Ceph中国社区 | IT大咖说 知识分享平台

The future of storage

Ceph-FS与大数据的恋爱史

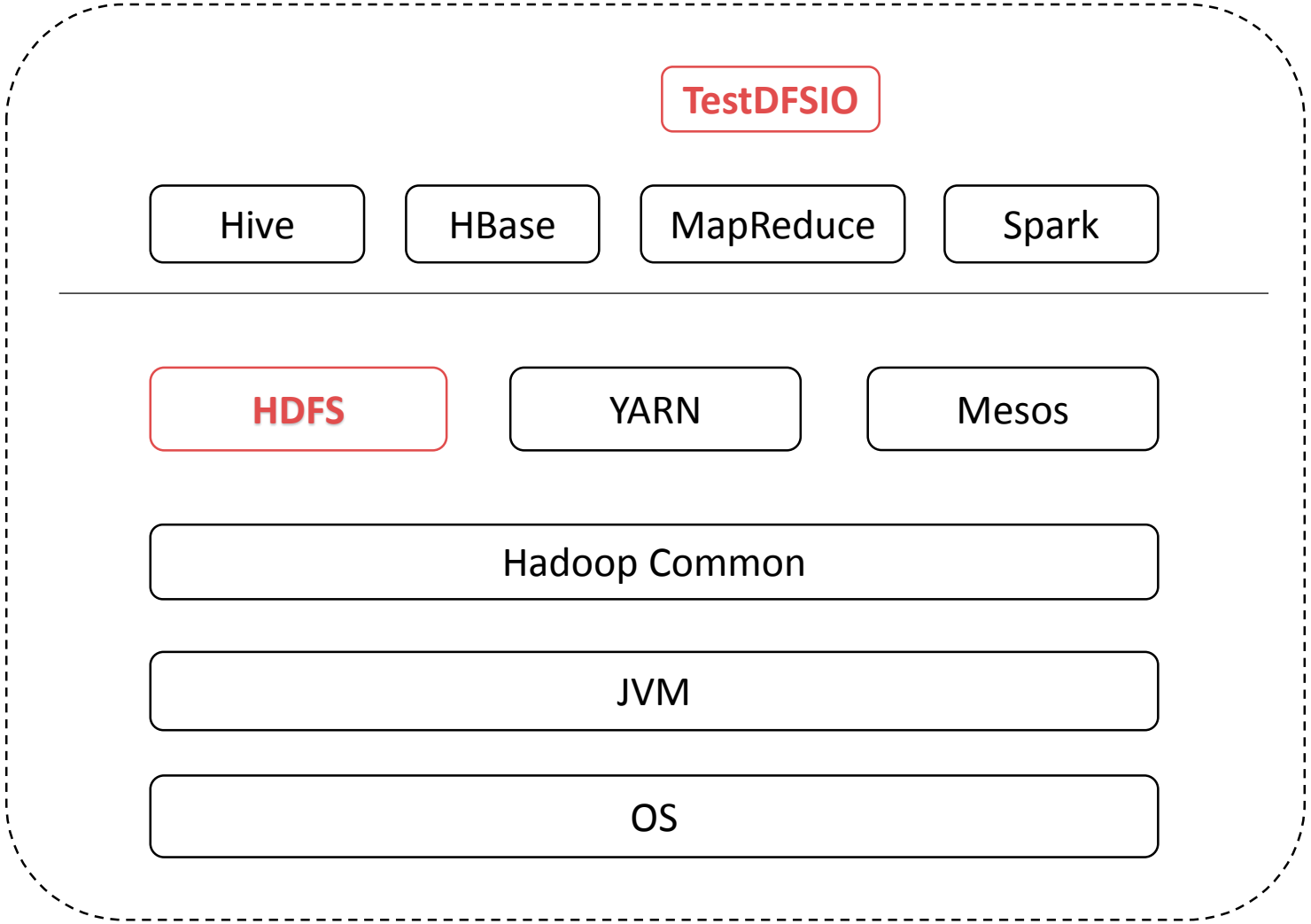
ZTE中兴

Ceph中国社区 | IT大咖说 知识分享平台

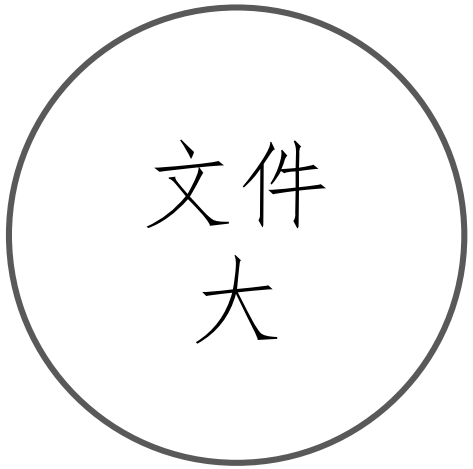
目·录

故事·背景

Hadoop

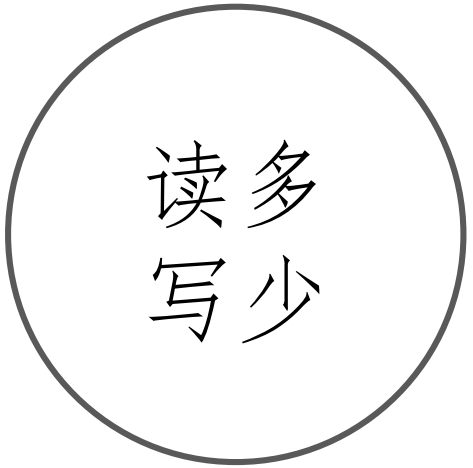

TestDFSIO
Hive
HBase
MapReduce
Spark
HDFS
YARN
Mesos
Hadoop Common
JVM
OS

典型大数据



GB-TB

千万级



一次写多次读

典型大数据 & HDFS

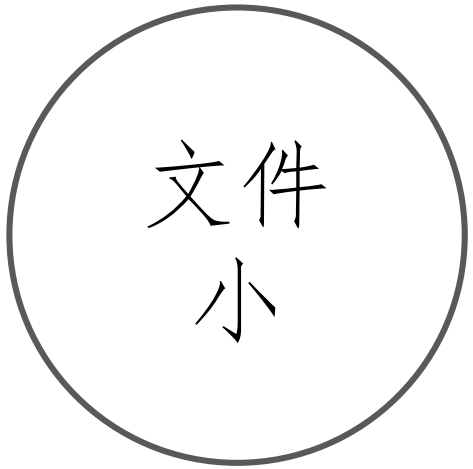
擅长处理大文件

支持千万量文件

读性能>写性能

天生一对

非典型大数据



KB-MB

亿级

非典型大数据 & HDFS

小文件性能差

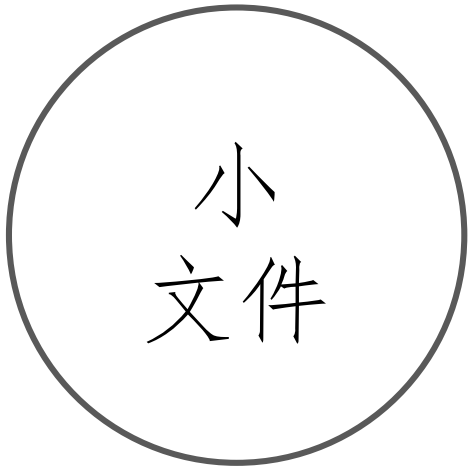
亿级文件搞不定

元数据内存消耗大

不合拍

HDFS Hold 不住，Ceph-FS 能 Hold 住吗？

Ceph-FS

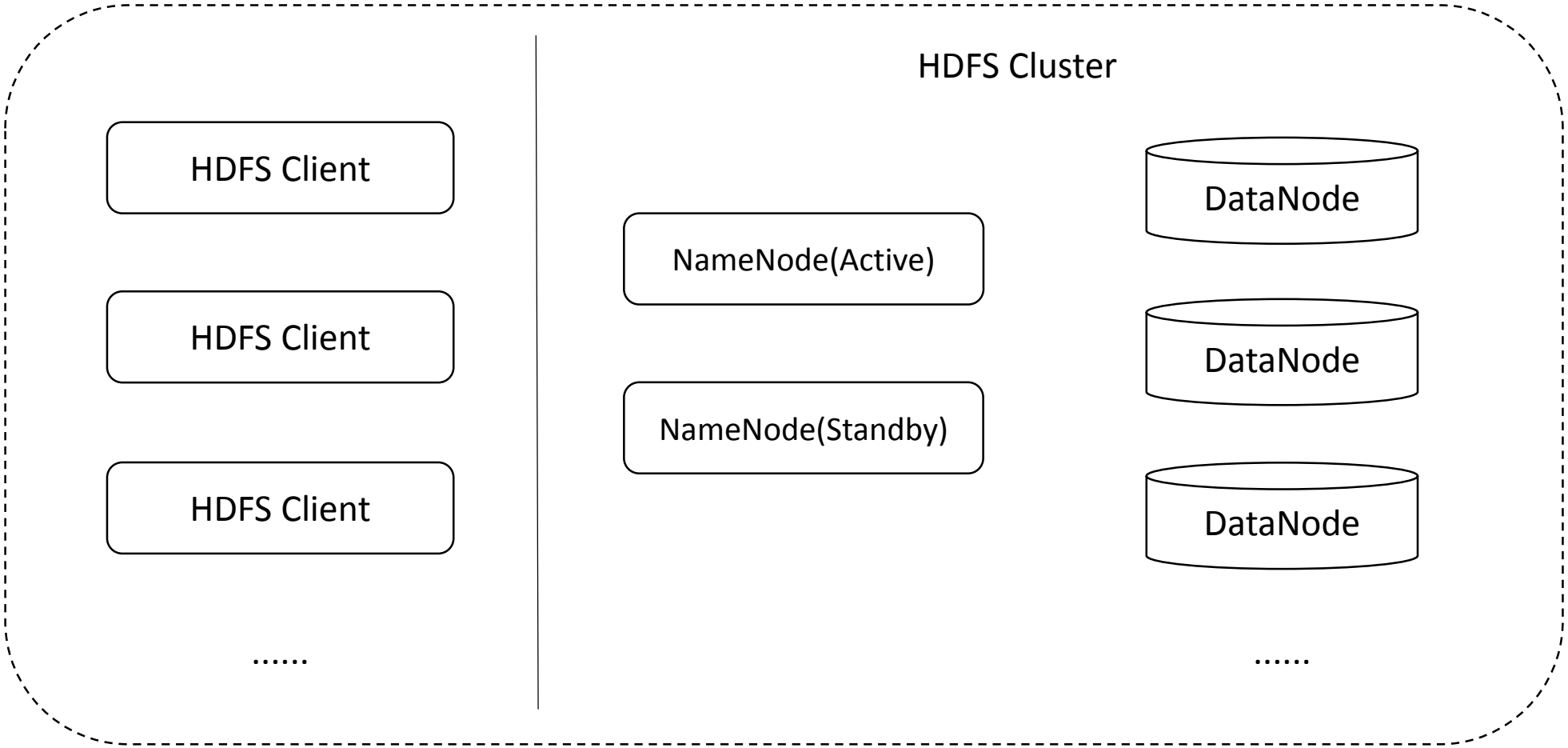


读写性能好

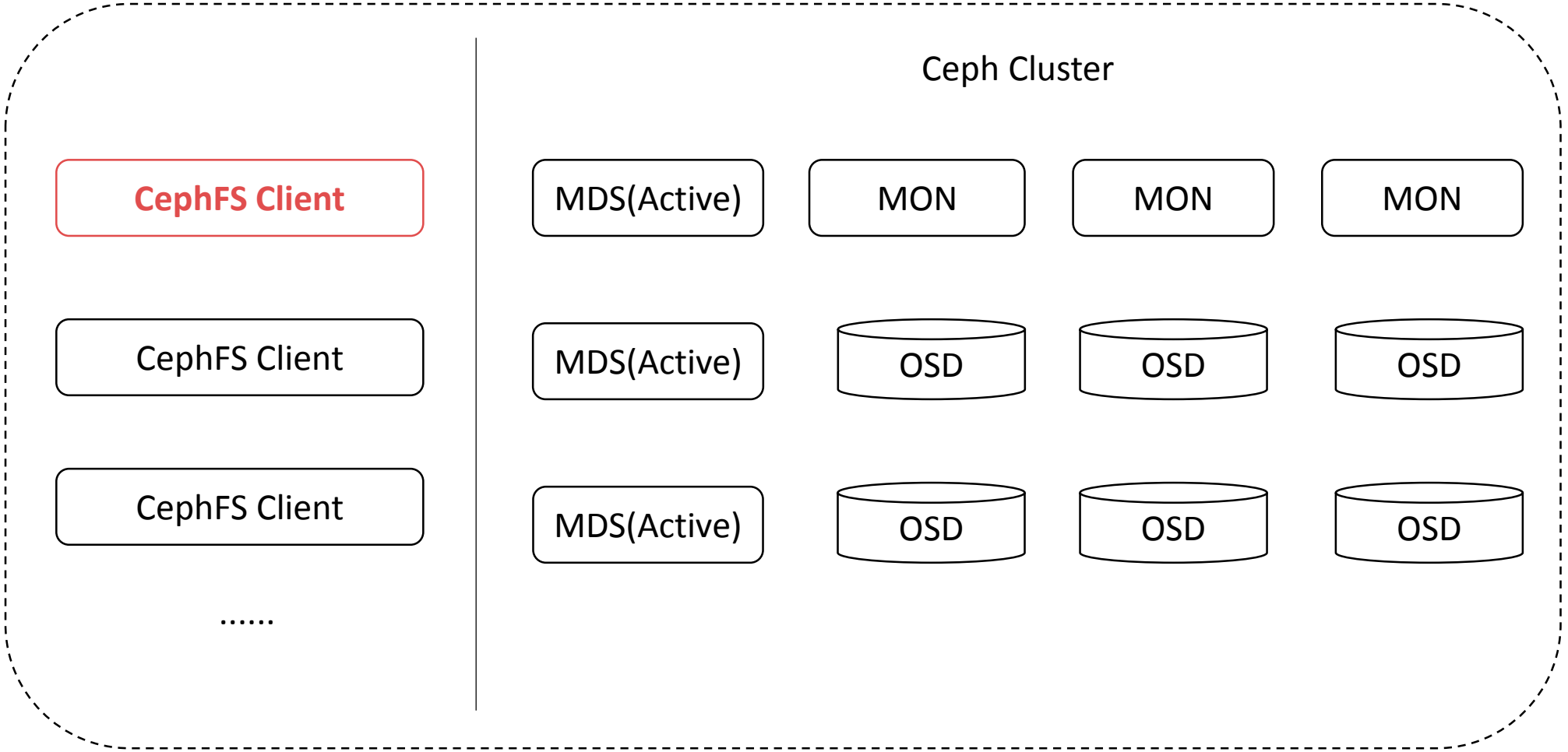
无限制

热点元数据

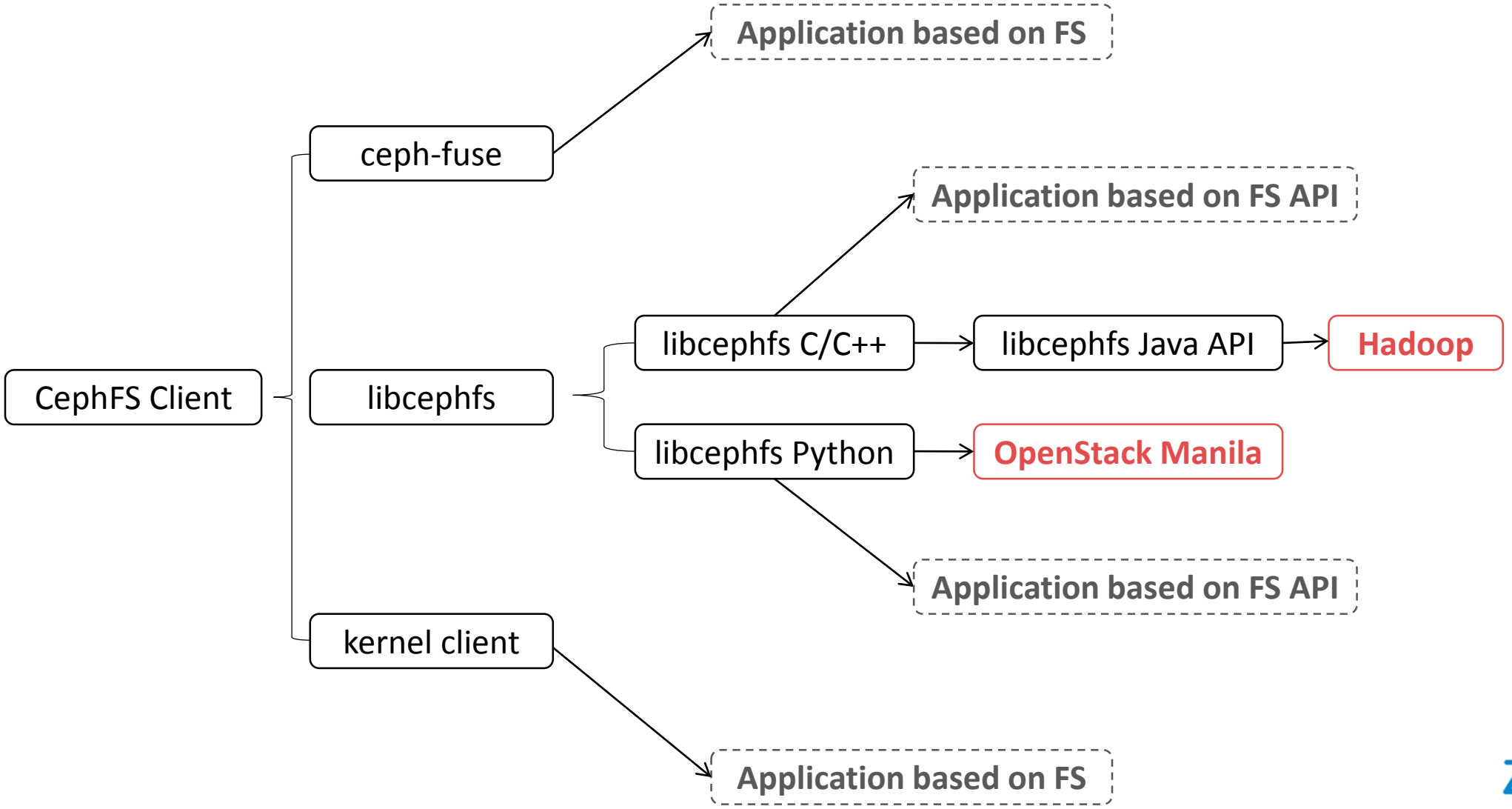
HDFS


HDFS Client
HDFS Client
HDFS Client
.....
HDFS Cluster
NameNode(Active)
NameNode(Standby)
DataNode
DataNode
DataNode
.....

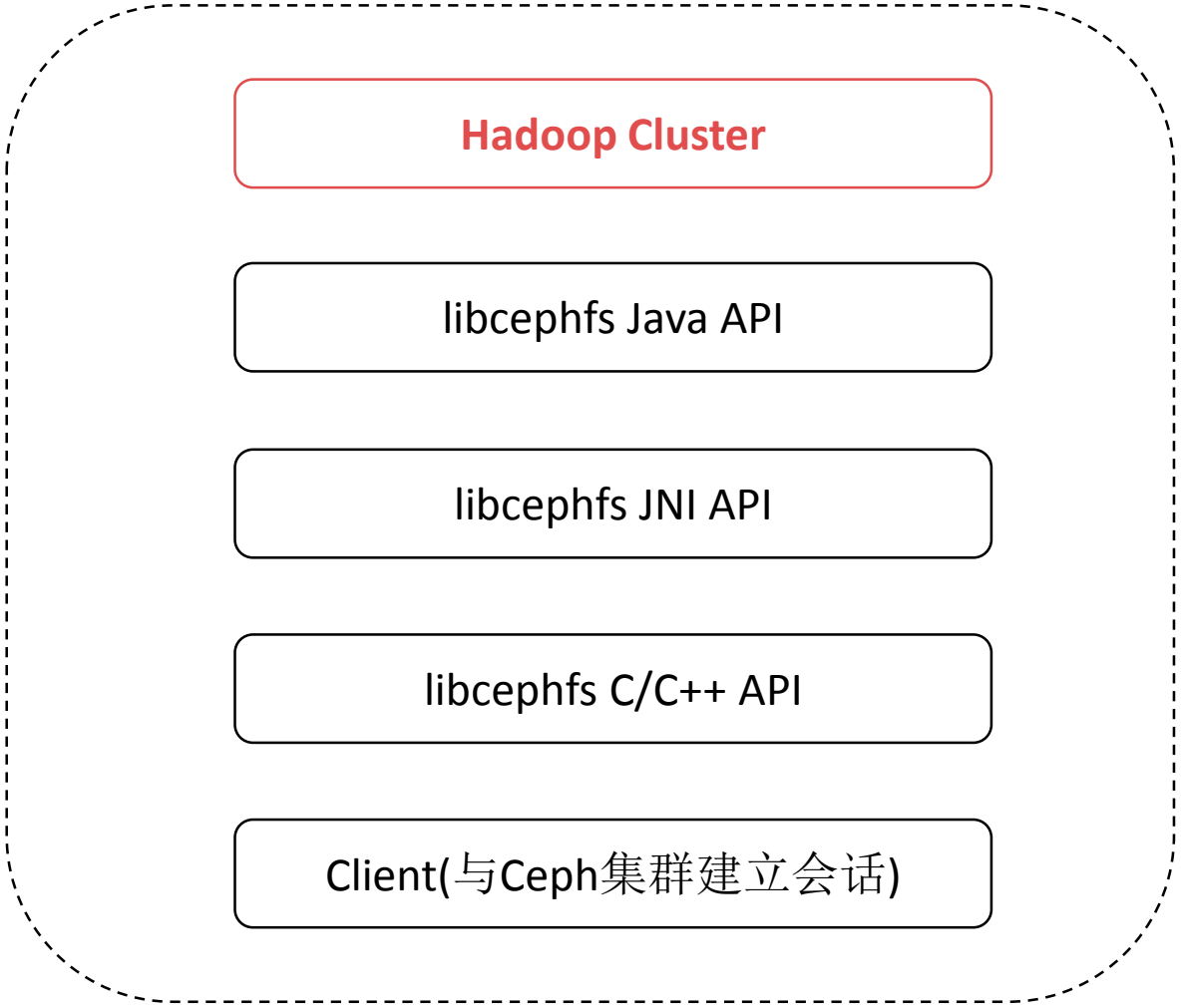
Ceph-FS


Ceph Cluster
CephFS Client
CephFS Client
CephFS Client
.....
MDS(Active)
MDS(Active)
MDS(Active)
MON
MON
MON
OSD
OSD
OSD
OSD
OSD
OSD

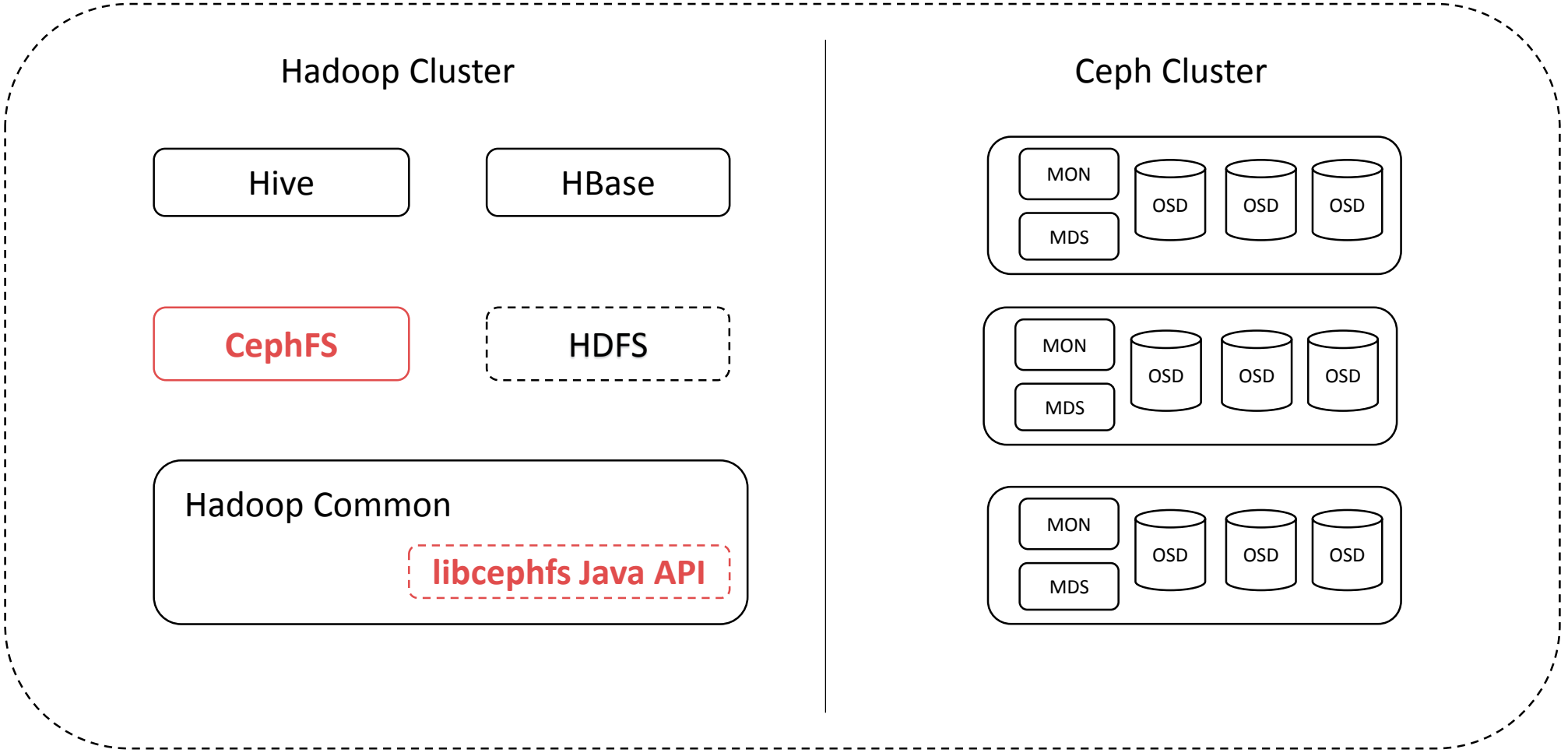
Ceph-FS


CephFS Client
ceph-fuse
libcephfs
kernel client
Application based on FS
Application based on FS API
libcephfs C/C++
libcephfs Java API
Hadoop
libcephfs Python
OpenStack Manila
Application based on FS API
Application based on FS

Ceph-FS

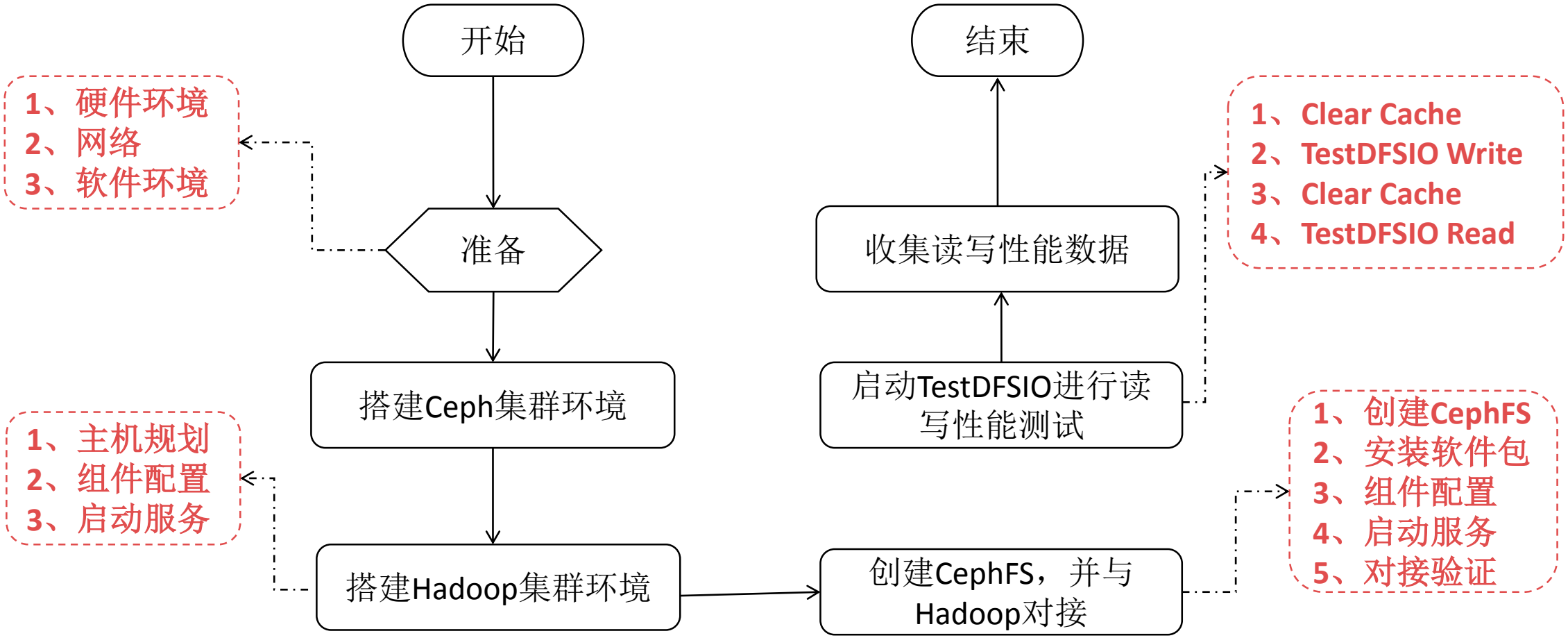

Hadoop Cluster
libcephfs Java API
libcephfs JNI API
libcephfs C/C++ API
Client(与Ceph集群建立会话)

Hadoop & Ceph-FS



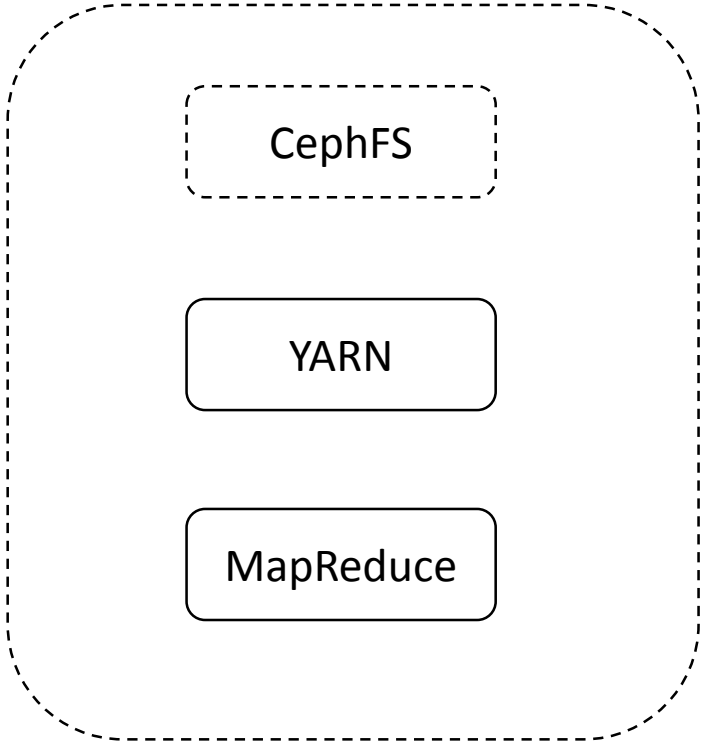
第一次·亲密接触

验证流程

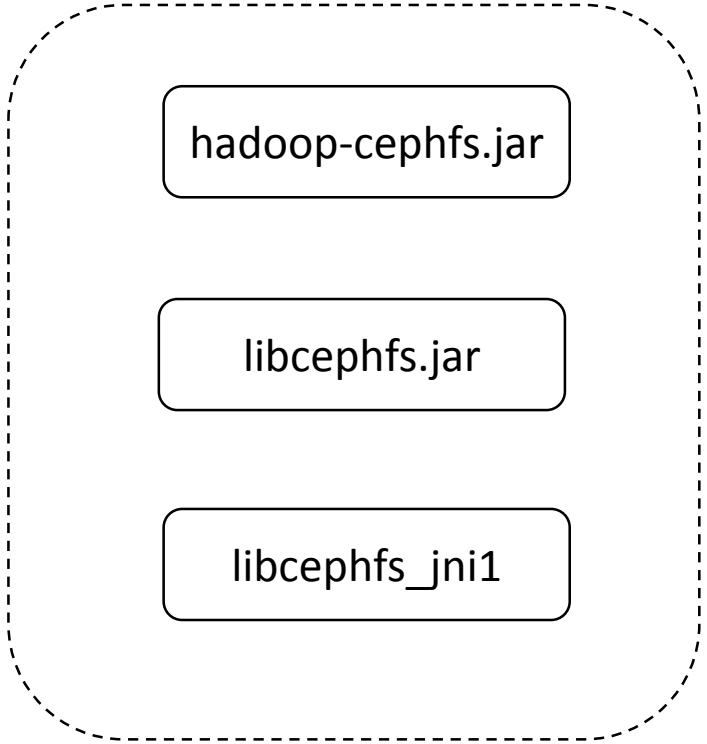

开始
准备
搭建Ceph集群环境
搭建Hadoop集群环境
创建CephFS，并与Hadoop对接
启动TestDFSIO进行读写性能测试
收集读写性能数据
结束
1、硬件环境
2、网络
3、软件环境
1、主机规划
2、组件配置
3、启动服务
1、创建CephFS
2、安装软件包
3、组件配置
4、启动服务
5、对接验证
1、Clear Cache
2、TestDFSIO Write
3、Clear Cache
4、TestDFSIO Read

验证准备

Hadoop集群组件



CephFS组件



验证障碍

遇到的问题：

- FileSystem implementation error - default port -1 is not valid
  hadoop-libcephfs.jar中未定义函数getDefaultPort，此函数用于获取默认端口。
  如果未定义，则使用0作为默认端口，在执行过程中进行合法性检查时，则抛出此异常。

- 写文件到Hadoop时，报访问被拒绝错误
  检查ceph权限配置：ceph auth list 查看client.hadoop是否有对osd、mon、mds执行相应操作的权限，在/etc/ceph/ceph.conf中设置：
  [client]
  client_mount_uid = $(id hadoop)
  client_mount_gid = $(id hadoop)

- 有些节点可以读写cephfs的文件，有些节点报Operation not permitted错误
  不同节点上hadoop用户的uid和gid不同引起，修改hadoop用户的uid和gid使其一致
  id hadoop
  usermod -u $uid hadoop
  groupmod -g $gid hadoop

功能验证

Hadoop FS Shell	CephFS	HDFS	Comments
cat	supported	supported	
chgrp	no effect	supported	*may* support
chmod	supported	supported	
chown	no effect	supported	*may* support
copyFromLocal	supported	supported	
copyToLocal	supported	supported	
count	supported	supported	
cp	supported	supported	
createSnapshot	unsupported	supported	*may* support
deleteSnapshot	unsupported	supported	*may* support
df	unsupported	supported	default values are normally displayed

功能验证

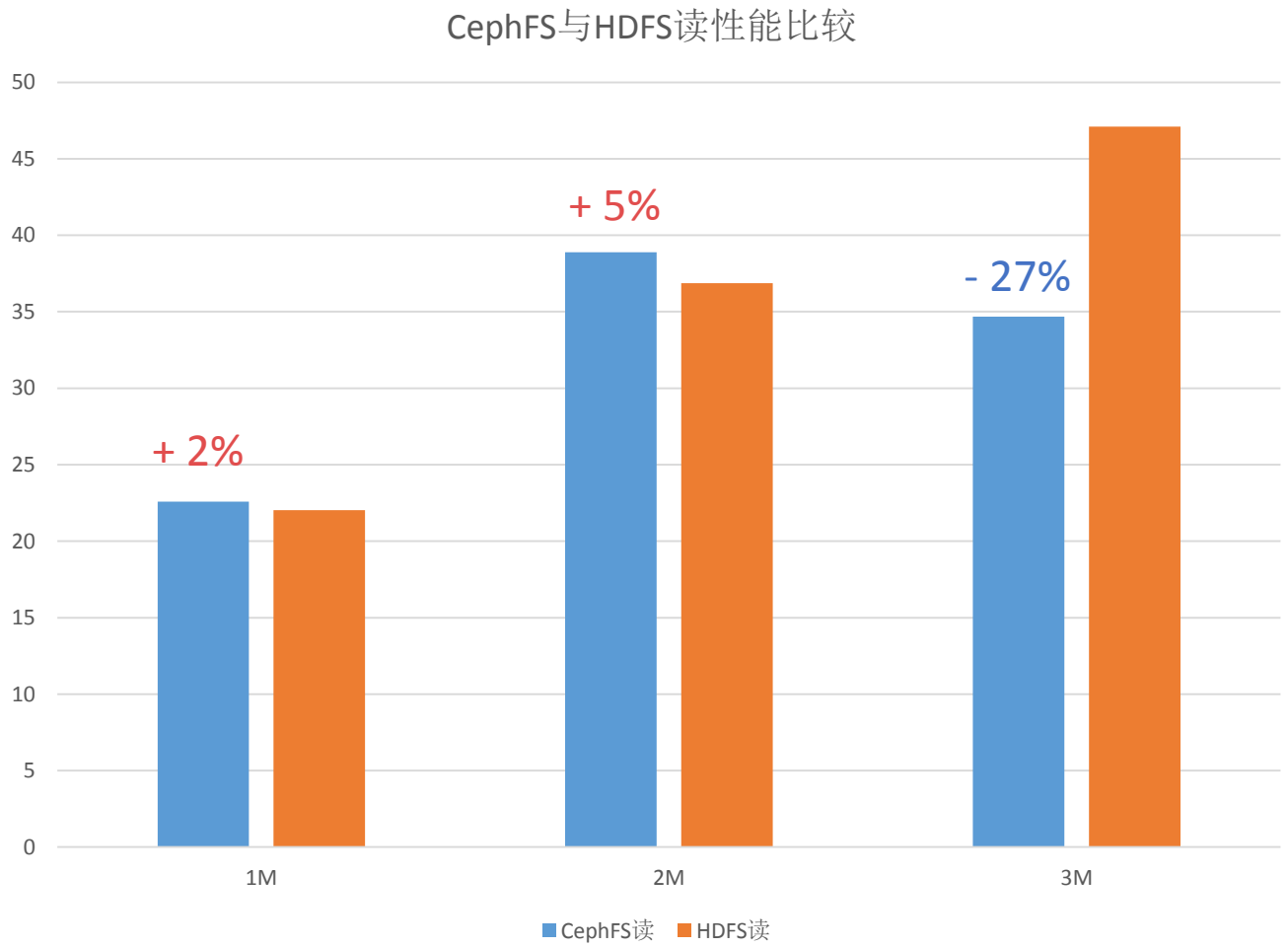
Hadoop FS Shell	CephFS	HDFS	Comments
find	supported	supported	
get	supported	supported	
getfacl	supported	supported	
getfattr	unsupported	supported	*may* support
ls	supported	supported	
mkdir	supported	supported	
moveFromLocal	supported	supported	
moveToLocal	unsupported	unsupported	
mv	supported	supported	
put	supported	supported	
renameSnapshot	unsupported	supported	*may* support

功能验证

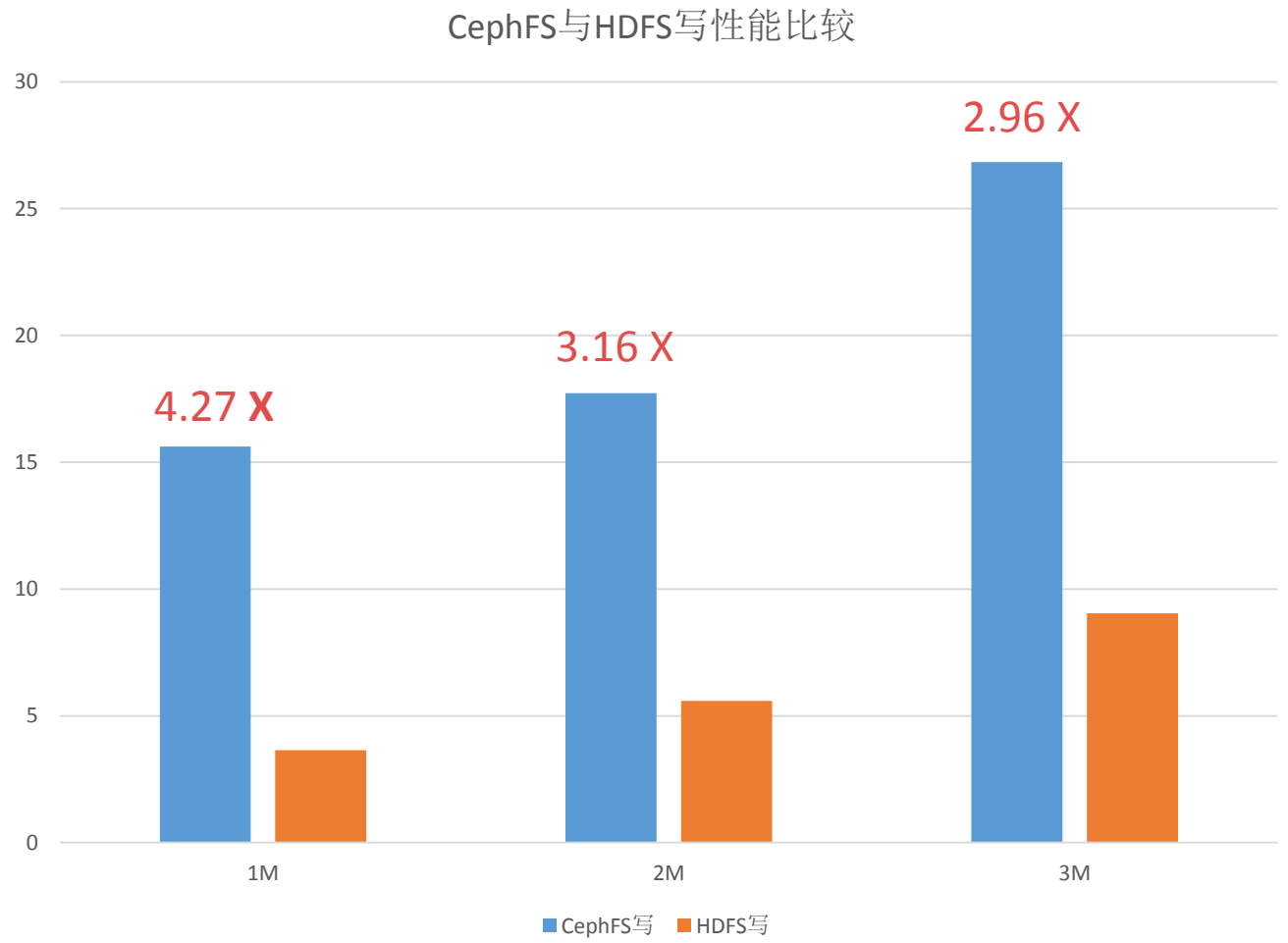
Hadoop FS Shell	CephFS	HDFS	Comments
rm	supported	supported	CephFileSystem.delete Method
rmdir	supported	supported	
setfacl	unsupported	supported	*may* support
setfattr	unsupported	supported	*may* support
stat	supported	supported	
tail	supported	supported	

http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html
https://github.com/ceph/cephfs-hadoop

性能验证

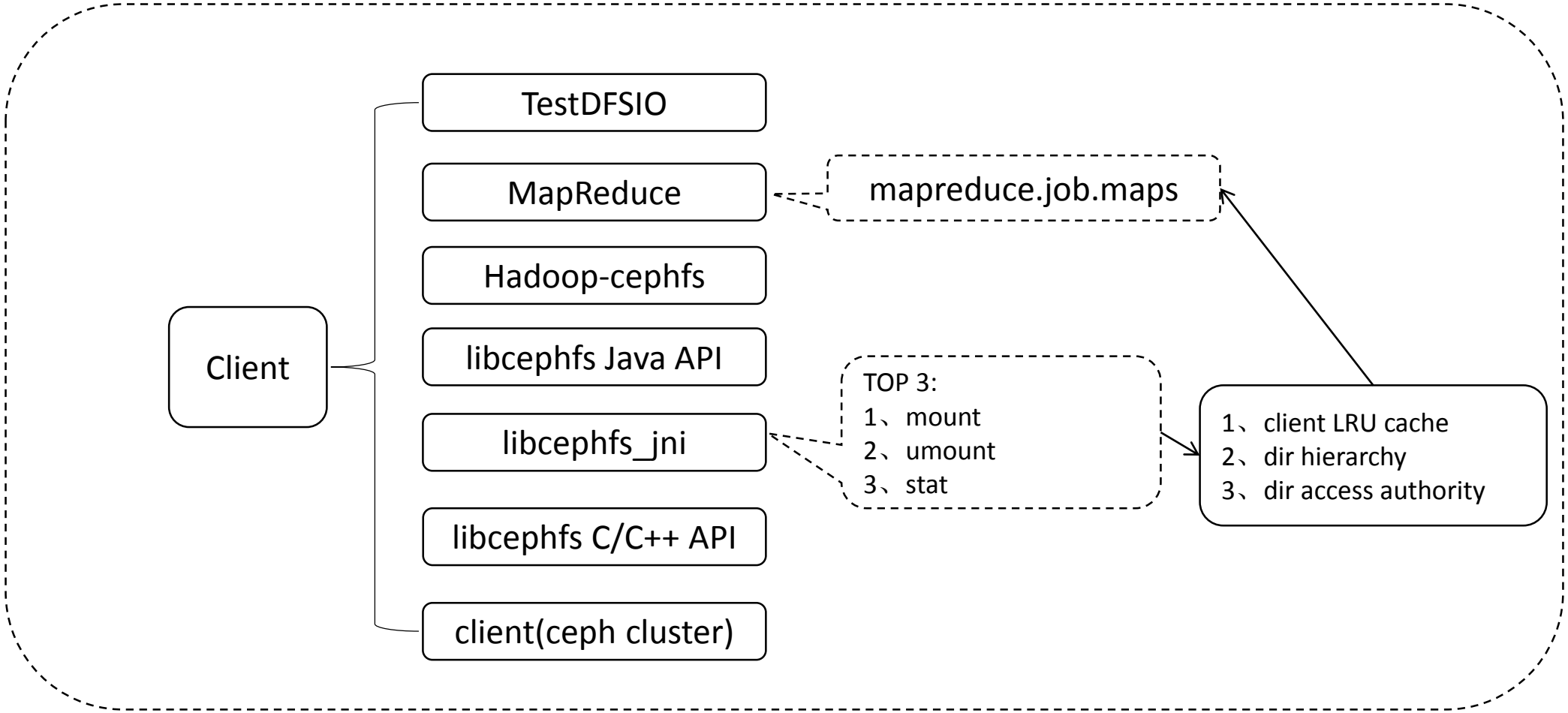

CephFS与HDFS读性能比较
50
45
40
35
30
25
20
15
10
5
0
+ 2%
+ 5%
- 27%
1M
2M
3M
CephFS读
HDFS读

性能验证

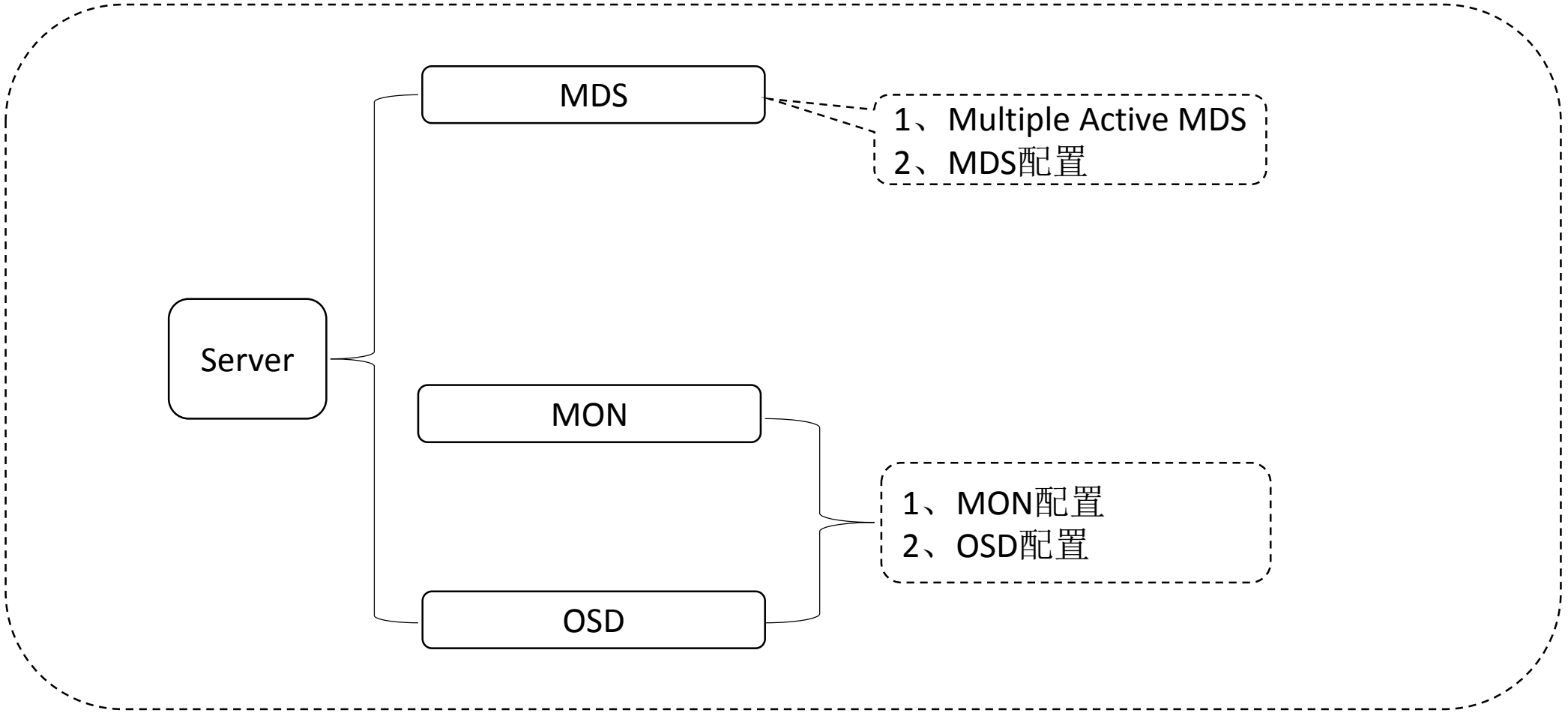

CephFS与HDFS写性能比较
30
25
20
15
10
5
0
4.27 X
3.16 X
2.96 X
1M
2M
3M
CephFS写
HDFS写

热·恋期

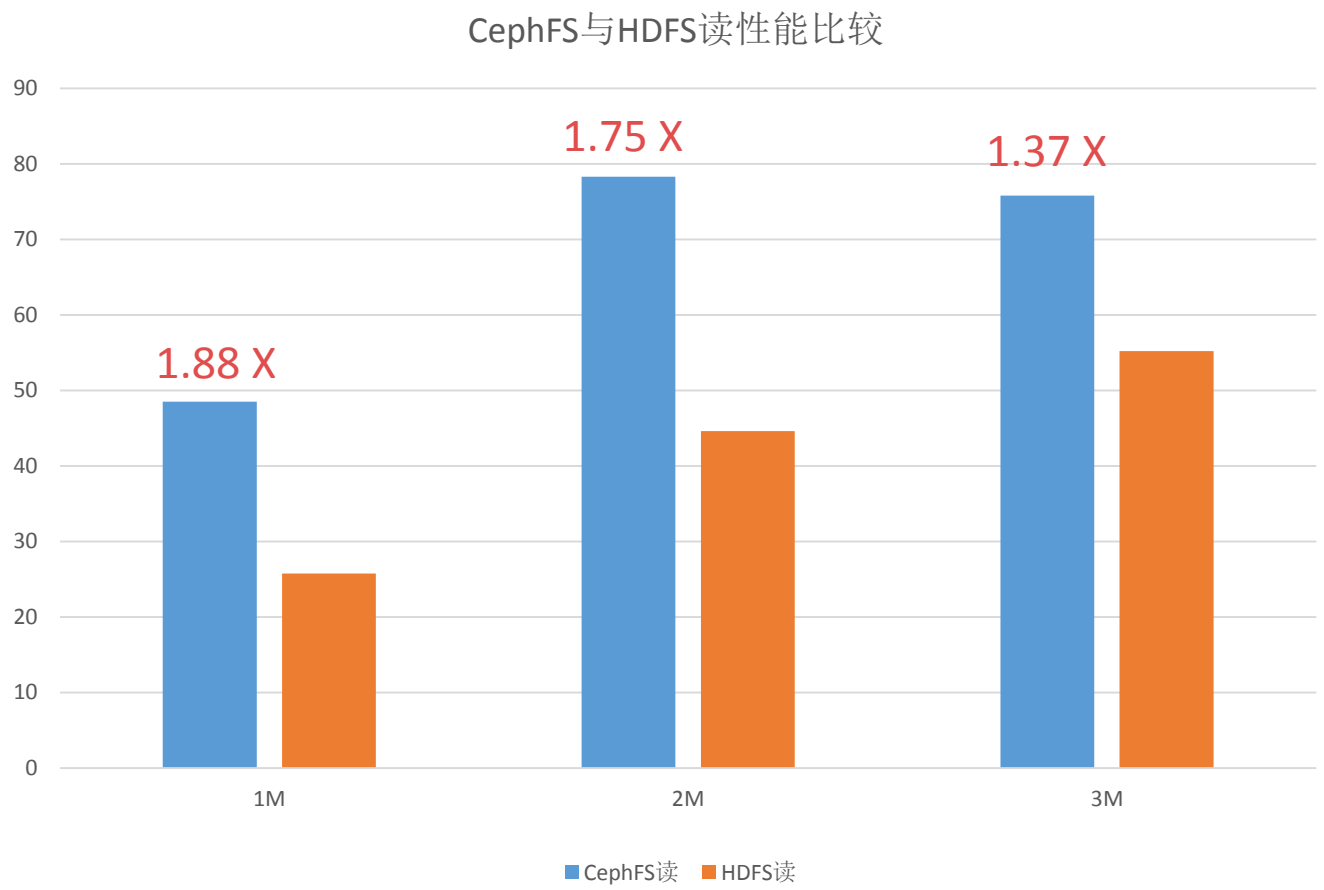
性能优化——瓶颈分析(Client)



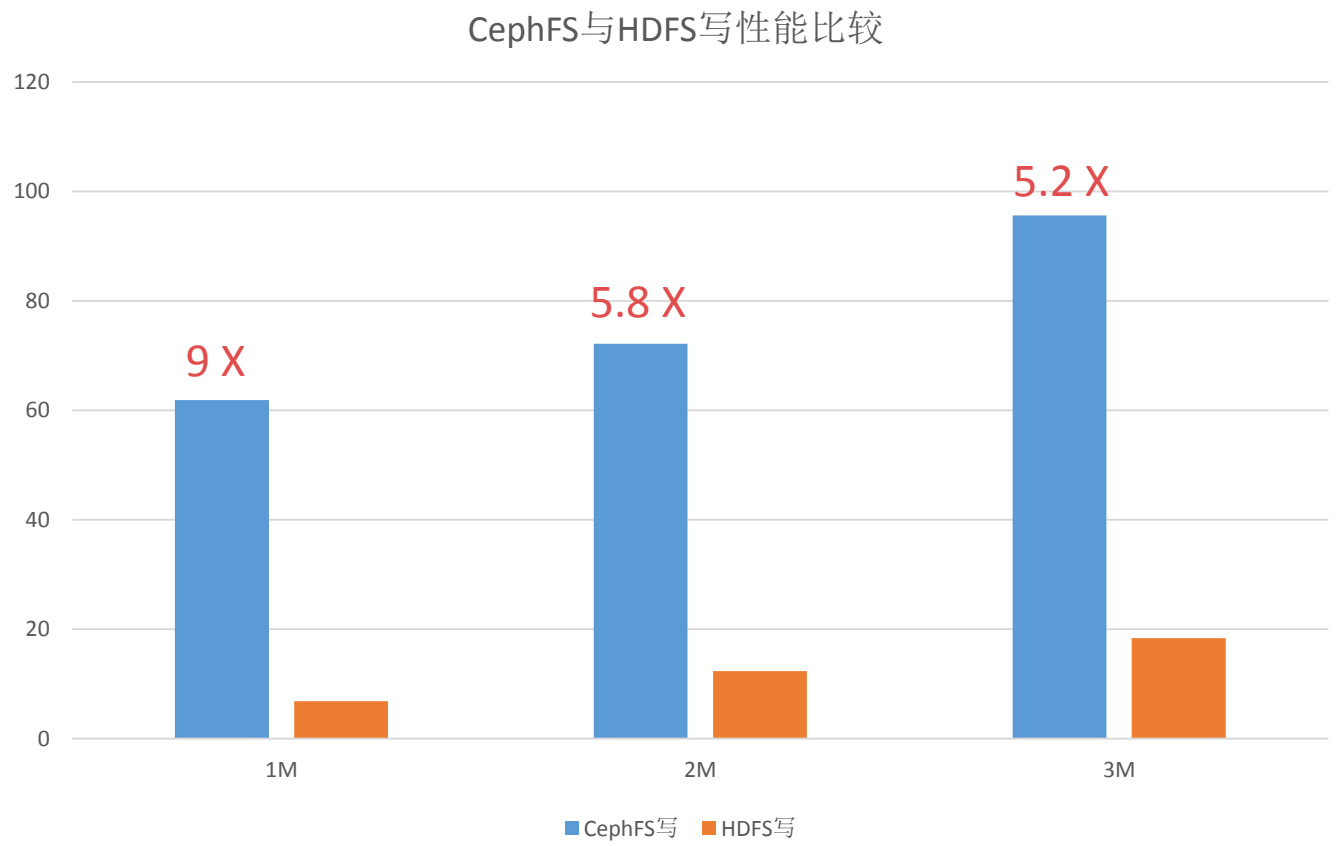
性能优化——瓶颈分析(Server)



优化后性能验证

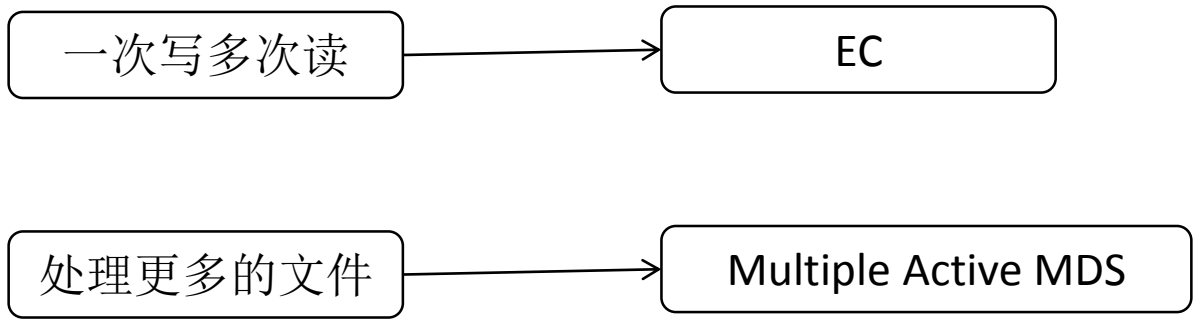

CephFS与HDFS读性能比较
90
80
70
60
50
40
30
20
10
0
1.88 X
1.75 X
1.37 X
1M
2M
3M
CephFS读
HDFS读

优化后性能验证


CephFS与HDFS写性能比较
120
100
80
60
40
20
0
9 X
5.8 X
5.2 X
1M
2M
3M
CephFS写
HDFS写

未完·待续

后续计划



ZTE中兴