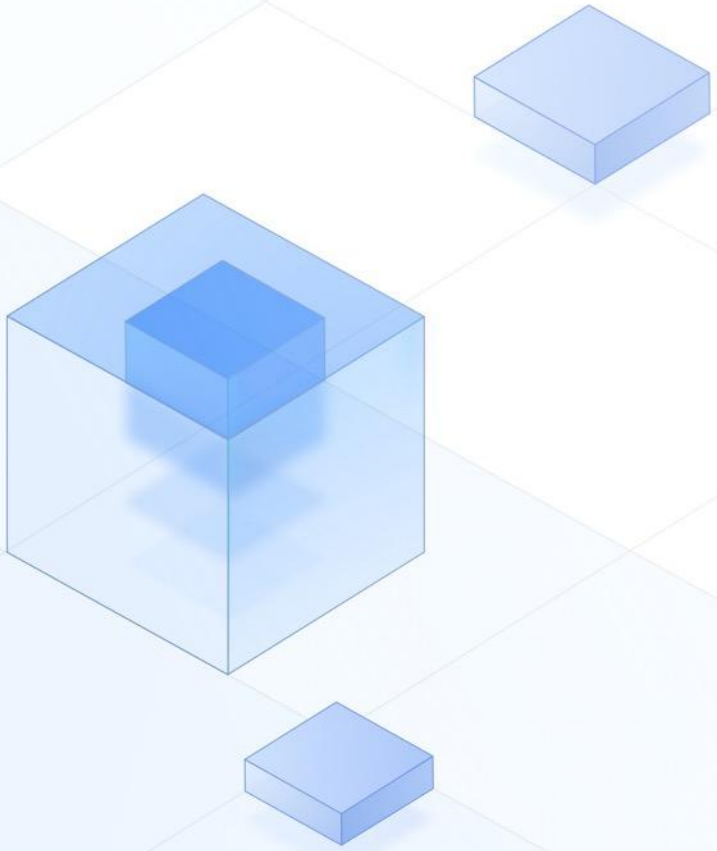
OceanBase 的 OLAP 能力提升实践

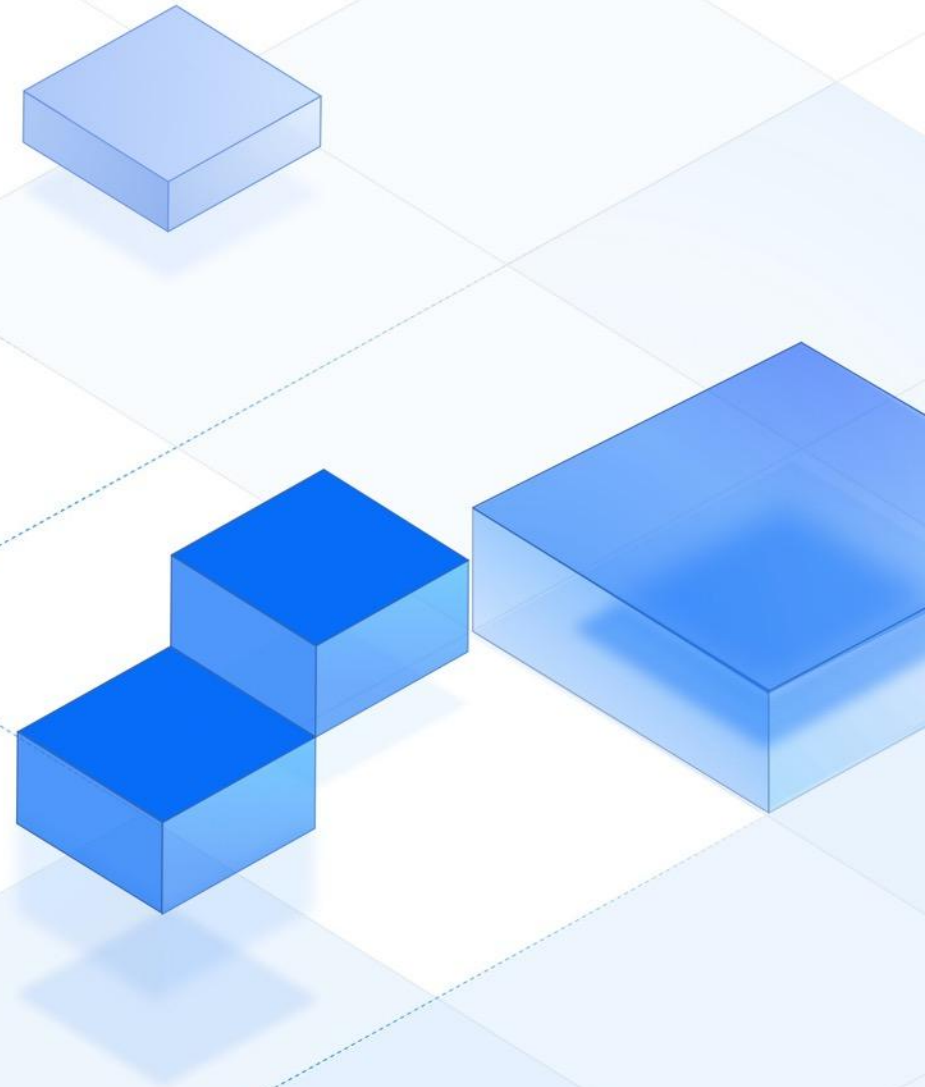
杨志丰（竹翁）

OceanBase 首席架构师

目录



OceanBase简介



OceanBase发展历程

OCEANBASE

- 自主研发，完整知识产权、核心能力100%掌控
- 企业级能力，多年支撑蚂蚁核心业务100%负载，数百家单位客户

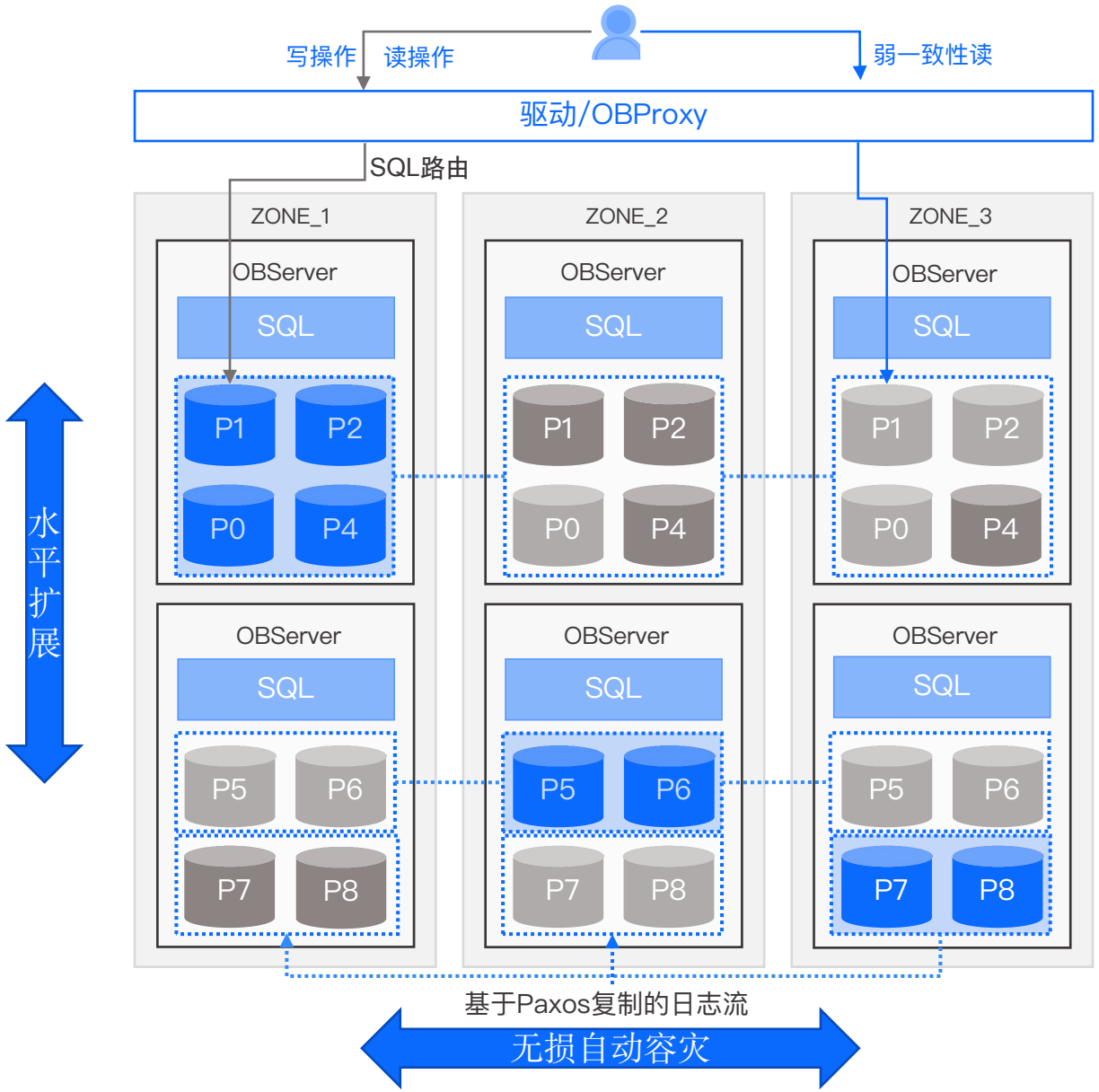
2010	2016	2017	2019	2020	2021	2022
产品立项	支付宝账务	多家金融客户	打破世界纪录	独立商业化	规模化推广	生态大拓展
第一个用户 淘宝 Taobao	核心账务 核心交易支付 金融核心业务 支付宝 网商银行	互联网核心系统 余额宝 南京银行 BANK OF NANJING	公有云服务 TPC-C 6100万 广东农信 GDRC 西安银行 BANK OF XI'AN 天津銀行 BANK OF TIANJIN	HTAP 引擎 TPC-C 7.07亿 中国工商银行 INDUSTRIAL AND COMMERCIAL BANK OF CHINA 中国建设银行 China Construction Bank 交通銀行 BANK OF COMMUNICATIONS	开源开放 TPC-H 30T 1526万 中国人寿 CHINA LIFE 中国石化 SINOPEC 国家电网 STATE GRID 数字江西	公有云 社区用户 Ctrip 携程 理想 致欧 TELECOM ITALIA kaspi China unicom中国联通

V0.1 分布式，三副本高可用 → V1.0 分布式事务，多租户 → V2.0 高度兼容，高性能 → V3.0 HTAP混合负载 → V4.0 单机分布式一体化

OceanBase 产品功能特性

- 高性能：TPC-C 7.07亿tpmC 世界第一
- 高可用：RPO=0，RTO<8s
- 高可扩展：水平和垂直扩展，自动负载均衡，弹性扩缩容
- Oracle/MySQL兼容：业务少量修改即可迁移到OB，自动评估和迁移工具
- 单机分布式一体化：高效单机和分布式，按需转换
- HTAP：同一套引擎同时支持OLTP和OLAP混合负载
- 低成本：LSM-Tree，编码压缩，存储空间MySQL 1/3
- 原生多租户：集中管理多个业务数据，适合微服务架构和SaaS行业应用
- 安全：透明加密、传输加密，安全审计，细粒度权限
- 国产化：适配鲲鹏、海光等芯片
- 完备产品体系：开发（ODC）、评估（OMA）、迁移（OMS）、运维（OCP）、诊断（OAS）

OceanBase 4.0整体架构



对等节点
- 无共享集群
- OBServer包含SQL、存储、事务

高可扩展性
- 按分区做数据分片扩展
- 多Zone多活扩展

单集群规模
- TPC-C使用1557节点

单机分布式一体化
- 日志流：数据库的所有变更
- 多个分区可共用一个日志流
- 单机内无分布式事务
- 低时延分布式处理技术

稳定可靠的金融级分布式数据库

OCEANBASE

以下数据来自于**实际**生产系统

6100
万次/秒
数据库峰值处理能力

>200
台
集群节点数

>6
PB
单库存储容量

>3200
亿行
单表行数

RPO=0,
RTO<8
秒
少数副本故障时

OLTP能力试金石：TPC-C世界第一

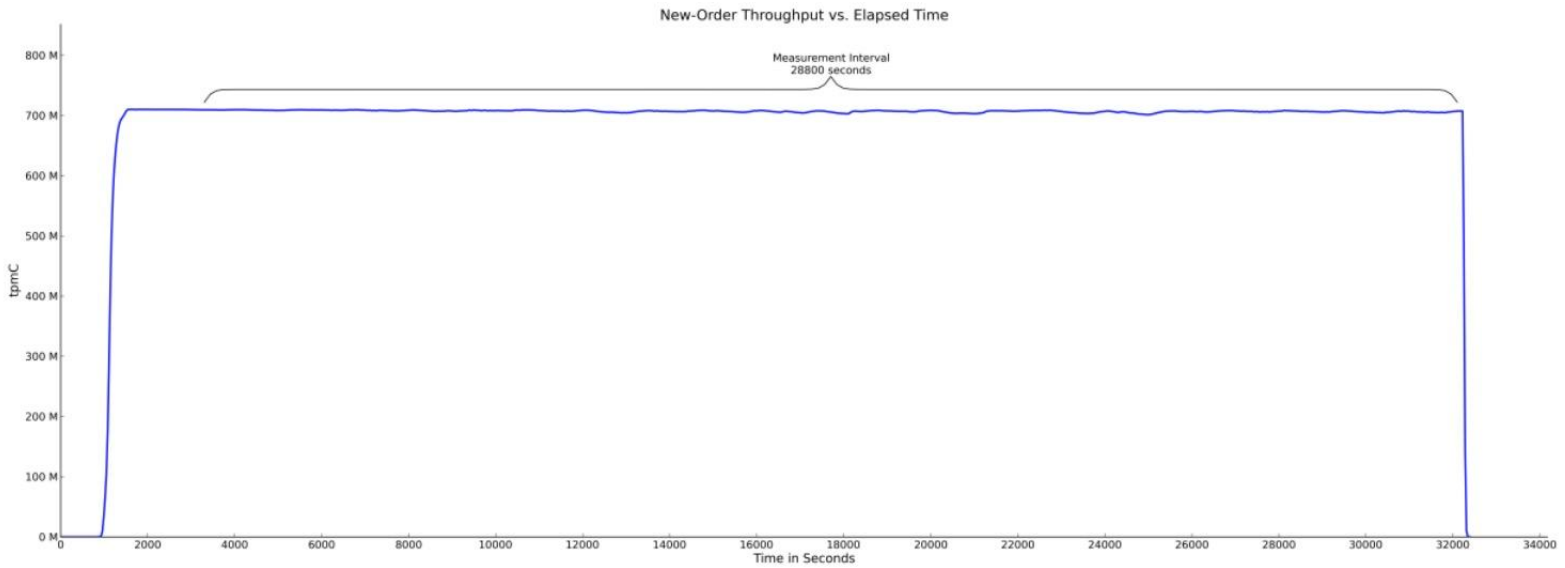


Figure 11 New-Order throughput versus Time

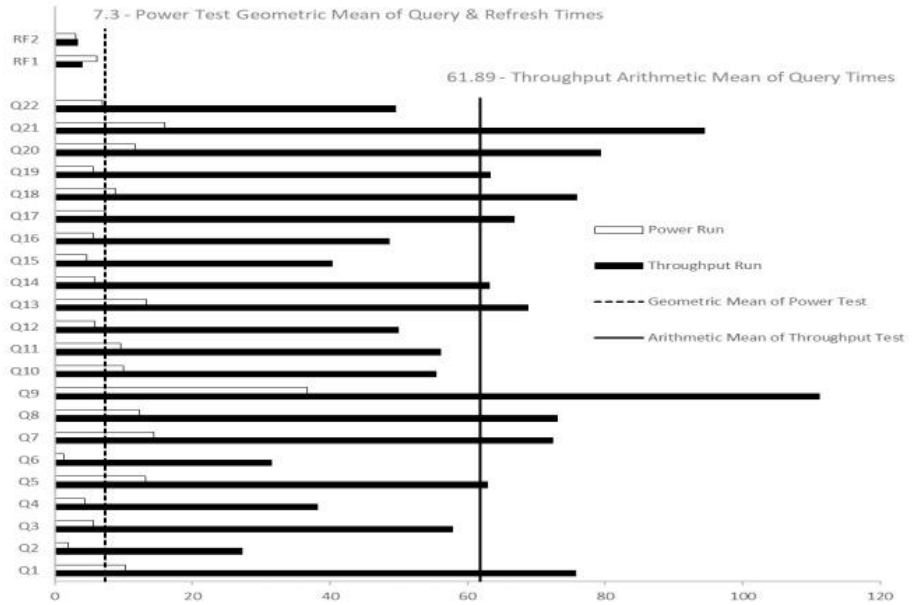
单机 Warehouse 数	36,000
单机理论最高 tpmC	460,800
数据库节点规模(台)	1557
8 小时压测 tpmC	707,351,007
压测 tpmC / 理论 tpmC	**0.987**

TPC-C世界第一

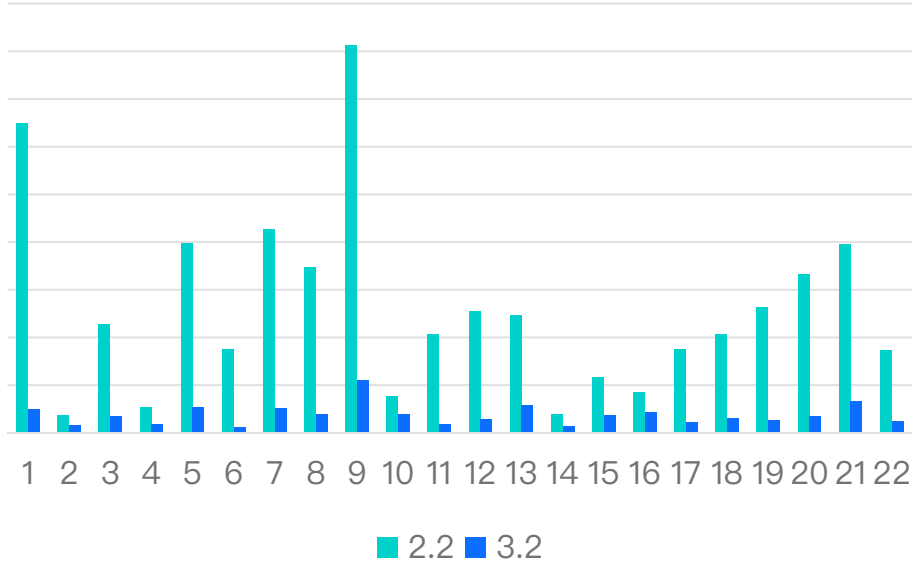
- TPC-C是国际最权威的OLTP评测
  - 严格ACID测试
  - **第一个**通过TPC-C的分布式数据
  - **第一个**通过TPC-C的中国数据库
- 事务模型
  - New-Order事务10%分布式
- 性能表现
  - 稳态运行8小时tpmC抖动小于1%
  - 平均23分钟完成一次快照

OLAP小试牛刀：TPC-H 30,000GB获得世界第一

OCEANBASE

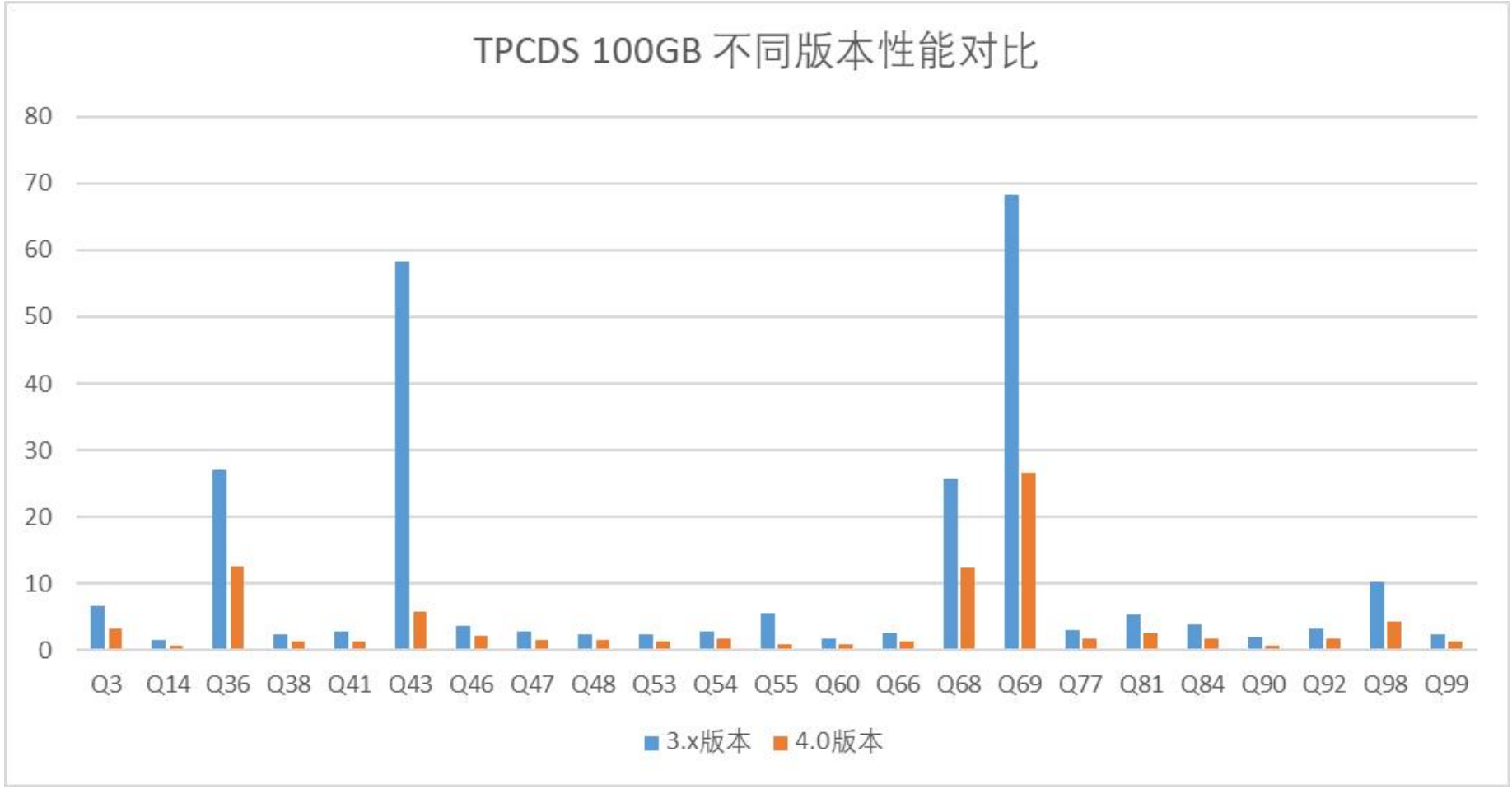


QphH	15,265,305
Price/kQphH	4,542 CNY
数据库节点规模	64
单机配置	80vCPUs+768GB
日期	05/19/21



- OceanBase 3.2 TPC-H整体性能提升620%
- 优化器
  - 优化时间提升10倍；新增改写规则
  - 直方图；统计信息管理
- 全新SQL执行引擎
  - Cache友好：强类型、向量化执行
- MPP&SMP并行执行框架（64节点4096并行度）
  - **并行DML、超大事务支持**

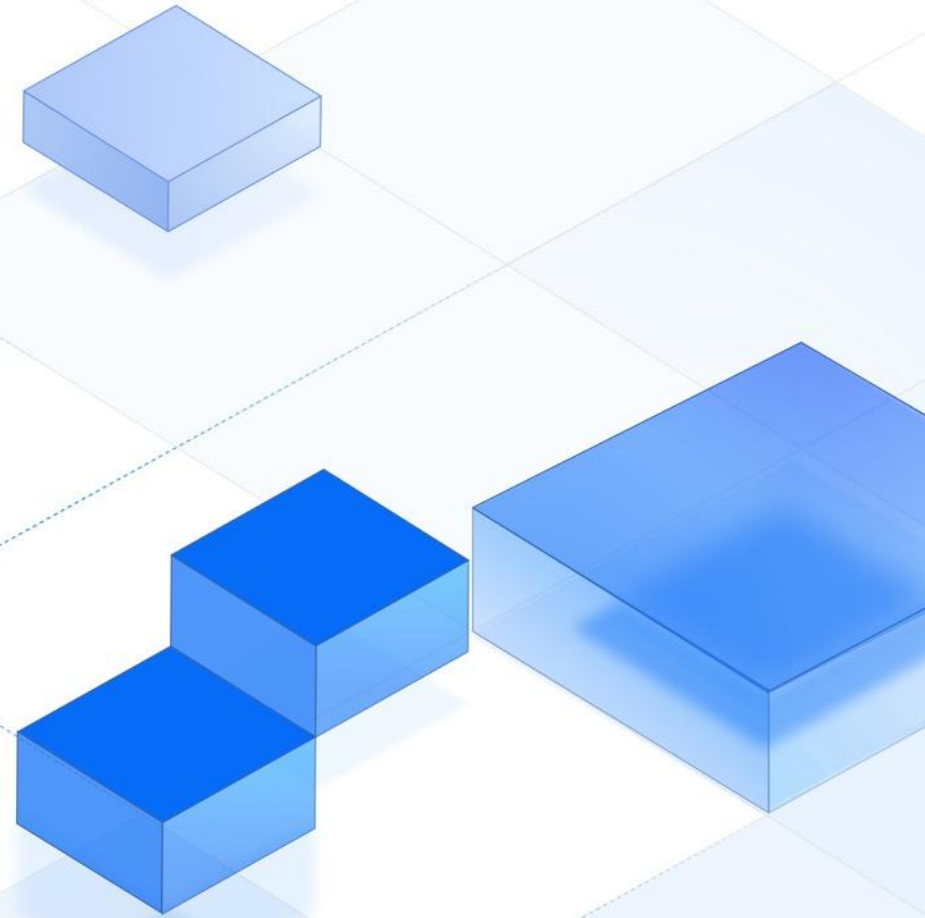
OceanBase 4.0 OLAP能力增强



性能提升3.4倍

- 918s –> 270s
- 一阶段分布式查询优化
- 自适应执行引擎
- 三阶段并行下压

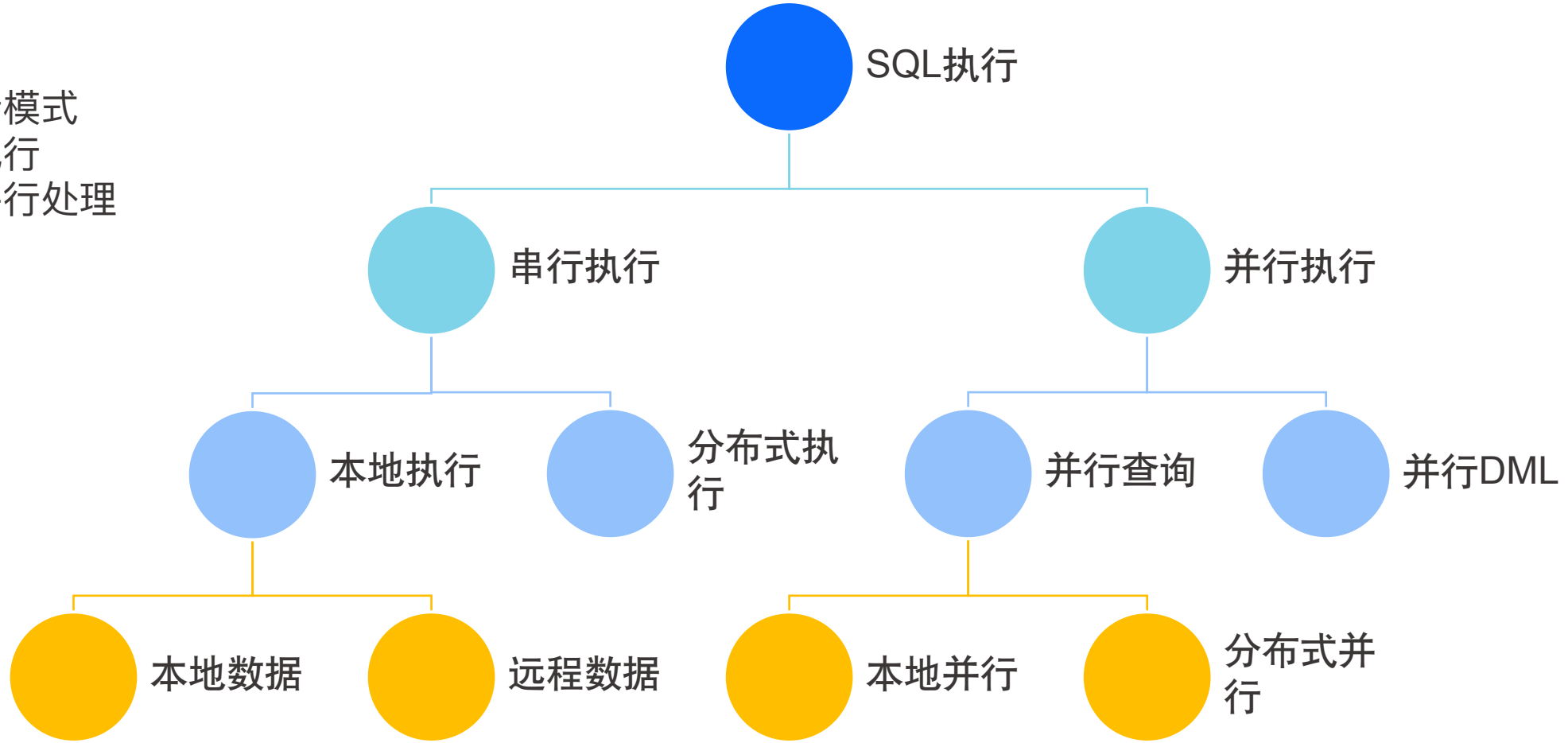
SQL并行执行



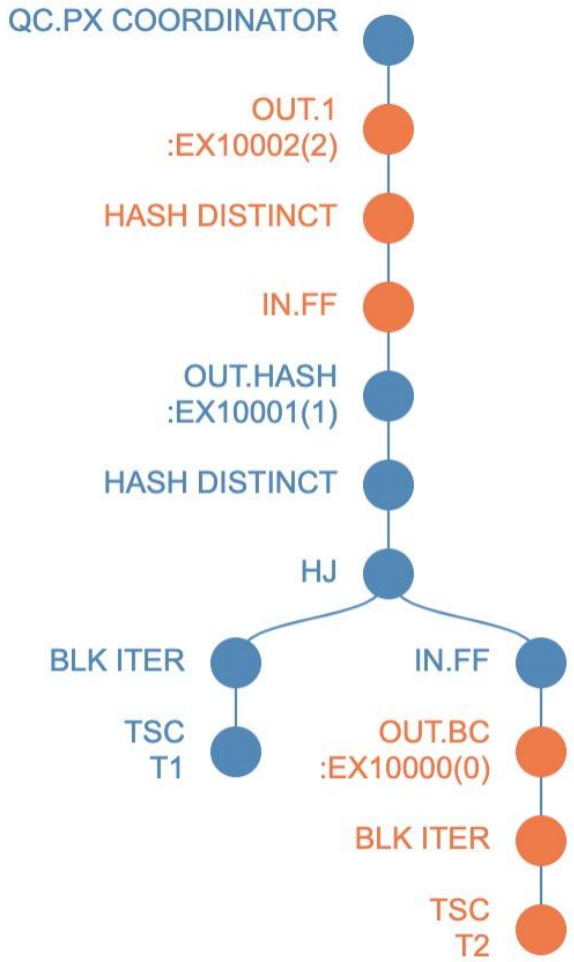
自适应TP+AP混合负载的执行引擎

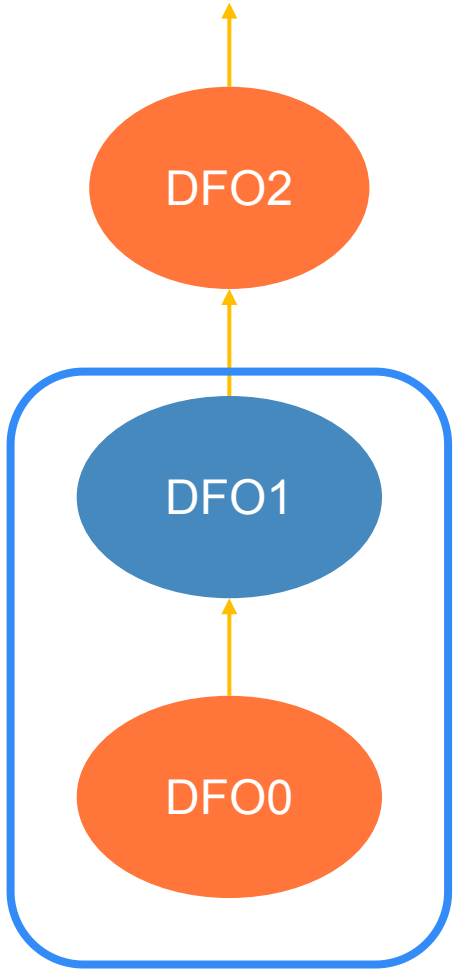
OCEANBASE

- 多种执行模式
- 向量化执行
- 大规模并行处理



并行执行调度


QC.PX COORDINATOR
OUT.1
:EX10002(2)
HASH DISTINCT
IN.FF
OUT.HASH
:EX10001(1)
HASH DISTINCT
HJ
BLK ITER
TSC
T1
IN.FF
OUT.BC
:EX10000(0)
BLK ITER
TSC
T2


DFO2
DFO1
DFO0

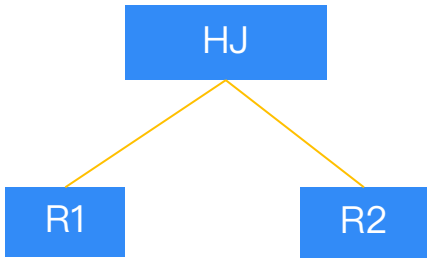
丰富的分布式执行策略

OCEANBASE

```
CREATE TABLE R1(a int, b int, c int) PARTITION BY HASH(b) PARTITIONS 4;
CREATE TABLE R2(a int, b int, c int) PARTITION BY HASH(b) PARTITIONS 4;
```

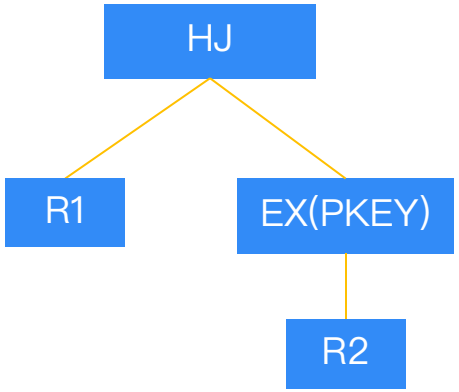
Partition Wise Join

R1.b = R2.b



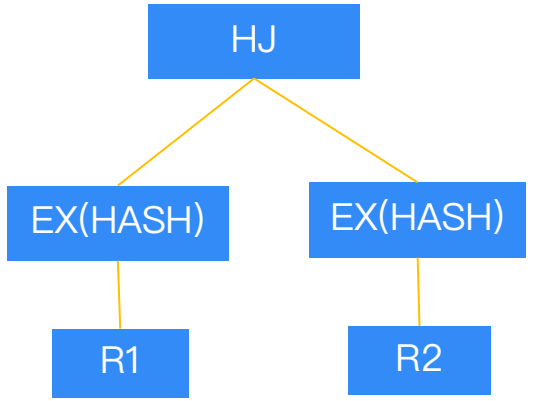
Partial Partition Wise Join

R1.b = R2.a



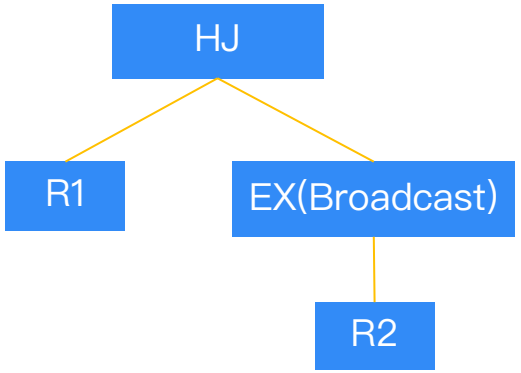
Hash–Hash Distribution Join

R1.a = R2.a



Broadcast Distribution Join

R1.a = R2.a



分布式连接算法

自适应执行

OCEANBASE

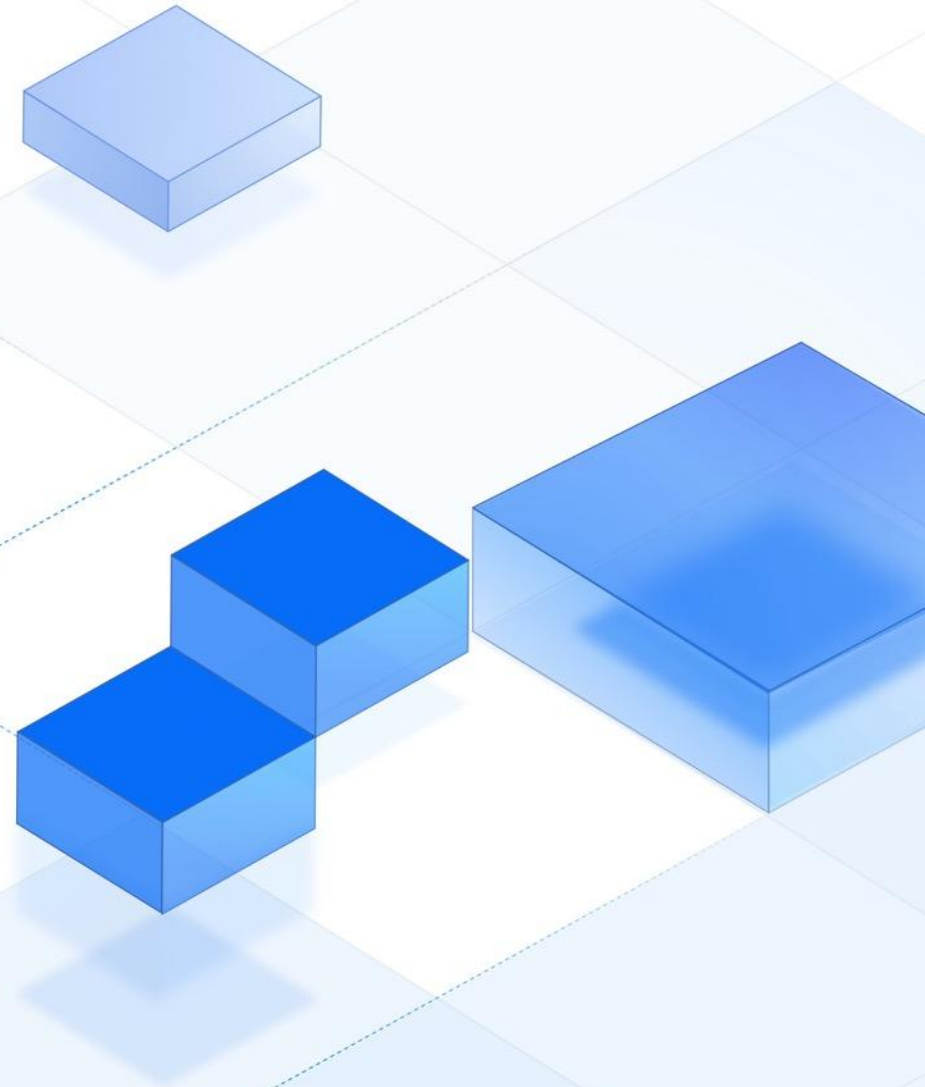
```
create table R1(a int primary key, b int, c int) partition by hash(a) partitions 4;
select b, sum(c) from R1 group by b;
```

```
==========================================================
|ID|OPERATOR                  |NAME     |EST. ROWS|COST|
----------------------------------------------------------
|0 |PX COORDINATOR            |         |1        |10  |
|1 | EXCHANGE OUT DISTR       |:EX10001|1        |10  |
|2 |  HASH GROUP BY           |         |1        |9   |
|3 |   EXCHANGE IN DISTR      |         |1        |9   |
|4 |    EXCHANGE OUT DISTR (HASH)|:EX10000|1        |8   |
|5 |     HASH GROUP BY        |         |1        |8   |
|6 |      PX PARTITION ITERATOR|         |1        |7   |
|7 |       TABLE SCAN         |r1       |1        |7   |
==========================================================
```

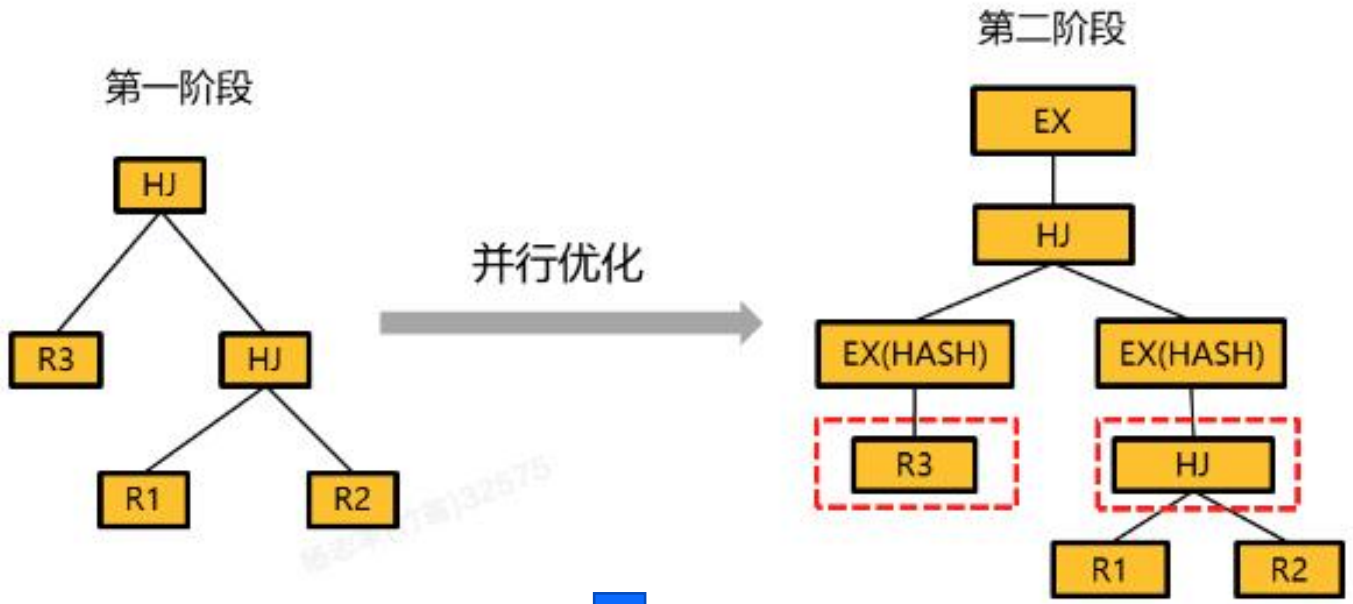
Group by/Distinct下压

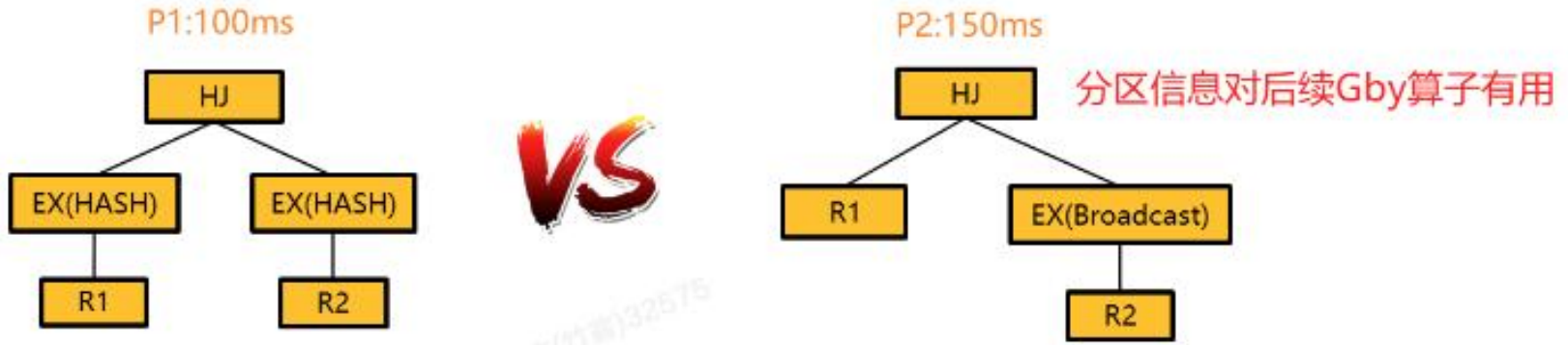
- 优化器总是下压
- 执行时基于实际数据特征决定是否跳过下压的算子

高级查询优化器



一阶段分布式查询优化





两阶段变为一阶段
- 避免不优的计划
- 执行计划包含分区位置信息
- 秒级完成50表连接的优化

并行下压

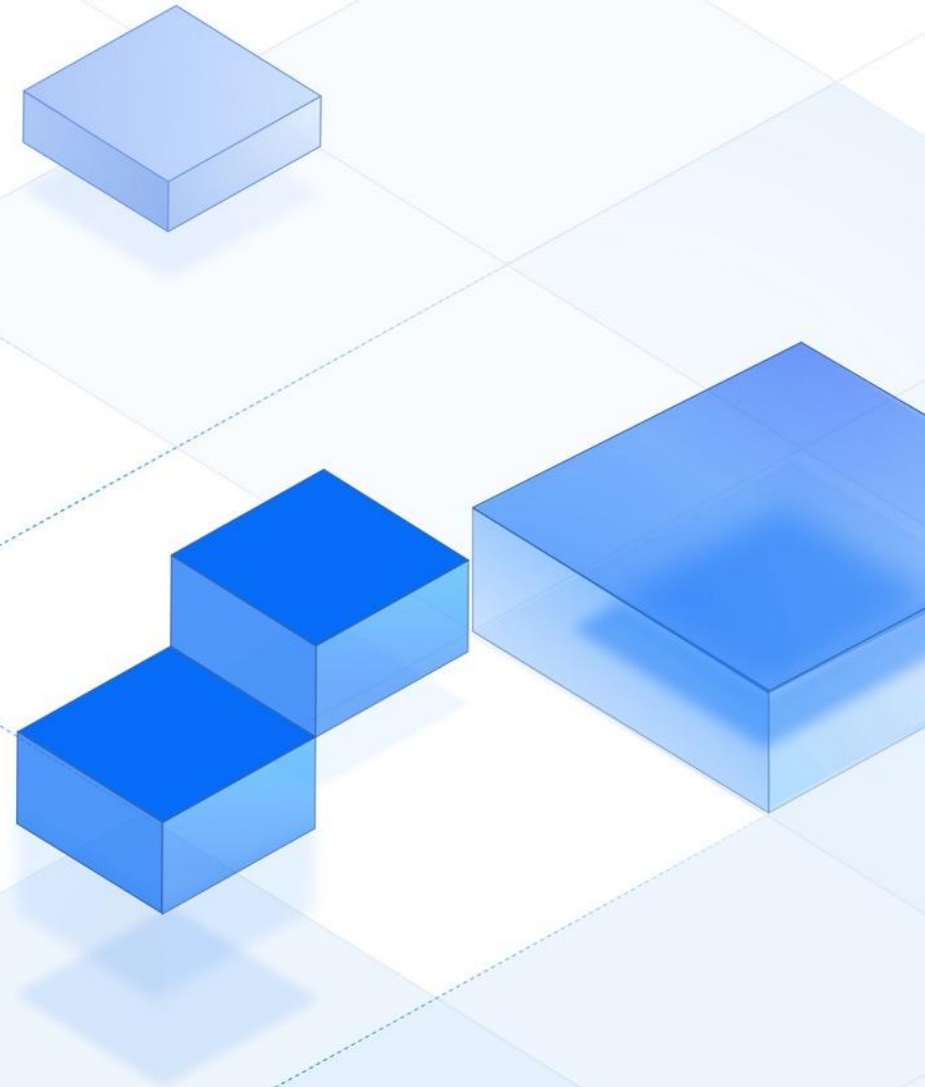
下压场景	例子	v3.2	v4.0
Group by, 无distinct去重的聚合函数	select a, sum(d) from t group by a;	支持	支持
Group By, 有distinct去重的聚合函数	select a, sum(distinct c),count(distinct d) from t group by a;	不支持	支持
Rollup	select a, sum(d) from t group by a rollup(b);	不支持	支持
Distinct	select distinct a from t;	支持	支持
Window Function	select a, b, sum(d) over (partition by c) from t;	不支持	支持

```
create table R1(a int, b int, c int, d int, primary key(a, b)) partition by hash(b) partitions 4;
select sum(distinct c) from R1 where a = 5;
```

```
=====================================
|ID|OPERATOR                |NAME    |
-------------------------------------
|0 |SCALAR GROUP BY         |        |
|1 | PX COORDINATOR         |        |
|2 |  EXCHANGE OUT DISTR    |:EX10000|
|3 |   PX PARTITION ITERATOR|        |
|4 |    TABLE SCAN          |r1      |
=====================================
```

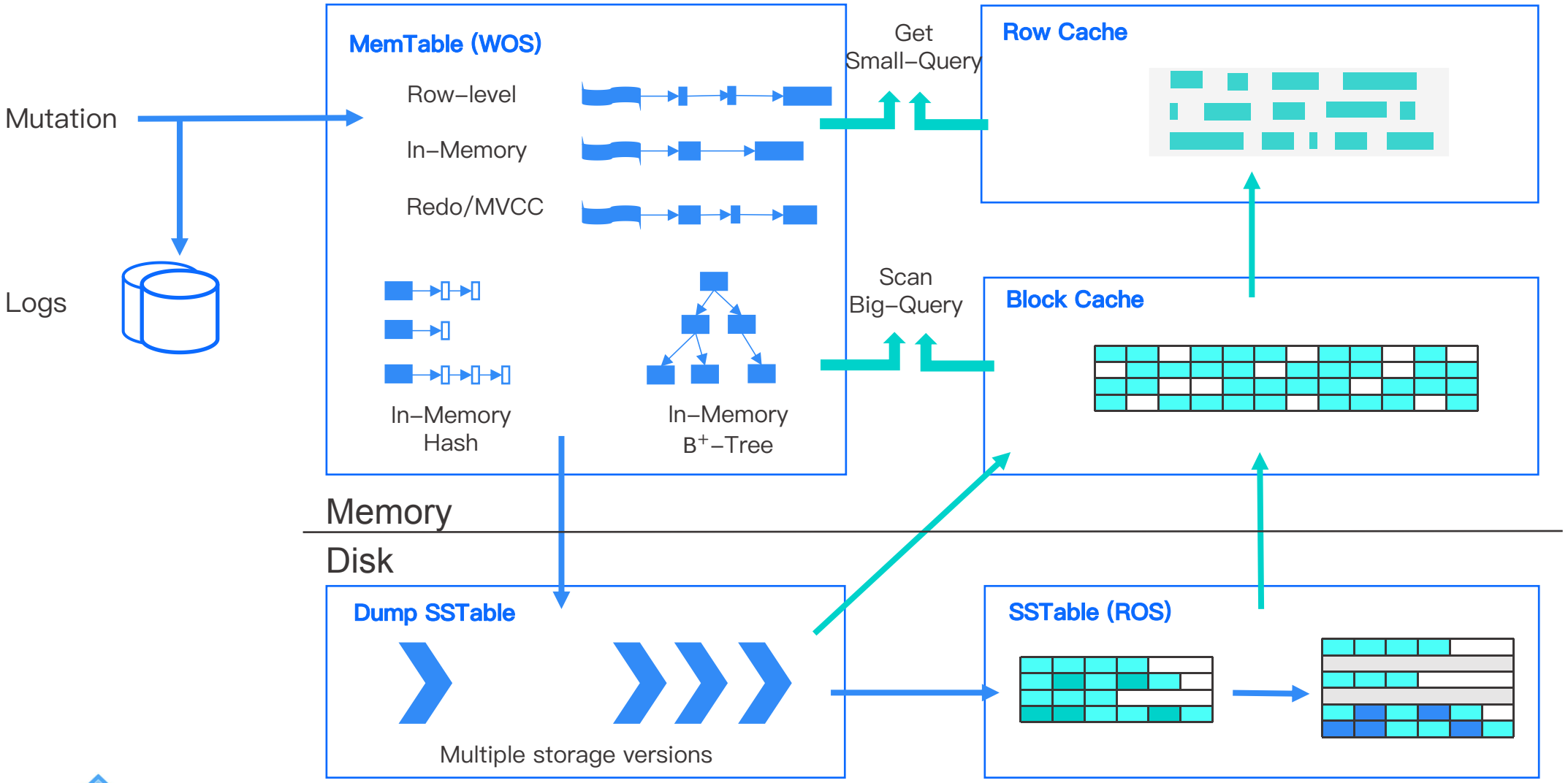
```
=============================================
|ID|OPERATOR                       |NAME    |
---------------------------------------------
|0 |SCALAR GROUP BY                |        |
|1 | PX COORDINATOR                |        |
|2 |  EXCHANGE OUT DISTR           |:EX10001|
|3 |   MERGE GROUP BY              |        |
|4 |    EXCHANGE IN DISTR          |        |
|5 |     EXCHANGE OUT DISTR (HASH)|:EX10000|
|6 |      HASH GROUP BY            |        |
|7 |       PX PARTITION ITERATOR   |        |
|8 |        TABLE SCAN             |r1      |
=============================================
```

行列混合存储



OceanBase存储引擎

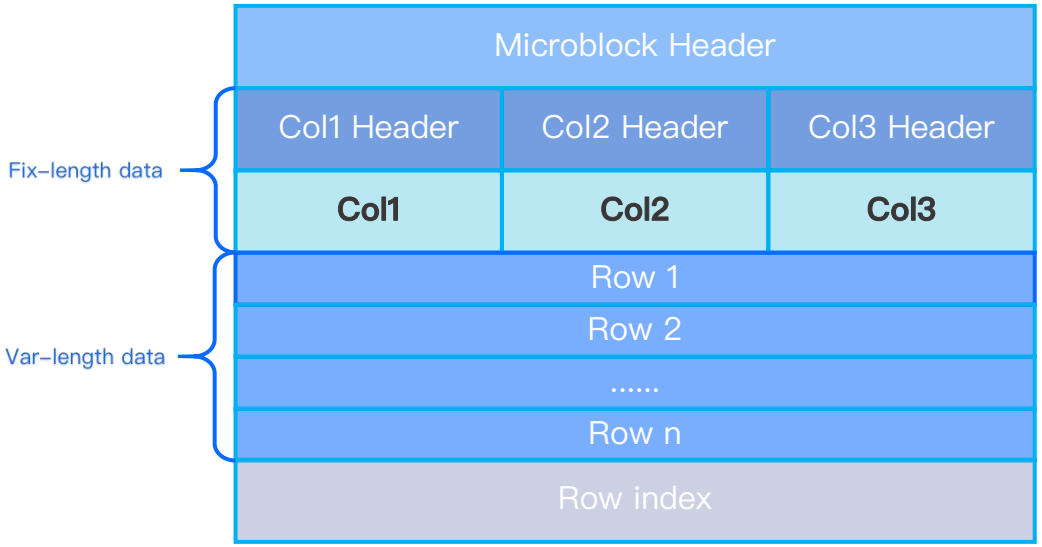
OCEANBASE



行列混合存储及编码压缩

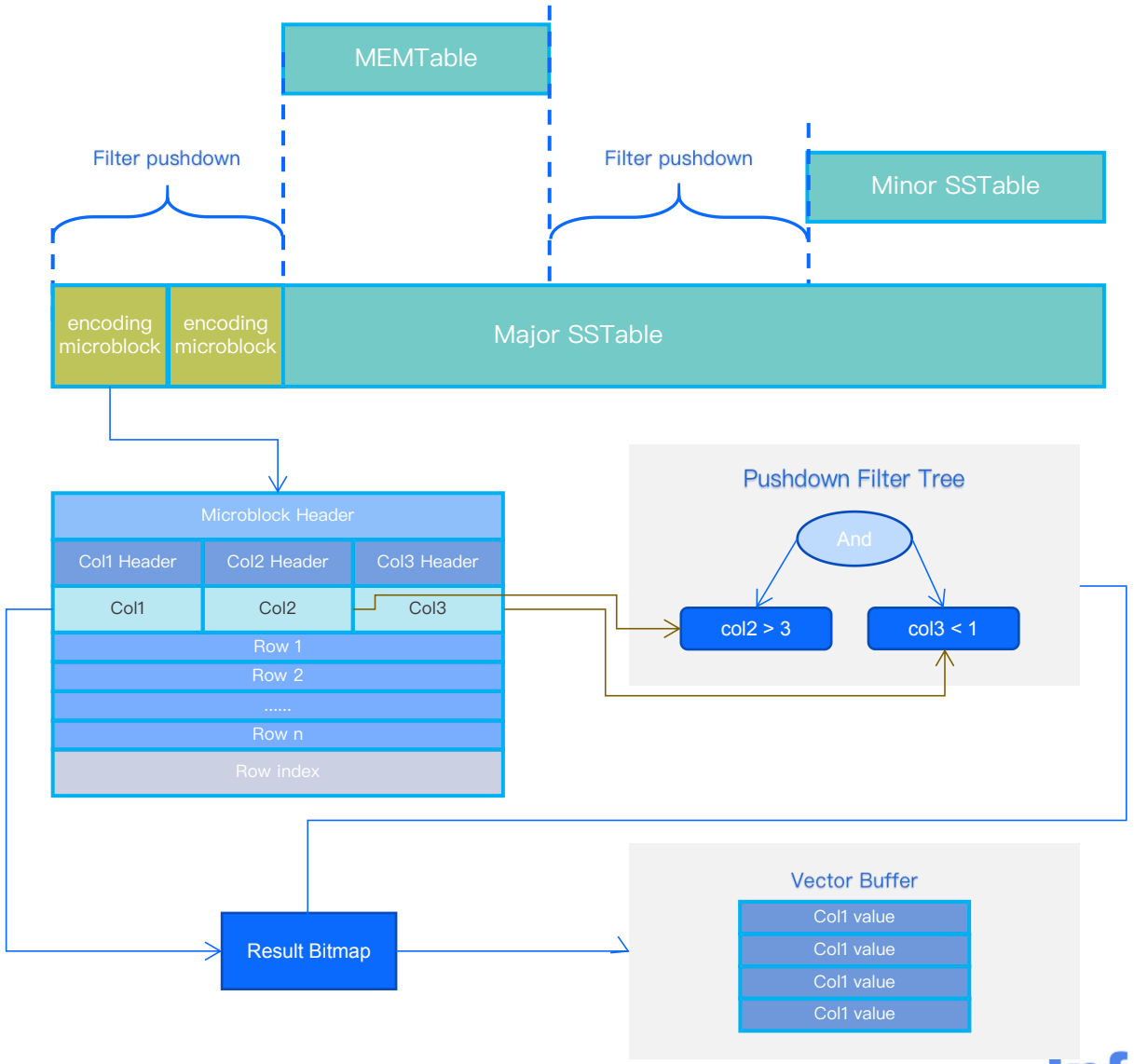
OCEANBASE

- 编码
  - 按列编码
  - 提升数据相似度
  - 规则发现
  - 微块自主选择
  - 经验推导
- 解码
  - 无需解压, 直接查询
- 效果
  - **存储空间是MySQL/Oracle 1/3**
  - 查询缓存使用效率提升

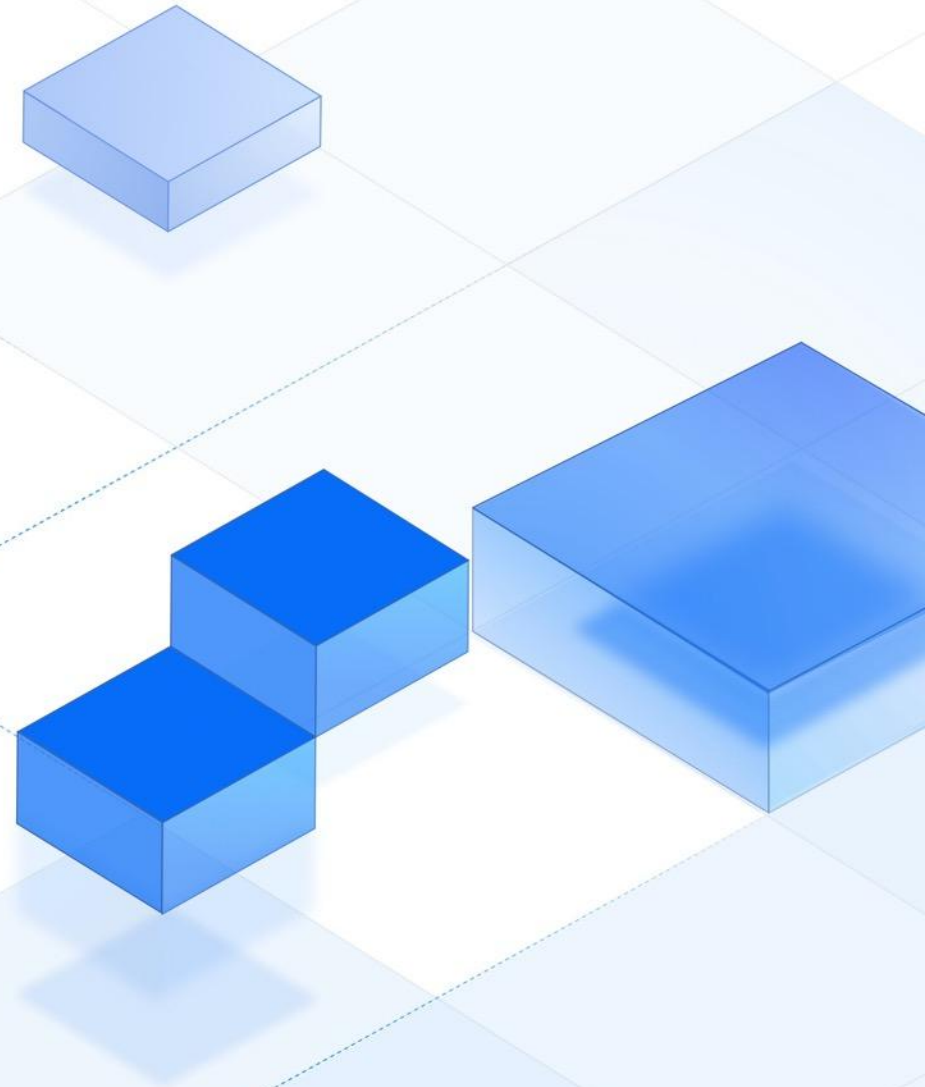


查询过滤下压

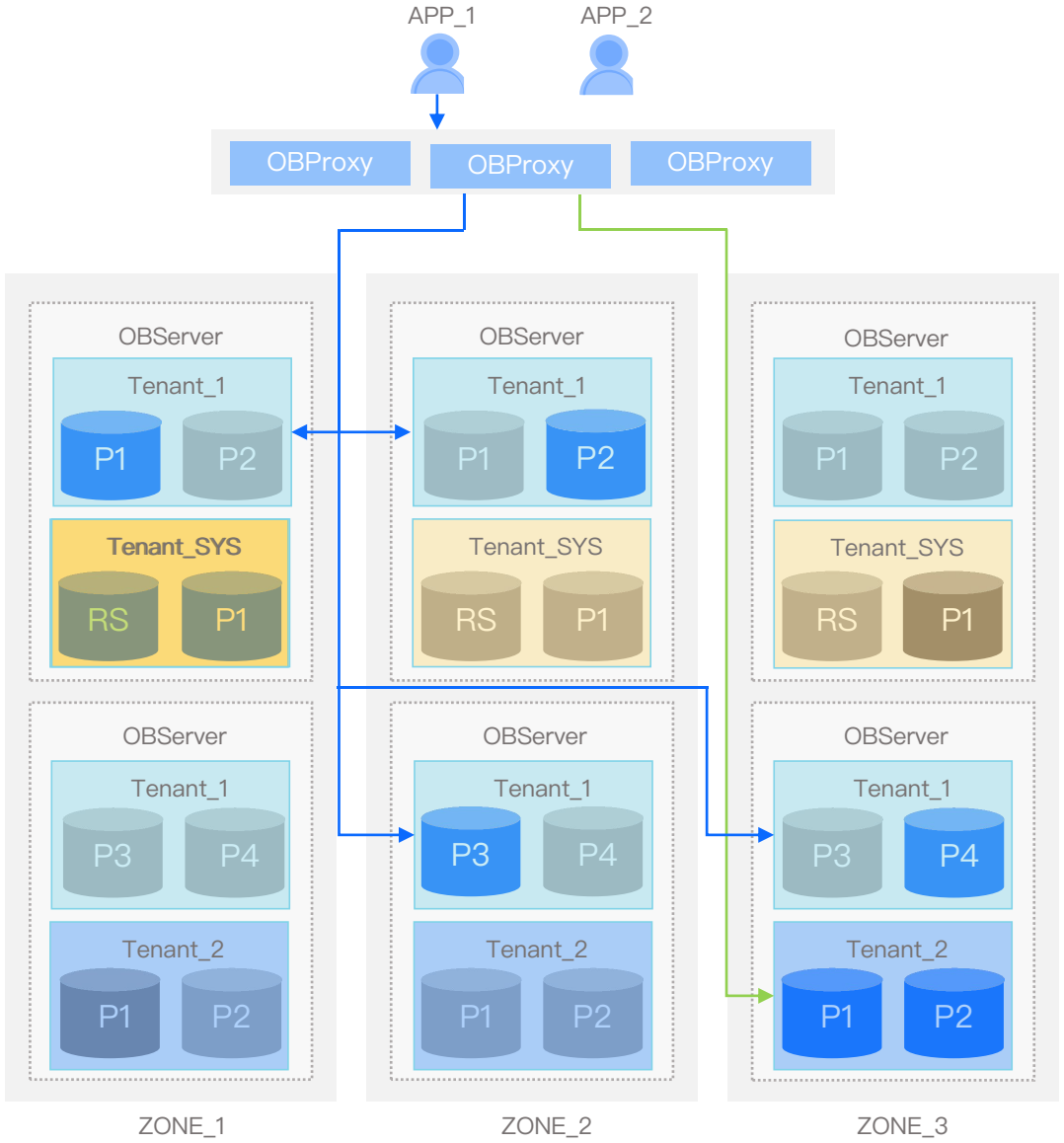
- 充分利用编码优势加速查询
- LSM-Tree难点
  - 增量数据交叉
- 谓词算子下压
  - 利用编码聚合信息快速过滤
  - 按列过滤充分利用剪枝
- 向量化
  - 按列批量解码
  - SIMD加速



HTAP资源隔离



多租户

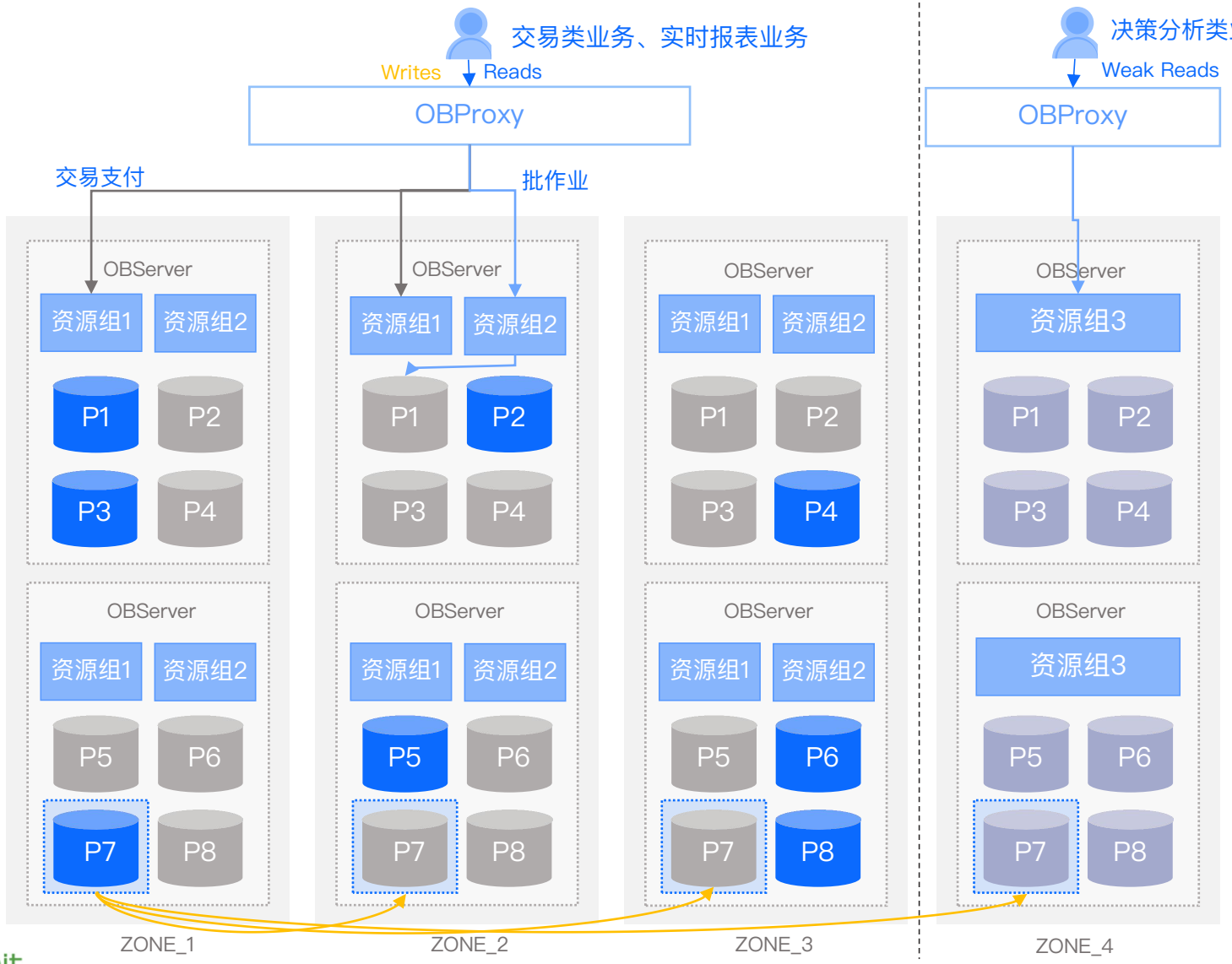


多租户实现	数据隔离	资源成本	资源隔离	运维复杂度
共享表	无（应用感知）	低	低（共享CPU,IO,网络）	低
schema	中（权限控制）	较低	低（共享CPU，IO，网络）	低
多数据库实例	高	高	高（独占CPU，IO，网络）	高
单实例多租户	高	中	中（独占CPU，共享IO、网络）	低

一个集群多个租户
- 多种租户类型并存
- 资源隔离与共享
- 大小租户独立扩缩容
- 统一运维管理

解决业务痛点
- 适合微服务应用架构
- 适合多租户SaaS服务
- 适合集团化数据管理

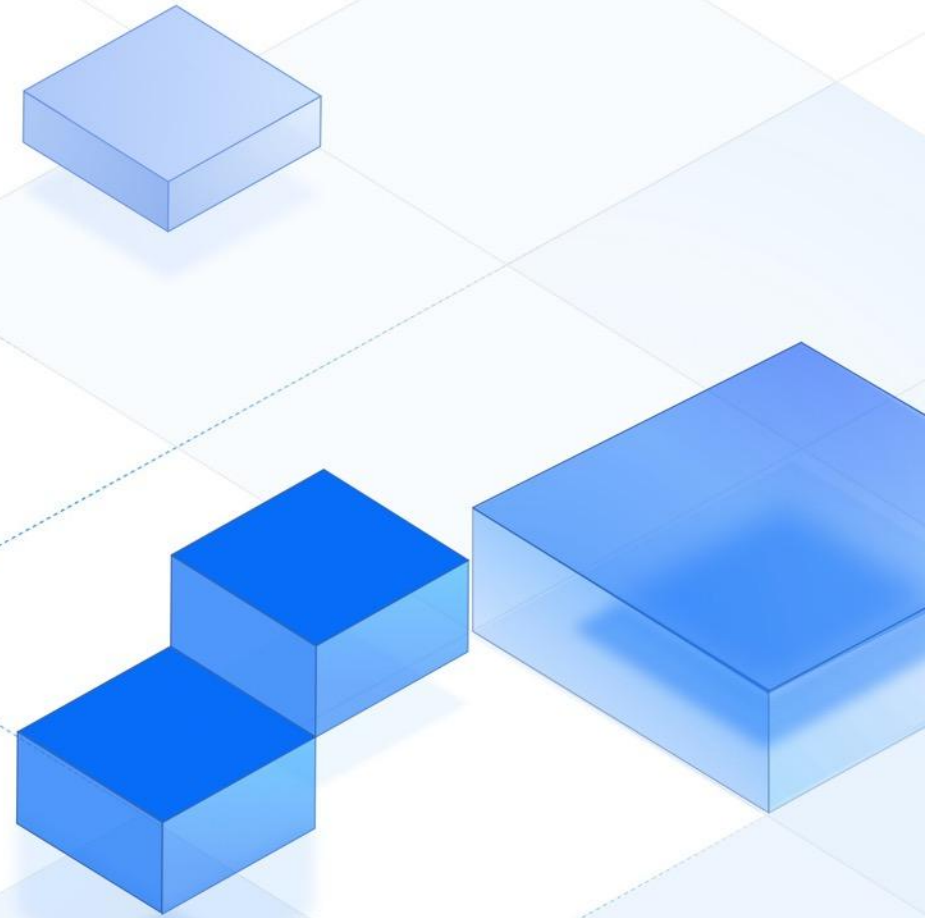
混合负载的资源隔离



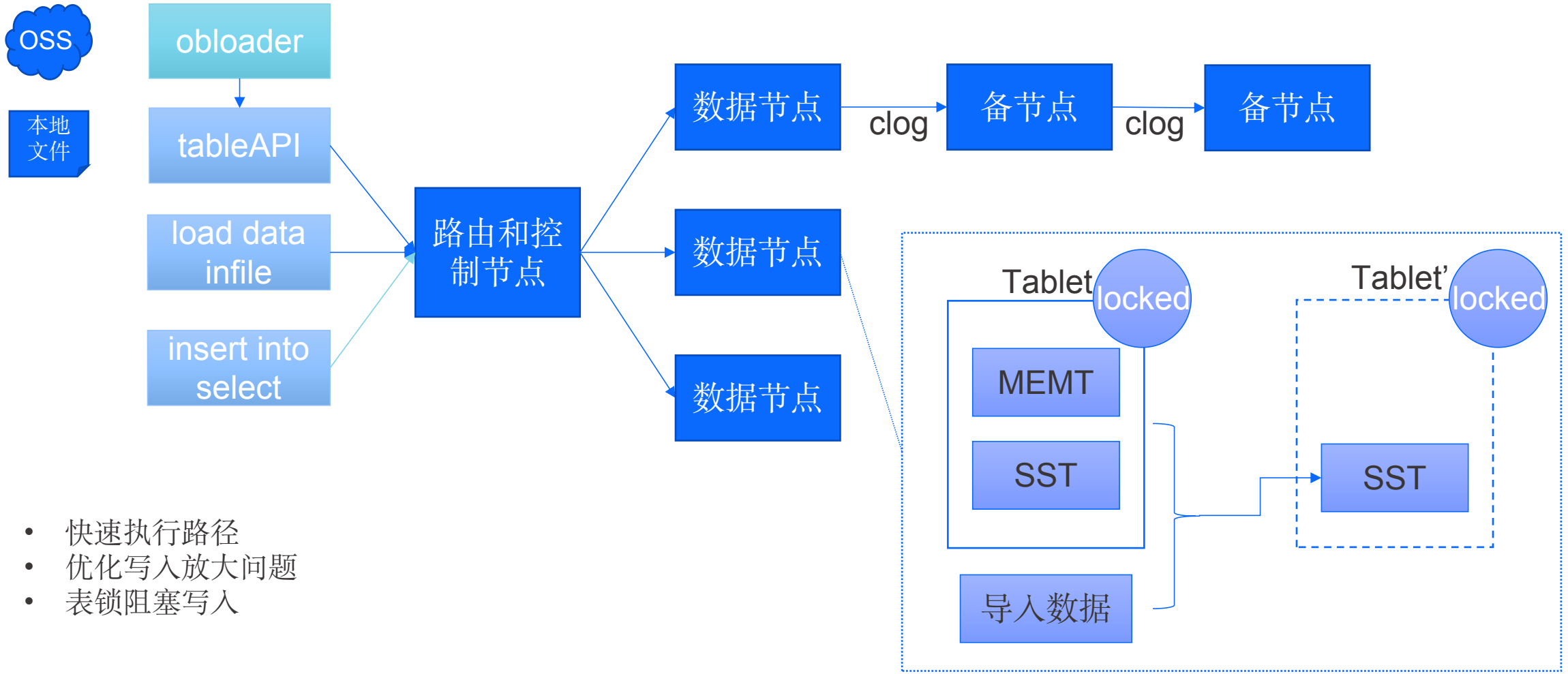
灵活的隔离机制

- 物理隔离
  - 弱一致性读读写分离
  - 多Zone读写分离
- 混合负载
  - 基于cgroup的资源组
    - 用户名匹配
    - SQL语句级匹配
  - 大查询自动隔离
    - 独立大查询队列

快速导入

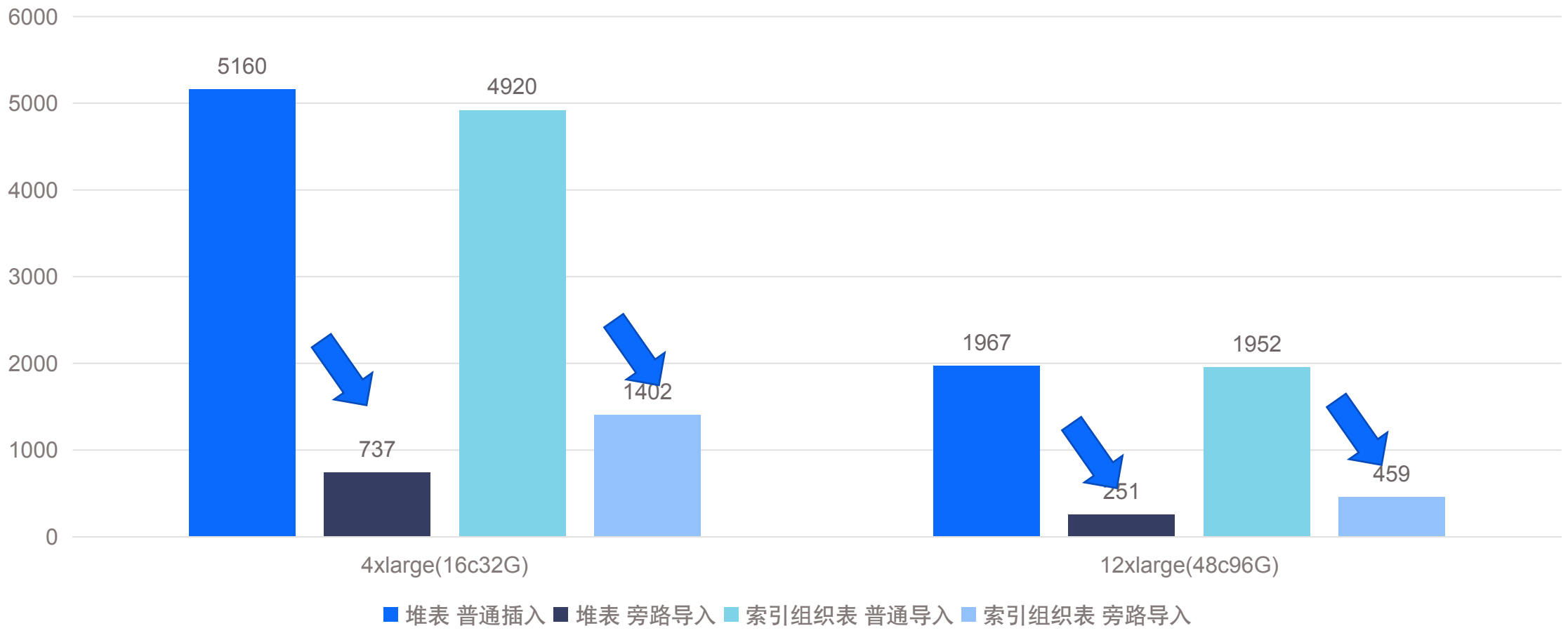


旁路（direct path）导入



- 快速执行路径
- 优化写入放大问题
- 表锁阻塞写入

旁路导入性能


6000
5000
4000
3000
2000
1000
0
5160
737
4920
1402
1967
251
1952
459
4xlarge(16c32G)
12xlarge(48c96G)
堆表 普通插入
堆表 旁路导入
索引组织表 普通导入
索引组织表 旁路导入

小结：OceanBase的OLAP能力和特性

OCEANBASE

基本能力

- 稳定可靠、高可扩展、高可用
- 并行执行引擎
- 高级查询优化器
- 低成本高性能行列混合存储
- 多租户与HTAP资源隔离

OLAP功能特性

- 复杂查询（大量表JOIN、复杂子查询）
- 分析函数（窗口函数、rollup）
- 层次查询（connect by）
- 表函数（from table）
  - 自定义管道函数（pipelined table）
- JSON、GIS类型
- 用户自定义函数UDF
  - 自定义聚集函数
- 异构数据库集成：dblink
- 导入：load data infile, obloader, 快速导入
- 导出：select into outfile, obdumper
- 联邦查询：外表

想一想，我该如何把这些技术应用在工作实践中？

THANKS