Storage

(A tale of three cities)

Itinerary

- Glasgow situation
  - Xrootd, Ceph
- QMUL situation
  - StoRM and WebDAV TPC
- Bristol situation
  - DPM-on-HDFS to Xrootd-on-HDFS

Glasgow Storage

(Sam Skipsey)

Some background

- "New" Storage:
  - Ceph (Nautilus) cluster
  - 3 x MON / MGRS
  - 18 x OSD [storage] hosts,
    - Each with 19 OSDs/HDDs @ 10TB each
    - 10Gb/s bandwidth
- Data stored as objects in low-level object store ("RADOS") using libradosstriper to chunk "files" into stripes of consistently sized objects.
- (RADOS layer itself is 8+2 or 8+3 EC coded, with host-level resilience)
- Access via gateway nodes running:
  - Gridftp with "gridftp-ceph" DSI
  - Xrootd with "xrd-ceph" OFS [with proxy-caches in the way]

Xrootd/ceph "issues"

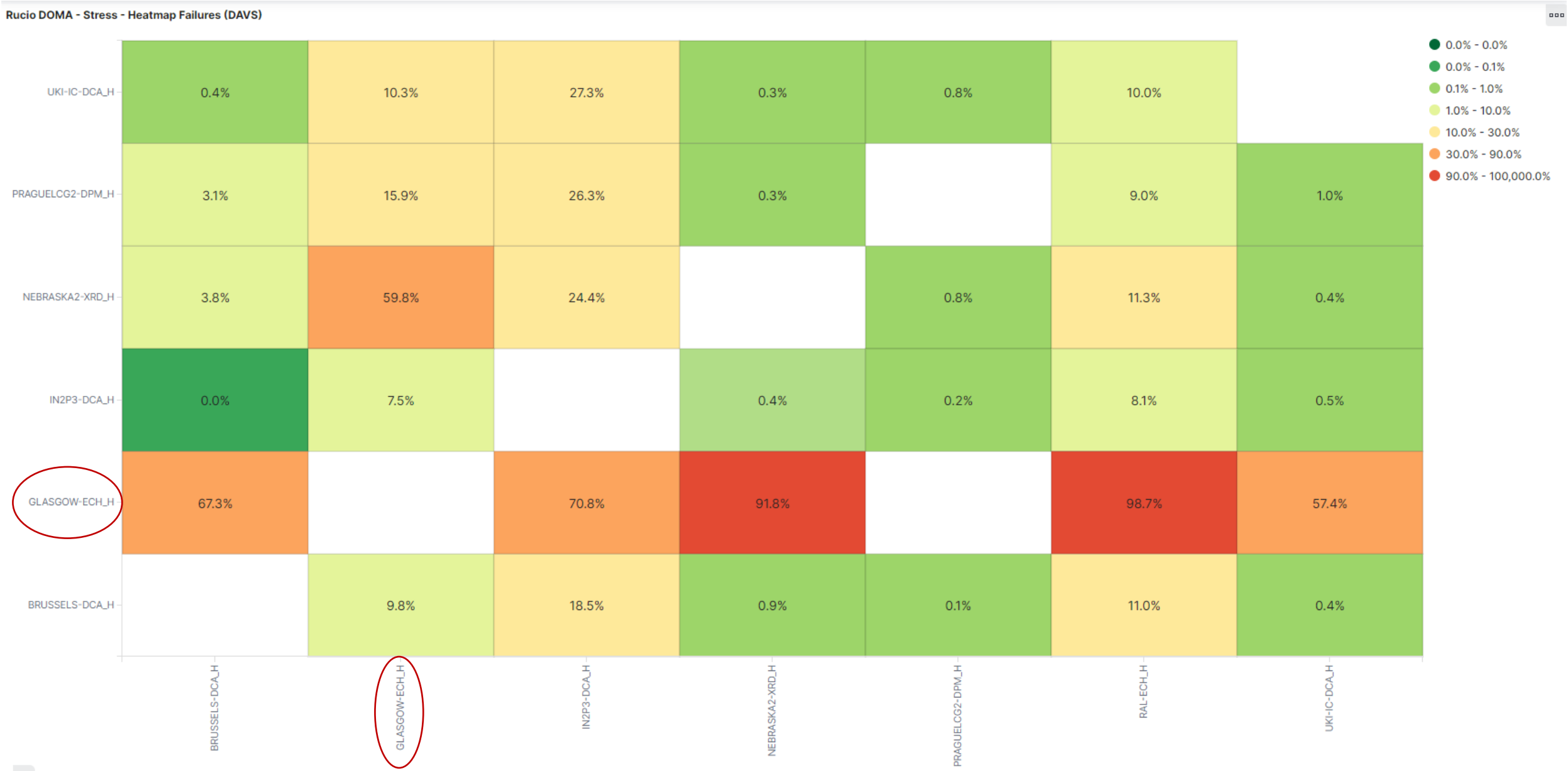
- Libradosstriper issue 1:
  - Lock exhaustion/overflow on many concurrent reads to same file.
- Libradosstriper issue 2:
  - "Junk data past end of file" when reading files with large block sizes.
- Xrootd (WebDAV) / RADOS impedance mis-matches:
  - Block sizes for RADOS should be large (on order of chunk size) - 16 to 64MB
  - WebDAV impl. in Xrootd has 1MB hardcoded buffers == write blocks.
  - Low read and write performance as a result.
- "POSIXness" assumptions in tools
  - gfal (for ex) needs to believe it can create directories, and will fail if returned NOT_SUPPORTED.
- OFS plugins no supported in XrdCeph?
  - Namespace TFC plugin can't be loaded by XrdCeph
  - you need a proxy in front of the server just to do TFC conversion for it…

WebDAV-TPC commissioning

- GridFTP-ceph gateways are reliable and efficient in our experience.
- However, GridFTP is deprecated, and WebDAV is the official preferred replacement.
- We provide this via the Xrootd (5) WebDAV* support.
- Achieving good performance as both a source and a destination has been more difficult than we expected…

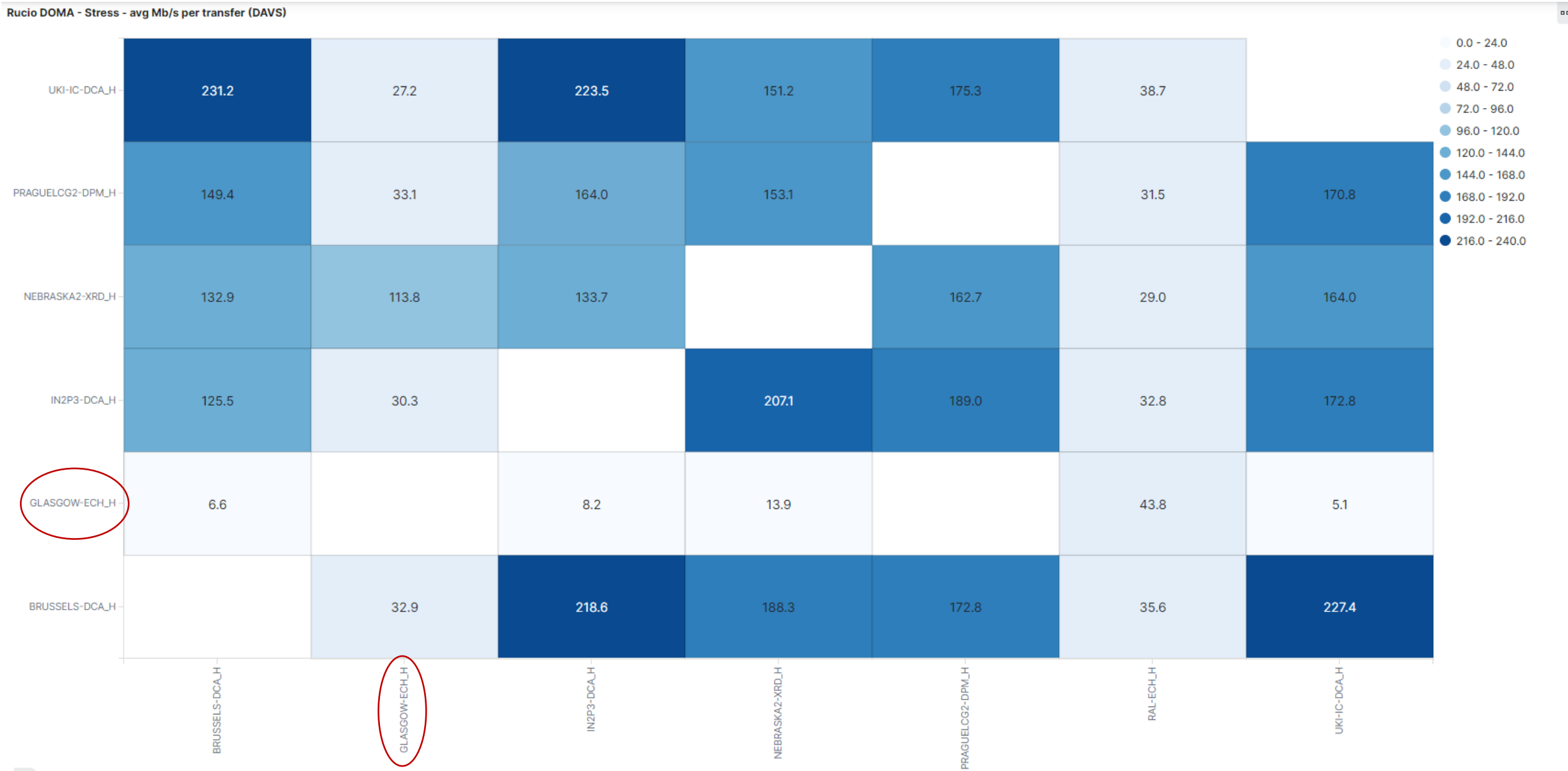
*There is some evidence to suggest that xrootd over Xrootd would be better performing.

Rucio DOMA tests [~last 3 weeks] - "Failures"



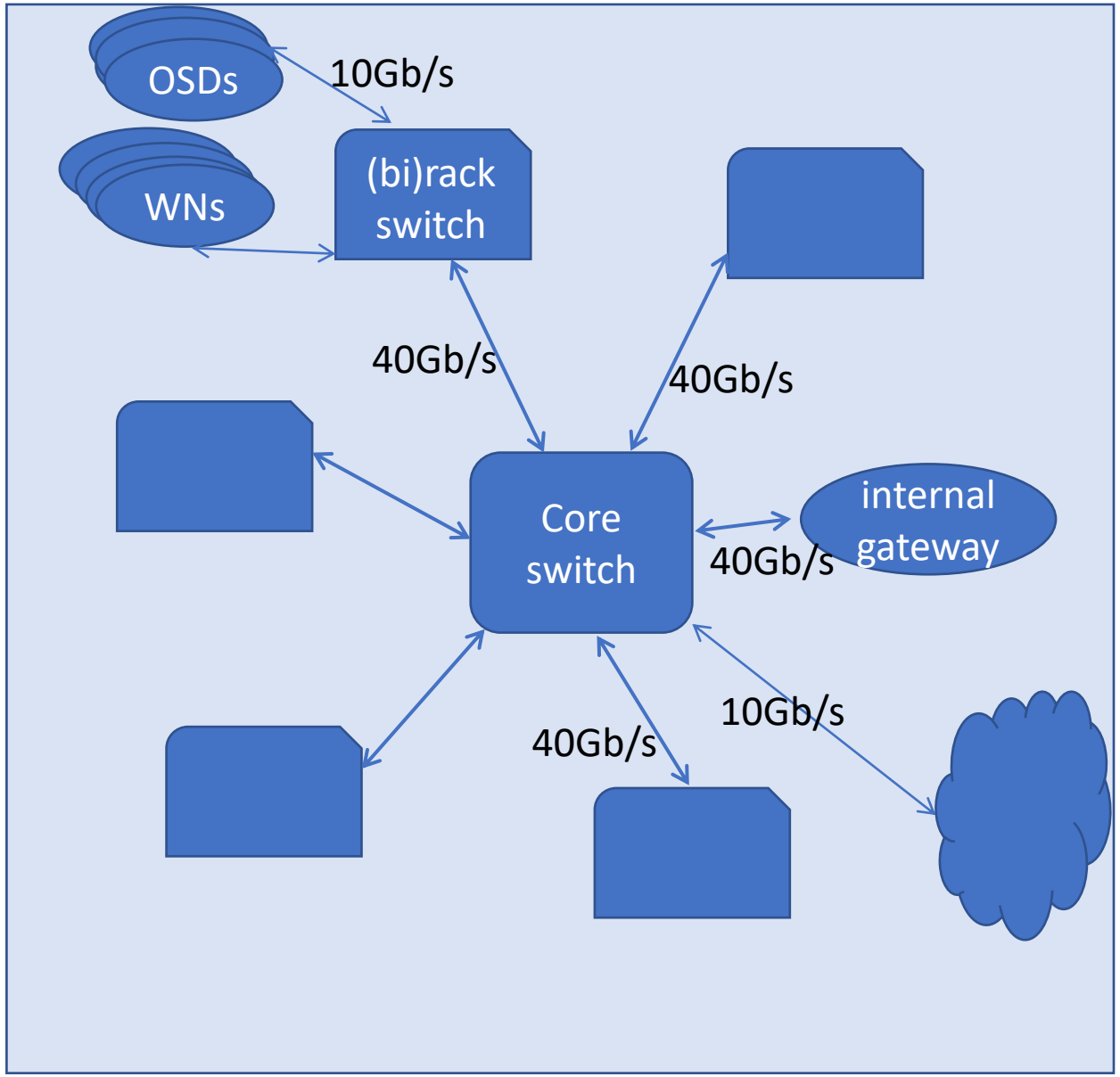
Source

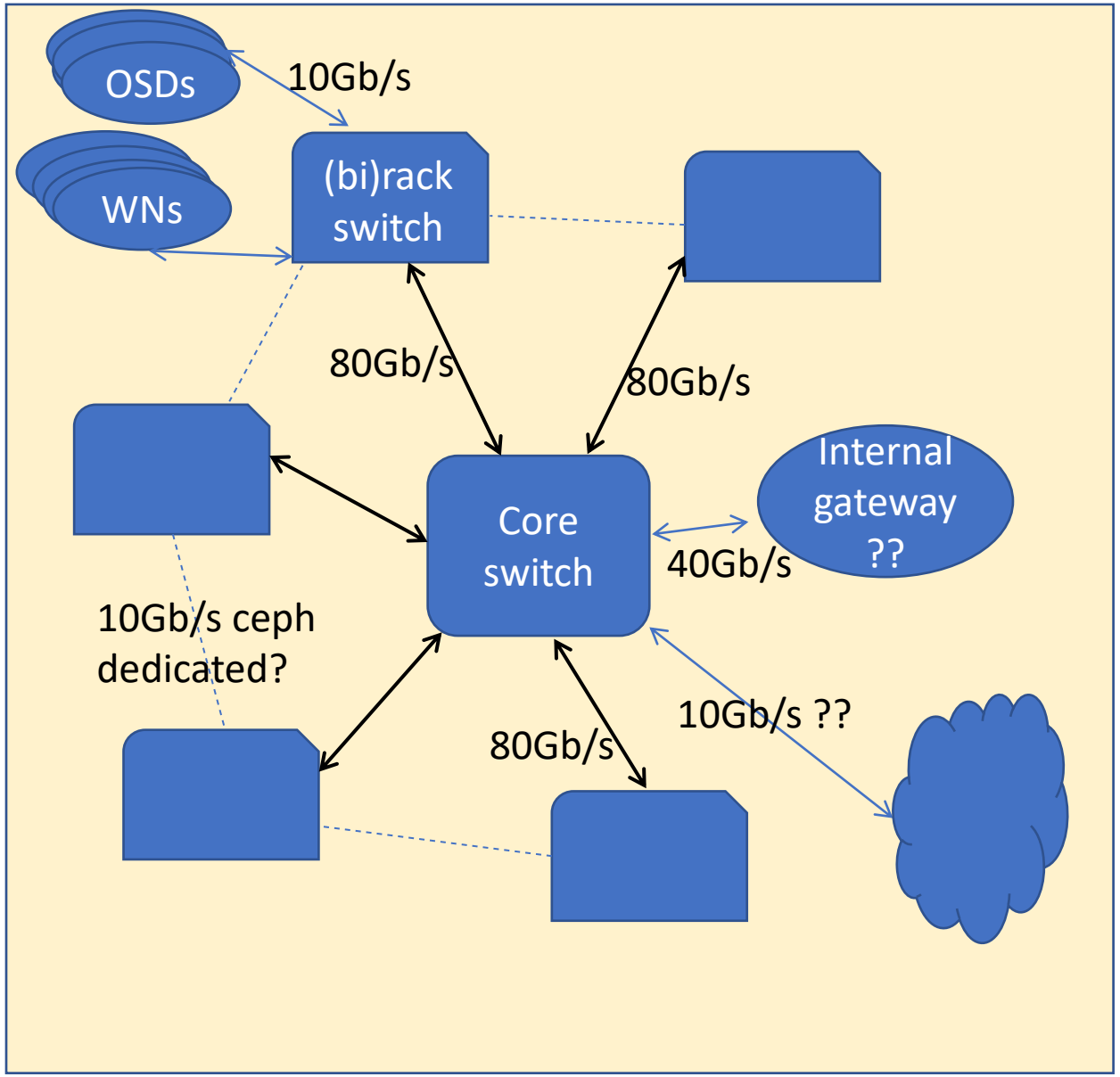
Rucio DOMA tests [~last 3 weeks] - Rates



Source

Glasgow network changes.





Transition to cephfs

- "New" tranche of storage is
  - 19 hosts @ 18 x 16TB disks (+ SSDs)
- Original plan was to just extend the existing cluster.

- NEW plan is to build a parallel cephfs [posix-like fs] cluster.
- Cephfs volume provides own striping, locking layer with better support than libradosstriper.
- Xrootd can provide access w/o specialised plugin (just posixish)
- WNs *could* mount cephfs directly, no longer needing an internal gateway

  - like StoRM/Lustre sites e.g.

QMUL Storage

(Dan Traynor)

QMUL Storage - software

- Present system is 5 PiB usable running Lustre 2.12.6 LTS, several point upgrades done.
- Next LTS release expected to be 2.15 with, hopefully, IPv6 support and Linux native client. Expect to migrate to this next year.
- Significant developments in pipeline. Most developments not that important for grid usage (more HPC focused) but we should see performance scale as networks get fatter and disk arrays get bigger (more parallelism and optimisation for PB sized disk arrays).
- Clear maintenance and support structure (regular LTS releases), DDN employ the main Lustre developers + others including national labs.

QMUL Storage -Hardware

- Hardware based mostly around 2U 16/24 disk servers, using Raid cards. Easy setup, easy monitoring and maintenance using manufacturers' software (e.g. Dell openmanage), stable running.
- Next tender almost out: Move to larger disk arrays (84/96 disks in a box) for lower power consumption, may also be able to increase rack density (400 to 500 disks a rack). Still use Raid cards rather than ZFS. Always ask for enterprise NLSAS disks not SATA. zero disk failure over 700 disks in last year (touch wood).
- Network over last 3 years has been almost all been upgraded to 25/100Gb/s Mellanox switches. New storage will be connected at 100Gb/s. Most other storage/compute connected at 25Gb/s.

Bye bye R510s
(1.5PB)

Hello R740XD2
(3.5PB)

QMUL Storage - StoRM

- Significant effort to enable TPC with WebDav: Complete change in config for CentOS7; Migrate from SRM for storage accounting to using Lustre project quotas (folder level quotas). Better in long term as we have better knowledge/control of what's happening in the file system (e.g. quota limits).
- StoRM ready for token-based authentication.
- Generally stable performance. Can usually sustain ~20Gb/s through one server.
- Still issues with webdav performance to some sites. In long term will need to scale out endpoints (more SEs) to improve throughput for webdav beyond 20Gb/s when 100Gb/s Wan lands.
- Some issue with file ACLs and checksums not always set properly.

QMUL Storage - Summary

- Happy with Lustre performance, stability, support. Plan to maintain as core storage for foreseeable future.
- StoRM a bit more rough around the edges but developers provide decent user support. As long as supported will continue to use it. Fall back could be dcache.
- Readonly xrootd instance will have to be decommissioned as lcmaps support is deprecated (logs show very little usage).
- Local ITS Research group building Ceph cluster could be interesting in the future.

Bristol DMLite Replacement

(Luke Kreczko)

Backstory

- Bristol uses HDFS as storage backend
  - Grid is secondary: storage primarily used for local users
  - Multi-VO: mostly CMS, but also DUNE, mu3e, LZ
- DMLite HDFS plugin obsolete for a while
  - Only alternative: xrootd + HDFS from OSG (note: HDFS plugin might go away > 3 years)
- xrootd SE fits requirements
  - Supports xrootd and HTTPS transfers
  - Supports macaroons and sci-tokens
  - Should we ever switch storage system ? likely to be supported
  - Fine-grained permissions in theory available (e.g. mapping grid users to local user accounts)

Implementation (xrootd > 5.0.0)

- Documentation mostly OSG
  - xrootd-lcmaps plugin is deprecated
  - No alternative documented (handled by magic OSG scripts?)
- Big help from Sam
  - Config, authorization file, etc
  - Using ARGUS in ban-only (but still xrootd-lcmaps)
- Sites in Germany try gridmapfile only approach
  - No reliance on xrootd-lcmaps?
  - Trying similar in Bristol

Work in progress

- Building up knowledge took time
  - Documentation, experts, GitHub issues to fill the gaps
- Slow implementation due to small fraction of time available
  - Coming back to the topic, reloading registers is not efficient
- Xrootd + lcmaps > local user accounts initially worked
  - Upgrades (xrootd, lcmaps?) broke things > needed alternative path
- Switched from VM on cluster to WLS 2 + Docker-compose on PC
  - Faster DEV cycles, easy to test `clustered` config
  - Can even set up a HDFS cluster for testing

(Overall) Thoughts

- Future of sites moving from DPM is still a little open.
  - See Matt's Lancaster talk for their thoughts
- XrootD becomes an increasingly important "critical software component", similar to where GridFTP was.
  - (Even EOS sites are implemented as
- As does distributed storage
  - Lustre, Ceph(fs), HDFS

- (Outside of scope of this talk: network capacity planning will also impact all of this.)