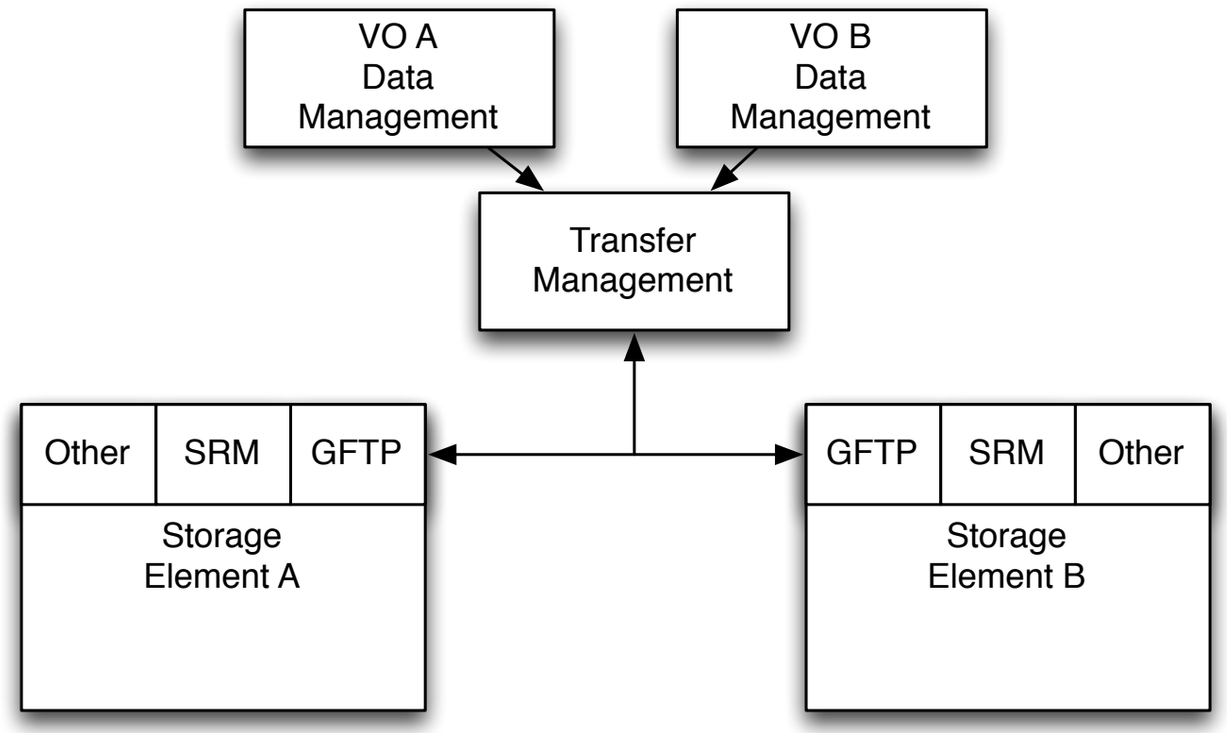
OSG Data Management Infrastructure

Evolutions, Revolutions, and Things Carrying On

Brian Bockelman - WLCG Workshop October 2016

Storage Element

- With the SE model, we have multiple services exposing a POSIX-like filesystem.
  - Each storage element acts independently.
- A higher-level transfer management layer moves files between SEs.
- VOs develop their own data management layer on top of that. Not quite so simple…



The Storage Element

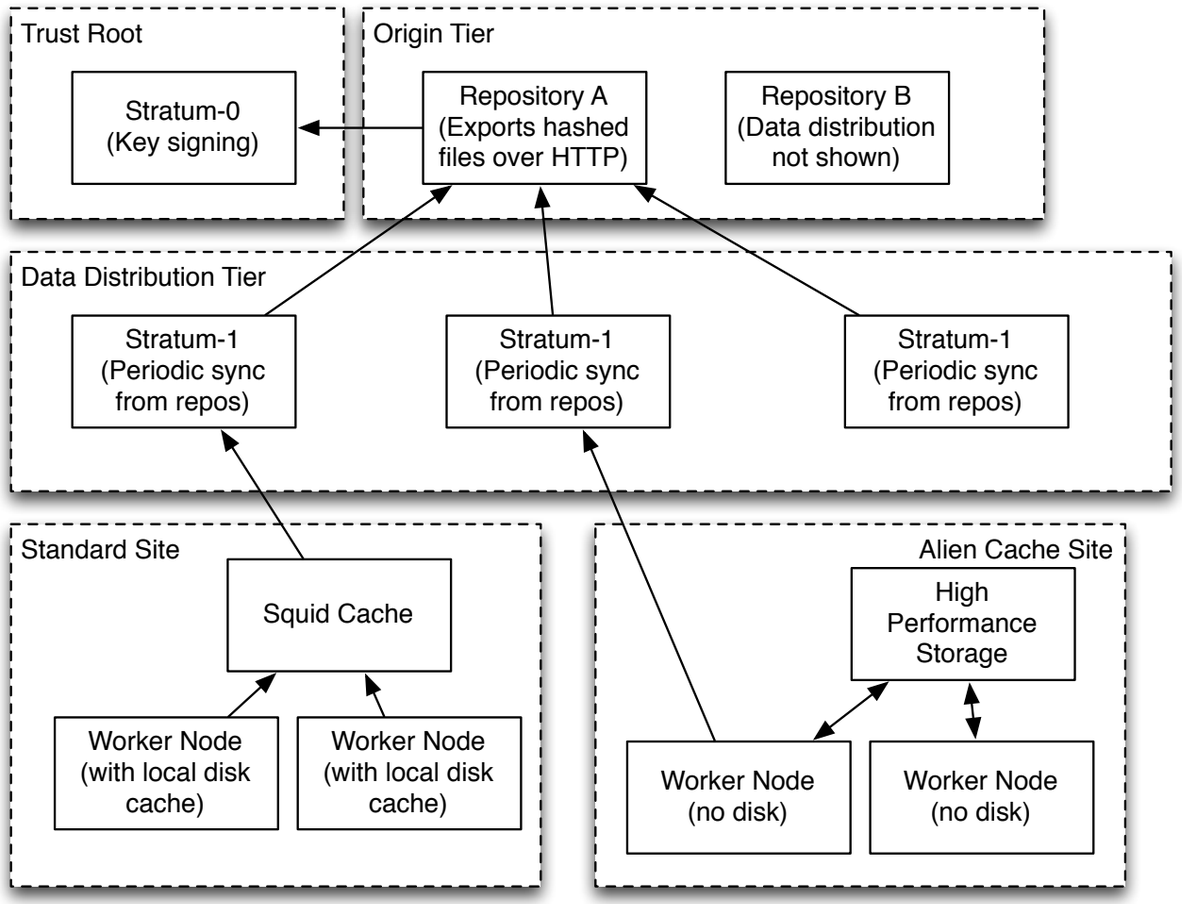
- In a storage element, the VO often assumes:
  - The SE is available and functioning.
  - The VO can keep track of the data it wrote to the SE.
  - The sysadmin responds to tickets and understands what the VO is asking.
  - Files, once written, are neither deleted nor corrupted.
- We've all seen counter-examples of the above…

A Different Paradigm: Caching

- The caching paradigm “fixes” many of the warts of the SE:
  - File loss - “cache eviction” - is common and acceptable. Admins can easily reclaim & repair storage.
  - No catalogs, simple-to-understand semantics.
- Downsides:
  - Caching is only useful if the **working set size** is smaller than the cache.
  - Caches do not add capacity to the overall system: 1PB of SEs + 500TB of cache = 1PB of storage.

Hypothesis: A significant number of LHC workflows and sites would benefit from caching

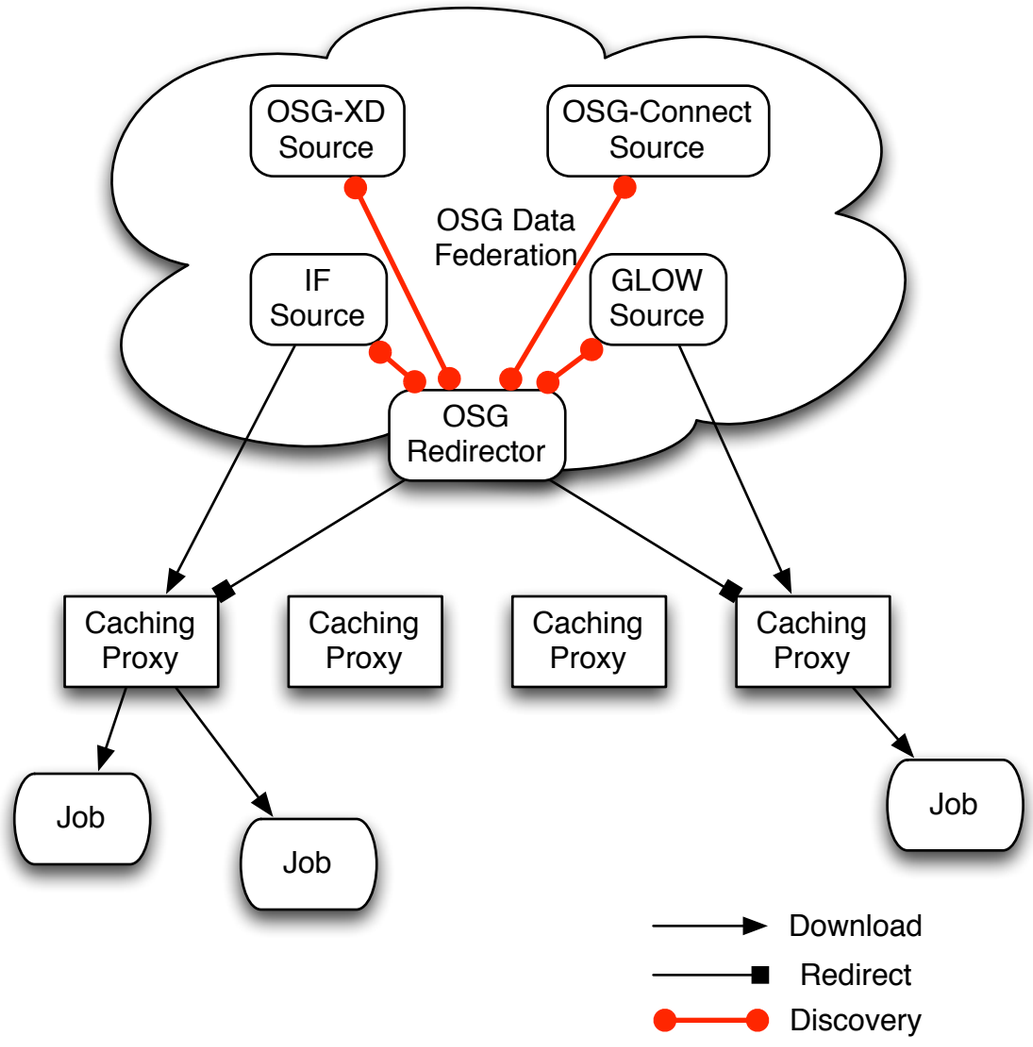
We Cache All The Time



- The “software distribution workflow” is a great example cache-friendly workflow. CVMFS’s has been very stable and performant.
    - OSG runs the data distribution tier, trust roots, and provides the origin for smaller VOs.

StashCache

- Caching infrastructure based on SLAC Xrootd server & xrootd protocol.
- Each VO has a origin server.
- OSG runs the redirector and caches.
- Jobs utilize “nearby” cache, for some definition of nearby.
- User interfaces:
  - “cp”-like: *stashcp.*
  - HTCondor file transfer
  - **POSIX**



stash.osgstorage.org

- We use CVMFS to provide a POSIX interface on top of StashCache.
  - We index and publish the contents of the OSG-VO “Stash” origin a few times per day.
  - From the login host, users write into their ~/public/ directory; waits until the files show up in /cvmfs. To users, analogous to a global read-only filesystem.
- With LIGO, we have created ligo.osgstorage.org that additionally adds GSI-based authorization on top of CVMFS.
- Have done a 1PB demo of publishing CMS data.
- Unifies various tools - data federations, caches, CVMFS - behind a POSIX interface.

A National Trend

- Universities consider **research computing a central service** similar to the library.

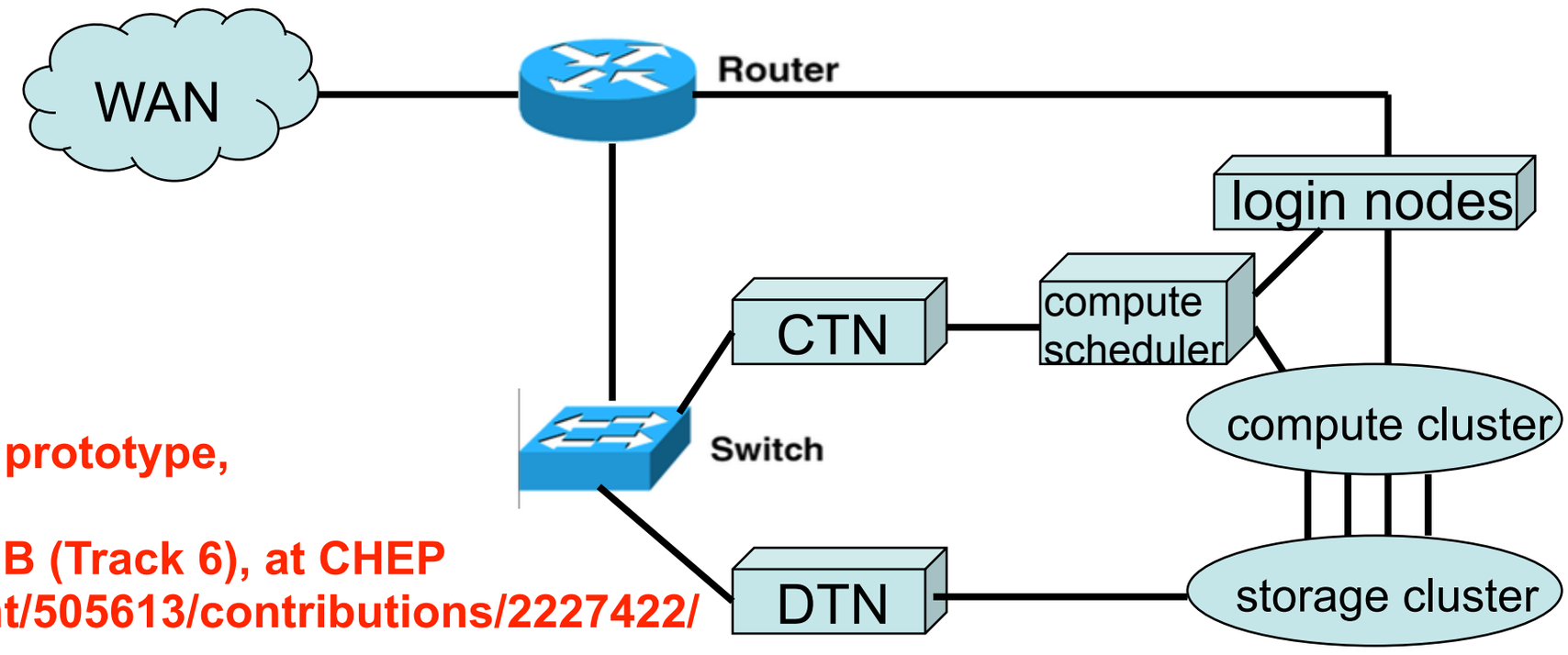
- Many US universities have “T1-scale” clusters they operate and own.

- Happy to help their local HEP professors: no background in grid computing, no specialized resources, and above all, *little time*.

HOW CAN OSG & WLCG BENEFIT?

Tier-3 In A Box

- Ask site to install CVMFS for software.
    - It appears that unprivileged container improvements in RHEL8 timeframe will allow us to do this ourselves.
- Xrootd cache to for file access.
- Xrootd server to export private analysis ntuples.
- Jobs can be submitted through a simple SSH connection.



For details on the existing prototype, see Jeff Dost's talk
October 13th, 14:00, Sierra B (Track 6), at CHEP
https://indico.cern.ch/event/505613/contributions/2227422/

Carrying On

- OSG is exploring caches to help “expand reach” of dedicated sites or university clusters. We certainly still must run “normal” Tier-2s and Tier-1!

- Looking forward:
  - Bestman2 support ends in approximately a year. At that point, SLAC Xrootd and Globus GridFTP will be our software platforms for data transfer.
    - Over the next year, we will ramp-up efforts to smooth the removal of bestman2.
  - Our supported storage solution, HDFS, will get a major upgrade to HDFS 3.0: sites running this will get a 35% boost in usable storage by switching to Erasure Codes.

Final thoughts

- Significant investments in decreasing complexity of running an OSG site. Very interested in the use case of large University clusters without dedicated staff.
    - Currently looking at doing this with cache-based technologies.
    - Getting ever-closer to having the entire stack (inc. CVMFS) work without needing special privileges.
- Still committed to evolving and maintaining existing storage approaches.