Accelerating the Future: Triton on Blackwell Architecture

Pradeep Ramani, NVIDIA

Thomas Raoux, OpenAI

Agenda

- Blackwell GPU
- NVIDIA-OAI collaboration
- Looking ahead
- Conclusions

NVIDIA Blackwell

AI Superchip
208B Transistors

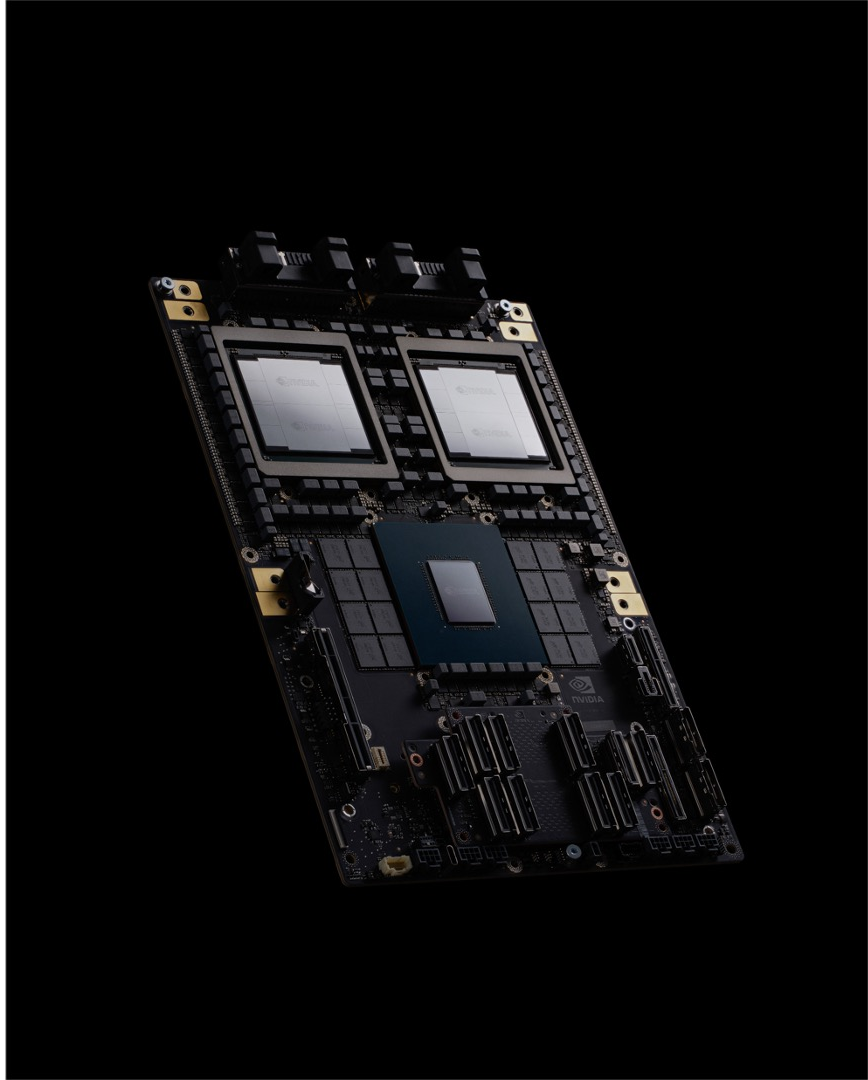
Transformer Engine
FP4/FP6 Tensor Core

Secure AI
Full Performance
Encryption and TEE

5th Generation NVLink
Scales to 576 GPUs

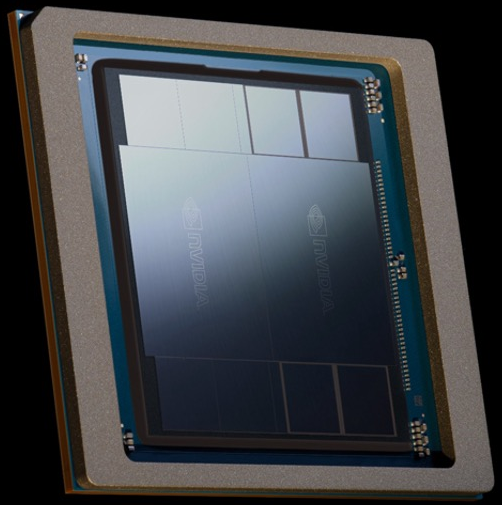
RAS Engine
100% In-System Self-Test

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 combined 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
- Built for :
  - Highest communication density
  - Lowest latency
  - Optimal energy efficiency

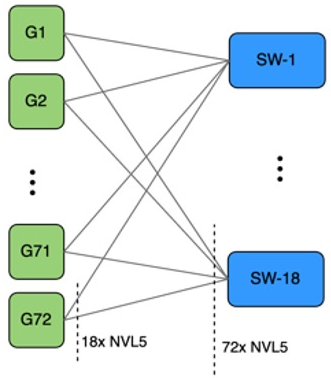


GB200 NVL72

Delivering new unit of compute

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

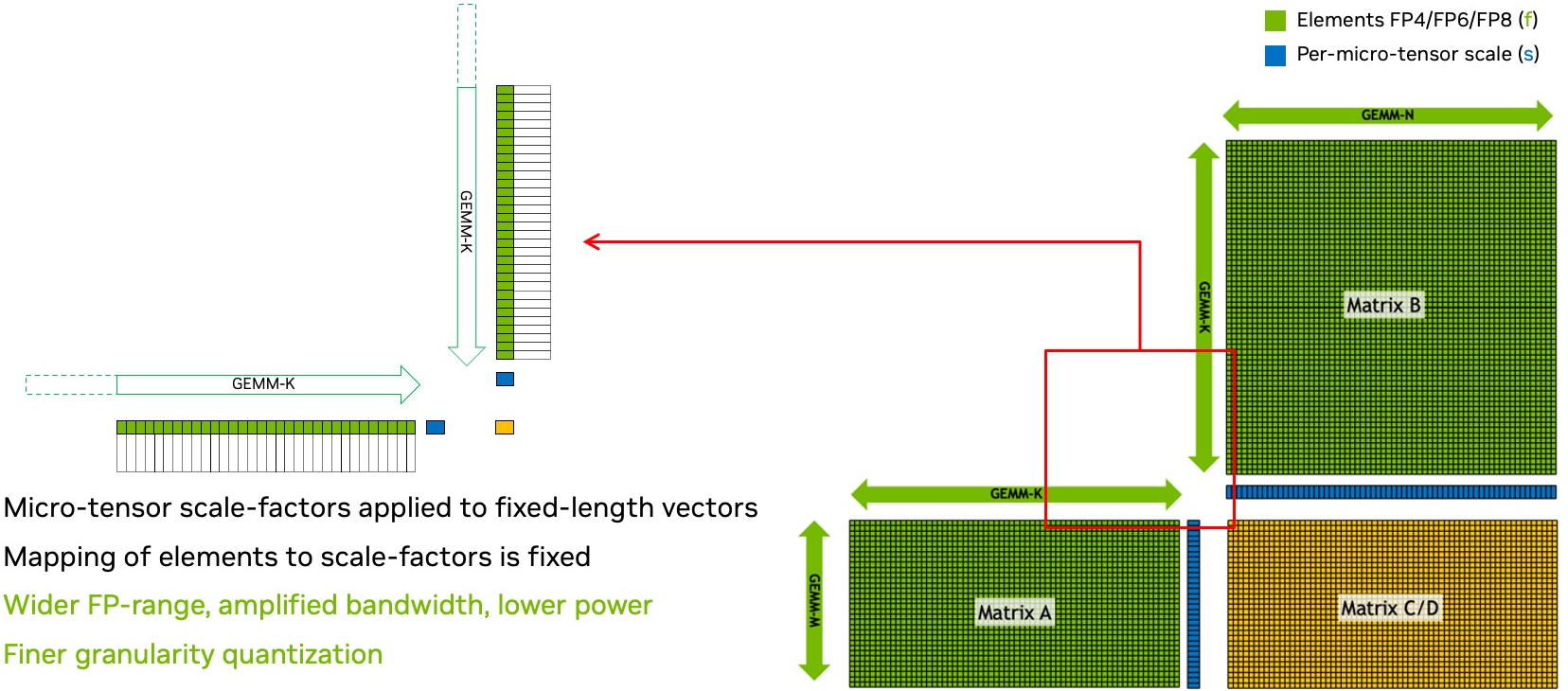
Training	720 PFLOPs
Inference	1,440 PFLOPs
NVL Model Size	27 Trillion params
Multi-Node Bandwidth	130 TB/s
Multi-Node All-Reduce	260 TB/s



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

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

Scaled FP4, FP6 and FP8 as defined by OCP



- 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

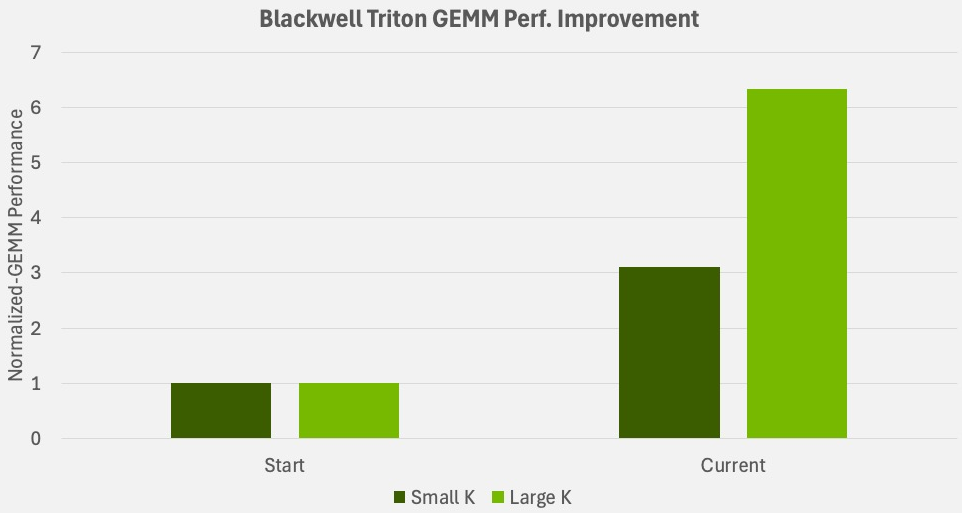
- New micro-scaled formats for FP4, FP6, and FP8
- 4x faster per-clock, per-SM FP4 vs. Hopper FP8
- Blackwell also increases operating frequency and SM count
- Supports dense and sparse variants

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 - OAI Triton Collaboration

Together we've improved performance on Blackwell GEMMs over 6x in just 4 weeks !

... and we are just getting started :)



Blackwell preliminary measured performance data with internal software + Triton Example #9
X Axis : Time ; Y Axis : Normalized Perf.
FP8 GEMM M=N=8192, Small-K = 512; Large-K = 8192;

Deep Collaboration to Deliver World-Class Performance

- Hopper Triton kernels are forward compatible with Blackwell
- TMA support with same language feature as Hopper & graduation out of experimental namespace
- Leverage new Blackwell 5th Generation Tensor Cores and latest architectural features
- Optimizations improve performance on both Hopper and Blackwell
  - Improved software pipelining for latency hiding
  - Enhanced instruction sequence in the main-loop
  - Better epilogue hiding
- Collaboration spans wide range of kernels to deliver peak workload performance

Summary

- Triton with Blackwell support planned for early 2025
- Users can expect efficient support for Micro Scaled Formats, Flash Attention, Mixed Input Precision GEMMs and more
- Triton performance on Blackwell continues to be tuned and we are pushing boundaries of performance
- Together we expect Blackwell will unlock over an order of magnitude performance for Trillion+ parameter models

ACKNOWLEDGEMENTS:
OpenAI : Pawel Szczerbuk, Philippe Tillet

NVIDIA : dePaul Miller, Samantha Hirsch, Wei Liu, Matthew Nicely, Tejash Shah, Meena Raj, Vartika Singh, Joe Delaere, Ashraf Eassa

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