Science and Technology Facilities Council

Use of XCache in UK (GridPP)

James Walder
(On behalf of UK / GridPP community)

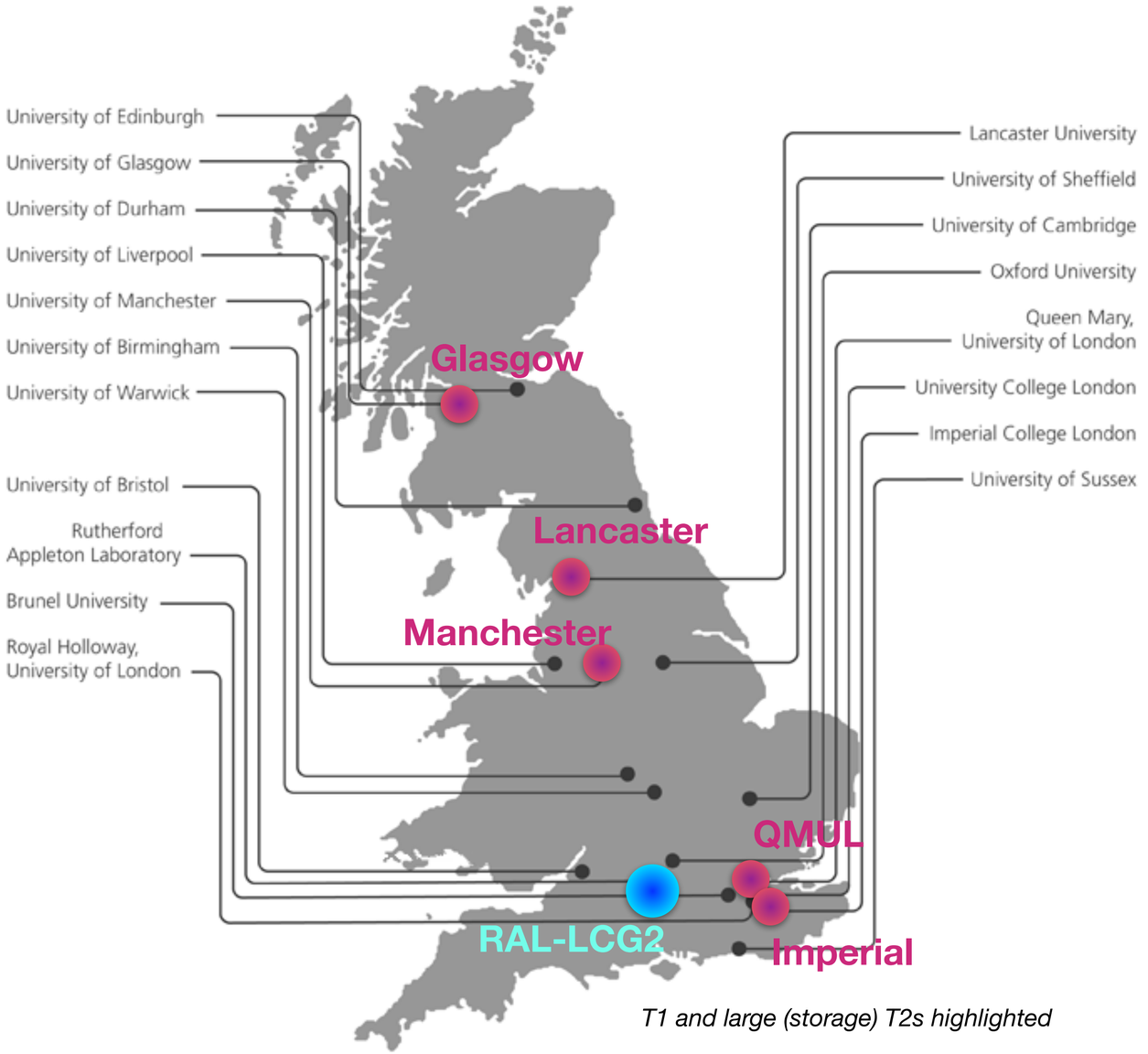
With thanks to Alastair Dewhurst, Alexander Rogovskiy et. al for material

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

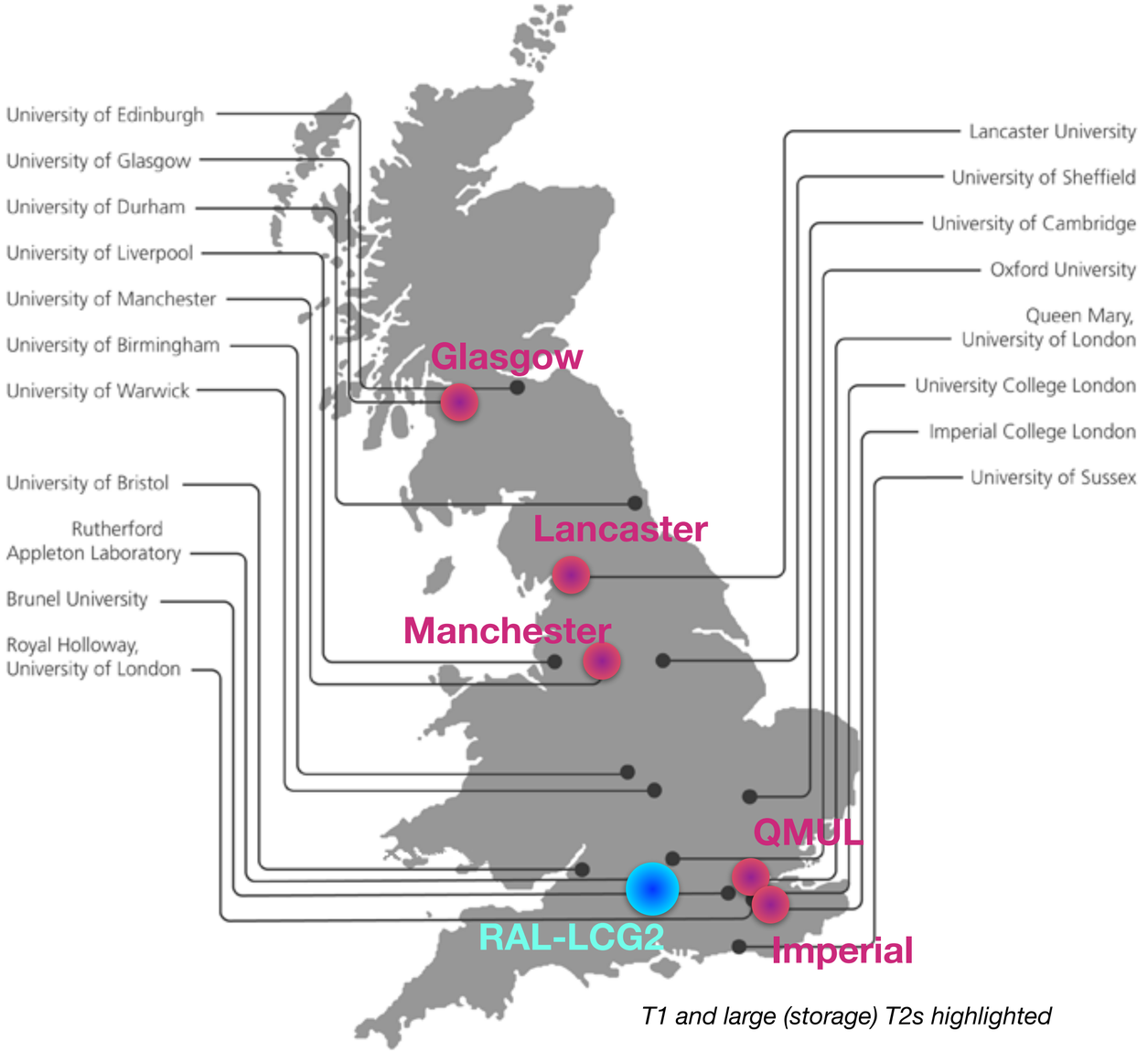


T1 and large (storage) T2s highlighted

Sites and storage configurations (GridPP)

- Sites, their storage technologies
- Links between Storageless sites and their respective SEs.

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	Mancheste r	ATLAS – 12PB	CephFS + XRootD	New Deployment
CORE	QMUL	ATLAS – 13PB	Lustre + StoRM (XRootD R/O)	Downtime for new DC
	Birmingha m	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	

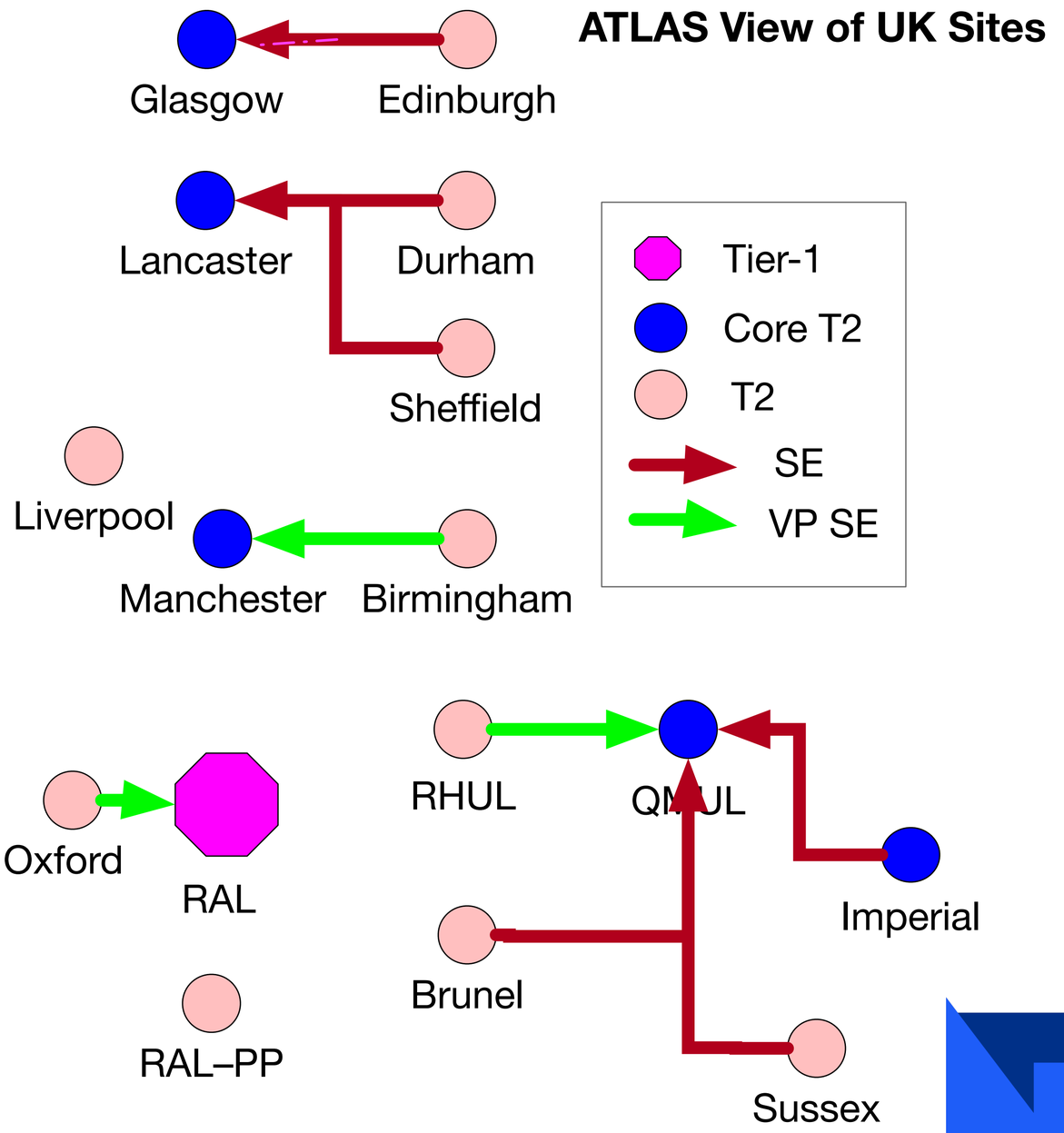


T1 and large (storage) T2s highlighted

Sites and storage configurations (GridPP)

- Sites, their storage technologies
- Links between Storageless sites and their respective SEs.

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	Mancheste r	ATLAS – 12PB	CephFS + XRootD	New Deployment
CORE	QMUL	ATLAS – 13PB	Lustre + StoRM (XRootD R/O)	Downtime for new DC
	Birmingha m	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	

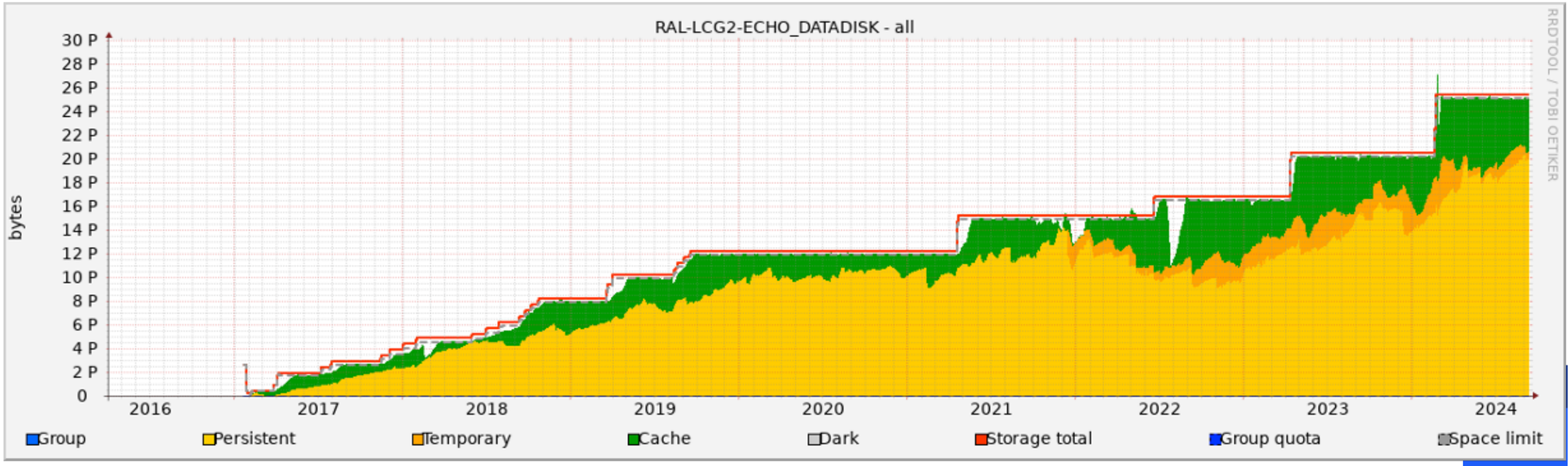


Evolution of Tier-1 Storage

- Migration in ~2017 to XRootD + Ceph Storage (Using the XrdCeph plugin and libradosstriper)
  - More recently GridFTP → Webdav for TPC
- Significant increase in usage
  - Increase in directI/O usage
- New Tape (CTA / Antares) service currently behind Echo (mainly using multi-hop TPC).

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

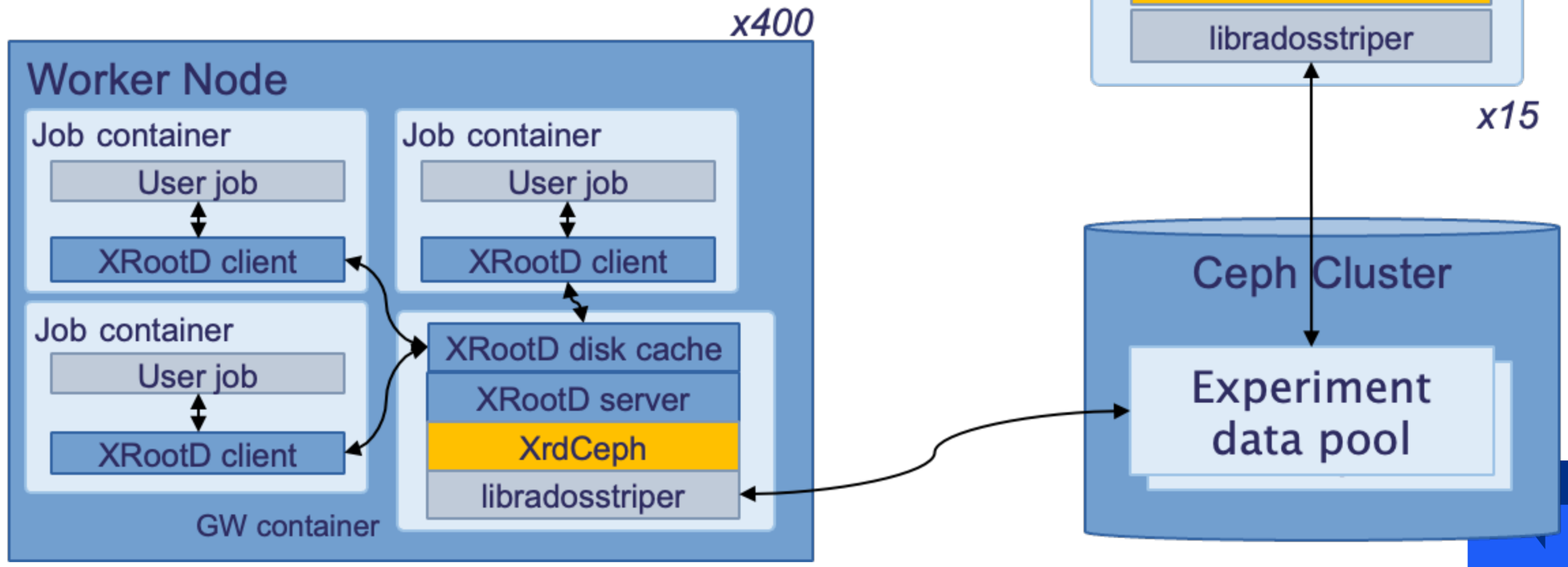
Echo provides 73PB of usable storage across 268 servers and more than 6000 HDD.



Data Access (Tier-1)

- TPC data access via External (XRootD) Gateways:
  - Dedicated gateways for ALICE and CMS AAA

- Each Worker Node also contains XRootD server and XCache
  - (R/O access to ECHO); Writes via External Gatways



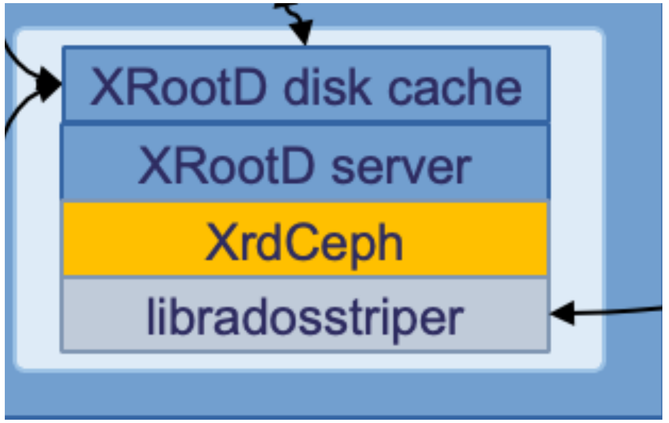
Caching required

- Ceph does not like small reads. To deal with it, need some caching:
  - i.e to read read data in large blocks instead of small chunks
  - WNs are using XRootD (XCache) proxy for this:

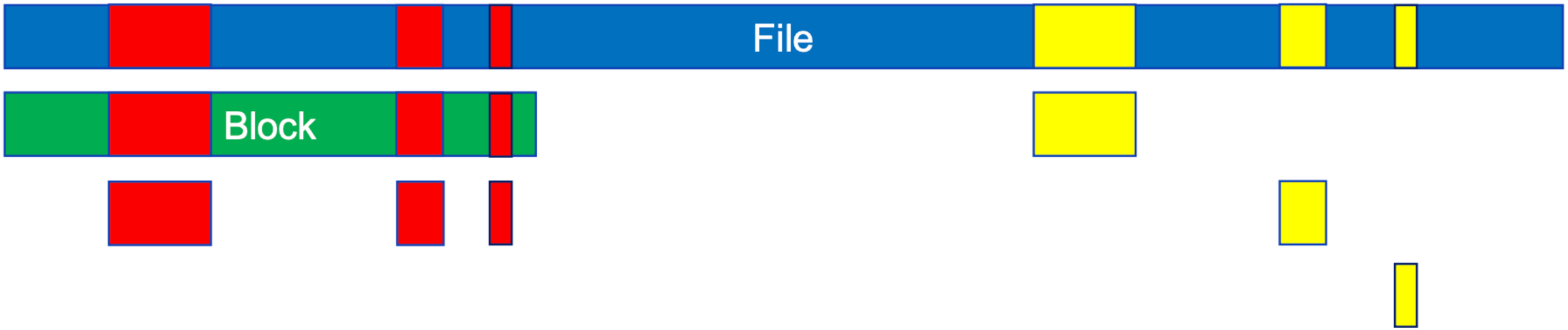
- WNs:
  - Increasing use of direct-IO (vector read) requests (wide range of usage patterns)
  - VOs use these requests to execute “Direct Access” jobs
    - i.e jobs that do not download input data, but access it directly from the storage

- External Gateways:
  - Mainly whole file sequential reads
  - Buffering layer added to XrdCeph plugin

Why XCache?



- Why XCache?
- Memory proxy would seem more suitable
  - Removes unnecessary copying to local disk
  - Unfortunately, memory proxy executes vector reads sequentially
- I.e. requests each chunk using ordinary Read requests one by one
- While XCache can extract necessary chunks from blocks

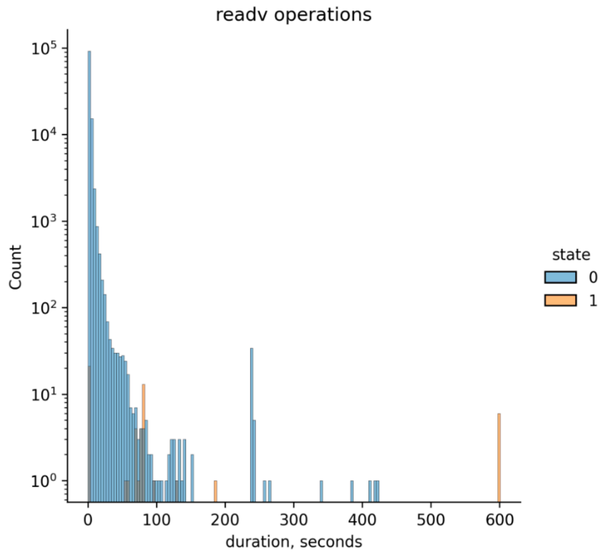


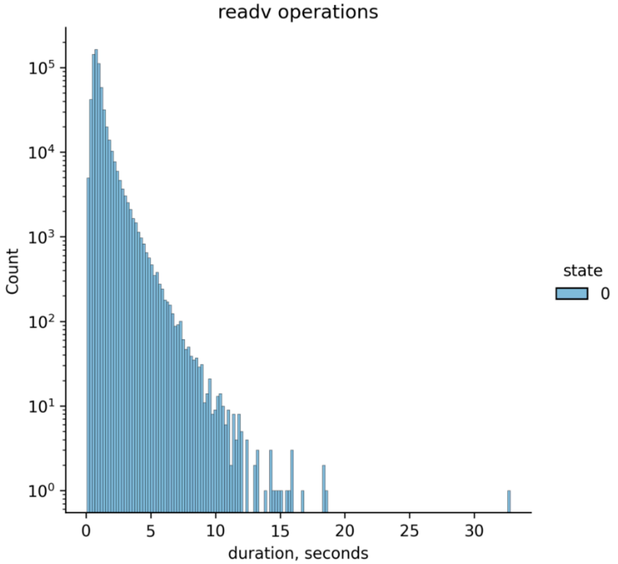
Aside: XrdCeph improvements

- We had two problems:
  - Vector Reads were executed sequentially by XrdCeph
  - All reads were relying on Rados striper and therefore were slow (locking overheads, due to ~ atomic correctness of libradosstriper)

- A new version of XrdCeph was developed
  - Does not use libradosstriper for synchronous reads
  - Uses Rados atomic reads
  - Does not generate any (read) locks (WORM – immutability)

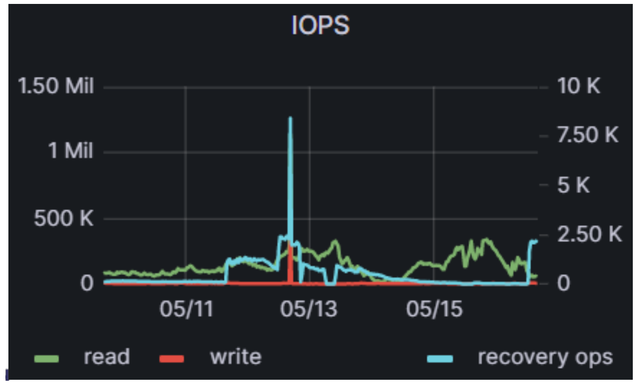
- Tests shown that the new XrdCeph version was better

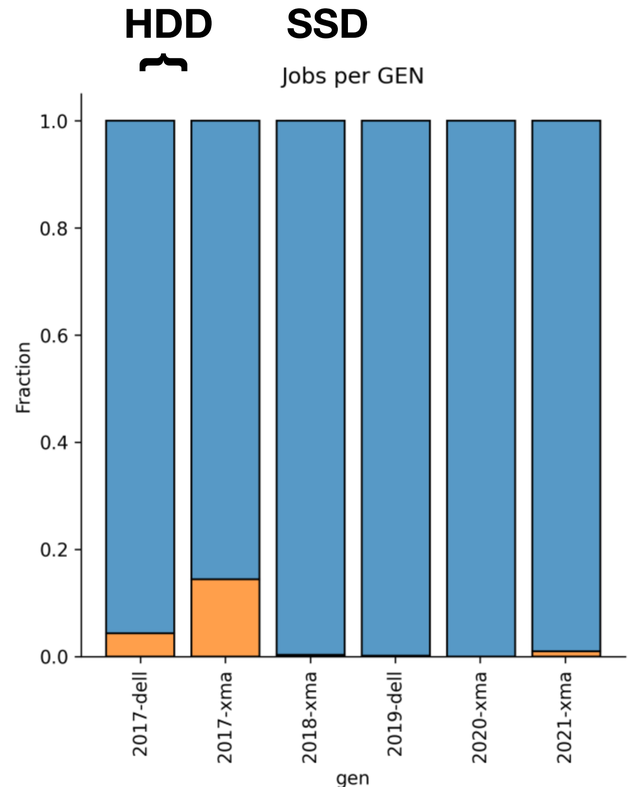




XrdCeph: deployment

- Improved XrdCeph code deployed, without XCache enabled on WNs:
  - Good performance for first days; XRootD + XrdCeph no longer the bottleneck
- However …. After a few days
  - ECHO overloaded with IOps with degraded overall performance
  - XCache reintroduced.
    - Some issues with readV's reappeared:
      - Resolved (for LHCb) by moving to SSD-only WNs.
- Final goal: either remove XCache, or, make more effective use of caching i.e. range coalescence.





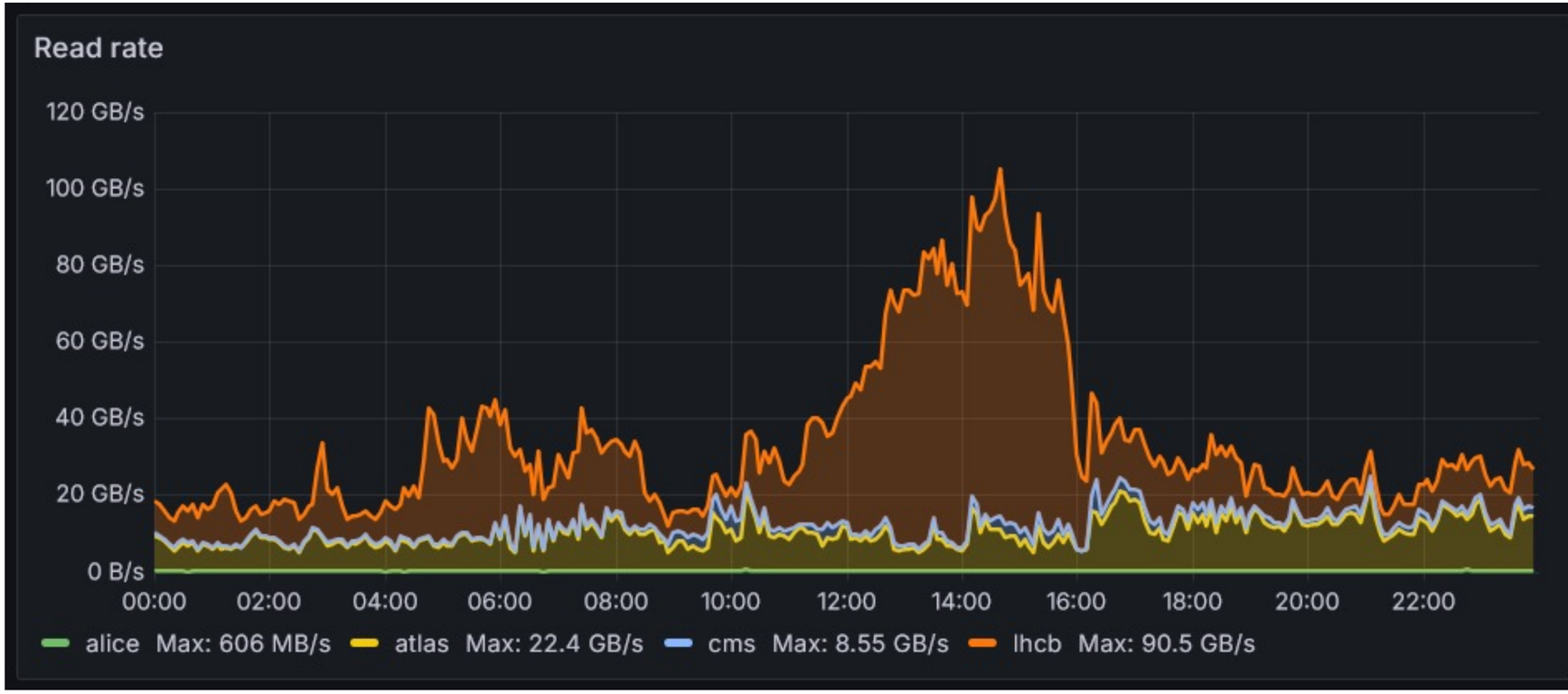
DirectI/O

- 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

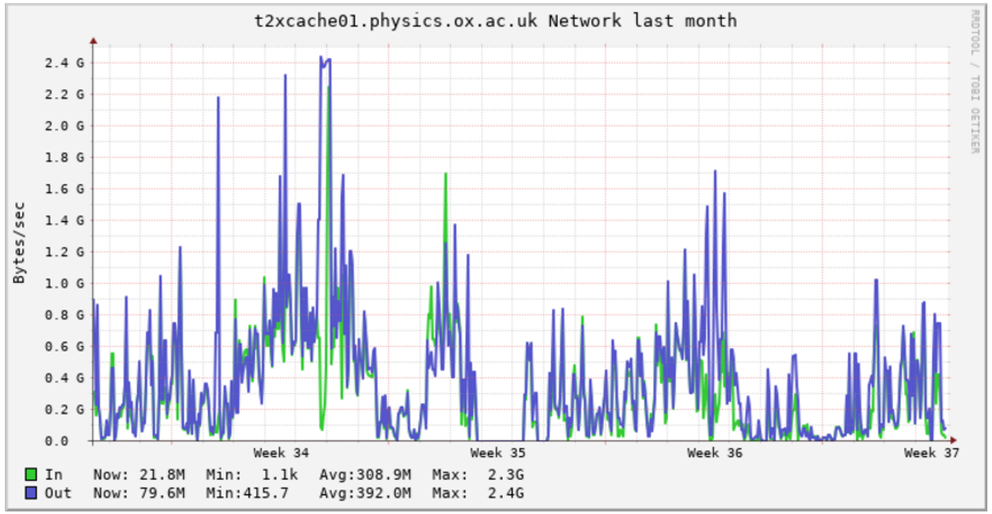


Other uses of XCache in UK

- Consolidation of Storage to ~ 5 core sites in UK:
  - Other institutions providing pledged Compute

- Oxford, Birmingham, RHUL XCaches are deployed (Virtual placement)
  - Run ATLAS jobs pointing at Core Tier-2s / Tier-1.

- Advantage:
  - Allows flexible configuration.
  - Acts as Buffer (e.g. latency hiding)



Oxford throughput, 24 x 4TB SSD

- However … Relatively little caching in production workflows

Virtual Placement

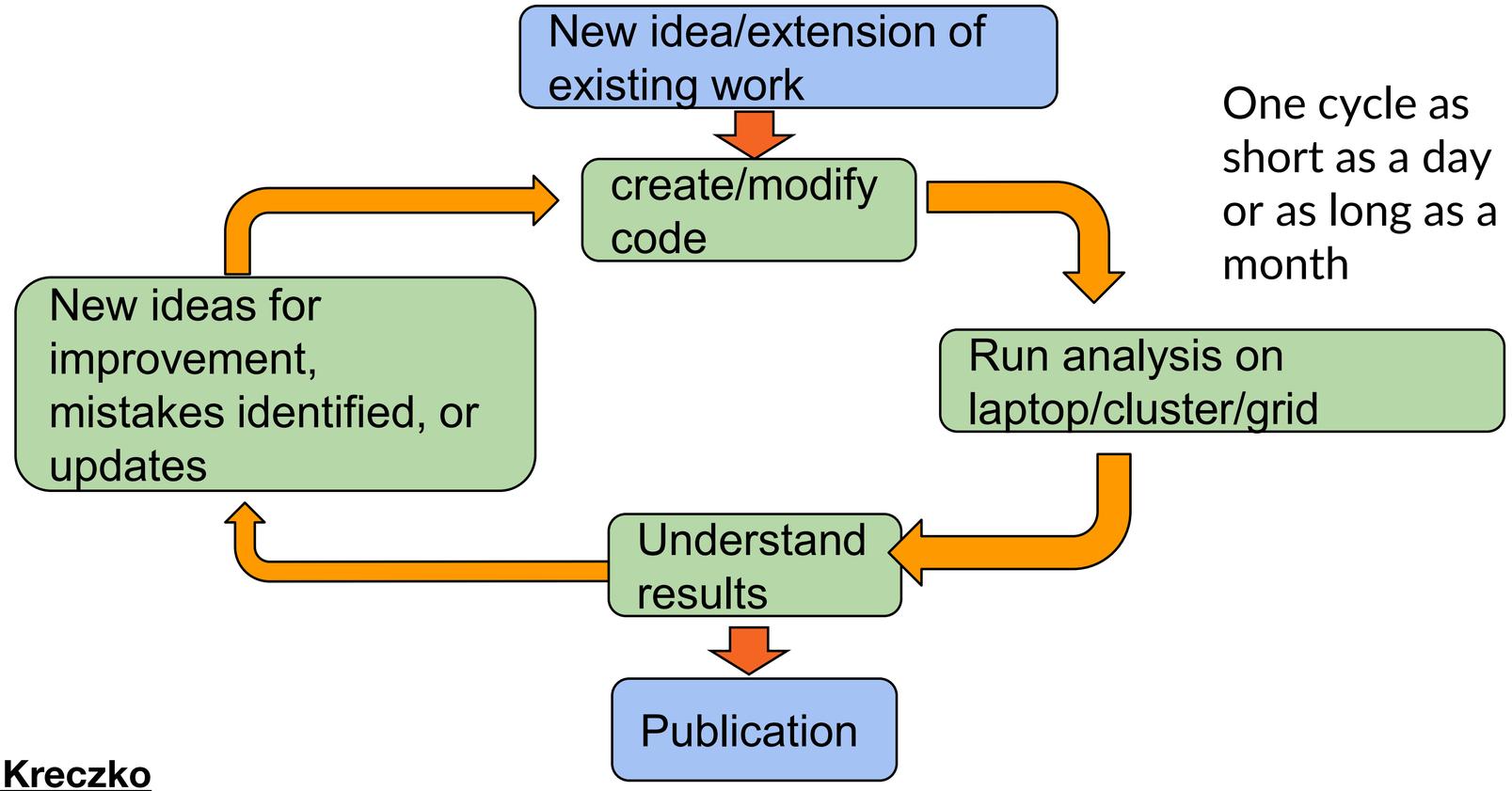
- VP developed with ATLAS / Rucio:
  - Use rucio to assign ‘virtual replicas’ at sites; aims to improve cache hit rates
  - Jobs can run at these sites with data ‘hopefully’ already cached:
    - Most benefit with analysis jobs (higher hit rates); less clear for production

- VP relies on GeoIP ordering of replicas as returned from RUCIO.
  - Known to have issues. Was “fixed” for DUNE several months ago.
  - Effort at Edinburgh on global fix.
    - Integration into RUCIO will likely be ~ 6months or so (?)

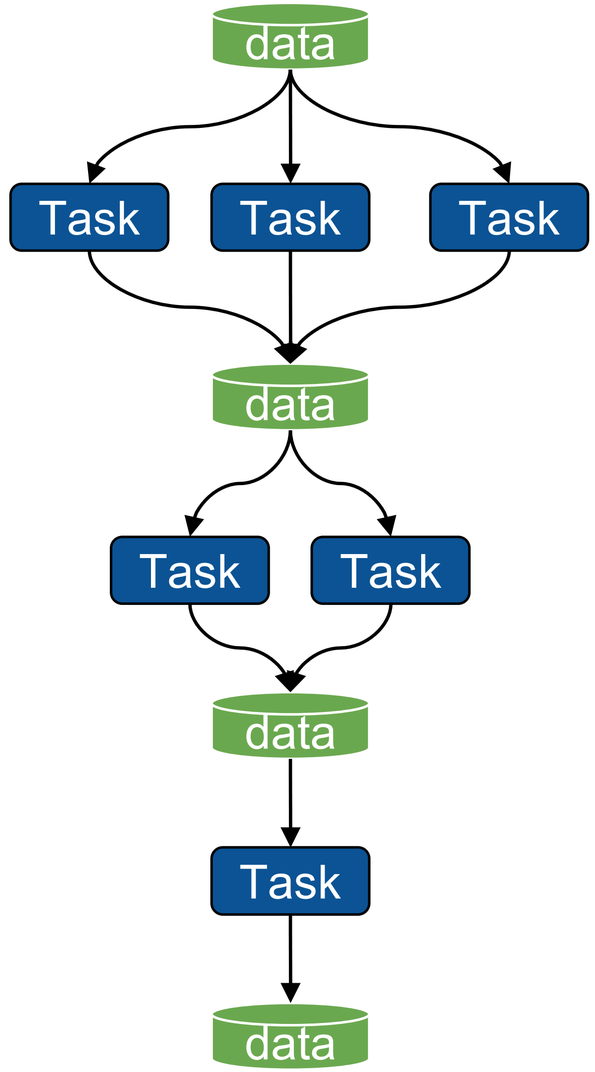
- Current (and future) needs for UK require efficient means to connect Storage at core sites to distributed Compute
  - VP (and XCache) are facilitating this.

XCache for Analysis

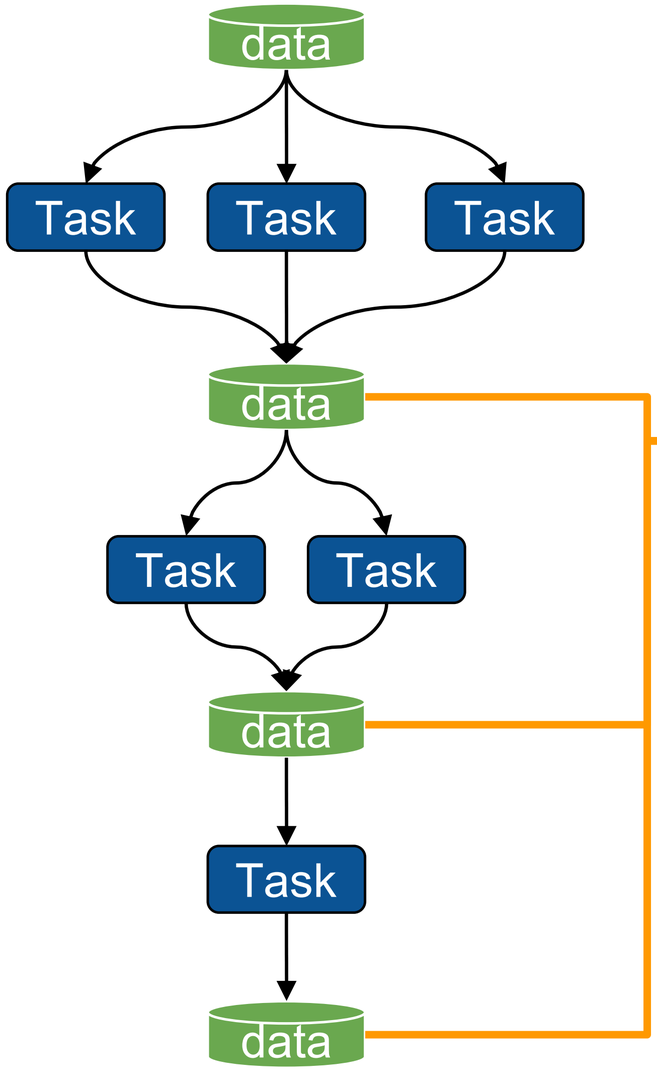
- Swift-HEP: UK contribution to the development of software for High Energy Physics experiments
  - Analysis Anatomy:



One cycle as short as a day or as long as a month

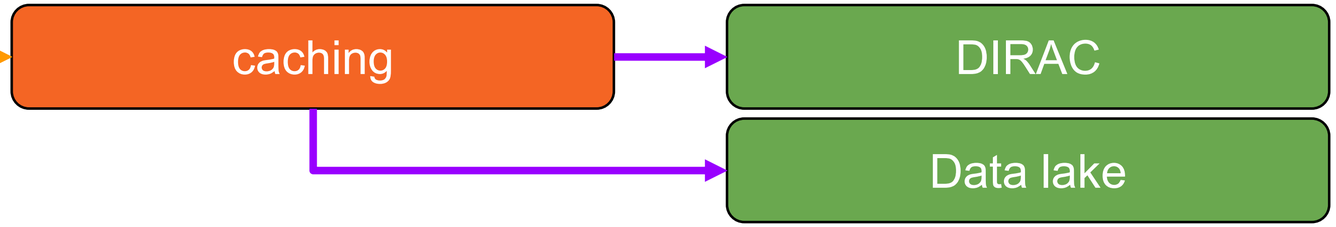


XCache for Analysis (2)



WP5

Analysis **step** output



<u>Caching Opportunities</u>:
The intermediate data at various stages of the workflow are opportune points for caching.

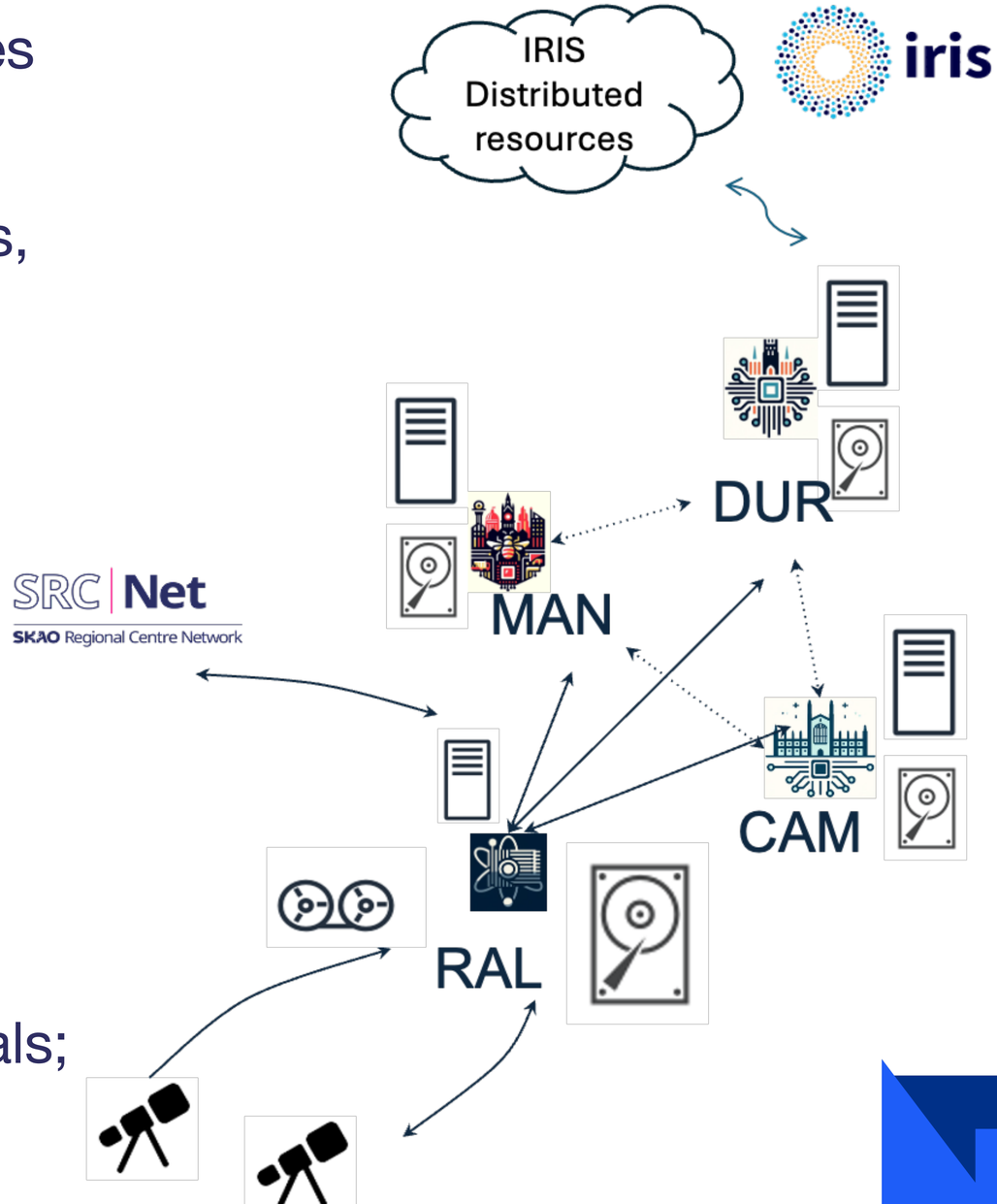
By storing these results, we can avoid redundant computations in iterative analysis, thus saving time and resources.

WP5 in a nutshell:
Run analysis workloads optimally** on distributed (GridPP) resources

** balanced between user-experience and computing efficiency

Caching for UKSRC (SKA)

- UK: to provision significant Storage / Compute resources
  - SRCNet v0.1 – RAL to provide storage and compute
- Design goal is to distribute the compute across UK sites, with some storage at each site;
  - Also aim to make opportunistic / planned usage of other HPC centres
- Caching (XCache, …) likely to have a role to play,
  - R&D to prototype and understand wrt. various workflows
- Understanding of whether to expose each site as ~ independent site in SRCNet, or encapsulate all internals;
  - Answer is likely somewhere in-between …

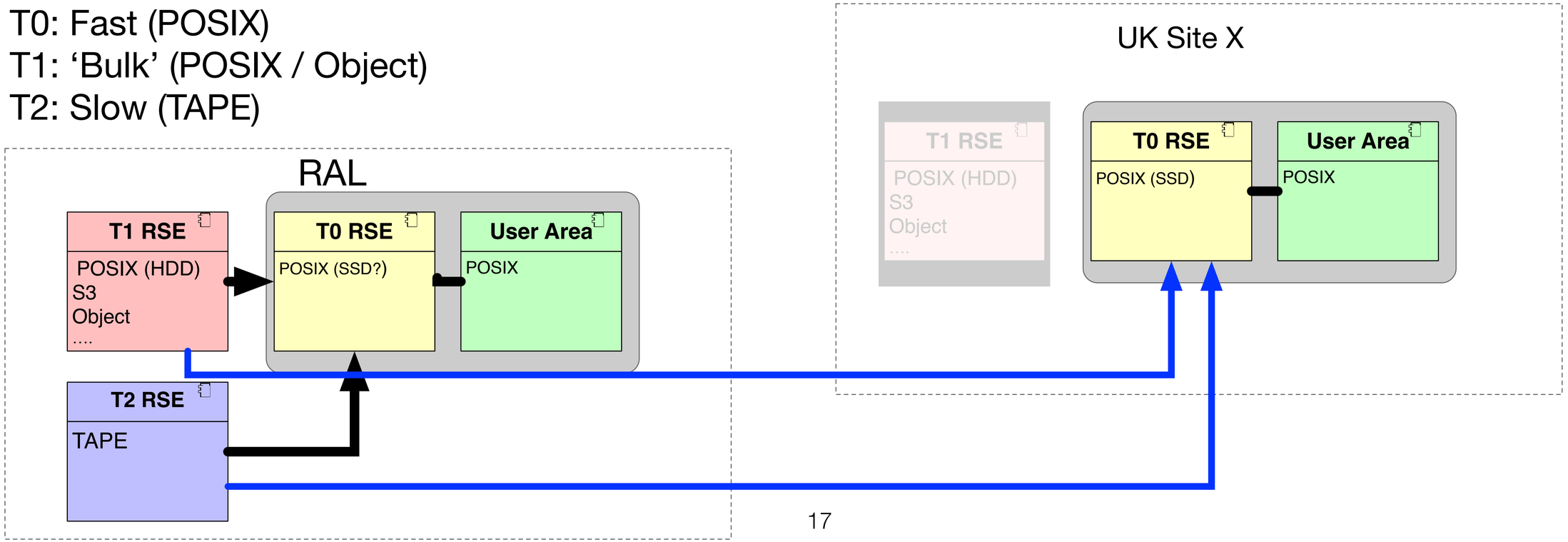


UK distributed proto-data model

- Current model to locate (majority) of bulk data and Tape at RAL:
  - Other sites will have local fast storage for user and ‘staging/cache’ area
  - Align with current SRCNet Tiering model

Storage Tiers:

T0: Fast (POSIX)
T1: ‘Bulk’ (POSIX / Object)
T2: Slow (TAPE)

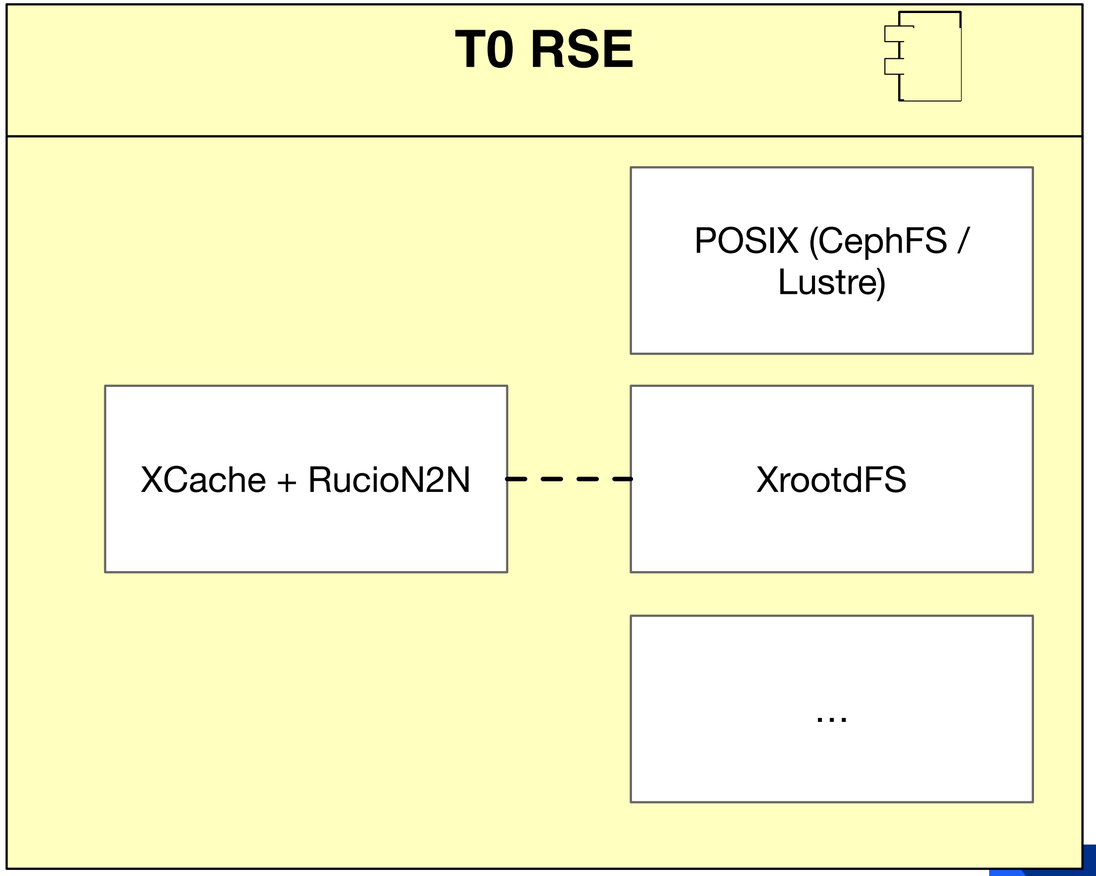


T0 RSE (?)

- Required to be – or rather, to provide – POSIX-like interface due to legacy codes
  - (May need to allow symlinked access from user area space)
- Is known to Rucio, to enable pre-staging of data and lifecycle management
- Must (in future) restrict access appropriately (i.e. embargo)

- If only partial file access, or high reuse, caching *should* be implemented.

- Tools, e.g. fsspec, s3fuse, … direct https
  - may provide mitigations going forward ?



Summary

- UK has significant experience with XRootD and XCache
  - XCache usage from ‘transparent’ site-based implementations,
    - to fully Rucio aware workflows (Virtual Placement)
- More details in talks (A. Dewhurst, A. Rogovskiy) from XRootD & FTS Workshop 2024

- Analysis workflows: fast turnaround, high data reuse
  - Ideal use-case for caching;
  - Swift-HEP working towards solutions for analysis facilities (in collaboration with, e.g. IRIS-HEP).

- UKSRC will be a distributed infrastructure (within the distributed infrastructure of SKA/ SRCNet)
  - Prototyping work to be started to understand the need / relevance of Caching in this environment