Dissecting the NVIDIA RTX Blackwell Architecture with Microbenchmarks

Aaron Jarmusch
University of Delaware
Newark, US

Nathan Graddon
University of Delaware
Newark, US

Sunita Chandrasekaran
University of Delaware
Newark, US

***Abstract*—We present the first microarchitectural analysis of NVIDIA's Blackwell (GB203) through PTX microbenchmarks, comparing against Hopper (GH100). Our measurements reveal Blackwell's unified INT32/FP32 cores reduce mixed-workload latency by 50%, while FP4 achieves 64% power reduction. However, Hopper maintains 4× higher GEMM throughput, exposing significant software optimization gaps. These findings provide actionable guidelines for GPU architecture selection.**

***Index Terms*—Blackwell, GPU, HPC, Microbenchmark**

I. Introduction

NVIDIA's Hopper (GH100) and Blackwell (GB203) represent contrasting designs: GH100 prioritizes AI training with HBM2e [1], while GB203 targets graphics and inference [2]. Prior work characterized earlier architectures [3]–[6], but Blackwell's unified execution units, FP4/FP6 Tensor Cores, and redesigned caches remain unexplored.

Contributions: (1) Discovery that unified cores eliminate 50% of mixed-workload latency; (2) First characterization of 5th-gen Tensor Cores revealing OMMA/QMMA instructions and 64% power savings with FP4; (3) Identification of L2 cache crossover where unified design outperforms partitioned under high concurrency; (4) Exposure of 4× GEMM gap despite hardware advantages, revealing software maturity issues.

II. Architecture and Methodology

Table I highlights architectural divergence. Blackwell's unified cores dynamically switch between INT32/FP32 each cycle versus GH100's concurrent execution. GH100's partitioned L2 (50 MB, 2 units) optimizes locality; GB203's monolithic design (65 MB) simplifies routing. We measure true latency (serialized) and completion latency (with ILP) using `%clock64` (1-cycle overhead GB203, 2-cycle GH100), averaging 1024 iterations on H100 PCIe and RTX 5080.

TABLE I
Architecture Comparison of Blackwell and Hopper GPUs.

Feature	GH100	GB203
FP32/INT32	Separate (128/64)	Unified (128)
FP64/SM	64	2
Tensor Core	4th gen	5th gen (FP4/FP6)
L1/SM	256 KB	128 KB
L2	50 MB (2 part.)	65 MB (1 part.)
Memory	80 GB HBM2e	16 GB GDDR7

III. Results

A. Unified Execution Units

Table II exposes a key pattern: pure INT32/FP32 achieve identical 4-cycle latency on both GPUs, but mixed workloads show 15.96 cycles (GB203) versus 31.62 (GH100)—50% reduction. Unified cores eliminate pipeline conflicts when both types compete for issue slots, while separate pipelines create scheduling bubbles. FP64 results validate design priorities: GH100's 64 units deliver 8.04 cycles; GB203's 2 units require 63.57 cycles, suggesting FP32 emulation rather than true FP64 execution.

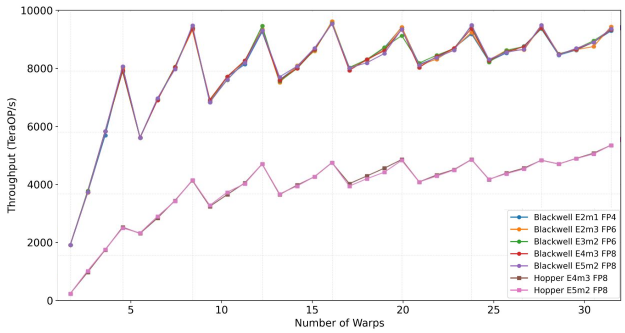

Fig. 1. Tensor throughput: GB203 achieves higher peak (11+ TFLOP/s) and better ILP scaling (max at ILP=6) versus GH100 (ILP=5).

TABLE II
Execution Latency (True/Completion in Cycles)

GPU	INT32	FP32	Mixed	FP64
GB203	4/16.97	4/7.97	15.96/14	63.57/11
GH100	4/16.69	4/7.86	31.62/16	8.04/13

B. Tensor Cores

Our investigation of Blackwell's Tensor Cores uncovered several surprises. SASS disassembly revealed Blackwell's instruction set is partially deployed: **QMMA** handles FP8/FP6, while **OMMA** targets FP4 with block scaling. However, compilers map non-scaled FP4 to QMMA as fallback, indicating maturing software. Figure 1 shows GB203 sustains 11+ TFLOP/s, peaking at ILP=6 versus GH100's ILP=5, with 1.21 vs. 1.66 cycle completion latency.

Figure 1 demonstrates that Blackwell's Tensor Cores extract more parallelism: sustained throughput exceeds 11 TFLOP/s and peaks at ILP=6 (25 active warps), compared to GH100's ILP=5 maximum (29 warps). Completion latency measurements show 1.21 cycles (GB203) versus 1.66 cycles (GH100), indicating more aggressive pipeline optimization. This higher ILP tolerance suggests Blackwell's scheduler can maintain more independent matrix operations in flight simultaneously.

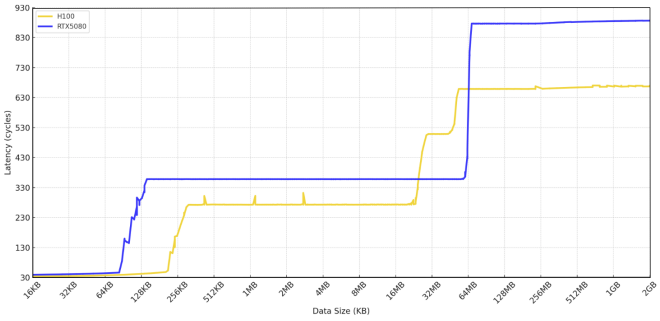

Fig. 2. Memory latency exposes L2 differences: GH100's partitioned design offers lower latency initially but degrades under pressure.

Table III reveals the power efficiency story behind low-precision formats. FP4 operations consume only 16.75W—a dramatic 64% reduction compared to FP8's 46W. FP6 formats occupy a middle ground at 39-46W, offering precision-power tradeoffs. Notably, Hopper maintains flat 55W consumption across all FP8 formats, while Blackwell's power scales with precision. This precision-aware power management explains why Blackwell excels in inference scenarios where dynamic precision mixing is common.

TABLE III
TENSOR CORE POWER CONSUMPTION (WATTS)

GPU	FP4 e2m1	FP6 e2m3	FP6 e3m2	FP8 e4m3	FP8 e5m2
GB203	16.75	39.38	46.72	46.66	46.81
GH100	N/A	N/A	N/A	55.82	55.79

C. Memory Hierarchy

Our pointer-chasing benchmarks (Figure 2) show GH100's partitioned design wins initially: 273-cycle latency versus GB203's 358 cycles. However, when both partitions saturate (31-45 MB), GH100 degrades to 508 cycles.

The crossover occurs under maximum concurrency: our 32-warp benchmark executing 1024 operations per thread measured 128.4k cycles on GB203 versus 128.9k on GH100. This reveals that GB203's unified 65 MB L2, despite higher base latency, delivers superior aggregate bandwidth when partition arbitration would otherwise bottleneck GH100. The monolithic design wins when all SMs simultaneously hammer the cache—exactly the scenario in high-batch inference workloads. Global memory bandwidth follows expected patterns (HBM2e: 15.8 TB/s read vs. GDDR7: 8.2 TB/s), with asymmetric write performance (2.2 vs. 1.6 TB/s) reflecting read-optimized designs.

D. Real-World Performance

Dense FP8 GEMM via cuBLASLt delivered the most surprising finding: GB203 achieves only 0.233 TFLOP/s at 8192^3 versus GH100's 0.887—4× lower (Table IV). GB203 also consumes 80-114W versus GH100's 58-68W, yielding worse performance-per-watt. This performance inversion points to immature kernel heuristics and tile size selection still tuned for Hopper.

cuBLAS kernel heuristics, tile size selection, and memory access patterns are likely still tuned for Hopper's architecture.

TABLE IV
DENSE GEMM THROUGHPUT (TFLOP/S, FP8)

Matrix Size	GH100	GB203
8192×8192×8192	0.887	0.233
2048×4096×8192	0.759	0.217
2048×2048×2048	0.554	0.191
1024×1024×1024	0.239	0.134

The multi-year lag between hardware release and software optimization represents a critical finding: theoretical hardware capabilities don't translate to performance without co-designed software stacks.

Dense FP8 GEMM via cuBLASLt delivered the most surprising finding: GB203 achieves only 0.233 TFLOP/s at 8192^3 versus GH100's 0.887—4× lower (Table IV). GB203 also consumes 80-114W versus GH100's 58-68W, yielding worse performance-per-watt.

TABLE V
TRANSFORMER INFERENCE POWER (WATTS)

Precision	FP32	FP16	FP8
GB203	58.82	47.78	45.14
GH100	57.64	57.64	57.69

IV. RECOMMENDATIONS AND CONCLUSION

Our microbenchmarking reveals architecture and software maturity are inseparable. **Choose Blackwell for:** mixed-precision inference with FP4/FP6 where power matters (<50W), graphics workloads under 60 MB. **Choose Hopper for:** FP64 scientific computing, large-scale training, and production GEMM until software matures.

We discovered unified cores deliver 50% latency reduction for mixed workloads, 5th-gen Tensor Cores achieve 64% power efficiency with FP4, and L2 design determines concurrent load behavior. However, the 4× GEMM gap exposes that hardware capabilities remain theoretical without mature software—a critical reminder as architectures evolve faster than optimization cycles.

ACKNOWLEDGMENT

We thank NVIDIA's Nikhil Jain. This work used University of Oregon resources and was supported by U.S. DOE S4PST DE-FOA-0003177.

REFERENCES

[1] NVIDIA Corporation, *NVIDIA H100 Tensor Core GPU Architecture*, NVIDIA, Mar. 2022. [Online]. Available: https://resources.nvidia.com/en-us-data-center-overview/gtc22-whitepaper-hopper

[2] ——, *NVIDIA RTX BLACKWELL GPU ARCHITECTURE*, NVIDIA, 2025. [Online]. Available: https://images.nvidia.com/aem-dam/Solutions/geforce/blackwell/nvidia-rtx-blackwell-gpu-architecture.pdf

[3] H. Wong, M.-M. Papadopoulou, M. Sadooghi-Alvandi, and A. Moshovos, "Demystifying gpu microarchitecture through microbenchmarking," in *2010 ISPASS*, 2010, pp. 235–246.

[4] Z. Jia, M. Maggioni, B. Staiger, and D. P. Scarpazza, "Dissecting the NVIDIA volta GPU architecture via microbenchmarking," *CoRR*, vol. 1804.06826, 2018. [Online]. Available: http://arxiv.org/abs/1804.06826

[5] Z. Jia, M. Maggioni, J. Smith, and D. P. Scarpazza, "Dissecting the nvidia turing T4 GPU via microbenchmarking," *CoRR*, vol. 1903.07486, 2019. [Online]. Available: http://arxiv.org/abs/1903.07486

[6] W. Luo, R. Fan, Z. Li, D. Du, H. Liu, Q. Wang, and X. Chu, "Dissecting the nvidia hopper architecture through microbenchmarking and multiple level analysis," 2025. [Online]. Available: https://arxiv.org/abs/2501.12084