Science and Technology Facilities Council

Echo 2022-23

Rob Appleyard and Jyothish Thomas

Echo – March 2023

- 57 PiB raw
- 41.5 PiB usable (for 8+3 pools)
- 16 XrootD gateway hosts
  - Unified WebDAV/vanilla XrootD type
- 2021 generation deployment imminent
  - 30PiB raw, 22 PiB usable

Ceph Team Staffing

- Staff turnover this year:
  - Matt (left last May)
  - Vijay (arrived May, left March)
  - Graduates (Josh and Bryce)

- Ceph Team is now Rob, Aidan, Jyothish and Tom
  - Rob – Manager/Ops/Legacy CASTOR
  - Aidan – Operations
  - Jyothish – Gateways/XrootD development (50%)
  - Tom – Systems Architect
  - Recruitment in progress

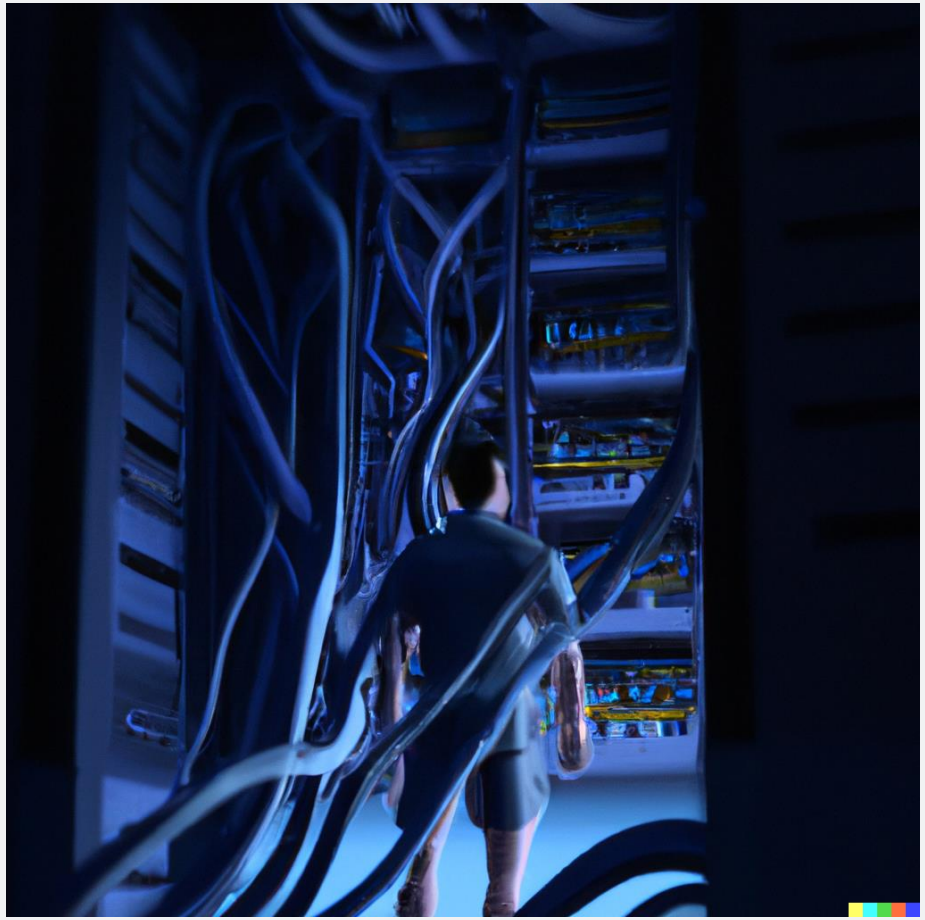


Image generated by Dall-E

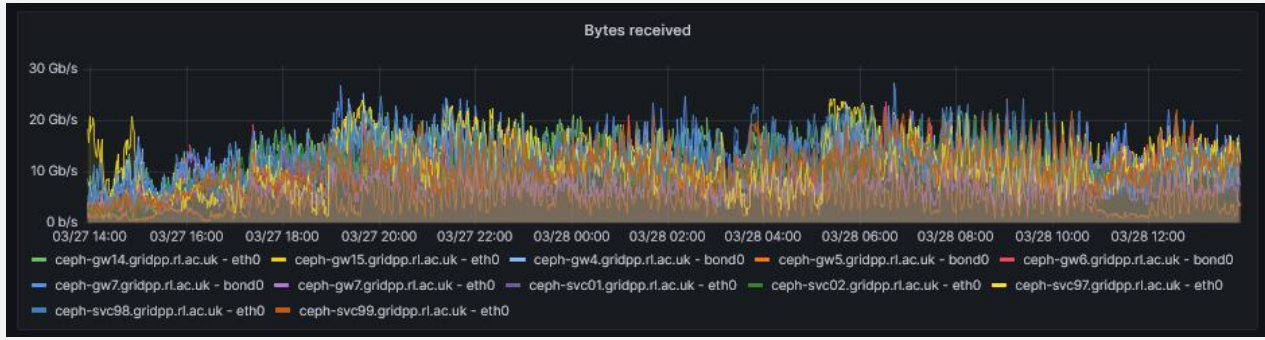
Science and Technology Facilities Council

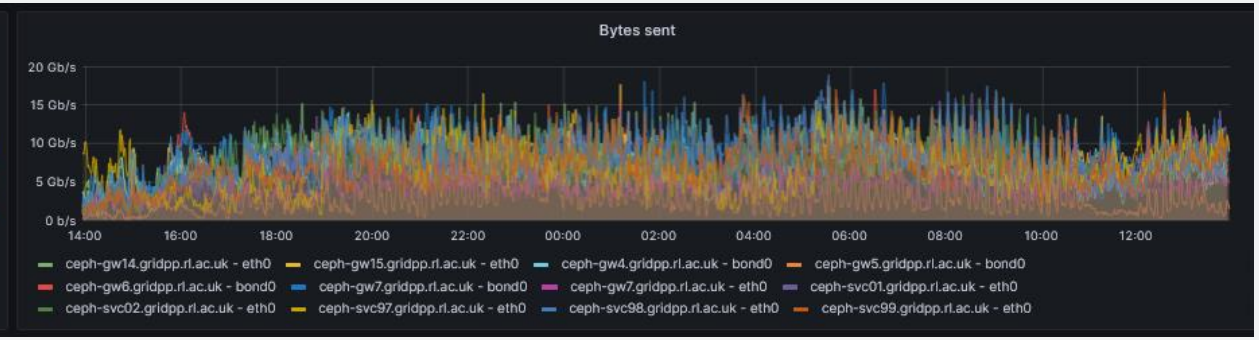
Echo Developments this year

- Retired grid-mapfile auth, added support for VOMS
  - Thanks to Matt and Vijay
- Retired 2015 hardware
  - First Echo storage generation, troublesome disks
- Deployed two large storage generations on the new network
- New test clusters

Echo on the New network

- (Almost) all new Echo hardware from the 2019 and 2020 generations is deployed on the new Tier 1 network
  - Some teething trouble, but now working well
- First gateways to be deployed on the new network in April





Science and Technology Facilities Council

Echo plans for the next year

- Deploy 2021…
  - SNs – expected June
  - Gateways – expected early May
  - Monitors – expected ~May
- Deploy 2022…
  - SNs – expected Q3
- Upgrade monitors to new hardware
  - Expected April
- High-level XrootD redirection
  - Jyothish will cover this

Science and Technology Facilities Council

Echo plans for the next year

- Regularise old-network 2020-generation SNs
- Upgrade Echo from Nautilus to Octopus
- Catch the White Whale
  - AKA ‘transition Echo to rack-level failure domains’
  - Rack-level redundancy
  - Fast, transparent reboots

Science and Technology Facilities Council

Other Ceph Business

- Our other clusters
  - Deneb
    - HDD CephFS for local facilities
  - Sirius
    - NVMe block storage for SCD Cloud
  - *NEW* Arided
    - SSD CephFS as supplementary storage for SCD Cloud
- Sirius and Deneb need to be upgraded to Octopus
- We use cephadm to manage Arided
  - We intend to migrate Sirius and Deneb to this too
  - …but not Echo

XRootD at RAL

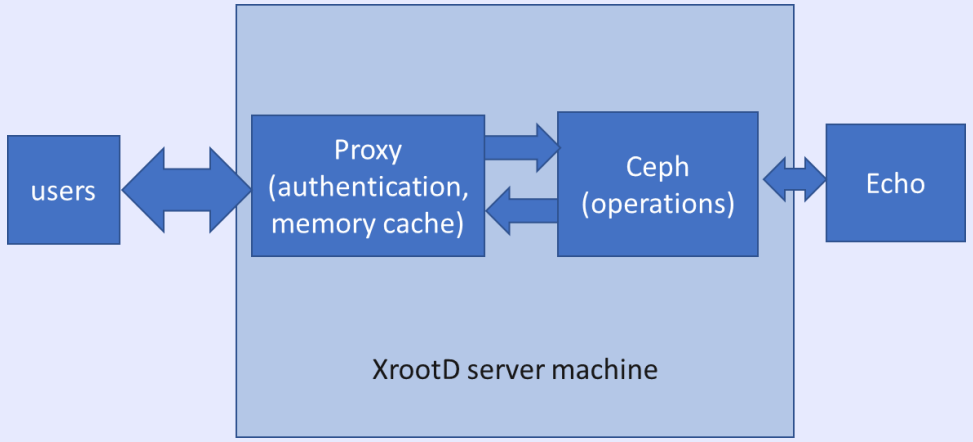
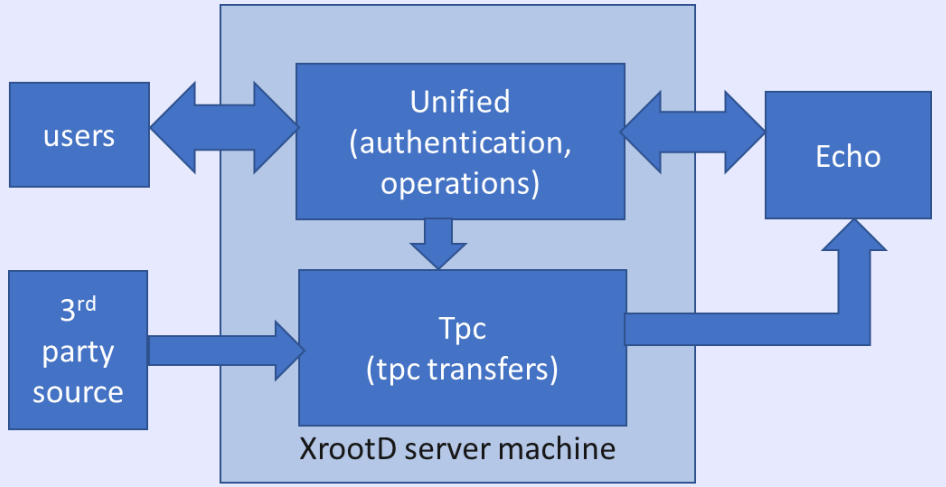
Jyothish Thomas
jyothish.thomas@stfc.ac.uk

Recent Developments

- Running XrootD 5.5
  - Gateways on 5.5.0, being upgraded to 5.5.3 shortly
  - Worker nodes temporarily on 5.3.3 while testing 5.5.3
  - XrootD 5.5.4 in testing
- Gateways are generally stable
  - All gateways now run in the unified/tpc configuration instead of proxy/ceph
- Vector read implementation initial testing passed (More in Alex's talk)
- Multistream read (xrdcp -S) issue identified and mitigated
- Issue with simultaneous reads identified and resolved
  - Implemented multi-buffers at XrdCeph layer
- Space information query functionality added

Science and Technology Facilities Council

Unified proxy and Ceph services

Before	After
• Cross-talk within xrootd • High CPU loads in split setups due to memory cache • Harder to debug/identify issues	• Easier to manage permissions • Better to manage logging • Prerequisite for some new features in XrdCeph (space reporting, speed up deletes)
	

Future Plans

Moving from resolving current (and old) problems to developing for the challenges of HL-LHCs data requirements

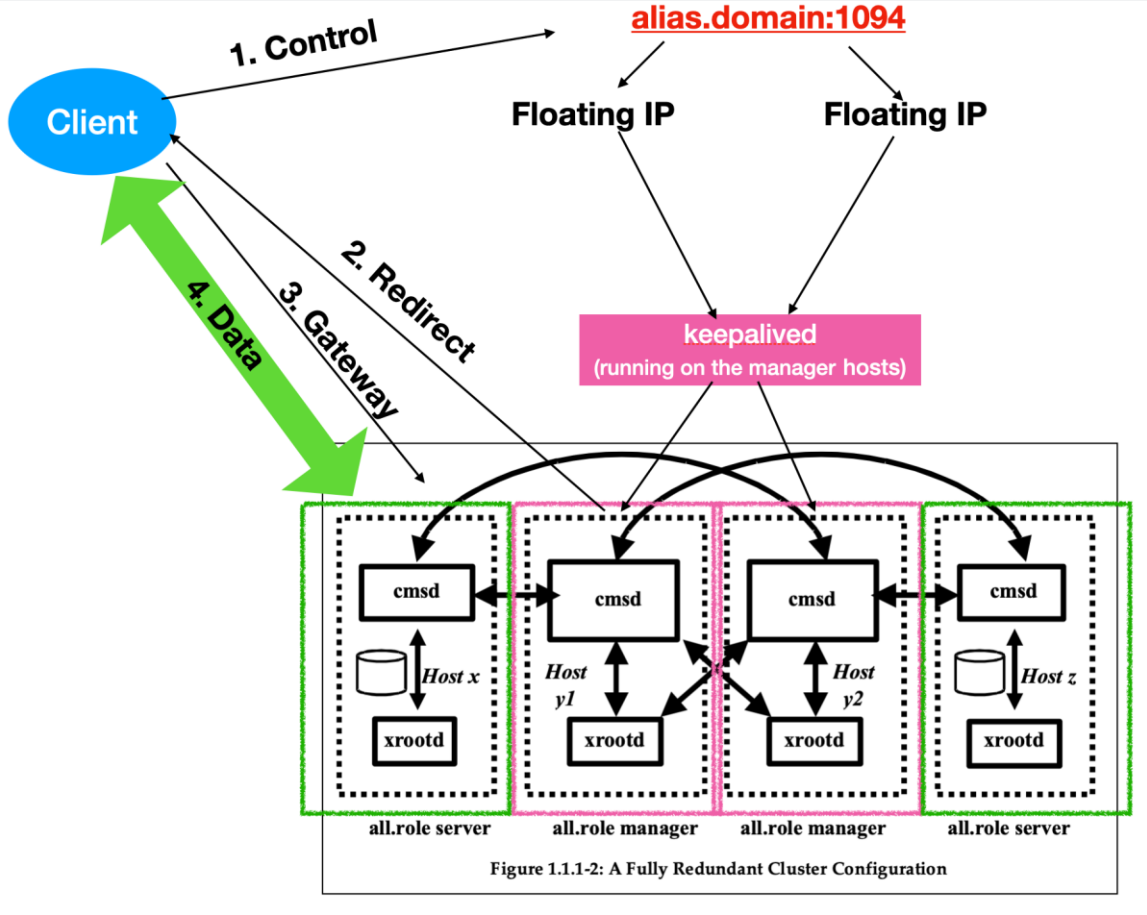
- XrootD Cluster Management Service redirector
  - Tested in pre-prod, to be deployed shortly in production
- XrootD 5.5.4
  - Increase socket and connection stability of the server

- Further integration of libradosstriper functionalities in XrdCeph
  - Improve I/O performance with atomic operations
  - Optimize for WORM(Write Once Read Many) type operations
- Automated testing and early deployment of future XrootD releases
- Test and prepare readiness for the Data Challenge ‘24 (25-30% of HL-LHC)

Science and Technology Facilities Council

CMSD – Site Level XrootD Redirection

DNS ROUND ROBIN	CMSD
• If a gateway fails then it remains in the alias until manually removed • Clients can bypass the round-robin by caching a particular gateway host. No active load-balancing is possible.	• Seamlessly deal with a failed gateway or intervene on individual gateways. • Evenly spread the load between the gateway hosts and automatically mitigate the pattern of 'hotspotting' • Reduce our dependence on the DNS provider. • Allow us to use a much longer TTL for our Echo alias, and so make Echo more resilient against any DNS issues



Checksum Server

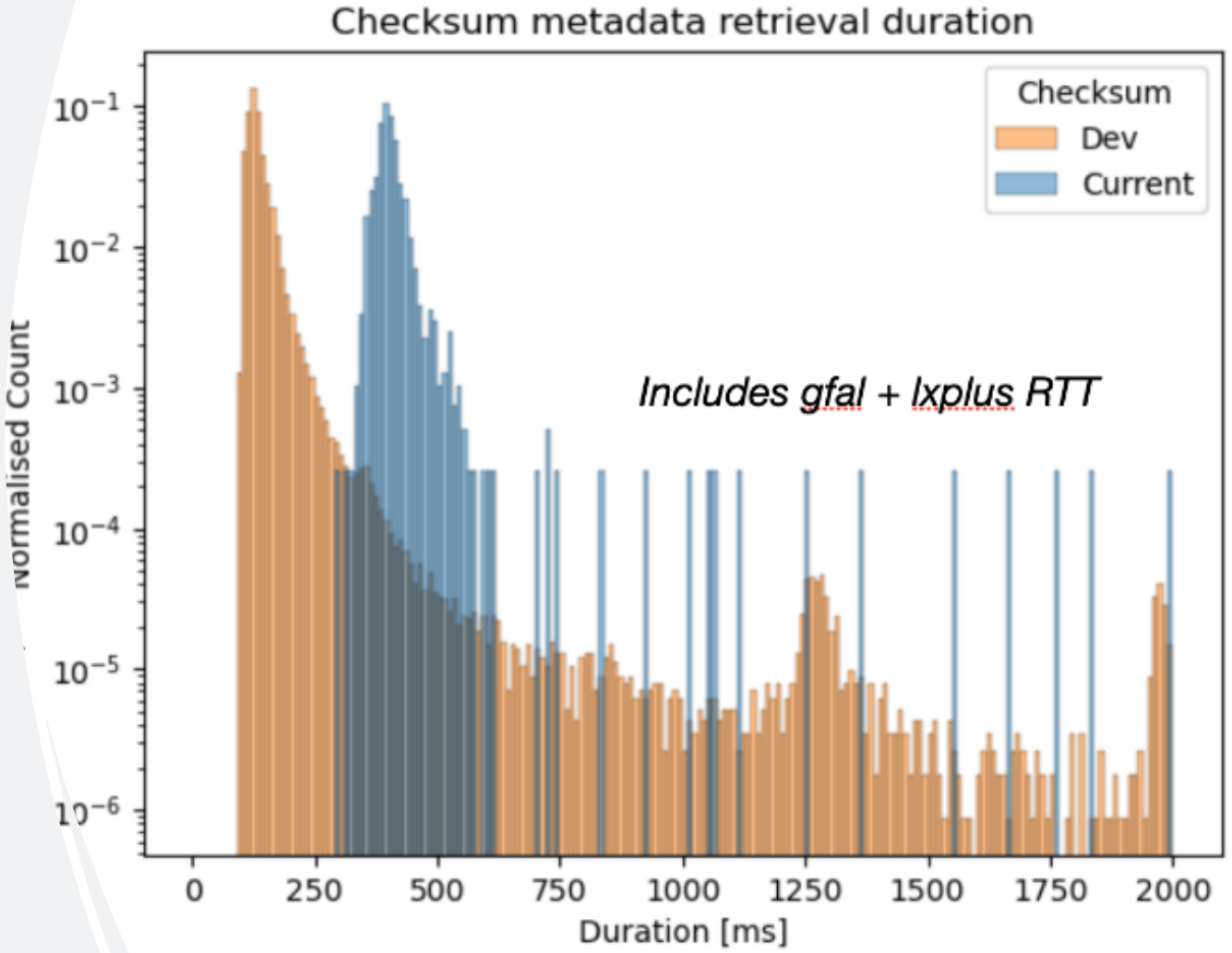
Motivation:

- Checksums script uses a significant amount of memory that is replicated every function call
- It also reinitializes a connection to the Ceph cluster on each call

The Checksum server will:

- Use a server to run the checksumming instead of a script
- Avoids reconnecting to the cluster for every checksum
- Reduce memory redundancy

ver	count	mean	std	min	25%	50%	75%	max
orig	2000.0	452.1	377.1	298.8	386.9	399.3	414.8	4259.8
test	572000.0	138.9	106.6	92.3	118.4	128.2	141.8	20618.5

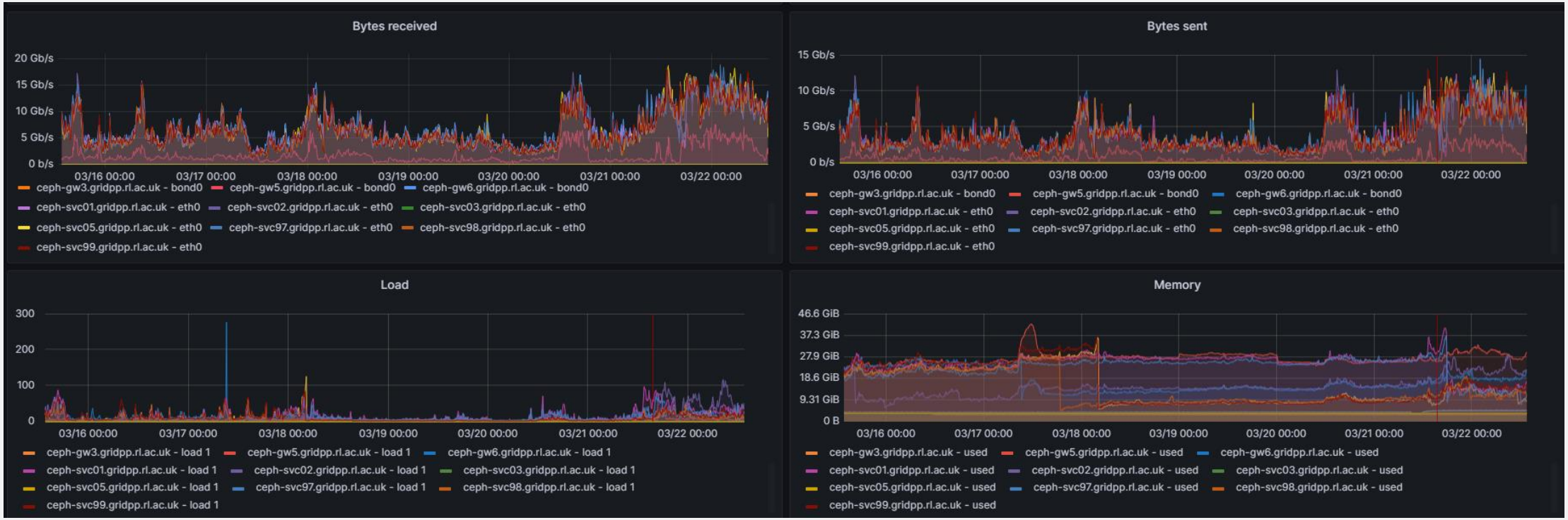


Science and Technology Facilities Council

Thank you

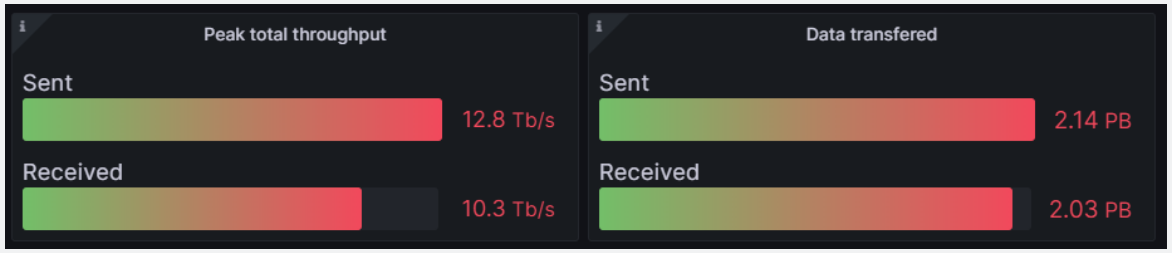
Science and Technology Facilities Council

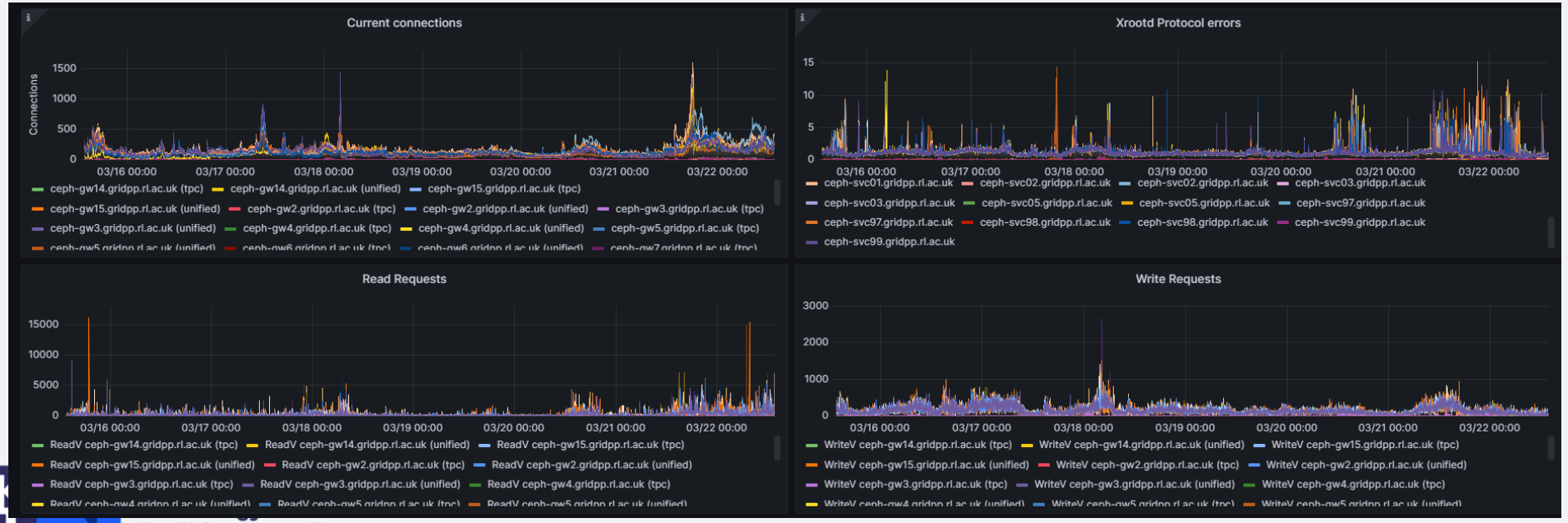
General monitoring



Science and Technology Facilities Council

Xrootd report





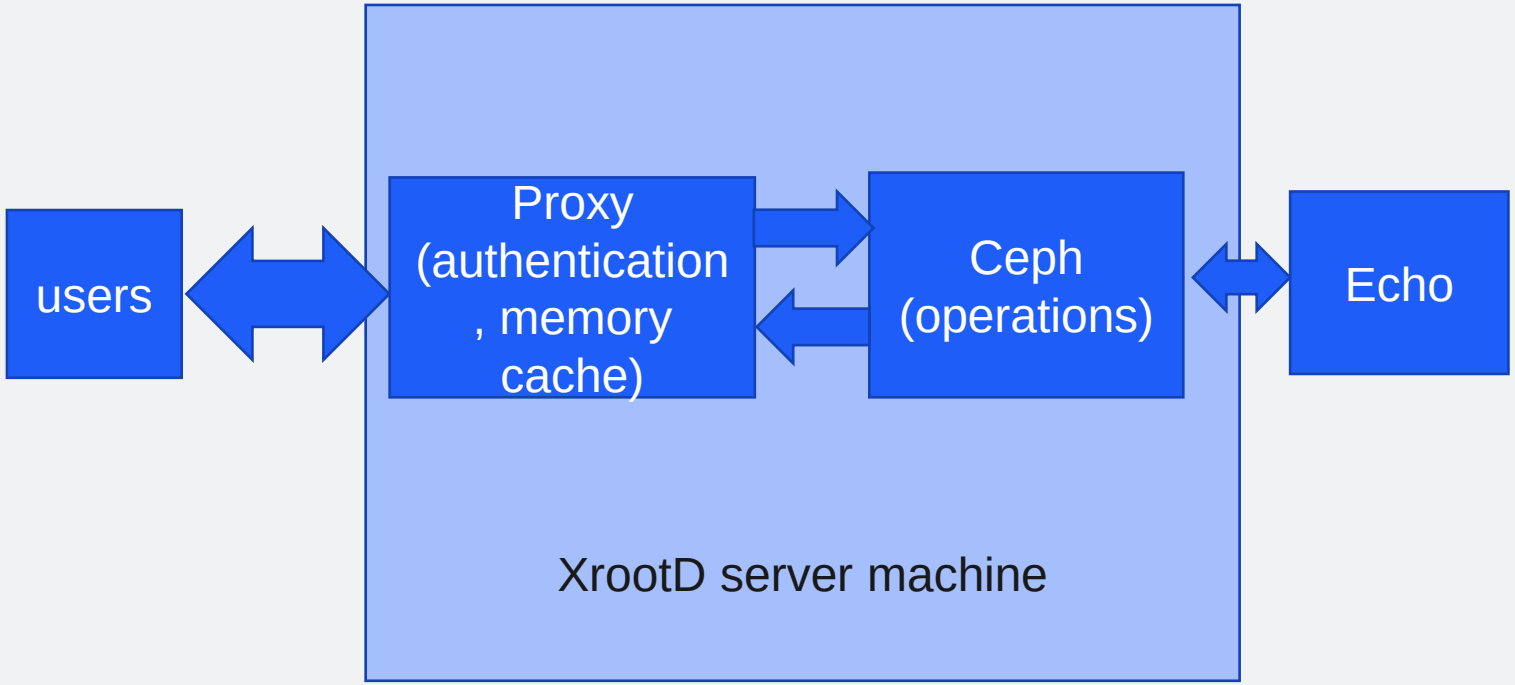
Facilities Council

XrootD-Ceph Buffered

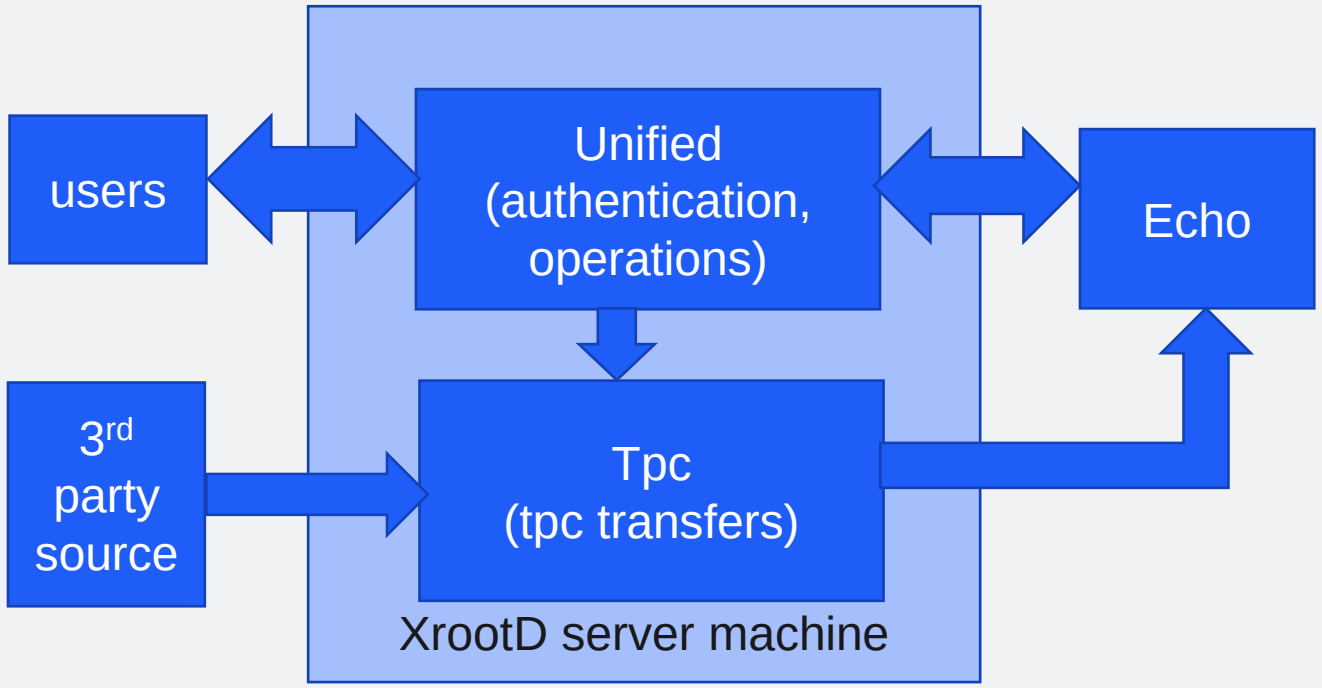
Asynch	Buffer	Root/webdav	Write	Read
No	No	root	28 MB/s	113MB/s
No	No	webdav	32MB/s	134MB/s
Yes	No	Root	27.7MB/s	146MB/s
Yes	No	webdav	37MB/s	128MB/s
No	Yes	Root	60.24MB/s	128MB/s
No	Yes	Webdav	61MB/s	146.3MB/s
Yes	Yes	Root	64MB/s	128MB/s
Yes	yes	Webdav	63MB/s	126MB/s

Tested on an EL7 VM with unified setup and XrootD 5.3.3. Asynch segsize of 67108864

The proxy/Ceph setup



The unified/tpc setup



Science and Technology Facilities Council