Analyzing the Impact of Architectural Design Decisions on Performance Across Generations of NVIDIA GPUs

Tavin Mistry*, Tobias Salamon*, Matthew Tindale*, Ian Karlin

*These authors contributed equal work

Introduction

Motivation: The high-performance computing (HPC) industry is moving past an era in which each new generation of devices provided uniform performance improvements across all workloads [1].

Goal: Evaluate how NVIDIA GPU architectural changes affect real-world application performance and energy efficiency.

- **3 GPU Generations:** Volta, Ampere, Hopper
- **8 Workloads:** AI inference, AI training, scientific computing, microbenchmarks
- **Measured:** Execution time, power, energy, GPU hardware counters
- **Analyzed:** which architectural improvements provide the greatest benefit and what workload characteristics determine their effectiveness [2].

Key Takeaway: Relative to transistor count growth:

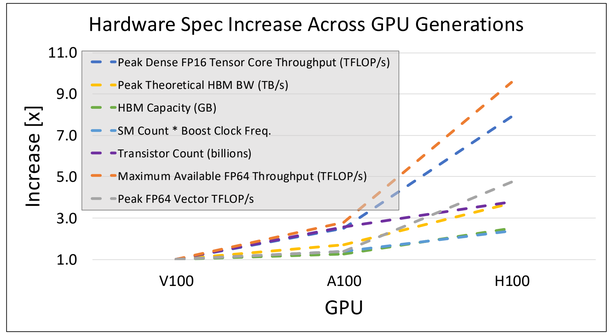
- **Significant** performance gains among Tensor Core-bound workloads
- **Expected** gains in memory-bound workloads
- **Below-expected** gains in vector throughput-bound workloads.

Application	Workload Category	Bottleneck
HPCG	Sparse Linear Solver Benchmark	HBM bandwidth
HPL	Dense linear algebra	FP64 Tensor throughput
LAMMPS	Molecular dynamics	SM + FP64 vector throughput
Megatron-LM Llama1B Training	Transformer training	Tensor Core throughput
MILC	Gauge field generation	HBM bandwidth + SM vector throughput
ResNet50 Training	CNN training	Tensor Core throughput + HBM bandwidth
STREAM	Memory bandwidth benchmark	HBM bandwidth
Inference – Llama7B	LLM inference	Tensor Core throughput + HBM bandwidth

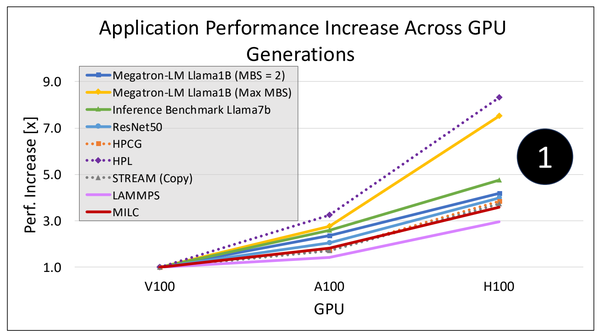
Testing Environment

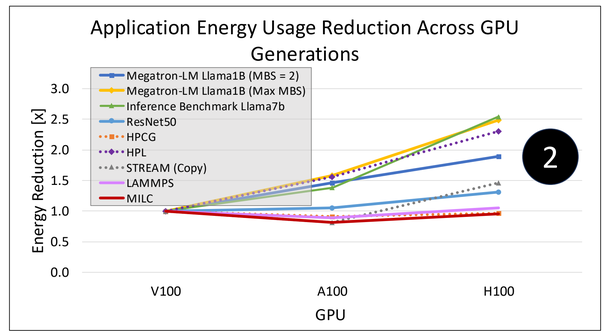
Experiments used a single GPU per run to isolate GPU performance and eliminate interconnect effects. Testing was conducted on clusters operated by the Digital Research Alliance of Canada (DRAC) and the Centre for Advanced Computing (CAC).

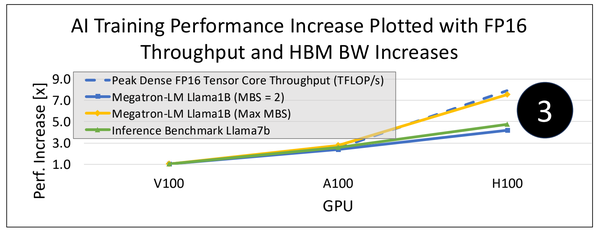
GPU Model	Form Factor	Memory Size	Cluster(s) Used
V100	PCIe	32GB	CAC – Frontenac
A100	SXM4	40GB	CAC – Frontenac, DRAC – Narval
H100	SXM5	80GB	DRAC – Rorqual, Nibi, Fir

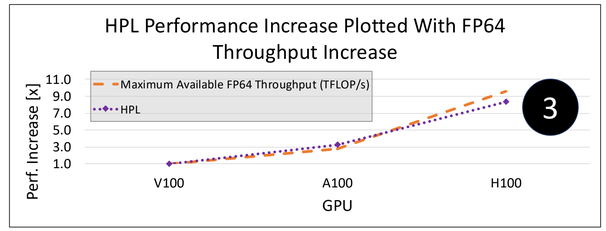


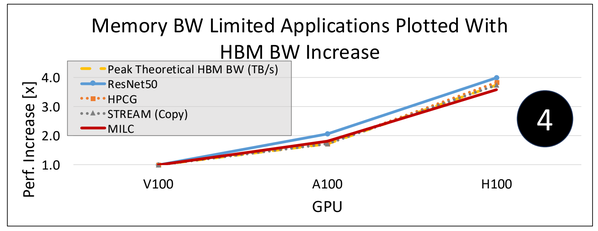
Results

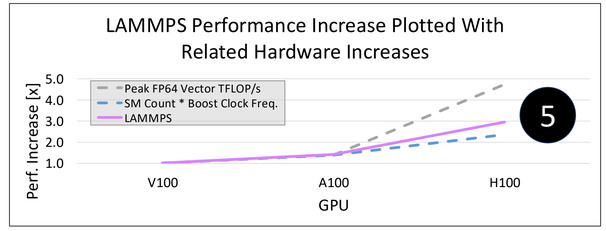












Analysis

1. **Workload dependence:** H100 speedups range from approximately 3-8x, showing that performance gains depend strongly on the workload.
2. **Energy efficiency:** Tensor Core heavy applications improve by about 2–2.5x, while bandwidth-bound workloads see smaller reductions because data-movement efficiency scales more slowly than compute efficiency [4].
3. **Tensor Core scaling:** Greater investment in Tensor Cores drives the largest gains; HPL tracks FP64 Tensor Core throughput, while Llama workloads benefit from increased FP16 tensor throughput.
4. **Memory bandwidth scaling:** STREAM, HPCG, MILC, and ResNet50 scale primarily with HBM bandwidth. ResNet50 also benefits from Tensor Cores, placing it between purely bandwidth-bound and Tensor Core-bound workloads.
5. **LAMMPS scaling**: LAMMPS improves the least, scaling with a combination of FP64 vector throughput and SM count * clock frequency, rather than Tensor Core throughput.

Conclusions

- **Memory-bound scaling:** From Volta to Hopper, increases in HBM BW match transistor gain, so memory bound apps' scale with transistor gain.
- **Tensor Core advantage:** Tensor Core-bound workloads have speedups roughly double the transistor count increase; Llama1B by 1.95x, Llama7B by 1.78x, HPL 2.19x reflecting increased investments in tensor throughput.
- **LAMMPS limitation:** LAMMPS is constrained by SM improvements and to a lesser extent FP64 vector scaling, leading to smaller gains than overall hardware technology improvements.

Future Work

- Add applications to workload suite.
- Collect data on **Pascal and Blackwell GPUs.**
- Analyze changes in **die-area investment** over time, including the fraction devoted to workload-critical hardware such as Tensor Cores [5].
- Run **frequency-capped experiments** to compare energy efficiency at the best performance-per-watt operating point.

References

[1] N. C. Thompson and S. Spanuth, "The decline of computers as a general purpose technology," Commun. ACM, vol. 64, p. 64–72, Feb. 2021.
[2] J. Domke, E. Vatai, A. Drozd, P. Chen, Y. Oyama, L. Zhang, S. Salaria, D. Mukunoki, A. Podobas, M. Wahib, and S. Matsuoka, "Matrix engines for high performance computing: A paragon of performance or grasping at straws?," in 2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 1056–1065, 2021.
[3] J. Tu, I. Karlin, J. Camier, V. Dobrev, T. Kolev, S. Henneking, and O. Ghattas, "Accelerating high-order finite element simulations at extreme scale with fp64 tensor cores," in ISC High Performance 2026 Research Paper Proceedings (41st International Conference), pp. 1–12, 2026.
[4] J. Shalf, "The future of computing beyond Moore's Law," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 378, no. 2166, p. 20190061, 2020.
[5] NVIDIA Corporation, NVIDIA Blackwell, Oct. 2025. Datasheet, document no. 4204213. Accessed: Aug. 5, 2026.