ESCAPE and the DataLake

Tommaso Boccali - INFN Pisa

What is ESCAPE?

- You know as much as I do:

E. Cuoco Chair of the General Assembly

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

Horizon 2020 funded project

EUROPEAN OPEN SCIENCE CLOUD

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)

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

This is "INFN"

This is "INFN"

This is EGO, but in a sense still partially "INFN"

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

WPs

Work package 1: **MIND - Management, Innovation, Networking and Dissemination**

Work package 2: **DIOS - Data Infrastructure for Open Science**

Work package 3: **OSSR - Open-source scientific Software and Service Repository**

Work package 4: **CEVO - Connecting ESFRI projects to EOSC through VO framework**

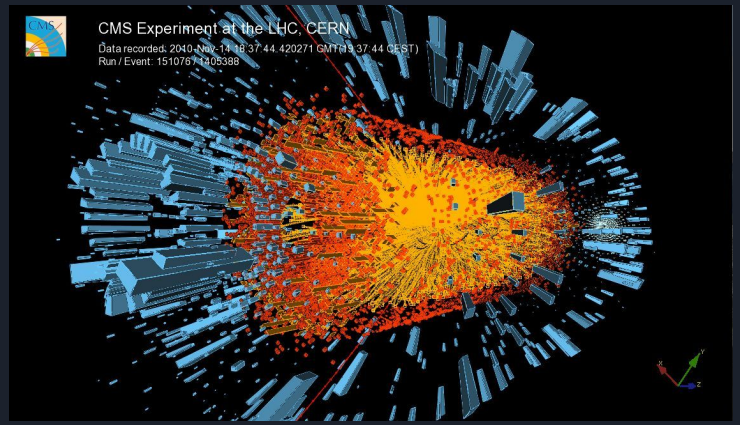
Work package 5: **ESAP - ESFRI Science Analysis Platform**

Work package 6: **ECO - Engagement and Communication**

In a nutshell: WP2 has the goal of providing open access and open science for the scientific communities encompassed within ESCAPE, which represent several very high-volume **data challenges**, as well as the needs of all of the communities in being able to **make the scientific data available in an accessible and transparent way across Europe**

Ok, enough with PR … What is WP2, the DataLake, etc

- All starts with LHC experiments and their evolution in HL-LHC
  - LHC: Large Hadron Collider, in 2018 was colliding some 10^{11} protons every 25 ns, generating 1 Billion collision events per second; this for some 150 days a year
  - Doing the math and considering the # of acquisition channels
    - 10 Billion events/y/experiment
    - To be compared with at least equivalent # of Simulated events
  - 2018 resource needs for the 4 major LHC experiments
    - It worked!

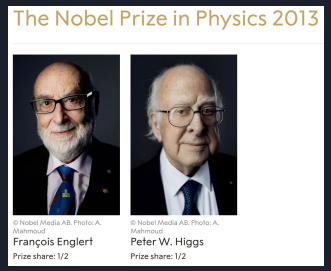


Tier	Pledge Type	SUM
Tier 0	CPU (HEP-SPEC06)	1,270,000
Tier 1	CPU (HEP-SPEC06)	2,302,398
Tier 2	CPU (HEP-SPEC06)	2,818,192
Tier 0	Disk (Tbytes)	96,700
Tier 1	Disk (Tbytes)	221,912
Tier 2	Disk (Tbytes)	210,615
Tier 0	Tape (Tbytes)	272,200
Tier 1	Tape (Tbytes)	499,899

~650k CPU cores

~530 PB disk

~770k PB tape



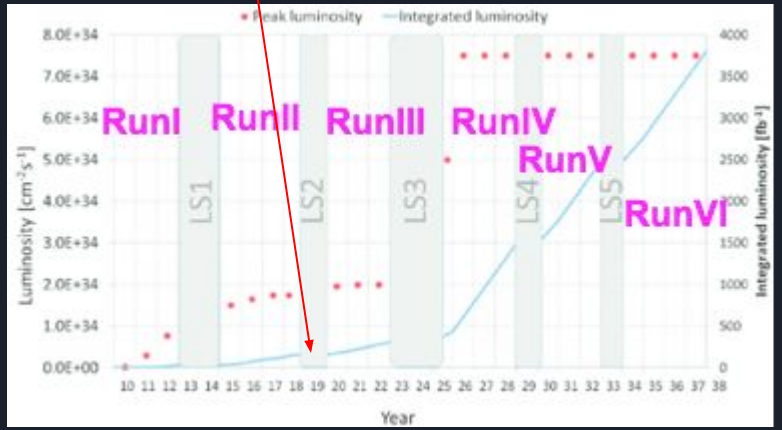
How is that handled today (storage aspect)?

- Hierarchy of Computing centers (from MONARC)
  - A **full copy** of RAW data at the collection site (CERN - Tier0)
  - A **shared second copy** at O(15) regional centers + Simulation (Tier1s)
  - Analysis and MC production facilities (~150 Tier-2s)
- Sites are handled via the Worldwide Lhc Computing Grid (WLCG), and have signed a MoU
- But:
  - Single sites are known to experiments, who have to handle the complexity
  - “I need to write 1 PB of data, where is there free space?”
  - “I have CPUs free in site X, and input data in site Y; what should I do?”
- This is going to get worse ...

HL-LHC

- Somewhere in 2026+
  - Factor up to 6x in event complexity
  - Experiments will need to collect O(10) the number of events
- This has consequences
  - “We collected 5% of the data LHC will give us in the planned future” - yet it seems already sooo long
  - A naive calculation says 60x more resources needed in 10 years
  - **Simply, we cannot afford it with the current model of owned resources / centers**

We are here



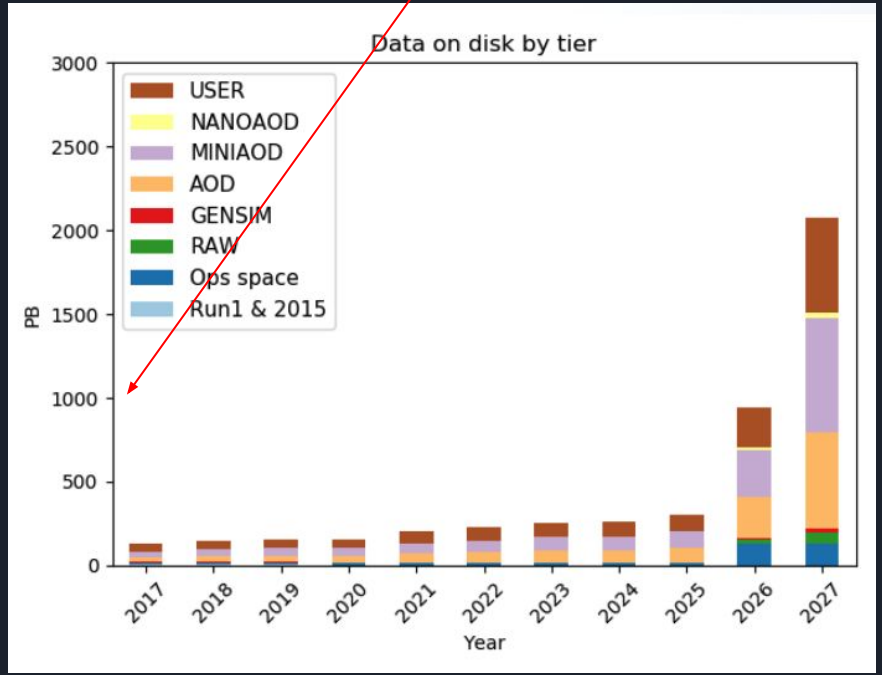
Accelerator schedule/plan:

- Red point are ~scaling with event complexity
- Blue line is ~scaling with total amount of collected data

Current plans

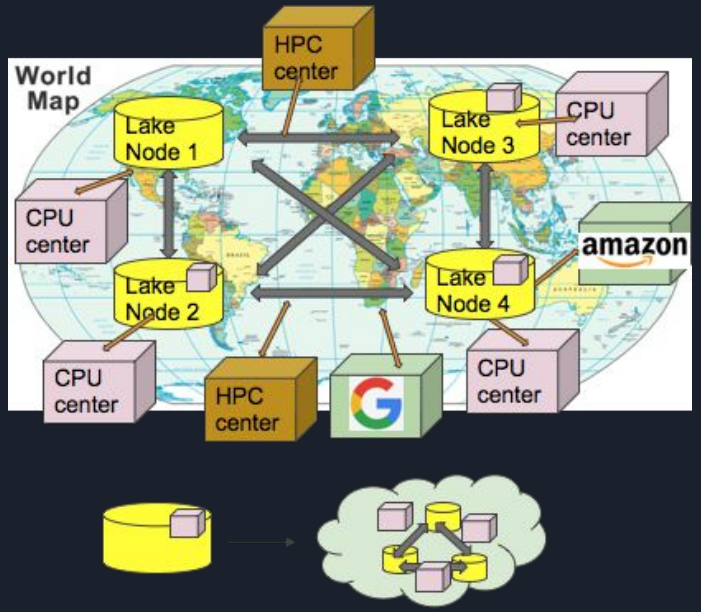
- We already tried to reduce needs
  - Fewer copies of analysis data
  - Fewer reprocessing (“do it well at the first try”)
  - Ideally use SuperComputers (HPC) and Commercial Clouds as a part of the resources
    - And be ready to use literally “any CPU you throw at us”
- But, new problems
  - If we get external CPUs, **how to feed them with data** ?
    - There is no opportunistic data utilization from good Samaritans
  - The numbers are still very frightening
- All in all
  - **Storage problem** is more complicated than CPU problem
  - Still data IS the LHC product: you must keep it safe!

This is 1 Exabyte == 1000 Petabytes



The DataLake

- A DataLake is today's prefered R&D direction for LHC; but it has nothing which prevents it to be used by other sciences
- Idea:
  - Build a **small number of owned data centers**, which can keep the data safe
  - Make them appear as **a logical single entity** (no need for the experiments to know exactly where a file is)
  - .. which means you need to be able and **serve efficiently data** to remote sites, possibly transparently
- The gain
  - The experiment sees fewer sites (at the limit, 1 big logical storage system)
  - **A single copy is ok** (for performance; still want 2 copies of irreproducible data)
- What is needed?
  - A lot of bandwidth to fake remote sites are "as local"
    - The ability to shield a "CPU only site" with caches if the network is not good enough
  - The capability to switch on/off certain route paths on demand

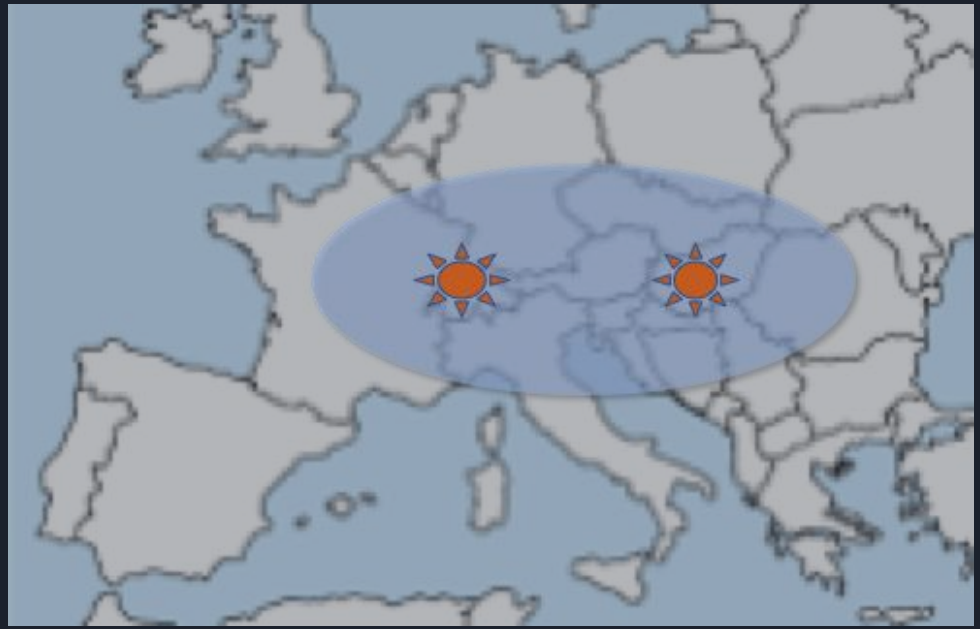


An example from the past/ today …

The main LHC computing center (CERN) has been co-located in Hungary and Geneva since 2015

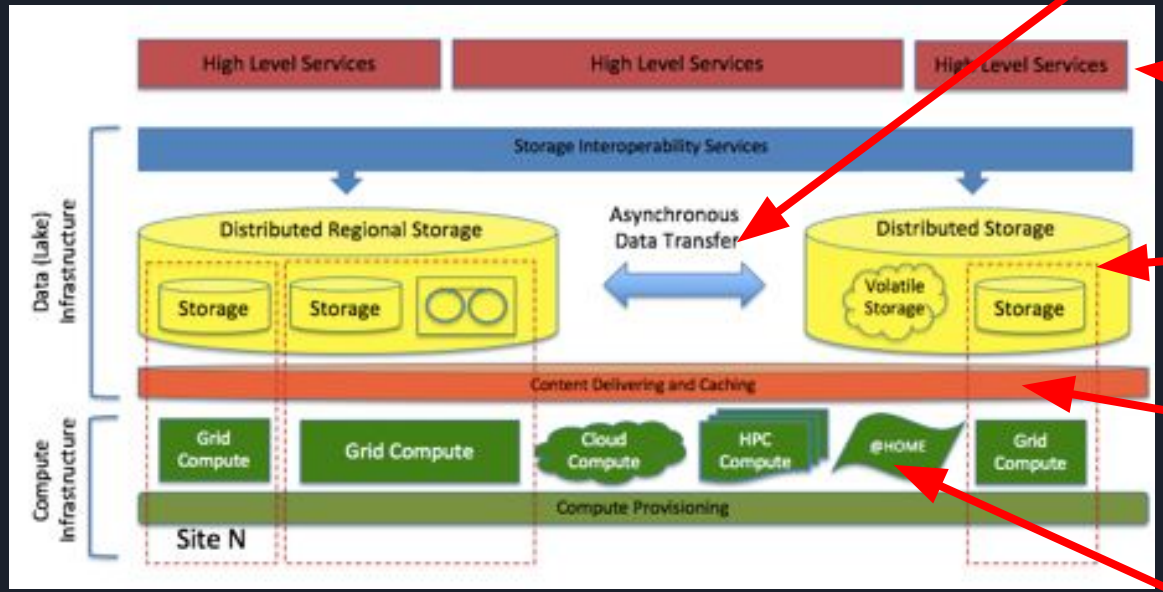
In principle, no need for experiments to care (in practice, not a complete success….)

(technically, the two sites linked with 2x100 Gbit/s; EOS making one disk copy per site)



ESCAPE WP2 - Data Infrastructure for Open Science

This is a Tb/s level connectivity internal to the lake



These are high level services (authentication, authorization, provisioning of network, …)

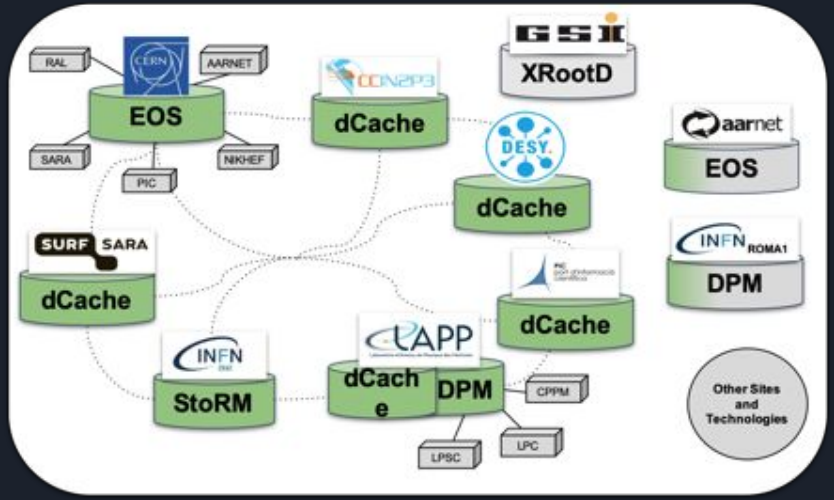
This is the DataLake

These are caches if needed, or direct remote connections

These are the sites / CPU resources (physical or logical)

A few important points

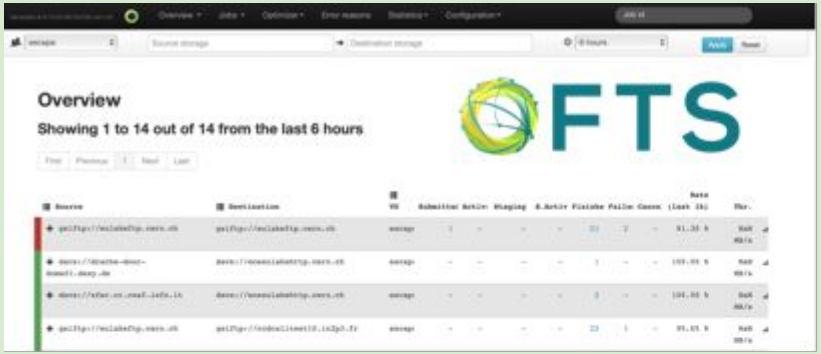
- No intention to force a storage technology
  - You do not need to reinstall anything (we envisage a thin layer on top of existing systems)
- No intention to disrupt legacy code
  - If you used Posix, you must be able to continue with that
- Authentication Authorization Infrastructure of last generation
  - Allows for your **legacy** methods like X509 proxies
- Quality of Service (QoS) is central to the design
  - *"Please save this file to the lake, making sure there are always 2 disk copies and one tape copy"*
  - *"Please make sure this dataset is available to be served at at least 10 GB/s"*
  - *"..."*



Currently deployed (**green**) and foreseen (**grey**) storage services in ESCAPE DataLake.

Some technical details on the proposed solutions

Asynchronous Data Transfer

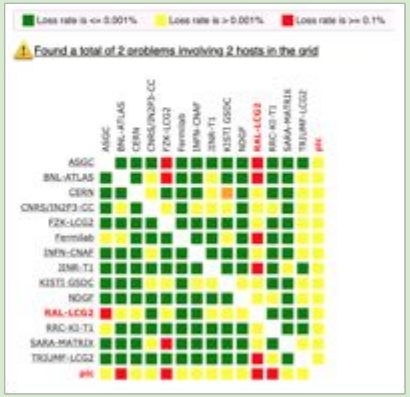


https://fts3-pilot.cern.ch:8449/fts3/ftsmon/#/?vo=escape

FTS is the workhorse for asynchronous point-to-point data transfer in the reference implementation

Planning for a gridFTP-free data lake: HTTP and xrootd

perfSONAR to provide low level network monitoring

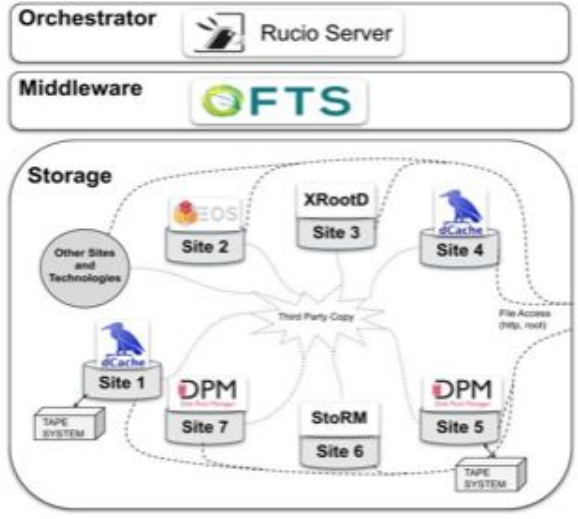


This is the WLCG OPN mesh.
We should build an ESCAPE one

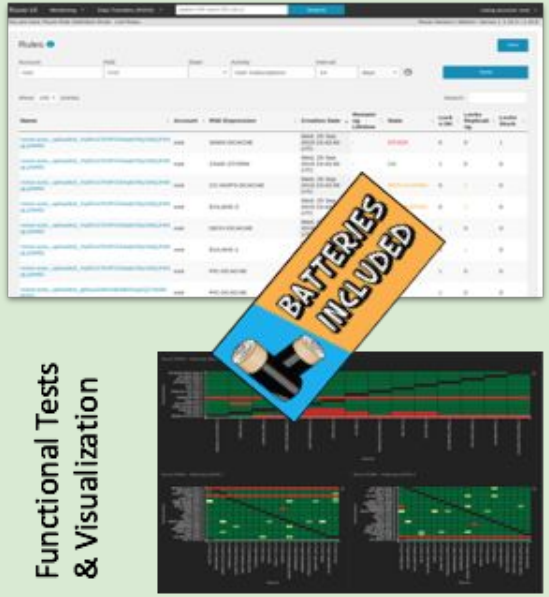
Orchestration Service

Rucio as the Orchestration service in the reference implementation

- File/dataset catalog, rule based engine



Rucio WebUI as interface for ESCAPE users

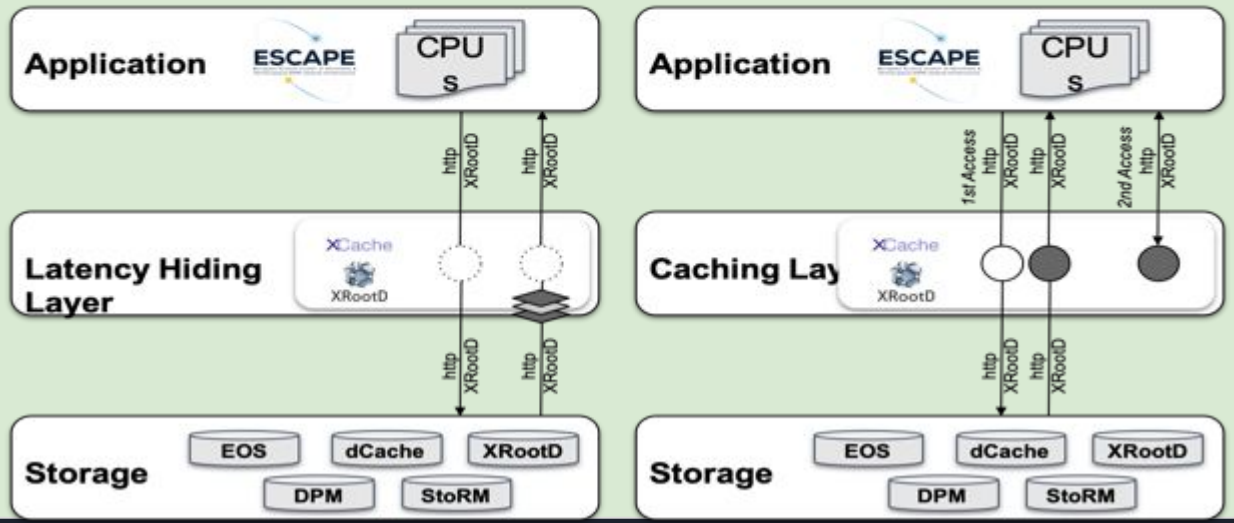


Functional Tests & Visualization

Caching and Latency hiding

xCache technology as reference implementation.

Support HTTP and xrootd protocols



Authentication, Authorization, Identity

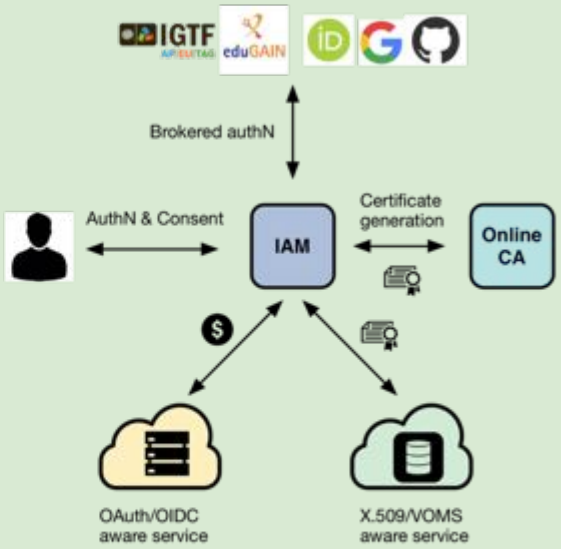
AAI has a much broader scope than WP2

IAM was chosen as reference implementation service for AAI

Token-based authentication rather than X509 as baseline. IAM enables translation X509-to-Token for coexistence period

X509-free DataLake is not realistic in the timescale of ESCAPE (IMHO)

Some fully token based use cases however are in range

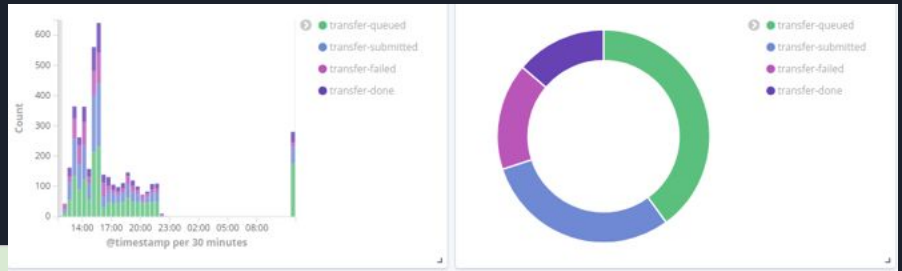


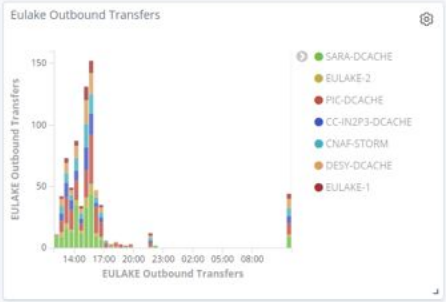
Leveraging the work in the Federated Identity Management For Research (FIM4R) initiative

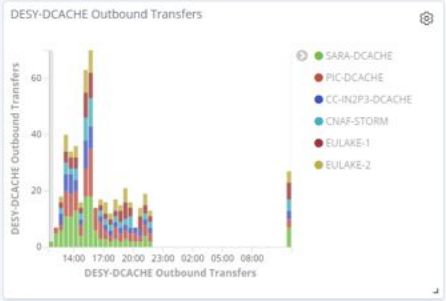
Monitoring

Plan to build an ESCAPE DataLake dashboard, based on open source tools

Most of the information (Rucio, FTS) are already collected in ElasticSearch and can be exposed e.g. via Kibana



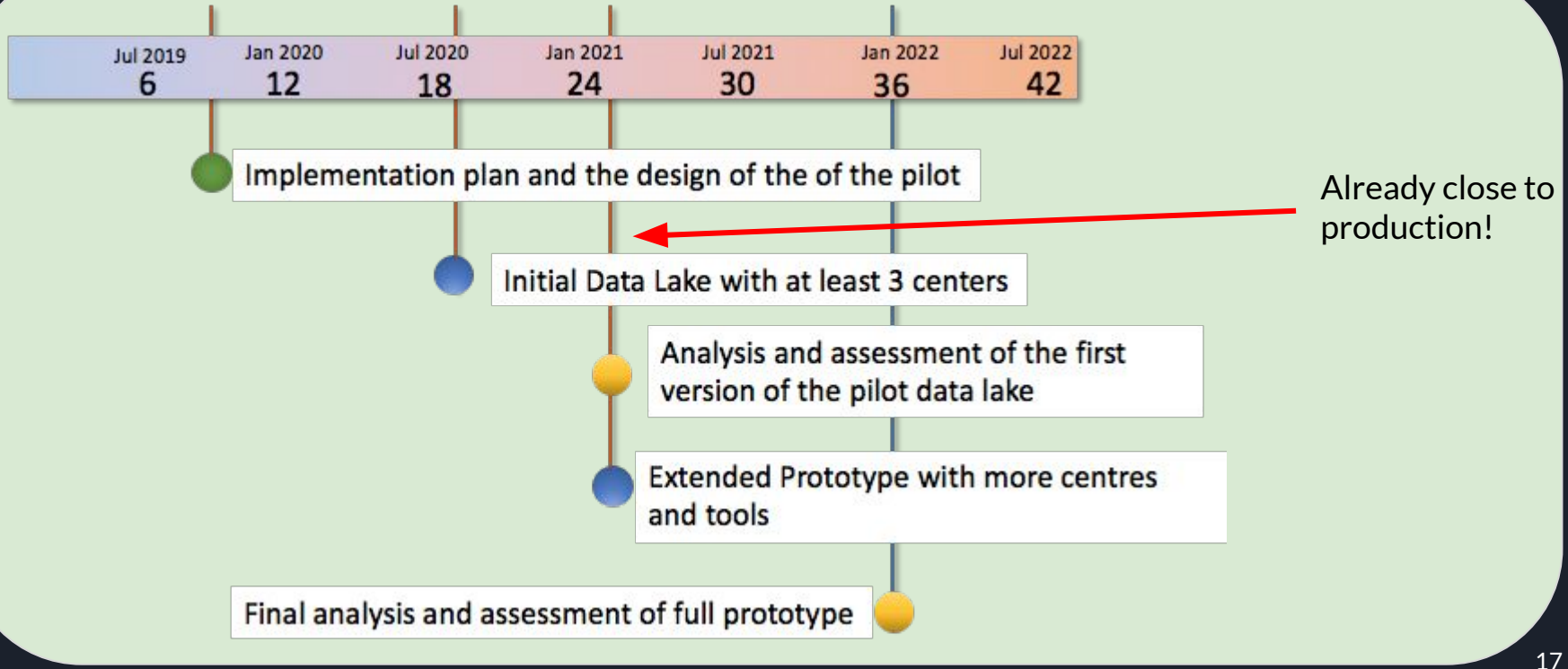




Transfer Error Logs

Source	Destination	payload.reason.keyword: Descending	Count
DESY-DCACHE	SARA-DCACHE	TRANSFER [13] TRANSFER ERROR: Copy failed with mode streamed, with error: Authentication error, reached maximum number of attempts	46
EULAKE-2	SARA-DCACHE	TRANSFER [13] TRANSFER ERROR: Copy failed with mode streamed, with error: Authentication error, reached maximum number of attempts	44
DESY-DCACHE	PIC-DCACHE	TRANSFER [110] DESTINATION OVERWRITE Connection timed out	43

WP2 important milestones



Conclusions

- Currently, the WP2 is operating with a subset of the ESFRIs as primary targets
  - **HL-LHC** and **SKA** are the driving design use cases, simply because their data rate is larger (by far)
- The design is expected to
  - Be endorsed in all the WLCG main sites (including Lyon and CNAF, serving also Virgo)
  - Be, if anything, over abundant for the Virgo use cases
- Not covered here, but OSG is following the same path with the DOMA project
  - All discussions are in common, and solutions are going to be by design interoperable when not identical -- implications on LIGO?
- **“We can speak”**