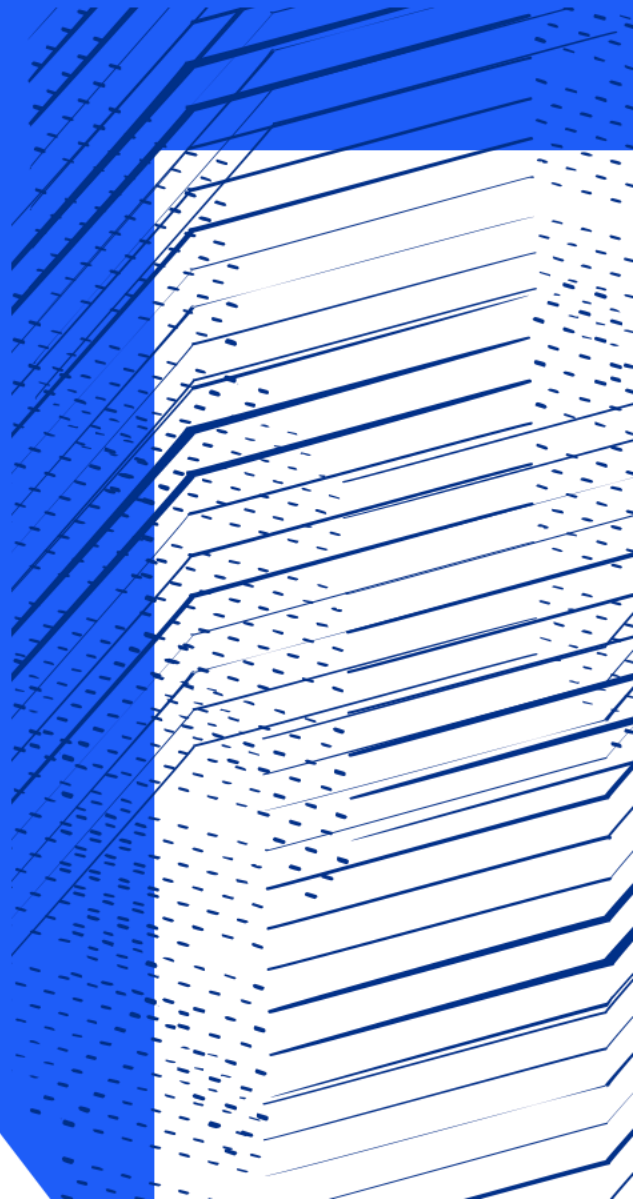
Science and Technology Facilities Council

UK Storage and XRootD

Alastair Dewhurst



GridPP

- GridPP is the project that provides the UK computing to LHC and HEP community.
- Led by Universities.
  - Want to bring benefits to their physicists.
  - Tier-2s often also co-host Tier-3 resources.
- Tier-1 has no local users.

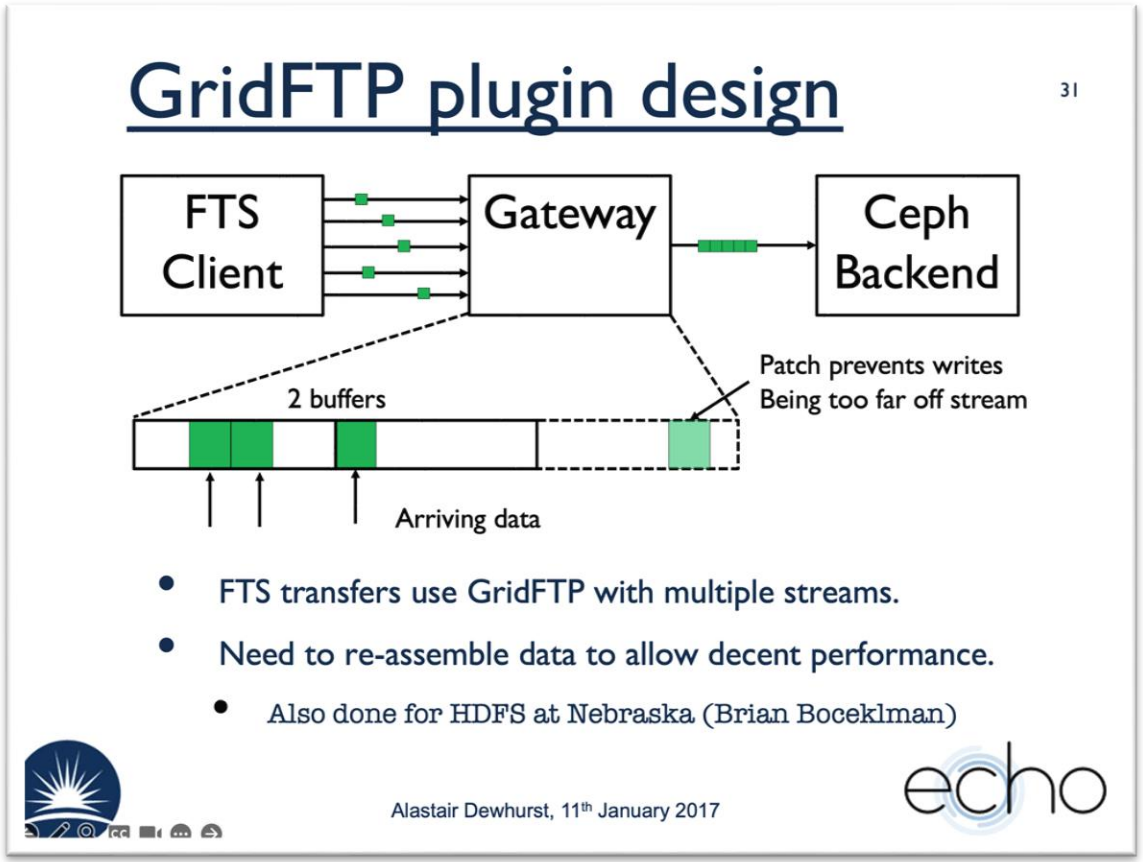
A Brief History…

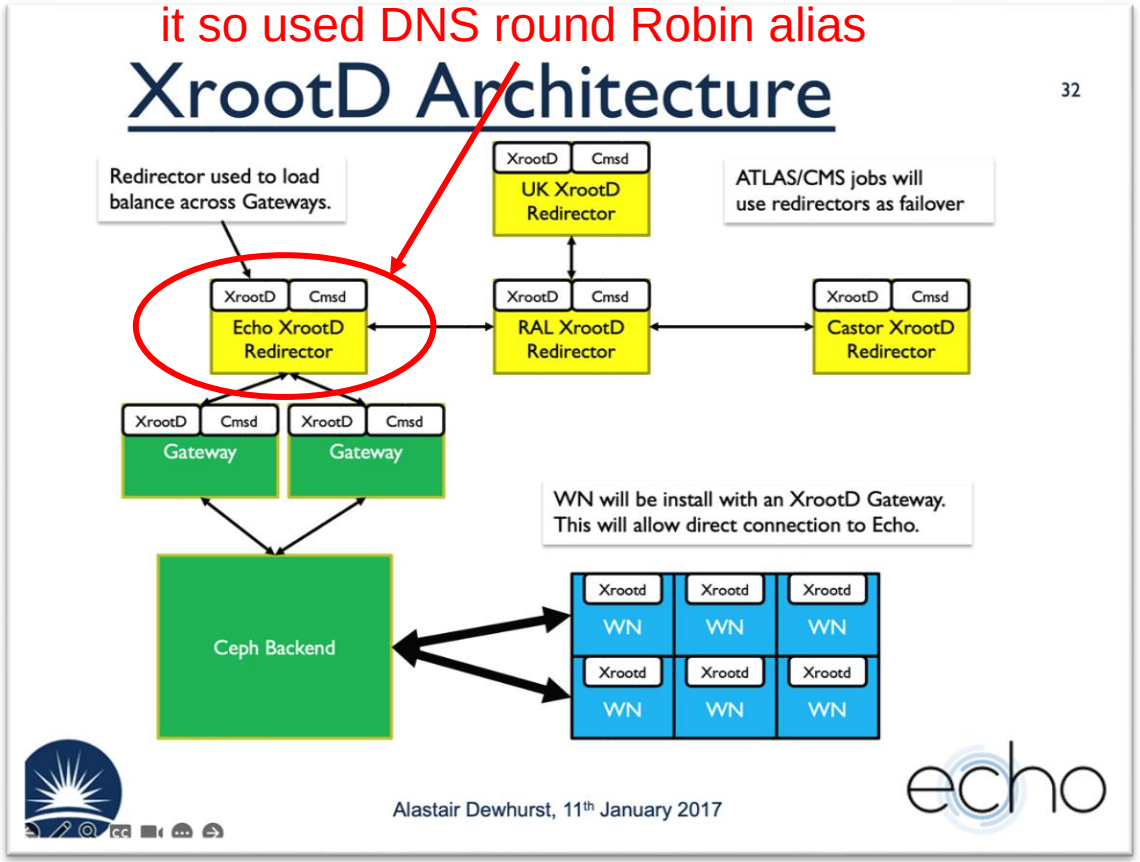
- In the beginning (< 2006):
  - RAL Tier-1 ran dCache
  - Most Tier-2 deployed DPM
- RAL couldn't get Tape to work so decided to switch to Castor.
  - Great Tape – Less Good Disk.
- When CERN switched to EOS we had a decision to make:
  - Worried about lack of community.
  - Wanted to avoid (Oracle) databases.
  - Wanted something that could handle hardware failure (without multiple replicas).
- Decision to use Ceph with XRootD layer on top of librados.
- Since 2018 there have been 5 core Tier-2 sites providing pledged storage.
  - Other Tier-2s can provide storage but officially only CPU.
  - End of DPM also meant many sites had to decide what to migrate to.
- CephFS + XRootD has become popular.

Ceph – Original Plan

- Initially relied heavily on GridFTP, many XRootD features not implemented.

Didn't have time to implement
it so used DNS round Robin alias

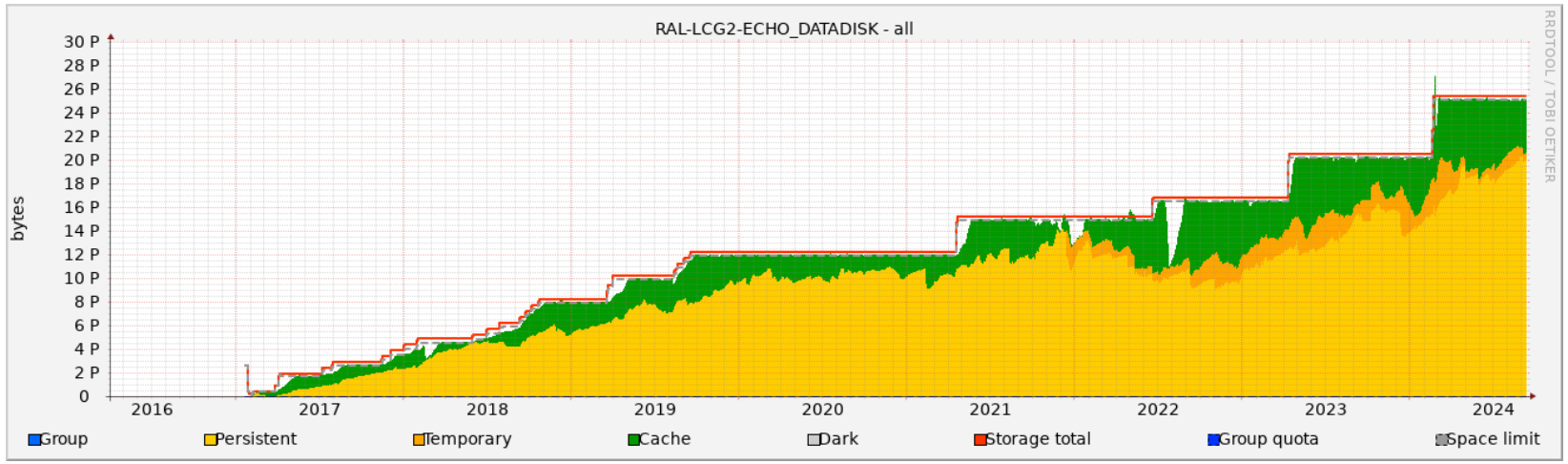




Storage Evolution

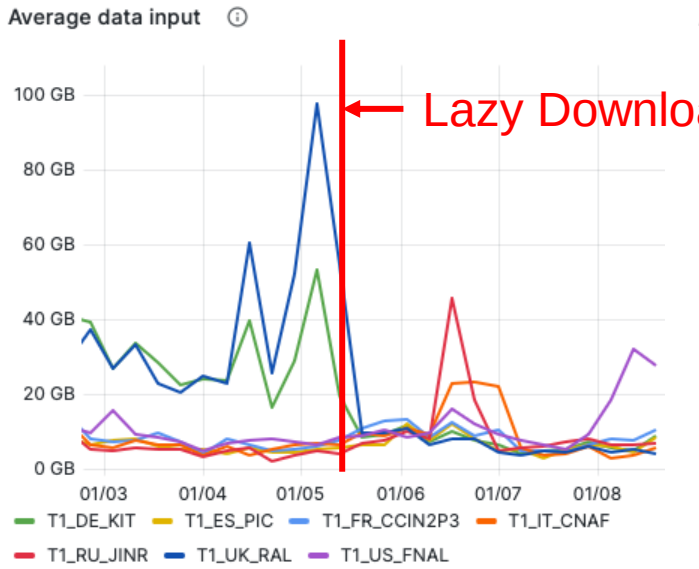
- Since 2017 some major evolution
  - GridFTP → Webdav for TPC
  - Significant increase in usage
  - Increase in directI/O usage
- New Tape service initially behind Echo (using multi-hop TPC).
  - Now evolving to separate endpoint.

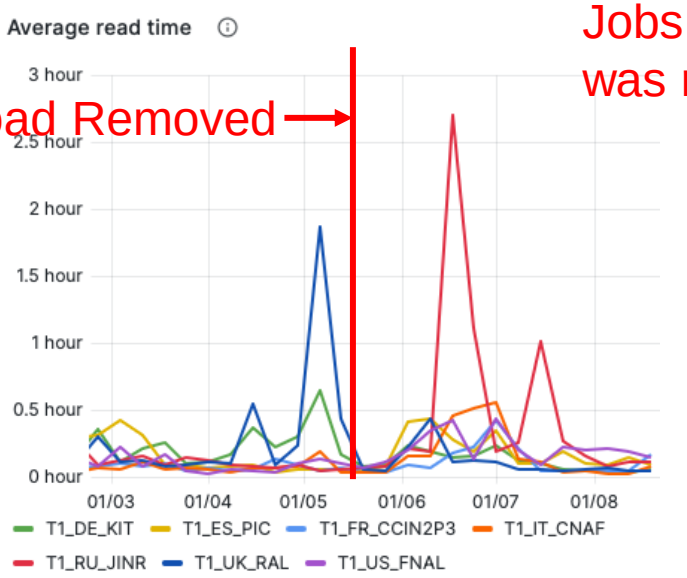
63PB of pledged Disk storage.
ALICE - 2PB
ATLAS – 26PB
CMS – 8PB
DUNE – 1PB
LHCb – 22PB
LSST – 2PB



Removal of Lazy Download (CMS)

- Lazy download, downloads entire CMS files in 128MB chunks while the job is running.
  - Assumed to work well with Echo 64MB chunk size.
- Improvements to directI/O meant it was no longer needed.





Jobs since Lazy Download was removed

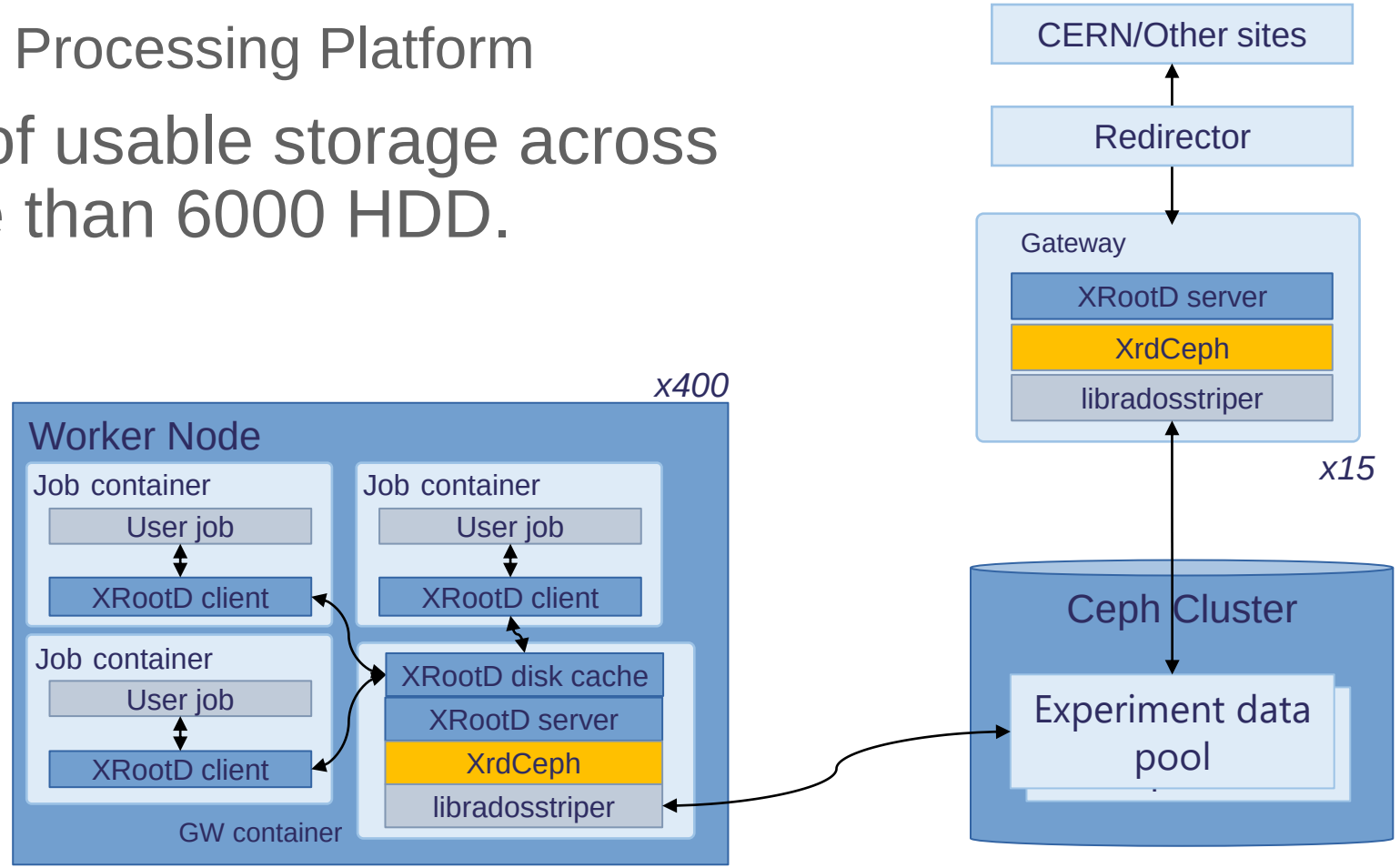
	CPU Eff	Failure rate
T1_US_FNAL	71.9%	19.89%
T1_UK_RAL	80.4%	6.11%
T1_RU_JINR	74.7%	8.05%
T1_IT_CNAF	76.2%	9.05%
T1_FR_CCIN2	81.1%	23.47%
T1_ES_PIC	67.7%	41.36%
T1_DE_KIT	77.8%	9.34%

Echo Today

- Tier-1 is very different from when Echo started.
  - Highly optimized Data Processing Platform
- Echo provides 73PB of usable storage across 268 servers and more than 6000 HDD.

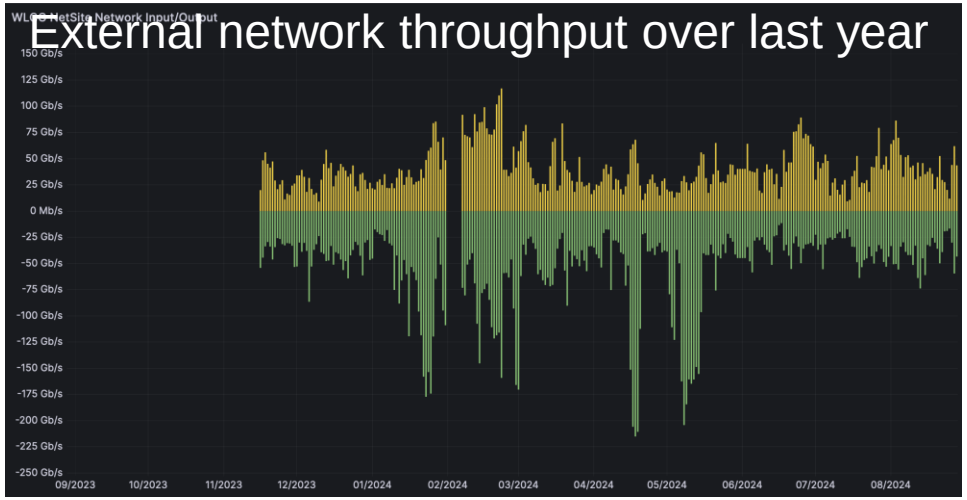
Significant XRootD expertise

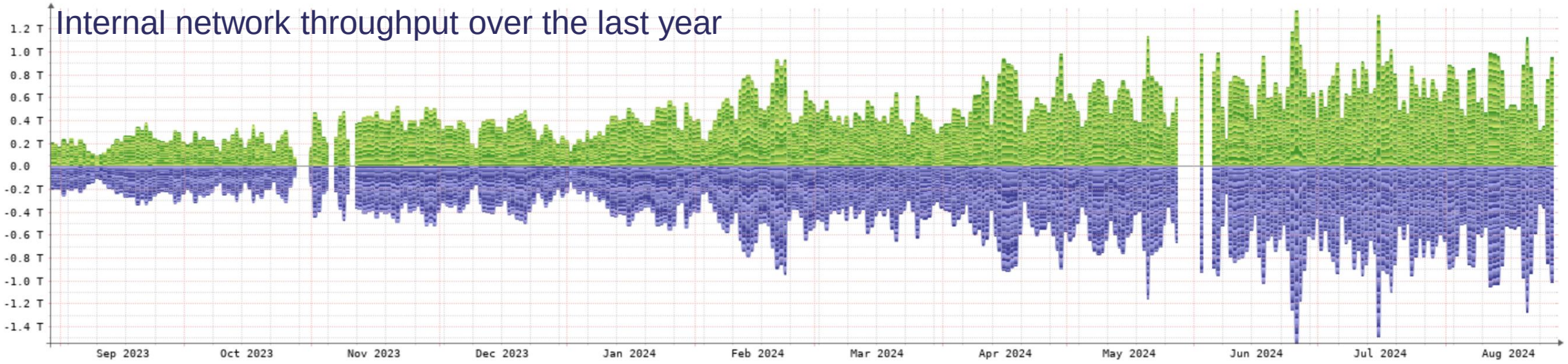
- Jyothish Thomas – XrootD Devop
- Ian Johnson – Developer (0.5FTE)
- Mariam Demir – Graduate (6 months)
- Alex Rogovksiy - LHCb Liaison
- James Walder - Former ATLAS Liason now UKSRC architect.
- Katy Ellis – CMS Liaison



Networking

- 200Gb/s link to CERN
- 400Gb/s link to JANET
- Leaf / Spine internal network provides non-blocking connectivity.





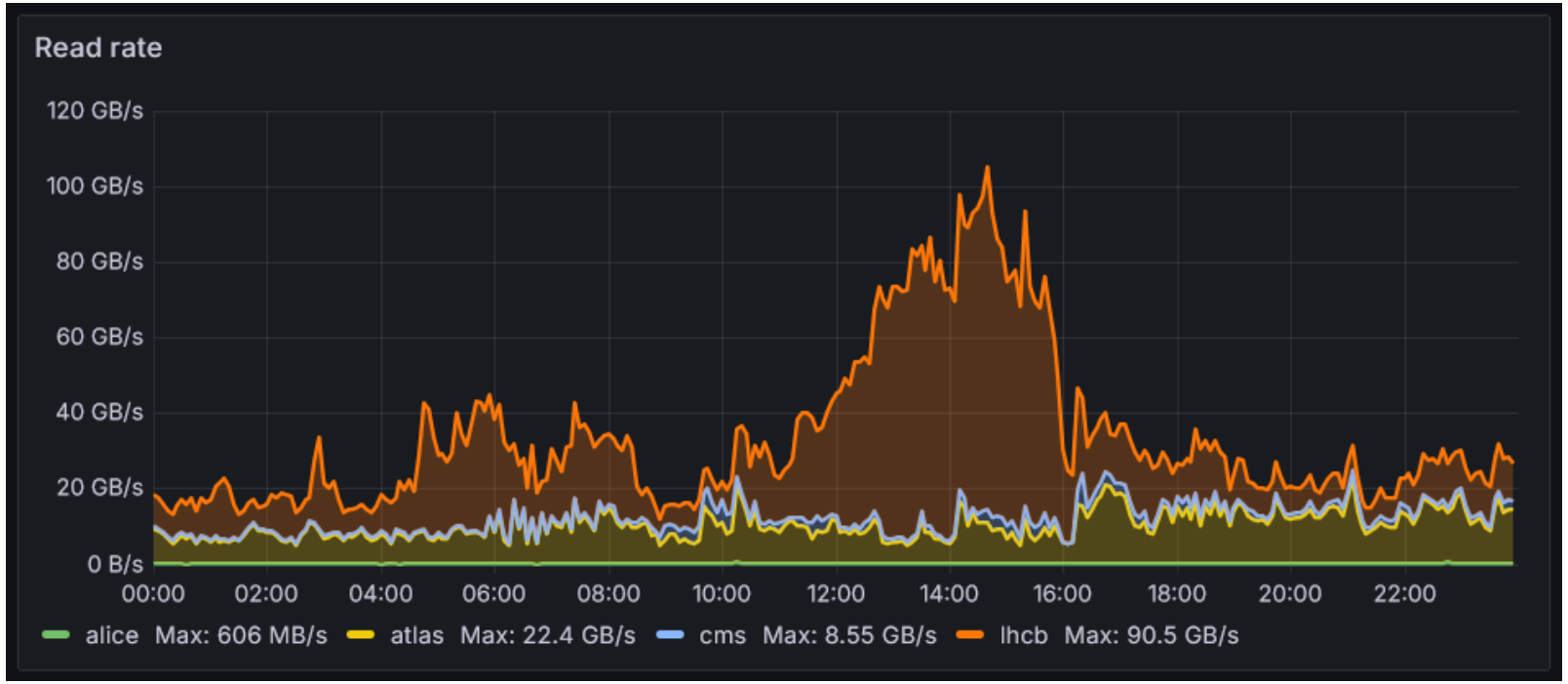
DirectI/O

- DirectI/O is now a common access method for data processing / analysis.
- Able to handle spikes of LHCb vector reads with very few jobs failures.

In the last 90 days:

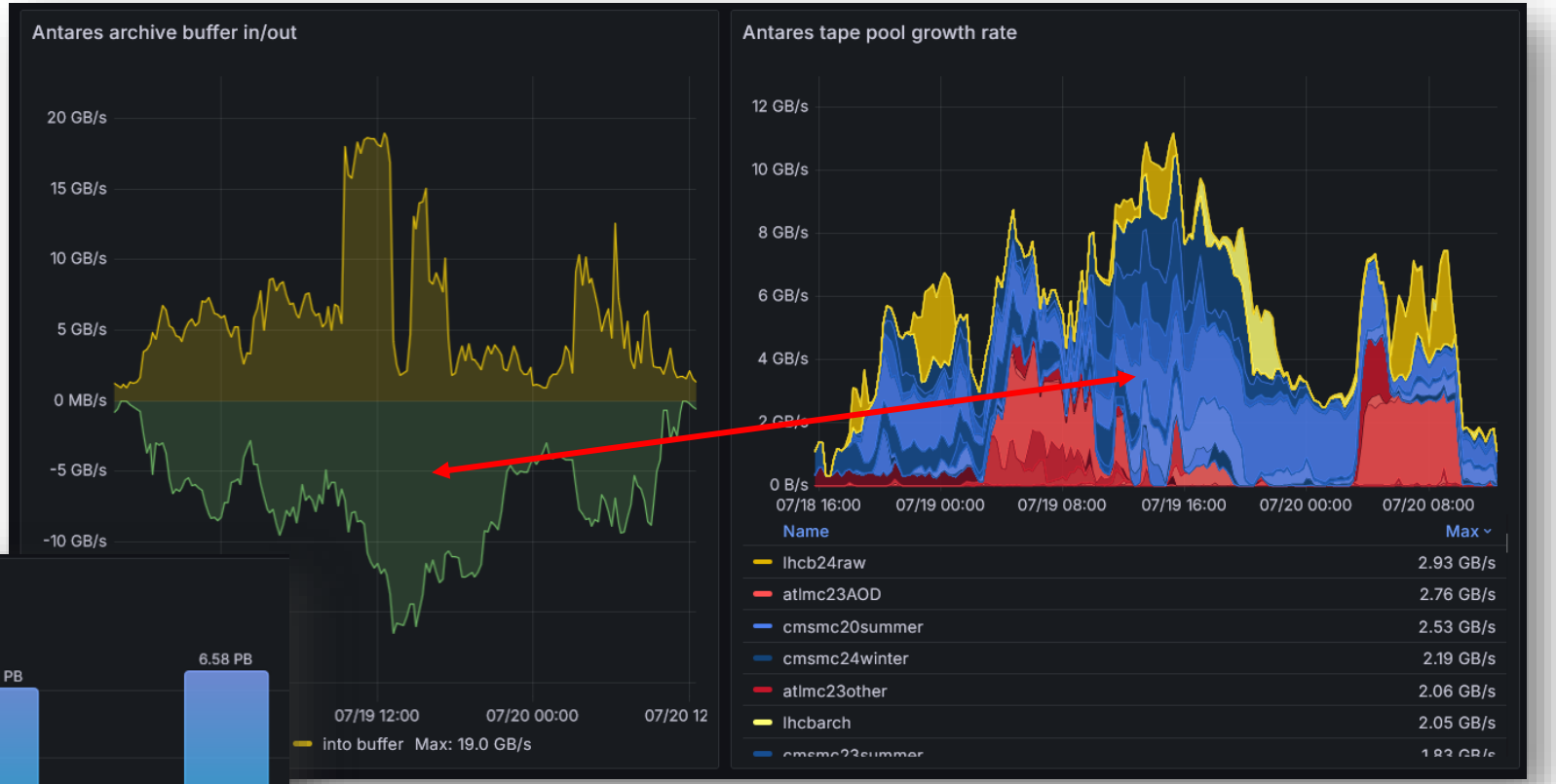
77.64PB
of data transferred

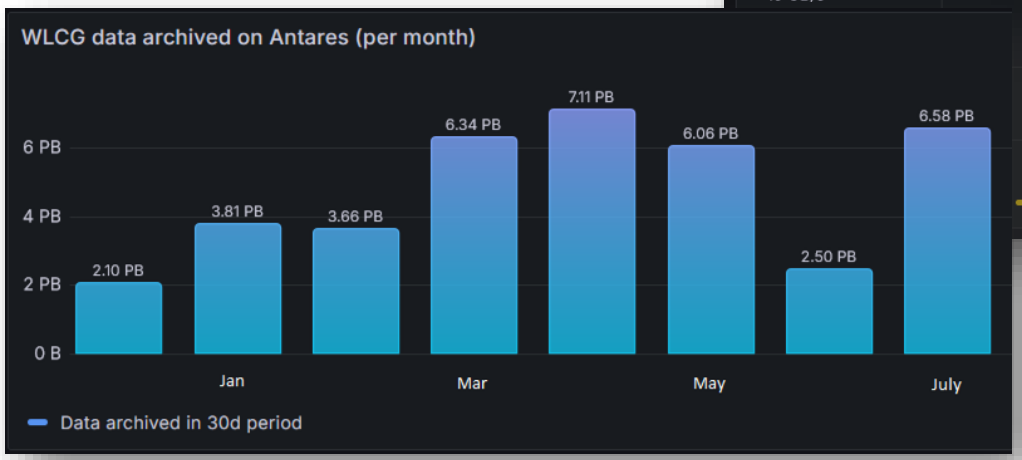
144,560,889
total transfers



Antares – EOS + CTA

Antares was initially designed to sit behind Echo.
XRootD to talk between the two.
With HTTP Tape API migrating to separate endpoint.





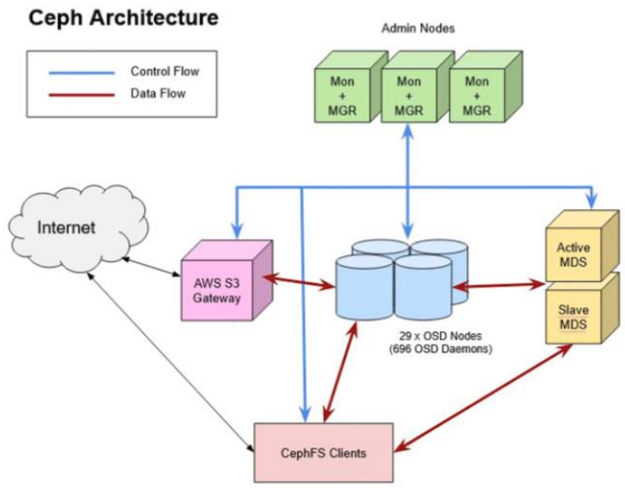
Averaging over 1PB a week written to tape since data taking restarted (last 6 months)

Tier-2 storage status

Category	Site	Primary VO	Storage	Notes
CORE	Glasgow	ATLAS - 10PB	Ceph + XrdCeph	
		Others	CephFS + XRootD	
CORE	Imperial	CMS - 23PB	dCache	
CORE	Lancaster	ATLAS – 10PB	CephFS + XRootD	
CORE	Manchester	ATLAS – 12PB	CephFS + XRootD	New Deployment
CORE	QMUL	ATLAS – 13PB	Lustre + StoRM (XRootD R/O)	Downtime for new DC
	Birmingham	ALICE	EOS	
		Others	XCache	
	Bristol	CMS	CephFS + XRootD	
	Brunel	CMS	CephFS + XRootD	
	Durham		CephFS + XRootD	
	Liverpool	ATLAS	dCache	Migrated from DPM
	RAL-PPD	CMS	dCache	

Lancaster – CephFS + XRootD

- Existing Ceph+XRootD installations used xrdceph as the interface between Ceph and XRootD. Lancaster chose to trailblaze the use of CephFS to interface to XRootD. https://indico.jlab.org/event/459/contributions/11358/
- After advice from Dan van der Ster we chose Ceph Pacific and configured our cluster with one active MDS server with a single rank backed up by a single standby.
- Data is stored using 8+3 Erasure Coding.
- Installation was done using cephadm.
- Single XRootD server.
- An S3 gateway - light usage.

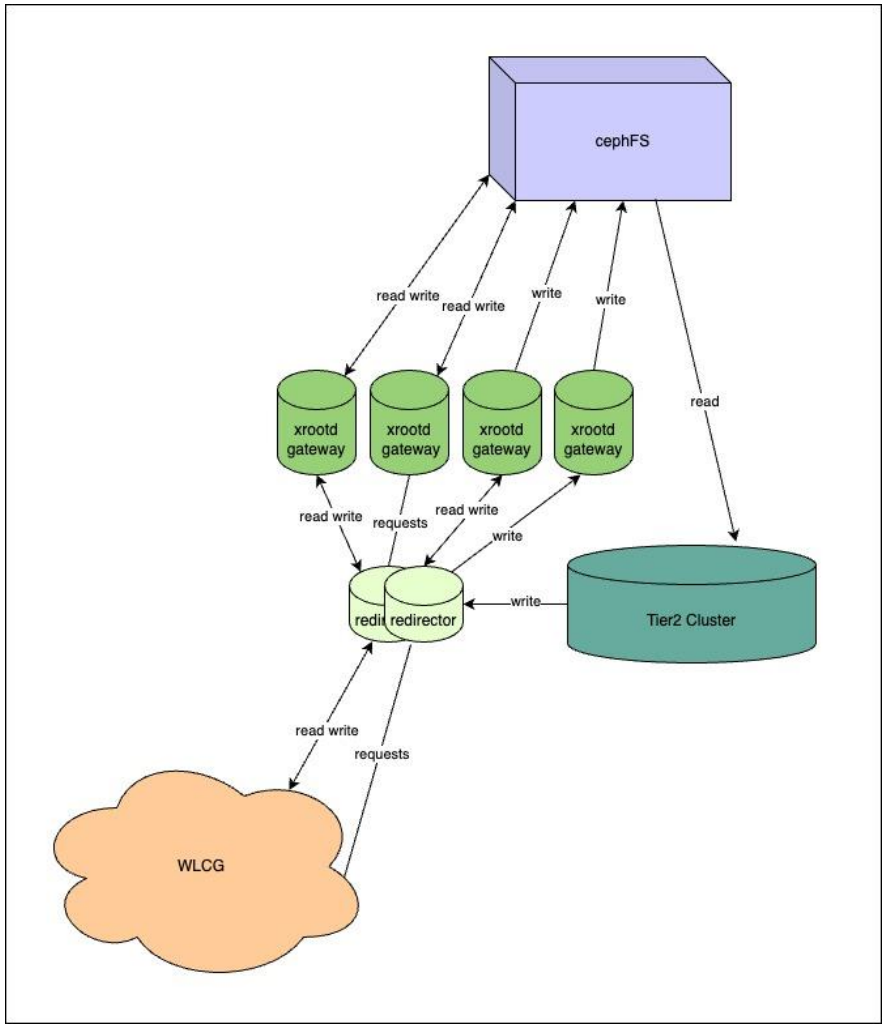


Lancaster University

g.hand@lancaster.ac.uk

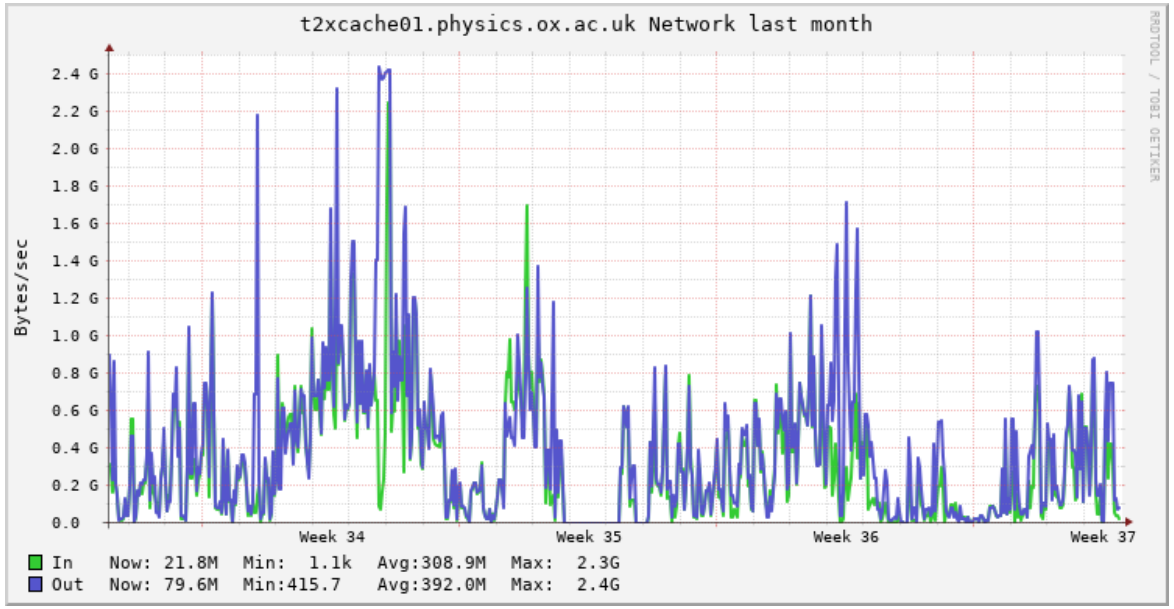
Manchester - CephFS + XRootD

- Ceph(fs) (reef on rocky 9)
  - Currently 6.4 PB spread across 28 servers bought explicitly for ceph
    - NVME disks for db osd
    - Standard hdd for data
    - 192 GB memory
    - Cephfs mds on same nodes
  - Future dpm hardware will be moved 6 additional PB, 58 older servers
    - hdd only
  - It will be interesting to see how ceph behaves
- Xrootd setup (5.7.0 on rocky 9)
  - 4 gateways
    - 2x25 Gb/s NICs
    - 256 GB mem
  - 2 redirectors on VMs on separate hosts
    - Failover configuration (for now)
- External transfers via https
- Local jobs read directly from cephfs but write access only via xrootd



XCache

- Oxford, Birmingham, RHUL XCaches are deployed.
- Run ATLAS jobs pointing at Core Tier-2s / Tier-1.
- Advantage:
  - Allows flexible configuration.
  - Acts as Buffer
- Relatively little caching.



Oxford throughput, 24 x 4TB SSD

When things break in Xcache huge amount of noise in logs:
240904 08:25:18 3703084 XrdPfc_File: error ProcessBlockResponse block 0x7fc8cc06a410, idx=9754, off=2556952576 error=-110
/atlas:datadisk/rucio/data22_13p6TeV/7c/28/DAOD_PHYS.39672532._000061.pool.root.1
240904 08:25:18 3703084 XrdPfc_File: error ProcessBlockError() io 0x7fcadc0cdff0, block 9754 finished with error 110 connection timed out
/atlas:datadisk/rucio/data22_13p6TeV/7c/28/DAOD_PHYS.39672532._000061.pool.root.1

Supporting XCache & VP

- Virtual Placement (VP) is when data is pushed to a Storage Element (or XCache) shortly before it is required by the job.
  - This would allow XCache to be even more effective.
- XCache is in use by multiple UK sites for Virtual Placement.
  - VP relies on GeoIP ordering of replicas as returned from RUCIO.
- Known to be broken. Was “fixed” for DUNE several months ago.
  - Effort at Edinburg on global fix.
  - Integration into RUCIO will likely be ~ 6months or so.

Future improvements

- RAL is deploying Gateways with 100Gb/s connections and will look into optimising their throughput.
- XRootD development:
  - Deletion rate in RADOS (RAL specific)
  - Transparent writable WN gateways
  - Containerisation and Orchestration of Externally facing XRootD gateways.
  - Improving buffer layer in XrdCeph
- The UK will try and document our experience with FileSystem + XRootD setups as Storage Endpoint.
- Checksumming improvements – Have to re-read the file to checksum it.

Science and
Technology
Facilities Council

Questions?