Achieving Cost/Performance Balance Ratio Using Tiered Storage Caching Techniques

Michael Poat, Jerome Lauret, Brookhaven National Laboratory

Case-study

- STAR CephFS storage cluster: 30 nodes, 4 x 2TB Seagate SAS HDD each. Can we use SSDs to speed up IO performance at low cost?
- From previous work: Tiered storage not compatible with CephFS, CephFS as fast as the slowest element

  M. Poat, J. Lauret – "Performance and Advanced Data Placement Techniques with Ceph's Distributed Storage System", J. Phys.: Conf. Ser. 223 – To be published.
- **Idea**: Implement low-level disk-caching techniques (bcache & dm-cache)

Method

- IOzone used for performance measurement, study IO threads scaling
- Tested bare drives HDD/SSD, bcache and dm-cache

Key observations

- dm-cache performance more stable than bcache (IO crash observed)
- Stand-alone test shows RAID0 as performant as SSD, dm-cache trailing
- In CephFS, outcome diverge – poor SSD performance observed, RAID0 not a gain over stock cluster (faster gain with more OSD), dm-cache provides no apparent gain
- Issue traced to IO competition between data and journaling ⇔ lack of in-memory cache (PLP) on cheap SSD forces flush/sync operations
- Moving the journal on a separate device changes outline: dm-cache wins (and only 1:1 configuration, not 1:3), followed by SSD and dm-cache without isolated journal ops

Lesson to learn: Test SSDs carefully, featureless SSDs not the best choice, enterprise SSD a better choice (at a cost)