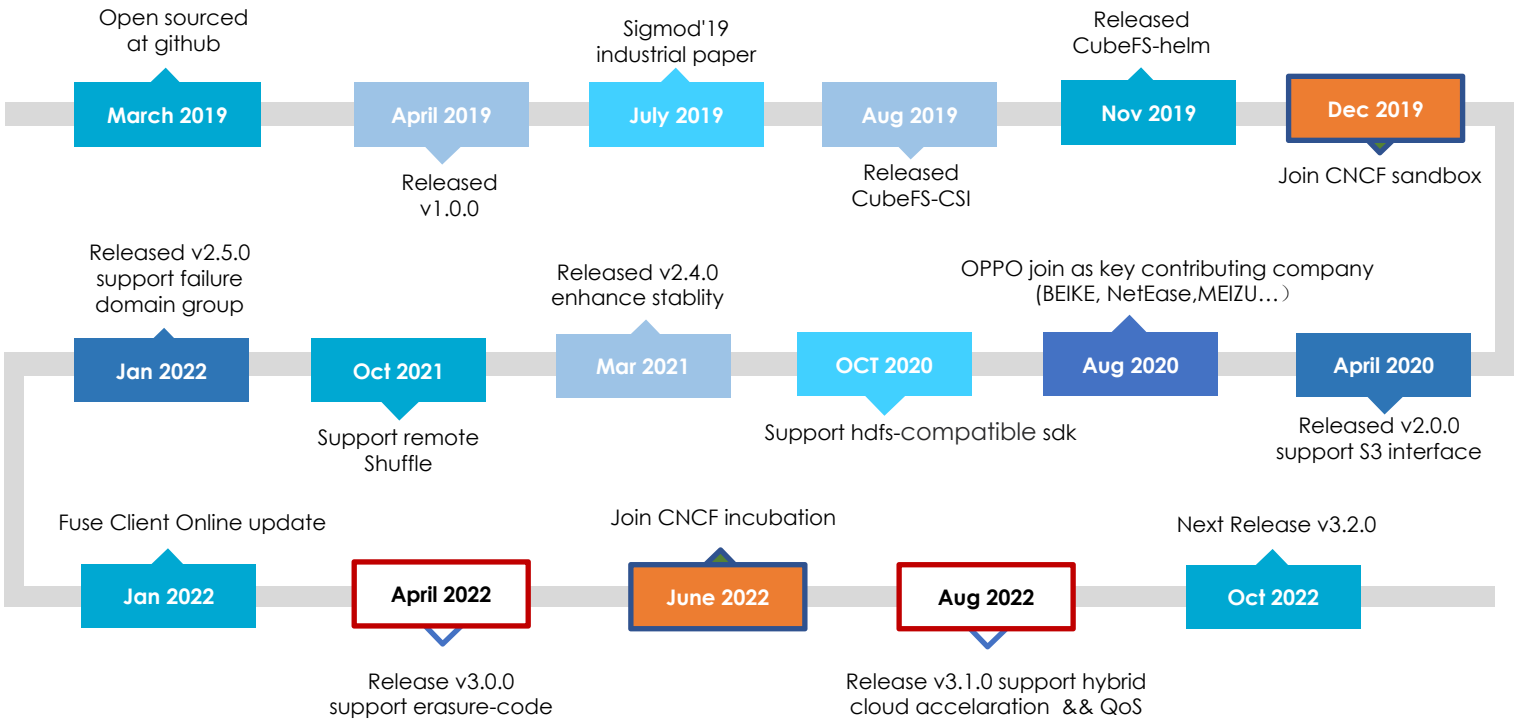
The development and evolution of CubeFS in OPPO

KubeCon | CloudNativeCon
North America 2022
BUILDING FOR THE ROAD AHEAD
DETROIT 2022

October 24-28, 2021

leon chang
Storage Technology
Architect, *OPPO*

CubeFS History



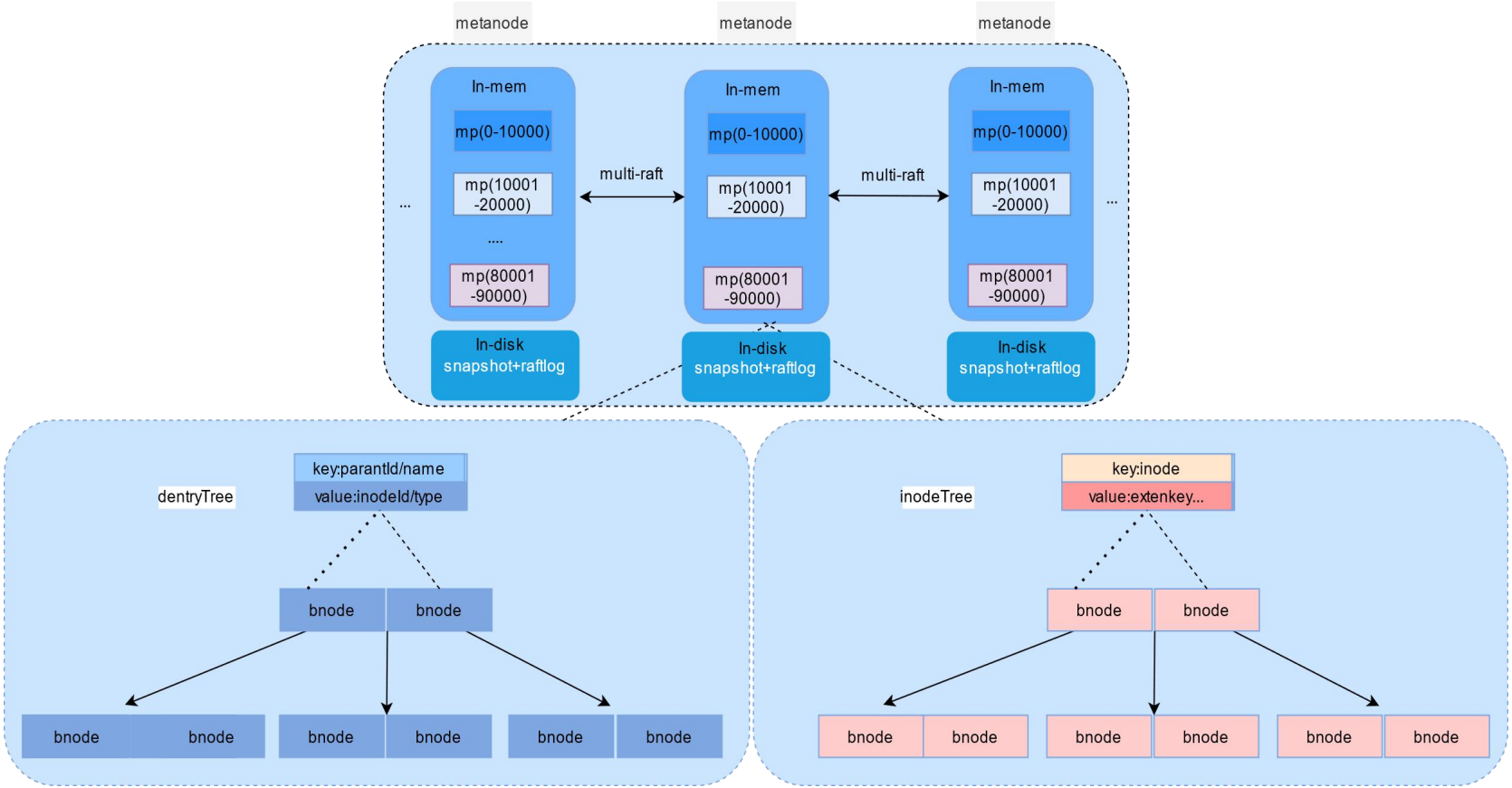
Community Growth

- Open sourced since March 2019
- Sandbox project since Dec 2019
- Incubation project since June 2022
- Statistics(https://chubaofs.devstats.cncf.io/):
  - before vs since joining sandbox
    - Commits: 516 -> 2814(inc 445%)
    - Code committers: 17 (5 companies) -> 96（15 companies)
    - Pull requests: 56 -> 937 (inc 1500%)
    - Contributors: 27 (5 companies) -> 137 (11 companies) (inc 248%)
    - Contributions: 1112 -> 5305 (inc 377%)
    - Forkers: 110 -> 335 (inc 200%)
    - Watchers: 851 -> 2212(inc 159%)

Background

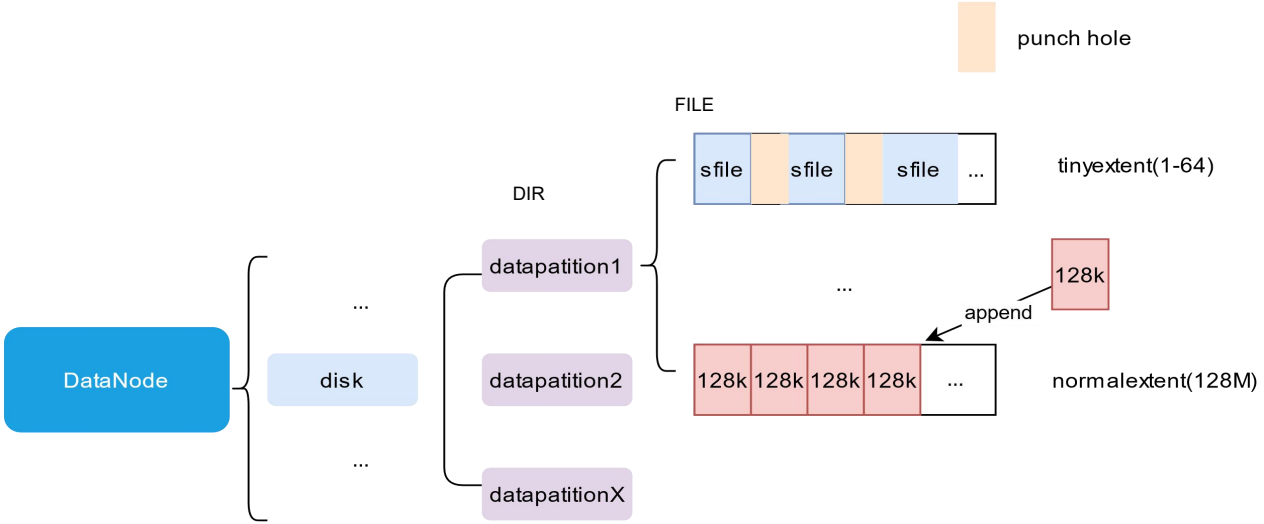
- CubeFS was initially developed to provide storage solution for containerized applications in large scale container platforms.
- Challenges
  - Large number of customers(volumes)
  - Storage usage hard to predict for a single user
  - Various file sizes ranging from KB to TB
  - Diverse read/write patterns, i.e. sequential or random
  - Data shared by upstream and downstream users

CubeFS-Elasticity and scalability for metadata

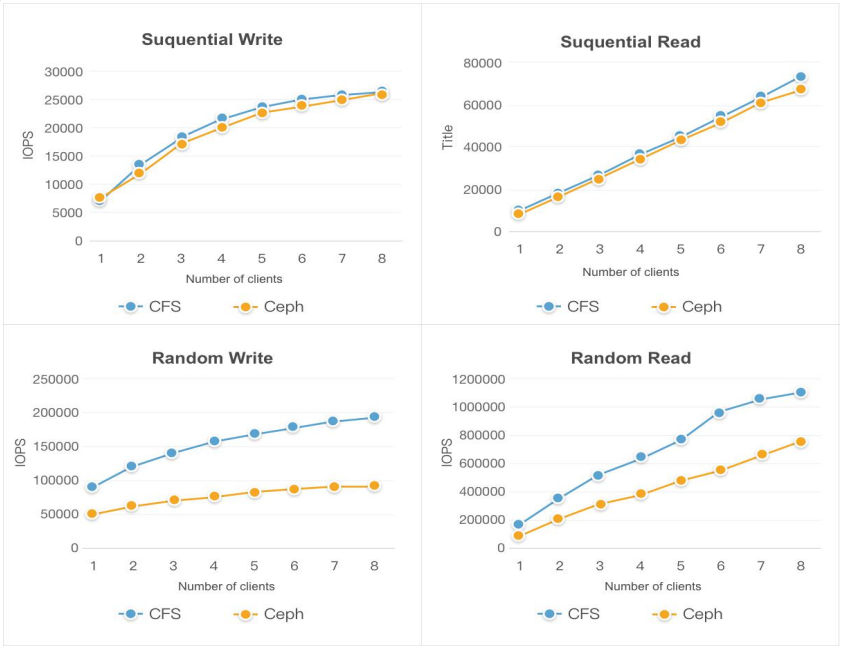


CubeFS-Optimized for small files

- Multiple small files are aggregated in one extent
- Efficient space reclamation: punch hole



CubeFS-performance comparison

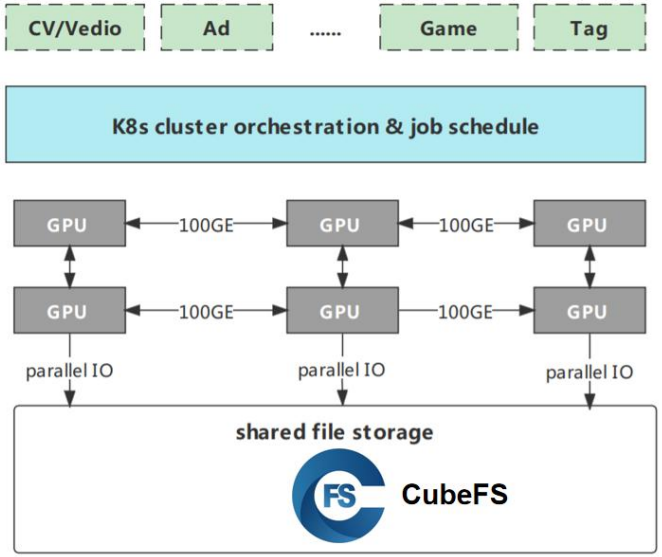


Large file read and write performance

Small file read and write performance

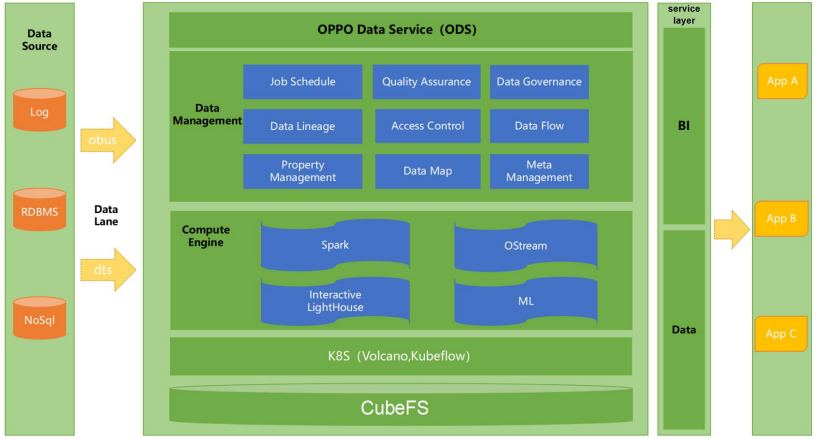
User Adoptions - OPPO

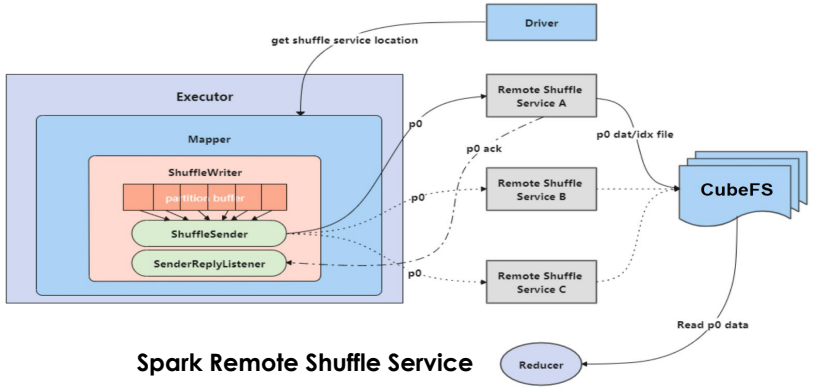
OPPO : A consumer electronics and mobile communications company.



AI Platform

Data Lake



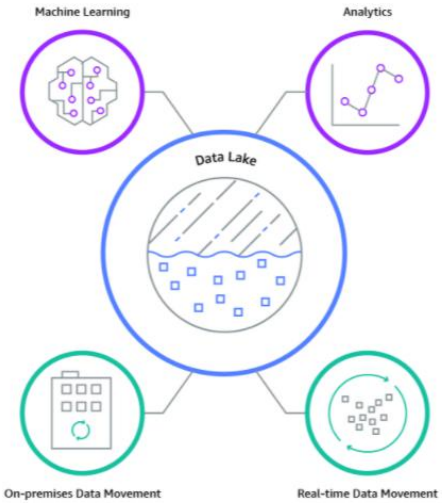


Spark Remote Shuffle Service

DataLake Solutions

What is a data lake?

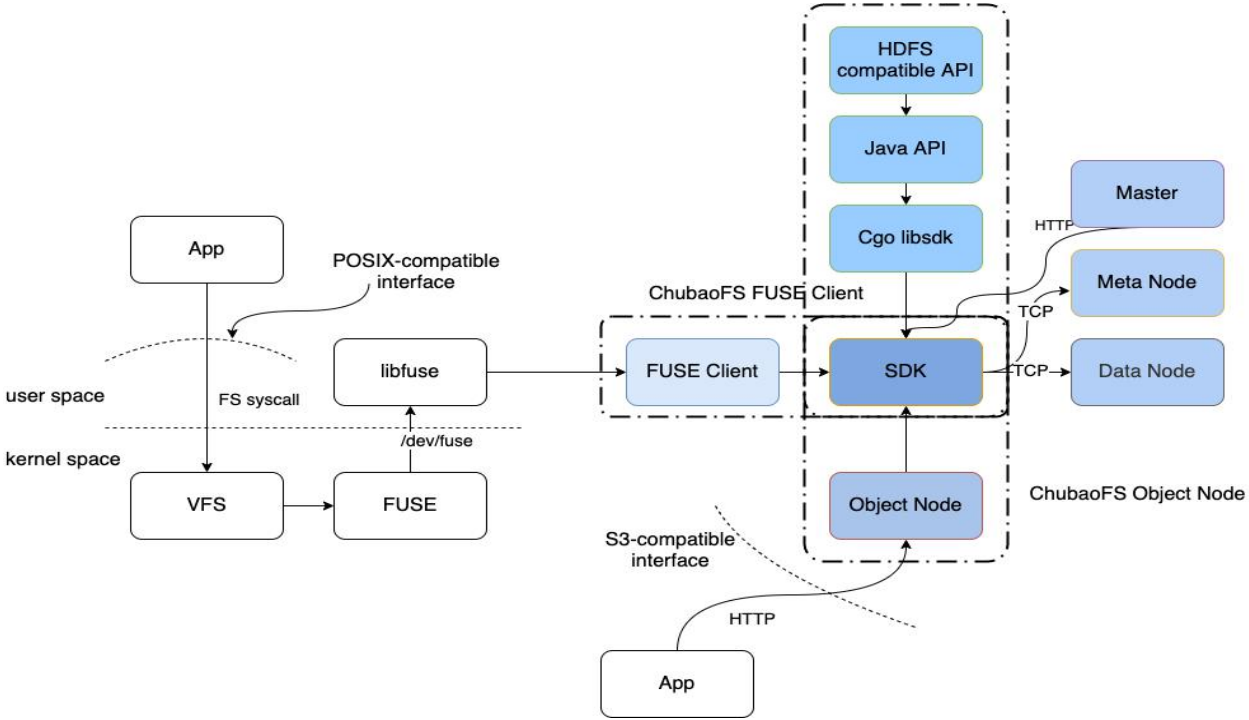
A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions.



- **Provide HDFS interface capability**
- **Erasure code storage to reduce cost**
- **Client local cache for performace**
- **Domain failure to enhance reliablity of massive data scenarios**
- **QOS flow control**

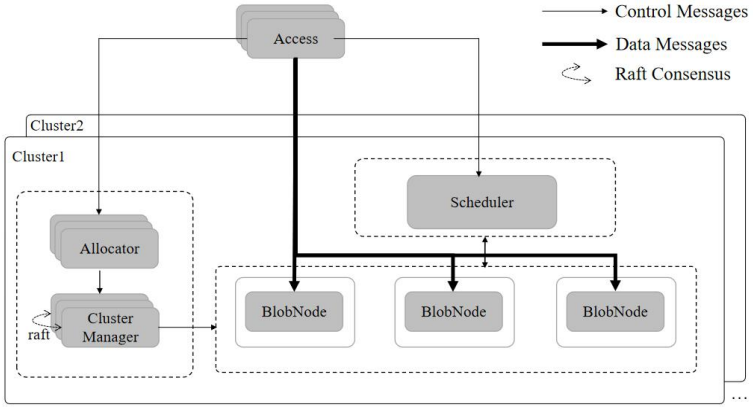
DataLake Solutions: Multi-protocal capability

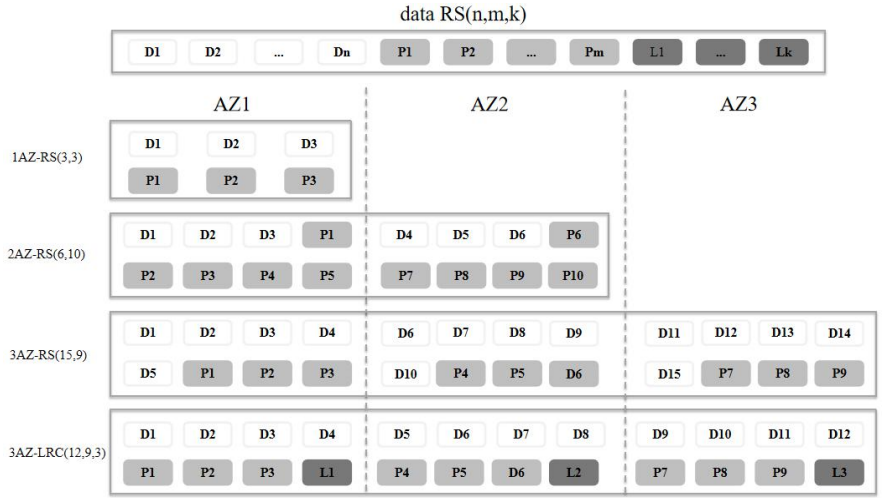
Converge filesystem、S3-compatible interfaces and hdfs interfaces



DataLake Solutions: Erasure Code Subsystem

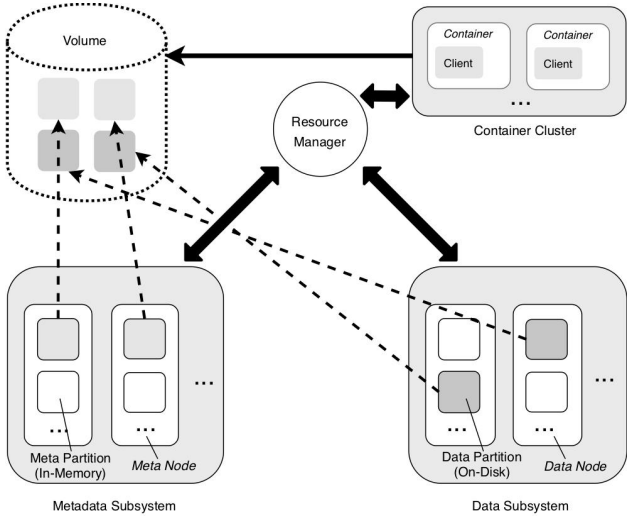
- Main feature
  - Larger cluster capacity
  - Higher durability with 11 9s
  - Lower TCO：reduces redundancy from 3x to 1.33x or less
  - Multi-AZ deployments
  - Multi-specification configurable erasure coding mode



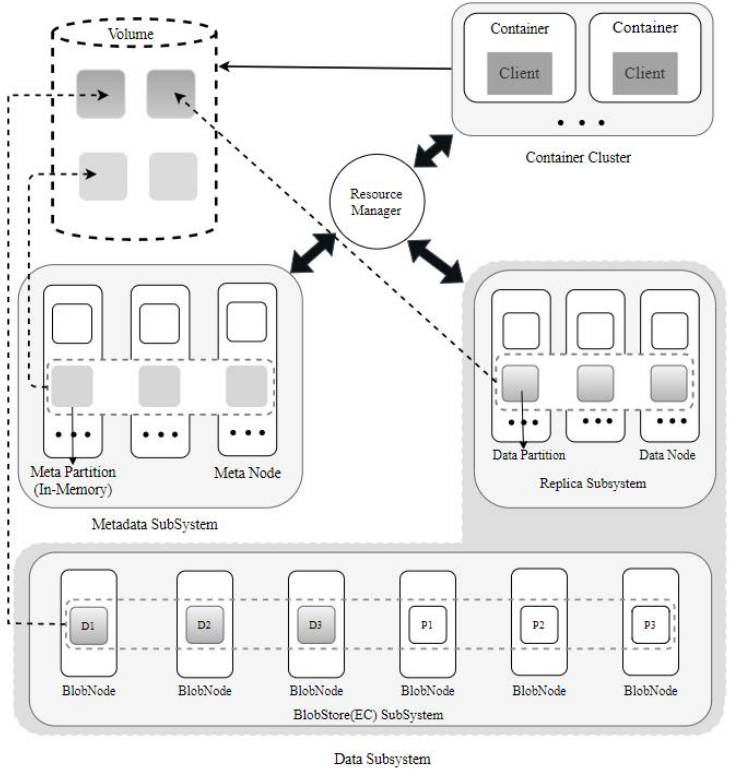


DataLake Solutions: Erasure Code Subsystem

Architecture upgrade



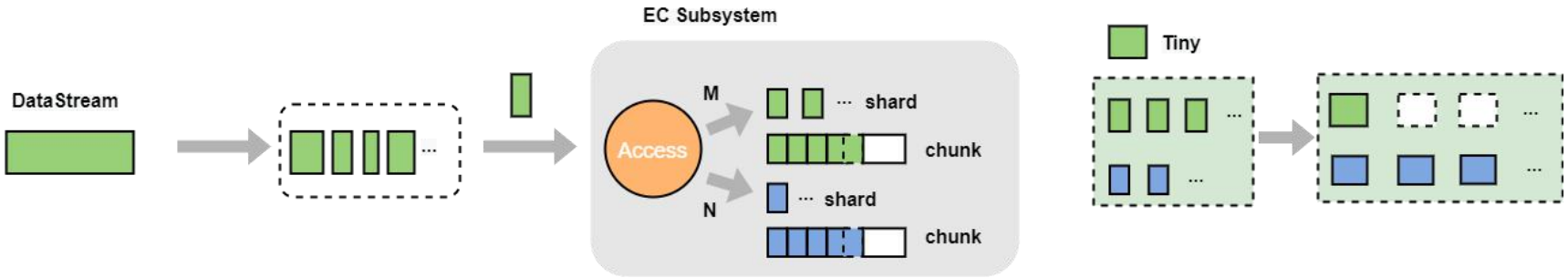
release-2.x



release-3.x

DataLake Solutions: Erasure Code Subsystem

optimazition



- Quorum mechanism: Allow certain write failures to effectively solve the problem of tailing delay.
- Small file EC optimization: trade space for time to improve read performance.
- Efficient garbage collection: Reclaim space through sparse semantics, reducing IO overhead.

Hybrid Cloud Acceration

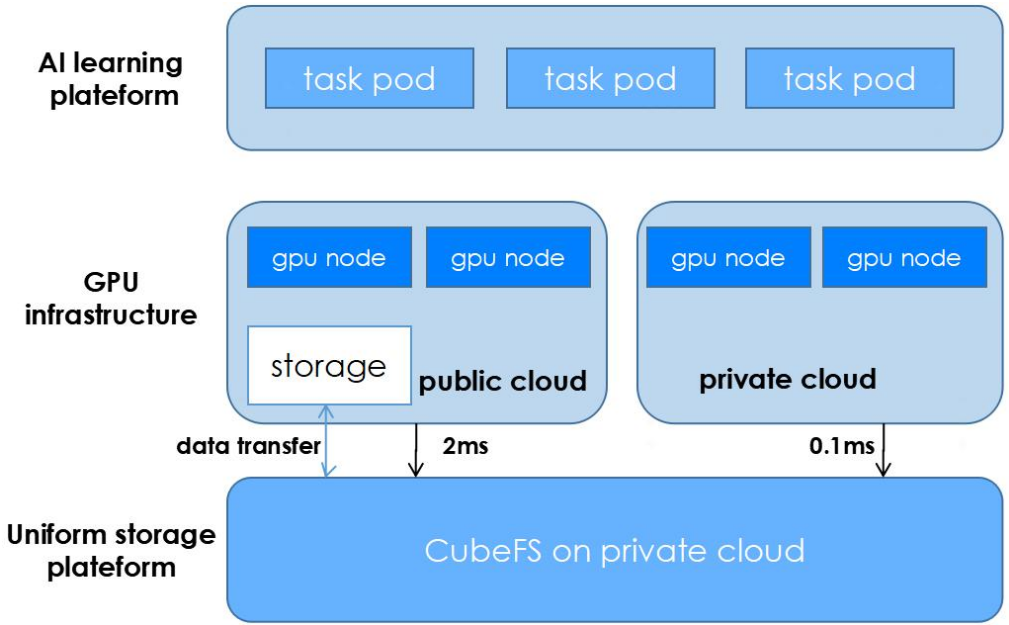
Challenges

- Performance problems in storage during cross regional
- High Cost of data migration
- Data security on public cloud

Under the AI hybrid cloud architecture, cache acceleration brings the following benefits

1) It is uniformly stored in CubeFS to achieve real flexibility for the GPU computing platform.

2) For business transparency, the use mode and performance are consistent with OPPO private cloud experience.



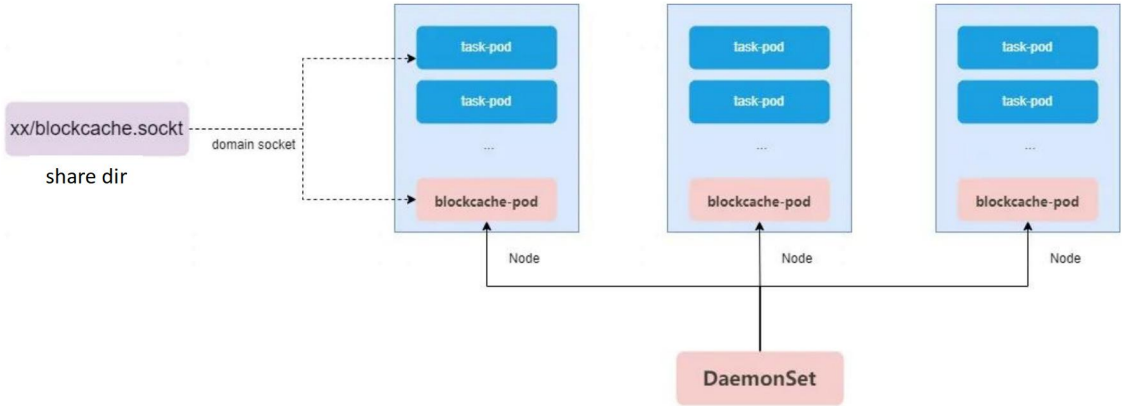
Hybrid Cloud Acceration

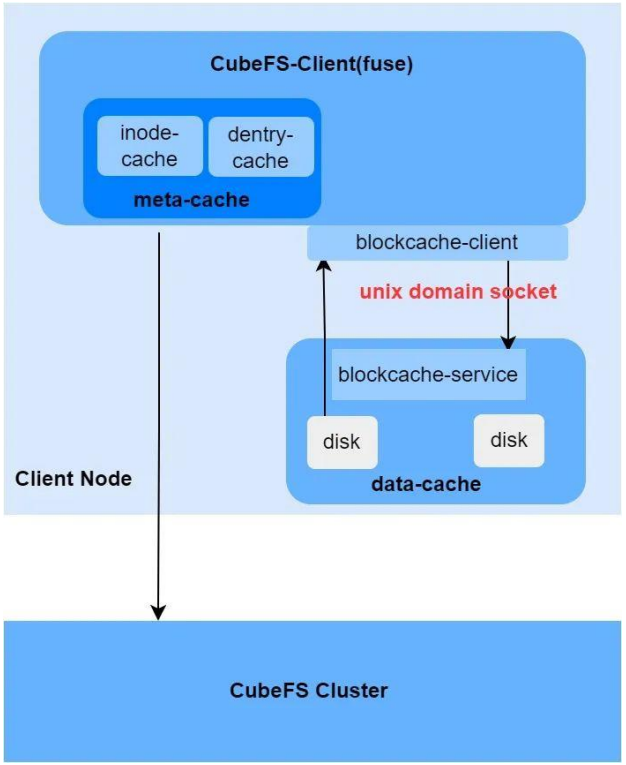
MetaCache:

- Cached in the memory of the CubeFS client.
- Caches inode and dentry metadata。

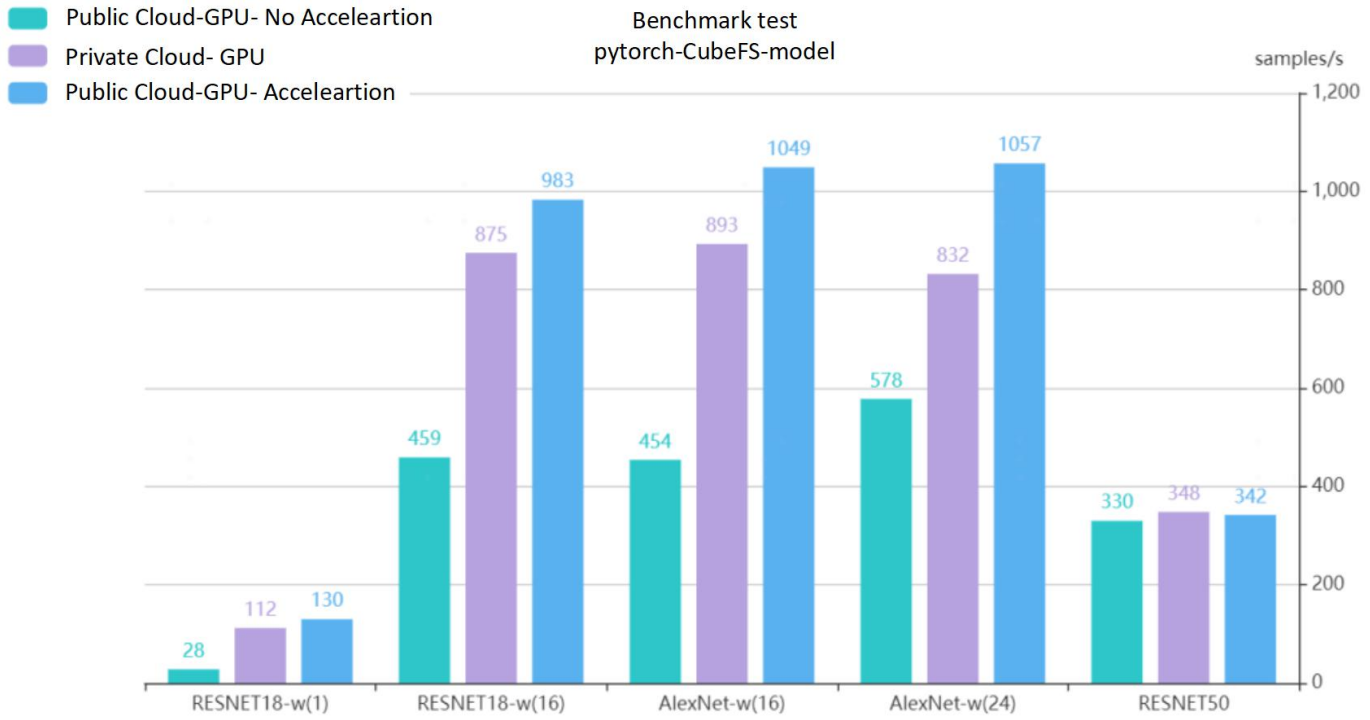
DataCache:

- Data cache service,need consider the resource limitation and generliariy
- Index management and data management





Hybrid Cloud Acceration



DataLake Solutions: QOS

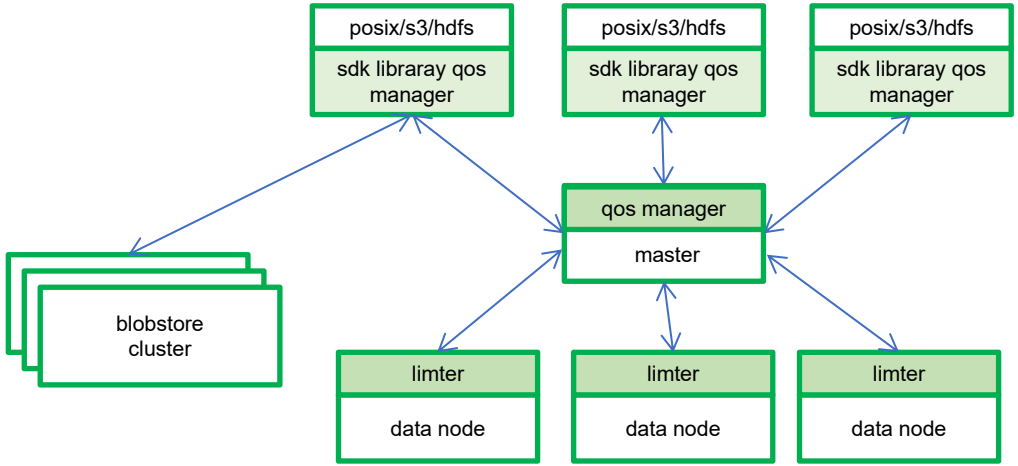
QOS flow control system

background

In multi-tenant scenarios, business has no control logic, io and traffic resources may be congested, and traffic bursts

feature

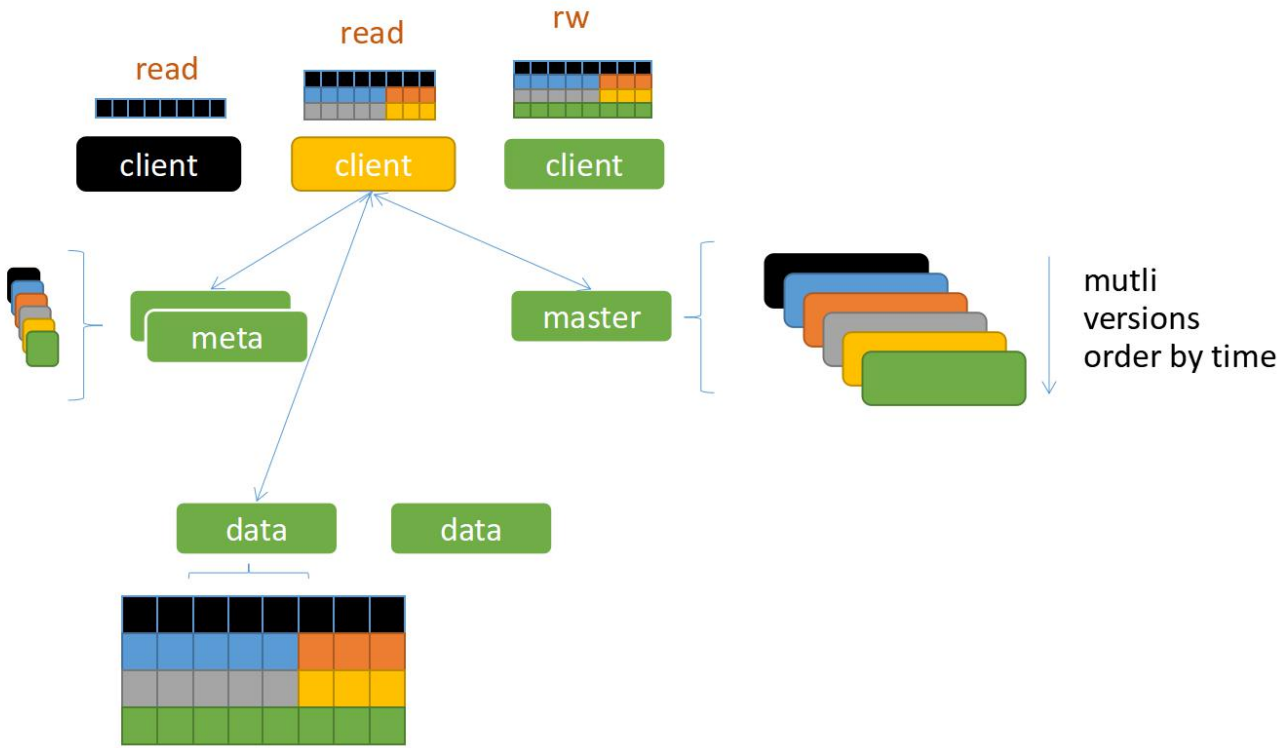
1. Does not depend on external components
2. Resource pre-allocation and dynamic adjustment
3. Dynamic adjustment of request period



AI Platform solution:snapshot

snapshot scenario

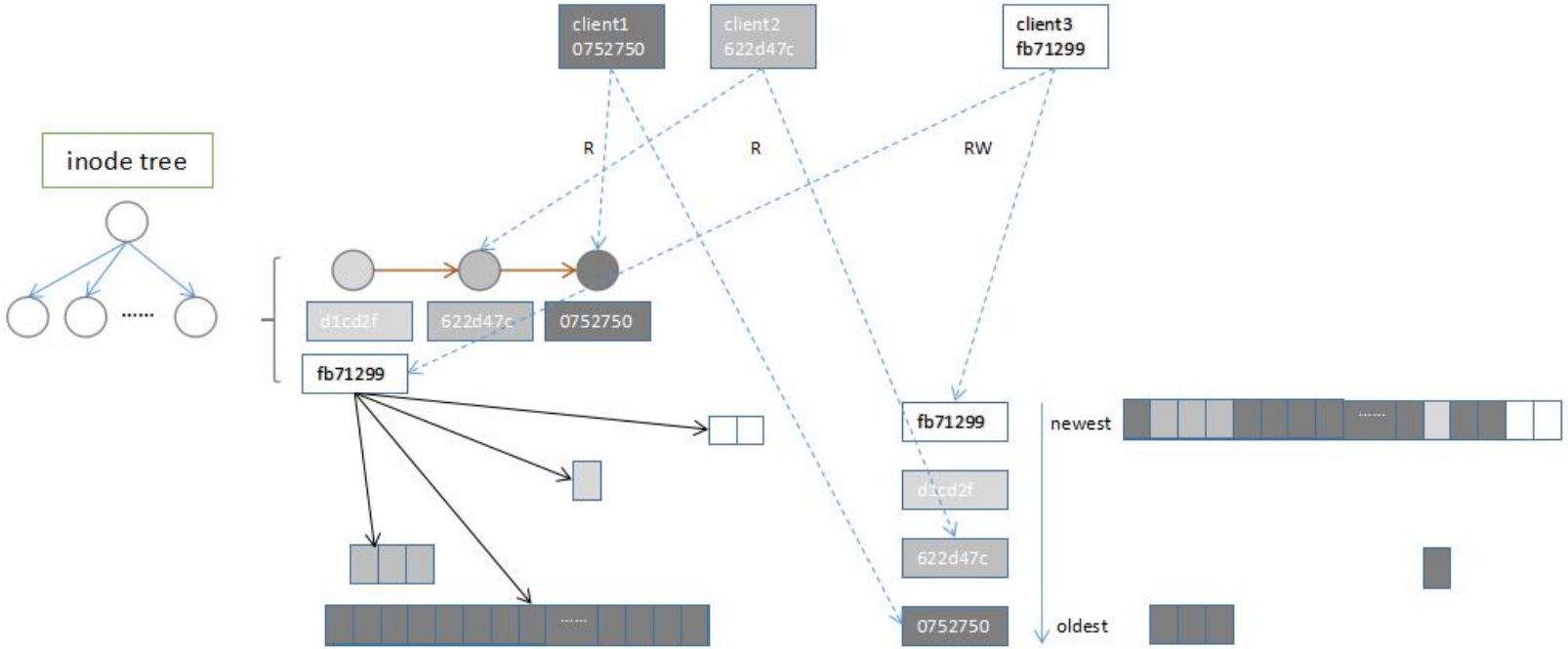
- Create snapshots in seconds
- No-lag snapshot version reads
- No write amplification
- Metadata, data without space redundancy
- Strong consistency



AI Platform solution:Snapshot

KubeCon CloudNativeCon North America 2022

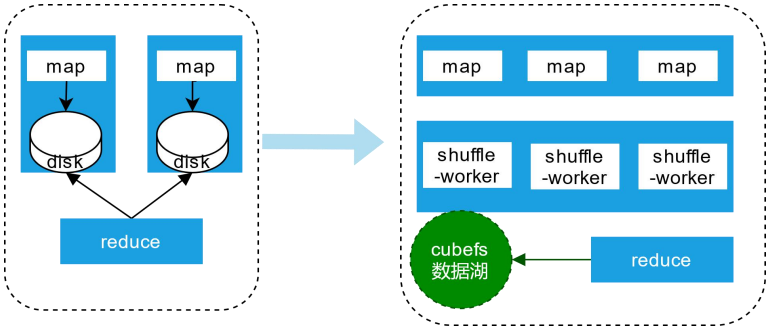
snapshot multi version index

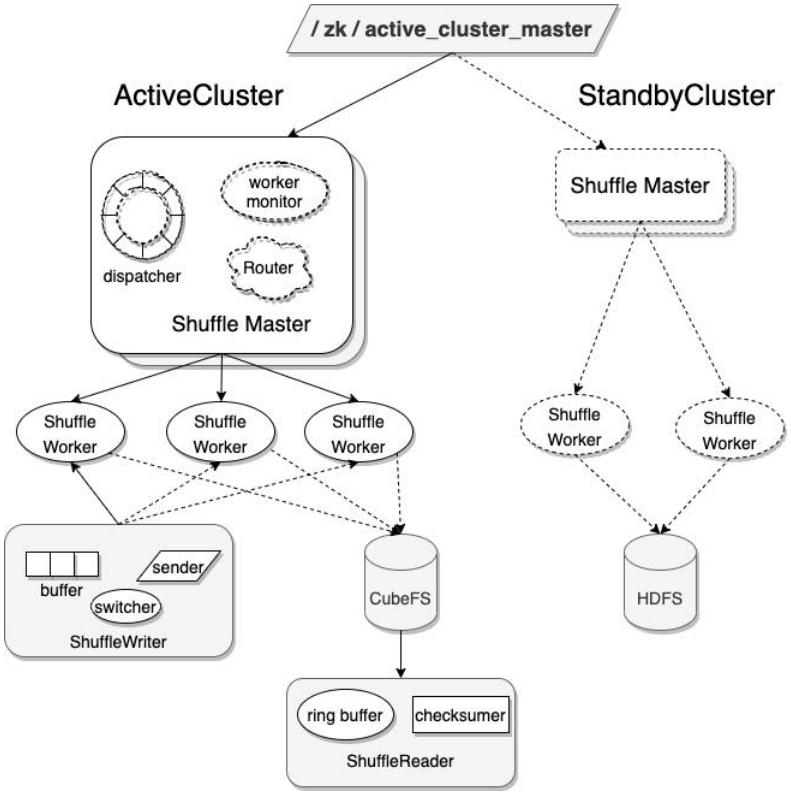


Remote shuffle solutions

Feature

- Local access to reduce cost
- Performace optimiztion
- Flexsible Replicas Strategy





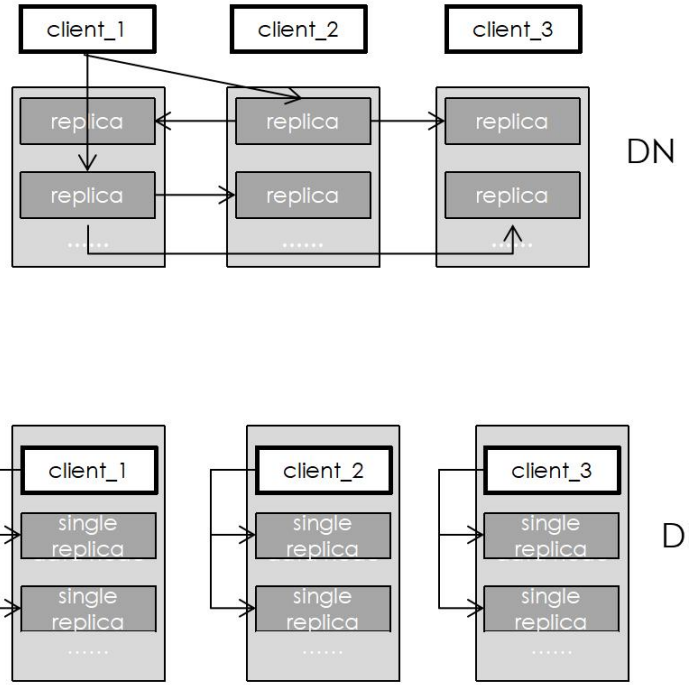
Remote shuffle solutions

Flexible replication strategy

Replica Options
- one to three replica

Single Replica Feature
- reduce TCO
- reduce network traffic
- reduce write latency

Single Replica Scenario
- service with low reliability requirements
- hadoop remote shuffle



RoadMap

Under Development:

- Cubekit: structured storage for mobile applications across devices
- Hdfs protocal compatible: Use Hdfs protocal directly access cubefs instead client jar package
- Atomic rename: reduce the middle status caused by abnormal

Release later:

- Snapshot
- Enrasure code reconstrucation

BUILDING FOR THE ROAD AHEAD

DETROIT 2022

THANK YOU!

https://github.com/cubefs/cubefs