RAISE Center of Excellence

HEP Benchmarking on HPC

David Southwick, Maria Girone, Eric Wulff, Eduard Cuba

In collaboration with HEPiX Benchmarking working group

Background

Efficient exploitation of HPC resources presents unique challenges. Scaling workload execution adds layers of complexity not captured in traditional compute environments

- Permissions:
  - Environment (containerization helps)
  - Monitoring (I/O, network, performance bottlenecks, etc)
- Connectivity:
  - isolated worker nodes
  - site connectivity (big data ingress/egress)

To successfully exploit HPC resources we need to understand efficiency both in terms of compute and data access.

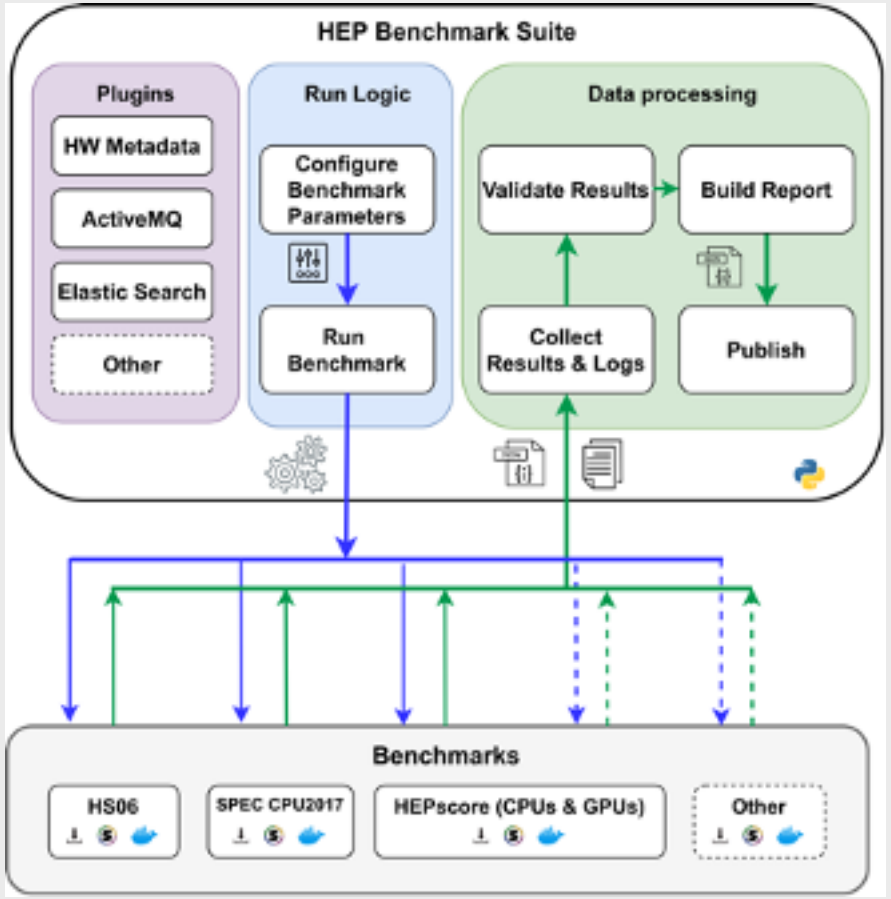
Context: Benchmarking at CERN

HEP Benchmark Suite: A benchmark orchestrator & reporting tool.

Executes an array of user-defined benchmarks & metadata collection

Refactored for HPC:

- Minimal dependencies (Python3 + OCI container)
- Automated result reporting (AMQ/Elastic)
- Scheduler agnostic, unprivileged
- Easily extendable to other sciences!

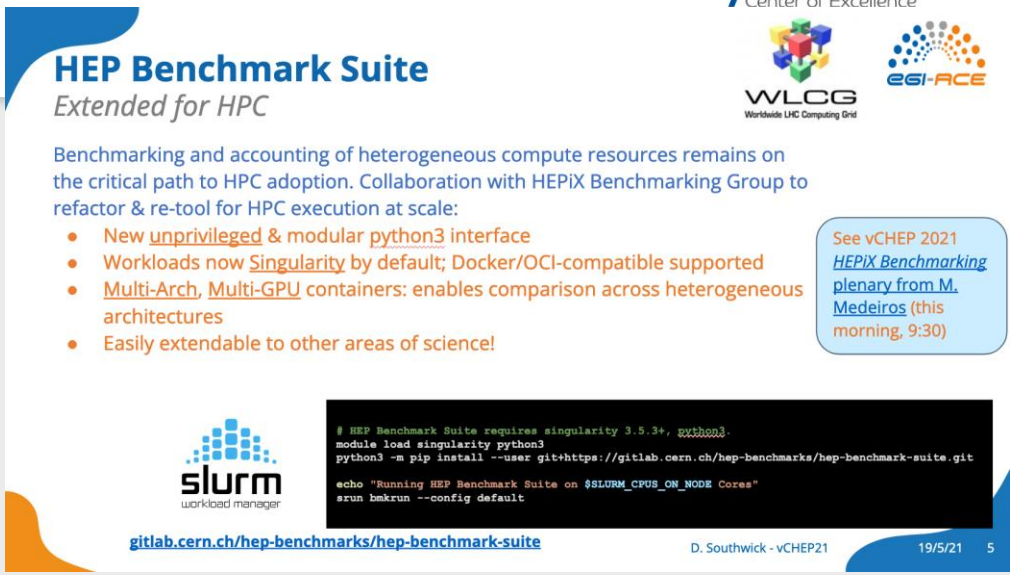


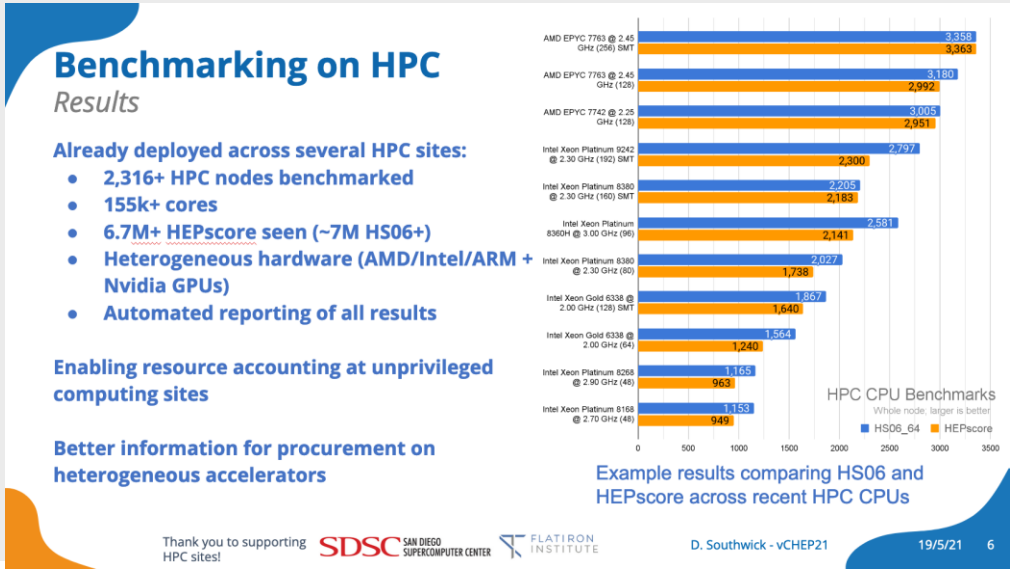
https://gitlab.cern.ch/hep-benchmarks/hep-benchmark-suite

Successes at HPC centers

HEPscore (executed by the HEP-Benchmark-Suite) has already been used for large scale deployments and studies at HPC sites:

- Initial experiences from vCHEP'21
- 200,000-core campaign with Run-2 production WLs
- Scale studies of new/upcoming AMD cpus





See vCHEP'21 presentation for full details

What's new?

First look of run3 workloads - many with heterogenous architectures!

- First ARM, IBM POWER, GPU development workloads
- GPU *vs* CPU *vs* GPU+CPU benchmarking studies
- Heterogenous partition studies (ARM+GPU, POWER+GPU, etc)
- ML / AI workload development (MPI scaled to ~200 GPUs)

Quality-of-Life updates:

- Batch uploading (post-run: supports "secure" worker nodes)
- GPU / accelerator meta-data inclusion
- CVMFS-attached benchmarking campaigns

Thank you to all partner HPC sites enabling this work!

Typical HPC single node resources:

2x AMD EPYC:	256 threads
4x Nvidia V100:	20,480 cuda cores
4x Nvidia A100:	27,648 cuda cores
4x Nvidia H100:	*67,584 cuda cores**

* https://www.nvidia.com/en-us/data-center

Conclusions

- HEP already benchmarking on HPC in heterogeneous partitions
- Automated reporting enables analysis for developers & operators
- First ML/AI workloads as repeatable benchmark
  - Containerized in similar manner to traditional CPU benchmarks
  - Support (multi) GPU accelerators for training/tuning
  - Examine events/second processed (same metric as HEPiX CPU jobs)

- First benchmarks of HPC “support” services
  - Characterize I/O requirements for generalized workflows
  - Development work to increase granularity of characterization
  - Automate profile generation (as much as reasonable)

drive. enable. innovate.

SDSC SAN DIEGO SUPERCOMPUTER CENTER

FLATIRON INSTITUTE

JÜLICH Forschungszentrum

RAISE Center of Excellence

The CoE RAISE project has received funding from
the European Union's Horizon 2020 –
Research and Innovation Framework Programme
H2020-INFRAEDI-2019-1 under grant agreement no. 951733

Follow us:

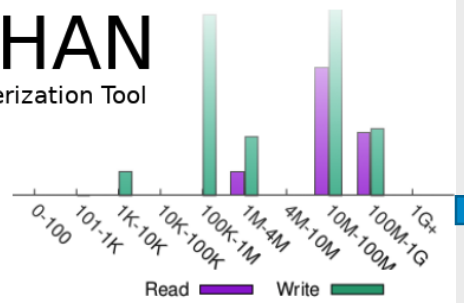
Workload I/O benchmark

Problem: Unclear how many data-driven workloads a given site may support without bottleneck shared resources

- Development of a *workload I/O benchmark*
- tune to the **I/O patterns of real workloads** to better inform reasonable scaling capabilities at a given HPC site
- More representative than sequential throughput metrics
- Uncover **I/O bottlenecks** (excessive file opens, read patterns, cache issues)
- Under development

HPC workload

DARSHAN
HPC I/O Characterization Tool

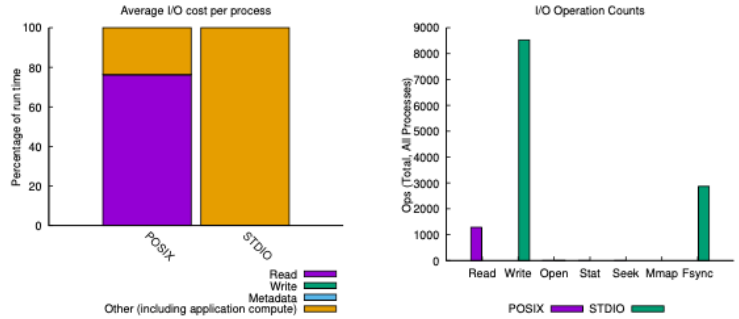


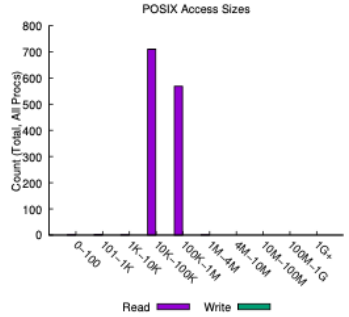
IoR
HPC benchmarks

https://github.com/hpc/ior
https://github.com/darshan-hpc/darshan

jobid: 2190289	uid: 1005	nprocs: 1	runtime: 6 seconds

I/O performance *estimate* (at the POSIX layer): transferred 172.4 MiB at 37.65 MiB/s
I/O performance *estimate* (at the STDIO layer): transferred 0.1 MiB at 63.62 MiB/s





Most Common Access Sizes (POSIX or MPI-IO)

	access size	count
POSIX	49284	141
	20873	3
	204628	3
	204758	2

File Count Summary (estimated by POSIX I/O access offsets)

type	number of files	avg. size	max size
total opened	2	950M	1.9G
read-only files	1	1.9G	1.9G
write-only files	1	69K	69K
read/write files	0	0	0
created files	1	69K	69K