Contribution ID: **46** Type: **Long talk**

Evaluating CephFS Performance vs. Cost on High-Density Commodity Disk Servers

Thursday, 20 May 2021 09:30 (30 minutes)

CephFS is a network filesystem built upon the Reliable Autonomic Distributed Object Store (RADOS). At CERN we have demonstrated its reliability and elasticity while operating several 100-to-1000TB clusters which provide NFS-like storage to infrastructure applications and services. At the same time, our lab developed EOS to offer high performance 100PB-scale storage for the LHC at extremely low costs, while also supporting the complete set of security and functional APIs required by the particle-physics user community. This work seeks to evaluate the performance of CephFS on this cost-optimized hardware when it is combined with EOS to support the missing functionalities. To this end, we have setup a proof-of-concept Ceph Octopus cluster on high-density JBOD servers (840 TB each) with 100Gig-E networking. The system uses EOS to provide an overlayed namespace and protocol gateways for HTTP(S) and XROOTD, and uses CephFS as an erasure-coded object storage backend. The solution also enables operators to aggregate several CephFS instances and adds features such as third-party-copy, SciTokens, and high-level user and quota management. Using simple benchmarks we measure the cost/performance tradeoffs of different erasure-coding layouts, as well as the network overheads of these coding schemes. We demonstrate some relevant limitations of the CephFS metadata server and offer improved tunings which can be generally applicable. To conclude, we reflect on the advantages and drawbacks related to this architecture, such as RADOS-level free space requirements and double-network penalties, and offer ideas for improvements in the future.

Primary authors: PETERS, Andreas Joachim (CERN); VAN DER STER, Dan (CERN)

Presenter: VAN DER STER, Dan (CERN)

Session Classification: Thurs AM Plenaries

Track Classification: Distributed Computing, Data Management and Facilities