京东分布式存储的持续研发

刘海锋

JD.京东 COM

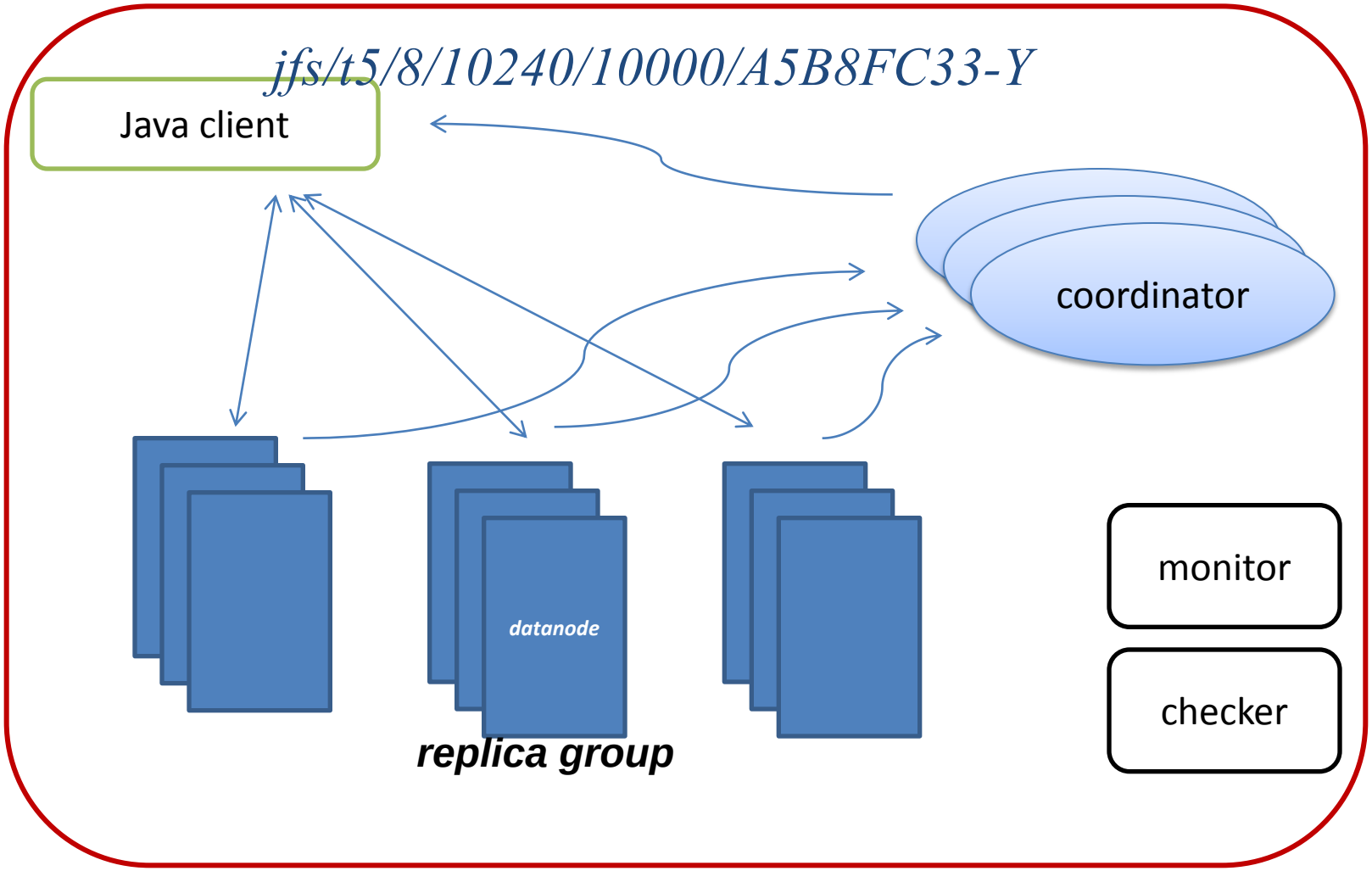
过去一年里所做的努力

- **自主研发**
  - 需求驱动，分期开展
  - 相互复用又高度定制的多个系统
- **具体产品**
  - 京东文件系统
  - 对象存储服务
  - 弹性块存储
  - 高速NoSQL平台

Jingdong File System

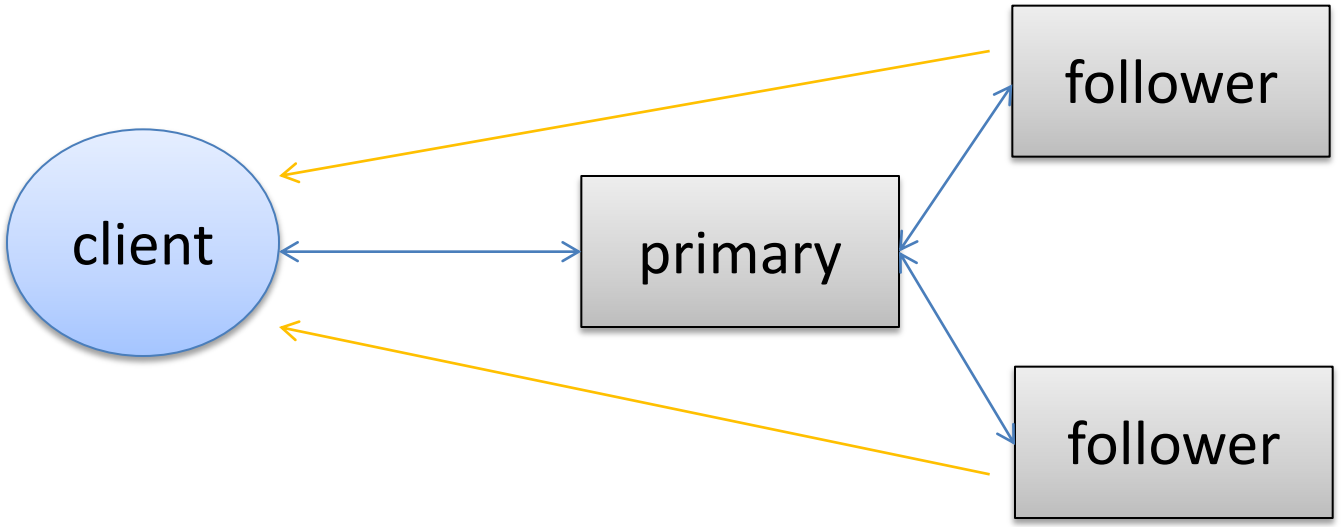
- 需求驱动
  - 特别针对海量小文件
  - 强可靠、强一致、高可用
  - 透明压缩以节省带宽与机器资源
- Scalable systemkey-file storage
  - 文件key由系统生成，应用自行保存

JFS-V1: immutable small files



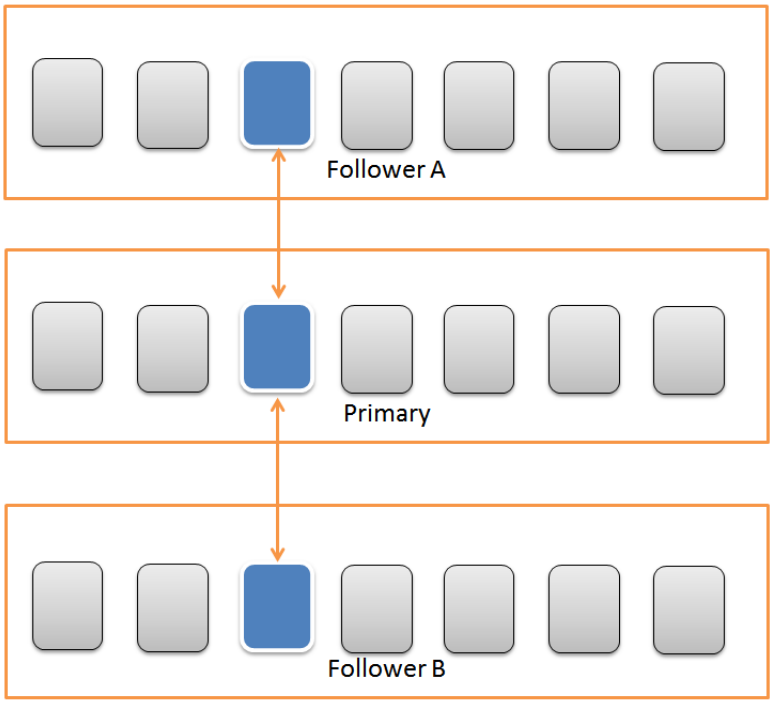
复制协议

- **Paxos变体**
  - 固定成员角色 – one primary + two followers
  - Full-quorum replication



单机存储引擎

- **一组append-only files, without in-RAM indexes**
  - chunkId/offset/size as the internal key



应用

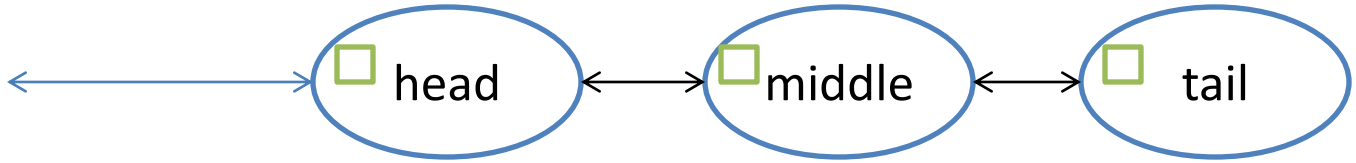
- **图片**
- **订单**
- **仓库流水记录**
- **数据库大字段**

...

- **极强的容错能力**
  - 机器或磁盘故障
  - 文件误删除或截断
- **优异的性能**
  - 追加写入
  - 单次seek读取

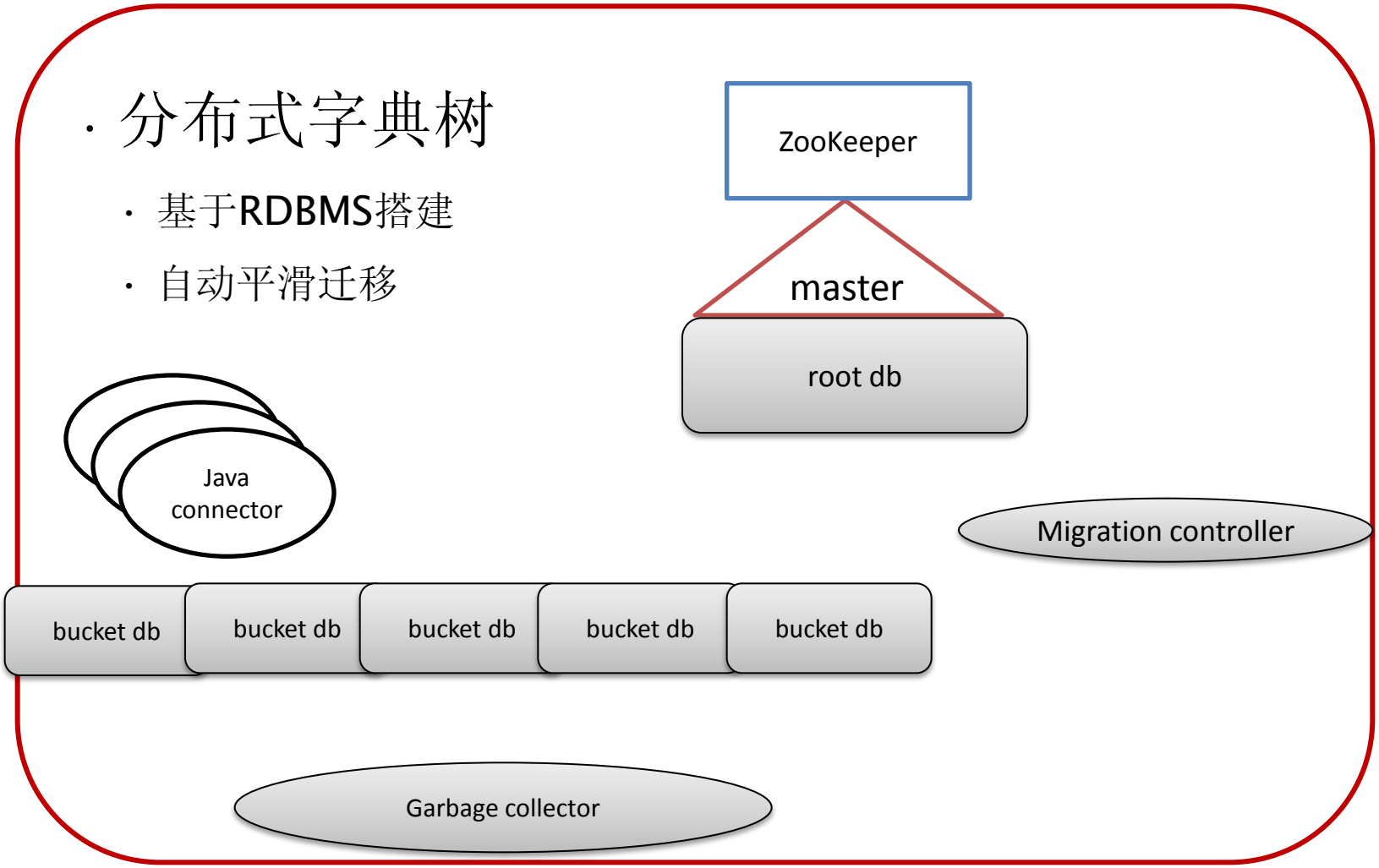
JFS-V2: 大文件支持

- 抽象
  - Append-Only File Segments
- 复制技术
  - Chain Replication



JFS-V3: scalable namespace management

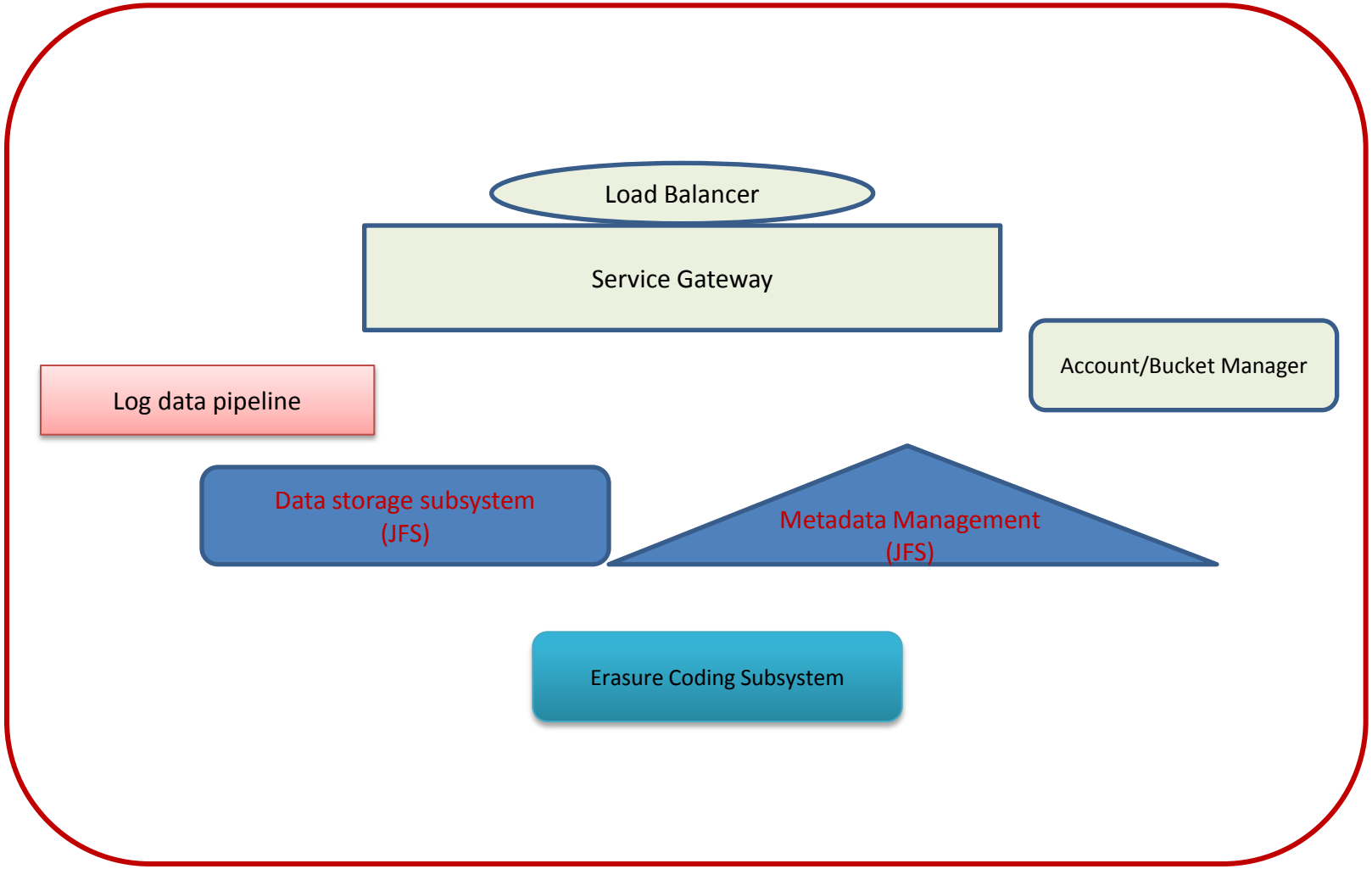
- 分布式字典树
  - 基于RDBMS搭建
  - 自动平滑迁移



Jingdong Storage Service

- 对象存储服务
  - 兼容S3 API
  - Accounts/Buckets/Objects
    - *name → metadata + data*

JSS内部架构

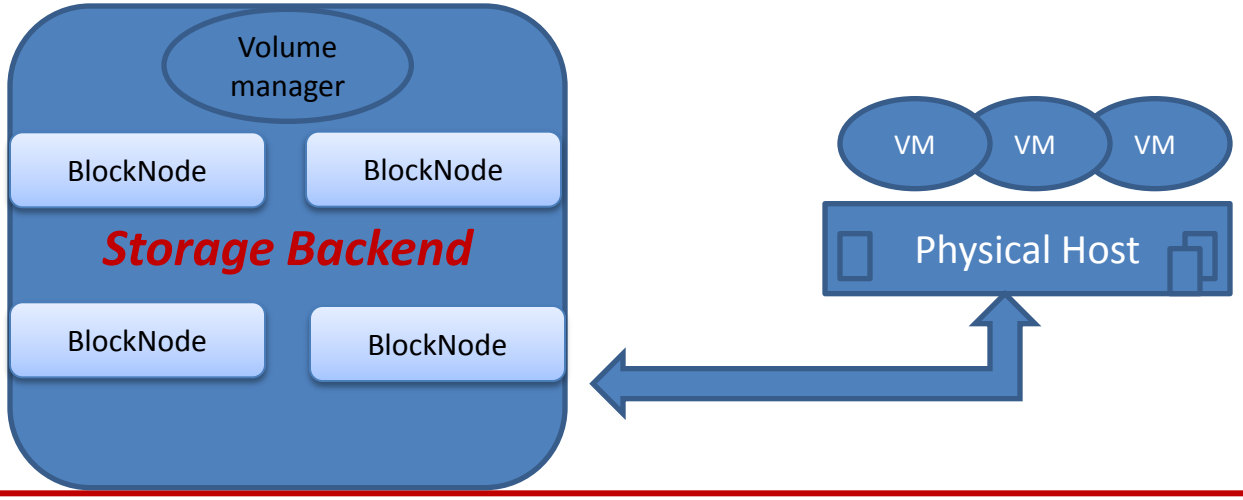

Load Balancer
Service Gateway
Account/Bucket Manager
Log data pipeline
Data storage subsystem
(JFS)
Metadata Management
(JFS)
Erasure Coding Subsystem

JSS应用

- **网盘、音乐、电子书、网站静态内容**
  - 极强的可靠性与可用性
- **数据共享、授权认证**
  - 内外网均可访问
- **冷数据永久归档**
  - 低成本保证 – 擦除编码

Jingdong Block Storage

- 持久、无故障的“云硬盘”抽象
  - 后端实现复制和高可用
  - 主要应用于虚拟化平台
- 技术挑战
  - 故障切换与恢复、性能调优



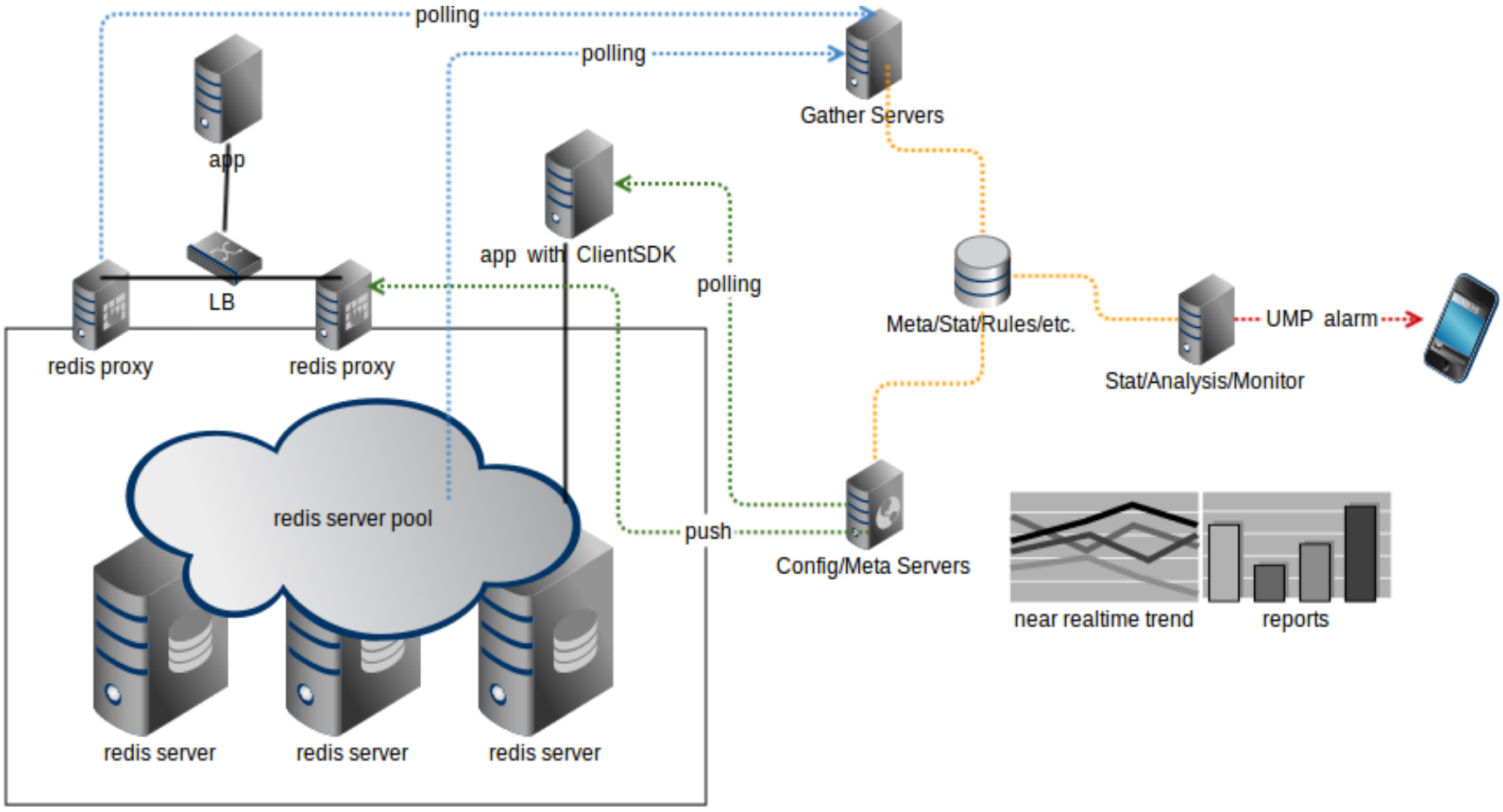
Jingdong In-Memory Store

“Memory is the new disk.”

演进历史

- Redis单实例
- Redis主从复制
- Pre-sharding
- 监控、报警
- 统一缓存平台
- JIMStore

统一缓存平台

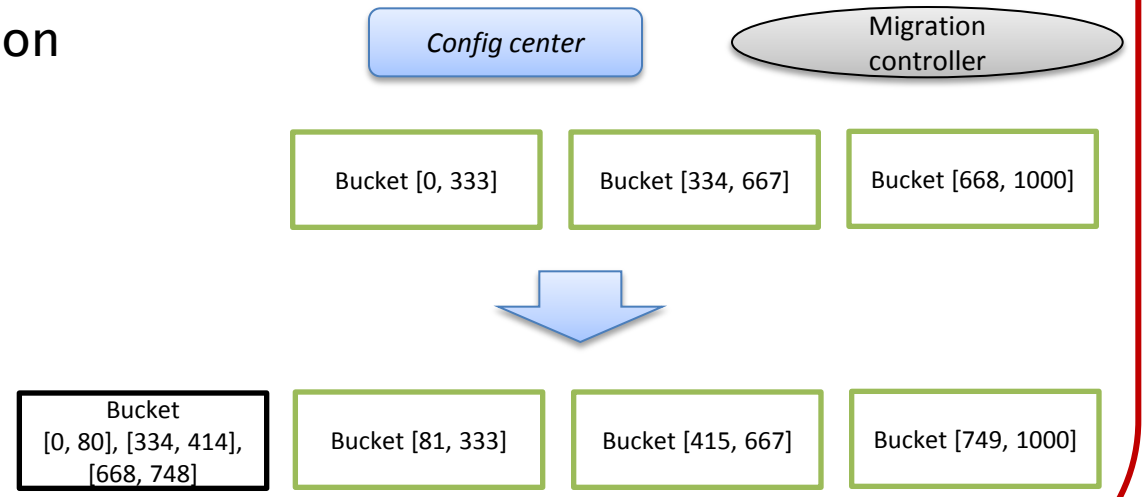

polling
polling
app
Gather Servers
app with ClientSDK
LB
polling
Meta/Stat/Rules/etc.
UMP alarm
redis proxy
redis proxy
Stat/Analysis/Monitor
redis server pool
push
Config/Meta Servers
near realtime trend
reports
redis server
redis server
redis server

JIMStore Features

- 精确的故障检测
  - Failure detector via distributed election
- 自动故障切换
  - Failover controller
- 多租户资源公平性
  - Traffic monitoring & control
- RAM/Disk两级存储
  - Our own Redis branch
- 横向扩容
  - Our own cluster mode

水平扩展

- 新增抽象
  - db, *buckets*, keys
- 新增操作
  - buckt-level *selective replication*
  - bucket deletion



总结

- **研发策略**
  - 分期开展，逐步推进
  - 围绕核心业务需求
- **挑战**
  - 软件质量保证
  - 运维管理

Thank You.

And we are hiring☺

JD.京东 COM