Ceph: a scalable, high-performance distributed file system, OSDI 2006

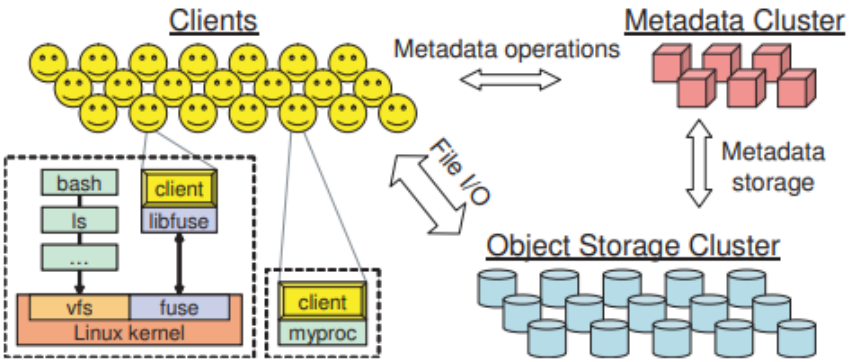
Ceph consists of a cluster of metadata servers that manage clients and metadata, and an cluster of object storage servers that store and manage data at the object level.

Ceph clients can access file systems mounted with FUSE and provide a file system interface that can be linked directly from an application.

Metadata server clusters manage security and system consistency as well as the file system namespace. Object storage server clusters store both file metadata and data.

In Ceph, to enhance the separation of file metadata and data, the object name is defined using the inode number of the file, and the object is distributed and stored in the object storage server through the method called CRUSH. This makes it easy to compute the name and location of an object using CRUSH anywhere in the Ceph component.

Ceph adopts dynamic subtree partitioning as metadata server clustering method in consideration of scalability of metadata server. This allows efficiently distribute work among metadata server clusters by dynamically distributing workload across multiple metadata servers or redistributing metadata based on workload patterns.



Pros

Ceph eliminates the centralized gateway and enable clients to interact with Ceph OSD Daemons directly. There is no single node failures.

I think most of performance of ceph is driven from effective metadata management. as they dynamically map subtrees of directory to metadata sever based on current workloads

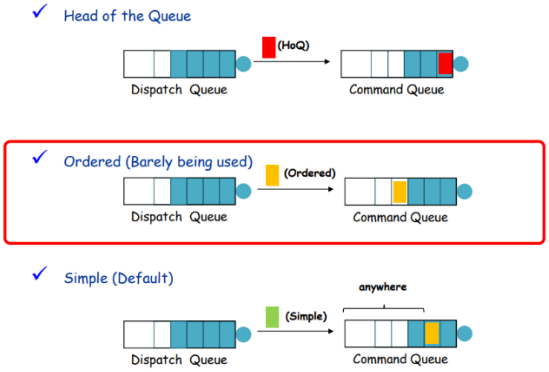
Cons & ideas

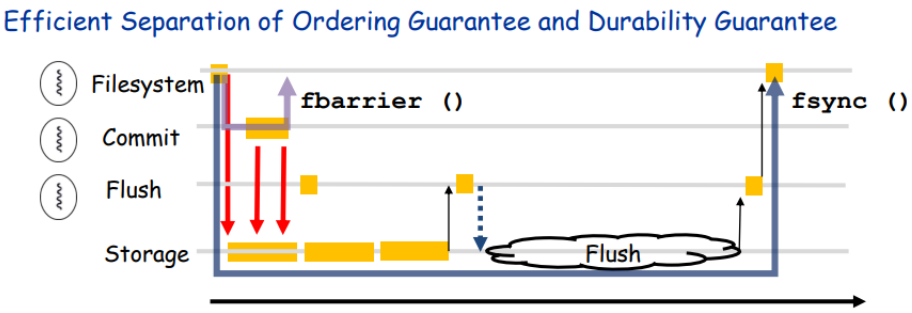
Too complicated to build wholes system. As they use EBOFS for optimized object storing performance, it adds client-side efforts for using system.

Barrier-enabled IO stack for flash storage, FAST 2018

This paper build barrier-enabled I/O stack which keep order of each designated write operation (barrier write) Using " barrier-enabled storage / Order-preserving Block Device Layer / Epoch Based IO scheduler" they suggest, the I/O order between barrier is kept for whole I/O stack. And by eliminate DMA transfer overhead

Using a barrier command that is supported by SCSI but rarely used, it creates a berry between the previous and subsequent write without flushing motion. It does not guarantee completion of io after actual return, but guarantees order





Pros

This system is designed to keep consistency of I/O system by keep a barrier and shows good performance enhance by utilizing existing commands without changing the overall semantic.

Cons & ideas

Currently, I/O stack and devices have expanded to a multi-queue structure. However, in this study, I/O is sent down using a single queue for the use of the barrier. While maintaining that barrier concept, I think this research would be better to exploit advantage multi-queue system. like usefulness of implementing streamed I/O.