

讲堂 □ [深入拆解 Java 虚拟机](#) □ [文章详情](#)

## 13 | Java内存模型

2018-08-20 郑雨迪



### 13 | Java内存模型

朗读人：郑雨迪 13'45" | 6.30M

0:00 / 0:00

我们先来看一个反常识的例子。

```
1 int a=0, b=0;
2
3 public void method1() {
4     int r2 = a;
5     b = 1;
6 }
7
8 public void method2() {
9     int r1 = b;
10    a = 2;
11}
```

□ 复制代码

这里我定义了两个共享变量 `a` 和 `b`，以及两个方法。第一个方法将局部变量 `r2` 赋值为 `a`，然后将共享变量 `b` 赋值为 1。第二个方法将局部变量 `r1` 赋值为 `b`，然后将共享变量 `a` 赋值为 2。请问 `(r1, r2)` 的可能值都有哪些？

在单线程环境下，我们可以先调用第一个方法，最终 `(r1, r2)` 为 `(1, 0)`；也可以先调用第二个方法，最终为 `(0, 2)`。

在多线程环境下，假设这两个方法分别跑在两个不同的线程之上，如果 Java 虚拟机在执行了任一方法的第一条赋值语句之后便切换线程，那么最终结果将可能出现 `(0, 0)` 的情况。

除上述三种情况之外，Java 语言规范第 17.4 小节 [1] 还介绍了一种看似不可能的情况 `(1, 2)`。

造成这一情况的原因有三个，分别为即时编译器的重排序，处理器的乱序执行，以及内存系统的重排序。由于后两种原因涉及具体的体系架构，我们暂且放到一边。下面我先来讲一下编译器优化的重排序是怎么回事。

首先需要说明一点，即时编译器（和处理器）需要保证程序能够遵守 `as-if-serial` 属性。通俗地说，就是在单线程情况下，要给程序一个顺序执行的假象。即经过重排序的执行结果要与顺序执行的结果保持一致。

另外，如果两个操作之间存在数据依赖，那么即时编译器（和处理器）不能调整它们的顺序，否则将会造成程序语义的改变。

```
1 int a=0, b=0;
2
3 public void method1() {
4     int r2 = a;
```

□ 复制代码

```
5 b = 1;
6 .. // Code uses b
7 if (r2 == 2) {
8 ..
9 }
10}
```

在上面这段代码中，我扩展了先前例子中的第一个方法。新增的代码会先使用共享变量 `b` 的值，然后再使用局部变量 `r2` 的值。

此时，即时编译器有两种选择。

第一，在一开始便将 `a` 加载至某一寄存器中，并且在接下来 `b` 的赋值操作以及使用 `b` 的代码中避免使用该寄存器。第二，在真正使用 `r2` 时才将 `a` 加载至寄存器中。这么一来，在执行使用 `b` 的代码时，我们不再霸占一个通用寄存器，从而减少需要借助栈空间的情况。

```
1int a=0, b=0;
2
3public void method1() {
4for (..) {
5int r2 = a;
6b = 1;
7.. // Code uses r2 and rewrites a
8}
9}
```

□ 复制代码

另一个例子则是将第一个方法的代码放入一个循环中。除了原本的两条赋值语句之外，我只在循环中添加了使用 `r2`，并且更新 `a` 的代码。由于对 `b` 的赋值是循环无关的，即时编译器很有可能将其移出循环之前，而对 `r2` 的赋值语句还停留在循环之中。

如果想要复现这两个场景，你可能需要添加大量有意义的局部变量，来给寄存器分配算法施加压力。

可以看到，即时编译器的优化可能将原本字段访问的执行顺序打乱。在单线程环境下，由于 `as-if-serial` 的保证，我们无须担心顺序执行不可能发生的情况，如  $(r1, r2) = (1, 2)$ 。

然而，在多线程情况下，这种数据竞争（data race）的情况是有可能发生的。而且，Java 语言规范将其归咎于应用程序没有作出恰当的同步操作。

## Java 内存模型与 happens-before 关系

为了让应用程序能够免于数据竞争的干扰，Java 5 引入了明确定义的 Java 内存模型。其中最为重要的一个概念便是 happens-before 关系。happens-before 关系是用来描述两个操作的内存可见性的。如果操作 `X` happens-before 操作 `Y`，那么 `X` 的结果对于 `Y` 可见。

在同一个线程中，字节码的先后顺序（program order）也暗含了 happens-before 关系：在程序控制流路径中靠前的字节码 happens-before 靠后的字节码。然而，这并不意味着前者一定在后者之前执行。实际上，如果后者没有观测前者的运行结果，即后者没有数据依赖于前者，那么它们可能会被重排序。

除了线程内的 happens-before 关系之外，Java 内存模型还定义了下述线程间的 happens-before 关系。

1. 解锁操作 happens-before 之后（这里指时钟顺序先后）对同一把锁的加锁操作。
2. `volatile` 字段的写操作 happens-before 之后（这里指时钟顺序先后）对同一字段的读操作。

3. 线程的启动操作（即 Thread.starts()）happens-before 该线程的第一个操作。
4. 线程的最后一个操作 happens-before 它的终止事件（即其他线程通过 Thread.isAlive() 或 Thread.join() 判断该线程是否中止）。
5. 线程对其他线程的中断操作 happens-before 被中断线程所收到的中断事件（即被中断线程的 InterruptedException 异常，或者第三个线程针对被中断线程的 Thread.interrupted 或者 Thread.isInterrupted 调用）。
6. 构造器中的最后一个操作 happens-before 析构器的第一个操作。

happens-before 关系还具备传递性。如果操作 X happens-before 操作 Y，而操作 Y happens-before 操作 Z，那么操作 X happens-before 操作 Z。

在文章开头的例子中，程序没有定义任何 happens-before 关系，仅拥有默认的线程内 happens-before 关系。也就是 r2 的赋值操作 happens-before b 的赋值操作，r1 的赋值操作 happens-before a 的赋值操作。

```
1 Thread1 Thread2
2 | |
3 b=1 |
4 | r1=b
5 | a=2
6 r2=a |
```

□ 复制代码

拥有 happens-before 关系的两对赋值操作之间没有数据依赖，因此即时编译器、处理器都可能对其进行重排序。举例来说，只要将 b 的赋值操作排在 r2 的赋值操作之前，那么便可以按照赋值 b，赋值 r1，赋值 a，赋值 r2 的顺序得到 (1, 2) 的结果。

那么如何解决这个问题呢？答案是，将 a 或者 b 设置为 volatile 字段。

比如说将 b 设置为 volatile 字段。假设 r1 能够观测到 b 的赋值结果 1。显然，这需要 b 的赋值操作在时钟顺序上先于 r1 的赋值操作。根据 volatile 字段的 happens-before 关系，我们知道 b 的赋值操作 happens-before r1 的赋值操作。

```
1 int a=0;
2 volatile int b=0;
3
4 public void method1() {
5     int r2 = a;
6     b = 1;
7 }
8
9 public void method2() {
10    int r1 = b;
11    a = 2;
12}
```

□ 复制代码

根据同一个线程中，字节码顺序所暗含的 happens-before 关系，以及 happens-before 关系的传递性，我们可以轻易得出 r2 的赋值操作 happens-before a 的赋值操作。

这也就意味着，当对 a 进行赋值时，对 r2 的赋值操作已经完成了。因此，在 b 为 volatile 字段的情况下，程序不可能出现 (r1, r2) 为 (1, 2) 的情况。

由此可以看出，解决这种数据竞争问题的关键在于构造一个跨线程的 happens-before 关系：操作 X happens-before 操作 Y，使得操作 X 之前的字节码的结果对操作 Y 之后的字节码可见。

## Java 内存模型的底层实现

在理解了 Java 内存模型的概念之后，我们现在来看看它的底层实现。Java 内存模型是通过内存屏障（memory barrier）来禁止重排序的。

对于即时编译器来说，它会针对前面提到的每一个 happens-before 关系，向正在编译的目标方法中插入相应的读读、读写、写读以及写写内存屏障。

这些内存屏障会限制即时编译器的重排序操作。以 volatile 字段访问为例，所插入的内存屏障将不允许 volatile 字段写操作之前的内存访问被重排序至其之后；也将不允许 volatile 字段读操作之后的内存访问被重排序至其之前。

然后，即时编译器将根据具体的底层体系架构，将这些内存屏障替换成具体的 CPU 指令。以我们日常接触的 X86\_64 架构来说，读读、读写以及写写内存屏障是空操作（no-op），只有写读内存屏障会被替换成具体指令 [2]。

在文章开头的例子中，method1 和 method2 之中的代码均属于先读后写（假设 r1 和 r2 被存储在寄存器之中）。X86\_64 架构的处理器并不能将读操作重排序至写操作之后，具体可参考 Intel Software Developer Manual Volume 3, 8.2.3.3 小节。因此，我认为例子中的重排序必然是即时编译器造成的。

举例来说，对于 volatile 字段，即时编译器将在 volatile 字段的读写操作前后各插入一些内存屏障。

然而，在 X86\_64 架构上，只有 volatile 字段写操作之后的写读内存屏障需要用具体指令来替代。（HotSpot 所选取的具体指令是 lock add DWORD PTR [rsp],0x0，而非 mfence[3]。）

该具体指令的效果，可以简单理解为强制刷新处理器的写缓存。写缓存是处理器用来加速内存存储效率的一项技术。

在碰到内存写操作时，处理器并不会等待该指令结束，而是直接开始下一指令，并且依赖于写缓存将更改的数据同步至主内存（main memory）之中。

强制刷新写缓存，将使得当前线程写入 volatile 字段的值（以及写缓存中已有的其他内存修改），同步至主内存之中。

由于内存写操作同时会无效化其他处理器所持有的、指向同一内存地址的缓存行，因此可以认为其他处理器能够立即见到该 volatile 字段的最新值。

## 锁，volatile 字段，final 字段与安全发布

下面我来讲讲 Java 内存模型涉及的几个关键词。

前面提到，锁操作同样具备 happens-before 关系。具体来说，解锁操作 happens-before 之后对同一把锁的加锁操作。实际上，在解锁时，Java 虚拟机同样需要强制刷新缓存，使得当前线程所修改的内存对其他线程可见。

需要注意的是，锁操作的 happens-before 规则的关键字是同一把锁。也就意味着，如果编译器能够（通过逃逸分析）证明某把锁仅被同一线程持有，那么它可以移除相应的加锁解锁操作。

因此也就不再强制刷新缓存。举个例子，即时编译后的 synchronized (new Object()) {}，可能等同于空操作，而不会强制刷新缓存。

`volatile` 字段可以看成一种轻量级的、不保证原子性的同步，其性能往往优于（至少不亚于）锁操作。然而，频繁地访问 `volatile` 字段也会因为不断地强制刷新缓存而严重影响程序的性能。

在 X86\_64 平台上，只有 `volatile` 字段的写操作会强制刷新缓存。因此，理想情况下对 `volatile` 字段的使用应当多读少写，并且应当只有一个线程进行写操作。

`volatile` 字段的另一个特性是即时编译器无法将其分配到寄存器里。换句话说，`volatile` 字段的每次访问均需要直接从内存中读写。

`final` 实例字段则涉及新建对象的发布问题。当一个对象包含 `final` 实例字段时，我们希望其他线程只能看到已初始化的 `final` 实例字段。

因此，即时编译器会在 `final` 字段的写操作后插入一个写屏障，以防某些优化将新建对象的发布（即将实例对象写入一个共享引用中）重排序至 `final` 字段的写操作之前。在 X86\_64 平台上，写屏障是空操作。

新建对象的安全发布（safe publication）问题不仅仅包括 `final` 实例字段的可见性，还包括其他实例字段的可见性。

当发布一个已初始化的对象时，我们希望所有已初始化的实例字段对其他线程可见。否则，其他线程可能见到一个仅部分初始化的新建对象，从而造成程序错误。这里我就不展开了。如果你感兴趣的话，可以参考这篇博客 [4]。

## 总结与实践

今天我主要介绍了 Java 的内存模型。

Java 内存模型通过定义了一系列的 happens-before 操作，让应用程序开发者能够轻易地表达不同线程的操作之间的内存可见性。

在遵守 Java 内存模型的前提下，即时编译器以及底层体系架构能够调整内存访问操作，以达到性能优化的效果。如果开发者没有正确地利用 happens-before 规则，那么将可能导致数据竞争。

Java 内存模型是通过内存屏障来禁止重排序的。对于即时编译器来说，内存屏障将限制它所能做的重排序优化。对于处理器来说，内存屏障会导致缓存的刷新操作。

今天的实践环节，我们来复现文章初始的例子。由于复现需要大量的线程切换事件，因此我借助了 OpenJDK CodeTools 项目的 `jcstress` 工具 [5]，来对该例子进行并发情况下的压力测试。具体的命令如下所示：

```
$ mvn archetype:generate -DinteractiveMode=false -DarchetypeGroupId=org.openjdk.jcstress 复制代码
1 DarchetypeArtifactId=jcstress-java-test-archetype -DarchetypeVersion=0.1.1 -DgroupId=org.sample -
  DartifactId=test -Dversion=1.0
2 $ cd test
3 $ echo 'package org.sample;
4 import org.openjdk.jcstress.annotations.*;
5 import org.openjdk.jcstress.infra.results.IntResult2;
6 @JCStressTest
7 @Outcome(id = {"0, 0", "0, 2", "1, 0"}, expect = Expect.ACCEPTABLE, desc = "Normal outcome")
8 @Outcome(id = {"1, 2"}, expect = Expect.ACCEPTABLE_INTERESTING, desc = "Abnormal outcome")
9 @State
10 public class ConcurrencyTest {
11     int a=0;
12     int b=0; // 改成 volatile 试试?
```

```
13@Actor
14public void method1(IntResult2 r) {
15r.r2 = a;
16b = 1;
17}
18@Actor
19public void method2(IntResult2 r) {
20r.r1 = b;
21a = 2;
22}
23}' > src/main/java/org/sample/ConcurrencyTest.java
24$ mvn package
25$ java -jar target/jcstress.jar
```

如果你想要复现非安全发布的情形，那么你可以试试这一测试用例 [6]。

- [1] <https://docs.oracle.com/javase/specs/jls/se10/html/jls-17.html#jls-17.4>
- [2] <http://gee.cs.oswego.edu/dl/jmm/cookbook.html>
- [3] <https://blogs.oracle.com/dave/instruction-selection-for-volatile-fences--mfence-vs-lock:add>
- [4] <http://vikan.com/blog/post/2014/02/14/java-safe-publication/>
- [5] <https://wiki.openjdk.java.net/display/CodeTools/jcstress>
- [6] <http://hg.openjdk.java.net/code-tools/jcstress/file/64f2cf32fa0a/tests-custom/src/main/java/org/openjdk/jcstress/tests/unsafe/UnsafePublication.java>

版权归极客邦科技所有，未经许可不得转载

写留言

## 精选留言



追梦

1

老师，求讲解下jvm中代码如何实现的内存屏障

2018-08-20

作者回复

即时编译器生成的代码里会使用CPU的内存屏障指令。HotSpot采用的lock前缀的指令，lock add DWORD PTR [rsp] 0。它也会刷缓存。

至于在即时编译器里禁止重排序所使用的“内存屏障”，就是一个特殊的编译器中间表达形式节点。

2018-08-20



黑崽

0

请教个问题。刚才有说不会把volatile放到寄存器，但是应该会在栈里面对吧。直接读取主内存，读写的是栈数据，然后利用堆内存和栈上数据是利用写缓存刷新同步的？

2018-08-22

作者回复

我可能没有在原文中讲清楚。这里指的是volatile变量不能被分配到寄存器中，但是计算还是加载到寄存器中来计算的。

所谓的分配到寄存器中，你可以理解为编译器将内存中的值缓存在寄存器中，之后一

直用访问寄存器来代表对这个内存的访问的。假设我们要遍历一个数组，数组的长度是内存中的值。由于我们每次循环都要比较一次，因此编译器决定把它放在寄存器中，免得每次比较都要读一次内存。对于会更改的内存值，编译器也可以先缓存至寄存器，最后更新回内存即可。

Volatile会禁止上述优化。

2018-08-22



Jerry Chan

博客在哪里啊？

2018-08-21

作者回复

0

文末[4]所指向的链接。

2018-08-22



qpm

老师我提个问题。一个共享对象的变量是非volatile的，那么这个变量的写入会先写到寄存器上，再写回内存吗？那么jvm是不是无论如何都不保证啥时候变量的值会写回内存。假如另一个线程加锁访问这个变量，是不是jvm也不保证它能拿到最新数据。

2018-08-20

作者回复

0

对的！如果即时编译器把那个变量放在寄存器里维护，那么另一个线程也没辙。

2018-08-22



大场镇车王

老师 为什么volatile内存屏障不允许所有写操作之前的读操作被重排序到写操作之后？前面不是说volatile的写操作happens before对用一字段的读操作吗

2018-08-20

作者回复

0

volatile字段的happens-before关系指的是在两个不同线程中，volatile的写操作 happens-before 之后对同一字段的读操作。这里有个关键字之后，指的是时间上的先后。也就是我这边写，你之后再读就一定能读得到我刚刚写的值。普通字段则没有这个保证。

屏障不允许重排序是针对即时编译器的。写后对同一字段的读，属于数据依赖，本来也不可以重排序的。

2018-08-20



Alex 1984

老师，我在一些技术文章里看到说 volatile 的变量是存在工作内存，这个工作内存是一个什么概念？

2018-08-20

作者回复

0

工作内存是JMM抽象出来的一个概念。你可以映射到实际的CPU缓存。

2018-08-22



godtrue

□ 10

恩，这节听了好几遍，也读了几遍，最后还是感觉蒙蒙的。

下面说下我的感受：

1:一图胜千言，尤其对于描述复杂的东西，这个建议其他同学也有提议的，希望雨迪采纳一下

2:感觉有些概念没有解释，比如：

2-1:Java内存模型，这节不就是要讲Java内存类型是什么？有什么特点？为什么这么设计嘛？不过我没看到这些内容，当然，特点是讲了的 happens-before 机制

2-2:内存屏蔽，这是什么意思？它怎么就能禁止重排序啦？还有有其引申出的各种屏蔽是怎么回事呢？也没完全明白

3:前面讲 as-if-serial 大概明白什么意思了，多处理器为了提高效率会采用流水线的方式来执行程序指令，但是同时要保证程序执行逻辑的正确性，所以，实际执行程序指令和代码层面上会有不同，这个是由编译器来保证的，虽然执行逻辑不同但是程序逻辑是没变的，单线程没问题，但是多线程就变得复杂一些了，光靠这种方式保证不了啦，所以引出了下文

4:happens-before，这个概念和作用比较容易理解，线程内有这种关系，线程间更有，这个机制我认为就是为了多线程环境下为了保证程序逻辑正确性的一种方式，不过它的具体实现细节感觉没理解，不清楚他是怎么办到的

5:volatile 这个关键字之前也学习过，作用是保证内容的修改对所有线程可见，原理是修改后同步更新所有的内容，这是因为内存和处理器直接还是存在距离的，比如：内存-一级缓存-二级缓存-各种寄存器-cpu，如果是内存-cpu，则不会有这种问题了，不过性能也就不行了

2018-08-23



啃yi嘴泥

□ 2

老师你好，关于指令重排序有点不太理解，指令重排序的粒度是方法级别的，还是整个源文件级别的。文中说道，b 加了volatile后，能够保证 b=1 先于r1=b，这个我能理解，但是如何保证不会因为指令重排导致 b=1 先于r2=a发生呢？文中虽然说了，同一个线程中，字节码顺序暗含了r2=a happen before b=1，但是文中也提到了，拥有happen-before关系的两对赋值操作之间没有数据依赖，处理器可以指令重排序。r2=a 和b=1之间没有数据依赖呀！不好意思，这块有点迷糊，老师能给详细解答下不？

2018-08-24

作者回复

首先，b加了volatile之后，并不能保证b=1一定先于r1=b，而是保证r1=b始终能够看到b的最新值。比如说b=1;b=2，之后在另一个CPU上执行r1=b，那么r1会被赋值为2。如果先执行r1=b，然后在另外一个CPU上执行b=1和b=2，那么r1将看到b=1之前的值。



在没有标记volatile的时候，同一线程中， $r2=a$ 和 $b=1$ 存在happens before关系，但因为没数据依赖可以重排列。一旦标记了volatile，即时编译器和CPU需要考虑到多线程happens-before关系，因此不能自由地重排序。

2018-08-30



冬末末末

□ 1

happen-before 在这里不能理解成在什么之前发生，它和时间没有任何关系。个人感觉解释成“生效可见于”更准确。

2018-09-27



Kenneth

□ 1

同求讲解该问题，谢谢老师！：

啃yi嘴泥

老师你好，关于指令重排序有点不太理解，指令重排序的粒度是方法级别的，还是整个源文件级别的。文中说道， $b$ 加了volatile后，能够保证 $b=1$ 先于 $r1=b$ ，这个我能理解，但是如何保证不会因为指令重排导致 $b=1$ 先于 $r2=a$ 发生呢？文中虽然说了，同一个线程中，字节码顺序暗含了 $r2=a$  happen before  $b=1$ ，但是文中也提到了，拥有happen-before关系的两对赋值操作之间没有数据依赖，处理器可以指令重排序。 $r2=a$ 和 $b=1$ 之间没有数据依赖呀！不好意思，这块有点迷糊，老师能给详细解答下不？

2018-08-24

2018-08-30

作者回复

嗯，我回到原问题上哈

2018-08-30



熊猫酒仙

□ 1

老师，既然有写缓存，是不是也有读缓存呢？强制刷新写缓存，无效化相应的内存数据，那么这些内存数据的读缓存也就失效了，需要重新加载最新数据，是否可以这样理解？

另外强制刷新写缓存是否也有粒度一说？就是我们加的锁也有粒度之分，那么所触发的强制刷新写缓存的区域可能不一样？

2018-08-22



魏全运

□ 0

读完全文我也有个和另一位读者相同的疑问：到底什么是内存模型？按照老师的解释感觉是说内存模型定义了变量的读写顺序，但还是没有直接对“内存模型”这个概念进行解释。那内存模型是不是和Java Jvm内存分类这种模块结构完全不同，内存模型只是一个理论上的定义，只是对内存访问操作的抽象而不是一个实际存在的或者说我们通常理解的模型，正如《深入理解java虚拟机：jvm高级特性与最佳实践》中提到的：Java内存模型的主要目标是定义程序中各个变量的访问规则，即在虚拟机中将变量存储到内存和从内存中取出变量这样的底层细节。

那么请问老师能否用总结性的语言定义下Java内存模型呢？

2018-10-14



影子

0

你好。这里在看jvm内存模型的时候遇到一个问题。java执行native方法时程序计数器为空。那执行完native方法之后是继续执行下去的呢？

2018-10-05



飞恒棋

0

老师，多线程的happen before操作其实都是发生在一个CPU中的吧，如果是另一个CPU，就没有效果了吧？我这样理解对吗？

2018-10-01



昵称而已

0

我写了个方法按照上面的代码循环执行了9999次，要么是1，0要么是0,2 一直没有出现1,2。这个怎么破？

2018-09-17

作者回复

你需要用jcmstress工具来进行压力测试。

如果跑自己写的程序的话，可以试试在电脑上跑多个满载进程，让操作系统不停切换进程。

2018-09-17



浩

0

您好 当工作线程需要的内存特别大 比如超过10m 那工作线程会copy10m的内存数据到工作线程嘛？

2018-09-03

作者回复

JMM里的工作内存是个虚拟概念，实际映射到的是CPU缓存。

如果你指工作线程会不会将用到的内存加载至CPU缓存中，那么答案是会的。

如果你指会不会另外开一片内存，来存储这部分用到的堆内存，那么答案是不会的。

2018-09-07



Rye

0

CPU无法直接操作内存地址，需要借助寄存器，volatile只是告诉编译器用的时候取最新的，但不能保证原子性，读，改，存至少3条CPU指令。

2018-08-28



Kyle

0

我个人理解的“JAVA内存模型”应该是包括两部分的内容：  
一是运行时数据区，  
二是定义了一组内存访问规则。

这里其实主要讲的是其中的第二部分内容。不知道是不是可以这样总结。

2018-08-25

作者回复

谢谢总结！确实，本文重点讲的是内存可见性规则。

JMM的工作内存，主内存这些概念都是抽象的，对应实际体系架构中的缓存和内存。本文切掉了抽象的那部分，直接用实际的体系架构来讲解。

2018-08-28



梦想启动的蜗牛

再次复习下 📖

2018-08-20

□ 0



agilejava

文字太多，最好能结合一些图形会更方便理解

2018-08-20

作者回复

□ 0

多谢建议！

2018-08-20