CERN openlab

High Performance Computing

David Southwick (IT-GOV-INN)

Openlab lecture series 2024

High Performance Computing

HPC centers are host to cutting-edge technologies that advance modern computing methodologies:

- AI/ML and scalable distributed workloads
- Heterogeneous technologies and topologies
- GPUS, compute accelerators (FPGAs, Quantum)
- Exascale infrastructure

CERN Openlab partners with industry and collaborates with organizations to further mutual HPC adoption:

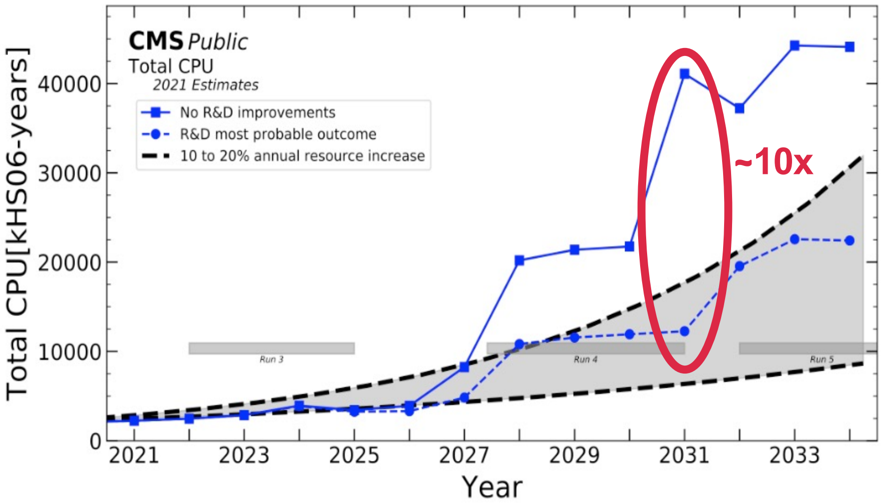
- Advancing HEP use cases via participation in EU projects
- Prototyping new and upcoming compute technologies
- Studying, Developing & Promoting novel software, methodologies, and toolkits
- **Building a community** with computing partners & projects

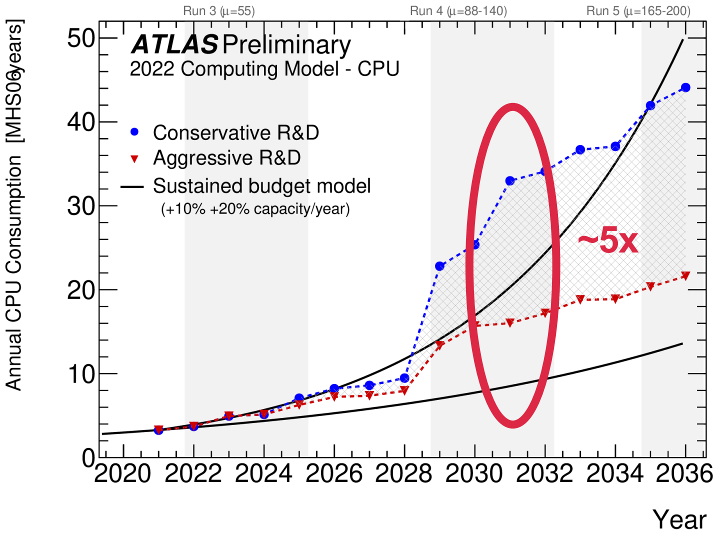
HEP Motivation

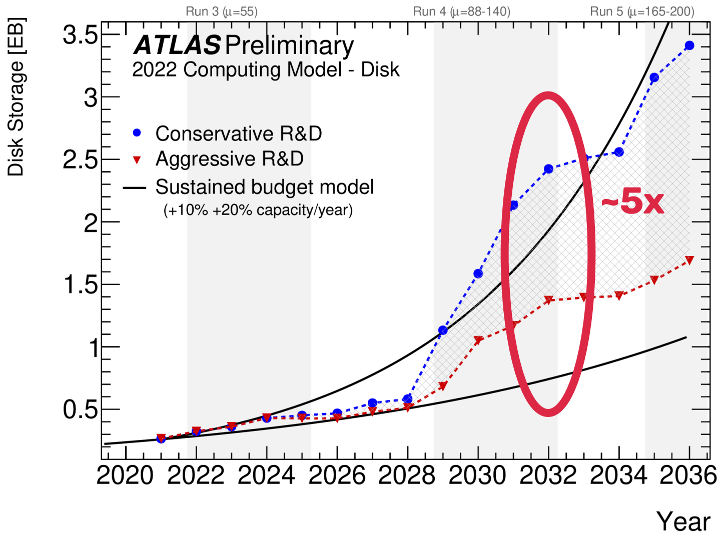
LHC expects more than exabyte of new data for <u>each year</u> of HL-LHC era from ~2029-2040.

This data must be exported in ~real time from CERN to compute sites.

CERN is not alone: SKAO expects similar requirements during similar period; other big-data sciences to follow







For more detail on how we get here see DAQ filtering:
https://indico.cern.ch/event/1386474

CMS: https://indico.jlab.org/event/459/contributions/11470/ https://cds.cern.ch/record/2815292
ATLAS: https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/UPGRADE/CERN-LHCC-2022-005/

Outline

- Intro & motivation

- What makes a HPC different?
- Software and Architectures
- Runtime Environments and Containers
- Provisioning

- Benchmarking and Accounting
- Data Processing and Access
- Authentication and Authorization
- Wide and Local Area Networking

CERN Computing

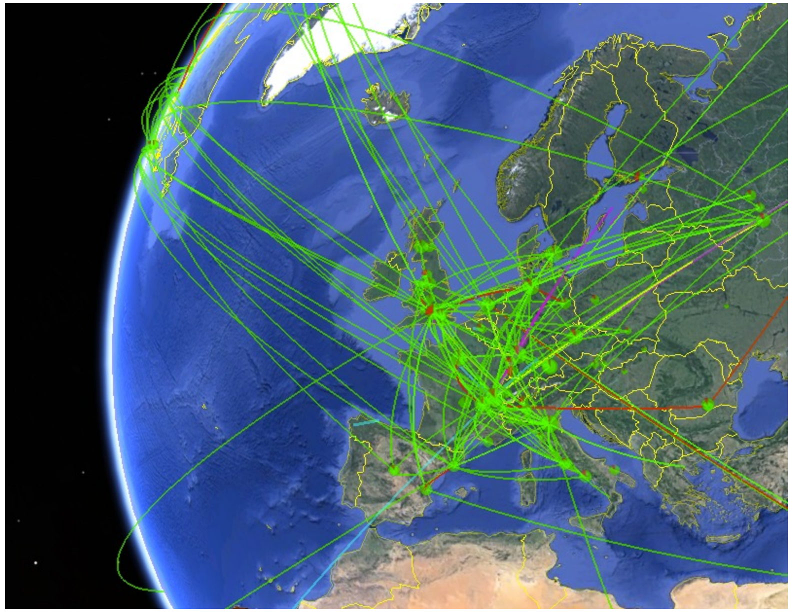
WLCG
Worldwide LHC Computing Grid

The Worldwide LHC Computing Grid (WLCG) is the distributed computing grid that provides ~12,000 physicists with ~local access to LHC data

- Around 1.5 Million CPU cores running 24/7
- 900 Petabyte disk, 1.4 Exabyte tape
- CERN provides ~20% of WLCG resources

WLCG sites provide a common "standard" set of resources

- Authorization/Accounting
- ~Homogenous Hardware / disk space
- Edge service (CVMFS, etc)
- Network and disk speed policies



NB! Compute performance based on Event Throughput (more about this later…)

Key Differences

HPC environments typically have three core components: compute processors, networking, and storage.

A core demand of HPC projects is reducing latency while leveraging as many resources as possible!

HPC centers differ from typical datacenters (like those operated here in the IT dept.) in several key areas:

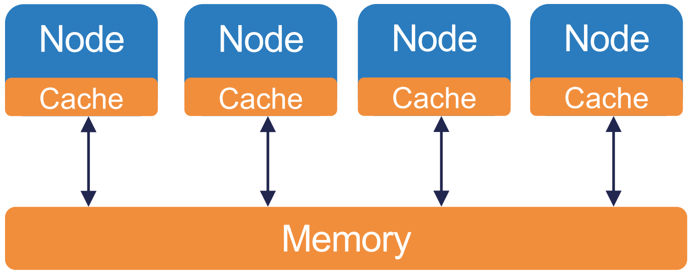
High Performance Computing (HPC)	High Throughput Computing (HTC)
Designed for maximum performance of a single task	Designed for maximum number of parallel tasks
Vertical scaling of resources for performance	Horizontal scaling for more task throughput
Resource management to schedule & distribute job across many nodes	Distributed resource manager(s)
Fault tolerance necessary	Tasks easily repeatable
Extreme network interconnect	Interconnect less important
Heterogeneous, high-performance hardware	Cost-effective hardware

HPC Topologies

Historically there have been two topologies:

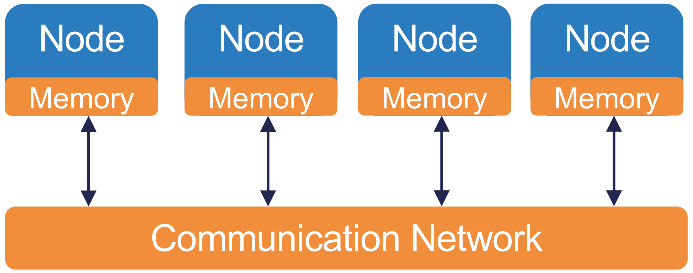
Shared-memory parallelization, with OpenMP

- All CPUs share a common physical address space, UMA or ccNUMA
- Implicit communication via memory operations
- 'Cache Coherence' protocols to avoid different CPUs modifying same values



Distributed memory parallel programming, with Message Passing Interface (MPI)

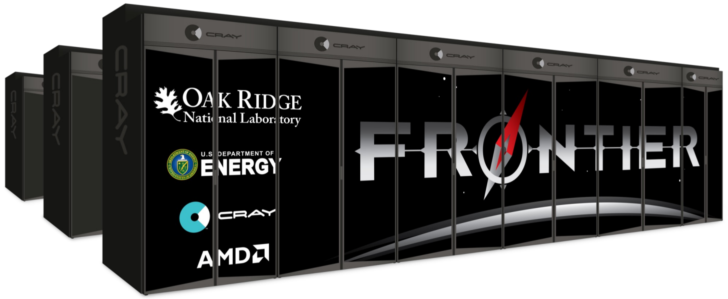
- CPUs communicate via network messages
- Private memory space
- Complicated programming, but most flexible



In practice, most systems provide hybrid of both, are considered 'programming models'

- This is changing with the introduction of quantum devices, neuromorphic devices

HPC Performance



Performance of HPC hardware is typically measured by how many Floating Point operations can be performed per second (FLOP/s).

Operations matter! – Nvidia RTX 4090 has 64x more FLOP/s for Single Precision (FP32)!

1 GigaFlop/s = 10^9 FLOP/s
1 TFlop/s = 10^{12} FLOP/s
1 PFlop/s = 10^{15} FLOP/s
1 ExaFlop/s = **10^{18}** FLOP/s

The Top 500 organization compares the performance of some of the fastest machines in the world based on double precision (FP64) performance. They also publish Green 500 comparing efficiency.

The current leaders are Frontier, a supercomputer with 9,472 EPYC 7453 CPUs and 37,888 Instinct MI250X GPUs, scoring 1.7 Exaflops/s, and JUPITER/JEDI with 272 GraceHopper superchips at 72.7 GFlop/Watt

JÜLICH Forschungszentrum

Name	FP64 (Gflop/s)	Notes
Raspberry PI 4 Model B	13	Inexpensive ARM SoC
Nvidia RTX 4090	1,142	Desktop GPU
PS5 GPU	643	Gaming console
AMD EPYC 9474F (96 threads)	5,530	Typical HPC CPU
Nvidia H100	33,500	HPC GPU

Today, the majority of HPC performance is delivered by GPUs!

HPC Networking

HPC communication networks (Fabrics) are more scalable and very low latency compared to typical Ethernet:

Specialized Hardware:

- 400+ Gbps interfaces, coupled with special non-blocking low-latency switches and NICs

Specialized Protocols:

- Full-path network protocols, most commonly Infiniband (IB)
- RDMA (Remote Direct Memory Access) allows direct access to remote memory without involving CPU or caches

Most HPC fabrics can achieve round-trip times of 1.3μs or lower!

Storage

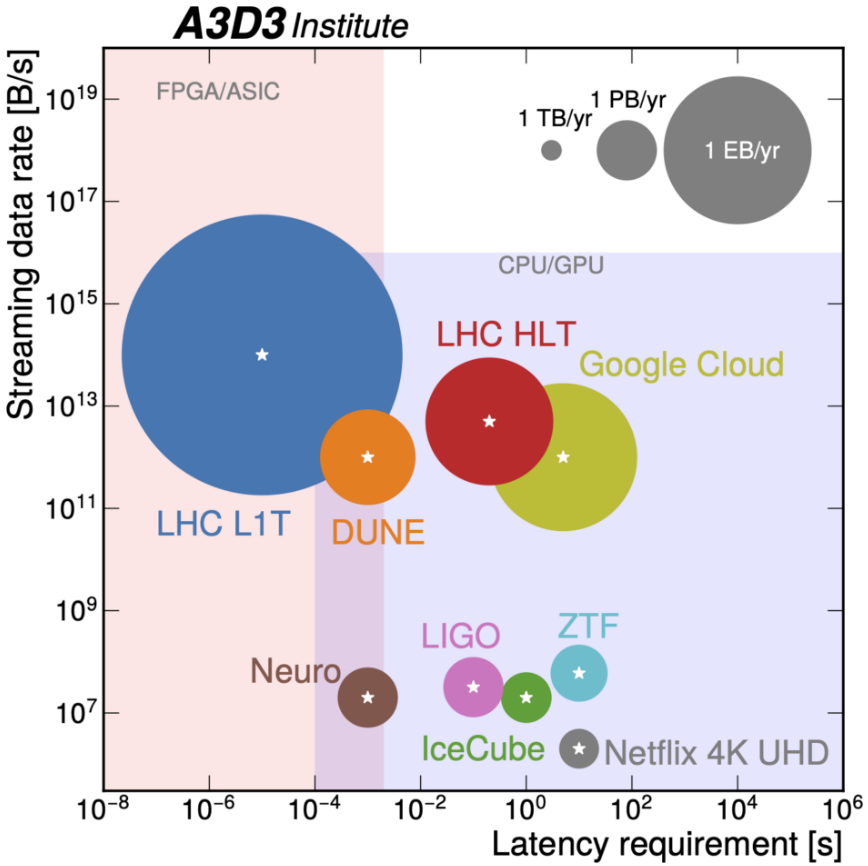
HPC storage must be performant enough to support the demands of high bandwidth, low-latency distributed workloads.

To operate at speed and scale network storage is typically composed of multiple metadata server and storage backends on the fabric.

Metadata servers act as directors mapping I/O requests to the closest available storage backend(s), tracking hot and cold data.

HPC sites generally offer several pools, based on performance need:

Object, Block, and File storage, often using a variety of storage platforms and filesystems.



See the dedicated talk on Storage: https://indico.cern.ch/event/1431367/

Software & Provisioning

The majority of HPC users today interact over SSH, via CLI or script.

SLURM (Simple Linux Utility for Resource Management) is the dominant resource manager for HPC. It is used both by site operators and users to:

- Allocate resources within a cluster
- Launch, manage, and monitor jobs
- Arbitrate resource contention, manage queues

SLURM allocations (jobs) are defined based on user constraints, then scheduled by cluster policy. Jobs run until completion, or according to the schedule policy (time, resources, etc.).

The `module` system is commonly used to expose the large number of software packages available on a HPC system. It manages a user's environment to prevent incompatibilities.

User software is often containerized via Apptainer (Docker, but built for HPC)

SLURM Basics

- `salloc` – request a resource allocation (job), prematurely end with `scancel`
- `srun` – lauch a job step (and request allocation if needed)
- `sbatch` – submit a script to the scheduler
- `sinfo` - displays the state of nodes, structured by partition
- `squeue` – displays state of the job queue
- `sacct` – display accounting data for past jobs, `seff` to view efficiency

All of these commands accept arguments – see the docs for details!

```
$ cat mpi_gpu_example.sbatch
#!/bin/bash
#SBATCH --job-name=examplejob # Job name
#SBATCH --partition=gpu       # partition name
#SBATCH --nodes=2             # Total number of nodes
#SBATCH --ntasks-per-node=8   # 8 MPI ranks per node, 16 total (2x8)
#SBATCH --gpus-per-node=8     # Allocate one gpu per MPI rank
module load openmpi
CPU_BIND="map_cpu:49,57,17,25,1,9,33,41"

export MPICH_GPU_SUPPORT_ENABLED=1

srun --cpu-bind=${CPU_BIND} <executable> <args>
```

```
srun --gpus=4 -C h100 --pty bash
```

HPC at CERN: https://batchdocs.web.cern.ch/linuxhpc/quickstart

use cases, progress, and challenges

Closing the computing gap with HPC adoption

HPC adoption

Today, nearly all areas of CERN are developing for HPC

Industry drove the convergence of AI and HPC with large model development and the need for faster insights to data.

Big-Data sciences (including HEP) have been investing in ML/AI development in diverse areas, often with many difficulties!

2nd CERN IT Machine Learning Workshop

- Common theme: Need for resources!

…but there is much more to HPC than only GPUs!

Status: AI is here to stay

ATLAS:

- Most simulation is still classical (but **Fast ML based on GAN is in production**)
- **Tagging is fully ML**, tracking classical, trigger mostly classical.
- Analysis is mostly classical or simple ML models
- **Expect 50% of ATLAS algorithms accelerated by GPU-based ML by 2030s**

ALICE:

- Multiple ML workloads with different data, training, deployment patterns
- **So far, smaller scale and simpler models than in ATLAS and CMS**

CMS:

- Multiple ML-based reconstruction **already in production**
- **Advanced use cases, highly customized**
- **Moving toward larger models (transformer - based)**
- **Extensive work** at the level of ML optimisation, frameworks (ML fully integrated in CMSSW),
- **At least 30% of CMS algorithms are ML-based today**

LHCb:

- Main use cases for online operations and trigger
- **Requirements at the analysis level are lower, given the data is simpler and luminosity lower than at ATLAS or CMS**

ATS:

- **Automation of the accelerators infrastructure** is the main scope for ML research
- In addition: **accelerator design** and **AI assistants (LLMs)**

HPC Opportunities and Challenges

Enormous computing resources that are far more heterogeneous than typical Grid sites

- Early adopters of technology, including accelerators
- Advanced low-latency networking
- Driving green computing

Complex to migrate from homogenous grid computing:

- Software and architecture adoption (workloads, schedulers, benchmarking, data handling infrastructures…)
- Authorization, Authentication, Accounting
- Networking
- Provisioning (opportunistic vs Pledged resources)

First outlined for HEP in 2020:

[Common challenges for HPC integration, M.Girone]()

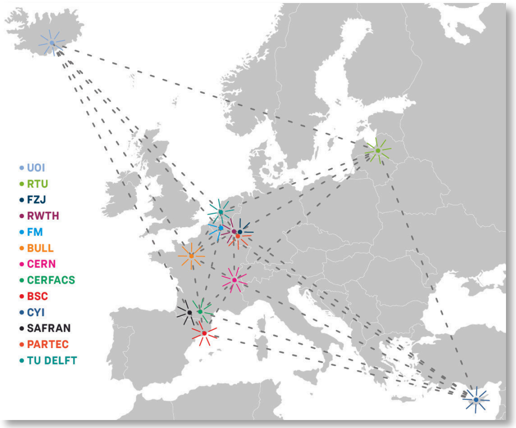
Collaboration promoting areas of work

See talk from Eric: https://indico.cern.ch/event/1392433

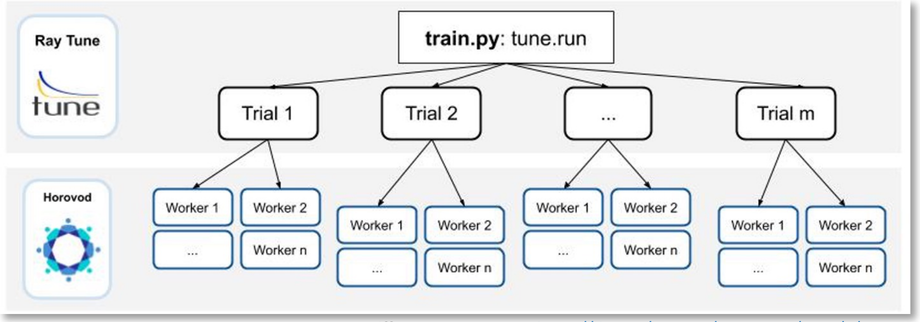
CoE RAISE

- CoE RAISE: Center of Excellence for Research on AI- and Simulation-Based Engineering at Exascale
    - Develops novel, scalable AI technologies along a wide range of scientific use-cases
- CERN leads WP4 on *Data-Driven Use-Cases towards Exascale* (lead by Dr. Maria Girone)
    - Task 4.1 on *Event reconstruction and classification at the HL-LHC* (lead by Eric Wulff)

CoE RAISE Partners

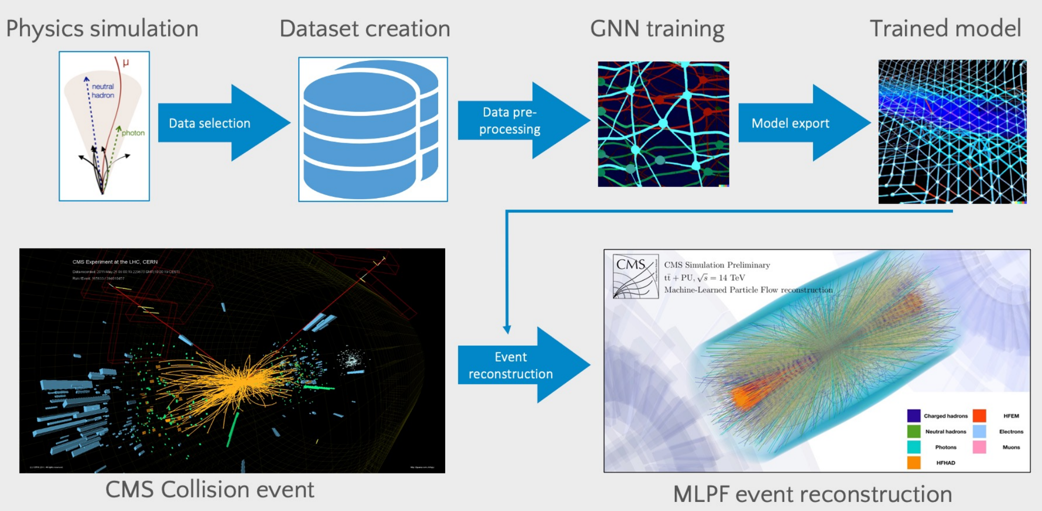


Large-scale distributed Hyperparameter Optimization on HPC



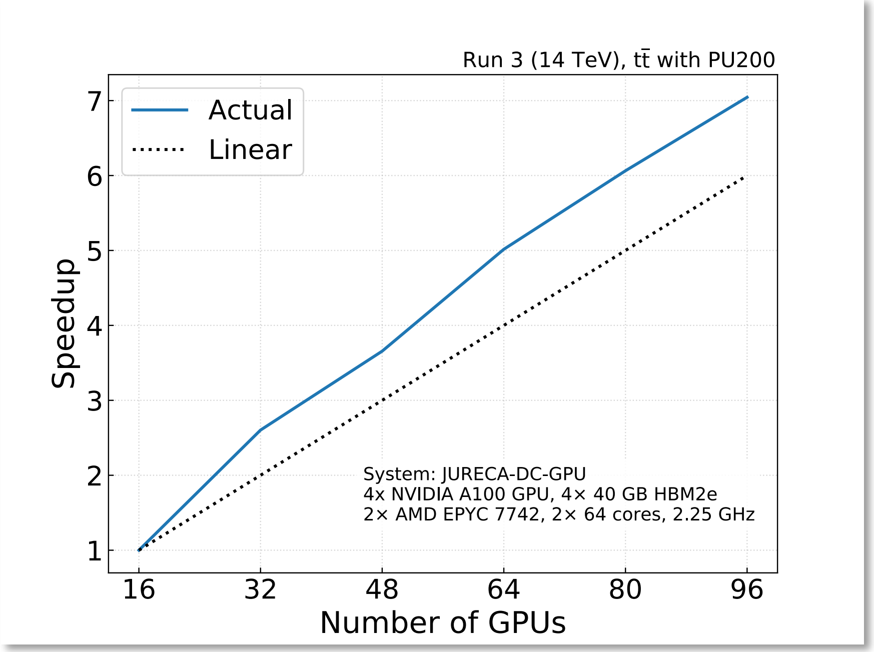
E.Wulff, M. Girone, J. Pata https://doi.org/10.1088/1742-6596/2438/1/012092

Deep Learning-based particle flow reconstruction workflow



Pata, J., Duarte, J., Mokhtar, F., Wulff, E., Yoo, J., Vlimant, J.-R., Pierini, M., Girone, M. (2022). *Machine Learning for Particle Flow Reconstruction at CMS*. Retrieved from http://arxiv.org/abs/2203.00330

Scaling of Hyperparameter Optimization using ASHA and Bayesian Optimization

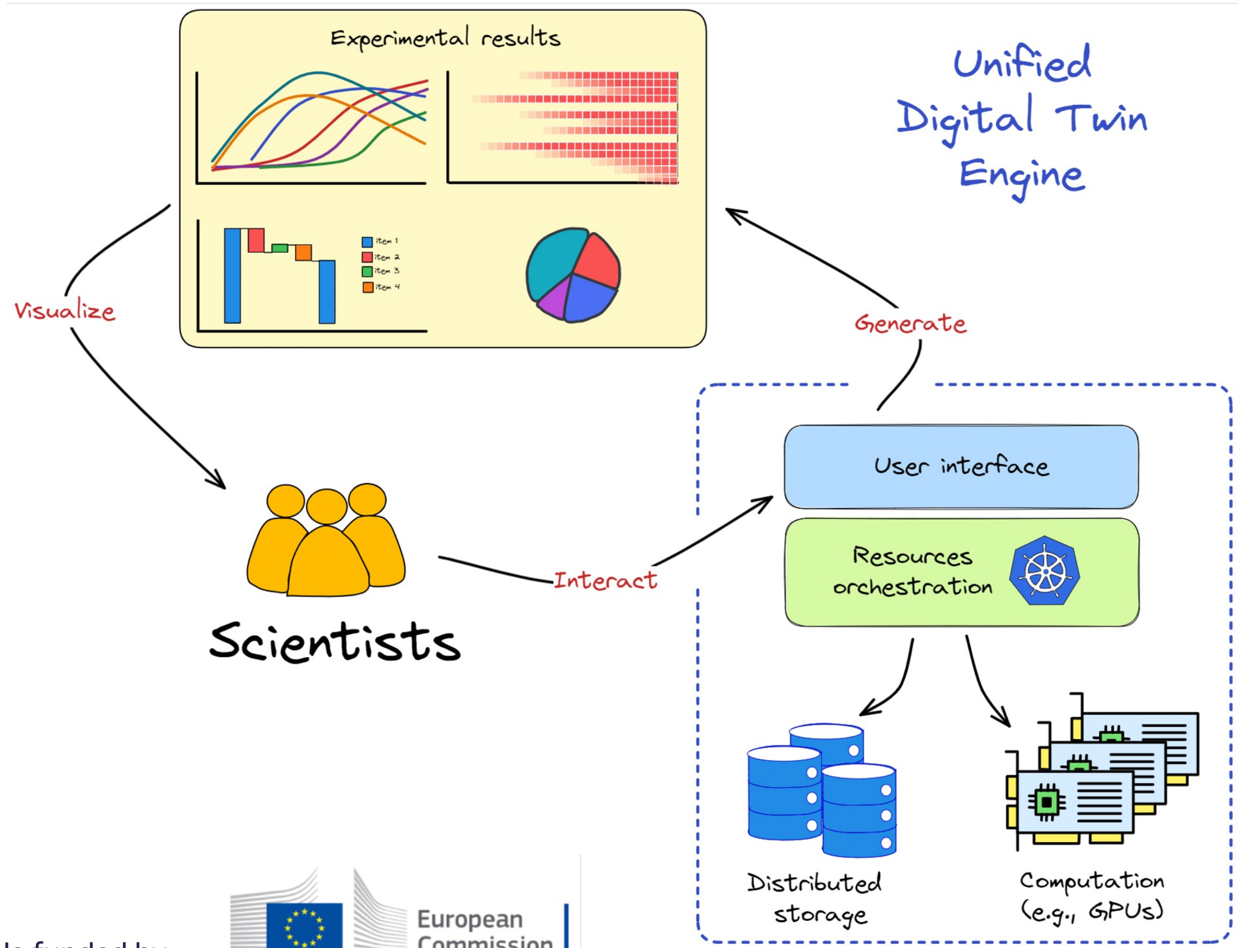


interTwin - Digital Twin Engine for science

Co-design and implement the prototype of an **interdisciplinary Digital Twin Engine**

Open-source platform based on **open standards**

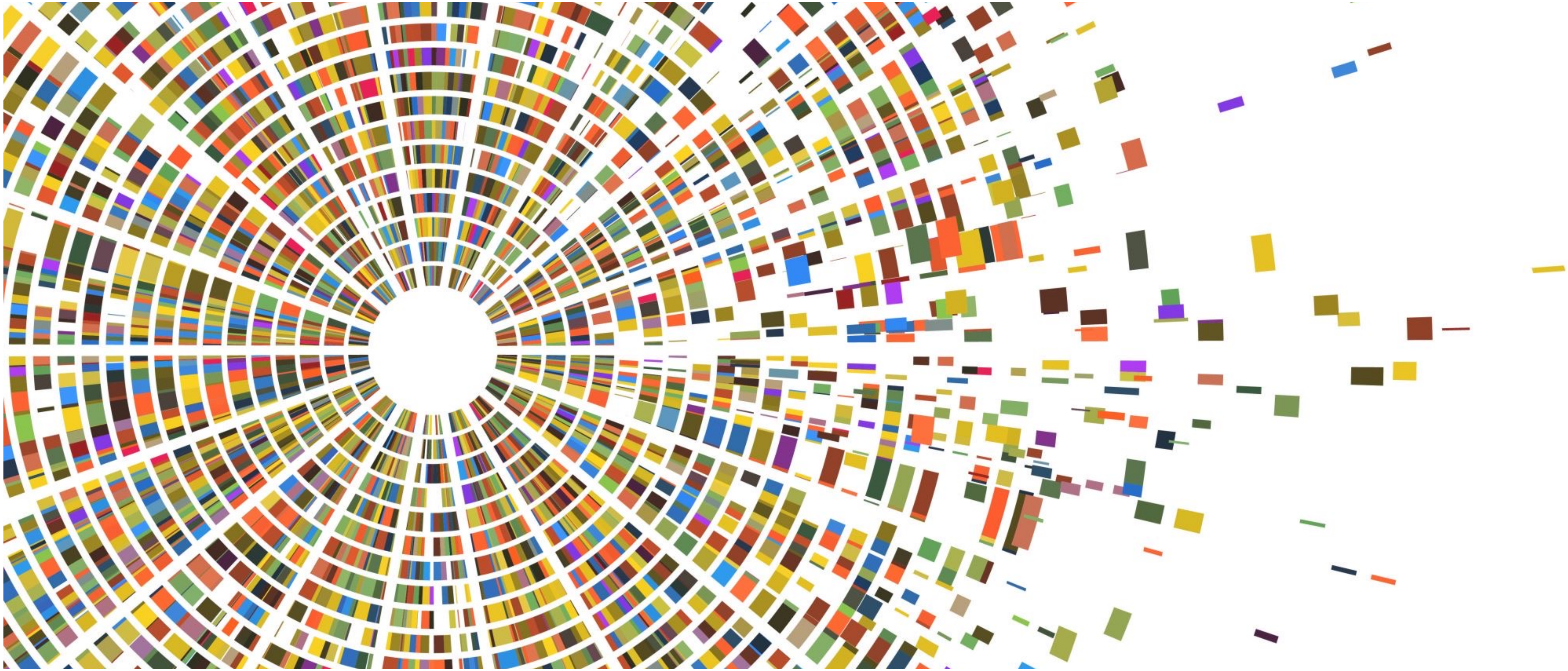
Large spectrum of **diverse use cases** from **physics** and **earth observation** sciences



interTwin Is funded by European Commission

https://indico.cern.ch/event/1392505

Benchmarking in HPC



Benchmarking and Accounting

Adopting HPC compute resources presents several new challenges beyond traditional x86 workload development:

- Diverse compute architectures (ARM, POWER, x86, RISC-V)
- Heterogenous accelerators (GPU, FPGA, Quantum*)

We must understand and account of all combinations of above to understand:

- Workload efficiency at runtime
- Efficiency of grant usage
- Mapping of users to resources

Benchmarking is used at CERN for:

- Efficiency
- Error detection
- Accounting
- Pledges
- Procurement

Contact with Industry KEY in this area of work

HPC Benchmarking

HEP Benchmarking Suite: The next generation of benchmarking for the WLCG , replacing HEPspec06 (over 15+ years use).

Historically benchmarking has been:

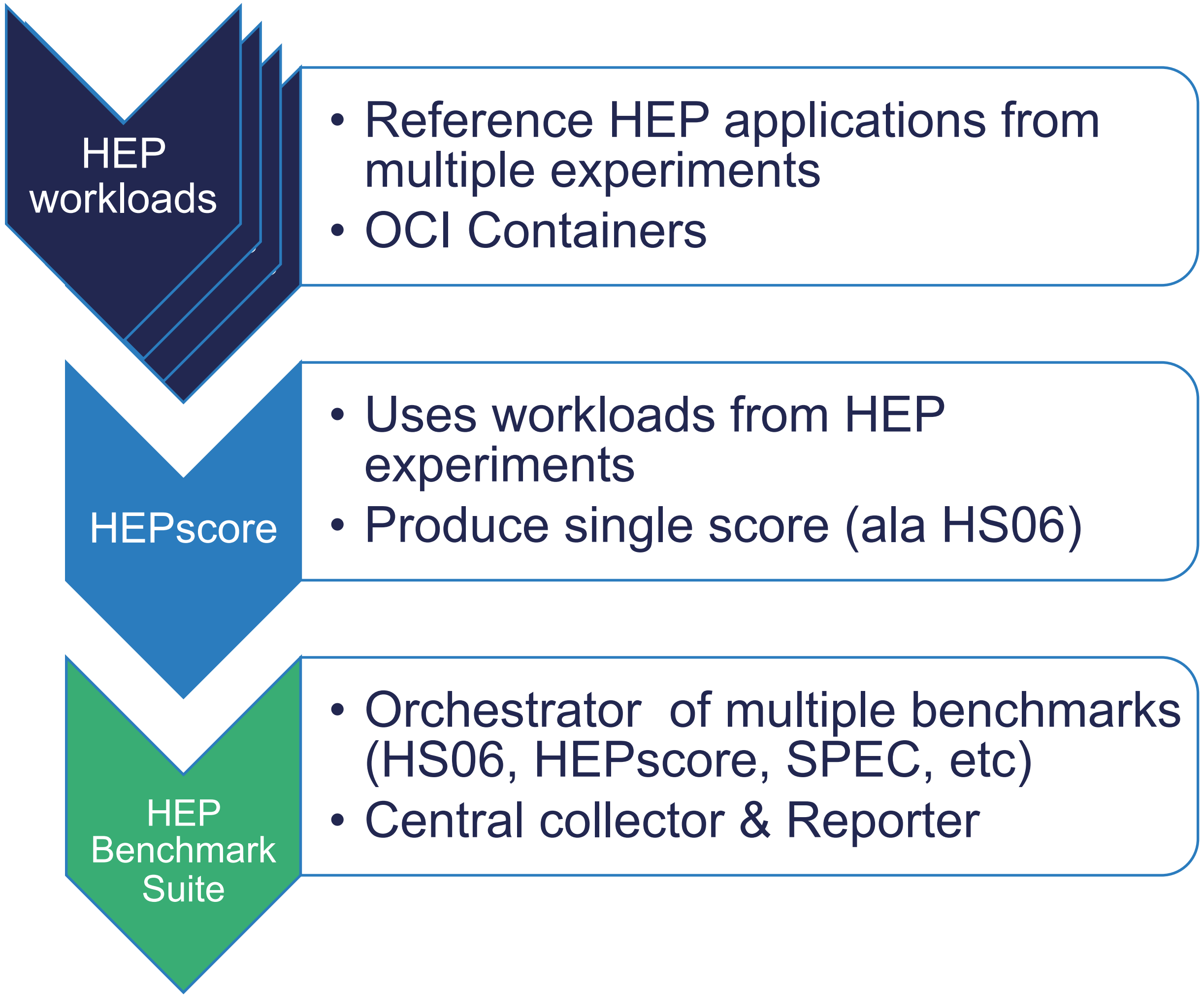
- Designed for WLCG compute environment
- Intended for procurement teams, site administrators
- First with VM containment, later nested docker images

None of these approaches are compatible with HPC!

- Refactor & re-tool for user execution at scale

- HEPscore ratified in April 2023 by the WLCG HEPscore Deployment Task Force as a replacement for HEPSPEC06

- https://w3.hepix.org/benchmarking.html



HEP Benchmark Suite

Minimal Dependencies
Python3 + container choice

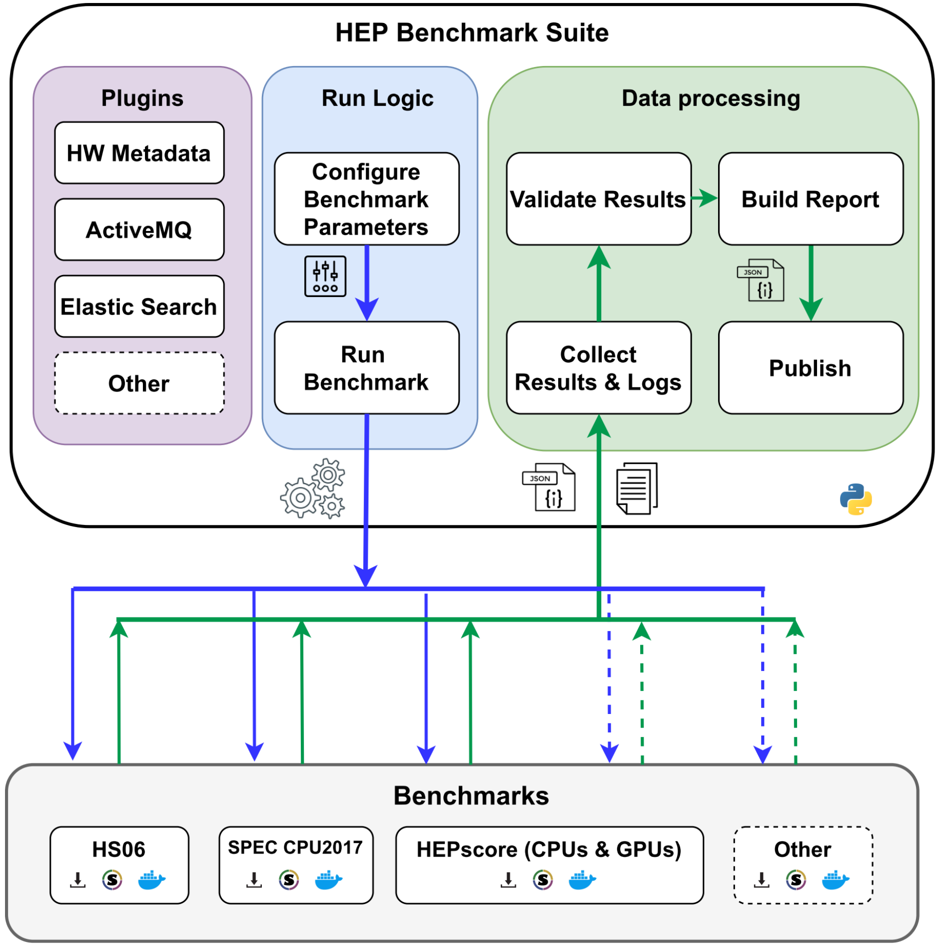
Modular Design
Snap-in workloads & modules

Repeatable & Verifiable
Declarative YAML config

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

Variety of containment choices
Singularity (incl. CVMFS Unpacked), Docker, Podman

Metadata + Analytics
Automated Reporting via AMQ



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

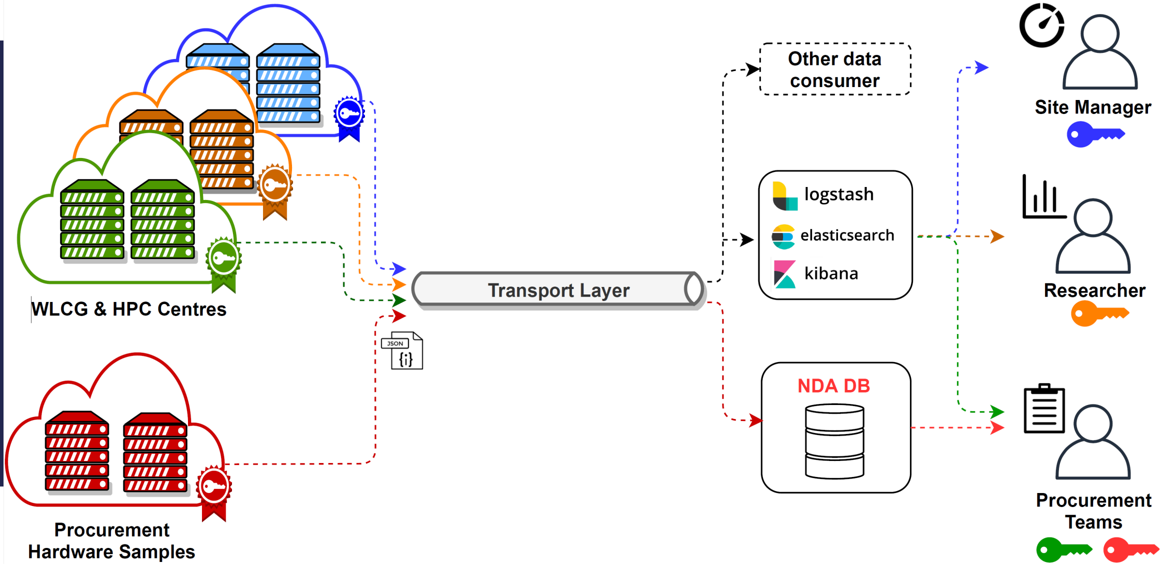
Automated HPC execution

Benchmarking Heterogeneous architectures

- Multi-arch as workloads become available (ARM, IBM Power …)
- GPU accelerators (Madgraph5, MLPF)

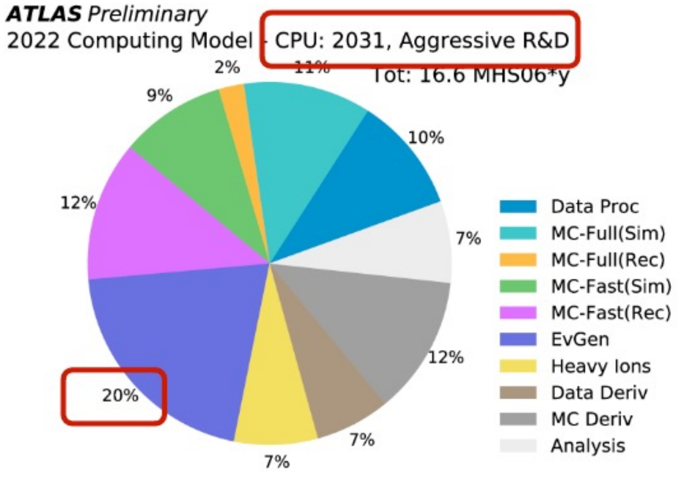
Simple integration with SLRUM, other job orchestrators

```
# HEP suite requires singularity/apptainer 3.5.3+, python3.
module load singularity python3

export RUNDIR=/tmp/HEP

echo "Running HEP Benchmark Suite on $SLURM_CPUS_ON_NODE Cores"
mkdir -p $RUNDIR
python3 -m pip install git+https://gitlab.cern.ch/hep-benchmarks/hep-
benchmark-suite.git

# Run suite
srun $HOME/.local/bin/bmkrun --config default --rundir $RUNDIR
```



Heterogeneous Benchmarking

- Combination of General-Purpose GPUs (GPGPU) and alternatives architectures targeted by experiments for Run 4
- GPU benchmarks for production workloads that operate on GPGPU and CPU+GPGPU
- ARM workloads
- MadGraph event generation for GPU and Vector CPUs
- Integration of non-x86 workloads into HEPscore



CERN-LHCC-2022-005

Event generation speedup, Nvidia A100

Process	Madevent 262 144 events			Standalone CUDA
	Total	Momenta+unweight	Matrix elm	ME Throughput
$e^+e^- \to \mu^+\mu^-$	17.9 s	10.2 s	7.8 s	$1.9 \times 10^6 s^{-1}$
+CUDA Tesla A100	10.0 s	10.0 s	0.02s	$633.8 \times 10^6 s^{-1}$
	1.8 x	1.0 x	390 x	334 x
$gg \to t\bar{t}gg$	209.3 s	7.8 s	201.5 s	$2.8 \times 10^3 s^{-1}$
+CUDA Tesla A100	8.4 s	7.8 s	0.6 s	$758.9 \times 10^3 s^{-1}$
	24.9 x	1.0 x	336 x	271 x
$gg \to t\bar{t}ggg$	2507.6 s	12.2 s	2495.3 s	$1.1 \times 10^2 s^{-1}$
+CUDA Tesla A100	30.6 s	14.1 s	16.5 s	$170.7 \times 10^2 s^{-1}$
	82.0 x	0.9 x	151 x	155 x

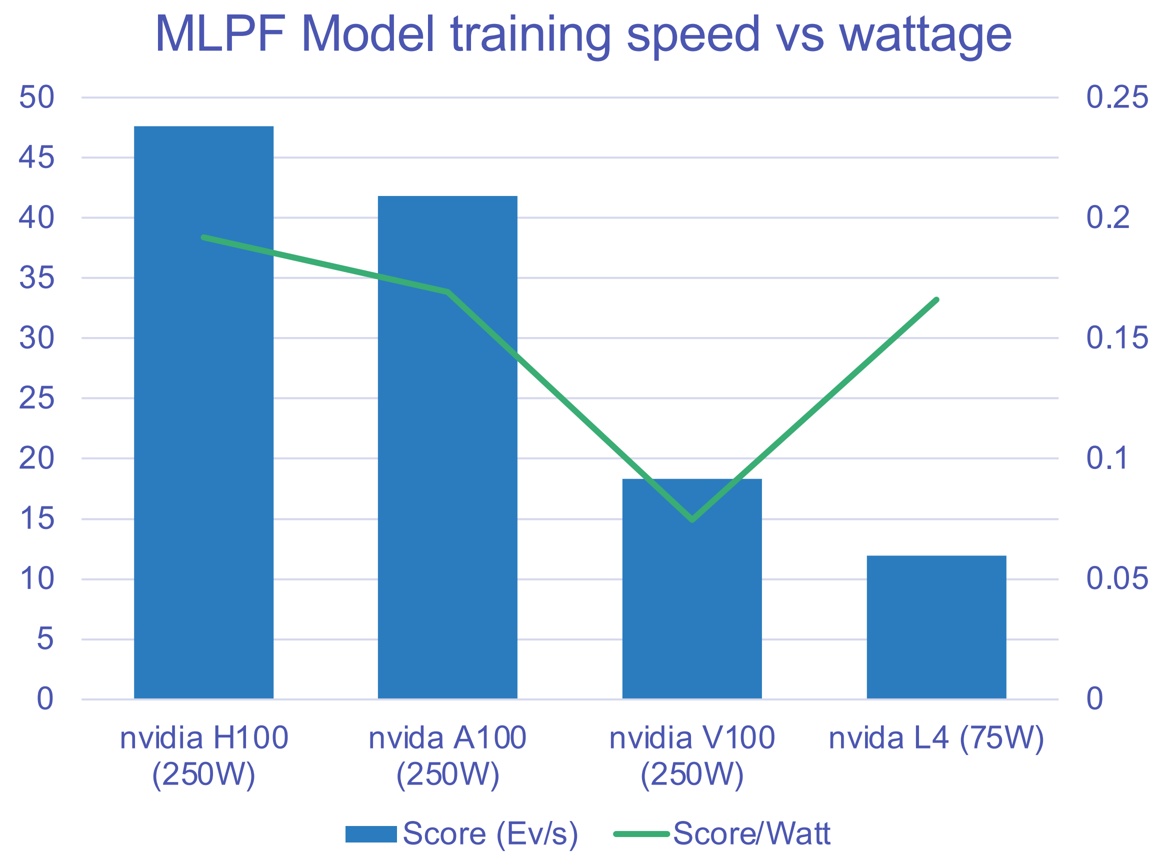
https://indico.jlab.org/event/459/contributions/11829/

ML/AI Benchmarking

Machine-learned particle-flow reconstruction algorithms (MLPF)

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)



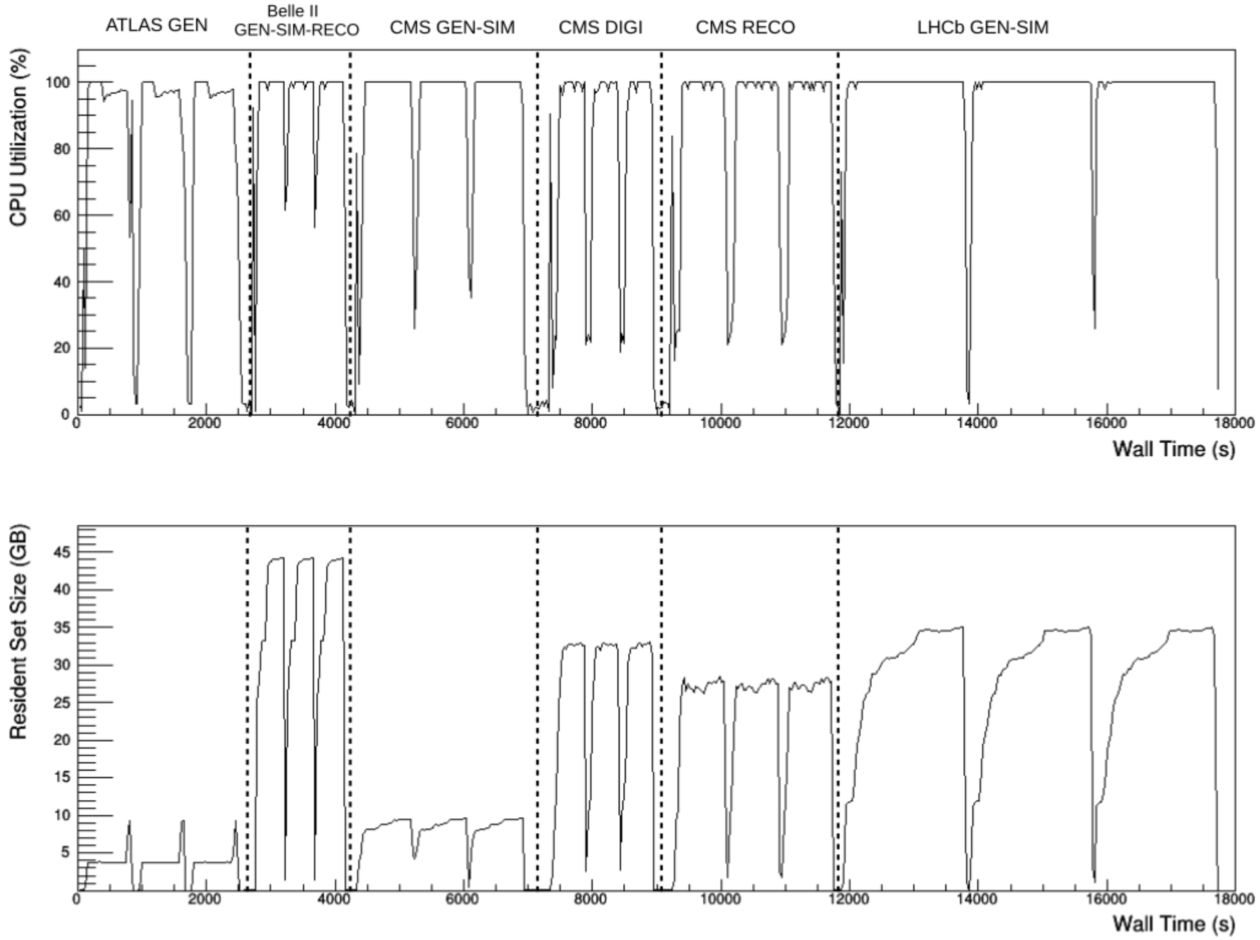
Understanding workload efficiency

Utilization at runtime is critical to benchmarking and production

- 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

Efficiency profile of typical HEPscore benchmark

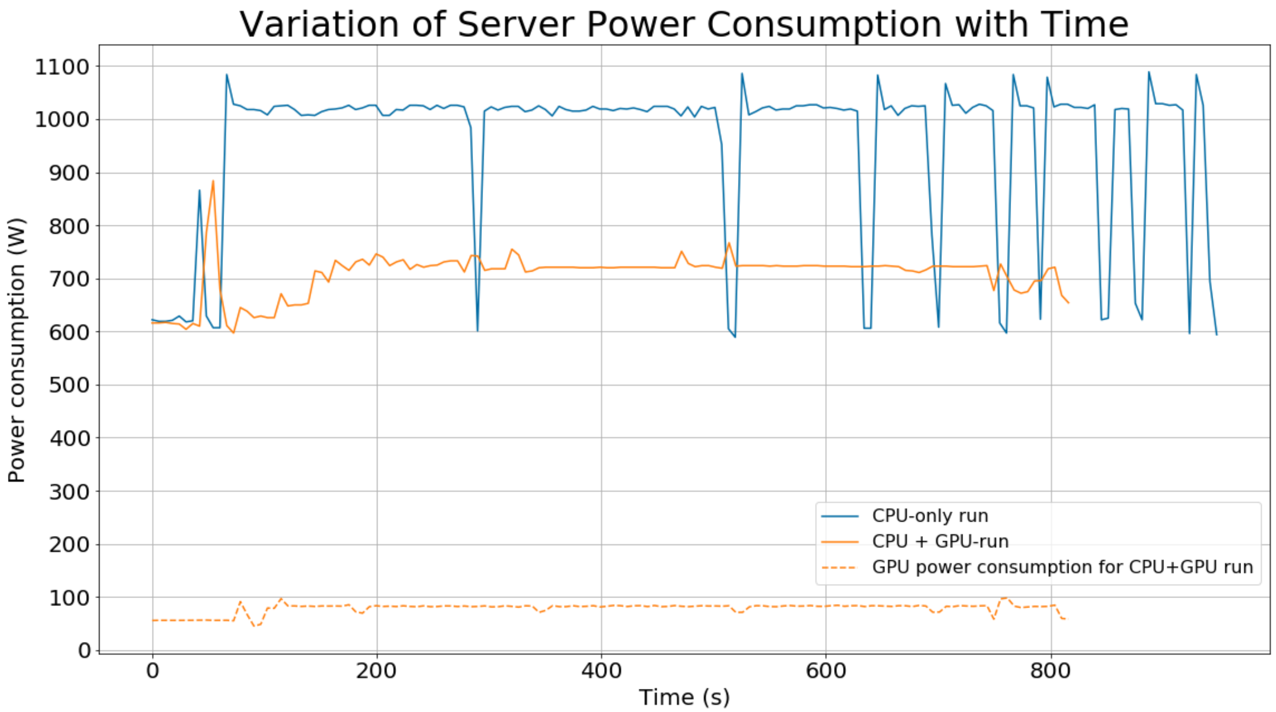


https://indico.cern.ch/event/1078853/contributions/4576275

Energy efficiency

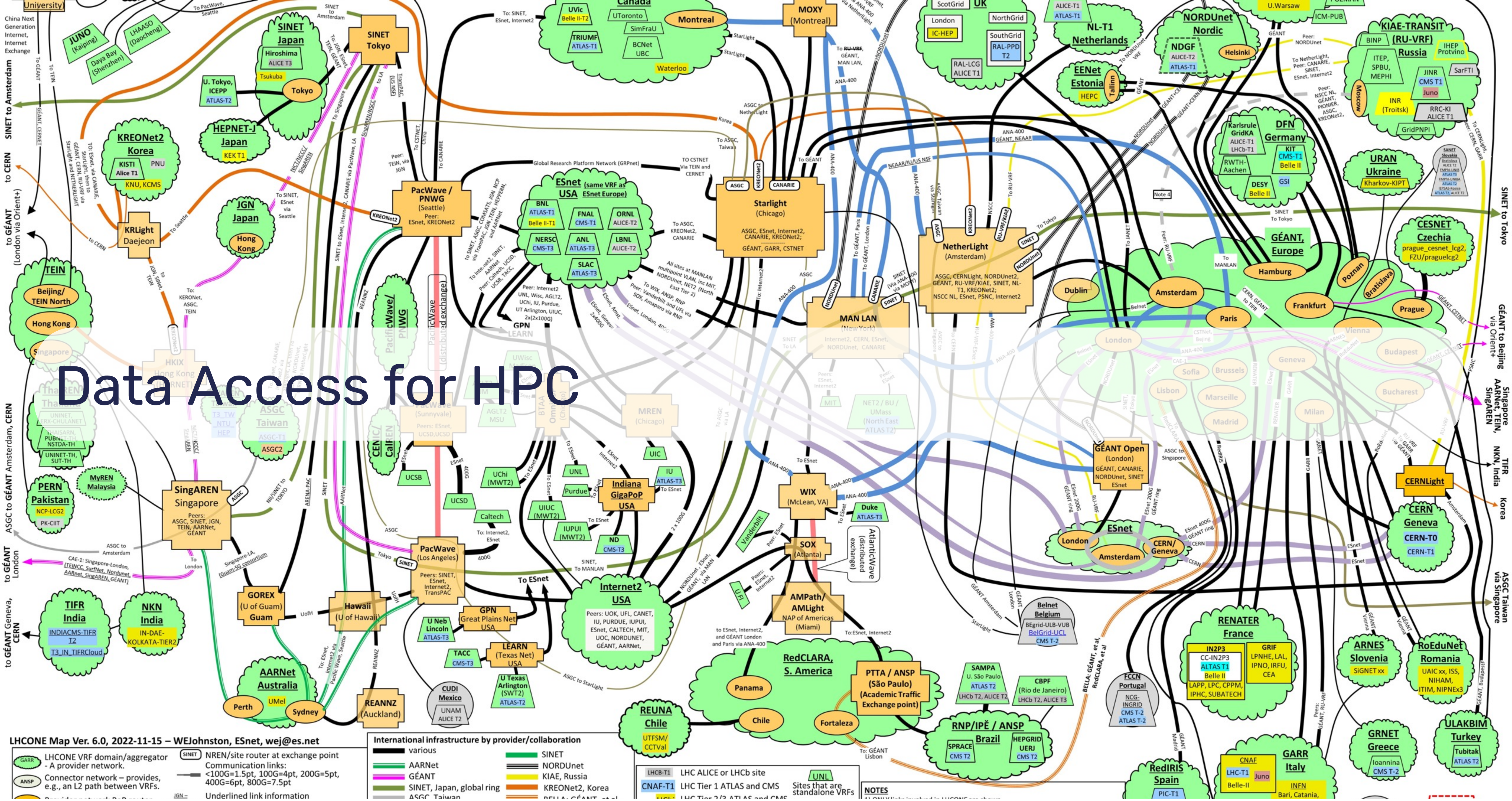
Energy efficiency is now included as a critical metric of performance

- Plugin to poll server power metrics (ipmi)
- Compare Nvidia-smi, ipmi & external metering
- BMK include energy metrics from CPU



K. Tuteja, openlab student program

Data Access for HPC



Storage

HPC storage is typically built from a common set of commercial building blocks.

Although standard, they are uniquely implemented at each site:

- Variable number of replications, metadata nodes, interconnect capabilities
- Little to no visibility into capabilities, usage, accounting, etc.

Lots of moving parts! Break it down into three general areas:

- Data ingress/egress from HPC centre
- Efficient usage of storage systems on site
- Dynamic scaling interaction between (1) and (2)

Some numbers

Initial HL-LHC models project **exabytes of data** production

HEP experiments will no longer be able to store all the produced data at a single site – it must be streamed in **~realtime.**

Structure HPC data challenge similar to WLCG Data Challenge:

HL-LHC goal to stream & process ~10 PB of physics data through a HPC site in a day:

- 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

HL-LHC Data Challenge

HPC Connectivity

Successfully exploiting opportunistic HPC allocation demands high connectivity for data-driven workloads. CERN current target **~5Tbps** connectivity by time of HL-LHC from CERN Tier0 to compute sites. WAN from HPC sites may be limiting factor for resource allocation without pre-placed data.

HPC Data challenge composed of EU Projects (CoE RAISE, InterTWIN), WLCG, and GÉANT to validate data-driven streaming and transfers

- Leverage GÉANT Data Transfer Nodes (DTNs) around EU for testing against backbone network
- Testing Unicore FTP (UFTP), FTS, Rucio for open science with HPC
- Currently exercising tests with Jülich, DE (200Gbps); SDSC, USA (400Gbps)

Shared filesystems

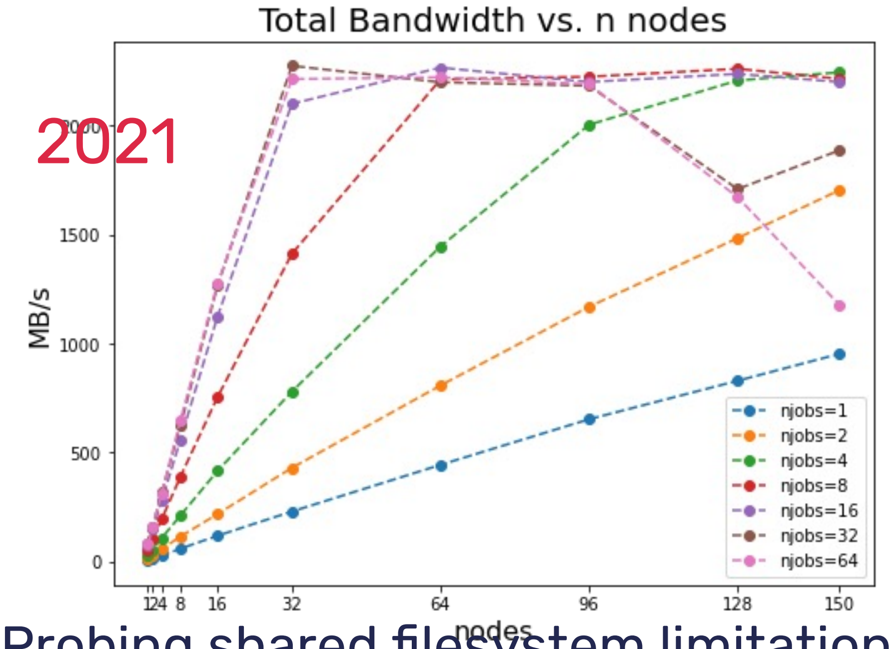
Traditional HPC workloads have low I/O demands – highly problematic running Big-Data workloads!

Compute-bound workloads dependent on shared file systems may be **effectively I/O bound** if scaled sufficiently

To avoid consuming a shared community resource, we need to understand what we can effectively scale to

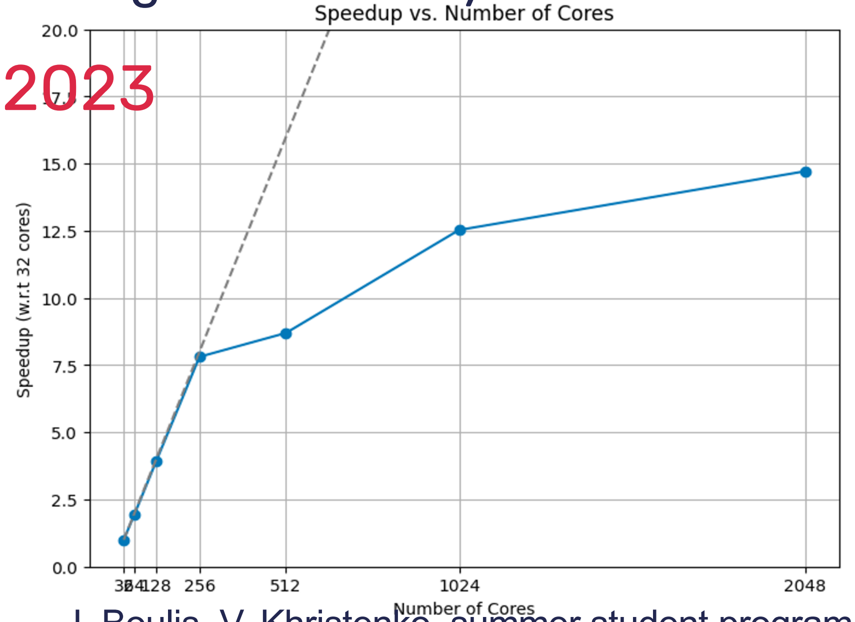
- Workload throughput O(100KB/s)-O(100MB/s)
- Many workloads per host

2021



Probing shared filesystem limitations

2023



J. Boulis, V. Khristenko, summer student program

Data formats

Data format drastically affects HPC storage efficiency:

- Writing data in storage format supporting parallel I/O
- Optimization: Tuning of parallel libraries to optimize the performance
- Adopting native object storage (HDF5) native to parallel IO
- Dramatically reduce random read during jobs

ROOT
Data Analysis Framework

Data Lakes

Separation of WLCG sites responsibilities to new “Data Lake” model for LHC data storage has introduced new standards and modernized capabilities. Leveraging better data access patterns to datasets with latency-hiding advancements of XrooD/Xcache greatly reduces data transfer requirements:

- RUCIO – a high level data management layer, coordinates file transfers over several protocols (HTTP/WebDAV, XrootD, S3, etc.)

- FENIX – Collaboration of HPC sites and ESCAPE to standardize data transfers

Authentication & Authorization

HPC and Authentication

HPC sites operate differently regarding account creation and access policies from from traditional CERN Grid:

- Varying levels of trust requirements
- Authentication methods (SSH, Certificate, tokens..)
- Not reasonable to expect importation/trust of CERN computing accounts (16k+)

AAI Transformation

WLCG transition from certificate-based authorization to token-based carries through into HPC .

Among several components of the ESCAPE project, AAI aims to bridge CERN AAI to HPC

- OIDC-token Authentication migration from X.509 Certificate – faster, easier for institutional trust
- Federated login AuthN/AuthZ for HPC via EduGAIN federation/Puhuri

ESCAPE IAM has been integrated into the EOSC AAI federation in collaboration with GÉANT,

EuroHyPerCon

ESCAPE

European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

ESCAPE project completed Summer 2022 after 42 months

Outlook

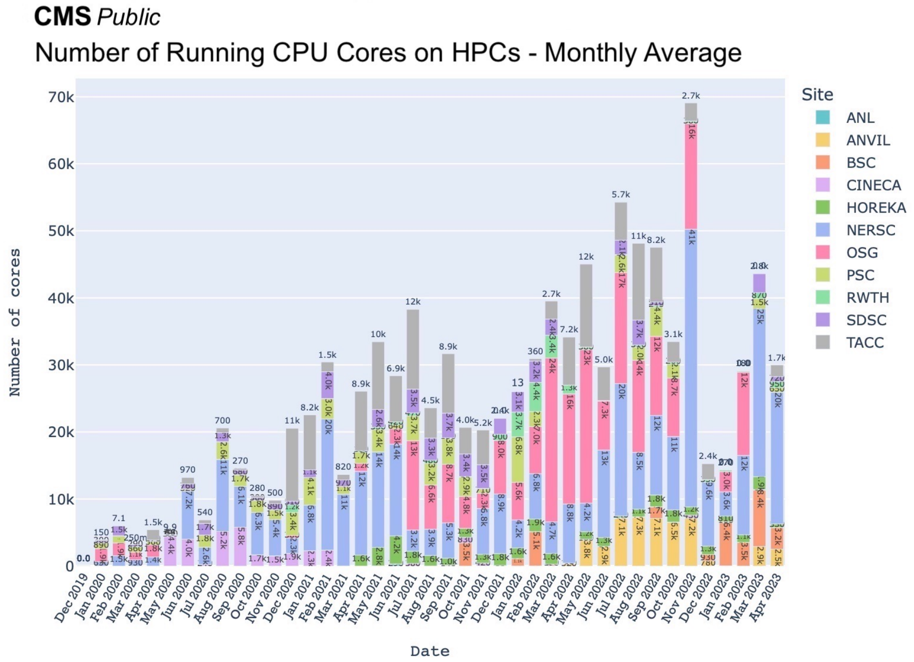
Ramping up

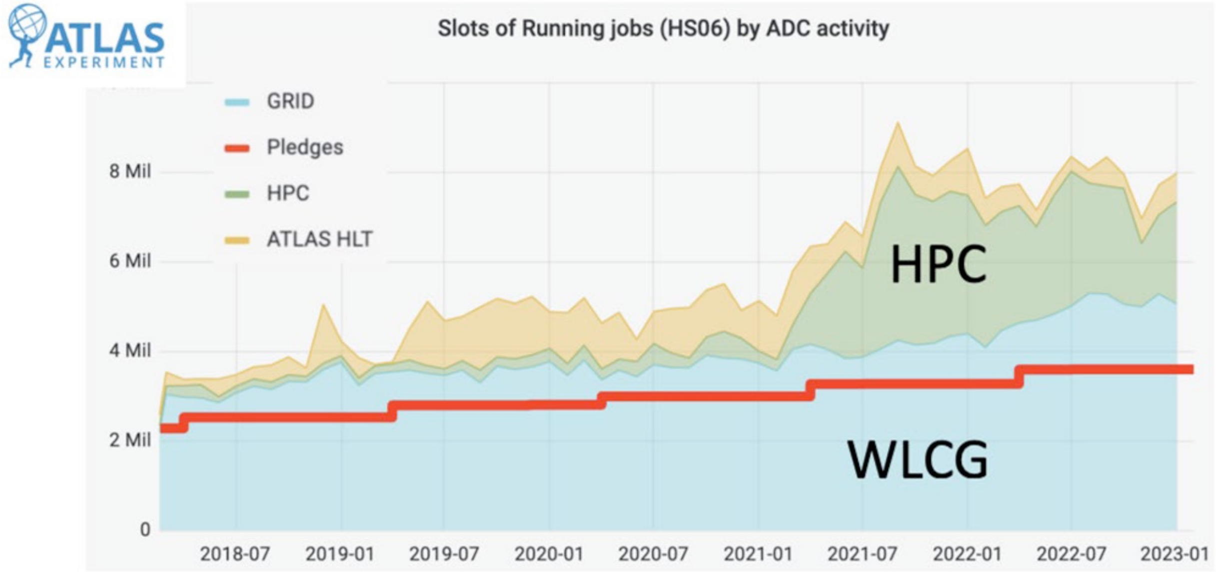
A complex problem with many moving parts – All feasible methods to close the computing gap are being pursued

- Including HPC!

Substantial technical investment, both for production and development in past years

HEP and Big-Data sciences can leverage potentially large benefits by exploiting HPCs



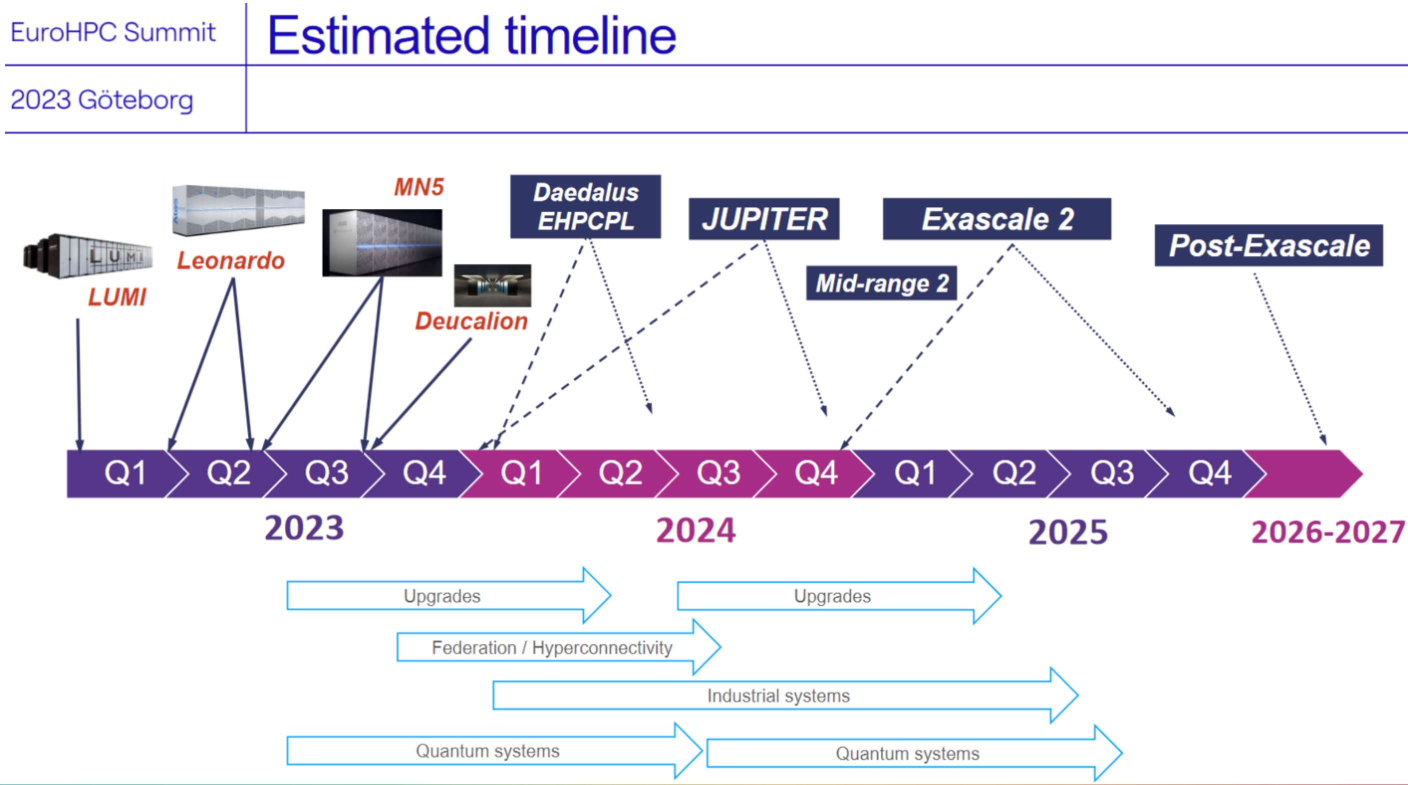


HPC is preparing for Big Data

European Pilot for Exascale

HPC communities (including HEP) inform future system design, drive convergence

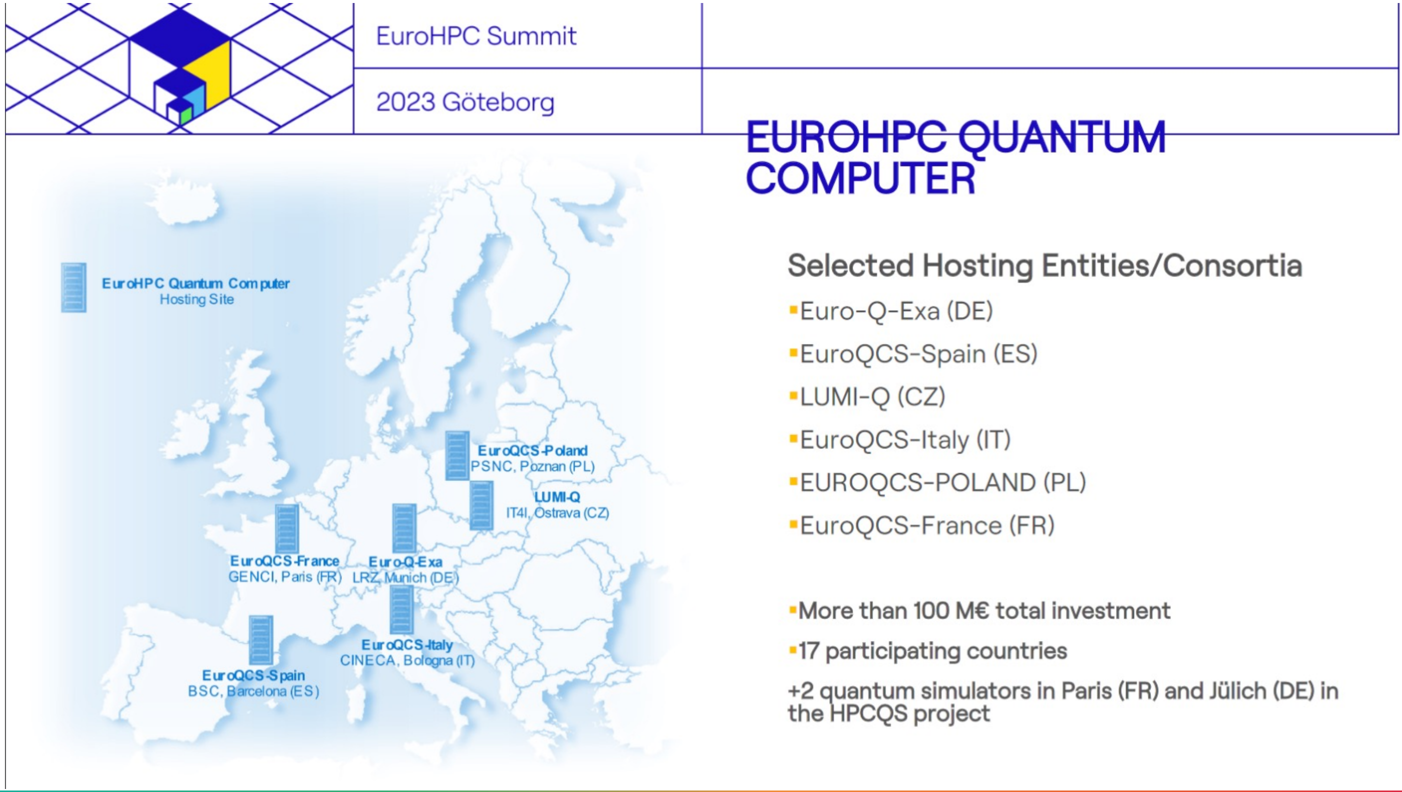
- EuroHPC call for tender for federation of hpc and quantum computers
- HPC roadmap for big-data workloads
- JUPITER procurement complete, 24' install
- HPC <-> Cloud connectors
- Upgrading WAN connectivity



Quantum + HPC

- HPC essential for quantum computing, massive computing needs for simulation & analysis research
- 2 quantum simulator sites (100+qubits each) at GENCI(FR), JSC(DE)
- 6 sites selected to host first European quantum computers

QUANTUM TECHNOLOGY INITIATIVE



What is SPECTRUM?

A project granted under the call HORIZON-INFRA-2023-DEV-01-05, which aims to prepare a Computing Strategy for Data-intensive Science Infrastructures in Europe for the High Energy Physics (HEP) and Radio Astronomy (RA) domains

Who is part of SPECTRUM?

SPECTRUM gathers selected stakeholders in the HEP and RA research domains, and at the same time experts from the e-infrastructures (HPC, Clouds, Quantum Computing). The former group brings **directions and future needs**, the latter bring **expectations** for new e-infrastructures about technical and policy aspects.

Expected outcomes

The realisation of a **Community of Practice (SPECTRUMCoP)** to gather and inform about future directions and needs in data-intensive research on the one side, and about future e-infrastructures on the other

A **Strategic Research, Innovation and Deployment Agenda (SRIDA)** and a Technical Blueprint about agreed processing models and solutions, to provide feedback on investment to funding agencies and policy makers

Why is SPECTRUM different from previous attempts?

Previous interactions between the research and e-infrastructure communities have been **a posteriori**, attempting to adapt scientific workflows to already operational facilities. This has been only partially successful due to technical (non-compliant system architectures, …) and policy (user access, …) incompatibilities.

SPECTRUM wants to move the handshaking process **a priori**, before the e-infrastructures are designed and deployed

https://www.spectrumproject.eu

Remaining Challenges

Much effort has been invested into HPC adoption in the past years, but challenges remain:

- Integrating independent machines as single entities (time/effort intensive)
- No common framework for Access/Usage policies, services, machine-lifetime (SPECTRUM will help!)
- Software deployment, edge services for data and workflow management
- Workflow/job orchestration – integration with data locality tracking, HTcondor, etc
  - e.g. “opportunistic” Data ingress/egress based on locality, compute resource & time constraints

Moving towards a common HPC interface

Addressing all HPC sites from an integrated platform

- Enable elastically expanding the resources available to big data sciences
- Interoperability of solutions in federated environment

Thank you!

CERN openlab

FREE ACCESS TO EUROHPC SUPERCOMPUTERS

WHO IS ELIGIBLE?

- Academic and research institutions (public and private)
- Public sector organisations
- Industrial enterprises and SMEs

→ Open to all fields of research

WHICH TYPES OF ACCESS EXIST?

- Regular access
- Extreme scale access
- Benchmark access
- Special access

Regular and extreme scale access calls are continuously open, with several cut-offs throughout the year triggering the evaluation of proposals.

WHAT ARE THE CONDITIONS FOR ACCESS?

Access is free of charge. Participation conditions depend on the specific access call that a research group has applied to.

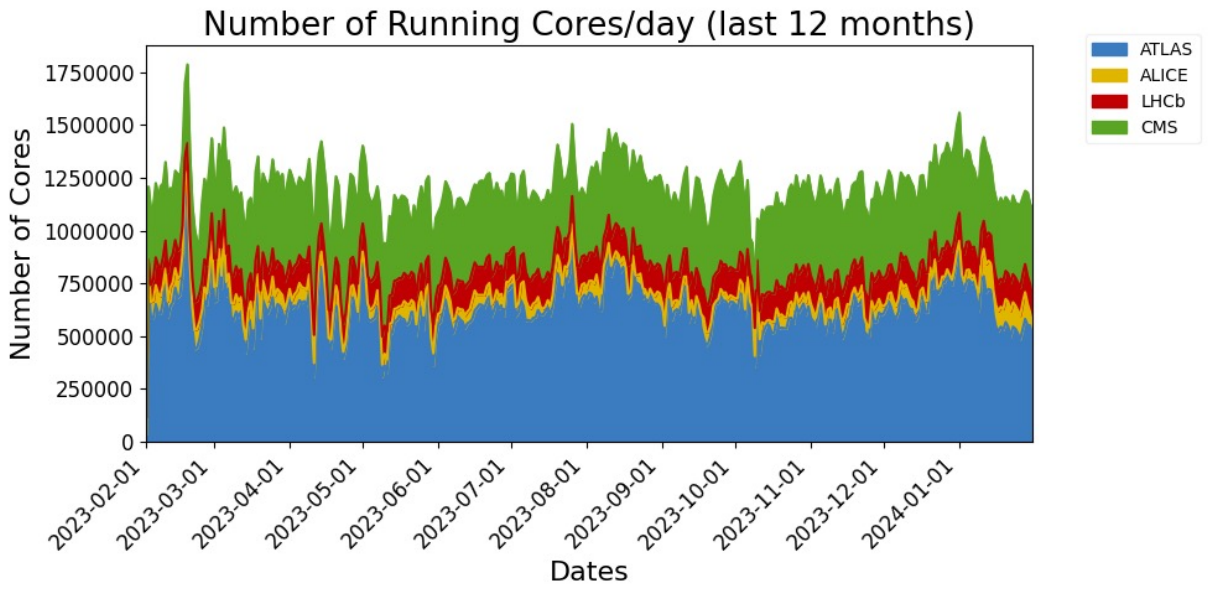
In general users of EuroHPC systems commit to:

- acknowledge the use of the resources in their related publications
- contribute to dissemination events
- produce and submit a report after completion of a resource allocation

More information on EuroHPC access calls available at: https://eurohpc-ju.europa.eu/participate/calls_en

Apples to ?

- 307 kHS06 avg by HPC first 7 mo. 2021 -> largest of CMS Tier-1 (FNAL) pledged 260kHS06 for 2020.
- Next generation of HPC machines (exascale) will provide more computational power than all WLCG sites combined



https://wlcg.web.cern.ch/using-wlcg/monitoring-visualisation/monthly-stats

Job Provisioning

SLURM scheduler used by HPC sites not immediately compatible with HTcondor

SLURM – push only, BATCH pull (pilot jobs)

Two ongoing efforts to extend batch schedulers to HPC:

- Extending HTCondor service (tested on connectivity-restricted sites)
- Dask + slurm plugin for submission/translation

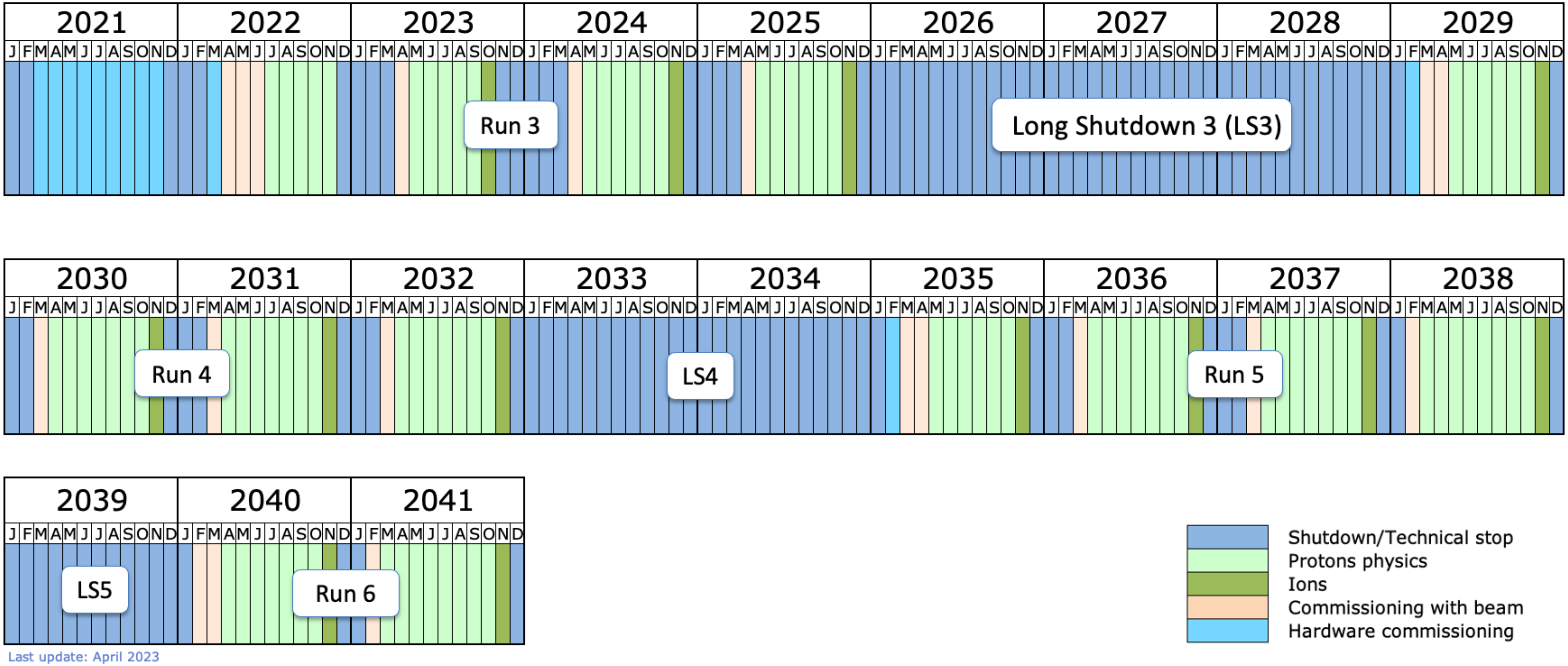
Portable frameworks

Experiments exploring several frameworks/languages to leverage heterogeneous compute

- Avoid vendor lock-in
- Leverage open source dev

	CUDA	Kokkos	SYCL	HIP	OpenMP	alpaka	std::par
NVIDIA GPU			intel/llvm compute-cpp	hipcc	nvc++ LLVM, Cray GCC, XL		nvc++
AMD GPU			openSYCL intel/llvm	hipcc	AOMP LLVM Cray		
Intel GPU			oneAPI intel/llvm	CHIP-SPV: early prototype	Intel OneAPI compiler	prototype	oneapi::dpl
x86 CPU			oneAPI intel/llvm computecpp	via HIP-CPU Runtime	nvc++ LLVM, CCE, GCC, XL		
FPGA				via Xilinx Runtime	prototype compilers (OpenArc, Intel, etc.)	protytype via SYCL	

CHEP 2023 https://indico.jlab.org/event/459/contributions/11807



https://lhc-commissioning.web.cern.ch/schedule/LHC-long-term.htm