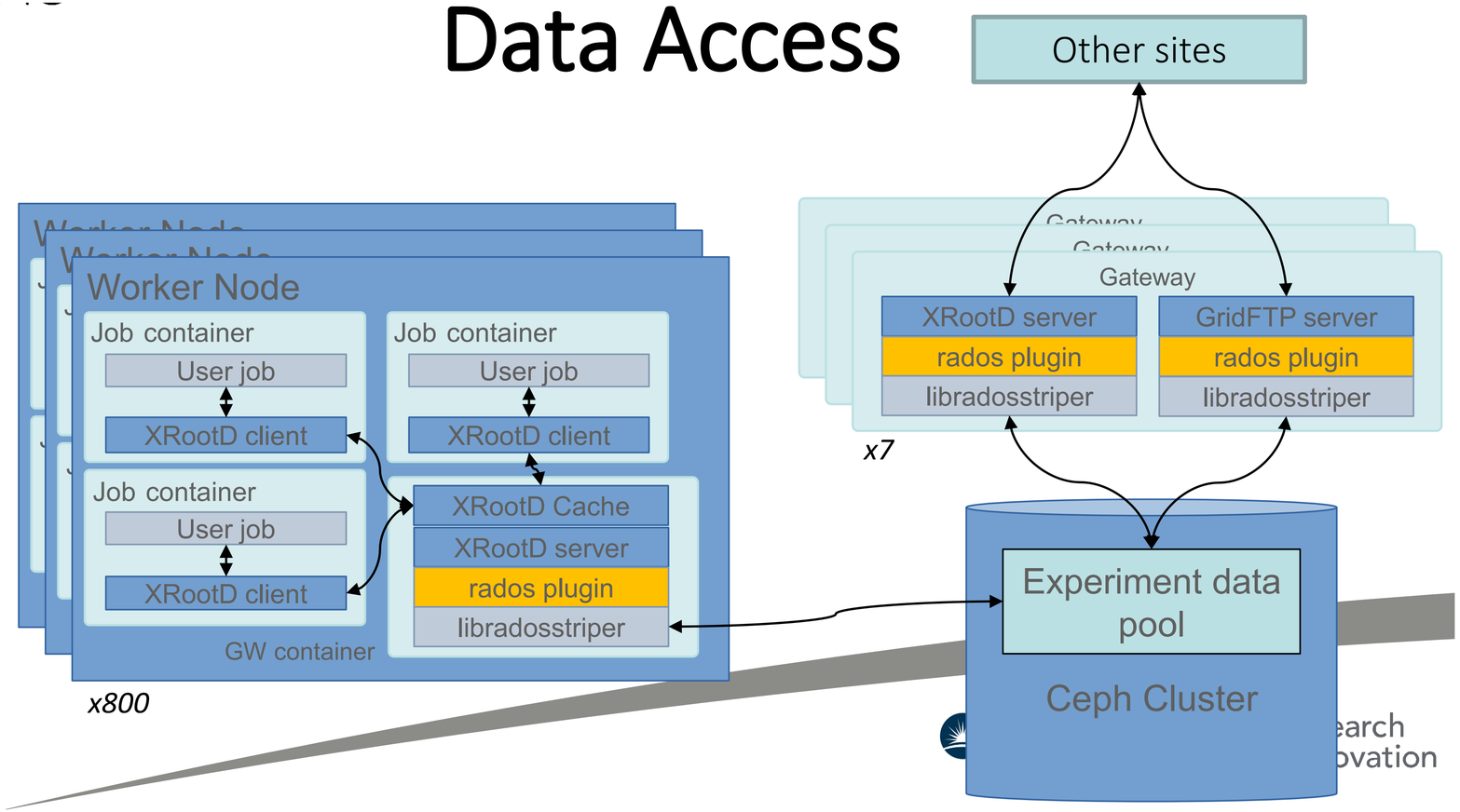
Science and Technology Facilities Council

Vector Read Development

On behalf of RAL T1 xrootd team

RAL Echo Configuration

Data Access



- At the scale we need to run, it is not sustainable to transfer data to worker nodes via the gateway cluster.
- We created gateways running in containers on every WN.
  - Entry in /etc/host directs transfer request to local gateway.
- A small cache on each WN allows pre-fetching of data.
  - This is aligned with the rados object size.
  - Reduces latency and improves throughput from Echo.
- Cache also offers some protection against pathological jobs.

Brief overview of the issue / status

- Noted that in (typically) cases of high load on worker nodes,
  - xrootd will not use the cache to fetch the requested streamed / direct-io data,
  - Instead passes the request through to Ceph resulting in inefficient requests, particularly in readV ops.
- Some mitigation work to alter xrootd buffer/cache sizes, without significant improvement observed.
- Development work to implement readV method within XrdCeph plugin;
  - Now done, however currently passes the ReadV request as a set of individual Read requests to libradosstriper.
  - Uses ceph_posix_pread (of XrdCeph) to call striper->read(fr->name, &bl, count, offset) method
  - https://github.com/stfc/xrootd-ceph/tree/vector_read (diff)

- While Rados appears to have support for sparse_reads,
  - Appears to remain on the ‘todo / improvements’ list of libradosstriper.

- Tentative work in merging readV ops into fewest number of Read requests; however is not a great solution
- Adding functionality into libradosstriper might present the long-term goal for performant operation but should be able to improve on current situation.