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

Smita Darmora[1] Michael E. Papka[1,2]
On behalf of the HEPiX Benchmarking Working Group
[1]University of Illinois Chicago, USA [2]Argonne National Laboratory, USA
sdarm@uic.edu

***Abstract*—The HEPiX Benchmarking Working Group is developing HEPScore4GPU, a GPU benchmark for High-Energy Physics (HEP) built from real HEP applications for the High Lumi-LHC program. A key question is how efficiently GPUs can be used for HEP today. We take three candidate applications under development, a Monte Carlo Generator, a detector full simulation and an AI-based end-to-end simulation application, and run them on NVIDIA A100 and H100 GPUs, measuring how throughput scales from one to four GPUs and how much power they use. In this contribution we will describe the measurement process and the understanding of the different behaviors, with the objective of providing input to the application developers to further improve the software and exploit the maximum power of the GPUs.**

***Index Terms*—HEP computing, GPU benchmarking, HEPScore, HEPScore4GPU, performance per watt, multi-GPU scaling, WLCG.**

I. Motivation

At CERN, the Large Hadron Collider (LHC) collides protons 40 million times a second to probe the fundamental particles and forces that make up the universe. Turning those collisions into physics is one of the largest sustained computing efforts in science. The Worldwide LHC Computing Grid is the backbone of CERN's success, as it provides more than a million CPU cores across roughly 170 sites all around the world. With the next upgrade of the accelerator, namely the High-Luminosity LHC, approaching, the demand for compute will drastically increase as the upgrade mainly aims for even more data to be collected. To keep up, the LHC experiments are more and more shifting towards heterogeneous compute architectures, such as GPUs, to improve data processing and cope with the expected requirements. This raises a key question: how efficiently GPUs can be used for High-Energy Physics (HEP) already today. To investigate this question, the HEPiX Benchmarking Working Group is preparing a first version of HEPScore4GPU, based on real HEP GPU applications. We study three of these applications, each of a different type: Pepper, a Monte Carlo generator that simulates the particle collisions; ATLAS FullSim with AdePT, a full detector simulation that models how the resulting particles pass through the ATLAS detector; and CMS FlowSim, an AI-based fast simulation that reproduces the detector response with a machine-learning model. Using these preliminary applications, we measured how each one scales from one to four GPUs and how much power it uses, to understand how efficiently they run on current GPUs and to give useful feedback to the developers.

II. Applications and Method

We ran each application on an NVIDIA A100 (PCIe, 80 GB, 300 W limit) and an H100 (SXM, 700 W limit) on bare-metal Chameleon nodes, measuring throughput at one, two, and four GPUs and logging GPU power at 1 Hz, reported as the mean. FlowSim and Pepper were run 15 times per configuration; the ATLAS FullSim runs used a reduced event count and are preliminary.

III. Results

FlowSim and Pepper scale almost linearly, giving close to two and four times the single-GPU throughput (1.00, 1.99, 3.98), with run-to-run spread under 0.5%. ATLAS FullSim, run with reduced statistics, scales sublinearly, as expected for a hybrid workload that depends on the CPU feeding the GPU.

At a single GPU the H100 clearly outperforms the A100. FlowSim, whose throughput is given relative to the HEPScore reference, scores 1.09 on the A100 and 2.83 on the H100 (about 2.6×); for Pepper it is about 1.5×, so the gain is application-dependent. Performance per watt is similar: FlowSim gives about 3.6 (A100) and 4.8 (H100) score per kilowatt (Table I).

The power profiles also differ. On the A100, FlowSim runs steadily at the 300 W limit, while Pepper reaches the limit only during compute, averaging about 152 W because part of each run is spent initializing with the GPU idle. On the H100, FlowSim draws about 590 W. ATLAS FullSim stays low, around 100 W.

TABLE I. Single-GPU performance and efficiency.

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×**

IV. Discussion

For HEPScore4GPU and for sites adopting GPUs, it matters that different HEP applications use a GPU in different ways. ATLAS FullSim is a hybrid CPU–GPU workload by design, so its throughput depends on the CPU that feeds the GPU and the GPU is not fully saturated — an expected property of this type of application. FlowSim and Pepper run almost entirely on the GPU: FlowSim sits at the A100 power cap, while Pepper reaches it during compute but averages less, owing to an idle initialization phase that points to where its GPU use could improve. The H100 is never at its power limit, but this does not hinder a direct comparison, since the raw performance already

provides a good measure; performance per watt adds energy efficiency. Pepper and FlowSim are well suited to compare raw performance and efficiency across GPUs, while ATLAS FullSim, being CPU-dependent, better reflects full-server performance. ATLAS FullSim is limited by the CPU, so it needs enough CPU cores per GPU to scale well. And reporting performance per watt alongside throughput helps sites judge efficiency, not just speed.

Acknowledgment

This work was carried out with the HEPiX Benchmarking Working Group, and in particular CERN's HEPScore team. The authors thank D. Giordano, R. Hofsaess, and N. D. Szczepanek, and the developers of the investigated workloads (CMS FlowSim, Pepper, and ATLAS FullSim/AdePT). Results presented in this paper were obtained using the Chameleon testbed supported by the National Science Foundation.