
CERN
openlab

Benchmarking Heterogeneous Architectures with HEPScore

Openlab Technical Workshop 2024

David Southwick

Benchmarking and Heterogeneous hardware

Computing at CERN has been dominated by x86 CPU operations for most of its history, but this environment is rapidly changing:

- Specialized compute devices (accelerators)
- Specialized architectures (registers, memory, extensions)

Necessitating a paradigm shift in software models to leverage **aggressive offloading to the most efficient device**

To understand this complex environment, new tooling was required to provide the flexibility to account many diverse devices in this increasingly fragmented and rapidly changing field

CERN Prévessin Data Center

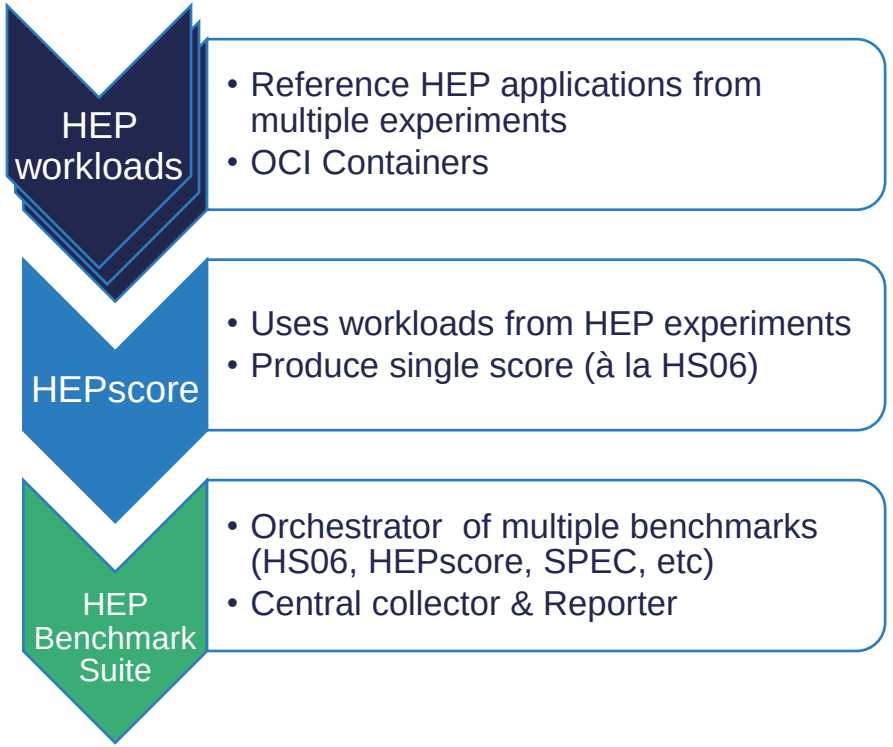
HEP Benchmarking

HEPscore: The next generation of benchmarking for the WLCG

- Intentional “single score” result
- Composable/updateable containerized workload array
- Extensible to any hardware that can run python

HEPscore ratified in April 2023 by the WLCG HEPscore Deployment Task Force as a replacement for HEPspec06

- Now progressively replacing HS06



Instructions & public results: https://w3.hepix.org/benchmarking.html

Heterogeneous execution… and more

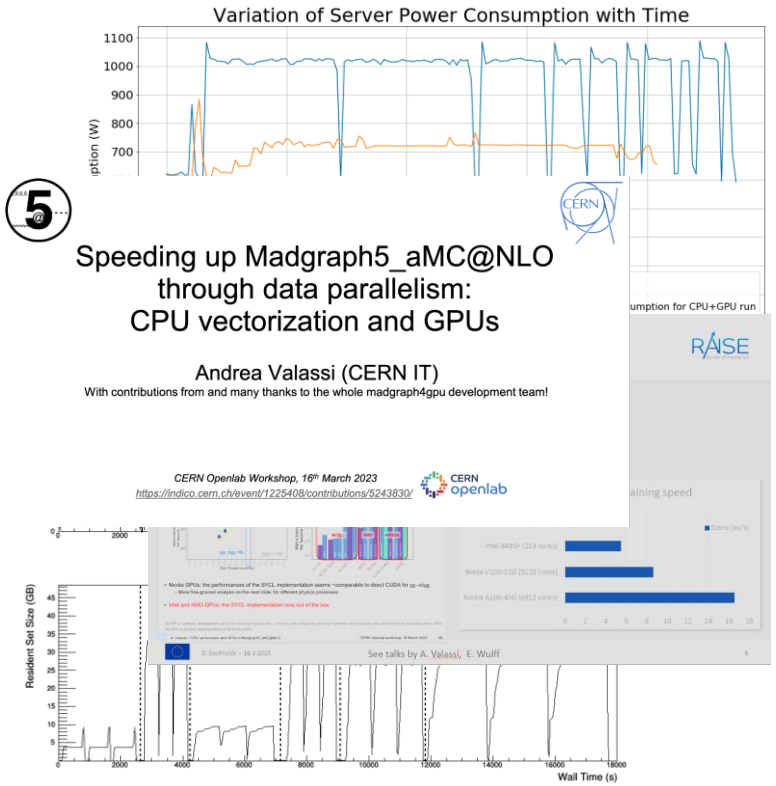
“HEPscore23” comprises a set of 7 production workloads:

- Available for x86_64 and **aarch64**

Additional (bonus!) workloads leveraging compute accelerators:

- MadGraph5 Event Generator (CPU vectorization, GPUs)
- CMS HLT (GPUs)
- MLPF (CPU vectorization, GPUs)

HEP Benchmark Suite includes plugins to measure energy consumption, load, CPU frequency; enabling heterogenous efficiency studies

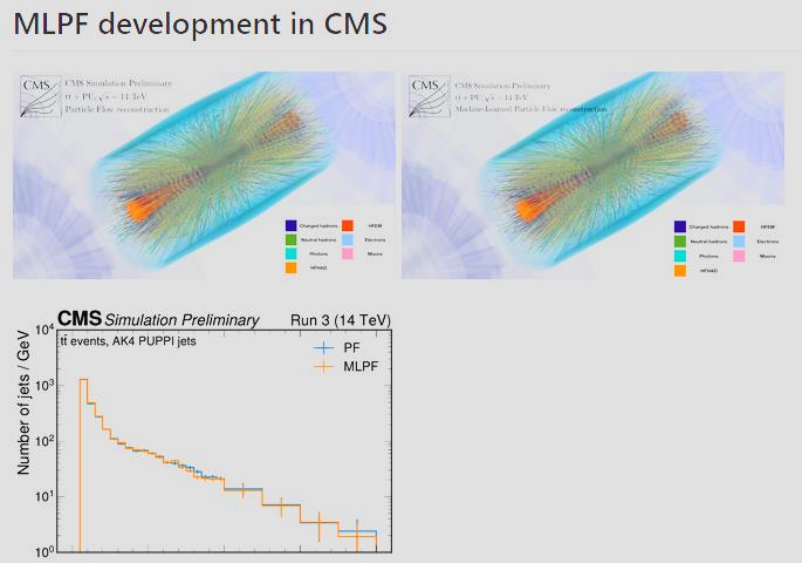


Machine Learned Particle Flow (MLPF)

An early example of a new generation of ML workloads being developed across all experiments

- Containerized and available as an optional benchmark in HEPscore
- Executes on any hardware supporting ML frameworks
- Aggressively leverages accelerator offloading

ML workloads such as this will take time before they replace their traditional deterministic (x86) counterparts, but this process has already begun in many experiments

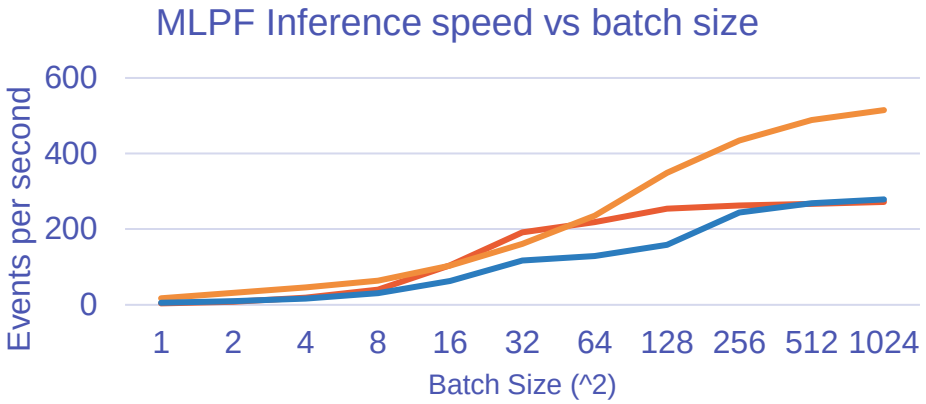


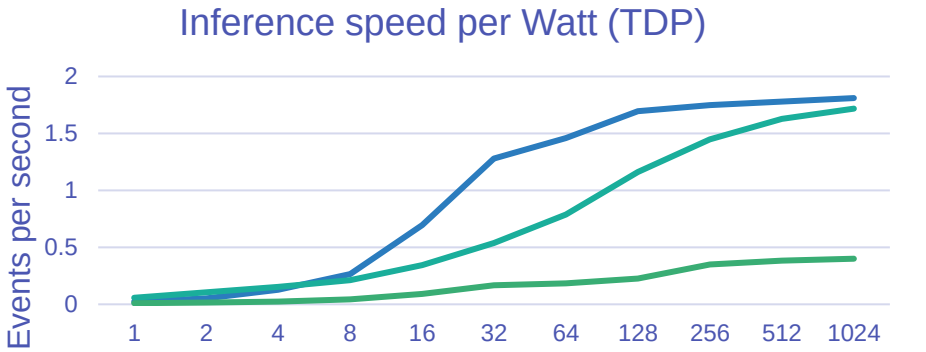
See talk by E. Wulff later today!

Recent heterogeneous accelerators

Understanding production ML performance (*inference*) critical as ML development matures in experiments.

- Leverage heterogenous testbed to study performance, efficiency (next talk!)
- Study modern GPU and CPU vector capabilities
- Inform procurement, pledging, accounting
- …Still complex & difficult!





Selection of current-gen accelerator performance

Outlook

Many developments in benchmarking to cover heterogeneous architectures:

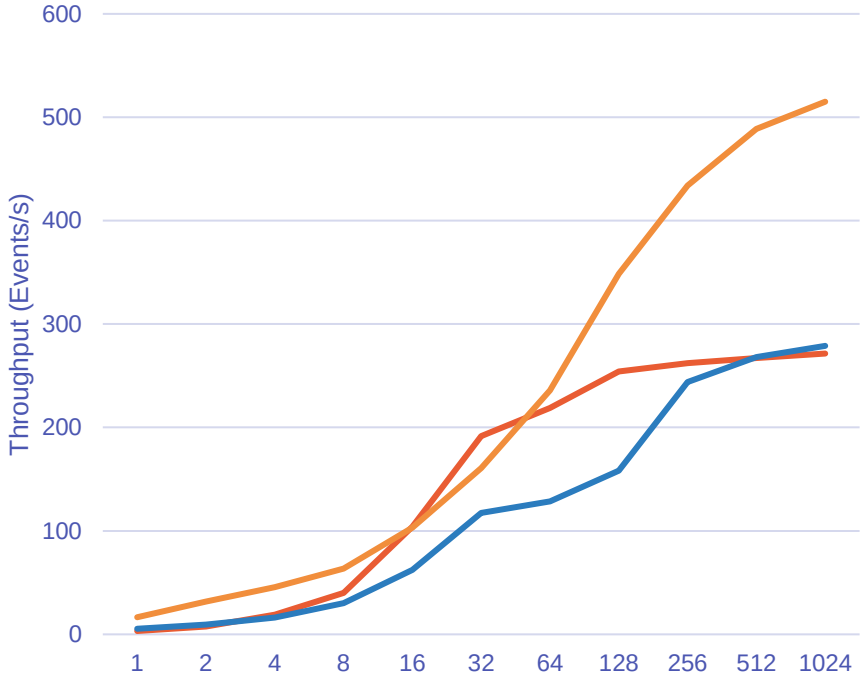
- Workloads supporting multiple compile targets
- Leveraging multiple types of compute accelerators and frameworks
- Major infrastructure tooling effort to support HEPscore rollout
- Focus on energy and compute efficiency across many devices

Preparing for upcoming heterogeneous workloads from experiments!


CERN
openlab

Energy Efficiency

Inference speed vs batch size (larger is better)



Inference speed per Watt (TDP)