SC19 Network Research Exhibition: Demonstration Abstract

Responsive Cache Tiering with Ceph and OSiRIS

Ben Meekhof, University of Michigan, bmeekhof@umich.edu

Abstract

Our project, OSiRIS, is a multi-institutional research storage platform comprised of a Ceph cluster spanning 3 major Michigan research institutions. The nature of a distributed platform like Ceph places certain latency-based limitations on the distance between cluster storage elements. In past Supercomputing conferences we have explored Ceph 'cache tiering' which allows us to place a more localized pool of storage as a transparent tier on top of a larger pool of storage localized in Michigan.

These tests certainly proved that performance for clients near the cache pool at SC could be improved while cache flushes via the higher-latency connection back to Michigan could occur 'out of band' after the client had completed I/O. They also showed the expected result that performance for clients not near the cache storage was correspondingly much lower due to the latency to reach the cache. Ceph has no mechanism to direct clients to one pool or another based on location - if there is a cache tier overlay then all clients are transparently directed to go through it.

We propose to experiment with the rapid creation and deletion of cache tiers to flexibly respond to client I/O needs in differing locations. As such we would create several cache tier pools on hardware at SC and perform various types of synthetic data I/O simulating clients at SC and also in Michigan which presumably will favor SC clients. We would then flush and delete these pools to arrive at a configuration more favorable to Michigan clients, and repeat for some number of iterations. Variables include the size of cache tier pools, tuning parameters that affect flushing to the backing pool, and potentially introducing a responsive component that automates this process based on different factors.

Coordination of cache tiers and network traffic monitoring will be handled by a plugin module for the Ceph manager daemon with configurable parameters for client proximity, cache locality, and pool / cache associations. The module will also include commands to internally simulate high traffic from any given location to not necessarily require a fully developed client traffic view for demo purposes. Tentatively we believe such information could be provided to the module by querying an Elasticsearch database fed by the Elastiflow sflow data gathering tool.

Goals

1. To explore potential problems with rapidly creating/destroying Ceph cache tiers as described
2. Determine an optimal size/configuration for cache tiers in the described usage
3. Demonstrate the viability of using cache tiering as a dynamic, responsive mechanism to compensate for higher latency in Ceph deployments
4. Implement a responsive mechanism for creating temporary cache tiers as indicated by client usage patterns, or at least have tools on which such a mechanism could be implemented.

Resources

Our cache tier pool(s) will be created on hardware located in our booth at SC, and backing storage will be located at our institutions in Michigan. Other requirements:

- High speed (100Gbit) WAN links between Michigan and our booth
- A block of IP addresses in the SC (SCInet) network to host local clients
- Power connection to the booth capable of supporting multiple data storage servers (30 amp circuit).

Involved Parties

- Shawn McKee, University of Michigan, smckee@umich.edu
- Ben Meekhof, University of Michigan, bmeekhof@umich.edu
- Patrick Gossman, Wayne State University, ac8456@wayne.edu
- Kenneth Merz, Michigan State University, merzjrke@msu.edu
- Andy Keen, Michigan State University, keenandr@msu.edu
- Michael Thompson, Wayne State University, Michael@wayne.edu