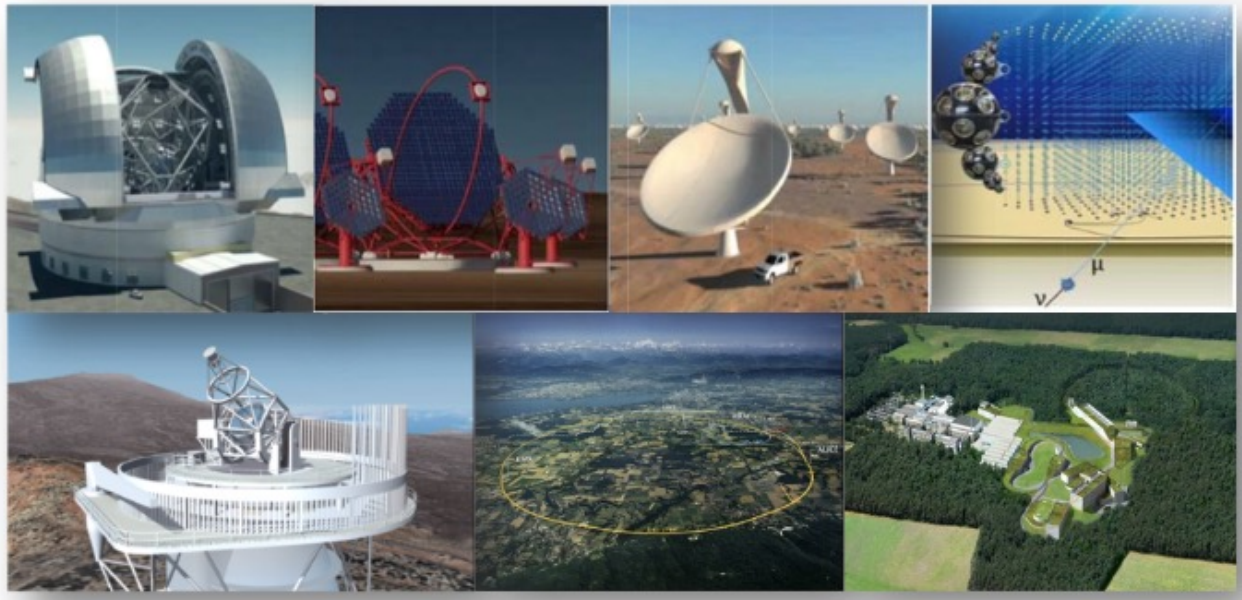
ESCAPE

European Science Cluster of Astronomy & Particle physics
ESFRI research infrastructures

Horizon 2020
funded project

EUROPEAN OPEN SCIENCE CLOUD



Data centres: CERN, INFN, DESY, GSI, Nikhef, SURFSara, RUG, CCIN2P3, PIC, LAPP, INAF

Goals:

Prototype an infrastructure adapted to the Exabyte-scale needs of the large science projects.

Ensure the sciences drive the development of the EOSC

Address *FAIR* data management

Science Projects

HL-LHC	SKA
FAIR	CTA
KM3Net	JIVE-ERIC
ELT	EST
EURO-VO (LSST)	EGO-VIRGO (CERN,ESO)

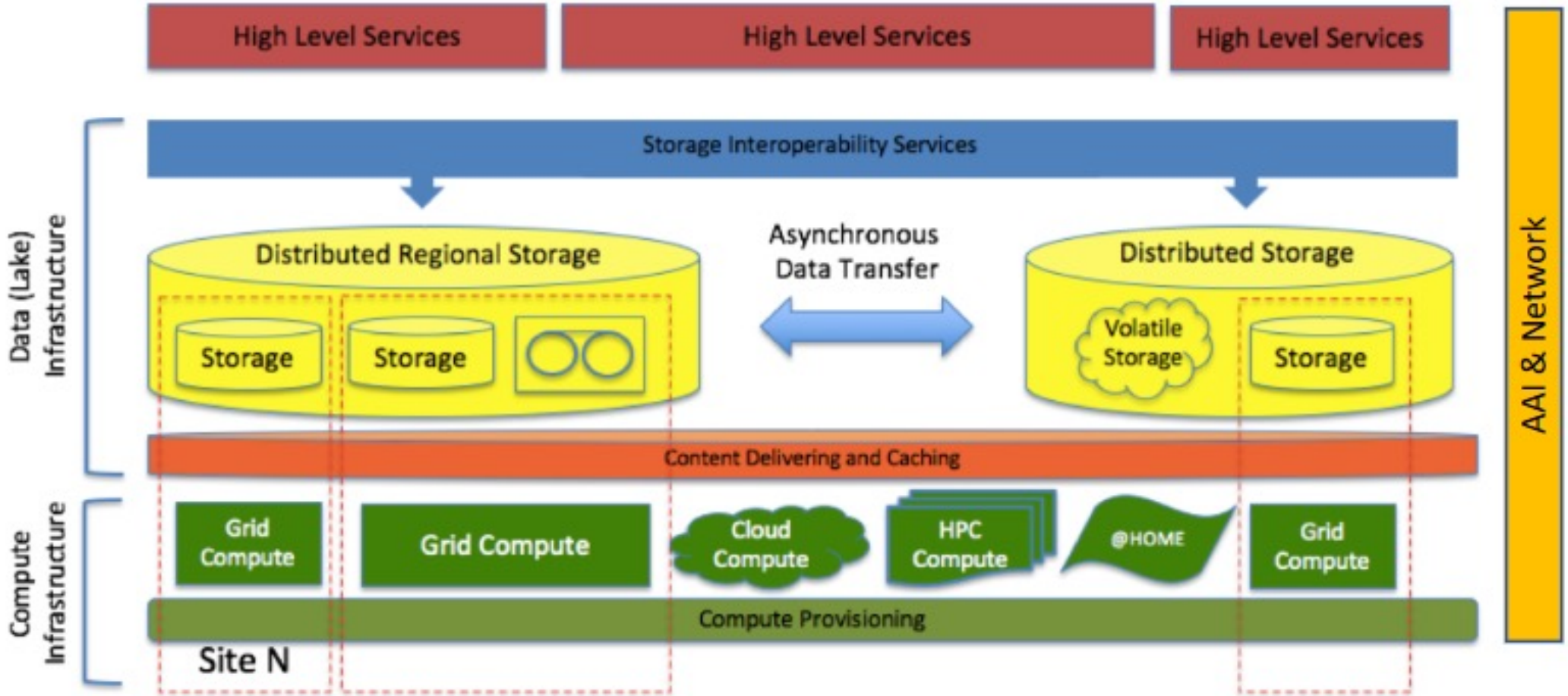
FAIR: findability, accessibility, interoperability, and reusability

WP2: Data Infrastructure for Open Science (DIOS)

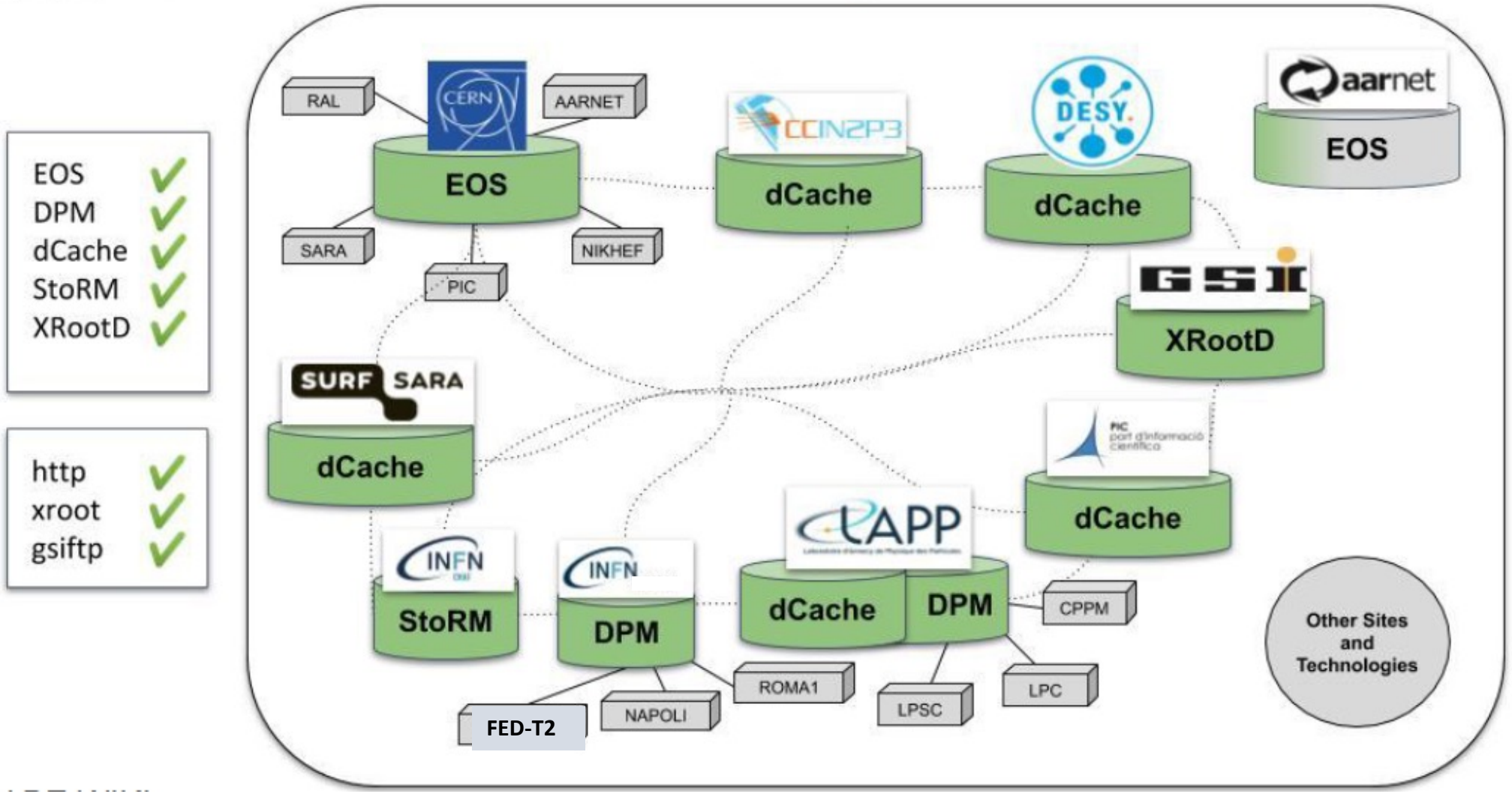
- DIOS is a federated data infrastructure of open access data that enables large national research data centres to work together and build a robust cloud-like service to scale up to multi-Exabyte needs.
- All of the communities are enabled to make the scientific data available in an accessible and transparent way across Europe
- DIOS works in synergy and complementing the work of WLCG DOMA: to define, integrate and commission an ecosystem of tools and services to build a data lake

The Scientific-Data Lake is a policy-driven, reliable, and distributed data infrastructure capable of managing Exabyte-scale data sets, and able to deliver data on-demand at low latency to all types of processing facilities.

A common set of tools is being adopted by different and multi-disciplinary scientific communities to stablish a coherent orchestration layer for the Research Infrastructures (RI) and for the sites providing resources to one or many of these RIs. The orchestration provided by these common set of tools include data management, data transfer and a common Identity Management, enabling a global vision of a single data pool easing the implementation of policies, rules and data life-cycles acting on the Scientific-Data Lake infrastructure as a whole, meaning all sites are working and seen as one.


High Level Services
High Level Services
High Level Services
Storage Interoperability Services
Data (Lake) Infrastructure
Distributed Regional Storage
Storage
Storage
Asynchronous Data Transfer
Distributed Storage
Volatile Storage
Storage
Content Delivering and Caching
Compute Infrastructure
Grid Compute
Grid Compute
Cloud Compute
HPC Compute
@HOME
Grid Compute
Compute Provisioning
Site N
AAI & Network

DIOS infrastructure – Current Status

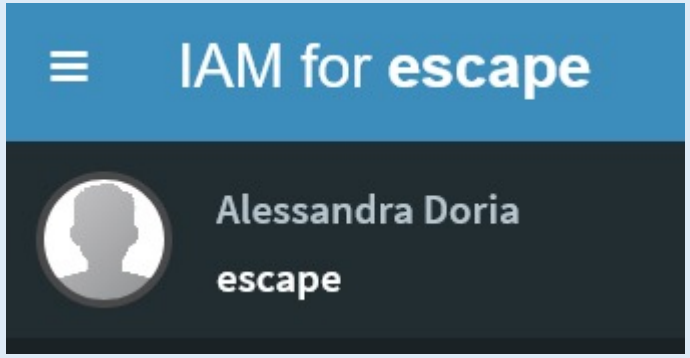


ESCAPE ha scelto Rucio per l'orchestrazione e il management dei dati (RUCIO CLI e Jupyter RUCIO extension) e FTS per i trasferimenti

src \ dst	DESY-DCACHE	SARA-DCACHE	PIC-DCACHE	EULAKE-1	LAPP-DCACHE	IN2P3-CC-	CNAF-STORM	ALPAMED-DPM	GSI-ROOT	INFN-NA-	LAPP-WEBDAV	INFN-NA-	INFN-ROMA1	FAIR-ROOT
DESY-DCACHE	NO DATA	100%	100%	85.7%	100%	100%	85.7%	87.5%	NO DATA	100%	0%	100%	100%	100%
SARA-DCACHE	100%	NO DATA	100%	100%	100%	100%	100%	100%	NO DATA	100%	0%	100%	100%	100%
PIC-DCACHE	100%	100%	NO DATA	100%	100%	100%	100%	100%	NO DATA	100%	0%	85.7%	100%	100%
EULAKE-1	100%	100%	100%	NO DATA	100%	100%	100%	100%	NO DATA	100%	0%	85.7%	100%	100%
LAPP-DCACHE	100%	100%	100%	100%	NO DATA	100%	100%	100%	NO DATA	100%	0%	100%	100%	100%
IN2P3-CC-DCACHE	100%	100%	100%	100%	100%	NO DATA	100%	83.3%	NO DATA	100%	0%	100%	100%	100%
CNAF-STORM	83.3%	100%	100%	83.3%	83.3%	100%	NO DATA	83.3%	NO DATA	100%	0%	83.3%	83.3%	100%
ALPAMED-DPM	100%	100%	83.3%	100%	83.3%	100%	83.3%	NO DATA	NO DATA	83.3%	0%	83.3%	83.3%	100%
GSI-ROOT	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	NO DATA
INFN-NA-DPM	100%	83.3%	83.3%	100%	100%	100%	100%	100%	NO DATA	NO DATA	0%	100%	100%	100%
LAPP-WEBDAV	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	NO DATA
INFN-NA-DPM-FED	100%	100%	100%	100%	100%	100%	83.3%	71.4%	NO DATA	100%	0%	NO DATA	100%	100%
INFN-ROMA1	100%	100%	100%	100%	100%	100%	100%	100%	NO DATA	100%	0%	100%	NO DATA	100%
FAIR-ROOT	100%	100%	100%	100%	100%	100%	100%	100%	NO DATA	100%	0%	100%	100%	NO DATA

L’interazione dei servizi con il Data Lake richiede l’autenticazione.

- Per AAI, ESCAPE ha scelto INDIGO Identity and Access Management (IAM), servizio gestito dal CNAF



- Sul nostro DPM, abilitata anche la Token Authentication con IAM, basata su protocollo OIDC (OpenID Connect), che può sostituire l’autenticazione con certificati X509.

- Content delivery and latency hiding: Xcache
- Data Lake Information System: CRIC

Periodicamente si svolgono dei Data Challenge
DAC21 appena terminato

- Run production **Data Management, Processing and Analysis workloads**
  - **Data Management**: acquisition, injection, replication, lifecycles
    - Feasibility of the replication of the samples and transfer between RSEs.
    - Reporting failure or bottlenecks.
    - Cleaning procedure of the data from the Datalake.
  - **Data Processing**: **Production** (= experiment "data preparation"). Use cases: reprocessing, reconstruction, calibration, etc. requiring input and output from/to the Data Lake. Activity requiring potentially "large" resources and scale up ability: computing farms, cloud provisioning, HPC, etc.
  - **Data processing**: **User Analysis.** Use cases: put the "prepared" data at the disposal of users, etc. Activity targeting, "small" resources: analysis platforms, online notebooks, personal computers, laptops, etc.
  - Demonstrate **Data Lake orchestration layer sustainability** after ESCAPE (towards EOSC)
    - Leverage, integrate and use experiment and site's dedicated installations, e.g. RUCIO and FTS.
  - Demonstrate integration and interplay of these several instances in a common Data Lake/storage infrastructure.
- **End-to-End AAI** proven for all sciences and workflows exercised at the DAC21.
  - Assessment ranging from experiments experts, to advanced users to newcomers and to sporadic web-based access (i.e. citizen science => evaluate possibility for WP6 collaboration on "people's science" project).
  - Token-based Authentication. Is it realistic as a target to have at least one workflow running 100% on tokens?

Per ATLAS test condotti da Arturo Sanchez (IN2P3)

We attempt to describe a series of exercises to perform during DAC21

https://docs.google.com/document/d/1ma8dRSc6wrGNKqxn7ZWNjvJ16QNgvQF6mgY7KsCxRck/

- Data “multiplication” where multiple version of the same data is generated, simulating a data-augmentation process
  - Allowing us to have a sustancial quantity data to be injected to the current analysis examples
- Writing of such “multiplied” data back to the Datalake
  - Defining different RSEs → e.g. LAPP/CNRS and Napoli/INFN
- Examples include the analysis of data stored in the Datalake
  - Writing back the results into the Datalake (small files of ~< 1Mb size each)
  - Analysis can be performed using CLI or the JupyterLab UI
- We are creating instructions for users that can be part of the challenge