RUCIO

Alessandro Di Girolamo

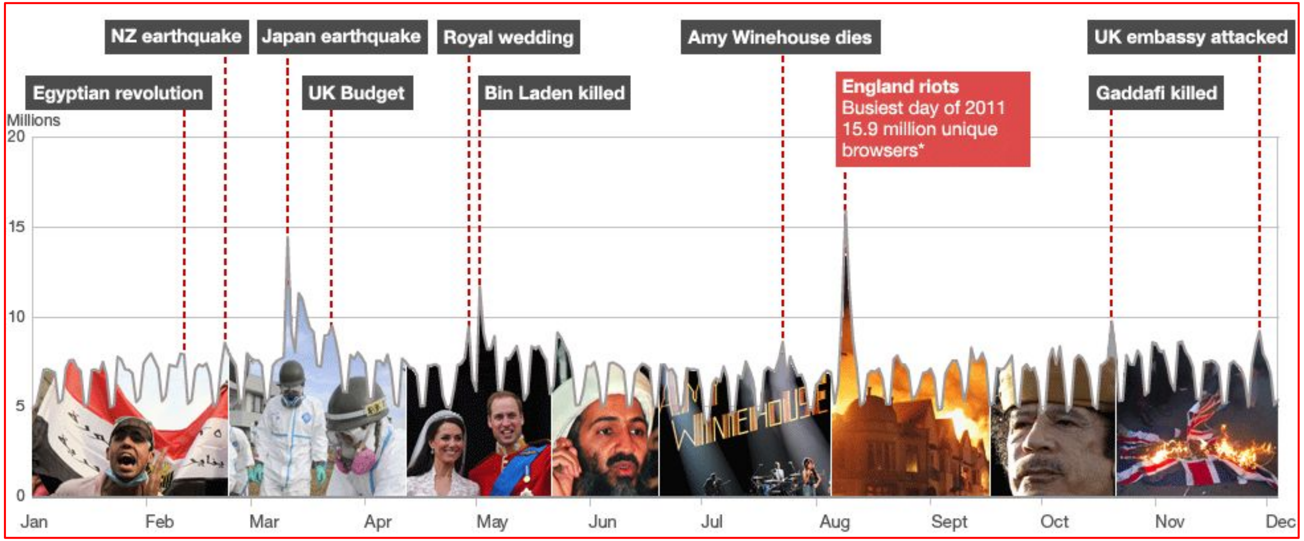
CERN IT (& ATLAS)

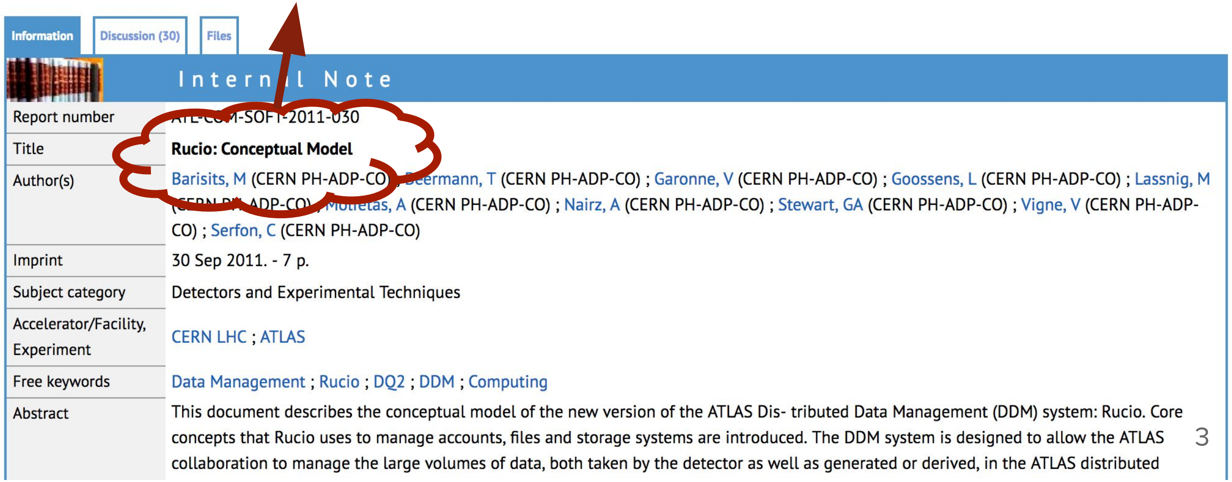
disclaimer

this talk is not about…

the technical details of Rucio

2011





Internal Note

Report number	ATL-COM-SOFT-2011-030
Title	**Rucio: Conceptual Model**
Author(s)	Barisits, M (CERN PH-ADP-CO) ; Beermann, T (CERN PH-ADP-CO) ; Garonne, V (CERN PH-ADP-CO) ; Goossens, L (CERN PH-ADP-CO) ; Lassnig, M (CERN PH-ADP-CO) ; Molfetas, A (CERN PH-ADP-CO) ; Nairz, A (CERN PH-ADP-CO) ; Stewart, GA (CERN PH-ADP-CO) ; Vigne, V (CERN PH-ADP-CO) ; Serfon, C (CERN PH-ADP-CO)
Imprint	30 Sep 2011. - 7 p.
Subject category	Detectors and Experimental Techniques
Accelerator/Facility, Experiment	CERN LHC ; ATLAS
Free keywords	Data Management ; Rucio ; DQ2 ; DDM ; Computing
Abstract	This document describes the conceptual model of the new version of the ATLAS Dis- tributed Data Management (DDM) system: Rucio. Core concepts that Rucio uses to manage accounts, files and storage systems are introduced. The DDM system is designed to allow the ATLAS collaboration to manage the large volumes of data, both taken by the detector as well as generated or derived, in the ATLAS distributed

... a bit of history

- From end 2011 till end 2013 the Rucio team gathered requirements, architected and developed Rucio.
  - Functional tests and commissioning of the various components started in 2013.
- From end 2013 till end 2014:
  - Stress tests, commissioning
  - Migration from DQ2 to Rucio
- From December 2014:
  - Rucio

(in Italy a Donkey is a stubborn slow animal... but it get things done!

*and it took me 3 years to learn how to say **Russsio**)*

Get a nice toy: now ride it!

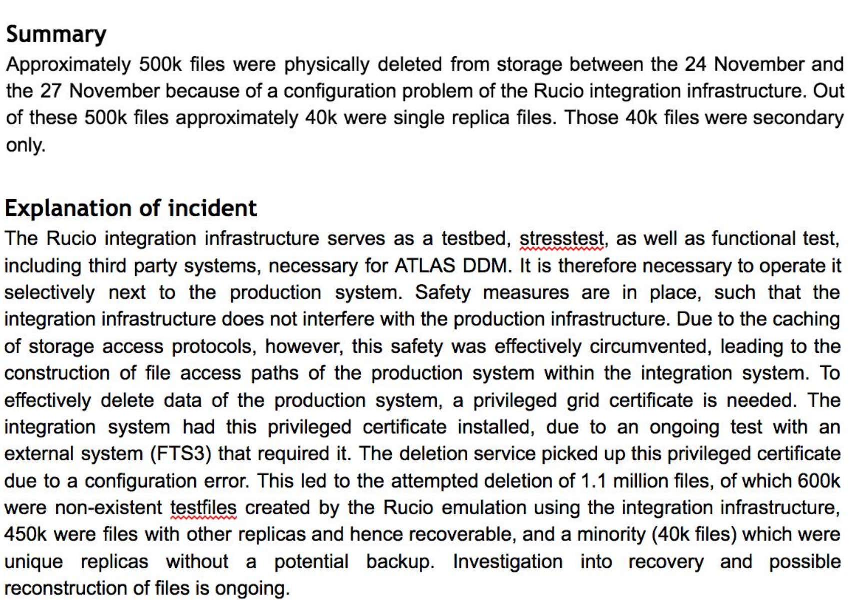
- Difference between driver(s) and pilot(s)
    - Improve the machine you're riding working hard with the engineers and the mechanics!
    - If you're a good pilot, you can make running also a Donkey

From theory to reality

- In Dec 2014 Rucio was (still) a newborn framework
  - Suffering of the problems of a completely new framework
- Part of it was too naive, too simplified, or not scaling
  - Despite the many variagate tests we run!
  - E.g. conveyor
- Some needed (old) features not present
  - E.g. Replica management at DDM Endpoint level
  - Really need them? not always clear which are the features you will be missing from one framework to another!
- The Rucio team was not Ops oriented (enough)
- In general it was all shiny, but just barely ready to walk, not really to rock!!

... ooopsss

- November 2014:
  - Just few days before the final migration



File Deletion: Incident Report

Summary

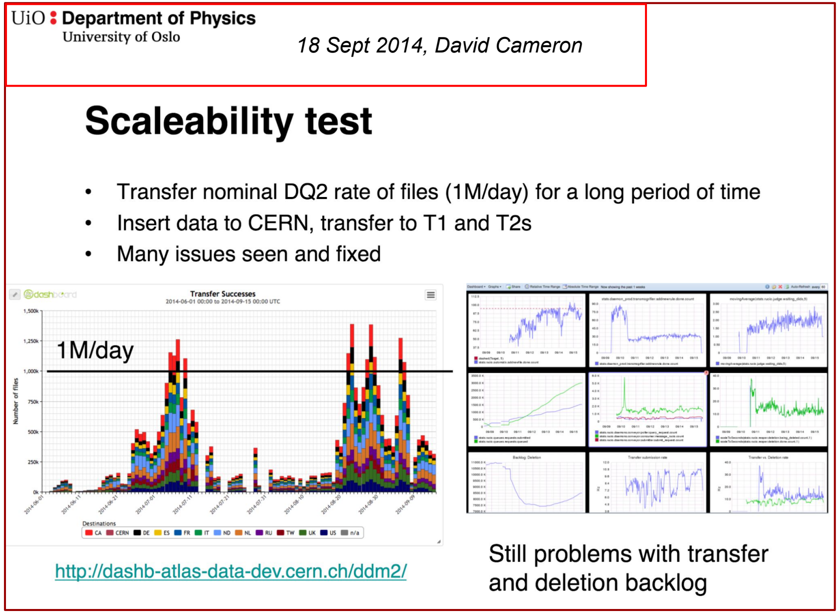
Approximately 500k files were physically deleted from storage between the 24 November and the 27 November because of a configuration problem of the Rucio integration infrastructure. Out of these 500k files approximately 40k were single replica files. Those 40k files were secondary only.

Explanation of incident

The Rucio integration infrastructure serves as a testbed, stresstest, as well as functional test, including third party systems, necessary for ATLAS DDM. It is therefore necessary to operate it selectively next to the production system. Safety measures are in place, such that the integration infrastructure does not interfere with the production infrastructure. Due to the caching of storage access protocols, however, this safety was effectively circumvented, leading to the construction of file access paths of the production system within the integration system. To effectively delete data of the production system, a privileged grid certificate is needed. The integration system had this privileged certificate installed, due to an ongoing test with an external system (FTS3) that required it. The deletion service picked up this privileged certificate due to a configuration error. This led to the attempted deletion of 1.1 million files, of which 600k were non-existent testfiles created by the Rucio emulation using the integration infrastructure, 450k were files with other replicas and hence recoverable, and a minority (40k files) which were unique replicas without a potential backup. Investigation into recovery and possible reconstruction of files is ongoing.

Changing gear…

- DevOps weekly meetings:
  - From 21 Jan 2014 till 2 days ago: 154 pages of minutes!


Transfer Successes
2014-06-01 00:00 to 2014-09-15 00:00 UTC
1M/day
Number of files
Destinations
CA CERN DE ES FR IT ND NL RU TW UK US n/a

UiO Department of Physics
University of Oslo

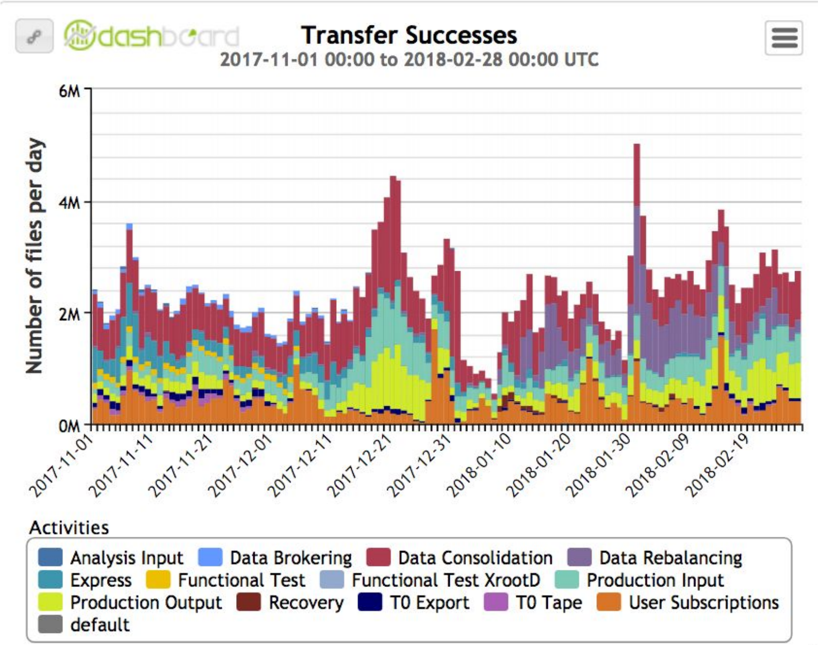
18 Sept 2014, David Cameron

Scaleability test

- Transfer nominal DQ2 rate of files (1M/day) for a long period of time
- Insert data to CERN, transfer to T1 and T2s
- Many issues seen and fixed

http://dashb-atlas-data-dev.cern.ch/ddm2/

Still problems with transfer and deletion backlog



on top of Rucio

- Replication policy to feed into subscriptions
- Storage quota/management policy for users/groups
- Storage dumps
- Deletion policy (watermark, greedy/non-greedy)
- Interaction with information system(s) (for ATLAS: AGIS, VOMS, REBUS, storage systems)
  - Static: RSEs definition, user information
  - Dynamic: storage space, downtime blacklisting
- **Most of the above is handled by external probes interacting with external systems and Rucio**
- Monitoring (in addition to Rucio UI):
  - Transfers (ATLAS DDM dashboard)
  - Accounting reports (ATLAS DDM accounting dashboard)
  - Storage space monitoring

One example: Rucio & Panda

- PanDA is the ATLAS workflow management system
  - Instructs Rucio to transfer jobs inputs and outputs data to the right place before and after jobs run
    - Temporary small (~10 files, overlapping) datasets are used for intermediate data locations
    - When input transfers complete Rucio sends a callback to PanDA so it knows to release the jobs
    - When the temporary datasets expire the data is left as cache (*secondary*) replicas
      - which could potentially be used by future jobs
  - Running jobs copy data from local storage using Rucio to lookup replica locations
  - Pilot is moving to use Rucio Movers, i.e. Rucio clients to manage(read/write) data.
    - Opening the door for more intelligence, e.g. smart source selection, usage of caches.

Not all GOLD, not all perfect

- a HUGE room for improvement
  - At many levels, in various part of it
  - … be more intelligent (what does intelligent mean??)
- BUT:
  - Rucio is flexible, allow external plugin to cleanly interact with it, and eventually, if needed/wanted, can be integrated into.
  - You can build your own intelligence

Challenge for you (for your hackathon tomorrow)!

- Develop a better than the present source selection algorithm for third party transfers!
- N.b.: FTS is also developing improved SRC selection algo, check their code, work with them, the most we are able to exploit with FTS the better it is!
- Tip: not so easy. Activities share, throughput GB/s of link, nfiles/hour, total SRC or DEST load, total queue …

Conclusions?

- Not really:
    - just the beginning
- 2 years of architecting and development, + 2 years of commissioning for ATLAS, turned Rucio into a mature product for ATLAS and for future users
    - I would not wish to my worst enemy a similar experience!
    - But I have to admit that I liked it, because eating 1Kg of salt together made us a real team
- Rucio is working:
    - flexible, solid, a lot of ancillary things all around it make it good for use it as it is, but also allow evolution

… and now?

- Would be a pity if other collaborations would not join the Rucio community:
  - Re-use, exploit the experience already gained!
  - Join the project, making it also their own!
  - Making it even more flexible and intelligent.
  - … believe me, there is space for everybody!

Underestimating the amount of work to do a working Data Management System could be lethal!…