xrootd monitoring

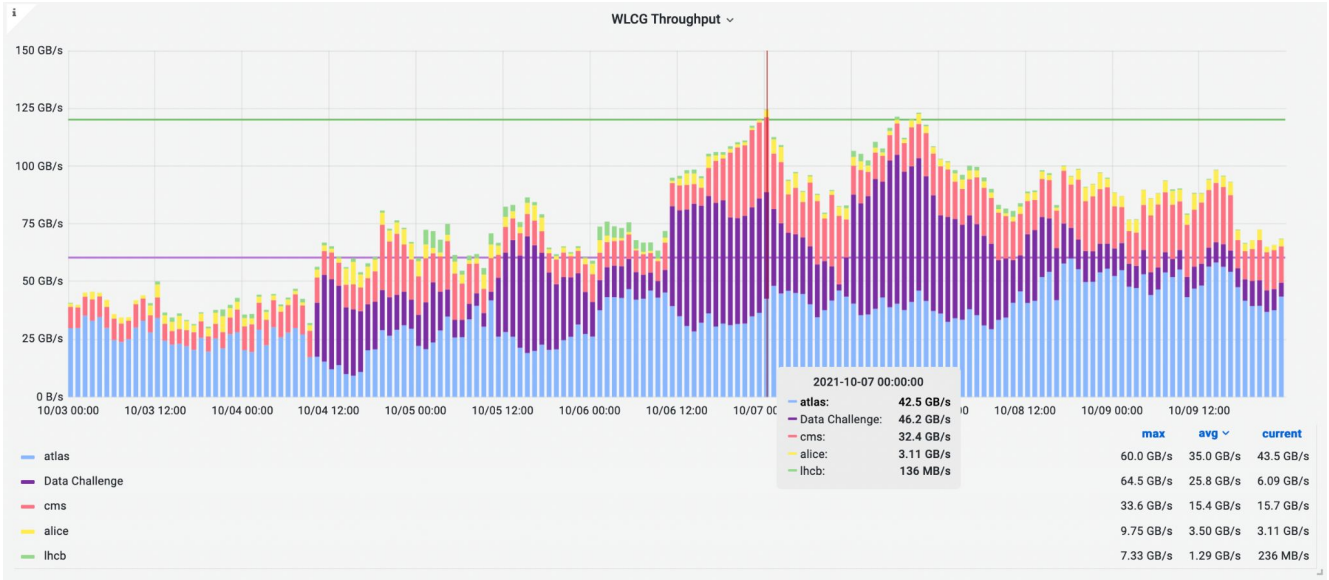
A. Forti, R. Currie

GridPP48

2 September 2022

Introduction

- DC Dashboard generally considered useful
  - Used for both DC and TC I & II
    - Required expertise in the data structures to plot consistent information
    - Many plots could only be static selections
    - xrootd traffic underestimated

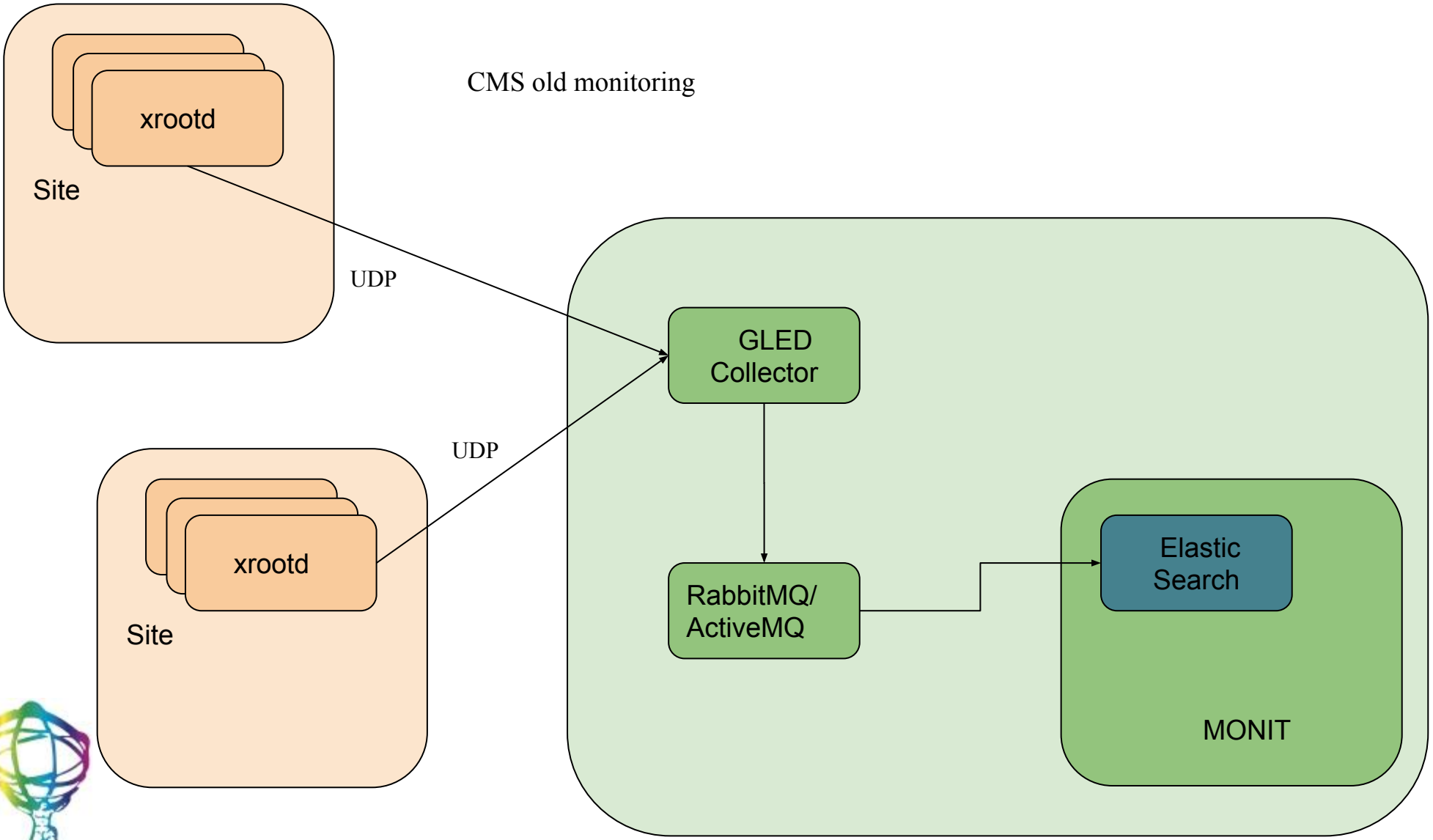


Example of static plot which pulls from different data sources and and took some expertise to put together a meaningful query

Monitoring TF

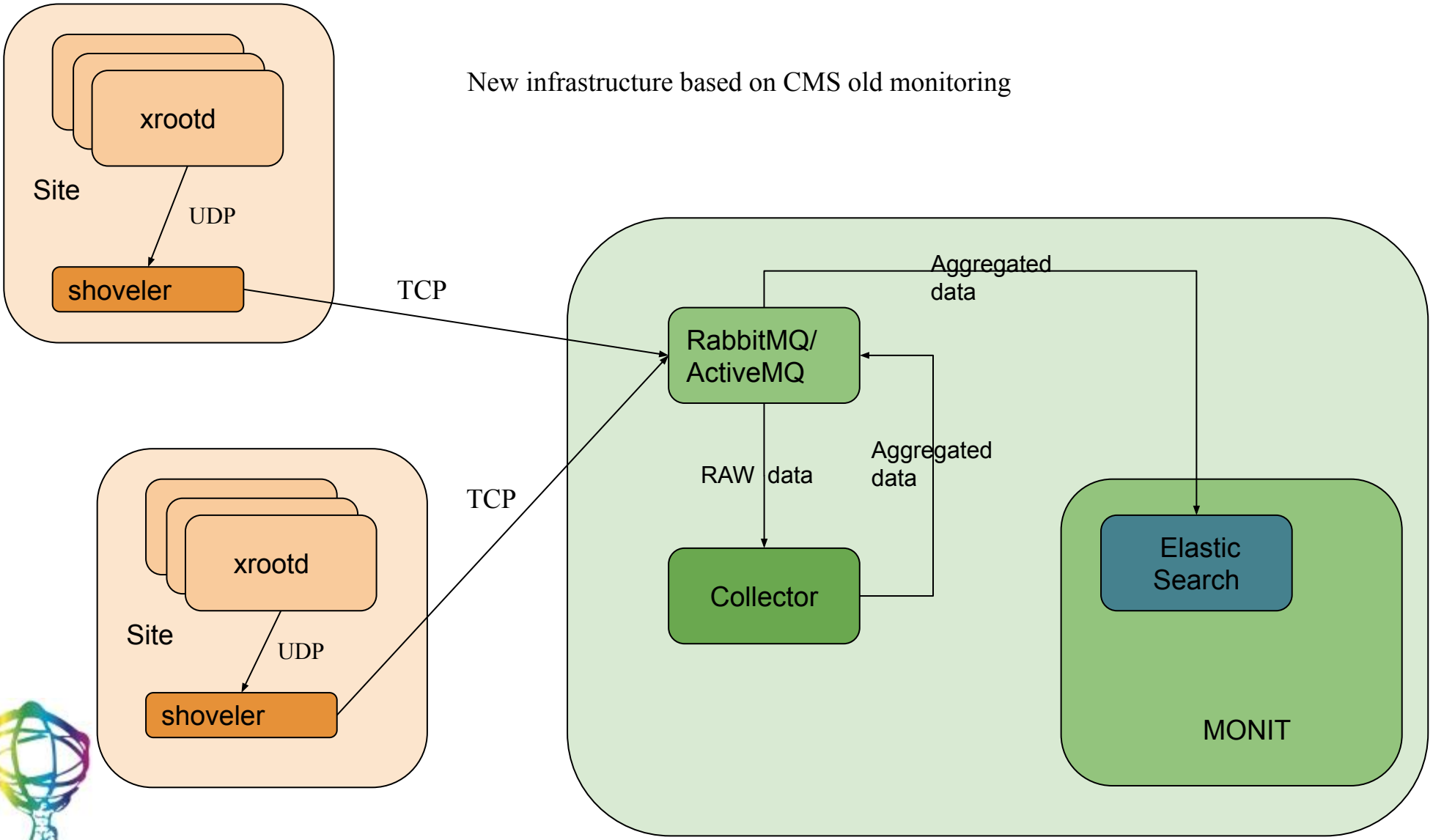
- Following DC recommandations WLCG monitoring TF
  - Re-structure xrootd monitoring infrastructure
    - Get the correct monitoring for xrootd protocol
    - Integrate Xcache traffic
  - Agree a common schema between the experiments, FTS & xrootd
    - i.e. consolidate in one data source that can be manipulated by users with selector buttons rather than experts
  - Refactor DC dashboard to use the new schema
  - Add site monitoring
- People
  - Borja Garrido, Rizart Dona, Julia Andreeva, Shawn McKee, Derek Weitzel, Alessandra Forti
  - Useful contribution from Rob Currie and Katy Ellis

Restructure Xrootd mon



Restructure Xrootd mon

New infrastructure based on CMS old monitoring



Shoveler

- UDP -> TCP translation service written in go
- Easy to install
  - Official notes use containers but….
  - 1 rpm, 1 yaml config file, 1 extra line on every xrootd server that needs to report
- Can send to 2 services (not clear why not more)
  - Local and remote monitoring
  - A limitation for multiple VOs that may want to send data to their monitoring from a shared storage
    - OSG doesn't care because they take care of it centrally

ActiveMQ Authz

- CERN ActiveMQ authz only accepts 2 methods
  - Username/password
  - Operator certificate
- Neither is convenient particularly for large scale publishing
- Need to discuss enabling tokens
  - OSG already uses them in their infrastructure
  - Good non user case to get acquainted with tokens
- It affects anything sites want to publish
  - Site network or tape monitoring
  - Benchmark results
  - …..

Collector

- Collector translates the shoveler output to a human readable format
- 73 monitoring fields
  - Originally designed for developers not for accounting
- Output tailored for Elastic Search
  - To add to prometheus you need to write a parser
- RAW data still needs to be massaged to be used for higher level plots in in dashboards like grafana
  - RAW data needs to be understood first!
- VO field information embedded in the user authz - No authz - no VO name
  - x509 connections are ok (Manchester tests)
  - tokens need to be tested
  - Alice on EOS and current RAL CMS data don't contain this information

```
# data.operation_time
# data.read_average
# data.read_bytes
# data.read_bytes_at_close
# data.read_max
# data.read_min
# data.read_operations
# data.read_sigma
# data.read_single_average
# data.read_single_bytes
# data.read_single_max
# data.read_single_min
# data.read_single_operations
# data.read_single_sigma
# data.read_vector_average
# data.read_vector_bytes
# data.read_vector_count_average
# data.read_vector_count_max
# data.read_vector_count_min
# data.read_vector_count_sigma
# data.read_vector_max
# data.read_vector_min
# data.read_vector_operations
# data.read_vector_sigma
data.remote_access
```

Common schema

- Final goal is to plot xrootd traffic & FTS traffic on the same plot
- xrootd traffic and FTS are inherently different
  - Different fields are needed to calculate the throughput
- xrootd monitoring also doesn't have topology information that needs to be added from CRIC
  - Not all the experiments fill the parameters needed in CRIC
- Dcache output different from native xrootd
  - Not all xrootd fields are needed
- Need to agree a common set of fields to build the data source

Attribute name	Meaning	Type	Comments and derived attributes
dst_hostname	Primary attribute which allows to identify transfer destination	String	Derived attributes: Destination site, country, experiment site, tier, federation Resolved via topology system CRIC
src_hostname	Primary attribute which allows to identify transfer source	String	Derived attributes: Source site, country, experiment site, tier, federation Resolved via topology system CRIC Dst_hostname and src_hostname also allow to define whether access is local or remote, that is remote_access boolean attribute
protocol	Used protocol	String	Should agree on a common naming

- Not yet agreed
  - Discussion at the next GDB 14/9
  - Deadline to agree on 30/9
  - Then other discussions with devs

UK sites contribution

- Original testing was done using Alice resources at CERN.
  - Clearly not enough because Alice is peculiar and because it was internal to CERN
- RAL: Katy installed the shoveler on the CMS production system it is the only visible traffic and the analysis she's doing with kibana poses useful questions
  - See her CMS talk
- Manchester: installed this on a DPM testbed not much traffic but the advantage is that I can send data as a number of VOs
- Edinburgh: xcache and stashcache

Cache monitoring work at Edinburgh

- Develop/understand XRootD monitoring at Edinburgh.
  - Use-case very similar but different to WLCG XRootD site monitoring.
- Interested in Cache performance, optimization and real-time monitoring.
- Currently working to understand and calibrate the monitoring data we get.
- XCache service at Edinburgh deployed on similar HW to StashCache, but with ~16TB of storage.

XCache at Edinburgh

- Managed about ~2.5PB of traffic to/from site WN in 6mo. (16TB total)
- Reduced required bandwidth from SE by 40%
- Planning to route all ATLAS job access through XCache using PANDA

Cache Monitoring

- To try and cross-check our monitoring we're dividing everything into 3 datasets which should agree.

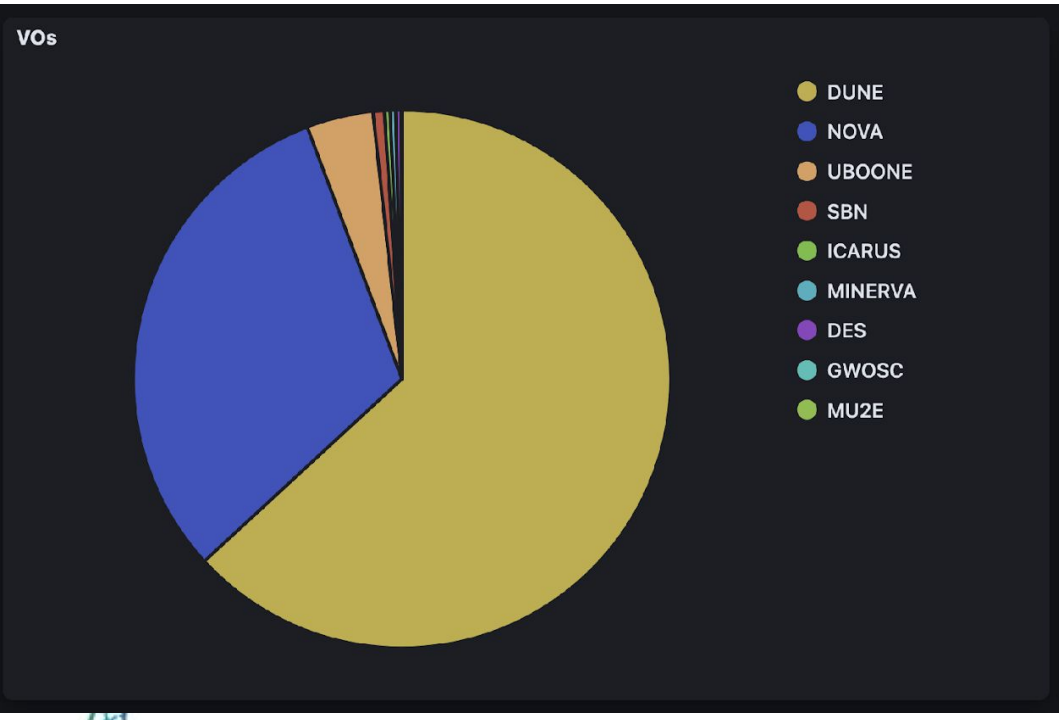
Node Exporter	XCache internal stats / logging	XRootD monitoring stream(s)
Proven industry standard tools using tested APIs. Reliable system and network monitoring stats.	"Must be True" data as this is the internal state of XRootD and the ProxyFileCache plugin. Internal state of the XRootD-PFC for each file is stored on disk in metadata. Have developed custom metadata tracker which records/reports state changes. Extracting transfer data from logs is valuable cross-check.	Closest to WLCG monitoring. Testing identified GDPR and configurability issues with WLCG shoveler service. – WIP Historically there have been concerns about accuracy of XRootD monitoring metrics.

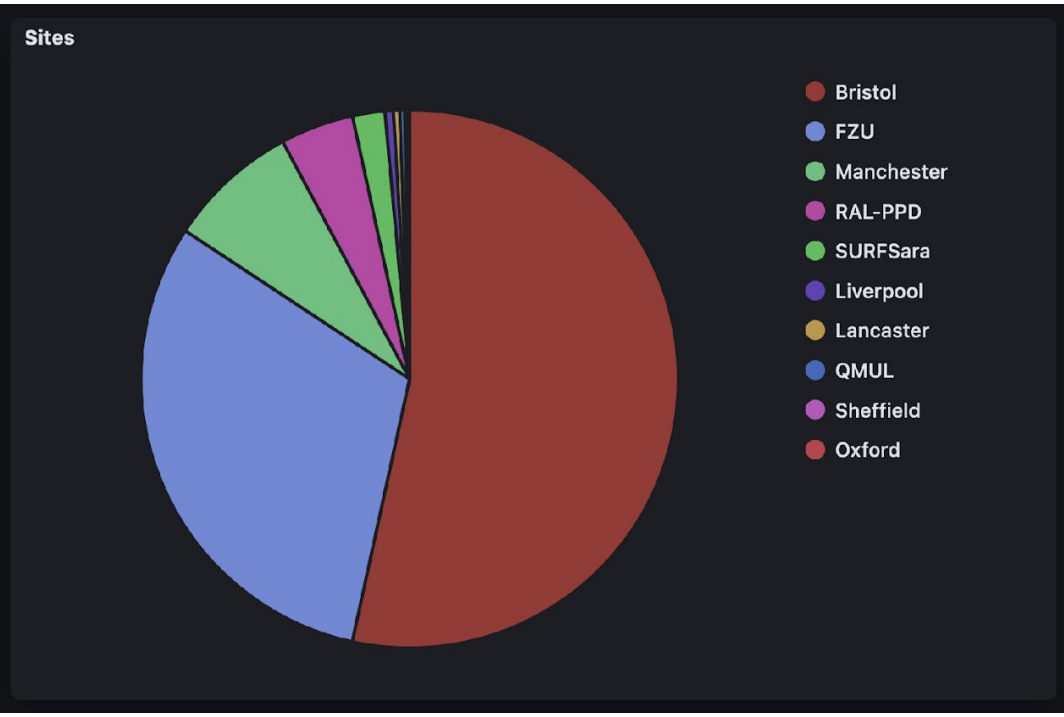
StashCache Service a Quick Summary

- A HTTP(S) based file caching network supporting partial file transfers.
- Like XRootD-PFC in some ways, like CVMFS in others.
- Service configuration via mostly “standard” XRootD config files.
- Access restricted to ports 1094,8000,8443,9619
- To join StashCache network Edinburgh registered as an “OSG site” in the UK with Geo-IP data.
- Currently StashCache service hosted on an old SE-node: 16Cores 24GB RAM ~10TB of storage with 10Gbps.

StashCache at Edinburgh

- Running for ~2months. Ingested ~2TB. Egressed ~120TB.
- Working to understand the impact of this work on final job efficiencies for DUNE and other VOs.





Cache Deployment and Monitoring

- Lots of metrics to track and being collected.
- Some metrics best monitored using Node+Prometheus.
- Other metrics require using OpenSearch+Dashboards.
- Now need to work out how to best to use/compare/display this data.
- Load on cache has been similar or higher to a well-tuned SE-node.
- These will need to be replaced/upgraded in time.

UK monitoring

- CERN is going to collect all the data from the storages and non-LHC VOs traffic can still be plotted
- Local monitoring from the storage to compare to the network monitoring is an important part
- Not clear if we want to have a centralised UK monitoring for non-LHC VOs
- The xrootd infrastructure as well as the monit infrastructure would have to be replicated in that case
  - OSG model

Conclusions

- Work is slowly progressing
- Infrastructure seemed the most important but there are a lot of details to iron out
- Multiple VOs difficult to handle if the VO name is not always included in the data
  - So far only x509 connections guarantueed
- UK infrastructure would be nice but lots of work without a monit infrastructure