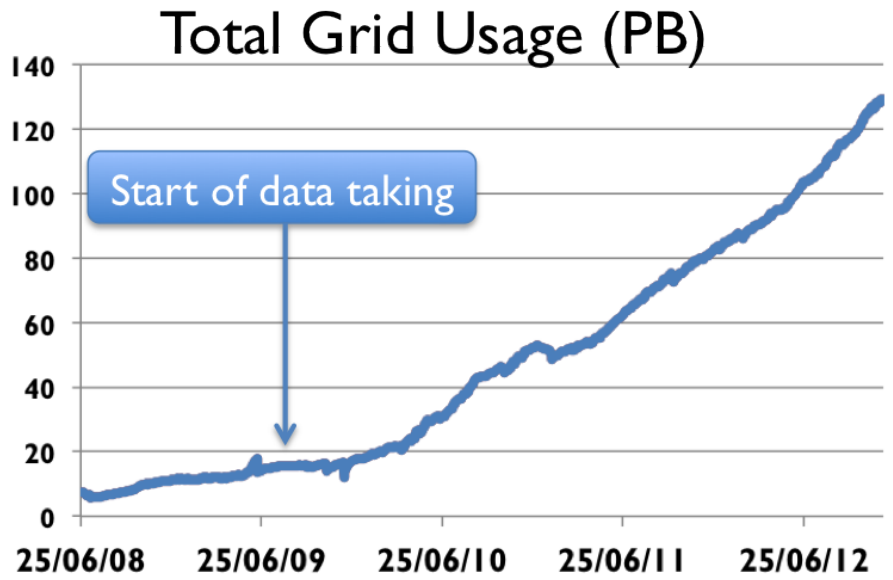
Rucio

Vincent Garonne, CERN

ATLAS Jamboree, Geneva, December 2012

Background

- The Distributed Data Management project manages ATLAS data on the grid
- The current system is Don Quijote 2 (DQ2)
  - 130 Petabytes
  - 600k datasets
  - 355 million files
  - 800 active users
  - 130 sites



- DQ2 works, but ...
  - Scaling problems, heavy operational burden and difficulties to add new features and technologies

The Next Version – Rucio

- Rucio is an evolution from DQ2 designed to ensure system scalability, reduce operational overhead and support new ATLAS use cases
  - The concepts are described in the Rucio Conceptual Model(v2) document [CDS Link]
  - The pilot service has been delivered in November 2012
- The target deployment and the decommissioning of DQ2 are scheduled for 2014 after the "Long Shutdown 1" (LS1)
- A plan for preliminary changes in DQ2 has been defined to facilitate the final migration

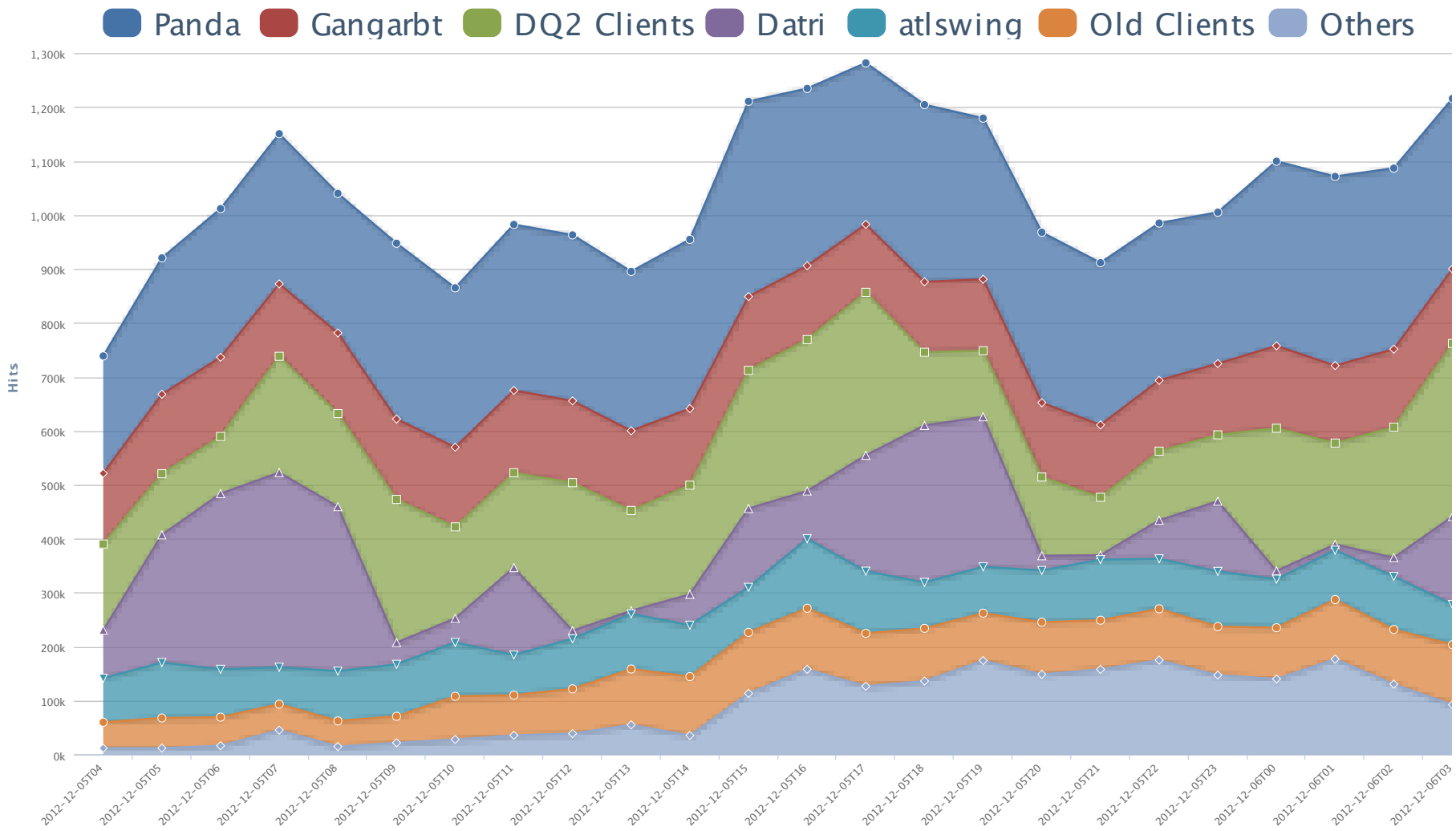
Accounts

- Rucio supports user, group and service accounts
  - Better management of users, physics groups, ATLAS activities, data ownership, permission, quota, etc.
- Lightweight and scalable token based authentication system which supports many types of credentials (X509, Kerberos, etc.) for read&write operations
- ATLAS grid users need to have a CERN account
  - The mapping {grid nickname - CERN ATLAS AFS/LXPLUS account} has been recently enforced
- Do site administrators need to access DDM ?
  - Who ? What ? Why ? How ?

Use Cases Collection

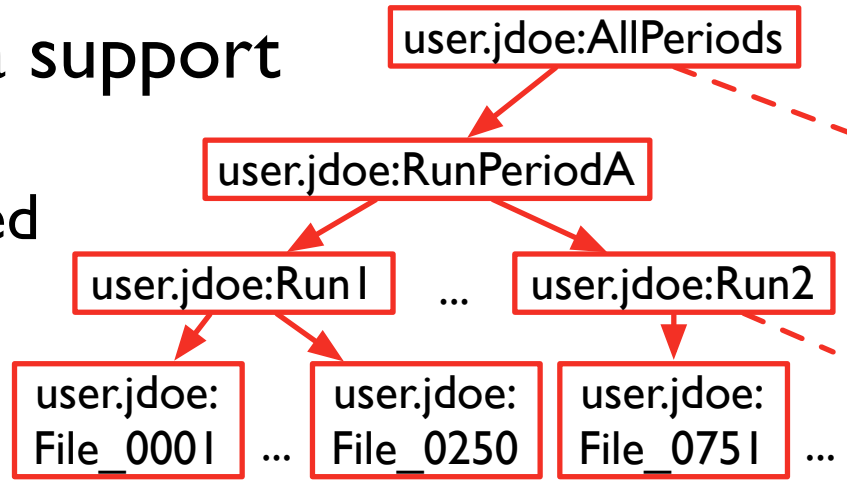
- The DQ2 load is extracted, mapped to use cases, and transformed into a Rucio load
  - Functional testing and performance evaluation of Rucio
  - Gradual migration of external applications (e.g., PanDA)
- Latest stable DQ2Clients (2.3.0) introduce Rucio accounts
  - Monitoring infrastructure based on Hadoop has been established to analyse central services traffic
- All sites must be upgraded to the latest stable
  - Automatic with CVMFS
  - Old clients will be blocked

Hits Per Application



Rucio Namespace

- Data hierarchy with metadata support
  - Files are grouped into datasets
  - Datasets/Containers are grouped in containers



- Files, datasets and containers are identified by <scope:name>
- The scope partitions and isolates the namespace into several sub-spaces, e.g.,

User : user.jdoe:004406.EXT0.00011.root
Group : group.phys-higgs:08.physicsD3PDSlimmed.root
Detector: data11_7TeV:AOD.491965._0042.pool.root.1

Replica Management

- Rucio Storage Element (RSE) uniquely identifies storage space with attributes
  - Name, supported protocols, QoS, space properties, etc.
- RSEs can be grouped in many logical ways by tagging, e.g., CLOUD=UK and Tier=1
- Accounts manage their data with replication rules defined on data identifiers and a list of RSEs
  - It gives the minimum number of replicas on the grid
  - e.g., User jdoe wants 2 copies of jdoe:dataset1 on cloud=UK and USERDISK
- Accounts are only charged for files on which they have set replication rules
  - Number of replicas requested, not physically existing

Storage Cache

- A cache is an RSE, tagged as volatile, for which Rucio doesn't control all file movements
  - e.g., Storage service keeping additional copies of files to reduce response time and bandwidth usage
- The application populating the cache must register and unregister file replicas in Rucio
  - The replica location on volatile RSEs can have a lifetime
  - Replicas on volatile RSEs are excluded from the Rucio replica management system
  - Explicit transfer requests can be made in order to populate the cache

Storage Interfaces

- Rucio Storage Element wrapper
  - High-level user abstraction
  - cf., Mario's talk
- Deterministic mapping between the logical file name and its path name in a scoped namespace to remove/decrease external file catalog lookups
  - End of use of the LCG File Catalog in 2014
  - e.g. mapping: <Scope>/??/??/<File name>
  - cf., Cedric's talk
- Plug-in interface using standard remote data access and control protocols
  - In addition to SRM

SRM Usage & Use Cases

ATLAS plans to migrate to a SRMless world ...

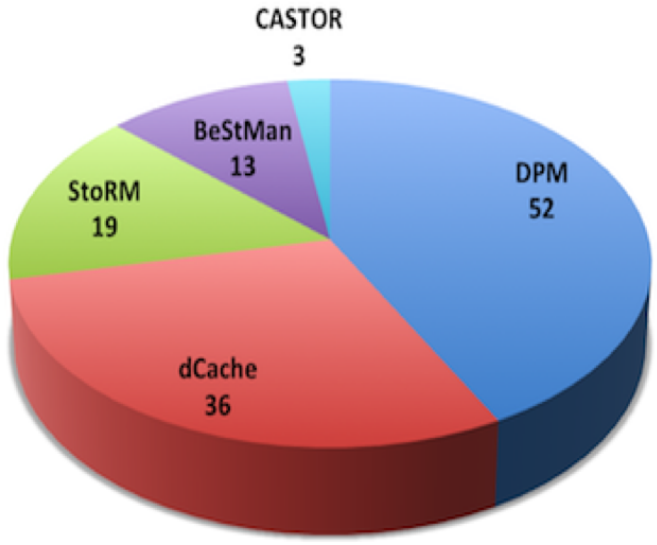
DDM Use cases	Clients
Copy	`lcg-cp (lcg-util)`
Redirection	`lcg-getturls (lcg_util)`
Third party transfer	`glite-transfer-* (FTS)`
Deletion	`gfal_deletesurls (gFal)`
Staging	`gfal_prestage* (gFal)`
Accounting[1]	`lcg_stmd (lcg_util)`
Renaming[2]	XX

[1] ... and consistency

[2] Functionality needed for the Rucio migration

Alternatives: DAV & xroot

- Copy, redirection, deletion and renaming use cases are possible with:
  - `WebDAV://` – Open source clients, e.g., wget, aria2c, etc.
    - Particularly relevant for the `dq2-get` use case
  - `xroot://` – e.g., Interactive data access from jobs
- Both protocols are supported by 85% of the sites
- The central Rucio migration infrastructure requires protocols that allow renaming
  - The migration will commission them
  - Load balanced front-ends should be published in BDII or AGIS



ATLAS Storage by type

SRM Space Tokens ?

- Can we get rid of them ? accounting ? ACLs ?
- Example of data organization on disks with Rucio

```
> ls -R rucio
rucio/data12_8TeV
rucio/group/...
rucio/group/perf-tau/
rucio/mc11_7TeV/
rucio/user/...
rucio/user/jdoe/
```

- ACLs should be defined at the scope directory level
- Online accounting needs for the root directory path
  - e.g., Results of an incremental `du` executed every 30 min. on `/rucio` in `/atlas/rucio/info/space-info.json`

Accounting & Consistency

- Fine grained accounting can be achieved with dumps
  - i.e., publication of a daily dump with everything under `/rucio` in `/atlas/rucio/info/namespace.csv`
  - Dumps can be collected remotely and map-reduced
- This also covers the consistency check use case to detect dark data
  - Data on disks but not in catalog (and vice-versa)
  - e.g., crosscheck of the content of `rucio/data12_8TeV` against the Rucio catalog in case of accounting numbers mismatch

Life Without SRM

- Third party transfers are possible with FTS and gridftp
  - Load balanced front-ends should be published in BDII or AGIS
  - More alternatives with FTS3
- Only one use case remaining: Tape recall
  - bringOnline and cache pinning
  - It concerns Tier0 and Tier1 sites
- With the proposed mechanisms, SRM can be dropped for disk sites

http://rucio.cern.ch