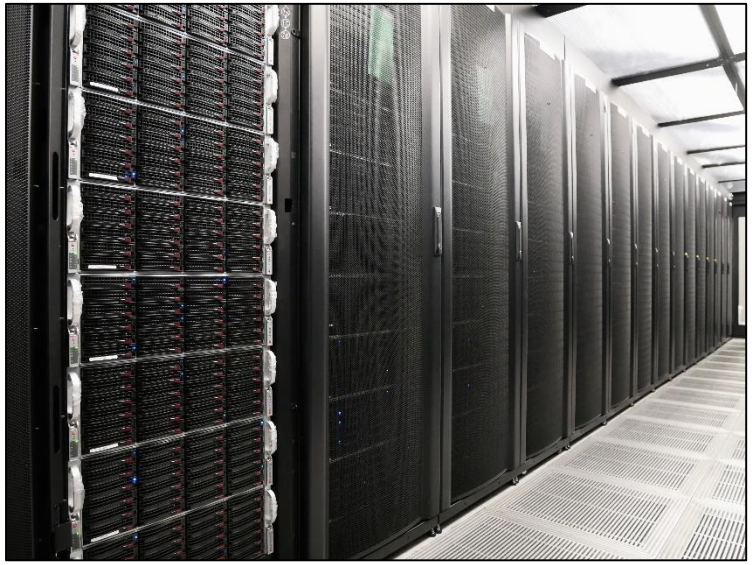
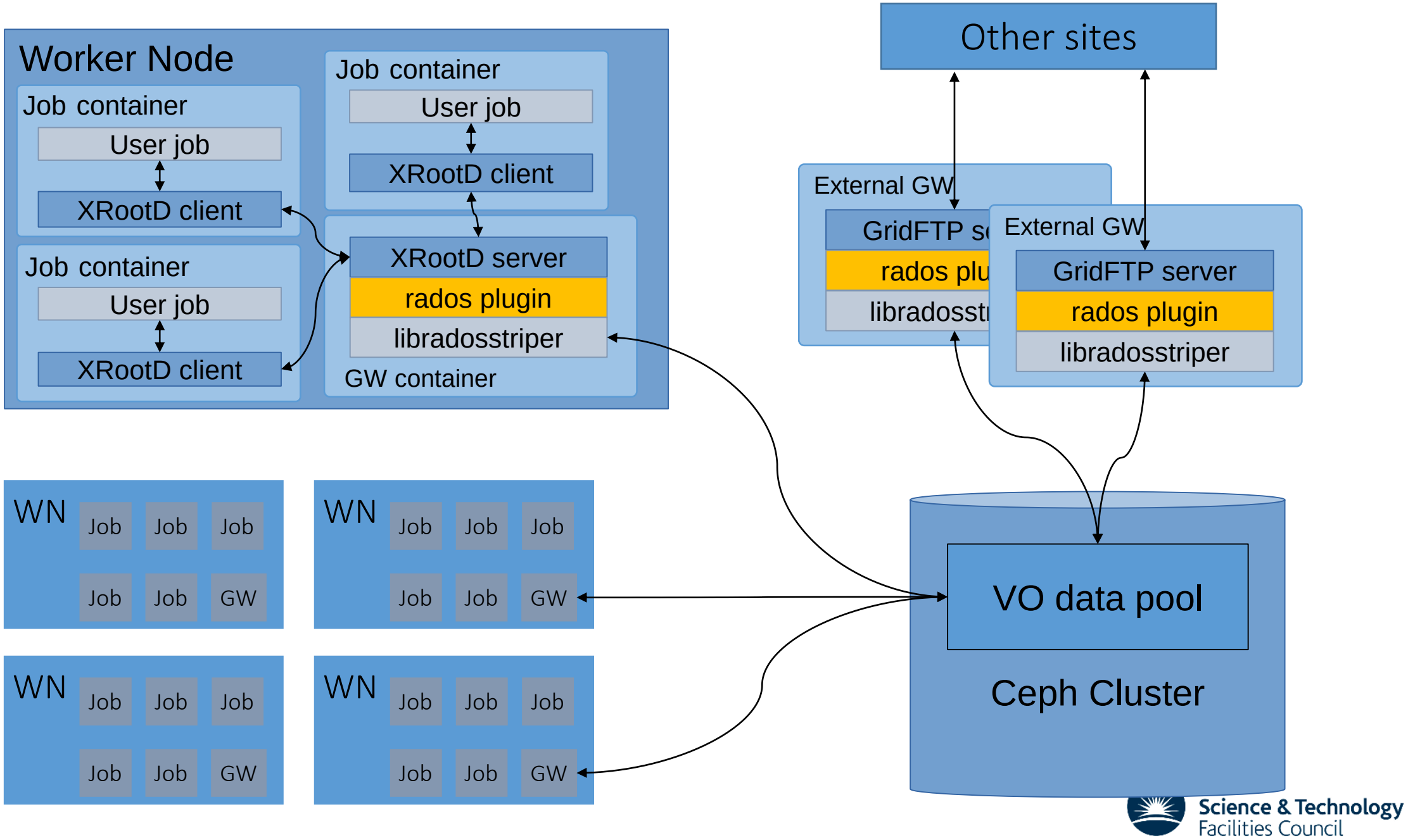
Ceph at RAL – 2019

Tom Byrne

Echo

- Big ceph cluster for WLCG Tier-1 object storage (and other users)
  - 181 Storage nodes
    - 4700 OSDs (6-12TB)
    - 36/28PB raw/usable – 16PB data stored
- Density and throughput over latency
  - EC 8+3
  - 64MB rados objects
- Been through 3 major Ceph versions
  - ~30% OSDs still filestore
  - Mon stores just moved onto RocksDB

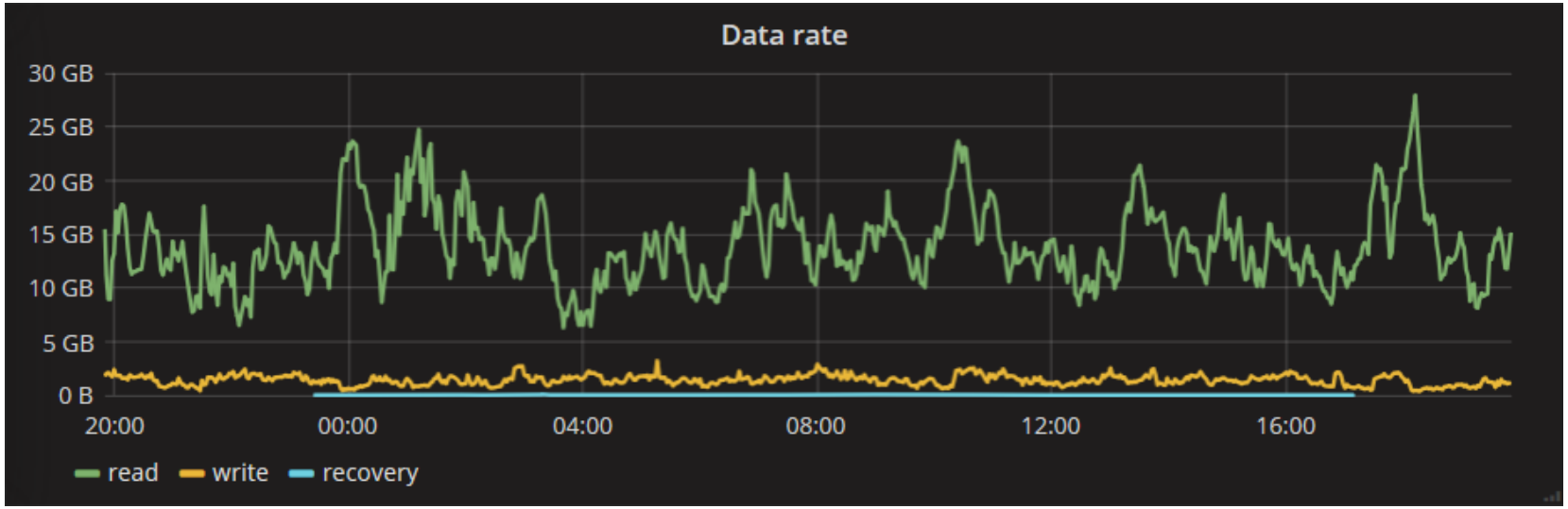



Worker Node
Job container
User job
XRootD client
Job container
User job
XRootD client
Job container
User job
XRootD client
XRootD server
rados plugin
libradosstriper
GW container
Other sites
External GW
GridFTP s
rados plu
libradosstr
External GW
GridFTP server
rados plugin
libradosstriper
WN
Job
Job
Job
Job
Job
GW
WN
Job
Job
Job
Job
Job
GW
WN
Job
Job
Job
Job
Job
GW
WN
Job
Job
Job
Job
Job
GW
VO data pool
Ceph Cluster
Science & Technology
Facilities Council

Echo – good bits

- Data loss incidents have been few and far between
  - Serious data loss incidents have never been directly due to Ceph bugs

- Few performance issues with Echo
  - Ceph performs well for this use case, no bottleneck found yet



Echo – bad bits

- Orchestration
  - Configuration mgmt. stops at ceph.conf.
  - Hardware addition is manageable for ~1000 disks
  - Reboots + upgrades are barely manageable
- Rados striper/custom plugin issues
  - Our main performance issue seems to be due to radosstriper/xrootd plugin interaction
  - Odd behaviour in gftp/xrd plugins and VO workflows have caused issues with VOs loosing files
    - Lots of work to track these down
- Scaling concerns
  - Current model relies on Ceph running well at 6000+ OSDs

Future – Echo

- Add depth to flat crush map
  - Getting racks into the crush map and new crush rules could allow speedy interventions
- Orchestration?
- Re-architect
  - Bigger does not necessarily equal better?

Sirius

- (Currently) Smaller ceph cluster for OpenStack (SCD Cloud)
- Exclusively block device storage
  - 11 hosts
    - 130 4TB HDD BlueStore OSDs
  - 3 replica
- ~15% of storage used, but we're out of IOPS
  - 25+ similar hosts ready to go in
    - Some with SSDs

Sirius – good bits

- Very low maintenance compared to Echo!
- Using ceph ‘as intended’ is remarkably easy ☺
- Fair performance for the hardware, with no tuning

Sirius – bad bits

- Running out of IOPS long before capacity
  - How to use capacity (without having a detrimental effect on RBD experience)

- Lack of transparency as to where IOPS are being used

Future – Sirius

- Lots more hardware, some of the same problems as Echo?
- SSD only nodes?
- CephFS/Manila

Manila

- File shares as a service are highly requested from SCD cloud users
- Works well in testing
  - NFS Ganesha needed for some use cases, high availability and scalability concerns.
- Currently only viewed as being able to provide scratch space
  - Persistent storage a requested feature from other departments