大规模存储系统的持续研发

刘海锋

www.jd.com

JD.COM 京东

Overview

JD.COM 京东

- 京东文件系统 - JFS
- 分布式缓存与高速KV服务 - Jimdb

Jingdong File System (JFS)

JD.COM 京东

- 愿景
  - The unified datacenter storage infrastructure (2013/7 – now)
- 小步快跑，分期开展
  - 海量小文件
  - 对象存储
  - 块存储
  - 新图片系统
  - 元数据表结构存储 (wip)
  - Hadoop集成 (wip)

海量小文件 – 需求与挑战

JD.COM 京东

- 商品订单
  - 365 * 数亿 * ~10KB
- 商品图片
  - 几十亿 * (20 ~200KB)
- 库房记录
  - 365 * 十亿 * (KB ~ MB)

- 各种方案
  - 关系数据库
    - Pains – 难以扩容、定期删除
  - 开源存储系统
    - Pains – 选型、维护、定制

自主研发

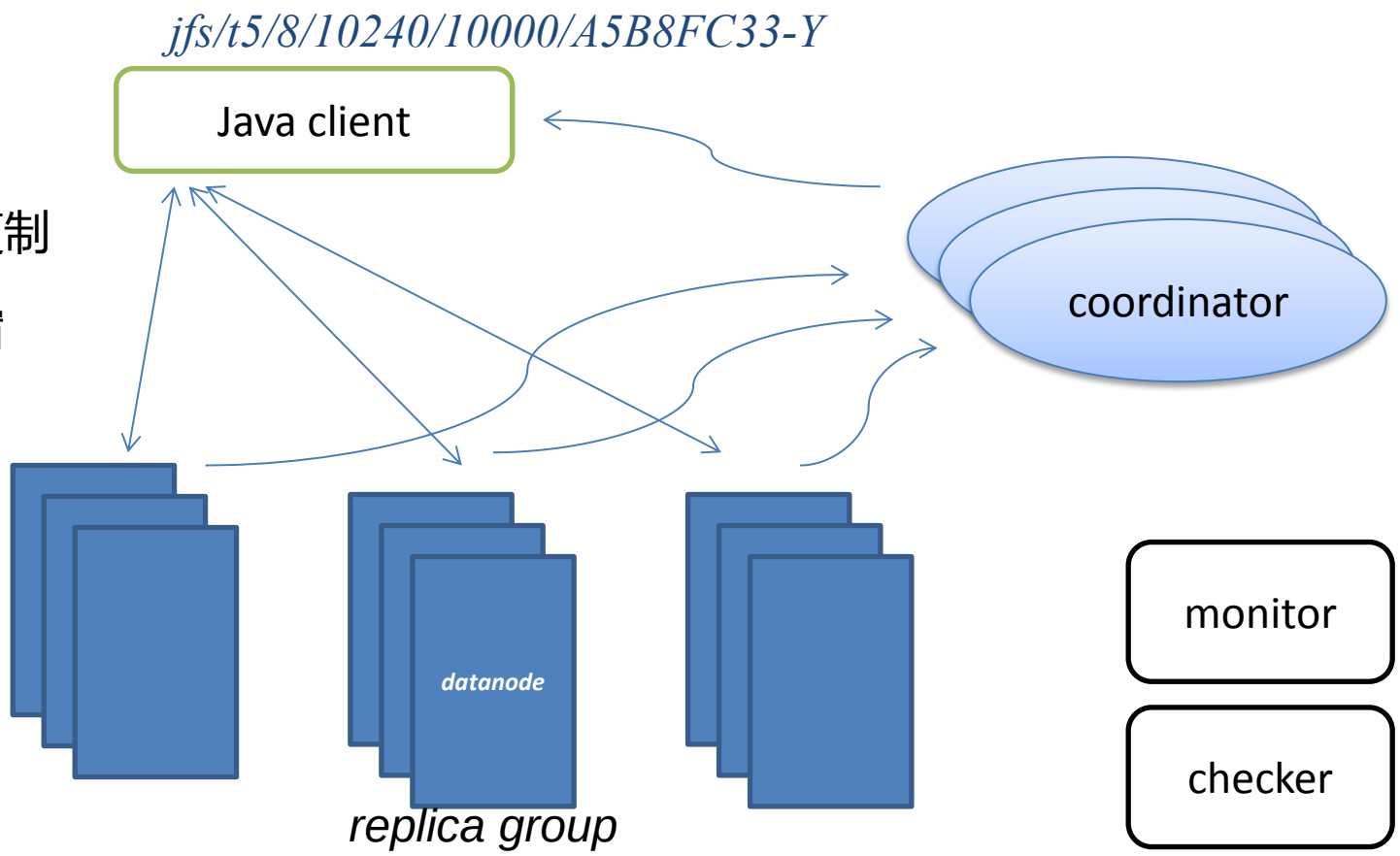
JD.COM 京东

- 规模推动自主研发
  - 拿来主义 → 开源定制 → 自研
  - 灵活可控、长期技术收益
- 挑战
  - 开发周期，稳定性，长期性
  - 小投入大产出
- 策略
  - 紧扣业务需求，高度定制，分期开展
  - 独辟蹊径，专注

JFS – 海量小文件

JD.COM 京东

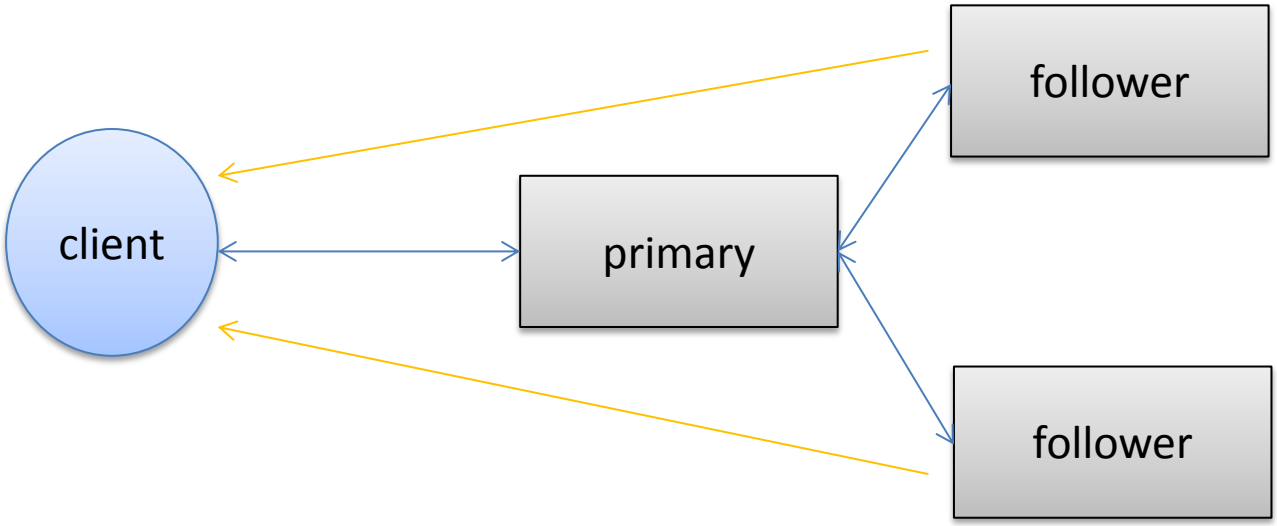
- 需求驱动
  - 在线数据多为小文件
- 系统特性
  - 系统命名
  - 强一致复制
  - 透明压缩



复制协议

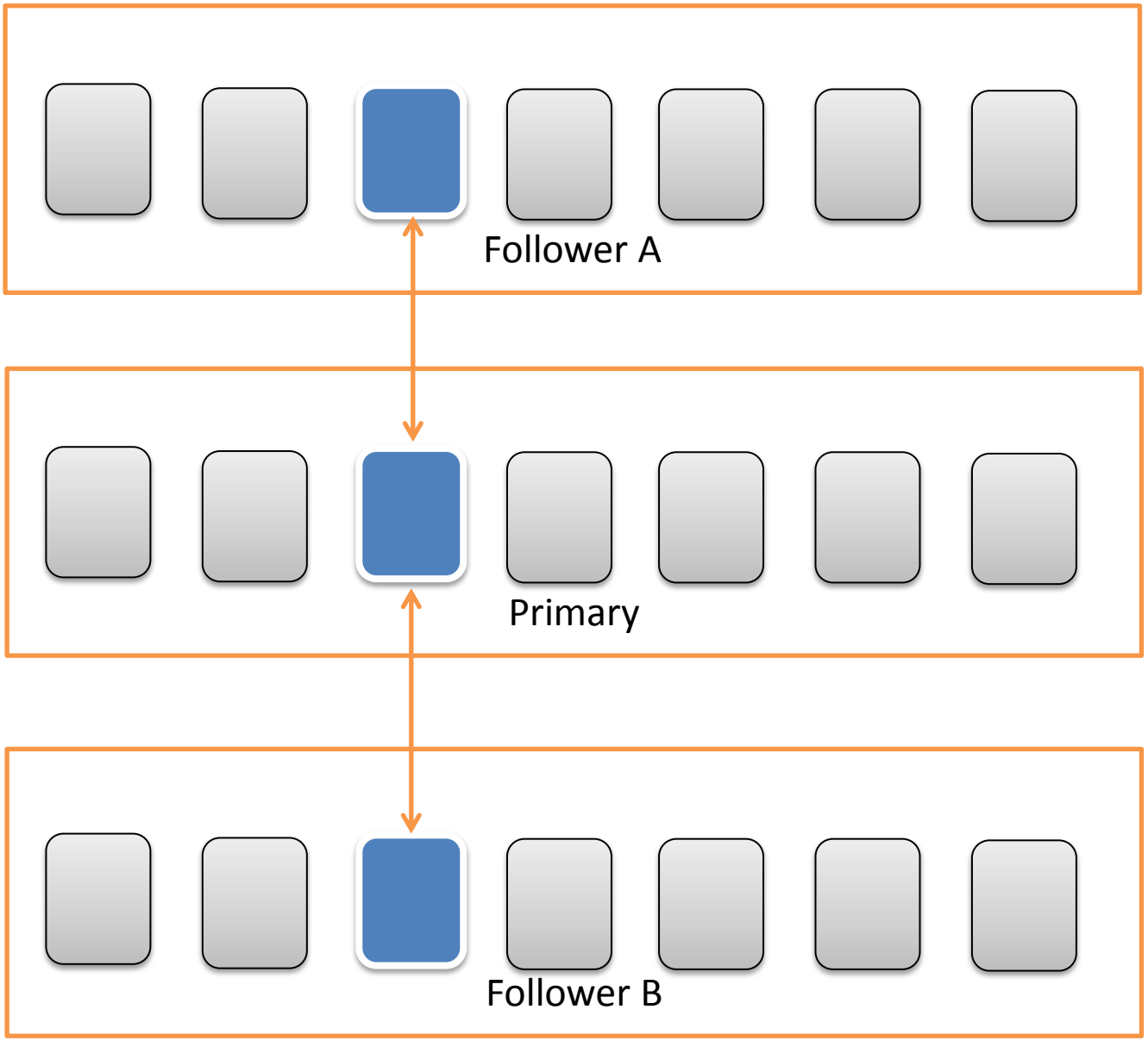
JD.COM 京东

- Paxos算法变体
  - 固定成员角色 – one primary + 2 followers
    - 不做majority-based leader election
  - Full-quorum replication
    - 二元状态机 - ReplGroupReady or ReplGroupSplit



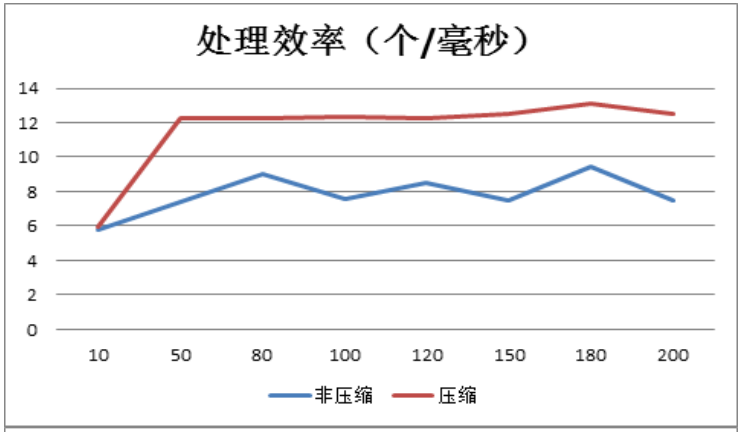
存储引擎

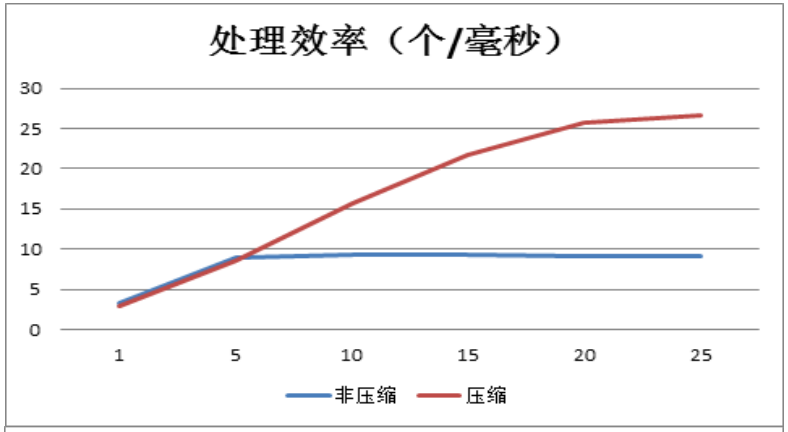
JD.COM 京东

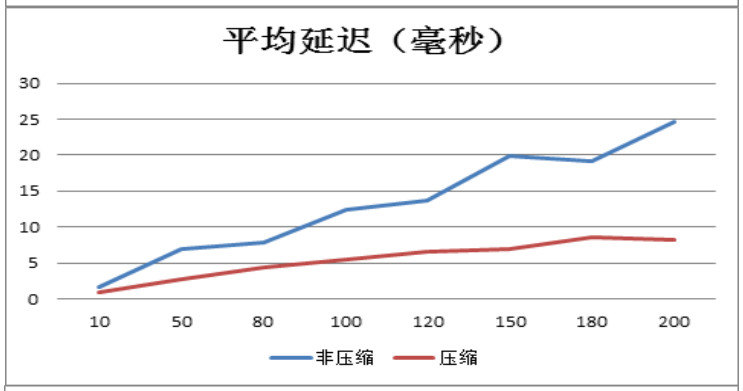


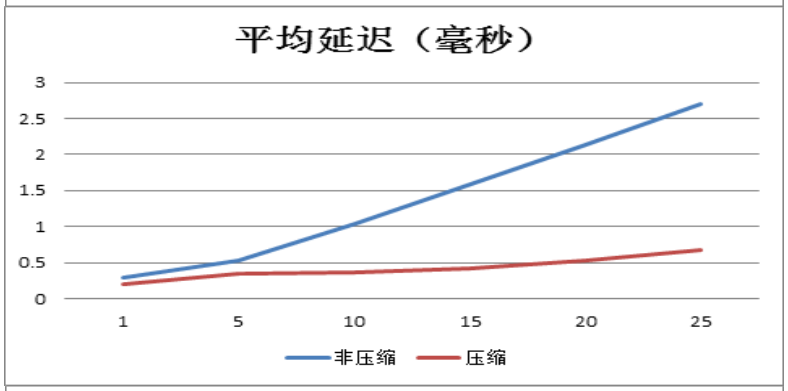
性能数字

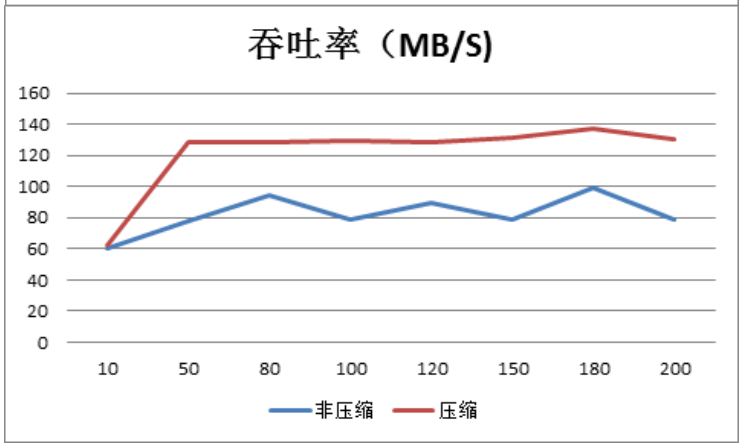
JD.COM 京东

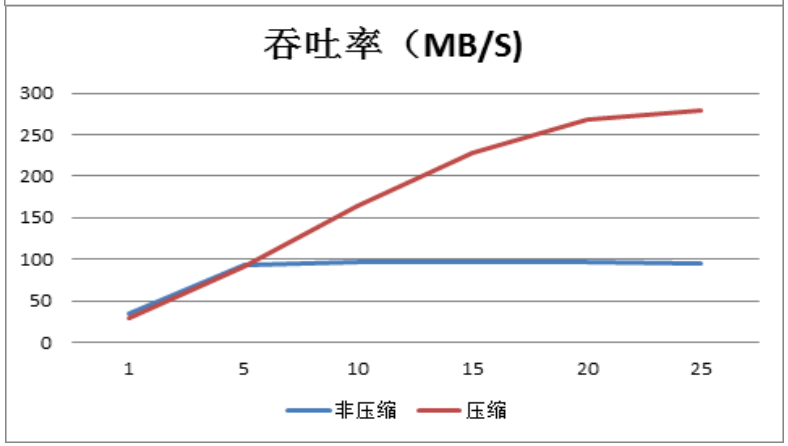












相关工作

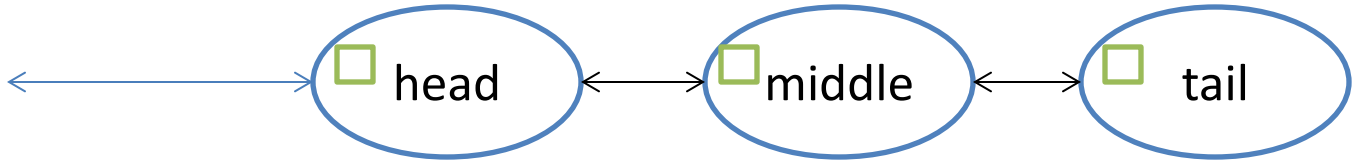
JD.COM 京东

- 类似系统
  - Facebook’s Haystack
  - Taobao’s TFS
  - FastDFS、Weed-FS、...
- JFS v1
  - 更重要的数据
  - 强一致性
  - 无单点故障
  - 无内存索引
  - 透明压缩, et al.

JFS – 大文件流式读写

JD.COM 京东

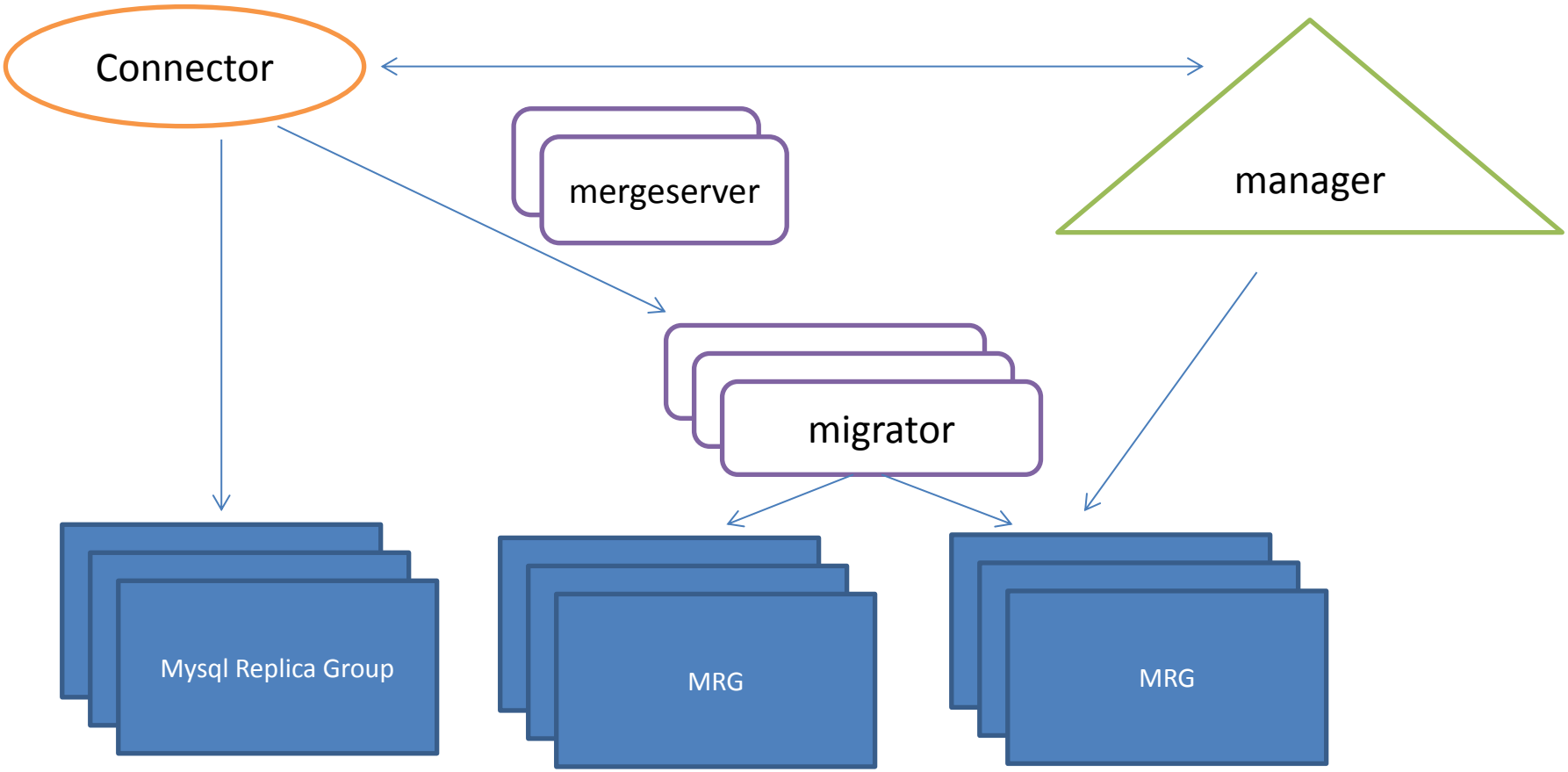
- 特别针对云存储服务（对象存储）
  - Chain Replication
  - Append-only extents
  - Pipelined write



JFS – 元数据表格系统

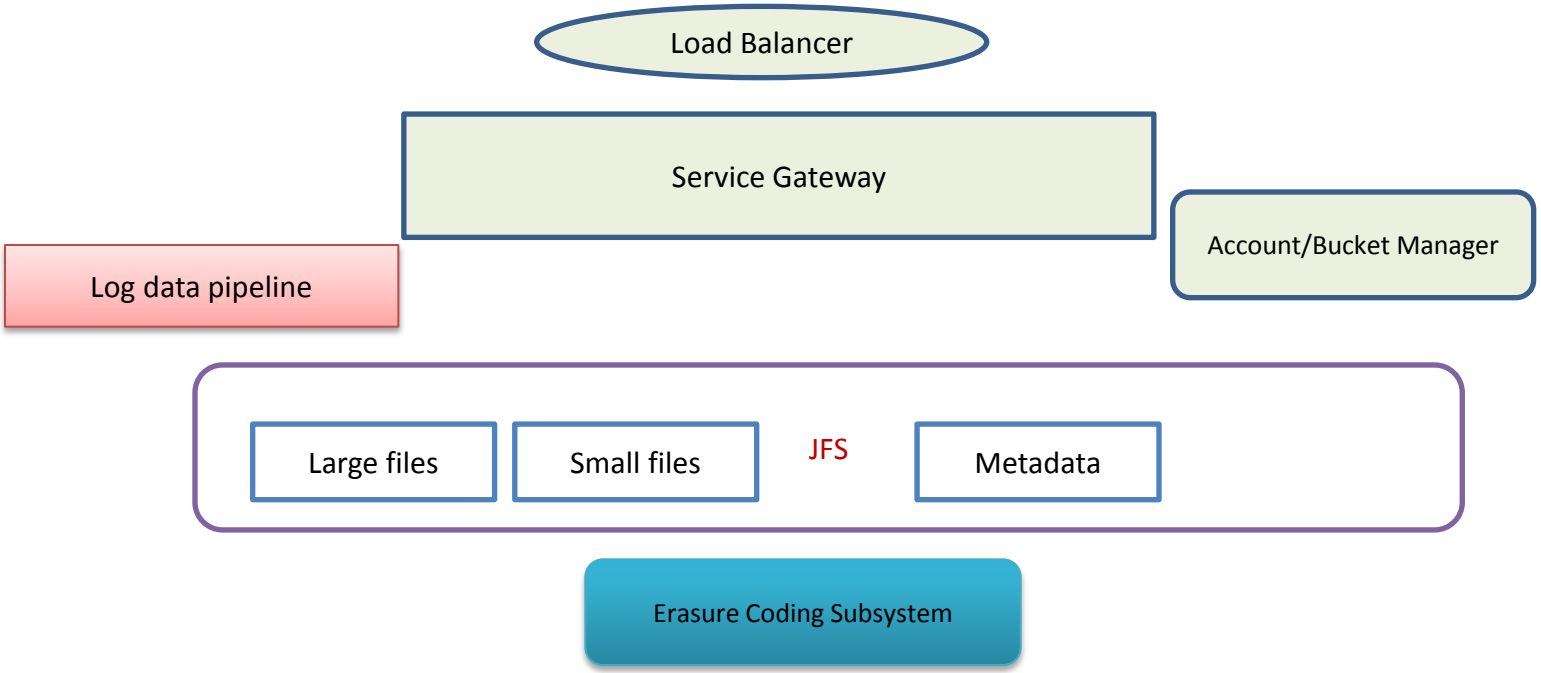
JD.COM 京东

- 分布式字典树



JFS – 云存储服务

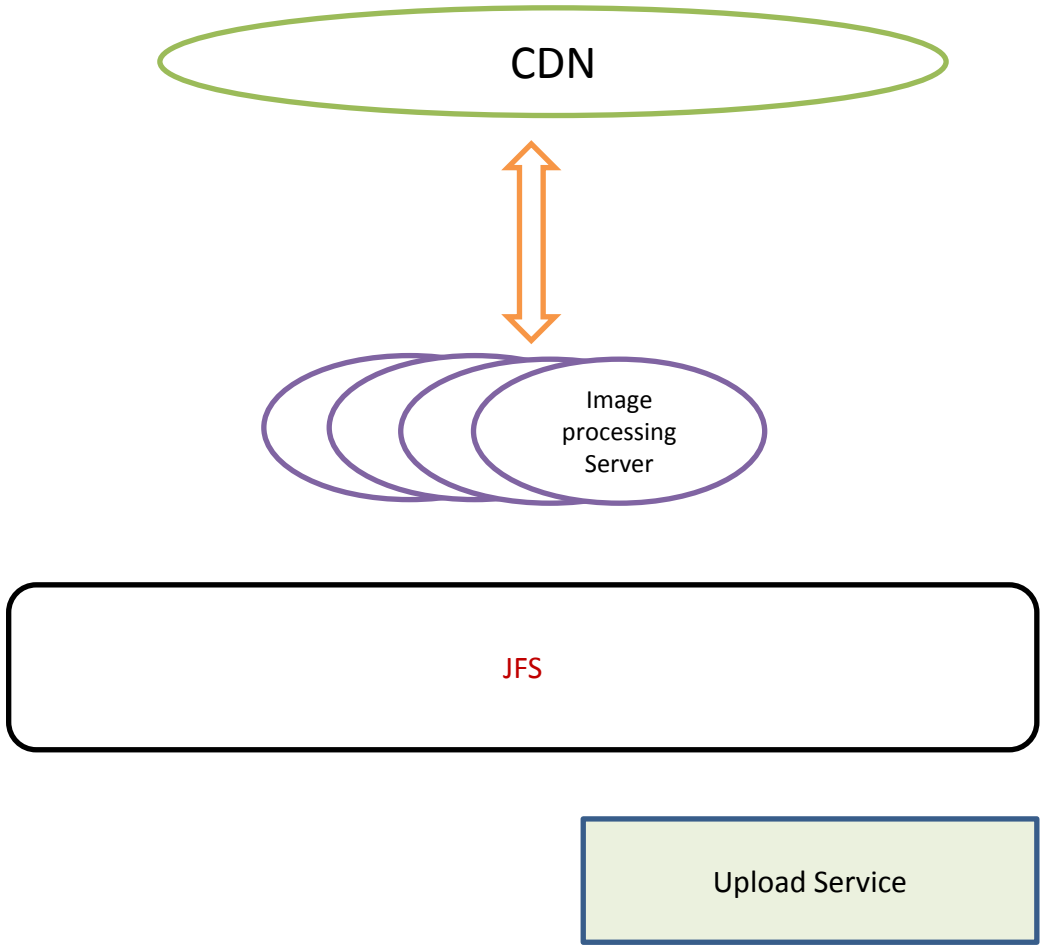
JD.COM 京东



JFS – 新图片系统

JD.COM 京东

- 重新搭建京东图片服务，从存储到展现
  - 商城主站与金融产品全部图片
- 技术
  - 基于JFS做底层存储
  - 重写在线缩放处理层



JFS – 业务影响

JD.COM 京东

- 多个集群
  - 图片、订单、仓库流水、内部云存储、公有云存储、网盘后端 ...
- 300个业务应用，PB规模

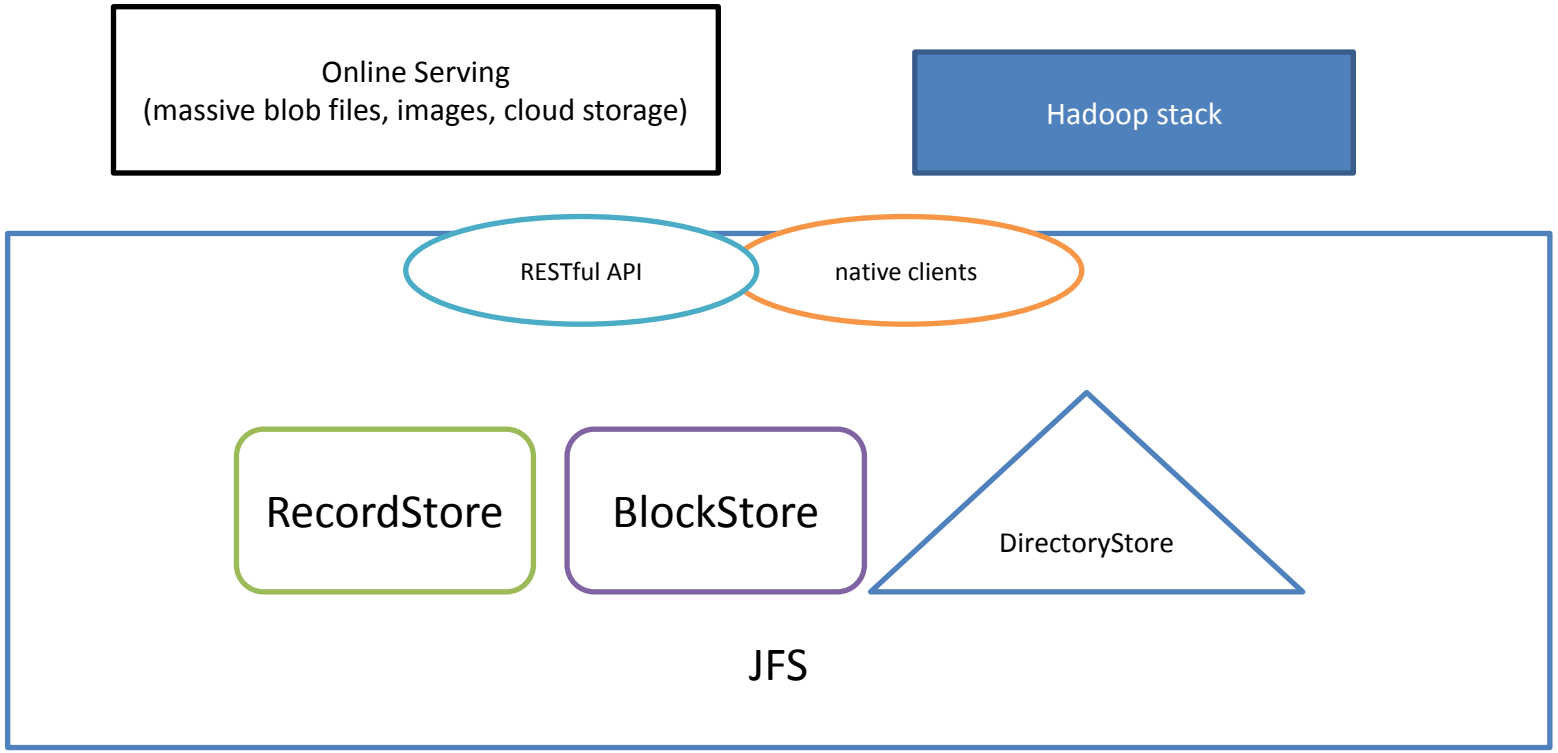
- 正在做的事情
  - 元数据表格系统二期
  - Hadoop集成
  - 多个子系统重构

JFS – the final picture

JD.COM 京东

Jingdong unified storage infrastructure for

* both small and large files
* both online serving and offline data processing



Overview

JD.COM 京东

- 京东文件系统 - JFS
- 分布式缓存与高速KV服务 - Jimdb

“Memory is the new disk.”

– Jim Gray

演进历史

JD.COM 京东

- 短平快
  - 分散管理独立Redis集群
- 统一平台
  - 服务化、自动化、完善监控
- 自主研发
  - 规模驱动、痛点驱动

No Pains, No Motivations

JD.COM 京东

- 故障检测
- 容错能力
- 内存超标
- 持久性不够
- 启动慢
- 难以扩展
- …

Jingdong In-Memory Store (Jimdb)

JD.COM 京东

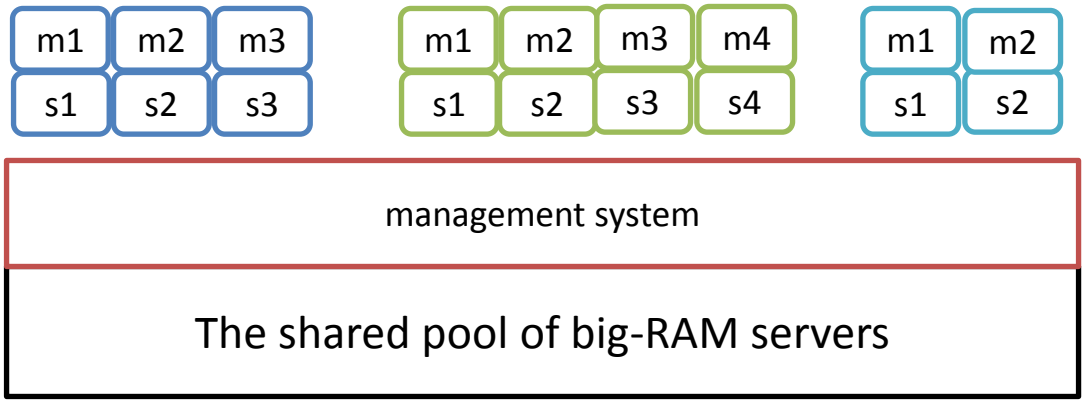
- 在原统一Redis平台基础上创新

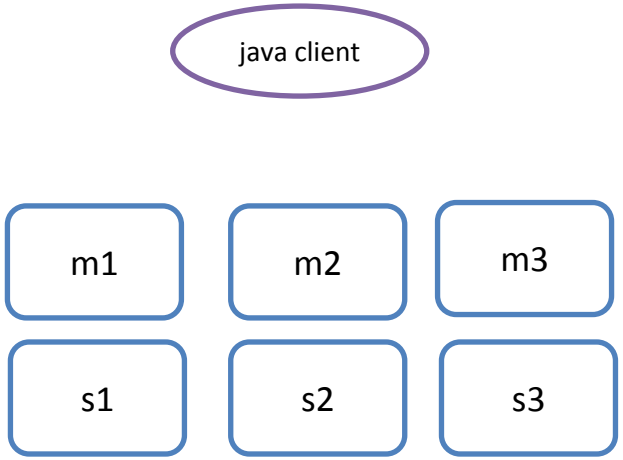
- 小步快跑，分期开展
  - 监控与运维工具
  - 精确故障检测
  - 自动故障切换
  - 两级存储层次
  - 在线纵向扩展
  - 在线横向扩容

规模

JD.COM 京东

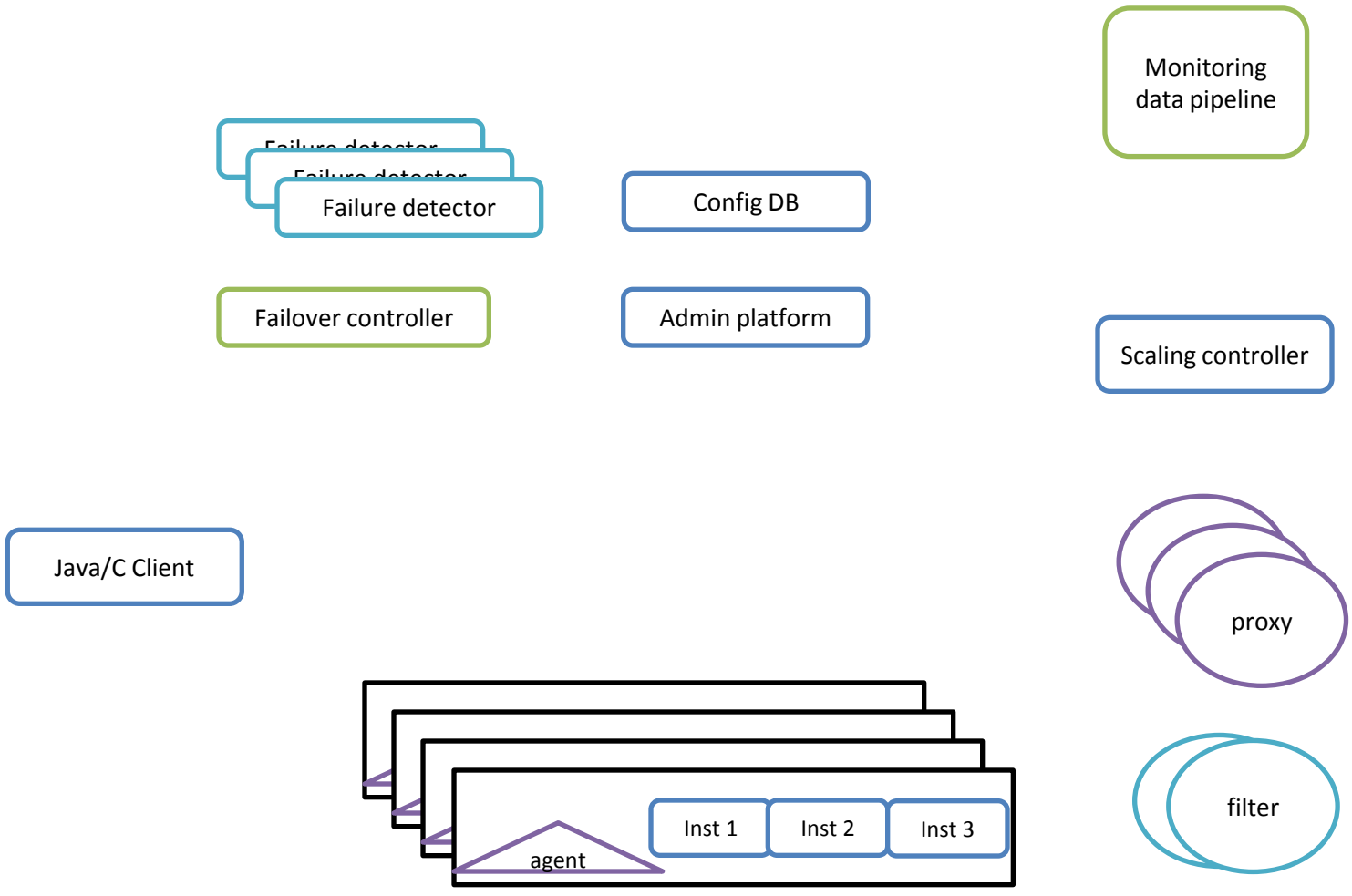
- 5500 Redis实例、1200服务器、400集群





Jimdb components

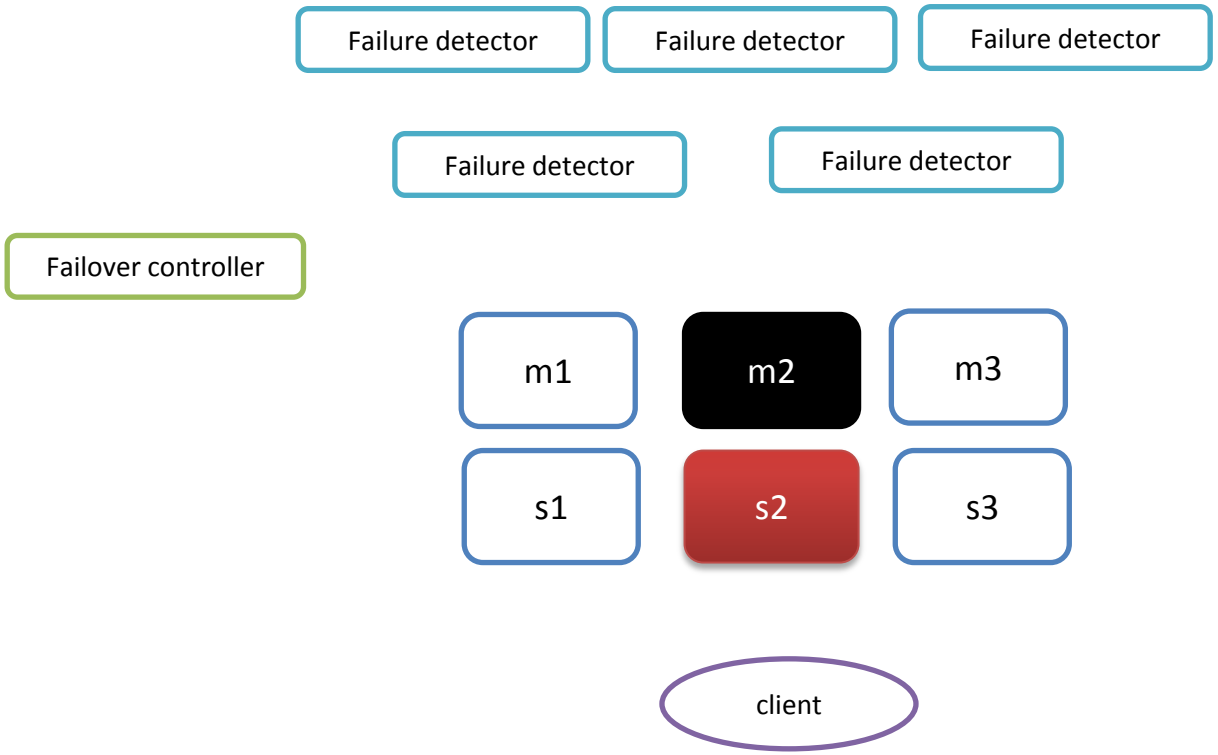
JD.COM 京东



Accurate failure detector & auto failover

JD.COM 京东

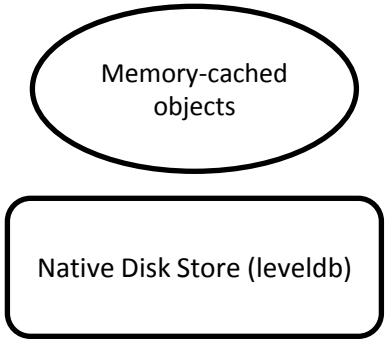
- 分布式投票检测故障
- 自动提升从实例为新主



Jimdb server internals

JD.COM 京东

- Redis协议兼容
- 两级存储结构
  - RAM + SSD/HDD
- 多线程读

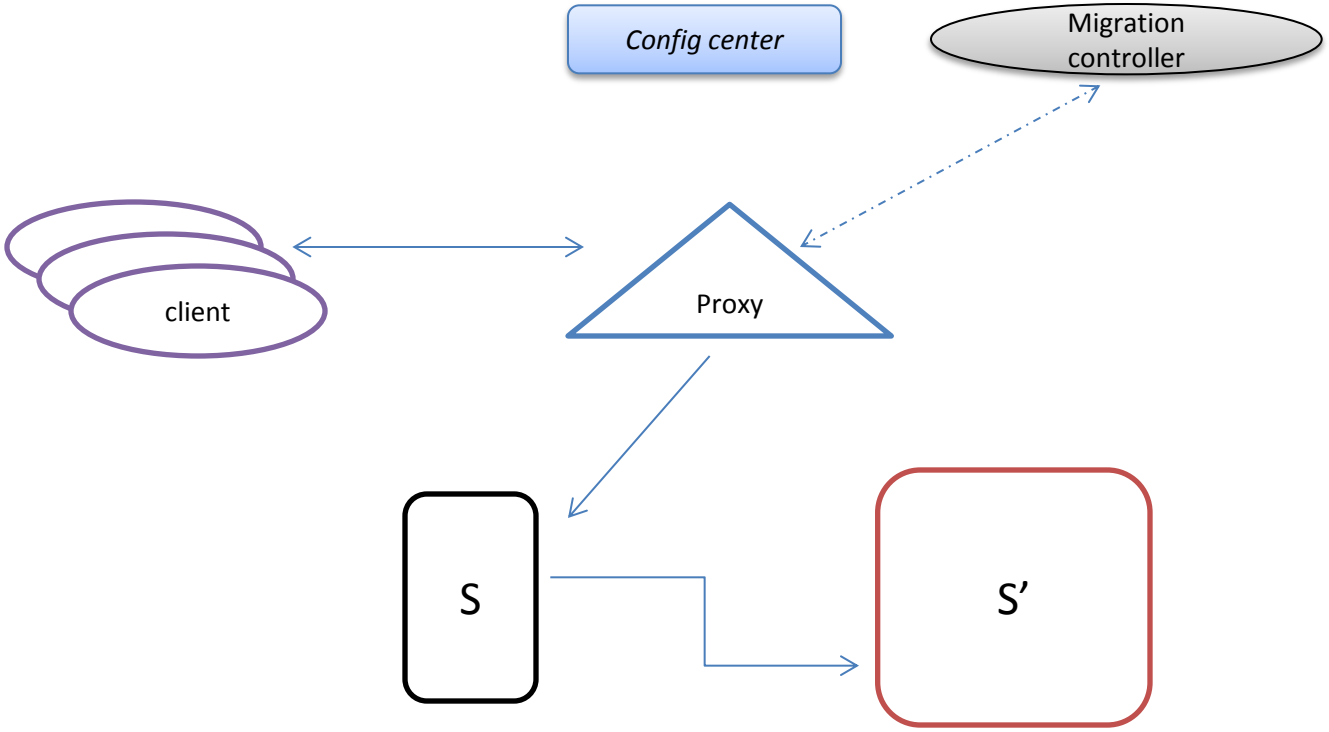


- 先后开发两个存储引擎
  - Leveldb vs. B+Tree
- 性能
  - 单实例，读写TPS近2万

Jimdb Scaling Up

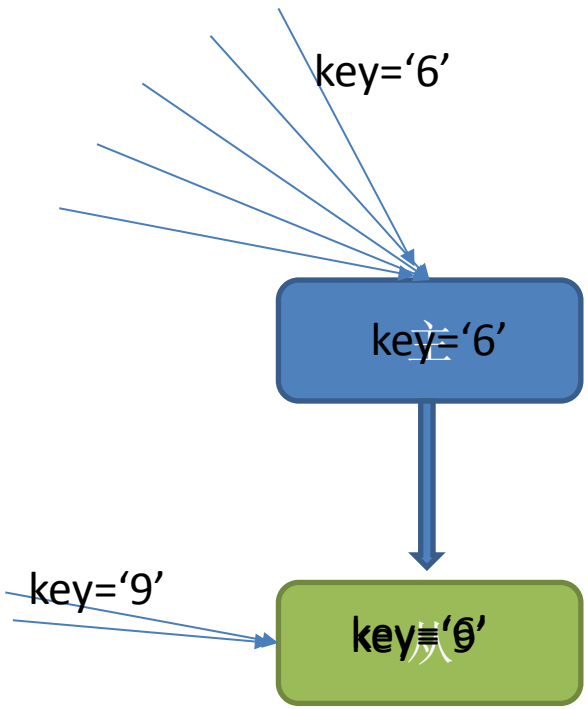
JD.COM 京东

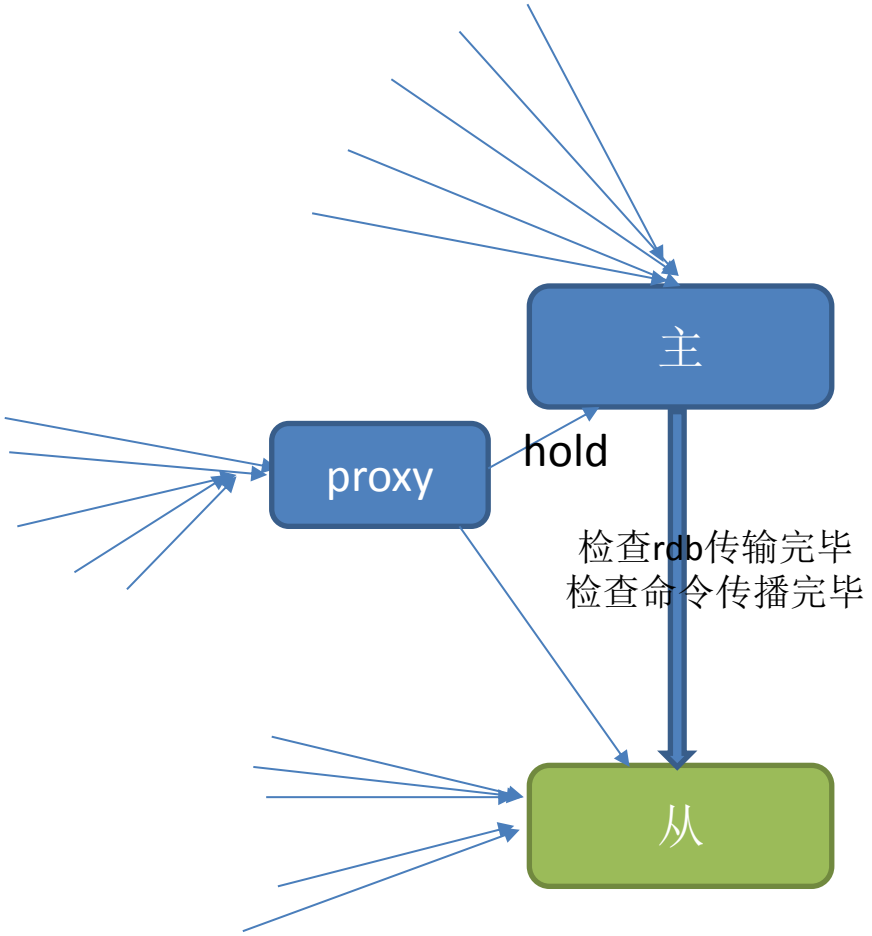
- Safe Online Switchover



Safe online switchover

JD.COM 京东

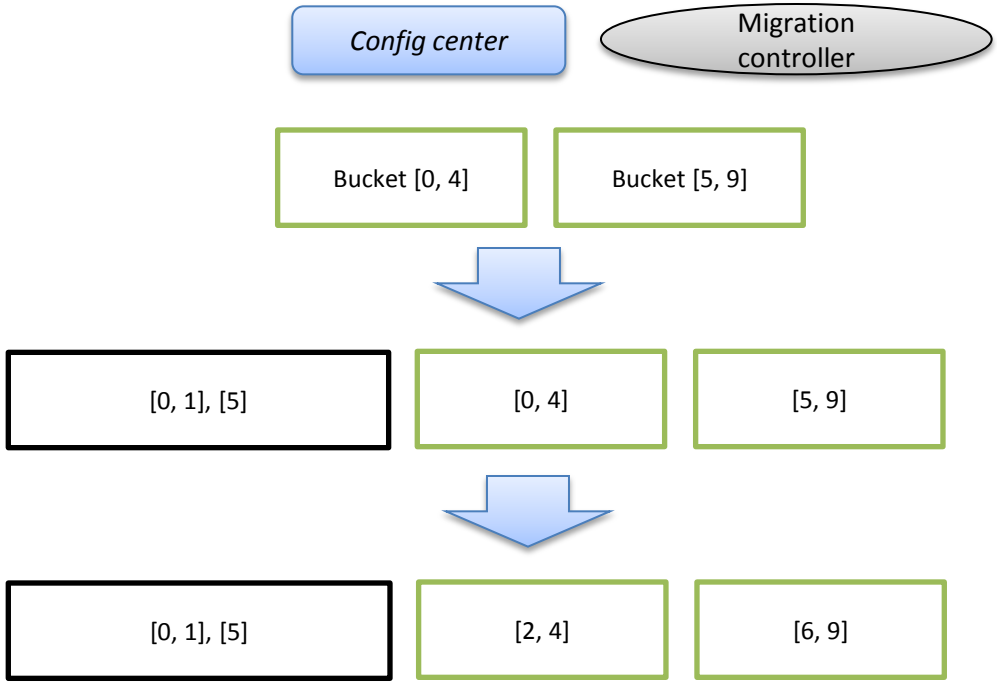




Jimdb Scaling Out

JD.COM 京东

- Safe Online Re-Sharding
  - Based on ***filtered replication***



未来工作

JD.COM 京东

* JFS

unified storage infrastructure for files/objects

*Jimdb

distributed cache & fast key-value store

FDB: structured data storage

总结

JD.COM 京东

- 规模驱动自主研发、持续研发
- 如果不能做得更好，自研其实没有意义
  - 好 != 铺摊子，好 = 适合公司需求

- 专注投入，用技术革新创造业务价值
- 不断改进甚至重构，保持系统的质量与活力

招聘广告

JD.COM 京东

岗位名称：系统技术部，分布式系统“攻城狮”
岗位要求：热爱技术，扎实积累，基情四射

工作职责：

- 运用扎实的系统、算法与编程技术，负责大规模存储、核心中间件、图片与视频平台的自主研发与工程实施；
- 与研发部各团队紧密合作，将自研系统广泛应用于业务，并通过技术手段持续完善。

- 支撑海量数据（图片、订单、物流、对象存储）的京东文件系统（JFS）
- 基于内存与SSD的分布式缓存与高速KV服务（Jimdb）
- 日均百亿消息传递的新消息平台（New MQ）
- 强有力的下一代RPC服务框架（New SAF）
- 持续优化图片系统，降低带宽成本并提升购物体验
- 建设视频分发网络(VDN)

谢谢！

刘海锋
bjliuhaifeng@jd.com

JD.京东.COM