一:任务调度与资源调度的区别

1. 任务调度是通过DAGScheduler、TaskSchedulerImpl、SchedulerBackend等进行的作业调度；
2. 资源调度是指应用程序如何获得资源；
3. 任务调度是在资源调度的基础上进行的；

二：资源调度

1. 因为Master负责资源调度和管理，所以资源调度的方法schedule方法位于Master.scala；
2. Schedule调用的时机：每次新的应用程序提交或者资源状态发生改变的时候（包括Executor增加或者减少，Worker增加或者减少等。）
3. 当前Master的状态是Alive的方式才进行资源的调度，如果不是Alive的状态会直接返回，也就是standby Master不会进行Application的资源调度

**if** (*state* != RecoveryState.*ALIVE*) { **return** }

1. 使用Random.shuffle把Master中保留的集群中所有Worker的信息随机打乱

**val** shuffledWorkers = Random.shuffle(*workers*) // Randomization helps balance drivers

其算法内部是循环随机交换所有Worker在Master缓存数据结构中的位置；

1. 接下来要判断所有Worker中哪些是Alive级别的Worker，Alive才能够参与资源的分配工作

**for** (worker <- shuffledWorkers **if** worker.*state* == WorkerState.*ALIVE*) {

1. 当SparkSubmit指定Driver在Cluster模式的情况下，此时Driver会加入waitingDrivers等待列表中，在每个DriverInfo的DriverDescription中有要启动Driver时候对Worker的内存及cores的要求的内容。在符合资源要求的情况下，然后采用随机打乱后的一个worker来启动Dirver，Master发指令给Worker，让远程的Worker启动Dirver。

**for** (driver <- *waitingDrivers*) {  
 **if** (worker.memoryFree >= driver.desc.mem && worker.coresFree >= driver.desc.cores) {  
 launchDriver(worker, driver)  
 *waitingDrivers* -= driver  
 }  
}

**private def** launchDriver(worker: WorkerInfo, driver: DriverInfo) {  
 logInfo("Launching driver " + driver.id + " on worker " + worker.id)  
 worker.addDriver(driver)  
 driver.*worker* = *Some*(worker)  
 worker.endpoint.send(*LaunchDriver*(driver.id, driver.desc))  
 driver.*state* = DriverState.*RUNNING*}

1. 先启动Drivers才会发生后续的一切的资源调度的模式
2. Spark默认为应用程序启动Executor的方式是FIFO的方式，也就是所有提交的应用程序都是放在调度的等待队列中，先进先出，只有满足了前面应用程序的资源分配的基础上才能够满足下一个应用程序资源的分配；
3. 为应用程序具体分配Executor之前要判断应用程序时候还需要分配core，如果不需要则不会为应用程序分配Executor；
4. 具体分配Executor之前要对要求Worker必须是Alive的状态且必须满足Application对每个Executor的内存和cores的要求，并且在此基础上进行排序，产生计算资源由大到小的usableWorkers数据结构，在FIFO的情况下默认是SpreadOutApps来让应用程序尽可能多的运行在所有的Node上。

**val** usableWorkers = *workers*.toArray.filter(\_.*state* == WorkerState.*ALIVE*)  
 .filter(worker => worker.memoryFree >= app.desc.memoryPerExecutorMB &&  
 worker.coresFree >= coresPerExecutor.getOrElse(1))  
 .sortBy(\_.coresFree).reverse  
**val** assignedCores = scheduleExecutorsOnWorkers(app, usableWorkers, *spreadOutApps*)

1. 为应用程序分配Executors有两种方式，第一种方式是尽可能在集群的所有worker上分配Executor，这种方式往往会带来潜在的更好的数据本地性。
2. 具体在集群上分配Cores的时候会尽可能的满足我们的要求

**val** enoughCores = usableWorkers(pos).coresFree - assignedCores(pos) >= minCoresPerExecutor

1. 准备具体要为当前应用程序分配的Executor信息后，Master要通过进程通讯发指令给Worker来具体启动ExecutorBackend进程

worker.endpoint.send(*LaunchExecutor*(*masterUrl*,  
 exec.application.id, exec.id, exec.application.desc, exec.cores, exec.memory)

1. 紧接着给我们应用程序的Driver发送一个ExecutorAdded的信息

exec.application.driver.send(  
 *ExecutorAdded*(exec.id, worker.id, worker.hostPort, exec.cores, exec.memory))