Curve

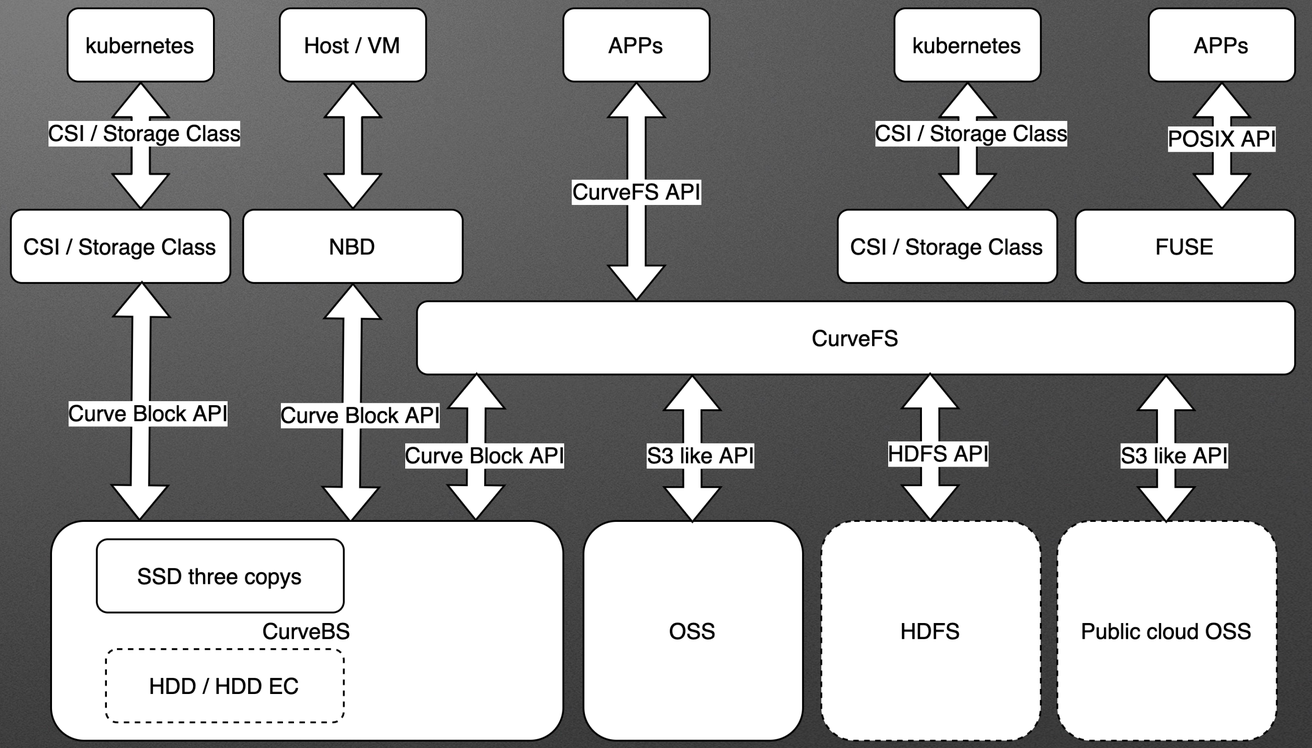
High performance Cloud native Distributed storage system
https://www.opencurve.io/

Agenda

- What is Curve
- Use Cases
- CurveBS
  - Key Features
  - Comparing to Ceph
- CurveFS
  - Comparing to Ceph
- Current Status
- Roadmap

What is Curve

- Curve is an distributed storage system
- Components
  - Curve Block Storage (CurveBS)
    - CurveBS: a high performance cloud native distributed block storage
  - Curve File System (CurveFS)
    - CurveFS: a high performance cloud native file system

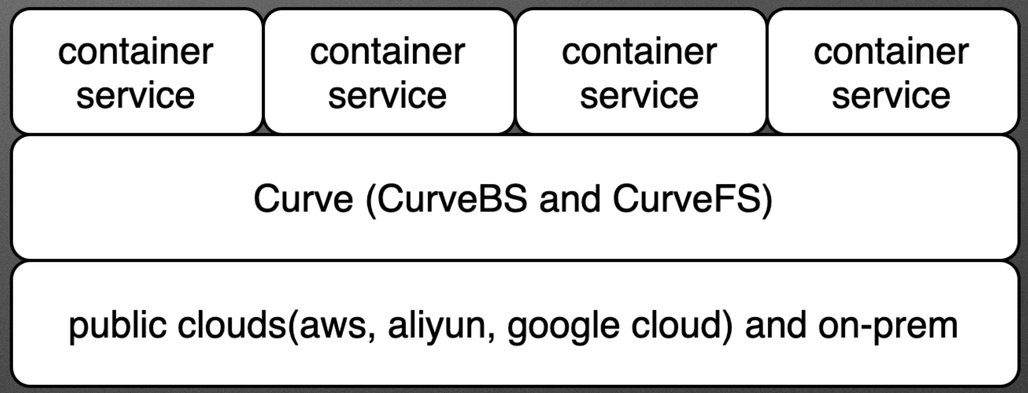


Use Cases

- Container
- Database
- Data apps(middleware/bigdata/ai)
- Data backup

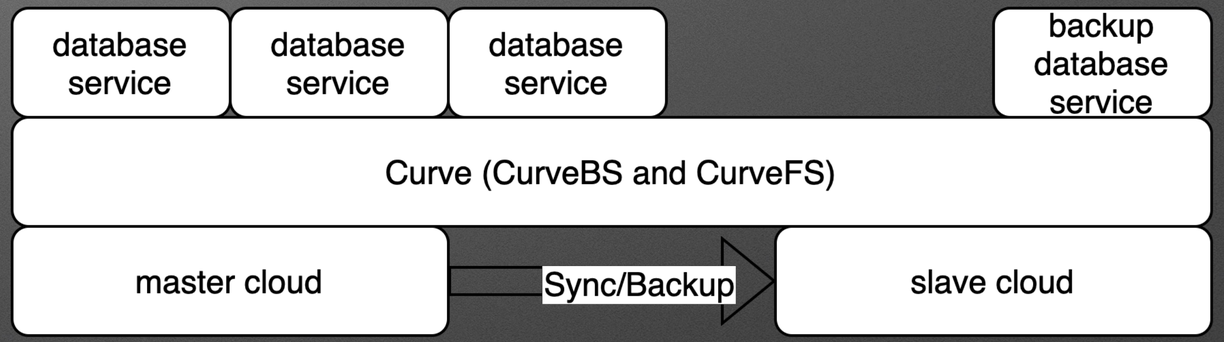
Container

- Aggregates underlying storage in the cloud (AWS EBS, AWS S3, AWS Glacier, aliyun EBS, aliyun OSS) or on-prem (baremetal, HDFS, OSS) and turns it into container-native storage



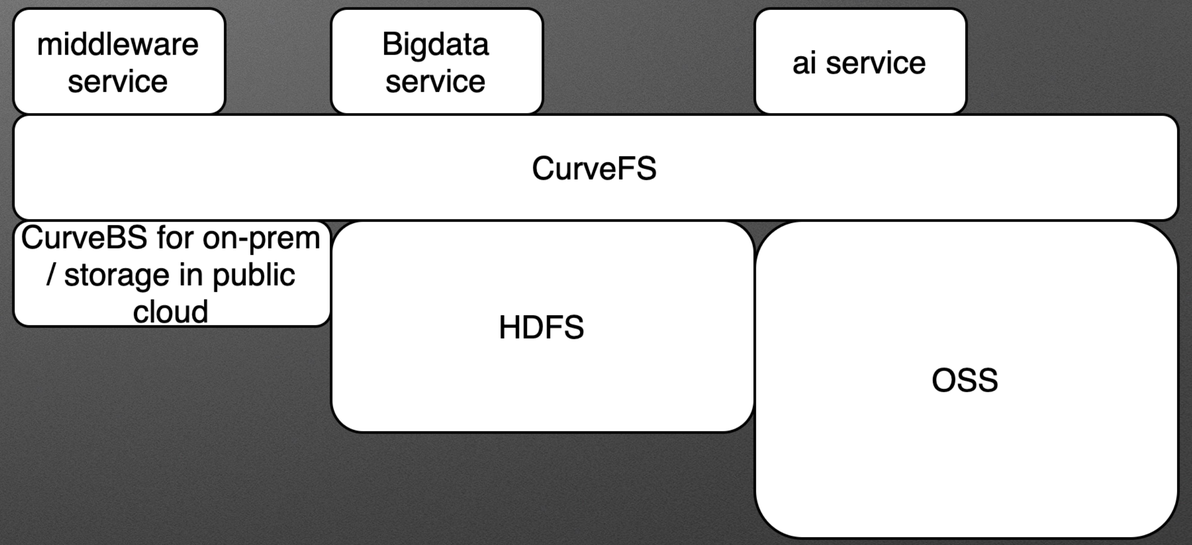
Database

- Database services orchestrated in the cloud
- Curve can backup / sync data to slave cloud
- When master cloud failure happens, Database service can move to the slave cloud



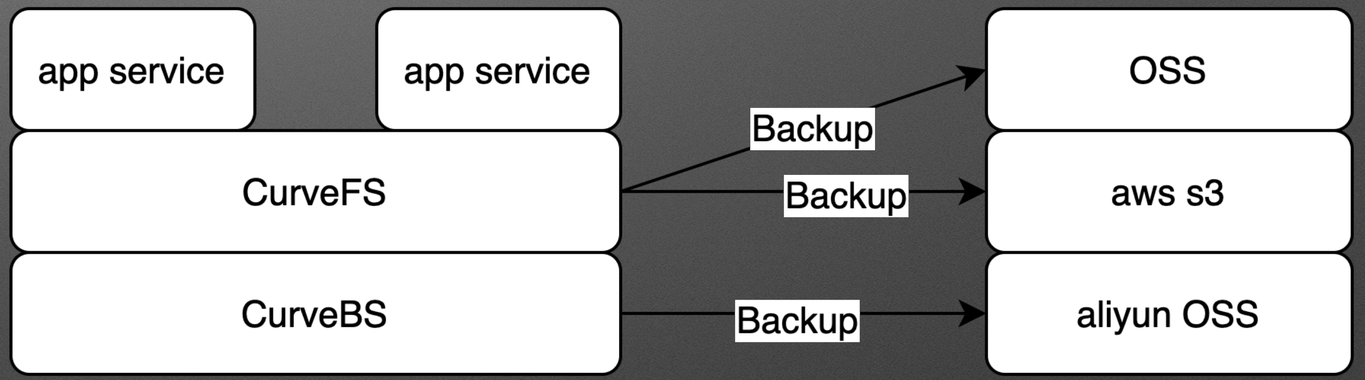
Data apps(middleware/bigdata/ai)

- CurveFS can manage different storages (HDFS, OSS, EBS) below
- Apps access data by POSIX interface
- Infrequent data is moved to OSS, and frequent data is moved to high speed storage transparently



Data backup

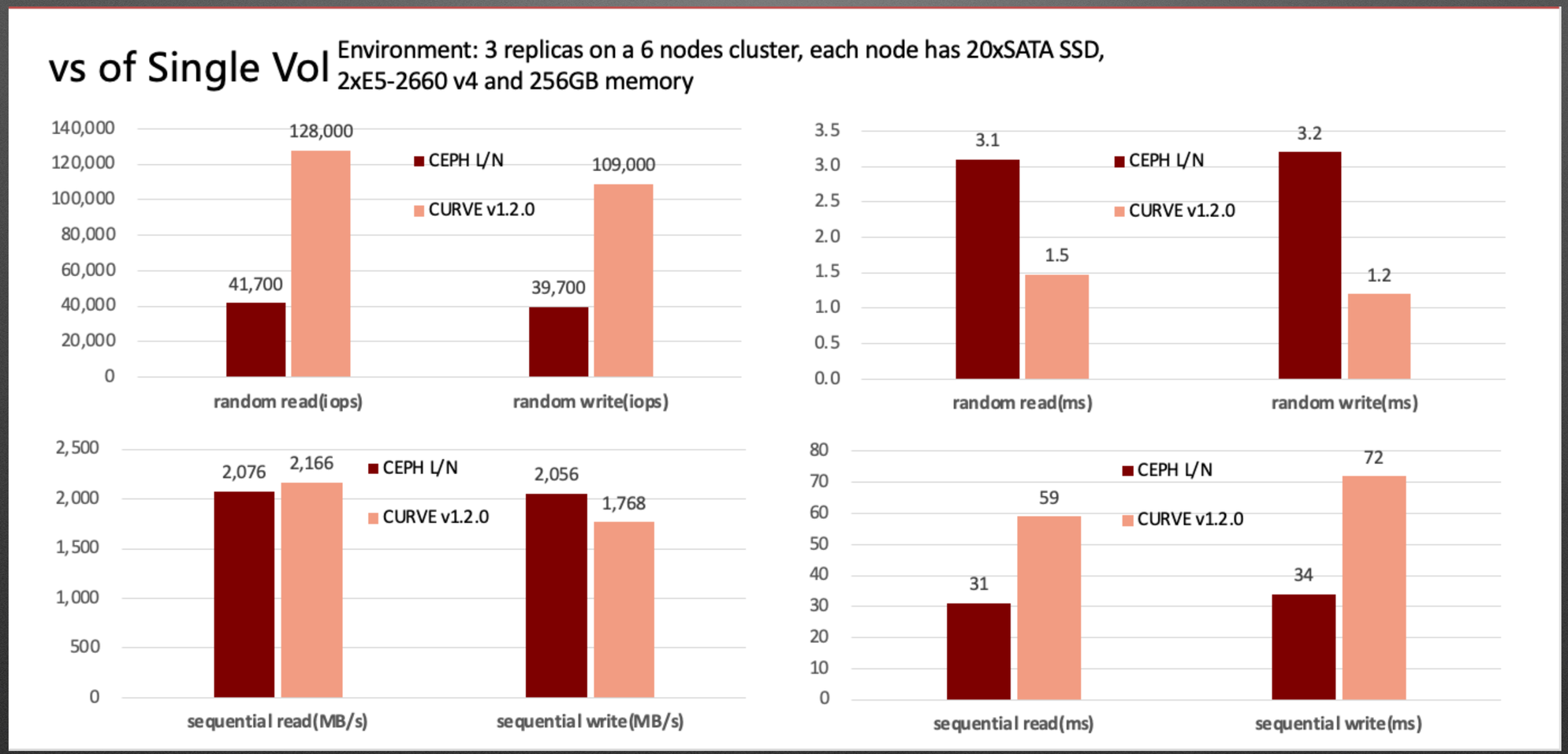
- Curve (CurveBS, CurveFS) can backup data to remote public cloud and to on-prem OSS



CurveBS

- high performance
- mainly used for (SSD, three replicas)
- csi / storage class for kubernete, nbd for HOST/VM

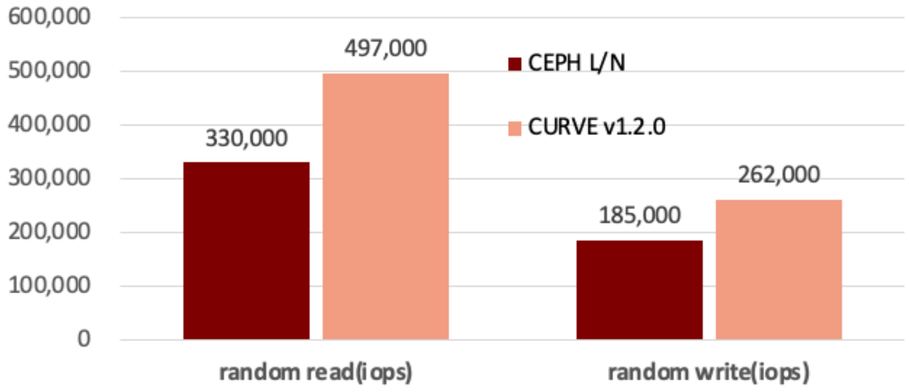
Performance (vs. Ceph RBD)

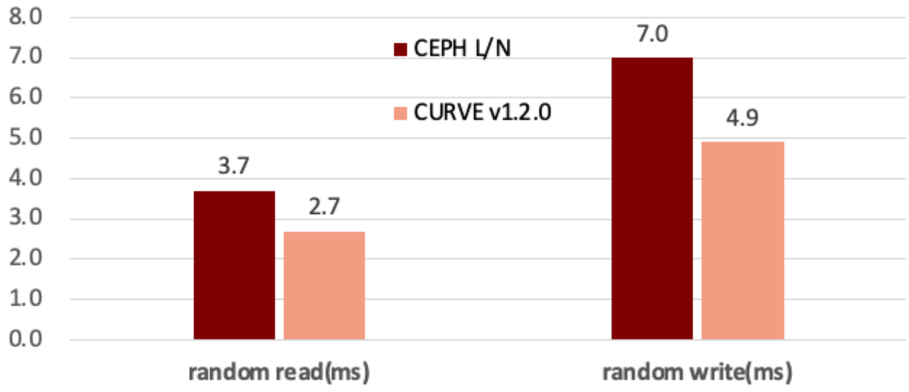


Performance (vs. Ceph RBD)

vs of Multi Vols

Environment: 3 replicas on a 6 nodes cluster, each node has 20xSATA SSD, 2xE5-2660 v4 and 256GB memory





Network bandwidth becomes a bottleneck in case of Sequential read and Sequential write

CurveBS Features

- RAFT for data consistency
  - minor impaction when chunk server fails
- Precreated chunk file for volume space mapping
- high performance framework
  - Use bthread (M bthread map N pthread) for scalability and performance on Multi-thread CPU
  - Lock free queue design
  - Memory zero copy design
- Cloud native support

Cloud native for CurveBS

- CSI plugin for CurveBS
- Deploy CurveBS as container service (in Plan)
- Config CurveBS by (Cluster and Pool CRDs) in Kubernetes (in Plan)
- Support Operator capability level 5 (in Plan)
  - horizontal / vertical scaling, auto config tuning, abnormal detection and schedule tuning

Storage Engine Comparison (vs. Ceph)

DATA CONSISTENT PROTOCOL	CURVE (RAFT)	CEPH
WRITE SUCCESS	majority write successful	all write successful
READ	Leader of copyset	Node in PG
SLOW STORAGE/DISK FAILURE INFLUENCE	without I/O disruption	I/O jitter occasionally
CAN SYNC WITH REMOTE DISK SERVER	Y	N

I/O Jitter (vs. Ceph)

3 replicas with 9 nodes cluster each node has 20 x SSD, 2xE5-2660 v4 and 256GB mem

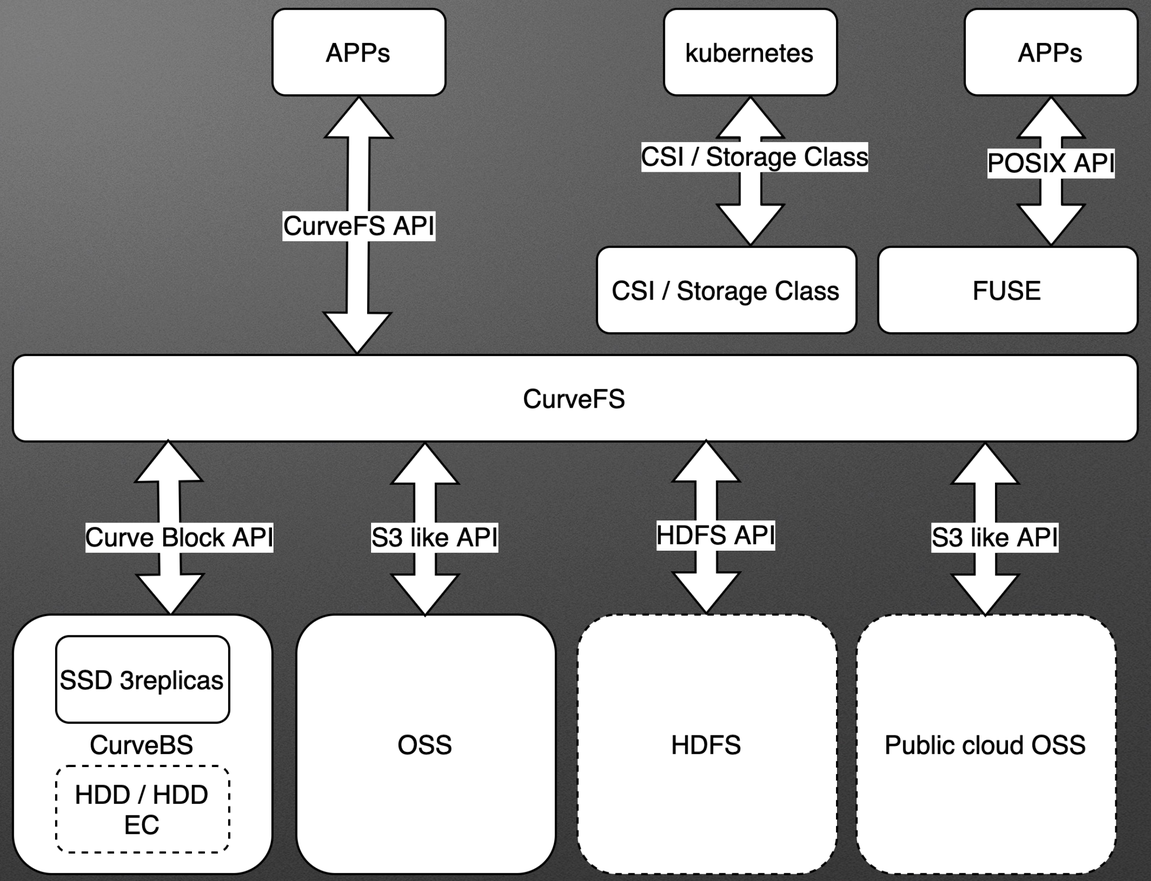
FAULTS CASE	CURVE I/O JITTER	CEPH I/O JITTER	COMMENT
ONE DISK FAILURE	4s	7s	
ONE SERVER FAILURE	4s	7s	
SERVER RESPONSE VERY SLOW	4s	unrecoverable	frequently delay of disk i/o are very long
NETWORK LATENCY 50MS	1s frequently	7s recently	

Storage Engine Comparison (vs. Ceph)

META MANAGEMENT	CURVE CHUNK SERVER	BLUESTORE
META	Precreate Chunk File Pool on ext4	RocksDB
META OVERHEAD	without ext4 meta overhead	increase read/write magnification
PERFORMANCE	High	Need to optimize rocksdb

CurveFS Features

- CurveFS can manage storages (open cloud storage and on-prem storage) and expose unified file space for app accessing
- RAFT for data consistency
- POSIX-compatiable
- Cloud native support



Cloud native plan for CurveFS

- CSI plugin for CurveFS (in Plan)
- Deploy CurveFS as container service (in Plan)
- Config CurveFS by (cluster and storage pools) CRDs in Kubernetes (in Plan)
- Support Operator capability level 5 (in Plan)
  - now support helm

Current Status

- Release 2 major version on CurveBS
  - v1.2 supporting QOS, Discard, data silent check
  - v1.3 some performance optimization
  - more details https://github.com/opencurve/curve/releases
- Now working on CurveFS

Roadmap

- CurveFS based on CurveBS
- POSIX-compatible and mountable
- Cache support on CurveFS
- CurveFS cloud native support
  - csi plugin for CurveFS
  - support operator capability level 2: automated application provisioning and configuration management and patch and minor version upgrads supported
- File meta data preallocate
- RAFT optimization
  - ParallelRaft for write
  - Reduce write magnification for file new write
- Cloud tiering support

Thanks