High Energy Physics HPC Pilot

Benchmarking, Containerization, and Data Access

David Southwick
Maria Girone

Dissemination level: Public

Disclosing Party: CERN

Recipient Party: EGI-ACE Conference 2021

EGI-ACE receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101017567.

Background & Motivation

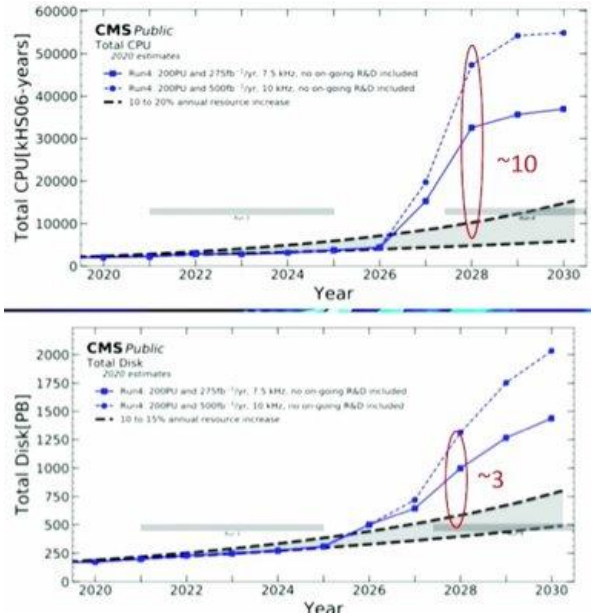
Big Data challenges in High Energy Physics

HL-LHC will produce more *computationally complex* physics events, with a *larger size* per event, and at *higher frequency:*

- Approaching the limits of what can be squeezed out of traditional CPUs for High Energy Physics (HEP) workloads
- Foreseen gap in computing resources far exceeds procurement feasibility by an order of magnitude

All avenues are being explored to enable processing and storage of full dataset:

- Aggressive event compression
- Aggressive event filtering
- Aggressive infrastructure expansion - **including HPC** compute sites (and **heterogeneous accelerators**)



Computing resource gap, CMS experiment projections

Pilot Concepts

CERN use case

Benchmarking heterogeneous resources

- Understanding and accounting compute accelerators and other architectures
- Understanding storage requirements when scaling jobs

Utilize heterogeneous compute resources & accelerators

- All experiments currently working to exploit accelerators (GPU/FPGA) and alt.arches
- Environments need to be packaged & mobile for shared computing

Data access & processing

- Enormous data volumes to stage, process, export from HPC sites
- Implicit authorization and authentication challenges
- Provisioning services for data management – both for dedicated storage site (Data lake) models and compute storage on HPC sites

Exploit synergies with other sciences!

HEP Benchmark Suite

A short history

HEP Benchmarking Suite: A benchmark orchestrator & reporting tool.
Provides an array of benchmarks, including HEPscore – the proposed solution for diverging HEPspec06 scores (over 15+ years use, EOL now)

- Collaboration with HEPiX Benchmarking Group to refactor & re-tool for **HPC** execution at scale!
- Importable, Extendable, and **architecture agnostic**
- Executes set of containerized workloads (Singularity, Docker, Podman, **uDocker***)
- Detailed report delivered in JSON via AMQ/Elastic Search
- Enables R&D benchmarking; comparison across heterogeneous architectures

Easily extendable to other areas of science!

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

Benchmarking HEP workloads on HPC

Project 2109 SoHPC

Supervised two students (Maria Herrero and Miguel Guerriero) with SURFsara (remote) for ~6 weeks on Benchmarking related tasks on HPC:

- Chose to investigate running **GPAW** benchmark from PRACE UEABS in a container at SURFsara with success
- Development work to integrate MPI-aware containerized benchmarks within HEP-Benchmark-Suite was out of time scale for program.
- Students tested and verified running HEP workloads using **uDocker** , eliminating the need for any software from HPC container solutions, enabling containerized benchmarking for any site (unrestricted from Singularity availability).

See uDocker documentation: https://indigo-dc.gitbook.io/udocker/user_manual

Benchmarking Heterogeneous Resources

Recent developments

PRACE preparatory access grant awarded for 6 months at two tier-0 sites for work on topics of benchmarking and data access

- CINECA Marconi 100 (25K core-hours) - IBM Power9, 4x NV V100 per node
- Juelich Juwels Booster (25K core-hours) - x86, 4x NV A100 per node

Heterogeneous computing investigations underway at these sites

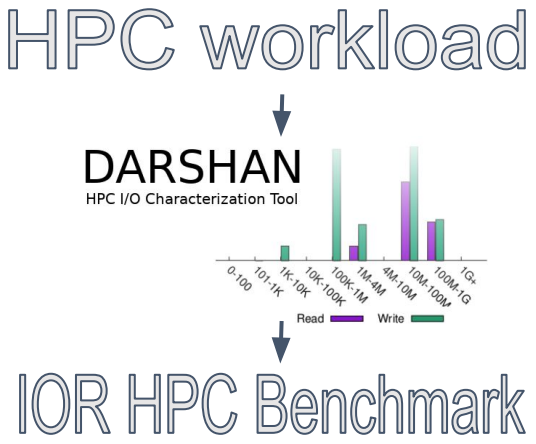
- Running CMS experiment workloads for POWER, x86, CUDA
- Containerization via Singularity, CVMFS where available
- Upcoming “run 3” software and ML/AI benchmarking on GPUs
- Comparison studies underway

Benchmarking File systems

New benchmark

Very little information is available about supporting file systems at HPC sites. Unclear how many data-driven workloads a given site may support

- Development of a *data access benchmark*
- tuned 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
- (more to come in next months)



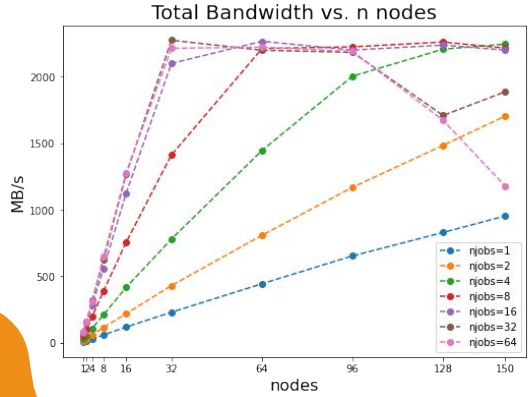
I/O characterization using Darshan

High throughput HEP data processing

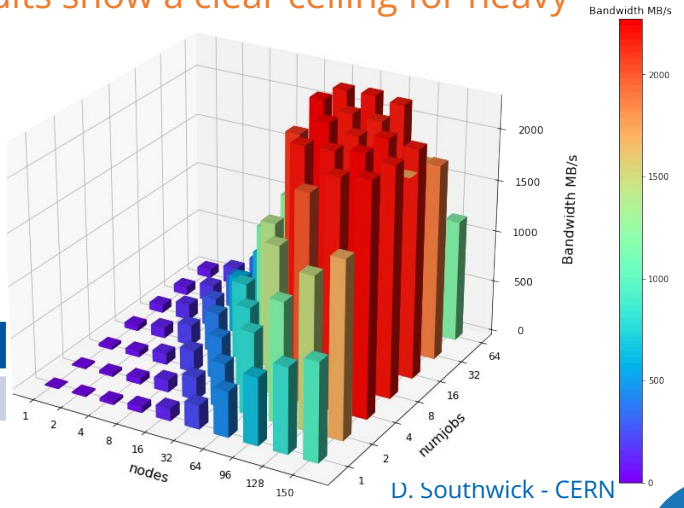
Project 2110 SoHPC

Students Carlos Cocha & Andraž Filipčič (mentored by Viktor Khristenko) investigated Data Access at CSCS:

- explored exercising the local shared file system using **FIO** and **IOR** synthetic benchmarks.
- Tested the storage solution by scaling parallel I/O threads and parallel nodes via MPI as provided by the IOR benchmark. Results show a clear ceiling for heavy parallel disk access.



Peak	Bandwidth
16 node	2.2 GB/s



Data Access

Exascale challenge

Upcoming run 4 (2027) expects **1 Exabyte physics data processing in 100 days**
Goal is to stream & process 10 PB of physics data through a HPC site in a day: several hundreds of Gbit/s continuously. HEP experiments can not store all the produced data at a single site.

- Challenge of increasing complexity: start with 10-20% goal (1PB), demonstrate management of hundreds of TBs data
- Maintain compute efficiency with high data rate in/out from/to storage & stream

Lots of moving parts! Break down challenge into three areas:

1. Data in/egress from HPC center
2. Efficient usage of storage systems on site
3. Dynamic scaling interaction between (1) and (2)

Throughput Investigations

Ingress/Egress capabilities

Collaboration with GÉANT and PRACE to perform distance throughput tests with workload-specific transfer protocols:

- GÉANT DTNs London/Paris to CINECA, Juelich
- GÉANT testbed service (GTS) permits containerized transfer tools
- Compare science-specific transfer tools (XrootD) alongside industry standard (iPerf, gftp, ethr, etc)

Look forward to update in coming weeks

Summary & Future work

CERN HPC Pilot

Lots of progress in these past months on HPC:

- Valuable experience gained for both students and mentors with SoHPC program
- Exploration of universal alternative to containerization: uDocker
- Development of benchmarking on heterogeneous architectures & GPUs
- Development of workload-driven storage benchmark
- Exploration of data access challenges both internal and external

AAI and throughput tests to be completed later this year!

See (draft) final results for SoHPC projects here
(final version to be published at https://summerofhpc.prace-ri.eu/)

Thank you!

Contact: **egi-ace-po@mailman.egi.eu**
Website: **www.egi.eu/projects/egi-ace**

EGI Foundation

@EGI_eInfra

CINECA

WLCG
Worldwide LHC Computing Grid

HPC Collaboration

GÉANT

SKAO

EGI-ACE receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101017567.

HEP Score

Experiment workload benchmark orchestrator

- Modular python3 “microservice” approach
- Importable, Extendable, and **architecture agnostic**
- Executes set of containerized workloads (Singularity, Docker, Podman, uDocker*)
- Workloads decided by experiment experts & WLCG teams
- Detailed report delivered in JSON/YAML via AMQ/Elastic Search
- **Simple to extend to other sciences**

4 large LHC experiments represented

Experiment	Name	Description	Experiment license	Readiness	Pipeline status
Alice	gen-sim	link	GNU GPL v3	w.i.p.	pipeline passed
Atlas	gen	link	Apache v2	Y	pipeline passed
Atlas	sim	link	Apache v2	Y	pipeline passed
Atlas	digi-reco	link	Apache v2	w.i.p.	pipeline passed
CMS	gen-sim	link	Apache v2	Y	pipeline passed
CMS	digi	link	Apache v2	Y	pipeline passed
CMS	reco	link	Apache v2	Y	pipeline passed
LHCb	gen-sim	link	GNU GPL v3	Y	pipeline passed
Belle2	gen-sim-reco	link	GNU GPL v3	Y	pipeline passed

https://gitlab.cern.ch/hep-benchmarks/hep-workloads

Benchmarking on HPC

Production workloads with HEP Benchmark Suite

Workload containers packaged as OCI-compatible docker/singularity images

- Multi-arch container workloads (x86_64, IBM Power, ARM, ...)
- Multi-GPU container workloads (Nvidia, AMD, Intel OneAPI)

Simple integration with SLURM & other job orchestrators

- Single dependency on Python3.6 + container service of your choice


```
# HEP Benchmark Suite requires singularity 3.5.3+, python3.
module load singularity python3
python3 -m pip install --user git+https://gitlab.cern.ch/hep-benchmarks/hep-benchmark-suite.git

echo "Running HEP Benchmark Suite on $SLURM_CPUS_ON_NODE Cores"
srun bmkrun --config default
```

Data Access (cont)

Exascale challenge

Data in/egress from HPC center
- Modernising transfer tools to be able to fill 100Gb/s wide-area netv

Efficient use of storage systems on site
- Reducing the local footprint of data management components
- Exploring the amount of local storage needed at each scale

Local data delivery to the processing nodes
- Performance of local storage
- Local network structure

Pilot program to test all these elements