百度大数据离线计算平台流式Shuffle服务

百度　张建伟

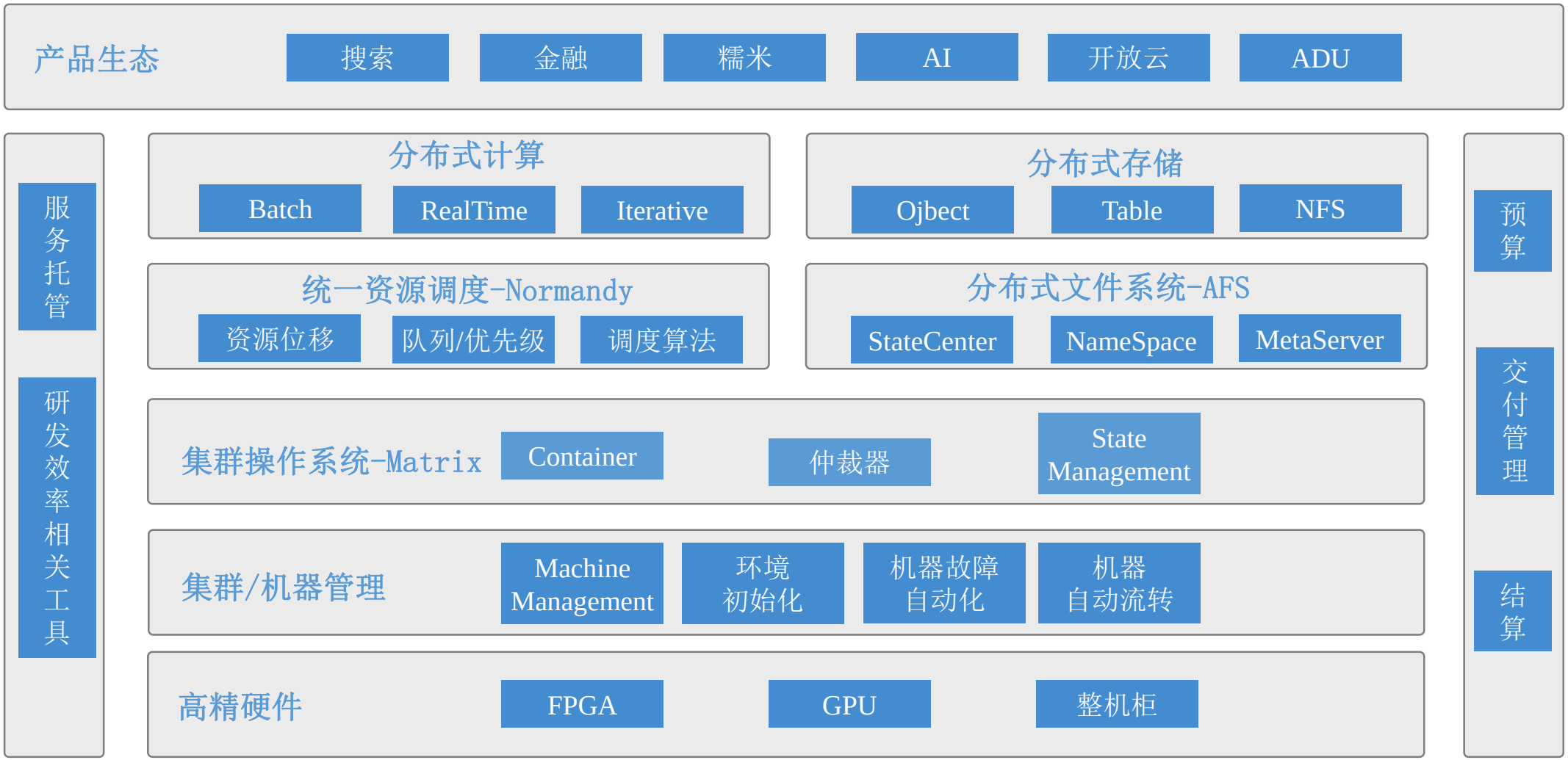
目录

目录

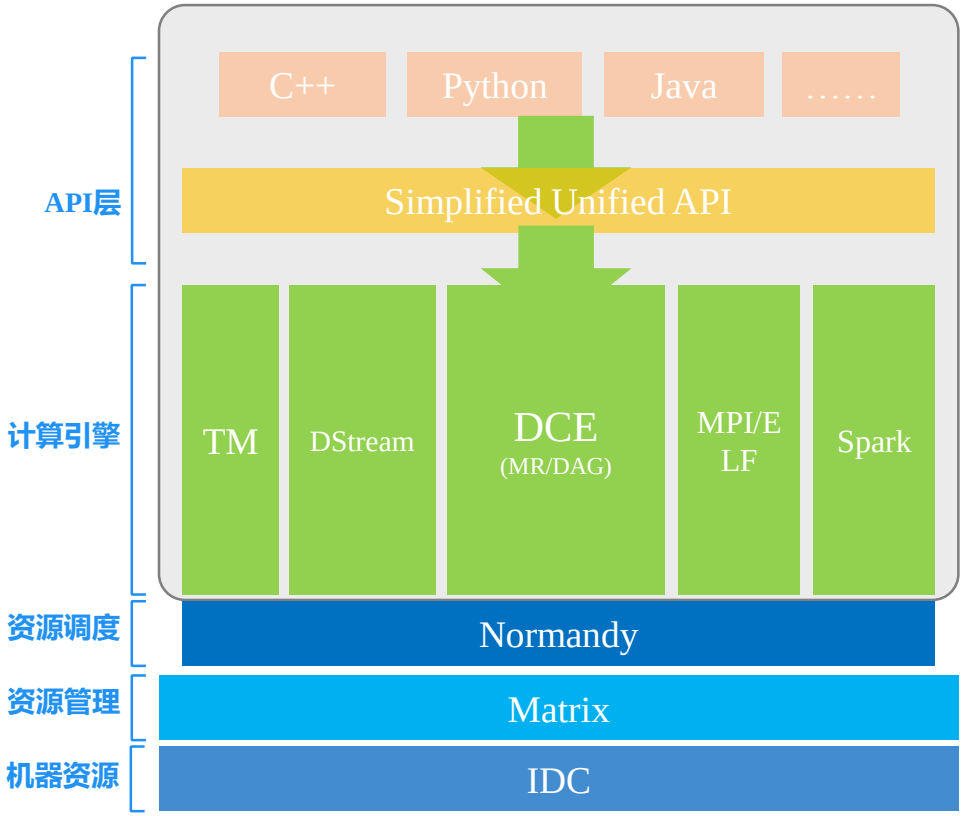
百度云

- 背景
- 架构
- 关键技术
- 收益与总结
- 下一步计划

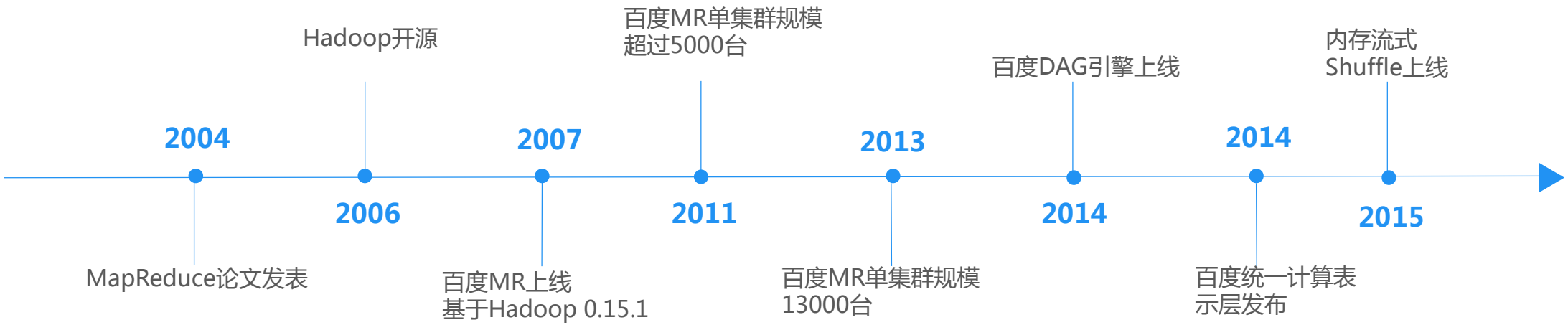
背景-百度私有云


产品生态
搜索
金融
糯米
AI
开放云
ADU
服务托管
研发效率相关工具
分布式计算
Batch
RealTime
Iterative
分布式存储
Ojbect
Table
NFS
统一资源调度-Normandy
资源位移
队列/优先级
调度算法
分布式文件系统-AFS
StateCenter
NameSpace
MetaServer
集群操作系统-Matrix
Container
仲裁器
State Management
集群/机器管理
Machine Management
环境初始化
机器故障自动化
机器自动流转
高精硬件
FPGA
GPU
整机柜
预算
交付管理
结算

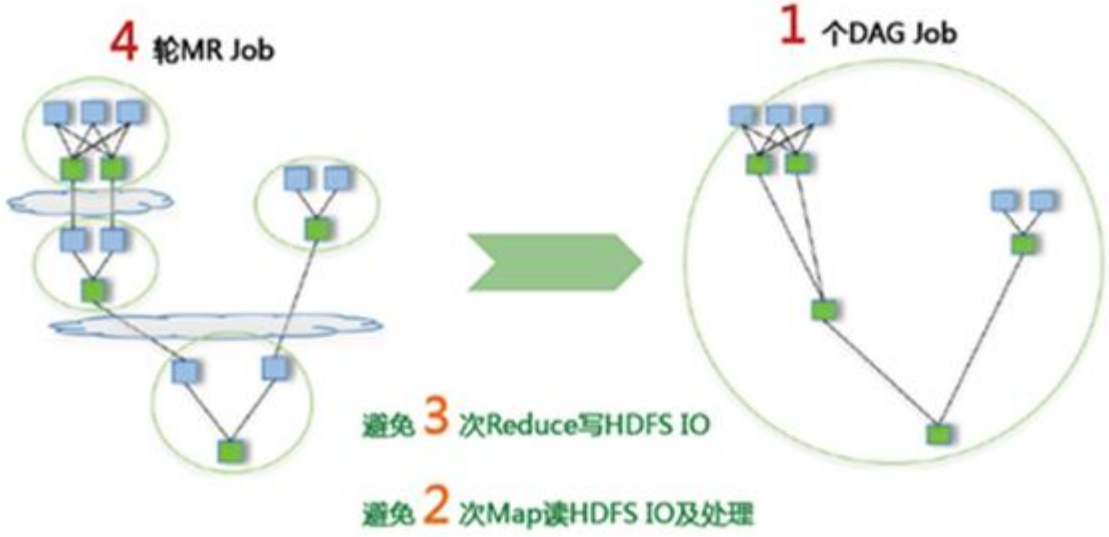
背景-百度大数据计算平台



背景-百度大数据离线计算平台发展历程



背景-DAG引擎

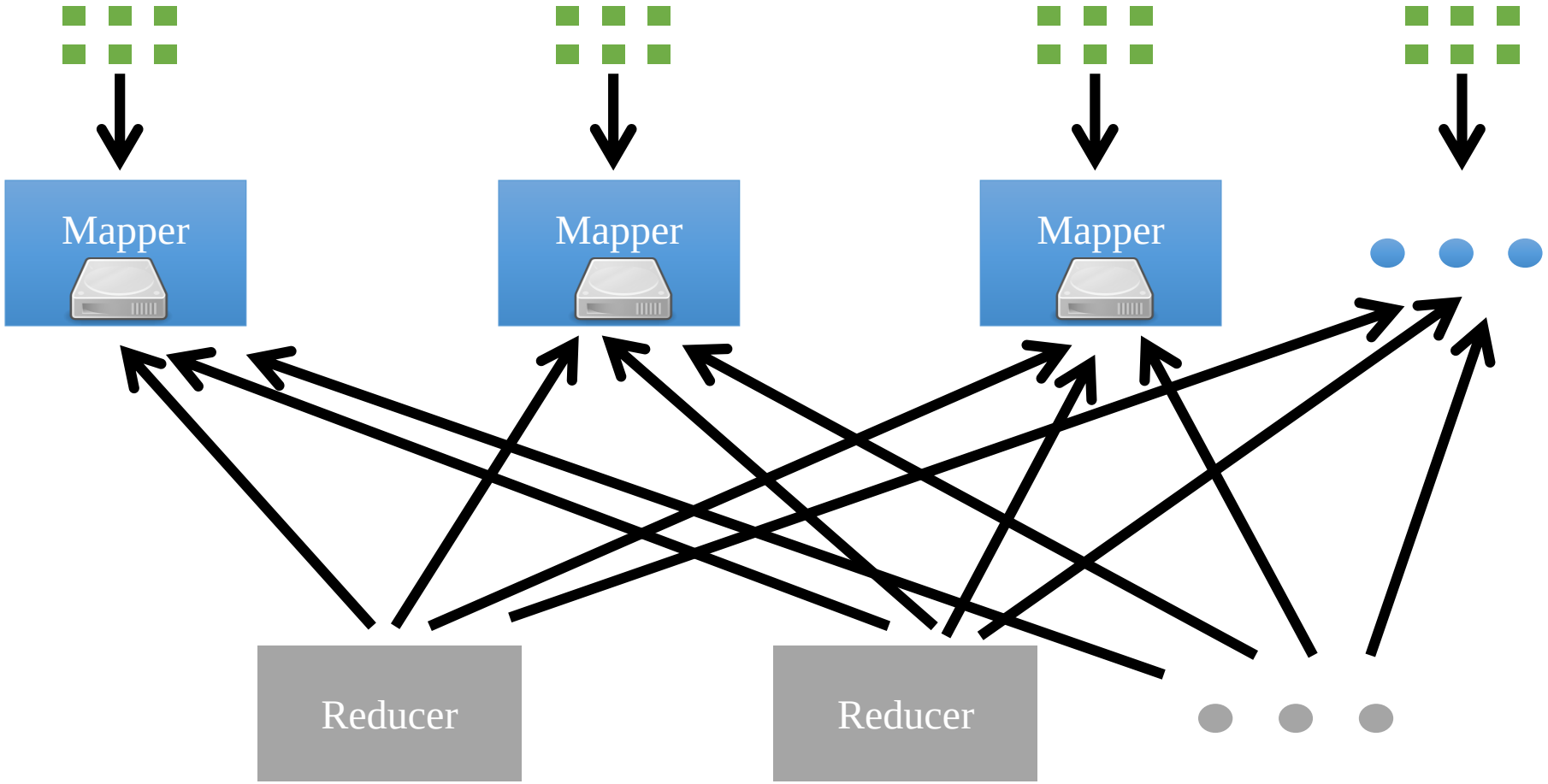


基于MR引擎，需要翻译成 *25个* MR Job

基于DAG引擎，只需要翻译成 *1个* DAG Job

运行时间从 *5小时* 缩减到 *1小时*

背景-一般的Shuffle模式

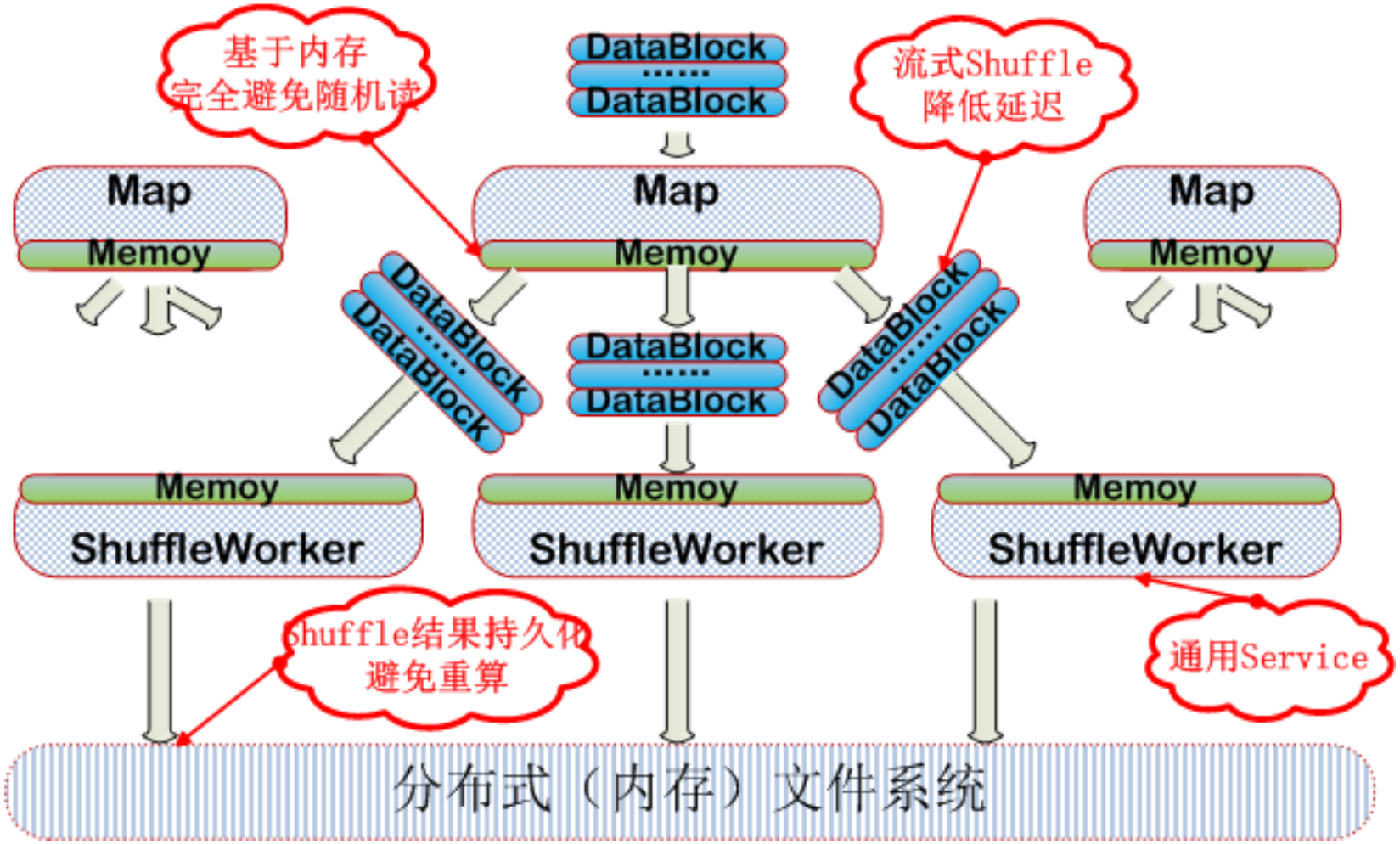


目录

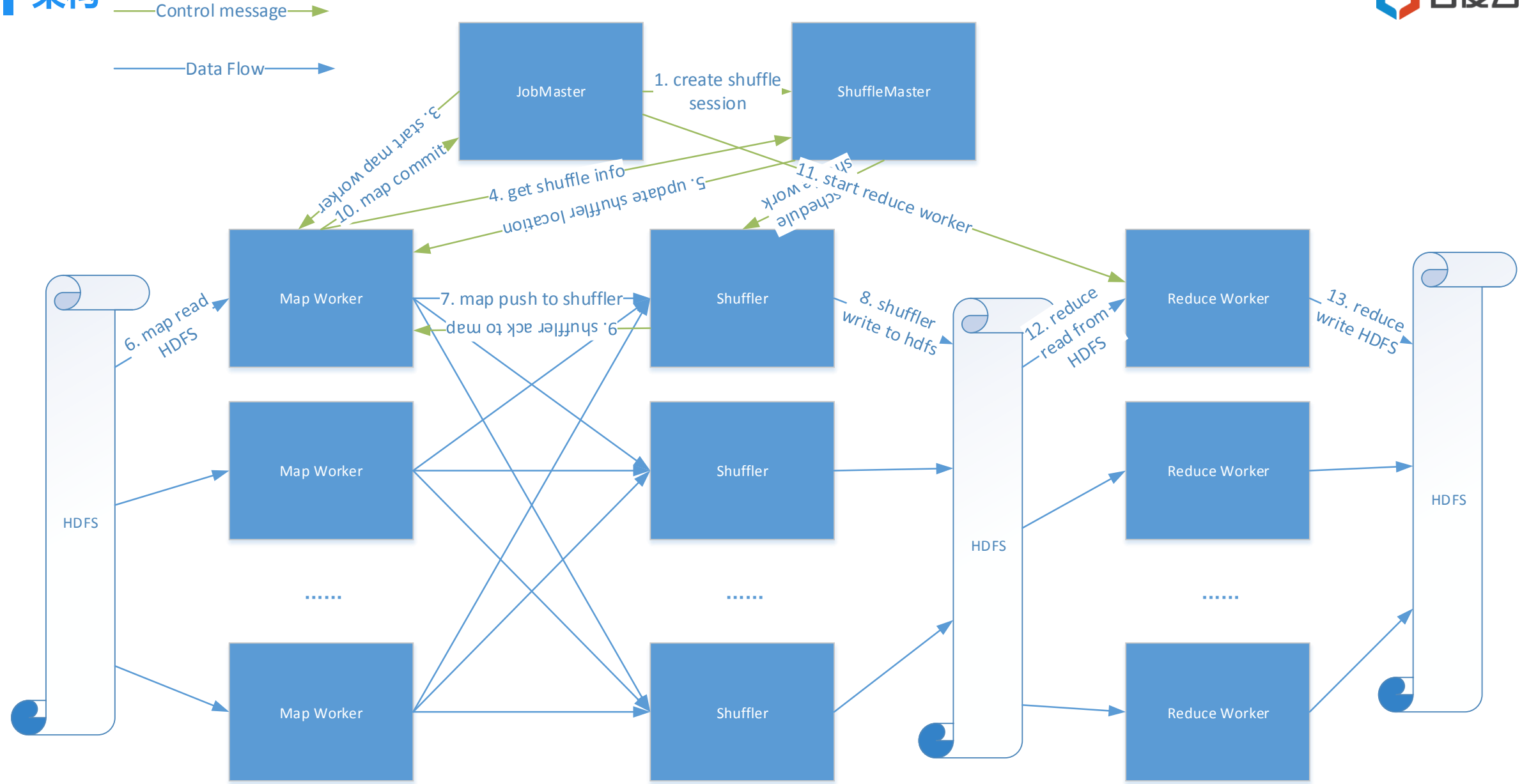
百度云

- 背景
- 架构
- 关键技术
- 收益与总结
- 下一步计划

架构


基于内存
完全避免随机读
DataBlock
.....
DataBlock
流式Shuffle
降低延迟
Map
Memoy
Map
Memoy
Map
Memoy
DataBlock
.....
DataBlock
DataBlock
.....
DataBlock
DataBlock
.....
DataBlock
Memoy
ShuffleWorker
Memoy
ShuffleWorker
Memoy
ShuffleWorker
huffle结果持久化
避免重算
通用Service
分布式（内存）文件系统

架构


Control message
Data Flow
JobMaster
ShuffleMaster
1. create shuffle session
2. start map worker
3. map commit
4. get shuffle info
5. update shuffler location
11. start reduce worker
Map Worker
Map Worker
Map Worker
Shuffler
Shuffler
Shuffler
6. map read HDFS
7. map push to shuffler
9. shuffler ack to map
8. shuffler write to hdfs
HDFS
HDFS
12. reduce read from HDFS
Reduce Worker
Reduce Worker
Reduce Worker
13. reduce write HDFS
HDFS

架构-基本概念

- ShuffleMaster
- Shuffler(Shuffle Worker)
- Writer
- Reader
- Session
- Shard

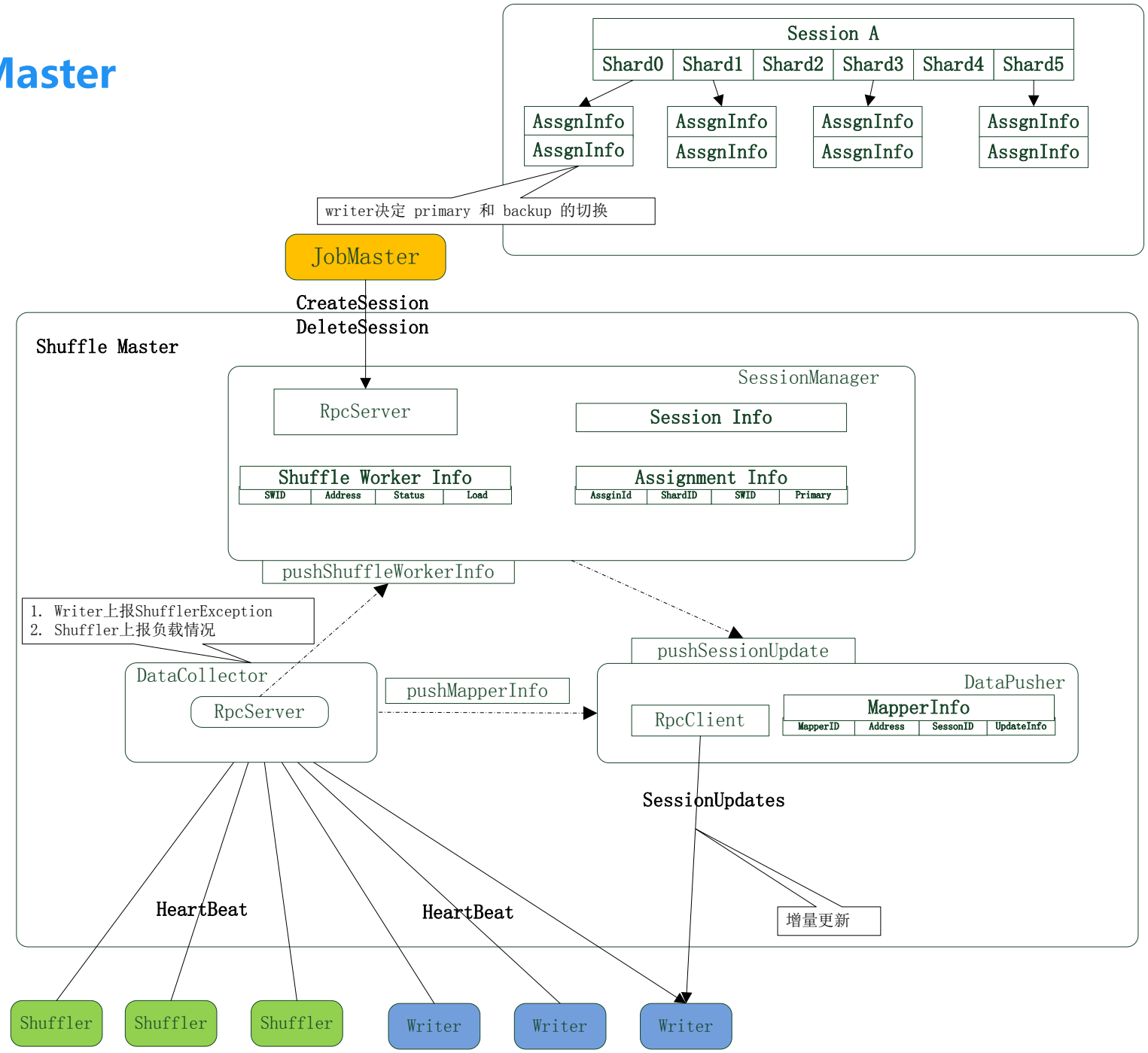
目录

百度云

- 背景
- 架构
- 关键技术
- 收益与总结
- 下一步计划

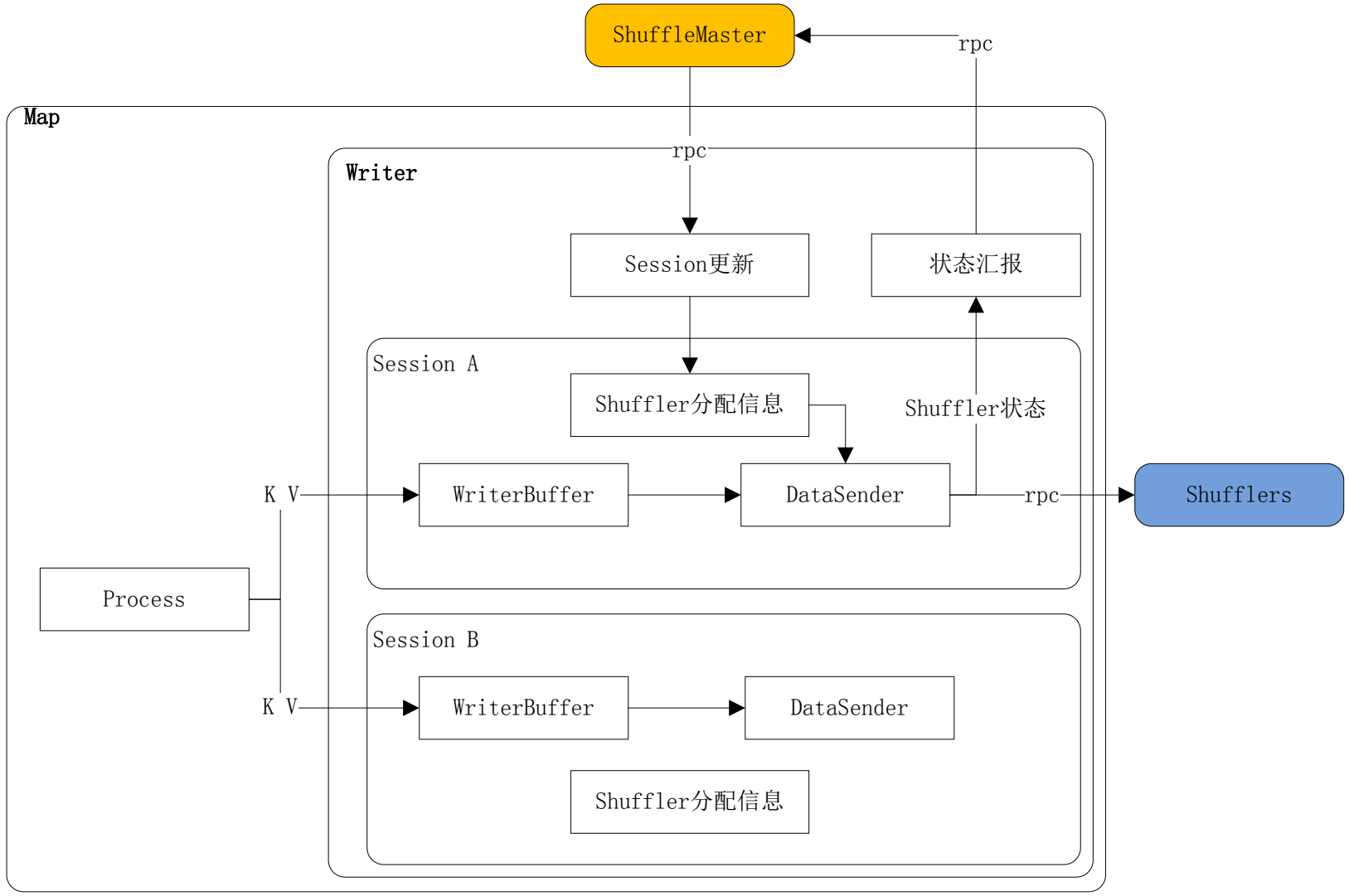
关键技术-ShuffleMaster

- 智能调度
  - 全局视图
  - 异常检测
  - 负载均衡
- 负载均衡
  - Shuffler
  - Shard



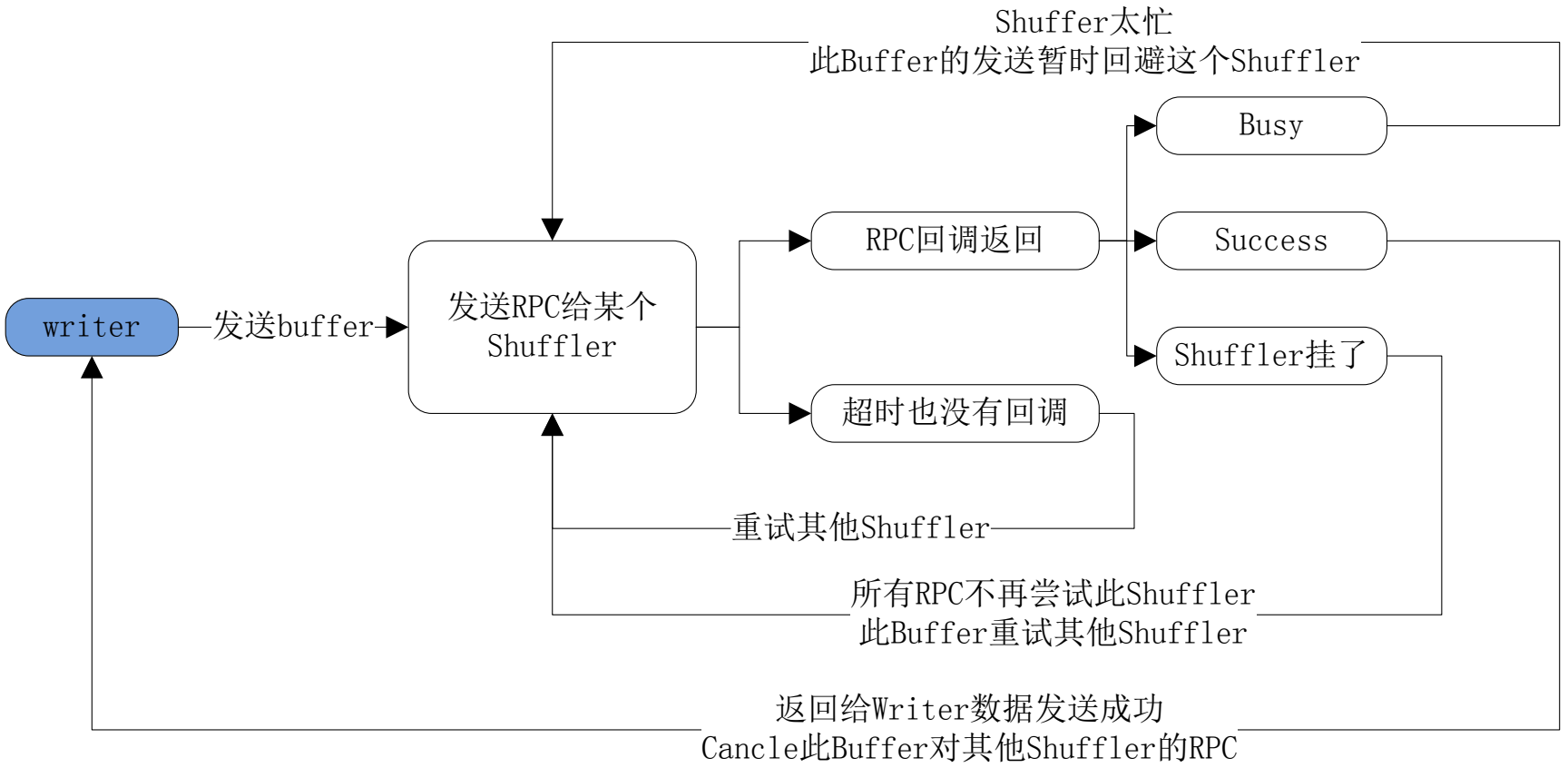
关键技术-Writer

- 数据缓存与异步发送
- 异常处理



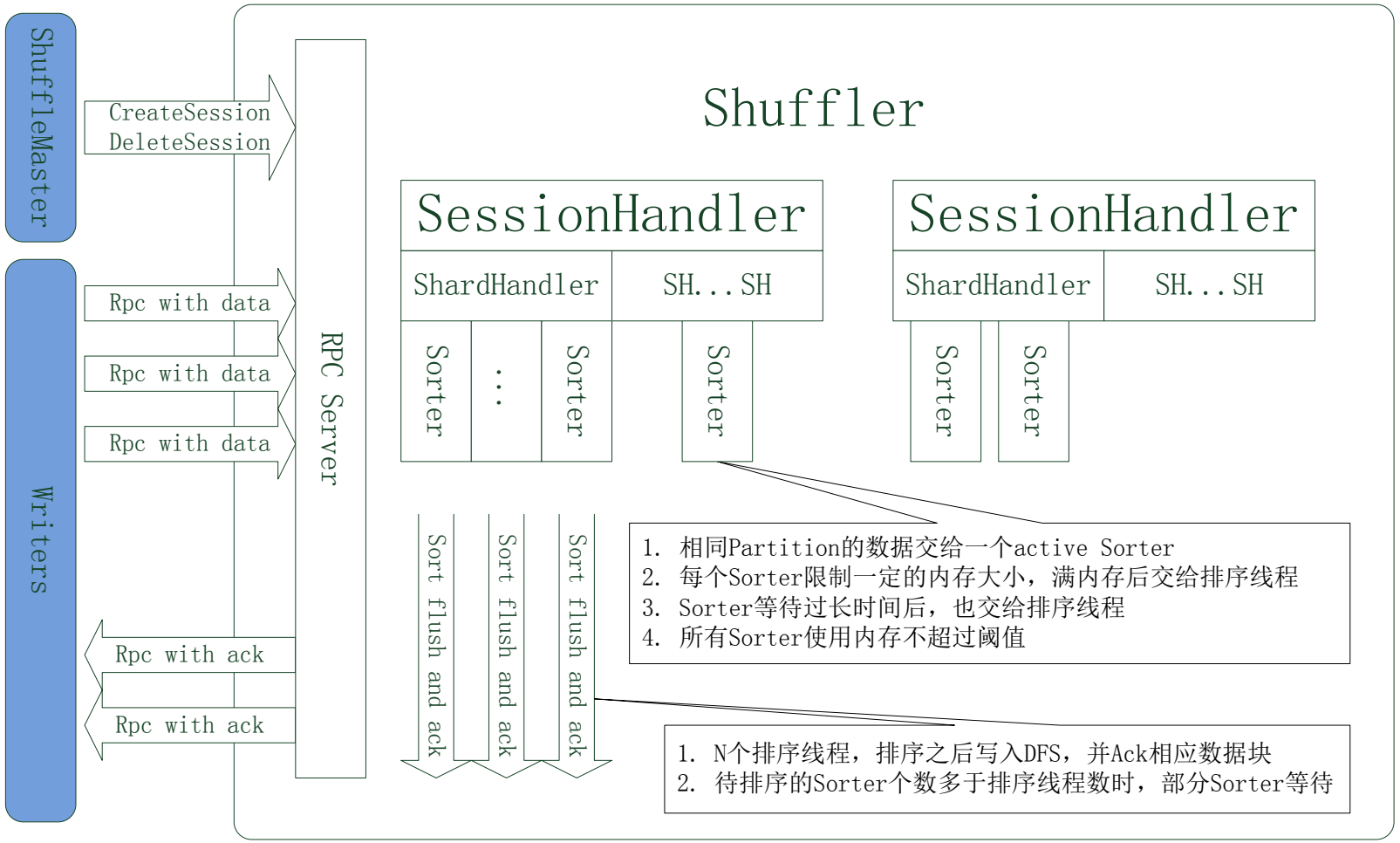
关键技术-Writer

- 异常处理



关键技术-Shuffler

- 内存聚合
- 流控
- Sort&Flush

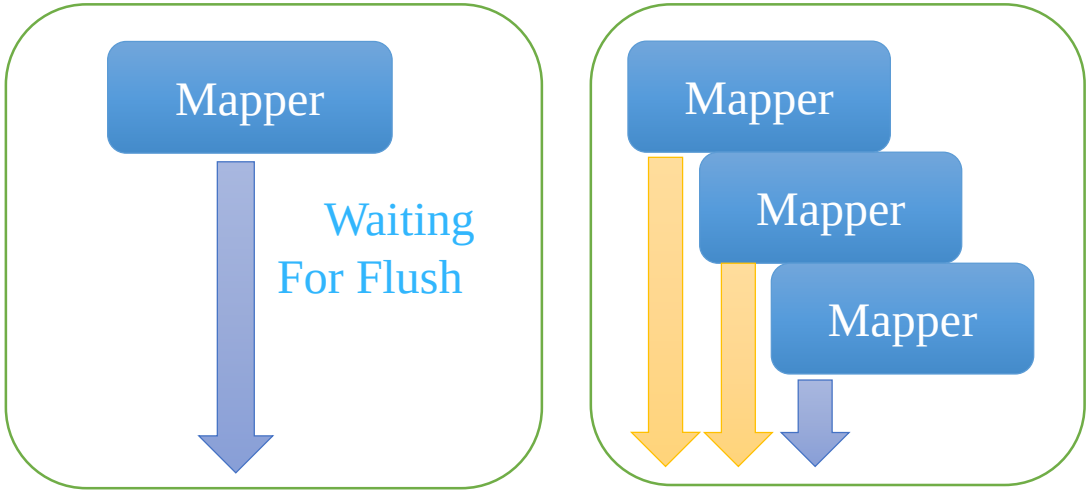


关键技术-Reader

- 所有Map完成后，调度Reduce
- 直接读取DFS排好序的数据
- 去重&数据验证
- 多路归并排序

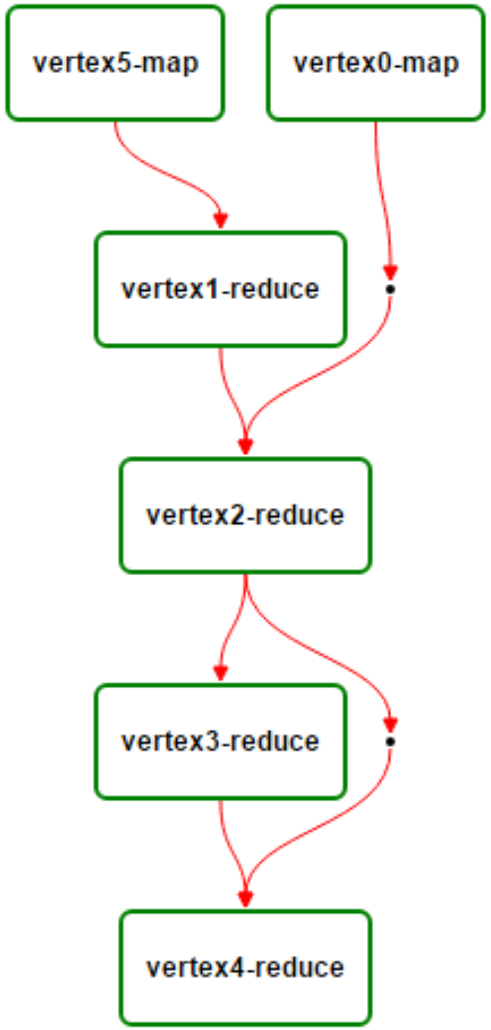
关键技术-Map Pipeline

- Map端Writer，要等Shuffler将数据持久化到DFS后，才能将发送的rpc buffer释放
- 所有发送的数据被Shuffler持久化后，Map才能安全退出



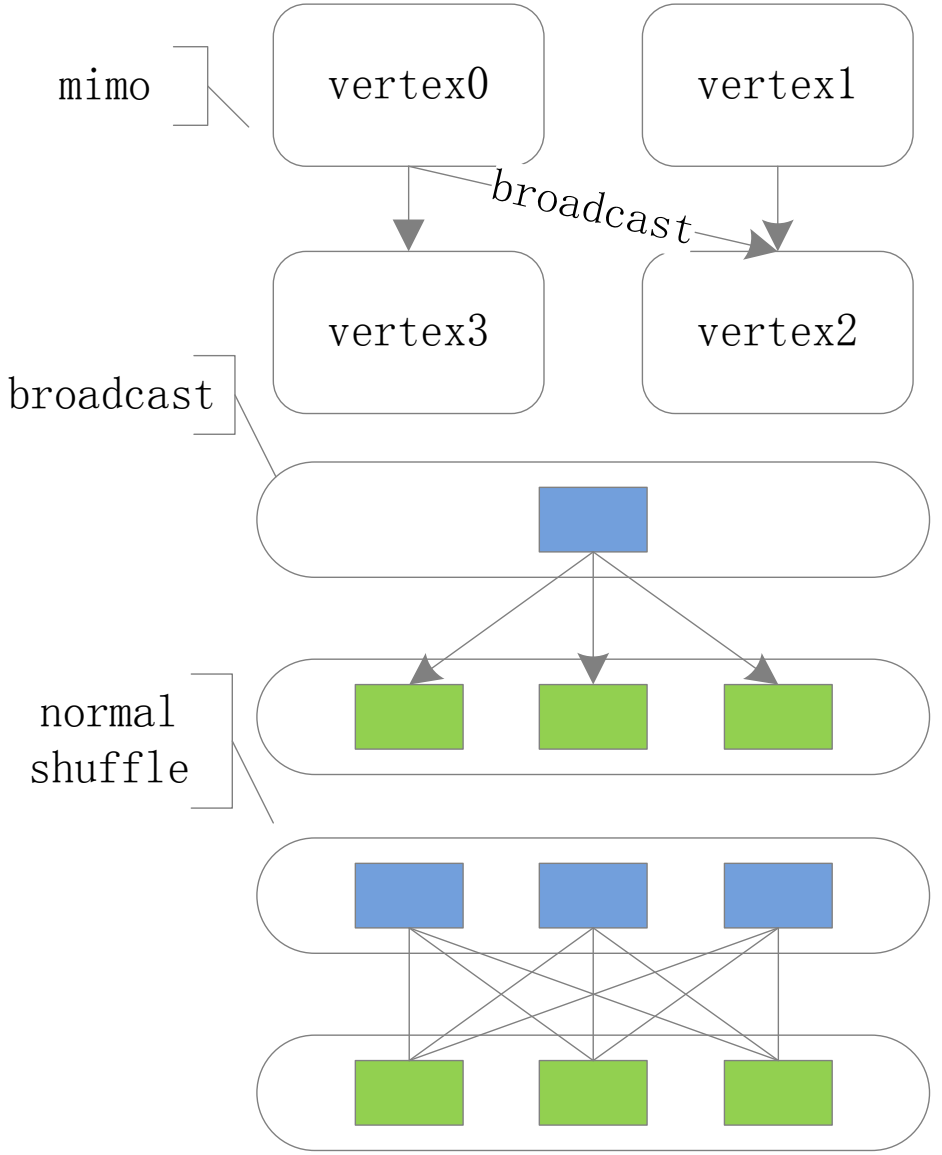
关键技术-MIMO

- 无MIMO时：Vetex3和vertex4收到vertex2的全部两路数据再做filter。此业务作业多shuffle近10T数据

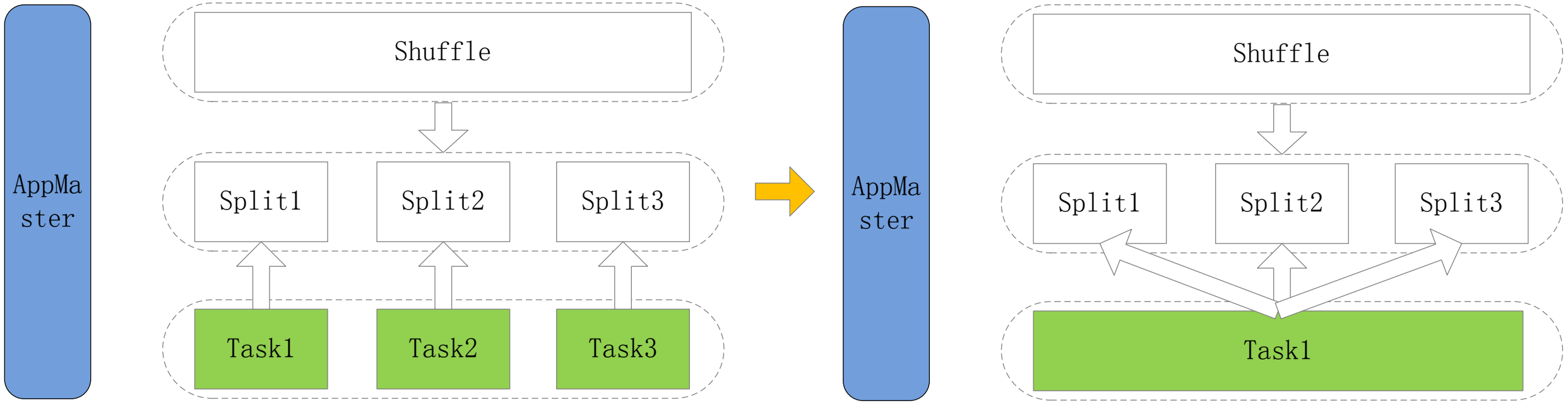


关键技术-MIMO

- 方案
  - 不同边可对应不同session
  - 不同session对应不同的dfs结果目录



关键技术-Runtime Partition



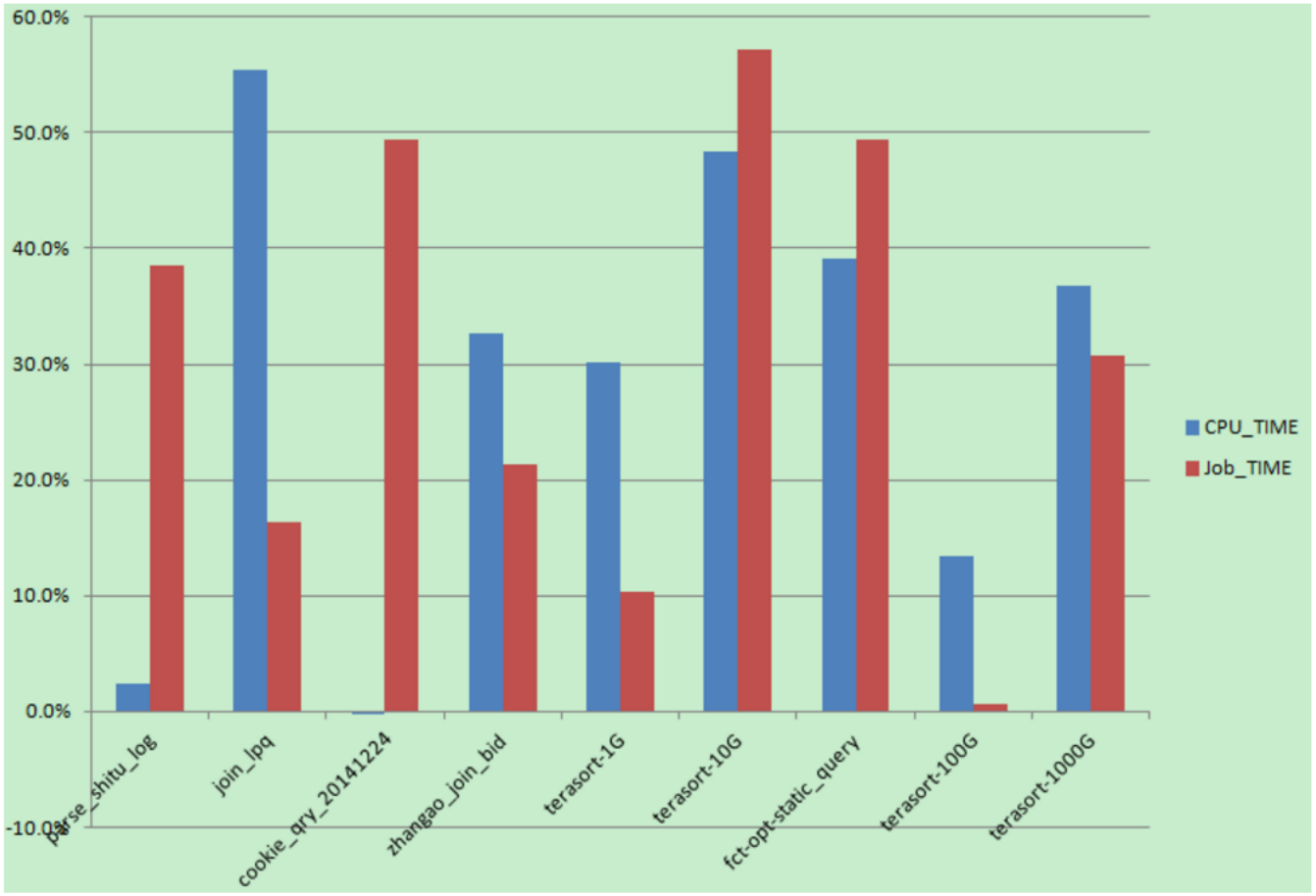
目录

百度云

- 背景
- 架构
- 关键技术
- 收益与总结
- 下一步计划

收益与总结

- 收益
  - 流式shuffle：减少旧shuffle map merge、reduce pull时间消耗
  - 内存Push：map端不落盘
  - Shuffler内存聚合：聚合度高，减少map端seek，减少reduce端merge路数，减少IO
  - Pipeline：大大提高中小作业map端运行速度
  - 中间数据持久化：避免重算（对dag作业尤为重要）



收益与总结

- 流式Shuffle服务
  - push instead of pull
  - Shuffle过程交给能拿到更多全局信息、更专业的shuffler来做
  - Map、Shuffler、Reduce，每个都做自己最适合做且擅长做的事情
  - Shuffle与Reduce解耦，简化Reduce，也为解决分桶不均问题提供可能
  - Pipeline。无缝的流水线，减少无谓的等待
- 问题
  - 更多的网络io（万兆网卡，网络不是瓶颈）
  - Shuffler资源共享，作业间可能互相影响（让Shuffler资源非瓶颈）

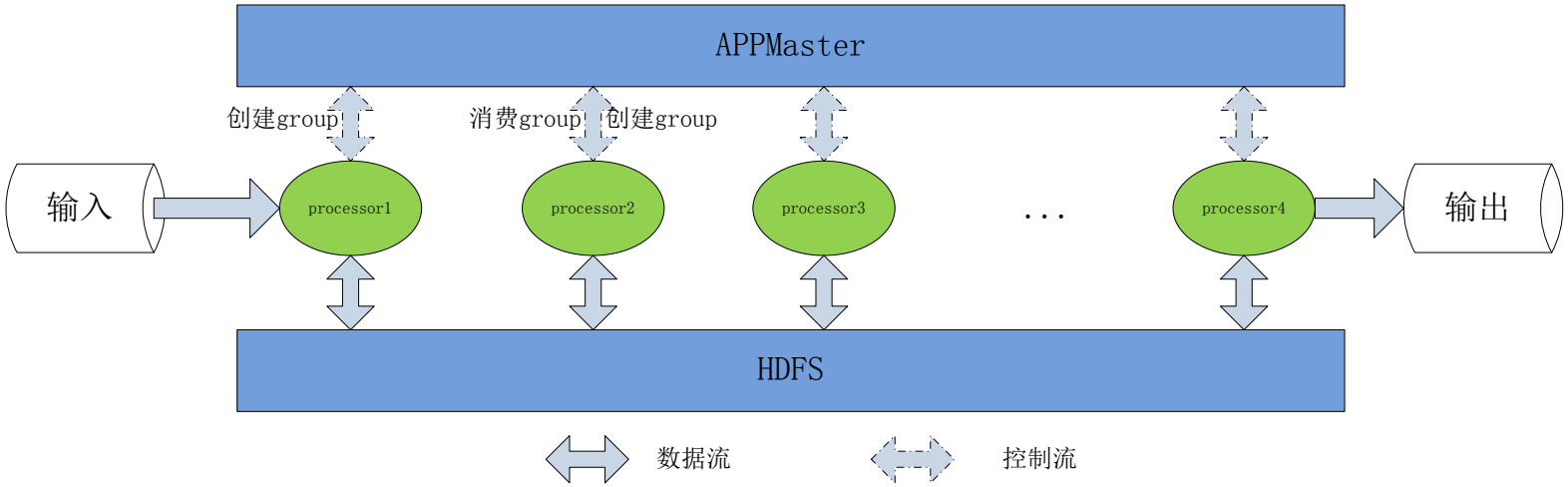
目录

百度云

- 背景
- 架构
- 关键技术
- 收益与总结
- 下一步计划

关键技术-下一步计划

- Hash Shuffle(not only sort-avoid shuffle)



- 动态调整DAG拓扑

THANK YOU

cloud.baidu.com