RAISE Center of Excellence

Benchmarking on HPC

David Southwick, Maria Girone, Eric Wulff, Eduard Cuba

Luca Atzori , Joaquim Santos, Krzysztof Mastyna

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.

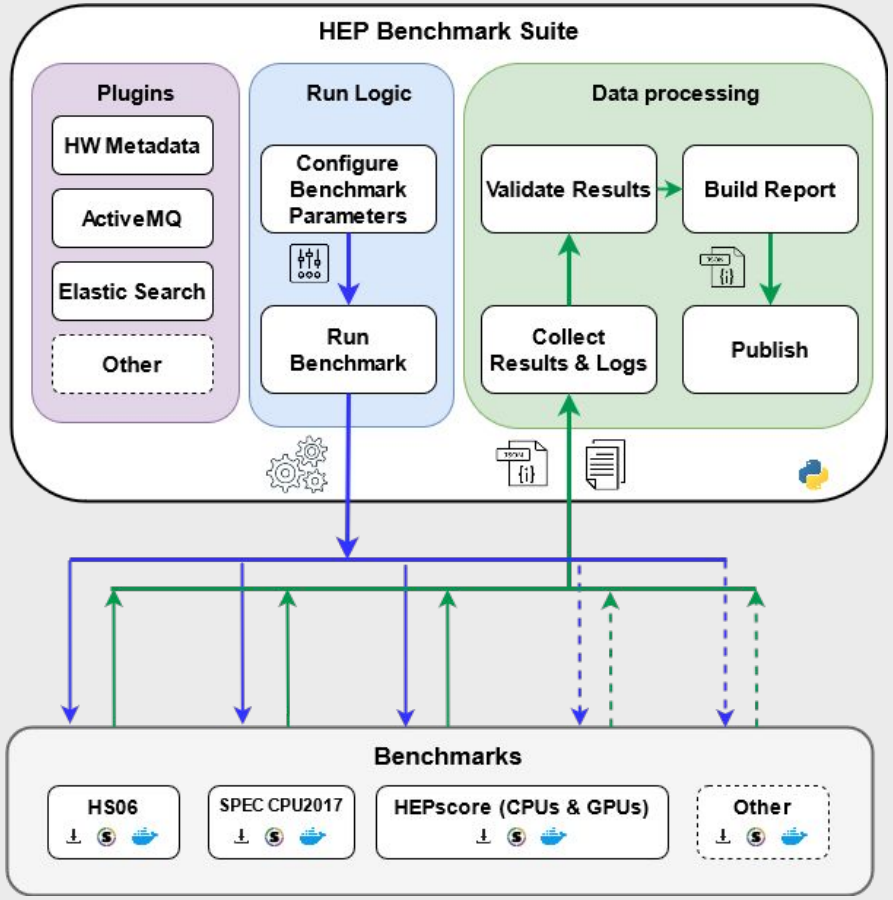
Context: Benchmarking in WLCG

HEP Benchmark Suite: A benchmark orchestrator & reporting tool. Benchmarking activity is driven by the *HEPiX Benchmarking WG*, whose role is to propose a new CPU/GPU benchmarks.

Executes an array of user-defined benchmarks & metadata collection

Support for HPC in v2:

- ➢ 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

Recent publication: https://doi.org/10.1007/s41781-021-00074-y

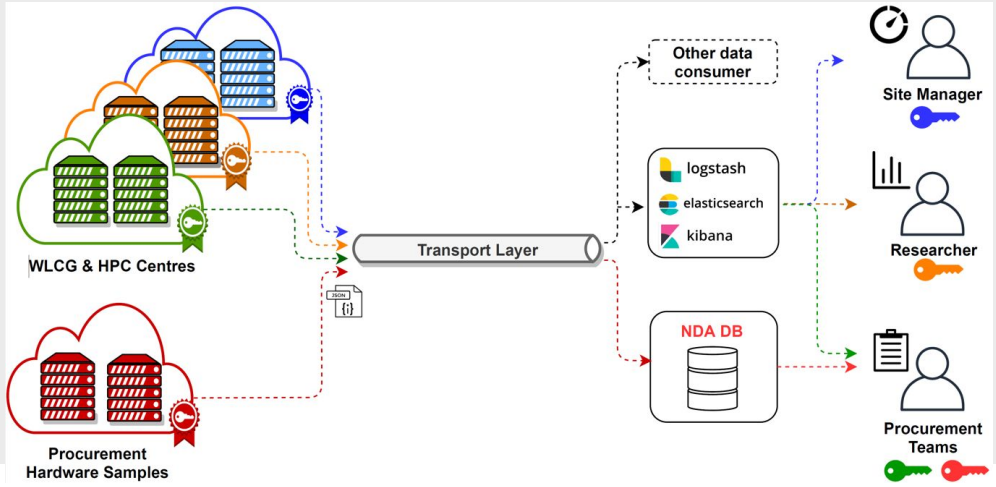
Aggregation & Analysis

RAISE Center of Excellence

Short benchmarking campaign ~120,000 cores

HPC site benchmarks with HEP Benchmark suite

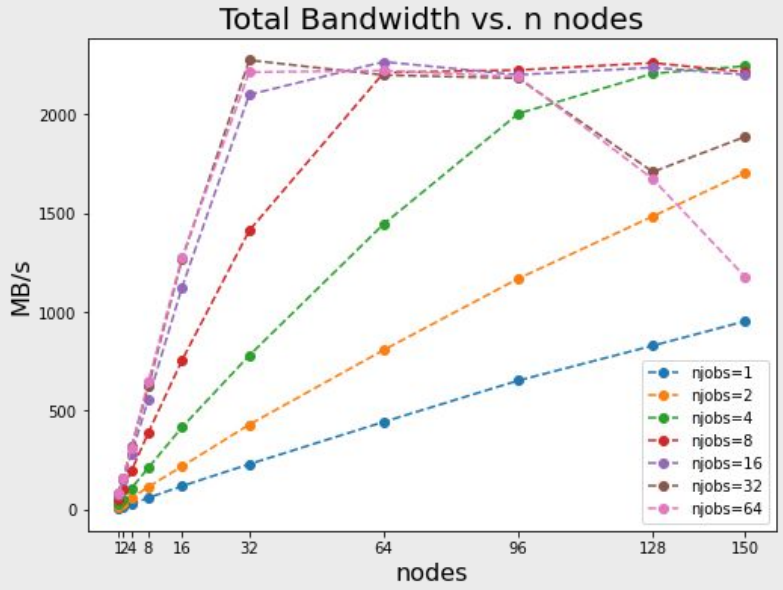
- Automated reporting/collection enables comparison & trend analysis
- Supports collection/reporting for compute nodes without WAN
- Performance fault identification & more



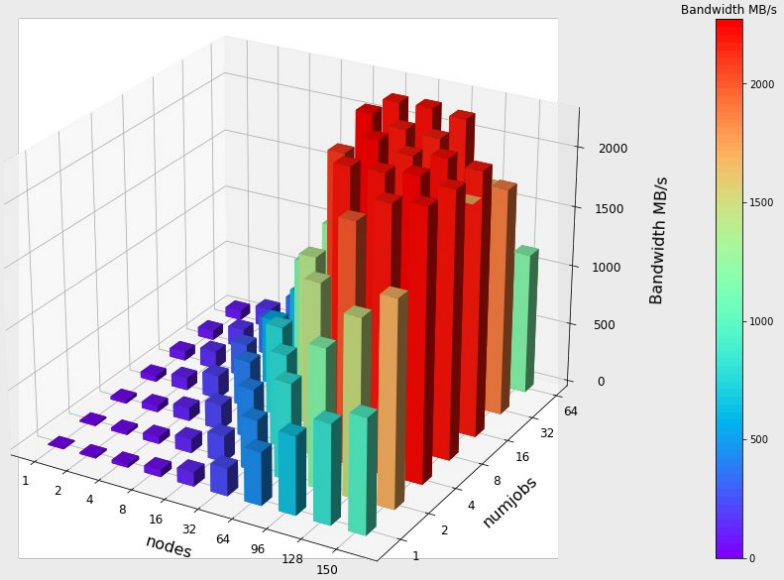
Example report view from HEPiX Benchmarking WG

Scaling and bottlenecks in Big Data

- Data-driven workloads demand performant storage and connectivity (which are shared!)
- Bottlenecks here significantly throttle job performance
- Capacity, capability, and monitoring not typically advertised by HPC sites



Peak	Bandwidth
16 node	2.2 GB/s

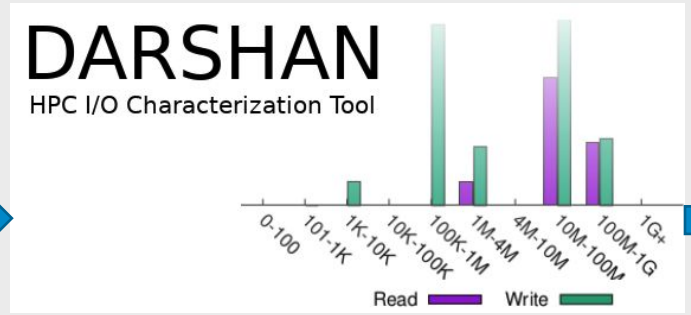


Workload I/O benchmark

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

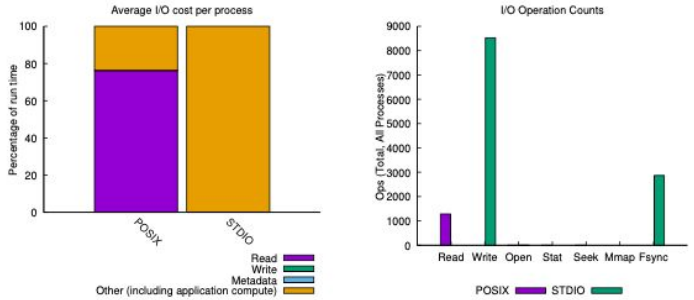


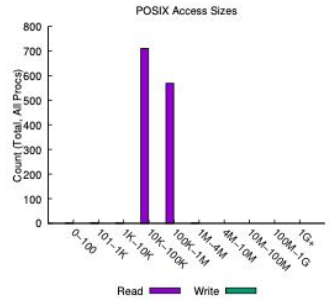
IoR
HPC benchmarks

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

Application: AI benchmarking

Approach 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)

Characterize I/O requirements for generalized workflows

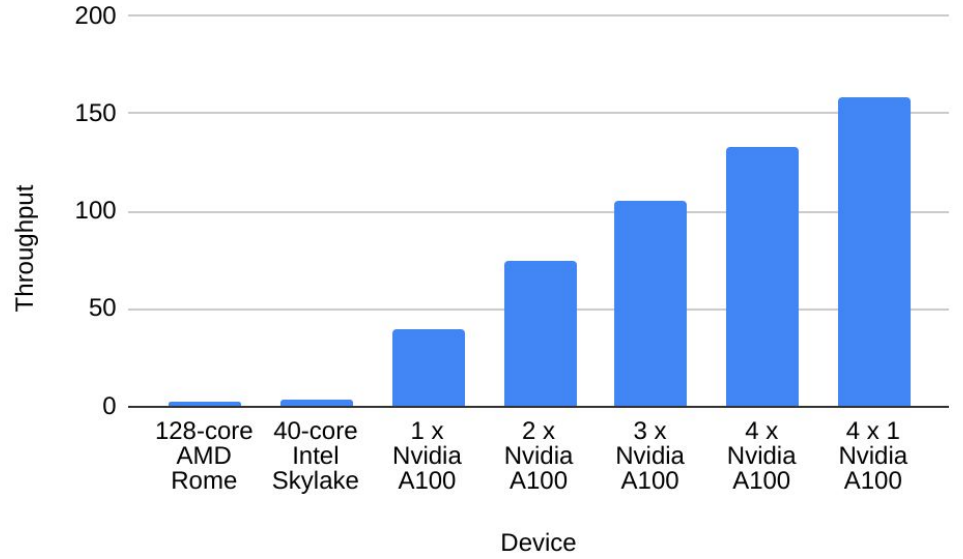
- Development work to increase granularity of characterization
- Automate profile generation (as much as reasonable)

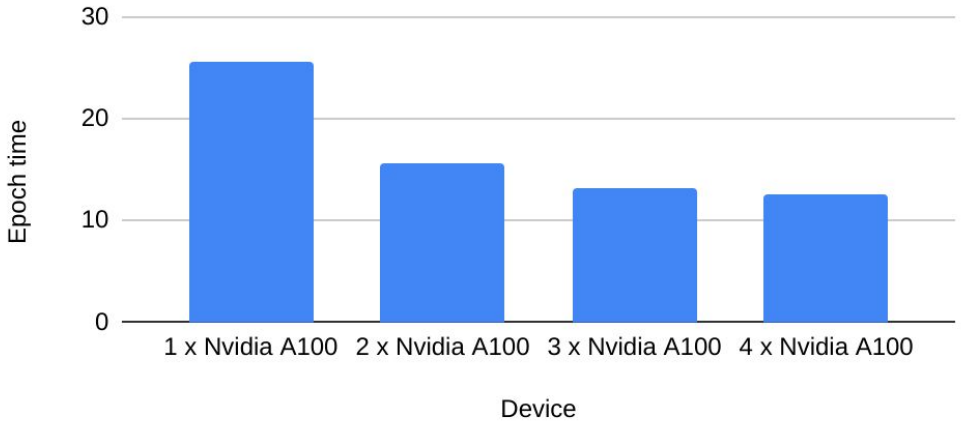
AI benchmarking

Machine-Learned Particle Flow (MLPF) from CMS
- GNN-based reconstruction algorithm
- Measure the training performance on GPUs
- Use event throughput as the metric

Integration into HEP-Score
- Support for multiple GPUs (single node)
- Containerized builds for Intel, Nvidia, AMD ROCm

Testing
- maximizing GPU throughput with multiple GPUs
- CPU parallelization and optimizations





Conclusions

- First ML/AI workloads for HEPiX benchmark working group introduced
- Growing support for heterogeneous workloads, accounting
- Testing feasibility of training / tuning HEP-driven AI applications on HPC hardware
- Generalized I/O characterization & benchmarking for HPC
- Development continues towards HPC computing, heterogeneous arch. support

drive. enable. innovate.

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:

Understanding workload CPU efficiency

- upcoming PRmon plugin to HEP benchmark suite enables profiling of CPU utilization
- Profile both native and containerized workloads
- Identify issues, acceptance testing, verification

PRmon source: https://github.com/HSF/prmon

https://indico.cern.ch/event/1078853/contributions/4576275

Supporting HEPiX WG paper:
https://doi.org/10.1007/s41781-021-00074-y