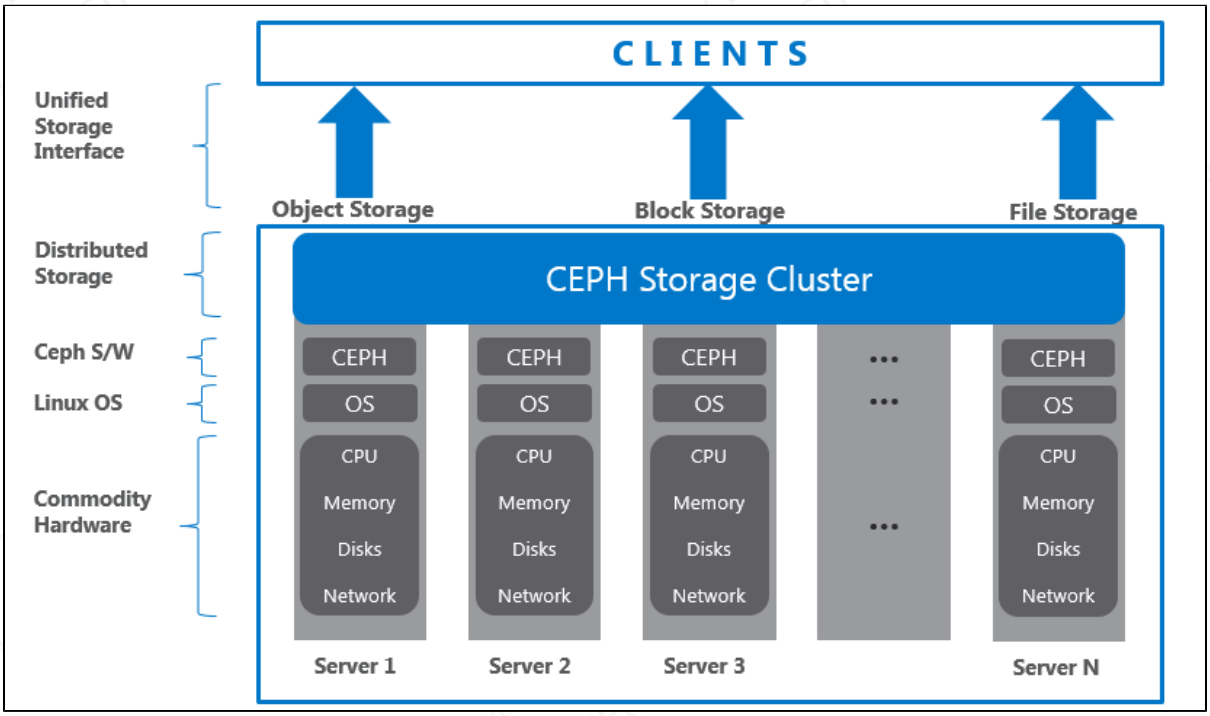
Shopee Ceph/S3 Overview

Introduction

This page provides an overview and introduction to our Ceph and S3 clusters

What is Ceph?

Ceph is an open-source software for Distributed File System/Software Define Storage



Ceph can run on common hardware (our H4_v2, I3_v2 and S1_v2 servers)

Main advantages of Ceph

- Free & Open source
- Production-ready: being used at CERN, RedHat, Huawei at very large scale (more than thousand nodes)
- Good performance
- No single point of failure
- Can handle PBs of data at ease

Disadvantages

- Hard to learn and maintain
- Complex architecture

Main use cases of Ceph

- **CephFS**: Like mounting filesystem to servers (similar to NFS)
- **S3**: Storing files (objects) like Dropbox and GG Drive
- **RBD**: for managing virtual volumes (K8s PVs, Openstack Volumes,...)

Ceph at Shopee

At Shopee, we mainly use Ceph as backend for S3 and CephFS

Some numbers

- Live Clusters: 14

- Data usage
  - Used Capacity: **2482TB**
  - Total Capacity: **7616TB**
- Number of files
  - Around nearly **1 billion objects** across all clusters

What is S3?

As mentioned above, S3 is a ***standard*** *originally from Amazon* for storing distributed files via HTTP

Can think of S3 like **Dropbox** or **GG Drive**, it is a place to store unstructured files, objects

Bucket is a place for storing files

One user can have one or more buckets to store file

S3 is widely used in Shopee as a place to store files: financial reports, user receipt images, service configurations, container images....

Teams which are using S3

- STO: Grafana, Mattermost, ALB, DNS, Cachecloud, Harbor, CertMS, SMAP, Sentry, Kafkacloud, TOC....
- DEEP: Paidads, Search, Recommendation, Platform...
- Business: Salesforce, Growth, Marketing,...
- DE/DI: Datanexus, Idata,....
- WSA
- Coreserver
- Backend
- ...

FAQs

Q1: What is the difference between CephFS, CephS3 and CephRBD?

A1: CephFS is a filesystem, rbd is a block device. CephFS is a lot like NFS;

- It's a filesystem shared over the network where different machines can access it all at the same time. RBD is more like a hard disk image, shared over the network.
- It's easy to put a normal filesystem (like ext2) on top of it and mount it on a computer, but if you mount the same RBD device on multiple computers at once then Really Bad Things are going to happen to the filesystem.
- In general, if you want to share a bunch of files between multiple machines, then CephFS is your best bet. If you want to store a disk image, perhaps for use with virtual machines, then you want RBD. If you want storage that is mostly compatible with Amazon's S3, then use radosgw.