Curve

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

Agenda

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

What is Curve

- Curve is an distributed storage system
  - High performance
  - Automated operation
  - Cloud native
- 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

High Performance

- CurveBS
  - Chunk File Pool to reduce write magnification
  - Data striping across copysets
  - Lock free queue design
  - Memory zero copy design
- CurveFS
  - Speed up write/read by cache
  - File meta preallocate
  - Raft for consistency

Storage Engine Comparison (vs. Ceph)

- Architecture Design
  - Use bthread (M bthread map N pthread) for scalability and performance on Multi-thread CPU
  - Lock free queue design
  - Memory zero copy design to reserve CPU resources

Storage Engine Comparison (vs. Ceph)

DATA CONSISTENT PROTOCOL	CURVE (RAFT)	BLUESTORE
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
IMPROVEMENT MEASURES	ParallelRaft for write	

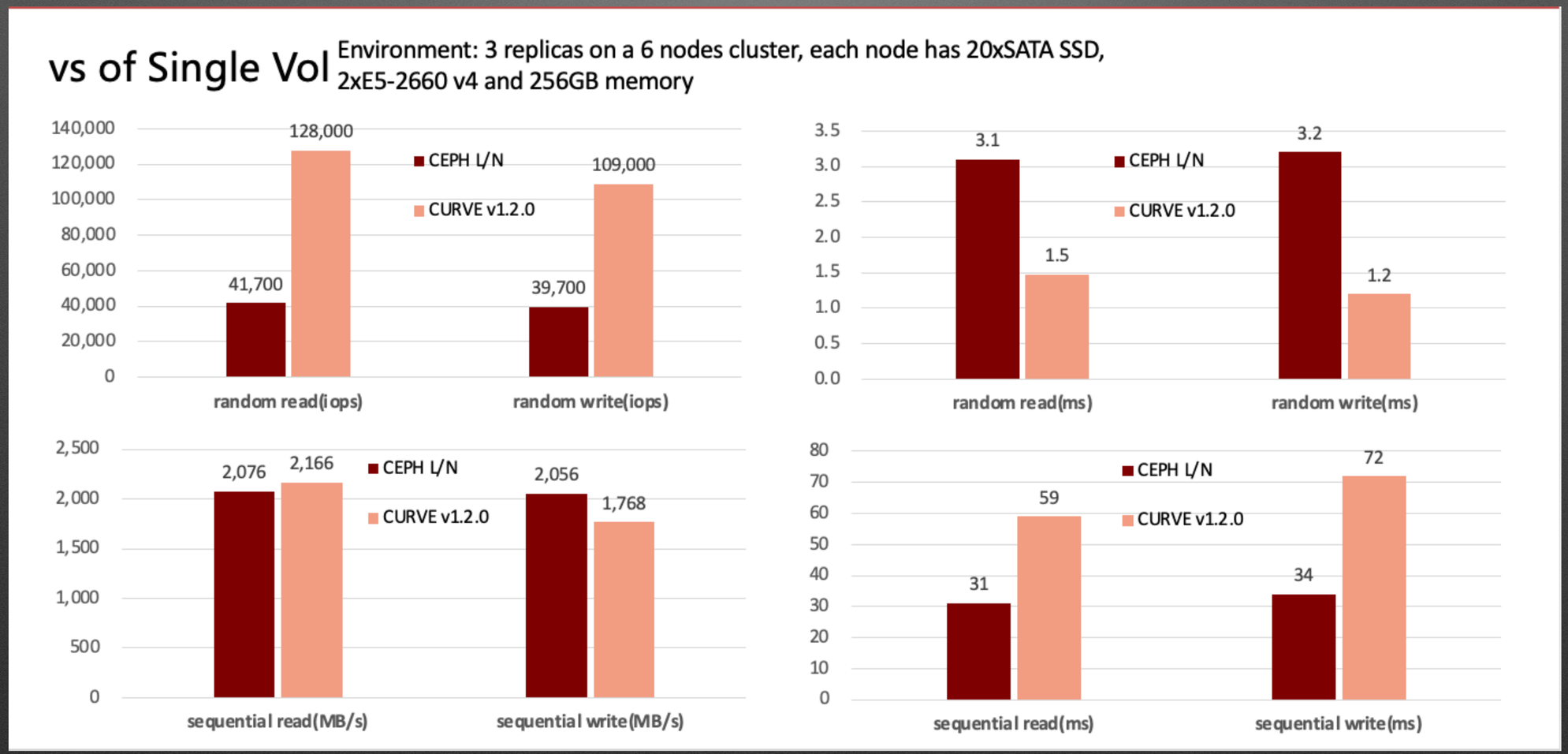
Storage Engine Comparison (vs. Ceph)

META MANAGEMENT	CURVE	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
IMPROVEMENT MEASURES	None overlay write of File no need write RAFT Log	

Storage Engine Comparison (vs. Ceph)

- Better performance
- Deployed on commodity hardware
- Less costs on CPU

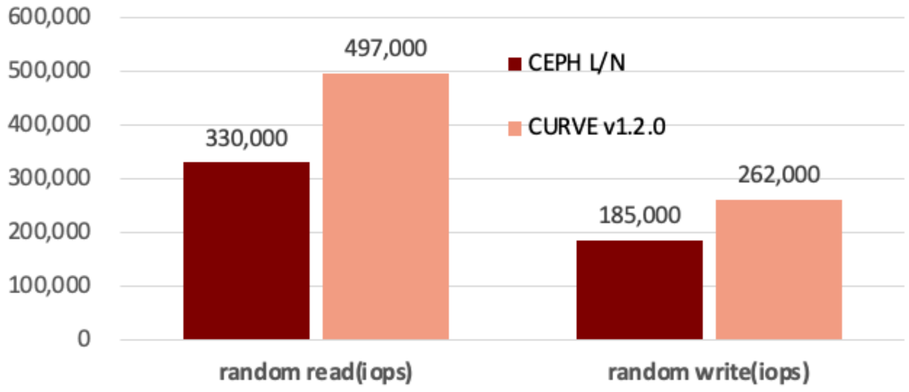
Performance (vs. Ceph RBD)

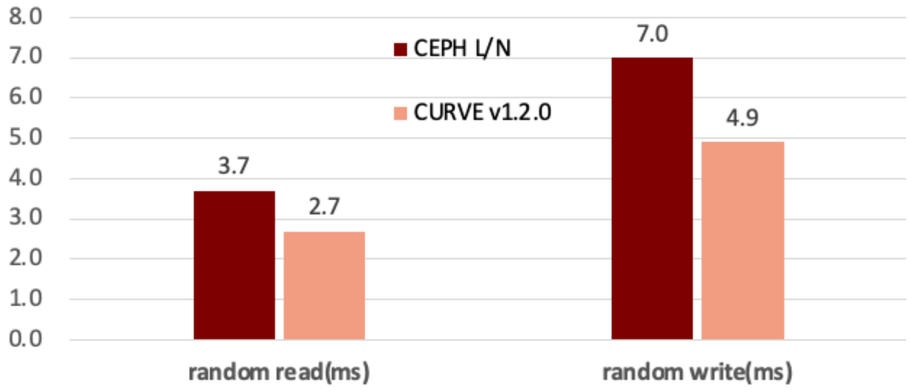


Performance (vs. Ceph RBD)

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

I/O Jitter (vs. Ceph)

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

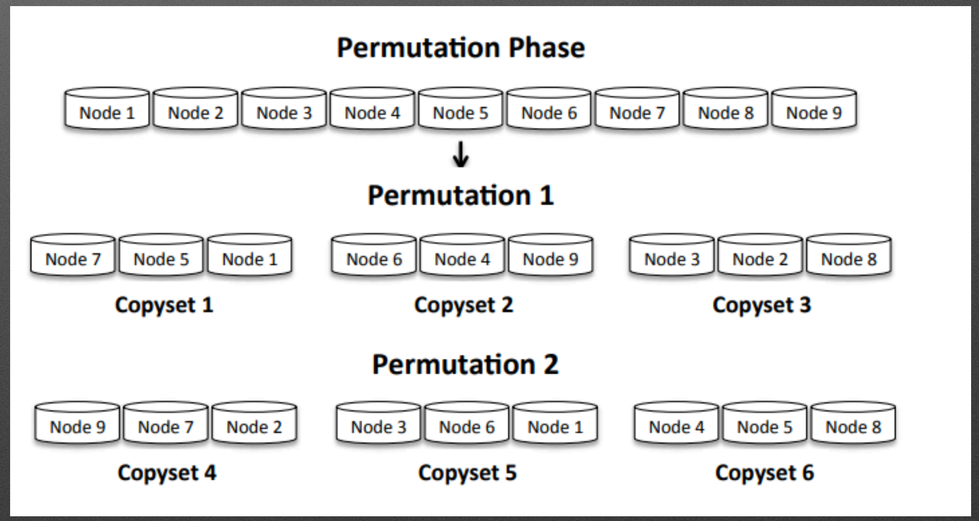
FAULTS CASE	CURVE I/O JITTER	CEPH I/O JITTER	COMMENT
ONE DISK FAILURE	4s	7s	
ONE SERVER FAILURE	4s	7s	
SERVER RESPONSE VERY SLOW	4s	unrecoverable	frequently delay of disk i/o are very long
NETWORK LATENCY 50MS	1s frequently	7s recently	

Automated Operation

STORAGE	UPGRADE DISRUPTIONS	CAPACITY EXPAND DISRUPTIONS	SLOW DISK DETECT/REPLACE	CAPACITY/LOAD BALANCE
CURVE	no disruptions	expanded by pool without performance degradation	No I/O jitter	balance in time / with no performance degradation
CEPH	minimal disruptions	expanded in pool with performance degradation	I/O jitter Occasionally	balance at specified period

Balance Strategy

- 3 replicas by MultiRAFT
- Nodes grouped by Copyset
  - Data in copyset syncs with RAFT
- Leader in charge of writing data to other members in Copyset and reading data
- Every node in cluster have similar number of Copysets
  - capacity balance
- Every node in cluster have similar number of Leaders
  - load balance

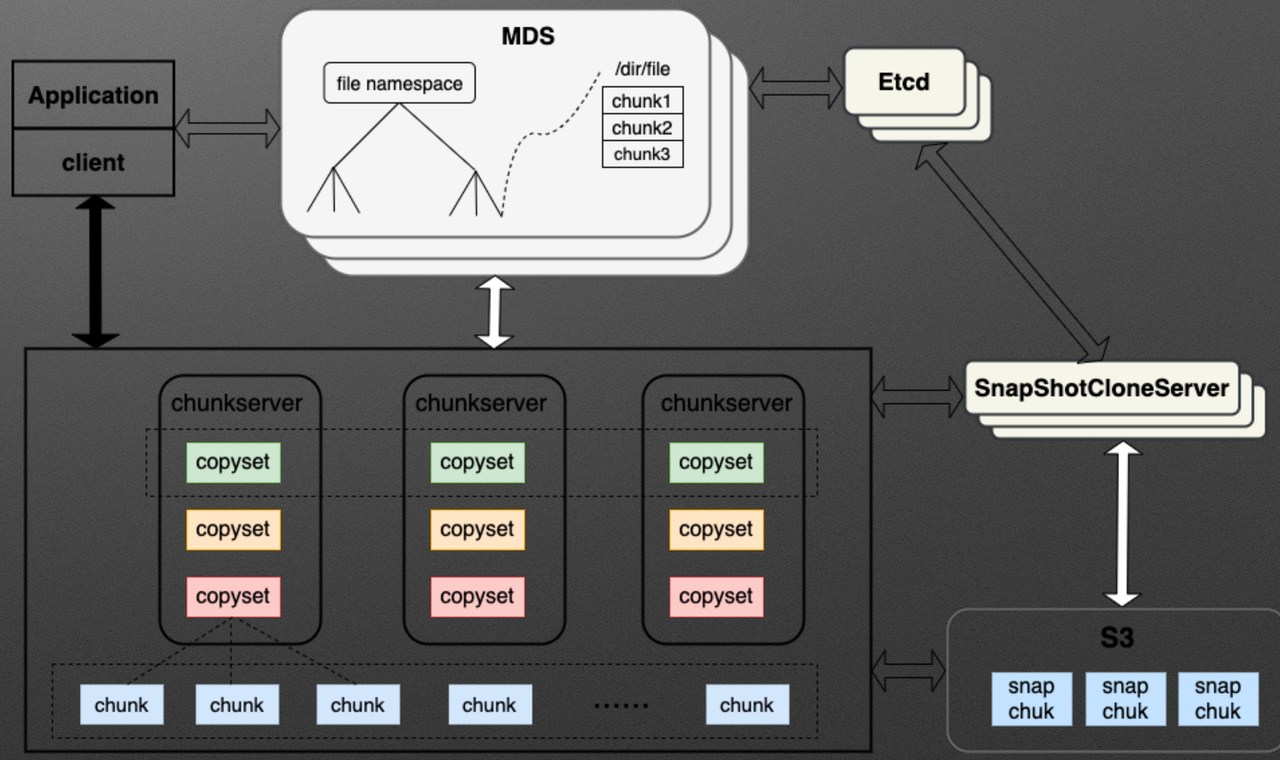


Cloud Native Storage

CLOUD NATIVE LEVEL	CURVE	CEPH + ROOK	COMMENT
BASIC INSTALL	Y	Y	automated application provisioning and configuration management
SEAMLESS UPGRADES	Y	Y	patch and minor version upgrads supported
FULL LIFECYCLE	On Roadmap	Y	app lifecycle, storage lifecycle(backup, failure, recovery)
DEEP INSIGHTS	On Roadmap		metrics, alerts, log processing and workload analysis
AUTO PILOT	On Roadmap		horizontal/vertical scaling, auto config tuning, abnormal detection, schedule tuning

CurveBS Features

- High performance distributed block storage
- support Snapshot / Clone
- Cloud Native (CSI / StorageClass)



CurveFS Features

- High performance filesystem
- POSIX-compatiable
- Cloud native support

CurveFS vs. CephFS

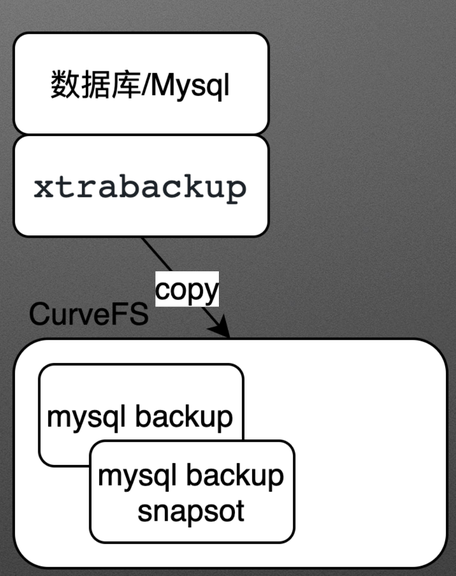
STORAGE	HIGH PERFORMANCE / CACHE	CLOUD NATIVE LEVEL	REPLICATION CONSISTENCY
CURVE(CURVEBS / CURVEFS)	Local/Remote (inode/ date) Cache	Auto pilot	RAFT
CEPH /CEPHFS / CEPH+ROOK	Remote data Cache	Full lifecycle	PAXOS

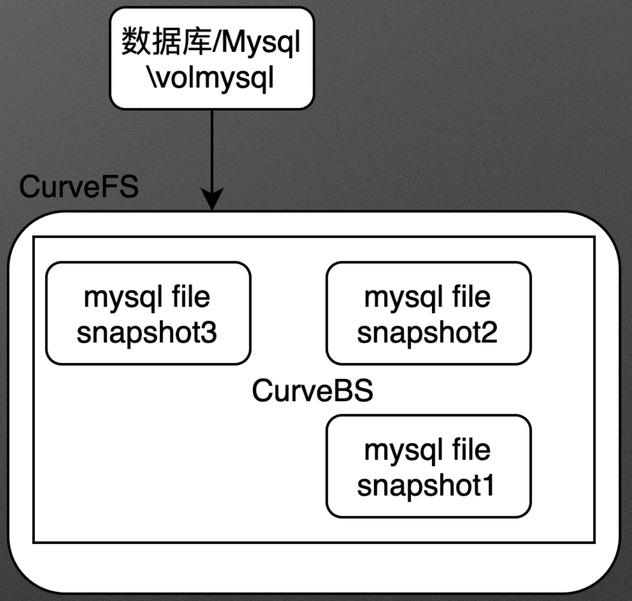
User Cases

- Database
- Middleware
- Big Data / AI
- File Shared
- Remote Data Synchronization

User Cases (Database)

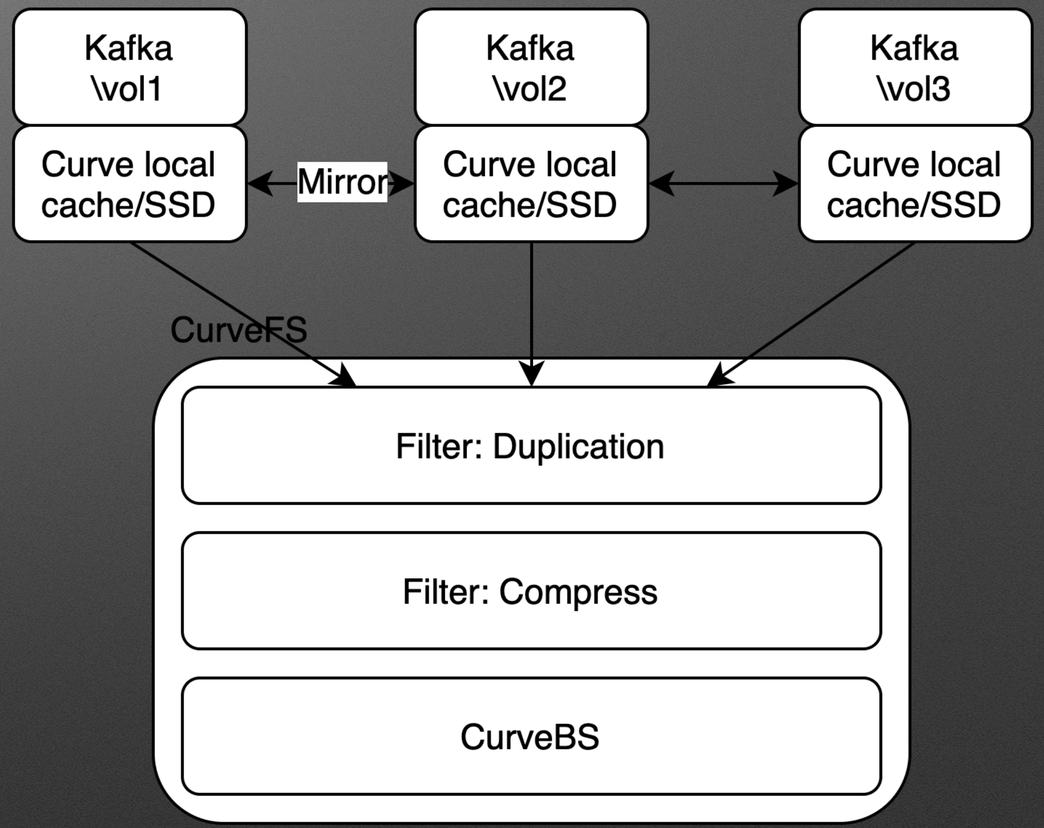
- Backup database to remote CurveFS: left picture.
- Use CurveFS for database, and create snapshots on CurveBS: right picture.





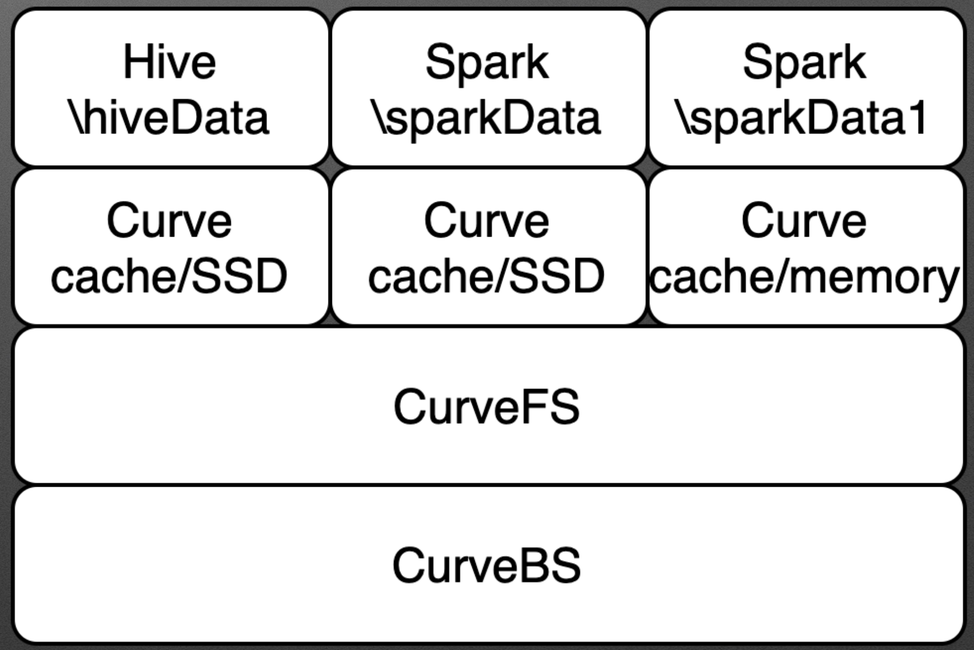
User Cases (Middleware)

- Support cache module to provide near local disk performance
- Use CurveBS storage as backend storage to provide high performance



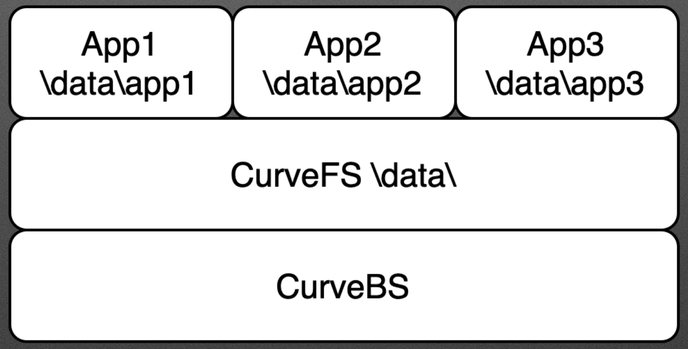
User Cases (Big Data / AI)

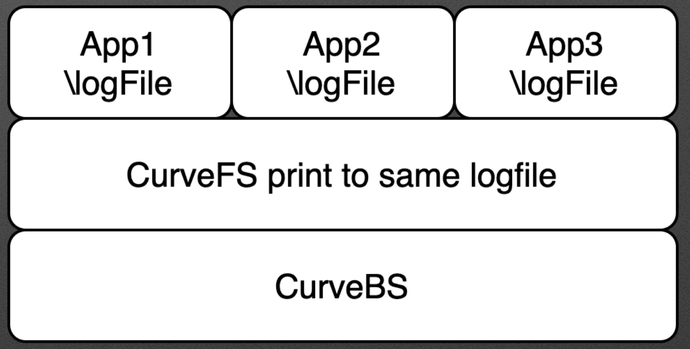
- Support cache module to provide near local disk performance
- Use CurveBS storage as backend storage to provide high performance



User Cases (File Shared)

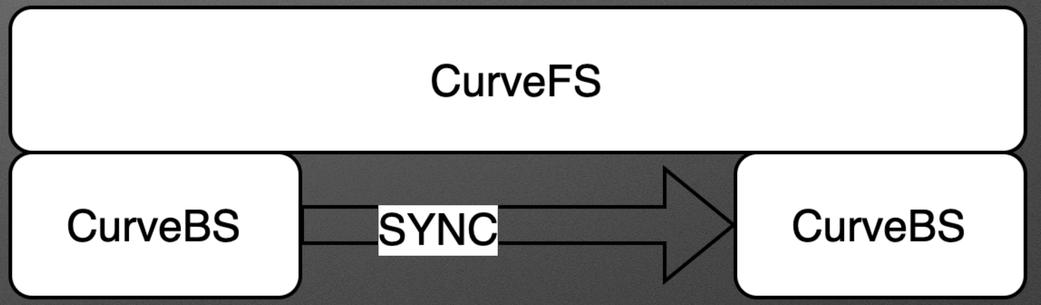
- Use CurveBS storage as backend storage to provide high performance





User Cases (Remote Data Synchronization)

- CurveFS can sync data among different places



Current Status

- Widely used in core business Units
  - NetEase YanXuan (China’s leading private-label e-commerce brand providing a premium selection of high-quality and cost-effective lifestyle products for Chinese consumers)
  - NetEase Cloud Music (China’s leading music streaming service with over 800 million users)
  - NetEase YouDao (a fully user-oriented full-chain educational technology company)
  - NetEase Games
- Mostly used scenarios
  - Block storage for KVM and Kubernetes
  - Storage for Spark and Kafka
- Developing for scenarios
  - Storage for Oracle / mysql
  - Storage for Cloud native database
  - New Object Storage Engine

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
- File meta data preallocate
- RAFT optimization
  - ParallelRaft for write
  - Reduce write magnification for file new write
- Cloud tiering support

Thanks