Alibaba Group 阿里巴巴集团

OceanBase
A Powerful Database System

支撑支付宝交易的分布式关系数据库

蒋志勇(恒谦)
zhiyong.jzy@alibaba-inc.com
2014-12

OceanBase
A Powerful Database System

Alibaba Group 阿里巴巴集团

内容

互联网应用对数据库的挑战

OceanBase的解决之道

小结

OceanBase A Powerful Database System

Why NOT 商业数据库？

Alibaba Group 阿里巴巴集团

对数据库并发性能的挑战

职员/柜员机
(几百,几千,几万)
操作数据库

网民/草根
(几十万/几百万)
操作数据库

OceanBase
A Powerful Database System

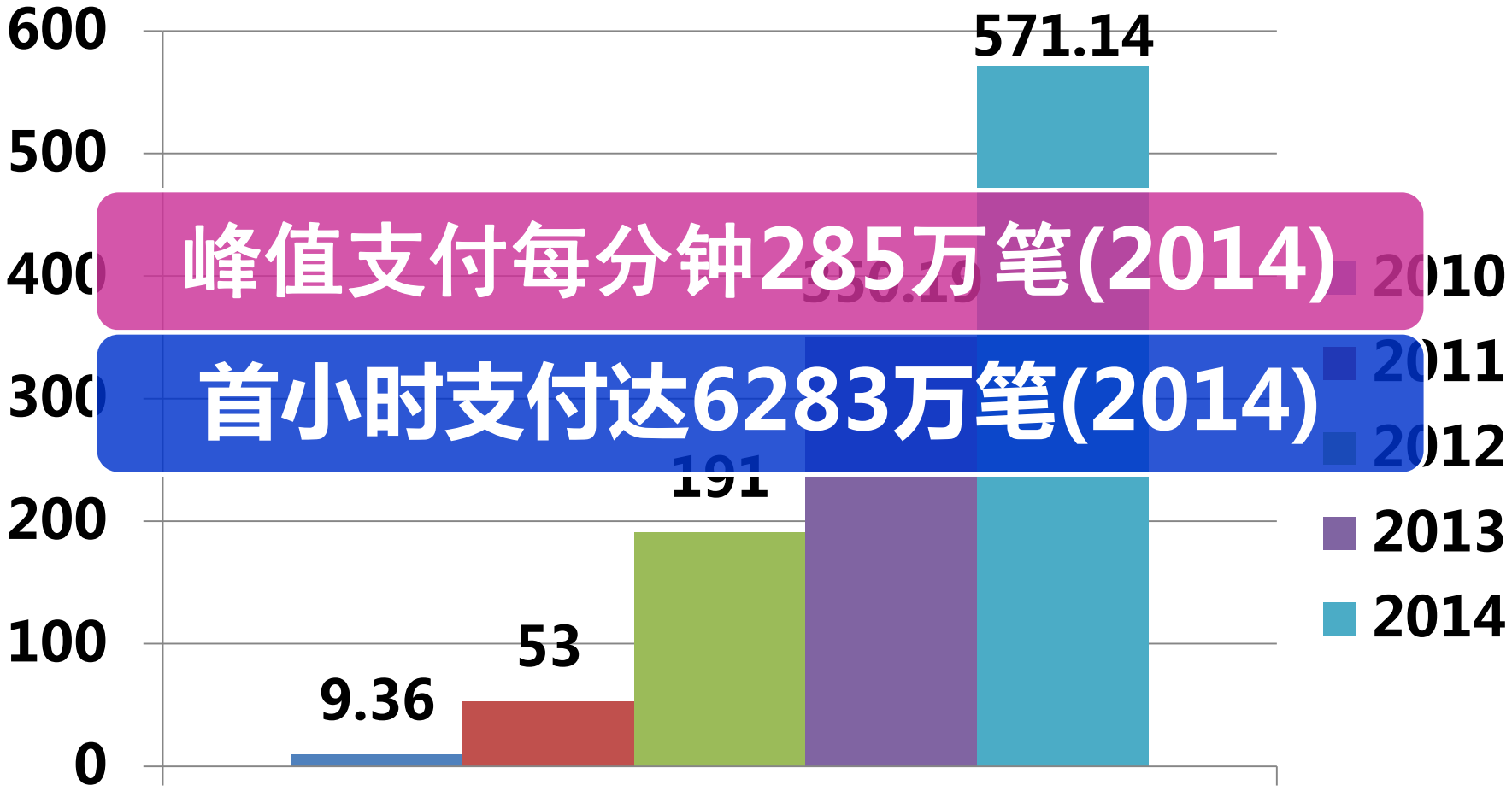
对数据库扩展能力的挑战

新建/扩建商场
数年/数月

2013.11.11
1700万人同时
在线

Alibaba Group 阿里巴巴集团

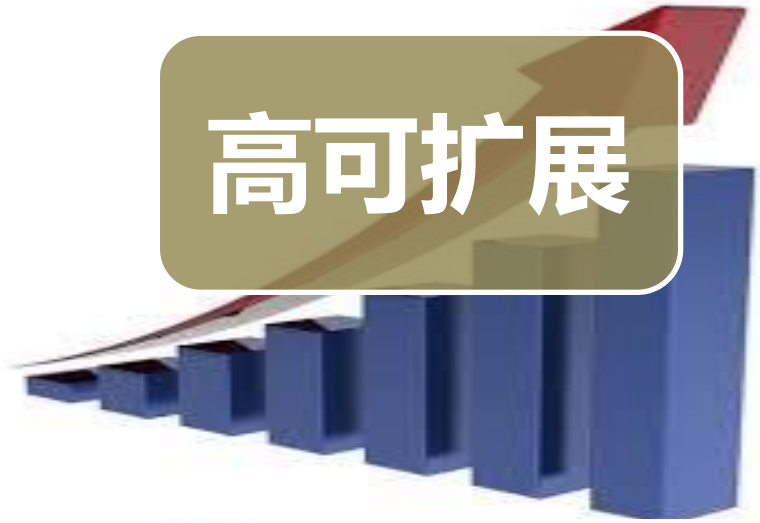
对数据库高可用的挑战



Alibaba Group 阿里巴巴集团

互联网需要的数据库

高性能



高可扩展

高可用

低成本

OceanBase
A Powerful Database System

Alibaba Group 阿里巴巴集团

关于OceanBase

- 一个分布式关系数据库

2011.2 (v0.1)

交易库

2014.10 (v0.5)

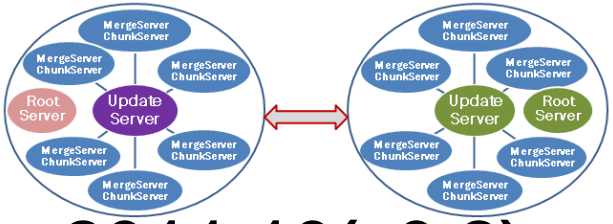
- 第一个用于金融核心系统的非商业数据库
  - 基于PC服务器，可在线水平扩展
  - 无共享存储
  - 性价比高于商业数据库。

OceanBase
A Powerful Database System

OceanBase 开发里程碑

2010.6

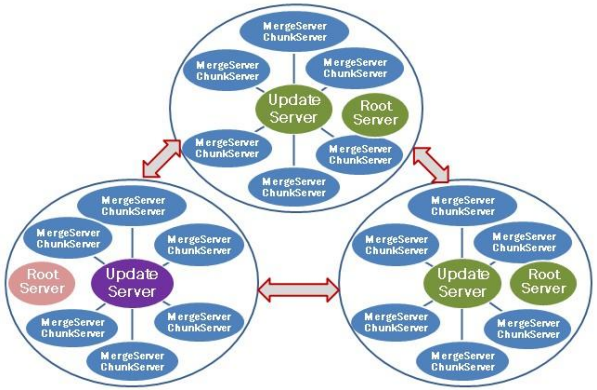
2011.2(v0.1)



2011.10(v0.2)

2012.4(v0.3)

2012.11(v0.4)



2014.2(v0.5)

OceanBase性能：如何超越

- 跟随先行者的足迹去追赶并超越？

.....

- 读(磁盘随机读)
  - ✖机械磁盘随机读：100-300次/秒(IOPS)
  - ✓固态盘(SSD)随机读：几万次/秒(IOPS)
- 写(磁盘随机写)
  - ✖机械磁盘随机写：100~300次/秒(IOPS)
  - ➢数据库写入放大：8KB块，每次修改100B➔80X
  - ➢固态盘：写入放大 & 写入前先擦除

Alibaba Group 阿里巴巴集团

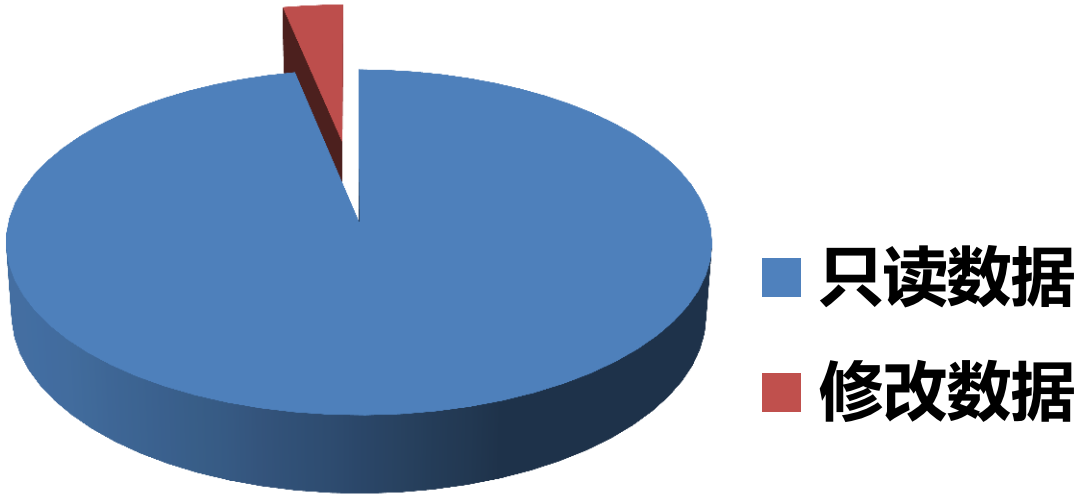
数据库应用实例

- 全国人口数据库
  - 14亿条记录
  - 增删改：出生、死亡、迁移……
- 支付宝交易数据库
  - 每笔交易一条或几条记录
  - 增删改：创建、买家已付款、卖家已发货、退货退款……
- 支付宝账务库
  - 每个人一条或几条记录
  - 增删改：付款、收款、账户变更……

记录总量大，而一天内增删改量占比很小。

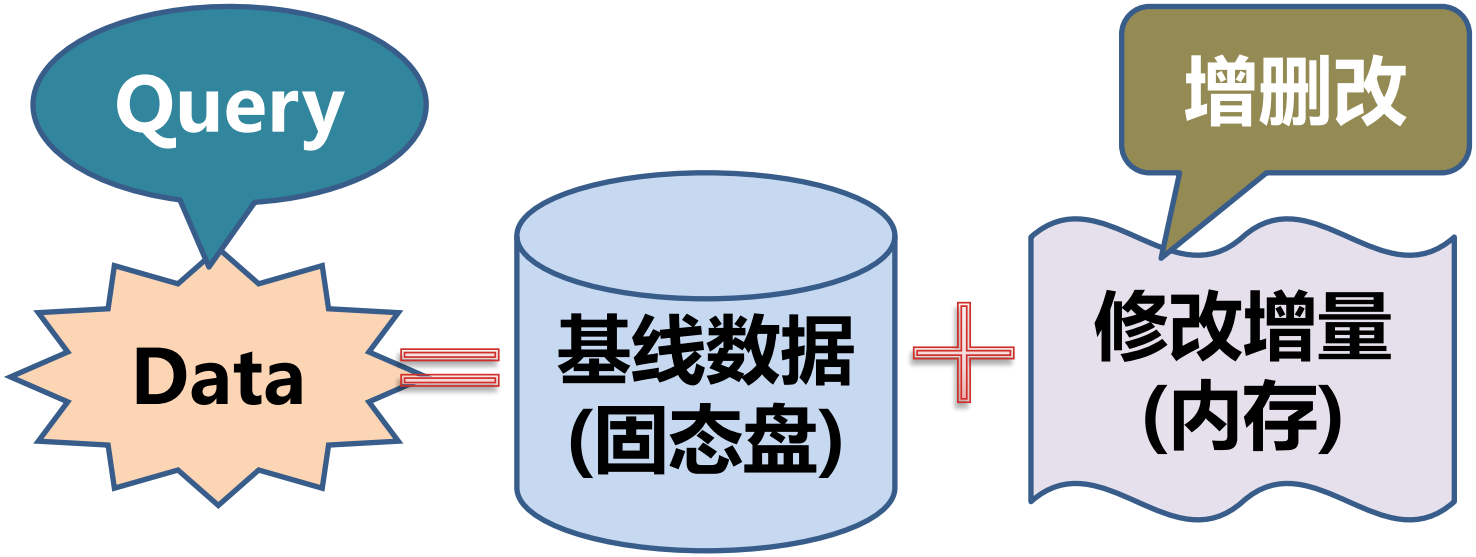
OceanBase
A Powerful Database System

- 数据库：总量大，增删改量少
  - 10亿笔写事务，100B/事务➔100GB

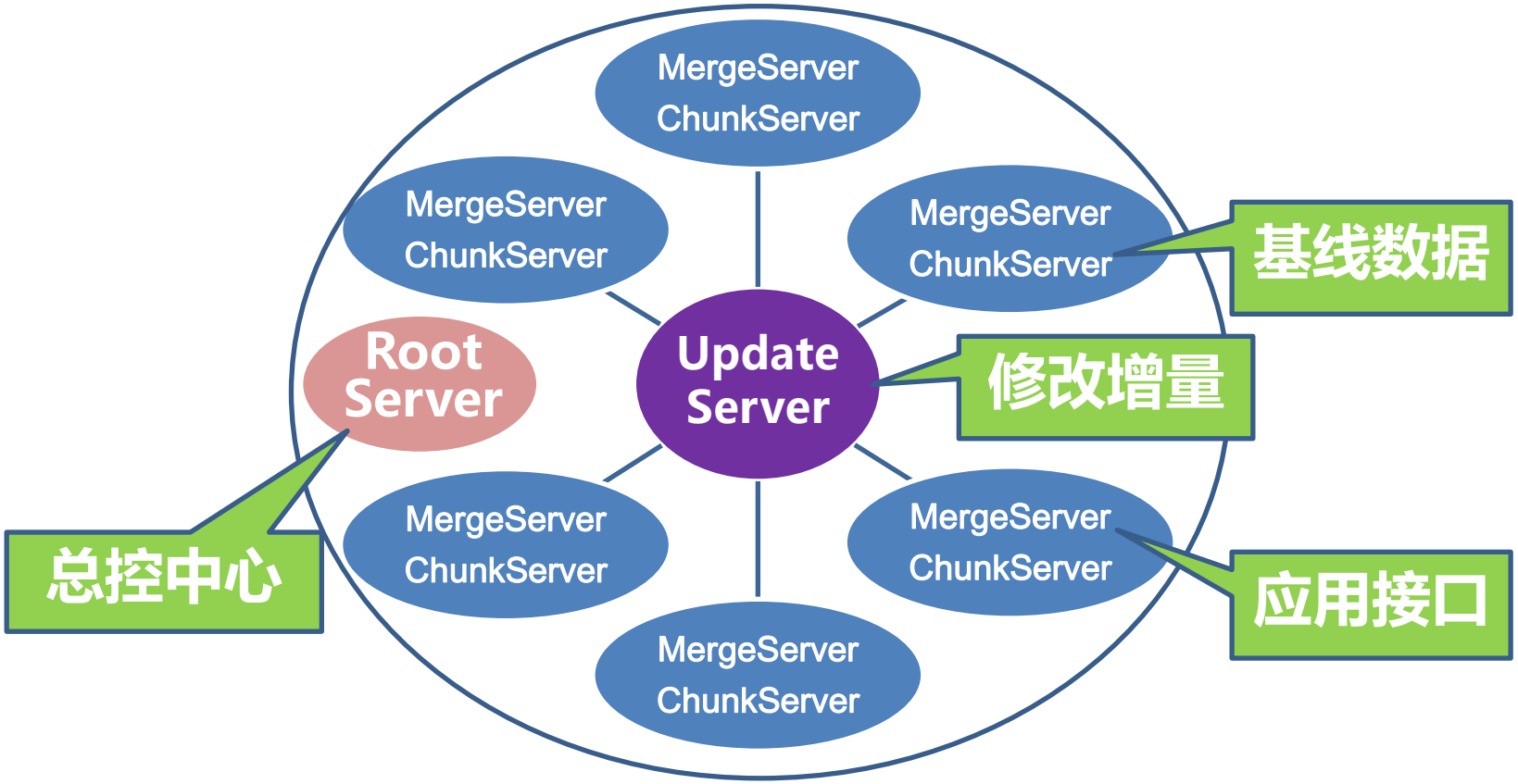


Alibaba Group 阿里巴巴集团

OceanBase的解决之道



OceanBase架构

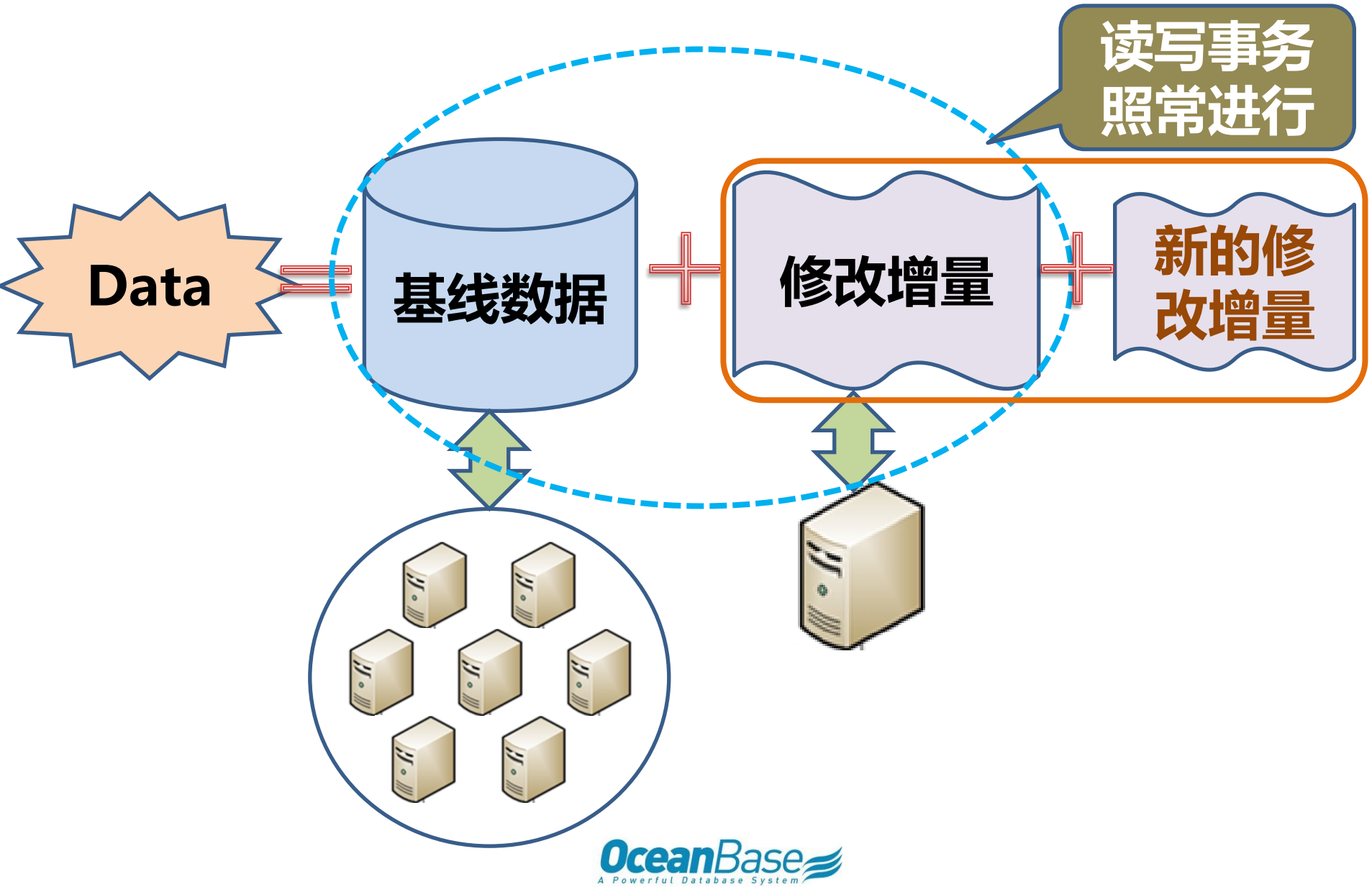


- 修改增量(增删改)置于内存：无随机磁盘写、性能高
- 单点写入：数据一致性好、性能&内存容量有瓶颈
- 修改增量与基线数据跨服务器

修改增量一直在内存？

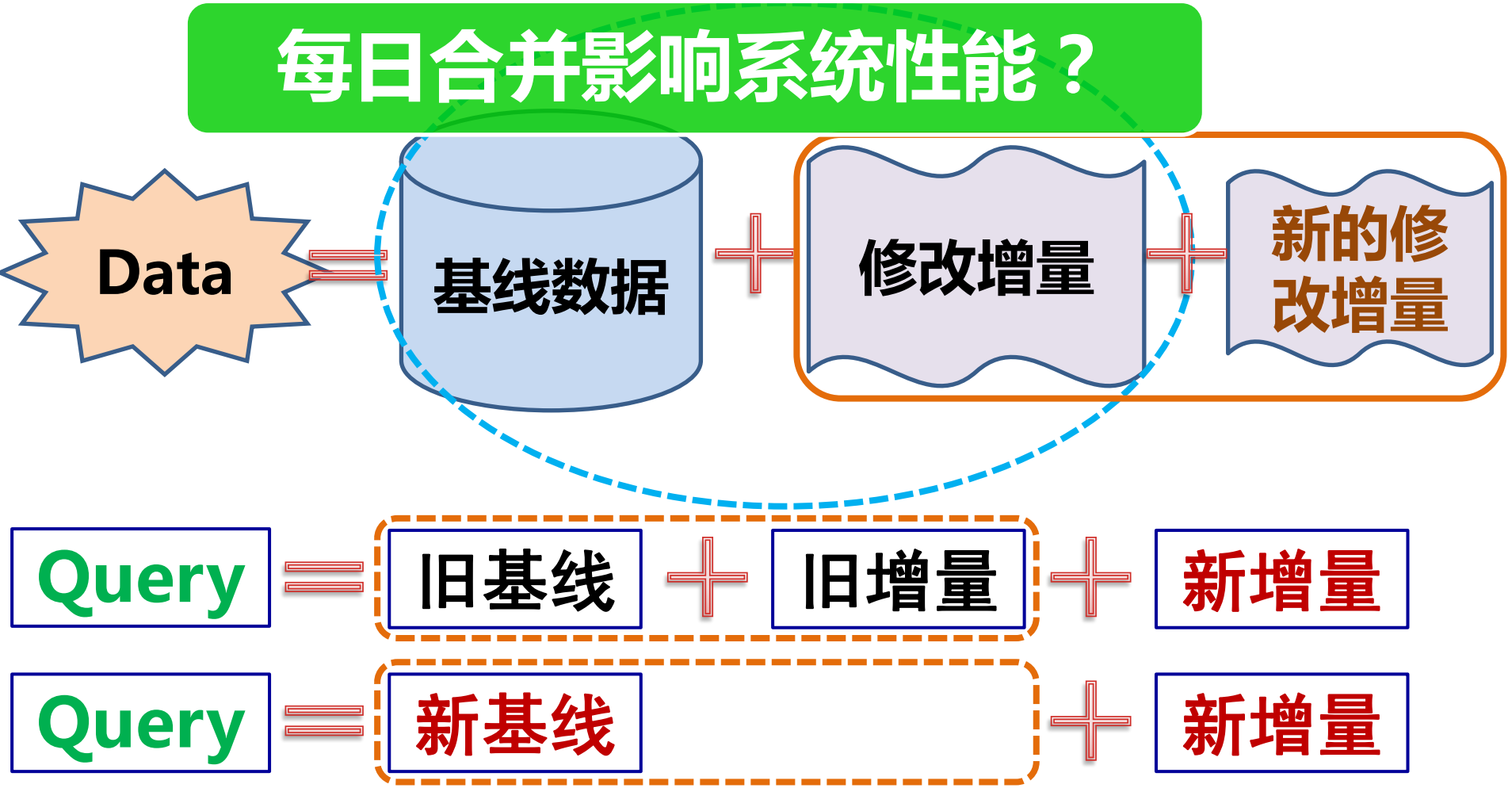
Alibaba Group 阿里巴巴集团

每日合并修改增量到基线



OceanBase
A Powerful Database System

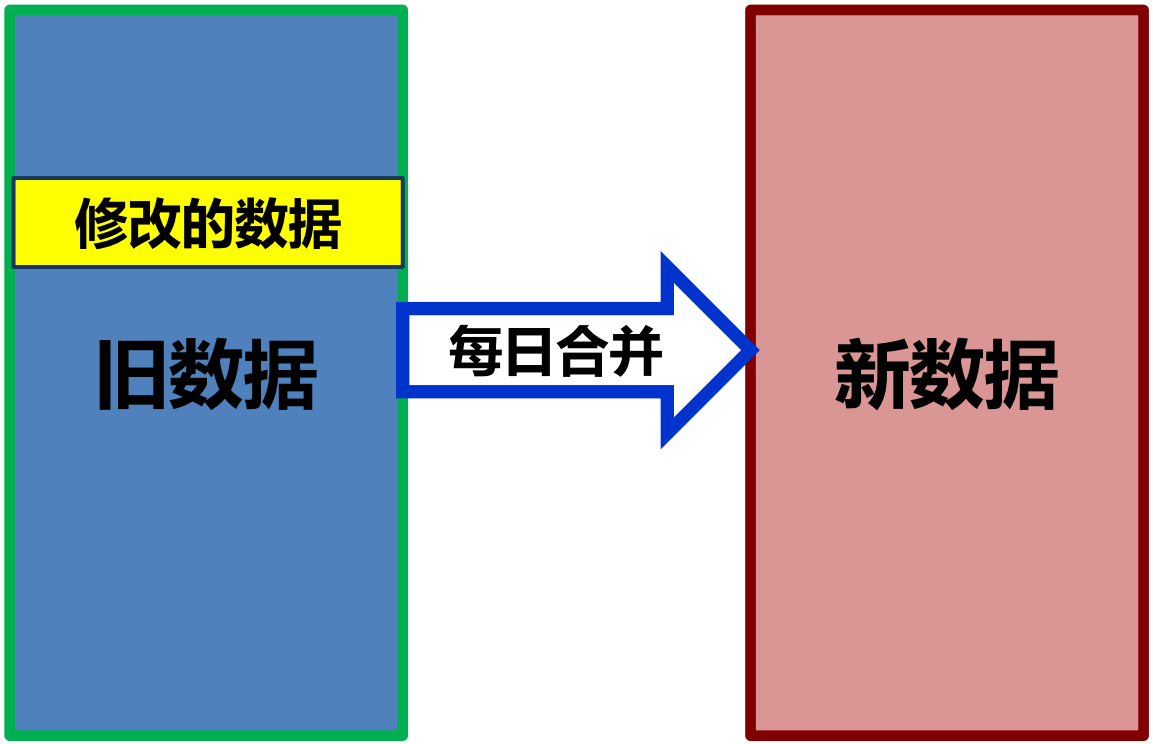
每日合并期间的query



- 使用新旧基线数据，查询结果相同

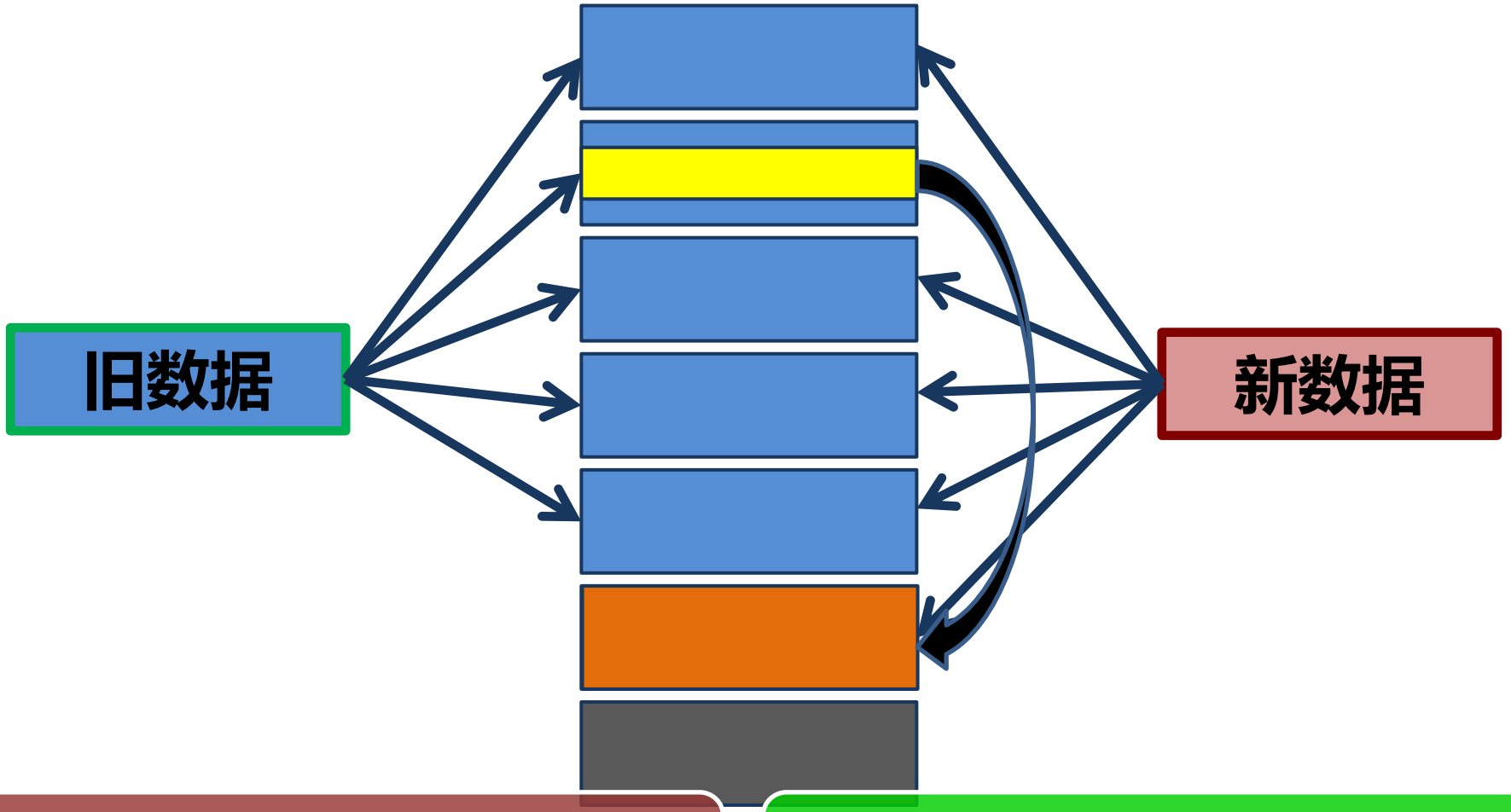
每日合并的耗时

- 一条记录修改，整个数据重写
- 50MB/s写入4TB盘，需4TB/50MB=80,000s



降低每日合并的耗时

- 数据分块(例如2MB)



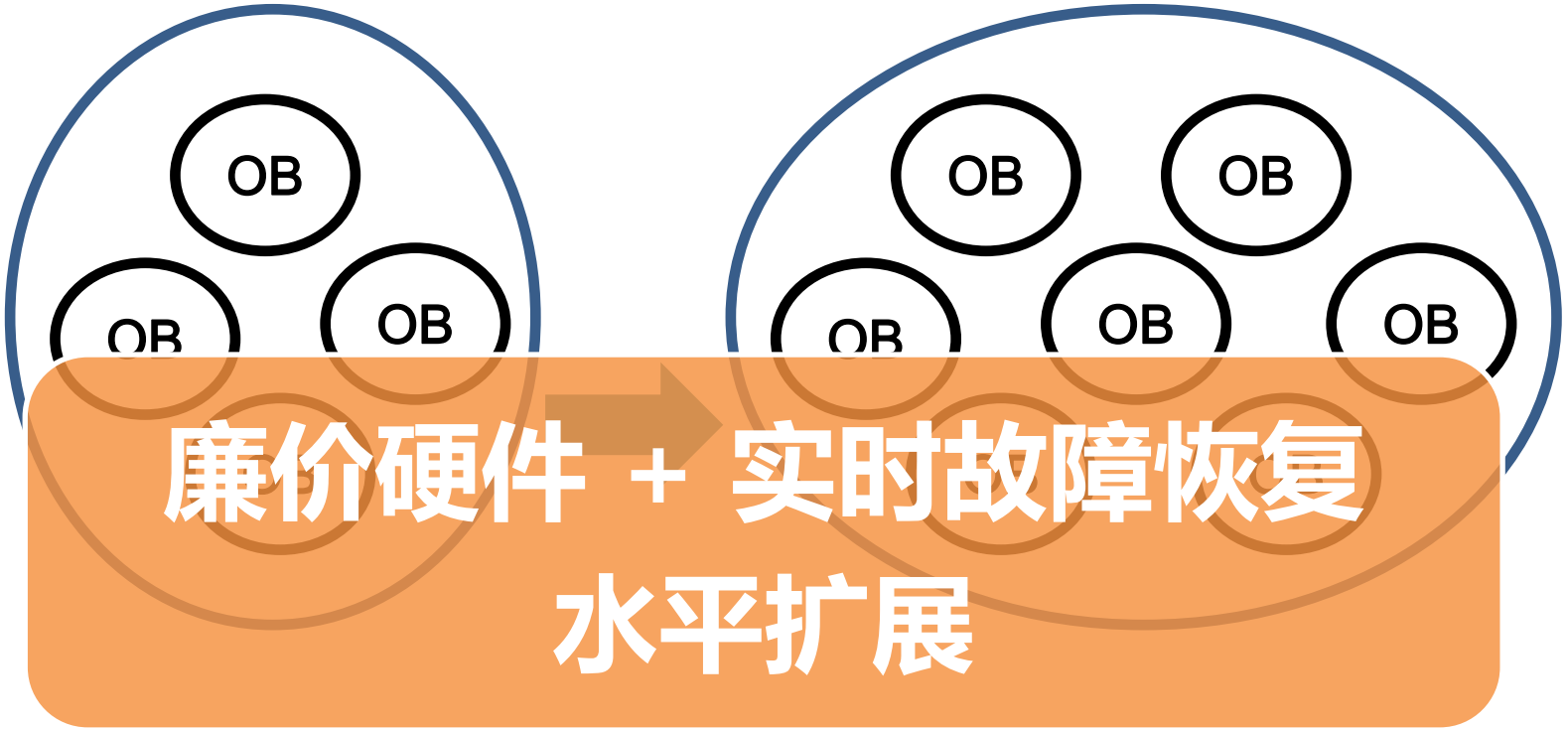
数据块越小越好？

写入导致磁盘磨损？

Alibaba Group 阿里巴巴集团

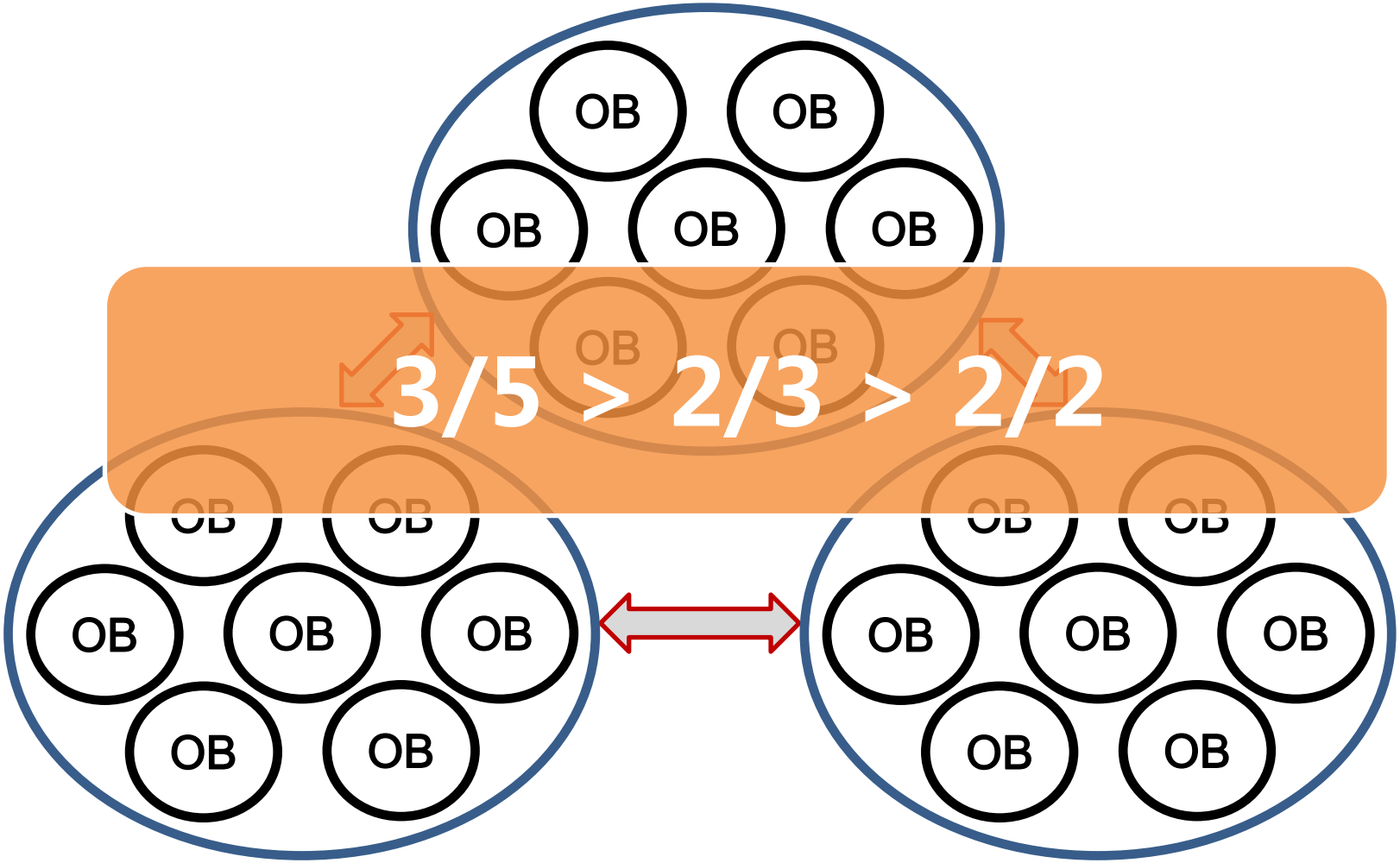
易于扩展

- 外部：一台虚拟大型机
- 内部：多台PC服务器组成
  - 按需扩容，自动迁移数据
  - 最大业务机群：200+台



OceanBase
A Powerful Database System

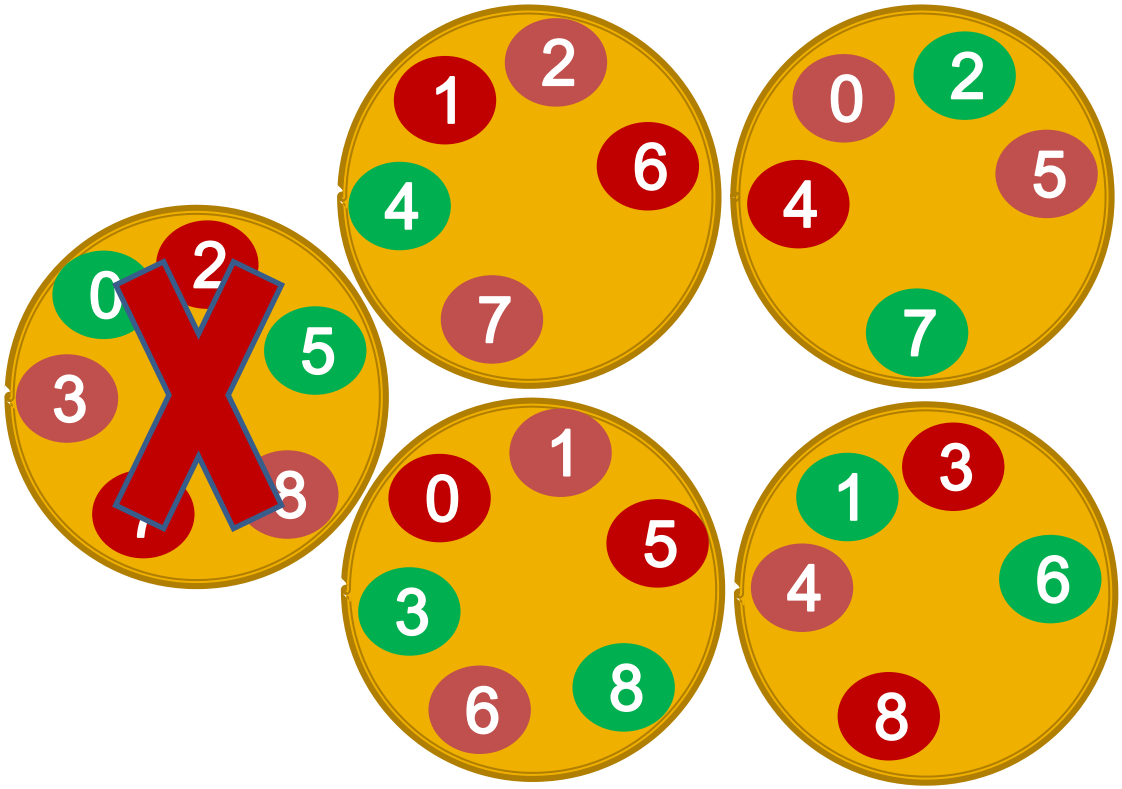
多机群：更高可用性



- 日志到达超过半数库(2/3,3/5…)事务成功

Alibaba Group 阿里巴巴集团

服务器故障自动恢复



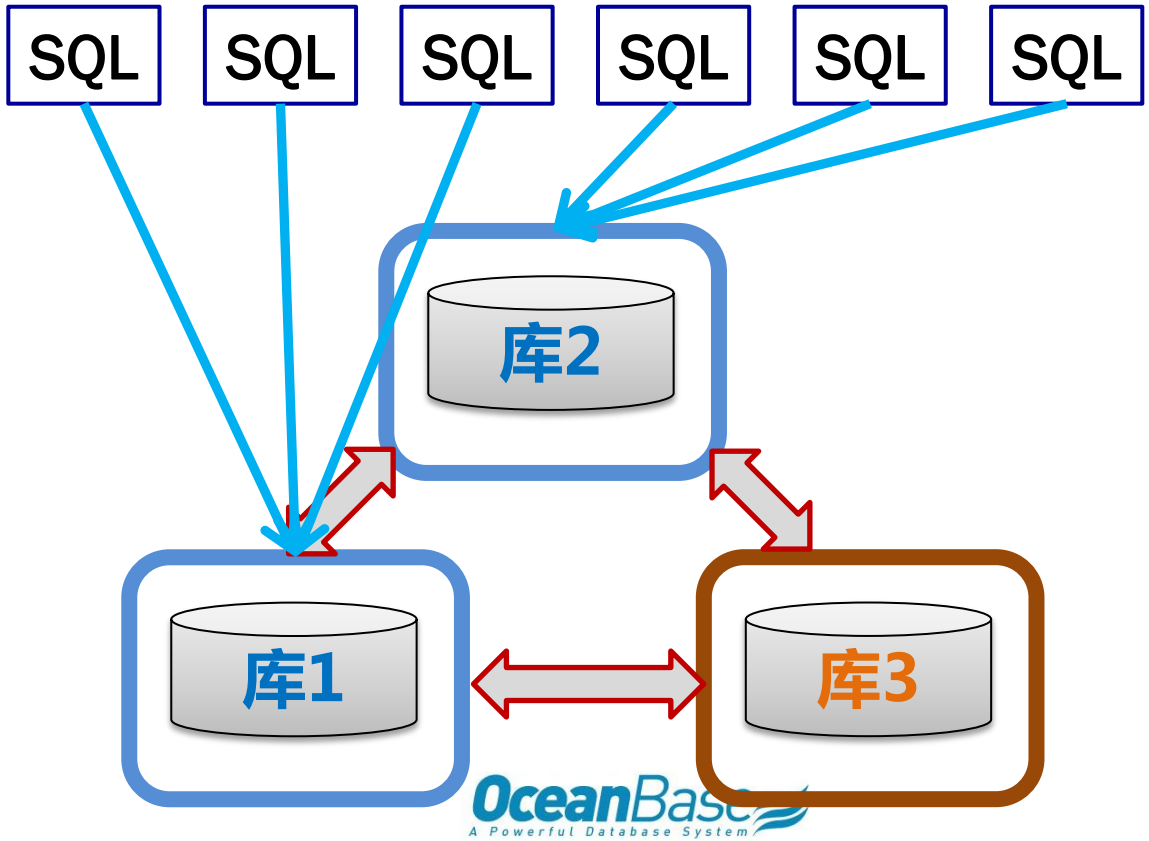
主库故障不丢数据

自动恢复(时间最长30s)

OceanBase
A Powerful Database System

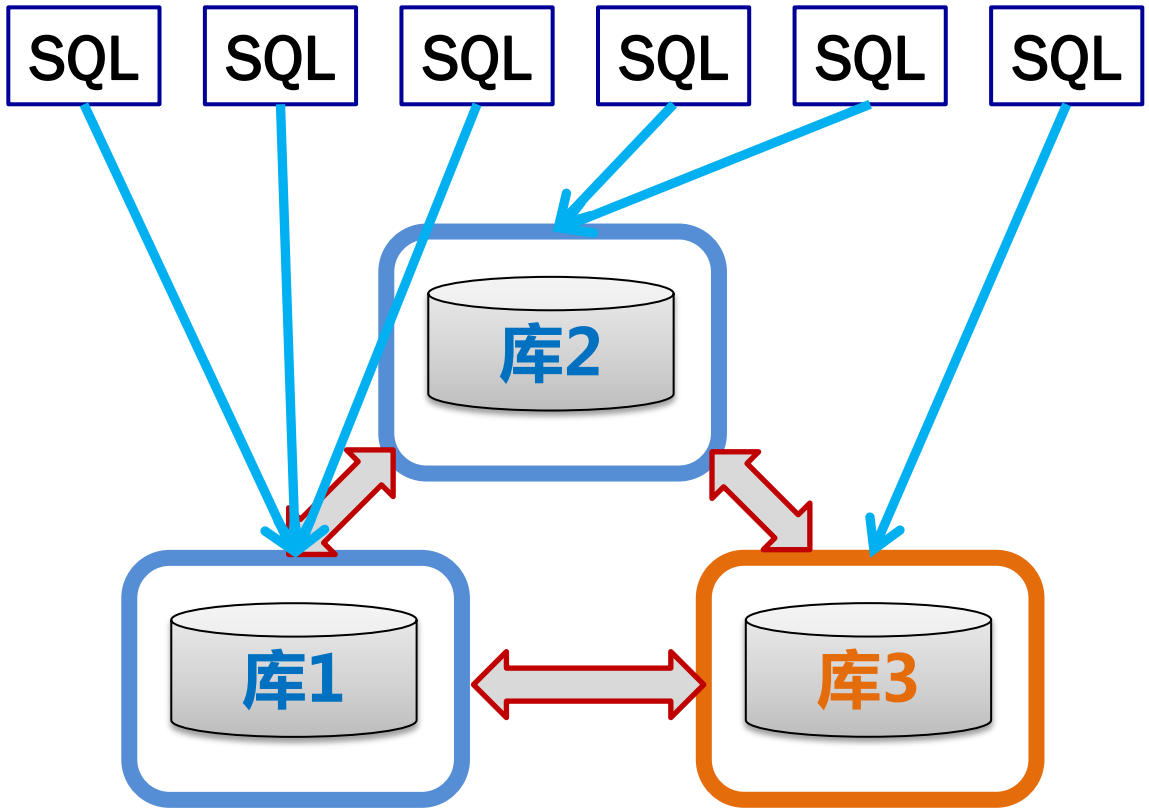
数据库灰度升级(1)

- 2014.8.19-30，美国国务院签证数据库异常(Oracle)
- 2013.6.23，中国工商银行数据库异常(DB2)
- OceanBase升级Step 1. 切走一个库流量并升级
  - 内置自动数据比对



数据库灰度升级(2)

- OceanBase升级Step 2.逐步导入流量
  - 白名单、1%、2%、5%.....
  - 继续数据比对，出现异常立即回滚



Alibaba Group 阿里巴巴集团

数据校验

- ✓ 磁盘读写：每条记录带64位checksum
- ✓ 网络传输：每个网络包带64位checksum
- ✓ 每个文件多副本(3~6)：每个文件都有64位checksum
- ✓ 修改增量多副本：带64位累积checksum
- ✓ Redo log：每条都带checksum及对应于UpdateServer内存的checksum
- ✓ 每个表每个列都带64位checksum

Alibaba Group 阿里巴巴集团

小结-互联网下的数据库

- ✓ 事务(ACID)：互联网 = 传统业务
- ✓ 可用性：互联网 = 传统业务
- ● 性能：互联网 >>传统业务
- ● 性价比：互联网 >>传统业务
- ● 扩展性：互联网 >>传统业务

OceanBase

Alibaba Group
阿里巴巴集团

Thanks

OceanBase
A Powerful Database System