

Storage and Data Movement Trends at CHEP

Bo Jayatilaka

Trends in storage and data movement in HEP in the coming decade will be driven by two key factors: the deluge of data expected in the HL-LHC and a computing model that will evolve beyond the relatively homogeneous platforms that evolved since the early 2000s. The former can be summed up simply by noting that the annual raw data production from all LHC experiments will grow from approximately 50 PB in 2016 to 600 PB in 2027¹. In the broadest sense, both of these challenges will require the philosophy behind data storage and management to move further in the direction it already has at the LHC experiments: making collocation of the data with computing optional and relying increasingly on automated data management at sites (including deletion).

Discussion² in the WLCG workshop around this topic centered mainly on whether a move to increased federation of data and including caching can both help reduce costs and levels of duplication. One extreme possibility of this is that the primary compute sites end up not being where the data is stored, resulting in an increased reliance on caching. Current efforts in caching were reported by OSG (federated xrootd-based caching) and RAL (S3/Ceph-based caching to “storageless” tier 2s). Most of these proposed solutions involve much heavier load on networks than we currently have.

CERN reported storage progress on several fronts: EOS, XRootD, Ceph, and some other general comments on data services³. With EOS⁴, CERN is moving away from AFS (which is being phased out CERN-wide) and developing direct FUSE mounts instead. CERN (specifically within openLab) is also exploring the use of Seagate Kinetic drives and has adapted EOS to work with this architecture (which depends less on data servers). CERN currently has 850M files in 150PB of storage in EOS. CERN has also deployed multiple Ceph⁵ pools, including a 30 PB pool that is primarily a testing area. With the 30PB pool, performance of up to 30GB/s was demonstrated (previously not achieved in pools > 10 PB).

Ceph is the focus of object-store setups outside of CERN as well. RAL has the last WLCG site to use Castor for disk storage and rather than move to EOS they moved to Ceph⁶. Performance in their case was network limited (10 Gbps). Meanwhile, the dCache team has implemented dCache as a service (namespace, authentication, doors) to sit on top of storage in the form of Ceph (or any other object store-- emphasis was placed on the Ceph-specific aspect as very

¹ [Bird \(WLCG\)](#)

² [Keeble et al \(Data management session, WLCG\)](#)

³ [Curull](#) “CERN Data Services for LHC Computing”

⁴ [Sindirallu](#) “EOS Developments”

⁵ [Curull](#) “CERN’s Ceph Infrastructure: OpenStack, NFS, CVMFS, CASTOR, and more!”

⁶ [Dewhurst](#) “The deployment of a large-scale object store at RAL Tier 1”

small)⁷. This will be available in dCache 3.0 but the initial sense I got was that performance was poor (not reported in the talk).

The LHC experiments reported on the status of their dynamic data management systems. ATLAS has upgraded their system from run 1 to use Rucio (developed at CERN) to handle management⁸. The new system also supports objectstores as an endpoint. The updated CMS system “Dynamo” now also handles automatically replicating popular datasets as well as deleting underused copies⁹. The LHC experiments have also explored using machine learning as a means for predicting dataset popularity, of which a talk was presented by LHCb¹⁰. Such techniques will likely need to be the norm during the HL-LHC (the ATLAS talk included a projection that while their current system will scale to Run 3, it won’t for Run 4). Coming back to the WLCG discussion, the possible pure-caching future would start with the existing caching setups. ATLAS and OSG have both deployed XrootD-based caches. The ATLAS caches are intended to be used by Tier3s, opportunistic resources, and some Tier2s¹¹. The OSG cache (StashCache) implements POSIX file access by way of CVMFS¹². CMS has gone a slightly different way, implementing HTTP-based caching in part to reach sites that may not support XrootD¹³. As with much of the shift towards relying on network to carry more of the load in a future with more caching, many of these tests emphasized saturating available links as one of their goals.

An interesting proposal was also made by Ian Fisk in his plenary talk about data storage and movement¹⁴. A dramatic shift in our computing model that could help reduce the footprint associated with data storage is to have specialized “data reduction centers” where data can be reduced by multiple orders of magnitude and then replicated to resources where more compute heavy analysis might take place.

⁷ [Mkrtchyan](#) “dCache - outsourced storage”

⁸ [Garonne](#) “Experiences with the new ATLAS distributed data management system”

⁹ [Iiyama](#) “Dynamo - The dynamic data management system for the distributed CMS computing system”

¹⁰ [Hushchyn](#) “GRID storage optimization in transparent and user-friendly way for LHCb datasets”

¹¹ [Gardner](#) “Caching servers for ATLAS”

¹² [Weitzel](#) “Accessing data federations with CVMFS”

¹³ [Vlimant](#) “HTTP as a data access protocol: Trials with XrootD in CMS’ AAA project”

¹⁴ [Fisk](#) “Future of distributed data and workflow management”