October 19 - 25, 2024
CHEP 2024
Conference on Computing in High Energy and Nuclear Physics

Operational experience from the Spanish CMS Analysis Facility at CIEMAT

A. Pérez-Calero Yzquierdo [1], M. Cárdenas-Montes, A. Delgado Peris, J. Flix[1], J. M. Hernández, C. M. Morcillo Pérez, F. J. Rodríguez Calonge
on behalf of the CMS Collaboration
CIEMAT and PIC ([1])

Outline of the talk

- Motivation and context for Analysis Facilities
- Infrastructure for the AF at CIEMAT
- The CIEMAT Analysis framework
- AF operational experience
- Conclusions and Outlook

Motivation for AF at CIEMAT

Analysis Facilities in the LHC context

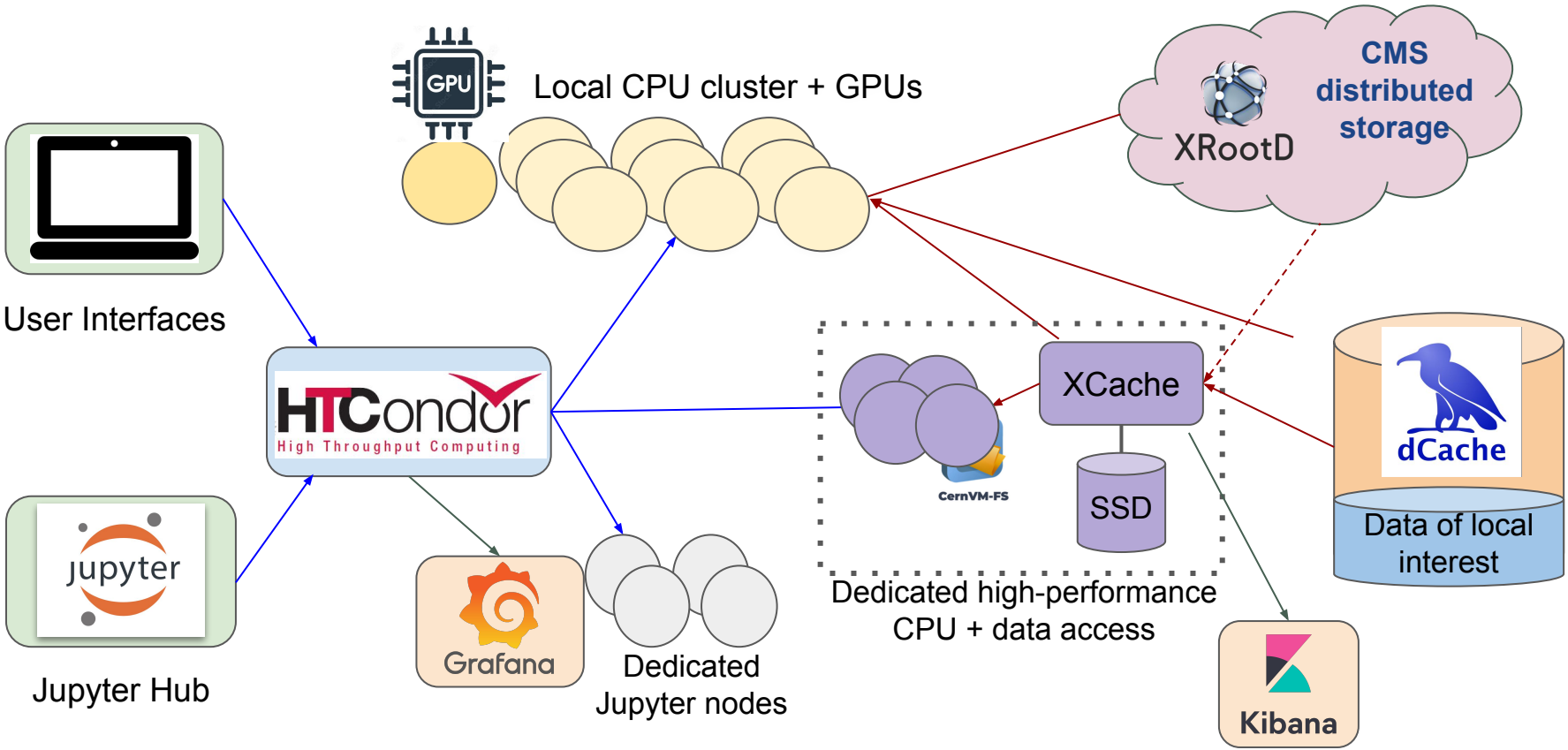
- The LHC Run 3 & High Luminosity phase challenge: more and busier events
  - Higher pile-up, higher trigger rate, more Monte Carlo simulated events
- From the laptop to the local facility
  - Growing dataset sizes demand a change of paradigm, where high-performance local facilities are used for final 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***)
  - 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.

CIEMAT and CMS Computing

- **WLCG**: Operation of the Spanish Tier-1 at PIC (Barcelona) and a Tier-2 at CIEMAT HQ (Madrid) over O(20) years
  - T1+T2 resources dedicated to central data-processing and simulation, plus CRAB analysis jobs
  - CIEMAT Scientific Computing unit deeply involved in CMS Computing and WLCG
- Analysis Facility project (started Sept. 2021) aims to provide CIEMAT physicists with enhanced support and resources
  - Adequate expansion of the analysis capacity
  - Adoption of innovative techniques.
  - Tighter collaboration between CIEMAT computing and analysis groups
    - Support, consultancy and collaborative planning

Infrastructure

CIEMAT Analysis Facility Architecture



Infrastructure (I)

User access

- Interactive login machines (*User Interfaces*)
- JupyterHub deployment (for CMS and other local communities)
- Local access only (a.t.m.) with Kerberos authentication

Processing power

- Two high-performance SSD-equipped servers for analysis
  - AF1: 180 TB SSD, 128 CPU cores, 3000 HS23
  - AF2: 180 TB SSD, 172 CPU cores, 5500 HS23
- Local batch CPU nodes
  - Few hundred CPU cores, co-located and managed with Tier-2 HTCondor pool
- GPU(s) accessible by HTCondor jobs and Jupyter notebooks
  - Server with NVIDIA HGX, 4 GPUs A100
- Plans to enable scaling out to additional clusters (local/HPC) for additional capacity

Infrastructure (II)

Storage areas

- dCache: CIEMAT massive storage solution
  - **Dedicated space (~1 PB)** for data of interest: Mini/NanoAOD, locally produced ntuples
- User data: AFS and NFS areas
  - Code, configuration, small/partial results.
- Analysis-only **XCache service**
  - Deployed on SSD space on AF1/2 servers (320 TB cache)

Access protocols

- Xrootd: CMS preferred option, required for XCache and CMS data federation
- NFS: Access to dCache from jobs/jupyter/UIs. Simplifies workflows (e.g. non-CMS)

Automatic data replication

- Most data relevant to local community will be made available locally
- Other data will be *XCached*

CIEMAT Analysis Framework

The CIEMAT analysis framework (link)

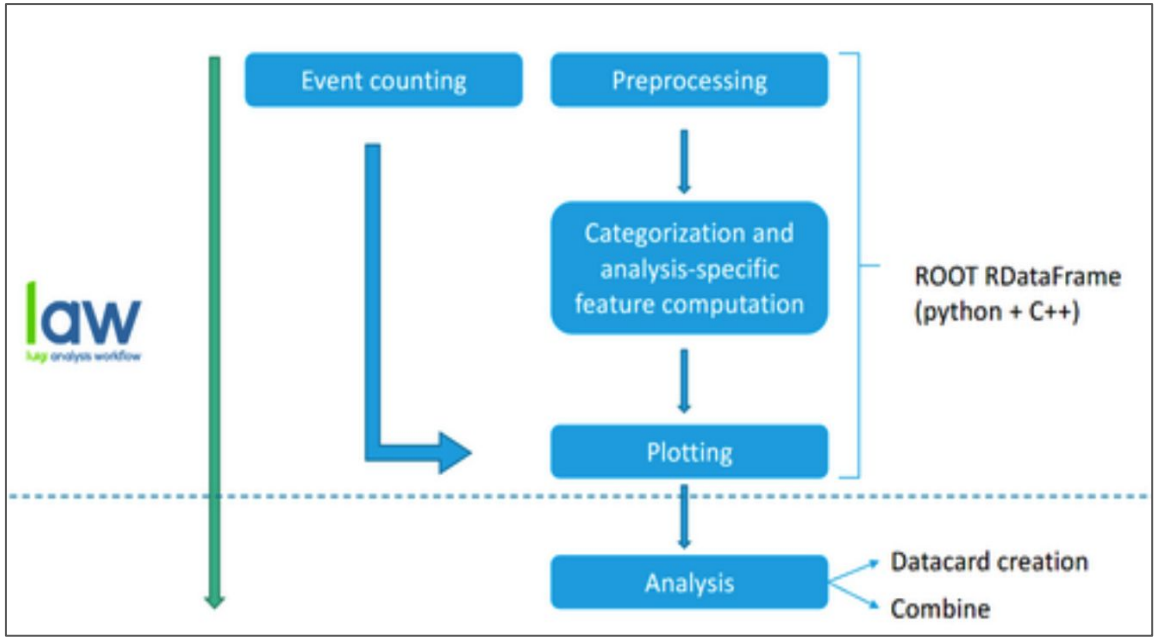
- Motivated originally by a single use case, then grown to be **general** and **flexible**
  - Initially reproduce $HH \rightarrow bb\tau\tau$ analysis using NanoAOD
  - Contributions from CIEMAT analysis and computing groups
  - Being also used by groups in other institutions
- Objectives: user-friendly, fast, general
- In line with general trends in CMS analysis tools
  - User code mostly in **Python** (although a lot of C++ is used)
  - Designed for **NanoAOD** (*or flat tuples*)
  - **CMSSW** and ROOT's **RDataFrame** at its core
- Built on Luigi Analysis Framework (*law*) (link)
  - Tasks organization, batch and file access support, CLI

The CIEMAT analysis framework

Support for:

- Local & HTCondor execution
- Local & xrootd input data files
- Built-in parallelization:
  - One task (job) per file,
  - RDataFrame multi-threading
- Version control as part of the standard workflow
- Tested at CERN and CIEMAT

Run a complete analysis workflow



CIEMAT AF operational experience

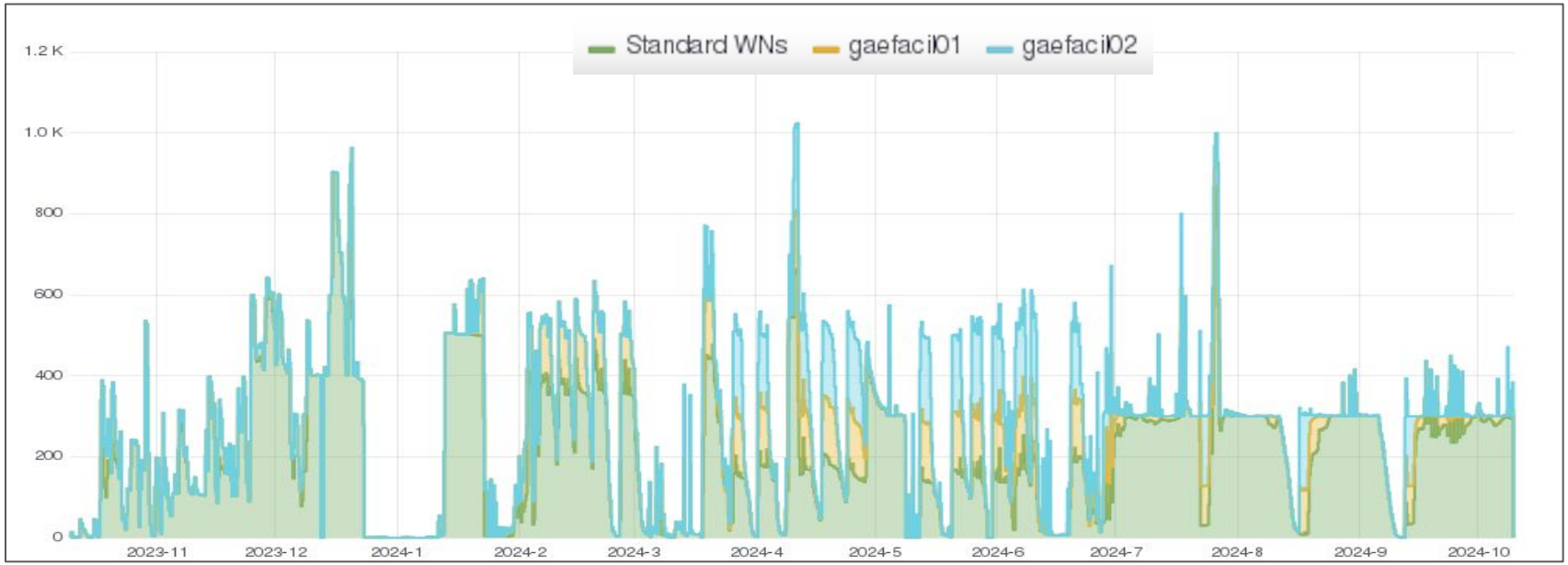
Usage of the AF infrastructure by analysts

- Our objective from the start was to **empower** users…
  - Reduce "*time to insight*"
  - Fight the "*I can only really run at lxplus*" feeling

- …and help/**encourage** them **transition to new tools/technologies**
  - Jupyter, NanoAOD, RDataFrame (and others), GPUs

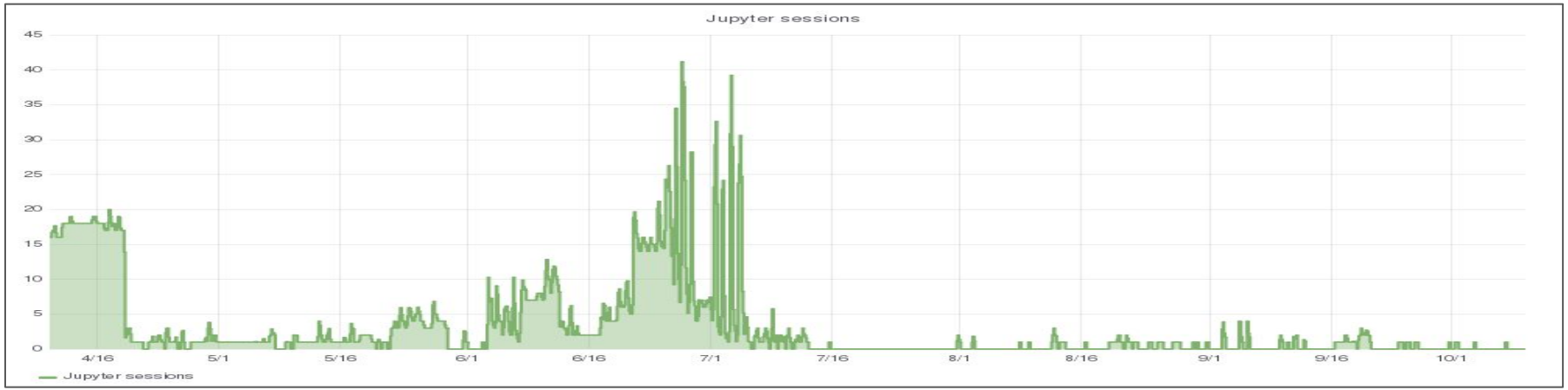
- Already making an **impact** on ongoing analysis!
  - From one early adopters (PhD student in last stage of her doctorate):
    - *I personally noticed the **change in efficiency** clearly. Without the use of these machines I could not have done all the tests and variations of my analysis that I have done, I could not have "tuned" it as much, without relying on lxplus or the previous local machines that took longer.*

CPU usage by local analysis jobs



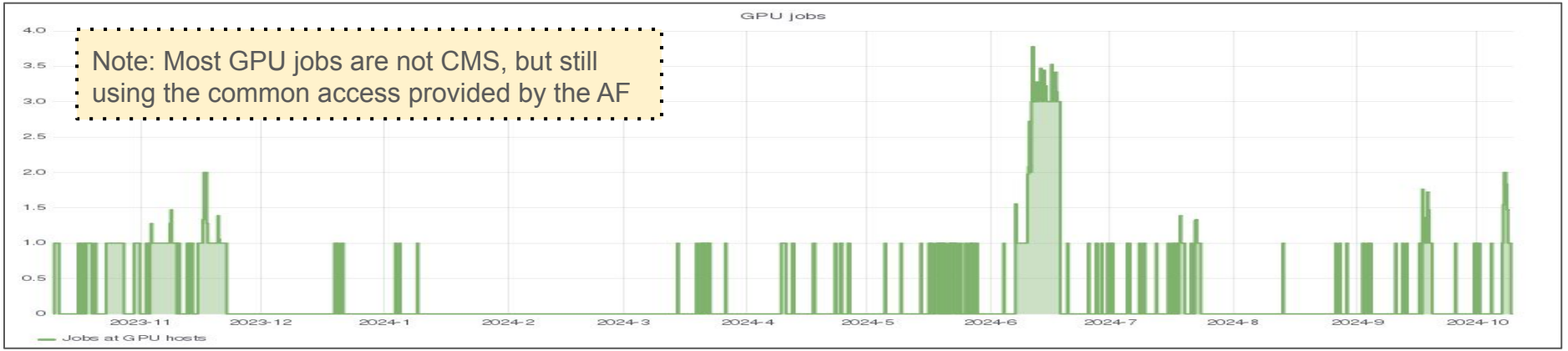
- Last year of **CPU utilization by local user jobs**, both at regular WNs and the two dedicated AF hosts (AF1 & AF2)
- Number of distinct **users** is ~15 (especially, the **youngest**), but **growing**
- Reasonable success so far; further publicity/pedagogy required to enlarge user base

Jupyterhub utilization



- About **10 users in Jupyter routinely** (the youngest), others apparently prefer to **log in a UI**
  - Observe Summer Physics School at CIEMAT!
- Q: **Is this just users preference? Or should the service be improved?**
  - Some users have declared they prefer the terminal
  - Ciemat analysis framework has not been adapted to Jupyter

GPU and Jupyterhub



- We do know that some members of the **CIEMAT CMS community** are using **machine learning** techniques
- But, ***not at CIEMAT*** (so far)
  - Our GPU availability is relatively limited and recent
- **Demand for GPUs expected to grow** as machine learning needs and awareness of available local resources increase within the team

Analysis data management (I)

- **Dataset replication driven by single user requests is very impractical**
- Instead, **general patterns of interesting data** were agreed upon with the CIEMAT-CMS analysts: `/SingleMuon/*/NANOAOD`
  `/SingleElectron/*/NANOAOD`
  `/EGamma/*/NANOAOD`
- Total space dedicated to analysis usage ~300 TB:
  - NanoAOD subscriptions around 65 TB
  - Adding other existing rules, up to nearly 180 TB of locally requested dataset
  - Plus 100 TB of pure user data (`/store/user`)
- Users may have not yet **assimilated** the actual scale of the available dedicated storage space (**1 PB**)

Analysis data management (II)

- **NanoAOD subscription: How?**
  - Create a general Rucio container `group.t2_es_ciemat:/data/NanoAOD_2023/USER`
  - Create a rule that binds that container to our RSE
  - List datasets matching our patterns (with `dasgoclient`), and add all/missing to the container with: `rucio attach <container> <dataset>`
- **Not entirely trivial**:
  - CMS Rucio admins had to create the *user scope* `group.t2_es_ciemat`, so that `t2_es_ciemat_local_users` was allowed to create and act on a container: https://github.com/dmwm/CMSRucio/issues/531
  - We found a bug preventing new datasets added to a container to be replicated due to an existing rule (solved!): https://github.com/dmwm/CMSRucio/issues/537

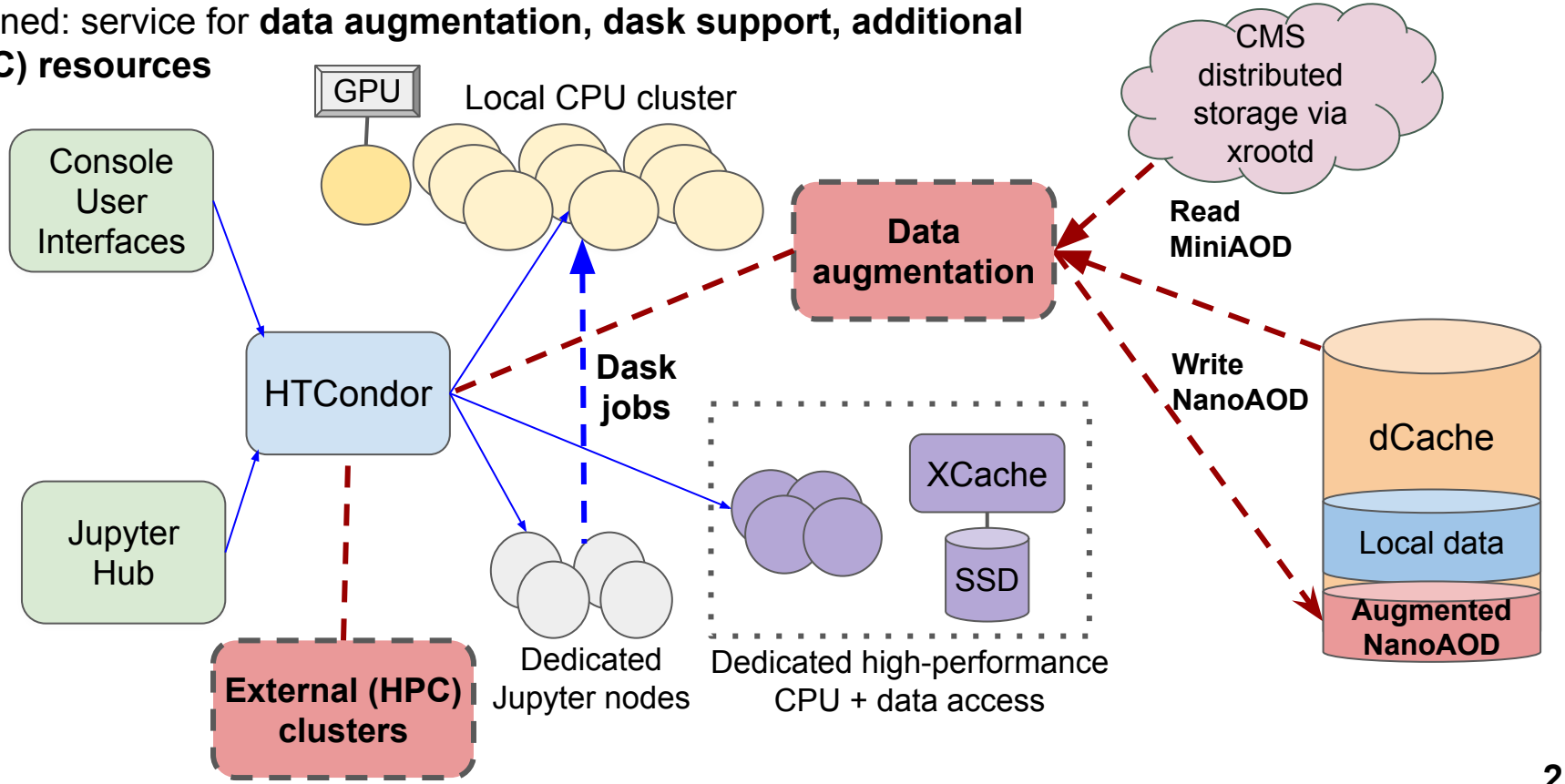
CIEMAT analysis framework: adoption

- Pretty **successful**, especially for new analysis (and new students)
  - Currently in used by several analysis
  - Some of them in collaboration with institutes beyond CIEMAT
- HH → bbττ
  - Several studies
- Long-lived particle analysis*
- Others:
  - Muon studies for POG
  - Wprime → muon+neutrino analysis
  - Charged Higgs analysis
  - W → cs (charm strange) studies
- These analysis are run at either CIEMAT or CERN infrastructures

(*) Jaime León Holgado, Imperial College, UK. Former CIEMAT member, original developer of the framework.

Next services for CIEMAT AF

Planned: service for **data augmentation, dask support, additional (HPC) resources**



Beyond CMS & HEP

- New infrastructure is a **friendly** environment for non-HEP scientific groups at CIEMAT
  - Jupyter interface found convenient by many groups
  - NFS makes massive storage (dCache) widely accessible to users
  - HTCondor allows to ask for required resources (GPU, high performance SSD nodes, etc.), both from command line or through Jupyter
- Several **ML-based studies** ongoing/planned
  - Classification of gravitational wave signals from binary black holes merge
  - Oncology images studies
  - Others: pollution forecast, dark matter detection in liquid argon

Conclusions & Outlook

Conclusions and outlook

- CIEMAT AF: hardware and services dedicated to CMS analysis
  - Progressively being adopted for analysis activities
  - Positive feedback on dedicated infrastructure applied to fast-turnaround analysis
- CMS-CIEMAT group adopting recommended CMS practices for analysis
  - Reduced data format, data caching, new software tools and paradigms
  - Developing user-friendly, fast and general software tools using NanoAOD data
- Outlook for an **ongoing project**
  - Services and infrastructure provided by the AF evolving according to user needs
  - Continuous communication between computing and analysis groups is essential
  - **Project extension approved** for +12m for a planned expansion of the CMS AF:
    - To diverse scientific communities in CIEMAT (e.g. ML applied to classification of GW images from binary black holes merge, oncology studies, pollution forecast)
    - To resources beyond the locally installed and managed (e.g. CIEMAT HPC cluster at CETA-CIEMAT)

Thanks!

Project co-funded by the Recovery and Resilience Facility (Spain), and the European Union – NextGenerationEU

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Financiado por la Unión Europea
NextGenerationEU

Plan de Recuperación, Transformación y Resiliencia

Research projects PID2019-110942RB-C21, PID2020-113807RA-I00, and PID2022-142604OB-C21 funded by:

MINISTERIO DE CIENCIA E INNOVACIÓN

Cofinanciado por la Unión Europea

AGENCIA ESTATAL DE INVESTIGACIÓN

Backup slides

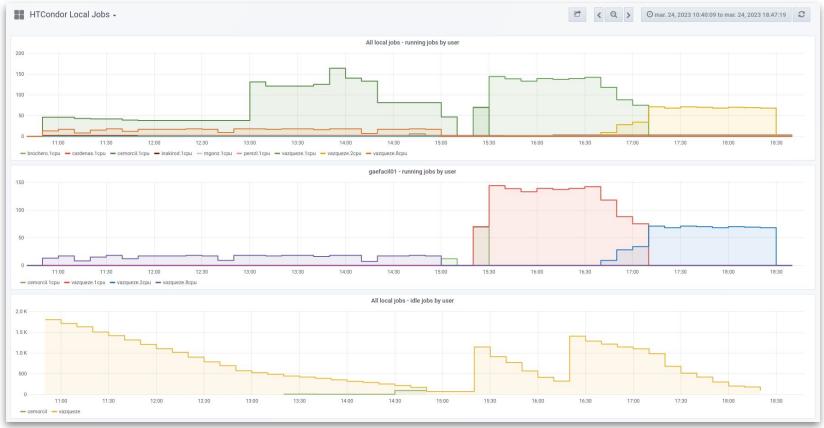
Infrastructure (III)

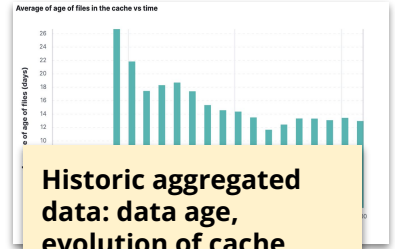
Monitoring

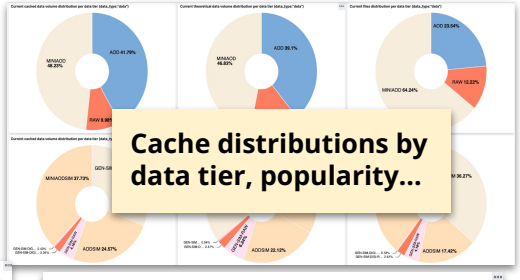
Grafana

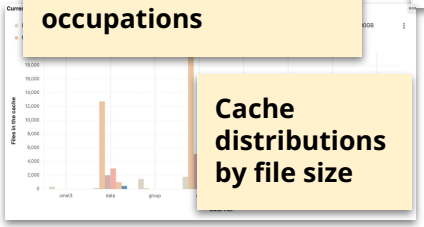
Kibana

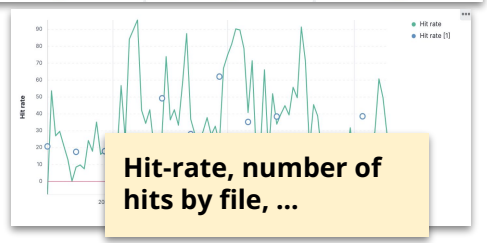
- Dashboards to monitor resource utilization
- Grafana on HTCondor activity on new AF hosts
- Kibana for XCache utilization and activity





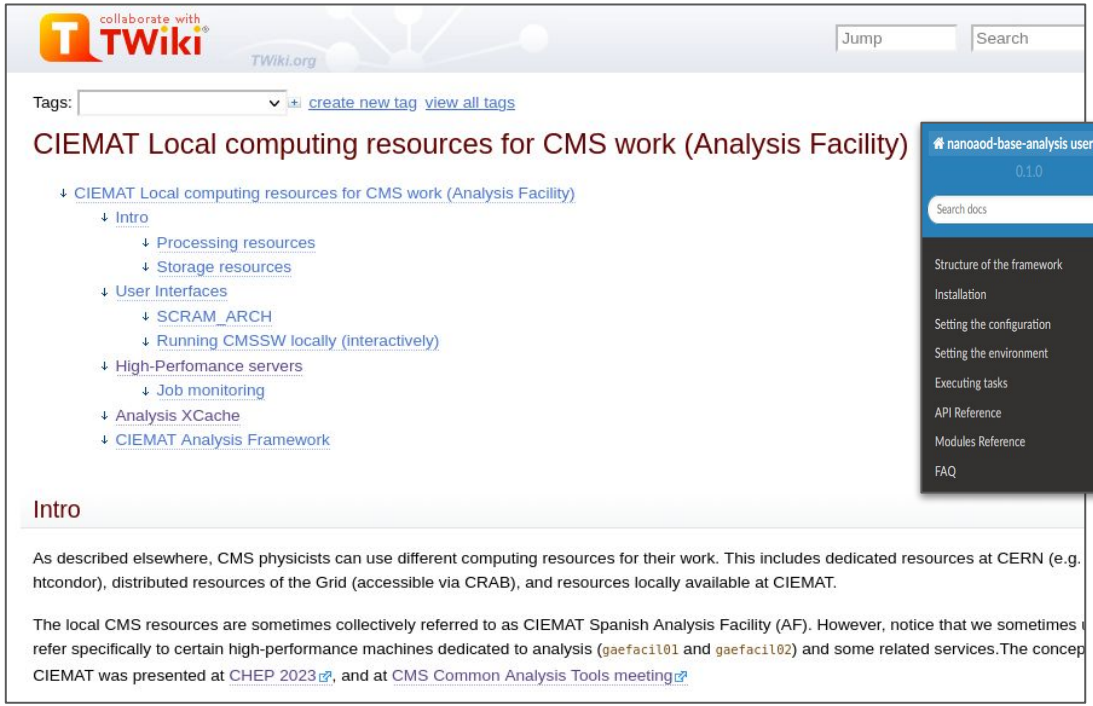


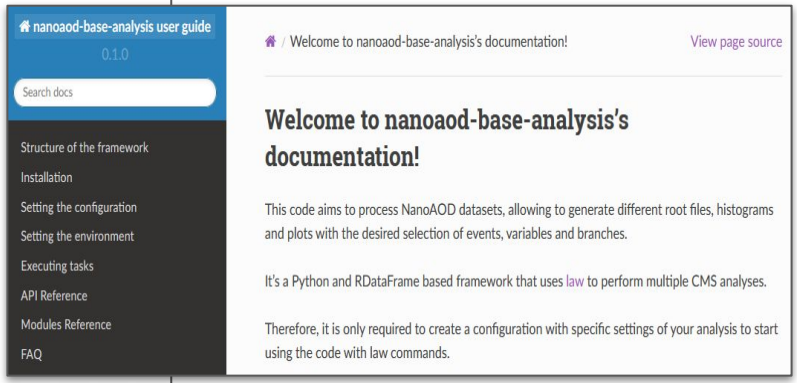




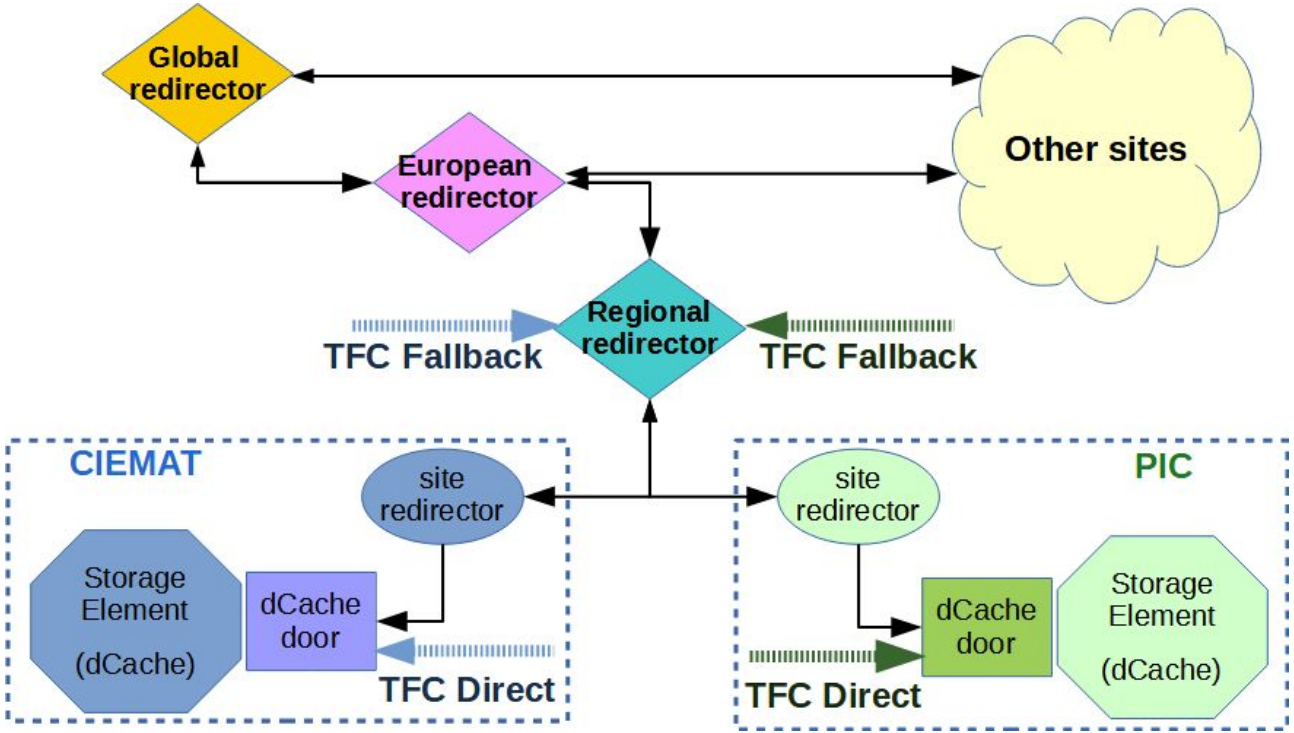
Documentation

- Essential for users to adopt the new tools and resources





Spanish xrootd federation



Lightweight site federation for CMS support, CMS Collaboration, C. Acosta-Silva, A.Delgado Peris, J. Flix, et al., CHEP 2019, Adelaide, Australia. Published in: EPJ Web Conf. 245 (2020), 03013