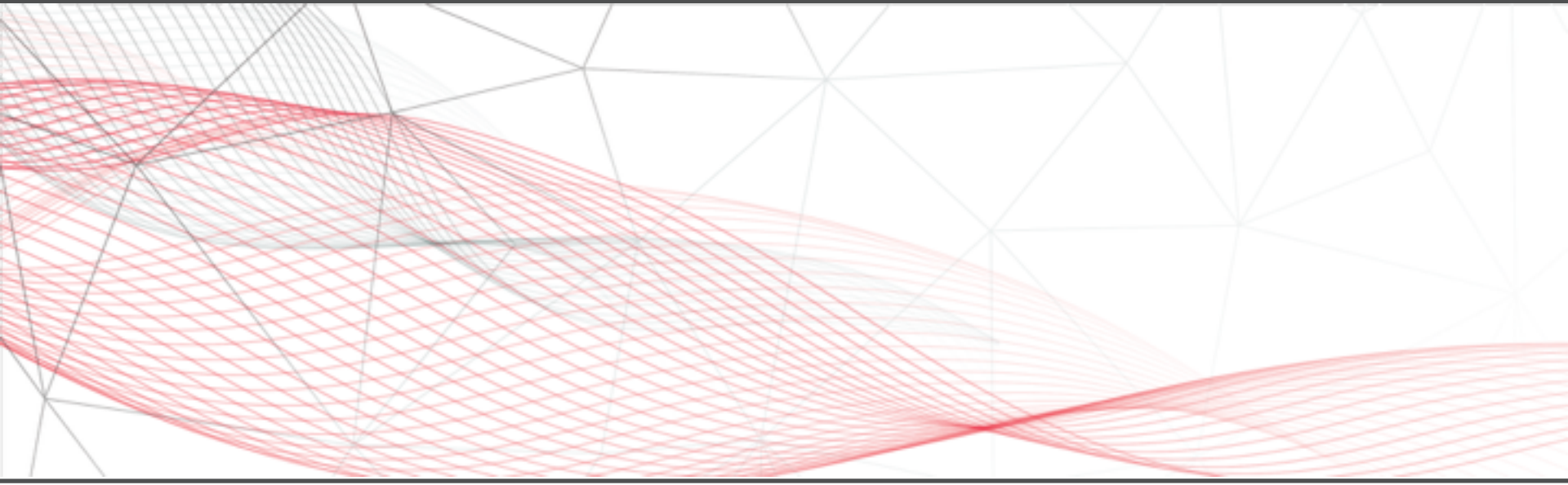


CEPH OSD BACKENDS & FILESYSTEMS

Bryan Apperson
Concurrent Computer Corporation
Storage Integration Engineer
4/12/2016

cephatl

Comparison of Features & Performance

- Is it performant? Comparisons for various Filestore filesystems.
- What are the benefits and features?
- What are the drawbacks?
- Is it battle-tested (production ready)?

cephatl

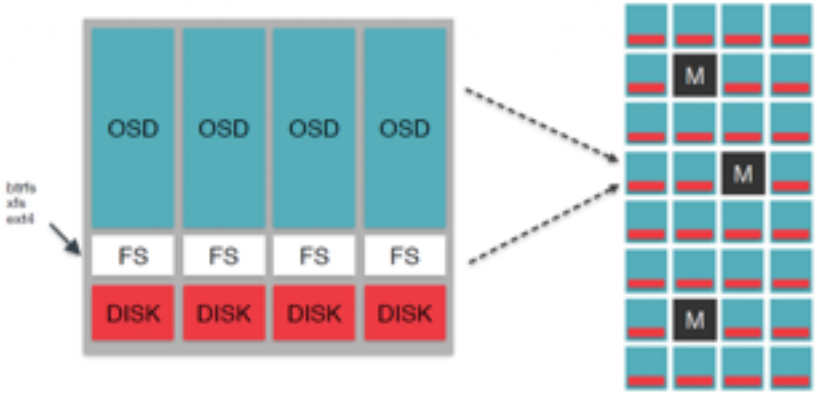
OSD Backends

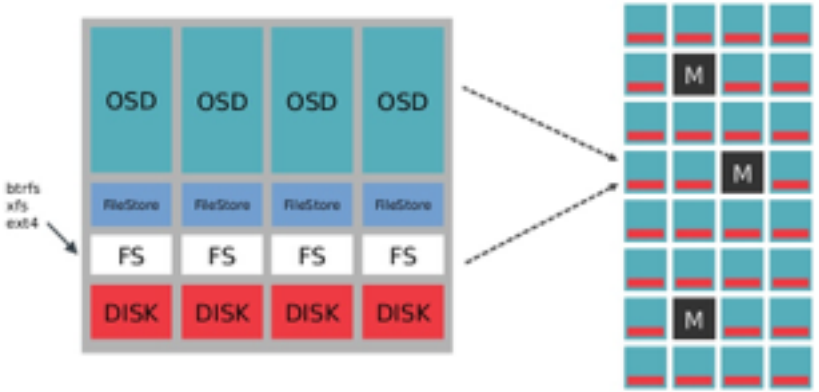
- Filestore: Presently the de-facto production OSD backend. Composed of one data partition and one journal partition. The data partition is layered above a POSIX filesystem (XFS, BTRFS, EXT4).

- Bluestore: Experimental backend as of Ceph Jewel. Eliminates double write penalty (from Filestore journal). Data is written direct to block.

cephatl

Filestore Architecture





cephatl

More about Filestore

- With the Filestore backend, Ceph writes objects as files on top of a POSIX filesystem such as XFS, BTRFS or EXT4. With the Filestore backend a OSD is composed of an un-formatted journal partition and an OSD data partition.

- One of the largest drawbacks with the OSD Filestore backend is the fact that all data is written twice, through the journal and then to the backing data partition.

cephatl

Filestore Filesystems

- XFS: Used at present in many production Ceph deployments. XFS was developed for Silicon Graphics, and is a mature and stable filesystem. The Filestore/XFS combination is well tested, stable and great for use today.
- BTRFS: A copy-on-write filesystem. It supports file creation timestamps and checksums that verify metadata integrity. BTRFS very interestingly supports transparent LZO and GZIP compression among and other features.
- EXT4: A solid, battle tested filesystem. However, with a maximum size of 16TB, it is not exactly future proof (considering that Samsung has already released a 16TB drive)

OSD Filestore Performance Test

All performance tests in this article were performed using Ceph Hammer. The specifications for the test machine are as follows:

- CPU: Intel(R) Core(TM) i7-6700K CPU @ 4.00GHz
- RAM: 32GB DDR4 2133 MHz
- OSDs: 5 x Seagate ST2000DM001 2TB
- OS: Fedora 23 (Kernel 4.3.3)
- OS Drive: 2 x Samsung 850 Pro SSD (BTRFS RAID1)

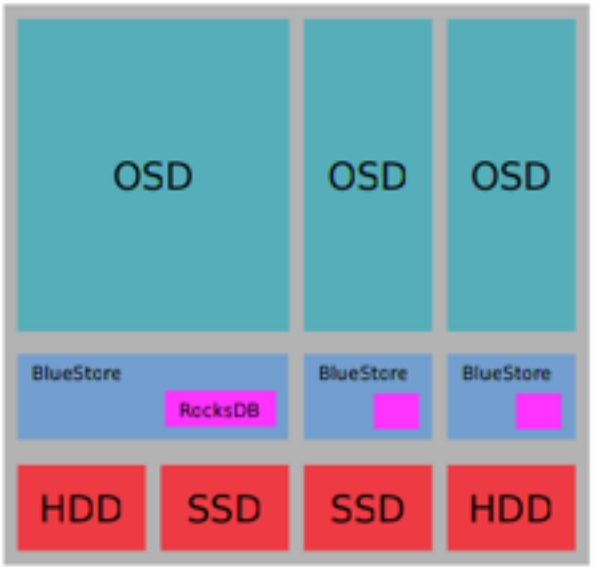
cephatl

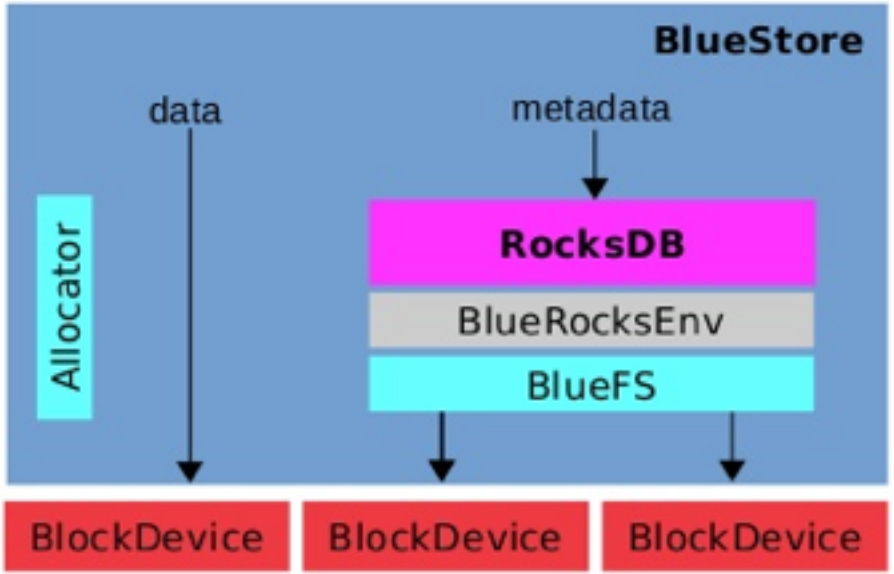
OSD Filestore Filesystem Performance

- XFS:
  - 4 MB
    - Write:
      - IOPS: 19.17
      - BW: 76.516 MB/s
      - Latency: 0.835348s
    - Read:
      - IOPS: 118.38
      - BW: 473.466MB/s
      - Latency: 0.134731s
  - 4 KB
    - Write:
      - IOPS: 203.124
      - BW: 0.790MB/s
      - Latency: 0.0789896s
    - Read:
      - IOPS: 209.33
      - BW: 0.812MB/s
      - Latency: 0.076963s

cephatl

Bluestone Architecture





cephatl

More About Bluestone

- Bluestore is set to release for experimental use in Jewel. The benefits of Bluestore are a direct to block OSD, without filesystem overhead or the need for a "double-write" penalty (associated with the Filestore journal). Bluestore utilizes RocksDB, which stores object metadata, a write ahead log, Ceph omap data and allocator metadata. Bluestore can have 2-3 partitions per, one for RocksDB, one for RocksDB WAL and one for OSD data (un-formatted - direct to block).

cephatl

Conclusion

- While the Filestore/XFS deployment scenario may be the stable way to go for production Ceph clusters
- Filestore/BTRFS is certainly the most feature rich.
- With the development of Bluestore this may change in the near future.

cephatl

Thanks for attending the April 12, 2016 Ceph ATL Meetup!

The accompanying blog article for this presentation can be found at:

http://bryanapperson.com/blog/ceph-osd-performance/

cephatl