DTCC

数，造，未，来

第十二届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2021

2021 年 10 月 18 日 - 20 日 | 北京国际会议中心

Spark
Hadoop
OLAP
HANA
MPP
Oracle
SQL

IT168.com ChinaUnix.net ITPUB

eBay HDFS架构的演进优化实践

林意群 eBay大数据平台工程师

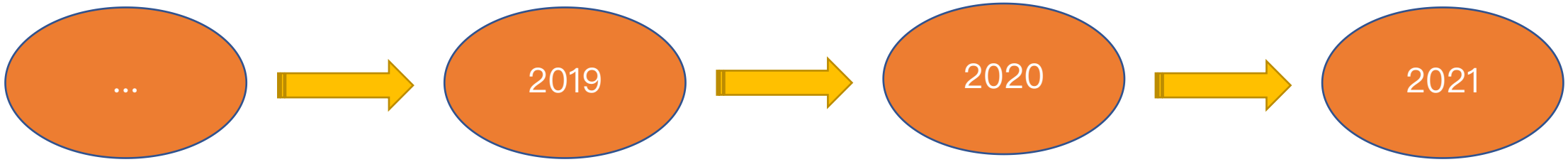
个人介绍-林意群

- Apache Hadoop PMC member
- Apache Ozone PMC member
- 多年大数据从业经验，19年加入eBay，主要负责eBay HDFS集群性能优化方面的工作。
- 参与开源社区多年，爱好技术分享。
- 《深度剖析Hadoop HDFS》作者

eBay Hadoop集群现状

10+	2W+	800PB +	100K+
集群	节点	存储	Job数

eBay HDFS集群的演进



独立集群模式

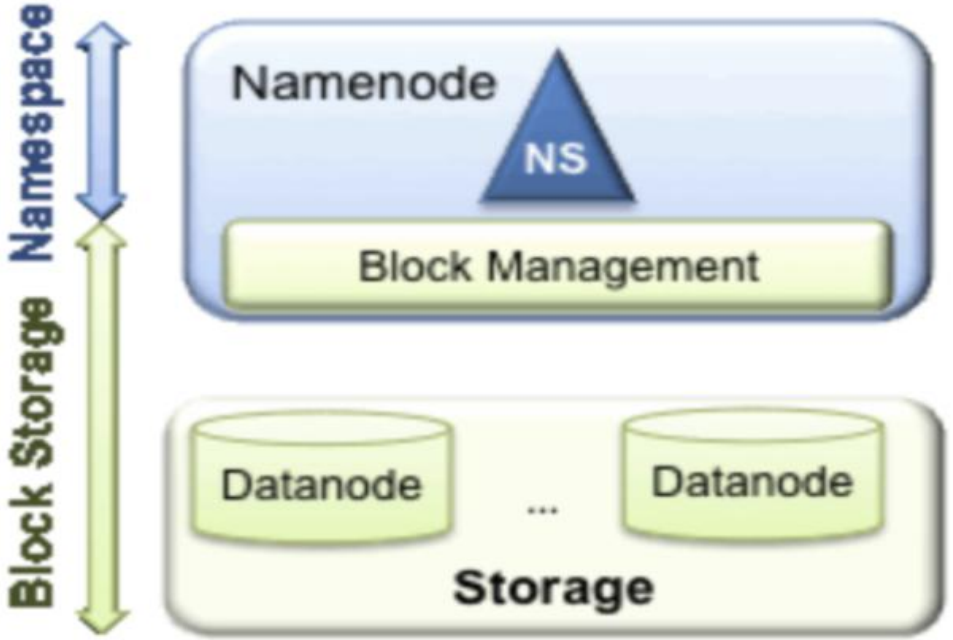
引入Viewfs Federation方案

NameNode单点瓶颈问题，
集群性能优化，拆分更多Federation集群。

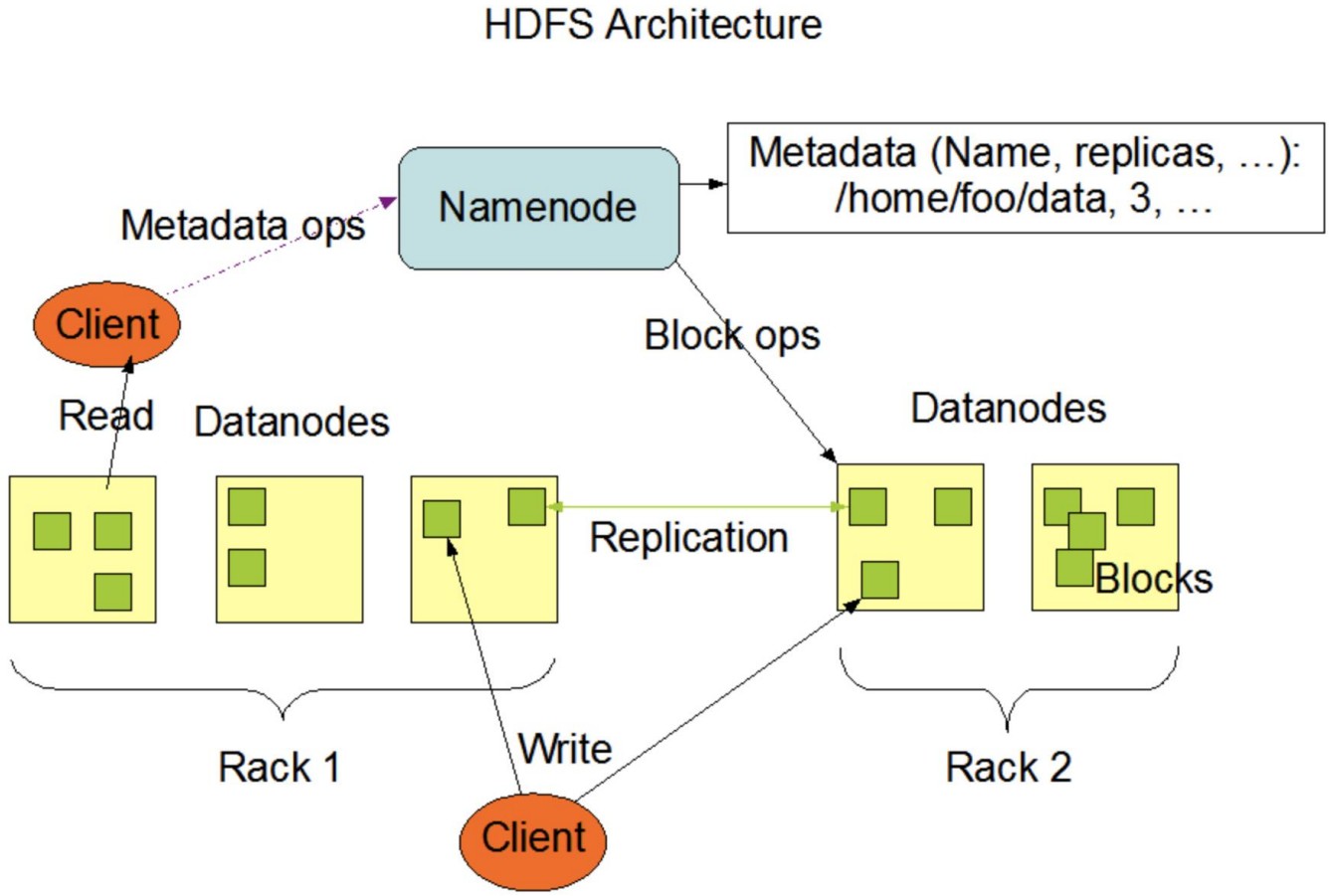
RBF方案逐步取代Viewfs Federation方式

初始HDFS架构模式

单集群模式



HDFS内部结构设计


HDFS Architecture
Namenode
Metadata (Name, replicas, …):
/home/foo/data, 3, …
Metadata ops
Client
Block ops
Read
Datanodes
Datanodes
Replication
Blocks
Rack 1
Write
Client
Rack 2

HDFS集群面临的挑战

1. 持续增长的数据存储压力，包括文件数据和元数据
2. NameNode服务的单点性能瓶颈
3. 多集群的运维管理，数据管理

HDFS性能调优

减少HDFS繁重API操作影响

- Balancer从Standby NameNode获取blocks操作
- Delete操作按照batch size执行的限制
- ListStatus操作忽略block location的获取
- Snapshot操作拆分为多子目录的管理

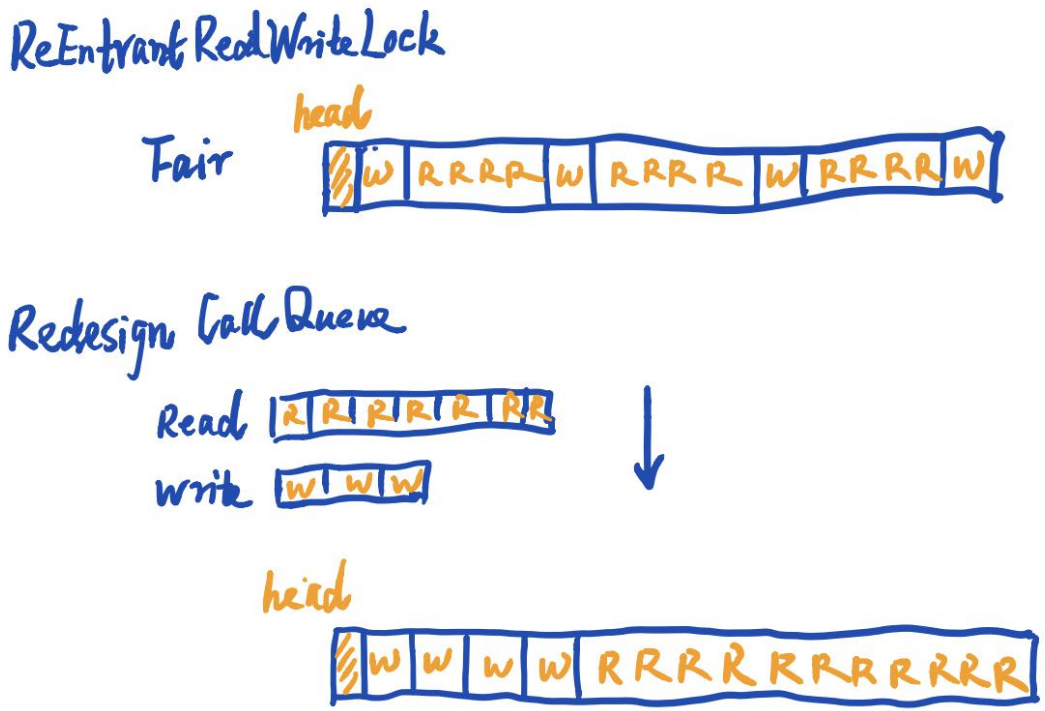
异步化RPC response

RPC的response阶段需要做加密操作，会造成一定的性能损耗，将此过程进行异步化地处理来提前释放NameNode的Handler资源(相关JIRA: HDFS-15486)。

NN锁优化处理

- 冗余目录锁的去除
- SetTimes操作写锁转读锁
- ReadWrite callqueue实现(相关JIRA: HDFS-15553)

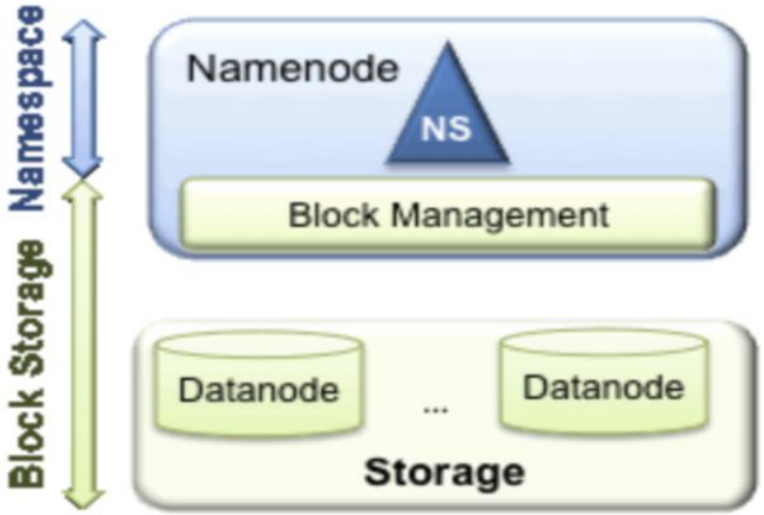
锁优化处理：ReadWrite callqueue



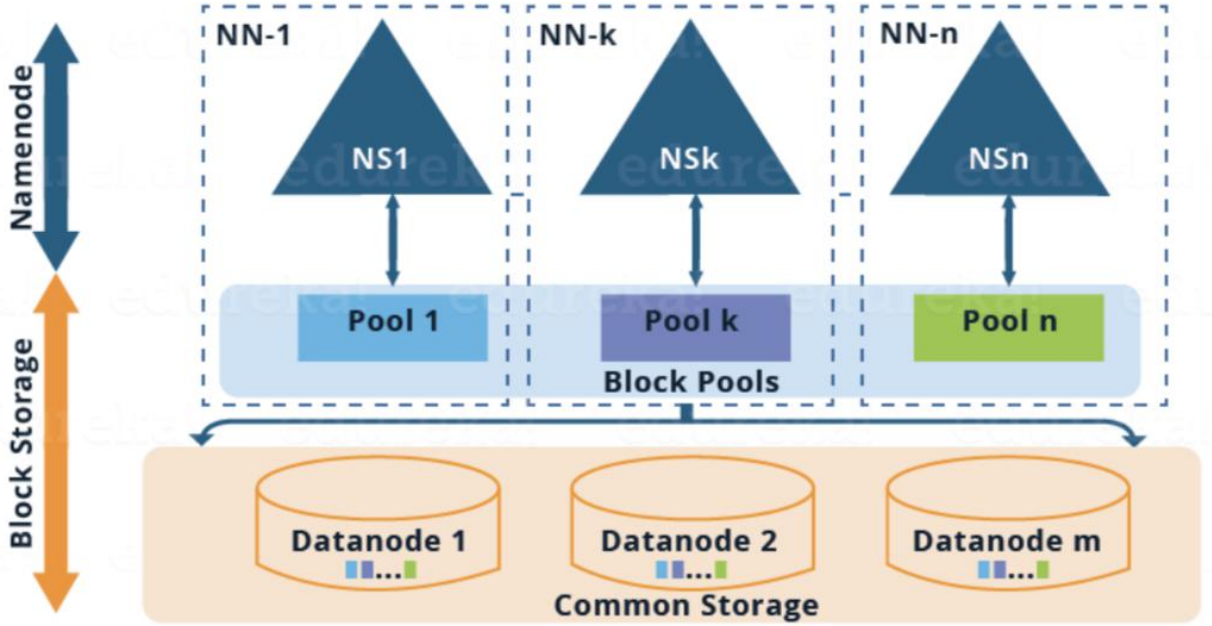
Ops数对比（读:写=30:1）

HDFS架构的演变

单集群模式

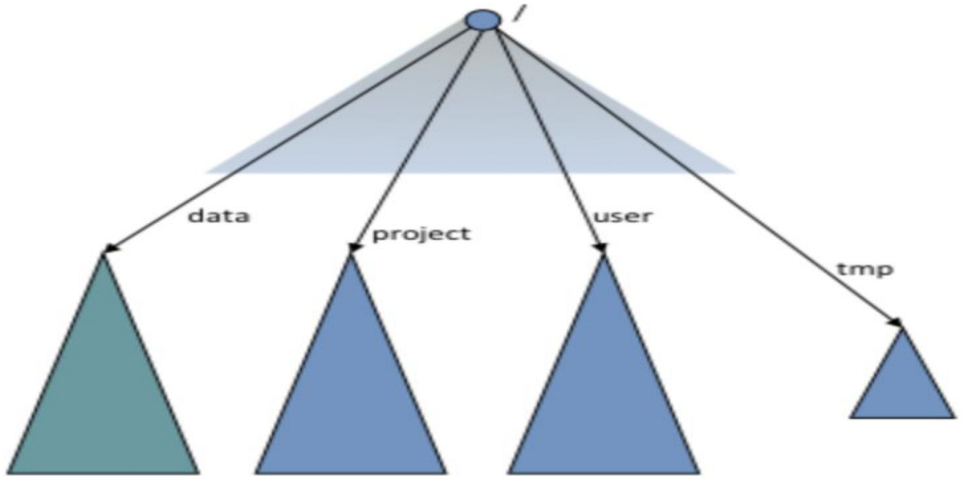


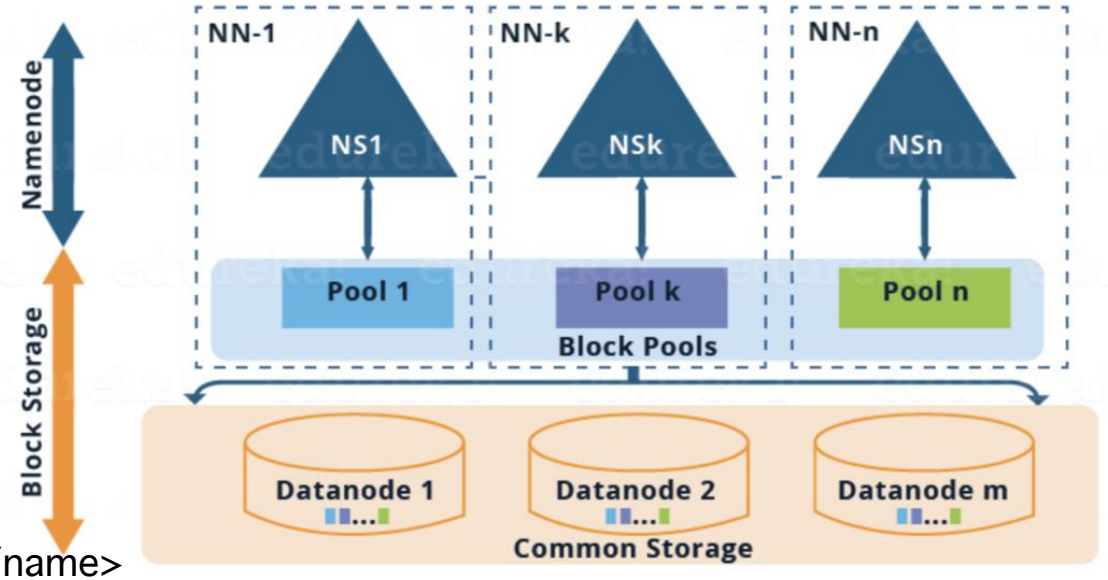
多集群Federation模式



基于Viewfs的Federation模式

客户端的mount table

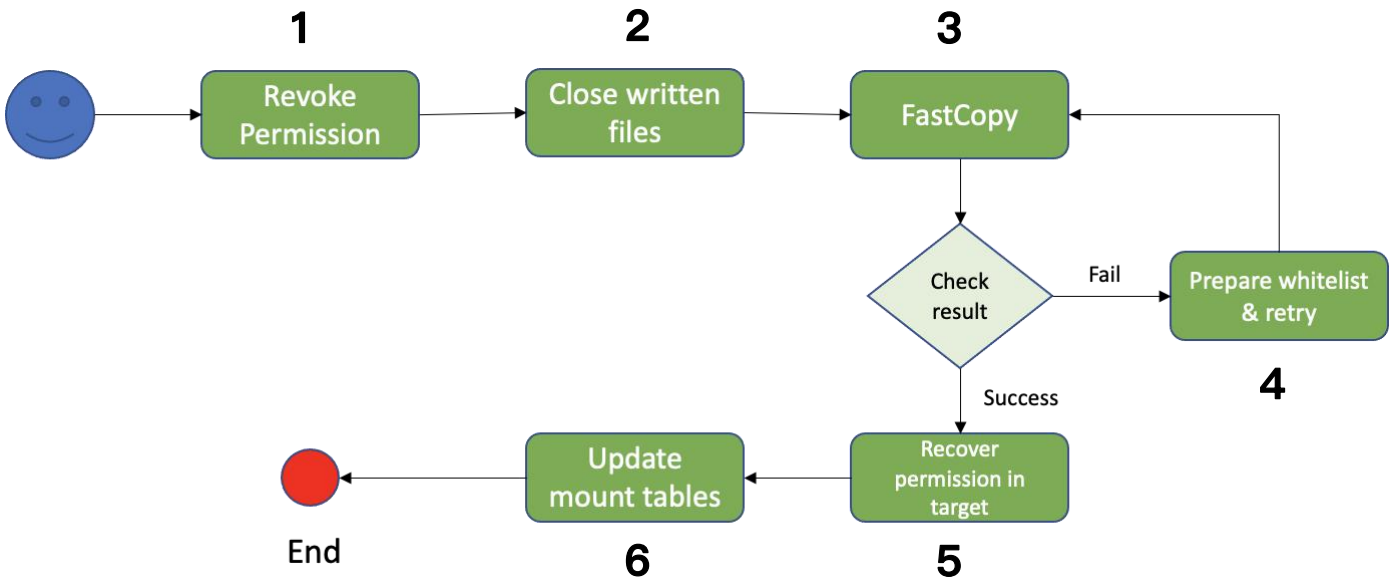




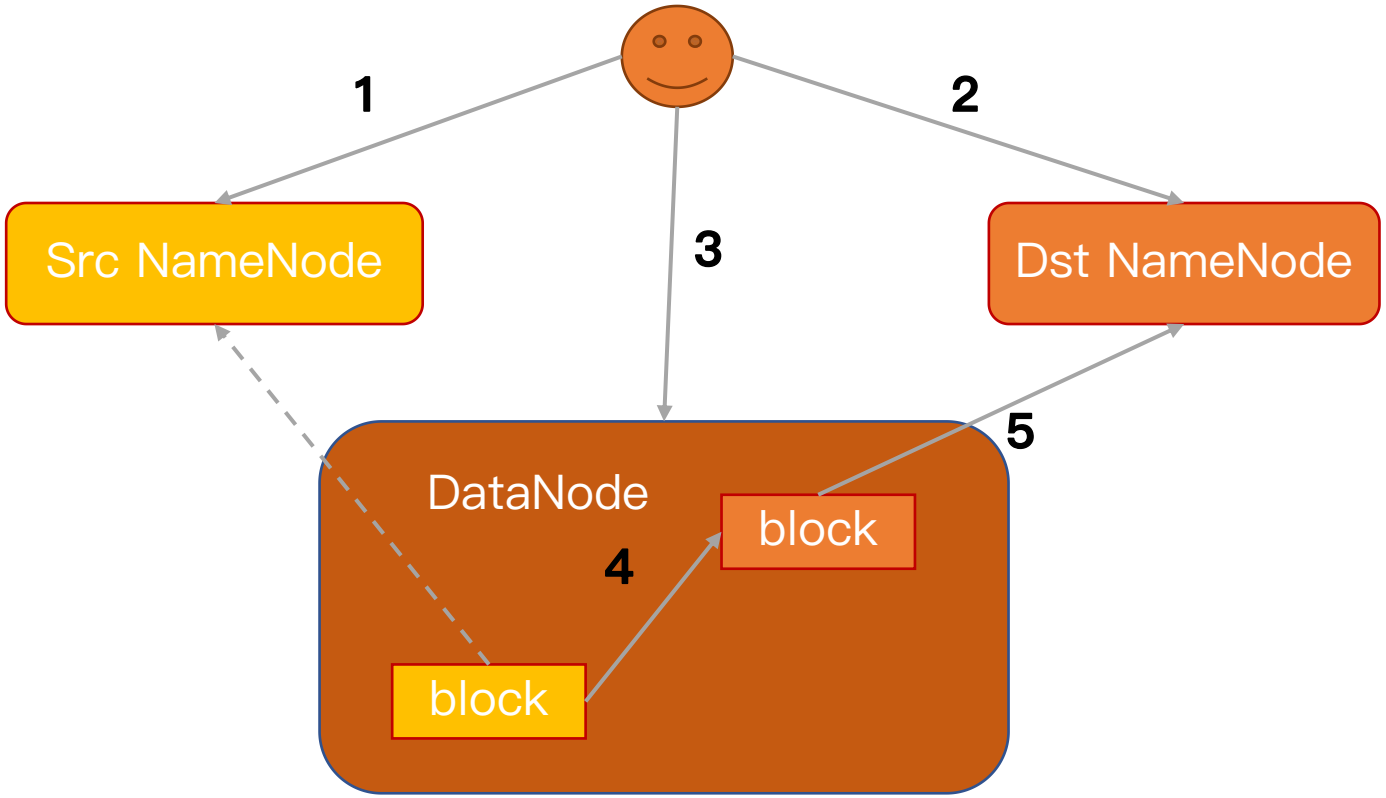

```
<property>
    <name>fs.viewfs.mounttable.{viewfs-name}.link./data</name>
    <value>hdfs://{cluster-name}/data</value>
</property>
```

基于Fastcopy的数据迁移

1. 收回目录权限
2. 关闭open中的文件
3. 使用Fastcopy进行数据的迁移
4. 如果3步骤失败，进行retry
5. 恢复权限
6. 更新mount table信息



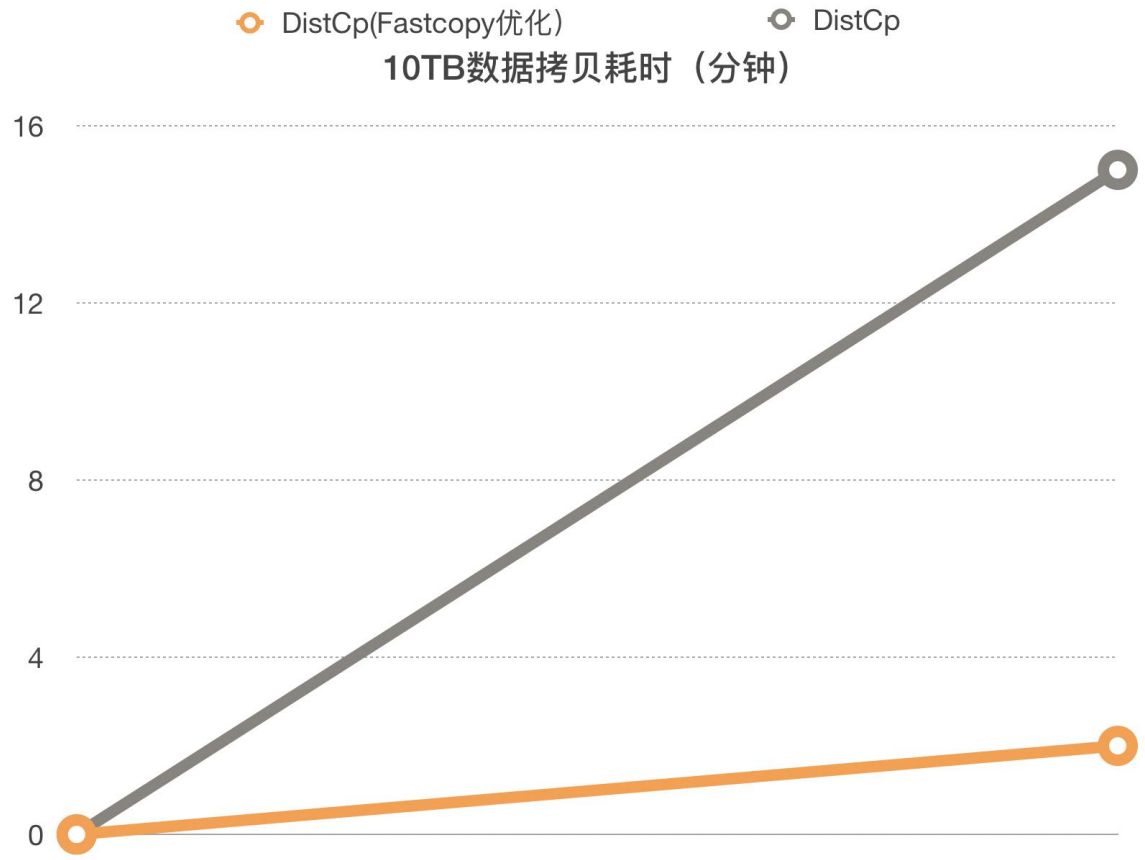
Fastcopy原理



1. Client向源NameNode查询文件block信息
2. 在目标NameNode上创建相应文件，block信息
3. 发送copy block请求到block所属DataNode
4. DataNode创建block，hard link到源block文件
5. 汇报block到目标NameNode

DistCp的Fastcopy集成

将Fastcopy功能集成进DistCp工具里，
性能提升近7倍

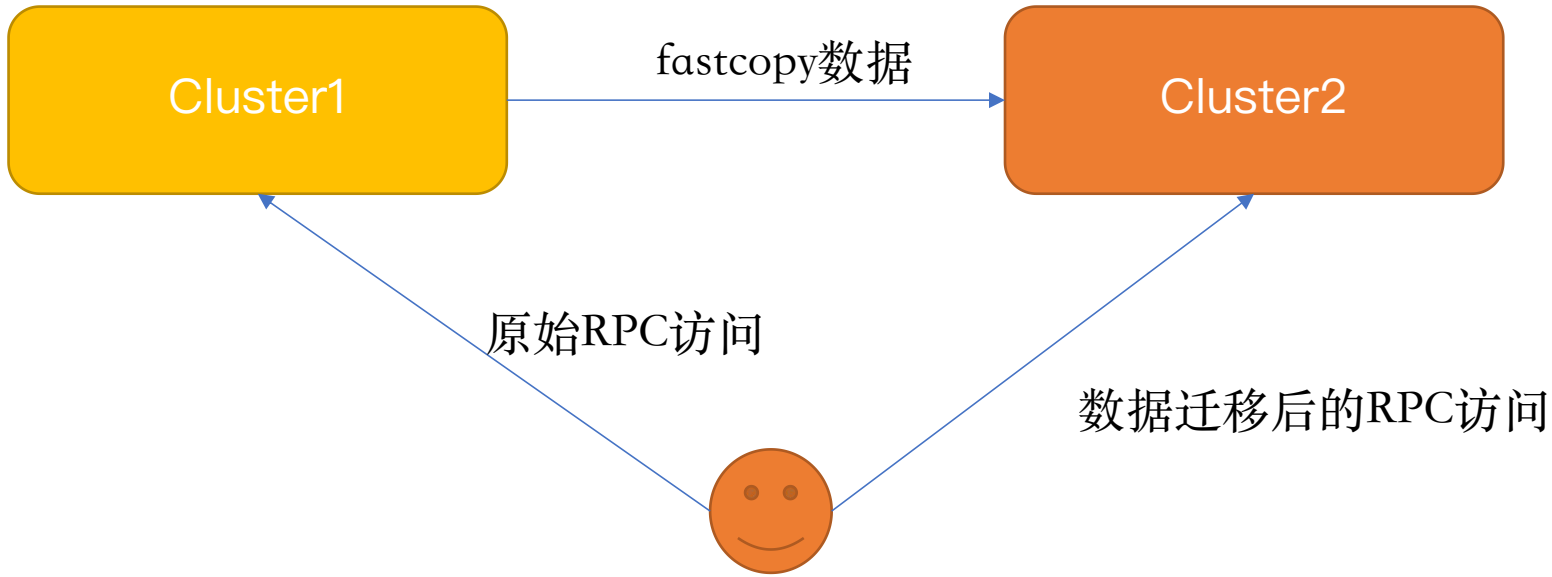


DistCp的其它改进优化

- 统一化大文件小文件的长度，避免出现长尾任务影响
- 目录ACL preserve操作的前置
- DistCp支持whitelist/exclude list的拷贝
- DistCp job的参数调优

集群RPC流量迁移

1. 分析用户数据访问行为，主要检查是否有rename操作的行为
2. Fastcopy数据从源cluster到目标cluster
3. 用户重定向到新cluster进行数据的访问

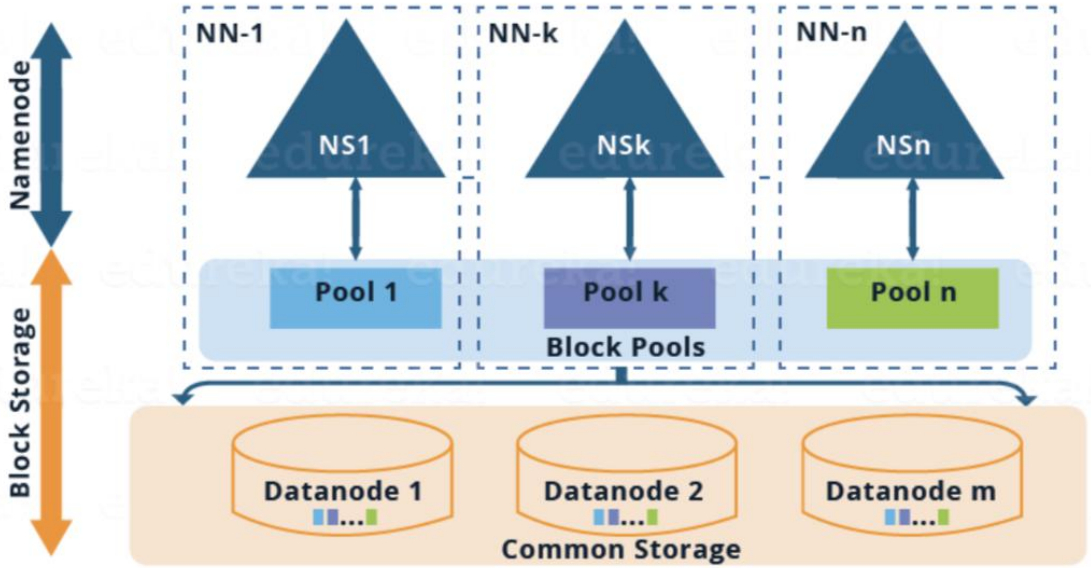


Viewfs Federation方式的问题

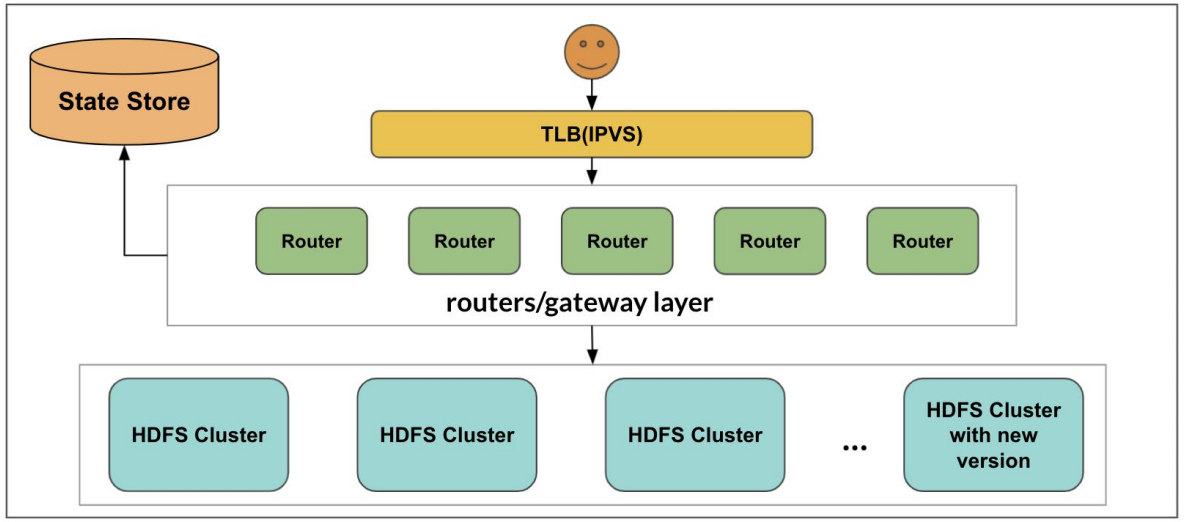
维护成本高	对客户端不透明
随着Federation集群变多，Viewfs的更新维护成本过高，需要在每个client端做更新。	Viewfs对客户端不透明，涉及到底层数据的迁移需要客户端的调整。

HDFS架构的演变

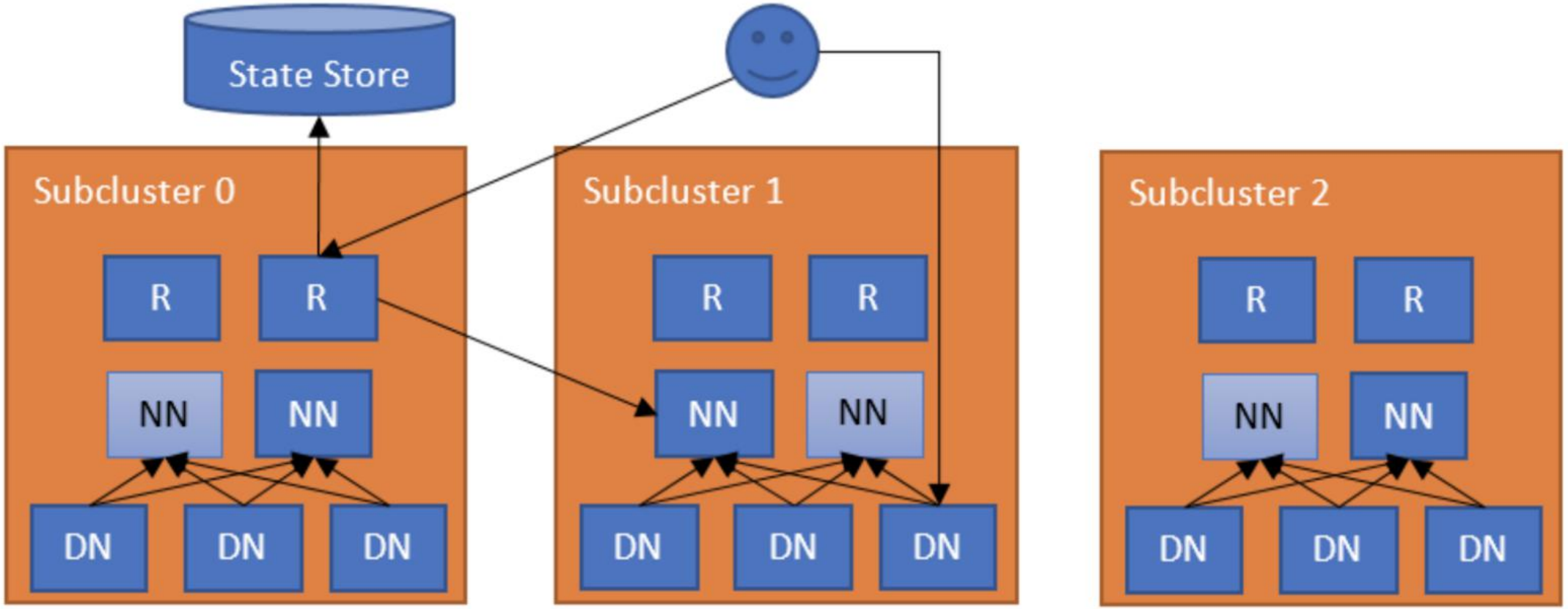
多集群Federation模式



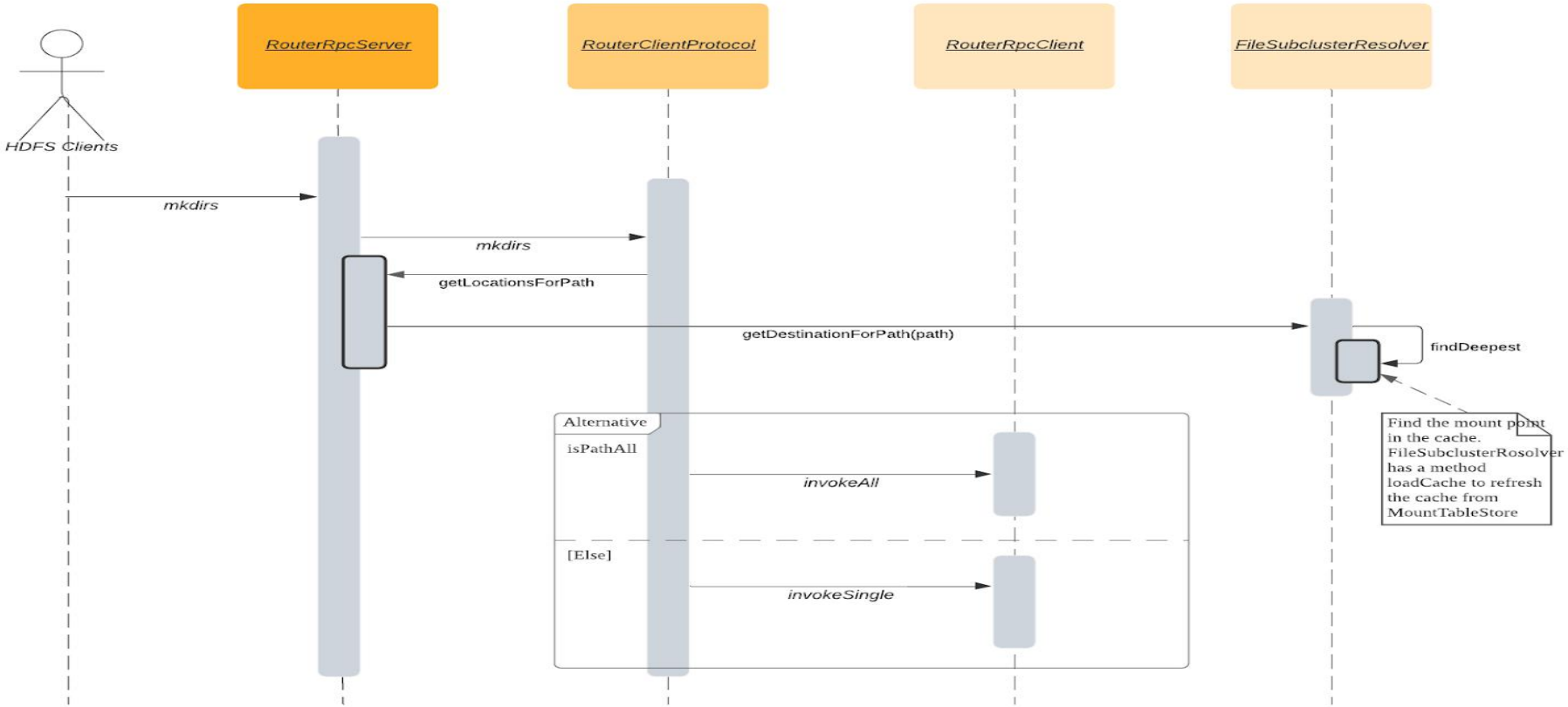
基于Router的Federation模式



HDFS Router-based Federation架构

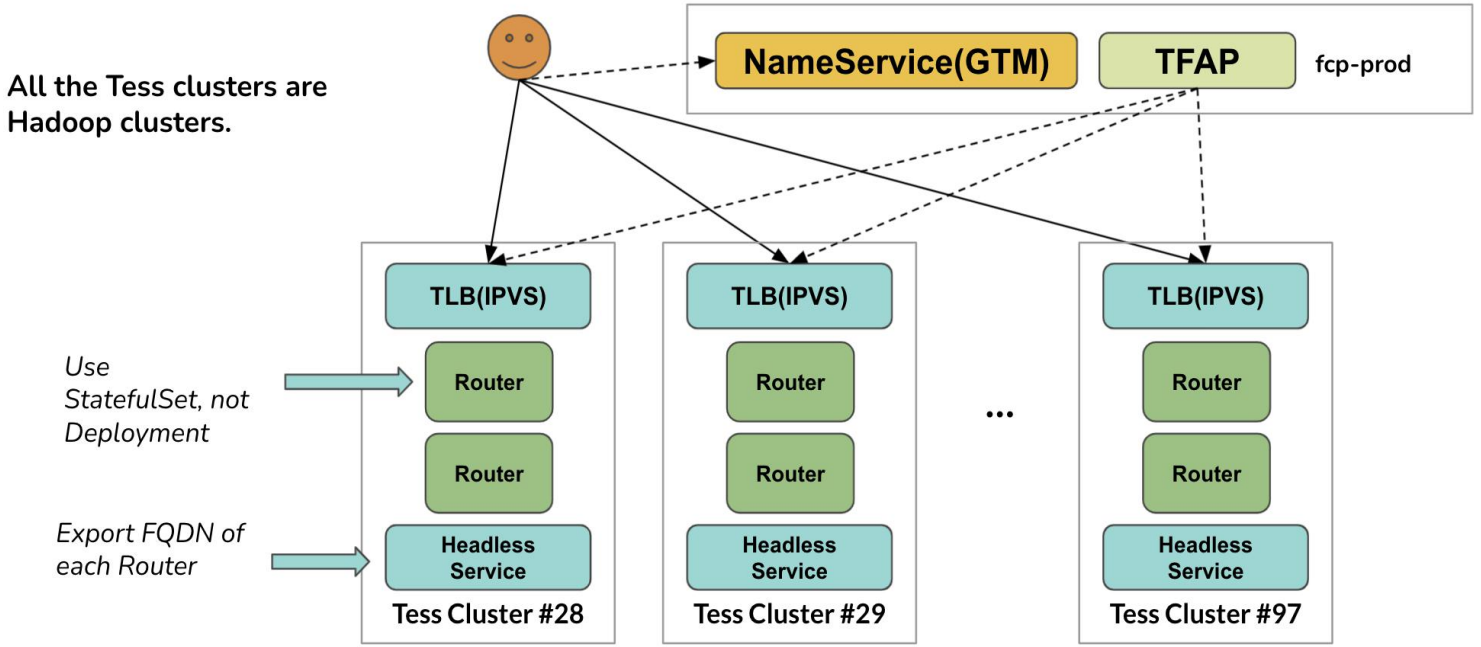


RBF的请求处理过程

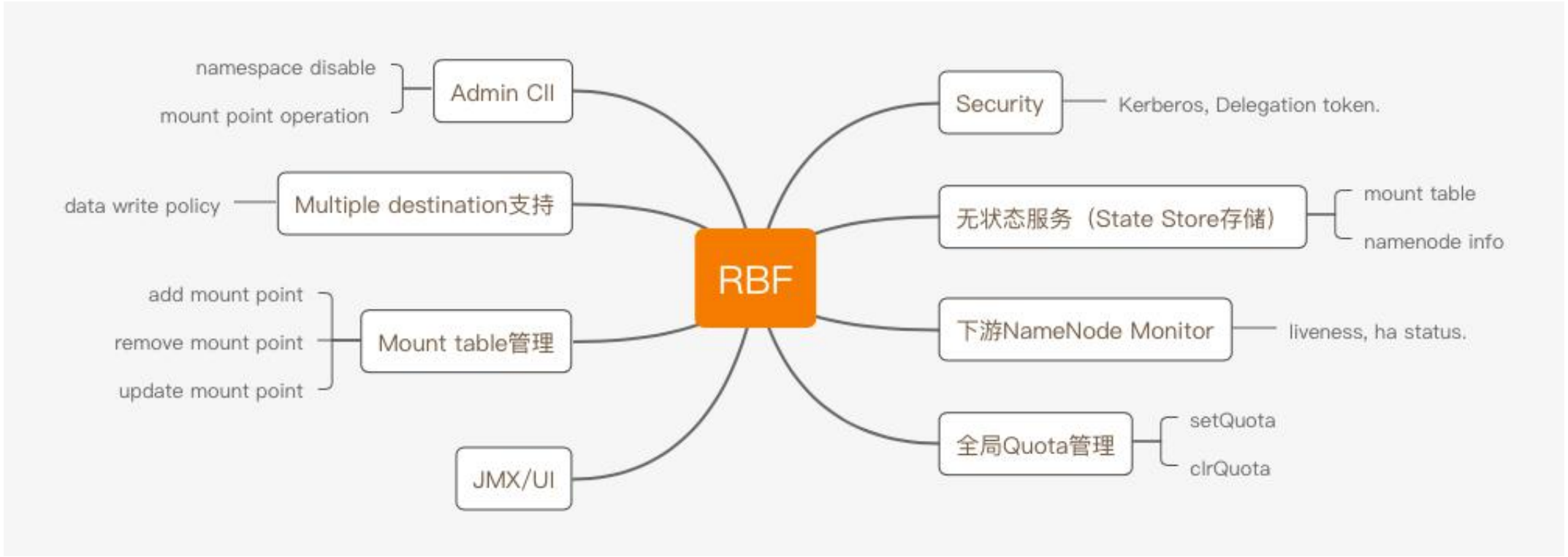

RouterRpcServer
RouterClientProtocol
RouterRpcClient
FileSubclusterResolver
HDFS Clients
mkdirs
mkdirs
getLocationsForPath
getDestinationForPath(path)
findDeepest
Find the mount point in the cache. FileSubclusterRosolver has a method loadCache to refresh the cache from MountTableStore
Alternative
isPathAll
invokeAll
[Else]
invokeSingle

RBF架构模式的优势

- 无状态的服务，cloud-native化部署，方便进行横向扩展
- Federation路径更新对用户完全透明，用户无需进行任何更新
- 可基于RBF架构做数据split拆分的方案



RBF的功能特性


RBF
Admin CII
namespace disable
mount point operation
Multiple destination支持
data write policy
Mount table管理
add mount point
remove mount point
update mount point
JMX/UI
Security
Kerberos, Delegation token.
无状态服务（State Store存储）
mount table
namenode info
下游NameNode Monitor
liveness, ha status.
全局Quota管理
setQuota
clrQuota

eBay RBF的优化

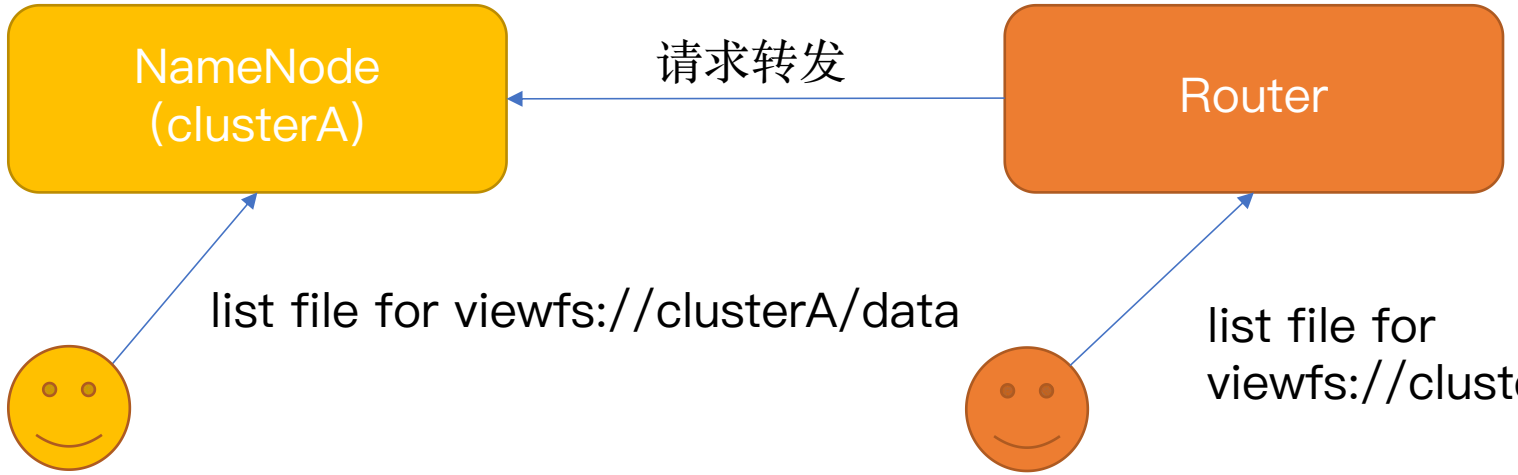
RBF的平滑部署

- Viewfs到RBF的兼容性支持
- YARN RM Security模式补token逻辑的改造

RBF的性能改造

- Router服务支持更大的RPC吞吐量
  - 解决Router内部的连接复用问题
  - 去掉Router和NameNode之间的Sasl加密操作
- Router支持客户端ip地址，clientId的保留，不会影响到任务data locality的读写
- 多挂载点模式下，moveToTrash文件删除问题的解决

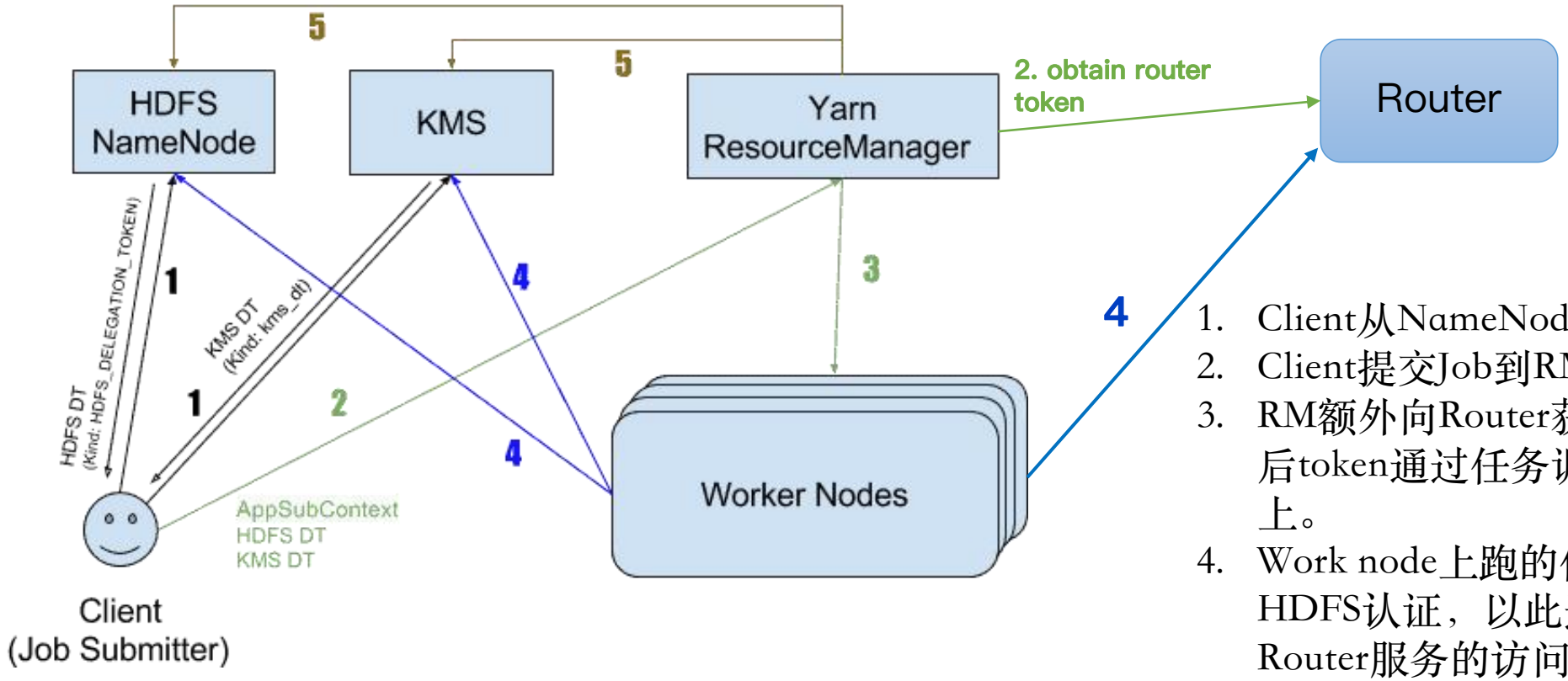
Viewfs到RBF的兼容性改造



Key	Value
fs.defaultFS	viewfs://clusterA
fs.viewfs.mounttable.clusterA.link./data	hdfs://clusterA/data

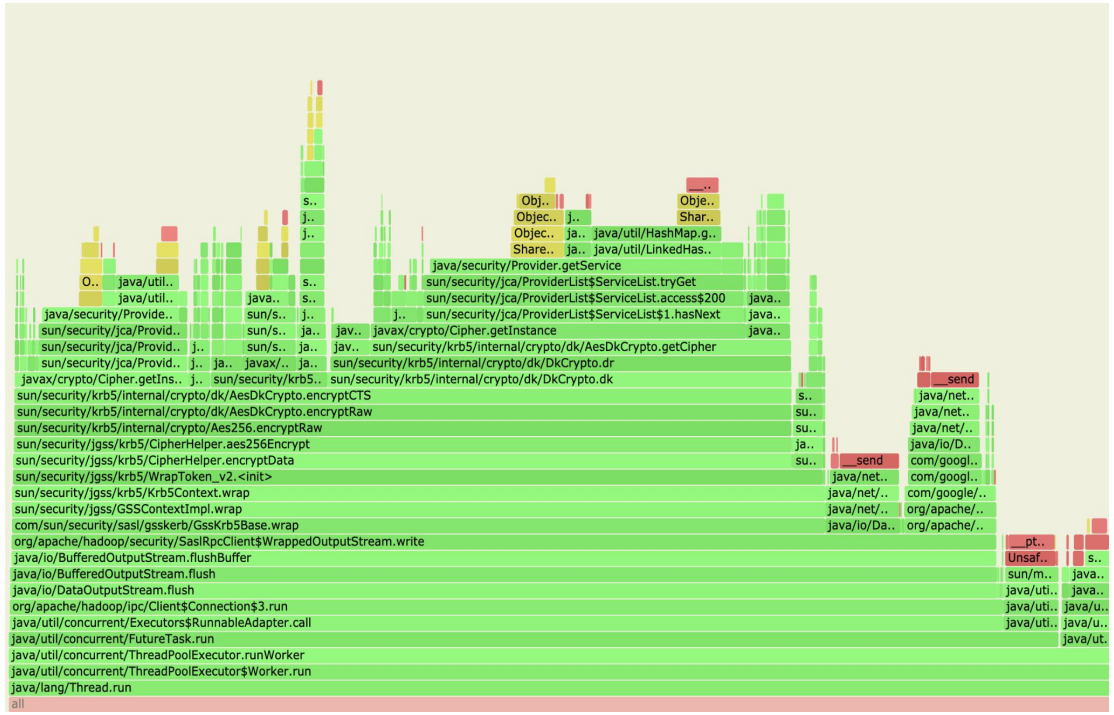
Key	Value
fs.defaultFS	viewfs://clusterA
fs.viewfs.clusterA.link	hdfs://router/
fs.viewfs.mounttable.clusterA.link./data	hdfs://clusterA/data

RBF补token改造

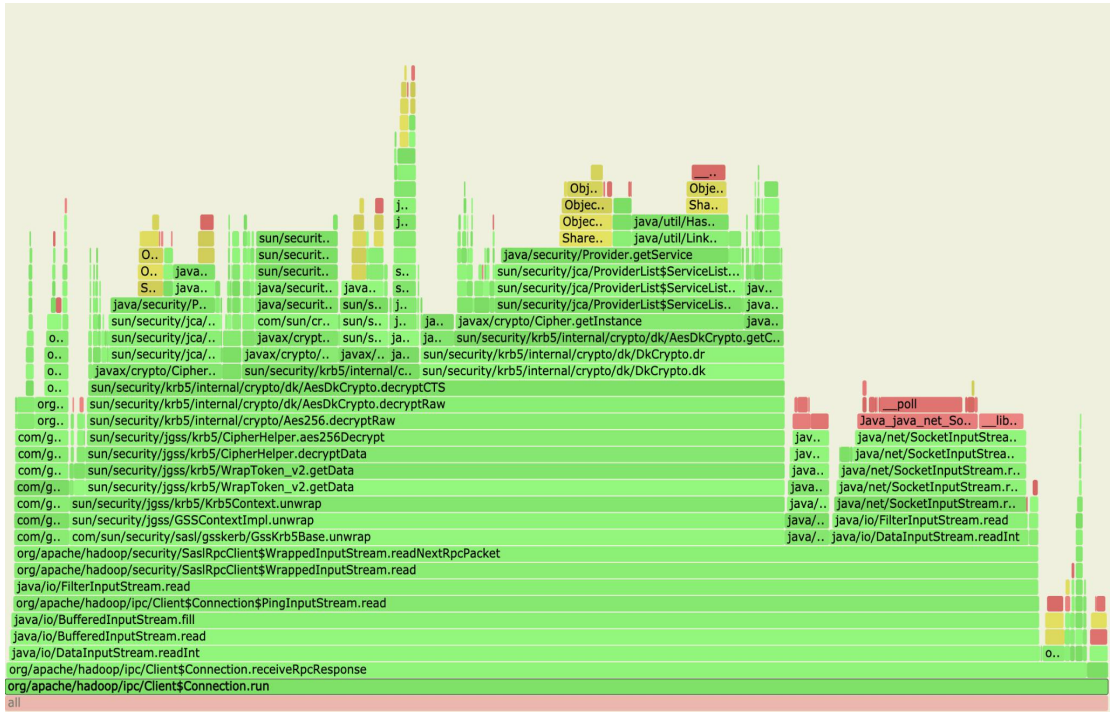


1. Client从NameNode获取token
2. Client提交Job到RM（携带token）
3. RM额外向Router获取Router token，随后token通过任务调度下发到work node上。
4. Work node上跑的任务通过token来做HDFS认证，以此进行HDFS数据和Router服务的访问。
5. Job运行结束，RM进行token的删除操作。

Router RPC的Sasl加密



RPC加密

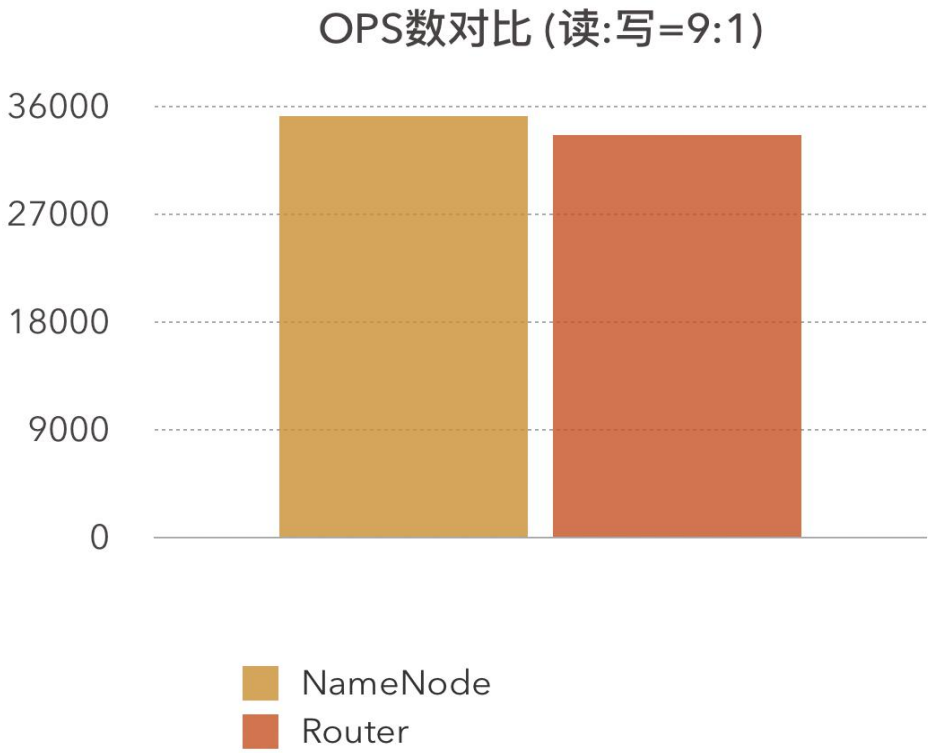


RPC解密

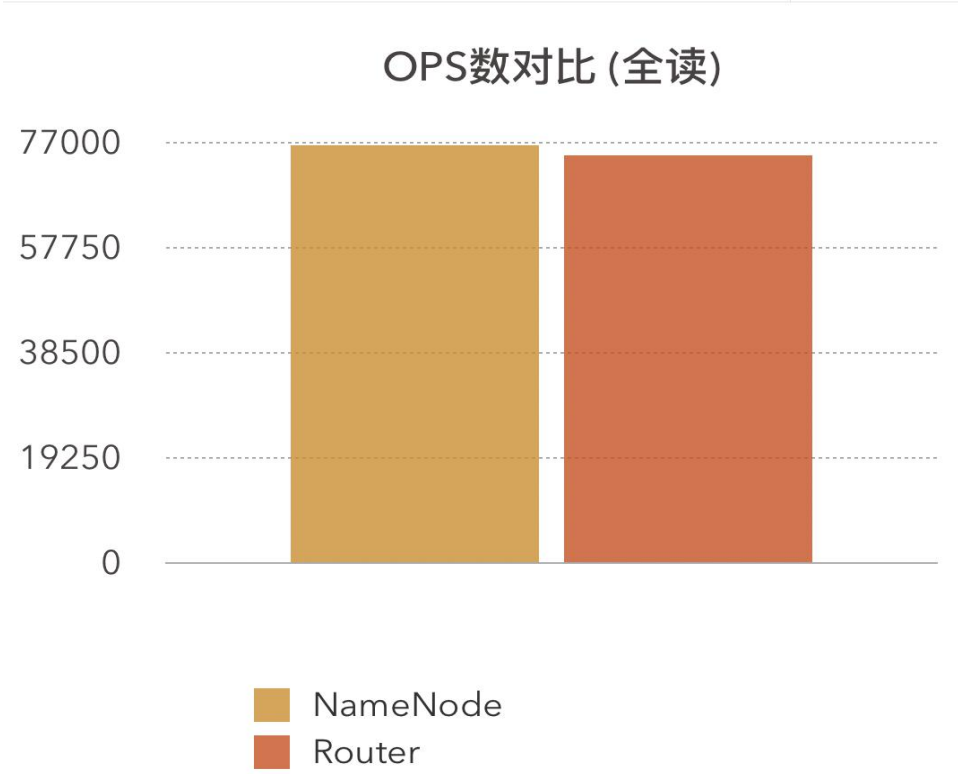
RBF性能测试

操作类型	操作数（直连NN）	操作数（RBF模式）
Open	11534	12095
getFileInfo	10919	11824
Create	10805	11012
Mkdirs	11647	11159
Delete	11005	11642
Rename	11880	11362

RBF性能测试



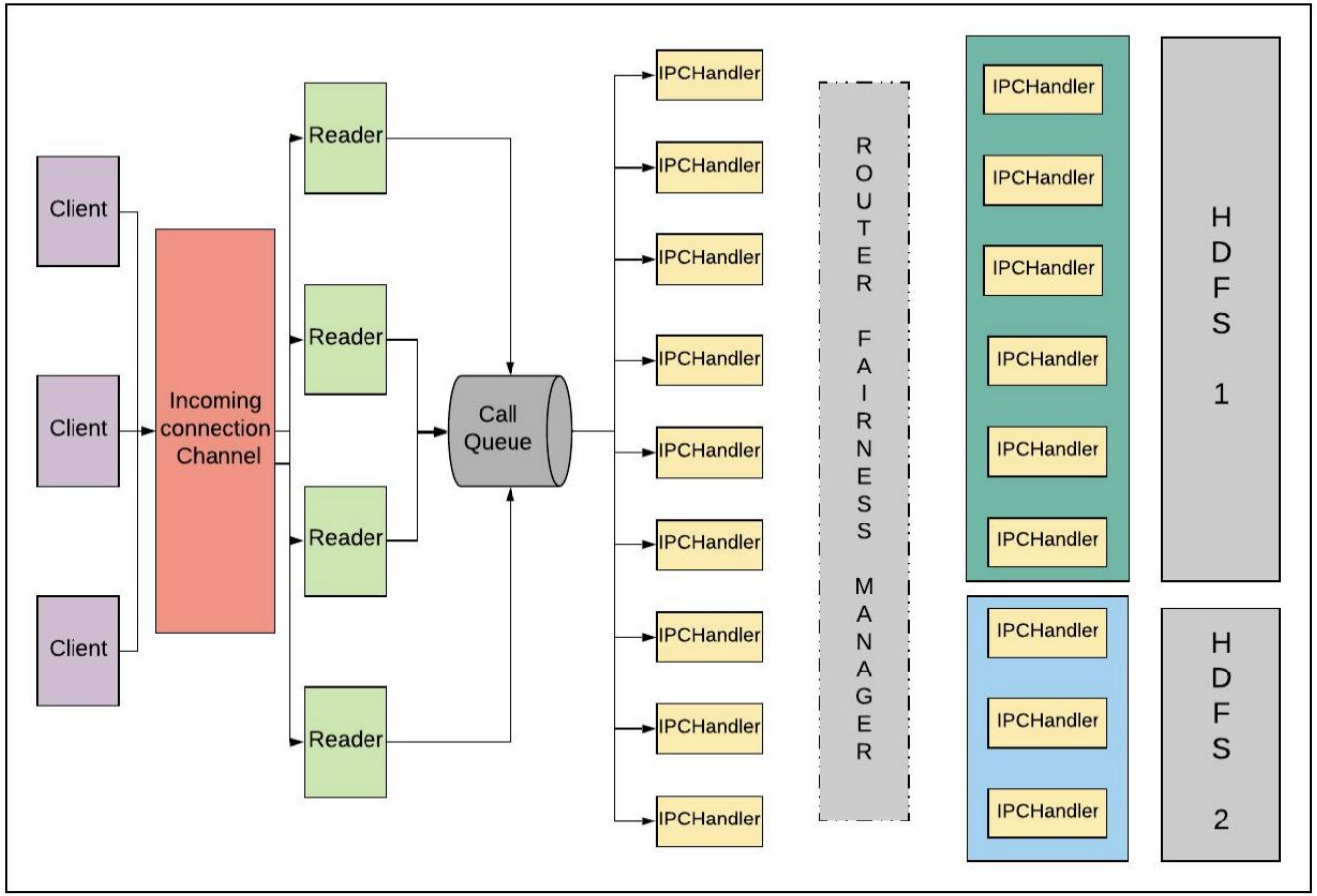
OPS数: 35.2k/33.6k(**4.6**%的差距)



OPS数: 76.5k/74.7k(**2.4**%的差距)

RBF未来展望

- RBF异步化RPC处理来进一步提升RPC吞吐量
- 基于RBF做更为自动化的数据split拆分
- 基于RBF模式下做Tiered Storage，提升集群存储的效率
- RBF对底层namespace间RPC处理的隔离




THANKS
时序
开源
自研
分布式
文档
Spark
HANA
SQL
OLAP
Oracle
Hadoop
MPP
DB2