DTCC

第十一届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2020

架构革新 高效可控

IT168.com ChinaUnix.net ITPUB

北京国际会议中心 | 2020/12/21-12/23

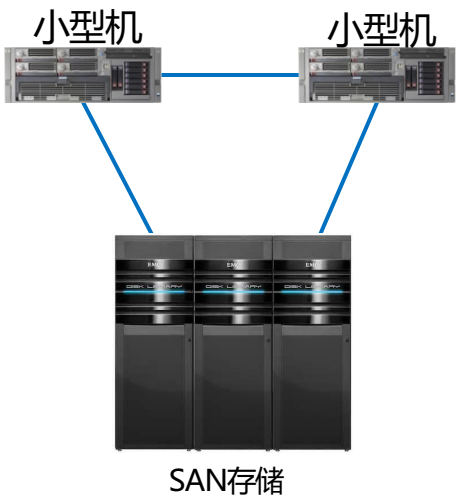
原生分布式数据库能力探讨

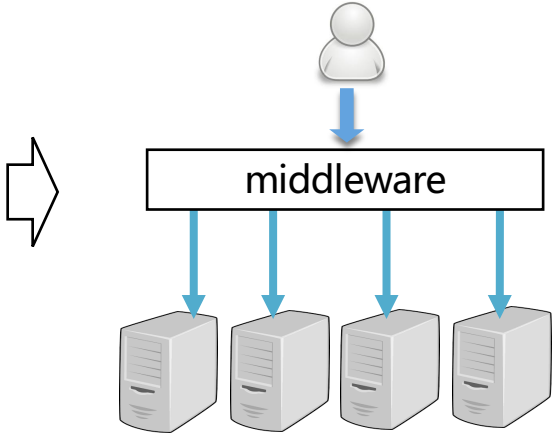
OceanBase原生分布式提供最佳使用体验

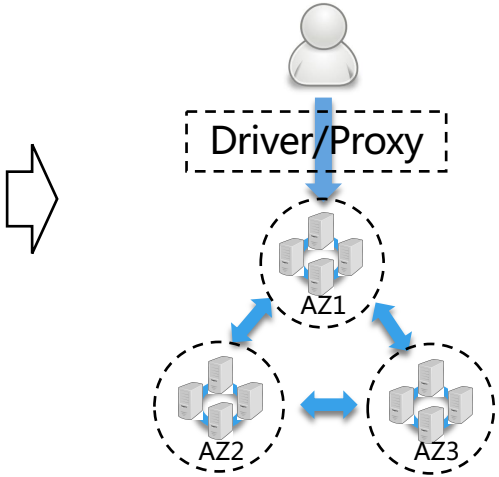
吴东昕（休伯）

蚂蚁集团 OceanBase资深解决方案架构师

数据库形态演进







- 传统架构数据库（Oracle、DB2）依赖高端硬件
- 事实上局限于单机处理能力，系统难于扩展
- 价格昂贵

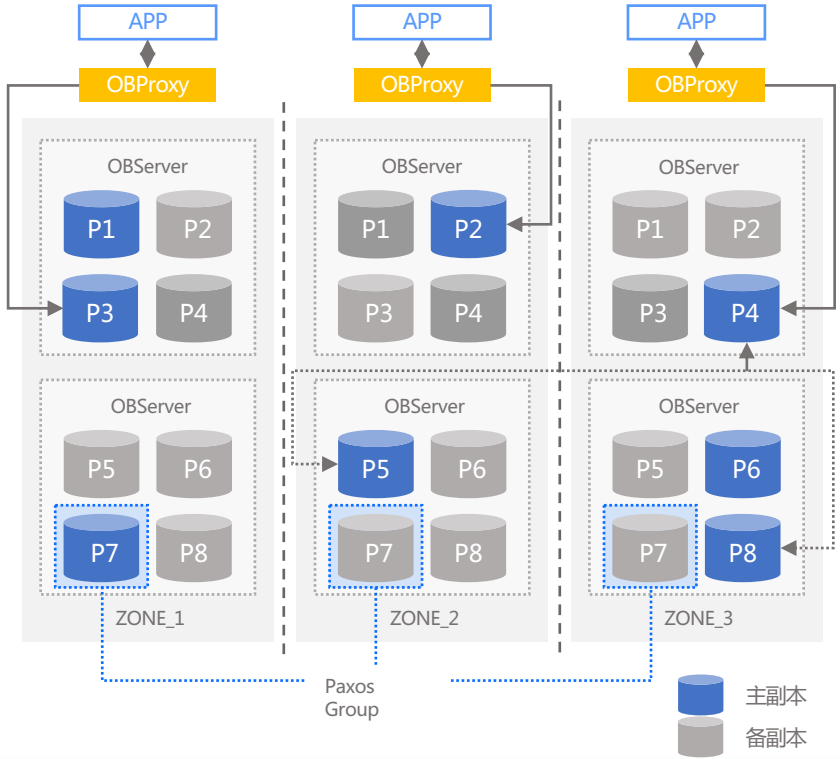
- 基于中间件的分库分表方案解决了扩展性的问题
- 亟待解决的问题
  - 跨库事务
  - 全局一致性
  - 负载均衡
  - 复杂SQL

- 分布式数据库解决了线性扩展问题
- 基于普通X86服务器，系统成本低
- 原生分布式查询支持
- 支持分布式事务，确保全局一致性
- 灵活的部署方式和负载均衡能力

OceanBase 数据库产品架构

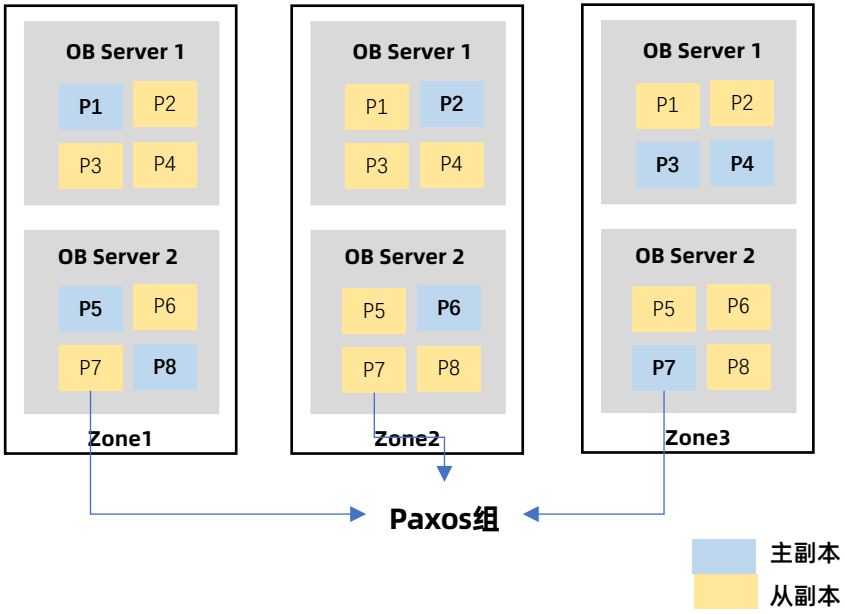
- Paxos协议 + 无共享架构 + 分区级高可用

- 多副本：一般部署为三/五个Zone，每个Zone由多个服务器节点（OBServer）组成
- 对等节点：每个节点均有自己的SQL引擎和存储引擎，自主管理各自承载的数据分区，TCP/IP 互通，协同服务，全部可读写，真正多活，没有硬件资源浪费
- 无需存储设备共享：数据分布在各个节点上，不基于任何设备级共享存储技术，不需要SAN网络
- 分区级可用性：分区是可靠性与扩展性的基本单元，自动实现访问路由、策略驱动负载均衡、自主故障恢复
- 高可用 + 强一致：多副本 + Paxos 分布式协议的高效高可靠工程实现， 确保数据（Clog日志）持久化在多数派节点成功

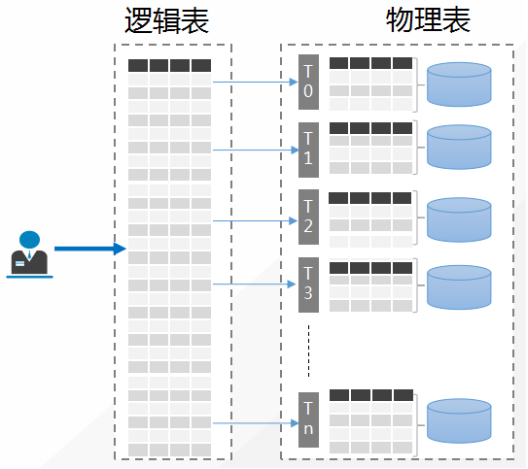


数据分布和多副本一致性协议

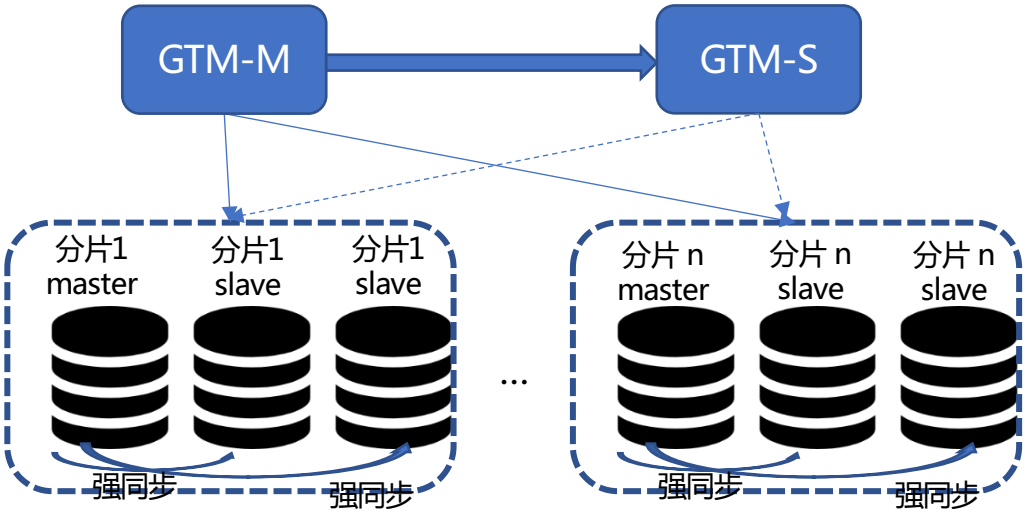
- **以表分区为单位组建Paxos协议组：**每个分区都有多份副本（Replica），基于clog物理复制，自动建立Paxos组，在分区级用多副本保证数据可靠性和服务高可用，数据管理更加灵活方便；
- **自动选举主副本**：OB自动生成多份副本，多副本自动选举主副本，主副本提供服务（如图黄色副本供应用访问，蓝色副本用于备份）；
- **表的分区方式：**完全由业务决定，以优化查询性能，利用分区修剪特性为依据；无需强制分布键做HASH分区



基于开源数据库单机拆库拆表非原生分布式数据库的特点



- 大表必须强制和分布式节点的拓扑进行强绑定
  HASH分布+分区
- 分区数量和节点数量相关



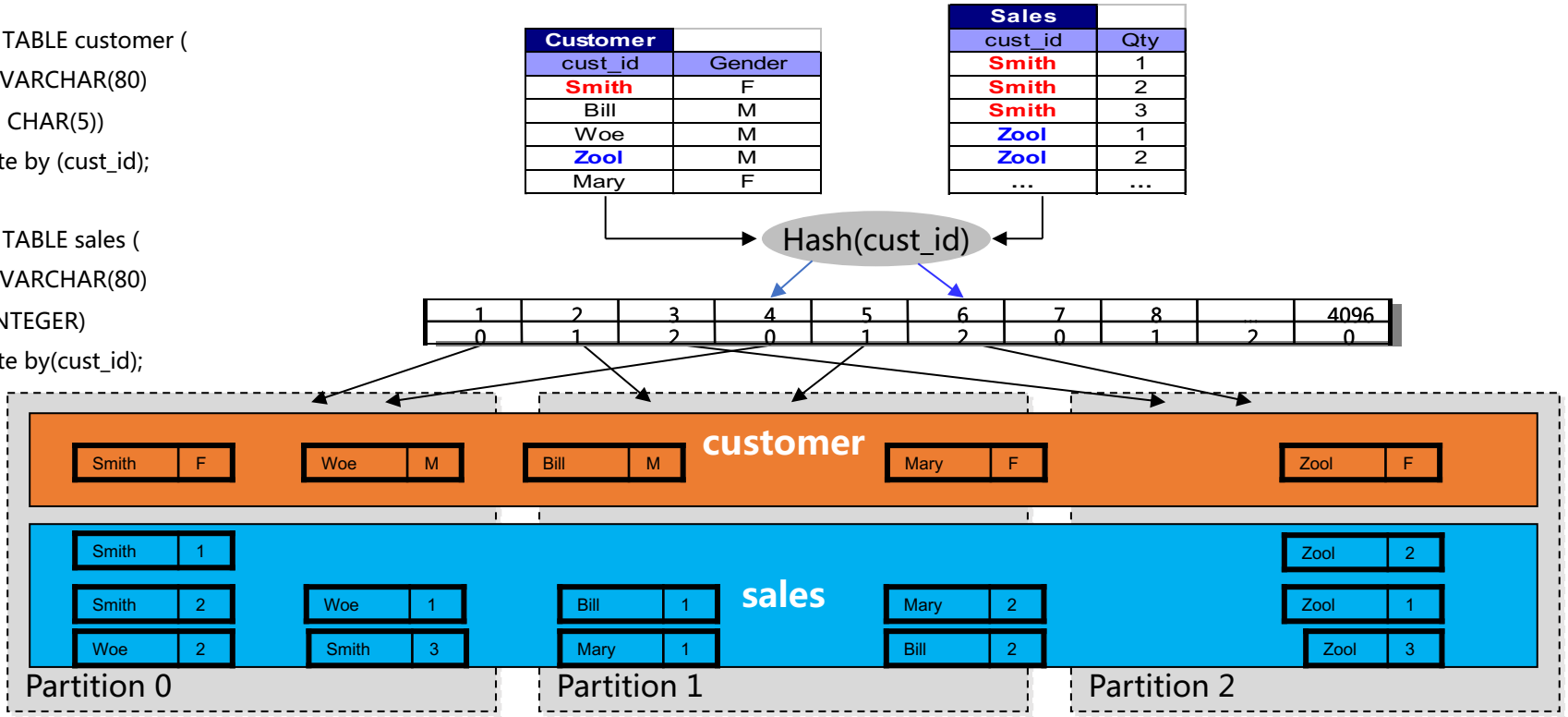
- 基于数据库分片（拆出来的库）的主从复制
- Slave节点计算能力仅仅用于接收日志和应用，硬件资源浪费

拆库拆表非原生分布式数据库的强制Hash分布

- 深度应用侵入式

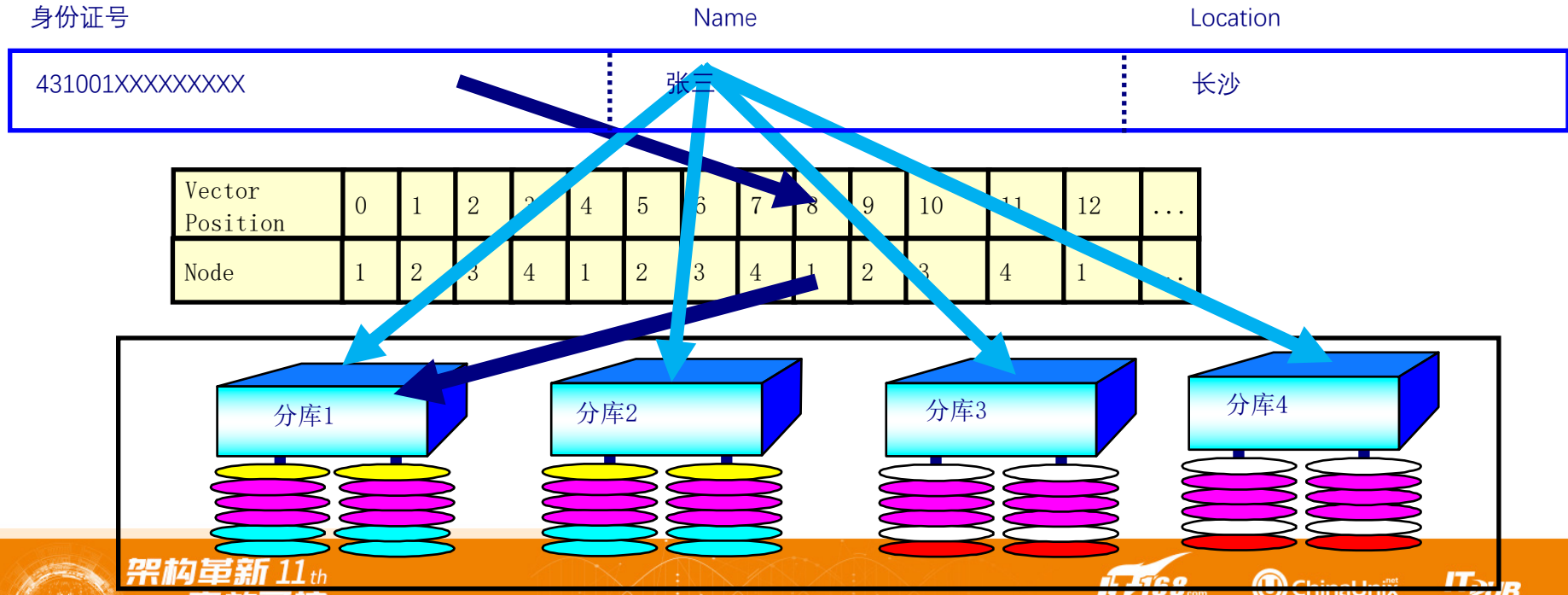
```
CREATE TABLE customer (
 cust_id VARCHAR(80)
,gender  CHAR(5))
Distribute by (cust_id);

CREATE TABLE sales (
 cust_id VARCHAR(80)
,qty     INTEGER)
Distribute by(cust_id);
```

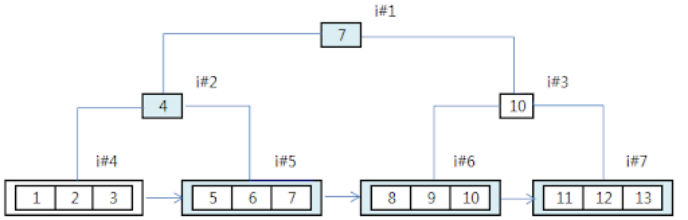


拆库拆表非原生分布式数据库访问局限

- 采用身份证号作为查询条件，可以确定确定记录所在节点
- 但是如果采用Name作为查询条件，或者非等值查询，必须广播到所有分区操作
  - 对于OLTP SQL，增加CPU消耗，并发度难以通过扩容分片数量实现



分库分表分布式面对非分区键查询、非等值查询条件的不可突破的障碍



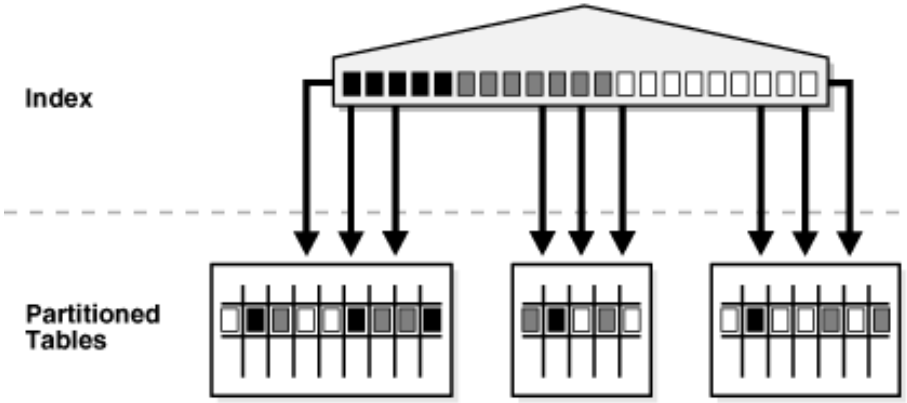
相同数据量多节点广播查询CPU消耗

- OLTP类查询一定是通过B（+/-）树的索引扫描定位数据获取数据
- 如果表中有n行数据，时间复杂度为O(log(n))
- 对于**强制分布键的拆库拆表架构**，如果不采用分布键进行索引扫描或者非等值扫描，必须广播到所有K个节点(k<<n)
  - 时间复杂度变成O(log(n/k))=O(log(n))-O(log(k))，在k<<n情况下和O(log(n))一致，因此对于查询加速可以忽略
  - 会发现在于CPU消耗变成k*（O(log(n))-O(log(k))），在k<<n**情况下几乎随着节点数增加而正比增加，逻辑读也随节点数量增加增加**
  - **这意味着增加节点带来额外的计算能力并都会被广播查询所消耗**
  - 如果采用一致性HASH分布，哈希桶数量m, m>>k，非等值查询CPU消耗变成m*（O(log(n))-O(log(m))），更加严重恶化CPU消耗
- 结论：对于大并发非分布键等值查询，拆库拆表非原生分布式数据库架构只会消耗更多CPU而导致无法通过增加节点来增加并发量

OceanBase通过全局索引解决广播问题


```
obclient> CREATE TABLE t1(a int PRIMARY KEY, b int, c int) PARTITION BY hash(a) partitions 5;

obclient> CREATE INDEX gkey ON t1(b) GLOBAL PARTITION BY range(b) (
  partition p0 VALUES less than (1),
  partition p1 VALUES less than (2),
  partition p2 VALUES less than (3)
);
```

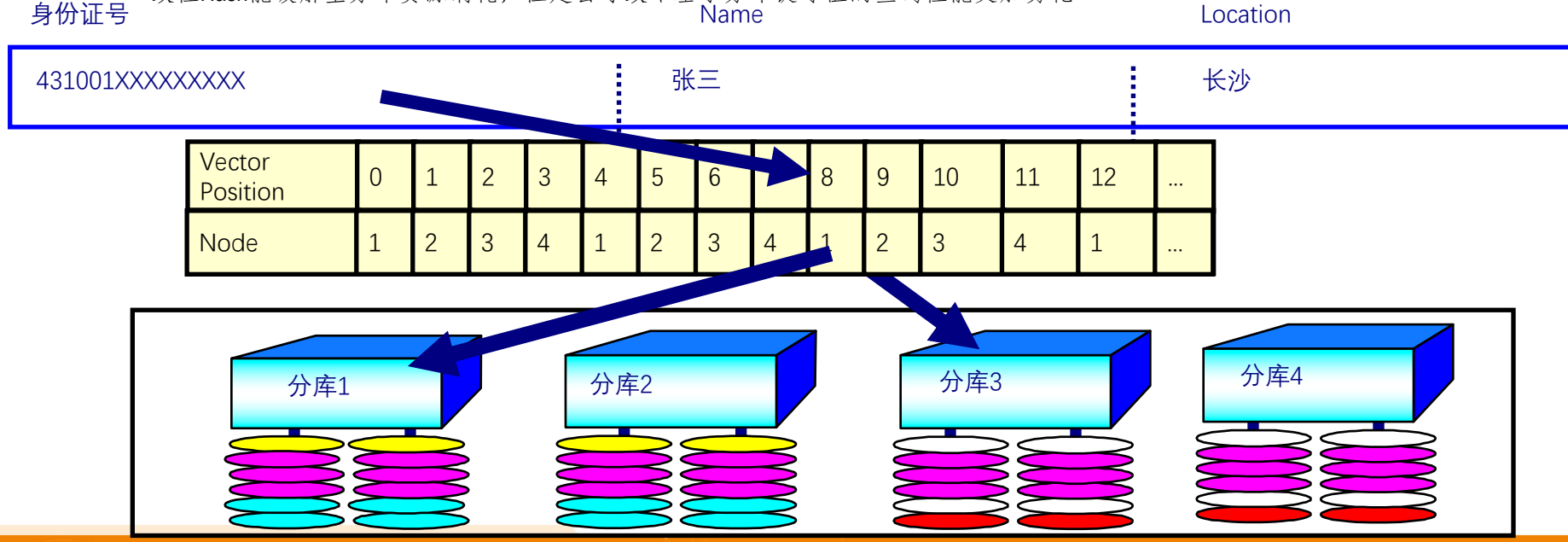


- 索引的分区和表的分区解耦
- 无需广播到所有的索引分区，进行索引扫描即可获得对应数据所在表的主键再获取数据
- 时间复杂度保持O(log(n))
- 和分区数量、节点数量无关

拆库拆表非原生分布式的在线扩容困难性

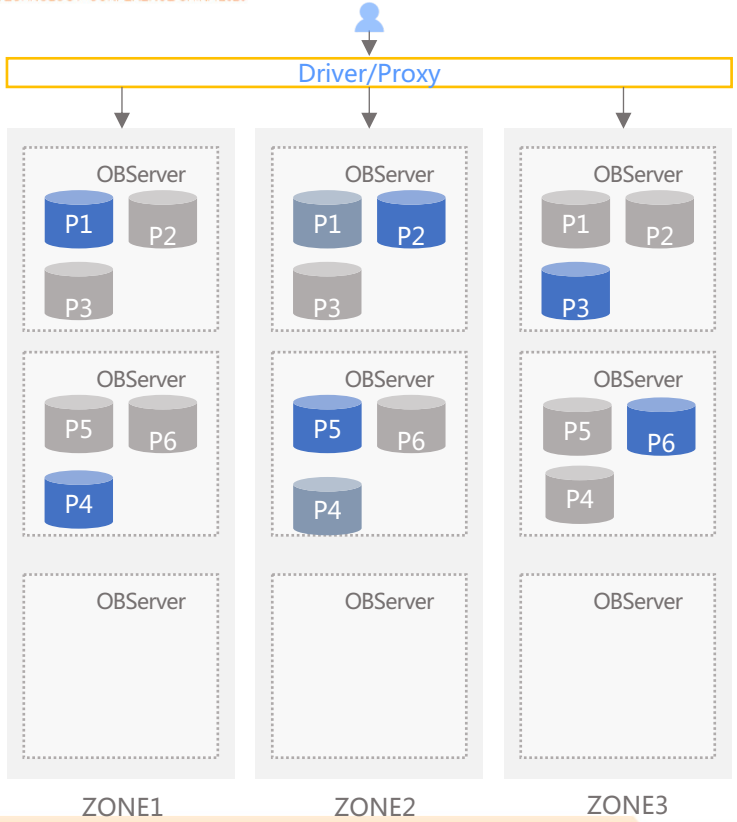
➢ 分片映射表

- 确定记录所在分片
- 增加节点（分片）需要重新hash每一行记录可以通过工具重新分布数据的位置，分布过程可能表锁定，通过Shadow Table等技术可以避免锁表，但是仍然需要Rehash
- 一致性Hash能缓解重分布资源消耗，但是会导致不基于分布键等值的查询性能更加劣化



OceanBase在线扩容ZONE内增加OBServer

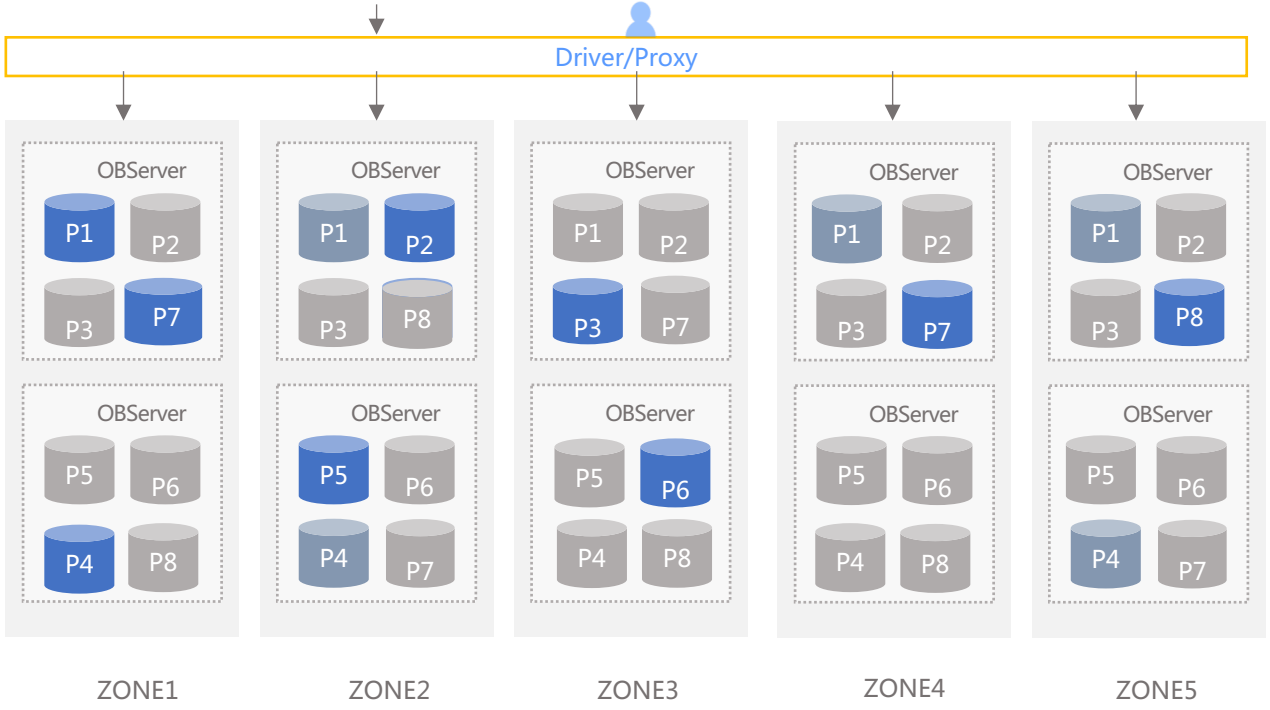
应用透明无感知



- ✓ 增加ZONE内部的OBSERVER硬件
- ✓ 对应的租户增加UNIT数量
- ✓ 系统自动将分区reblance到新增加的Observer
  - ✓ 物理拷贝，单节点>500MB/s
  - ✓ 无需重新hash每条记录
  - ✓ 应用无感知
- ✓ ZONE内减少OBSERVER也同样也能应用透明平滑支持

OceanBase在线增加ZONE实现扩容

应用透明无感知



- ✓ 增加副本（ZONE）数量
- ✓ 无需调整租户UNIT数量
- ✓ 在新的集群重新自动选主
  - ✓ 不同ZONE优先级可以手工干预
- ✓ 业务压力分散到新的ZONE当中
- ✓ 分区级别级别物理拷贝
  - ✓ 单节点>500MB/s
- ✓ 应用透明
- ✓ 对于缩容（减少ZONE数量）同样平滑支持

库内分布式事务（业务透明）

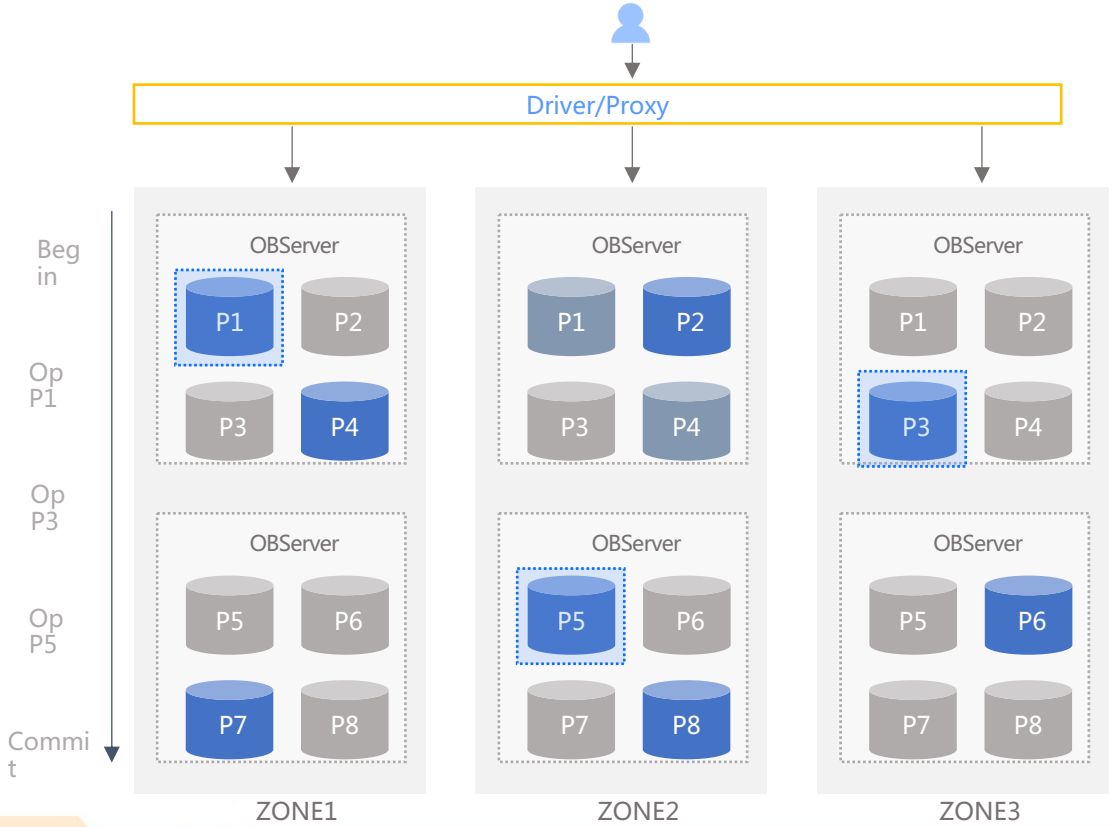
全局时间戳

内部自动两阶段提交

异步提交提升性能

TableGroup减少分布式事务

继续对外再提供XA接口

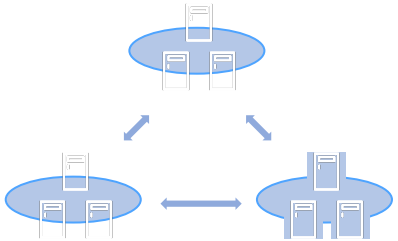


OceanBase原生分布式VS基于拆库拆表分布式数据库

项目	OceanBase原生分布式	拆库拆表分布式
Shading/distribution Key强制要求	无	必须
表分区形式	基于应用优化的设计	强制基于数据Hash分布键的设计
应用性能优化要求	基本和单机数据库一样，利用table group应用非侵入式地优化	OLTP类SQL的Where条件必须带shading key等值条件，RDBMS退化成KEY-VALUE
SQL透明性	和单机数据库基本一样，利用全局（分区）索引实现	跨分片访问数据有各种限制，缺乏全局索引等特性
扩容缩容	基于分区级别的物理迁移，单机500MB/s，TableGroup级别同步迁移，保持Co-location特性	基于行的hash重新分布，需要消耗大量算力，很难带业务进行，几十MB/s，迁移是还会破坏Co-location特性

OceanBase内置的业务连续性

OceanBase



写事务到达超过半数库

少数库异常不影响业务

两地三中心多活

灰度升级

基于Paxos协议的典型三副本部署：

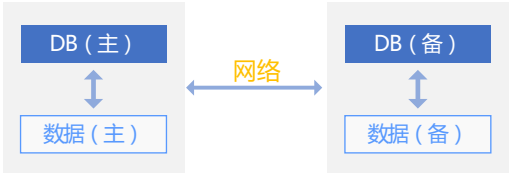
- 数据强一致性
- 持续可用
- 主备自动切换，对上层业务透明
- 单机、机房、城市级故障：自动故障切换，不停服务，不丢数据

RPO = 0(少数成员故障)

RTO < 30秒

超越国标灾难恢复能力6级

传统数据库



主库+备库

跨站点切换对应用很难透明

无内置自动选取正确的主副本功能

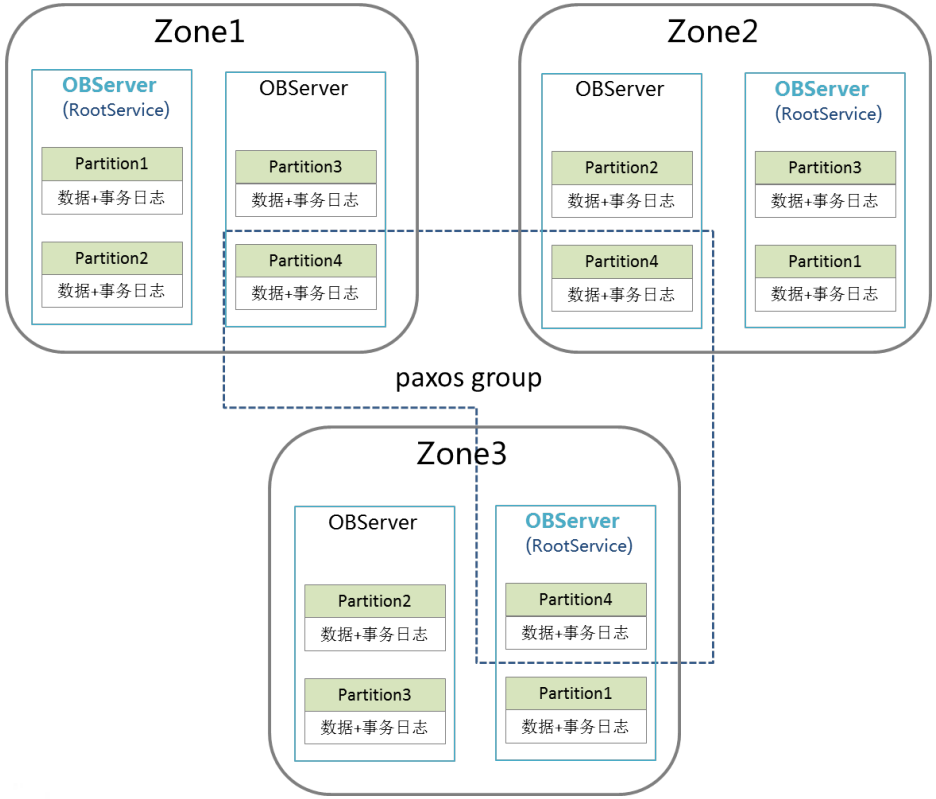
RTO/RPO与灾难恢复能力等级关系

（GB/T 20988-2007）

灾难恢复能力等级	RTO（回复时间目标）	RPO（恢复点目标）
1	2天以上	1天至7天
2	24小时以上	1天至7天
3	12小时以上	数小时至1天
4	数小时至2天	数小时至1天
5	数分钟至2天	0至30分钟
6	数分钟	0

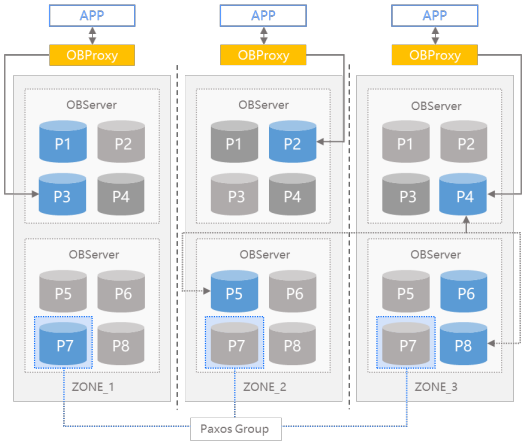
OceanBase保证数据库、业务高可用的基础

- **复制对象：数据分区**
  - 支持数据分区（partitioning）
  - 各分区独立选主、写日志
- **高可用&强一致**
  - PAXOS协议保证物理数据（日志Clog）同步到多数机器
  - 故障时自动切主
  - 自动切换功能内置于数据库内部，无需借助外部工具或者外挂
  - 复制的强校验clog
  - 本地数据读取强校验



总结：原生分布式数据库OceanBase带来使用上的极大便利性

OCEANBASE



- ✓ 作为独创的原生分布式数据库，OceanBase在让您享受分布式数据库的横向扩展红利的同时保持应用透明、低侵入特性，您可以像使用单机数据库一样来使用OceanBase开发应用
- ✓ 内置的基于分区的物理复制和Paxos多活保证了节点物理资源的充分利用和真正多活，高可用切换的完全透明特性
- ✓ 基于分区的数据分布真正让在线扩容缩容成为现实

OceanBase 微信公众号

OceanBase 官网

THANKS