Ceph storage

Lester Vecsey

1 Background

I describe the current Ceph cluster that I have and how the various pieces are interconnected.

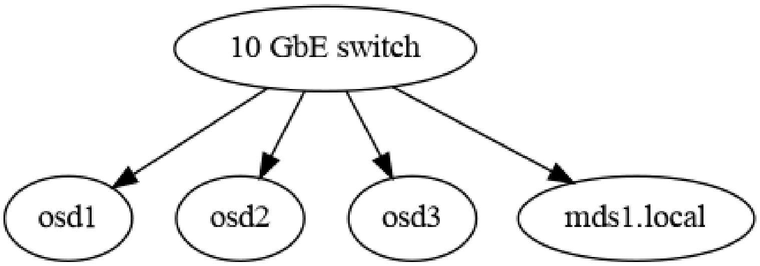
This is a really great way to setup reliable data storage.

I also talk about an additional file share I set up, using Samba (SMB) and ZFS as the file system.

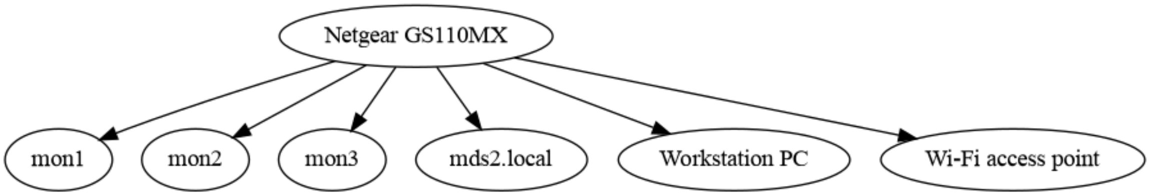
2 Systems

- Ceph storage backends. Intel NUC as osd1, osd2, osd3.
- Raspberry PI computers are the three monitor servers, mon1, mon2, mon3.
- R86S small form computer, as the mds1.local server.
- Seedless small form computer, as the mds2.local server (standby).

3 Network switches



The osd's, and mds1.local are connected to a 10 GbE switch. Those run at 2.5 GbE for the NUC's and 10 GbE for the mds server.



Finally I have another switch, a **Netgear GS110MX** which interconnects with the above. It has the Raspberry PI computers on it. It also connects to the mds2.local at 2.5 GbE.

4 How does this work?

The workstation is configured to talk to the monitor servers (mon1, mon2, mon3) to find out more about the Ceph system. For the Ceph FS or file system, it talks to the mds server. It is a metadata server which is tasked with serving the file system.

Once communication is established, the workstation will actually communicate directly with the storage backends (osd1, osd2, osd3). The underlying file objects will be read or written directly to and from these backends.

5 Mount point

Ubuntu 24.04 is running on a Workstation PC, which can mount the **/mnt/ceph** partition. It does this through the active mds server.

This system can also be rebooted to Windows 11, however it can no longer mount a ceph partition. The ceph versions I am running are the **squid** release.

As an interim solution I set up additional ZFS storage connected to the mds2.local system, and I'm able to share a folder on that one with SMB protocol. Even though I have a large amount of storage attached, I just use this **shared** folder as a temporary location to copy files between systems.

Finally, there is a Wireless network as well and a System76 laptop running Pop OS. That one can mount the /mnt/ceph partition as well.

6 Health and repairing

It is important to monitor the health of the cluster. This can be done through the **ceph status** or **ceph health detail** commands. A response of HEALTH_OK means everything is fine.

There is actually a whole data repair process which can be used in case one of the nodes breaks or goes missing. Actually, Ceph can self-heal itself so if you rebuild a node it will just refill itself with data.

However if you don't operate the cluster as intended and some form of data corruption forms, then you may need to make use of the data repair process. Think of it as a glorified file system check.

7 What's next?

I plan to add a Kubernetes cluster. It will be described in a future blog post.

One of the things it will be able to do is to spin up a container that will mount the /mnt/ceph location, and serve it out to the network over SMB protocol. This way it can fully reach the system booted as Windows 11.