ESCAPE Data Lake and HPC integration: CMS and CINECA

D. Ciangottini on behalf of the CMS and CNAF team

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

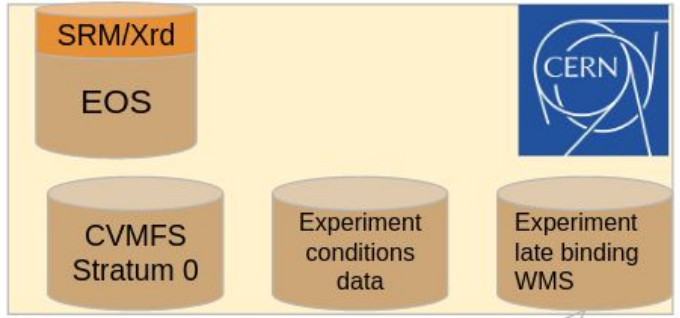
Outline

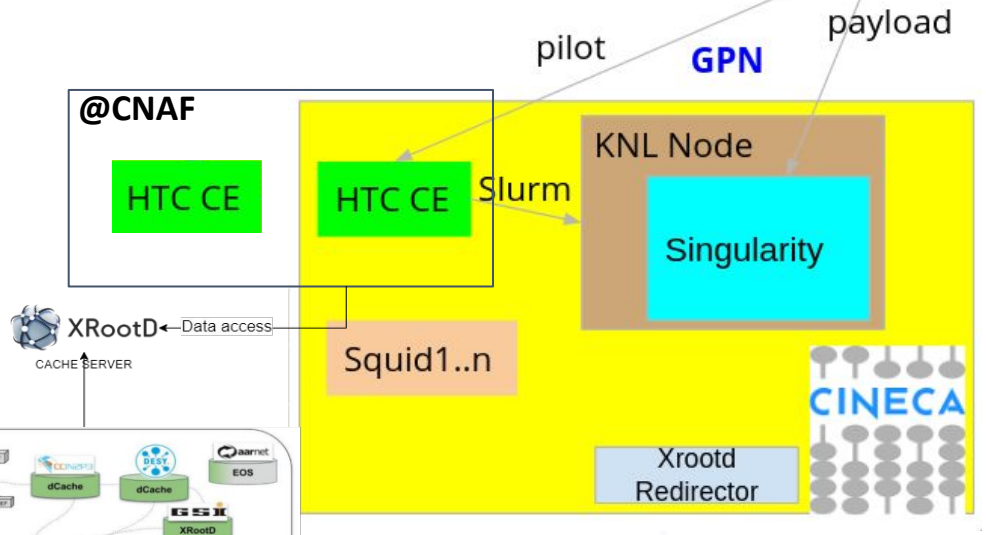
- Data-lake access for CINECA HPC resources
  - CMS-CINECA integration
  - XCache for network fan-out demonstration
  - AuthN/Z and storage configuration
- CMS workflow tests @CINECA
- Results and conclusion

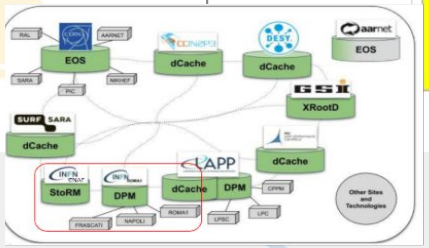
Funded by the European Union's
Horizon 2020 - Grant N° 824064

CMS-CINECA WMS integration

- SW dependencies on CVMFS
- Distributing payloads:
  - Accept / submit workloads which fit the RAM / walltime and IO bandwidth
- We went for **(site) customizable pilots** inside CMS, which allow to **accept** at WN level incoming tasks based on regexps





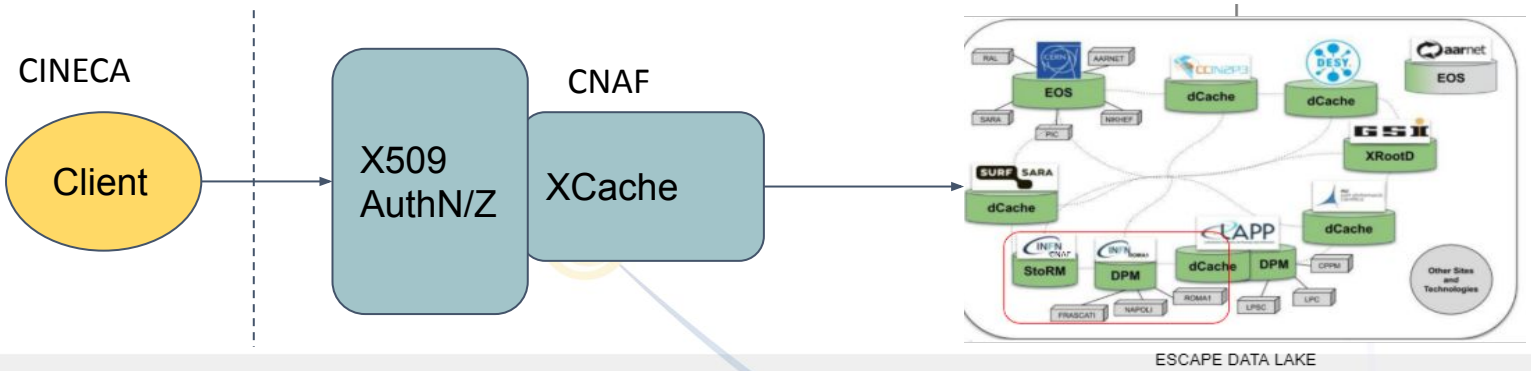


ESCAPE DATA LAKE

Funded by the European Union's Horizon 2020 - Grant N° 824064

Using ESCAPE XCache@CNAF to serve data for HPC

- We wanted to demonstrate that in this setup we can **seamlessly read from ESCAPE data-lake data** in workflows running **at CINECA HPC center**
- In particular, XCache solution can **enable "fan-out" connectivity toward the lake** that is one of the problem when using HPC resources
  - In other words, CINECA trusts and allows only the connection with the XCache at CNAF, but through the cache it can see any other endpoint of the lake



Funded by the European Union's Horizon 2020 - Grant N° 824064

XRootD@INFN-T1 for the ESCAPE embargoed data-lake

The setup currently consists in:

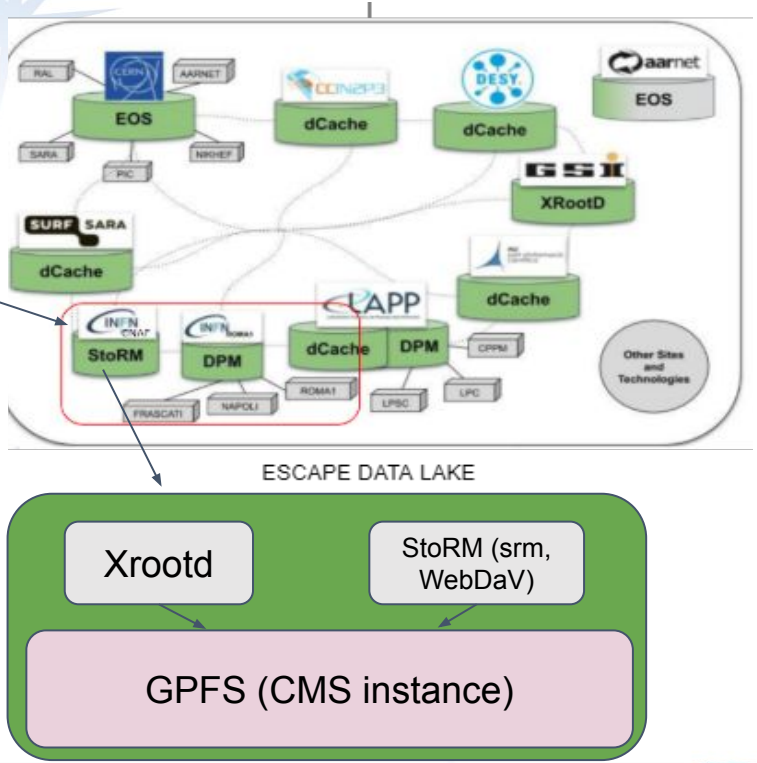
- one endpoint acting as a remote custodial site (origin) running an xrootd server exposing: an HTTP/WebDAV + token service and one with xrootd + x509
- another endpoint working as an XCache instance fetching and storing data from the custodial site
- capability-based AuthN/Z model managed with Escape IAM access tokens, and identity-based one used by xrootd endpoint

* Of interest for today

Funded by the European Union's Horizon 2020 - Grant N° 824064

Additional technical details

- Both origin and XCache run in two Docker images on the same server *xs-102-03-41.cr.cnaf.infn.it*, which is a CMS StoRM WebDAV server
- The origin exposes a dedicated fileset in CMS fs *gpfs_tsm_cms* @CNAF
- XCache uses a dedicated fileset in a fs *gpfs_cache* hosting caches.
- **Dedicated to embargoed data** → not in Rucio machinery yet



Funded by the European Union's Horizon 2020 - Grant N° 824064

CINECA-ESCAPE test: setup

- We **first populated** the ESCAPE datalake origin at CNAF with a **CMS HammerCloud dataset**
- We prepared a CRAB (the CMS distributed analysis tool) task to process the dataset
- A task is formed of **multiple jobs**, each one **accessing a different portion of a dataset**
- The job is directed to CNAF
  - The CINECA KNL “Marconi A2” partition is available via a PRACE Project Access Grant # 2018194658 (PI: Tommaso Boccali) to LHC-Italy since early 2019
  - Machines are intel xeon phi 7250, with 68x4 cores per node, and 96 GB of RAM
  - **The partition is seen by CMS WM as an extension of CNAF Tier1 - driven by CMS decision**

CINECA

Marconi cluster

- Based on Omnipath
- ~19 Pflop/s
- 17 PB of local storage

Marconi A2 Partition

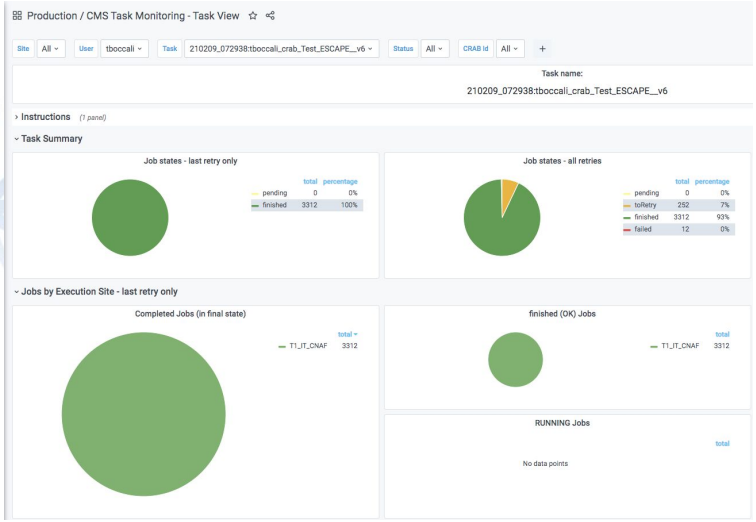
- 3600 nodes with 1 Xeon Phi 2750 (KNL) at 1.4 GHz and 96 GB of RAM
- 68 cores/node, 244800 cores
- Peak Performance: ~11 Pflop/s

Marconi A3 Partition

- 3216 nodes with SkyLake at 2.1 GHz
- Peak Performance: ~8 Pflop/s

Funded by the European Union’s Horizon 2020 - Grant N° 824064

CINECA-ESCAPE test: results



- Jobs running physically at CNAF read the dataset via the standard CMS storage setup
- Jobs **running at CINECA access the files from the XCache ESCAPE endpoint**
  - Specifically configured to verify HPC functionality

- The task (about 3300 unique jobs) has run to completion
  - 40% of the jobs run at CINECA, 60% at CNAF
- Around 250 jobs required a second attempt - quite typical for a CRAB test
- CMS monitoring indeed shows 100% success rate

Funded by the European Union's Horizon 2020 - Grant N° 824064

Conclusions

- Standard **CMS workflows ran without any issue on CINECA resources** reading data from ESCAPE data-lake storage endpoints
- No dedicated tunes on neither storage nor WN side. **Working out of the box with X509 legacy authN/Z**
- Useful tests for **validating the setup for storing CMS embargoed data**
  - We'd like to **test the "token way" soon**

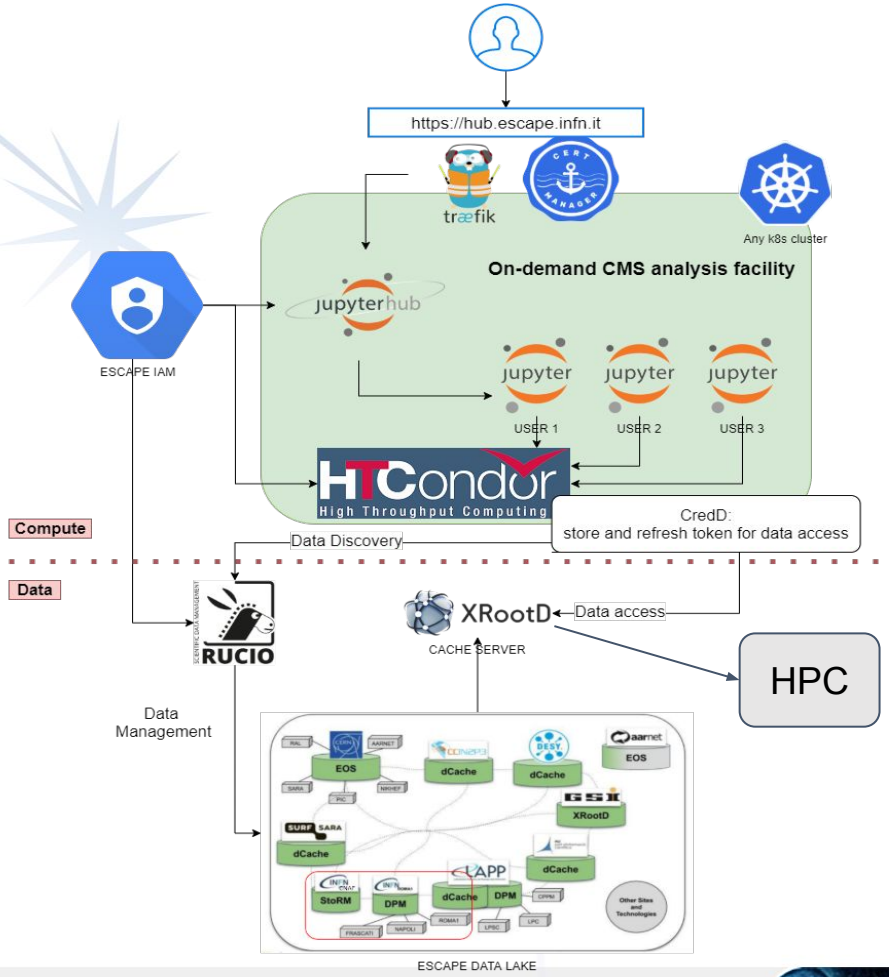
Funded by the European Union's Horizon 2020 - Grant N° 824064

Backup

CMS use case

- Objectives:
  - On demand analysis facility to deploy on any k8s cluster
    - Stateless and reproducible
  - IAM based AuthN/Z for both data and compute resources access
    - X509 free
  - Support both interactive and batch workflows
  - Support data access for multiple kind architecture and resources

Today is about this



Funded by the European Union's Horizon 2020 - Grant N° 824064