Data Management

(Aspirations and Reality)

28 Mar 2023

GridPP49/SWIFT-HEP

Sam Skipsey (he/they)

#translivesmatter

From last time (Nov 2022)

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 soon!]
- 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

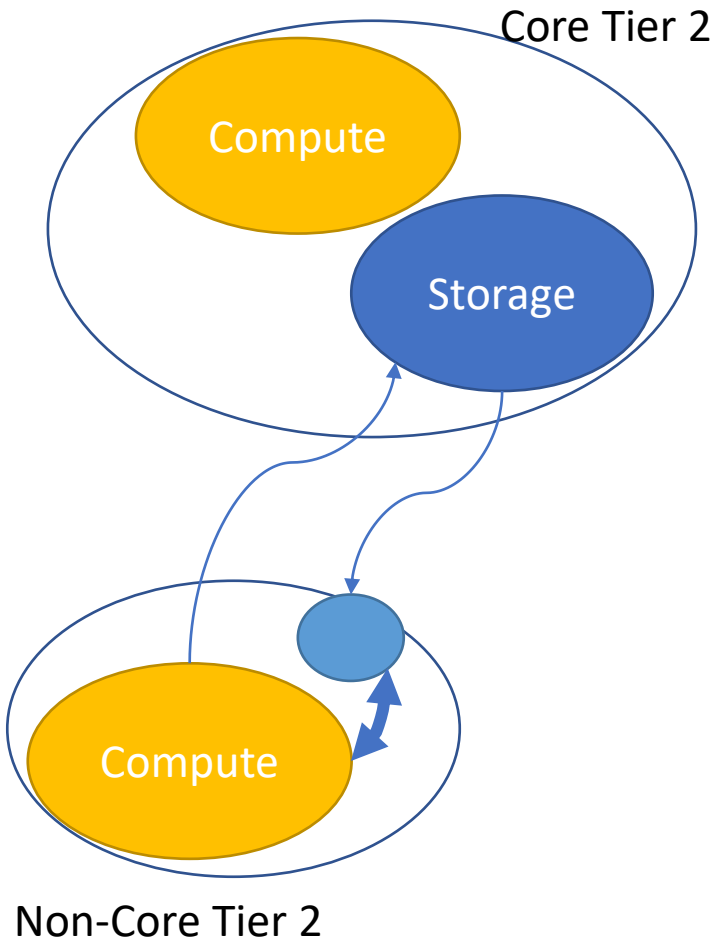
- (Core Tier-2s also exploring new technologies - xrootd/cephfs + xrootd/rados)

GridPP status ~now

Site	Storage (now)	Storage (if changing)		Network ** [Gb/s]
RAL-LCG2 (T1)	Echo (XRootD+Ceph)			~2x100Gbps (LHCOPN), LHCONE, Redundant 200Gb/s for RAL site to Janet
UKI-LT2-Brunel	DPM	XRootD+CephFS		40
UKI-LT2-IC-HEP	dCache			100
UKI-LT2-QMUL	StoRM (lustre)			100
UKI-LT2-RHUL	DPM	Storageless (SE – QMUL)		10
UKI-NORTHGRID-LANCS-HEP	XRootD+CephFS (+ DPM)	XRootD+CephFS (+dCache)		40
UKI-NORTHGRID-LIV-HEP	DPM	dCache	?	10
UKI-NORTHGRID-MAN-HEP	DPM	XRootD+CephFS		40
UKI-NORTHGRID-SHEF-HEP	Storageless (SE – RAL-LCG2)			10
UKI-SCOTGRID-DURHAM	DPM	(TBD)	?	10
UKI-SCOTGRID-ECDF	DPM	dCache	?	10
UKI-SCOTGRID-GLASGOW	Echo (XRootD+Ceph) + CephFS			20 (testing)
UKI-SOUTHGRID-BHAM-HEP	Storageless (SE – MAN + VP)			10
UKI-SOUTHGRID-OX-HEP	Storageless (SE – RAL-LCG2)			10
UKI-SOUTHGRID-RALPP	dCache			20
UKI-SOUTHGRID-SUSX	Storageless (SE – QMUL)			10

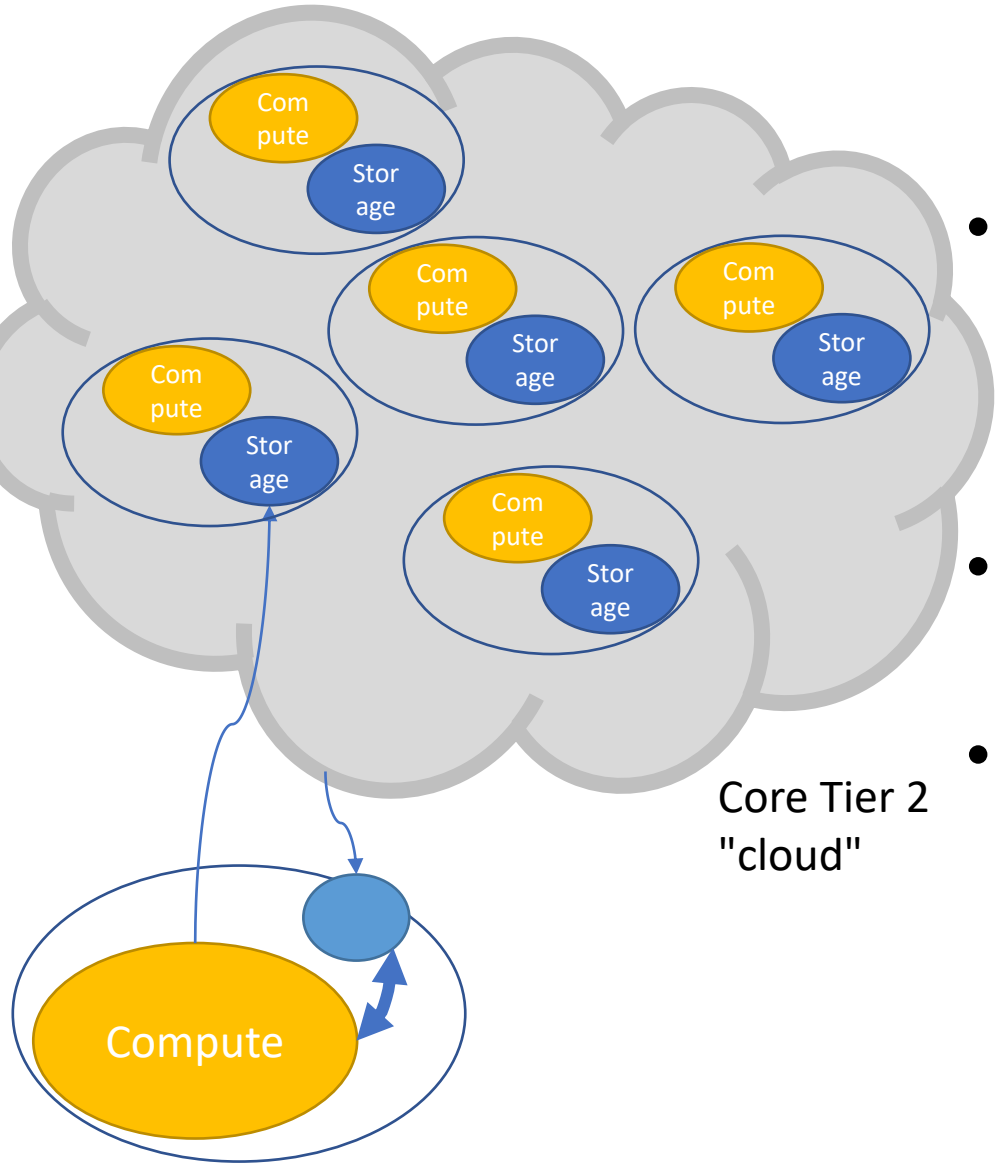
(This slide derived from James Walder's Xrootd +FTS Workshop talk slides)

GridPP Data Management "Aspirations"



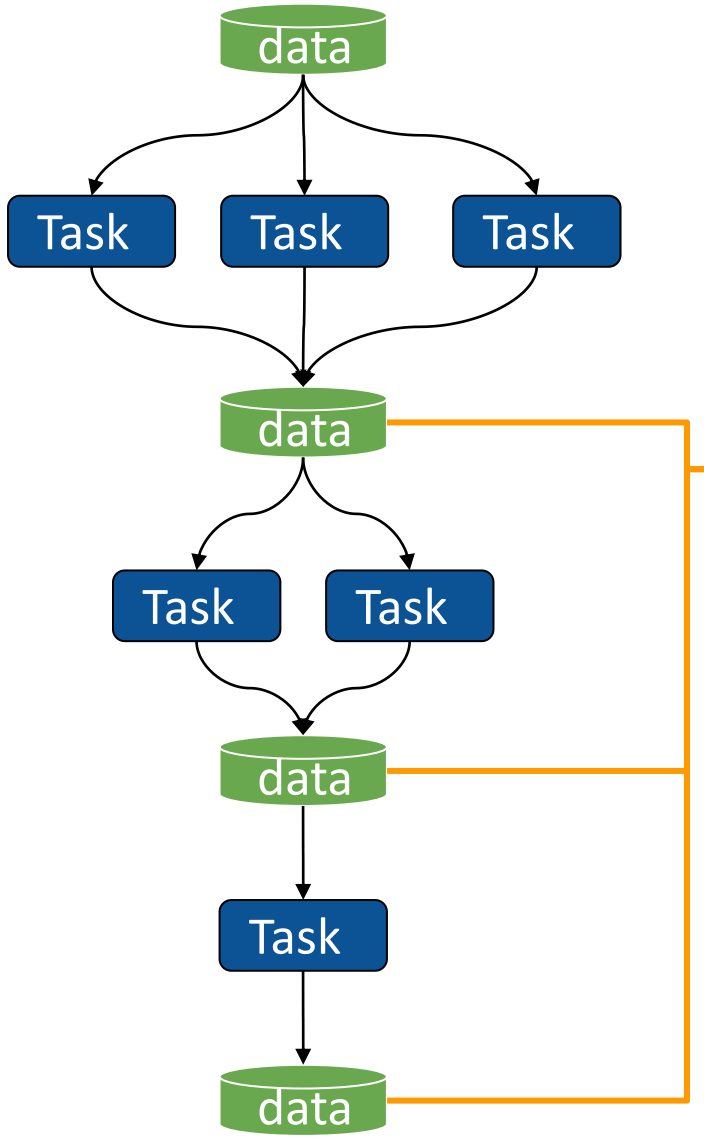
- More efficient storage use, with Core Tier-2s holding local storage volumes (~10PB+)
- More numerous non-Core Tier-2s host a local "cache"/"volatile storage".
  - Most job traffic occurs between local compute and "cache"
  - (stage outs need to happen to a Core Tier-2)
  - (cache needs filled from one or more remote storage elements)
    - Efficiency here is key - prestaging / prewarming cache significantly improves potential gains.

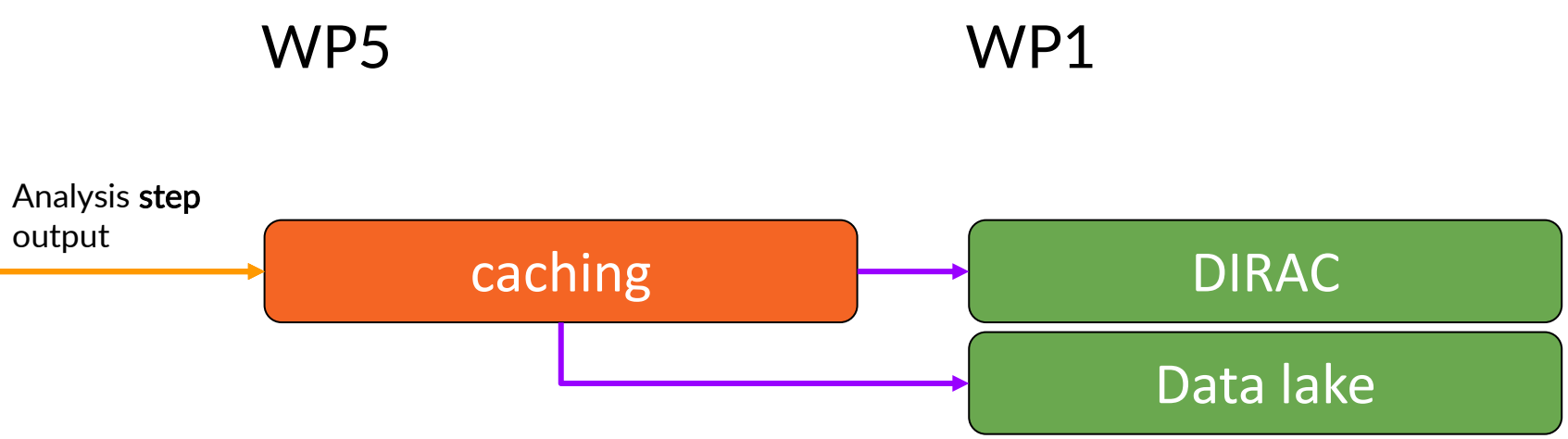
GridPP Data Management "Aspirations"



- Previous model risks overloading paired Core Tier-2s
  - Also increases risk for Core Tier-2 downtimes, as affects paired non-cores.
- This model diffuses risk… but needs support from Experiment tools to move data.
- ATLAS/Rucio "Virtual Placement"
  - Data moved from "anywhere"
  - *Prestaged so our caches are actually efficient (only cache useful data)*

SWIFT-HEP WP 5 data management (aspirations, thanks Luke K)





In a nutshell:

Want to store both intermediate and final analysis products in a sharable way on data lake.

Intermediate results should automatically be cleaned up after X days.

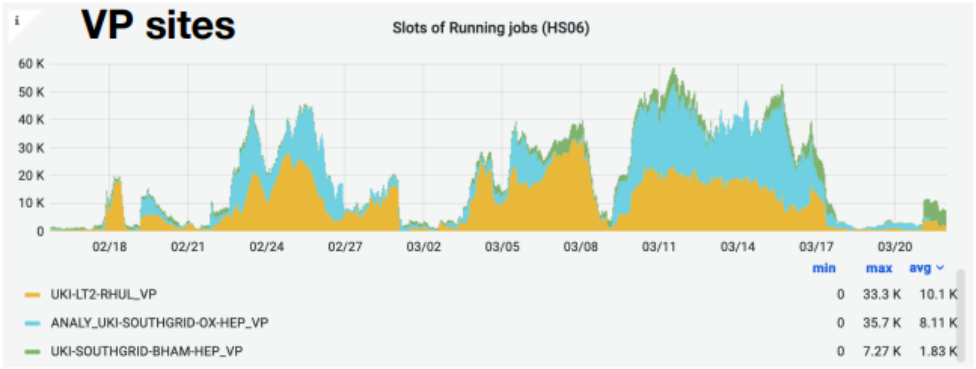
Technologies

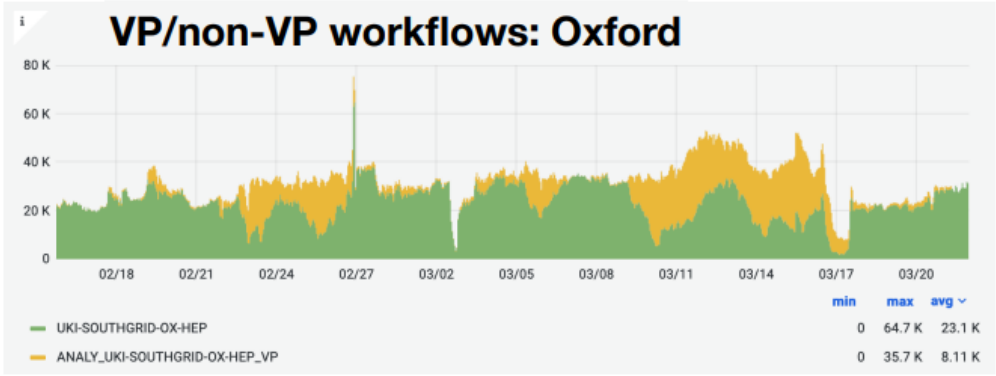
- GridPP:
  - Xrootd caches - essentially all our "new storage tech" relies heavily on Xroot.
    - "Xcache" [disk backed caching proxy]
    - Caching proxies [memory backed caching proxy]
    - Virtual Placement (working with *Rucio* plugin for ATLAS)
  - StashCache @ Edinburgh
- SWIFT-HEP @ RAL:
  - multiVO Rucio

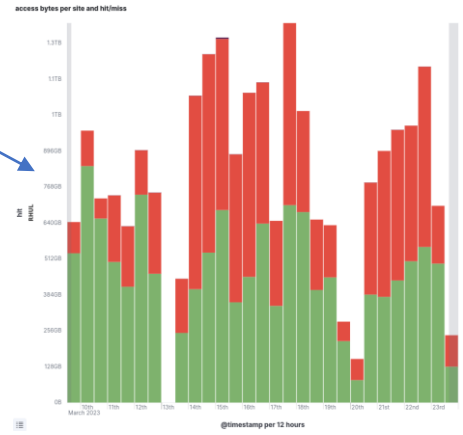
UK: Cache usage (VP and XCache)

- XCaches used:
  - (Currently) internally on each ECHO WN (at RAL)
  - Internally at a few sites
  - Stashcache (ECDF)
  - On the ingress for sites that have / are transitioning to become storageless (more likely useful for latency, than hit rate)
  - Also exploring the usage of Virtual Placement for ATLAS:
    - Analysis workflows, using partial file reads
    - Not using SLATE (for setup) (docker-compose, or manual)
- Example (last 21 days); For Oxford Xcache, usage from normal production workflows included

Access type	first accesses		following accesses	
Site	UKI-SOUTHGRID-OX-HEP	RHUL	UKI-SOUTHGRID-OX-HEP	RHUL
Count	408,641	38,837	166,415	125,474
Sum of b_hit	275.8TB	1.8TB	241.1TB	14.6TB
Sum of b_miss	92TB	894.6GB	4.4TB	10.4TB
Sum of b_bypass	0B	11.5GB	0B	3.9GB
Average percentage_read	96.972%	6.141%	75.159%	14.807%
Average rate	10.43	0.23	123.319	0.688
Average sparseness	96.843%	7.561%	90.07%	52.779%

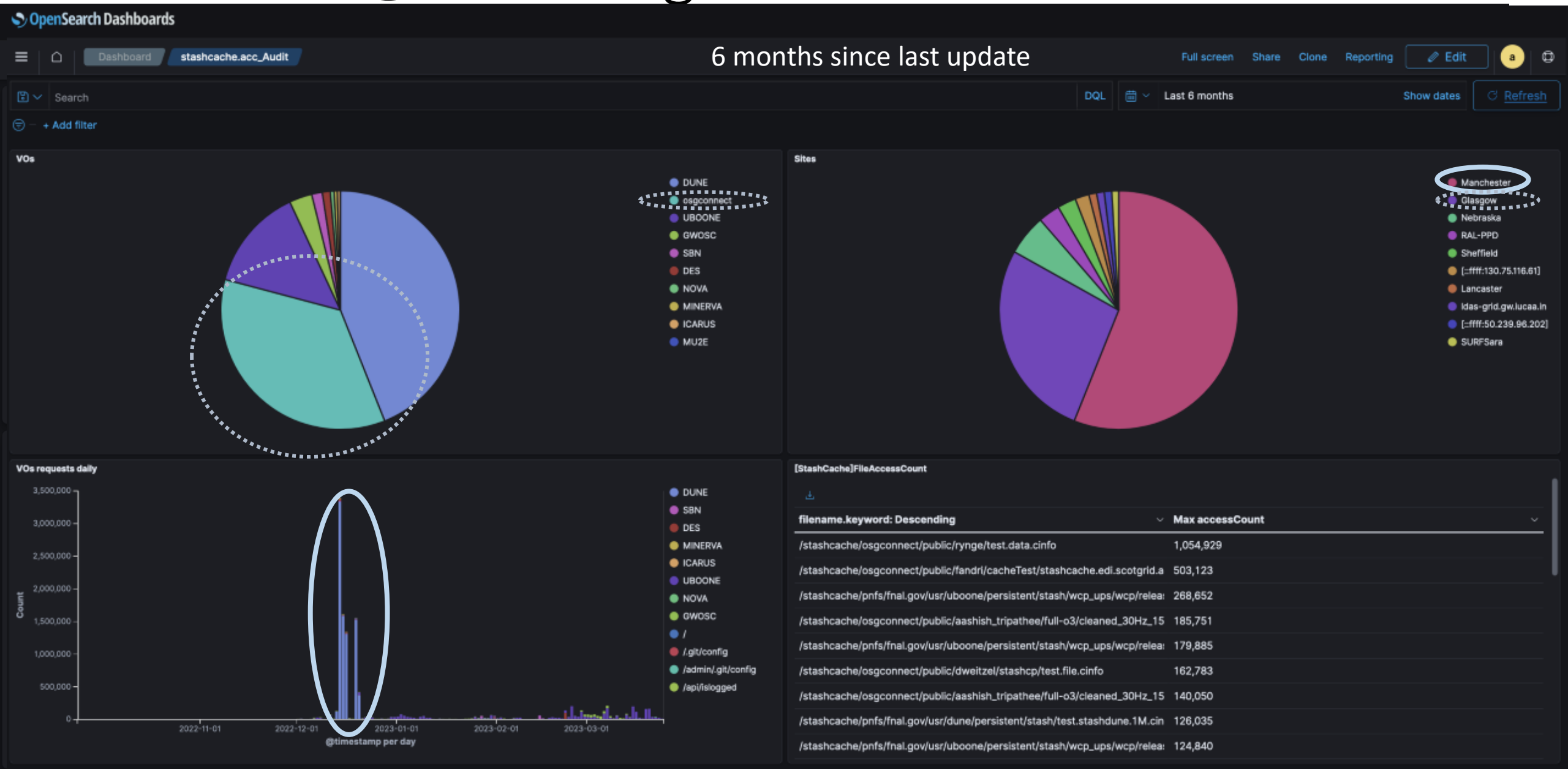






(This slide derived from James Walder's Xrootd +FTS Workshop talk slides)

StashCache @ Edinburgh



Multi-VO Rucio WebUI [George Matthews]

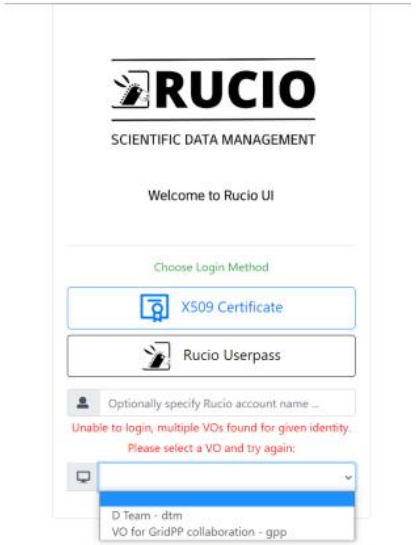
WebUI 2.0

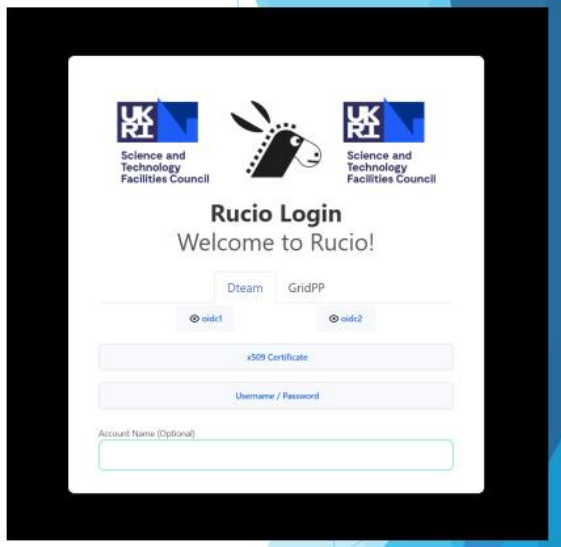
- Move to TypeScript and React
  - Typescript is a superset of JavaSc
  - React is a front-end JavaScript lil
- Move to a fully REST'ful architect
  - Remove direct access to the data
- Ensuring compatibility with Multi
  - Support Multi-VO login
  - Check other pages for compatibil
- Test Multi-VO components render

Science and Technology Facilities Council

Multi-VO login

- Tab selection for Vos
  - Clear and easy to use
  - Custom display names set as variables
- Request headers Multi-VO
  - Tabs change which VO is set in the X-Rucio-VO header
- Separate list of OIDC providers for each VO
- Jest Tests






```
[multi-vo]
REACT_APP_multi_vo_enabled = true
REACT_APP_vos = atl,ops,dtm
REACT_APP_vo_dtm = Dteam
REACT_APP_vo_atl = ATLAS
REACT_APP_vo_ops = Operations

REACT_APP_oidc_providers_dtm = oidc1,oidc2
```

Science and Technology Facilities Council

Multi-VO Rucio K8S Deployment (RAL)

- External access for Server final piece of puzzle
  - Ensuring secure external access to Rucio
  - Issues converting NodePort access (easy to setup and use) to DNS based ingress
- Otherwise stable deployment with supporting infrastructure deployed on the cluster for secret management, monitoring, messaging queues.

Thanks: Tim Noble

Multi-VO Rucio Database improvements (RAL)

- Current Rucio DB used connections are between 55-75 for a low-level deployment
- These are low in use frequency and low in load when used (TN has seen between 5 and 17 active at one time with current use) writing one to 100 lines at a time
- Working with Database Team to move the DB to a new DB deployment capable of increased number of connections to support scale tests in future

Thanks: Tim Noble

The other bit we don't talk about (QoS)

- SWIFT-HEP Rucio work also should look at "Quality of Service" awareness, transitions, management at sites.

- The problem is that we don't have a good model for what Quality of Service categories / distinctions the user community wants or cares about.
- (This includes WLCG who also don't have a finalised set of mappings…)

- Some input would be great from the assembled attendees!

Future Work (from November)

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

- Non-ATLAS solutions: VP is being integrated directly into Rucio, so should be available for any other Experiment using it. (via MultiVO Rucio?)
- 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]

Summary

- GridPP continues to transition (increasingly rapidly) to "storageless" sites configuration.
  - Driven partly by the ongoing DPM transition
- This results in a dependency on Xrootd in our destination configurations. (But Virtual Placement is a nice benefit of this w/ Rucio)
- SWIFT-HEP multiVO Rucio work complements this nicely, assuming VP works with it.
- Some unanswered questions remain: for example, whence QoS?