HDFS的透明压缩存储

刘景龙
邮箱：baggioss@gmail.com
twitter：baggioss

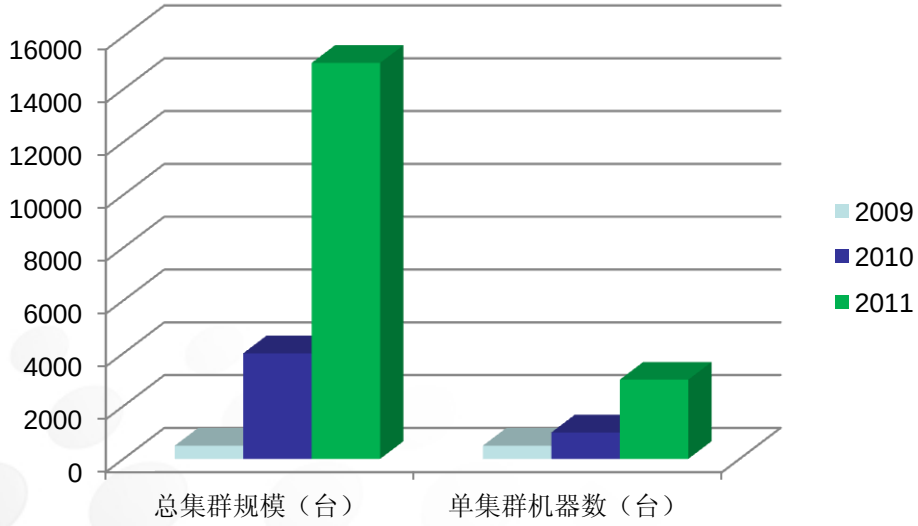
主要内容

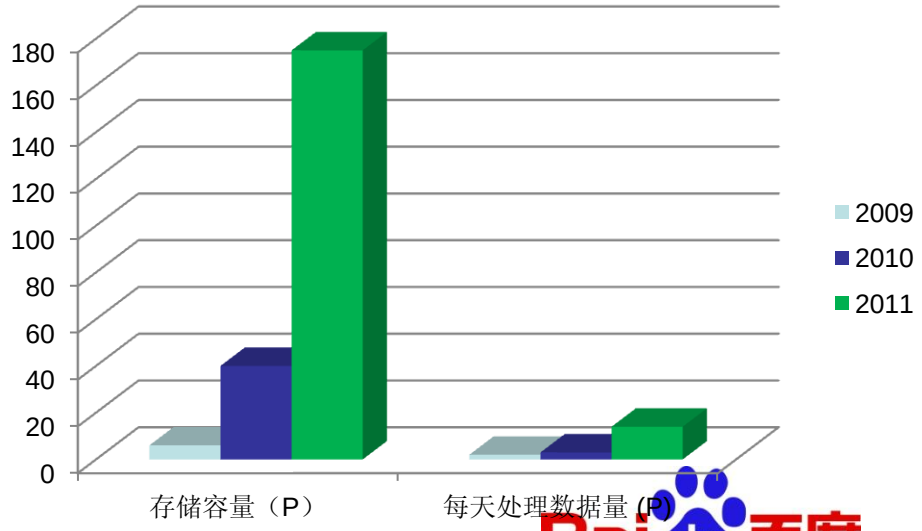
- Hadoop @baidu
  - 过去一年的工作
  - 进行中的项目
- 透明压缩
  - 目标
  - 实现
  - 如何解决核心问题
  - 如何规避风险
  - 未来计划

Baidu 百度

Hadoop@baidu

- 16000+ 机器，10个集群
- 最大集群机器数3000台
- 存储 127.2PB/174.5PB 72%
- 处理 17PB+ 数据/每天
- 平均CPU使用率 55 %, 峰值80 % - 90 %





过去一年的工作

- HDFS：
  - 规模问题改进：
    - Namenode 启动优化
      - 并行加载fsimage
    - Namenode rpc优化
      - registerChannel 锁优化 HADOOP-7105
      - 使用独立线程RegisterChannel 和cleanup
  - 数据安全问题改进
    - 块复制机制改进

过去一年的工作

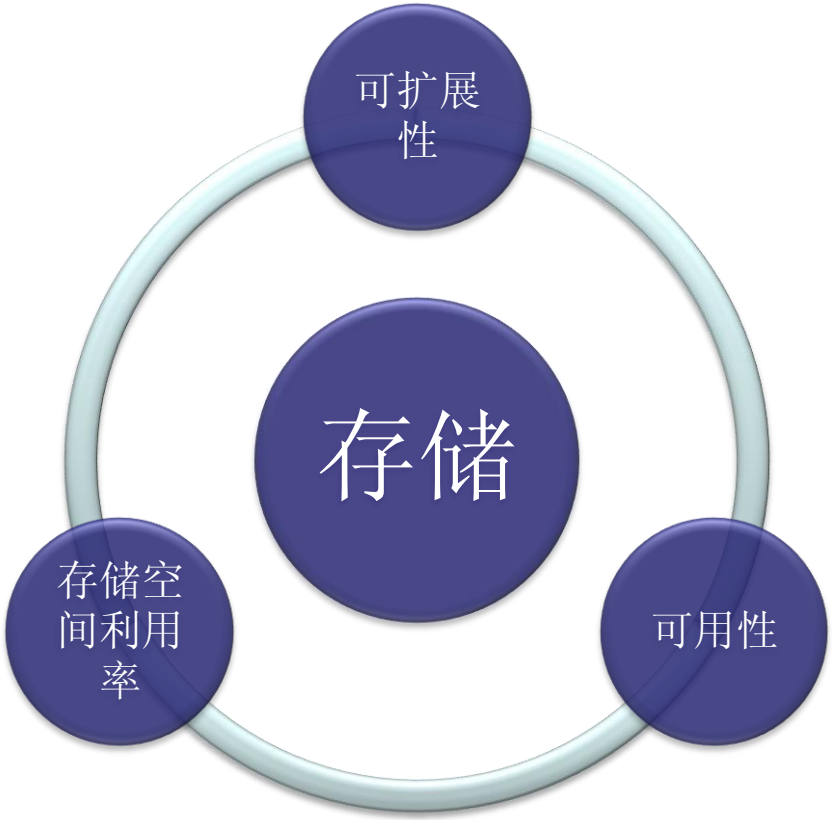
- Hard link
  - Why not symbol link?
- 跨机房优化
  - 跨机房提交作业
    - listStatus + getBlockLocation = too many rpc ?
  - 跨机房数据传输
    - dfs.send.socket.buffer.size (datanode, client)
    - dfs.datanode.recv.buffer.size (datanode)

过去一年的工作

- Mapred:
  - Shuffle独立
  - Hce 2.0
    - Hce基础上支持streaming 接口
  - 作业断点重启

进行中的项目

- 存储



主要内容

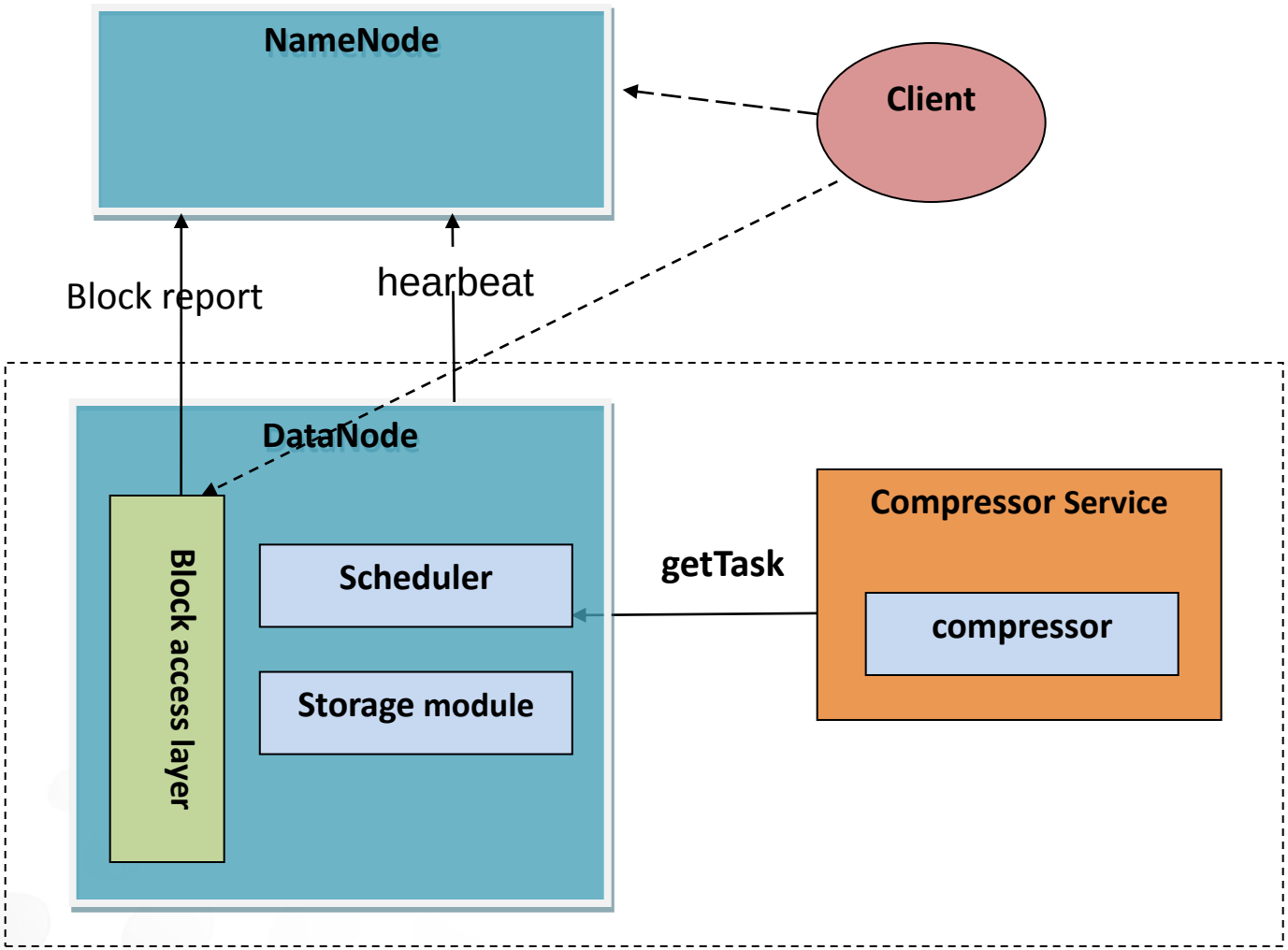
- Hadoop @baidu
  - 过去一年的工作
  - 进行中的项目
- 透明压缩
  - 目标
  - 实现
  - 如何解决关键问题
  - 如何规避风险
  - 未来计划

Baidu 百度

目标

- 节省存储空间
- 避免压缩影响计算作业
- 用户透明

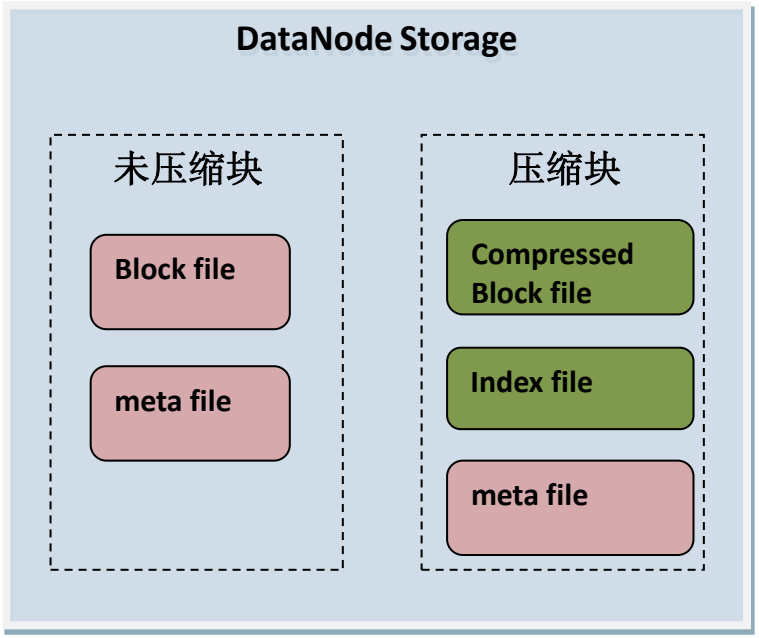
实现

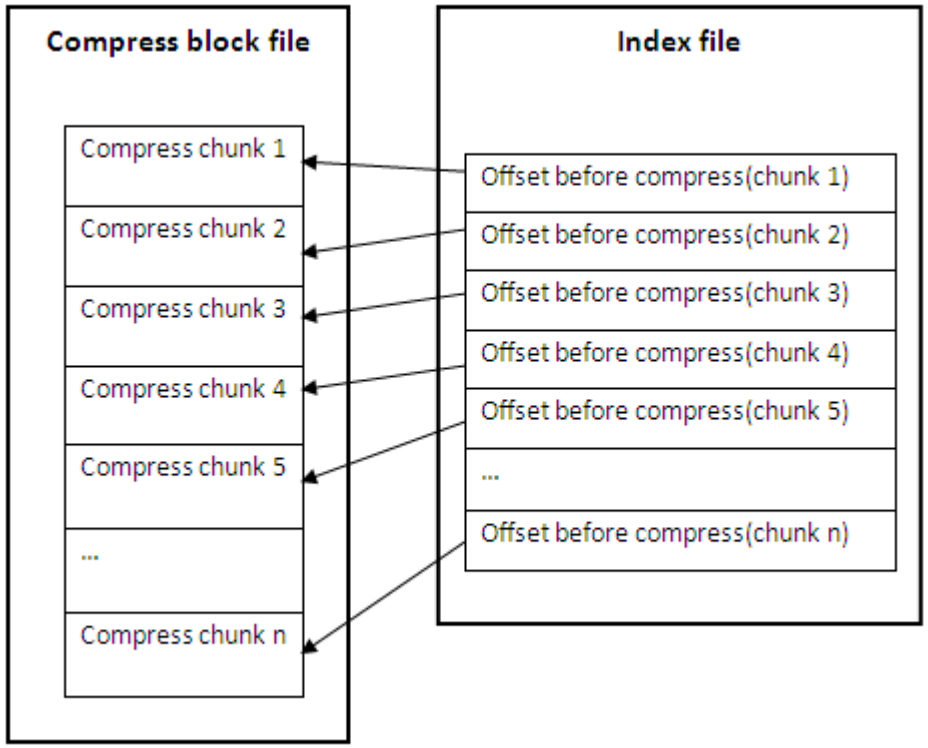

NameNode
Client
Block report
hearbeat
DataNode
Block access layer
Scheduler
Storage module
getTask
Compressor Service
compressor

如何解决关键问题？

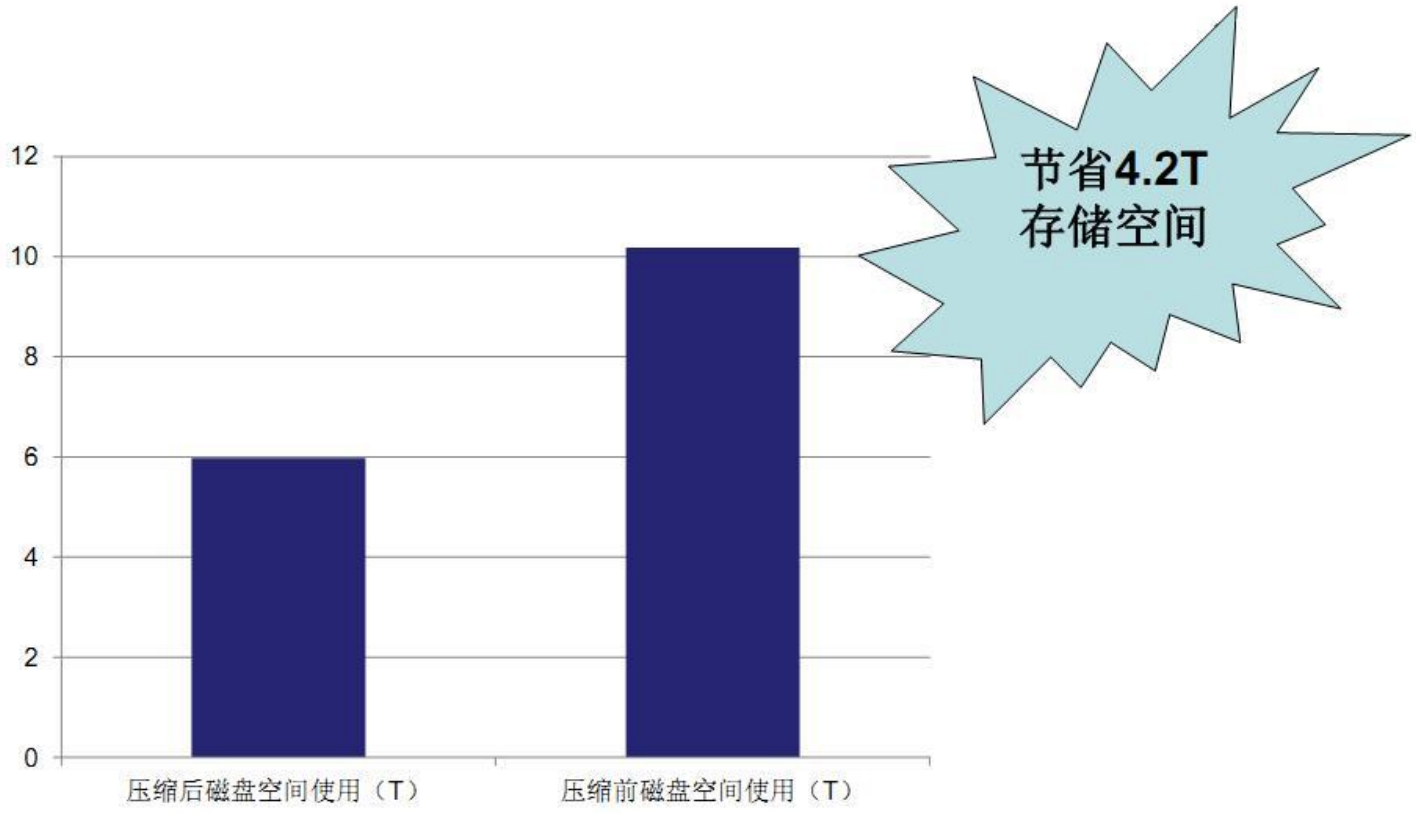
- 如何控制资源使用
  - Datanode 任务分配 (Xceiver 数)
  - 2.6.32 内核进程/ io 优先级调度
- 如何确定冷数据
  - 增加block的atime，1周没有访问？
- 如何处理特殊操作
  - append
  - 随机读

存储结构


DataNode Storage
未压缩块
Block file
meta file
压缩块
Compressed Block file
Index file
meta file


Compress block file
Compress chunk 1
Compress chunk 2
Compress chunk 3
Compress chunk 4
Compress chunk 5
...
Compress chunk n
Index file
Offset before compress(chunk 1)
Offset before compress(chunk 2)
Offset before compress(chunk 3)
Offset before compress(chunk 4)
Offset before compress(chunk 5)
...
Offset before compress(chunk n)

收益



总块数 21W， 可压缩块数10.2W，可压缩块占48.6 %
可压缩块压缩比3.28， 平均压缩比1.71

如何规避风险

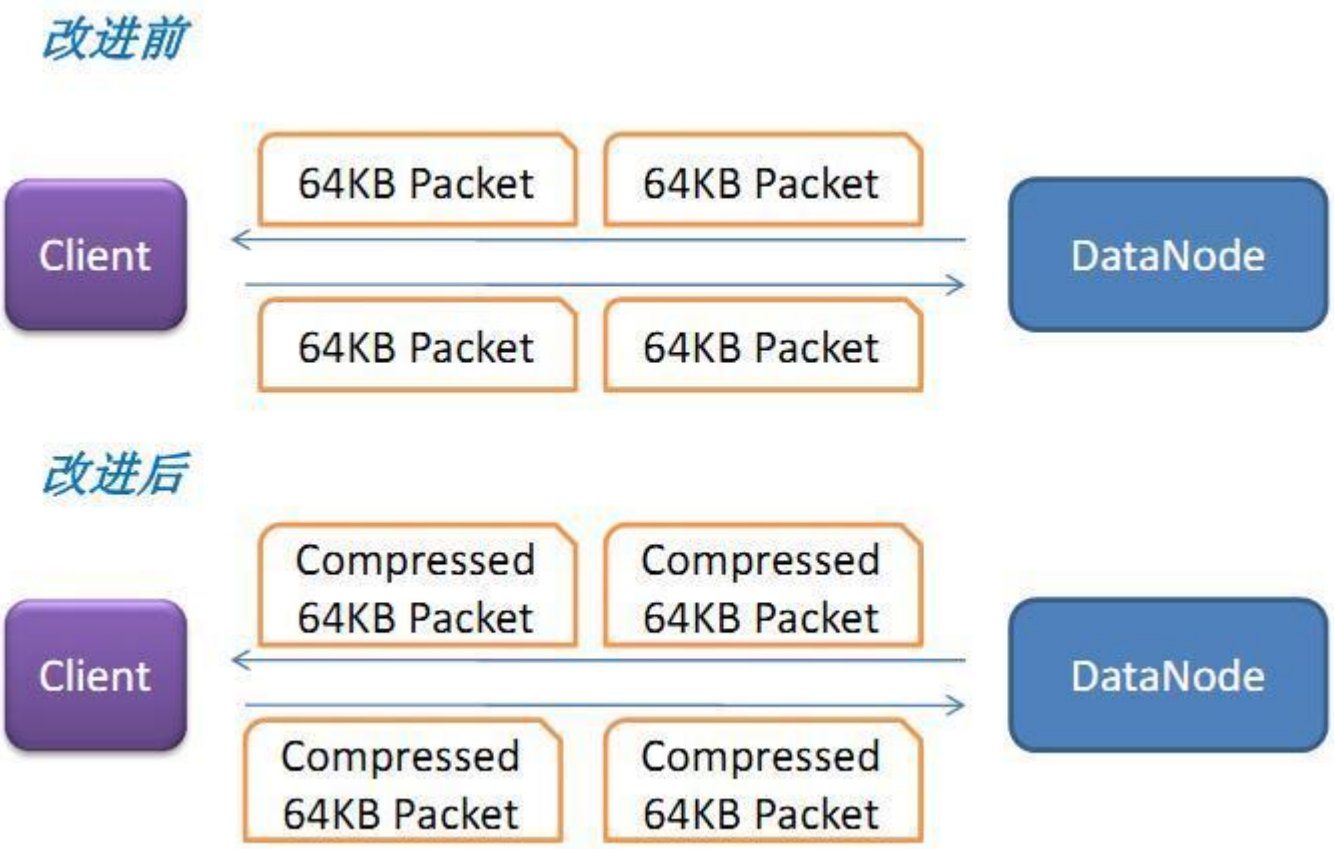
- 尝试解压
  - 目的： 规避压缩算法bug
- 小流量上线
  - 目的：上线一个机架datanode，避免透明压缩bug导致数据丢失
- 黑白名单

In the future

- 开源
  - https://issues.apache.org/jira/browse/HDFS-2542
- 多出的Quota分给谁
- 协处理器应用

In the future

- 透明压缩传输

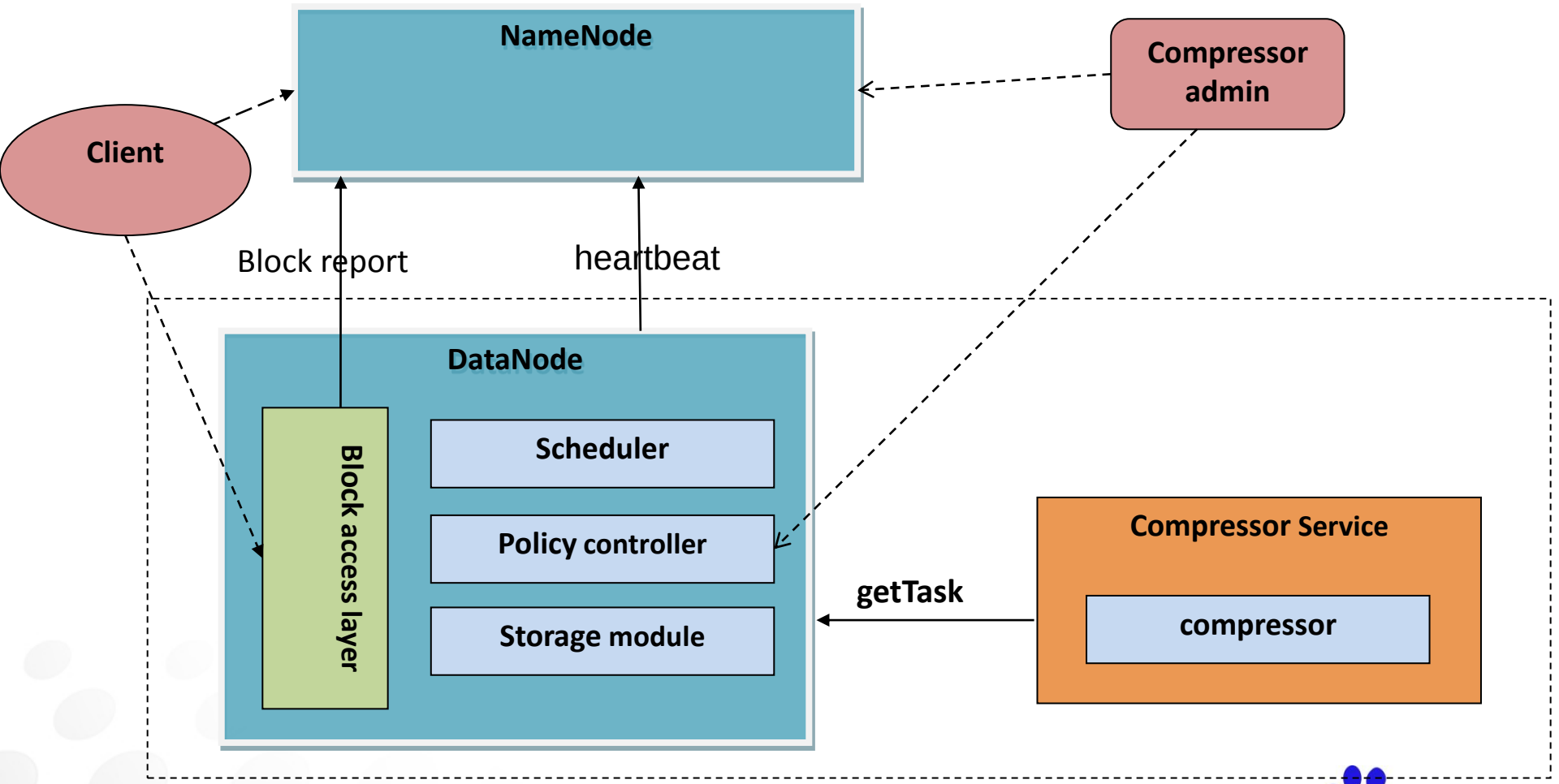


Q & A

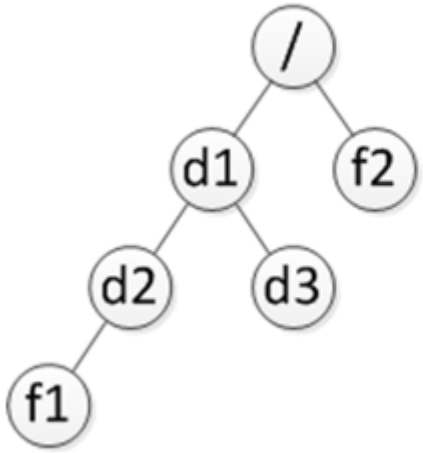
谢　谢！

Baidu 百度

透明压缩黑名单实现

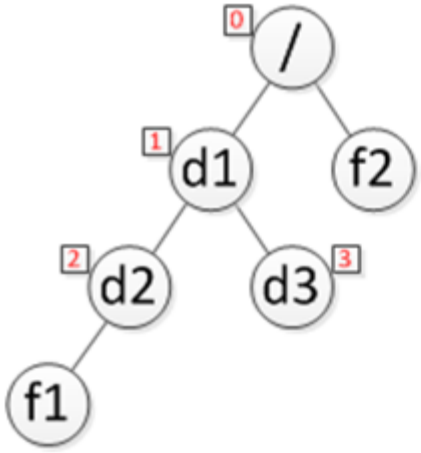

NameNode
Compressor admin
Client
Block report
heartbeat
DataNode
Block access layer
Scheduler
Policy controller
Storage module
getTask
Compressor Service
compressor

Fsimage并行加载



is dir	local name	num children
True	/	2
True	d1	2
True	d2	1
False	f1	-
True	d3	0
false	f2	-

HDFS-1070 短路径优化



dir-image

local name	num children	array of subdir indices
/	2	0
d1	2	0,1
d2	1	-1
d3	0	-

file-image

local name	parent serial num	index in parent
f1	2	0
f2	0	1

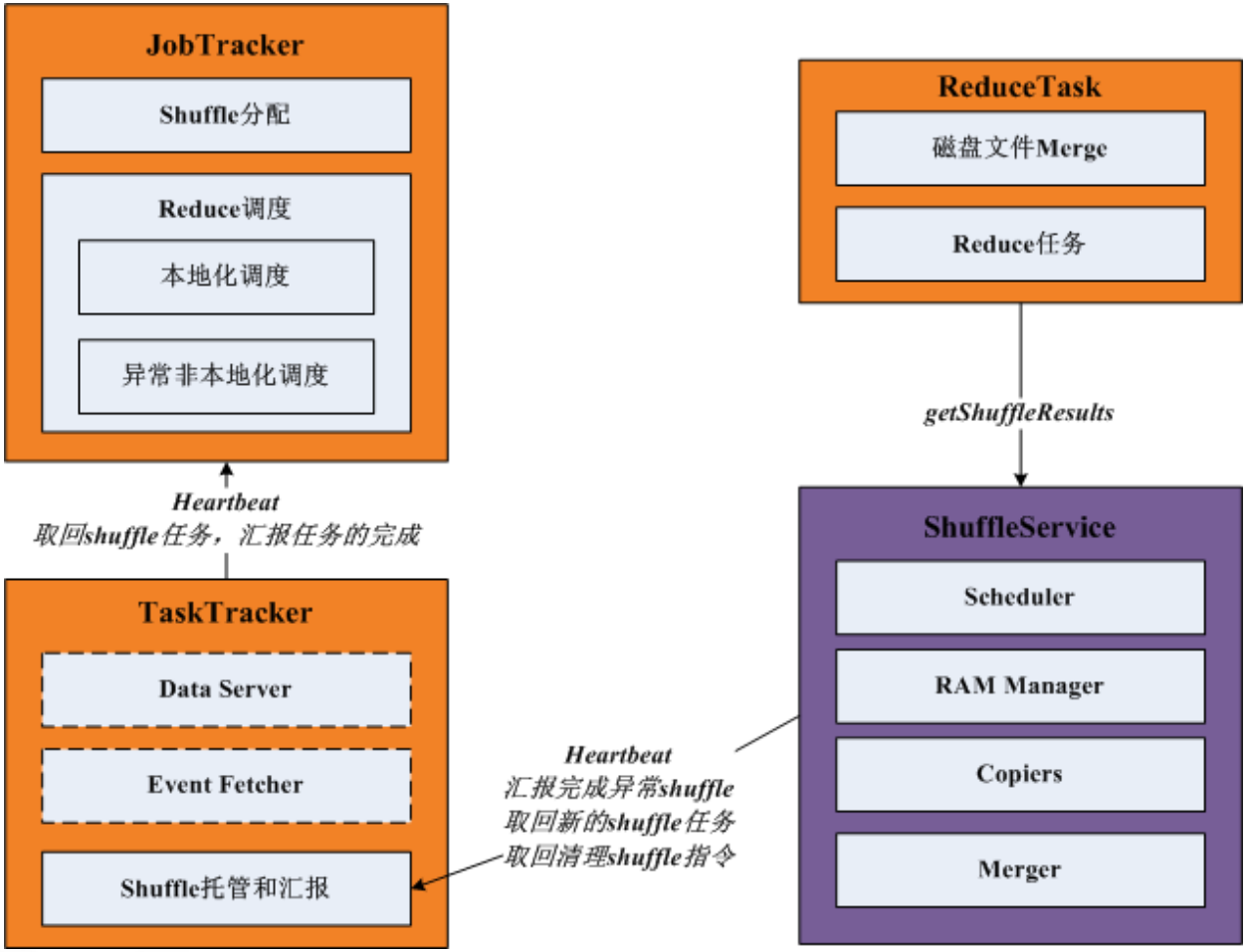
并行加载fsimage

Baidu 百度

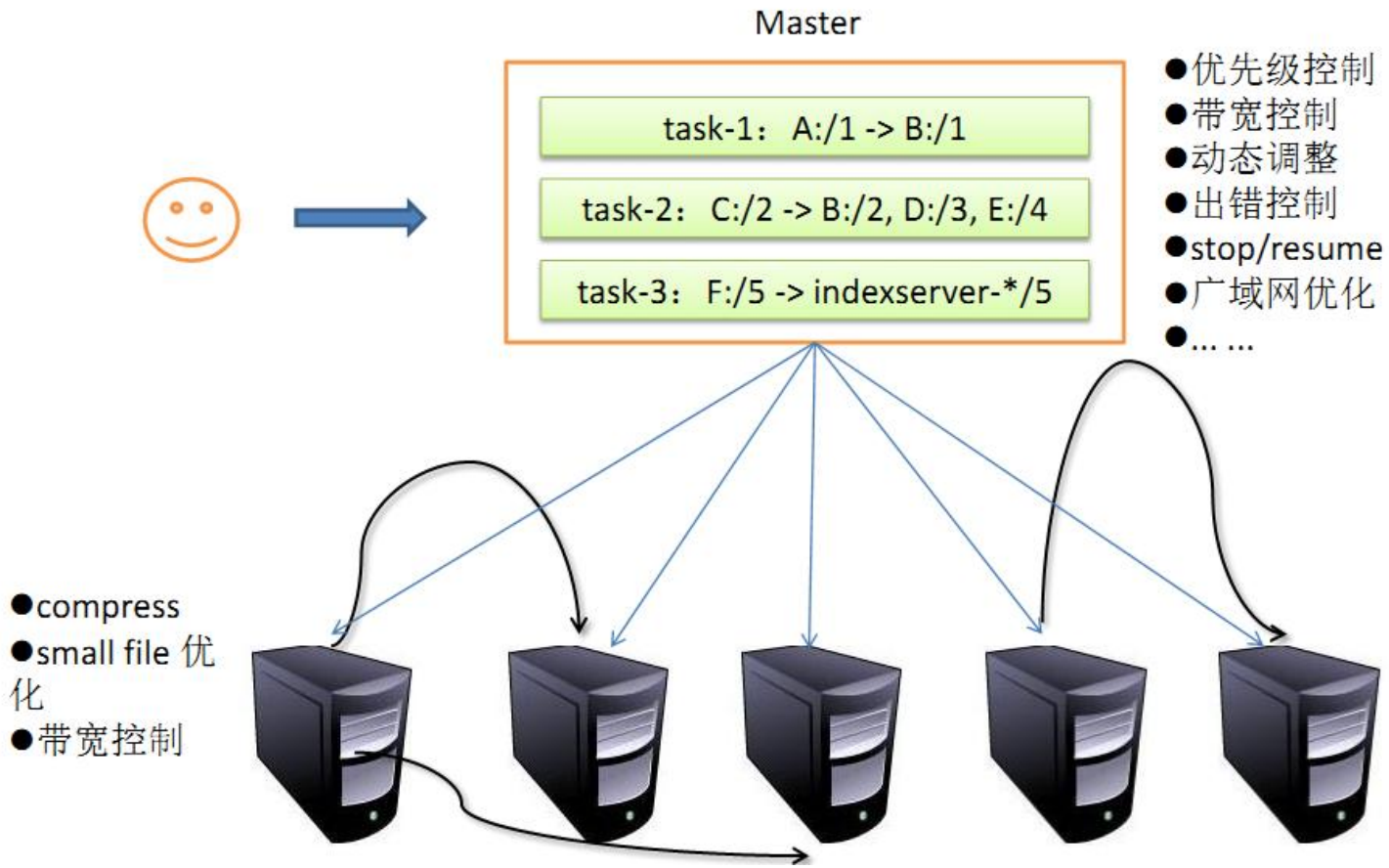
Shuffle独立- 解决问题

- map/reduce 槽位隔离，槽位利用率低
- shuffle占用 reduce槽位，资源利用率低
- shuffle和reduce串行，对大作业，运行时间长
- Shuffle/reduce自身的问题，内存利用率不高，连接数打满

Shuffle独立 – 结构



传输项目


Master
task-1： A:/1 -> B:/1
task-2： C:/2 -> B:/2, D:/3, E:/4
task-3： F:/5 -> indexserver-*/5
●优先级控制
●带宽控制
●动态调整
●出错控制
●stop/resume
●广域网优化
●... ...
●compress
●small file 优化
●带宽控制