Toward HEPScore4GPU: Scaling and Power of Real HEP GPU Applications on the A100 and H100

Smita Darmora*, Michael E. Papka**

*University of Illinois Chicago, **Argonne National Laboratory

Throughput is only half the story :
scaling and power of real physics workloads on the A100 and H100

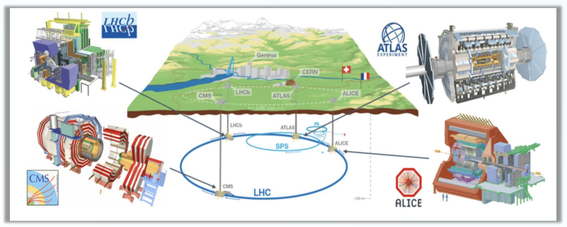
Abstract

HEPiX Benchmarking WG — the international effort that benchmarks computing for **CERN's LHC experiments** — is building HEPScore4GPU, a GPU benchmark for High-Energy Physics (HEP), built from real applications.

- How efficiently GPUs run HEP today — guiding HEPScore4GPU and the application developers.

Why GPU benchmarking matters

- **Scale & demand** — >1M cores, ~170 sites → High-Luminosity LHC pushes far higher → shift to GPUs
- **Fair comparison** — a shared benchmark from real applications → compare GPUs, plan procurement
- **Efficiency, not just speed** — measure power and efficiency, not only throughput



Large Hadron Collider: ~17-mile ring spanning France and Switzerland.

Benchmark Applications

Pepper — Monte Carlo generator: simulates particle collisions *(GPU)*
ATLAS FullSim — detector simulation: how particles cross the detector *(CPU + GPU)*
CMS FlowSim — AI-based fast simulation: reproduces the detector response *(GPU)*

Methodology & Experimental Setup

- **Benchmark suite:** HEPScore / HEP Benchmark Suite (CERN Apptainer containers)
- **Hardware:** NVIDIA A100 (PCIe, 80 GB, 300 W) & H100 (SXM, 700 W) — bare-metal, Chameleon testbed
- **Configuration:** 1 / 2 / 4 GPUs; identical conditions per run
- **Runs**: FlowSim & Pepper — 15 each; ATLAS — reduced event count
- **Measured:** throughput (scaling 1 → 4 GPUs) · GPU power (1 Hz, mean)

Results — Multi-GPU Scaling

- FlowSim & Pepper scale near-linearly — 1.00 / 1.99 / 3.98× as GPUs go 1 → 2 → 4
- Run-to-run spread < 0.5% → stable, reproducible
- ATLAS FullSim scales sublinearly — it's CPU-fed

GPU-resident workloads add up cleanly; the hybrid one is limited by the CPU feeding it

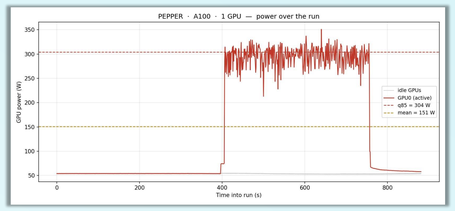
Results — H100 vs A100

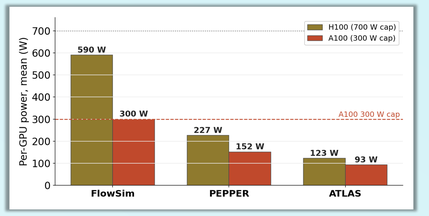
- **H100 draws more power** (700 W cap vs A100's 300 W) — but this **does not invalidate the comparison**
- **Raw performance** = the primary, valid measure of each GPU as deployed
- **Performance per watt** *extends* it — adds the energy-efficiency view (it doesn't "rescue" the comparison)

Metric (single GPU)	A100	H100	H100/A100
FlowSim score (vs reference)	1.09	2.83	**2.6×**
FlowSim perf. per watt (score/kW)	3.6	4.8	**1.3×**
Pepper throughput (10^6 ev/s)	0.83	1.24	**1.5×**

Results — Power

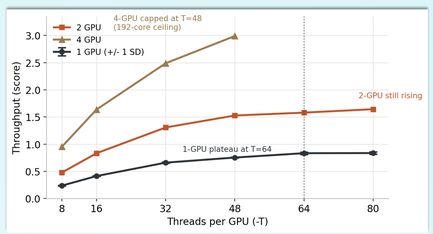
- Power is workload-driven — varies sharply by application
- **FlowSim** — near the caps: A100 ~300 W, H100 ~590 W
- **ATLAS** — light (~100 W): waits on the CPU
- **Pepper** — mean ~152 W: idle start-up, then burst at the 300 W cap
- Report the mean (whole-run energy), not just peak





Results — ATLAS Threads

- CPU-fed — throughput rises with CPU threads
- 1-GPU plateau at T = 64 (GPU fully fed)
- 4 GPUs: cores run out — T=48 (H100), T=40 (A100) → no plateau
- Lesson: needs enough CPU cores per GPU to scale



Discussion & Conclusion

- GPU-bound → near-linear scaling; CPU-fed → need enough CPU cores per GPU
- Newer GPUs are faster — but at scale, performance per watt is the metric
- Guides GPU benchmarking & provisioning for large-scale science → HEPScore4GPU
- Next: more workloads & GPUs → a production HEPScore4GPU (performance + efficiency)

Acknowledgements & References: Carried out with the **HEPiX Benchmarking Working Group**, in particular CERN's HEPScore team. Thanks to **D. Giordano, R. Hofsaess, and N. D. Szczepanek**, and the developers of CMS FlowSim, Pepper, and ATLAS FullSim/AdePT. Results were obtained using the **Chameleon testbed, supported by the National Science Foundation.**