GridPP Data & Workload Projects

(A rapidly constructed update.)

Sam Skipsey

Thanks to James Walder, Rob Currie, Wenlong Yuan, Vip Davda, Emanuele Simili for slides I have used, remixed or otherwise taken for this presentation

Items

- DPM Retirement + Storage evolution (a context)
- Xrootd "caches" (Xcache or memory-backed)
  - Oxford
  - Edinburgh
  - (ATLAS) Virtual Placement @ Birmingham
- StashCache and other work
  - Edinburgh
- HEPSCORE + power efficiencies
- Future work + conclusions

GridPP6, DPM retirement + tokens

- GridPP context: GridPP6 consolidates storage at 5-6 Tier-2s, down from ~all 17 in previous GridPPs.
- Most of the sites transitioning away from storage use DPM
  - DPM is also being dropped as a storage solution by WLCG [timescale ~mid24]
- Move to token auth from x509 also driving this (as DPM does not support this).
- Currently exploring ways of efficiently running "storageless" sites at non-core Tier-2s
  - Xrootd caches, Virtual Placement
- (Core Tier-2s also exploring new technologies - xrootd/cephfs + xrootd/rados)

XCaches

- The easiest way to deal with the problem of storage is to… stop having storage.
- Several sites in the UK have already made this transition either entirely (Sheffield, Oxford) or for some subset of their client experiments (QMUL for CMS [retaining ATLAS], Imperial for ATLAS [retaining CMS], Birmingham for ATLAS [retaining ALICE]).

- Oxford + Birmingham have been testing storage-backed xrootd proxies ("Xcaches") for GridPP for some time now.
- Edinburgh, Glasgow (and RAL) have also explored internal caching xrootd proxies for various cases.

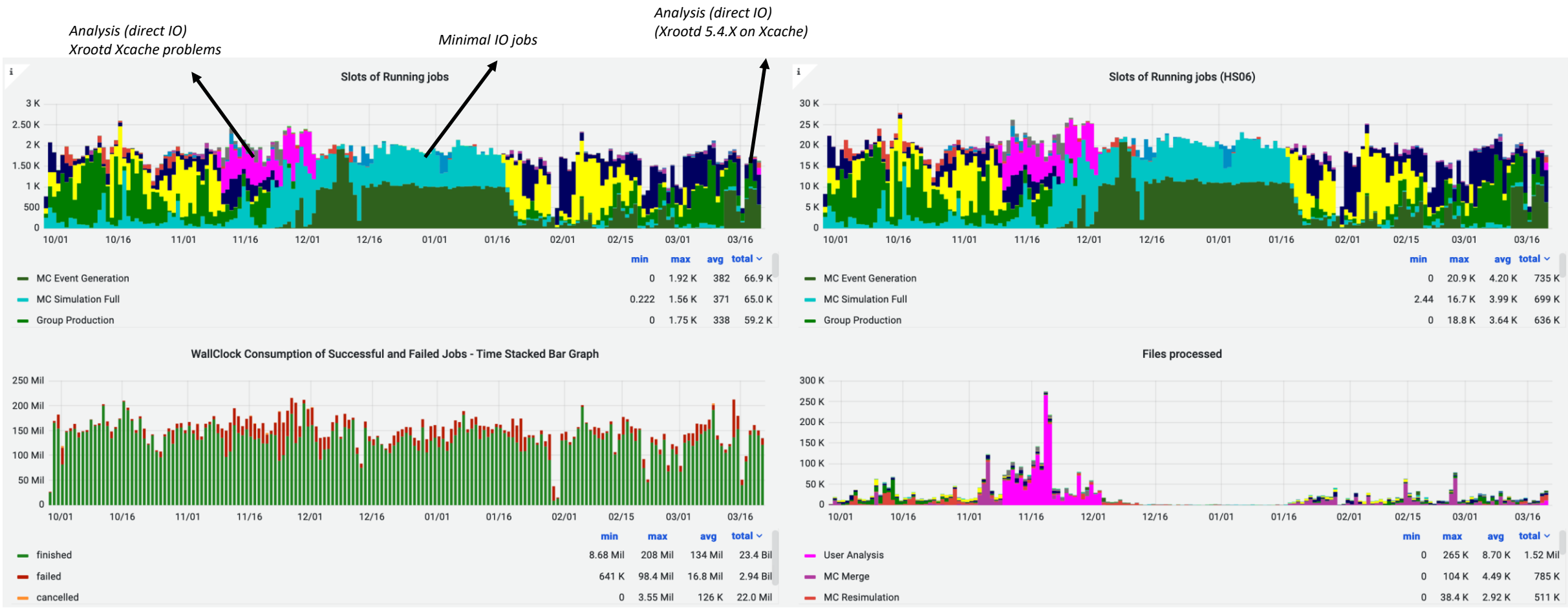
Oxford

- Configured as a simple caching proxy (RAL as the remote host).
  - ATLAS workload management assumes a single "close SE" for compute.

- Initial hardware was a repurposed disk server. This needed tuning to provide useful Iops - RAID6 not capable of write performance.
  - Numerous issues with various xrootd release updates causing performance regressions.
  - Failure coupling: issues at RAL cause issues at Oxford.
- Oxford now has a second Xcache host, loaned by RAL
  - High performance, modern, SSD backed storage server.
- Initial plan is to compare cache effectiveness between this and old cache
  - Demonstrates effect of "removing" IO bottlenecks on the cache hardware itself.

ATLAS view of Oxford (22Q1)

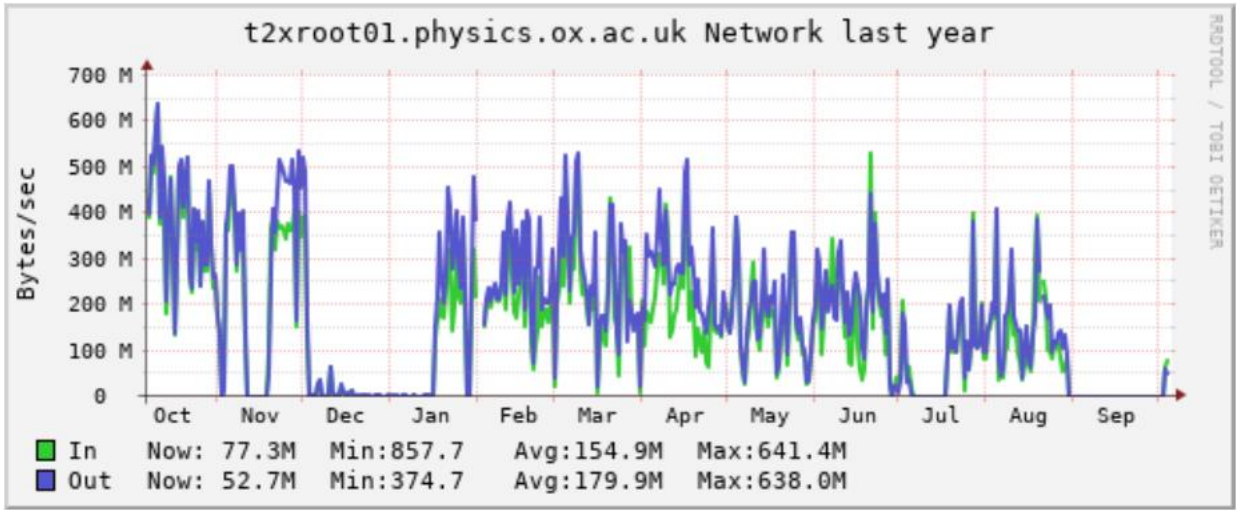
- Various tests of different ATLAS / Xcache (on/off) configurations attempted.



- Initially large problems with direct-io; traced to issues with XrootD, and since fixed in 5.4.1

Oxford

- Xcaches do not have to cache all of a file accessed through them.
  - Partial reads are possible (especially useful for some analysis workflows which sparsely read their input files).
  - This is not tunable for different file types, however:
    - Setting "prefetch" to whole file would make analysis jobs less efficient
    - Setting "prefetch" to 1 block would make production jobs inefficient (as they benefit mostly from having the entire file local / latency hiding)



Network traces on cache suggest that most of the time it is buffering more than caching (network in ~= network out) due to mostly production workflow

Virtual Placement @ Birmingham

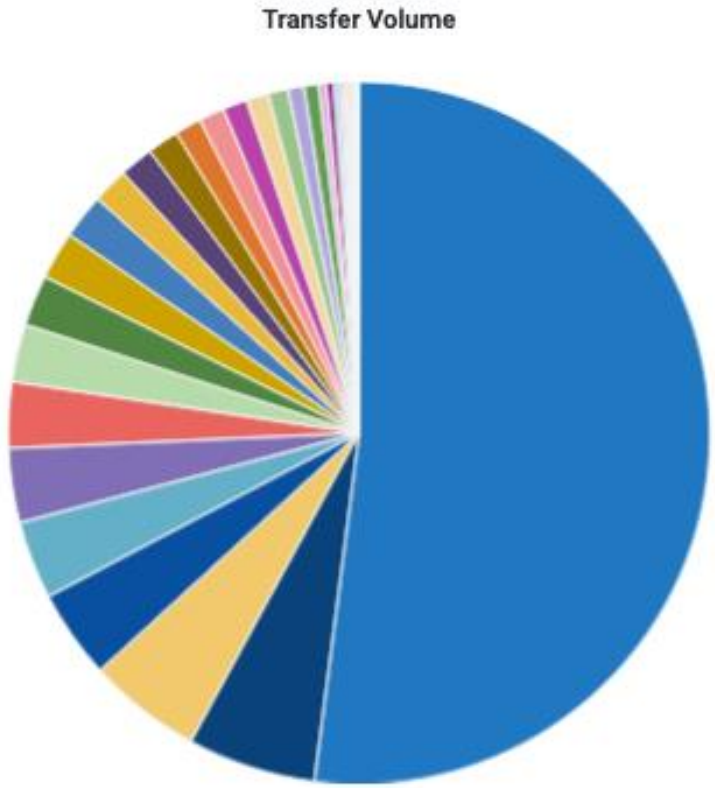
- (see Alessandra's AF presentation yesterday)
- Initially simple Xcache [paired with Manchester remote]
- Mature Virtual Placement setup by 1 October this year:
  - Tighter integration with ATLAS workflow management - data is preplaced into cache before jobs that require it arrive.
  - (So hit-rate is much higher than for an opportunistic cache).

Site	Accesses	Cache hits	Cache misses	File read [%]
Birmingham	70,127	39TB	1.4TB	~ 75%

Comparison to simple XCache

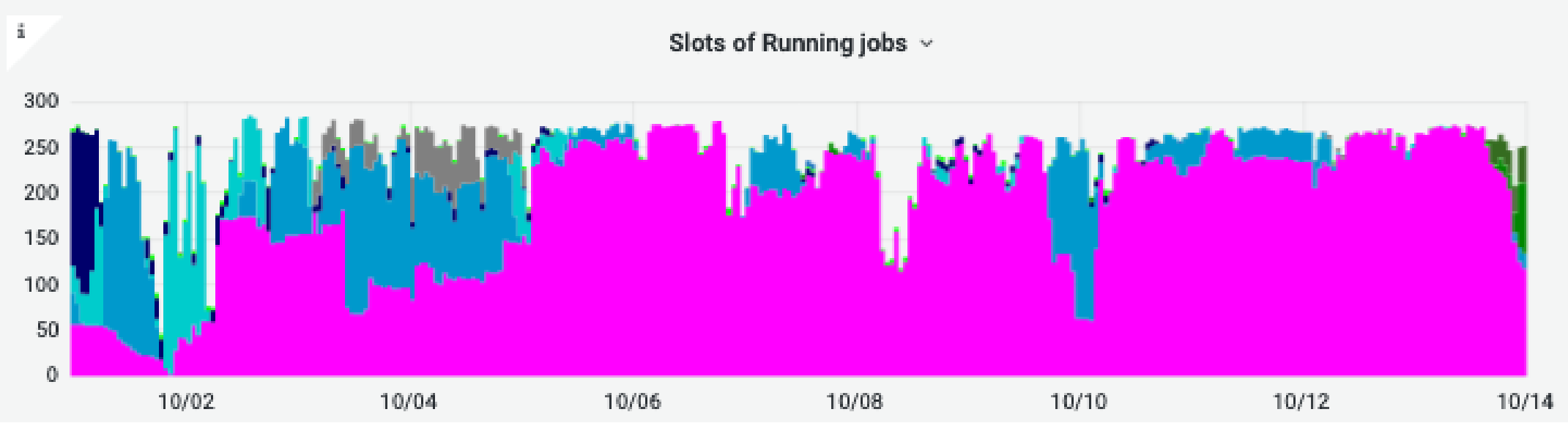
- As well as an intrinsically higher hit-rate, VP configuration has other advantages over naïve XCaches.
- The cache here is prefilled from the entire cloud, not just Manchester

	total	percentage
UKI-NORTHGRID-MAN-HEP	25 TB	52%
RRC-KI-T1	3 TB	6%
CERN-PROD	3 TB	5%
FZK-LCG2	2 TB	4%
IN2P3-CPPM	2 TB	4%
BNL-ATLAS	2 TB	3%
CA-SFU-T2	1 TB	3%
NDGF-T1	1 TB	3%
UKI-LT2-QMUL	1 TB	2%
Taiwan-LCG2	1 TB	2%
TOKYO-LCG2	1 TB	2%
pic	806 GB	2%
RAL-LCG2	753 GB	2%
AGLT2	705 GB	1%
IN2P3-CC	590 GB	1%



Comparison to simple XCache

- As well as an intrinsically higher hit-rate, VP configuration has other advantages over naïve XCaches.
- The cache here is prefilled from the entire cloud, not just Manchester
- And jobs here are analysis (potentially high-io) workloads.



Data Management Work @ Edinburgh

- Edinburgh is currently configured with an internal cache (between their DPM and WNs).
  - This does get good hit rates, *with appropriate tuning.*
  - About 40% traffic reduction relative to cacheless.
- Previous work @ Edinburgh suggested that many disadvantageously cached files are simply identifiable (by, eg, suffix)
- Current work underway to test custom caching library, to avoid caching such files.

- Edinburgh is also working on Xrootd monitoring topics.

StashCache @ Edinburgh

- DUNE utilizes CVMFS as it provides a read-only POSIX interface to StashCache
- CVMFS client is the first choice for StashCache, as the most redundant features, including
  - built-in GeoIP locating
  - rate monitoring
  - fallback in failures
- DUNE use StashCache to deliver larger payloads such as flux files and shower libraries to grid jobs
- DUNE CVMFS StashCache eliminates need to copy files to every single job all the time, reduced FNAL dCache load

/cvmfs/dune.osgstorage.org

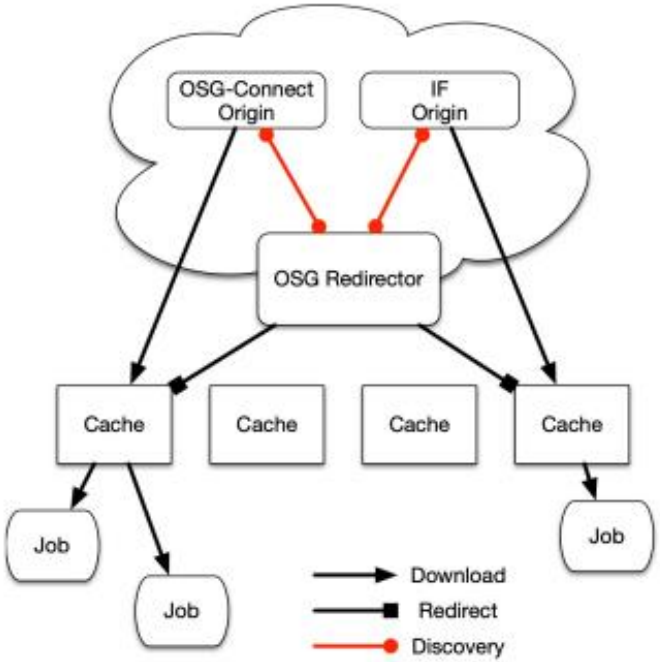


Figure 1: StashCache Architecture: Jobs request data from caches, which in turn query the redirector for the data location. Data is transferred from the origin to the cache, and then to the job.

arXiv:1905.06911

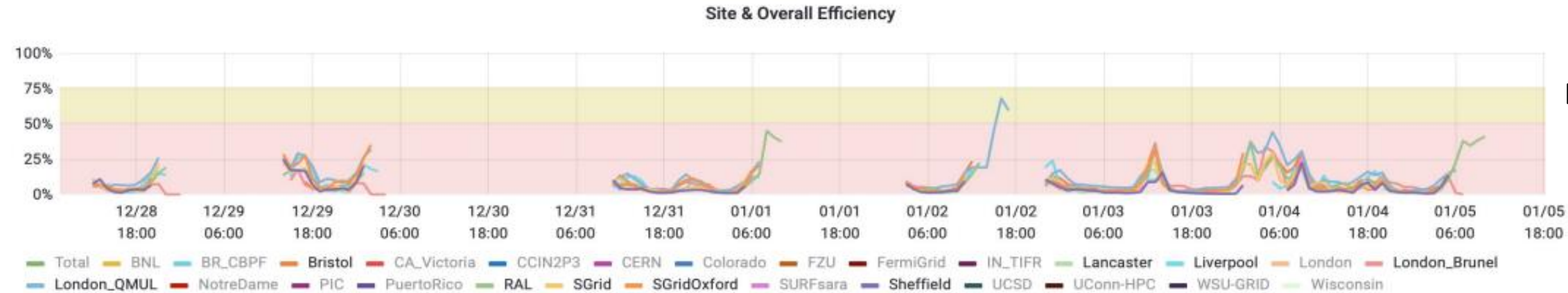
StashCache implementation at Edinburgh

- Deploying and commissioning since April 2022
  - Installing the Open Science Data Federation Cache (OSDF Cache) from RPM
    - https://opensciencegrid.org/docs/data/stashcache/install-cache/
    - straightforward setting up
  - Registering in the OSG
    - https://osg-htc.org/docs/common/registration/#registering-resources
    - https://github.com/opensciencegrid/topology/blob/master/topology/University%20of%20Edinburgh/Scotgrid%20ECDF/UKI-SCOTGRID-CDF.yaml
  - StashCache is OSG Xcache
    - utilizes CVMFS
    - A HTTP(S) based file caching network

Edinburgh registered as an OSG Cache site in the UK with Geo-IP data

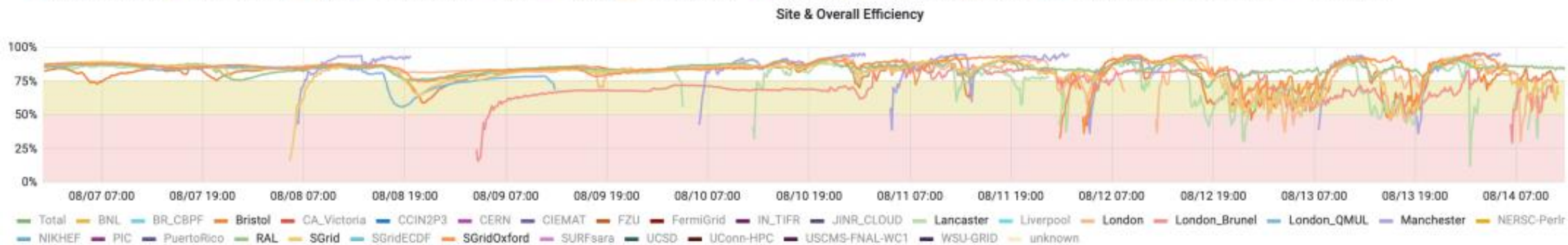
DUNE Jobs in August

- DUNE UK sites Job Eff. are >60% in average in August
- Edinburgh StashCache solved the flux file low efficiency problem



Before Cache

Dec 21' - Jan 22'



After Cache

Aug 22'

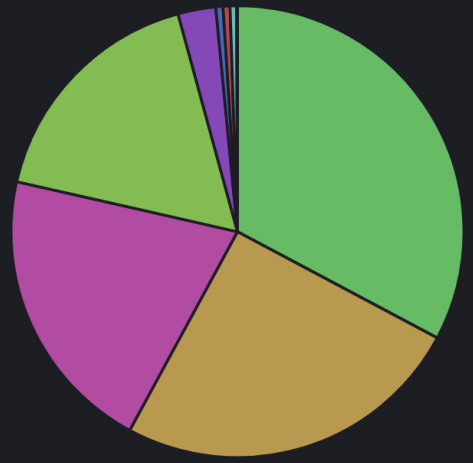
Metrics as of 1 Oct 2022 (~6 months running). Approximately 18x traffic reduction based on hits.

OpenSearch Dashboards

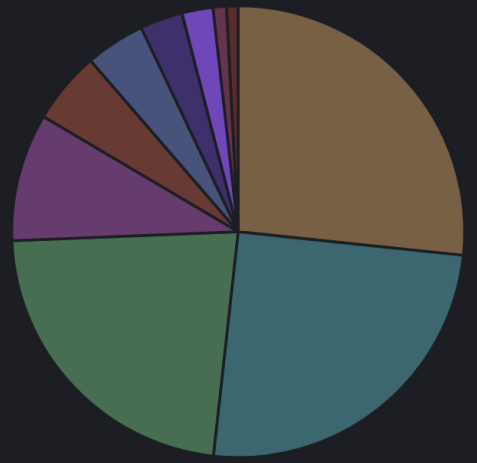
stashcache.acc_Audit

Full screen | Share | Clone | Reporting | Edit

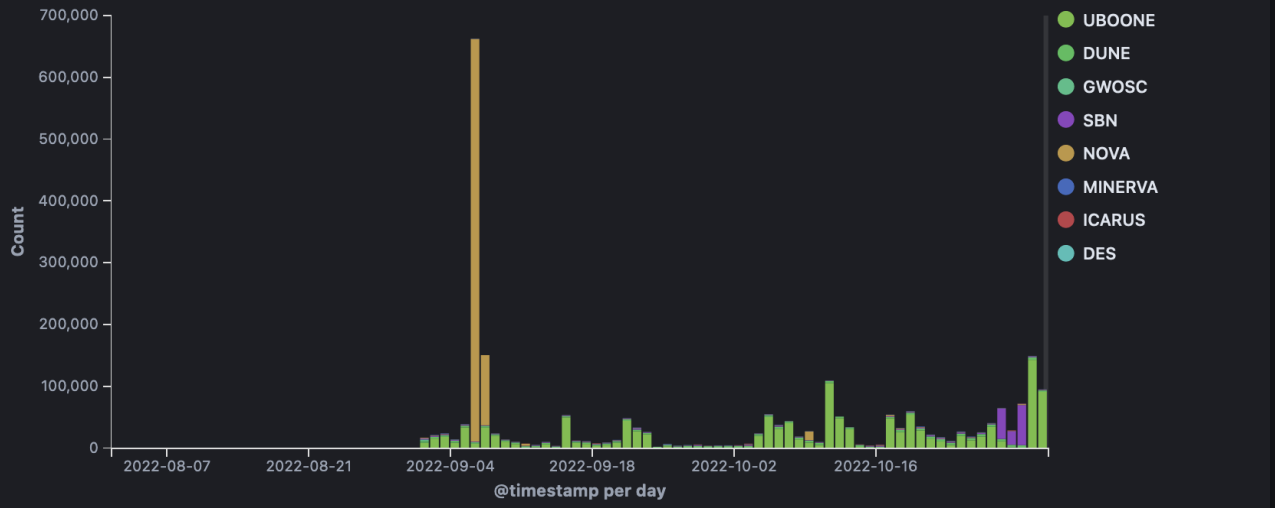
VOs



Sites



VOs requests daily



[StashCache]FileAccessCount

filename.keyword: Descending	Max accessCount
/stashcache/osgconnect/public/rynge/test.data.cinfo	629,009
/stashcache/osgconnect/public/fandri/cacheTest/stashcache.edi.scotgrid.ac.uk.cinfo	290,136
/stashcache/pnfs/fnal.gov/usr/uboone/persistent/stash/wcp_ups/wcp/releases/tag/v00_10_00/input_data_files /XGB_nue_seed2_0923.xml.cinfo	124,516
/stashcache/pnfs/fnal.gov/usr/uboone/persistent/stash/wcp_ups/wcp/releases/tag/v00_10_00/uboone_photon_library.root.cinfo	97,026
/stashcache/osgconnect/public/dweitzel/stashcp/test.file.cinfo	94,319
/stashcache/pnfs/fnal.gov/usr/dune/persistent/stash/test.stashdune.1M.cinfo	70,926
/stashcache/pnfs/fnal.gov/usr/uboone/persistent/stash/wcp_ups/wcp/releases/tag/v00_10_00/input_data_files/scn_vtx/t48k-m16-l5-lr5d-res0.5-CP24.pth.cinfo	54,706
/stashcache/pnfs/fnal.gov/usr/nova/persistent/stash/test.stashnova.1M.cinfo	54,349
/stashcache/pnfs/fnal.gov/usr/sbnd/persistent/stash/test.stashsbnd.1M.cinfo	53,174
/stashcache/pnfs/fnal.gov/usr/minerva/persistent/stash/test.stashminerva.1M.cinfo	53,007
/stashcache/pnfs/fnal.gov/usr/sbn/persistent/stash/test.stashsbn.1M.cinfo	52,497

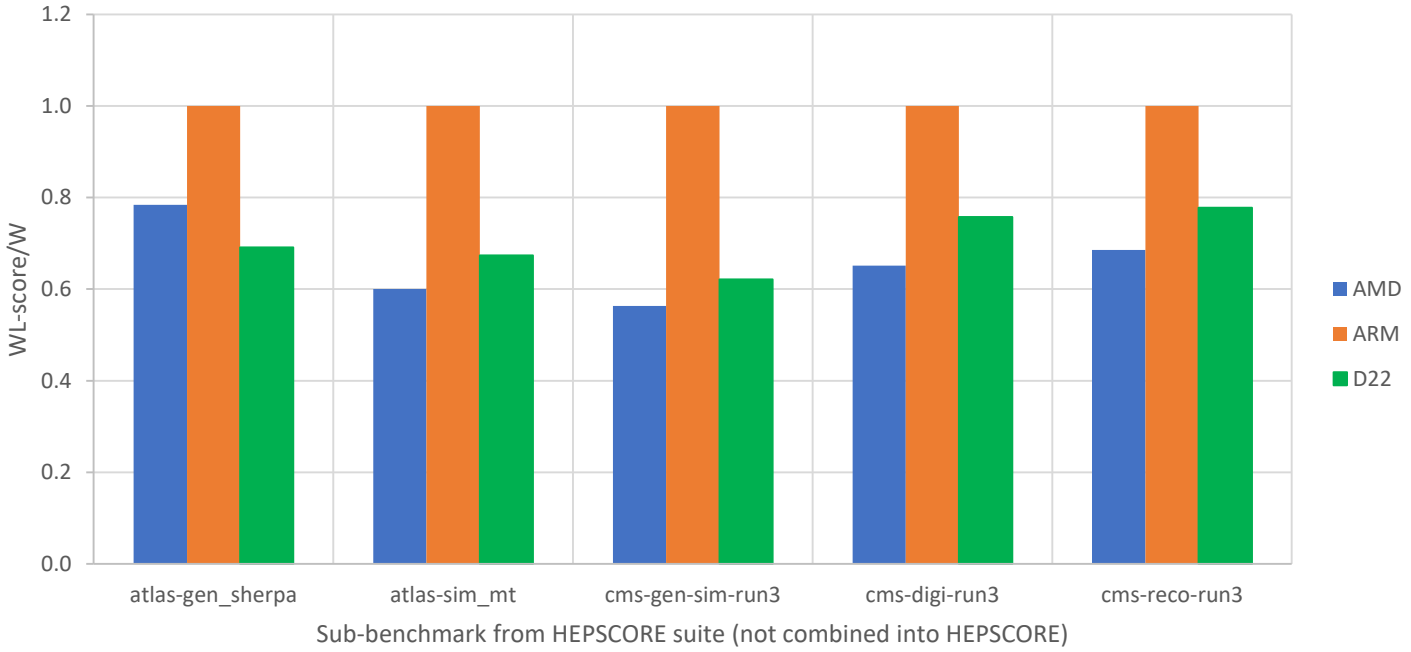
HEPSCORE + power efficiency

- Current "standard WLCG compute benchmark" HS06 has never been a perfect fit
  - Licensed, synthetic (no real HEP workloads), aging.
- WLCG "HEPSCORE" project developing a HEP-focused compute benchmark.

- RAL has been benchmarking new procurements with (draft) HEPSCORE benchmarks for > 1 year.

HEPSCORE @ Glasgow

- Glasgow has developing work in power-efficiency.
- Recent HEPSCORE suite [and other] benchmarking of matched ARM + x86_64 systems shows significant energy savings for ARM.
  - Little or no time cost is incurred as a result - for most benchmarks, the ARM system also performed better.



Future Work

- Comparison (side by side?) of VP and Xcaching at Oxford.
- Evaluation of the storage system scaling for VP services at Site (wrt site HEPSCORE or other compute capacity measure)

- Non-ATLAS solutions: VP is being integrated directly into Rucio, so should be available for any other Experiment using it.
- More sites moving to cache or low-storage solutions over EoY, start of next.

- HEPSCORE roll-out to other UK sites.
- Power-efficiency work beyond benchmarking [watch this space]