(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111033479 B
(45)授权公告日 2023.07.25

(21)申请号 201880052360.2
(22)申请日 2018.08.10
(65)同一申请的已公布的文献号
申请公布号 CN 111033479 A
(43)申请公布日 2020.04.17
(30)优先权数据
15/680,577 2017.08.18 US
(85)PCT国际申请进入国家阶段日
2020.02.12
(86)PCT国际申请的申请数据
PCT/IB2018/056028 2018.08.10
(87)PCT国际申请的公布数据
WO2019/034976 EN 2019.02.21
(73)专利权人 国际商业机器公司
地址 美国纽约
(72)发明人 L•古普塔 K•J•阿什 K•安德森
(74)专利代理机构 北京市中咨律师事务所 11247
专利代理师 李永敏 于静
(51)Int.Cl.
G06F 12/08(2016.01)
(56)对比文件
US 2017068624 A1,2017.03.09
US 2017068624 A1,2017.03.09
US 2017097793 A1,2017.04.06
CN 103562887 A,2014.02.05
CN 103946790 A,2014.07.23
US 2014304479 A1,2014.10.09
US 2017091106 A1,2017.03.30
US 9384143 B1,2016.07.05
审查员 杨玉玲

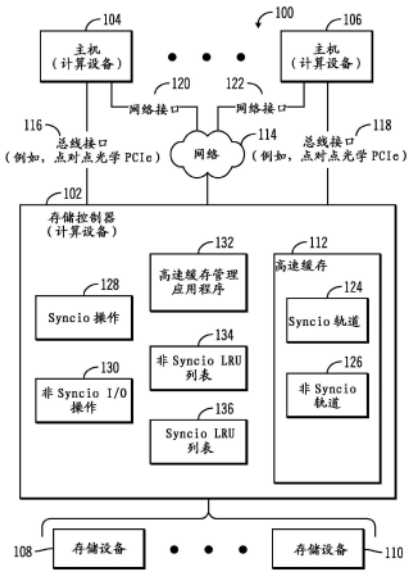
权利要求书3页 说明书10页 附图10页

(54)发明名称

高速缓存管理

(57)摘要

生成高速缓存中第一类型轨道的列表。生成高速缓存中第二类型轨道的列表，其中，相对于第二类型轨道，对第一类型的轨道的I/O操作相对更快地完成。确定是从第一类型的轨道的列表还是从第二类型轨道的列表降级轨道。



CN 111033479 B

1.一种高速缓存管理方法,包括:

在高速缓存中生成第一类型轨道的列表;在所述高速缓存中生成第二类型轨道的列表,其中,基于对轨道执行的最后I/O操作是第一类型I/O操作还是第二类型I/O操作,所述轨道被添加到所述第一类型轨道的所述列表或所述第二类型轨道的所述列表中;以及

确定是从所述第一类型轨道的所述列表还是从所述第二类型轨道的所述列表降级轨道,

其中:

为了对第一类型轨道执行所述第一类型I/O操作,将应用线程保持在旋转循环中,以等待所述第一类型I/O操作完成;以及

为了对第二类型轨道执行所述第二类型I/O操作,避免将所述应用线程保持在所述旋转循环中。

2.如权利要求1所述的方法,其中,所述第一类型轨道是Syncio轨道。

3.如权利要求2所述的方法,其中,所述第二类型轨道是非Syncio轨道。

4.如权利要求3所述的方法,其中,基于所述第一类型轨道的所述列表中的每个轨道的最近被使用的时间来对所述第一类型轨道的所述列表中的轨道进行排序,并且其中,基于所述第二类型轨道的所述列表中的每个轨道的最近被使用的时间来对所述第二类型轨道的所述列表中的轨道进行排序。

5.如权利要求4所述的方法,所述方法还包括:

响应于确定所述第二类型轨道的所述列表中的轨道的读取命中率小于所述第一类型轨道中的轨道的读取命中率的预定因子,将轨道从所述第二类型轨道的所述列表中降级。

6.如权利要求5所述的方法,其中,基于所述第一类型轨道的所述列表和所述第二类型轨道的所述列表的预定底部的读取命中来计算所述第一类型轨道的所述列表中的轨道的读取命中率和所述第二类型轨道的所述列表中的轨道的读取命中率。

7.如权利要求5所述的方法,所述方法还包括:

调整所述预定因子以增加输入/输出(I/O)操作的速率。

8.一种用于高速缓存管理的系统,包括:

存储器;以及

耦合至所述存储器的处理器,其中,所述处理器执行操作,由所述处理器执行的所述操作包括:

在高速缓存中生成第一类型轨道的列表;

在所述高速缓存中生成第二类型轨道的列表,其中,基于对轨道执行的最后I/O操作是第一类型I/O操作还是第二类型I/O操作,将所述轨道添加到所述第一类型轨道的所述列表或所述第二类型轨道的所述列表中;以及

确定是从所述第一类型轨道的所述列表还是从所述第二类型轨道的所述列表降级轨道,

其中:

为了对第一类型轨道执行所述第一类型I/O操作,将应用线程保持在旋转循环中,以等待所述第一类型I/O操作完成;以及

为了对第二类型轨道执行所述第二类型I/O操作,避免将所述应用线程保持在所述旋

转循环中。

9.如权利要求8所述的系统,其中,所述第一类型轨道是Syncio轨道。

10.如权利要求9所述的系统,其中,所述第二类型轨道是非Syncio轨道。

11.如权利要求10所述的系统,其中,基于所述第一类型轨道的所述列表中的每个轨道的最近被使用的时间来对所述第一类型轨道的所述列表中的轨道进行排序,并且其中,基于所述第二类型轨道的所述列表中的每个轨道的最近被使用的时间来对所述第二类型轨道的所述列表中的轨道进行排序。

12.如权利要求11所述的系统,所述操作还包括:

响应于确定所述第二类型轨道的所述列表中的轨道的读取命中率小于所述第一类型轨道中的轨道的读取命中率的预定因子,将轨道从所述第二类型轨道的所述列表中降级。

13.如权利要求12所述的系统,其中,基于所述第一类型轨道的所述列表和所述第二类型轨道的所述列表的预定底部的读取命中来计算所述第一类型轨道的所述列表中的轨道的读取命中率和所述第二类型轨道的所述列表中的轨道的读取命中率。

14.如权利要求12所述的系统,所述操作还包括:

调整所述预定因子以增加输入/输出(I/O)操作的速率。

15.一种用于高速缓存管理的计算机可读存储介质,所述计算机可读存储介质具有包含在其中的计算机可读程序代码,所述计算机可读程序代码被配置为执行操作,所述操作包括:

在高速缓存中生成第一类型轨道的列表;

在所述高速缓存中生成第二类型轨道的列表,其中,基于对轨道执行的最后I/O操作是第一类型I/O操作还是第二类型I/O操作,将所述轨道添加到所述第一类型轨道的所述列表或所述第二类型轨道的所述列表中;以及

确定是从所述第一类型轨道的所述列表还是从所述第二类型轨道的所述列表降级轨道,

其中:

为了对第一类型轨道执行所述第一类型I/O操作,将应用线程保持在旋转循环中,以等待所述第一类型I/O操作完成;以及

为了对第二类型轨道执行所述第二类型I/O操作,避免将所述应用线程保持在所述旋转循环中。

16.如权利要求15所述的计算机可读存储介质,其中,所述第一类型轨道是Syncio轨道。

17.如权利要求16所述的计算机可读存储介质,其中,所述第二类型轨道是非Syncio轨道。

18.如权利要求17所述的计算机可读存储介质,其中,基于所述第一类型轨道的所述列表中的每个轨道的最近被使用的时间来对所述第一类型轨道的所述列表中的轨道进行排序,并且其中,基于所述第二类型轨道的所述列表中的每个轨道的最近被使用的时间来对所述第二类型轨道的所述列表中的轨道进行排序。

19.如权利要求18所述的计算机可读存储介质,所述操作还包括:

响应于确定所述第二类型轨道的所述列表中的轨道的读取命中率小于所述第一类型

轨道中的轨道的读取命中率的预定因子,将轨道从所述第二类型轨道的所述列表中降级。

20.如权利要求19所述的计算机可读存储介质,其中,基于所述第一类型轨道的所述列表和所述第二类型轨道的所述列表的预定底部的读取命中来计算所述第一类型轨道的所述列表中的轨道的读取命中率和所述第二类型轨道的所述列表中的轨道的读取命中率。

21.如权利要求19所述的计算机可读存储介质,所述操作还包括:

调整所述预定因子以增加输入/输出(I/O)操作的速率。

22.一种用于高速缓存管理的存储控制器,包括:

高速缓存;以及

维护在所述存储控制器中的高速缓存管理应用程序,其中,所述高速缓存管理应用程序执行操作,所述操作包括:

在高速缓存中生成第一类型轨道的列表;

在所述高速缓存中生成第二类型的轨道的列表,其中,基于对轨道执行的最后I/O操作是第一类型I/O操作还是第二类型I/O操作,将所述轨道添加到所述第一类型轨道的所述列表或所述第二类型轨道的所述列表中;以及

确定是从所述第一类型轨道的所述列表还是从所述第二类型轨道的所述列表降级轨道,

其中:

为了对第一类型轨道执行所述第一类型I/O操作,将应用线程保持在旋转循环中,以等待所述第一类型I/O操作完成;以及

为了对第二类型轨道执行所述第二类型I/O操作,避免将所述应用线程保持在所述旋转循环中。

23.如权利要求22所述的存储控制器,其中,所述第一类型轨道是Syncio轨道。

24.如权利要求23所述的存储控制器,其中,所述第二类型轨道是非Syncio轨道。

25.如权利要求24所述的存储控制器,其中,基于所述第一类型轨道的所述列表中的每个轨道的最近被使用的时间来对所述第一类型轨道的所述列表中的轨道进行排序,并且其中,基于所述第二类型轨道的所述列表中的每个轨道的最近被使用的时间来对所述第二类型轨道的所述列表中的轨道进行排序。

高速缓存管理

技术领域

[0001] 实施例涉及基于输入/输出(I/O)操作的类型的高速缓存管理。

背景技术

[0002] 在某些存储系统环境中,存储控制器(或存储控制器复合体)可以包括彼此耦合的多个存储服务器。该存储控制器允许主机计算系统对由该存储控制器控制的存储设备执行输入/输出(I/O)操作,其中该主机计算系统可以被称为主机。

[0003] 存储控制器包括存储数据的高速缓存,从而可以更快地满足来自主机的对该数据的未来请求。与写入存储设备或从存储设备读取数据相比,将数据写入高速缓存或从高速缓存中读取数据要快得多。当可以在高速缓存中找到主机请求的数据时,发生高速缓存命中,而在高速缓存中找不到请求的数据时,发生高速缓存未命中。通过从高速缓存中读取数据来为高速缓存命中提供服务,这比从与存储控制器耦合的存储设备中读取数据要快。

[0004] 当高速缓存的空间或分段不足时,则高速缓存需要从高速缓存中逐出数据项,并且从高速缓存中逐出数据项称为降级。高速缓存替换策略可以用于确定高速缓存中的哪些数据项将被降级以为新的数据项腾出空间。最近最少使用(LRU)策略首先将最近最少使用的数据项降级。

发明内容

[0005] 提供一种用于高速缓存管理的方法、系统和计算机程序产品,其中生成高速缓存中的第一类型轨道的列表。生成高速缓存中第二类型轨道的列表,其中,相对于第二类型轨道,对第一类型轨道的I/O操作相对更快地完成。确定是从第一类型轨道的列表还是从第二类型轨道的列表降级轨道。作为确定的结果,降级的轨道增加了I/O操作的速率。

[0006] 在进一步的实施例中,第一类型轨道是Syncio轨道,其中为了对Syncio轨道执行I/O操作,将应用程序线程保持在旋转循环中以等待I/O操作完成。结果是,在某些实施例中,为了增加I/O操作的速率,可能不希望降级Syncio轨道。

[0007] 在又一些实施例中,第二类型轨道是非Syncio轨道,其中,为了对于非Syncio轨道执行I/O操作,避免将应用程序线程保持在旋转循环中,并且其中根据在轨道上执行的最后I/O操作是包括Syncio操作的第一类型I/O操作还是包括非Syncio操作的第二类型I/O操作,将轨道添加到第一类型轨道的列表或第二类型轨道的列表中。结果是,通过构建两种不同类型的列表来提高I/O操作的速率。

[0008] 在进一步的实施例中,基于最近如何使用第一类型轨道的列表中的每个轨道来对第一类型轨道的列表中的轨道进行排序,并且其中,基于最近如何使用第二类型轨道的列表中的每个轨道来对第二类型轨道的列表中的轨道进行排序。结果是,通过建立两种不同类型的最近最少使用的轨道列表,提高了I/O操作的速度。

[0009] 在某些实施例中,响应于确定所述第二类型轨道的所述列表中的轨道的读取命中率小于所述第一类型轨道的轨道中的读取命中率的预定因子,将轨道从所述第二类型轨道

的所述列表中降级。由于第二类型轨道的降级，I/O操作的速率增加了。

[0010] 在另外的实施例中，基于所述第一类型轨道的所述列表和所述第二类型轨道的所述列表的预定底部的读取命中来计算所述第一类型轨道的所述列表中的轨道的读取命中率和所述第二类型轨道的所述列表中的轨道的读取命中率。结果是，轨道的降级是基于对一组最近使用的轨道的收集统计数据。

[0011] 在其他实施例中，调整预定因子以增加输入/输出(I/O)操作的速率。结果是，无论如何，可能非Syncio轨道的LRU列表中的轨道降级，而不是Syncio轨道的LRU列表中的轨道降级，除非作为这种降级的结果是系统中I/O操作的速率降低。

附图说明

[0012] 现在参考附图，其中相同的附图标记始终表示相应的部分：

[0013] 图1示出了根据某些实施例的包括耦合到一个或多个主机和一个或多个存储设备的存储控制器的计算环境的框图，其中基于I/O操作的Syncio从主机到存储控制器发生；

[0014] 图2示出了根据某些实施例的框图，该框图示出了如何在存储控制器中维护最近最少使用的(LRU)轨道列表；

[0015] 图3示出了根据某些实施例的第一流程图，该第一流程图示出了从非Syncio LRU列表或从Syncio LRU列表进行的轨道降级；

[0016] 图4示出了根据某些实施例的第二流程图，该第二流程图示出了从非Syncio LRU列表或从Syncio LRU列表进行的轨道降级；

[0017] 图5示出了根据某些实施例的流程图，该流程图示出了如何计算读取命中率；

[0018] 图6示出了根据某些实施例的框图，该框图示出了如何计算轨道降级的预定因子；

[0019] 图7示出了根据某些实施例的基于生成两种不同类型的轨道的列表的轨道降级的流程图；

[0020] 图8示出了根据某些实施例的云计算环境的框图；

[0021] 图9示出了根据某些实施例的图8的云计算环境的更多细节的框图；以及

[0022] 图10示出了根据某些实施例的计算系统的框图，该计算系统示出了可以包括在存储控制器或主机中的某些元件，如图1-9所述。

具体实施方式

[0023] 在下面的描述中，将参考形成本文一部分并且示出了几个实施例的附图。应该理解，可以利用其他实施例，并且可以进行结构和操作上的改变。

[0024] Syncio(也称为同步I/O)包括用于计算设备的附件硬件和协议。Syncio被设计用于极低延迟的随机读取和小块顺序写入。计算设备之间的Syncio连接可能是点对点光学外围组件互连高速(PCIe)接口。Syncio操作与传统I/O的行为有所不同，因为主机计算设备可以在等待I/O操作完成的同时，将应用程序线程保持在旋转循环中。这避免了需要处理器周期来执行传统I/O的两次上下文交换，避免执行使I/O线程进入睡眠状态，然后重新分派I/O线程的操作以及避免对I/O中断。

[0025] 计算设备中的代码路径需要被极大地优化以满足Syncio操作的时间要求。延迟Syncio操作的任何条件(例如高速缓存未命中)都可能生成指示无法执行Syncio操作的状

态，并且可能必须重试该Syncio操作。由于仅当存在高速缓存命中时才可以成功地执行Syncio操作，因此提供了某些实施例以增加用于Syncio操作的高速缓存命中。在这样的实施例中，通过维护用于Syncio轨道和非Syncio轨道的分离的最近最少使用的(LRU)列表，并且基于分离的LRU列表使用降级机制，与仅单个LRU列表被维护的实施例相比，Syncio操作更快地完成。在某些实施例中，只要可能，非Syncio轨道的LRU列表中的轨道被降级，而不是Syncio轨道的LRU列表中的轨道被降级，除非系统中的I/O操作的速率由于这种降级而降低。

[0026] 示例性实施例

[0027] 图1示出了根据某些实施例的计算环境100的框图，该计算环境100包括耦合到一个或多个主机104、106以及一个或多个存储设备108、110的存储控制器102。存储控制器102允许多个主机104、106执行由存储控制器102维护的具有逻辑存储的输入/输出(I/O)操作。可以在一个或多个存储设备108、110和/或存储控制器102的高速缓存112(例如，存储器)中找到与逻辑存储相对应的物理存储。

[0028] 存储控制器102和主机104、106可以包括任何合适的计算设备，包括本领域中当前已知的那些，例如，个人计算机、工作站、服务器、大型机、手持计算机、掌上电脑顶级计算机、电话设备、网络设备、刀片计算机、处理设备等。存储控制器102、主机104、106和存储设备108、110可以是任何合适的网络114，例如，存储区域网络、广域网、互联网、内联网中的元件。在某些实施例中，存储控制器102、主机104、106和存储设备108、110可以是包括计算环境100的云计算环境中的元件。存储设备108、110可以由存储磁盘、磁带驱动器、固态存储器等组成，并且可以由存储控制器102控制。

[0029] 在某些实施例中，主机104、106可以经由总线接口(例如，点对点光学PCIe接口)116、118和网络接口120、122耦合到存储控制器102。来自主机104、106的Syncio操作可以通过总线接口116、118执行。来自主机104、106的传统I/O操作(即，非Syncio操作)可以通过网络接口120、122执行。总线接口116、118可以包括比网络接口120、122更快的I/O访问通道。可以使用扩展总线接口116、118的其他总线接口技术，包括PCIe扩展电缆或组件(例如分布式PCIe交换机)以允许通过以太网(例如使用ExpEther技术)进行PCIe。

[0030] 在某些实施例中，高速缓存112可以包括被划分为一个或多个等级的读/写高速缓存，其中每个等级可以包括一个或多个存储轨道。高速缓存112可以是本领域已知或将来开发的任何合适的高速缓存。在一些实施例中，可以利用易失性存储器和/或非易失性存储器来实现高速缓存112。高速缓存112可以存储修改的和未修改的数据。高速缓存112可以将数据存储在包括Syncio轨道124和非Syncio轨道126的多个轨道中。可以从主机104、106在存储控制器102上执行的Syncio操作128来读取和写入Syncio轨道 124，并且可以从主机104、106在存储控制器102上执行的非Syncio I/O操作130读取和写入非Syncio轨道126。高速缓存管理应用程序132生成在高速缓存112中未修改的Syncio轨道的非Syncio最近使用的(LRU)列表134和在高速缓存112中的未修改的非Syncio轨道的Syncio LRU列表136，其中未修改的Syncio轨道和未修改的非Syncio轨道是从高速缓存112降级的候选者以释放高速缓存112中的空间。在某些实施例中，高速缓存管理应用程序132可以以软件、硬件、固件或其任何组合来实现。

[0031] 图2示出了根据某些实施例的框图200，该框图示出了如何在存储控制器102中维

护轨道的LRU列表。

[0032] 非Syncio LRU列表134是可以对其执行非Syncio操作的高速缓存112中未修改的非Syncio轨道的列表，并且Syncio LRU列表136是可以对其执行Syncio操作的高速缓存112中的未修改的Syncio轨道的列表。

[0033] 非Syncio LRU列表134中的条目是高速缓存112中未经修改的非Syncio轨道，将从最近最多使用的202到最近最少使用的204从上到下放置。非Syncio LRU列表134的底部206的多个预定条目可以用于计算非Syncio LRU列表134上的读取命中率。通过仅选择底部206(例如，非Syncio LRU列表134中的1000个最近最少使用的非Syncio轨道，)来计算非Syncio LRU列表134上的读取命中率，确定读取命中率以用于从高速缓存112降级的潜在轨道。

[0034] Syncio LRU列表136中的条目是从最近最多使用的208到最近最少使用的210从上到下放置在高速缓存112中的未修改的Syncio轨道。在Syncio LRU列表136的底部212处的多个预定条目可以用于计算Syncio LRU列表136上的读取命中率。通过仅选择底部212(例如，Syncio LRU列表136中的1000个最近最少使用的Syncio轨道)来计算Syncio LRU列表136上的读取命中率，确定读取命中率以用于从高速缓存112降级的潜在轨道。

[0035] 图3示出了根据某些实施例的第一流程图300，该第一流程图300示出了从非Syncio LRU列表或从Syncio LRU列表中的轨道降级。图3中所示的操作可以由在存储控制器102中执行的高速缓存管理应用程序132执行。

[0036] 控制在框302开始，在框302中，高速缓存管理应用程序132生成并维护Syncio LRU列表136。高速缓存管理应用程序132确定(在框304处)Syncio LRU列表136中轨道的读取命中率，并且控制返回框302。

[0037] 与框302、304中所示的操作并行，高速缓存管理应用程序132生成并维护非Syncio LRU列表134(在框306处)。高速缓存管理应用程序132确定(在框308处)非Syncio LRU列表134中的轨道的读取命中率，并且控制返回到框306。

[0038] 在执行框302，304，306，308的操作时，高速缓存管理应用程序132继续或开始执行(在框310)用于从高速缓存112降级轨道的操作。高速缓存管理应用程序132确定(在框312)非Syncio LRU列表134中的轨道的读取命中率是否比Syncio LRU列表136中的轨道的读取命中率小于预定因子(例如，倍数)。例如，如果非Syncio LRU列表134中的轨道的读取命中率是0.2，Syncio LRU列表136中的轨道的命中率是0.1，预定因子是3，那么Syncio LRU列表134中的轨道的命中率乘以预定因子为0.3。非Syncio LRU列表134中的轨道的读取命中率是0.2，其小于0.3。因此，在该示例中，非Syncio LRU列表134中的轨道的读取命中率比Syncio LRU列表136中的轨道的读取命中率小于预定因子。

[0039] 如果在框312中，确定非Syncio LRU列表134中的轨道的读取命中率比Syncio LRU列表136中的轨道的读取命中率小于预定因子(“是，分支314”)，则高速缓存管理应用程序132从非Syncio LRU列表134降级(在框316)轨道，并且Syncio操作的性能可以在Syncio轨道被降级的情况下得到增强。非Syncio LRU列表134中的轨道被使用得比Syncio LRU列表136中的轨道少得多，因此期望从非Syncio LRU列表134中降级。

[0040] 如果在框312确定非Syncio LRU列表134中的轨道的读取命中率比Syncio LRU列表136中的轨道的读取命中率不小于预定因子(“否”，分支318)，则高速缓存管理应用程序132从Syncio LRU列表136降级(在框320)。非Syncio LRU列表134中的轨道比Syncio LRU列

表136中的轨道被充分利用,因此,期望从Syncio LRU列表136降级。在Syncio LRU列表136中的轨道上读命中可能相对较少,因此希望从Syncio LRU列表136降级。

[0041] 因此,图3描述了其中通过维护用于Syncio和非Syncio轨道的不同LRU列表的实施例,与仅维护单个LRU列表的实施例相比,可以更快地执行Syncio操作。

[0042] 图4示出了根据某些实施例的第二流程图,该第二流程图示出了从非Syncio LRU列表134或从Syncio LRU列表136进行轨道降级。图4中所示的操作可以由在存储控制器102中执行的高速缓存管理应用程序132执行。

[0043] 控制从框402开始,在框402中,高速缓存管理应用程序132确定Syncio LRU列表136是否为空。如果是(“是”,分支404),则控制前进至框406,其中高速缓存管理应用程序132从非Syncio LRU列表134降级轨道。否则(“否”,分支408),控制进行至框410,其中高速缓存管理应用程序132确定非Syncio LRU列表134是否为空。

[0044] 如果在框410中,高速缓存管理应用程序132确定非Syncio LRU列表134为空(“是”,分支412),则高速缓存管理应用程序132从Syncio LRU列表136降级轨道。否则(““否”,分支416)控制前进到框418,其中高速缓存管理应用程序132确定非Syncio LRU列表134的最老轨道是否比Syncio LRU列表136的最老轨道早,其中最老轨道是列表中所有其他轨道的最后一次使用之前最后一次使用的轨道(即,该轨道上发生读取命中),并且是在最后一次使用第二次之前最后一次使用的,早于第二轨道的第一条轨道。

[0045] 如果在框418,高速缓存管理应用程序132确定非Syncio LRU列表134的最老轨道早于Syncio LRU列表136的最老轨道(“是”分支420),则控制进行到框422,其中,高速缓存管理应用程序132从非Syncio LRU列表134降级轨道。否则(“否”,分支424),高速缓存管理应用程序132计算(在框426上)以下内容的读取命中率:(a)Syncio LRU列表212的底部;(b)非Syncio LRU列表206的底部部分。

[0046] 将Syncio LRU列表212的底部的读取命中率称为SIO_BOTTOMHITS,将非Syncio LRU列表206的底部的读取命中率称为NONSIO_BOTTOMHITS。

[0047] 控制从框426前进到框428,其中高速缓存管理应用程序132确定NONSIO_BOTTOMHITS是否比SIO_BOTTOMHITS乘以预定因子小。如果是这样(“是”,分支430),则高速缓存管理应用程序132从非Syncio LRU列表134降级轨道。否则(“否”,分支434),高速缓存管理应用程序132从Syncio LRU列表136降级轨道(在框436)。

[0048] 因此,图4示出了某些实施例,其中,非Syncio LRU列表134和Syncio LRU列表136的至少底部206、212用于读取命中的比较,从而对非Syncio LRU列表134和Syncio LRU列表136中的相对较老的轨道上的读取命中进行比较。轨道的降级来自这些相对较老的轨道,因此针对这些相对较老的轨道计算读取命中率。

[0049] 图5示出了流程图500,其示出了根据某些实施例的如何计算读取命中率。图5中所示的操作可以由在存储控制器102中执行的高速缓存管理应用程序132执行。

[0050] 控制在框501开始,其中高速缓存管理应用程序132将Syncio LRU列表212的底部的读取命中计数器和非Syncio LRU列表206的底部的读取命中计数器设置为零。控制进行到框502,其中读取操作在高速缓存112上生成读取命中,并且控制并行进行到框504和506。在框504,高速缓存管理应用程序132确定是否在Syncio LRU列表212的底部读取命中,如果是(“是”,分支508),则递增Syncio列表212的底部的读取命中计数器(在框510处)。否则

(“否”,分支512),控制返回到框502。

[0051] 在框506处,高速缓存管理应用程序132确定是否在非Syncio LRU列表206的底部上读取命中,并且如果是(“是”,分支514),则递增在非Syncio LRU列表206的底部上的读取命中计数器(在框516)。否则(“否”分支518),控制返回到框502。重复执行框501、502、504、506、508、510、512、514、516、518的操作一段时间(例如800毫秒)(如参考数字520所示)。

[0052] 在执行经由附图标记520所示的操作的同时,执行框522和524所示的操作。在框522,高速缓存管理应用程序132通过将最后N个时间间隔的读取命中相加并除以那些时间间隔的累积时间,来计算Syncio列表212的底部的读取命中率。在框524,高速缓存管理应用程序132通过将最后N个时间间隔的读取命中相加并除以那些时间间隔的累积时间来计算非Syncio列表206的底部的读取命中率。在某些实施例中,N可以是诸如30的数字。

[0053] 因此,图5示出了用于计算非Syncio LRU列表134和Syncio LRU列表136底部的读取命中率的某些实施例,以用于图4所示的操作。

[0054] 图6示出了框图600,该框图根据某些实施例的示出了如何计算轨道降级的预定因子。在图3的框312和图4的框428中已经示出了示例性预定因子。

[0055] 在某些实施例中,预定因子可以基于系统的性能基准(如通过附图标记602所示)。在其他实施例中,预定因子可以由希望Syncio工作负载优于非Syncio工作负载的性能的客户来配置(如通过附图标记604所示)。在又一些实施例中,可以基于在存储控制器102中发生的I/O操作的速率来动态地改变预定因子(如通过附图标记606所示)。例如,如果I/O操作的速率随着预定因子的增加而增加,则高速缓存管理应用程序132继续增加预定因子,并且如果I/O操作的速率随着预定因子的增加而降低,则高速缓存管理应用程序132继续减小预定因子。

[0056] 图7示出了根据某些实施例的流程图700,该流程图700显示了基于生成两种不同类型的轨道的列表来进行轨道降级。图7中所示的操作可以由在存储控制器102中执行的高速缓存管理应用程序132执行。

[0057] 控制从框702开始,在框702中,生成高速缓存112中的第一类型轨道的列表(例如Syncio LRU列表136)。控制进行到框704,在框704中,生成高速缓存112中的第二类型轨道的列表(非Syncio LRU列表134),其中第一类型轨道相对于第二类型轨道而言,I/O操作相对更快地完成,并且其中根据在轨道上执行的最后I/O操作是第一类型I/O操作(例如,Syncio操作)还是第二类型I/O操作(例如,非Syncio操作),将轨道添加到第一类型轨道的列表或第二类型轨道的列表。例如,如果在轨道上执行的最后I/O操作是Syncio操作,则将该轨道添加到Syncio LRU列表136,并且如果在轨道上执行的最后I/O操作是非Syncio操作,则轨道被添加到非Syncio LRU列表134。做出关于是否从第一类型轨道的列表或从第二类型轨道的列表降级的轨道的确定(在框706)。第一类型轨道可以是Syncio轨道,其中为了对Syncio轨道执行I/O操作,将应用程序线程保持在旋转循环中,等待I/O操作完成。第二类型轨道可以是非Syncio轨道,其中对于非Syncio轨道执行I/O操作,可以避免将应用程序线程保持在旋转循环中。

[0058] 在某些实施例中,基于最近如何使用第一类型轨道的列表中的每个轨道(例如,如参考标号208、210所示)来对第一类型轨道的列表中的轨道进行排序,其中基于最近如何使用第二类型轨道的列表中的每个轨道来对第二类型轨道的列表中的轨道(例如,如通过附

图标记202、204所示)进行排序,其中如果轨道上发生读取命中则轨道被称为被使用。

[0059] 在某些实施例中,响应于确定第二类型轨道的列表中的轨道的读取命中率小于第一类型轨道中的轨道的读取命中率的预定因子,从第二类型轨道的列表中将轨道降级(在框708)。基于在第一类型和第二类型轨道的列表的预定底部上的读取命中和来计算第一类型轨道的列表中的轨道的读取命中率和第二类型轨道的列表中的轨道的读取命中率。调整预定因子以增加输入/输出(I/O)操作的速率。

[0060] 因此,图1-7示出了某些实施例,其中通过维护用于Syncio轨道和非Syncio轨道的分离的LRU列表并使用基于分离的LRU列表的降级机制,与仅维护单个LRU列表的实施例相比,Syncio操作可以更快地完成。在某些实施例中,只要有可能,就将非Syncio轨道的LRU列表(非Syncio LRU列表134)中的轨道降级,而不是将Syncio轨道的LRU列表(Syncio LRU列表136)中的轨道降级,除非由于这种降级,系统中的I/O操作减少。

[0061] 云计算环境

[0062] 云计算是一种模型,用于实现对可配置计算资源(例如,网络、服务器、存储器、应用程序和服务)的共享池的方便、按需网络访问,这些资源可以用最少的管理工作或者提供商的互动快速配置和发布。

[0063] 现在参考图8,其中显示了示例性的云计算环境50。如图所示,云计算环境50包括云计算消费者使用的本地计算设备可以与其相通信的一个或者多个云计算节点10,本地计算设备例如可以是个人数字助理(PDA)或移动电话54A,台式电脑54B、笔记本电脑54C和/或汽车计算机系统54N。云计算节点10之间可以相互通信。可以在包括但不限于如上所述的私有云、共同体云、公共云或混合云或者它们的组合的一个或者多个网络中将云计算节点10进行物理或虚拟分组(图中未显示)。这样,云的消费者无需在本地计算设备上维护资源就能请求云计算环境50提供的基础架构即服务(IaaS)、平台即服务(PaaS)和/或软件即服务(SaaS)。应当理解,图8显示的各类计算设备54A-N仅仅是示意性的,云计算节点10以及云计算环境50可以与任意类型网络上和/或网络可寻址连接的任意类型的计算设备(例如使用网络浏览器)通信。

[0064] 现在参考图9,其中显示了云计算环境50(图8)提供的一组功能抽象层。首先应当理解,图9所示的组件、层以及功能都仅仅是示意性的,本发明的实施例不限于此。

[0065] 硬件和软件层60包括硬件和软件组件。硬件组件的示例包括在一个示例IBM zSeries系统中的大型机;在一个示例IBM pSeries系统中的基于RISC(精简指令集计算机)体系结构的服务器;IBM xSeries系统;IBM BladeCenter系统;存储设备;网络和网络组件。软件组件的示例包括网络应用服务器软件,在一个示例中为IBM WebSphere应用服务器软件。和在一个示例中是IBM DB2数据库软件的数据库软件。(IBM、zSeries、pSeries、xSeries、BladeCenter、WebSphere和DB2是国际商业机器公司(International Business Machines Corporation)在全球许多司法管辖区注册的商标)

[0066] 虚拟层62提供一个抽象层,该层可以提供下列虚拟实体的例子:虚拟服务器、虚拟存储、虚拟网络(包括虚拟私有网络)、虚拟应用和操作系统,以及虚拟客户端。

[0067] 在一个示例中,管理层64可以提供下述功能:资源供应功能:提供用于在云计算环境中执行任务的计算资源和其它资源的动态获取;计量和定价功能:在云计算环境内对资源的使用进行成本跟踪,并为此提供帐单和发票。在一个例子中,该资源可以包括应用软件

许可。安全功能:为云的消费者和任务提供身份认证,为数据和其它资源提供保护。用户门户功能:为消费者和系统管理员提供对云计算环境的访问。服务水平管理功能84:提供云计算资源的分配和管理,以满足必需的服务水平。服务水平协议(SLA)计划和履行功能:为根据SLA预测的对云计算资源未来需求提供预先安排和供应。

[0068] 工作负载层66提供云计算环境可能实现的功能的示例。在该层中,可提供的工作负载或功能的示例包括:地图绘制与导航;软件开发及生命周期管理;虚拟教室的教学提供;数据分析处理;交易处理;以及;和降级操作的处理68,如图1-8所示。

[0069] 附加实施例细节

[0070] 所描述的操作可以被实现为使用标准编程和/或工程技术以产生软件、固件、硬件或其任何组合的方法、装置或计算机程序产品。因此,实施例的各个方面可以采取完全硬件实施例,完全软件实施例(包括固件、常驻软件、微代码等)或结合了软件和在本文中通常称为“电路”“模块”或者“系统:的硬件方面的实施例的形式。此外,实施例的各方面可以采取计算机程序产品的形式。计算机程序产品可以包括计算机可读存储介质,其上载有用于使处理器实现本发明的各个方面的计算机可读程序指令。

[0071] 计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是——但不限于——电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括:便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式压缩盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身,诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波(例如,通过光纤电缆的光脉冲)、或者通过电线传输的电信号。

[0072] 这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备,或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令,并转发该计算机可读程序指令,以供存储在各个计算/处理设备中的计算机可读存储介质中。

[0073] 用于执行本发明操作的计算机程序指令可以是汇编指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、集成电路配置数据或者以一种或多种编程语言的任意组合编写的源代码或目标代码,所述编程语言包括面向对象的编程语言—诸如Smalltalk、C++等,以及过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。在一些实施例中,通过利用计算机可读

程序指令的状态信息来个性化定制电子电路，例如可编程逻辑电路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA)，该电子电路可以执行计算机可读程序指令，从而实现本发明的各个方面。

[0074] 这里参照根据本发明实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述了本发明的各个方面。应当理解，流程图和/或框图的每个框以及流程图和/或框图中各框的组合，都可以由计算机可读程序指令实现。

[0075] 这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器，从而生产出一种机器，使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时，产生了实现流程图和/或框图中的一个或多个框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中，这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作，从而，存储有指令的计算机可读介质则包括一个制造品，其包括实现流程图和/或框图中的一个或多个框中规定的功能/动作的各个方面的指令。

[0076] 也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上，使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤，以产生计算机实现的过程，从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个框中规定的功能/动作。

[0077] 附图中的流程图和框图显示了根据本发明的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个框可以代表一个模块、程序段或指令的一部分，所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中，框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个框、以及框图和/或流程图中的框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0078] 图10示出了根据某些实施例的可以包括在存储控制器102或主机104、106或其他计算设备中的某些元件的框图。系统1000可以包括电路1002，其在某些实施例中可以至少包括处理器1004。系统1000还可以包括存储器1006(例如，易失性存储设备)和存储器1008。存储器1008可以包括非存储器1008。易失性存储设备(例如，EEPROM、ROM、PROM、闪存、固件、可编程逻辑等)，磁盘驱动器、光盘驱动器、磁带驱动器等。存储器1008可以包括内部存储设备，连接的存储设备和/或网络可访问存储设备。系统1000可以包括包含代码1012的程序逻辑1010，该代码1012可以被加载到存储器1006中并由处理器1004或电路1002执行。在某些实施例中，包括代码1012的程序逻辑1010可以被存储在存储器1008中。在某些其他实施例中，程序逻辑1010可以在电路1002中实现。系统1000中的一个或多个组件可以经由总线或经由其他耦合或连接1014进行通信。因此，尽管图10从其他元件单独示出了程序逻辑1010，程序逻辑1010可以在存储器1006和/或电路1002中实现。

[0079] 某些实施例可以针对一种用于由人部署计算指令或将计算机可读代码集成到计算系统中的自动处理的方法，其中使代码与计算系统结合能够执行所描述的实施例的操作。

[0080] 术语“一实施例”、“实施例”，“多个实施例”、“该实施例”、“该多个实施例”、“一个或多个实施例”，“一些实施例”和“一个实施例”是指本发明的“一个或多个(但不是全部)实施例”，除非另有明确说明。

[0081] 除非另外明确指出，术语“包括”，“包含”，“具有”及其变体表示“包括但不限于”，除非另有明确说明。

[0082] 所列举的项目清单并不意味着任何或所有项目是互斥的，除非另有明确说明。

[0083] 除非另外明确指出，术语“一个”，“一种”和“该”表示“一个或多个”。

[0084] 除非另外明确指出，否则彼此通信的设备不必彼此持续通信。另外，彼此通信的设备可以通过一个或多个中间设备直接或间接通信。

[0085] 具有多个彼此通信的组件的实施例的描述并不暗示需要所有这样的组件。相反，描述了各种可选的组件以说明本发明的各种可能的实施例。

[0086] 此外，尽管可以按顺序描述处理步骤、方法步骤、算法等，但是可以将这样的处理、方法和算法配置为以替代顺序工作。换句话说，可以描述的步骤的任何序列或顺序不一定表示要求以该顺序执行步骤。本文描述的过程的步骤可以以任何实际顺序执行。此外，可以同时执行一些步骤。

[0087] 当在本文描述单个设备或物品时，将显而易见的是，可以使用一个以上的设备/物品(无论它们是否协作)来代替单个设备/物品。类似地，在本文描述了一个以上的设备或物品的情况下(无论它们是否协作)，很明显可以使用单个设备/物品代替一个以上的设备或物品或者可以使用不同数量的设备/物品而不是设备或程序所示的数量。设备的功能和/或特征可以可替换地由一个或多个未明确描述为具有这种功能/特征的其他设备来体现。因此，本发明的其他实施例不需要包括设备本身。

[0088] 已经在图中示出的至少某些操作示出了以特定顺序发生的特定事件。在替代实施例中，某些操作可以以不同的顺序执行、修改或移除。而且，可以将步骤添加到上述逻辑并且仍然符合所描述的实施例。此外，本文描述的操作可以顺序地发生，或者可以并行地处理某些操作。另外，操作可以由单个处理单元或由分布式处理单元执行。

[0089] 为了说明和描述的目的，已经给出了本发明的各种实施例的前述描述。其并非旨在穷举或将本发明限制为所公开的精确形式。根据以上教导，许多修改和变化是可能的。意图在于本发明的范围不由该详细描述限制，而是由所附权利要求书限制。以上说明、示例和数据提供了对本发明的组成的制造和使用的完整描述。因为可以在不脱离本发明范围的情况下做出本发明的许多实施例。

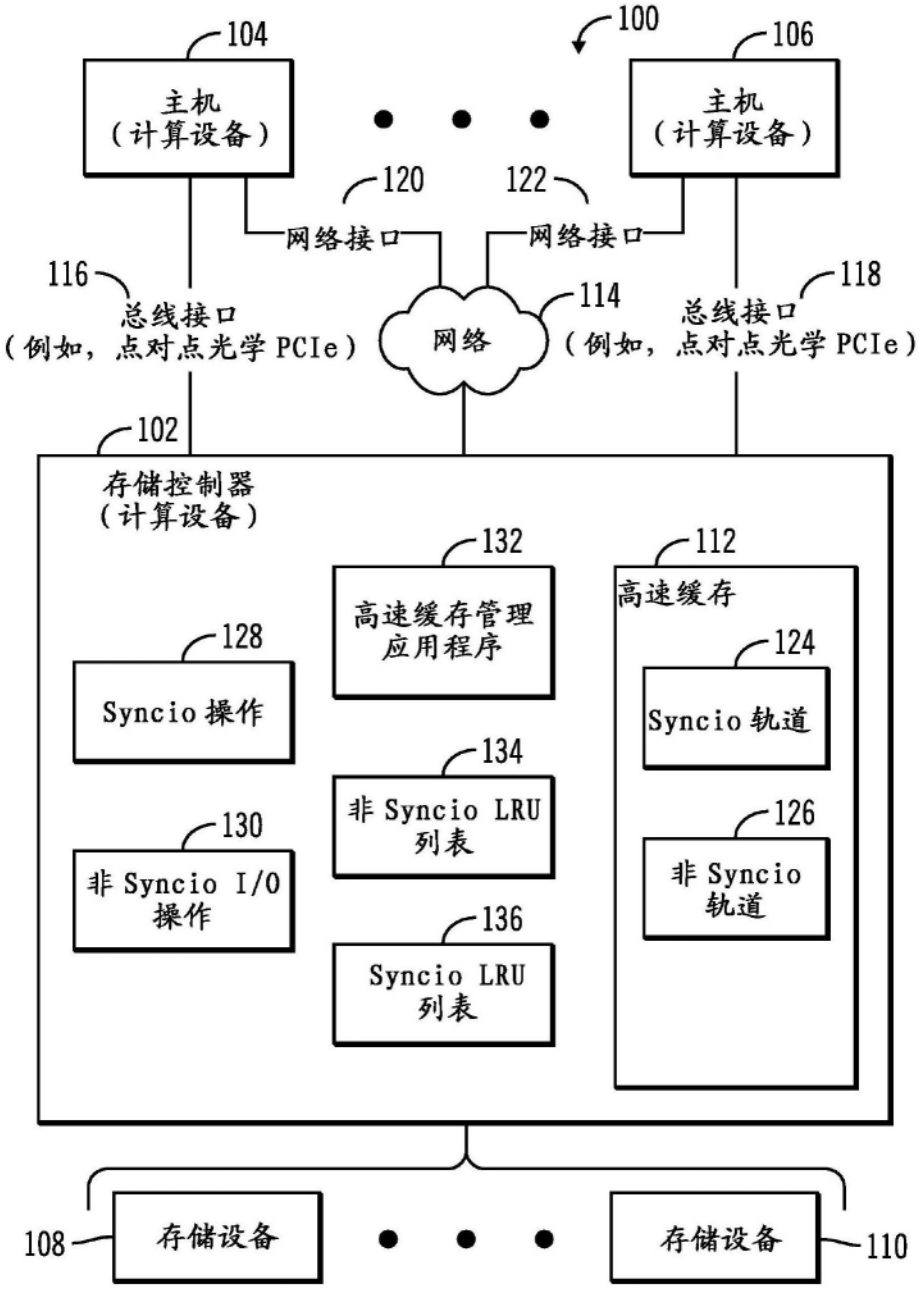

100
104
主机
（计算设备）
106
主机
（计算设备）
120
122
网络接口
网络接口
116
总线接口
（例如，点对点光学 PCIe）
网络
114
118
总线接口
（例如，点对点光学 PCIe）
102
存储控制器
（计算设备）
132
高速缓存管理
应用程序
112
高速缓存
128
Syncio 操作
124
Syncio 轨道
134
非 Syncio LRU
列表
130
非 Syncio I/O
操作
126
非 Syncio
轨道
136
Syncio LRU
列表
108
存储设备
存储设备
110

图1

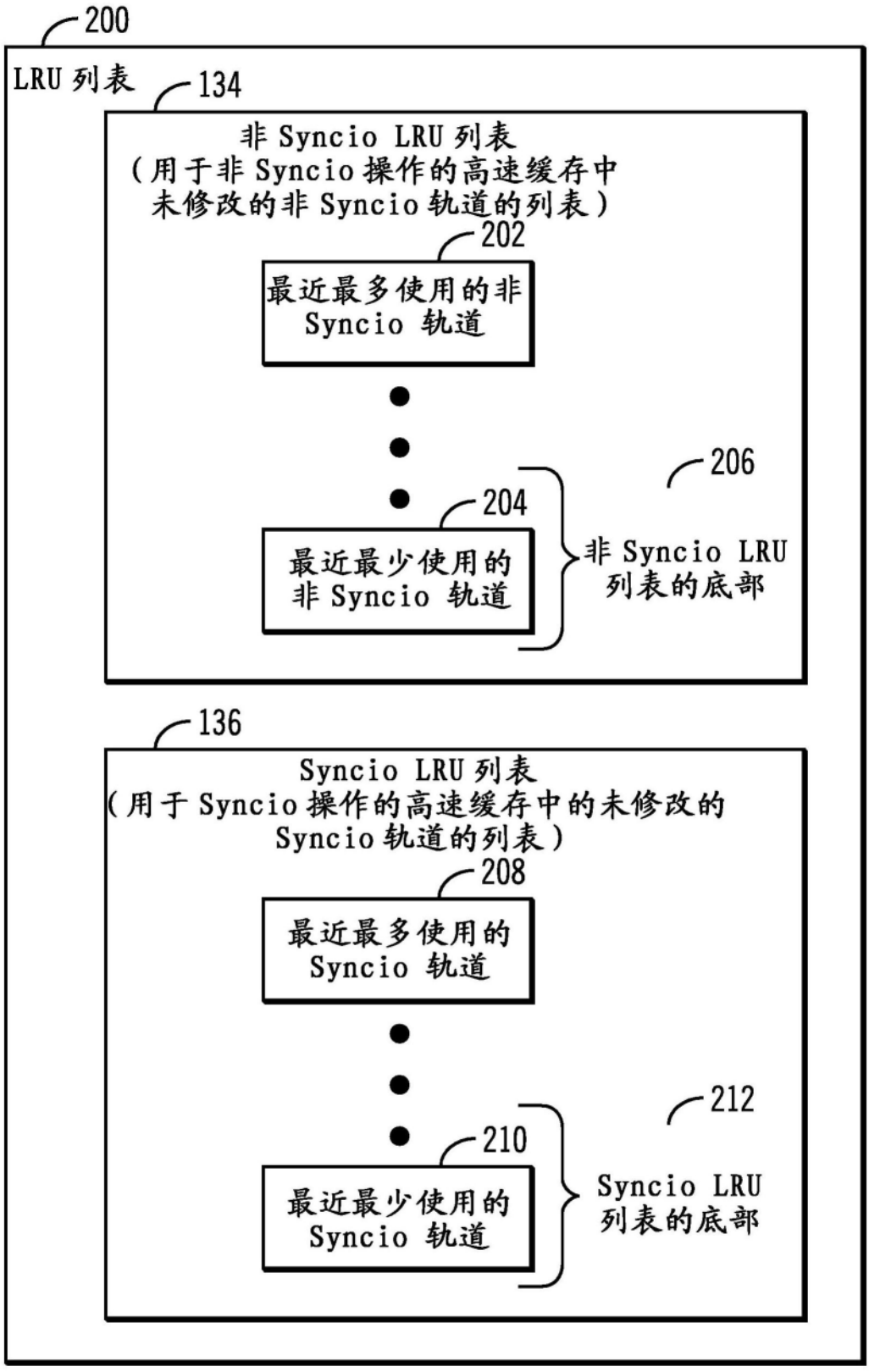

200
LRU 列表
134
非 Syncio LRU 列表
（用于非 Syncio 操作的高速缓存中
未修改的非 Syncio 轨道的列表）
202
最近最多使用的非
Syncio 轨道
206
204
最近最少使用的
非 Syncio 轨道
非 Syncio LRU
列表的底部
136
Syncio LRU 列表
（用于 Syncio 操作的高速缓存中的未修改的
Syncio 轨道的列表）
208
最近最多使用的
Syncio 轨道
212
210
最近最少使用的
Syncio 轨道
Syncio LRU
列表的底部

图2

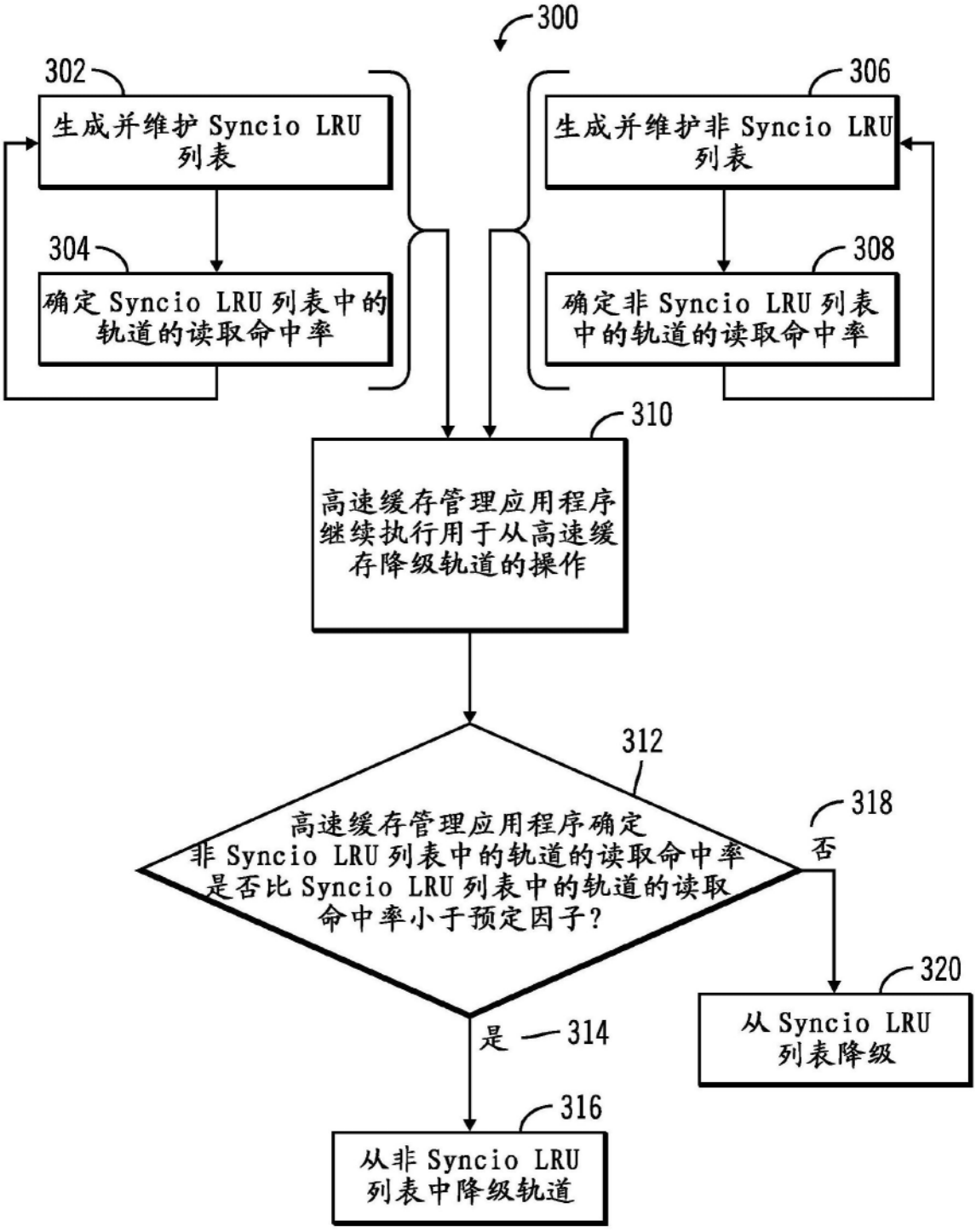

300
302
生成并维护 Syncio LRU 列表
304
确定 Syncio LRU 列表中的轨道的读取命中率
306
生成并维护非 Syncio LRU 列表
308
确定非 Syncio LRU 列表中的轨道的读取命中率
310
高速缓存管理应用程序继续执行用于从高速缓存降级轨道的操作
312
高速缓存管理应用程序确定非 Syncio LRU 列表中的轨道的读取命中率是否比 Syncio LRU 列表中的轨道的读取命中率小于预定因子?
318
否
320
从 Syncio LRU 列表降级
是
314
316
从非 Syncio LRU 列表中降级轨道

图3

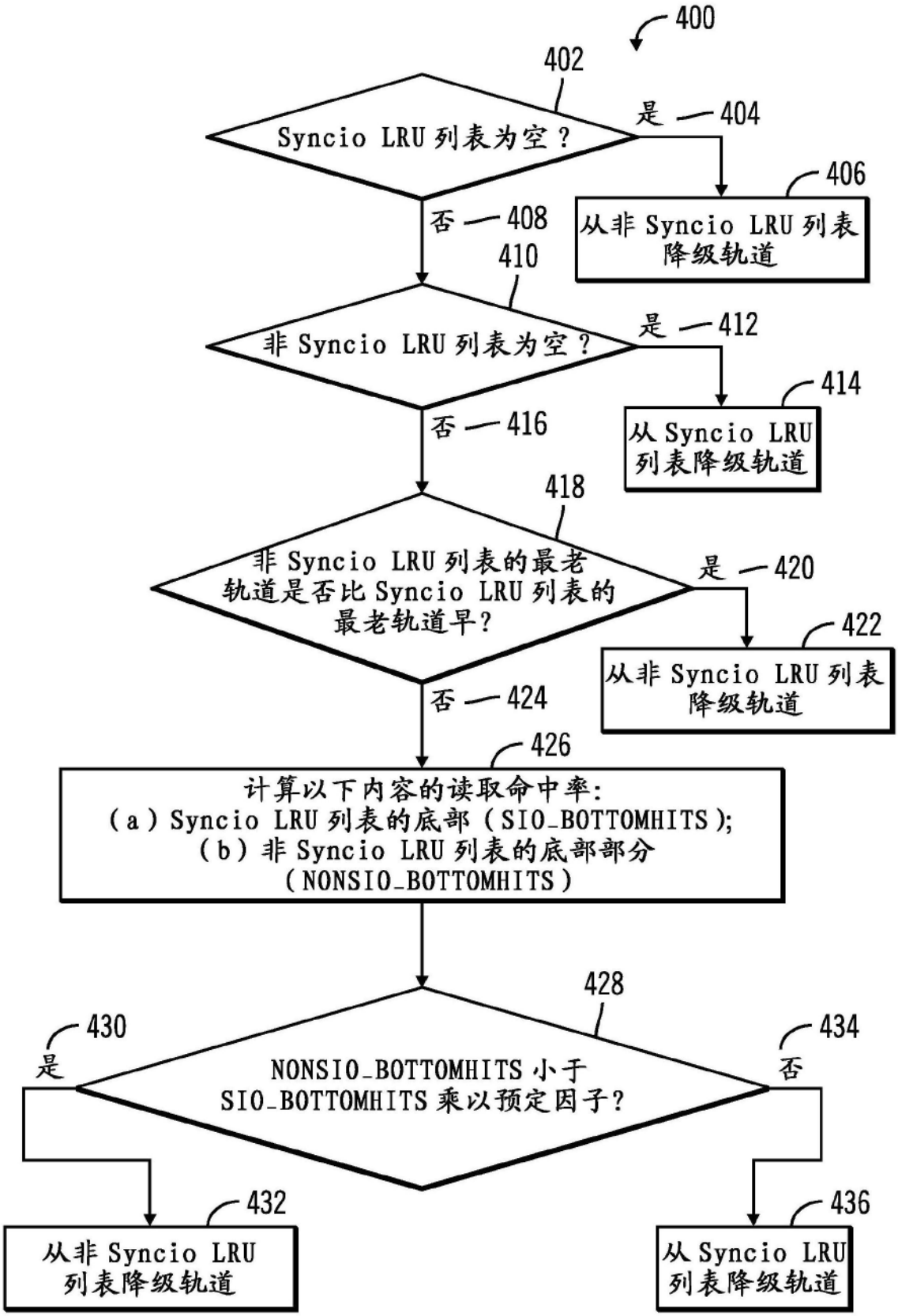

400
402
Syncio LRU 列表为空？
是
404
406
从非 Syncio LRU 列表降级轨道
否
408
410
非 Syncio LRU 列表为空？
是
412
414
从 Syncio LRU 列表降级轨道
否
416
418
非 Syncio LRU 列表的最老轨道是否比 Syncio LRU 列表的最老轨道早？
是
420
422
从非 Syncio LRU 列表降级轨道
否
424
426
计算以下内容的读取命中率：
（a）Syncio LRU 列表的底部（SIO_BOTTOMHITS）；
（b）非 Syncio LRU 列表的底部部分（NONSIO_BOTTOMHITS）
428
NONSIO_BOTTOMHITS 小于 SIO_BOTTOMHITS 乘以预定因子？
是
430
否
434
从非 Syncio LRU 列表降级轨道
432
从 Syncio LRU 列表降级轨道
436

图4

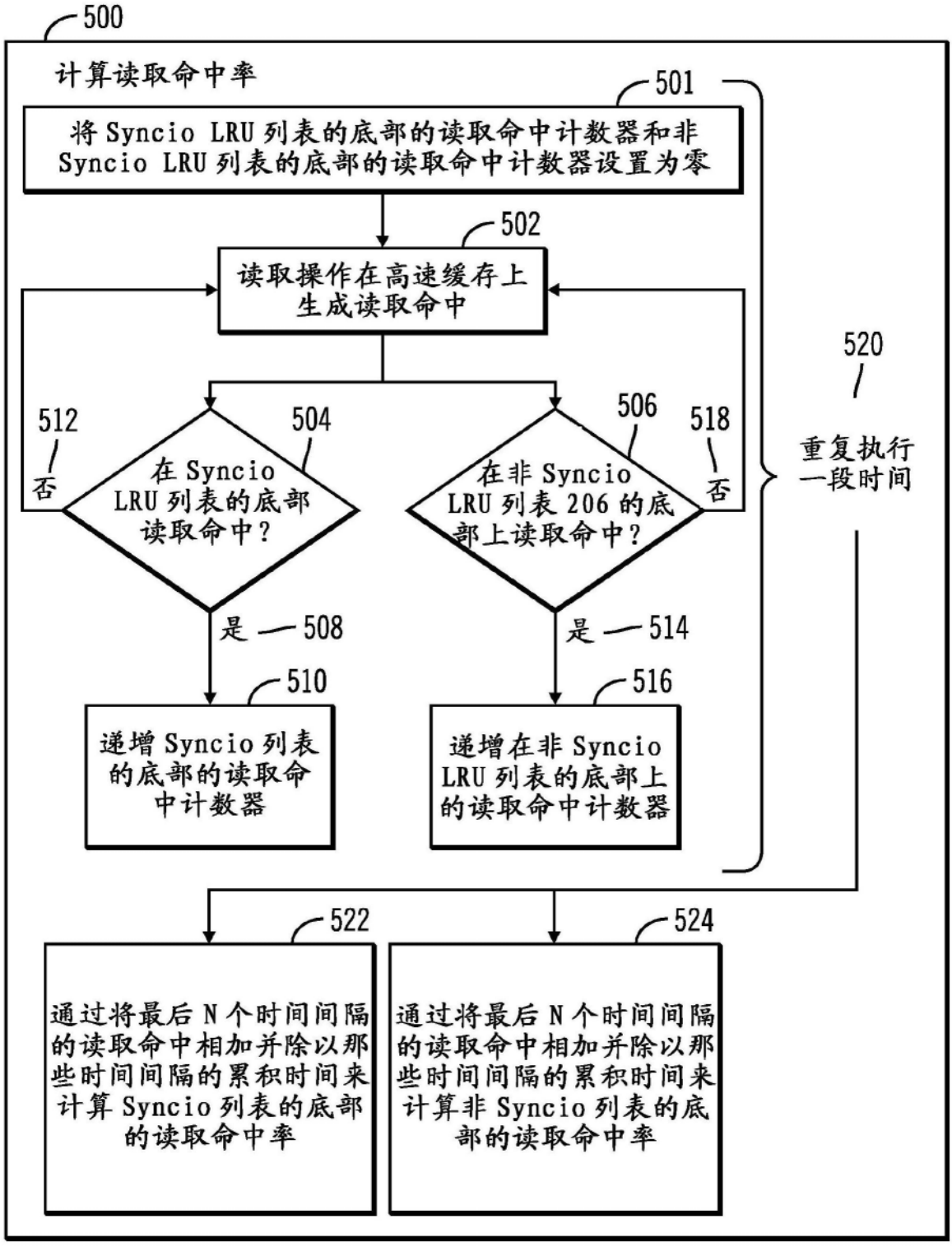

500
计算读取命中率
501
将 Syncio LRU 列表的底部的读取命中计数器和非 Syncio LRU 列表的底部的读取命中计数器设置为零
502
读取操作在高速缓存上生成读取命中
520
重复执行一段时间
512
否
504
在 Syncio LRU 列表的底部读取命中？
506
在非 Syncio LRU 列表 206 的底部上读取命中？
518
否
是
508
是
514
510
递增 Syncio 列表的底部的读取命中计数器
516
递增在非 Syncio LRU 列表的底部上的读取命中计数器
522
通过将最后 N 个时间间隔的读取命中相加并除以那些时间间隔的累积时间来计算 Syncio 列表的底部的读取命中率
524
通过将最后 N 个时间间隔的读取命中相加并除以那些时间间隔的累积时间来计算非 Syncio 列表的底部的读取命中率

图5

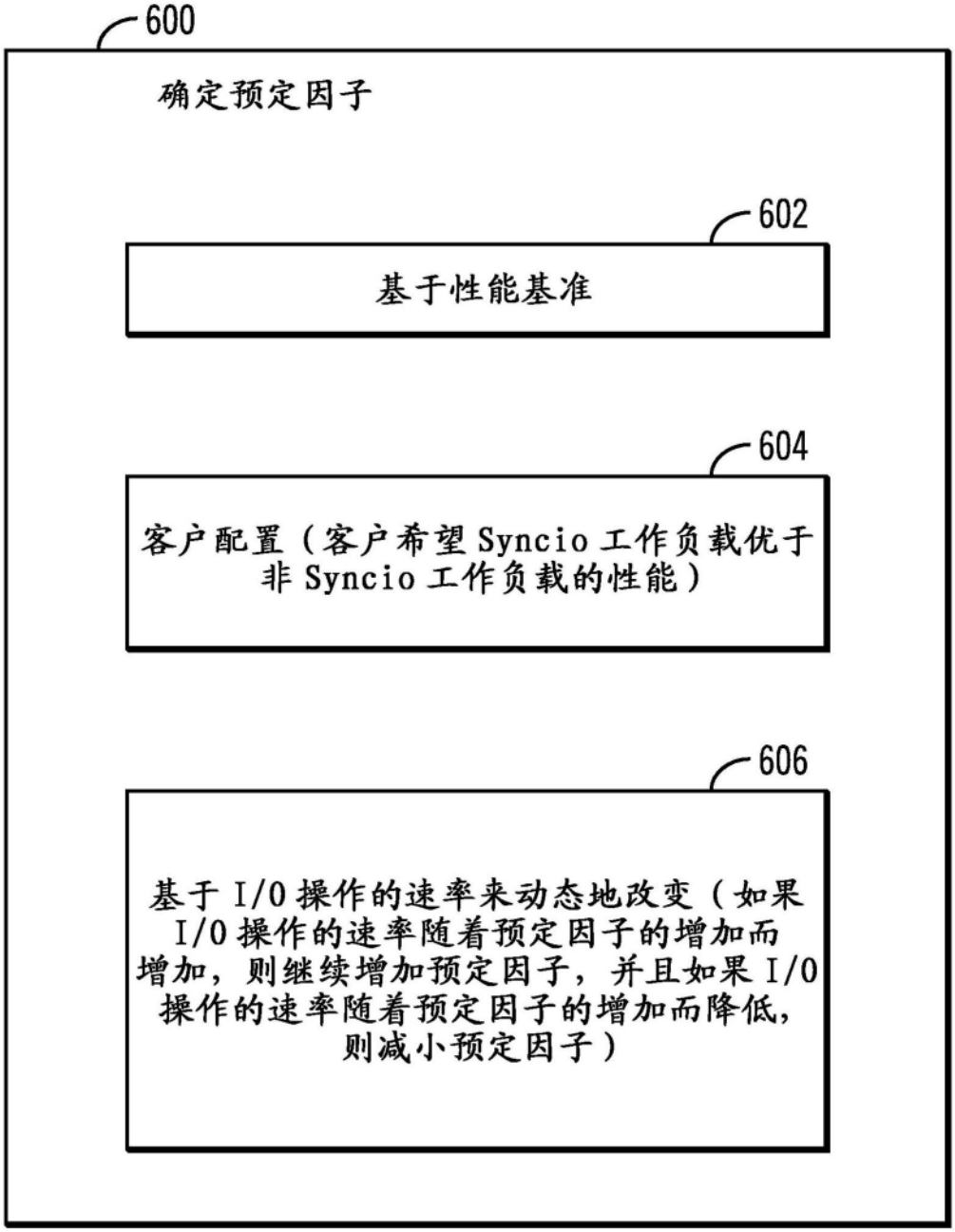

600
确定预定因子
602
基于性能基准
604
客户配置（客户希望Syncio工作负载优于非Syncio工作负载的性能）
606
基于I/O操作的速率来动态地改变（如果I/O操作的速率随着预定因子的增加而增加，则继续增加预定因子，并且如果I/O操作的速率随着预定因子的增加而降低，则减小预定因子）

图6

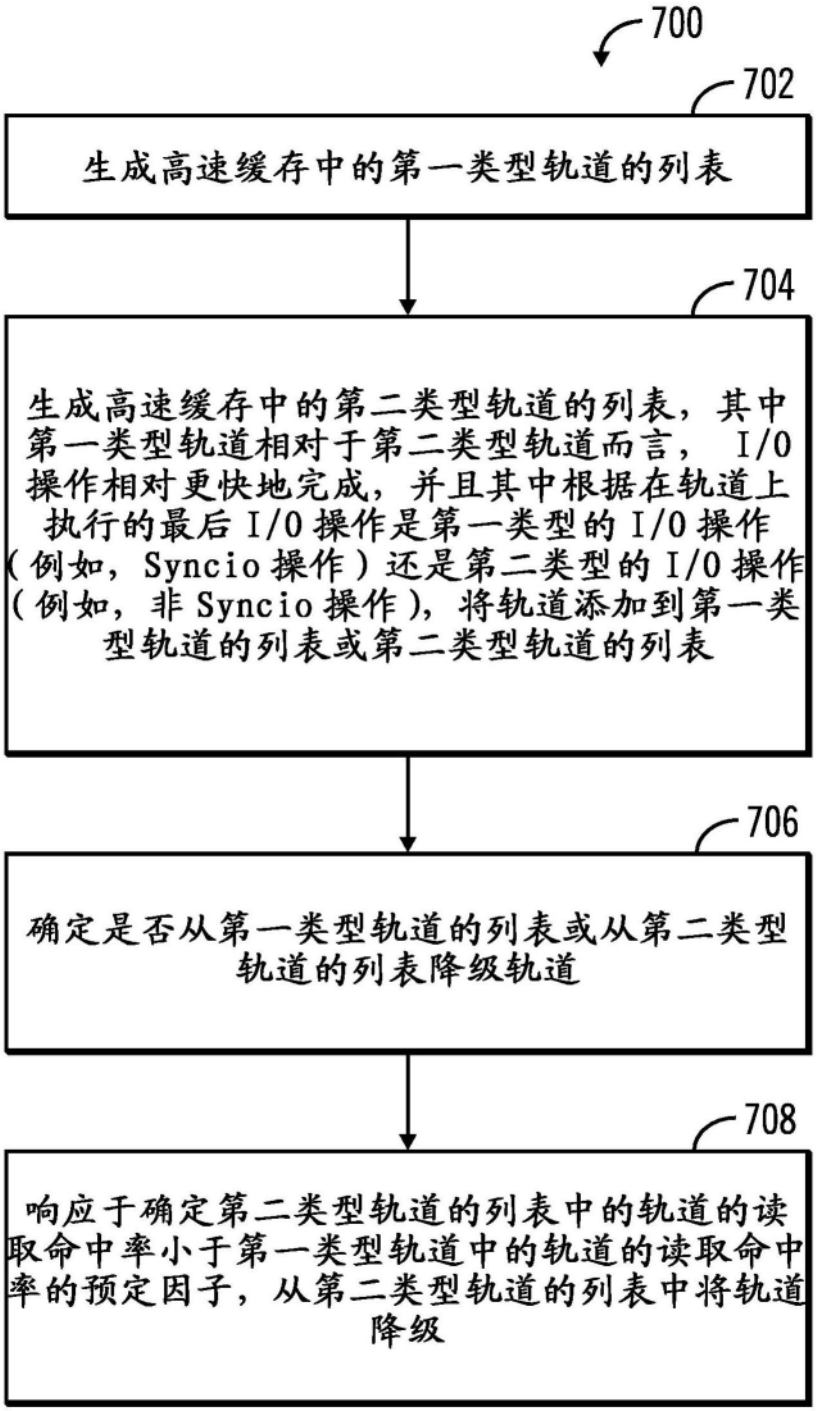

700
702
生成高速缓存中的第一类型轨道的列表
704
生成高速缓存中的第二类型轨道的列表，其中第一类型轨道相对于第二类型轨道而言， I/O操作相对更快地完成，并且其中根据在轨道上执行的最后 I/O 操作是第一类型的 I/O 操作（例如，Syncio 操作）还是第二类型的 I/O 操作（例如，非 Syncio 操作），将轨道添加到第一类型轨道的列表或第二类型轨道的列表
706
确定是否从第一类型轨道的列表或从第二类型轨道的列表降级轨道
708
响应于确定第二类型轨道的列表中的轨道的读取命中率小于第一类型轨道中的轨道的读取命中率的预定因子，从第二类型轨道的列表中将轨道降级

图7

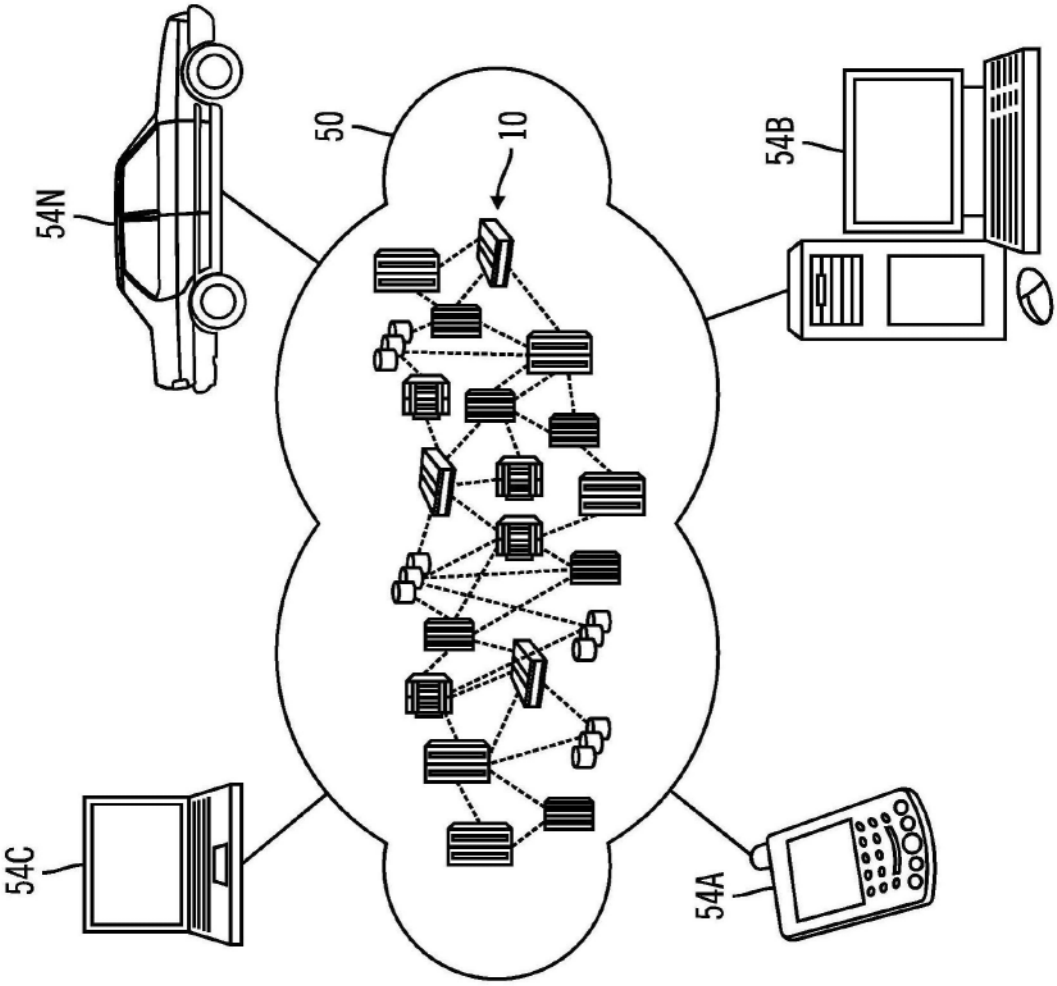


图8

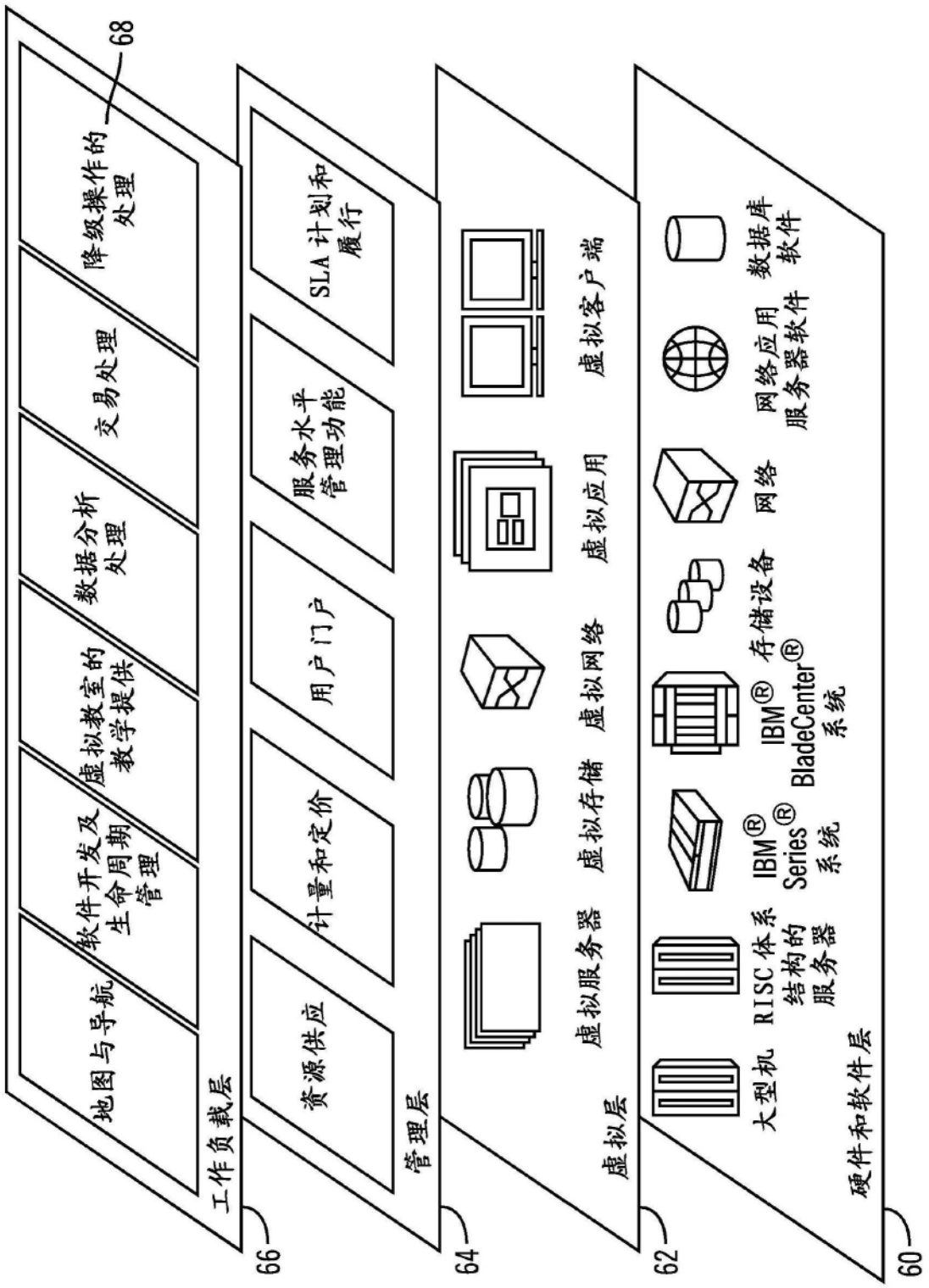

68
地图与导航
软件开发及生命周期管理
虚拟教室的教学提供
数据分析处理
交易处理
降级操作的处理
工作负载层
66
资源供应
计量和定价
用户门户
服务水平管理功能
SLA 计划和履行
管理层
64
虚拟服务器
虚拟存储
虚拟网络
虚拟应用
虚拟客户端
虚拟层
62
大型机
RISC 体系结构的服务器
IBM® Series® 系统
IBM® BladeCenter® 系统
存储设备
网络
网络应用服务器软件
数据库软件
硬件和软件层
60

图9

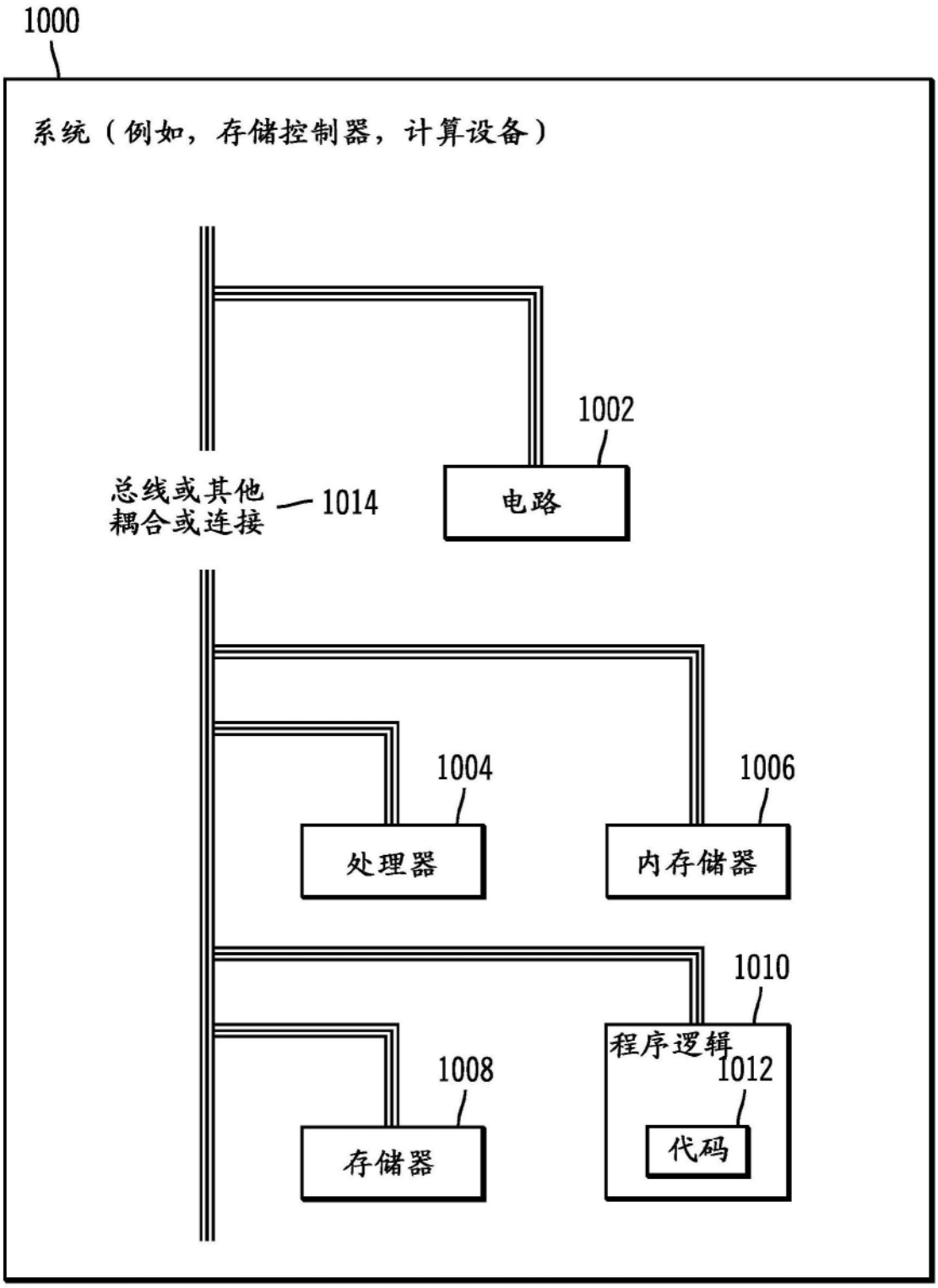

1000
系统（例如，存储控制器，计算设备）
1002
电路
总线或其他
耦合或连接
1014
1004
处理器
1006
内存储器
1010
程序逻辑
1012
代码
1008
存储器

图10