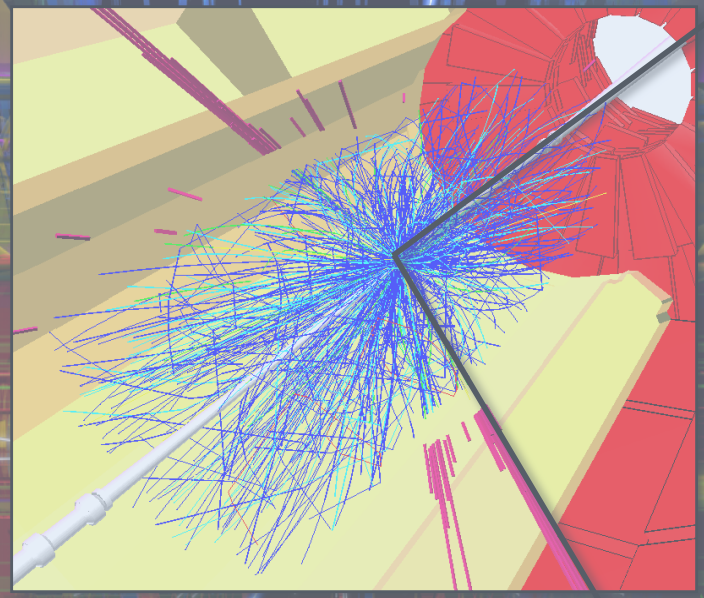
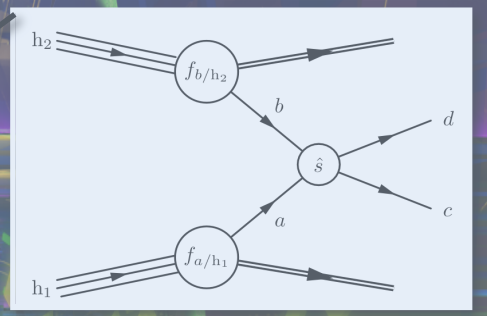
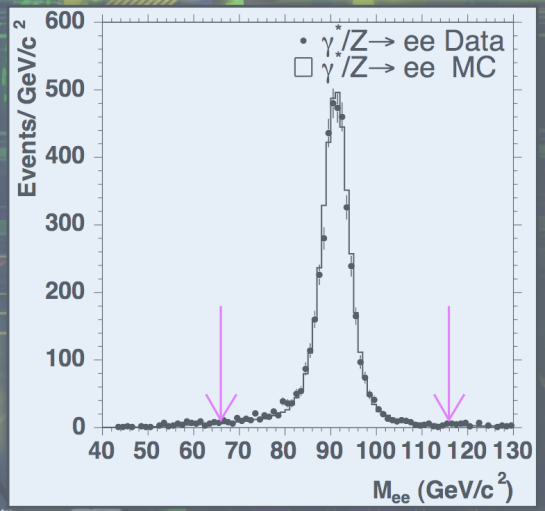
Информационные технологии в физике элементарных частиц







Национальный Исследовательский Ядерный Университет «МИФИ»

март, 2020

Алексей Климентов

Лекция III

- Лекция II
  - Суперкомпьютеры
  - Интеграция суперкомпьютеров и грид
  - «другие» (не Intel x86) архитектуры
  - *Computing model evolution*

Evolution of HEP computing

HSF-CWP-2017-01
December 15, 2017

arXiv:1712.06982v5 [physics.comp-ph] 19 Dec 2018

A Roadmap for HEP Software and Computing R&D for the 2020s

HEP Software Foundation[1]

Abstract: Particle physics has an ambitious and broad experimental programme for the coming decades. This programme requires large investments in detector hardware, either to build new facilities and experiments, or to upgrade existing ones. Similarly, it requires commensurate investment in the R&D of software to acquire, manage, process, and analyse the shear amounts of data to be recorded. In planning for the HL-LHC in particular, it is critical that all of the collaborating stakeholders agree on the software goals and priorities, and that the efforts complement each other. In this spirit, this white paper describes the R&D activities required to prepare for this software upgrade.

[1]Authors are listed at the end of this report.

https://doi.org/10.1007/s41781-018-0018-8

WLCG-LHCC-2018-001
05 April 2018

WLCG Strategy towards HL-LHC

Executive Summary

The goal of this document is to set out the path towards computing for HL-LHC in 2026/7. Initial estimates of the data volumes and computing requirements show that this will be a major step up from the current needs, even those anticipated at the end of Run 3. There is a strong desire to maximise the physics possibilities with HL-LHC, while at the same time maintaining a realistic and affordable budget envelope. The past 15 years of WLCG operation, from initial prototyping through to the significant requirements of Run 2, show that the community is very capable of building an adaptable and performant service, building on and integrating national and international structures. The WLCG and its stakeholders have continually delivered to the needs of the LHC during that time, such that computing has not been a limiting factor. However, in the HL-LHC era that could be very different unless there are some significant changes that will help to moderate computing and storage needs, while maintaining physics goals. The aim of this document is to point out where we see the main opportunities for improvement and the work that will be necessary to achieve them.

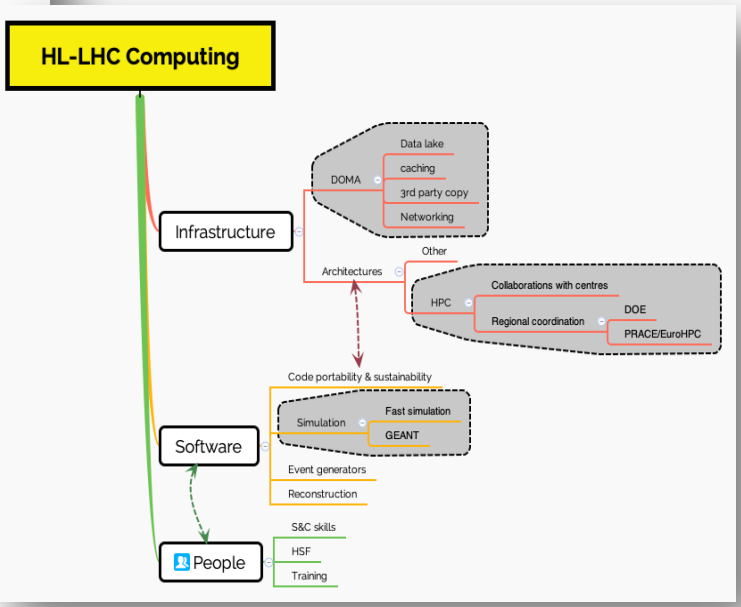
During 2017, the global HEP community has produced a white paper - the Community White Paper (CWP), under the aegis of the HEP Software Foundation (HSF). The CWP is a ground-up gathering of input from the HEP community on opportunities for improving computing models, computing and storage infrastructures, software, and technologies. It covers the entire spectrum of activities that are part of HEP computing. While not specific to LHC, the WLCG gave a charge to the CWP activity to address the needs for HL-LHC along the lines noted above. The CWP is a compendium of ideas that can help to address the concerns for HL-LHC, but by construction the directions set out are not all mutually consistent, not are they prioritised. That is the role of the present document - to prioritise a program of work from the WLCG point of view, with a focus on HL-LHC, building on all of the background work provided in the CWP, and the experience of the past.

At a high level there are a few areas that clearly must be addressed, that we believe will improve the performance and cost effectiveness of the WLCG and experiments:

- **Software:** With today's code the performance is often very far from what modern CPUs can deliver. This is due to a number of factors, ranging from the construction of the code, not being able to use vector or other hardware units, layout of data in memory, and end-end I/O performance. With some level of code re-engineering, it might be expected to gain a moderate factor (x2) in overall performance. This activity was the driver behind setting up the HSF, and remains one of the highest priority activities. It also requires the appropriate support and tools, for example to satisfy the need to fully automate the ability to often perform physics validation of software. This is essential as we must be adaptable to many hardware types and frequent changes and optimisations to make the best use of opportunities. It also requires that the community develops a level of understanding of how to best write code for performance, again a function of the HSF.

1

https://cds.cern.ch/record/2621698



Дальнейшее развитие компьютерной модели

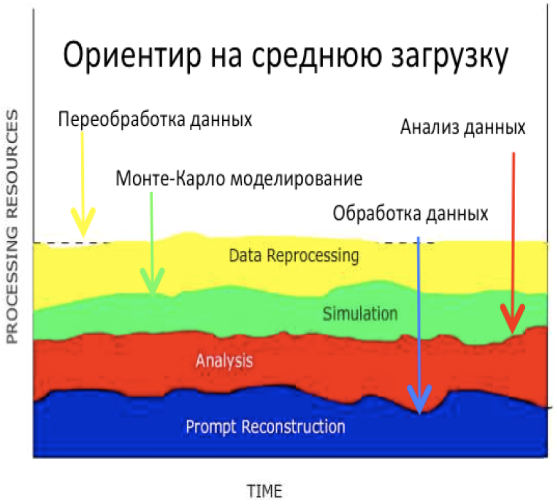
Фундаментальным вопросом для развития компьютерной модели в области физики элементарных частиц является вопрос : «как новые данные будут обрабатываться, анализироваться и моделироваться через 7-10 лет ?».

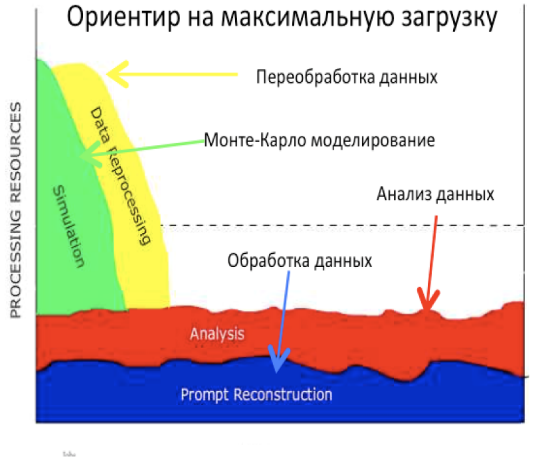
До последнего времени модель строилась в предположении, что эксперименты являются “собственниками” вычислительного ресурса.

Вариантами решения могут быть :

1. Эксперименты ФВЭ и ЯФ будут продолжать покупать необходимое аппаратное обеспечение и расширять свою компьютерную инфраструктуру;
   - Очевидное преимущество - это преимущество “собственника” ресурса, ресурс м.б. использован и доступен в любой момент;
   - Это преимущество надо учитывать только в случае, если есть достаточный ресурс в момент максимальной загрузки, в остальное время вычислительный ресурс не будет использован в полном объеме;
2. Эксперименты ФВЭ и ЯФ будут покупать мощности у тех, кто их предоставляет на коммерческой основе.
   - Преимущество такого подхода состоит в том, что капитальные затраты несет третья сторона;
   - Недостатком является отсутствие гарантии, что ресурс будет доступен для использования, когда это потребуется; А также необходимость “доверия” к третьей стороне и предоставления ей доступа к данным международной коллаборации.
3. Компромиссным является вариант, когда базовые ресурсы принадлежат экспериментам, а в момент максимальной нагрузки также “используются” поставщики вычислительных услуг и сервисов.

Дальнейшее развитие компьютерной модели. Смена парадигмы





- Ландшафт современных вычислительных ресурсов и потребности в них драматически отличаются от ситуации 20 летней давности, когда приложения ФВЭ и ЯФ были одним из основных “потребителей” вычислительных мощностей в глобальном мире ИТ
- В настоящее время существует больший пул вычислительных ресурсов за пределами ФВЭ и ЯФ . В первую очередь это коммерческие ресурсы и суперкомпьютерные центры. Так вычислительный ресурс гигантов ИТ индустрии : Яндекс, Google, Amazon, Microsoft в сотни раз превышает мощности консорциума WLCG, ресурс суперкомпьютера Titan (остановлен в июле 2019 года) превышал весь ресурс WLCG.
  - Это позволяет и требует пересмотра “усредненного” подхода к использованию вычислительных мощностей и смены модели с ориентацией использования максимального вычислительного ресурса на момент пиковой нагрузки и соответствующее планирование потоков заданий. При таком сценарии классы потоков заданий могут быть переориентированы соответственно.

Common challenges

- Management of Exabyte- scale science data
  - And associated tools, networks, infrastructure
- Move from “simple” x86-like clusters to very heterogenous resources
  - Use of HPC and Exascale computing resources
- Infrastructures & centres likely to be common between HEP & Astronomy, Astroparticle, GW, etc.
- Software challenge – associated with the above
  - How to easily move code between various compute resources, validate correctness, adapt to new architectures, etc.
- Develop and retain skills in software and computing
  - In the scientific community – as well as with specialists
  - Issue of recognition in academic environments

ESFRI Science Projects

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

ESCAPE

European Science Cluster of Astronomy & Particle physics
ESFRI research infrastructures

EUROPEAN OPEN SCIENCE CLOUD

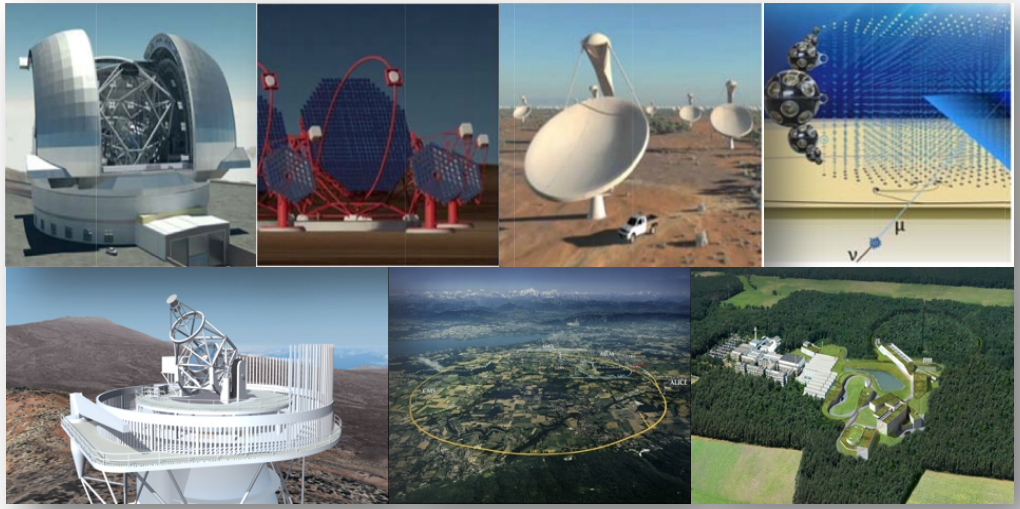
Horizon 2020 funded project

Goals:

Prototype an infrastructure for the EOSC that is adapted to the Exabyte-scale needs of the large ESFRI science projects.

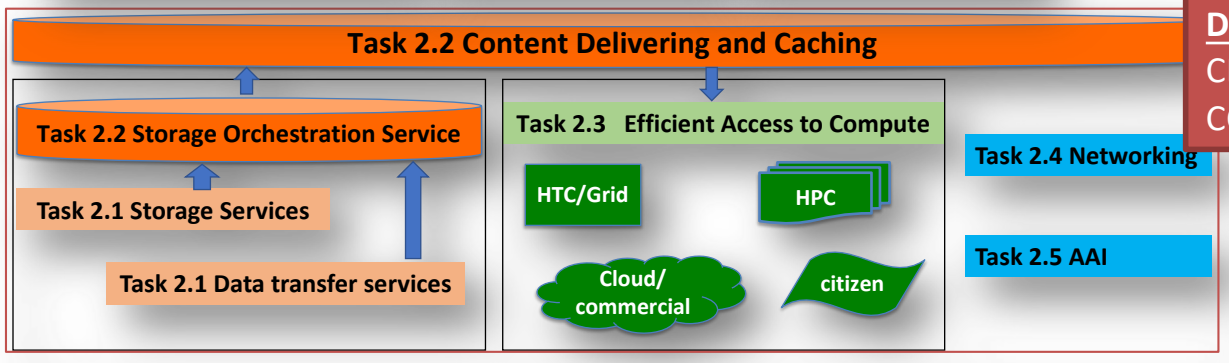
Ensure that the science communities drive the development of the EOSC.

Has to address *FAIR* data management, long term preservation, open access, open science, and contribute to the EOSC catalogue of services.



Work Packages

WP2 – Data Infrastructure for Open Science
WP3 – Open-source scientific Software and Service Repository
WP4 – Connecting ESFRI projects to EOSC through VO framework
WP5 – ESFRI Science Analysis Platform



Data centres (funded in WP2)
CERN, INFN, DESY, GSI, Nikhef, SURFSara, RUG, CCIN2P3, PIC, LAPP, INAF

Data Infrastructure

DOMA project
(Data Organization, Management, Access)

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

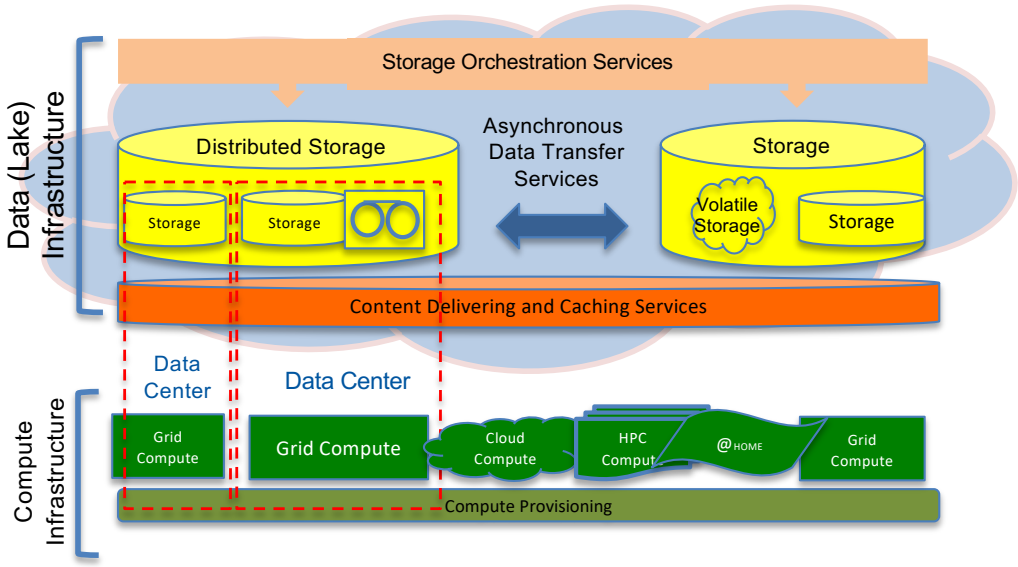
Idea is to localize bulk data in a cloud service (Tier 1's ➔ data lake):
minimize replication, assure availability; policy driven

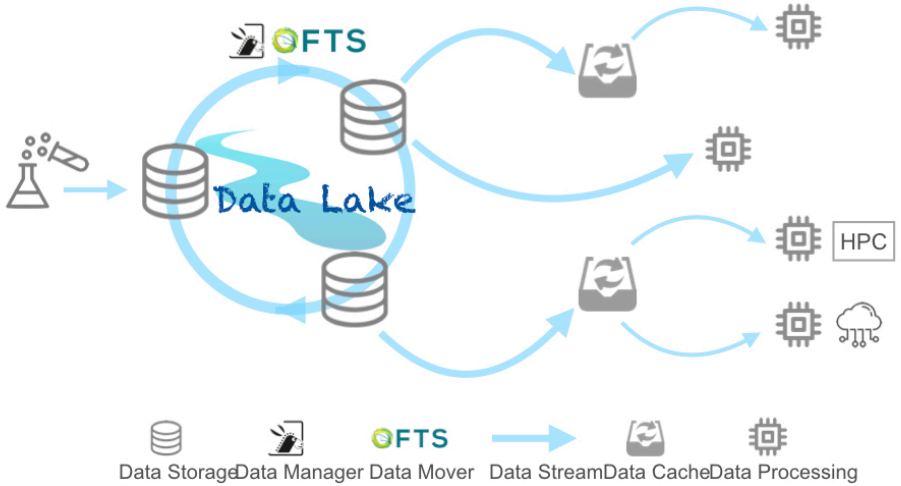
Serve data to remote (or local) compute – grid, cloud, HPC, etc.

- Simple caching is all that is needed at compute site

Works at national, regional, global scales

Model to integrate private and commercial storage – in a "RAID" configuration across sites

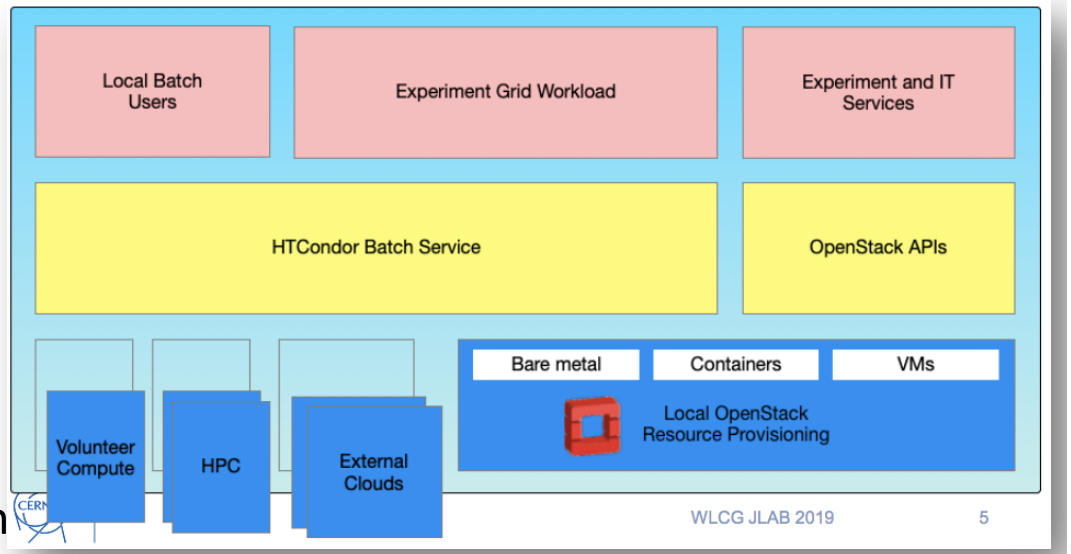




ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures

Heterogenous compute reslources

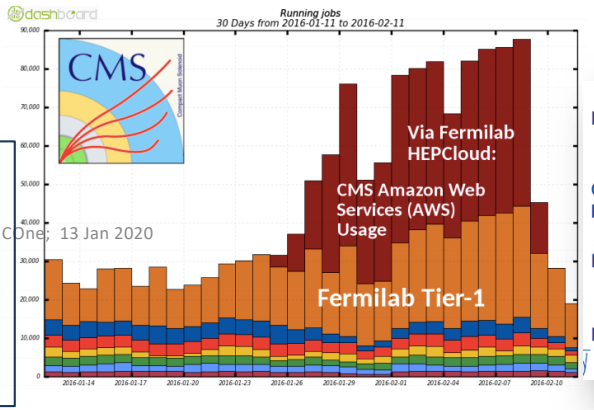
- Requires:
  - Common provisioning mechanisms, transparent to users
  - Facilities able to control access (cost), appropriate use, etc
- HPC, Clouds, HLT will not have (affordable) local storage service (in the way we assume today)
  - Must be able to deliver data to them when they are in active use

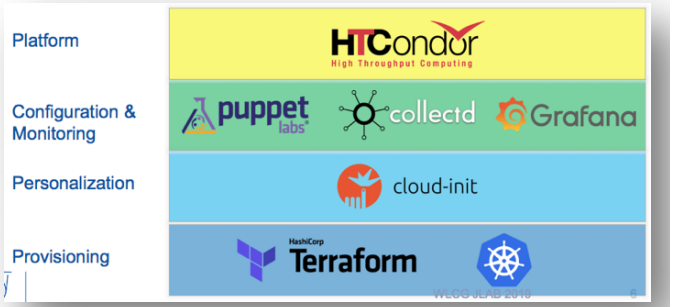


HELIX NEBULA THE SCIENCE CLOUD

Deployed in a hybrid cloud mode:

- Procurers' data centres
- commercial cloud service providers
- GEANT network and EduGAIN Federated Identity Management

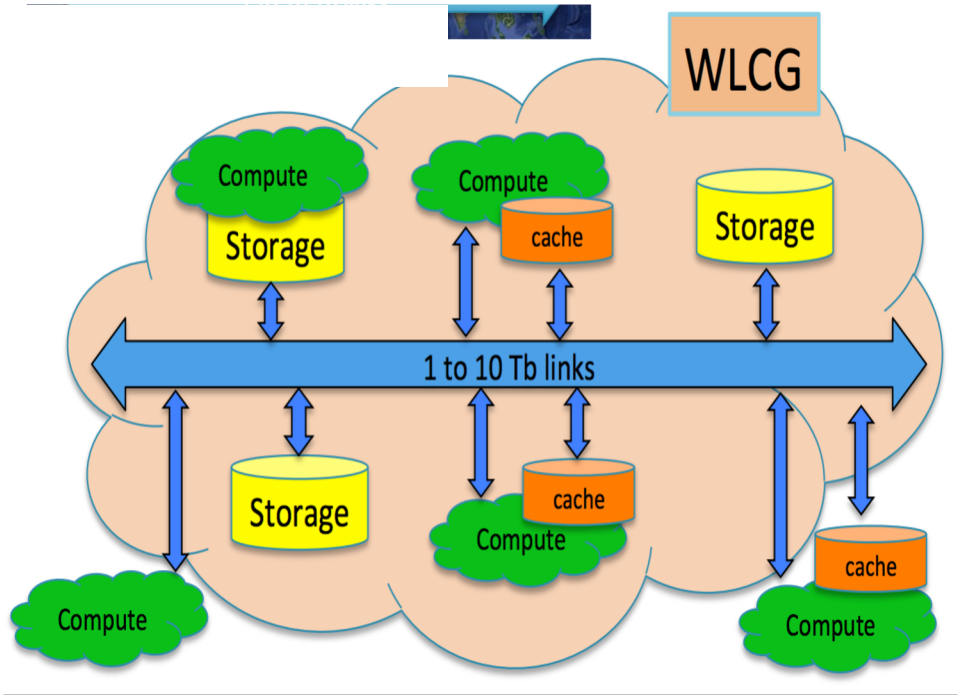




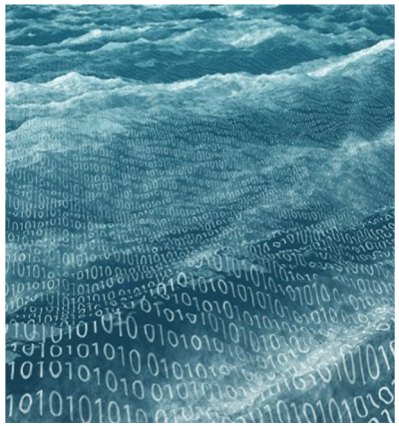
Infrastructure challenges

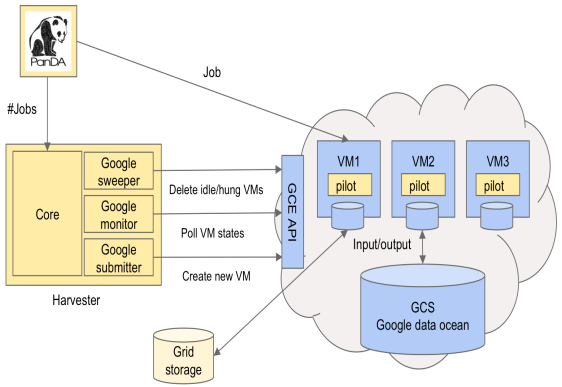
- A federated data infrastructure that:
  - Enables policy driven wide area data replication across a “virtual data centre”
    - == “Data Cloud” or “Data Lake”
    - Want it to appear as a single data repository although distributed
    - Avoid having small *managed* storage service everywhere
  - Is able to feed data to heterogenous compute resources distributed at processing centres
    - Traditional grid/HTC; HPC, Commercial cloud, citizen scientists
    - Streaming, latency hiding, caching, etc.
  - Can integrate owned and commercial resources
- Hopefully a lot in common between HEP and other related sciences with similar needs
- Avoid adding complexity to the system –
  - today it is much simpler than original design; this has decreased the operational cost significantly

Дальнейшее развитие компьтерной модели («озеро/океан данных»)



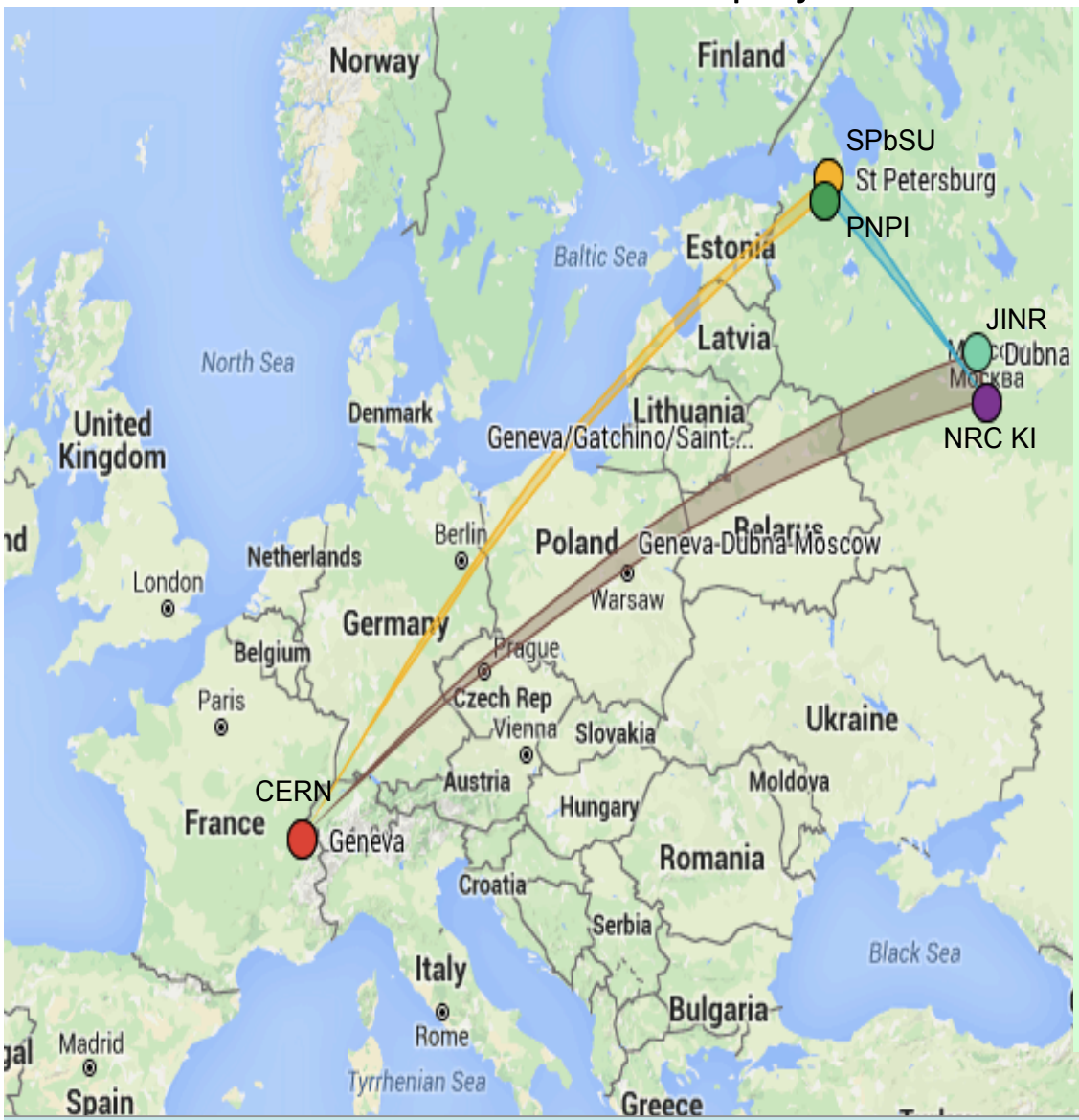
Gb – гигабит
Tb - терабит





Пример интеграции системы
для обработки данных
для прототипа «океан данных»

Federated data and “data lake” R&D projects in HEP



R&D Project Motivation

Computing models for the Run3 and HL-LHC era anticipate a growth of storage needs.

The reliable operation of large scale data facilities need a clear economy of scale.

A distributed heterogeneous system of independent storage systems is difficult to be used efficiently by user communities and couples the application level software stacks with the provisioning technology at sites.

- Federating the data centers provides a logical homogeneous and consistent reliable resource for the end users

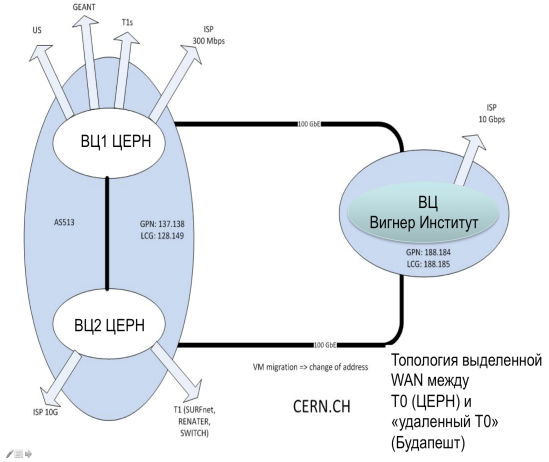
Small institutions have no enough people to support fully-fledged software stack.

- In our project we try to analyze how to set up distributed storage in one region and how it can be used from Grid sites, from HPC, academic and commercial clouds,

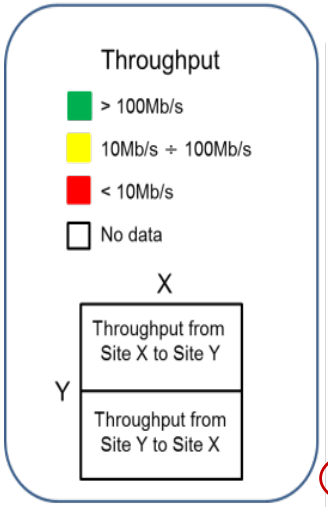
Tested : T2 (ATLAS, PNPI, Gatchina), T2 (ALICE, SPbSU, Petergof), CERN
Ready for tests : T2 (ATLAS, PNPI, Gatchina), T2 (ALICE, SPbSU, Petergof), T1 (NRC-KI, Moscow)
Planned : T0 (JINR, Dubna, NICA), T0/T1 (NRC KI, Moscow, FAIR), T0 (CERN, LHC), SINP
A.Kirianov, A.Klimentov, D.Krasnopevstev, A.Kryukov, N.Kutovskij, A.Petrosyan, E.Ryabinkin

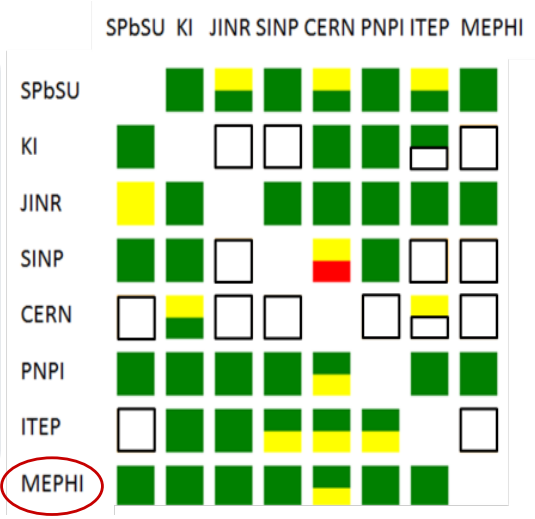
2016/18 – the first prototype

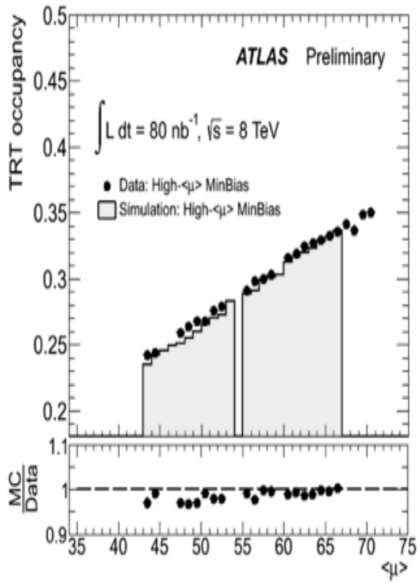
Дальнейшее развитие компьютерной модели. Создание федеративного дискового пространства в рамках гетерогенной киберинфраструктуры

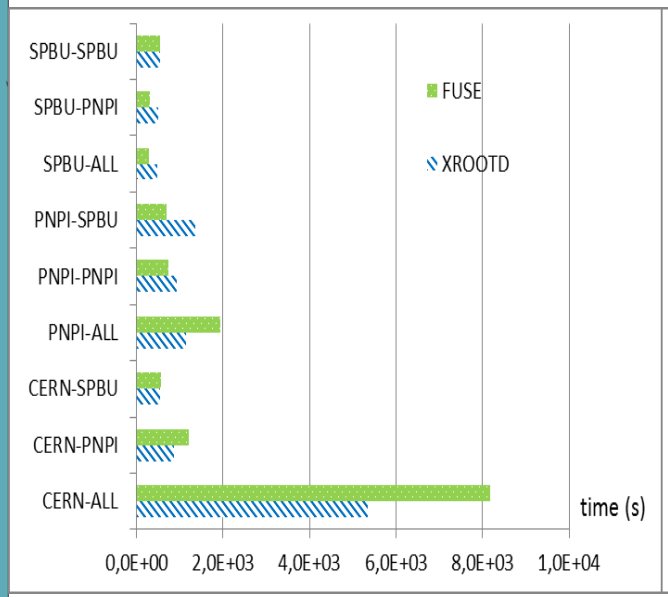


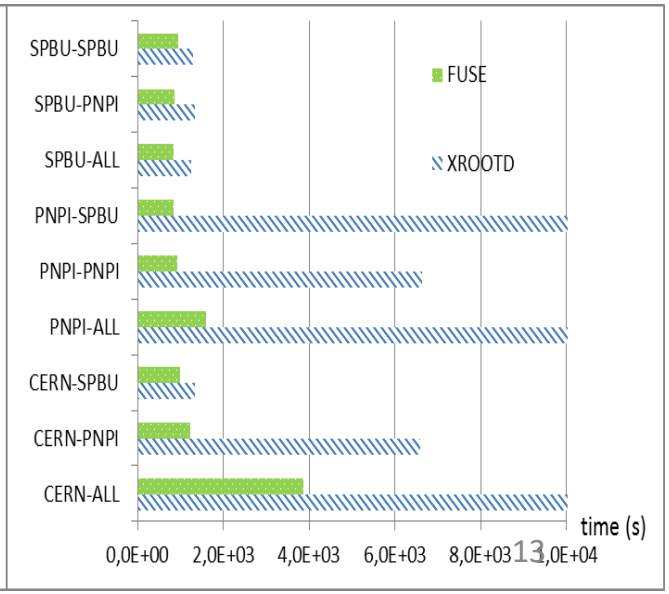
Топология выделенной WAN между T0 (ЦЕРН) и «удаленный T0» (Будапешт)





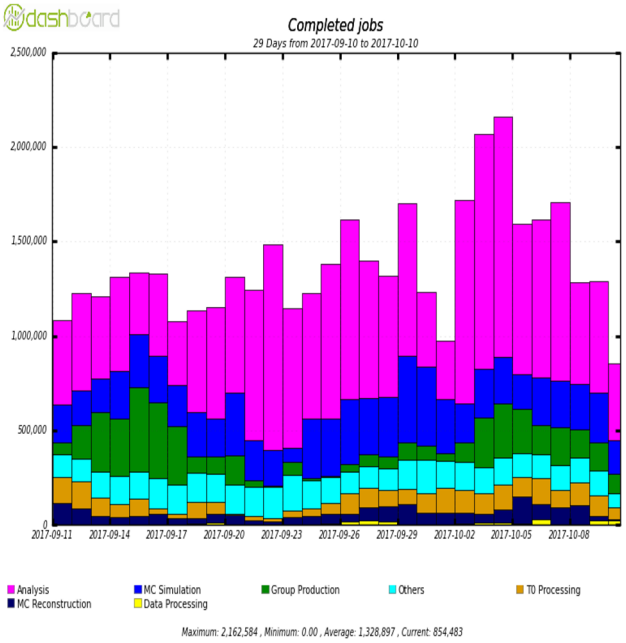




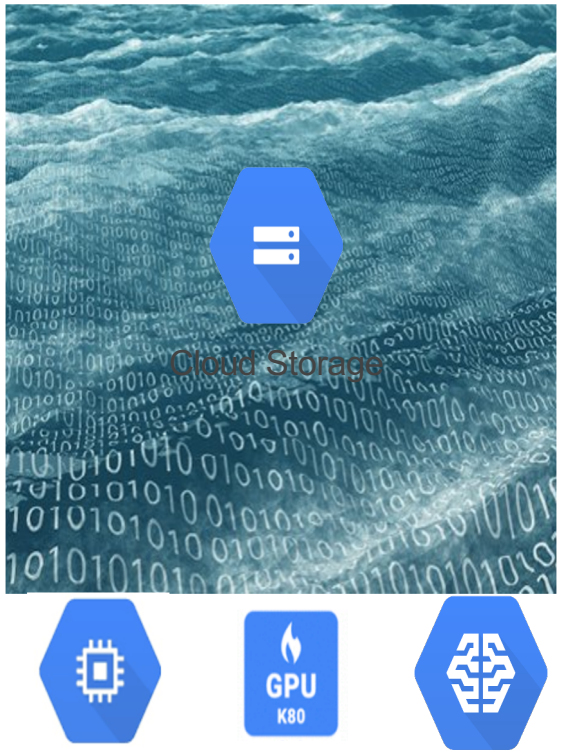


HENP-Google. Three ideas

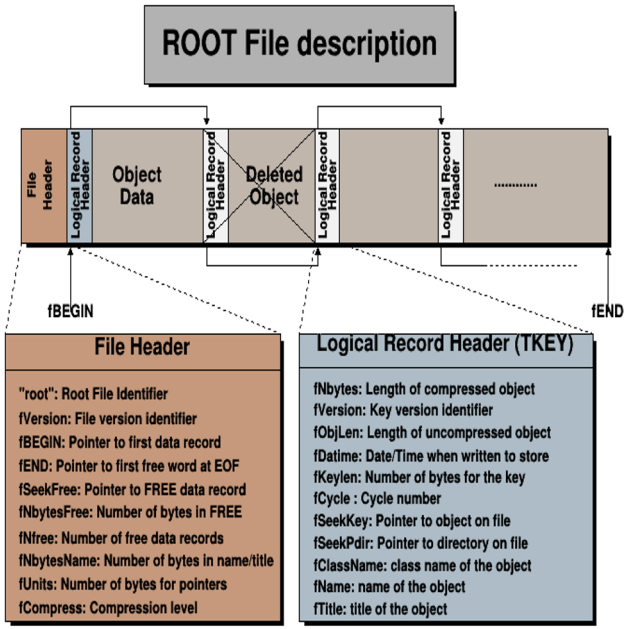
User Analysis



Data Analysis, Replication and Placement



Data Streaming



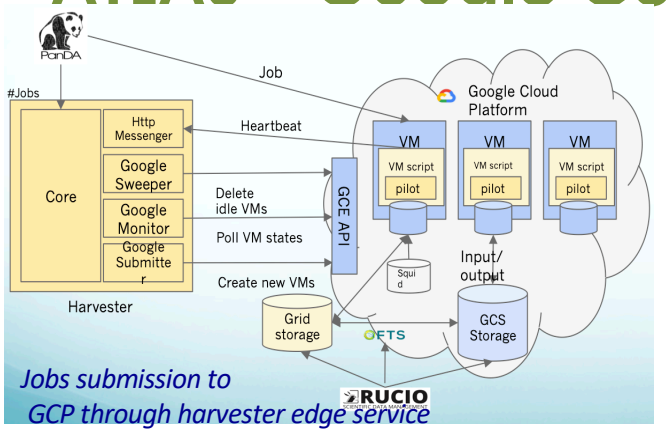
Google Cloud

Confidential & Proprietary

R&D Project Motivation

- IT landscape has changed dramatically since end of XX century. At the end of 90s HEP was a major computer user, and at late 90s
  - Google name was not registered until Sep 1998
  - Amazon had been selling books online
- Today HEP is not the main IT customer, most of innovations are driven by society requests (including social networks)
- Today commercial technology sector is recognized as world IT leaders
  - Amazon, Google, Microsoft, Oracle,… - already play significant role in worldwide scientific computing. Companies are investing in many scientific projects (LSST, MD, genomics)
- LHC data intensive computing challenges are (and have been) at the cutting edge of technology development
- Foster partnerships with IT industries in research and development – and not just as late stage product adopters
- The huge challenges at the HL-LHC have spurred new efforts in ATLAS to collaborate with technology partners
- We proposed to start a new front in LHC R&D, with companies willing to invest in open source solutions

ATLAS – Gooale Compressed Story



Jobs submission to GCP through harvester edge service

white paper v3.0, Mar 2019

US-ATLAS Collaboration with Google for High Energy Physics Applications in the HL-LHC Era

J. Elmsheuser, A. Klimentov, T. Wenaus
Brookhaven National Laboratory

P. Calafiura
Lawrence Berkeley National Laboratory

D. Malon, P. Van Gemmeren
Argonne National Laboratory

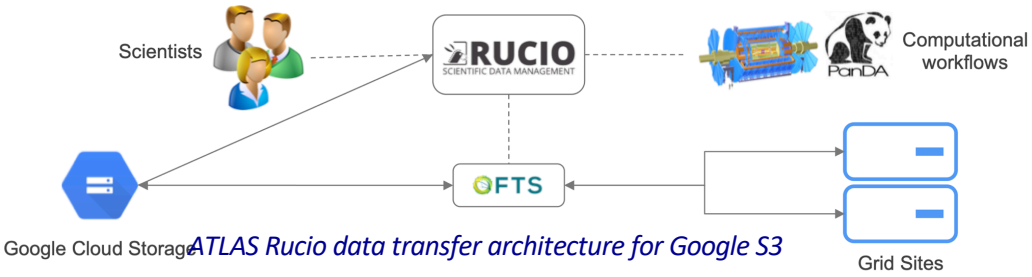
A. Hanushevsky
Stanford Linear Accelerator Laboratory

R. Gardner
University Chicago

F. Barreiro, K. De
University Texas at Arlington

K. Bhatia, K. Kissel
Google

J. Wells
Oak Ridge National Laboratory



ATLAS Rucio data transfer architecture for Google S3

2013 - 2014: BigPanDA project. Extending the scope to cloud computing. Collaborate with Google Compute Engine preview project to run ATLAS jobs at scale

2017: Discussions at Smoky Mountains and Supercomputing conferences. Intensive discussions in Oct-Dec to define Data Ocean Project scope.

Dec : Data Ocean R&D ATLAS internal note.

2018 :

Jan : Google presentation of Proof of Concept Data Ocean project at ATLAS SW&C week.

Feb-Sep : Data Ocean project : User's analysis, data placement, GCP/PanDA integration

May-Aug : bi-weekly technical meetings between US Labs / Universities and Google to discuss potential R&D projects. Six WG with target date Aug 31st for white paper draft. U Tokyo / Google collaboration.

Oct : draft white paper submitted to DOE

Dec : WLCG Mgmt, ATLAS, Google, OpenLab Technical Meeting at CERN

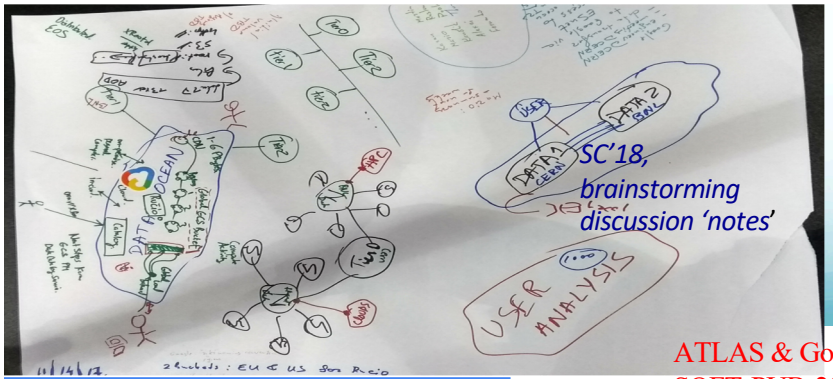
2019 :

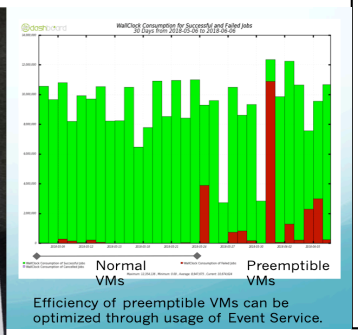
Feb-Mar : v1.0, v2.0 and v3.0 of white paper

Apr : US ATLAS Ops Program Director's review. R&Ds Partnership with Industry talk (KD)

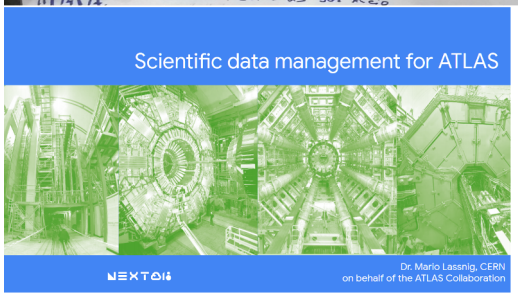
Jun 14th : US-ATLAS, Google meeting with J.Siegrist and DOE HEP and ASCR offices reps.

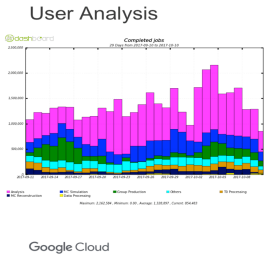
Jun 24/28 : ATLAS Google splinter meeting and discussions during SW&C week @NYU

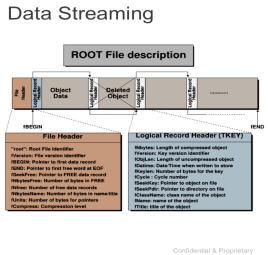




ATLAS & Google — "Data Ocean" R&D Project, ATLAS note ATL-SOFT-PUB-2017-002 https://cds.cern.ch/record/2299146/, 29 Dec 2017



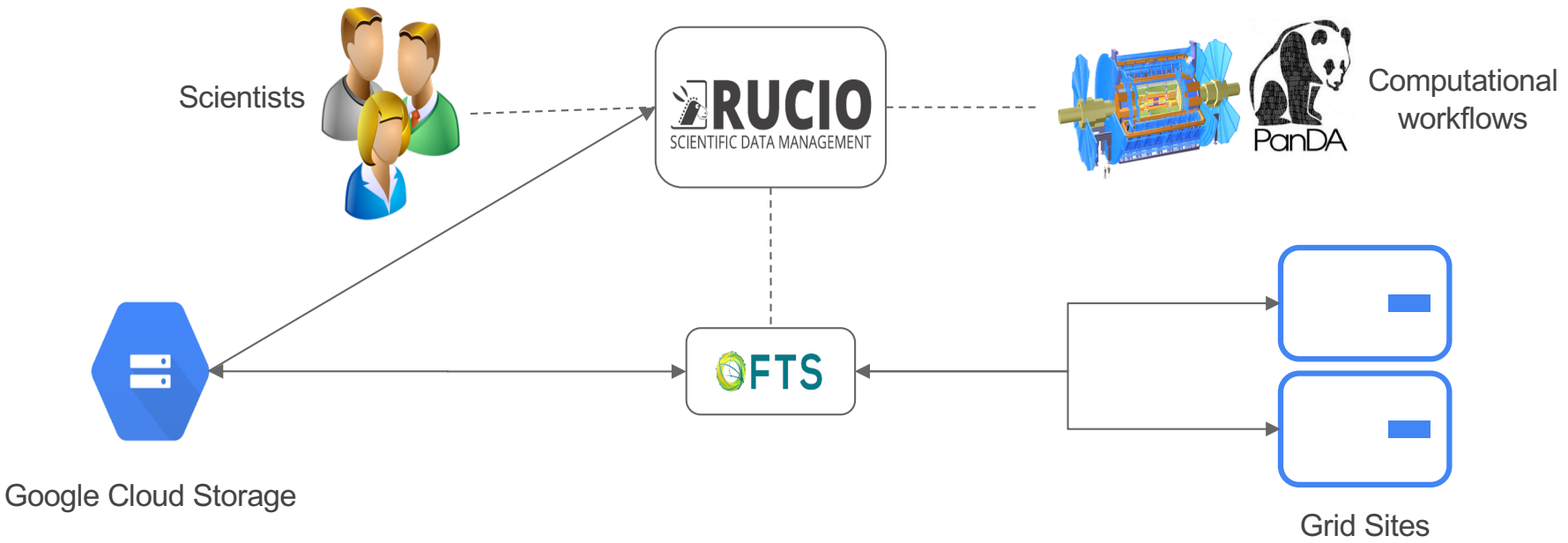




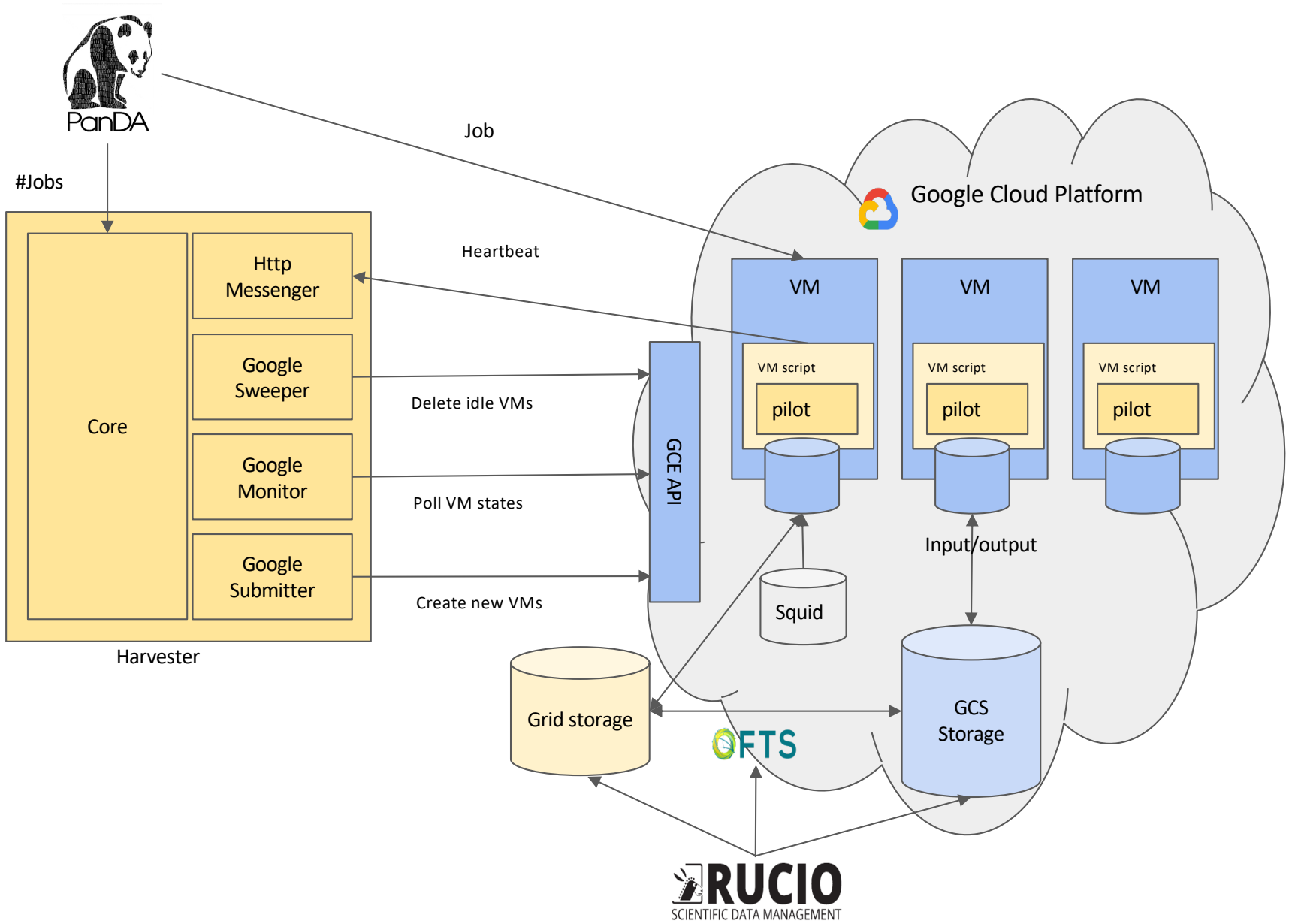
M. Lassnig

Getting data into Google Cloud Storage

- The ATLAS Data Management system *Rucio* orchestrates all experiment transfers
  - S3 used in the first iteration, since support is already available from both sides
  - Tests successful, however not usable for client-based access (key distribution, server-side signing)
  - Parallel third-party copy is rate-limited to 100MB/sec because we were not using the native GCS API
- Decision to move to GCP-native client-side signed URLs

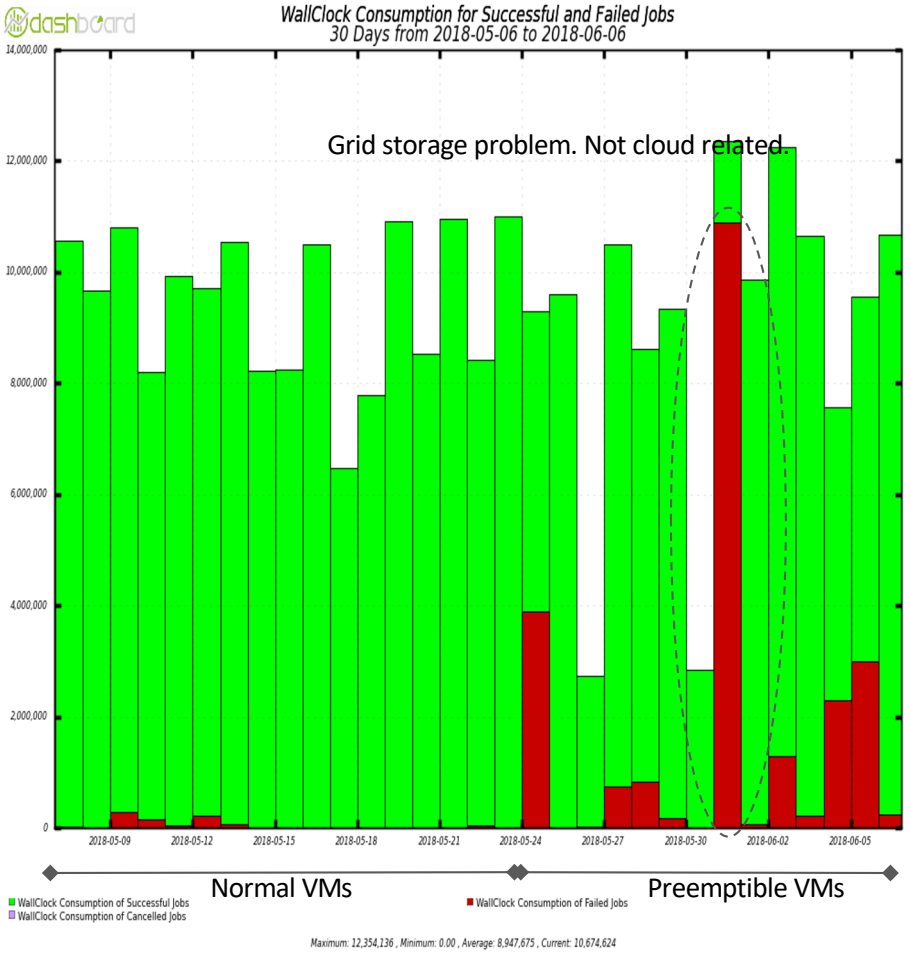


Workload Management and Google Cloud Platform



Compute evaluation for simulation

- Operated a 120 core cluster running standard **simulation** jobs for 1.5 months
  - I/O to CERN storage
  - Excellent success rate (<<5% errors) using normal VMs
- Preemptible VMs
  - Significantly higher error rate (~20%, including a Grid storage outage)
  - Still gain on a $/event basis

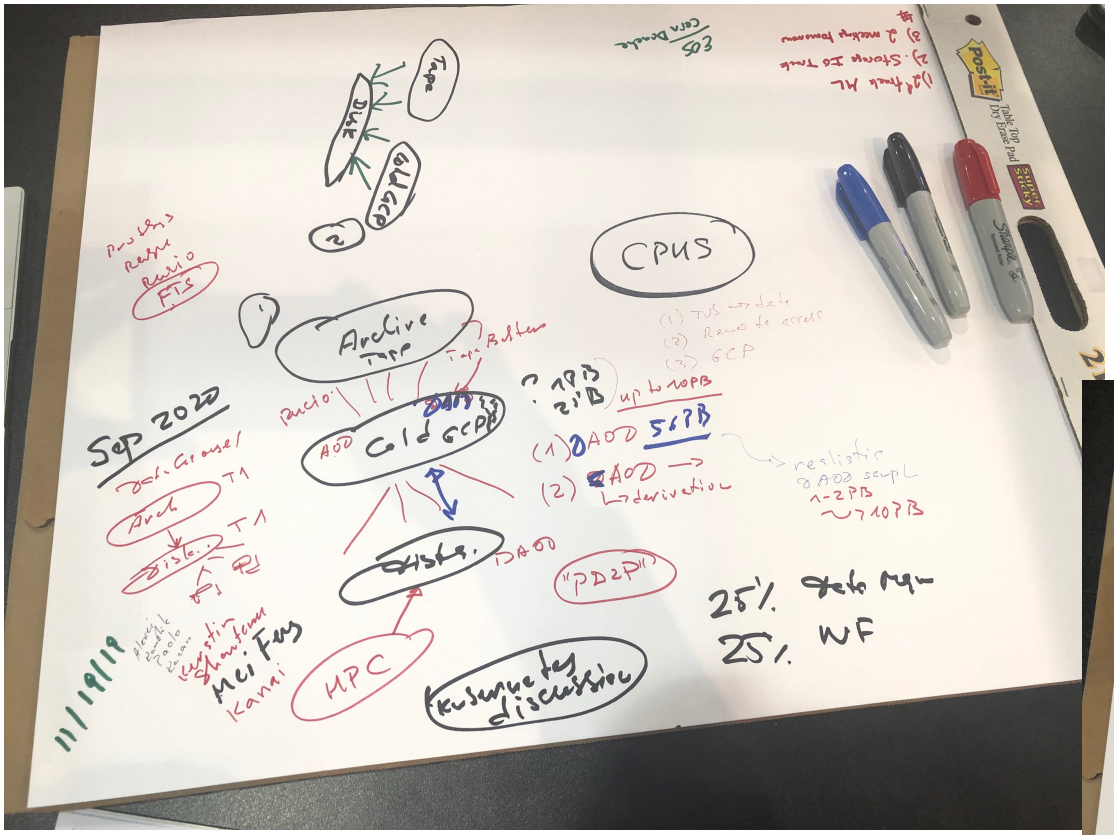


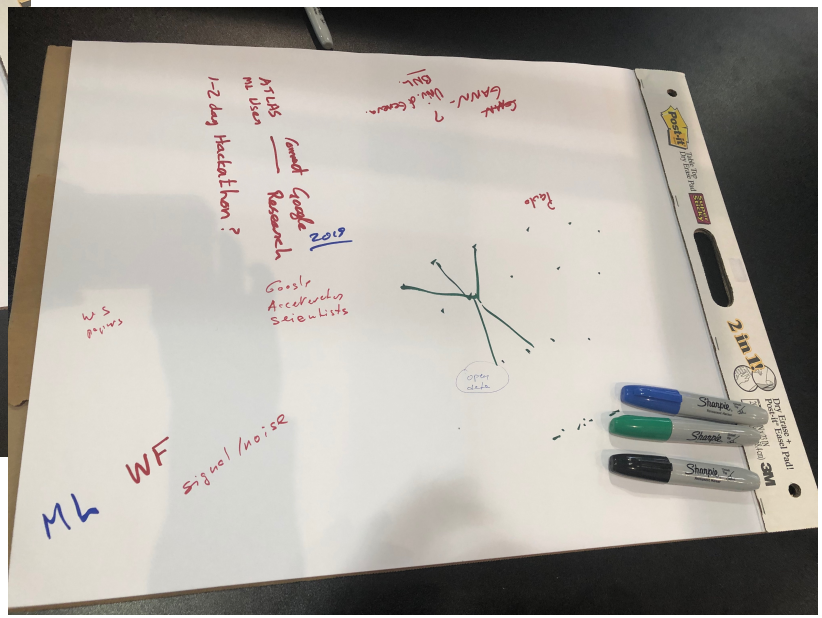
Efficiency of preemptible VMs can be optimized through usage of Event Service.

End-User Analysis use case

- First cloud exercise with native use of cloud storage
  - Pre-placement of datasets to Google Storage
- Full user analysis succeeded
  - ~1M events, 450 GB of input, and 63 GB output
- It was more challenging than simulation
  - Incompatibilities between CernVM4 and Athena pre 21 releases
    - Requires generating and pre-upload of shared libraries
  - Preemptible VMs create confusion
  - OOM reaper killing payloads
  - File corruption errors when using direct I/O: need to stage-in full files

Nov 19, 2019; Super Computing Conference
US ATLAS/Google brainstorming
Alexei, Ema, Kanai, Karan,
Kaushik, Kerstin, Meifeng,
Miles, Paolo, Shantenu

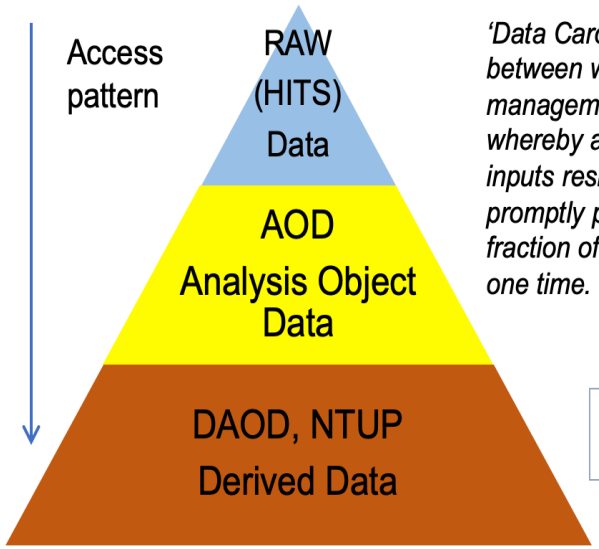




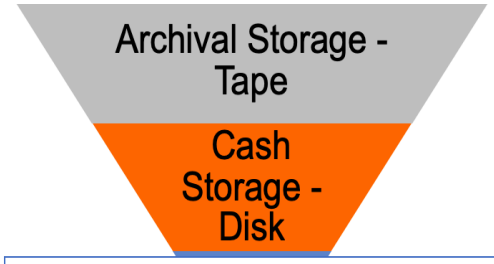
ATLAS Google Collaboration

- "Proof of Concept" project success has led to expanded work plan
  - Geared towards HL-LHC, leveraging Google expertise
  - Expanded technical teams, both within ATLAS and Google experts
    - Five areas of collaboration identified in white paper (after ~4 months of technical discussions). They are in various stages from planning to active technical work. They are attracting interest (of different level) from HEP and WLCG communities and funding agencies.

Track 1	Data Management across Hot/Cold storage
Track 2	Machine learning and quantum computing
Track 3	Optimized I/O and data formats
Track 4	Worldwide distributed analysis
Track 5	Elastic computing for WLCG facilities



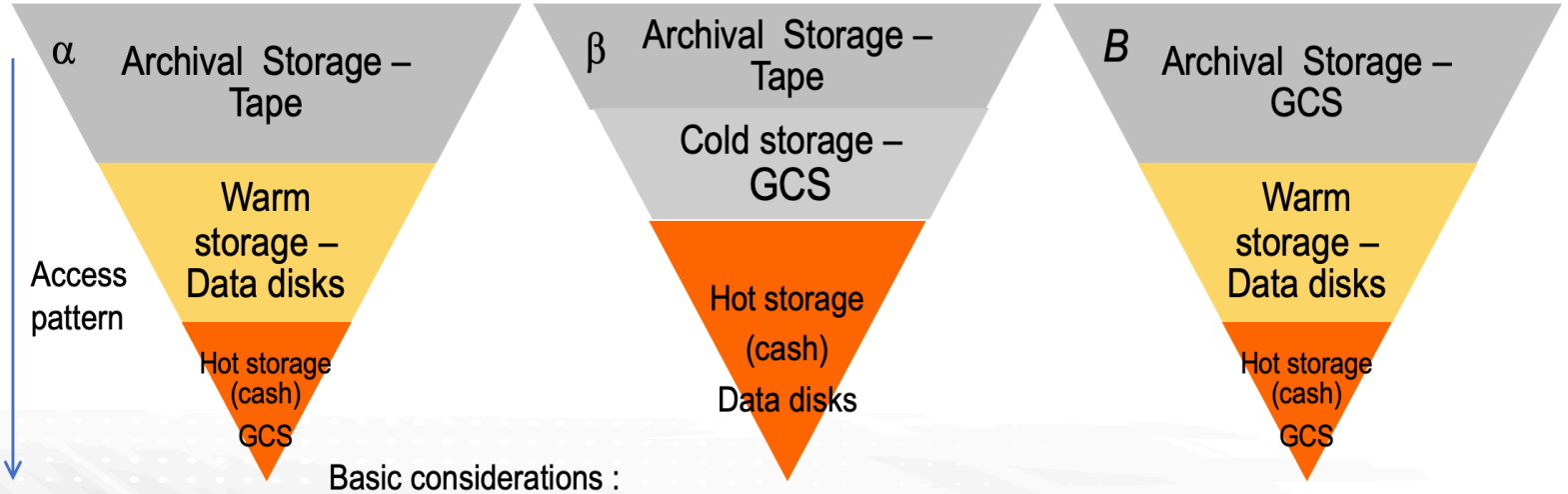
'Data Carousel' we mean an orchestration between workflow management (WFMS), data management (DDM/Rucio) and tape services whereby a bulk production campaign with its inputs resident on tape, is executing and promptly processing of inputs. Only a small fraction of inputs are pinned on disk at any one time.



Data Carousel Model. Automatic data migration between disk and tape

Hot/Cold Storage Model

Hot/Cold storage model gives us more flexibility with data handling .
We can archive ALL data on tape and keep on disk and cash the most popular data
Plan A (α, β): Data will migrate between hot/warm/cold storage automatically
Plan B : will address the case when tape drives market will be in danger



Basic considerations :

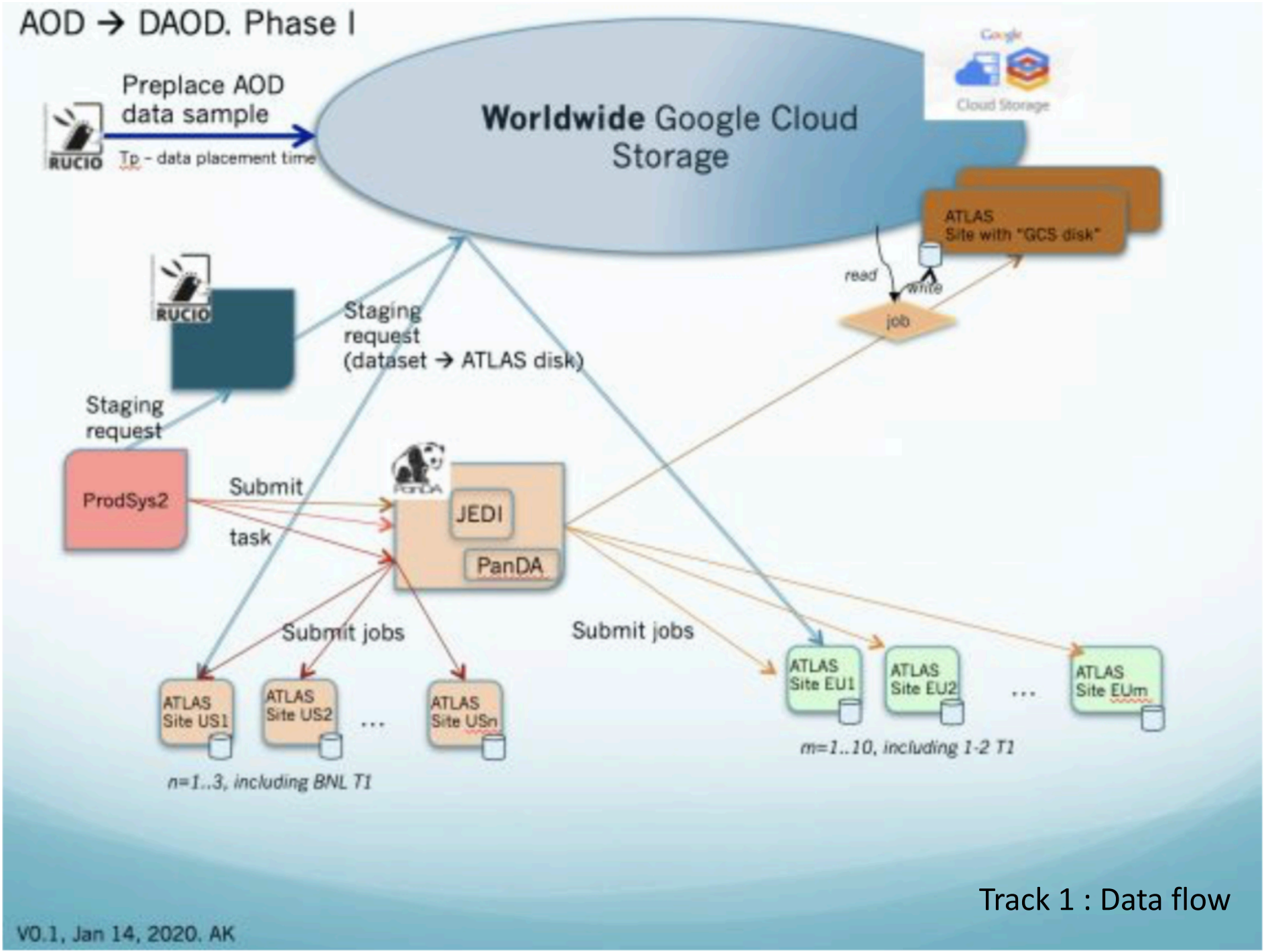
1. Access pattern
2. Cost, performance and capability
   1. Capability = functionality.. How well requirements are managed
   2. Performance = data availability, retrieval speed and data access speed

U.S. DEPARTMENT OF ENERGY

8/06/19

BROOKHAVEN NATIONAL LABORATORY

Track 1 : Hot/Cold storage

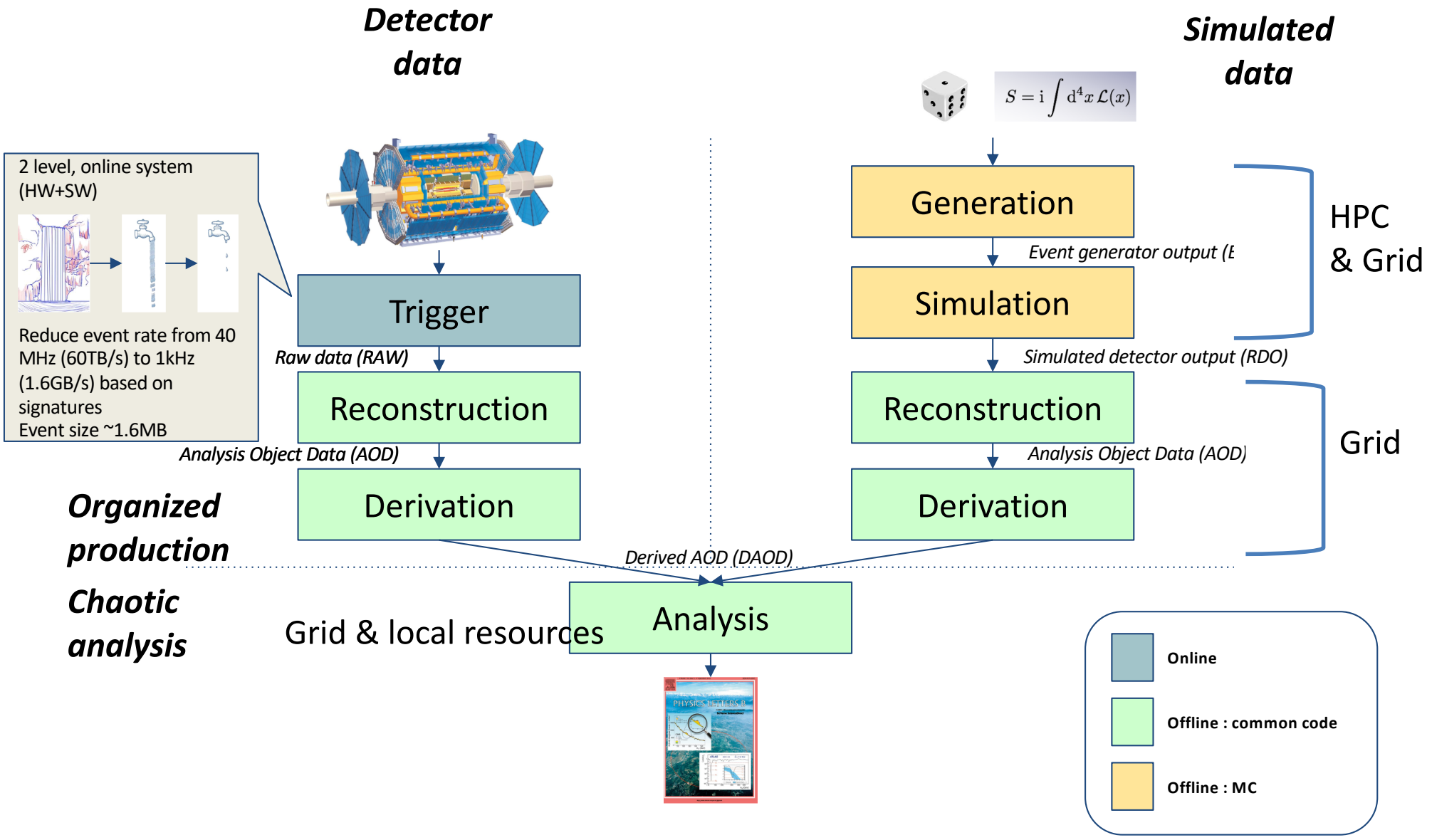


Track 1 : Data flow

Data and Workflow/load Management

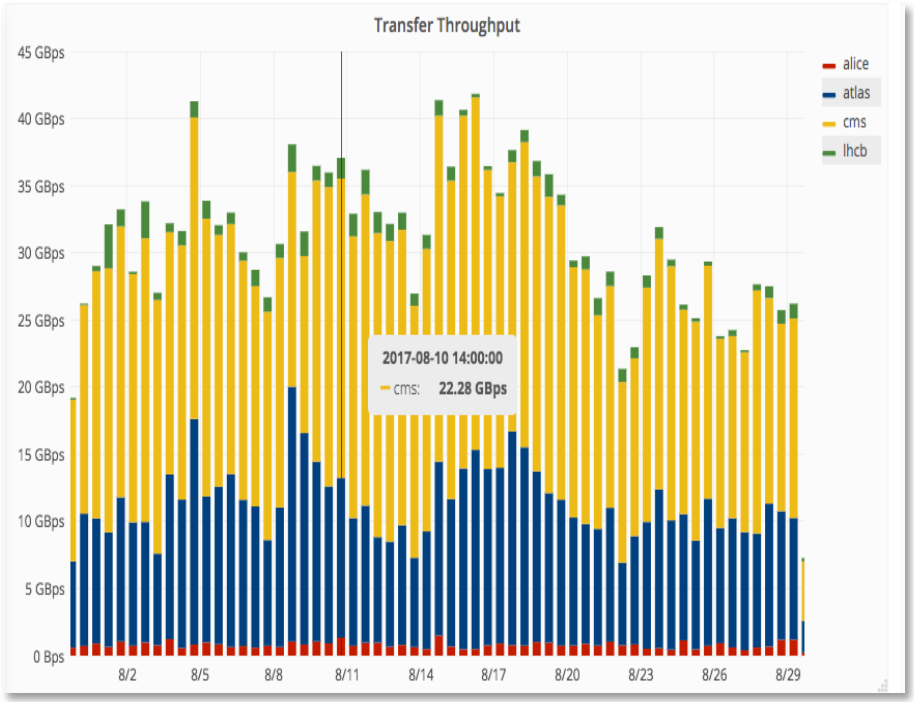
- Data management (DDM – Distributed Data Management)
- Workflow and Workload management (WMS – Workload Management System)
- Monitoring
- WMS evolution
- Beyond ATLAS and HEP

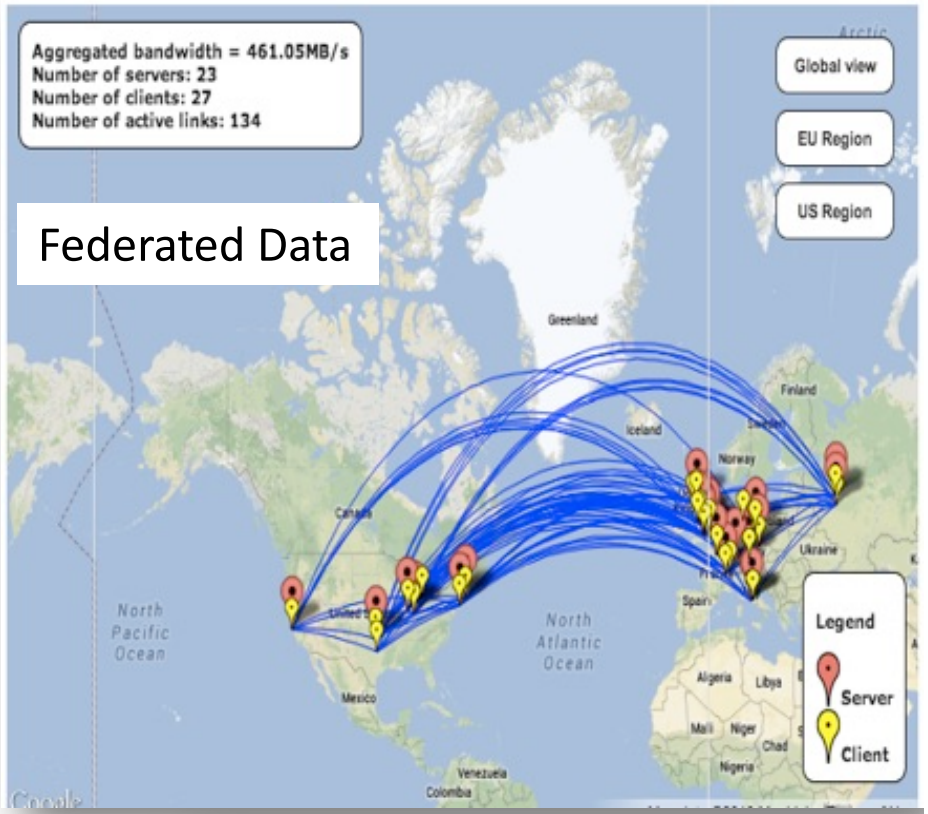
The data processing chain



Data distribution

- Global transfer rates increased to 30-40 GB/s (>2 x Run1)





Increased performance everywhere:

- Data acquisition >10PB / month
- Data transfer rates > 35 GB/s globally

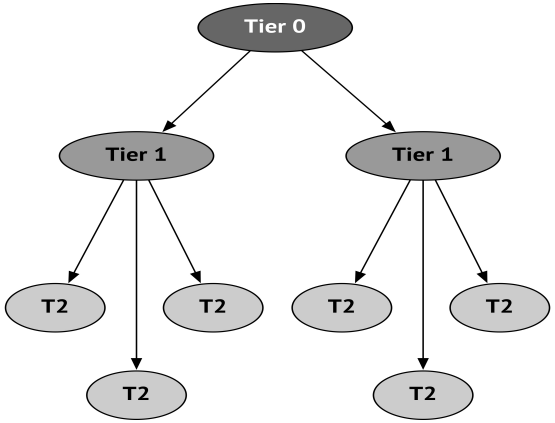
Regular transfers of >80 PB/month with ~100 PB/month during July-October (many billions of files)

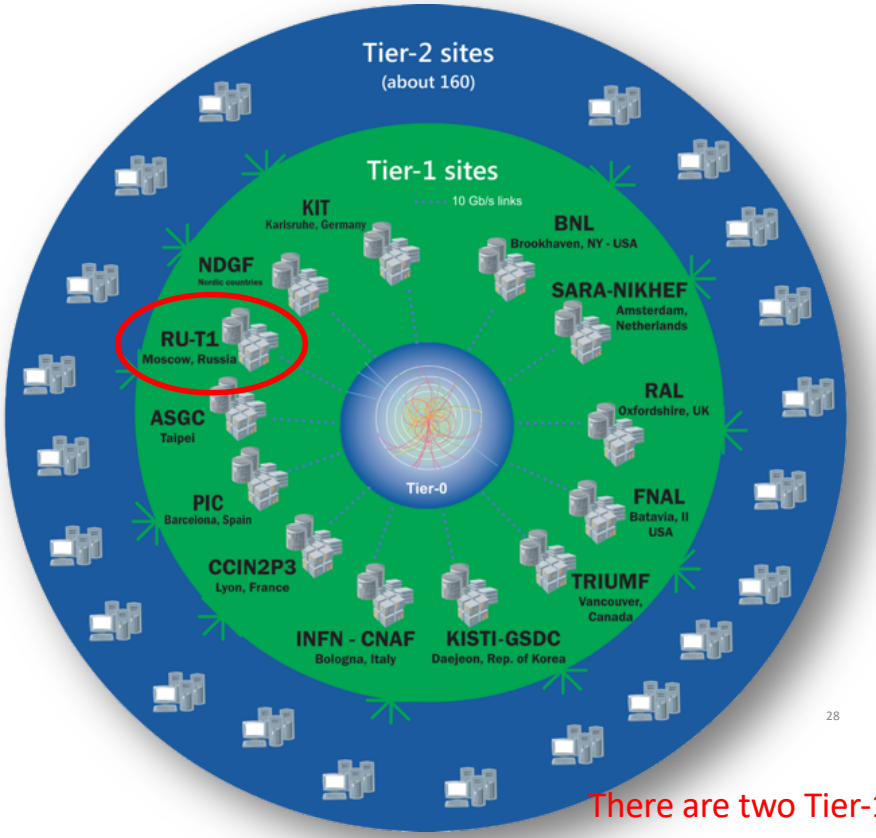
The Worldwide LHC Computing Grid

Tier-0 15%
(CERN and *Hungary*):
data recording, reconstruction and distribution

Tier-1 40%: permanent storage, re-processing, Analysis
T0 spill-over
HLT
MC Simulation
Derivation production

MONARC - Models of Networked Analysis at Regional Centres for LHC Experiments.





Tier-2 45%: Simulation, end-user analysis
Re-processing
Derivation production

~170 sites, 42 countries

~750k CPU cores

~1 EB of storage

> 2 million jobs/day

10-100 Gb links

There are two Tier-1s in Russia : JINR and NRC KI

WLCG:
An International collaboration to distribute and analyse LHC data

Integrates computer centres worldwide that provide computing and storage resource into a single infrastructure accessible by all LHC physicists

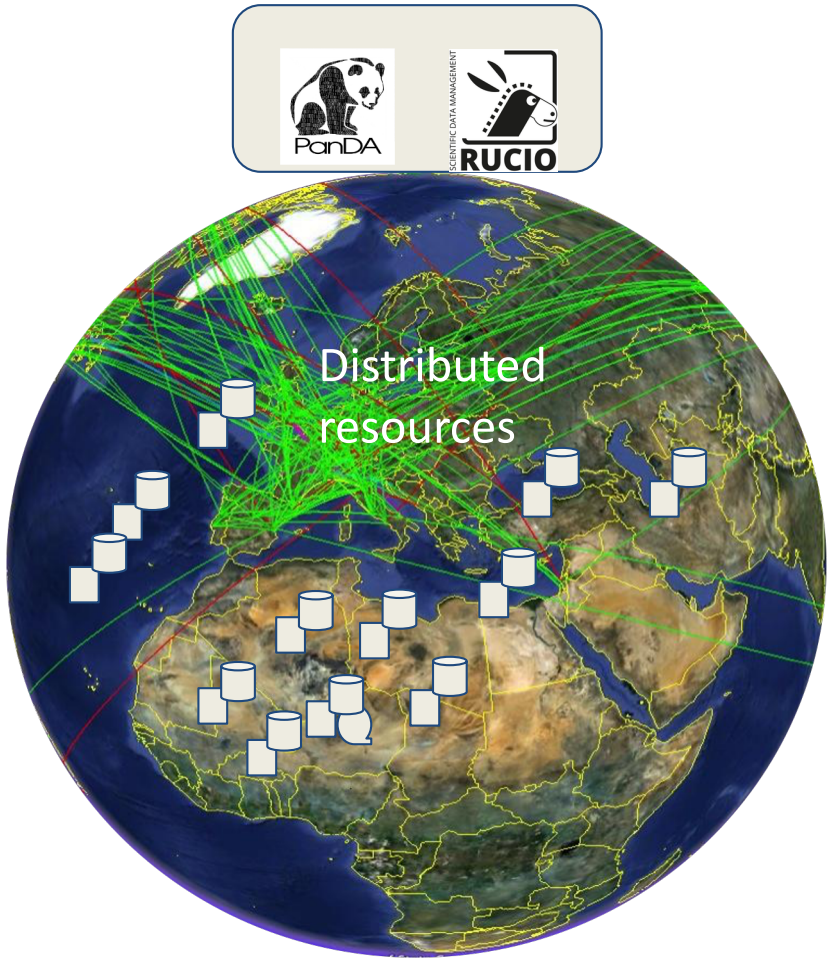
© 2016 Google

Primary distributed computing software tools

Workflow Management:
"translates" physicist requests into production tasks

Workload Management:
submission and scheduling of jobs & tasks

Monitoring production jobs & tasks, shares, users



Data Management:
bookkeeping and distribution of files & datasets

Information System
(ORACLE backend)
Queues and resources description

Databases: Conditions and data processing (ORACLE, mySQL, PostgresSQL)

Paradigm shift in Particle Physics Computing in XXI century

Old paradigms	New ideas
• Distributed resources are **independent entities**	• Distributed resources are seamlessly **integrated worldwide** through a single submission system • **Hide middleware** while supporting diversity
• Groups of users utilize **specific resources** (whether locally or remotely)	• Access to all **resources** may be granted to all users
• Fair shares, priorities and policies are managed **locally**, for each resource	• **Global** fair share, priorities and policies allow efficient management of resources
• **Uneven user experience** at different sites, based on local support and experience	• **Automation, error handling, and other features** improve user experience • Central support coordination
• Privileged users have access to **special resources**	• All users have access to **same resources**

Outline

- Data management
- Workload management
- Monitoring
- WMS evolution

Distributed Data Management in a nutshell

- Stores and manages all the experiment's data across a distributed environment following the computing model principles
  - Computing model determines the number and location of copies of different types of data
- High-level requirements:
  - **Data bookkeeping**
    - Location of files and datasets
    - Relationship between files and datasets
    - Owner, checksum and other metadata
  - **Data transfers**
  - **Data deletion**
  - **Data consistency**
    - Are the files really where we think they are?

Rucio
(ATLAS)

PhEDEx
(CMS)

AliEn
(Alice)

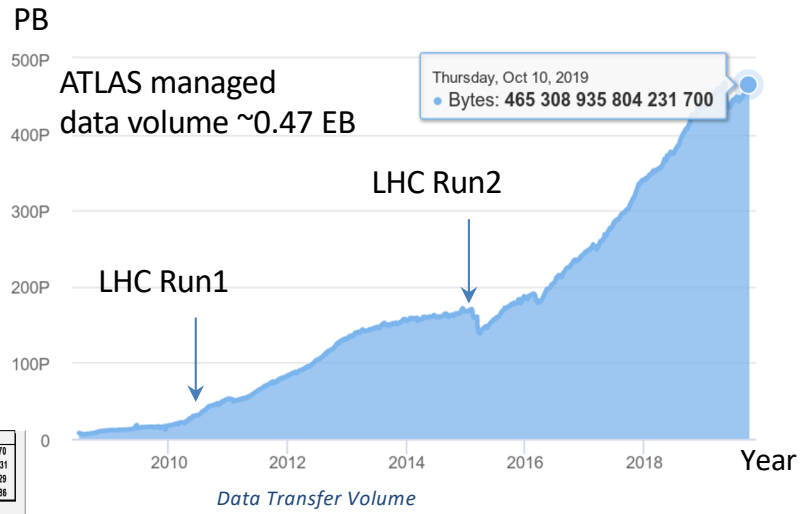
DIRAC
(LHCb)

- Each experiment has their own with slightly different features: we will focus on ATLAS Rucio (http://rucio.cern.ch/ developed by CERN and Univ. Oslo)

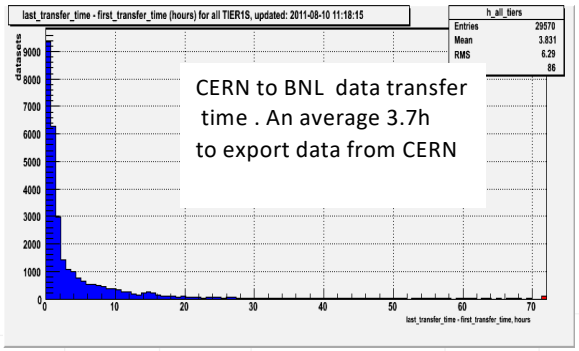
Data management. Rucio

- A few numbers to set the scale
  - 1B+ files, 460+ PB of data, 400+ Hz interaction
  - 120 data centres, 5 HPCs, 2 clouds, 1000 users
  - 500 Petabytes/year transferred & deleted
  - 2.5 Exabytes/year downloaded & uploaded

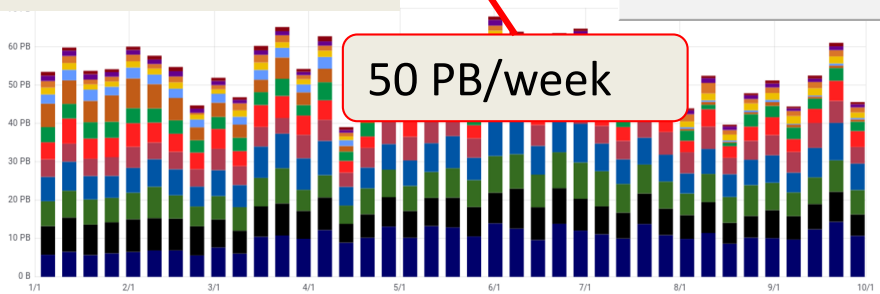
Rucio is evaluating or already in use for many experiments including Belle II, CMS , SKA, AMS



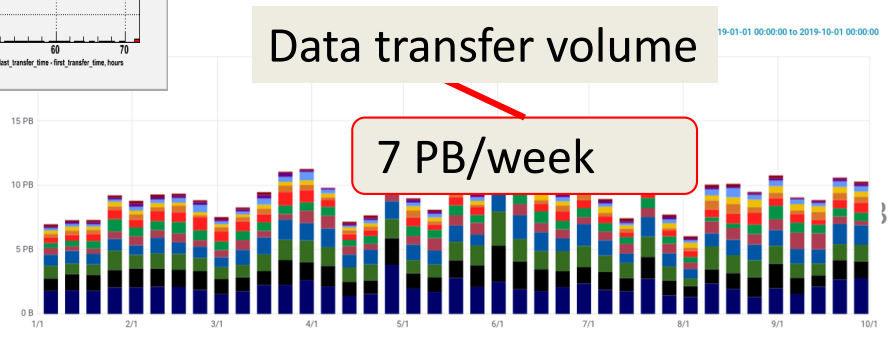
Data Transfer Volume



Data access volume



Data transfer volume



First exascale scientific data management system today

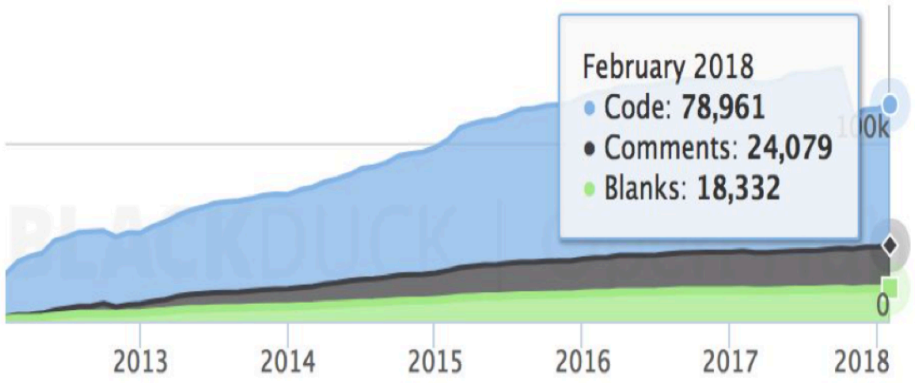
Data Management Tools

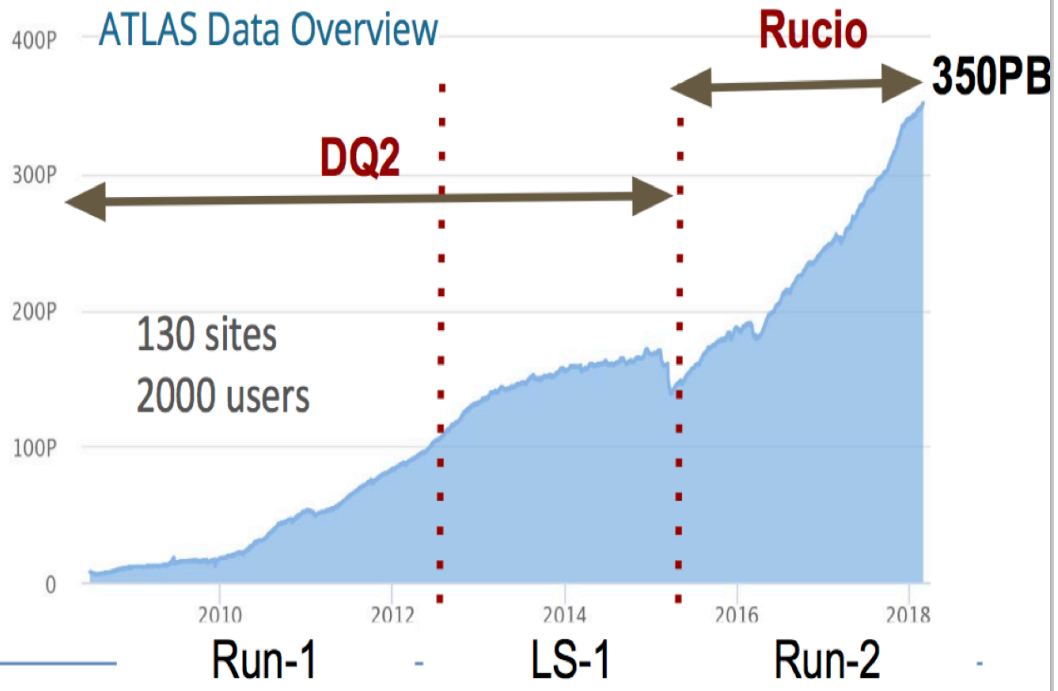
- At time of inception, no global/commercial solution for the distributed computing available for our 'Big Data' handling
  - A data intensive instrument which generates unprecedented data volumes
  - Facilities are distributed at multiple locations under different administrative domains
  - Data is produced at many locations where it is neither stored, nor analyzed by researchers nor archived
- ATLAS developed its own tools
  - The first implementation of the data management system was Don Quijote 2 (DQ2)
  - In production from 2006 : Originally designed as a transfer system
  - 2007-2013: Many new features added during LHC Run-1

Rucio Development & Commissioning

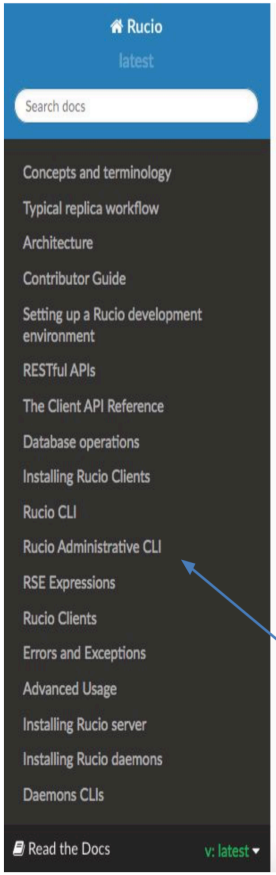
- Long initial process:
  - 2012: User surveys, technical studies & design phase ~1 year
  - 2012-2014: Initial development ~2 years
  - 2015: Commissioning & gradual migration from predecessor system DQ2 ~1 year

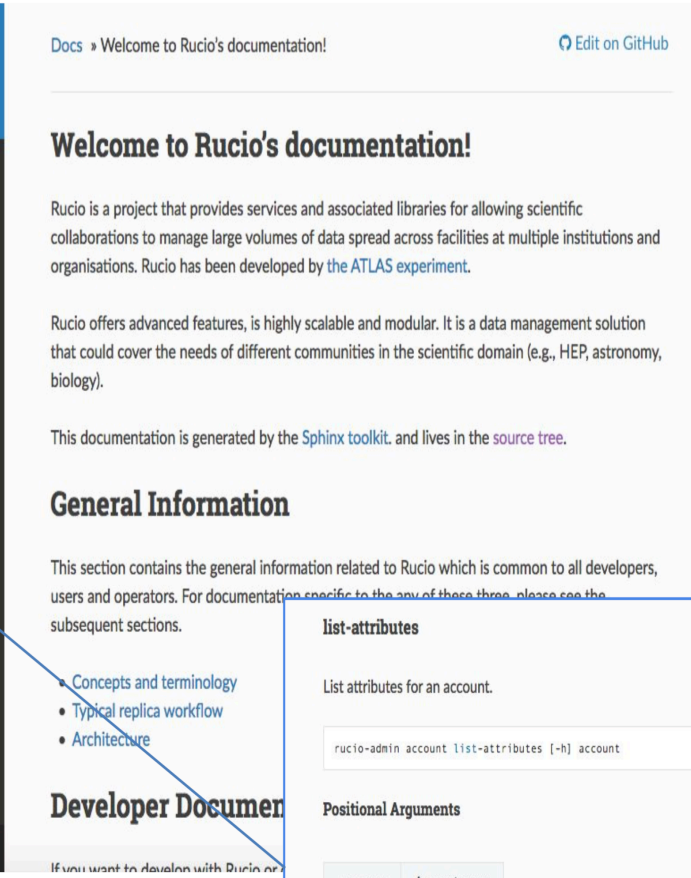
Lines of Code

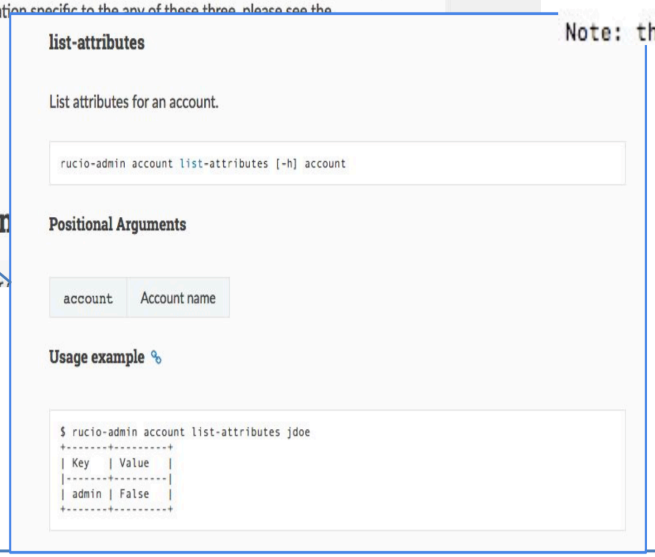




Documentation



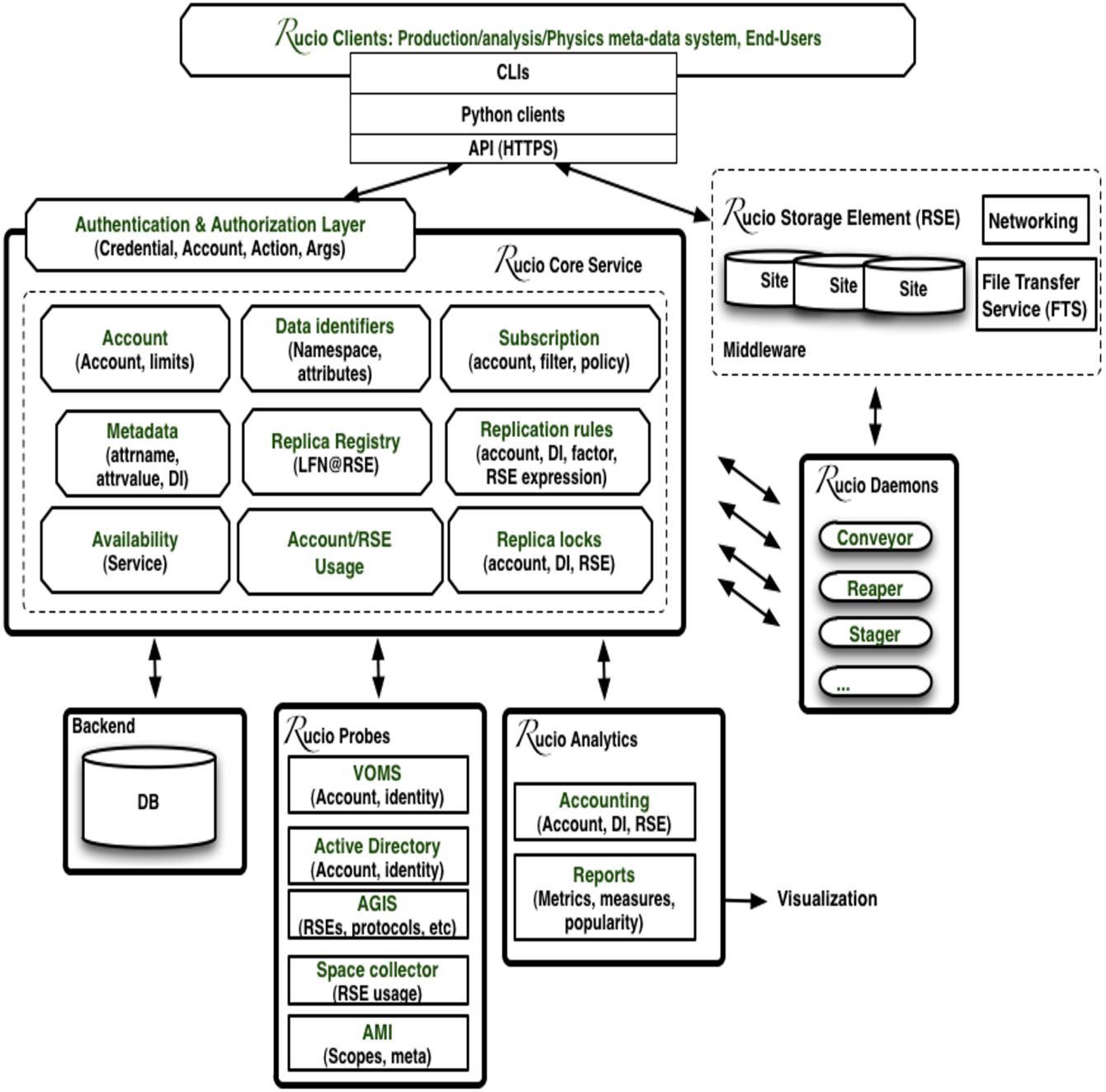





```
[root@rucio-nagios-prod-02 ~]# rucio-admin account list-attributes -h
usage: rucio-admin account list-attributes [-h] account

positional arguments:
  account     Account name

optional arguments:
  -h, --help  show this help message and exit

Usage example
"""""""""""""

    $ rucio-admin account list-attributes jdoe
    +-------+---------+
    | Key   | Value   |
    |-------+---------|
    | admin | False   |
    +-------+---------+

Note: this table empty in most cases.
```

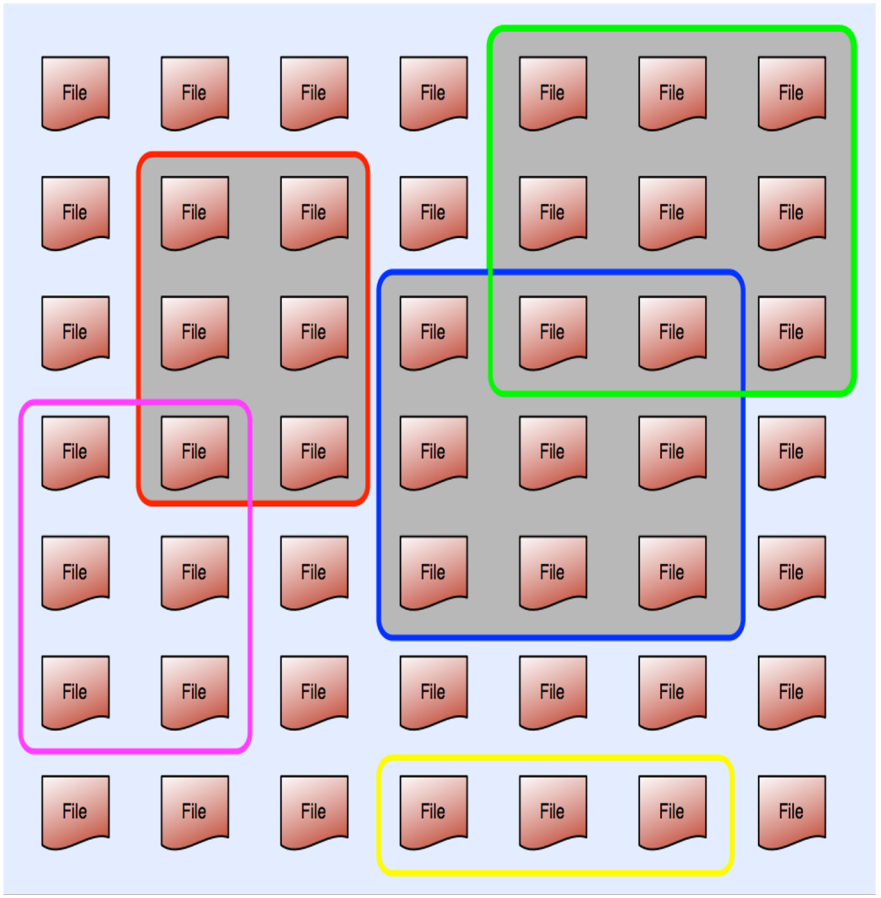
https://rucio.readthedocs.io/

Data Management: Rucio architecture



Rucio data concepts

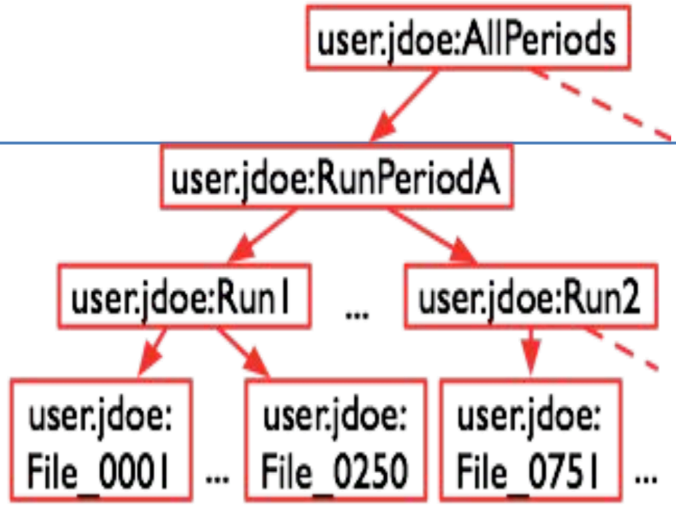
- Events: collisions
- Files: Collections of events (e.g. C++ objects)
- Datasets: logical grouping of files
  - Units of replication

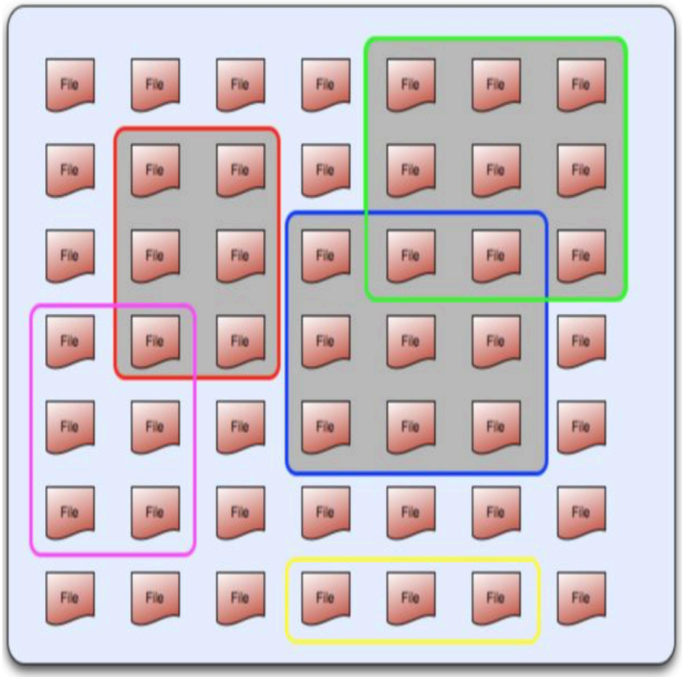


Data Hierarchy

- At the heart of everything is a file*
- Files are grouped into datasets
- Datasets are grouped into containers
  - Datasets only hold files
- Containers are grouped into containers
  - containers only hold datasets or containers
- Collections can be organised freely
  - Files can be in multiple datasets
  - datasets can be in multiple containers
  - containers can be in multiple containers

* sub-file support being explored



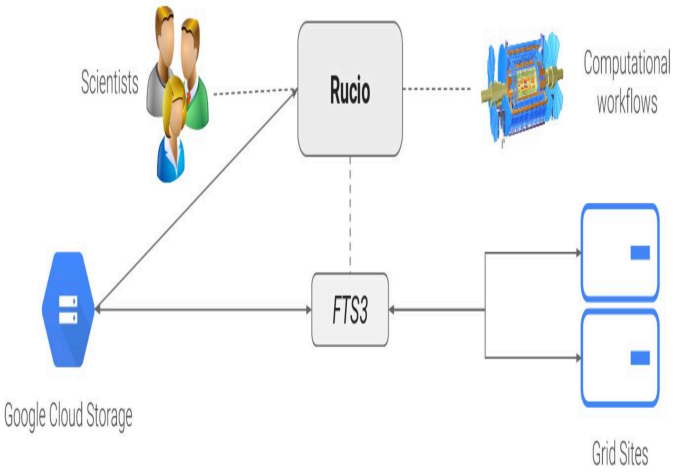


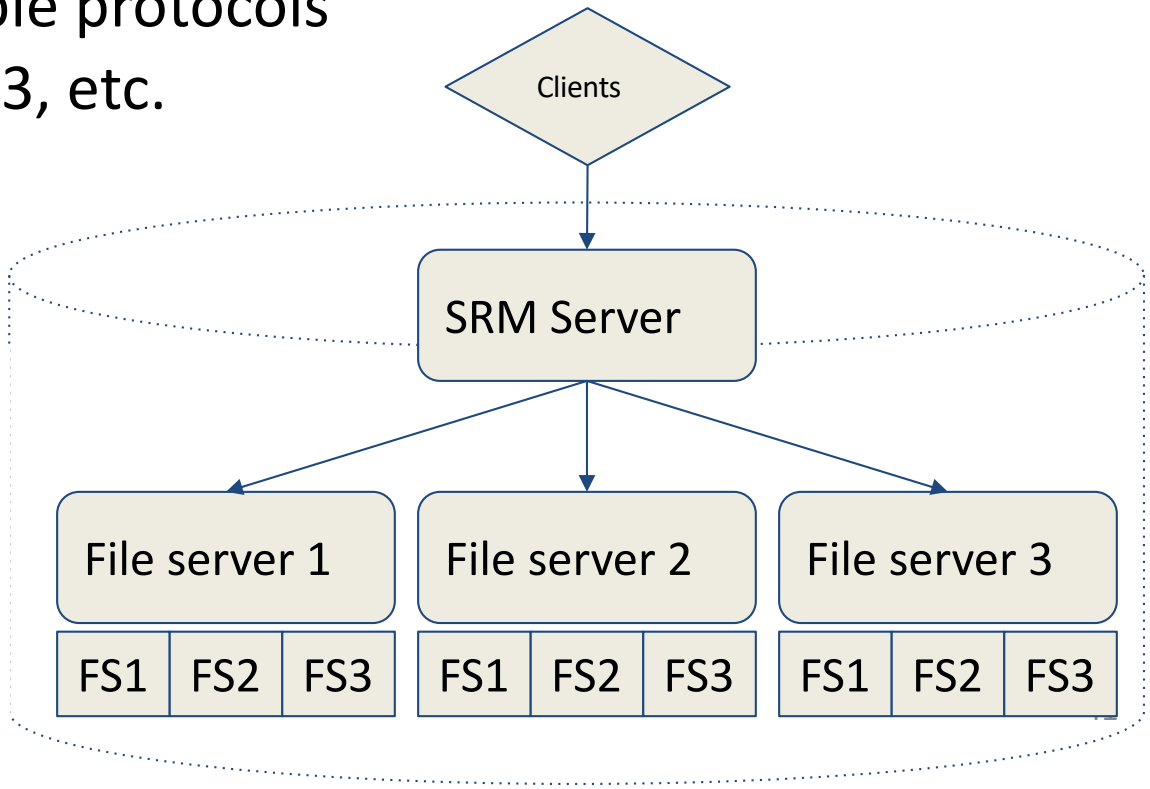
Metadata

- Metadata are custom attributes on data identifiers
  - Support for arbitrary metadata being explored
- Rucio supports different kinds of metadata
  - System-defined, e.g., size, checksum, did_type, is_open, created_at
  - Physics, e.g., number of events, GUID
  - Workflow management system, e.g., which task or job produced the file
  - Data management, necessary for the organisation of data
- Metadata provides another namespace
  - Datasets are searchable by name and metadata

Rucio Storage Element (RSE)

- Software abstraction for a storage end-point
  - E.g. CERN_DATADISK, MEPHI_DATADISK,...
- A deterministic mapping between the logical file name and its path can be used to remove file catalog lookups
- RSEs support multiple protocols
  - GridFTP, HTTP, S3, etc.

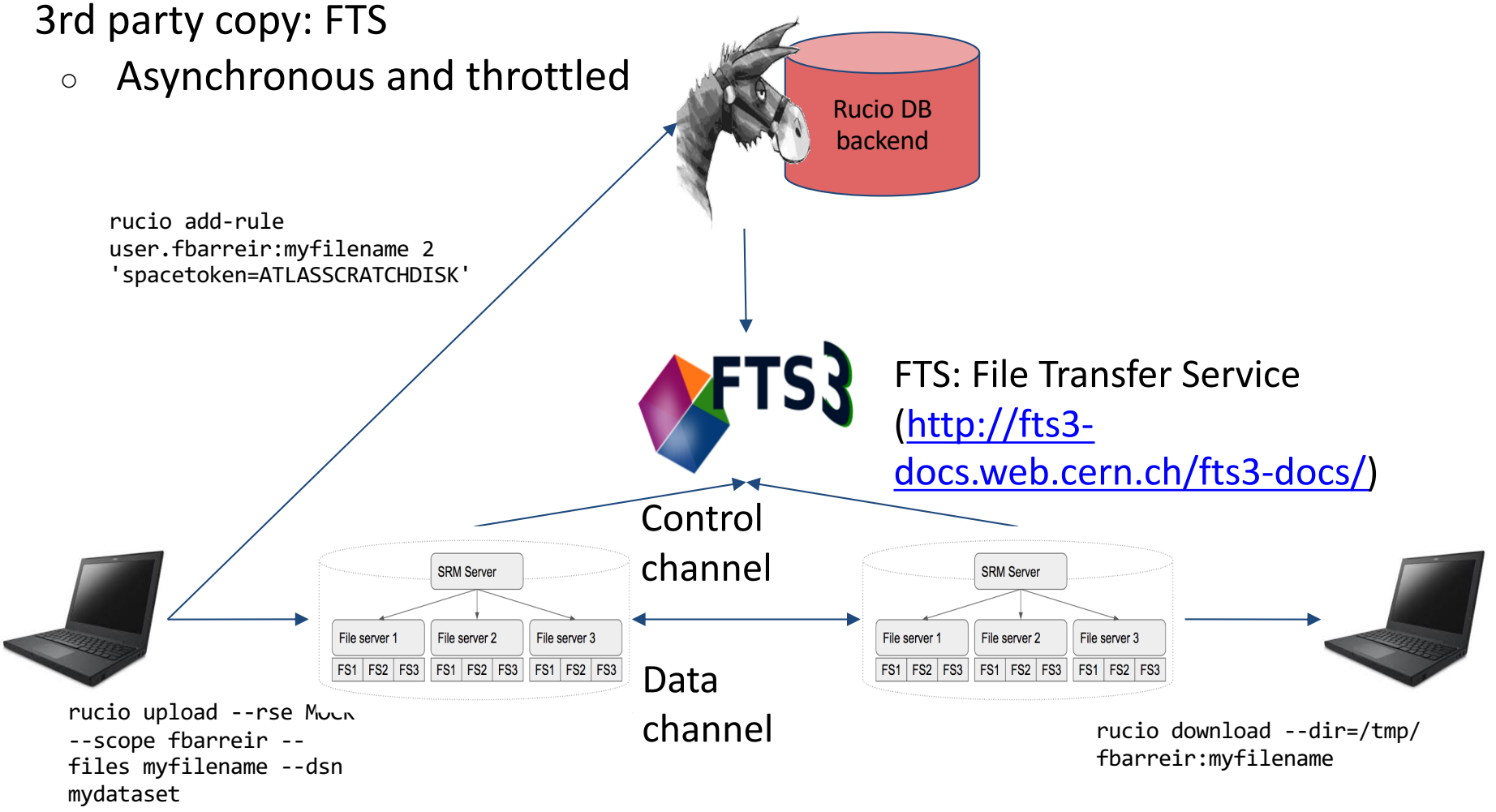




SRM: Storage Resource Manager

Interaction with the data

- Upload and download
  - Synchronous
- 3rd party copy: FTS
  - Asynchronous and throttled



Rucio hides all the complicated details (paths, protocols, hosts) from the users!

Listing, copying and removing files.

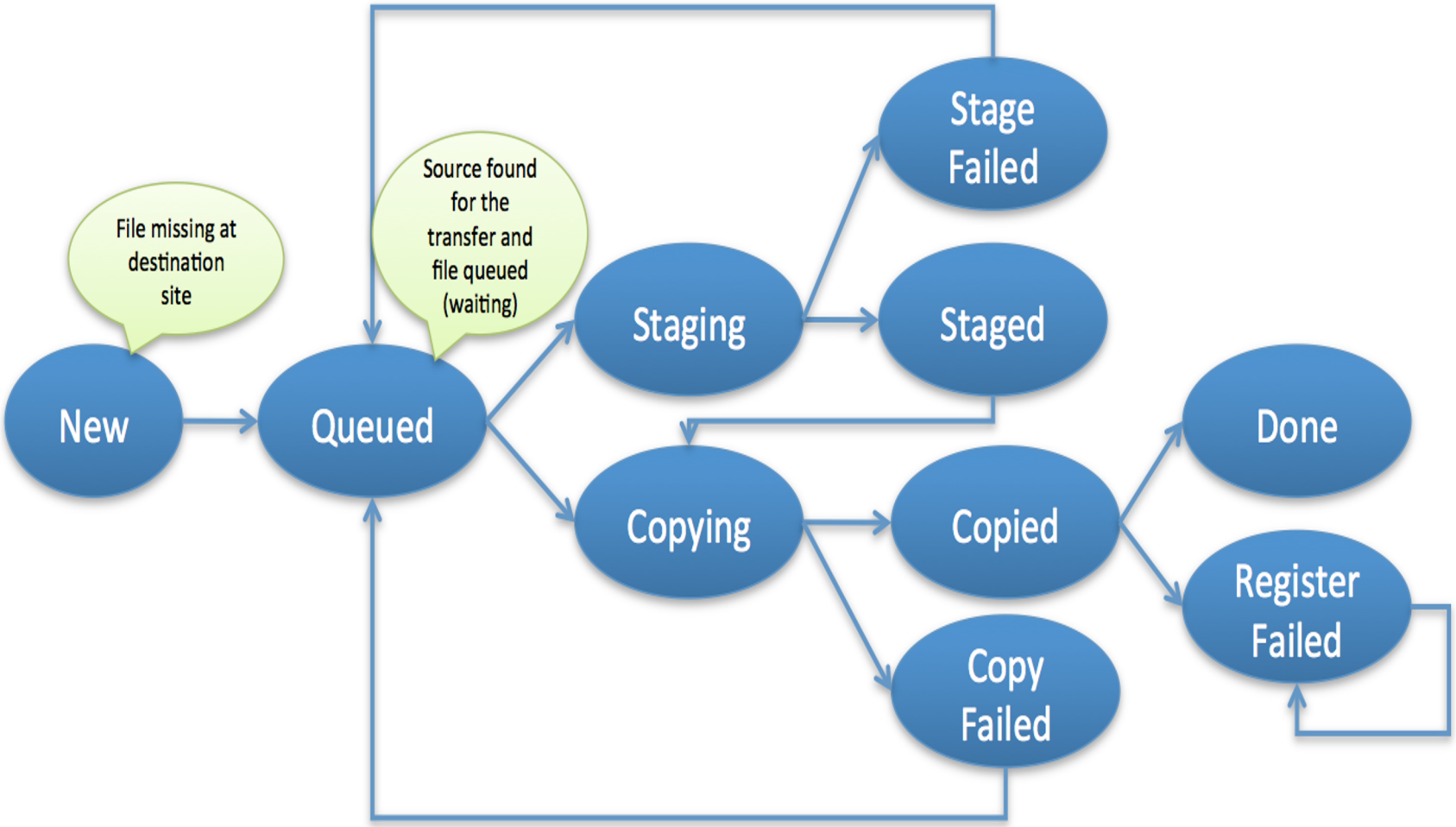
```
[ui03] > edg-gridftp-ls --verbose gsiftp://i2g-se01.lip.pt/flatfiles/itut
total 4
drwxrwxr-x    3 itut         4096 Nov  8 15:06 generated
drwxrwxr-x    2 itut         4096 Nov  8 18:32 tut-14-11-07
```

```
[ui03] > globus-url-copy -vb file:/home/tutorial/user01/data-manag/dm-file-user01
gsiftp://i2g-se01.lip.pt/flatfiles/itut/tut-14-11-07/dm-file-user01
      1048576 bytes       329.49 KB/sec avg       329.49 KB/sec inst
```

```
[ui03] > edg-gridftp-ls --verbose gsiftp://i2g-
se01.lip.pt/flatfiles/itut/tut-14-11-07
total 9412
-rw-rw-r--    1 itut      9621413 Nov  8 18:33 dm-file-user01
```

```
[ui03] > edg-gridftp-rm gsiftp://i2g-
se01.lip.pt/flatfiles/itut/tut-14-11-07/dm-file-user01
```

State machines: file transfer



Dark data and consistency checks

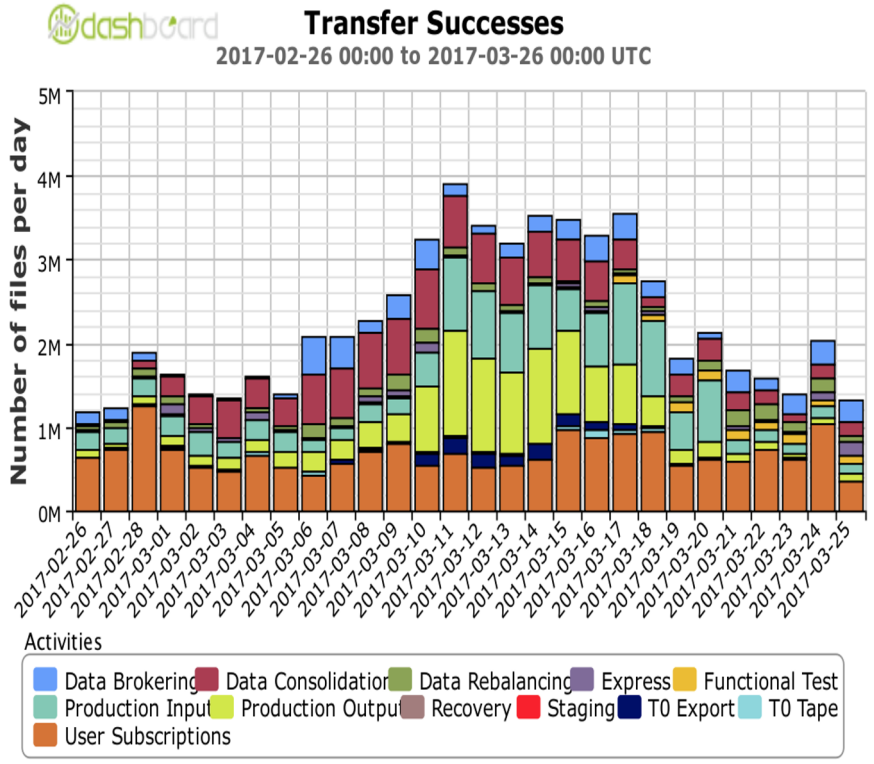
- Consistency between the Rucio Storage Elements and the Rucio DB
  - **Lost Files:** Files in the catalog(s) but not physically on the SE
  - **Dark Data:** Files on the SE but not registered in the catalog(s)
- Automatic consistency check is based on comparison of information dumps
  - Each site provides storage elements dumps on a regular basis (monthly or quarterly)
  - Rucio dumps of expected file replicas generated every day
- Comparison times scale from **few seconds** for small sites to **few hours** for the biggest one (70M files)
  - **Dark Data** is automatically collected and deleted by another daemon
  - **Lost Files** are reported to site support for investigation
    - Confirmed Lost Files are
      - Copied from other SEs if other copies exist
      - Notified to the owner and deleted from the dataset

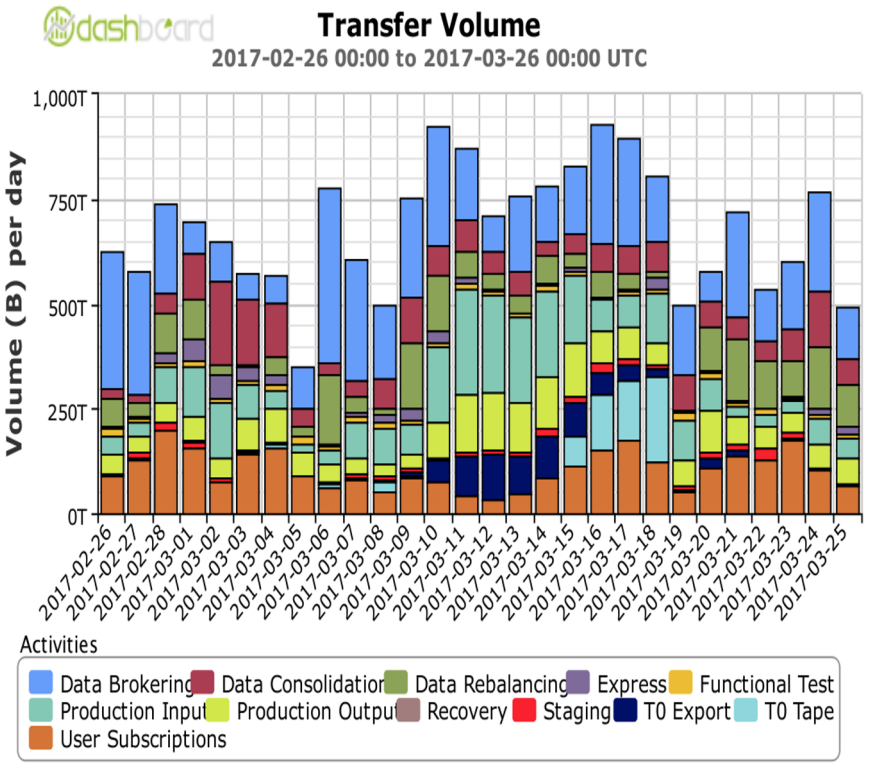
Traces, data popularity and analytics

- Common questions
  - Which files/datasets are popular in the system?
  - Which files/datasets are not used at all?
  - Statistics on transfers times, deletion times, etc.
- Traces: each event is sent to HDFS (Hadoop File System)
  - Important information: event type, file, dataset, source/destination, user, size, transfer time
  - 6M json dictionaries per day (~5GB)
- **Data reduction:** redundant, unused copies of old data can be removed
- **Data pre-placement:** popular data can be replicated to facilitate usage
- **Network map:** measure current bandwidth between sites

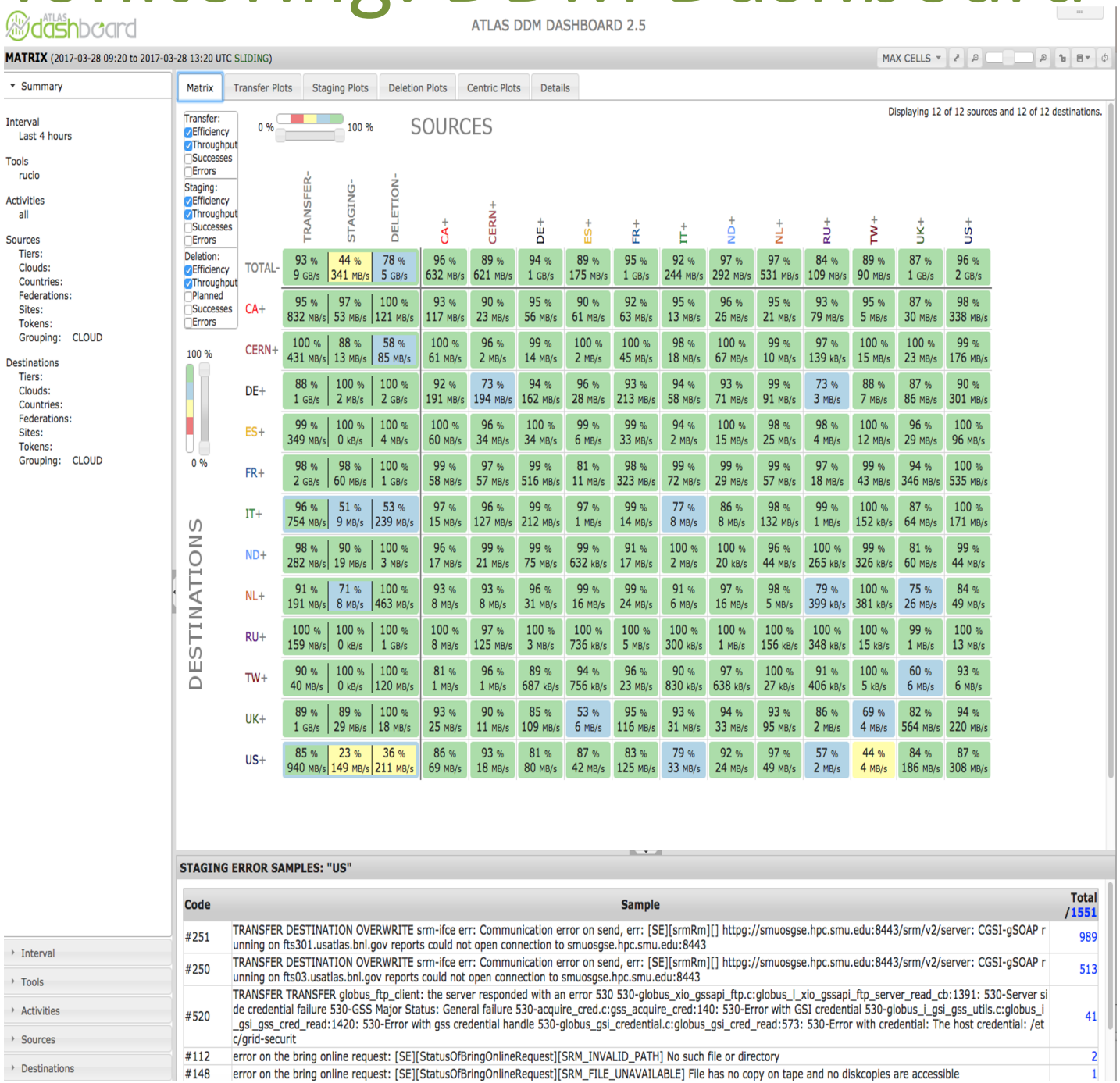
Data Management: some metrics

- Transfers
  - >40M files/month
  - Up to 40 PB/month
- Deletion
  - 100M files/month
  - 40 PB/month
-

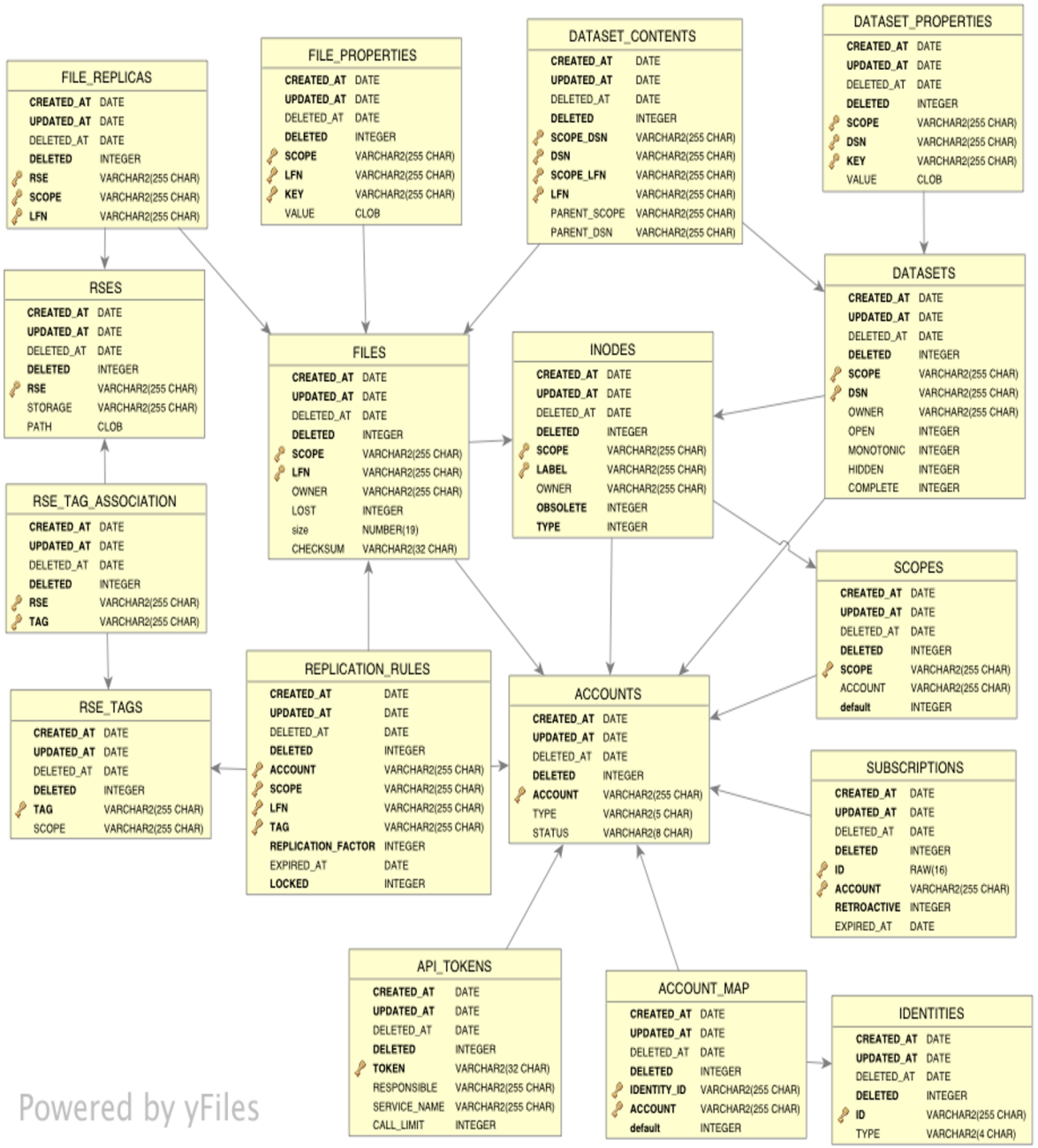




Monitoring: DDM Dashboard



Database schema



http://rucio.cern.ma.png

Powered by yFiles