# Computing Infrastructure at LHC Run 3 and HL-LHC

## XVI CPAN Days

Miguel Villaplana
Institut de Física Corpuscular (IFIC)
on behalf of the WLCG-ES sites

November 19, 2024

# Worldwide LHC Computing Grid (WLCG)

**The challenge: data scale**
The LHC detectors generate an enormous amount of data that needs to be stored, distributed across hundreds of scientific institutions worldwide and then analysed by thousands of scientists.

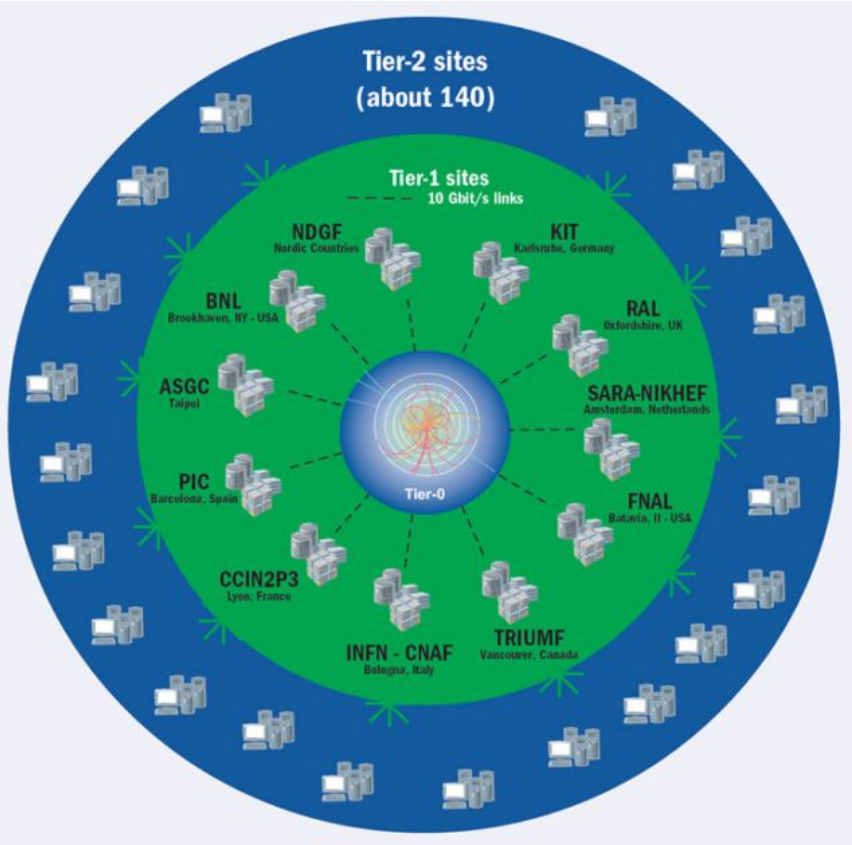
- +80 Petabytes transferred monthly.
- ~700 million computing jobs/year, consuming 5 billion CPU hours.

**Global Collaboration:**

- 42 countries involved.
- 170 computing centres.
- Over 1 million computer cores and 2 exabytes of storage.

**A key tool for physics:**

- Seamless access to data storage, processing power, sensors, and visualization tools.
- The most sophisticated data-taking & analysis system ever built for science, providing near real-time access to LHC data.

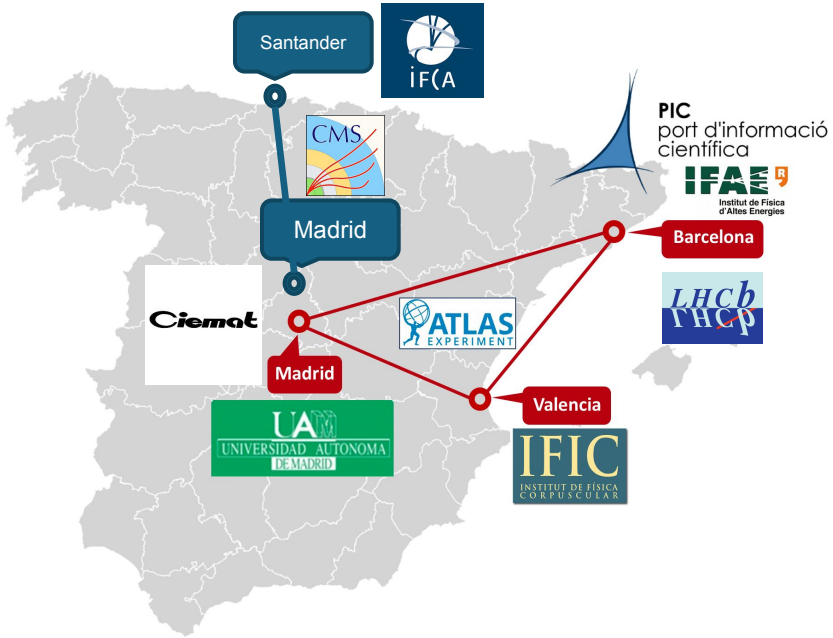


# Spanish computing sites for the LHC (WLCG-ES sites)

Tier 1 (ATLAS, CMS, LHCb): PIC-Barcelona

Federated Tier2s:

- 60% IFIC-Valencia
- 25% IFAE-Barcelona
- 15% UAM-Madrid
- 75% Ciemat-Madrid
- 25% IFCA-Santander




- LHC sites pledges in the last 5 years:
  https://wlcg-cric.cern.ch/core/federation/list/
- Integrated in the WLCG project and **strictly following the experiments computing models.**
- Goal: represent the 5% of the total Tier-2s and Tier-1s resources by 2026
  - With 50% of CPU resources provided by BSC (See Josep Flix talk tomorrow)

# PIC

- **Spanish WLCG Tier-1 centre** → **~80% of resources**
  - Provides ~4% of Tier1 data processing of CERN's LHC detectors ATLAS, CMS and LHCb
- **¼ of the Spanish ATLAS Tier-2** and **a Tier-3 ATLAS data analysis facility**
  - **~10% of resources**
- T2K [neutrinos], MAGIC and CTA [gamma-ray astronomy], PAU and EUCLID [cosmology], VIP [instrumentation], opportunistic access to LIGO/VIRGO and DUNE, among others…

Nov. 2024

**CPU:** 169 kHS23
**Disk:** 20.2 PB
**Tape:** 63.2 PB


- **18 GPUs available:** via JupyterHub and (direct) acces by some VOs, also available through Grid
- PIC is part of the national **Data services RES nodes**, which also allow us to exploit storage resources as well for the LHC experiments

# Resources: PIC

- Well above the required pledge for 2024.
- Goal: to represent the 5% of the total Tier-2 and Tier-1 resources by 2026.

Tier-1 pledged resources at PIC

| Experiment | Resource | 2023 | 2024 | 2025 | 2026 |
|---|---|---|---|---|---|
| ATLAS T1 | CPU [kHS23] | 57.200 | 60.640 | 73.575 | 89.350 |
| | Disk [PB] | 5.440 | 6.520 | 8.369 | 10.050 |
| | Tape [PB] | 14.120 | 18.080 | 25.245 | 35.350 |
| CMS T1 | CPU [kHS23] | 32.000 | 37.200 | 49.500 | 60.000 |
| | Disk [PB] | 3.920 | 4.880 | 6.390 | 8.200 |
| | Tape [PB] | 12.640 | 15.200 | 20.025 | 26.650 |
| LHCb T1 | CPU [kHS23] | 28.280 | 22.880 | 41.760 | 74.350 |
| | Disk [PB] | 2.420 | 2.448 | 4.046 | 6.335 |
| | Tape [PB] | 6.280 | 5.332 | 8.766 | 13.600 |
| WLCG T1 Share | | 4% | 4% | 4.5% | 5% |

- **Tape:** new equipment requires a relative humidity (RH) below 50% for optimal performance
  - Construction of a new humidity-controlled enclosure for the tape system expected to start in Spring 2025
- **Estimated network needs for January 2030:** Overall: 600/800 Gbps
  - **Tier-1 → 300/400 Gbps for LHCONE and 300/400 Gbps for LHCOPN**
  - Science DMZ: Dedicated 100 Gbps for RES and CIEMAT projects

# Resources: ATLAS Tier-2

**IFAE has 25% of the Spanish ATLAS Tier2 resources**

IFAE is co-located with the Spanish Tier-1 at PIC sharing the resources dynamically. The site uses 12% of the PIC batch system: **1566 cores** and **1.75 PB** of disk space based on dCache.

**IFIC has 60% of the Spanish ATLAS Tier2 resources**

IFIC dedicates 61 computing servers with 1943 cores with no hyper-threading, 15 storage servers with **4 PB** of disks, and several machines for file transfer control, information systems and management

**UAM has 15% of the Spanish ATLAS Tier2 resources**

UAM dedicates 21 computing servers with **576 cores**, 44 storage nodes with capacities ranging from 11 TB to 120 TB totaling **1.1 PB** and over 10 machines dealing with transfer metadata, GRID job submission and other auxiliary tasks

| Resource | 2023 | 2024 | 2025 | 2026 |
|---|---|---|---|---|
| CPU [kHS23] | 69.9 | 73.4 | 99.0 | 126.5 |
| Disk [PB] | 6.7 | 7.9 | 10.2 | 12.5 |
| % of ATLAS | 3.0% | 4.0% | 4.5% | 5.0% |

Goal: to represent the 5% of the total Tier-2 and Tier-1 resources by 2026

# Resources: CMS Tier-2

## IFCA provides 25% of the CMS Spanish Tier-2 resources

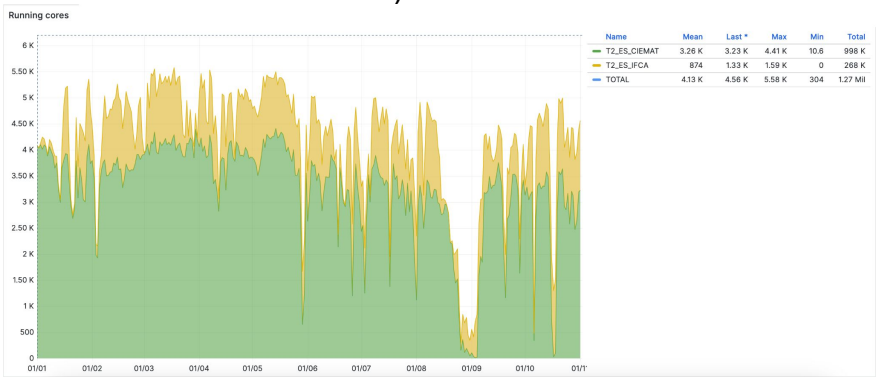
- **Virtualized Infrastructure** (Openstack):
  - Cloud System including both CPU and GPU resources
  - **HPC System: ~2000 cores dedicated to CMS.**
- Opportunistic running on the HPC Altamira node.
- **Storage**: **800 TB for CMS**
- **Network Upgrade**: Migrating from **10Gbps** to **25Gbps**, new compute nodes with Infiniband HDR 200Gbps.

**CMS Spanish Tier-2 Pledges**:

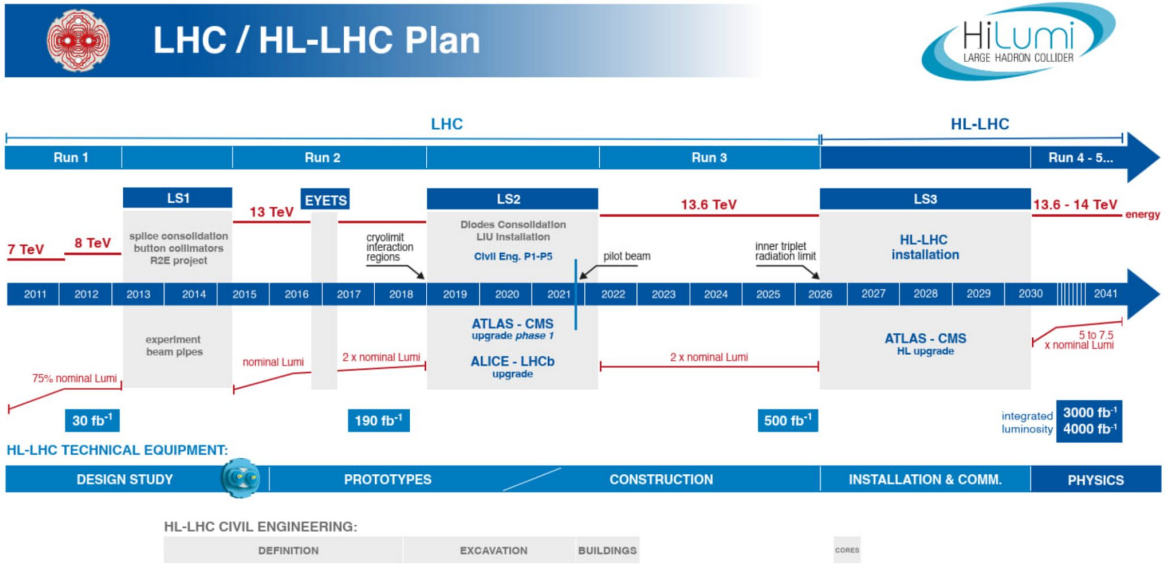
| Site | Resource | 2023 | 2024 | 2025 |
|---|---|---|---|---|
| CIEMAT | CPU [kHS23] | 40.5 | 48.0 | 64.1 |
| | Disk [PB] | 3.5 | 4.5 | 5.9 |
| IFCA | CPU [kHS23] | 13.5 | 16.0 | 21.4 |
| | Disk [PB] | 1.2 | 1.5 | 2.0 |

## Ciemat hosts 75% of the CMS Spanish Tier-2 resources

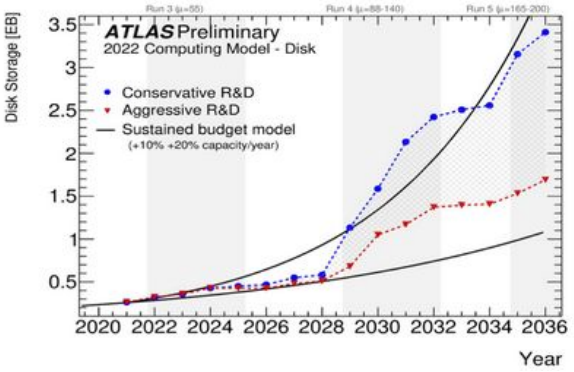
- **CPU**: ~4500 cores for CMS
- **Storage**: **5.3 PB** (above pledge!), dCache v2.27 (dual-stack IPv4/IPv6, TPC enabled for HTTPS)
- **Network**:
  - 2x10 Gbps WAN (LHCOne + Internet)
  - Upgrade to **100 Gbps WAN** pending (RedIRIS and CIEMAT)

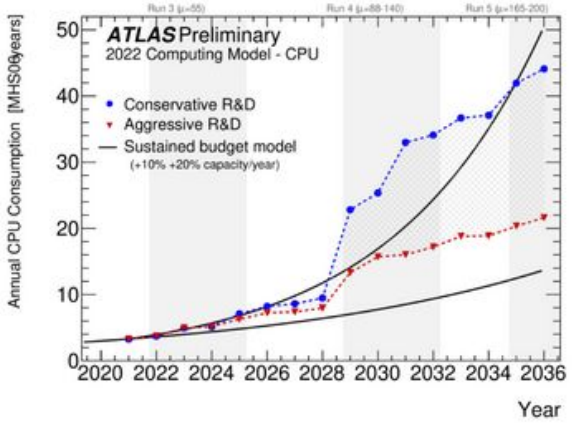


# The HL-LHC context




https://hilumilhc.web.cern.ch/content/hl-lhc-project



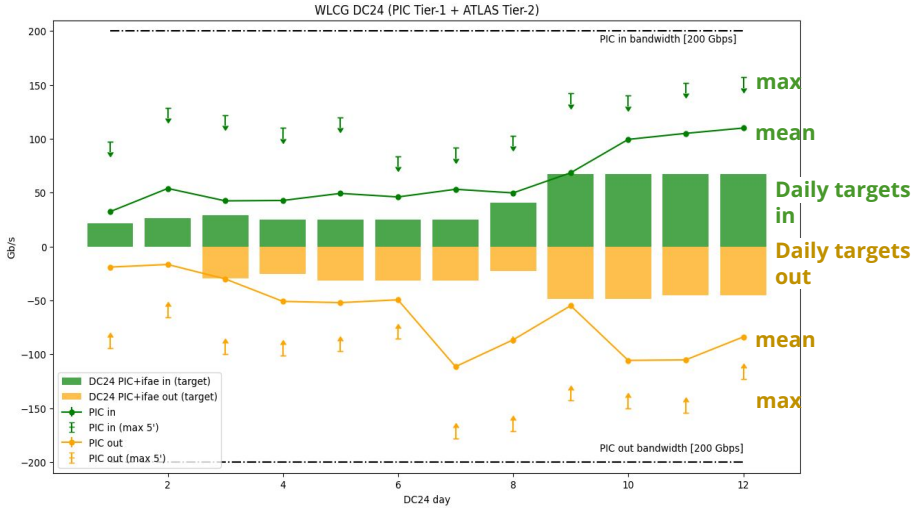




Currently expanding and adapting the WLCG to accommodate the **increased data processing demands expected at the HL-LHC era**

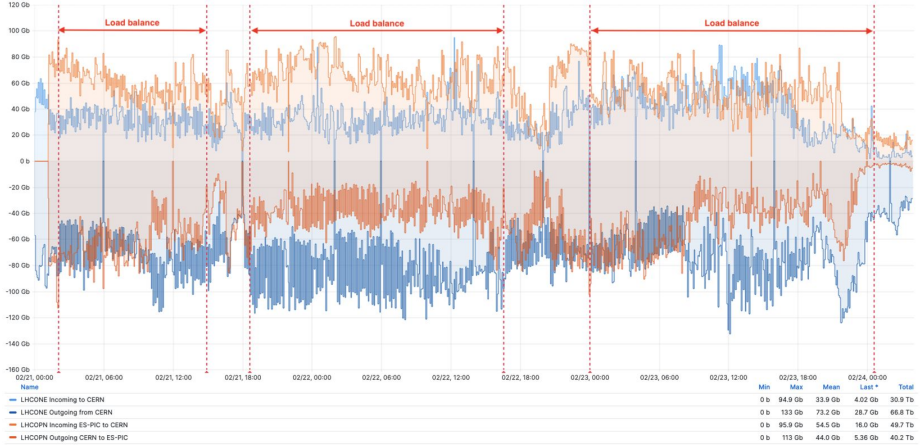
- Storage is the most expensive resource to deploy and operate in the WLCG
- Emphasis on the need for **cost-effective solutions** to manage the growing volume of data
- Moving to a more network-centric model: datalakes, reduce data size and duplication, analysis via caches, …

# Data Challenge ‘24: results at PIC

- During the DC24 testing days, the PIC maintained an average incoming (outgoing) traffic of **63 (64) Gbps**, with peak values reaching **157 (178) Gbps**, respectively, measured in five-minute intervals




- **NOTED** (**Network On-demand Traffic Engineering Demonstration**) was successfully activated at the PIC:
  - Demonstration of **load balancing implementation between LHCOPN and LHCONE**.
  - It activates during high traffic on LHCOPN to prevent saturating the available 100 Gbps bandwidth




❑ ES-PIC load balancing between LHCOPN and LHCONE (from 21st to 23rd of February 2024)

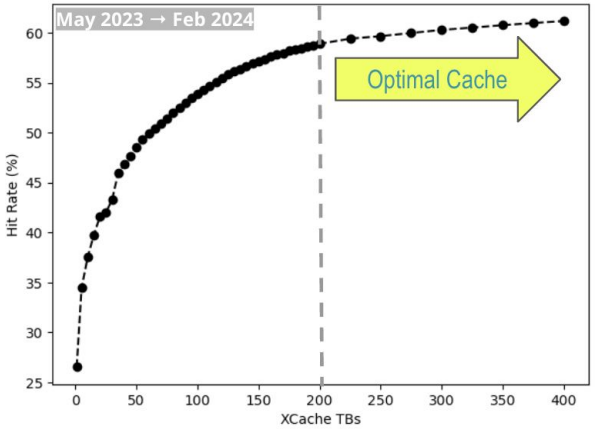
# R&D at PIC: CMS regional cache

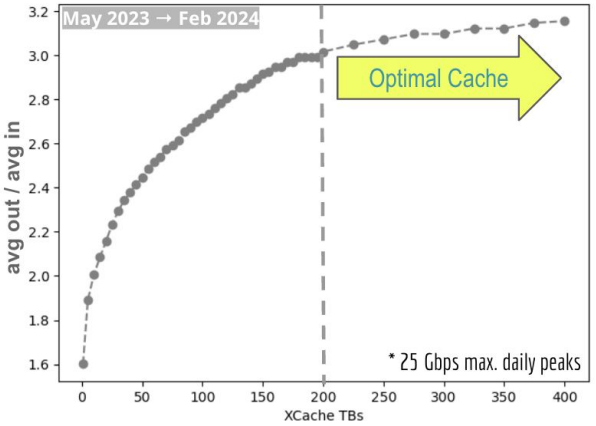
Processing campaigns are run where data is placed… but many CMS user analysis tasks read data remotely, hence **these would benefit with caching**

- Evaluate caching benefits to optimize CMS task execution, especially for remote data access.
- **xCache** deployed in the PIC **can efficiently serve data** just as effectively as if it were read locally at each of the Spanish CMS centers (sites within 10 ms RTT)
  - Observed a +10%  in CPU efficiency for analysis tasks in Spain
  - Introducing data **caches reduces the XRootD traffic generated by 'user jobs'** remote reads by (at least) a factor of 3

**Optimal cache size**

- The **data access details** from all of the Spanish centers play a crucial role in determining the **optimal requirements for cache size** and **network connectivity requirements.**
- Simulating various cache sizes can identify the most efficient option for serving the region, determined by factors such as the **cumulative Hit Rate** (accesses to cached files over total accesses) and **network considerations**
- Deletion from the cache follows the **Least Recently Used (LRU) algorithm**.
  - Even if the LRU mechanism is efficient, more than 65% of the files are accessed once. **Machine learning** can help in improving Hit Rate in caches



# R&D at IFIC: The EventIndex Project

EventIndex is the catalogue of all the ATLAS real and simulated events

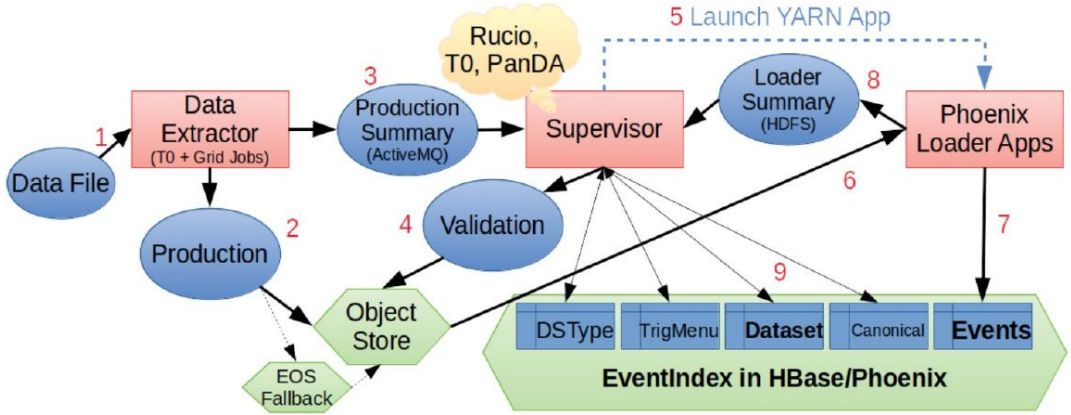
It follows a producer-consumer structure based on message passing and feeding a NoSQL database with event metadata

- Use cases:
  - Event picking
  - Counts or selections
  - Overlaps
    - Of triggers in a dataset
    - Of events in derivations
  - Production checks

The Data Collection is the process that:

- Data extractor obtains the information from the events of the datasets (metadata)
- Supervisor program validates it, assuring correctness and completeness
- Phoenix Loader Apps stores the metadata into the EventIndex back-end

EventIndex back-end on HBase compatible with Phoenix.
Scala with Spark loaders use Resilient Distributed Dataset API.




IFIC is responsible for Data Extractor, Supervisor & Phoenix Loader Apps

**+1306** billion events processed (stored in a single HBase table)

**+614K** datasets

**+48** million files

# R&D at UAM:

## Using MongoDB as a Grafana Backend

- **Grafana** helps visualize site data, such as machine load, storage usage, and network activity.
- The data is traditionally stored in an **InfluxDB** database.
- We're developing a **proxy** (gtom) to store this data in **MongoDB**, providing flexibility and scalability.
- This approach **reduces costs** compared to commercial alternatives while leveraging the benefits of MongoDB.

## OpenTofu: Infrastructure as Code

- Grafana visualizations need to be **defined programmatically**.
- Using the **REST API** over HTTP to define them can be cumbersome.
- Cloud tools like **Terraform** can help automate Grafana provisioning.
- We'll use **OpenTofu** due to its **open-source** nature.
- The result: We can define all our **monitoring dashboards** and **alerts** through simple text files!

## utok: A Microscopic OpenID Connect Client

- Authentication on the grid is moving from **traditional certificates** to **OpenID Connect**.
- We need clients to generate **tokens** for authentication.
- We developed a **minimal client** for testing purposes, which is also **production-ready**.
- It has been successfully tested with the **WLCG's Indigo IAM project**.
- This helps us become more familiar with **OpenID Connect** for future integration.

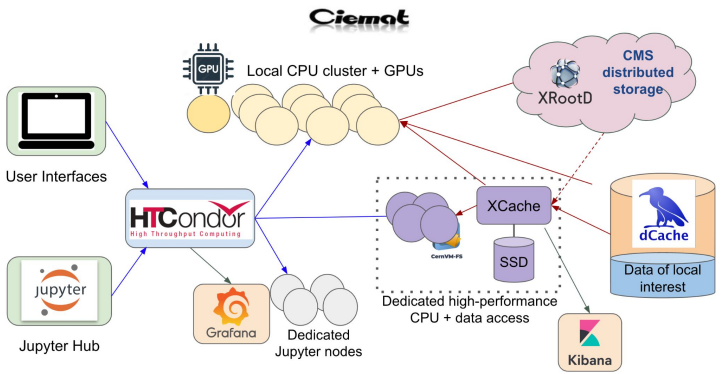
# Analysis Facilities in the LHC context

- The LHC Run 3 & High Luminosity phase challenge: more and busier events
  - Higher pile-up, higher trigger rate, more Monte Carlo simulated events
- From the laptop to the local facility
  - Growing dataset sizes demand a change of paradigm, where high-performance local facilities are used for final analysis (instead of laptops)
  - Key objectives: ease of use, performance, scalability, sustainability
- Evolution of analysis software and paradigms
  - New centrally produced data formats of reduced size (***NanoAOD, DAOD_PHYSLITE***)
  - Use of modern programming interfaces and tools (declarative, **columnar**)
  - Enable trivial/implicit parallelization of the code
  - Growing use of Python libraries, Jupyter notebooks, Machine Learning, etc.
- Key features:
  - 1. **Local access to data:** reduced formats stored **on-site**.
  - 2. **Wealth** (and **quality**) of **resources**: CPUs, GPUs, RAM, SSDs…
  - 3. **Efficient** and **elastic** infrastructure: LRMSs (HTCondor), peak absorption on HPC/Cloud.
  - 4. **Modernisation**: **SW ecosystems**, **UIs**, **expert** on-call, analysis logic **reuse**…

# Analysis Facilities

Typically consist of:

- Dedicated storage resources of order of several 100s of TB (~1PB at CIEMAT!)
  - HDD for local access to reduced data formats and produced ntuples
  - SSD for low latency access to data with catching
- SW delivery mostly via CVMFS with increasing presence of containers
- GPU resources available in most sites, but often not dedicated.
  - IFIC: **ARTEMISA** infrastructure (http://artemisa.ific.uv.es)
  - IFCA: **IaaS access**
- CPU resources used interactively and/or via a batch system (mostly HTCondor)




- **Modernised UIs**
  - Users are reluctant to get into the local batch system, keeping execution of analysis tasks in the UIs. As the need for resources grows, the pressure for them to use HTCondor grows too!
  - We need UI to be designed in a flexible way, with user-friendly interfaces that do not discourage users
    - JupyterHub deployed at CIEMAT (for CMS and other communities) and at IFAE and IFIC (kore.ific.uv.es).
- **Fast** network **backbone** for **LAN** and **WAN** data access.
- Take note: **Dataset Management at CIEMAT:** Dataset replication driven by single user requests is very impractical.
  - Instead, general patterns of interesting data were agreed upon with the CIEMAT-CMS analysts.

# Conclusions

- We have developed a **great infrastructure and community within WLCG-ES**!
  - Expertise and infrastructure built upon decades of work to service the LHC experiments
  - Also supports other experiments HEP/astro/cosmo

- **Data is our distinctive asset**!
  - Currently expanding and adapting the WLCG to accommodate **increased data processing demands expected at the HL-LHC era**
  - Emphasis on the need for **cost-effective solutions** to manage the growing volume of data

- Quite some R&D underway
  - Try to reduce cost of storage infrastructure: **moving to a more network-cetric model.**
    - **excellent results from DC'24**, where we tested load balancing between LHCOPN and LHCONE.
    - **XCache** at PIC serves data as efficiently as local access at Spanish CMS centers (RTT < 10 ms).
    - **Data caches simulation** used to identify optimal **cache size** and **network** connectivity requirements, and to explore various cache **deletion mechanisms**
  - CPU from outside WLCG resources (HPC, GPUs, Clouds)
  - Tabular data formats for easy massive filtering operations
  - Recent improvements in the monitoring and alerts system aim at reducing the operational effort

- **Analysis Facilities key to competitive analysis**
  - Flexible design with a user-friendly interface
  - Specialized for highly parallel local and interactive analysis
  - With local access to reduced data formats

**Acknowledgements**

The authors of this work express their gratitude to the PIC, CIEMAT, IFAE, IFIC, UAM, URL, IFCA and BSC teams for their support in the integration of BSC resources for LHC activities in the Spanish region. This work is partially financed by the Spanish Ministry of Science, Innovation and Universities MCIN AEI/10.13039/501100011033 through the grants with references PID2022-142604OB-C21/C22, PID2022-136323OB-C21/C22, and PID2022-139519OB-C21, which include FEDER funds from the European Union, and also the Catalan government under contract 2021 SGR 00574.


**Authors: Villaplana, M., Acin, V., Acosta Silva, C., Acción García, E., Burriel Navarro, H., Cabrillo, I., Cantero, J., Collado Soto, P., Cuevas, J., Delgado, A., del Peso, J., Fernández, A., Flix Molina, J., García Navarro, J., García, M. J., González de la Hoz, S., Hernández, J. M., Lloret, Lara, Matorras, F., Merino, G., Martínez, F., Pacheco Pagès, A., Pérez-Calero, A., Planas Teruel, E., Salt, J., Sánchez Martínez, F., Saborido, J., Torro Pastor, E., Uzum, R., Vilasís, X.**

# Thanks