
CERN
openlab

HPC integration in data intensive science

CERN, SKAO, GÉANT, PRACE: The European Consortium on Advances on HPC and applications to Fundamental Research

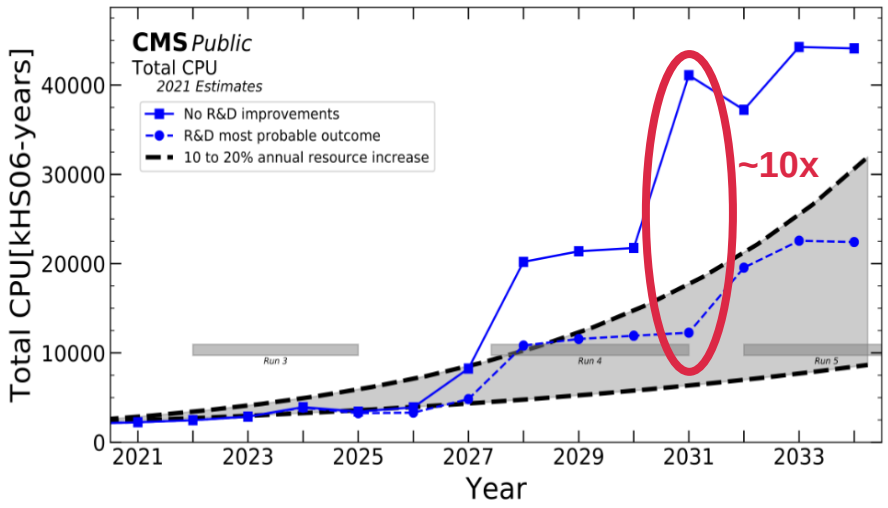
David Southwick (CERN)

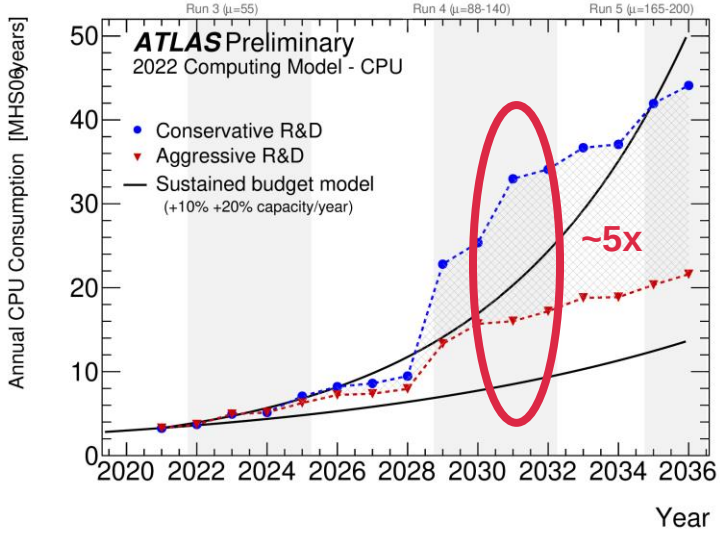
Motivation

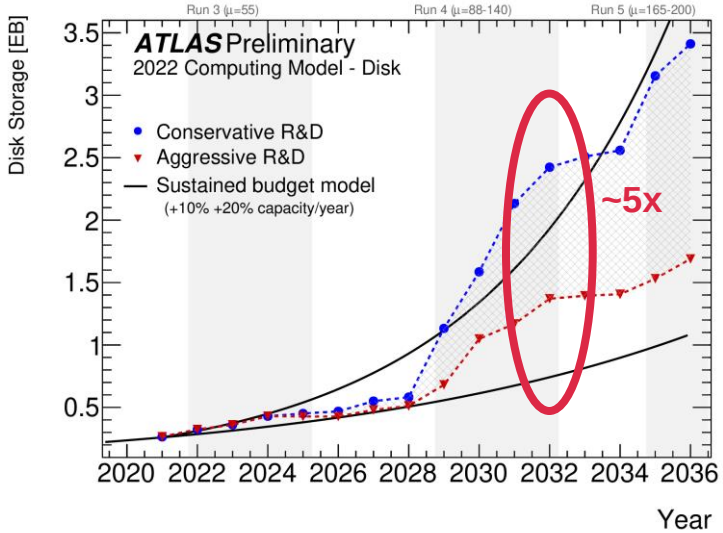
LHC expects more than exabyte of new data for each year of HL-LHC era from 2029-2040.

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

SKAO expects similar requirements during similar period.

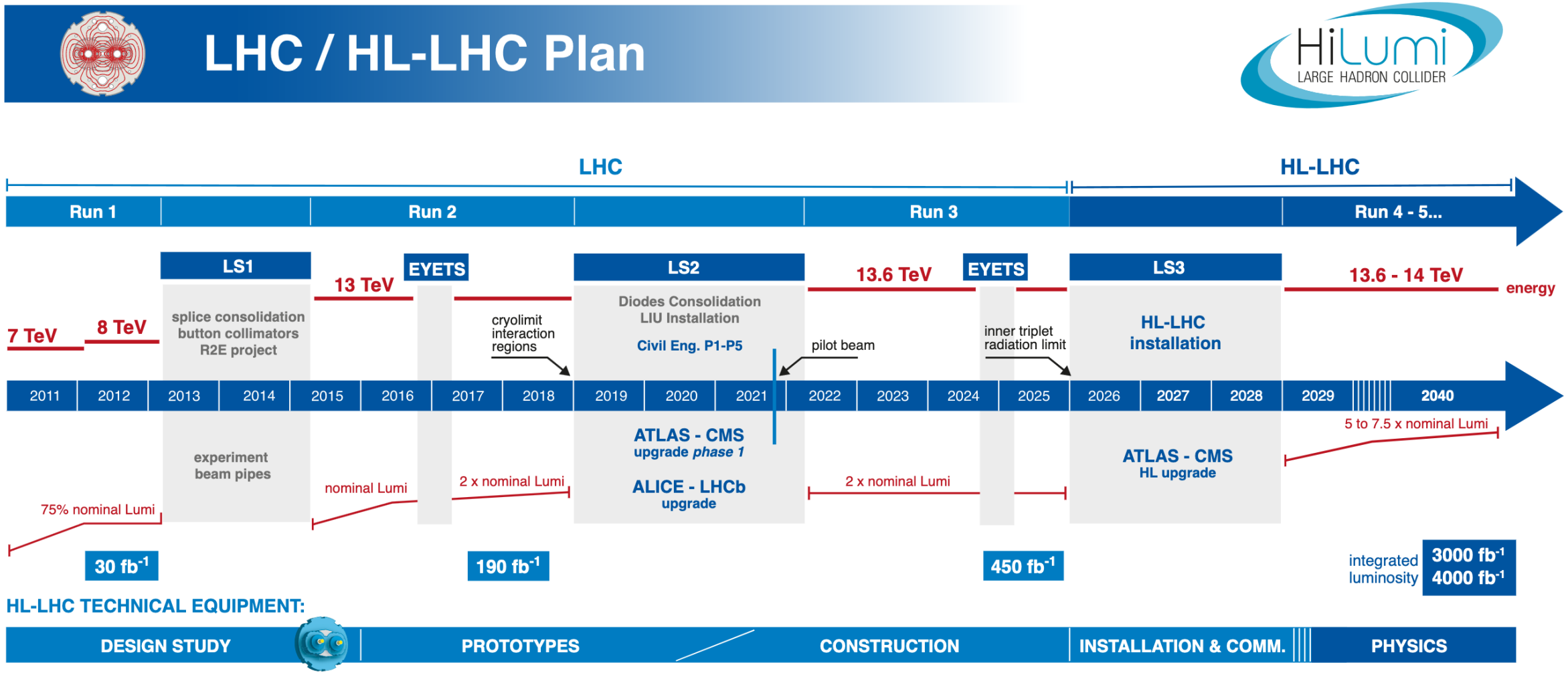






ATLAS https://indico.jlab.org/event/459/contributions/11470/ https://cds.cern.ch/record/2815292
CMS https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/UPGRADE/CERN-LHCC-2022-005/

Schedule


LHC / HL-LHC Plan
HiLumi
LARGE HADRON COLLIDER
LHC
HL-LHC
Run 1
Run 2
Run 3
Run 4 - 5...
LS1
EYETS
LS2
EYETS
LS3
7 TeV
8 TeV
13 TeV
13.6 TeV
13.6 - 14 TeV
energy
splice consolidation
button collimators
R2E project
cryolimit
interaction
regions
Diodes Consolidation
LIU Installation
Civil Eng. P1-P5
pilot beam
inner triplet
radiation limit
HL-LHC
installation
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2040
experiment
beam pipes
ATLAS - CMS
upgrade phase 1
ALICE - LHCb
upgrade
ATLAS - CMS
HL upgrade
5 to 7.5 x nominal Lumi
75% nominal Lumi
nominal Lumi
2 x nominal Lumi
2 x nominal Lumi
30 fb-1
190 fb-1
450 fb-1
integrated
luminosity
3000 fb-1
4000 fb-1
HL-LHC TECHNICAL EQUIPMENT:
DESIGN STUDY
PROTOTYPES
CONSTRUCTION
INSTALLATION & COMM.
PHYSICS

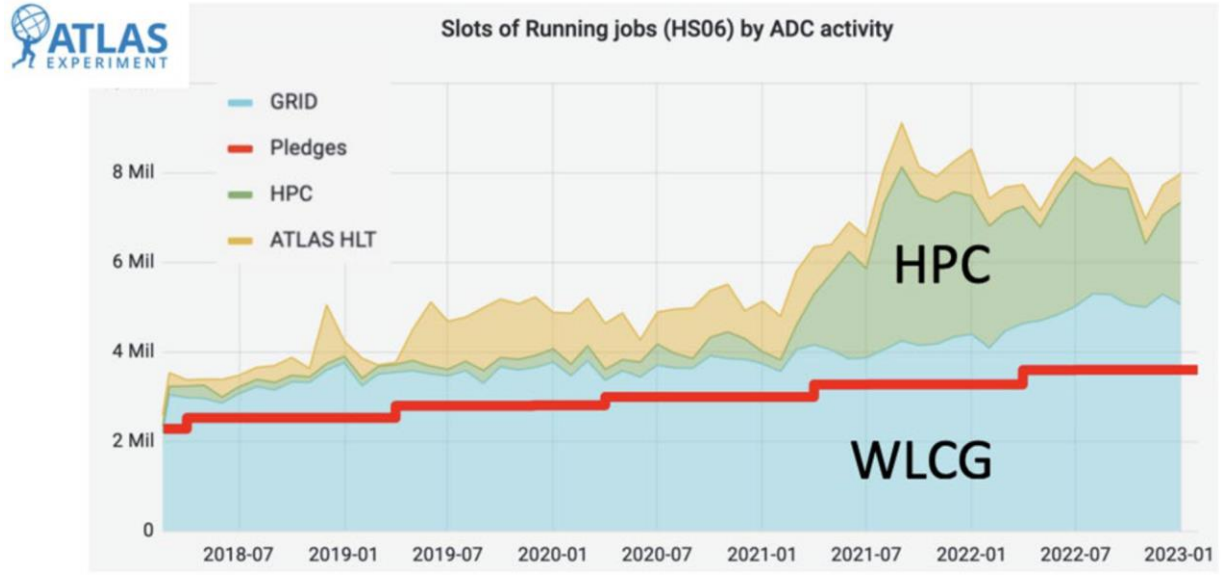
Ramping up

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

- Including HPC!

Astronomy and HEP see potentially large benefits in exploiting HPCs

Substantial technical investment during the last years which increased its usage



Xavier Espinal, EuroHPC 23'

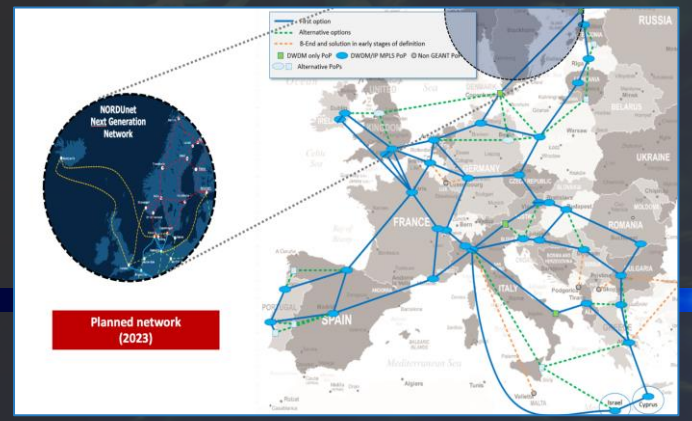
As we adapt

- 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

INFIERI 2021

Signature Ceremony

From the HPC Collaboration Kick-off-Workshop



CERN, SKAO, GÉANT, PRACE Consortium

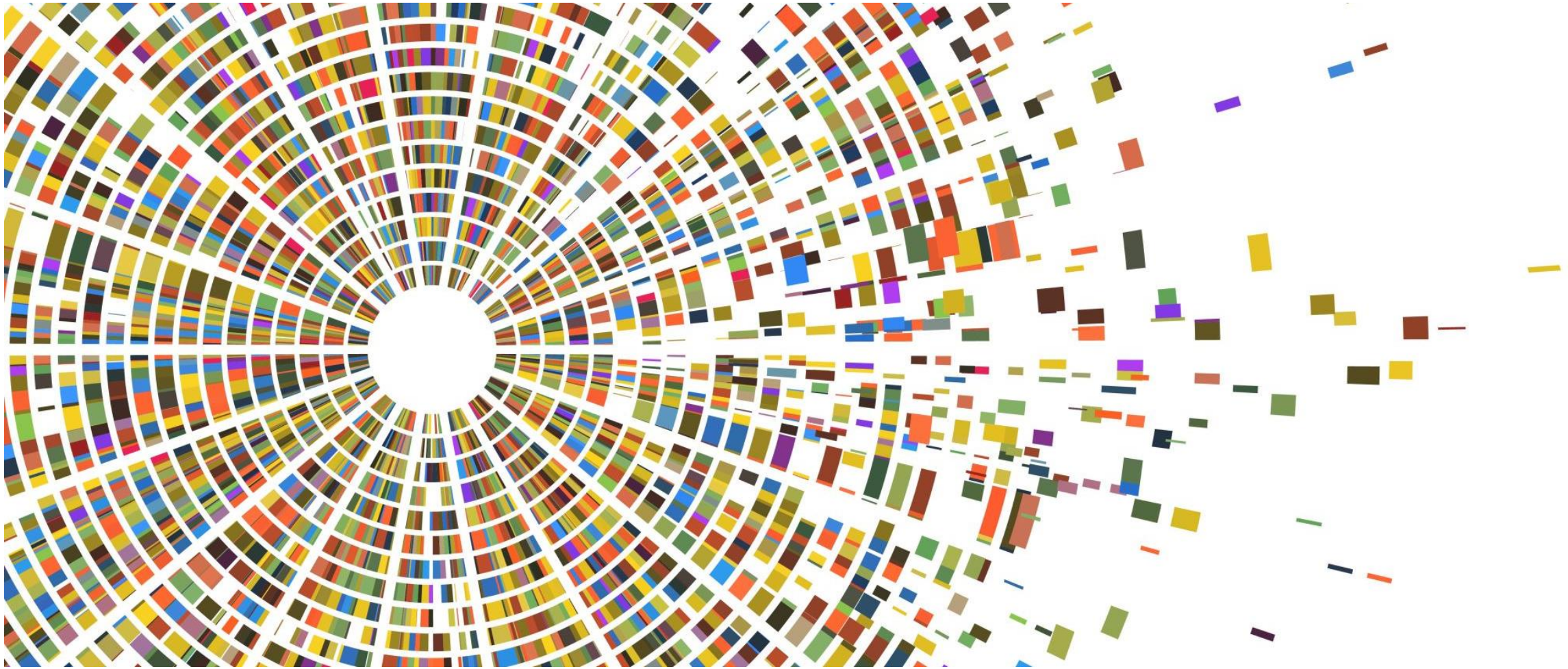
- **Consortium completed after 18 months (Dec. 2021)**
- Four areas of work identified as foundational; continue to guide development since 2021:
  - **Benchmarking**
  - **Data Access**
  - **Authentication and Authorization**
  - **Building a Common Center of Expertise**

The Four Pillars of the Collaboration

Areas of work

- Benchmarking and Accounting
- Data Processing and Access
- Authentication and Authorization
- Software and Architectures
- Runtime Environments and Containers
- Provisioning
- Wide and Local Area Networking

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

HPC Benchmarking

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

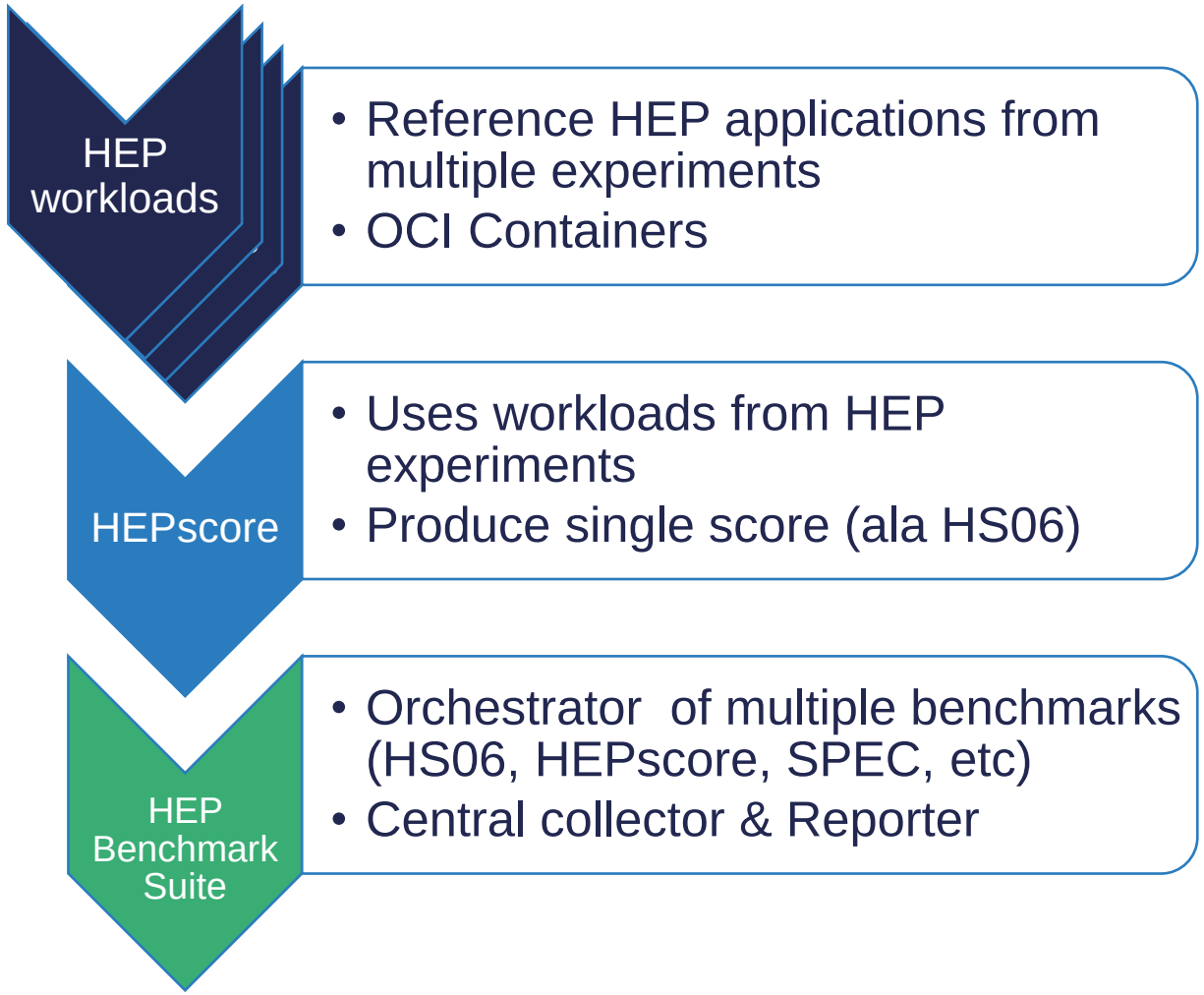
Historically benchmarking has been:

- 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!

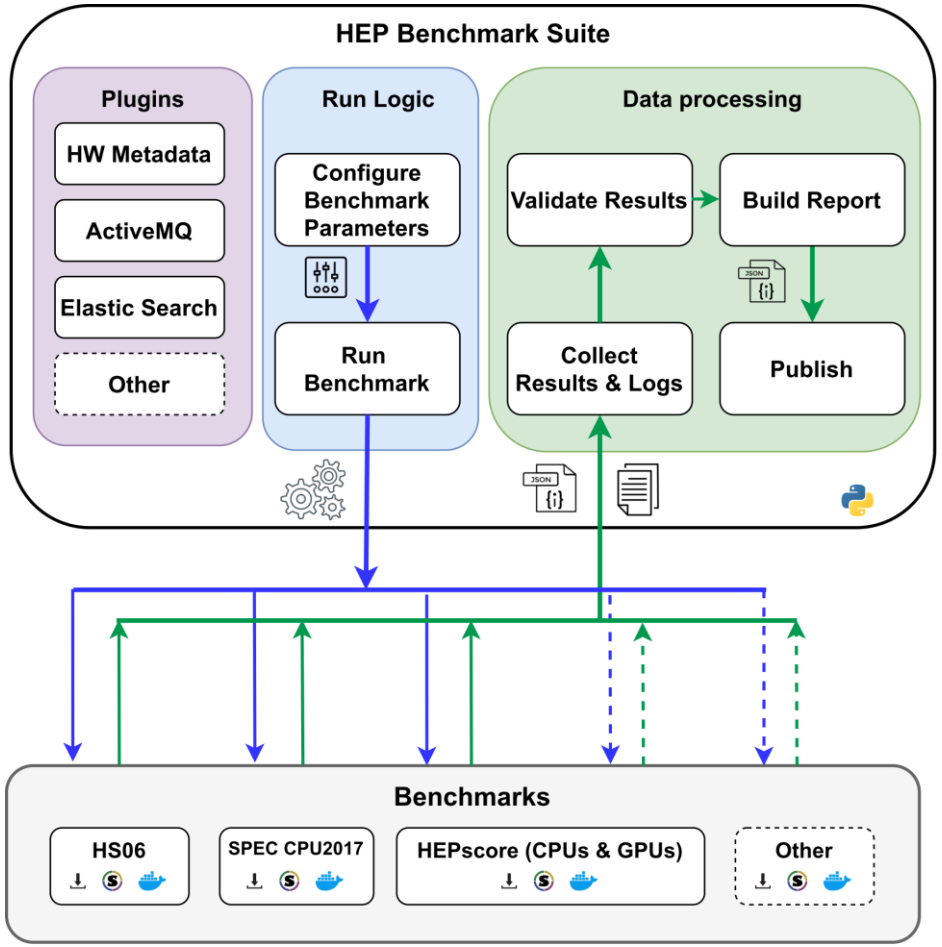
- Refactor & re-tool for user execution at scale

- HEPscore now in transition phase to replace HS06
- 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

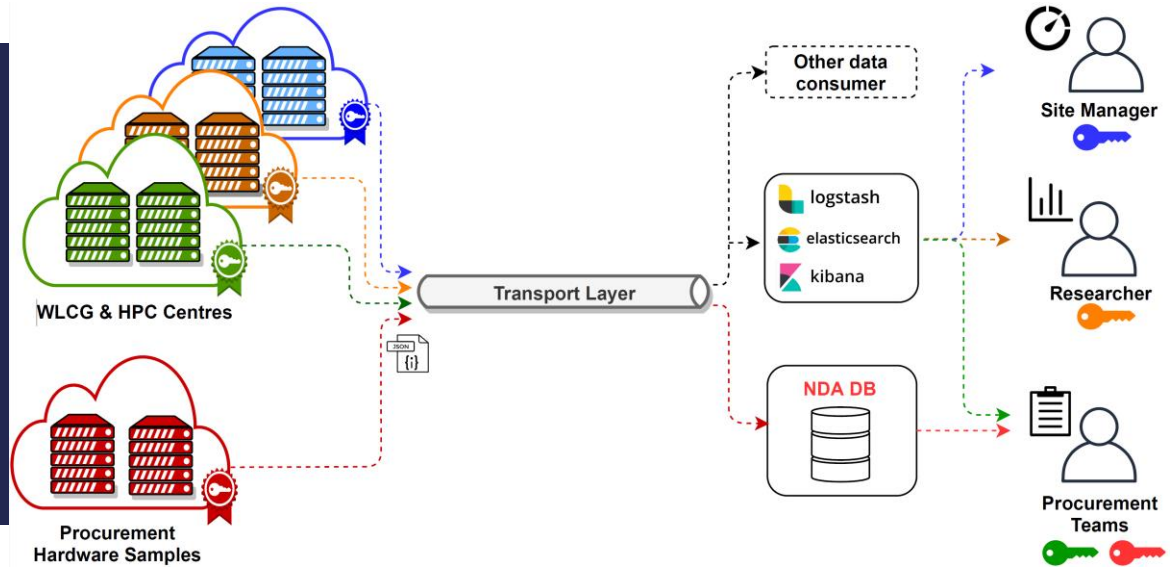
Automated HPC execution

Benchmarking Heterogeneous architectures

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

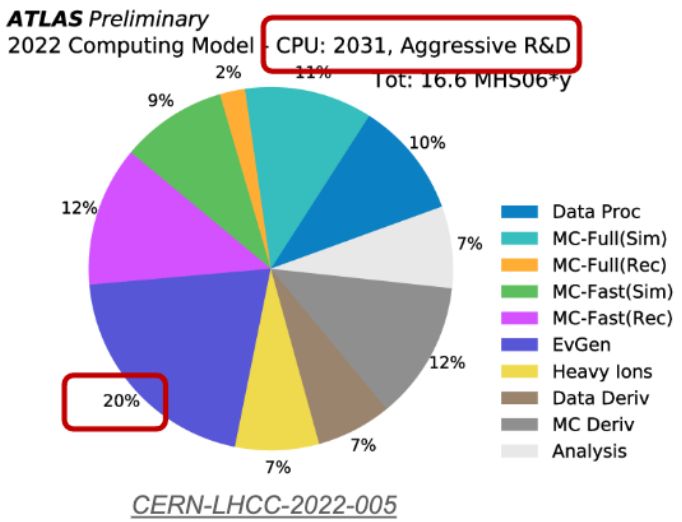
Simple integration with SLRUM, other job orchestrators

```
module purge
# 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 --user 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



CERN-LHCC-2022-005

Event generation speedup, Nvidia A100

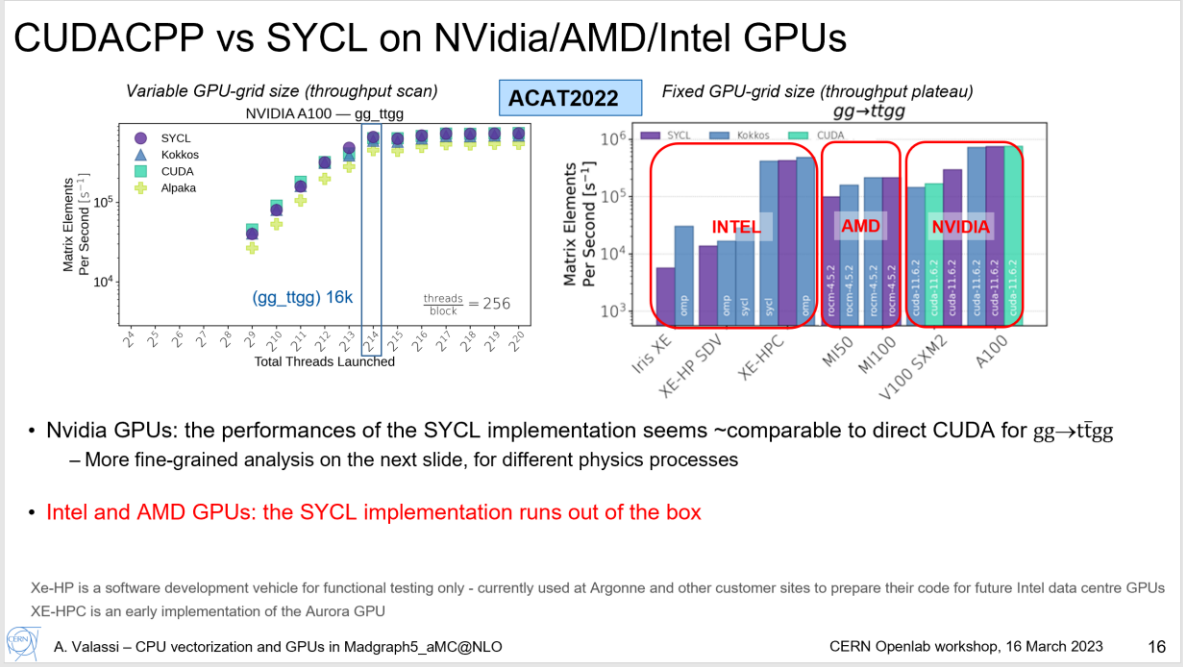
Process	Madevent 262 144 events			Standalone CUDA
	Total	Momenta+unweight	Matrix elm	ME Throughput
$e^+e^- \to \mu^+\mu^-$	17.9 s	10.2 s	7.8 s	$1.9 \times 10^6 s^{-1}$
+CUDA Tesla A100	10.0 s	10.0 s	0.02s	$633.8 \times 10^6 s^{-1}$
	1.8 x	1.0 x	390 x	334 x
$gg \to t\bar{t}gg$	209.3 s	7.8 s	201.5 s	$2.8 \times 10^3 s^{-1}$
+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

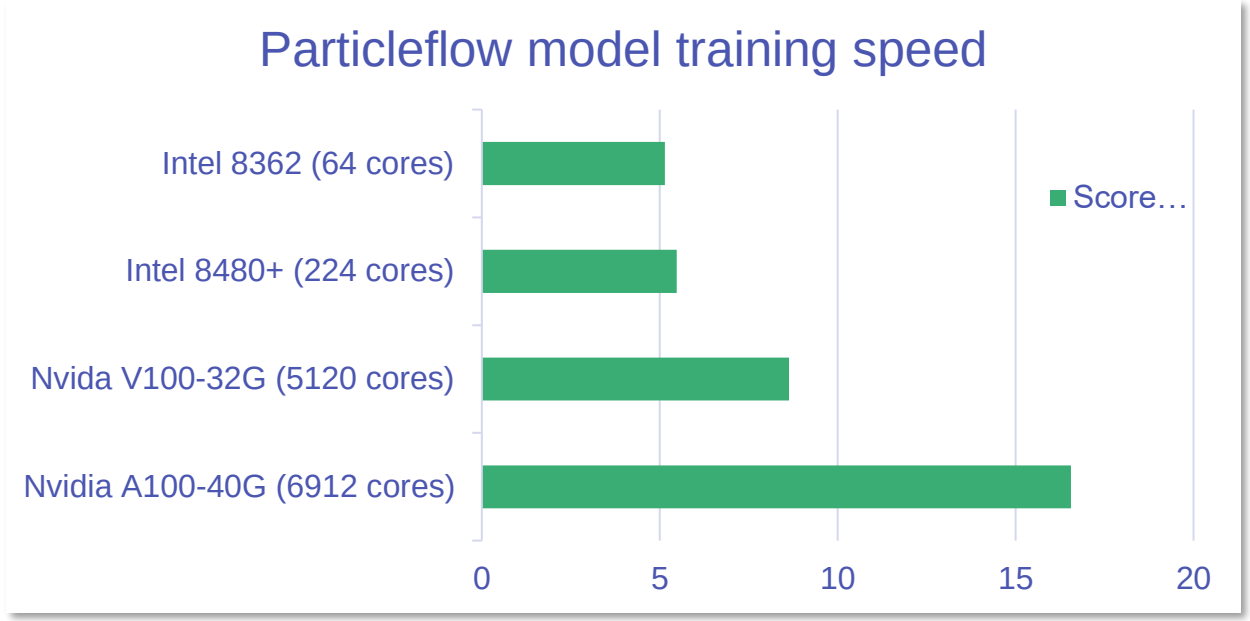
https://indico.jlab.org/event/459/contributions/11829/

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)





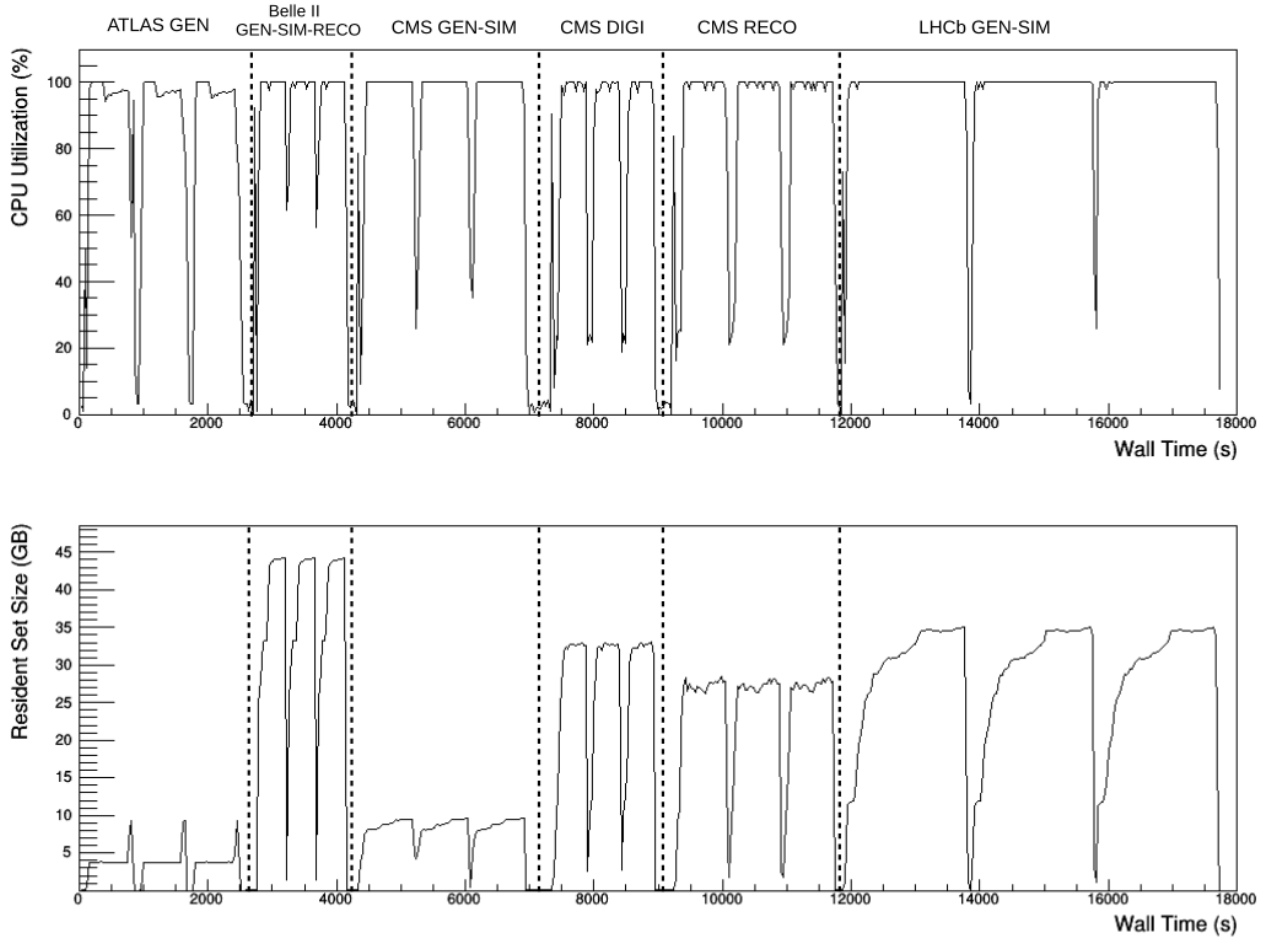
Understanding workload efficiency

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

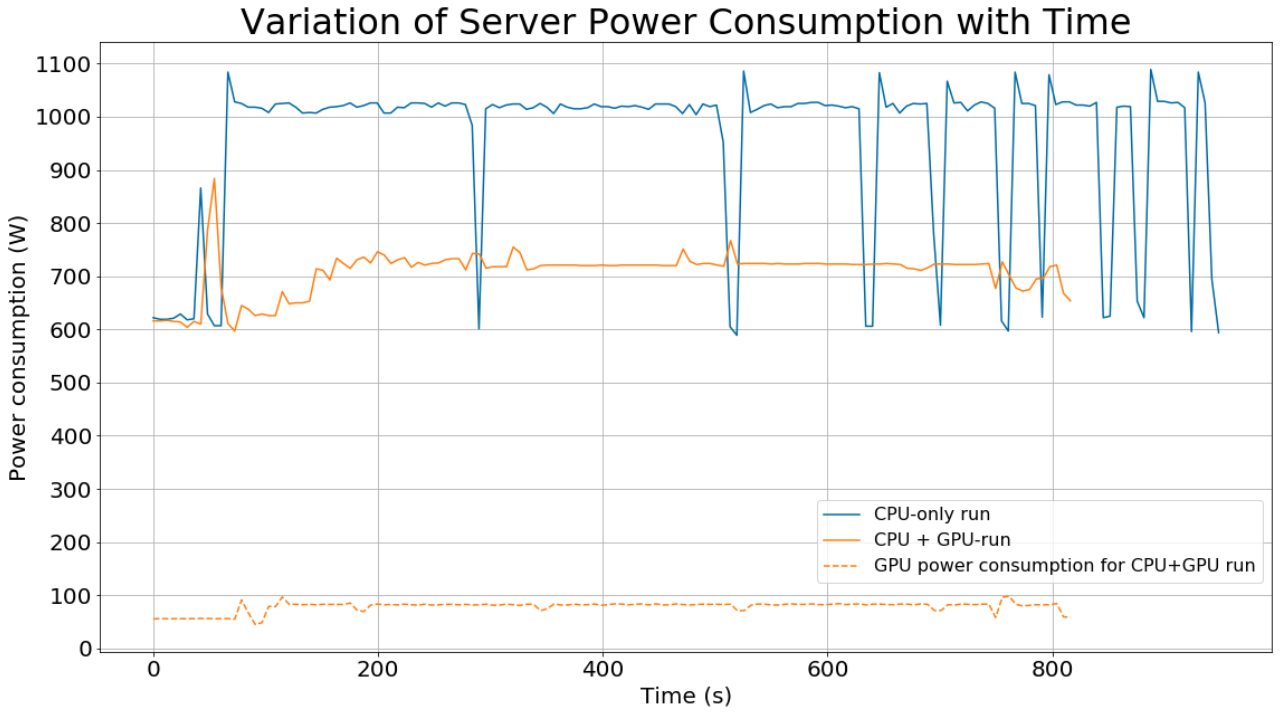


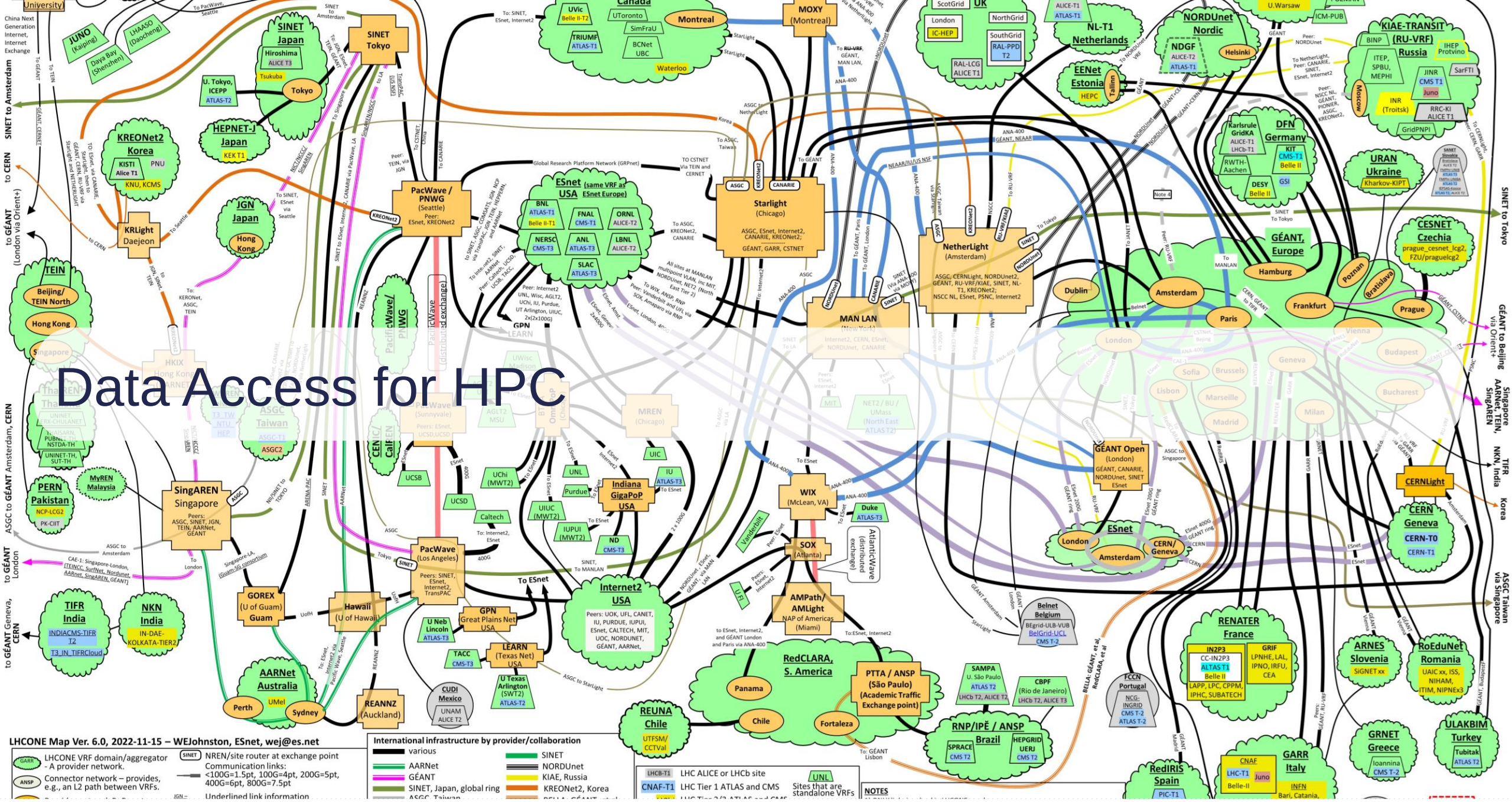
https://indico.cern.ch/event/1078853/contributions/4576275

Energy efficiency

Energy efficiency is now considered a critical metric of performance

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




Data Access for HPC
LHCONE Map Ver. 6.0, 2022-11-15 – WEJohnston, ESnet, wej@es.net
International infrastructure by provider/collaboration
various
AARNet
GÉANT
SINET, Japan, global ring
SINET
NORDUnet
KIAE, Russia
KREONet2, Korea
LHC ALICE or LHCb site
LHC Tier 1 ATLAS and CMS
Sites that are standalone VRFs
NOTES
LHCONE VRF domain/aggregator - A provider network.
Connector network – provides, e.g., an L2 path between VRFs.
NREN/site router at exchange point
Communication links: <100G=1.5pt, 100G=4pt, 200G=5pt, 400G=6pt, 800G=7.5pt
Underlined link information

Some numbers

Initial models expect **1 Exabyte physics data processing in 100 days.**

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

Goal is to stream & process 10 PB of physics data through a HPC site in a day: several hundreds of Gbps continuously.

- 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

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 center
- Efficient usage of storage systems on site
- Dynamic scaling interaction between (1) and (2)

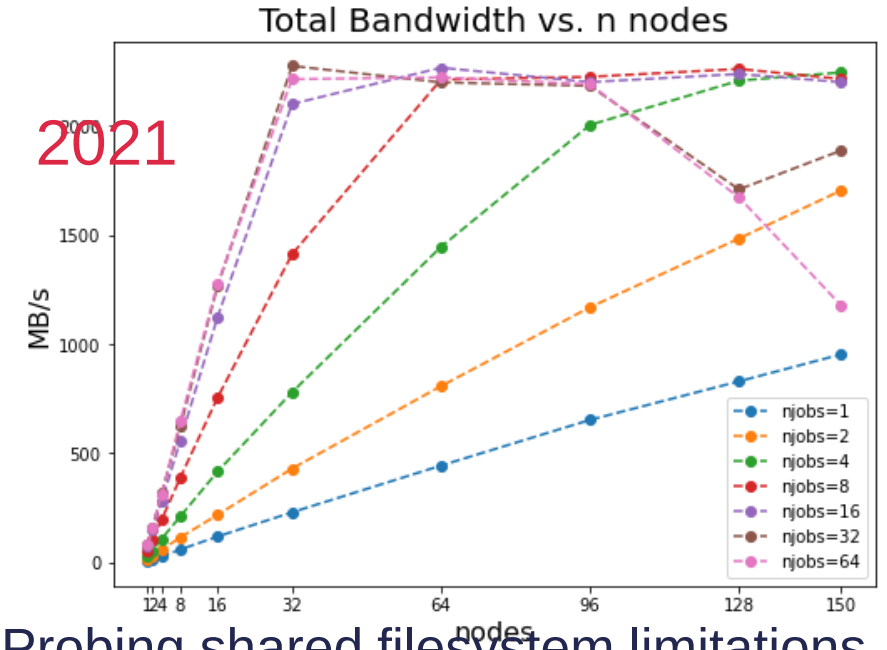
Shared filesystems

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

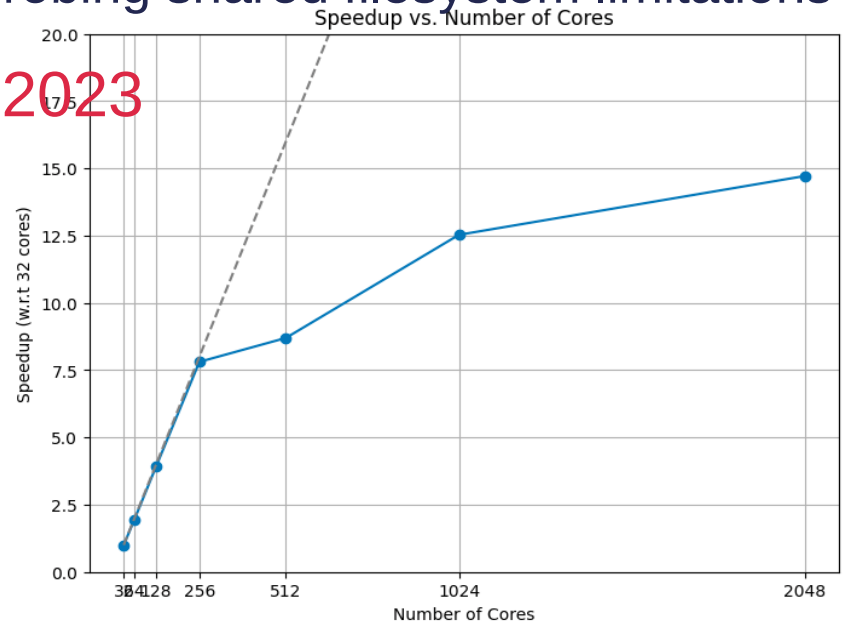
Compute-bound workloads dependent on shared file systems may be **effectively I/O bound** if scaled sufficiently

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



Probing shared filesystem limitations



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, GridFTP, S3)
- FENIX – Collaboration with HPC sites and ESCAPE to standardize data transfers

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 200Gbps tests with Jülich HPC Centre, DE

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 project completed Summer 2022 after 42 months

Outlook

Future Direction

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

- Integrating independent machines as single entities, requiring specific integration
- Access and usage policies, available services, system architectures and machine-lifetime.
- Software deployment, edge services for data and workflow management,

Moving towards a General Purpose HPC – addressing HPC as a common machine

- Enable flexibly and elastically expanding the resources available to big data sciences

SPECTRUM

Computing Strategy for Data-Intensive Science Infrastructures in Europe

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

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

CERN
openlab

Portable frameworks

	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