Seqdb存储引擎

DBA 王剑英

2013.4.20

Agenda

- 业务的需求
- 现有引擎的性能
- 改进方案
- Seqdb引擎架构
- Seqdb引擎的性能
- 部署情况及经验

业务需求（1）

- 场景（1）
  数据量100GB ， 读20W /S ,写200/S.

- 场景（2）
  数据量1TB ， 主键随机读写， 写3W/S ,读1W/S.

- 场景（3）
  展示类，数据量500GB ,读3000/S ,写1000/S. SLA要求 100ms以下99.99%.

服务器配置：

CPU：2*4核心.

内存：48GB ，64GB，128GB可选.

存储：(1) HDD 容量1.5T 300 IOPS, 访问延迟5ms.

(2) SSD 800GB 3.2万IOPS ， 访问延迟50us.

业务需求（2）

- 场景（1）
  数据量100GB ， 读20W /S ,写200/S.
  4 cache server + 1 MySQL（HDD）

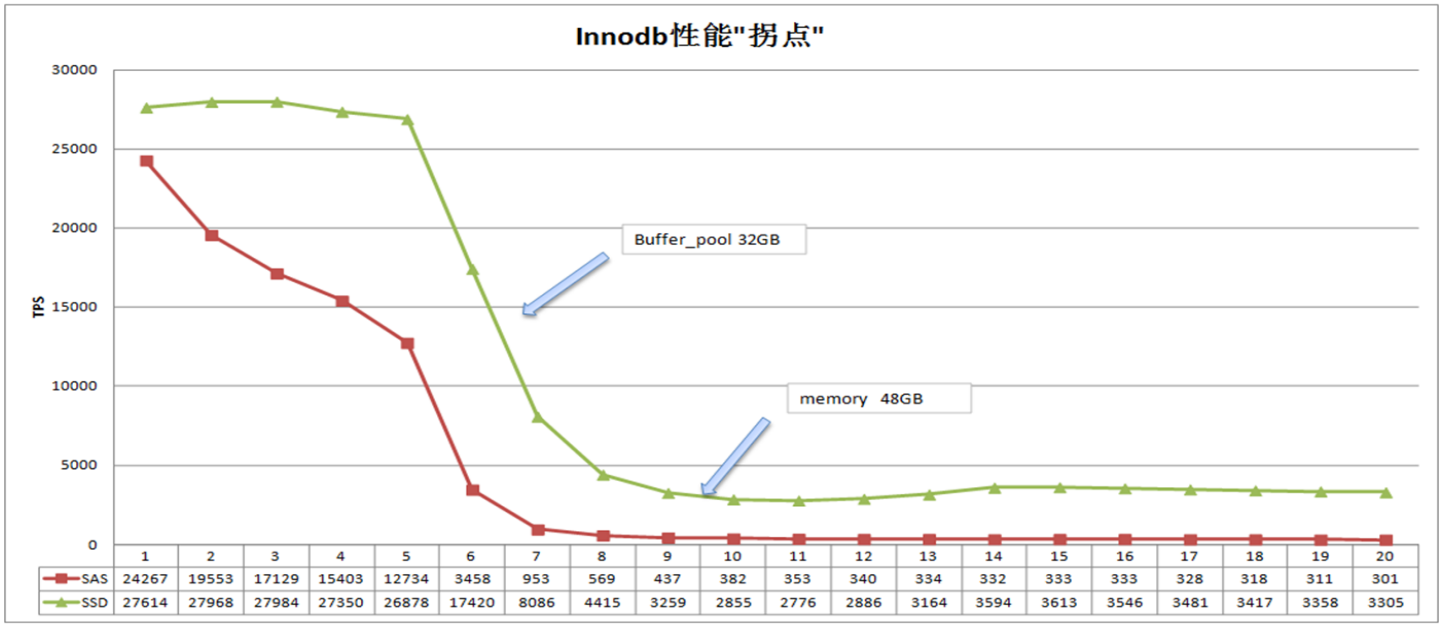
- 场景（2）
  数据量1TB ， 主键随机读写，写3W/S 读1W/S.
  MySQL(Innodb)+SSD ?
  MySQ (memcached plugin，handler socket）+SSD?

- 场景（3）
  数据量500GB ,读3000/S ,写1000/S.KV类访问, SLA要求 100ms以下 99.99%.
  MySQL(Innodb)+SSD ?

Agenda

- 业务的需求
- 现有引擎的性能
- 改进方案
- **Seqdb**引擎架构
- **Seqdb**引擎的性能
- 部署情况及经验

innodb的性能“拐点”



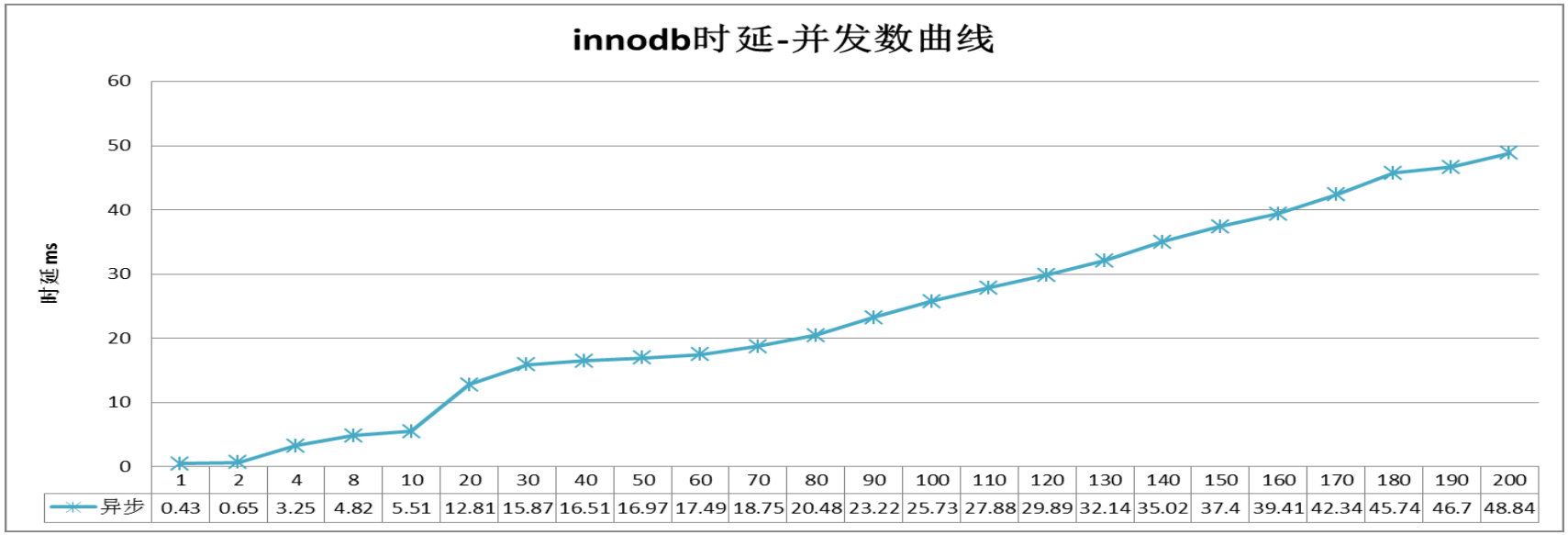
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
SAS	24267	19553	17129	15403	12734	3458	953	569	437	382	353	340	334	332	333	333	328	318	311	301
SSD	27614	27968	27984	27350	26878	17420	8086	4415	3259	2855	2776	2886	3164	3594	3613	3546	3481	3417	3358	3305

2 颗四核心CPU，48GB内存. intel 160GB MLC*6 RAID5 官方数据3.2万 IOPS. 单行长度512byte.数据5GB到100GB,update性能

1. 数据量超过内存，IO bound时，依赖IOPS.（handler socket？ ）
2. Innodb不能发挥高性能IO设备的潜能.

满足场景（2）需要10个MySQL cluster（Innodb+SSD）.

Innodb的时延曲线



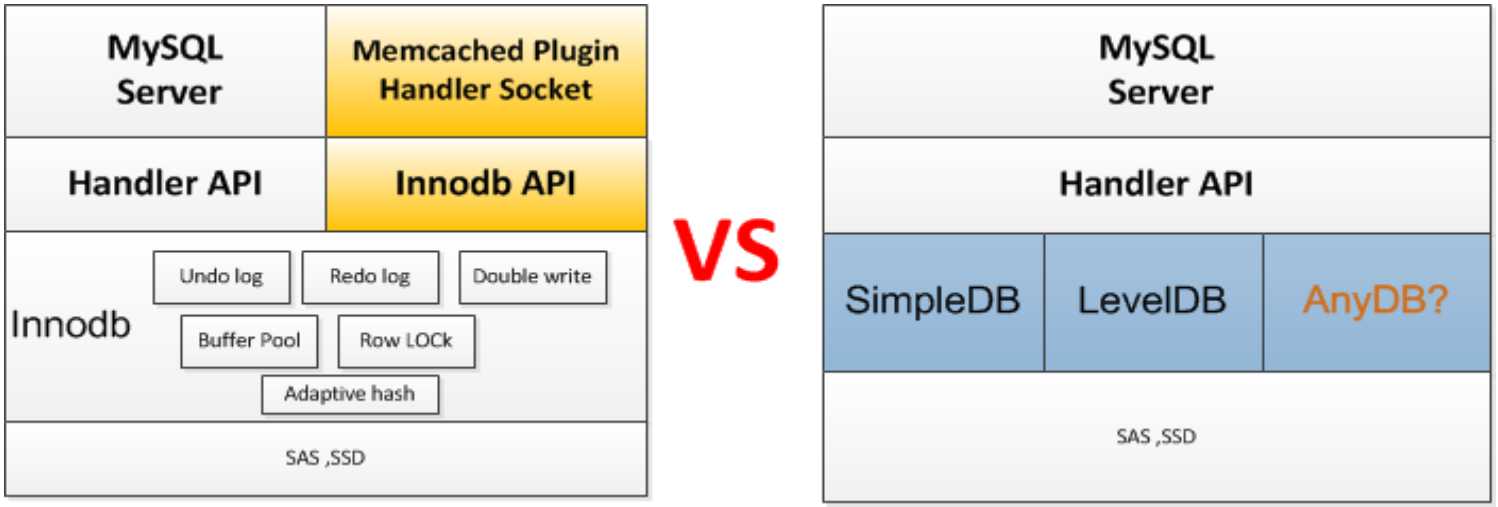
	1	2	4	8	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
异步	0.43	0.65	3.25	4.82	5.51	12.81	15.87	16.51	16.97	17.49	18.75	20.48	23.22	25.73	27.88	29.89	32.14	35.02	37.4	39.41	42.34	45.74	46.7	48.84

innodb不可能达到 100ms以下99.99%的SLA，达不到场景（3）的要求.

Agenda

- 业务的需求
- 现有引擎的性能
- 改进方案
- Seqdb引擎架构
- Seqdb引擎的性能
- 部署情况及经验

方案选择



目标：

（1）完全发挥SSD潜能， QPS>=IOPS,低延迟.

（2）SQL接口，经验可复用.

功能选择。

1. 只支持基于主键的增删改查，可映射为存储层的KV操作.

```
select * from tb where  col1=XXX_key;
insert into values (),();
update  ***  where col1=XXX_key;
delete from tb where col1=XXX_key;
```

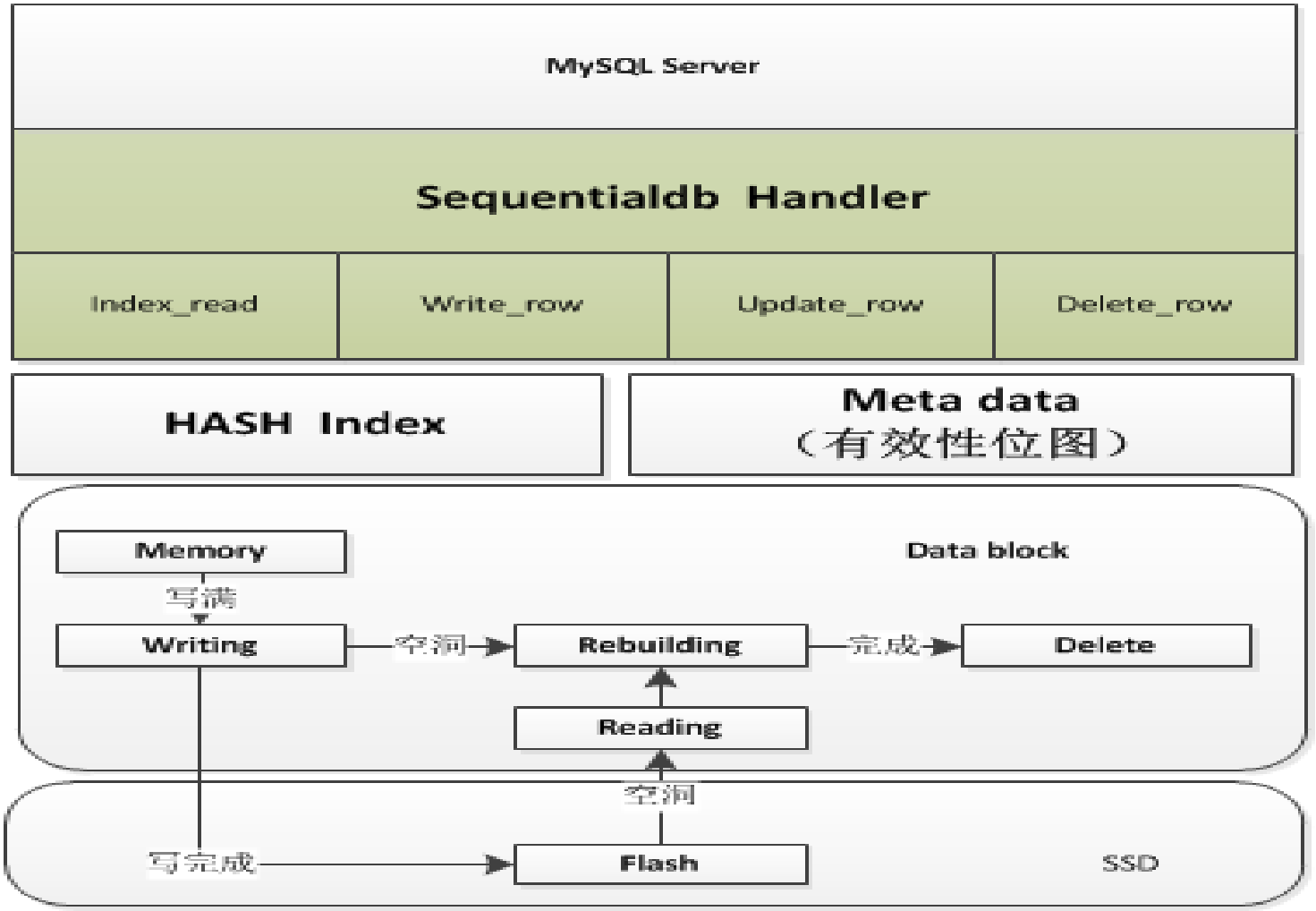
2. 放弃join，range，group by等复杂查询功能，不做全功能引擎，简化复杂度.

20%的投入，解决80%的需求.

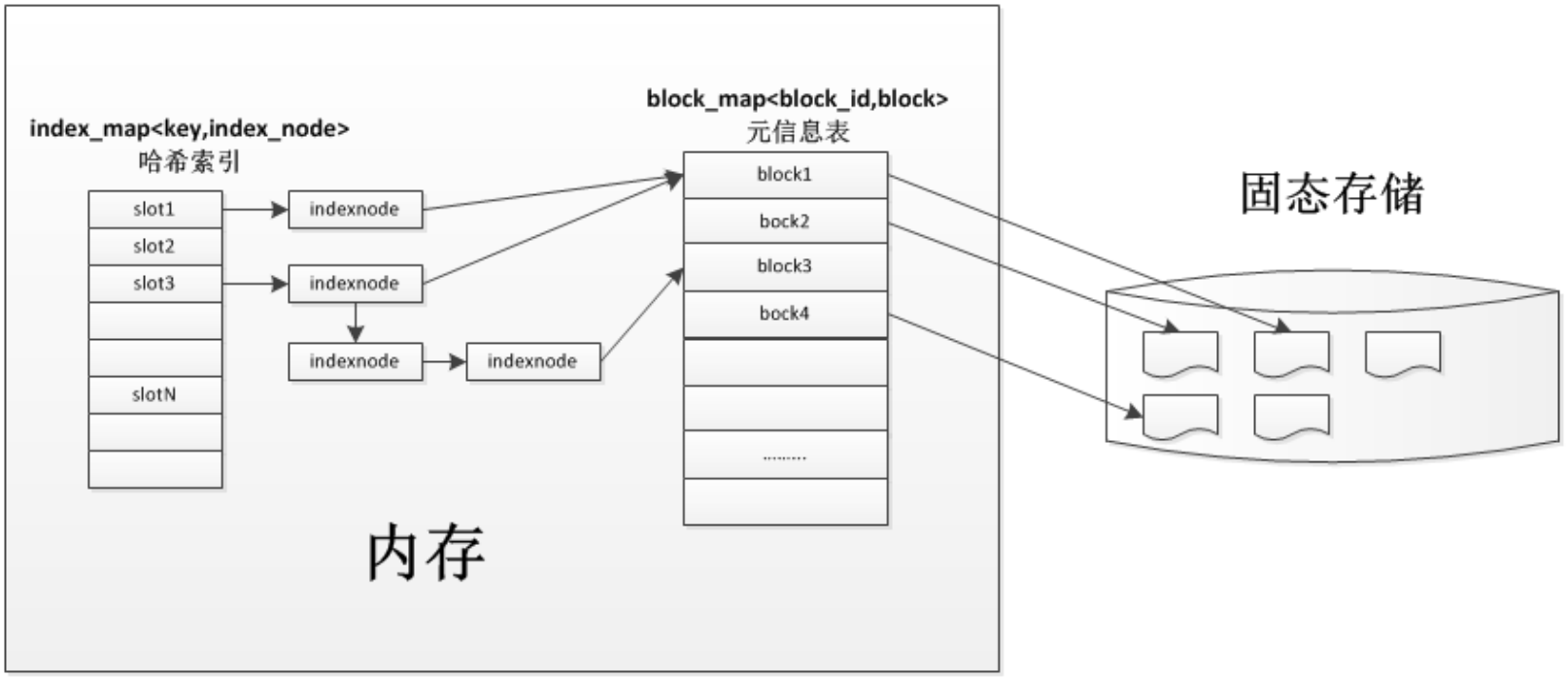
Agenda

- 业务的需求
- 现有引擎的性能
- 改进方案
- **Seqdb引擎架构**
- **Seqdb引擎的性能**
- 部署情况及经验

Sequentialdb引擎：总体架构

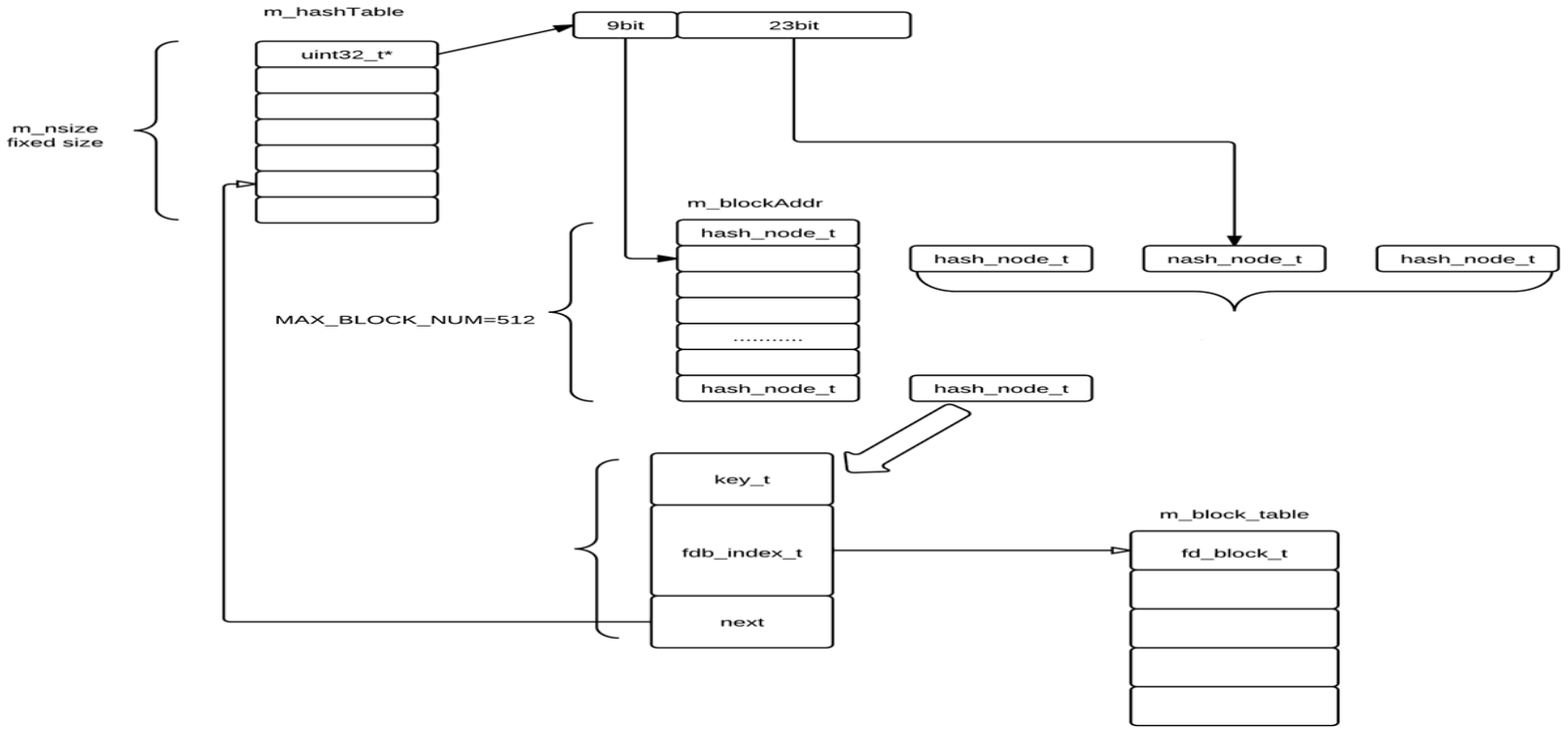


Sequentialdb引擎：访问模型



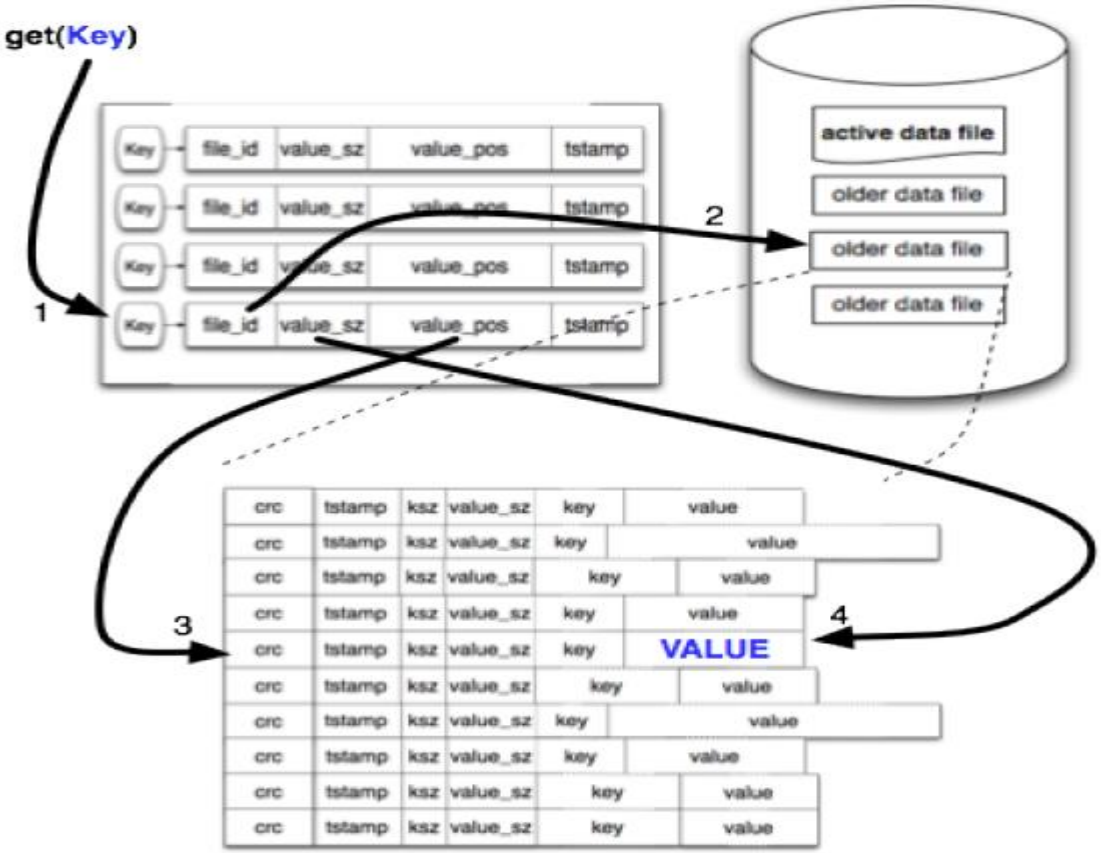
1. key求签名查找哈希表，索引节点冲突拉链，顺外链指针遍历。
2. 依据索引节点信息得到: a.block有效性位图 b.磁盘文件及偏移

Sequentialdb引擎：索引结构



1. 索引使用一个高效的hash容器，4字节间接寻址。HanoNode存放于HashNodeBlock。
2. HashNode包含Kev-value和外链指针，索引Node的分配和删除通过外链指针进行。

Sequentialdb引擎：数据查找



Sequentialdb引擎：垃圾回收

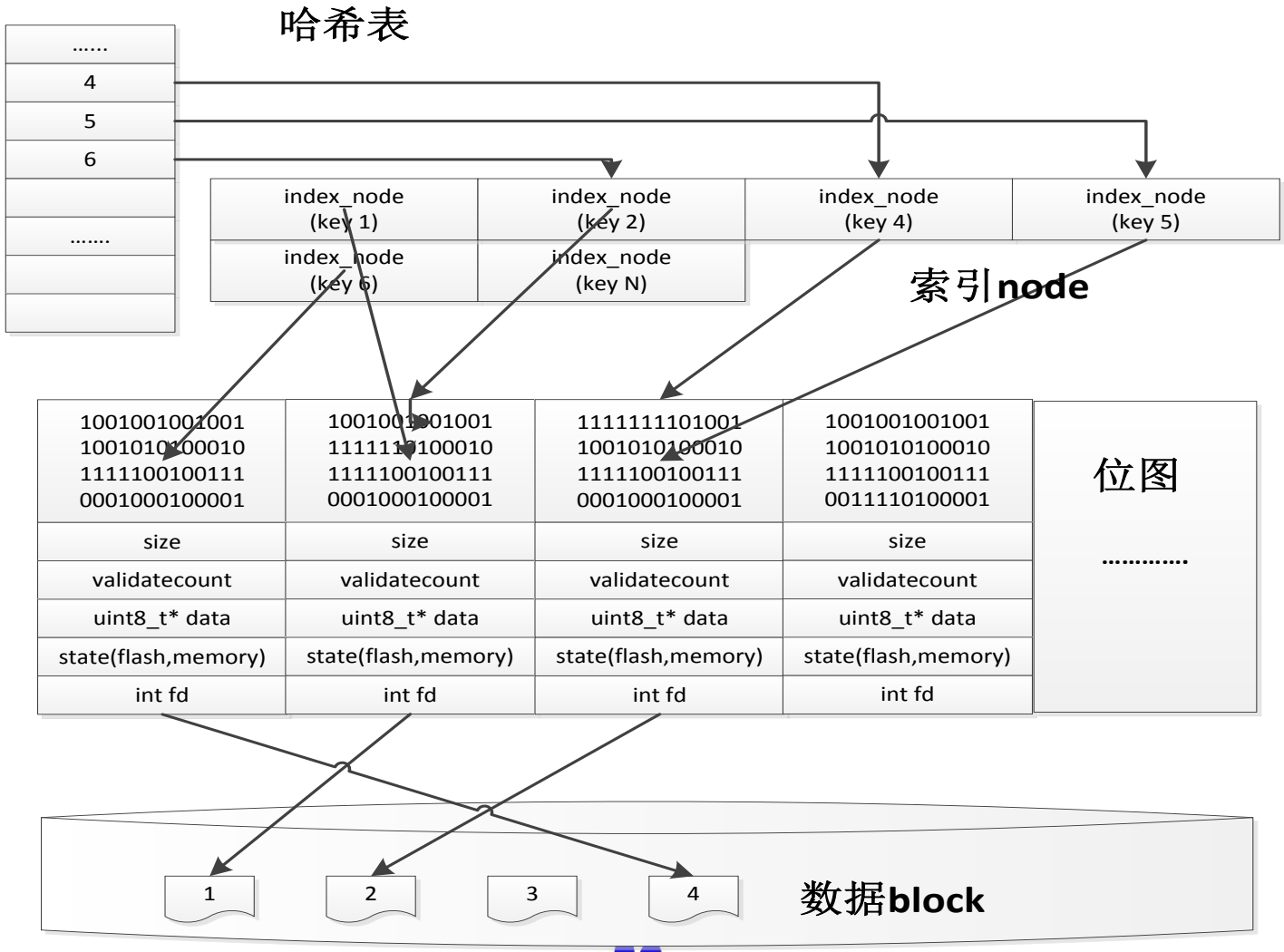
(1).block的使用率=有效数据条目/总数据条目

(2).删除一条数据时校验block的有效数目比例，降到一定百分比时进行rebuild。

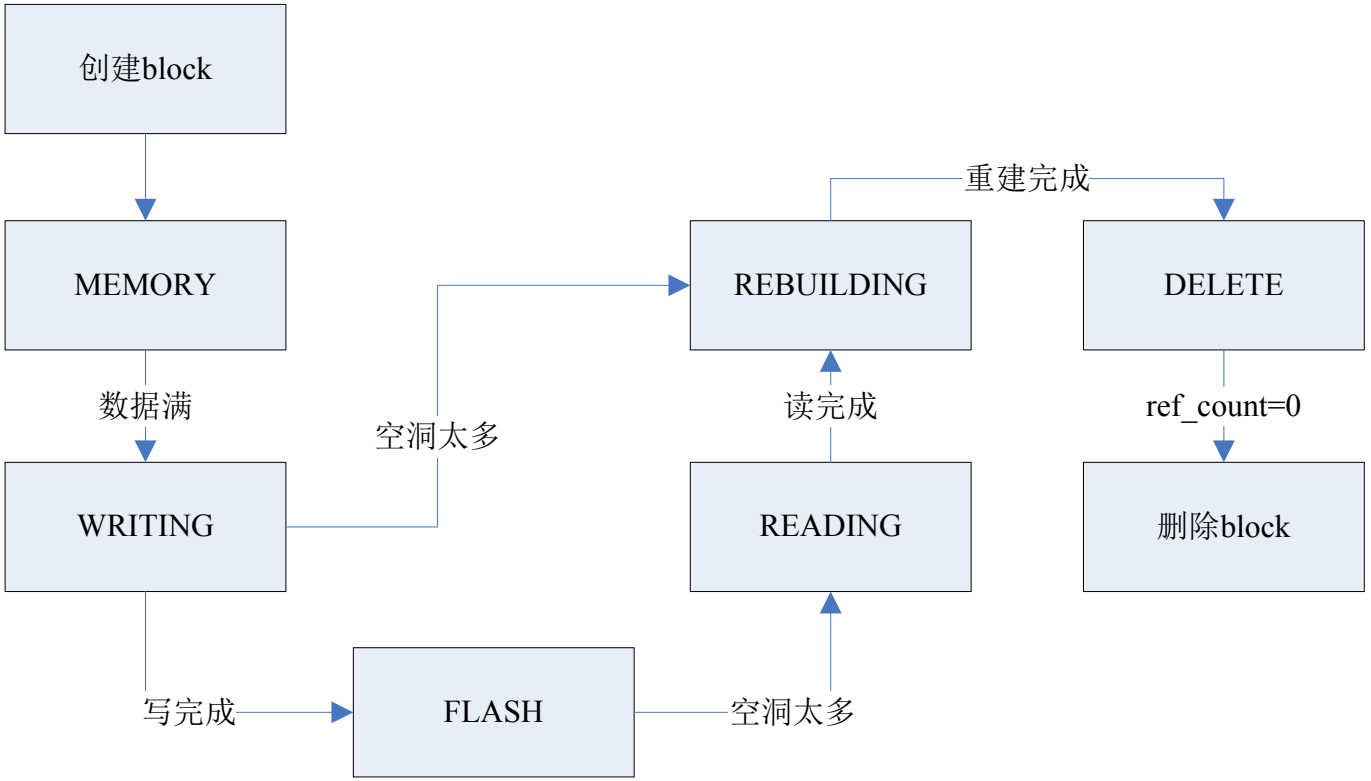
(3).被删除block中的数据回写入当前block。

(4).四个后台worker线程负责写block及垃圾回收。

Sequentialdb引擎：元信息位图



Sequentialdb引擎：BLOCK管理

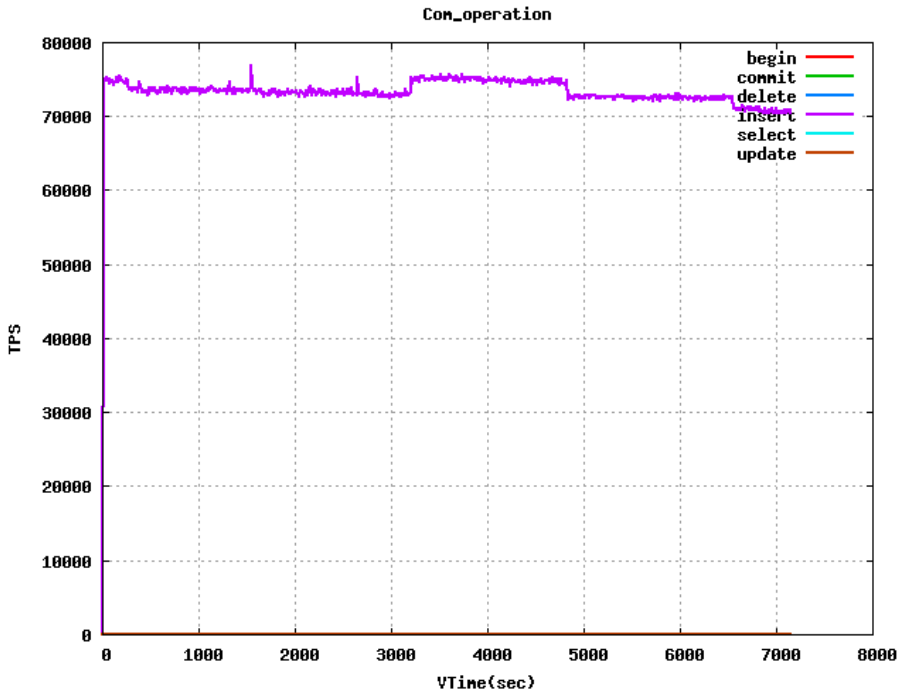


Agenda

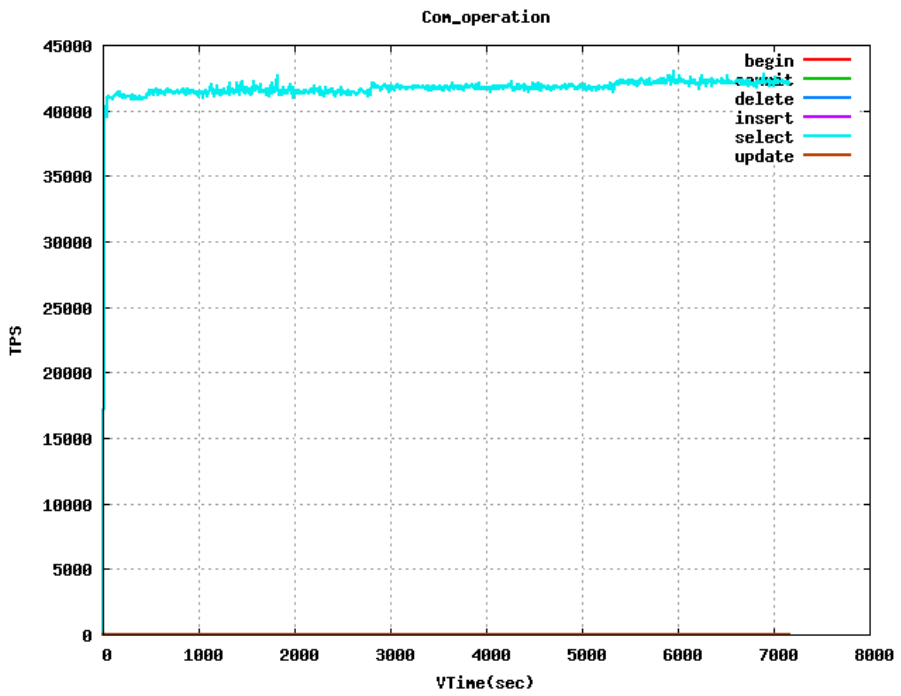
- 业务的需求
- 现有引擎的性能
- 改进方案
- Seqdb引擎架构
- Seqdb引擎的性能
- 部署情况及经验

Sequentialdb引擎：性能

insert



update

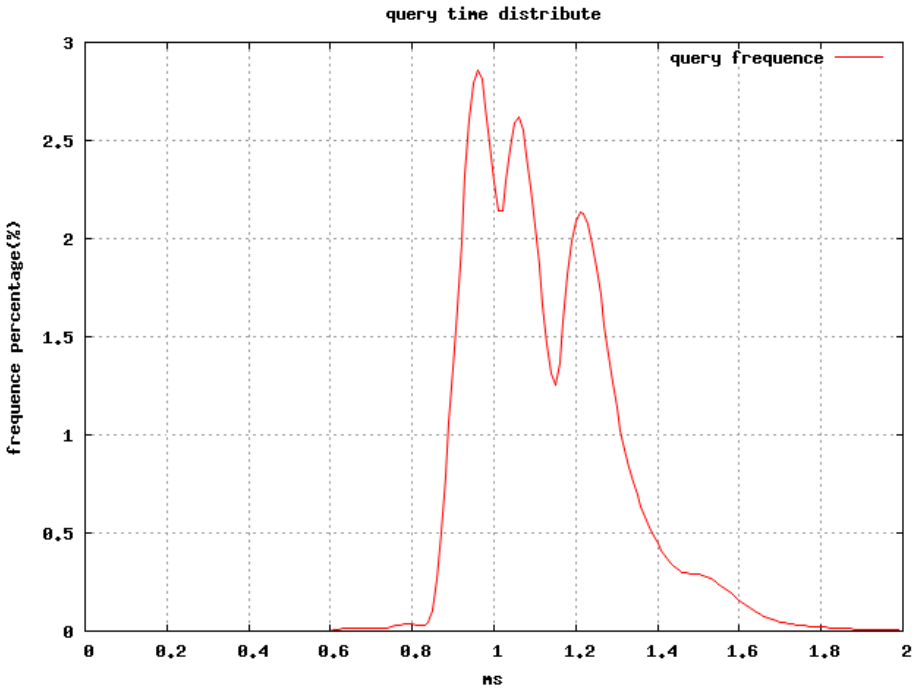


数据量5亿,单行长度200bytes,30 C线程 随机主键访问.

场景（2）问题解决,2个Seqdb的cluster能hold住(4台SSD套餐机型）

Sequentialdb引擎：时延

➢ 访问时延



环境：
dbproxy+MySQL（主+从，跨机房部署）线上评估测试
读写比 20:1

长耗时统计（4000万次查询中）
超过5ms的查询 4784次
超过10ms的查询 2421次
超过50ms的查询 448次
超过100ms的查询 282次
200ms以上为失败

场景3，问题解决，一个seqdb cluster.（2台SSD套餐机型）

Agenda

- 业务的需求
- 现有引擎的性能
- 改进方案
- Seqdb引擎架构
- Seqdb引擎的性能
- 部署情况及经验

Sequentialdb引擎：线上部署

- 产品线

1. 在线展示。
2. 离线计算，结合分布式数据库ddbs.

- 集群规模

MySQL实例 1000+.
单应用数据量百亿级.

- 性能

吞吐导向 -> 单集群 300W TPS.
时延导向 -> 响应水平 99.99% <100ms 95% < 2ms.

Sequentialdb引擎：经验

- 索引常驻内存安全吗？
  一主多从，内存的安全性是足够的。索引放在磁盘上访问成本太高
  稳定运行一年无维护操作。

- 单机故障
  SSD损坏（月概率1‰以下，远低于HDD），数据损毁。

- 时延敏感型应用，瓶颈在写binlog。
  （1）调整sync_binlog。
  （2）Binlog迁移到SSD盘。

- 毫秒级别超时API


QA

百度技术沙龙

关注我们：t.baidu-tech.com

资料下载和详细介绍：infoq.com/cn/zones/baidu-salon

“畅想•交流•争鸣•聚会”是百度技术沙龙的宗旨。 百度技术沙龙是由百度与InfoQ中文站定期组织的线下技术交流活动。目的是让中高端技术人员有一个相对自由的思想交流和交友沟通的的平台。主要分讲师分享和OpenSpace两个关键环节，每期只关注一个焦点话题。

讲师分享和现场Q&A让大家了解百度和其他知名网站技术支持的先进实践经验，OpenSpace环节是百度技术沙龙主题的升华和展开，提供一个自由交流的平台。针对当期主题，参与者人人都可以发起话题，展开讨论。

InfoQ 策划·组织·实施

关注我们：weibo.com/infoqchina