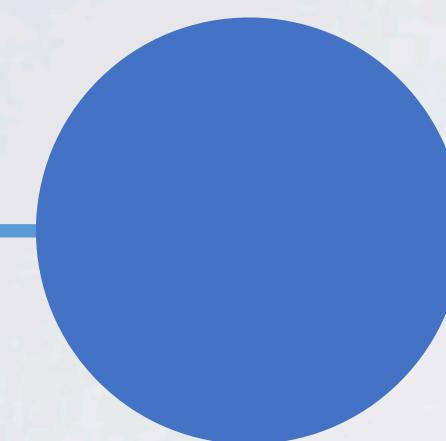




小米HDFS扩展性解决方案

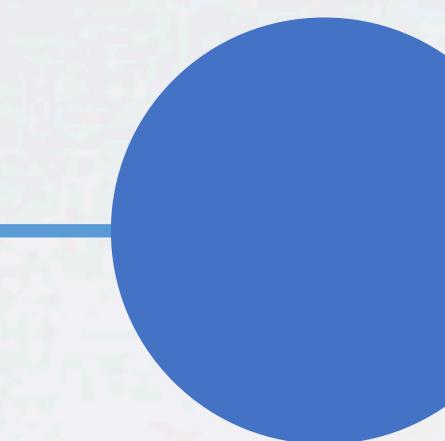
2017

单NameNode部署
到达瓶颈



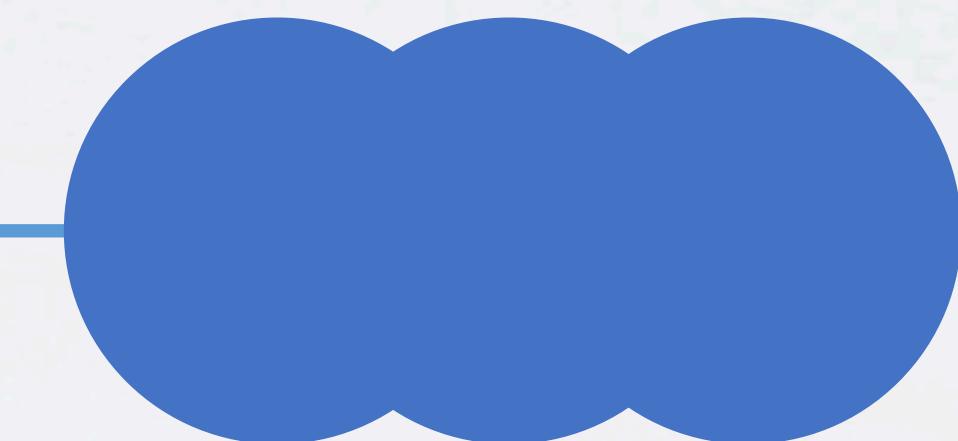
2018

国内主力集群完成
Federation改造



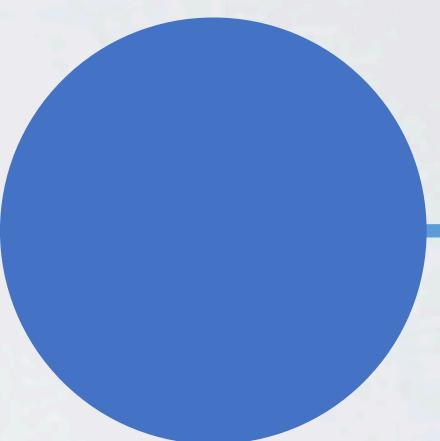
2019

NameNode负载
不均，内存不足



2020

RPC排队时间显著
变长，影响作业执
行



2019

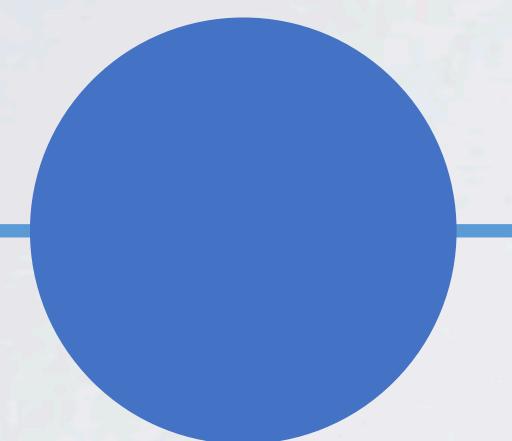
Federation集群
改造RBF

2019

单机房扩展性受限

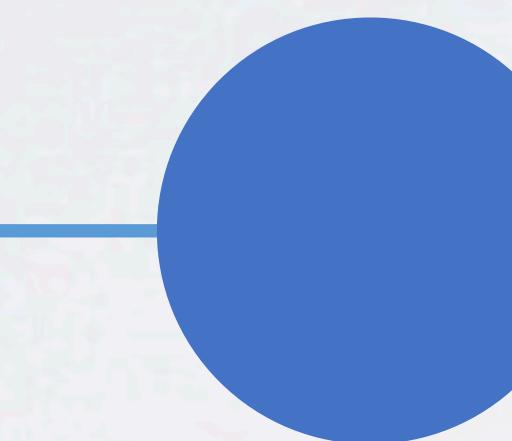
2017

单NameNode部署
到达瓶颈



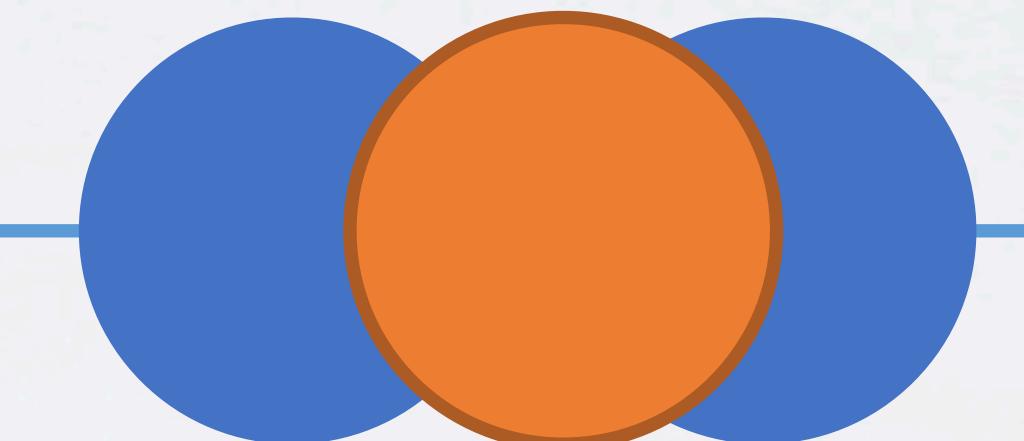
2018

国内主力集群完成
Federation改造



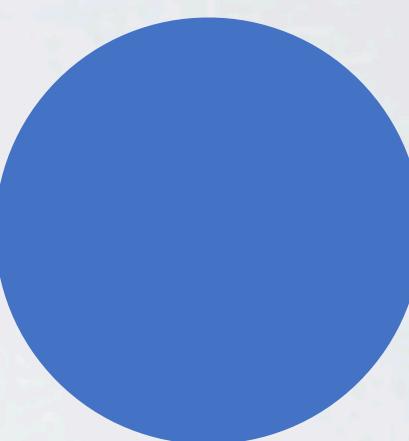
2019

NameNode负载
不均，内存不足



2020

RPC排队时间显著
变长，影响作业执
行



2019

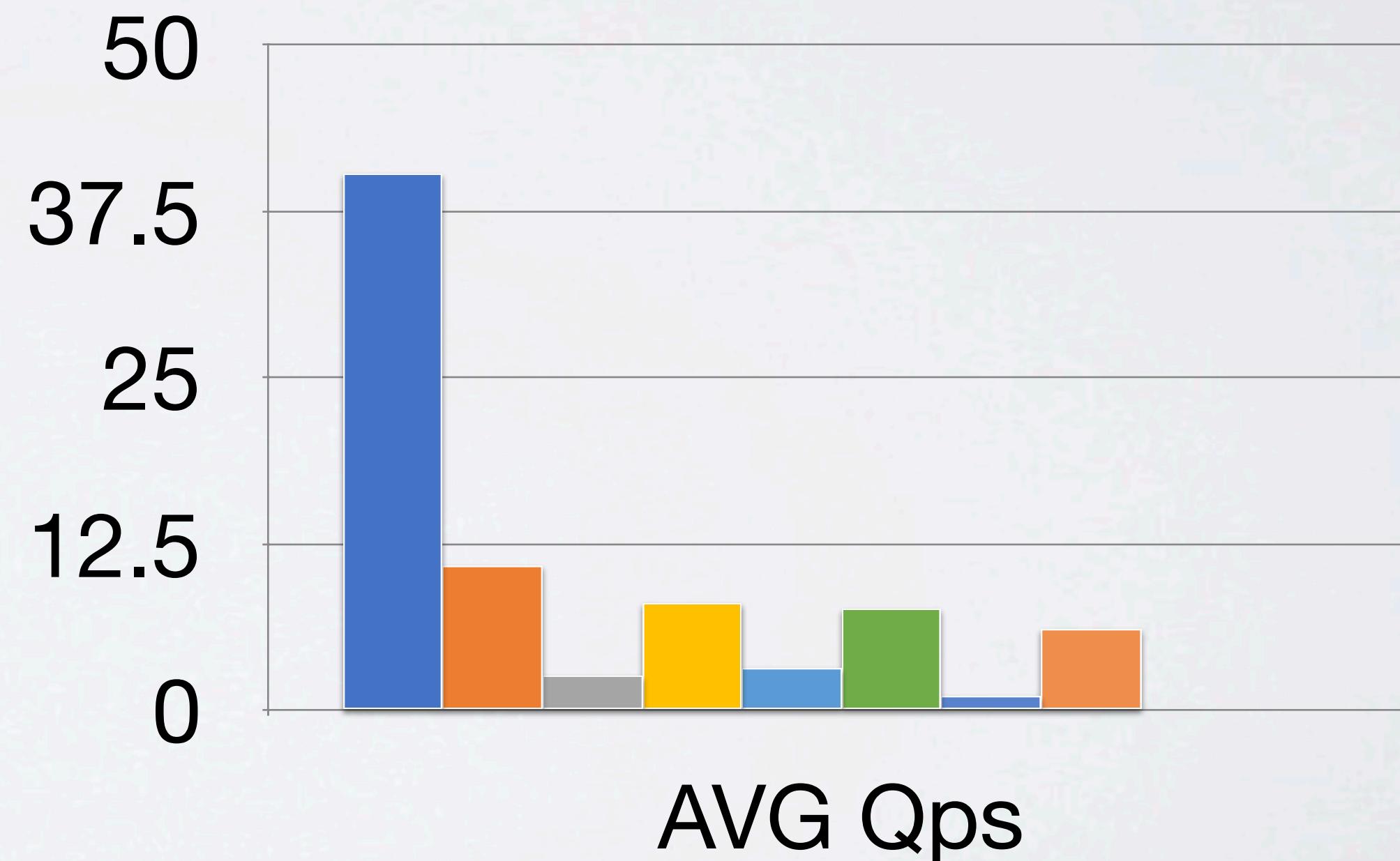
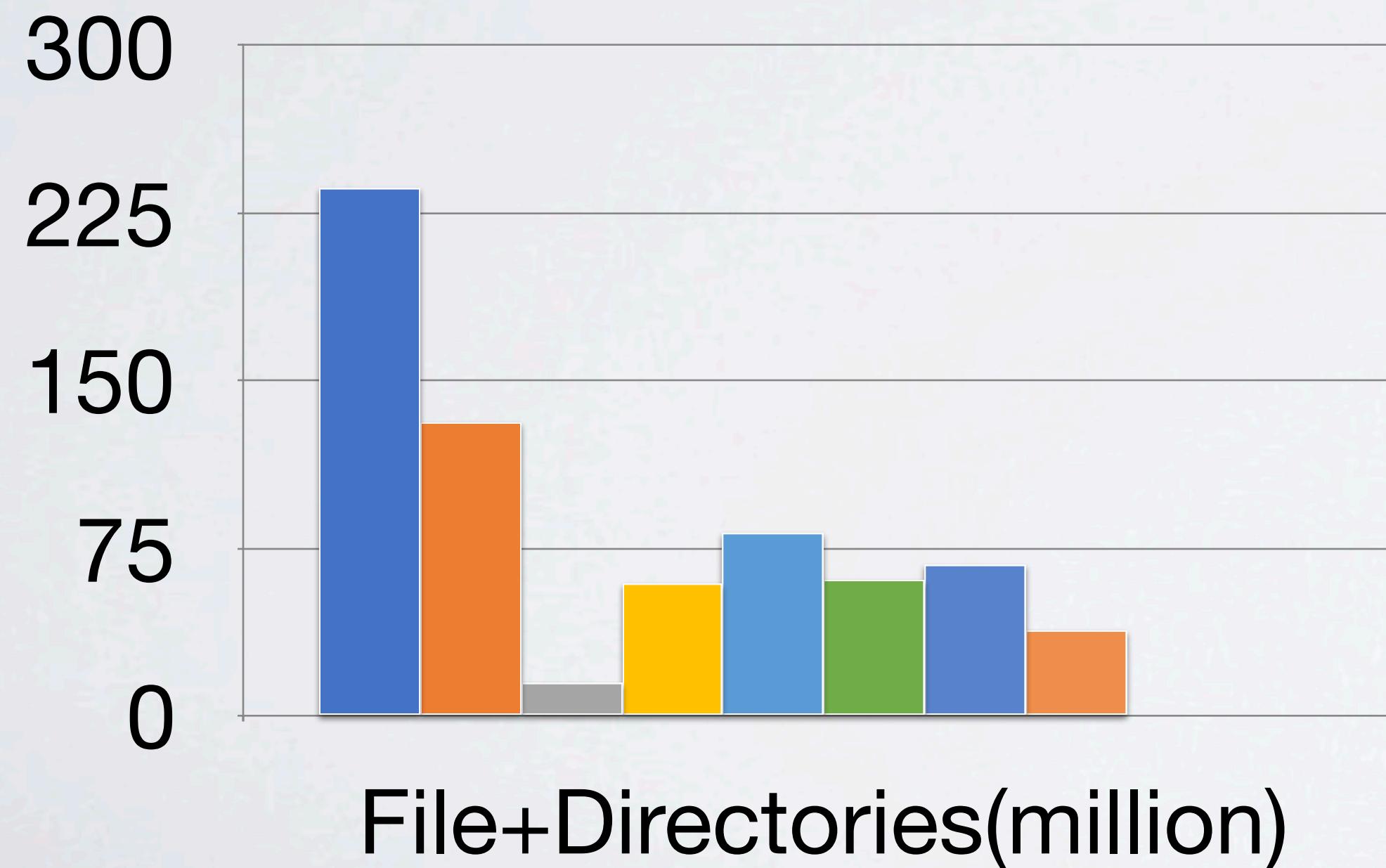
Federation集群
改造RBF

2019

单机房扩展性受限

多组Namespace负载不均衡

- 堆大小接近128GB
- RPC未来很可能成为瓶颈



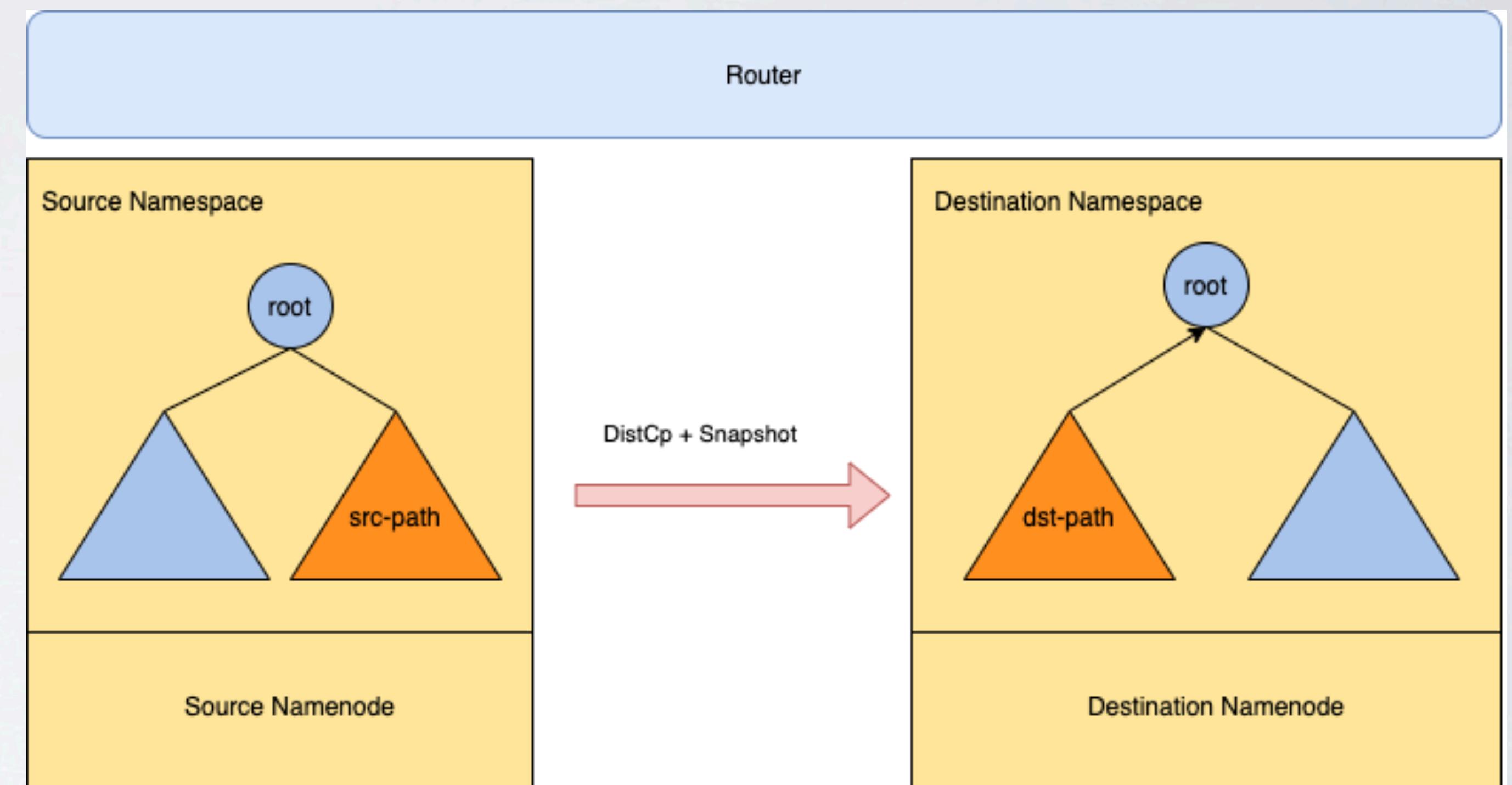
Federation Balance方案

问题

- 怎么迁移数据?
- 怎么让Client找到新Namespace?

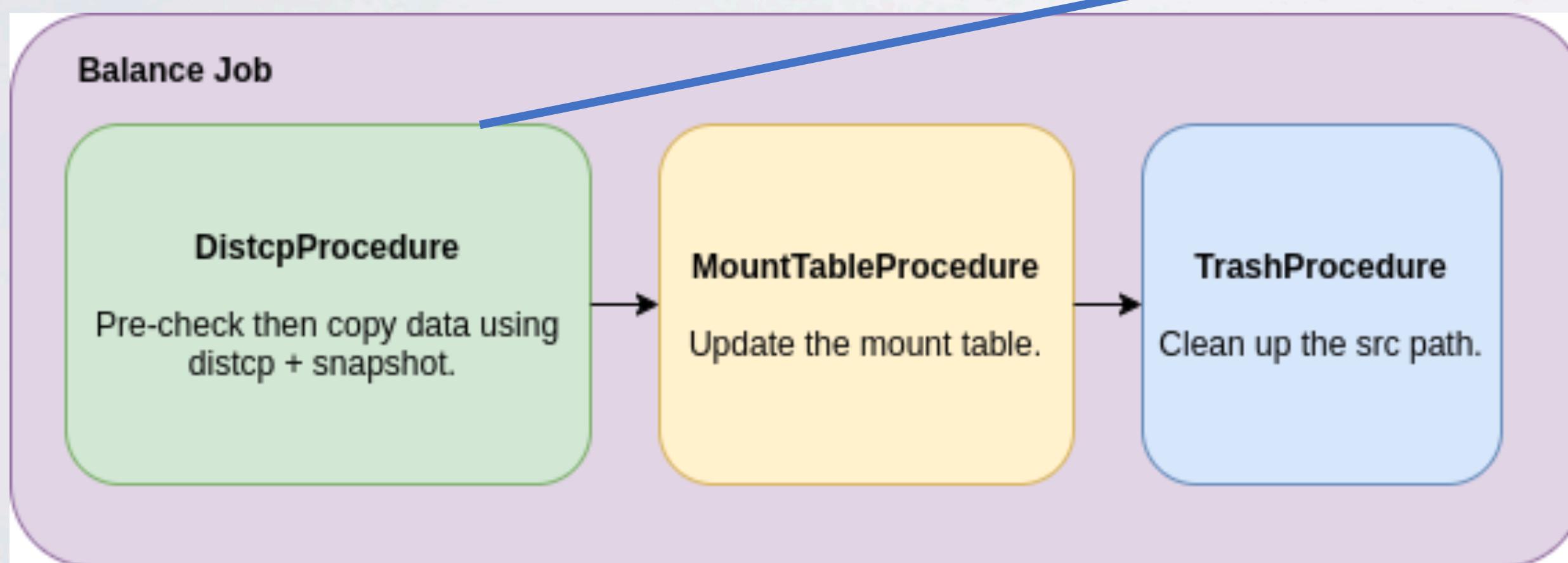
方案

- 使用Balance Job整合Balance过程
- 使用Procedure Scheduler调度Job



Balance Job

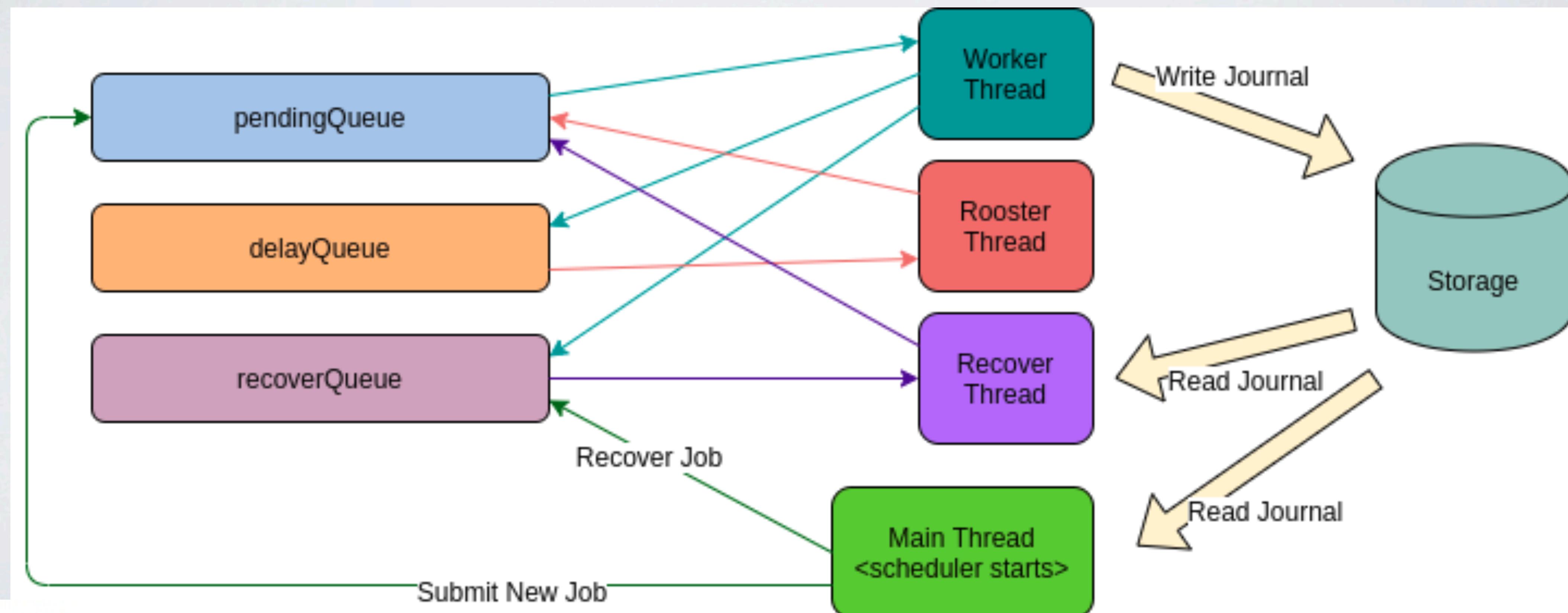
- Balance Job: DistCpProcedure、MountTableProcedure、TrashProcedure。
- 每个Procedure可以包含一个或多个Phase。



PRE_CHECK	预检查工作
INIT_DISTCP	第一轮DistCp
DIFF_DISTCP	一轮一轮DistCp直到没有diff
DISABLE_WRITE	禁止写操作
FINAL_DISTCP	最后一轮DistCp
FINISH	完成

ProcedureScheduler

- ProcedureScheduler是一个状态机，管理Job的整个生命周期：提交、执行、重试、错误恢复

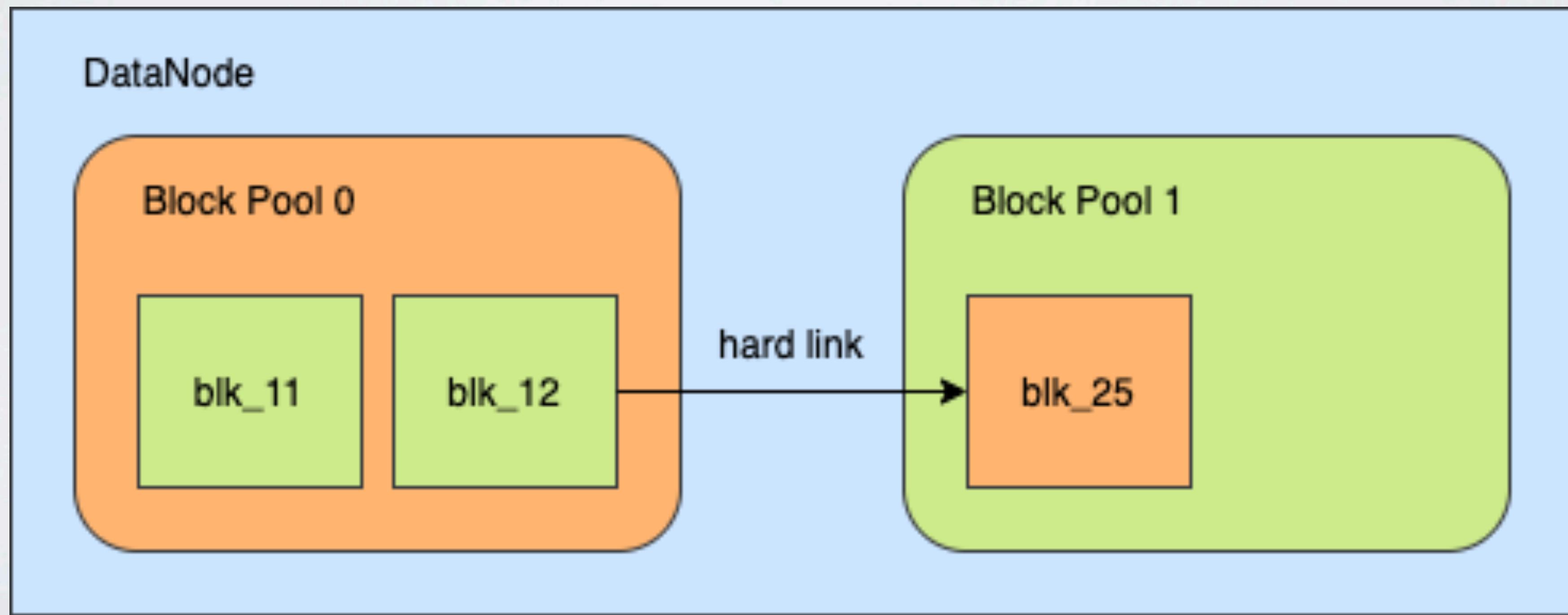


Federation Balance Tool

- Federation balance tool: <https://issues.apache.org/jira/browse/HDFS-15294>

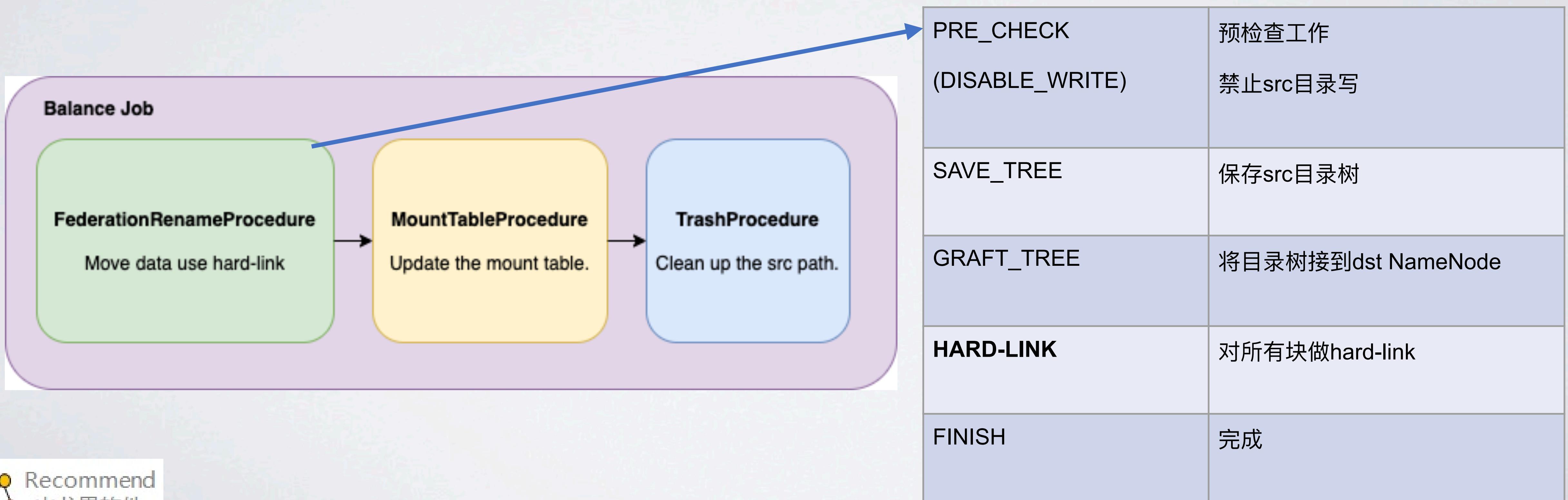
能更快一点吗？

- NameNode支持元数据迁移
- 使用**hard-link**拷贝data



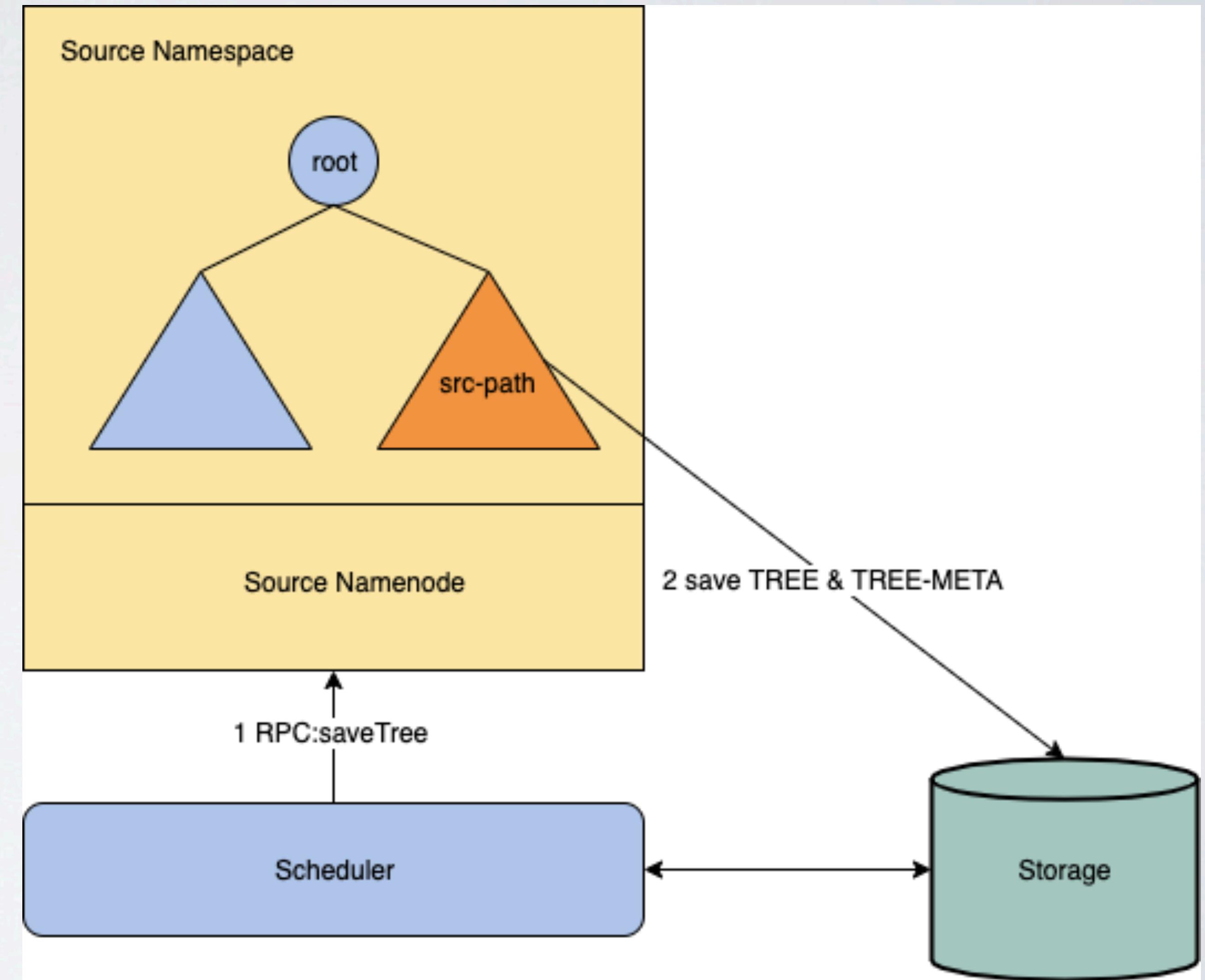
Balance Job (hard-link)

- DistCpProcedure -> FederationRenameProcedure



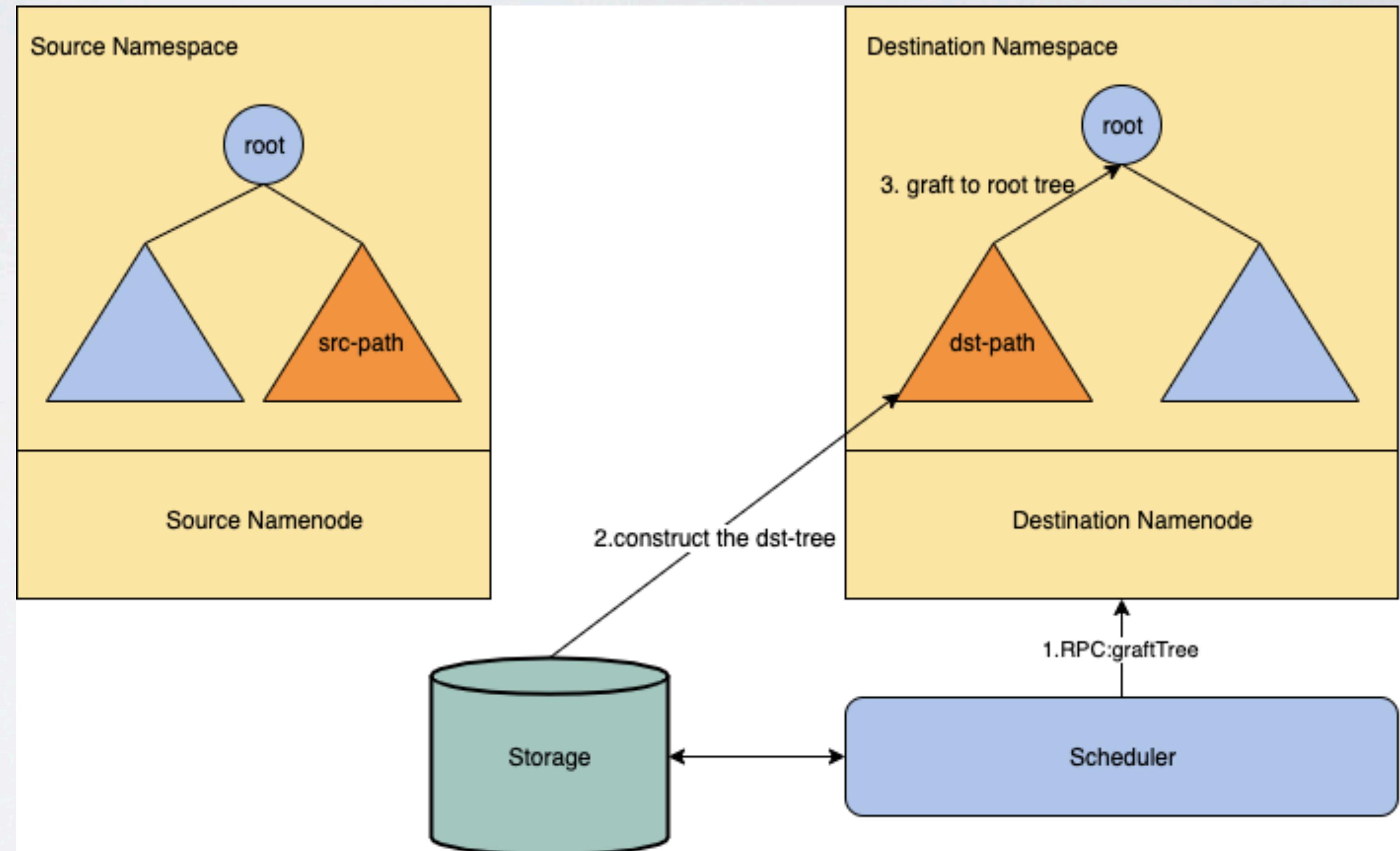
FederationRenameProcedure

- SaveTree



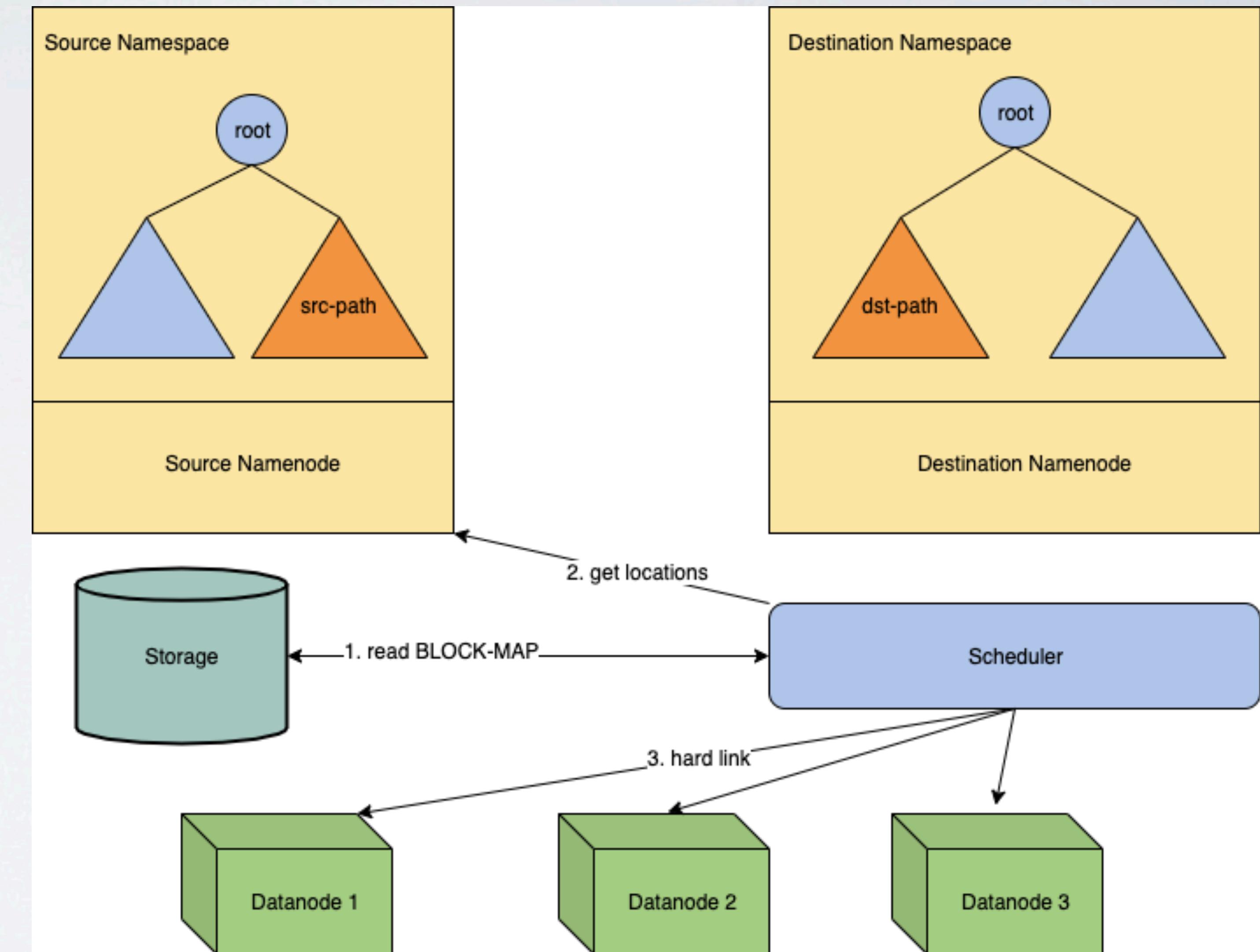
FederationRenameProcedure

- GraftTree



FederationRenameProcedure

- HardLink



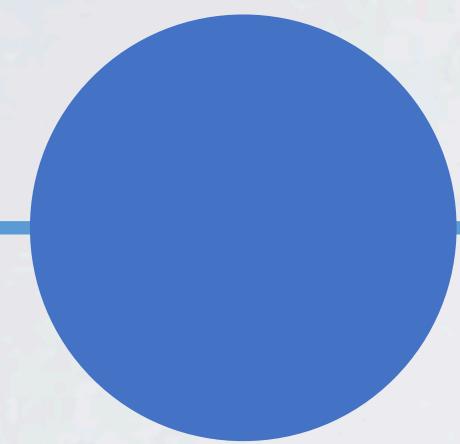
FederationRenameProcedure Performance

- 2个NameSpace, 14个DataNode, 3副本
- Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz, 12 cores.
- 128GB RAM, 4T x 12 HDD
- Linux 2.6.32, JDK 8

Data sets	Directories	Files	Blocks	Time costs/ms	File Size(BlkSize=256MB)
set 7-7	19608	117649	117649	18,001	28.72TB
set 7-8	37449	262144	262144	30,890	64TB
set 8-9	597871	4782969	4782969	577,360	1.14PB

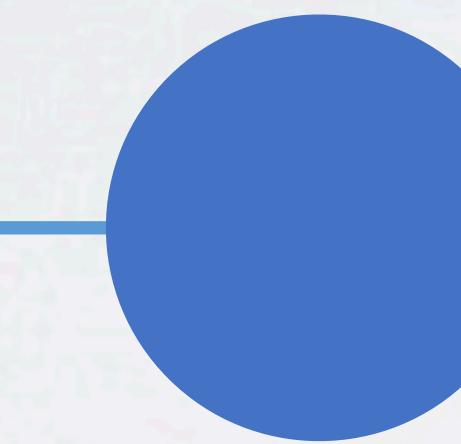
2017

单NameNode部署
到达瓶颈



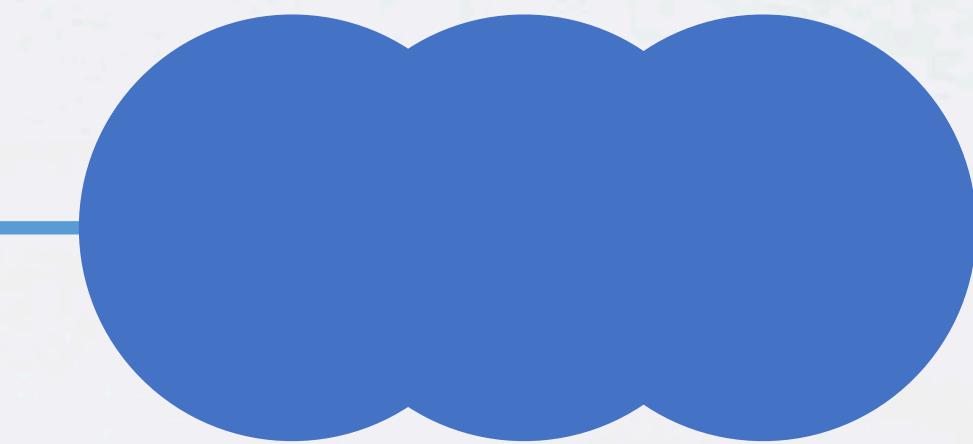
2018

国内主力集群完成
Federation改造



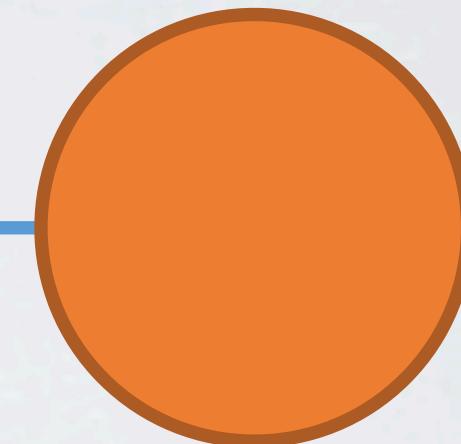
2019

NameNode负载
不均，内存不足



2020

RPC排队时间显著
变长，影响作业执
行



2019

Federation集群
改造RBF

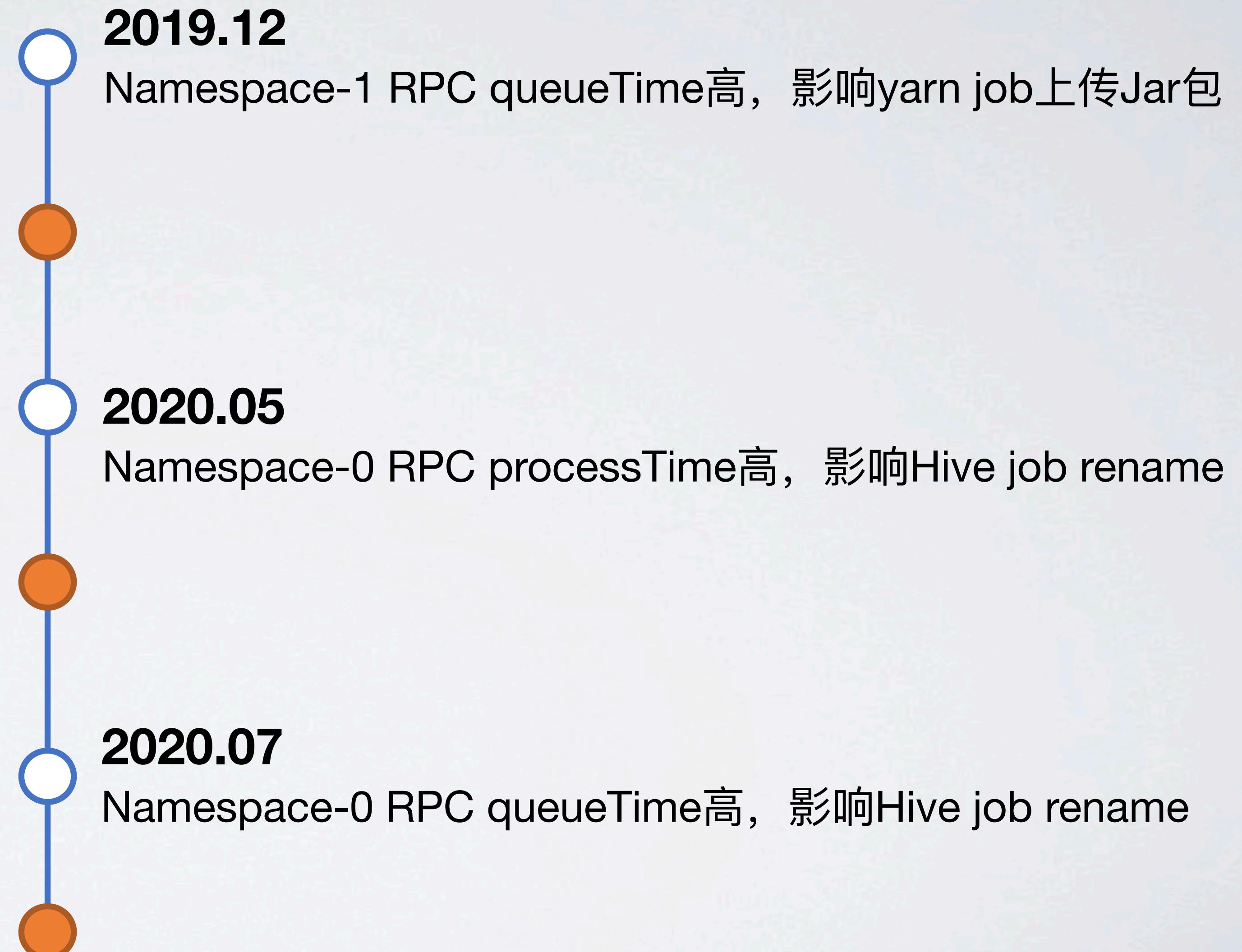
2019

单机房扩展性受限

解决办法
迁移yarn job log到独立Namespace

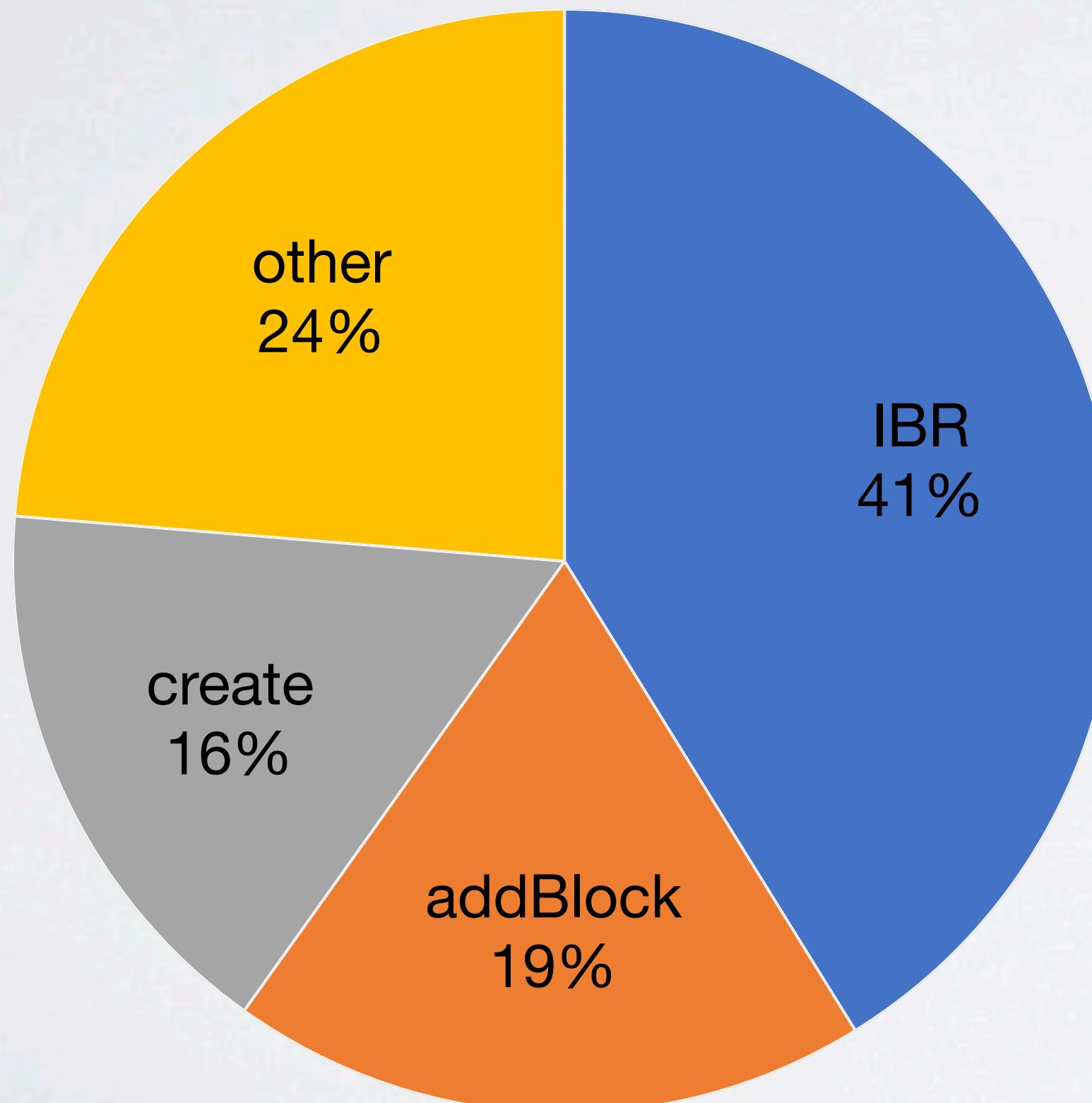
解决办法
定位读RPC突增原因，推动用户优化

解决办法
RBF多挂载点改造

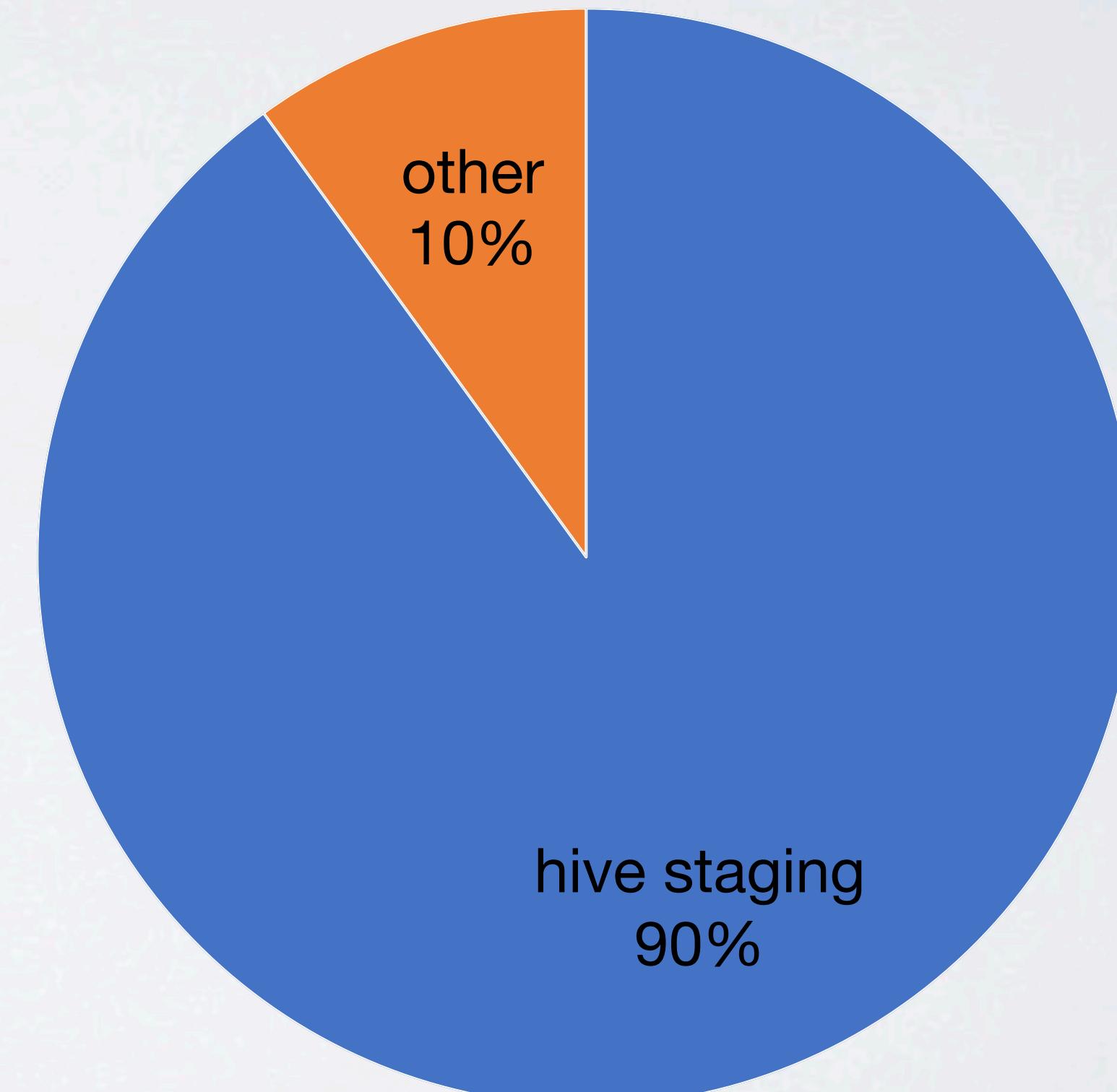


RPC来源

主力NS写RPC qps

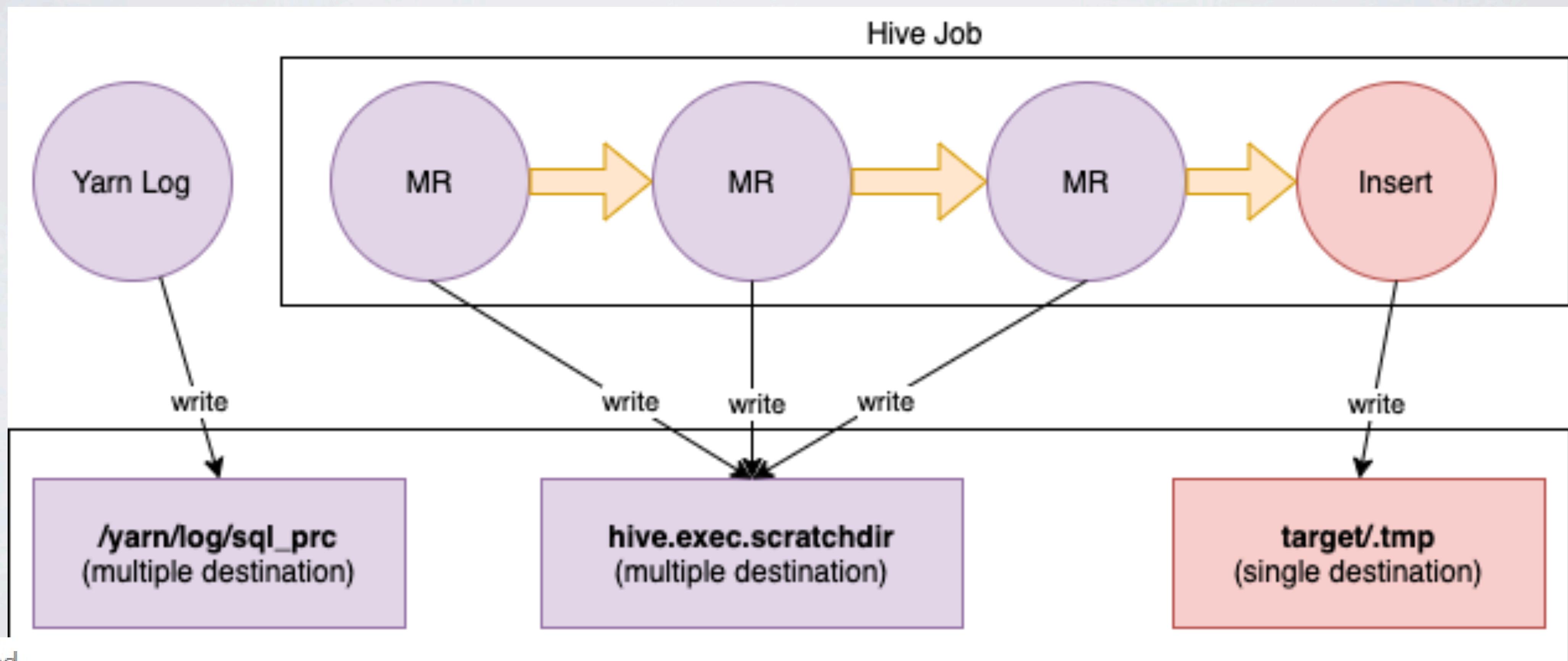


主力NS写目录占比



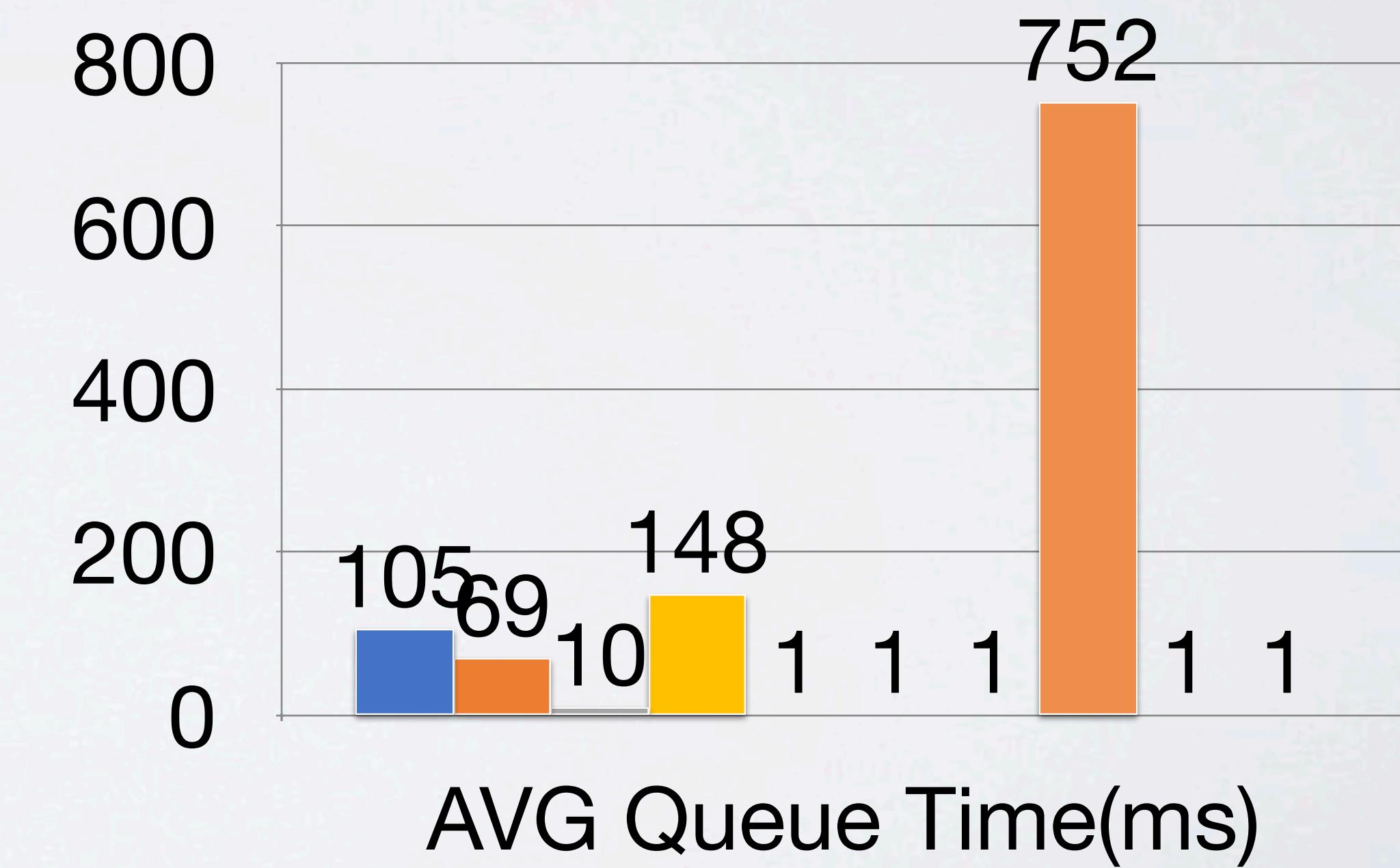
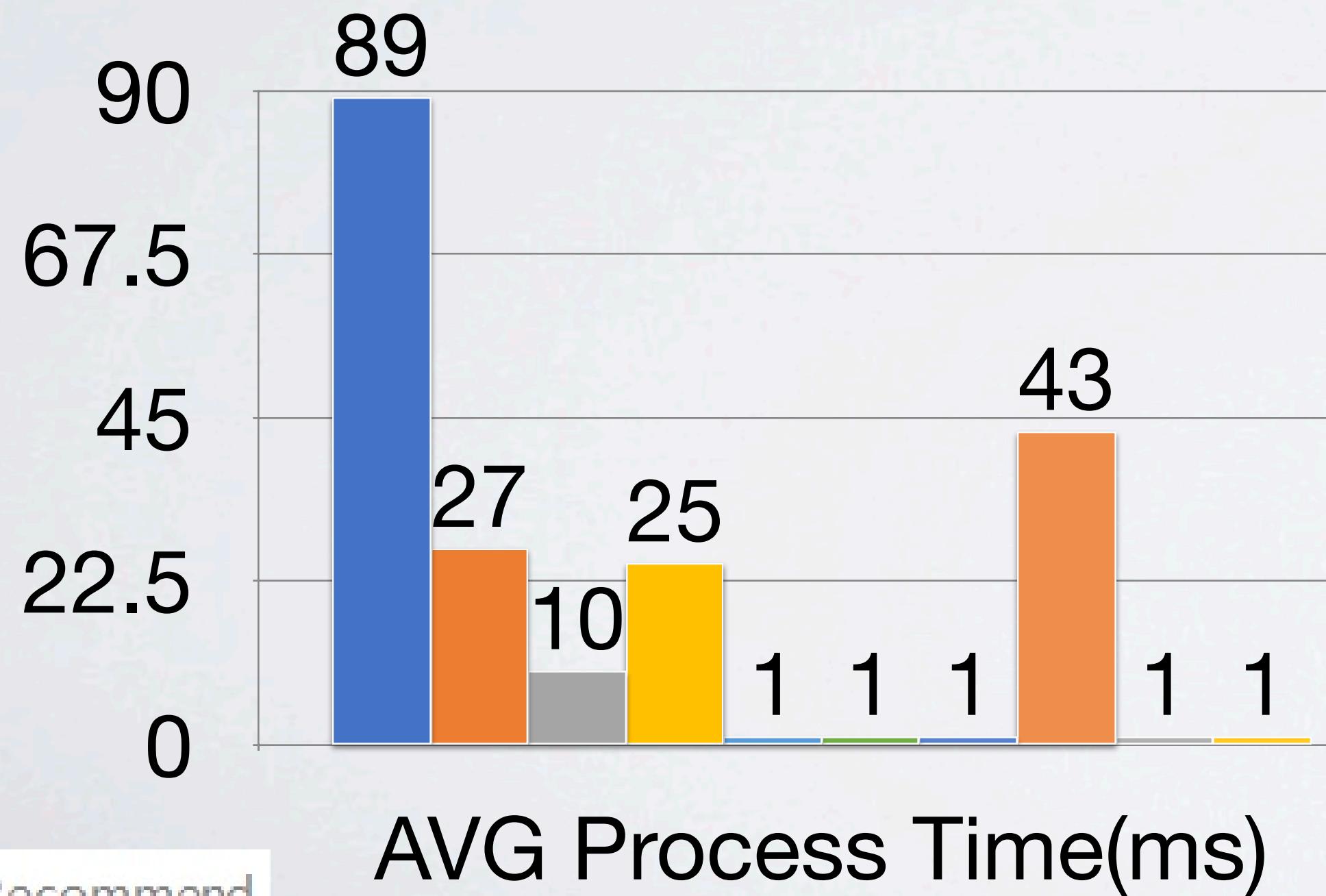
多挂载点改造

- MultipleDestinationMountTableResolver + HashFirstResolver
- Hive Staging + Yarn log

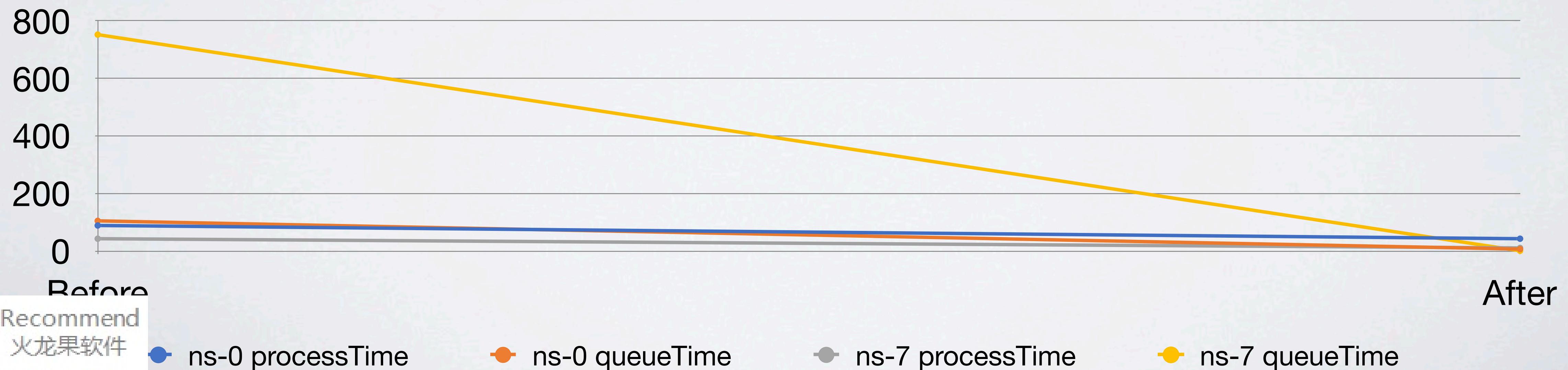
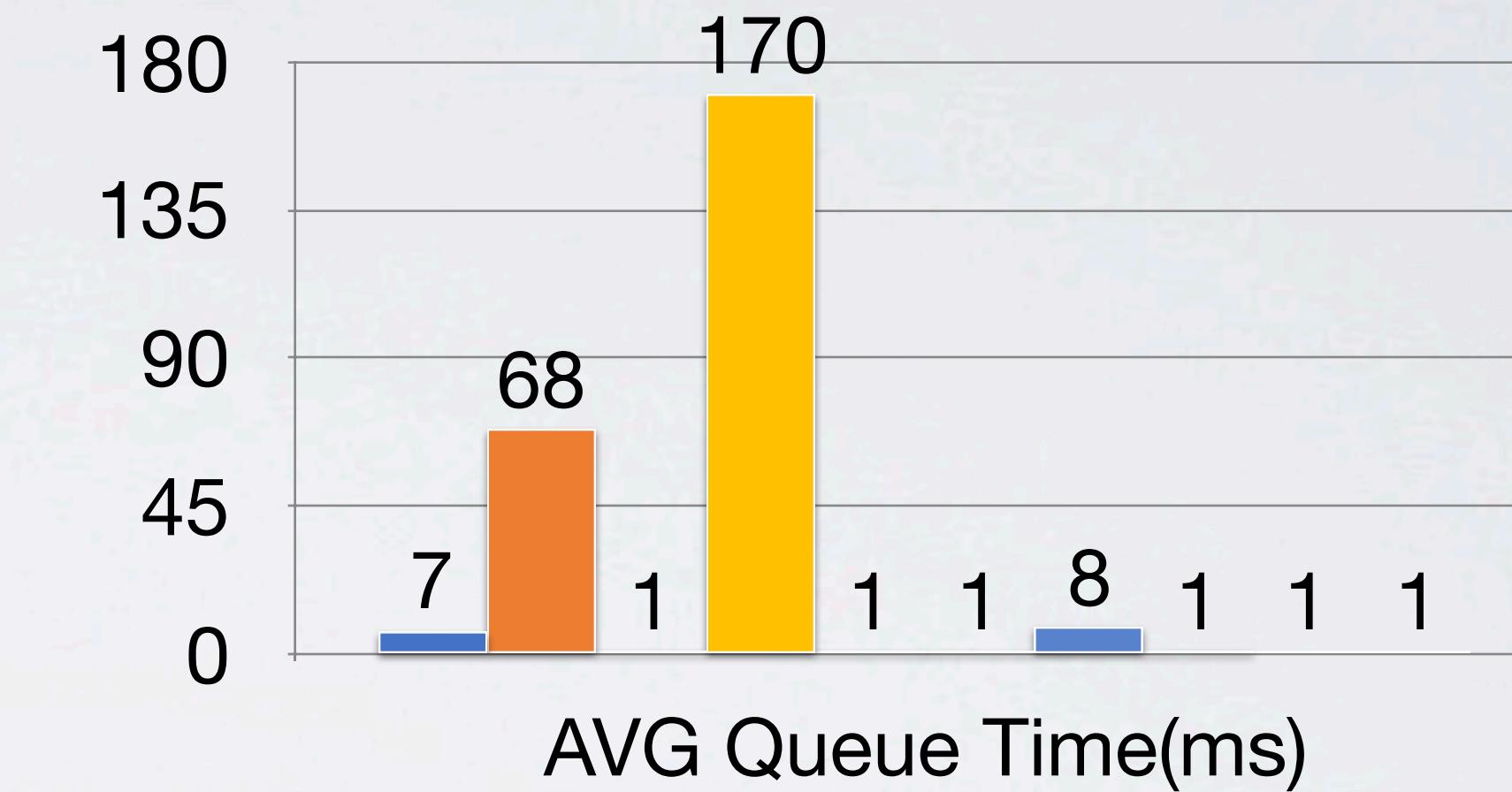
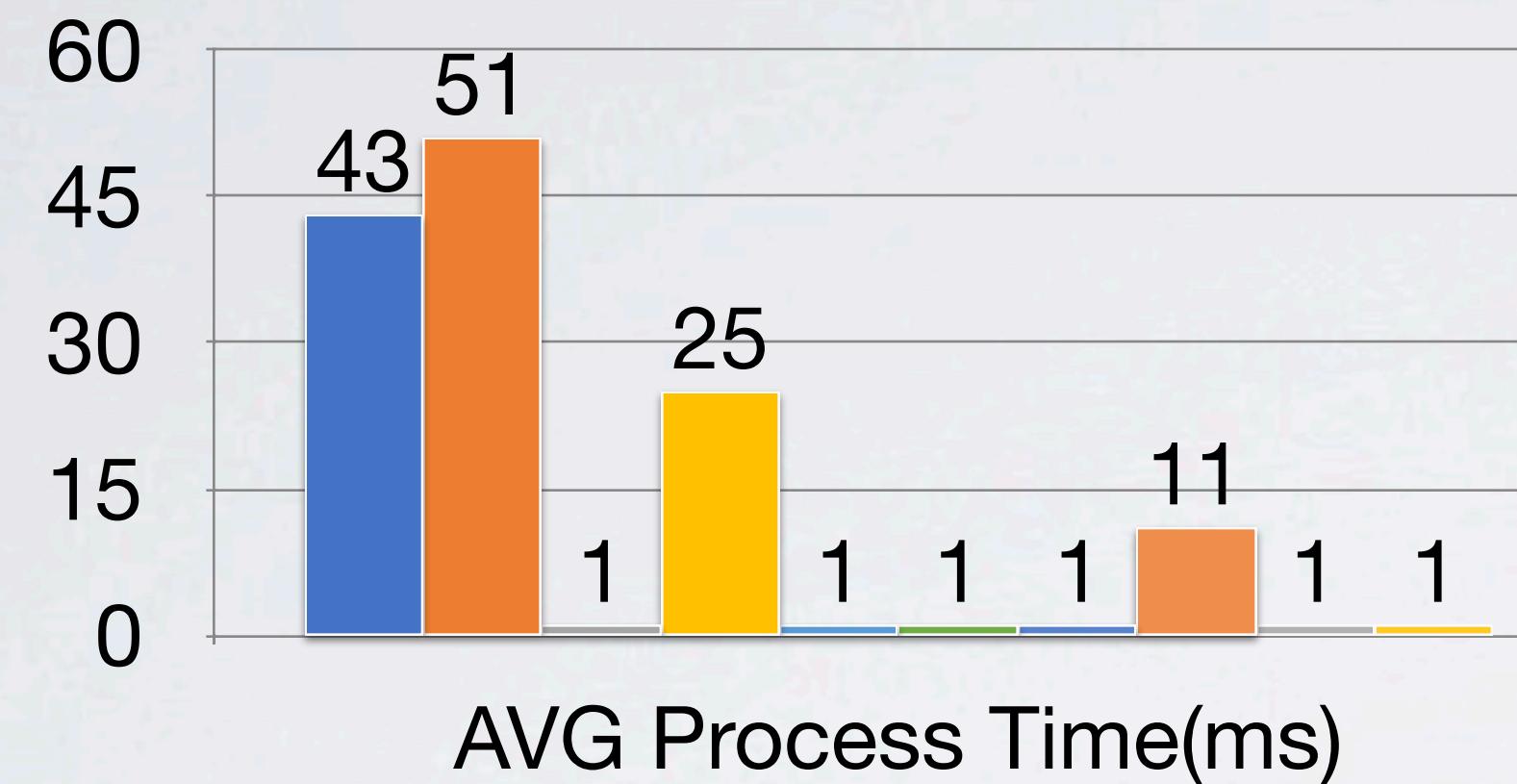


改造前状况

- 各namespace RPC处理延迟非常不均衡
- 主力namespace平均延迟194ms, 峰值接近1s, 峰值写qps接近4000

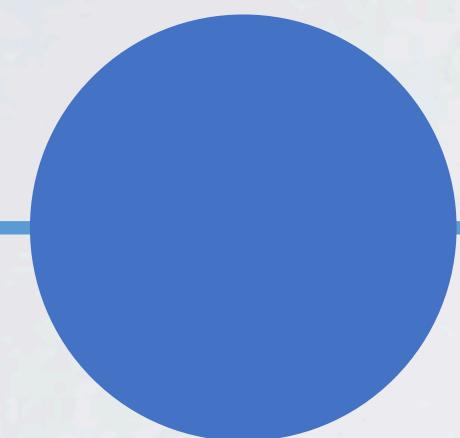


多挂载点改造效果



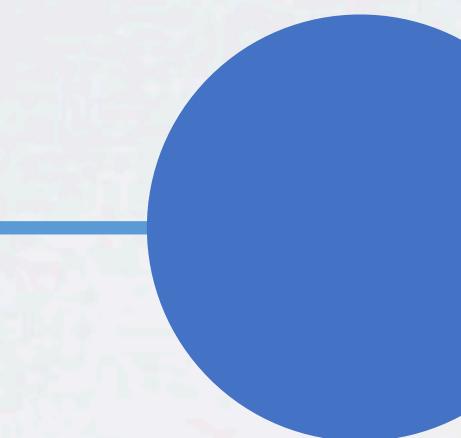
2017

单NameNode部署
到达瓶颈



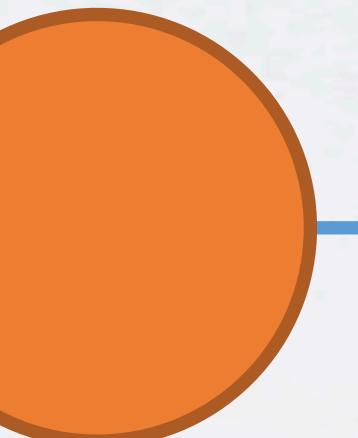
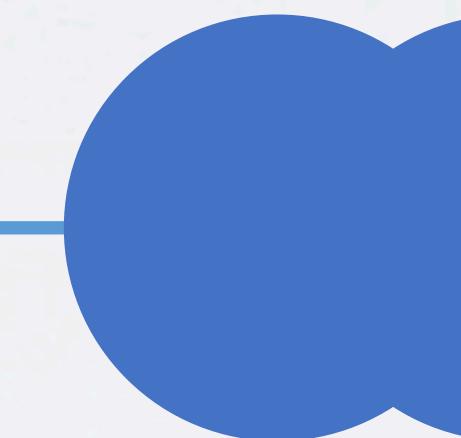
2018

国内主力集群完成
Federation改造



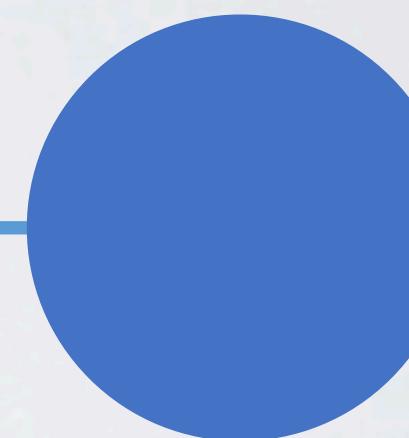
2019

NameNode负载
不均，内存不足



2020

RPC排队时间显著
变长，影响作业执
行



2019

Federation集群
改造RBF

2019

单机房扩展性受限

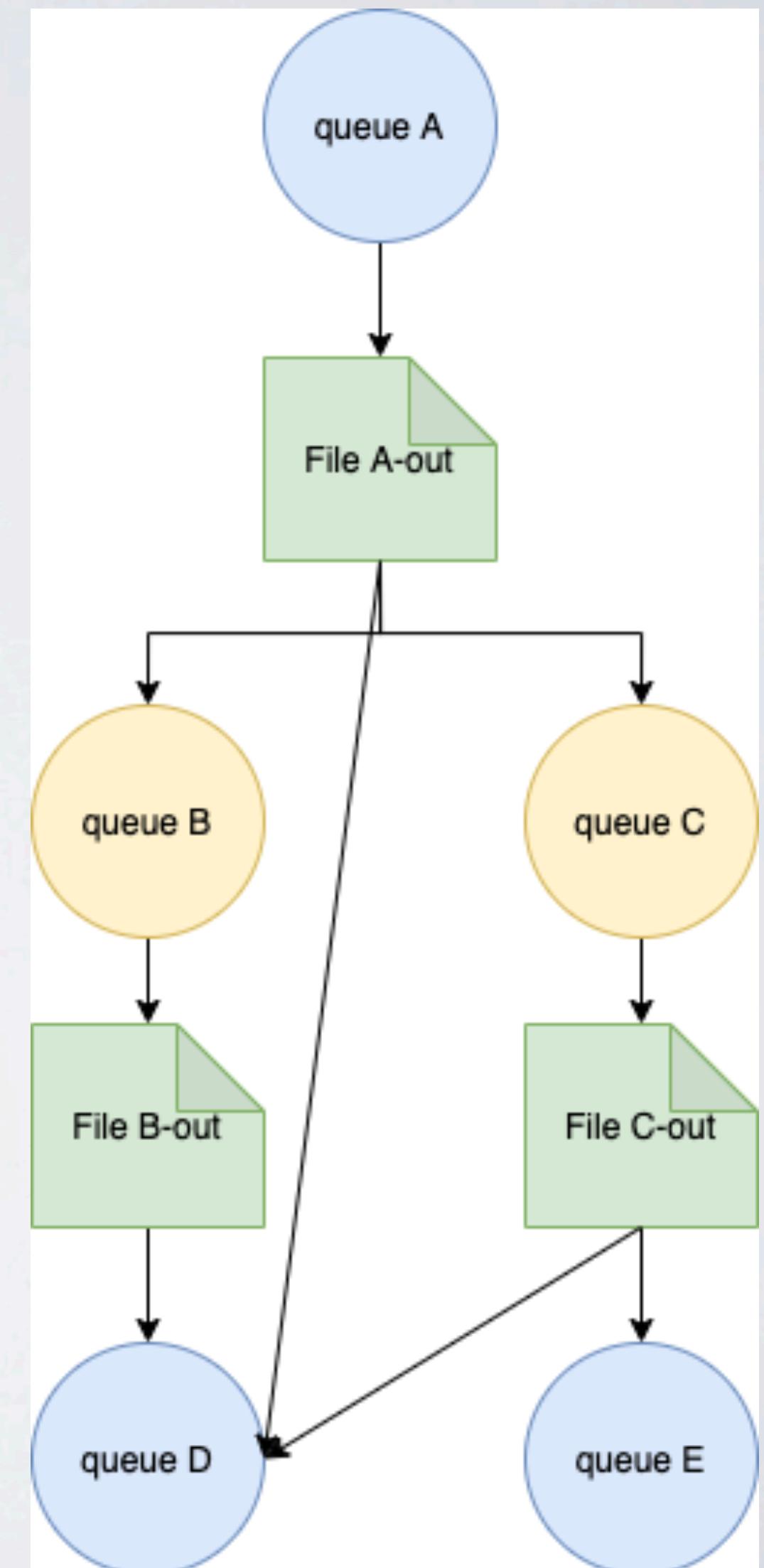
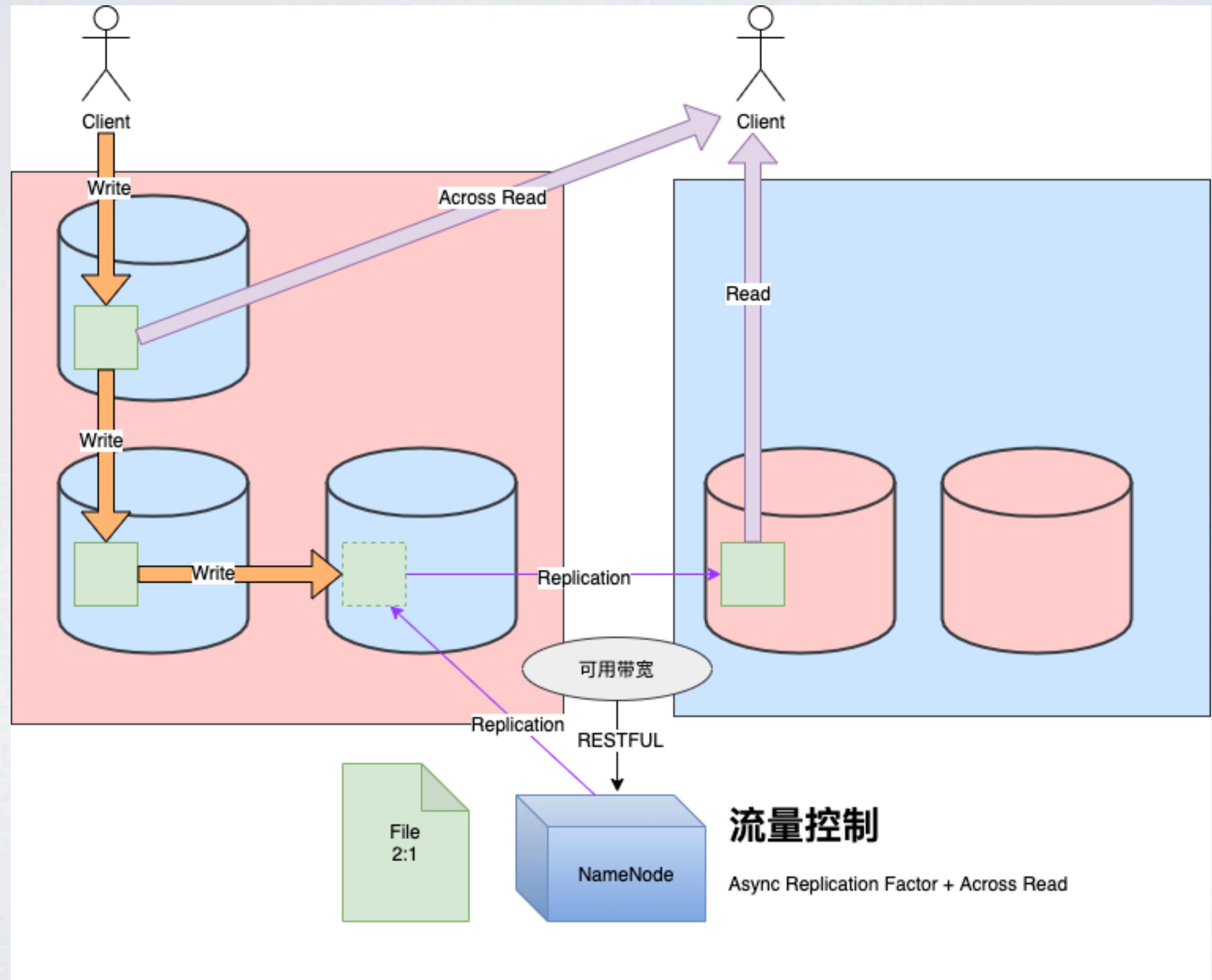
跨机房需求

- 单机房扩展性
- 海外成本
- 多云迁移
- 机房容灾



跨机房方案

- 流量控制



THANKS!