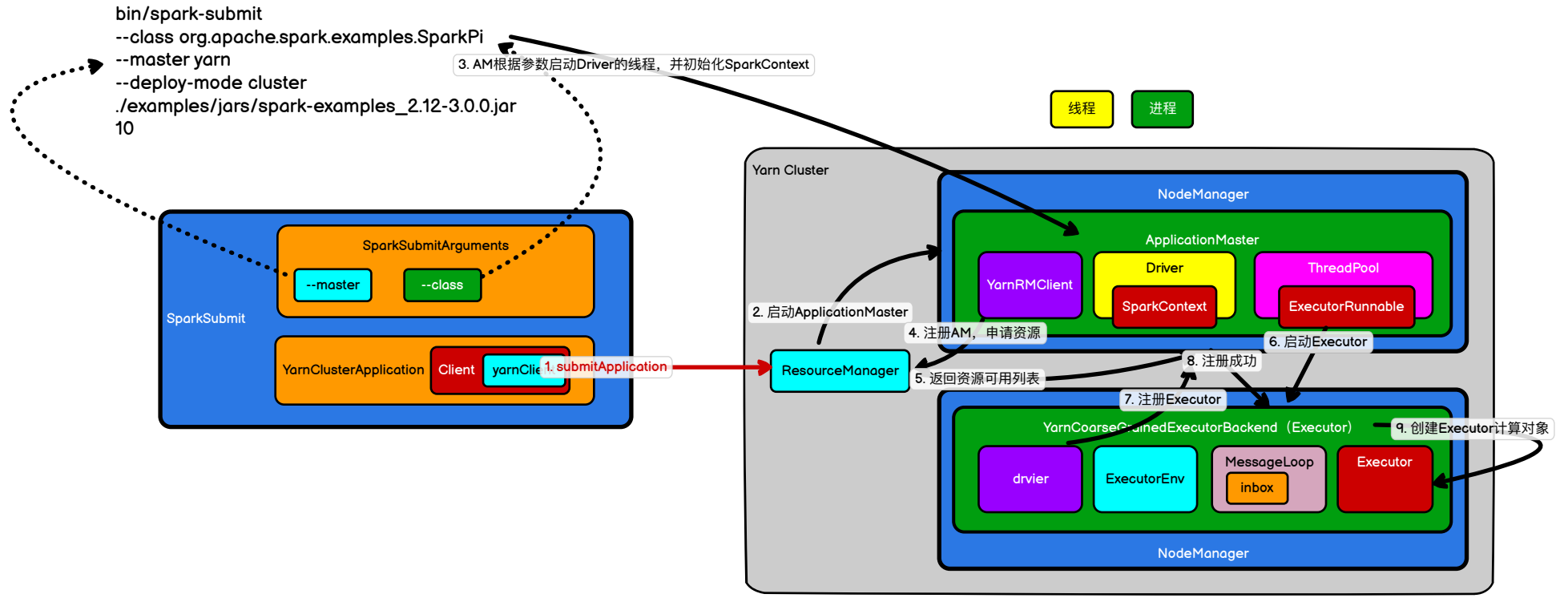
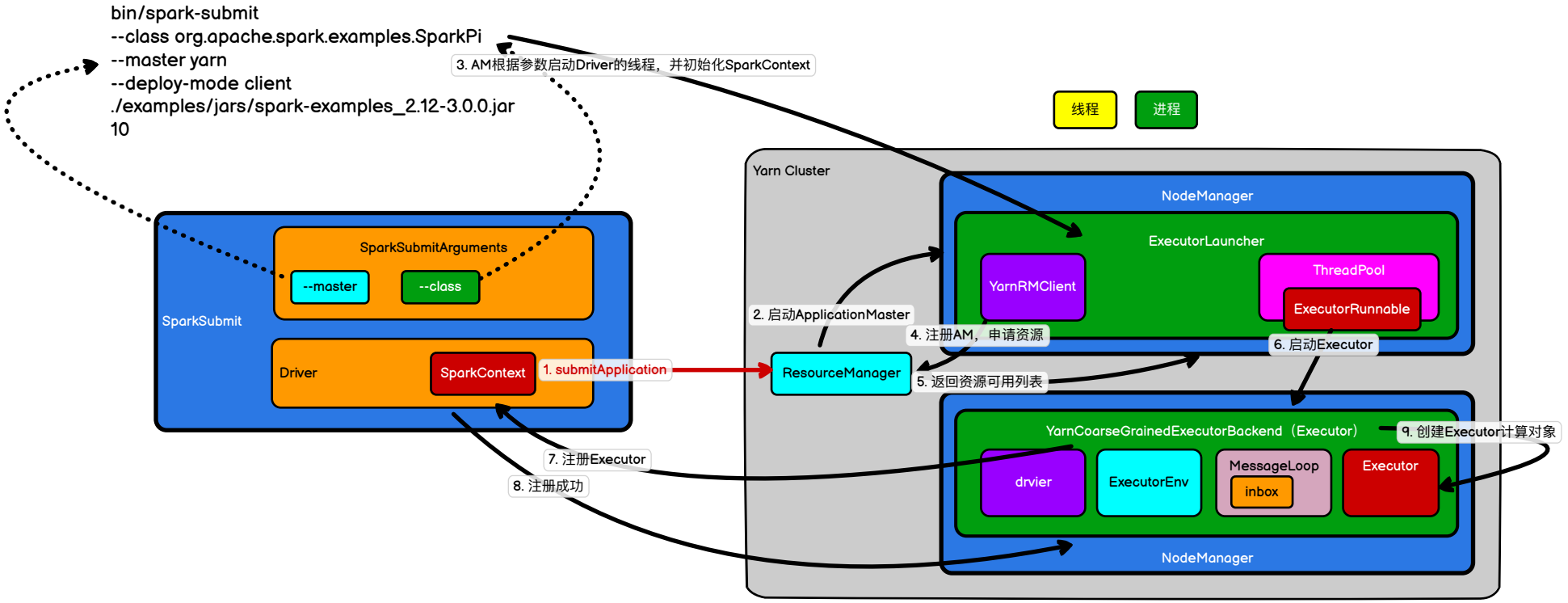

bin/spark-submit
--class org.apache.spark.examples.SparkPi
--master yarn
--deploy-mode cluster
./examples/jars/spark-examples_2.12-3.0.0.jar
10
3. AM根据参数启动Driver的线程，并初始化SparkContext
线程
进程
Yarn Cluster
NodeManager
ApplicationMaster
YarnRMClient
Driver
SparkContext
ThreadPool
ExecutorRunnable
SparkSubmit
SparkSubmitArguments
--master
--class
YarnClusterApplication
Client
yarnClie
1. submitApplication
ResourceManager
2. 启动ApplicationMaster
4. 注册AM，申请资源
5. 返回资源可用列表
6. 启动Executor
7. 注册Executor
8. 注册成功
YarnCoarseGrainedExecutorBackend（Executor）
9. 创建Executor计算对象
drvier
ExecutorEnv
MessageLoop
inbox
Executor
NodeManager


bin/spark-submit
--class org.apache.spark.examples.SparkPi
--master yarn
--deploy-mode client
./examples/jars/spark-examples_2.12-3.0.0.jar
10
3. AM根据参数启动Driver的线程，并初始化SparkContext
线程
进程
Yarn Cluster
NodeManager
ExecutorLauncher
YarnRMClient
ThreadPool
ExecutorRunnable
SparkSubmit
SparkSubmitArguments
--master
--class
Driver
SparkContext
1. submitApplication
ResourceManager
2. 启动ApplicationMaster
4. 注册AM，申请资源
5. 返回资源可用列表
6. 启动Executor
7. 注册Executor
8. 注册成功
YarnCoarseGrainedExecutorBackend（Executor）
9. 创建Executor计算对象
drvier
ExecutorEnv
MessageLoop
inbox
Executor
NodeManager

Netty： 通信框架，AIO

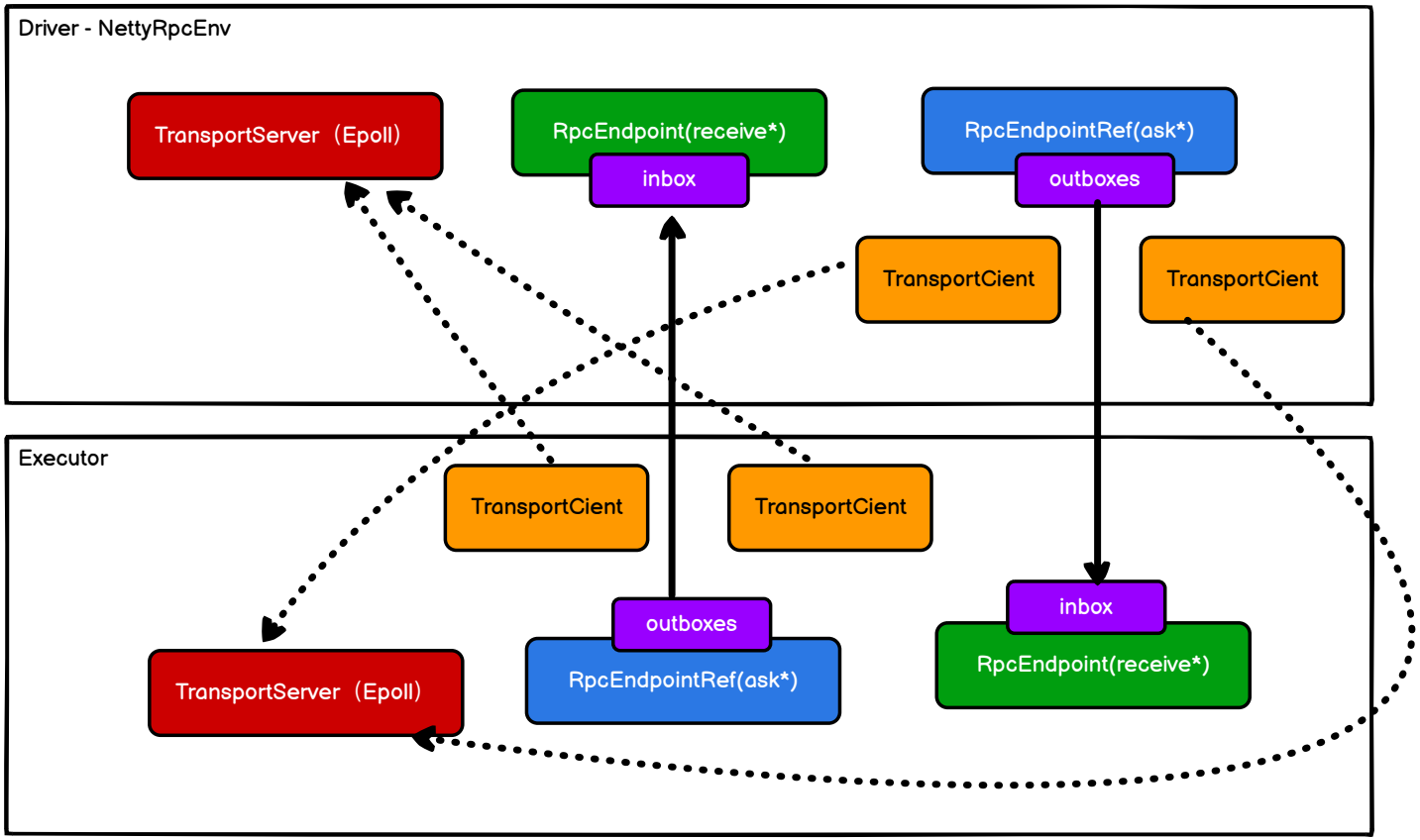
BIO： 阻塞式IO

NIO： 非阻塞式IO

AIO： 异步非阻塞式IO

Linux对AIO支持不够好，Windows支持好

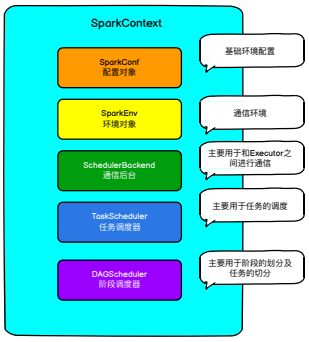
Linux采用Epoll方式模仿AIO操作

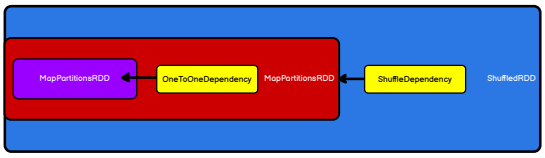


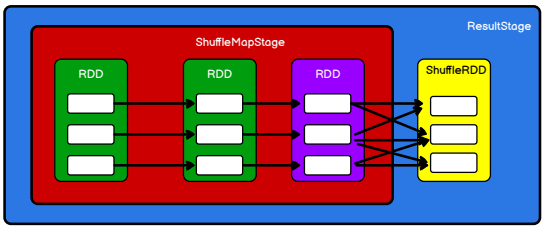
RPCEnv : 通信环境

Backend： 后台

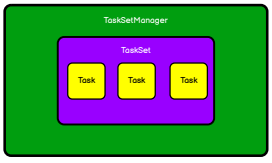
Endpoint： 终端

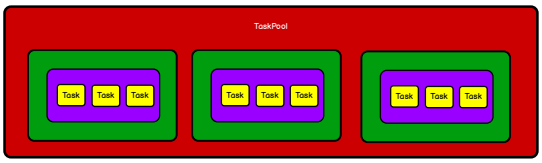


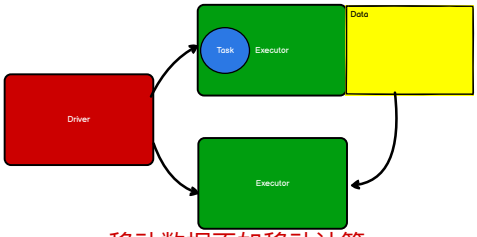




Spark中阶段的划分等于shuffle依赖的数量 + 1







移动数据不如移动计算

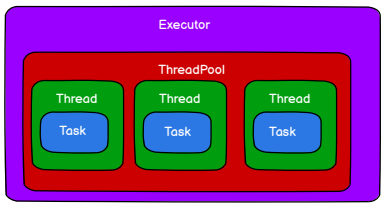
计算和数据的位置存在不同的级别，这个级别称之为本地化级别

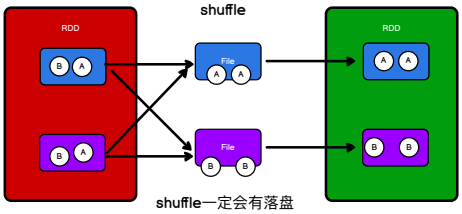
进程本地化：数据和计算在同一个进程中

节点本地化：数据和计算在同一个节点中

机架本地化：数据和计算在同一个机架中

任意

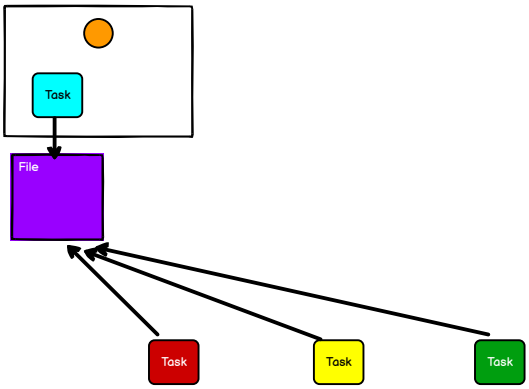


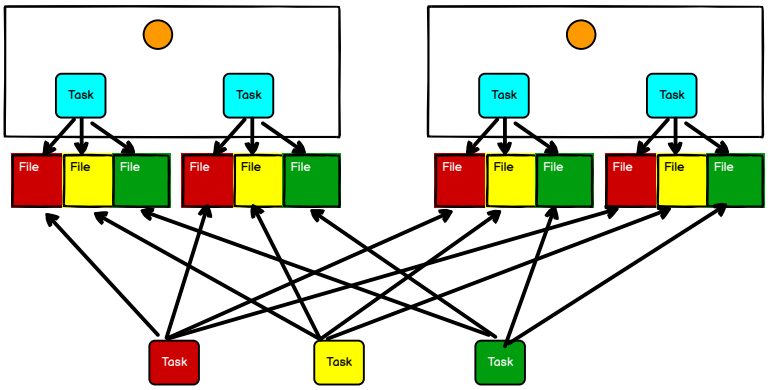


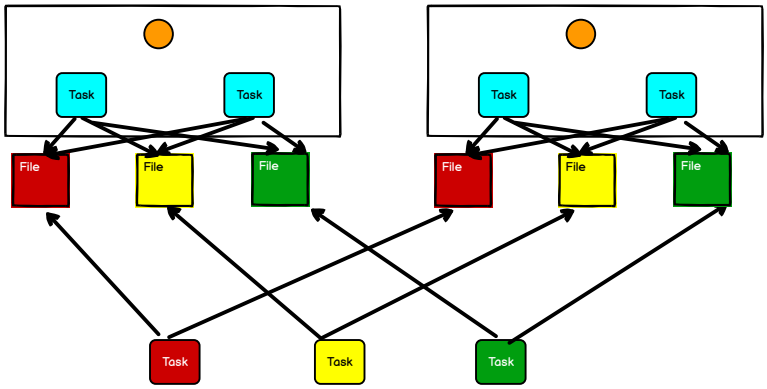
shuffle一定会有落盘

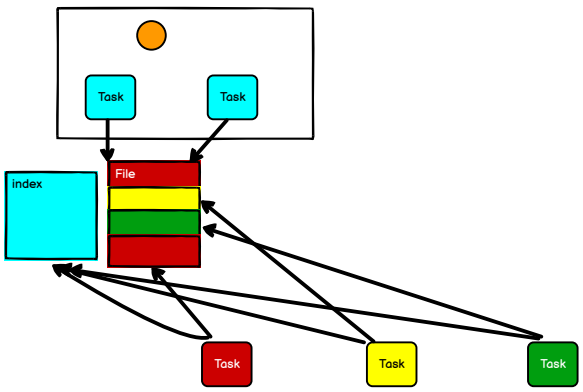
如果shuffle过程中落盘数据量减少，那么可以提高性能

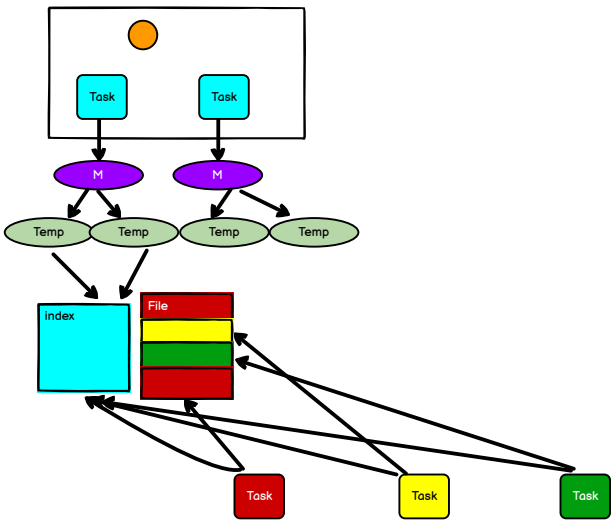
算子如果存在预聚合功能，可以提高shuffle的性能

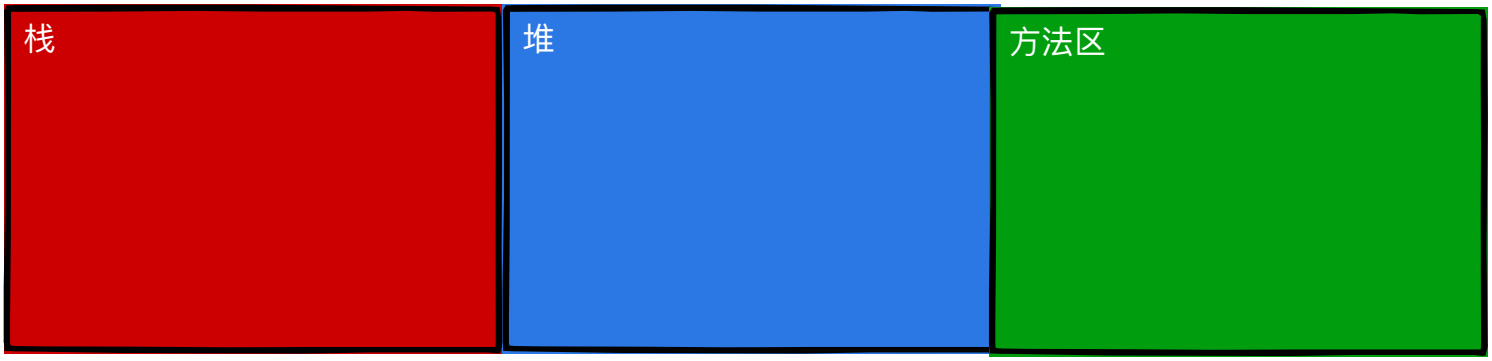

Task
File
Task
Task
Task

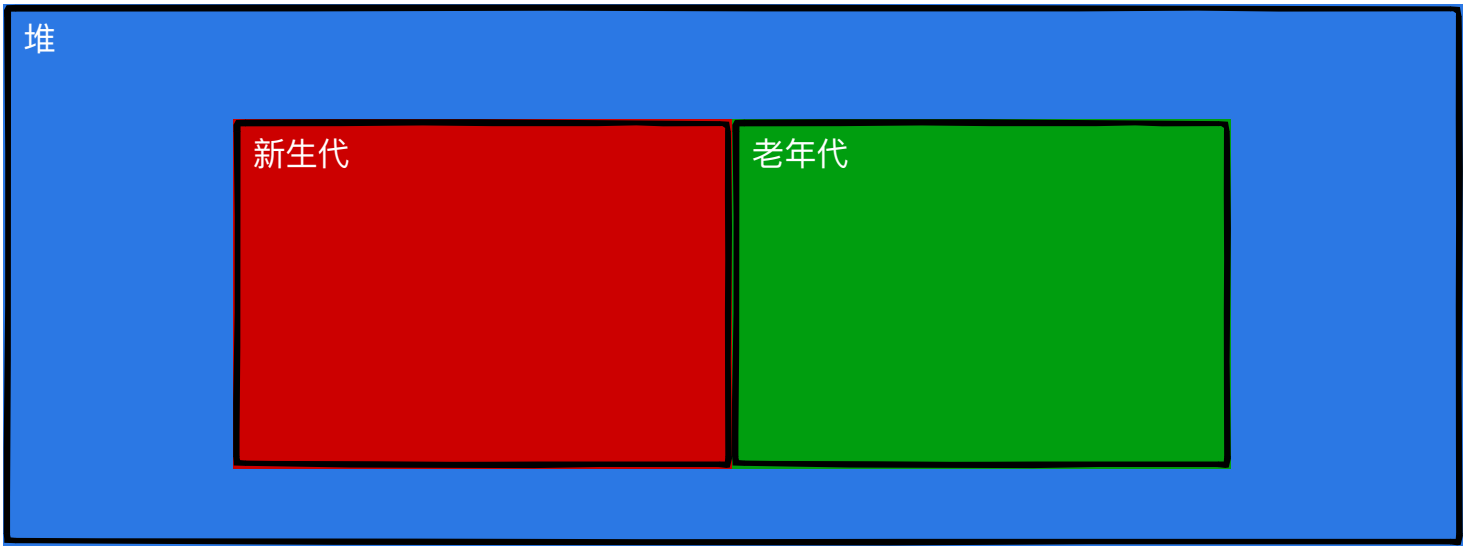

Task
Task
Task
Task
File
File
File
File
File
File
File
File
File
File
File
File
Task
Task
Task


Task
Task
Task
Task
File
File
File
File
File
File
Task
Task
Task

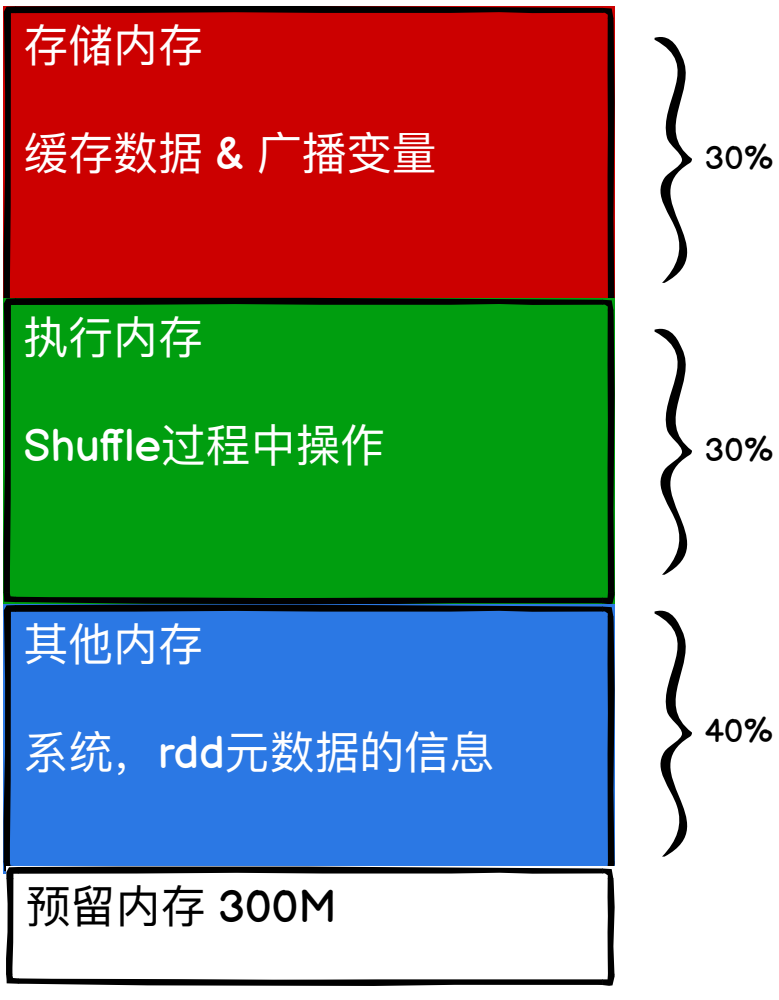

Task
Task
File
index
Task
Task
Task

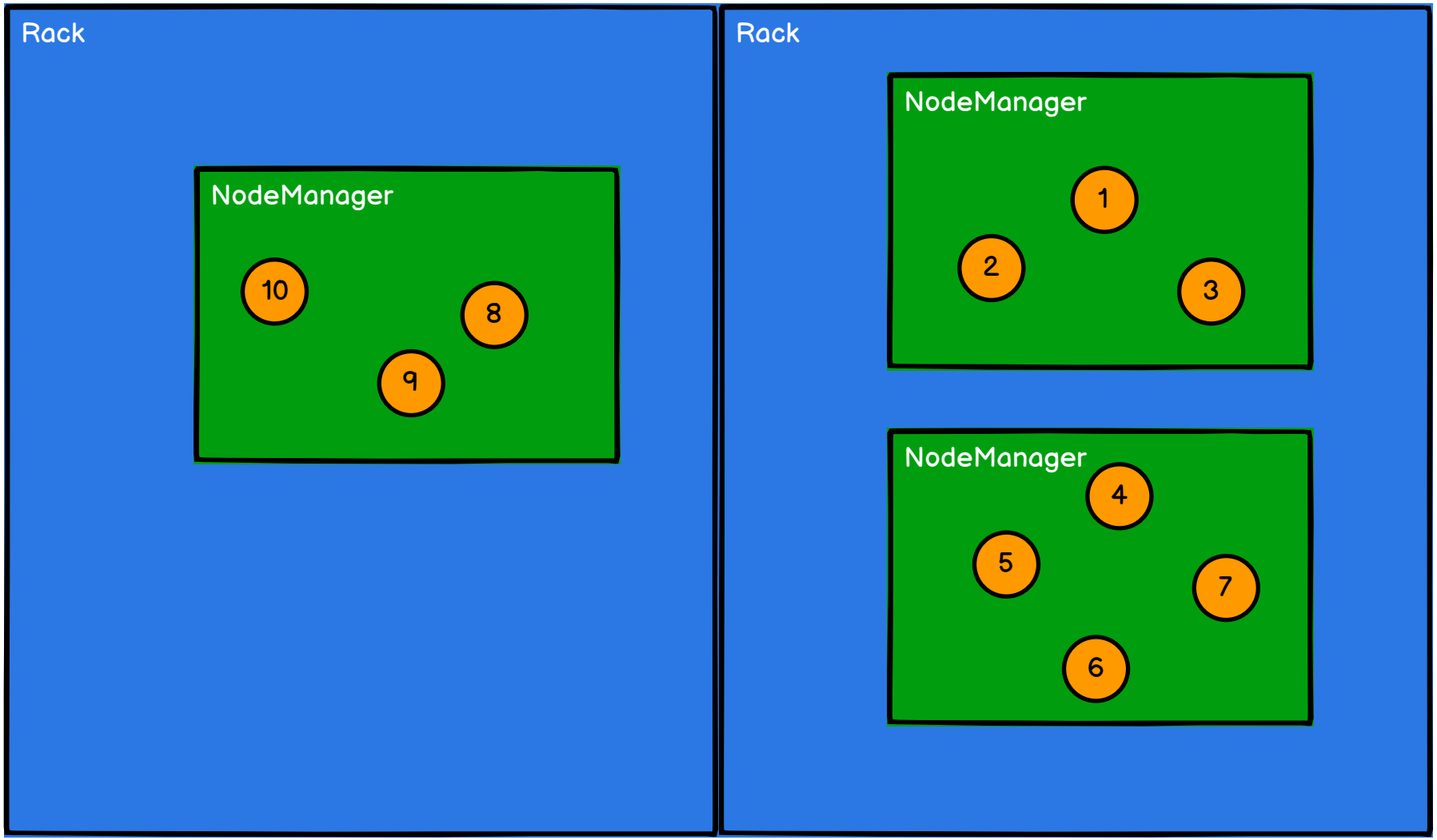

Task
Task
M
M
Temp
Temp
Temp
Temp
File
index
Task
Task
Task


栈
堆
方法区


堆
新生代
老年代

统一内存管理


存储内存
缓存数据 & 广播变量
30%
执行内存
Shuffle过程中操作
30%
其他内存
系统，rdd元数据的信息
40%
预留内存 300M


Rack
NodeManager
10
8
9
Rack
NodeManager
1
2
3
NodeManager
4
5
7
6

首选位置