Any Data, Anytime, Anywhere for the Open Science Grid Communities

XRootD workshop @ UCSD

January 29th, 2014

Frank Würthwein

OSG Use in 2014

- 67 Institutions within US and South America provided 789 Million hours of compute time total.
  - This translates into an average of 90k cores 24x7.
  - 68% for ATLAS & CMS
  - 16% for the 5 main opportunistic VOs
    - they operate across all sites that provide them resources.
    - 10 sites provided 75% of resources in 2014 for these.
  - 16% for all other VOs
    - most of them operate only at specific sites they have social ties to.

Big Data on OSG

- To good approximation, only ATLAS & CMS do Big Data on OSG

Can we use XRootD to Open Big Data for all of Science on OSG ?

What's Big ?

- Today, all we can do is whatever can be moved around with the application, or stuffed into OASIS (i.e. libraries etc.)
- Even 1TB dataset capability would be a x100 step forward.

Support a few communities with needs between 10GB and 10TB each.

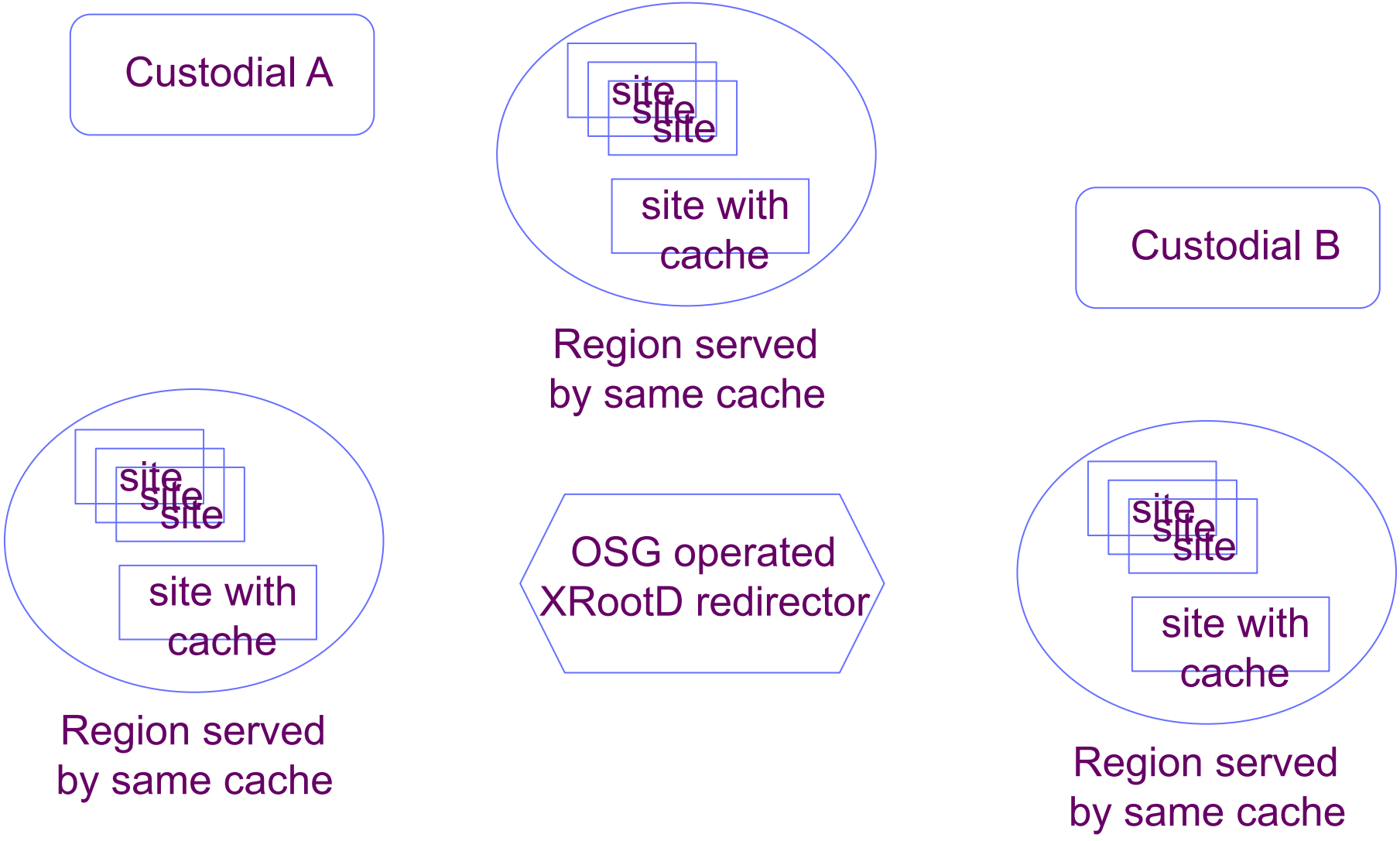
Conceptual Idea (I)

- Single PIs served by OSG-Connect or XD allocation.
  - PIs import their data into their allocated space at SDSC or U. Chicago
    - Let's call this their "custodial copy", even though it is temporary and disk resident.
- VOs have disk resident "custodial" centers
  - e.g. Intensity Frontier Experiments @ FNAL

Conceptual Idea (II)

- OSG & its partners distribute XRootD Caching Proxies across OSG
  - start with a handful at sites that provide most of the opportunistic cycles.
- OSG establishes uniform runtime environment for applications that does not require any knowledge of XRootD API
  - Fuse ? XRootD POSIX Layer ? What is the right implementation ?

A picture is worth …



Thoughts on Scale

- A cache should be no more than ~ $5k
  - 12x4TB SATA disks @ ~$200 plus modest server.
  - Total size per cache ~ 48TB. No RAID needed.
  - Plenty of space not to worry about thrashing
- Caches should be 10Gbps connected to WAN to serve region.
- Expect typical system use not to exceed ~10k cores. Peak use no larger than ~30k cores across all caches.
  - Each cache is hit with only a modest size hammer.

Thoughts on Operations

- Do we need to run multiple independent data federations across the same hardware but different ports?
  - Achieve some degree of separation ?
- Do we operate XRootD config centrally via remote config options?
  - sites that host cache need only operate hardware.
- Fuse or POSIX pre-load library or ???

Are we ready to give this a shot ?