21st International Conference on Computing in High Energy and Nuclear Physics **CHEP2015** Okinawa Japan: April 13 - 17, 2015

Contribution ID: **443** Type: **oral presentation**

Engineering the CernVM-FileSystem as a High Bandwidth Distributed Filesystem for Auxiliary Physics Data

Monday, April 13, 2015 5:30 PM (15 minutes)

Fermilab has several physics experiments including NOvA, MicroBooNE, and the Dark Energy Survey that have computing grid-based applications that need to read from a shared set of data files. We call this type of data Auxiliary data to distinguish it from (a) Event data which tends to be different for every job, and (b) Conditions data which tends to be the same for each job in a batch of jobs. Conditions data also tends to be relatively small per job (100s of Megabytes) where both Event data and Auxiliary data are larger per job (Gigabytes), but unlike Event data, Auxiliary data comes from a limited working set (10s to 100s of Gigabytes) of shared files. Since there is some sharing of the Auxiliary data, it appeared at first that the CernVM-Filesystem (CVMFS) infrastructure built for distributing software to the grid would also be the best way to distribute Auxiliary data. However, because grid jobs tend to be started in large batches running the same code, the software distribution use case of CVMFS has a very high cache hit ratio, so the bandwidth requirements on the per-site http proxy caches (squids) is quite low. As a result those proxy caches have been engineered with relatively low bandwidth, and they are easily overwhelmed by Auxiliary data with its relatively low cache hit ratio. A new approach was needed. We are taking advantage of a CVMFS client feature called “alien cache” to cache data on site-local high-bandwidth data servers that were engineered for storing Event data. This site-shared cache replaces caching CVMFS files on both the worker node local disks and on the site-local squids. We have tested this alien cache with the dCache NFSv4.1 interface, Lustre, and Hadoop-FS-fuse, and found that they all perform well. In addition, we use high-bandwidth data servers at central sites to perform the CVMFS Stratum 1 function instead of the low-bandwidth web servers deployed for the CVMFS software distribution function. We have tested this using the dCache http interface. As a result, we have an end-to-end high-bandwidth widely distributed caching read-only filesystem, using existing client software already widely deployed to grid worker nodes and existing file servers already widely installed at grid sites. Files are published in a central place and are soon available on demand throughout the grid and cached locally on the site with a convenient POSIX interface. This paper discusses the details of the architecture and reports performance measurements.

Primary author: DYKSTRA, Dave (Fermi National Accelerator Lab. (US))

Co-authors: BOCKELMAN, Brian Paul (University of Nebraska (US)); BLOMER, Jakob (CERN); HERNER, Ken (Fermilab); SLYZ, Marko (Fermilab); LEVSHINA, Tanya (Fermilab)

Presenter: BLOMER, Jakob (CERN)

Session Classification: Track 3 Session

Track Classification: Track3: Data store and access