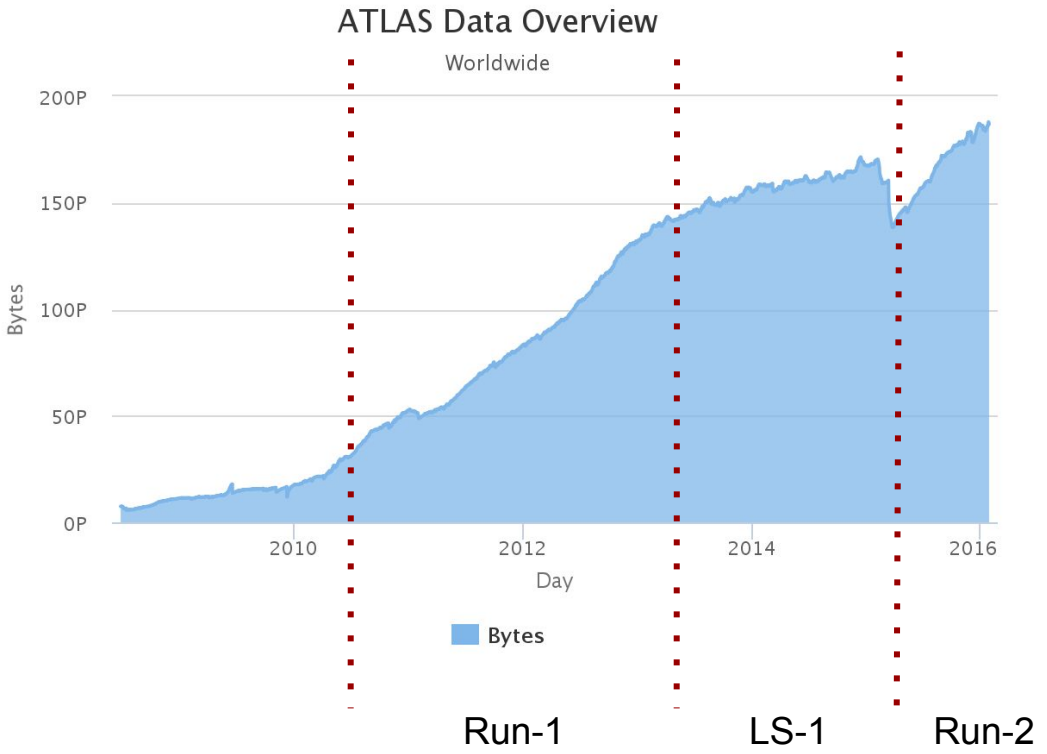
ATLAS Data Management: Status & Evolution

Vincent Garonne & DDM

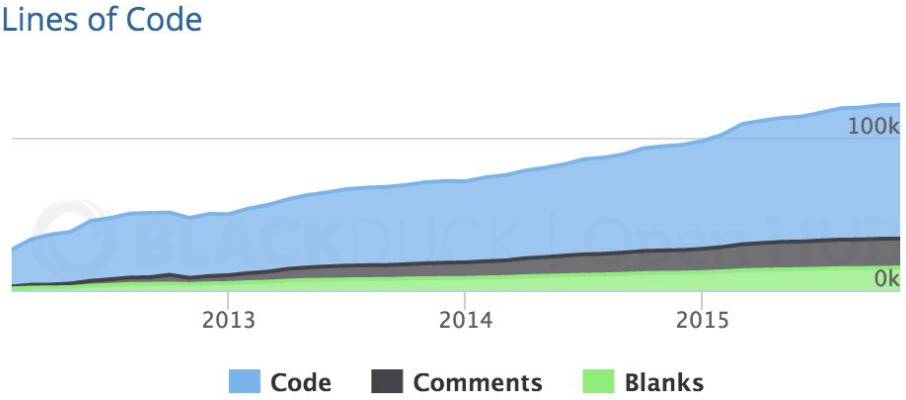
ATLAS DDM Status

After one year in production, The ATLAS DDM System/Rucio has demonstrated very large scale data management:

- Almost 200 PB on 130 sites
- 1B file replicas
- 40M file transfers/Month
- 20 PB data transfer/Month
- 100M deleted files/Month
- 30 PB deleted data /Month



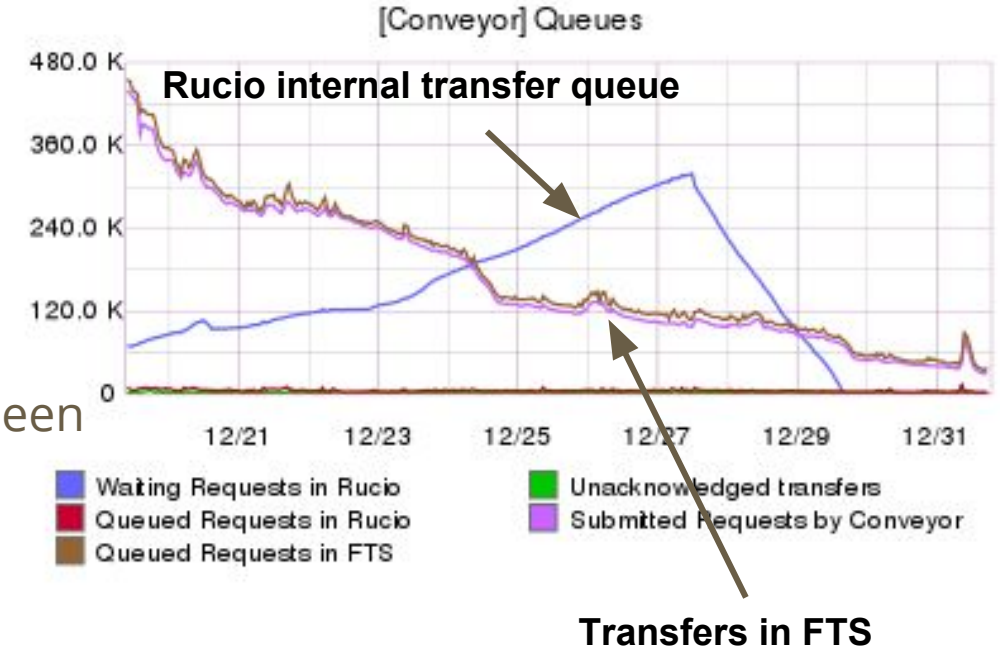
Data Taking: Experience



- During data taking, we focused on scalability, performance and automation
  - In less than one year, we went from release 0.3.* to 1.3.* : 30 releases !
  - The last rucio release was the culmination of 50 months of work with almost 3,300 commits done by 26 contributors (8 institutes) !
- New components have been successfully deployed in production and new ones are coming soon
    - E.g,, Rucio WebUI, R2D2, consistency (Cf. Fernando's talk), cache, etc
- One of the biggest challenge has been to offer new features, support new use cases and fix critical bugs while offering an high availability service
  - Festina lente !

Data Taking: Observations

- Data volume increases a lot faster than the available capacity !
- We have to deal with bursty load and transient data: reprocessing, obsolescence campaign, lifetime model, data rebalancing activity, etc.
- This bursty load put a lot of stress on the underlying resources, like storage or FTS, which can break or misbehave
- The System's architecture has been flexible enough to offer some protection for such situation
  - E.g., delay, timeout, retry, internal queues

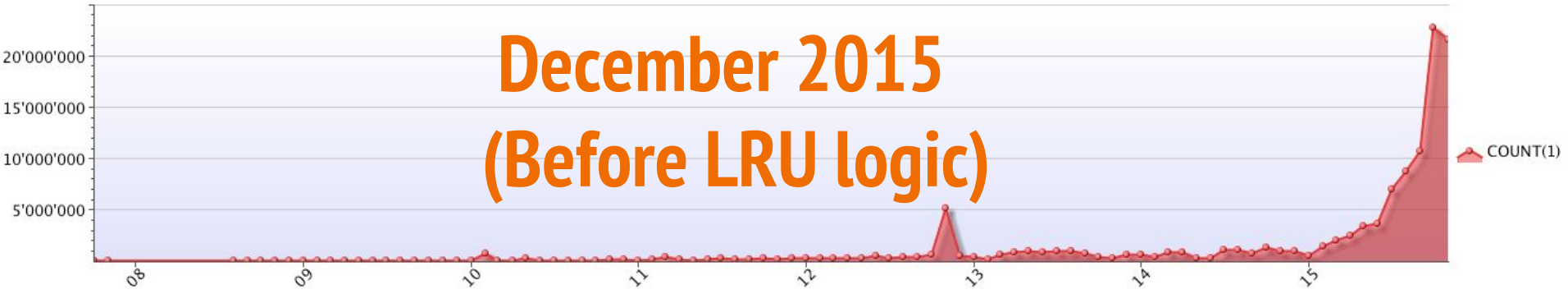


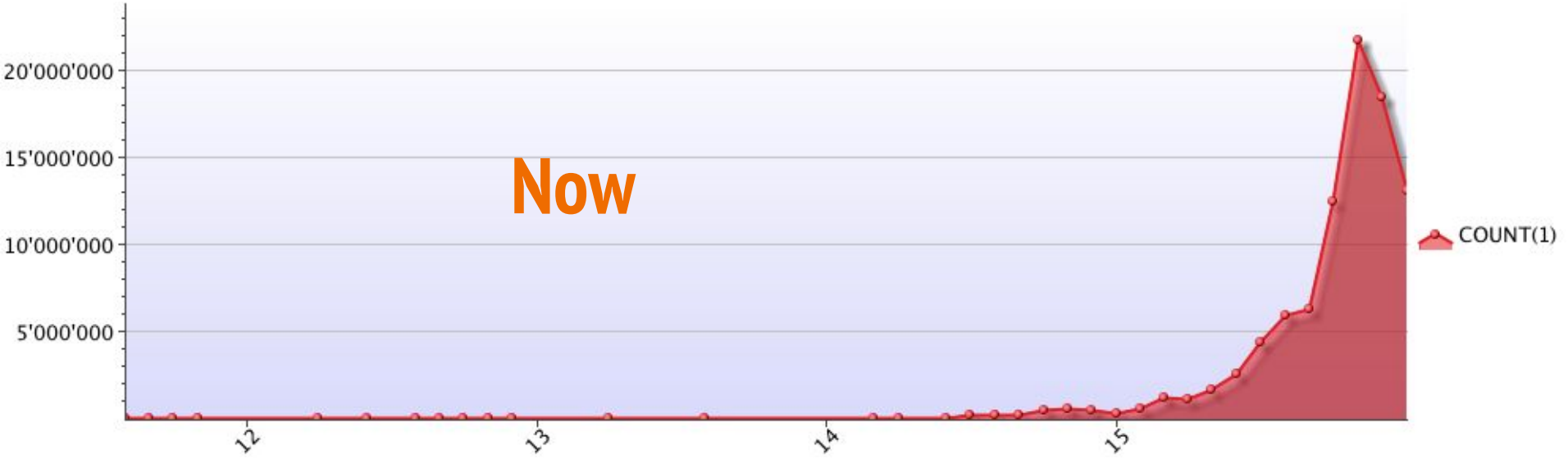
Optimization & Tunings: Deletion Policy

- We have an unbalanced usage of certain sites, mainly Tiers-2
- We are working now on fine and complex tuning of the system
- For example, we tuned the deletion policy to have a cache LRU logic for secondary data
- The goal is to keep as much as possible ‘interesting’ secondary data on disk
  - Interesting data: recently created and/or recently touched data
  - Less interesting data: Old and unused data

Years

Secondary data on disk (number of files) ordered by creation date or last access date for all sites





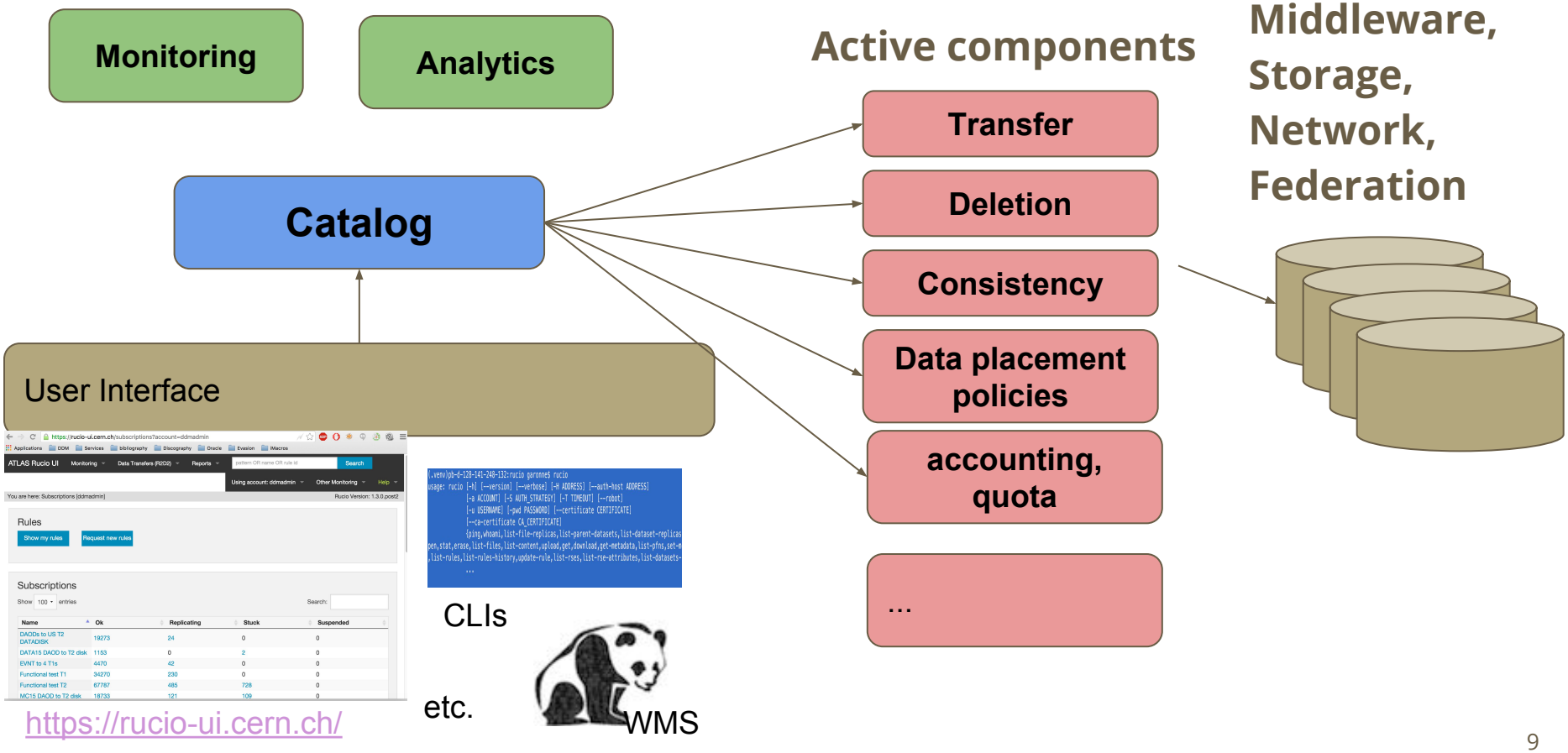
Optimization & Tunings: Replication Policy

- We are also working on replicating automatically ‘interesting’ data within the C3PO project
  - C3PO is the pd2p successor
- The main idea is to replicate ‘interesting' data on ‘unused' sites to offload busy sites, avoid hotspots and speed-up the job response time
  - Interesting data: Data which will be used by jobs or possibly popular
  - First 'dry run’ model in place (T.Beermann)
- This requires to collect performance data, analyze it, use analytical tools and techniques like prediction
  - Tight integration with other ADC projects, e.g., PanDA, Agis and Network aware brokerage
  - Cf. Tadashi’s Talk
  - Cf. Analytics’ Talk

Other Next Steps

- The next focus is to leverage Rucio's new features for ATLAS computing
- Few examples
  - Distributed datasets to move less inputs
  - Consolidation of group tape endpoints
  - Localgroupdisk management (Cf. Martin & Thomas talk)
  - SRM-less site (Cf. Cedric's talk)
  - Rucio Cache (Cf. Cedric's talk)
  - Object Stores (Cf. Wen's talk)

DDM Evolution: Current Logical Overview



Future of Catalog ?

- Most of DDM implementations for the LHC are based on catalog
  - It's convenient to have one global and fast index for job scheduling
  - Easier to manage, few misses and availability > 99
  - Flexible design with no dependence on particular implementation
  - Follow the advances in databases, open and standard technologies
- DDM has to scale with the (cumulative) number of data objects and operations
  - Data object can be event(s), file(s), dataset(s), containers(s)
  - Horizontal scalability as a strong requirement
  - Today: 2.5 M transferred files/Day , Tomorrow: ? a factor 10 ?
  - Complement RDBMS database with key-value stores ?
- I have observed a general trend to have more and more (physics) metadata in DDM to facilitate data selection, discovery and analytics
  - DDM and the metadata part strongly coupled ? Yes (IMO)

Storage, (Regional) Federation & Middleware

- Integration of new storage types is a constant need
  - New protocols
  - Authentication mechanisms
  - Objects store, volatile storage Cloud/HPC to increase the total storage capacity, etc.
- DDM systems are complemented with storage federations (e.g., for failover, storageless computing sites, etc.)
  - Ideally we want to a (regional) federation as one DDM end-point
  - More details in the Federation and cache talk
- The biggest predictable gain will come from network (X200) and will strongly influence the experiments computing models
  - Cf. Network's talk

Summary

- Rucio is now fully in production for ATLAS since 1st December 2014
- The performance meets the expectations and Rucio is at a much larger scale than DQ2
- The focus now is on using the new features and optimizing the system while working on the long and medium term evolution