ADMIN
Network & Security

Discover CephFS

TECHNICAL REPORT

SPONSORED BY

SUSE

Image © vlastas, 123RF.com

Discover CephFS

TECHNICAL REPORT

The CephFS filesystem combines the power of object storage with the simplicity of an ordinary Linux filesystem.

Ceph is a powerful object storage technology that delivers reliable, fault-tolerant, and scalable storage. If your organization is due for a data upgrade, and you're looking for a solution that is less expensive and more efficient than the conventional block storage alternatives, you're probably already considering the benefits of Ceph. SUSE Enterprise Storage is a Ceph-based storage solution that brings enterprise-grade certification and support to the Ceph environment. The deployment and management tools available through SUSE Enterprise Storage let you create a Ceph cluster in as few as two hours.

Part of your Ceph solution is to configure an access protocol that will serve as an interface to the Ceph cluster. Ceph has several interface alternatives that allow it to serve a number of different roles on your network. Your Ceph system can appear to the network as a block storage device, REST gateway, or iSCSI network device. One popular option for deploying Ceph is to mount it as a filesystem. CephFS is a network-ready filesystem that lets you seamlessly integrate the Ceph cluster with legacy applications and other components that require a conventional filesystem interface.

CephFS is a fully functioning, POSIX-compliant filesystem. You can mount the CephFS filesystem as you would any other Linux filesystem, storing data in the familiar arrangement of files and directories. Behind the scenes, CephFS is a fully functioning interface to the Ceph cluster (Figure 1).

CephFS up Close

Your CephFS configuration starts with a working Ceph cluster. See the "SUSE Enterprise Storage 5 Deployment Guide" [1] for more on setting up a Ceph cluster.

CephFS requires at least one metadata server (MDS) daemon. The MDS has the role of mapping the conventional file and directory

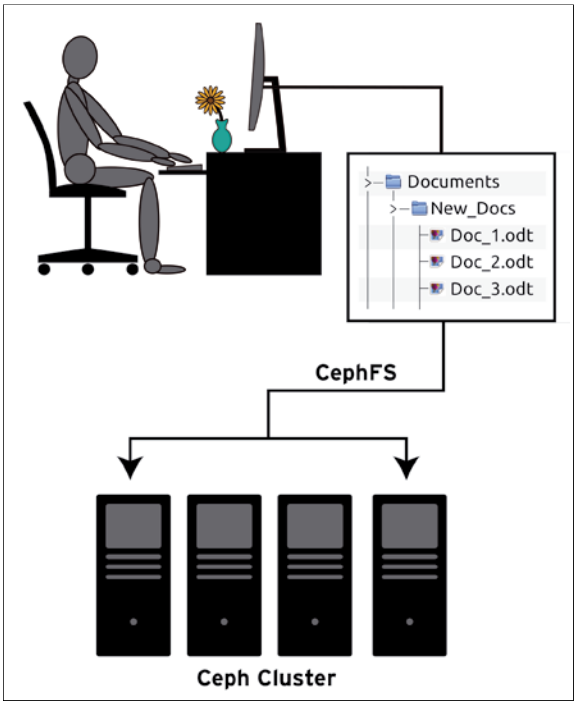

Figure 1: CephFS appears to the client as an ordinary filesystem, with files and folders arranged in the familiar hierarchical structure. Behind the scenes, the CephFS filesystem acts as an interface to the Ceph object store.

structure used with a Linux filesystem to the underlying object storage environment of the Ceph cluster. In other words, an MDS lets CephFS clients efficiently execute common commands such as `ls` or `find` without having to extract the data directly from the cluster. Previous versions of SUSE Enterprise Storage only allowed for one active MDS; however, SUSE Enterprise Storage 5 lets you define multiple active MDSs to optimize workload and improve performance. Deploying multiple MDS servers also lets you avoid a single point of failure and thus improves availability. You must define a metadata pool, which the MDS will use for storing and retrieving metadata. You will also need to define one or more data pools that will hold the filesystem data (Figure 2).

Out on the network, a kernel-based filesystem client application interfaces with the cluster. Multiple clients can mount CephFS and access the cluster simultaneously. The CephFS client runs on Linux. If you need direct access to the Ceph cluster from a Windows computer, you are currently better off using the iSCSI gateway or another interface instead of CephFS. As you will learn later in this paper, you also have the option of configuring a gateway to let clients on SMB and NFS networks access CephFS data.

The CephFS metadata pool and data pools can coexist with other interface technologies on the Ceph cluster. In other words, you can devote one part of the cluster to CephFS and still use other parts of the cluster for data stored through the object storage interface, iSCSI, the REST gateway, or other interfaces. Note that you can't access the same data from multiple interfaces – the data pools designated for CephFS can only be used with CephFS.

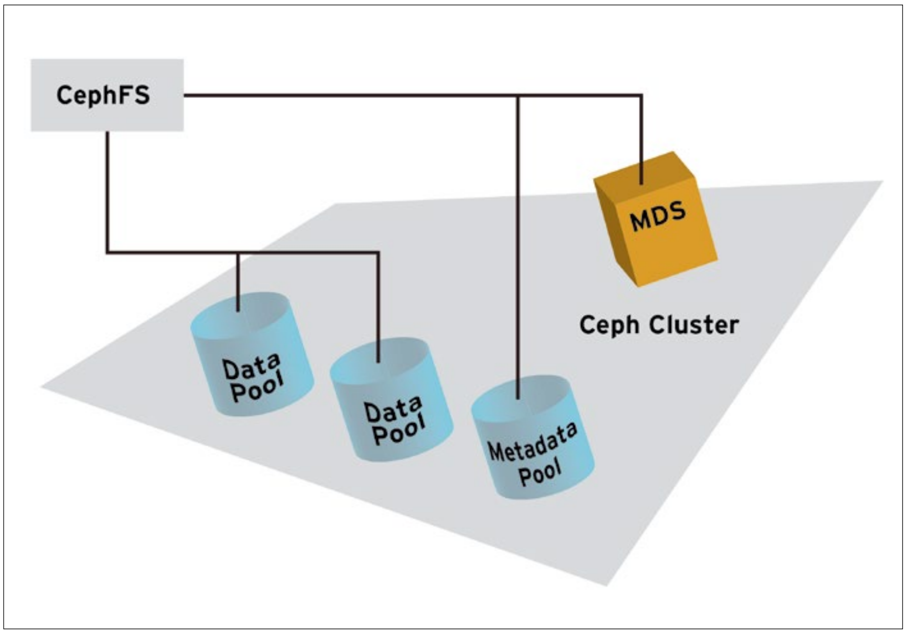

Figure 2: CephFS requires a metadata server (MDS) daemon, a metadata pool, and one or more data pools.

CephFS's support for multiple data pools leads to some powerful options for optimizing the storage environment. For instance, you can map a directory for archive data to a data pool that uses inexpensive, spinning storage disks and map another directory for low latency data to a pool that uses faster SSD storage.

Setting up CephFS

CephFS requires a working Ceph cluster [2]. You can set up the MDS when you are deploying your cluster (see the "SUSE Enterprise Storage 5 Deployment Guide" [1]) or add it later (see the "SUSE Enterprise Storage 5 Administration Guide" [2]).

Once you have a working cluster with an operational MDS, you just have to:

1. Create a metadata pool and one or more data pools on the Ceph cluster.

2. Enable the filesystem.

3. Mount the filesystem on the client using the CephFS kernel client.

The pools you use with CephFS will not be accessible through other interface methods, so you'll need a plan for which systems within the cluster you will dedicate to CephFS. Create the pools with the following commands:

```
ceph osd pool create cephfs_data <pg_num>
ceph osd pool create cephfs_metadata <pg_num>
```

These commands create a data pool called `cephfs_data` and a metadata pool called `cephfs_metadata`. The `<pg_num>` variable is a number related to the number of placement groups associated with the pool – see the Ceph documentation for more on the `<pg_num>` value. As mentioned previously, you can use multiple data pools with CephFS, so create additional data pools as needed.

After you have created the pools, you can set up the CephFS filesystem with a single command

```
ceph fs new filesystem_name <metadata> <data>
```

where `filesystem_name` is the name you are giving to the filesystem, `<metadata>` is the name of the metadata pool, and `<data>` is the name of the data pool.

To create a CephFS filesystem called `My_CephFS` using the pools created above, enter:

```
ceph fs new My_CephFS cephfs_metadata cephfs_data
```

If the CephFS filesystem starts successfully, the MDS should enter an active state, which you can check using the `ceph mds stat` command:

```
$ ceph mds stat
e5: 1/1/1 up {0=a=up:active}
```

Once the CephFS system is started and the filesystem is mounted, you can treat it as you would any other Linux filesystem: Create, delete, and modify directories and files within the directory tree. Behind the scenes, CephFS stores the data safely within the Ceph cluster.

CephFS must interface with the security system of the underlying Ceph cluster, so you'll need to supply a username and secret key in order to mount the CephFS filesystem.

Access the key for the user in a keyring file:

```
cat /etc/ceph/ceph.clinet.admin.keyring
```

The key will look similar to the following:

```
[client.admin]
key = AQCj2YpRiAe6CxAA7/ETt7Hc19IyxyYciVs47w==
```

Copy the key to a secret file that includes the name of the user as part of the filename (e.g., `/etc/ceph/admin.secret`), and then paste the key into the contents of the file. You will reference this file as part of the `mount` command.

Create a mount point on the Linux system you wish to use as the client:

```
/mnt/cephfs
```

To mount the CephFS filesystem, enter

```
sudo mount -t ceph ceph_monitor1:6789:/ ➘
/mnt/cephfs -o name=admin, ➘
secretfile=/etc/ceph/admin.secret
```

where `/mnt/cephfs` is the mount point, `ceph_monitor1` is a monitor host for the Ceph cluster, `admin` is the user, and `/etc/ceph/admin.secret` is the secret key file.

The monitor host is a system that holds a map of the underlying cluster. The CephFS client will obtain the CRUSH map from the monitor host and thus obtain the information necessary to interface with the cluster. The Ceph monitor host listens on port 6789 by default.

If your Ceph cluster has more than one monitor host, you can specify multiple monitors in the `mount` command. Use a comma-separated list:

```
sudo mount -t ceph ceph_monitor1,ceph_monitor2,➘
ceph_monitor3:6789/ /mnt/cephfs ...
```

Specifying multiple monitors provides failover in case one monitor system is down.

Linux views CephFS as a regular filesystem, so you can use all the standard mounting techniques used with other Linux filesystems. For instance, you can add your CephFS filesystem to the /etc/fstab file to mount the filesystem at system startup.

Gateways

CephFS appears as a standard POSIX filesystem. You can mount it on a Linux computer using all the usual Linux mount commands. But what if you want to make the CephFS filesystem accessible to a wider network of mixed clients (Linux, Windows, and Mac systems)? SUSE Enterprise Storage 5 offers two optional gateways for interfacing CephFS with the network:

- NFS Ganesha – an interface for the Network File System (NFS).
- Samba Gateway – an interface for the SMB/CIFS file service protocols.

The NFS Ganesha gateway is fully supported in SUSE Enterprise Storage 5. The Samba gateway is included as a technical preview. Most modern operating systems provide some form of client support for either NFS or SMB/CIFS.

Use the NFS or Samba gateway to let remote users mount CephFS as a local filesystem (Figure 3). See the "SUSE Enterprise Storage 5 Deployment Guide" [1] and the "SUSE Enterprise Storage 5 Administration Guide" [2] for more on configuring a gateway for network access to CephFS directories.

Configuration and Management Tools

CephFS comes with a collection of command-line tools for configuring and managing CephFS filesystems. The ceph fs command is a general-purpose configuration tool with several options for managing file layout and location. A collection of utilities available with SUSE Enterprise Storage 5 provide functionality equivalent to the Unix/Linux fsck filesystem consistency check utility. The tools cephfs-data-scan, cephfs-journal-tool, and cephfs-table-tool help an expert user discover and repair damage to the filesystem journal and metadata. These tools are meant for disaster recovery and should not be used on an active filesystem.

Extended Attributes

CephFS's extended attributes feature adds additional metadata attributes to the CephFS filesystem. Extended attributes let you track statistics on filesystem usage. You can also use the extended attributes feature to change the

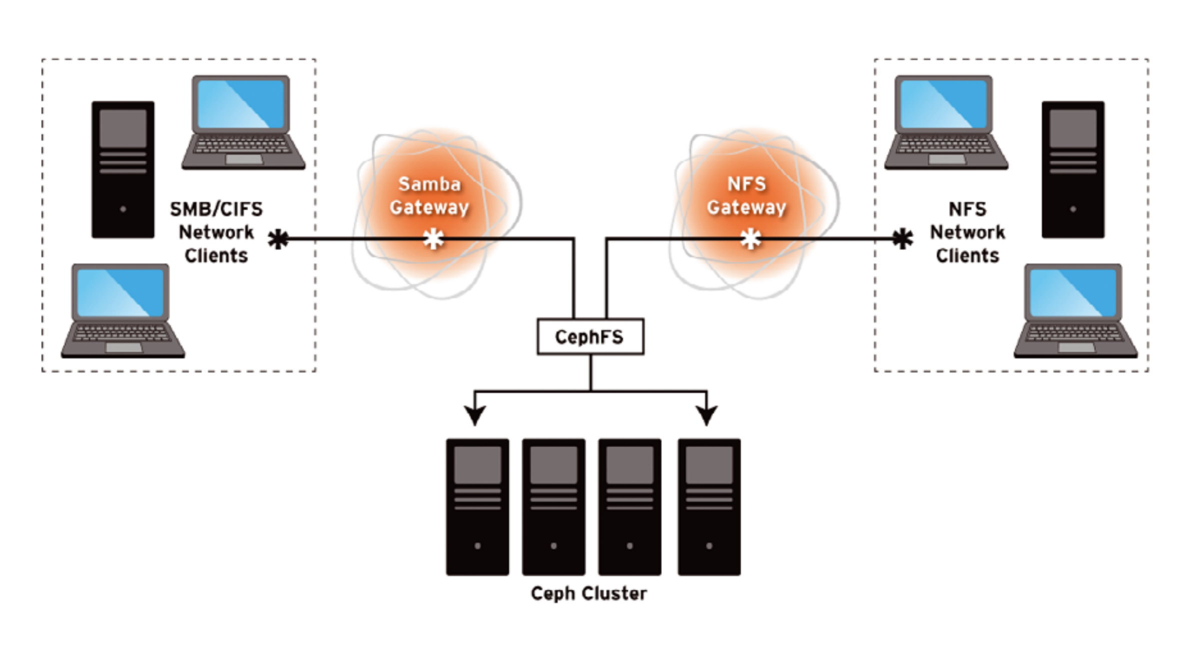

FIGURE 3: **CephFS provides optional gateway interfaces for network access from NFS or SMB/CIFS client systems.**

behavior of Ceph's CRUSH algorithm, customizing the way the filesystem maps to the object store. As described previously, you could configure CephFS to store one directory in a pool with high-performing SSD drives and another directory in a pool with slower, archive media. Extended attributes also let you direct certain directories to a pool dedicated to a specific department or branch or to a pool with specific replication characteristics.

Erasure Coding

Ceph supports an optional fault tolerance technique known as *erasure coding*. Earlier versions of Ceph did not extend erasure coding support to the CephFS filesystem; however, the version of CephFS included with SUSE Enterprise Storage 5 does offer erasure coding support.

The choice of whether to base your fault tolerance strategy on erasure coding or a more conventional form of replication may depend on your needs and the details of your network configuration. In general, erasure coding tends to be slower, but it makes more efficient use of disk resources.

Practical Notes

The following CephFS features are not currently supported in SUSE Enterprise Storage:

- CephFSsnapshots
- The CephFS FUSE module

See the box entitled "Ceph and SUSE Enterprise Storage" for more on feature support in SUSE Enterprise Storage and the Ceph project.

Conclusion

CephFS is a Linux filesystem that serves as an interface to a Ceph cluster. The CephFS filesystem lets you store data in the familiar file and directory structure known to Linux users and applications. If you have legacy applications that expect to interact with a conventional filesystem, or if you want to minimize the disruption of your current network configuration, you can use CephFS to present the Ceph cluster to your network as a conventional filesystem.

CephFS comes with a set of repair tools for finding and fixing data errors, as well as gateway interfaces that support the NFS and SMB/CIFS network file sharing protocols. For more information on working with CephFS, see the Ceph project documentation [3]. ■

Ceph and SUSE Enterprise Storage

Ceph is an open source project served by a large, global community of developers, testers, and users. SUSE Enterprise Storage is an enterprise-ready storage system based on Ceph. The SUSE team provides hardware certification and customer support, shaping the Ceph framework to improve stability and reliability for enterprise environments. The professional testing and tuning offered through SUSE Enterprise Storage means some features that are still considered experimental by the greater Ceph community are fully implemented and ready for use in SUSE Enterprise Storage. Other tools available through the Ceph open source project are not considered ready and reliable enough for the enterprise and are, therefore, not supported in SUSE Enterprise Storage.

Info

[1] SUSE Enterprise Storage 5 Deployment Guide: [https://www.suse.com/documentation/suse-enterprise-storage-5/singlehtml/book_storage_deployment/book_storage_deployment.html]

[2] SUSE Enterprise Storage 5 Administration Guide: [https://www.suse.com/documentation/suse-enterprise-storage-5/singlehtml/book_storage_admin/book_storage_admin.html]

[3] Ceph Documentation: [http://docs.ceph.com/docs/master/]