SparkExcutorn内存管理

首先我们知道在执行 Spark 的应用程序时，Spark 集群会启动 Driver 和 Executor 两种 JVM 进程，前者为主控进程，负责创建 Spark 上下文，提交 Spark 作业（Job），并将作业转化为计算任务（Task），在各个 Executor 进程间协调任务的调度，后者负责在工作节点上执行具体的计算任务，并将结果返回给 Driver，同时为需要持久化的 RDD 提供存储功能。由于 Driver 的内存管理相对来说较为简单，本文主要对 Executor 的内存管理进行分析，下文中的 Spark 内存均特指 Executor 的内存。

另外，Spark 1.6 之前使用的是静态内存管理 (StaticMemoryManager) 机制，StaticMemoryManager 也是 Spark 1.6 之前唯一的内存管理器。在 Spark1.6 之后引入了**统一内存管理 (UnifiedMemoryManager) 机制**，UnifiedMemoryManager 是 Spark 1.6 之后默认的内存管理器，1.6 之前采用的静态管理（StaticMemoryManager）方式仍被保留，可通过配置 spark.memory.useLegacyMode 参数启用。这里仅对统一内存管理模块 (UnifiedMemoryManager) 机制进行分析。

> ### Executor内存总体布局

默认情况下，Executor不开启堆外内存，因此整个 Executor 端内存布局如下图所示:

我们可以看到在Yarn集群管理模式中，Spark 以 Executor Container 的形式在 NodeManager 中运行，其可使用的内存上限由 yarn.scheduler.maximum-allocation-mb 指定，我们称之为 MonitorMemory。

整个Executor内存区域分为两块：

1、**JVM堆外内存**

大小由**spark.yarn.executor.memoryOverhead**参数指定。默认大小为 executorMemory * 0.10, with minimum of 384m。

此部分内存主要用于JVM自身，字符串, NIO Buffer（Driect Buffer）等开销。此部分为用户代码及 Spark 不可操作的内存，不足时可通过调整参数解决。

2、**堆内内存（ExecutorMemory）**

大小由 Spark 应用程序启动时的 **--executor-memory/spark.executor.memory** 参数配置，即JVM 最大分配的堆内存 (-Xmx)。Spark为了更高效的使用这部分内存，对这部分内存进行了逻辑上的划分管理。我们在下面的统一内存管理会详细介绍。

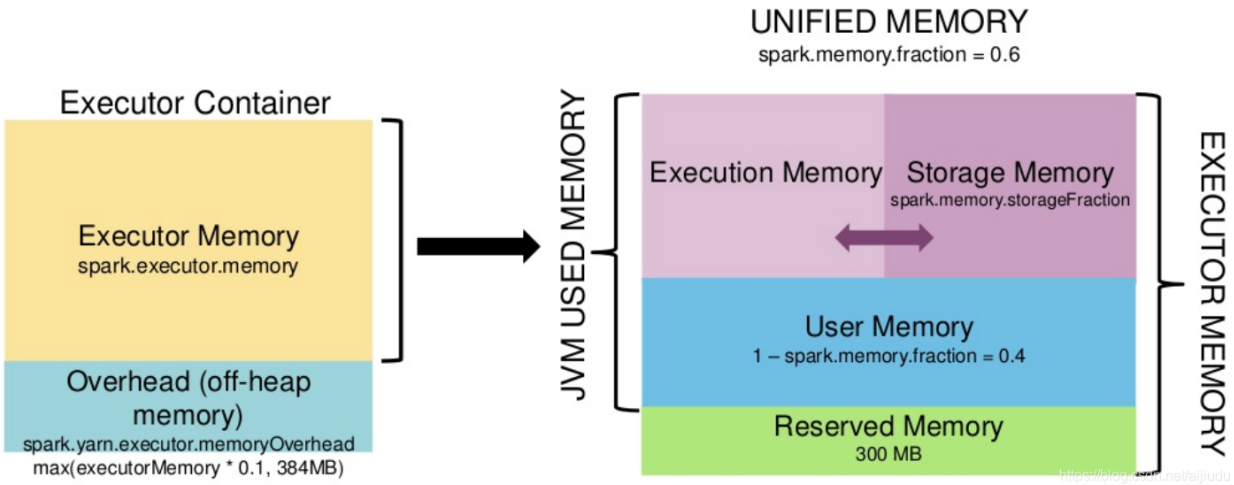
```

对于Yarn集群，存在: ExecutorMemory + MemoryOverhead <= MonitorMemory，若应用提交之时，指定的 ExecutorMemory 与 MemoryOverhead 之和大于 MonitorMemory，则会导致 Executor 申请失败；若运行过程中，实际使用内存超过上限阈值，Executor 进程会被 Yarn 终止掉 (kill)。

```

> ## 统一内存管理

Spark 1.6之后引入了统一内存管理，包括了**堆内内存 (On-heap Memory)** 和**堆外内存 (Off-heap Memory)** 两大区域，下面对这两块区域进行详细的说明。

> ### 堆内内存 (On-heap Memory)

默认情况下，Spark 仅仅使用了堆内内存。Spark 对堆内内存的管理是一种逻辑上的“规划式”的管理，Executor 端的堆内内存区域在逻辑上被划分为以下四个区域:

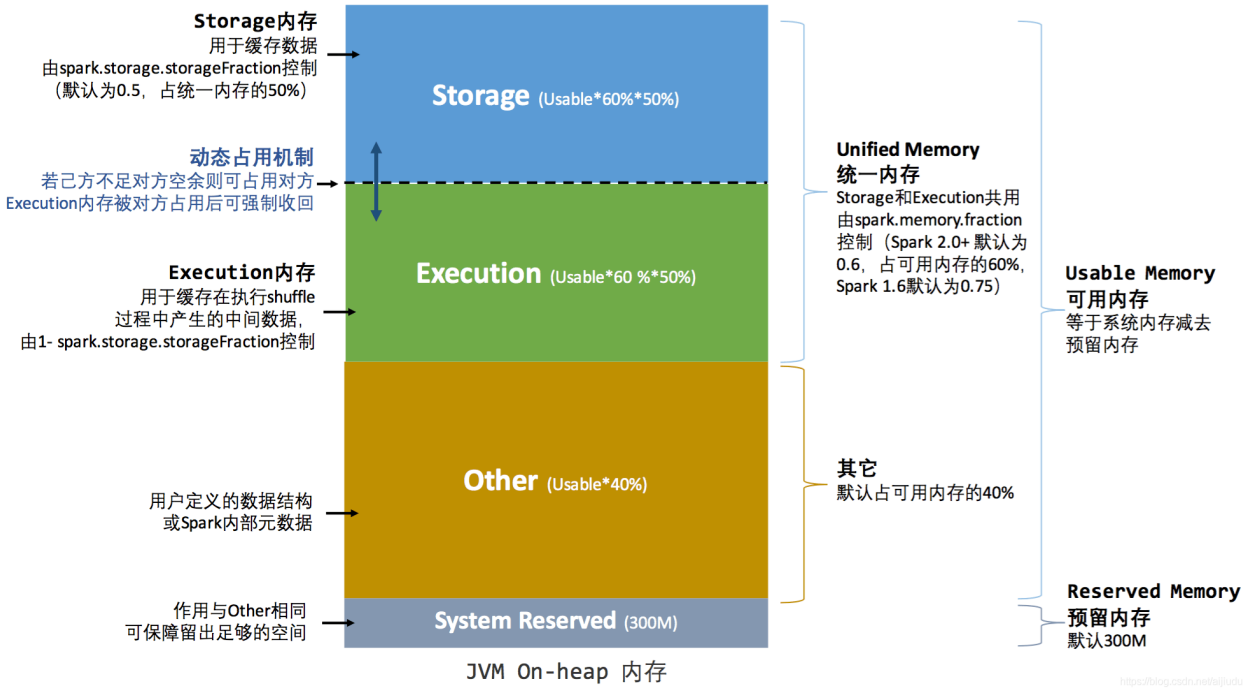


\1. 执行内存 (Execution Memory) : 主要用于存放 Shuffle、Join、Sort、Aggregation 等计算过程中的临时数据；

\2. 存储内存 (Storage Memory) : 主要用于存储 spark 的 cache 数据，例如RDD的缓存、unroll数据；

\3. 用户内存（User Memory）: 主要用于存储 RDD 转换操作所需要的数据，例如 RDD 依赖等信息；

\4. 预留内存（Reserved Memory）: 系统预留内存，会用来存储Spark内部对象。



\1. 预留内存 (Reserved Memory)

系统预留内存，会用来存储Spark内部对象。其大小在代码中是写死的，其值等于 300MB，这个值是不能修改的（如果在测试环境下，我们可以通过 spark.testing.reservedMemory 参数进行修改）；如果Executor分配的内存小于 1.5 * 300 = 450M 时，Executor将无法执行。

\2. 存储内存 (Storage Memory)

主要用于存储 spark 的 cache 数据，例如 RDD 的缓存、广播（Broadcast）数据、和 unroll 数据。内存占比为 UsableMemory * spark.memory.fraction * spark.memory.storageFraction，Spark 2+ 中，默认初始状态下 Storage Memory 和 Execution Memory 均约占系统总内存的30%（1 * 0.6 * 0.5 = 0.3）。在 UnifiedMemory 管理中，这两部分内存可以相互借用，具体借用机制我们下一小节会详细介绍。

\3. 执行内存 (Execution Memory)

主要用于存放 Shuffle、Join、Sort、Aggregation 等计算过程中的临时数据。内存占比为 UsableMemory * spark.memory.fraction * (1 - spark.memory.storageFraction)，Spark 2+ 中，默认初始状态下 Storage Memory 和 Execution Memory 均约占系统总内存的30%（1 * 0.6 * (1 - 0.5) = 0.3）。在 UnifiedMemory 管理中，这两部分内存可以相互借用，具体借用机制我们下一小节会详细介绍。

\4. 其他/用户内存 (Other/User Memory) : 主要用于存储 RDD 转换操作所需要的数据，例如 RDD 依赖等信息。内存占比为 UsableMemory * (1 - spark.memory.fraction)，在Spark2+ 中，默认占可用内存的40%（1 * (1 - 0.6) = 0.4）。

其中，usableMemory = executorMemory - reservedMemory，这个就是 Spark 可用内存。

NOTES

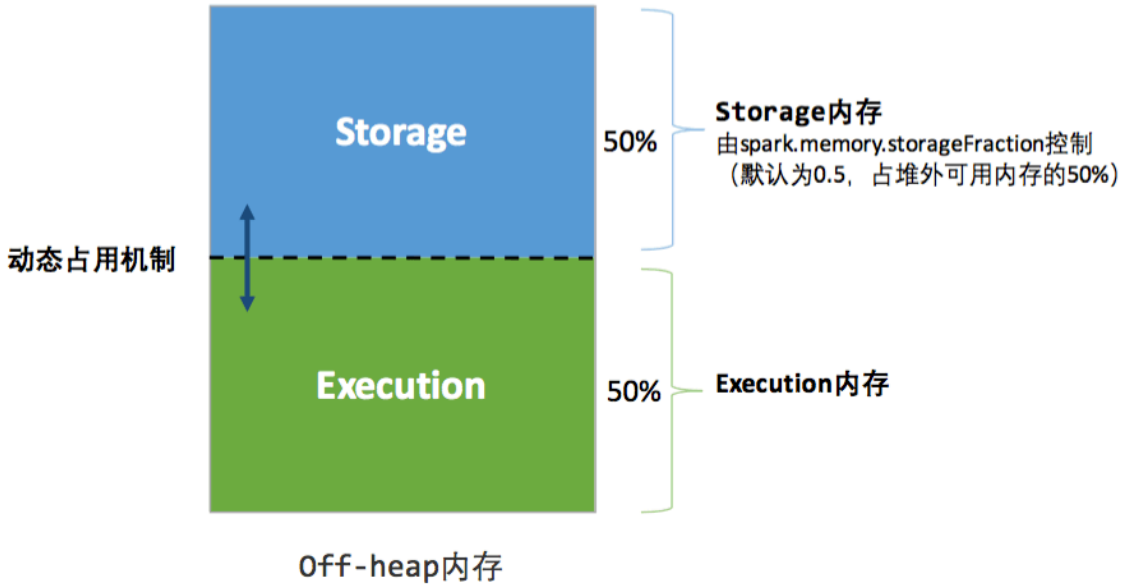
```

1、为什么设置300M预留内存

统一内存管理最初版本other这部分内存没有固定值 300M 设置，而是和静态内存管理相似，设置的百分比，最初版本占 25%。百分比设置在实际使用中出现了问题，若给定的内存较低时，例如 1G，会导致 OOM，具体讨论参考这里 Make unified memory management work with small heaps。因此，other这部分内存做了修改，先划出 300M 内存。

2、spark.memory.fraction 由 0.75 降至 0.6

spark.memory.fraction 最初版本的值是 0.75，很多分析统一内存管理这块的文章也是这么介绍的，同样的，在使用中发现这个值设置的偏高，导致了 gc 时间过长，spark 2.0 版本将其调整为 0.6，详细谈论参见 Reduce spark.memory.fraction default to avoid overrunning old gen in JVM default config。

```

> ### 堆外内存 (Off-heap Memory)

Spark 1.6 开始引入了 Off-heap memory (详见SPARK-11389)。这种模式不在 JVM 内申请内存，而是调用 Java 的 unsafe 相关 API 进行诸如 C 语言里面的 malloc() 直接向操作系统申请内存。这种方式下 Spark 可以直接操作系统堆外内存，减少了不必要的内存开销，以及频繁的 GC 扫描和回收，提升了处理性能。另外，堆外内存可以被精确地申请和释放，而且序列化的数据占用的空间可以被精确计算，所以相比堆内内存来说降低了管理的难度，也降低了误差。，缺点是必须自己编写内存申请和释放的逻辑。

默认情况下Off-heap模式的内存并不启用，我们可以通过 spark.memory.offHeap.enabled 参数开启，并由 spark.memory.offHeap.size 指定堆外内存的大小，单位是字节（占用的空间划归 JVM OffHeap 内存）。

如果堆外内存被启用，那么 Executor 内将同时存在堆内和堆外内存，两者的使用互补影响，这个时候 Executor 中的 Execution 内存是堆内的 Execution 内存和堆外的 Execution 内存之和，同理，Storage 内存也一样。其内存分布如下图所示：



相比堆内内存，堆外内存只区分 Execution 内存和 Storage 内存：

\1. 存储内存 (Storage Memory)

内存占比为 maxOffHeapMemory * spark.memory.storageFraction，Spark 2+ 中，默认初始状态下 Storage Memory 和 Execution Memory 均约占系统总内存的50%（1 * 0.5 = 0.5）。在 UnifiedMemory 管理中，这两部分内存可以相互借用，具体借用机制我们下一小节会详细介绍。

\2. 执行内存 (Execution Memory)

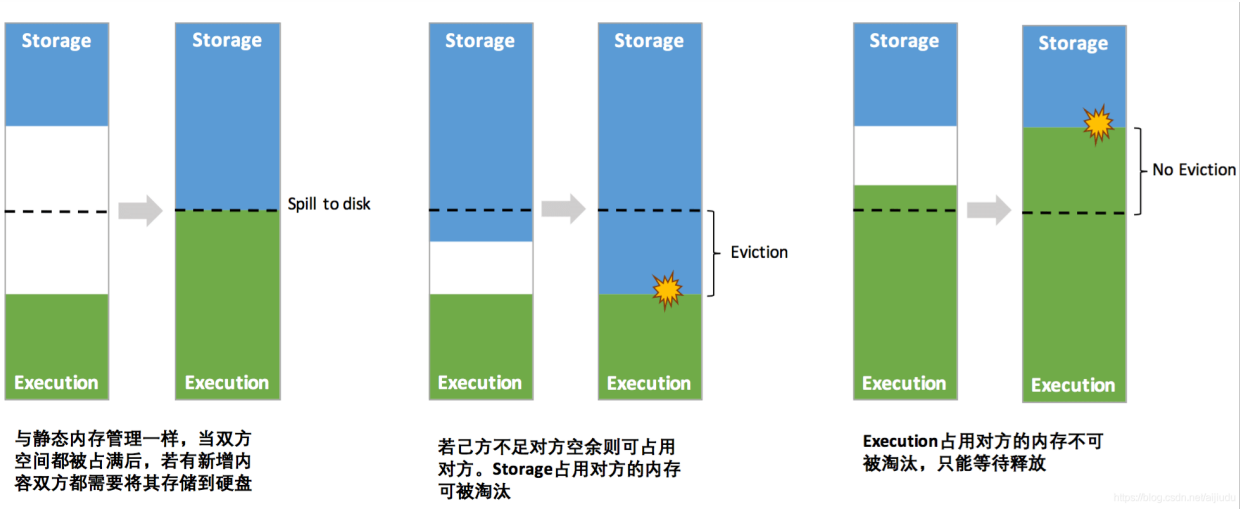
内存占比为 maxOffHeapMemory * (1 - spark.memory.storageFraction)，Spark 2+ 中，默认初始状态下 Storage Memory 和 Execution Memory 均约占系统总内存的50%（1 * (1 - 0.5) = 0.5）。在 UnifiedMemory 管理中，这两部分内存可以相互借用，具体借用机制我们下一小节会详细介绍。

> ### Execution 内存和 Storage 内存动态占用机制

在 Spark 1.5 之前，Execution 内存和 Storage 内存分配是静态的，换句话说就是如果 Execution 内存不足，即使 Storage 内存有很大空闲程序也是无法利用到的；反之亦然。

静态内存管理机制实现起来较为简单，但如果用户不熟悉 Spark 的存储机制，或没有根据具体的数据规模和计算任务或做相应的配置，很容易造成”一半海水，一半火焰”的局面，即存储内存和执行内存中的一方剩余大量的空间，而另一方却早早被占满，不得不淘汰或移出旧的内容以存储新的内容。

统一内存管理机制，与静态内存管理最大的区别在于存储内存和执行内存共享同一块空间，可以动态占用对方的空闲区域：



其中最重要的优化在于动态占用机制，其规则如下：

- 程序提交的时候我们都会设定基本的 Execution 内存和 Storage 内存区域（通过 spark.memory.storageFraction 参数设置）。我们用 onHeapStorageRegionSize 来表示 spark.storage.storageFraction 划分的存储内存区域。这部分内存是不可以被驱逐(Evict)的存储内存（但是如果空闲是可以被占用的）。

- 当计算内存不足时，可以借用 onHeapStorageRegionSize 中未使用部分，且 Storage 内存的空间被对方占用后，需要等待执行内存自己释放，不能抢占。

- 若实际 StorageMemory 使用量超过 onHeapStorageRegionSize，那么当计算内存不足时，可以驱逐并借用 StorageMemory – onHeapStorageRegionSize 部分，而 onHeapStorageRegionSize 部分不可被抢占。

- 反之，当存储内存不足时（存储空间不足是指不足以放下一个完整的 Block），也可以借用计算内存空间；但是 Execution 内存的空间被存储内存占用后，是可让对方将占用的部分转存到硬盘，然后"归还"借用的空间。

- 如果双方的空间都不足时，则存储到硬盘；将内存中的块存储到磁盘的策略是按照 LRU 规则进行的。

> ### 任务内存管理（Task Memory Manager）

Executor 中任务以线程的方式执行，各线程共享JVM的资源（即 Execution 内存），任务之间的内存资源没有强隔离（任务没有专用的Heap区域）。因此，可能会出现这样的情况：先到达的任务可能占用较大的内存，而后到的任务因得不到足够的内存而挂起。

在 Spark 任务内存管理中，使用 HashMap 存储任务与其消耗内存的映射关系。每个任务可占用的内存大小为潜在可使用计算内存（ 潜在可使用计算内存为: 初始计算内存 + 可抢占存储内存）的 1/2n ~ 1/n，当剩余内存为小于 1/2n 时，任务将被挂起，直至有其他任务释放执行内存，而满足内存下限 1/2n，任务被唤醒。其中 n 为当前 Executor 中活跃的任务树。

比如如果 Execution 内存大小为 10GB，当前 Executor 内正在运行的 Task 个数为5，则该 Task 可以申请的内存范围为 10 / (2 * 5) ~ 10 / 5，也就是 1GB ~ 2GB 的范围。

任务执行过程中，如果需要更多的内存，则会进行申请，如果存在空闲内存，则自动扩容成功，否则，将抛出 OutOffMemroyError。

每个 Executor 中可同时运行的任务数由 Executor 分配的 CPU 的核数 N 和每个任务需要的 CPU 核心数 C 决定。其中:

```

N = spark.executor.cores

C = spark.task.cpus

```

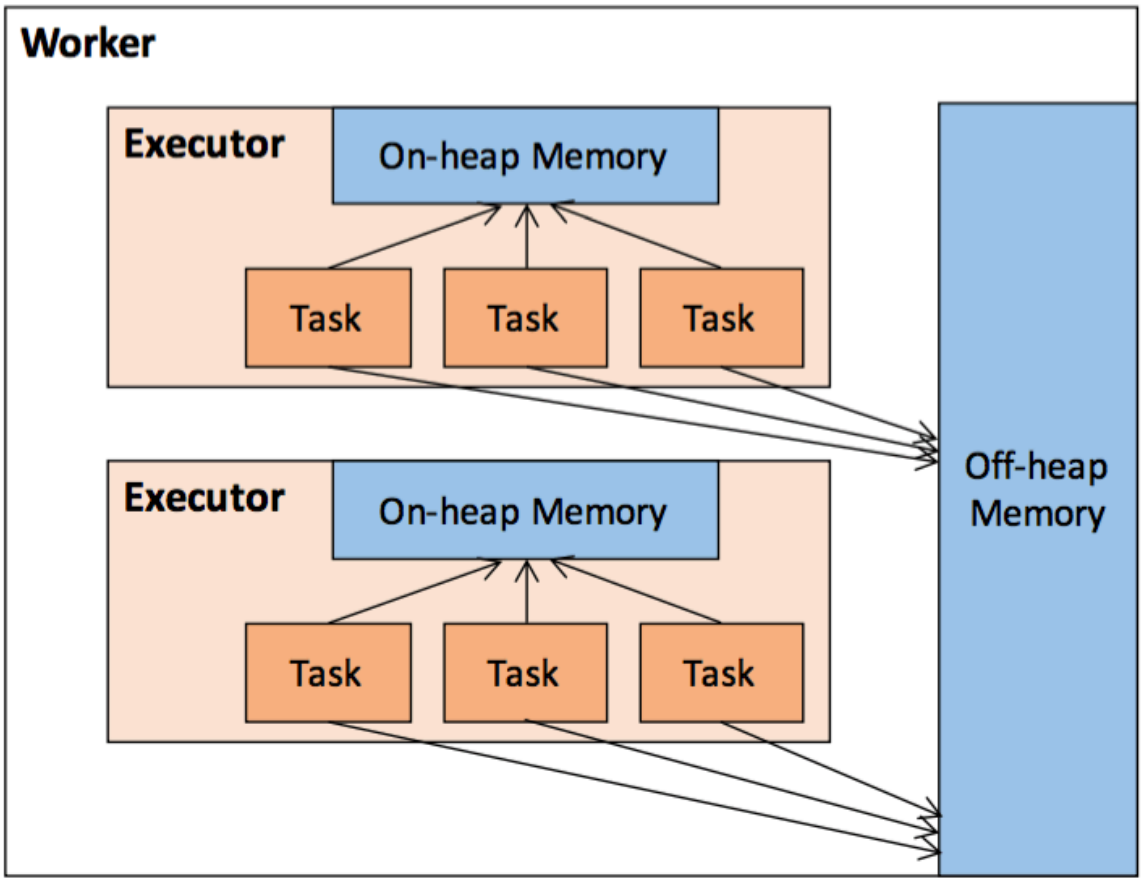
由此每个 Executor 的最大任务并行度可表示为 : TP = N / C 。

其中，C 值与应用类型有关，大部分应用使用默认值 1 即可，因此，影响 Executor 中最大任务并行度（最大活跃task数）的主要因素是 N。

依据 Task 的内存使用特征，前文所述的 Executor 内存模型可以简单抽象为下图所示模型：

其中，Executor 向 yarn 申请的总内存可表示为 : M = M1 + M2 。

如果考虑堆外内存则大概是如下结构：



> ## Executor内存参数调优

\1. Executor JVM Used Memory Heuristic

```

现象：配置的executor内存比实际使用的JVM最大使用内存还要大很多。

原因：这意味着 executor 内存申请过多了，实际上并不需要使用这么多内存。

解决方案：将 spark.executor.memory 设置为一个比较小的值。
```

```

\2. Executor Unified Memory Heuristic

```

现象：分配的统一内存 (Unified Memory = Storage Memory + Execution Memory) 比 executor 实际使用的统一内存大的多。

原因：这意味着不需要这么大的统一内存。

解决方案：降低 spark.memory.fraction 的比例。

```

\3. Executor OOM类错误 （错误代码 137、143等）

```

该类错误一般是由于 Heap（M2）已达上限，Task 需要更多的内存，而又得不到足够的内存而导致。因此，解决方案要从增加每个 Task 的内存使用量，满足任务需求 或 降低单个 Task 的内存消耗量，从而使现有内存可以满足任务运行需求两个角度出发。因此有如下解决方案:

法一：增加单个task的内存使用量

增加最大 Heap值，即上图中 M2 的值，使每个 Task 可使用内存增加。

降低 Executor 的可用 Core 的数量 N , 使 Executor 中同时运行的任务数减少，在总资源不变的情况下，使每个 Task 获得的内存相对增加。当然，这会使得 Executor 的并行度下降。可以通过调高 spark.executor.instances 参数来申请更多的 executor 实例（或者通过 spark.dynamicAllocation.enabled 启动动态分配），提高job的总并行度。

法二： 降低单个Task的内存消耗量

降低单个Task的内存消耗量可从配置方式和调整应用逻辑两个层面进行优化:

一、配置方式

减少每个 Task 处理的数据量，可降低 Task 的内存开销，在 Spark 中，每个 partition 对应一个处理任务 Task，因此，在数据总量一定的前提下，可以通过增加 partition 数量的方式来减少每个 Task 处理的数据量，从而降低 Task 的内存开销。针对不同的 Spark 应用类型，存在不同的 partition 配置参数 :

P = spark.default.parallism (非SQL应用)

P = spark.sql.shuffle.partition (SQL 应用)

通过增加 P 的值，可在一定程度上使 Task 现有内存满足任务运行。注: 当调整一个参数不能解决问题时，上述方案应进行协同调整。

二、调整应用逻辑

Executor OOM 一般发生 Shuffle 阶段，该阶段需求计算内存较大，且应用逻辑对内存需求有较大影响，下面举例就行说明：

1、选择合适的算子，如 groupByKey 转换为 reduceByKey

一般情况下，groupByKey 能实现的功能使用 reduceByKey 均可实现，而 ReduceByKey 存在 Map 端的合并，可以有效减少传输带宽占用及 Reduce 端内存消耗。

2、避免数据倾斜 (data skew)

Data Skew 是指任务间处理的数据量存大较大的差异。

如左图所示，key 为 010 的数据较多，当发生 shuffle 时，010 所在分区存在大量数据，不仅拖慢 Job 执行（Job 的执行时间由最后完成的任务决定）。 而且导致 010 对应 Task 内存消耗过多，可能导致 OOM。

而右图，经过预处理（加盐，此处仅为举例说明问题，解决方法不限于此）可以有效减少 Data Skew 导致的问题。

```

\4. Execution Memory Spill Heuristic

```

现象: 在 stage 3 发现执行内存溢出。Shuffle read bytes 和 spill 分布均匀。这个 stage 有 200 个 tasks。

原因: 执行内存溢出，意味着执行内存不足。跟上面的 OOM 错误一样，只是执行内存不足的情况下不会报 OOM 而是会将数据溢出到磁盘。但是整个性能很难接受。

解决方案: 同 3。

```

\5. Beyond ... memory, killed by yarn.

```

出现该问题原因是由于实际使用内存上限超过申请的内存上限而被 Yarn 终止掉了, 首先说明 Yarn 中 Container 的内存监控机制：

Container 进程的内存使用量 : 以 Container 进程为根的进程树中所有进程的内存使用总量。

Container 被杀死的判断依据 : 进程树总内存（物理内存或虚拟内存）使用量超过向 Yarn 申请的内存上限值，则认为该 Container 使用内存超量，可以被“杀死”。

因此，对该异常的分析要从是否存在子进程两个角度出发。

1、不存在子进程

根据 Container 进程杀死的条件可知，在不存在子进程时，出现 killed by yarn 问题是于由 Executor(JVM) 进程自身内存超过向 Yarn 申请的内存总量 M 所致。由于未出现上一节所述的 OOM 异常，因此可判定其为 M1 (Overhead) 不足，依据 Yarn 内存使用情况有如下两种方案:

法一、如果，M (spark.executor.memory) 未达到 Yarn 单个 Container 允许的上限时，可仅增加 M1（spark.yarn.executor.memoryOverhead），从而增加 M；如果，M 达到 Yarn 单个 Container 允许的上限时，增加 M1，降低 M2。

注意二者之各要小于 Container 监控内存量，否则伸请资源将被 yarn 拒绝。

法二、减少可用的 Core 的数量 N，使并行任务数减少，从而减少 Overhead 开销

2、存在子进程

Spark 应用中 Container 以 Executor（JVM进程）的形式存在，因此根进程为 Executor 对应的进程，而 Spark 应用向Yarn申请的总资源 M = M1 + M2，都是以 Executor(JVM) 进程（非进程树）可用资源的名义申请的。申请的资源并非一次性全量分配给 JVM 使用，而是先为 JVM 分配初始值，随后内存不足时再按比率不断进行扩容，直致达到 Container 监控的最大内存使用量 M。当 Executor 中启动了子进程（如调用 shell 等）时，子进程占用的内存（记为 S）就被加入 Container 进程树，此时就会影响 Executor 实际可使用内存资源（Executor 进程实际可使用资源变为: M - S），然而启动 JVM 时设置的可用最大资源为 M，且 JVM 进程并不会感知 Container 中留给自己的使用量已被子进程占用，因此，当 JVM 使用量达到 M - S，还会继续开劈内存空间，这就会导致 Executor 进程树使用的总内存量大于 M 而被 Yarn 杀死。

典形场景有:

PySpark（Spark已做内存限制，一般不会占用过大内存）

自定义Shell调用

其解决方案分别为:

1) PySpark场景:

如果，M 未达到 Yarn 单个 Container 允许的上限时，可仅增加 M1 ，从而增加 M；如果，M 达到 Yarn 单个 Container 允许的上限时，增加 M1，降低 M2。

减少可用的 Core 的数量 N，使并行任务数减少，从而减少 Overhead 开销

2) 自定义 Shell 场景:（OverHead 不足为假象）

调整子进程可用内存量 (通过单机测试，内存控制在 Container 监控内存以内，且为 Spark 保留内存等留有空间）。

```
```