HTC to HPC transition activities

Investigating Benchmarking, Containerization, and Data Access in EGI-ACE

David Southwick
Maria Girone

Dissemination level: Public

Disclosing Party: CERN

Recipient Party: WLCG Grid Deployment Board (GDB)

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

Background

Big Data challenges in High Energy Physics

Participated last year as HPC pilot case in *EGI-ACE* (WP 7.3), as well as member *CERN-PRACE-GEANT-SKA* consortium. Very busy year with much progress on CERN tasks for HTC-HPC integration:

- Build on improvements to HEP Benchmark Suite (and hepscore) for HPC (collaboration with *HEPIX Benchmarking wg*)
- Participation in PRACE Summer of HPC student program
- Preparatory Access grant - investigated workload scaling, benchmark development
- Data access investigations on HPC, internal & external

Challenges

Approaching HPC as High Throughput Computing (HTC)

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 alternative archetectures
- 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

Benchmarking

Extending to HPC

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). Previously:

- Designed for WLCG homogeneous compute environment
- Intended for procurement teams, site administrators
- First with VM containment, later nested docker images

None of these approaches are compatible with HPC:

- Collaboration with HEPiX Benchmarking Group to refactor & re-tool for **HPC** execution at scale!
- Enables R&D benchmarking; comparison across heterogeneous architectures

HEP Benchmark Suite

HEP Benchmark Suite 2.0: Now with 100% more HPC!

Minimal dependencies
- Python3 + container runtime choice

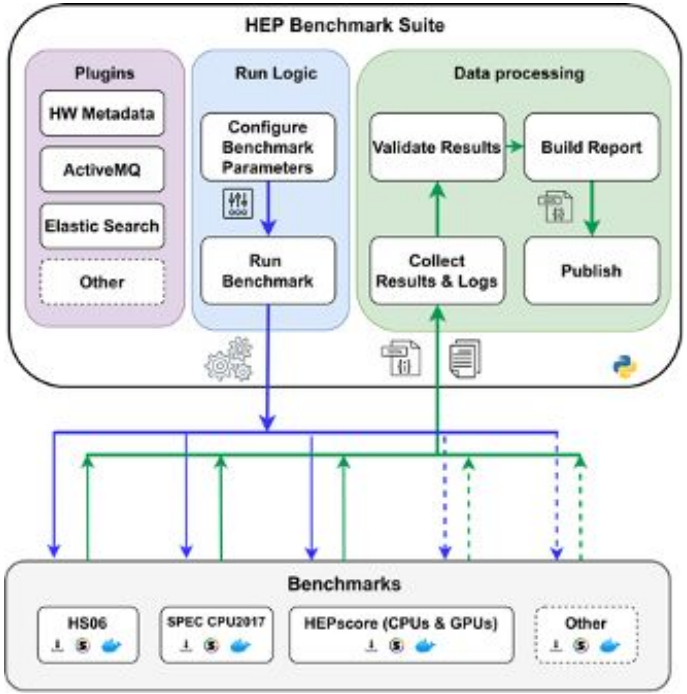
Modular Design
- Snap-in workloads & modules

Repeatable & Verifiable
- Declarative YAML config, hashed config in report

Designed for Ease-of-Use
- Simple integration with any job scheduler

Variety of containment choices
- Singularity (incl. CVMFS Unpacked), Docker, Podman*, uDocker* soon

Metadata + Analytics
- **Improved** Automated Reporting via AMQ



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

HEP Score

Experiment workload benchmark orchestrator

- **NEW** containerization options for HPC: uDocker. Switch to Singularity by default (HPC) Podman coming soon
- **NEW** HPC reporting over AMQ improvements - simplify reporting for constrained nodes
- **NEW** run3 workloads coming online, much activity to integrate these new environments (and alt. Arch. workloads as HPC demonstrators)
- **NEW** ML/AI PoC benchmark based on ParticleFlow (single node available, MPI to follow!)

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

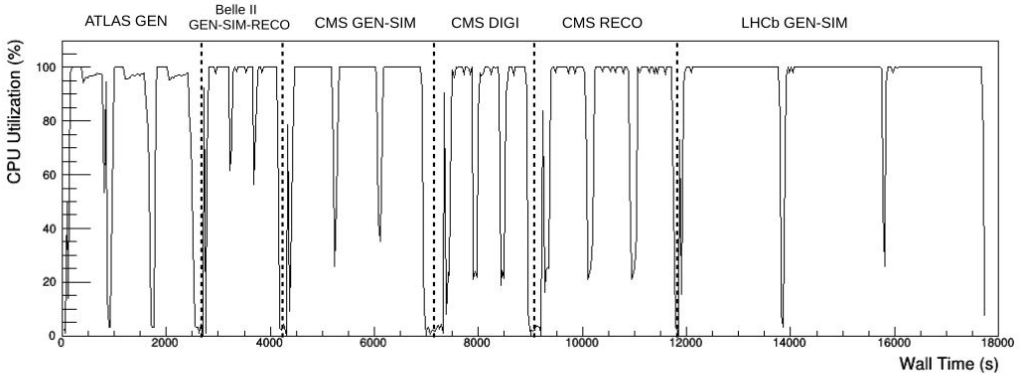
https://gitlab.cern.ch/hep-benchmarks/hep-workloads

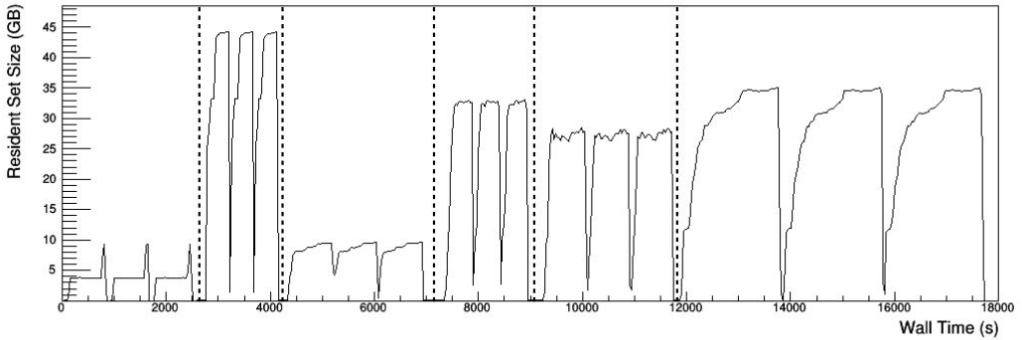
Understanding benchmark efficiency

HEP Benchmark Suite - prmon plugin

PRmon plugin to HEP benchmark suite enabling profiling of CPU utilization

- Profile containerized benchmarks (Singularity, others soon) + native
- Identify issues, acceptance testing, verification





See recent update to HEPiX workshop (26 Oct.) by Domenico Giordano: https://indico.cern.ch/event/1078853/contributions/4576275/

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

Benchmarking HPC

Summer Student activities

Mentored two PRACE Summer of HPC student projects (Jul-Aug)

- Benchmarking HPC workloads on HPC (2109)
- High Throughput HEP data processing (2110)

Two student per project, participating (remote) with mentoring HPC sites at **SURFsara** and **CSCS.**

Benchmarking students:

- Chose to investigate running **GPAW** benchmark from PRACE UEABS in a container at SURFsara with success
- 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

Ongoing benchmarking efforts

PRACE preparatory access program at two sites for work on topics of benchmarking and data access

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

NB: development application for HPC access will soon change:
https://prace-ri.eu/benchmark-and-development-access-information-for-applicants/

Data Access

Exascale Challenge

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 cannot 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 (members of CERN-GÉANT-PRACE-SKA collab) to perform distance throughput tests with workload-specific transfer protocols:

- GÉANT DTNs London/Paris to CINECA, Juelich, others
- GÉANT testbed service (GTS) permits containerized transfer tools
- Compare science-specific transfer tools (XrootD) alongside industry standard (iPerf, gftp, ethr, etc)
- Investigate requirements for edge networks (NAT/DMZ), caching on both side of link, interface with Data Lakes (RUCIO)

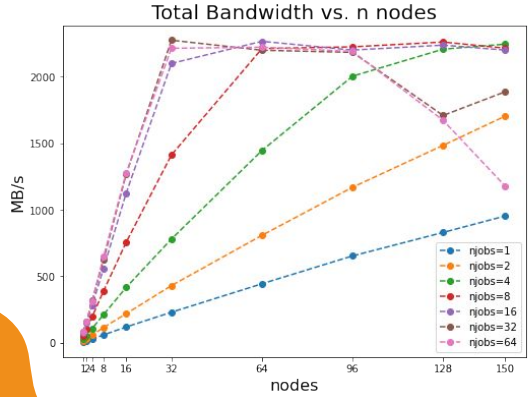
CINECA only has 1x 40Gbps link, shared and throttled (no DMZ for xrootd) - but good partnership with CNAF
Juelich very familiar with XrootD, high connectivity
Testing will continue at other HPC sites this year

High throughput HEP data processing

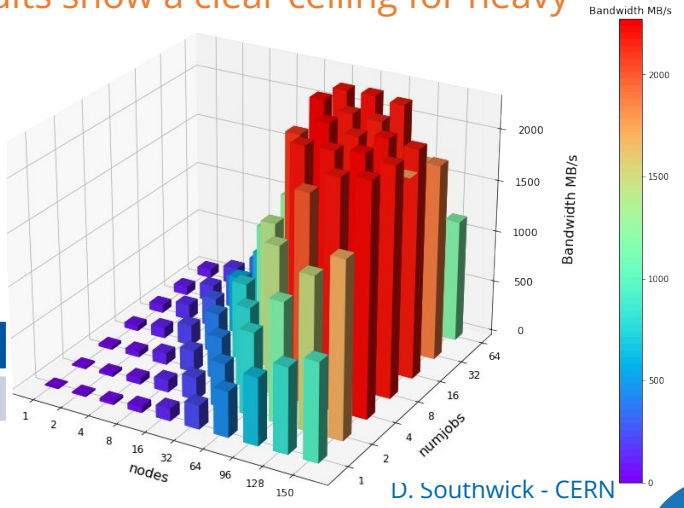
Project 2110 SoHPC

Students Carlos Cocha & Andraž Filipčič (mentored by V. 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



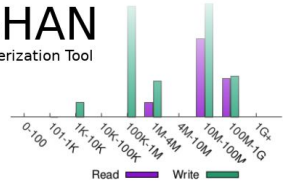
Benchmarking File systems

Big-data scalability benchmark

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

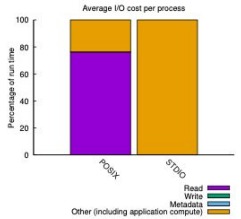
- Development of a *data access benchmark*
- tuned to the **I/O patterns of real workloads** to better inform reasonable scaling capabilities at a given HPC site
- More representative than sequential throughput metrics
- Uncover **I/O bottlenecks** (excessive file opens, read patterns, cache issues)
- Basic version available now, testing & development underway

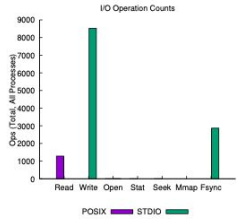
HPC workload → DARSHAN HPC I/O Characterization Tool → IOR HPC benchmark

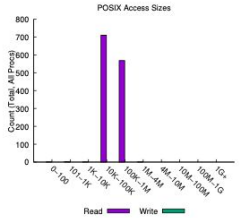


jobid: 2190289	uid: 1005	nprocs: 1	runtime: 6 seconds

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

I/O characterization using Darshan

Conclusions

Big Data challenges in High Energy Physics

Participated last year as HPC pilot case in *EGI-ACE* (WP 7.3), as well as member *CERN-PRACE-GEANT-SKA* consortium. Very busy year with much progress on CERN-side!

- Built on improvements to HEP Benchmark Suite (and hepscore) **for HPC** (collaboration with *HEPIX Benchmarking wg*)
- Participation in PRACE Summer of HPC student program
- HTC job requirements (disk, throughput) make difficult to “shop” for - HPC sites poor visibility for this job type
- Throughput tests setup, to continue in 2022

Summary & Future work

CERN HPC Pilot program

Lots of progress in these past months for HPC:

- Overhaul of Benchmarking approach for HPC
- Development of heterogeneous container bmk for CPUs & GPUs
- Characterization tooling for HPC jobs (both benchmarking + MPI)
- Approaching HPC storage as a heterogeneous resource (I/O benchmark)
- Exploration of data access challenges both internal and external

...and lots of work still remains!

- AAI and throughput tests ongoing
- Scaling from prototypes to production tests
- Data-Intensive workflows investigation with CoE RAISE
- To be continued ... :-)

Thank you!

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

EGI Foundation

@EGI_eInfra

HPC Collaboration

CERN GÉANT PRACE SKAO

SUMMER OF HPC

EGI-ACE

CERN openlab

WLCG Worldwide LHC Computing Grid

ROOT Data Analysis Framework

JÜLICH FORSCHUNGSZENTRUM

CINECA

FLATIRON INSTITUTE

SDSC SAN DIEGO SUPERCOMPUTER CENTER

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

Backup

Early adopter experiences for procurement

Thanks to RAL (Alastair D.), Nikhef (Andew P.), IJCLab (Emmanouil V.)

https://indico.cern.ch/event/1072128/