DTCC 13 2010 2022

第十三届中国数据库技术大会

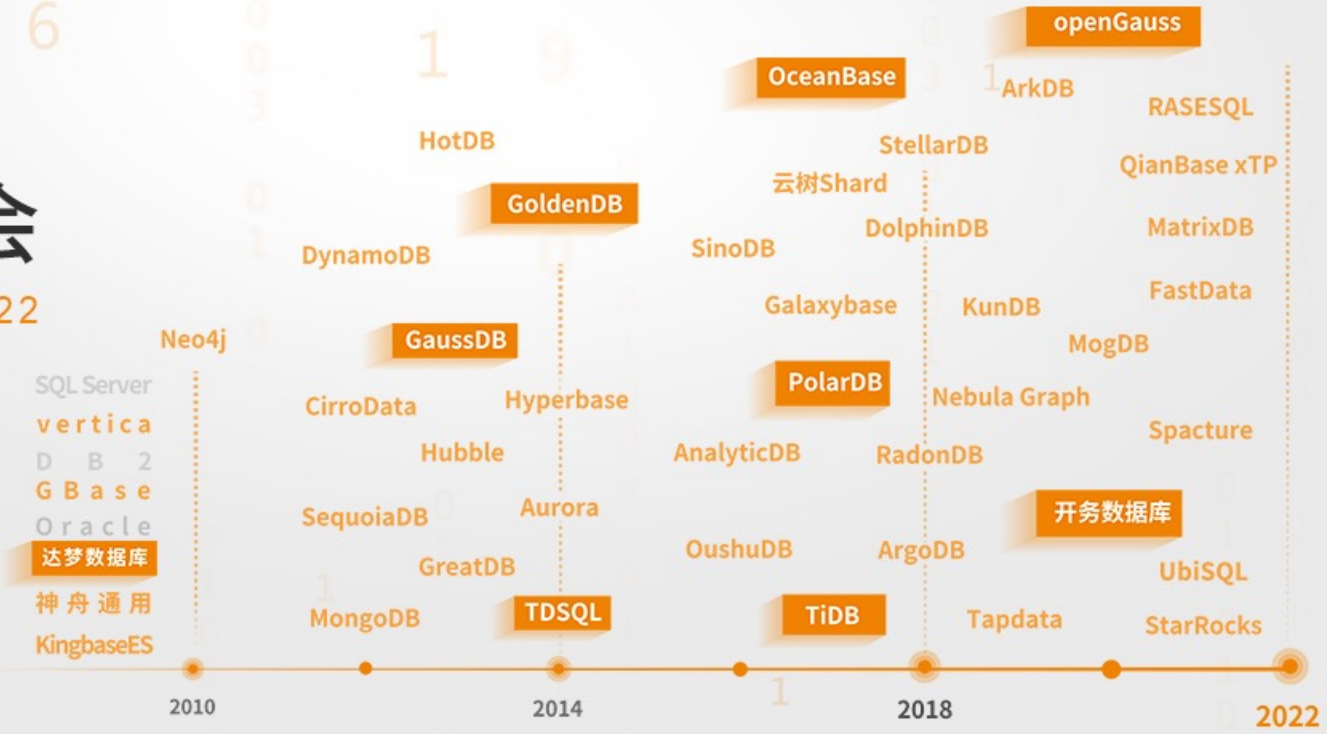
DATABASE TECHNOLOGY CONFERENCE CHINA 2022

数据智能 价值创新

IT168.com ChinaUnix.net ITPUB

线上直播 | 2022/12/14-16

数据来源：数据库产品上市商用时间



百度沧海.存储万亿级元数据底座TafDB设计和实践

曹彪 百度 云存储高级架构师

目录

01 云存储元数据面技术演进趋势

02 云存储元数据底座的技术选型

03 云存储元数据底座TafDB关键设计和实践

04 TafDB应用效果

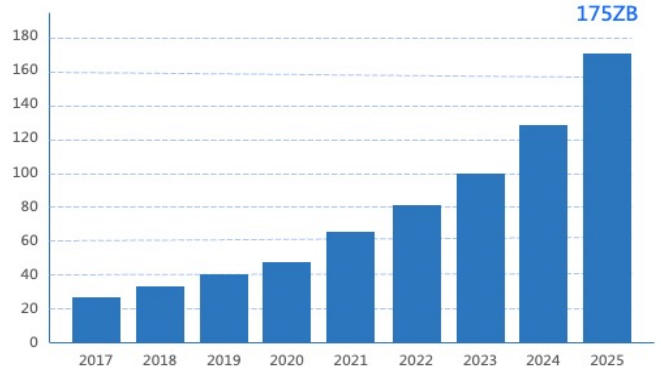
05 后续规划

01 云存储元数据面技术演进趋势

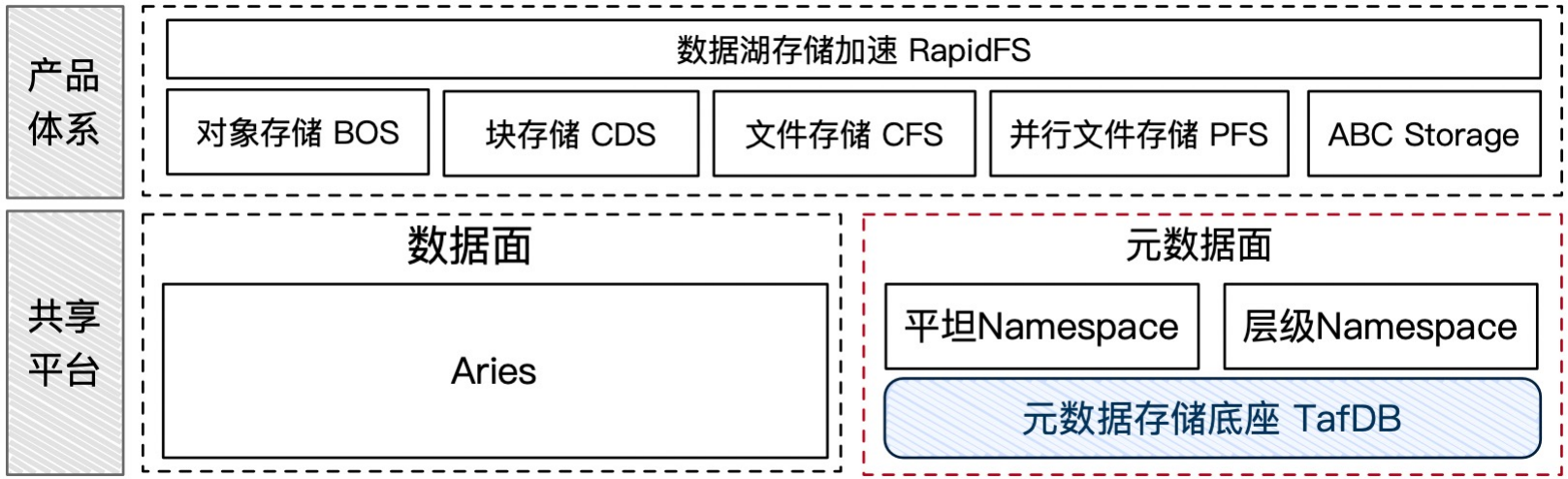
云存储元数据面的扩展性面临挑战

数据持续爆发式增长，对云存储系统的扩展性提出了更高要求

- IDC预测，全球数据量从2018年的33ZB将增长至2025半年的175ZB



云存储系统由元数据面和数据面构成，元数据面的扩展性影响到整个存储系统的扩展性

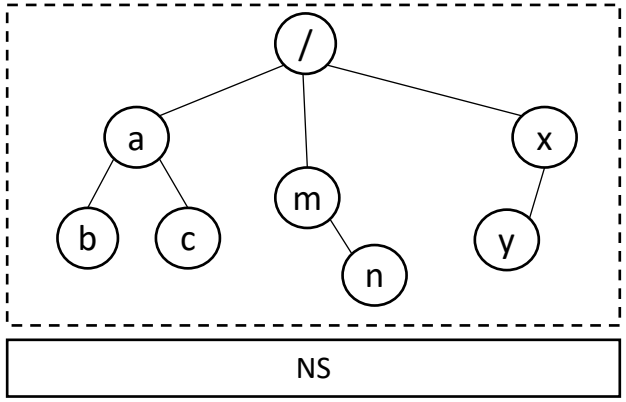


百度沧海.存储 整体架构

元数据面技术演进趋势 – 层级Namespace

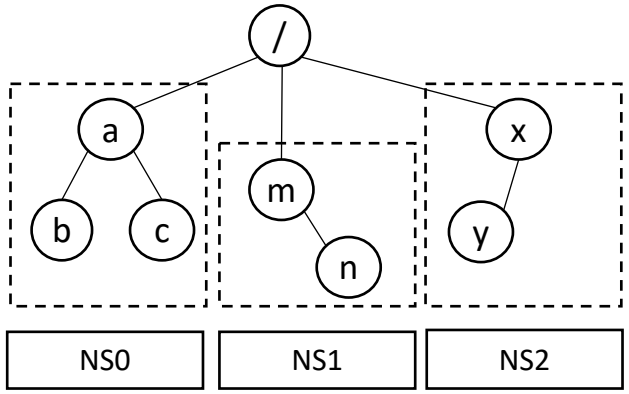
分布式文件系统namespace架构演进路线

① 单机全内存目录树架构



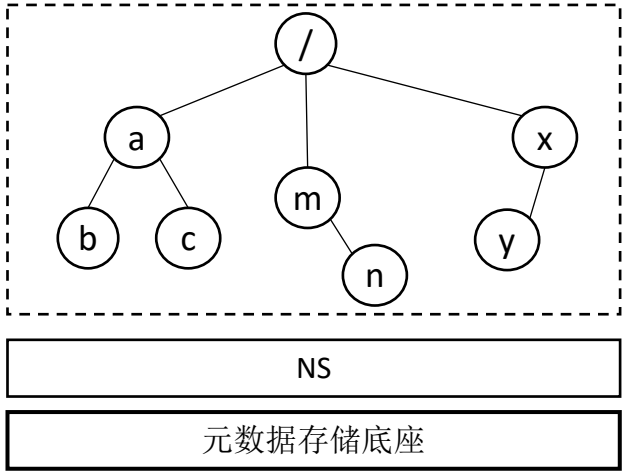
- 10亿级别，无法横向扩展

② 静态子目录树划分



- 单个静态子树存在上限，最大10亿级别
- 易产生热点，负载均衡代价高
- 不支持跨子树的rename

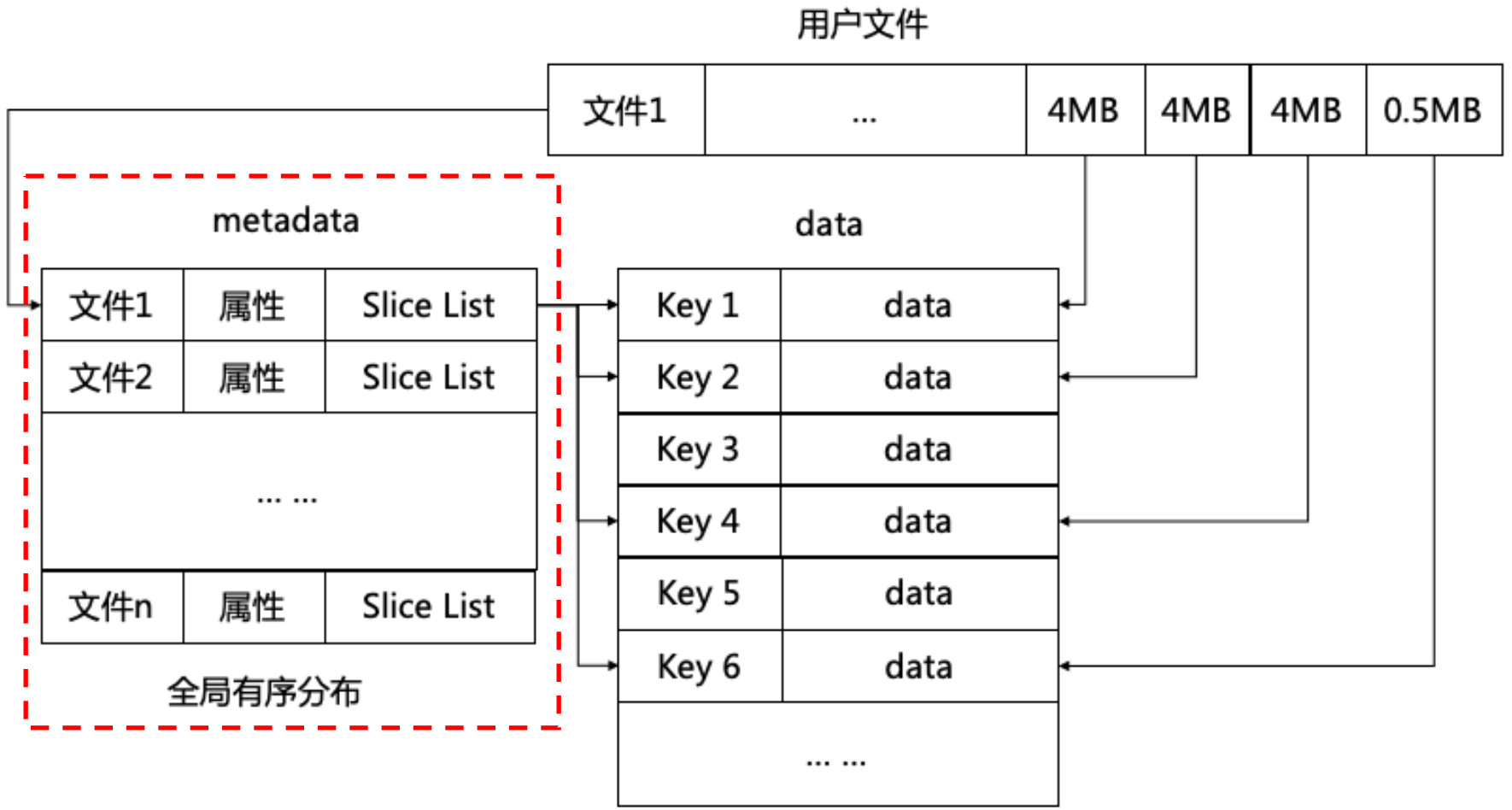
③ 基于共享存储



- 无限扩展

元数据面技术演进趋势 – 平坦Namespace

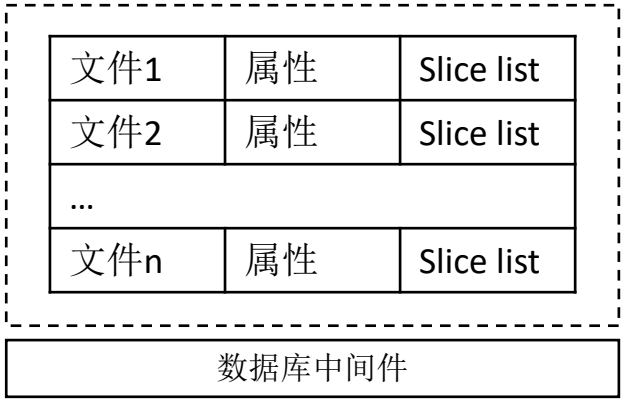
对象存储平坦namespace介绍



元数据面技术演进趋势 – 平坦Namespace

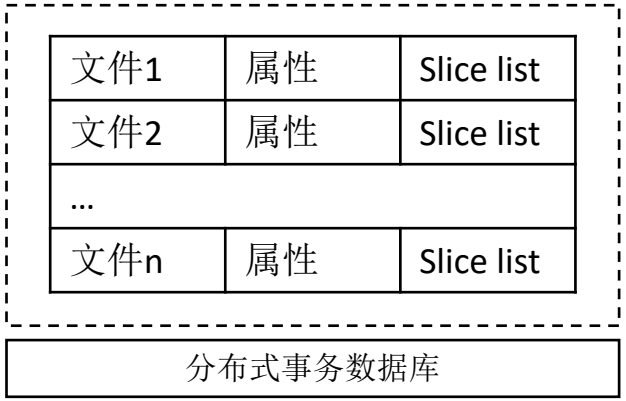
平坦namespace架构演进路线

① 基于数据库中间件



- 扩容只能倍扩，对成本造成很大压力
- 对跨库的分布式事务支持不友好

② 基于分布式事务数据库



- 从根本上解决了数据库中间件扩展性问题

02 云存储元数据底座的技术选型

元数据底座设计目标

规模

支撑万亿级纪录存储

性能

百万QPS，毫秒级延迟

运维

数据均衡，扩缩容简单

数据库特性

事务、索引、备份、CDC

元数据底座技术选型

调研

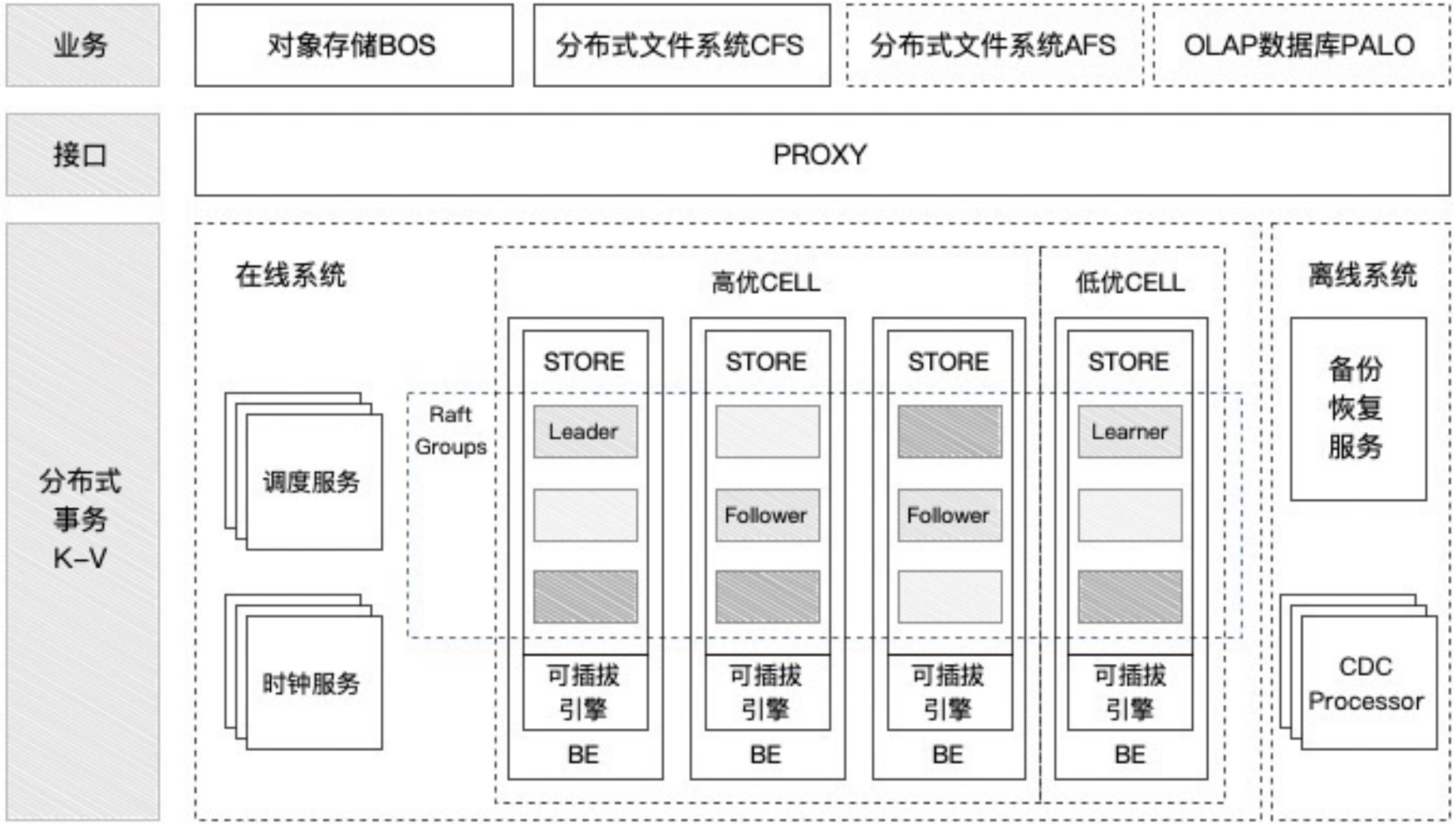
技术流派	适应场景
Spanner	面向通用**OLTP**场景
Calvin	面向确定性事务**OLTP**场景
FoundationDB	面向对延迟要求不高的**OLTP**场景

结论

自主研发：打造一个类Spanner架构的NoSQL with ACID的系统

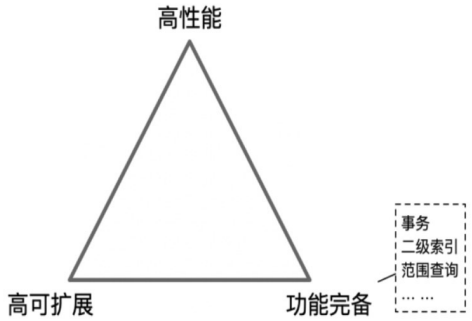
03 TafDB关键设计和实践

TafDB系统架构


业务
对象存储BOS
分布式文件系统CFS
分布式文件系统AFS
OLAP数据库PALO
接口
PROXY
分布式
事务
K-V
在线系统
高优CELL
低优CELL
离线系统
调度服务
时钟服务
Raft
Groups
STORE
Leader
Follower
Learner
可插拔
引擎
BE
备份
恢复
服务
CDC
Processor

TafDB工程挑战

挑战



挑战1：如何在保证元数据ACID 操作的同时，避免2PC事务的高额开销？

-- 解决事务、索引功能和系统性能的矛盾

挑战2：如何在大量删除场景保证LSM-Tree范围操作的性能?

-- 解决连续删除 + 范围查询和性能的矛盾

挑战3: 如何消除数据流程的单点，提供极致的扩展性和可用性？

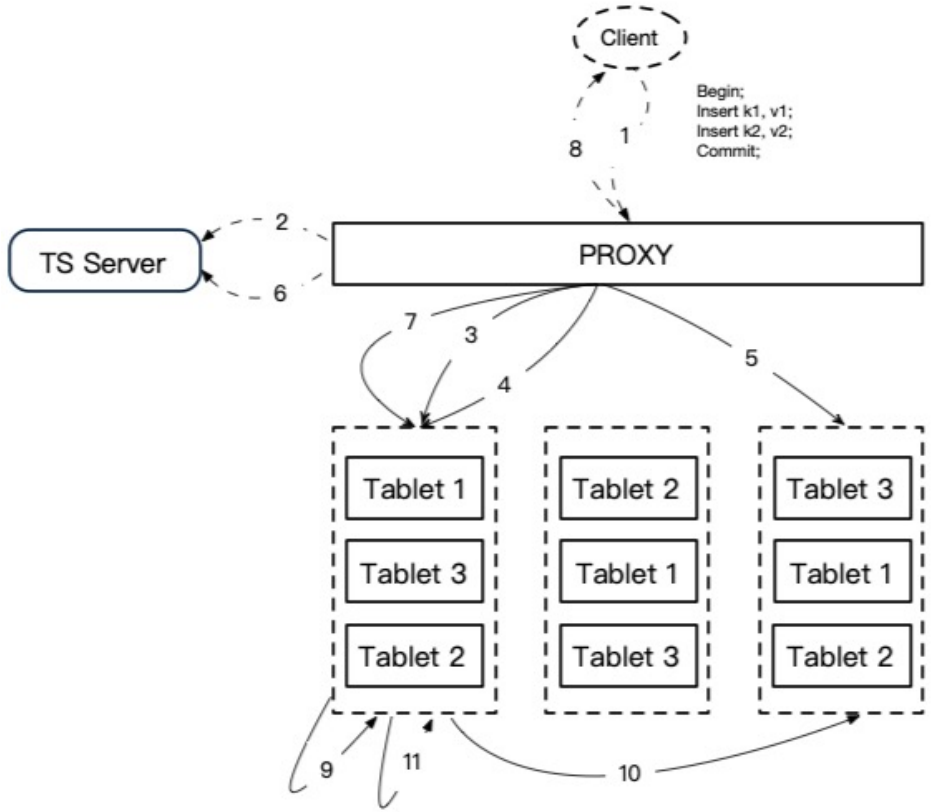
-- 解决事务功能和高可扩展性&可用性的矛盾

TafDB工程挑战 – 整体解决思路

挑战	解决思路
挑战1：如何在保证元数据ACID操作的同时避免2PC事务的高额开销？ -- 解决事务、索引功能和系统性能的矛盾	消除系统中不必要的跨分片事务： ① 异步索引 ② 自定义分裂策略
挑战2：如何保证LSM-Tree范围操作的性能? -- 解决lsm-tree连续删除 + 范围查询和性能的矛盾	① Scale out：打散删除到多台机器 ② Scale up: 提升单机墓碑消除速度，在确定性时间内消除墓碑
挑战3：如何消除数据流程的单点来提供极致的扩展性和可用性？ -- 解决事务、索引功能和系统性能的矛盾	① 设计分布式时钟方案

挑战1：如何在保证元数据操作ACID的同时，避免2PC事务高额开销？

背景：跨分片事务性能开销极大



2PC事务流程

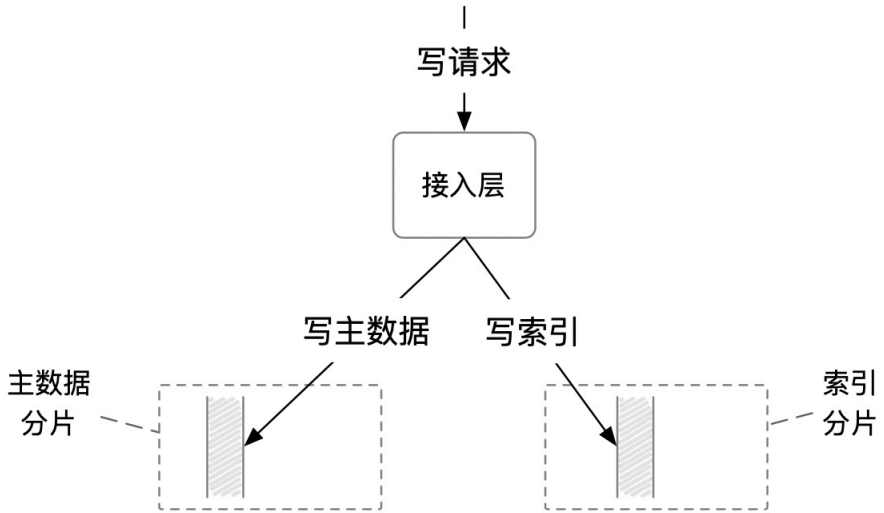
- 未经优化的2PC模型会带来N倍的写放大

跨2个分片事务的写放大次数	
写Raft log次数	写rocksdb次数
7	7

挑战1：如何在保证元数据操作ACID的同时，避免2PC事务高额开销？

痛点：元数据场景可能触发大量跨分片事务

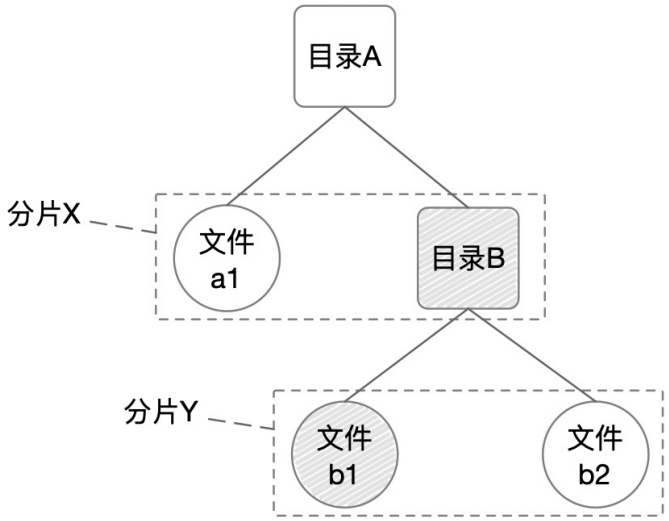
平坦Namespace场景



全局二级索引：

- 主键数据和索引数据在不同的分片上
- 依赖分布式事务保证主键和索引写入的原子性

层级Namespace场景



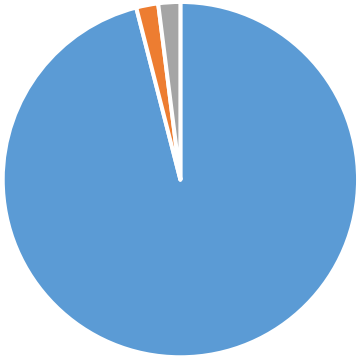
目录树操作：

- 要修改的节点通常在不同分片上
- 依赖分布式事务保证跨分片操作的原子性

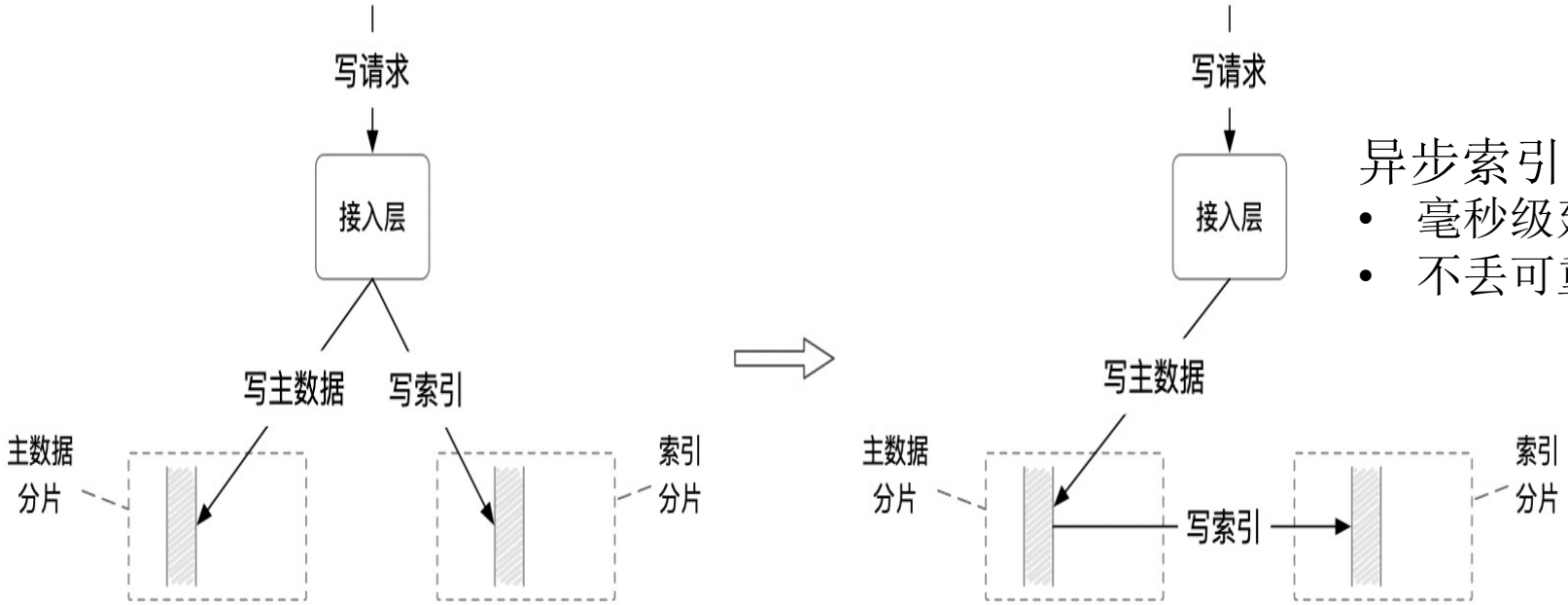
挑战1：如何在保证元数据操作ACID的同时，避免2PC事务高额开销？

针对平坦Namespace解决方案：支持异步二级索引

BOS写请求类型占比



	主键和索引不要求原子写入
	主键和索引要求原子写入
	Rename



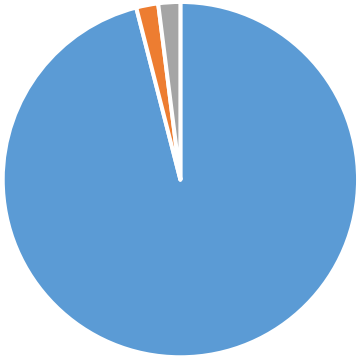
异步索引系统特性：

- 毫秒级延迟
- 不丢可重

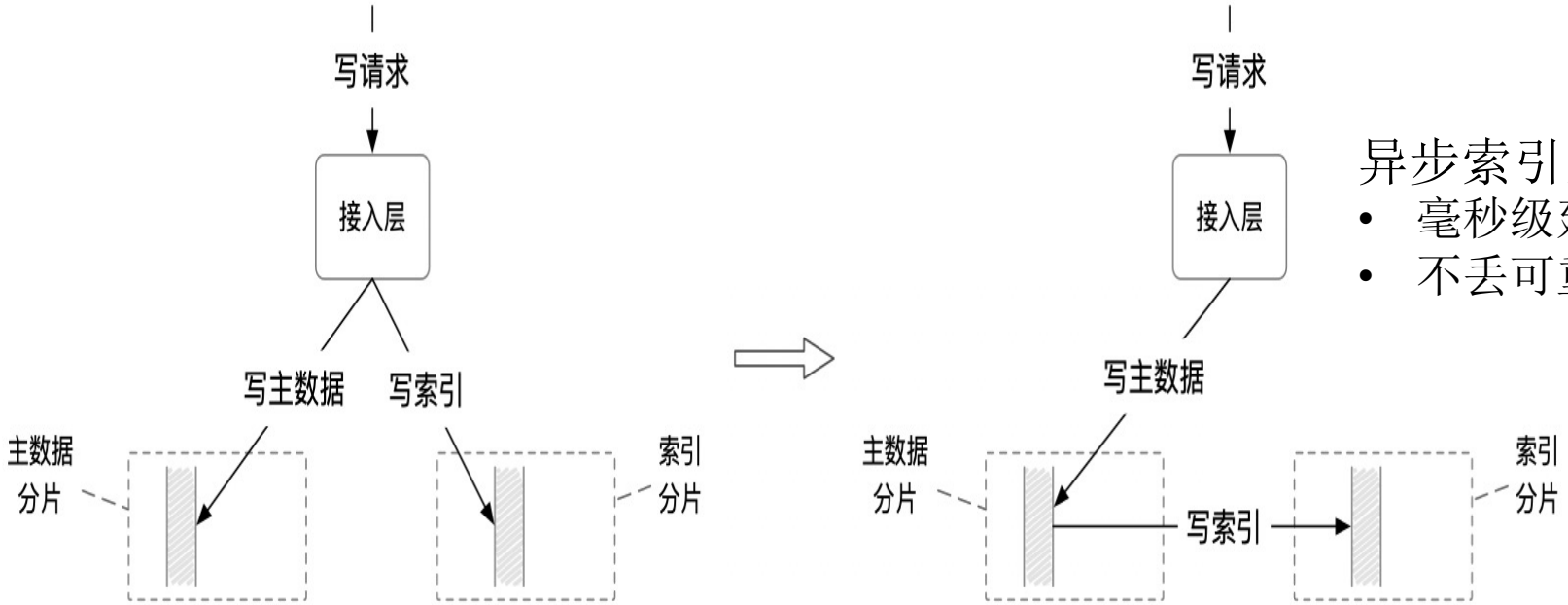
挑战1：如何在保证元数据操作ACID的同时，避免2PC事务高额开销？

针对平坦Namespace解决方案：支持异步二级索引

BOS写请求类型占比



	主键和索引不要求原子写入
	主键和索引要求原子写入
	Rename



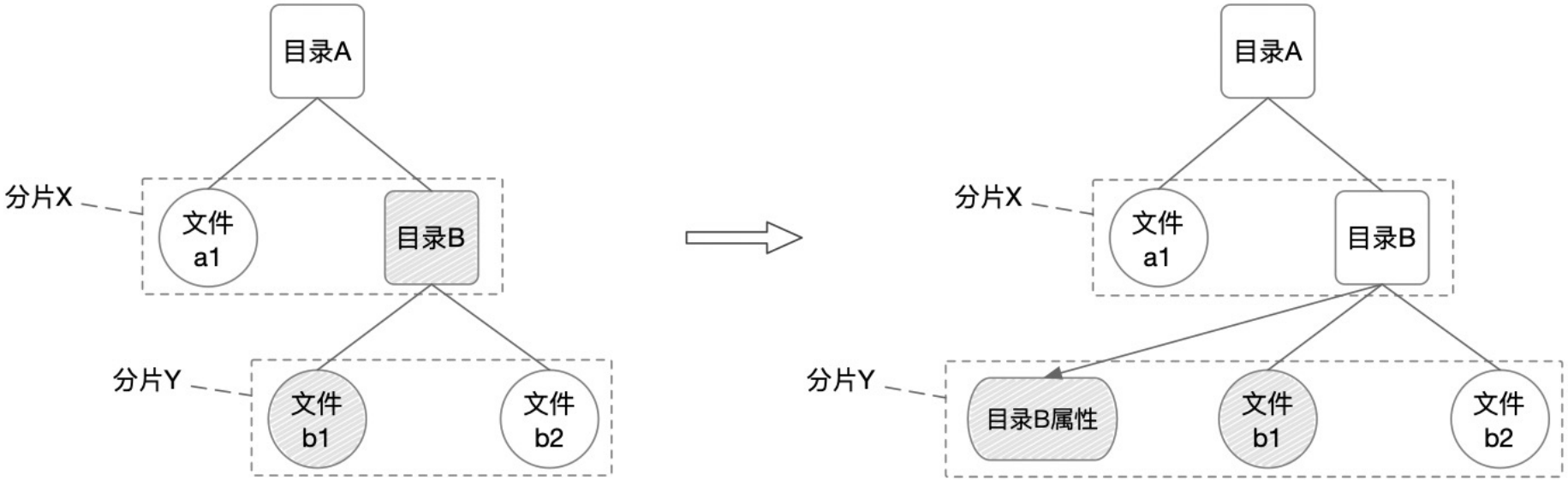
异步索引系统特性：

- 毫秒级延迟
- 不丢可重

挑战1：如何在保证元数据操作ACID的同时，避免2PC事务高额开销？

针对层级Namespace解决方案：支持自定义分裂策略

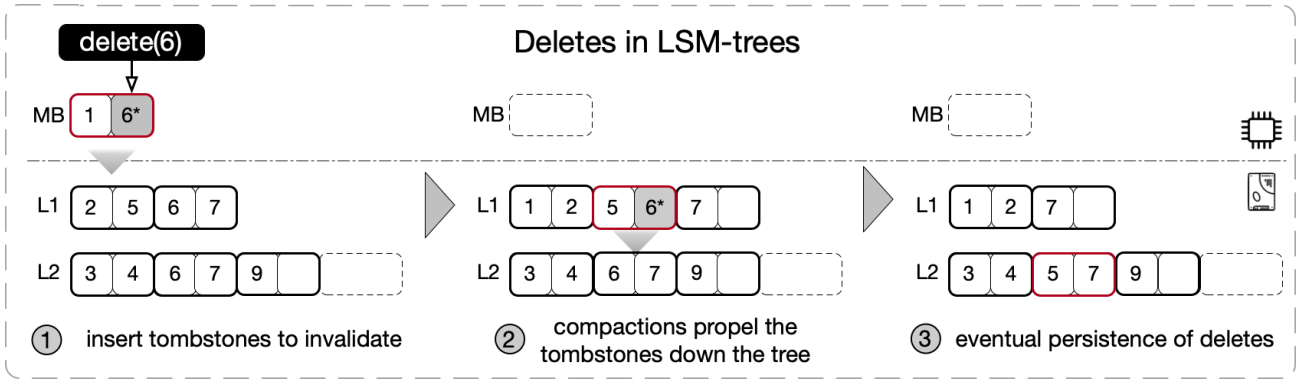
- 关键设计：
  - 支持自定义分裂策略：同目录下的所有子节点控制在同一个数据分片
  - 单分片事务优化：业务系统调整表结构，控制大部分操作要更改的节点都在同层目录



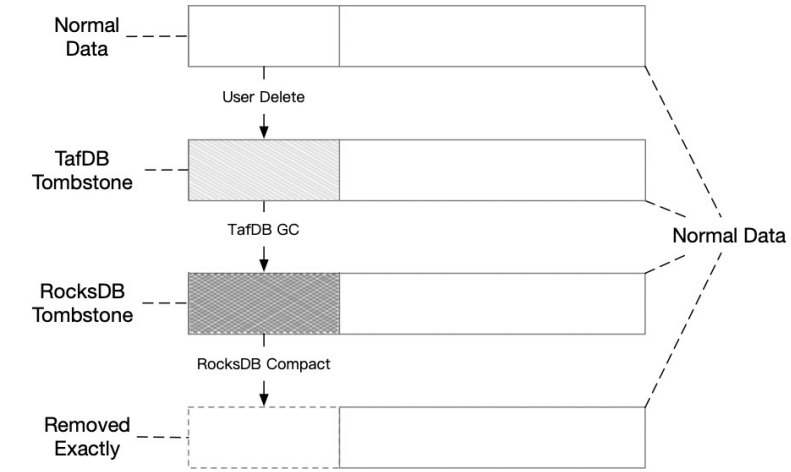
挑战2：如何在大量删除场景保证LSM-Tree范围查询的性能?

背景

- LSM-Tree对大量删除后的范围操作不友好
  - ① Rocksdb Tombstone存在时长无法控制
  - ② 如范围扫描的区间存在大量墓碑会导致扫描性能低下



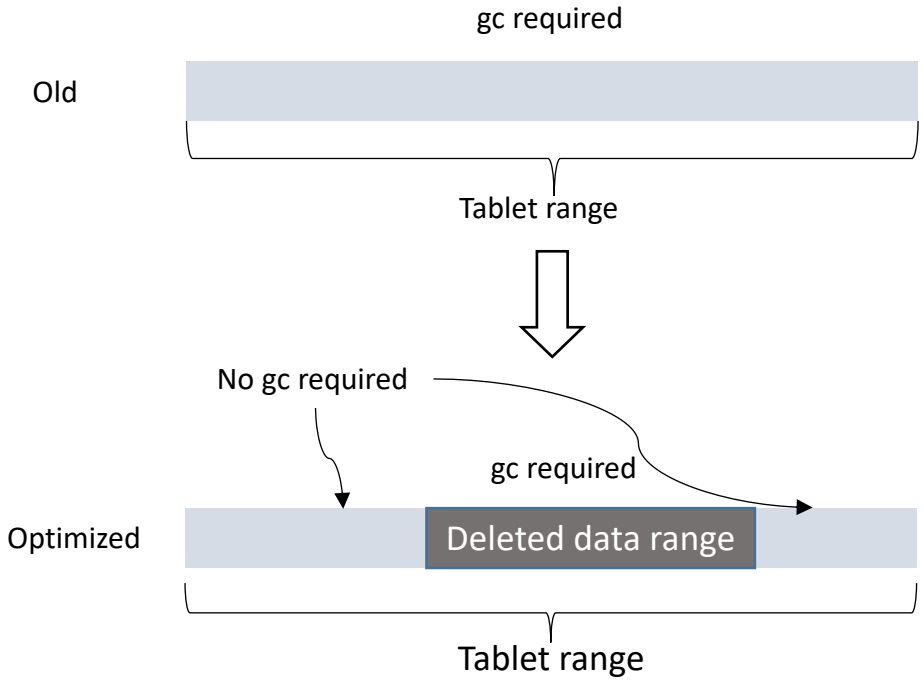
- TafDB两层墓碑机制进一步加剧了此问题
  - ① 为实现MVCC，TafDB在删除时使用墓碑机制
  - ② 全量MVCC GC的方式导致TafDB tombstone存在时长无法控制



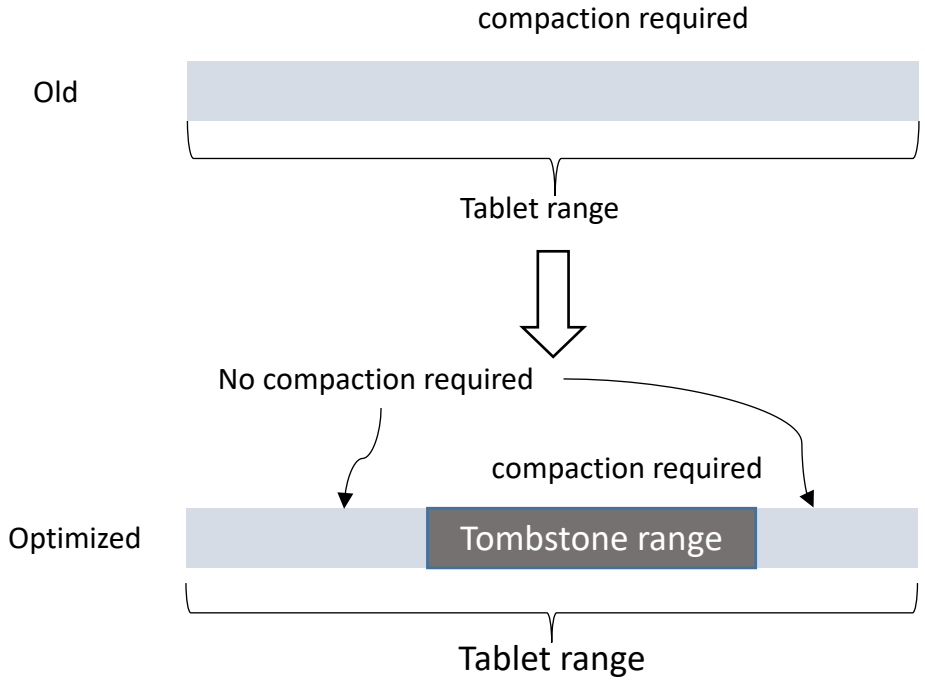
挑战2：如何在大量删除场景保证LSM-Tree范围查询的性能?

解决方案：在确定性时间内消除两层墓碑

- 支持多层级的MVCC GC机制：
  ① FastGC：优先级感知的MVCC GC机制，对存在大量删除的range进行精确的GC
  ② FullGC：周期性全量MVCC GC

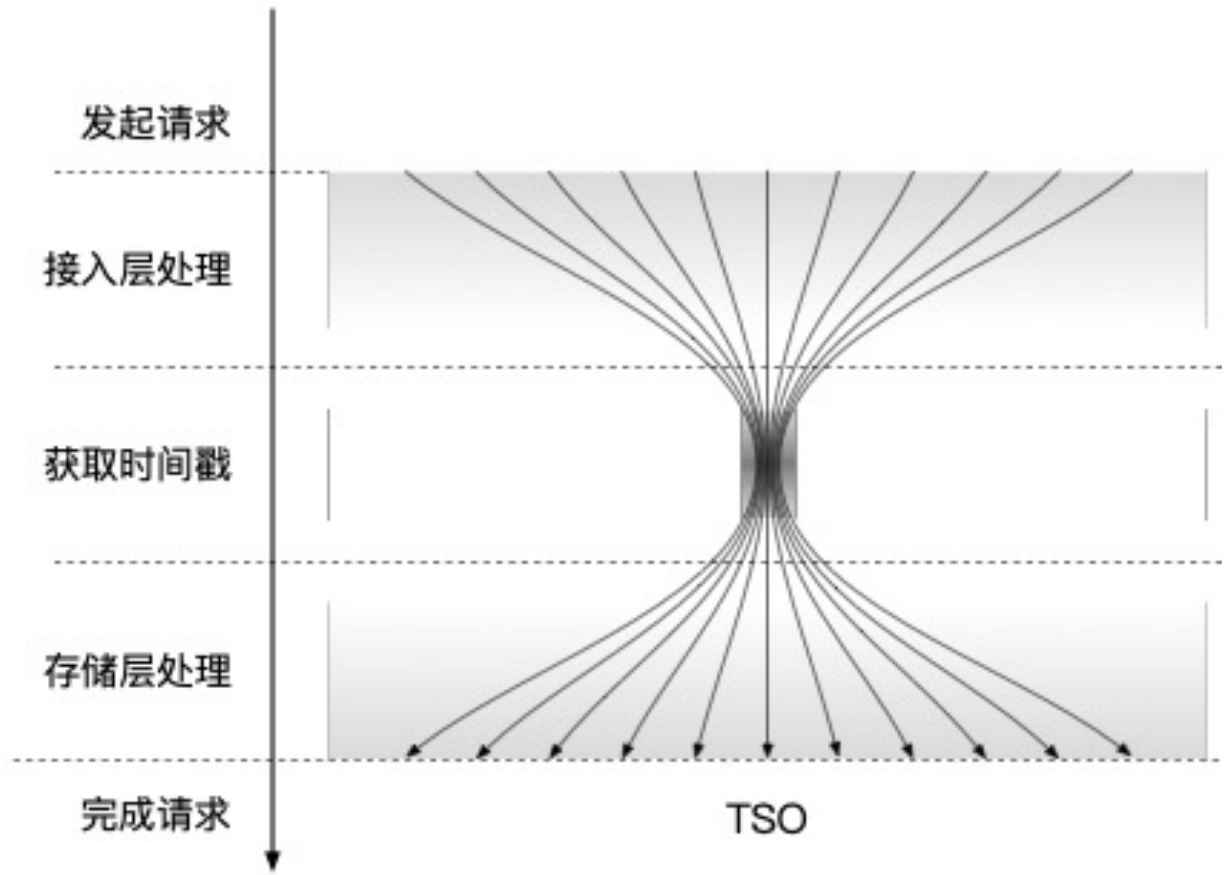


- Compaction精细化调度：
  ① 精确感知删除Range，针对这一段Range触发compact_range



挑战3：如何消除数据流程的单点，提供极致的扩展性和可用性？

痛点：全局时间戳服务器是TafDB 扩展性和可用性的瓶颈



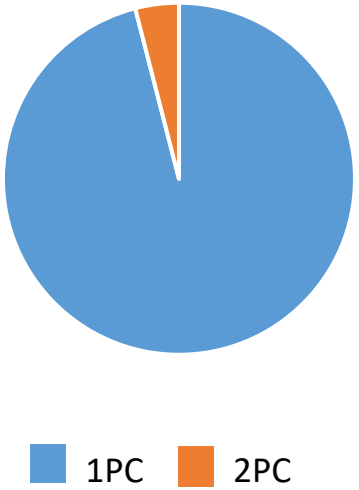
挑战3：如何消除数据流程的单点，提供极致的扩展性和可用性？

业界分布式数据库时钟方案			
方案	系统	优点	缺点
全局时间戳服务器	TafDB当前	实现简洁	时钟服务成为性能和可用性瓶颈
HLC	CockroachDB	无中心化的性能和可用性瓶颈	• 时钟和DB逻辑耦合 • 调试复杂度高

挑战3：如何消除数据流程的单点，提供极致的扩展性和可用性？

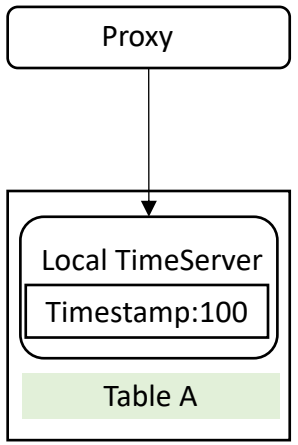
元数据写场景

写请求类型占比



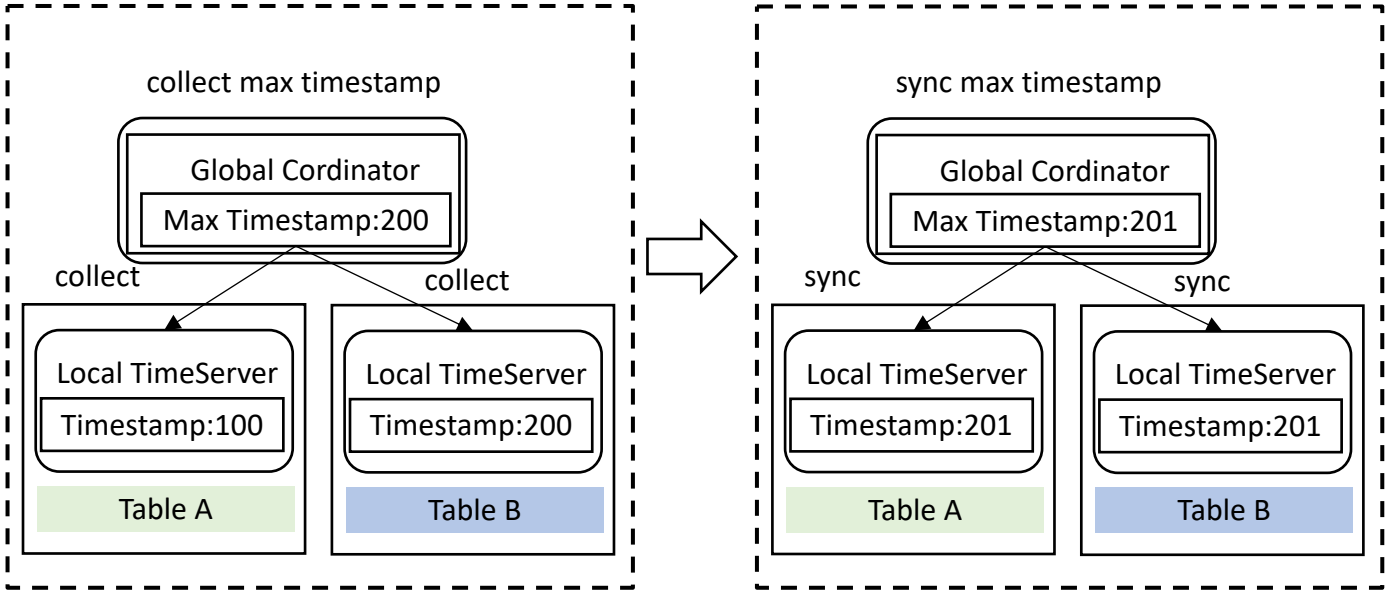
TafDB分布式时钟方案

单分片事务



直接使用本地时钟服务

跨分片事务



通过广播协调保证因果序

04 TafDB应用效果

TafDB应用效果

系统定位：百度沧海.存储统一元数据底座

TafDB应用效果

业务	上线效果
对象存储	• 扩展性提升：单 Bucket 容量从 百亿级别提升到万亿级别
	• 性能提升：小文件延迟降低42%
文件系统	• 扩展性提升：单集群容量从十亿级别提升到万亿级
	• 性能提升：写延迟2ms，读延迟百us级

05 后续规划

后续规划

目标：打造业界领先的元数据存储底座

- 更高性能 – 面向元数据场景设计最优的读写路径
- 更稳定 – 打造零运维的存储系统
- 更易用 – 提供更丰富的“Layer”：NamespaceLayer、EtcdLayer


百度智能云


openGauss
OceanBase
ArkDB
RASESQL
HotDB
StellarDB
QianBase xTP
云树Shard
GoldenDB
DolphinDB
MatrixDB
DynamoDB
SinoDB
FastData
Galaxybase
KunDB
SQL Server
vertica
D B 2
GBase
Oracle
达梦数据库
Spacture
SequoiaDB
开炎数据库
OushuDB
ArgoDB
GreatDB
UbiSQL
神舟通用
MongoDB
TDSQL
TiDB
Tapdata
StarRocks
KingbaseES
2010
2014
2018

THANKS