SC19 Network Research Exhibition: Demonstration Abstract

Dynamic Traffic Management for Ceph and OSiRIS

Ezra Kissel, Indiana University, ezkissel@iu.edu

Abstract

The Open Storage Research Infrastructure (OSiRIS) is a Ceph-based multi-institutional storage platform spanning three core deployment sites at Michigan research universities. The Network Management Abstraction Layer (NMAL) incorporates and extends PerfSONAR capabilities to include automated network topology discovery and tracking and incorporates Software Defined Networking (SDN) into overall operations of the OSiRIS infrastructure.

With OSiRIS expanding to sites with members that do not have bandwidth parity with the core deployment, one of the concerns with Ceph is effectively managing replication and recovery traffic that can potentially overload bottleneck links. This type of Ceph traffic is crucial to the distributed and resilient operation of the cluster but is often difficult to manage effectively using application tuning alone.

There are two main concerns we will investigate:

1. Replication and backfill operations can essentially create a denial-of-service situation at a bandwidth-limited site to the point where client operations are affected.

2. Transport protocols like TCP used by Ceph can behave poorly in circumstances where you have network capacity mismatches, especially at higher latencies.

We propose to address these issues with a combination of SDN-based quality-of-service (QoS) techniques using OpenVSwitch, network monitoring, and declarative, programmatic control of the networking using a domain-specific language called Flange that are incorporated into NMAL. At SC, we will manage the Ceph traffic between our local SC OSiRIS deployment and the WAN links to our core deployment in Michigan. Traffic shaping queues will be used to limit the rate of traffic to the SC cluster, and priority queues will be used to ensure that client traffic is given enough bandwidth to ensure responsiveness of service. Network monitoring will inform the active Flange program to dynamically adapt the shaping parameters as appropriate.

Goals

1. Quantify the benefits of selective traffic shaping and priority queuing on an extended Ceph cluster over a WAN deployment.
2. Determine the usefulness and level of dynamic control needed (via a reactive Flange program) to adjust the shaping parameters based on network conditions.
3. Explore the viability of NMAL to discover the local SC cluster resources and integrate available monitoring capabilities.
4. Demonstrate and develop a model for deploying a working QoS solution for Ceph through containerization, host integration, and DevOps practices.

Resources

This proposal will leverage the hardware and networking capabilities within the University of Michigan booth at SC and co-exist with the proposed *Responsive Cache Tiering with Ceph and OSIRIS* NRE submission. Other requirements specific to this proposal:

- Interaction with the SCinet Measurement Team to collect WAN and booth connection metrics.
- The ability to have or induce alternate WAN paths with different characteristics (capacity, asymmetry) to explore dynamic selection and adaptation at the booth side. Specifically, we plan to have a 100G WAN path back to Michigan for SC19 but would like to have additional (shared) access to alternative paths we could use to orchestrate our traffic.

Involved Parties

- Jeremy Musser, Indiana University, jemusser@iu.edu
- Grant Skipper, Indiana University, gskipper@iu.edu
- Paventhan Vivekenandan, pvivekan@iu.edu
- Ezra Kissel, Indiana University, ezkissel@iu.edu
- Martin Swany, Indiana University, swany@iu.edu
- Ben Meekhof, University of Michigan, bmeekhof@umich.edu
- Shawn McKee, University of Michigan, smckee@umich.edu