



(12)发明专利

(10)授权公告号 CN 104412241 B

(45)授权公告日 2017.06.27

(21)申请号 201380032838.2

(72)发明人 D·兹罗继安尼斯 P-A·拉森

(22)申请日 2013.06.13

(74)专利代理机构 上海专利商标事务有限公司 31100

(65)同一申请的已公布的文献号

申请公布号 CN 104412241 A

代理人 陈斌

(43)申请公布日 2015.03.11

(51)Int.Cl.

G06F 12/02(2006.01)

(30)优先权数据

G06F 17/30(2006.01)

13/529,865 2012.06.21 US

G06F 12/08(2016.01)

(85)PCT国际申请进入国家阶段日

2014.12.22

(56)对比文件

US 2006155791 A1,2006.07.13,

(86)PCT国际申请的申请数据

US 5577246 A,1996.11.19,

PCT/US2013/045749 2013.06.13

US 2009327621 A1,2009.12.31,

(87)PCT国际申请的公布数据

CN 101981545 A,2011.02.23,

W02013/192020 EN 2013.12.27

CN 101124554 A,2008.02.13,

(73)专利权人 微软技术许可有限责任公司

审查员 周丹丹

地址 美国华盛顿州

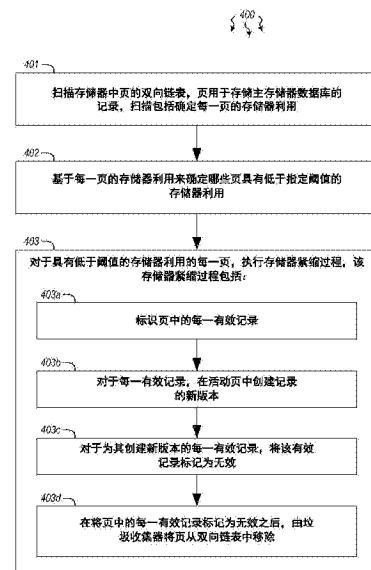
权利要求书1页 说明书10页 附图8页

(54)发明名称

用于主存储器数据库的存储器紧缩机制

(57)摘要

本发明涉及用于主存储器数据库中执行存储器紧缩的方法、系统和计算机程序产品。主存储器数据库将记录存储在页内，页以双向链表组织在分区堆中。存储器紧缩过程使用准更新来将来自要被清空的页的记录移动到分区堆中的活动页。准更新在活动页中创建记录的新版本，该新版本具有与记录的旧版本相同的数据内容。创建新版本可以使用一事务来执行，该事务采用针对依赖性的等待以在该事务创建新版本的同时允许记录的旧版本被读取，藉此最小化存储器紧缩过程对主存储器数据库中的其它事务的影响。



1. 一种由存储主存储器数据库的计算机系统执行的用于在所述主存储器数据库内执行存储器紧缩的方法，所述方法包括：

扫描存储器中的页的双向链表，所述页用于存储所述主存储器数据库的记录，扫描包括确定每一页的存储器利用(401)；

基于每一页的存储器利用来确定哪些页具有低于指定阈值的存储器利用(402)；

对于具有低于所述阈值的存储器利用的每一页，执行存储器紧缩过程，所述存储器紧缩过程包括(403)：

标识所述页中的每一有效记录(403a)；

对于每一有效记录，在活动页中创建所述记录的新版本(403b)；

对于为其创建新版本的每一有效记录，将所述记录标记为无效(403c)；以及

在将所述页中的每一有效记录标记为无效之后，将所述页清空并且从所述双向链表中移除(403d)。

2. 如权利要求1所述的方法，其特征在于，所述存储器紧缩过程使用悲观事务来执行，所述悲观事务采用针对依赖性的等待以允许在所述悲观事务创建和提交所述记录的新版本的同时旧版本被其它事务读取。

3. 如权利要求1所述的方法，其特征在于，扫描所述页的双向链表包括从所述页的头部读取每一页的存储器利用。

4. 如权利要求1所述的方法，其特征在于，所述存储器紧缩过程进一步包括在要被处理的每一页的头部中设置一标志以指示所述页正被清空。

5. 如权利要求4所述的方法，其特征在于，所述存储器紧缩过程进一步包括：

在活动页中创建每一记录的新版本之后，访问所述页中的存储器块的列表，其中在设置所述标志的同时接收到针对所述存储器块的释放存储器调用；以及

释放所述列表中的每一存储器块。

6. 如权利要求1所述的方法，其特征在于，所述扫描由工作者线程响应于接收到工作项来执行。

7. 如权利要求1所述的方法，其特征在于，所述双向链表包括分区堆中所有部分为空的页。

8. 如权利要求7所述的方法，其特征在于，所述分区堆涉及表或数据库之一。

9. 如权利要求1所述的方法，其特征在于，标识有效记录包括标识具有被设为无穷大的结束时间戳的记录，并且将记录标记为无效包括将所述记录的结束时间戳设为创建所述记录的新版本的事务的提交时间戳。

用于主存储器数据库的存储器紧缩机制

[0001] 背景

[0002] 1. 背景和相关技术

[0003] 计算机系统及相关技术影响社会的许多方面。的确,计算机系统处理信息的能力已转变了人们生活和工作的方式。现在,计算机系统通常执行在计算机系统出现以前手动执行的许多任务(例如,文字处理、日程安排、帐目管理等)。最近,计算机系统彼此耦合并耦合到其他电子设备以形成计算机系统及其他电子设备可在其上传输电子数据的有线和无线计算机网络。因此,许多计算任务的执行分布在多个不同的计算机系统和/或多个不同的计算环境中。

[0004] 许多计算机系统采用数据库来存储数据。一种类型的数据库(称为主存储器数据库或存储器内数据库)被存储在主存储器中。换言之,主存储器数据库的表被完整地存储在主存储器中,而不是存储在盘或其它类型的辅助存储上。由于存储器是计算机系统中有限且有价值的资源,以尽可能最高效的方式利用存储器是很重要的。

[0005] 在实现主存储器数据库时产生的一个问题是存储器碎片。图1示出了存储器碎片能够如何在主存储器数据库中发生。图1示出了包含各个页101a-101d的存储器100。一般而言,页是存储器中已经被分配用于存储数据库记录的一部分。在图1的左侧,这些页被示为充满的,这意味着每一页包含最大数目的记录,并且这些记录是有效的。左侧相应地表示数据库对存储器的有效利用。

[0006] 相反,在图1的右侧,页101a-101d被示为近乎是空的。具体来说,每一页被示为包含很少的有效记录。几乎为空的页未能高效地利用存储器,因为该页的所分配尺寸中仅一小部分被用来存储有效数据(并且因为该页存储至少一个有效记录,该页无法被回收以供后续使用)。页101a-101d可能通过各种方式变得几乎为空,诸如在删除页中的记录时。

[0007] 当主存储器数据库的页未被高效利用时,主存储器数据库的性能遭受损失。例如,较差的存储器利用可能导致数据库占据超过必要限度的存储器,从而限制了能够被有效地实现为主存储器数据库的数据库的大小。

[0008] 简要概述

[0009] 本发明涉及用于实现主存储器数据库中的存储器紧缩的方法、系统和计算机程序产品。存储器紧缩指的是将来自具有较差空间利用的页的记录重新定位到另一页,从而将记录“紧缩”到较少的页中。以此方式,具有较差空间利用的页(即,包含相对较少有效记录的页)能够被清空以允许这些页被回收以供其它使用。

[0010] 根据本发明的主存储器数据库能够使页内的记录结构化成以分区堆内的双向链表来组织。主存储器数据库可包括许多分区堆。可以通过将来自分区堆的一个页的记录移动到分区堆的活动页以藉此清空来自该页的所有有效记录来在分区堆上执行存储器紧缩。移动记录可包括对记录执行准更新。准更新指的是在活动页中创建记录的新版本但该新版本包含与旧版本相同的用户数据的事实。

[0011] 根据本发明的一个实施例,存储器紧缩可以通过扫描存储器中页的双向链表来执行。页存储主存储器数据库的记录。扫描包括确定每一页的存储器利用。基于每一页的存储

器利用,确定哪些页具有低于指定阈值的存储器利用。对于具有低于阈值的存储器利用的每一页,存储器紧缩过程通过以下操作来执行: (1) 标识页中的每一有效记录; (2) 对于每一有效记录,在活动页中创建该记录的新版本; (3) 对于为其创建新版本的每一有效记录,将该有效记录标记为无效; 以及 (4) 在页中的每一有效记录被标记为无效之后,通过垃圾收集器将该页从双向链表中移除。

[0012] 根据另一实施例,存储器紧缩线程确定应当在主存储器数据库的分区堆中的至少一个分区堆上执行存储器紧缩。存储器紧缩线程创建工作项,该工作项包含用于在分区堆上执行存储器紧缩的指令。存储器紧缩线程将工作项推送到关于在分区堆上执行存储器紧缩的工作者线程的队列中。

[0013] 工作者线程访问队列中的工作项以确定该工作项指令工作者线程在分区堆上执行存储器紧缩。工作者线程访问分区堆中双向链表页中每一页的头部中的存储器利用值。对于具有低于阈值的存储器利用值的每一页,工作者线程执行以下存储器紧缩过程: (1) 用于创建页中的每一有效记录的新版本,该新版本被创建在分区堆的活动页中; 以及 (2) 用于将页中为其创建新版本的每一有效记录标记为无效以藉此允许垃圾收集器随后清空该页的所有记录并且回收该页以供其它使用。

[0014] 提供本概述是为了以简化的形式介绍将在以下详细描述中进一步描述的一些概念。本概述不旨在标识出所要求保护的主题的关键特征或必要特征,也不旨在用于帮助确定所要求保护的主题的范围。

[0015] 本发明的附加特征和优点将在以下描述中叙述,并且其一部分根据本描述将是显而易见的,或者可通过对本发明的实践来获知。本发明的特征和优点可通过在所附权利要求书中特别指出的工具和组合来实现和获得。本发明的这些以及其它特征、优点和特征将根据以下描述和所附权利要求而变得更显而易见,或者可通过如此后阐述的对本发明的实践而获知。

附图说明

[0016] 为了描述可获得本发明的上述和其他优点和特征的方式,将通过参考附图中示出的本发明的具体实施例来呈现以上简要描述的本发明的更具体描述。可以理解,这些附图仅描述本发明的典型实施例,从而不被认为是对其范围的限制,本发明将通过使用附图用附加特征和细节来描述和说明,在附图中:

[0017] 图1示出了存储器碎片的示例;

[0018] 图2A示出了示例性主存储器数据库系统;

[0019] 图2B-2D示出了主存储器数据库中使用的示例性存储结构;

[0020] 图3示出了用于实现存储器紧缩的各种组件;

[0021] 图4示出了用于在主存储器数据库中执行存储器紧缩的示例性方法的流程图; 以及

[0022] 图5示出了用于在主存储器数据库中执行存储器紧缩的另一示例性方法的流程图。

具体实施方式

[0023] 本发明涉及用于实现主存储器数据库中的存储器紧缩的方法、系统和计算机程序产品。存储器紧缩指的是将来自具有较差空间利用的页的记录重新定位到另一页，从而将记录“紧缩”到较少的页中。以此方式，具有较差空间利用的页（即，包含相对较少有效记录的页）能够被清空以允许这些页被回收以供其它使用。

[0024] 根据本发明的主存储器数据库能够使页内的记录结构化成以分区堆内的双向链表来组织。主存储器数据库可包括许多分区堆。可以通过将来自分区堆的一个页的记录移动到分区堆的活动页以藉此清空来自该页的所有有效记录来在分区堆上执行存储器紧缩。移动记录可包括对记录执行准更新。准更新指的是在活动页中创建记录的新版本但该新版本包含与旧版本相同的用户数据内容的事实。

[0025] 根据本发明的一个实施例，存储器紧缩可以通过扫描存储器中页的双向链表来执行。页存储主存储器数据库的记录。扫描包括确定每一页的存储器利用。基于每一页的存储器利用，确定哪些页具有低于指定阈值的存储器利用。对于具有低于阈值的存储器利用的每一页，存储器紧缩过程通过以下操作来执行：(1)标识页中的每一有效记录；(2)对于每一有效记录，在活动页中创建该记录的新版本；(3)对于为其创建新版本的每一有效记录，将该记录标记为无效；以及(4)在将页中的每一记录标记为无效之后，允许通过现有的垃圾收集过程将该页从双向链表中移除。

[0026] 根据另一实施例，存储器紧缩线程确定应当在主存储器数据库的分区堆中的至少一个分区堆上执行存储器紧缩。存储器紧缩线程创建工作项，该工作项包含用于在分区堆上执行存储器紧缩的指令。存储器紧缩线程将工作项推送到关于在分区堆上执行存储器紧缩的工作者线程的队列中。

[0027] 工作者线程访问队列中的工作项以确定该工作项指令工作者线程在分区堆上执行存储器紧缩。工作者线程访问分区堆中双向链表页中每一页的头部中的存储器利用值。对于具有低于阈值的存储器利用值的每一页，工作者线程执行以下存储器紧缩过程：(1)用于创建页中的每一有效记录的新版本，该新版本被创建在分区堆的活动页中；以及(2)用于将页中为其创建新版本的每一有效记录标记为无效以藉此允许垃圾收集器移除记录并且藉此清空该页的所有记录并且回收该页以供其它使用。

[0028] 示例性环境

[0029] 本发明的各实施例可包括或利用专用或通用计算机，该专用或通用计算机包括诸如举例而言一个或多个处理器和系统存储器的计算机硬件，如以下更详细讨论的。本发明范围内的各实施例还包括用于承载或存储计算机可执行指令和/或数据结构的物理和其他计算机可读介质。这样的计算机可读介质可以是可由通用或专用计算机系统访问的任何可用介质。存储计算机可执行指令的计算机可读介质是计算机存储介质（设备）。承载计算机可执行指令的计算机可读介质是传输介质。由此，作为示例而非限制，本发明的各实施例可包括至少两种显著不同的计算机可读介质：计算机存储介质（设备）和传输介质。

[0030] 计算机存储介质（设备）包括RAM、ROM、EEPROM、CD-ROM、固态驱动器（SSD）（如基于RAM）、闪存、相变存储器（PCM）、其他类型的存储器、其他光盘存储、磁盘存储或其他磁存储设备、或可用于存储计算机可执行指令或数据结构形式的所需程序代码装置且可由通用或专用计算机访问的任何其他介质。相应地，要求保护的本发明可以被实现为存储在一个或多个计算机存储设备上的计算机可执行指令。

[0031] “网络”被定义为使得电子数据能够在计算机系统和/或模块和/或其它电子设备之间传输的一个或多个数据链路。当信息通过网络或另一个通信连接(硬连线、无线、或者硬连线或无线的组合)传输或提供给计算机时,该计算机将该连接适当地视为传输介质。传输介质可包括可用于携带计算机可执行指令或数据结构形式的所需程序代码装置并可由通用或专用计算机访问的网络和/或数据链路。上述的组合也应被包括在计算机可读介质的范围内。

[0032] 此外,在到达各种计算机系统组件之后,计算机可执行指令或数据结构形式的程序代码装置可从传输介质自动传输到计算机存储介质(设备)(或反之亦然)。例如,通过网络或数据链接接收到的计算机可执行指令或数据结构可被缓存在网络接口模块(例如,“NIC”)内的RAM中,然后最终被传输到计算机系统RAM和/或计算机系统处的较不易失性的计算机存储介质(设备)。因而,应当理解,计算机存储介质(设备)可被包括在还利用(甚至主要利用)传输介质的计算机系统组件中。

[0033] 计算机可执行指令例如包括,当在处理器处执行时使通用计算机、专用计算机、或专用处理设备执行某一功能或某组功能的指令和数据。计算机可执行指令可以是例如二进制代码、诸如汇编语言之类的中间格式指令、或甚至源代码。尽管用结构特征和/或方法动作专用的语言描述了本主题,但可以理解,所附权利要求书中定义的主题不必限于上述特征或动作。更具体而言,上述特征和动作是作为实现权利要求的示例形式而公开的。

[0034] 本领域的技术人员将理解,本发明可以在具有许多类型的计算机系统配置的网络计算环境中实践,这些计算机系统配置包括个人计算机、台式计算机、膝上型计算机、消息处理器、手持式设备、多处理器系统、基于微处理器的或可编程消费电子设备、网络PC、小型计算机、大型计算机、移动电话、PDA、平板、寻呼机、路由器、交换机等等。本发明也可在其中通过网络链接(或者通过硬连线数据链路、无线数据链路,或者通过硬连线和无线数据链路的组合)的本地和远程计算机系统两者都执行任务的分布式系统环境中实施。在分布式系统环境中,程序模块可以位于本地和远程存储器存储设备二者中。

[0035] 本发明可以在其中实现的特定示例环境是云环境。云环境包括互联计算组件的集合(例如,服务器计算系统的群集)。相应地,根据本发明的主存储器数据库可以被存储在云环境的主存储器中并在其中执行。

[0036] 存储器紧缩

[0037] 图2A示出了主存储器数据库系统200的示例性结构。系统200可包括任何合理数目的数据库,但在图2A中被示为包含三个数据库(DB1、DB2和DB3)。出于简化说明的目的,仅示出DB2中的表。DB2被示为包括存储固定大小的记录的三个表(T1、T2和T3)。尽管未示出,DB2还可包括用于存储可变大小的记录的各种表。单个存储器分配器(分别为VA1、VA2和VA3)用于为三个表中的每一表管理存储器分配。第四存储器分配器(VAg)用于为在DB2中存储可变大小的记录(未示出)的任何表管理存储器分配。

[0038] 存储器管理可以在存储器分配器级处进行监视。例如,对于存储器分配器VA1,表T1的存储器利用可以被监视以确定表T1是否正展现出高效的存储器利用,如下文将进一步描述的。

[0039] 图2B示出了主存储器数据库的表201的示例性存储结构。表201可以表示图2A中示出的表T1、T2或T3中的任一个。存储器中用于存储表201的记录的页被安排成各个分区堆,

这些分区堆包括图2B中示出的分区堆202和203。所使用的分区堆的数目可以变化并且可以是可配置的参数。

[0040] 在图2B中,分区堆202和203中的每一个被示为包括被安排成由每一页之间的双相箭头表示的双向链表的多个页。尽管在每一分区堆中仅示出三个页,但任何合理数目的页可存在于特定分区堆中,如省略号所表示的。

[0041] 在分区堆中使用双向链表存在若干优点。首先,页插入和移除能够在列表中的不同位置处发生,从而在常规操作期间降低了争用并且提高了可缩放性。这些操作可以被高效地执行而不要求从任一端遍历列表。第二,空的页可以被立即从列表移除并且释放给操作系统。作为对比,在使用堆栈数据结构时,空的页必须出现在堆栈的顶部以便被释放被操作系统。

[0042] 在一些实施例中,双向链表可以是无锁数据结构。无锁意味着该数据结构可以被访问(例如,读访问或写访问)而不要求数据结构上的锁。这些类型的锁常常也被称为锁存器。注意到,在此意义上的锁不同于下文在多版本并发性控制方案中描述的事务性锁(即,写锁)。事务性锁是记录或记录群上的锁,而非整个数据结构上的锁(即锁存器)。然而,双向链表不必以无锁的方式来实现。存储器紧缩可以用对这一数据结构的基于锁的实现来实现。然而,无锁可以更好地缩放并且因此尤其适用于存储器内数据库。

[0043] 每一页包含一个或多个记录并且包括头部。头部可以存储各种类型的信息,但出于本公开的目的,头部包括对该页的存储器利用的指示。例如,头部可以指示存储在该页中的记录的数目。这一指示可以具有各种形式,诸如具体数字、百分比等。如下文进一步描述的,这一指示允许标识出展现低存储器利用并且因此是存储器紧缩的良好候选的页。

[0044] 图2C示出了图2B中的表201添加了活动页210a和211a。一般而言,每一分区堆包括活动页(即,分区堆中向其添加了新记录的页)。活动页210a是分区堆202中向其添加新记录的当前页。类似地,活动页211a是分区堆203中向其添加新记录的当前页。

[0045] 活动页210a被示为已经包括两个记录210a1和210a2,这两个记录是已经被新添加到表201的记录。根据本发明的技术,记录202b1-1和202c1-1(分别是记录202b1和202c1的新版本)被添加到活动页210a。类似地,活动页211a被示为在创建记录203c1-1(其是记录203c1的新版本)之前未包括任何记录。

[0046] 如上文参考图2B所讨论的,每一页存储标识该页的存储器利用的信息。基于该信息,页能够被标识以供紧缩。例如,展现出低于某一阈值的存储器利用的页能够被选择用于存储器紧缩。在图2C的示例中,该阈值被假定为两个记录(这意味着存储少于两个有效记录的任何页将遭受紧缩)。页202b、202c和203c的存储器利用各自落入低于该阈值。相应地,来自这些页中的每一页的记录被示为重新定位到对应的活动页。

[0047] 具体来说,根据本发明,记录202b1、202c1和203c1分别从页202b、202c和203c“移动”到对应的活动页。“移动”的意思是在恰当的活动页中创建每一记录的新版本。该移动包括对记录的准更新,因为该记录的用户数据内容(常常被称为有效载荷)保持不变。相应地,记录202b1-1、202c1-1和203c1-1的内容分别与页202b、202c和203c中的记录202b1、202c1和203c1的内容相同。

[0048] 如下文将进一步描述的,记录202b1、202c1和203c1因而成为记录的旧版本,如网状线所指示的。无效记录的实际移除被延迟,直到它们不再对任何活动事务可见。一旦垃圾

收集器从页移除了所有记录,该页就可以被释放。图2D示出了在移动记录的这一过程发生之后页202b、202c和203c发生了什么的细节。

[0049] 如图2D中所示,页202b、202c和203c已经分别从随分区堆的其相应的双向链表中移除(出于清楚起见未示出活动页210a和211a)。这些页由垃圾收集器代理更新双向链表中的指针(例如,以使得页202a被链接到页202d(现在移除的页202c之前在右侧链接到的页))来从双向链表中被移除。类似地,分区堆203被修改,以使得页203d在双向链表中链接到页203b。注意到,尽管在该具体示例中使用双向链表,但也可使用其它数据结构。然而,双向链表是优选结构,因为其促进了页列表在分区堆内的整合(即,在两个方向上重新链接页),并且其允许在任何时候释放空的页(不同于在使用堆栈存储结构的情况下当页在堆栈顶部时才能释放页)。

[0050] 存储器紧缩过程可以在周期性的基础上(例如,每x秒)、响应于触发(例如,当检测到过多具有低存储器利用的页时)、按需(例如,响应于用户输入或来自另一过程的输入)等来执行。在一些实施例中,可以维护关于系统中每一存储器分配器的存储器利用的统计数据以允许检测展现出较差存储器利用的存储器分配器(即,涉及存储器分配器的分区堆)。

[0051] 图3示出了可用于实现参考图2B和2C描述的过程的各种组件。图3示出了如在图2B中出现的表201。图3还包括各种其它组件:主线程301(称为存储器紧缩(MC)线程)、工作者线程302和303、队列302a和303a、以及工作项310和311。

[0052] MC线程301执行如参考图2B所描述的检测过程。具体来说,MC线程301被配置成周期性地或以其它方式确定哪些存储器分配器正展现出低存储器利用。如上所述,这可以通过访问所存储的关于特定存储器分配器的存储器利用(即,涉及存储器分配器的分区堆的存储器利用)的统计数据来完成。

[0053] 当MC线程301检测到展现出低存储器利用的存储器分配器时,MC线程301可以创建一个或多个工作者项。在一些实施例中,可以针对涉及存储器分配器的每一个分区堆创建一个工作者项。图3显式地示出了已经为分区堆202和203中的每一个创建的一个工作者项。每一工作者项包括足够的指令以指令工作者线程在特定分区堆上执行记录的重新定位。

[0054] 在图3中描绘的示例中,工作者项310对应于分区堆202,而工作者项311对应于分区堆203。MC线程301将工作者项310和311分别插入到队列302a和303a(即,每一工作者线程具有对应的队列,MC线程301将工作者项插入到该队列中供该工作者线程进行处理)中。尽管图3中示出了两个工作者线程,但也可以向单个工作者线程指派工作者项310和311两者。

[0055] 例如,工作者线程可以按照各种方式来分区。这些分区可包括每核一个工作者线程,每套接口一个工作者线程,每非统一存储器接入(NUMA)节点一个工作者线程,以及每机器一个工作者线程。相应地,特定分区堆的工作项可被指派给被指派给该特定分区堆的工作者线程。

[0056] 每一工作者线程在处理其对应队列中的工作者项以执行存储器紧缩。每一工作者线程扫描工作者项中标识的分区堆以标识展现出低存储器利用的页。在标识展现出低存储器利用的页时,工作者线程能够将展现出低存储器利用的该页中的记录“移动”到活动页中。例如,工作者线程302能够在活动页210a中创建记录202b1和202c1的新版本,而工作者线程303能够在活动页210a中创建记录203c1的新版本。一旦创建了记录的新版本,旧版本可被相应地标记以指示该记录不再有效,藉此导致垃圾收集器清空该页以允许该页被回收。

[0057] 为了确保工作者线程正在处理的页不变成活动页,工作者线程可以在页头部中设置标志。该标志可以被设置为指示该页正在紧缩中。任何部分为空的页可能被存储器分配器选为活动页。相应地,工作者线程使用标志来向对应的存储器分配器指示在紧缩中的页不应当被选为活动页。

[0058] 该标志还可用来确保存储器紧缩过程不干扰存储器的释放或者阻塞或延迟垃圾收集过程。在设置该标志并且对正被处理的页中的存储器块作出释放存储器调用时,要被释放的存储器块被添加到链表,从而允许释放存储器调用返回(而不会释放存储器块)。相应地,存储器紧缩过程不会导致释放调用阻塞。一旦存储器紧缩过程完成,工作者线程检查链表以确定在紧缩过程期间是否有任何存储器块被添加到列表。工作者线程随后执行对添加到链表的存储器块的全部释放。

[0059] 多版本并发性控制方案

[0060] 将记录从一页重新定位到另一页能够以各种公知或其它已知的方式来执行。采用使用事务来创建更新的记录的多版本并发性控制方案的一种方式在共同拥有共同待审的题为“EFFICIENT MULTI-VERSION LOCKING FOR MAIN MEMORY DATABASES(用于主存储器数据库的高效多版本锁定)”的申请No.:13/042,269中公开,该申请整体通过援引纳入于此。采用使用乐观事务来创建更新的记录的多版本并发性控制方案的另一种方式在共同拥有共同待审的题为“OPTIMISTIC SERIALIZABLE SNAPSHOT ISOLATION(乐观可串行化快照隔离)”的申请No.:12/641,961中公开,该申请整体通过援引纳入于此。本发明的存储器紧缩过程可以使用这些多版本并发性控制方案中的任一个或者任何其它恰适的多版本并发性控制方案以在活动页中创建准更新的记录。准指的是创建记录的新版本但新版本的用户数据内容与旧版本的内容相同的事。以此方式,在存储器紧缩期间重新定位记录对其它事务的影响被最小化。尽管以上公开涉及乐观和悲观事务,但准更新对记录的重新定位并不取决于这些事务是乐观的还是悲观的。相反,可以使用任何多版本化方案,其中更新创建完全新的版本并且不原地更新。

[0061] 说明性地,可以对要被准更新的每一记录使用一事务来创建记录的新版本。例如,当以上描述的工作者线程确定记录要被重新定位时,其可创建新的事务以重新定位记录。相应地,在以下基于以上引用的申请no.:13/042,269和申请no.:12/641,961的公开的描述中,在存储器紧缩过程期间被创建用来将记录重新定位到活动页的事务是更新事务的一个示例(被称为TU)。

[0062] 在一个多版本并发性控制方案中,事务被给予分别指示事务的开始和结束事件的逻辑时间的两个唯一的时间戳。时间戳定义事务事件之间的整体排序。时间戳,如此处所使用的,可以是从单调递增计数器接收的值,且不限于时钟值。

[0063] 例如,当事务开始时,它可通过读取并递增时间戳计数器来接收时间戳。这一开始时间戳唯一地标识了事务并且因此在一些实施例中能够充当事务id。当事务终止时,事务能够通过读取时间戳计数器并且递增该计数器来接收结束时间戳。如果事务通过提交而终止,则这一结束时间戳还可充当其提交时间戳。时间戳的这一使用使得多版本化方案能够保留并发事务之间的可串行化性。

[0064] 主存储器数据库中的记录被版本化以允许多个事务的并发访问。时间戳还用于标识记录的各版本以及它们的有效时间。例如,记录的提交版本包含两个时间戳,开始时间戳

和结束时间戳。事务指定了每一次读取的逻辑读取时间，该读取时间通常等于事务的开始时间戳。仅在读取时间落入版本的有效时间内时版本才对事务可见；所有其它版本均被忽略。

[0065] 提交版本的开始时间戳等于创建该版本的事务的提交时间。例如，如果事务T1在其处理期间创建记录的一版本(诸如通过修改现有的记录或创建新的记录)，则所创建的版本将接收与事务T1的提交时间相同的开始时间戳。针对本公开的存储器紧缩过程，当记录被重新定位时，活动页中新创建的记录能够接收执行该重新定位的事务(即，工作者线程创建的事务)的开始时间戳。

[0066] 版本的结束时间戳最初被设为指示该时间戳尚未确定并且要被解读为无穷大的特殊值。然而，当另一事务T2提交对该版本的修改时(无论是对该版本的更新(因此创建新的版本)、还是对该版本的删除)，该版本的结束时间戳被设为事务T2的提交时间戳。换言之，一旦T2提交(并因此制作它的该记录的新版本或持久地删除该记录)，该记录的前一版本就不再是最新版本。一旦终止了所有当前活动的事务，旧版本将变得过时，即不再对任何事务可见并且可被丢弃。

[0067] 在T2提交之前，该版本的结束时间戳被设为T2的事务ID，因为T2的提交时间尚未可知。出于相同原因，这一相同的事务ID最初还用作新版本的开始时间戳。因此，当事务创建新版本时，它将其事务ID分配给正被修改的版本的结束时间戳和新版本的开始时间戳。一旦T2提交，它写入其提交时间戳作为旧版本的结束时间戳以及作为新版本的开始时间戳。为对包含有效时间戳的版本以及使临时事务ID被分配为其时间戳的版本之间进行区分，可以使用标志。

[0068] 参考本公开的存储器紧缩过程，一旦在活动页中创建更新记录的事务提交，活动页中新创建的记录接收事务的提交时间戳作为其开始时间戳，而被重新定位的记录的版本(即，在其中的记录被重新定位的页中的记录)接收事务的提交时间戳作为其结束时间戳。

[0069] 由于记录的旧版本具有指示其是“旧”版本的结束时间戳(例如，结束时间戳表示相对于时钟的过去时间，或结束时间戳是低于用于时间戳的单调递增计数器的当前值的数字)，在该记录不再对任何活动或将来事务可见的意义上而言知晓该记录是过时的。以此方式，一旦在活动页中创建经受紧缩的页中的每一记录的新版本，可以确定该页是空的(即，不包含记录，除了哪些过时记录之外)并且因此可被回收(在垃圾收集期间)。

[0070] 总结来说，在本发明的存储器紧缩过程中，工作者线程可以根据以上描述的多版本并发性控制方案来创建重新定位记录的事务。具体来说，工作者线程可以使用单独的事务(例如，如上所述的TU)来更新每一记录。相应地，为了“重新定位”记录，一般而言，工作者线程获得记录上的写锁定，如果在记录上存在任何读标记则针对依赖性而等待，在活动页中创建记录的新版本，并且一旦针对依赖性的任何等待(或提交依赖性)被移除则提交。

[0071] 图4示出了用于在主存储器数据库中执行存储器紧缩的示例性方法400的流程图。方法400将参考图2B-2D和3描述。

[0072] 方法400包括扫描存储器中的页的双向链表的动作401。页存储主存储器数据库的记录。扫描包括确定每一页的存储器利用。例如，分区堆202中的每一页可被扫描以确定分区堆202中的每一页的存储器利用。该扫描可以由工作者线程(例如工作者线程302)响应于从主存储器紧缩线程(例如MC线程301)接收到工作项来执行。

[0073] 方法400包括基于每一页的存储器利用来确定哪些页具有低于指定阈值的存储器利用的动作402。例如,如果指定阈值是两个有效记录,则可以确定分区堆202中的页202b和202c展现出低于指定阈值的存储器利用。

[0074] 方法400包括对于具有低于阈值的存储器利用的每一页执行存储器紧缩过程的动作403。存储器紧缩过程可以由指派给该分区堆的工作者线程(例如,分区堆202的工作者线程302)执行。

[0075] 动作403的存储器紧缩过程包括标识页中的每一有效记录的子动作403a。例如,对于页202b,可以标识有效记录202b1,而对于页202c,可以标识有效记录202c1。在实现多版本并发性控制方案时,这一标识可以通过读取页中的每一记录的结束时间戳来执行。如果记录的结束时间戳指示其是有效的(例如,如果结束时间戳是无穷大或者某一其它值以指示该记录不是“旧”的),则工作者线程将确定该记录是有效的。

[0076] 动作403的存储器紧缩过程包括对于每一有效记录在活动页中创建记录的新版本的子动作403b。例如,可以在活动页210a中创建针对记录202b1的新版本202b1-1,并且可以在活动页210a中创建针对记录202c1的新版本202c1-1。这些新版本可以各自使用工作者线程创建的单独事务来创建。该单独事务可以是悲观事务或乐观事务,这取决于数据库中实现的多版本并发性控制方案的类型。

[0077] 动作403的存储器紧缩过程包括对于为其创建新版本的每一有效记录将该有效记录标记为无效的子动作403c。例如,在实现多版本并发性控制方案时,记录202b1的结束时间戳可被设为创建新版本202b1-1的事务的提交时间戳。这一相同的提交时间戳还可被设为新版本202b1-1的开始时间戳以指示新版本202b1-1'的有效时间的开始。

[0078] 在实现悲观事务方案时,创建新版本的这一动作可包括等待直到针对依赖性的任何等待已经被移除才提交事务。例如,如果在更新事务在记录202b1上具有写锁定之前或同时另一事务获得记录202b1上的读标记,则更新事务被迫等待直到这些读标记被释放更新事务才能够提交(藉此使得记录202b1-1有效而记录202b1无效)。类似地,如果在更新事务的预提交阶段期间另一事务读取记录202b1-1(从而给予更新事务一提交依赖性),则更新事务被要求在更新事务提交时通知其它事务(即,递减其它事务的提交依赖性计数)。

[0079] 动作403的存储器紧缩过程包括在将页中的每一有效记录标记为无效之后将页从双向链表中移除(例如,通过垃圾收集器)的子动作403d。例如,页202b和202c可以从分区堆202的双向链表中移除。由于使用双向链表来组织页,页202a中指向列表中的下一页的指针可以被更新为指向页202d,而页202d中指向列表中的前一页的指针可被更新为指向页202a(如图2C所示)。相应地,由于页202b和202c已经被清空了所有有效记录(且由于双向链表中没有任何其它页引用这些页),页202b和202c可以通过垃圾收集被回收。

[0080] 图5示出了用于在主存储器数据库中执行存储器紧缩的另一示例性方法500的流程图。方法500将参考图2B-2D和3来描述。

[0081] 方法500包括存储器紧缩线程确定应当在主存储器数据库的分区堆中的至少一个分区堆上执行存储器紧缩的动作501。例如,MC线程301可以接收自动周期性请求,或来自用户或另一过程用于初始化存储器紧缩的按需请求。替换地,MC线程301可被配置成收集关于整个主存储器数据库(或主存储器数据库的一个或多个表)的存储器利用的统计数据,并且基于该统计数据来确定是否应当执行存储器紧缩。

[0082] 方法500包括存储器紧缩线程创建包含用于在分区堆上执行存储器紧缩的指令的工作项的动作502。例如,MC线程301可以创建包含在表201的分区堆202上执行存储器紧缩的指令的工作项310。

[0083] 方法500包括存储器紧缩线程将工作项推送到关于在分区堆上执行存储器紧缩的工作者线程的队列中的动作503。例如,MC线程301可以确定工作者线程302具有在分区堆202上执行存储器紧缩的任务并且可以将工作项310推送到关于工作者线程302的队列302a中。

[0084] 方法500包括工作者线程访问队列中的工作项以确定该工作项指令工作者线程在分区堆上执行存储器紧缩的动作504。例如,工作者线程302可以访问队列302a中的工作项310以确定工作者线程302要在分区堆202上执行存储器紧缩。

[0085] 方法500包括工作者线程访问分区堆中页的双向链表中每一页的头部中的存储器利用值的动作505。例如,工作者线程302可以访问分区堆202中的被安排为双向链表的每一页(例如,页202a、202b、202c等)的头部中的存储器利用值。

[0086] 方法500包括对于具有低于阈值的存储器利用值的每一页,工作者线程执行存储器紧缩过程以创建页中每一有效记录的新版本的动作506,该新版本被创建在分区堆的活动页中,并且将页中为其创建了新版本的每一有效记录标记为无效以藉此清空页的所有有效记录。例如,工作者线程302可以在页202b和202c上执行存储器紧缩过程,包括在活动页210a中分别创建记录202b1和202c1的新版本202b1-1和202c1-1。

[0087] 在一些实施例中,创建每一记录的新版本可以通过使用此处描述的多版本并发性控制方案的技术来创建用于创建新版本的事务来执行。

[0088] 本发明可具体化为其它具体形式而不背离其精神或本质特征。所描述的实施例在所有方面都应被认为仅是说明性而非限制性的。因此,本发明的范围由所附权利要求书而非前述描述指示。落入权利要求书的等效方案的含义和范围内的所有改变应被权利要求书的范围所涵盖。

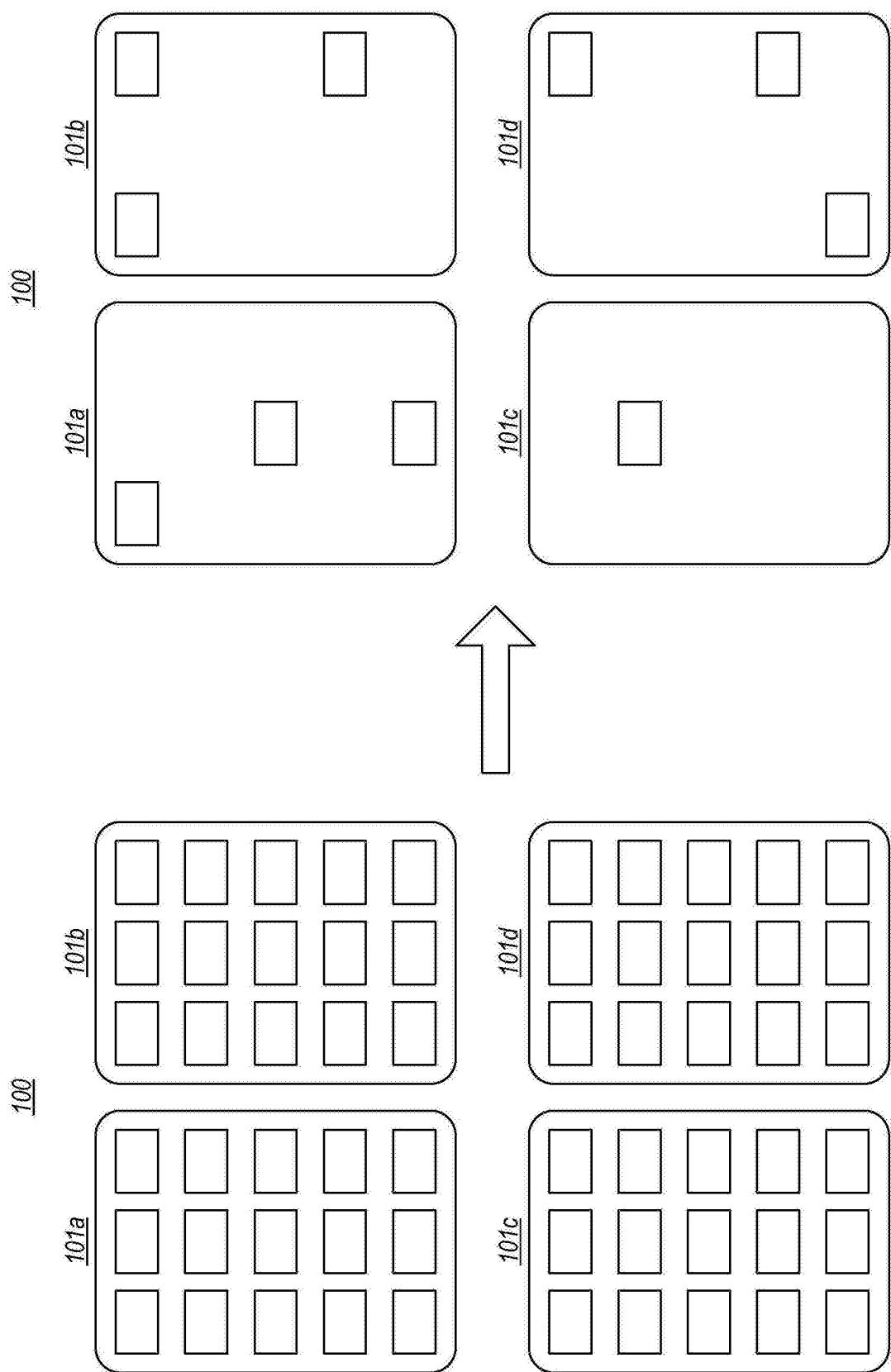


图1

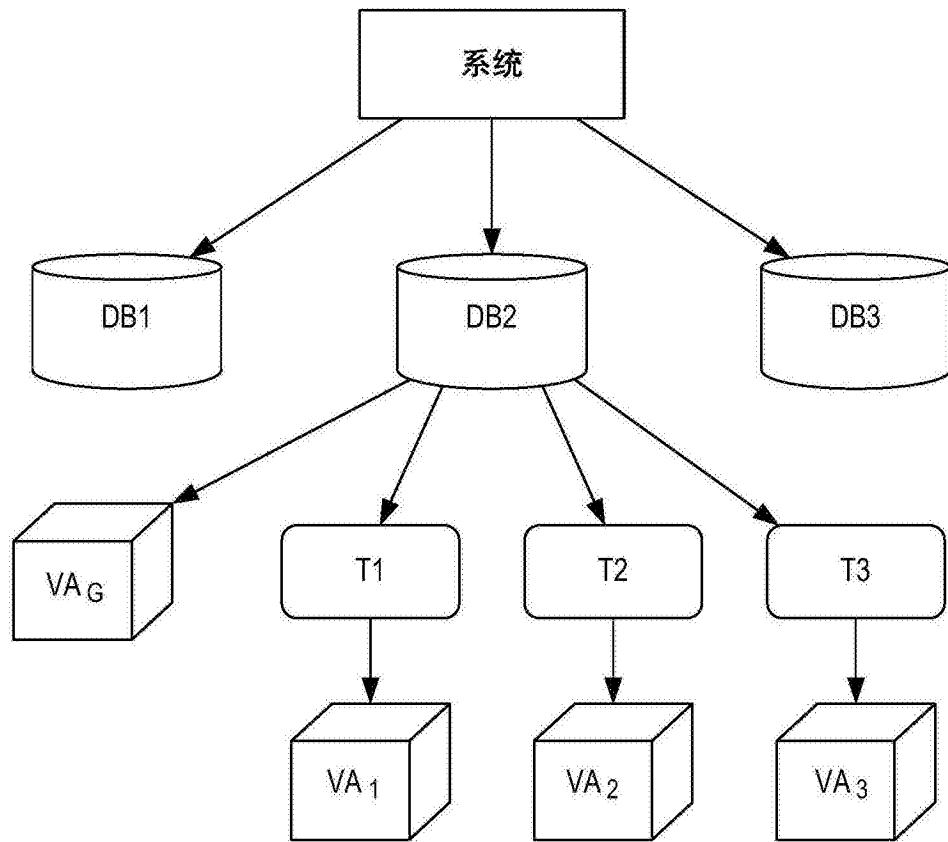


图2A

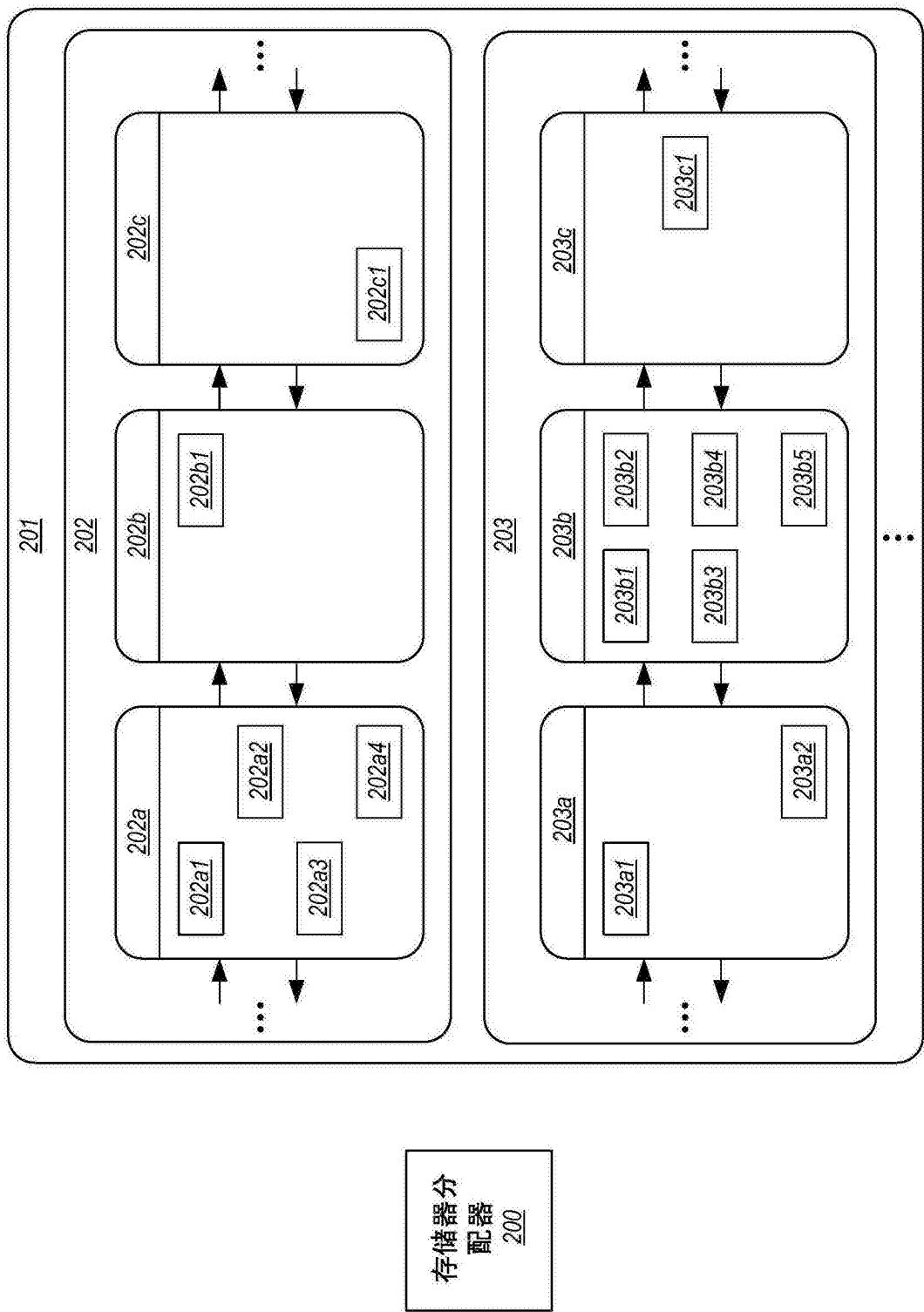


图 2B

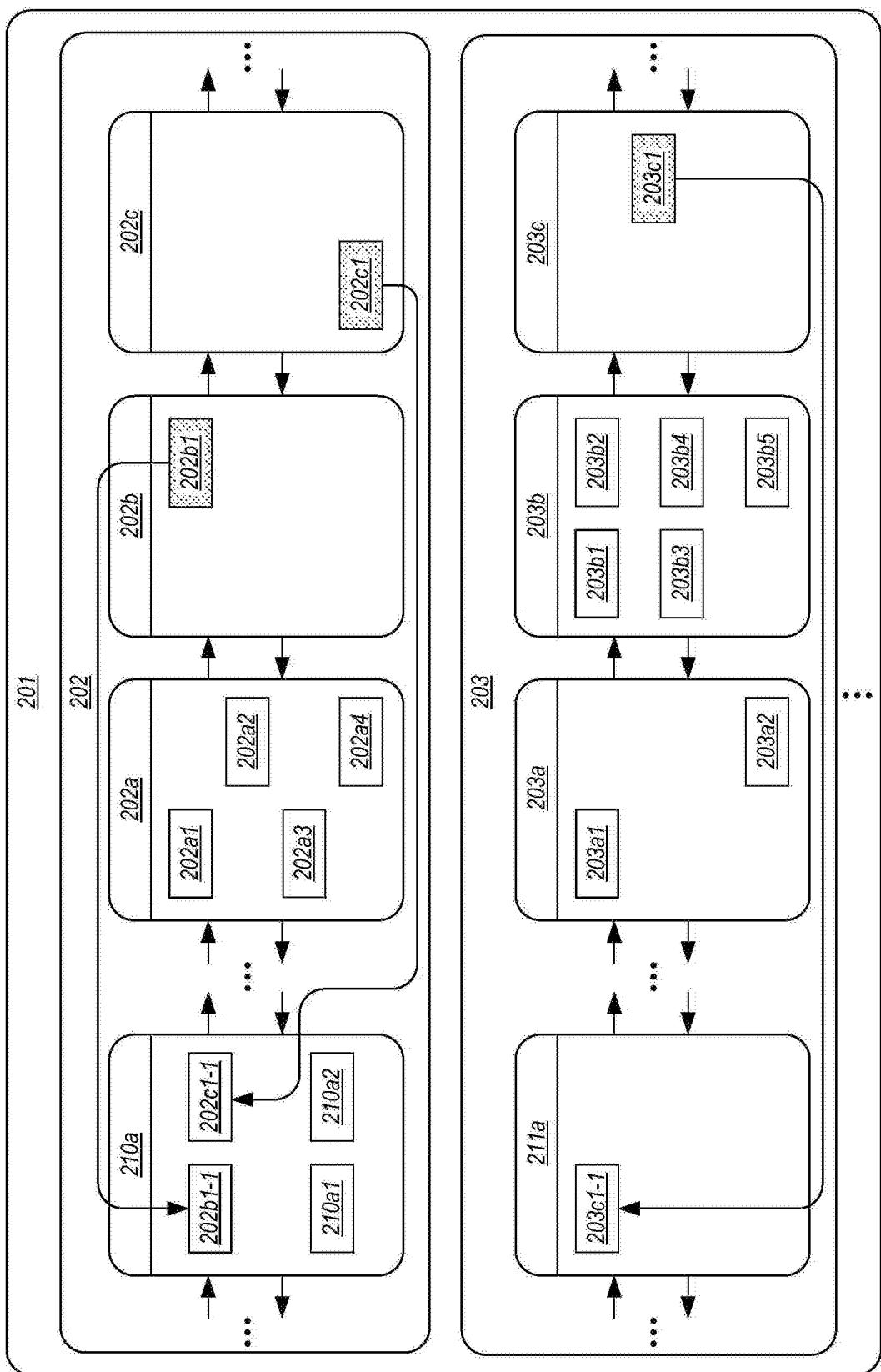


图2C

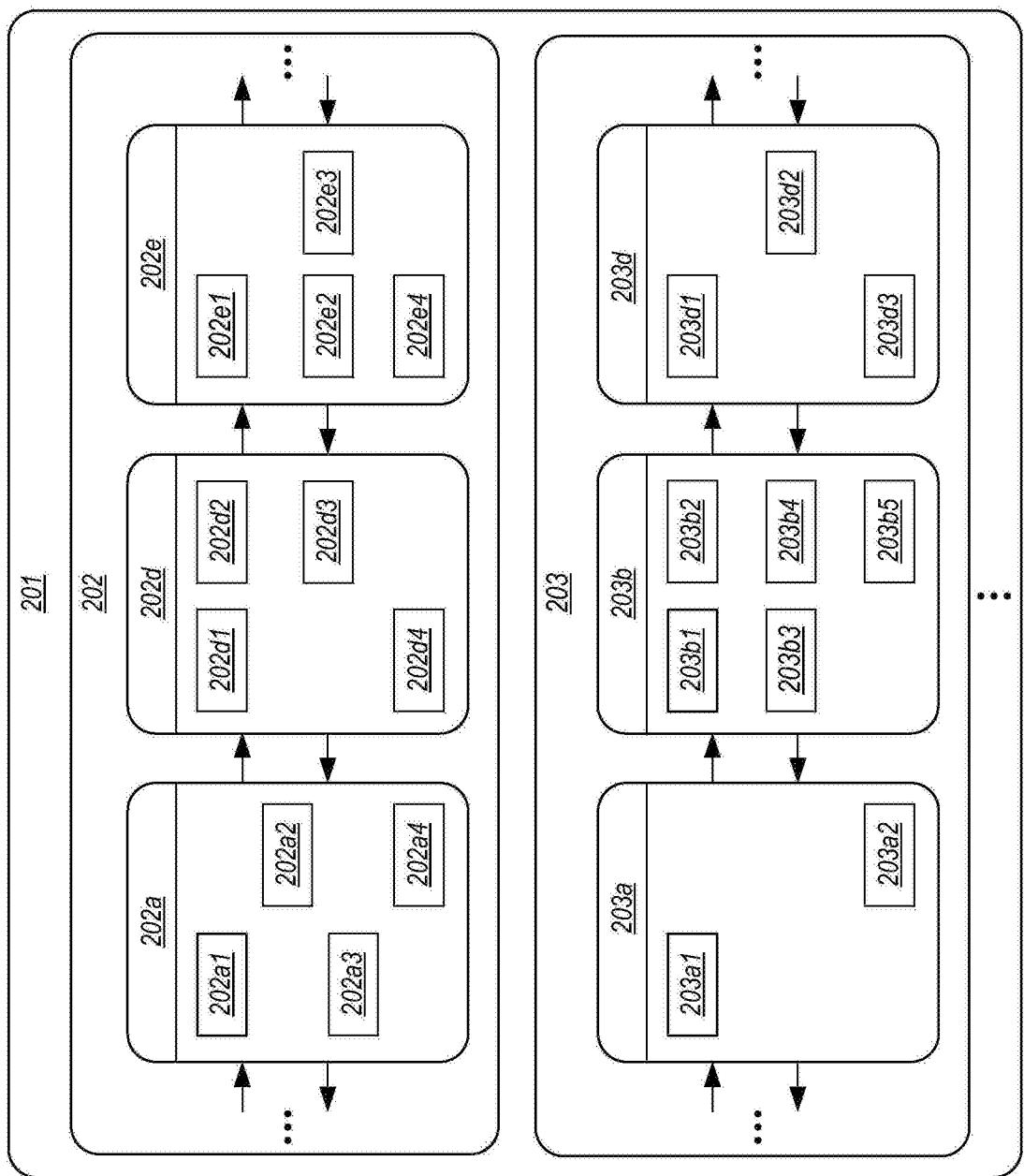


图2D

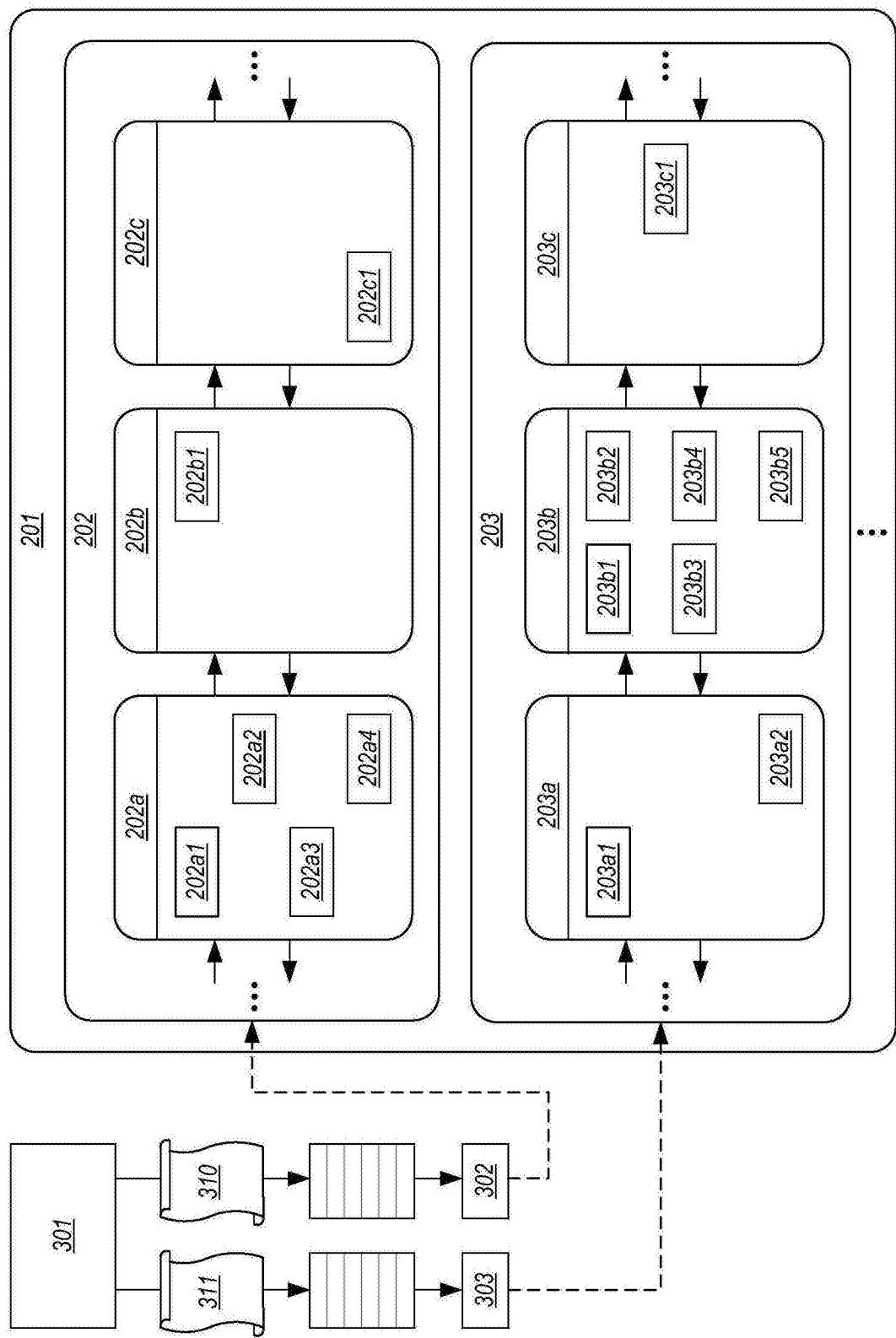


图3

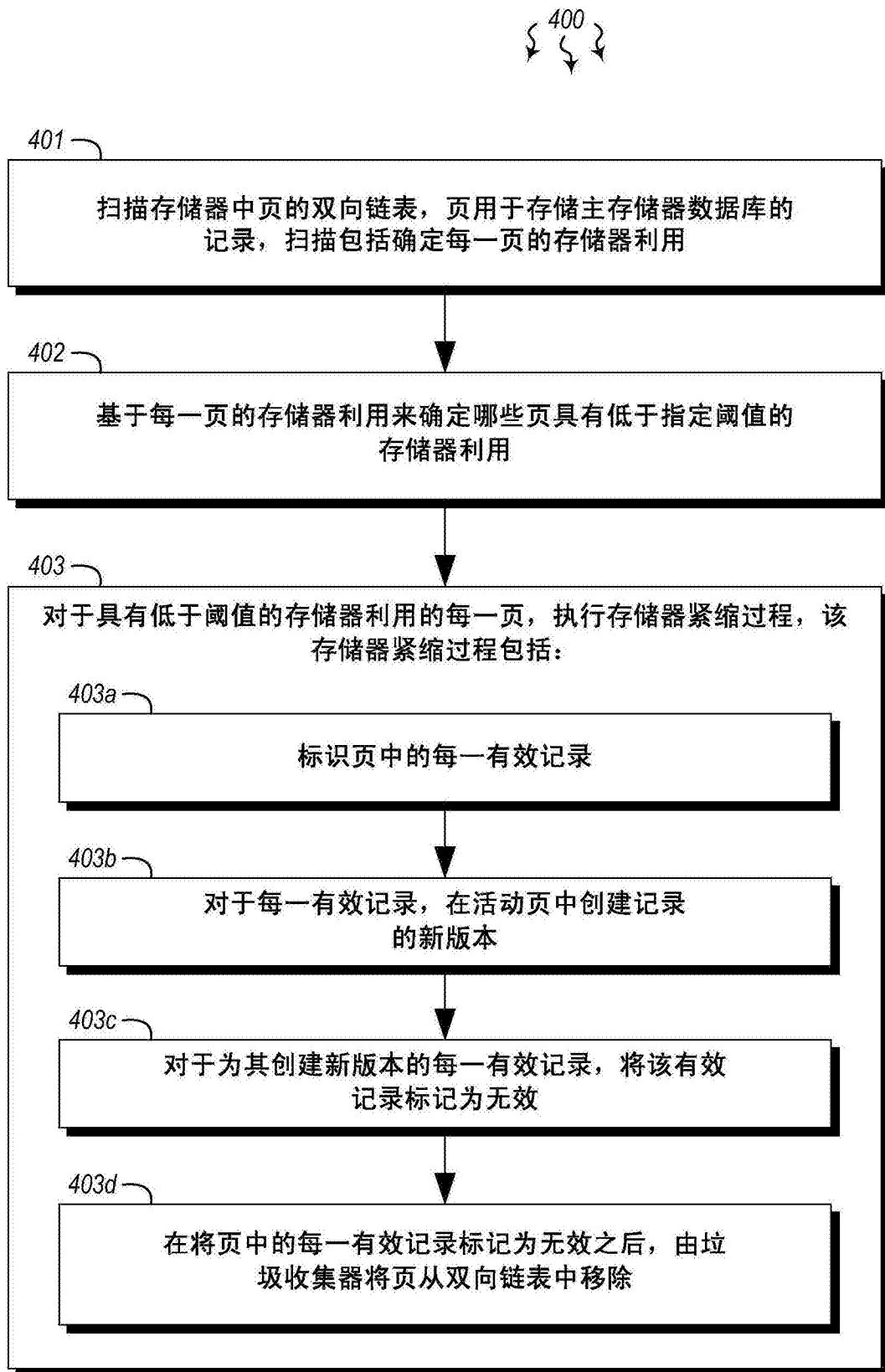


图4

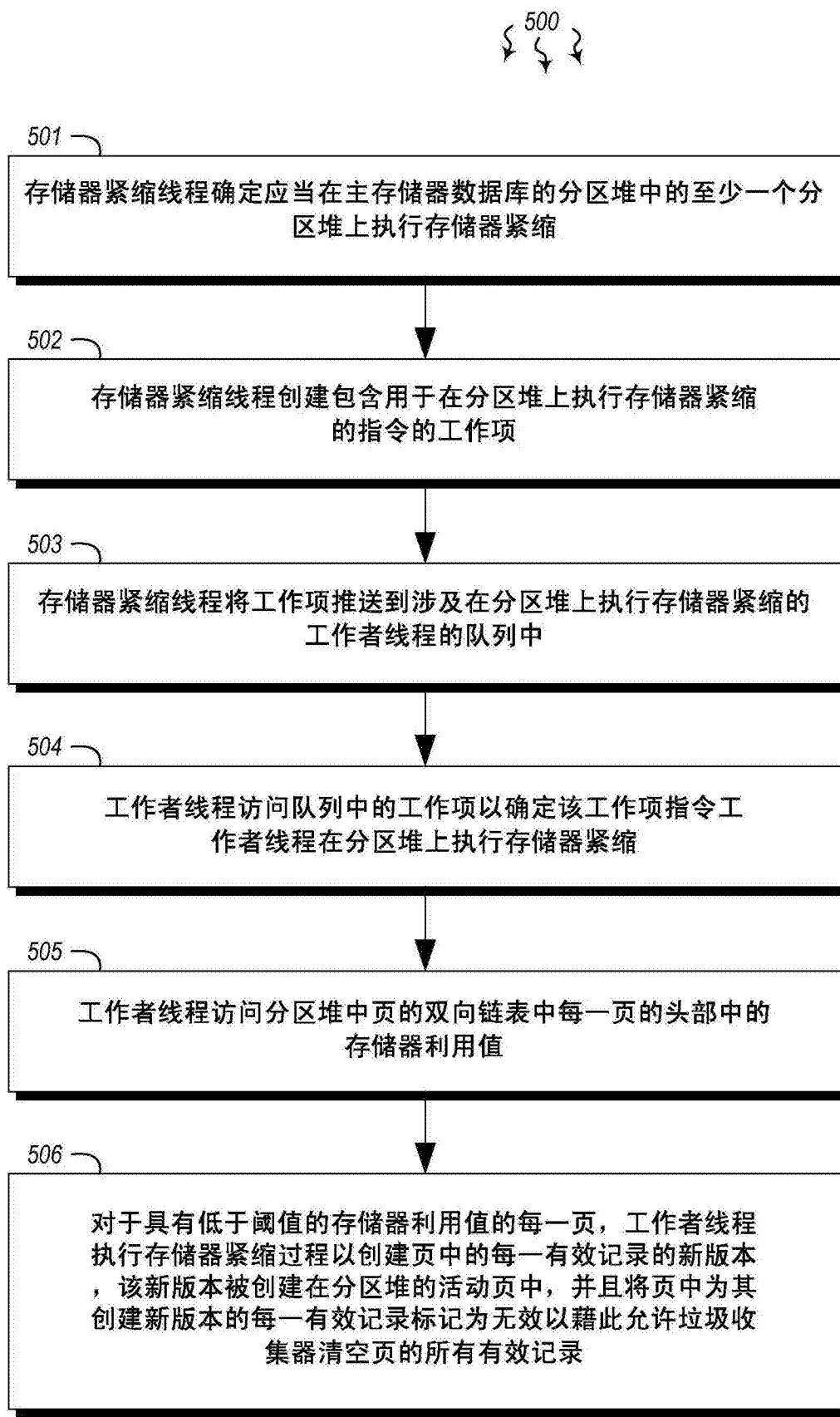


图5