Tier 3 Data Management, Tier 3 Rucio Caches

Doug Benjamin
Duke University

Current situation

Using information from AGIS

- SRM based sites – 9 sites
  - ANLASC (HPC frontend), Bellarmine-T3, IllinoisHEP, Lucille, NERSC (local groupdisk only), OUHEP, SMU (local group disk only), Penn (local group disk only), Wisc (localgroup disk only)
- Gridftp only endpoints – 8 +1 sites
  - ANLASC (for ANL Tier 3), Duke, Indiana (online?), Nevis, NYU (still online?), Stony Brook, UC Santa Cruz (not on line), UT Dallas, (SLAC has a gridftp only site but list in AGIS as Tier 3)

- Gridftp only sites – Not officially supported with ATLAS ADC. DDM team helps best effort – Transfers monitored on test Dashboard
- http://dashb-atlas-data-soup-tbed.cern.ch/dashboard/request.py/dataset?site=NYU-ATLAS_GRIDFTP

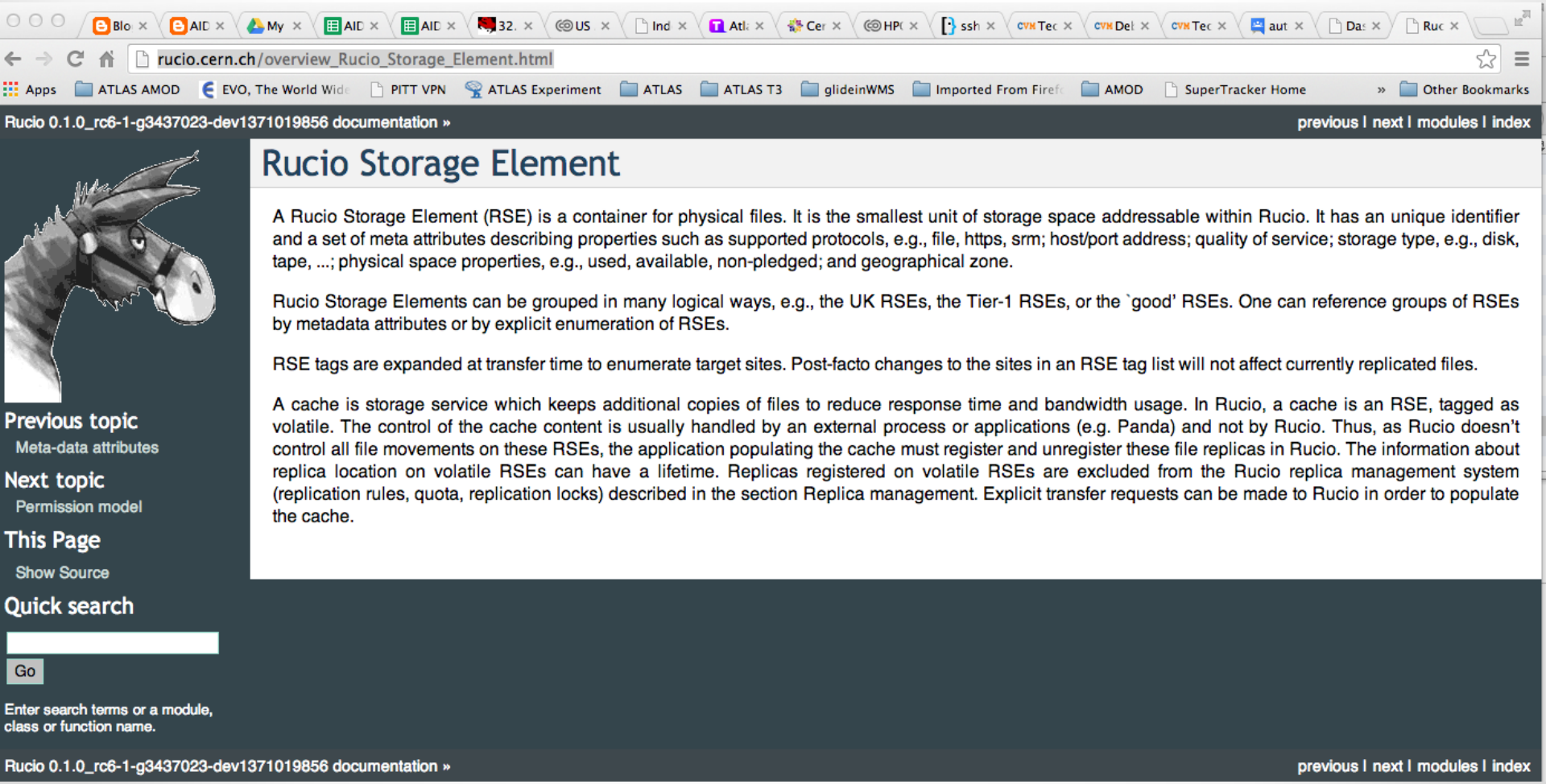
Why gridftp only endpoints in first place?

- Facilities need controlled way to transfer data
- Users want simple efficient way to transfer the data
- SLAC moves a lot of data through its gridftp only endpoint
- No worries about consistency between files on site and central LFC or Rucio catalogs
- File cleanup a local issue (no deletion service)
- No way to broker PANDA jobs against this storage (blessing and a curse)

- Can we do better?

Rucio Cache

- http://rucio.cern.ch/overview_Rucio_Storage_Element.html



Rucio Cache(2)

"A cache is storage service which keeps additional copies of files to reduce response time and bandwidth usage. **In Rucio, a cache is an RSE, tagged as volatile. The control of the cache content is usually handled by an external process or applications (e.g. Panda) and not by Rucio.** Thus, as Rucio doesn't control all file movements on these RSEs, **the application populating the cache must register and unregister these file replicas in Rucio.** The information about replica location on volatile RSEs can have a lifetime. **Replicas registered on volatile RSEs are excluded from the Rucio replica management system (replication rules, quota, replication locks)** described in the section Replica management. Explicit transfer requests can be made to Rucio in order to populate the cache."

Rucio Cache (RSE) development

- Have had several meetings in person and remotely with Rucio development team to identify the issues
- Currently technologies under development
  - Xrootd FRM (file redicency manager)
  - GRIDFTP server
  - Web DAV server (after first 2)
- Rucio team determining best way to get the file population and file depopulation information from Cache.
  - Current Baseline is to use call outs to ActiveMQ server at CERN
  - Need to agree on API (or at least the message format)
  - Xrootd solution looks straight forward
  - Just started to look at globus DSI (data interface) – Wei has given my the xrootd example for the DSI libraries
  - Also some though about publishing Cache content (ala ARC cache)

Can we use the Federated Storage at Tier 1 and Tier 2 ?

- US ATLAS has the potential to have decent amount of storage in Local Group Disk at Tier 1 and Tier 2 sites.
- This storage has excellent network connectivity
- Many US ATLAS institutions have been given NSF funds for significant network upgrades to the edge of the campus (Duke for example is going to many 10's Gbs WAN, other places even higher – 100 Gbe)
- Many US ATLAS insitutions are close in network time (others are not) to a Tier 1 or Tier 2 site.

Fax can be part of the Tier 3 data handling solution But….

- Need a consistent way to reference the data – With the new Rucio N2N system is much more scalable.
- Testing needs to identify the problems with the system before the end user
- For example – Recently – while doing data analysis I discovered a severe incompatibility between the MWT2 networking configuration and Xrootd.
  - Pilot error was blamed initially because the current level of testing did not show the problem. The current testing is not doing what users are doing some of the time.
- Worry that if the system is not more robust – far fewer failures using root code from a Tier 3 then initial users might walk away from the system and give it bad press.

Next Steps

- Need to identify more effort for the various development activities
- Will need help from Rucio team but they are focused on getting the system output (as they should be)
- Expect prototype Rucio Cache before the summer
- Expect a production ready part for Tier 3's by end of September
- More labor from facilities would help.
- Likely should think about better network monitoring at the Tier 3 site site (perfsonar boxes from Tier 1/Tier 2 cast off machines?)

Summary

- Need to consider end to end solution for data at the Tier 3
- From Grid to final plots for talk/ conference and paper
- All Solutions need to be trivial/ robust and set and forget.
- Need to minimize the support load on End users and local Site system admins.
  - Time administering the system reduces the time for Scientific Discovery
- Federated Storage and new Tools (Rucio Cache for example) will play a crucial role in Run 2 and we need to be prepared and ready by the data challenge next Summer if at possible