2019
05
08-10
北京新云南皇冠假日酒店

数据风云十年变迁

DTCC 第十届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2019

IT168 ChinaUnix ITPUB

数据库压缩技术在百度网盘的应用

百度数据库架构师：陈辉
百度数据库资深运维工程师：高佳

大纲

- 百度网盘简介
- 问题与挑战
- 阶段一：InnoDB压缩（100G压缩到60G）（2014~2015）
- 阶段二：TokuDB压缩（100G压缩到35G）（2016~2018）
- 阶段三：下一步（MyRocks & 冷热分离）（2018~2019）

百度网盘发展历史

萌芽探索期 → 快速发展期 → 瓶颈洗牌期 → 成熟理性期

2005

- 2005, GmailDrive，网易邮箱个人文件夹功能成为云盘的雏形；
- 2009年，Dropbox用户数突破百万，华为网盘、115网盘等产品出现。

2012

- 各厂商纷纷进入个人云盘领域，竞争日趋白热化。
- **百度网盘发布。**

2016

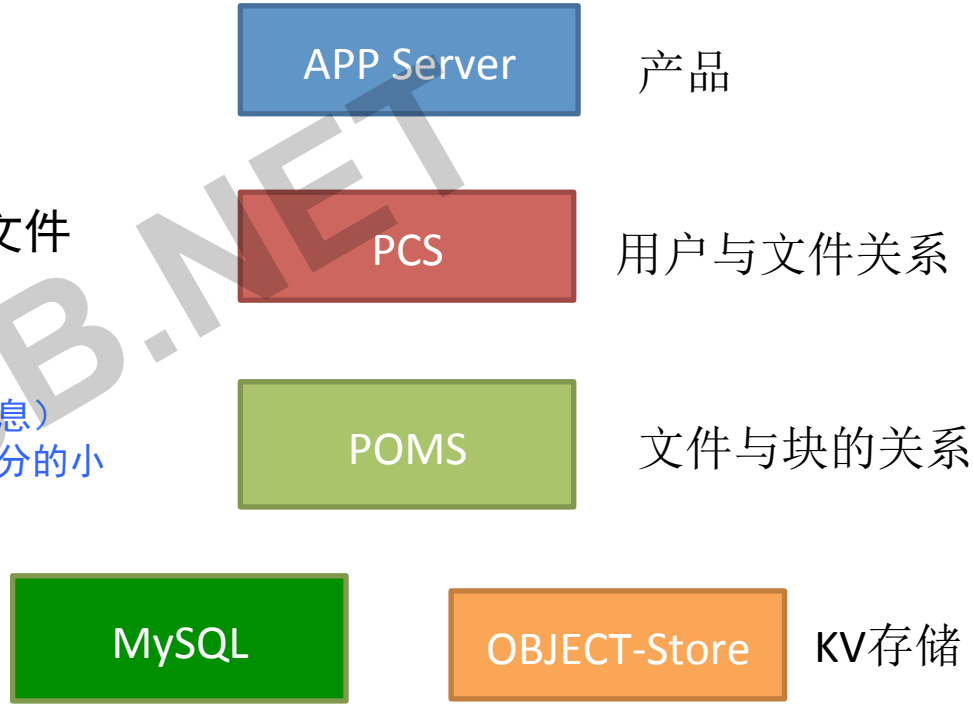
- 行业政策监管趋严，个人云盘盈利困难，包括360等多家服务商关停网盘服务。
- 百度云坚持下来，大量用户转到百度云。

2017

- 2017年，行业洗牌逐渐完成，商业模式形成。
- 百度网盘（6亿用户）、DropBox（5亿用户）。

百度网盘架构介绍

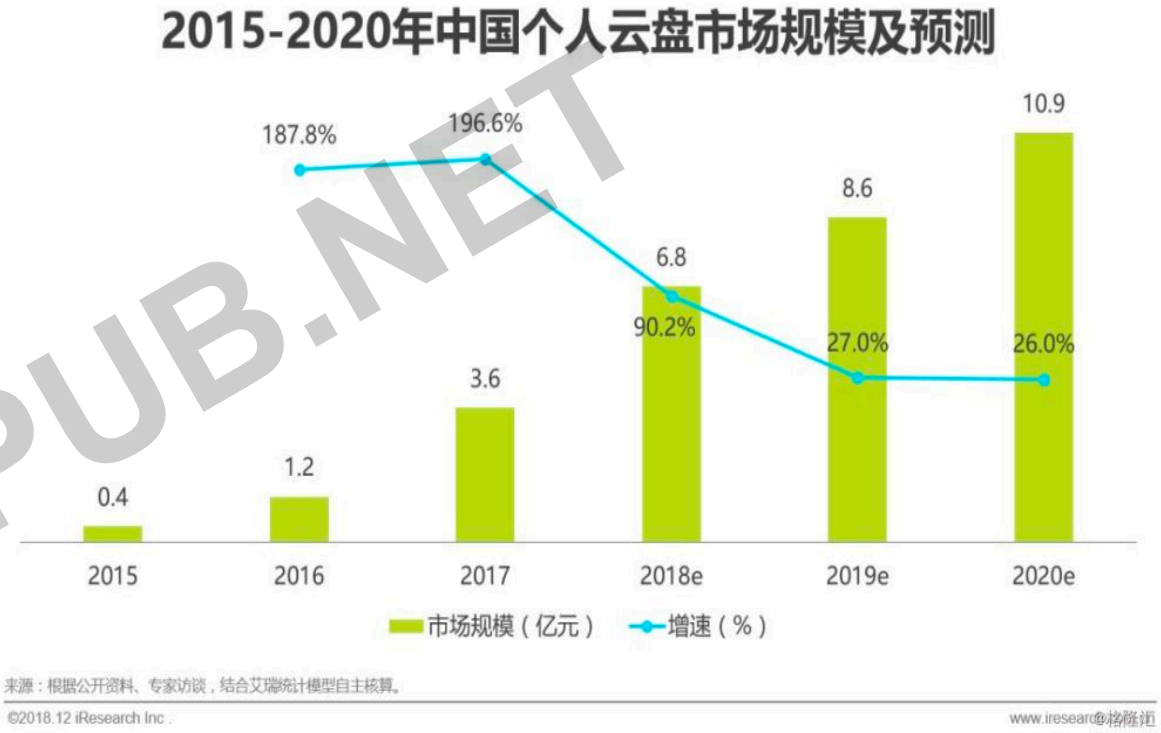
- PCS：管理用户和文件
- POMS： 管理文件和块的关系
- OBJECT：分布式KV存储，存储最终的文件块。
- MySQL：
  - 1. 存储用户目录信息（用户与文件的MAP信息）
  - 2. 存储文件拆分后的索引信息（按照4MB拆分的小文件）
- 为什么要用MySQL？
  - 需要SQL接口
  - 大量order by, group by, distinct等



MySQL问题和挑战

问题与挑战：

- 数据量大
  - 整个网盘估算：10G*6亿 = 10G * 600,000,000 = 6ZB??
  - MySQL数据库PB级别
- MySQL集群规模大
  - 千级别机器和实例
- MySQL磁盘利用率高
  - 达到60%左右
- 不断快速增长！



大纲

- 百度网盘简介
- 问题与挑战
- 阶段一：InnoDB压缩（100G压缩到60G）（2014~2015）
- 阶段二：TokuDB压缩（100G压缩到35G）（2016~2018）
- 阶段三：下一步（MyRocks & 冷热分离）（2018~2019）

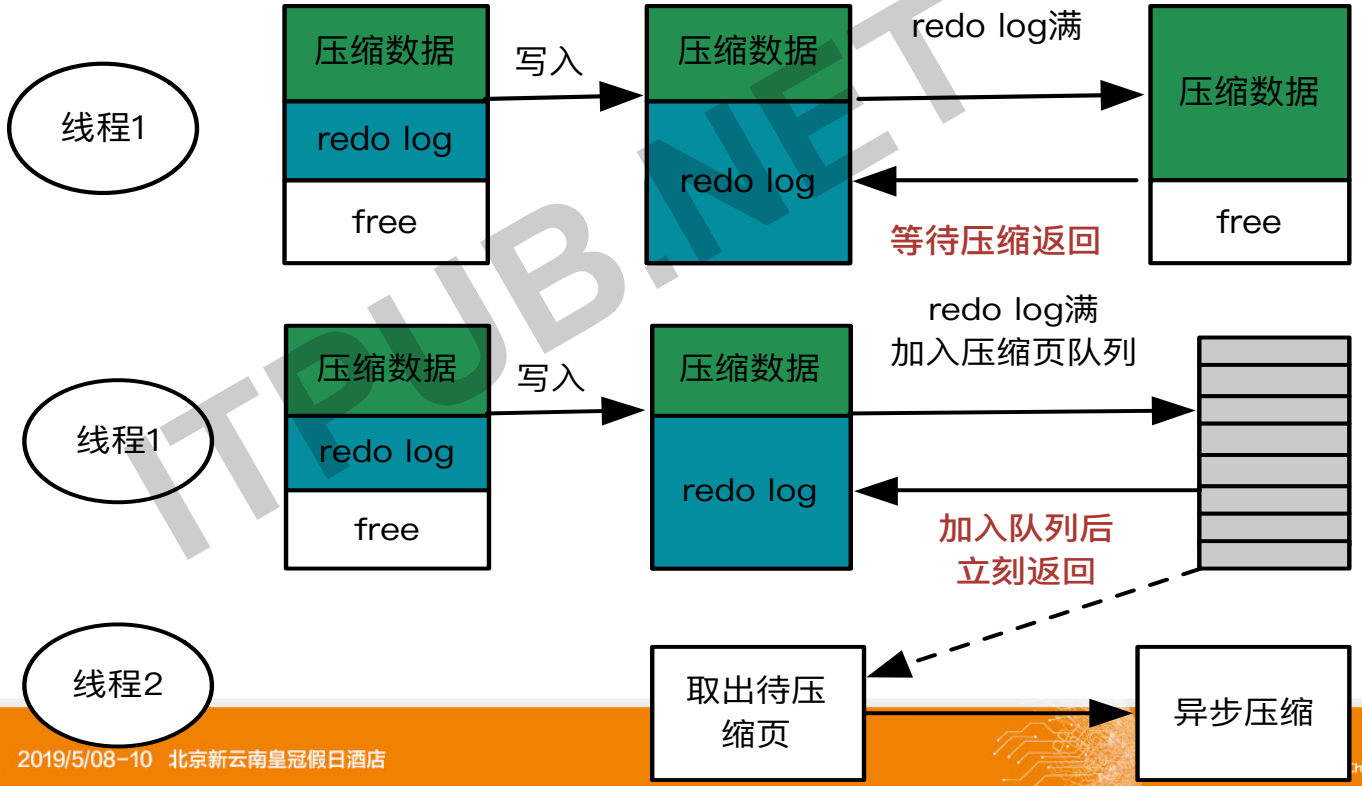
阶段一：InnoDB压缩

社区版压缩功能不能直接使用。

- 问题1：压缩后写性能是未开启压缩性能的1/3，业务无法接受。
- 问题2：压缩后读性能没有提升。

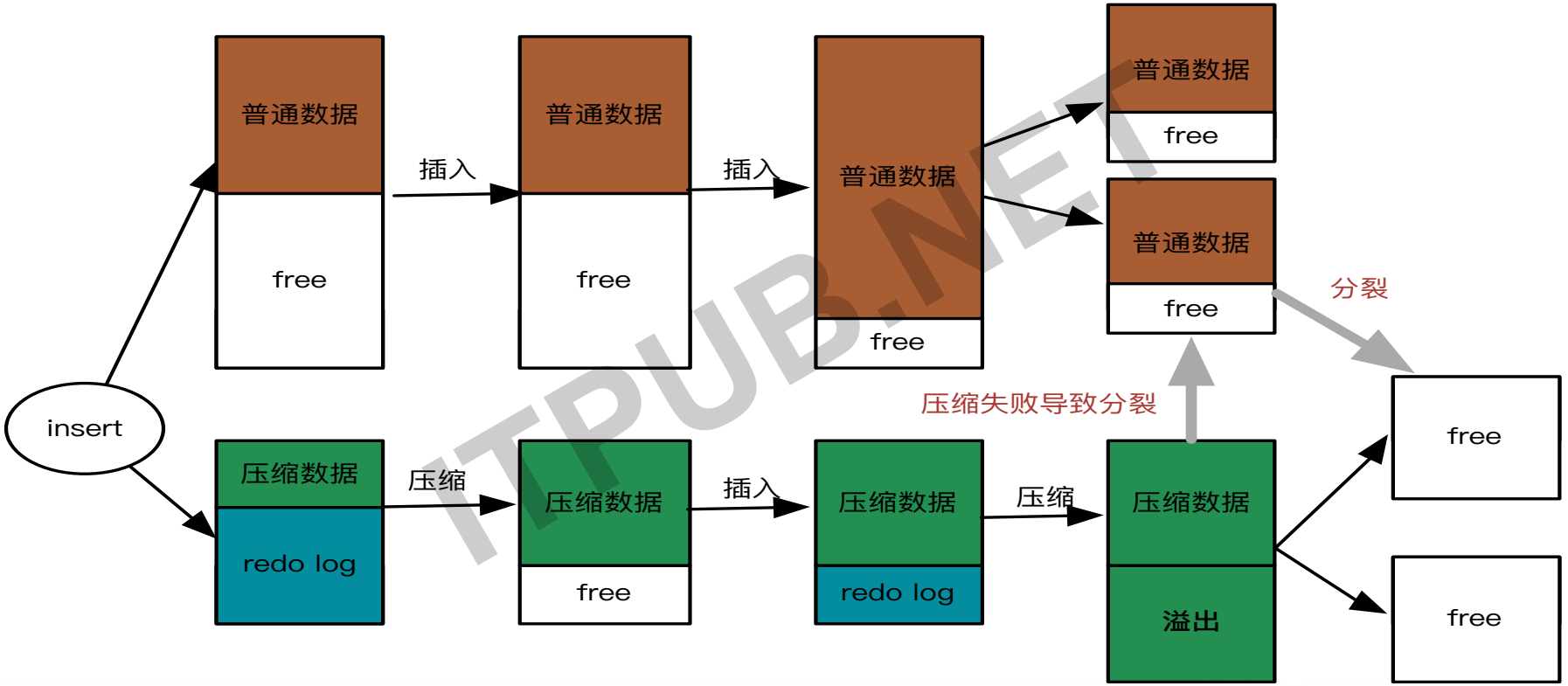
写性能差原因1：同步压缩

- 原因1：写入时增加的同步压缩操作。

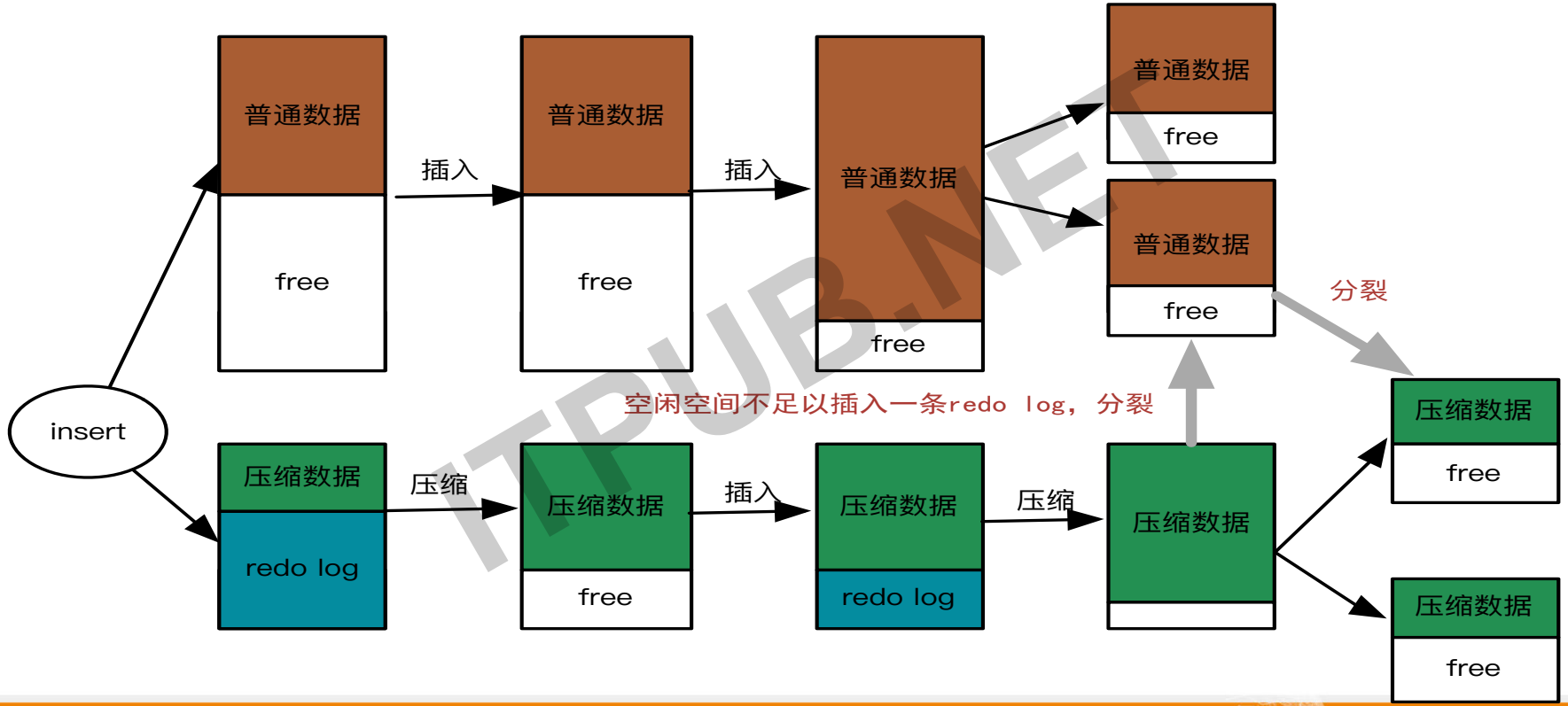


- 解决方案：
  异步压缩

写性能差原因2：压缩失败页分裂



解决方案：提前分裂



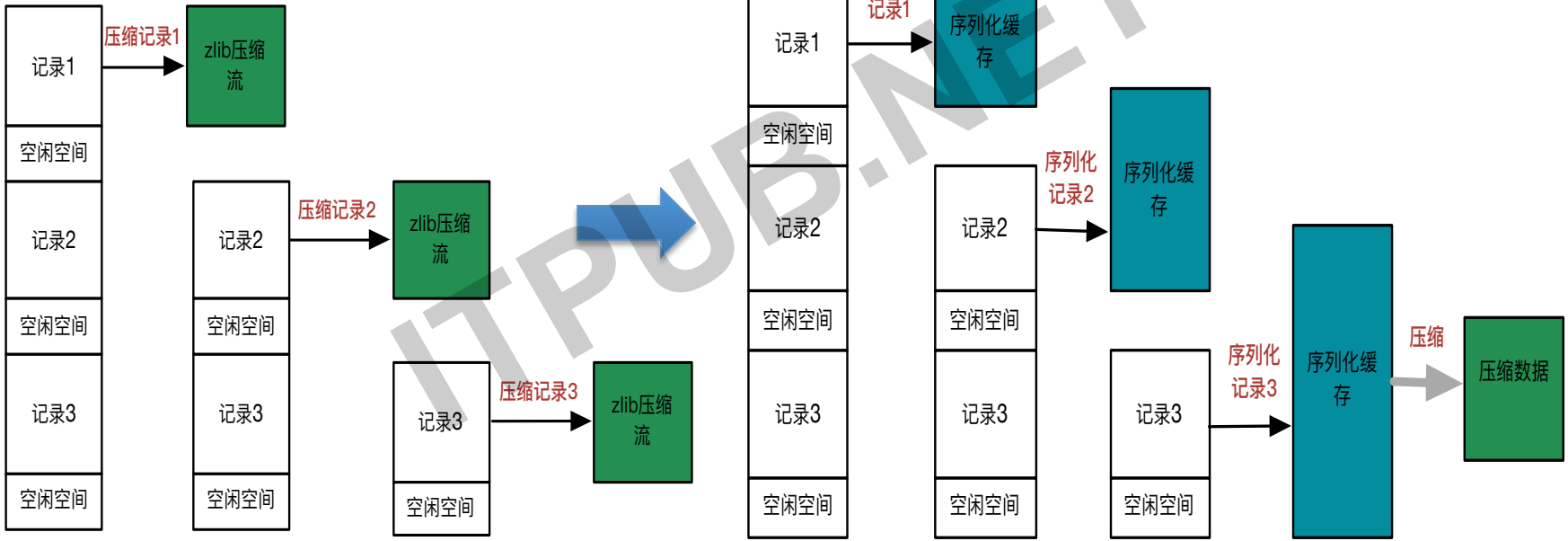
问题：提升读性能

- 原因：ssd上读取速度和innodb zlib解压速度相当。节省的io被解压操作抵消了。
- 解决方案：替换压缩算法LZ4

算法	压缩率	压缩速度	解压速度
zlib	0.211724	29.060595	236.911942
lzo	0.306656	387.374542	711.443909
snappy	0.314193	241.024979	859.810669
lz4	0.297518	385.175537	1618.197998

问题：提升读性能

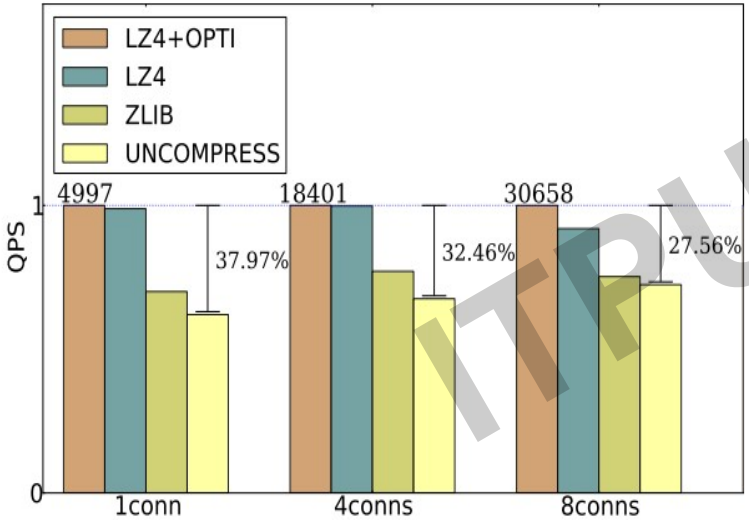
- LZ4基于块压缩，ZLIB基于流式压缩
- 需要将流式记录压缩转化为块压缩



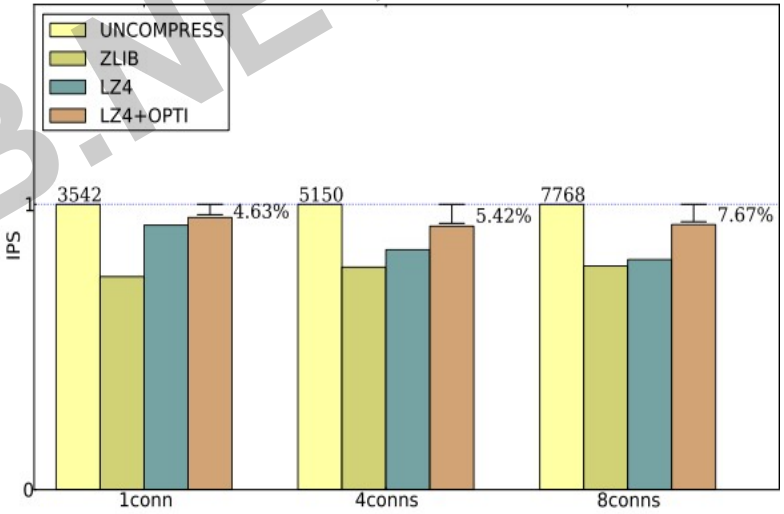
优化成果

- Sysbench压测

Read，读性能提升20~30%



Write，写性能下降10%以内



大纲

- 百度网盘简介
- 问题与挑战
- 阶段一：InnoDB压缩（100G压缩到60G）（2014~2015）
- 阶段二：TokuDB压缩（100G压缩到35G）（2016~2018）
- 阶段三：下一步（MyRocks & 冷热分离）（2018~2019）

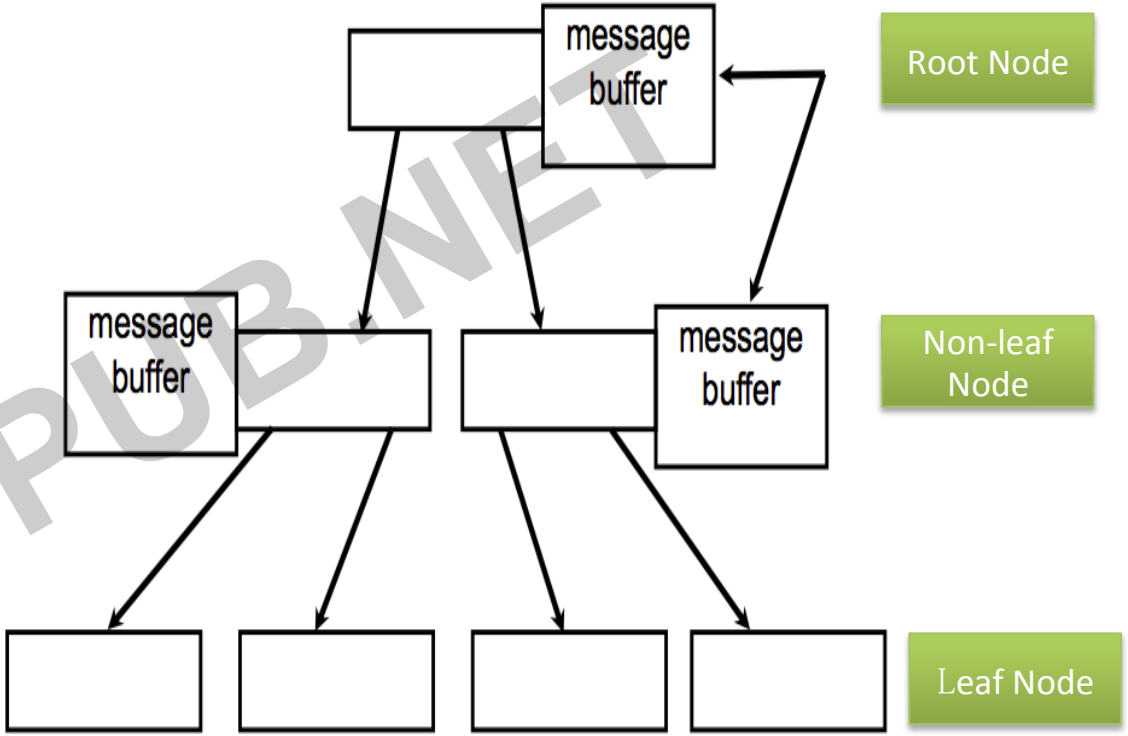
阶段二：tokudb压缩

特点：

- Fractal Tree
- B+ tree + Message Buffer
- Big nodes（4MB vs. ~16KB）

性能：

- 优化了写入行为，同步改异步。
- 读性能下降，每次查询请求都需要对同条链路上的异步消息缓存消息进行下推。



主要的工作

1. 压缩比和性能测试
2. ZSTD新压缩算法引入
3. 海量数据迁移
4. 稳定性工作

1. 压缩比和性能测试

- 压缩测试：模拟线上环境，采用的是网盘PCS表的数据
- 性能测试：sysbench压测，表个数是10张，每张表1000W行记录，OLTP环境

压缩算法	数据库版本	压缩率	并发数	TPS	读写请求	平均耗时
LZ4(InnoDB)	MySQL5.6	58.5%	10	243/s	4474/s	41.3ms
			20	485/s	8481/s	41.6ms
			50	1156/s	20813/s	43.6ms
QUICKLZ(TokuDB)	Percona5.6	34.5%	10	785/s	15141/s	11.6ms
			20	1239/s	22309/s	16.2ms
			50	1248/s	22470/s	40.1ms

Tokudb引擎优势

- Innodb的压缩受到页对齐影响，理论上线不会超过50%。
- innodb引擎的压缩率接近58.5%，tokudb引擎压缩率可达到34.5%，空间节省70%。
- Tokudb在低并发（<=20）场景下性能有明显优势。

2. ZSTD压缩算法引入

压缩名称	压缩比例	压缩速度	解压速度
ZSTD 1.3.4-1	2.877	470MB/S	1380MB/S
ZLIB 1.2.11-1	2.743	110MB/S	400MB/S
QUICKLZ 1.5.0	2.238	550MB/S	710MB/S
LZ4 1.8.1	2.101	750MB/S	3700MB/S
SNAPPU 1.1.4	2.091	530MB/S	1800MB/S
LZF 3.6-1	2.077	400MB/S	860MB/S

1. 综合权衡压缩率和性能，ZSTD最优。
2. ZSTD和tokudb默认QUICKLZ算法相比，压缩率提升20%左右，性能提升7%左右。
3. ZSTD 支持多个压缩级别，从level=1到level=22。level值越大，压缩效果越好，但占用的资源和压缩时间也就越长。Level=6

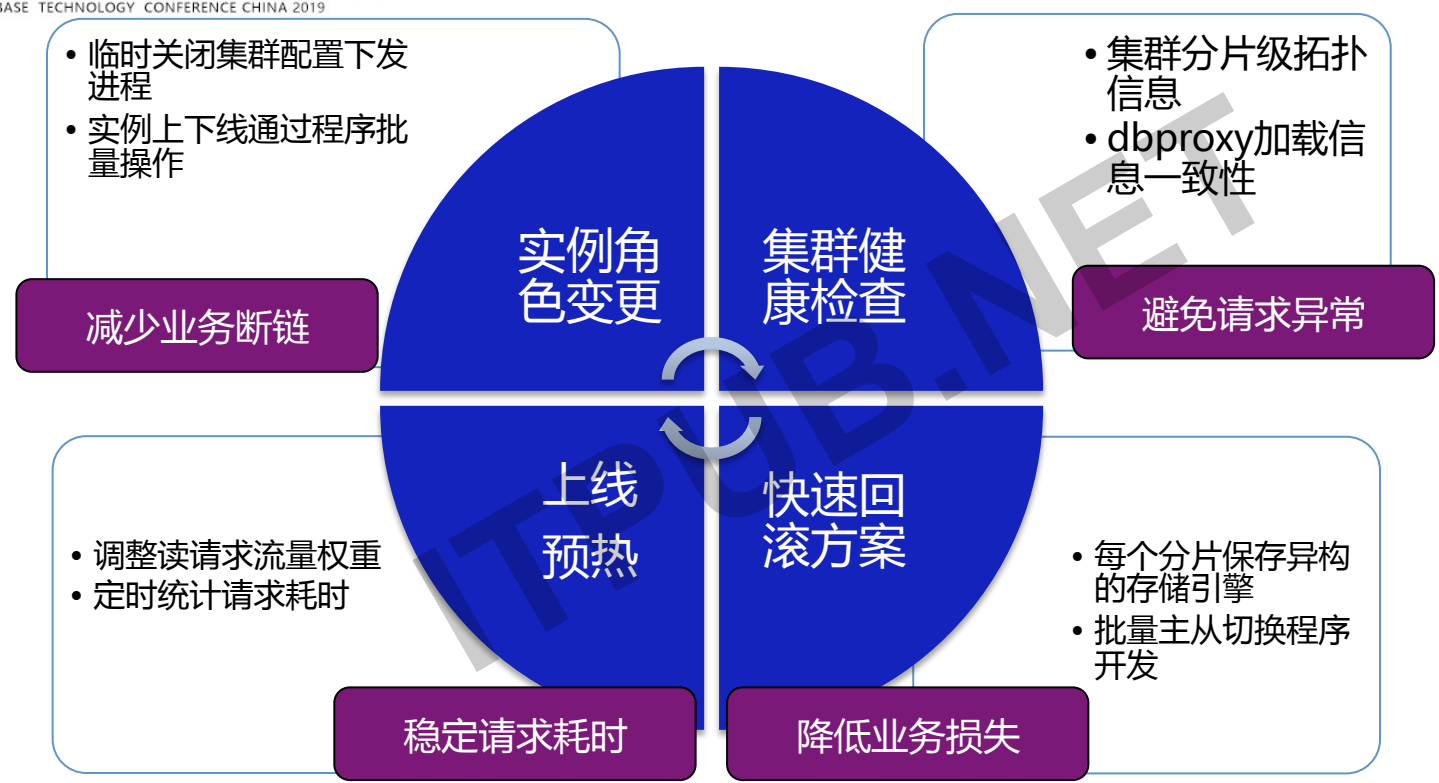
ZSTD效果

压缩算法	存储容量	并发数	QPS读	SQL平均相应	CPUIDLE	IOUTIL
QUICKLZ	561GB	10	5670/s	1.18ms	64%	9%
		20	9000/s	1.42ms	36%	10.2%
		30	11800/s	1.64ms	28%	18.3%
ZSTD_level_1	478GB	10	5640/s	1.15ms	67%	11.2%
		20	9400/s	1.31ms	39%	13.2%
		30	11500/s	1.52ms	30%	15.1%
ZSTD_level_6	463GB	10	5710/s	1.26ms	66%	7.2%
		20	9200/s	1.35ms	33%	9%
		30	12000/s	1.53ms	26%	16.4%

ZSTD效果

引擎名称	压缩算法	存储容量	并发数	TPS	读写请求	平均耗时
LZ4(InnoDB)	MySQL5.6	58.5%	10	243/s	4474/s	41.3ms
			20	485/s	8481/s	41.6ms
			30	1156/s	20813/s	43.6ms
QUICKLZ(TokuDB)	Percona5.6	34.5%	10	785/s	15141/s	11.6ms
			20	1239/s	22309/s	16.2ms
			50	1248/s	22470/s	40.1ms
ZSTD(TokuDB)	MySQL5.6	28.5%	20	比QuickLZ提升7%	比QuickLZ提升7%	——

3. 数据迁移



4. 稳定性相关

- Bug修复
  - 一致性约束失效
  - tokudb引擎统计信息失效
  - 关闭binlog导致MySQL实例crash
  - 执行MySQL部分操作慢
  - TokuDB中ALTER TABLE可能产生的线程阻塞
  - jemalloc库版本bug
  - Etc..
- 监控项
- xtrabackup

压缩技术总体收益

- 存储容量：网盘数据库数据量减少PB级别
- 成本节省：
  - 数据库压缩技术每年为网盘节省近千台服务器
  - 网盘数据库单GB存储成本减少70%。
- 其他收益：
  - 百度网盘的压缩整体方案只是一个开始。
  - 数据库压缩技术已经在百度所有的MySQL业务上推广起来。

大纲

- 百度网盘简介
- 问题与挑战
- 阶段一：InnoDB压缩（100G压缩到60G）（2014~2015）
- 阶段二：TokuDB压缩（100G压缩到35G）（2016~2018）
- 阶段三：下一步（MyRocks & 冷热分离）（2018~2019）

不同存储引擎压缩算法适配场景

innodb（LZ4）	tokudb（ZSTD）	rocksdb（ZSTD）
优势： ✧ 数据压缩效果一般 ✧ 请求类型支持广泛 ✧ 事务支持完善	优势： ✧ 数据压缩效果最佳 ✧ 表字段以字符串为主 ✧ 数据碎片化少	优势： ✧ 数据压缩效果较好 ✧ 数据类型基于KV存储
劣势： ✧ 随机写性能较差 ✧ 实例崩溃恢复速度慢 ✧ 回滚段存在空间浪费	劣势： ✧ 不支持外键约束 ✧ 读请求耗时稍长 ✧ 不适合范围查询	劣势： ✧ 事务支持不完善 ✧ 排序操作性能差 ✧ 不支持在线表变更

- 多种引擎并存
  - INNODB：对事务要求较高的业务：金融类、商业订单类。
  - Tokudb： 已经成熟
  - MyRocks： 正在成熟落地

MyRocks是未来方向

优势：

- LSM-Tree写性能大大优于B+树，FTL树。
- RocksDB的压缩率和TokuDB接近。
- 社区非常活跃，各IT厂商纷纷采用。
- 学术界不断有各种论文和研究成果推出：读写性能提升，空间放大优化。

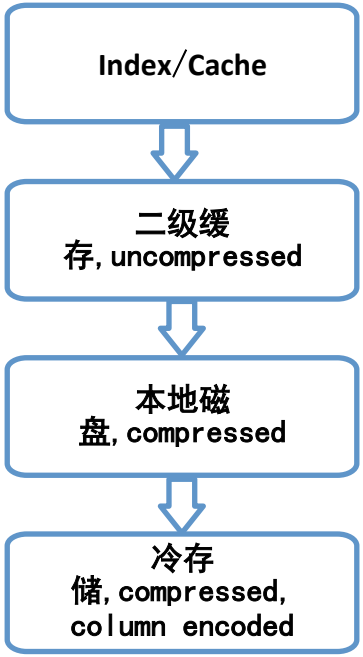
计划：软硬结合，下一代高性能引擎

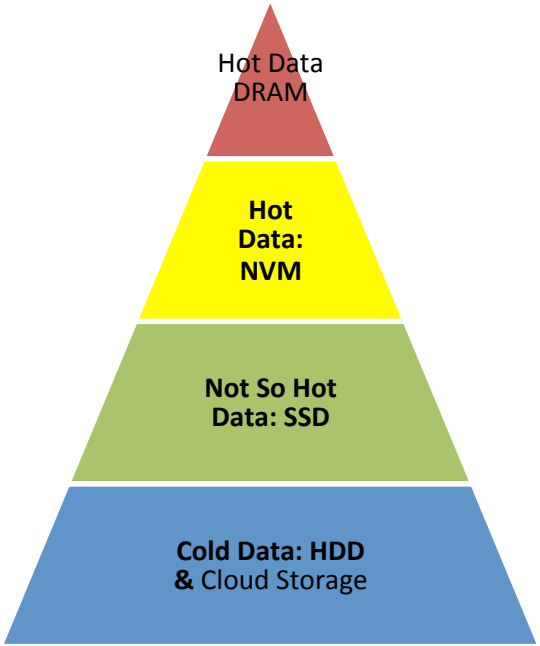
- 将CPU密集的操作offload到FPGA/GPU。
- 使用Open Channel SSD深度定制针对MySQL业务特点的磁盘。
- 使用NVM技术提升读写性能。
- 在软件层面，优化LSM-Tree，进一步提升读写性能。

新硬件推动了数据库分层存储模式

- 经常被访问的数据存储在内存、NVM中，提供最高的性能。
- 历史数据存储在SSD/HDD磁盘上，甚至是云存储上。
- 自动分析数据的冷热特征，动态调度，弹性扩缩容量。

介质	SSD	DRAM（内存）	NVM（非易失性内存）
性能	低	高	SSD < NVM < DRAM
易失性	非易失	易失	非易失
成本	低	高	SSD < NVM < DRAM






THANKS