
- Built on our decades-long innovation
- Ever-expanding set of algorithms

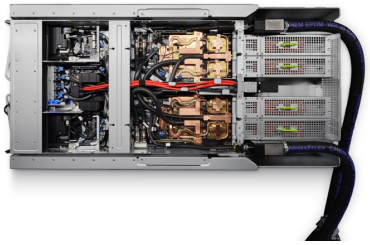


NVIDIA Blackwell Platform

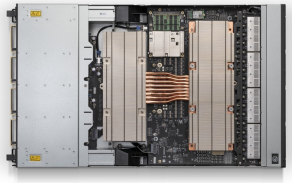
NVSwitch Tray



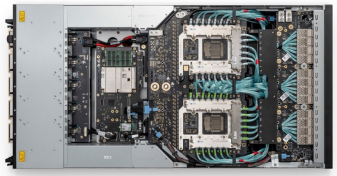
Compute Tray



Spectrum-X800



Quantum-X800



Blackwell GPU

NVIDIA Blackwell

AI Superchip
208B Transistors

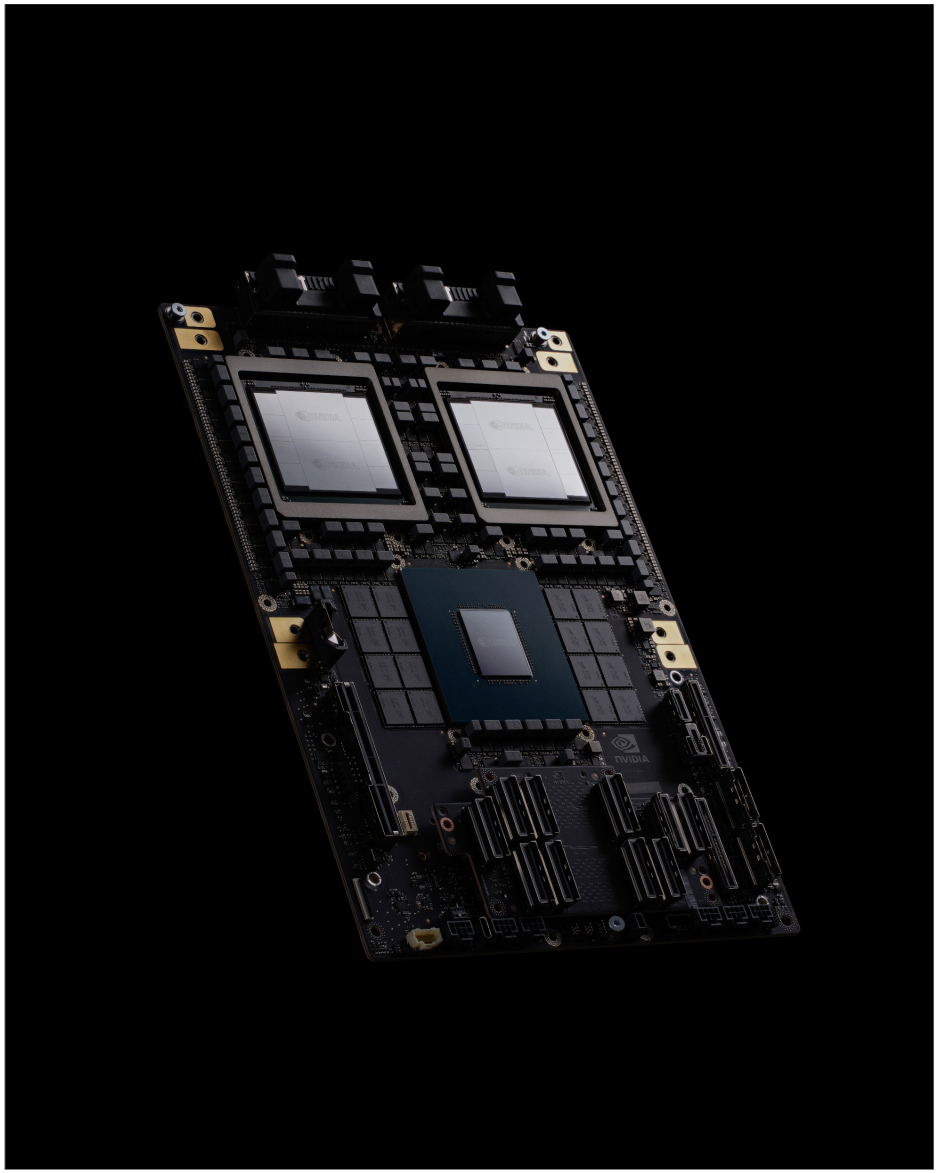
Transformer Engine
FP4/FP6 Tensor Core

Secure AI
Full Performance
Encryption and TEE

5th Generation NVLink
Scales to 576 GPUs

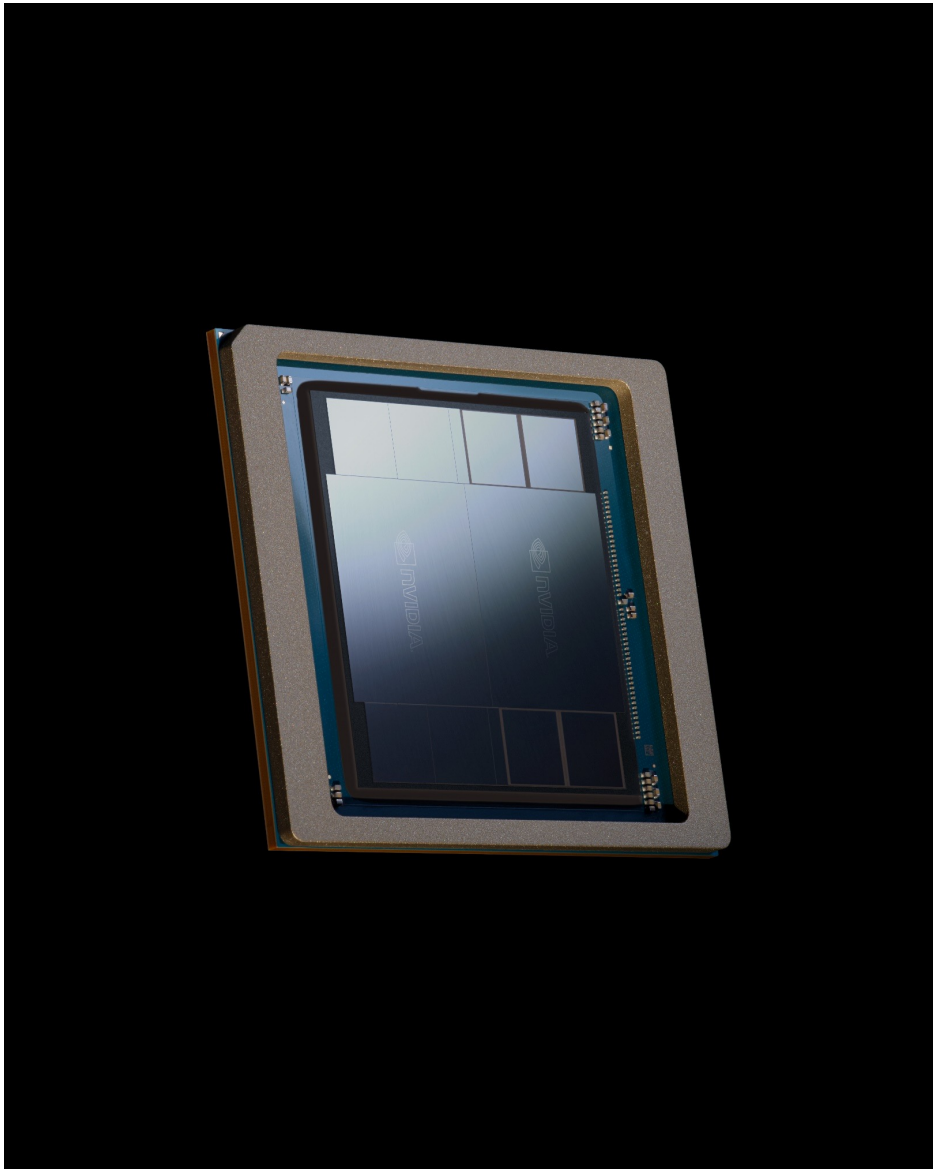
RAS Engine
100% In-System Self-Test

Decompression Engine
800 GB/sec



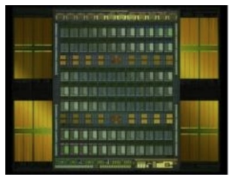
NVIDIA Blackwell GPU

- Highest AI compute, memory bandwidth, and interconnect bandwidth ever in a single GPU
- Two reticle-limited GPUs merged into one:
  - 208B transistors in TSMC 4NP
  - 20 PetaFLOPS FP4 AI
  - 8 TB/s Memory Bandwidth | 8-site HBM3e
  - 1.8 TB/s Bidirectional NVLink bandwidth
  - High-speed NVLink-C2C Link to Grace CPU



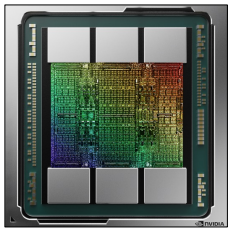
Highest AI Compute in a Single GPU

- Build each GPU to reticle limit as intra-GPU communication provides:
  - Highest communication density
  - Lowest latency
  - Optimal energy efficiency



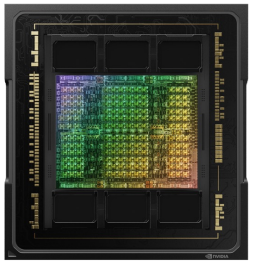
Volta

>21 billion transistors
815mm^2
TSMC 12nm FFN



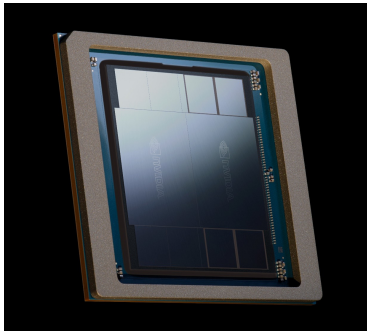
Ampere

>54 billion transistors
826 mm^2
TSMC N7



Hopper

>80 billion transistors
814 mm^2
TSMC 4N



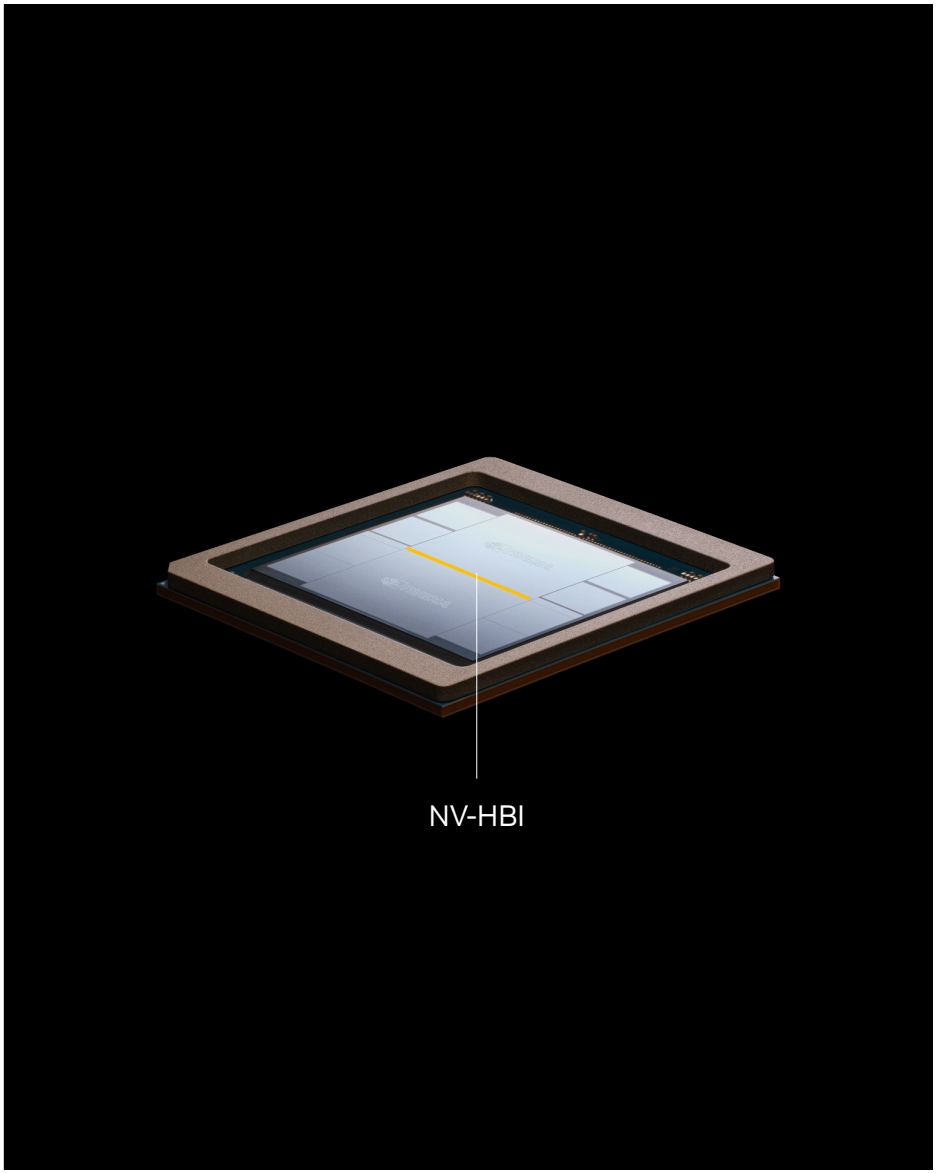
Blackwell

>208 billion transistors
>1600 mm^2
TSMC 4NP

NVIDIA High-Bandwidth Interface (NV-HBI)

Building unified GPU with full performance

- 10 TB/s bi-directional bandwidth across single edge
- Low energy per bit
- Coherent link between GPUs
- Great performance, no compromise



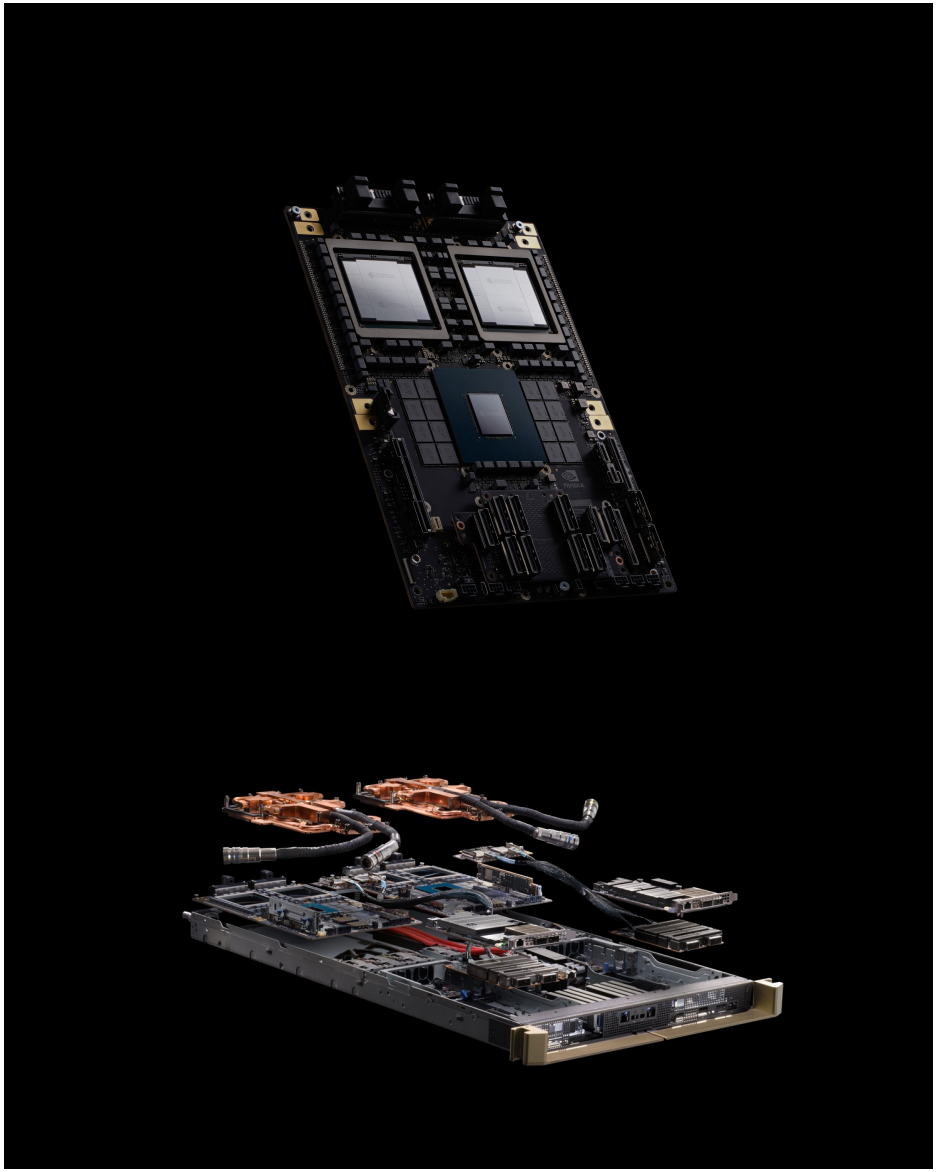
NVIDIA GB200 Grace Blackwell Superchip

- GB200 Grace Blackwell Superchip
  - 1 Grace CPU and 2 Blackwell GPUs
  - NVLink-C2C interconnect
  - 40 PetaFLOPS FP4 | 20 PetaFLOPS FP8
- Grace Blackwell Compute Tray
  - 2 Grace CPUs and 4 Blackwell GPUs

High Bandwidth / Low Latency

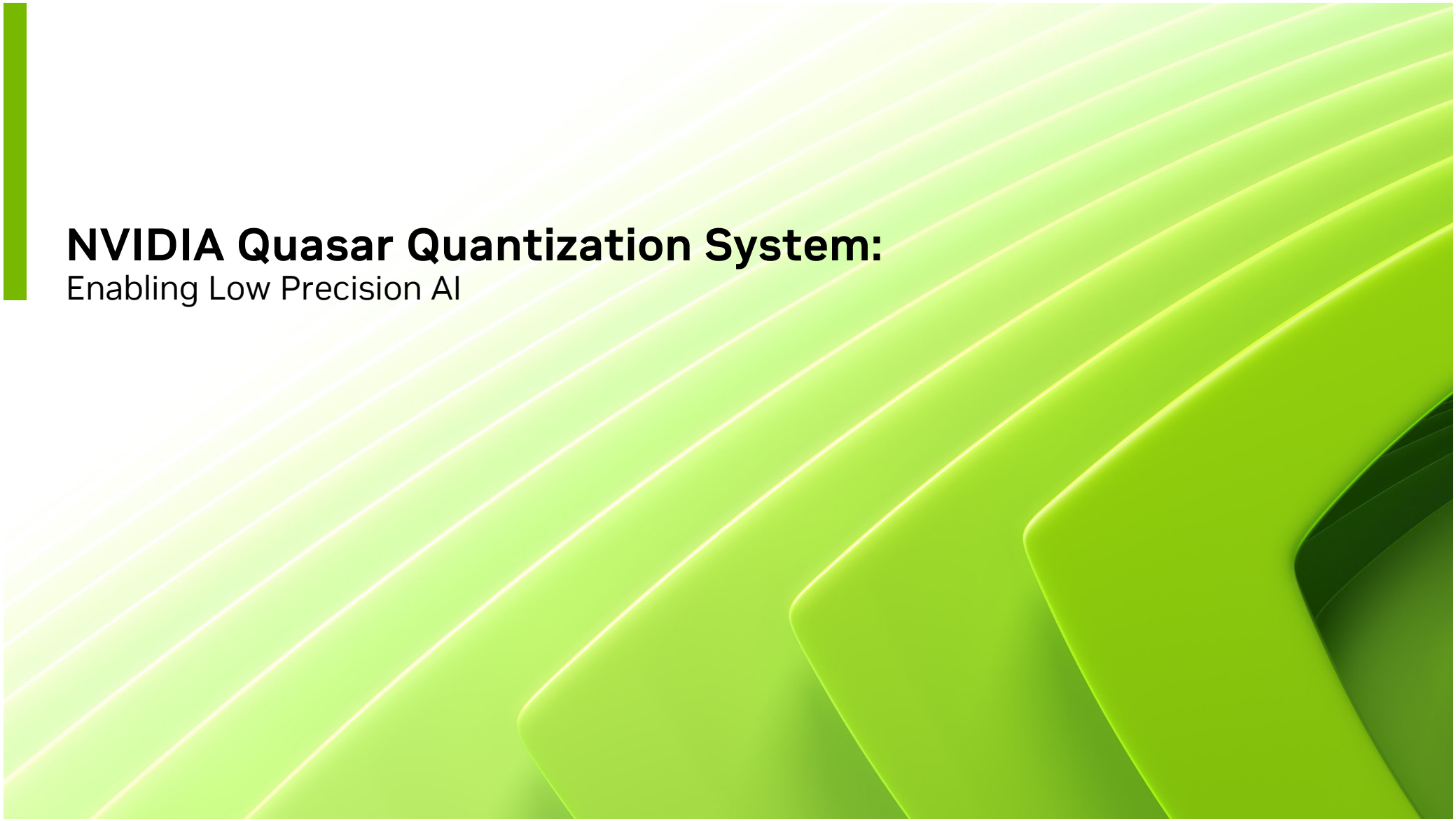
Low-Power High Density

Key Value (KV) Caching



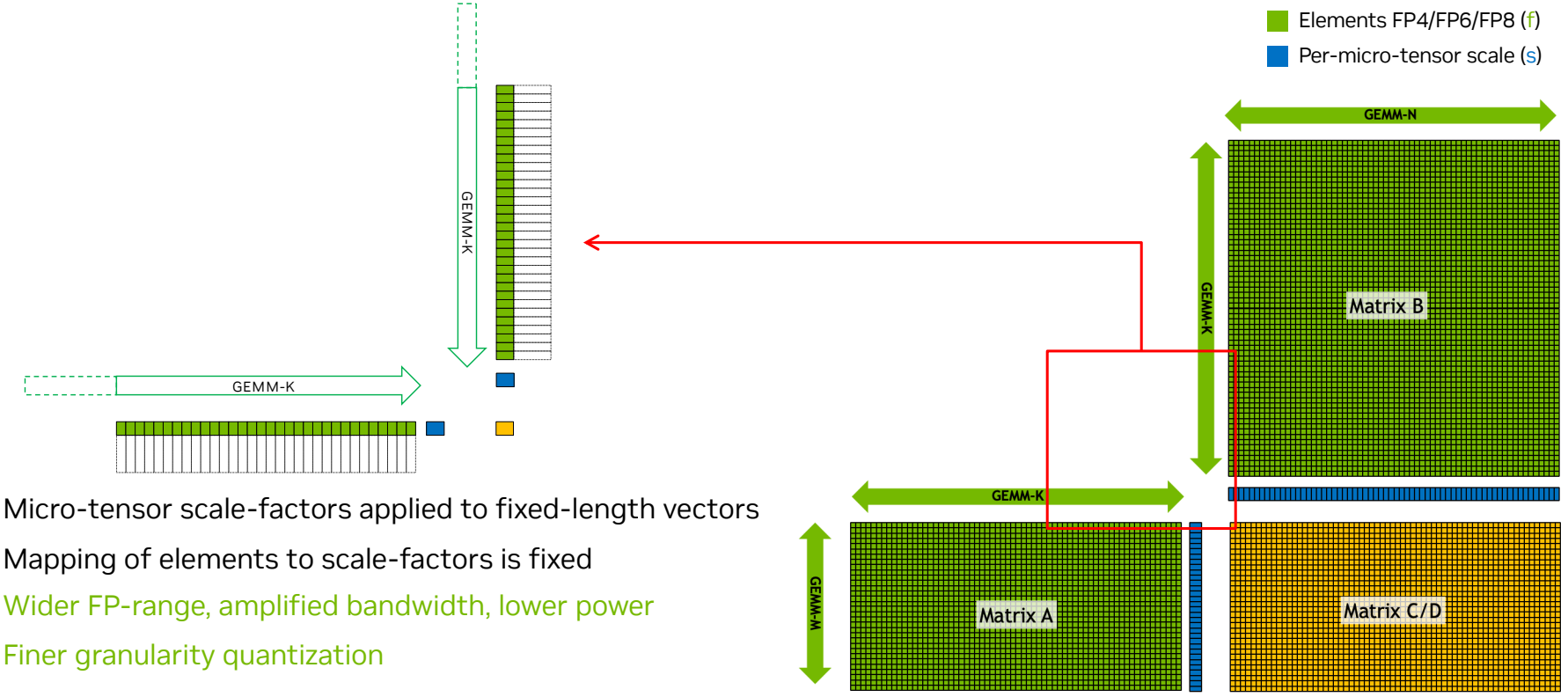
NVIDIA Quasar Quantization System:

Enabling Low Precision AI



5th Gen Tensor Core — New Micro-Tensor Scaled FP Formats

Scaled FP4, FP6 and FP8



- Micro-tensor scale-factors applied to fixed-length vectors
- Mapping of elements to scale-factors is fixed
- Wider FP-range, amplified bandwidth, lower power
- Finer granularity quantization

5th Generation Tensor Cores — FP Formats Summary

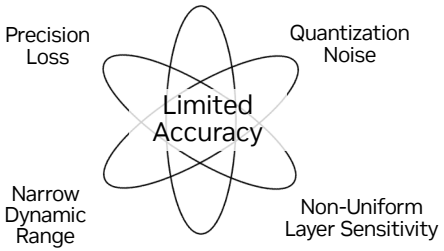
- New FP4 and FP6 precision support
- New micro-scaled formats for FP4, FP6, and FP8
- 4x faster per-clock, per-SM FP4 vs. Hopper FP8
- Blackwell also increases operating frequency and SM count

Format	Hopper SM MACs/clock		Blackwell SM MACs/clock		Blackwell Speedup per clock per SM
	Dense	Sparse	Dense	Sparse	
FP16	2048	4096	4096	8192	2x
BF16	2048	4096	4096	8192	2x
FP8 (+μ-scale)	4096	8192	8192	16384	2x
FP6 (+μ-scale)	-		8192	16384	New! 2x of Hopper FP8
FP4 (+μ-scale)	-		16384	32786	New! 4x of Hopper FP8

NVIDIA Quasar Quantization System and Research

Pushing low precision AI compute beyond limits

Low Precision Challenges



High Accuracy with Low Precision

TRANSFORMER ENGINE HW and SW

Blackwell Tensor Cores
Micro Tensor Scaling
Adaptive Range Calculations

LIBRARIES

TensorRT Model Optimizer
TensorRT
TensorRT-LLM
Megatron-Core
Megatron-LM
cuDNN
.
.
.

LOW PRECISION NUMERICAL ALGORITHMS

Sensitivity Based Layer Selection
Dynamic Range Regularization
Adapters for Reduced Precision
Input Tensor Normalization
Staged Model Quantization

NVIDIA Quasar Quantization System

FP4 Inference Accuracy

- Data from Blackwell silicon with quantized FP4
- Excellent MMLU* scores for LLMs
- Same accuracy even for Nemotron-4 340B model
- Achieved with full-stack hardware, software and algorithm co-design

*Massive Multitask Language Understanding

Blackwell FP4 and NVIDIA Quasar Quantization System Measured MMLU Scores

Model	BF16	Quantized FP4
Nemotron-4 15B	64.2	64.5
Nemotron-4 340B	81.1	81.1

Nemotron-4 340B Technical Report
https://d1qx31qr3h6wln.cloudfront.net/publications/Nemotron_4_340B_8T_0.pdf

Image Generation on Blackwell Silicon

Using FP4-quantized model

Model using FP16

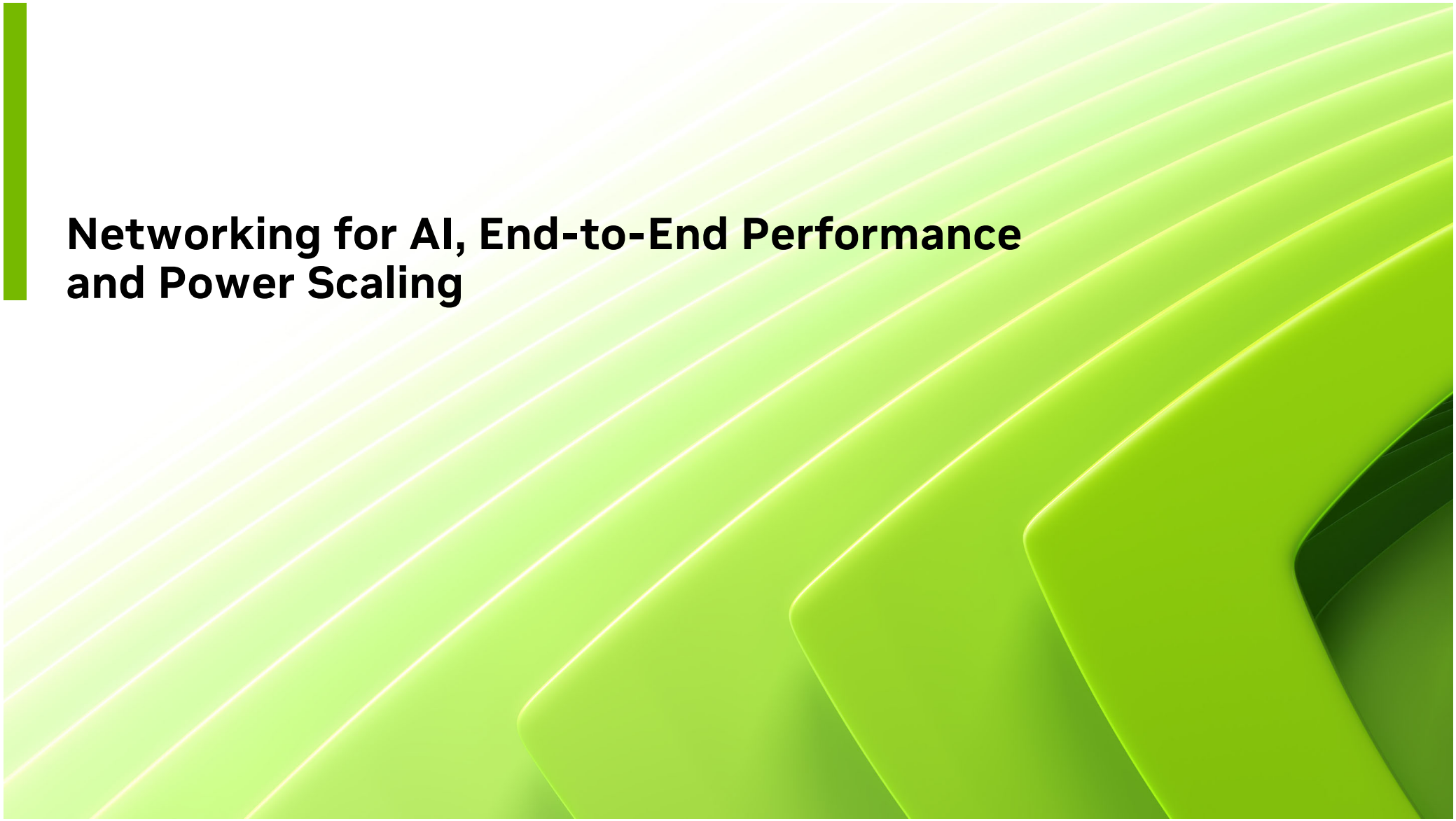
Model using FP4

Prompt: Close up photo of a rabbit, forest in spring, haze, halation, bloom, dramatic atmosphere, centered, rule of thirds, 200mm 1.4f macro shot

Outputs generated using SDXL 1.0 Base. FP4 result generated with NVIDIA Quasar Quantization System.

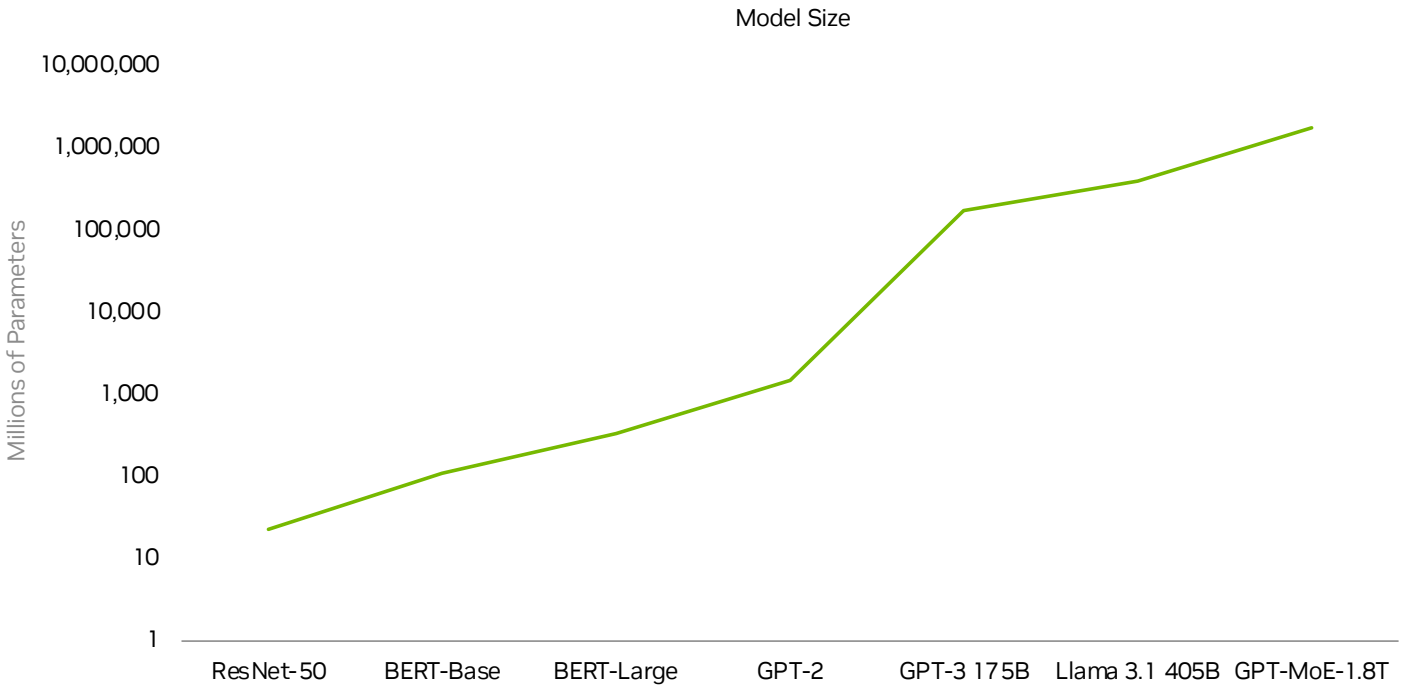
NVIDIA

Networking for AI, End-to-End Performance and Power Scaling



AI Models Growing Exponentially

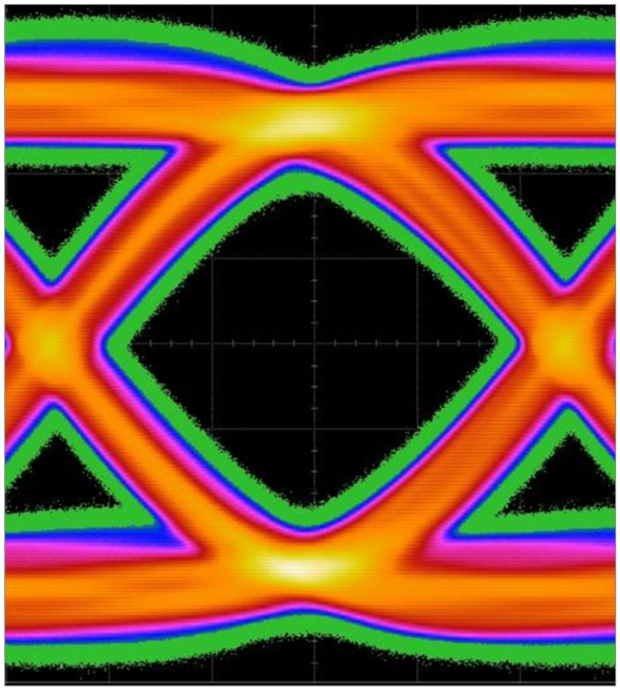
Need for multi-GPU inference at scale



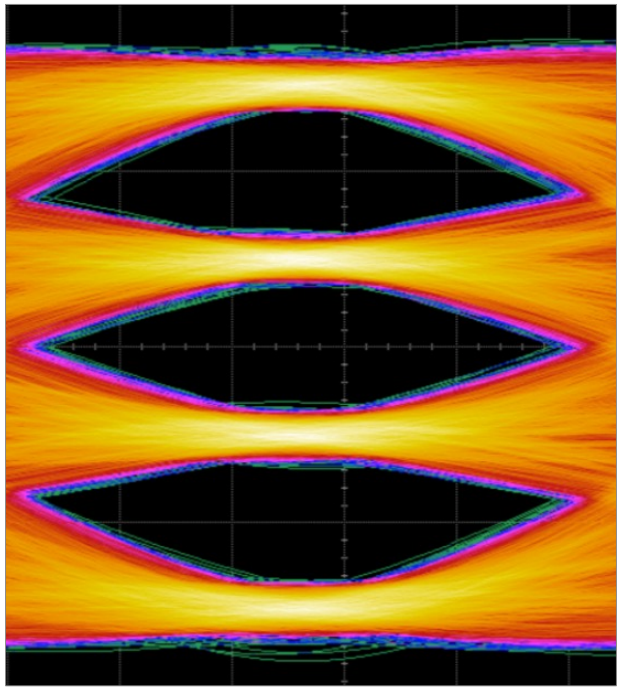
New Capabilities | Trillions of Parameters | 70,000X Growth in a Decade

World-Class NVLink PHY Performance

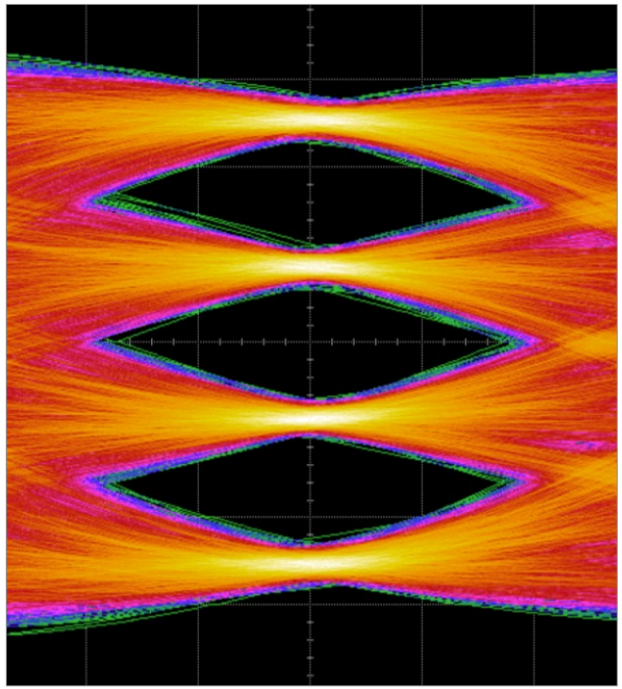
5th generation NVLink PHY drives high-speed interconnect



Ampere | NVLink3
12 NVLinks | 50GB/s each
x4@50Gbps-NRZ
600GB/s total



Hopper | NVLink4
18 NVLinks | 50GB/s each
x2@100Gbps-PAM4
900GB/s total

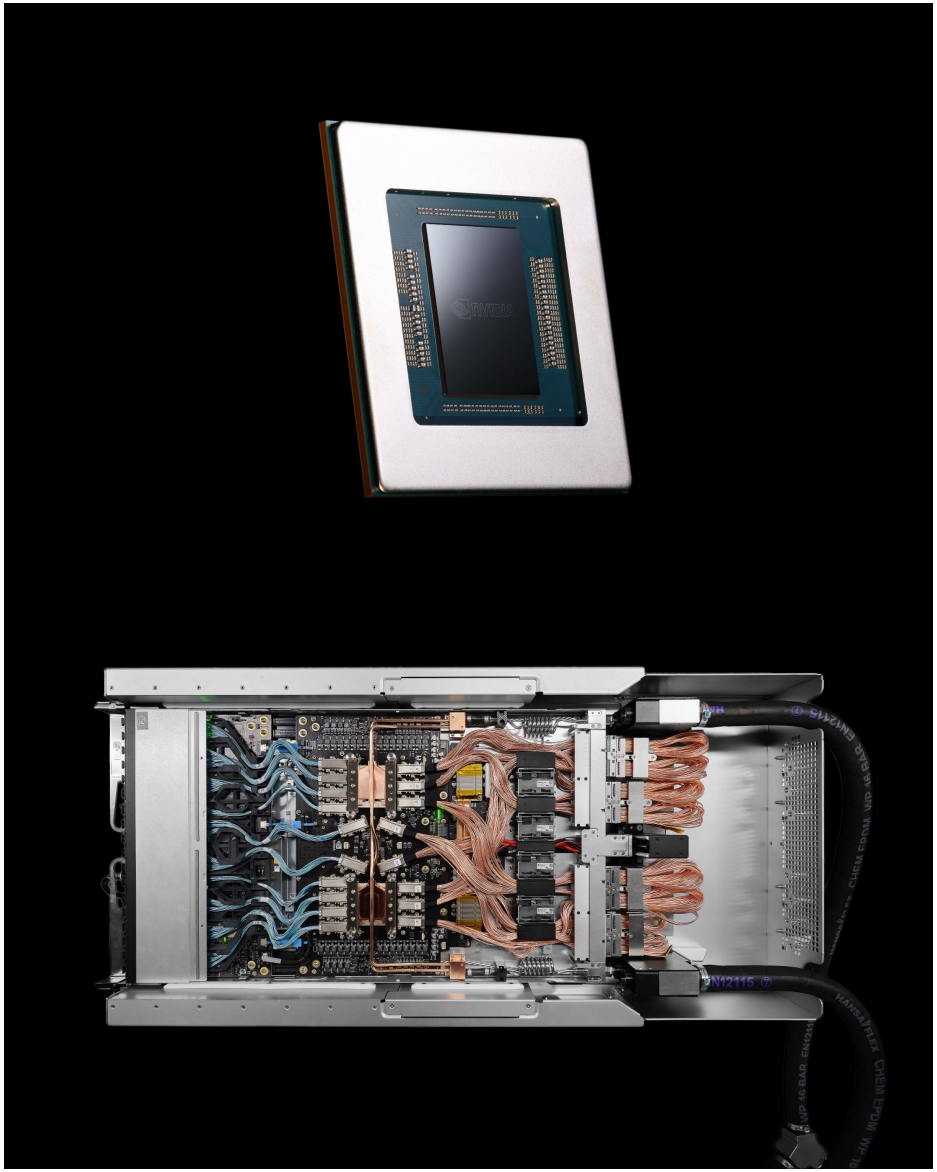


Blackwell | NVLink5
18 NVLinks | 100GB/s each
x2@200Gbps-PAM4
1800GB/s total

4th Generation NVLink Switch Chip and NVLink Switch Tray

Faster scale-up interconnect

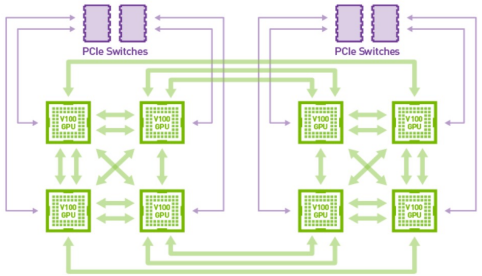
- NVLink Switch Chip
  - >800 mm2 in TSMC 4NP
  - Extends NVLink to 72 GPUs on GB200 NVL72
  - 7.2 TB/s full all-to-all bidirectional BW over 72 ports
  - SHARP* In-Network Compute — 3.6 TFLOPS
- NVLink Switch Tray
  - 2x NVLink Switch chips
  - 14.4 TB/s total bandwidth



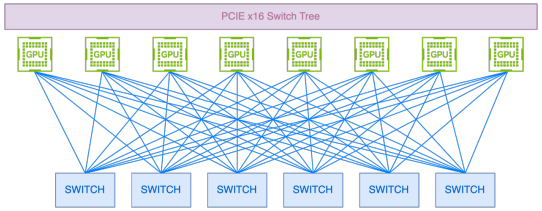
** Scalable Hierarchical Aggregation and Reduction Protocol*

NVLink Switch and NVLink Domain

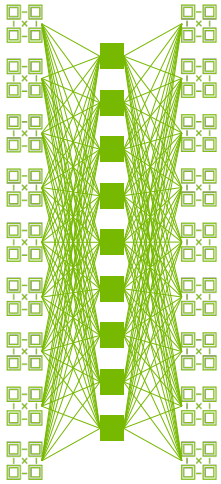
Critical for multi-GPU inference



2016
Hybrid Cube Mesh NVLink technology



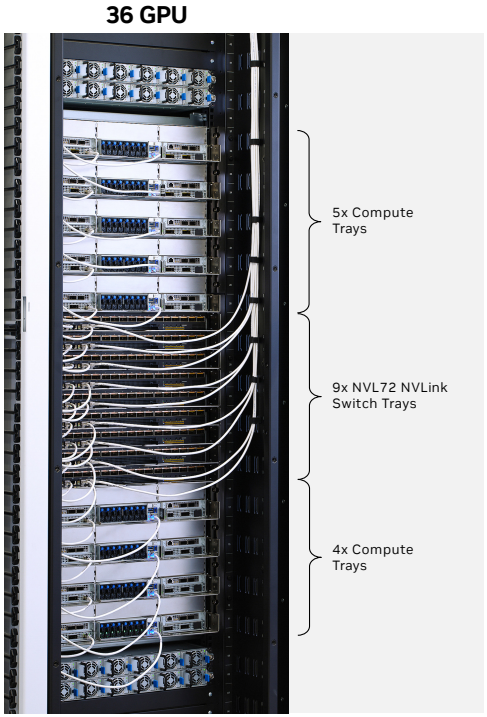
2022
3rd Gen NVLink Switch
All-to-all connection among NVLink domain of 8 GPU



2024
4th Gen NVLink Switch Chip
All-to-all connection among NVLink domain of 72 GPU

GB200 NVL72 & NVL36

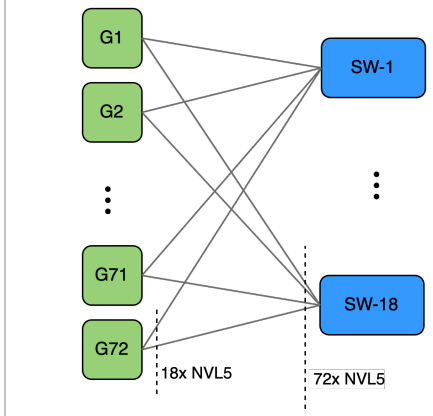
Delivering new unit of compute



GB200 NVL72	36 Grace CPUs 72 Blackwell GPUs Fully connected NVLink Switch rack

Training	720 PFLOPs
Inference	1,440 PFLOPs
NVL Model Size	27 Trillion params
Multi-Node Bandwidth	130 TB/s
Multi-Node All-Reduce	260 TB/s

GB200 NVL36	18 Grace CPUs 36 Blackwell GPUs Fully connected NVLink Switch rack

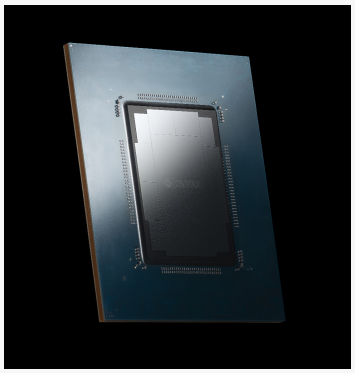


- 72 GPUs fully connected to 18 NVLink Switch
- 18 NVL5 ports per GPU
- 72 NVL5 ports per Switch

Spectrum-X

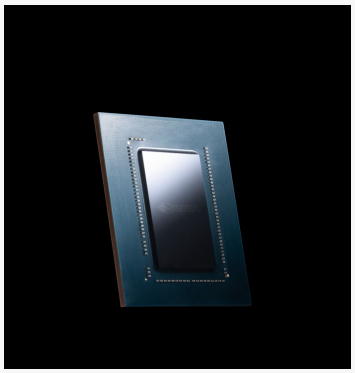
World's first Ethernet fabric built for AI

- End-to-end optimized for cloud AI workloads
- RoCE adaptive routing over lossless Ethernet
- Congestion control, multi-tenant traffic isolation
- Increase effective bandwidth from 60% to 95%



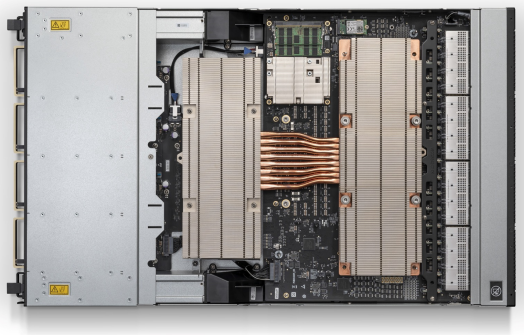
Spectrum-4

- 100B transistors
- 51.2T bandwidth
- 64 X 800G, 128 X 400G



Bluefield-3

- 16 Arm A78 Cores
- 16 Core/ 256 Thread Datapath Accelerator
- 400Gb/s Ethernet



Spectrum-X800

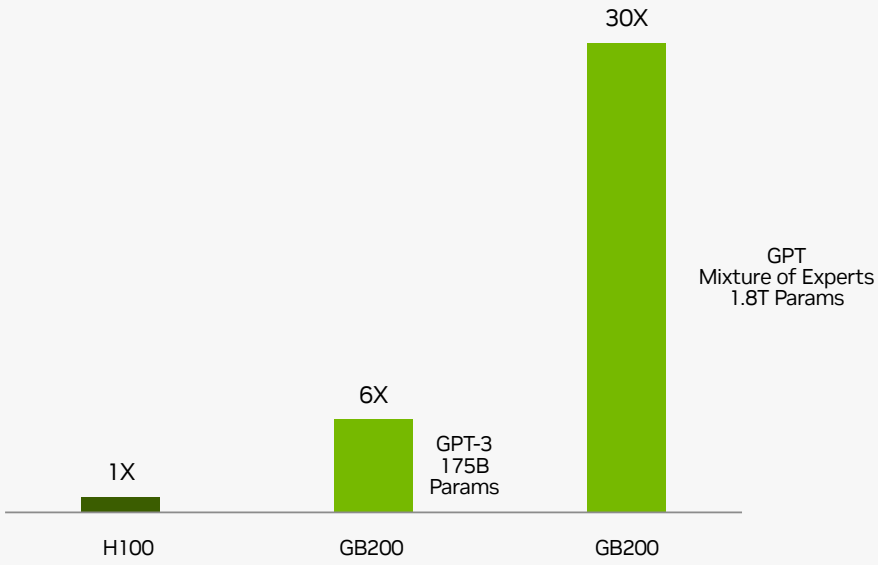
GB200 NVL72 Enabling Trillion Parameter AI

30x real-time MoE inference,
25x improved energy efficiency

30X Higher Throughput

25X Energy Efficiency

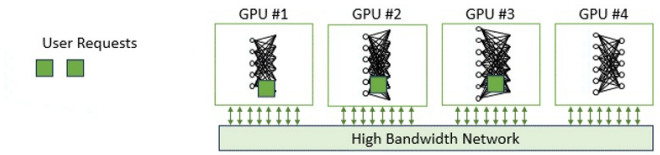
25X Lower TCO



Projected performance subject to change
Token-to-token latency (TTL) = 50 milliseconds (ms) real time
GPT-3 175B: First token latency (FTL) 2s; input sequence length = 2,048, output sequence length = 128, 4 HGX H100 air-cooled 400GB IB Network vs 2 GB200 Superchips liquid-cooled NVLink; per GPU performance comparison,
GPT-MoE-1.8T: FTL = 5s; input sequence length = 32,768, output sequence length = 1,024 , 8 HGX H100 air-cooled 400GB IB Network vs 18 GB200 Superchips liquid-cooled NVL36; per GPU performance comparison

Splitting Work Across Multiple GPUs — Pipeline and Tensor Parallel

Pipeline Parallel

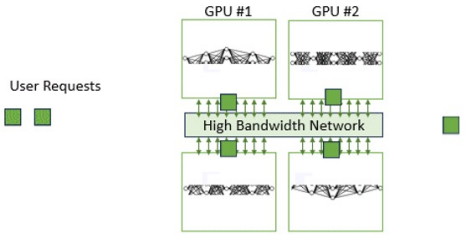


Each network layer runs in a distinct GPU

Latency is sum of execution and data transfer to adjacent GPU

Seldom used alone for inference

Tensor Parallel



Each layer split across multiple GPUs

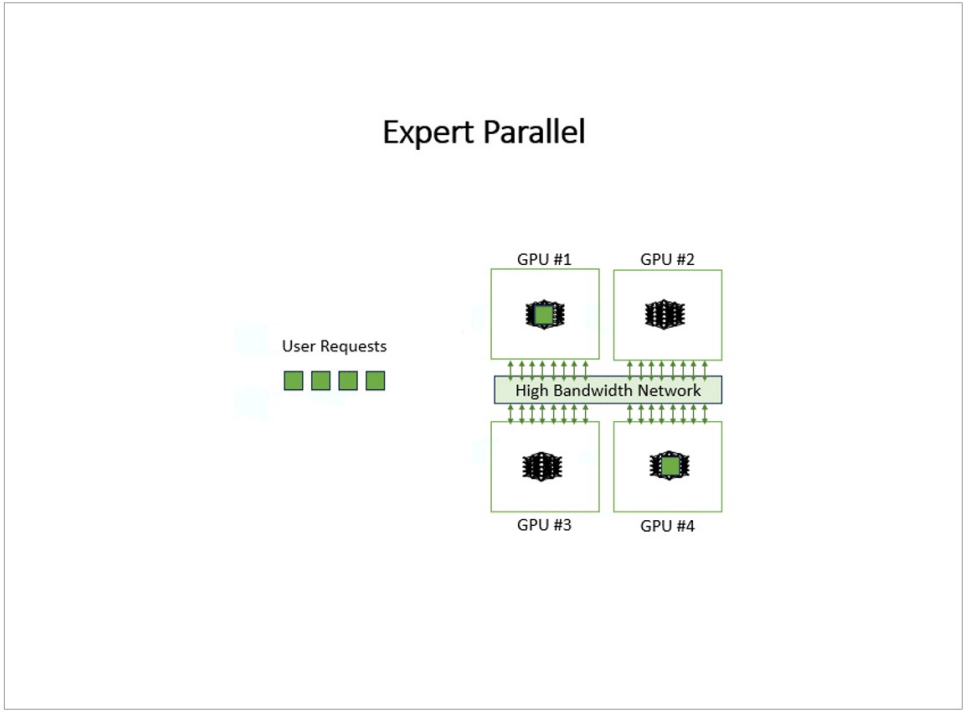
Recombined over a high bandwidth network

Improves interactivity

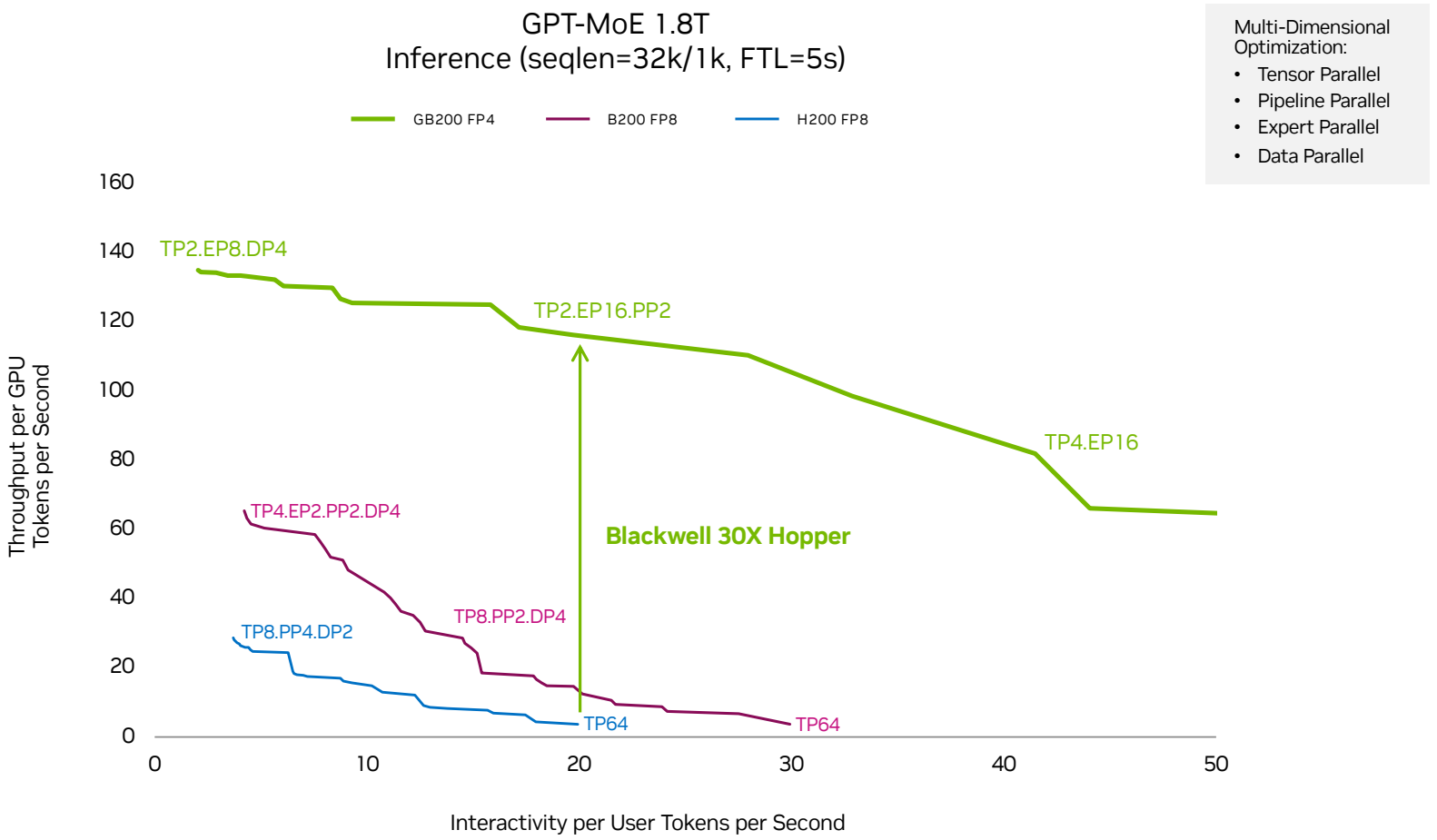
Primary strategy for Hopper at desired latencies

Splitting Work for Large Models: Expert Parallel

GB200 NVL72 unlocks GPT 1.8T MoE EP16



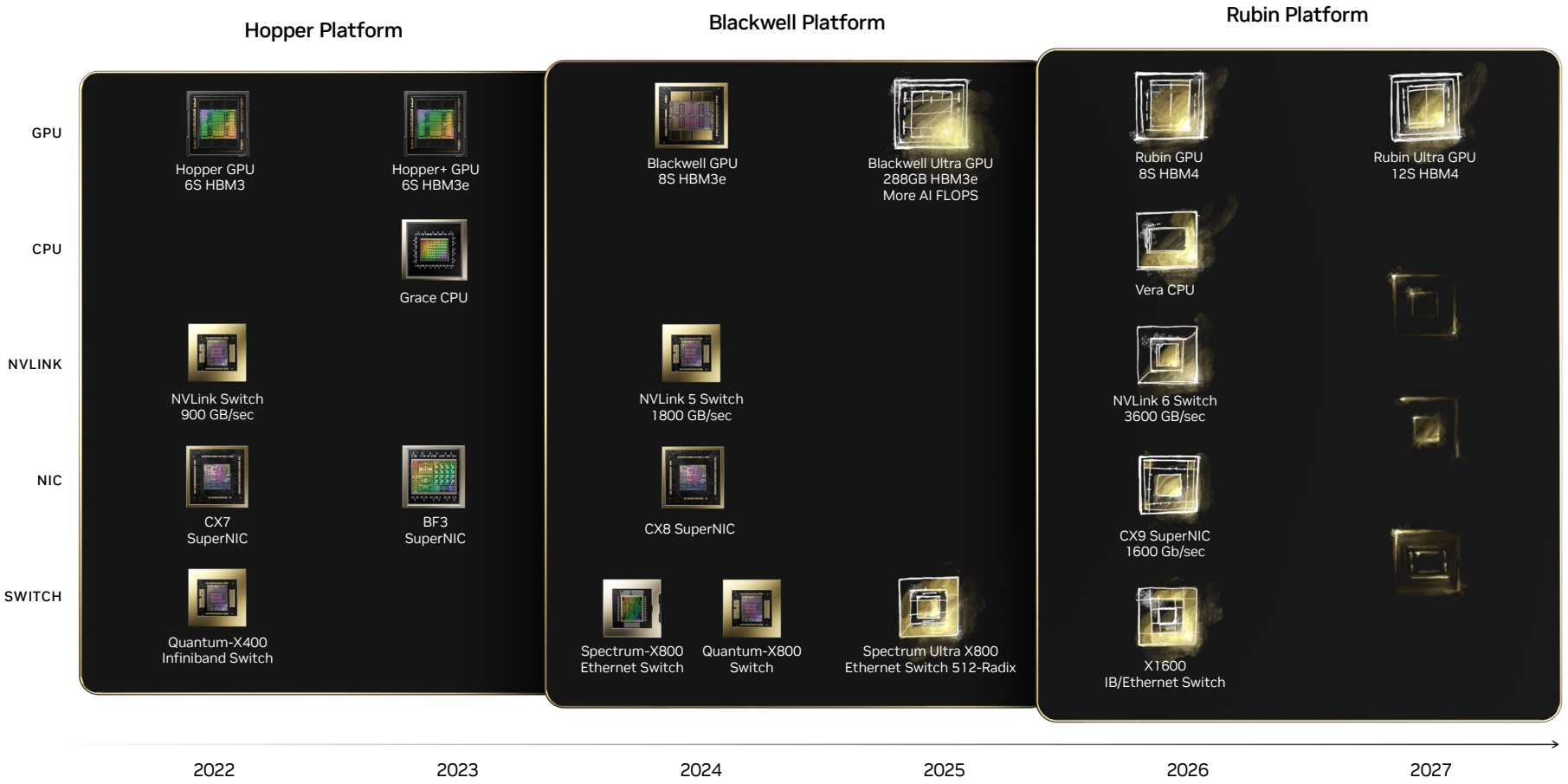
- Blackwell enables entire expert in one GPU
- Needs high bandwidth, low latency interconnect
- More efficient math utilization than tensor parallel
- Primary strategy for Blackwell at desired latency



LLM inference: TTL = 50 milliseconds (ms) real time, FTL = 5s, 32,768 input/1,024 output, NVIDIA HGX™ H200 and HGX B200 scaled over InfiniBand (IB) vs. GB200 NVL72. Projected performance subject to change.

Conclusion

Datacenter Scale | One-Year Rhythm | Technology Limits | One Architecture



Blackwell Summary

- Full-stack, data center scale platform : GPU, CPU, NVSwitch, DPU, NIC, Spectrum and Quantum switches
- NVIDIA Quasar Quantization System brings hardware-software codesign, libraries and algorithmic innovations together for low precision AI
- Over an order-of-magnitude more performance for real-time trillion parameter LLM inference
- Significant performance and power improvements for AI training, inference and accelerated computing

ACKNOWLEDGEMENT:
Thousands of NVIDIA employees made Blackwell happen.
Sincere thanks and gratitude to present on their behalf.