京东文件系统：从2013到2015

京东商城高级架构师 桂创华

DTCC

2015中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2015

大数据技术探索和价值发现

IT168.com ChinaUnix ITPUB

Overview

- **JFS大、小文件系统**
- JFS在京东的应用
- JFS V2 & V3
- JFS展望

Jingdong File System

- 愿景
  - The unified datacenter storage infrastructure (2013/7 – now)
- 小步展开，分期快跑
  - 小文件存储
  - 大文件存储
  - 块存储
  - 对象存储
  - JFS V2
  - JFS V3

海量小文件-需求与挑战

- 交易订单：千万/天
- 商品图片：千万/天
- 库房记录：亿/天
- 电子签收：千万/天

各种方案

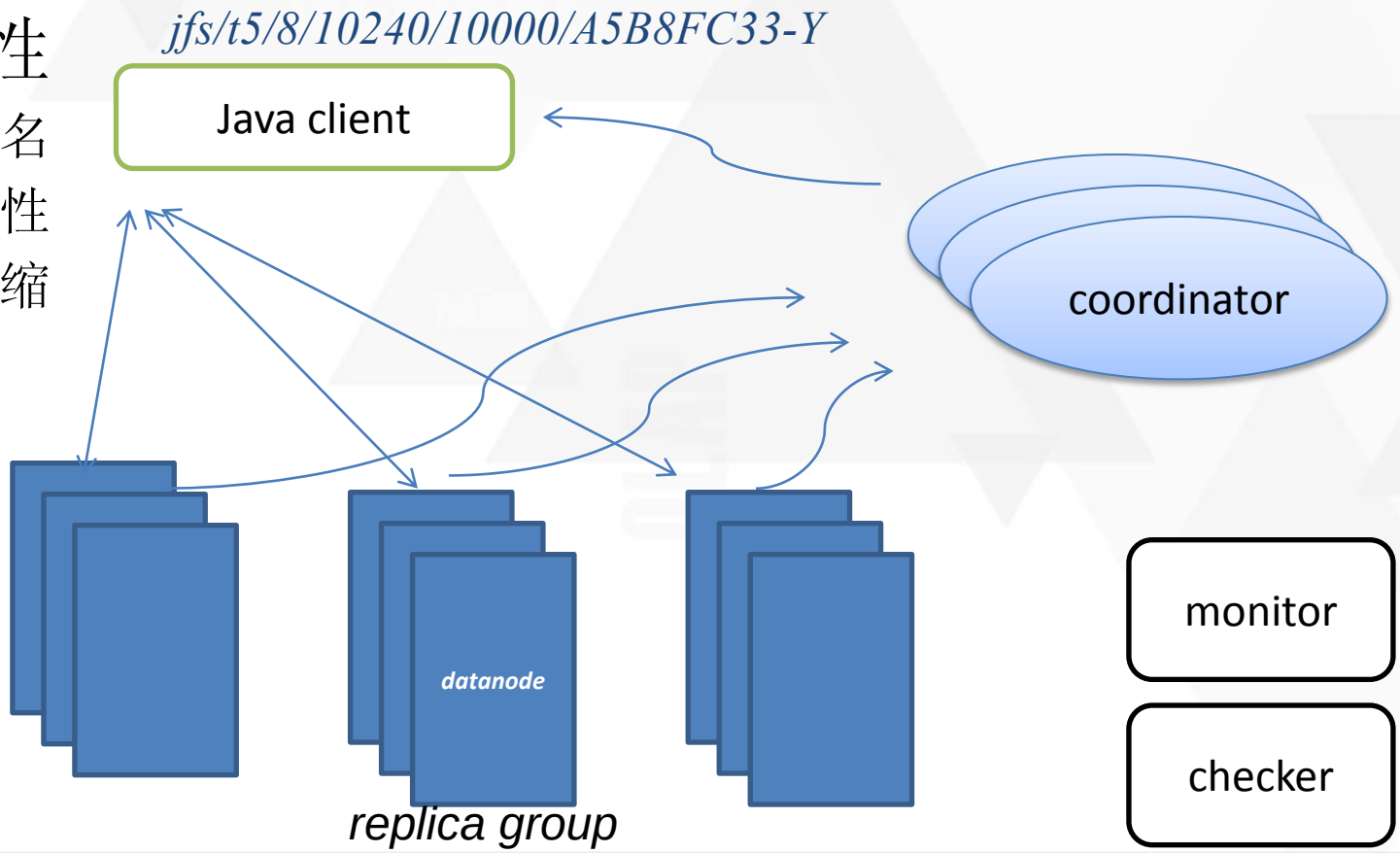
- 关系型数据库
  - Mysql
  - Pains：难以扩容，性能瓶颈
- 开源存储系统
  - HDFS，FastDFS
  - Pains：选型、维护、定制

自主研发

- 收益
  - 灵活可控、长期技术收益
  - 紧扣业务、高度定制
- 挑战
  - 开发周期
  - 稳定性
  - 长期性

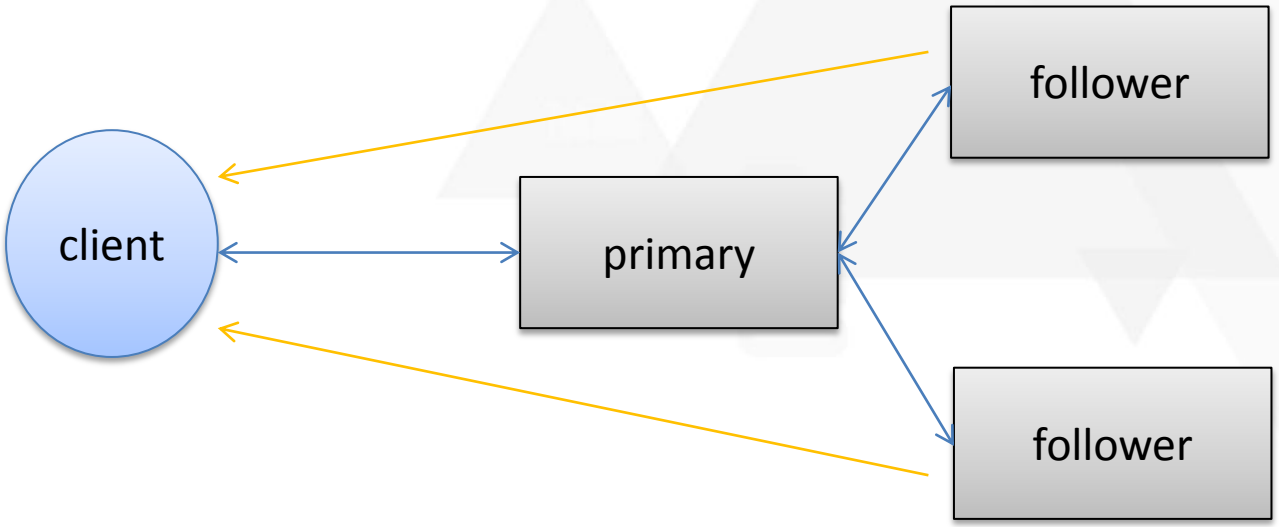
JFS小文件存储系统

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

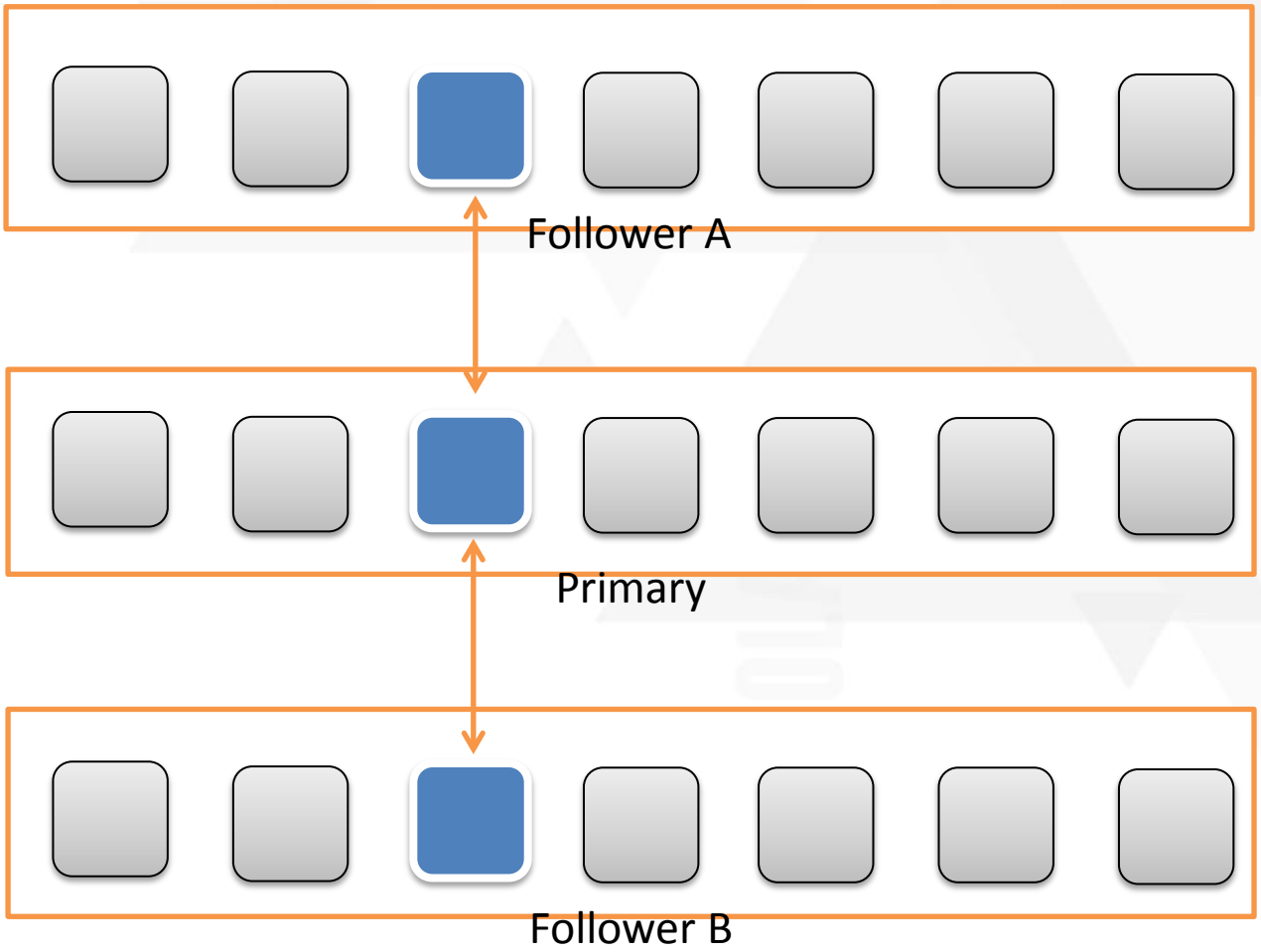


复制协议

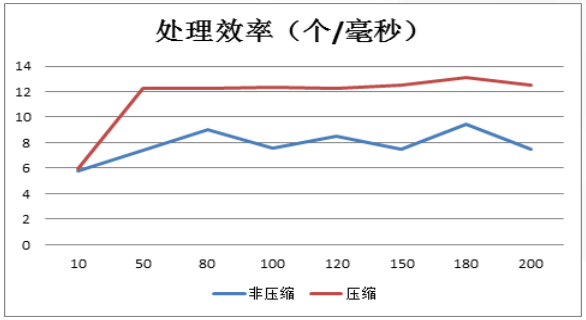
- 固定成员角色
  - One Primary + 2 follower
  - 强一致性复制
- 交叉混布

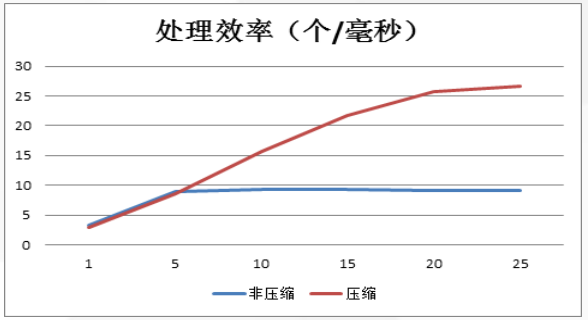


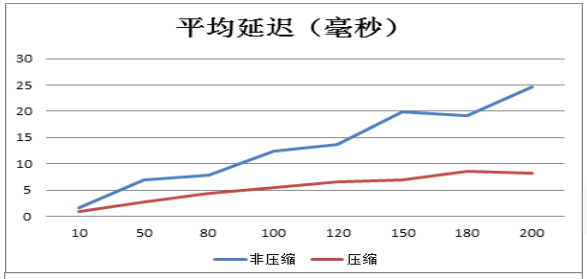
存储引擎

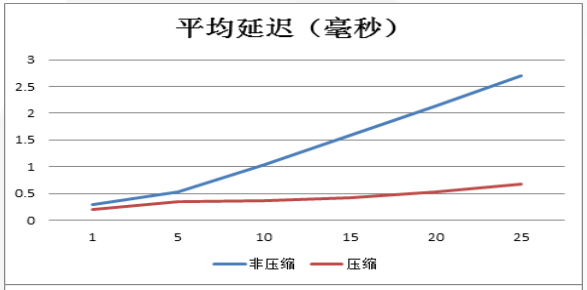


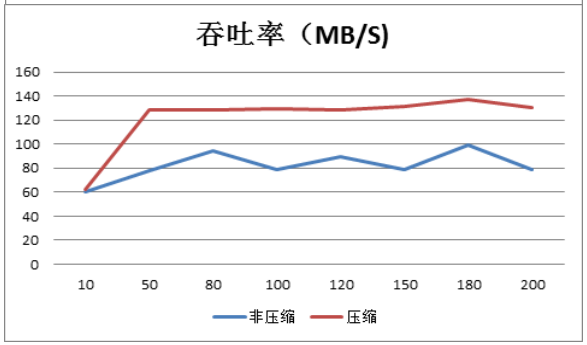
性能

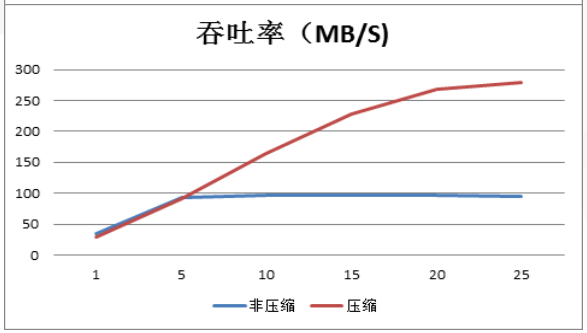

处理效率（个/毫秒）
14
12
10
8
6
4
2
0
10 50 80 100 120 150 180 200
非压缩
压缩


处理效率（个/毫秒）
30
25
20
15
10
5
0
1 5 10 15 20 25
非压缩
压缩

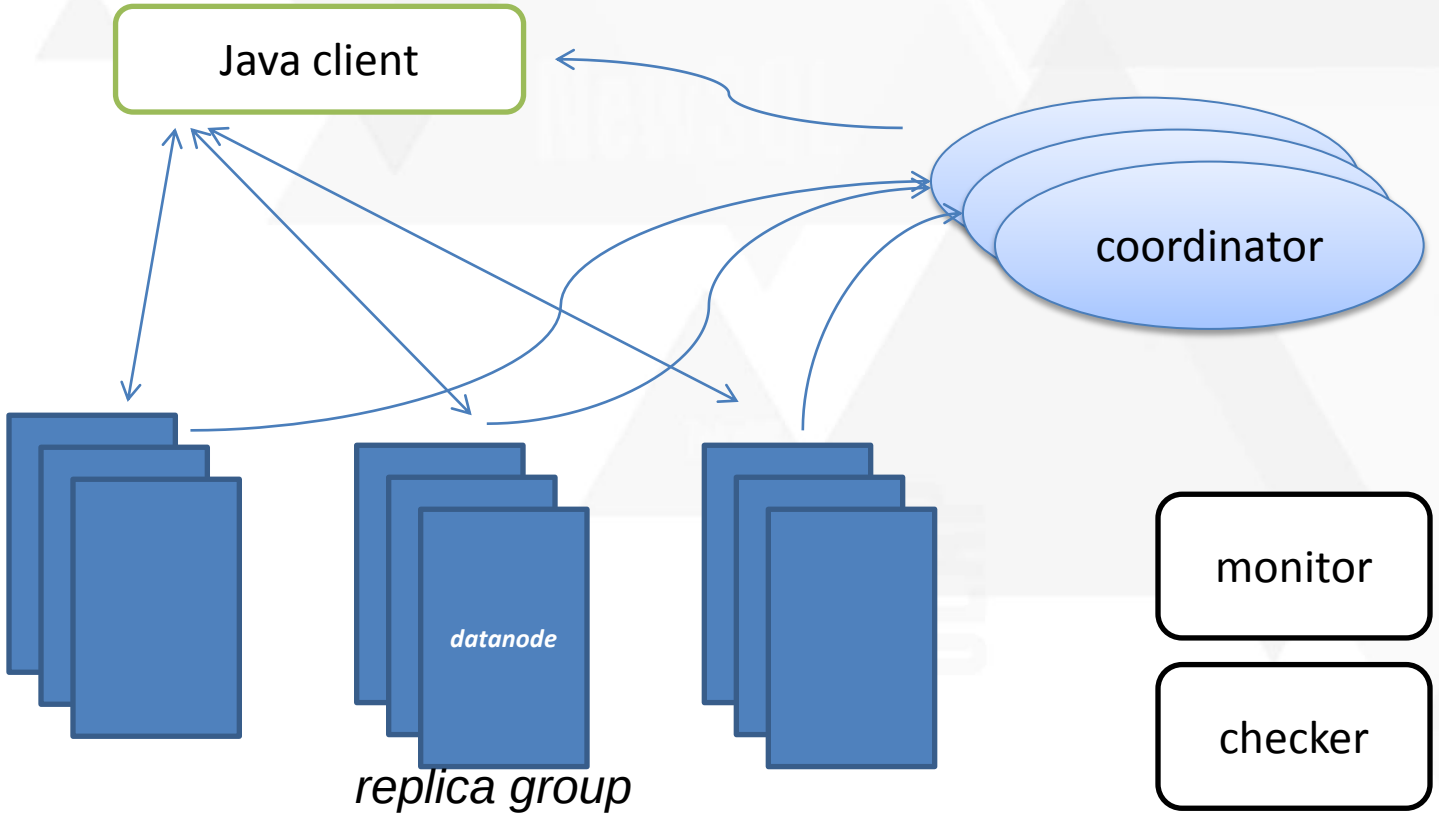

平均延迟（毫秒）
30
25
20
15
10
5
0
10 50 80 100 120 150 180 200
非压缩
压缩


平均延迟（毫秒）
3
2.5
2
1.5
1
0.5
0
1 5 10 15 20 25
非压缩
压缩


吞吐率（MB/S)
160
140
120
100
80
60
40
20
0
10 50 80 100 120 150 180 200
非压缩
压缩

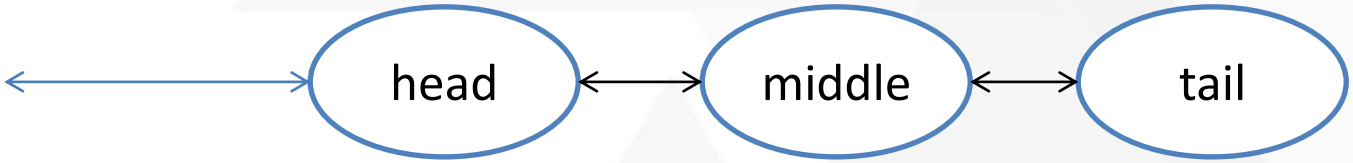

吞吐率（MB/S)
300
250
200
150
100
50
0
1 5 10 15 20 25
非压缩
压缩

JFS大文件存储系统


Java client
coordinator
datanode
replica group
monitor
checker

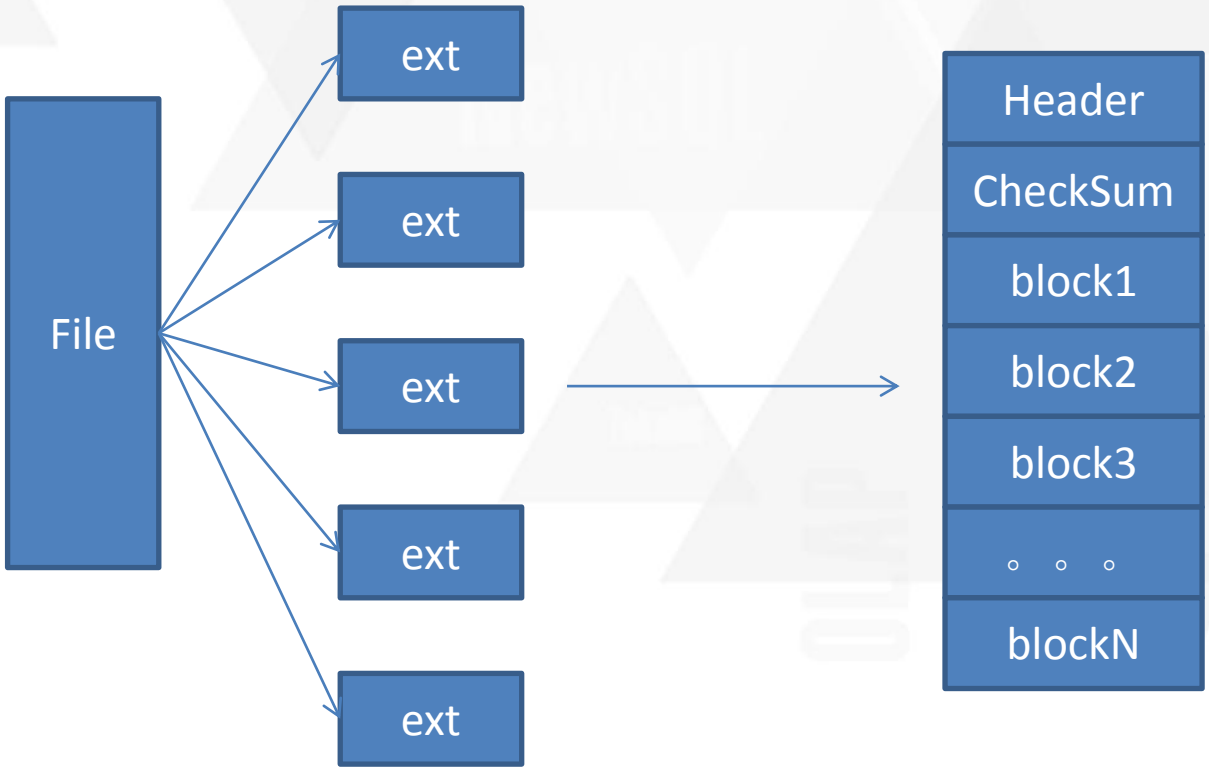
复制协议

- 链式复制



- 流式读写

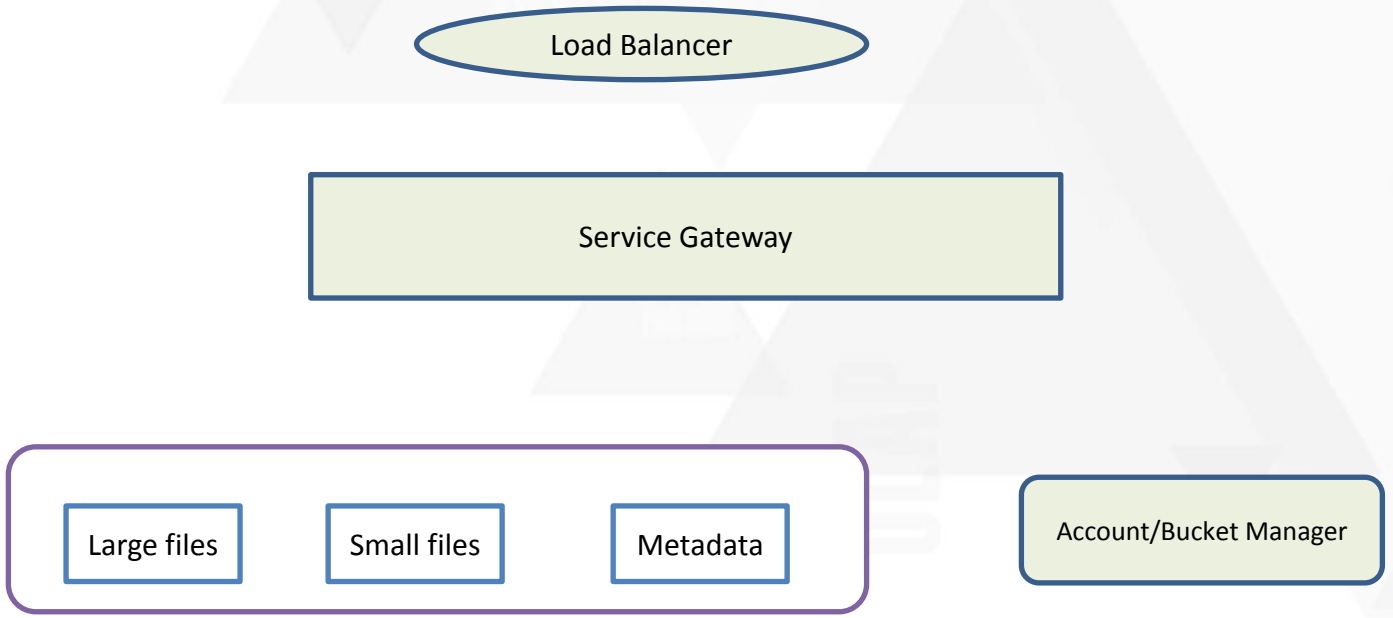
存储引擎


File
ext
ext
ext
ext
ext
Header
CheckSum
block1
block2
block3
。 。 。
blockN

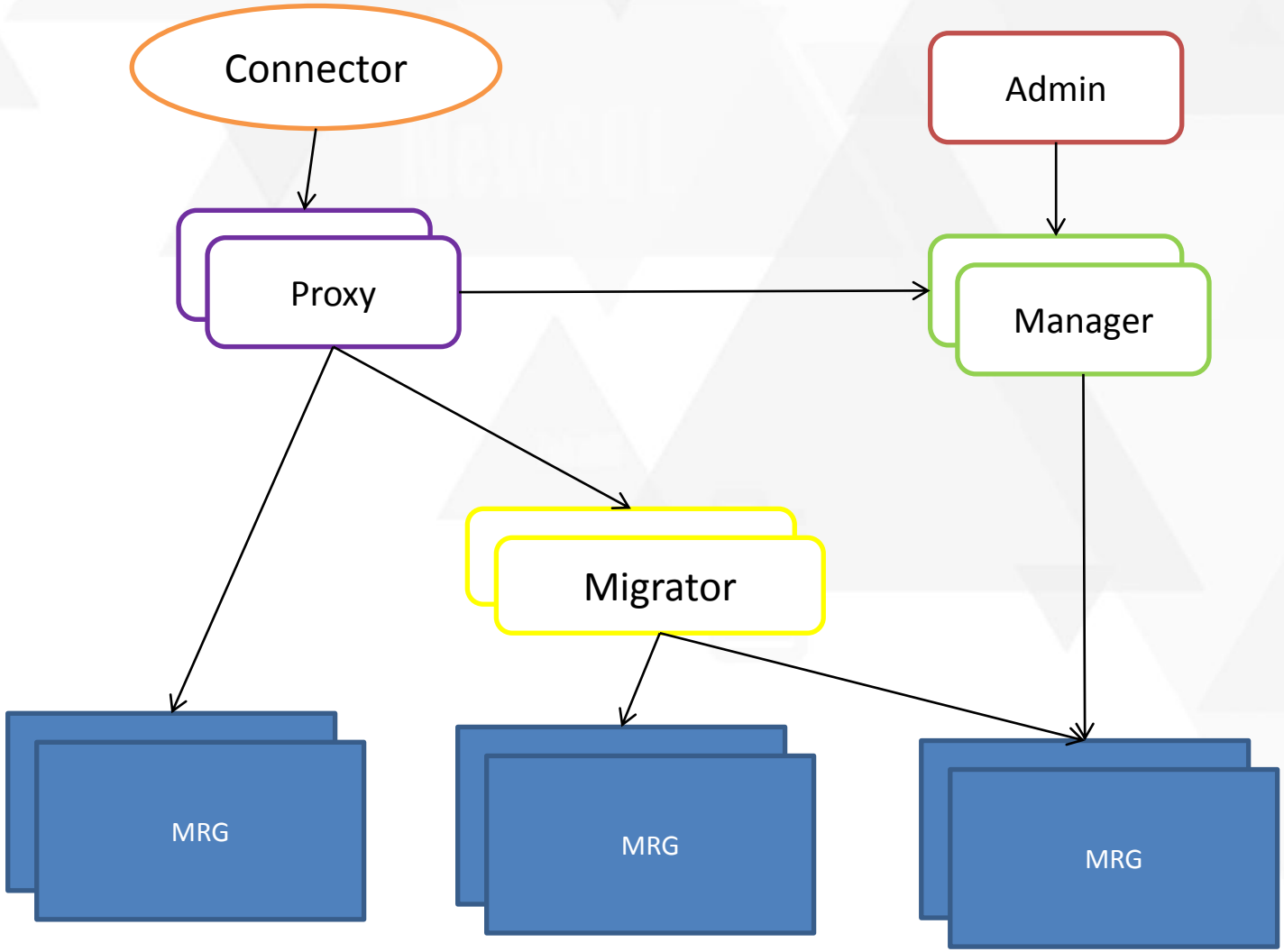
Overview

- JFS大、小文件系统
- **JFS在京东的应用**
- JFS V2 & V3
- JFS展望

云存储服务JSS

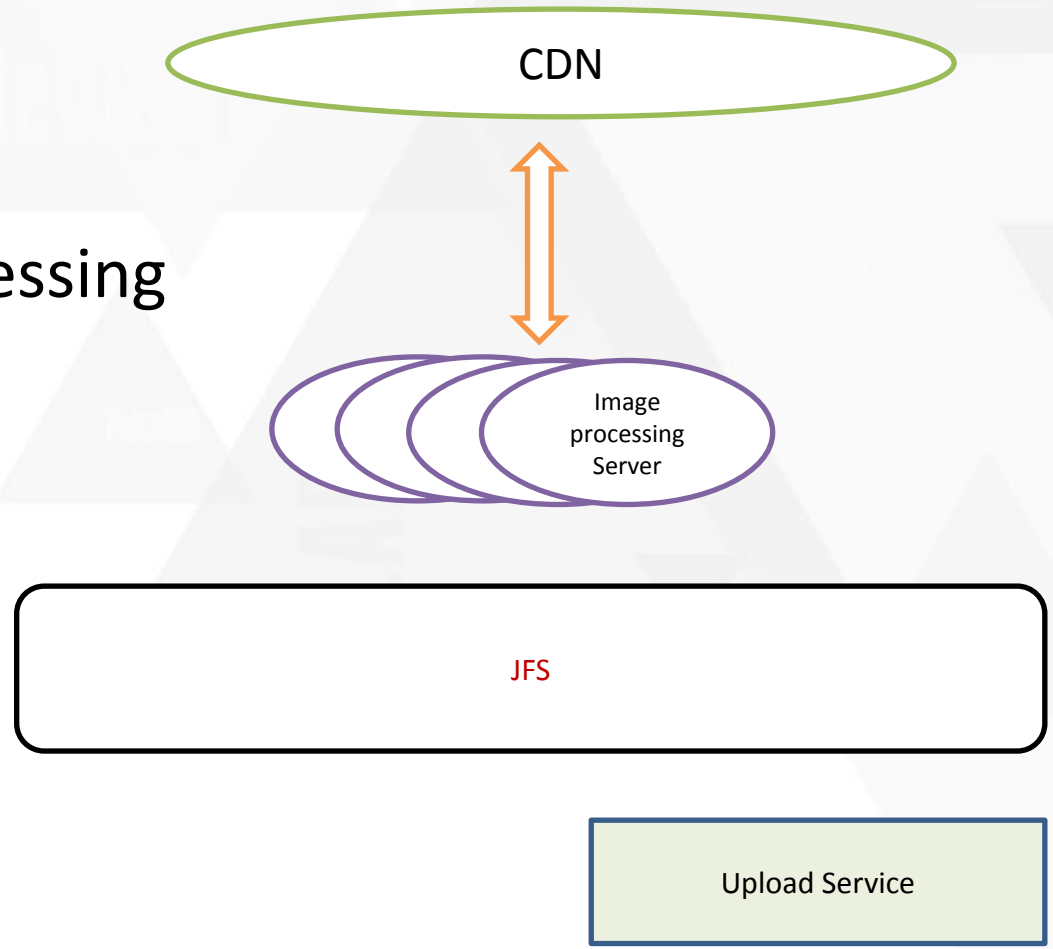

Load Balancer
Service Gateway
Large files
Small files
Metadata
Account/Bucket Manager

JFS元数据表格系统



JFS新图片系统

- 重构图片服务
  - JFS做底层存储
  - 重写Image Processing
  - 图片性能优化
  - WebP



JFS业务影响

- 多个集群
  - 京东图片、OFC、内部云存储、公有云存储、拍拍主图和商描、电子签收
- 300多个业务应用，PB规模

Overview

- JFS大、小文件系统
- JFS在京东的应用
- **JFS V2 & V3**
- JFS展望

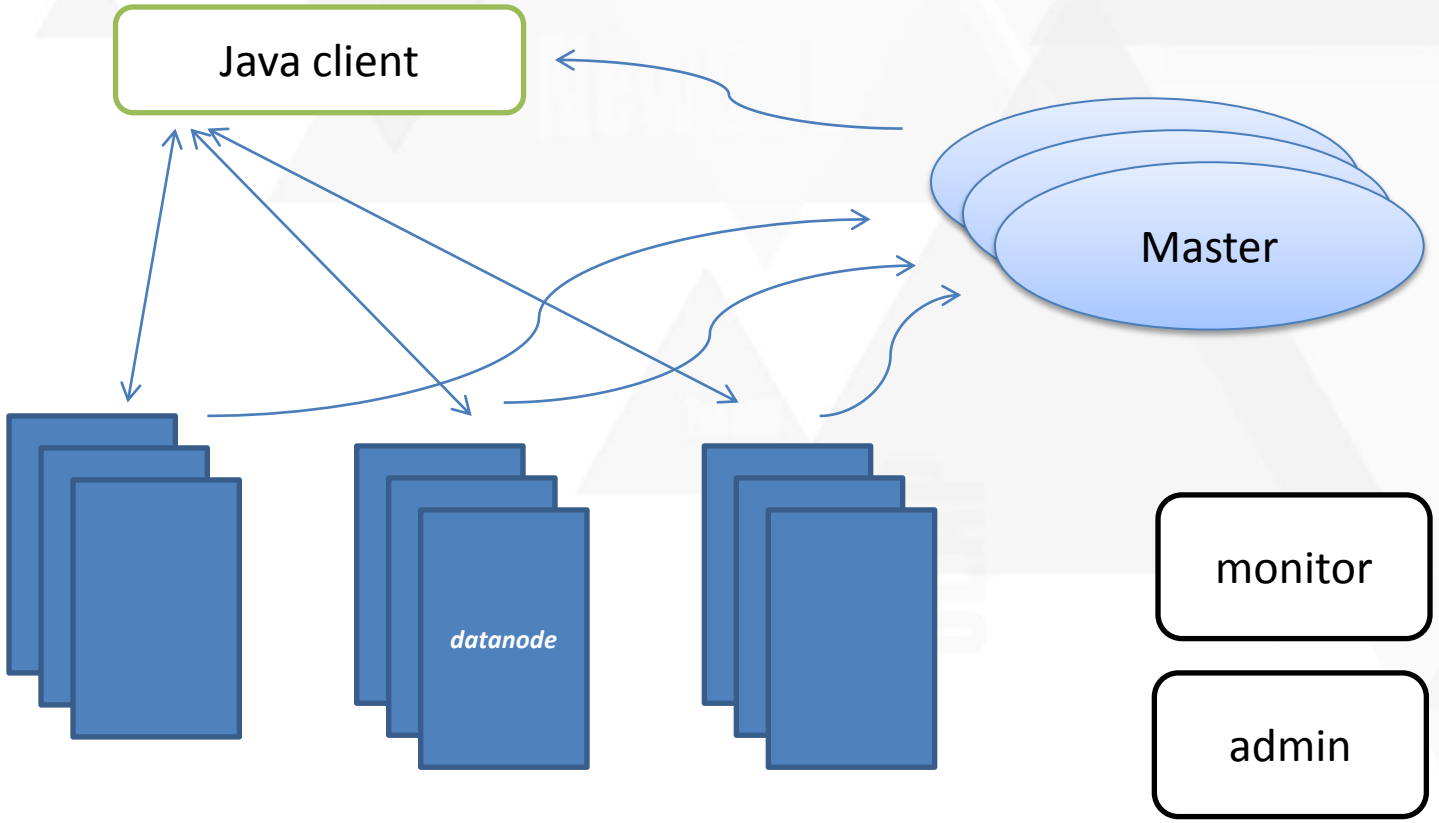
JFS V1痛点

- 小文件和大文件存储系统能否统一
- 磁盘故障修复太慢
- 副本增、删不自动
- Client太重
- 无法自动升级
- 监控系统弱

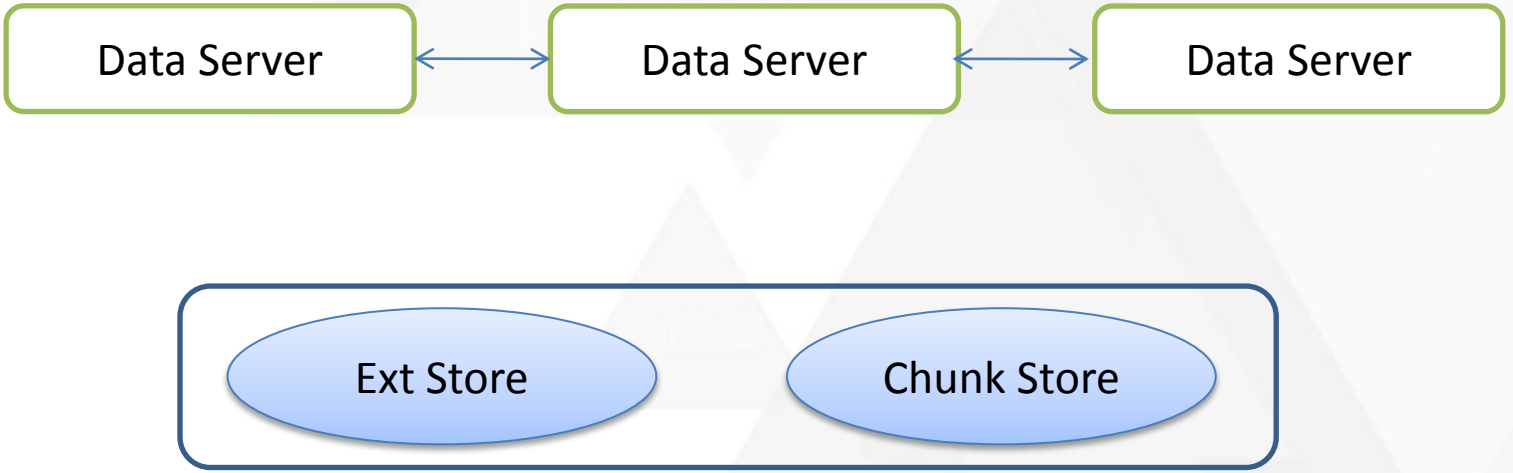
JFS V2

- 支持小文件的整读整写
- 支持大文件的整读整写、随机读、追加写
- 支持副本的动态增、减
- 支持系统的自动升级
- 支持磁盘故障的分钟级修复
- 支持机器级别、volume级别关键指标的实时监控
- 强大的管理端

JFS V2



统一的Server架构



Master

- 资源调度和均衡
- 自动横向扩容
- 弹性增加、减少副本
- 副本对比和修复

监控管理系统

- 监控和统计
  - 机器维度：CPU、磁盘、网络、TCP连接数等等
  - Volume维度：读TPS、写TPS、读平均延时、写平均延时、读超时请求比例、写超时请求比例、error等等
  - 业务维度：Volume总数、入请求数、出请求数等等
- 统计
  - 业务Volume视图
  - 机器volume视图
- 告警
  - 告警分级
  - 自定义告警

故障快速修复

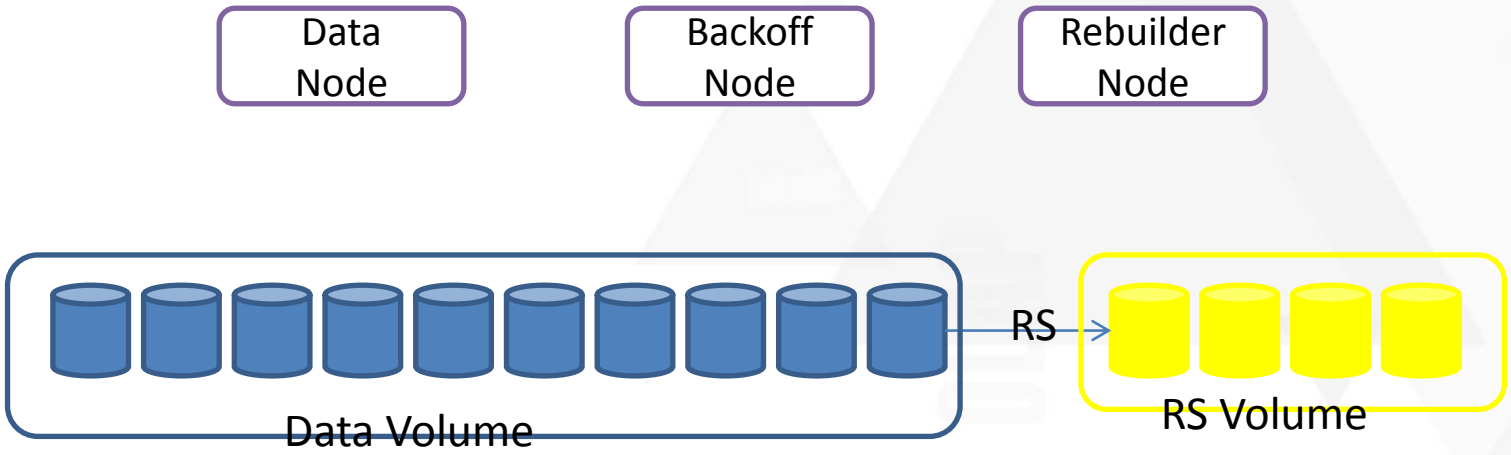
- 磁盘粒度-----JFS V1
  - 优点：Master元数据小
  - 缺点：磁盘故障修复速度慢
- 文件粒度-----HDFS
  - 优点：故障修复速度快
  - 缺点：Master元数据大
- Volume粒度-----JFS V2

JFS V3

- Hadoop集成
- 可擦除编码
- 跨机房

可擦除编码

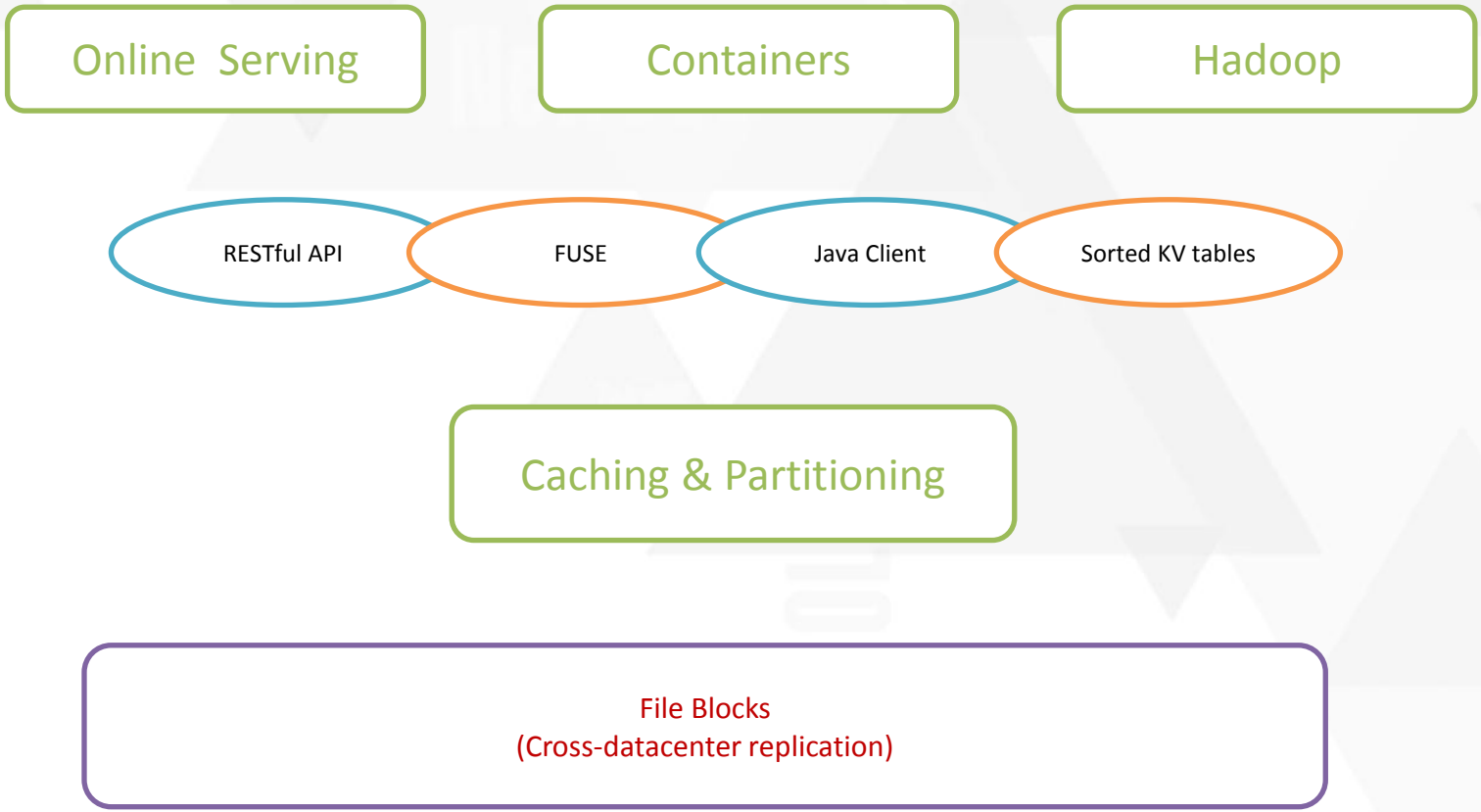
- Warm & Cold Data



Overview

- JFS大、小文件系统
- JFS在京东的应用
- JFS V2 & V3
- **JFS展望**

JFS展望


Online Serving
Containers
Hadoop
RESTful API
FUSE
Java Client
Sorted KV tables
Caching & Partitioning
File Blocks
(Cross-datacenter replication)


ChinaUnix
THANKS