一致性和复制

分布式系统中的一个重要问题是数据的复制。通常复制数据是为了增强可靠性或提高性能。其中一个主要问题是保持副本的一致性。非正式地说，这意味着当一个副本被更新时，我们需要确保其他副本也被更新;否则，副本将不再相同。在本章中，我们将详细讨论复制数据的一致性实际上意味着什么，以及实现一致性的各种方法。

我们从一个一般性的介绍开始，讨论为什么复制是有用的，以及它与可伸缩性的关系。然后我们继续关注一致性到底意味着什么。一致性模型的一个重要假设是多个进程同时访问共享数据。当读取和更新共享数据时，知道其他人也在访问该数据时，可以根据进程的期望来制定这些情况的一致性。

在大规模分布式系统中，共享数据的一致性模型通常很难有效地实现。此外，在许多情况下可以使用更简单的模型，这些模型通常也更容易实现。一个特定的情况是由以客户端为中心的一致性模型构成的，该模型从单个(可能是移动的)客户端角度关注一致性。

一致性只是问题的一半。我们还需要考虑一致性实际上是如何实现的。我们基本上需要考虑两个或多或少独立的问题。首先，我们将重点放在管理副本上，它不仅考虑副本服务器的位置，还考虑内容如何分布到这些服务器。

第二个问题是副本如何保持一致。在大多数情况下，应用程序需要很强的一致性。非正式地说，这意味着更新将在副本之间或多或少地立即传播。有多种实现强一致性的方法，我们将在单独的部分中讨论。此外，还注意了缓存协议，它是一致性协议的一种特殊情况。

作为最大的分布式系统，我们分别关注基于web的系统中的缓存和复制，特别是内容交付网络和边缘服务器缓存技术。

7.1 简介

在本节中，我们首先讨论复制数据的重要原因。我们专注于将复制作为一种实现可伸缩性的技术，并解释了为什么一致性的推理如此重要。

**复制的原因**

复制数据有两个主要原因。首先，复制数据以提高系统的可靠性。如果一个文件系统已经被复制，那么在一个副本崩溃之后，只要切换到另一个副本，就可以继续工作。此外，通过维护多个副本，可以更好地保护数据免受损坏。例如，假设有一个文件的三个副本，每个副本都执行读写操作。通过将至少两个副本返回的值视为正确的值，我们可以防止一次失败的写操作。

复制数据的另一个原因是性能。当分布式系统需要按大小或所覆盖的地理区域进行伸缩时，性能复制非常重要。当越来越多的进程需要访问但一个服务器时，就会产生性能问题。在这种情况下，可以通过复制服务器并在访问数据的进程之间分配工作负载来改进性能。

相对于地理区域的扩展也可能需要复制。其基本思想是，通过将数据副本放在使用它们的进程附近，访问数据的时间就会减少。因此，该进程所感知到的性能会增加。这个例子还说明，复制对性能的好处可能很难评估。尽管客户机进程可能会感知到更好的性能，但也可能会消耗更多的网络带宽来保持所有副本都是最新的。

如果复制有助于提高可靠性和性能，那么谁会反对呢？不幸的是，在复制数据时要付出代价。复制的问题是，拥有多个副本可能会导致一致性问题。每当修改一个副本时，该副本就与其他副本不同。因此，必须对所有副本进行修改，以确保一致性。究竟什么时候以及如何进行这些修改决定了复制的价格。

要理解这个问题，可以考虑改进Web页面的访问时间。如果不采取特殊措施，从远程Web服务器获取页面有时甚至需要几秒钟。为了提高性能，Web浏览器通常在本地存储以前获取的Web页面的副本。如果用户再次需要该页面，浏览器将自动返回本地副本。用户感知的访问时间非常好。然而，如果用户总是想要一个页面的最新版本，这种方法可能不会生效。问题是，如果同时修改了页面，修改将不会传播到缓存副本，从而使这些副本过期。

将过期副本返回给用户的一个解决方案是首先禁止浏览器保存本地副本，从而有效地让服务器完全负责复制。但是，如果在用户附近没有放置副本，则此解决方案仍然可能导致较差的访问时间。另一个解决方案是让Web服务器使每个缓存副本失效或更新，但这要求服务器跟踪所有缓存并向它们发送消息。这进而可能降低服务器的整体性能。下面我们回到性能与可伸缩性的问题。

在接下来的文章中，我们将主要关注性能方面的复制。第8章讨论了可靠性的复制。

**Replication as scaling technique**

为提高性能而进行的复制和缓存作为扩展技术得到了广泛的应用。可伸缩性问题通常以性能问题的形式出现。将数据副本放在使用它们的进程附近可以通过减少访问时间提高性能，从而解决可伸缩性问题。

需要做出的一个可能的权衡是，使副本保持最新可能需要更多的网络带宽。考虑一个进程P每秒访问一个本地副本N次，而副本本身每秒更新M次。假设更新完全刷新本地副本的前一个版本。如果N << M，也就是说，访问更新比非常低，我们就会遇到这样的情况:P永远不会访问本地副本的许多更新版本，这使得这些版本的网络通信毫无用处。在这种情况下，最好不要安装接近P的本地副本，或者应用不同的策略来更新副本。

然而，一个更严重的问题是，保持多个副本的一致性本身可能存在严重的可伸缩性问题。直观地说，当副本总是相同时，副本集合是一致的。这意味着在任何副本上执行的读取操作总是返回相同的结果。因此，当对一个副本执行更新操作时，无论在哪个副本上启动或执行该操作，应该在后续操作发生之前将更新传播到所有副本。

这种类型的一致性有时称为非正式地(而且不精确地)一致性，由同步复制提供。(在第7.2节中，我们将提供一致性的精确定义，并介绍一系列一致性模型。) 关键思想是，更新作为单个原子操作或事务在所有副本上执行。不幸的是，当操作也需要快速完成时，实现涉及大量副本的原子性本质上是困难的，这些副本可能广泛分布在大型网络中。

困难在于我们需要同步所有副本。实际上，这意味着所有副本首先需要就何时在本地执行更新达成一致。例如，副本可能需要使用Lamport时间戳决定操作的全局顺序，或者让协调器分配这样的顺序。全局同步只需要大量的通信时间，特别是当副本分布在广域网中时。

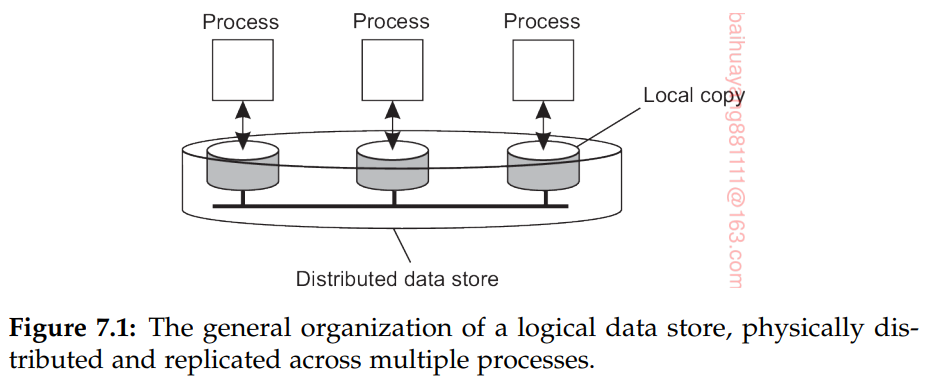
我们现在面临着进退两难的局面。一方面，通过应用复制和缓存可以缓解可伸缩性问题，从而提高性能。另一方面，为了保持所有副本的一致性，通常需要全局同步，这在性能方面有内在的代价。

在许多情况下，唯一真正的解决方案是放松一致性约束。换句话说，如果我们能够放松更新需要作为原子操作执行的要求，我们就能够避免(瞬时)全局同步，从而获得性能。付出的代价是，副本可能并不总是在所有地方都一样。事实证明，可以在多大程度上放松一致性，在很大程度上取决于对复制数据的访问和更新模式，以及使用这些数据的目的。

有一系列的一致性模型，以及通过所谓的分布和一致性协议实现模型的许多不同方法。

**7.2以数据为中心的一致性模型**

传统上，一致性是在共享数据的读写操作上下文中讨论的，这些操作通过(分布式)共享内存、(分布式)共享数据库或(分布式)文件系统来实现。这里，我们使用更广泛的术语**数据存储**。数据存储可以物理地分布在多台机器上。特别是，可以假定每个数据都存有本地副本，所以每个进程都可以进行访问。写操作被传播到其他副本，如图7.1所示。数据操作在更改数据时被分类为写操作，否则被分类为读操作。



**一致性模型**本质上是进程和数据存储之间的约定。它说，如果进程同意遵守某些规则，存储的数据承诺将正常工作。通常，对数据项执行读操作的进程期望该操作返回一个值，该值显示上次对该数据执行写操作的结果。

在没有全局时钟的情况下，很难精确地定义哪个写操作是最后一个写操作。作为一种替代方法，我们需要提供其他定义，从而产生一系列一致性模型。每个模型都有效地限制数据项上的读取操作可以返回的值。正如预期的那样，具有主要限制的应用程序很容易使用，例如在开发应用程序时，而具有次要限制的应用程序通常被认为很难在实践中使用。当然，代价是，易于使用的模型的性能不如难使用的模型。生活就是这样。Such is life~

**连续的一致性**

没有复制数据的最佳解决方案。复制数据会带来一致性问题，而这些问题通常无法有效地解决。只有我们放松一致性，才能有希望获得有效的解决方案。不幸的是，也没有放松一致性的一般规则:确切的说，只有高度依赖的应用程序才是可以容忍一致性的。

应用程序有不同的方法来指定它们可以容忍哪些不一致。Yu和Vahdat[2002]采用了一种通用的方法，通过区分三个独立的轴来定义不一致性:副本之间数值的偏差、副本之间陈旧程度的偏差以及更新操作顺序的偏差。他们将这些偏差称为形成**连续一致性**范围。

数据具有数值语义的应用程序可以使用数值偏差来测量不一致性。一个明显的例子是复制包含股票市场价格的记录。在这种情况下，应用程序可以指定两个副本的偏差不应超过0.02美元，这将是一个绝对的数值偏差。或者，可以指定一个相对的数值偏差，说明两个副本之间的差异不应超过0.5%。在这两种情况下，如果股票上涨(并且其中一个副本立即更新)而没有违反指定的数值偏差，那么副本仍然被认为是相互一致的。

数值偏差也可以理解为已应用于给定副本但尚未被其他人看到的更新的数量。例如，Web缓存可能没有看到由Web服务器执行的一批操作。在本例中，值的相关偏差也称为其权重。

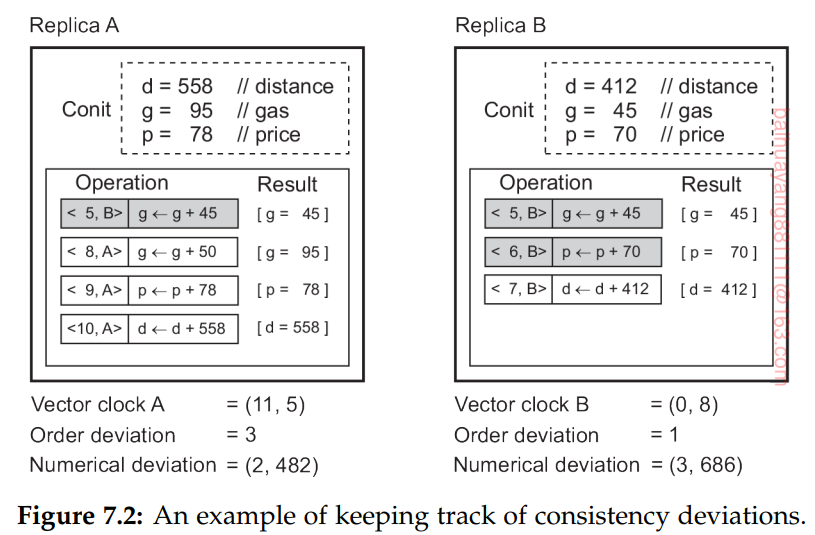
新旧偏差与副本更新的最后时间有关。对于某些应用程序，只要旧数据不太旧，就可以容忍副本提供旧数据。例如，天气报告通常在一段时间内保持相当准确，比如几个小时。在这种情况下，主服务器可能会收到及时的更新，但是可能决定只在一段时间内将更新传播到副本。

最后，在一些 应用程序类中，允许在不同的副本上执行不同的更新次序，只要这些差别是有限度的。查看这些更新的一种方法是，它们暂时应用于本地副本，等待所有副本的一起同步。因此，一些更新可能需要回滚并以不同的顺序应用，然后才能成为永久性的。直观地说，顺序偏差比其他两个一致性度量更难理解。

The notion of a conit

为了定义不一致性，Yu和Vahdat引入了一个**一致性单元**，缩写为**conit。**conit指定要测量一致性的单位。例如，在我们的股票交易所示例中，conit可以定义为表示单个股票的记录。另一个例子是单独的天气报告。

为了给出一个conit的例子，并同时说明数值和顺序偏差，请考虑跟踪车队的情况。特别是，车队老板想知道他平均为汽油付多少钱。为此，每当司机加满汽油时，他都会报告加满的汽油量(记为g)、支付的价格(记为p)以及自上次加满以来的总距离(记为变量d)。从技术上讲，g、p和d这三个变量构成了一个conit。这个conit被复制到两个服务器上，如图7.2所示，司机通过单独更新每个变量，定期向其中一个服务器报告他的汽油使用量(无需进一步考虑问题车辆)。



服务器的任务是保持conit的“一致性”被复制。为此，每个复制服务器维护一个二维矢量时钟。我们使用符号<T,R>， Ri来表示由副本R在(其)逻辑时间T执行的操作。在本例中，我们看到两个副本操作一个conit，其中包含本例中的数据项g、p和d。假设所有变量都初始化为0。副本A从副本B收到操作：



我们用灰色阴影表示该操作，表示A已将该操作提交到其本地存储。换句话说，它已经成为永久性的，不能回滚。副本A还列出了三个暂定的更新操作：<8,A>,<9,A>和<10,A>。就连续一致性而言，A有三个待提交的暂定操作，order deviation是3。类似地，在总共进行了三次操作(其中两次操作)的情况下，B的顺序为1。

从这个例子中，我们看到A的逻辑时钟值现在是11。我们假设向量的第一个分量用于A，第二个分量用于B。因为A接收到的B的最后一个操作的时间戳是5，所以A处的向量时钟将是(11,5)。同理，B处的逻辑时钟为(0,8)。在一个复制R处的数值偏差由两个部分组成:在所有其他尚未被R看到的复制上的操作数，以及相应的缺失值的总和(当然，也可能有更复杂的方案)。（就是别的机器的操作还没看到）。在我们的例子中，A看不到<6,B>和<7,B>操作，总共两个操作，值的和为70+412=486，所以值差为(2,482)。同样的，B也看不到A的三个操作，总差值是686，所以B的值差为(3,686)。

使用这些概念，可以指定特定的一致性方案。例如，我们可以通过指定一个可接受的最大值来限制这些偏差。同样，我们可能希望两个副本的数值偏差不超过1000。拥有这样的一致性方案确实需要一个副本知道它与其他副本的差异有多大，这意味着我们需要单独的通信来通知副本。基本的假设是，这种通信比保持副本同步的通信要便宜得多。诚然，如果这个假设也适用于我们的例子，那就值得怀疑了（怀疑啥？）。

虽然从概念的角度来看conits是捕获一致性需求的一种吸引人的方法，但是在将它们投入实际使用之前，有两个重要的问题需要处理。首先，为了加强一致性，我们需要协议。本章后面将讨论连续一致性协议。第二个问题是程序开发人员必须为他们的应用程序指定一致性需求。实践表明，获得这些需求可能极其困难。程序员通常不习惯处理复制，更不用说理解提供关于一致性的详细信息意味着什么。因此，必须有简单易懂的编程接口。

**操作顺序一致性**

在过去的几十年里，有大量以数据为中心的一致性模型的工作。一类重要的模型来自并行编程领域。在并行和分布式计算中，多个进程需要同时共享资源和访问这些资源，面对这一事实，研究人员试图在复制共享资源时表达并发访问的语义。我们在这里讨论的模型都处理对共享、复制数据的一致排序操作。原则上，当需要提交副本上的临时更新时，副本需要就全局(即更新的一致顺序)达成一致，因此模型增强了连续一致性。

**顺序一致性**

在下面，我们将使用一个特殊的符号，在这个符号中，我们沿着时间轴绘制进程的操作。时间轴总是水平绘制的，时间从左到右递增。我们使用符号Wi(x)来表示进程Pi将值a写入数据项x。类似地，Ri(x)b表示进程读取x并返回值b。我们假设每个数据项的初始值为NIL。当没有混淆哪个进程正在访问数据时，我们从符号W和R中省略索引。

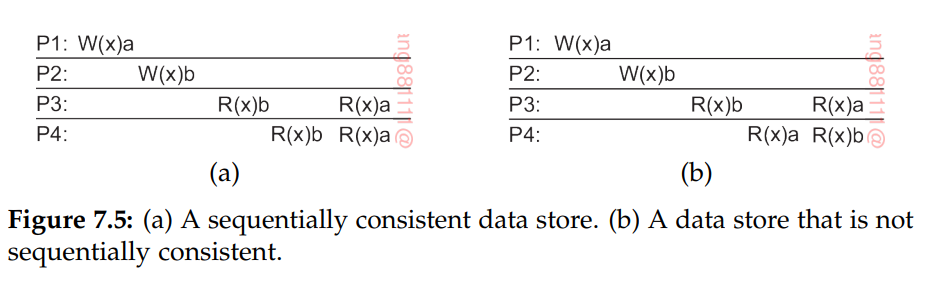
在图7.4中，p1向x写入数据a。注意，根据我们的系统模型，W1(x)a操作首先在P1本地的数据存储副本上执行，然后才传播到其他本地副本。在我们的示例中，P2稍后读取值NIL，在此之后的一段时间内读取值a(从存储的本地副本)。我们在这里看到的是，将x的更新传播到P2花费了一些时间，这完全可以接受。

**顺序一致性**是一种重要的以数据为中心的一致性模型，Lamport[1979]在多处理器系统共享内存的上下文中首次定义了该模型。当数据存储满足以下条件时，它被称为顺序一致的。

*The result of any execution is the same as if the (read and write) operations by all processes on the data store were executed in some sequential order and the operations of each individual process appear in this sequence in the order specified by its program.*

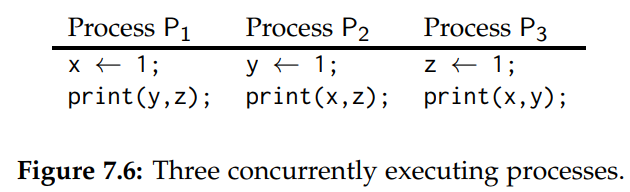
这个定义的意思是，当进程在(可能)不同的机器上并发运行时，任何有效的读写操作交错都是可接受的行为，但是所有进程都看到相同的操作交错（还是不懂，后面例子会讲解）。注意，这里没有提到时间;也就是说，数据项上没有对“最近”写操作的引用。此外，进程可以通过自己的读操作看到其他所有进程的写操作。

从图7.5中可以看出，时间没有发挥作用。考虑在相同数据项x上运行的四个进程。在图7.5（a），进程P1首先执行对x的写操作，写入a。之后（绝对时间上），进程P2同样执行x的写操作，写入b。然而，P3和P4都是先读到b值，然后读到a值。换句话说，写入b的操作先于写入a的操作。

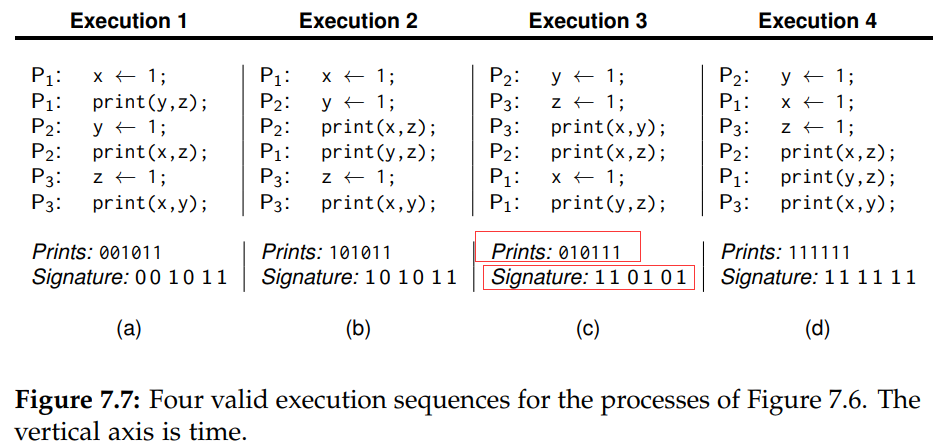


相反，图7.5(b)违反了顺序一致性，因为不是所有进程都看到相同的写操作交错。特别是在处理P3时，数据项似乎首先被更改为b，然后又被更改为a。另一方面，P4会得出最终的值是b。

为了使顺序一致性的概念更加具体，考虑三个并发执行的进程P1、P2和P3，如图7.6所示。本例中的数据项由三个整数变量x、y和z组成，它们存储在(可能是分布式的)共享的顺序一致的数据存储中。我们假设每个变量都初始化为0。在本例中，赋值对应于写操作，而print语句对应于它的两个参数的同时读取操作。所有的语句都被认为是不可分割的。



各种交错执行序列是可能的。使用6个独立语句，可能有720(6!)个可能的执行序列，尽管其中一些违反了程序顺序。考虑从x1开始的120(5!)个序列。其中一半在y =1之前已经打印(x,z)，因此违反了程序顺序。其中一半在z =1之前已经打印(x,y)，因此违反了程序顺序。因此只剩下四分之一，即30组有效的。另外30个有效序列可以从y =1开始，另外30个可以从z= 1开始，总共有90个有效的执行序列。其中的4种如图7.7所示。



在图7.7(a)中，这三个进程依次运行，首先是P1，然后是P2，然后是P3。另外三个例子展示了在时间上的交错不同但同样有效的语句。这三个进程中的每个进程打印两个变量。由于每个变量只能接受初值(0)或赋值(1)，所以每个进程生成一个2位字符串。打印后的数字是出现在输出设备上的实际输出。

如果我们按照这个顺序将P1、P2和P3的输出连接起来，就会得到一个6位的字符串，它表示语句之间的特定交错。这是作为图7.7中的签名列出的字符串。下面我们将通过每个订单的签名而不是打印输出来描述每个订单。

并不是所有64个签名模式都被允许。作为一个简单的例子，00 00 00是不允许的，因为这意味着print语句在赋值语句之前运行，违反了按程序顺序执行语句的要求。一个更微妙的例子是00 10 01。前两位00表示p打印时y和z都是0。只有当P1在P2或P3 \*之前执行两个语句时才会出现这种情况。接下来的两位，10，意味着P2必须在P1启动之后，P3启动之前运行。最后两位，01，意味着P3必须在P1开始之前完成，但我们已经看到P1必须先完成。因此，00 10 01是不允许的。

简而言之，90种不同的有效语句排序会产生各种不同的程序结果(不过小于64)，这些结果在序列一致性的假设下是允许的。进程与分布式共享系统之间的约定是进程必须接受这些结果作为有效的结果。换句话说，进程必须接受图7.7中所示的四个结果和所有其他有效结果作为有效的结果，并且必须在其中任何一个出现时正确工作。对其中一些结果有效而对其他结果无效的程序违反了与数据存储的约定，并且是不正确的。

**因果一致性**

因果一致性模型[Hutto and Ahamad, 1990]代表了弱化的顺序一致性，因为它区分了可能与因果相关的事件和不相关的事件。在前一章讨论向量时间戳时，我们已经遇到了因果关系。

如果事件b是由较早的事件a引起或影响的，因果关系要求其他人首先看到a，然后看到b。

考虑通过分布式共享数据库进行的简单交互。假设进程P1写了一个数据项x。进程p2读取x并写入y。这里读x和写y可能有因果关系，因为y的计算可能依赖于P2读x时的值。

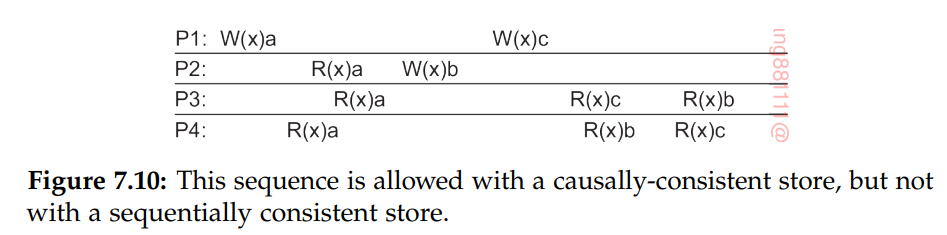
另一方面，如果两个进程同时自发地编写两个不同的数据项，则它们之间没有因果关系。没有因果关系的操作称为并发操作。要使数据存储被认为是因果一致的，必须使存储遵守以下条件：

*Writes that are potentially causally related must be seen by all processes*

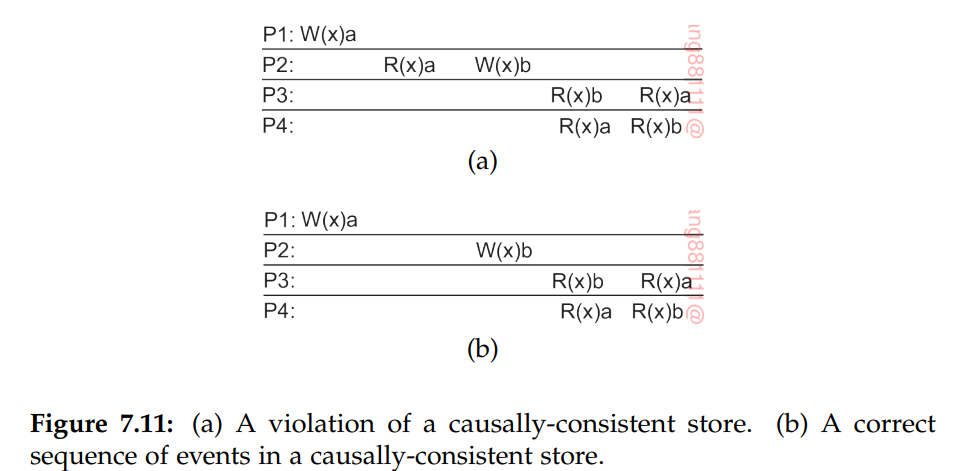
*in the same order. Concurrent writes may be seen in a different order on*

*different machines.*

作为因果一致性的一个例子，请考虑图7.10。在这里，我们有一个事件序列，它允许使用与原因一致的存储，但是禁止使用顺序一致的存储或严格一致的存储。需要注意的是，写入W2(x)b和W1(x)c是并发的，所以并不要求所有进程都以相同的顺序查看它们：



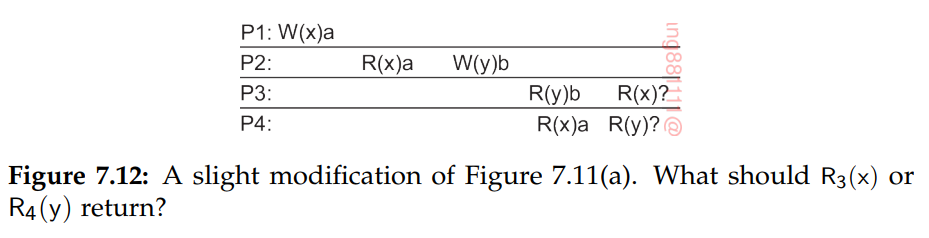
现在考虑第二个例子。在图7.11(a)中，我们有W2(x)b，它可能依赖于W1(x)a，因为将值b写到x中可能是R2(x)a先前读取值的计算结果。这两个写是因果相关的，所以所有进程必须以相同的顺序查看它们（即现在写入是p1->x(a) p2->x(b)，其他进程读取x 必须是 先a 后 b）。因此，图7.11(a)是不正确的。（3,4 必须以相同顺序读取a和b）。另一方面，在图7.11(b)中已经删除了读操作，所以W1(x)a和W2(x)b现在是并发写操作。原因一致的存储不需要对并发写进行全局排序，因此图7.11(b)是正确的。注意，图7.11(b)反映了顺序一致的存储不能接受的情况。



实现因果一致性需要跟踪哪些流程看到了哪些写入。有许多微妙的问题需要考虑。为了说明这一点，假设我们将图7.11(a)中的W2(x)b替换为W2(y)b，同样地，R3(x)b替换为R3(y)b。这种情况如图7.12所示。

首先来看R3(x)。进程P3在R3(y)b之后执行此操作。在这一点上我们确定W(x)a发生在W(y)b之前。特别是W(x)a->R (x)a->W(y)b，意思是如果我们要保持因果关系，从y中读取b后读取x只能返回a。如果系统将nil返回到P3，就违反了因果关系的保存。

R4 (y)呢?它能否返回y的初始值，即NIL? 答案是肯定的:虽然我们在没有从y中读取b的情况下 有W(x)a->W(y)b这层关系，进程P4仍然可以合理地观察到W(x)a的发生独立于y的初始化。



在实现方面，保留因果关系引入了一些有趣的问题。例如，考虑图7.12中的中间件基础进程P3。当这个中间件从读取y返回值b时，它必须知道关系W(x)a ->W (y) b。换句话说，当y的最新值传播到P3的中间件时，至少也应该传播关于y依赖关系的元数据（这里指W(x)a->W(y)b）。或者，传播也可以与更新P3节点上的x一起完成。总的来说，底线是我们需要一个依赖关系图，哪个操作依赖于哪个操作。当相关数据也在本地存储时，可以删除这样的图。

注：和上文非结果依赖顺序一致性区别：

非结果依赖，顺序一致性强调所有的进程必须执行出来的结果都保持相同的顺序，可以a b 也可以 b a，而结果依赖如7.11(a)，由于p2读取到了p1的写入值，则p2的写入b必须位于p1的写入a之后，则 p3 p4 读取的值 必须保持 先a后b。