Computation Institute

Caching FAX accesses

Ilija Vukotic

ADC TIM - Chicago

October 28, 2014

THE UNIVERSITY OF CHICAGO

efi.uchicago.edu
ci.uchicago.edu

Caching – why and where?

- Straight Tier3
  - Most of the disk space is devoted to input data. The input data is almost always from downloaded from the grid
    - A lot of stale data
    - Tedious cleanups (mails asking people to clean up)
    - Different file paths
    - Have to worry about the data sizes
- Tier3 with a nearby Tier2 or Tier1
  - Users advised to DaTRI data to a localgroupdisk and use it from there. That solves all the problems above but it bloats the localgroupdisk.

Caching – why and where?

- Tier2
  - Most of the jobs accessing FAX data are overflown. These already come from the optimal place (thanks to cost matrix).
  - Any cache would have a very small hit rate.
- DDM Endpoint-less Tier2 – but with cache disk
  - Like UCL or a cloud based Tier2
  - Would have a very small hit rate – unless …
  - We specialize it:
    - Only certain physics group jobs?
    - Only certain type of jobs?
    - Only high priority stuff?

XRootD caching proxy

- Alja & Matevz caching plugin
  - Presented at the Federated Storage Workshop @SLAC*
  - Tested to hundreds of concurrent reads/writes good enough to saturate a 100Gb/s link
- Basics:
  - File level (pre-fetching)
  - Sub-file level caching
  - Caches blocks of configurable size.
  - Supports vector reads
  - Purging based on High/Low watermark
- But never tried in production environment

*https://indico.fnal.gov/getFile.py/access?contribId=38&sessionId=17&resId=0&materialId=slides&confId=7207

Configuration

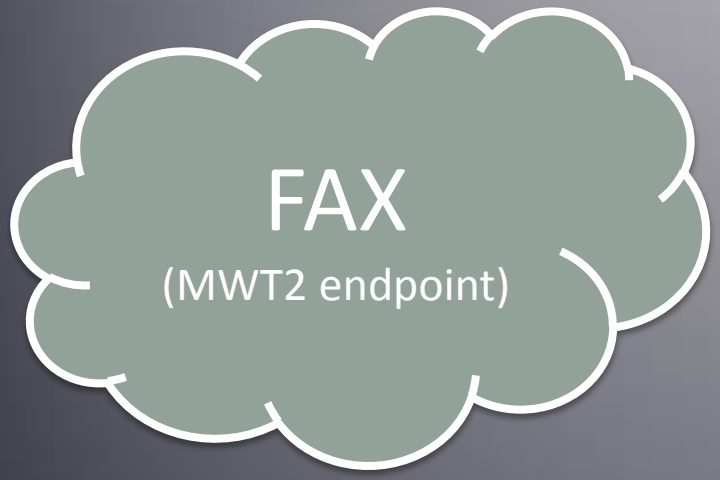
Original servers:

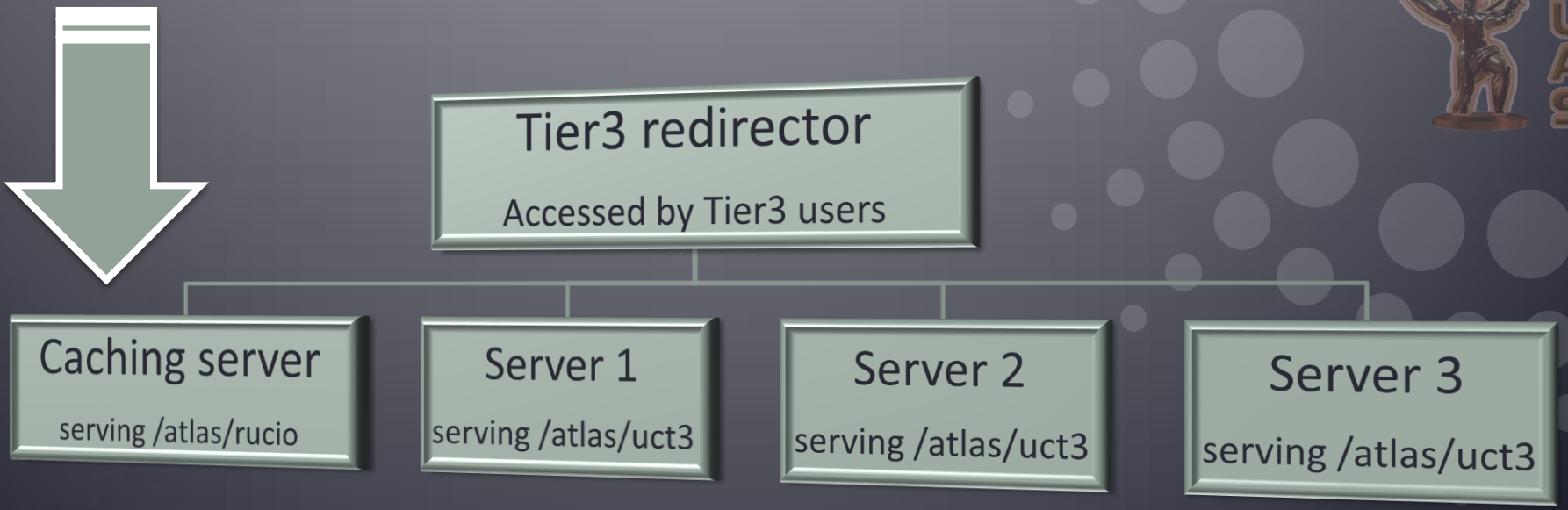
- 113 TB HDDs in 5 Dell shelves RAID 6
- Native xrootd
- Also used as interactive nodes

Caching proxy:

- 28TB HDD in 2 shelves RAID 6
- 2 x 160GB SDD in RAID 0, fronting HDDs
- Custom kernel + rebuilt tools from SL7 (bcache)

Straightforward configuration. (xrootd 4.0.4)
Zero downtime.





Performance

- One of the xAOD analysis tutorial lessons*
  - 200 input files (all available at MWT2)
  - Simple cut and plot example
  - rcSetup Base,2.0.14
  - ROOT 5.34.18, no TTC

Empty cache	1:25
Full file cached	1:07
Sub file	0:29

- Now open for end-users. Waiting for their feedback.

* https://ci-connect.atlassian.net/wiki/display/AC/xAOD+analysis+tutorial

Conclusion

- For Tier3 storage XRootD cache solves most of the issues
  - Admin friendly
    - Simple deployment
    - High performance
    - 30/70 storage/cache split and sub-file level caching recommended
  - User friendly
    - Provides more effective storage (stale files, files of long gone users, …)
    - Sub-file level caching can be seen as a one free skim/slim stage for everybody
- T2 usage still to be investigated upon longer Tier3 testing