



(21)申请号 201310039200.3

(22)申请日 2013.01.31

(65)同一申请的已公布的文献号

申请公布号 CN 103226519 A

(43)申请公布日 2013.07.31

(30)优先权数据

13/362,099 2012.01.31 US

(73)专利权人 安华高科技通用IP(新加坡)公司

地址 新加坡新加坡市

(72)发明人 马克·艾施 阿南特·巴德丁尼

加里·J·斯梅尔顿

(74)专利代理机构 北京律盟知识产权代理有限公司

责任公司 11287

代理人 王田

(51)Int.Cl.

G06F 12/0871(2016.01)

G06F 11/16(2006.01)

(56)对比文件

US 5586291 A, 1996.12.17, 说明书第5-9
栏,附图1-2,附图5-6.

US 5586291 A, 1996.12.17, 说明书第5-9
栏,附图1-2,附图5-6.

US 6279078 B1, 2001.08.21, 说明书第1-2
栏.

US 6574709 B1, 2003.06.03, 全文.

CN 102122235 A, 2011.07.13, 全文.

审查员 李娇

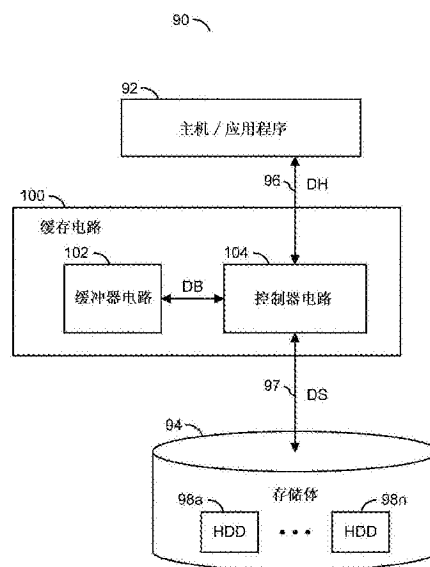
权利要求书2页 说明书9页 附图11页

(54)发明名称

冗余缓存数据的弹性缓存

(57)摘要

本发明涉及冗余缓存数据的弹性缓存,具体提供了一种用于冗余缓存数据的弹性缓存的设备。该设备可具有多个缓冲器和电路。该电路可被配置为(i)从主机接收写入请求以在存储体中存储写入数据,(ii)基于与写入请求相关联的冗余组织,在缓冲器中分配多个区域,以及(iii)在多个区域中存储写入数据,其中,(a)多个区域中的每一个位于缓冲器中的不同一个内,以及(b)响应写入请求,多个区域被动态链接在一起。



1. 一种用于冗余缓存数据的弹性缓存的设备, 包括:

多个缓冲器; 以及

电路, 被配置以 (i) 从主机接收写入请求以在存储体中存储写入数据, (ii) 基于所述写入请求从多个冗余组织中选择特定冗余组织, (iii) 基于所述写入数据的大小和所述特定冗余组织两者, 在所述缓冲器中分配多个区域, 以及 (iv) 在所述多个区域中存储所述写入数据, 其中, (a) 所述多个区域中的每一个别区域被分配在所述缓冲器中的不同缓冲器内, (b) 响应所述写入请求, 所述多个区域被动态链接在一起, 以及 (c) 两个以上链接跨越所述缓冲器之间的一个或多个边界。

2. 根据权利要求1所述的设备, 其中, 所述电路还被配置以 (i) 将所述写入数据的小于所有所述写入数据的一部分从所述缓冲器复制到所述存储体, 以及 (ii) 响应于复制所述写入数据的所述一部分, 解链所述多个区域中的至少一个但小于所有所述多个区域的区域。

3. 根据权利要求1所述的设备, 其中, 所述多个区域通过改变与所述多个区域中的每一个相对应的相应参数来动态链接。

4. 根据权利要求3所述的设备, 其中, 所述相应参数中的每一个被改变为指向分配给所述写入数据的所述多个区域中的另一区域。

5. 根据权利要求1所述的设备, 其中, 所述电路还被配置以 (i) 基于从所述主机接收到的另一请求, 选择非冗余组织, 以及 (ii) 基于所述非冗余组织分配一个或多个区域。

6. 根据权利要求1所述的设备, 其中, (i) 所述电路还被配置以基于与从所述主机接收到的另一写入请求相关联的所述冗余组织的另一冗余组织, 分配另一多个区域, 以及 (ii) 所述另一冗余组织不同于所述特定冗余组织。

7. 根据权利要求1所述的设备, 其中, 所述特定冗余组织包括以所述存储体中的独立盘的冗余阵列使用的组织。

8. 根据权利要求1所述的设备, 其中, 所述设备被实施为一个或多个集成电路。

9. 一种用于冗余缓存数据的弹性缓存的方法, 包括以下步骤:

(A) 从主机接收对设备的写入请求以在存储体中存储写入数据,

(B) 基于所述写入请求从多个冗余组织中选择特定冗余组织,

(C) 基于所述写入数据的大小和所述特定冗余组织两者, 在多个缓冲器中分配多个区域; 以及

(D) 在所述多个区域中存储所述写入数据, 其中, (i) 所述多个区域中的每一个别区域位于所述缓冲器中的不同缓冲器内, 以及 (ii) 响应所述写入请求, 所述多个区域被动态链接在一起, 以及 (iii) 两个以上链接跨越所述缓冲器之间的一个或多个边界。

10. 根据权利要求9所述的方法, 还包括以下步骤:

将所述写入数据的小于所有所述写入数据的一部分从所述缓冲器复制到所述存储体, 以及响应于复制所述写入数据的所述一部分, 解链所述多个区域中的至少一个但小于所有所述多个区域的区域。

11. 根据权利要求9所述的方法, 其中, 所述多个区域通过改变与所述多个区域中的每一个相对应的相应参数来动态链接。

12. 根据权利要求11所述的方法, 其中, 所述相应参数中的每一个被改变为指向分配给所述写入数据的所述多个区域中的另一区域。

13. 根据权利要求9所述的方法,还包括以下步骤:

基于从所述主机接收到的另一请求,选择非冗余组织;以及基于所述非冗余组织分配一个或多个区域。

14. 根据权利要求9所述的方法,还包括以下步骤:

基于与从所述主机接收到的另一写入请求相关联的所述冗余组织的另一冗余组织,分配另一多个区域,其中,所述另一冗余组织不同于所述特定冗余组织。

15. 根据权利要求9所述的方法,其中,所述特定冗余组织包括以所述存储体中的独立盘的冗余阵列使用的组织。

冗余缓存数据的弹性缓存

技术领域

[0001] 本发明总体上涉及存储装置缓存,更具体地,涉及用于实施冗余缓存数据的弹性缓存的方法和/或设备。

背景技术

[0002] 在常规存储系统中,整个系统和应用程序的性能通过缓存经常存取的数据来提高。在许多应用中,若缓存使用回写模式,则工作负荷受益于此且性能得到提高。保持在回写模式中的缓存数据有时通过提供冗余来保护以防止数据丢失。冗余通常利用缓存控制器中的两个闪存模块来实施。两个闪存模块利用镜像永久配对,有效减少了将缓存数据保持在单个闪存模块中的有效空间。假定仅有某些主机数据受益于冗余,则使闪存模块中有效空间被低效使用。

[0003] 将期望实施冗余缓存数据的弹性缓存。

发明内容

[0004] 本发明涉及用于冗余缓存数据的弹性缓存的设备。该装置可具有多个缓冲器和电路。该电路可被配置为(i)从主机接收写入请求以在存储体中存储写入数据,(ii)基于与写入请求相关联的冗余组织,在缓冲器中分配多个区域(延伸区,extent),以及(iii)在多个区域中存储写入数据,其中,(a)多个区域中的每一个位于缓冲器中的不同一个内,以及(b)响应写入请求,多个区域被动态链接在一起。

[0005] 本发明的目标、特征和优势包括提供了一种用于实施冗余缓存数据的弹性缓存的方法和/或设备,该冗余缓存数据可以(i)提供附加有效缓存空间,(ii)与常规技术相比,提供了对给定数量的缓存介质的更好缓存覆盖,(iii)使用现有区域结构,(iv)使用现有管理基础设施,(v)为缓存冗余实施各种RAID级,(vi)与常规技术相比,减少了对于给定数量的主机输入/输出活动的缓存写入数量,(vii)增强了对于给定主机输入/输出活动的闪存缓冲器的寿命,和/或(viii)实施固态驱动器作为用于缓存数据的大缓冲器。

附图说明

[0006] 根据以下详细描述和所附权利要求以及附图,本发明的这些和其他目标、特征和优势将变得显而易见,其中:

[0007] 图1是系统的实例性实施的框图;

[0008] 图2是根据本发明优选实施方式的系统中缓存电路的实例性实施的框图;

[0009] 图3是区域的实例性结构的示意图;

[0010] 图4是实例性缓存数据的示意图;

[0011] 图5是区域的几个实例性组的示意图;

[0012] 图6是具有弹性镜像缓存的实例性缓存组织的示意图;

[0013] 图7是用于服务写入请求的实例性方法的流程图;

- [0014] 图8是用于服务读取请求的实例性方法的流程图；
- [0015] 图9是用于分配区域的实例性方法的流程图；
- [0016] 图10是用于去分配区域的实例性方法的流程图；以及
- [0017] 图11是示出在缓存电路中来自弹性镜像脏写(elastically mirroring dirty writes)的空间益处的图表示意图。

具体实施方式

[0018] 本发明的某些实施方式可弹性使用有效缓存空间,以提供写入缓存空间的部分的冗余。弹性方式通常利用增加缓存空间整体有效部分来缓存读取数据和/或未保护(例如,非冗余)写入数据的智能管理技术来实施。此外,弹性方法可被模型化为示出其他益处,诸如聚集(例如,某些数据可以每个区域为基础在单独物理装置上镜像,使得若存储和缓存通过多个节点(服务器)共享,则被镜像区域的每个部分可驻留在单独节点上)以及策略选择(例如,某些存储磁盘使用受保护的(或冗余的)缓存数据和基于选择策略,某些不保护缓存数据)。

[0019] 参照图1,示出了系统90的实例性实施的框图。该系统(或设备)90通常包括块(或电路)92、块(或电路)94、网络(或总线)96、网络(或总线)97和块(或电路)100。电路94通常包括多个块(或电路)98a-98n。电路100通常包括块(或电路)102和块(或电路)104。电路92-104可表示可被实施为硬件、软件、硬件和软件的结合或其他实施的模块和/或块。

[0020] 双向信号(例如,DH)可经由网络96在电路92与电路100之间交换。信号DH可被用于在电路92与电路100之间传送数据(例如,读取数据和/或写入数据)、请求和状态信息。双向信号(例如,DB)可在电路102与电路104之间交换。信号DB可在电路102与电路104之间传递缓存数据和元数据。双向信号(例如,DS)可经由网络97在电路94与电路100之间交换。信号DS可在电路100与电路94之间传送数据和关联命令(例如,读取命令,写入命令和配置命令)。

[0021] 电路92可实施一个或多个主机电路(或服务器)和/或一个或多个软件应用程序。电路92通常可操作以通过信号DH和DS经由电路100和网络96从电路94读取数据和/或向电路94写入数据。电路92还可操作以通过信号DH沿相应地址信息(例如,逻辑块地址(LBAs))向电路100提出请求(例如,读取请求和写入请求)。此外,电路92可操作以通过信号DH向电路100发出命令(例如,预取、刷新、配置)。信号DH还可由电路100控制以将来自电路100的状态信息(例如,缓存命中(hit),缓存错失)回传到电路92。

[0022] 电路94可实施一个或多个存储体(storage volume)。电路94响应写入命令通常可操作以存储从电路100经由信号DS接收到的数据。电路94响应读取命令还可操作以经由信号DS向电路100提供数据。存储体可被实施为逻辑体、虚拟体和/或物理体。电路94可以是存储区域网络(SAN)、网络附加存储(NAS)和/或磁盘阵列子系统(DAS)架构的一部分。

[0023] 网络96可实施一个或多个数字通信网络和/或总线。网络96通常可操作以提供电路92与电路100之间的通信。网络96的实施可包括但不限于互联网、以太网、光纤网络、无线网络、有线网络、射频通信网络和/或底板总线中的一个或多个。

[0024] 网络97可实施一个或多个数字通信网络和/或总线。网络97通常可操作以提供电路94与电路100之间的通信。网络97的实施可包括但不限于互联网、以太网、光纤网络、无线

网络、有线网络、射频通信网络和/或底板总线中的一个或多个。

[0025] 每个电路98a-98n可实施大容量存储驱动器。在某些实施方式中,每个电路98a-98n可实施硬盘驱动器(例如,HDD)。电路98a-98n通常可操作以经由信号DS从电路100存储数据和向电路100提供数据。电路98a-98n可被布置和操作为一个或多个存储体。在某些实施方式中,电路98a-98n可被组织和操作为一个或多个冗余和独立的磁盘(例如,RAID)配置。RAID组织可包括但不限于RAID0、RAID1、RAID2、RAID3、RAID4、RAID5、RAID6、RAID7、RAID10、RAID53和/或RAID0+1的组织。可实施其他RAID组织级别以满足特定应用程序的标准。此外,多个不同RAID组织可由电路98a-98n同时表示。

[0026] 电路100可实施缓存电路。电路100通常可操作以缓存在电路92与电路94之间被交换的数据。对于从电路92向电路94移动的写入数据,缓存可实施回写(或后写)策略。脏写入数据可利用在多个介质之间传播的冗余组织缓存到电路100,冗余组织和传播可被配置为使得在单个介质失效导致失效介质中部分或全部的写入数据变为不可访问和/或丢失之后,保持对脏写入数据的访问。电路100可实施位于电路92、电路94中或在中间点处的缓存。在某些实施方式中,电路100可被实施为一个或多个集成电路。

[0027] 电路102可实施多个缓冲器电路。电路102可操作以存储在电路92与电路94之间传送的缓存(或缓冲)数据。电路102还可操作以存储与缓存数据相关联的元数据。缓存数据可在区域中布置。每个区域可表示为特定数据文件或软件程序保留的存储空间连续块。元数据通常包括用于管理区域的信息。在某些实施方式中,电路102可整个被实施在电路100的内部(如图所示)、整个在电路100的外部、部分在电路100的内部和外部、或跨多个节点(或服务器)而聚集。

[0028] 电路104可实施控制器装置。电路104可操作以从电路92接收在电路94中存储写入数据的写入请求,基于与写入请求相关联的冗余组织在电路102的多个缓冲器中分配多个区域,并在多个区域中存储写入数据。多个区域中的每个区域通常位于不同缓冲器中。响应写入请求,多个区域可被动态链接在一起。响应从缓冲器向电路94复制写入数据,电路104还可操作以解链多个区域中的至少一个。在某些实施方式中,电路104可被实施为一个或多个集成电路。

[0029] 参照图2,示出了根据本发明优选实施方式的电路100的实例性实施的框图。电路102通常包括多个块(或电路)106a-106n。电路104通常包括块(或模块)108、块(或模块)110和块(或模块)112。电路106a-106n可表示可被实施为硬件、软件、硬件和软件的结合或其他实施的模块和/或块。模块108-112可被实施为可执行控制电路104的固件(或软件或程序指令或者代码)。

[0030] 信号DH可通过接口114由电路104发送和接收。信号DS可通过接口116由电路104发送和接收。信号DB可被实施为多个信号(例如,DBa-DBn)。每个信号DBa-DBn通常建立电路104的接口118a-118n与电路106a-106n的相应接口120a-120n之间的通信。

[0031] 每个电路106a-106n可实施缓冲器电路。每个电路106a-106n可操作以存储缓存数据和关联元数据。在某些实施方式中,每个电路106a-106n可被实施为固态驱动器(例如,SSD)。固态驱动器的常见大小范围可从1特字节(terabyte)到2特字节。在其他实施方式中,每个电路106a-106n可被实施为闪存驱动器。闪存的常见大小范围可从1吉字节到32吉字节。在又一实施方式中,每个电路106a-106n可被实施为双倍数据速率(例如,DDR)存储器电

路。DDR存储器的常见大小范围可从1吉字节到64吉字节。可实施其他大小以满足特定应用程序的标准。电路102中的电路106a-106n的数量可多于、等于或少于电路94中的电路98a-98n的数量。

[0032] 模块108可实施区域控制器。模块108可操作以在电路102中分配和去分配用于服务从电路92接收到的请求(例如,读取请求和/或写入请求)的一个或多个区域。模块108还可操作以改变在与电路102中的每个区域相关联的元数据中的参数来创建分组空间(例如,读取空间和/或写入空间)。分组可通过将区域链接在一起以形成可在冗余保护组织中容纳缓存数据的更大存储器空间(或超级区域)来实现。将区域分配(或链接)到分组空间通常包括改变关联区域中的每个参数以指向分组中的下一区域。指针可在从一个或多个初始区域到一个或多个最终区域的链路中继续。最终区域的参数可指回初始区域以闭合链接。模块108还可改变元数据中的参数以将分组的空间向后分解成单独区域和/或更小分组空间。在缓存数据被缓冲在非冗余组织中的情况下,模块108可将一个或多个区域的参数改变为分配(空)值,以指示当前该区域被用于缓冲缓存数据。当区域不再用于缓冲缓存数据时,模块108可将关联参数改变为未分配(或未使用)值。

[0033] 模块110可实施缓存控制器。模块110可操作以执行标准缓存操作和标准缓存策略。缓存操作通常包括但不限于将数据写入电路102,从电路102读取缓存(或缓冲)数据,将元数据写入电路102,从电路102读取元数据,搜索缓存数据的标签,当发现匹配标签时指示缓存命中,当发现不匹配标签时指示缓存错失,跟踪电路102中的脏写入数据,将脏写入数据刷新(例如,复制)到电路94,跟踪有效和无效缓存行,跟踪有效和无效缓存字等。模块110还可操作以实施关于缓存数据的保留和去除的一个或多个缓存策略。缓存策略可包括但不限于用于脏写入数据的回写模式,用于某些写入数据类型的连续写入模式,某些写入数据类型的冗余(或保护)缓冲,其他写入数据类型的非冗余(或非保护)缓冲,用于脏写入数据的一个或多个刷新策略,清除电路102内的空间的一个或多个去除策略等。

[0034] 模块112可实施RAID控制器。模块112通常可操作以基于分配给数据的RAID策略在各种RAID配置中组织(或布置或者格式化)缓冲在电路102和存储在电路94中的数据。

[0035] 当应保护的写入数据从电路92接收时,模块112可将写入数据重新安排到RAID组织间的给定组织中。重新安排的写入数据可基于写入数据的大小(例如,写入数据跨过多少区域边界)和对应于写入请求的RAID组织(例如,RAID0、RAID3、RAID5等)随后被存储在两个或更多个电路106a-106n中。可实施其他数据保护形式以满足特定应用程序的标准。

[0036] 当受保护的缓存数据从电路102向电路92提供时,模块112可操作以检查数据错误以及如果可用则修正错误。若数据已被条形化(stripe),则模块112还可在将被请求数据提供给电路92之前,将柄(stipe)收集回相应的块。

[0037] 可在将数据保持在RAID组织中的同时执行RAID保护的写入数据从电路102向电路94的传送。例如,若数据在被镜像RAID1组织中,则数据可经由信号DB和DS从两个电路106a-106n(例如,106a和106b)被传送到相应的两个电路98a-98n(例如,98f和98g)。在另一实例中,若数据在利用带有两个条区域的奇偶区域的RAID3组织中,则数据可经由网络97从三个电路106a-106n(例如,106a、106b和106c)被传送到相应的三个电路98a-98n(例如,98d、98g和98k)。

[0038] 某些RAID保护存储数据从电路94向电路102的传送(例如,读取数据)可在从电路

94到模块112的RAID组织和从模块112到电路102的非冗余组织中执行。例如,存储在RAID3配置中的数据可从电路94被复制到模块112作为三个并行传送(例如,两个区域和奇偶区域)。模块112可与正常一样执行错误检查和错误纠正,并将被存储数据的单个非冗余副本传送到电路102。

[0039] 参照图3,示出了区域140的实例性结构的示意图。在电路102中的缓存数据和在电路94中的存储数据可以区域(或窗口)的形式组织。每个区域通常表示数据连续块的集合,该数据连续块作为用于分配、去分配、跟踪和管理数据的单个单元来处理。实例性区域140通常包括多个数据线142a-142n。每个数据线142a-142n可包括多个数据块144a-144n。每个数据块144a-144n通常包括多个字节、字或其他数据单元。此外,每个数据块144a-144n可具有相应逻辑块地址(例如,LBAA至LBAN)。可实施其他区域组织以满足特定应用程序的标准。

[0040] 参照图4,示出了实例性缓存数据的示意图。该实例示出了用于冗余的RAID1组织。为简化起见,仅示出了电路106a-106n中的两个电路(例如,106a-106b)。每个电路106a-106b可被组织为一个或多个区域和相应元数据。某些区域(例如,写入区域)可承载未交给电路94的脏写入数据。其他区域(例如,读取区域)可基于某些基本缓存标准承载来自电路94的缓存数据。在电路106a或106b失效的情况下,承载脏写入数据的写入区域可利用冗余保护来缓冲。承载缓存读取的所有其他区域可在没有冗余保护的情况下缓冲,因为在电路106a或106b失效时丢失的缓存数据可在失效发生后从电路94向工作电路106a和106b重新填充。

[0041] 参照图5,示出了在电路102内的几个实例性区域组的示意图。在所示实例中,电路102可包括电路106a-106d。可实施更少或更多电路106a-106n以满足特定应用程序的标准。如图所示,多个区域组、多个组类型(例如,RAID1、RAID3、RAID4、RAID5、RAID10、RAID0+1等)和/或多个未分组区域(例如,读取和写入区域)可同时存在于电路102中。

[0042] 每个区域和相应元数据可利用现有方法(或技术)管理。具体地,每个区域可具有用于跟踪和识别区域的唯一区域识别码。对于每个主机输入/输出请求,LBA可被映射到区域以定位到正确区域,从而服务主机请求。参数(或字段)可被添加到每个区域的元数据以跟踪用于冗余保护的另区域识别码。

[0043] 当在单个区域内的写入数据由电路100从电路92接收时,两个或更多个区域可被分配给一组。相应参数可被更新以在环路中指向(或链接)彼此。将区域分配给组可被配置为使得单个组中的每个区域被分配在不同电路106a-106n中。

[0044] 在写入数据跨多个区域和/或横跨区域边界的某些实施方式中,两个或更多个区域的集合可被处理为大区域。两个或更多个大区域可被分配给一组,且相应参数可被更新以在环路中指向彼此。大区域中的每个集合可在不同电路106a-106n中分配。例如,RAID1组150通常包括电路106b中两个区域的集合和电路106c中两个区域的另一集合。电路106b中的两个区域的每个参数可指向电路106c中的相应区域。类似地,电路106c中的两个区域的每个参数可指向电路106b中的相应区域。

[0045] 在写入数据跨多个区域和/或横跨区域边界的其他实施方式中,邻接区域可位于不同电路106a-106n中且被每个冗余组织链接。例如,在RAID3组152中的数据可具有缓冲在电路106b中的初始区域,缓冲在电路106c中的下一区域和缓冲在电路106d中的奇偶区域。

[0046] 基于缓存策略,某些写入数据可被缓冲在非冗余组织154内的电路102中。在由于

电路102失效而导致数据丢失的情况下,这种未保护写入数据可由电路92取代。未保护写入数据还可具有适于潜在数据丢失的低重要性。

[0047] 对于由于读取失误而执行的每个缓存填充操作,可根据要重新填充电路102中的数据的位置和数量,一个或多个区域的集合可被分配给一个或多个电路106a-106n。被分配集合中的每个参数可被改变(或程序化)为指示相应区域被分配到非冗余组织中的值(例如,空值或分配值)。

[0048] 在区域组中的脏写入数据被刷新到电路94之后,组中的区域可不被分配以获得有效空间用于进一步分配。可通过更新相邻区域的参数而不分配区域,以将未分配区域从组中解链。未分配区域可由用于保护写入数据的冗余组织来确定。例如,由于两个被镜像区域包括所有写入数据,所以具有RAID1组织的区域组可去分配被镜像区域中的任一个。具有RAID3组织的区域组可解链被分配以缓冲奇偶数据的区域。缓冲实际数据的剩余区域可被保留在组中。

[0049] 当电路106a-106n失效时,标准恢复技术可被用于服务来自幸存区域的主机请求。例如,在失效电路106a-106n(例如,电路106a)中丢失块的具有RAID5组织的区域组156可使用幸存块和在幸存电路106a-106n(例如,106b和106d)中的奇偶块来重构丢失的块。重构的块可随后用于满足读取/写入请求。

[0050] 参照图6,示出了带有弹性镜像缓存的实例性缓存组织160的示意图。该实例通常示出了利用两个电路106a-106n(例如,106a和106b)的镜像组织。区域1和区域10两者可被分配给共用冗余写入空间162。该分配可通过改变以下两者来实现,即(i)与指向区域10的识别码的区域1相关联的参数,以及(ii)与指向区域1的识别码的区域10相关联的参数。缓冲在电路106a中的区域4可通过将关联参数改变为空值(或分配值)而被分配给读取空间164。缓冲在电路106b中的另一区域8可通过将关联参数更新为空值(或分配值)而被分配给另一读取空间166。其他区域(例如,带有附图标记168的区域12)可不驻留数据,且因此相应参数可被设置为未分配(或未使用)值。用于位于电路106a中的区域(例如,区域1-6)的元数据可被缓冲在元数据空间170中。用于位于电路106b中的区域(例如,区域7-12)的元数据可被缓冲在元数据空间172中。

[0051] 参照图7,示出了用于服务写入请求的实例性方法180的流程图。方法(或过程)180可在电路100中实施。方法180通常包括步骤(或状态)182、步骤(或状态)184、步骤(或状态)186、步骤(或状态)188、步骤(或状态)190以及步骤(或状态)192。步骤182-192可表示可被实施为硬件、软件、硬件和软件的结合或其他实施的模块和/或块。

[0052] 在步骤182中,电路104可经由信号DH从电路92接收写入请求。对缓冲在电路102中的缓存数据的搜索可在步骤184中由电路104(例如,模块110)执行,以确定用于写入数据的适当区域是否已存在于电路102中。若适当区域在电路102中不可用(例如,缓存错失),则电路104(例如,模块108)可向电路106a-106n分配电路102中未使用的区域,以服务步骤186中的写入请求。

[0053] 一旦区域组可用于服务写入请求,则模块110可在步骤188中从电路94复制对应于写入区域的存储数据。若从电路94接收到的存储数据已不在受保护组织(或格式)中,则电路104(例如,模块112)可将数据安排到受保护组织中,作为步骤188的一部分。在步骤188中,数据可由模块110连续缓冲到合适的电路106a-106n。在步骤190中,模块110和112可由

受保护组织来组织从电路92接收到的写入数据,并将该写入数据存储到电路102中。由于写入数据尚未被复制回电路94中(例如,回写策略),所以模块110可更新相应元数据以标记新写入的数据为脏写入数据。

[0054] 若在电路102中发现写入数据(例如,缓存命中),则模块108可将附加区域分配和链接到导致缓存命中的现有一个或多个区域。此后,在步骤188中,区域组可以与已发生缓存错失时的相同方式从电路94填充。在步骤190中,新写入数据可由模块110添加到电路106a-106n。

[0055] 参照图8,示出了用于服务读取请求的实例性方法200的流程图。方法(或过程)200可在电路100中实施。方法200通常包括状态(或步骤)202、状态(或步骤)204、状态(或步骤)206、状态(或步骤)208和状态(或步骤)210。步骤202-210可表示可被实施为硬件、软件、硬件和软件的结合或其他实施的模块和/或块。

[0056] 在步骤202中,电路104可经由信号DH从电路92接收读取请求。在步骤204中,对缓冲在电路102中的缓存数据的搜索可由电路104(例如,模块110)来执行。该搜索可确定用于读取数据的合适区域是否已存在于电路102中。若合适区域在电路102中不可用(例如,缓存错失),则电路104(例如,模块108)可在电路106a-106n中分配一个或多个未使用区域,以服务步骤206中的读取请求。

[0057] 一旦所分配区域可用于服务读取请求,则模块110可在步骤208中从电路94复制对应于读取区域的存储数据。在步骤208中,接收数据可由模块110连续缓冲到合适电路106a-106n。在步骤210中,模块110可将被请求读取数据的副本从该区域传送到电路92。在步骤204中,若在电路102中发现读取数据(例如,缓存命中),则方法200可继续步骤210,其中,电路104将被请求读取数据传送到电路92。

[0058] 参照图9,示出了用于分配区域的实例性方法220的流程图。方法(或过程)220可在电路100中实施。方法220通常包括步骤(或状态)222、步骤(或状态)224、步骤(或状态)226、步骤(或状态)228、步骤(或状态)230、步骤(或状态)232、步骤(或状态)234以及步骤(或状态)236。步骤222-236可表示可被实施为硬件、软件、硬件和软件的结合或其他实施的模块和/或块。

[0059] 在步骤222中,电路(例如,模块108)可确定应被分配以服务请求的多个区域。区域的数量可基于被缓冲的数据量和用于数据的组织类型(例如,冗余保护或非冗余)。若在步骤224中确定组织类型为非冗余,则方法220可继续步骤226。电路104(例如,模块108)可基于与从电路92接收到的读取请求相关联的非冗余组织来分配一个或多个区域。在步骤226中,该分配通常包括改变从未分配值到分配值(或空值)的区域的相应参数。所产生的一个或多个区域(例如,图6中区域164)通常准备填充来自电路94的数据。

[0060] 若在步骤224中确定组织类型为冗余,则方法220可继续步骤228。在步骤228-236期间,多个区域可通过改变对应于多个区域中的每一个的相应参数以指向多个区域中的另一区域而动态链接在一起。在步骤228中,模块108可初始化计数器(例如,N=1)。在步骤230中,初始分配区域的相应参数可被更新(或改变)以指向下一被分配区域。计数器在步骤232中可由模块108更新。对更多区域的检查可在步骤234中执行。若更多区域应被链接到组中,则模块108可更新下一区域的参数以在步骤230中指向其后区域。可继续从步骤230到234的循环,直到所有区域已被添加到组中。在步骤236中,模块108可通过更新最后区域的参数以

指回初始区域来闭合链接。所产生的区域组(例如,图6中的组162)可准备填充数据。

[0061] 参照图10,示出了用于去分配区域的实例性方法240的流程图。方法(或过程)240可在电路100中实施。方法220通常包括步骤(或状态)242、步骤(或状态)244、步骤(或状态)246以及步骤(或状态)248。步骤242-248可表示可被实施为硬件、软件、硬件和软件的结合或其他实施的模块和/或块。

[0062] 在步骤242中,电路104(例如,模块110)可将脏写入数据从电路102刷新到电路94中。在某些情况下,写入数据可从电路102读取到电路104、被纠错、被恢复到原始形式、被布置到RAID配置中、以及从电路104被传送到电路94。在其他情况下,脏写入数据可被复制为类似从电路106a-106n中的每一个被复制到相应电路98a-98n。例如,当区域被组织为两个或更多个数据块和奇偶块时,每个数据块和奇偶块可无变化地从适当电路106a-106n经由信号DB和DS复制到相应电路98a-98n。

[0063] 在步骤244中,电路104(例如,模块108)可确定哪个区域可从刚刷新的区域组中被解链和不分配。例如,若写入数据由镜像(例如,RAID1组织)保护,则由于两个区域包括所有写入数据,所以任一区域可被解链和不分配。在另一实例中,若写入数据以RAID4格式被保护,则缓冲奇偶数据的区域可被解链和不分配。在步骤246中,模块108可通过改变相邻区域中的相应参数以旁路所选择区域来解链所选择区域。在步骤248中,模块108可将所选择区域的参数改变为未分配值。

[0064] 参照图11,示出了说明来自电路102中的弹性镜像脏写的空间益处的图表260的示意图。沿垂直轴的空间益处指示可用于缓冲非镜像数据的总区域的百分比。如图表260所示,空间益处可增加为使更少的区域表示数据的写入部分。当所有区域被镜像以保护脏写入数据时(例如,100%),可实现0百分比空间节省。在区域的更低百分比被用于缓冲镜像数据时,空间益处动态(或弹性)增加。当没有区域被用于缓冲镜像数据时(例如,0%),正常将被用于保持镜像数据的电路106a-106n中的50%可被替换以用于保持非脏写入数据和/或读取数据。

[0065] 根据被缓冲到电路100中的数据的读/写混合,可通过区域的弹性分组和取消分组来释放附加缓存空间。对于缓存介质(例如,电路106a-106n)中的给定投入,自由缓存空间中的动态变化通常提供了更好的缓存覆盖。因此,具有扩展频繁访问数据的服务器(例如,电路92)可利用附加有效缓存空间。

[0066] 本发明的某些实施方式一般利用在缓存装置中正常发现的大多数区域和元数据管理架构。用于弹性分配和不分配区域的附加参数可构建在现有架构的顶端。用于创建、修改和破坏区域组的技术可从镜像保护(例如,RAID1级别)扩展到其他RAID级别。此外,对于给定的主机输入/输出数据,由于冗余可限制脏数据,所以该技术通常减少向电路102中的写入的数量。在电路106a-106n实施有闪存时,对于给定主机输入/输出活动,写入数量的减少可提高闪存介质的寿命。

[0067] 由图1-图11的示意图执行的功能可利用根据本说明书的教导程序化的常规通用处理器、数字计算机、微处理器、微控制器、RISC(精简指令集计算机)处理器、CISC(复杂指令集计算机)处理器、SIMD(单指令多数据)处理器、信号处理器、中央处理单元(CPU)、算术逻辑单元(ALU)、视频数字信号处理器(VDSP)和/或类似计算机中的一个或多个来实施,这对相关领域技术人员将是显而易见的。熟练程序员可基于本公开的教导很容易地准备适

当软件、固件、代码、例程、指令、操作码、微代码和/或程序模块,这对相关领域技术人员也将是显而易见的。来自一个或几个介质中的软件通常由机器实施的一个或多个处理器执行。

[0068] 本发明还可通过ASIC(专用集成电路)、平台ASIC、FPGA(现场可编程门阵列)、PLD(可编程逻辑器件)、CPLD(复杂可编程逻辑器件)、门海、RFIC(射频集成电路)、ASSP(专用标准产品)、一个或多个单片集成电路、布置成倒装芯片模块和/或多芯片模块的一个或多个芯片或管芯的制备来实施,或者通过互联常规部件电路的适当网络来实施,如本文所述,其修改对于本领域技术人员将是显而易见的。

[0069] 因此,本发明还可包括计算机产品,其可以是存储介质或介质和/或传输介质或包括指令的介质,该指令可被用于对机器编程以执行根据本发明的一个或多个过程或方法。由机器执行包括在计算机产品中的指令以及外围线路的操作,可将输入数据转换成存储介质上的一个或多个文件和/或表示物理对象或物质的一个或多个输出信号,诸如音频和/或视频描述。存储介质可包括但不限于任何类型的盘,包括软盘、硬盘驱动器、磁盘、光盘、CD-ROM、DVD和磁光盘以及诸如ROM(只读存储器)、RAM(随机存取存储器)、EPROM(可擦写可编程ROM)、EEPROM(电可擦写可编程ROM)、UVPRM(紫外可擦写可编程ROM)、闪存、磁卡、光卡和/或适于存储电子指令的任何介质类型的电路。

[0070] 本发明的元件可形成一个或多个装置、单元、部件、系统、机器和/或设备的一部分或全部。装置可包括但不限于服务器、工作站、存储阵列控制器、存储系统、个人计算机、膝上型计算机、笔记本电脑、掌上电脑、个人数字终端、便携式电子装置、电池供电装置、机顶盒、编码器、解码器、代码转换器、压缩器、解压缩器、预处理器、后处理器、发送器、接收器、收发器、密码电路、蜂窝电话、数码相机、定位和/或导航系统、医疗设备、头戴式显示器、无线装置、录音、音频存储和/或音频回放装置、视频记录、视频存储和/或视频回放装置、游戏平台、外设和/或多芯片模块。相关领域技术人员将理解,本发明的元件可以其他器件类型来实施以满足特定应用的标准。如本文所使用,术语“同时”是指描述共享某些共用时间段的事件,但该术语并非限于在相同时间点开始、在相同时间点结束的事件或具有相同时间段的事件。

[0071] 尽管已参照其优选实施方式示出并描述了本发明,但本领域技术人员将理解,在不偏离本发明范围的情况下,可做出形式和细节上的各种变更。

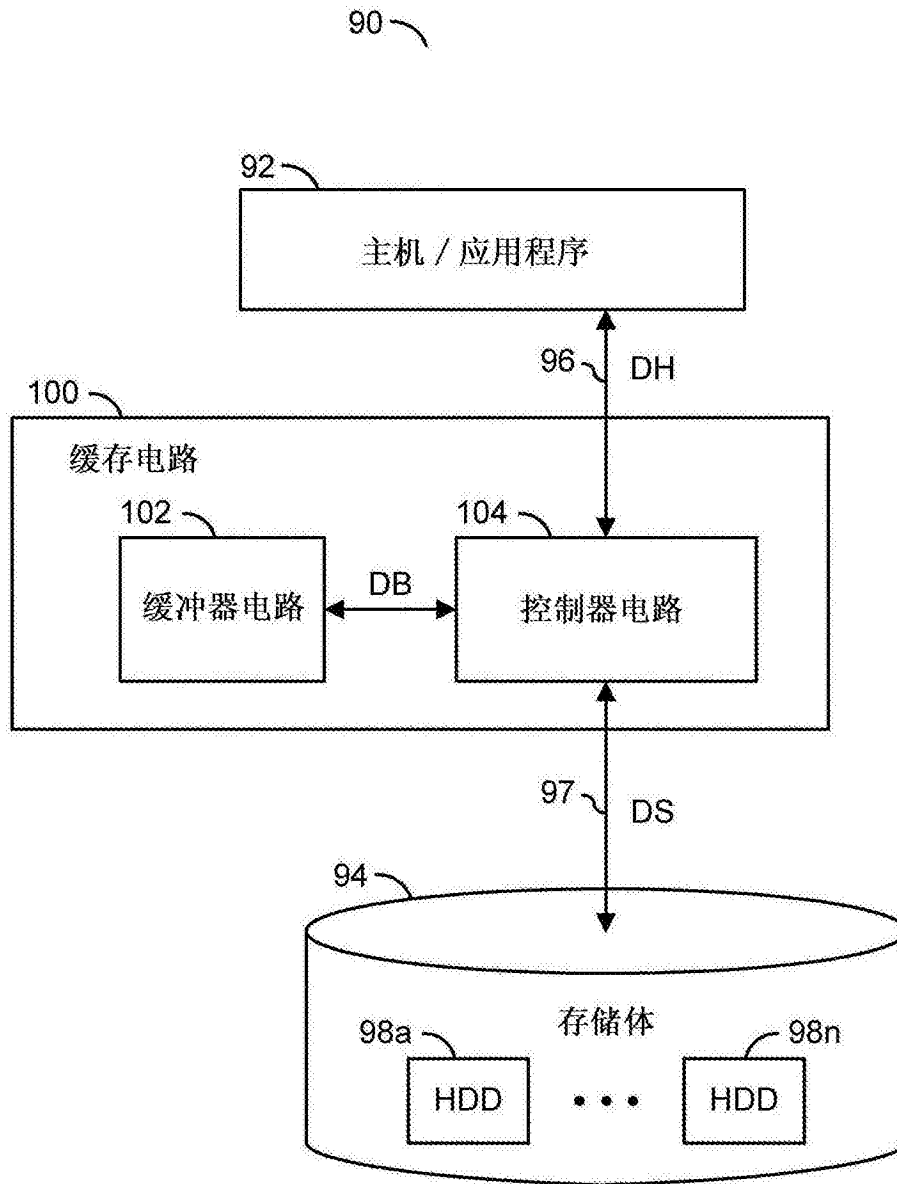


图1

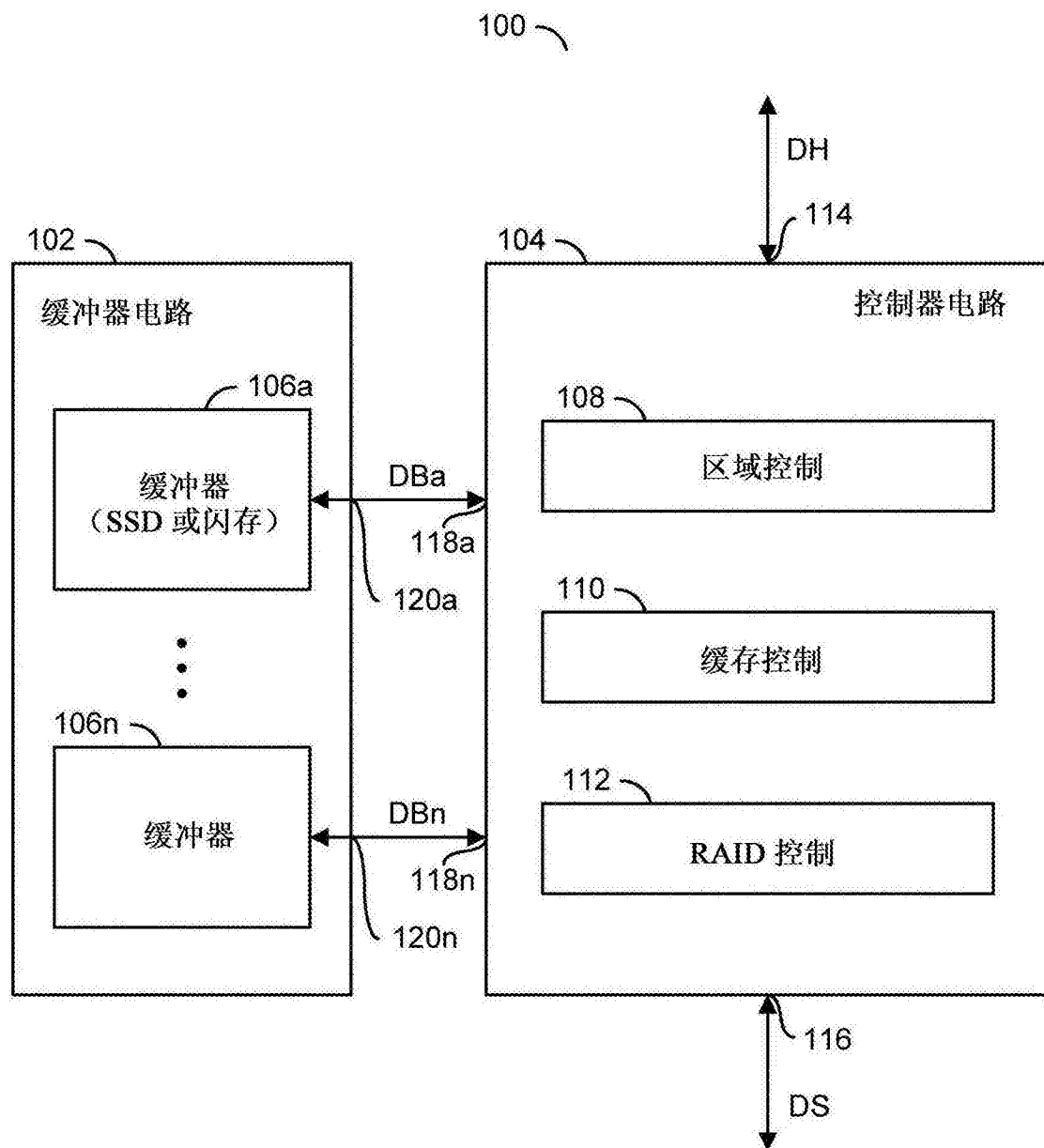


图2

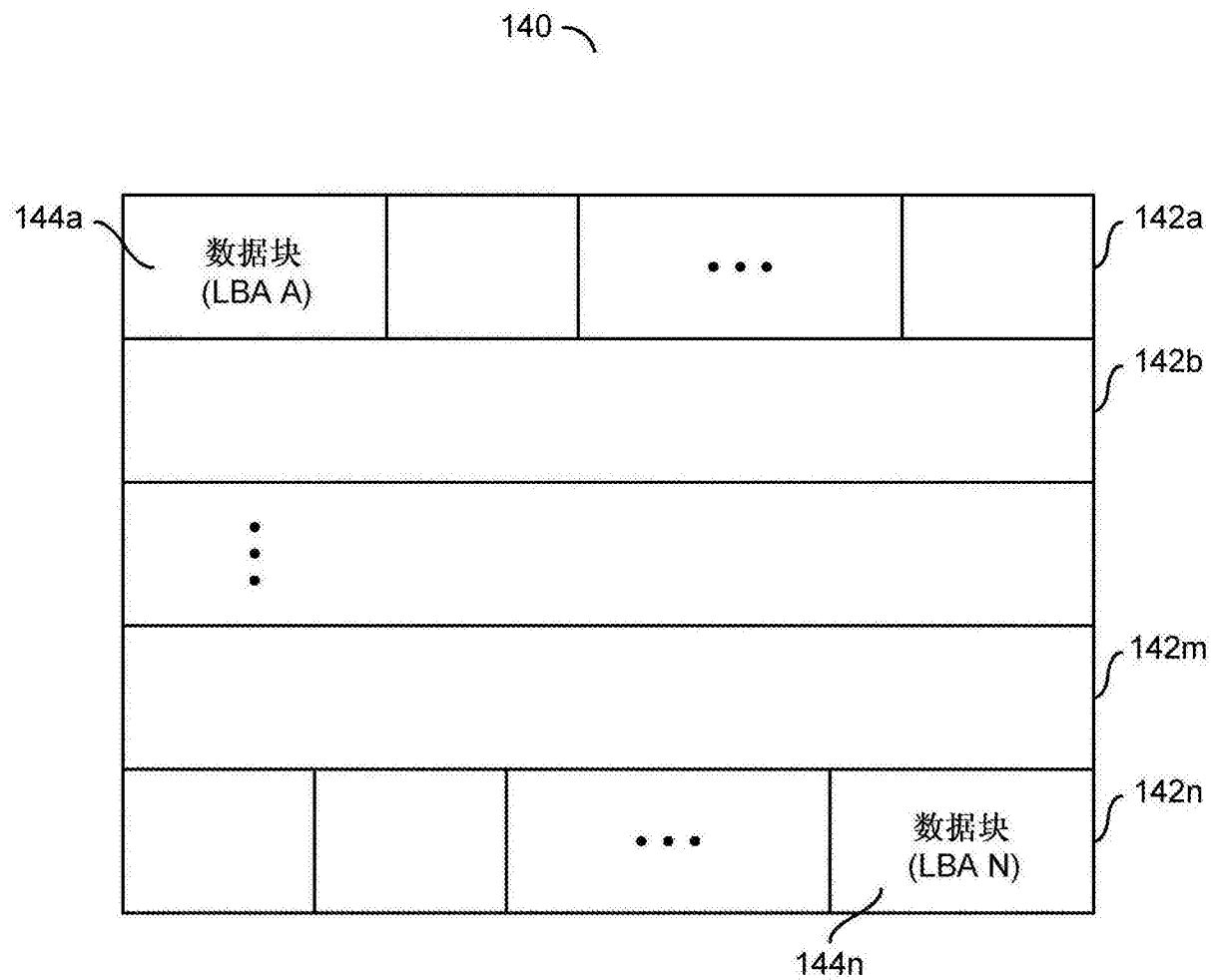


图3

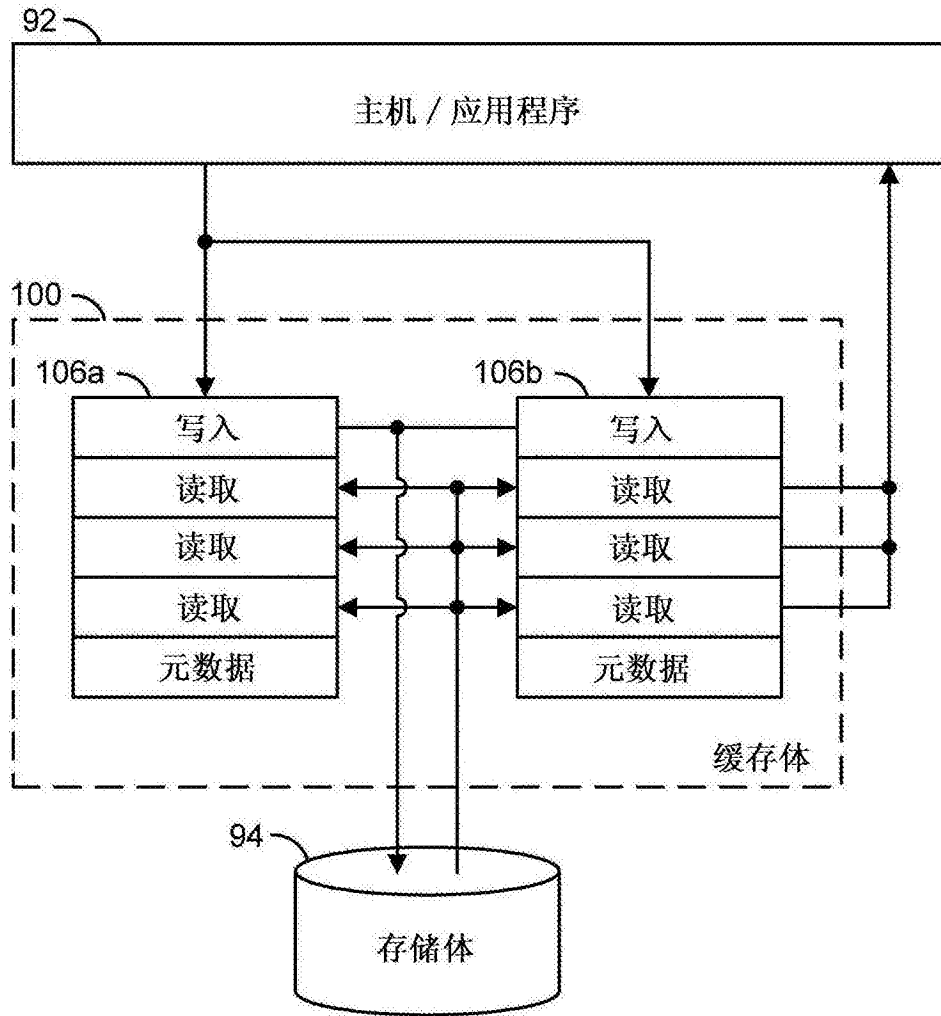


图4

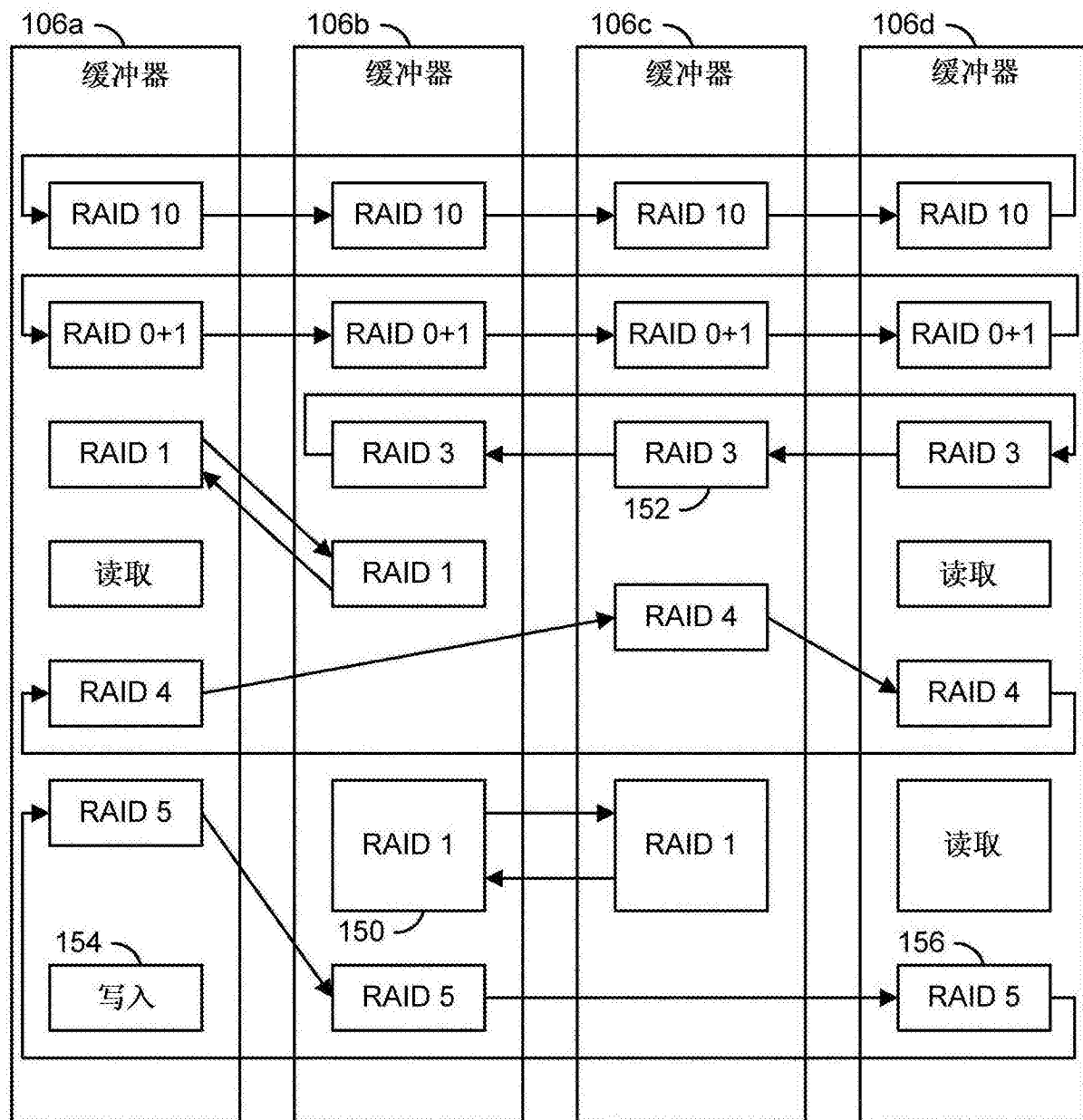


图5

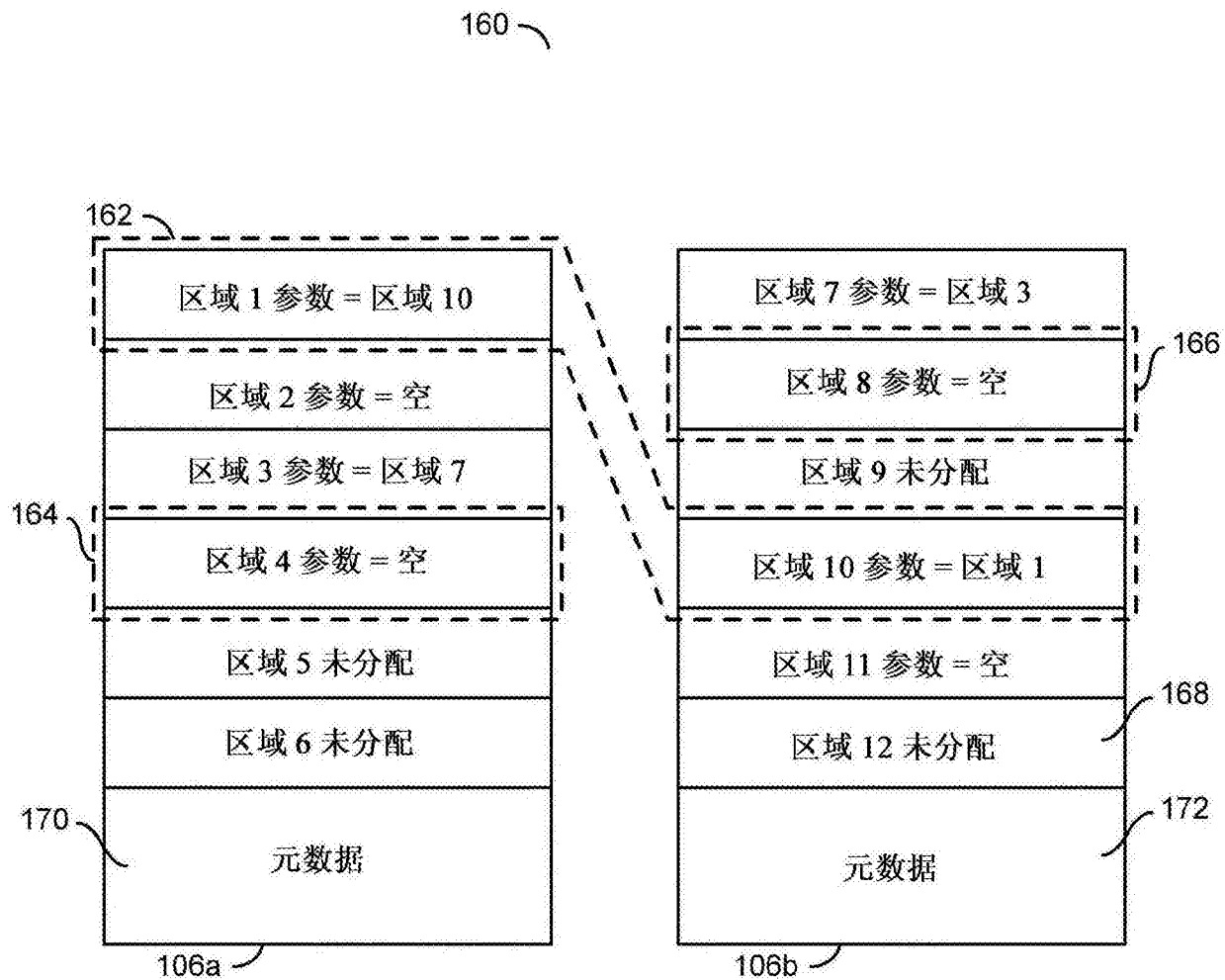


图6

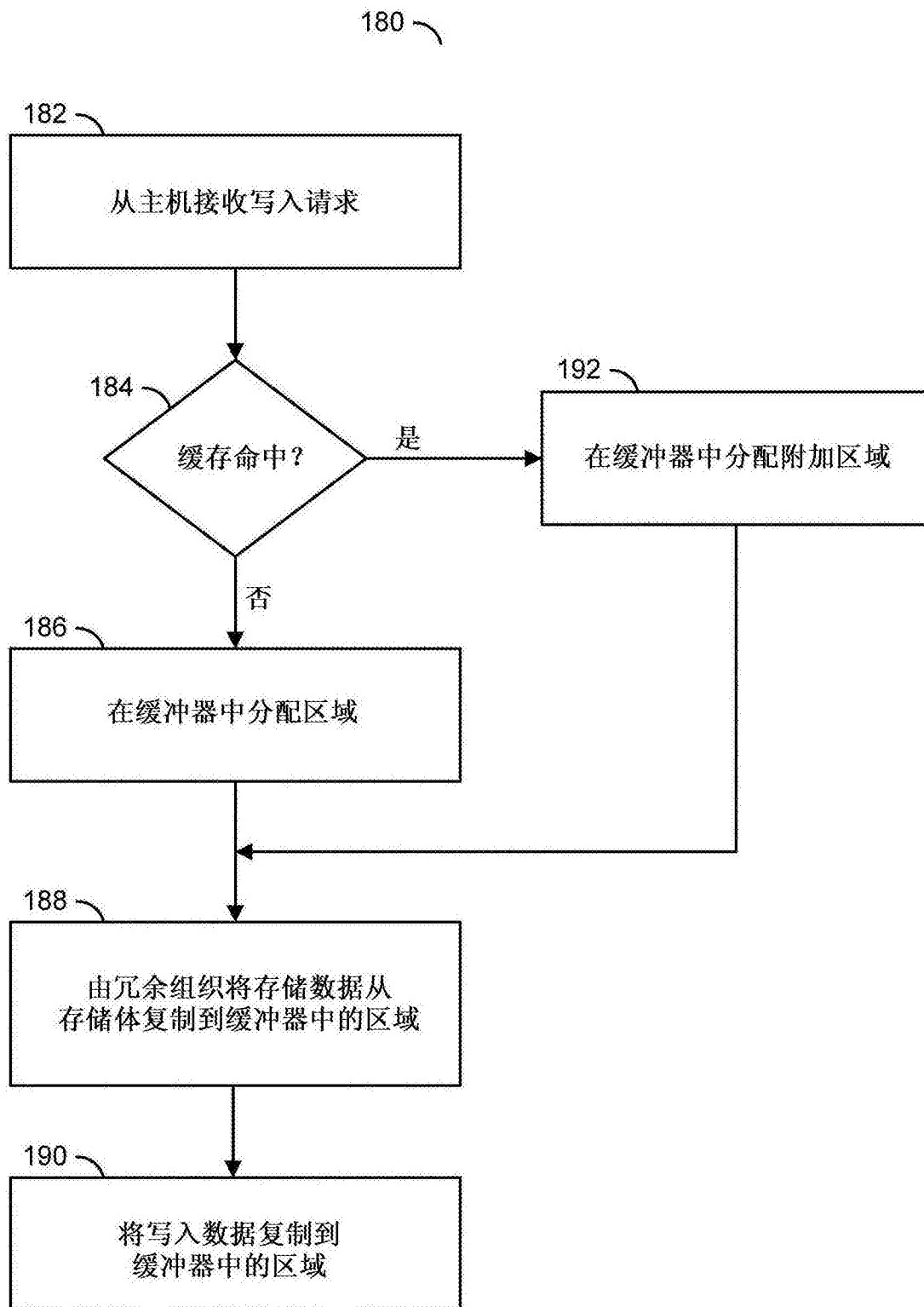


图7

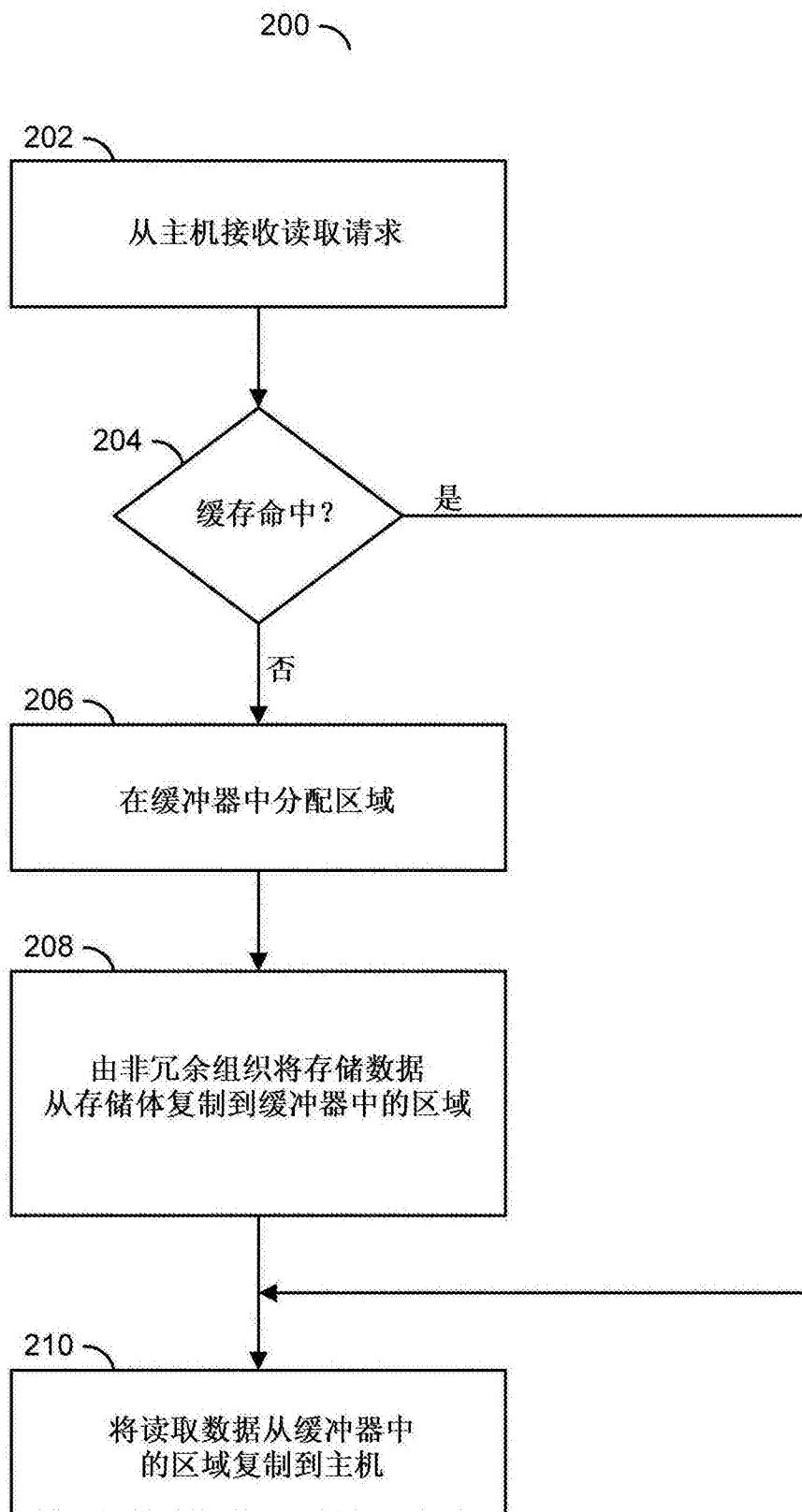


图8

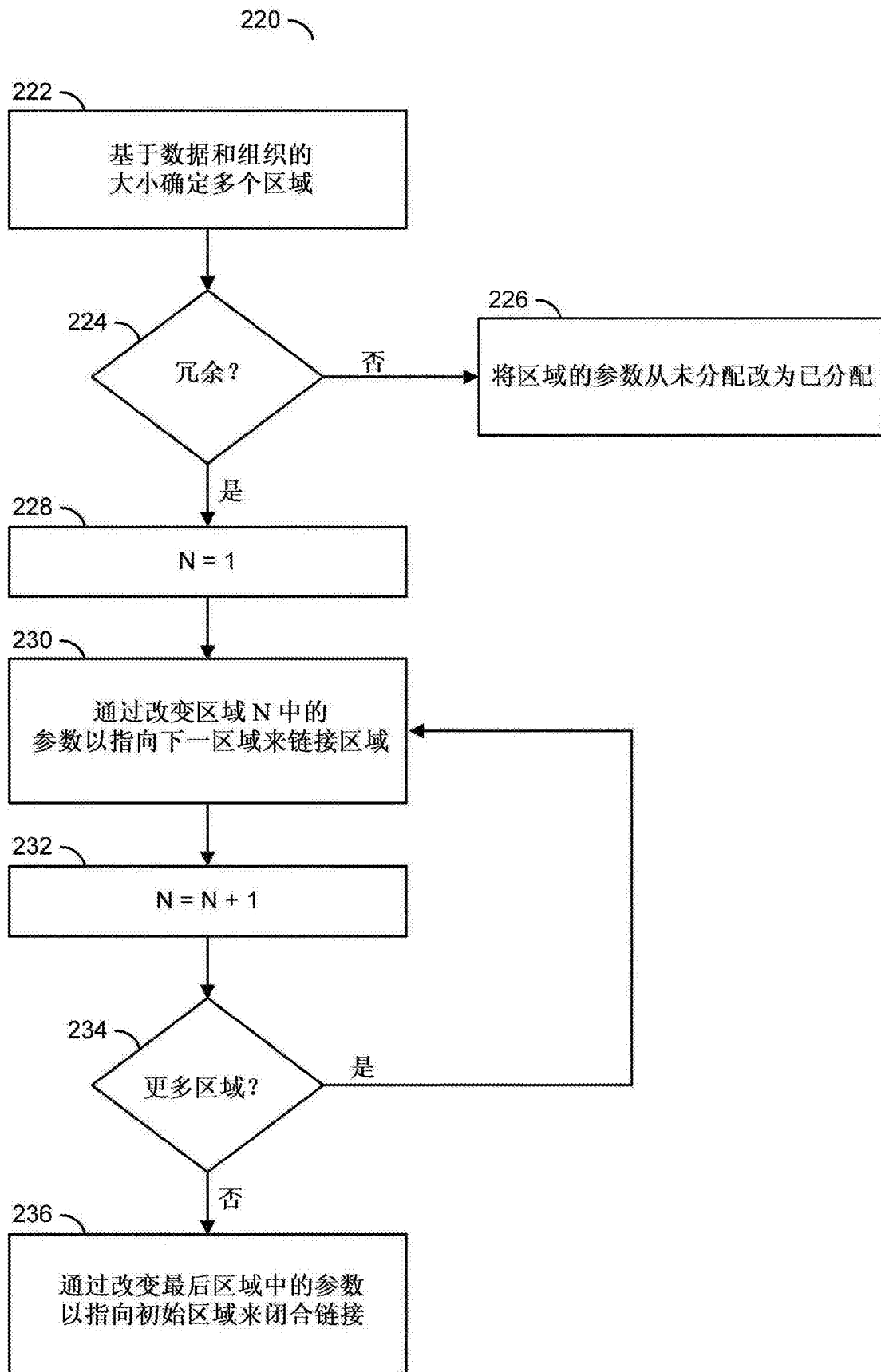


图9

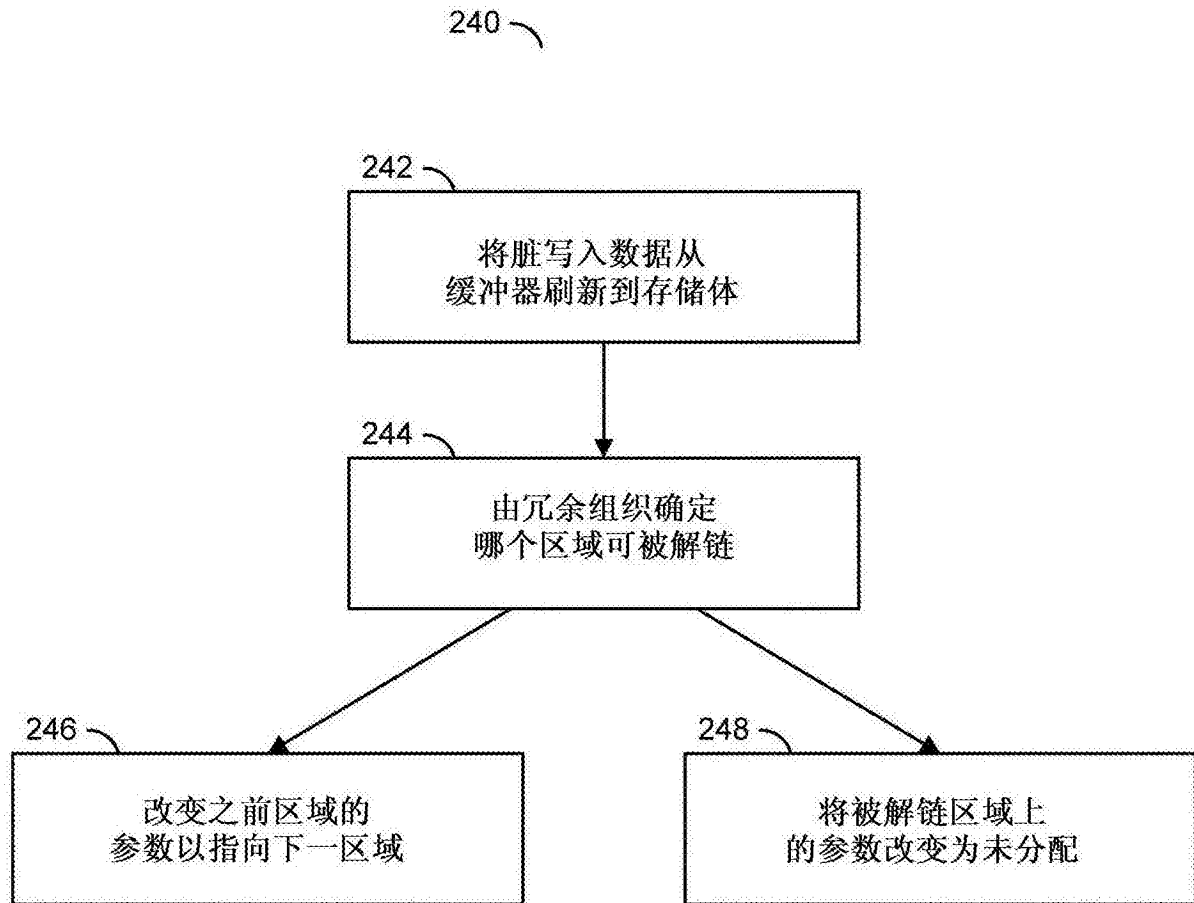


图10

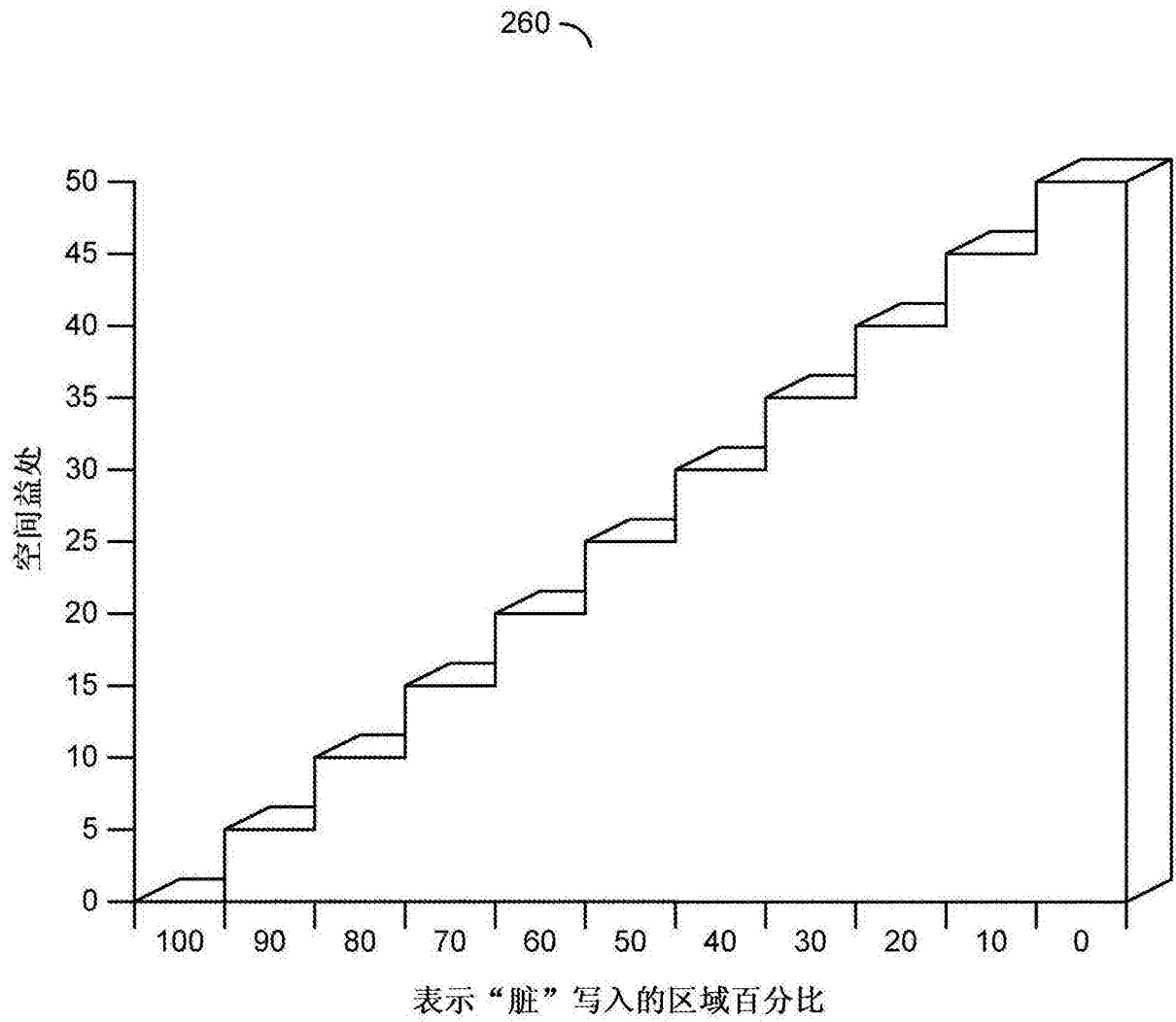


图11