MultiNN: 基于HDFS的多NameNode分布式文件系统

何军权, 张 欢, 熊 劲
中国科学院 计算技术研究所

背景

Hadoop自推出以来已获得广泛应用，HDFS作为Hadoop的底层文件系统，也得到广泛关注和研究。HDFS 采用的单元数据服务器架构大大简化系统复杂性，但作为代价集群性能受到单一中心节点的制约，当面对海量数据元数据服务时会出现性能瓶颈。HDFS架构的显著问题在于：

单点故障：整个系统只有一个名字节点提供服务，当名字节点宕机，整个集群停止提供服务。

单节点性能瓶颈：系统性能会受制于单一的名字节点，当需求变大，系统会达到瓶颈。

采用去中心化元数据服务架构可以解决单点故障与单点性能瓶颈问题，但是也引入了一些新的问题。

关键问题

- 名字空间如何划分？
- 如何保证文件系统的局部性？
- 怎样保证元数据的一致性？

现有工作

系统架构

- 单个元数据服务器架构
- 元数据服务器集群架构
- P2P架构

一致性

- 共享存储池
- 两阶段或基于两阶段提交

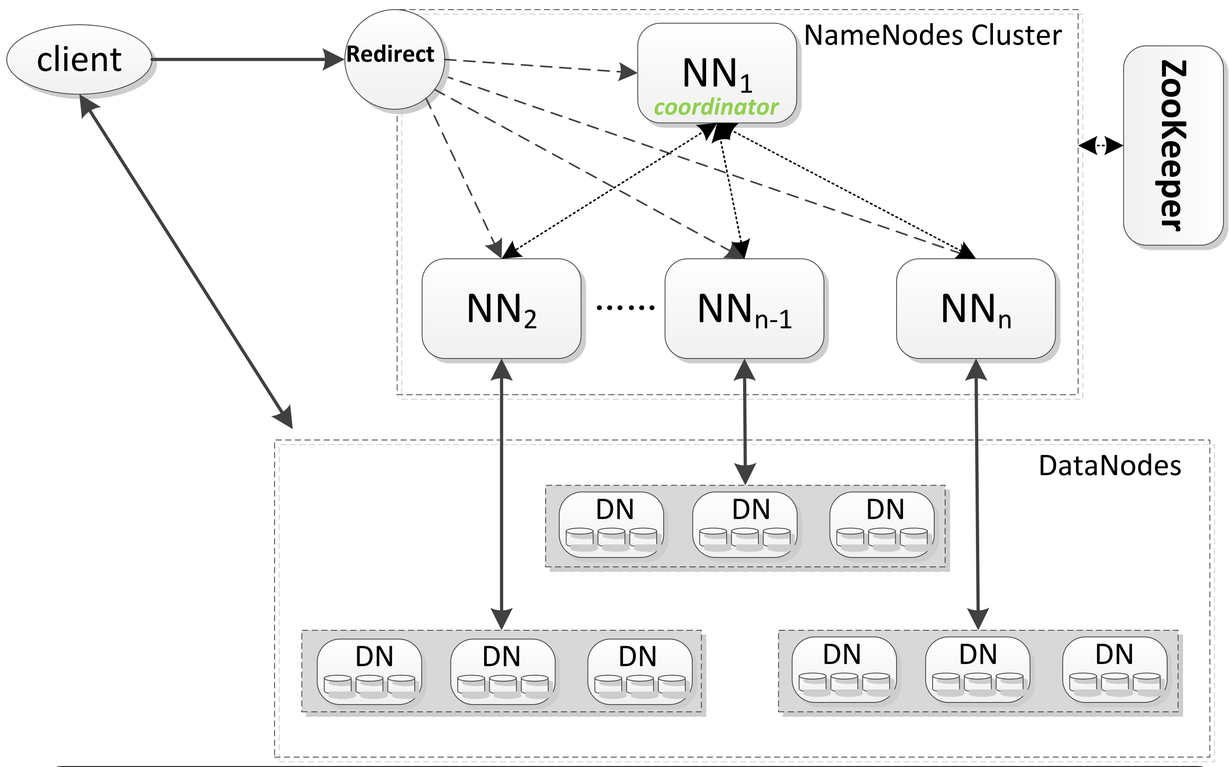
名字空间划分与局部性

- 静态/动态子树划分
- 基于哈希划分
- 基于表结构的划分： range partition
- 多副本机制

MultiNN设计

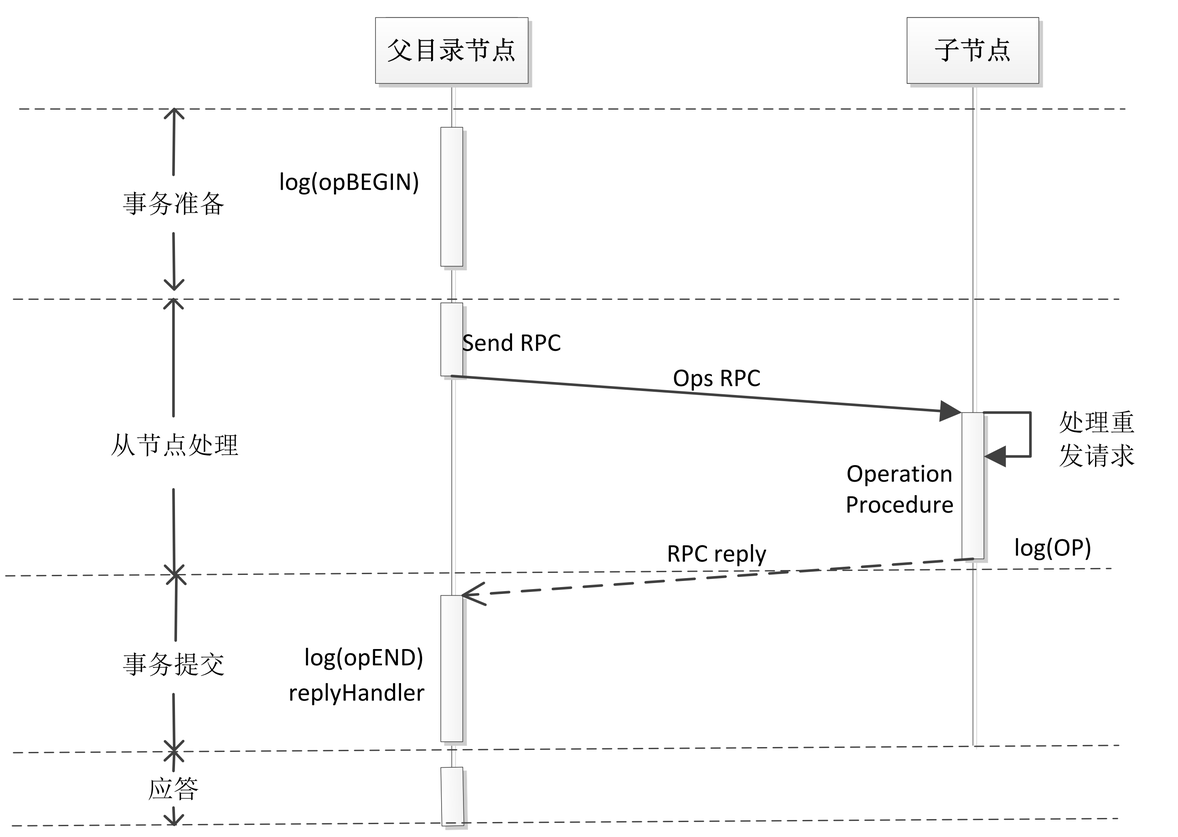
- 系统架构:
  - 采用名字节点集群提供元数据服务
  - 每个名字节点管理一组数据节点
- 名字空间划分
  - 划分方法：基于哈希划分
  - 划分粒度：基于目录粒度
- 事务处理
  - 1-NN操作(e.g. create)：日志机制
  - 2-NN 操作(mkdir)：两阶段操作
  - 多-NN操作：中心化管理机制

系统架构

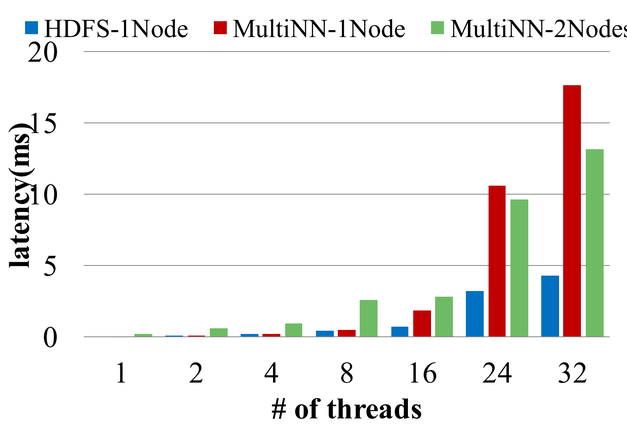


文件定位流程:
(1). 根据目录路径选择名字节点； (2). 定位目录桶(dirSlots)；
(3). 在dirList中查找对应目录索引节点； (4).在目录孩子列表中 获取文件inode.

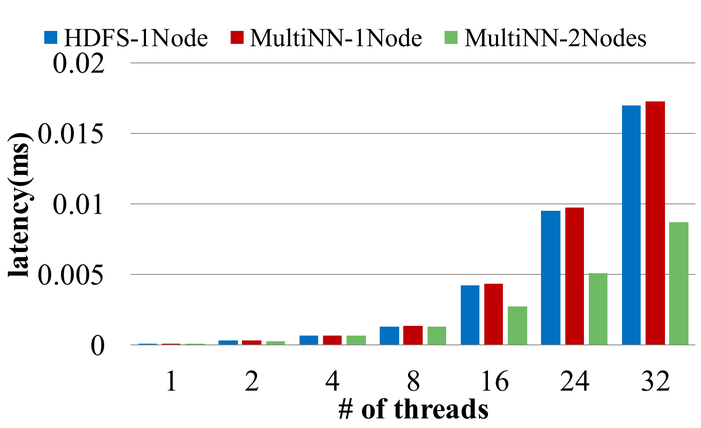
事务操作



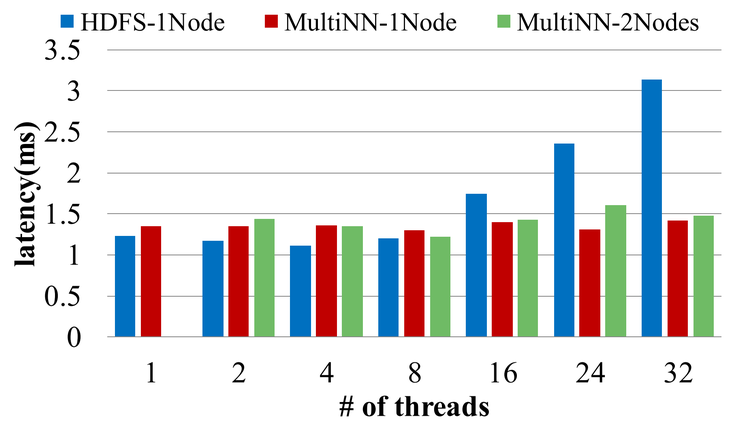
评测结果



Mkdir操作时延



Create操作时延



Stat操作时延

中国科学院计算技术研究所
Institute of Computing Technology, Chinese Academy of Sciences