CERN openlab

HPC and CERN:
Challenges and Integration

David Southwick, Maria Girone, Eric Wulff, Matteo Bunino, Alexander Zoechbauer

In collaboration with

WLCG Benchmarking WG, ROOT WG

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)
- Exa-scale infrastructure

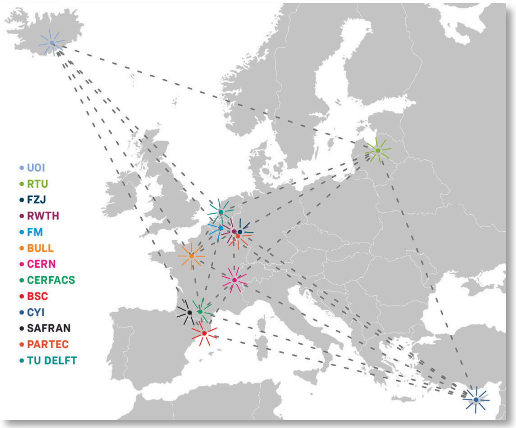
CERN Openlab partners with industry and collaborates with organizations to further mutual HPC adoption:

- 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

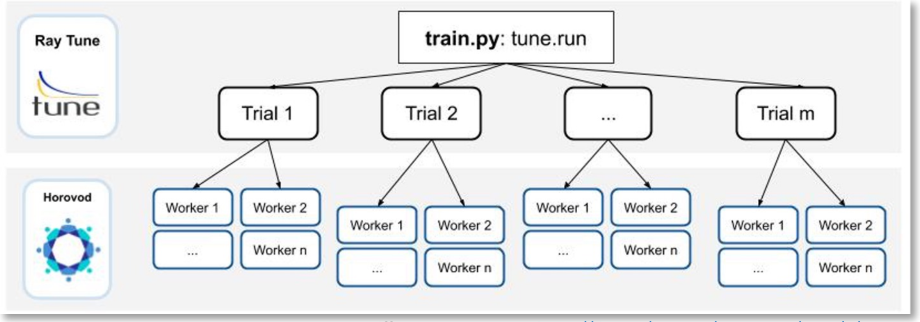
CoE RAISE

- 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

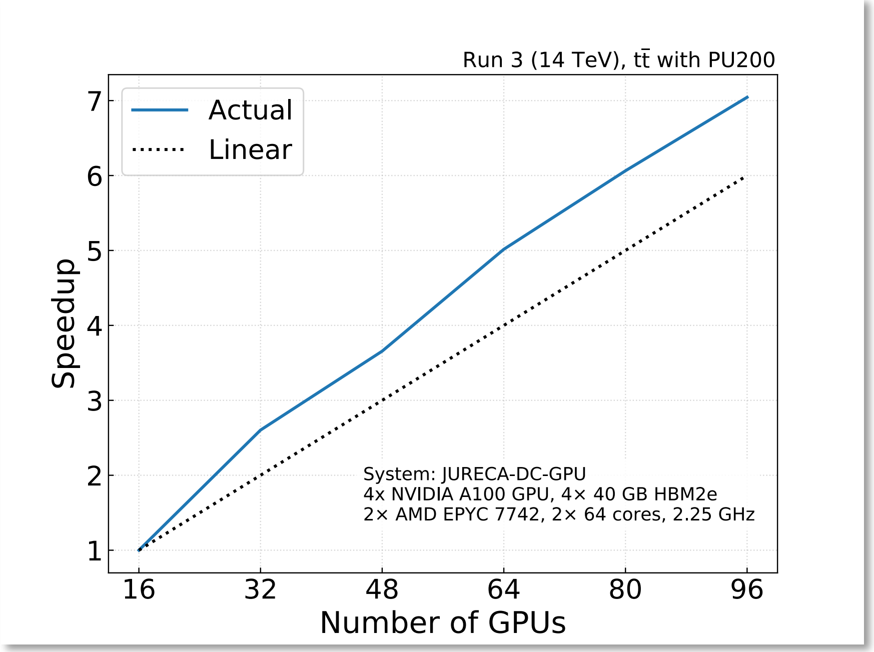


Large-scale distributed Hyperparameter Optimization on HPC

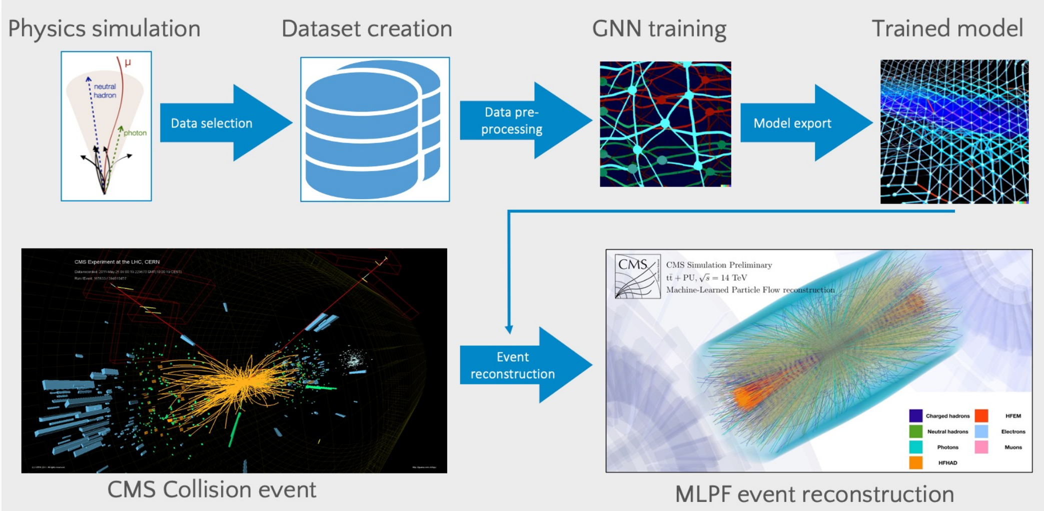


E.Wulff, M. Girone, J. Pata https://doi.org/10.1088/1742-6596/2438/1/012092

Scaling of Hyperparameter Optimization using ASHA and Bayesian Optimization



Deep Learning-based particle flow reconstruction workflow



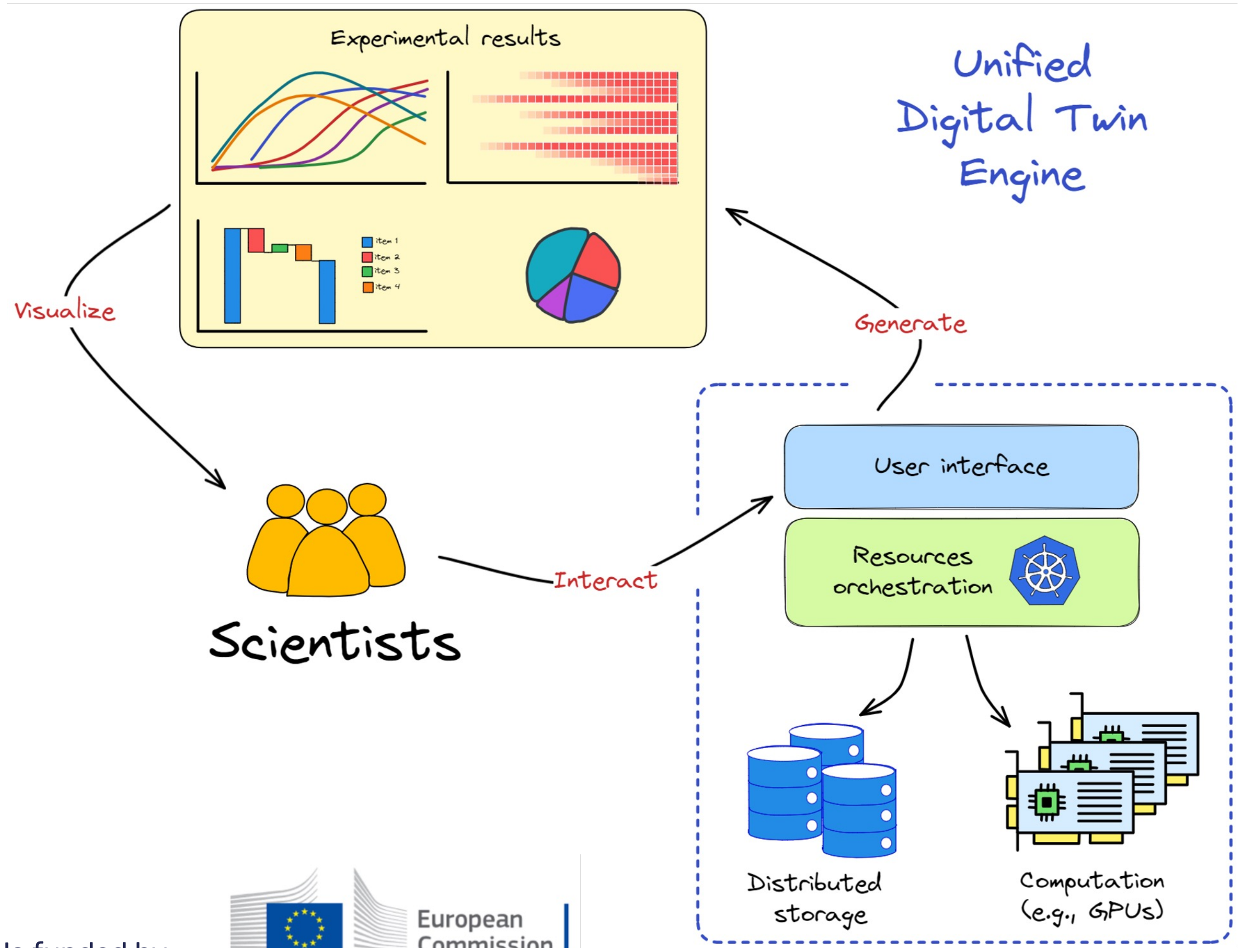
Pata, J., Duarte, J., Mokhtar, F., Wulff, E., Yoo, J., Vlimant, J.-R., Pierini, M., Girone, M. (2022). *Machine Learning for Particle Flow Reconstruction at CMS*. Retrieved from http://arxiv.org/abs/2203.00330

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

HPC adoption

Today, at most HPC sites, GPUs account for the majority of a site's total computing power.

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!

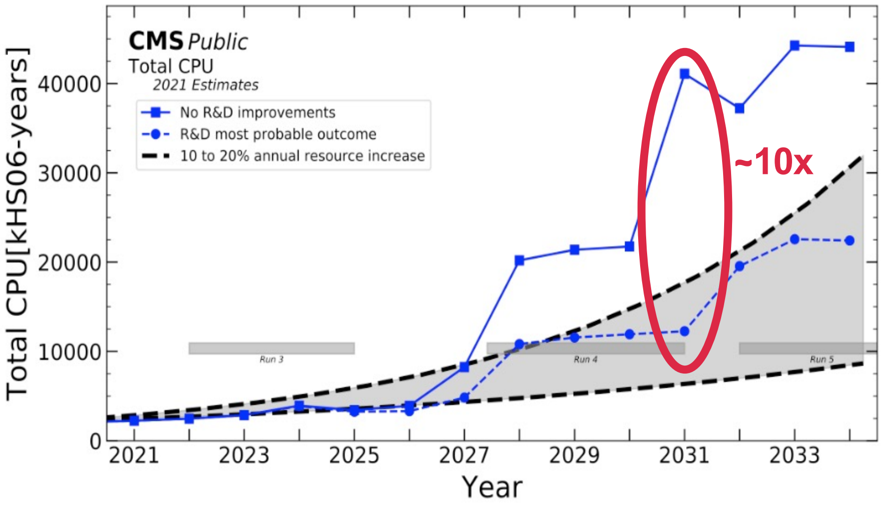
…but there is much more to HPC than only GPUs!

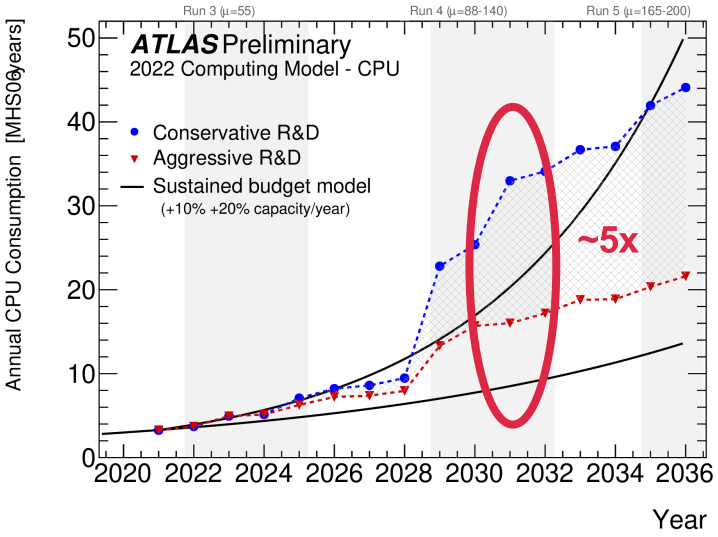
HEP Motivation

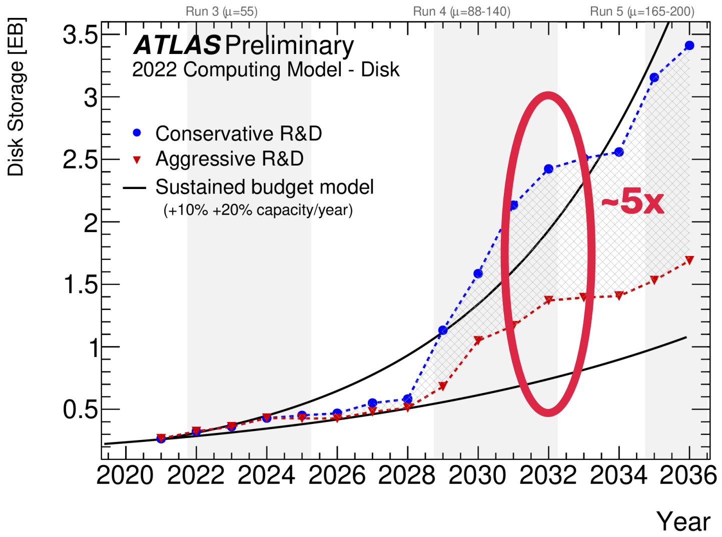
LHC expects more than exabyte of new data for each year of HL-LHC era from ~2029-2040.

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







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

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

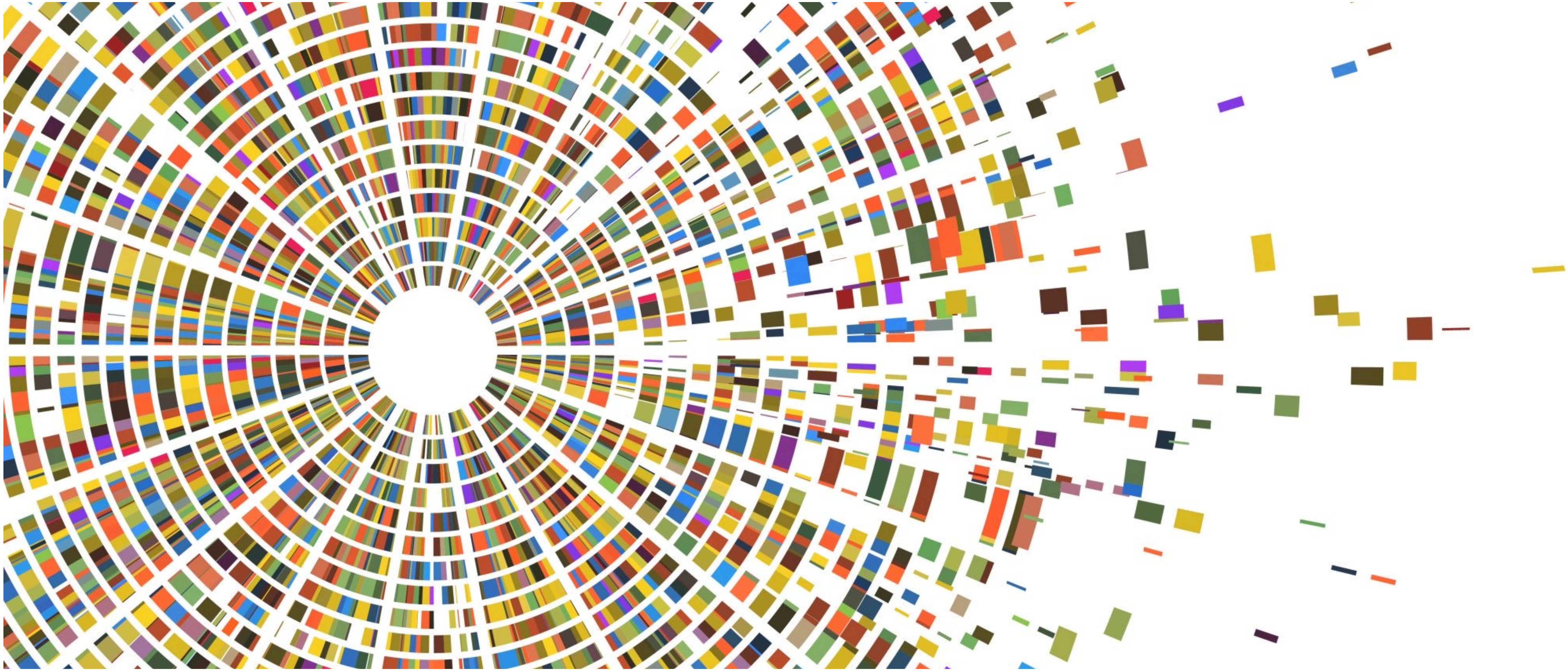
Outline

- Intro & motivation

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

- Software and Architectures
- Runtime Environments and Containers
- Provisioning

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:

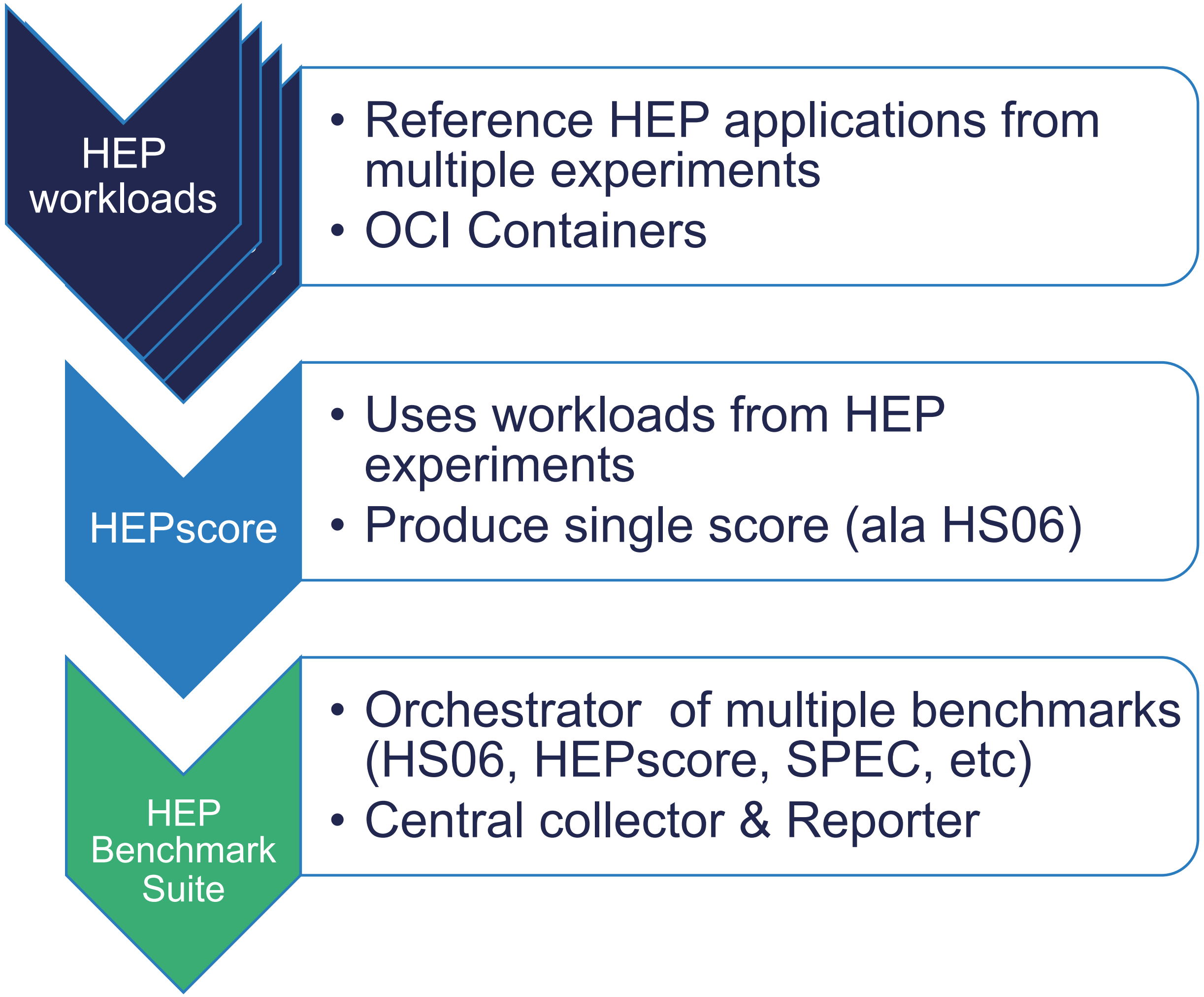
- Designed for WLCG compute environment
- Intended for procurement teams, site administrators
- First with VM containment, later nested docker images

None of these approaches are compatible with HPC!

- Refactor & re-tool for user execution at scale

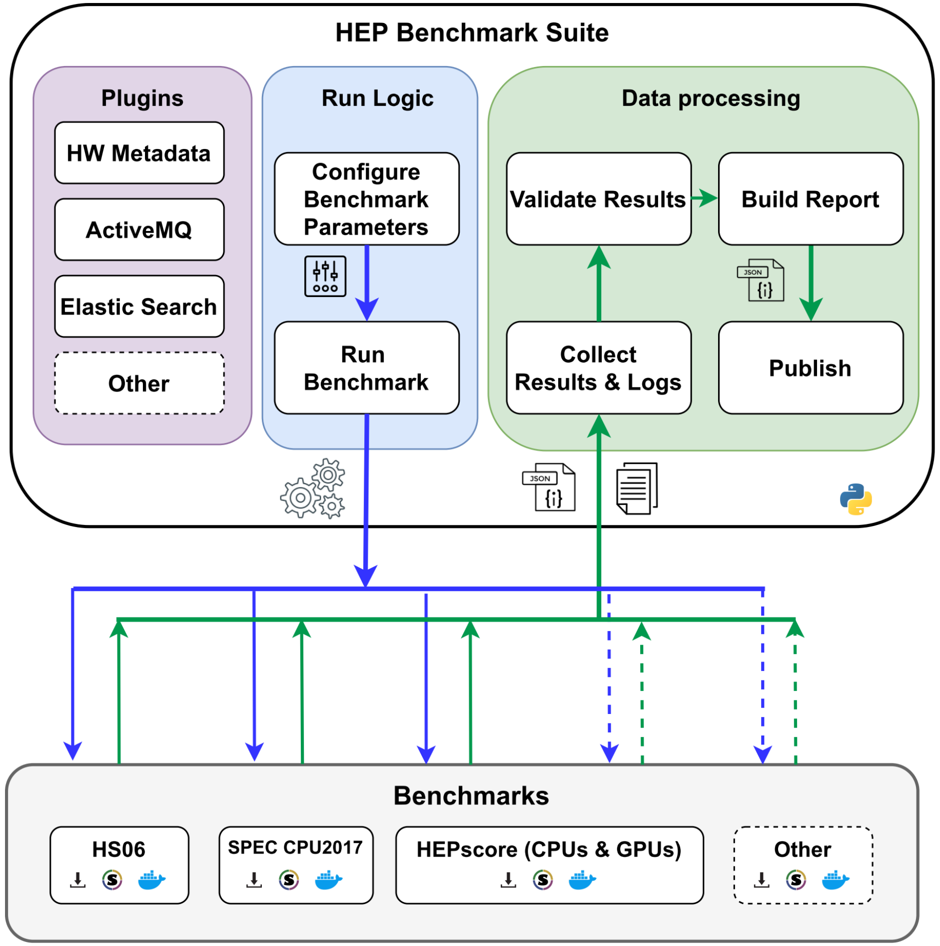
- HEPscore ratified in April 2023 by the [WLCG HEPscore Deployment Task Force](#) as a replacement for HEPSPEC06

- 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

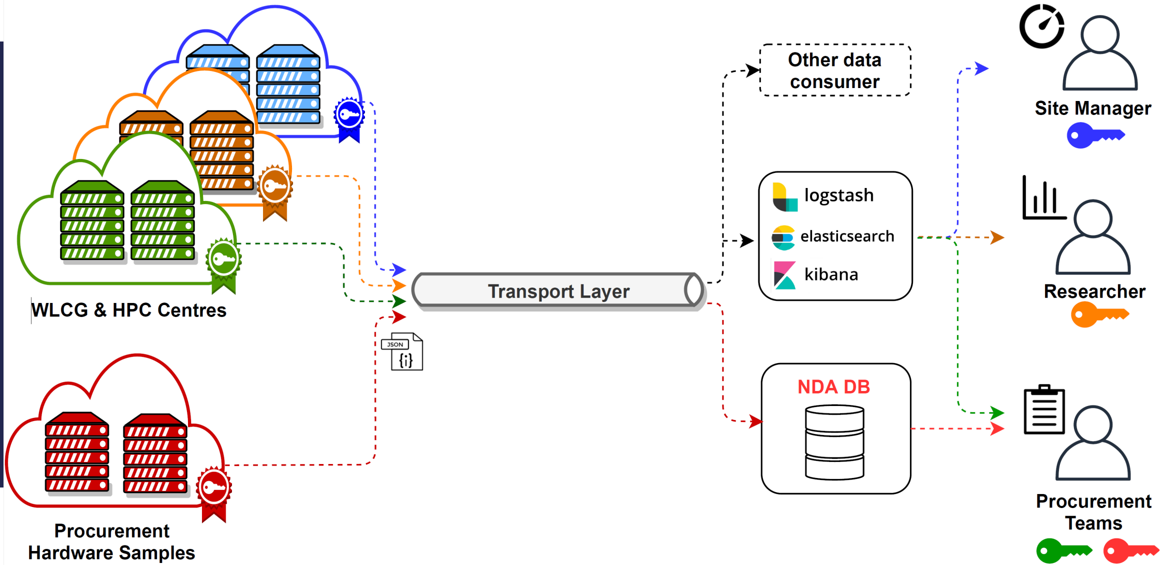
Automated HPC execution

Benchmarking Heterogeneous architectures

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

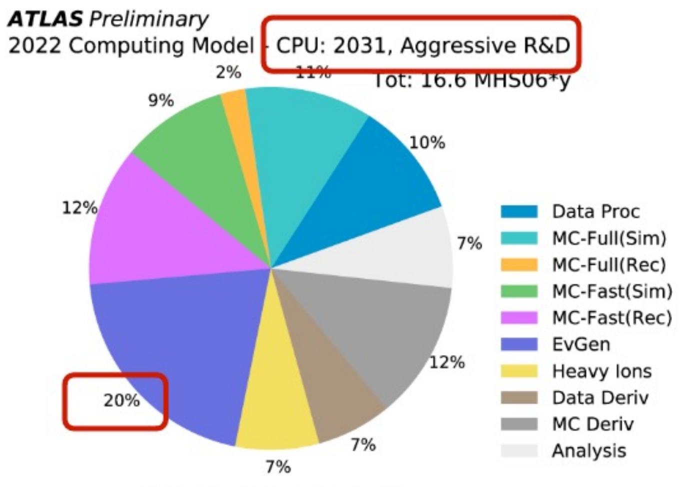
Simple integration with SLRUM, other job orchestrators

```
# 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

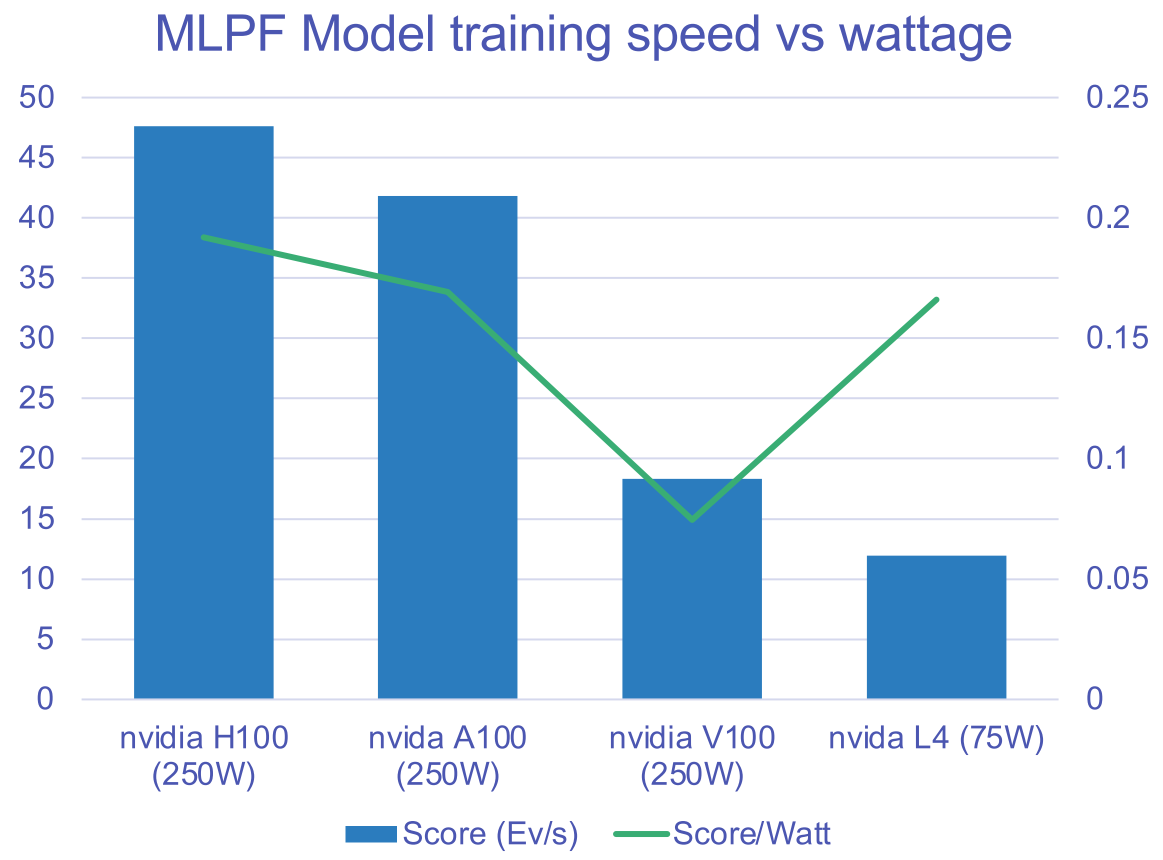
https://indico.jlab.org/event/459/contributions/11829/

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)



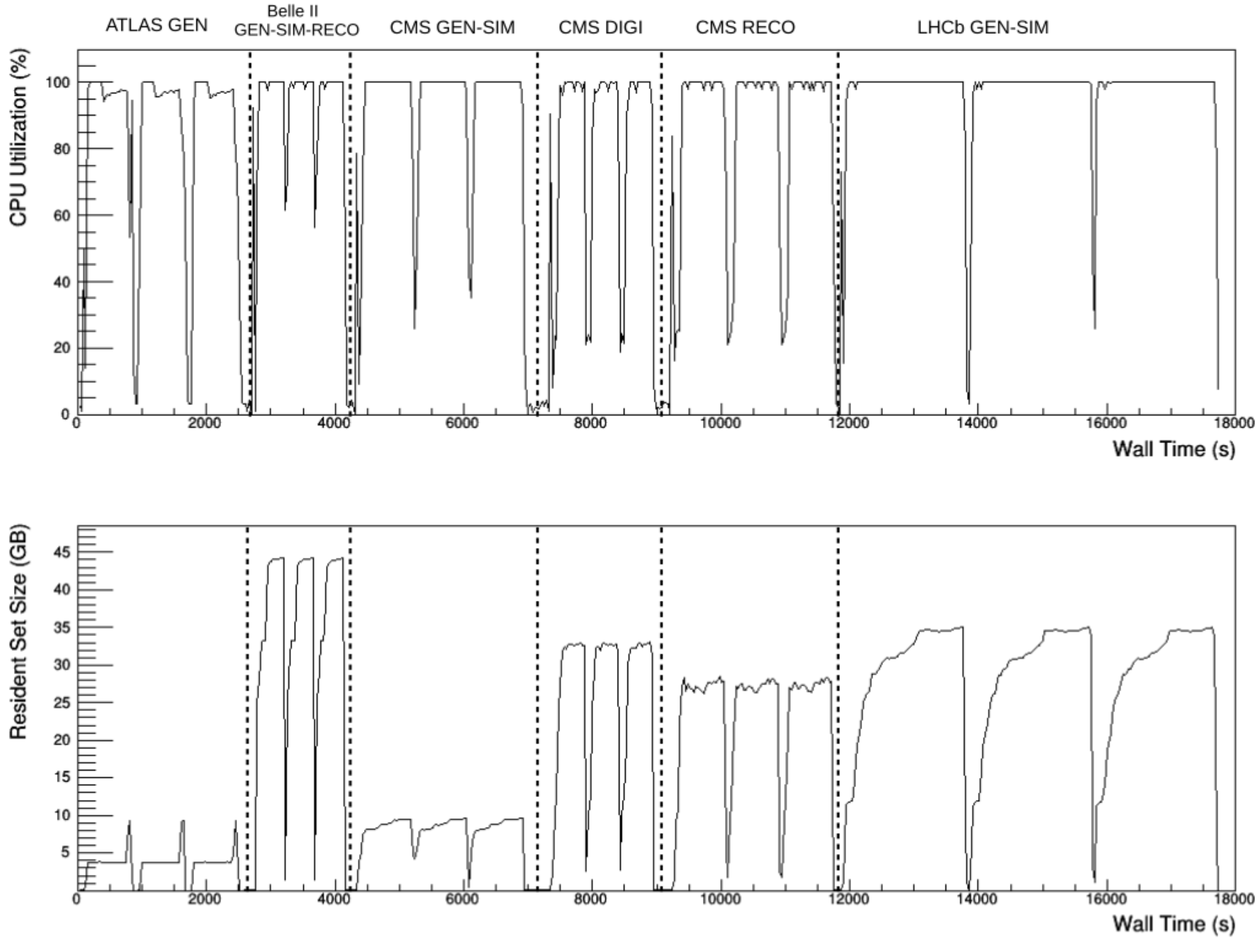
Understanding workload efficiency

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

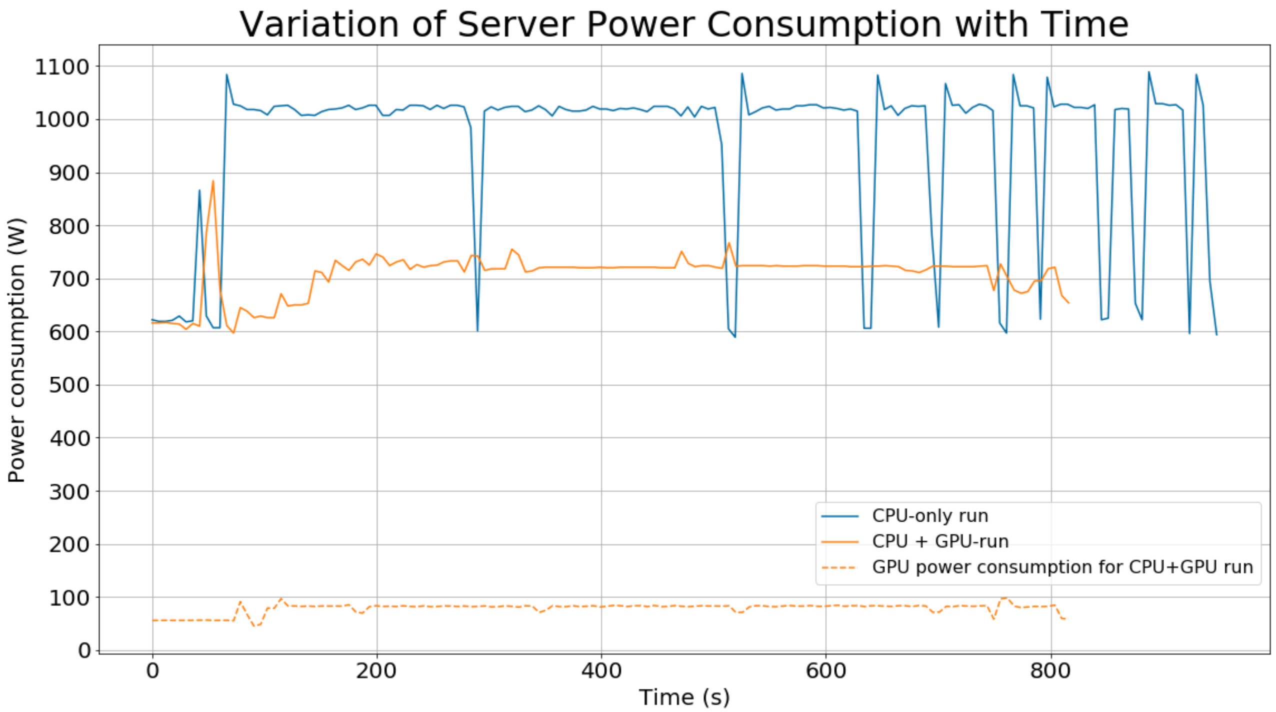


https://indico.cern.ch/event/1078853/contributions/4576275

Energy efficiency

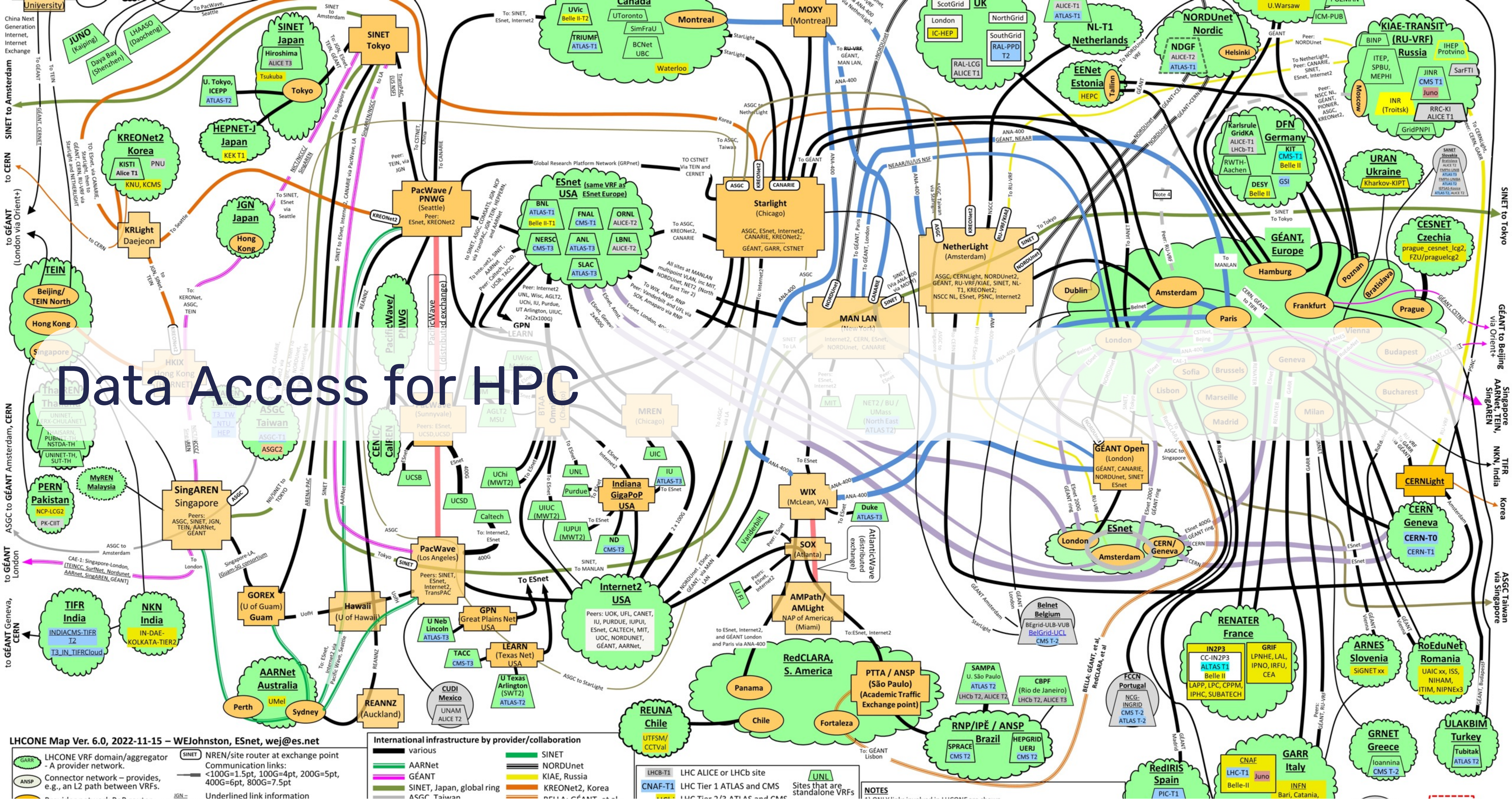
Energy efficiency is now included as a critical metric of performance

- 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



Some numbers

Initial HL-LHC models project **10s of 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

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)

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

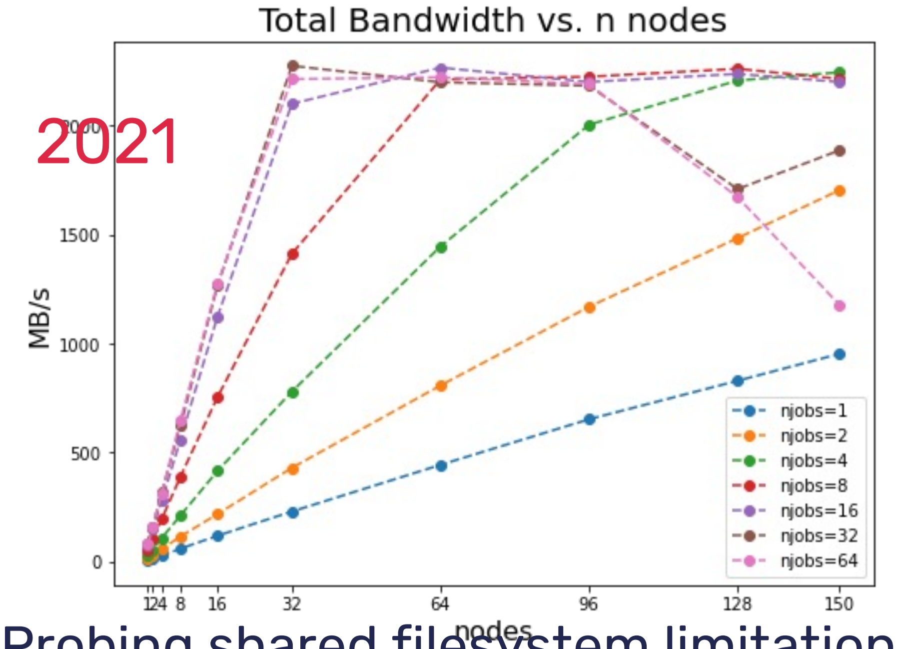
Traditional HPC workloads have low I/O demands – highly problematic running Big-Data workloads!

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

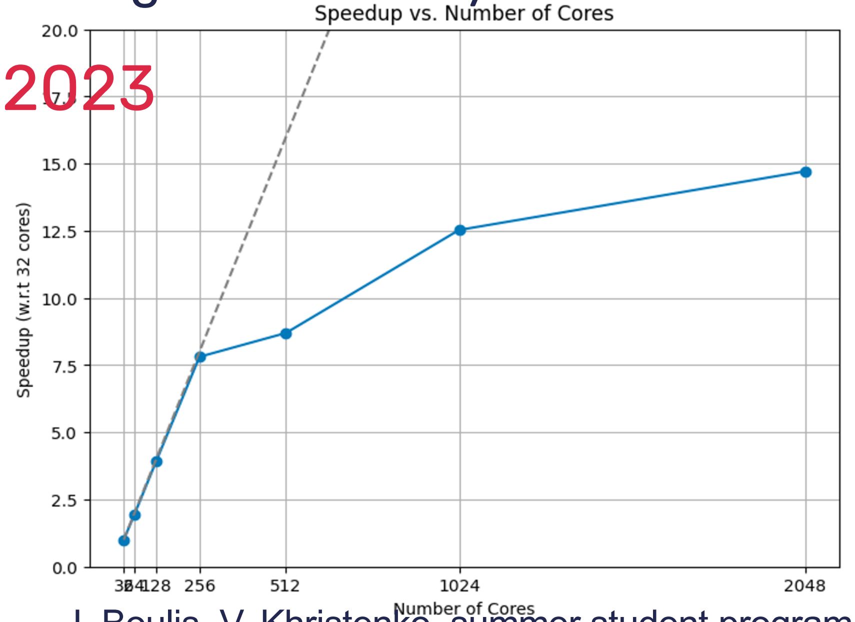
- Workload throughput O(100KB/s)-O(100MB/s)
- Many workloads per host

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

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

FENIX
RESEARCH INFRASTRUCTURE

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

Authentication & Authorization

HPC and Authentication

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

- 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,

ESCAPE

European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

ESCAPE project completed Summer 2022 after 42 months

Outlook

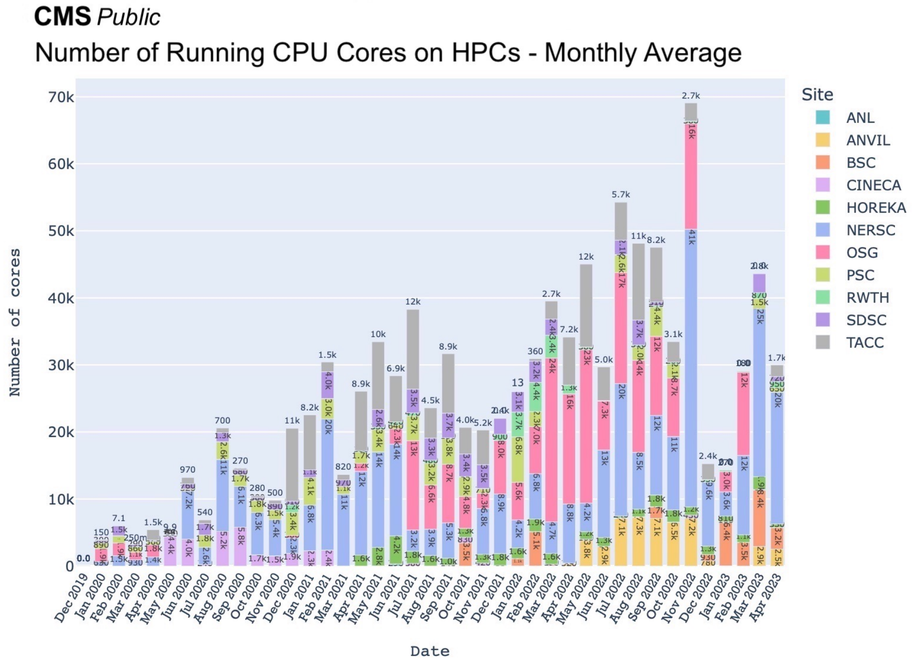
Ramping up

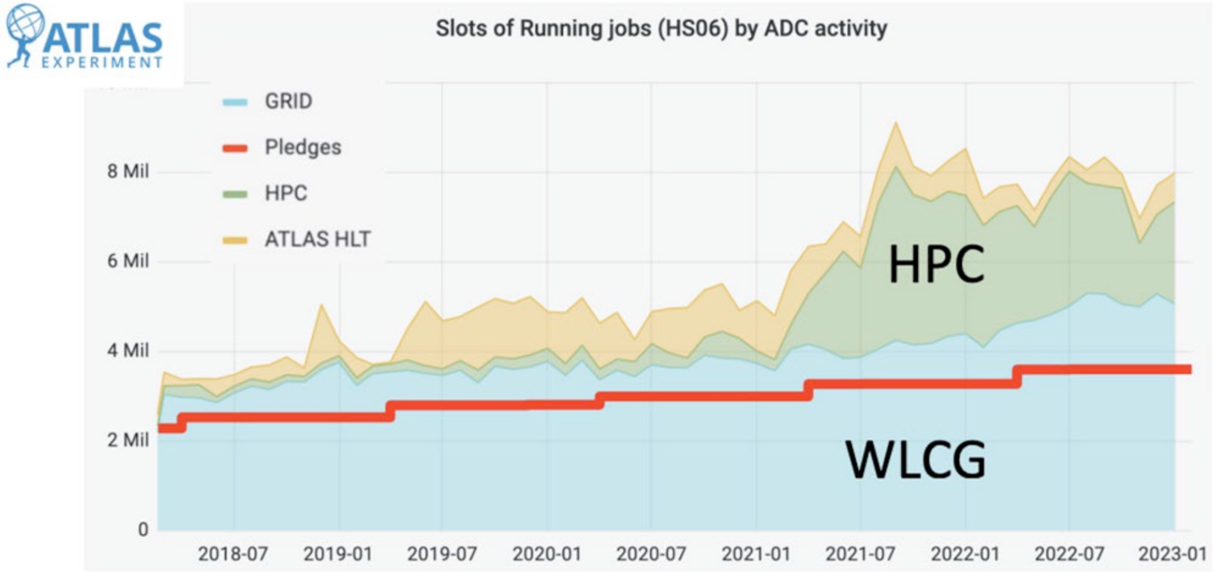
A complex problem with many moving parts – All feasible methods to close the computing gap are being pursued

- 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



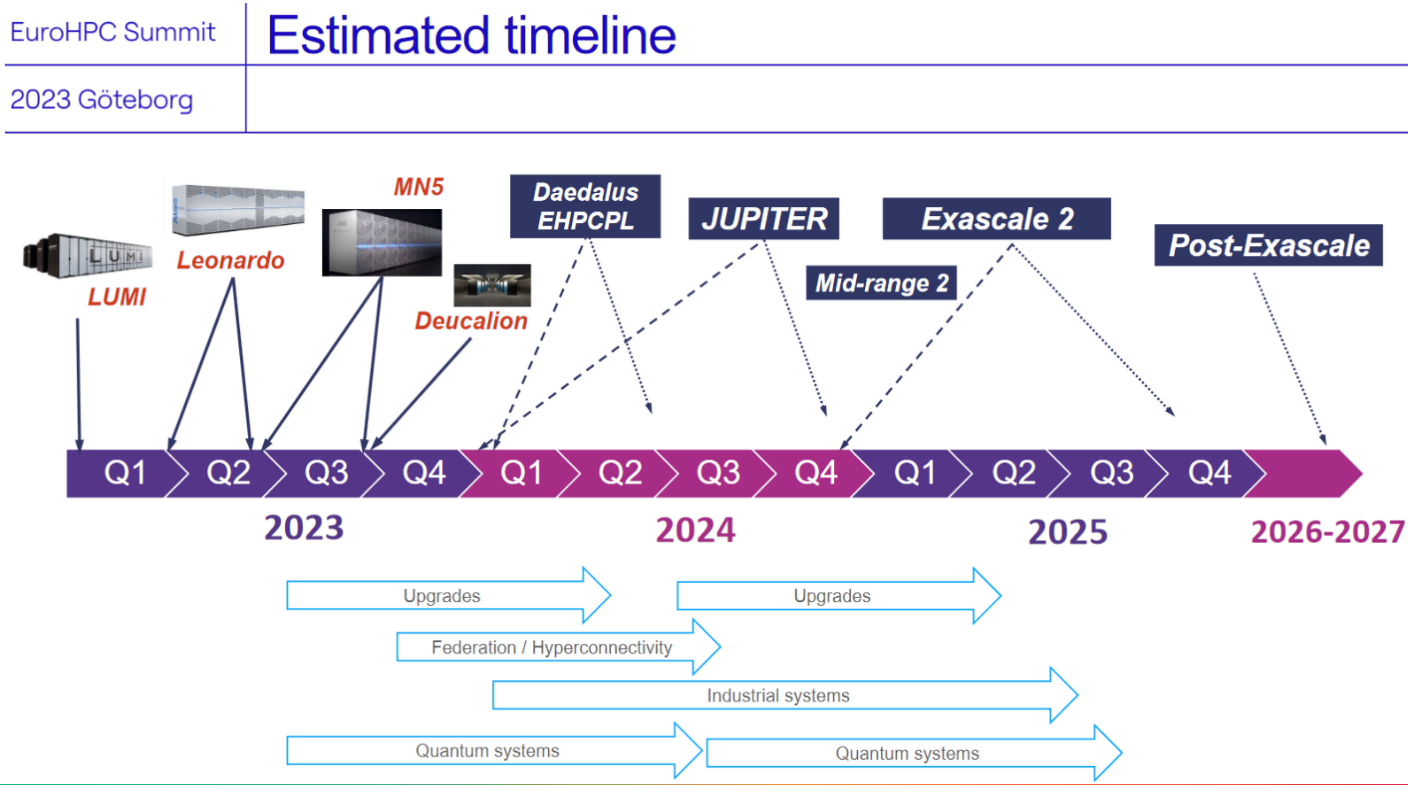


HPC is preparing for Big Data

EUPEX European Pilot for Exascale

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

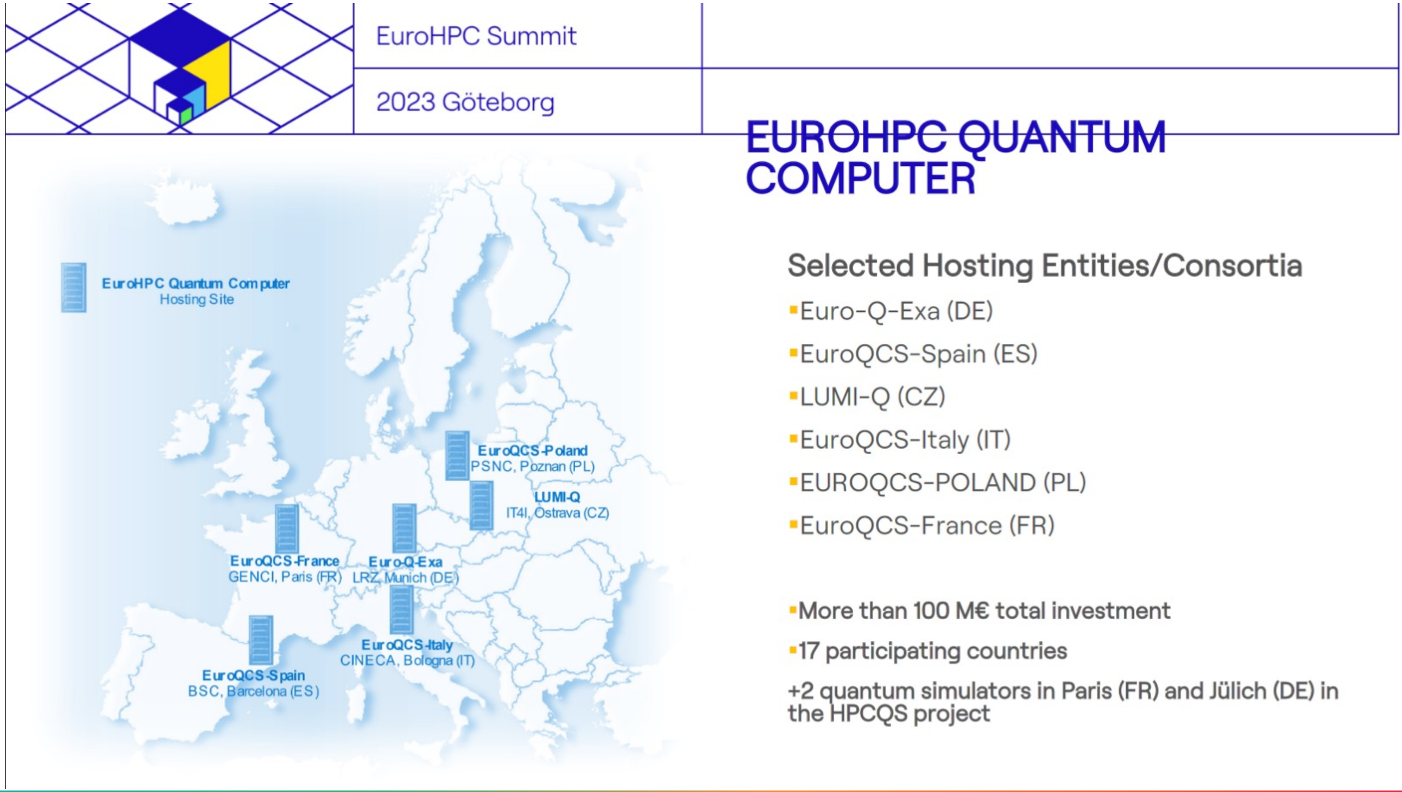
- EuroHPC call for tender for federation of hpc and quantum computers
- HPC roadmap for big-data workloads
- JUPITER procurement complete, 24’ install
- HPC <-> Cloud connectors
- Upgrading WAN connectivity



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



SPECTRUM

Computing Strategy for Data-Intensive Science Infrastructures in Europe

Approved for 2024!

Objective:

Deliver a Strategic Research, Innovation and Deployment Agenda (SRIDA) which defines the vision, overall goals, main technical and non-technical priorities, investment areas and a research, innovation and deployment roadmap for data-intensive science and infrastructures during 2025-2035

Vision:

Data-intensive scientific collaborations have access to a European exabyte-scale research data federation and compute continuum

Duration:

From 2024, 30 Months

Members:

EGI, CERN, SKAO, INFN, LOFAR, CNRS/JPV, EuroHPC (FZJ, CINECA, SURF), Other partners being contacted

horizon europe
Le programme européen pour la recherche et l'innovation

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!

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

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

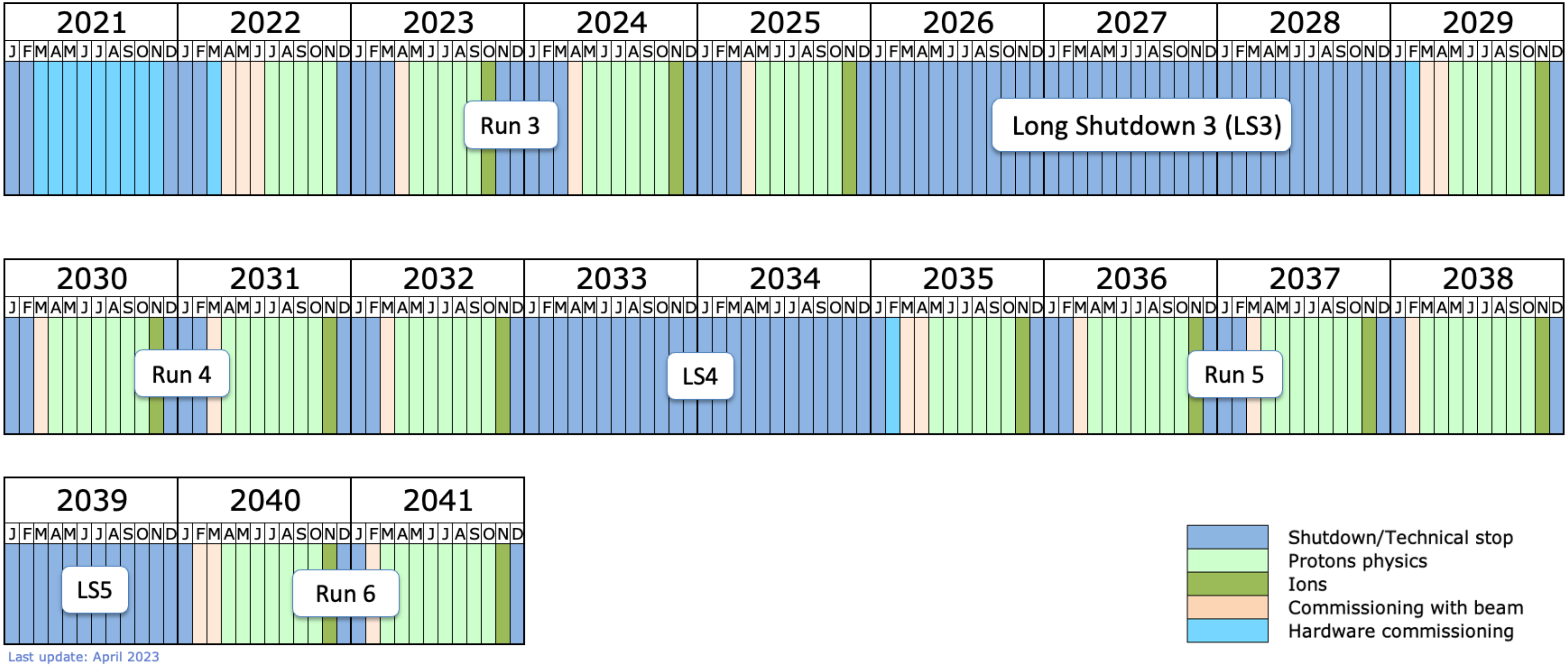
Portable frameworks

Experiments exploring several frameworks/languages to leverage heterogeneous compute

- Avoid vendor lock-in

	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



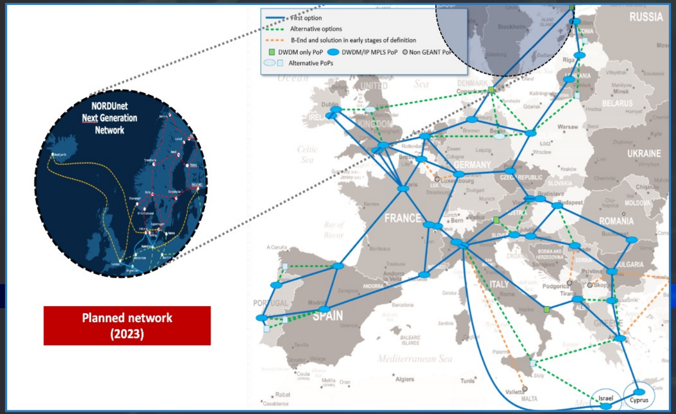
https://lhc-commissioning.web.cern.ch/schedule/LHC-long-term.htm

As we adapt

- Our consortium is ideally composed
  - HL-LHC and SKA have a burning physics need and in depth knowledge of the algorithms employed
  - PRACE provide considerable experience in the system adaptation of software environments
  - GEANT provides the infrastructure to take the computing to the many nodes that are needed to tackle the demand

[Signature Ceremony](#)

[From the HPC Collaboration Kick-off-Workshop](#)



CERN, SKAO, GÉANT, PRACE Consortium

Maria Girone
CERN openlab CEO