

SelectDB Cloud 与 Apache Doris 存算分离的思考与实践

杨勇强

飞轮科技 技术副总裁

Apache Doris PMC 成员

个人介绍



杨勇强

Apache Doris PMC Member

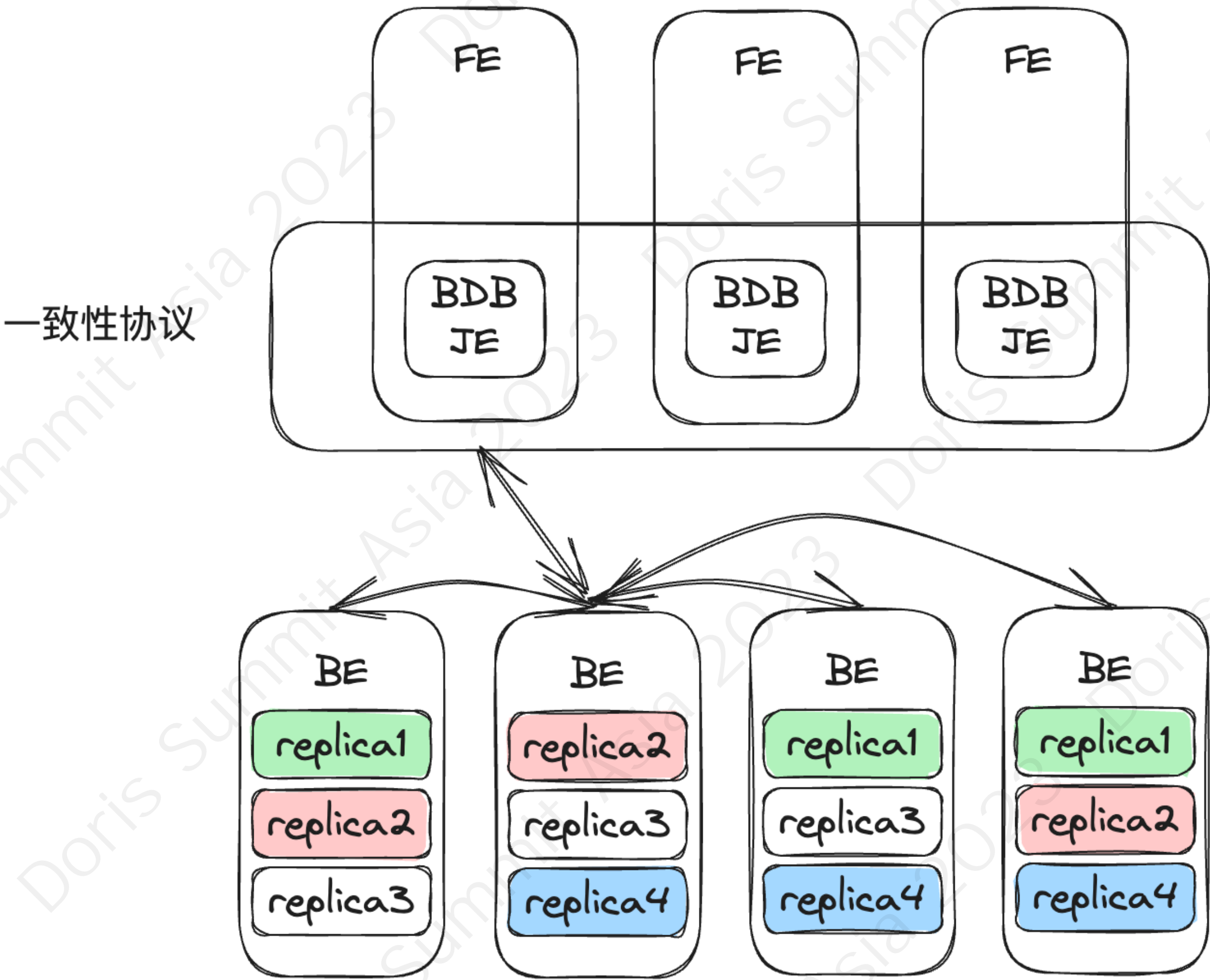
- SelectDB 技术副总裁，主要负责云原生版本产品研发
- 曾就职于百度智能云存储部主任架构师（T9）
- 具有 10 年云存储产品与架构经验
- 本科毕业于南开大学 计算机科学与技术专业，硕士毕业于中国科学院计算技术研究所

目录

1. 存算一体 vs 存算分离
2. Apache Doris 存算分离架构实现方案
3. 存算分离功能示例
4. 存算分离和存算一体适用场景

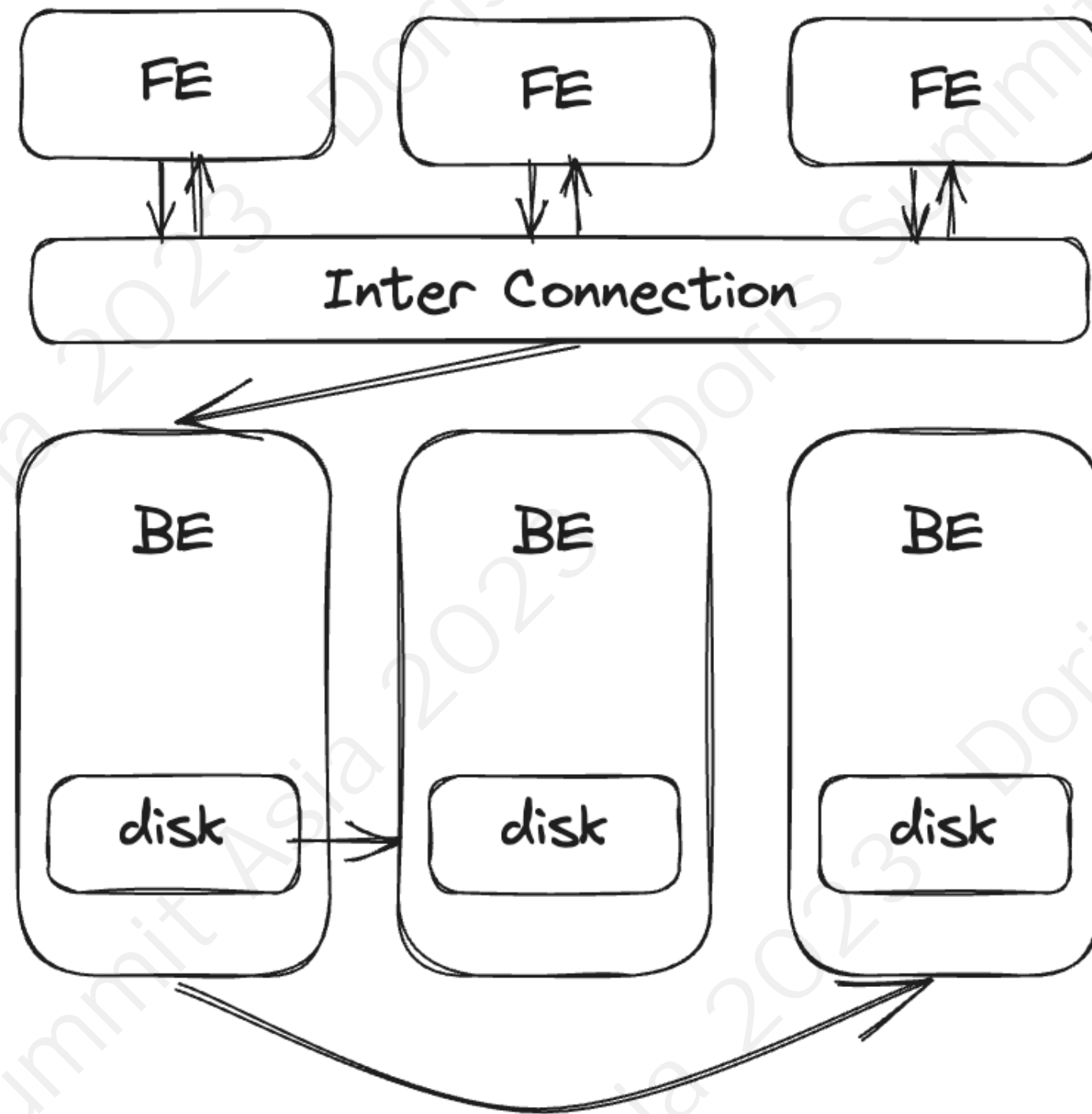
1 存算一体 vs 存算分离

Apache Doris 系统架构

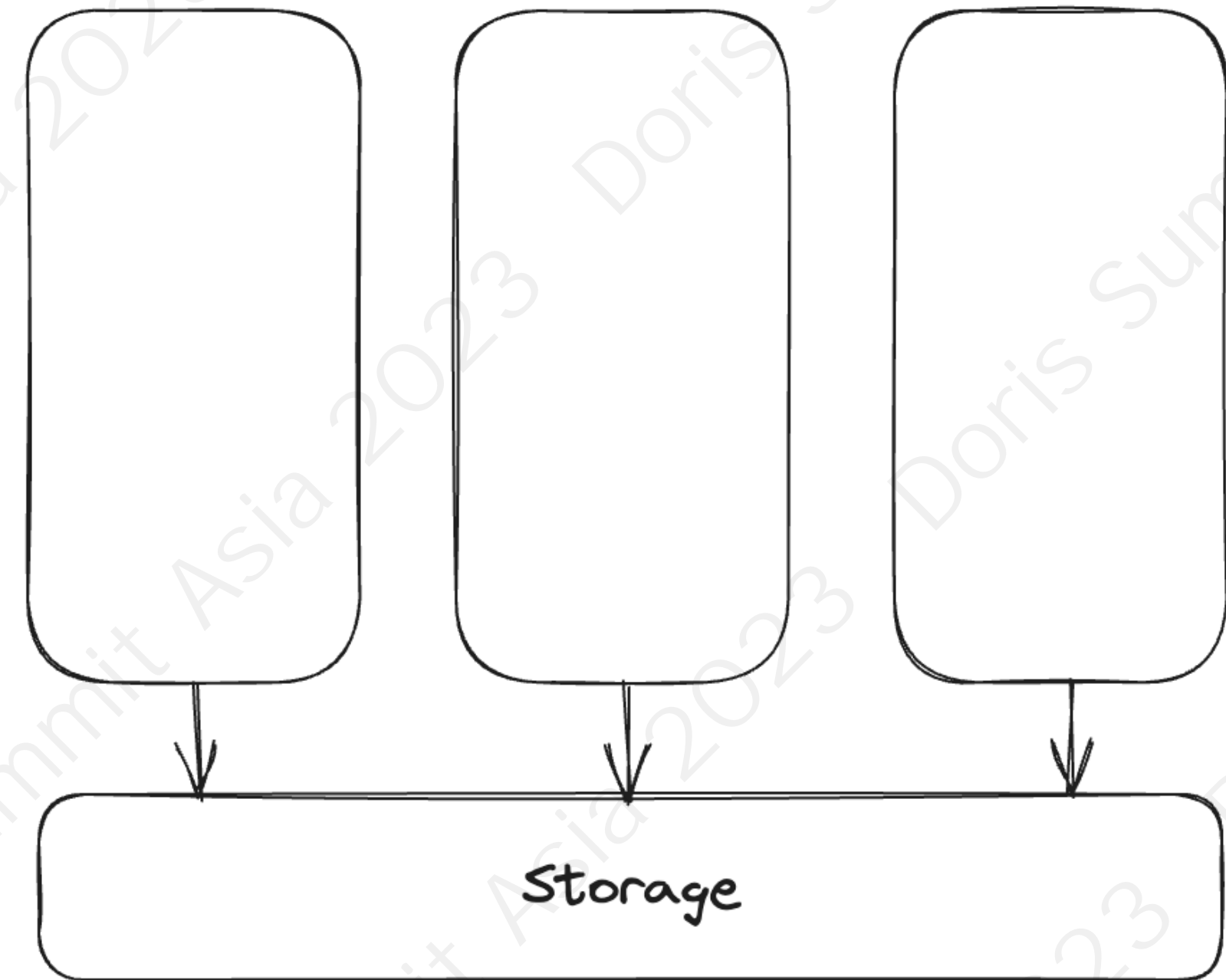


- 低成本：SSD->HDD
- 扩缩容：节点间数据副本自动均衡
- 负载隔离：Resource Group、Workload Group

存算一体 vs 存算分离



- 部署简单：仅FE与BE进程，BE和FE都可以单独扩容
- 稳定可靠：不依赖共享存储系统
- 性能优异：计算节点访问本地存储



- 为什么还需要存算分离？
- 什么样的存算分离架构更能贴合用户需求？
- 要克服那些挑战？

为什么需要存算分离

低成本与资源弹性

- 计算和存储解绑，单独扩缩容
- 计算资源波谷波峰，灵活弹性
- 数据存储冷热效应明显

可靠性与负载隔离

- 读写任务分离
- 更彻底的业务隔离，解决不同业务间的相互影响以及资源抢占问题

数据共享

- 单一数据面向不同的分析负载使用
- 数据快速移动、快速备份恢复

云基础设施的成熟

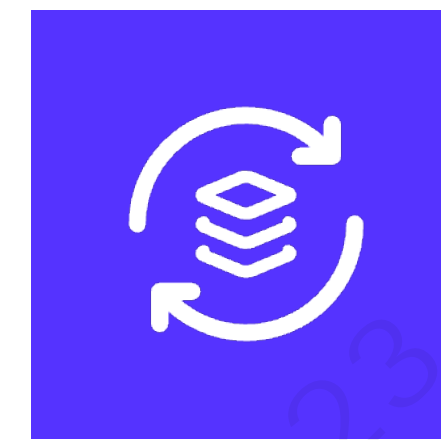
- 云上基础设施逐步完善，提供可靠的共享存储
- 完全按量付费，资源消耗更低

从实际需求出发，2.0 版本降低成本提升弹性



低成本

- 引入对象存储节省冷数据资源
- 增加弹性计算节点

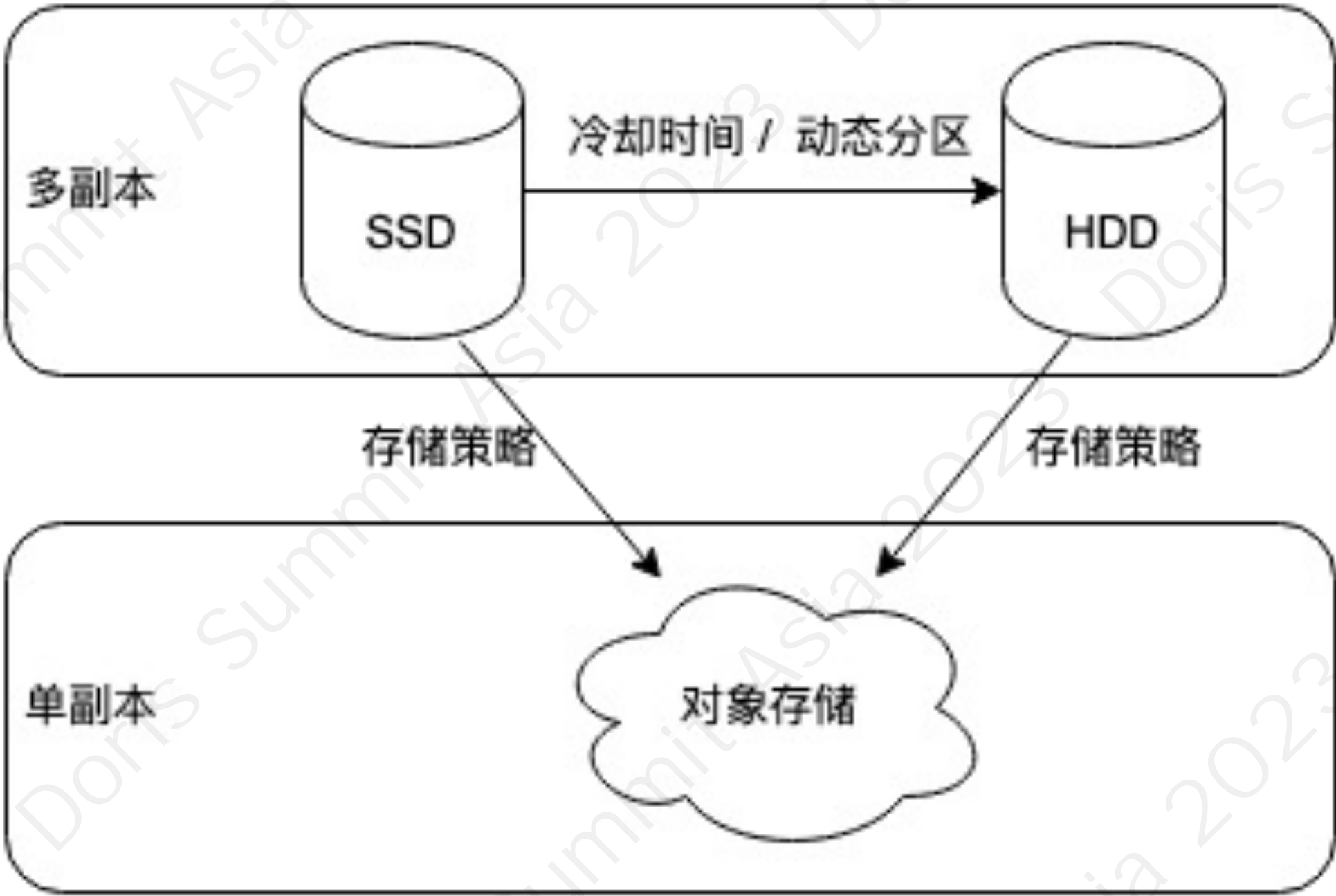


逐步迭代

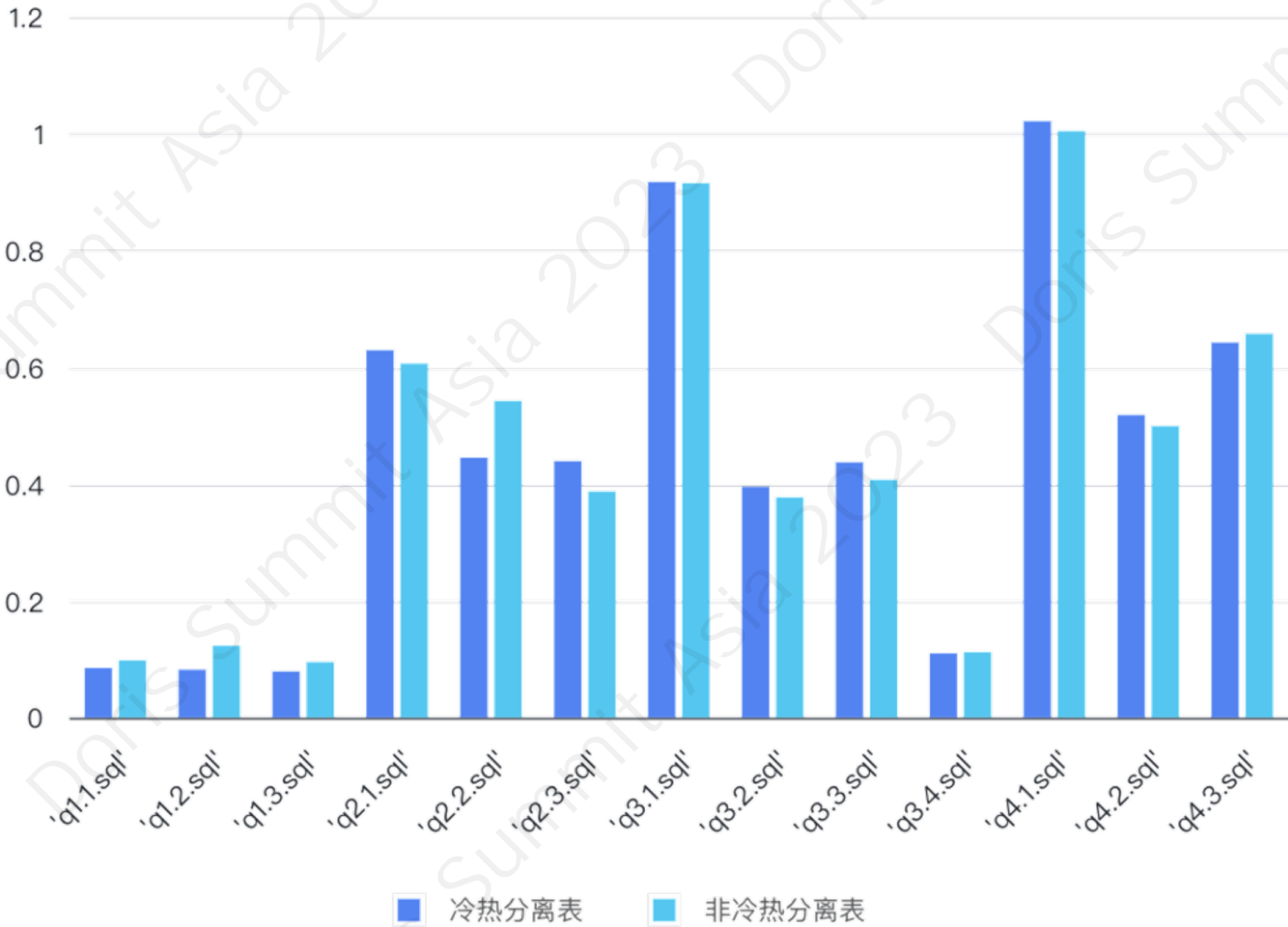
- 绝大多数用户已采取存算一体架构
- 升级过程中需要保证对已有架构的兼容

2.0 冷热数据分层，实现冷数据存储分离

- 根据将冷热数据分别存储在成本不同的存储介质上，从原本的 SSD->HDD 增加到 SSD -> HDD ->OS 三层；
- 云磁盘的价格通常是对象存储的 5-10 倍，如果可以将 80% 的冷数据保存到对象存储中，**存储成本至少可降低 70%**；
- 通过冷数据 Compaction 实现数据的高效压缩，提供冷数据 Cache 加速冷热数据查询，降低成本的同时保持性能不受影响；

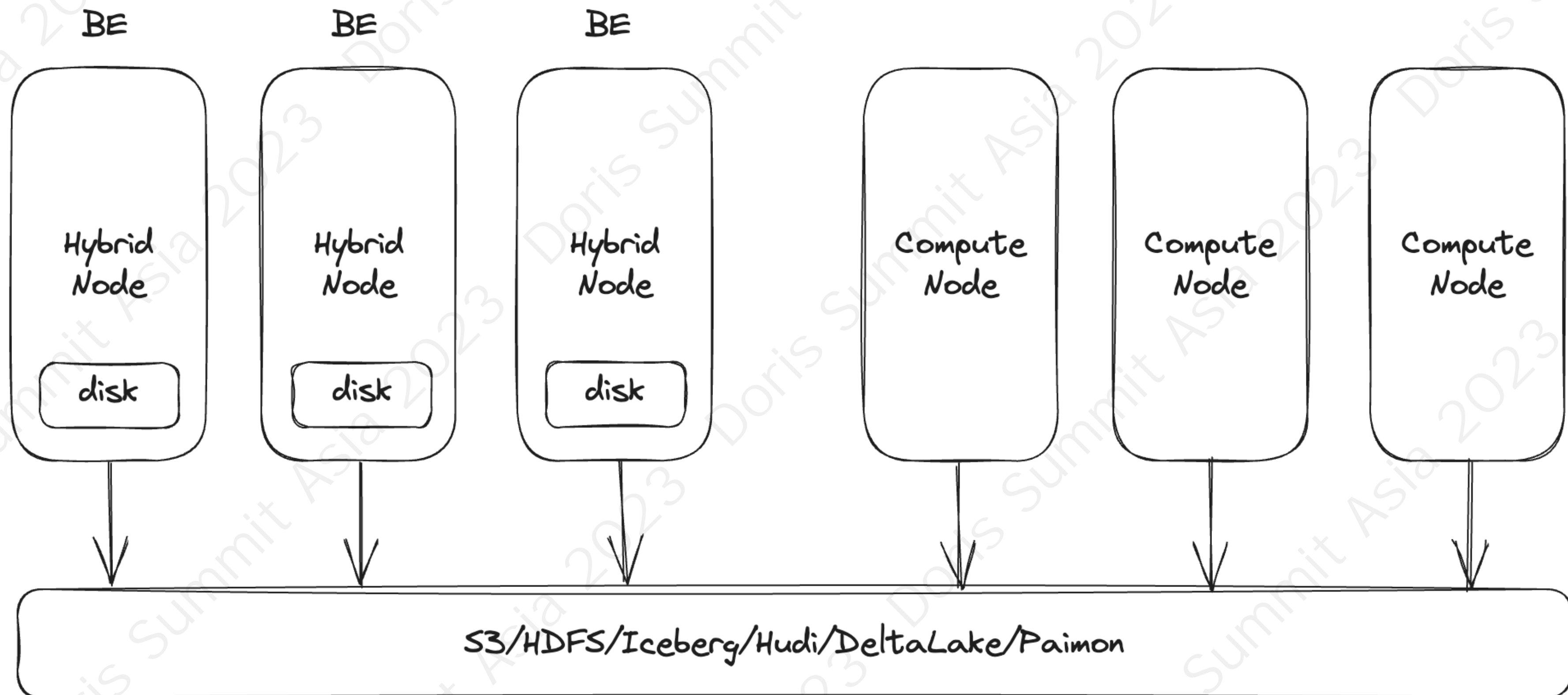


查询性能对比



2.0 弹性计算节点，实现计算分离

- 引入了无状态的计算节点 Compute Node，专门用于执行计算分析；
- 不保存任何数据，集群扩缩容时无需进行数据分片的负载均衡，具有明显高峰的场景中可以灵活扩容、快速加入集群分摊计算压力；
- 用户数据往往存储远端存储中，执行查询时查询任务会优先调度到 Compute Node 执行，以避免内表与外表查询之间的计算资源抢占。



2 Apache Doris 存算分离架构实现方案

整体改造思路



负载隔离

读写分离

业务隔离

内部负载隔离



低成本

存储成本大幅下降

计算和存储可以独立弹性

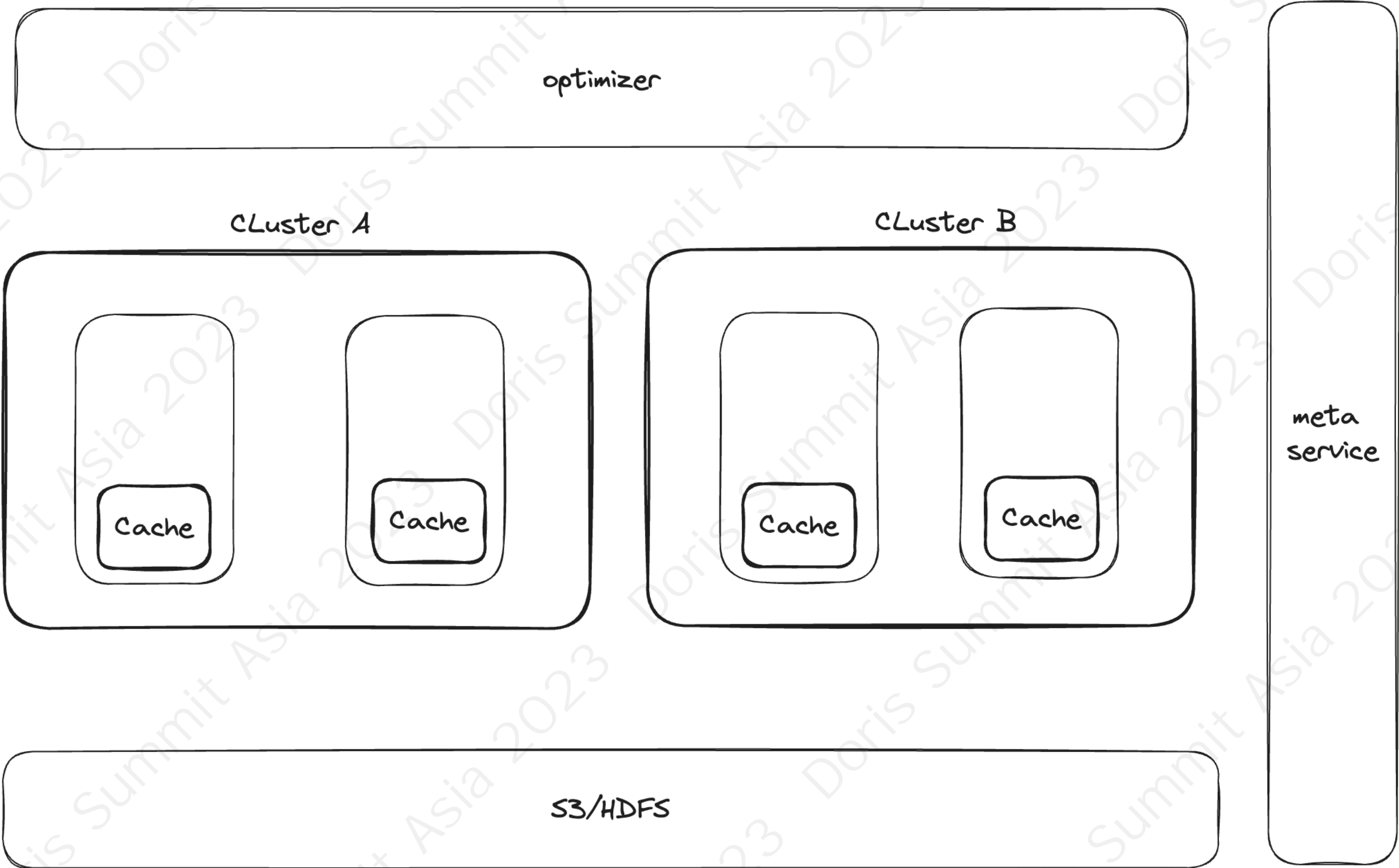
使用业务的波峰波谷调整计算资源



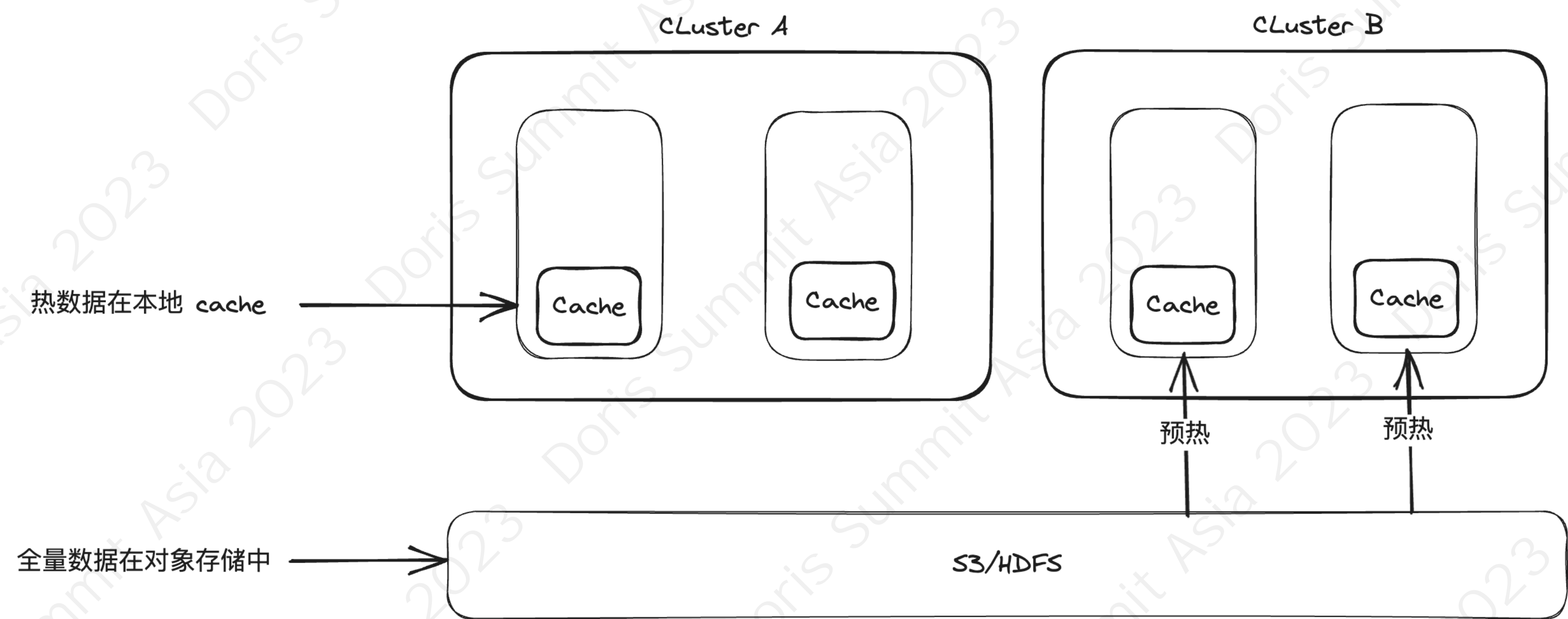
数据共享

元数据服务

存算分离整体架构



热数据本地 Cache，全量数据在对象存储



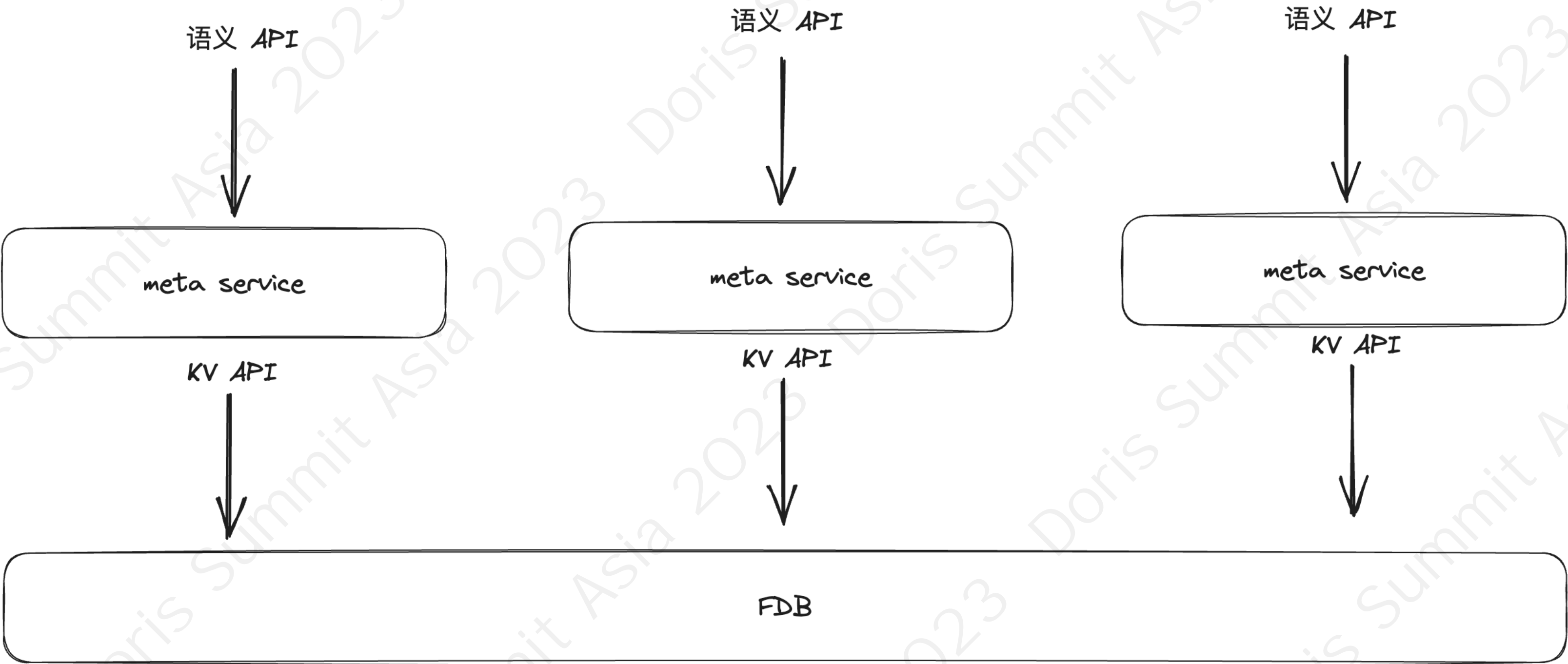
成本最高降低90%

- 存算一体：全量数据 * 3 * 块存储价格
- 存算分离：热数据 * 1 * 块存储价格 + 全量数据 * 对象存储价格
- 最高可以节省90%以上

灵活的Cache管理

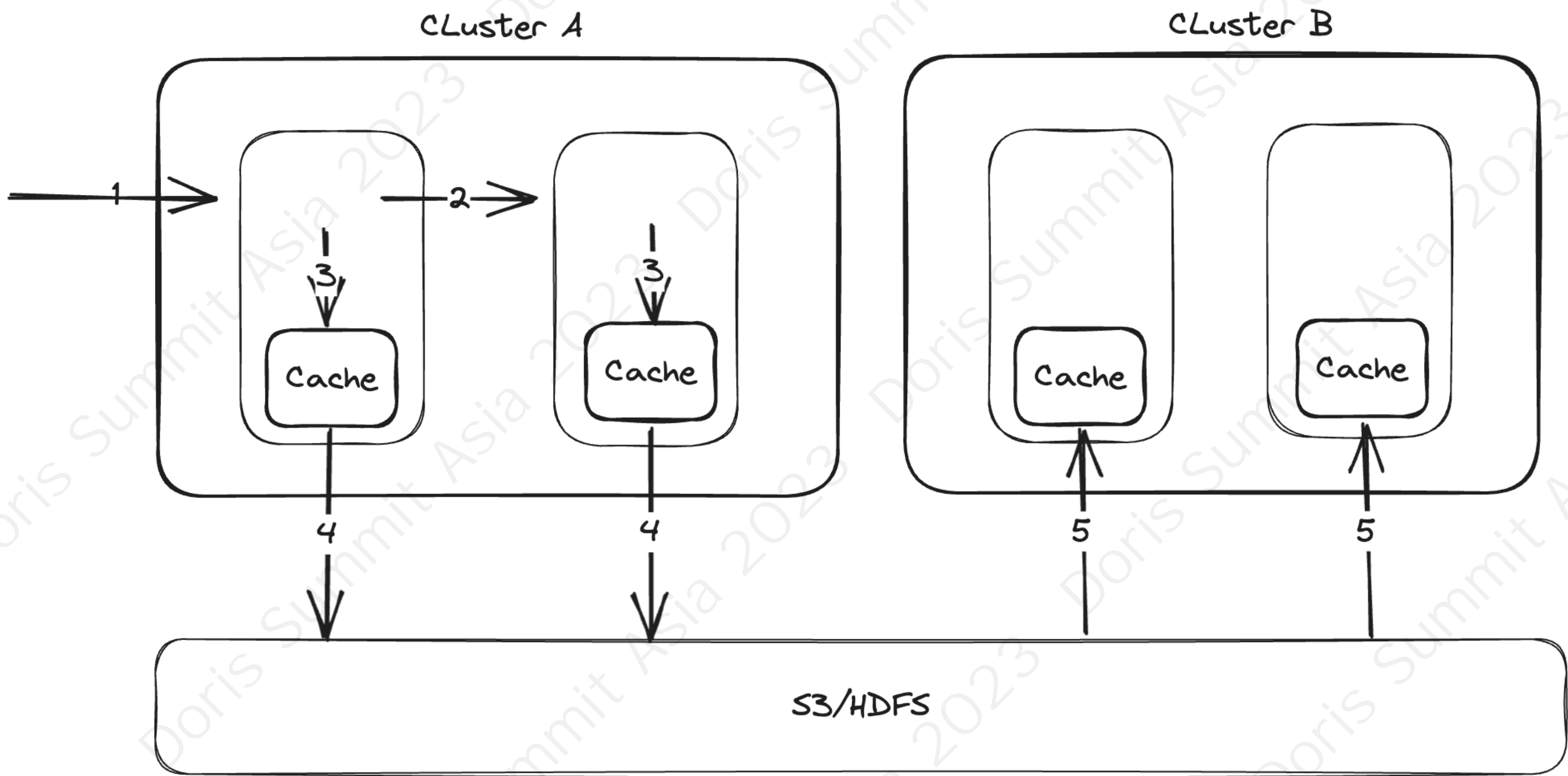
- 手动预热
- 读写预热

元数据层



- 高可靠
- 高可用
- 高性能
- 语义 API 更加模块化和内聚
- meta service 无状态
- 写入事务依赖 FDB，缩短了导入路径

数据导入

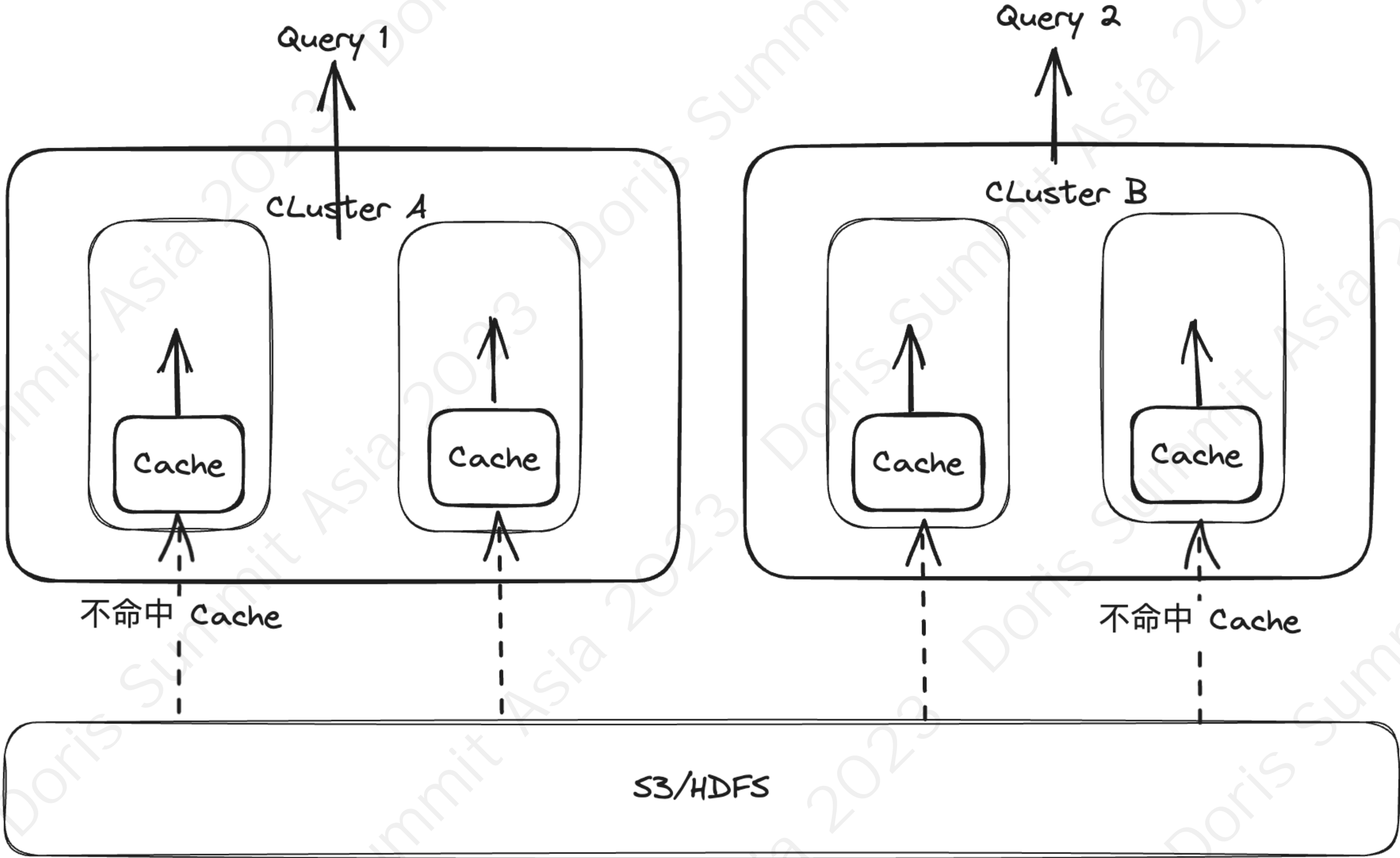


- 1. 数据进入协调者 BE
- 2. 数据分发到多个 BE
- 3. 数据写入 Cache
- 4. 数据写入 S3
- 5. 读写分离 Cluster 预热 Cache

数据导入效率更高

- 只需处理单副本数据
- 数据和BE没有固定的关系
- 没有 publish，写入流程更短

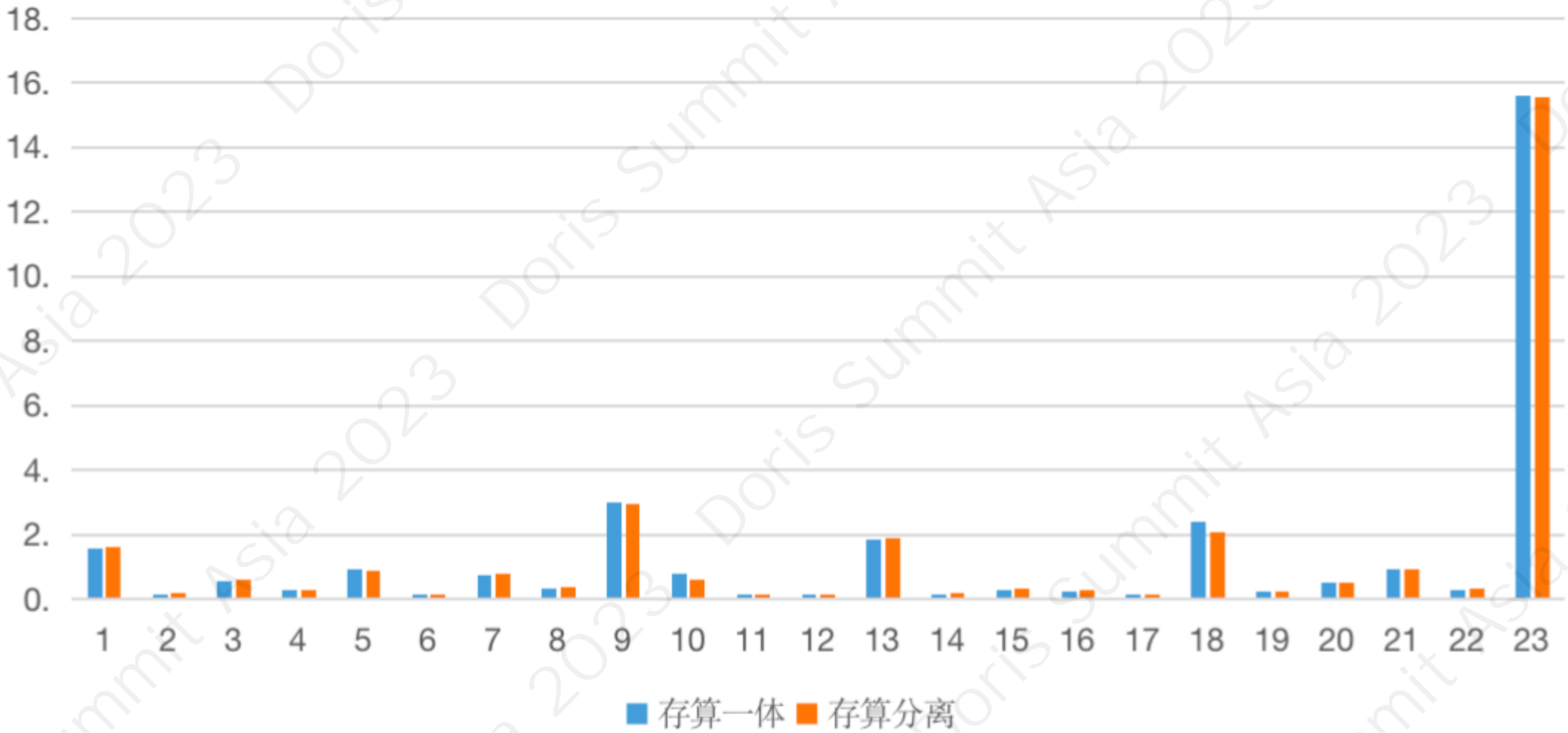
数据查询



- 多个 cluster 独立
- 不命中 Cache 时从 S3 读数据
- 命中时从本地 Cache 读数据
- 弹性资源大幅降低成本

性能

TPCH-SF100



3 存算分离功能示例

存算分离 - 多计算集群实现工作负载隔离



存算分离 – 计算的弹性扩缩容

手动弹性伸缩 | 分时弹性伸缩

计费方式

按量计费

当前计算资源

4 vCPU, 32 GB 内存

当前缓存空间

200 GB

目标计算资源 (vCPU)



默认每个集群最多 1024 vCPU，如需更高配额，请 [联系我们](#) 申请

目标内存大小

128 GB

目标缓存空间

800 GB

手动弹性伸缩 | 分时弹性伸缩

分时规则 ☒

+ 添加

规则 ?	执行周期	执行时间 ?	目标 vCPU ?	操作
规则 1	每天	00:00	256	编辑 删除
规则 2	每天	01:00	128	编辑 删除

- 手动：可以选择 cpu 数据进行伸缩
- 分时：可以指定一条扩容规则和一条缩容规则

存算分离 – 计算的弹性扩缩容 – 集群自动停机

< 新建集群

集群名称 *

请输入

必须字母开头，最多 64 个字符，可以使用字母（大小写不敏感）、数字和

集群计算资源 (vCPU)

4

256

512

768

1024

-

4

+

默认每个集群最多 1024 vCPU，如需更高配额，请 [联系我们](#) 申请

集群内存大小

32 GB

集群缓存空间

200 GB

计费方式

按量计费

包年包月

自动启停

自动停机触发条件： 闲置时长 ② 大于 10 分钟

自动启动触发条件： 暂不支持，需要手动启动。

存算分离 - 性能

< 新建集群

集群名称 *

请输入

必须字母开头，最多 64 个字符，可以使用字母（大小写不敏感）、数字和_

集群计算资源 (vCPU)

42565127681024

-

4

+

默认每个集群最多 1024 vCPU，如需更高配额，请 [联系我们](#) 申请

集群内存大小

32 GB

集群缓存空间

200 GB

计费方式

按量计费

包年包月

自动启停

自动停机触发条件：

闲置时长 ^② 大于 10 分钟 [↗]

自动启动触发条件：

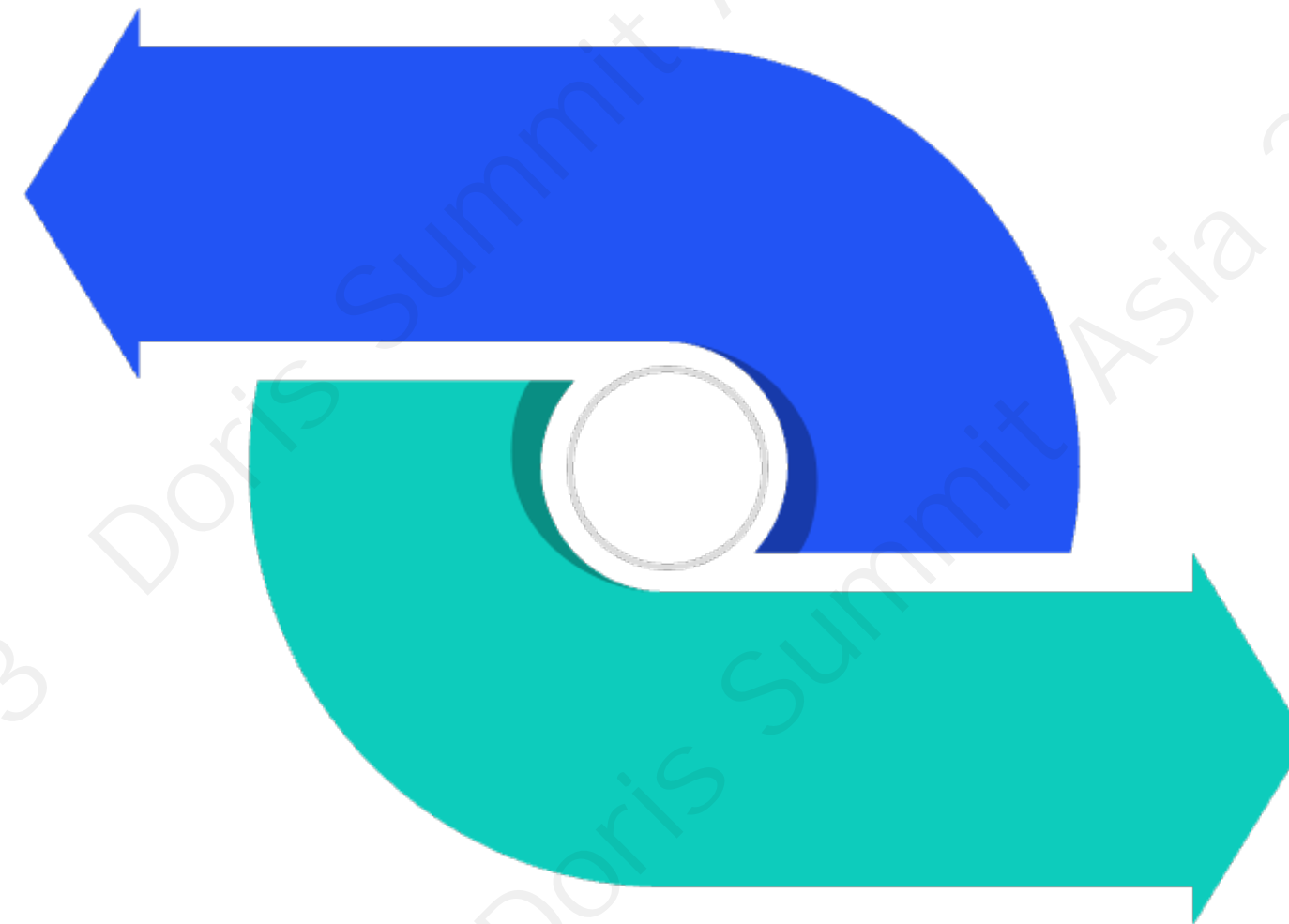
暂不支持，需要手动启动。

4 存算分离和存算一体适用场景

两种架构的适用场景

存算一体

- 对性能要求更高、对查询时延更敏感
- 不需要极致弹性扩缩容
- 没有可靠的共享存储可用
- 仅简单体验或开发测试
- 业务线独立运行Doris，无专职DBA
-



- 成本敏感型，追求性价比
- 已经在公有云上构建服务
- 拥有可靠的共享存储服务
- 要求极致弹性，需要运行在容器平台上
- 有专职的团队维护数据仓库平台

存算分离

后续规划

- Resource Group、Workload Group 与多计算集群的融合
- 实现共享的高速缓存，与计算节点进一步分离
- 时间旅行
- 统一 catalog



获取更多社区动态与最佳实践

Apache Doris 官方平台:

- Apache Doris 官网: doris.apache.org
- Apache Doris GitHub: github.com/apache/doris/

获取更多峰会资料:

- Doris Summit 峰会官网: doris-summit.org.cn
- Doris Summit 峰会回放: <https://space.bilibili.com/1196172099/channel/collectiondetail?sid=1824324>