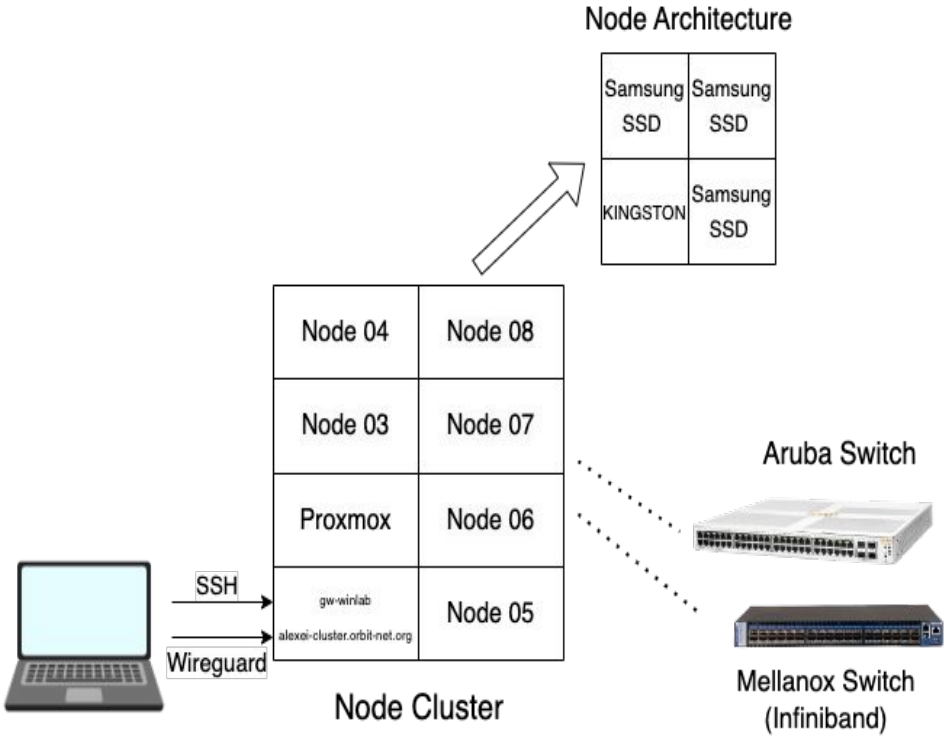
Distributed Data Infrastructure

By: Anna, Samyak, Keshav, Jason, Deepika

Project Goal

- Test the performance of CephFS
- Note how configuration changes affect performance
- Compare Ceph performance with other distributed file systems

Hardware Architecture



Gateway (Node 01):

- Provides gateway (gw-winlab), wireguard vpn, DHCP (dynamic host control protocol)
- Also hosts FOG, and Debian .iso sharing

Clients (Node 02):

- 8 Linux containers (lxc01-lxc08) on Proxmox serve as clients to access the storage clusters.
- Gitlab, Slurm, Database, Grafana

Cluster File Servers (Node 03 - 08):

- Each server contains:
  - 1 KINGSTON SA400S3 (447 GiB)
  - 3 Samsung SSD 870 (466 GiB)

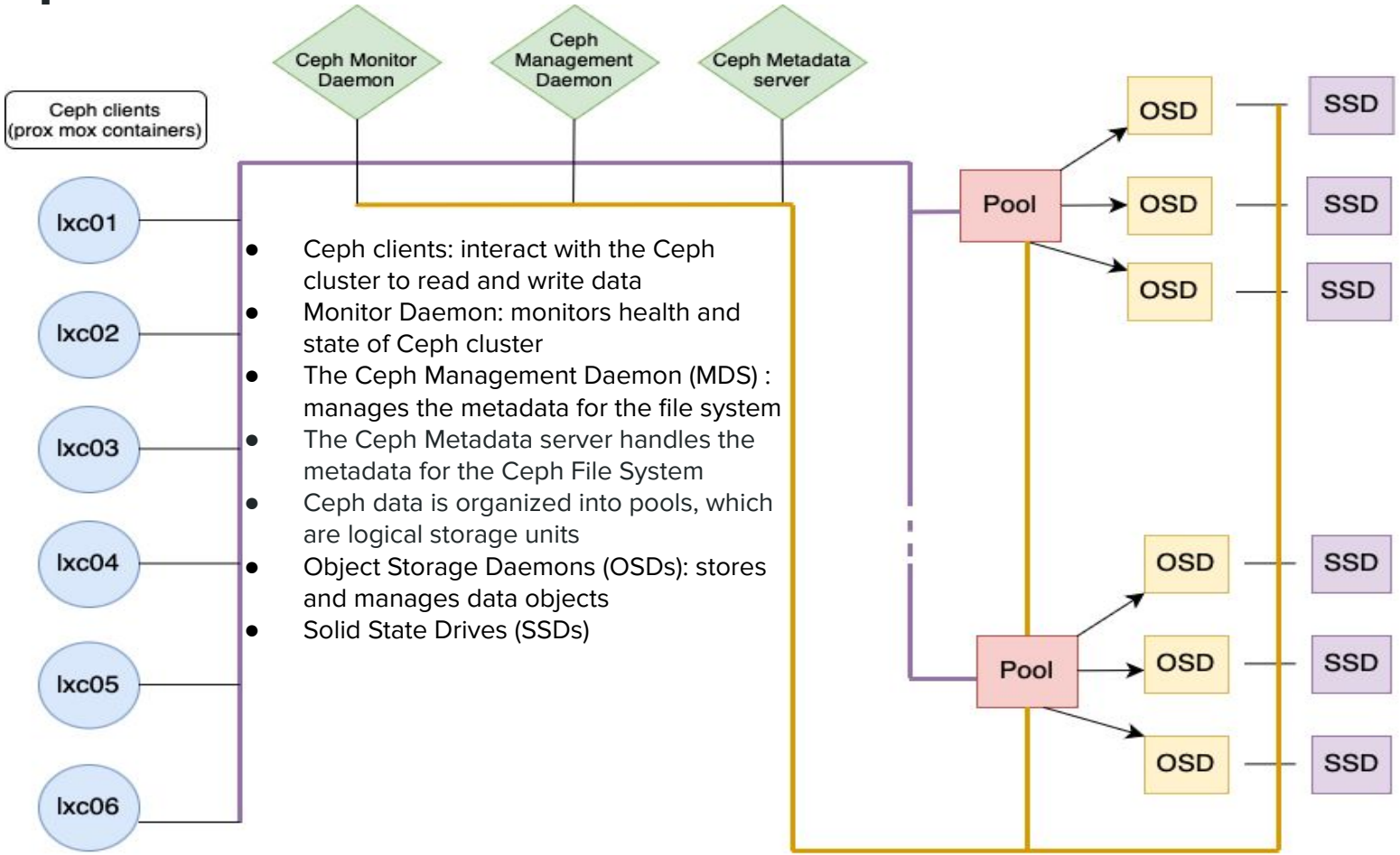
Aruba Switch:

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

Mellanox Switch:

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

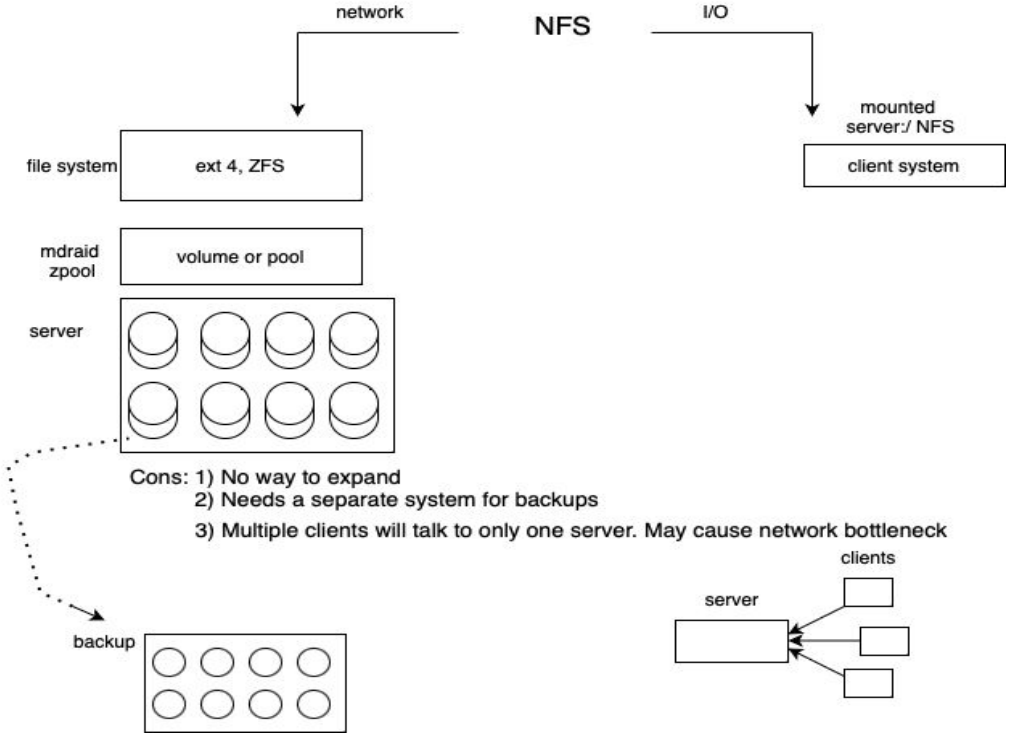
Ceph Overall Architecture



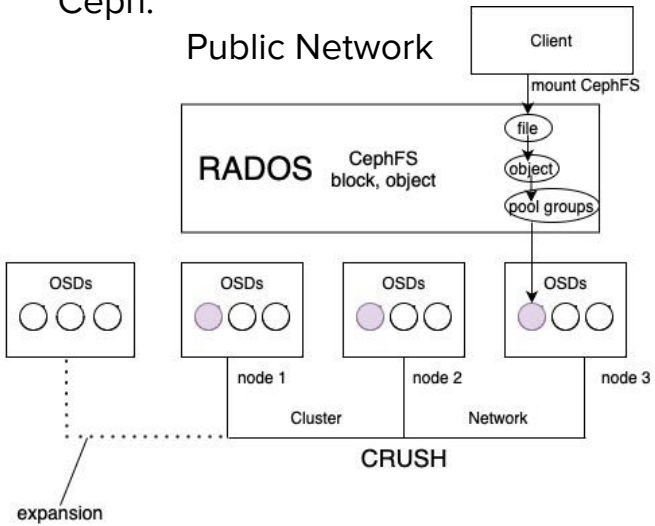
- Ceph clients: interact with the Ceph cluster to read and write data
- Monitor Daemon: monitors health and state of Ceph cluster
- The Ceph Management Daemon (MDS) : manages the metadata for the file system
- The Ceph Metadata server handles the metadata for the Ceph File System
- Ceph data is organized into pools, which are logical storage units
- Object Storage Daemons (OSDs): stores and manages data objects
- Solid State Drives (SSDs)

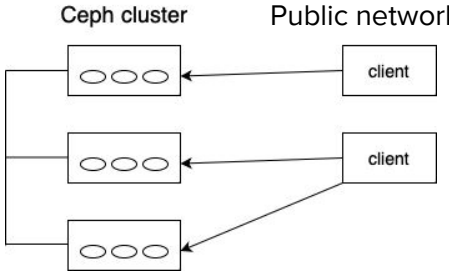
Ceph Vs. Classic File System:

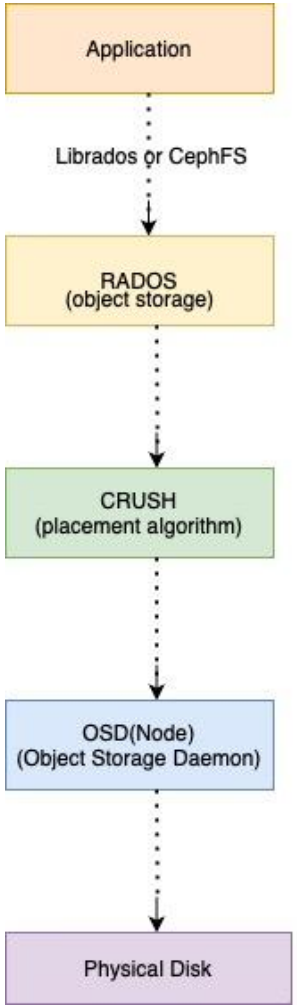
Classical file system like NFS:



Ceph:







Rados and Crush in Ceph

1. Application

- Interacts with Ceph through Librados and CephFS.
- Sends read and write requests to RADOS to store and retrieve data.

2.RADOS (Reliable Autonomic Distributed Object Store): Manages data storage and retrieval

3.CRUSH (Controlled Replication Under Scalable Hashing):

- Placement Data Calculation
- Uses crush map to map data into OSDs

4.OSD (Object Storage Daemon)

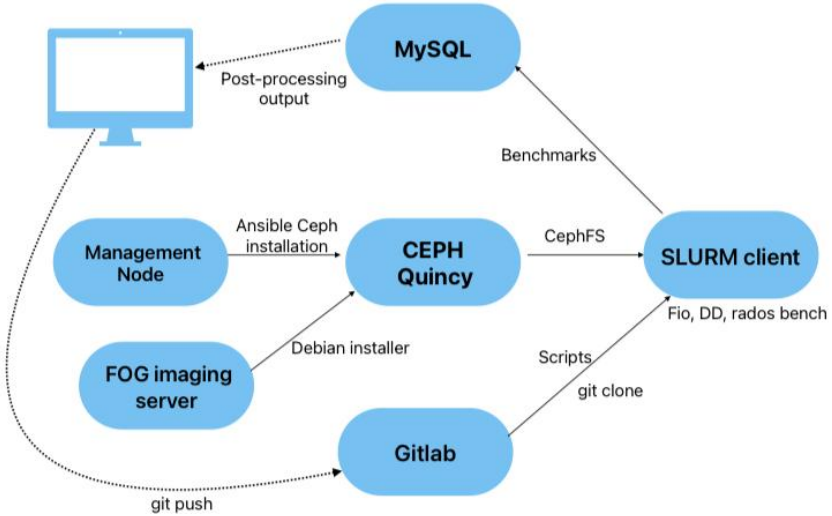
5.Physical Disk:

- Where the data is stored.
- OSDs manage data placement, replication, and recovery on these disks.

Workflow

Automated Workflow-

- Clean Debian install by booting into Fog installer
- Ansible playbooks to setup and configure Ceph
- Gitlab to store our jobfiles and scripts
- SLURM to schedule benchmarking jobs
- MySQL to store our benchmarks output



Benchmarking tools

- DD: used to perform basic I/O operations
- Fio: it is used to simulate more complex I/O patterns, block sizes, read/write ratio, queue depth, etc.
- Rados Bench: it is specific to Ceph.

Redundancy

Replication-

- Data is replicated and stored in form of objects
- Ceph uses RADOS to distribute objects among OSDs
- RADOS divides objects into placement groups
- CRUSH is used to determine how data is distributed and replicated

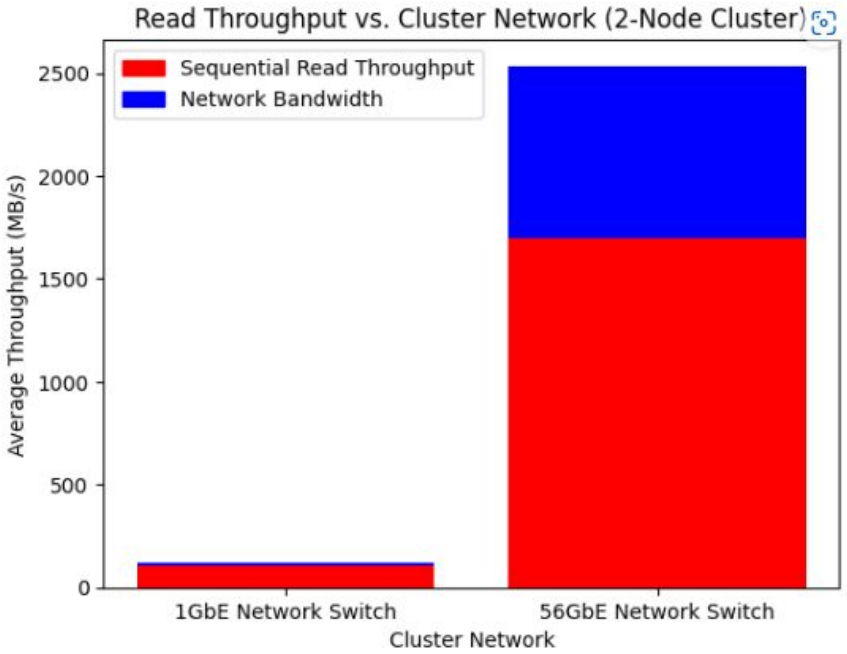
Erasure Coding-

- Offers higher storage efficiency than replication at increased computational cost
- Breaks data into smaller fragments, generates parity bits and are distributed across all OSDs
- Parity bits are used to regain lost data in case of drive failure or any data loss

Results

1 GbE vs 40 Gb IPoIB Network Switch

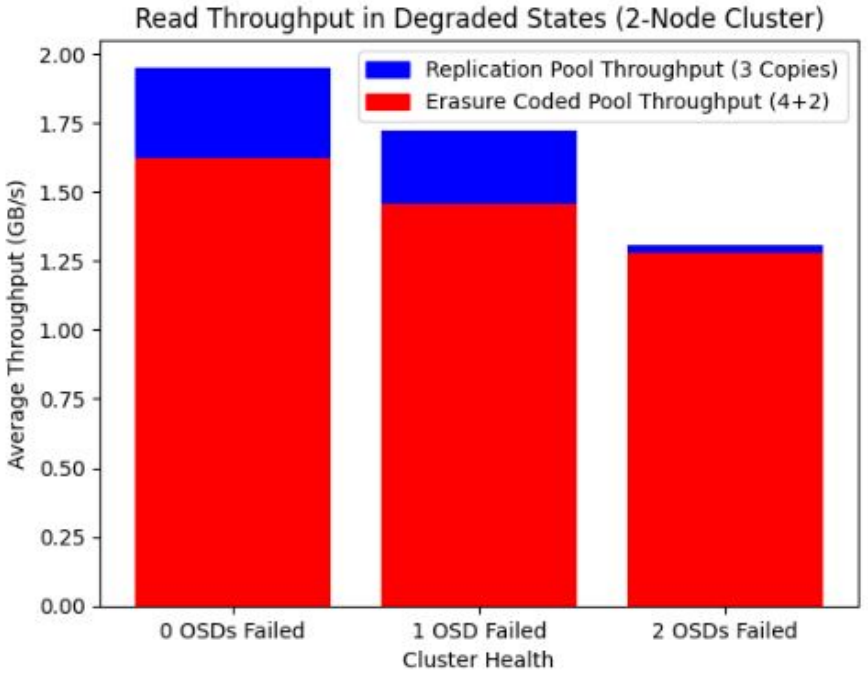
- To test the impacts of network switches, we utilized iperf and rados bench to compare the network bandwidth vs file system throughput when using different switches
- On 1 GbE Aruba Switch, throughput is close to network bandwidth (105.3 MB/s vs 117.5 MB/s)
- On 40 GB IPoIB Mellanox switch, there is a gap between CephFS and network bandwidth (1.65 GB/s vs 2.45 GB/s)



Results-2

Erasure Coding vs Replication in Disaster Recovery

- Disaster Recovery occurs when an OSD or node fails
- When looking at the impact on throughput for erasure-coded and replication pools:
  - In clean states, replication outperforms erasure-coded pools
  - As OSDs fail, erasure-coded pools experience a smaller drop off in throughput



Future Work

- Further explore Ceph performance in relation to machine learning workflows of the Nverses Capital and to continue to optimize the system’s performance for its application.
- We have already done some analysis on performance testing with up to 3 OSD failures. We need to explore how Ceph handles more than 3 OSD failures.
- We need to also see how Ceph handles the failure of entire nodes with quorum voting.

Our group would like to acknowledge the people at nVerses Capital, Kinesin Data Technologies, our advisor Alexei Kotelnikov, Jenny Shane, Ivan Seskar, Jakub Kolodziejski and the rest of the WINLAB staff for making this project possible.