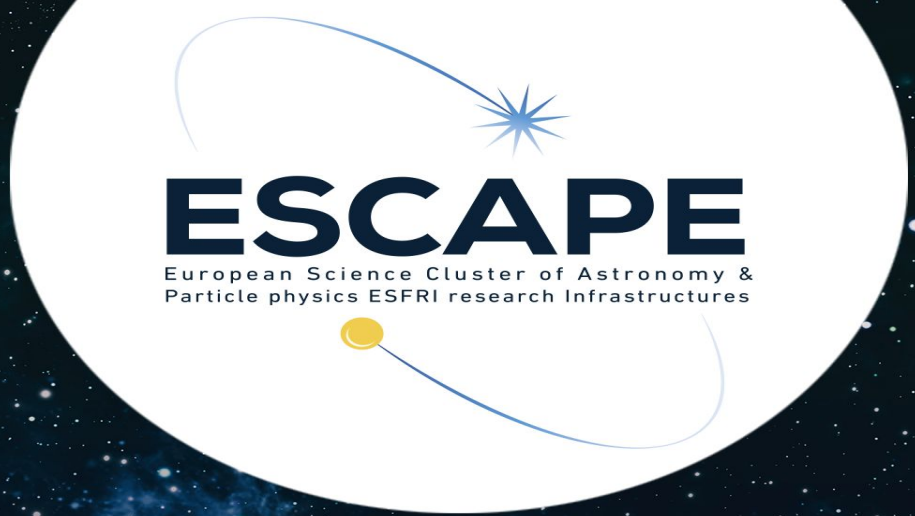


DIOS
Task 2.1 Datalake infrastructure

Xavier Espinal

(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.

T2.1 Datalake infrastructure - status

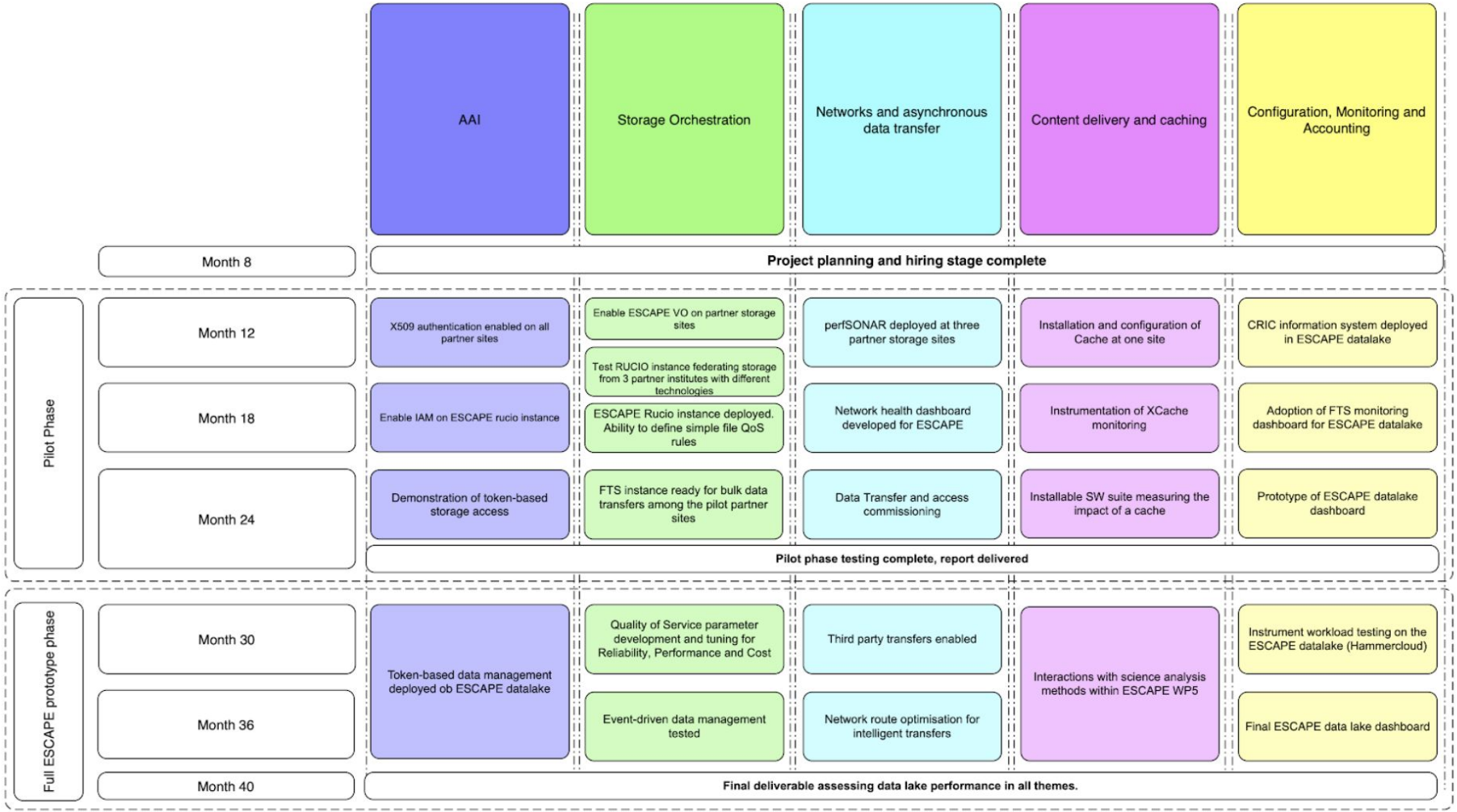
- **Infrastructure**
  - Storage endpoints deployed on 7 sites (out of 9): CERN, INFN, DESY, SARA/Nikhef, IN2P3, LAPP, IFAE/PIC
    - Covering 4 out of 5 storage technologies: DPM, dCache, SToRM and EOS
  - Work in progress: GSI and RUG (xrootd storage)
- **Data transfer and data management services**
  - Dedicated RUCIO instance with on Rucio Storage Element (RSE) per storage endpoint
    - QoS: discussions with some sites and RUCIO team about how to implement it
  - File Transfer Service (FTS) integration (service running at the FTS-pilot instance at CERN)
    - ESCAPE VO recognised, dedicated IAM instance based on x509, token integration WIP
  - Ability to move files with xrootd/http/gsiftp with RUCIO+FTS and CLI/WebUi
    - Simple automated tests ongoing: continuous push+distribute transfers, results collected and displayed
- **Monitoring**: FTS metrics, RUCIO events collected on an ES cluster, visualized through Kibana
- **Caching technologies**: first investigations on XCache (stand-alone installation and monitoring)
- **Information system**: CRIC integration in progress

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

T2.1: Planning (proposal)

- Short-term milestones (month 12, Feb 2020)
  - Storage endpoints **enabled at all 11 participating sites**
  - **Functional tests** enabled (data orchestration done with FTS and RUCIO)
    - Transfer matrix in place, monitoring and error reporting/debugging
  - XCache PoC deployed at one site, able to start collecting measures (ie. cache on/off effects)
  - First version of a centralised information system (CRIC)
- Mid-term milestones (month 18, Aug 2020)
  - Datalake performance tests (efficiencies, bandwidth, scaling, protocols, etc.)
  - Start identification of workloads (Hammercloud)
  - PoC: integration of SA and AUS sites into the datalake infrastructure
  - Prototyping of the overall datalake monitoring dashboard

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


AAI
Storage Orchestration
Networks and asynchronous data transfer
Content delivery and caching
Configuration, Monitoring and Accounting
Month 8
Project planning and hiring stage complete
Pilot Phase
Month 12
X509 authentication enabled on all partner sites
Enable ESCAPE VO on partner storage sites
perfSONAR deployed at three partner storage sites
Installation and configuration of Cache at one site
CRIC information system deployed in ESCAPE datalake
Test RUCIO instance federating storage from 3 partner institutes with different technologies
Month 18
Enable IAM on ESCAPE rucio instance
ESCAPE Rucio instance deployed. Ability to define simple file QoS rules
Network health dashboard developed for ESCAPE
Instrumentation of XCache monitoring
Adoption of FTS monitoring dashboard for ESCAPE datalake
Month 24
Demonstration of token-based storage access
FTS instance ready for bulk data transfers among the pilot partner sites
Data Transfer and access commissioning
Installable SW suite measuring the impact of a cache
Prototype of ESCAPE datalake dashboard
Pilot phase testing complete, report delivered
Full ESCAPE prototype phase
Month 30
Token-based data management deployed ob ESCAPE datalake
Quality of Service parameter development and tuning for Reliability, Performance and Cost
Third party transfers enabled
Interactions with science analysis methods within ESCAPE WP5
Instrument workload testing on the ESCAPE datalake (Hammercloud)
Month 36
Event-driven data management tested
Network route optimisation for intelligent transfers
Final ESCAPE data lake dashboard
Month 40
Final deliverable assessing data lake performance in all themes.