

DTCC

2013中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2013

大数据 数据库架构与优化 数据治理与分析

SequeMedia
盛拓传媒

IT168.com

ITPUB

ChinaUnix

Database
BDaas
flowingdata
DB2
NoSQL MySQL
Oracle Big Data

厚德（HOLD）载物

腾讯在线交易平台统一数据层

高一致性解决方案

腾讯数据平台部-雷海林



目录

1. 为什么是厚德(hold)

2. 原理解析

2.1、总体架构

2.2、灾难检测

2.3、容灾切换

2.4、自动扩容

3. 性能优化

4. 总结与未来

为什么是厚德—业务场景



账户名	账户描述	现有账户量&请求量
TBOSS	各大钻、会员包月	<10亿账户，最高并发8000/s(读)
个账	QBQD	<10亿账户，最高并发2000/s(读)
云账户	游戏账户，积分账户	10亿账户，最高并发8000/s(读写)
安全策略中心	账户类安全限制	20亿账户，最高并发15000/s(读写)



为什么是厚德—最初梦想

● 并发不足！

● 耗时高！

接入层

● 容灾方案多种多样！

● 数据层扩容费劲！

账户逻辑层

● 运维成本太大！！

厚德平台

为什么是厚德—需求

账户的特点 & 系统要求

- 高价值 —— 具有高一致性，不间断的读写服务，遵循一个原则：宁可不服务，不能错账
- 高可用性 —— 遇到灾难，在尽可能短的时间内，恢复读写服务
- 数据层扩容 —— 需要不间断服务的数据层扩容方案
- 具有优秀的读写性能 —— 系统架构的最底层，高并发，低延迟<10ms
- 数据间没有关联关系

我们需要什么样的厚德平台

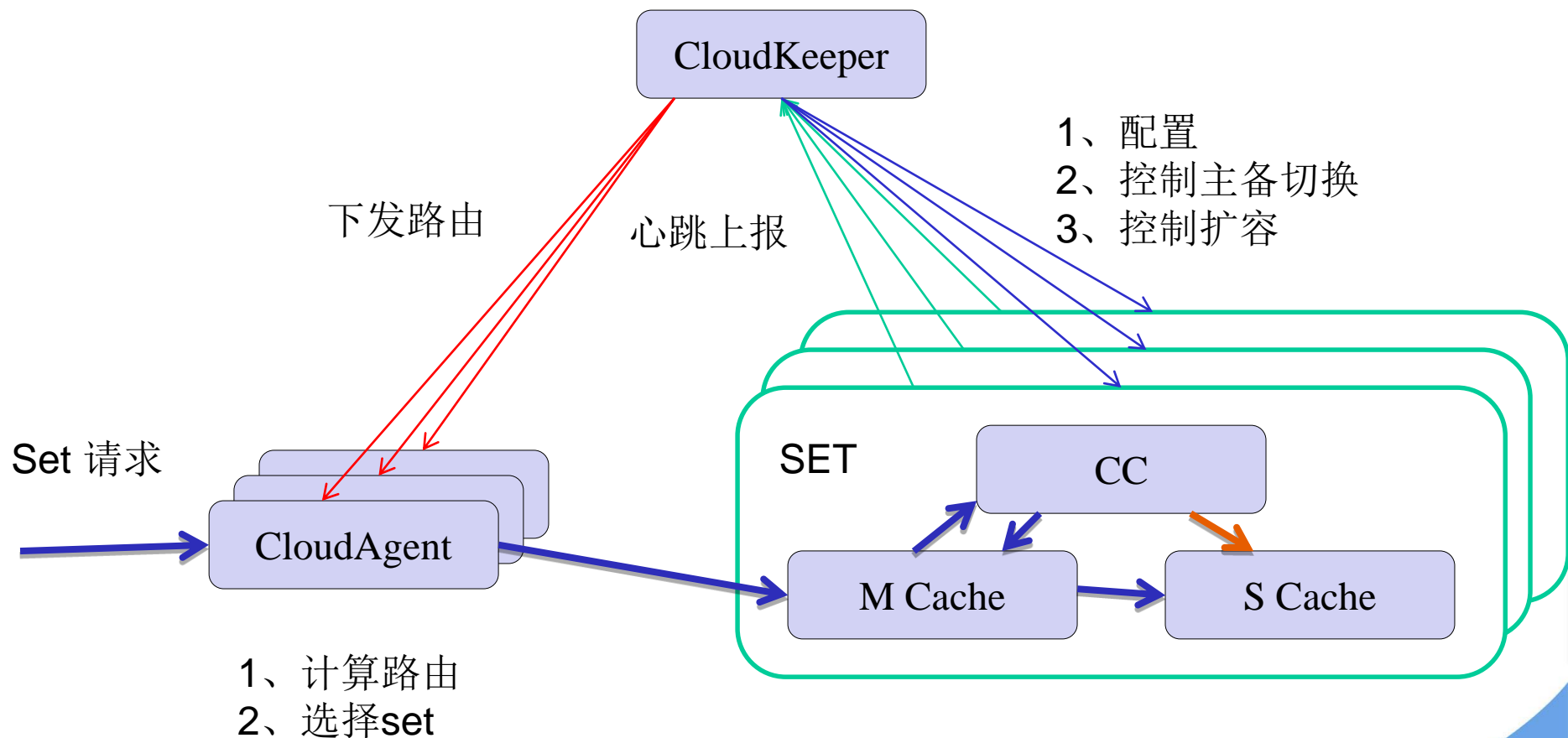
- 对于容灾
 - 准确判断
 - 高一致性切换
 - 容易回切
- 对于扩容
 - 不间断服务
 - 数据无损
- 支持海量读写
- 可以通过DB/Bindump/Binlog等方式实现数据落地

高一致性 分布式cache

原理分析

不自动解决双重灾难
宁可拒绝服务，也不能有数据错乱

总体架构



总体架构

数据一致性

-  记录版本号保证最终一致性,解决冲突

数据安全

-  用表把不同的数据隔离,对表进行鉴权

数据恢复

-  镜像/binlog/DB . . .

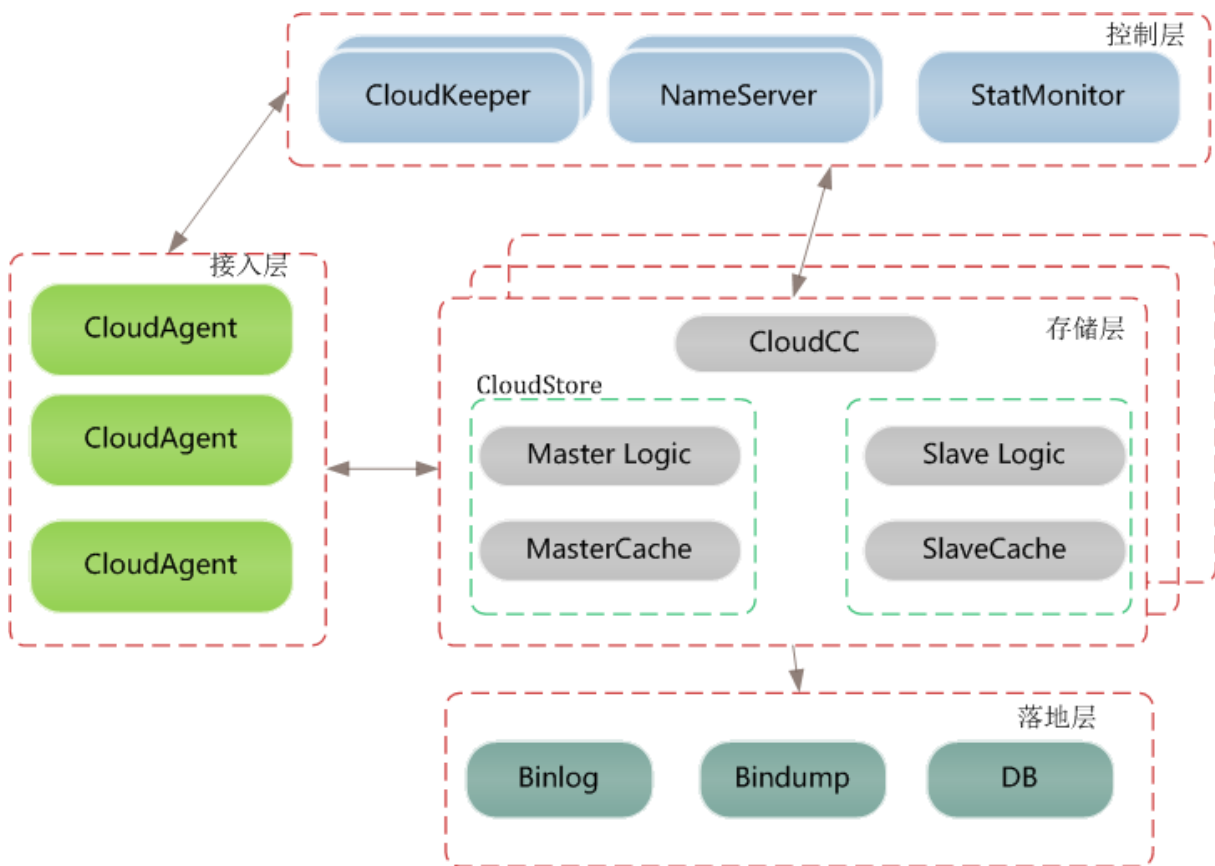
容灾

-  跨IDC容灾,跨城容灾

扩容

-  不停服务、数据无损

总体架构



控制层：

配置下发、状态监控、容灾流程控制、扩容流程控制、名字服务、状态数据展示

接入层：

路由计算、数据分发、权限校验、流量控制

存储层：

A、逻辑层：角色管理、binlog同步、主备同步、黑名单控制

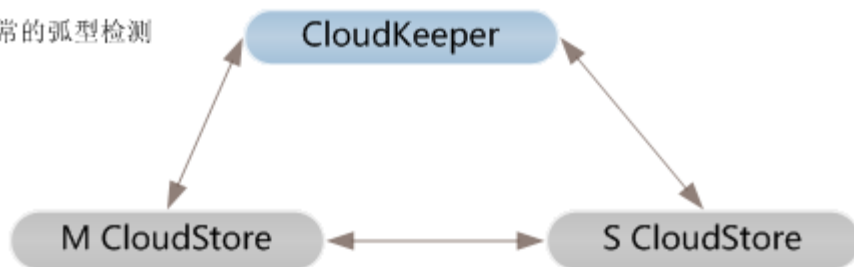
B、Cache层：单纯的读写操作、高性能

落地层：

把内存中的数据通过镜像和binlog的方式做持久化存储，同时可以根据数据分析的需要，把数据导入DB

容灾切换—孤岛检测

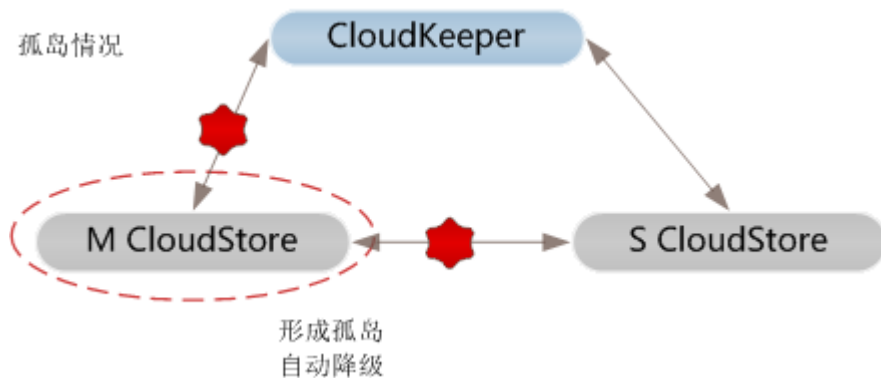
正常的弧型检测



正常弧形检测：

每个CloudStore除了上报自己的状况外，还会上报同Set另外一个CloudStore的情况

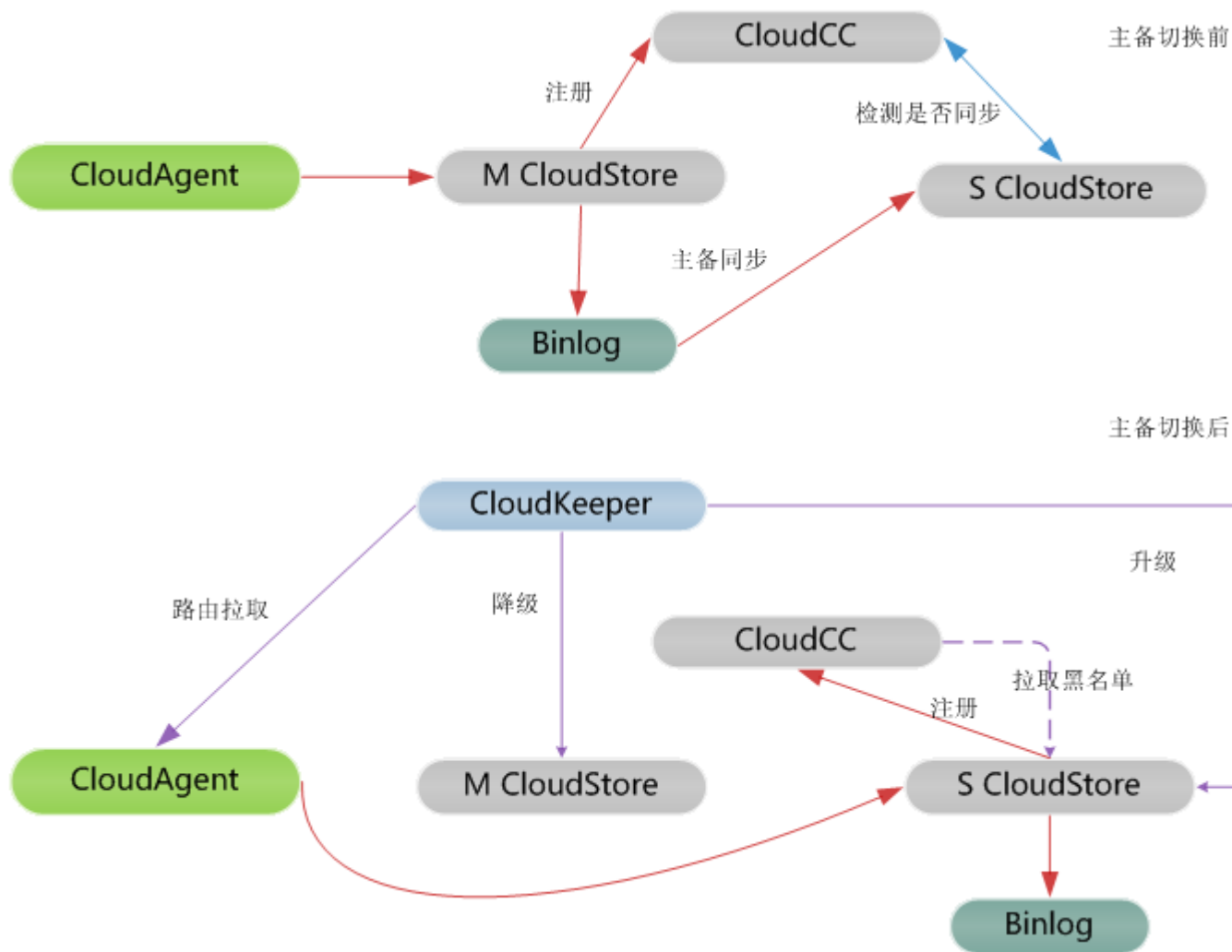
孤岛情况



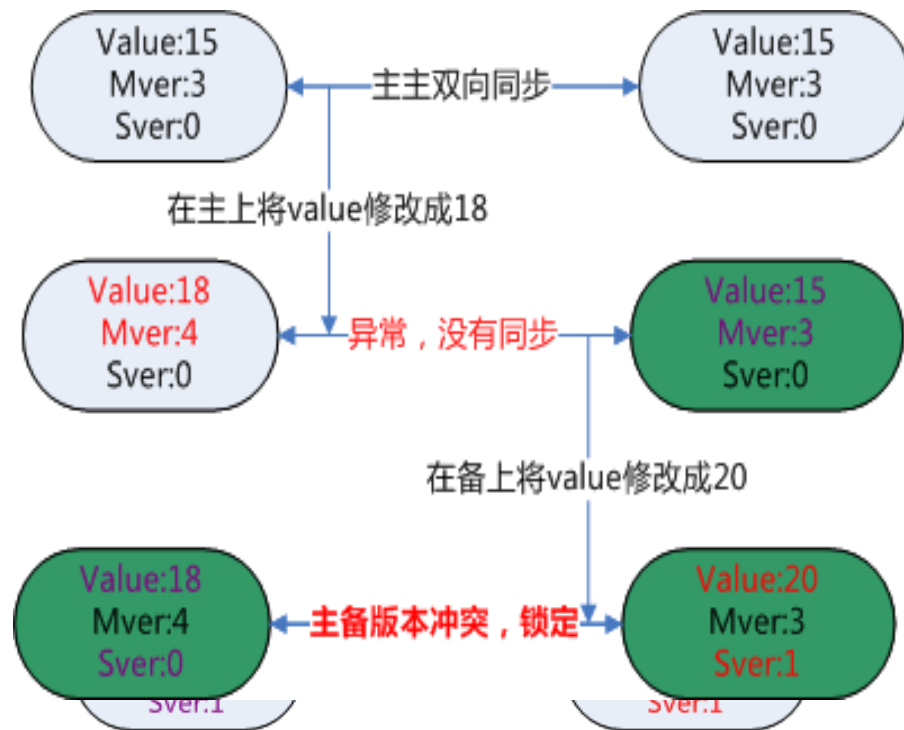
孤岛判定：

只有CloudKeeper通过两个CloudStore都无法获得健康的状态后才会判定该CloudStore已经形成孤岛，启动切换流程

容灾切换——切换流程

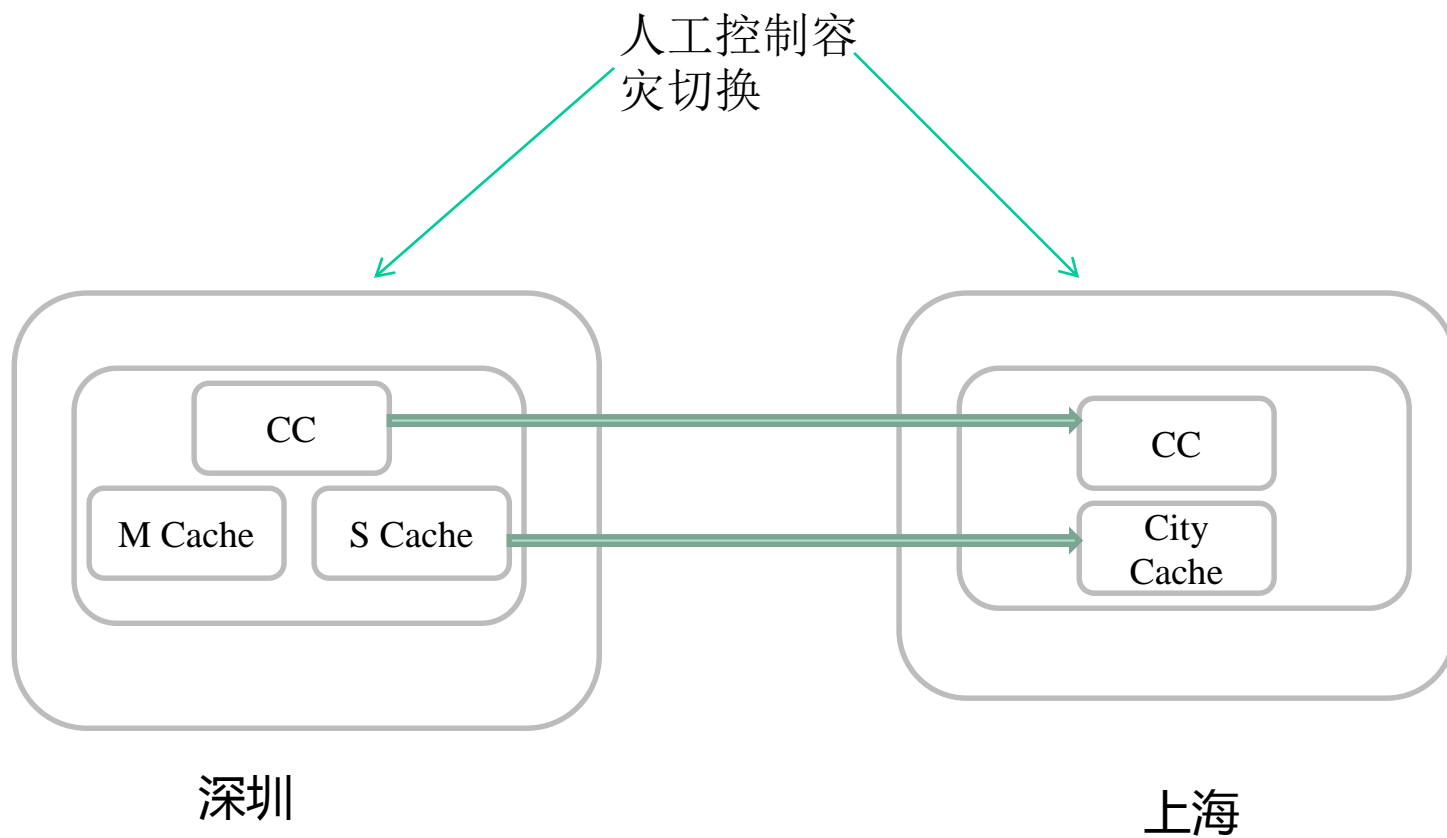


容灾切换——主备双版本号

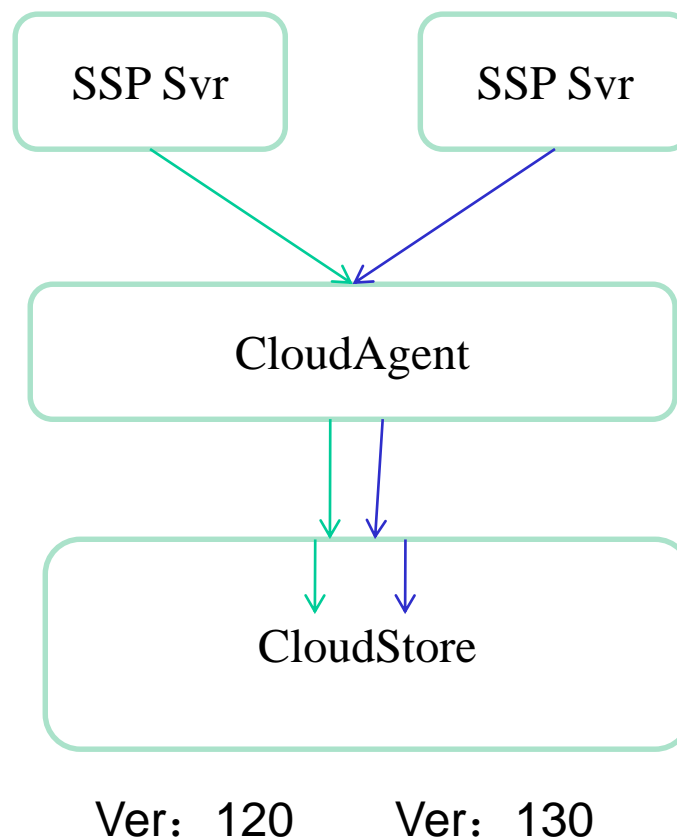
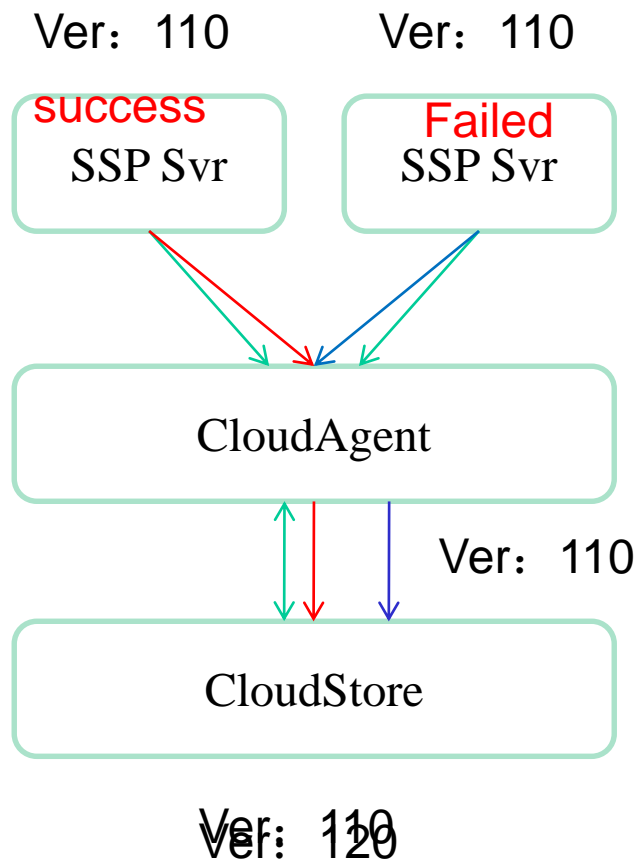


1. 便于跟踪数据在主、备上的更新流水
2. 如果主备因为各种未知bug出现错乱，同步的时候会出现版本号冲突，能锁定异常的帐户并容易修复

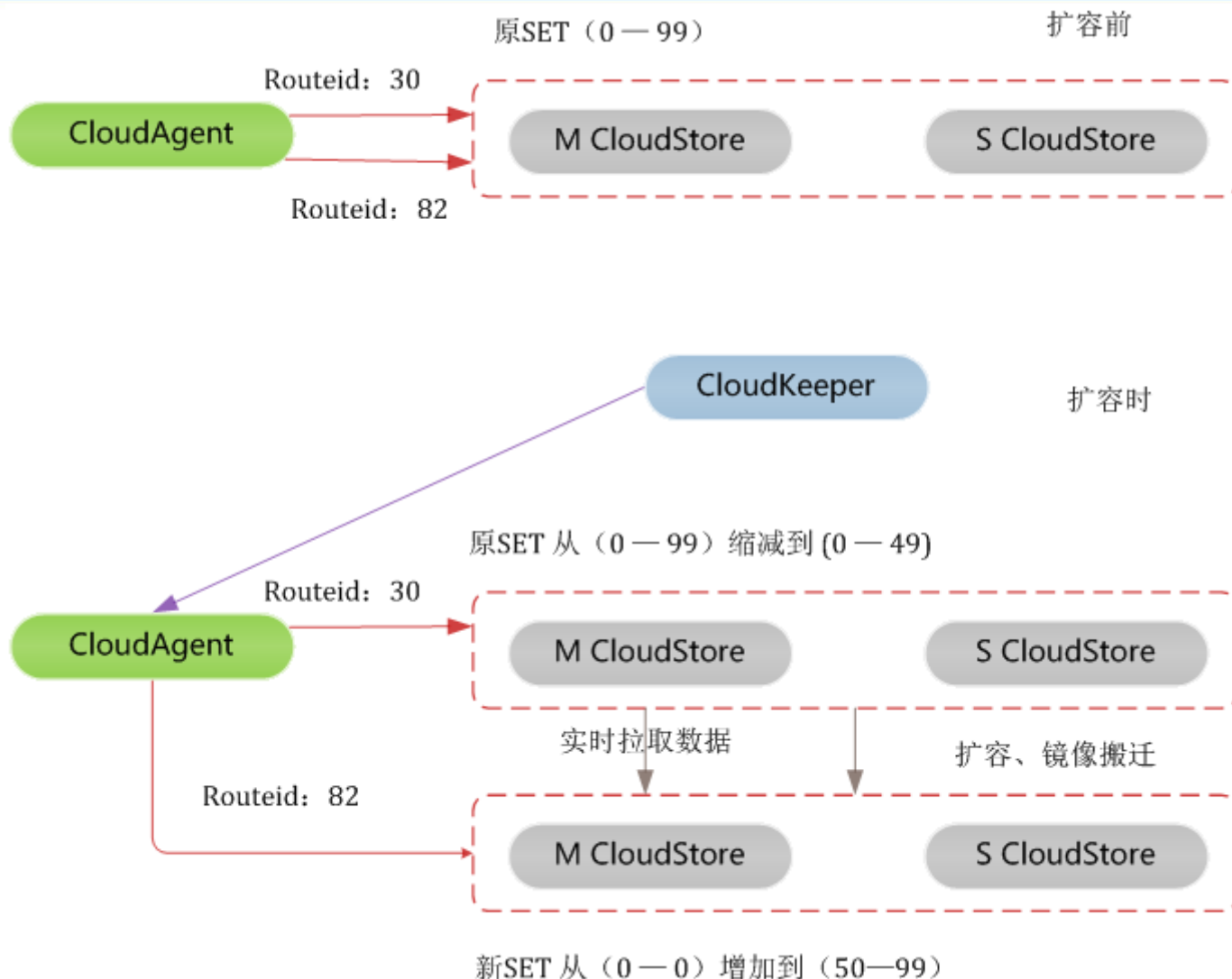
跨城容灾



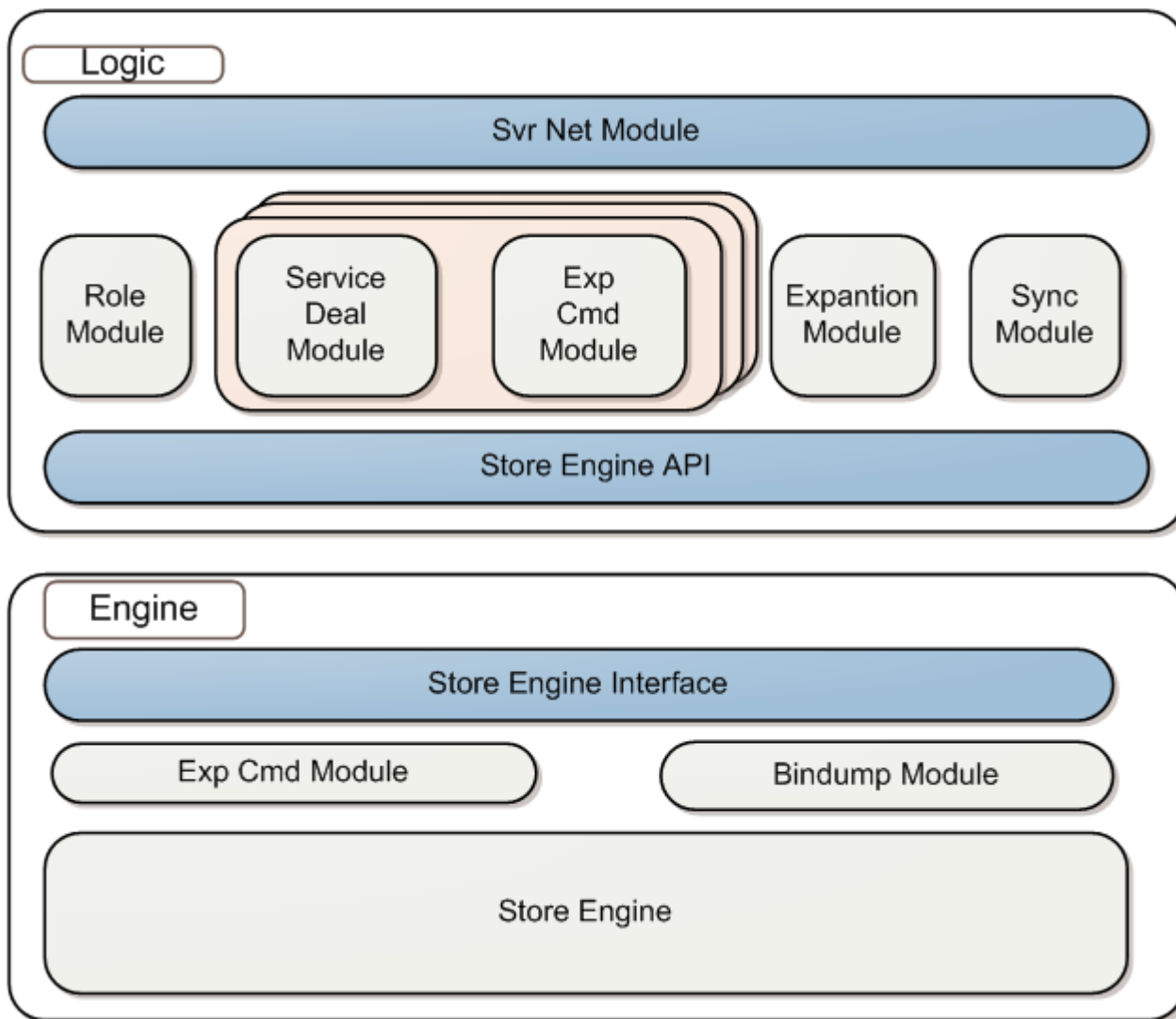
扩展命令——解决高并发串行化问题



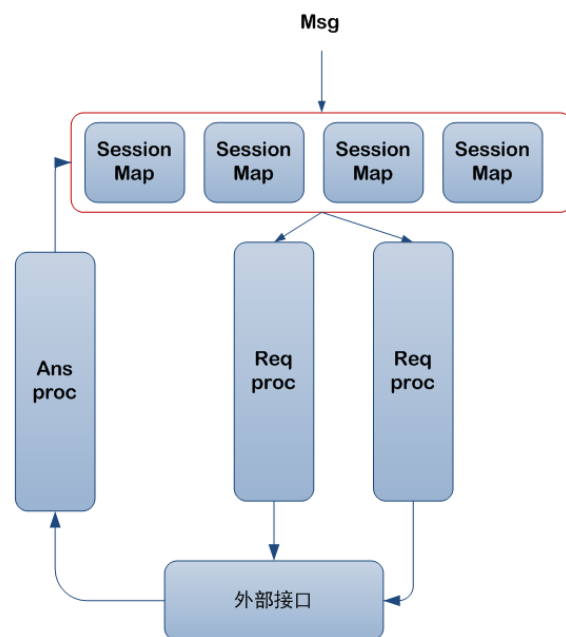
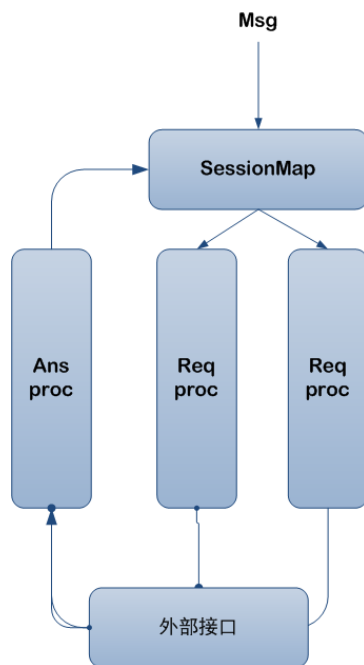
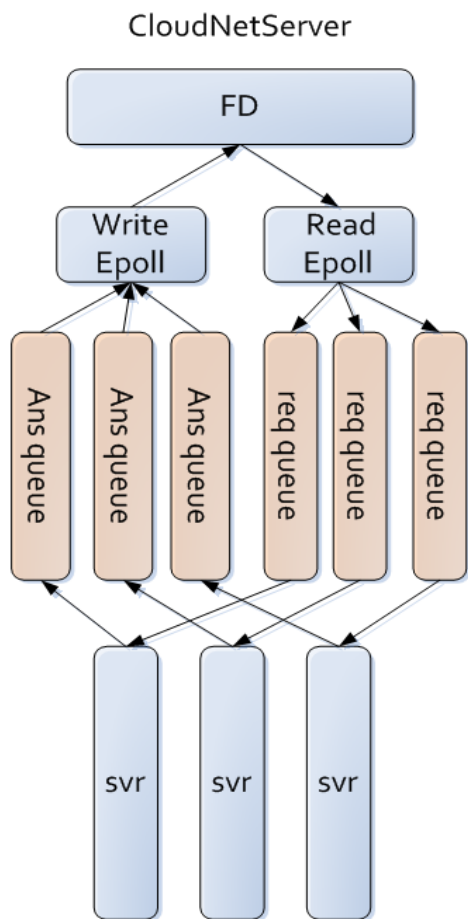
自动扩容



性能提升



性能提升—减少锁竞争



性能提升

- **减少内核态调用**

- 用自写函数替代gettimeofday

- 使用原子所 atomic.h

- **多应用无锁编程的思想**


- **部分耗时的函数重写，比如snprintf,inet_pton之类**

- **内存管理部分用tcmalloc代替标准的ptmalloc**

- **阻塞操作全部异步化**

关于性能

 B6(2CPU8核2.1G , 64G内存) , 4实例 , 150B

 高一致性

项目	处理(笔/s)	耗时(ms)	CPU耗用
Get	300K	8.164	78.40%
10:1混合	230K	9.627	90.38%

总结

主要支持服务

1 : 1主备组合、镜像、binlog、冷备保证数据安全
用户感知不到数据底层的容灾切换和数据扩容
高性能，不在为并发和耗时而烦恼


未来

更多的存储引擎


-  LevelDB

-  MySQL (SSD、fusion io)

标准化的集群管理

-  标准化配置

-  完善的运维前台管理

-  丰富的监控数据展示

Q&A
Thank you all !

欢迎莅临

2013中国数据库技术大会

Database
BDaaS
flowingdata
DB2
NoSQL MySQL
Oracle Big Data