三大分布式数据库主要技术特性

OceanBase&TiDB&KunlunDB

何革新

泽拓科技（深圳）有限责任公司

目录

CONTENTS

数据分布原理

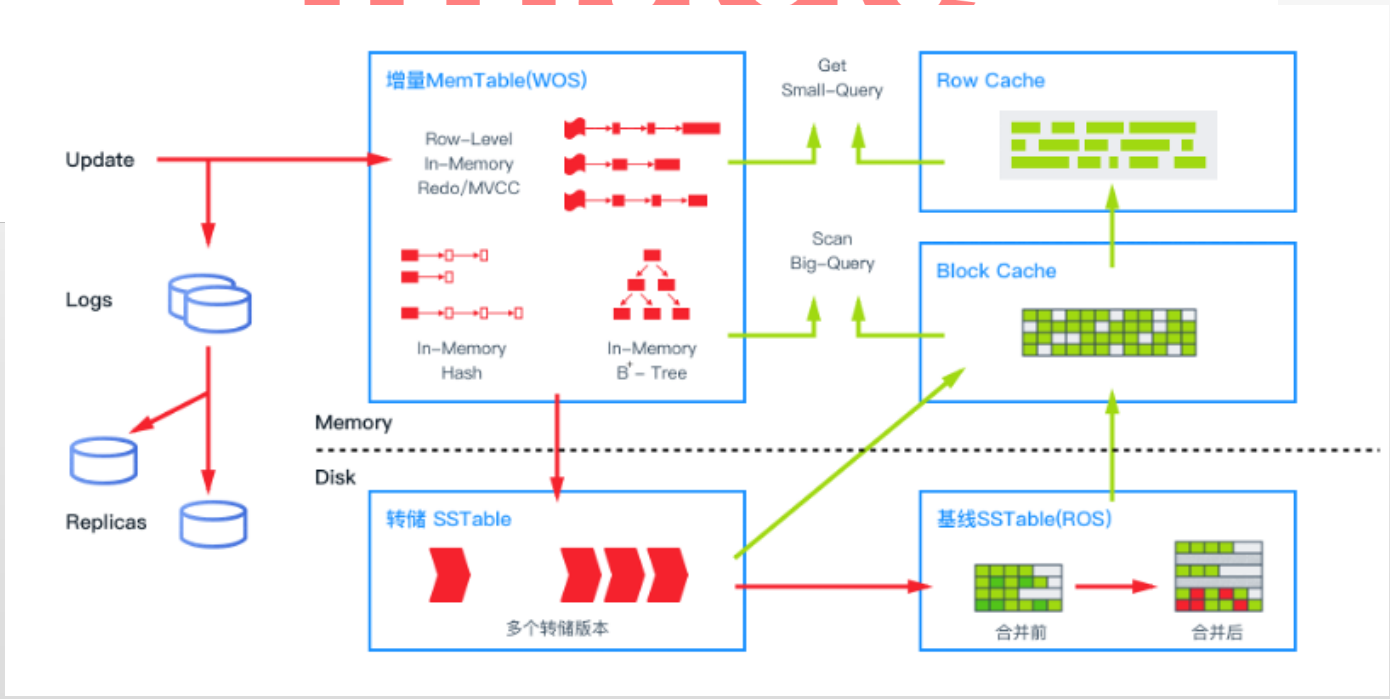
数据库	数据库分布原理	副本	负载均衡	特点
OceanBase	以分区为单位分片	多种副本， 每个副本位于不同的Zone	通过分区迁移实现负载均衡	对应用程序透明，支持范围， hash , list 等分区策略
TiDB	数据以 Region 为单位分布	默认 最少需要三个副本 默	动态调整Region范围，以 Region 为单位在存储节点间做迁移	数据在以 region 为单位分布的基础上支持分区表
KunlunDB	以分区为单位分片	可以以集群为单位自行定义副本数量及副本策略	根据负载，以分区为单位在存储节点间做迁移	对应用程序透明，支持range，hash , list 等分区策略

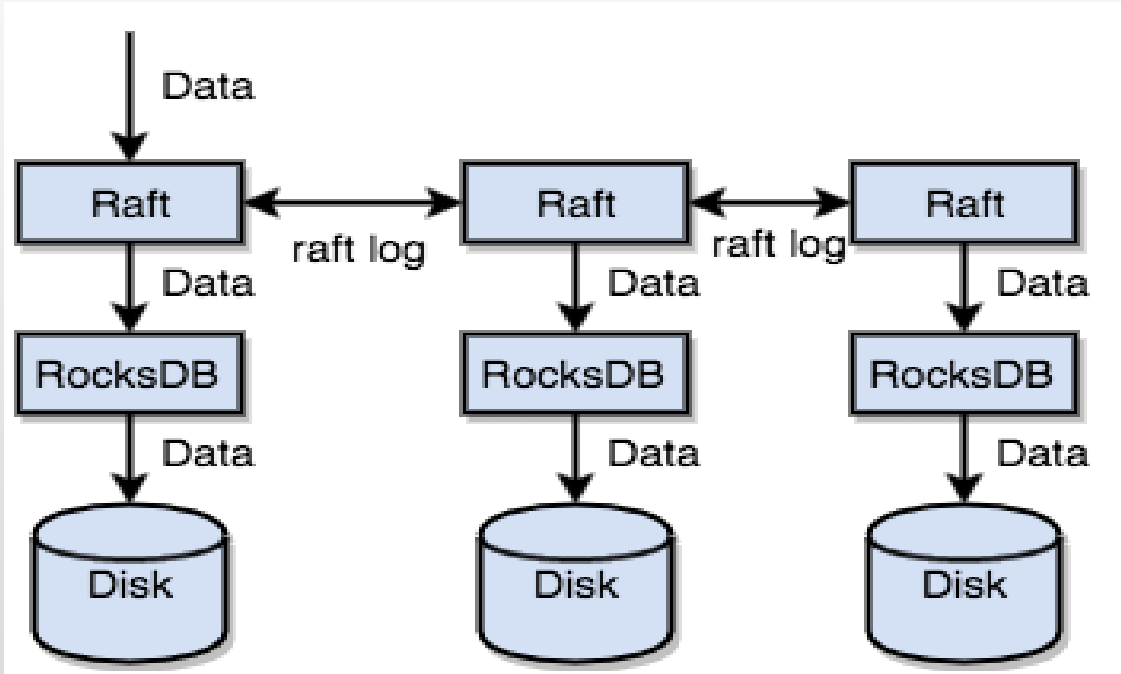
数据分片的好处

- 精准定位分区查询数据，不需要全表扫描， 提高效率
- 并行查询， 跨多个磁盘查询， 提高吞吐量
- 索引变小，读写效率提升
- 突破单节点数据库服务器的能力， 提高扩展性
- 提高大数据表的管理性， 如以月为单位删除或移动数据
- 基于数据库内核的分区分库技术，对应用内程序透明， 不会提高业务的复杂性

Oceanbase

Tidb

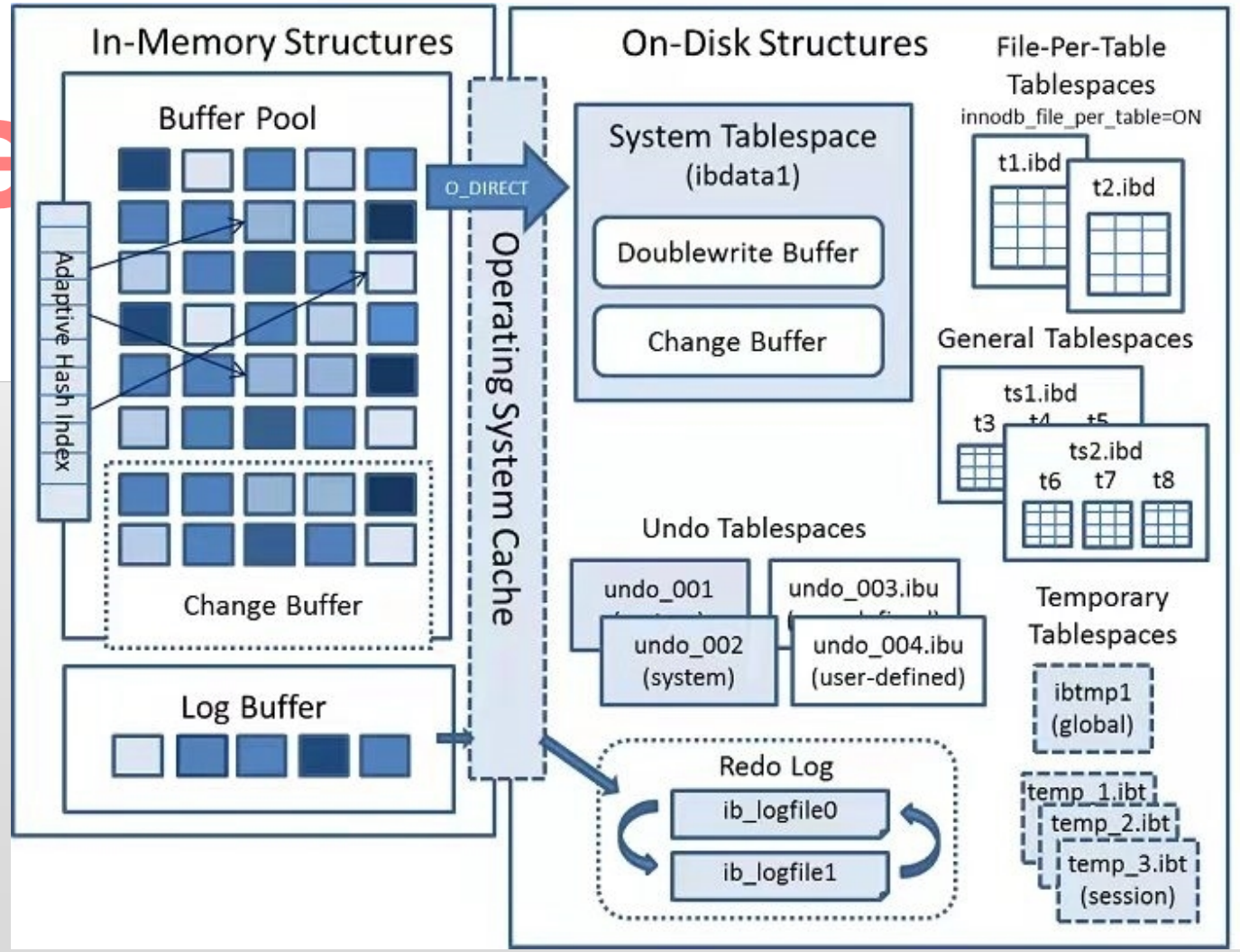




这种结构的写入，全部都是以Append的模式追加，不存在删除和修改

KunlunDB
(innodb)

大部分主流数据库都采用
B TREE 或B+ TREE 的存储
引擎， 如 Oracle， mysql 等

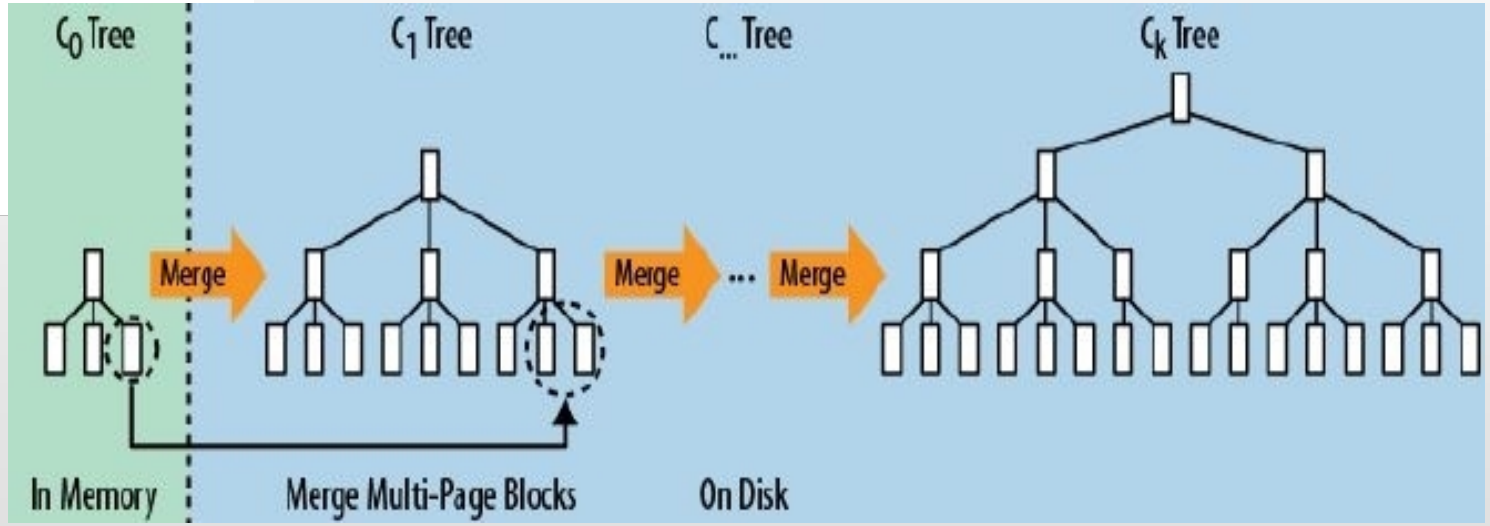


No Image

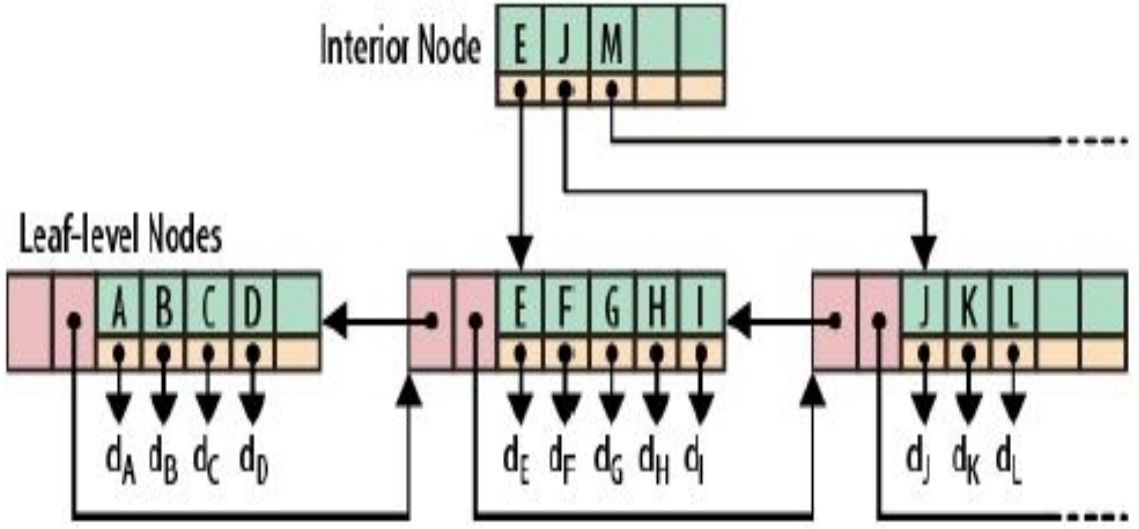
原理：把一颗大树拆分成N棵小树， 它首先写入到内存中（内存没有寻道速度的问题，随机写的性能得到大幅提升），在内存中构建一颗有序小树，随着小树越来越大，内存的小树会flush到磁盘上。当读时，由于不知道数据在哪棵小树上，因此必须遍历所有的小树，但在每颗小树内部数据是有序的.

优势：利用了磁盘的顺序写,写入速度相对比较块

劣势：一次查询可能需要多次单点查询，稍微慢一些



B+树是B树的一种变形，它把数据都存储在叶子节点，内部只存关键字（其中叶子节点的最小值作为索引）和孩子指针，简化了内部节点；B+树的遍历高效，将所以叶子节点串联成链表即可从头到尾遍历，



B+树的优&缺点：

1. 非叶子节点不会带上ROWID，这样，一个块中可以容纳更多的索引项，一是可以降低树的高度。二是一个内部节点可以定位更多的叶子节点。

2. 叶子节点之间通过指针来连接，范围扫描将十分简单，而对于B树来说，则需要在叶子节点和内部节点不停的往返移动。

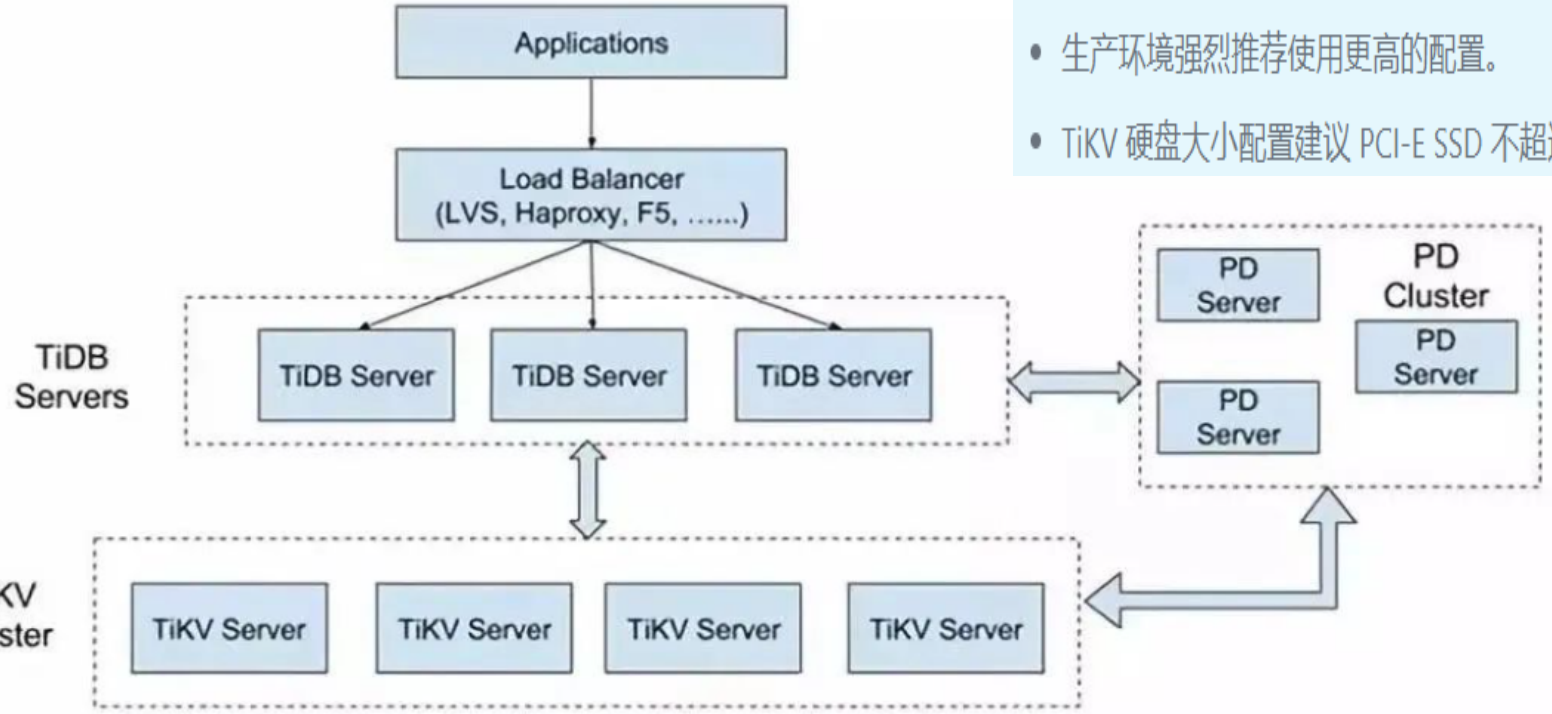
B+树最大的性能问题是会产生大量的随机IO，随着新数据的插入，叶子节点会慢慢分裂，逻辑上连续的叶子节点在物理上往往不连续，甚至分离的很远，但做范围查询时，会产生大量读随机IO

OceanBase 数据库的整体架构如下图所示。

服务器应满足的最低配置要求如下表所示：

服务器类型	数量	功能最低配置	性能最低配置
OCP 管控服务器	1台	32C，128 G，1.5 TB 存储（包含 OAT 与 ODC 所需资源）	32C，128 G，1.5 TB SSD 存储，万兆网卡（包含 OAT 与 ODC 所需资源）
OceanBase 计算服务器	3台	32C，128 G，1.2 TB 存储	32C，256 G，2 TB SSD 存储，万兆网卡
OBProxy 计算服务器	3 台，可复用 OBServer 服务器	4C，8 GB 内存，200 GB 存储	N/A
OAT 部署服务器	1 台，可复用 OCP 管控服务器	• X86 X64 架构：8C，16 GB 内存 • ARM aarch 64 架构：8C，16 GB 内存	N/A

- 生产环境中的 TiDB 和 PD 可以部署和运行在同服务器上，如对性能和可靠性有更高的要求，应尽可能分开部署。
- 生产环境强烈推荐使用更高的配置。
- TiKV 硬盘大小配置建议 PCI-E SSD 不超过 2 TB，普通 SSD 不超过 1.5 TB。

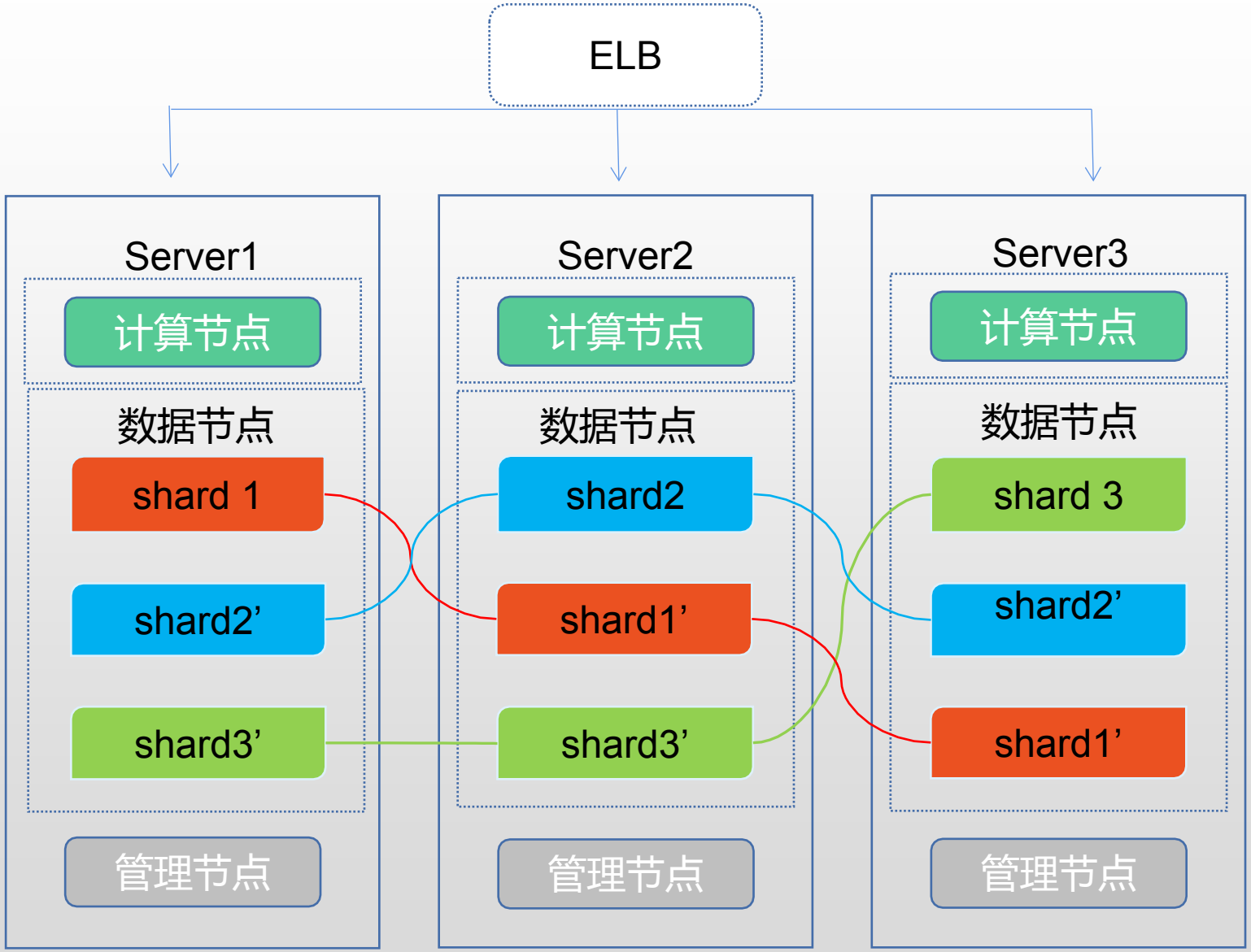


生产环境

组件	CPU	内存	硬盘类型	网络	实例数量(最低要求)
TiDB	16 核+	32 GB+	SAS	万兆网卡（2 块最佳）	2
PD	4核+	8 GB+	SSD	万兆网卡（2 块最佳）	3
TiKV	16 核+	32 GB+	SSD	万兆网卡（2 块最佳）	3
TiFlash	48 核+	128 GB+	1 or more SSDs	万兆网卡（2 块最佳）	2
TiCDC	16 核+	64 GB+	SSD	万兆网卡（2 块最佳）	2
监控	8 核+	16 GB+	SAS	千兆网卡	1

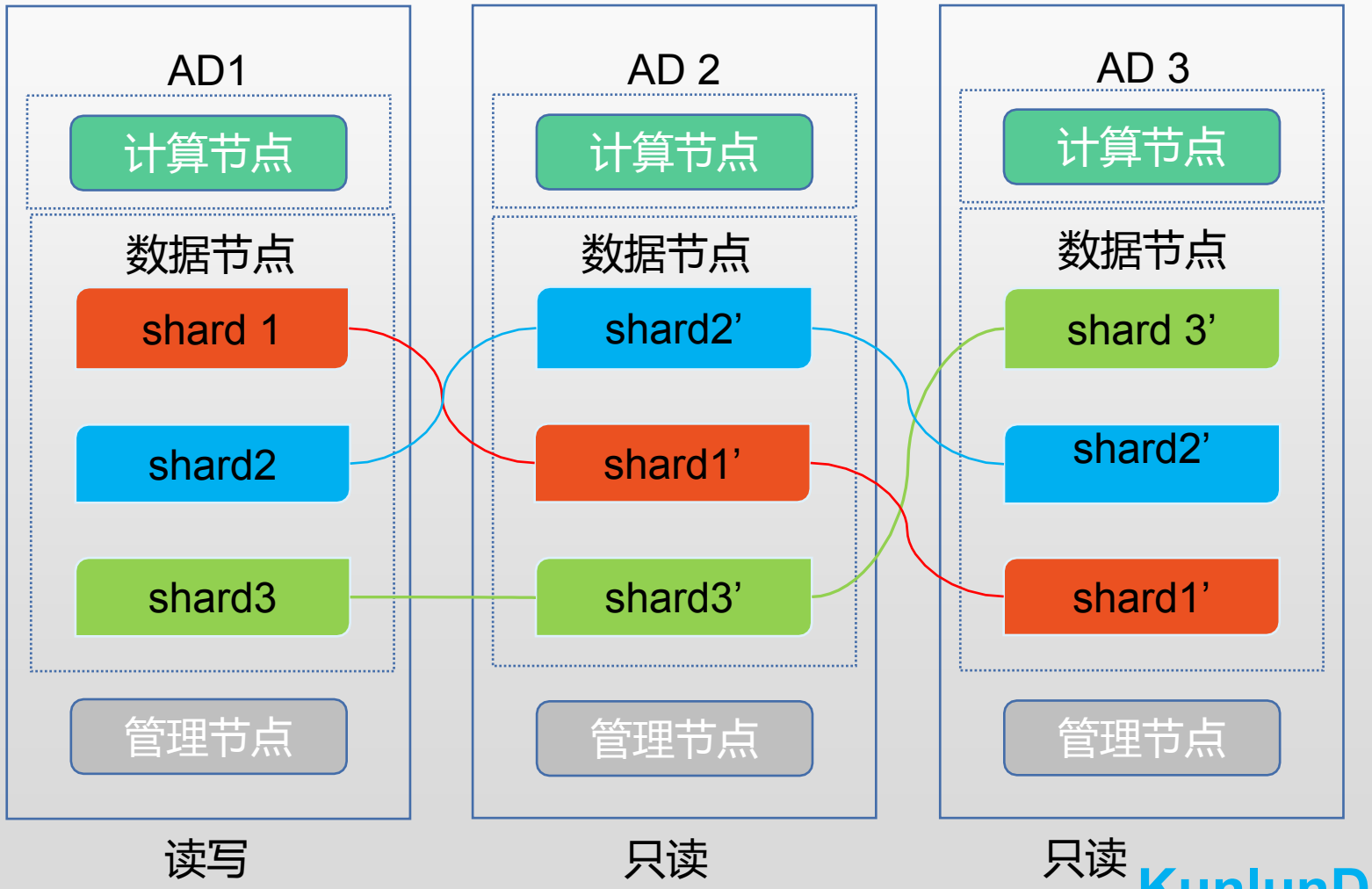
部署架构-KunlunDB对等部署方案

- 服务器：每台服务器是一个shard集群，集群数据库的主从副本散列在各个服务器上，实现负载均衡
- 服务器之间是对等关系，但每个分区数据的主从副本不存储在同一台服务器里
- 每台服务器部署全部的集群组件：计算节点， 存储节点和管理节点
- 功能最低基础配置(每台）：4 核 ,16G 内存, 普通硬盘
- 性能最低（每台）：16核 ,64G 内存, 普通硬盘或 SSD.



部署架构-多AD 部署方案

- AD（可用区）：每个可用区是一个服务器&网络等基础设施的共用区，处于同一个可用区的服务器具有一致的可靠性特性，每个可用区部署多个计算节点及多个存储节点。可用区可以是一个机架单位，或一个机房单位。同一AD内的服务器可以纵向扩展
- 不同的可用区是处于对等的单位的部署位置，数据在可用区之间互为冗余实现高可靠性
- 每个可用区的计算节点运行KunlunServer实列数目根据负载动态增减
- 数据节点运行kunlun-storage服务器进程，每个kunlun-storage实列是一个shard
- 管理节点运行meta-storage服务器进程作为元数据管理



易用性

KunlunDB	TiDB	OceanBase
基于Postgresql&mysql 入门无障碍，对于有 MySQL 基础的工程师即学即用	兼容Mysql 协议，较完整的文档体系， 通过培训后可以基于工具做日常管理	同时支持mysql &oracle 协议（租户二选一），文档系统完整，培训后可以基于工具做日常管理
Prometheus+ Grafana	Prometheus+ Grafana	黑屏工具，白屏工具
社区知识库+mysql 和PG 生态有丰富的故障处理知识库	TUG 社区获取知识	原厂支持
生态工具：丰富的第三方工具：percona 系列 mysql 工具，各种数据迁移工具如canal	TiUp,DM 等	

Q&A

谢谢观看~