Computing at the HL-LHC and Beyond

Rob Gardner

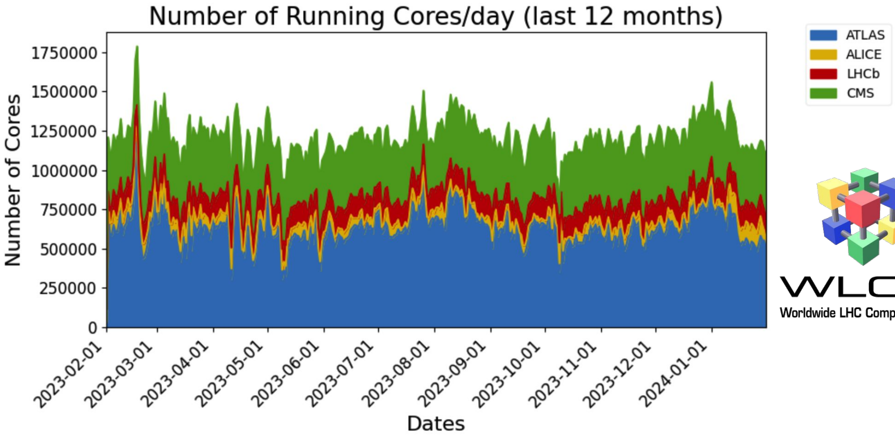
University of Chicago

On behalf of the ALICE, ATLAS, CMS, and LHCb collaborations

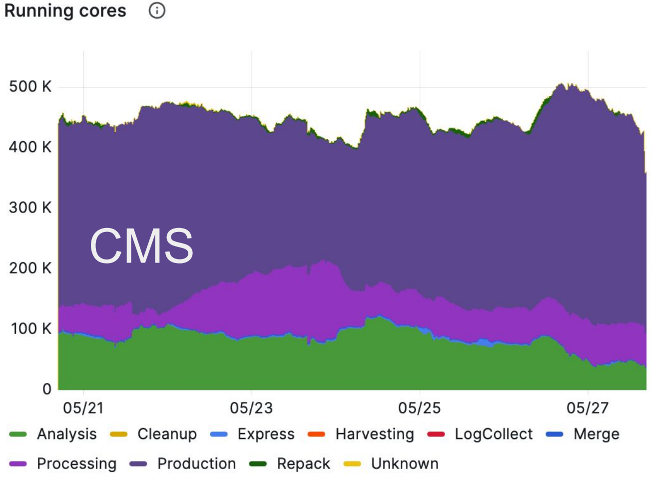
LHCP Boston 2024

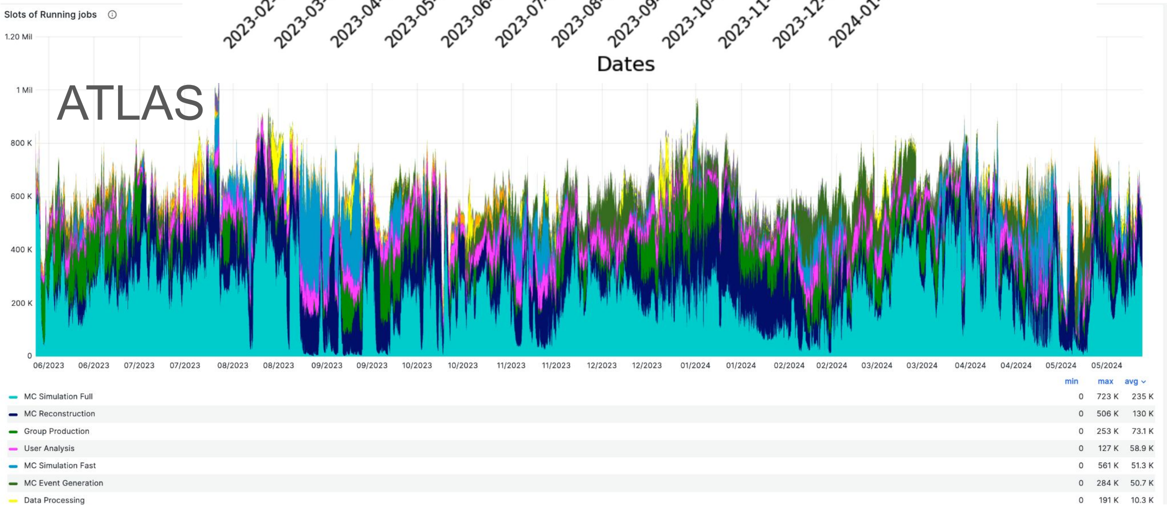
LHC computing today

- Over 1M running jobs steadily between the four experiments
- Variety of job types on 170 sites
- Grid, HPC and Cloud resources
- 24x7x365 operations

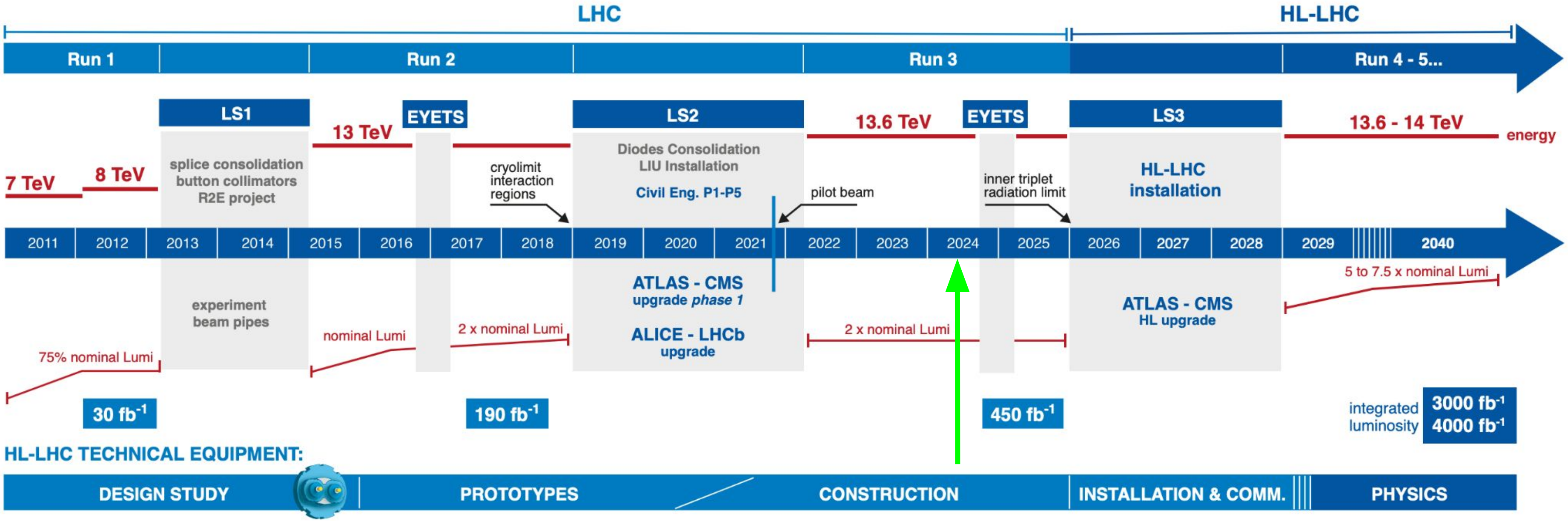


WLCG
Worldwide LHC Computing Grid





The computing road ahead: Run3 => Run4



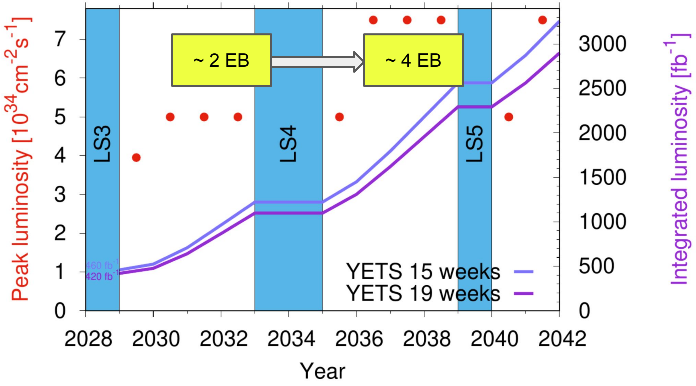
Software R&D activities: (c.f. David Shope's talk tomorrow)

Distributed Computing R&D: managing exabyte scale data, heterogeneous resources

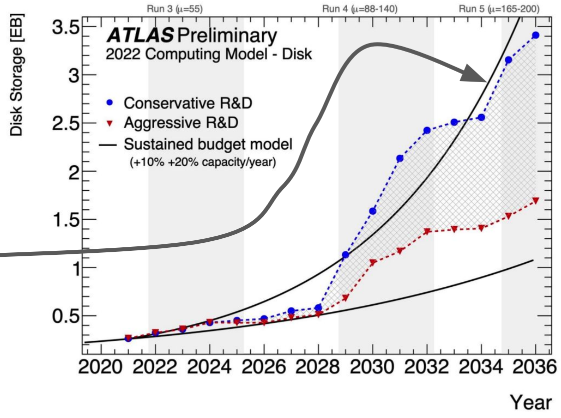
Facility R&D activities: flexible infrastructure for data delivery, analysis platforms

Resource Estimate Projections - ATLAS

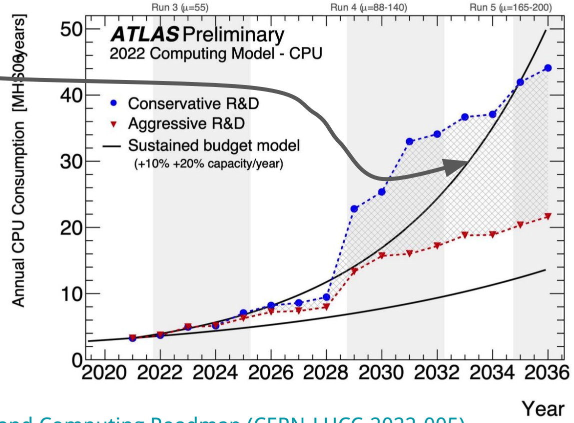
[HL-LHC luminosity](#) forecast by year



estimate resources



with flat budgets, these curves are what keep some of us awake at night



R&D program with many demonstrators, including

- Integration of non-x86 resources
  - We expect an increasing heterogeneous environment - HPCs, GPUs, ARM
- and cloud resources (e.g. ATLAS Google Project)
- Optimizing tape and disk access
- Sustainability modelling (gCO2 impact)

D. South, R&D in ATLAS Distributed Computing towards HL-LHC

ATLAS Software and Computing Roadmap (CERN-LHCC-2022-005)

Similarly, CMS

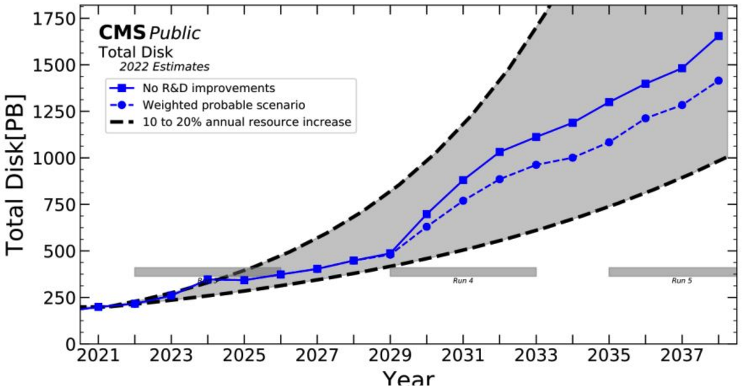
US CMS has organized R&D activities around four major grand challenges (https://arxiv.org/abs/2312.00772):

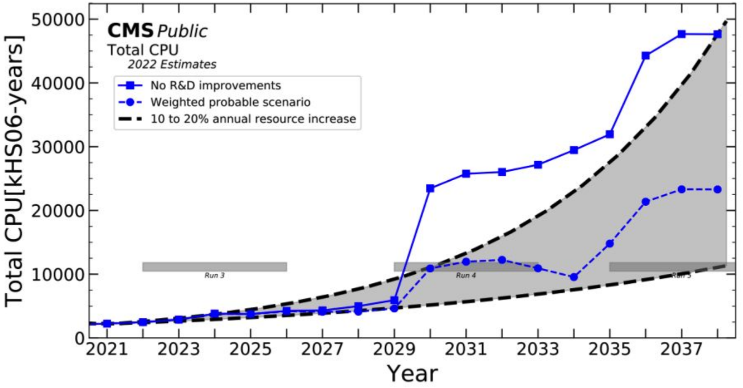
(1) **Modernizing Physics Software and Improving Algorithms** Develop innovative algorithms exploiting machine learning/AI, optimal use of modern hardware and accelerators.

(2) **Building Infrastructure for Exabyte-Scale Datasets** Build infrastructure to archive, store, transfer, and provide access to exabyte-scale datasets.

(3) **Transforming the Scientific Data Analysis Process** Leverage industry advances in data science; facility and software infrastructure to support thousands of physicists analyzing exabytes of data.

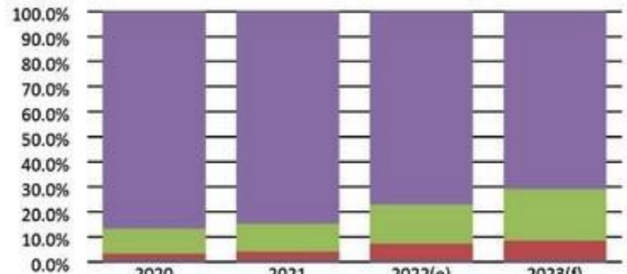
(4) **Transition from R&D to Operations** The R&D program will contribute to several advances in infrastructure, analysis facilities and networking/storage.





CMS Phase-2 Computing Model: Update Document (July 2022)

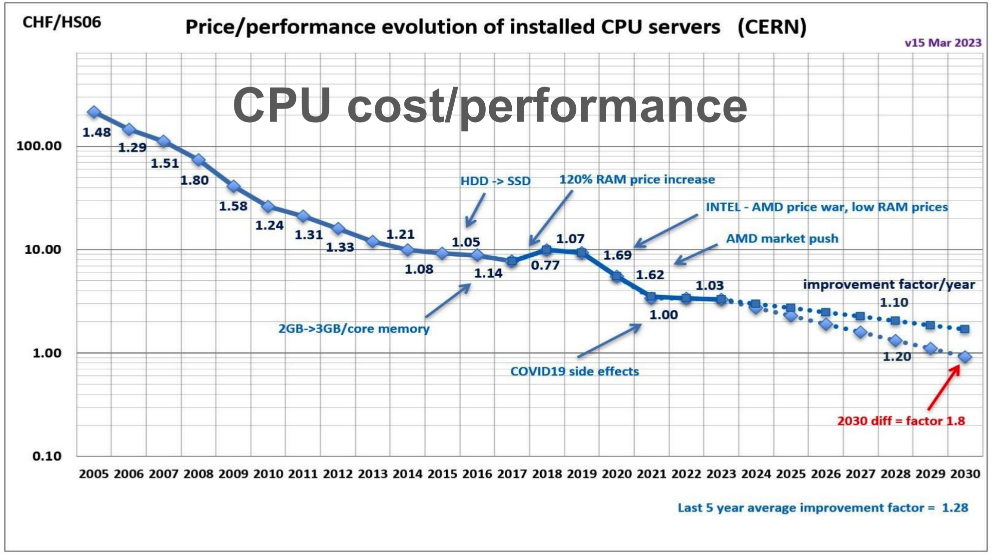
Technology and Markets - the world we live in



	2020	2021	2022(e)	2023(f)
Intel	86.6%	84.7%	77.0%	70.9%
AMD	10.1%	11.2%	15.6%	20.5%
Arm	2.6%	3.5%	6.8%	8.1%
Others	0.7%	0.6%	0.6%	0.5%

x86 and ARM CPUs by vendor. Likely will see ARM coming into WLCG this year

There are obvious risks… only 3 companies in the world capable of fabricating leading-edge chips: Samsung, TSMC & Intel



- **Expert opinion is that there are no technical obstacles for Run-4, but the overall uncertainties are in the 20% to x2 range for CPU, disk & tape**

B. Panzer-Steindel, Computing Technology and Market Evolution with a view on Run 4 (HL-LHC)

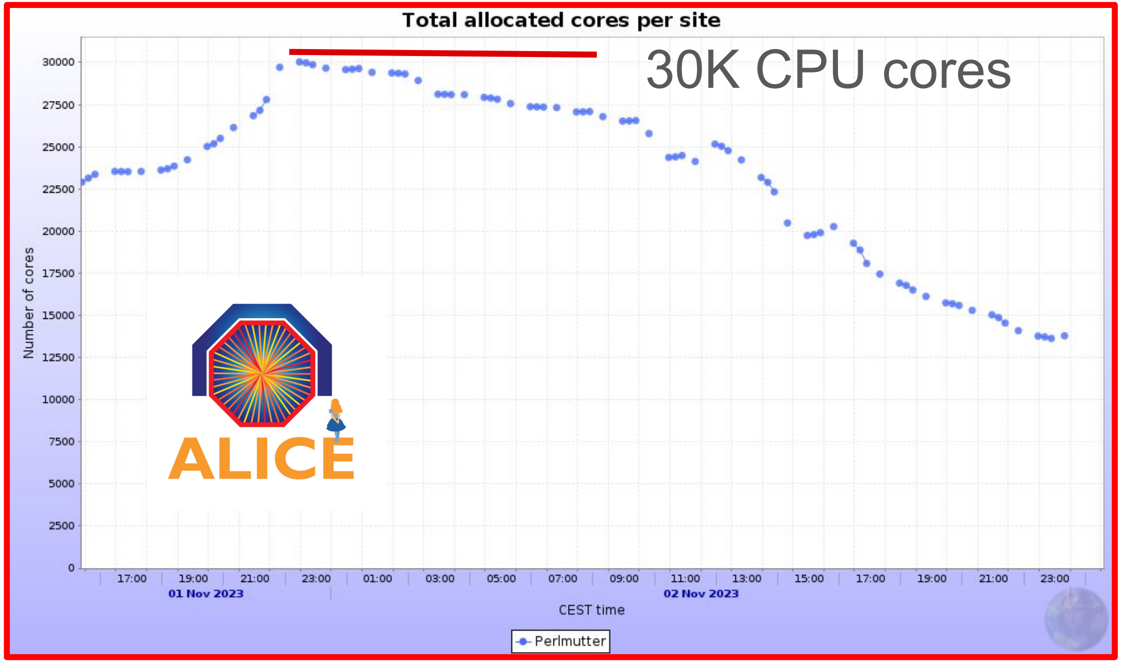
HPC - where all the cores are, just not how we like them (today)

All four LHC experiments have effectively integrated national scale HPC facilities into their workload systems, and its getting easier:

- Containers and CVMFS
- Large x86 CPU partitions have become available

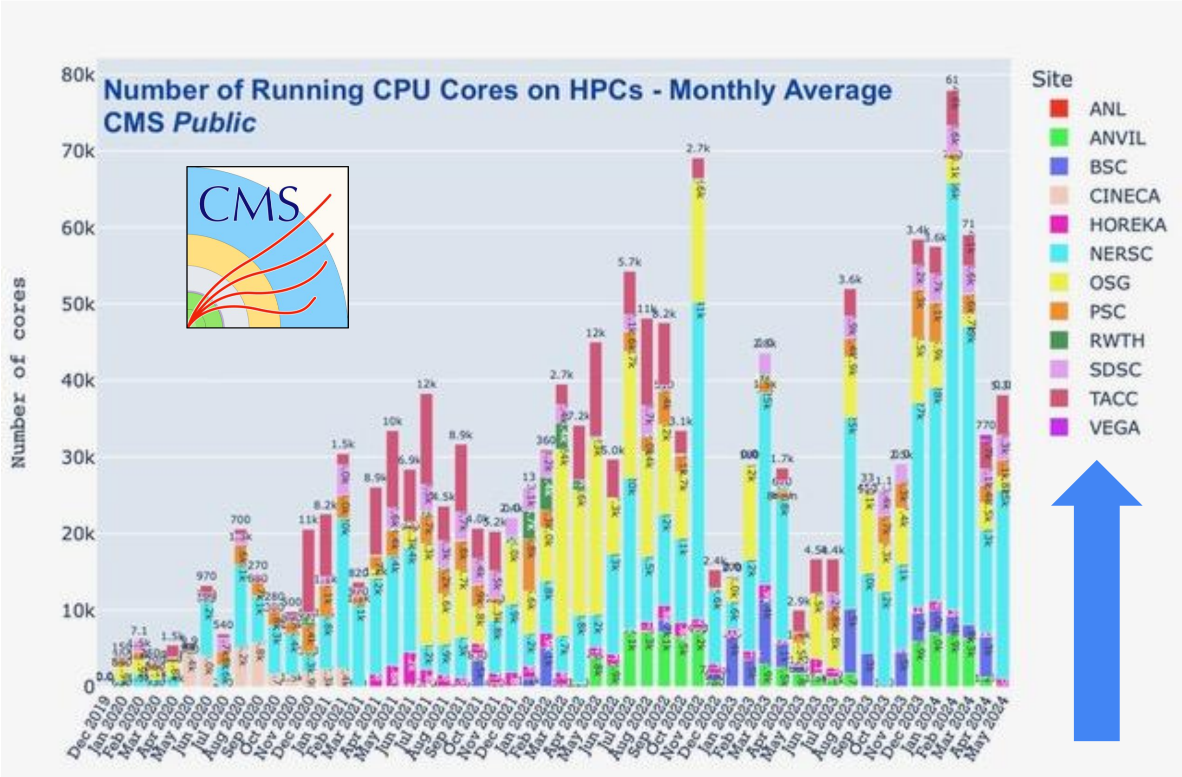
Excellent for now, but..

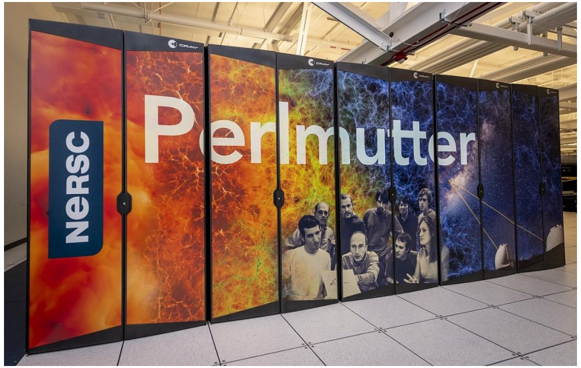
- GPUs will continue to dominate the bulk of processing capacity in large scale HPC centers

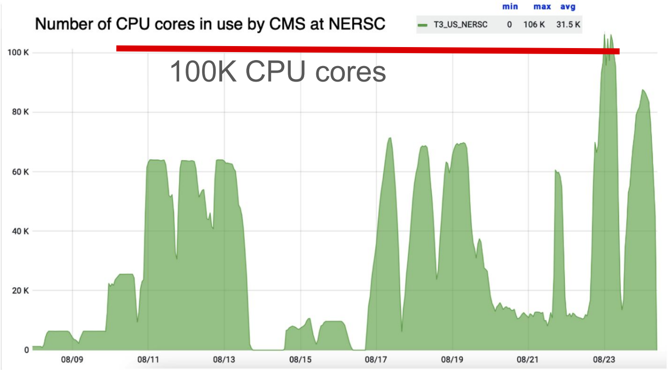


US ATLAS & US CMS whitepaper on HPC and Cloud Integration

CMS is deploying on many HPC clusters



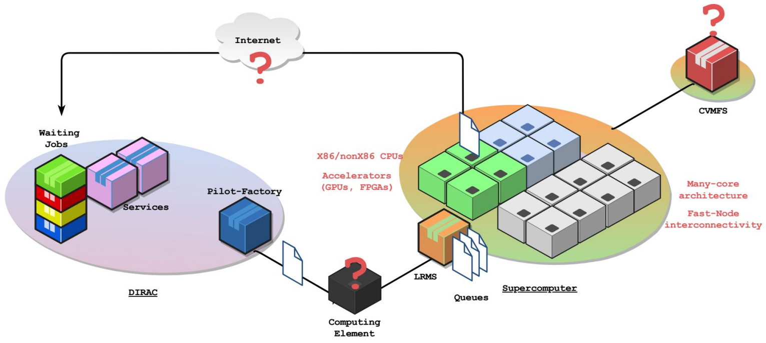


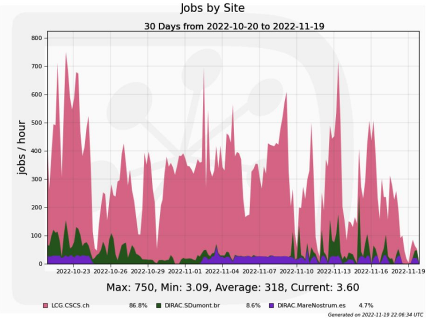


[HPC resources for CMS offline computing: an integration and scalability challenge for the Submission Infrastructure, CMS Collaboration, CHEP '23]

as is LHCb...

- Majority of LHCb offline capacity is dedicated to MC simulation on WLCG
- Additionally there are HPCs
  - **But with the same challenges**: software access, worker node connectivity, local scheduler configuration
- A number of technical solutions have been devised
  - cvmfs-exec, installing pilot on edge, installing special agents to communicate to external services, install a "cvmfs-builder", etc.
- And a diverse set of HPCs are now accessible by LHCb
  - PizDaint, SantosDumont & MareNostrum





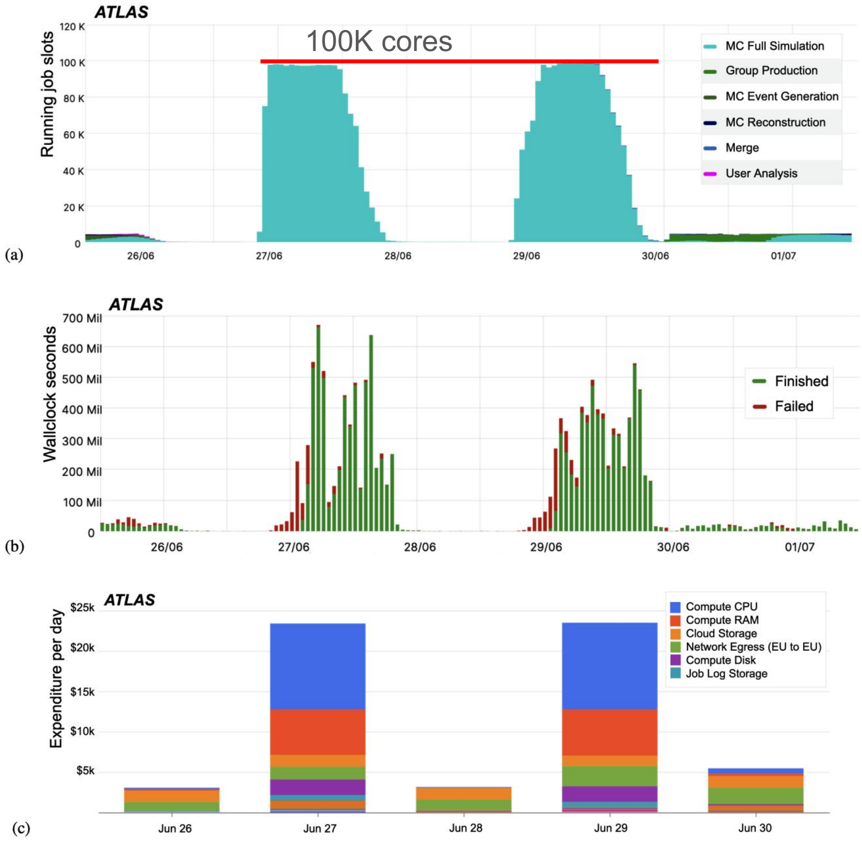
Available supercomputers process 300 jobs/hour on average
vs
WLCG grid resources process 14,000 jobs/hour on average.

[Alexander Boyer, Integrating LHCb workflows on Supercomputers: State of Practice (CHEP 23)]

Public cloud offers burst capability & versatility

- Google site for ATLAS
  - Independent, or as a Tier2 extension
  - Support all workflows and data management functionality
- Subscription pricing model
  - Cost components & comparison to grid sites
- On-demand resources
  - An excellent bursting resource for simulation
- Versatility
  - GPU and ARM queues easy to setup
- Total cost of ownership study

Total cost of ownership and evaluation of Google cloud resources for the ATLAS experiment at the LHC ATLAS Collaboration, May 2024



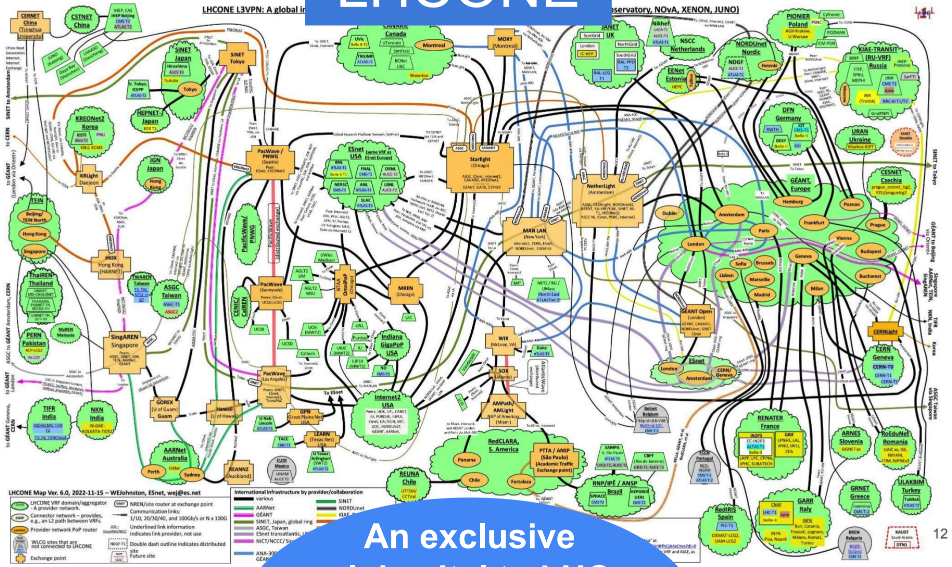
The network - thankfully its been a step ahead

- Our most reliable component!
- Well managed by a strong community of R&E network professionals
- Many sites have 100 Gbps and are planning for 400 Gbps
- Nearly always a shared resource -- ***Large scale science ensures continued investment***

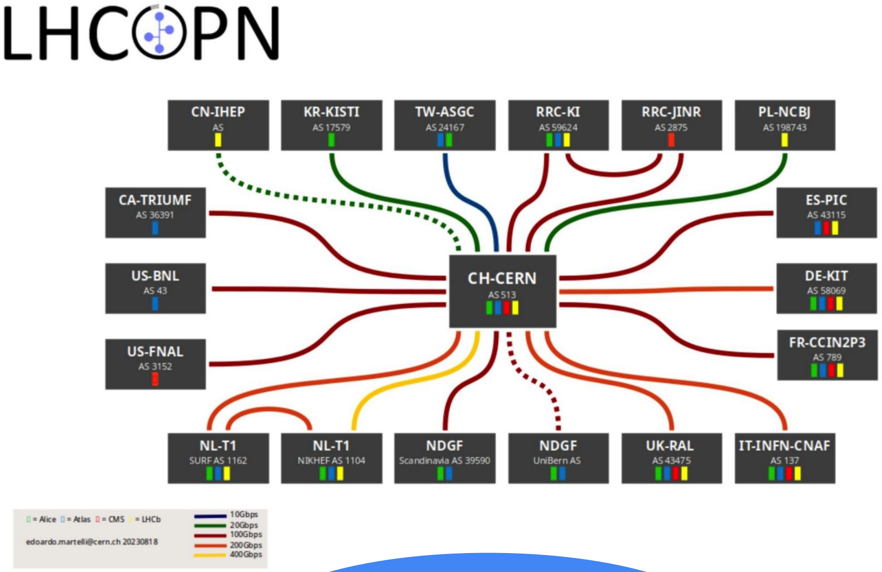
WLCG/DOMA Data Challenge 2024 Final Report

But we'll need more bandwidth capacity & optimization tools

LHCONE



An exclusive club, vital to LHC physics, 14 years on.. crucial for HL-LHC



Dedicated links from CERN to national centers

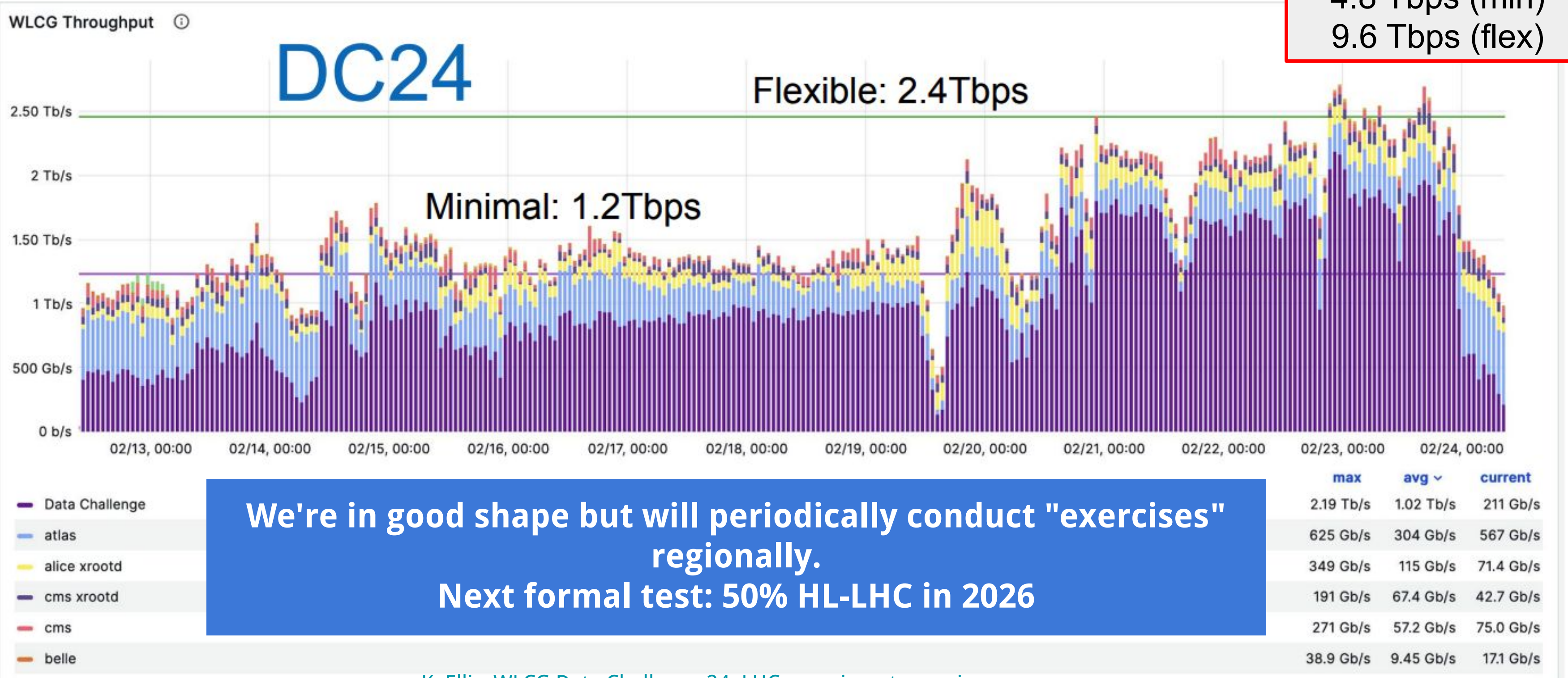
M. Babik, Distributed Computing Challenges at the LHC and HL-LHC, 4th GLOBAL RESEARCH PLATFORM WORKSHOP Oct 2023

R&D activities in the backup slides

Data Challenges to test & plan for HL-LHC scale data

Demonstrated **25% HL-LHC throughput** across the WLCG tiers

HL-LHC
4.8 Tbps (min)
9.6 Tbps (flex)



We're in good shape but will periodically conduct "exercises" regionally.
Next formal test: 50% HL-LHC in 2026

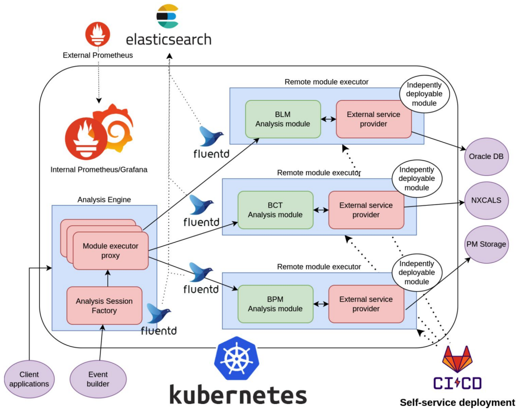
K. Ellis, WLCG Data Challenge 24: LHC experiment experiences

Commercial cloud technology in our community

Use of **containerization** and **"cloud tech"** on sites

- **Kubernetes** at CERN ([R. Rocha](#))
  - Analysis Reproducibility - REANA
  - Machine Learning and Kubeflow
  - GitLab, GitLab CI/CD
  - Notebook Servers and SWAN
  - CERN IT central monitoring
  - LHCb Dirac - Workload Management
  - **Accelerator control system & Technical Network**
- R&D efforts at Tier2 sites

a totally modern accelerator controls network at CERN… how fitting

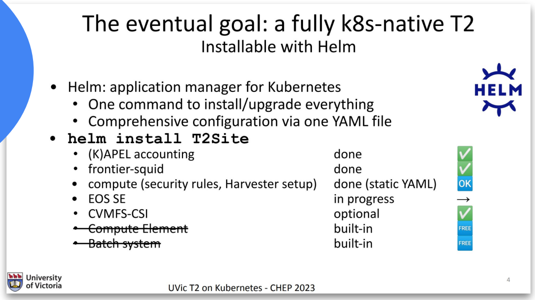


An Update on the CERN Journey from Bare Metal to Orchestrated Containerization for Controls: https://inspirehep.net/literature/2754625

Motivations:

- **Agility**: hardware management, software deployment and validation.
- **Optimize**: infrastructure resources in terms of cost and energy.
- **Align:** with DevOps practices in industry

Transforming WLCG sites to be more versatile



The eventual goal: a fully k8s-native T2

Installable with Helm

- Helm: application manager for Kubernetes
  - One command to install/upgrade everything
  - Comprehensive configuration via one YAML file
- `helm install T2Site`

(K)APEL accounting	done	✅
frontier-squid	done	✅
compute (security rules, Harvester setup)	done (static YAML)	OK
EOS SE	in progress	→
CVMFS-CSI	optional	✅
~~Compute Element~~	built-in	FREE
~~Batch system~~	built-in	FREE

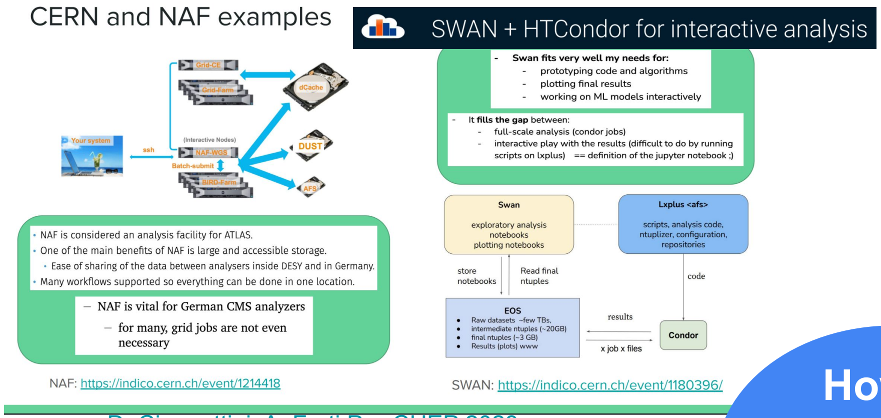
University of Victoria

UVic T2 on Kubernetes - CHEP 2023

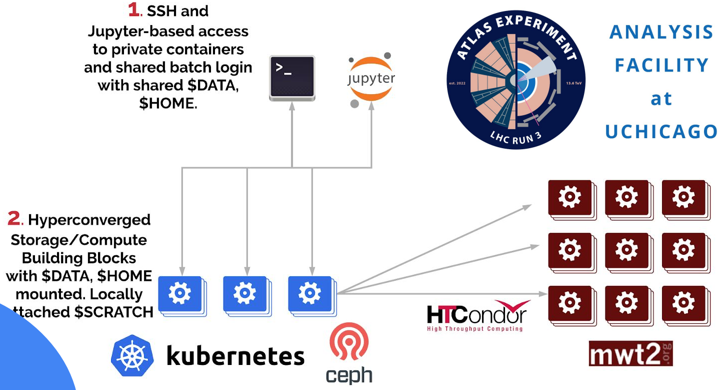
4

Facility R&D efforts at Tier2s: [Using Kubernetes as an ATLAS computing site](#) and [A grid site reimagined: Building a fully cloud-native ATLAS Tier 2 on Kubernetes](#) and [Operational Experience and R&D results using the Google Cloud for High Energy Physics in the ATLAS experiment](#)

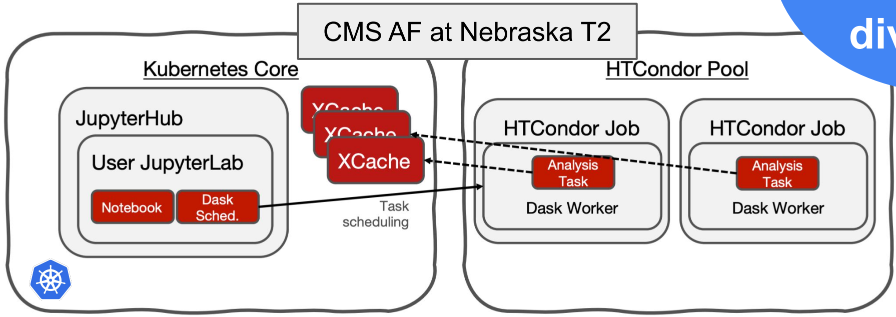
Traditional and new approaches for Analysis Facilities



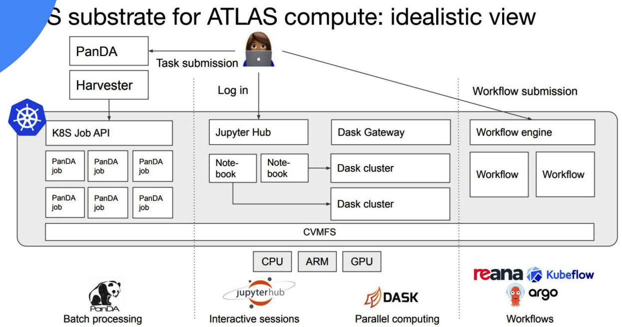
D. Ciangottini, A. Forti Pre-CHEP 2023
E. Tejedor, Analysis Faclity at CERN



How to benefit from the diversity?

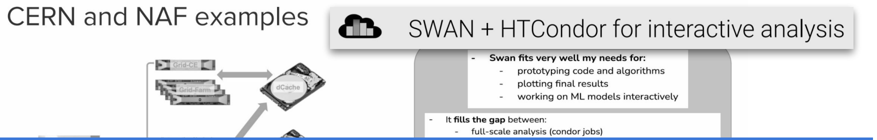


B. Bockelman, IRIS-HEP 200 Gbps challenge



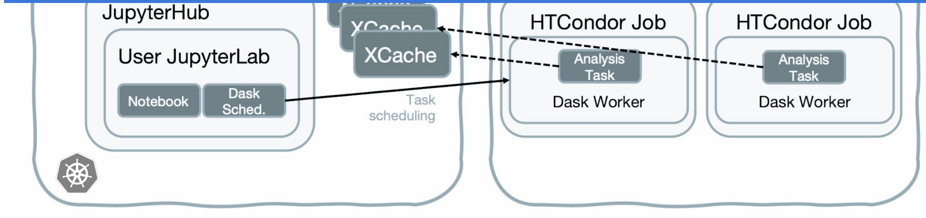
Fernando Barreiro Megino, Lukas Heinrich, KubeCon Oct 2022

Essentially we are designing and **building** platforms



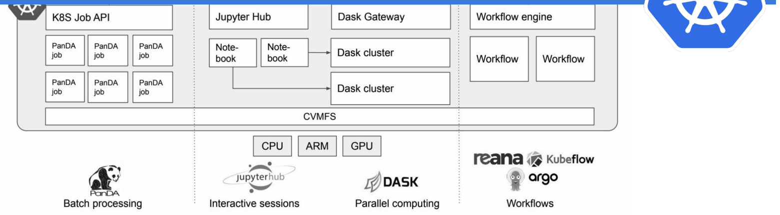
"Behind every great product is a great factory"

Solomon Hykes, Docker founder (ref)



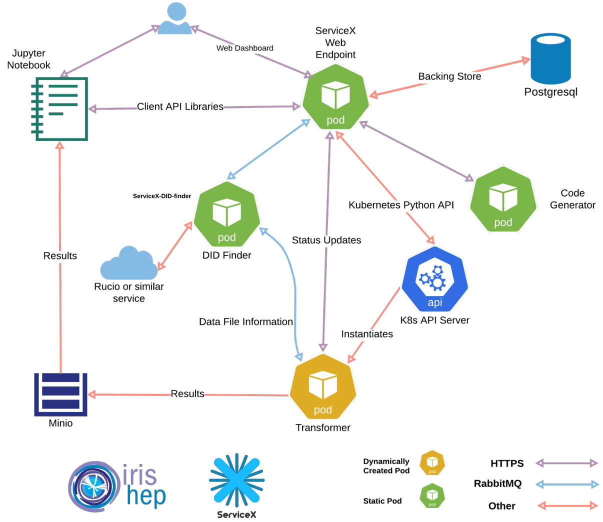
Continuously improve the factory and the product together

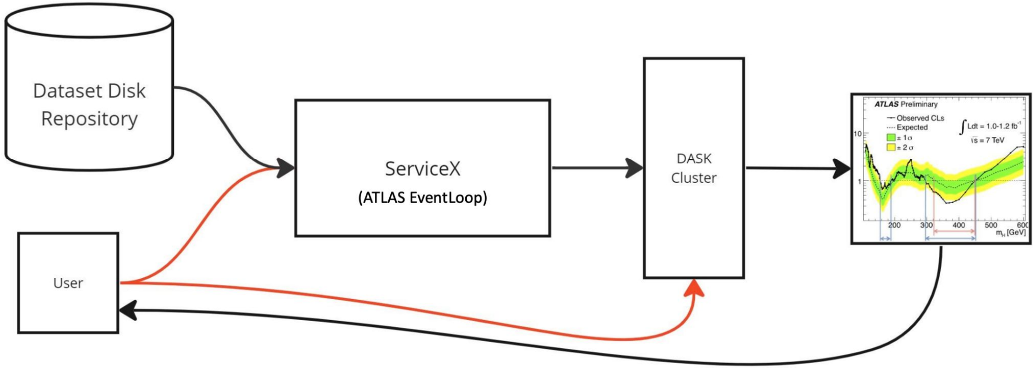
.. and leverage the platform building community
https://landscape.cncf.io/



Easier to deploy new capabilities in different places

- **Declarative, CI/CD tools** reduce the need for specialized expertise at every site
- **New types of capability** IRIS-HEP ServiceX data delivery & Coffea Casa, a Dask-based processing service



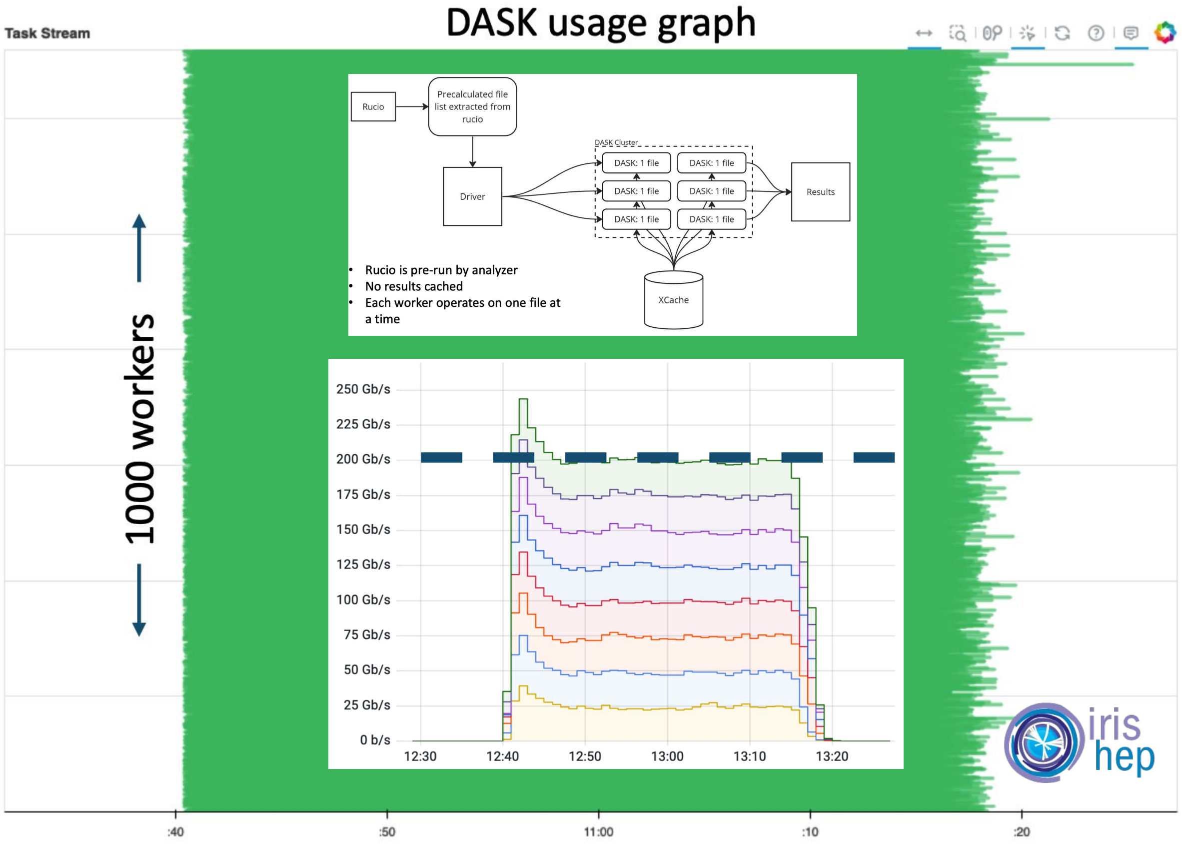


A capable K8s DevOps engineer (i.e. upskilled WLCG site admin) can deploy this in minutes!

Testing it out with Analysis Grand Challenges

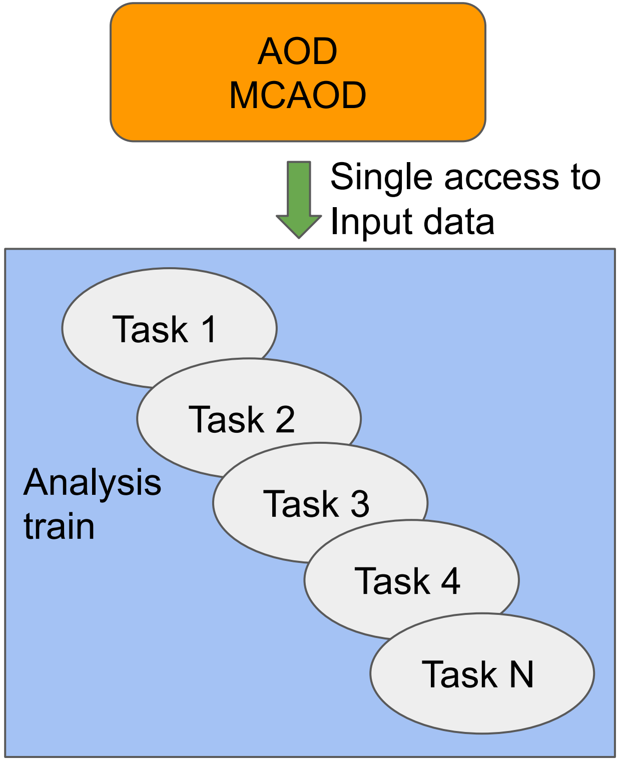
- "Grand challenge"
  - A framing device to focus effort and take stock of what is possible today
- IRIS-HEP, ATLAS, CMS
  - Analysis I/O at 25% HL-LHC scale
  - → 200 Gbps disk to CPU
- Identify facility bottlenecks and test the scalability of software

In the ballpark for a single user and lessons for a large multi-user facility



[Demonstrator Analysis 200 Gbps, WLCG/HSF Workshop, May 2024, B. Bockelman]

ALICE Analysis Facilities (AFs)

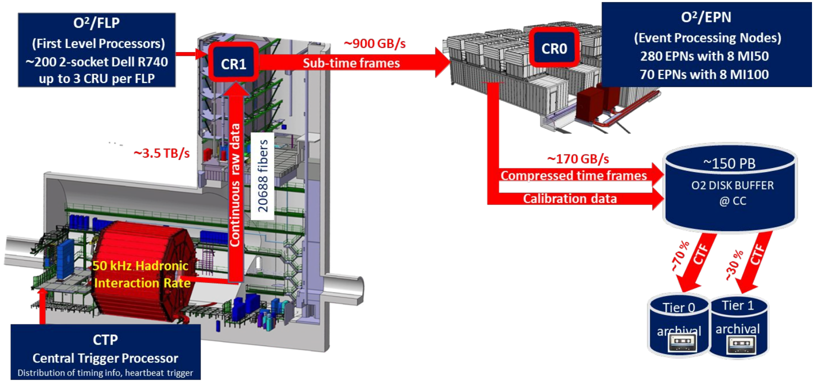


- Essential element of the Run3+ computing model
- Subset of data transferred to AF from T0/T1s/T2s
- Goals
  - Provide a location with comprehensive data samples from asynchronous and MC data processing at ~10% statistics
  - Fast tuning of analysis algorithms - once ready, run on full sample on the Grid
  - Analysis turnaround in less than 24 hours
  - First data and low statistics analysis (if compatible)
- Incorporated in the Grid framework
- Sites tuned for fast I/O between storage and CPU
  - Approximate total size 6-8k cores, 10PB storage
  - ~15MB/s/core throughput
- As of today - GSI Darmstadt, KFKI Budapest, LBNL Berkeley
  - 10 PB storage and 12,000 CPU cores

Technical Design Report for the Upgrade of the Online-Offline Computing System

ALICE O2 facility for Run 3 and 4:

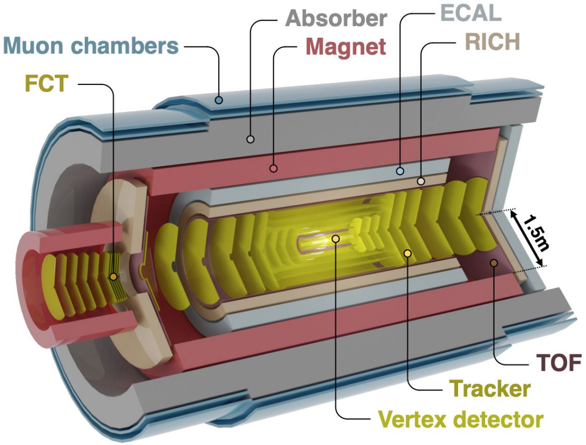
- Major upgrades during Long Shutdown 2
- No trigger for main detectors, continuous readout
- Store all Pb-Pb collisions up to 50 kHz interaction rate:
  - 3.5 TB/s raw detector data
  - Fast online compression during data taking leveraging heterogeneous architecture:
    - FPGAs in FLP: 3.5 TB/s -> 900 GB/s
    - GPUs in EPN: 900 GB/s -> 170 GB/s



See also Gabriele Cimador's Friday talk on GPU performance in Run3 ALICE online/offline reconstruction and Technical Design Report for the upgrade to the OnlineOffline Computing System for Run 3 and 4

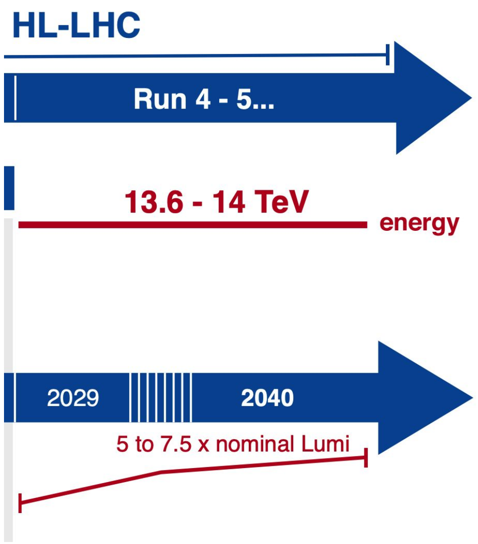
ALICE 3: novel and innovative detector concept

- Compact and lightweight all-silicon tracker
- Large acceptance with particle identification
  → without TPC, much lower data volume / event
- Continuous read-out and online processing
- Target interaction rates x2 in Pb-Pb and x50 in pp (24 MHz)
  → data throughput will be dominated by pp
  → Plan to maintain the online compression scheme of Run 3 and 4 in Run 5, leveraging technological speedup to handle higher pp rates



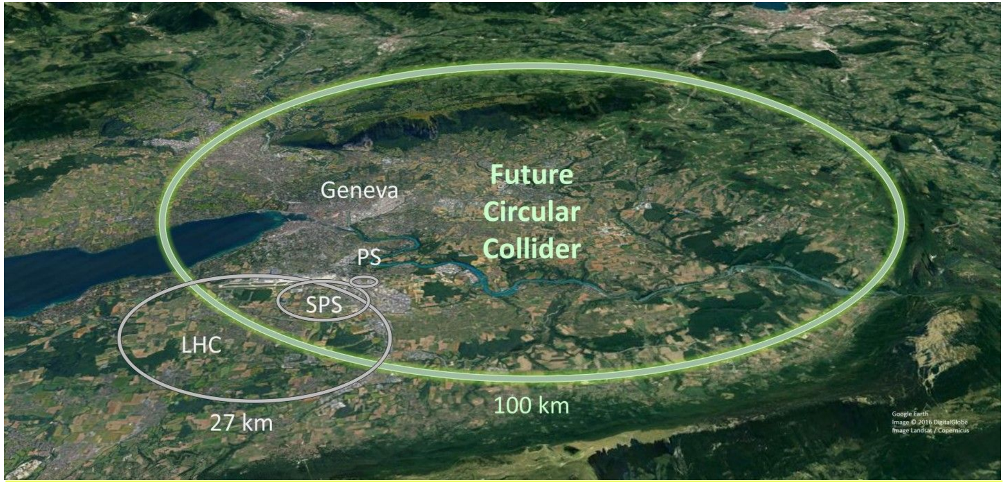
ALICE 3 LoI: CERN-LHCC-2022-009

… and beyond



Technology landscape in 20 years impossible to predict

Drivers today are the hyperscaler public clouds, GenAI. In 2040?



=> the most effective means of reducing risk is to invest in our workforce!!

Summary and Conclusions

- The HL-LHC presents a significant, but manageable computing challenge
  - In the midst of a rapidly changing technology landscape
- The experiments have launched vigorous **facility R&D** programs
  - to complement **software R&D** efforts
  - and give an added safety margin
- **Demonstrators and grand challenges** have been devised prove out capabilities
- **CERN Tier0, national Tier1s, Tier2s**
  - Planning computing infrastructure upgrades
- Will come into focus with Computing TDRs for HL-LHC in the next year

Early production deployments in Run3 will help guide the way!

"HL-LHC computing landscape in 2030 and beyond" Image source: ChatGPT-4o

backup

Networking R&D Activities

NOTED Monitor links & predict the behavior of applications. Integrate with file transfer services (FTS)

AutoGOLE/SENSE VPNs between routers to enforce paths, implement network QoS.

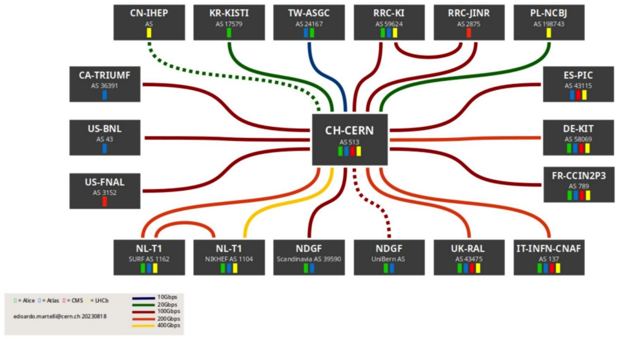
ALTO/TCN Application-Layer Traffic Optimization to obtain network information for long-term schedules (FTS/Rucio)

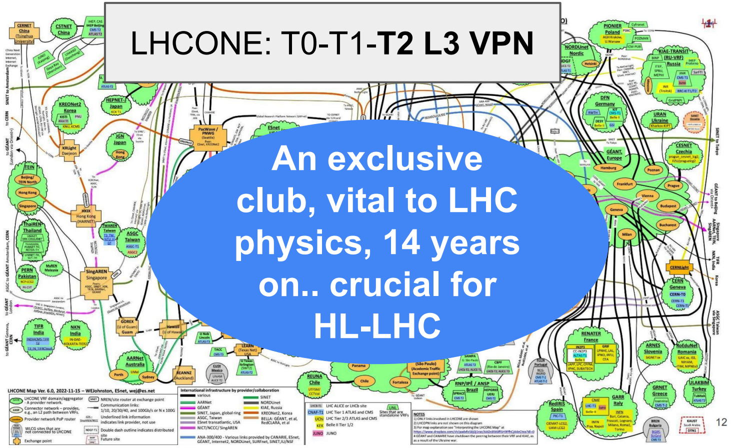
Packet marking Identify traffic at the network layer (by experiment and activity)

Network throughput studies Packet pacing & Jumbo frames

LHCOPN

Dedicated links from CERN to national centers





M. Babik, [Distributed Computing Challenges at the LHC and HL-LHC](), 4th GLOBAL RESEARCH PLATFORM WORKSHOP Oct 2023

Analysis Facilities for the HL-LHC

In the past few years much attention has been given to so-called *Analysis Facilities*.

- HEP Analysis Ecosystem 2017 workshop
- IRIS-HEP Analysis Systems R&D on Scalable Platforms 2019
- WLCG pre-CHEP 2019 workshop
- IRIS-HEP Future Analysis Systems 2020 blueprint workshop
- HSF Analysis Ecosystems II 2022 workshop
- WLCG pre-CHEP 2023 workshop
- WLCG/HSF May 2024 workshop

HSF Analysis Facilities Whitepaper was published in April covering:

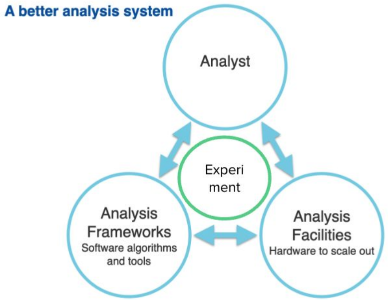
- User perspectives, compute and data access, consistency across infrastructures, continuous integration deployment and other features of current AFs

Yet significant questions remain:

- What are the use cases, analysis model differences Run 3 to Run 4, organization, benchmarks, and dedicated hardware needed

An LHCC review is forthcoming in 2024

"Arguing about the definition of analysis facility" Image source: ChatGPT-4o



A. Forti, WLCG/HSF May 2024

> The LHCC **recommends** that experiments engage in the process of developing and defining the structure of the future Analysis Facilities and requests they produce a document which defines the use cases in order to establish realistic benchmarks. This process should be coordinated with the HL-LHC Computing and SW review panel. The document is expected to be regularly updated in the process towards HL-LHC.

Follow-up on the Focus Session on Analysis Facilities held at the 154th LHCC in June 2023