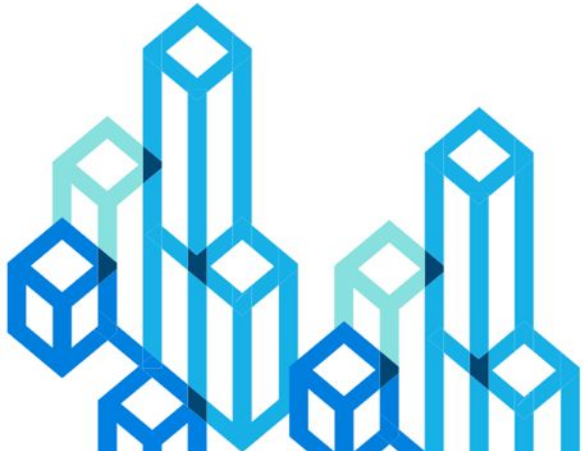
The Evolution of TiDB Architecture



About Me

- 张建，开源爱好者，TiDB Maintainer
- TiKV 产品负责人@PingCAP
- 前 MaxCompute 执行引擎研发@Alibaba
- 关注数据库、分布式系统、开源软件与社区等领域
- Email: zhangjian@pingcap.com

Agenda

01 What's TiDB

02 The Way to TiDB 5.0

03 Future Work

What's TiDB

What's TiDB

TiDB：全新一栈式实时 HTAP 数据库

- 一键水平扩容或者缩容
- 金融级高可用
- 实时 HTAP
- 云原生的分布式数据库
- 兼容 MySQL 5.7 协议和 MySQL 生态

助推全球 1500+ 企业互联网化和数字化

10+ 银行与头部金融企业

- 网联交易、网贷核算
- 互联网业务中台
- 实时风控、反欺诈

三大运营商

- 支付与风控
- 征信与账单系统
- 精准营销

覆盖中国互联网独角兽企业的 80%

- 万亿量级数据毫秒响应
- 支付、秒杀等场景
- 单个集群1000+ 节点

物流与零售

- 实时统计报表
- 数据中台
- 交易与报表业务隔离

美国、日本、东南亚头部支付企业与电商

- 在线支付
- 用户与订单管理
- 风控与营销

OLTP 规模化的头部用户

面向零售高增长交易场景

全球领先餐饮巨头

- **实时高并发**，高效支撑短时高峰交易及数据访问，**1.5亿用户**
- **弹性高扩展**，TiDB on K8s 随时随地在线扩展
- **安全高可用**，可靠放心的高可用容灾保障能力

在线核心联机支付交易平台

日本头部金融支付企业

- 在线支付联机交易，钱包等**高增长交易支撑，3000万用户**
- **水平弹性扩展**和几乎线性的扩展能力
- **友好的开发界面**，应用几乎无需改动

面向金融敏态交易场景

平安人寿在线金融联机服务

- 支撑平安人寿“金管家”单日交易额破**1000亿，超1亿用户**
- **高可靠、低延迟、可快速扩展**
- **大大降低敏态应用开发复杂度**，加快应用上线速度
- **在线弹性扩缩容**满足不确定业务需求

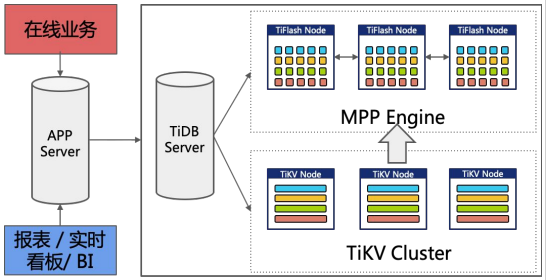
TiDB HTAP：一栈式数据服务生态

混合负载场景

汽车之家 看车·买车·用车

交易处理和在线分析同步规模化

报表业务不影响在线交易

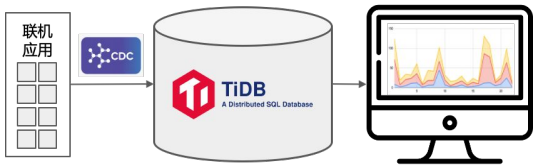


流式计算场景

360 www.360.cn

实时到账、实时统计

缩短业务等待周期，提升用户体验

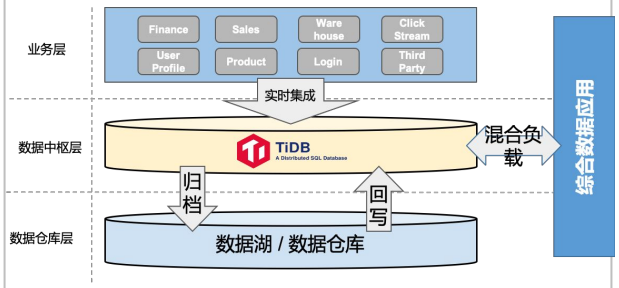


数据中枢场景

贝壳 找房大平台

多源汇聚，统一数据服务

节省系统和数据聚合代码开发时间

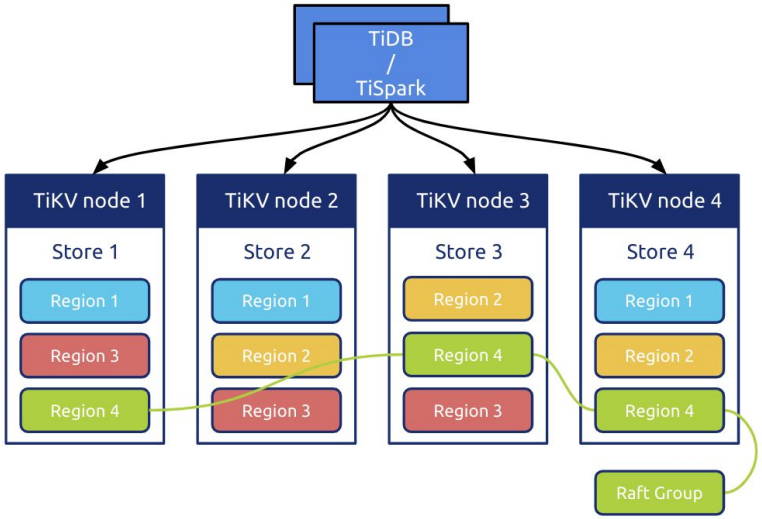


写 SQL 就行了

一个TiDB 系统，一个访问入口，一份数据

The Way to TiDB 5.0

TiDB 的初心



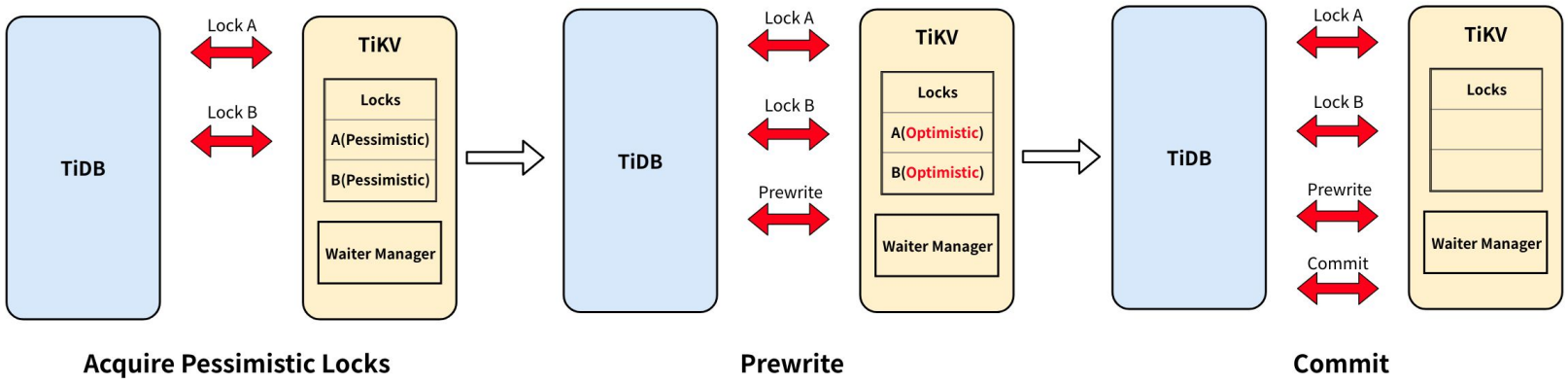
- 从 DBA 熬夜的日子出发
- 设计像 iPhone 一样简单
- 所有能够分离的都分离
- 回归经典数据库体验

金融核心的第一个挑战：悲观锁

乐观事务：中间的 DML 语句不检查锁，Commit 时再检查索冲突

不足：

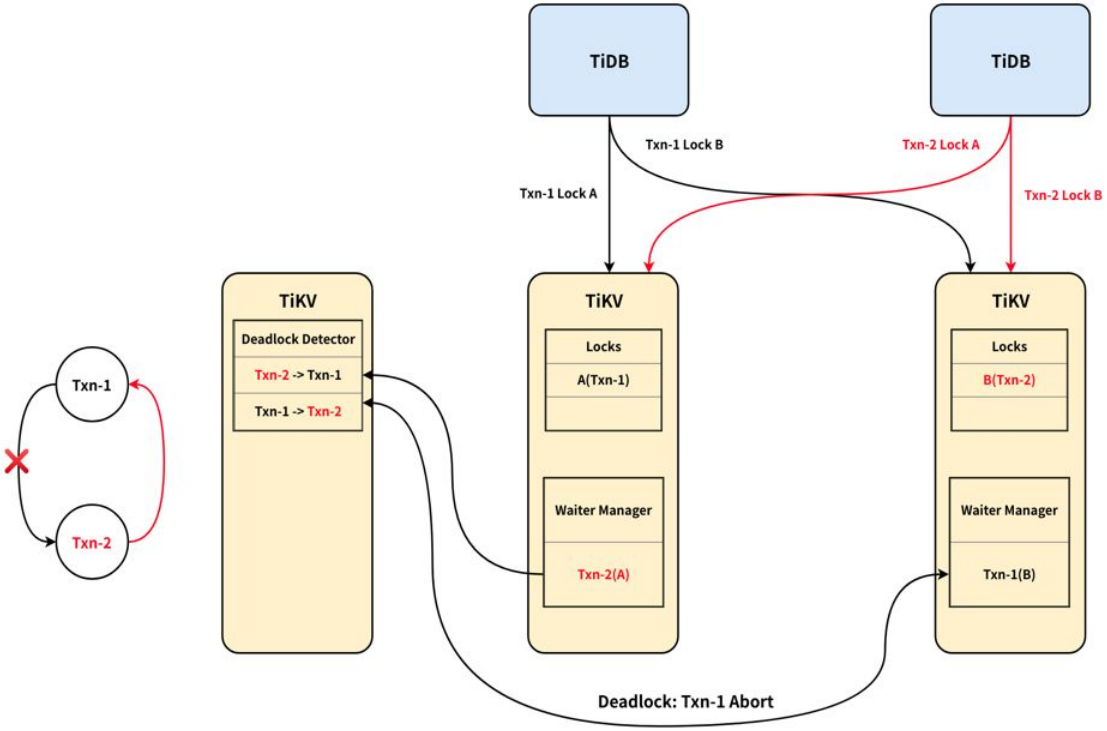
1. 高并发冲突的情况下，性能严重下降
2. 大量应用基于悲观锁设计，学习和改造成本高



死锁检测

TiDB 新特性漫谈：悲观事务

- 动态选举出一个 TiKV 负责死锁检测
- 悲观事务在等锁的时候，异步进行死锁检测

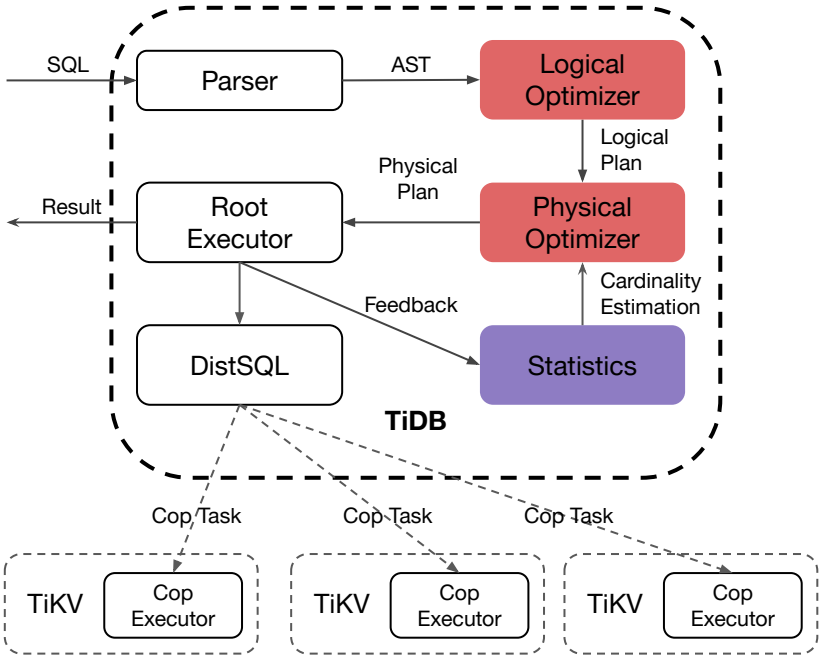


挑战二：执行计划的稳定性

场景一：线上 SQL 忽然执行计划变了，选错了索引导致慢查询，业务运行异常

场景二：升级后，因为优化器的改动导致执行计划变化，升级后业务运行异常

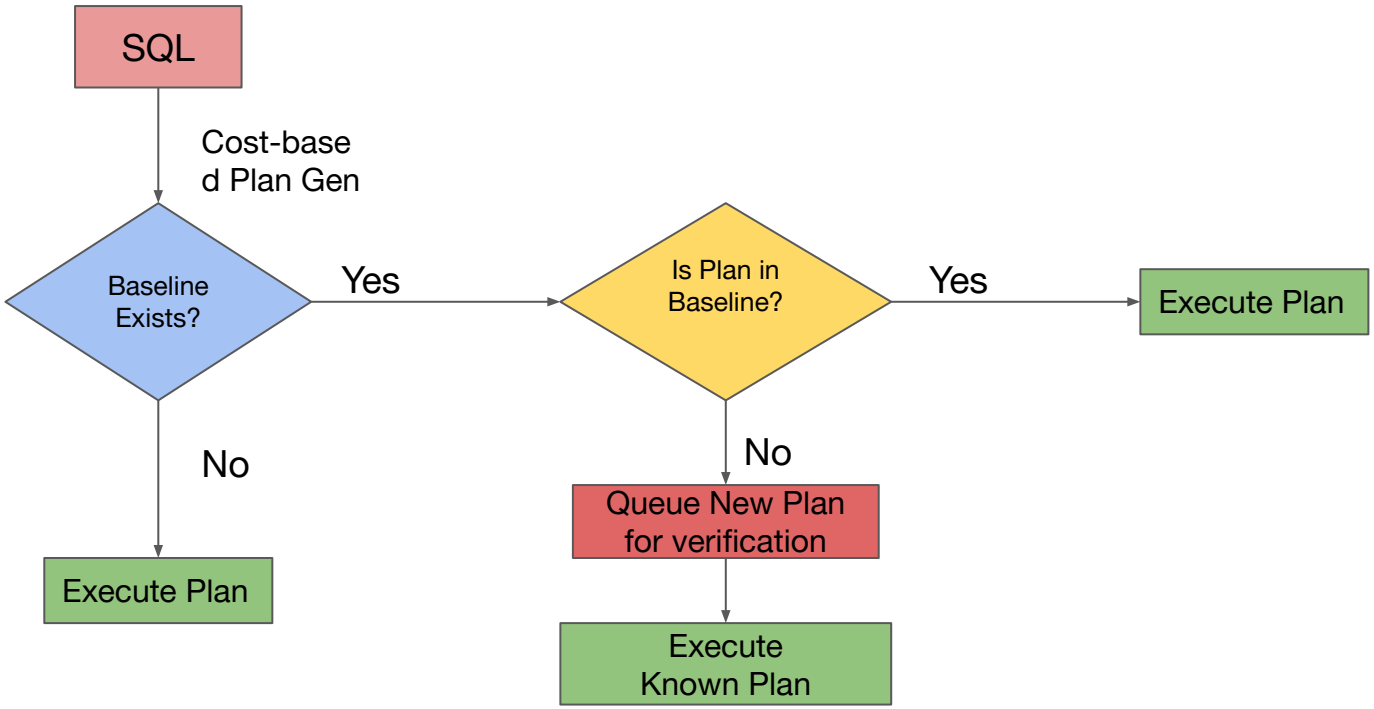
...



综合问题需要综合治理

先提供规避和手工纠正的能力：

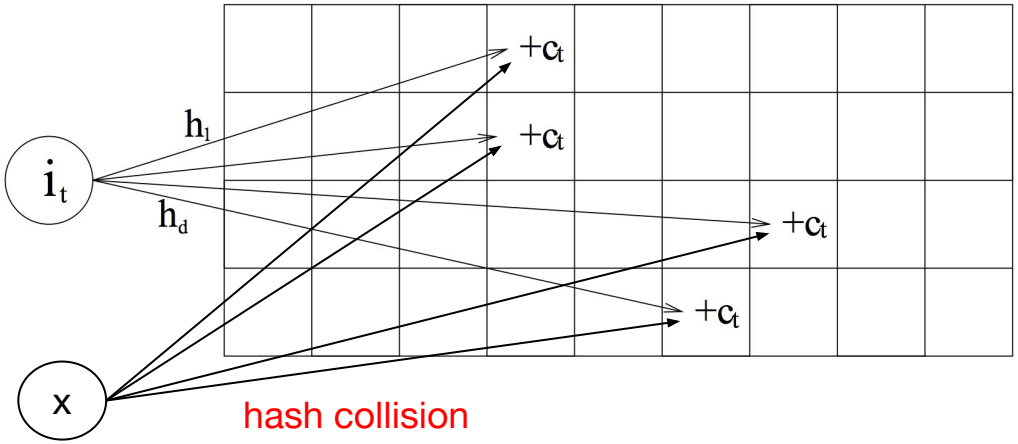
- 需要改业务 SQL：增加 SQL Hint，从 TiDB v1.0 加到 v5.0，形成了一套体系
- 不需要修改业务：从 SQL Binding（TiDB v3.0）
- 不想一条条加：SPM（TiDB v4.0）



执行计划的稳定性:综合问题需要综合治理

优化内功:

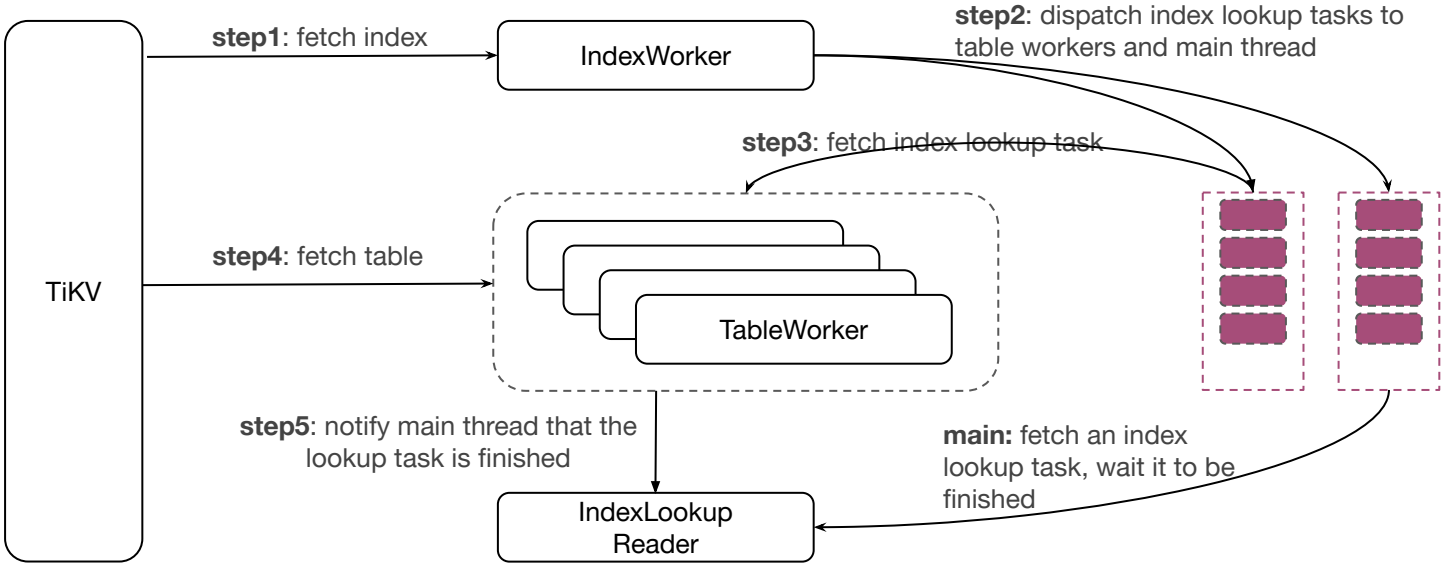
- 降低犯错的概率:引入 Skyline Pruning 裁剪不优索引，引入更多启发式规则
- 提升做对的能力:重构统计信息，降低基数估算误差
- 建立优化器测试评估体系



挑战三：内存管理

情况一：单个大查询

情况二：高并发中等查询



挑战三：内存管理

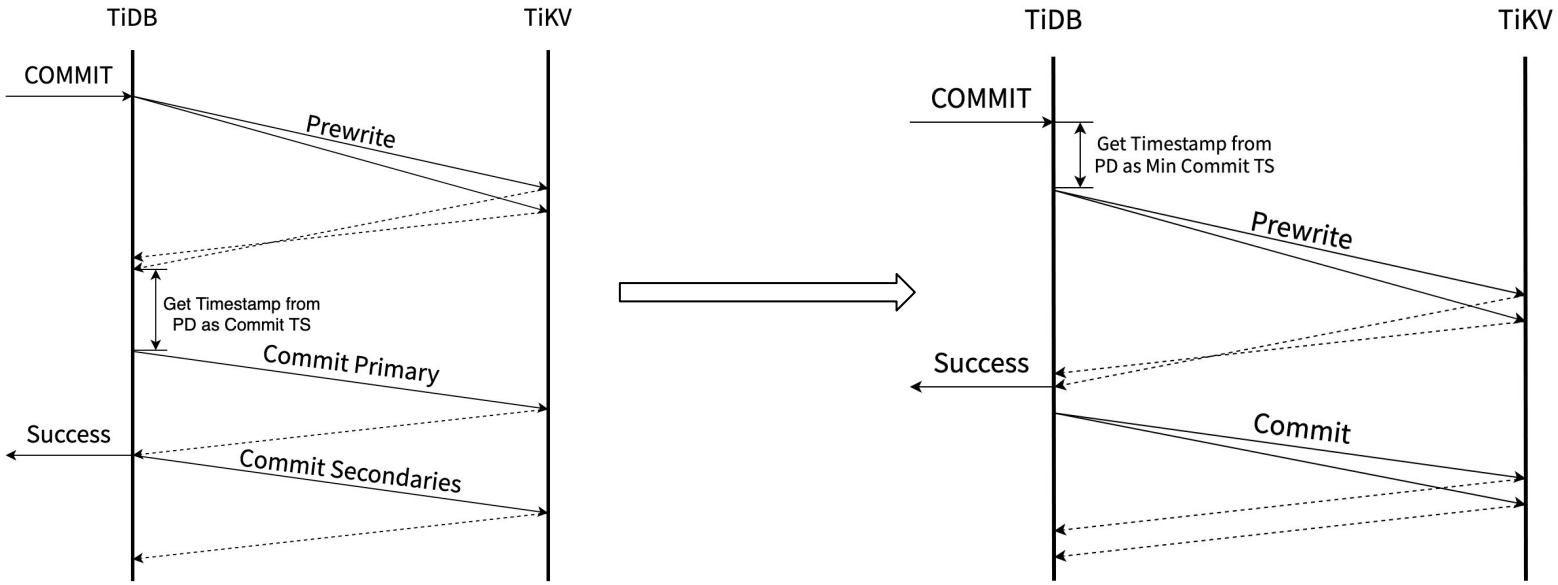
情况一：单个大查询

情况二：高并发中等查询

- 隔离火势：追踪内存使用，当超过阈值时，能够将该查询 kill 掉，不影响其他查询
- 灭火：通过算子落盘，让内存占用过多的大查询不被 kill 顺利执行完

挑战四：事务延迟

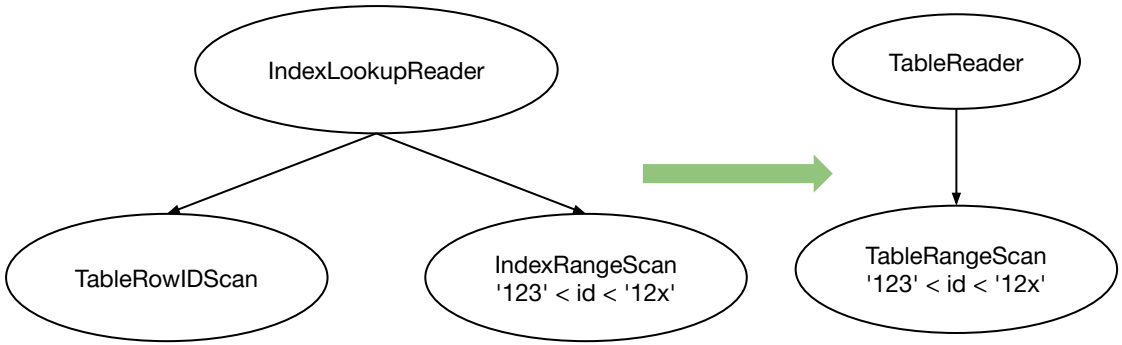
Async Commit（TiDB v5.0）：把 2PC 变为 1PC



挑战四：事务延迟

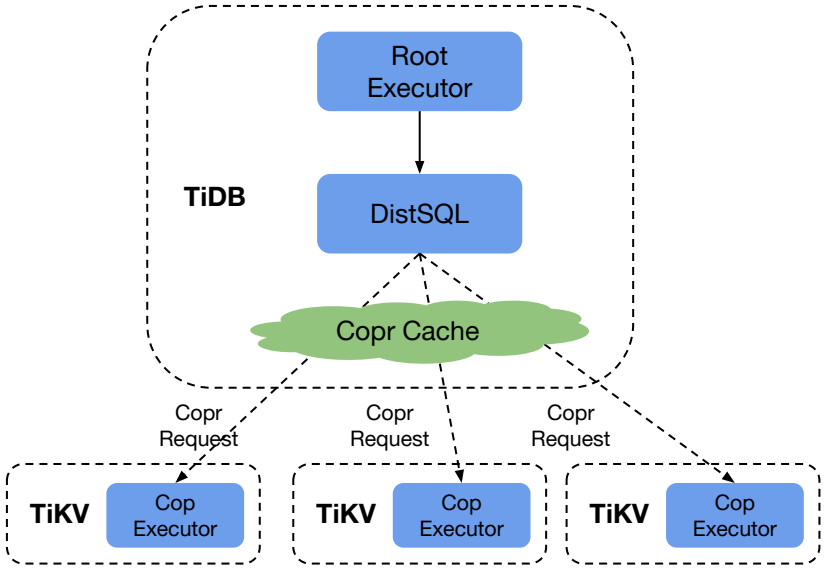
聚簇索引（TiDB v5.0）：减少回表次数，让查询时间减半

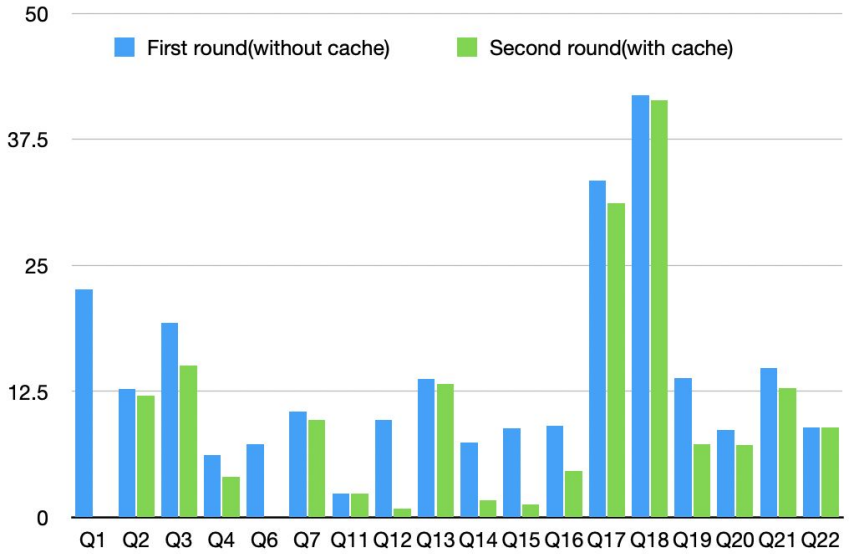
```
CREATE TABLE `user` (
  `id` varchar(20) NOT NULL,
  `name` varchar(20) DEFAULT NULL,
  PRIMARY KEY (`id`) CLUSTERED
)

SELECT * FROM user
WHERE id > '123'
  AND id < '12x';
```



挑战四：事务延迟

Coprocessor Cache（TiDB v5.0）：减少下推的计算量，降低重复查询的延迟





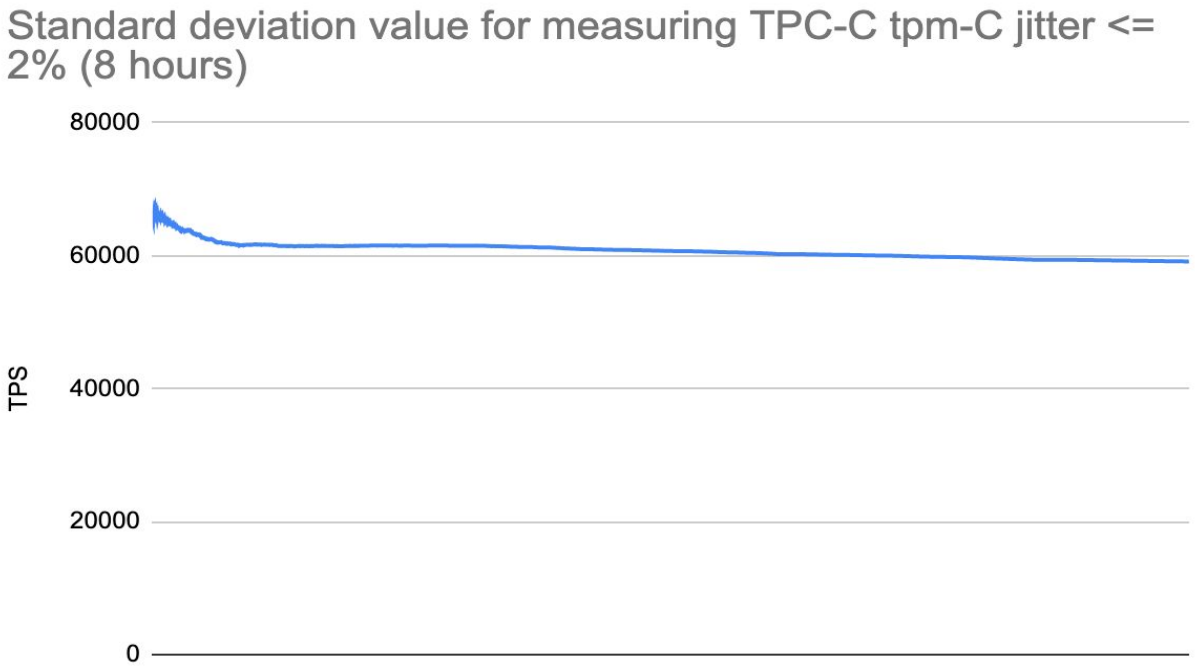
挑战五：后台任务对用户查询的影响

TiDB 有哪些后台任务：

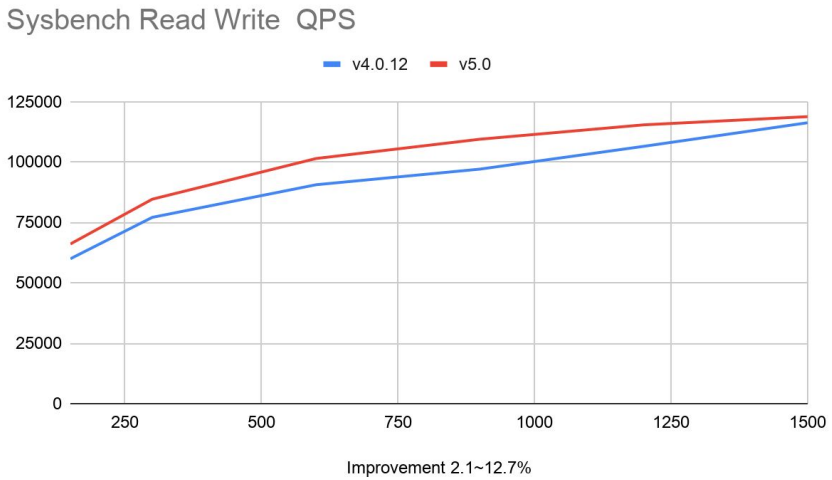
1. MVCC GC
2. RocksDB Compaction
3. Region Scheduling
4. Analyze Table
5. Update Global Variables
6. ...

挑战五：后台任务对用户查询的影响

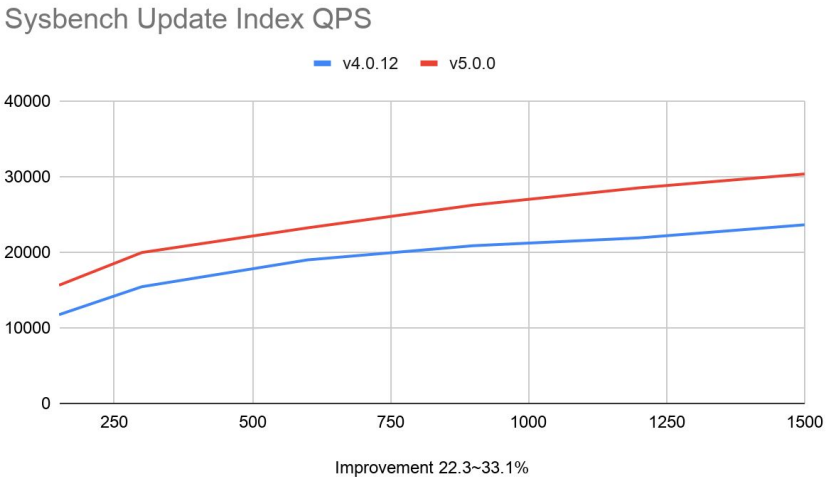
- MVCC GC in Compaction Filter
- RocksDB’s IO Rate Limiter
- Avoid redundant region rebalance caused by empty regions



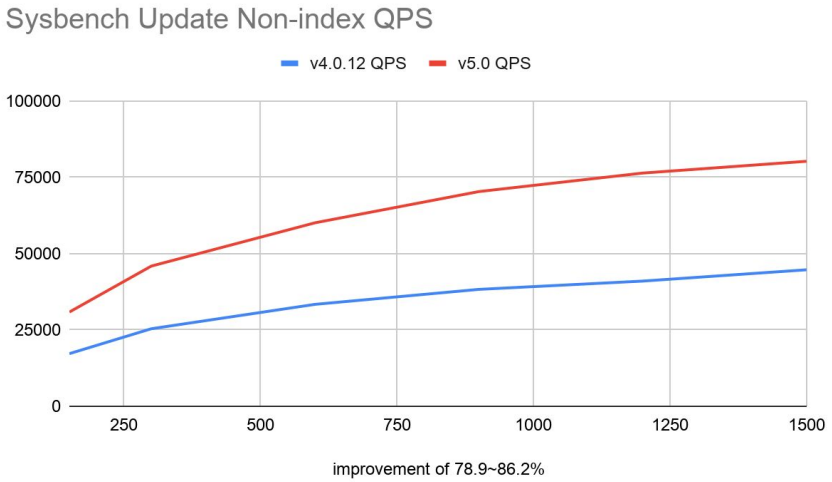
TiDB 5.0 Sysbench



Improvement 2.1~12.7%

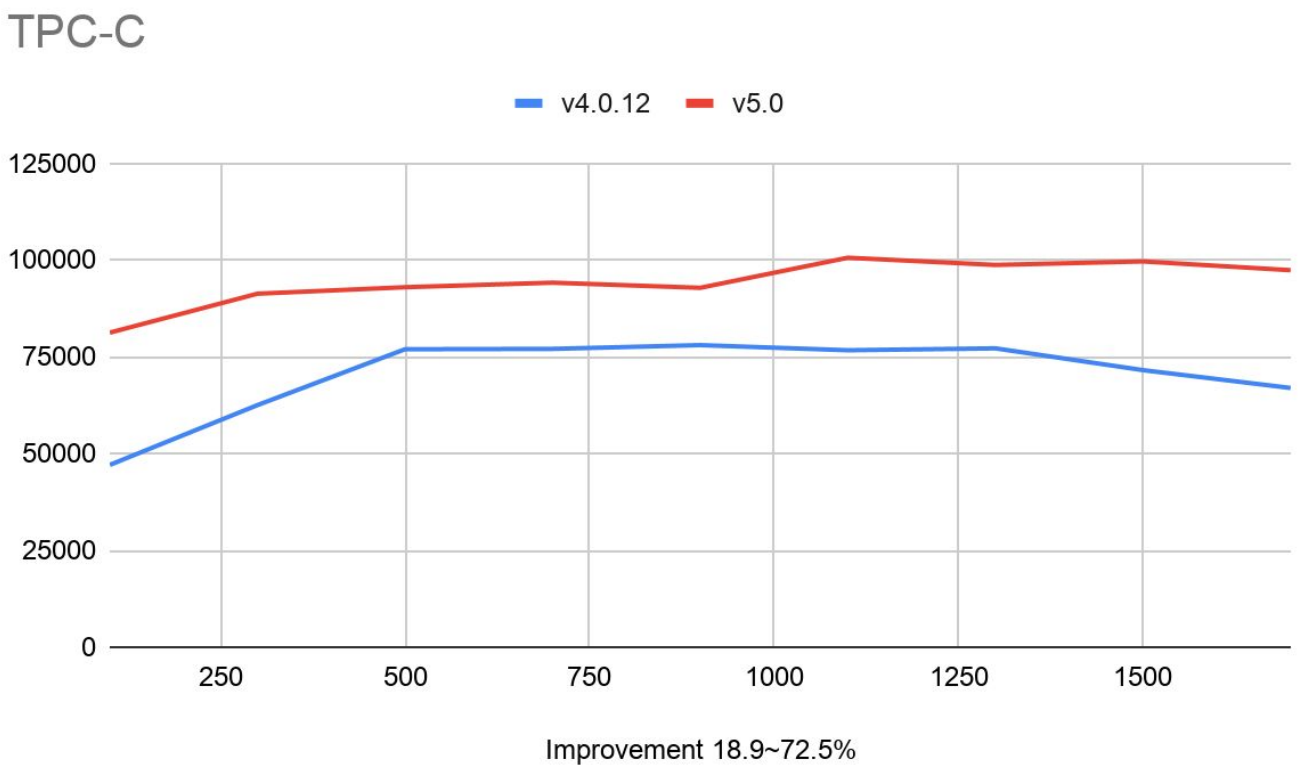


Improvement 22.3~33.1%

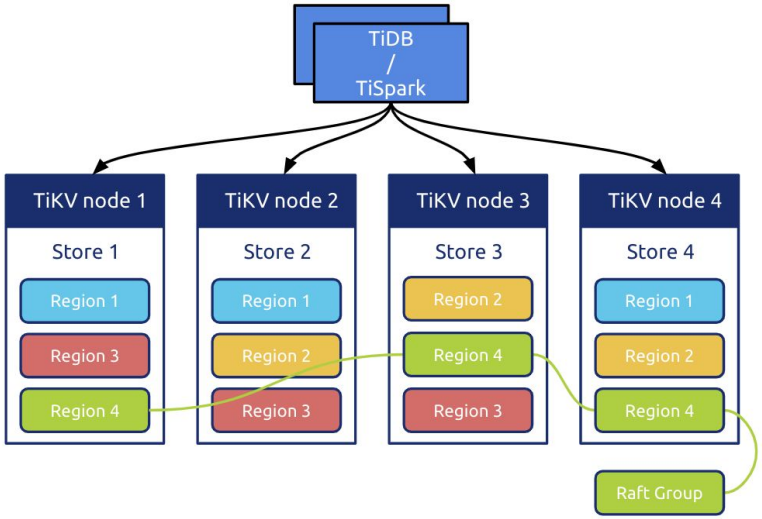


improvement of 78.9~86.2%

TiDB 5.0 TPC-C

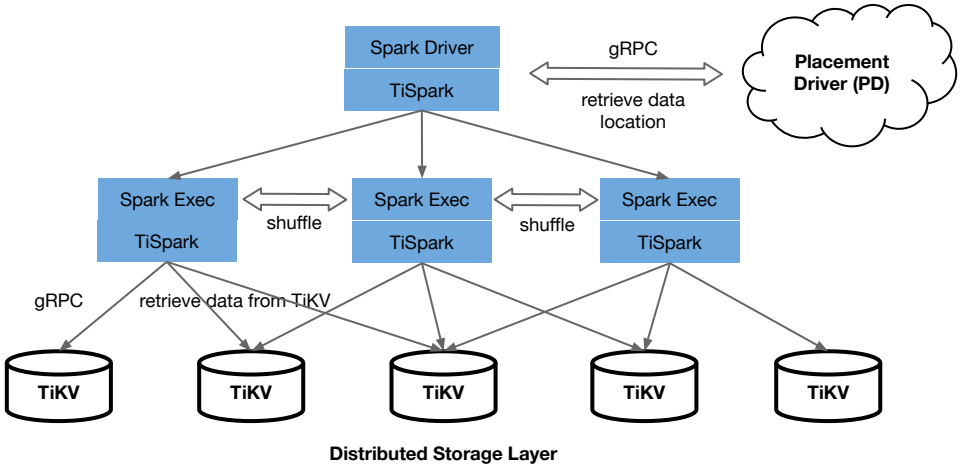


Improvement 18.9~72.5%



让我们回来看看 HTAP 的发展

TiSpark

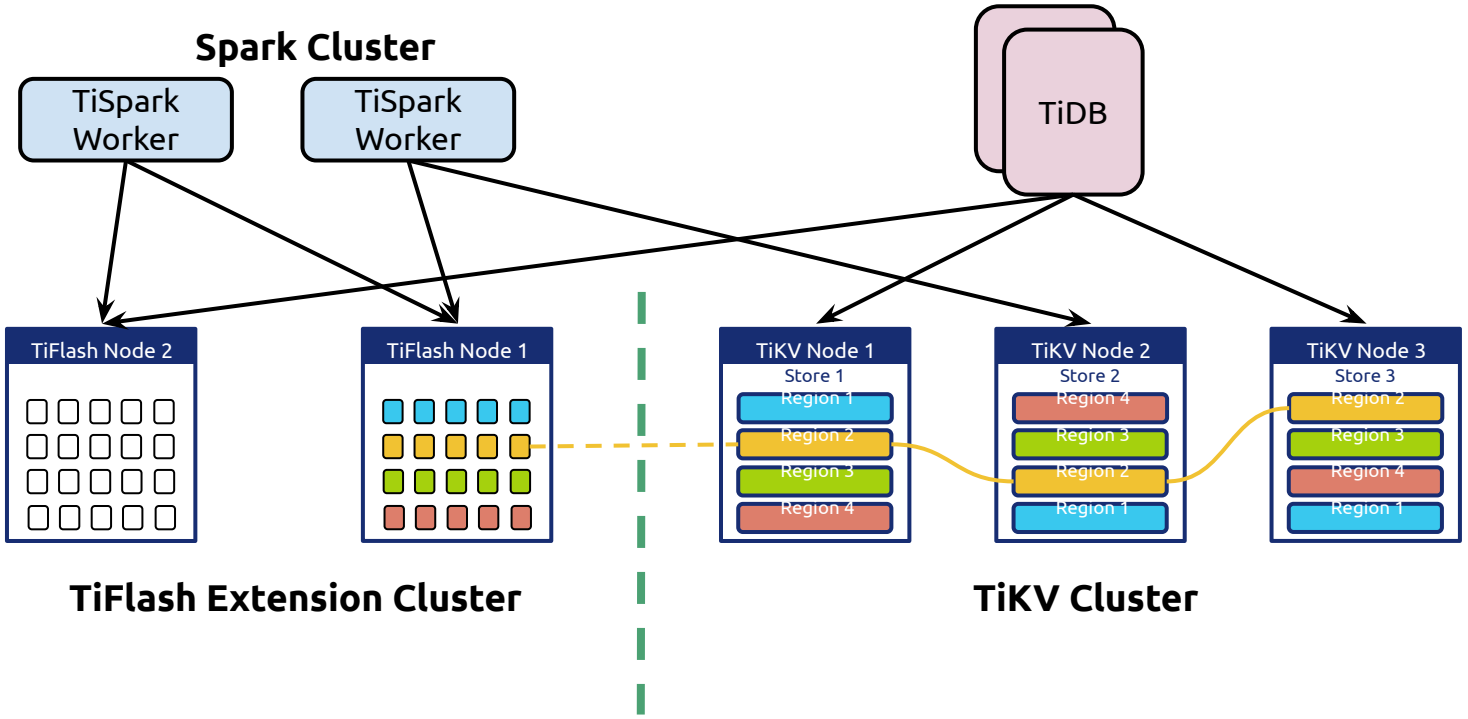


- 使用存储在 TiKV 的同一份数据
- 用 Spark 做分析查询

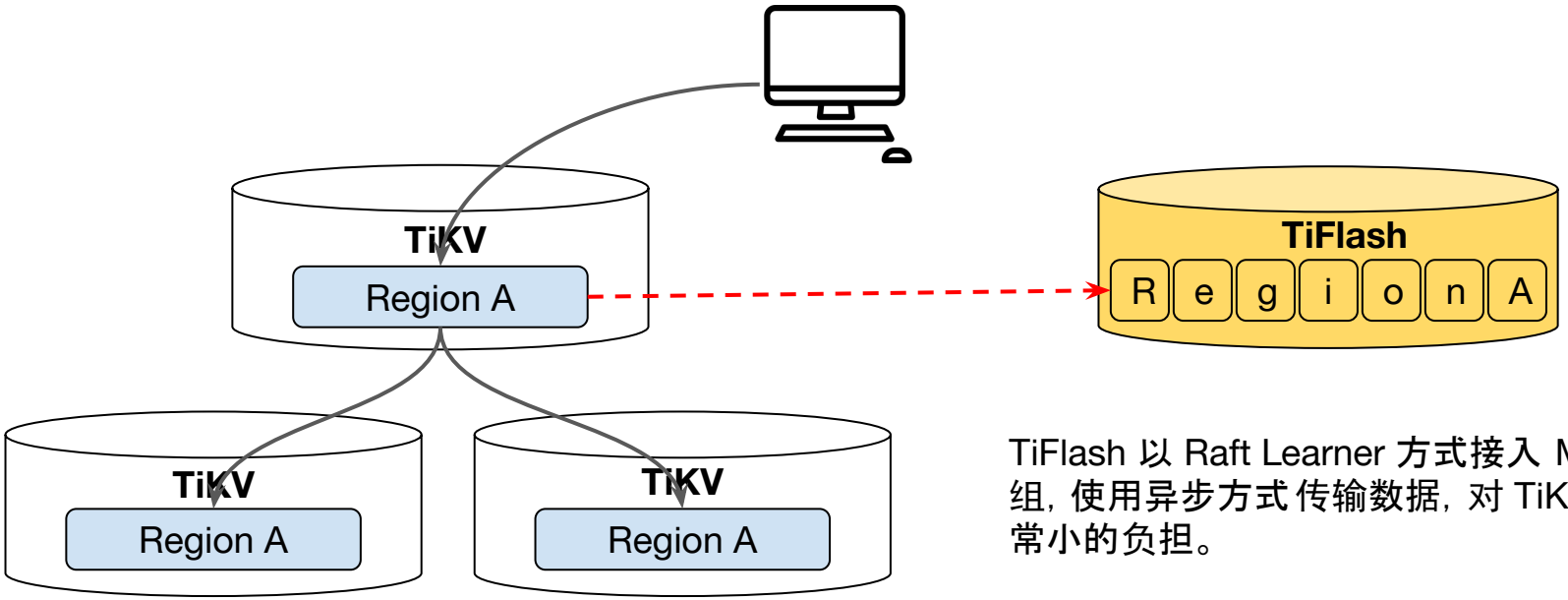
TiSpark 是否就足够了？

- 用户在很多场合下需要高并发中小规模 AP 能力
- TiKV 这样的 **行存** 无法满足分析场景需要的 **性能**
- TiKV 上跑在线业务时跑 OLAP 负载，没有 **资源隔离** 简直是个灾难

引入列存，与 TiKV 物理上隔离



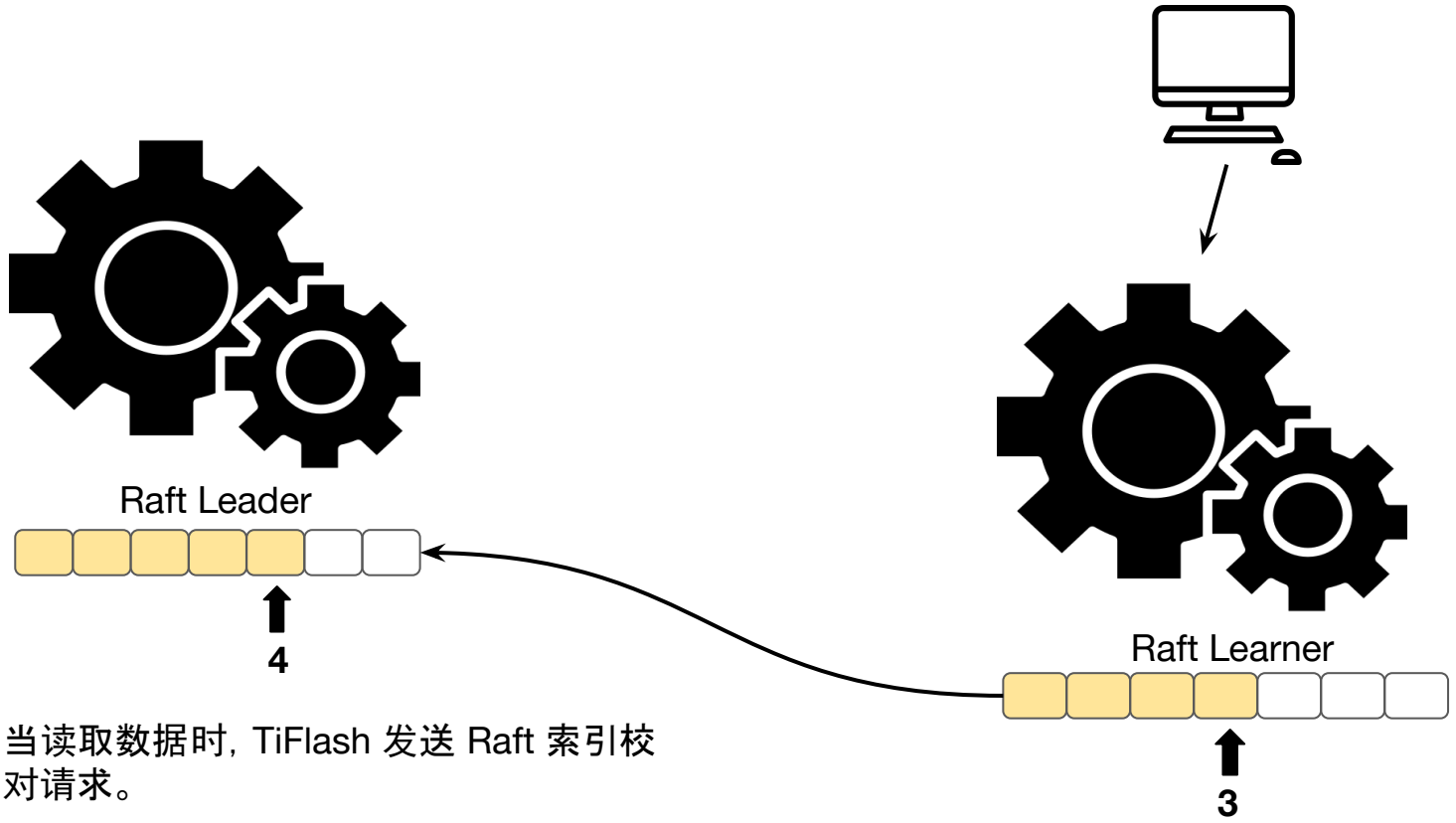
Raft Learner - Sync



TiFlash 以 Raft Learner 方式接入 Multi-Raft 组，使用异步方式传输数据，对 TiKV 产生非常小的负担。

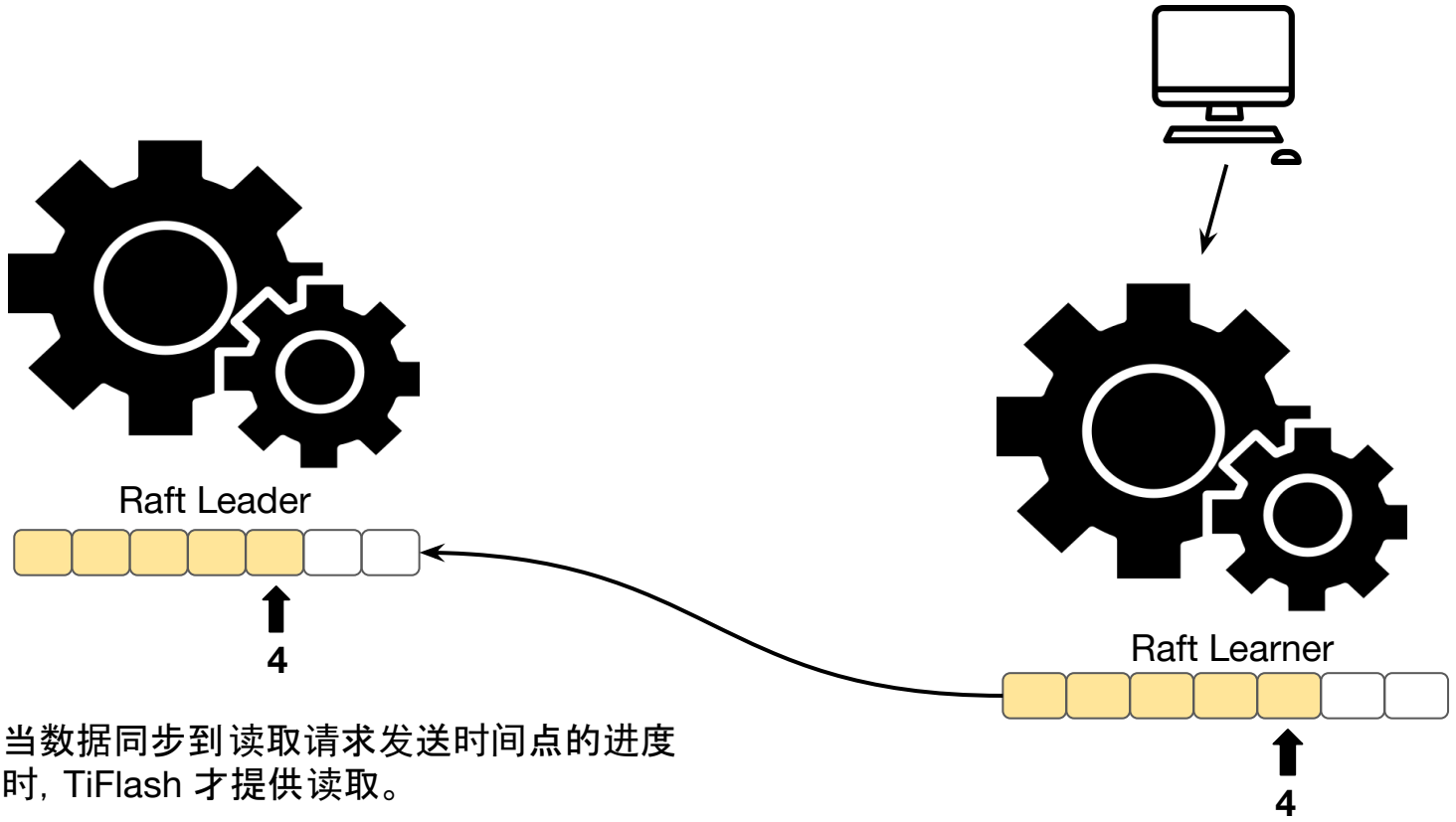
当数据同步到 TiFlash 时，会被从行格式拆解为列格式。

Raft Learner - Read



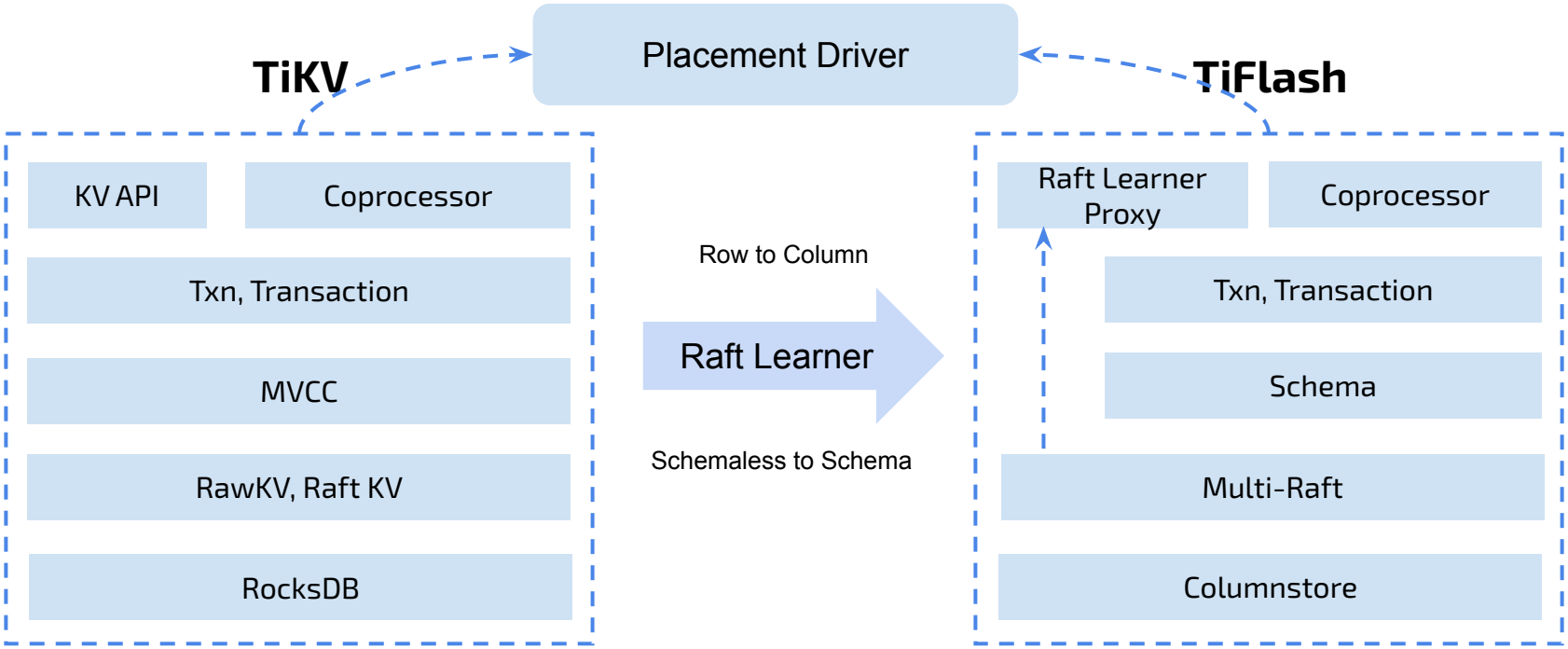
当读取数据时，TiFlash 发送 Raft 索引校对请求。

Raft Learner - Read



当数据同步到读取请求发送时间点的进度时，TiFlash 才提供读取。

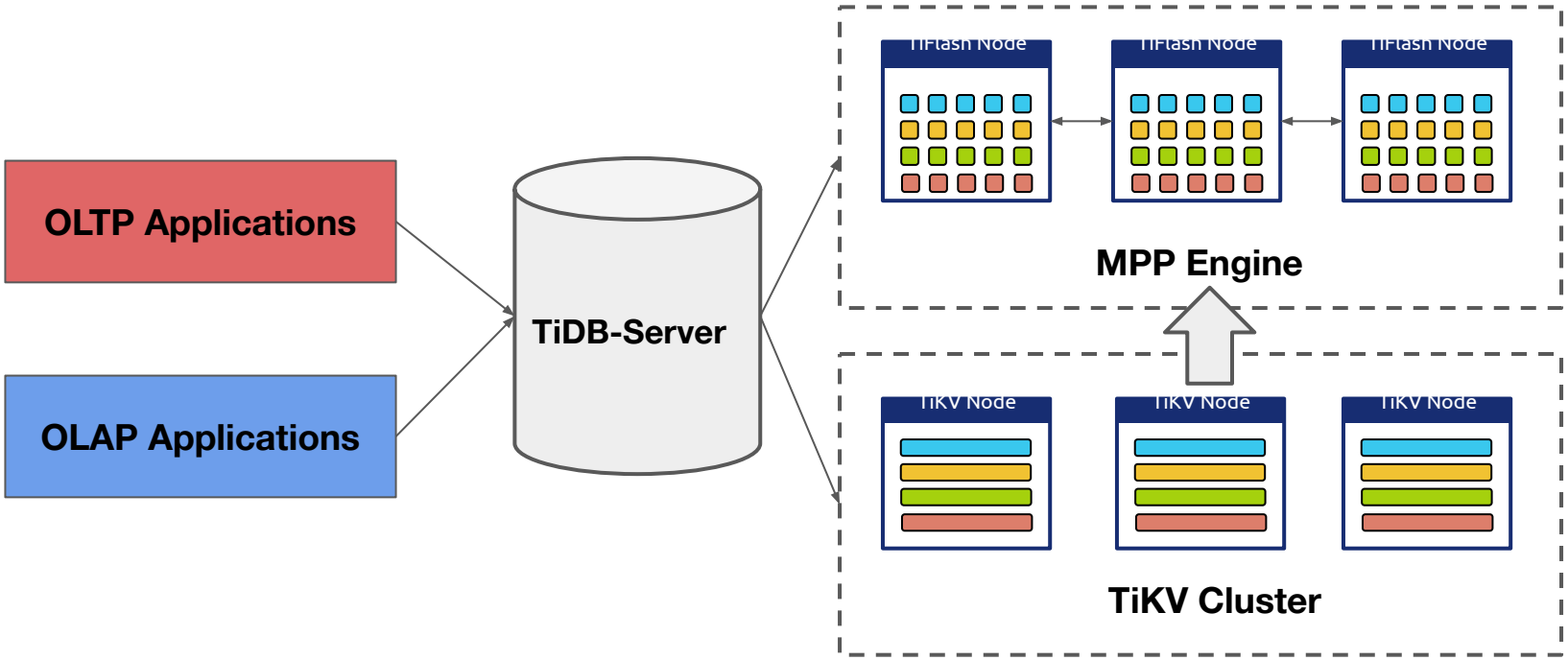
列存改造后的 TiFlash



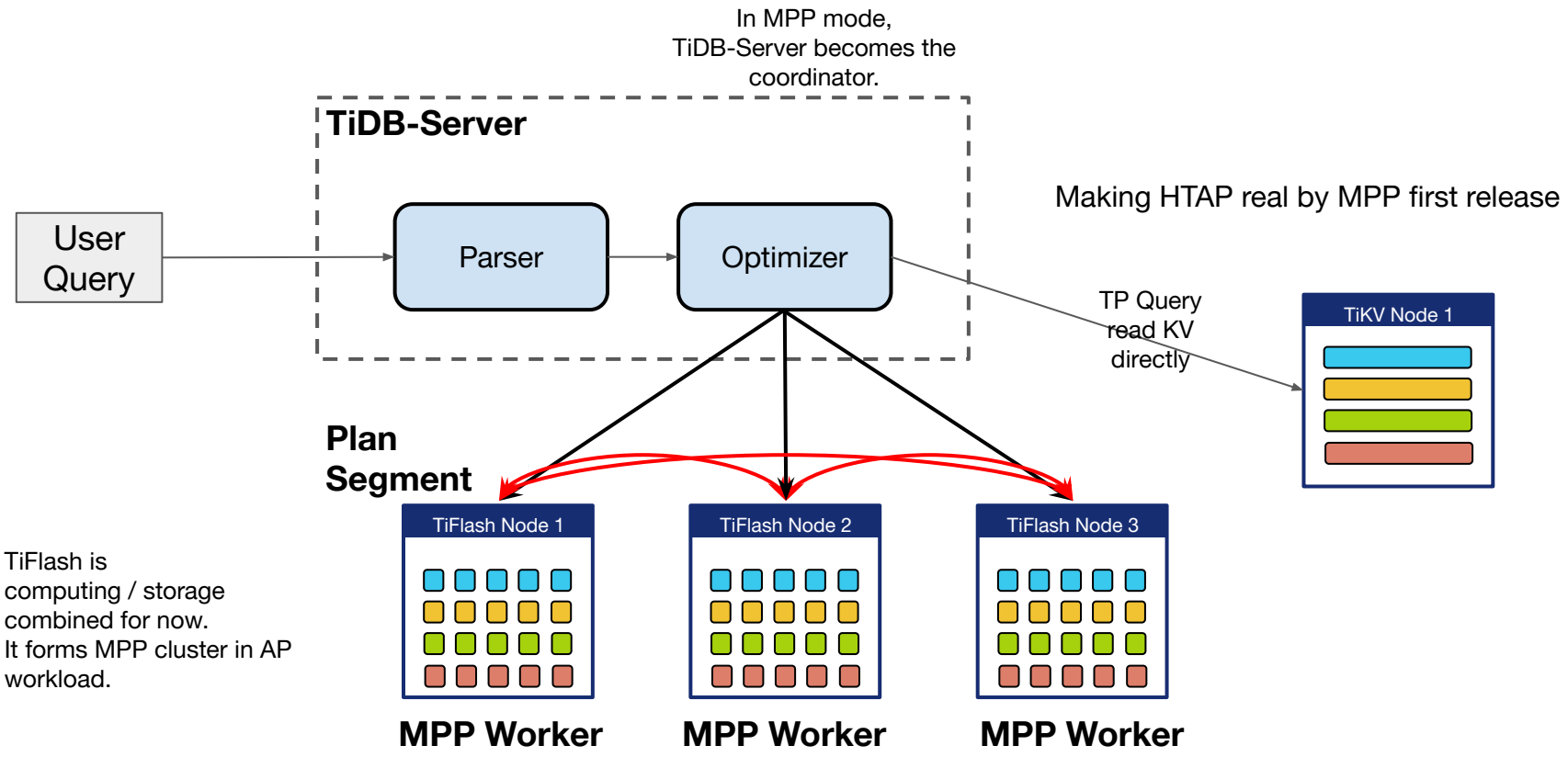
其他剩下的问题

- TiSpark 作为唯一分布式计算引擎
  - 缺少中规模快速查询的解决方案
  - 略重的模型(MR 模型)- 仍需要 MPP 引擎
- 写入需要通过 TiKV
  - 大批量写入速度吞吐不够
  - 副本必须先以行存方式写入再同步为列存

先做 MPP，解决中规模快速查询的问题

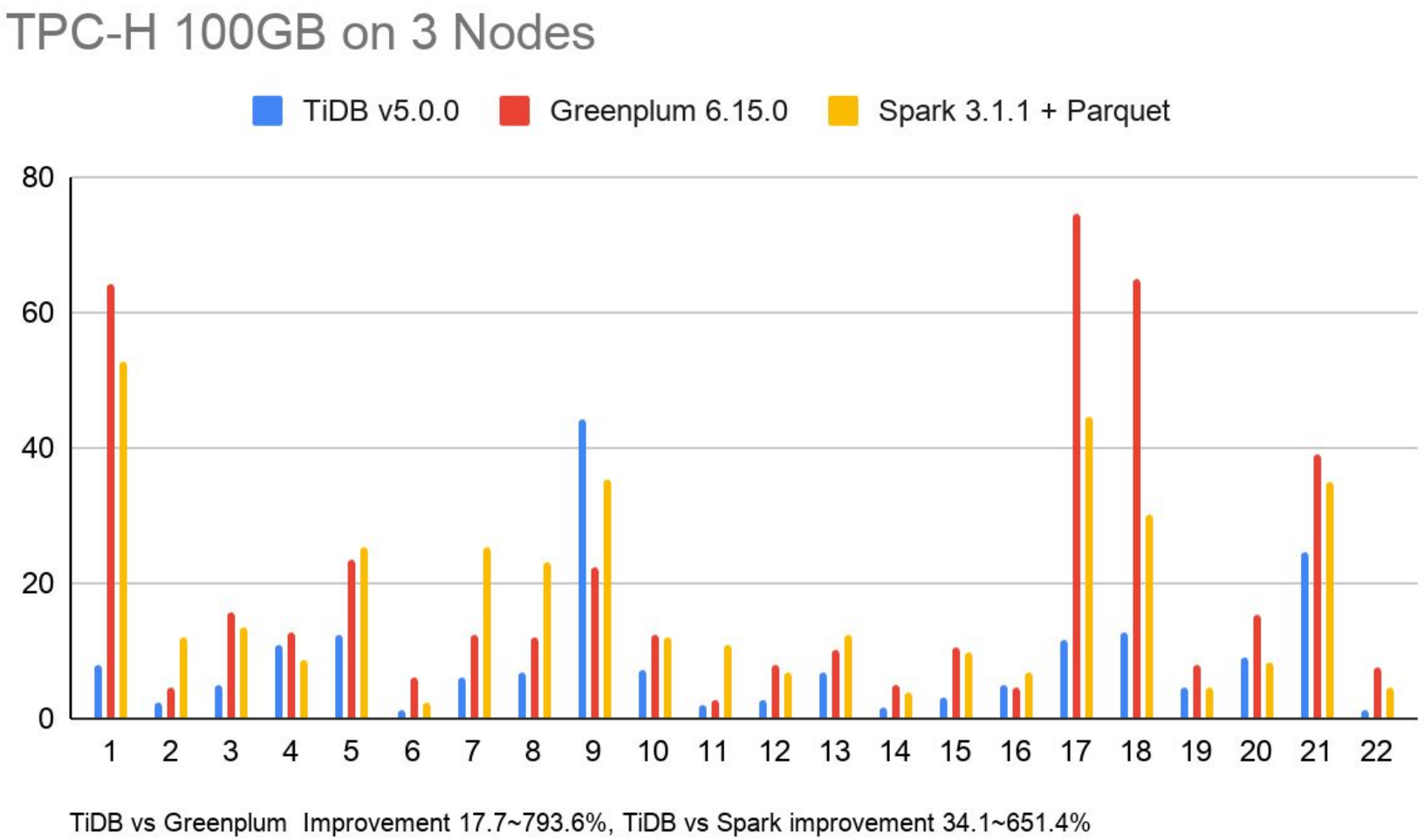


先做 MPP，解决中规模快速查询的问题



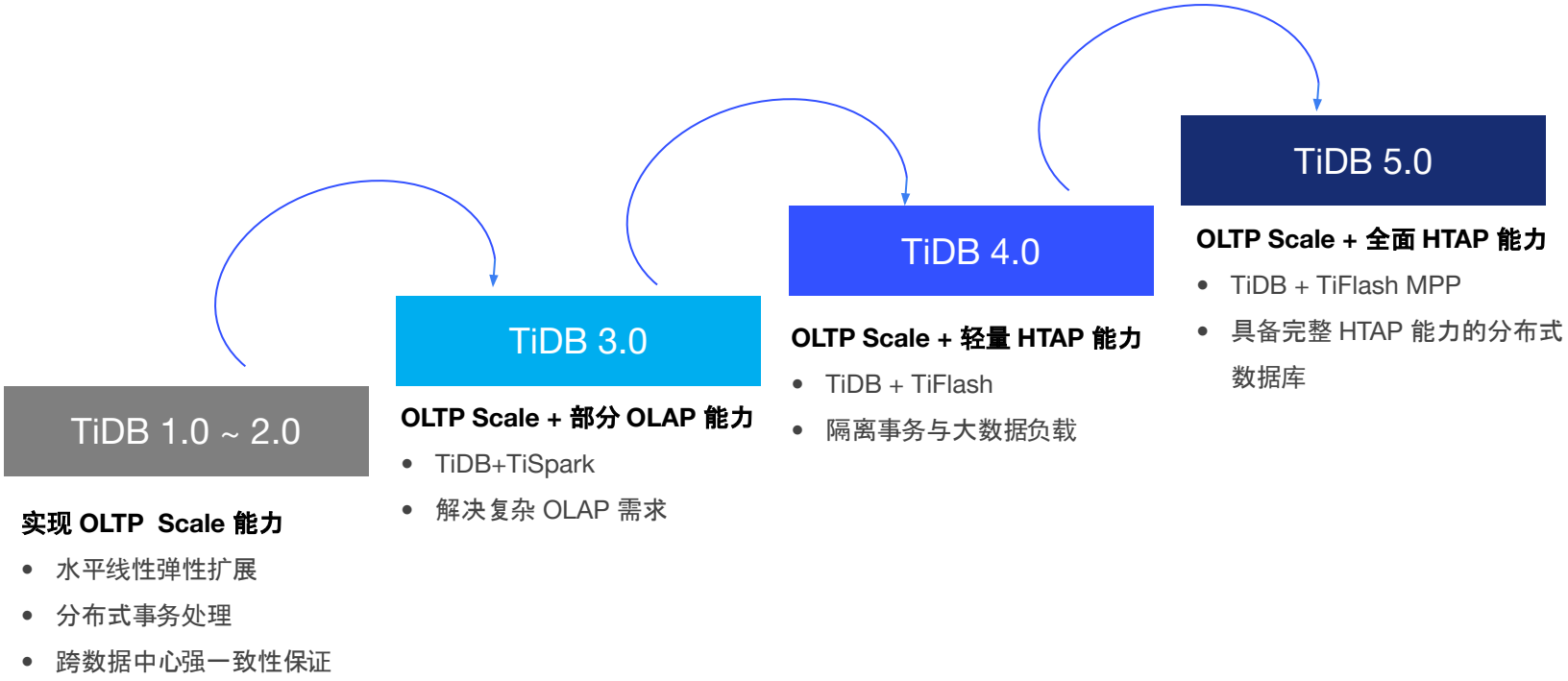
TiFlash is
computing / storage
combined for now.
It forms MPP cluster in AP
workload.

MPP 性能表现



TiDB vs Greenplum Improvement 17.7~793.6%, TiDB vs Spark improvement 34.1~651.4%

TiDB 架构演进总结


TiDB 1.0 ~ 2.0
实现 OLTP Scale 能力
• 水平线性弹性扩展
• 分布式事务处理
• 跨数据中心强一致性保证
TiDB 3.0
OLTP Scale + 部分 OLAP 能力
• TiDB+TiSpark
• 解决复杂 OLAP 需求
TiDB 4.0
OLTP Scale + 轻量 HTAP 能力
• TiDB + TiFlash
• 隔离事务与大数据负载
TiDB 5.0
OLTP Scale + 全面 HTAP 能力
• TiDB + TiFlash MPP
• 具备完整 HTAP 能力的分布式数据库

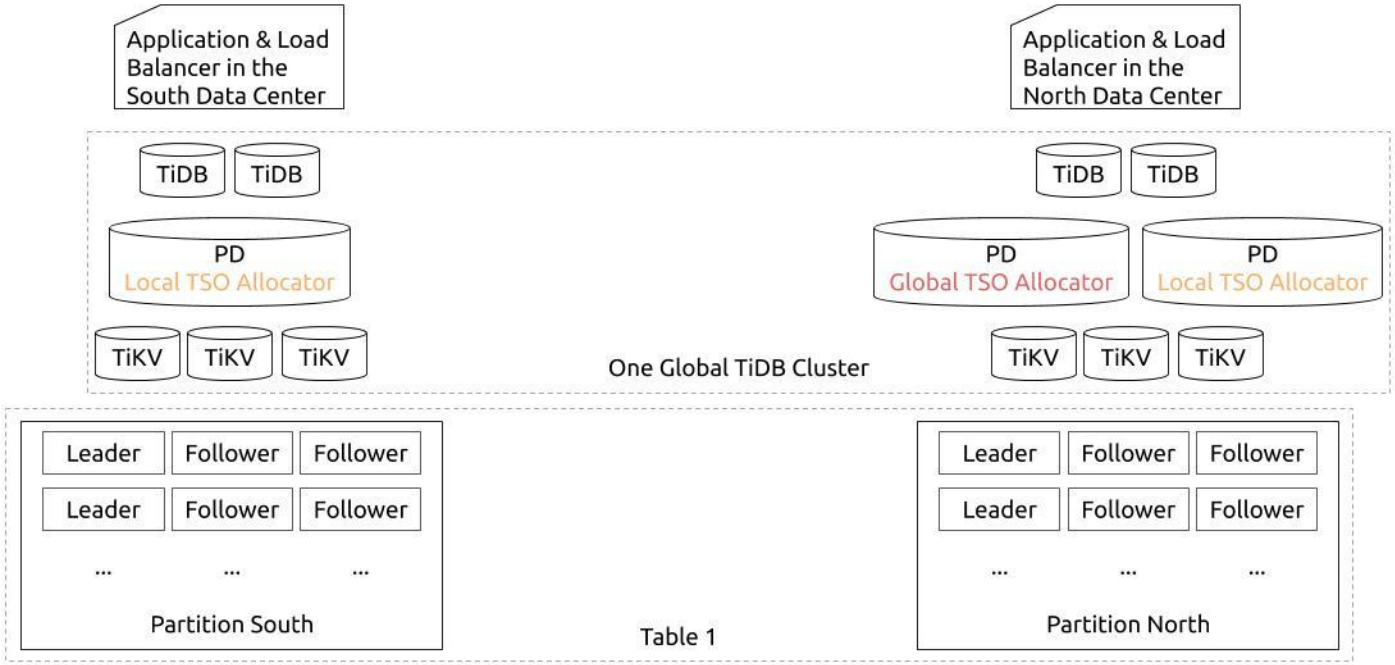
Future Work

Future Work

OLTP Scale

- Big Cluster
- Plan Stability
- Memory Management & Query Performance
- Geo-Partition

Future Work - Geo Partition



Future Work

OLTP Scale

- Big Cluster
- Plan Stability
- Memory Management & Query Performance
- Geo-Partition

Real-time HTAP

- Direct Write
- Bigdata Ecosystem Integration

麦思博(msup)有限公司是一家面向技 术型企业的培训咨询机构，携手2000余位中外客座导师，服务于技术团队的能力提升、软件工程效能和产品创新迭代，超过3000余家企业续约学习，是科技领域占有率第1的客座导师品牌，msup以整合全球领先经验实践为己任，为中国产业快速发展提供智库。

高可用架构公众号主要关注互 联网架构及高可用、可扩展及高性能领域的知识传播。订阅用户覆盖主流互联网及软件领域系统架构技术从业人员。

高可用架构系列社群是一个社区 组织，其精神是“分享+交流”，提倡社区的人人参与，同 时从社区获得高质量的内容。