(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 106164870 B
(45)授权公告日 2019.05.28

(21)申请号 201580016557.7

(22)申请日 2015.03.10

(65)同一申请的已公布的文献号
申请公布号 CN 106164870 A

(43)申请公布日 2016.11.23

(30)优先权数据
61/972,082 2014.03.28 US
14/530,354 2014.10.31 US

(85)PCT国际申请进入国家阶段日
2016.09.27

(86)PCT国际申请的申请数据
PCT/US2015/019587 2015.03.10

(87)PCT国际申请的公布数据
WO2015/148100 EN 2015.10.01

(73)专利权人 甲骨文国际公司
地址 美国加利福尼亚

(72)发明人 Z·拉多维奇 P·罗文斯坦 J·G·约翰逊

(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 边海梅

(51)Int.Cl.
G06F 11/07(2006.01)

(56)对比文件
CN 101923499 A,2010.12.22,
CN 103098034 A,2013.05.08,
US 2013036332 A1,2013.02.07,
US 2013013843 A1,2013.01.10,
US 2013191330 A1,2013.07.25,

审查员 宫玉龙

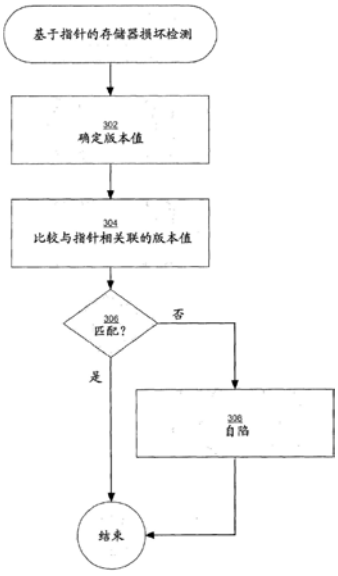
权利要求书2页 说明书11页 附图7页

(54)发明名称

对分布式共享存储器应用的存储器损坏检测支持

(57)摘要

分布式节点系统中的节点被配置为当存储器在节点之间被共享时支持存储器损坏检测。分布式节点系统中的节点在本文称为“共享高速缓存行”的存储器单元中共享数据。节点将版本值与共享高速缓存行中的数据相关联。版本值和数据可以存储在节点的主存储器中的共享高速缓存行中。当节点执行存储器操作时,它可以使用版本值来确定是否已发生存储器损坏。例如,指针可以与版本值相关联。当指针被用来访问存储器时,指针的版本值可以指示在存储器位置处的期望版本值。如果版本值不匹配,则已发生存储器损坏。



CN 106164870 B

1.一种用于检测存储器损坏的方法,包括:

在本地节点的存储器中,生成复制的高速缓存行,该复制的高速缓存行用于维护源节点上的源高速缓存行的副本,其中所述复制的高速缓存行包括版本位和数据位,所述版本位被设置为版本值;

生成指向所述复制的高速缓存行的指针,所述指针包括所述版本值;

利用所述指针在所述复制的高速缓存行上执行存储器操作,其中执行存储器操作包括:

将包括在所述指针中的版本值与所述复制的高速缓存行的版本位被设置为的版本值进行比较;及

基于比较确定是否已发生存储器损坏。

2.如权利要求1所述的方法,其中生成复制的高速缓存行包括:

确定所述源高速缓存行的版本值;及

将所述版本位设置为所述源高速缓存行的版本值。

3.如权利要求2所述的方法,其中所述版本值由所述源节点响应于存储器分配请求而生成。

4.如权利要求1所述的方法,其中对所述版本值进行比较包括:

确定所述复制的高速缓存行是否是无效的;及

响应于确定所述复制的高速缓存行是无效的,将所述源高速缓存行复制到所述复制的高速缓存行。

5.如权利要求4所述的方法,其中所述方法还包括以下步骤:

所述本地节点在存储缓冲区中存储对所述复制的高速缓存行的还没有被传播到所述源高速缓存行的一个或多个更新;及

还响应于确定所述复制的高速缓存行是无效的,将所述一个或多个更新传播到所述复制的高速缓存行。

6.如权利要求1所述的方法,还包括:

如果已发生存储器损坏,则执行自陷操作。

7.如权利要求6所述的方法,其中执行自陷操作包括:

通知应用已发生存储器损坏。

8.如权利要求6所述的方法,其中执行自陷操作包括:

终止所述存储器操作。

9.一种计算机系统,包括:

一个或多个计算节点,其中所述一个或多个计算节点中的每个计算节点被配置为:

在所述每个计算节点的存储器中,生成复制的高速缓存行,所述复制的高速缓存行用于维护属于所述一个或多个计算节点的源节点上的源高速缓存行的副本,其中所述复制的高速缓存行包括版本位和数据位,所述版本位被设置为版本值;

生成指向所述复制的高速缓存行的指针,所述指针包括所述版本值;

使用所述指针在所述复制的高速缓存行上执行存储器操作,其中所述存储器操作包括:

将包括在所述指针中的版本值与所述复制的高速缓存行的版本位被设置为的版本值

进行比较;及

基于比较确定所述复制的高速缓存行是否已被损坏。

10.如权利要求9所述的系统,其中为了生成复制的高速缓存行,所述一个或多个计算节点中的每个计算节点被配置为:

确定所述源高速缓存行的版本值;及

将所述版本位设置为源高速缓存行的版本值。

11.如权利要求10所述的系统,其中对于所述一个或多个计算节点中的每个计算节点,所述版本值由被配置为响应于存储器分配请求而生成版本值的源节点生成。

12.如权利要求9所述的系统,其中对于所述一个或多个计算节点中的每个计算节点,为了对所述版本值进行比较,每个计算节点被配置为:

确定所述复制的高速缓存行是否是无效的;及

如果所述复制的高速缓存行是无效的,则将所述源高速缓存行复制到所述复制的高速缓存行。

13.如权利要求9所述的系统,其中对于所述一个或多个计算节点中的每个计算节点,每个计算节点被配置为:

在存储缓冲区中存储对所述复制的高速缓存行的还没有被传播到所述源高速缓存行的一个或多个更新;及

还响应于确定所述复制的高速缓存行是无效的,将所述一个或多个更新传播到所述复制的高速缓存行。

14.如权利要求9所述的系统,其中对于所述一个或多个计算节点中的每个计算节点,为了执行自陷操作,每个计算节点被配置为如果已发生存储器损坏则执行自陷操作。

15.如权利要求14所述的系统,其中,对于所述一个或多个计算节点中的每个计算节点,为了执行自陷操作,每个计算节点被配置为通知应用已发生存储器损坏。

16.如权利要求14所述的系统,其中为了执行自陷操作,所述一个或多个计算节点中的每个计算节点被配置为终止所述存储器操作。

17.一种或多种存储指令的非临时性存储介质,所述指令在被一个或多个处理器执行时,引起如权利要求1-8中任何一项所述的方法的执行。

对分布式共享存储器应用的存储器损坏检测支持

[0001] 对相关申请的交叉引用;权益要求

[0002] 本申请要求由Zoran Radovic等人于2014年3月28日提交的标题为“Memory Corruption Detection Support For Distributed Shared Memory Applications”的美国临时申请No.61/972,082的优先权,其内容通过引用被结合于此。本申请涉及于2013年3月15日提交的标题为“MEMORY BUS PROTOCOL TO ENABLE CLUSTERING BETWEEN NODES OF DISTINCT PHYSICAL DOMAIN ADDRESS SPACES”、代理人案号为50277-4032的美国专利申请No.13/838,542;于2013年3月15日提交的标题为“REMOTE-KEY BASED MEMORY BUFFER ACCESS CONTROL MECHANISM”、代理人案号为50277-4091的美国专利申请No.13/839,525;以及于2013年3月14日提交的标题为“MEMORY SHARING ACROSS DISTRIBUTED NODES”、代理人案号为50277-4072的美国专利申请No.13/828,555;在本段中的每个申请的内容通过引用被结合于此。

技术领域

[0003] 本公开内容一般涉及用于检测分布式节点系统中的存储器损坏的技术。

背景技术

[0004] 在互联网上或在企业网络上可用的许多功能和服务由分布式计算节点中的一个或多个集群提供。例如,用来运行大规模业务的数据库可以由在形成集群的多个分布式计算节点上运行的多个数据库服务器来维护并且通过其变得可用。利用计算节点的集群提供功能或服务会具有许多优点。例如,利用集群,相对容易添加另一个节点来增加系统的能力,以满足不断增长的需求。集群也使得有可能在各种节点之间进行负载均衡,使得如果一个节点变得负担过重,则工作可以被分配给其它节点。另外,集群使得有可能容忍故障,使得如果一个或多个节点失效,则功能或服务仍然可用。此外,集群中的节点可以能够共享信息,以便例如一起工作和执行事务、负载均衡、实现故障预防和恢复等。

[0005] 对于在集群上运行的应用,可能需要存储器损坏检测。当存储器位置被不恰当地访问或修改时,发生存储器损坏。存储器损坏的一个例子发生在当应用试图推动指针变量超出为特定数据结构分配的存储器时。这些存储器错误会导致程序崩溃或意外的程序结果。

[0006] 对于单机应用存在存储器损坏检测方案。单机存储器损坏检测方案允许计算机在运行时跟踪应用指针并且就存储器错误通知用户。

[0007] 但是,在集群上运行的应用比单机应用更难调试。对于调试在集群上运行的应用存在一些解决方案。这些调试解决方案可能包括内部工具支持、运行时支持、或校验和方案。不幸的是,这些解决方案使编程模型变得复杂并且向系统增加了性能开销,并且可能无法检测到存储器损坏。

[0008] 在本部分中描述的方法是可以实行的方法,但不一定是先前已构想或实行的方法。因此,除非另外指出,否则不应当假定在本部分中描述的任何方法仅仅凭其包括在本部

分中就有资格作为现有技术。

附图说明

[0009] 在附图中：

[0010] 图1是绘出实施例中的示例分布式节点系统的框图；

[0011] 图2示出了根据实施例的其中分布式节点系统中的一些节点是共享存储器的例子；

[0012] 图3是绘出在实施例中用于检测节点中的存储器损坏的过程的流程图；

[0013] 图4是绘出在实施例中用于在检测到存储器损坏时在加载高速缓存行时更新高速缓存行的过程的流程图；

[0014] 图5A是绘出在实施例中用于在远程节点中执行存储的过程的流程图；

[0015] 图5B是绘出在实施例中用于从远程节点向源节点传播存储的过程的流程图；

[0016] 图6是绘出在实施例中用于在源节点中执行存储的过程的流程图；

[0017] 图7是示出其上可以实现本发明的实施例的计算机系统的框图。

具体实施方式

[0018] 在下面的描述中，出于解释的目的而阐述了许多具体细节，以便提供对本发明的透彻理解。但是，将显而易见，本发明也可以在没有这些具体细节的情况下实践。在其它情况下，众所周知的结构和设备以框图的形式示出以避免不必要地模糊本发明。

[0019] 总体概述

[0020] 根据本文描述的实施例，分布式节点系统中的节点被配置为当存储器在节点之间被共享时支持存储器损坏检测。分布式节点系统中的节点在本文称为“共享高速缓存行”的存储器单元中共享数据。节点将版本值与共享高速缓存行中的数据相关联。版本值和数据可以存储在节点的主存储器中的共享高速缓存行中。当节点执行存储器操作时，它可以使用版本值来确定是否已发生存储器损坏。例如，指针可以与版本值相关联。当指针被用来访问存储器时，指针的版本值可以指示在存储器位置处的期望版本值。如果版本值不匹配，则已发生存储器损坏。

[0021] 如该术语在本文中所使用的，指针是包含指向存储器中存储的另一个值的存储器位置的地址的值。该值可加载到处理器的寄存器中。根据实施例，指针包含两个单独的值即版本值和虚拟地址，其中虚拟地址被转换为物理地址以用于执行存储器操作。

[0022] 分布式节点系统中的节点与系统中的其它节点共享它们主存储器的部分。节点(“源节点”)使其主存储器的一部分可用于与系统中的其它节点共享，并且另一个节点(“远程节点”)在其自己的主存储器中复制该共享存储器部分。存储器部分可以包括一个或多个共享高速缓存行。远程节点创建复制的高速缓存行，它是源节点中的源高速缓存行的副本。

[0023] 在实施例中，共享高速缓存行包括版本位和数据位。共享高速缓存行的版本位指示与共享高速缓存行相关联的版本值。配置为指向共享高速缓存行的指针也包含版本值。当指针被用来在共享高速缓存行上执行存储器操作时，节点比较指针的版本值和由共享高速缓存行的版本位指示的版本值。

[0024] 在实施例中，源节点响应于存储器分配请求而生成版本值。例如，如果应用为数据

结构分配存储器，则源节点可以生成要与那个数据结构相关联的版本值。生成的版本值和相关联的数据结构可以被复制在本地节点的主存储器中。

[0025] 在实施例中，存储器操作是通过应用请求的。如果节点检测到已发生存储器破坏，则节点可以就错误通知应用。节点也可以终止存储器操作而不是执行它。

[0026] 在另一种实施例中，节点使用版本值来维护节点之间的一致性。例如，在远程高速缓存行中的版本值可以指示该远程高速缓存行已过时。远程节点然后可以根据对应的源高速缓存行更新该远程高速缓存行。在实施例中，一个或多个版本值被保留用于指示何时复制的高速缓存行是无效的。当节点响应于存储器分配请求而生成版本值时，一个或多个保留的版本值不被使用。

[0027] 系统概述

[0028] 图1示出了在实施例中的示例分布式节点系统100的框图。分布式节点系统100包括三个节点：节点1 102A、节点2 102B和节点3 102C。虽然在本说明中示出了三个节点，但是系统100可以包括更多或更少的节点。

[0029] 每个节点102包括主存储器108。主存储器108包括一个或多个共享高速缓存行106。在实施例中，共享高速缓存行106包括版本位112和数据位114。数据被存储在数据位114中。版本位112指示与共享高速缓存行106相关联的版本值。共享高速缓存行106可以是相同的大小或者大小可以不同。

[0030] 节点102可以使其主存储器108的一部分可用于与其它节点共享(“共享存储器部分”)。另一个节点102可以分配其主存储器108的一部分(“复制的存储器部分”)用于复制共享存储器部分的内容。在实施例中，节点102既可以使其主存储器108的一部分可用于共享，又可以复制由另一个节点102变得可用的主存储器108的一部分。对于本发明的目的，节点102可以共享任何数量的存储器部分(零个或多个)，并且可以复制任意数量的共享存储器部分(零个或多个)。每个存储器部分可以包括一个或多个共享高速缓存行106。在实施例中，共享或复制主存储器108的一部分分别包括共享或复制一个或多个共享高速缓存行106。

[0031] 作为例子，在图2中，节点2 102B正在使其主存储器108B的一部分可用于与其它节点共享。节点1和3正在复制共享存储器部分202。因此，节点1 102A在其主存储器108A中具有存储器部分204A，它是共享存储器部分202的副本，并且节点3 102C在主存储器108C中具有存储器部分204C，它是共享存储器部分202的副本。节点3 102C也正在使其主存储器108C的一部分可用于与其它节点共享。节点1和节点2正在复制共享存储器部分206。因此，节点2 102B具有存储器部分208B，它是共享存储器部分206的副本，并且节点1 102A具有存储器部分208A，它是共享存储器部分206的副本。在示出的例子中，节点2和节点3既在共享存储器部分，又在复制来自另一个节点的共享存储器部分。节点1正在从两个节点复制存储器部分，但是没有共享存储器部分。

[0032] 在实施例中，节点102可以包括目录210。对于每个共享存储器部分，目录210指示系统100中哪些节点包含那个共享存储器部分的副本。在实施例中，目录210包含用于共享存储器部分中每个源高速缓存行的条目。即，目录210包含用于其中节点102是源节点的每个共享高速缓存行的条目。

[0033] 在实施例中，节点102可以包括索引212。对于每个共享存储部分，索引212指示目

录在该共享存储器部分的主存储器108中的位置。对于每个复制的存储器部分，索引212还指示共享了该存储器部分的源节点和该共享存储器部分在源节点的主存储器中的位置。在实施例中，索引212包含用于主存储器108中每个共享高速缓存行的条目。对于复制的存储器部分中的每个共享高速缓存行，索引212指示共享了源高速缓存行的源节点和该源高速缓存行在源节点的主存储器中的的位置。

[0034] 系统初始化

[0035] 为了准备系统100中的节点102以共享存储器，节点102被初始化。在实施例中，节点102可以以下面描述的方式被初始化。节点102可以共享任何数量的存储器部分，并且可以复制任意数量由其它节点共享的存储器部分。取决于节点102决定做什么，它可以执行所描述的操作中的一些操作、全部操作或者不执行任何所描述的操作。

[0036] 在初始化期间，节点102确定它是否希望使其主存储器108的任何部分可用于与系统100中的其它节点共享。如果它确实希望，则节点102向其它节点102广播信息，指示其愿意共享其主存储器的一部分。广播的信息可以包括关于节点102、共享存储器部分202的大小、以及存储器部分202在主存储器108上位于哪里的信息。该信息向系统100中的其它节点指示在哪里访问共享存储器位置。

[0037] 节点102可以接收指示另一个节点希望共享其主存储器的一部分的广播信息。响应于接收到广播信息，节点102可以决定是否要复制或不复制共享存储器部分202。如果节点102决定复制该共享存储器部分，则节点将分配足以存储共享存储器部分的副本的复制的存储器部分。

[0038] 在实施例中，节点102不利用数据填充分配的存储器。即，节点只分配存储器，但不从共享存储器部分复制数据。节点将用于复制的存储器部分中的每个复制的高速缓存行的版本值设置为指示复制的高速缓存行是无效的值。在实施例中，直到应用请求数据，节点102才将来自共享存储器部分的数据复制到其存储器部分的副本中。当节点试图执行针对复制的高速缓存行的操作时，版本值将向节点指示该共享高速缓存行是无效的。节点然后可以将源高速缓存行从共享存储器部分复制到复制的存储器部分中的复制的高速缓存行中。

[0039] 在实施例中，如果节点102正共享其主存储器108的一部分，则该节点在主存储器108中分配存储器用于存储目录结构210。目录结构210指示哪些节点包含由节点102共享的每个存储器部分的副本。在实施例中，目录结构210包含用于在共享存储器部分中每个共享高速缓存行的目录条目。换句话说，每个源高速缓存行与目录条目相关联。因此，对于每个源高速缓存行，目录条目指示哪些其它节点具有应该是那个源高速缓存行的副本的复制的高速缓存行。在实施例中，目录条目也可以指示远程节点中每个复制的高速缓存行是否是有效(最新)副本。在实施例中，目录条目可以包括串行化对目录条目的访问的锁(lock)。

[0040] 在实施例中，节点102为索引结构212在其主存储器108中分配存储器。索引结构212包含用于主存储器108中每个共享高速缓存行的索引条目。如果节点102正在共享共享存储器部分中的共享高速缓存行，则索引条目为共享高速缓存行指示目录条目在主存储器108中的位置。如果共享高速缓存行处于复制的存储器部分中，则索引条目指示共享了共享存储器部分的源节点和对应的源高速缓存行在源节点的主存储器中的位置。在实施例中，如果节点102决定在从源节点接收到广播信息时复制共享存储器部分，则它更新索引结构

212。从源节点接收到的信息可以对应于存储在索引结构212中的信息。

[0041] 示例性存储器分配

[0042] 在实施例中，当存储器被分配时，节点102给存储器位置指派版本值。例如，当应用执行malloc请求时，节点102分配所请求的存储器量、生成与所分配的存储器相关联的版本值、并且将指针返回给应用。在实施例中，所分配的存储器位置包括一个或多个共享高速缓存行。版本值可以由每个共享高速缓存行的版本位指示。

[0043] 在实施例中，版本值由应用的堆管理器(heap manager)生成。版本值可以从有效值的范围中选择。在实施例中，一个或多个版本值被用来指示何时共享高速缓存行是无效的，并且不包括在从中选择的有效值的范围内。版本值的格式可以取决于实现方式而不同。例如，版本值可以是四位长，从而导致十六种可能的值。在另一个例子中，版本值可以是44位的时间戳。

[0044] 版本值也与指向所分配的存储器的指针相关联。在实施例中，指针包括版本值和虚拟地址。例如，节点可以使用44位寄存器来存储指针，但是虚拟地址不使用该整个44位。版本值可以存储在44位寄存器的额外未使用位中。

[0045] 如果所分配的存储器作为共享存储器部分的一部分被共享，则其它节点102可以将在所分配的存储器位置中的共享高速缓存行复制到其各自的复制的存储器部分中。在实施例中，复制共享高速缓存行包括复制相关联的版本值。其它节点102也可以生成指向复制的共享高速缓存行的指针。版本值可以与每个生成的指针相关联被存储。

[0046] 基于指针的存储器损坏检测

[0047] 图3是示出用于利用与指针相关联的版本值检测节点102中存储器损坏的过程的流程图。该过程可以当执行涉及由其中与版本值相关联的指针引用的共享高速缓存行的存储器操作时执行。该过程可以在下文中被称为基于指针的存储器损坏检测。

[0048] 例如，节点102从应用接收命令。该命令可以是例如执行存储器操作的请求，诸如加载或存储命令。在命令的执行期间，节点102执行用于检测存储器损坏的步骤。命令可以包括指向主存储器108中的共享高速缓存行的指针。如以上所讨论的，在实施例中，当节点102向应用分配存储器时，节点返回与版本值相关联的指针。

[0049] 在步骤302中，节点102确定与包括在命令中的指针相关联的版本值。在实施例中，指针包括版本值。与指针相关联的版本值可以指示命令期望要与所请求的共享高速缓存行相关联的版本值。例如，如果命令在使用指针访问数据结构，则版本值可以与该数据结构相关联。

[0050] 在步骤304中，节点102将指针的版本值与和所请求的共享高速缓存行相关联的版本值进行比较。在实施例中，共享高速缓存行的版本位指示与该共享高速缓存行相关联的版本值。该方法然后前进到决定框308。

[0051] 在决定框308处，如果指针的版本值与和所请求的共享行相关联的版本值不匹配，则存储器损坏被检测到。在实施例中，执行自陷(trap)操作。自陷操作可以包括向应用指示存储器损坏被检测到。自陷操作也可以包括终止存储器操作的执行。可替代地，该过程结束并且存储器操作继续。

[0052] 如果指针的版本值与和所请求的共享行相关联的版本值匹配，则该过程结束并且存储器操作继续。

[0053] 在图3中示出的用于利用与指针相关联的版本值检测存储器损坏的过程可以在执行各种存储器操作时执行。将进一步详细描述这些存储器操作。

[0054] 节点之间的一致性

[0055] 在实施例中，共享高速缓存行中的版本值也可以被用来管理节点之间的共享高速缓存行的一致性。当源节点更新源高速缓存行时，远程节点中的复制的高速缓存行将会过时。但是，远程节点可以不立即更新其复制的高速缓存行。而是，每个复制的高速缓存行的版本值被设置以指示该复制的高速缓存行是无效的。以后，如果远程节点试图访问该复制的高速缓存行，则节点将看到该复制的高速缓存行是无效的并且将更新该复制的高速缓存行。

[0056] 在实施例中，当节点102执行存储命令时，它可以执行自陷操作。在实施例中，取决于目标共享高速缓存行是源高速缓存行还是复制的高速缓存行，节点102将执行不同的步骤。如果目标共享高速缓存行是复制的高速缓存行，则节点102将把存储传播到源节点中的源高速缓存行。在实施例中，在将存储发送到源节点之前，远程节点可以在存储缓冲区中记录该存储。

[0057] 在实施例中，节点102包含索引212。如果所请求的共享高速缓存行是复制的高速缓存行，则索引条目将指示源节点和用于复制的高速缓存行的源高速缓存行的位置。因此，节点102可以引用索引212来确定所请求的共享高速缓存行是复制的高速缓存行还是源高速缓存行。基于该确定，节点102可以确定采取哪些步骤来执行存储命令。

[0058] 远程节点加载

[0059] 在实施例中，当源节点更新源高速缓存行时，节点不更新对应的复制的高速缓存行。当复制的高速缓存行在节点处被加载时，节点可以只更新复制的高速缓存行。指示复制的高速缓存行无效的版本值触发更新。当复制的高速缓存行被更新时，存储器损坏检测被执行。图4是示出当在节点102中请求复制的高速缓存行时用于更新共享高速缓存行的过程的流程图。

[0060] 在步骤402中，节点102从应用接收命令。例如，命令可以是涉及加载操作的存储器操作，诸如加载命令。

[0061] 命令可以包括指向主存储器108中的共享高速缓存行的指针。如以上所讨论的，在实施例中，当节点102向应用分配存储器时，节点返回与版本值相关联的指针。出于本说明的目的，将假定包含在命令中的指针与版本值相关联。

[0062] 在步骤404中，节点102确定版本值是否指示共享高速缓存行是无效的。在实施例中，至少一个版本值被用来指示共享高速缓存行是无效的并且不在存储器分配期间被使用。在实施例中，共享高速缓存行是复制的高速缓存行。如果例如复制的高速缓存行还没有用来自源高速缓存行的数据填充，则版本值可以指示共享高速缓存行是无效的。所请求的共享高速缓存行可以是复制的高速缓存行或可以不是复制的高速缓存行。

[0063] 在一个例子中，共享高速缓存行不是复制的高速缓存行。在实施例中，不在复制的存储器部分中的共享高速缓存行被假定为始终是有效的。

[0064] 在另一个例子中，共享高速缓存行是复制的高速缓存行。共享高速缓存行中的数据可能会过时。即，复制的高速缓存行中的数据与源高速缓存行中的数据不同。这会例如当源节点存储数据到源高速缓存行时发生。

[0065] 该方法然后前进到决定框406。在决定框406处，如果版本值指示共享高速缓存行是有效的，则节点102继续过程的执行并且前进到步骤410，在该步骤中，基于指针的存储器损坏检测被执行。

[0066] 如果版本值指示共享高速缓存行是无效的，则方法前进到步骤408。在步骤408中，节点使命令的执行挂起并且执行自陷操作。

[0067] 在实施例中，自陷操作包括将源高速缓存行复制到复制的高速缓存行。复制源高速缓存行可以包括复制源高速缓存行的版本位和数据位。因此，在复制被执行之后，复制的高速缓存行的版本值被设置为来自源高速缓存行的版本值。在复制的高速缓存行中的数据被设置为在源高速缓存行中包含的最近数据，如由远程节点对复制的高速缓存行做出的还没有被传播到源高速缓存行的记录到存储缓冲区中的任何存储所修改的那样。因此，节点能够更新共享高速缓存行中的数据，以便维护与其它节点的一致性。

[0068] 在实施例中，节点包含索引212。节点可以使用对应于所请求的共享高速缓存行的索引条目，以便确定哪些源节点包含对应的源高速缓存行以及对应的源高速缓存行位于源节点的主存储器中哪里。

[0069] 在实施例中，源节点包含目录210。当远程节点更新其复制的高速缓存行时，源节点可以更新用于对应的源高速缓存行的目录条目，以指示在远程节点处的副本是有效副本。

[0070] 远程节点存储

[0071] 如前面提到的，在实施例中，在存储被发送到源节点之前，远程节点使用存储缓冲区来记录该存储。图5A是示出由分布式节点系统100中的远程节点102执行的存储的流程图。该存储可以被执行，以执行存储命令。命令可以包括指向主存储器108中复制的高速缓存行的指针。指针可以与版本值相关联。

[0072] 在步骤502中，节点102使命令的执行挂起并且执行自陷操作来执行以下步骤。

[0073] 在步骤504处，存储被记录在存储缓冲区中。记录在存储缓冲区中的信息可以指示向其执行存储的存储器位置和要存储什么数据。在存储缓冲区中记录存储可以包括指示源节点和应该向其执行存储的源节点的主存储器中源高速缓存行的位置、存储线程、以及与(一个或多个)存储相关联的版本号。

[0074] 在实施例中，节点102包含索引212。节点可以使用对应于所请求的共享高速缓存行的索引，以便确定哪个源节点包含对应的源高速缓存行以及对应的源高速缓存行位于源节点的主存储器中哪里。

[0075] 在步骤506中，节点102确定复制的高速缓存行的版本值是否指示复制的高速缓存行是无效的。如果版本值指示共享高速缓存行是无效的，则存储不对该共享高速缓存行执行。如果该值指示复制的高速缓存行有效，则该方法前进到步骤508。

[0076] 在508处，节点102执行基于指针的存储器损坏检测。如果由节点102执行的基于指针的存储器损坏检测没有检测到存储器损坏，则该方法前进到步骤510。

[0077] 在步骤510处，节点102将数据存储在它的共享高速缓存行中。

[0078] 自陷操作结束。

[0079] 更新传播

[0080] 在实施例中，远程节点将存储记录在其存储缓冲区中，但不将存储发送到包含对

应的源高速缓存行的源节点。在节点将存储记录在其存储缓冲区之后，该存储需要被传播到源节点。传播该存储可以作为与记录存储缓冲区相同的过程的一部分来执行，或者它可以被单独地执行。在实施例中，节点可以接收包括传播存储操作的命令。例如，存储命令可以包括传播该存储的指令。存储可以在自陷操作完成之后作为恢复存储命令的执行的一部分被传播。在另一种实施例中，节点102可以在向共享高速缓存行写入之前为条目检查存储缓冲区。图5B是示出分布式节点系统100中的存储传播的流程图。存储可以由另一个执行的线程异步传播。

[0081] 在步骤522处，节点从存储缓冲区中检索条目。条目可以包括指示源节点、应该对其执行存储的源高速缓存行、要被存储的数据、与(一个或多个)存储相关联的版本号以及存储线程的信息。

[0082] 在步骤524处，节点102向源节点请求用于源高速缓存行的远程节点列表。在接收到这一信息之后，该方法前进到步骤526。

[0083] 在实施例中，响应于该请求，源节点引用用于那个共享高速缓存行的目录条目。该目录条目指示哪些节点包含源高速缓存行的副本。系统100中任何数量的节点可以包含源高速缓存行的副本。在实施例中，当访问用于所请求的共享高速缓存行的目录条目时，源节点锁定该目录条目。在实施例中，源节点只共享包含源高速缓存行的有效副本的远程节点列表。目录条目可以被更新，以指示所有远程节点包含无效副本。

[0084] 在步骤526处，节点102使包含源高速缓存行的副本的其它远程节点将其复制的高速缓存行标记为无效。该节点向保持各自复制的高速缓存行的每个节点指示源高速缓存行中的数据已被改变。在远程节点处的复制的高速缓存行的版本值被改变，以指示复制的高速缓存行是无效的。

[0085] 在步骤528处，节点102通知源节点来执行存储。通知可以包括源高速缓存行在源节点的主存储器中的位置、要被存储在源高速缓存行中的数据以及版本号。在执行存储之前，源节点将来自存储缓冲区的版本号与在各自源高速缓存行中的版本号进行比较。如果检测到版本不匹配，则源节点不执行存储并且可以例如经由异步自陷通知发起线程。

[0086] 在步骤530处，存储的数据从存储缓冲区中去除。

[0087] 在实施例中，对存储缓冲区中的每个条目重复这些步骤。

[0088] 在可替代的实施例中，远程节点不将存储记录在存储缓冲区中。而是，远程节点在自陷操作执行期间执行更新传播步骤，来代替向存储缓冲区写入。

[0089] 源节点存储

[0090] 在实施例中，源节点在不使用存储缓冲区的情况下执行存储命令来存储共享高速缓存行。图6是示出由源节点102执行的在分布式节点系统100中执行存储命令的步骤的流程图。存储命令可以包括指向主存储器108中的源高速缓存行的指针。指针可以与版本值相关联。

[0091] 在步骤602处，节点102使存储命令的执行挂起，并且执行自陷操作。

[0092] 在步骤604处，节点102为源高速缓存行执行基于指针的存储器损坏检测。如果没有检测到存储器损坏，则该方法前进到步骤606。如果检测到存储器损坏，则该方法退出自陷操作而不执行存储。

[0093] 在步骤606中，节点102指示远程节点使其各自复制的高速缓存行无效。节点102向

每个远程节点指示源高速缓存行中的数据已被改变。在远程节点处复制的高速缓存行的版本值被改变，以指示复制的高速缓存行是无效的。

[0094] 在实施例中，源节点引用用于那个共享高速缓存行的目录条目。该目录条目指示哪些节点包含源高速缓存行的副本。系统100中任何数量的节点可以包含源高速缓存行的副本。该节点向正在复制源高速缓存行的每个节点指示数据已被改变。在其它节点处复制的高速缓存行的版本值被改变，以指示源高速缓存行的副本是无效的。

[0095] 在实施例中，使源高速缓存行无效被记录并且到远程节点的无效化指令被延迟地发送。例如，除执行存储的线程之外的其它线程发现无效源高速缓存行的记录，并且发送指令到远程节点，以使源高速缓存行的复制的高速缓存行无效。

[0096] 在步骤608处，源节点在源高速缓存行上执行存储。

[0097] 源节点完成自陷操作。

[0098] 硬件概述

[0099] 根据一种实施例，本文所描述的技术由一个或多个专用计算设备实现。专用计算设备可以是硬连线的以执行所述技术，或者可以包括诸如被永久性地编程以执行所述技术的一个或多个专用集成电路(ASIC)或现场可编程门阵列(FPGA)的数字电子设备，或者可以包括编程为按照固件、存储器、其它存储装置或者其组合中的程序指令执行所述技术的一个或多个通用硬件处理器。这种专用计算设备还可以组合定制的硬连线逻辑、ASIC或FPGA与定制的编程来实现所述技术。专用计算设备可以是台式计算机系统、便携式计算机系统、手持式设备、联网设备或者结合硬连线和/或程序逻辑来实现所述技术的任何其它设备。

[0100] 例如，图7是说明本发明的实施例可以在其上实现的计算机系统700的框图。计算机系统700包括总线702或者用于传送信息的其它通信机制，以及与总线702耦合用于处理信息的硬件处理器704。硬件处理器704可以是例如通用微处理器。

[0101] 计算机系统700还包括耦合到总线702用于存储信息和要由处理器704执行的指令的主存储器706，诸如随机存取存储器(RAM)或其它动态存储设备。主存储器706还可以用于在要由处理器704执行的指令执行期间存储临时变量或其它中间信息。当存储在处理器704可访问的非临时性存储介质中时，这种指令使计算机系统700变成为被定制以执行指令中所指定的操作的专用机器。

[0102] 计算机系统700还包括耦合到总线702的只读存储器(ROM)708或者其它静态存储设备，用于为处理器704存储静态信息和指令。提供了诸如磁盘、光盘或固态驱动器的存储设备710，并且存储设备710耦合到总线702，用于存储信息和指令。

[0103] 计算机系统700可以经总线702耦合到显示器712(诸如阴极射线管(CRT))，用于向计算机用户显示信息。包括字母数字和其它键的输入设备714耦合到总线702，用于向处理器704传送信息和命令选择。另一种类型的用户输入设备是光标控件716，诸如鼠标、轨迹球或者光标方向键，用于向处理器704传送方向信息和命令选择并且用于控制光标在显示器712上的运动。这种输入设备通常具有在两个轴即第一个轴(例如，x)和第二个轴(例如，y)中的两个自由度，以允许设备在平面内指定位置。

[0104] 计算机系统700可以利用定制的硬连线逻辑、一个或多个ASIC或FPGA、固件和/或程序逻辑来实现本文所述的技术，这些与计算机系统相结合使计算机系统700或者把计算机系统700编程为专用机器。根据一种实施例，本文的技术由计算机系统700响应于处理器

704执行包含在主存储器706中的一条或多条指令的一个或多个序列而执行。这种指令可以从另一存储介质(诸如存储设备710)读到主存储器706中。包含在主存储器706中的指令序列的执行使处理器704执行本文所述的过程步骤。在可替代的实施例中，硬连线的电路系统可以代替软件指令或者与其结合使用。

[0105] 如在本文所使用的，术语“存储介质”指存储使机器以特定方式操作的数据和/或指令的任何非临时性介质。这种存储介质可以包括非易失性介质和/或易失性介质。非易失性介质包括例如光盘、磁盘，或固态驱动器，诸如存储设备710。易失性介质包括动态存储器，诸如主存储器706。存储介质的常见形式包括例如软盘、柔性盘、硬盘、固态驱动器、磁带或者任何其它磁性数据存储介质、CD-ROM、任何其它光学数据存储介质，任何具有孔模式的物理介质、RAM、PROM和EPROM、FLASH-EPROM、NVRAM、任何其它存储器芯片或盒式磁带。

[0106] 存储介质与传输介质不同但是可以与其结合使用。传输介质参与在存储介质之间传送信息。例如，传输介质包括同轴电缆、铜线和光纤，包括包含总线702的配线。传输介质还可以采取声波或光波的形式，诸如在无线电波和红外线数据通信中产生的那些。

[0107] 各种形式的介质可以涉及把一条或多条指令的一个或多个序列携带到处理器704供执行。例如，指令最初可以承载在远程计算机的磁盘或固态驱动器上。远程计算机可以把指令加载到其动态存储器中并且利用调制解调器经电话线发送指令。位于计算机系统700本地的调制解调器可以在电话线上接收数据并且使用红外线发送器把数据转换成红外线信号。红外线检测器可以接收在红外线信号中携带的数据并且适当的电路系统可以把数据放在总线702上。总线702把数据携带到主存储器706，处理器704从主存储器706检索并执行指令。由主存储器706接收到的指令可以可选地在被处理器704执行之前或之后存储在存储设备710上。

[0108] 计算机系统700还包括耦合到总线702的通信接口718。通信接口718提供耦合到网络链路720的双向数据通信，其中网络链路720连接到本地网络722。例如，通信接口718可以是综合业务数字网络(ISDN)卡、电缆调制解调器、卫星调制解调器，或者提供到对应类型电话线的数据通信连接的调制解调器。作为另一个例子，通信接口718可以是提供到兼容的局域网(LAN)的数据通信连接的LAN卡。也可以实现无线链路。在任何此类实现中，通信接口718发送和接收携带表示各种类型的信息的数字数据流的电信号、电磁信号或光信号。

[0109] 网络链路720通常通过一个或多个网络向其它数据设备提供数据通信。例如，网络链路720可以通过本地网络722提供到主计算机724或者到由互联网服务提供商(ISP)726操作的数据设备的连接。ISP 726又通过现在通常称为“互联网”728的全球分组数据通信网络提供数据通信服务。本地网络722和互联网728两者都使用携带数字数据流的电信号、电磁信号或光信号。通过各种网络的信号以及在网络链路720上并通过通信接口718的信号是传输介质的示例形式，其中这些信号把数字数据携带到计算机系统700或者携带来自计算机系统700的数字数据。

[0110] 计算机系统700可以通过(一个或多个)网络、网络链路720和通信接口718发送消息和接收数据，包括程序代码。在互联网例子中，服务器730可以通过互联网728、ISP 726、本地网络722和通信接口718发送对应用程序的请求代码。

[0111] 接收到的代码可以由处理器704在它被接收到时执行、和/或存储在存储设备710或其它非易失性存储装置中用于以后执行。

[0112] 在前面的说明书中，本发明的实施例已经参考许多具体细节进行了描述，这些细节可以从一种实现到另一种实现而不同。因此，说明书和附图应当在说明性而不是限制性的意义上加以考虑。本发明范围的唯一且排他指示以及申请人预期要作为本发明范围的是由本申请产生的权利要求集合的字面和等效范围，以这种权利要求产生的具体形式，包括任何后续的校正。

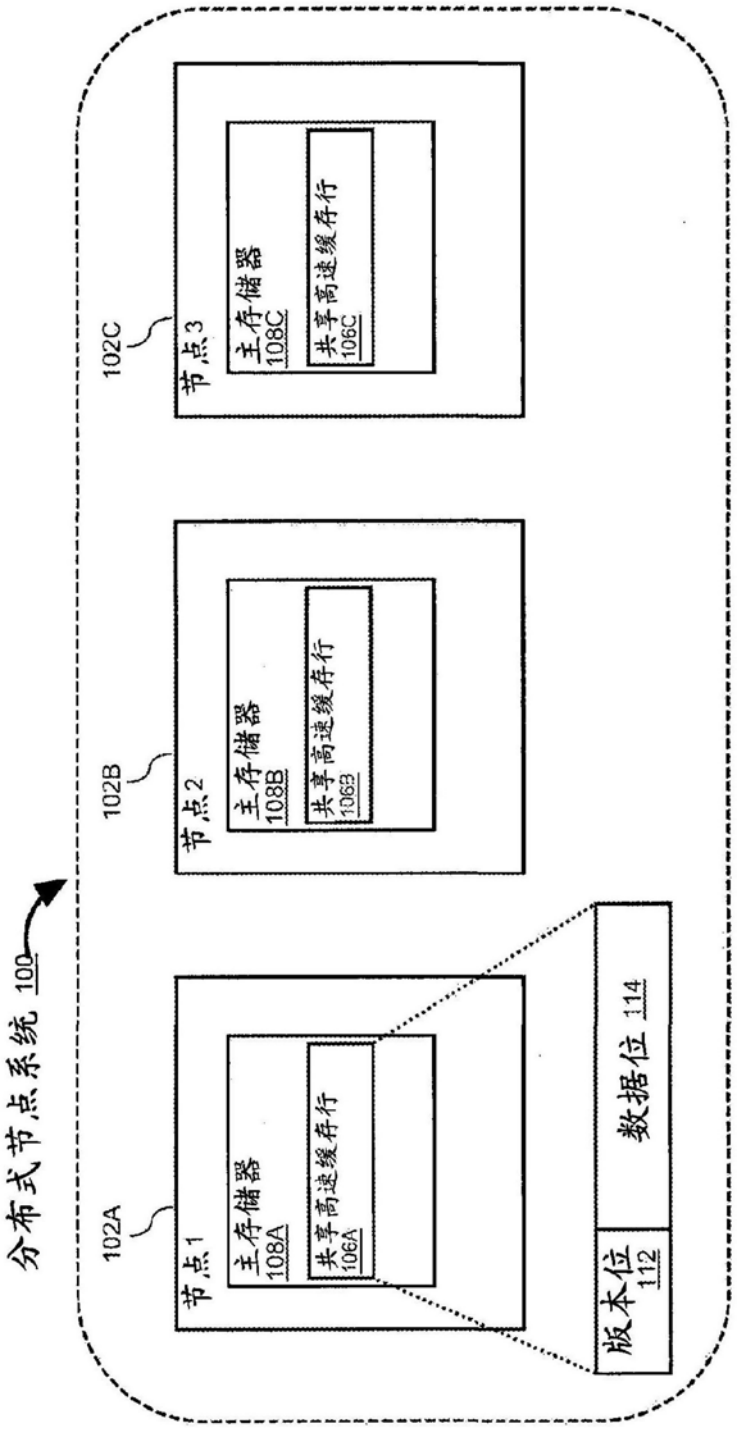

分布式节点系统 100
102A
节点1
主存储器 108A
共享高速缓存行 106A
102B
节点2
主存储器 108B
共享高速缓存行 1063
102C
节点3
主存储器 108C
共享高速缓存行 106C
版本位 112
数据位 114

图1

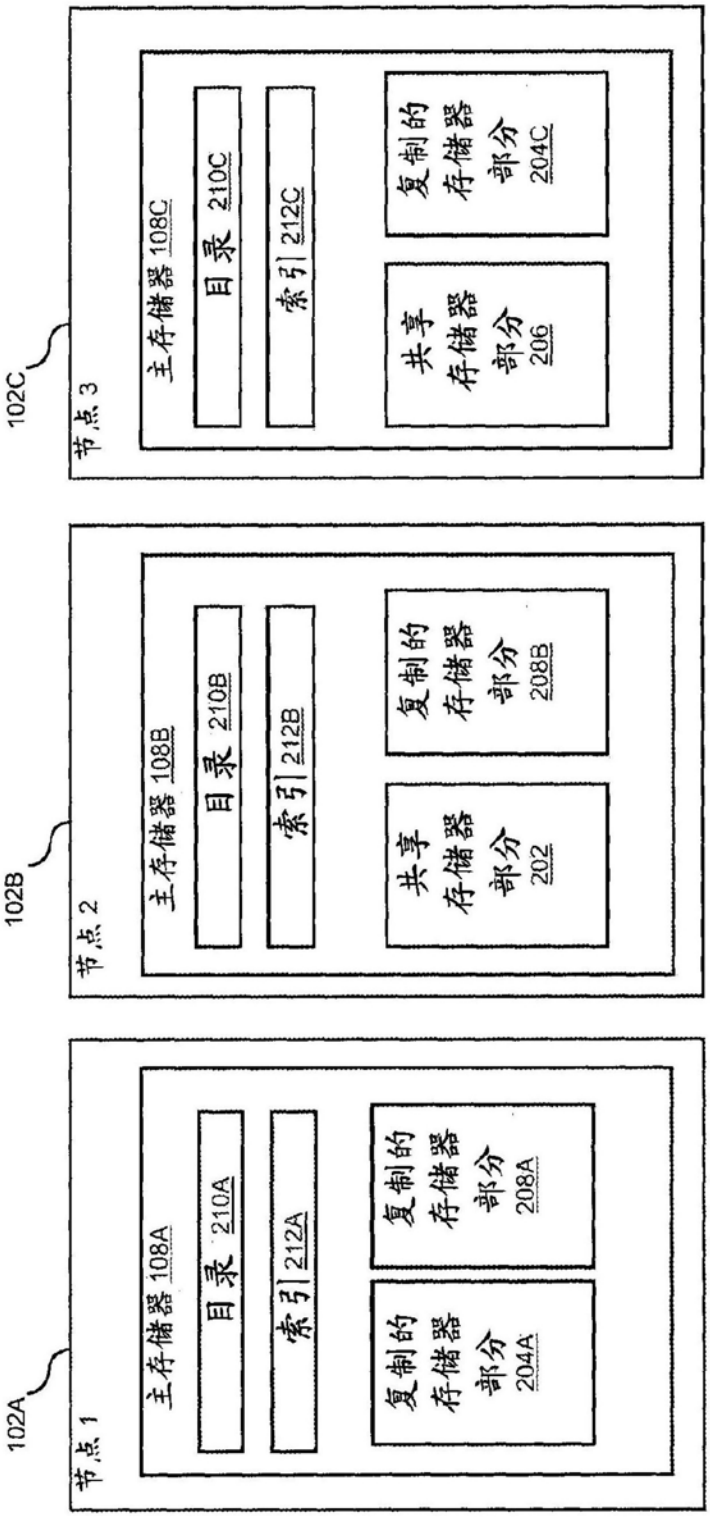

102A
节点 1
主存储器 108A
目录 210A
索引 212A
复制的存储器部分 204A
复制的存储器部分 208A
102B
节点 2
主存储器 108B
目录 210B
索引 212B
共享存储器部分 202
复制的存储器部分 208B
102C
节点 3
主存储器 108C
目录 210C
索引 212C
共享存储器部分 206
复制的存储器部分 204C

图2

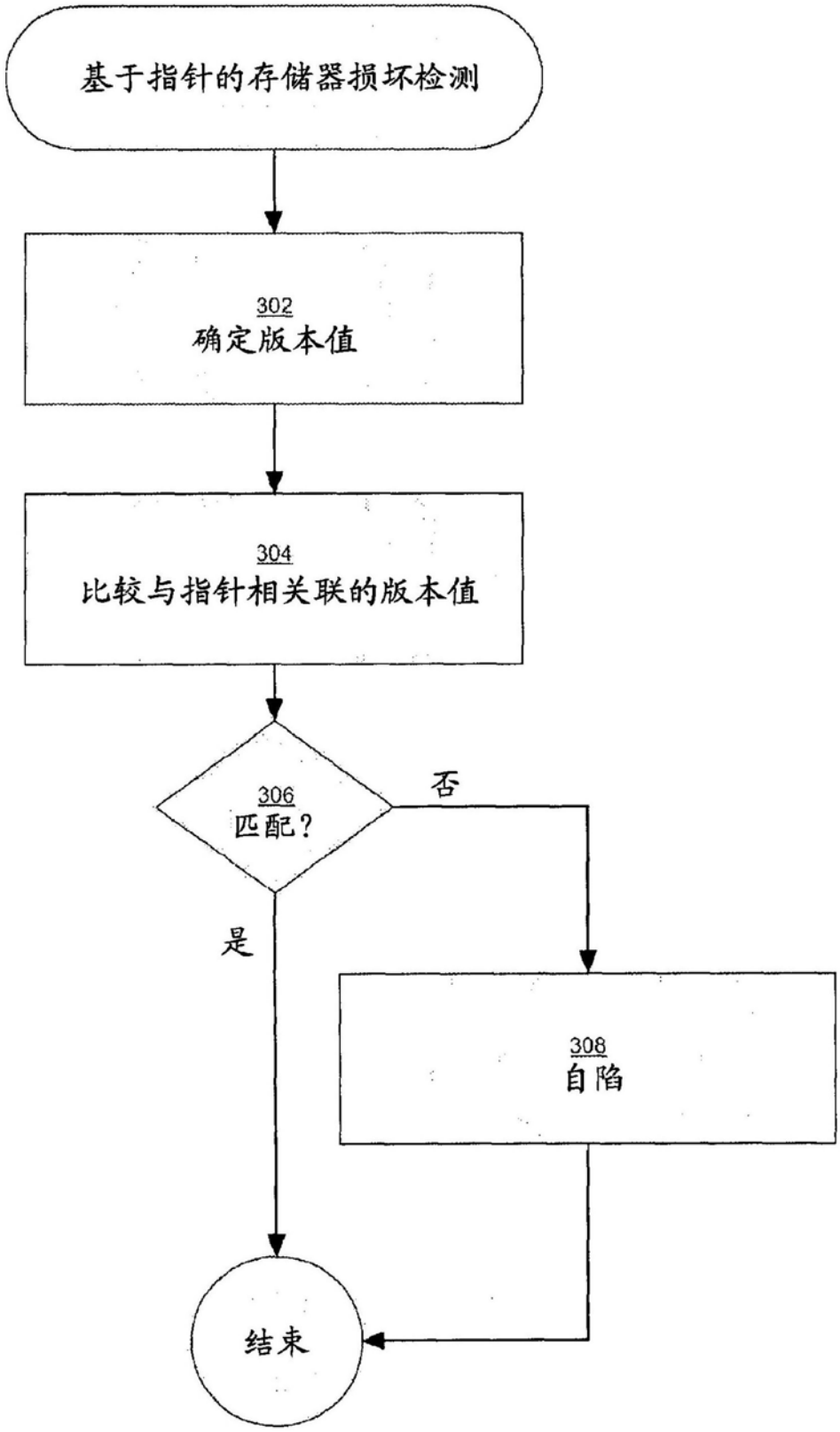


图3

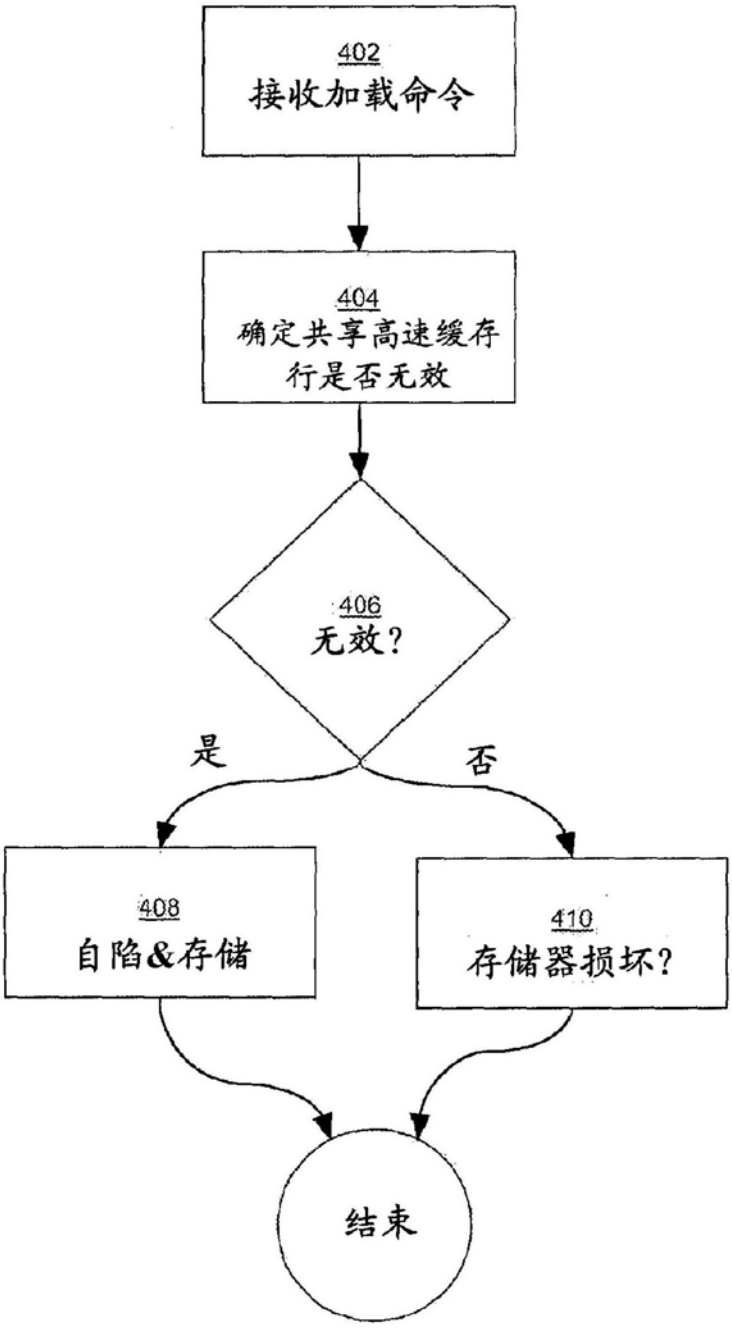


图4

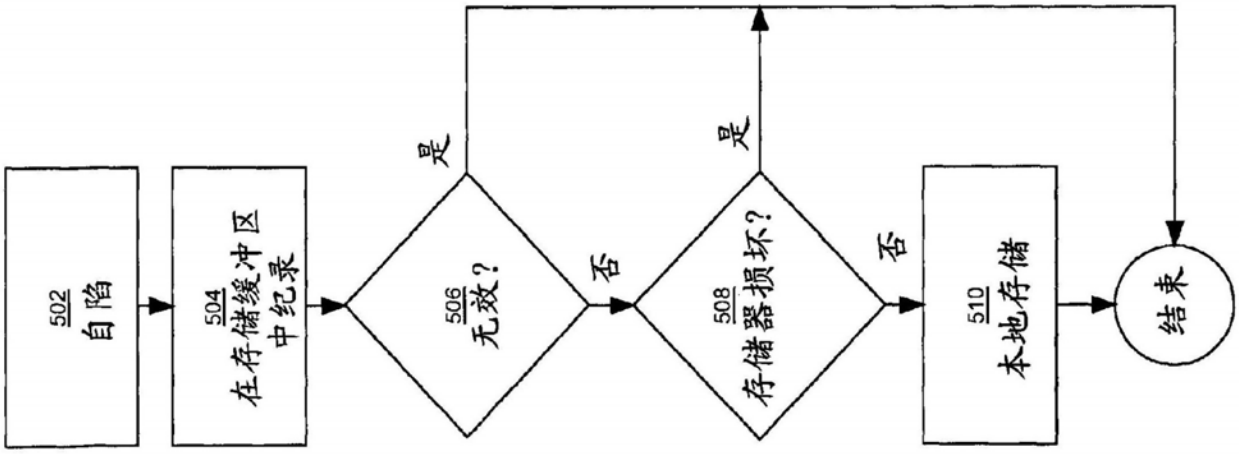

502
自陷
504
在存储缓冲区中纪录
506
无效?
是
否
508
存储器损坏?
是
否
510
本地存储
结束

图5A

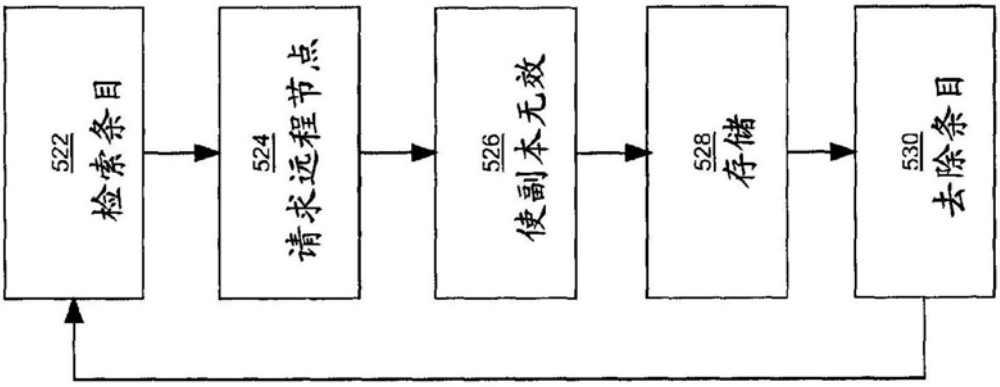

522
检索条目
524
请求远程节点
526
使副本无效
528
存储
530
去除条目

图5B

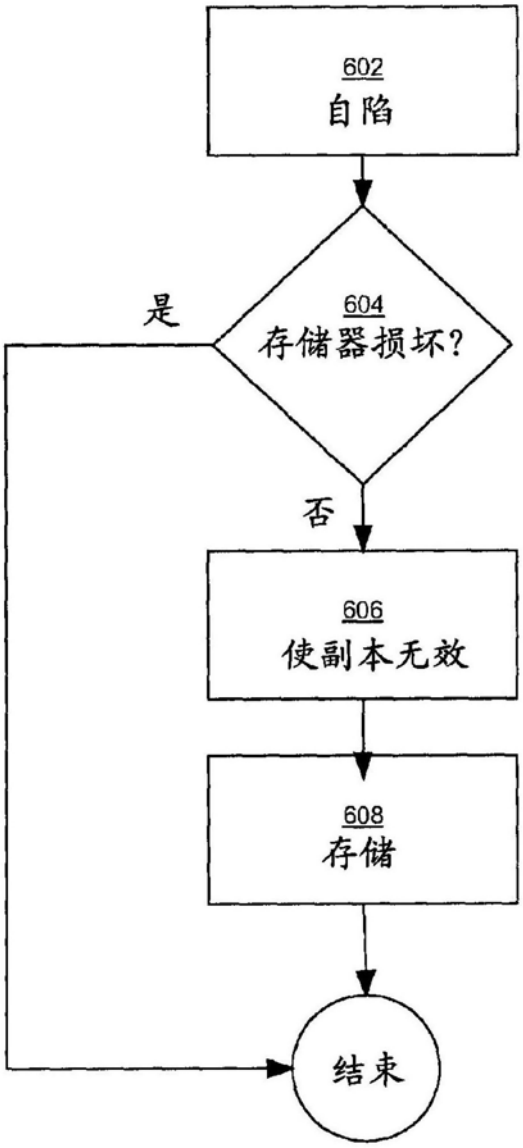


图6

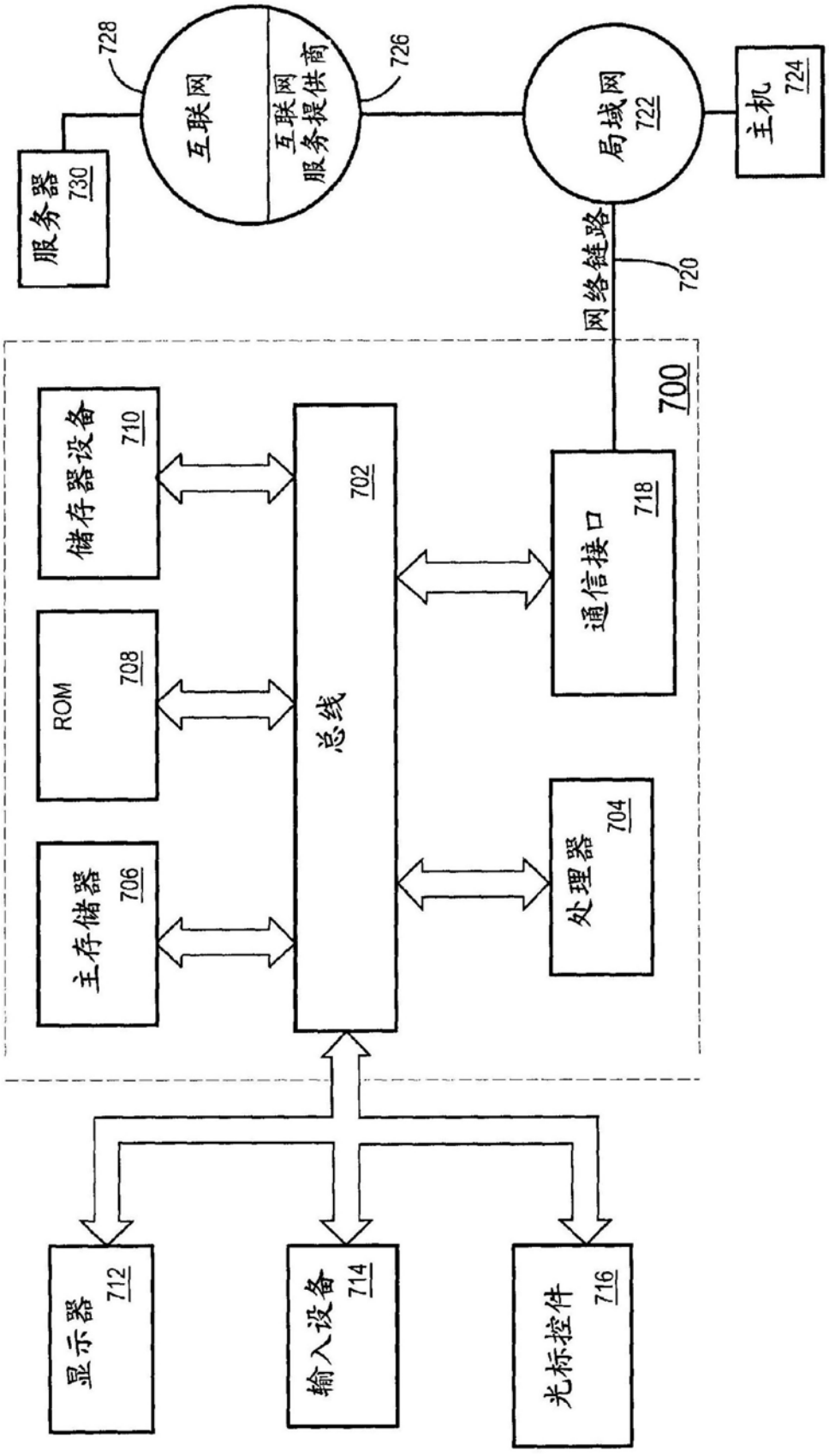

728
互联网
互联网
服务提供商
726
局域网
722
主机
724
服务器
730
网络链路
720
700
储存器设备
710
702
通信接口
718
ROM
708
总线
主存储器
706
处理器
704
显示器
712
输入设备
714
光标控件
716

图7