DTCC

2016中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2016

数 据 定 义 未 来

SequeMedia 盛拓传媒　IT168.com　ChinaUnix　ITPUB

360云盘底层Cassandra+Storm介绍

郭东东、倪传雷

- Cassandra+Storm整体架构及现状
- Cassandra系统改进及实践
- Storm系统改进及实践

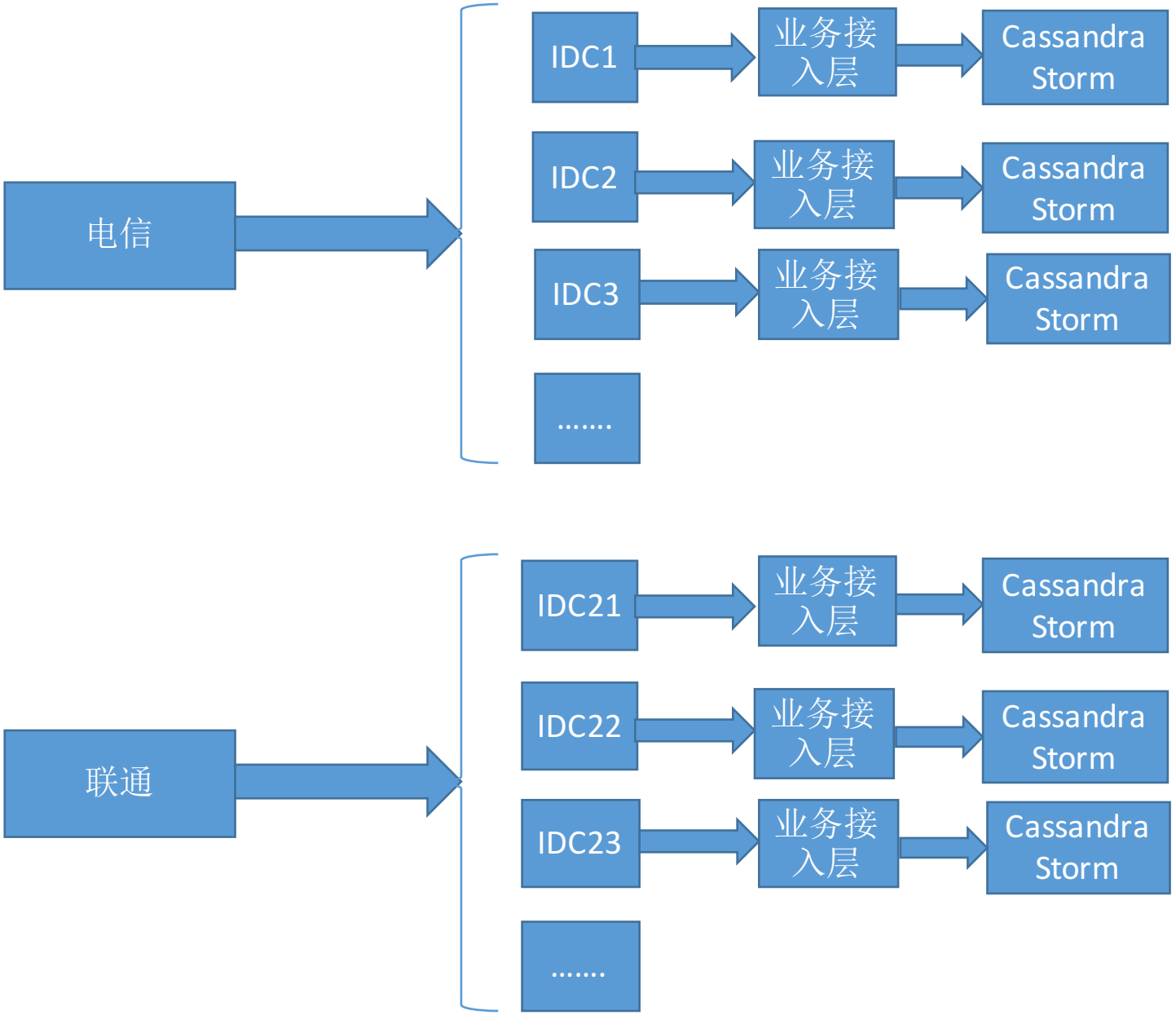
Cassandra+Storm集群现状

- 集群规模

主机规模	备份规模	数据规模	单日新增
13000台	8000台	400PB	500TB

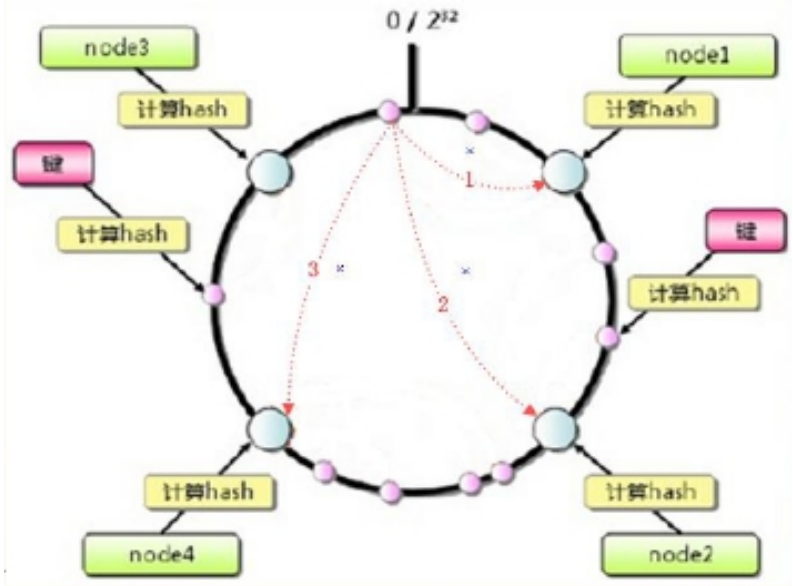
单集群规模	集群容量
150台（24/3T）	9P
300台（24/3T）	18P
……	……
600台（36/4T）	**79P**

Cassandra+Storm集群现状



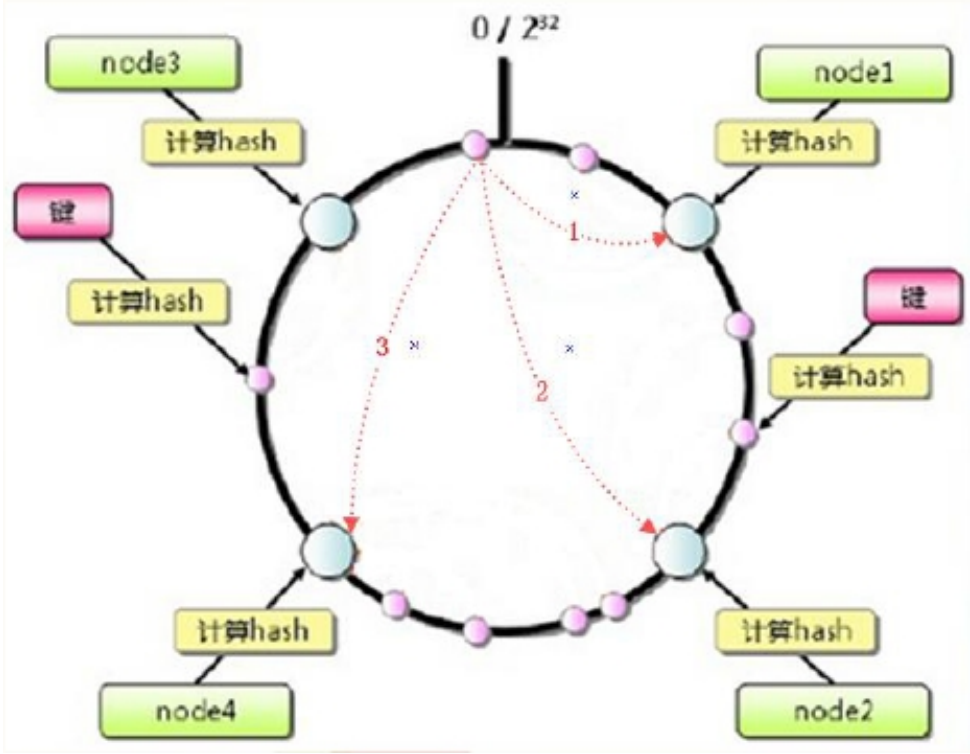
Cassandra系统改进-数据可靠性

- 主要问题：
  - 扇区、磁盘故障、主机故障导致副本缺失
  - 新写入数据副本可能不足（ONE/QUORUM）
  - 系统自带机制不能保证副本及时修复：
    - 读修复、Hinthandoff、Repair操作局限性
    - 损坏的SSTable在内存索引中，但磁盘数据读异常



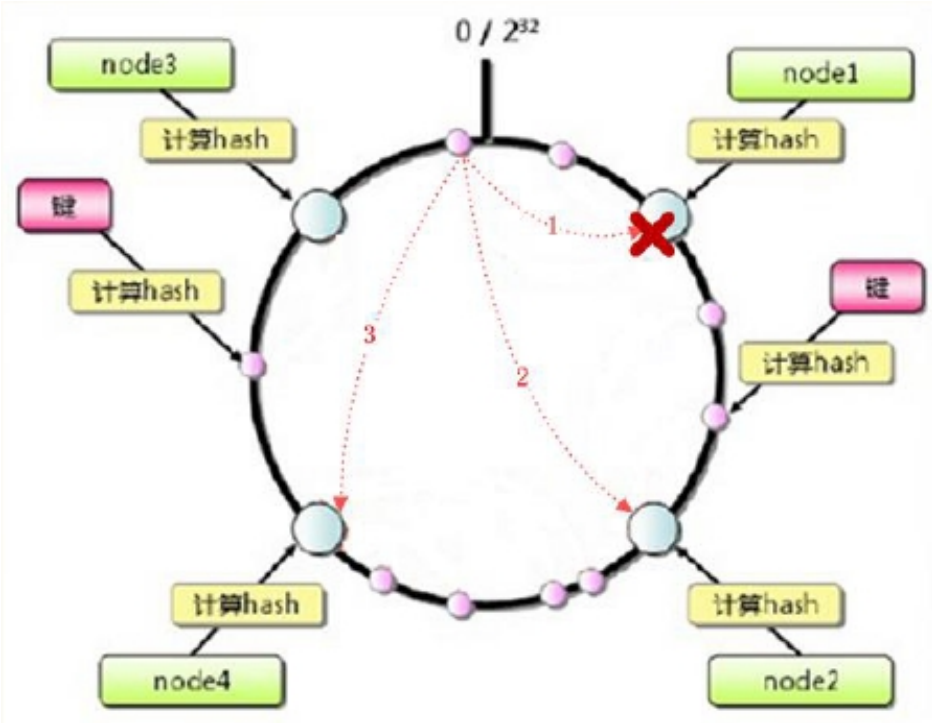
Cassandra系统改进-数据可靠性

- 文件/磁盘自动摘除
  - 目的：
    - 去腐生肌
    - 消除影响
  - 基于统计
    - 文件异常访问次数
    - 摘除文件比例



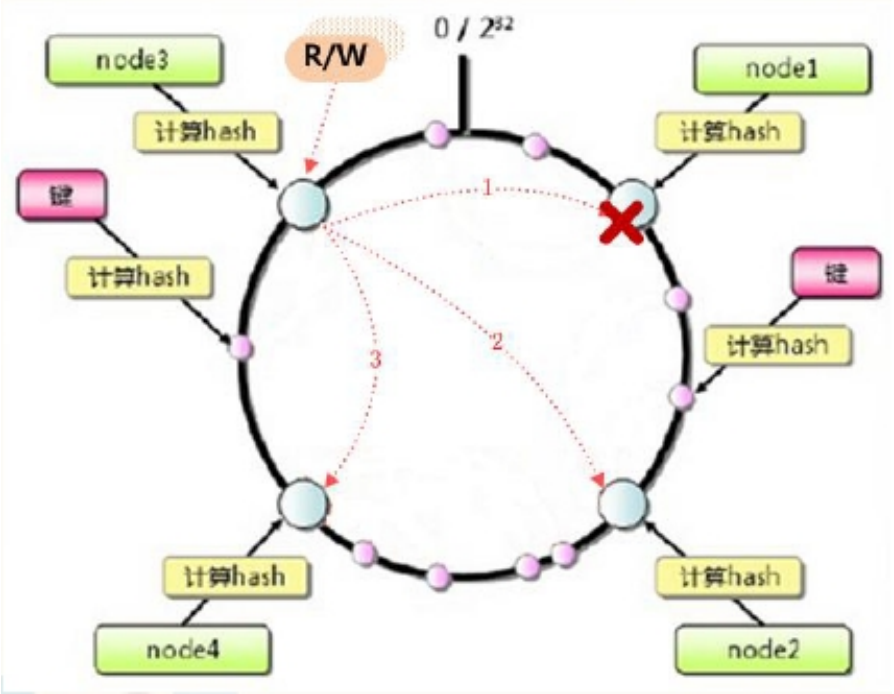
Cassandra系统改进-数据可靠性

- 数据节点定期扫描修复
  - 目的：
    - 磁盘/文件故障摘除即启动修复
    - 尽快恢复全副本的状态
  - 修复方式：
    - 确定故障所属Range
    - RowScan + Diff
    - KeyScan + Read（ALL）



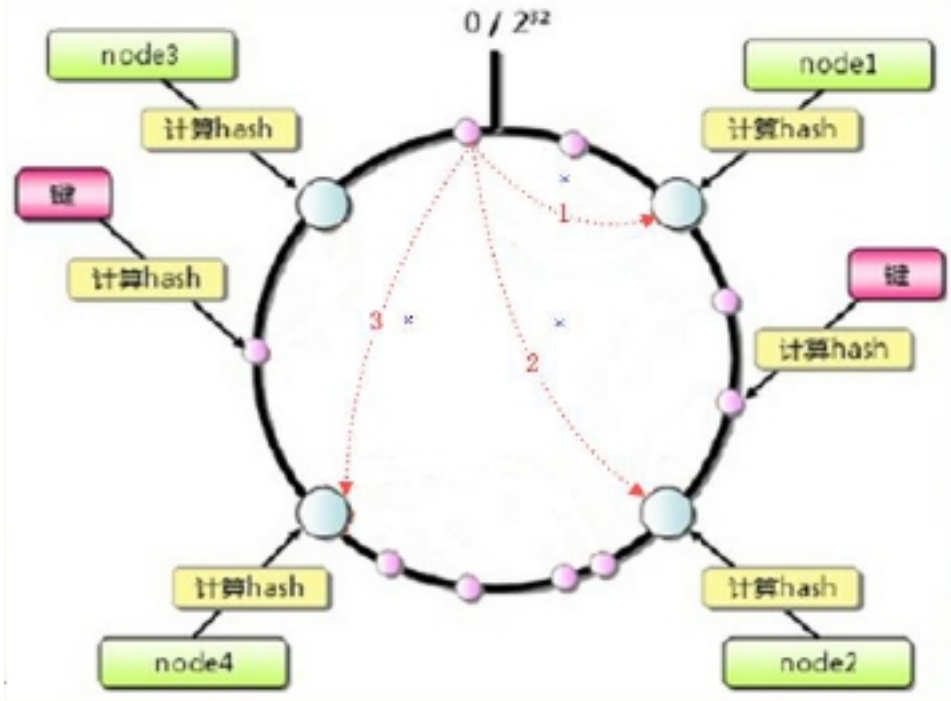
Cassandra系统改进-数据可靠性

- 增量数据的检查修复
  - 目的：
    - 保证新写入数据副本数足够
    - 解决hinthandoff缺点
  - 处理方式：
    - 新增辅助表：proxycheck
    - 副本不足记入辅助表
      - 数据节点写失败：超时/拒绝
      - 数据节点停机
    - 读修复



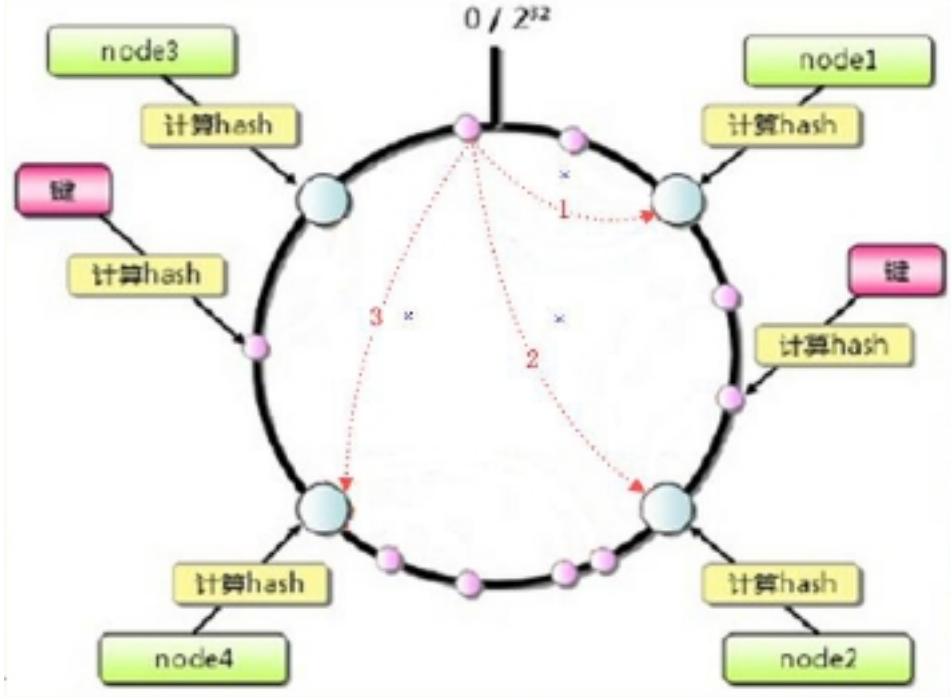
Cassandra系统改进-数据分配策略

- 主要问题：
  - 基于集群的数据Partition策略，不灵活
  - 不同Keyspace根据数据类型及访问需求，需要不同的Partition策略
  - 部分KeySpace 有范围Scan的需求



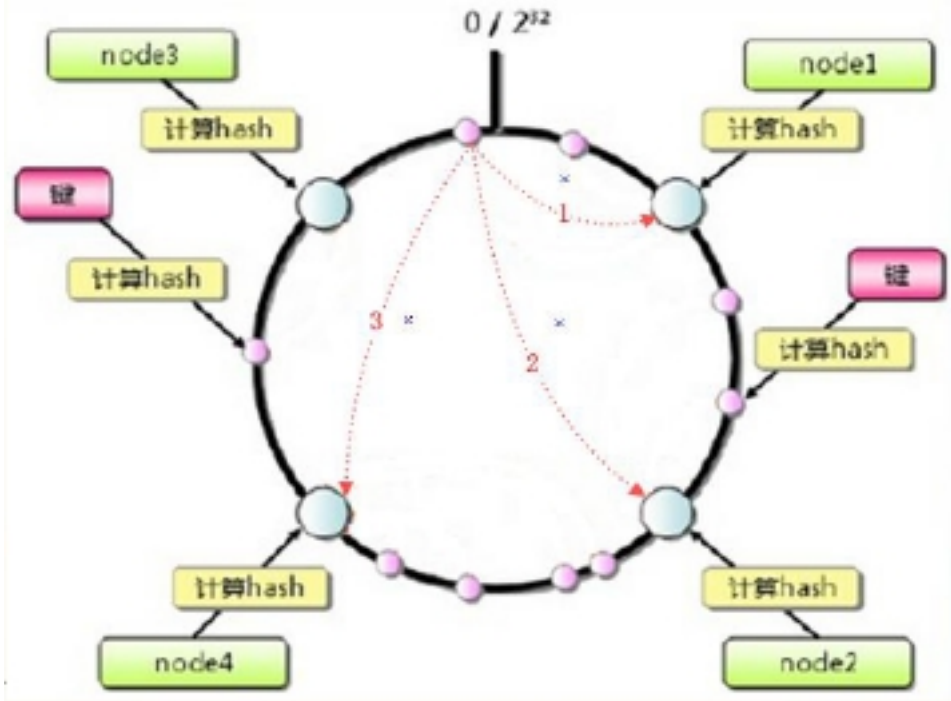
Cassandra系统改进-数据分配策略

- 基于KeySpace的Partition策略：
  - 支持RandomPartitioner和ByteOrderedPartitioner
  - 修改表的Meta信息存储，增加Partition策略
  - 增加不同Partition的Token对应关系
  - 修改访问接口内部实现，实现不同Partition的兼容性



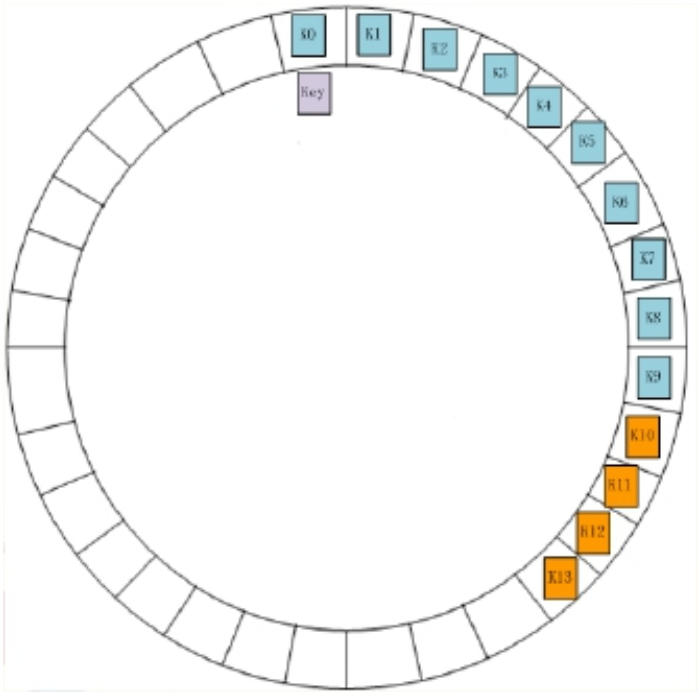
Cassandra系统改进-EraserCode

- 主要问题：
  - 数据规模原来越大，储存成本越来越高
  - SimpleStrategy ,NetworkTpStrategy 存储3副本成本太高



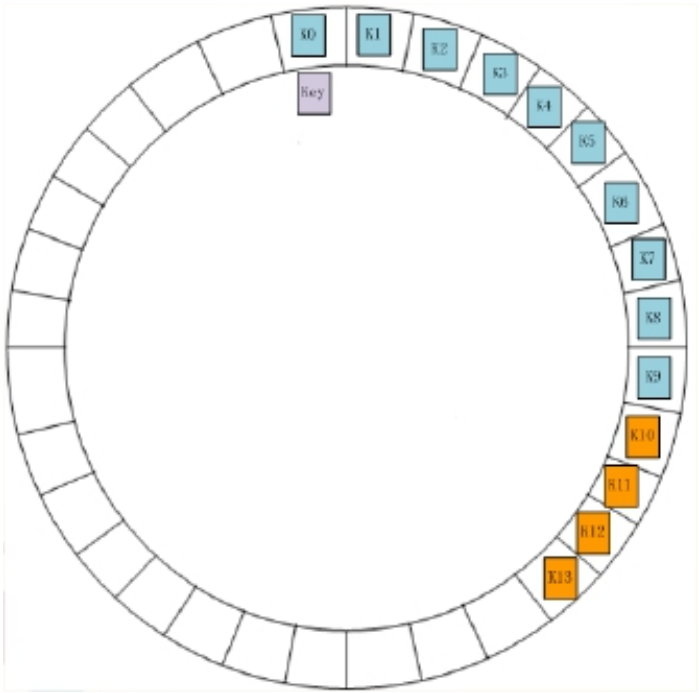
Cassandra系统改进-EraserCode

- 基于Stripe的EC存储策略：
  - 数据切分成N段子数据
  - N段子数据计算出P个校验值
  - N+P段数据依序存储在环上



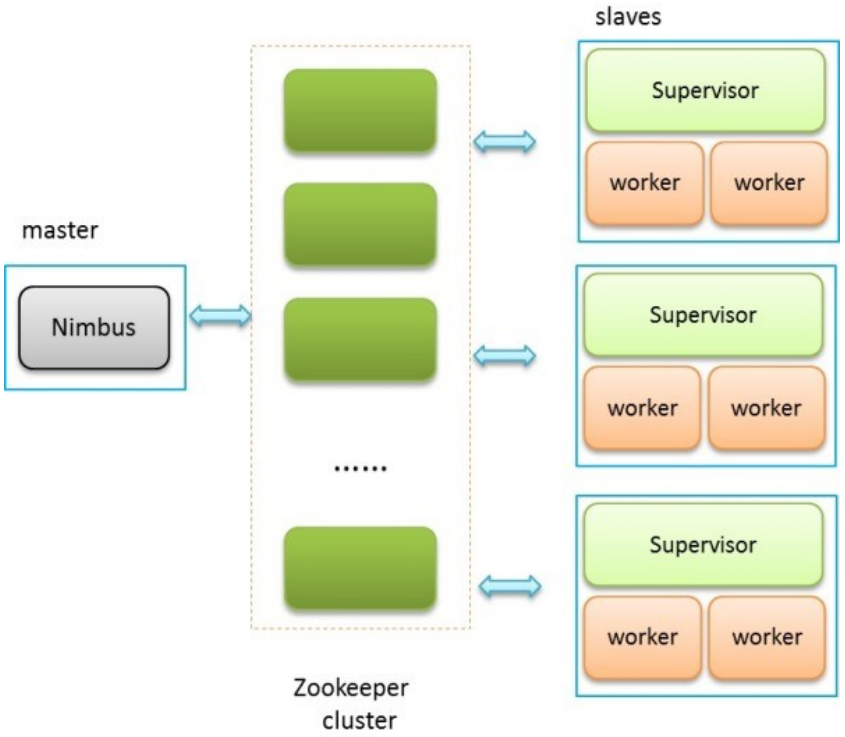
Cassandra系统改进-EraserCode

- EC存储数据切分及存储原理：
  - 数据内容按照大小切分成N个字段
  - 子段的Key，有前缀+Value的Md5
  - 原始Key只存储子Key列表



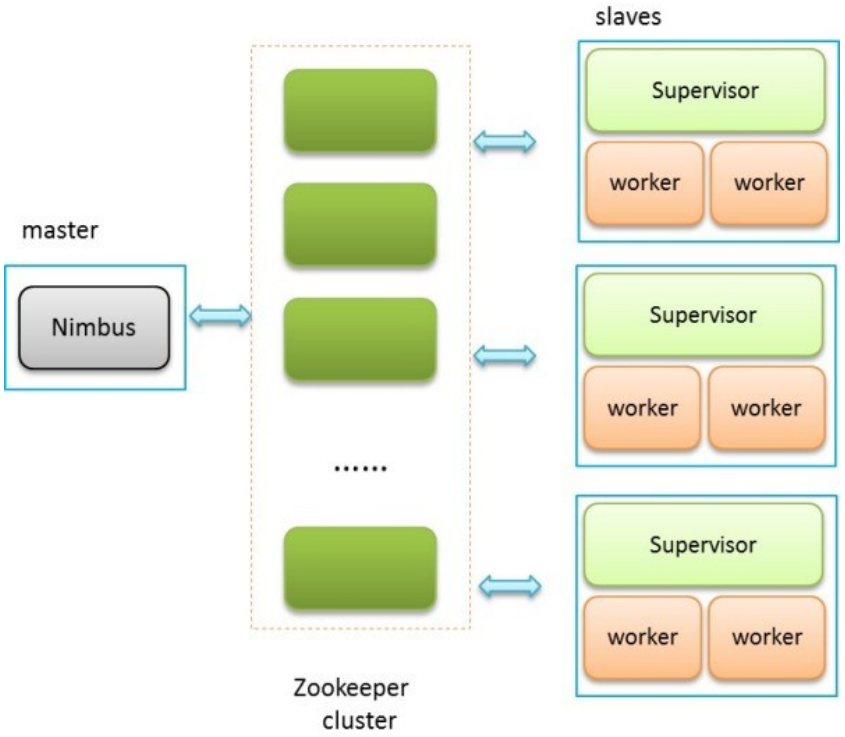
Storm系统改进-大文件缓存

- 主要问题：
  - 部分Storm topology依赖数据量较大，如机器学习模型
  - 规模达数百兆，导致topology启动时间过长
  - 数据变化不大



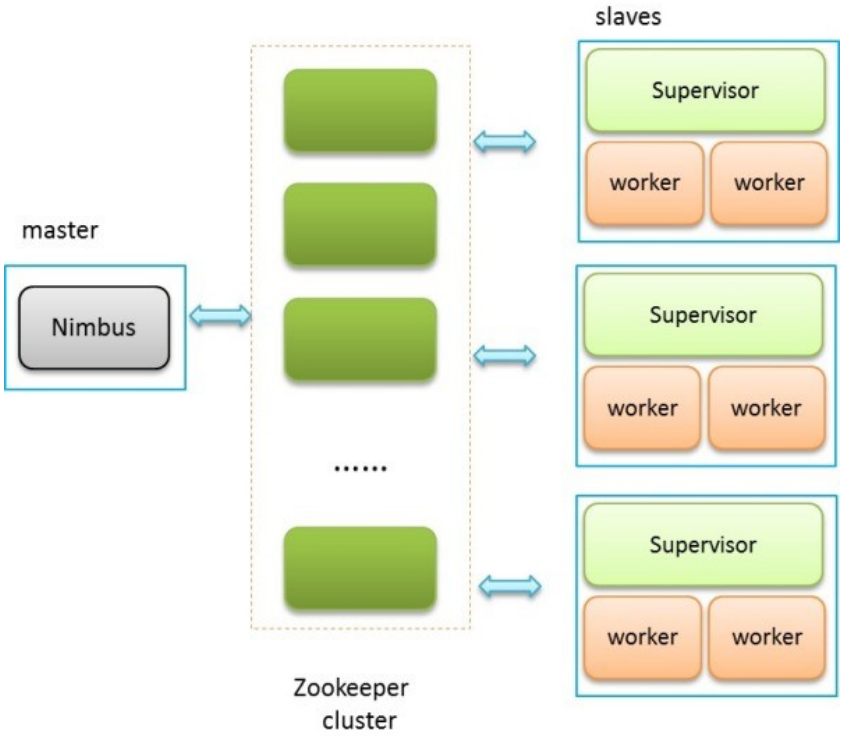
Storm系统改进-大文件缓存

- 大文件缓存：
  - 大文件采用类似Jar 管理方式，由supervisor管理
  - topology下线之后并不立即删除，长时间不用才释放



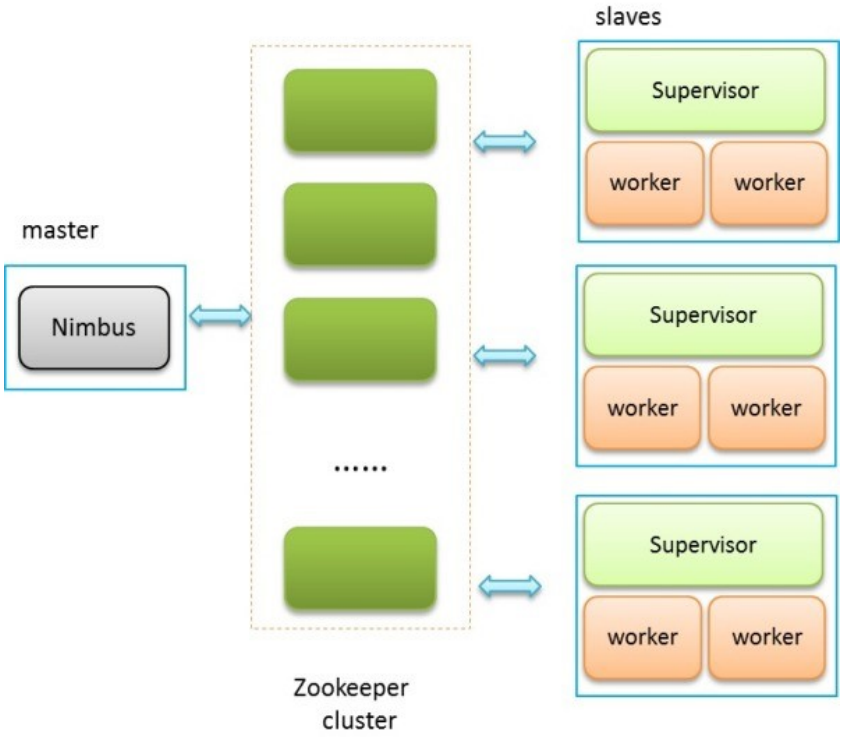
Storm系统改进-应用Jar包P2P分发

- 主要问题：
  - topology的jar包比较大
  - worker 数量在上千台服务器上
  - topology启动耗时比较长



Storm系统改进-应用Jar包P2P分发

- P2P软件分发：
  - Storm集成P2P分发软件功能
  - Nimbus到Supervisor节点软件包的快速分发



THANKS

SequeMedia 盛拓传媒 IT168.com ChinaUnix ITPUB