Alibaba Group 阿里巴巴集团

OceanBase
A Powerful Database System

OceanBase 0.5版本开发实践

杨传辉，2014-04
rizhao.ych@alipay.com

Alibaba Group 阿里巴巴集团

OceanBase介绍

- 2010.5 ~ ...
- 服务器总量：1000+
- 业务情况
  - 业务数：50+
  - 三淘：收藏夹、天猫评价、店铺装修、P4P报表/量子统计、库存中心、SNS、淘足迹、推荐系统，。。。。
  - 支付宝：会员视图、移动通知、钱包公众账号、实时数据平台，。。。。
  - B2B：B2B收藏夹
- 单集群
  - 单表最大记录数：OLTP（167亿）/ OLAP（2130亿）
  - 单集群最大TPS / QPS：6.8万TPS / 5.1万 QPS
  - 单集群最大一天写入行数 / 总量：7亿 / 246G

OceanBase A Powerful Database System

Alibaba Group 阿里巴巴集团

议程

- 整体架构回顾
- 0.5架构升级
- 0.5开发测试
- 经验与后续规划

OceanBase A Powerful Database System

Alibaba Group 阿里巴巴集团

议程

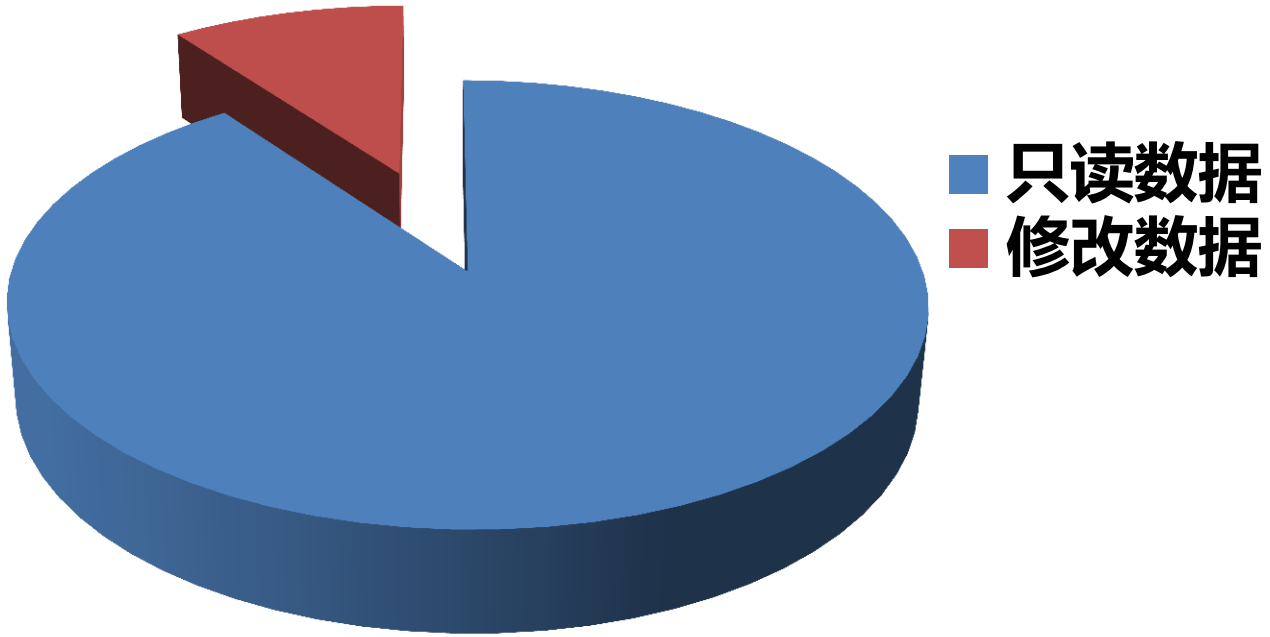
- 整体架构回顾
- 0.5架构升级
- 0.5开发测试
- 经验与后续规划

OceanBase A Powerful Database System

Alibaba Group 阿里巴巴集团

设计假设

- 数据库中每天修改的数据只占整体数据的一小部分

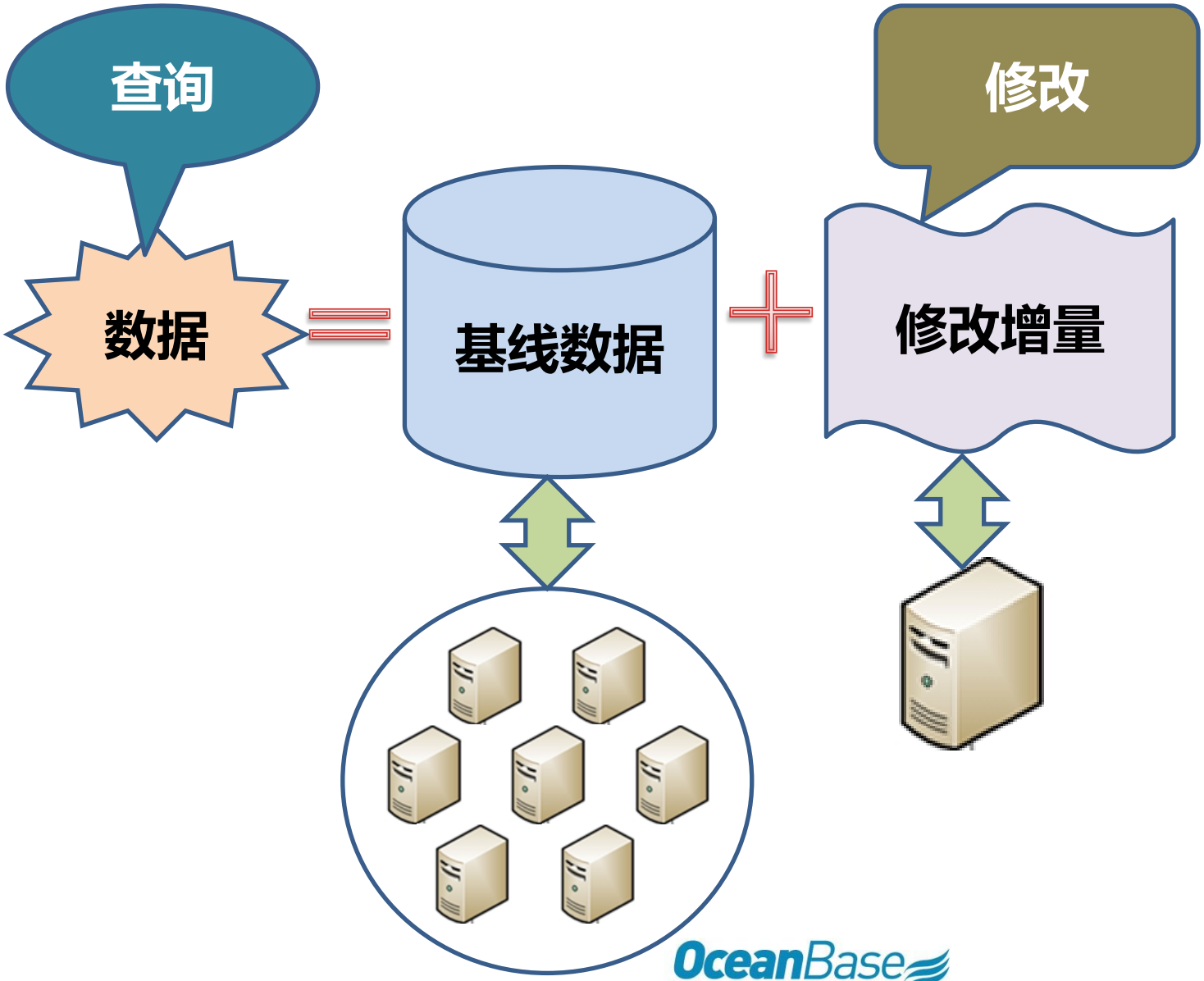


- 假设一天产生2亿笔事务，每笔事务需要10个DML语句，平均一个DML语句 500个字节:
  - 一天修改量: 0.2G * 10 * 500= 1 TB
  - 假设单个节点存储200G，总共只需要5个节点

OceanBase A Powerful Database System

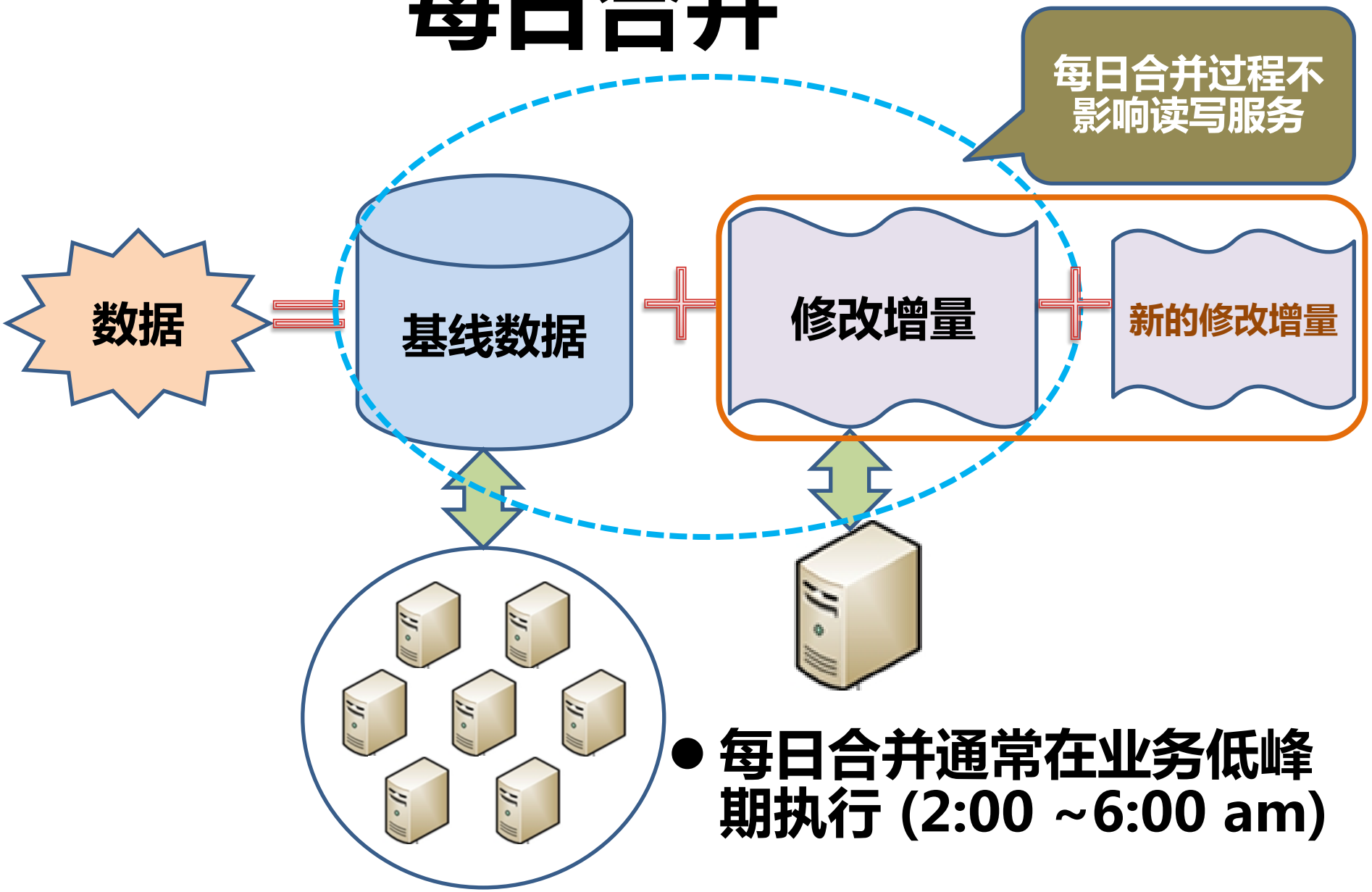
Alibaba Group 阿里巴巴集团

分布式存储引擎



OceanBase
A Powerful Database System

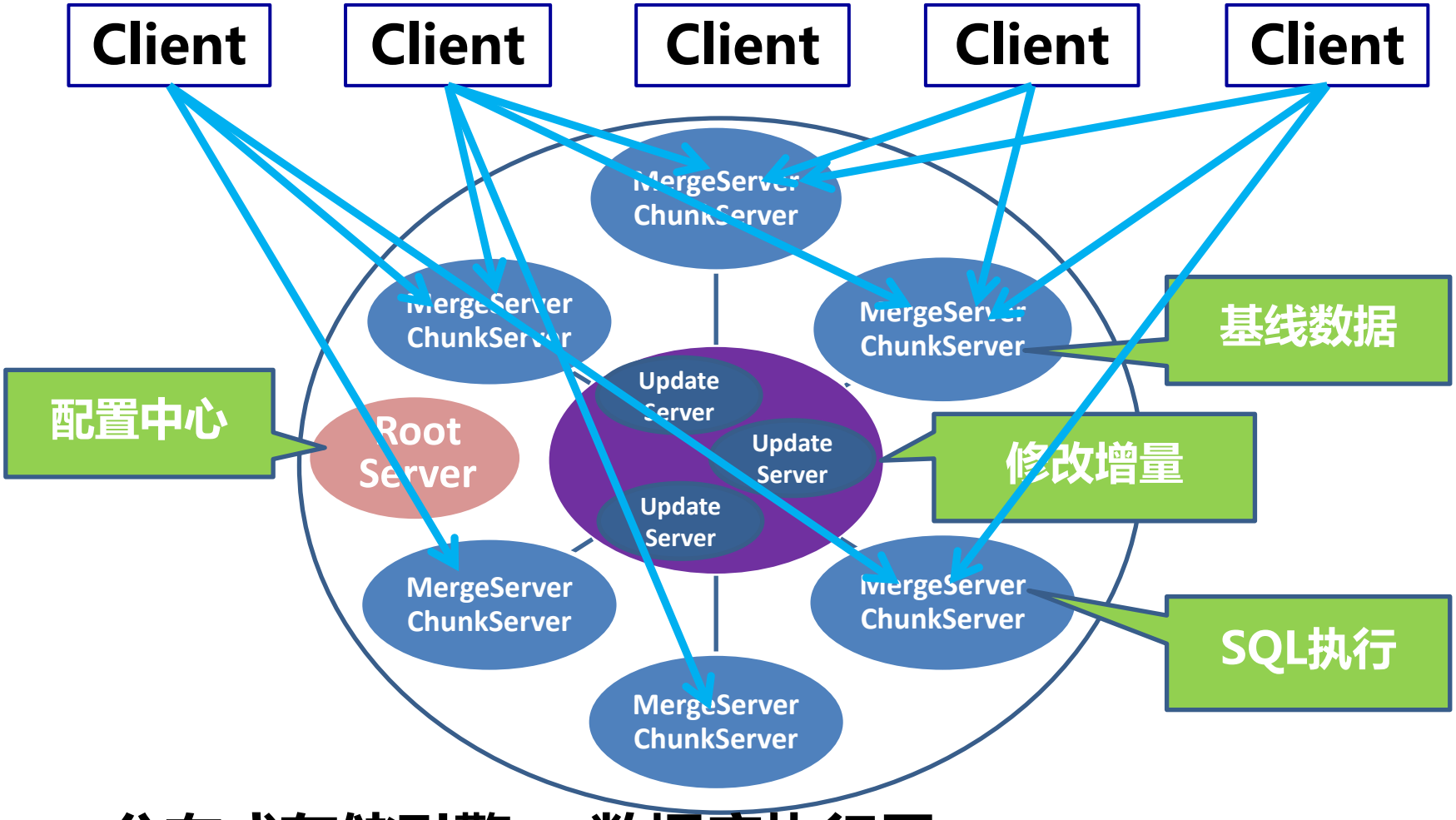
每日合并



- 每日合并通常在业务低峰期执行 (2:00 ~6:00 am)

Alibaba Group 阿里巴巴集团

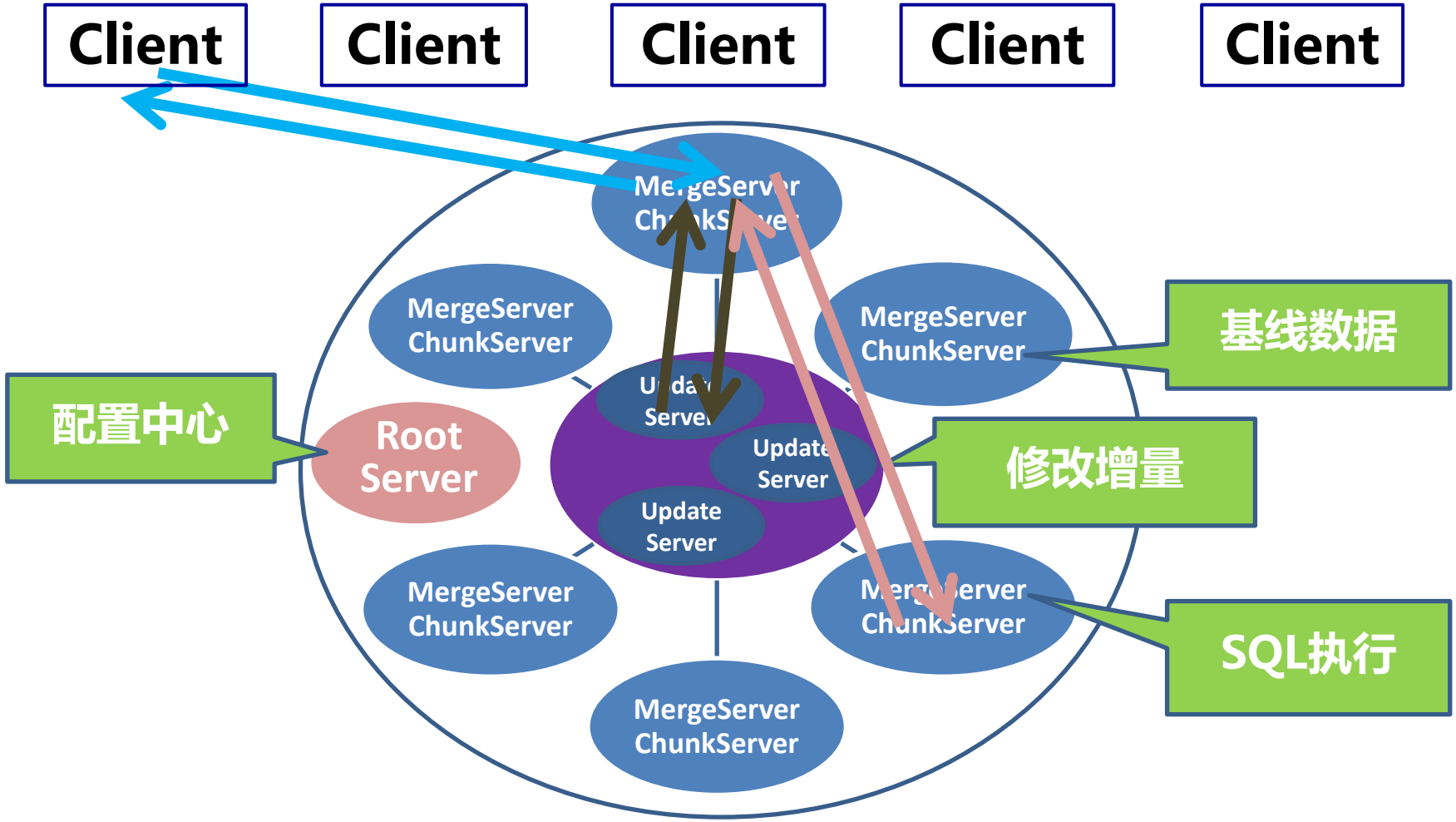
整体架构



- 分布式存储引擎 + 数据库执行层
- Client与MergeServer之间采用MySQL协议

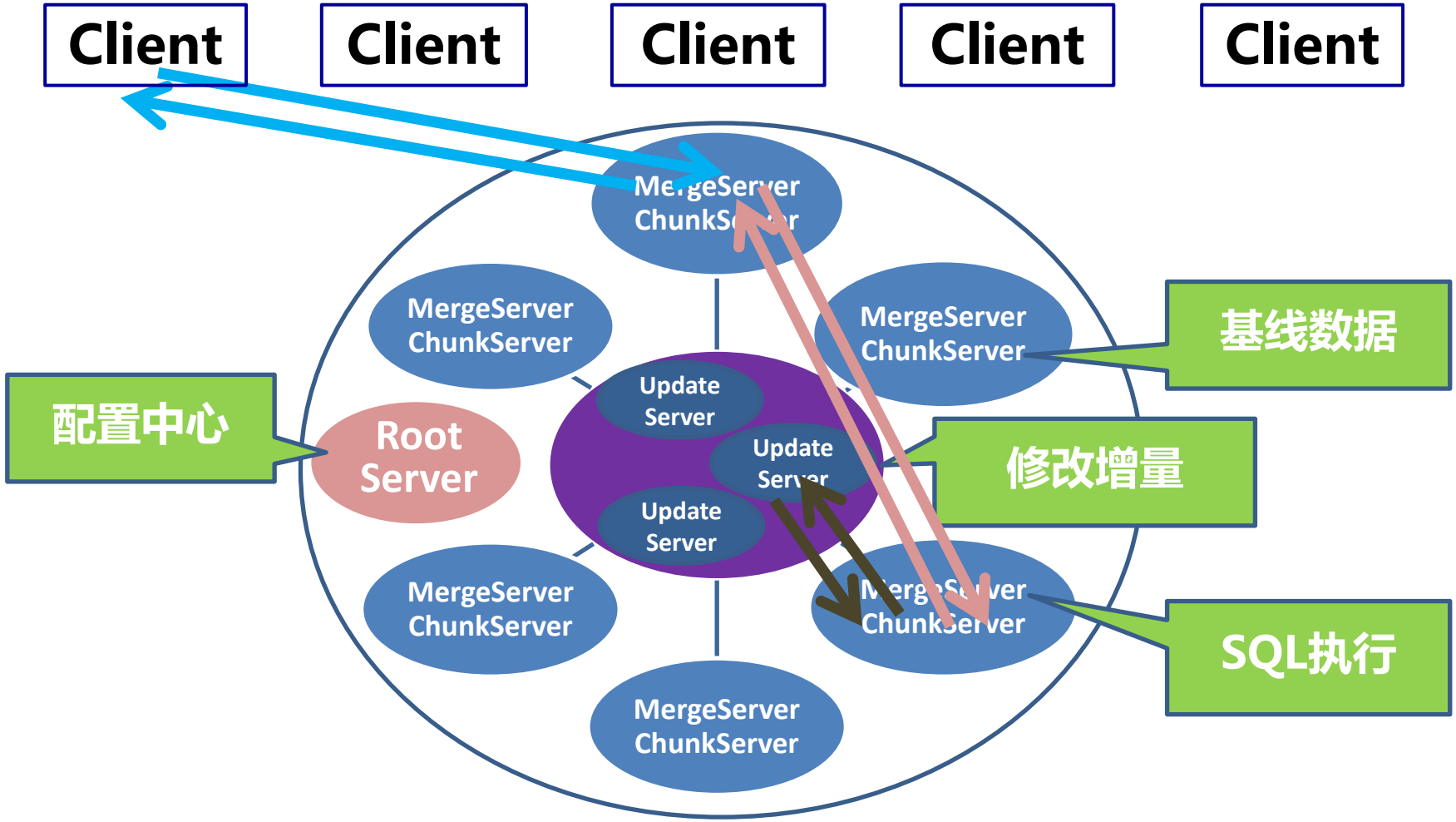
OceanBase A Powerful Database System

写事务



- 1. MS请求CS得到基线数据
- 2. MS将基线数据与物理执行计划发送给UPS

读事务



- 1. MS将读事务物理执行计划发送给CS
- 2. CS请求UPS获取修改增量并与基线数据融合

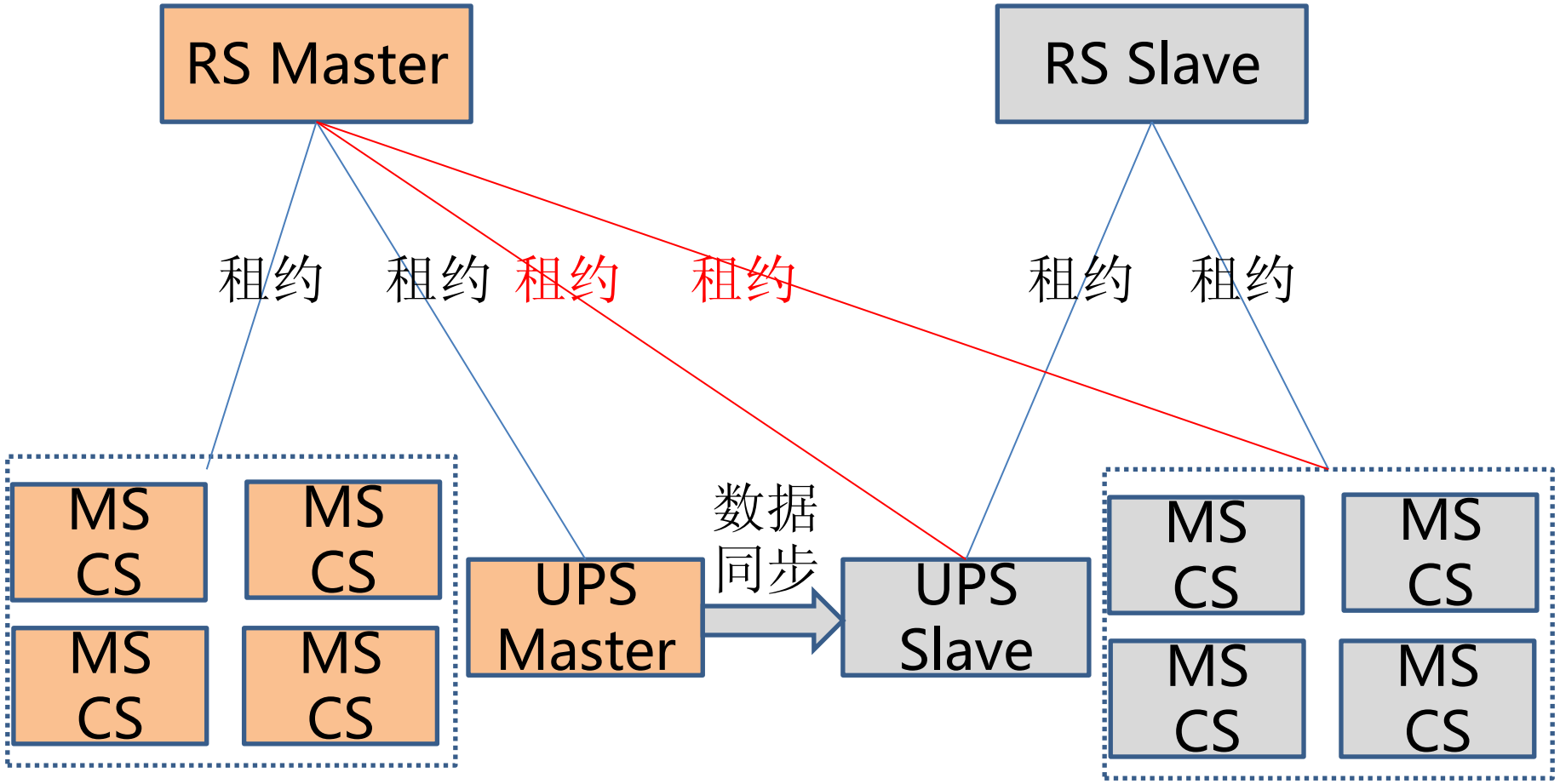
Alibaba Group 阿里巴巴集团

议程

- 整体架构回顾
- 0.5架构升级
- 0.5开发测试
- 经验与后续规划

OceanBase A Powerful Database System

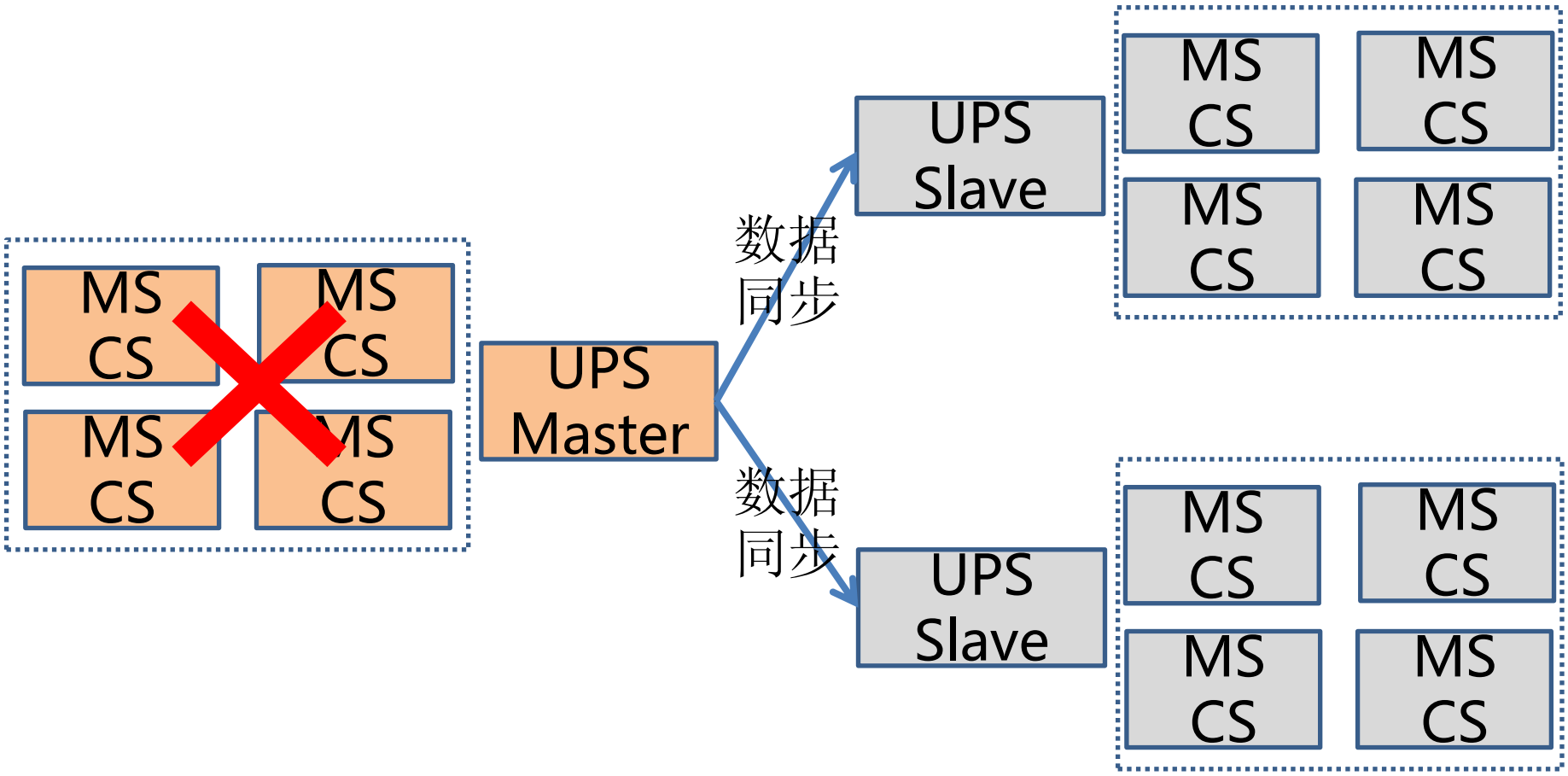
大集群



- **自动错峰合并：每日合并将流量自动切走，合并完切回**

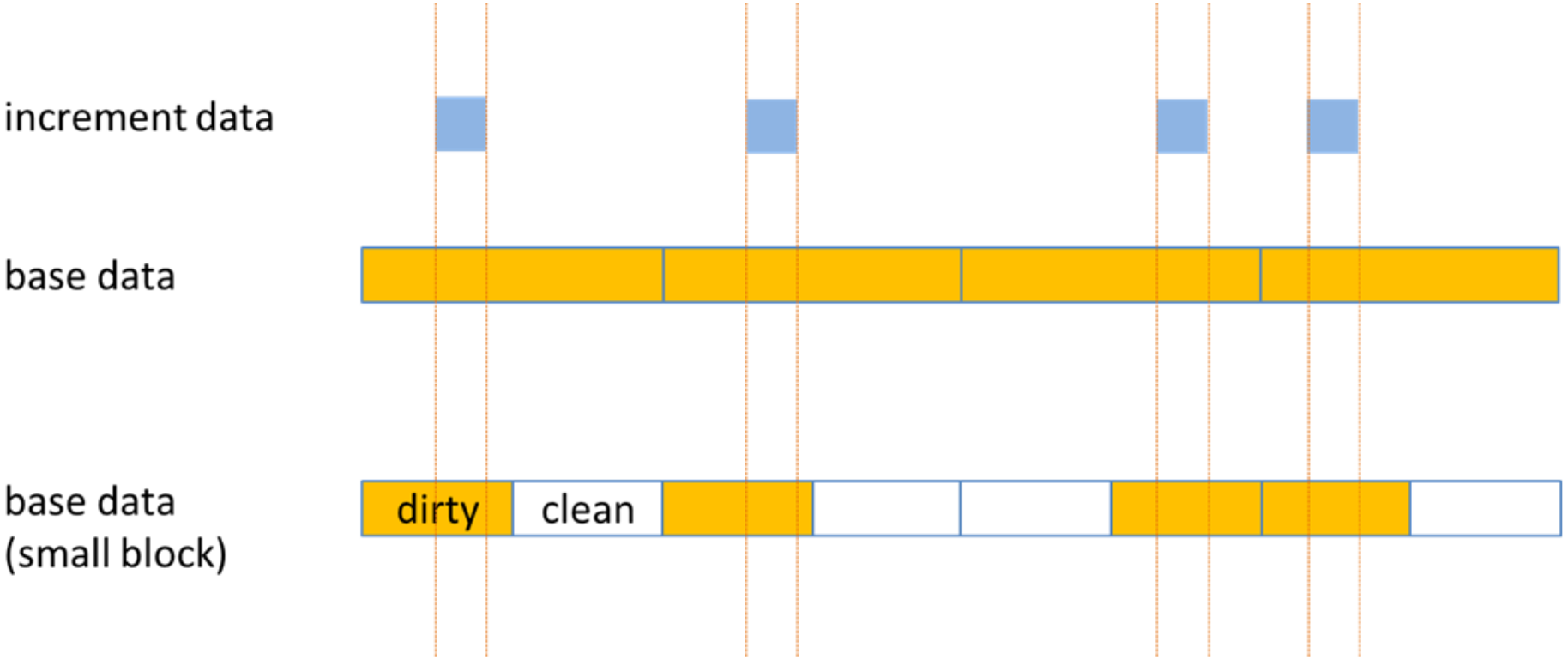
主备同步

- 两机房主备 => 三机房选举



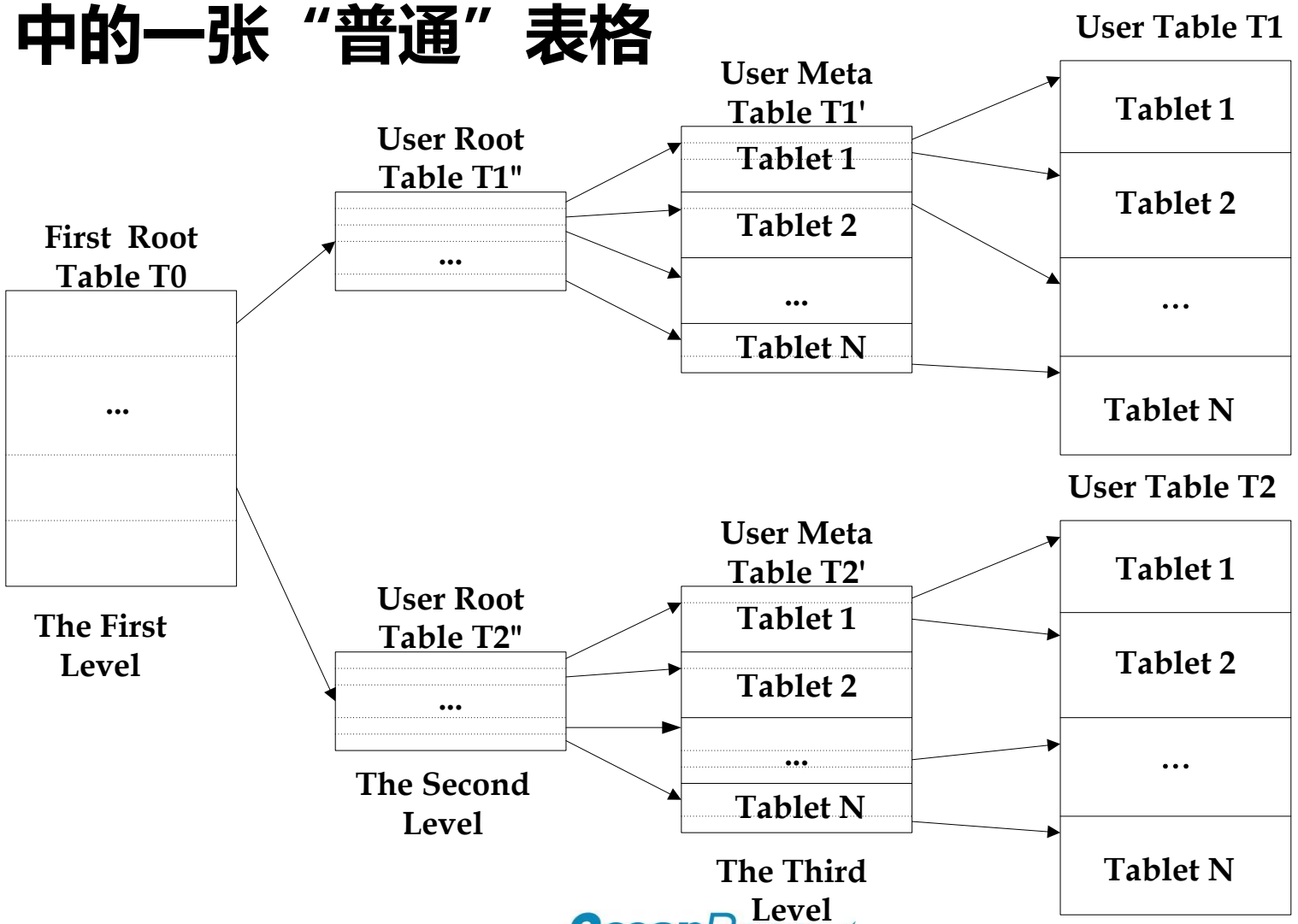
新存储系统

- 分块单元由Tablet(64~256MB) => 2MB宏块



元数据分布式化

- 元数据：RootServer内存 => 分布式存储引擎中的一张“普通”表格



Alibaba Group 阿里巴巴集团

SQL功能增强

- SQL模块重构
  - 引入类MySQL“range analysis”
  - 表达式内存管理、SQL模块对象管理代码重构
- SQL功能增强
  - 数据类型：Number
  - 操作：多行DML，nest loop join
  - 函数：日期时间函数，nvl，case...when等
  - 更多系统表
- 复杂SQL功能
  - 二级索引
  - 复杂OLAP运算

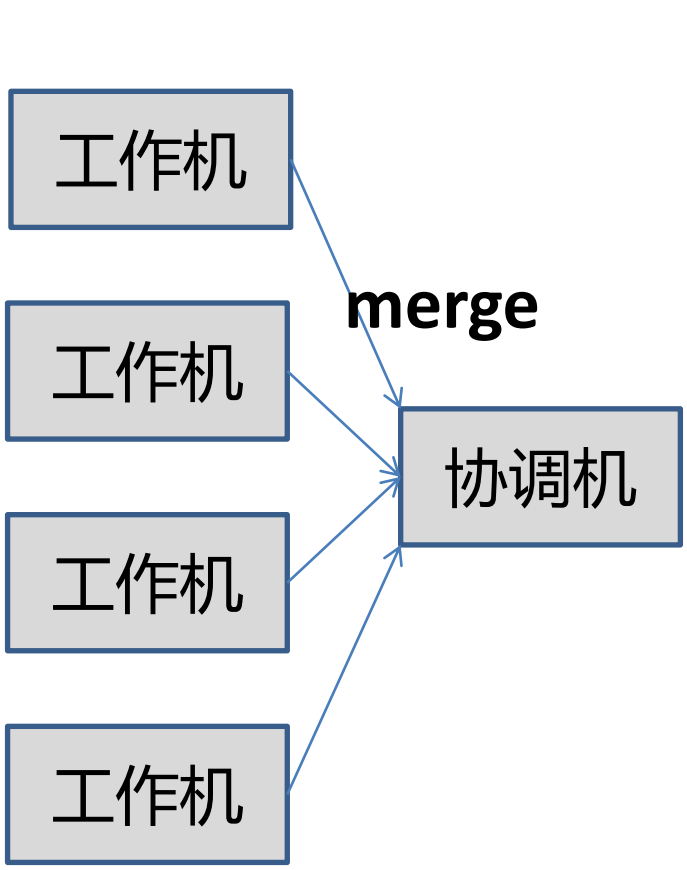
Alibaba Group 阿里巴巴集团

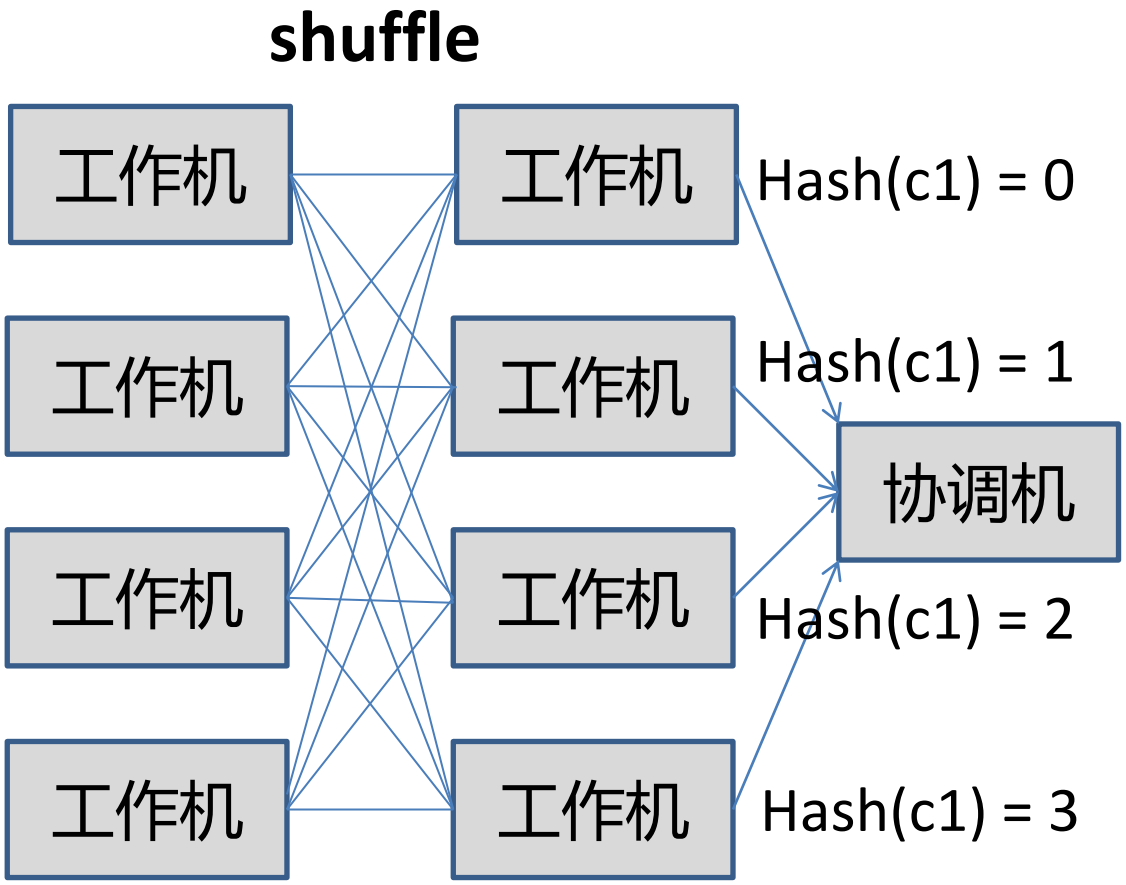
分布式索引

- 问题1：如何对基线数据构建分布式索引？
  - 数据表主键为(k1, k2），索引列为（c1, c2)
  - 如何生成一张按照索引列(c1, c2)有序组织的分布式表格？
- 问题2：如何维护数据表和索引表的一致性？
  - 机器1和机器2：机器1认为数据表有索引，机器2认为没有索引
  - 机器2修改数据表，未修改索引表；机器1查询使用了索引表 => 数据不一致
- 问题3：查询时如何选择合适的索引？
  - 有多个索引供查询选择时，选择哪一个？

OLAP功能升级

- select sum(c1) from table group by c1;





Alibaba Group 阿里巴巴集团

议程

- 整体架构回顾
- 0.5架构升级
- 0.5开发测试
- 经验与后续规划

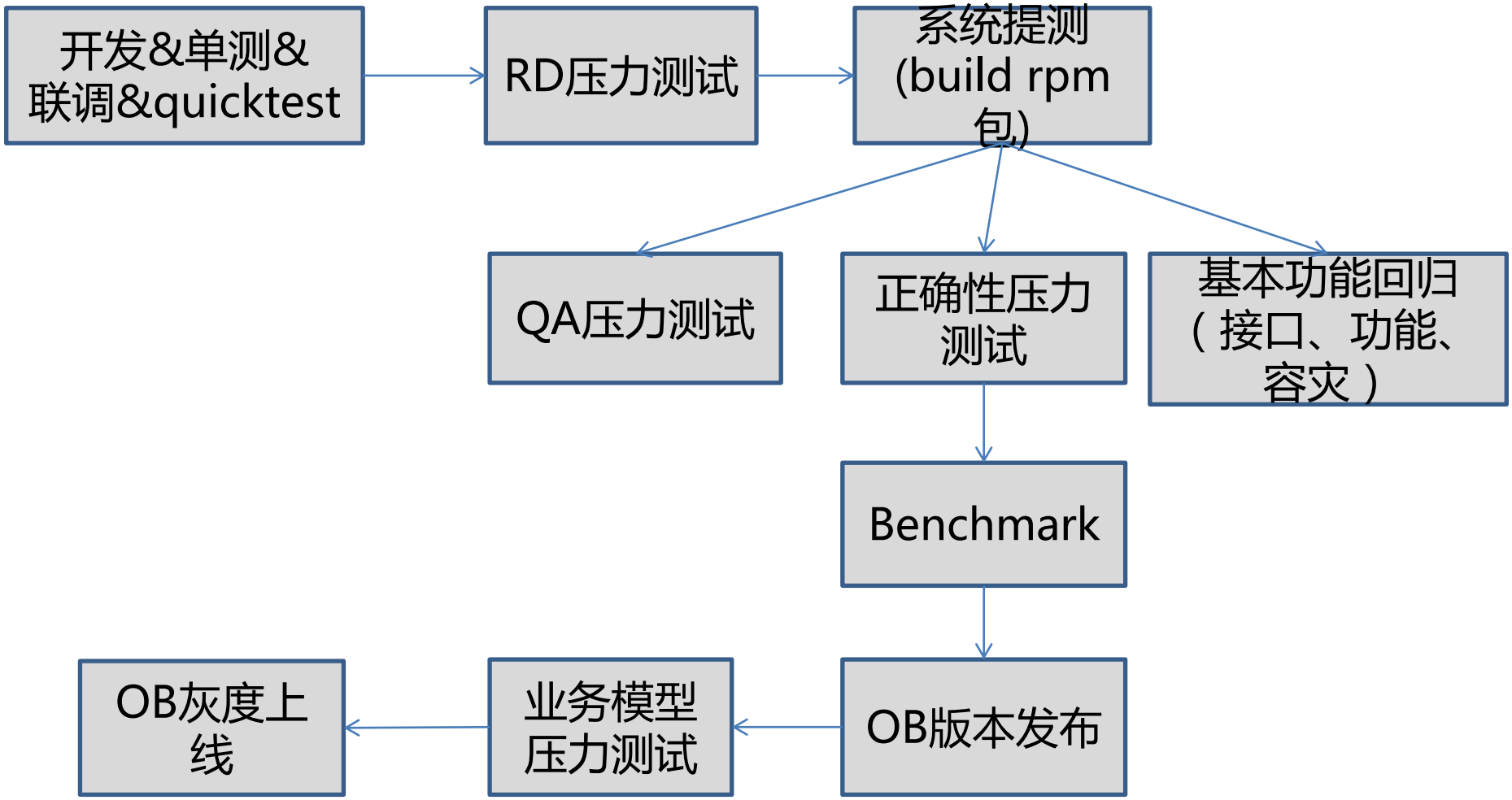
OceanBase A Powerful Database System

工程量

- OB整体代码行数：50W source code + 20W test code
- 本次改造量：10W+ source code
- Bug数量：1000+
- 开发过程
  - 2013.4 ~ 2013.12 预热期：分布式索引、新存储引擎、OLAP、元数据分布式化等主要功能开发
  - 2013.12 ~ 2014.4月底 突破期：遗留功能开发、联调、压测等

Alibaba Group 阿里巴巴集团

质量保证



OceanBase A Powerful Database System

Alibaba Group 阿里巴巴集团

敏捷开发

```
#部署一个OceanBase集群
deploy ob1=OBI(cluster=1211);
deploy ob1.reboot;
sleep 10;

#连接到其中一台MergeServer（ms0）
deploy ob1.connect conn1 ms0 admin admin test;
connection conn1;
```

```
create table t1(pk int primary key, c1 varchar);
insert into t1 values(2,'2_abc'), (3,'3_abc'),
(4,'4_abc'), (5,'5_abc');
update t1 set c1='5_UPDATE' where pk=5;
delete from t1 where pk=2;
#读取表格内容
select * from t1;
```

- 所有测试用例都是文本化的，要求每个RD和QA都会写测试用例
- 测试框架自动化，要求每个RD都会使用测试框架
- 结果驱动而不是角色驱动：部分测试用例由RD开发或者联合开发

OceanBase A Powerful Database System

Alibaba Group 阿里巴巴集团

议程

- 整体架构回顾
- 0.5架构调整
- 0.5开发测试
- 经验与后续规划

OceanBase A Powerful Database System

开发经验总结

- 数据校验：在各个环节校验crc checksum
  - OB 0.5 crc checksum线下发现5+严重bug
- 版本质量：代码bug的第一owner是RD，而不是QA
- 敏捷开发：开发、测试并行，通过外围业务做线上验证
- 兼容性：系统设计之初想好以后如何做无缝升级
- 权衡业务需求和基础架构：抽象业务需求，遵守SQL标准和主流厂商惯用法，避免根据业务定制

Alibaba Group 阿里巴巴集团

后续规划

- 多UpdateServer
- 跨机房架构
- 更多SQL功能
- 持续不断的性能优化
- ……

OceanBase A Powerful Database System

Alibaba Group 阿里巴巴集团

谢谢!

开源网址(GPLv2)

https://github.com/alibaba/oceanbase

OceanBase
A Powerful Database System