再有人问你Java内存模型是什么，就把这篇文章发给他。(http://www.hollischuang.com/archives/2550)

2018-07-01 分类：Java (http://www.hollischuang.com/archives/category/java) 阅读(44542) 评论(12)

前几天，发了一篇文章，介绍了一下JVM内存结构、Java内存模型以及Java对象模型之间的区别(http://www.hollischuang.com/archives/2509)。有很多小伙伴反馈希望可以深入的讲解下每个知识点。Java内存模型，是这三个知识点当中最晦涩难懂的一个，而且涉及到很多背景知识和相关知识。

网上有很多关于Java内存模型的文章，在《深入理解Java虚拟机》和《Java并发编程的艺术》等书中也都有关于这个知识点的介绍。但是，很多人读完之后还是搞不清楚，甚至有的人说自己更懵了。本文，就来整体的介绍一下Java内存模型，目的很简单，让你读完本文以后，就知道到底Java内存模型是什么，为什么要有Java内存模型，Java内存模型解决了什么问题等。

本文中，有很多定义和说法，都是笔者自己理解后定义出来的。希望能够让读者可以对Java内存模型有更加清晰的认识。当然，如有偏颇，欢迎指正。

为什么要有内存模型

在介绍Java内存模型之前，先来看一下到底什么是计算机内存模型，然后再来看Java内存模型在计算机内存模型的基础上做了哪些事情。要说计算机的内存模型，就要说一下一段古老的历史，看一下为什么要有内存模型。

内存模型，英文名Memory Model，他是一个很老的老古董了。他是与计算机硬件有关的一个概念。那么我先给你介绍下他和硬件到底有啥关系。

CPU和缓存一致性

我们应该都知道，计算机在执行程序的时候，每条指令都是在CPU中执行的，而执行的时候，又免不了要和数据打交道。而计算机上面的数据，是存放在主存当中的，也就是计算机的物理内存啦。

刚开始，还相安无事的，但是随着CPU技术的发展，CPU的执行速度越来越快。而由于内存的技术并没有太大的变化，所以从内存中读取和写入数据的过程和CPU的执行速度比起来差距就会越来越大,这就导致CPU每次操作内存都要耗费很多等待时间。

> 这就像一家创业公司，刚开始，创始人和员工之间工作关系其乐融融，但是随着创始人的能力和野心越来越大，逐渐和员工之间出现了差距，普通员工原来越跟不上CEO的脚步。老板的每一个命令，传到到基层员工之后，由于基层员工的理解能力、执行能力的欠缺，就会耗费很多时间。这也就无形中拖慢了整家公司的工作效率。

可是，不能因为内存的读写速度慢，就不发展CPU技术了吧，总不能让内存成为计算机处理的瓶颈吧。

所以，人们想出来了一个好的办法，就是在CPU和内存之间增加高速缓存。缓存的概念大家都知道，就是保存一份数据拷贝。他的特点是速度快，内存小，并且昂贵。

那么，程序的执行过程就变成了：

当程序在运行过程中，会将运算需要的数据从主存复制一份到CPU的高速缓存当中，那么CPU进行计算时就可以直接从它的高速缓存读取数据和向其中写入数据，当运算结束之后，再将高速缓存中的数据刷新到主存当中。

> 之后，这家公司开始设立中层管理人员，管理人员直接归CEO领导，领导有什么指示，直接告诉管理人员，然后就可以去做自己的事情了。管理人员负责去协调底层员工的工作。因为管理人员是了解手下的人员以及自己负责的事情的。所以，大多数时候，公司的各种决策，通知等，CEO只要和管理人员之间沟通就够了。

而随着CPU能力的不断提升，一层缓存就慢慢的无法满足要求了，就逐渐的衍生出多级缓存。

按照数据读取顺序和与CPU结合的紧密程度，CPU缓存可以分为一级缓存（L1），二级缓存（L3），部分高端CPU还具有三级缓存（L3），每一级缓存中所储存的全部数据都是下一级缓存的一部分。

这三种缓存的技术难度和制造成本是相对递减的，所以其容量也是相对递增的。

那么，在有了多级缓存之后，程序的执行就变成了：

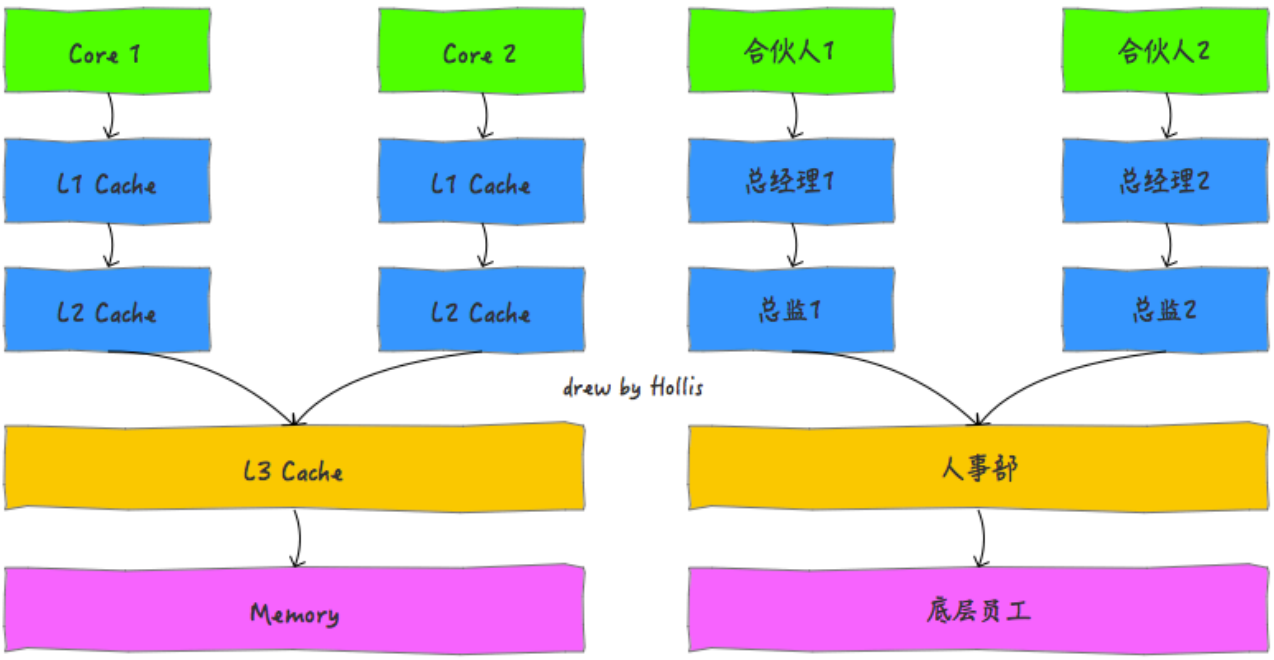
当CPU要读取一个数据时，首先从一级缓存中查找，如果没有找到再从二级缓存中查找，如果还是没有就从三级缓存或内存中查找。

> 随着公司越来越大，老板要管的事情越来越多，公司的管理部门开始改革，开始出现高层，中层，底层等管理者。一级一级之间逐层管理。

单核CPU只含有一套L1，L2，L3缓存；如果CPU含有多个核心，即多核CPU，则每个核心都含有一套L1（甚至和L2）缓存，而共享L3（或者和L2）缓存。

> 公司也分很多种，有些公司只有一个大Boss，他一个人说了算。但是有些公司有比如联席总经理、合伙人等机制。
>
> 单核CPU就像一家公司只有一个老板，所有命令都来自于他，那么就只需要一套管理班底就够了。
>
> 多核CPU就像一家公司是由多个合伙人共同创办的，那么，就需要给每个合伙人都设立一套供自己直接领导的高层管理人员，多个合伙人共享使用的是公司的底层员工。
>
> 还有的公司，不断壮大，开始差分出各个子公司。各个子公司就是多个CPU了，互相之前没有共用的资源。互不影响。

下图为一个单CPU双核的缓存结构。



随着计算机能力不断提升，开始支持多线程。那么问题就来了。我们分别来分析下单线程、多线程在单核CPU、多核CPU中的影响。

单线程。cpu核心的缓存只被一个线程访问。缓存独占，不会出现访问冲突等问题。

单核CPU，多线程。进程中的多个线程会同时访问进程中的共享数据，CPU将某块内存加载到缓存后，不同线程在访问相同的物理地址的时候，都会映射到相同的缓存位置，这样即使发生线程的切换，缓存仍然不会失效。但由于任何时刻只能有一个线程在执行，因此不会出现缓存访问冲突。

多核CPU，多线程。每个核都至少有一个L1 缓存。多个线程访问进程中的某个共享内存，且这多个线程分别在不同的核心上执行，则每个核心都会在各自的caehe中保留一份共享内存的缓冲。由于多核是可以并行的，可能会出现多个线程同时写各自的缓存的情况，而各自的cache之间的数据就有可能不同。

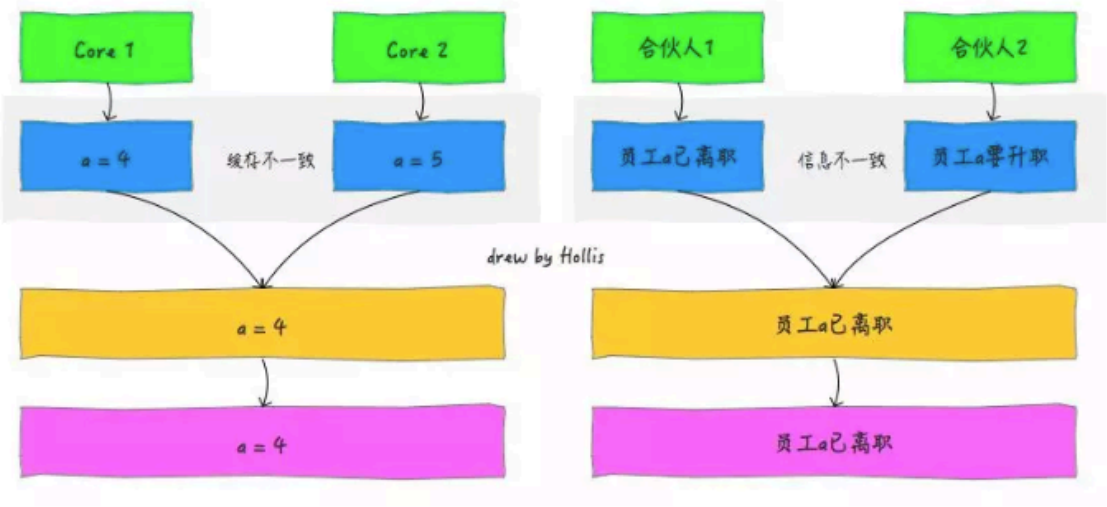
在CPU和主存之间增加缓存，在多线程场景下就可能存在**缓存一致性问题**，也就是说，在多核CPU中，每个核的自己的缓存中，关于同一个数据的缓存内容可能不一致。

如果这家公司的命令都是串行下发的话，那么就没有任何问题。

如果这家公司的命令都是并行下发的话，并且这些命令都是由同一个CEO下发的，这种机制是也没有什么问题。因为他的命令执行者只有一套管理体系。

如果这家公司的命令都是并行下发的话，并且这些命令是由多个合伙人下发的，这就有问题了。因为每个合伙人只会把命令下达给自己直属的管理人员，而多个管理人员管理的底层员工可能是公用的。

比如，合伙人1要辞退员工a，合伙人2要给员工a升职，升职后的话他再被辞退需要多个合伙人开会决议。两个合伙人分别把命令下发给了自己的管理人员。合伙人1命令下达后，管理人员a在辞退了员工后，他就知道这个员工被开除了。而合伙人2的管理人员2这时候在没得到消息之前，还认为员工a是在职的，他就欣然的接收了合伙人给他的升职a的命令。



处理器优化和指令重排

上面提到在在CPU和主存之间增加缓存，在多线程场景下会存在**缓存一致性问题**。除了这种情况，还有一种硬件问题也比较重要。那就是为了使处理器内部的运算单元能够尽量的被充分利用，处理器可能会对输入代码进行乱序执行处理。这就是**处理器优化**。

除了现在很多流行的处理器会对代码进行优化乱序处理，很多编程语言的编译器也会有类似的优化，比如Java虚拟机的即时编译器（JIT）也会做**指令重排**。

可想而知，如果任由处理器优化和编译器对指令重排的话，就可能导致各种各样的问题。

> 关于员工组织调整的情况，如果允许人事部在接到多个命令后进行随意拆分乱序执行或者重排的话，那么对于这个员工以及这家公司的影响是非常大的。

并发编程的问题

前面说的和硬件有关的概念你可能听得有点蒙，还不知道他到底和软件有啥关系。但是关于并发编程的问题你应该有所了解，比如原子性问题，可见性问题和有序性问题。

其实，原子性问题，可见性问题和有序性问题。是人们抽象定义出来的。而这个抽象的底层问题就是前面提到的缓存一致性问题、处理器优化问题和指令重排问题等。

这里简单回顾下这三个问题，并不准备深入展开，感兴趣的读者可以自行学习。我们说，并发编程，为了保证数据的安全，需要满足以下三个特性：

原子性是指在一个操作中就是cpu不可以在中途暂停然后再调度，既不被中断操作，要不执行完成，要不就不执行。

可见性是指当多个线程访问同一个变量时，一个线程修改了这个变量的值，其他线程能够立即看得到修改的值。

有序性即程序执行的顺序按照代码的先后顺序执行。

有没有发现，**缓存一致性问题**其实就是**可见性问题**。而**处理器优化**是可以导致**原子性问题**的。**指令重排**即会导致**有序性问题**。所以，后文将不再提起硬件层面的那些概念，而是直接使用大家熟悉的原子性、可见性和有序性。

什么是内存模型

前面提到的，缓存一致性问题、处理器器优化的指令重排问题是硬件的不断升级导致的。那么，有没有什么机制可以很好的解决上面的这些问题呢?

最简单直接的做法就是废除处理器和处理器的优化技术、废除CPU缓存，让CPU直接和主存交互。但是，这么做虽然可以保证多线程下的并发问题。但是，这就有点因噎废食了。

所以，为了保证并发编程中可以满足原子性、可见性及有序性。有一个重要的概念，那就是——内存模型。

为了保证共享内存的正确性（可见性、有序性、原子性），内存模型定义了共享内存系统中多线程程序读写操作行为的规范。通过这些规则来规范对内存的读写操作，从而保证指令执行的正确性。它与处理器有关、与缓存有关、与并发有关、与编译器也有关。他解决了CPU多级缓存、处理器优化、指令重排等导致的内存访问问题，保证了并发场景下的一致性、原子性和有序性。

内存模型解决并发问题主要采用两种方式：**限制处理器优化**和**使用内存屏障**。本文就不深入底层原理来展开介绍了，感兴趣的朋友可以自行学习。

什么是Java内存模型

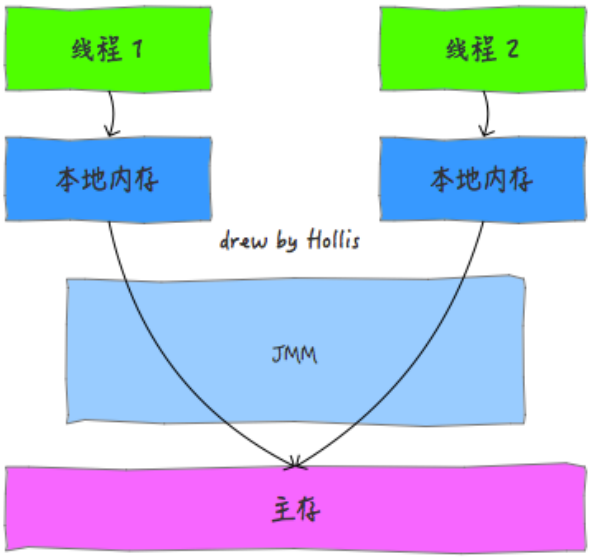
前面介绍过了计算机内存模型，这是解决多线程场景下并发问题的一个重要规范。那么具体的实现是如何的呢，不同的编程语言，在实现上可能有所不同。

我们知道，Java程序是需要运行在Java虚拟机上面的，**Java内存模型（Java Memory Model ,JMM）就是一种符合内存模型规范的，屏蔽了各种硬件和操作系统的访问差异的，保证了Java程序在各种平台下对内存的访问都能保证效果一致的机制及规范。**

提到Java内存模型，一般指的是JDK 5 开始使用的新的内存模型，主要由JSR-133: JavaTM Memory Model and Thread Specification (http://www.cs.umd.edu/~pugh/java/memoryModel/jsr133.pdf) 描述。感兴趣的可以参看下这份PDF文档（http://www.cs.umd.edu/~pugh/java/memoryModel/jsr133.pdf）

Java内存模型规定了所有的变量都存储在主内存中，每条线程还有自己的工作内存，线程的工作内存中保存了该线程中是用到的变量的主内存副本拷贝，线程对变量的所有操作都必须在工作内存中进行，而不能直接读写主内存。不同的线程之间也无法直接访问对方工作内存中的变量，线程间变量的传递均需要自己的工作内存和主存之间进行数据同步进行。

而JMM就作用于工作内存和主存之间数据同步过程。他规定了如何做数据同步以及什么时候做数据同步。



这里面提到的主内存和工作内存，读者可以简单的类比成计算机内存模型中的主存和缓存的概念。特别需要注意的是，主内存和工作内存与JVM内存结构中的Java堆、栈、方法区等并不是同一个层次的内存划分，无法直接类比。《深入理解Java虚拟机》中认为，如果一定要勉强对应起来的话，从变量、主内存、工作内存的定义来看，主内存主要对应于Java堆中的对象实例数据部分。工作内存则对应于虚拟机栈中的部分区域。

所以，再来总结下，JMM是一种规范，目的是解决由于多线程通过共享内存进行通信时，存在的本地内存数据不一致、编译器会对代码指令重排序、处理器会对代码乱序执行等带来的问题。

Java内存模型的实现

了解Java多线程的朋友都知道，在Java中提供了一系列和并发处理相关的关键字，比如 `volatile`、`synchronized`、`final`、`concurren` 包等。其实这些就是Java内存模型封装了底层的实现后提供给程序员使用的一些关键字。

在开发多线程的代码的时候，我们可以直接使用 `synchronized` 等关键字来控制并发，从来就不需要关心底层的编译器优化、缓存一致性等问题。所以，**Java内存模型，除了定义了一套规范，还提供了一系列原语，封装了底层实现后，供开发者直接使用。**

本文并不准备把所有的关键字逐一介绍其用法，因为关于各个关键字的用法，网上有很多资料。读者可以自行学习。本文还有一个重点要介绍的就是，我们前面提到，并发编程要解决原子性、有序性和一致性的问题，我们就再来看下，在Java中，分别使用什么方式来保证。

原子性

在Java中，为了保证原子性，提供了两个高级的字节码指令 `monitorenter` 和 `monitorexit` 。在synchronized的实现原理 (http://www.hollischuang.com/archives/1883)文章中，介绍过，这两个字节码，在Java中对应的关键字就是 `synchronized` 。

因此，在Java中可以使用 `synchronized` 来保证方法和代码块内的操作是原子性的。

可见性

Java内存模型是通过在变量修改后将新值同步回主内存，在变量读取前从主内存刷新变量值的这种依赖主内存作为传递媒介的方式来实现的。

Java中的 `volatile` 关键字提供了一个功能，那就是被其修饰的变量在被修改后可以立即同步到主内存，被其修饰的变量在每次是用之前都从主内存刷新。因此，可以使用 `volatile` 来保证多线程操作时变量的可见性。

除了 `volatile`，Java中的 `synchronized` 和 `final` 两个关键字也可以实现可见性。只不过实现方式不同，这里不再展开了。

有序性

在Java中，可以使用 synchronized 和 volatile 来保证多线程之间操作的有序性。实现方式有所区别：

volatile 关键字会禁止指令重排。 synchronized 关键字保证同一时刻只允许一条线程操作。

好了，这里简单的介绍完了Java并发编程中解决原子性、可见性以及有序性可以使用的关键字。读者可能发现了，好像 synchronized 关键字是万能的，他可以同时满足以上三种特性，这其实也是很多人滥用 synchronized 的原因。

但是 synchronized 是比较影响性能的，虽然编译器提供了很多锁优化技术，但是也不建议过度使用。

总结

在读完本文之后，相信你应该了解了什么是Java内存模型、Java内存模型的作用以及Java中内存模型做了什么事情等。关于Java中这些和内存模型有关的关键字，希望读者还可以继续深入学习，并且自己写几个例子亲自体会一下。

可以参考《深入理解Java虚拟机》和《Java并发编程的艺术》两本书。

参考资料

cpu缓存与多线程 (https://www.cnblogs.com/gtarcoder/p/5295281.html)

浅析内存模型 (http://valleylord.github.io/post/201606-memory-model/)

内存模型和缓存一致性 (https://ljalphabeta.gitbooks.io/a-primer-on-memory-consistency-and-cache-coherenc/content/%E7%AC%AC%E4%B8%80%E7%AB%A0-consistency.html)

并发编程——原子性，可见性和有序性 (https://blog.csdn.net/eff666/article/details/66473088)

《深入理解Java虚拟机》 (https://book.douban.com/subject/24722612/)

(全文完)