HEP workloads on HPC resources

David Southwick, Maria Girone, HEPiX Benchmarking working group
additional support from CERN, WLCG

ACAT 2021

Introduction

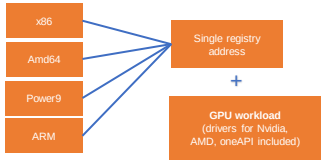
Approaching HPC site with High Throughput Computing (HTC) workloads presents unique challenges not found in homogenous WLCG compute sites. We present work enabling the benchmarking of HEP workloads on heterogeneous resources found at HPC centers (in collaboration with the HEPiX Benchmarking working group), as well as initial investigations into the topic of data access. We share the development and testing of a benchmark for HPC shared storage systems which may be used as a metric to guide job scheduling policy.

Benchmarking on HPC

Collaborating with the *HEPiX benchmarking working group*, we have extended the functionality of a candidate (HEPscore) replacement for HS06 in order to meet the demands of execution at scale on HPC centers.

Moving from nested Docker containers to unprivileged OCI-compatible Singularity containers by default enables root-less execution and compatibility with several containerization services available at HPC sites.

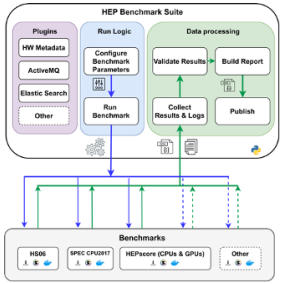
In addition, workloads utilizing heterogeneous compute accelerators and alternative architectures are fully supported via two new methods in container technologies: multi-arch container manifests, where the client chooses the architecture, and multi-gpu container builds, where the container includes all necessary drivers. Both approaches may be combined.



Benchmarking at HPC scale

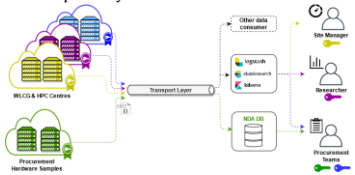
In addition to works enabling root-less, containerized heterogeneous workloads, the automated execution and collection of workload results has been completely rewritten for the HPC environment, enabling execution across partitions at scale with automated reporting.

- Architecture-agnostic Python3
- Modular plug-and-play framework
- Singularity containers by default (Podman, uDocker dev support)
- Declarative, repeatable running conditions via simple YAML
- Report standard JSON to choice of endpoints
- Automated reporting via AMQ
- Open source
- **Easily extended to other sciences!**



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

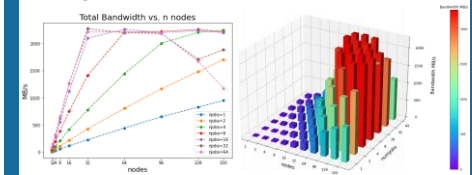
Development for uniquely constrained environments, such as where HPC compute nodes have no external network connectivity have been contributed by means of batch uploading of node results after execution from a node with connectivity, and pre-caching of workload images and python wheels required by the suite.



Refactoring for the more contrained HPC environment has resulted in a more robust and yet flexible, modular, architecture-agnostic benchmarking framework that is easily extended to run a large array of benchmarks both in and outside of big-data sciences!

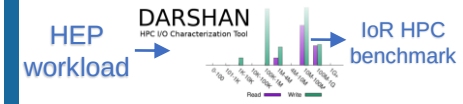
Shared storage system usage

Scaling simultaneous data-driven workloads is dependant on the performance of the supporting shared filesystem, and is not characteristic of "Traditional" HPC workloads. Given this, and the lack of information generally available about HPC storage performance, a new I/O benchmark based on HEP workload I/O characteristics is under development.



Initial testing showing example throughput ceiling of 16 nodes @ 2.2GB/s

To avoid the need to test by trial and error at each unique HPC storage setup, we have developed a rapid I/O benchmark, built on industry standard disk benchmarks. Using Darshan, we record HEP Workload I/O characteristics, which is then used to inform and customize industry standard disk benchmarks (IoR, FiO, mdtest) to reproduce a characteristic disk usage pattern, which may then be scaled against any given storage solution, without the need for tooling, setup, execution, and possible disruption of other users with an otherwise lengthy investigation.



Data Ingress/Egress

Collaboration with GÉANT and PRACE as members of CERN-GÉANT-PRACE-SKA collab. is underway to perform distance throughput tests with workload-specific transfer protocols:

- GÉANT DTNs London/Paris to CINECA, Jülich, and others
- GÉANT testbed service (GTS) permits containerized transfer tools
- Compare science-specific transfer tools (XrootD) alongside industry standard (iPerf, gftp, ethr, etc)
- Developing requirements for NAT/DMZ
- Better caching on both side of link, informed by scaled workload I/O requirements from previous I/O benchmark

Testing will continue as HPC sites upgrade their connectivity to NRENs

Thank you to supporting partners!

WLCG Worldwide LHC Computing Grid
HEPiX
JÜLICH FORSCHUNGSZENTRUM
CINECA
CERN
PRACE
GÉANT
SKAO

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