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

Project Title	European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructure
Project Acronym	ESCAPE
Grant Agreement No	824064
Instrument	Research and Innovation Action (RIA)
Topic	Connecting ESFRI infrastructures through Cluster projects (INFRA-EOSC-4-2018)
Start Date of Project	01.02.2019
Duration of Project	42 Months
Project Website	projectescape.eu

Milestone 8: Initial pilot data lake with at least 3 core data centres

Work Package	WP2, DIOS
Lead Author (Org)	Xavier Espinal, Rosie Bolton
Contributing Author(s) (Org)	Andrea Ceccanti, Paul Millar, Yan Grange
Due Date	31.07.2020
Date	30.07.2020
Version	1.0

Dissemination Level

☒ PU: Public
☐ PP: Restricted to other programme participants (including the Commission)
☐ RE: Restricted to a group specified by the consortium (including the Commission)
☐ CO: Confidential, only for members of the consortium (including the Commission)

Versioning and contribution history

Version	Date	Authors	Notes
0.1	16.07.2020	Xavier Espinal (CERN)	Skeleton
1.0	29.07.2020	Rosie Bolton (SKAO)	Final Draft

Disclaimer

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 WP2 project milestone 2.2: Initial datalake pilot with at least 3 sites.

Explanation of the work carried out

A Pilot Datalake. In the first 6 months of the project, WP2 defined the architecture of the datalake based on the needs of the ESFRIs in terms of data management according to FAIR principles. We reviewed the existing technologies from different European projects and the ESFRIs themselves and understood if and how they would be suitable for implementing building blocks of the architecture reference implementation. The architecture document and implementation plan was delivered in September 2019 and the implementation of the data lake pilot started after that. The pilot data lake relies on services provided by the data centres of WP2 partners and orchestrates them through a common service layer for all ESFRIs.

The work package is well on target - indeed, we have ten sites in our datalake, well above our target of 3 by this stage, and we are developing the teams, tools and technologies needed to test and improve the way we use these sites to meet our experimental use cases.

A working pilot. The pilot datalake advanced quickly and to date (July 2020) includes **ten storage endpoints**, covering the majority of the partners: INFN-CNAF, INFN-ROMA, INFN-Napoli, DESY, SURF-SARA, IN2P3-CC, CERN, IFAE-PIC, CNRS-LAPP and GSI. This is very good, as it allows us to start far more ambitious data transfer and management tests than a simple three-site datalake (i.e. the milestone target) would have done. The ESCAPE datalake harnesses a variety of storage technologies through the orchestration layer: dCache, DPM, XRootD, EOS, StoRM covering also distributed and federated storage systems as well as traditional/local installations. More information and details about the storage systems involved can be found at [1]. For the pilot datalake, the orchestration layer has been consolidated, a dedicated RUCIO instance for Data Management.

Rucio is the key data management element for the WP2 datalake - it is an open source software (https://rucio.github.io/) with a proven track record in managing billions of files across over 100 data centres at the half-exabyte capacity (for the CERN ATLAS experiment). We have deployed our own ESCAPE rucio instance, accessible to ESCAPE partners, and our work going forward is to test the suitability of this technology to perform for the different use cases and data storage paradigms that the different experiments in ESCAPE require.

Data transfers from within our Rucio instance are made using an FTS instance at CERN (FTS - the File Transfer Service enables third party copies using a range of different protocols - see https://cern.service-now.com/service-portal?id=service_element&name=file-transfer). We are also able to use FTS outside Rucio, so we can understand performance tradeoffs associated with the use of Rucio.

The ESCAPE datalake is currently populated with some real data from several experiments: LOFAR, LSST, ATLAS and CMS data (in moderate data volumes). Workflow and pipeline implementation to test data access is progressing and has started exercising data access. A simple implementation of storage Quality of Service (QoS) across the datalake is in place and data lifecycles being implemented.

Monitoring and operations. Due to the increasing scale of the datalake and the active involvement from the partners and experiments it was clear the scale of operations needed to be coordinated. We set-up a team to look after operations and deployment, not only for the short term challenges but also to ensure sustainability of the datalake infrastructure and tools after the project finishes. The idea behind is to foster knowledge transfer to sites and experiments, so that all WP2 partners can become experienced in the technologies supporting WP2. This will maximise the likelihood of the ESCAPE initiatives to be adopted in the future by the ESFRIs (see the original proposal at [2]). Monitoring is being developed to reflect the actual status of the datalake infrastructure and give the necessary information to the DepOps team to follow-up activities, chase technical issues and follow-up new deployments and initiatives. See [Fig. 1] shows a snapshot of the main monitoring page, this actually comprises. Live monitoring available at [3,4] :

- Data transfers view: site efficiency based on data in/out using different protocols
- Data volume and number of files transferred
- Network status: a tool to show real time network performance is in place, this is based on perfSONAR server that has been deployed at the majority of sites.
- Site reporting tool being evaluated for deployment to show the storage volume and occupancy provided to the datalake, this tool is based on a Storage Reporting JSON mechanism.

Figures 1 and 2 show examples of the DepOps team's monitoring dashboards (which are still under development) - we are building dashboards to enable rapid and easy monitoring of the overall datalake status, the network health and the success of test transfers.

Data, content delivery and caching: the prototyping phase is well on track and three sites deployed a proof of concept to evaluate XCache technology to integrate compute resources with the datalake. XCache provides the needed functionality and we are currently testing its scalability and multi-VO usability. The content facilitates data processing from compute resources inside the datalake but also enables inclusion of storage-less sites, including cloud resources and HPC centres (see figure 4). There has been an effort made towards a vanilla installation (experiment-unbiased) caching service, easily deployable by the partners' sites, and a monitoring dashboard is available (figure 5). Testbeds have been deployed at CERN, CC-IN2P3, LAPP, and INFN-CNAF.
Real data processing from/to the datalake have been demonstrated at:

- LAPP for ATLAS Open Data, with a simple analysis example using data in the ESCAPE datalake to search for the Higgs boson in the H --> yy (Higgs into two photons) decay channel. This was disseminated through the ESCAPE communication team [5]
- INFN-CNAF demonstrated the ability to process data from /to the datalake for CMS, using an Open Data approach but also via a prototype that covers the use case of embargoed data (only accessible by a few people, i.e. CMS experiment) [6]

Addressing experiment/ESFRI requirements:

- IFAE-PIC: a prototype has been deployed for a data injector to store data from an experiment source into the datalake. This is targeting the gamma-ray telescope use-case with telescopes located on the island of La Palma (CTA and MAGIC telescopes). This exercise is an excellent proof of concept (PoC) for data upload and distribution from external storage with particular demands (e.g. remote locations with non-deterministic LFNs) to the datalake infrastructure. Work on this PoC is being regularly presented at the WP2 coordination meetings [7])

Information system (CRIC). A centralised information system has been set up. This is the central place for the definition of storage endpoints (URLs), protocols supported at the sites, storage access preferences, etc. It is interfaced with RUCIO, meaning that sites can deploy new storage endpoints by defining them in CRIC and then they are synchronised with RUCIO, making the datalake aware of these new resources (Fig. 3)

Authentication, Authorization and Identity management (AAI): the ESCAPE IAM. An instance of IAM (Identity and Access Management) in support of the ESCAPE Data lake prototype has been deployed at INFN, allowing us to give and control access to the ESCAPE data lake resources. The instance, providing support for token-based and legacy voms-based authorization, is deployed on a Kubernetes cluster running at INFN CNAF and is available at [9]. The ESCAPE IAM instance is configured to support authentication via EduGAIN, Google and X.509 certificates. Documentation describing the IAM setup is provided at [10].

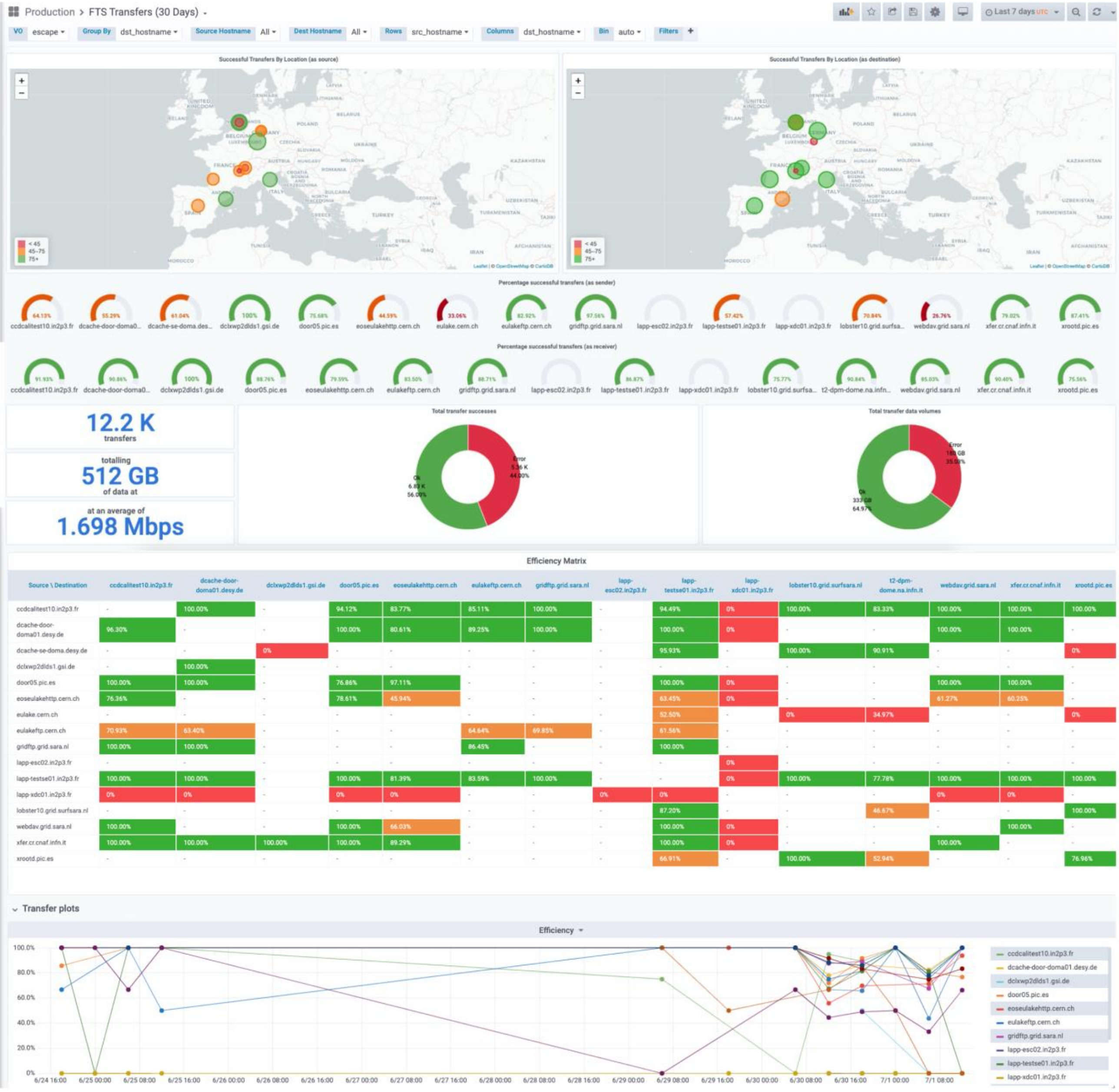
Notable features have been implemented in IAM in support of LHC and other communities' authentication and authorization requirements, with the highlights being:

- Allow disabling local authentication (or limiting only to VO administrators), which allows the IAM to be configured to completely rely on brokered authentication from an external identity provider;

- Requiring external authentication on the registration endpoint, which requires applicants to authenticate against an external identity provider before being allowed to submit a registration request;
- User account end time management and pluggable account validation support, used to properly implement integration between IAM and the CERN Human Resource Database;
- Support for multiple token profiles, which allows the same IAM instance to support different token profiles depending on the configuration defined at client level.

These features are planned to be released in IAM v1.6.0 scheduled for the end of July 2020, but are already deployed in production for some IAM instances hosted at CNAF and at CERN.

Figures



Production > FTS Transfers (30 Days)

VO escape · Group By dst_hostname · Source Hostname All · Dest Hostname All · Rows src_hostname · Columns dst_hostname · Bin auto · Filters +

Last 7 days UTC

12.2 K transfers

totalling 512 GB of data at

at an average of 1.698 Mbps

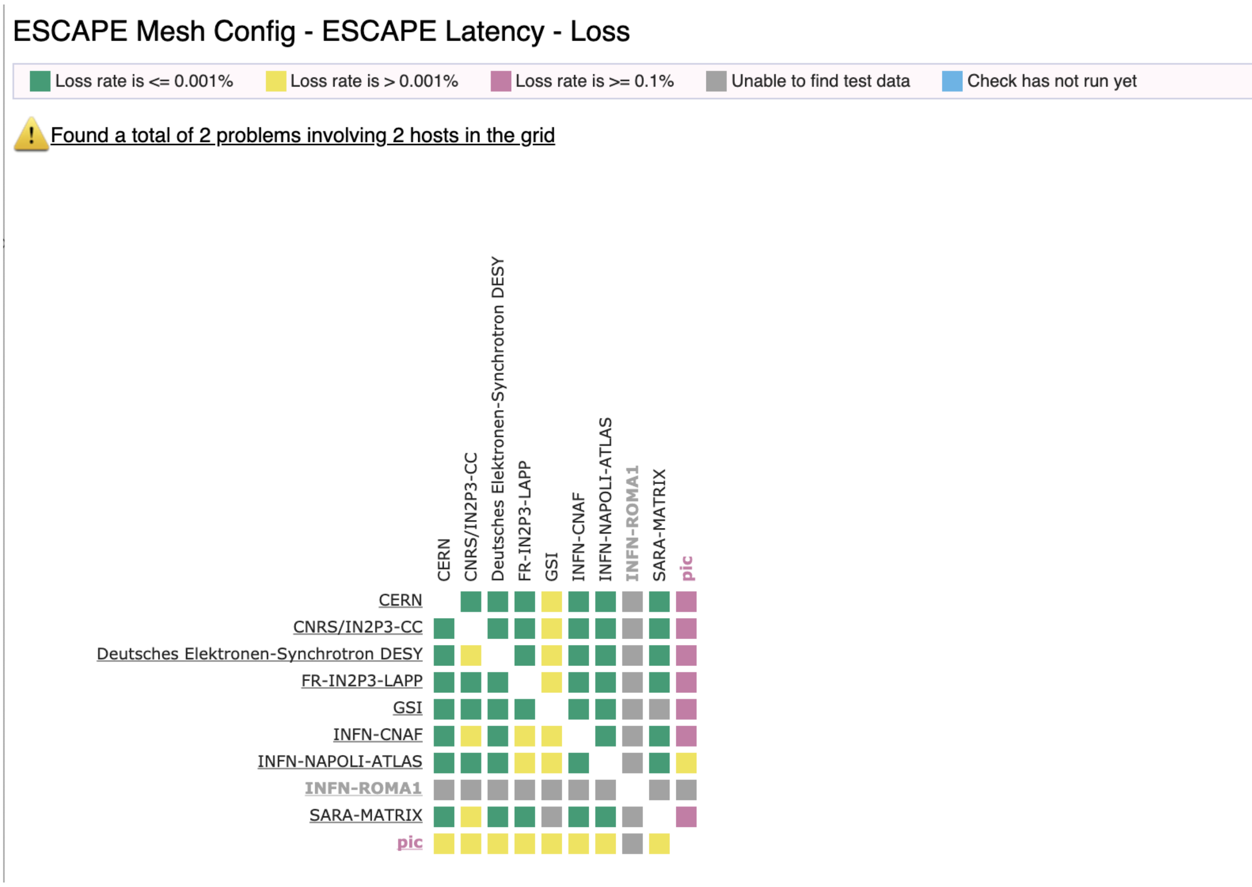
Efficiency Matrix

Source \ Destination	ccdcalitest10.in2p3.fr	dcache-door-doma01.desy.de	dclxwp2dlds1.gsi.de	door05.pic.es	eoseulakehttp.cern.ch	eulakeftp.cern.ch	gridftp.grid.sara.nl	lapp-esc02.in2p3.fr	lapp-testse01.in2p3.fr	lapp-xdc01.in2p3.fr	lobster10.grid.surfsara.nl	t2-dpm-dome.na.infn.it	webdav.grid.sara.nl	xfer.cr.cnaf.infn.it	xrootd.pic.es
ccdcalitest10.in2p3.fr	-	100.00%	-	94.12%	83.77%	85.11%	100.00%	-	94.49%	0%	100.00%	83.33%	100.00%	100.00%	100.00%
dcache-door-doma01.desy.de	96.30%	-	-	100.00%	80.61%	89.25%	100.00%	-	100.00%	0%	-	-	100.00%	100.00%	-
dcache-se-doma.desy.de	-	-	0%	-	-	-	-	-	95.93%	-	100.00%	90.91%	-	-	0%
dclxwp2dlds1.gsi.de	-	100.00%	-	-	-	-	-	-	-	-	-	-	-	-	-
door05.pic.es	100.00%	100.00%	-	76.86%	97.11%	-	-	-	100.00%	0%	-	-	100.00%	100.00%	-
eoseulakehttp.cern.ch	76.36%	-	-	78.61%	45.94%	-	-	-	63.45%	0%	-	-	61.27%	60.25%	-
eulake.cern.ch	-	-	-	-	-	-	-	-	52.50%	-	0%	34.97%	-	-	0%
eulakeftp.cern.ch	70.93%	63.40%	-	-	-	64.64%	69.85%	-	61.56%	-	-	-	-	-	-
gridftp.grid.sara.nl	100.00%	100.00%	-	-	-	86.45%	-	-	100.00%	-	-	-	-	-	-
lapp-esc02.in2p3.fr	-	-	-	-	-	-	-	-	-	0%	-	-	-	-	-
lapp-testse01.in2p3.fr	100.00%	100.00%	-	100.00%	81.39%	83.59%	100.00%	-	-	0%	100.00%	77.78%	100.00%	100.00%	100.00%
lapp-xdc01.in2p3.fr	0%	0%	-	0%	0%	-	-	0%	0%	-	-	-	0%	0%	-
lobster10.grid.surfsara.nl	-	-	-	-	-	-	-	-	87.20%	-	-	46.67%	-	-	100.00%
webdav.grid.sara.nl	100.00%	-	-	100.00%	66.03%	-	-	-	100.00%	0%	-	-	-	100.00%	-
xfer.cr.cnaf.infn.it	100.00%	100.00%	100.00%	100.00%	89.29%	-	-	-	100.00%	0%	-	-	100.00%	-	-
xrootd.pic.es	-	-	-	-	-	-	-	-	66.91%	-	100.00%	52.94%	-	-	76.96%

Transfer plots

Figure 1: Snapshot of one of our ESCAPE datalake dashboards (still under development). The ESCAPE pilot datalake in action, transfer volume, site transfer efficiencies, etc. (Note that, though some sites are showing as unresponsive in this snapshot, this is not generally the case - the point is that we are able to quickly identify issues with sites and follow up as needed.)

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.



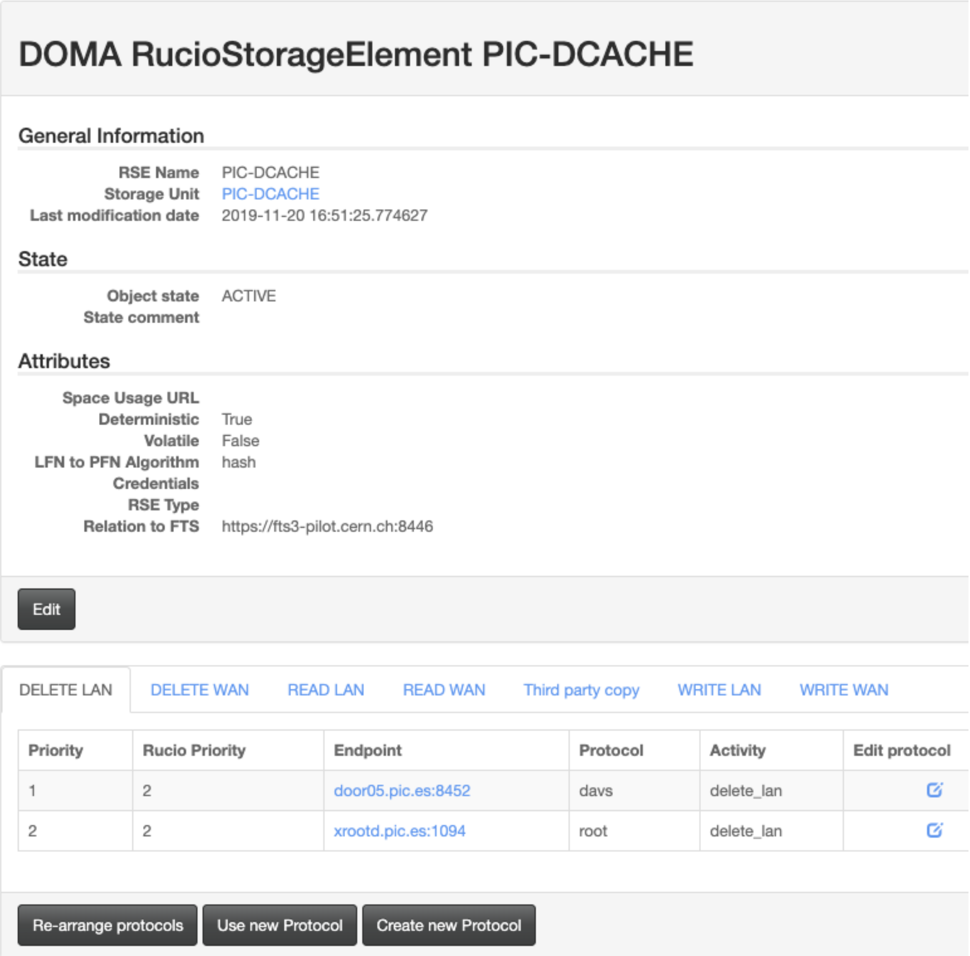
ESCAPE Mesh Config - ESCAPE Latency - Loss

🟩 Loss rate is <= 0.001% 🟨 Loss rate is > 0.001% 🟪 Loss rate is >= 0.1% ▨ Unable to find test data 🟦 Check has not run yet

⚠ Found a total of 2 problems involving 2 hosts in the grid

	CERN	CNRS/IN2P3-CC	Deutsches Elektronen-Synchrotron DESY	FR-IN2P3-LAPP	GSI	INFN-CNAF	INFN-NAPOLI-ATLAS	INFN-ROMA1	SARA-MATRIX	pic
CERN		🟩	🟩	🟩	🟨	🟩	🟩	▨	🟩	🟪
CNRS/IN2P3-CC	🟩		🟩	🟩	🟨	🟩	🟩	▨	🟩	🟪
Deutsches Elektronen-Synchrotron DESY	🟩	🟨		🟩	🟨	🟩	🟩	▨	🟩	🟪
FR-IN2P3-LAPP	🟩	🟩	🟩		🟨	🟩	🟩	▨	🟩	🟪
GSI	🟩	🟩	🟩	🟩		🟩	🟩	▨	▨	🟪
INFN-CNAF	🟩	🟨	🟩	🟨	🟨		🟩	▨	🟩	🟪
INFN-NAPOLI-ATLAS	🟩	🟩	🟩	🟨	🟨	🟩		▨	🟩	🟨
INFN-ROMA1	▨	▨	▨	▨	▨	▨	▨		▨	▨
SARA-MATRIX	🟩	🟨	🟩	🟩	▨	🟩	🟩	▨		🟪
pic	🟨	🟨	🟨	🟨	🟨	🟨	🟨	▨	🟨	

Figure 2: Example taken from the ESCAPE PerfSONAR dashboard (http://maddash.aglt2.org/maddash-webui/index.cgi?dashboard=ESCAPE%20Mesh%20Config) showing that we have ten sites with perfSONAR boxes installed, and with automated tests running (results here show latency, but throughput and traceroute statistics are also available on the same dashboard).



DOMA RucioStorageElement PIC-DCACHE

General Information

RSE Name	PIC-DCACHE
Storage Unit	PIC-DCACHE
Last modification date	2019-11-20 16:51:25.774627

State

Object state	ACTIVE
State comment	

Attributes

Space Usage URL	
Deterministic	True
Volatile	False
LFN to PFN Algorithm	hash
Credentials	
RSE Type	
Relation to FTS	https://fts3-pilot.cern.ch:8446

Priority	Rucio Priority	Endpoint	Protocol	Activity	Edit protocol
1	2	door05.pic.es:8452	davs	delete_lan	✎
2	2	xrootd.pic.es:1094	root	delete_lan	✎

Figure. 3 Snapshot taken from the ESCAPE information system (CRIC). This snapshot shows the view of one of the storage endpoints in the ESCAPE datalake. URL and prioritized supported protocols.

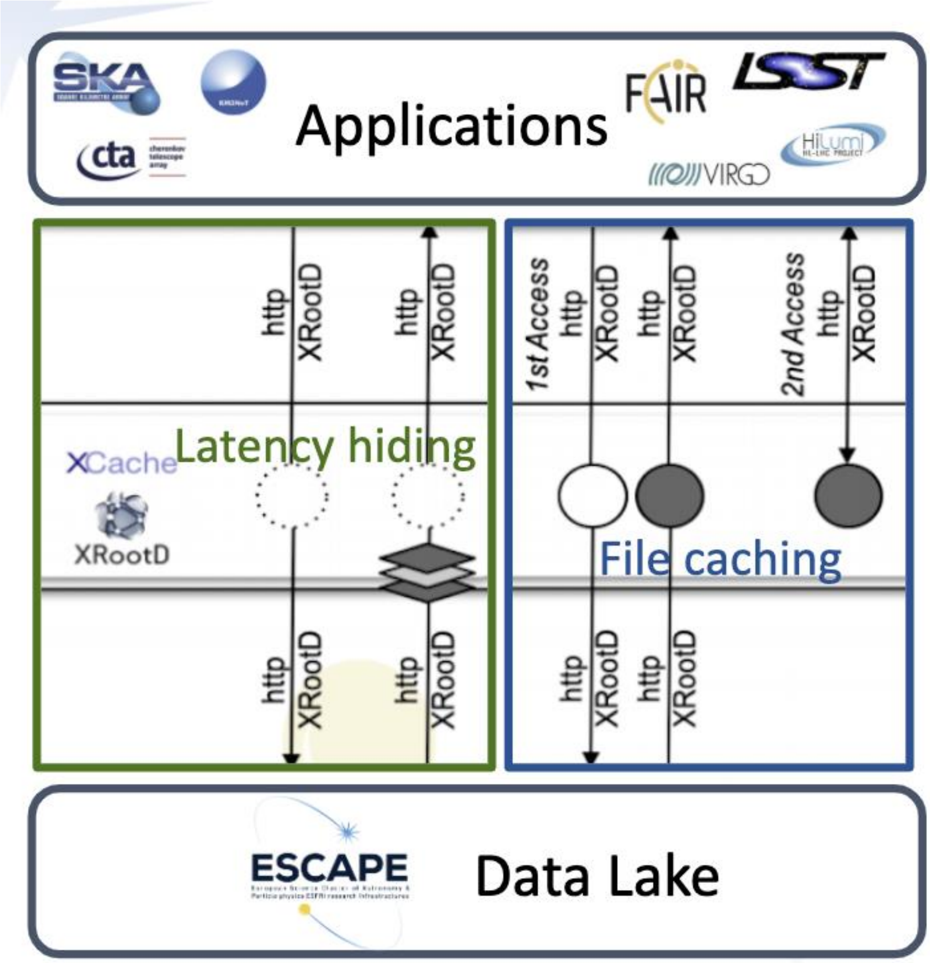


Figure 4: Caching infrastructure and the two main benefits to operate caches: latency hiding and file reusability.

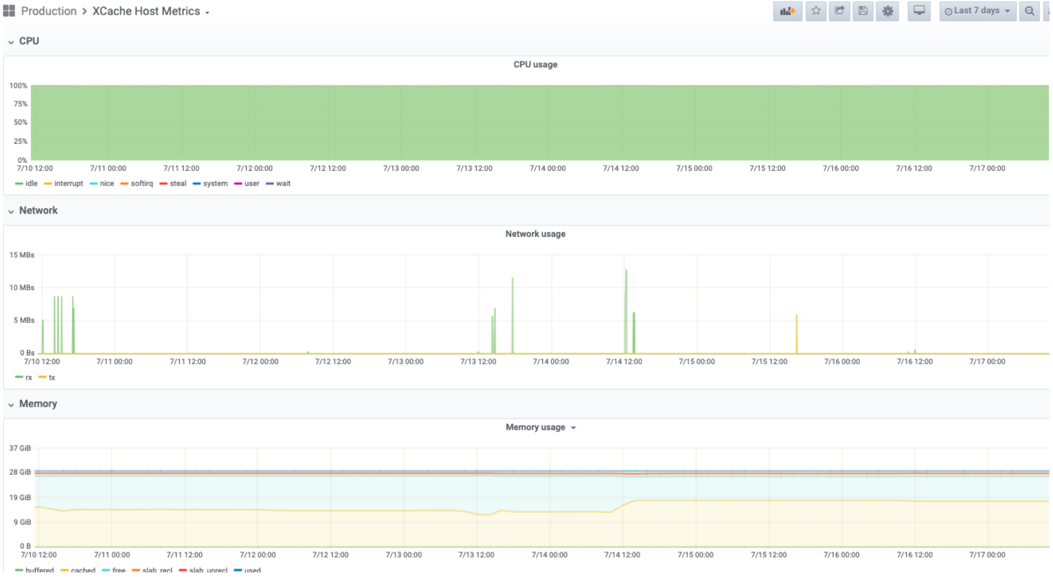


Figure. 5 Caching services monitoring are ready. Metrics are collected from the XCache hosts.

References:

[1] Datalake Components, storage endpoints, protocols
https://wiki.escape2020.de/index.php/WP2_-_DIOS#3_Datalake_Components_and_Reference_Imp
[2] DepOps team proposal:
https://indico.in2p3.fr/event/21321/lementation
[3] ESCAPE Pilot Datalake Live Monitoring: https://monit-grafana.cern.ch/d/cHBQ2NjWz/escape-home
[4]Live "view" of a datalake activity: transfers in-flight, data volume, throughputs, etc.
https://monit-grafana.cern.ch/d/000000420/fts-transfers-30-days?orgId=51 (access details here)

[5] First analysis experience with the ESCAPE data lake: https://projectescape.eu/news/first-analysis-experience-escape-data-lake
[6] Datalake data access: CMS experiment use case for open and embargoed data:
https://indico.in2p3.fr/event/19942
[7] PIC-IFAE gamma-ray telescope datalake data injector: https://indico.in2p3.fr/event/19943
and https://indico.in2p3.fr/event/21820
[8] ESCAPE Information System (CRIC): http://escape-cric.cern.ch/
[9] ESCAPE IAM instance: https://iam-escape.cloud.cnaf.infn.it/
[10] ESCAPE IAM documentation: https://indigo-iam.github.io/escape-docs/
[11] Caching activities presentation at the ESCAPE progress meeting 26-27 Feb 2020 (Brussels)
https://indico.in2p3.fr/event/20203/contributions/80150/attachments/57599/77032/RiccardoDiMaria_H2020ESCAPEProgressMeeting_XCache.pdf