# Java

## jdk8特性

### Metaspace（元空间）

JDK 1.8 中， HotSpot 已经没有 “PermGen space”这个区间了，取而代之是一个叫做 Metaspace（元空间） 的东西。

移除永久代的工作从JDK1.7就开始了。JDK1.7中，存储在永久代的部分数据就已经转移到了Java Heap或者是 Native Heap。但永久代仍存在于JDK1.7中，并没完全移除，譬如符号引用(Symbols)转移到了native heap；字面量(interned strings)转移到了java heap；类的静态变量(class statics)转移到了java heap。

元空间的本质和永久代类似，都是对JVM规范中方法区的实现。不过元空间与永久代之间最大的区别在于：元空间并不在虚拟机中，而是使用本地内存。因此，默认情况下，元空间的大小仅受本地内存限制，但可以通过以下参数来指定元空间的大小：

-XX:MetaspaceSize，初始空间大小，达到该值就会触发垃圾收集进行类型卸载，同时GC会对该值进行调整：如果释放了大量的空间，就适当降低该值；如果释放了很少的空间，那么在不超过MaxMetaspaceSize时，适当提高该值。

-XX:MaxMetaspaceSize，最大空间，默认是没有限制的。

## JVM(Java虚拟机)

### 运行时数据区



#### 所有线程共享的数据区（堆空间线程共享）

##### 方法区(Method Area)

存放被虚拟机加载的类信息、常量、静态变量、编译后的代码数据，也称为Permanent Generation。

-XX:MaxPermSize=16m ##设置持久代最大为16m。

##### 堆(Heap)

存放对象实例

-Xmx3500m ##设置堆最大可用内存。

-Xms3500m ##设置堆最小可用内存。

-Xmn2g

##设置年轻代大小（整个堆大小=年轻代大小+年老代大小+持久代大小）。

##持久代一般固定大小为64m，所以增大年轻代后，将会减小年老代大小。

##此值对系统性能影响较大，Sun官方推荐配置为整个堆的3/8。

-XX:NewRatio=4

##设置年轻代（包括Eden和两个Survivor区）与年老代的比值（不包括持久代）。设置为4，则年轻代与年老代所占比值为1：4，年轻代占整个堆栈的1/5。

-XX:SurvivorRatio=4

##设置年轻代中Eden区与Survivor区的大小比值。

设置为4，则两个Survivor区与一个Eden区的比值为2:4，一个Survivor区占整个年轻代的1/6。

-XX:MaxTenuringThreshold=0

##设置垃圾最大年龄。

如果设置为0的话，则年轻代对象不经过Survivor区，直接进入年老代 。对于年老代比较多的应用，可以提高效率。如果将此值设置为一个较大值，则年轻代对象会在Survivor区进行多次复制，这样可以增加对象在年轻代的存活时间 ，增加在年轻代即被回收的概率。

#### 线程隔离的数据区（栈空间线程私有）

生命周期与线程相同。

##### 虚拟机栈(Virtual Machine Stack)

Java方法执行的内存模型。

Stack Frame(栈帧)用于存储局部变量表、操作数栈、动态链接、方法出口等信息。

每一个方法从调用直至执行完成的过程，就对应着一个栈帧在虚拟机栈中入栈到出栈的过程。

局部变量表存放编译可知的各种基本数据类型、对象引用类型和returnAddress类型(指向一条字节码指令的地址)。

long和double占用2个局部变量空间(Slot)，其余占用1个。

局部变量表所需内存空间在编译期间完成分配，当进入方法时，这个方法需要在帧中分配多大局部变量空间是完全确定的，方法在运行期间不会改变局部变量表的大小。

如果线程请求的栈深度大于虚拟机所允许的深度，将抛出StackOverflowError异常；如果虚拟机栈可以动态扩展（当前大部分的Java虚拟机都可动态扩展，只不过Java虚拟机规范中也允许固定长度的虚拟机栈），如果扩展时无法申请到足够的内存，就会抛出OutOfMemoryError异常。

-Xss128k

##设置每个线程的堆栈大小。

##根据应用的线程所需内存大小进行调整。

##在相同物理内存下，减小这个值能生成更多的线程。

##但是操作系统对一个进程内的线程数还是有限制的，不能无限生成，经验值在3000~5000左右。

##### 本地方法栈(Native Method Stack)

与虚拟机栈类似。

虚拟机栈为虚拟机执行Java方法（字节码）服务。

本地方法栈为虚拟机使用的Native方法服务。

有的虚拟机(例如Sun HotSpot)直接把本地方法栈和虚拟机栈合二为一。

##### 程序计数器(Program Counter Register)

当前线程所执行的字节码的行号指示器，记录正在执行的虚拟机字节码指令的地址，如果正在执行Native方法，这个计数器值为空(Undefined)

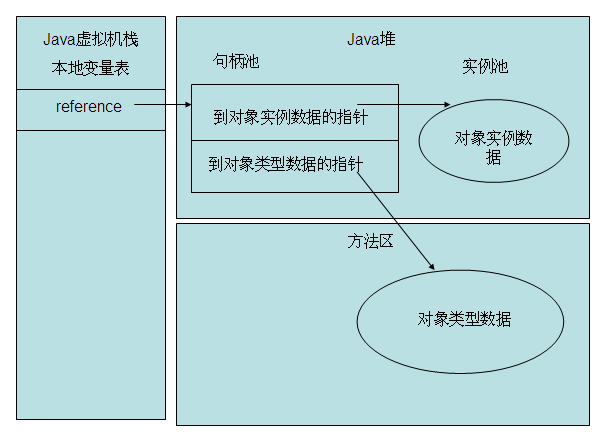
唯一一个在Java虚拟机规范中没有规定任何OutOfMemoryError情况的区域。

各个线程之间计数器互不影响，独立存储。

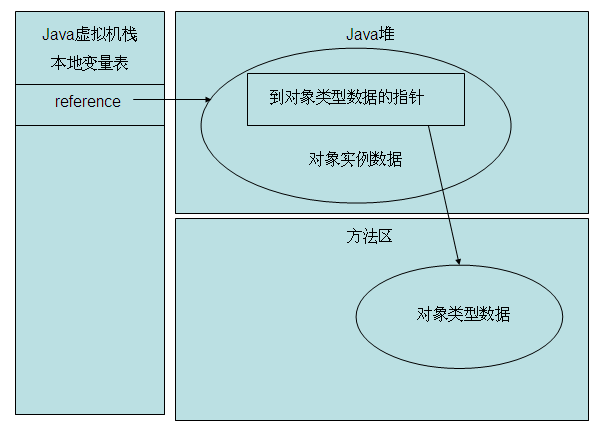
### 对象访问的两种方式

主流的两种访问方式:使用句柄和直接指针。(HotSpot虚拟机就是使用直接指针的访问方式)

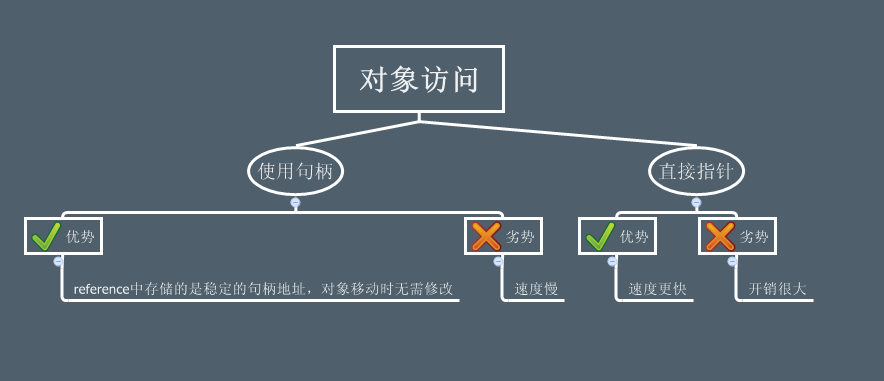
#### 使用句柄访问



#### 使用直接指针访问



优缺点比较



### 堆分代

1、minor gc 之后，存活于survivor 区域的对象的age会+1，当超过（默认）15的时候，转移到老年代。

2、动态对象，如果survivor空间中相同年龄所有的对象大小的综合和大于survivor空间的一半，年级大于或等于该年级的对象就可以直接进入老年代。

以上采用分代垃圾收集的思想，对一个对象从存活到死亡所经历的历程。期间，在新生代的时刻，会用到复制算法，在老年代时，有可能会用到标记-清楚算法（mark-sweep）算法或者标记-整理算法，这些都是垃圾回收算法基于不同区域的实现，我们看下这几种回收算法的实现原理。

#### Eden

创建的新对象在这里分配内存

#### FromSurvivor和ToSurvivor

### jvm参数说明

#### java -XX:+PrintFlagsInitial

查看本机的初始化参数

#### 堆内存

##### -Xms

初始堆大小，默认值物理内存的1/64(<1GB)，默认(MinHeapFreeRatio参数可以调整)空余堆内存小于40%时，JVM就会增大堆直到-Xmx的最大限制.

##### -Xmx

最大堆大小，物理内存的1/4(<1GB)，默认(MaxHeapFreeRatio参数可以调整)空余堆内存大于70%时，JVM会减少堆直到 -Xms的最小限制

##### -Xmn

年轻代大小(1.4or lator),注意：此处的大小是（eden+ 2 survivor space).与jmap -heap中显示的New gen是不同的。整个堆大小=年轻代大小 + 年老代大小 + 持久代大小.增大年轻代后,将会减小年老代大小.此值对系统性能影响较大,Sun官方推荐配置为整个堆的3/8

##### -XX:NewSize

设置年轻代大小(for 1.3/1.4)

##### -XX:MaxNewSize

年轻代最大值(for 1.3/1.4)

##### -XX:NewRatio

年轻代(包括Eden和两个Survivor区)与年老代的比值(除去持久代)，-XX:NewRatio=4表示年轻代与年老代所占比值为1:4,年轻代占整个堆栈的1/5，Xms=Xmx并且设置了Xmn的情况下，该参数不需要进行设置。

##### -XX:SurvivorRatio

Eden区与Survivor区的大小比值，设置为8,则两个Survivor区与一个Eden区的比值为2:8,一个Survivor区占整个年轻代的1/10。

##### -XX:MaxTenuringThreshold

垃圾最大年龄，如果设置为0的话,则年轻代对象不经过Survivor区,直接进入年老代. 对于年老代比较多的应用,可以提高效率.如果将此值设置为一个较大值,则年轻代对象会在Survivor区进行多次复制,这样可以增加对象再年轻代的存活 时间,增加在年轻代即被回收的概率，该参数只有在串行GC时才有效.

#### 堆回收器

##### -XX:+UseParallelGC

Full GC采用parallel MSC(此项待验证)

选择垃圾收集器为并行收集器.此配置仅对年轻代有效.即上述配置下,年轻代使用并发收集,而年老代仍旧使用串行收集.( 不可与CMS收集同时使用)( 系统吞吐量优先)

##### -XX:+UseParNewGC

设置年轻代为并行收集, 可与CMS收集同时使用

JDK5.0以上,JVM会根据系统配置自行设置,所以无需再设置此值

##### -XX:+UseConcMarkSweepGC

设置年老代为并发收集,使用CMS内存收集, 测试中配置这个以后,-XX:NewRatio=4的配置失效了,原因不明.所以,此时年轻代大小最好用-Xmn设置

#### 方法区(永久代)

##### -XX:PermSize

设置持久代(perm gen)初始值，默认物理内存的1/64

##### -XX:MaxPermSize

设置持久代最大值,默认物理内存的1/4

#### 线程栈

##### -Xss

每个线程的堆栈大小，JDK5.0以后每个线程堆栈大小为1M,以前每个线程堆栈大小为256K.根据应用的线程所需内存大小进行调整.在相同物理内存下,减小这个值能生成更多的线程.但是操作系统对一个进程内的线程数还是有限制的,不能无限生成,经验值在3000~5000左右

一般小的应用， 如果栈不是很深， 应该是128k够用的 大的应用建议使用256k。这个选项对性能影响比较大，需要严格的测试。

和threadstacksize选项解释很类似,官方文档似乎没有解释,在论坛中有这样一句话:"”

-Xss is translated in a VM flag named ThreadStackSize”

一般设置这个值就可以了。

#### gc日志

##### -XX:+PrintGC

##### -XX:+PrintGCDetails

##### -XX:+PrintGCTimeStamps

##### -XX:+PrintHeapAtGC

打印GC前后的详细堆栈信息

##### -Xloggc:filename

与上面几个配合使用，把相关日志信息记录到文件以便分析。

##### -XX:+UseGCLogFileRotation

gc日志文件转存

##### -XX:GCLogFileSize=200k

gc日志最大文件size

##### -XX:NumberOfGCLogFiles=10

gc日志最多文件数

#### -XX:LargePageSizeInBytes

内存页的大小不可设置过大， 会影响Perm的大小，=128m

#### jdk8

从JDK8开始，永久代(PermGen)的概念被废弃掉了，取而代之的是一个称为Metaspace的存储空间。Metaspace使用的是本地内存，而不是堆内存，也就是说在默认情况下Metaspace的大小只与本地内存大小有关。当然你也可以通过以下的几个参数对Metaspace进行控制：

##### -XX:MetaspaceSize=N

这个参数是初始化的Metaspace大小，该值越大触发Metaspace GC的时机就越晚。随着GC的到来，虚拟机会根据实际情况调控Metaspace的大小，可能增加上线也可能降低。在默认情况下，这个值大小根据不同的平台在12M到20M浮动。使用java -XX:+PrintFlagsInitial命令查看本机的初始化参数，-XX:Metaspacesize为21810376B（大约20.8M）。

##### -XX:MaxMetaspaceSize=N

这个参数用于限制Metaspace增长的上限，防止因为某些情况导致Metaspace无限的使用本地内存，影响到其他程序。在本机上该参数的默认值为4294967295B（大约4096MB）。

##### -XX:MinMetaspaceFreeRatio=N

当进行过Metaspace GC之后，会计算当前Metaspace的空闲空间比，如果空闲比小于这个参数，那么虚拟机将增长Metaspace的大小。在本机该参数的默认值为40，也就是40%。设置该参数可以控制Metaspace的增长的速度，太小的值会导致Metaspace增长的缓慢，Metaspace的使用逐渐趋于饱和，可能会影响之后类的加载。而太大的值会导致Metaspace增长的过快，浪费内存。

##### -XX:MaxMetasaceFreeRatio=N

当进行过Metaspace GC之后， 会计算当前Metaspace的空闲空间比，如果空闲比大于这个参数，那么虚拟机会释放Metaspace的部分空间。在本机该参数的默认值为70，也就是70%。

##### -XX:MaxMetaspaceExpansion=N

Metaspace增长时的最大幅度。在本机上该参数的默认值为5452592B（大约为5MB）。

##### -XX:MinMetaspaceExpansion=N

Metaspace增长时的最小幅度。在本机上该参数的默认值为340784B（大约330KB为）。

### 垃圾回收算法

#### 标记-清除算法(Mark-Sweep)

最基础的收集算法是“标记-清除”（Mark-Sweep）算法，如同它的名字一样，算法分为“标记”和“清除”两个阶段：首先标记出所有需要回收的对象，在标记完成后统一回收所有被标记的对象，它的标记过程其实在前一节讲述对象标记判定时已经介绍过了。之所以说它是最基础的收集算法，是因为后续的收集算法都是基于这种思路并对其不足进行改进而得到的。它的主要不足有两个：一个是效率问题，标记和清除两个过程的效率都不高；另一个是空间问题，标记清除之后会产生大量不连续的内存碎片，空间碎片太多可能会导致以后在程序运行过程中需要分配较大对象时，无法找到足够的连续内存而不得不提前触发另一次垃圾收集动作。

#### 复制算法(Copying)

为了解决效率问题，一种称为“复制”（Copying）的收集算法出现了，它将可用内存按容量划分为大小相等的两块，每次只使用其中的一块。当这一块的内存用完了，就将还存活着的对象复制到另外一块上面，然后再把已使用过的内存空间一次清理掉。这样使得每次都是对整个半区进行内存回收，内存分配时也就不用考虑内存碎片等复杂情况，只要移动堆顶指针，按顺序分配内存即可，实现简单，运行高效。只是这种算法的代价是将内存缩小为了原来的一半，未免太高了一点。

现在的商业虚拟机都采用这种收集算法来回收新生代，IBM公司的专门研究表明，新生代中的对象98%是“朝生夕死”的，所以并不需要按照1:1的比例来划分内存空间，而是将内存分为一块较大的Eden空间和两块较小的Survivor空间，每次使用Eden和其中一块Survivor。当回收时，将Eden和Survivor中还存活着的对象一次性地复制到另外一块Survivor空间上，最后清理掉Eden和刚才用过的Survivor空间。HotSpot虚拟机默认Eden和Survivor的大小比例是8:1，也就是每次新生代中可用内存空间为整个新生代容量的90%（80%+10%），只有10%的内存会被“浪费”。当然，98%的对象可回收只是一般场景下的数据，我们没有办法保证每次回收都只有不多于10%的对象存活，当Survivor空间不够用时，需要依赖其他内存（这里指老年代）进行分配担保（Handle Promotion）。

#### 标记-整理算法(Mark-Compact)

复制收集算法在对象存活率较高时就要进行较多的复制操作，效率将会变低。更关键的是，如果不想浪费50%的空间，就需要有额外的空间进行分配担保，以应对被使用的内存中所有对象都100%存活的极端情况，所以在老年代一般不能直接选用这种算法。

根据老年代的特点，有人提出了另外一种“标记-整理”（Mark-Compact）算法，标记过程仍然与“标记-清除”算法一样，但后续步骤不是直接对可回收对象进行清理，而是让所有存活的对象都向一端移动，然后直接清理掉端边界以外的内存。

#### 分代收集算法(Generational Collection)

当前商业虚拟机的垃圾收集都采用“分代收集”（Generational Collection）算法，这种算法并没有什么新的思想，只是根据对象存活周期的不同将内存划分为几块。一般是把Java堆分为新生代和老年代，这样就可以根据各个年代的特点采用最适当的收集算法。在新生代中，每次垃圾收集时都发现有大批对象死去，只有少量存活，那就选用复制算法，只需要付出少量存活对象的复制成本就可以完成收集。而老年代中因为对象存活率高、没有额外空间对它进行分配担保，就必须使用“标记—清理”或者“标记—整理”算法来进行回收。

### 垃圾回收器

#### 串行回收器

只适用于小数据量的情况。

#### 并行回收器（吞吐量优先）

并行收集器主要以到达一定的吞吐量为目标，适用于科学技术和后台处理等。

XX:+UseParallelGC

##选择垃圾收集器为并行收集器。此配置仅对年轻代有效。即上述配置下，年轻代使用行发收集，而年老代仍旧使用串行收集。

-XX:ParallelGCThreads=20

##配置并行收集器的线程数，即：同时多少个线程一起进行垃圾回收。此值最好配置与处理器数目相等。

-XX:+UseParallelOldGC

##配置年老代垃圾收集方式为并行收集。

-XX:MaxGCPauseMillis=100

##设置每次年轻代垃圾回收的最长时间，如果无法满足此时间，JVM会自动调整年轻代大小，以满足此值。

-XX:+UseAdaptiveSizePolicy

##设置此选项后，并行收集器会自动选择年轻代区大小和相应的Survivor区比例，以达到目标系统规定的最低相应时间或者收集频率等，此值建议使用并行收集器时，一直打开。

#### 并发回收器(响应时间优先)

并发收集器主要是保证系统的响应时间，减少垃圾收集时的停顿时间。适用于应用服务器、电信领域等。

-XX:+UseConcMarkSweepGC

##CMS(Concurrent Mark-Sweep)

##设置年老代为并发收集。测试中配置这个以后，-XX:NewRatio=4的配置失效了，原因不明。所以，此时年轻代大小最好用-Xmn设置。

-XX:+UseParNewGC

##设置年轻代为并行收集。可与CMS收集同时使用。jdk5.0以上，JVM会根据系统配置自行设置，所以无需再设置此值。

-XX:CMSFullGCsBeforeCompaction=5

##由于并发收集器不对内存空间进行压缩、整理，所以运行一段时间以后会产生“碎片”，使得运行效率降低。此值设置运行多少次GC以后对内存空间进行压缩、整理。

-XX:+UseCMSCompactAtFullCollection

##打开对年老代的压缩。可能会影响性能，但是可以消除碎片。

### 辅助信息

#### -XX:+PrintGC

##输出形式：[GC 118250K->113543K(130112K), 0.0094143 secs]

                [Full GC 121376K->10414K(130112K), 0.0650971 secs]

#### -XX:+PrintGCDetails

##输出形式：[GC [DefNew: 8614K->781K(9088K), 0.0123035 secs] 118250K->113543K(130112K), 0.0124633 secs]

                [GC [DefNew: 8614K->8614K(9088K), 0.0000665 secs][Tenured: 112761K->10414K(121024K), 0.0433488 secs] 121376K->10414K(130112K), 0.0436268 secs]

#### -XX:+PrintGCTimeStamps

##可与上面两个混合使用。

##输出形式：11.851: [GC 98328K->93620K(130112K), 0.0082960 secs]

#### -XX:+PrintGCApplicationConcurrentTime

##打印每次垃圾回收前，程序未中断的执行时间。可与上面混合使用

##输出形式：Application time: 0.5291524 seconds

#### -XX:+PrintGCApplicationStoppedTime

##打印垃圾回收期间程序暂停的时间。可与上面混合使用

输出形式：Total time for which application threads were stopped: 0.0468229 seconds

#### -XX:+PrintHeapAtGC

##打印GC前后的详细堆栈信息

#### -Xloggc:filename

##与上面几个配合使用，把相关日志信息记录到文件以便分析。

### perf

### 内存分析工具

#### jps

jps是JDK提供的一个小工具，上面的命令会把操作系统里面的java应用都展示出来，显示PID，启动类或者JAR，VM参数。可以通过jps -help进一步了解详细信息。当然也可以用操作系统的netsat查询PID。

##### jps -vl

#### jmap

##### jmap -heap pid

jmap查看内存使用情况与生成heapdump

jmap -heap 1234  (1234为进程号)

-heap：打印heap空间的概要，这里可以粗略的检验heap空间的使用情况。

例：

jmap -heap 12345

输出：

Attaching to process ID 2657, please wait...

Debugger attached successfully.

Client compiler detected.

JVM version is 1.5.0\_16-b02

using thread-local object allocation.

Mark Sweep Compact GC

Heap Configuration:

MinHeapFreeRatio = 40

MaxHeapFreeRatio = 70

MaxHeapSize = 67108864 (64.0MB)

NewSize = 655360 (0.625MB)

MaxNewSize = 4294901760 (4095.9375MB)

OldSize = 1441792 (1.375MB)

NewRatio = 12

SurvivorRatio = 8

PermSize = 8388608 (8.0MB)

MaxPermSize = 67108864 (64.0MB)

Heap Usage:

New Generation (Eden + 1 Survivor Space):----------------------------------------新生代区

capacity = 4521984 (4.3125MB)

used = 1510200 (1.4402389526367188MB)

free = 3011784 (2.8722610473632812MB)

33.39684527853261% used

Eden Space:--------------------------------------------------------------------伊甸园区

capacity = 4063232 (3.875MB)

used = 1495992 (1.4266891479492188MB)

free = 2567240 (2.4483108520507812MB)

36.81778446320565% used

From Space:-------------------------------------------------------------------年轻代（幸存者乐园1）

capacity = 458752 (0.4375MB)

used = 14208 (0.0135498046875MB)

free = 444544 (0.4239501953125MB)

3.0970982142857144% used

To Space:-----------------------------------------------------------------------------年轻代（幸存者乐园2）

capacity = 458752 (0.4375MB)

used = 0 (0.0MB)

free = 458752 (0.4375MB)

0.0% used

concurrent mark-sweep generation:-------------------------------------------------老年代

capacity = 8589934592 (8192.0MB)

used = 0 (0.0MB)

free = 8589934592 (8192.0MB)

0.0% used

Perm Generation:----------------------------------------------------------------------永久代

capacity = 11796480 (11.25MB)

used = 11712040 (11.169471740722656MB)

free = 84440 (0.08052825927734375MB)

99.28419325086806% used

以上的输出很简单，第四行起开始输出此进程我们的JAVA使用的环境。

Heap Configuration，指在我们启动时设置的一些JVM参数。像最大使用内存大小，年老代，年青代，持久代大小等。有这个可以很简单的查看本进程的内存使用情况。也许进程占用的总内存比较多，但我们在这里可以看到真正用到的并没有多少，很多都是"Free"。内存使用的堆积大多在老年代，内存池露始于此，所以要格外关心“tenured generation”。

##### jmap -dump:live,format=b,file=fileName pid

产生一个HeapDump文件

##### jmap -histo [pid]

查看JVM堆中对象详细占用情况

#### jinfo

##### jinfo -flags pid

显示JVM的参数，包括显示设置的和系统默认的。比如所用的垃圾回收器，堆的最大值等。

##### jinfo -sysprops pid

显示System.getProperties()的内容

#### jstat

##### jstat -gc pid

显示JVM的各个内存区使用情况（容量和使用量），GC的次数和耗时。

##### jstat -class pid

查看class的加载情况。

#### jstack [pid]

线程的所有堆栈信息，包括死锁进程

线程状态的分析

死锁，Deadlock（重点关注）

执行中，Runnable

等待资源，Waiting on condition（重点关注）

          如果发现有大量的线程都在处在 Wait on condition，从线程 stack看， 正等待网络读写，这可能是一个网络瓶颈的征兆。因为网络阻塞导致线程无法执行。一种情况是网络非常忙，几 乎消耗了所有的带宽，仍然有大量数据等待网络读 写；另一种情况也可能是网络空闲，但由于路由等问题，导致包无法正常的到达。所以要结合系统的一些性能观察工具来综合分析，比如 netstat统计单位时间的发送包的数目，如果很明显超过了所在网络带宽的限制 ; 观察 cpu的利用率，如果系统态的 CPU时间，相对于用户态的 CPU时间比例较高；

等待获取监视器，Waiting on monitor entry（重点关注）

暂停，Suspended

对象等待中，Object.wait() 或 TIMED\_WAITING

阻塞，Blocked（重点关注）

停止，Parked

TIMED\_WAITING (parking)：这里指定了时间，到达指定的时间后自动退出等待状态；parking指线程处于挂起中

#### jstack pid

查看线程运行情况，检测是否有死锁。

#### jconsole

#### jvisualvm

#### Eclipse Memory Analysis

##### 安装

在Eclipse help -> Eclipse Marketplace下搜索Memory

##### Shallow Size

Shallow Size是对象本身占据的内存的大小，不包含其引用的对象。对于常规对象（非数组）的Shallow Size由其成员变量的数量和类型来定，而数组的ShallowSize由数组类型和数组长度来决定，它为数组元素大小的总和。

##### Retained Size

Retained Size=当前对象大小+当前对象可直接或间接引用到的对象的大小总和。(间接引用的含义：A->B->C,C就是间接引用) ，并且排除被GC Roots直接或者间接引用的对象

换句话说，Retained Size就是当前对象被GC后，从Heap上总共能释放掉的内存。

不过，释放的时候还要排除被GC Roots直接或间接引用的对象。他们暂时不会被被当做Garbage。

#### JProfiler

### [linux下查找java进程占用CPU过高的线程](http://www.cnblogs.com/liusxg/p/5811085.html)

top查看进程占用资源情况

使用top -H -p <pid>查看线程占用情况

将线程id转换成十六进制

#printf %x 15664

#3d30

jstack <pid> | grep -a 线程id（十六进制）

### JVM性能调优实战

如何在高性能服务器上进行JVM调优？

为了充分利用高性能服务器的硬件资源，有两种JVM调优方案，它们都有各自的优缺点，需要根据具体的情况进行选择。

1. 采用64位操作系统，并为JVM分配大内存

我们知道，如果JVM中堆内存太小，那么就会频繁地发生垃圾回收，而垃圾回收都会伴随不同程度的程序停顿，因此，如果扩大堆内存的话可以减少垃圾回收的频率，从而避免程序的停顿。

因此，人们自然而然想到扩大内存容量。而32位操作系统理论上最大只支持4G内存，64位操作系统最大能支持128G内存，因此我们可以使用64位操作系统，并使用64位JVM，并为JVM分配更大的堆内存。但问题也随之而来。

堆内存变大后，虽然垃圾收集的频率减少了，但每次垃圾回收的时间变长。如果对内存为14G，那么每次Full GC将长达数十秒。如果Full GC频繁发生，那么对于一个网站来说是无法忍受的。

因此，对于使用大内存的程序来说，一定要减少Full GC的频率，如果每天只有一两次Full GC，而且发生在半夜， 那完全可以接受。

要减少Full GC的频率，就要尽量避免太多对象进入老年代，可以有以下做法：

确保对象都是“朝生夕死”的

一个对象使用完后应尽快让他失效，然后尽快在新生代中被Minor GC回收掉，尽量避免对象在新生代中停留太长时间。

提高大对象直接进入老年代的门槛

通过设置参数-XX:PretrnureSizeThreshold来提高大对象的门槛，尽量让对象都先进入新生代，然后尽快被Minor GC回收掉，而不要直接进入老年代。

注意：使用64位JDK的注意点

64位JDK支持更大的堆内存，但更大的堆内存会导致一次垃圾回收时间过长。

现阶段，64位JDK的性能普遍比32位JDK低。

堆内存过大无法在发生内存溢出时生成内存快照

若将堆内存设为10G，那么当堆内存溢出时就要生成10G的大文件，这基本上是不可能的。

相同程序，64位JDK要比32位JDK消耗更大的内存

2. 使用32位JVM集群

针对于64位JDK种种弊端，我们更多选择使用32位JDK集群来充分利用高性能机器的硬件资源。

如何实现？

在一台服务器上运行多个服务器程序，这些程序都运行在32位的JDK上。然后再运行个服务器作为反向代理服务器，由它来实现负载均衡。

由于32位JDK最多支持2G内存，因此每个虚拟结点的堆内存可以分配1.6G，一共运行10个虚拟结点的话，这台物理服务器可以拥有16G的堆内存。

有啥弊端？

多个虚拟节点竞争共享资源时容易出现问题

如多个虚拟节点共同竞争IO操作，很可能会引起IO异常。

很难高效地使用资源池

如果每个虚拟节点使用各自的资源池，那么无法实现各个资源池的负载均衡。如果使用集中式资源池，那么又存在竞争的问题。

每个虚拟节点最大内存为2G

别忘了直接内存也可能导致内存溢出！

问题描述

有个小型网站，使用32位JDK，堆1.6G。运行期间发现老是出现内存溢出。为了判断是否是堆内存溢出，在程序运行前添加参数：-XX:+HeapDumpOnOutOfMemeryError(添加这个参数后当堆内存溢出时就会输出异常日至)。但当再次发生内存溢出时，没有生成相关异常日志。从而可以判定，不是堆内存发生溢出。

问题分析

我们可以发现，在32位JDK中，将1.6G分配给了堆，还有一部分分配给了JVM的其它内存，只有少于0.4G的内存为非JVM内存。我们知道，如果使用了NIO，那么JVM会在JVM内存之外分配内存空间，这部分内存也叫“直接内存”。因此，如果程序中使用了NIO，那么就要小心“直接内存”不足时发生内存溢出异常了！

直接内存的垃圾回收过程

直接内存虽然不是JVM内存空间，但它的垃圾回收也有JVM负责。直接内存的垃圾回收发生在Full GC时，只有当老年代内存满时，垃圾收集器才会顺便收集一下直接内存中的垃圾。

如果直接内存已满，但老年代没满，这时直接内存先是抛出异常，相应的catch块中调用System.gc()。由于System.gc()只是建议JVM回收，JVM可能不马上回收内存，那么这时直接内存就抛出内存溢出异常，使得程序终止。

JVM崩溃的原因

当内存溢出时，JVM仅仅会终止当前运行的程序，那么什么时候JVM会崩溃呢？

什么是异步请求？

我们知道，Web服务器和客户端采用HTTP通信，而HTTP底层采用TCP通信。异步通信就是当客户端向服务器发送一个HTTP请求后，将这个请求的TCP连接委托给其它线程，然后它转而做别的事，那条被委托的线程保持TCP连接，等待服务器的回信。当收到服务器回信后，再将收到的数据转交给刚才的线程。这个过程就是异步通信过程。

异步请求如何造成JVM崩溃？

如果一个Web应用使用了较多的异步请求(AJAX)，每次主线程发送完请求后都将TCP连接交给一条新的线程去等待服务器回信，那么如果网络不流畅时，这些受委托的线程迟迟等不到服务器的回信，因此保持着TCP连接。当TCP连接过多时，超过JVM的承受能力，JVM就发生崩溃。

如何处理大对象？

大对象对于JVM来说是个噩耗。如果对象过大，当前新生代的剩余空间装不下它，那么就需要使用分配担保机制，将当前新生代的对象都复制到老年代中，给大对象腾出空间。分配担保涉及到大量的复制，因此效率很低。

那么，如果将大对象直接放入老年代，虽然避免了分配担保过程，但该对象只有当Full GC时才能被回收，而Full GC的代价是高昂的。如果大对象过多时，老年代很快就装满了，这时就需要进行Full GC，如果Full GC频率过高，程序就会变得很卡。

因此，对于大对象，有如下几种处理方法：

1. 在写程序的时候尽量避免大对象

从源头降低大对象的出现，尽量选择空间利用率较高的数据结构存储。

2. 尽量缩短大对象的有效时间

对象用完后尽快让它失效，好让垃圾收集器尽快将他回收，避免因在新生代呆的时间过长而进入老年代。

### Minor GC,Major GC,Full GC

Minor GC清理年轻带内存

Major GC清理永久代

Full GC清理整个堆空间—包括年轻代和永久代，fullGC会触发 stop the world。

许多 Major GC 是由 Minor GC 触发的，所以很多情况下将这两种 GC 分离是不太可能的

**Minor GC ，Full GC 触发条件**

Minor GC触发条件：当Eden区满时，触发Minor GC。

Full GC触发条件：

（1）调用System.gc时，系统建议执行Full GC，但是不必然执行

（2）老年代空间不足

（3）方法区空间不足

（4）通过Minor GC后进入老年代的平均大小大于老年代的可用内存

（5）由Eden区、From Space区向To Space区复制时，对象大小大于To Space可用内存，则把该对象转存到老年代，且老年代的可用内存小于该对象大小

## 集合

### HashMap的实现原理

HashMap是基于哈希表的Map接口的非同步实现。此实现提供所有可选的映射操作，并允许使用null值和null键。此类不保证映射的顺序，特别是它不保证该顺序恒久不变。

#### HashMap的数据结构

在Java编程语言中，最基本的结构就是两种，一个是数组，另外一个是模拟指针（引用），所有的数据结构都可以用这两个基本结构来构造的，HashMap也不例外。HashMap实际上是一个“链表散列”的数据结构，即数组和链表的结合体。

HashMap底层就是一个数组结构，数组中的每一项又是一个链表。当新建一个HashMap的时候，就会初始化一个数组。

#### HashMap的存取实现

##### 存储

当我们往HashMap中put元素的时候，先根据key的hashCode重新计算hash值，根据hash值得到这个元素在数组中的位置（即下标）， 如果数组该位置上已经存放有其他元素了，那么在这个位置上的元素将以链表的形式存放，新加入的放在链头，最先加入的放在链尾。如果数组该位置上没有元素，就直接将该元素放到此数组中的该位置上。

##### 读取

从HashMap中get元素时，首先计算key的hashCode，找到数组中对应位置的某一元素，然后通过key的equals方法在对应位置的链表中找到需要的元素。

#### 归纳

简单地说，HashMap 在底层将 key-value 当成一个整体进行处理，这个整体就是一个 Entry 对象。HashMap 底层采用一个 Entry[] 数组来保存所有的 key-value 对，当需要存储一个 Entry 对象时，会根据hash算法来决定其在数组中的存储位置，在根据equals方法决定其在该数组位置上的链表中的存储位置；当需要取出一个Entry时，  
也会根据hash算法找到其在数组中的存储位置，再根据equals方法从该位置上的链表中取出该Entry。

### 队列(Queue)

#### BlockingQueue

BlockingQueue支持两个附加操作的Queue：1）当Queue为空时，获取元素线程被阻塞直到Queue变为非空；2）当Queue满时，添加元素线程被阻塞直到Queue不满。BlockingQueue不允许元素为null，如果入队一个null元素，会抛NullPointerException。常用于生产者消费者模式。

　　BlockingQueue对于不能满足条件的操作，提供了四种处理方式：

　　　　1）直接抛异常，抛出异常。如果队列已满，添加元素会抛出IllegalStateException异常；如果队列为空，获取元素会抛出NoSuchElementException异常；

　　　　2）返回一个特殊值（null或false）；

　　　　3）在满足条件之前，无限期的阻塞当前线程，当队列满足条件或响应中断退出；

　　　　4）在有限时间内阻塞当前线程，超时后返回失败。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 抛出异常 | 返回特殊值 | 阻塞 | 超时 |
| 入队 | add(e) | offer(e) | put(e) | offer(e,time,unit) |
| 出队 | remove() | poll(e) | take(e) | poll(time,unit) |
| 检查 | element() | peek() |  |  |

## IO

### BIO与NIO、AIO的区别

IO的方式通常分为几种，同步阻塞的BIO、同步非阻塞的NIO、异步非阻塞的AIO。

一、BIO

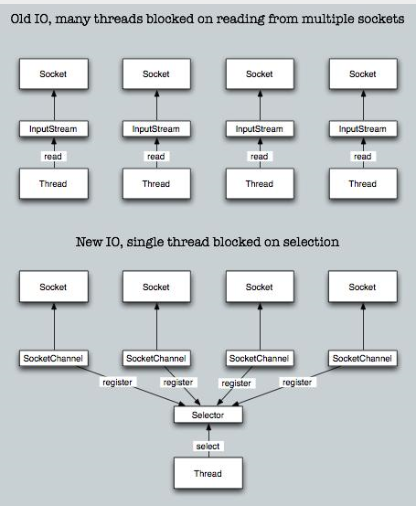
     在JDK1.4出来之前，我们建立网络连接的时候采用BIO模式，需要先在服务端启动一个ServerSocket，然后在客户端启动Socket来对服务端进行通信，默认情况下服务端需要对每个请求建立一堆线程等待请求，而客户端发送请求后，先咨询服务端是否有线程相应，如果没有则会一直等待或者遭到拒绝请求，如果有的话，客户端会线程会等待请求结束后才继续执行。

二、NIO

    NIO本身是基于事件驱动思想来完成的，其主要想解决的是BIO的大并发问题： 在使用同步I/O的网络应用中，如果要同时处理多个客户端请求，或是在客户端要同时和多个服务器进行通讯，就必须使用多线程来处理。也就是说，将每一个客户端请求分配给一个线程来单独处理。这样做虽然可以达到我们的要求，但同时又会带来另外一个问题。由于每创建一个线程，就要为这个线程分配一定的内存空间（也叫工作存储器），而且操作系统本身也对线程的总数有一定的限制。如果客户端的请求过多，服务端程序可能会因为不堪重负而拒绝客户端的请求，甚至服务器可能会因此而瘫痪。

    NIO基于Reactor，当socket有流可读或可写入socket时，操作系统会相应的通知引用程序进行处理，应用再将流读取到缓冲区或写入操作系统。  也就是说，这个时候，已经不是一个连接就要对应一个处理线程了，而是有效的请求，对应一个线程，当连接没有数据时，是没有工作线程来处理的。

   BIO与NIO一个比较重要的不同，是我们使用BIO的时候往往会引入多线程，每个连接一个单独的线程；而NIO则是使用单线程或者只使用少量的多线程，每个连接共用一个线程。



      NIO的最重要的地方是当一个连接创建后，不需要对应一个线程，这个连接会被注册到多路复用器上面，所以所有的连接只需要一个线程就可以搞定，当这个线程中的多路复用器进行轮询的时候，发现连接上有请求的话，才开启一个线程进行处理，也就是一个请求一个线程模式。

      在NIO的处理方式中，当一个请求来的话，开启线程进行处理，可能会等待后端应用的资源(JDBC连接等)，其实这个线程就被阻塞了，当并发上来的话，还是会有BIO一样的问题。

　　HTTP/1.1出现后，有了Http长连接，这样除了超时和指明特定关闭的http header外，这个链接是一直打开的状态的，这样在NIO处理中可以进一步的进化，在后端资源中可以实现资源池或者队列，当请求来的话，开启的线程把请求和请求数据传送给后端资源池或者队列里面就返回，并且在全局的地方保持住这个现场(哪个连接的哪个请求等)，这样前面的线程还是可以去接受其他的请求，而后端的应用的处理只需要执行队列里面的就可以了，这样请求处理和后端应用是异步的.当后端处理完，到全局地方得到现场，产生响应，这个就实现了异步处理。

三、AIO

     与NIO不同，当进行读写操作时，只须直接调用API的read或write方法即可。这两种方法均为异步的，对于读操作而言，当有流可读取时，操作系统会将可读的流传入read方法的缓冲区，并通知应用程序；对于写操作而言，当操作系统将write方法传递的流写入完毕时，操作系统主动通知应用程序。  即可以理解为，read/write方法都是异步的，完成后会主动调用回调函数。  在JDK1.7中，这部分内容被称作NIO.2，主要在java.nio.channels包下增加了下面四个异步通道：

* AsynchronousSocketChannel
* AsynchronousServerSocketChannel
* AsynchronousFileChannel
* AsynchronousDatagramChannel

其中的read/write方法，会返回一个带回调函数的对象，当执行完读取/写入操作后，直接调用回调函数。

BIO是一个连接一个线程。

NIO是一个请求一个线程。

AIO是一个有效请求一个线程。

先来个例子理解一下概念，以银行取款为例：

* 同步 ： 自己亲自出马持银行卡到银行取钱（使用同步IO时，Java自己处理IO读写）；
* 异步 ： 委托一小弟拿银行卡到银行取钱，然后给你（使用异步IO时，Java将IO读写委托给OS处理，需要将数据缓冲区地址和大小传给OS(银行卡和密码)，OS需要支持异步IO操作API）；
* 阻塞 ： ATM排队取款，你只能等待（使用阻塞IO时，Java调用会一直阻塞到读写完成才返回）；
* 非阻塞 ： 柜台取款，取个号，然后坐在椅子上做其它事，等号广播会通知你办理，没到号你就不能去，你可以不断问大堂经理排到了没有，大堂经理如果说还没到你就不能去（使用非阻塞IO时，如果不能读写Java调用会马上返回，当IO事件分发器会通知可读写时再继续进行读写，不断循环直到读写完成）

Java对BIO、NIO、AIO的支持：

* Java BIO ： 同步并阻塞，服务器实现模式为一个连接一个线程，即客户端有连接请求时服务器端就需要启动一个线程进行处理，如果这个连接不做任何事情会造成不必要的线程开销，当然可以通过线程池机制改善。
* Java NIO ： 同步非阻塞，服务器实现模式为一个请求一个线程，即客户端发送的连接请求都会注册到多路复用器上，多路复用器轮询到连接有I/O请求时才启动一个线程进行处理。
* Java AIO(NIO.2) ： 异步非阻塞，服务器实现模式为一个有效请求一个线程，客户端的I/O请求都是由OS先完成了再通知服务器应用去启动线程进行处理，

BIO、NIO、AIO适用场景分析:

* BIO方式适用于连接数目比较小且固定的架构，这种方式对服务器资源要求比较高，并发局限于应用中，JDK1.4以前的唯一选择，但程序直观简单易理解。
* NIO方式适用于连接数目多且连接比较短（轻操作）的架构，比如聊天服务器，并发局限于应用中，编程比较复杂，JDK1.4开始支持。
* AIO方式使用于连接数目多且连接比较长（重操作）的架构，比如相册服务器，充分调用OS参与并发操作，编程比较复杂，JDK7开始支持。

另外，I/O属于底层操作，需要操作系统支持，并发也需要操作系统的支持，所以性能方面不同操作系统差异会比较明显。

在高性能的I/O设计中，有两个比较著名的模式Reactor和Proactor模式，其中Reactor模式用于同步I/O，而Proactor运用于异步I/O操作。

    在比较这两个模式之前，我们首先的搞明白几个概念，什么是阻塞和非阻塞，什么是同步和异步,同步和异步是针对应用程序和内核的交互而言的，同步指的是用户进程触发IO操作并等待或者轮询的去查看IO操作是否就绪，而异步是指用户进程触发IO操作以后便开始做自己的事情，而当IO操作已经完成的时候会得到IO完成的通知。而阻塞和非阻塞是针对于进程在访问数据的时候，根据IO操作的就绪状态来采取的不同方式，说白了是一种读取或者写入操作函数的实现方式，阻塞方式下读取或者写入函数将一直等待，而非阻塞方式下，读取或者写入函数会立即返回一个状态值。

 一般来说I/O模型可以分为：同步阻塞，同步非阻塞，异步阻塞，异步非阻塞IO

同步阻塞IO：在此种方式下，用户进程在发起一个IO操作以后，必须等待IO操作的完成，只有当真正完成了IO操作以后，用户进程才能运行。JAVA传统的IO模型属于此种方式！

同步非阻塞IO:在此种方式下，用户进程发起一个IO操作以后边可返回做其它事情，但是用户进程需要时不时的询问IO操作是否就绪，这就要求用户进程不停的去询问，从而引入不必要的CPU资源浪费。其中目前JAVA的NIO就属于同步非阻塞IO。

异步阻塞IO：此种方式下是指应用发起一个IO操作以后，不等待内核IO操作的完成，等内核完成IO操作以后会通知应用程序，这其实就是同步和异步最关键的区别，同步必须等待或者主动的去询问IO是否完成，那么为什么说是阻塞的呢？因为此时是通过select系统调用来完成的，而select函数本身的实现方式是阻塞的，而采用select函数有个好处就是它可以同时监听多个文件句柄，从而提高系统的并发性！

 异步非阻塞IO:在此种模式下，用户进程只需要发起一个IO操作然后立即返回，等IO操作真正的完成以后，应用程序会得到IO操作完成的通知，此时用户进程只需要对数据进行处理就好了，不需要进行实际的IO读写操作，因为真正的IO读取或者写入操作已经由内核完成了。目前Java中还没有支持此种IO模型。

1,同步和异步是针对应用程序和内核的交互而言的。   
  
2,阻塞和非阻塞是针对于进程在访问数据的时候，根据IO操作的就绪状态来采取的不同方式，说白了是一种读取或者写入操作函数的实现方式，阻塞方式下读取或者写入函数将一直等待，而非阻塞方式下，读取或者写入函数会立即返回一个状态值。    
  
由上描述基本可以总结一句简短的话，同步和异步是目的，阻塞和非阻塞是实现方式。   
  
编号 名词 解释 举例   
1 同步 指的是用户进程触发IO操作并等待或者轮询的去查看IO操作是否就绪 自己上街买衣服，自己亲自干这件事，别的事干不了。   
2 异步 异步是指用户进程触发IO操作以后便开始做自己的事情，而当IO操作已经完成的时候会得到IO完成的通知（异步的特点就是通知） 告诉朋友自己合适衣服的尺寸，大小，颜色，让朋友委托去卖，然后自己可以去干别的事。（使用异步IO时，Java将IO读写委托给OS处理，需要将数据缓冲区地址和大小传给OS）   
3 阻塞 所谓阻塞方式的意思是指, 当试图对该文件描述符进行读写时, 如果当时没有东西可读,或者暂时不可写, 程序就进入等待 状态, 直到有东西可读或者可写为止 去公交站充值，发现这个时候，充值员不在（可能上厕所去了），然后我们就在这里等待，一直等到充值员回来为止。（当然现实社会，可不是这样，但是在计算机里确实如此。）   
4 非阻塞 非阻塞状态下, 如果没有东西可读, 或者不可写, 读写函数马上返回, 而不会等待， 银行里取款办业务时，领取一张小票，领取完后我们自己可以玩玩手机，或者与别人聊聊天，当轮我们时，银行的喇叭会通知，这时候我们就可以去了。

## 多线程(Thread)

### 进程与线程的区别

进程：每个进程都有独立的代码和数据空间（进程上下文），进程间的切换会有较大的开销，一个进程包含1--n个线程。（进程是资源分配的最小单位）。

线程：同一类线程共享代码和数据空间，每个线程有独立的运行栈和程序计数器(PC)，线程切换开销小。（线程是cpu调度的最小单位）。

线程和进程一样分为五个阶段：创建、就绪、运行、阻塞、终止。

多进程是指操作系统能同时运行多个任务（程序）。

多线程是指在同一程序中有多个顺序流在执行。

### Thread和Runnable的区别

如果一个类继承Thread，则不适合资源共享。但是如果实现了Runable接口的话，则很容易的实现资源共享。

总结：

实现Runnable接口比继承Thread类所具有的优势：

适合多个相同的程序代码的线程去处理同一个资源。

可以避免java中的单继承的限制。

增加程序的健壮性，代码可以被多个线程共享，代码和数据独立。

线程池只能放入实现Runable或Callable类线程，不能直接放入继承Thread的类。

### 线程状态转换



**新建状态（New）**：新创建了一个线程对象。

**就绪状态（Runnable）**：线程对象创建后，其他线程调用了该对象的start()方法。该状态的线程位于可运行线程池中，变得可运行，等待获取CPU的使用权。

**运行状态（Running）**：就绪状态的线程获取了CPU，执行程序代码。

**阻塞状态（Blocked）**：阻塞状态是线程因为某种原因放弃CPU使用权，暂时停止运行。直到线程进入就绪状态，才有机会转到运行状态。阻塞的情况分三种：

（一）、**等待阻塞**：运行的线程执行wait()方法，JVM会把该线程放入等待池中。(wait会释放持有的锁)

（二）、**同步阻塞**：运行的线程在获取对象的同步锁时，若该同步锁被别的线程占用，则JVM会把该线程放入锁池中。

（三）、**其他阻塞**：运行的线程执行sleep()或join()方法，或者发出了I/O请求时，JVM会把该线程置为阻塞状态。当sleep()状态超时、join()等待线程终止或者超时、或者I/O处理完毕时，线程重新转入就绪状态。（注意,sleep是不会释放持有的锁）

**死亡状态（Dead）**：线程执行完了或者因异常退出了run()方法，该线程结束生命周期。

### synchronized和volatile

synchronized关键字可以实现操作的原子性，其本质是通过该关键字所包括的临界区（Critical Section）的排他性保证在任何一个时刻只有一个线程能够执行临界区中的代码，这使得临界区中的代码代表了一个原子操作。这一点，读者可能已经很清楚。但是，synchronized关键字所起的另外一个作用——保证内存的可见性（Memory Visibility）线程调度。

CPU在执行代码的时候，为了减少变量访问的时间消耗可能将代码中访问的变量的值缓存到该CPU的缓存区（如L1 Cache、L2 Cache等）中。因此相应代码再次访问某个变量时，相应的值可能是从CPU缓存区而不是主内存中读取的。同样地，代码对这些被缓存过的变量的值的修改也可能仅是被写入CPU缓存区，而没有被写回主内存。由于每个CPU都有自己的缓存区，因此一个CPU缓存区中的内容对于其他CPU而言是不可见的。这就导致了在其他CPU上运行的其他线程可能无法“看到”该线程对某个变量值所做的更改。这就是所谓的内存可见性。

synchronized关键字的另外一个作用就是它保证了一个线程执行临界区中的代码时所修改的变量值对于稍后执行该临界区中的代码的线程来说是可见的。这对于保证多线程代码的正确性来说非常重要。

而volatile关键字也能够保证内存可见性。即一个线程对一个采用volatile关键字修饰的变量的值的更改对于其他访问该变量的线程而言总是可见的。也就是说，其他线程不会读到一个“过期”的变量值。因此，有人将volatile关键字与synchronized关键字所代表的内部锁做比较，将其称为轻量级的锁。这种称呼其实并不恰当，volatile关键字只能保证内存可见性，它并不能像synchronized关键字所代表的内部锁那样能够保证操作的原子性。volatile关键字实现内存可见性的核心机制是当一个线程修改了一个volatile修饰的变量的值时，该值会被写入主内存（即RAM）而不仅仅是当前线程所在的CPU的缓存区，而其他CPU的缓存区中存储的该变量的值也会因此而失效（从而得以更新为主内存中该变量的相应值）。这就保证了其他线程访问该volatile修饰的变量时总是可以获取该变量的最新值。

volatile关键字的另外一个作用是它禁止了指令重排序（Re-order）。编译器和CPU为了提高指令的执行效率可能会进行指令重排序，这使得代码的实际执行方式可能不是按照我们所认为的方式进行的。

#### 线程优先级

Java线程有优先级，优先级高的线程会获得较多的运行机会。

Java线程的优先级用整数表示，取值范围是1~10，Thread类有以下三个静态常量：

MAX\_PRIORITY

          ##线程可以具有的最高优先级，取值为10。

MIN\_PRIORITY

          ##线程可以具有的最低优先级，取值为1。

NORM\_PRIORITY

          ##分配给线程的默认优先级，取值为5。

#### Thread.sleep

线程睡眠，使线程转到阻塞状态。millis参数设定睡眠的时间，以毫秒为单位。当睡眠结束后，就转为就绪（Runnable）状态，**不释放锁**。

#### Object.wait

线程等待，导致当前的线程等待，直到其他线程调用此对象的 notify() 方法或 notifyAll() 唤醒方法。这个两个唤醒方法也是Object类中的方法。**wait释放锁，必须在对象锁代码块里调用,否则会抛出java.lang.IllegalMonitorStateException**。

#### Object.notify(Object.notifyAll)

Object类中的notify()方法，唤醒在此对象监视器上等待的单个线程。如果所有线程都在此对象上等待，则会选择唤醒其中一个线程。选择是任意性的，并在对实现做出决定时发生。线程通过调用其中一个 wait 方法，在对象的监视器上等待。 直到当前的线程放弃此对象上的锁定，才能继续执行被唤醒的线程。被唤醒的线程将以常规方式与在该对象上主动同步的其他所有线程进行竞争；例如，唤醒的线程在作为锁定此对象的下一个线程方面没有可靠的特权或劣势。类似的方法还有一个notifyAll()，唤醒在此对象监视器上等待的所有线程。**notify和notifyAll()必须在对象锁代码块里调用,否则会抛出java.lang.IllegalMonitorStateException**。

#### Thread.yield

线程让步，暂停当前正在执行的线程对象，把执行机会让给相同或者更高优先级的线程。

yield应该做的是让当前运行线程回到可运行状态，以允许具有相同优先级的其他线程获得运行机会。因此，使用yield()的目的是让相同优先级的线程之间能适当的轮转执行。但是，实际中无法保证yield()达到让步目的，因为让步的线程还有可能被线程调度程序再次选中。

#### Thread.join

线程加入，等待其他线程终止。在当前线程中调用另一个线程的join()方法，则当前线程转入阻塞状态，直到另一个进程运行结束，当前线程再由阻塞转为就绪状态。

join是Thread类的一个方法，启动线程后直接调用，即join()的作用是：“等待该线程终止”，这里需要理解的就是该线程是指的主线程等待子线程的终止。也就是在子线程调用了join()方法后面的代码，只有等到子线程结束了才能执行。

#### Thread.**interrupt**

不要以为它是中断某个线程！它只是向线程发送一个中断信号，让线程在无限等待时（如死锁时）能抛出异常，从而结束线程，但是如果你吃掉了这个异常，那么这个线程还是不会中断的！

#### **wait和sleep区别**

共同点：

他们都是在多线程的环境下，都可以在程序的调用处阻塞指定的毫秒数，并返回。

wait()和sleep()都可以通过interrupt()方法 打断线程的暂停状态 ，从而使线程立刻抛出InterruptedException。

如果线程A希望立即结束线程B，则可以对线程B对应的Thread实例调用interrupt方法。如果此刻线程B正在wait/sleep/join，则线程B会立刻抛出InterruptedException，在catch() {} 中直接return即可安全地结束线程。

需要注意的是，InterruptedException是线程自己从内部抛出的，并不是interrupt()方法抛出的。对某一线程调用 interrupt()时，如果该线程正在执行普通的代码，那么该线程根本就不会抛出InterruptedException。但是，一旦该线程进入到 wait()/sleep()/join()后，就会立刻抛出InterruptedException 。

不同点：

Thread类的方法：sleep(),yield()等

Object的方法：wait()和notify()等

每个对象都有一个锁来控制同步访问。Synchronized关键字可以和对象的锁交互，来实现线程的同步。

sleep方法没有释放锁，而wait方法释放了锁，使得其他线程可以使用同步控制块或者方法。

wait，notify和notifyAll只能在同步控制方法或者同步控制块里面使用，而sleep可以在任何地方使用

sleep，wait必须捕获异常，而notify和notifyAll不需要捕获异常。

所以sleep()和wait()方法的最大区别是：

　　　　sleep()睡眠时，保持对象锁，仍然占有该锁；

　　　　而wait()睡眠时，释放对象锁。

　　但是wait()和sleep()都可以通过interrupt()方法打断线程的暂停状态，从而使线程立刻抛出InterruptedException（但不建议使用该方法）。

sleep()方法

sleep()使当前线程进入停滞状态（阻塞当前线程），让出CUP的使用、目的是不让当前线程独自霸占该进程所获的CPU资源，以留一定时间给其他线程执行的机会;

　　 sleep()是Thread类的Static(静态)的方法；因此他不能改变对象的机锁，所以当在一个Synchronized块中调用Sleep()方法是，线程虽然休眠了，但是对象的机锁并木有被释放，其他线程无法访问这个对象（即使睡着也持有对象锁）。

　　在sleep()休眠时间期满后，该线程不一定会立即执行，这是因为其它线程可能正在运行而且没有被调度为放弃执行，除非此线程具有更高的优先级。

wait（）方法

wait()方法是Object类里的方法；当一个线程执行到wait()方法时，它就进入到一个和该对象相关的等待池中，同时失去（释放）了对象的机锁（暂时失去机锁，wait(long timeout)超时时间到后还需要返还对象锁）；其他线程可以访问；

　　wait()使用notify或者notifyAlll或者指定睡眠时间来唤醒当前等待池中的线程。

　　wiat()必须放在synchronized block中，否则会在program runtime时扔出java.lang.IllegalMonitorStateException异常。

### 常见线程名词解释

#### 主线程

JVM调用程序main()所产生的线程。

#### 当前线程

这个是容易混淆的概念。一般指通过Thread.currentThread()来获取的进程。

#### 后台线程（守护线程）

指为其他线程提供服务的线程，也称为守护线程。JVM的垃圾回收线程就是一个后台线程。**用户线程和守护线程的区别在于，是否等待主线程依赖于主线程结束而结束**

#### 前台线程

是指接受后台线程服务的线程，其实前台后台线程是联系在一起，就像傀儡和幕后操纵者一样的关系。傀儡是前台线程、幕后操纵者是后台线程。由前台线程创建的线程默认也是前台线程。可以通过isDaemon()和setDaemon()方法来判断和设置一个线程是否为后台线程。

### 线程池

#### ThreadPoolExecutor

##### 构造方法参数说明

corePoolSize：核心池的大小。

这个参数跟后面讲述的线程池的实现原理有非常大的关系。在创建了线程池后，默认情况下，线程池中并没有任何线程，而是等待有任务到来才创建线程去执行任务，除非调用了prestartAllCoreThreads()或者prestartCoreThread()方法，从这2个方法的名字就可以看出，是预创建线程的意思，即在没有任务到来之前就创建corePoolSize个线程或者一个线程。默认情况下，在创建了线程池后，线程池中的线程数为0，当有任务来之后，就会创建一个线程去执行任务，当线程池中的线程数目达到corePoolSize后，就会把到达的任务放到缓存队列当中；

maximumPoolSize：线程池最大线程数。

这个参数也是一个非常重要的参数，它表示在线程池中最多能创建多少个线程；

**keepAliveTime**：表示线程没有任务执行时最多保持多久时间会终止。

默认情况下，只有当线程池中的线程数大于corePoolSize时，keepAliveTime才会起作用，直到线程池中的线程数不大于corePoolSize，即当线程池中的线程数大于corePoolSize时，如果一个线程空闲的时间达到keepAliveTime，则会终止，直到线程池中的线程数不超过corePoolSize。但是如果调用了allowCoreThreadTimeOut(boolean)方法，在线程池中的线程数不大于corePoolSize时，keepAliveTime参数也会起作用，直到线程池中的线程数为0；

**unit**：参数keepAliveTime的时间单位。

有7种取值，在TimeUnit类中有7种静态属性：

TimeUnit.DAYS; //天

TimeUnit.HOURS; //小时

TimeUnit.MINUTES; //分钟

TimeUnit.SECONDS; //秒

TimeUnit.MILLISECONDS; //毫秒

TimeUnit.MICROSECONDS; //微妙

TimeUnit.NANOSECONDS; //纳秒

workQueue：一个阻塞队列，用来存储等待执行的任务。

这个参数的选择也很重要，会对线程池的运行过程产生重大影响，一般来说，这里的阻塞队列有以下几种选择：

ArrayBlockingQueue;

PriorityBlockingQueue;

LinkedBlockingQueue;

SynchronousQueue;

ArrayBlockingQueue和PriorityBlockingQueue使用较少，一般使用LinkedBlockingQueue和SynchronousQueue。线程池的排队策略与BlockingQueue有关。

threadFactory：线程工厂，主要用来创建线程。

handler：表示当拒绝处理任务时的策略。

有以下四种取值：

ThreadPoolExecutor.AbortPolicy：丢弃任务并抛出RejectedExecutionException异常。

ThreadPoolExecutor.DiscardPolicy：也是丢弃任务，但是不抛出异常。

ThreadPoolExecutor.DiscardOldestPolicy：丢弃队列最前面的任务，然后重新尝试执行任务（重复此过程）。

ThreadPoolExecutor.CallerRunsPolicy：由调用线程处理该任务。

#### ThreadPoolExecutor继承关系

ThreadPoolExecutor(类)<- AbstractExecutorService(抽象类)<- ExecutorService(接口)<- Executor(接口)

Executor是一个顶层接口，在它里面只声明了一个方法execute(Runnable)，返回值为void，参数为Runnable类型，从字面意思可以理解，就是用来执行传进去的任务的；

ExecutorService接口继承了Executor接口，并声明了一些方法：submit、invokeAll、invokeAny以及shutDown等；

抽象类AbstractExecutorService实现了ExecutorService接口，基本实现了ExecutorService中声明的所有方法；

#### 线程池方法说明

##### execute()

execute()方法实际上是Executor中声明的方法，在ThreadPoolExecutor进行了具体的实现，这个方法是ThreadPoolExecutor的核心方法，通过这个方法可以向线程池提交一个任务，交由线程池去执行。

##### submit()

submit()方法是在ExecutorService中声明的方法，在AbstractExecutorService就已经有了具体的实现，在ThreadPoolExecutor中并没有对其进行重写，这个方法也是用来向线程池提交任务的，但是它和execute()方法不同，它能够返回任务执行的结果，去看submit()方法的实现，会发现它实际上还是调用的execute()方法，只不过它利用了Future来获取任务执行结果。

##### shutdown()

关闭线程池

##### shutdownNow()

马上关闭线程池

##### getQueue()

##### getPoolSize()

##### getActiveCount()

##### getCompletedTaskCount()

#### 线程池状态

在ThreadPoolExecutor中定义了一个volatile变量，另外定义了几个static final变量表示线程池的各个状态：

volatile int runState;

static final int **RUNNING** = 0;

static final int **SHUTDOWN** = 1;

static final int **STOP** = 2;

static final int **TERMINATED** = 3;

runState表示当前线程池的状态，它是一个volatile变量用来保证线程之间的可见性；

下面的几个static final变量表示runState可能的几个取值。

当创建线程池后，初始时，线程池处于**RUNNING**状态；

如果调用了shutdown()方法，则线程池处于**SHUTDOWN**状态，此时线程池不能够接受新的任务，它会等待所有任务执行完毕；

如果调用了shutdownNow()方法，则线程池处于**STOP**状态，此时线程池不能接受新的任务，并且会去尝试终止正在执行的任务；

当线程池处于**SHUTDOWN**或**STOP**状态，并且所有工作线程已经销毁，任务缓存队列已经清空或执行结束后，线程池被设置为**TERMINATED**状态。

## spring

### Spring Bean的生命周期

1. 实例化一个Bean，也就是我们通常说的new

2. 按照Spring上下文对实例化的Bean进行配置，也就是IOC注入

3. 如果这个Bean实现了BeanNameAware接口，会调用它实现的setBeanName(String beanName)方法，此处传递的是Spring配置文件中Bean的ID

4. 如果这个Bean实现了BeanFactoryAware接口，会调用它实现的setBeanFactory(BeanFactory beanFactory)，传递的是Spring工厂本身（可以用这个方法获取到其他Bean）

5. 如果这个Bean实现了ApplicationContextAware接口，会调用setApplicationContext(ApplicationContext)方法，传入Spring上下文，该方式同样可以实现步骤4，但比4更好，以为ApplicationContext是BeanFactory的子接口，有更多的实现方法

6. 如果这个Bean关联了BeanPostProcessor接口，将会调用postProcessBeforeInitialization(Object obj, String s)方法，BeanPostProcessor经常被用作是Bean内容的更改，并且由于这个是在Bean初始化结束时调用After方法，也可用于内存或缓存技术

7. 如果这个Bean在Spring配置文件中配置了init-method属性会自动调用其配置的初始化方法

8. 如果这个Bean关联了BeanPostProcessor接口，将会调用postAfterInitialization(Object obj, String s)方法

注意：以上工作完成以后就可以用这个Bean了，那这个Bean是一个single的，所以一般情况下我们调用同一个ID的Bean会是在内容地址相同的实例

9. 当Bean不再需要时，会经过清理阶段，如果Bean实现了DisposableBean接口，会调用其实现的destroy方法

10. 最后，如果这个Bean的Spring配置中配置了destroy-method属性，会自动调用其配置的销毁方法

以上10步骤可以作为面试或者笔试的模板，另外我们这里描述的是应用Spring上下文Bean的生命周期，如果应用Spring的工厂也就是BeanFactory的话去掉第5步就Ok了

### @Autowired

默认是按照byType进行注入的，如果有多个，按照byName方式比对，如果还有多个，则报出异常

## Spring Boot

### 最简单的SpringBoot Web工程配置

#### pom.xml

<parent>

<groupId>org.springframework.boot</groupId>

<artifactId>spring-boot-starter-parent</artifactId>

<version>1.5.6.RELEASE</version>

<relativePath />

</parent>

<properties>

<project.build.sourceEncoding>UTF-8</project.build.sourceEncoding>

<project.reporting.outputEncoding>UTF-8</project.reporting.outputEncoding>

<java.version>1.8</java.version>

</properties>

<dependencies>

<!-- 上边引入 parent，因此 下边无需指定版本 -->

<dependency>

<groupId>org.springframework.boot</groupId>

<artifactId>spring-boot-starter-web</artifactId>

</dependency>

</dependencies>

<build>

<plugins>

<plugin>

<groupId>org.springframework.boot</groupId>

<artifactId>spring-boot-maven-plugin</artifactId>

</plugin>

</plugins>

</build>

#### Controller

@RestController

public class TestController {

@GetMapping("/helloworld")

public String helloworld() {

return "helloworld";

}

}

#### SpringbootApplication

@SpringBootApplication

public class SpringbootApplication {

public static void main(String[] args) {

SpringApplication.run(SpringbootApplication.class, args);

}

}

### 配置详解

#### application.properties

##### server.port=8888

指定端口

##### spring.application.name=hello

指定应用名(SpringCloud中会被注册为服务名)

#### application.yml

environments:

dev:

url: http://dev.bar.com

name: Developer Setup

prod:

url: http://foo.bar.com

name: My Cool App

server:

port:8881

---

spring:

profiles:test

server:

port:8882

---

spring:

profiles:prod

server:

port:8883

#### 自定义参数

除了可以在Spring Boot的配置文件中设置各个Starter模块中预定义的配置属性，也可以在配置文件中定义一些我们需要的自定义属性。比如在application.properties中添加：

book.name=SpringCloudInAction

book.author=ZhaiYongchao

然后，在应用中可以通过@Value注解来加载这些自定义的参数，比如：

@Component

public class Book {

@Value("${book.name}")

private String name;

@Value("${book.author}")

private String author;

// 省略getter和setter

}

@Value注解加载属性值的时候可以支持两种表达式来进行配置，如下所示。

·一种是上面介绍的PlaceHolder方式，格式为${...}，大括号内为PlaceHolder。

·另一种是使用SpEL表达式（Spring Expression Language），格式为#{...}，大括号内为SpEL表达式。

#### 参数引用

在application.properties中的各个参数之间可以直接通过使用PlaceHolder的方式来进行引用，就像下面的设置：

book.name=SpringCloud

book.author=ZhaiYongchao

book.desc=${book.author} is writing《${book.name}》

book.desc参数引用了上文中定义的book.name和book.author属性，最后该属性的值就是ZhaiYongchao is writing《SpringCloud》。

#### 使用随机数

在一些特殊情况下，我们希望有些参数每次被加载的时候不是一个固定的值，比如密钥、服务端口等。在Spring Boot的属性配置文件中，可以通过使用${random}配置来产生随机的int值、long值或者string字符串，这样我们就可以容易地通过配置随机生成属性，而不是在程序中通过编码来实现这些逻辑。

${random}的配置方式主要有以下几种，读者可作为参考使用。

# 随机字符串

com.didispace.blog.value=${random.value}

# 随机int

com.didispace.blog.number=${random.int}

# 随机long

com.didispace.blog.bignumber=${random.long}

# 10以内的随机数

com.didispace.blog.test1=${random.int(10)}

# 10~20的随机数

com.didispace.blog.test2=${random.int[10,20]}

该配置方式可以设置应用端口等场景，以避免在本地调试时出现端口冲突的麻烦。

#### 命令行参数

使用命令java -jar来启动的方式。该命令除了启动应用之外，还可以在命令行中指定应用的参数，比如java -jar xxx.jar --server.port=8888，直接以命令行的方式来设置server.port属性，并将启动应用的端口设为8888。

在用命令行方式启动Spring Boot应用时，连续的两个减号--就是对application.properties中的属性值进行赋值的标识。所以，java -jar xxx.jar--server.port=8888命令，等价于在application.properties中添加属性server.port=8888。

通过命令行来修改属性值是Spring Boot非常重要的一个特性。通过此特性，理论上已经使得应用的属性在启动前是可变的，所以其中的端口号也好、数据库连接也好，都是可以在应用启动时发生改变的，而不同于以往的Spring应用通过Maven的Profile在编译器中进行不同环境的构建。Spring Boot的这种方式，可以让应用程序的打包内容贯穿开发、测试以及线上部署，而Maven不同Profile的方案为每个环境所构建的包，其内容本质上是不同的。但是，如果每个参数都需要通过命令行来指定，这显然也不是一个好的方案，所以下面我们看看如何在Spring Boot中实现多环境的配置。

#### 多环境配置

我们在开发应用的时候，通常同一套程序会被应用和安装到几个不同的环境中，比如开发、测试、生产等。其中每个环境的数据库地址、服务器端口等配置都不同，如果在为不同环境打包时都要频繁修改配置文件的话，那必将是个非常烦琐且容易发生错误的事。

对于多环境的配置，各种项目构建工具或是框架的基本思路是一致的，通过配置多份不同环境的配置文件，再通过打包命令指定需要打包的内容之后进行区分打包，Spring Boot也不例外，或者说实现起来更加简单。

在Spring Boot中，多环境配置的文件名需要满足application-{profile}.properties的格式，其中{profile}对应你的环境标识，如下所示。

·application-dev.properties：开发环境。

·application-test.properties：测试环境。

·application-prod.properties：生产环境。至于具体哪个配置文件会被加载，需要在application.properties文件中通过spring.profiles.active属性来设置，其值对应配置文件中的{profile}值。如spring.profiles.active=test就会加载application-test.properties配置文件内容。

下面，以不同环境配置不同的服务端口为例，进行样例实验。

·针对各环境新建不同的配置文件application-dev.properties、application-test.properties、application-prod.properties。

·在这三个文件中均设置不同的server.port属性，例如，dev环境设置为1111，test环境设置为2222，prod环境设置为3333。

·application.properties中设置spring.profiles.active=dev，意为默认以dev环境设置。

·测试不同配置的加载。

·执行java -jar xxx.jar，可以观察到服务端口被设置为1111，也就是默认的开发环境（dev）。·执行java -jar xxx.jar --spring.profiles.active=test，可以观察到服务端口被设置为2222，也就是测试环境的配置（test）。

·执行java -jar xxx.jar --spring.profiles.active=prod，可以观察到服务端口被设置为3333，也就是生产环境的配置（prod）。

按照上面的实验，可以如下总结多环境的配置思路：

·在application.properties中配置通用内容，并设置spring.profiles.active=dev，以开发环境为默认配置。

·在application-{profile}.properties中配置各个环境不同的内容。

·通过命令行方式去激活不同环境的配置。

#### 加载顺序

在上面的例子中，我们将Spring Boot应用需要的配置内容都放在了项目工程中，已经能够通过spring.profiles.active或是通过Maven来实现多环境的支持。但是，当团队逐渐壮大，分工越来越细致之后，往往不需要让开发人员知道测试或是生产环境的细节，而是希望由每个环境各自的负责人（QA或是运维）来集中维护这些信息。那么如果还是以这样的方式存储配置内容，对于不同环境配置的修改就不得不去获取工程内容来修改这些配置内容，当应用非常多的时候就变得非常不方便。同时，配置内容对开发人员都可见，这本身也是一种安全隐患。对此，出现了很多将配置内容外部化的框架和工具，后续将要介绍的Spring Cloud Config就是其中之一，为了后续能更好地理解Spring Cloud Config的加载机制，我们需要对Spring Boot对数据文件的加载机制有一定的了解。

为了能够更合理地重写各属性的值，Spring Boot使用了下面这种较为特别的属性加载顺序：

1. 在命令行中传入的参数。

2. SPRING\_APPLICATION\_JSON中的属性。SPRING\_APPLICATION\_JSON是以JSON格式配置在系统环境变量中的内容。

3. java:comp/env中的JNDI属性。

4. Java的系统属性，可以通过System.getProperties()获得的内容。5. 操作系统的环境变量。

6. 通过random.\*配置的随机属性。

7. 位于当前应用jar包之外，针对不同{profile}环境的配置文件内容，例如application-{profile}.properties或是YAML定义的配置文件。

8. 位于当前应用jar包之内，针对不同{profile}环境的配置文件内容，例如application-{profile}.properties或是YAML定义的配置文件。

9. 位于当前应用jar包之外的application.properties和YAML配置内容。

10. 位于当前应用jar包之内的application.properties和YAML配置内容。

11. 在@Configuration注解修改的类中，通过@PropertySource注解定义的属性。

12. 应用默认属性，使用SpringApplication.setDefaultProperties定义的内容。

优先级按上面的顺序由高到低，数字越小优先级越高。

可以看到，其中第7项和第9项都是从应用jar包之外读取配置文件，所以，实现外部化配置的原理就是从此切入，为其指定外部配置文件的加载位置来取代jar包之内的配置内容。通过这样的实现，我们的工程在配置中就变得非常干净，只需在本地放置开发需要的配置即可，而不用关心其他环境的配置，由其对应环境的负责人去维护即可。

### 监控与管理

#### 开启HTTP security认证

<dependency>

**<groupId>**org.springframework.boot**</groupId>**

**<artifactId>**spring-boot-starter-security**</artifactId>**

</dependency>

application.properties

security.user.name=admin

security.user.password=123456

management.security.enabled=true

management.security.role=ADMIN

#### spring-boot-starter-actuator

##### ****关闭验证****

默认情况下很多端点是不允许访问的，会返回401:Unauthorized。

application.properties

management.security.enabled=false

##### /health

{

"status" : "UP",

"diskSpace" : {

"status" : "UP",

"total" : 107374178304,

"free" : 19108171776,

"threshold" : 10485760

}

}

##### /autoconfig

## SpringCloud

spring cloud 为开发人员提供了快速构建分布式系统的一些工具，包括配置管理、服务发现、断路器、路由、微代理、事件总线、全局锁、决策竞选、分布式会话等等。

### 服务治理

**服务治理**可以说是微服务架构中最为核心和基础的模块，它主要用来实现**各个微服务实例的自动化注册与发现**。为什么我们在微服务架构中那么需要服务治理模块呢？微服务系统没有它会有什么不好的地方吗？

在最初开始构建微服务系统的时候可能服务并不多，我们可以通过做一些静态配置来完成服务的调用。比如，有两个服务A和B，其中服务A需要调用服务B来完成一个业务操作时，为了实现服务B的高可用，不论采用服务端负载均衡还是客户端负载均衡，都需要手工维护服务B的具体实例清单。但是随着业务的发展，系统功能越来越复杂，相应的微服务应用也不断增加，我们的静态配置就会变得越来越难以维护。并且面对不断发展的业务，我们的集群规模、服务的位置、服务的命名等都有可能发生变化，如果还是通过手工维护的方式，那么极易发生错误或是命名冲突等问题。同时，对于这类静态内容的维护也必将消耗大量的人力。

为了解决微服务架构中的服务实例维护问题，产生了大量的服务治理框架和产品。这些框架和产品的实现都围绕着服务注册与服务发现机制来完成对微服务应用实例的自动化管理。

#### 服务注册

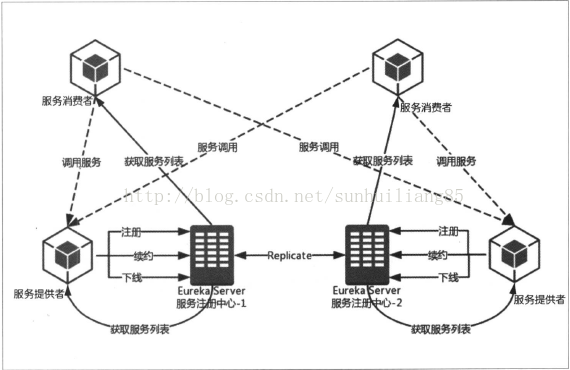
在服务治理框架中，通常都会构建一个注册中心，每个服务单元向注册中心登记自己提供的服务，将主机与端口号、版本号、通信协议等一些附加信息告知注册中心，注册中心按服务名分类组织服务清单。比如，我们有两个提供服务A的进程分别运行于192.168.0.100:8000和192.168.0.101:8000位置上，另外还有三个提供服务B的进程分别运行于192.168.0.100:9000、192.168.0.101:9000、192.168.0.102:9000位置上。当这些进程均启动，并向注册中心注册自己的服务之后，注册中心就会维护类似下面的一个服务清单。另外，服务注册中心还需要以心跳的方式去监测清单中的服务是否可用，若不可用需要从服务清单中剔除，达到排除故障服务的效果。

#### 服务发现

由于在服务治理框架下运作，服务间的调用不再通过指定具体的实例地址来实现，而是通过向服务名发起请求调用实现。所以，服务调用方在调用服务提供方接口的时候，并不知道具体的服务实例位置。因此，调用方需要向服务注册中心咨询服务，并获取所有服务的实例清单，以实现对具体服务实例的访问。

比如，现有服务C希望调用服务A，服务C就需要向注册中心发起咨询服务请求，服务注册中心就会将服务A的位置清单返回给服务C，如按上例服务A的情况，C便获得了服务A的两个可用位置192.168.0.100:8000和192.168.0.101:8000。当服务C要发起调用的时候，便从该清单中以某种轮询策略取出一个位置来进行服务调用，这就是后续我们将会介绍的客户端负载均衡。这里我们只是列举了一种简单的服务治理逻辑，以方便理解服务治理框架的基本运行思路。

实际的框架为了性能等因素，不会采用每次都向服务注册中心获取服务的方式，并且不同的应用场景在缓存和服务剔除等机制上也会有一些不同的实现策略。



#### Spring Cloud Eureka

Spring Cloud Eureka是Spring Cloud Netflix微服务套件中的一部分，它基于Netflix Eureka做了二次封装，主要负责完成微服务架构中的服务治理功能。Spring Cloud通过为Eureka增加了Spring Boot风格的自动化配置，我们只需通过简单引入依赖和注解配置就能让Spring Boot构建的微服务应用轻松地与Eureka服务治理体系进行整合。

Spring Cloud Eureka，使用Netflix Eureka来实现服务注册与发现，它既包含了服务端组件，也包含了客户端组件，并且服务端与客户端均采用Java编写，所以Eureka主要适用于通过Java实现的分布式系统，或是与JVM兼容语言构建的系统。但是，由于Eureka服务端的服务治理机制提供了完备的RESTful API，所以它也支持将非Java语言构建的微服务应用纳入Eureka的服务治理体系中来。只是在使用其他语言平台的时候，需要自己来实现Eureka的客户端程序。不过庆幸的是，在目前几个较为流行的开发平台上，都已经有了一些针对Eureka注册中心的客户端实现框架，比如.NET平台的Steeltoe、Node.js的eureka-js-client等。

##### Eureka服务端

我们也称为服务注册中心。它同其他服务注册中心一样，支持高可用配置。它依托于强一致性提供良好的服务实例可用性，可以应对多种不同的故障场景。如果Eureka以集群模式部署，当集群中有分片出现故障时，那么Eureka就转入自我保护模式。它允许在分片故障期间继续提供服务的发现和注册，当故障分片恢复运行时，集群中的其他分片会把它们的状态再次同步回来。以在AWS上的实践为例，Netflix推荐每个可用的区域运行一个Eureka服务端，通过它来形成集群。不同可用区域的服务注册中心通过异步模式互相复制各自的状态，这意味着在任意给定的时间点每个实例关于所有服务的状态是有细微差别的。

##### Eureka客户端

主要处理服务的注册与发现。客户端服务通过注解和参数配置的方式，嵌入在客户端应用程序的代码中，在应用程序运行时，Eureka客户端向注册中心注册自身提供的服务并周期性地发送心跳来更新它的服务租约。同时，它也能从服务端查询当前注册的服务信息并把它们缓存到本地并周期性地刷新服务状态。

##### 搭建注册服务中心

首先，创建一个基础的Spring Boot工程，命名为eureka-server，并在pom.xml中引入必要的依赖内容，代码如下：<parent>

<groupId>org.springframework.boot</groupId>

<artifactId>spring-boot-starter-parent</artifactId>

<version>1.3.7.RELEASE</version>

<relativePath/>

</parent>

<dependencies>

<dependency>

<groupId>org.springframework.cloud</groupId>

<artifactId>spring-cloud-starter-eureka-server</artifactId>

</dependency>

</dependencies>

<dependencyManagement>

<dependencies>

<dependency>

<groupId>org.springframework.cloud</groupId>

<artifactId>spring-cloud-dependencies</artifactId>

<version>Brixton.SR5</version>

<type>pom</type>

<scope>import</scope>

</dependency>

</dependencies>

</dependencyManagement>

通过@EnableEurekaServer注解启动一个服务注册中心提供给其他应用进行对话。这一步非常简单，只需在一个普通的Spring Boot应用中添加这个注解就能开启此功能，比如下面的例子：

@EnableEurekaServer

@SpringBootApplication

public class Application {

public static void main(String[] args) {

new SpringApplicationBuilder(Application.class).web(true).run(args);

}

}

在默认设置下，该服务注册中心也会将自己作为客户端来尝试注册它自己，所以我们需要禁用它的客户端注册行为，只需在application.properties中增加如下配置：

server.port=1111

eureka.instance.hostname=localhost

eureka.client.register-with-eureka=false

eureka.client.fetch-registry=false

eureka.client.serviceUrl.defaultZone=http://${eureka.instance.hostname}:${serve

r.port}/eureka/

由于后续内容也都会在本地运行，为了与后续要进行注册的服务区分，这里将服务注册中心的端口通过server.port属性设置为1111。

eureka.client.register-with-eureka：由于该应用为注册中心，所以设置为false，代表不向注册中心注册自己。

eureka.client.fetch-registry：由于注册中心的职责就是维护服务实例，它并不需要去检索服务，所以也设置为false。

在完成了上面的配置后，启动应用并访问http://localhost:1111/。可以看到如下图所示的Eureka信息面板，其中Instances currently registered with Eureka栏是空的，说明该注册中心还没有注册任何服务。

##### 注册服务提供者

在完成了服务注册中心的搭建之后，接下来我们尝试将一个既有的Spring Boot应用加入Eureka的服务治理体系中去。

可以使用上一章中实现的快速入门工程来进行改造，将其作为一个微服务应用向服务注册中心发布自己。首先，修改pom.xml，增加Spring Cloud Eureka模块的依赖，具体代码如下所示：

<dependencies>

<dependency>

<groupId>org.springframework.boot</groupId>

<artifactId>spring-boot-starter-web</artifactId>

</dependency>

<dependency>

<groupId>org.springframework.boot</groupId>

<artifactId>spring-boot-starter-test</artifactId>

<scope>test</scope>

</dependency>

<dependency>

<groupId>org.springframework.cloud</groupId>

<artifactId>spring-cloud-starter-eureka</artifactId>

</dependency>

</dependencies>

<dependencyManagement>

<dependencies>

<dependency>

<groupId>org.springframework.cloud</groupId>

<artifactId>spring-cloud-dependencies</artifactId>

<version>Brixton.SR5</version>

<type>pom</type>

<scope>import</scope>

</dependency>

</dependencies>

</dependencyManagement>

接着，改造/hello请求处理接口，通过注入DiscoveryClient对象，在日志中打印出服务的相关内容。

@RestController

public class HelloController {

private final Logger logger = Logger.getLogger(getClass());

@Autowired

private DiscoveryClient client;

@RequestMapping(value = "/hello", method = RequestMethod.GET)

public String index() {

ServiceInstance instance = client.getLocalServiceInstance();

logger.info("/hello, host:" + instance.getHost() + ", service\_id:" +

instance.getServiceId());

return "Hello World";

}

}

然后，在主类中通过加上@EnableDiscoveryClient注解，激活Eureka中的DiscoveryClient实现（自动化配置，创建DiscoveryClient接口针对Eureka客户端的EurekaDiscoveryClient实例），才能实现上述Controller中对服务信息的输出。

@EnableDiscoveryClient

@SpringBootApplication

public class HelloApplication {

public static void main(String[] args) {

SpringApplication.run(HelloApplication.class, args);

}

}

最后，我们需要在application.properties配置文件中，通过spring.application.name属性来为服务命名，比如命名为hello-service。再通过eureka.client.serviceUrl.defaultZone属性来指定服务注册中心的地址，这里我们指定为之前构建的服务注册中心地址，完整配置如下所示：

spring.application.name=hello-service

eureka.client.serviceUrl.defaultZone=http://localhost:1111/eureka/

下面我们分别启动服务注册中心以及这里改造后的hello-service服务。在hello-service服务控制台中，Tomcat启动之后，com.netflix.discovery.DiscoveryClient对象打印了该服务的注册信息，表示服务注册成功。

s.b.c.e.t.TomcatEmbeddedServletContainer : Tomcat started on port(s): 8080 (http)

c.n.e.EurekaDiscoveryClientConfiguration : Updating port to 8080

com.didispace.HelloApplication : Started HelloApplication in 7.218

seconds (JVM running for 11.646)

com.netflix.discovery.DiscoveryClient : DiscoveryClient\_HELLO-SERVICE/PC-

201602152056:hello-service-registration status: 204

而此时在服务注册中心的控制台中，可以看到类似下面的输出，名为hello-service的服务被注册成功了。

c.n.e.registry.AbstractInstanceRegistry : Registered instance

HELLO-SERVICE/PC-201602152056:hello-service with status UP (replication=true)

我们也可以通过访问Eureka的信息面板，在Instances currently registered with Eureka一栏中看到服务的注册信息。

通过访问http://localhost:8080/hello，直接向该服务发起请求，在控制台中可以看到如下输出：

com.didispace.web.HelloController : /hello, host:PC-201602152056,

service\_id:hello-service

这些输出内容就是之前我们在HelloController中注入的DiscoveryClient接口对象，从服务注册中心获取的服务相关信息。

##### Region和Zone

对于访问实例的选择，Eureka中有Region和Zone的概念，一个Region中可以包含多个Zone，每个服务客户端需要被注册到一个Zone中，所以每个客户端对应一个Region和一个Zone。在进行服务调用的时候，优先访问同处一个Zone中的服务提供方，若访问不到，就访问其他的Zone

## 设计模式

## Maven

## Gradle