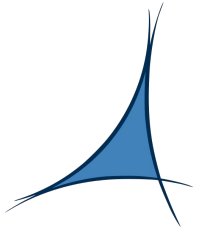


PIC
port d'informació
científica

Distributed Computing

COMCHA school, Zaragoza, April 8th 2026

G. Merino, PIC/CIEMAT

IFAE Institut de Física d'Altes Energies
Gobierno de España Ministerio de Ciencia e Innovación
Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas
Unió Europea Fons Europeu Next Generation
Gobierno de España
Plan de Recuperación, Transformación y Resiliencia
Next Generation Catalunya
Generalitat de Catalunya Departament de Recerca i Universitats

Outline

PIC
port d'informació
científica

How distributed computing came to be

The Worldwide LHC Computing Grid

Some successful Distributed Computing paradigms

An ever increasing need: the energy efficiency factor

Computing Models for large experiments

Summary

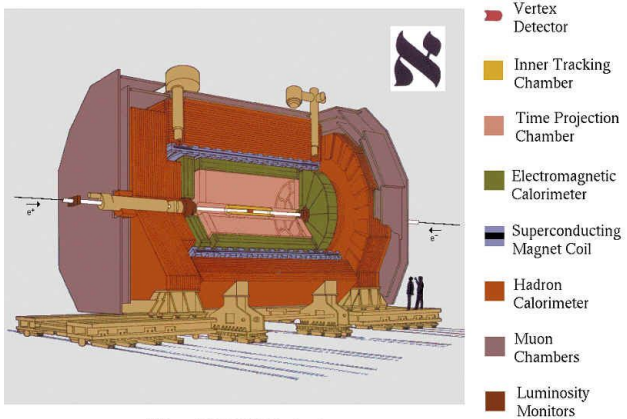
HEP computing in the 90s

PIC
port d'informació
científica

The ALEPH Experiment took data from the CERN e+e– collider LEP from 1989 to 2000

International Collaboration of ~500 scientists

The ALEPH legacy data is a **30TB dataset.**



The ALEPH Detector

Massive computing was needed to analize this data.

All **centralized** at CERN computing centre.

CERN computing centre in 1998

History: computing capacity at CERN

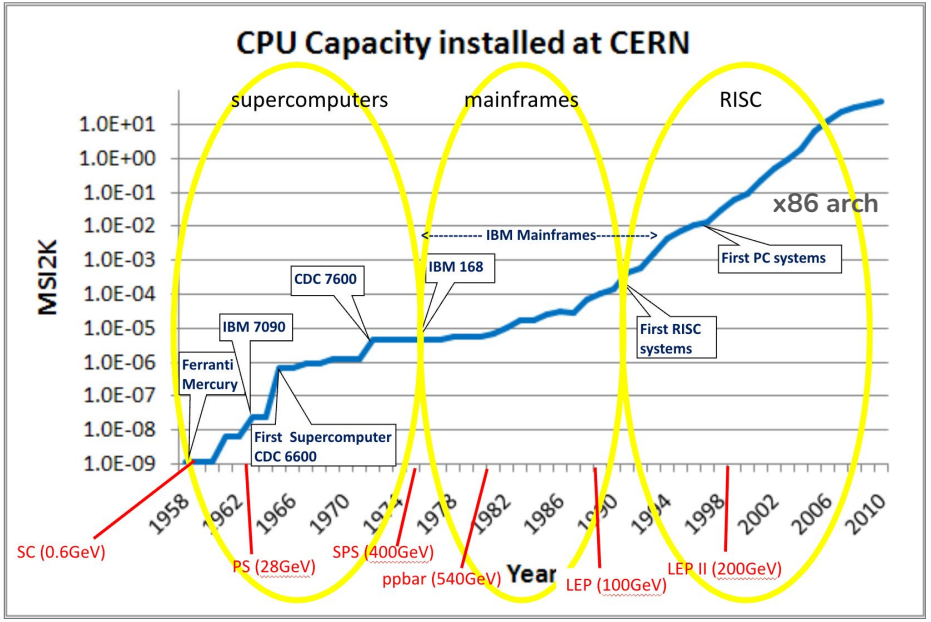
PIC
port d'informació
científica

1996: Intel Pentium Pro - start of the dominance of the x86 architecture, thanks to the massive PC (windows) market.

- Before that, mostly UNIX workstations with RISC processors (Digital Alpha, Sun SPARC, HP PA-RISC ...)

10 million SI2K in 2006

- Equivalent to ~2000 CPU cores today (5-10 CPU servers)



Source: L. Robertson, https://indico.cern.ch/event/1525790

Distributed infrastructure for the LHC

PIC
port d'informació
científica

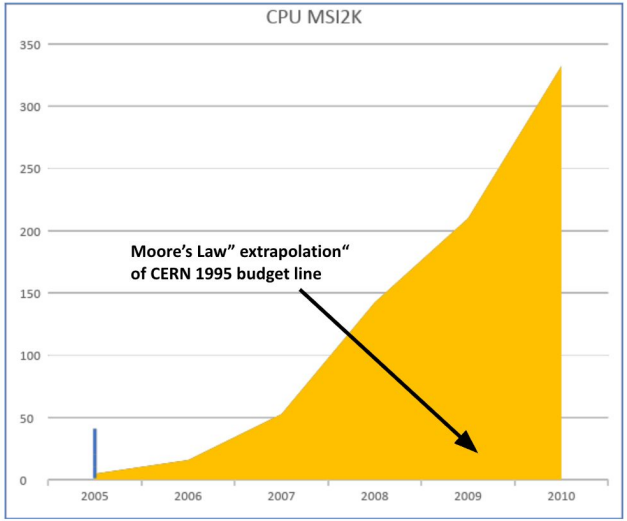
LCG TDR, 2005:

- The LHC (expected to start in 2007) would require an order of magnitude more computing than the CERN computing budget would support.

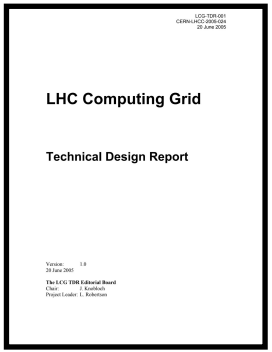
CERN would only be able to provide 10–20% of the total needed capacity for the LHC

⇒ a distributed infrastructure was **mandatory**.

Estimated requirements for the LHC offline computing



Source: L. Robertson, https://indico.cern.ch/event/1525790


LHC Computing Grid

Technical Design Report

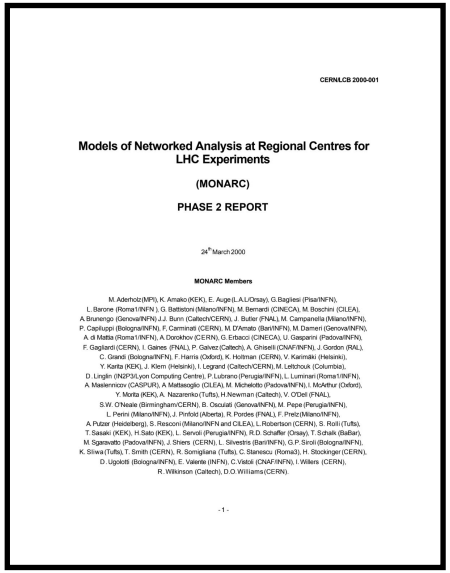
The MONARC model

PIC
port d'informació
científica

In 1997, the MONARC project was set up to find a solution for addressing the 10x increase in LHC computing needs.

Key conclusions:

- A distributed hierarchical model (CERN - Tier-1s - Tier-2s) is feasible .
  - Network requirements: ~30 MBytes/s (622 Mbps) needed between CERN and each Tier-1 for data replication and analysis.
- Willingness in many national communities to participate in developing, funding and deploying this model.
- Integration with GRID will require middleware development (job/data management, resource allocation).
  - Expect GRID projects to provide these fundamental tools in coming years.



CERN/LCB 2000-001

Models of Networked Analysis at Regional Centres for LHC Experiments

(MONARC)

PHASE 2 REPORT

24th March 2000

MONARC Members

M. Aderholz(MPI), K. Amako(KEK), E. Auge(LAL/Orsay), G.Bagliesi (Pisa/INFN), L. Barone (Roma1/INFN), G. Battistoni (Milano/INFN), M. Bernardi (CINECA), M. Boschini (CILEA), A.Brunengo (Genova/INFN),J.J. Bunn (Caltech/CERN), J. Butler (FNAL), M. Campanella (Milano/INFN), P. Capiluppi (Bologna/INFN), F. Carminati (CERN), M. D'Amato (Bari/INFN), M. Damen (Genova/INFN), A. di Mattia (Roma1/INFN), A. Dorokhov (CERN), G. Erbacci (CINECA), U. Gasparini (Padova/INFN), F. Gaglardi (CERN), I. Gaines (FNAL), P. Galvez (Caltech), A. Ghiselli (CNAF/INFN), J. Gordon (RAL), C. Grandi (Bologna/INFN), F. Harris (Oxford), K. Holtman (CERN), V. Karimäki (Helsinki), Y. Kanta (KEK), J. Klem (Helsinki), I. Legrand (Caltech/CERN), M. Letchcuk (Columbia), D. Linglin (IN2P3/Lyon Computing Centre), P. Lubrano (Perugia/INFN), L. Luminari (Roma1/INFN), A. Maslennicov (CASPUR), A. Mattaeoglio (CILEA), M. Michelotto (Padova/INFN), I. McArthur (Oxford), Y. Morita (KEK), A. Nazarenko (Tufts), H.Newman (Caltech), V. O'Dell (FNAL), S.W. O'Neale (Birmingham/CERN), B. Osculati (Genova/INFN), M. Pepe (Perugia/INFN), L. Perini (Milano/INFN), J. Pinfold (Alberta), R. Pordes (FNAL), F. Prelz(Milano/INFN), A.Putzer (Heidelberg), S. Resconi (Milano/INFN and CILEA), L. Robertson (CERN), S. Rolli (Tufts), T. Sasaki (KEK), H.Sato (KEK), L. Servoli (Perugia/INFN), R.D. Schaffer (Orsay), T. Schalk (BaBar), M. Sgaravatto (Padova/INFN), J. Shiers (CERN), L. Silvestris (Bari/INFN), G.P. Siroli (Bologna/INFN), K. Sliwa (Tufts), T. Smith (CERN), R. Somigliana (Tufts), C. Stanescu (Roma3), H. Stockinger (CERN), D. Ugolotti (Bologna/INFN), E. Valente (INFN), C.Vistoli (CNAF/INFN), I. Willers (CERN), R. Wilkinson (Caltech), D.O. Williams (CERN).

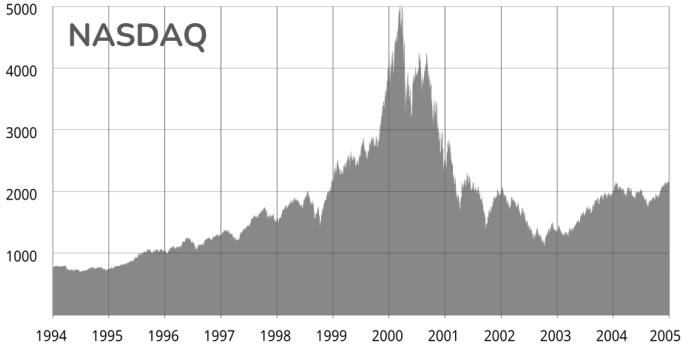
- 1 -

Dot-com bubble and network availability

PIC
port d'informació
científica

Late 90s: rapid market growth fueled the widespread adoption of the World Wide Web and the Internet.

- Investors poured billions into telecom companies, racing to lay thousands of Km of fiber-optic cable across the globe, assuming that internet traffic would double every 100 days, forever.

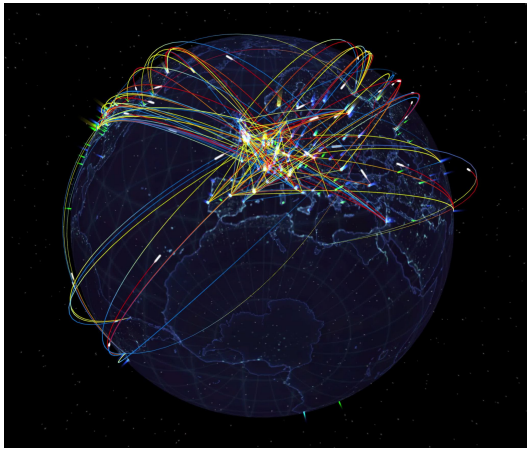


Source: Wikipedia

2001 - The dot-com bubble busted, and the world became hyper-connected.

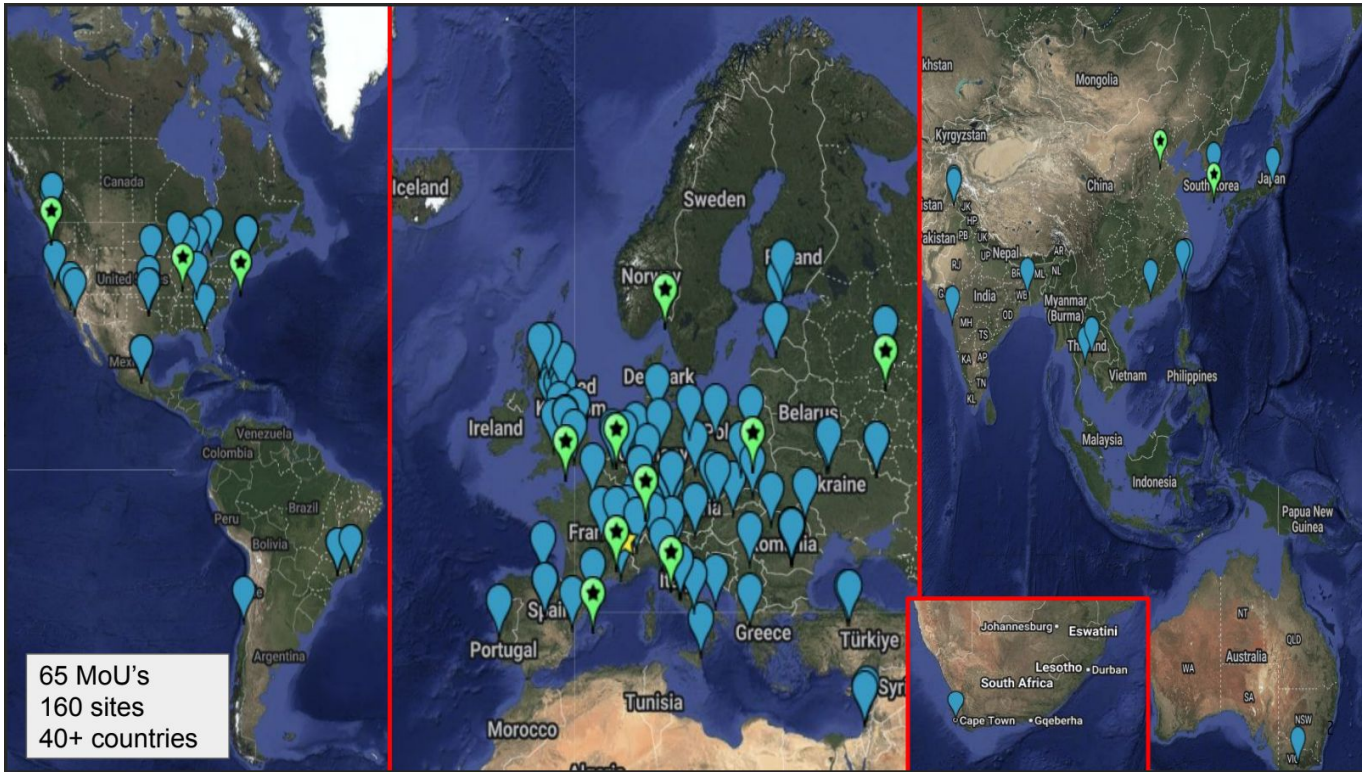
- By 2002, it was estimated that only 1% to 3% of the fiber optic cable buried in the ground was actually being used

The availability of WAN for the LHC was much higher than originally expected.



The Worldwide LHC Computing Grid today

PIC
port d'informació
científica



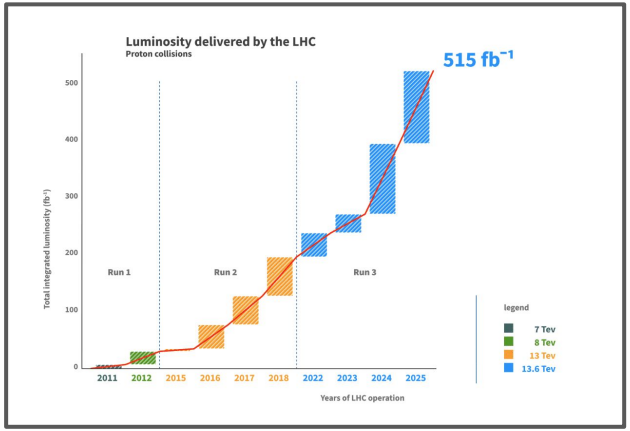
Source: T. Boccali, LHCC Open session Nov 2025

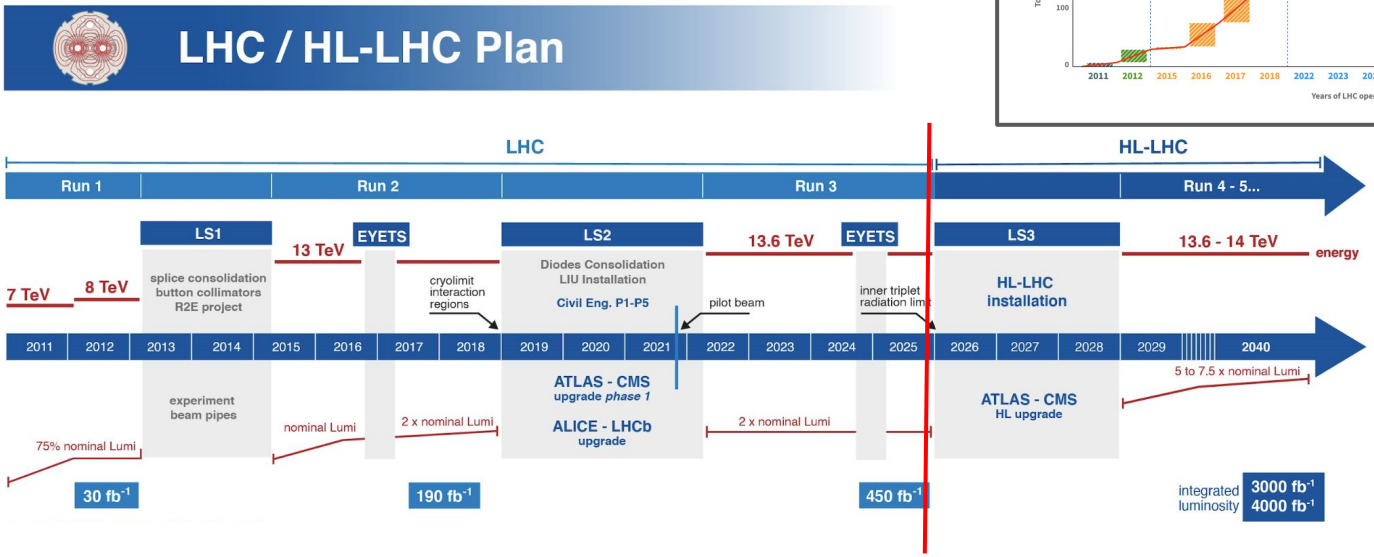
The LHC program: 2010-2026

PIC
port d'informació
científica

Three runs with increasing energy and luminosity

Two Long Shutdown periods for upgrades





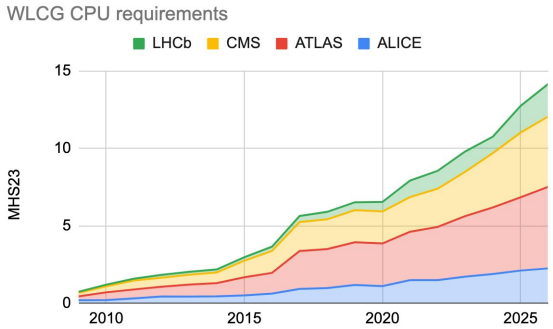
WLCG Capacity

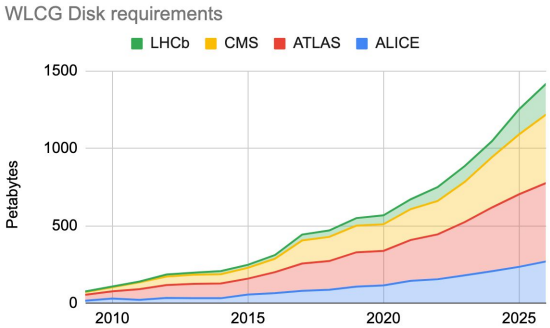
PIC
port d'informació
científica

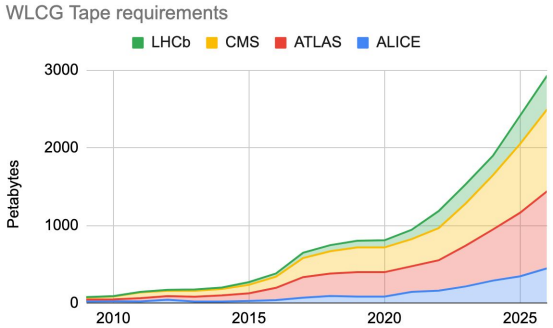
The WLCG system has continuously increased its capacity, following the requirements of the experiments.

Computing does not stop even if experiments are not taking data.

- Continuous simulation and re-processing activities







Distributed Computing Paradigms

PIC
port d'informació
científica

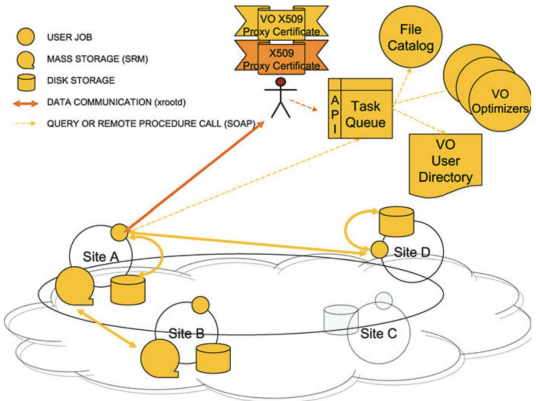
1. Workload Management: the pilot job model (“pull” paradigm)
2. Sotware Distribution: CVMFS success story
3. Data Distribution and Access: push vs. pull paradigms
4. Data scaling: tape vs. disk

Pilot jobs and the pull paradigm

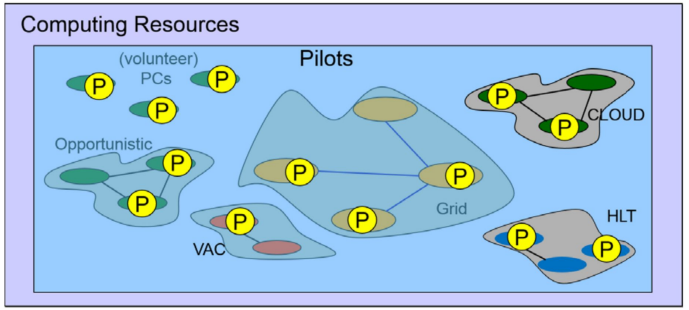
The LHC experiments developed a WMS architecture for Grid computing to enhance resiliency against unstable execution nodes.

Unlike traditional Grid middleware that "pushed" tasks, the pilot job model "pulls" them.

- Experiments deploy "pilot jobs" (simple placeholders)
- After landing on a worker node, the pilot inspects the environment, contacts a central Task Queue, and pulls a real workload for just-in-time scheduling



Source: F. Carminati & P. Buncic DOI 10.1007/978-3-642-23157-5



Pilots form an overlay layer hiding the underlying diversity

Source: DOI 10.1088/1742-6596/898/9/092024

CERN VM Filesystem - CVMFS

PIC
port d'informació
científica

The Problem: distributing massive sw stacks to 10.000s of computers worldwide.

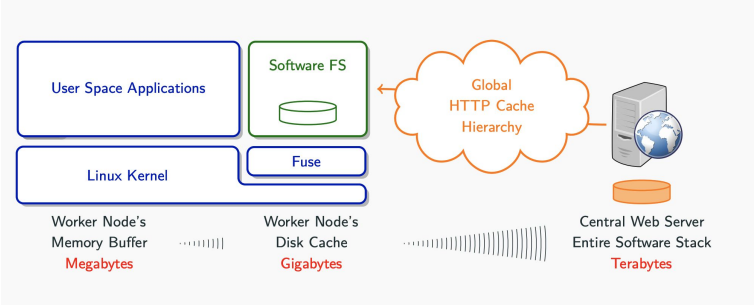
- Original method of "installation jobs": sw was pushed to every site via the Grid.

CVMFS: Originally developed at CERN ~ 2007 for the CernVM project to support running in the cloud and avoid “fat” VM images.

- Adopted by all LHC experiments for sw distribution ahead of LHC Run 1.

Today, over 5 billion files distributed globally (1.5 PB unique, compressed data)

- Allows 1000s of developers to test nightly builds without overwhelming the network



Source: J. Blomer, CERN

Ingredients:

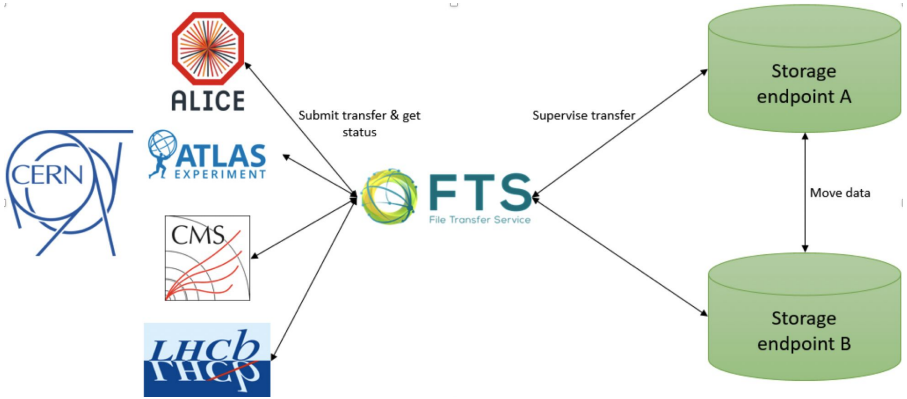
- Single point of publishing
- Content-addressable storage
  - Automatic de-duplication
  - Data integrity
- HTTP transport
- Caching on all levels

File Transfer Service

PIC
port d'informació
científica

FTS is a bulk data movement service

- Responsible for moving files from one site to another efficiently
- Distributes the majority of LHC data across the WLCG infrastructure



https://fts.web.cern.ch/fts/

Key Features:

- Robust: Checksums and retries provided per transfer
- Intelligent Scheduling:
  - Runtime optimization (file concurrency and nr. streams)
  - Fair-share based on activities and roles
- Flexible: Multiprotocol support
- Easy to use: REST API and web monitoring

ATLAS Rucio: data management at scale

PIC
port d'informació
científica

Rucio uses FTS for file movement but adds crucial data lifecycle management functionality:

- **Rule-Based Replication:** Daemons enforce rules (e.g., "keep 5 copies in Europe"), triggering transfers if sites fail.
- **Data Popularity:** Tracks usage. Dynamically replicates highly accessed data and moves unused data to cheaper Tape storage.
- **Network Information:** Interacts with the FTS & PerfSONAR to monitor the current network status of the grid.
  - Throughput Aware: Reroutes data or waits if links are congested or down.
  - Cost-Aware: Prioritizes cheaper links (e.g., research network over commercial cloud) unless data is high-priority.
  - SDN Integration: Uses SDN to request dedicated high-bandwidth "circuits" for massive transfers without affecting general internet traffic.

CMS Any data, Anytime, Anywhere

PIC
port d'informació
científica

Data access paradigm to break the "data locality" rule that dominated Grid computing.

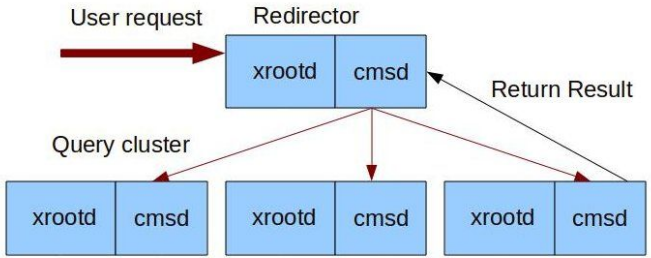
- Traditionally, a computing job was sent to the site where the data is physically stored. If it was busy or the storage was down, the job waited in the queue.
- AAA: a job running at Site A could transparently stream data from Site B over the network.

Creates a global namespace of all data on disk. The data looks as if it is in the local HD.

Common use cases: fallback mechanism (non-available data), opportunistic CPUs.

Relies on the Xrootd system:

- Global redirectors
- Xrootd protocol: optimized for high-latency WAN links (readahead, parallel streams)



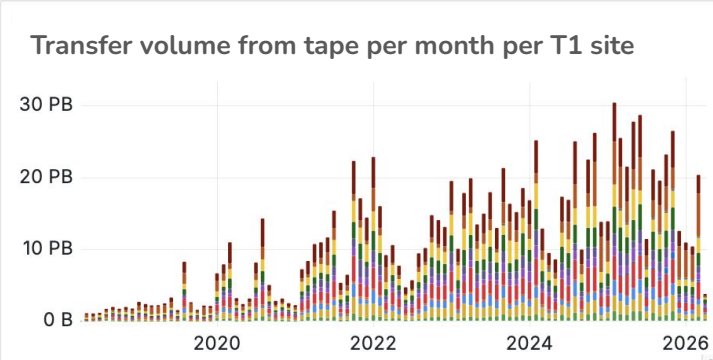
Exploiting tape: ATLAS Data Darousel

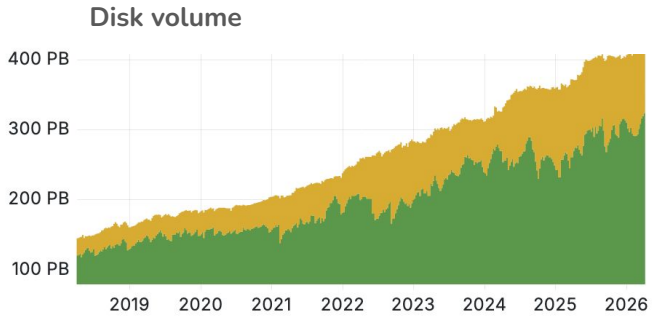
PIC
port d'informació
científica

A tape-driven workflow system that orchestrates data access directly from tape storage for production jobs, minimizing reliance on expensive disk space.

- Started as an R&D project in 2018, it integrates ATLAS workload management (PanDA/ProdSys2), data management (Rucio), and tape systems at Tier-0/Tier-1 sites.

Also useful to enable stringent policies for removing unused data from disk. ATLAS keeps below 4% its data fraction that has not been read in 12 months.

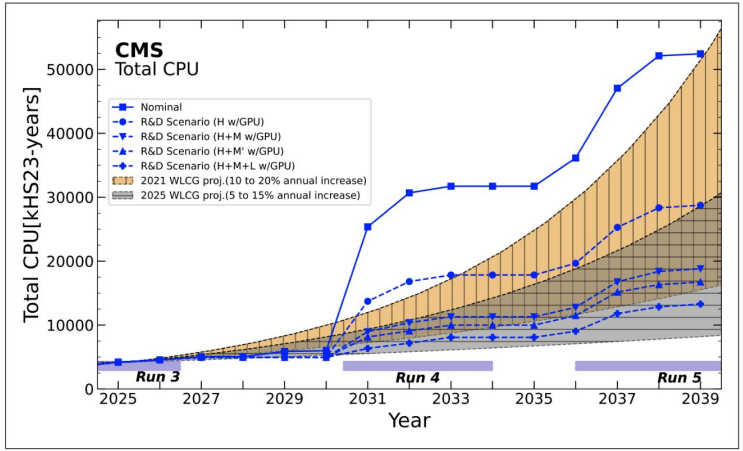




Projections for HL-LHC

Both ATLAS and CMS are expecting a massive data rate increase for Run4. Computing needs were estimated in 2022, and this year will be updated and formally reviewed.

CMS recently published their updated estimations in a Conceptual Design Report detailing their R&D plans to improve their current model and meet the higher needs.



https://cds.cern.ch/record/2957472

Projections for several “R&D intensity” scenarios.

Compared to “flat-budget”

- 10-20%/yr (orange) still there, although it was updated to 5-15% (grey) last year.
- Recent (6 months ago) hardware “cost explosion” due to the AI-boom still not considered - a looming risk…

A clear message emerges: **HL-LHC is not a solved computing problem. Sustained R&D is mandatory.**

Energy Consumption

Data centers are a growing concern in the clean energy transition due to their significant energy consumption.

Data centers currently consume about 1.5% of global electricity, rising by 12% annually.

Projections show this consumption is expected to more than double to 945 TWh by 2030.

The main driver today is the energy demands of accelerated computing for AI.

European Commission

EN

Energy, Climate change, Environment

Energy

Home | Topics | Strategy | Data and analysis | Energy explained | Resources | Events | News

Home > News > In focus: Data centres – an energy-hungry challenge

In focus: Data centres – an energy-hungry challenge

NEWS ARTICLE

17 November 2025 — Directorate-General for Energy — 4 min read

©AdobeStock/(left)Spia100/(Right)ConnectWorld

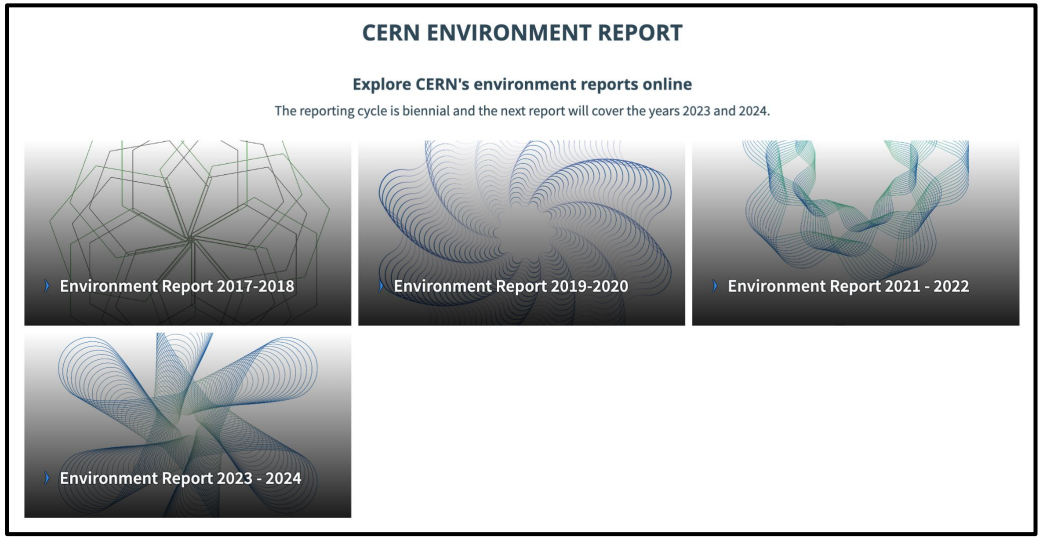
Data centres are a vital infrastructure supporting our ever-growing use of cloud storage, social media, AI, streaming services and more.

They're also an increasingly hot topic of the clean transition, as they consume significant amounts of energy.

Energy Consumption

PIC
port d'informació
científica

Large labs now regularly publish environment reports, with measurements, metrics and goals of improvement. CERN is doing it since 2019, every 2 years:



https://hse.cern/environment-report

LHC computing energy needs

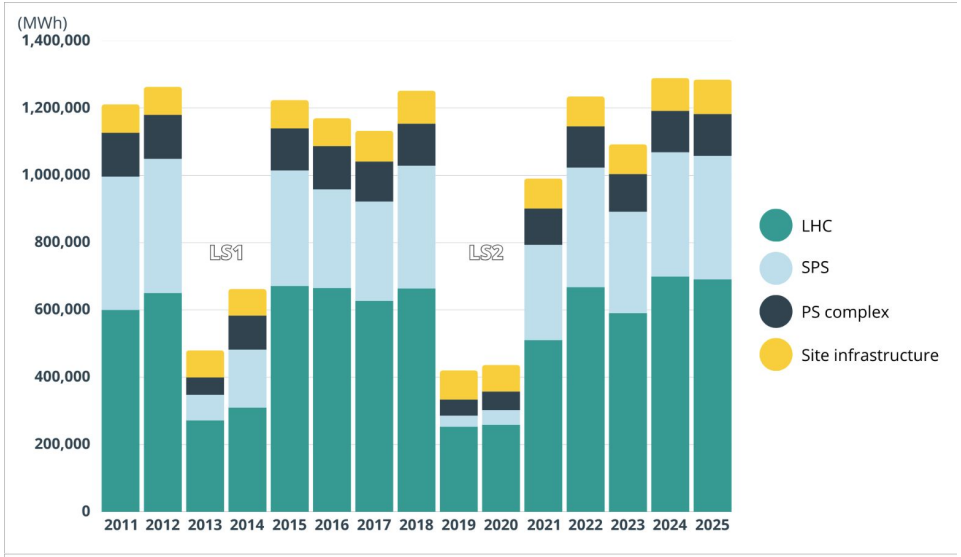
PIC
port d'informació
científica

The CERN energy consumption is ~1.25TWh/year during LHC runs

The CERN IT infrastructure contributes <5% in average

CERN provides ~20% of the WLCG resources

In terms of energy needs for the LHC program, the impact of computing is non negligible.



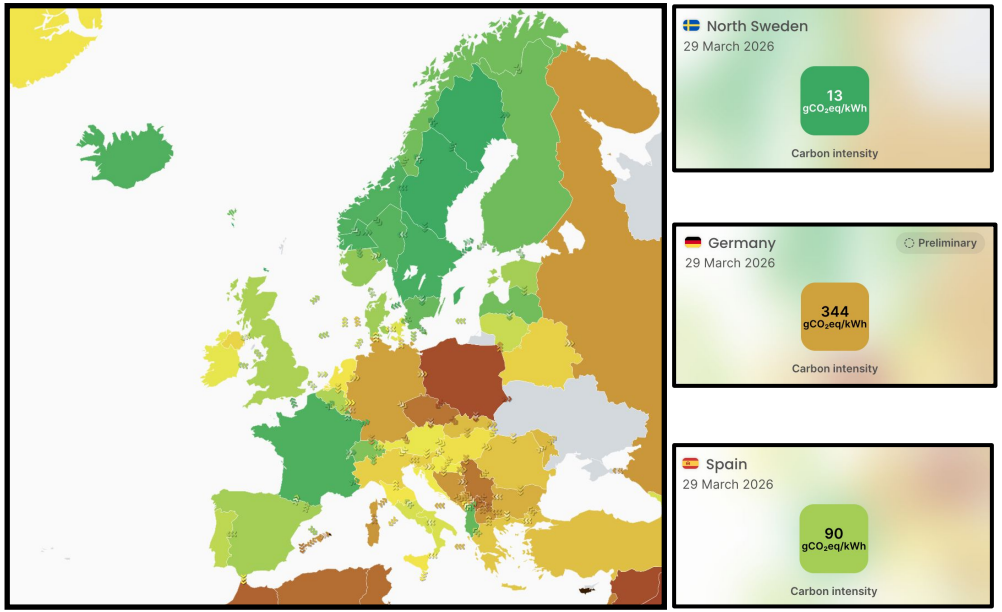
CERN's electricity consumption 2011-2025.

https://hse.cern/content/energy-management

Carbon footprint

The carbon footprint of electricity consumption depends on the specific mix in every country

- Quite small at CERN, since electricity comes from France, mainly nuclear.



https://app.electricitymaps.com/

Interesting trade-off:

Recall that distributed infrastructure was necessary for distributed funding.

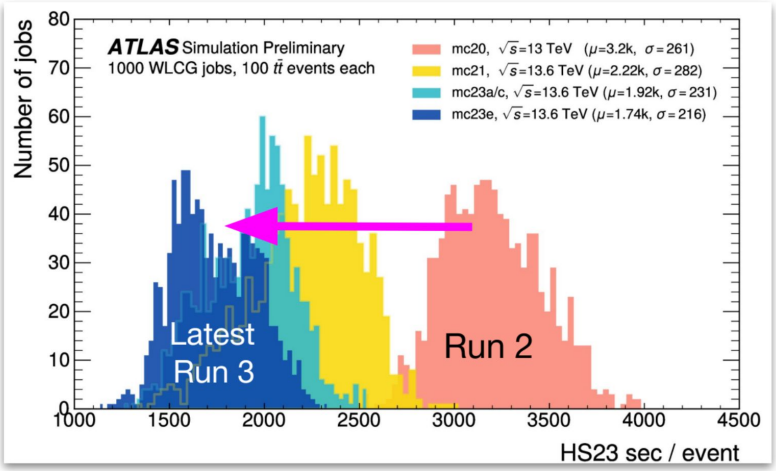
Now, to minimize carbon footprint, should we run all jobs in Sweden&Norway?

Methods to improve energy efficiency

PIC
port d'informació
científica

Software performance engineering: Reducing an application's "time-to-solution" is the first step toward efficiency.

- Using performance analysis tools to identify bottlenecks in communication, computation, and I/O.



ATLAS full simulation is 30% of the total CPU consumption (the largest single consumer).

Significant improvement in FullSim:

- Average time now 1800 HS23s/event, compared to 2900 HS23s/event at the start of Run 3

Methods to improve energy efficiency

PIC
port d'informació
científica

Using newer CPUs: energy efficiency is improved in new generations of CPUs.

- On average, a 50% improvement spec/Watt in 5 years has been measured.
- On the other hand, servers' embodied carbon (emissions during manufacturing, shipping, materials) can be as high as operating emissions for 4-5 years.

Using GPUs: highly parallelizable algorithms can benefit from specialized accelerators.

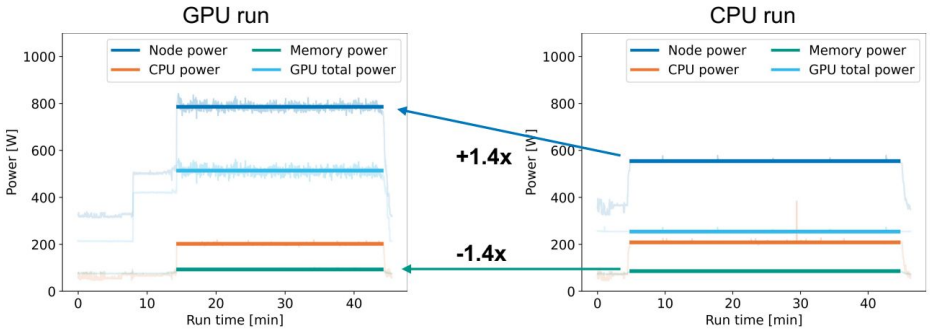
- Matrix Element Evaluation in event generators are very well suited for parallelization
  - The PEPPER program has achieved speed-up factors between 10 and 100
    (Similar developments also underway for MadGraph5_aMC@NLO)
- Reconstruction: most LHC experiments use GPUs in their HLT, processing from +50% to +100% more events per kwh.

Methods to improve energy efficiency

Full simulation: the activity where experiments spend most of their CPU. A fundamentally stochastic process, so it is not great for GPU parallelization.

- Still, a big effort is underway to port parts of GEANT4 to GPUs: AdePT/Celeritas for electromagnetic showers.

Energy-efficiency of Athena on Perlmutter



Source: Severin Diederichs, WLCG GPU workshop Dec'25

First dedicated measurements indicate that the speed-up is similar to the increase in energy.

A speed-up of 2.6 would be needed to reach effective energy savings per event.

- And this is still without comparing the purchasing cost of the GPU vs CPU …

Computing models for large experiments

PIC
port d'informació
científica

LHC experiments pioneered large-scale distributed computing models. Several other physics experiments with massive computing needs are now developing their own models. While each is unique, all leverage common WLCG infrastructure.

Experiment	Physics Domain	Expected Raw Data (Annual)	Operational Status
LHC	Particle Physics	~50 PB	Currently Running (Run 3)
HL-LHC	Particle Physics	~400 PB	Expected 2031
SKAO	Radio Astronomy	~700 PB	Full array for 2029-2030
CTAO	Gamma-ray Astronomy	~10 PB	Full array for 2026-2030
ePIC (at EIC)	Nuclear Physics	~70 PB	Expected 2030s
DUNE	Neutrino Physics	~30 PB	First neutrino beam 2031
LVK (LIGO/Virgo/Kagra)	Gravitational Waves	~2 – 3 PB	Currently Running (O4/O5)
Vera Rubin (LSST)	Cosmology	~7.3 PB (20 TB/night)	First Light ~2025

The ePIC detector at EIC

ePIC Echelons are similar concept to LHC Tiers.

- Somewhat more specialized, and hierarchical.

Minimal nr. of Echelon1s:

- Just two, at the host Labs.

Echelon2s:

- sites pledging resources through a formal process: Resources Review Board.

The overall model, streaming DAQ to global processing

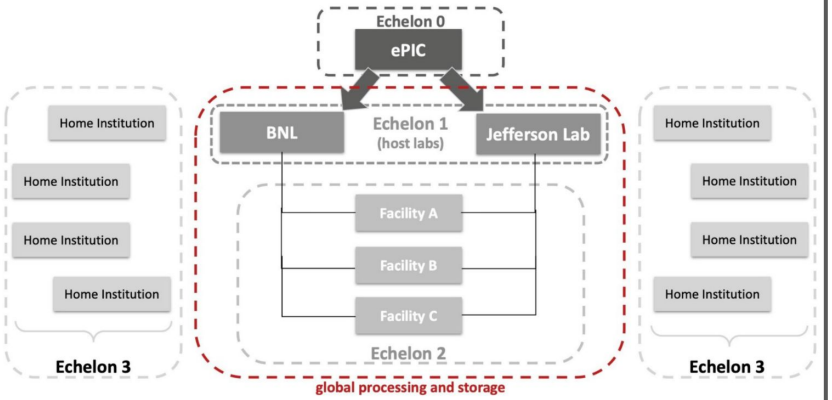
Four tiers:

Echelon 0:
ePIC experiment

Echelon 1:
Two host lab facilities

Echelon 2:
Global processing and data management

Echelon 3:
Home institutes:
where the analyzers are



Source: T. Wenaus, ePIC S&C Apr 2024

Rucio will be used for data transfer&replication between Echelon1-Echelon2

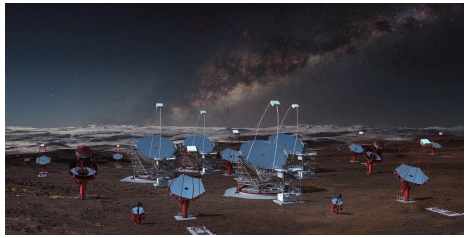
CTA Observatory

Two (remote) sites, sending the data to (just) four data centers.

- 6 PB/yr of data per site

Four data centers will provide data redundancy and services replication for high-availability.

Rucio will be used to manage data replication among DCs and for data transfer from the arrays.



SKA Observatory

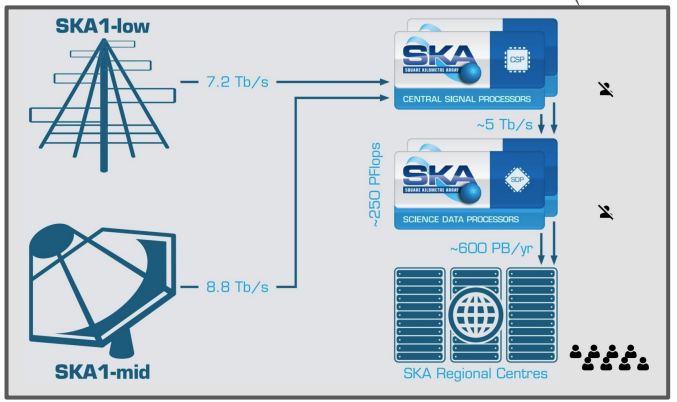
Two supercomputers to do the prompt reconstruction

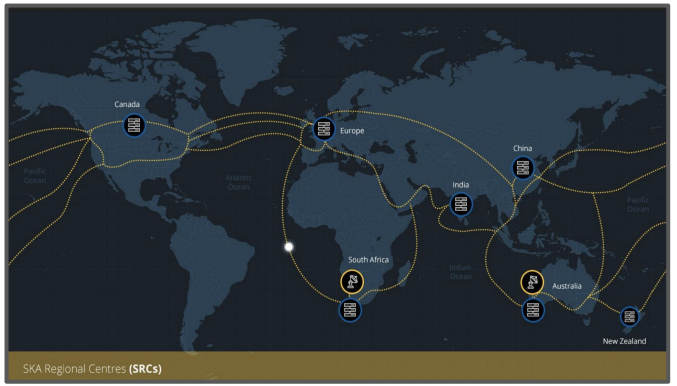
700 PB of data sent to the SKA Regional Centers (SRCNet) - distributed model.

Use Rucio for distributing data to the SRCNet

Users will run their analysis at SRCs

- No data download
- Send the jobs to the data





LIGO/VIRGO/Kagra

PIC
port d'informació
científica

Joint operation of four gravitational wave interferometers.

International collaboration of over 2500 scientists.

2-3 PB of RAW data annually, dominated by 100.000 auxiliary monitoring channels.

- Calibrated strain (analysis) data just 10TB/yr
- Custom Apache/kafka based data streaming service for low-latency (alert) analyisis

Common tools with LHC:

- Rucio for replicating data from detectors to main data centers: Caltech (US) & CC-IN2P3 (FR)
- CVMFS/xrootd used to distribute strain data to ~20 analysis sites
- HTCondor pilot jobs used to distribute offline workload.

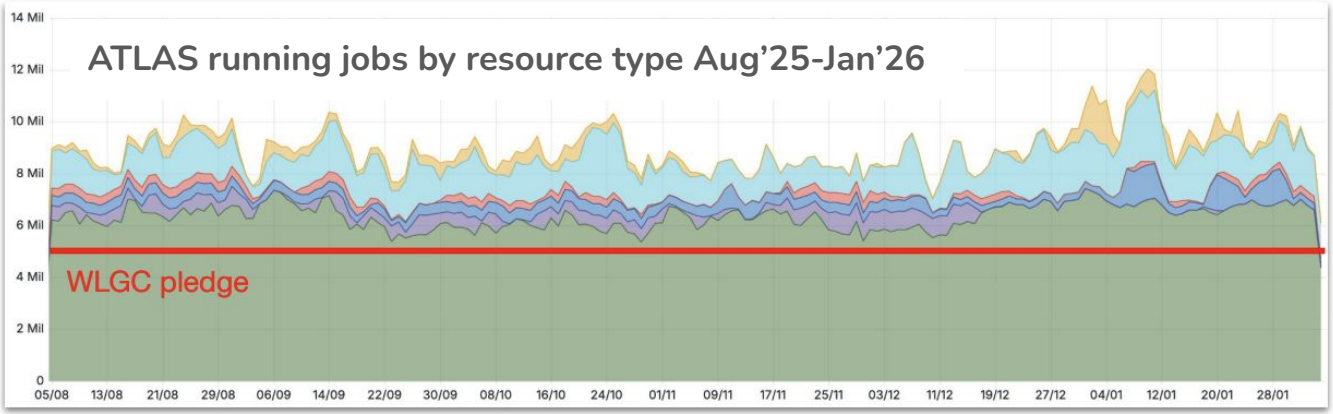
Integration of Supercomputers

PIC
port d'informació
científica

HPC systems have traditionally been developed as standalone systems, to maximize peak performance - a “parallel world” to distributed computing.

This trend is now changing and both worlds are converging and aiming for integration.

- Research Infrastructures get larger and demand more computing
- AI workflows & datasets rapidly growing, demanding a truly Federated AI Factories ecosystem



Name	Mean	Max
GRID	6.32 Mil	7.43 Mil
HPC	1.41 Mil	2.77 Mil
Perlmutter + TACC	269 K	1.81 Mil
HLT farm	434 K	1.39 Mil
Tier-0	319 K	759 K
BOINC	189 K	406 K

Commercial Clouds - a new computing model?

PIC
port d'informació
científica

ATLAS R&D project 2020-2023 to evaluate integration with Google Cloud Platform

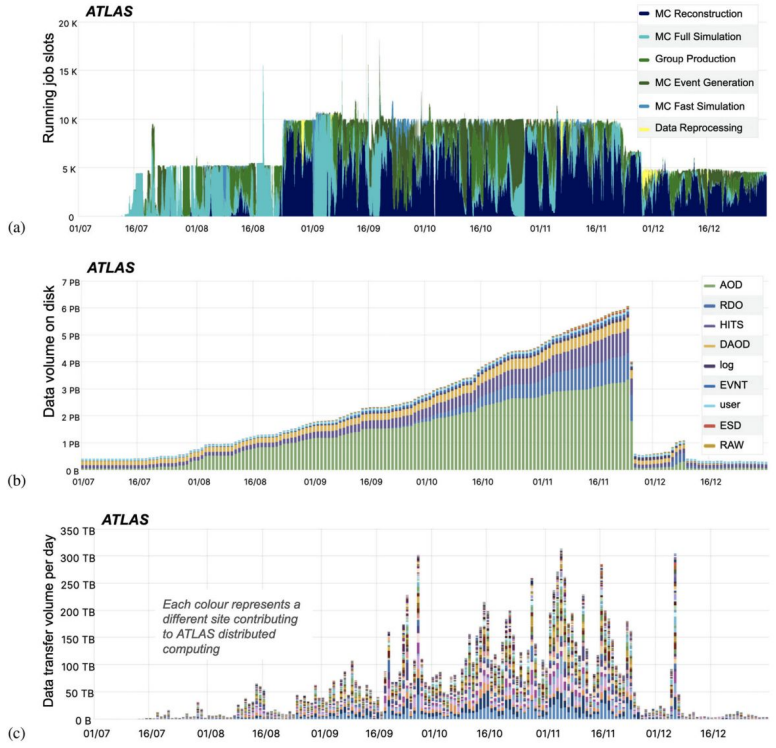
- integration of PanDA, Rucio with GCP
- demonstration of GCP for ATLAS analysis

"Production" phase: running a Cloud site of ∽2.5% size in production for about 1 year.

- GCP fixed-price "subscription agreement": unlocking [a]70% discount w.r.t. list prices

Conclusion: integration is feasible, unlocking Cloud elasticity.

- BUT: Is cost predictable? Is this model feasible for international collaborations?



Summary

PIC
port d'informació
científica

Distributed computing was born out of necessity

- CERN could only fund 10–20% of LHC computing needs, forcing the creation of the WLCG: a hierarchical, globally distributed infrastructure now spanning 160 sites in 40+ countries.

Key paradigms emerged and proved their worth

- The pilot job model, CVMFS for sw distribution, and data management systems like Rucio and FTS solved core challenges of running heterogeneous, large-scale workloads reliably across the grid.

HL-LHC computing is an unsolved problem

- the tenfold data rate increase expected for Run 4 cannot be met by hardware improvements alone; sustained R&D in algorithms and heterogeneous architectures exploitation is mandatory.

Energy efficiency is becoming a first-class concern

- data centers consume a growing share of global electricity, and the physics community must actively pursue software optimization, hardware modernization, and carbon-aware scheduling.

The LHC computing model is being adopted (and developed!) broadly

- experiments like SKAO, CTAO, LIGO/Virgo, DUNE and others are building on WLCG tools