HPC Benchmarking for Exascale

openlab technical workshop 2023

David Southwick (CERN)

Heterogeneous Benchmarking in HPC

Benchmarking for deployment at HPC-scale presents new challenges external to traditional benchmarking endeavors.

Capturing a more complete snapshot of an HPC system's capabilities requires the combination of many elements – an array that will continue to grow as new technologies become available.

To successfully exploit HPC resources for Big Data workloads, we need to understand the capabilities of not only the compute nodes, but the attached accelerators and supporting file systems and networks.

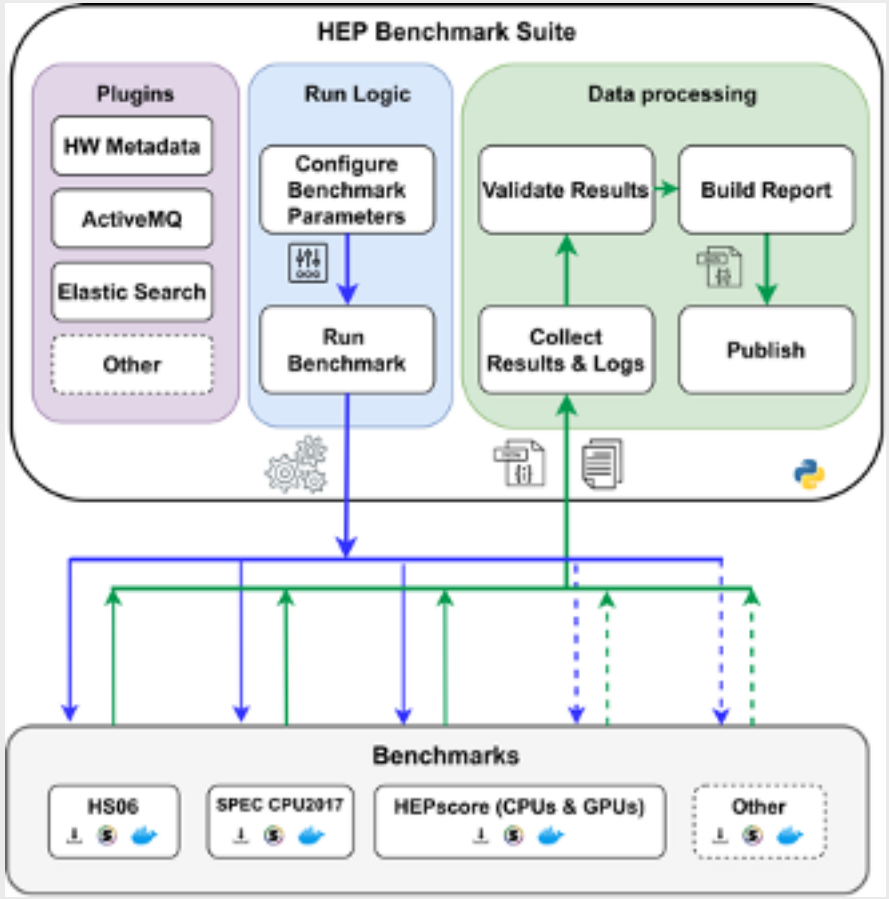
Context: Benchmarking at CERN

HEP Benchmark Suite: A benchmark orchestrator & reporting tool.

Executes an array of user-defined benchmarks & metadata collection

Support for HPC:

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



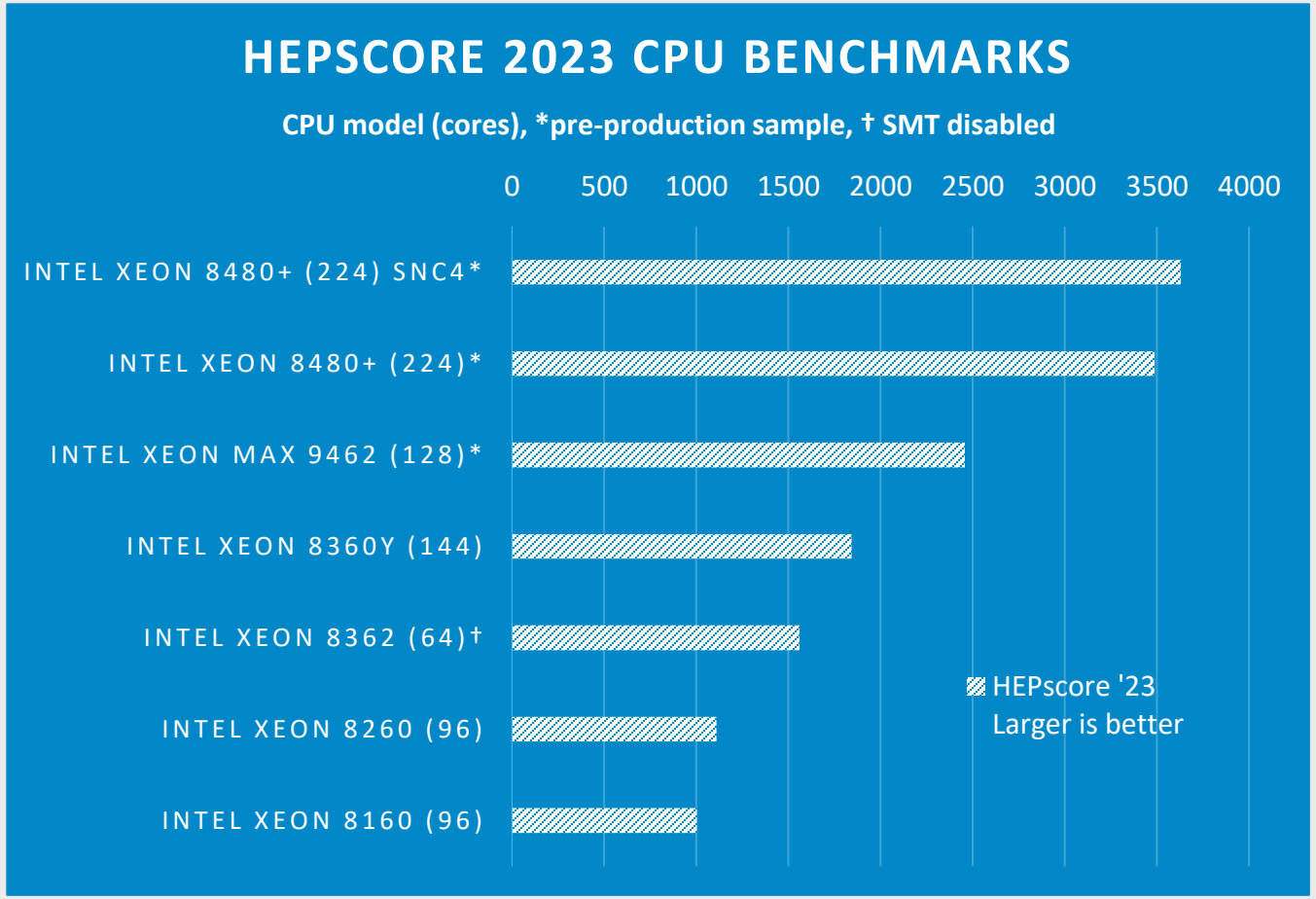
https://gitlab.cern.ch/hep-benchmarks/hep-benchmark-suite

Compute benchmarking

CPU Benchmarking

Traditional CPU benchmarking with HEPscore 2023:

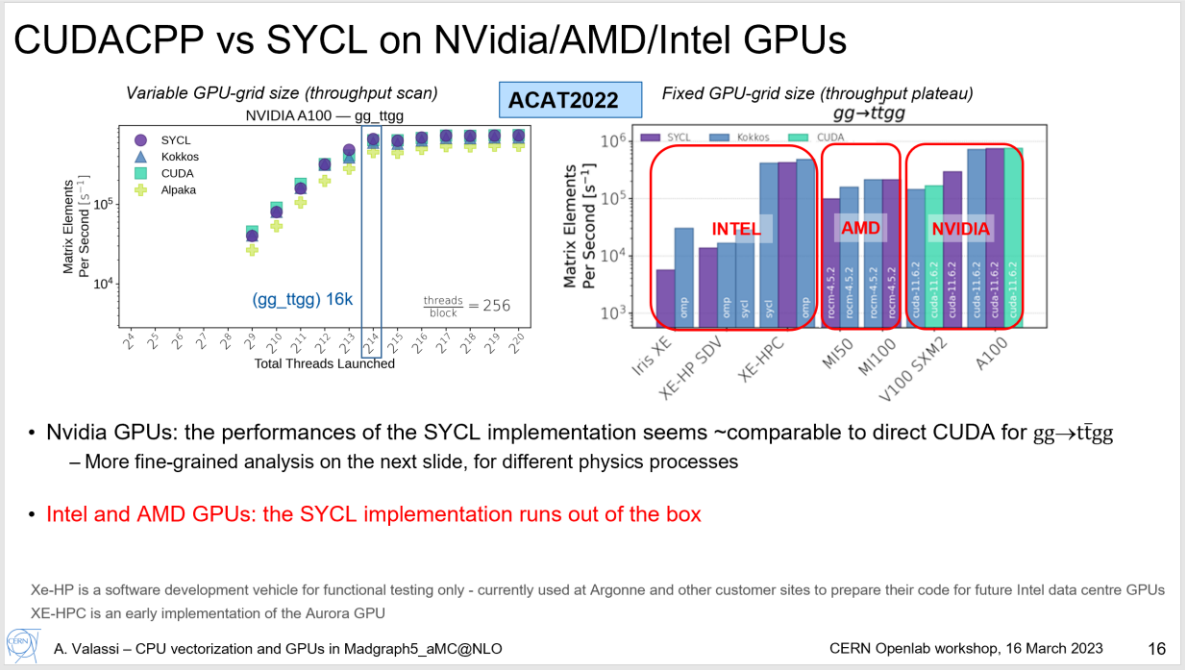
- Production HEP workloads
- Single number result (score)
- Controlled correlation to production
- Permits direct comparison across models and generations
- Arm, Power workloads in development

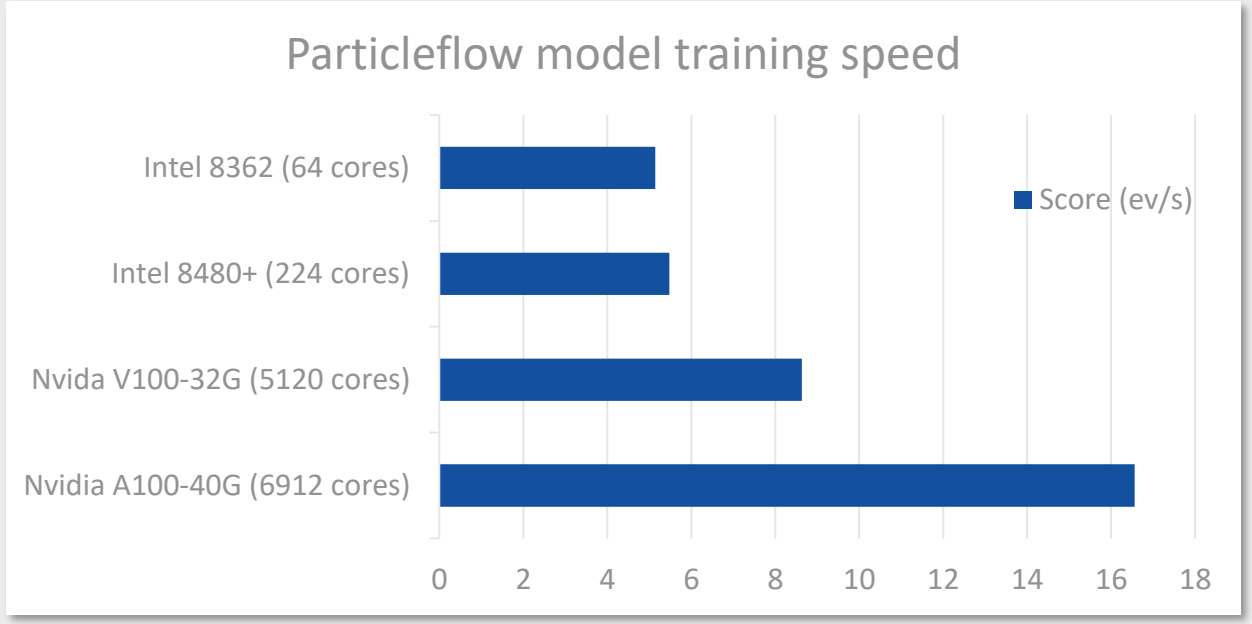


See talk by D. Giordano

GPU Benchmarking

- Approach GPU 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)



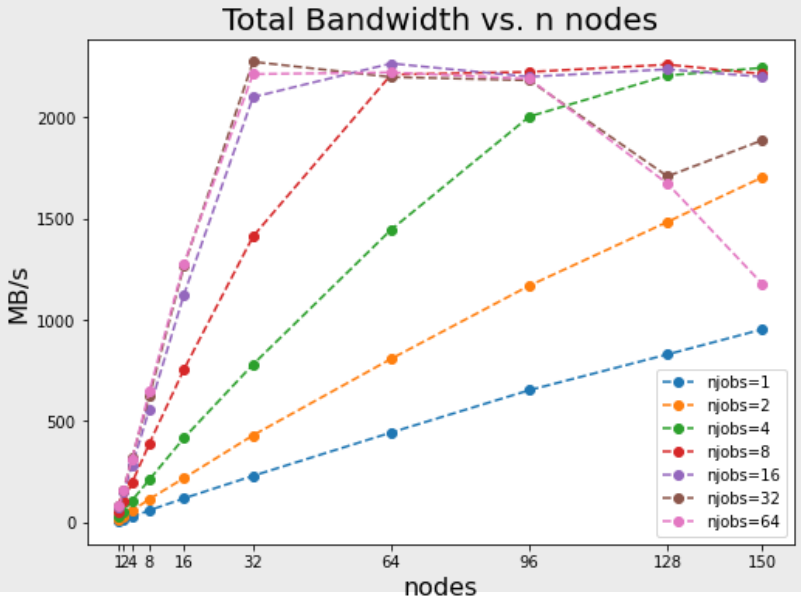


See talks by A. Valassi, E. Wulff

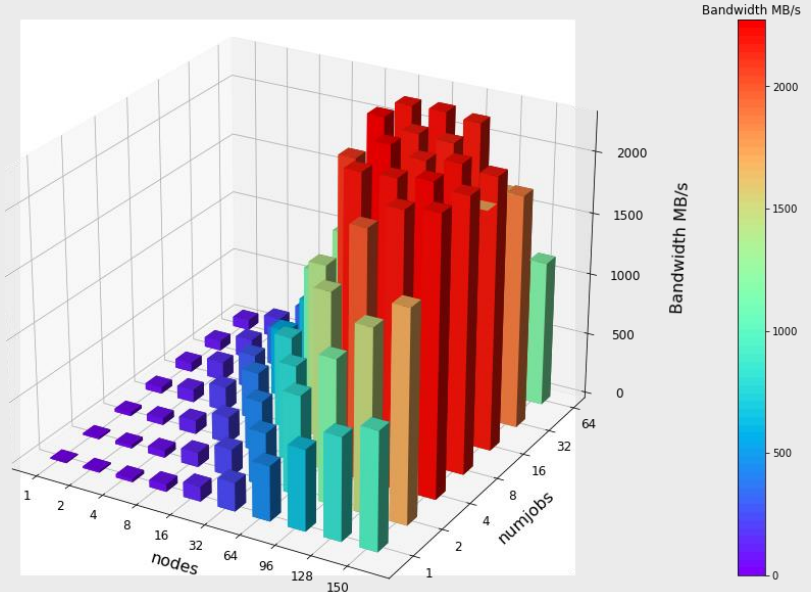
Non-compute benchmarking

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



Summer of HPC student HPC I/O investigation C. Cocha, A. Filipcic, V. Khristenko

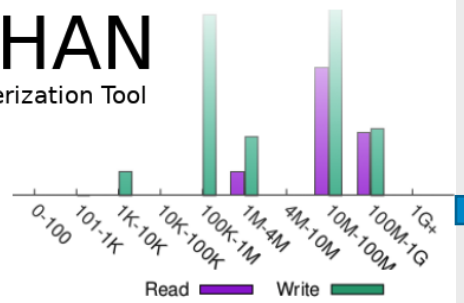
Workload I/O benchmark

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

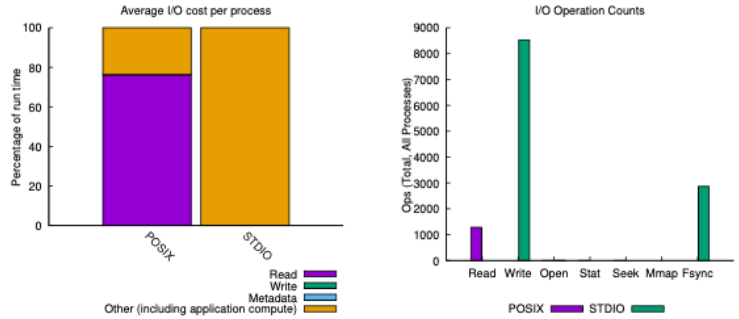
DARSHAN
HPC I/O Characterization Tool

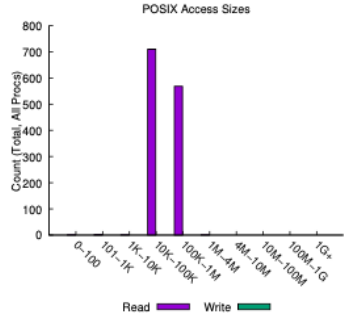


IoR
HPC benchmarks

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

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

Ongoing work

New formats and architectural structures may offer significant speedups on modern hardware that support them:

- Development and benchmarking of reduced precision (mixed precision) ML training for bfloat16. Testing on latest GPUs + CPUs (where supported)
- Sub-NUMA clustering studies on recent processors and accelerators
- Network I/O studies between HPC sites, CERN, GEANT network testbed
- Filesystem I/O studies for VAST NFS platform

Conclusions

Benchmarking efforts continue to grow, yielding a more complete system representation

- GPU workloads for HEP benchmarking maturing as studies continue
- Growing support for heterogeneous workloads and accelerators
- Filesystem & network benchmarks representative of HEP workloads
- We look forward to results of ongoing studies on the path to Exascale

drive. enable. innovate.

FLATIRON INSTITUTE

LUMI

CSCS

WLCG
Worldwide LHC Computing Grid

JÜLICH FORSCHUNGSZENTRUM

SDSC SAN DIEGO SUPERCOMPUTER CENTER

CERN openlab

RAISE Center of Excellence

intel

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: