RUTGERS
WINLAB | Wireless Information Network Laboratory

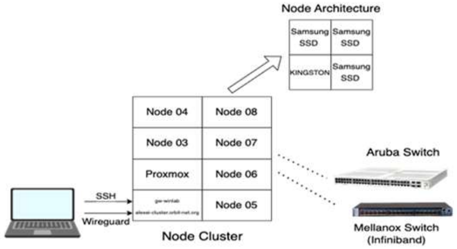
Distributed Data Infrastructure

Samyak Agarwal, Keshav Subramaniyam, Anna Kotelnikov, Gyana Deepika Dasara, Jason Zhiyuan Zhang
Advisor: Professor Alexei Kotelnikov

Project Goal

- Test the performance of CephFS, an open source distributed file system
- Note how changes to different configurations (ie. number of placement groups, redundancy algorithm, etc.) affect performance

Hardware



Gateway (Node 01):

- Provides gateway), wireguard vpn, DHCP
- Hosts FOG, and Debian .iso sharing

Clients (Node 02):

- 8 Linux containers (lxc01-lxc08) on Proxmox serve as clients to access the storage clusters.

Cluster File Servers (Node 03-08):

Each server contains:

- 1 KINGSTON SA400S3 (447 GiB)
- 3 Samsung SSD 870 (466 GiB)

Aruba Switch:

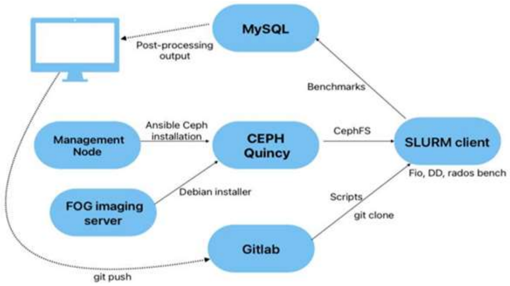
- Version: Aruba Instant On 1930 48G 4SFP/SFP+ Switch (JL685A)
- Line Rate: 1 GbE

Mellanox Switch:

- Version: Mellanox MLNX-OS SX6036
- Line Rate: 40 Gb IPoIB
- Offers InfiniBand support

Workflow

Automated workflow using Ansible playbooks to install and configure CEPH and SLURM to schedule benchmarking tasks.



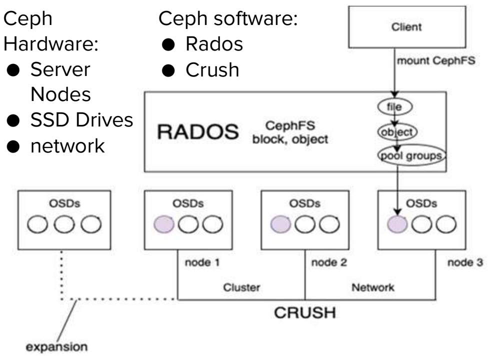
Ceph

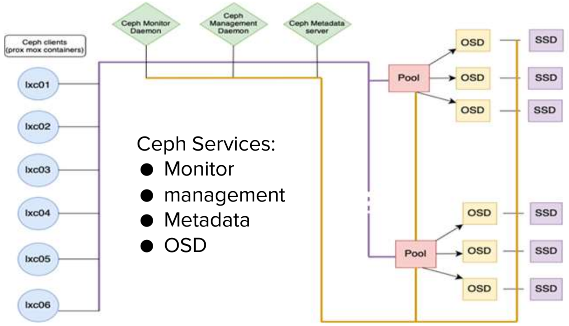
Ceph Hardware:

- Server Nodes
- SSD Drives
- network

Ceph software:

- Rados
- Crush





Ceph Services:

- Monitor
- management
- Metadata
- OSD

Redundancy/Expandability

Ceph ensures data redundancy through replication and erasure coding

Replication

- For each piece of data, several copies are generated and stored
- High performance at cost of low disk usage efficiency

Erasure Coding

- Breaks data into smaller fragments, generates parity bits to compute lost data in case of drive failure or any data loss
- Offers higher storage efficiency than at increased computational cost

With Ceph, we can also expand clusters extremely quickly and easily. With OSDs abstracting disks, we can replace an OSD with any drive, and expand the cluster infinitely by adding more servers to the cluster

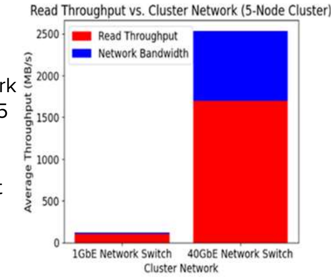
Results

1 GbE vs 40 Gb IPoIB Switch

Using iperf and rados bench,

- On 1 GbE switch, read throughput is similar to network bandwidth (105.3 MB/s vs 117.5 MB/s)
- On 40 Gb switch, significant gap between read throughput and network bandwidth (1.61 GB/s vs 2.45 GB/s).

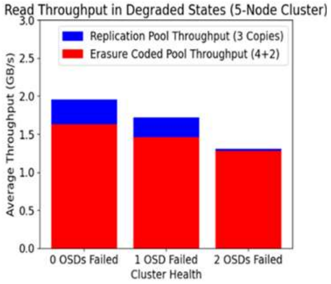
Need 40Gb+ network switch to avoid network bottleneck



Erasure Coding vs Replication in Disaster Recovery

Disaster Recovery occurs when an OSD or node fails.

- In clean states, generally replication pools have better throughput than erasure-coded pools
- As OSDs fail, erasure-coded pools experience a smaller dropoff in throughput



Future Work

Application Specific Performance

- Our research was sponsored by, Nverses Capital, a hedge fund which utilizes machine learning.
- We hope to explore CephFS' performance for computationally intensive and machine learning applications.

Large Scale Disaster Recovery

- We only tested failure of individual OSDs
- In the future, we hope to explore how Ceph handles the failure of entire nodes with the quorum voting

Acknowledgement

Our group would like to acknowledge NVerses Capital for their financial support as well as JB Kim, Ian Roddis, Sam Silver and our advisor Alexei Kotelnikov for their guidance and technical contributions.

WINLAB