# 背景

RocksDB是FaceBook起初作为实验性质开发的，旨在充分实现快存上存储数据的服务能力。由Facebook的Dhruba Borthakur于2012年4月创建的LevelDB的分支，最初的目标是提高服务工作负载的性能，最大限度的发挥闪存和RAM的高度率读写性能。

Key和value是任意大小的字节流支持原子的读和写。除此外，RocksDB深度支持各种配置，可以在不同的生产环境（纯内存、Flash、hard disks or HDFS）中调优，RocksDB针对多核CPU、高效快速存储（SSD)、I/O bound workload做了优化，支持不同的数据压缩算法、和生产环境debug的完善工具。

RocksDB的主要设计点是在快存和高服务压力下性能表现优越，所以该db需要充分挖掘Flash和RAM的读写速率。 例如360开源的Pika， 是由360 DBA 和基础架构组联合开发的类 Redis 存储系统，完全支持 Redis 协议，用户不需要修改任何代码，就可以将服务迁移至 Pika，Pika底层存储引擎用的就是Rocksdb。

RocksDB需要支持高效的point lookup和range scan操作，需要支持配置各种参数在高压力的随机读、随机写或者二者流量都很大时性能调优，基于LSM树数据结构( log-structured merge-tree)，由C++编写并官方提供C、C++、Java(官方提供的称为RocksJava)三种语言的API，社区提供了不少第三方API，如python、go等。

尽管RocksDB不是一个SQL 数据库，但是有facebook有修改了代码的MyRocks存储引擎作为MySQL的存储引擎。和其他的NoSQL类似，RocksDB不提供关系型数据模型、不支持SQL查询，没有直接对辅助索引(secondary indexes)支持。

## 写多读少

## 原地写/追加写

在基于磁盘的 kv 型存储组件中，针对于写操作的实现方案包括原地写和追加写两种，其区别主要体现在更新数据的操作流程当中：

* 原地写：

倘若要针对一组 kv 数据执行更新操作，首先要找到 kv 老数据的所在位置，再在其基础之上执行进行更新。这个过程涉及到磁盘的随机IO，因此性能较差。

与之相对，在执行读操作时，可以根据 k 寻找到 kv 数据所在位置并直接拿到查询结果，因此读操作效率相对较高，且具有着不错的空间利用率。

* 追加写：

追加写类型的写操作中，无须区分本次写操作是插入还是更新，而是选择将 kv 对以追加的形式直接插入到文件的末尾位置。因此不涉及磁盘的随机 IO，只需要执行顺序 IO 操作，在写流程中的执行性能相较于原地写而言有较大的提升。

然而大家应该也注意到了，追加写策略在提升写操作效率的同时，所付出的代价是导致同一组 kv 对可能产生多份冗余数据，而除了最新记录外，此前的数据记录实际上都是无用的，因此会存在空间浪费的情况。

也正是因为这一原因，追加写策略下的读流程性能是比较差的。每次根据 k 查询数据时，都需要沿文件末尾向前反向遍历追溯，直到找到第一笔满足条件的 kv 对数据为止。在这种模式下，查询操作变成了线性时间复杂度，是无法接受的。

# 概述

参考：

<https://blog.csdn.net/weixin_43618070/article/details/102317769>

<https://www.cnblogs.com/xueqiuqiu/articles/9700198.html>

<https://www.jianshu.com/nb/48664459>

mongorocks引擎原理解析：<http://mysql.taobao.org/monthly/2018/04/02/>

WAL介绍：<http://mysql.taobao.org/monthly/2018/04/09/>

RocksDB原理：<https://zhuanlan.zhihu.com/p/616209332?utm_id=0>

## 简介

[RocksDB](https://github.com/facebook/rocksdb) 是由Facebook基于LevelDB开发的一款提供键值存储与读写功能的LSM-tree架构引擎（嵌入式存储引擎）。用户写入的键值对会先写入磁盘上的WAL (Write Ahead Log)，然后再写入内存中的跳表（SkipList，这部分结构又被称作MemTable）。LSM-tree引擎由于将用户的随机修改（插入）转化为了对WAL文件的顺序写，因此具有比B树类存储引擎更高的写吞吐。

内存中的数据达到一定阈值后，会刷到磁盘上生成SST文件（Sorted String Table），SST又分为多层（默认至多6层），每一层的数据达到一定阈值后会挑选一部分SST合并到下一层，每一层的数据是上一层的10倍（因此90%的数据存储在最后一层）。

Rocksdb目前已经运用在许多知名的项目中，例如TiKV，MyRocks，CrockRoach等。

说明：RocksDB作为高效发挥硬件存储性能的高性能嵌入式数据库，与其他数据库不同的是，不能提供网络服务，只负责存储服务。RocksDB有针对硬件的一些优化。

## 高性能/高可靠

RocksDB使用了许多技术来实现其高性能和可靠性，下面是一些主要的技术：

### LSM树

1、LSM树：LSM树（Log-Structured Merge Tree）是一种基于日志结构的数据结构，能够高效地存储和更新键值数据。它将数据分为多个层，每一层都是一个有序的键值存储文件，其中较旧的数据位于较低的层，较新的数据位于较高的层。当数据被写入时，它首先被写入到一个内存中的结构，称为内存表（MemTable），然后在后台异步地将内存表与磁盘上的某个层合并，最终生成新的文件。这种设计使得RocksDB能够高效地处理大量写入操作，并支持快速的范围查询。

### 压缩

2、压缩：RocksDB使用了多种压缩算法来压缩数据文件，减小了磁盘空间的占用，提高了存储效率。压缩算法包括LZ4、Snappy、Zlib等。

### 并发控制

3、并发控制：RocksDB使用多种技术来实现并发控制，以支持高并发读写操作。例如，它使用锁、读写锁、CAS等机制来保证多线程并发的正确性和一致性。

### 内存管理

4、内存管理：RocksDB使用了多种技术来管理内存，以保证高效的内存使用和低延迟的响应。例如，它使用了对象池、内存池等技术来减少内存分配和释放的开销，使用了缓存技术来缓存热点数据，使用了内存映射技术来快速加载数据文件等。

### 日志系统

5、日志系统：RocksDB使用了可插拔的日志系统，可以将日志输出到不同的目标，例如文件、控制台、网络等，以支持不同的日志需求。

### 文件格式

6、文件格式：RocksDB使用了一种自定义的文件格式，可以高效地存储键值数据，并支持快速的数据访问和查询。这种格式将数据划分为多个块，每个块包含多个键值对，每个块都有一个索引来支持快速的查找和范围查询。

说明：这里自定义的文件格式，包括内存memtable的分层和跳表数据结构设计，磁盘sstable的分层和数据结构设计。

## 功能特性

### ColumnFamily

参考：

ColumnFamily简介：<http://mysql.taobao.org/monthly/2018/06/09/>

RocksDB允许用户创建多个ColumnFamily ，这些ColumnFamily各自拥有独立的内存跳表以及SST文件，但是共享同一个WAL文件，这样的好处是可以根据应用特点为不同的ColumnFamily选择不同的配置，但是又没有增加对WAL的写次数。

写操作先写WAL，再写memtable，memtable达到一定阈值后切换为Immutable Memtable，只能读不能写。后台Flush线程负责按照时间顺序将Immu Memtable刷盘，生成level0层的有序文件（SST）。后台合并线程负责将上层的SST合并生成下层的SST。Manifest负责记录系统某个时刻SST文件的视图，Current文件记录当前最新的Manifest文件名。每个ColumnFamily有自己的Memtable，SST文件，所有ColumnFamily共享WAL、Current、Manifest文件，用户可以基于RocksDB构建自己的column families。很多应用程序把RocksDB当做库（libary），尽管他提供server或者CLI接口。

column family的具体使用场景，官方并没有给出非常好的说明，rocksdb支持跨column family的原子写操作，说明官方认为column family的数据之间还是有一定的关系的，而column family的数据文件是分割开的，包括sst文件和memtable都是不会共用的，所以官方还是希望不同column family的数据进行隔离。那么不同column family之间会怎样互相干涉呢？

Q：多column family的flush是怎么进行线程分布的？

A：可以共用一个线程池，所以写的多的会多占用线程。也可以分开配置线程池（指定不同的env），这样就会不互相干扰。

Q：多column family的内存是怎么分配的？

A：内存主要有两方面，第一方面是write buffer，第二方面是block cache。

write buffer是每个column family单独享有的，block cache可以配置成column family级别，也可以配置成整个db的column family共享一个block cache，也可以配置成多个db共享一个block cache。

参考：<https://www.cnblogs.com/xueqiuqiu/articles/10111070.html>

MANIFEST文件介绍：<http://mysql.taobao.org/monthly/2018/05/08/>

### Memtable

可插拔memtable，RocksDB的memtable的默认实现是一个skiplist。skiplist是一个有序集，当工作负载使用range-scans并且交织写入时，这是一个必要的结构。然而，一些应用程序不交织写入和扫描，而一些应用程序根本不执行范围扫描。对于这些应用程序，排序集可能无法提供最佳性能。因此，RocksDB支持可插拔的API，允许应用程序提供自己的memtable实现。

开发库提供了三个memtable：skiplist memtable，vector memtable和前缀散列（prefix-hash）memtable。Vector memtable适用于将数据批量加载到数据库中。每个写入在向量的末尾插入一个新元素; 当它是刷新memtable到存储的时候，向量中的元素被排序并写出到L0中的文件。前缀散列memtable允许对gets，puts和scans-within-a-key-prefix进行有效的处理。

### SSTFile/SSTTable

RocksDB在磁盘上的file结构sstfile由block作为基本单位组成，一个sstfile结构由多个data block和meta block组成，其中data block就是数据实体block，meta block为元数据block，其中data block就是数据实体block，meta block为元数据block。

sstfile组成的block有可能被压缩（compression），不同level也可能使用不同的compression方式。sstfile如果要遍历block，会逆序遍历，从footer开始。RocksDB是一个嵌入式的K-V（任意字节流）存储。所有的数据在引擎中是有序存储，可以支持Get(key)、Put（Key）、Delete（Key）和NewIterator()。RocksDB的基本组成是memtable、sstfile和logfile。memtable是一种内存数据结构，写请求会先将数据写到memtable中，然后可选地写入事务日志logfile。logfile是一个顺序写的文件。当内存表溢出的时候，数据会flush到sstfile中，然后这个memtable对应的logfile也会安全地被删除。sstfile中的数据也是有序存储以方便查找。

## 特点

<https://zhuanlan.zhihu.com/p/162052214>

### 优点

1. LevelDB是一个持久化存储的KV系统，和Redis这种内存型的KV系统不同，LevelDB不会像Redis一样狂吃内存，而是将大部分数据存储到磁盘上。
2. LevleDB在存储数据时，是根据记录的key值有序存储的，就是说相邻的key值在存储文件中是依次顺序存储的，而应用可以自定义key大小比较函数。
3. LevelDB支持数据快照（snapshot）功能，使得读取操作不受写操作影响，可以在读操作过程中始终看到一致的数据。
4. LevelDB还支持数据压缩等操作，这对于减小存储空间以及增快IO效率都有直接的帮助。

### 缺点

rocksdb是通过wal来保证数据的持久性的，当rocksdb出现问题当机后，可以通过重做wal来使rocksdb恢复到当机前的状态。但是这里存在两个问题：

1、当你为了读性能把memtable设置的足够大时，WAL也可能变得很大（Flush频率下降），此时如果发生当机，rocksdb需要足够长的时间来恢复。

2、如果机器硬盘出现损坏，wal被破坏，那么会出现数据损坏（这里没有double write机制？）。即使你做了Raid，也需要很长时间来恢复数据。

因此可用性是个大问题。

另外由于LSM架构，rocksdb的读性能存在问题。

怎么解决？

为了一致性，大部分的解决方案都是将一致性协议置于rocksdb之上，每份数据通过一致性协议提交到多个处于不同机器的rocksdb实例中，以保证数据的可靠性和服务的可用性。

## 对比/优化点

### RocksDB vs LevelDB

RocksDB与LevelDB对比：

1. 增加了column family，这样有利于多个不相关的数据集存储在同一个db中，因为不同column family的数据是存储在不同的sst和memtable中，所以一定程度上起到了隔离的作用。
2. 采用了多线程同时进行compaction的方法，优化了compact的速度。
3. 增加了merge operator，优化了modify的效率。
4. 将flush和compaction分开不同的线程池，能有效的加快flush，防止stall。
5. 增加了对write ahead log(WAL)的特殊管理机制，这样就能方便管理WAL文件，因为WAL是binlog文件。

### RocksDB vs MyRocks

# 架构

# 原理

<https://www.zhihu.com/question/270732348/answer/356254676>

## LSM设计思想

首先明确目标，我们需要解决的是写多读少的使用场景，在此之上提出的核心诉求包括：

* 由于写操作是最核心的使用场景，因此需要以提高写效率作为首要目标
* 读效率可以适当牺牲，但也需要保证在有限的容忍度范围以内

在此基础之上，我们很自然地想到使用追加写策略是较优的选择。于是我们需要考虑的问题就是，在传承顺序写所带来的优势的同时，如何尽可能去规避或者弱化其存在于读流程以及空间利用率方面的劣势。

### 追加写存在的问题

首先再梳理一下追加写（顺序写）策略所存在的问题：

* 数据冗余（空间浪费）

顺序写无视 k 此前的存在状况，简单粗暴地追加的方式完成数据的写入或者更新操作，因此不可避免地存在一组 kv 对对应多份冗余记录的情况。并且，除了最后一笔记录之外，此前多笔老数据都属于无用的冗余数据. 考虑一个最极端的场景，在对顺序写策略不执行任何优化改进的前提下，只需要一笔kv对数据，采用无限次写操作即可打满磁盘空间。

* 读性能低

顺序写模式下，执行一次查询操作需要反向追溯，直到找到第一笔满足条件的 kv 数据记录才能返回。因此，其最坏的时间复杂度是线性的 O(N)，显然无法满足使用诉求。

明确了存在的问题后，接下来我们就以这两项核心问题为主线展开推演和优化过程，尝试一步步构建出 lsm tree 的完整架构。

### 数据合并

首先考虑如何解决数据冗余的问题。

既然同一组 kv 对可能存在多份冗余数据，那我们自然能想到采用压缩合并的方式来解决这一问题。

于是，在读写主流程之外，我们可以异步启动一个负责压缩合并的线程，持续对重复的 kv 对数据进行合并，只保留最新记录，消除此前无用的老数据。

但是一旦这样做，新的问题也就随之产生了。这是因为针对同一文件的写入操作和压缩操作是需要互斥的，否则可能产生一系列严重的并发问题。

然而，我们一旦采取了互斥操作，那么在文件压缩合并期间，写入操作就会被阻塞，这样会严重影响到使用性能。

### 文件分块

在前面数据合并的基础之上，我们进一步采取的对策是对文件进行拆分。

倘若我们把一个大文件拆分为一系列的小文件（table），每次写操作只会追加到最新的new table中，而异步的压缩合并操作则只面向于老文件old table，这样两个流程之间就能够实现解耦，写操作不再有陷入阻塞的风险，同时压缩操作和写操作之间也能存在清晰明确的界限。

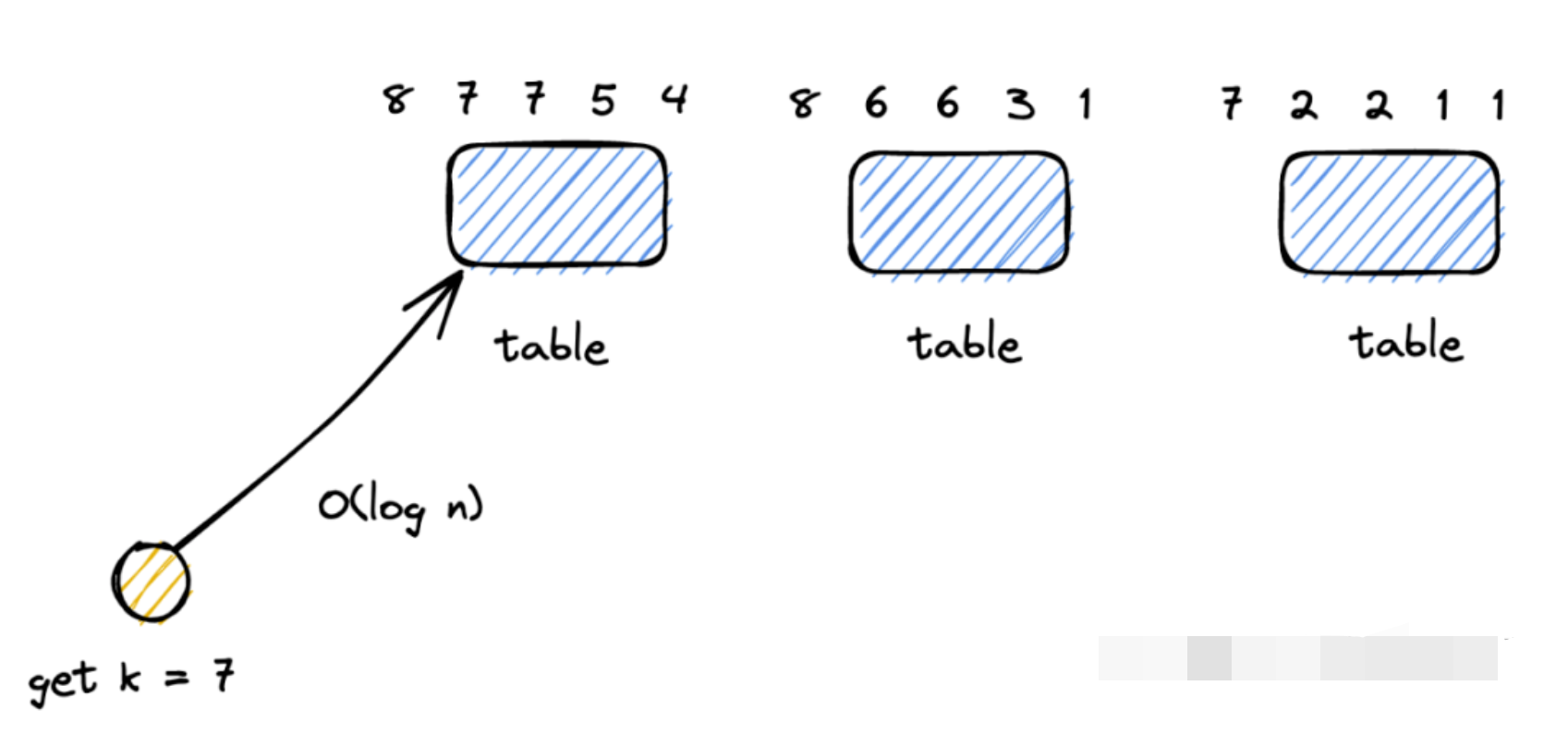
### 数据有序存储

数据冗余的优化探讨我们暂时先告一段落，接下来我们聊聊顺序写存在的另一个问题：很差的读性能。

目前现状是，每查询一个k，需要在全量数据范围内执行逆序遍历操作，直到找到满足条件的第一笔 kv 对记录位置。

要想降低读流程的时间复杂度，就需要在 kv 对的存储组织结构上做做文章。

这里我们可以采取的策略是，在组织每个 table 内的数据时，事先根据 k 进行数据排序，那么在查询环节中，我们就无需承受线性遍历的代价，而是能通过二分的方式在对数级的时间复杂度下获得我们想要的结果（即跳表的设计）。



但是大家需要注意，一旦 kv 数据在 table 中的存放顺序有了限制，这一点其实就和我们探讨这一问题的前提条件自相矛盾，已经打破了追加写策略的基础设定。

### 内存+磁盘

为了解决数据有序存储中遍历查找的问题，我们在磁盘文件disktable的基础之上，额外引入在内存中维护的memtable结构。

说明：提高查询的速率，肯定不是在磁盘上进行遍历查找，因此引入基于内存的数据结构memtable，并且这个结构需要是有序的，并且可以实现就地写，因为在内存中，可以容忍就地写带来的性能波动。

首先明确几个事项：

* memtable 在内存中缓存
* memtable 本身基于k进行有序存储
* memtable 中的写操作统一采用就地写
* memtable 是用户写操作的唯一入口
* memtable 数据量达到阈值后溢写到磁盘，成为disktable

重要的事情，我们再次强调一遍，**内存中的 memtable 采用就地写操作，而非追加写**。因为数据是维护在内存中的，因此哪怕写操作需要承受随机 IO 的代价，也处在可以接受的范围以内。

此外，当memtable中的数据量达到阈值后，再一次性flush到磁盘中成为 disktable。这个过程中实际上是以table为粒度，在磁盘中执行了我们所谓的“追加写”操作。

这样做还带来的一点好处是，由于所有的disk table都是由内部有序的memtable生成的，因此能做到disktable文件内部天然就是局部有序的。

然而，引入缓存之后，由此也引发出了三个新的问题：

* 内存是易失性存储，倘若 memtable 在溢写磁盘前就宕机，那么导致的数据丢失问题如何解决？

说明：预写（WAL）日志，将memtable的操作先持久化到磁盘。

* 在 memtable 溢写磁盘的过程中，外部的写操作需要阻塞，所带来的性能问题如何解决？

说明：通过memtable的分层实现，即active和inmutabletable实现。

* memtable 作为有序存储结构，内部采用什么样的数据结构进行实现？

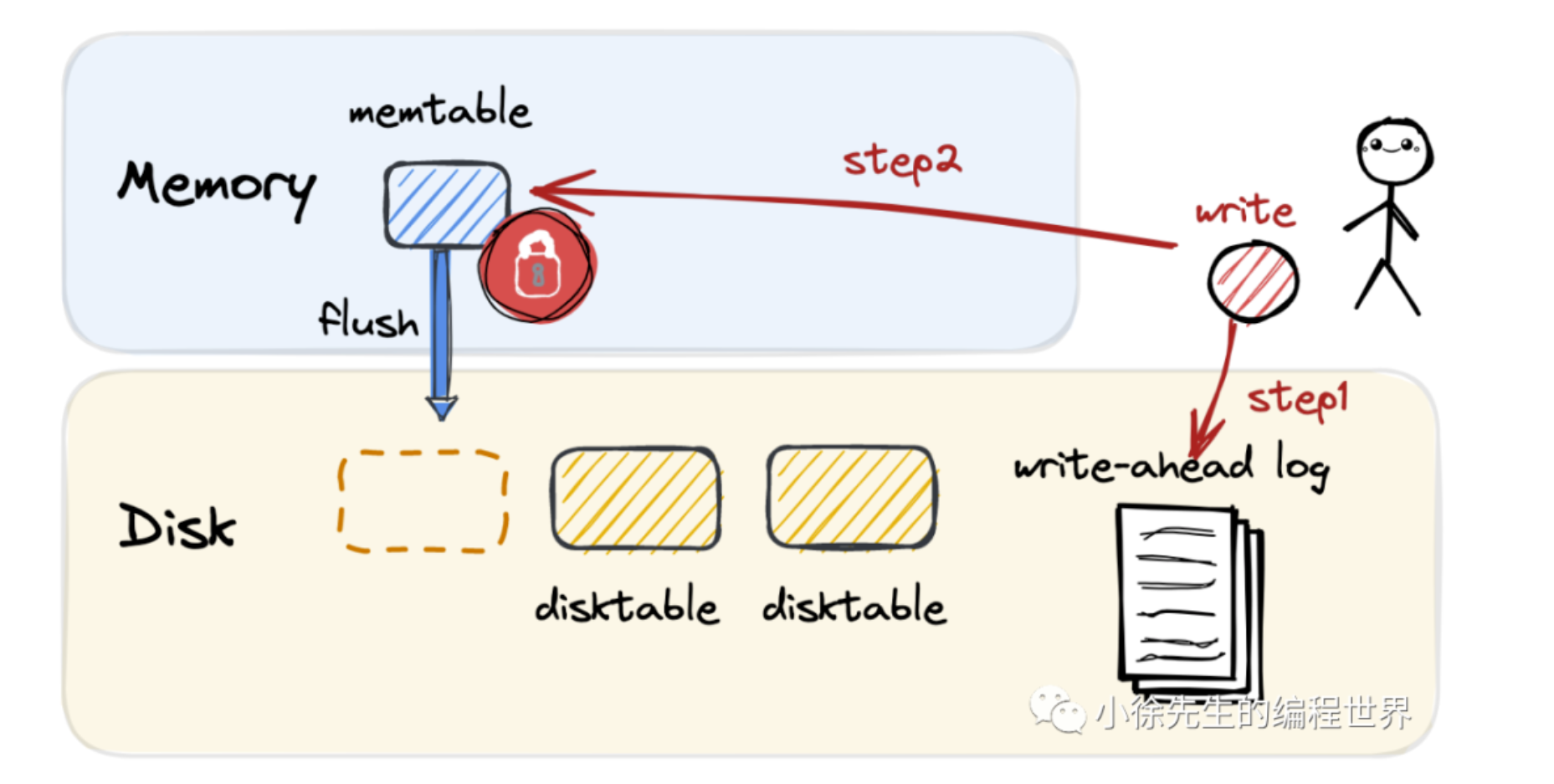
说明：采用跳表实现。

### 预写日志

首先面对第一个问题，如何避免因宕机导致memtable数据丢失。

解决这个问题的手段就是WAL（write-ahead log）预写日志技术：在将数据写入memtable，先通过追加写的方式，将操作记录到处于磁盘的WAL当中，这样哪怕宕机导致内存数据丢失，也能通过重放WAL的方式，重新恢复memtable 的数据。

此外，WAL和memtable可以建立对应关系，每当一个memtable被溢写到磁盘中成为disktable，其发生数据丢失问题的风险也就随之消除，因此对应的 WAL也就可以删除了。并且，由于WAL中也是追加写的操作，属于磁盘顺序IO，因此性能不会成为瓶颈。

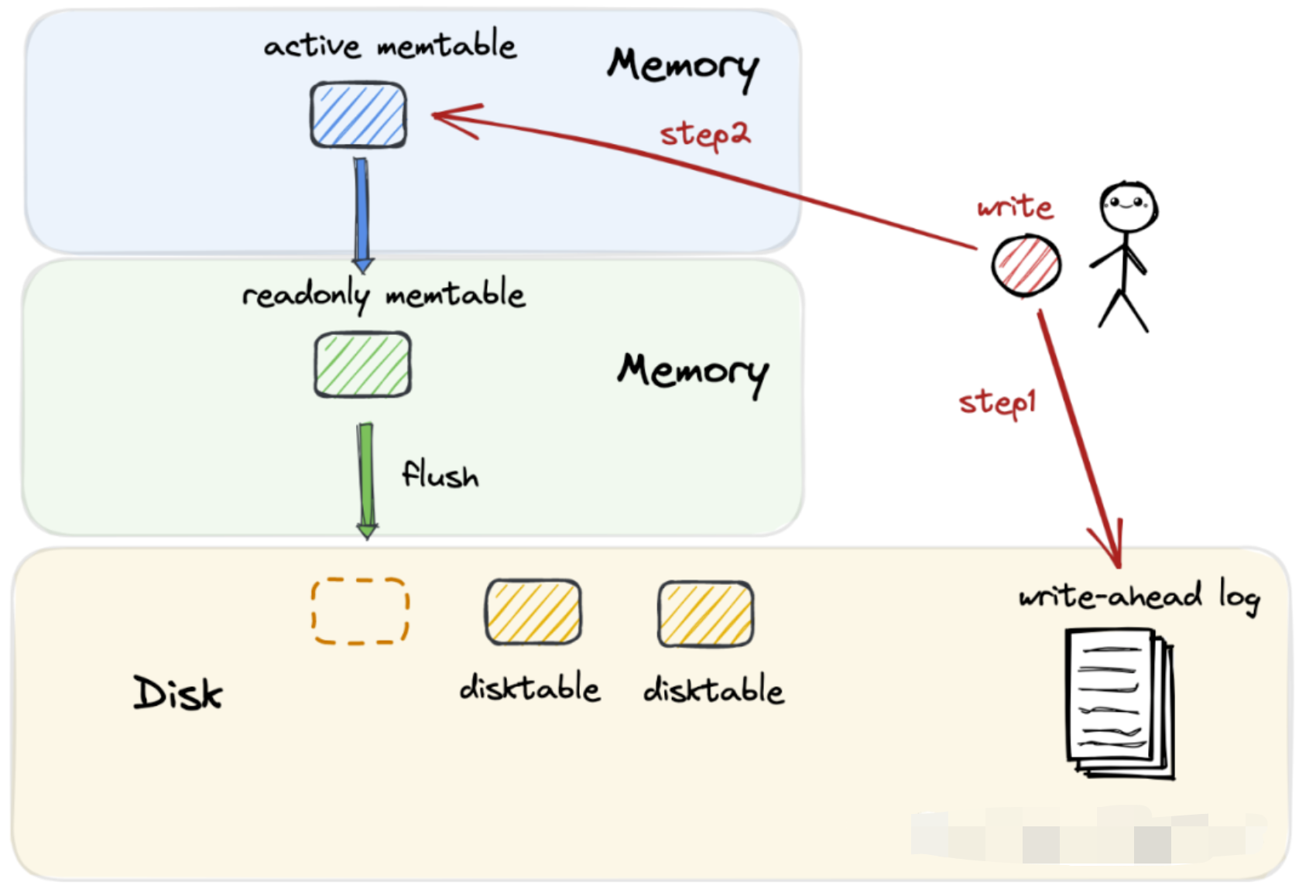


### 内存读写+只读分层

接下来面对第二个问题，在memtable 溢写磁盘期间，到来的写操作如何处理？

这里的解决思路是“金蝉脱壳”。每当 memtable 需要溢写时，就将其一分为二，将已有的旧数据归属到readonly memtable部分，成为一个只读的数据结构，专注于执行将其溢写到磁盘的流程；于此同时，建立一个全新空白的 active memtable，作为写操作的新入口，这样两个流程之间就实现了解耦，写操作不再需要阻塞。

由此可见，active memtable 持有的是最新的数据，readonly memtable 则次之，已经落到磁盘的 disktable 则再次。

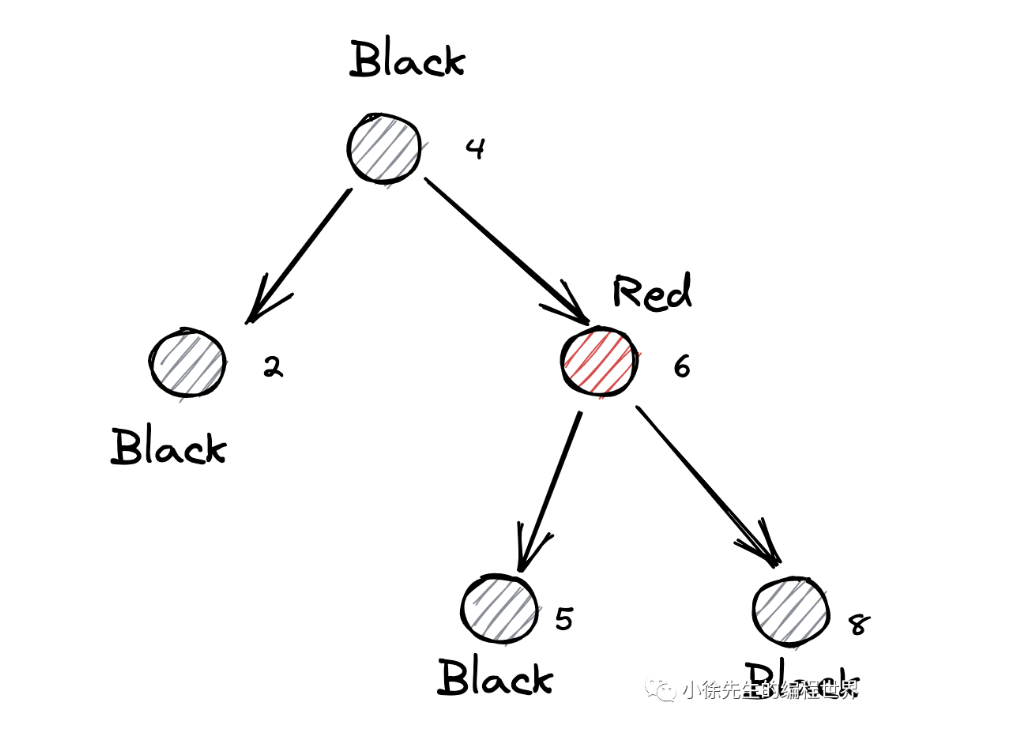


### 内存数据结构/memtable

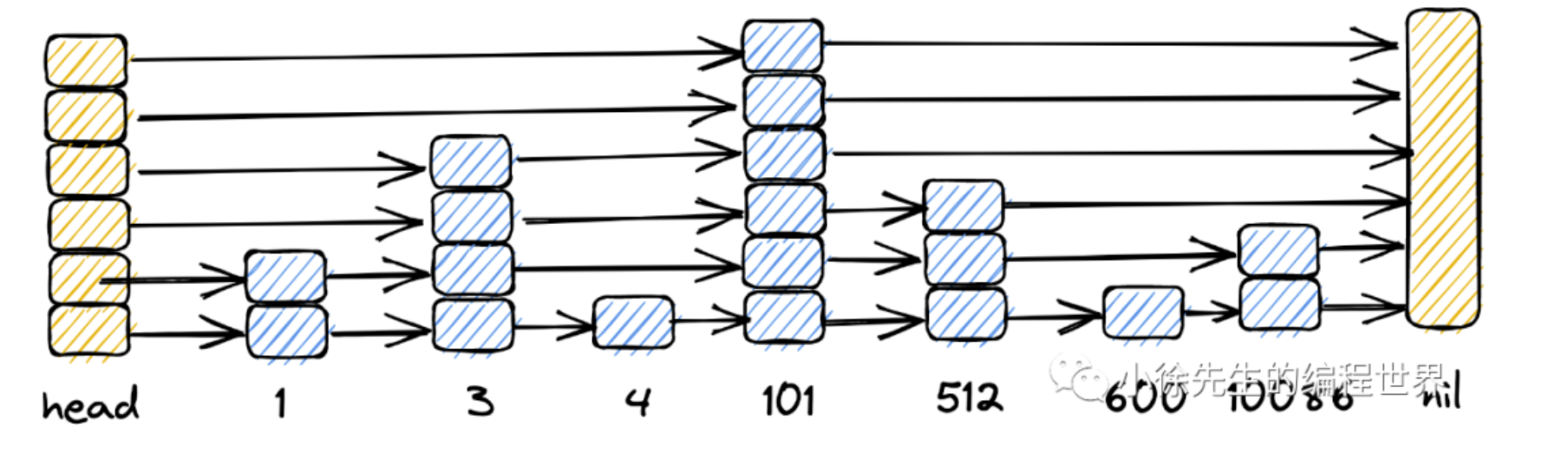
再聊第三个问题，内存中的 memtable 采用什么样的数据结构进行实现，其背后的核心要求是作为数据结构需要是一种有序表，能基于 k 进行数据的有序存储，从而保证所有的读操作和写操作都能在O(logN) 的时间复杂度之内拿下。

在这些诉求之下，进入视野的候选项包括红黑树（red-black tree）和跳表（skip list）两种。

红黑树结构示意如下：



跳表结构示意如下



在读写性能上，跳表和红黑树的性能表现不分伯仲。但跳表相比于红黑树具有两大核心优势：

* 更简单的实现
* 更细的并发锁粒度

作为存储组件，memtable 中的有序数据不可避免地会被用户并发读写访问，因此需要加锁保证临界资源的安全性和一致性。

在锁粒度上，红黑树由于自身染色的机制，每次写操作时都需要对对整棵树的（全量数据）进行加锁，相对笨重；

而跳表则不同，在执行写操作时，可以只针对插入节点局部范围内以及对跳表全局最大高度进行加锁，拥有着更细的锁粒度，在很多场景中是可以做到并发写的。

综上，跳表无疑是更好的选择。同时，也是 rocksdb 中默认采用的 memtable 数据结构。

### 磁盘分层

磁盘文件需要要分块细化粒度，由此诞生出了一系列的disktable。

所有disktable是由内存中的memtable溢写得到的，这样disktable就天然具有两大优势：

* disktable内部不存在重复kv对数据，因为memtable执行的是就地写操作
* disktable内部的kv对数据是有序的，因为memtable数据本身就是有序的

然而，disktable间还存在一个局限：由于每个memtable视野有限，只能做到自身范围内的k去重和排序。因此，不同disktable之间可能存在重复冗余的 kv 对数据，且不同disktable之间的数据无法做到全局有序。

理清了现状，我们再在此基础之上，进一步引入磁盘disktable分层的概念。

* 首先，我们将磁盘整体分为level0-levelk共计k+1层
* 每个level层中的disktable数量保持一致
* level(i+1)中disktable 的容量大小固定为level(i)的T倍，T为常量，通常取值为10左右
* 数据流向是由浅入深，层层递进，即由level(i) -> level(i+1)
* memtable溢写的数据落到level0
* levelk作为兜底
* 当某个level内数据总量达到达到阈值时，会发起level(i) -> level(i+1)的归并操作
* 数据从level(i)流向level(i+1)过程中，通过归并操作进行去重和排序，保证level(i+1)中kv数据无重复且全局有序

结合上述设定，我们可以得出以下结论：

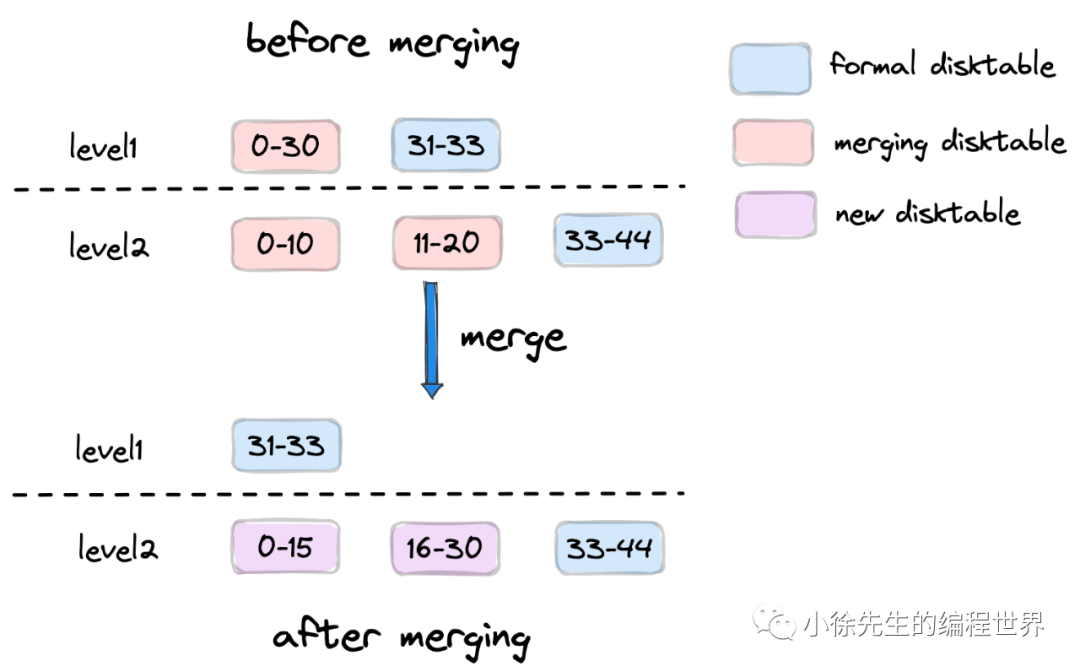
* level0 是特殊的，其中 disktable 之间可能存在冗余的 kv 对数据且不保证全局有序，因为其数据来自 memtable
* level1~levelk 中单层之内没有冗余的 kv 对数据，且保证全局有序
* 不同 level 层之间可能存在冗余的 kv 数据
* 较热（最近写入）的数据位于浅层，较冷（更早写入）的数据位于深层
* levelk 作为最深的一层，整体沉淀的数量达到全局的百分之九十左右

下面举一例，说明一下从 level1 到 level2 的数据归并过程。

假设此时 level1 层的数据总量已经达到阈值，接下来需要发起 level(1) -> level(2) 的归并操作：

* 从level1中随机选择一个 disktable，尝试将其合并到 level2. 由于数据是有序的，我们可以拿到其中 k 的取值范围. 假设其中最小的 key k\_min = 3，最大的 key k\_max = 30，记为[3,30]
* 假设 level2 中有 2 个 disktable 的 k 范围和待合并文件存在重叠，分别为 [0,16] 和 [17,32]
* 将 level1 的 [3,30] 与 level2 的 [0,10] 和 [11,20] 合并，这个过程本质上是个归并排序操作
* 新生成的 disktable [0,30] 不急于插入 level2 ，会根据 level2 中 disktable 的大小规模将其拆分为合适的数量
* 假设拆分得到的两个新的 disktable 分别为 [0,15] 和 [16,30]，将其插入到 level2
* 对应的老数据 level1 的[3,30] 以及 level2 的 [0,10] 和 [11,20] 都被被新数据替代，因此需要删除

值得一提的是，倘若因为这一合并操作，导致level2的数据容量又超出阈值，则会进一步引起level 2到level 3的数据合并操作，以此类推，层层递进。



### 磁盘数据结构/sstable

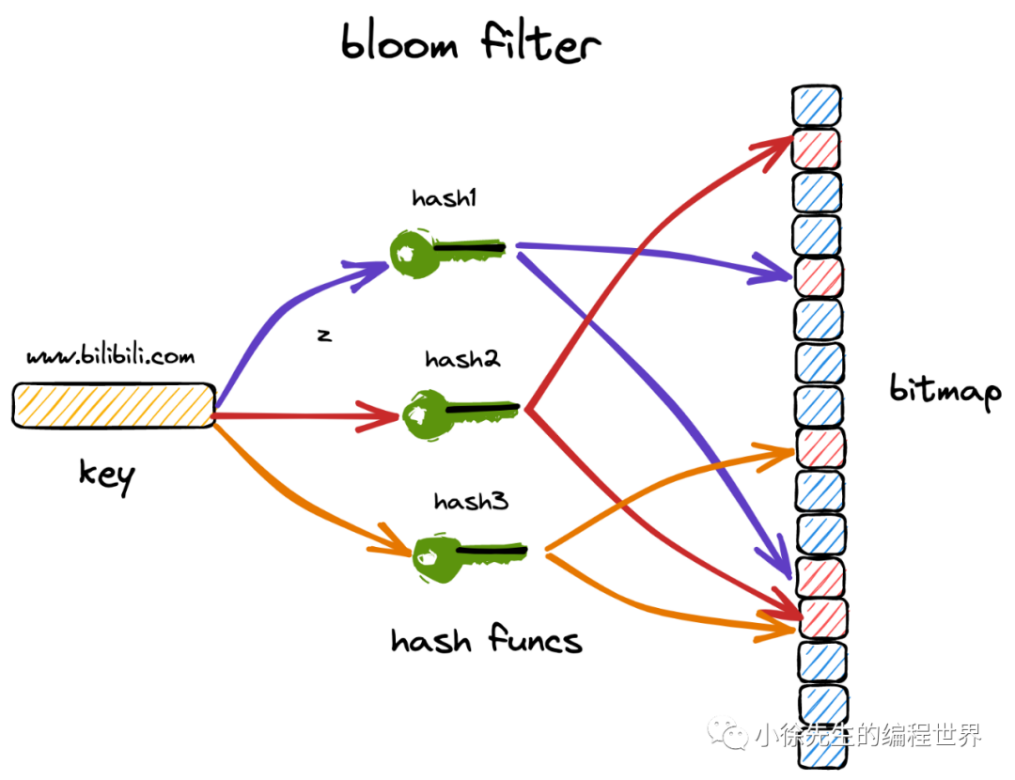
事实上，在 lsm tree的设定中，对于前文提到的每个磁盘文件块 disktable，设计了一类专门的数据结构sstable（sorted string table）。

每个 level 的 sstable 容量大约是上一层的10倍，因此一旦到了深层，sstable 的容量可能很大，对应展开的读操作会略显笨重。

sstable 在此基础上进行了优化：

* 首先 sstable 内部会进一步将 table 拆分为多个 block 块，其在逻辑意义上从属于同一个 sstable；
* 其次，sstable 中会额外维护一个索引信息，其中记录了每个 block 的 k\_min 和 k\_max 以及每个块中各行的 k\_max 和 k\_min，便于辅助的查询操作
* 此外，lsm tree 还维护着一个全局索引信息，记录着不同 level 中，每个 sstable 对应的k\_max 和 最小 k k\_min 的范围
* 最后，每个 sstable 还维护着一个布隆过滤器 bloomfilter，用于快速判断一个 k 是否存在于当前 sstable 中.

由于 bloom filter 具有假阳性的特点，因此判定不存在的 k 是必然不存在的，然而判定为存在的 k 也可能存在着一定的失误概率。

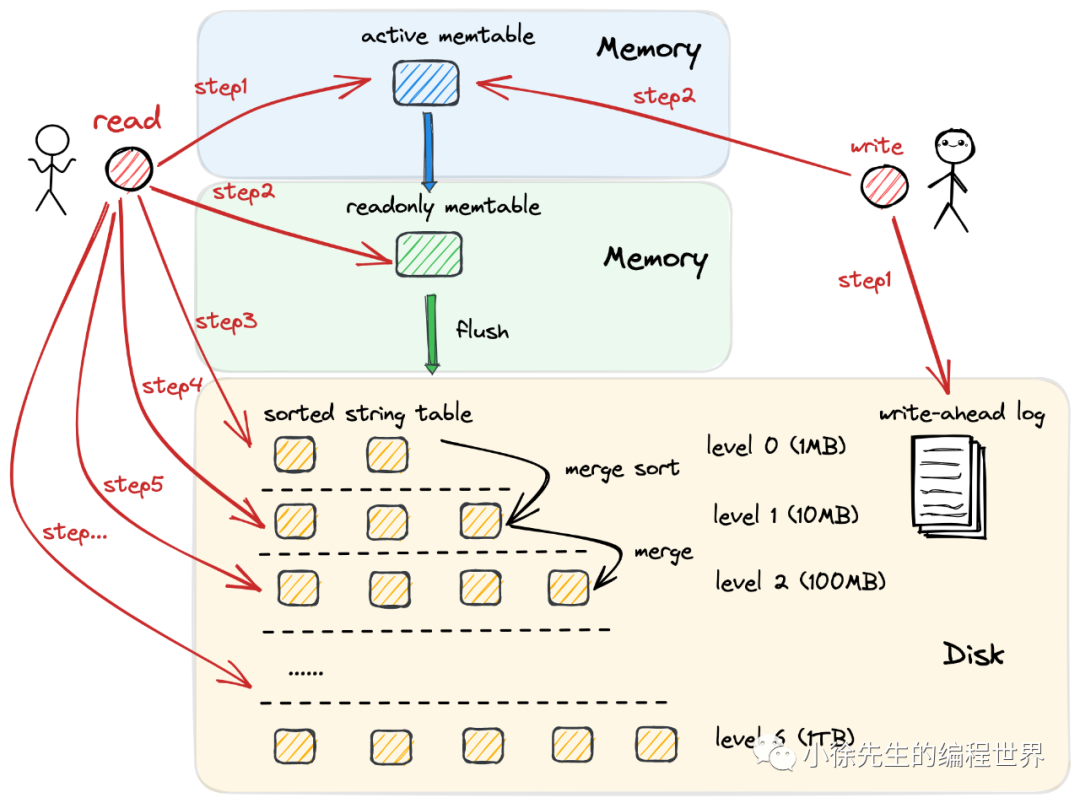


### 纵览lsm tree

到此为止，有关于 lsm tree 的拼图碎片已经集齐了，下面我们回过头来做个拼接和总览。

lsm tree全称 Log Structure Merge Tree，其核心设定如下：

* 存储介质主要依赖磁盘（sstable），但上层也会借助内存的辅助（memtable）
* 内存（memtable）就地写，磁盘（sstable）顺序写
* 写入口为可读可写的 active memtable
* 达到阈值后 active memtable 转为只读的 readonly memtable
* memtable 保证有序，默认基于跳表实现
* 由于 sstable 来自 memtable，每个sstable 内部无冗余数据且有序
* 磁盘文件分层（level0~levelk），上层为近期写入的热数据，下层为较早写入的冷数据
* level(i+1) sstable 大小恒定为 level(i) 层 T 倍
* level0 sstable 之间存在冗余数据
* level1~levelk 单层内无冗余数据且全局有序
* 数据沿着 level 0 -> level k 的方向合并，自顶向下流动



## 数据库文件

主要有以下几种类型sst文件，CURRENT文件，manifest文件，log文件，LOG文件和LOCK文件。

sst文件存储的是落地的数据，CURRENT文件存储的是当前最新的是哪个manifest文件，manifest文件存储的是Version的变化，log文件是rocksdb的write ahead log，就是在写db之前写的数据日志文件，LOG文件是一些日志信息，是供调试用的，LOCK是打开db锁，只允许同时有一个进程打开db。

### sst文件

### CURRENT文件

### manifest文件

### log文件

### LOG文件

### LOCK文件

## Memtable

参考：

Memtable写入逻辑：<http://mysql.taobao.org/monthly/2018/08/08/>

解析memtable、跳表和内存池Arena：

<https://www.cnblogs.com/JayL-zxl/p/15025932.html>

### SkipList/跳跃表

### Pipeline write

#### 概述

To enable pipelined write, simply set Options.enable\_pipelined\_write=true. db\_bench benchmark shows 20% write throughput improvement with concurrent writers and WAL enabled, when DB is stored in ramfs and compaction throughput is not the bottleneck.

要启用流水线写入，只需设置 Options.enable\_pipelined\_write = true。当DB存储在ramfs中并且压缩吞吐量不是瓶颈时，db\_benchbenchmark显示启用并发编写器和WAL可提高20%的写入吞吐量。

**pipeline write实现了哪些功能：**

流水线会负责把多个写请求打包成一个WriteGroup，把这些写请求的wal日志打包成连续的wal日志一起写下去，然后负责通过其它线程，把各自的数据插入到metmable当中。

流水线本身也实现了switch metmable与 metmable insert两种请求的barrier本身也兼顾了stalling write机制。

图示

描述已自动生成

这里pipeline的主要意思，就是上一轮的Group写入wal结束以后，下一轮新的Group又可以开始写wal。由不同的线程（Group）像流水线一样接替写wal日志。

图表, 箱线图

描述已自动生成

而pipeline的主要逻辑就是实现上述功能，如何在并发的写请求里面挑出来GroupLeader，GroupFollower，如何实现多个Group之间流水线一样的写入wal日志。

同时pipeline还要兼容 switchmemtable与写入metmable的并发。

#### 关键对象

WriteBatch 每一个写请求，都会带自己的WriteBatch，它里面记录了事务写入的key value数据。

WriteThread::Writer： 代码一个写请求，在并发纬度，也代表当前写请求所在的线程。多个Writer会形成一个双向链表。

WriteThread 每一个DB只有一个object，多线程（多个写请求）并发过程中如何实现pipeline写入的。

WriteThread ::WriteGroup 多个顺序的 WriteThread::Writer 的打包，顺序按照 WriteThread::Writer 调用 WriteThread::JoinBatchGroup 的先后顺序（严格意义上应该是指 WriteThread::Writer 链表的顺序）

#### 样例

下图展示了一个打包wal的样例， Writer1 把Writer1 Writer2 Writer3 打包成一个WriteGroup。

1、初使状态，当前的DB还没有任务一个写请求进来。

2、第一个写请求(Writer1)进来，因为 newest\_writer\_ 为空，Writer1（它所在的线程）被委派成了 STATE\_GROUP\_LEADER 角色，负责把多个writer的wal日志进行打包，并写入。 而 newest\_writer\_ 指向了 Writer1。

3、第二个写请求进来(Writer2)，直接通过link\_older指针挂到Writer1的后面，因为当前已经有STATE\_GROUP\_LEADER ，所以它不会成STATE\_GROUP\_LEADER ,等待被委派其它角色。

4、第三个写请求进来(Writer3)，直接通过link\_older指针挂到Writer1的后面，因为当前已经有STATE\_GROUP\_LEADER ，所以它不会成STATE\_GROUP\_LEADER ,等待被委派其它角色。

5、Wrtier1 所在EnterAsBatchGroupLeader，把Writer1 Writer2 Write3 要写入的数据打包成一个WriteGroup，叫做wal\_write\_group, WriteGroup的leader 指向 Writer1, WriteGroup的last\_writer\_ 指向Writer3。

6、Writer1 所在的线程，把wal\_write\_group写入wal 日志（rocksdb的流程，tdsql 3.0 不需要）

7、Writer1 ExitAsBatchGroupLeader

case1: 如果这个过程中，没有新来的write请求，也就是说 7.1 先于7.2 发生，那么 newest\_writer\_会被置为nullptr，Writer4 在linkOne的时候，发现newest\_writer\_为nullptr，也会像Writer1 一样，被委派成 STATE\_GROUP\_LEADER 角色

case2: 如果 7.2 先于 7.1发生，ExitAsBatchGroupLeader会把Writer4 置为 STATE\_GROUP\_LEADER

图形用户界面

中度可信度描述已自动生成

#### writer角色的转换

WriteThread::Writer 通过 WriteThread::JoinBatchGroup 后，会被赋予各种角色，下面是同一个writer在各个角色之间的转换

图示

描述已自动生成

#### 关键代码分析

##### EnterAsBatchGroupLeader

该函数是一个write抢到了GroupLeader的角色，开始设置Group的边界，把哪几个Write打包成一起。

这里有两个条件会限制group的大小：

1. 这个group包含所有的WriteBatch的大小
2. 找到了最后一个writer了

图示

描述已自动生成

##### WriteThread::LinkOne

尝试把当前writer的old\_linker 连接到 newest\_writer\_ 如果中间没有其它并发linkone，那么年上一次 newest\_writer\_ 是不是nullptr，如果是的话，返回true : link as leader

图示

描述已自动生成

/\*

\* 尝试把当前writer的old\_linker 连接到 newest\_writer\_

\* 如果中间没有其它并发linkone，那么年上一次 newest\_writer\_ 是不是nullptr，如果是的话，返回true : link as leader

如果它链接到已经有了 leader了，它就简单的连接到 newest\_writer\_ 的后面。

\*/

bool WriteThread::LinkOne(Writer\* w, std::atomic<Writer\*>\* newest\_writer) {

assert(newest\_writer != nullptr);

assert(w->state == STATE\_INIT);

Writer\* writers = newest\_writer->load(std::memory\_order\_relaxed);

// w（当前的writer） 来抢 newest\_writer 这个位置，如果抢占成功了。而且上一次 newest\_writer 是nullptr，

// 那么当就是 leader writer, return true

// 如果 newest\_writer 上个值有人了，那说明它没有抢到leder

while (true) {

// If write stall in effect, and w->no\_slowdown is not true,

// block here until stall is cleared. If its true, then return

// immediately

if (writers == &write\_stall\_dummy\_) {

if (w->no\_slowdown) {

w->status = Status::Incomplete("Write stall");

SetState(w, STATE\_COMPLETED);

return false;

}

// Since no\_slowdown is false, wait here to be notified of the write

// stall clearing

{

MutexLock lock(&stall\_mu\_);

writers = newest\_writer->load(std::memory\_order\_relaxed);

if (writers == &write\_stall\_dummy\_) {

stall\_cv\_.Wait();

// Load newest\_writers\_ again since it may have changed

writers = newest\_writer->load(std::memory\_order\_relaxed);

continue;

}

}

}

w->link\_older = writers;

if (newest\_writer->compare\_exchange\_weak(writers, w)) {

return (writers == nullptr);

}

}

}

#### memtable writer 与 memtable switch的barrier

写memtable 与 memtable switch是不能并发的，这里pipeline构造了它们的barrier。

case1: 如果在pipeline write中调用了switch memtable，那么只有GroupLeader可以调用switch memtable，然后再调用 WriteThread::WaitForMemTableWriters 函数，等上一个Group所有的memtable writer都退出，然后再switch memtable，作为当前的groupLeader，当前是不会写memtable，只要保证上一个Group的memtable writer都退出，就可以保证当前没有并发任务写memtable

waitForMemTableWriters 函数分析

图示

描述已自动生成

case2: 不在pipeline write调用switch metmable，通过WriteThread::EnterUnbatched 抢一个空的pipeline的GroupLeader，然后再等上一个Group的memtable writer退出，然后再switch memtable，然后调用WriteThread::ExitUnbatched 让出自己GroupLeader的位置。

static WriteThread::AdaptationContext eu\_ctx("EnterUnbatched");

void WriteThread::EnterUnbatched(Writer\* w, InstrumentedMutex\* mu) {

assert(w != nullptr && w->batch == nullptr);

mu->Unlock();

bool linked\_as\_leader = LinkOne(w, &newest\_writer\_);// 尝试去抢一个GroupLeader

if (!linked\_as\_leader) {//

STORAGE\_LOG\_INFO("linked\_as\_follower WriteThread::EnterUnbatched::Wait enable\_pipelined\_write:%d",

enable\_pipelined\_write\_);

TEST\_SYNC\_POINT("WriteThread::EnterUnbatched:Wait");

// Last leader will not pick us as a follower since our batch is nullptr

AwaitState(w, STATE\_GROUP\_LEADER, &eu\_ctx);//没有抢成功，就等自己被设置为GroupLeader

} else { // 抢GroupLeader 成功

STORAGE\_LOG\_INFO("linked\_as\_leader WriteThread::EnterUnbatched not wait enable\_pipelined\_write:%d",

enable\_pipelined\_write\_);

}

if (enable\_pipelined\_write\_) {

WaitForMemTableWriters();//等上一个Group的memtbale writer都退出

}

mu->Lock();

}

### Flush

参考：

Memtable flush分析：<http://mysql.taobao.org/monthly/2018/09/04/>

Flush是指将memtable的数据导入到sst中，变成持久化存储，就不怕数据丢失了。

1、首先在memtable的add的时候，会检测是否memtable的大小达到了max write buffer，如果是就将should\_flush\_置为true，并会在WriteBatch的Handler里面调用CheckMemtableFull，将当前column family加入flush\_scheduler。

2、在Write的时候，调用ScheduleFlushes，将需要flush的column family的memtable切换一个新的，同时将原来的memtable加入cfd的imm中，如果这个column family data的imm数量大于min\_write\_buffer\_number\_to\_merge，并启动一个新的线程调用BGWorkFlush。

由于真正的Flush过程是在另一个线程完成的，所以这个地方并不会block写过程。

另外，如果total\_log\_size大于max\_wal\_log\_size并且不是只有一个column family，也会触发flush，因为flush能将memtable持久化到磁盘上，同时对应的wal就可以删除了。

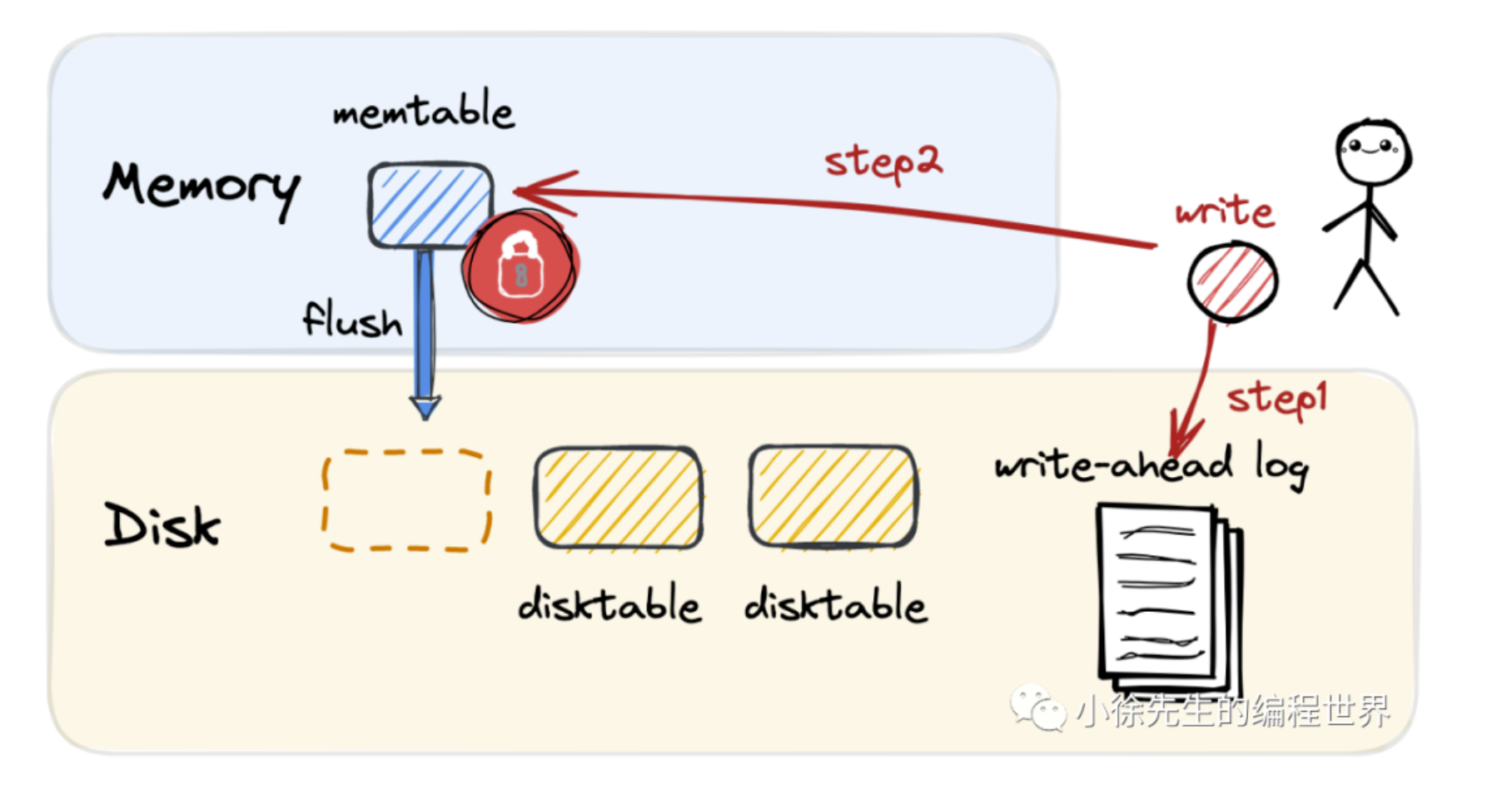
## WAL

### 概述

如何避免因宕机导致 memtable 数据丢失。

解决这个问题的手段就是WAL（write-ahead log）预写日志技术：在将数据写入memtable，先通过追加写的方式，将操作记录到处于磁盘的WAL当中，这样哪怕宕机导致内存数据丢失，也能通过重放 WAL 的方式，重新恢复 memtable 的数据。

此外，WAL和memtable可以建立对应关系，每当一个memtable被溢写到磁盘中成为disktable，其发生数据丢失问题的风险也就随之消除，因此对应的 WAL 也就可以删除了。并且，由于WAL中也是追加写的操作，属于磁盘顺序IO，因此性能不会成为瓶颈。



### 原理

1. 每次写操作，rocksdb会先写write ahead log，然后才会写db。
2. write ahead log可以配置到单独的空间，并且可以配置WAL文件的单独的删除机制。这种原因是为了保存WAL文件，达到特殊的目的，比如，其他sst文件放在不可靠存储里面，而WAL放到可靠存储里面。

## SSTable

### 版本管理

### Compaction

参考：

Level Compact 分析：<http://mysql.taobao.org/monthly/2018/10/08/>

#### Rocksdb CompactionIterator实现分析

#### Rocksdb compaction files的选取逻辑

##### 概述

RocksDB的Compaction流程可以保证不同sst文件之间key-value之间的有序性，数据的压缩存储，清理过时数据。以下内容主要介绍的是Compaction过程对于Compaction FileMetaData的文件选取逻辑。

下图所列出来的是Compaction的整体流程，包括三个阶段：Prepare key， Process key 以及 Write key.

图形用户界面

描述已自动生成

Prepare key： 主要是计算每一层的Score，确定需要参与Compaction的file， 然后将参与Compaction的文件分解为多个SubCompaction（SubCompaction主要是针对L0向其他层进行compaction时，也能够并行化）。

Process key:将SubCompaction中的file构建一个MergingIterator，MergingIterator实现逻辑是小顶堆，遍历小顶堆，根据不同的key\_type, 对key value 数据进行合并以及排序操作。

Write key: 将key-value数据写入对应的block 数据之中。

##### 实现

以下部分主要内容为Prepare key实现

图示

描述已自动生成

上图为Compaction prepare key 较详细的实现，主要内容为以下几部分

根据Function Score 计算每一层的Score， 这一步的计算方式因level\_compaction\_dynamic\_level\_bytes的值而不同。

遍历每一层level，取出该层参与Compaction的文件 （Score>1），根据Cleancut取出input\_level 以及 output\_level 需要参与此次Compaction 的FileMetaData.

将以上取出来的FileMetaData的key边界取出来，边界排序去重，计算SubCompaction的个数，确定最终Subcompaction key的边界范围。

为每一个SubCompaction创建一个线程，进入下一个阶段（Process key）

###### Compaction触发条件

Rocksdb的compaction的触发条件有两类：

1、某一个level的数据太多

VersionStorageInfo的compaction\_score\_的计算方法是level0的是当前文件数目/level0\_file\_num\_compaction\_trigger，其他层是该层当前文件大小总和/该层的配置的允许文件总和最大值。

基于level的存储的compaction总的来说，就是一次挑选某一个level的一个文件，然后将该文件和高level的多个相交文件merge，最后生成多个高level的文件。具体的细节是：每次会挑选compaction score最高的一个level，并在这个level中找到一个文件大小最大，并且上一个level的相交文件没有在compaction的一个文件。

2、seek太多

疑问：

如果option里面soft\_rate\_limit设置的为0.0和1之间，compaction并不会触发，但是会触发write delay，这是为什么？

以下是判断是否发生Compaction的主体函数：

1. 是否存在超时的sst
2. files\_marked\_for\_periodic\_compaction\_ 、return bottommost\_files\_marked\_for\_compaction\_ 以及files\_marked\_for\_compaction\_ 文件是否为空。
3. 遍历每一层，判断其Score 是否大于1

###### Compaction Score的计算

###### Comapaction每一层level大小的计算过程

###### 挑选参与Compaction的FileMetaData文件

###### Compaction job根据获取到数据分配compaction线程

#### Rocksdb compaction流程示意图

图示

描述已自动生成

### 写stall

在DBImpl也就是db的实例里面有一个WriteController，同时在ColumnFamilySet里面也有这个WriteController的指针，这个数据结构会控制db的写stall行为。

在ColumnFamily进行SuperVersion变更的时候（增加新memtable，flush增加sst，compaction）都会查看需不需要stall Write，stall的条件是：

1）imm的数量大于等于option允许的最大数目

2）level0的文件的数量大于option允许的数目

如果没有满足上面两项，但是compaction score比较大会delay写

## 读写放大

### 写放大

RocksDB的写会写WAL（Write Ahead Log），如果sync的话，会写一次磁盘，然后会写memtable，写rocksdb的时候有可能会卡住。

说明：即sync刷盘可能会拖慢写操作，导致写放大。

### 读放大

RocksDB的读，会首先读memtable，如果memtable没有找到的话，会读下面level的数据，由于level0的多个sst会有交叠，所以每个sst都会通过filemeta判断在不在最小和最大的范围内，如果在就需要读这个sst的文件内容，来查看，其他level的sst文件不会有数据交叠的情况，所以只会有一个文件可能含有这个数据。

可以看出来读放大还是比较严重的。rocksdb为了减少读放大，增加了cache。

#### cache

**读cache：**

rocksdb的读cache分为两部分：table cache和block cache。这两个都是LRUCache。

block cache存储的block，包括index block和filter block（通过options可以配置）。

table cache存储的是table，是整个文件的meta信息和Foot信息。table\_cache\_size的消耗内存的大小是有Options里面的max\_open\_files决定的。

#### Bloomfilter

**Bloomfilter：**

bloomfilter的增加并不能减少写放大，因为bloomfilter是table范围的或者block范围的，而且bloomfilter是存储在文件中的，那么必须把这些从文件里面读出来后才能起到作用。

## 操作接口

### 写流程

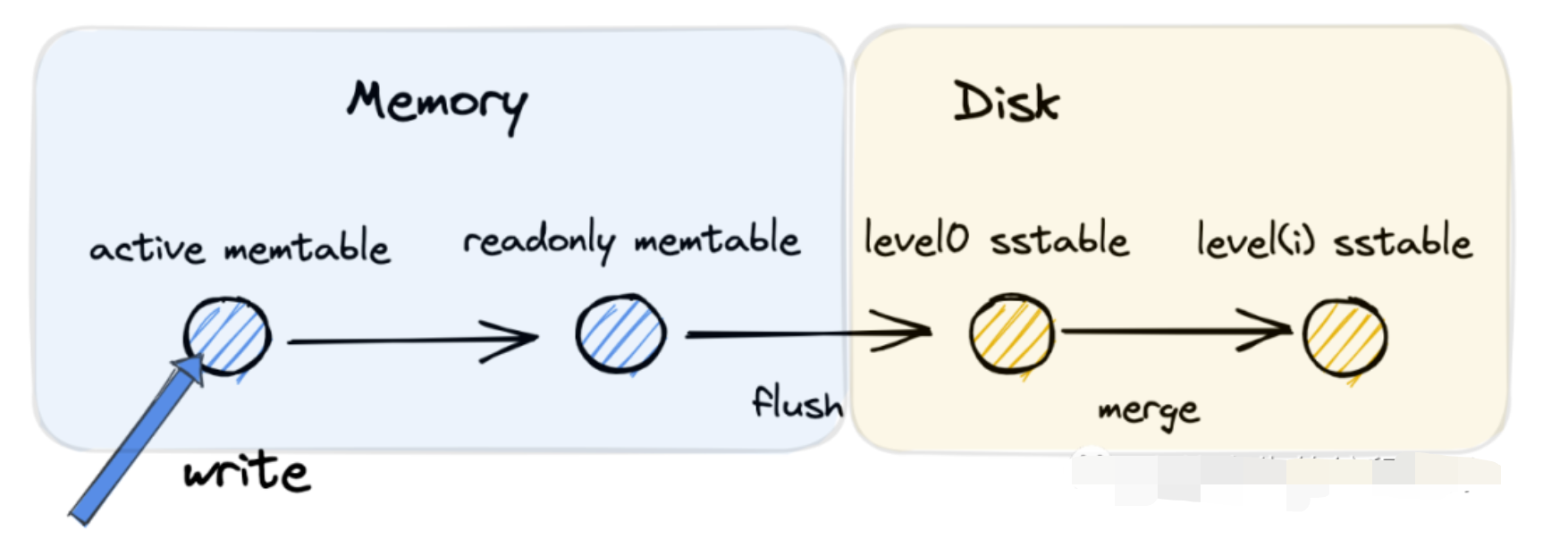
参考：

写入逻辑的实现：<http://mysql.taobao.org/monthly/2018/07/04/>

Write Prepared Policy：<http://mysql.taobao.org/monthly/2018/08/02/>

rocksdb写入时，直接以append方式写到log文件以及memtable，随即返回，因此非常快速。

memtable/immute memtable触发阈值后，flush到Level0 SST，Level0 SST触发阈值后，经合并操作(compaction)生成level 1 SST，level1 SST合并操作生成level 2 SST，以此类推，生成level n SST。



* 基于就地写模式（这里是优化为就地写而非顺序写，但是这里是基于内存而非磁盘的就地写，因此性能损耗非常小），写入内存中的active memtable
* active memtable达到阈值后转为只读的readonly memtable
* readonly memtable会flush到磁盘，成为level0的sstable
* level(i)层数据容量达到后，会基于归并的方式合并到level(i+1)，以此类推

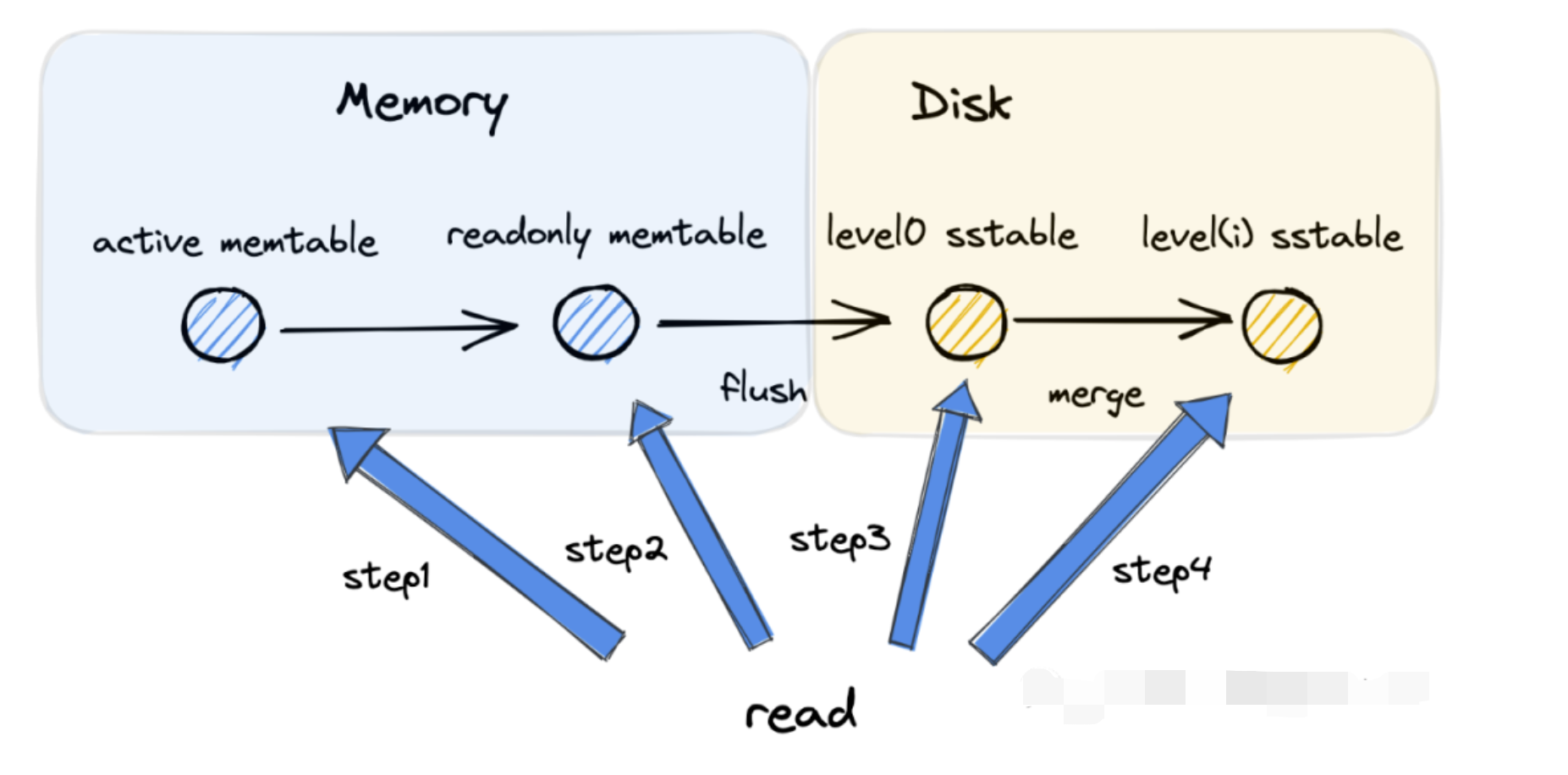
### 读流程

参考：

数据的读取：<http://mysql.taobao.org/monthly/2018/11/05/>

<http://mysql.taobao.org/monthly/2018/12/08/>

按照memtable --> Level 0 SST–> Level 1 SST --> … -> Level n SST的顺序读取数据。这和记录的新旧顺序是一的。因此只要在当前级别找到记录，就可以返回。



* 尝试读 active memtable
* 尝试读 readonly memtable
* 尝试读 level0，需要按照溢写顺序进行倒序，依次读 level0 中的每个 sstable（level0 sstable间数据可能冗余）
* 根据全局的索引文件，依次读level1~levelk，每个level最多只需要读一个sstable
* 读一个sstable时，借助内部的bloom filter和索引，加速查询流程

综上所述，在lsm tree架构下，一次读操作可以在常数级别的IO次数下完成，同时每次 IO 操作中则需要承受对应 table 内数据量对数级别的查询时间复杂度.

# 源码

## RocksDB的初始化

## create table

## select

## insert

## explain

## commit

## rollback

# 事务

参考：

TransactionDB 介绍：<http://mysql.taobao.org/monthly/2018/10/09/>

RocksDB TransactionDB事务实现分析：https://yq.aliyun.com/articles/272166/

## 基本概念

### LSN (log sequence number)

Q1：此处的LSN和时间戳的关系？在分布式系统下，一个事务的多个节点上的LSN之间应该是什么关系？

A1：TDStore已经修改LSN的获取方式为从MC来全局获取，事务开始时获取一次，事务提交时获取一次。

RocksDB中的每一条记录(KeyValue)都有一个LogSequenceNumber(后面统称lsn)，从最初的0开始，每次写入加1。该值为逻辑量，区别于InnoDB的lsn为redo log物理写入字节量。

这个lsn在RocksDB内部的memtable中是单调递增的，在WriteAheadLog(WAL)中以WriteBatch为单位递增(count(batch.records)为单位)。

Q2：WriteBatch，是否类似于OCC的写集？如果是，则读集是否应该维护一个类似WriteBatch的东东？

但后面谈到锁时，表明WriteBatch不等同于OCC的写集，因为写写并发提前被封锁机制禁止了。

WriteBatch是一次RocksDB::Put()的原子操作集合，不同的WriteBatch间是遵循ACID特性(要么完全成功要么完全失败，并且相互隔离)，结构如下：

WriteBatch :=

sequence: fixed64

count: fixed32

data: record[count]

从RocksDB外部能看到的LSN是按WriteBatch递增的(LeaderWriter(或LastWriter)最后一次性更新)，所以进行snapshot读时，使用的就是此lsn。

Q3：考虑RDB的snapshot和MVCC中的snapshot的异同点。

同时考虑，对于分布式事务中同一个事务的不同的节点上，子节点的快照的建立，依据的是谁（是子节点自己的LSN—不是，计划使用HLC）？

A3：分布式事务开始前，SQLEngine从mc获取开始时间戳，各个子节点以该开始时间戳作为snapshot

注意: 在WAL中每条WriteBatch的lsn并不严格满足以下公式(比如2pc情况下):

lsn(WriteBatch[n]) < lsn(WriteBatch[n+1])，可能相等

Q4：考虑为什么会出现此处说的“可能相等的情况”？

一个事务，是否可以使用多个WriteBatch，以实现对于大事务的支持？

A4：数据只有写入到memtable中，lsn才会加一，而并不是所有的writebatch都会向memtable里写数据，比如事务2pc中prepare阶段的prepare writebatch，只会写wal，而不会写memtable（写memtable的时机为commit阶段）。很少情况使用多个WriteBatch，对大事务的支持，在同一region上使用多个WriteBatch，未必有效果哈。从资源消耗上来讲，所有WriteBatch都是存储在内存中的，分成多个WriteBatch并不会降低内存的消耗。从实现方式上来讲，原生RocksDB通过WriteBatch保证事务的原子性，如果大事务使用多个WriteBatch，则这些WriteBatch之间还需要使用额外的手段保证原子性，会加大事务提交的开销。

Q4.1 回复还需要细致一点儿。

1、如一个WriteBatch最大支持多大的写事务量？即单个事务的更新量的极限是多少？

2、“很少情况使用多个WriteBatch”—是一定没有还是个别情况有？如果有，特别需要明确说明。

3、可以深入说明LSN对于数据可见性的价值。标志了一个事务的提交，是写入数据到memtable还是LSN加一，还是如图1中有commit的一个标志然后才写入数据到memtable？

`

### Snapshot

Snapshot是RocksDB的快照，实际存储的就是一个lsn.

class SnapshotImpl {

public:

// 当前的lsn

SequenceNumber number\_;

private:

SnapshotImpl\* prev\_;

SnapshotImpl\* next\_;

SnapshotList\* list\_;

// unix时间戳

int64\_t unix\_time\_;

// 是否属于Transaction(用于写冲突)

bool is\_write\_conflict\_boundary\_;

};

查询时如果设置了snapshot为某个lsn, 那么对于此snapshot的读来说，只能看到lsn(key)<=lsn(snapshot)的key，大于该lsn的key是不可见的。

Q5：为什么是一个双向列表？这样做有什么好处？

一个snapshot不是隶属于一个事务？

A5：维护snapshot双向列表的目的是，方便定位最旧的snapshot，以避免进行compaction操作时，将被该snapshot可见的中间版本清除掉，造成数据不一致

Q5.1：如果链表中间有个SI对应的事务已经提交，则compaction是否清理该SI对应的数据版本？

Q5.2： 一定需要双向链表，才可以找到最旧的SI？ 且一定需要该双向链表吗？

Q5.3：该双向链表是RDB的原生设计？

snapshot的创建和删除都需要由一个全局的DoubleLinkList (DBImpl::SnapshotList)管理，天然的根据创建时间(同样也是lsn大小)的关系排序，使用之后需要通过DBImpl::ReleaseSnapshot释放。snapshot还用于在RocksDB事务中实现不同的隔离级别。

### 隔离级别

Q6：隔离级别是如何实现的（这里没有具体说明）？实现效率可进行比较。优先级，中

A6：TDStore使用rocksdb提供的乐观事务，未提交事务所写入的数据会存储在WriteBatch中，对其它事务不可见，保证了Read Committed。在此基础上，RocksDB提供了SI隔离级别，事务开始时可以通过为事务设置snapshot，保证Repeatable Read。

为了实现事务下的一致性非锁定读(读可以并发)，不同的数据库(引擎)实现了不同的读隔离级别。SQL规范标准中定义了如下四种：

ReadUncommited ReadCommited RepeatableRead Serializable

Oracle No Yes No Yes

MySQL Yes Yes Yes Yes

RocksDB No Yes Yes No

ReadUncommitted 读取未提交内容，所有事务都可以看到其他未提交事务的执行结果。存在脏读。

ReadCommitted读取已提交内容

，事务只能看见其他已经提交事务所做的改变，多次读取同一个记录可能包含其他事务已提交的更新。

RepeatableRead 可重读，确保事务读取数据时，多次操作会看到同样的数据行(InnoDB通过NextKeyLocking对btree索引加锁解决了幻读)。

Serializable串行化，强制事务之间进行排序，不会互相冲突。

大部分数据库(如MySQL InnoDB、RocksDB)，通过MVCC都可以实现上述的在非排它锁锁定情况下的多版本并发读。

## RocksDB Transaction

简单的例子:

// 基本配置,事务相关操作需要TransactionDB句柄

Options options;

options.create\_if\_missing = true;

TransactionDBOptions txn\_db\_options;

TransactionDB\* txn\_db;

// 用支持事务的方式opendb

TransactionDB::Open(options, txn\_db\_options, kDBPath, &txn\_db);

// 创建一个事务上下文, 类似MySQL的start transaction

Transaction\* txn = txn\_db->BeginTransaction(write\_options);

// 直接写入新数据

txn->Put("abc", "def");

// ForUpdate写，类似MySQL的select ... for update

s = txn->GetForUpdate(read\_options, "abc", &value);

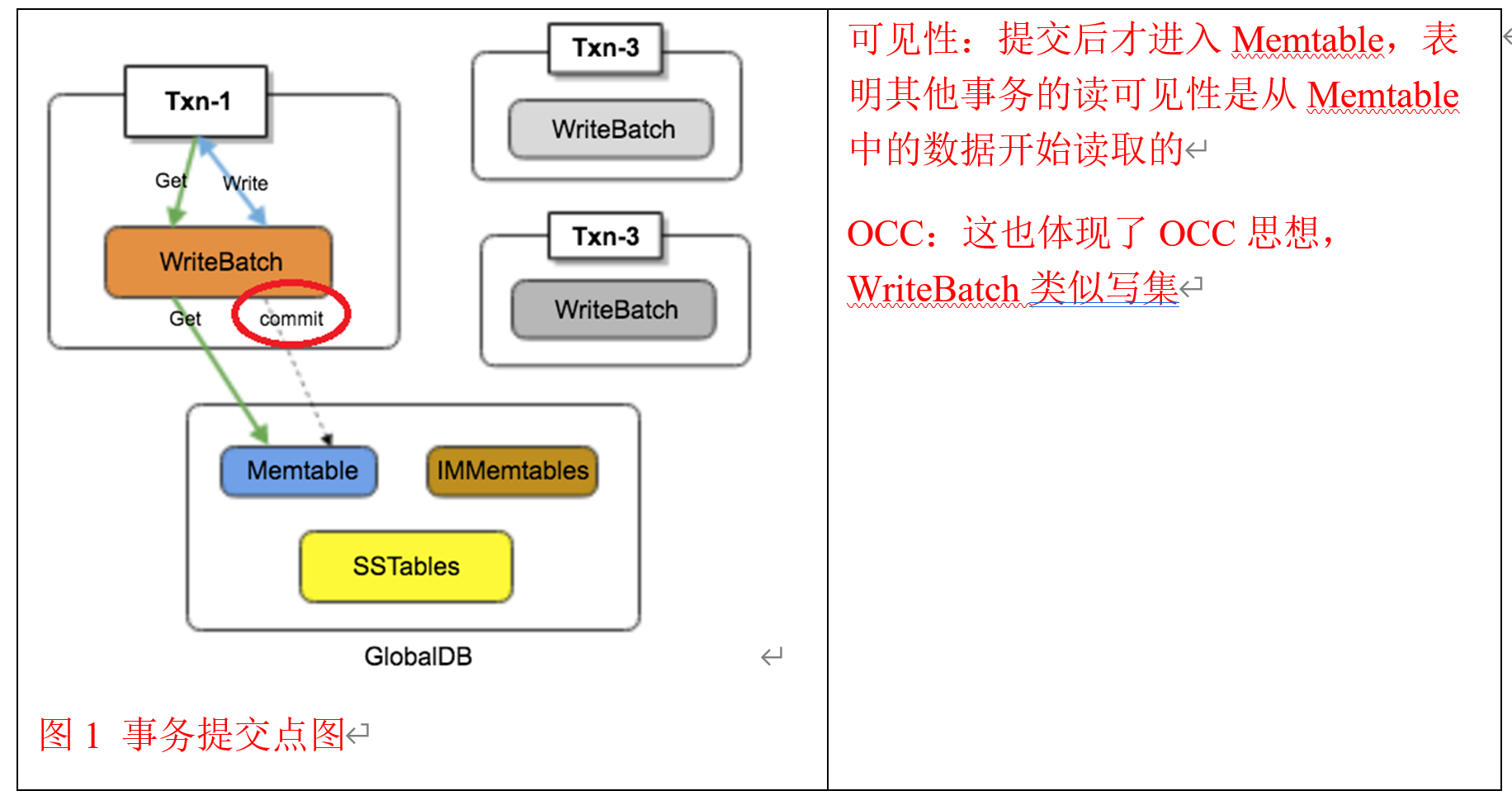
txn->Commit(); // or txn->Rollback();

如图1，RocksDB的一个事物操作，是通过事物内部申请一个WriteBatch实现的，所有commit之前的读都优先读该WriteBatch(保证了同一个事务内可以看到该事务之前的写操作)，写都直接写入该事务独有的WriteBatch中，提交时在依次写入WAL和memtable，依赖WriteBatch的原子性和隔离性实现了ACID。

Q7：分布式事务，父事务和子事务之间，需要传递什么信息以维系父子事务之间的关系？数据结构应该怎么修改？

A7：读写阶段，SQLEngine维护当前分布式事务涉及到的region列表；2PC阶段，SQLEngine将region列表同步给协调者，协调者再把region列表同步给各个参与者。同一个分布式事务中，各个region上的子事务trans id相同

Q7.1“协调者再把region列表同步给各个参与者”—为什么这么做？参与者什么情况下需要知道region列表中的所有内容?



有些单独写操作也可以通过TransactionDB直接写

txn\_db->Put(write\_options, "abc", "value");

txn\_db->Get(read\_options, "abc", &value);

用TransactionDB::Put()，内部会直接生成一个auto transaction，将这个单独的操作封装成一个transaction，并自动commit。所以在TransactionDB中，所有的入口内部都会转化成trasaction(所以显示的transaction是可以马上读取到了外面TransactionDB::Put()的数据，注意这不属于脏读)这个和MySQL的形式是类似的，默认每个SQL都是个auto transaction。但这种transaction是不会触发写冲突检测。

Q8: 讨论“auto transaction”的价值。

“用TransactionDB::Put()，内部会直接生成一个auto transaction，将这个单独的操作封装成一个transaction，并自动commit。”，这说明，一个put算是一个事务，那我们在一个事务中写入多条KV，则不能使用put？而是有另外的接口？

A8：这里我们用TransactionDB::BeginTransaction创建一个新的Transaction，再调用Transaction::Put，在该Transaction提交之前，写入的数据都在内存的WriteBatch里，对其它事务不可见，在提交阶段再将数据刷入到memtable中，以此保证事务多次写入的原子性。

### GetForUpdate

类似MySQL的select ... for update，RocksDB提供了GetForUpdate接口。区别于Get接口，GetForUpdate对读记录加独占写锁，保证后续对该记录的写操作是排他的。所以一般GetForUpdate会配合snapshot和SetSnapshotOnNextOperation()进行读，保证多个事务的GetForUpdate都可以成功锁定，而不是一个GetForUpdatech成功其他的失败。尤其是在一些大量基于索引更新的场景上。

### 事务并发

不同的并发事务之间，如果存在数据冲突，会有如下情况：

* 事务都是读事务，无论操作的记录间是否有交集，都不会锁定。
* 事务包含读、写事务：

所有的读事务不会锁定，读到的数据取决于snapshot设置。

写事务之间如果不存在记录交集，不会锁定。

写事务之间如果存在记录交集，此时如果未设置snapshot，则交集部分的记录是可以串行提交的。如果设置了snapshot，则第一个写事务(写锁队列的head)会成功，其他写事务会失败(之前的事务修改了该记录的情况下)。

### 独占写锁和写冲突

RocksDB事务写锁是基于Key Locking行锁的(实现上锁力度会粗一些)，所以在多个Transaction同时更新一条记录，会触发独占写锁定。如果还设置了snapshot的情况下，会触发写冲突分析。每个写操作(Put/Delete/Merge/GetForUpdate)开始之前，会进行写锁定，见TransactionLockMgr代码。如果存在记录有交集，写锁定会锁住一片key保证只有一个事物会独占写。

|  |  |
| --- | --- |
| image.png  图2 封锁机制 | Q9：数据经过TransactionLockMgr代码才会进入WriteBatch，即对于写写冲突会进行封锁，这违背了OCC的原则。  思考是否需要改进？怎么改进？  A9：目前RocksDB同时提供了乐观事务和悲观事务，TDStore是在乐观事务基础上改的，而TransactionLockMgr是悲观事务用到的。  Q9.1 这个回答是想表明图2中的机制，在TDStore的实现中，是被跳过去不采纳了，是不？ |

内部实现还是比较精炼的，全局有个LockMaps结构，里面按照ColumnFamily级别和num\_strips(默认16)级别做了shard进一步降低冲突(此处RocksDB还针对每个LockMap做了ThreadLocal优化)。最底层是一个ColumnFamily下某一个strip的LockMapStripe结构

Q10：如下内容，如果是所有key共用OS锁，则需要改进，如按range分配锁，或固定X个锁，自动划分等。优先级中。

struct LockMapStripe {

// 当下所有keys共用的os锁

std::shared\_ptr<TransactionDBMutex> stripe\_mutex;

std::shared\_ptr<TransactionDBCondVar> stripe\_cv;

// key -> 记录key， value -> 每个key对应的LockInfo结构

// map中所有的key共享上述os锁，作者这里提到了未来会有更细粒度的锁

// TODO(agiardullo): Explore performance of other data structures.

std::unordered\_map<std::string, LockInfo> keys;

};

struct LockInfo {

// 是否是独占锁(也可以是共享锁)

bool exclusive;

// 等待这个key的所有事务链表

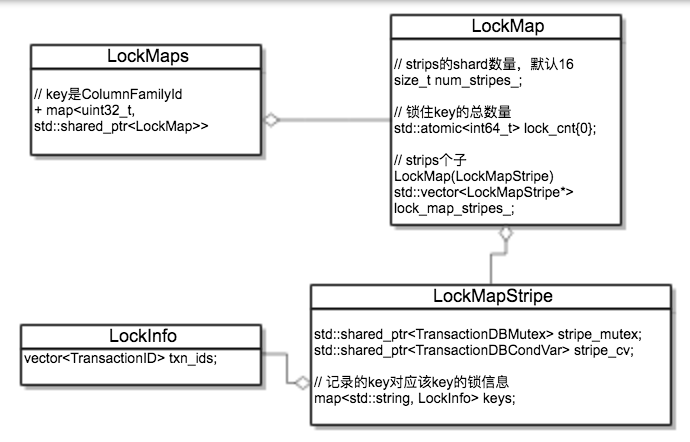
autovector<TransactionID> txn\_ids;

// 锁超时时间

uint64\_t expiration\_time;

};

关系图



针对每一个LockMapStripe里所有的key，有一个LockInfo(包含是否是排它锁，这个key挂的事务ID列表,超时时间)的map，所有落在这个map里的key如果存在并发写的情况，则会等待写锁释放。这里有个粒度问题，两个不相关的key如果落在同一个map里，也会等写锁。不如InnoDB的页锁冲突小，RocksDB作者在注释里提到之后会有更好的方案

加锁代码：

Status TransactionImpl::TryLock(ColumnFamilyHandle\* column\_family,

const Slice& key, bool read\_only,

bool exclusive, bool untracked) {

// tracked\_keys\_cf记录着当前事务中所有操作的key(涉及所有ColumnFamily)

auto iter = tracked\_keys\_cf->second.find(key\_str);

if (iter == tracked\_keys\_cf->second.end()) {

// 没找该key说明之前该事务之前一定没有独占锁定这个key

previously\_locked = false;

} else {

if (!iter->second.exclusive && exclusive) {

// 如果之前是共享锁，现在申请独占锁，则进行锁升级

lock\_upgrade = true;

}

previously\_locked = true;

current\_seqno = iter->second.seq;

}

if (!previously\_locked || lock\_upgrade) {

// 通过全局的LockMgr独占锁定该key(内部使用os锁)，如果没有其他事务操作该key(也可

// 能不同的key命中同一个LockMapStrip)，则TryLock理解返回并持有该key独占写锁。否则，

// TryLock需要等待其他事务释放该key的独占写锁，或者等待其他事务锁超时

s = txn\_db\_impl\_->TryLock(this, cfh\_id, key\_str, exclusive);

}

......

// 如果没有设置snapshot方式(可以通过创建事务的TransactionOptions指定snapshot或者

// 调用Transaction的SetSnapshot()方法)，则直接获取最新的lsn

if (untracked || snapshot\_ == nullptr) {

......

} else {

// 如果设置了snapshot，需要通过ValidateSnapshot判断是否有其他事务对该key进行了

// 更改(如该事务等待TryLock独占写锁时，其他获得了该锁的事务更新了该key)。具体实现

// 就是是在memtable，immemtable以及sst中取得该key最大的lsn对应的记录(通过

// DBImpl::GetLatestSequenceForKey)，看该lsn是否大于当前snapshot的lsn，

// 大于则写冲突。

if (s.ok()) {

s = ValidateSnapshot(column\_family, key, current\_seqno, &new\_seqno);

........

}

}

if (s.ok()) {

// 将当前key写入tracked\_keys\_cf

TrackKey(cfh\_id, key\_str, new\_seqno, read\_only, exclusive);

}

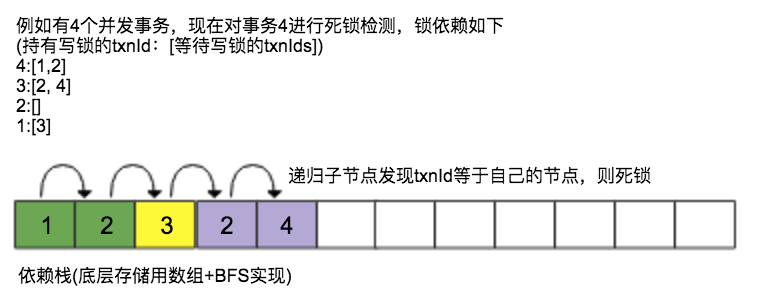
return s;

}

Q11：上述代码，优先级中。需要确认过程

### 死锁检测/超时

创建事务时 TransactionOptions.deadlock\_detect 选项可以支持死锁检测(默认不开启，性能影响较大，尤其是热点记录场景下。依赖timeout机制解决死锁)。如果多个事务之间发生死锁，则当前检测到死锁的事物失败(可以回滚)。死锁检测是通过刚才提到的LockInfo中全局事物ID列表以和当前事务ID进行环检测实现，通过广度优先递归遍历当前事务ID依赖的事务ID，判断其是否指向自己，如果能递归的找到自己的ID则说明有环，发生死锁。deadlock\_detect\_depth参数可以指定检测的深度，防止过深的依赖。



### Optimistic Transaction

相较于悲观锁，RocksDB也实现了一套乐观锁机制的OptimisticTransaction，接口上和Transaction是一致的。不过在写操作(Put/Delete/Merge/GetForUpdate)时，不会触发独占写锁和写冲突检测，而是在事务commit时("乐观"锁)，写入WAL时判断是否存在写冲突，而commit失败。这种方式的好处时，更新操作或者GetForUpdate()时，不用加独占写锁，省去了加锁的代价，乐观的认为没有写冲突，推迟到事务提交时一次性提交所有写入的key进行判断。

Q12：即与Q9不同，此是TDStore目前采用的方式，即没有经过TransactionLockMgr代码的代码。其流程，是否可以按前面锁相关的，写出来？

A12：乐观事务流程：

- 事务开始时，会传入一个开始时间戳start\_ts

-当事务写数据时，会将数据写入到内存中的WriteBatch里，WriteBatch保证事务之间的隔离性，写入到WriteBatch里的数据不会被其它事务读取

-当事务读取数据时：

 假如事务写过该数据项，则从WriteBatch中读取之前写过的版本

 否则，会选取commit\_ts＜当前事务start\_ts的事务进行读取。假如该数据项上，有正在提交的事务加了锁，那么需要等待锁释放之后再读取。

- 当事务（参与者）进入到2pc的prepare阶段时：

 先以key为单位，对所有写过的数据项加锁，假如某个数据项已经被其它事务加了锁，那么认为出现写写冲突，回滚当前事务 ---- Q12.1 确认此处是排序后加锁吧？如不是，有死锁可能。

 若加锁成功，则获取该数据项上的最新版本commit\_ts，若最新版本commit\_ts ＞ 当前事务start\_ts，也认为出现了写写冲突，回滚当前事务 ---- Q12.2 数据项上的commit\_ts记录在哪里？

 若未出现写写冲突，则将WriteBatch中的数据编码成日志，同步到raft中（数据被编码成redo log，同时后面还会附加上prepare log，prepare log相当于一个tag，表示全部数据都已经被编码为redo log）

* 当事务（参与者）进入到2pc的commit阶段时：

 将commit log同步到raft中

 将数据落盘到memtable中 ---- Q12.3 为什么是“落盘”到memtable中？是把一个事务的WriteBatch中的数据按key逐条按序拷贝到memtable中（有个排序merge的过程）？

 释放prepare阶段中对数据项加的锁

- 当事务（参与者）进入到2pc的clear阶段时（TDStore为实现无状态协调者所引入的新阶段）：

 将clear log同步到raft中 ---- Q12.4 有什么作用？

 释放事务上下文

## MVCC

RocksDB实现的ReadCommited和RepeatableRead隔离级别，类似其他数据库引擎，都使用MVCC机制。例如MySQL的InnoDB，通过undo page实现了行记录的多版本，这样可以在不同的隔离级别下，看到不同时刻的行记录内容。不过undo需要undo页的存储空间以及redo日志的保护(redo写undo)，这跟其btree的in-place update有关，而RocksDB依靠其天然的AppendOnly，所有的写操作都是后期merge，自然地就是key的多版本(不同版本可能位于memtable,immemtable,sst)，所以RocksDB首先MVCC是很容易的，只需要通过snapshot(lsn)稍加限制即可实现。

例如需要读取比某个lsn小的历史版本，只需要在读取时指定一个带有这个lsn的snapshot，即可读到历史版本。所以，在需要一致性非锁定读读取操作时，默认ReadCommited只需要按照当前系统中最大的lsn读取(这个也是默认DB::Get()的行为)，即可读到已经提交的最新记录(提交到memtable后的记录一定是已经commit的记录，未commit之前记录保存在transaction的临时buffer里)。在RepeatableRead下读数据是，需要指定该事务的读上界(即创建事务时的snapshot(lsn)或通过SetSnapshot指定的当时的lsn)，已提交的数据一定大于该snapshot(lsn)，即可实现可重复读。

txn = txn\_db->BeginTransaction(write\_options);

// ReadCommited (default)

txn->Get(read\_options, "abc", &value);

txn = txn\_db->BeginTransaction(write\_options, txn\_options);

txn\_options.set\_snapshot = true;

// RepeatableRead

read\_options.snapshot = txn->GetSnapshot();

s = txn->Get(read\_options, "abc", &value);

可见snapshot对于MVCC有着很重要的意义：

1、snapshot可以实现不同隔离级别的非锁定读

2、snapshot可以用于写冲突检测

3、snapshot由全局的snapshot链表进行管理，在compaction时，会保留该链表中snapshot不被回收

Q13：为什么“snapshot由全局的snapshot链表进行管理”？类似Q5，但这里在问：全局的snapshot链表存在的价值在哪里？

A13：Snapshot的价值体现在compaction过程中，中间版本的数据不会被回收。这里举个例子：

- memtable中某key上目前有两个版本，分别为1时刻写的“a”、3时刻写的“b”。

- 在4时刻，某读事务T1获取Snapshot

- 在5时刻，某写事务T2写“c”，并提交

- 随后，rocksdb对memtable中的数据进行compaction。

Compaction操作会将数据项上的多个版本合并成一个，下沉到immutable memtable或sst中。这里，假如compaction将中间版本"b"会回收，只留下最新版本“c”，那么事务T1将无法读取到正确的版本。

所以，为了避免这种情况，rocksdb在对某数据项进行compaction时，会依靠snapshot链表，找到时间戳最早的snapshot，仅对该snapshot可见的版本之前的历史版本进行合并。在该例中，rocksdb仅会将1时刻写的“a”回收，不会回收“b”，保证事务T1读数据的正确性。

Q13.1如果目的只是为了找到最旧的SI，则记录最旧的SI即可，为什么要用双向的链表数据结构？ 双向的链表是有序的吧？最旧且已经完成的事务的SI，才可被从双向的链表中淘汰吧？能否一次淘汰最旧的n个SI？

## DQL、DML操作

### 查询操作

前述了数据可见性判断的方式，不再单独表述。

### 更新操作

### 插入操作

### 删除操作

CRDB实现方式：Tombstone：一种特殊的数据版本，Value为空，事务读到Tombstone版本则意味着当前记录此时已被删除

## 2PC两阶段提交

Q14：TDStore的2PC与原生2PC的差异在哪里？

A14：差异主要体现在：

- TDStore将原生的WAL去掉了，改成同步Raft Log，以支持多副本

- 原生2PC只提供了参与者的Prepare和Commit阶段逻辑，而TDStore在此基础上添加了协调者的逻辑：

 实现了协调者和参与者之间的通信

 TDStore的协调者是无状态的，为了实现协调者宕机恢复，TDStore为每个参与者添加了Clear阶段，保证了协调者的状态是可恢复的

 实现了异步提交机制

RocksDB除了实现了基本类型的事务，还实现了2pc(https://github.com/facebook/rocksdb/wiki/Two-Phase-Commit-Implementation。某种程度上看，需求来自于MySQL的MyRocks引擎，binlog和引擎日志(redolog、wal)有一个XA的约束，防止出现写一个日志成功，另一个失败的情况。所以需要引擎日志实现2pc来支持binlog和引擎日志的原子提交。

详细文档可参见：

<https://github.com/facebook/rocksdb/wiki/Two-Phase-Commit-Implementation>

两阶段提交在原有的Transaction基础之上，在写记录和commit之间增加了一个Prepare操作：

BeginTransaction;

Put()

Delete()

.....

Prepare(xid)

Commit(xid) // or Rollback(xid)

### 2PC实现原理

前面几个步骤和普通的Transaction基本都是一直的，主要是后面Prepare和Commit有所区别。首先，2pc的事务有一个全局的事务表，所有2pc的事务都要有一个name，在设置name的同时，将该事务注册到全局事务表里：

Status TransactionImpl::SetName(const TransactionName& name) {

if (txn\_state\_ == STARTED) {

......

// 向事务管理器注册事务

txn\_db\_impl\_->RegisterTransaction(this);

......

}

* prepare阶段

// 设置事务状态为开始PREPARE

txn\_state\_.store(AWAITING\_PREPARE);

// PREPARE之后不允许事务超时, 可能会遇到2pc的通病????

expiration\_time\_ = 0;

WriteOptions write\_options = write\_options\_;

write\_options.disableWAL = false;

// MarkEndPrepare会将当前batch开头和结尾写入PREPARE标记

// 正常的WriteBatch格式一般是:

// Sequence(0);NumRecords(2);Put(a,1);Delete(b);

// MarkEndPrepare之后:

// Sequence(0);NumRecords(4);BeginPrepare();Put(a,1);Delete(b);EndPrepare(transaction\_id);

// 对WriteBatch开始和结束分别加入Begin/End，标识是个PREPARE

WriteBatchInternal::MarkEndPrepare(GetWriteBatch()->GetWriteBatch(), name\_);

// 将更改之后的WriteBatch写入db，这里只写WAL，不写memtable

s = db\_impl\_->WriteImpl(write\_options, GetWriteBatch()->GetWriteBatch(),

/\*callback\*/ nullptr, &log\_number\_, /\*log ref\*/ 0,

/\* disable\_memtable\*/ true);

if (s.ok()) {

.......

txn\_state\_.store(PREPARED);

}

整个过程将修正后的prepared writebatch只是写入WAL日志，并不会更新memtable，这样保证了其他的普通事务和2pc事务是不能访问到该2pc事务的记录(memtable不可见)，保证了隔离性。这里有个点需要注意，大部分RocksDB的写操作都是一定写memtable和WAL(可以disable)的，所以全局的LSN就会递增。但prepare步骤是不写入memtable的，所以LSN不会增加，这就解释了文章开头说的WAL中LSN并不一定满足lsn(WriteBatch(n)) < lsn(WriteBatch(n+1))。

* commit阶段

Q15：为什么“这也就是说2pc是多个WriteBatch”？ 为什么2pc会有多个WriteBatch？不能就使用OCC操作阶段的WriteBatch作为提交时的WriteBatch吗？

A15：OCC操作阶段的WriteBatch中包含事务所写入的数据，以及prepare标识，这些数据在prepare阶段写入WAL。而这里说的“临时”WriteBatch，是在commit阶段，又向WAL写入的commit log，表示事务的状态已经变更为“提交”。事务提交过程中，写WAL是分为多次进行的，所以用到多个WriteBatch Q15.1 WriteBatch既可以放用户数据，也可以放WAL，只是一块私有内存，是不是？

// 设置事务状态为准备commit

txn\_state\_.store(AWAITING\_COMMIT);

// 获取临时的一个WriteBatch buffer，区别于prepare之前的操作的WriteBatch

// 所以commit的WriteBatch和prepare的WriteBatch是单独分开的，这也就是说2pc

// 是多个WriteBatch所以需要额外保证原子性。

WriteBatch\* working\_batch = GetCommitTimeWriteBatch();

// 写入commit标识和事务ID

WriteBatchInternal::MarkCommit(working\_batch, name\_);

// WAL终止点(暂没想到更好的叫法)，后续写入的数据，WAL会全部忽略

working\_batch->MarkWalTerminationPoint();

// 将包含prepare的全部数据追加到WriteBatch里，这些数据是供memtable写入用的

WriteBatchInternal::Append(working\_batch, GetWriteBatch()->GetWriteBatch());

// 数据写入memtable(包含prepare)，并将commit事件写入WAL

s = db\_impl\_->WriteImpl(write\_options\_, working\_batch, nullptr, nullptr,

log\_number\_);

if (!s.ok()) {

return s;

}

// 从全局事务表里删除该事务

txn\_db\_impl\_->UnregisterTransaction(this);

commit阶段主要做两件事：

1、将commit标识写入WAL

2、将数据写入memtable(让其他事务可以访问到)---特别注意：事务的数据从此可见了。相当于事务的提交标志

整体回顾整个2pc提交的流程，prepare阶段生成BeginPrepare/EndPrepare相关的WAL记录，并写入WAL持久化(这里可以防止crash时，仍旧可以构建出来该事务)，但为了保证隔离性，不会写入memtable。commit阶段将Commit的WAL记录写入WAL，并写入memtable，让其他事务可见。这里用了多个WriteBatch，打破了RocksDB默认的单WriteBatch原子性的保证，所以需要在WAL记录中增加额外标识，并在crash时，重建内存2pc事务状态。

### 2PC Recovery

RocksDB的2pc是跨WriteBatch实现的prepare和commit，所以可能存在中间态，比如prepare之后commit之前crash了。这时候系统启动时要重建所有的正在执行的事务(仅2pc事务，普通事务通过单个WriteBatch已经保证了原子性)。MemtableInserter作为处理WriteBatch中每一条记录，在遇到BeginPrepare/EndPrepare时，会在内存中重建事务的上下文，具体可见MemtableInserter代码本文不赘述。

## MyRocks

RocksDB的TransactionDB支持了大部分MySQL对事务的规范，整体接口形式和行为基本一致，有些细节比如online ddl、gap locking的支持、需要binglog开启row模式等有差别。

具体可见 <https://github.com/facebook/mysql-5.6/wiki>

# 应用场景

参考：

<https://www.jianshu.com/p/3302be5542c7>

RocksDB的典型场景（低延时访问）:

1. 需要存储用户的查阅历史记录和网站用户的应用
2. 需要快速访问数据的垃圾检测应用
3. 需要实时scan数据集的图搜索query
4. 需要实时请求Hadoop的应用
5. 支持大量写和删除操作的消息队列