Experiences with the new ATLAS Distributed Data Management System

Vincent Garonne on behalf of the ATLAS collaboration

CHEP, SLAC/LBNL, 2016

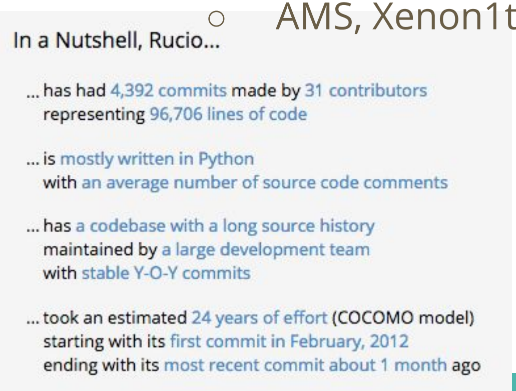
DDM in a Nutshell

- The **D**istributed **D**ata **M**anagement project is charged with managing all ATLAS data
- All for the purpose of helping the ATLAS collaboration to store, manage and process LHC data in a heterogeneous distributed environment
- Requirements:
  - Discover data
  - Transfer data to/from sites
  - Delete data from sites
  - Ensure data consistency at sites
  - Enforce ATLAS computing model

➥ The current DDM system relies on the Rucio software project, developed during Long Shutdown 1 to address the challenges of run2

Rucio - gitlab.cern.ch/rucio01/rucio

- Rucio exploits commonalities between experiments and other data intensive sciences to address HEP experiments needs and scaling requirements
- Rucio is an evolution from our previous Data management system, DQ2, used during Run-1
- One of our goal is to build a broader support community
  - AMS, Xenon1t, etc.

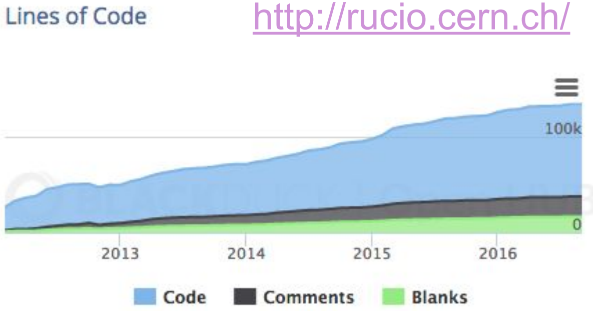

In a Nutshell, Rucio...

... has had 4,392 commits made by 31 contributors
representing 96,706 lines of code

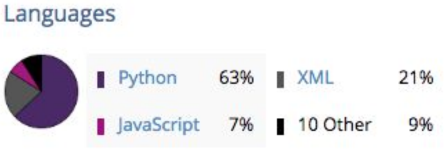
... is mostly written in Python
with an average number of source code comments

... has a codebase with a long source history
maintained by a large development team
with stable Y-O-Y commits

... took an estimated 24 years of effort (COCOMO model)
starting with its first commit in February, 2012
ending with its most recent commit about 1 month ago



http://rucio.cern.ch/



ATLAS DDM: The Scale

The ATLAS DDM System has demonstrated very large scale data management

Total:

- 1B file replicas
- 230 PB on 130 sites

Transfers:

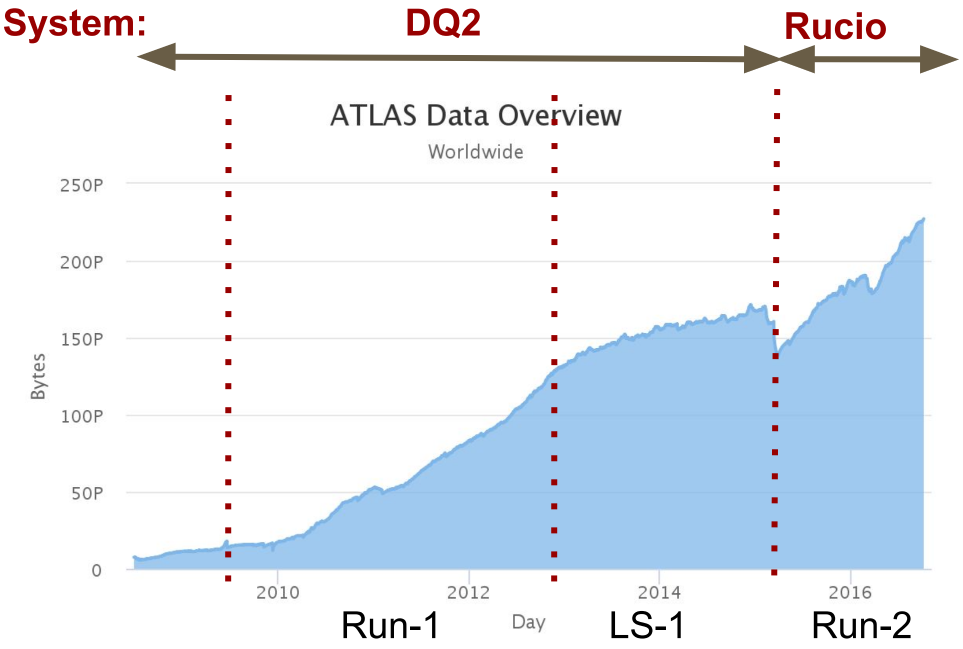
- 40M files/Month
- 40 PB/Month

Download:

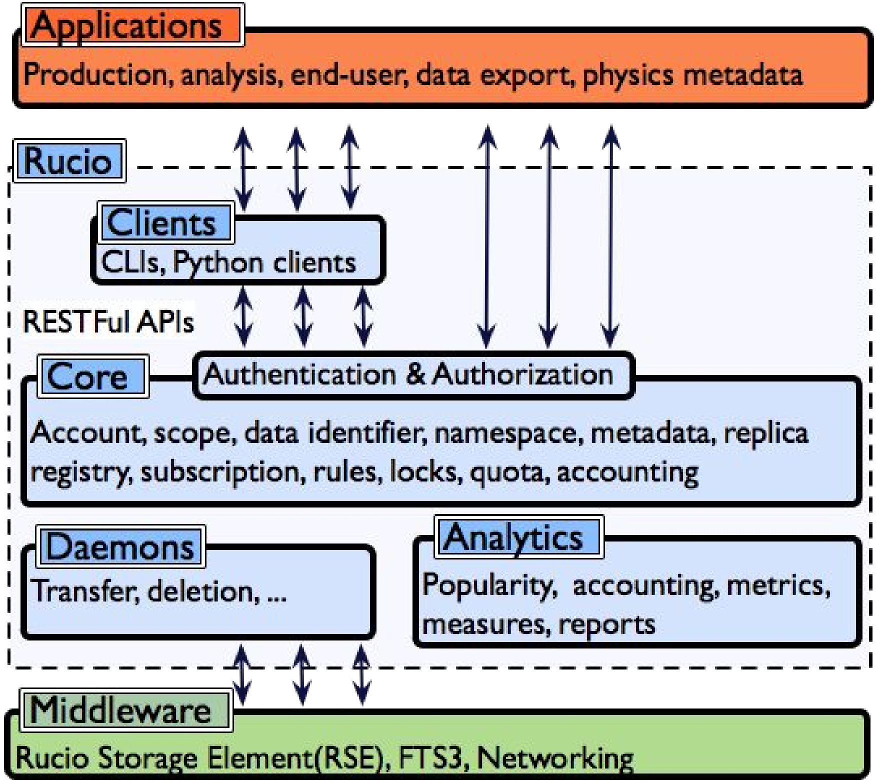
- 150 M files/Month
- 50 PB/Month

Deletion:

- 100M files/Month
- 40 PB /Month



Rucio - SW Stack Overview


Applications
Production, analysis, end-user, data export, physics metadata
Rucio
Clients
CLIs, Python clients
RESTFul APIs
Core
Authentication & Authorization
Account, scope, data identifier, namespace, metadata, replica registry, subscription, rules, locks, quota, accounting
Daemons
Transfer, deletion, ...
Analytics
Popularity, accounting, metrics, measures, reports
Middleware
Rucio Storage Element(RSE), FTS3, Networking

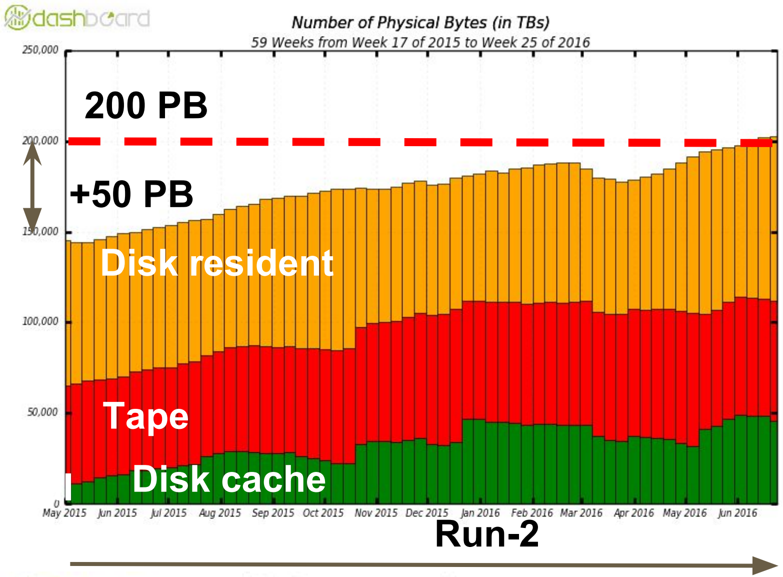
Open and standard technologies:

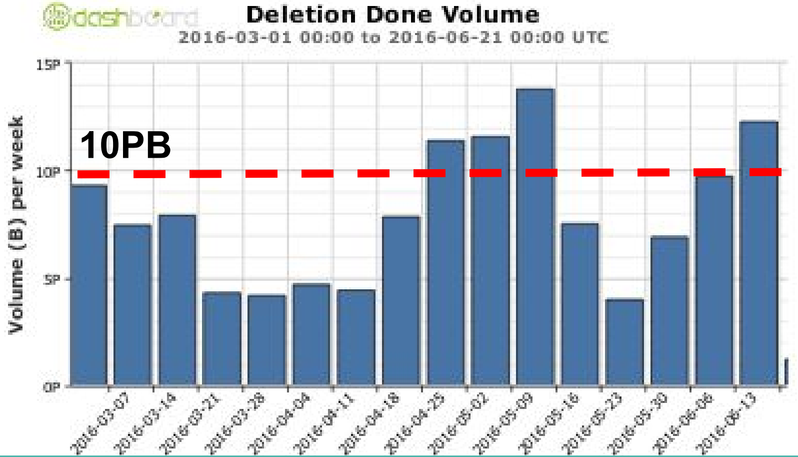
- WSGI server
- Caching
- Token-based authentication
- New middleware capabilities

Run-1 vs. Run-2

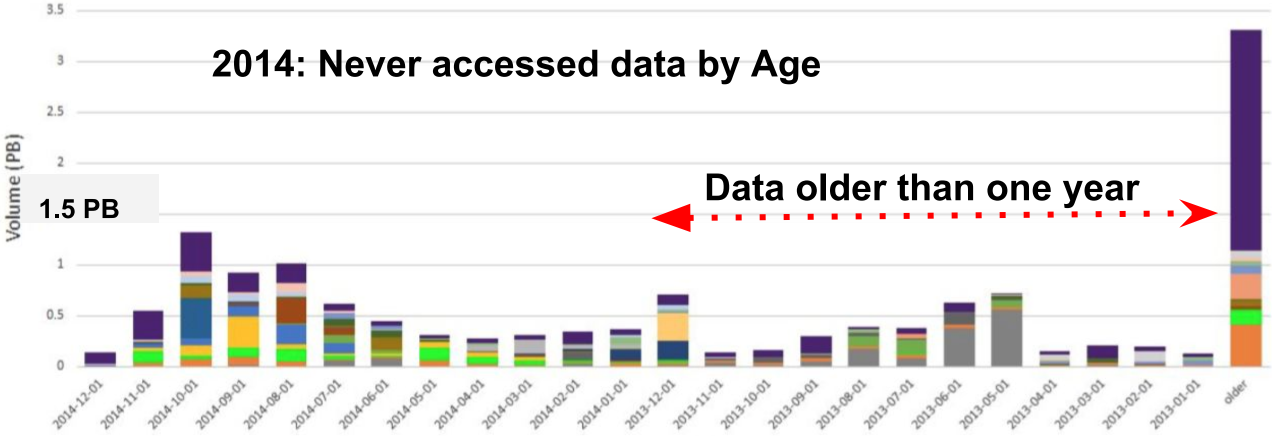
Rucio is scalable, robust and reliable.
It keeps up nicely with the load increase:

- Transfer
  - 2M transfers/day
  - Equivalent to Run-1 but bigger files with peak at 40GB/s
  - More load/hotspots on the network
- Deletion
  - 8M deleted files/day
  - Factor 4 increase since Run-1
  - Much more pressure on disk space





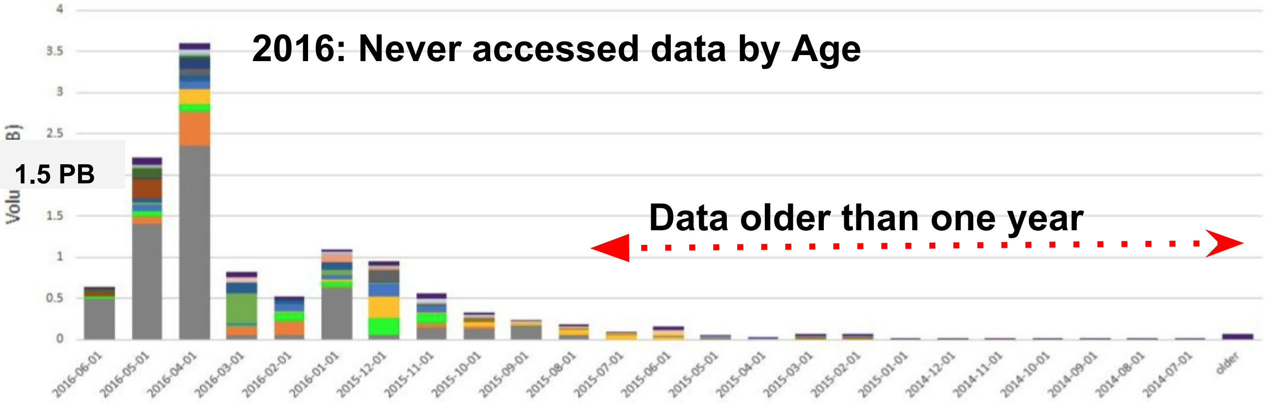
Disk Usage: Improvements



2014: Never accessed data by Age

1.5 PB

Data older than one year



2016: Never accessed data by Age

1.5 PB

Data older than one year

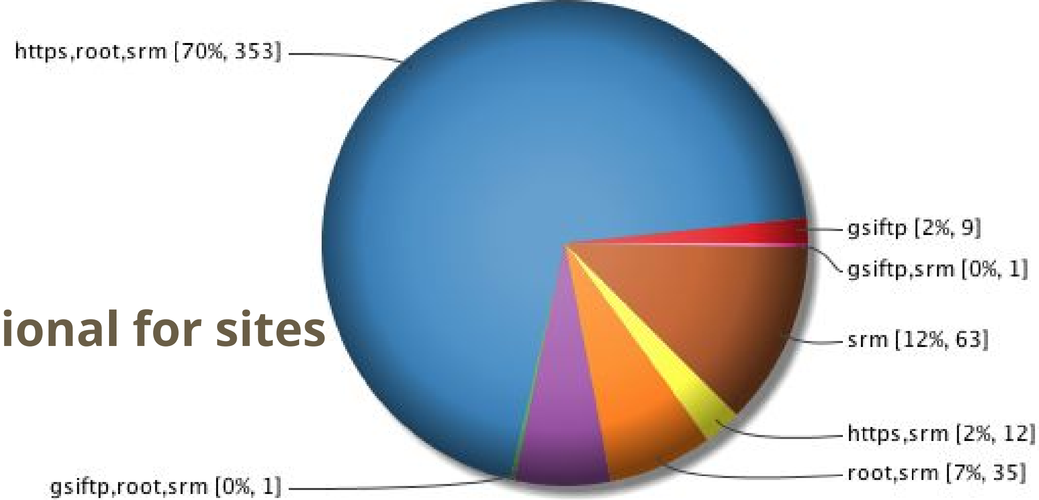
Thanks to:

- Better space monitoring
- Lifetime model
- ATLAS Policies/actions
- strategies to keep recent and popular data on disks (LRU deletion) and to avoid data duplication

More automation in place, Cf. :

- Rucio Auditor - Consistency in the ATLAS Distributed Data Management System
- C3PO - A Dynamic Data Placement Agent for ATLAS Distributed Data Management

SRM Alternatives



- **Achieved goal: Make SRM optional for sites**
  - Caveat: Not for tapes !
  - We now have sites without SRM !
- DDM/Rucio supports natively multiple protocols
  - But requires some work to support them: FTS, plugin, swift, etc
- We proposed alternatives for all SRM functionalities
  - E.g., gsiftp/xroot for third party copy, space reporting with a JSON file
- We are gradually moving to SRM alternatives
  - Deletion, upload/download, third party copy

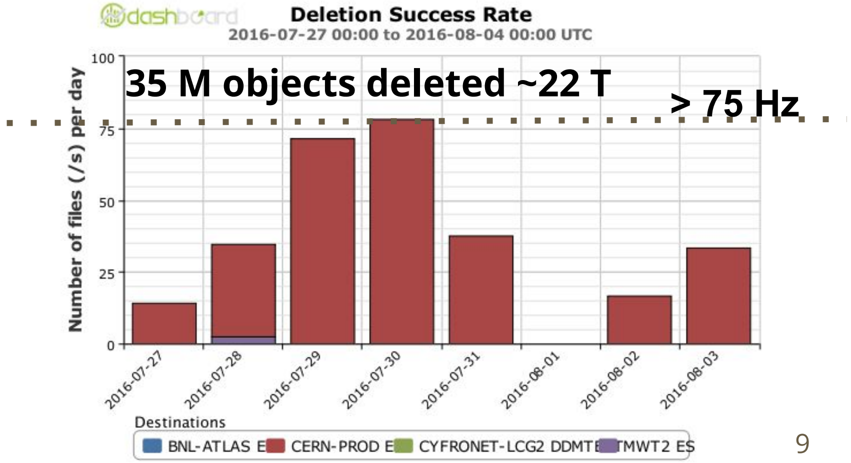
Object Store Support

DDM can use objectstore as a standard Storage endpoint

- BNL (Ceph), Lancaster (Ceph), RAL (Ceph), CERN (Ceph), MTW2(Ceph)

Two use cases are supported in production:

- Log files: Upload/Download are transparently supported in the rucio clients
- ATLAS Event Service (AES): deletion
  - 300k events deleted per day [Link]

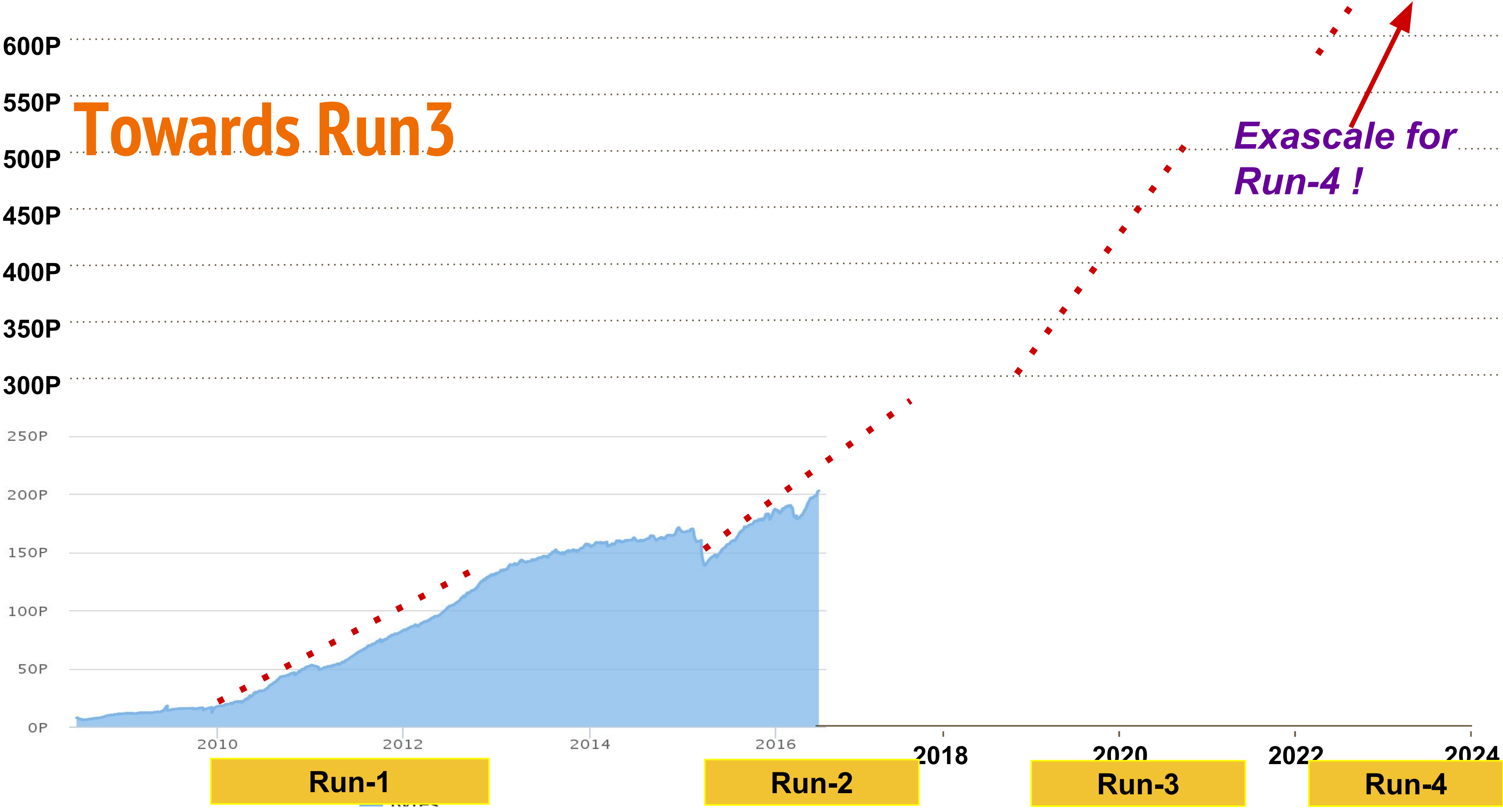


35 M objects deleted ~22 T

> 75 Hz

Cf. Object-based storage integration within the ATLAS DDM system

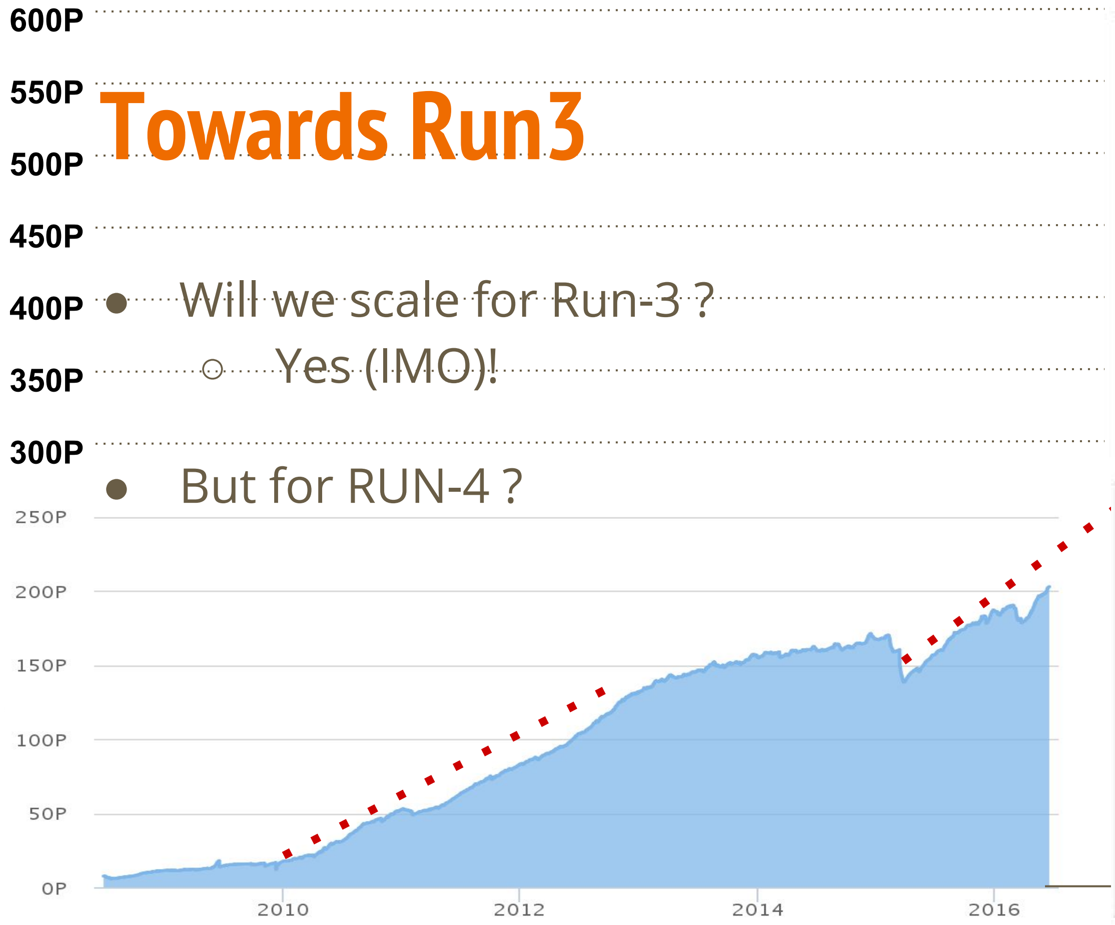
Towards Run3



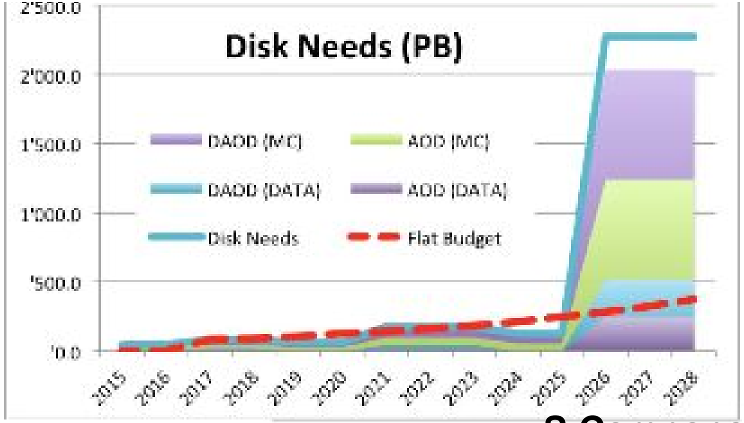
Exascale for Run-4 !

Towards Run3

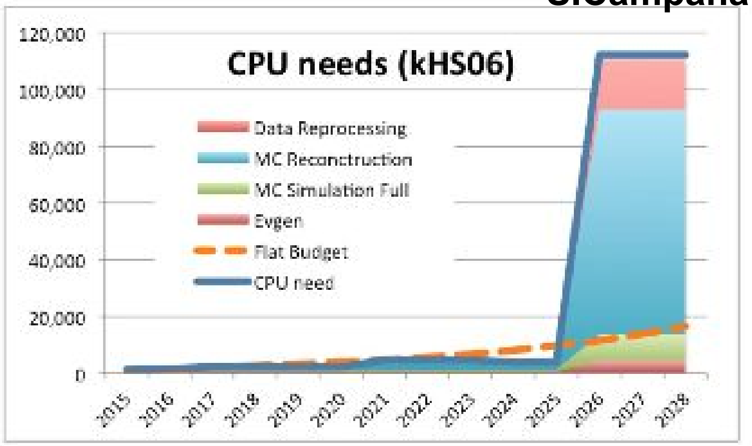
- Will we scale for Run-3 ?
  - Yes (IMO)!
- But for RUN-4 ?



Initial studies of computing for HL-LHC



S.Campana



Run-1

Run-2

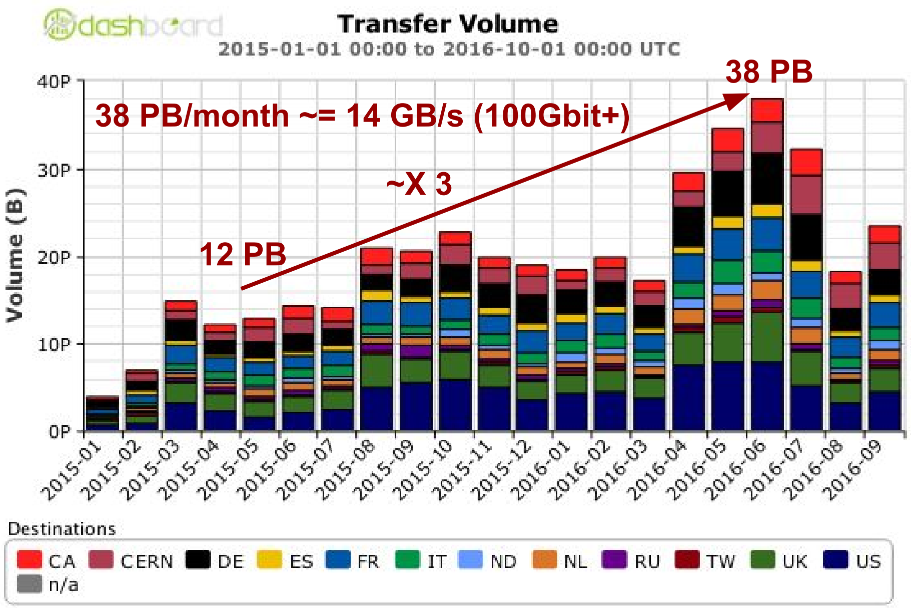
Run-3

Run-4

Network Evolution

- We’ll be more and more reliant on our foundation of the network
- It’s coherent with our approach to use it more and more
  - E.g., remote i/o
- By 2020, 800 Gbps waves should be possible but not from everywhere..

Network Use in ATLAS



38 PB

38 PB/month ~= 14 GB/s (100Gbit+)

~X 3

12 PB

- Cf. Using machine learning algorithms to forecast network and system load metrics for ATLAS Distributed Computing

Summary

- DDM is in good shape
  - It has been operating robustly, stably and effectively since beginning of 2016
  - We are safe with Run-2 data taking
  - ATLAS is using all the available resources at full scale
  - We need to keep an eye on disk spaces
- The target keeps moving with challenging development work ahead
- Evolution or Revolution for Run-4?
  - We need to gain one order of magnitude in computing capability !
- R&D planning
  - We will do it collaboratively with others (WLCG, HEP Software Foundation, community white paper, cross experiment working groups)

ATLAS DDM: CHEP Contributions

Rucio WebUI - The Web Interface for the ATLAS Distributed Data Management

C3PO - A Dynamic Data Placement Agent for ATLAS Distributed Data Management

Object-based storage integration within the ATLAS DDM system

Rucio Auditor - Consistency in the ATLAS Distributed Data Management System

Using machine learning algorithms to forecast network and system load metrics for ATLAS Distributed Computing