TDSQL升级版引擎架构和关键技术介绍

韩硕 / shuohan
腾讯云数据库高级工程师

大纲

TDSQL升级版引擎架构

- 计算模块 SQLEngine
- 存储模块 TDStore
- 管控模块 TDMetaCluster

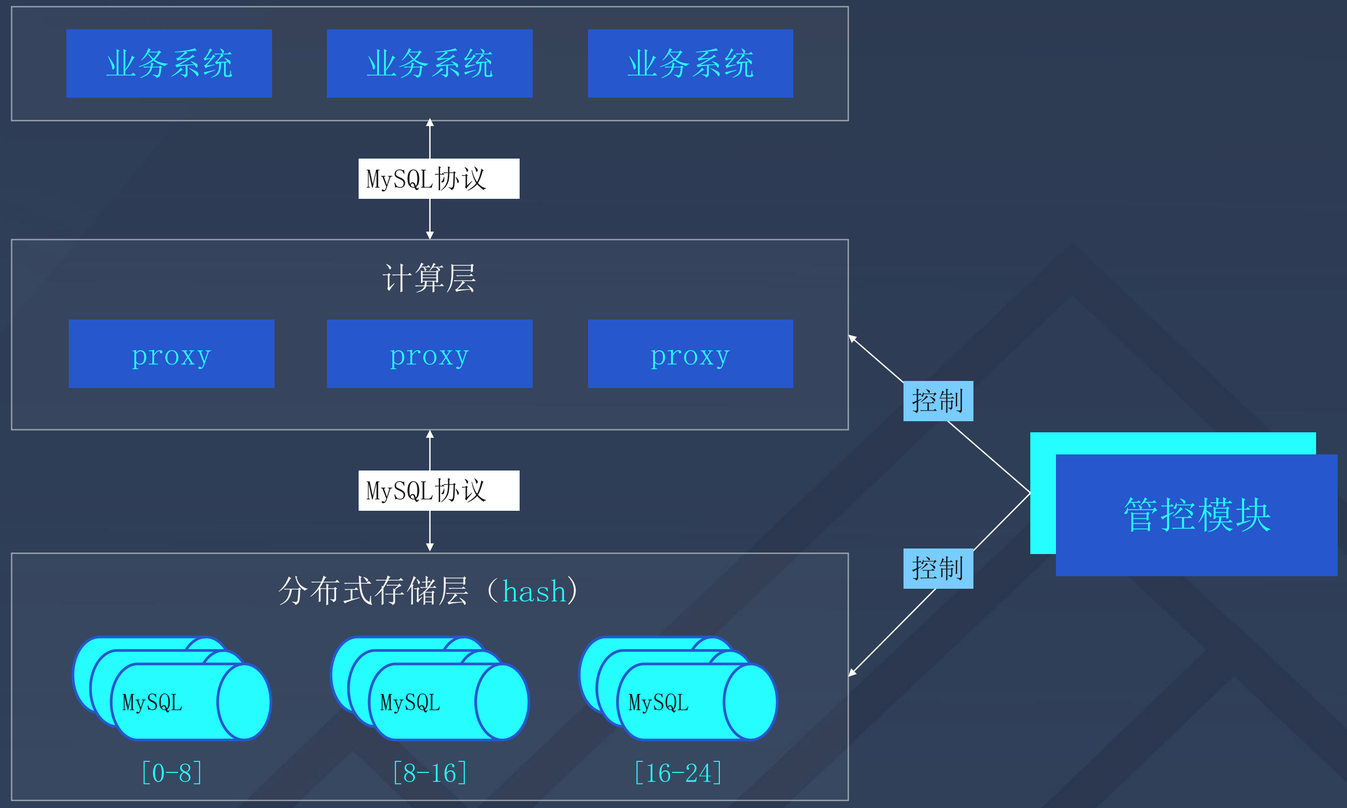
关键技术介绍

- 分布式事务
- 无感知扩缩容
- 数据存储与迁移

为什么要做升级版引擎

- TDSQL是腾讯自研的高性能、高可用的企业级分布式数据库解决方案，国内TOP20的银行近半在使用TDSQL

- 业务敏态发展对底层基础技术提出了具备敏态能力的要求：
  - 兼容性：建表需要指定shardkey
  - 运维：存储层扩容，需要DBA发起，部分事务会中断
  - 模式变更：online DDL 依赖pt等工具



TDSQL升级版引擎架构

- 目标：业务像使用单机数据库一样使用分布式数据库
- 功能特性：MySQL完全兼容 + 全局一致性 + 无感知扩缩容+ 在线表模式变更

TDSQL升级版引擎技术亮点

MySQL兼容、分布式、低成本

- MySQL兼容：兼容原生 MySQL 语法，业务层无入侵；
- 分布式：数据以 Key Range 打散和路由，业务层无须手动分库分表；
- 低成本：存储层采用 LSM-tree 结构，具有更好压缩比，适合大规模数据量业务；

高可扩展：计算/存储资源弹性扩缩容

- 计算层：多主模式，每个 SQLEngine 均可读写；无状态化设计，可根据业务流量随时灵活添加或移除任意数量的计算节点；
- 存储层：可根据业务数据存储量需求，添加或移除 TDStore 节点，通过数据自动迁移，实现容量弹性伸缩，业务层无感知；

全局读一致性

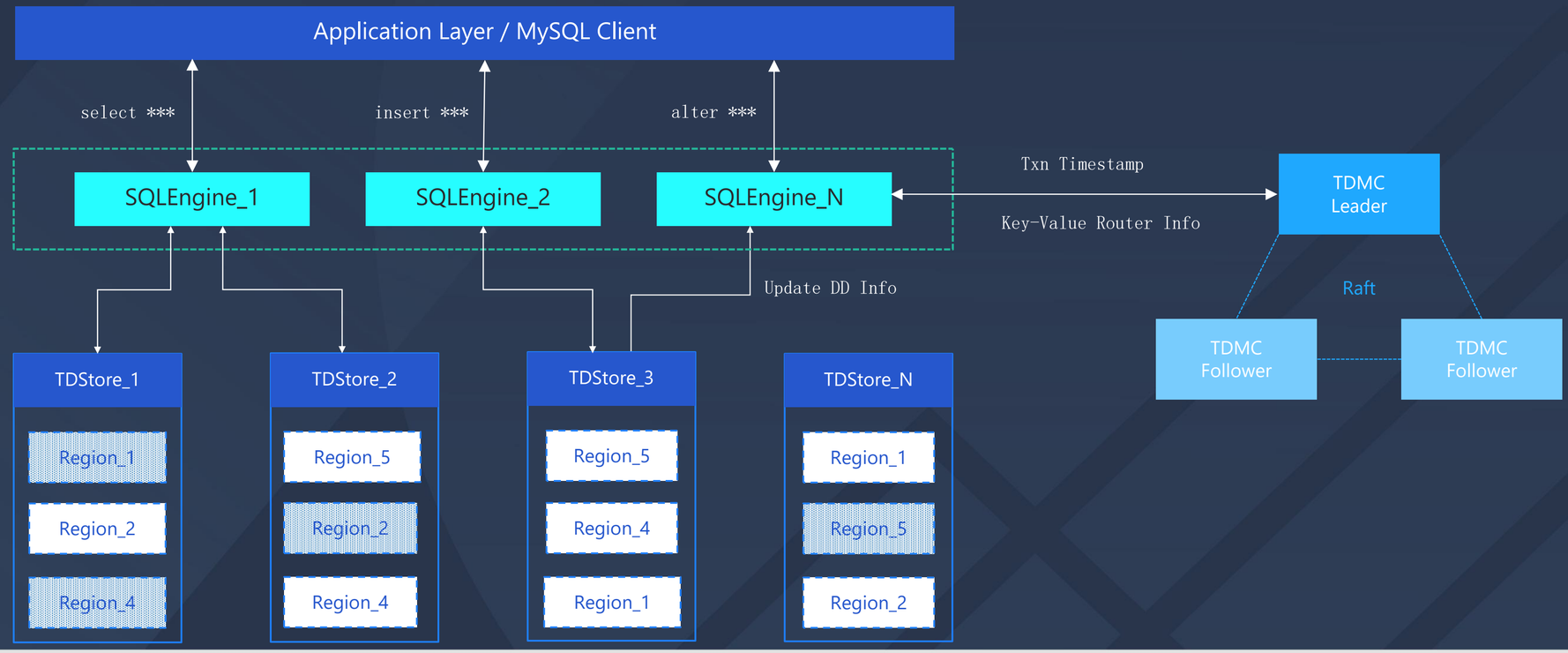
- TDMetaCluster 统一分配全局唯一递增事务时间戳
- TDStore 层基于数据多版本和事务时间戳判定数据可见性

Online DDL

- 支持在线加减列操作
- 支持在线加减索引
- 支持大部分 DDL 操作以 Online 方式执行

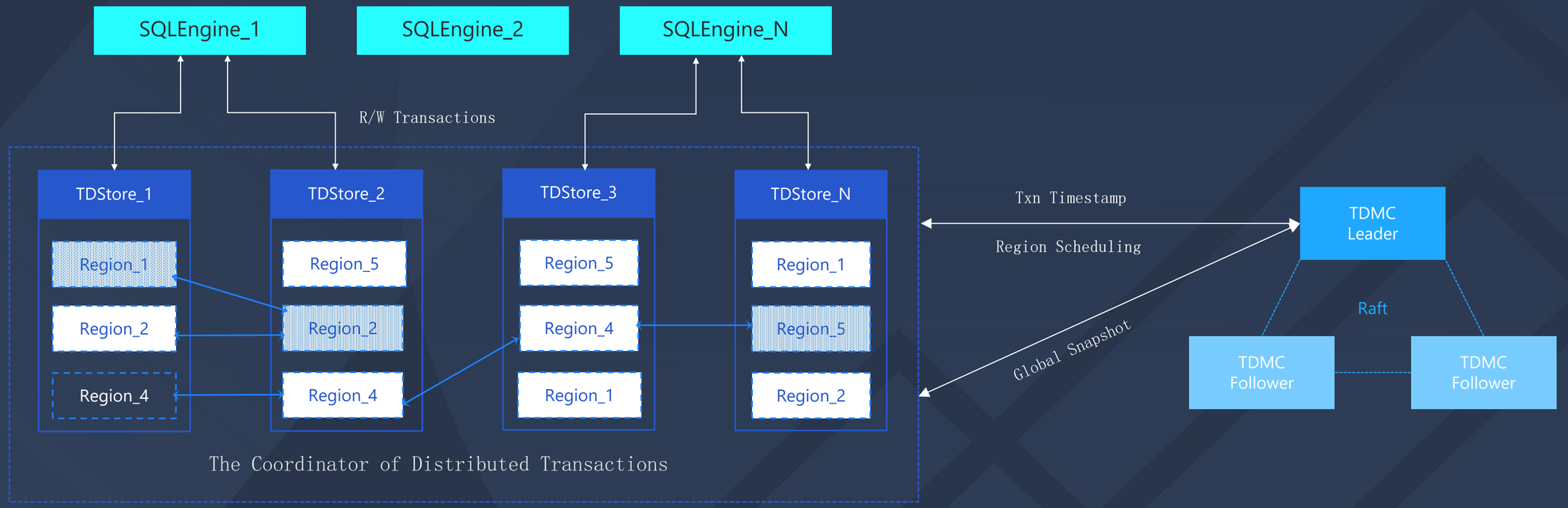
计算模块 SQLEngine

- 内核：完全兼容MySQL8.0；
- 架构：计算层为多主架构，每个 SQLEngine 节点均可读写，SQLEngine 之间通过一定方式刷新表结构变更等信息；
- 改造：无状态化设计，移除各种有状态化的数据信息（如锁、本地表结构，binlog 等）；多线程框架替换为协程框架；
- 交互：SQLEngine 从 MC 获取全局事务时间戳和路由信息，然后与 TDStore 进行事务的交互，向客户端返回结果；



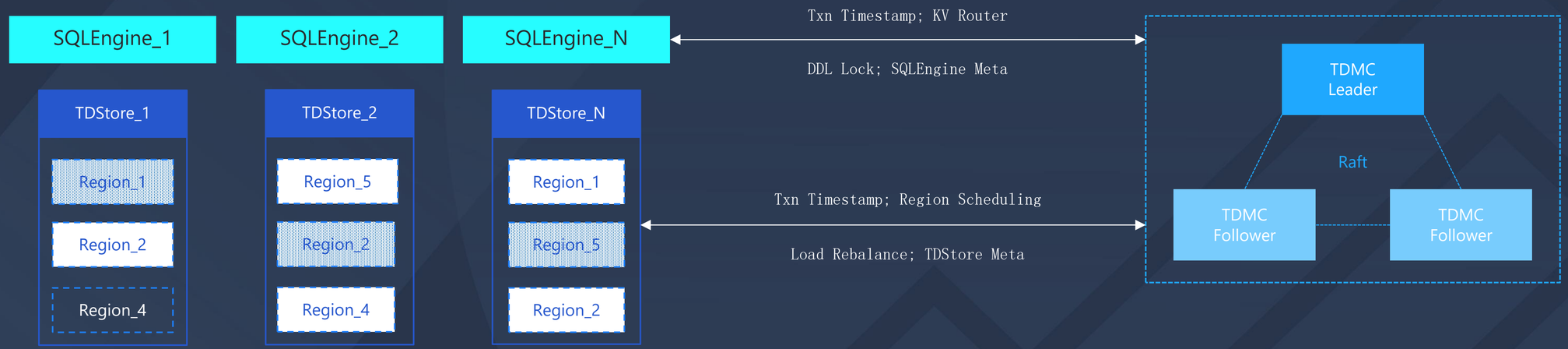
存储模块 TDStore

- 架构：基于 LSM-Tree 和 Multi-Raft 的分布式 KV 存储引擎；
- 数据：Region 是基于Raft同步的多副本的存储管理单元，数据根据Key范围分布在不同 Region上；Region TDMC 调度下可发生分裂、合并、迁移、切主等操作；
- 交互：TDStore 接收来自 SQLEngine 的事务请求，充当分布式事务的协调者角色，处理后返回结果；每个 Region 的主副本负责接收和处理读写请求；



管控模块 TDMetaCluster

- 架构：基于 Raft 的一主两备的元数据管理集群，由 Leader 提供服务；
- 数据：1）分配全局唯一且递增的事务ID；2）管理 TDStore 和 SQLEngine 元数据；3）管理 Region 数据路由信息； 4）全局 MDL 锁管理；
- 管控：1）调度 Region 的分裂、合并、迁移、切主 ；2）存储层的扩缩容调度； 3）存储层的负载均衡调度； 4）各维度的异常事件告警；



大纲

TDSQL升级版引擎架构

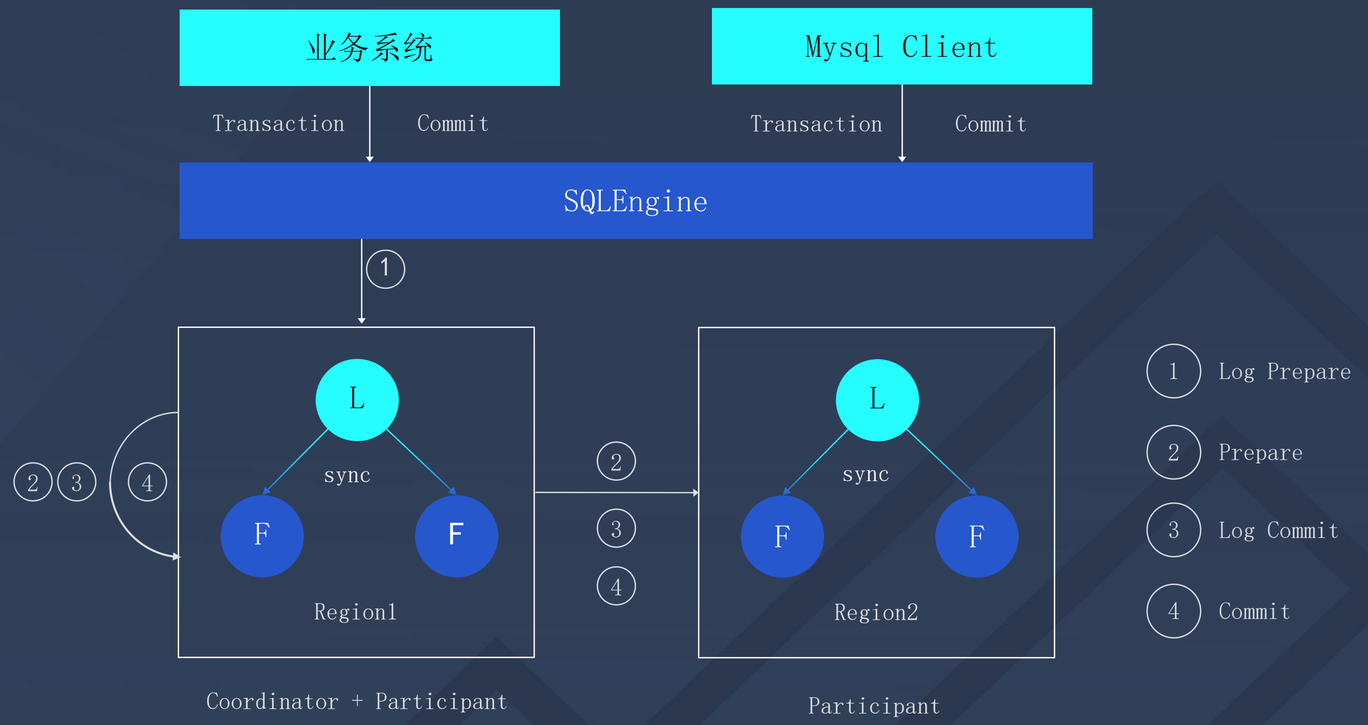
- 计算模块 SQLEngine
- 存储模块 TDStore
- 管控模块 TDMetaCluster

关键技术介绍

- 分布式事务
- 无感知扩缩容
- 数据存储与迁移

TDSQL升级版 – 分布式事务

- 协调者下沉，且不记录日志，降低延迟
- 故障恢复的时采用参与者协商的方法，协调者需要收集一次所有参与者状态



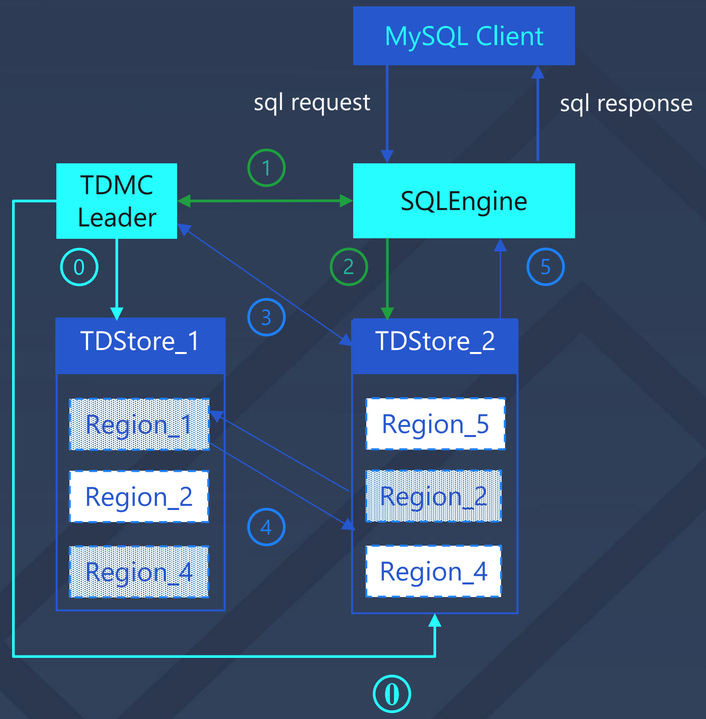
TDSQL升级版– 分布式事务

- Step 0：TDMC周期性向所有TDStore下发全局最小快照点，确保读操作的全局一致性；
- Step 1：SQLEngine接收到SQL语句，开启事务，从TDMC 取回事务开启时间戳 begin_ts；在自身缓存中查询SQL语句对应的 Region 路由信息;若无，则将对应的 key 范围查询请求发送给TDMC，由TDMC告知最新的路由信息；
- Step 2：SQLEngine将请求发送到对应的Region的Leader副本所在的TDStore 上；
- Step 3：Leader 副本作为协调者开启两阶段事务，从TMDC 取回prepare_ts；
- Step 4：所有参与者Prepare就绪后，从TDMC取回commit_ts，提交事务；若过程中，参与事务的其他 Region 发生切主， 则从TDMC 查询最新 Leader 信息；
- Step 5：提交事务，成功后，将结果返回至 SQLEngine，继而将结果返回到客户端；

SQLEngine：根据数据路由发起事务

TDMC：提供事务时间戳

TDStore：事务两阶段提交协调/参与者

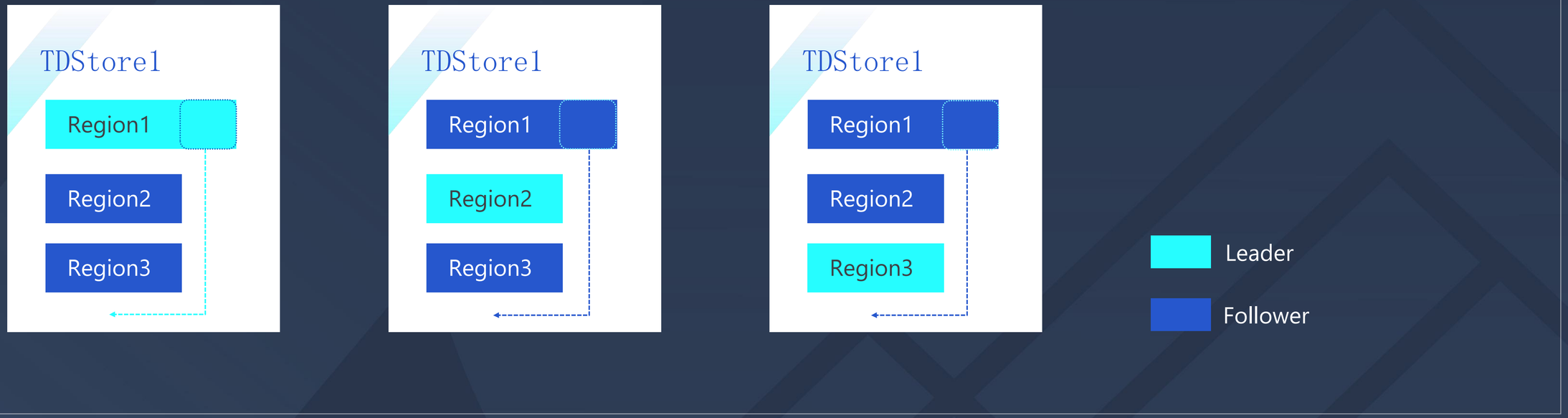


TDSQL升级版– 无感知扩缩容

- TDStore中的数据分段管理在Region中，数据调度是通过Region调度来做到的
- Region的调度：分裂、迁移、切主

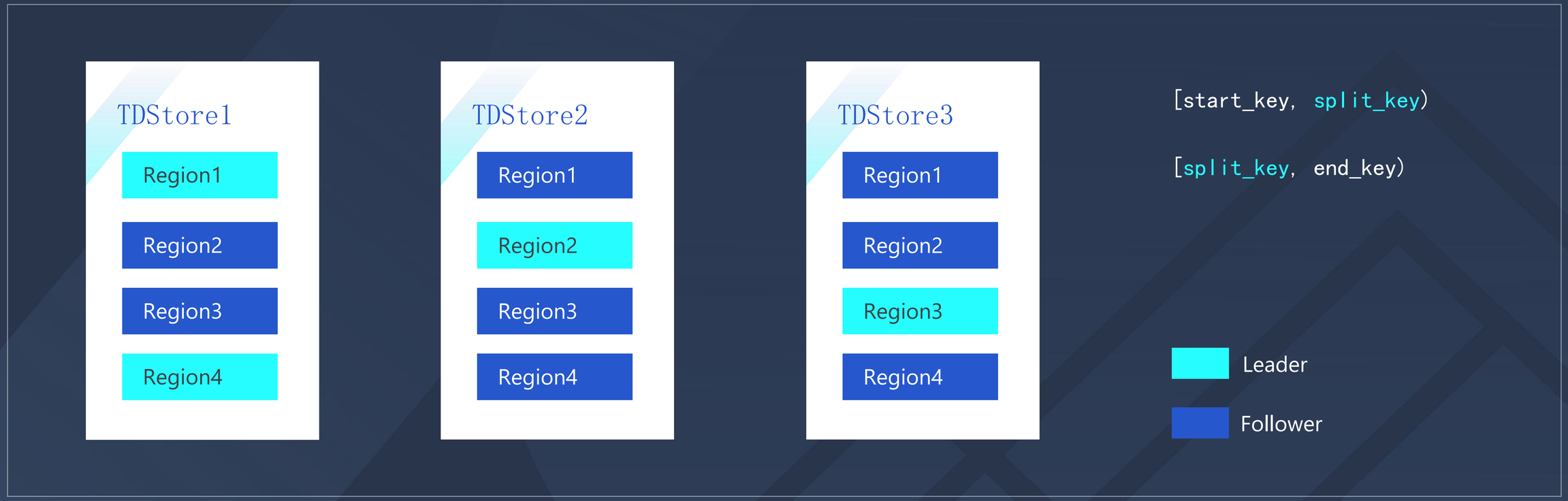
TDSQL升级版– 无感知扩缩容

- Region的调度：分裂、迁移、切主
  - 2PC流程，MC作为协调者，region的所有副本作为参与者，保证全员步调一致，避免部分成功部分失败造成的不一致



TDSQL升级版– 无感知扩缩容

- Region的调度：分裂、迁移、切主
  - 分裂点的选取：容量估算函数

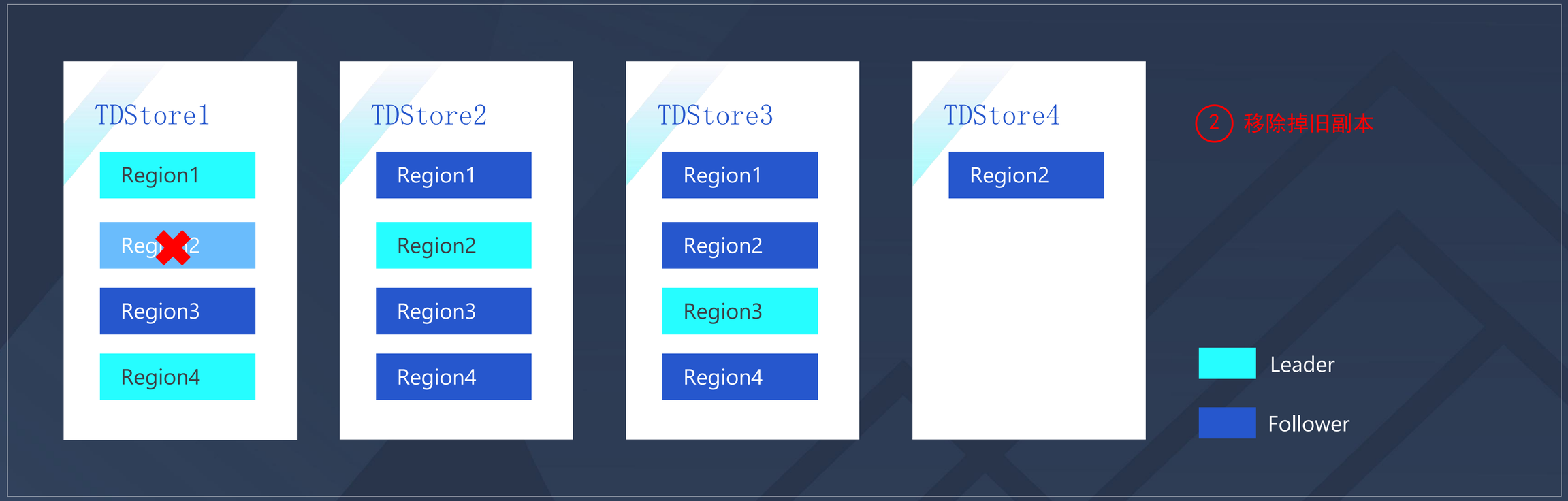


TDSQL升级版– 无感知扩缩容

- Region的调度：分裂、迁移、切主
  - 与Raft相结合，增减副本的方式来做到迁移

TDSQL升级版– 无感知扩缩容

- Region的调度：分裂、迁移、切主
  - 与Raft相结合，增减副本的方式来做到迁移

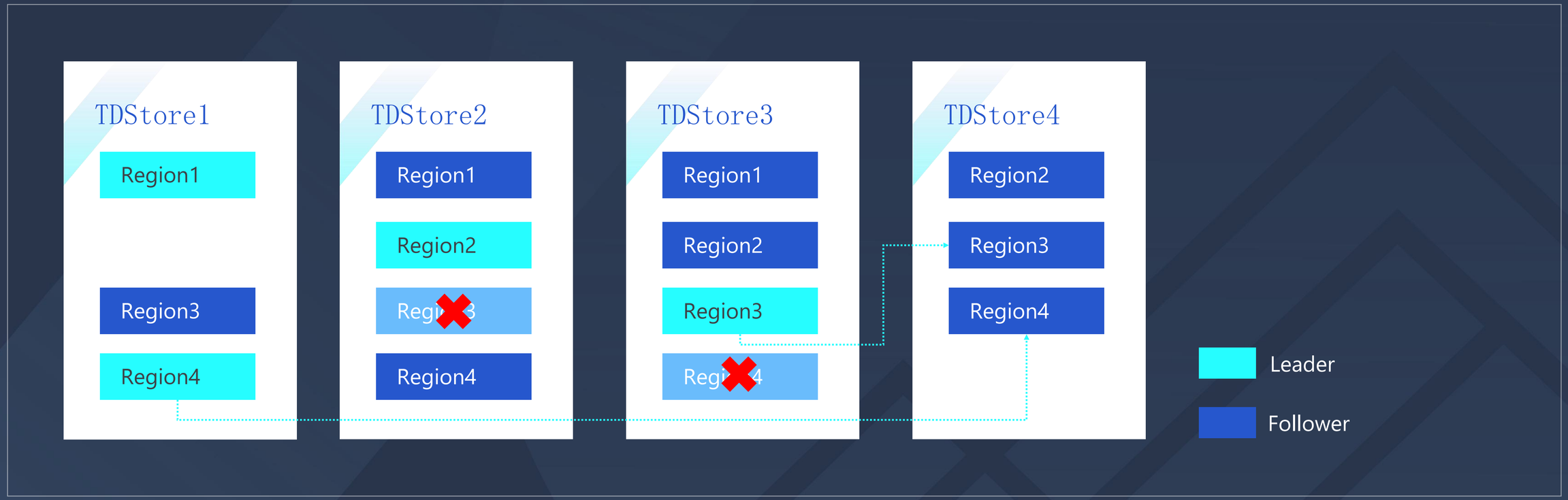


TDSQL升级版– 无感知扩缩容

- Region的调度：分裂、迁移、切主
  - 需要拷贝传输Region的持久化kv数据 —— 性能瓶颈点
  - 对事务透明，不阻塞事务

TDSQL升级版– 无感知扩缩容

- Region的调度：分裂、迁移、切主
  - 需要拷贝传输Region的持久化kv数据 —— 性能瓶颈点
  - 对事务透明，不阻塞事务



TDSQL升级版– 无感知扩缩容

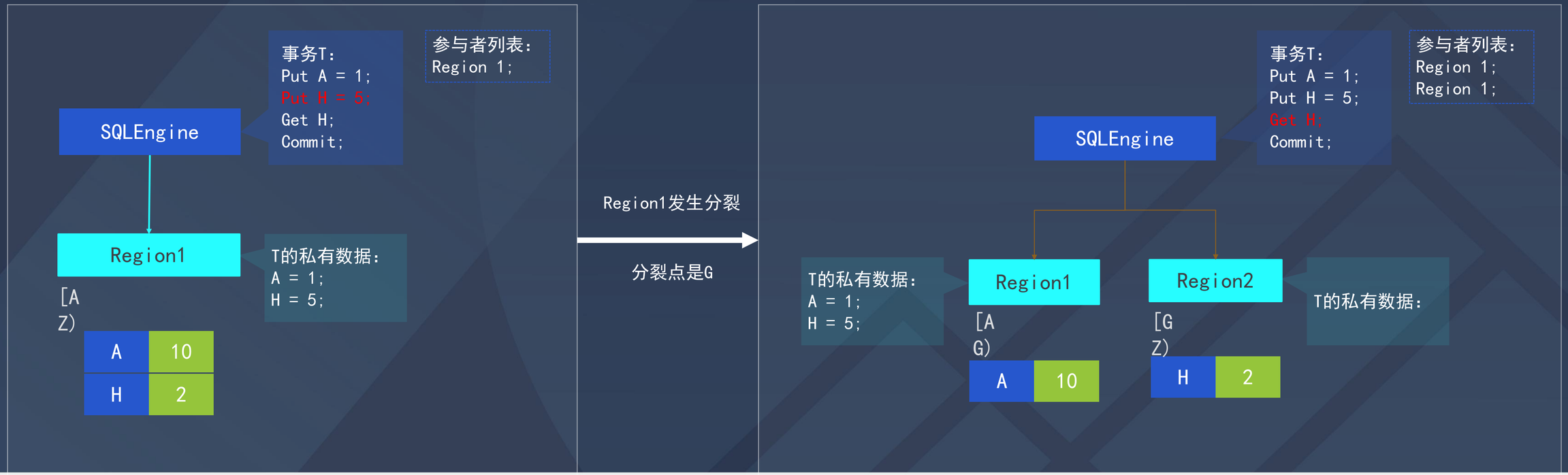
- Region的调度：分裂、迁移、切主
  - 均衡Region Leader分布，调整热点，均衡负载

TDSQL升级版– 无感知扩缩容

- Region的调度：分裂、迁移、切主
  - 主动切主不杀事务：已进入2PC阶段的事务，由新主继续推进；未进入2PC的事务，切主前需要将事务数据传输到新主上

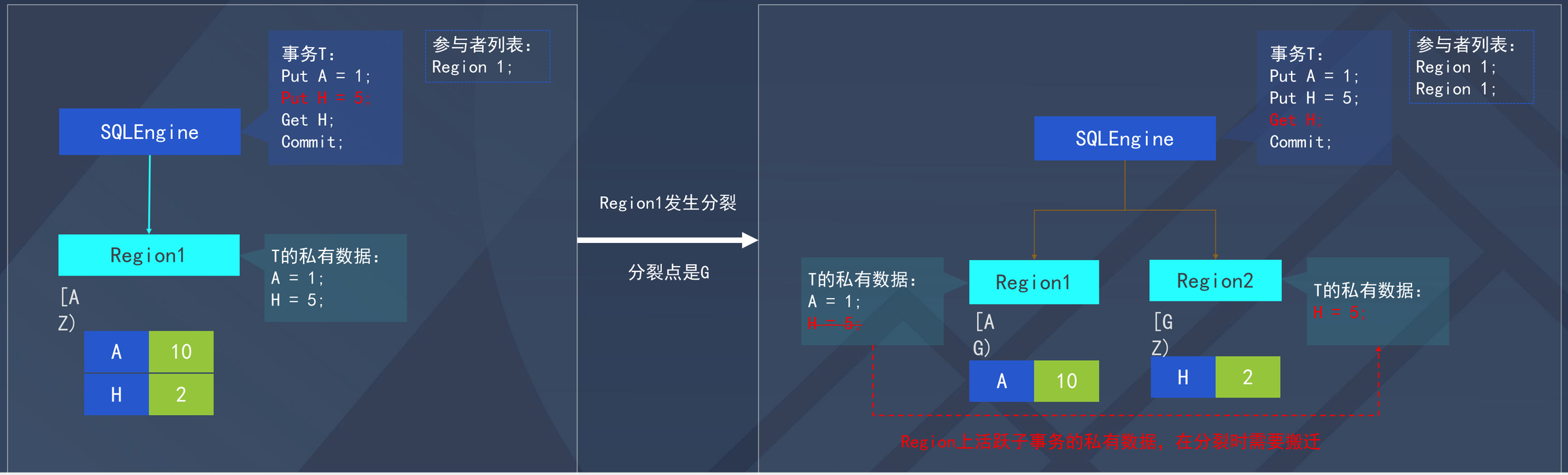
TDSQL升级版– 无感知扩缩容

- Region调度与事务并发：
  - 热点调度和自动伸缩的前提：业务层不能感知到服务中断
  - 迁移是通过Raft增减副本的方式进行的，与提供服务的Leader无直接关系
  - 分裂和切主都是在Leader节点上发起执行的，与事务不可避免的存在并发
  - 事务的生命周期要跨越分裂和切主



TDSQL升级版– 无感知扩缩容

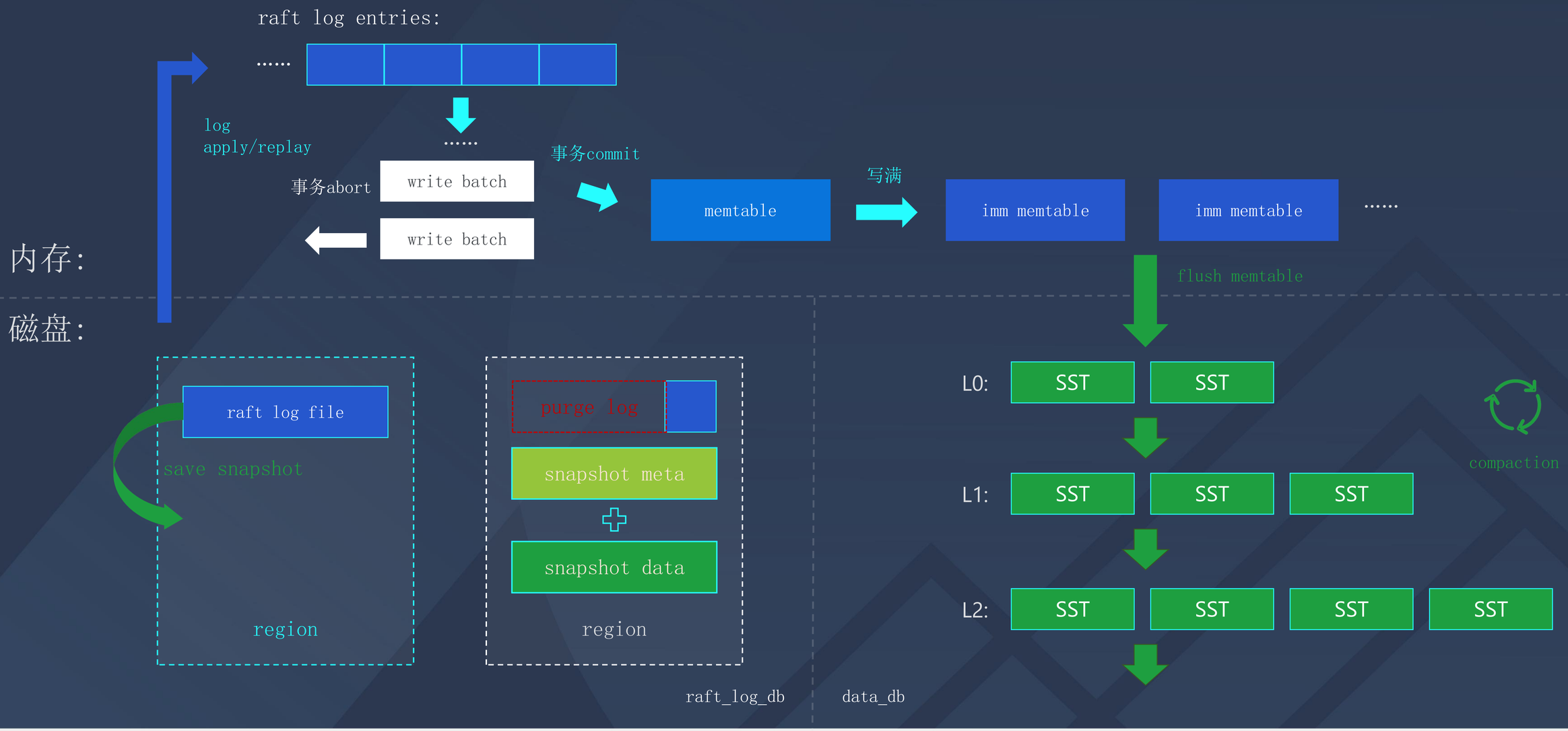
- Region调度与事务并发：
  - 热点调度和自动伸缩的前提：业务层不能感知到服务中断
  - 迁移是通过Raft增减副本的方式进行的，与提供服务的Leader无直接关系
  - 分裂和切主都是在Leader节点上发起执行的，与事务不可避免的存在并发
  - 事务的生命周期要跨越分裂和切主



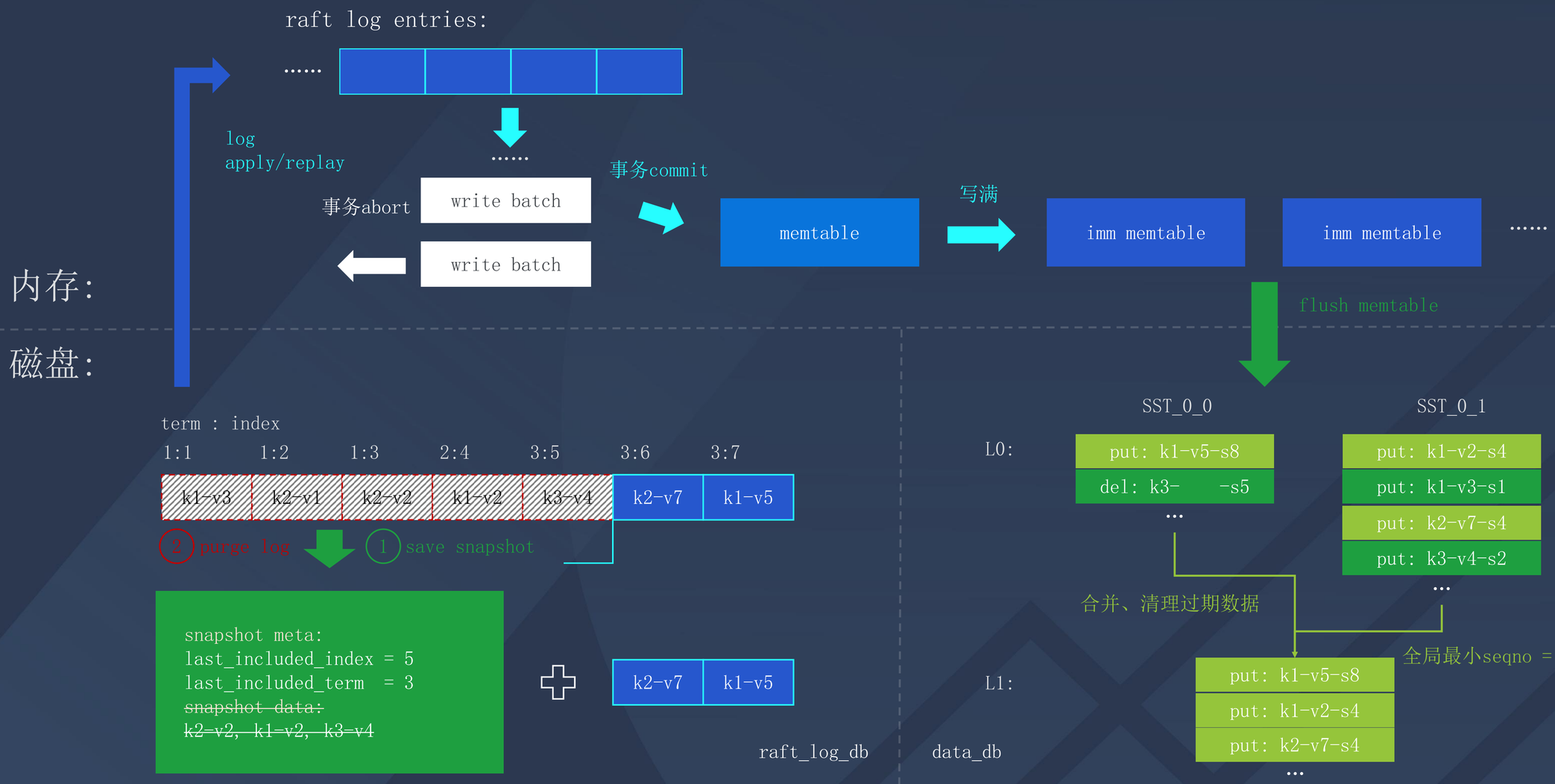
TDSQL升级版– 无感知扩缩容

- Region调度与事务并发：
  - 热点调度和自动伸缩的前提：业务层不能感知到服务中断
  - 迁移是通过Raft增减副本的方式进行的，与提供服务的Leader无直接关系
  - 分裂和切主都是在Leader节点上发起执行的，与事务不可避免的存在并发
  - 事务的生命周期要跨越分裂和切主

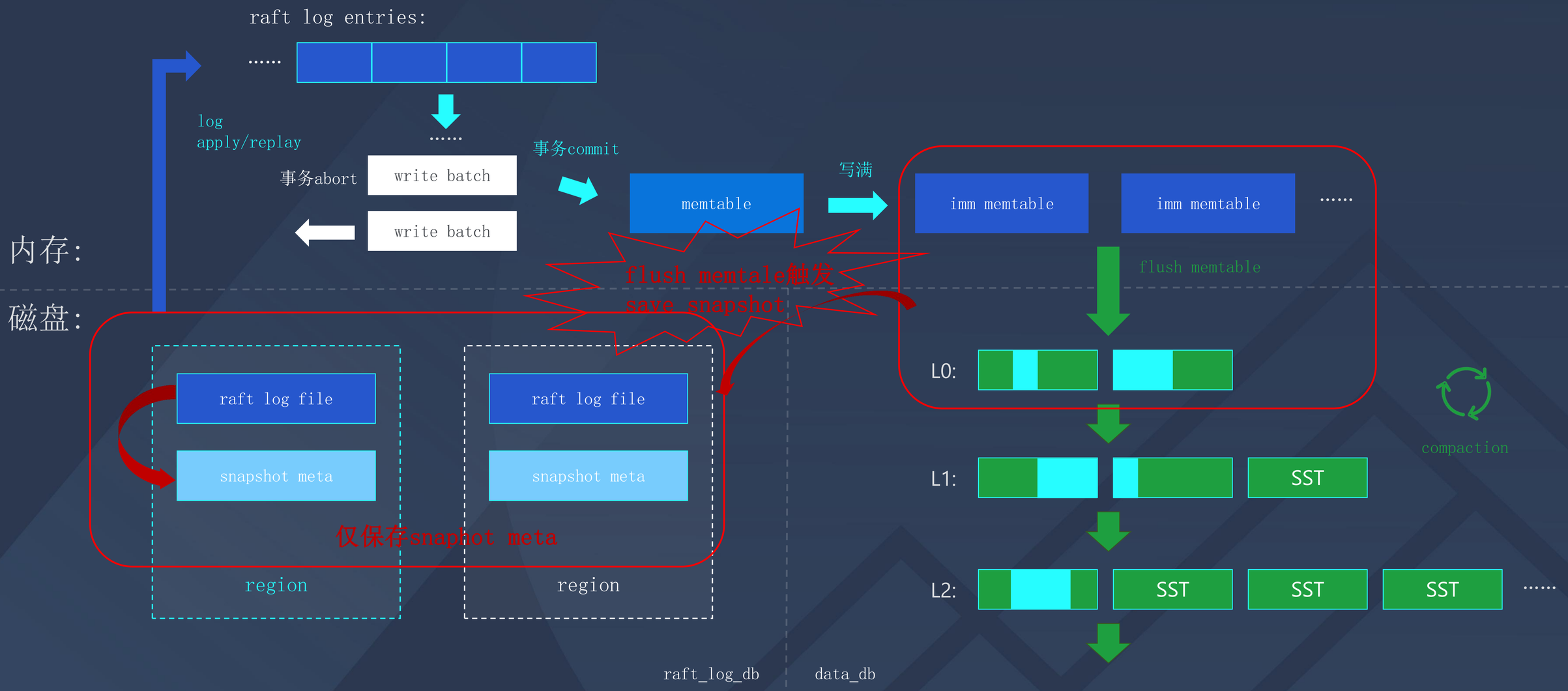
TDStore – 数据存储与迁移



TDStore – 数据存储与迁移

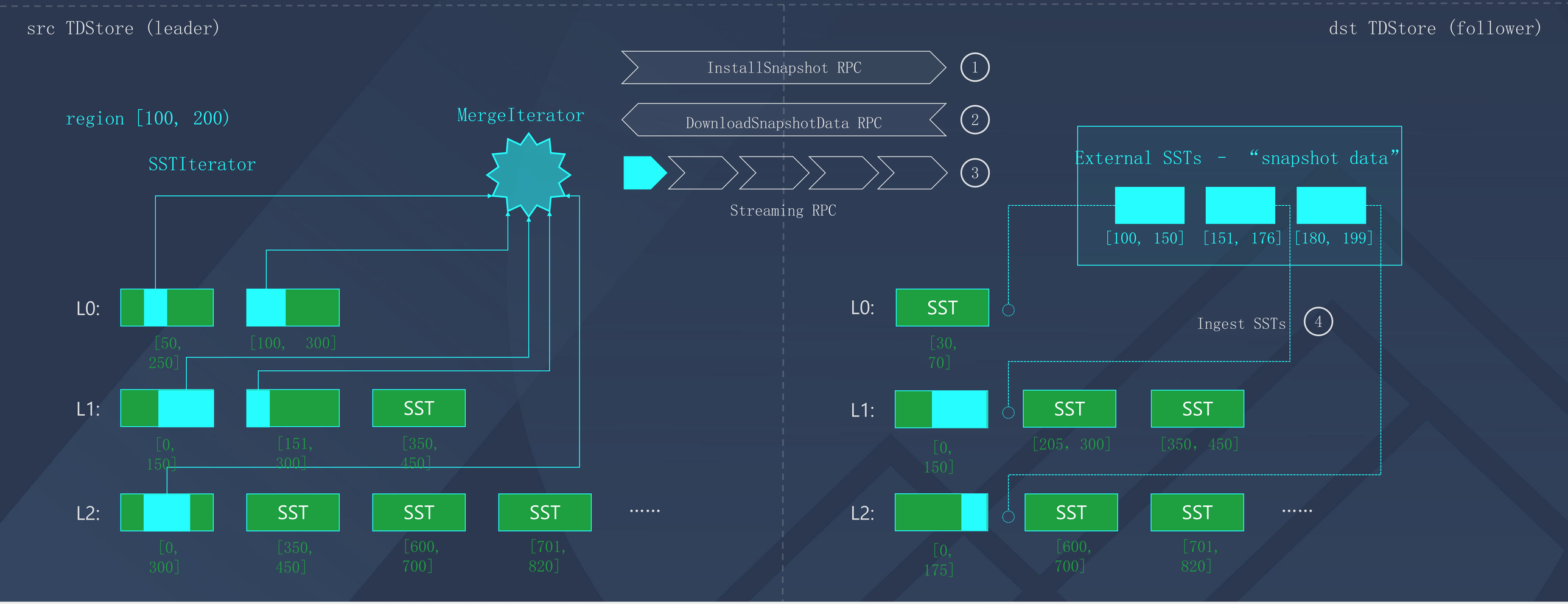


TDStore – 数据存储与迁移



TDStore – 数据存储与迁移

install snapshot（region迁移）流程



总结与未来规划

- TDSQL升级版特性：
  - MySQL完全兼容
  - 无限扩展的计算能力和存储容量
  - 无需指定shared key
  - 高可靠、高可用
  - 业务无感知的扩缩容

无限扩展性的单机数据库

THANKS

Global

Architect Summit