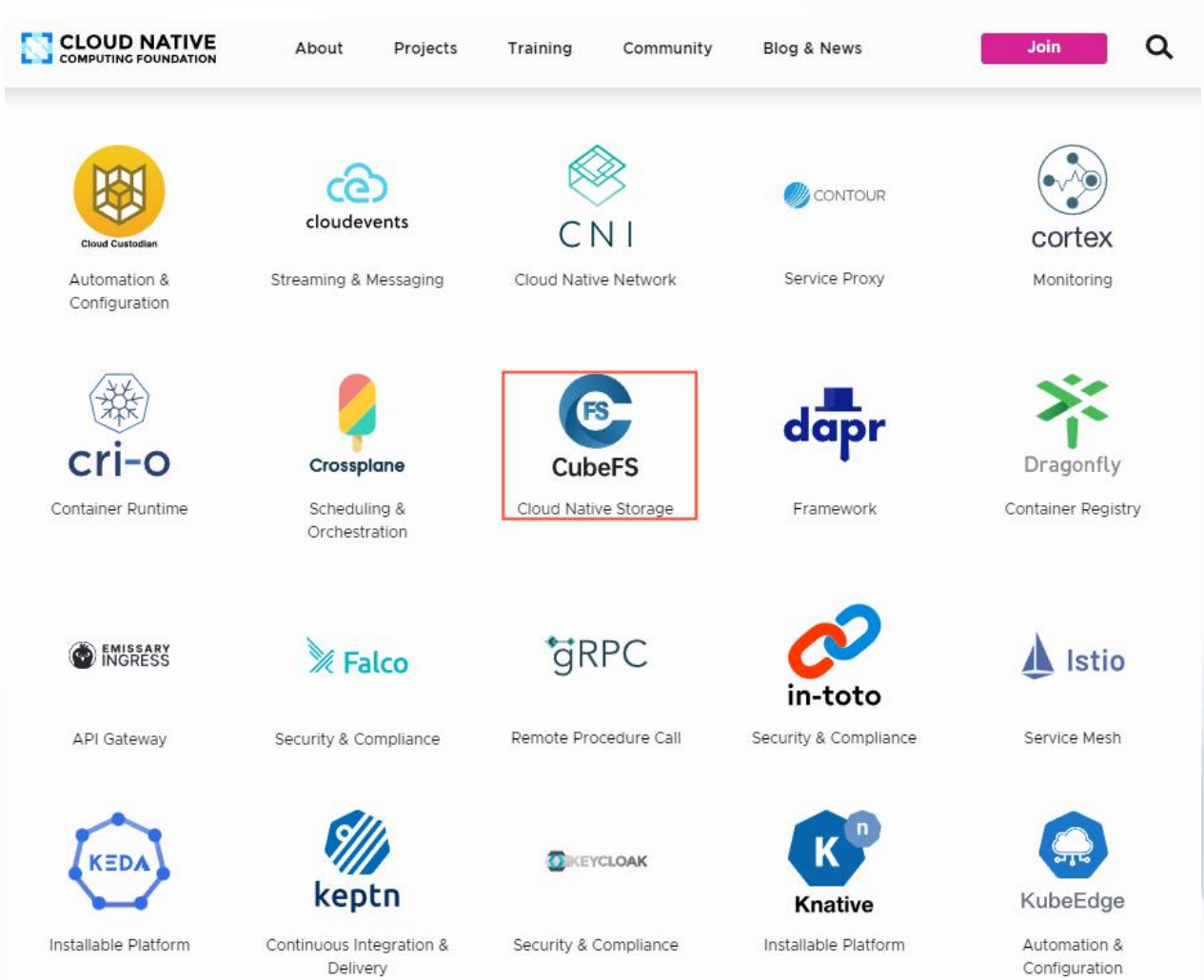

GOTC 2023
全球开源技术峰会
THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE
OPEN SOURCE, INTO THE FUTURE
「Cloud Native Summit」专场
本期议题：The Best Practice of Machine Learning Platform Storage Based on CubeFS
常亮 2022年5月28日

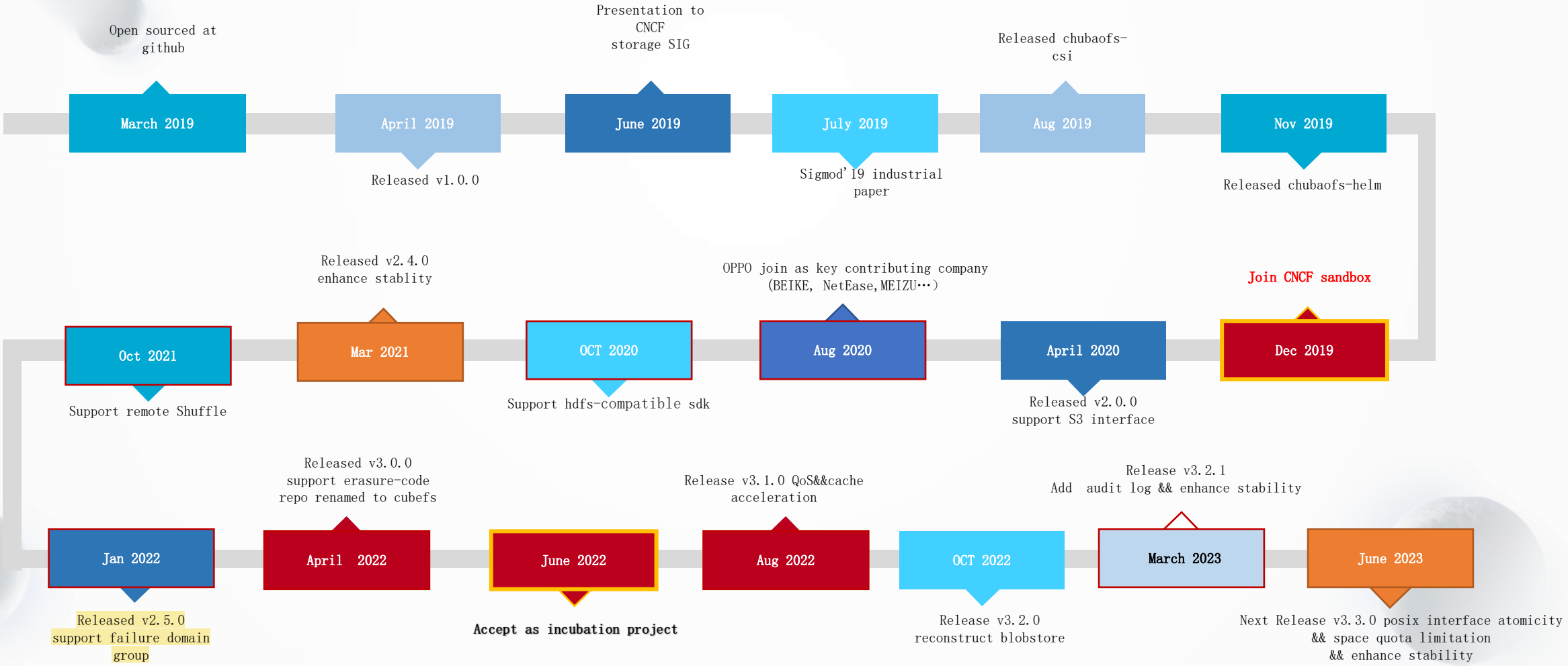
CubeFS Introduction

CubeFS is a new generation of cloud-native open source storage product hosted by the Cloud Native Computing Foundation (CNCF). Currently in the incubation stage, CubeFS has complete file and object storage capabilities.

Product website: https://cubefs.io



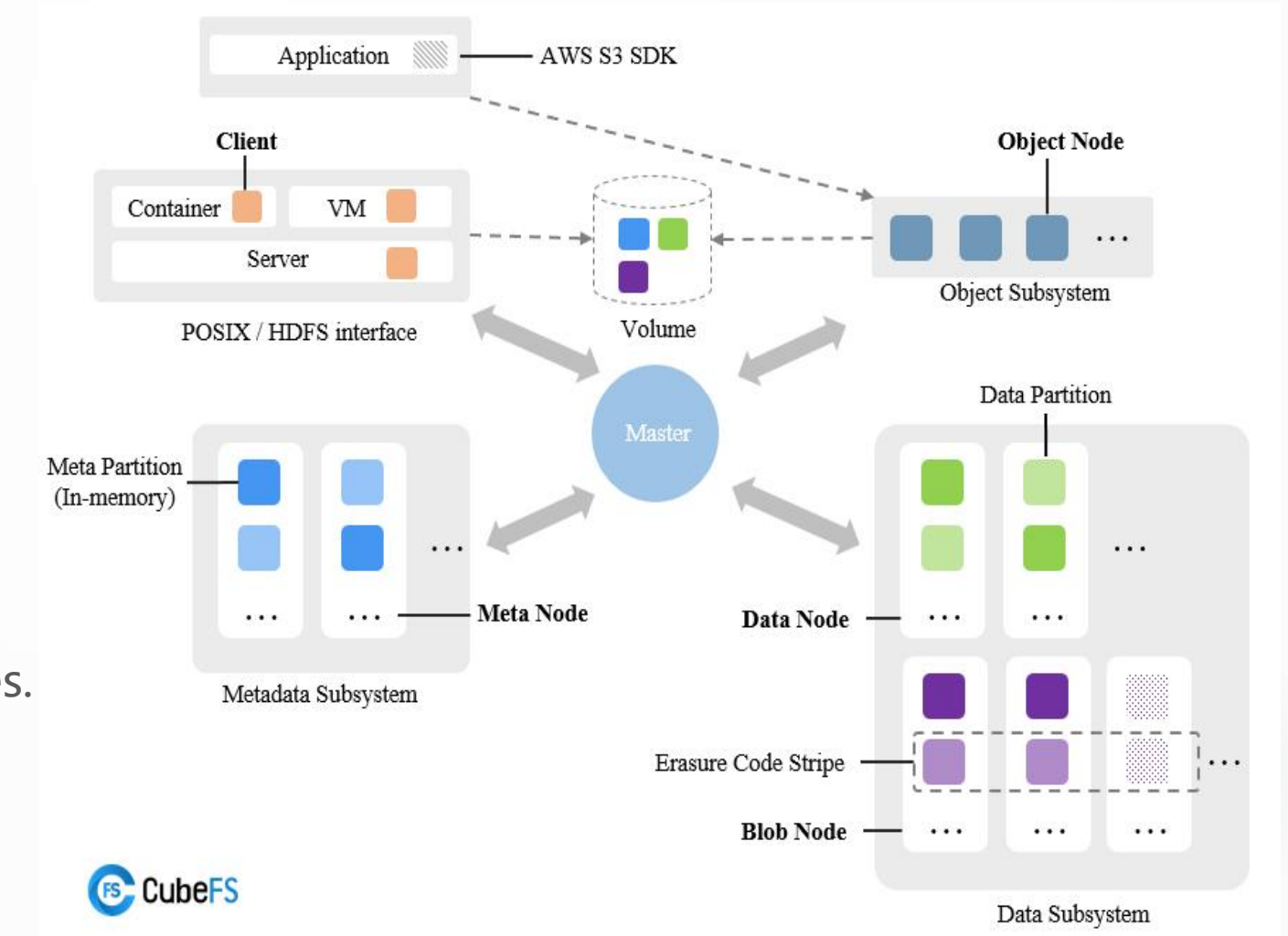
CubeFS History



Architecture

Key features

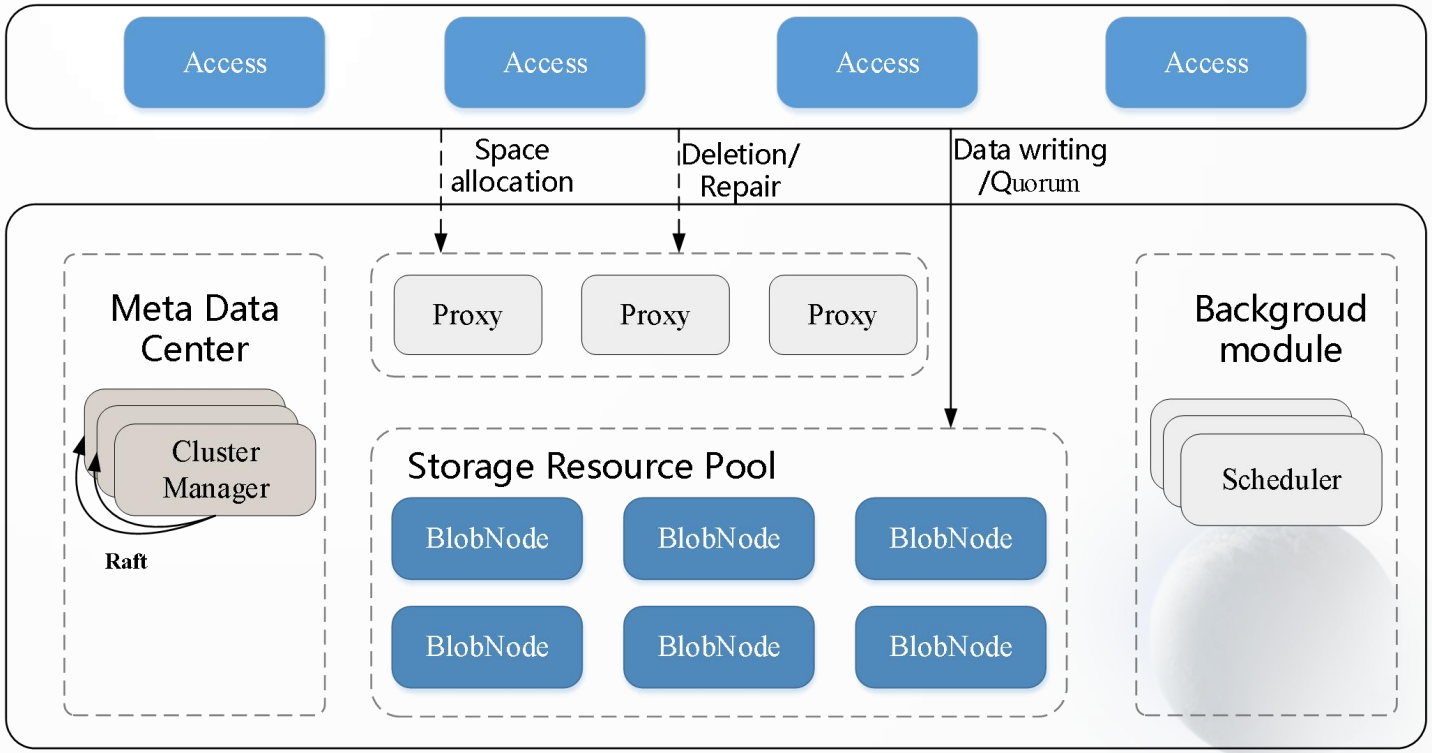
- Compatible with various access protocols such as S3, POSIX, and HDFS
- Multi-engine(Multi-replicas and erasure coding)
- Multi-tenant
- Highly Scalable
- High-performance
- Cloud-native,based on the CSI plugin, CubeFS can be quickly used on Kubernetes.



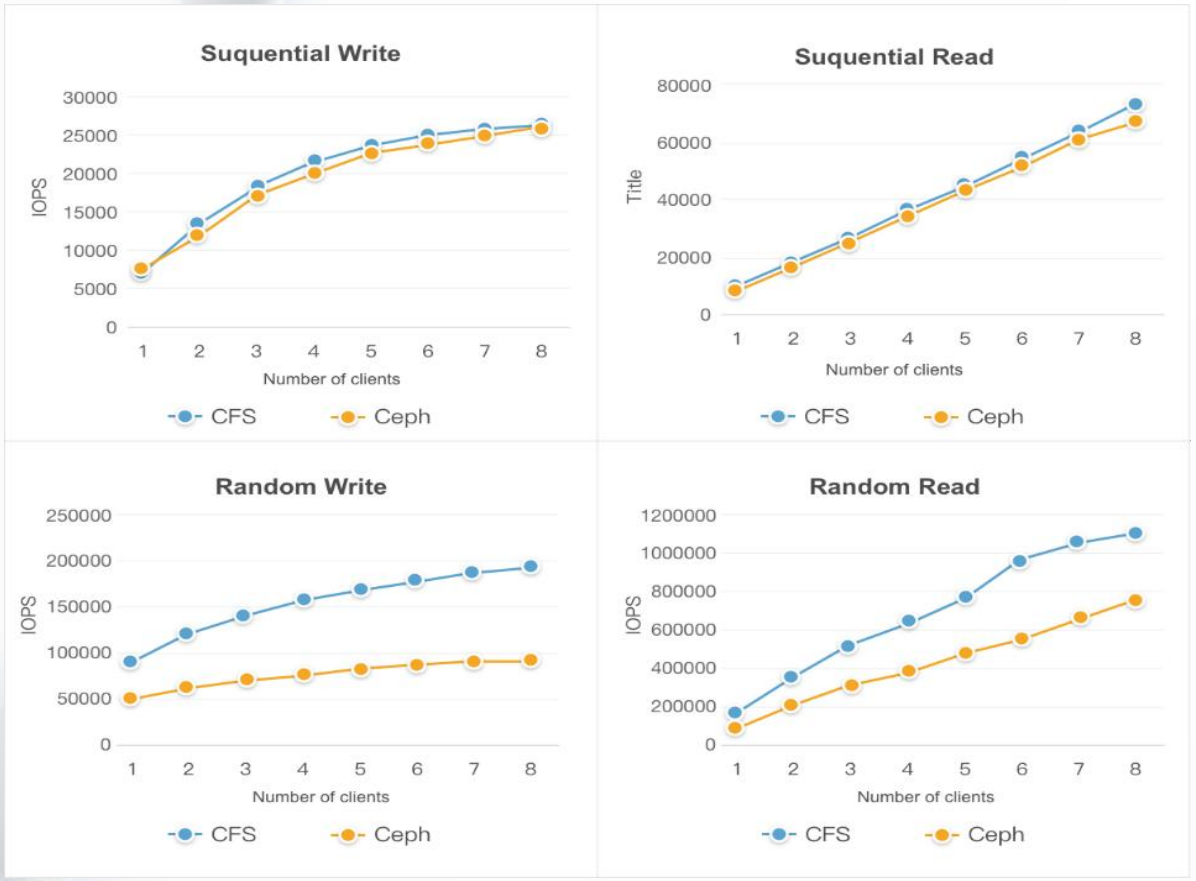
Architecture: Erasure Coding Engine

Key tips

- Online encoding
  - Access layer calculates the erasure code directly online and writes it into storage node.
- High availability
  - Raft ensures high availability of metadata service with second-level switching.
- High reliability
  - Background services such as data inspection, data repair, and bad disk detection ensure high reliability.
- Multi-AZ deployment
  - Supports 1, 2, and 3 AZ deployments, with AZ-level disaster recovery support.



CubeFS-performance comparison



Large file read and write performance

Small file read and write performance

Big Challenges for AI/ML Platform

- Large Number of Small Files
  - Tens of billions files: including images、videos、and text.
- Super large directory
  - Many datasets directory contain a large number of files (for example ImageNet contains 14 million images)
- Hot Spot Directory
  - The access to public data by multi-user parallel training tasks can easily make the data node a performance bottleneck and cannot make full use of the cluster performance.
- High performance
  - AI/ML training clusters require very high bandwidth and low latency to reduce job completion time.

Problems with existing storage systems

HDFS

- Weak extensibility of metadata
- Global locks lead to poor performance
- Not friendly to small files
- Poor tenant isolation and many other pain points

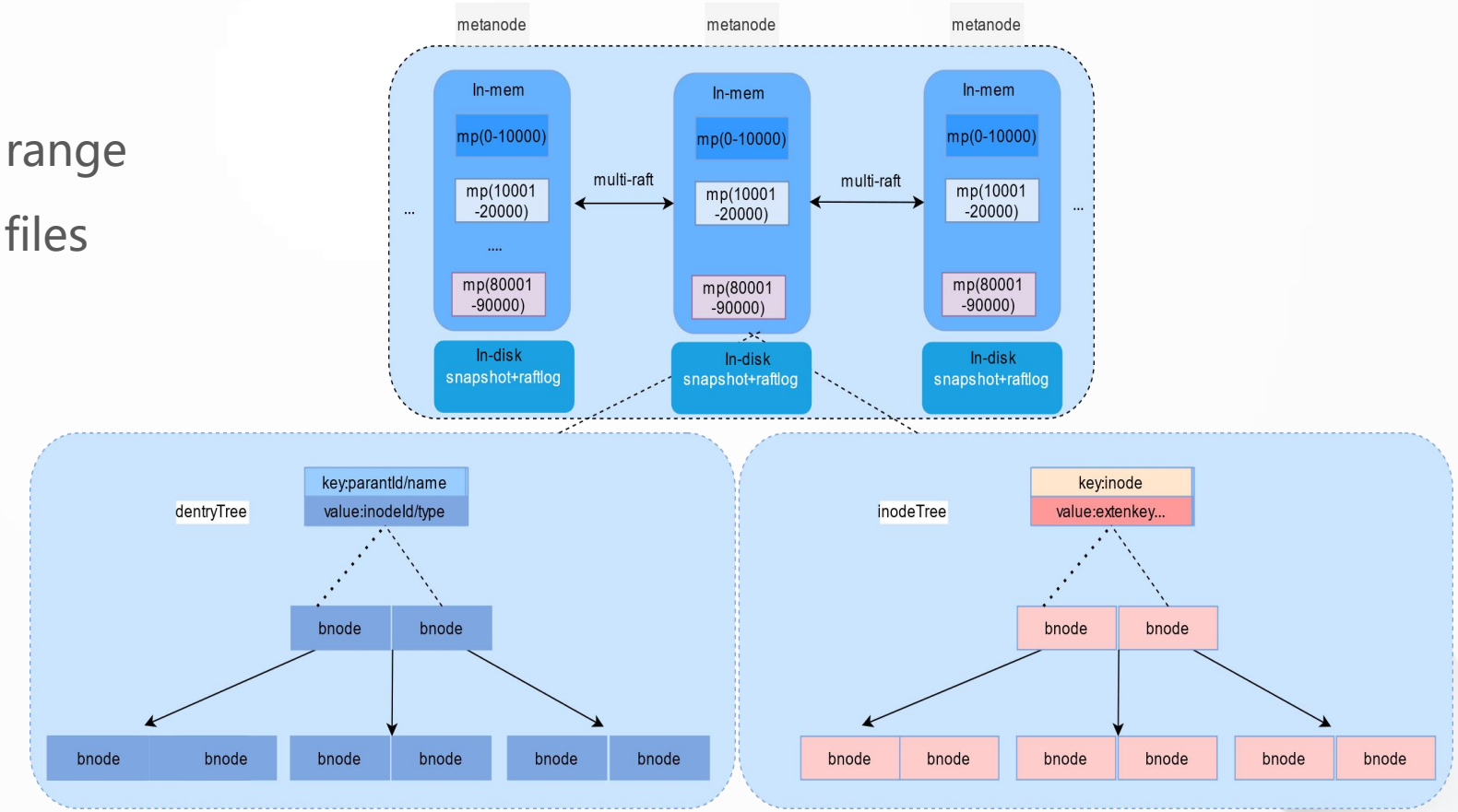
CephFS

- Poor stability caused by mds
- Weak performance on small files storage and random write
- High storage costs

CubeFS-Elasticity and scalability for metadata

Key tips

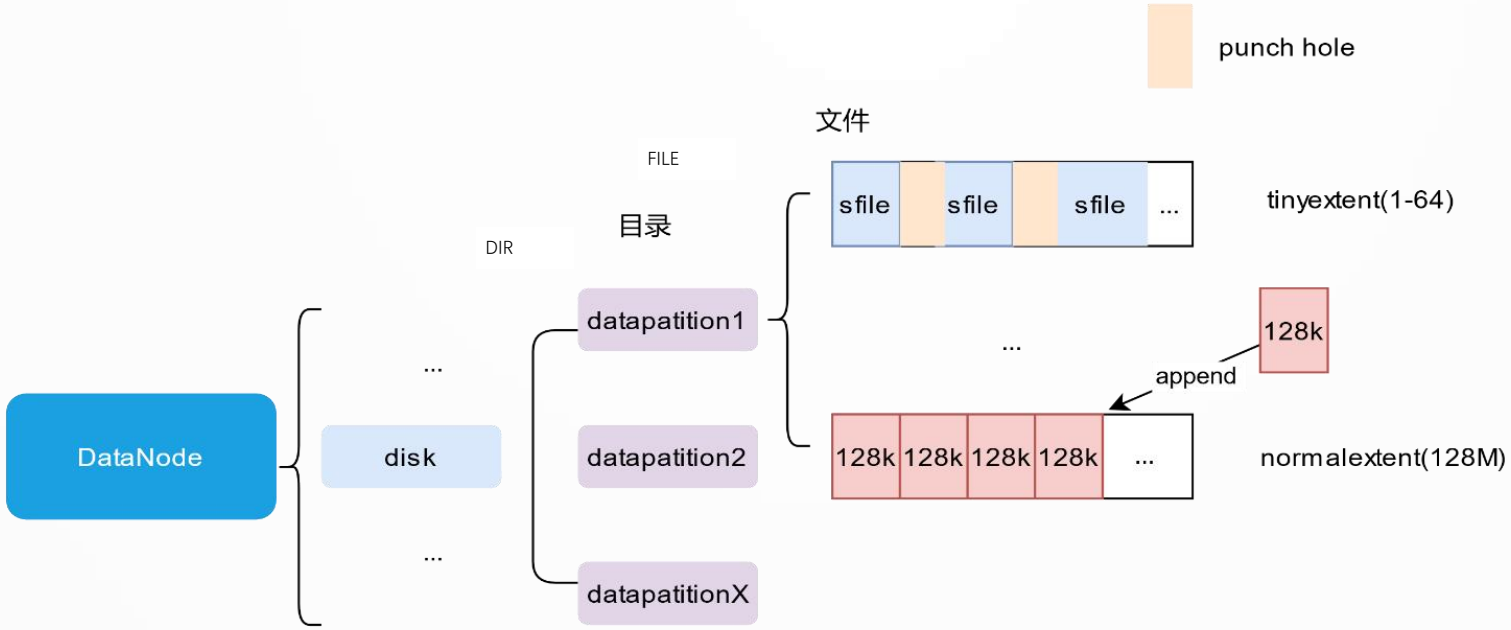
- All metadata cached in memory
- File’ s dentry and inode split by range
- Single directory: tens of millions files
- A single cluster supports tens of billions of files



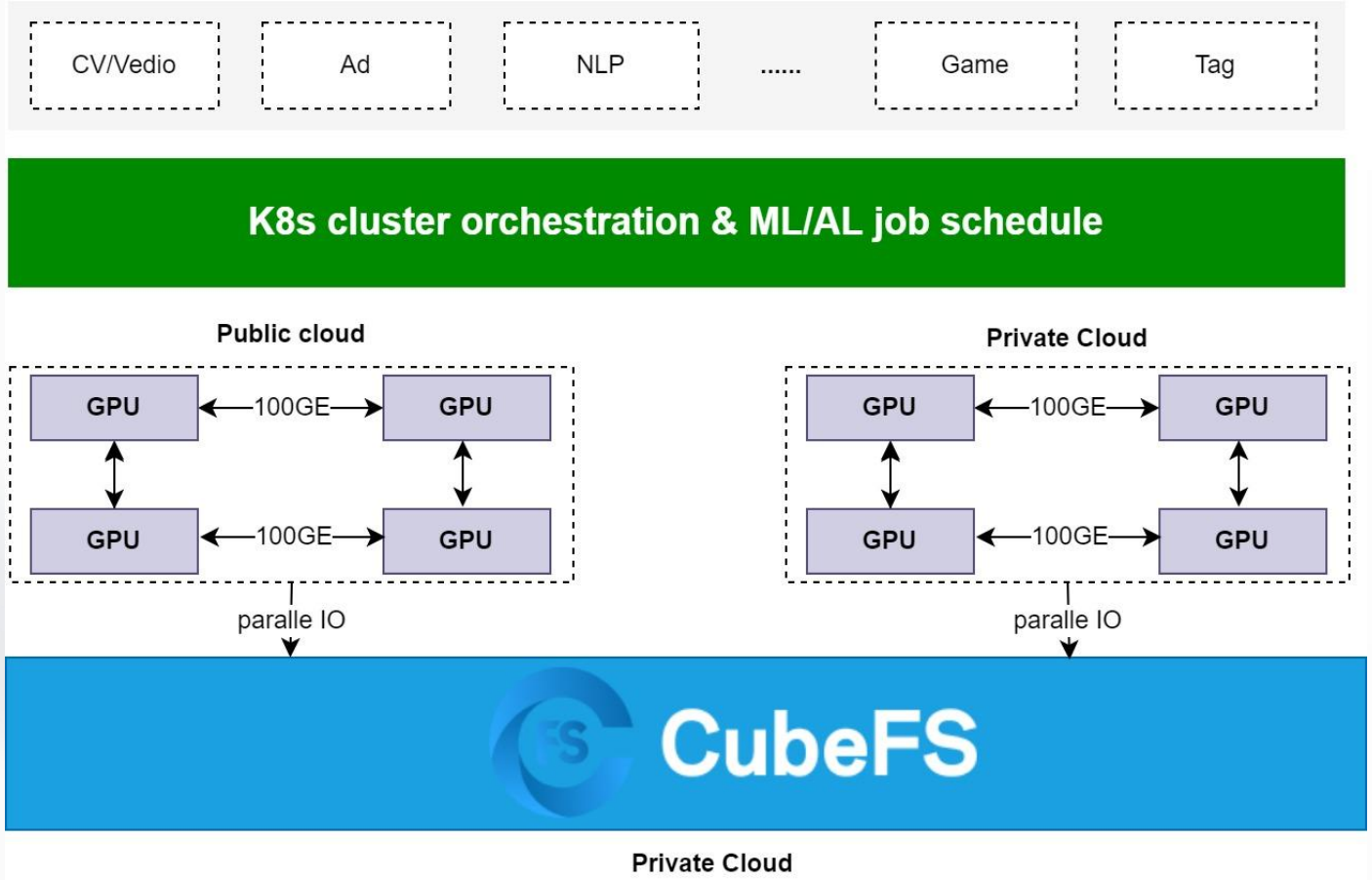
CubeFS-Optimized for small files

Key tips

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



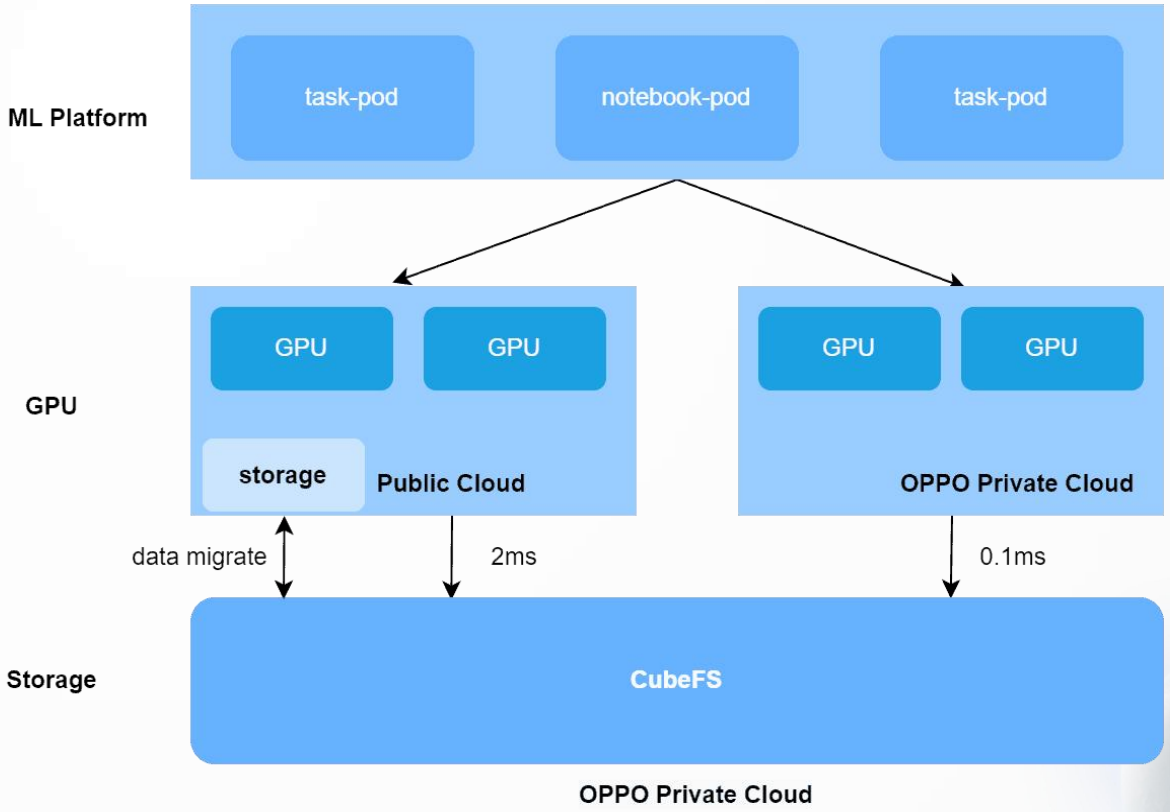
AI/ML Platform Unified storage based on CubeFS



AI/ML Platform solution:hybrid cloud acceration

Challenges

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



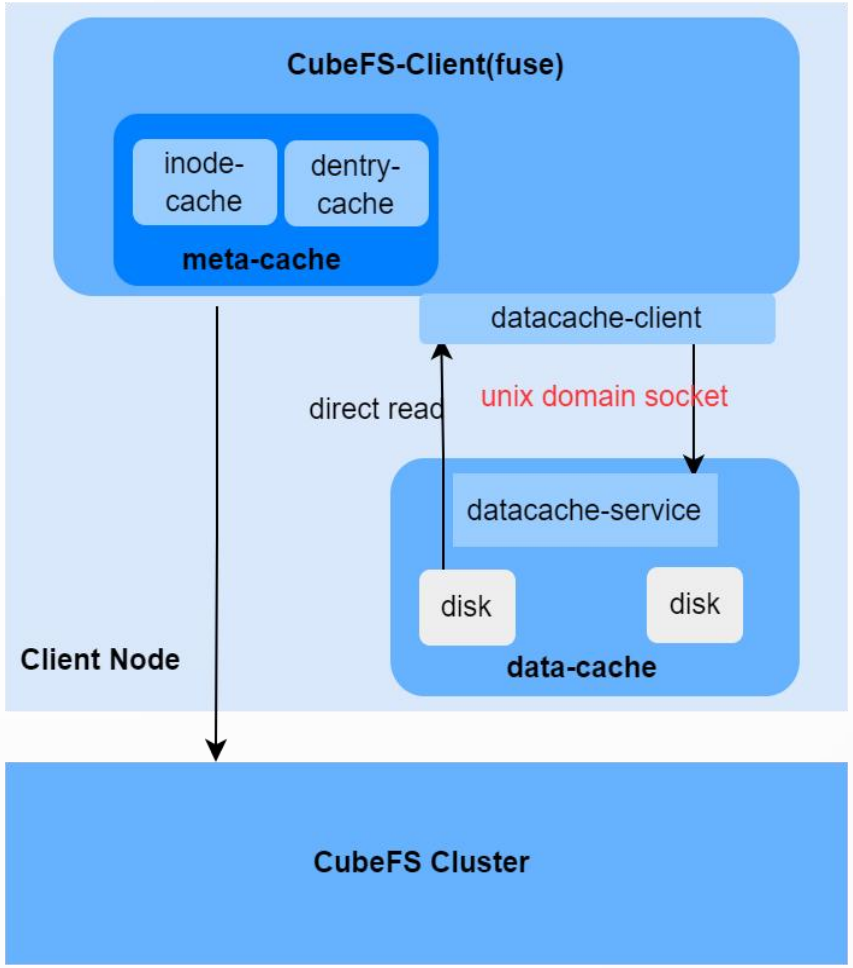
AI/ML Platform solution:hybrid cloud acceration

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



AI/ML Platform solution:hybrid cloud acceration

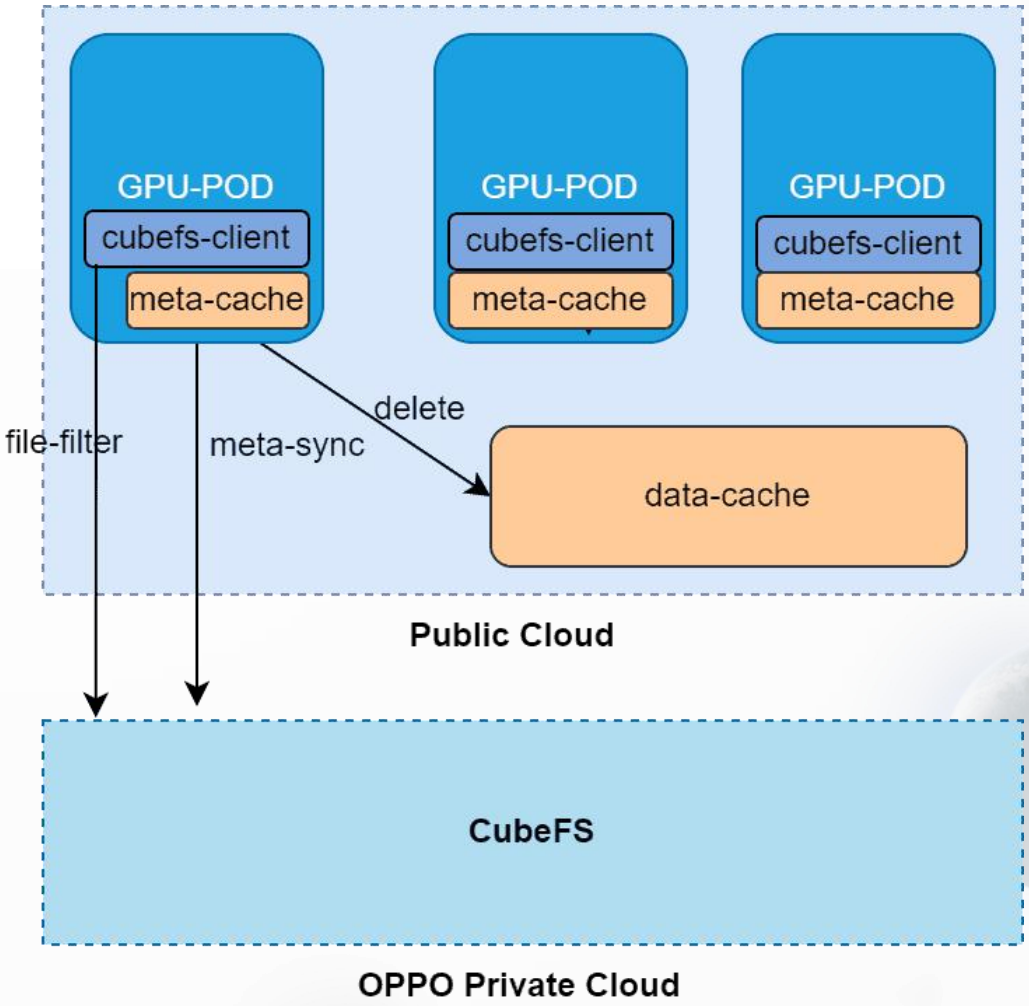
Cache consistent

Strong consistency

- The client is configured based on file extensions, filtering out file types that do not require access to the cache, such as for application program files and configuration files.

Eventual consistency

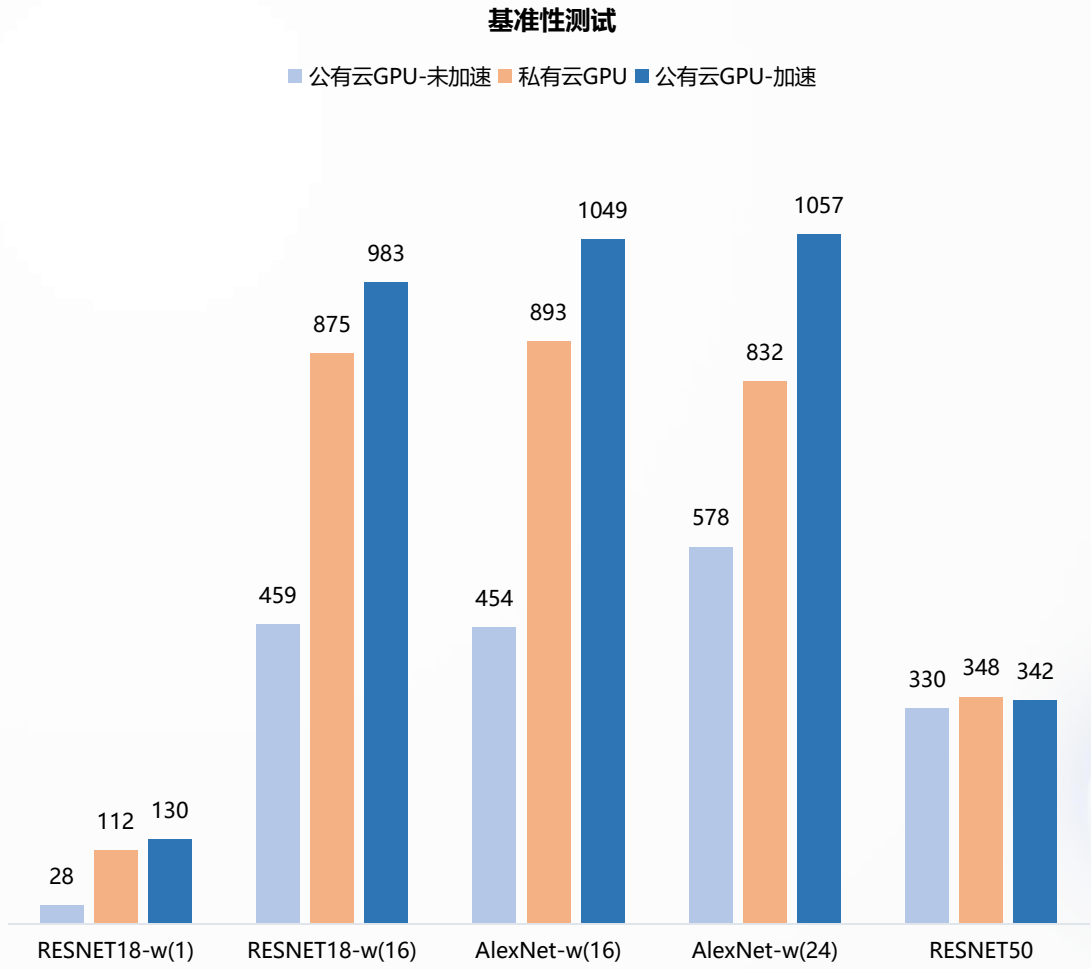
- CubeFS client starts the meta sync task, scans the metadata of all files in the cache, queries all updated cache data during the scan cycle, updates the metadata cache if the file is changed.



AI/ML Platform solution:hybrid cloud acceration

BenchMark

- RESNET18:performance improvements of 360% and 114% respectively with one and 16 Dataloader workers.
- AlexNet shows performance improvements of 130% and 80% respectively with 16 and 24 Dataloader workers.
- Compared to private cloud deployment, there is also a performance improvement of 12% to 27%.



AI/ML Platform solution: QoS

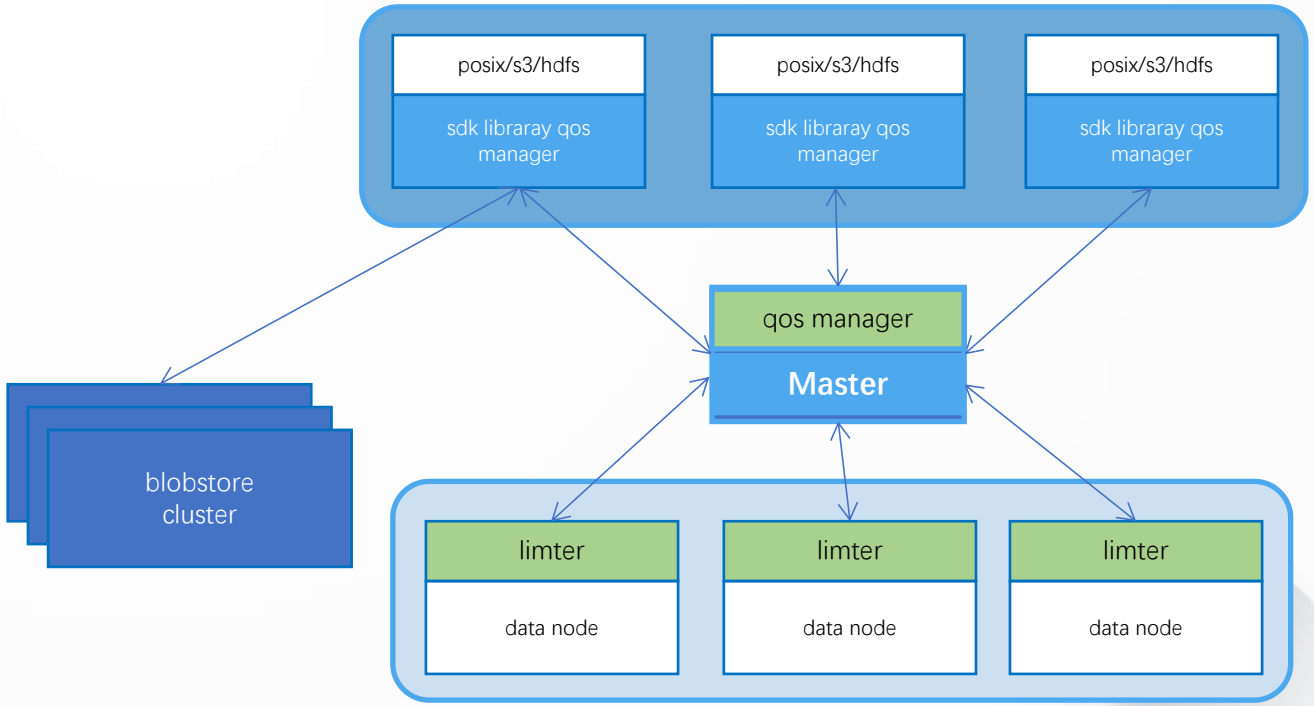
QoS flow control system

Background

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

Feature

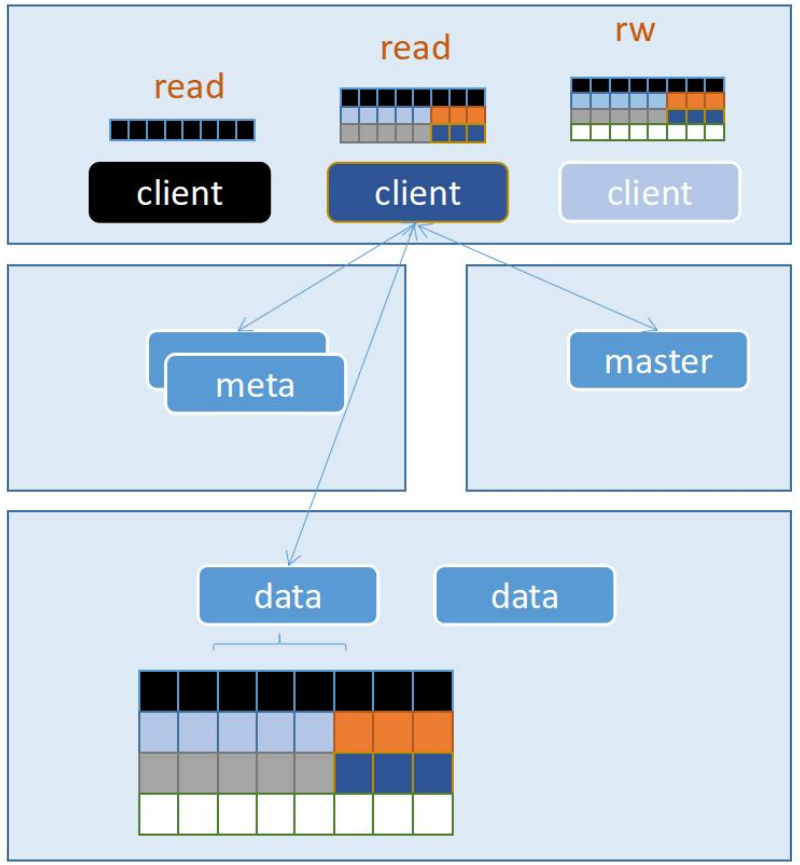
- Does not depend on external components
- Resource pre-allocation and dynamic adjustment
- Dynamic adjustment of request period

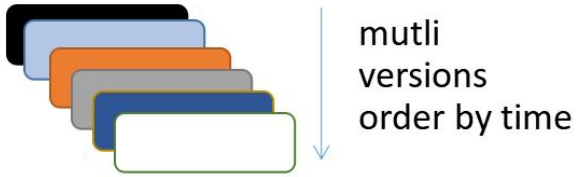


AI/ML Platform solution: Snapshot

Redirect-on-Write, ROW

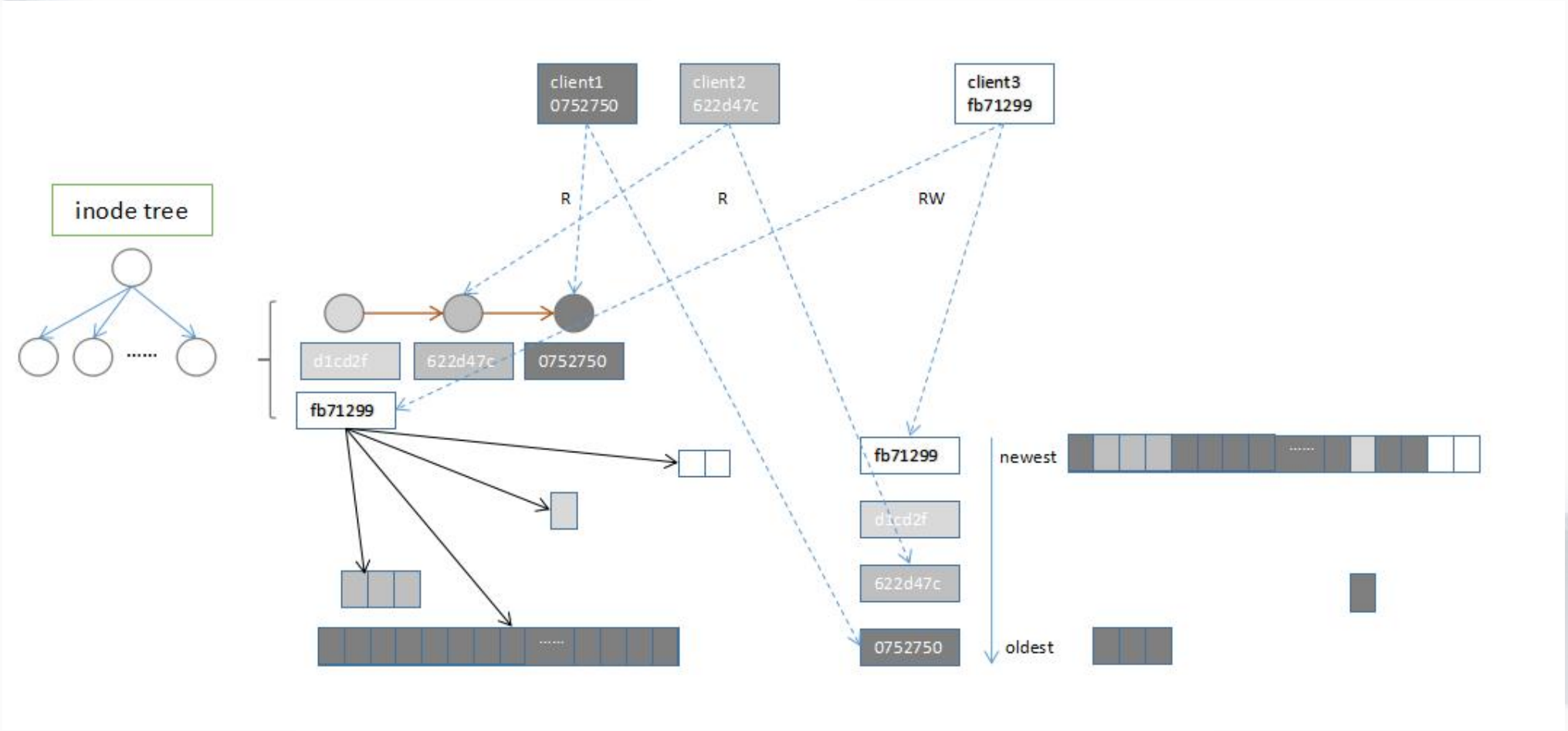
1. Create snapshots in seconds
2. No-lag snapshot version reads
3. No write amplification
4. Metadata, data without space redundancy
5. Strong consistency





AI/ML Platform solution: Snapshot

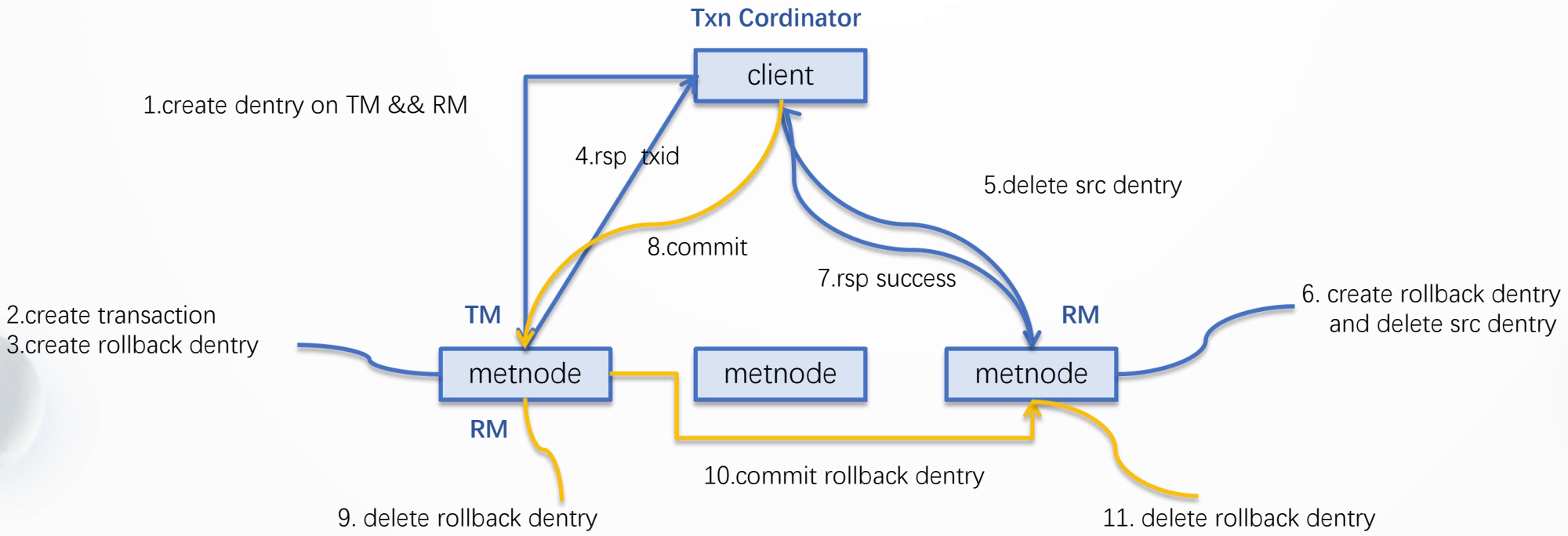
Snapshot multi version index



AI/ML Platform solution:POSIX Interface Atomicity

Rename

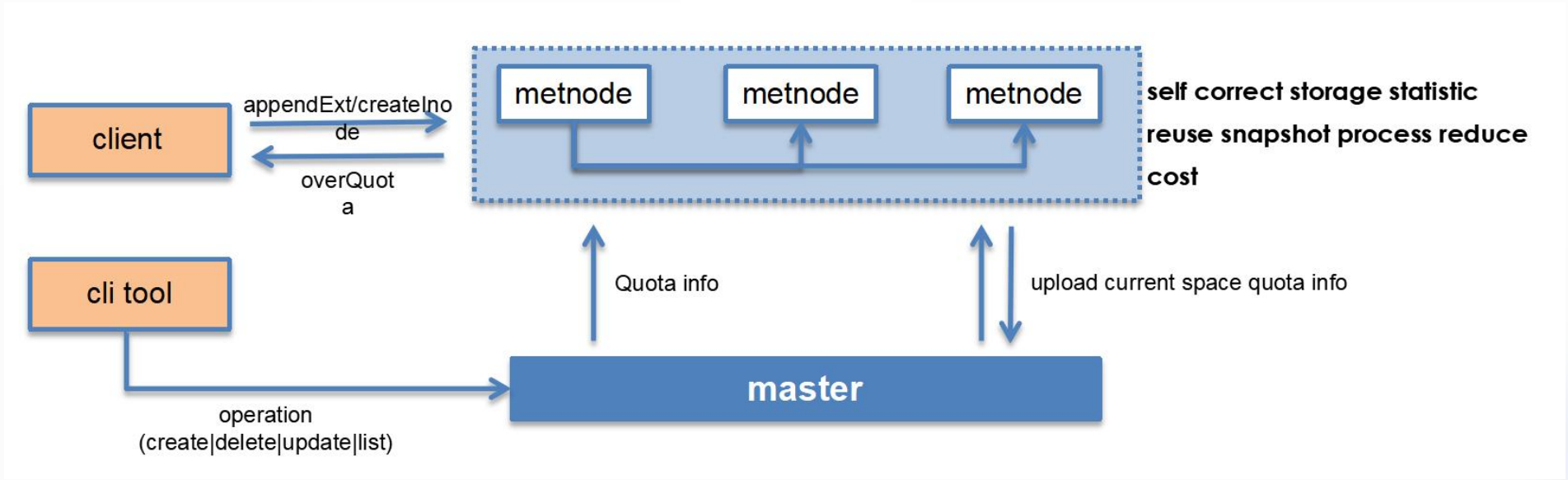
Txn Cordinator —— Client
Resource Manager(RM) —— MetaPartition
Txn Manager(TM) —— MetaPartition



AI/ML Platform solution:Quota management

1.Directory Quota Management

2.Uid Space Management



Community Development

- Broad Content Platform

  WeChat official account : 17 articles，reading volume of 6000+ (a MoM growth of 56%) and gained 420+ new followers.

- Developer Activities

  Organized the developer event && Participated in the "Summer of Code in Space"

- Developer Community

  WeChat community group：1000+ new users with 30+ new sub-groups established

- Ecological Collaboration

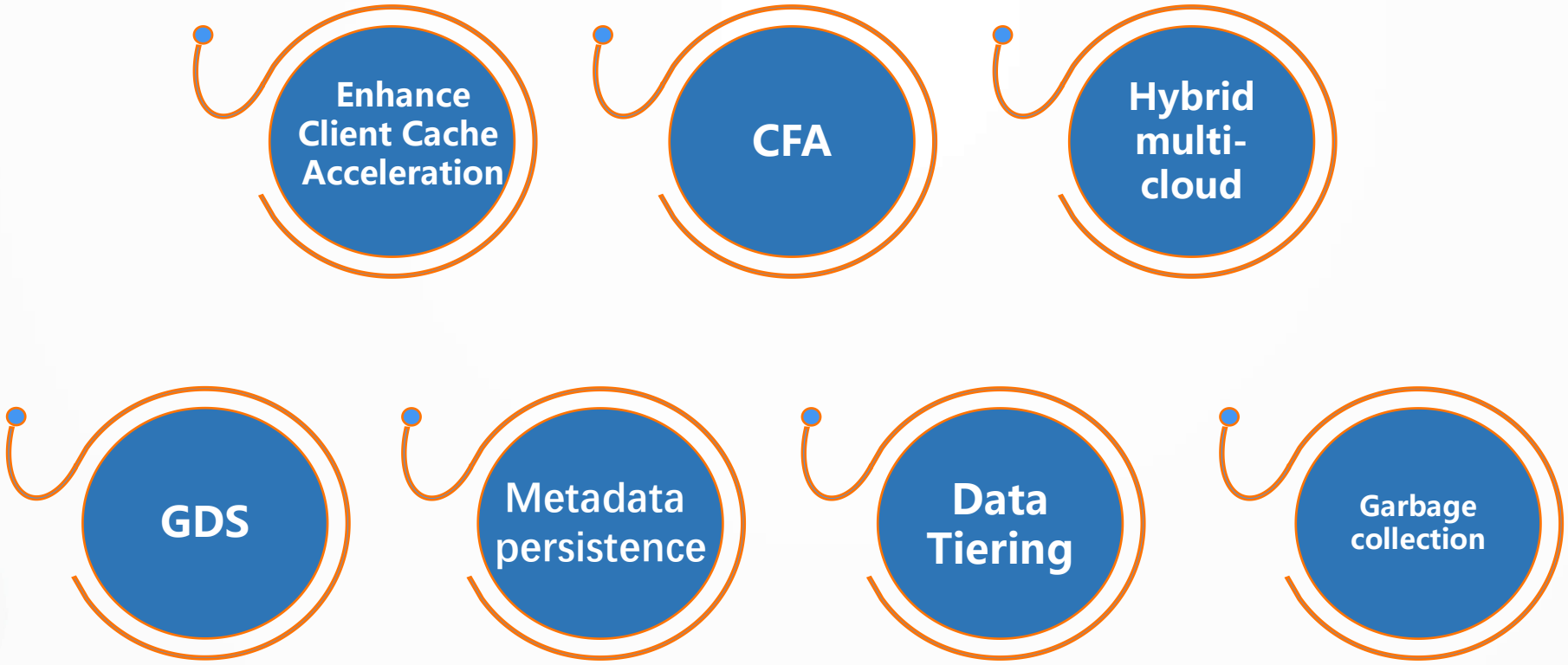
  Huazhong University of Science && Technology and University of Science and Technology of China

- Multi-cloud Deployment

  Aliyun 、AWS.

- Operator-based management of CubeFS is now supported and progressing as planned.

Future Planning



GOTC

THANKS

https://cubefs.io/

cubefs helper

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE