百度分布式计算技术发展

连林江
lianlinjiang@baidu.com
2012.07.08

我

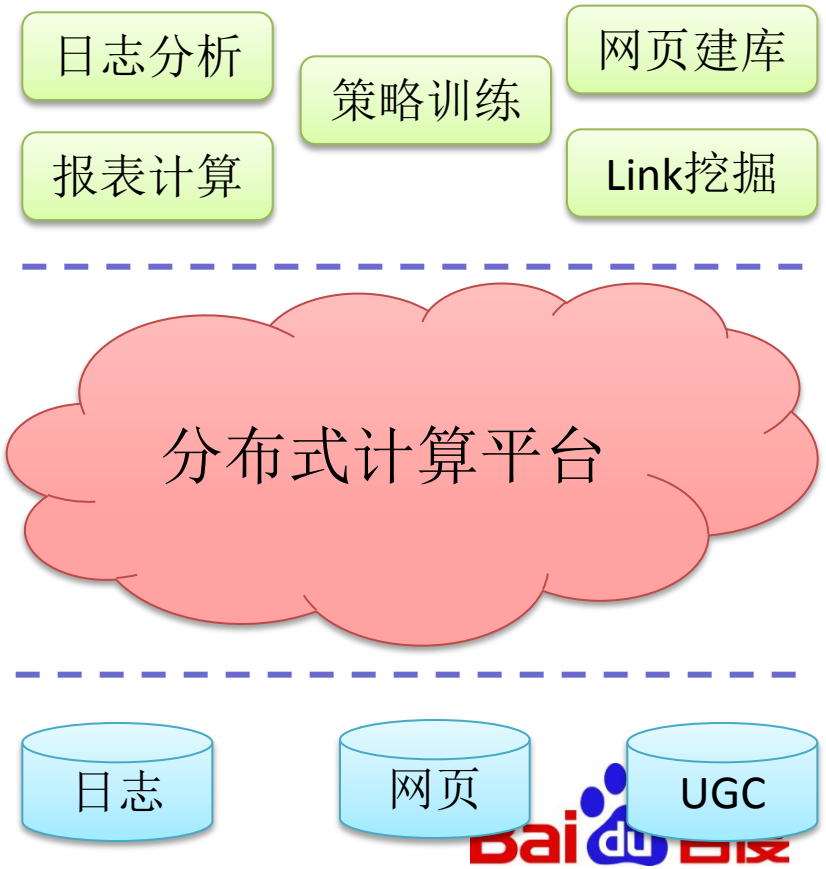
- 基础架构部
  - 项目经理
- 负责分布式计算团队
  - HDFS
  - MapReduce及其他批量计算模型
  - Resource Management System

大纲

- 分布式计算平台
- 我们的挑战
- 分布式计算技术2.0
- 展望

分布式计算平台

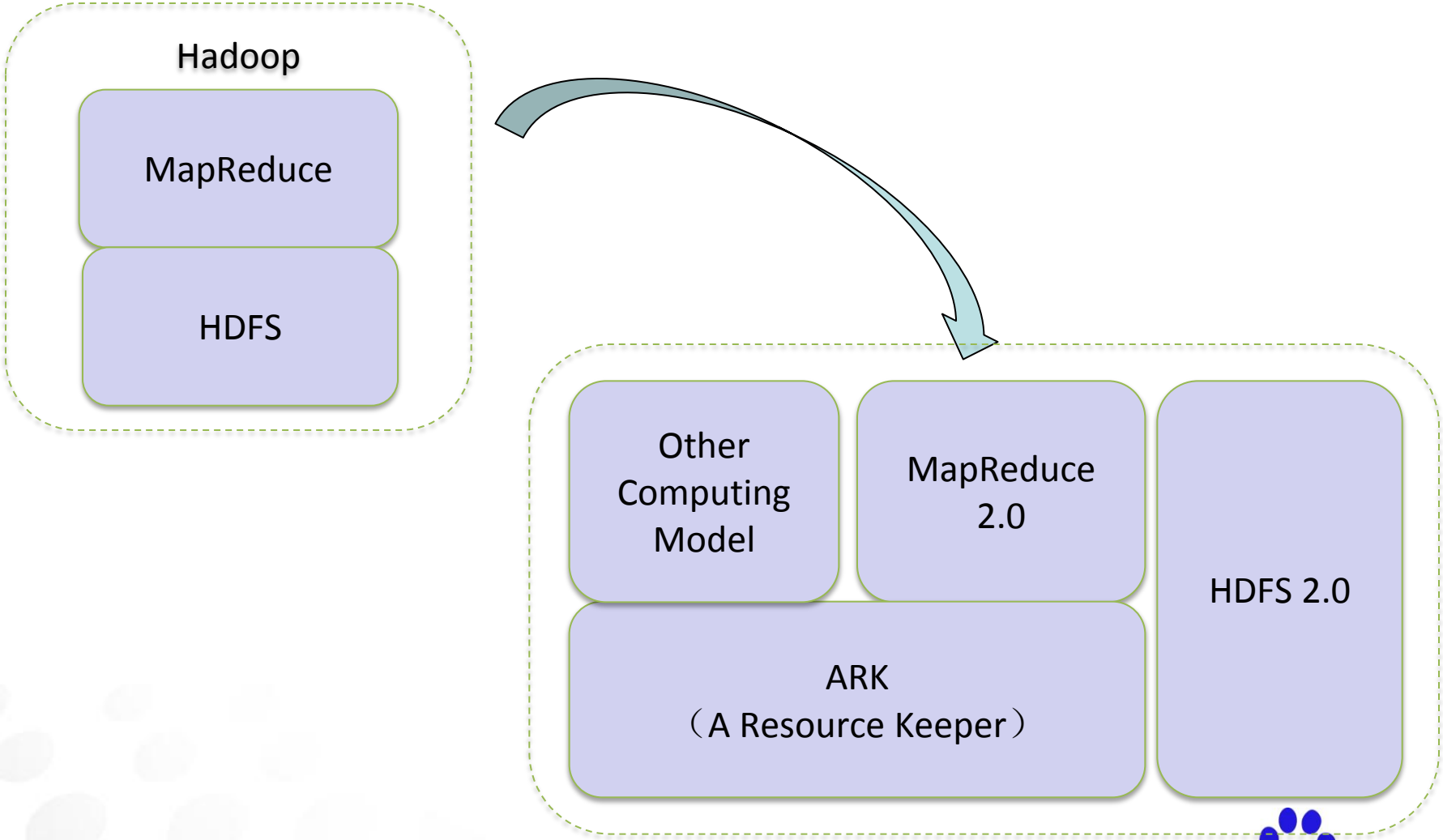
- 2008
  - 开始于Hadoop v0.18/0.19
  - 300台机器，2个集群
- Now
  - 总规模2W以上
  - 最大集群接近4,000节点
  - 每日处理数据20PB+
  - 每日作业数120,000+



我们的挑战

- 规模
  - 单集群1000→2000→3000→5000→10000
- 效率
  - 资源利用率（cpu/mem/io）—高峰vs平均
  - 存储利用—无压缩、冷数据
  - 存储与计算资源使用均衡问题
- 服务可用
  - 随着规模增大问题变得突出
    - 3K+节点升级或异常小时级中断
    - 用户影响面：在可用99.9%下用户容忍度变低

分布式计算技术2.0


Hadoop
MapReduce
HDFS
Other Computing Model
MapReduce 2.0
HDFS 2.0
ARK
（A Resource Keeper）

HDFS 2.0--Scalability

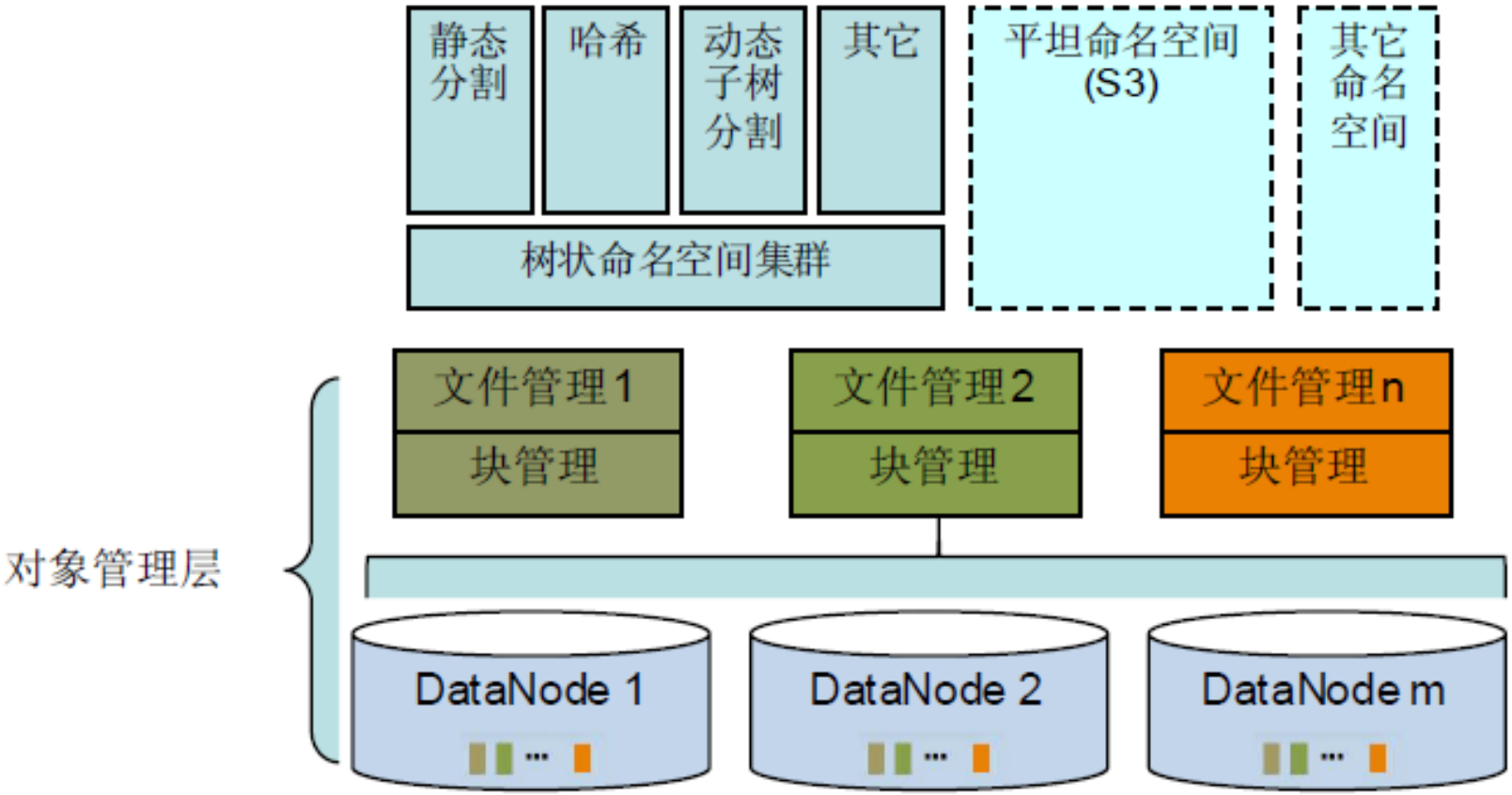
- 1.0面临问题
  - 内存可扩展性
    - 1.5亿文件/1.2亿块，内存占用90GB

Cluster Summary

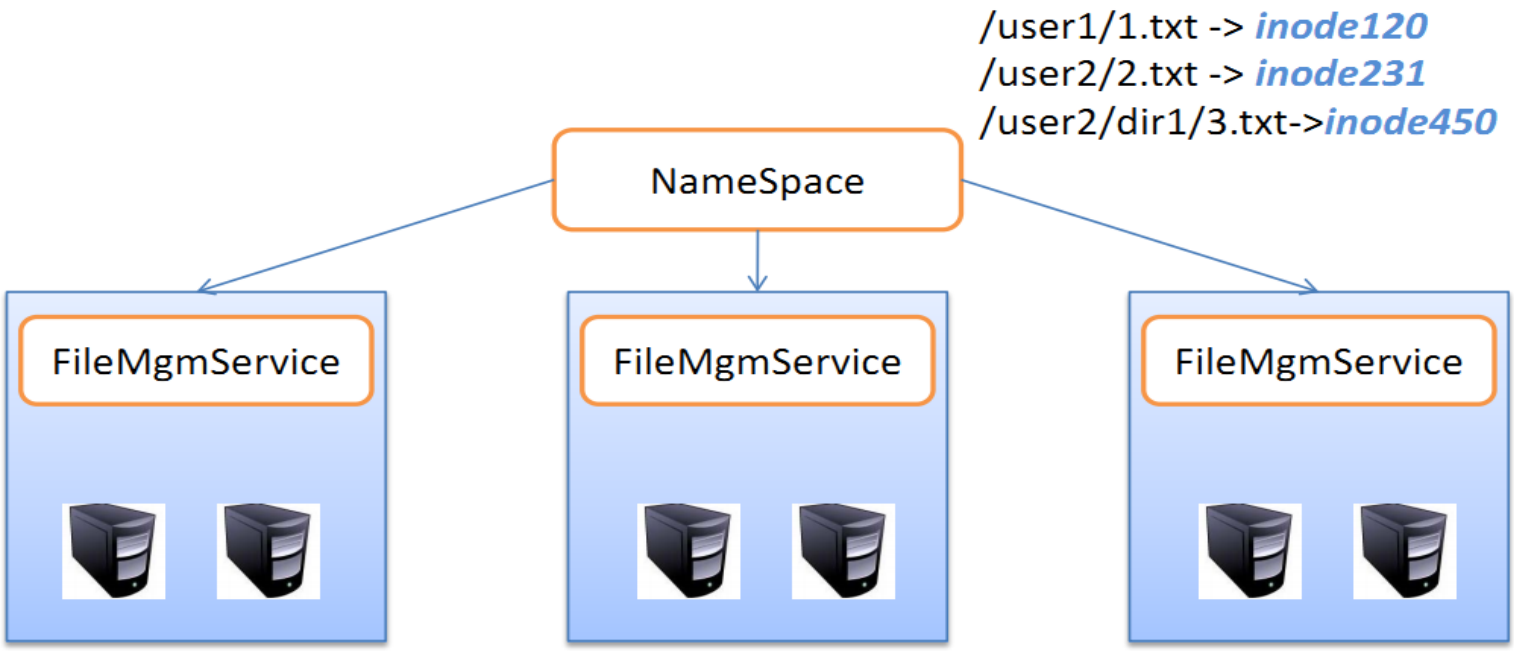
156146781 files and directories, 120438664 blocks = 276585445 total.
Heap Memory used 89.51 GB is 69% of Commited Heap Memory 127.95 GB. Max Heap Memory is 127.95 GB.
Non Heap Memory used 29.99 MB is 66% of Commited Non Heap Memory 45.42 MB. Max Non Heap Memory is 132 MB.

  - 负载可扩展性
    - 集群规模扩大→单点NameNode请求压力增大
    - 3000节点：连接超时/拒绝，有时操作响应延迟高

HDFS 2.0--Scalability



HDFS 2.0--Scalability



inode120->
user,group, rwxr-x---, size, repl
blkX, blkY, blkZ

inode450->
user,group, rwxr-x---, size, repl
blkX, blkY, blkZ

inode231 -> xxxxxx

inodexxx -> xxxxxx

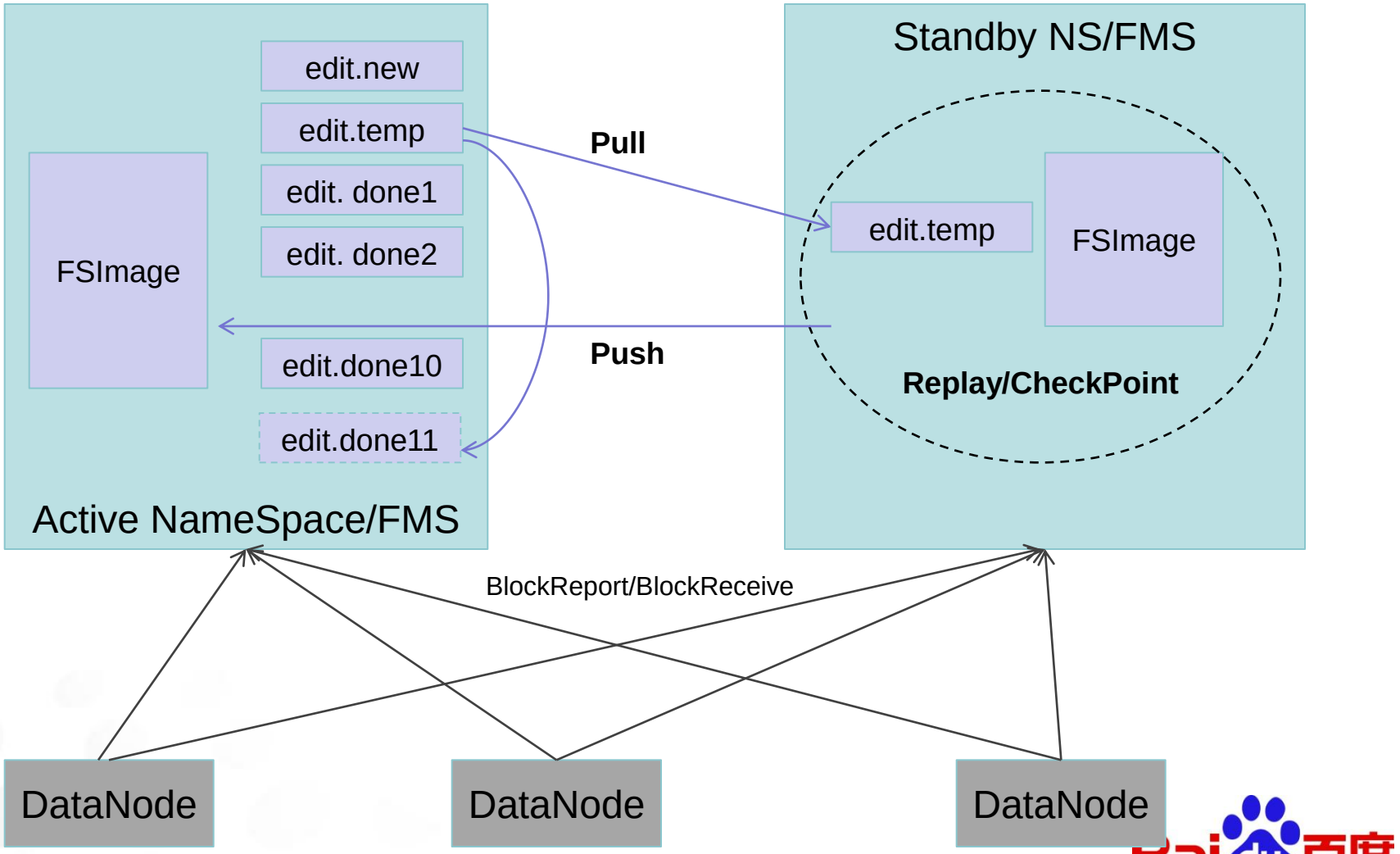
HDFS 2.0--Scalability

- 内存负载: 10亿文件，10亿块
  - Namespace：66GB文件数据+1GB目录，单节点管理
- 请求负载
  - 13.7%耗cpu操作 →Namespace
  - Namespace不再维护块信息，大部分操作都不需要加全局锁，可以更充分利用CPU资源
- 吞吐
  - 按照我们的负载读写比例 x5~10

HDFS 2.0--Availability

- 1.0面临问题
  - NameNode单点/手工Failover
  - 启动/升级时间长
    - 2亿文件/3K节点，启动时间40-50分钟（百度）

HDFS 2.0--Availability



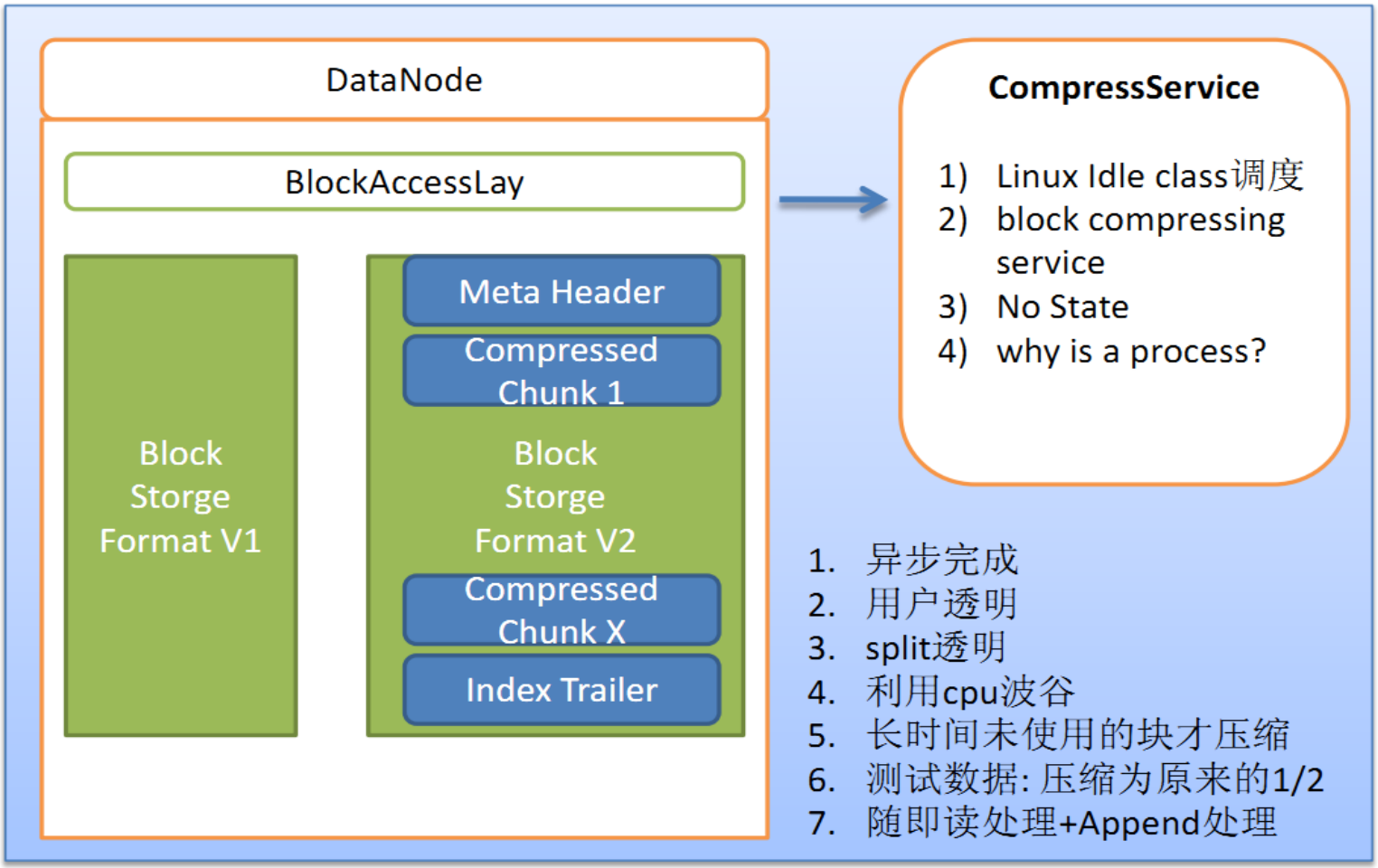
HDFS 2.0--Availability

- 热备支持
- 分钟级别切换
- 最坏情况，应用可能丢失1分钟级数据

HDFS 2.0--透明压缩

- 存储压力很大？
  - 很多是存储决定预算
  - 70-80%使用率
- 为什么不压缩？
  - 应用层压缩后，造成无法对数据split来分布式计算
  - 使用可分割的压缩算法，使用非常复杂
  - 压缩需要同步耗费CPU
  - 用户希望透明
- 冷数据
  - 使用不频繁
  - 量很大
  - 存储成本较高

HDFS 2.0--透明压缩


DataNode
BlockAccessLay
Block
Storge
Format V1
Meta Header
Compressed
Chunk 1
Block
Storge
Format V2
Compressed
Chunk X
Index Trailer
CompressService
1) Linux Idle class调度
2) block compressing service
3) No State
4) why is a process?
1. 异步完成
2. 用户透明
3. split透明
4. 利用cpu波谷
5. 长时间未使用的块才压缩
6. 测试数据: 压缩为原来的1/2
7. 随即读处理+Append处理

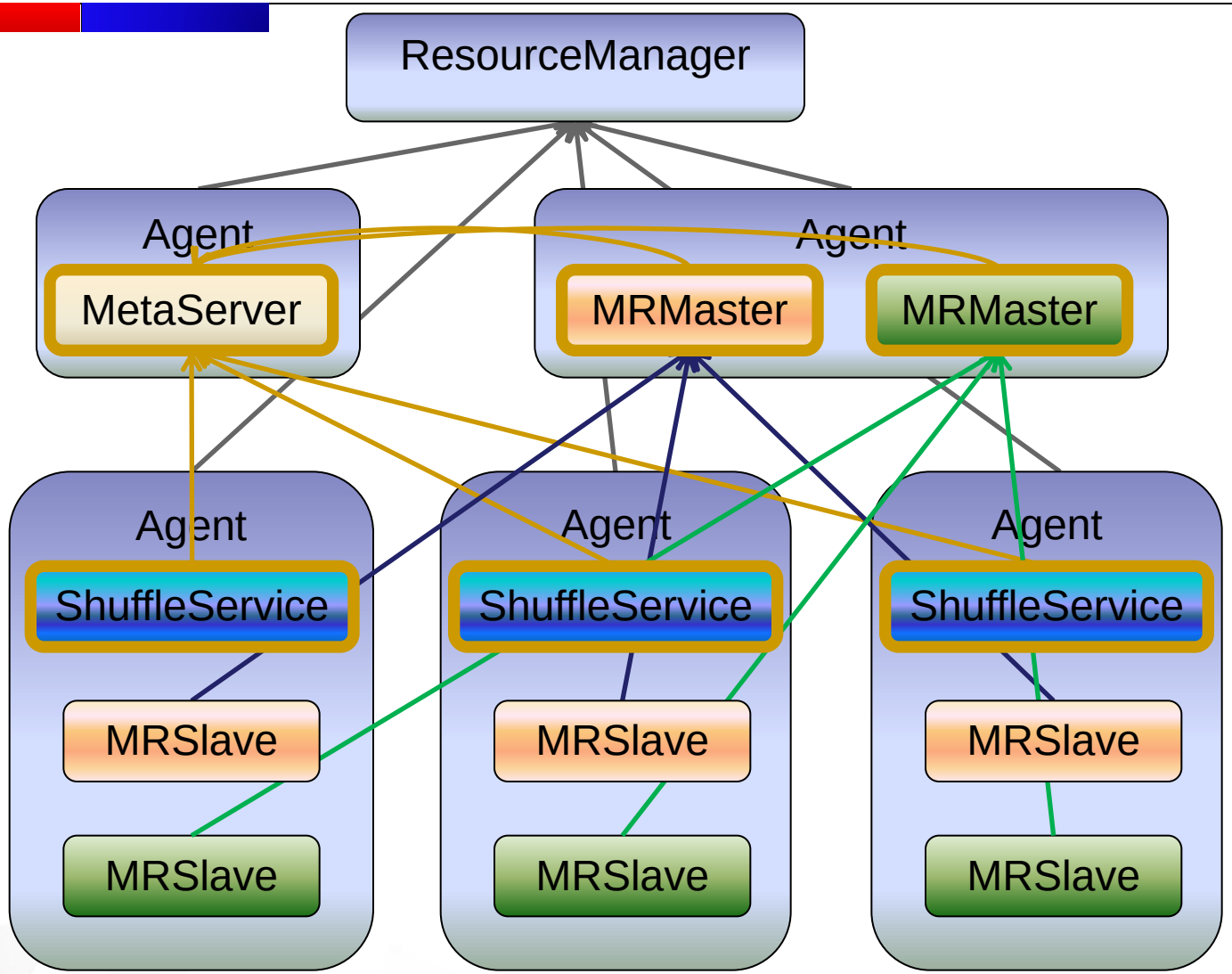
HDFS 2.0--透明压缩

- 改进效果
  - 节省存储空间30%+，增加Quota 40%+
  - 进一步的高压缩算法启用会有更大收益

MapReduce 2.0

- 1.0面临问题
  - JobTracker单点
    - 负载太重，扩展性受限→1W
    - 故障/升级中断服务重跑作业
  - 资源粒度过粗
    - slot（cpu、mem）
  - 资源利用不高
    - Shuffle+Reduce，空占slot

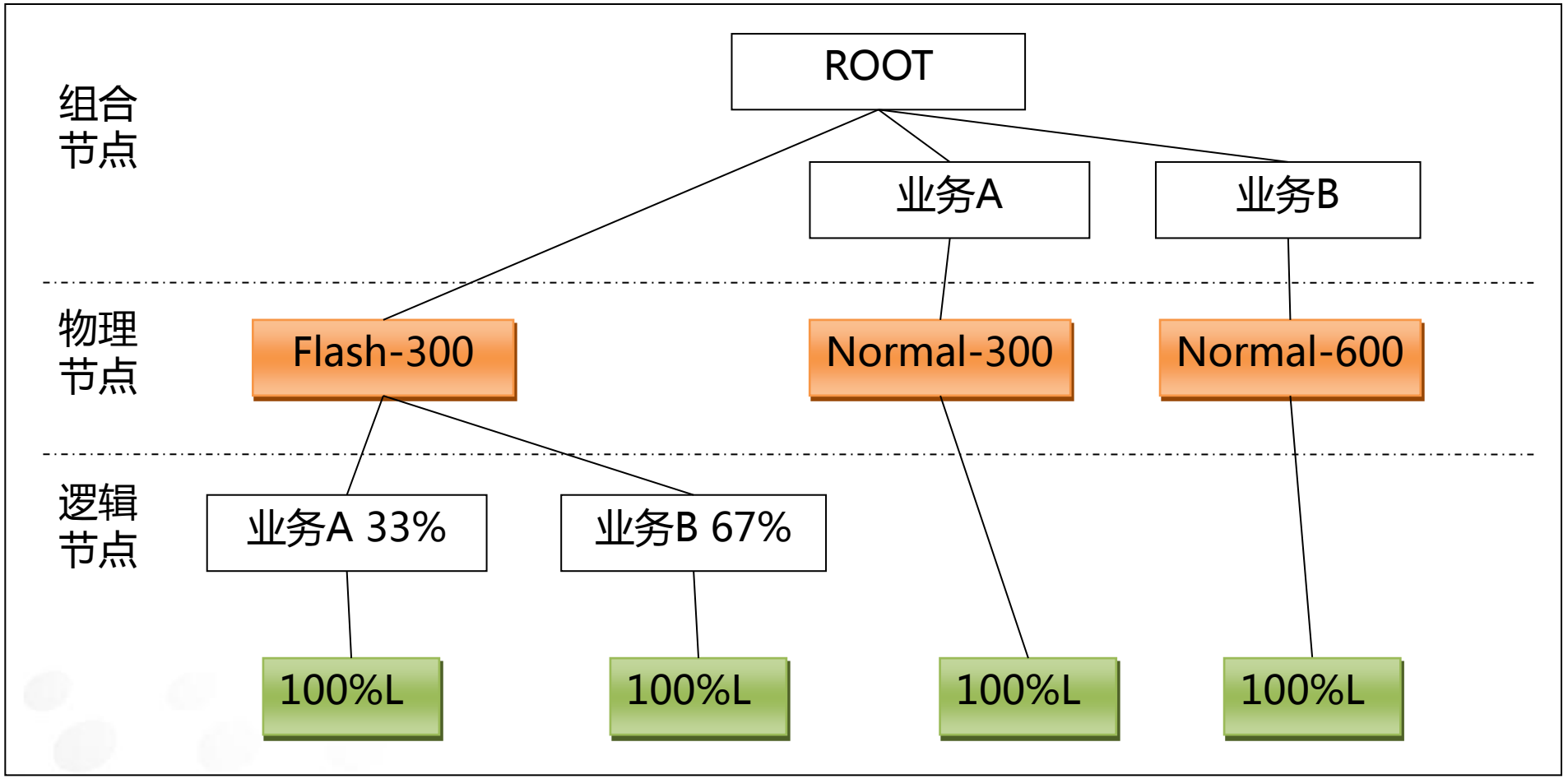
MapReduce 2.0



MapReduce 2.0 - 架构优势

- 可扩展性W台以上
- 架构松耦合，支持多种计算模型
- 可支持热升级
- 更精细的资源调度
- MR优化：Shuffle独立/Task同质调度

MapReduce 2.0 -资源模型



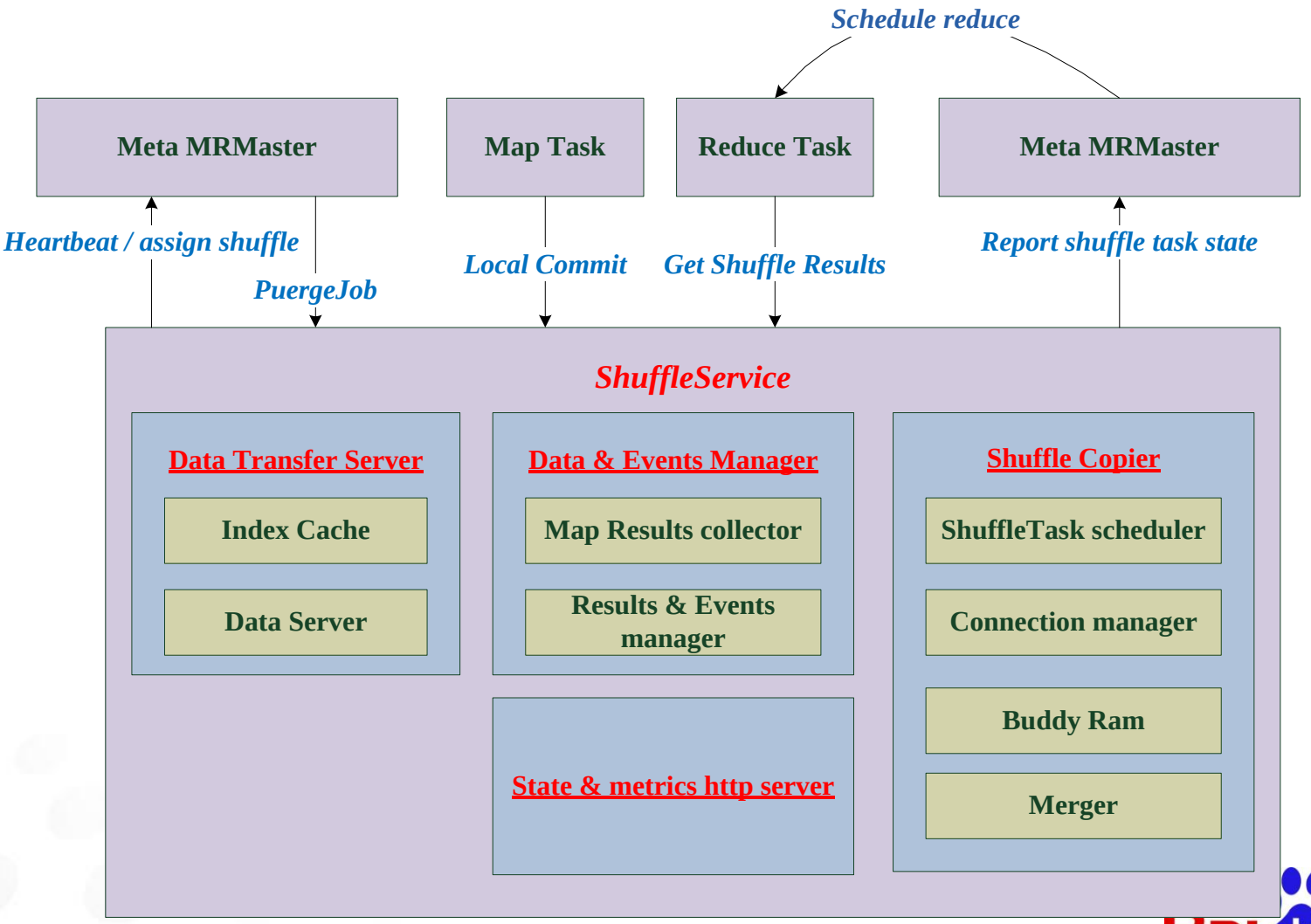
MapReduce 2.0 –资源模型

- 资源需求用一个多元组表示，目前使用(cpu, mem)，后续可以变成(cpu, mem, disk, disk io, net io)
- 调度
  - 资源的共享与抢占
  - 作业的优先级
  - 资源的物理分组与逻辑分组

MapReduce 2.0 - 资源模型优势

- 资源充分共享
- 灵活的优先级控制
- 管理方便

MapReduce 2.0 – Shuffle独立



Shuffle Service Admininstration

Ram Manager

shuffleBufferMegabytes	maxSingleShuffleLimit	startMergePercent(%)	PercentUsed(%)	waitForMemoryThreads	numStarted	numClosed	maxInMemOutputs	m
1024	33554432	80.0	0.0	0	0	0	20000	0.

Shuffle Result

ShuffleID	JobPriority	ResultStatus	ResultFileNum	ResultFileList
job_201108291638_0001-195	NORMAL	DONE	2	&file:/home/disk4/dcmmapred/shuffleService/job_201108291638_0001/195/output/map_283.out& /shuffleService/job_201108291638_0001/195/output/map_46.out
job_201108291638_0001-198	NORMAL	DONE	2	&file:/home/disk6/dcmmapred/shuffleService/job_201108291638_0001/198/output/map_351.out& /shuffleService/job_201108291638_0001/198/output/map_2.out
job_201108291638_0001-197	NORMAL	DONE	2	&file:/home/disk7/dcmmapred/shuffleService/job_201108291638_0001/197/output/map_176.out& /shuffleService/job_201108291638_0001/197/output/map_46.out
job_201108291638_0001-199	NORMAL	DONE	2	&file:/home/disk8/dcmmapred/shuffleService/job_201108291638_0001/199/output/map_352.out& /shuffleService/job_201108291638_0001/199/output/map_46.out

ShuffleWorks

shuffleWork	totalMaps	remainingMaps	usedMem	inMemorySegments	onDiskSegments	totalFailures	pendingHosts	fetchFailedMaps	penalties

ShuffleCopiers

Copier-Id	Running ShuffleWork
1	null
2	null
3	null
4	null
5	null

展望

- W台以上大集群
  - 高吞吐高资源利用率
- HDFS
  - 压缩传输&分级压缩
  - Untility Storage
- MapReduce
  - DAG
- IDLE计算平台

Baidu 百度

Q & A

谢　谢！

Baidu 百度