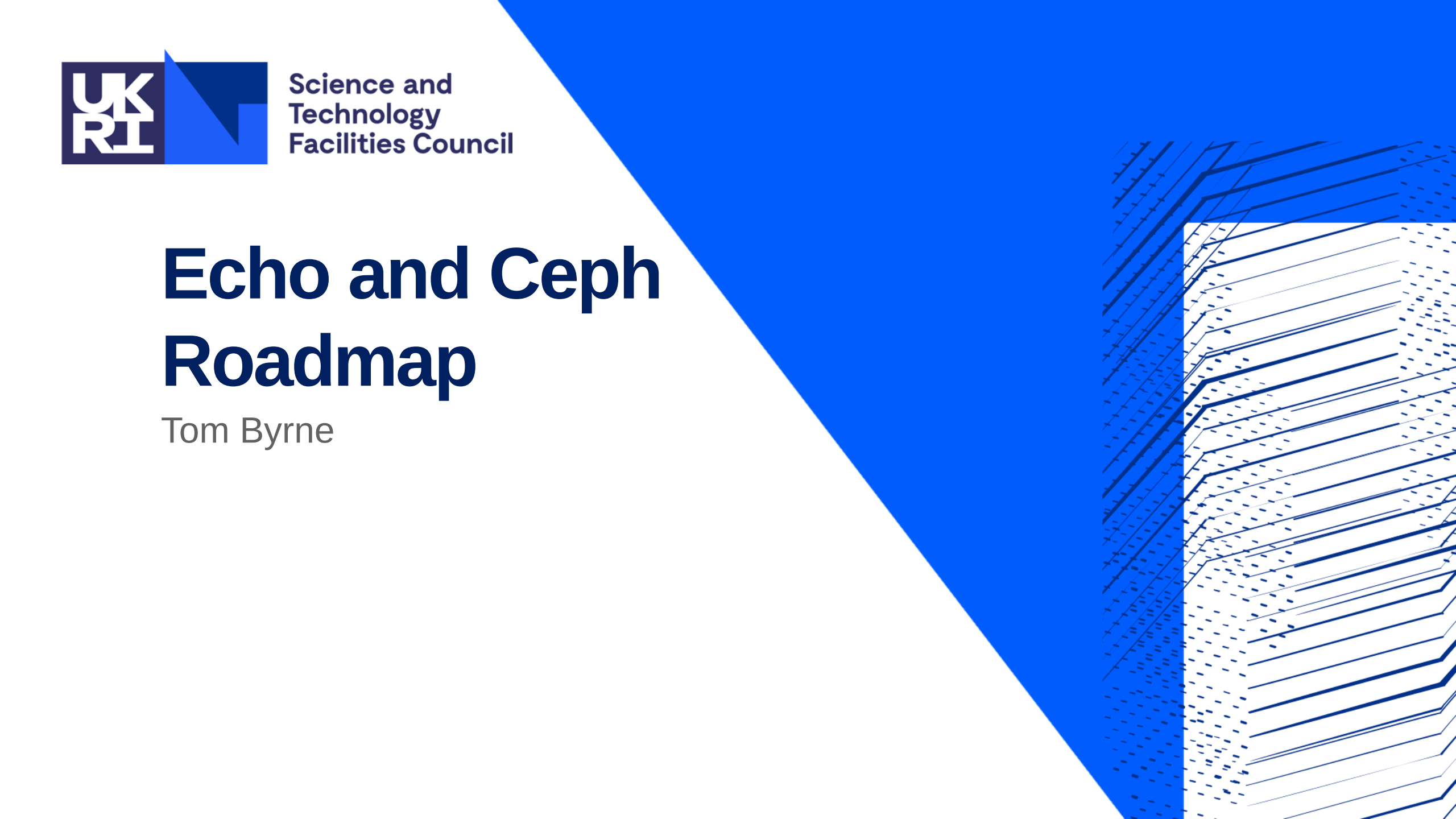


Echo and Ceph Roadmap

Tom Byrne

Introduction

- The Echo project started in 2015 to replace Castor for Disk.
  - Strategic goals: Industry Standard backend with thinnest layer of Grid middleware ontop.
- A Ceph cluster providing 34PB of usable storage
  - supports the LHC experiments
  - and many other other non LHC experiments and organisations
- Uses thin plugins for XRootD and GridFTP to translate requests directly to low level Ceph commands
  - SRM-less WLCG disk storage since 2017
    - Used SRR for storage reporting from the start

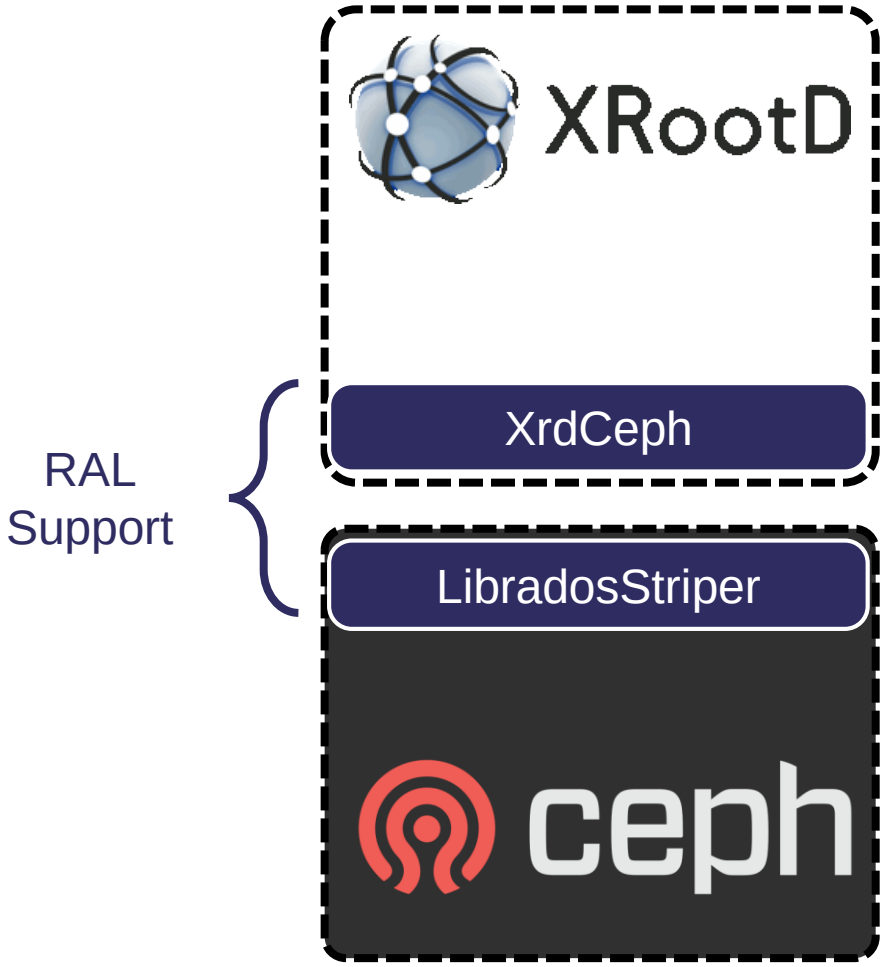
Science and Technology Facilities Council

XRootD

XrdCeph

- An XRootD filesystem plugin that uses libradosstriper to talk to a Ceph cluster
  - Lower level then CephFS, RGW, RBD etc.

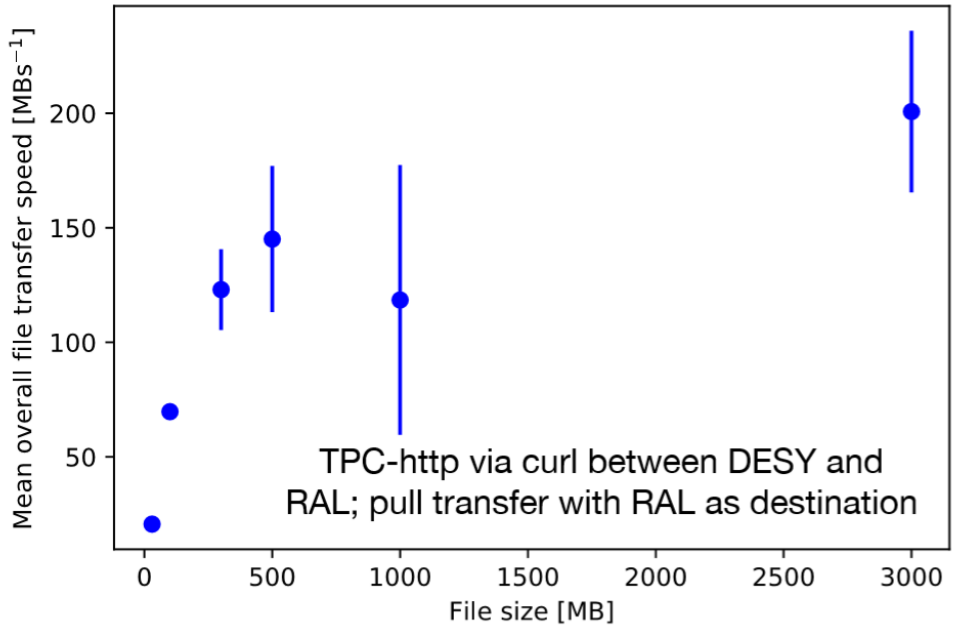
- People involved in XrdCeph:
  - Ian Johnson (Primary contact)
  - George Patargias (Secondary, CTA expert)
  - Sam Skipsey (GridPP Storage Coordinator)
  - James Walder (ATLAS Software Expert)
  - Tom Byrne (Ceph Expert)



Third Party Copies

- Echo will support XRootD and http TPCs.
- Being an object store Echo lacks certain operations (e.g. mkdir, ls, mv)
  - Problems can occur with higher layer wrappers because they can make assumptions about lower layers.
- XRootD TPC are passing all Smoke tests.
  - Not focusing on performance as preference for http.
- Majority of http TPC working.
  - Issues with some permutations of SRC/DST, push/pull being work on.
- http TPC showing good single transfer performance.
- http TPC also demonstrated to work when placed under high load.

TPC-http	RAL acting as:	Copy mode	Result
Curl / gfal (non-TPC)	DST	upload, download, delete	✓
Curl (COPY)	SRC/DST	push/pull	✓
Davix	SRC/DST	push/pull	✓
FTS	SRC	push	✓
FTS	SRC	pull	✗
FTS	DST	push	✗
FTS	DST	pull	✓



Science and Technology Facilities Council

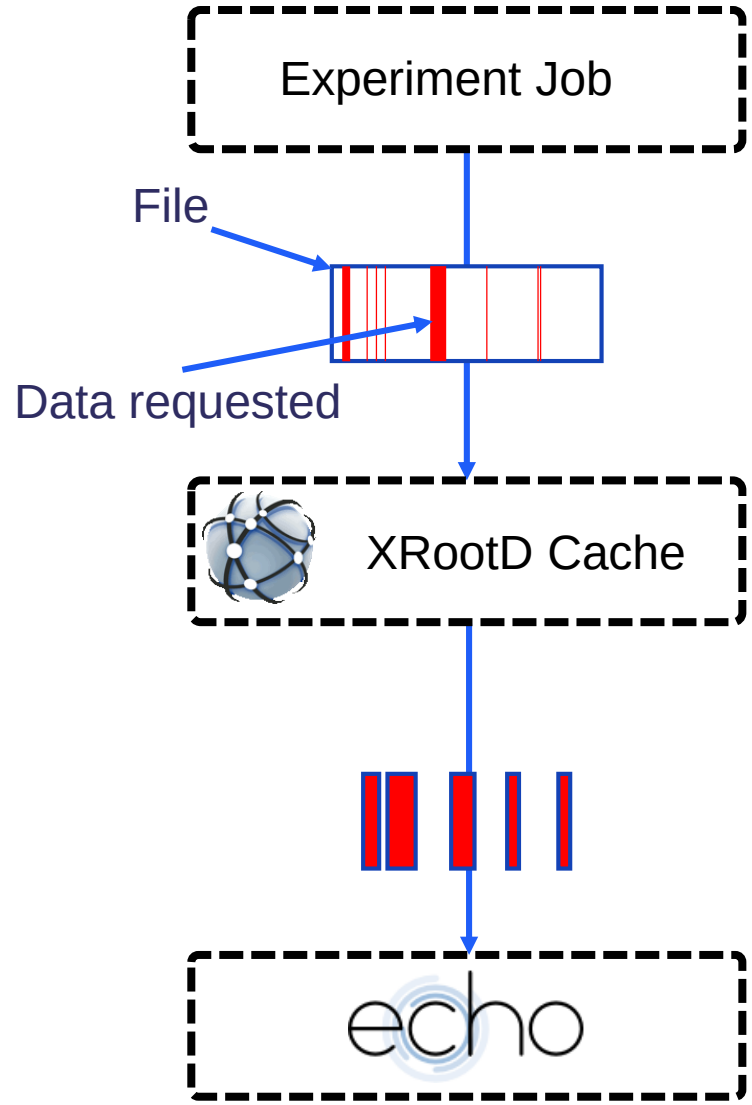
echo

AAI

- Relying on Authn/z from mainline XRootD
- Would like to keep a consistent authn/z layer between our services, currently limited to gridmap by CASTOR
  - Many more possibilities when moving to CTA
- See XRootD Roadmap talk:
  - https://indico.cern.ch/event/941278/contributions/4088026/

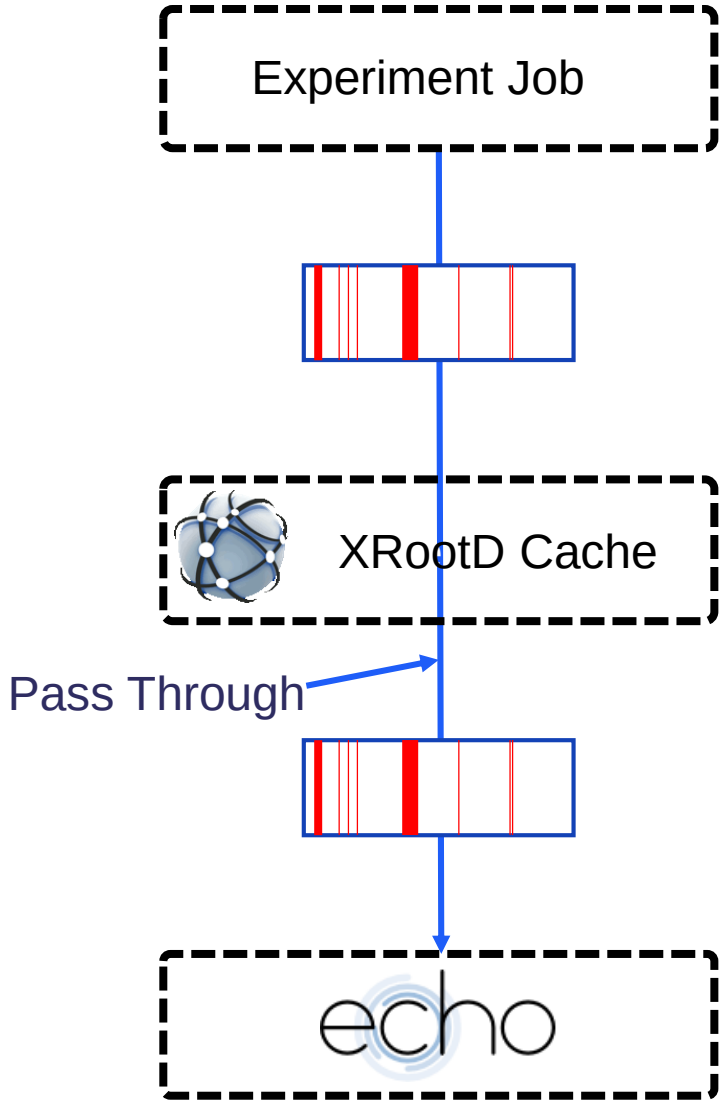
Vector Reads

- When Echo was originally deployed it didn't have support for Vector reads.
- We thought it would be better to request larger blocks of data from Ceph.
  - Similar to CMS' Lazy download.
- Small Caches are deployed on every WN that can turn many small requests into more organized blocks.
  - This appears to work very well…
- VOs reported higher than expected failures rates for Direct I/O jobs.



Vector Reads (2)

- (X)Caches are designed to accelerate work.
  - If they are busy they will pass through the request (so as not to cause problems!).
- Using improved monitoring with XRootD 5, XRootD team* looked at jobs accessing 50k files and performing 2.4million vector reads.
  - If the cache passed through the request, it would take between 400 - 1000 times longer to process.
  - Causing jobs to time out and reducing overall efficiency.
- We are working on implementing Vector reads.
  - Will be deployed as soon as ready, aim to merge into XRootD 5.2.x



*Thanks to Matevž Tadel - https://github.com/xrootd/xrootd/issues/1259

Plans

- RAL development has moved to XRootD 5.
  - Aiming for TPC and Vector read code to be included in XRootD 5.2.x
- TPC:
  - http (and xrootd) into production in 2021 Q1.
  - Assuming no short term blocking issues with XRootD 5, relatively straight forward.
- Vector Reads:
  - Aiming for code to be finished in 2021 Q1.
  - Still in development so it is not possible to say it will definitely fix all problems.
  - Note: Next years Capacity CPU procurement will be in production 2021 Q1 and that will mean ~80% of RAL's pledged capacity will be SSD backed.
    - SSD back CPUs have a significantly lower incidence of this issue.
- AAI:
  - In conjunction with CTA deployment we will deploy a consistent more modern AAI.

Ceph

Ceph in the scientific community

- Ceph is a hugely popular storage technology, with a large community
  - Officially supported Red Hat storage technology
  - 1000s of large deployments globally
- A growing scientific community across many disciplines
  - Monthly user gatherings
- Ceph usage in the HEP Community has also exploded
  - More on this later…

Ceph - Quality of Service

- Conceptually simple to implement different storage types in Ceph
  - Mixtures of device types and resilience methods possible
  - Understanding what is useful is key!
- Investigating adding faster (flash based) tiers of storage to Echo
  - Via S3 for non WLCG communities currently
- Not exploring reduced redundancy for bulk storage
  - EC storage overhead already low, and added administration effort not worth it
- All data pools currently 8+3 Erasure Coded, giving acceptable overhead and excellent ‘administrative flexibility’
  - See recording of Alastair’s talk on Friday for more info
    - https://indico.cern.ch/event/941278/contributions/4104604/

Ceph Development – Ease of Use

- Large scale storage is complicated!
  - Efforts in 2019 to make Ceph easier to manage
- Ceph orchestration
  - Adding support for deployment tasks within Ceph, supports multiple ‘orchestrators’
    - K8s
    - Bare metal (SSH)
  - Massive quality of life improvements for ‘day 2’ operations
- Configuration can now be centrally managed, configuration store kept by the mons
  - Reducing the need to managing configuration files across the whole cluster
  - `ceph config set <who> <what> <value>`
  - Daemons check for config when booting, and during runtime (if on the fly change possible)

Ceph Development – Project Crimson

- Ceph consumes raw block devices, no filesystem layer used
  - The ‘object storage daemon’ does everything from the raw block device access up to the network IO
- Current OSD code based on traditional multi-threading model
  - When storage is fast, context switching is expensive
  - Ceph is becoming increasingly CPU bound as storage becomes faster
- Complete IO path rewrite
  - Using seastar, a modern C++ framework designed for high-performance server applications on modern hardware.
  - One thread per core, no locking and blocking etc
  - Huge IOPS/Core improvements seen in early proofs of concepts

Ceph HEP community updates

CERN

- Has been running Ceph in production for 7 years
- Ceph backs their cloud, container, and HPC activities
- Dan van der Ster is a huge force in the Ceph community, and is on the board of the Ceph Foundation

Ceph Clusters at CERN (Sept-2020)		Size	Version
Block Storage for OpenStack	2 rooms avail.	6.4PB	nautilus
Hyperconverged: OpenStack + Ceph on same hosts		250TB	nautilus
CephFS for HPC/OpenStack/OpenShift	10x MDS	1.1PB	luminous
Pre-prod testing, 3x MDS		166TB	nautilus
Hyperconverged HPC: SLURM + Ceph on same hosts, 2x MDS		356TB	nautilus
S3 Object Storage	Erasure coded objects	1.9PB	luminous
CASTOR Tape System	Erasure coded objects	5.5PB	nautilus
CERN Tape Archive Metadata		800GB	nautilus

Uni Bonn

- Using XRootD on top of CephFS to support HEP users analysis workloads
- Several clusters to support wildly different use cases
  - Managed by a comparatively small number of staff

Main use cases of Ceph at Uni Bonn Tier3

HTC Cluster (ATLAS Tier 3): CephFS

- cluster with 4288 logical CPU cores, > 0.7 PB eff. CephFS
- $>$ 150 local users (ATLAS, Belle II, hadron physics,...), Grid jobs
- `LOCALGROUPDISK`: XRootD on OSDs + Redirector as VM
- Erasure coding ($k = 4$, $m = 2$), Snappy compression, IP over IB

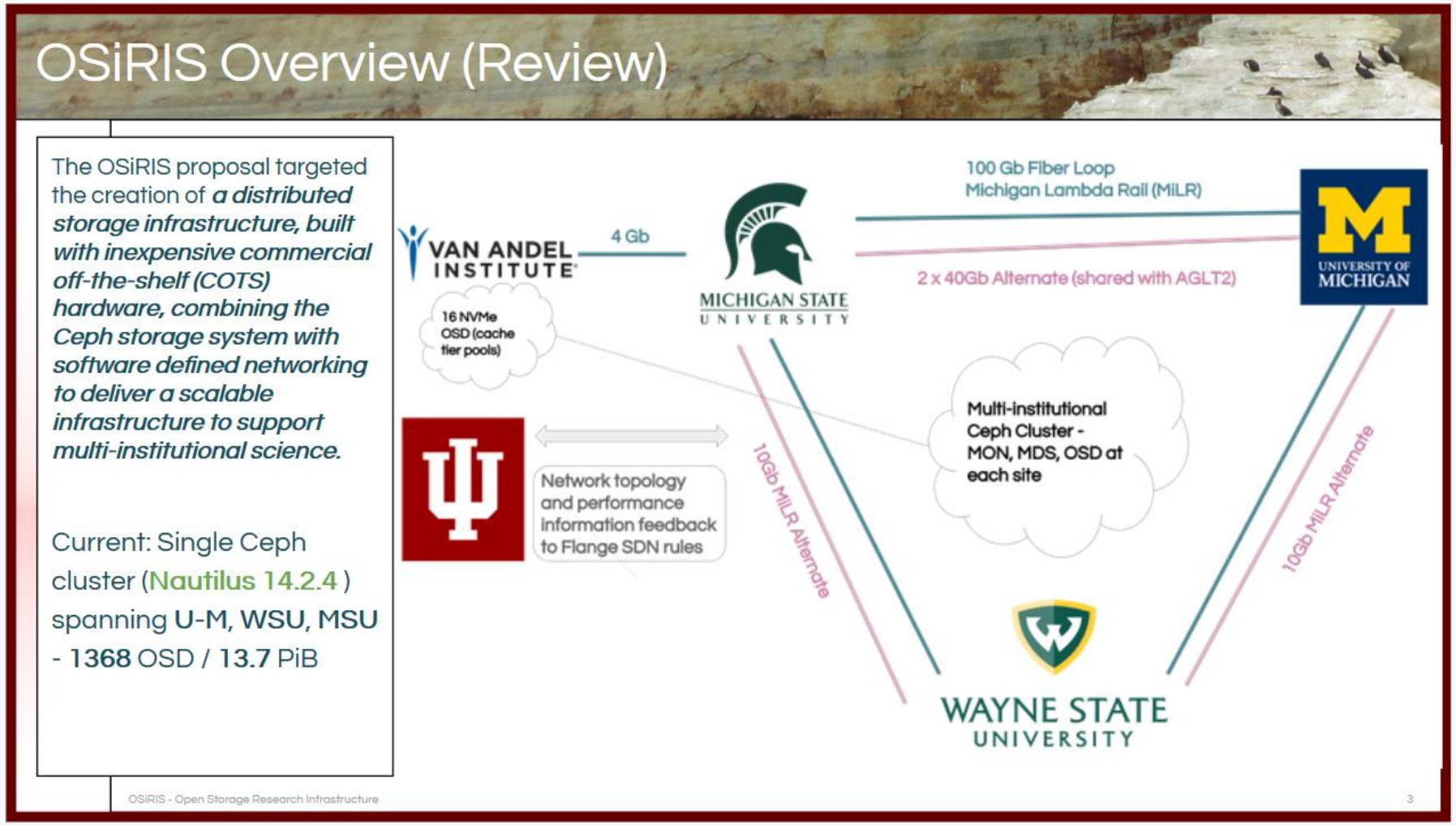
Virtualization Cluster: Rados Block Devices (RBD)

- 13 hypervisors, 78 VMs (growing)
- using libvirt & QEMU / KVM (managed via Foreman)
- *15 TB effective storage, 3 replicas across 3 buildings, all SSDs*

Backup System: Rados Gateway (RGW)

- for user data, device backups and mirroring of RBDs of VMs
- almost all disks 10 years old: 'Make old hardware great again!'
- *64 TB effective storage, 3 replicas across 3 buildings*

University of Michigan



- Multi site, 13.7 PiB CephFS cluster
  - split between three research institutions in the state of Michigan
- Supporting many scientific domains, including HEP
  - facilitate data sharing between researchers at the institutions

Science and Technology Facilities Council

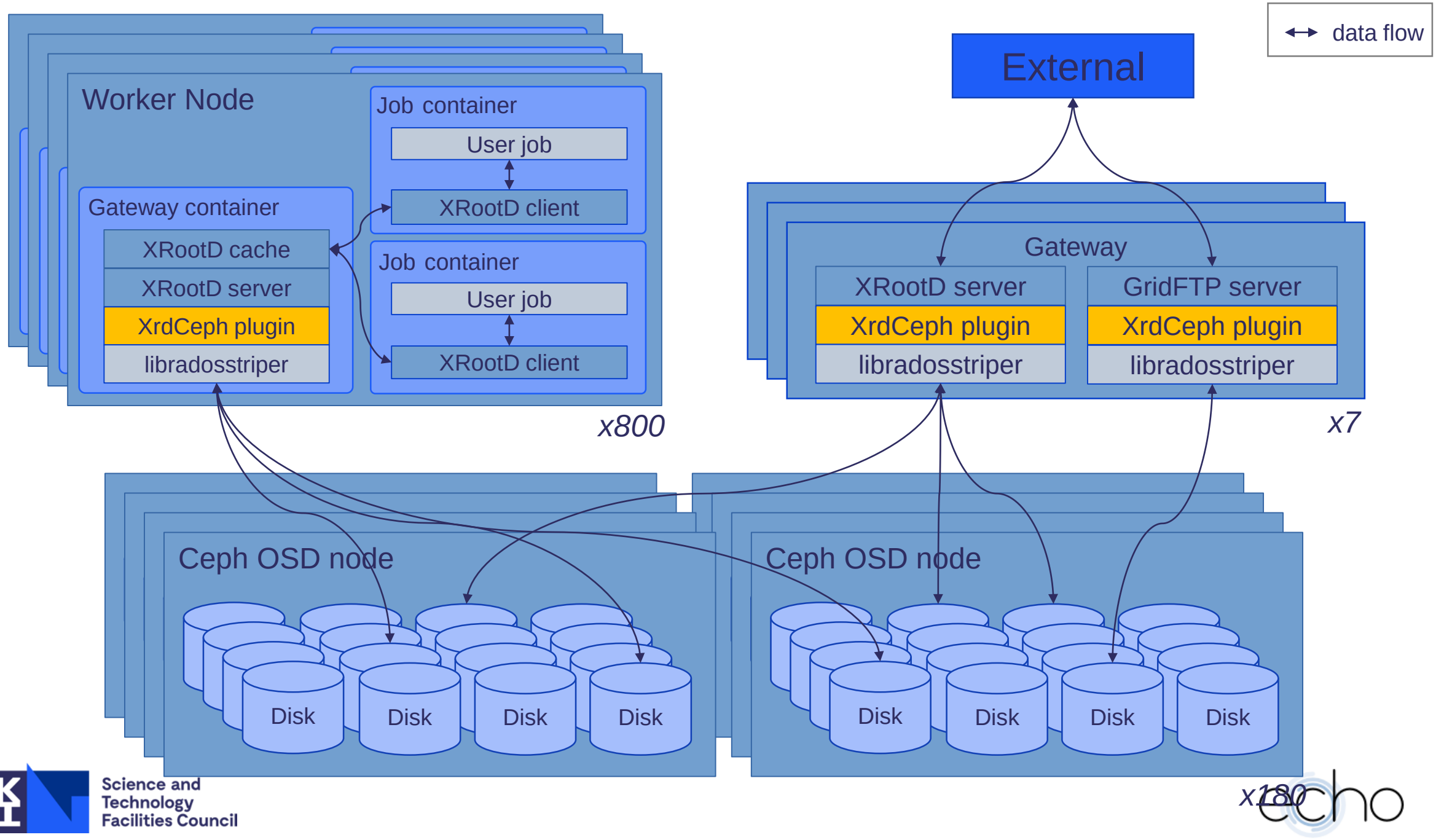
echo

Summary

- RAL will continue to develop XrdCeph to support ongoing WLCG use cases
- Ceph has an exciting development roadmap
  - and large commercial interest
  - and a large, friendly community!
- Usage of Ceph in High energy physics continues to grow in new and exciting ways

Science and
Technology
Facilities Council

Questions?



Science and Technology Facilities Council

echo