

小米结构化存储系统 及融合云平台的 设计与实践

林尚泉

linshangquan@xiaomi.com



大纲

- 小米结构化数据存储
 - 融合云平台



结构化存储

分布式NoSQL数据库服务 对标AWS DynamoDB

弹性可扩展

高可用

低延时

稳定可靠



结构化存储

SDS

(Structured Data Store)

基

无

简

多

于
H
B
A
S

状
态
公
网
访

化
认
证
和
配

语
言
S
D

多租户

功能扩展

ACL

流
量
控
制

数
据
类
型

二
级
索
引

S
T
R
E
A
M

软
删
除

数
据
冷
备

结构化存储

小米内部业务



生态链业务



结构化存储

- 应用规模

国际化

- 北京，天津，美国，新加坡

机器规模

- 100+

业务数

- 20+

数据量

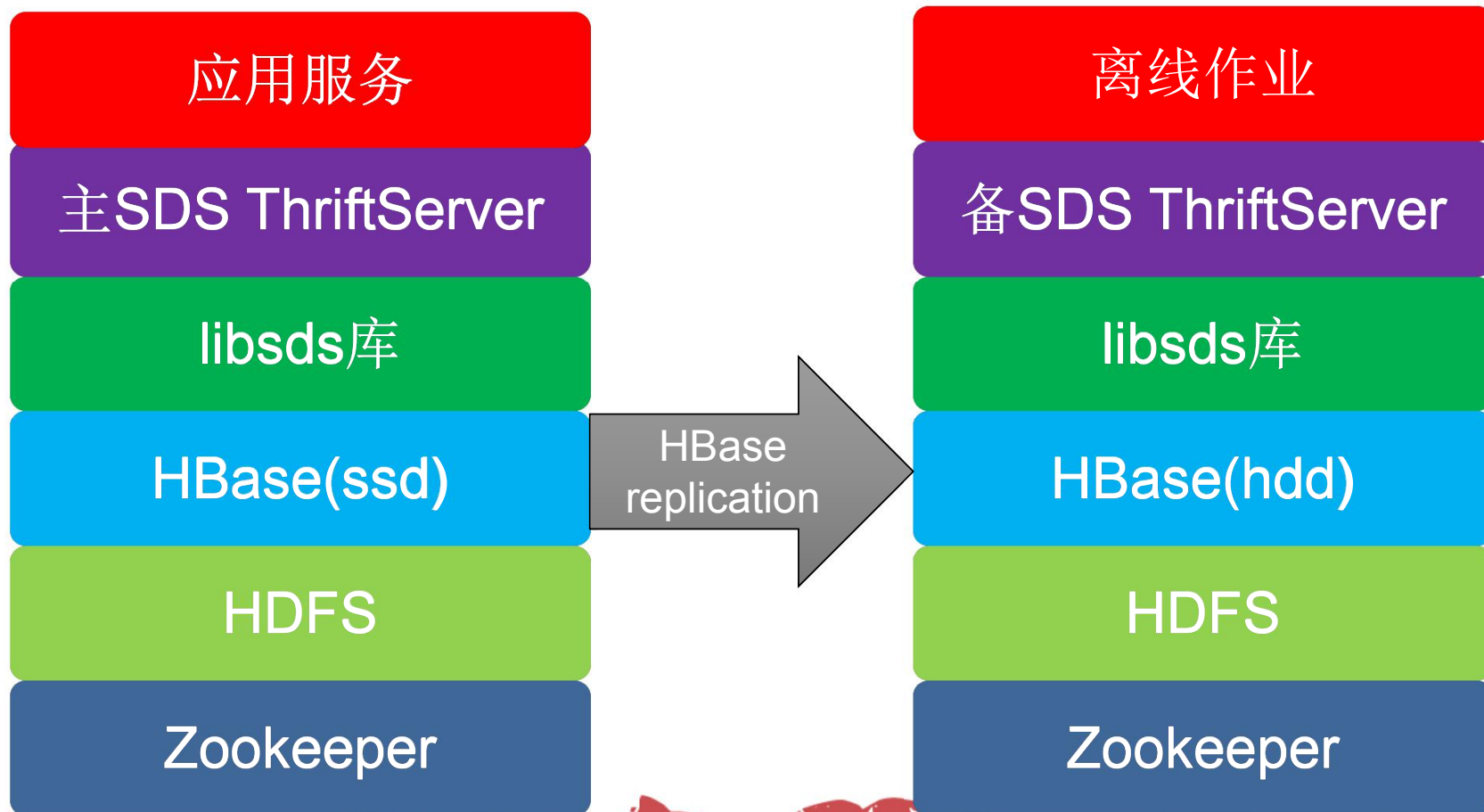
- 100+TB，数千亿行

单集群QPS

- 10万级

结构化存储

典型部署



结构化存储

- libsdsc库

libsdsc

规范
化的
数据
模型

内建
数据
类型
支持

局部
二级
索引

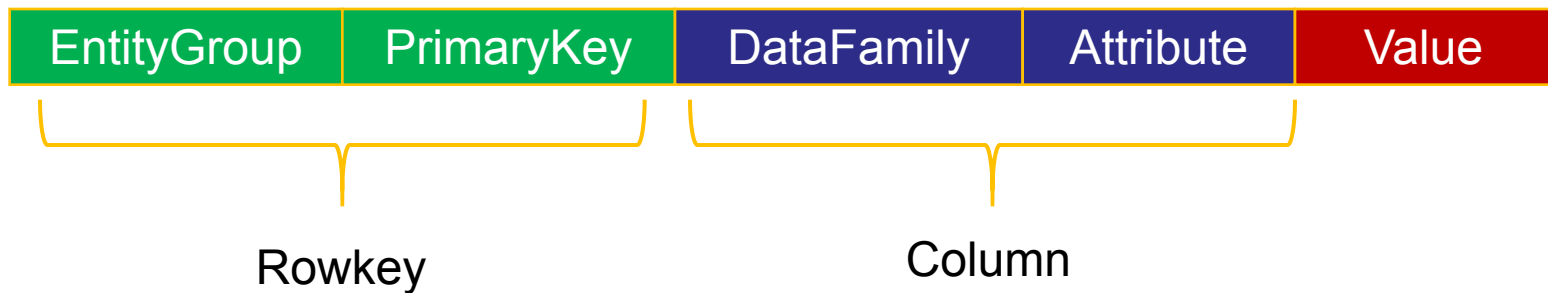
全局
二级
索引

stream

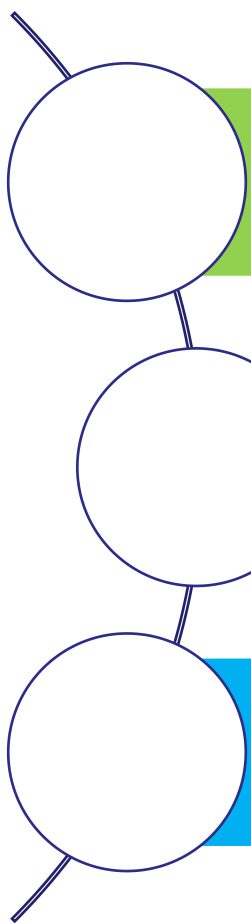
结构化存储

- 规范化数据模型

- Rowkey前缀：实体组键，支持哈希分布
- Rowkey后缀：主键
- Column Qualifier：数据列



结构化存储



数据类型支持

- bool/int8/int16/int32/int64/float/double/string/binary/set

HBASE-8201和SQLite编码方案¹

- 顺序保持编码

支持逆序

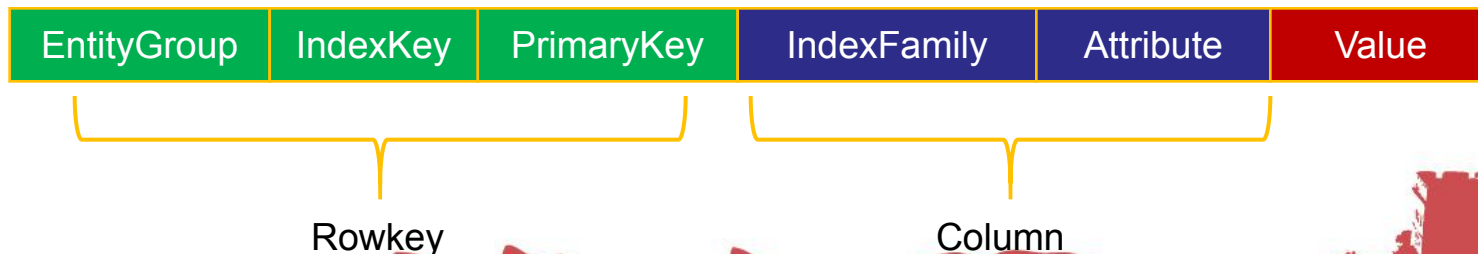
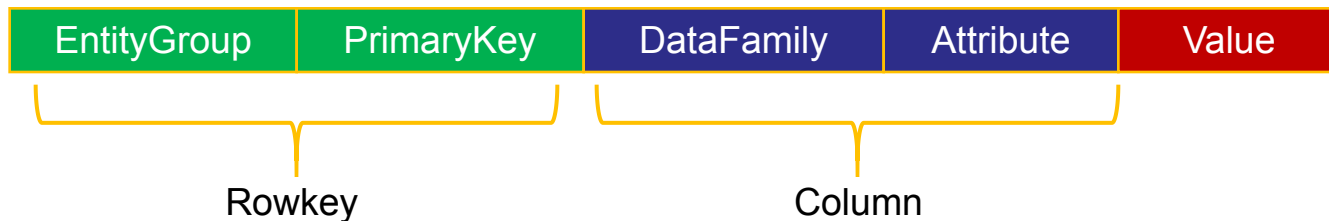
- 按位取反

¹http://sqlite.org/src4/doc/trunk/www/key_encoding.wiki



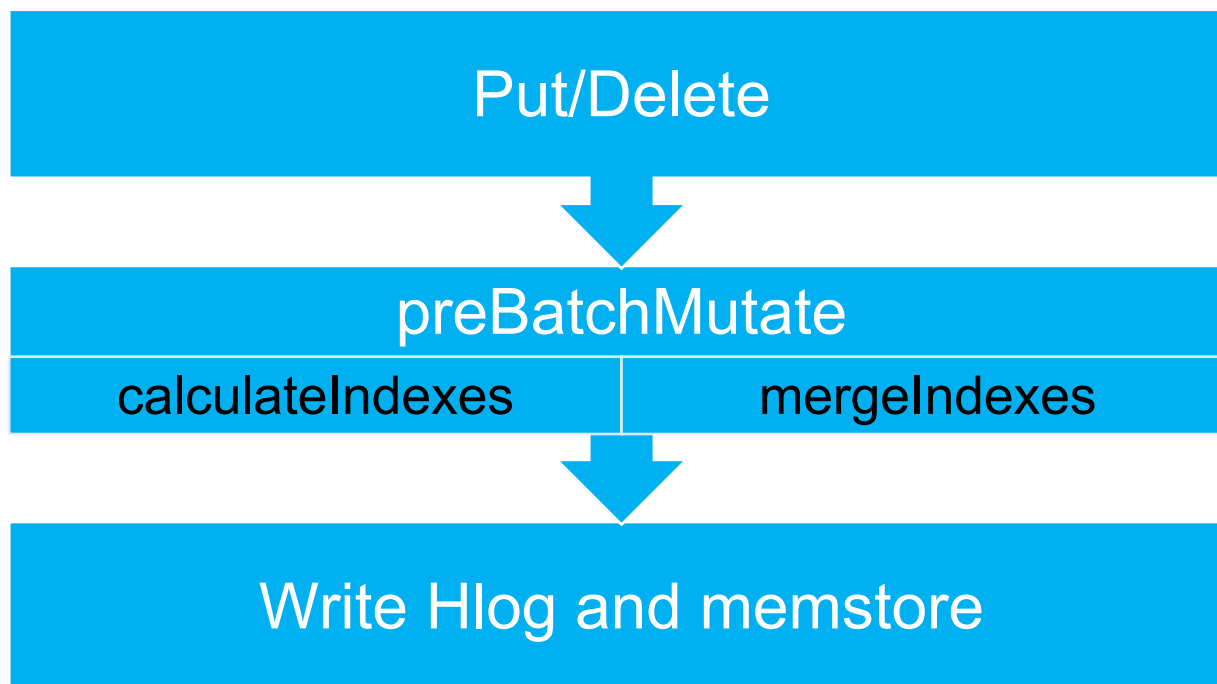
结构化存储

- 局部二级索引：实体组内部
 - 同一张表，不同Column Family
 - 基于正则表达式的前缀分割策略(RegexPrefixRegionSplitPolicy):
保证同一个实体组键数据不能split



结构化存储

- 局部二级索引：HBase
Coprocessor——IndexObserver



结构化存储

- 多种类型局部二级索引

EAGER

更新删除时同时删除失效索引，适合写少读多

LAZY

读取时判断索引有效性，更新时不做额外操作，适合写多读少

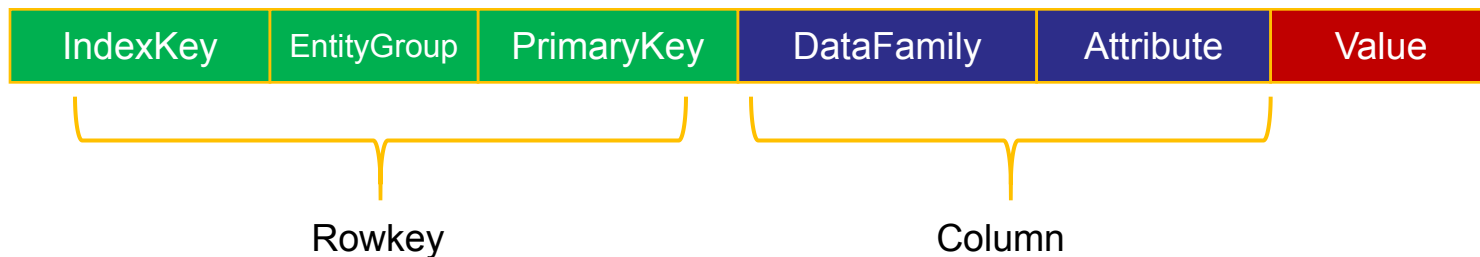
IMMUTABLE

适合只读数据一次性写入，读写均无需额外判断

结构化存储

- 全局二级索引

- 使用单独的HBase 表存储索引
- Google percolator²的 HBase实现: Themis³, 保证跨表更新的原子性
- Chronos⁴全局单调递增时间戳



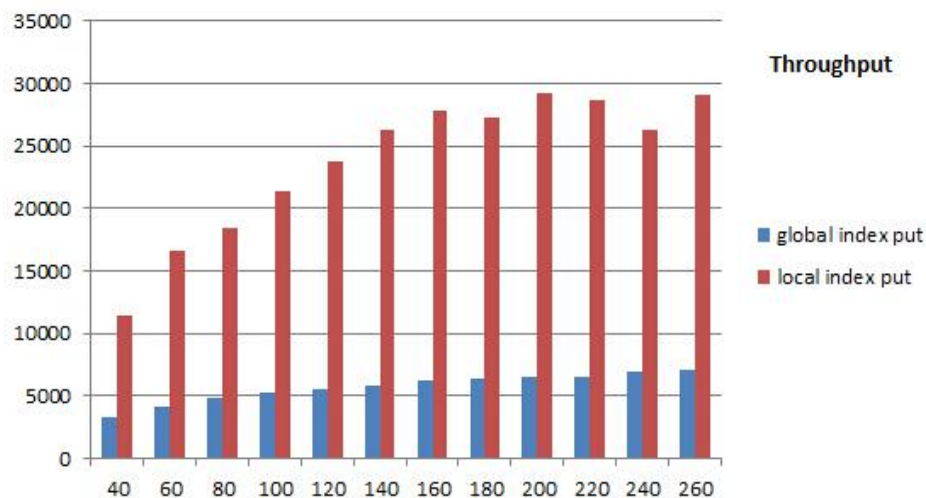
²Large-scale Incremental Processing Using Distributed Transactions and Notifications

³<https://github.com/xiaomi/themis>

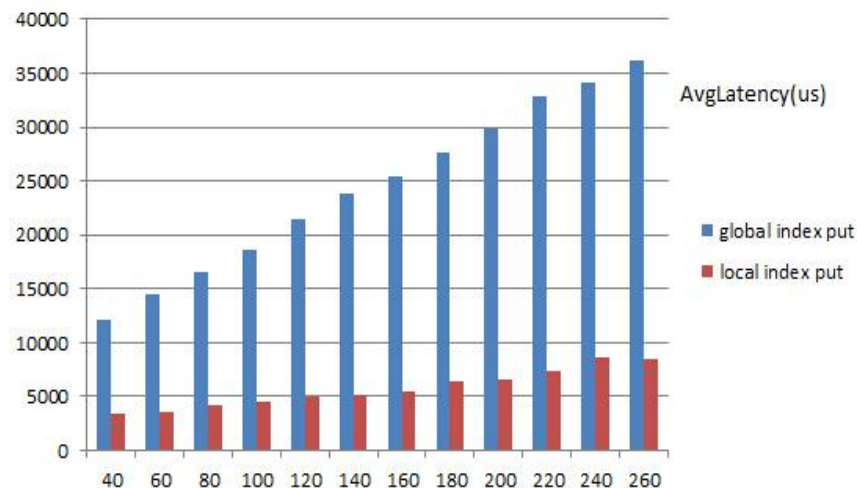
⁴<https://github.com/xiaomi/chronos>

结构化存储

- 局部二级索引 VS. 全局二级索引(写)



线程数

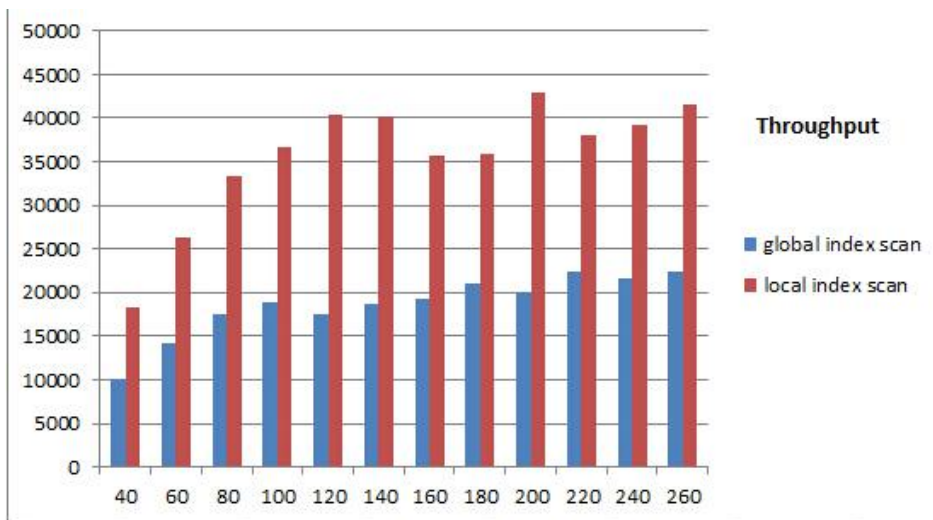


线程数

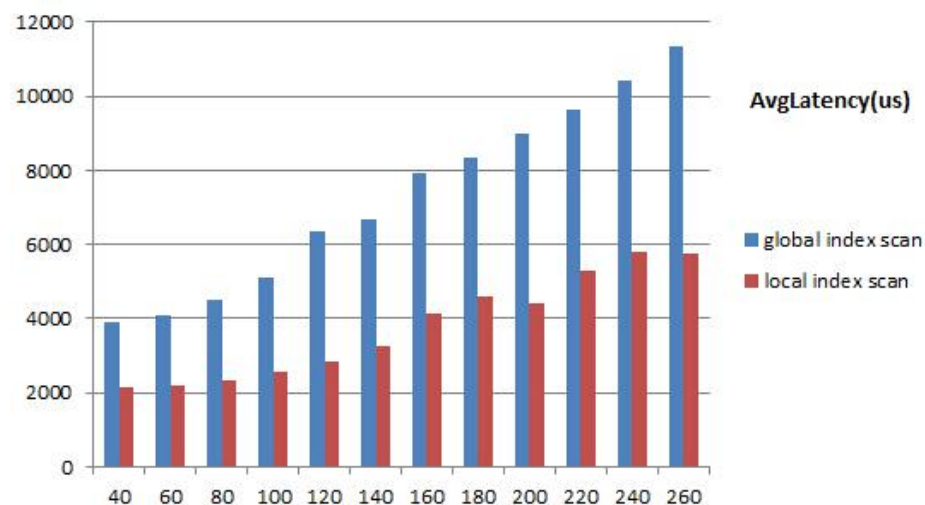


结构化存储

- 局部二级索引 VS. 全局二级索引(读)



线程数



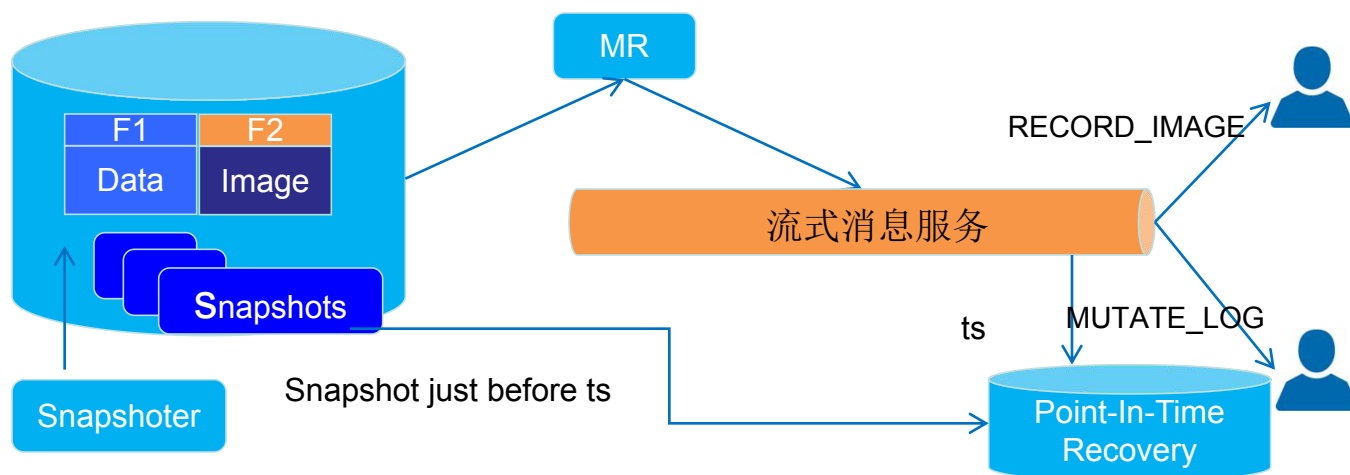
线程数



结构化存储

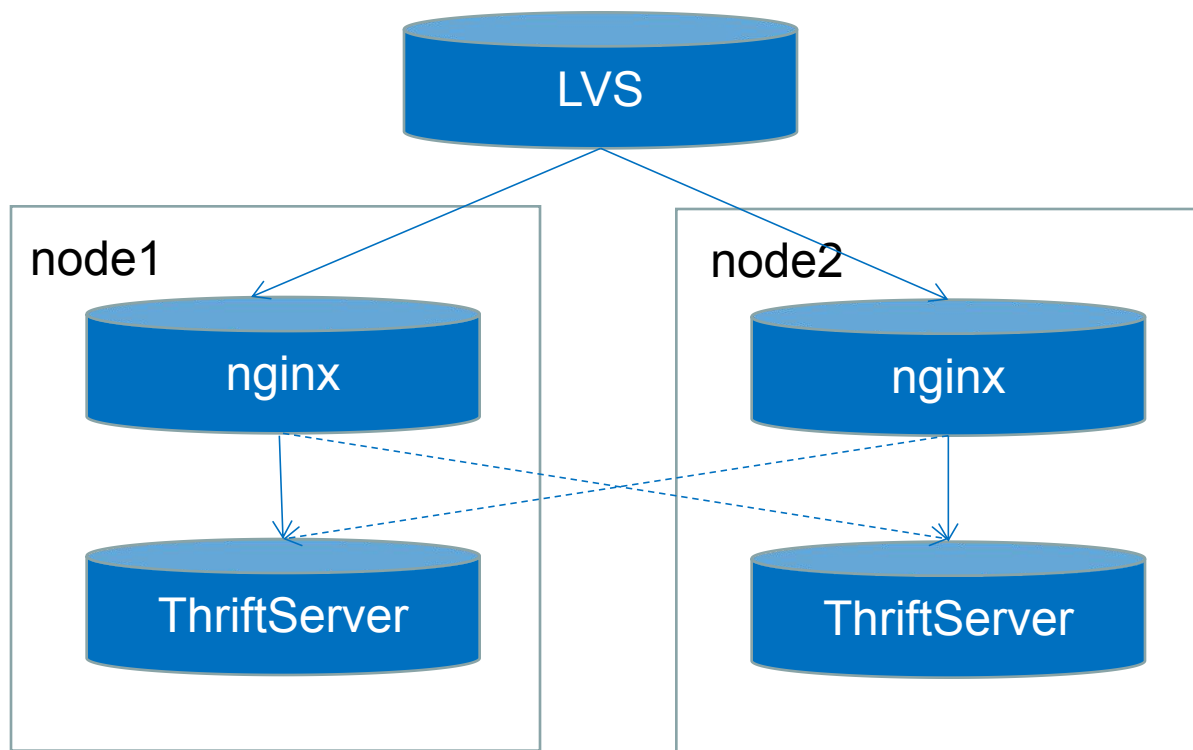
- stream

- RECORD_IMAGE: 增量备份
- MUTATE_LOG: Point-In-Time Recovery



结构化存储

- SDS ThriftServer: 无状态 Http Server



结构化存储

SDS ThriftServer

Http 无
状态公
网访问

简化认
证和配
置

多语言
SDK

多租户

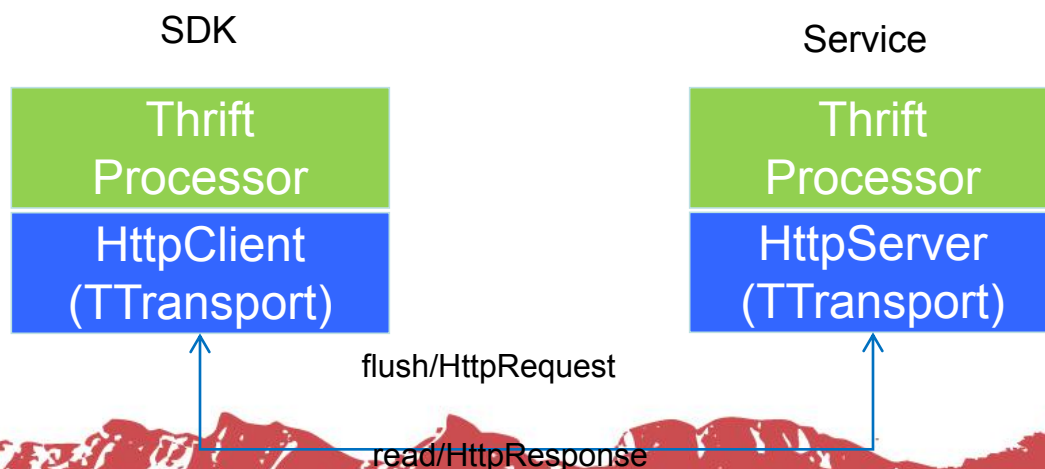
ACL

流量
控制

结构化存储

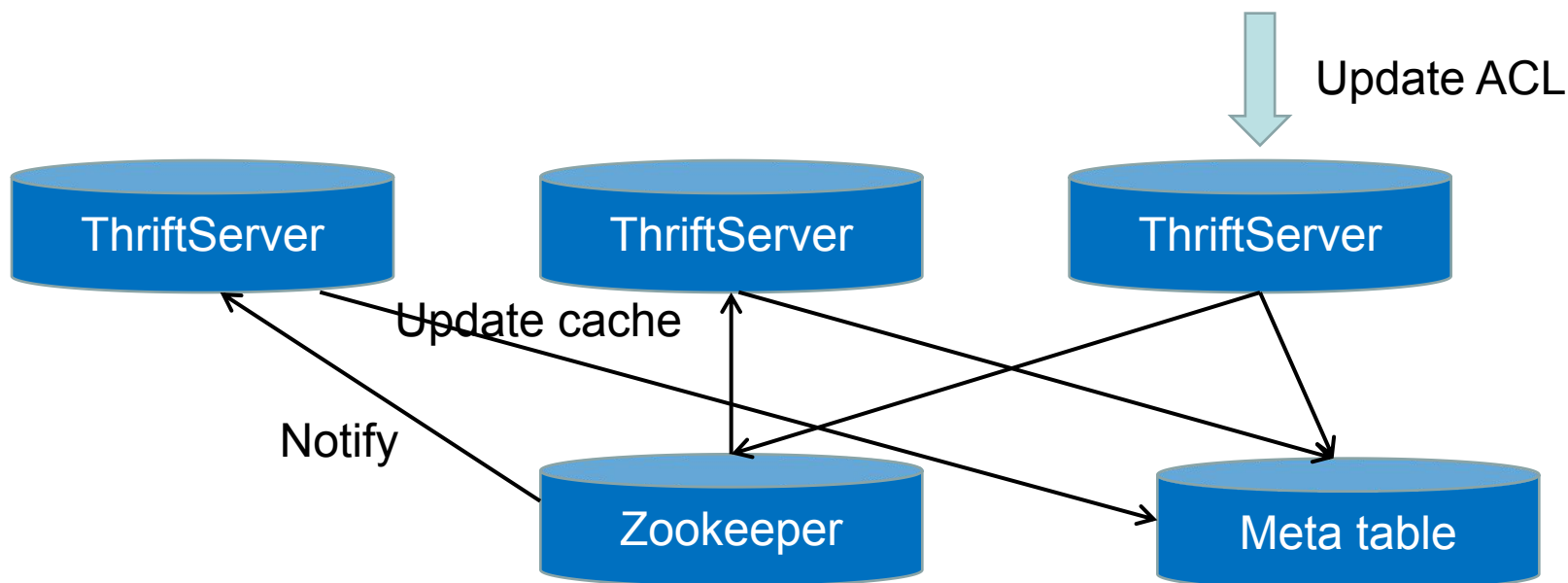
- SDS ThriftServer

- Http 无状态公网访问
- 简化认证和配置：使用小米融合云或开放平台认证，对外屏蔽了Zookeeper配置和Kerberos认证
- 多语言SDK(java/python/php/node.js/c++/go/javascript)



结构化存储

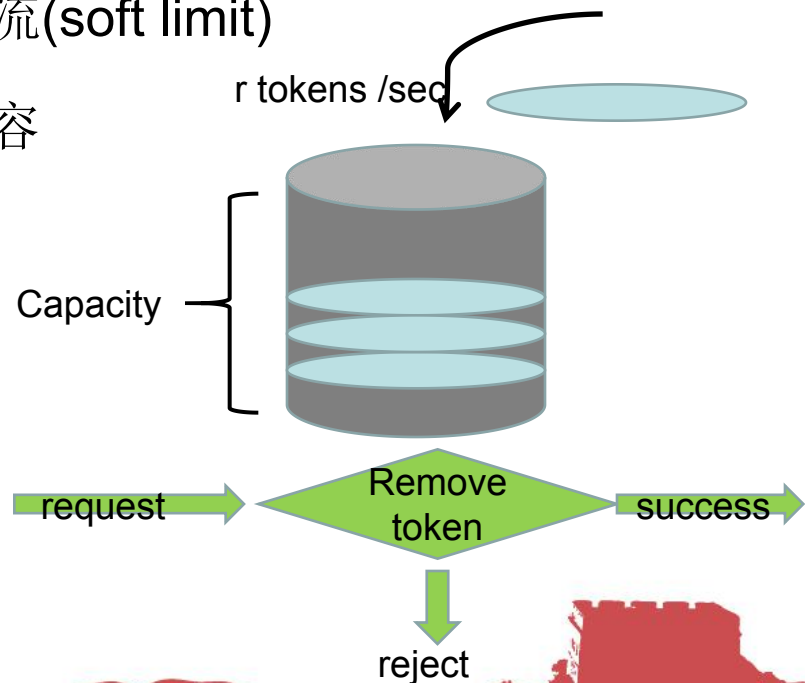
- SDS ThriftServer (ACL)
 - 元数据表: tableId => {ACL信息}
 - 本地cache元数据, zookeeper watcher 同步



结构化存储

• SDS ThriftServer (流量控制)

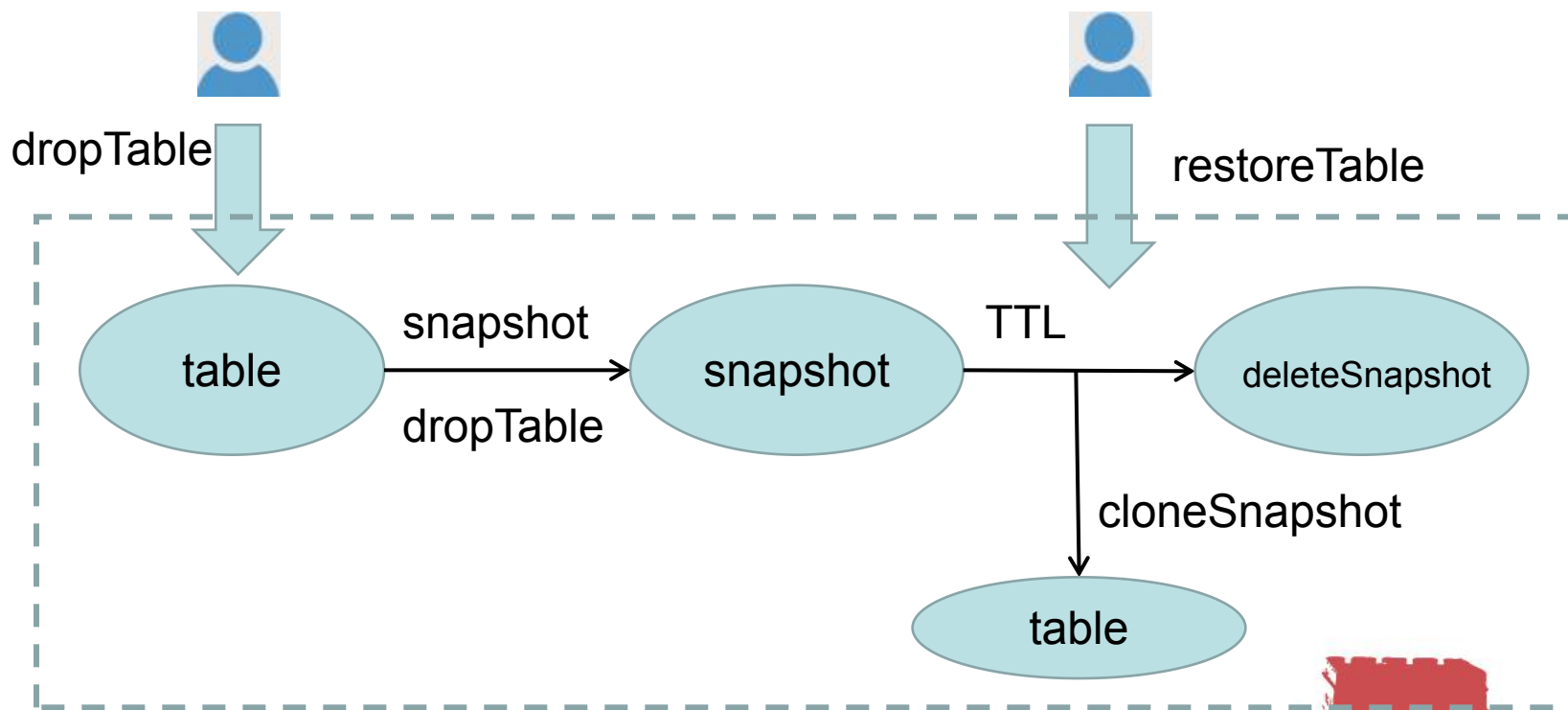
- 用户设置表的读写quota，设置表quota时检查集群能力
- 基于token bucket算法⁵进行限流(soft limit)
- 集群能力使用超80%时提醒扩容



⁵https://en.wikipedia.org/wiki/Token_bucket

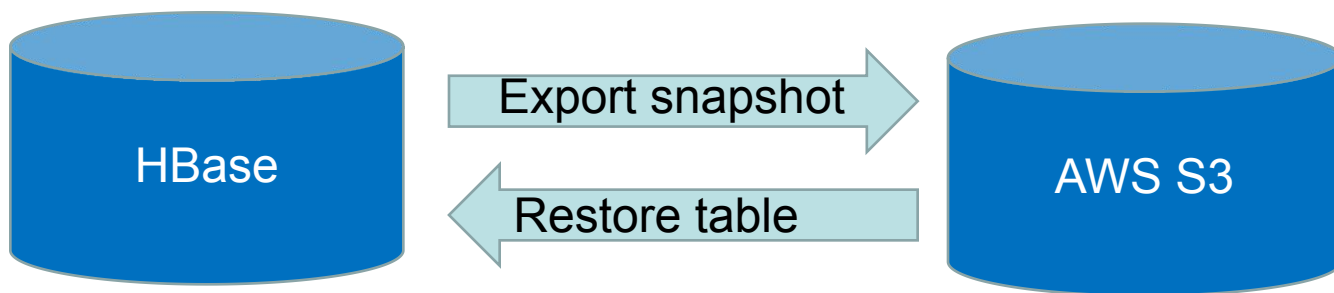
结构化存储

- 软删除



结构化存储

- 数据冷备



大纲

- 小米结构化数据存储
 - 融合云平台



融合云



小米融合云

产品

文档



团队管理



Docker 镜像



结构化存储



文件存储



HBASE



HDFS



消息队列



流式消息队列



LCS



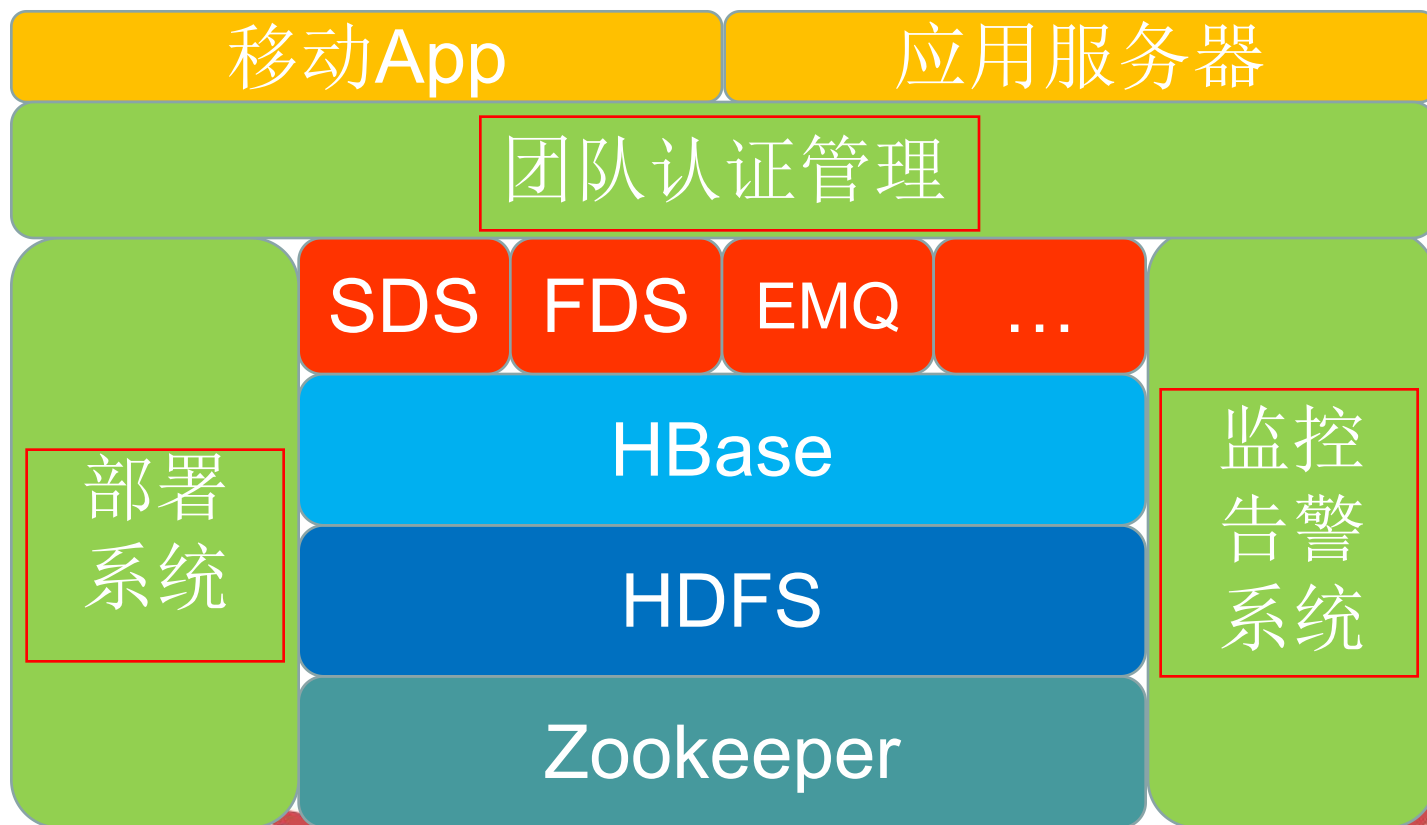
OLAP



深度学习

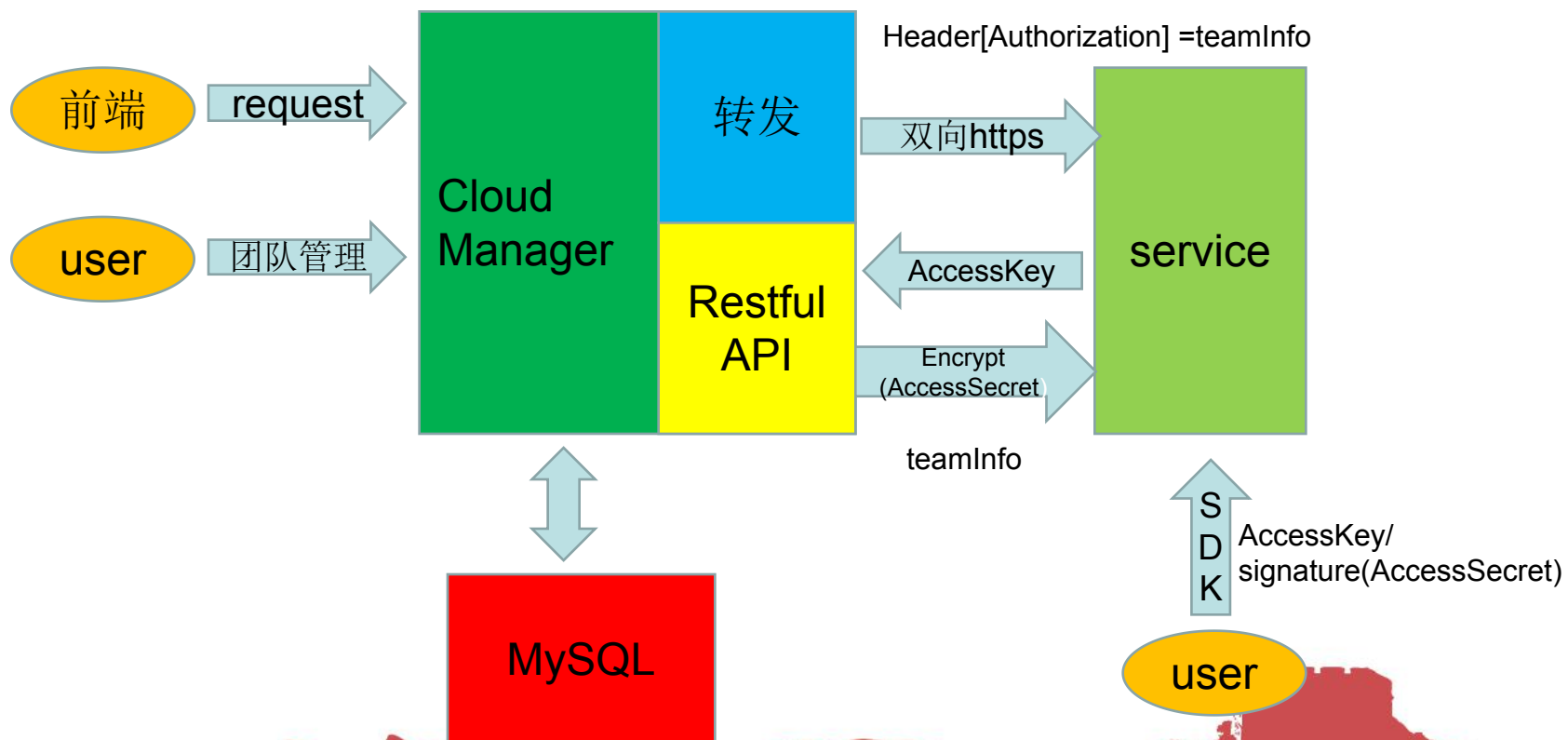
融合云

- 架构



融合云

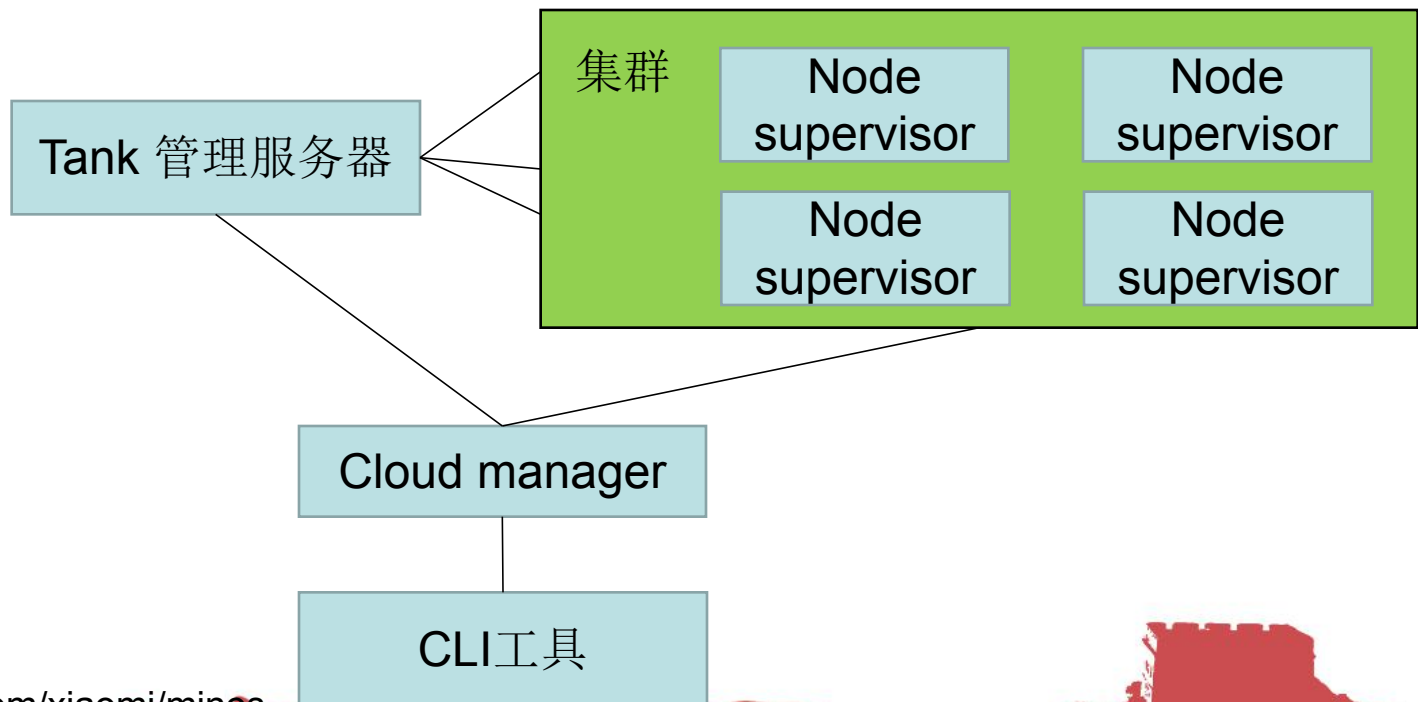
• 团队认证管理——CloudManager



融合云

- 部署系统——Minos⁶ 2.0

➤增加了认证授权模块



⁶<https://github.com/xiaomi/minos>

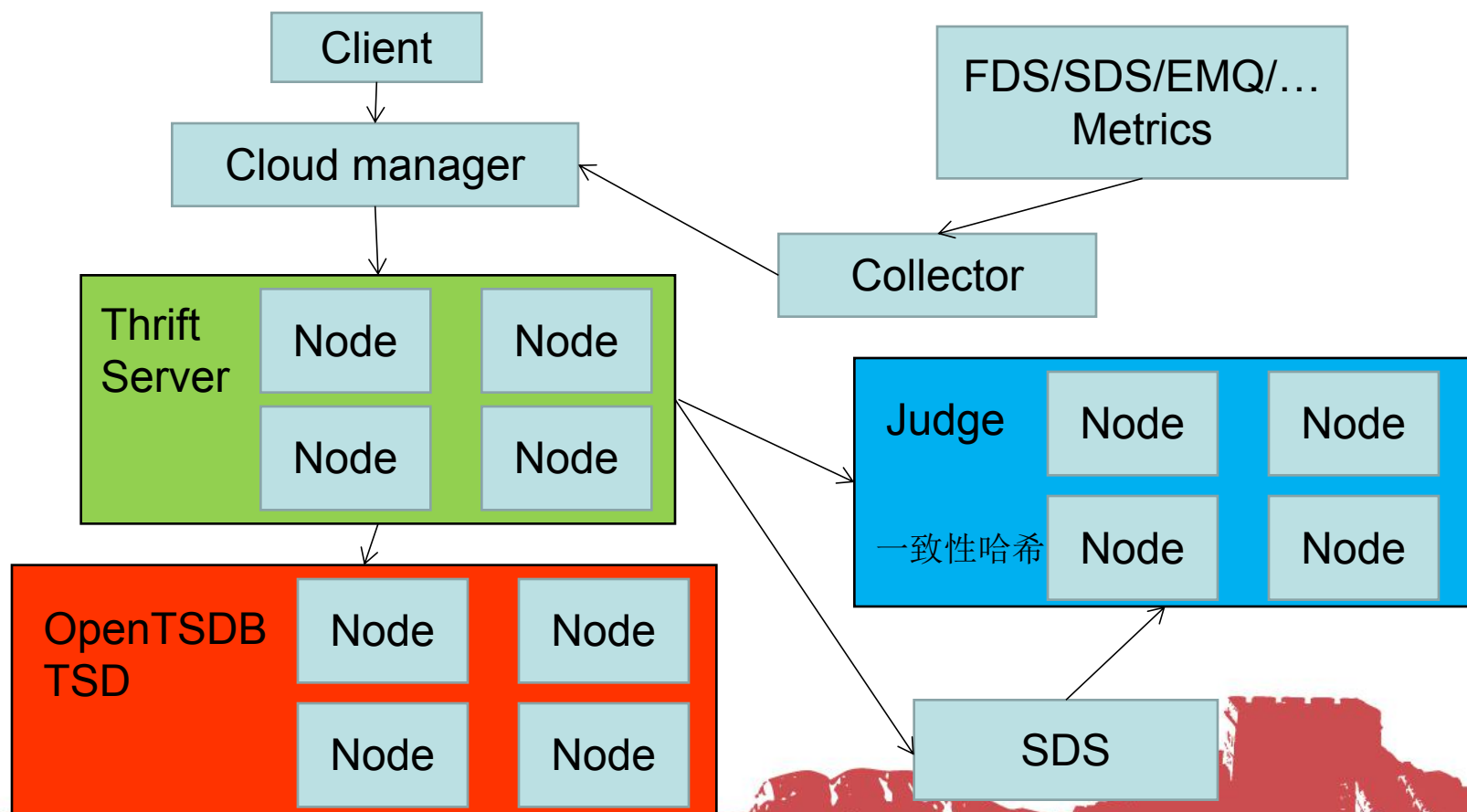
融合云

- 监控告警系统



融合云

- 监控告警系统——架构





THANK YOU

