DOMA status

LHCC Computing Referees meeting

S. Campana (CERN)

DOMA in a nutshell

DOMA project
(Data Organization, Management, Access)

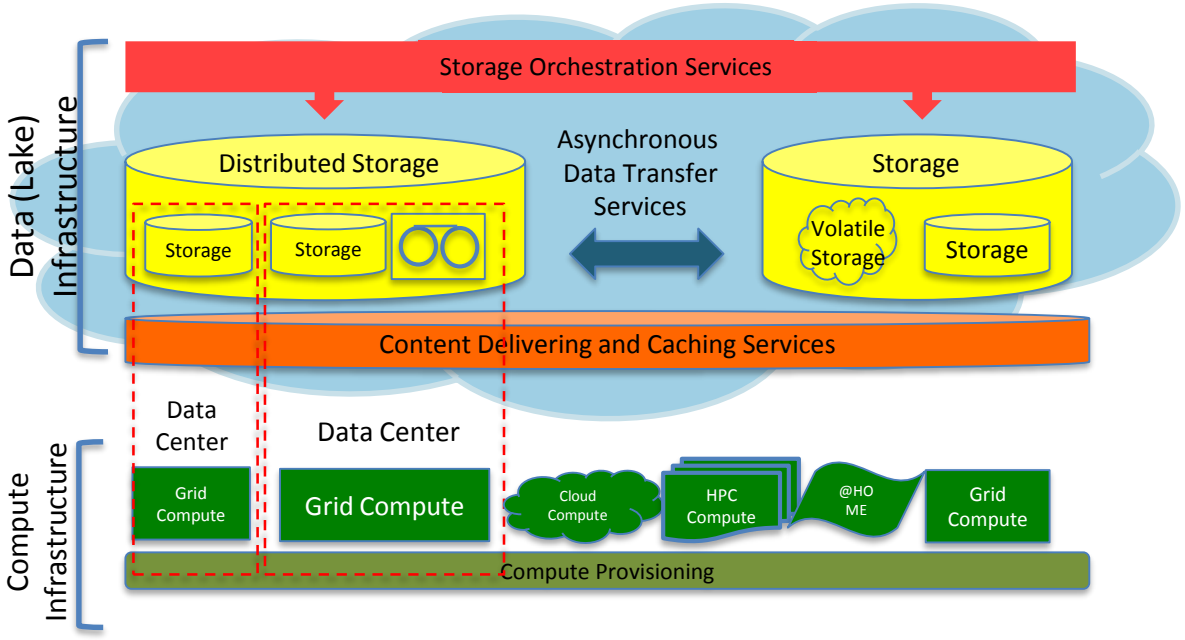
https://twiki.cern.ch/twiki/bin/view/LCG/DomaActivities

A set of R&D activities evaluating components and techniques to build a common HEP data cloud

Three Working Groups

- ACCESS for Content Delivery and Caching
- TPC for Third Party Copy
- QoS for storage Quality of Service

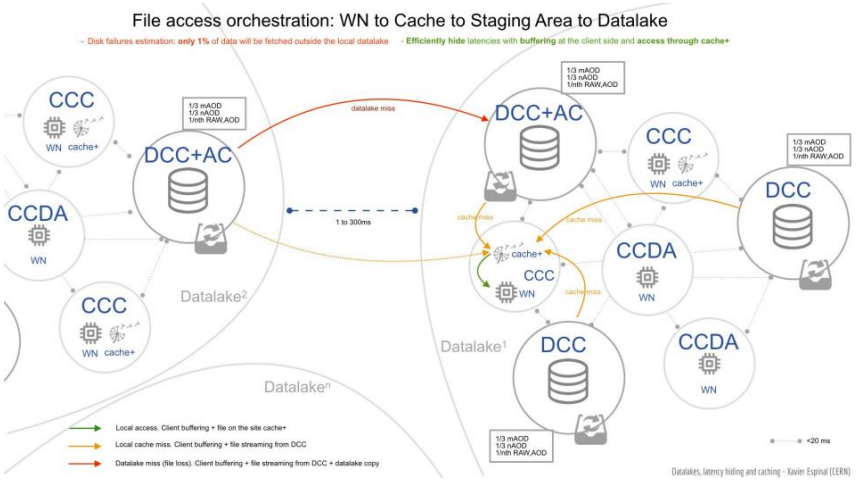
And many activities, reporting regularly



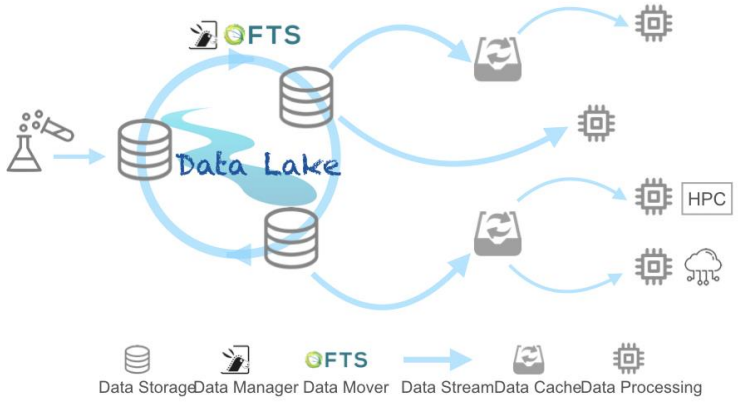
ACCESS

At the last LHCC …

- Data Access, Content Delivery, Caching, Latency Hiding: bridging CPUs and data
  - A huge amount of performance measurements, technologies evaluation, workflows studies
  - A strawman model proposed for the data analysis use case
  - Xcache emerges as the most promising caching and latency hiding technology
- Solutions are being prototyped by sites and experiment.



Detailed and Simplified Strawman Model



ACCESS: xCache performance

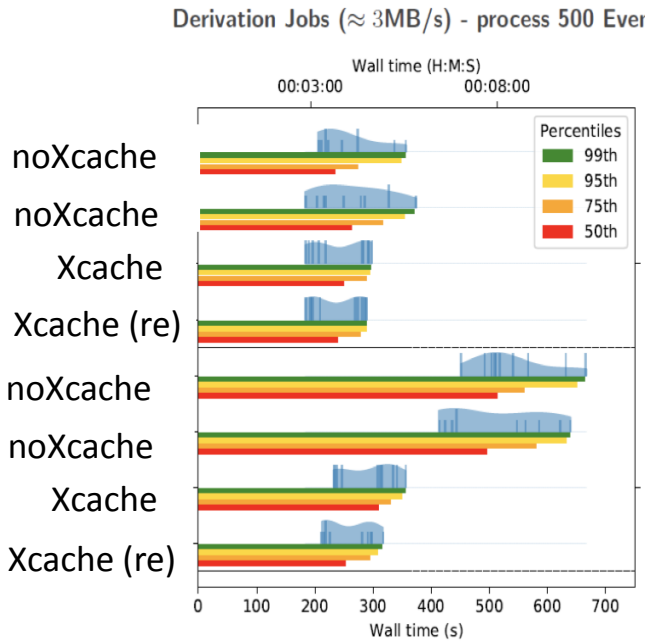
ATLAS Derivation jobs. Metric: WallTime / 500 Evts

Compares direct read from storage (directIO) with read through xCache in Munich

Conclusion:

- xCache hides latency for high RTT. Data access seen as “quasi-local”
- Further benefit in case of re-use (caching)

Processing Nodes in Munich

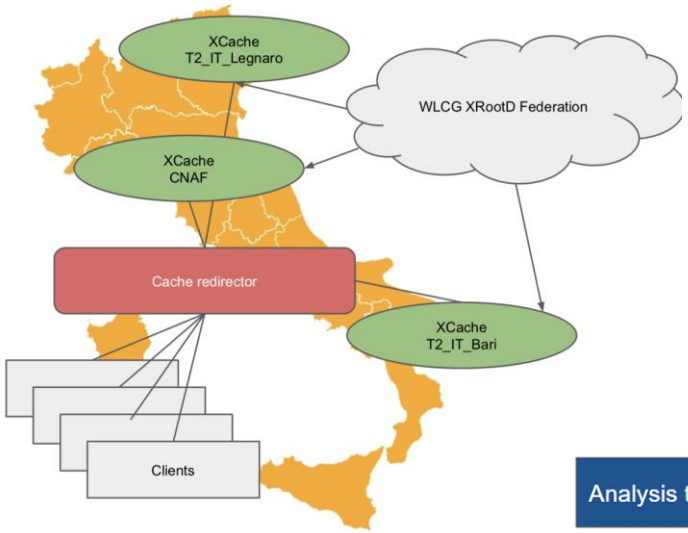


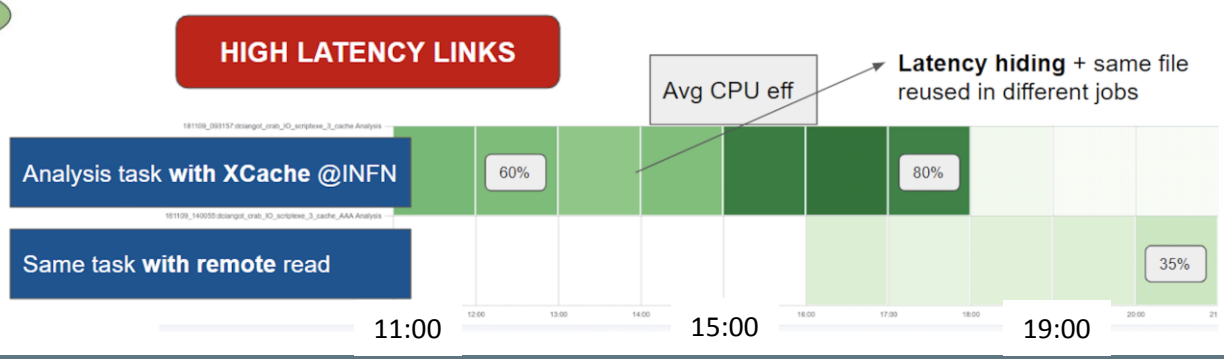
data from DESY

data from Beijing

ACCESS: caching layer prototype

A distributed caching system in INFN



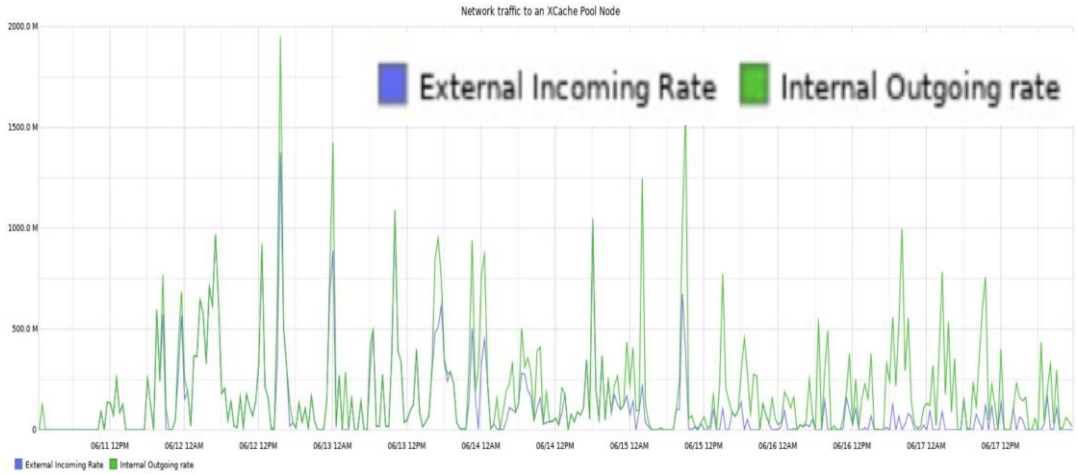


ACCESS: caching in production

Birmingham has no pledged storage any longer. Data source for the BH Worker Nodes is Manchester.

Simple direct read was overloading Manchester SE. Deployed xCache in Birmingham

xCache network traffic



Conclusion:

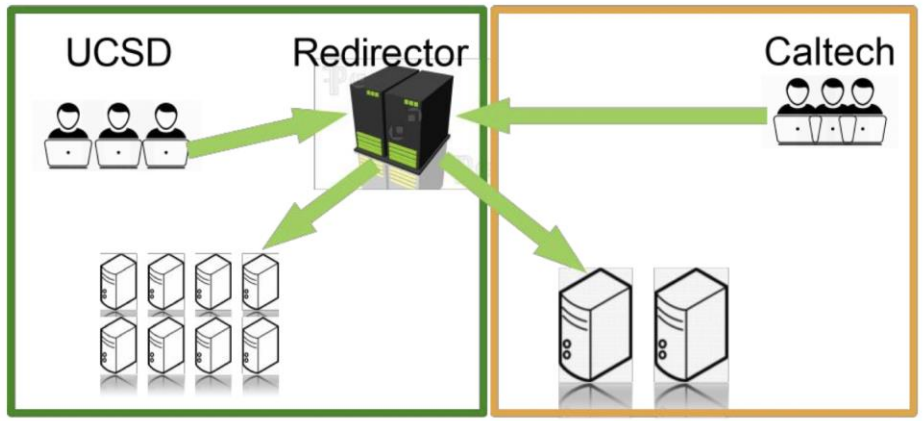
- Caching works as expected
- Files reused ~3 times
- Significant saving in network traffic

ACCESS: SoCal

Courtesy David Lange
Present Model of CMS
HL-LHC resource planning

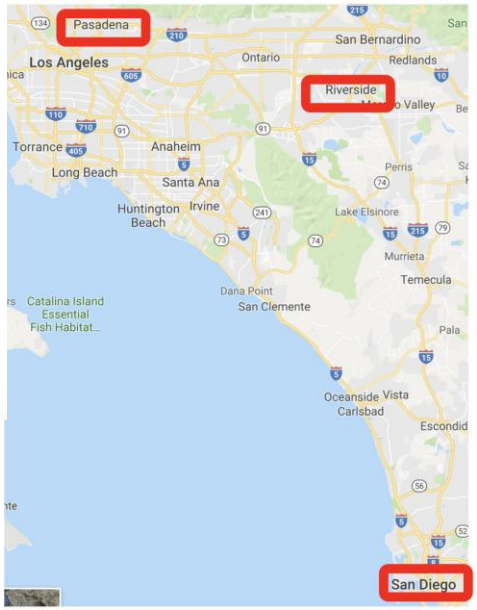
Data Tier	Data
RAW [MB]	7.4
AOD [MB]	2.0
MiniAOD [kB]	200
NanoAOD [kB]	4

SoCal XRootD Cache



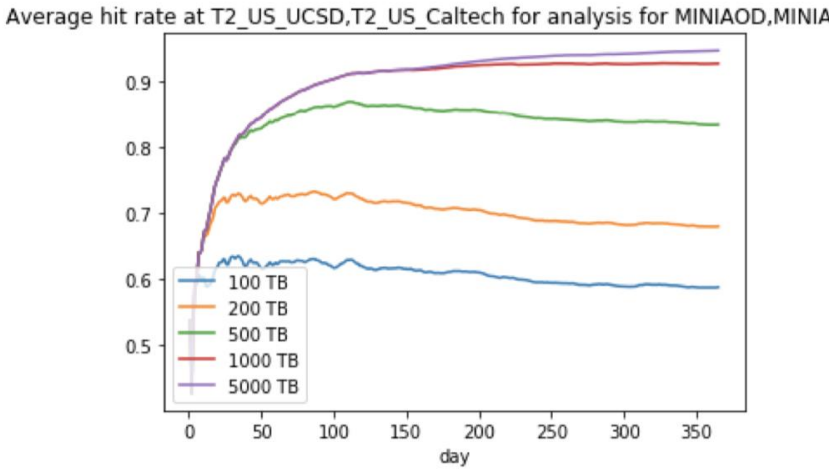
	UCSD	Caltech
Nodes	11 (10 more coming)	2
Disk Capacity per node	12x2TB = 24TB	30x6TB (HGST Ultrastar 7K600)
Network Card per node	10 Gbps	40 Gbps
Total Disk Capacity	264 TB	360TB

Last Year



SoCal serves data to CPUs in all South California CMS sites

ACCESS: SoCal



A 1PB cache in SoCal filled with Mini/Nano AODS has 90% hit rate

From Frank Wuerthwein:

- Want CMS to switch to Buffer & Cache mode.
  - Buffer that assumes nothing in buffer needs to stay there for longer than a week, to keep buffer small.
- Want to operate only JBODs
- Want CMS to be responsible for dealing with data losses due to disk losses.

Overall, want to decrease total cost of ownership.

From Frank Wuerthwein:

UCSD

Summary & Conclusions

CMS

- I clearly see advantages for my T2 operations from the Data Lake straw proposal presented at DOMA session at HOW.
  - Less operational burden
  - Less money spent on disk that is rarely accessed.
- We can get started immediately with existing Xcache software.
  - Slowly increase the money spend on CPU vs disk, thus reversing the opposite trend.
- There is a lot that can be improved going forward.
  - Smarter treatment of cache misses.
  - Smarter placement of jobs given knowledge of cache content.
  - Better production workflows such that data spends less time in T2 buffers.

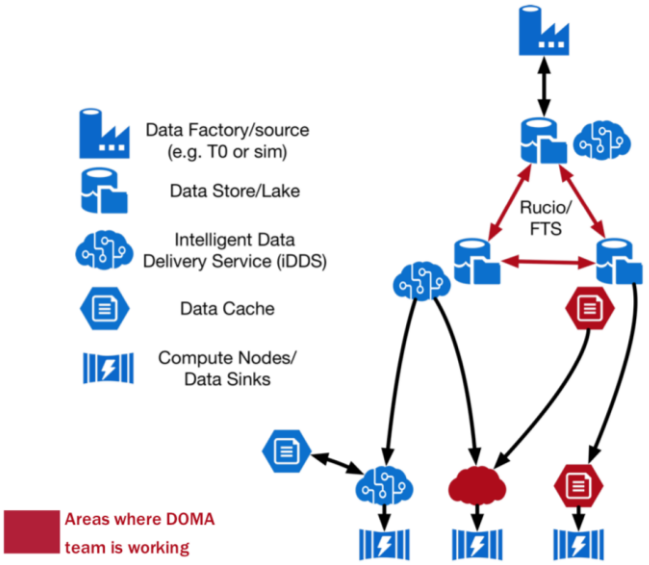
Use Run3 to put the improvements in place over time!

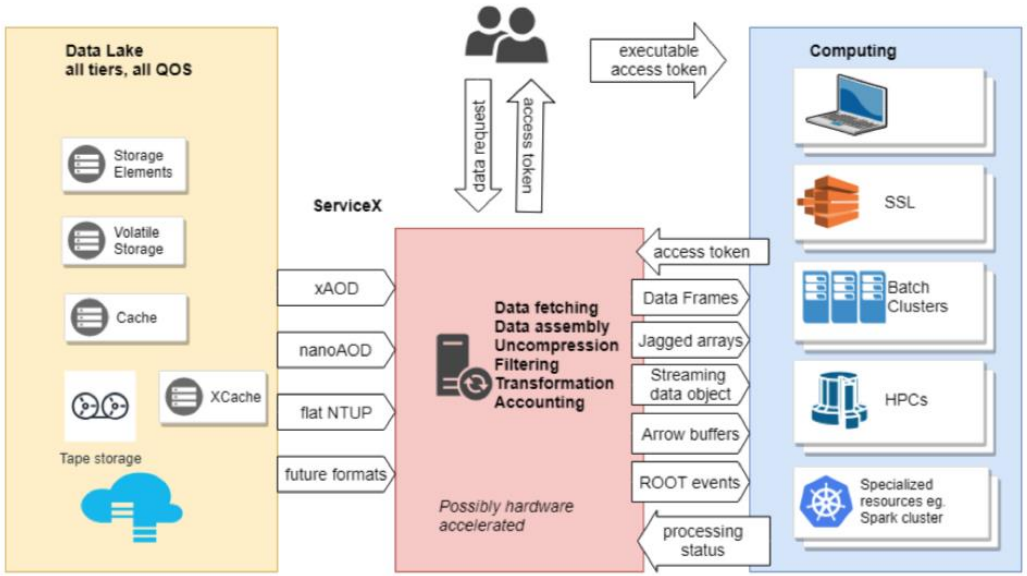
14

IDSS R&D

Intelligent Data Delivery Service: a R&D project being developed in the context of IRIS-HEP

An active layer that “prepares” the data for you





The SLATE spin-off

Welcome to SLATE CI Portal

Services Layer at the Edge and the Mobility of Capability

The primary goal of SLATE is to accelerate collaborative science. SLATE augments the canonical **Science DMZ pattern** with a secure container orchestration platform and federated trust model.

The platform permits hosting of containerized services needed for higher-level capabilities such as **data transfer nodes**, software and data caches, workflow services and science gateway components.

SLATE is using Kubernetes as an underlying technology to implement these capabilities.

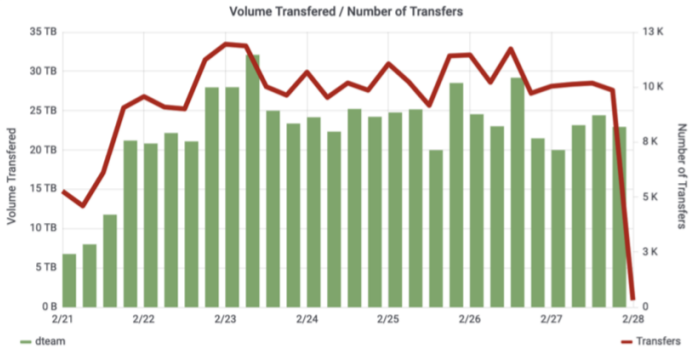
SLATE is being looked at in the context of xCache deployment
A WLCG discussion started on security and policies

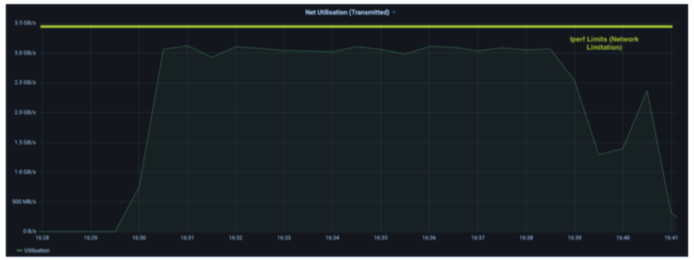
TPC

Goal: commission non-gridFTP protocols for asynchronous data transfer (Third Party Copy)

- Phase-2 (deadline June 2019): all sites providing > 3PB of storage to WLCG should provide a non gridFTP endpoint in production

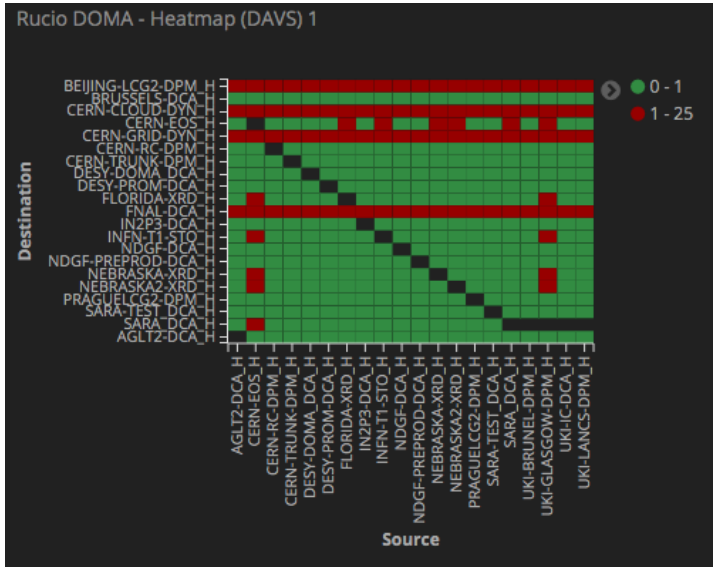
Functional and Stress testing



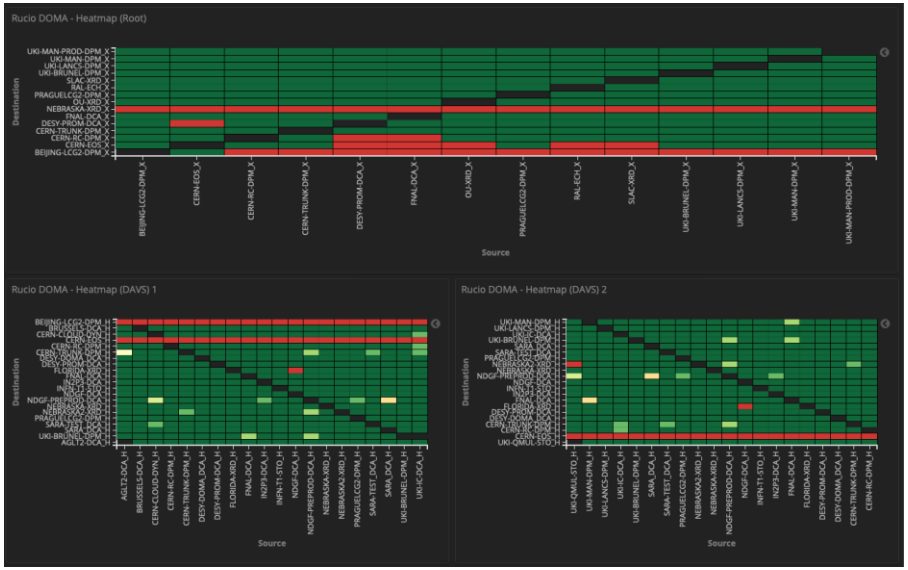


Capable to fill available bandwidth

Functional Tests in June



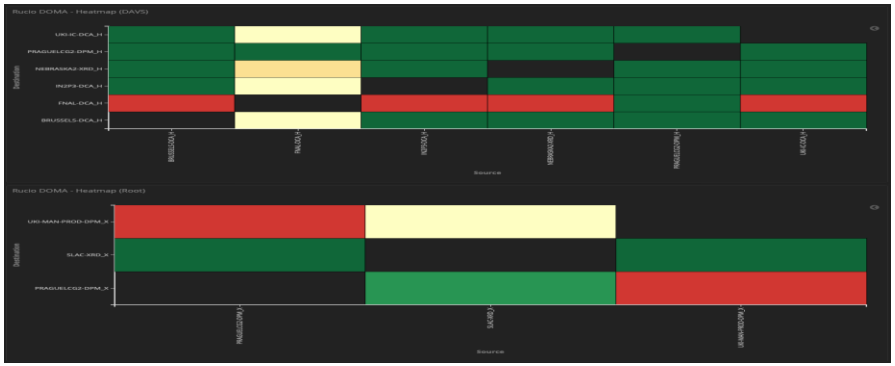
Functional tests today



Now setting targets for non gridFTP production traffic. Already tested in limited time windows

Spin-off: progressing toward a SRM-less world

Stress tests today



TPC next phase

- Phase-3 (deadline Dec 2019): all sites to have a non-gridFTP endpoint

Gave an opportunity to review WLCG storage deployment: some features needed for TPC are available only in decently recent versions of storage

Many sites conservatively did not upgrade storage in Run-2 to favor stability. Which is fine.

Upgrade campaign now ongoing and will take several months, as experiments are all but idle.

- Probably need to shift the Phase-3 target date by a few months

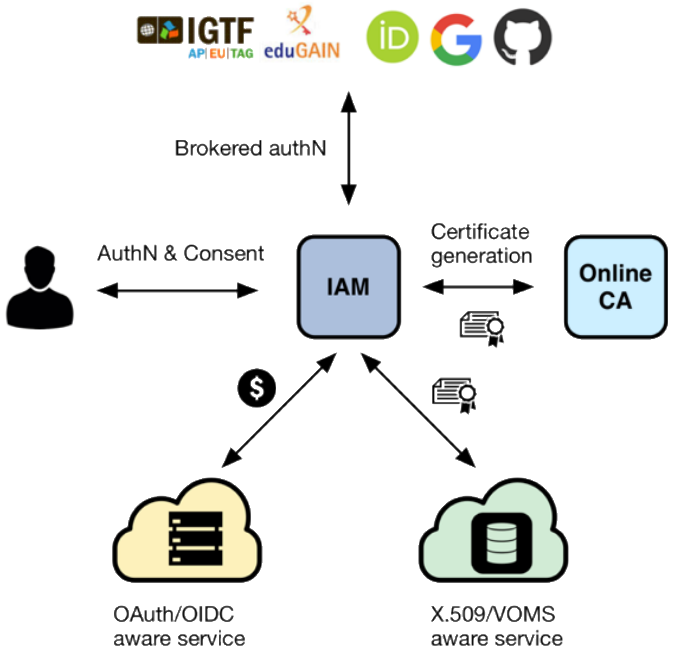
TPC and AAI

WLCG is planning to evolve AAI toward token based Auth/AuthZ and Federated Identities

The WLCG task force is finalizing the token profile as last item

While this is has a much broader scope than DOMA, TPC offers a well confined use case to start with

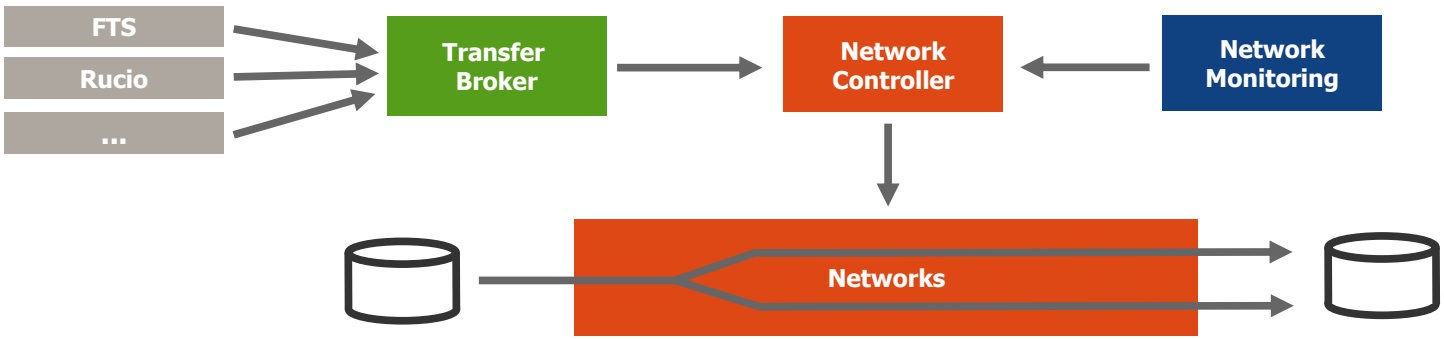
Rucio is integrating tokens. Storage is preparing to manage them.



NOTED: shaping networks

Implement a **Transfer broker:**
- Identify upcoming and on-going substantial data transfers
- get information from transfer services (FTS, Rucio …)
- map transfers to network endpoints
- make transfers info available to network providers

Demonstrate a **Network Controller:**
- takes input from Transfer Broker
- modify network behavior to increase transfer efficiency
- take into account real-time network status information



ATLAS Data Carousel

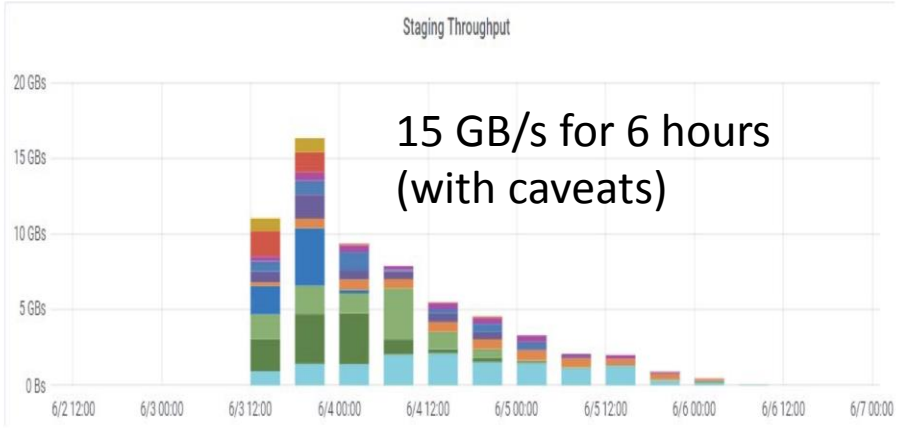
End-to-End data processing from high latency media (TAPE). Not simply a tape staging throughput test.

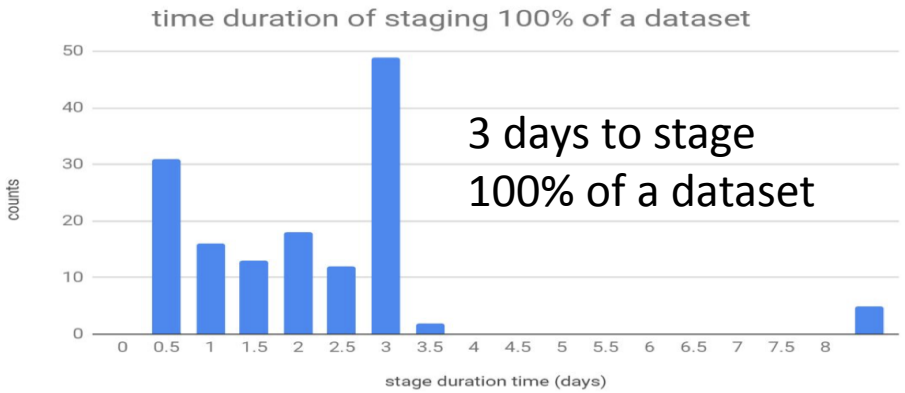
Focus on AOD->DAOD (derivations). Goal is to reduce AOD footprint on disk

Staging throughput looks very promising. Achieving desired End2End performance requires more work

More tests in summer 2019. Results being analyzed

Tests from June 2019





Tape-less archive storage

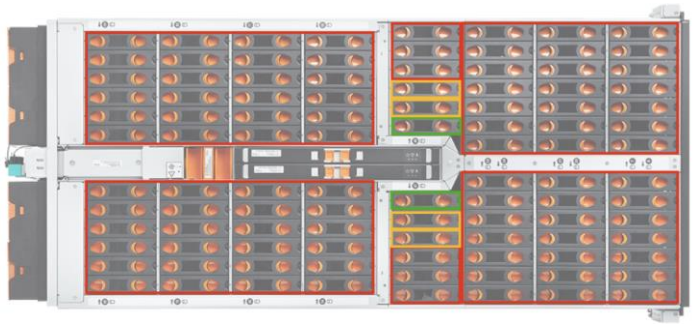
R&D launched by KISTI and Alice one year ago. Reduce cost of operating an archive storage and avoid vendor lock-in. Discussed and approved in WLCG Overview Board

Based on EOS technology

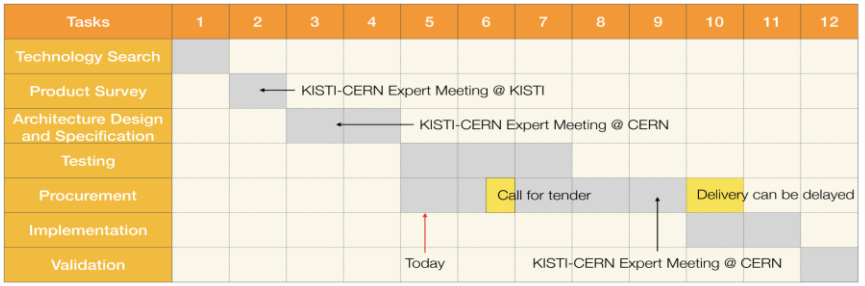
- State-of-the-art JBOD technology: high density disk loads (up to 102), SAS 12Gb/s transfer speed, SAS dual-port disks
- 2-D data protection: erasure coding implemented in EOS (RAIN) and ZFS (RAID-Z3)
- Tunable QoS classes: Usable capacity vs. Data protection

102 Disks JBOD

x4 ZFS Group - RAID-Z3
21 Data disks + 3 Parity Disks + 1 Hotspare
+2 Global HS



Schedule (2019)



Final Messages

The DOMA activities focus on two aspects:

- Filling the gap between needed and expected resources in HL-LHC
- Modernization of infrastructure, tools and services for long term sustainability

Very good collaboration between experiments, service developers and service

Preparing for HL-LHC is an adiabatic process. New technologies and workflows are maturing to production. Run-3 will offer the opportunity to commission them at increasing scale