



KubeCon



CloudNativeCon



China 2024

CubeFS 助力AI效能提升

OPPO Chi.He

01

CubeFS项目介绍

02

在OPPO机器学习平台的应用实践

03

CubeFS未来展望

项目简介



China 2024

CubeFS是托管在云原生计算基金会(CNCF)的新一代云原生开源存储产品，具备完整的文件和对象存储能力，目前项目毕业工作正在收尾阶段。

官网地址: <https://cubefs.io/>



ArtifactHUB
Application Definition &
Image Build



Backstage
Application Definition &
Image Build



Buildpacks.io
Application Definition &
Image Build



CERT
MANAGER
Security & Compliance



Chaos Mesh
Chaos Engineering



Cloud Custodian
Automation &
Configuration



CNI
Cloud Native Network



CONTOUR
Service Proxy



cortex
Observability



Crossplane
Scheduling &
Orchestration



CubeFS
Cloud Native Storage



dapr
Application Definition &
Image Build



Dragonfly
Container Registry



EMISSARY
INGRESS
API Gateway



gRPC
Remote Procedure Call



in-toto
Security & Compliance



KARMADA
Scheduling &
Orchestration



keptn
Continuous Integration &
Delivery



KEYCLOAK
Security & Compliance



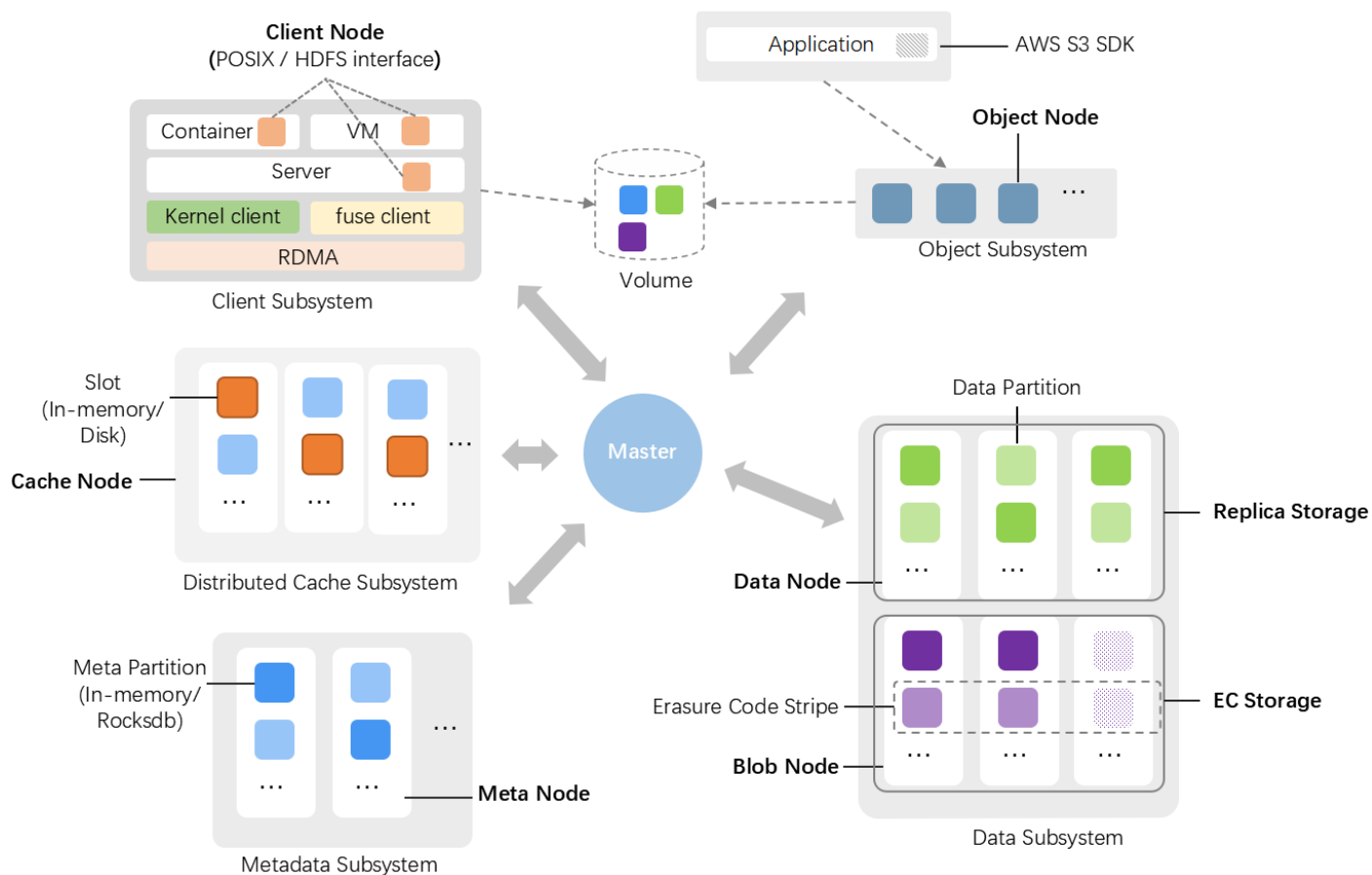
Knative
Scheduling &
Orchestration



总体架构



China 2024



Master: 资源管理子系统，用于管理集群中的资源元信息。

Volume: 卷，客户端可以从容器访问数据的文件系统。

Data Partition: 数据分区，文件数据分片的最小管理单位。

Replica Subsystem: 副本子系统，管理集群中的数据分区。

Erasure Code Subsystem: 纠删码子系统，管理集群中纠删码条带。

Meta Partition: 元数据分区，文件元信息的最小管理单位。

Metadata Subsystem: 元数据子系统，管理集群中的元数据分区。

Object Subsystem: 对象网关，兼容标准s3语义的对象网关。

Client: 客户端子系统，提供挂载文件系统的访问接口。



元数据子系统



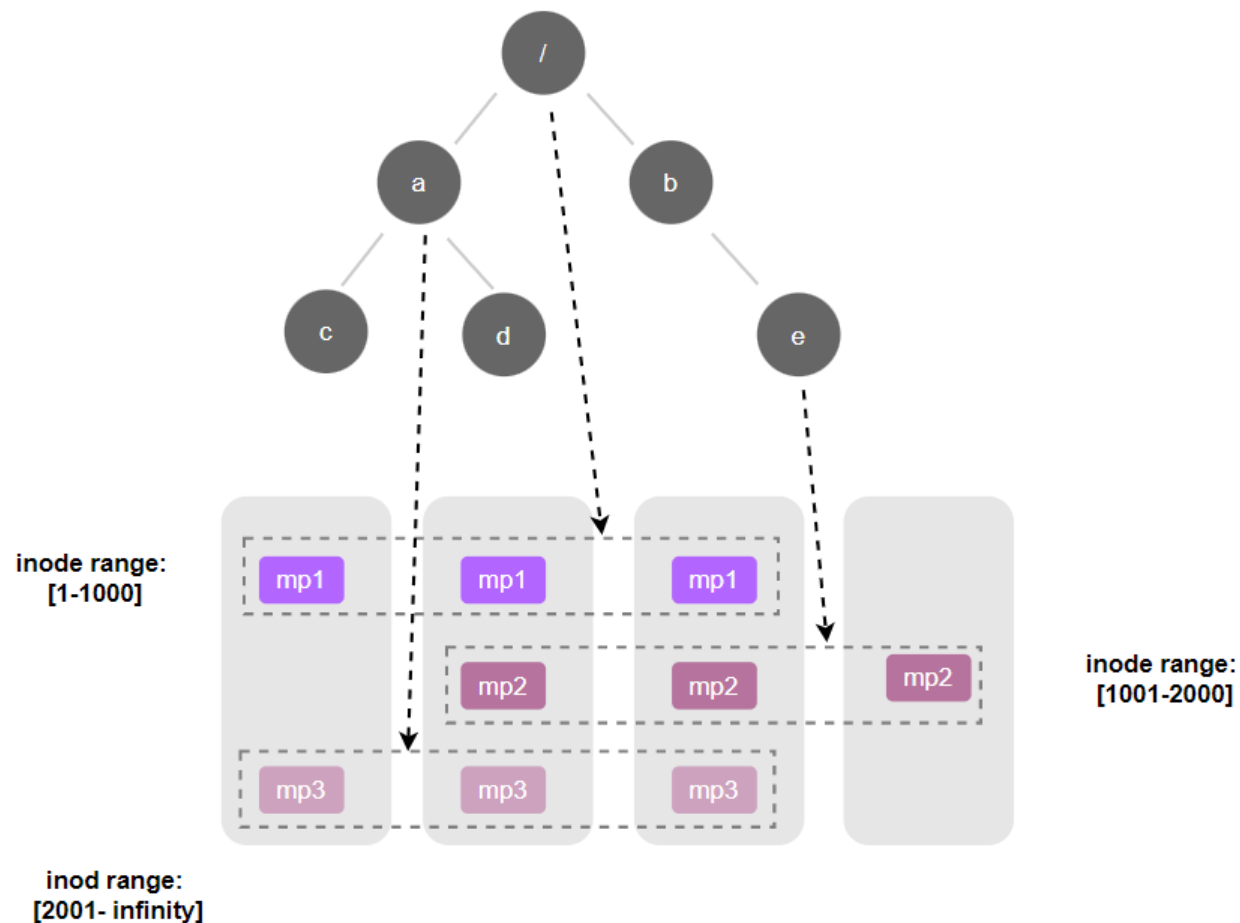
China 2024

元数据分区分裂:通过将拆分元数据分区的管理范围,实现动态扩容;不会触发数据迁移任务。

全内存缓存策略: 提高元数据的访问速度。

Multi-raft: 保证数据强一致性以及高可用。

定期快照: 元数据以分区为单位定期持久化到磁盘,用于备份和恢复。



副本子系统



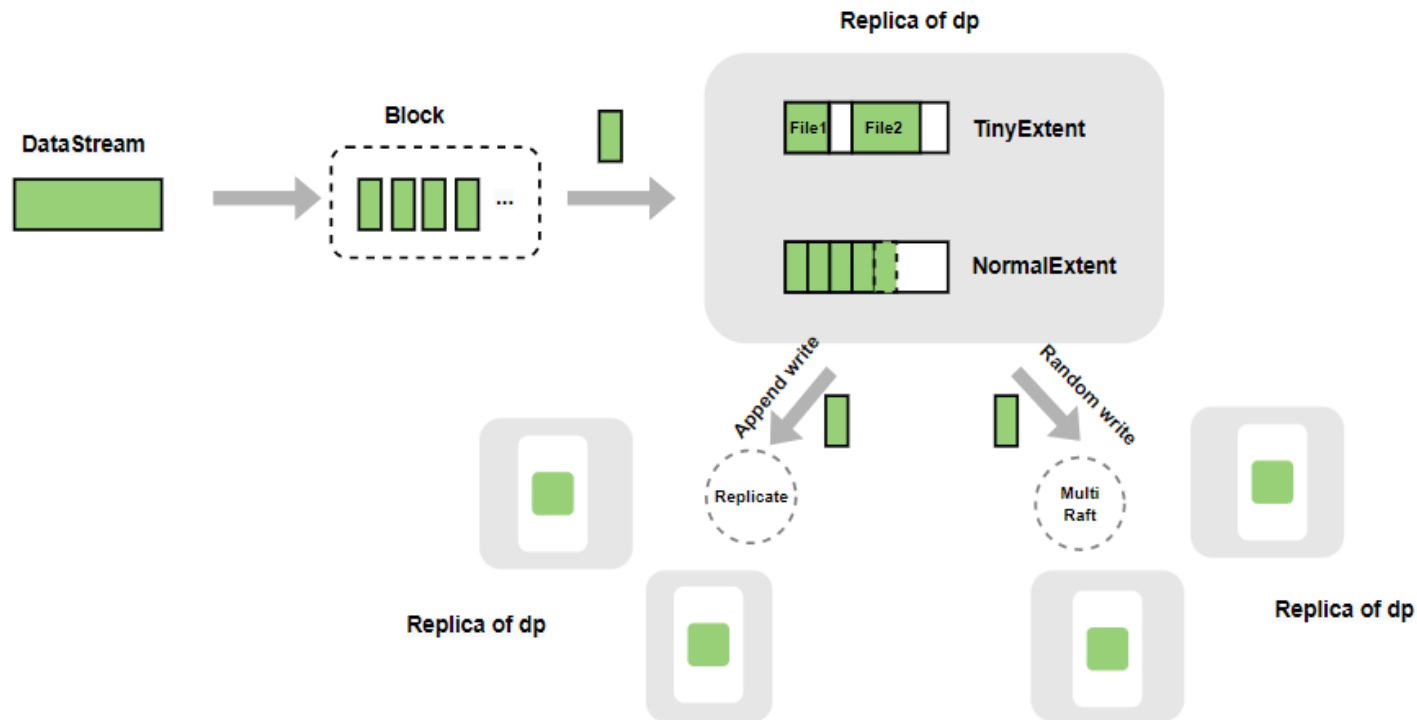
China 2024

大/小文件存储优化: 通过分片/聚合的方式存储; 提前分配好TinyExtent, 降低网络开销。

场景感知复制: 副本之间根据不同的写入方式采取不同的复制策略, 提高复制效率。

坏盘自动迁移: 下线过程具备原子性, 不需要人工干预。

异常副本自愈: 自动修复异常副本, 保证数据高可靠性。



纠删码子系统



KubeCon



CloudNativeCon



China 2024

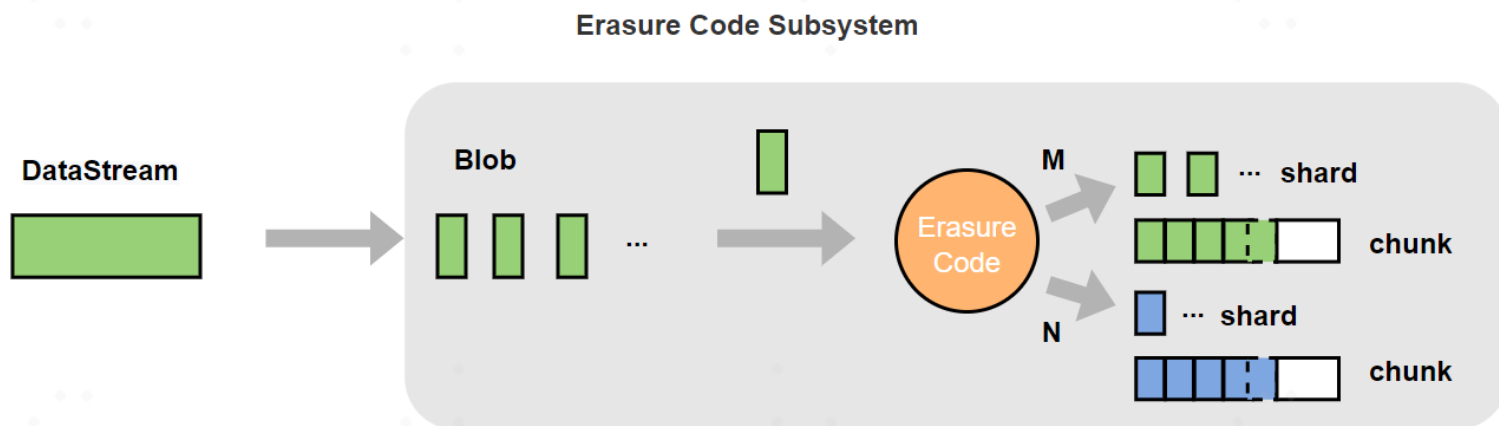


AI_dev

Quorum机制: 允许一定写入失败，有效解决拖尾时延问题。

多AZ部署: 持1,2,3AZ部署，支持AZ级别容灾。

数据巡检: 保证数据高可用性。



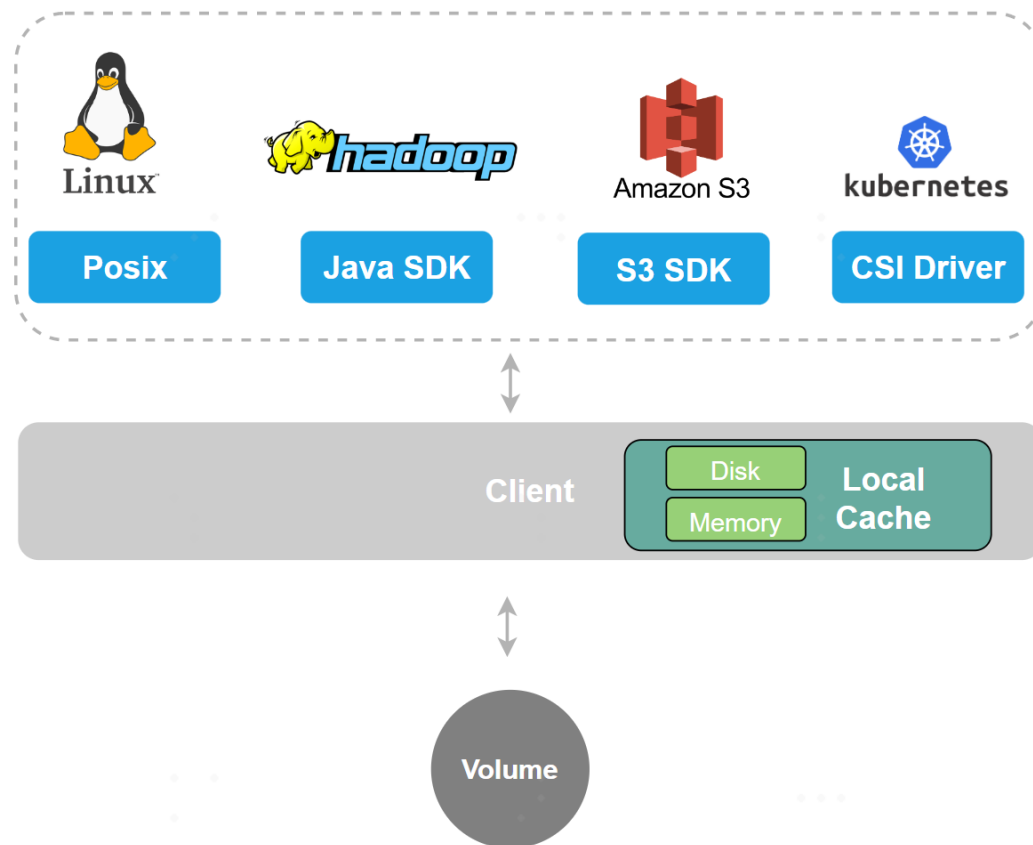
客户端子系统



China 2024

多协议互通: 支持Posix, HDFS等应对不同的业务场景,提高数据利用率。

数据和元数据缓存: 提高数据读取效率。



特性小结



China 2024



多协议

支持Posix、S3、HDFS等多种协议，共享业务数据。



双引擎

根据业务需求，灵活选择多副本或者纠删码存储引擎。



易扩展

元数据和数据支持水平扩展，轻松构建PB或者EB级别存储。



高性能

元数据全内存缓存，客户端本地缓存加速访问效率。



多租户

多租户管理，隔离用户数据和资源。



云原生

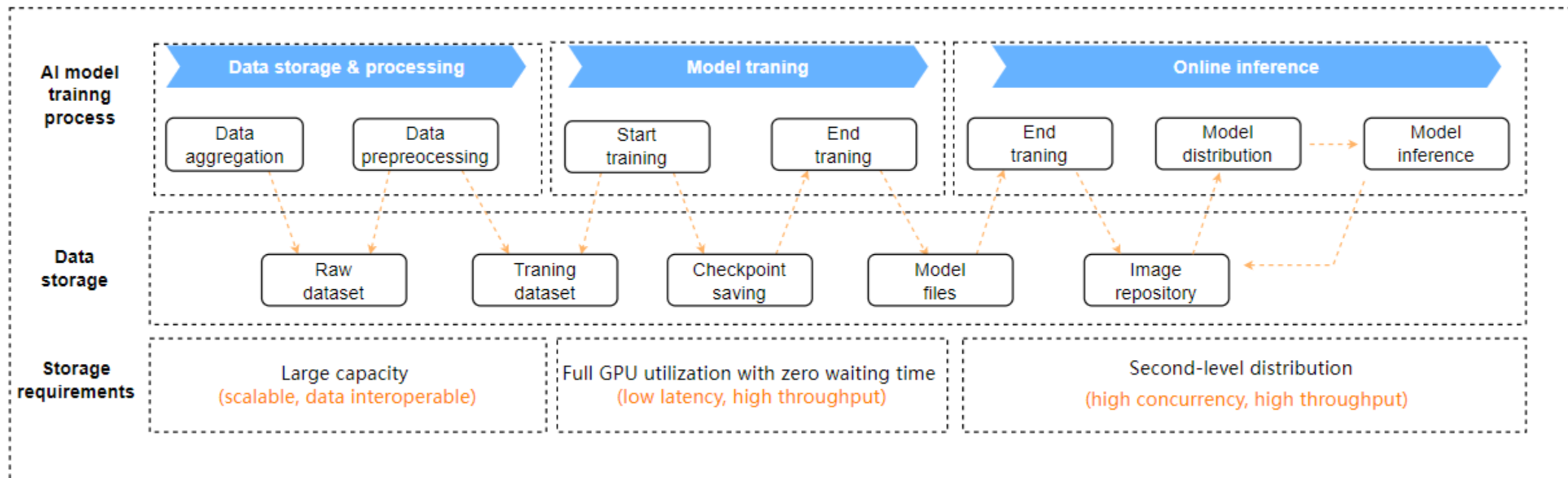
基于CSI插件速度在K8S上使用CubeFS。



AI模型训练链路



China 2024



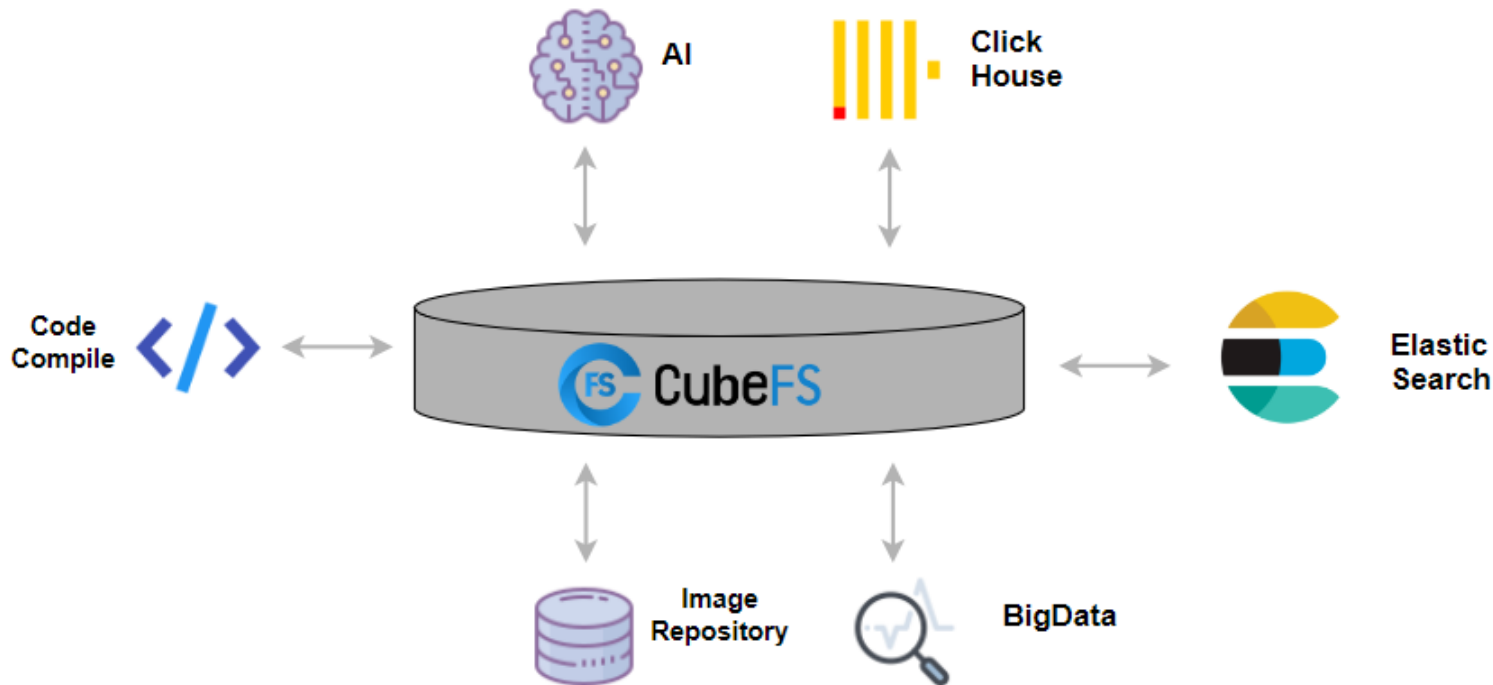
数据互通



China 2024

统一存储底座：不同业务系统的数据的第一个落脚点。

数据互通：多协议共享一套数据，提高数据流转效率。



智能生命周期管理

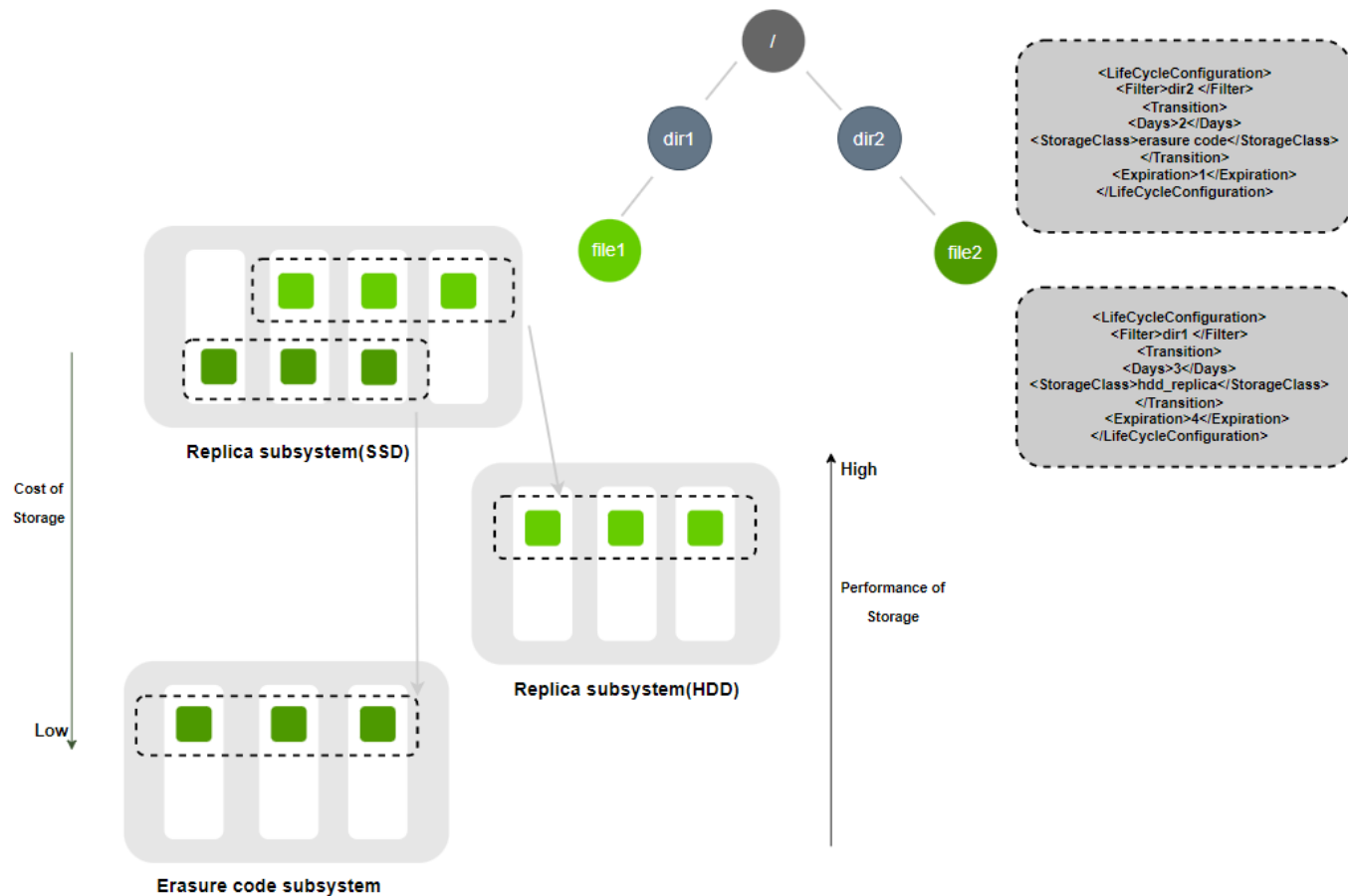


China 2024

平衡计算性能和存储成本: 热数据保留在性能更高的存储介质，冷数据转移到存储成本更低的介质。

基于目录的生命周期策略配置: 操作简单，降冷过程无需人工干预，节省运维人力成本。

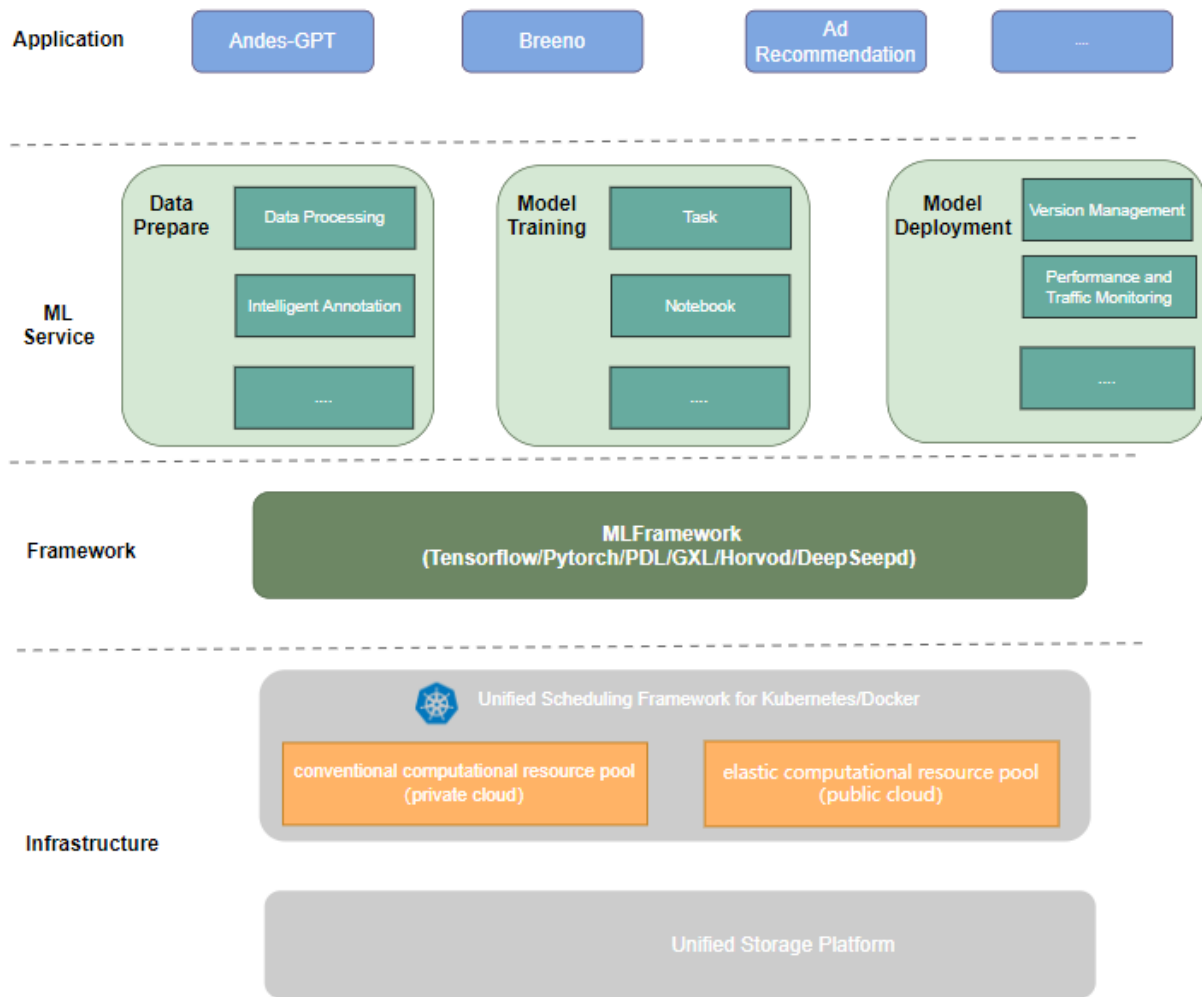
基于租约的降冷策略: 降冷过程不影响业务对存储的访问。



StarFire架构



China 2024



可用性

元数据子系统采用三副本的策略, 可用性从99.9% 提升至 99.99%。

运维成本

简洁的架构设计, 运维成本大幅度降低, 扩容更简便。

元数据性能

全内存的元数据策略, 平均时延缩短至1ms。

StarFire is OPPO's self-developed end-to-end machine learning service platform.

混合云的挑战

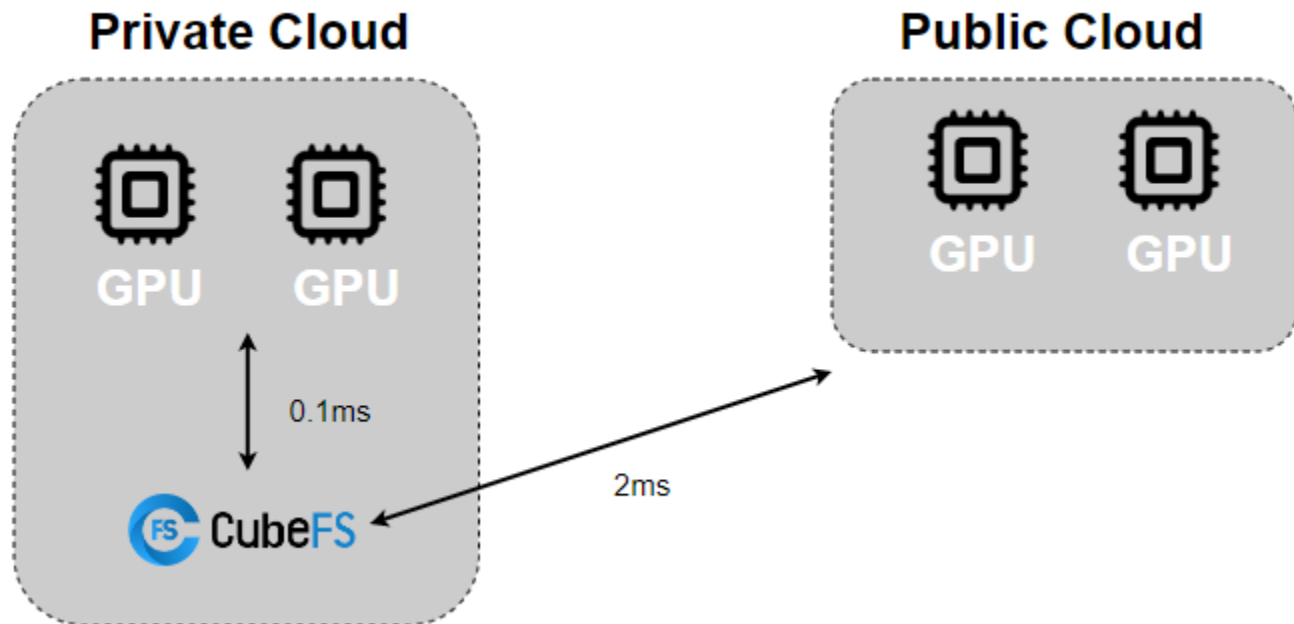


China 2024

存储访问性能下降：公有云GPU资源与私有云的存储资源网络延时在2ms左右，导致GPU利用率低，影响AI训练效率。

存储相对更难弹性化：迁移成本高，公有云和私有云的数据一致性无法保证。

数据隐私安全：将数据保存在公有云有泄露风险。



缓存加速方案



China 2024

Shuffle训练数据: 涉及readdir元数据操作。

读取训练数据: 涉及open/close元数据操作和read数据操作。

训练特点: 单机/多机训练都是反复对同一批数据执行epoch。

元数据/数据缓存: 充分利用计算节点的空闲内存和磁盘，提升训练效率。

```
// Training iterates through multiple epochs.
for each in epoch {
    List files in the dataset and shuffle

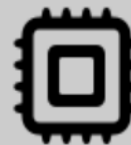
    // Training by reading data in batches
    for each in batch {
        Read files of the batch
        Training
    }

    //Periodically saving checkpoints for fault recovery
    Checkpoint if necessary
}
```

Private Cloud



Public Cloud



GPU



GPU

CubeFS-Client

Metadata Cache

Data Cache

Read from private cloud
during the first epoch



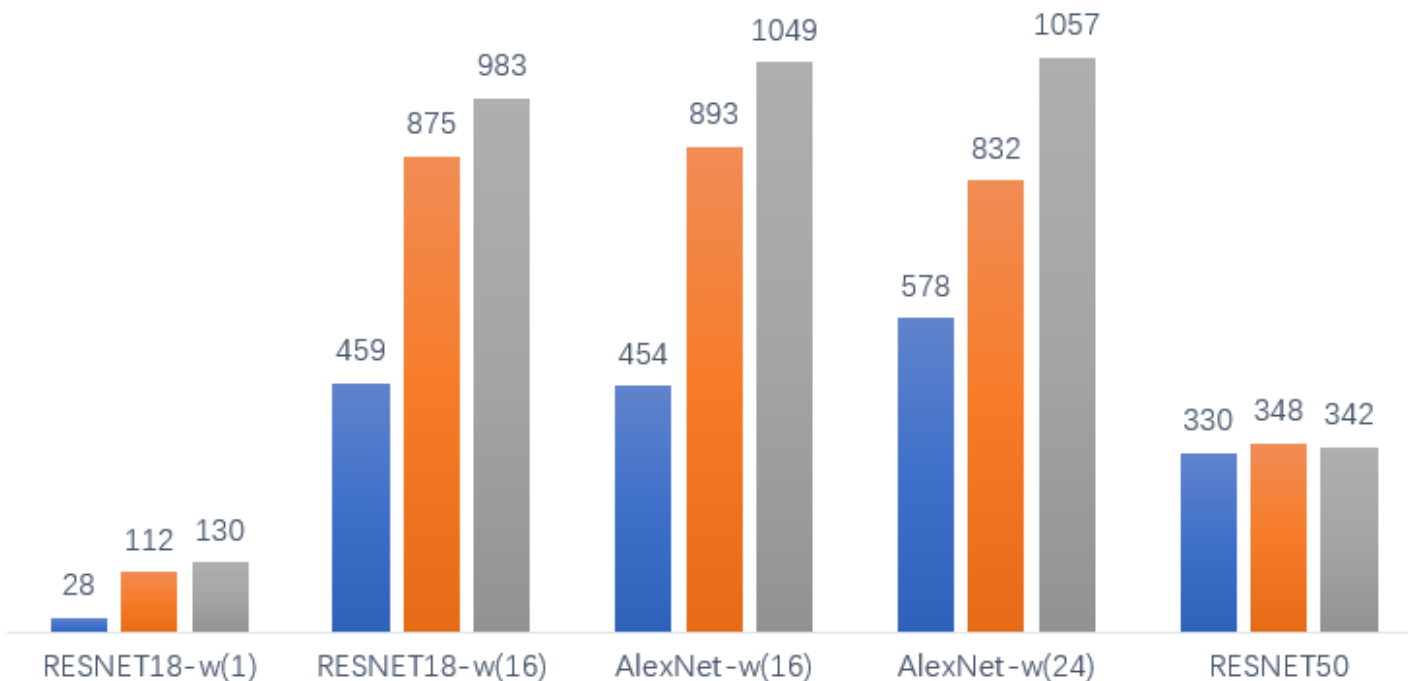
缓存加速收益



China 2024

Baseline Test

■ Public cloud GPU - Unaccelerated ■ Private cloud GPU ■ Public cloud GPU - Accelerated



RESNET18在Dataloader worker=1、16时，性能分别提升**360%**、**114%**。

AlexNet在Dataloader worker=16、24时，性能分别提升**130%**、**80%**。

相比私有云部署，性能也有**12%-27%**的提升。

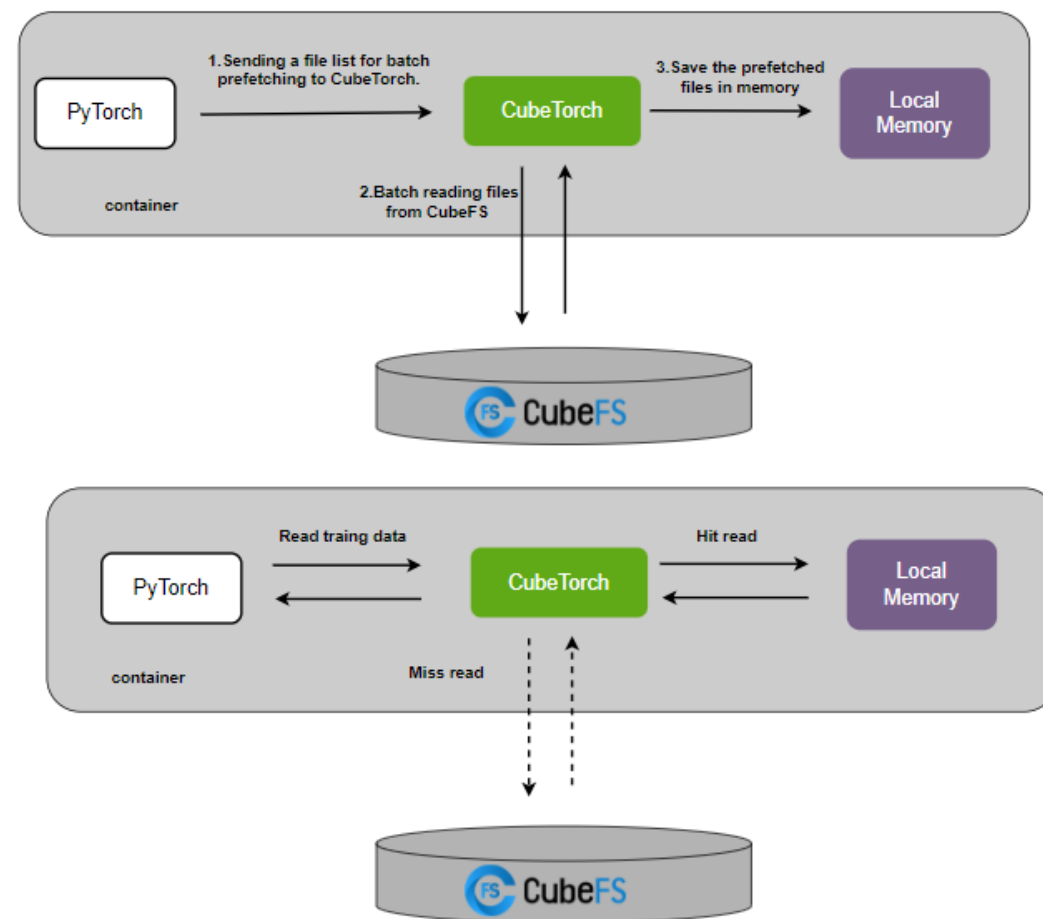
CubeTorch



China 2024

DataLoader: 让训练数据的读取过程和模型的训练过程并行起来，从而提升GPU的训练效率。

训练数据预加载:通过批量下载提前将下一批次的训练数据加载到内存；绕过内核。



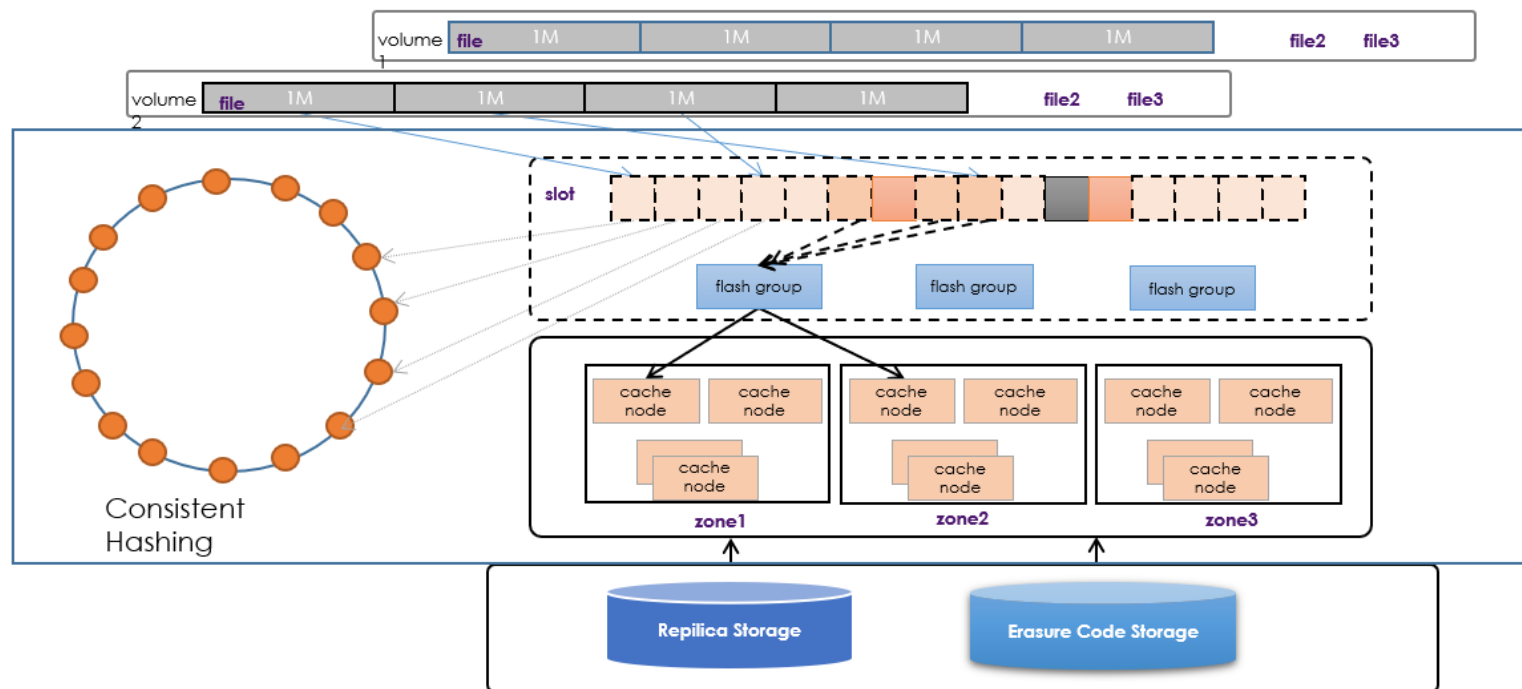
分布式缓存

距离感知: 计算节点选择网络延时最低的缓存节点进行访问。

动态扩容: 为用户提供可弹性扩展的吞吐能力。

弹性副本: 根据业务需求配置，均衡热点数据的访问请求。

多级缓存: 和计算节点的本地缓存组成多级缓存，进一步提高访问性能。



RDMA



China 2024

共享内存环: 数据拷贝到sendbuf中, 再通过RDMA write直接高效的写入到recvBuf中。

更高的数据写入效率: 数据的传输过程绕过内核和协议层之间的数据拷贝, 全程不需要CPU接入。



模型分发

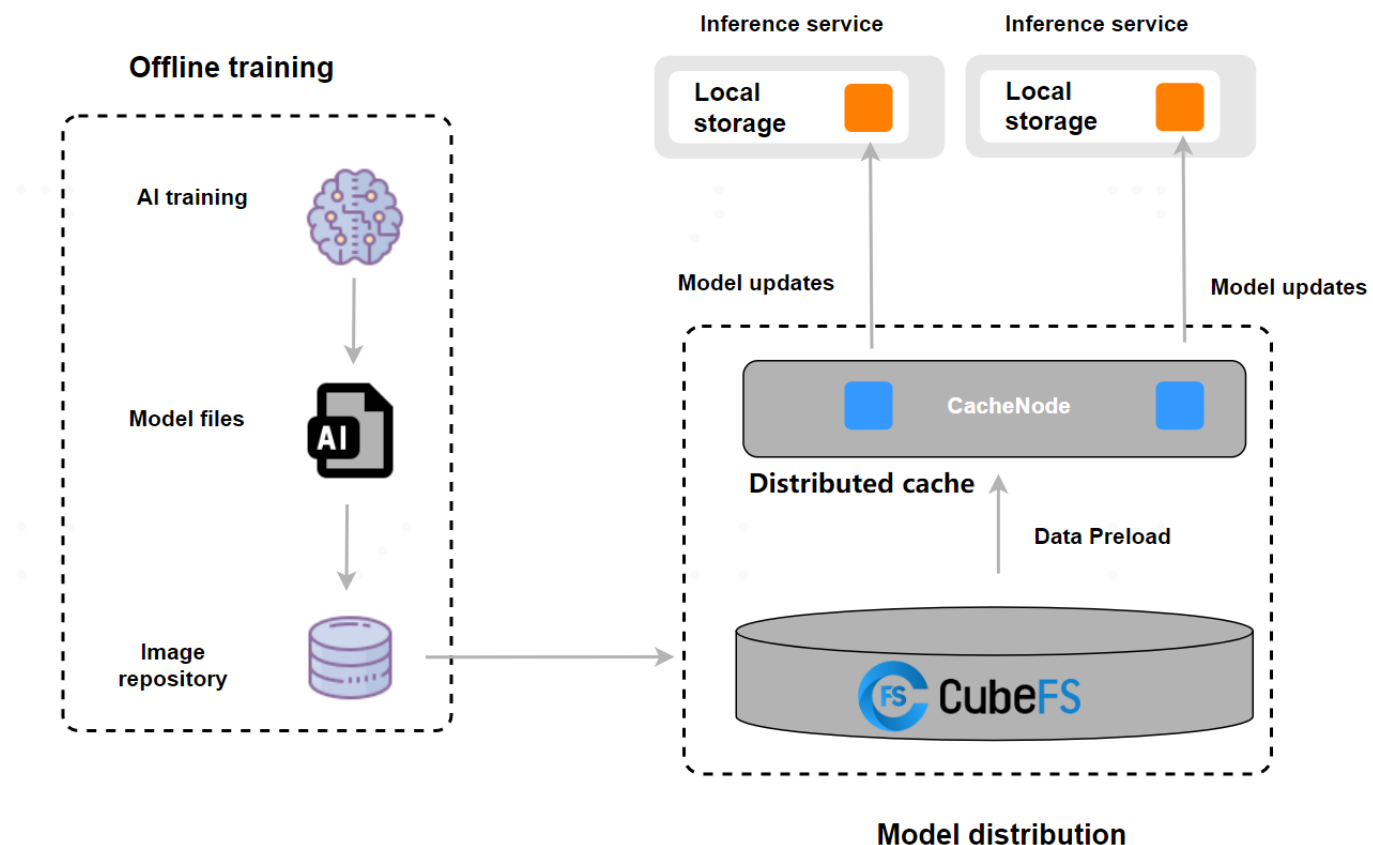


China 2024

数据预热: 将模型文件提前加载到CacheNode。

区域感知: 就近读取缓存数据。

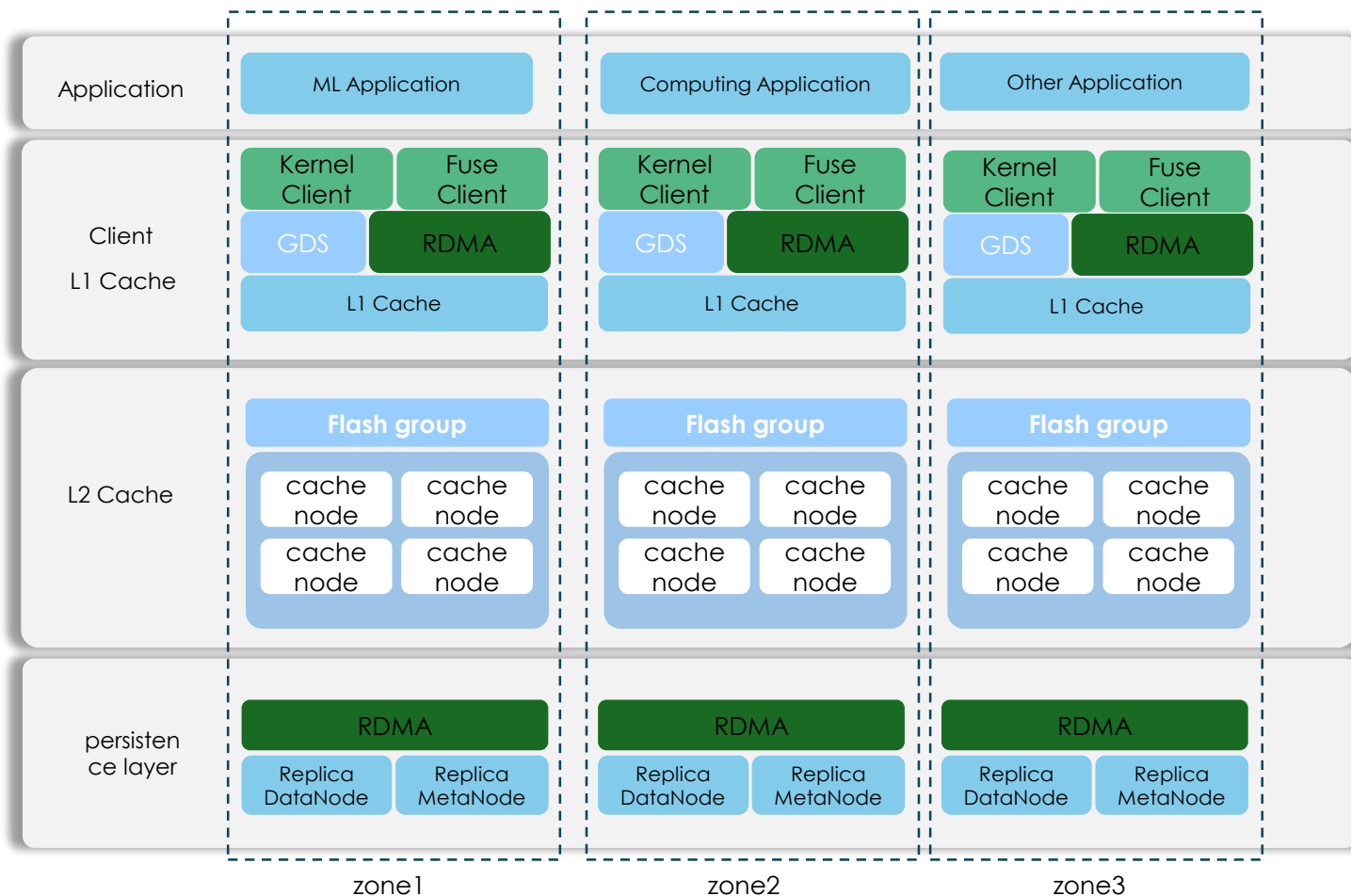
弹性副本: 提高模型文件的吞吐能力。



未来展望



China 2024



GDS

客户端GPU直接访问后端存储。



内核客户端

构建内核文件系统。



Ecosystem Engagement



Prometheus



Kubernetes



Vitess



Helm



Harbor



Elastic Search



ClickHouse

Contributors and Users

Contributors as well as users:



Users:





KubeCon



CloudNativeCon

THE LINUX FOUNDATION



AI_dev
Open Source GenAI & ML Summit

China 2024
