Full Dress Rehearsal Exercise on the ESCAPE Pilot DataLake

Riccardo Di Maria

CERN

January 13th, 2021 - WLCG Grid Deployment Board, CERN

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.

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

Science Projects

HiLumi HL-LHC PROJECT
SKA SQUARE KILOMETRE ARRAY
VIRGO
FAIR
KM3NeT
JIVE Joint Institute for VLBI ERIC
cta cherenkov telescope array
LSST
EST european solar telescope
ESO
CERN

ESCAPE

European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

EUROPEAN OPEN SCIENCE CLOUD

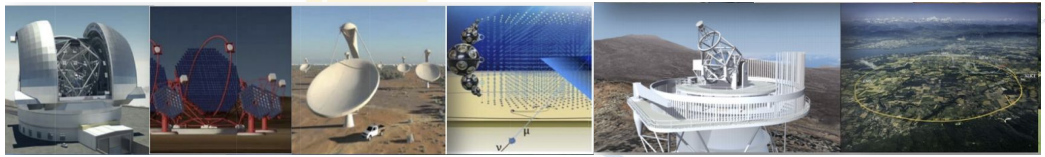
Horizon2020
European Union Funding
for Research & Innovation

Data Centres

CERN
INFN
GSI
DESY
Nikhef
SURF SARA
rijksuniversiteit groningen
CCIN2P3
CNRS
PIC port d'informació científica
LAPP Laboratoire d'Annecy de Physique des Particules
INAF ISTITUTO NAZIONALE DI ASTROFISICA NATIONAL INSTITUTE FOR ASTROPHYSICS

Project Goals

- Prototype an infrastructure adapted to exabyte-scale needs of large science projects.
- Ensure sciences **drive** the development of EOSC.
- Address FAIR data management principles.



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

The ESCAPE Project Work Packages

- **Management, Innovation, Networking and Dissemination (MIND)**: coordination and management.
- **Data Infrastructure for Open Science (DIOS)**: a scalable federated data infrastructure (DataLake) as the basis of an open science for the ESFRI projects within ESCAPE.
- **Open-source scientific Software and Service Repository (OSSR)**: the repository of scientific software services of the research infrastructures concerned by the ESCAPE project.
- **Virtual Observatory - connecting ESFRI projects to EOSC through VO framework (VO)**: astronomical high-level products archive and related services.

@rucio @swan

- **ESFRI Science Analysis Platform (ESAP)**: a flexible science platform for the analysis of open access data.
- **Citizen Science - engagement and communication (CS)**: an open gateway dedicated to the public through Citizen Science and communication actions.

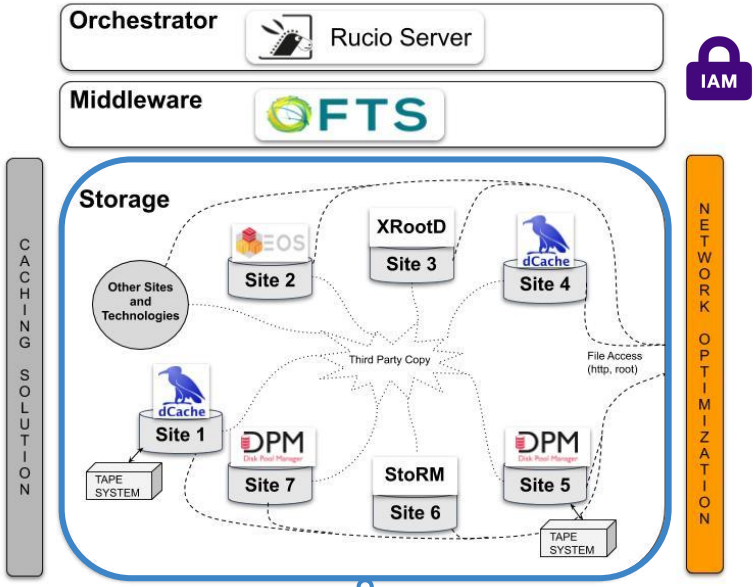
Funded by the European Union's Horizon 2020 - Grant N° 824064

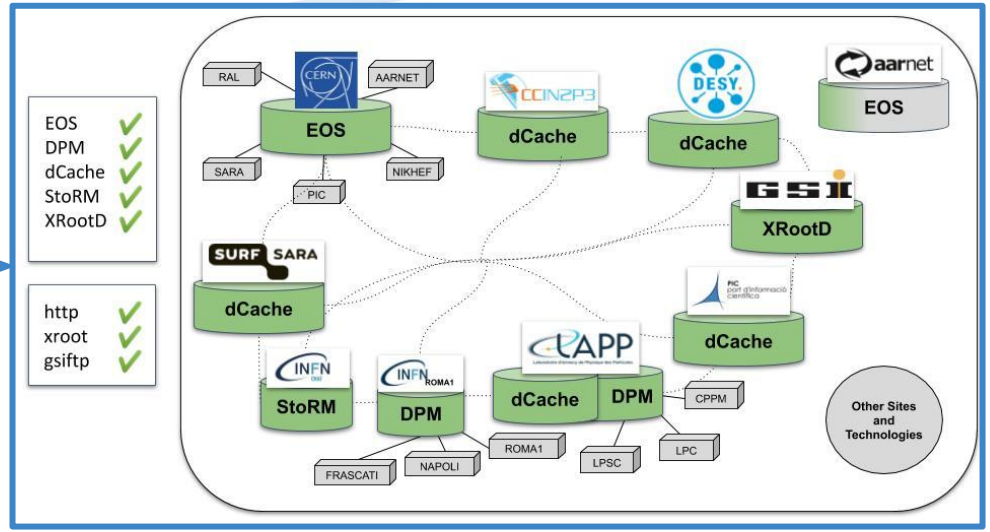
ESCAPE DIOS

- Deliver a Data Infrastructure for Open Science, a non HEP specific implementation of the DataLake concept (HSF Community White Paper + WLCG Strategy Document for HL-LHC).
- ESCAPE sciences at different phases of evolution, all with special interest on data storage, organisation, management and access (**DOMA**).
- Backbone consists of services operated by the partner institutes and connected through reliable networks, **leveraging the existing expertise in WLCG**.
  - e.g. RUCIO, FTS, XRootD-XCache, CRIC, AAI X.509 and Tokens (Indigo IAM), WLCG storage technologies.
    - Development, QoS integration, access-tokens, stress-testing, multi-VO.
  - Supporting various access protocols (HTTP, XRootD and GridFTP) to serve the data to heterogeneous facilities, from conventional Grid sites to HPC centres and Cloud providers.

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

ESCAPE DataLake



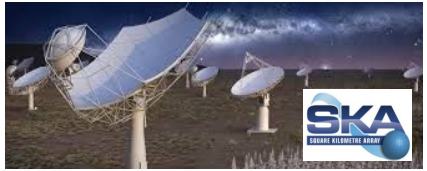


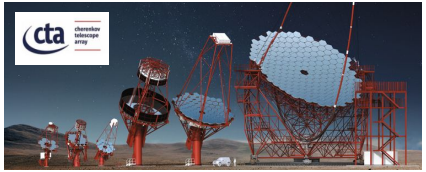
- Hiding complexity and providing transparent access to data.
- Heterogeneous federated storage and operations model.
- Some centers joining even if not funded by ESCAPE.

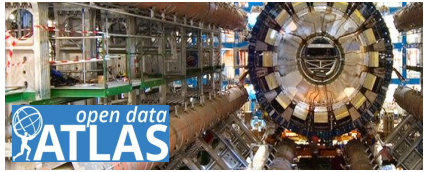
Further info: https://wiki.escape2020.de/index.php/WP2_-_DIOS#Datalake_Status

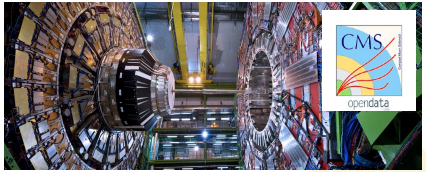
Funded by the European Union's Horizon 2020 - Grant N° 824064

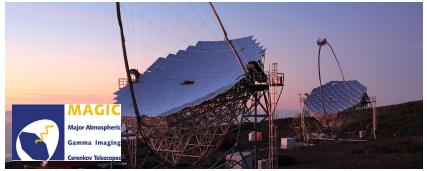
ESCAPE DataLake











RSE	Quota	WM
ALPAMED-DPM	100 TB	10 TB
CNAF-STORM	10 TB	1 TB
DESY-DCACHE	40 TB	4 TB
EULAKE-1	300 TB	30 TB
GSI-ROOT	1 TB	10 GB
IN2P3-CC-DCACHE	60 TB	1 TB
INFN-NA-DPM	68 TB	5 TB
INFN-NA-DPM-FED	46 TB	5 TB
INFN-ROMA1	2 TB	200 GB
LAPP-DCACHE	10 TB	1 TB
LAPP-WEBDAV	100 GB	90 GB
PIC-DCACHE	28 TB	27.99 TB
PIC-INJECT	28 TB	27.99 TB
SARA-DCACHE	98 TB	140 GB

ESCAPE DataLake

- Total Quota:

  891 TB

- Watermark:

  113.44 TB

- 10+ RSEs
- 9 sciences
- 50+ accounts

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

DataLake **24-hour Full Dress Rehearsal** Preparation

- The goal of the FDR exercise is to cover **experiment data workflow** needs on a single day.
  - Perspective from **scientists** and from **sites**.
  - Assessment of the ESCAPE DataLake tools and services under pseudo-production conditions: RUCIO, FTS, CRIC, IAM, perfSONAR, monitoring, QoS, clients, etc.
- Bi-weekly Data Injector Demonstrators meetings pivotal for FDR exercise.
  - Pilot infrastructure at the disposal of very different scientific communities in a cross-collaboration environment.
    - Upsize tasks aming to have basic data management operations known to everyone.
    - Tailored realistic workflows and data lifecycles.
  - Bring (new) sites on board.
- Improving and deploying (new) Kubernetes/Rucio features/functionalities.

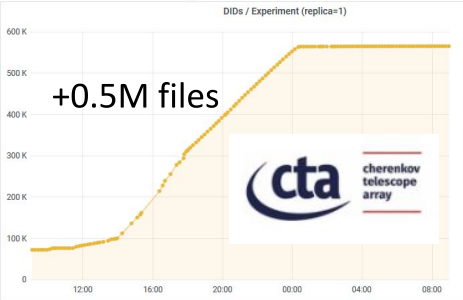
Funded by the European Union's Horizon 2020 - Grant N° 824064

Rucio on Kubernetes Cluster @ CERN

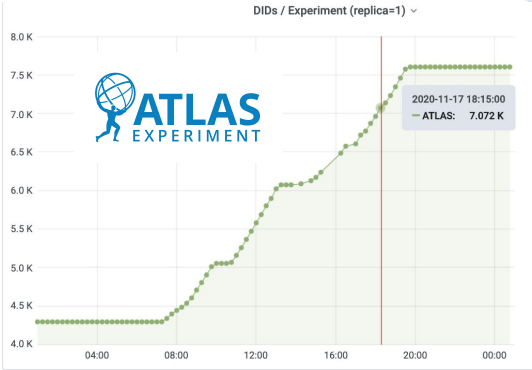
- Fruitful extended collaboration with teams and experts of the various components within and beyond ESCAPE.
  - e.g. MonIT, CERN Cloud, OracleDB, Kubernetes, as well as Rucio, IAM, FTS, CRIC, etc.
- ESCAPE was able to deploy a set of functional services on top of a container orchestrator (Kubernetes) to be tested at experiments/sciences needs.
- Documenting preparation and FDR itself (as deliverables/milestones) is a key objective in the ESCAPE project.
  - Beyond ESCAPE term, different sciences will be able to deploy and manage the subset of services they will want to run and/or customise at their convenience.
- Rucio/JupyterLab Integration Project within CERN-HSF Google Summer of Code (M. Aditya Hilmy) and used by LOFAR during FDR to analyse data.

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

DL **24-hour Full Dress Rehearsal** Takeaway → Workshop

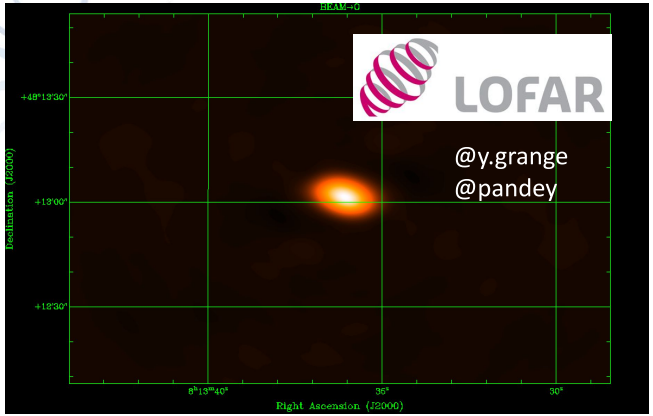


CTA - Simulate a night-data-captured from telescope in Canary Island for 6h: 500 datasets of 10 files ingest.

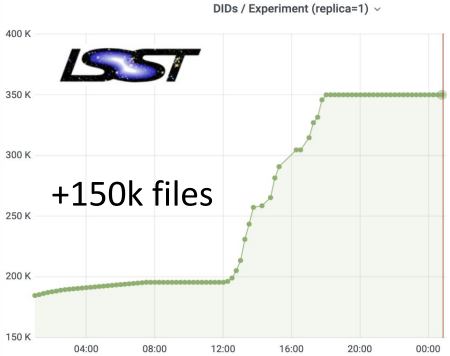


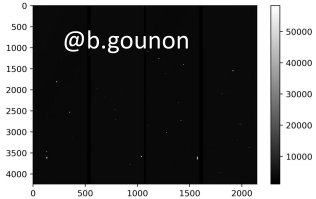
ATLAS - Storage QoS functionality tests: upload files from LAPP cluster to ALPAMED-DPM (FR) and INFN-NA-DPM (IT).

Request transfer to 1 RSE **QoS=SAFE** and 2 RSEs **QoS=CHEAP-ANALYSIS**.



LOFAR - Astronomical radio source 3C196 image using LOFAR data. The raw visibility data were downloaded via rucio from the EULAKE-1 and processed on OpenNebula at SURFsara using the container based LOFAR software.

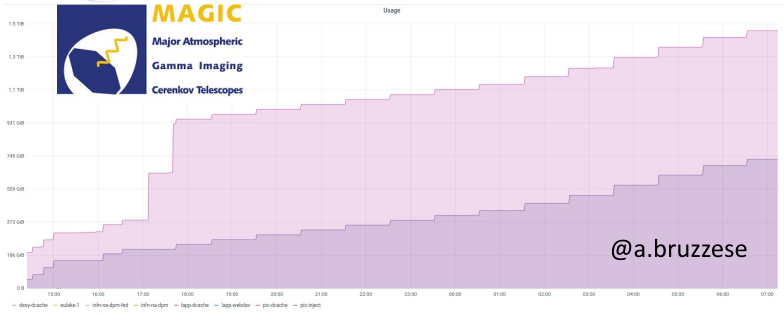




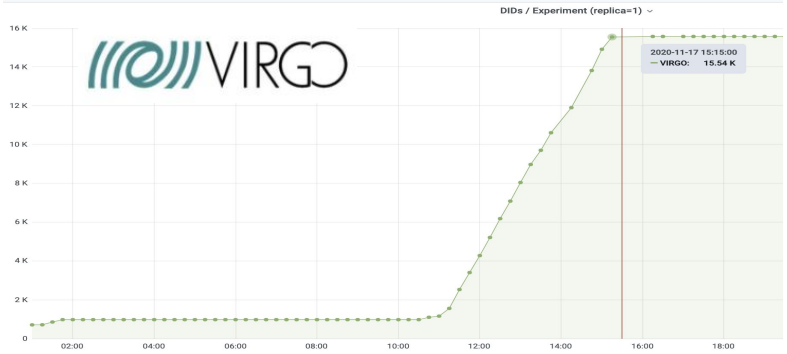
LSST - Simulate production conditions: ingest the HSC RC2 dataset from CC-IN2P3 local storage to the DataLake, **at a realistic LSST data rate** (20TB/24h); **confirm integrity and accessibility of the data via a notebook**.
The image is a reconstruction drawn within a Jupyter Notebook accessing the data used in the Full Dress Rehearsal.

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

DL **24-hour Full Dress Rehearsal** Takeaway → Workshop



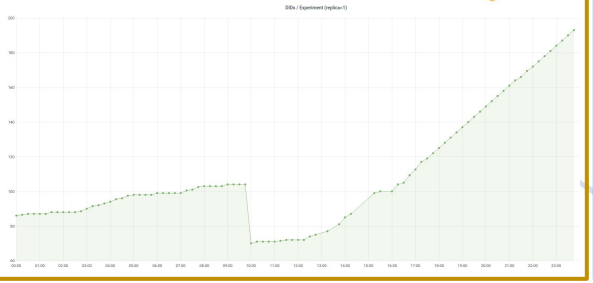
MAGIC - Mimics a real MAGIC observation use case. Remote storage (DataLake aware) **next to the telescope** acts as a buffer for subsequent data injection to the ESCAPE DataLake (and local deletion after success).

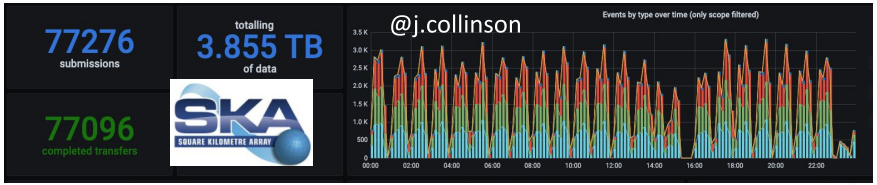


EGO/VIRGO - Upload 4h of VIRGO public data sampled at 4 kHz from an EGO server to the DataLake. Download data to CNAF-STORM. Data are split into 1s samples. Making available the real-time strain data to pipelines and tools assessing the data quality.

FAIR - Upload 1 file (1 GB) every 10 minutes for the whole duration of the rehearsal. Request 2 replicas in **QOS=SAFE** and 1 replica in **QOS=CHEAP-ANALYSIS**.

File size and QoS tagging approximate data ingestion from CBM (i.e. the FAIR experiment expected to produce the largest volume of raw data).





SKA - Pulsar Observations injection test. For 4 hours at any point during the 24h, injecting new group of files in a dataset every 10 minutes. Files fall into two containers, representing different SKA Projects.
24h test moving data on basis of QoS class.

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

Rucio Events & Stats

monit-grafana/rucio-stats

monit-grafana/rucio-events

@rizart
@alba

Stats

472497 submissions	totalling 7.979 TB of data
450766 completed transfers	18403 failed transfers

Used Storage per Experiment (replica=1)

ATLAS 2020-11-17T23:57:45.000Z	1.63 TB
CMS 2020-11-17T23:57:45.000Z	408.33 GB
CTA 2020-11-17T23:57:45.000Z	5.22 TB
FAIR 2020-11-17T23:57:45.000Z	198.92 GB
LOFAR 2020-11-17T23:57:45.000Z	42.06 TB
LSST 2020-11-17T23:57:45.000Z	6.45 TB
MAGIC 2020-11-17T23:57:45.000Z	472.46 GB
SKA 2020-11-17T23:57:45.000Z	8.81 TB
VIRGO 2020-11-17T23:57:45.000Z	1.34 GB

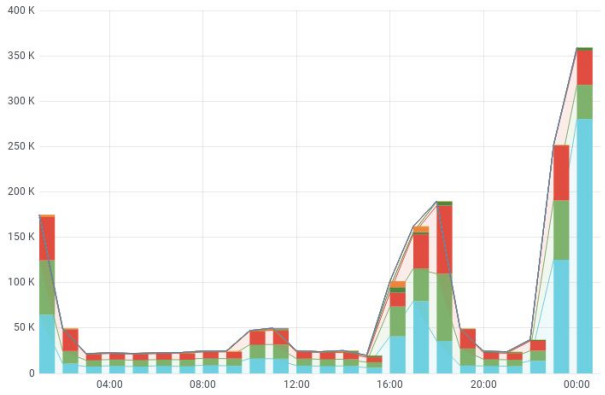
DIDs per Experiment (replica=1)

Experiment	Number of DIDs	Number of files	Number of datasets	Number of containers	Average Filesize ↓
LOFAR	25.3 K	25.2 K	5	0	1.666 GB
FAIR	194	192	2	0	1.036 GB
CMS	401	398	3	0	1.026 GB
MAGIC	13.5 K	824	12.6 K	18	573 MB
ATLAS	7.604 K	6.952 K	652	0	235 MB
LSST	350 K	350 K	13	0	18.5 MB
CTA	564 K	563 K	1.458 K	0	9.273 MB
SKA	2.736 Mil	2.703 Mil	33.0 K	25	3.259 MB
VIRGO	15.6 K	15.6 K	10	0	86.4 kB

Transfer Matrix: transfer-done/transfer-submitted

src \ dst	DESY-DCACHE	SARA-DCACHE	PIC-DCACHE	EULAKE-1	LAPP-DCACHE	IN2P3-CC-DCACHE	CNAF-STORM	ALPAMED-DPM	GSI-ROOT	INFN-NA-DPM	LAPP-WEBDAV	INFN-NA-DPM-FED	INFN-ROMA1
DESY-DCACHE	NO DATA	100%	51%	100%	104%	100%	100%	93%	35%	98%	100%	100%	NO DATA
SARA-DCACHE	100%	NO DATA	98%	100%	100%	100%	98%	88%	26%	98%	98%	96%	NO DATA
PIC-DCACHE	100%	100%	NO DATA	99%	100%	100%	100%	100%	23%	100%	100%	96%	NO DATA
EULAKE-1	100%	75%	47%	NO DATA	100%	100%	100%	100%	42%	100%	100%	100%	NO DATA
LAPP-DCACHE	100%	100%	99%	100%	NO DATA	98%	100%	98%	16%	98%	94%	96%	NO DATA
IN2P3-CC-DCACHE	100%	100%	89%	100%	100%	NO DATA	100%	91%	35%	98%	100%	100%	NO DATA
CNAF-STORM	100%	100%	98%	100%	100%	97%	NO DATA	100%	18%	100%	100%	100%	NO DATA
ALPAMED-DPM	28%	94%	100%	100%	100%	100%	100%	NO DATA	49%	93%	100%	100%	NO DATA
GSI-ROOT	100%	99%	94%	100%	99%	100%	100%	89%	NO DATA	100%	97%	95%	NO DATA
INFN-NA-DPM	100%	100%	100%	100%	99%	100%	99%	90%	45%	NO DATA	98%	NO DATA	NO DATA
LAPP-WEBDAV	100%	100%	100%	100%	100%	100%	98%	100%	100%	100%	NO DATA	100%	NO DATA
INFN-NA-DPM-FED	100%	100%	96%	100%	93%	100%	96%	81%	40%	NO DATA	96%	NO DATA	NO DATA
INFN-ROMA1	NO DATA	NO DATA	NO DATA	NO DATA	NO DATA	NO DATA	NO DATA	NO DATA	NO DATA	NO DATA	NO DATA	NO DATA	NO DATA

Events by type over time (only scope filtered)



	total
transfer-queued	796.0 K
transfer-submitted	474.2 K
transfer-done	452.3 K
deletion-done	25.5 K
transfer-failed	18.5 K
create_dts	10.0 K
detach	688
register_cnt	661
erase	17
transfer-submission_failed	9
create_cnt	4

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

Full Dress Rehearsal Takeaway (1/2)

- Infrastructure should be resource-aware for the project sustainability (**minimal env.**).
  - Sciences at different scale and trying to address multiple future use case, including experiments with smaller data management load than ATLAS and CMS.
  - Full Dress Rehearsal proved:
    - ESCAPE Rucio needs less than **30 CPUs** and **40 GiB** for 29 k8s-pods on 6 [8 CPU|16 GiB] OpenStack VMs.
- FDR highlighted limits of the current configuration - important lesson for ESCAPE and Rucio:
  - Synergy with **Rucio team** allowed to solve the encountered issues and tailor the infrastructure to cope with the needs → exploring new Rucio phase space.
    - Main and Auth servers limits: **SOLVED** on-the-fly.
    - Abacus Collection Replica Daemon: **SOLVED** post FDR.
    - Judge (Injector, Evaluator, Repairer) Daemons 1M-file rule: **SOLVED** post FDR (new algorithm! → ATLAS).

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

Full Dress Rehearsal Takeaway (2/2)

- k8s/CERN-GitOps for R&D work and stable production environment: **ON-GOING**.
- DB (devdb19u) problematic: **SOLVED** post FDR - moved to PROD+DEV.

- Sites involved and responsive.
  - GSI-ROOT RSE on a VM single disk **SOLVED** on-the-fly → now better XRootD endpoint.
- Sciences and experiments strongly involved and committed.
  - Contributing with more and more realistic use cases and workflows.
  - LSST batch issue immediately **SOLVED** with a workaround.
- LAPP pipe filled due to ATLAS data movement clash (ESCAPE-WLCG overlap)
  → workload orchestration to be minded especially for mid-size multi-VO sites.

Injected: 20+ TB / 800+ k files → <u>25 MB average file size</u>
Transferred: 8+ TB

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

Conclusion and Next Steps

- ESCAPE managed to pilot a DataLake infrastructure that could fulfil the functional data management needs of flagship ESFRIs from several scientific disciplines.
  - Sensible technologies choice, conceived in WLCG environment and LHC experiments.
- FDR played a pivotal role to test model, concepts, and pilot infrastructure, and more importantly to enroll Astronomy, High Energy Physics, and Astro-Particle Physics sciences to deploy workflows into into a common data management infrastructure, identifying and addressing infrastructural and service bottlenecks.
  - Chosen technologies offer the right functionality for a broader set of communities.
  - ESCAPE contributing to broaden the scope of some of those technologies according to partners needs (in line/collaboration with providers plans).
- ESCAPE work is complementary and supports to the WLCG direction of broadening the scope of the infrastructure to other sciences/experiments, strengthening relationship for future collaboration.

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

Conclusion and Next Steps

- ESCAPE is mature to move towards a more mature phase (prototype).
  - e.g. fine-tuned QoS, continuous stress-testing and monitoring, ability to plug heterogeneous clouds (commercial) and HPC.
  - Fine-tuning interactions with science analysis methods through content delivery and caching → XCache@CERN.
- ESCAPE end in 2022 → necessity to address long term sustainability.
  - Adopting components from established scientific contexts.
  - Leveraging services supported by large open source communities.
  - Documenting know-how on integration and deployment.
  - Ensuring services become part of EOSC-core.
- New FDR exploiting evolved infrastructure will happen in early 2022… *STAY TUNED!*
  - e.g. FAIR data management vs. embargoed (Open Data policy), fully multi-VO, implementation of token-based AAI → complementing existing efforts in WLCG, ESCAPE is perfect environment to test disruptive changes willing to be prototyped in WLCG within HL-LHC scope.

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

Backup

- ESCAPE presentation at GDB from Xavier Espinal on May 6th, 2020

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

Milestones

Start 02/2019

September 2019

Pilot DataLake design and implementation plan ✓

August 2020

Initial DataLake implementation with at least 3 centers ✓

November 2020

Pilot DataLake performance testing and analysis - Workshop ✓

January 2021

Analysis and assessment of the pilot DataLake ✓

February 2021

Expanded Prototype. Verify RI data access from compute platforms (inc. clouds)

May 2022

Full analysis and assessment of the full DataLake prototype

End 08/2022

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

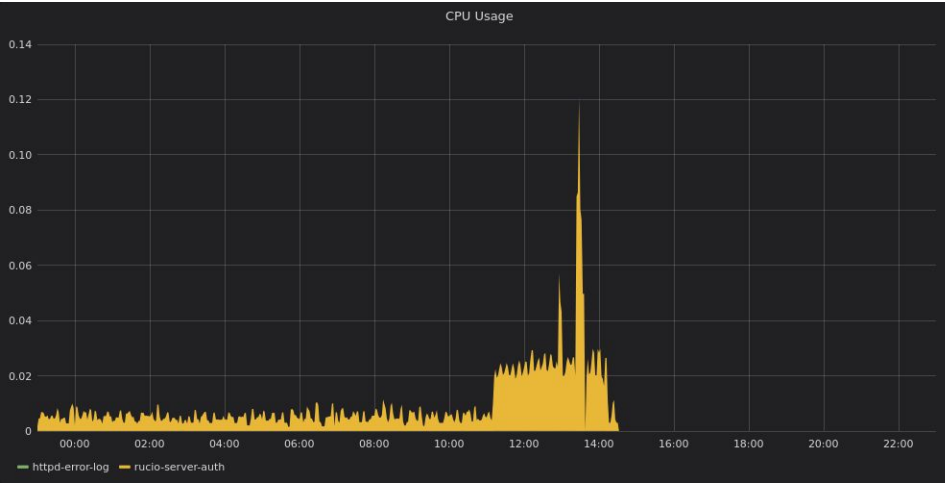
Next Steps: from Pilot to Prototype

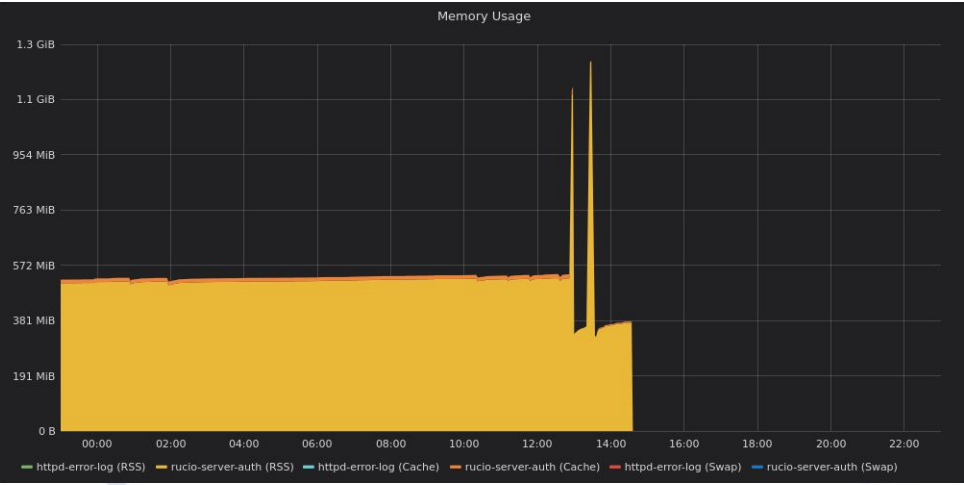
- **AAI:** token-based data management deployed on the DataLake.
- **Storage Orchestration:** QoS parameter development and tuning for reliability, performance, and cost; event-driven data management tested.
- **Network and Asynchronous Data Transfer:** third party transfers enabled; network route optimisation for intelligent transfers.
- **Content Delivery and Caching:** interactions with science analysis methods within ESCAPE WP5.
  - Real data distribution and analysis for non-HEP RI (LOFAR, CTA, LSST, MAGIC).
  - Rucio/JupyterLab Integration Project by Muhammad Aditya Hilmy (GSOC Student) presented at August WP2 Fortnightly Meeting.
  - Investigate data corruption.
  - *XCache@CERN*.
- **Configuration, Monitoring, and Accounting:** instrument workload testing on the DataLake; final DataLake dashboard.
  - Ability to plug heterogeneous clouds (commercial) and HPC.
  - HammerCloud to run realistic research infrastructure workloads.
  - Enable or develop more features, e.g. Rucio multi-VO, tokens, etc..

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

Auth - before setting resources requests/limits

Service	Resources Requests [CPU | Memory]	Resources Limits [CPU | Memory]	FDR Usage of Resources [CPU (peak) | Memory (peak)]	Restarts/Comments
server-auth	-	-	0.02 (0.12) | 550 MiB (1.25 GiB)	Errors and restart due to no limits set.

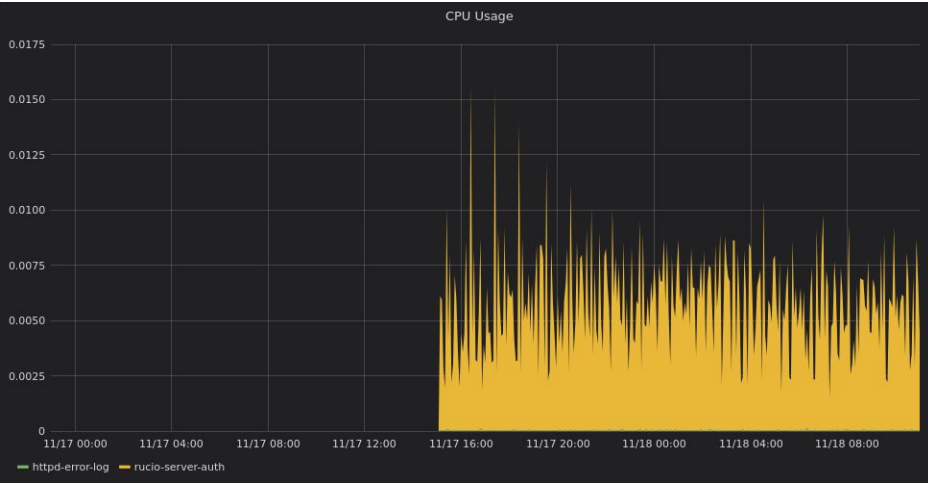


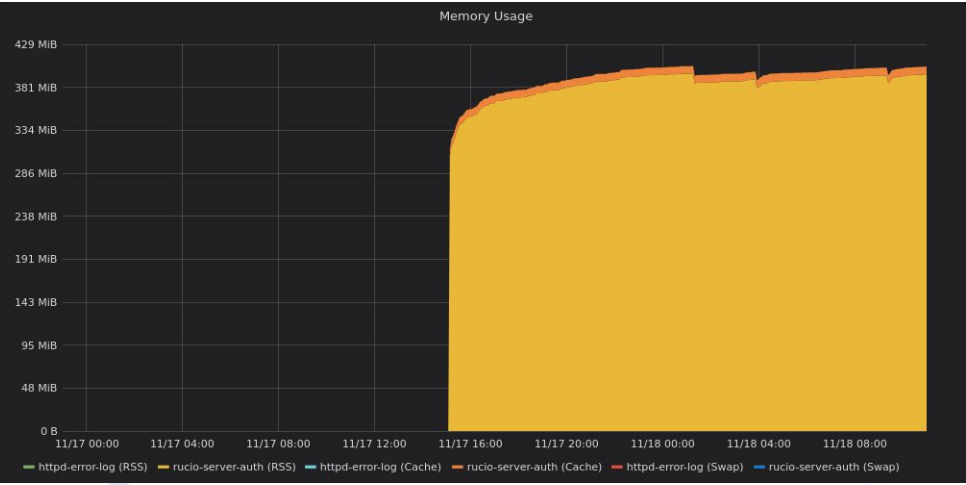


Funded by the European Union's Horizon 2020 - Grant N° 824064

Auth - after setting resources requests/limits

Service	Resources Requests [CPU | Memory]	Resources Limits [CPU | Memory]	FDR Usage of Resources [CPU (peak) | Memory (peak)]	Restarts/Comments
server-auth	4 | 2500 MiB	4 | 2500 MiB	0.02 | 500 MiB	OK





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

Kubernetes Cluster @ CERN

- OpenStack VMs:
  - 1 master: 4 CPU, 8 GB RAM
  - 10 nodes: 8 CPU, 16 GB RAM
- K8s cluster:
  - filebeat (per node) and logstash for cluster monitoring
  - rucio-client with root account and admin privileges for DataLake managing
  - escape-crons pod
    - IAM-Rucio sync
    - IAM-Gridmap (EOS) sync
    - CRIC-Rucio sync
    - noise production (100MB file upload per RSE + add rule per RSE)
    - Gfal SAM and FTS tests
- OracleDB (devdb19u):
  - quota raised from 15 GB to 50 GB

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

Kubernetes Cluster @ CERN

- Rucio (HELM-charts-based):
  - UI (escape-rucio.cern.ch)
  - Auth Server
  - Main Server (2)
  - Daemons:
    - **Abacus Account** [updating account (counter) usages]
    - **Abacus Collection Replica** [updating collection replica]
    - **Abacus RSE** [updating RSE (counter) usages]
    - **Conveyor Submitter** (3 x 4 threads) [managing non-tape file transfers - preparing and submitting jobs]
    - **Conveyor Poller** (3 x 4 threads) [checking status of submitted transfers]
    - **Conveyor Finisher** (2 threads) [updating Rucio internal state for finished transfers]
    - **Hermes** [delivering messages via STOMP to a message broker]
    - **Judge Injector** (2) [asynchronously injecting replication rules]
    - **Judge Evaluator** (3 x 3 threads) [executing and reevaluating replication rules]
    - **Judge Repairer** (2 x 5 threads) [repairing stuck replication rules]
    - **Judge Cleaner** (2 x 5 threads) [cleaning expired replication rules]
    - **Reaper2** (2 x 4 threads) [deleting replicas]
    - **Transmogrifier** [creating replication rules for DIDs matching a subscription]
    - **Undertaker** [managing (deleting) expired DIDs]

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

FDR Takeaway

Injected: 20+ TB / 800+ k files → 25 MB average file size
Transferred: 8+ TB

- Rucio → (#replicas) [CPU (*limits)|Memory]:
  - UI → (1) [0.1|500 (*800) MiB];
  - Auth Server → (2) [0.2 (*1)|0.5 (*1) GiB];
  - Main Server → (2) [2 (*4)|2 (*4) GiB];
  - Daemons:
    - **Abacus Account** → (1) [0.1|150 MiB];
    - **Abacus Collection Replica** → (1) [0.4|200 MiB];
    - **Abacus RSE** → (1) [0.1|150 MiB];
    - **Conveyor Submitter** → (3 x 4 threads) [0.8|400 MiB];
    - **Conveyor Poller** → (3 x 4 threads) [0.5|250 MiB];
    - **Conveyor Finisher** → (1 x 2 threads) [1(*1.5)|250 (*500) MiB];
    - **Hermes** → (1) [0.1|200 MiB];
    - **Judge Injector** → (2) [0.1 (*0.8)|200 (*400) MiB];
    - **Judge Evaluator** → (3 x 3 threads) [2|3 GiB];
    - **Judge Repairer** → (2 x 5 threads) [1|0.8 (*6) GiB];
    - **Judge Cleaner** → (2 x 5 threads) [1|400 MiB];
    - **Reaper2** → (2 x 4 threads) [0.4|400 (*800) MiB];
    - **Transmogrifier** → (1) [0.1|200 MiB];
    - **Undertaker** → (1) [1|400 MiB].
  - Total → (29) [21.3|21.60 GiB].
    → *28.8|38.75 GiB
  - OpenStack VMs → (6 nodes) [8|16 GiB].

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