(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 102782683 B
(45)授权公告日 2013.08.21

(21)申请号 201180012126.5
(22)申请日 2011.03.02
(30)优先权数据
12/717,139 2010.03.04 US
(85)PCT申请进入国家阶段日
2012.09.03
(86)PCT申请的申请数据
PCT/US2011/026930 2011.03.02
(87)PCT申请的公布数据
WO2011/109564 EN 2011.09.09
(73)专利权人 微软公司
地址 美国华盛顿州
(72)发明人 C·张 S·克里希纳穆希
G·I·雷内亚 A·韦尔比茨基
(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 顾嘉运
(51)Int.Cl.
G06F 17/30(2006.01)
G06F 12/00(2006.01)
审查员 徐波

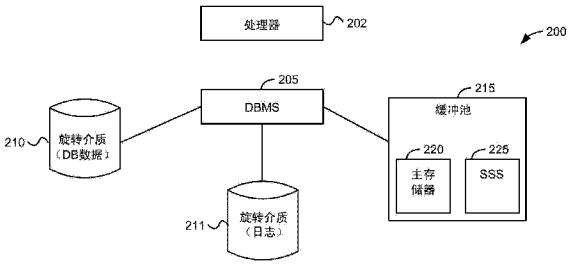
权利要求书2页 说明书9页 附图5页

(54)发明名称

用于数据库服务器的缓冲池扩展

(57)摘要

此处所描述的主题的各方面涉及用于数据库系统的缓冲池。在各方面，诸如固态存储的次存储器被用于扩展数据库系统的缓冲池。可以通过采样算法来确定诸如火热、热门和冷门的阈值，所述阈值用于基于页面的访问历史来分类页面。当数据库系统需要释放主存储器中的缓冲池中的空间时，可以基于该页面被如何分类以及次存储器或其他存储的条件来将一个页面驱逐到次存储器中的缓冲池或其他存储。



CN 102782683 B

1. 一种至少部分地由计算机实现的方法，所述方法包括：

访问数据库缓冲池的页面的元数据，所述数据库缓冲池包括被存储在主存储器中的页面、存储在固态存储器中的页面以及存储在主存储器和固态存储器两者之中的页面；

从所述元数据中确定用于确定页面是否是冷门的冷门阈值，其中如果将一个函数应用于所述页面的元数据的访问数据返回的是小于或等于所述冷门阈值的值，则所述页面是冷门的；

从所述元数据中确定用于确定页面是否是热门的热门阈值，其中如果将所述函数应用于所述页面的元数据的访问数据返回的是大于所述冷门阈值且小于或等于所述热门阈值的值，则所述页面是热门的；以及

将页面从所述主存储器或所述固态存储器驱逐出去以为一个或多个其他页面释放空间；

其中所述将页面驱逐出所述固态存储器包括将多个页面在单个写入操作中从所述主存储器复制到所述固态存储器。

2. 如权利要求 1 所述的方法，其特征在于，其中访问数据库缓冲池的页面的元数据包括对从所述数据库缓冲池中随机选择的小于所述数据库缓冲池的所有页面的数目的页面的元数据进行采样。

3. 如权利要求 2 所述的方法，其特征在于，其中访问所述数据库缓冲池的页面的元数据包括从所述页面的元数据中为每个经采样的页面获得一个或两个时间戳，其中所述一个或两个时间戳对应于访问经采样的页面的最近一个或多个时间，对至少所述一个或两个时间戳施加所述函数以生成指示所述页面是火热、热门或冷门的值。

4. 如权利要求 1 所述的方法，其特征在于，其中驱逐页面包括如果所述固态存储器具有可用的空闲空间并且如果在所述固态存储器中不存在所述冷门页面的副本，则将所述冷门页面复制到所述固态存储器中。

5. 如权利要求 1 所述的方法，其特征在于，进一步包括改变用于所述数据库缓冲池的所述固态存储的量。

6. 如权利要求 1 所述的方法，其特征在于，还包括确定是否已经达到所述固态存储的 I/O 阈值，并且如果达到，则将所述页面复制到除所述主存储器和所述固态存储器之外的存储上。

7. 一种在计算环境中的系统，包括：

一个或多个存储设备集，包括在操作上移动以提供对所述存储设备的数据的访问的组件；

主存储器，是易失的且操作上无需所述主存储器的任意组件的物理移动就能提供对其上存储的数据的访问；

次存储器，是非易失的且操作上无需所述次存储器的任意组件的物理移动就能提供对其上存储的数据的访问；

一个或多个处理器，在操作上执行对应于数据库管理系统的指令，所述数据库管理系统在操作上管理所述主存储器和次存储器中的缓冲池中的页面，并基于确定页面是否具有对应于至少三种分类的访问来驱逐所述页面，如果对所述页面的访问在第一百分比范围之内则所述页面具有对应于第一分类的访问，如果对所述页面的访问在第二百分比范围之内

则所述页面具有对应于第二分类的访问，如果对所述页面的访问在第三百分比范围之内则所述页面具有对应于第三分类的访问；

其中基于确定页面是否具有对应于至少三种分类的访问来驱逐所述页面包括将多个所述页面在单个写入操作中从所述主存储器复制到所述次存储器。

8. 如权利要求 7 所述的系统，其特征在于，所述一个或多个存储设备集包括一个或多个硬盘，所述主存储器包括随机存取存储器，而所述次存储器包括固态存储器。

9. 如权利要求 7 所述的系统，其特征在于，所述数据库管理系统在操作上还扫描所述缓冲池的随机选择的页面的元数据以确定所述第一、第二和第三范围，所述元数据包括指示每个页面被访问的最近一个或多个时间的数据。

10. 如权利要求 7 所述的系统，其特征在于，所述数据库管理系统在操作上驱逐页面包括：所述数据库管理系统在操作上将被确定为处于第二百分比范围内的页面从所述主存储器驱逐到所述次存储器上。

11. 如权利要求 7 所述的系统，其特征在于，所述数据库管理系统在操作上驱逐页面包括：在将所选中的页面复制到所述存储设备中的一个或多个之前，所述数据库管理系统在操作上将被确定为处于第三百分比范围内的所选中的页面从所述次存储器复制到所述主存储器上。

12. 如权利要求 7 所述的系统，其特征在于，其中所述第一百分比范围包括百分之 0- 百分之 5，所述第二百分比范围包括百分之 5- 百分之 25，而所述第三百分比范围包括百分之 25- 百分之 100。

13. 一种至少部分地由计算机实现的方法，包括：

接收对数据库数据的请求；

确定所述数据驻留在除主存储器之外的存储中；

确定所述主存储器中的缓冲池是满的；

基于一函数选择要从主存储器中的所述缓冲池中驱逐出的候选页面，所述函数根据基于对所述候选页面的访问的至少三种分类中的一个对所述候选页面分类；以及

将所述候选页面从主存储器中驱逐出去；

其中，将所述候选页面从主存储器中驱逐出去包括将多个所述候选页面在单个写入操作中从所述主存储器复制到固态存储器。

14. 如权利要求 13 所述的方法，其特征在于，选择要从主存储器中的所述缓冲池中驱逐出的候选页面包括查找处于所述分类中的第三种分类中的页面，所述分类中的第三种分类表示与其他分类中任一分类中的页面相比较少频率地被访问的页面。

用于数据库服务器的缓冲池扩展

背景技术

[0001] 对于数据库，当数据库页面的工作集被保持在主存储器中时，存在良好的性能。不幸地是，由于许多大型数据库的大小原因，要将工作集保持在主存储器中并不是切实可行的。对于一个大型数据库而言，这意味着该数据库的大多数数据页面驻留在具有足够容量的 I/O 子系统上，所述子系统通常是使用诸如磁盘的旋转介质来构建。这样的子系统是昂贵的，会占据大量的空间并消耗大量的能耗。这些子系统经常成为数据库的瓶颈，因为在相同速率时，旋转介质的性能与主存储器和处理器相比并不领先。

[0002] 在此要求保护的主题不限于解决任何缺点或仅在诸如上述环境中操作的各个实施例。相反，提供该背景仅用以示出在其中可实践在此描述的部分实施例的一个示例性技术领域。

发明内容

[0003] 简言之，此处所描述的主题的各方面涉及用于数据库系统的缓冲池。在各方面，诸如固态存储的次存储器被用于扩展数据库系统的缓冲池。可以通过采样算法来确定诸如火热、热门和冷门的阈值，所述阈值用于基于页面的访问历史来分类页面。当数据库系统需要释放主存储器中的缓冲池中的空间时，可以基于该页面被如何分类以及次存储器或其他存储的条件来将一个页面驱逐到次存储器中的缓冲池或其他存储。

[0004] 提供本发明内容是为了简要地标识在以下详细描述中进一步描述的主题的一些方面。本发明内容并不旨在标识出所要求保护的主题的关键特征或必要特征，也不旨在用于限制所要求保护的主题的范围。

[0005] 除非上下文清楚地指出，否则短语“此处所描述的主题”指的是具体实施方式中描述的主题。术语“方面”被当作“至少一个方面”。标识具体实施方式中所描述的主题的各方面不旨在标识所要求保护的主题的关键特征或必要特征。

[0006] 上述各方面和此处所描述的主题的其它方面是借助于示例说明的，并且不受附图限制，附图中相同的标号指示相似的元素。

附图说明

[0007] 图 1 是表示其中可结合本文所描述主题的各方面的示例性通用计算环境的框图；

[0008] 图 2 是概括地表示此处所描述的主题的各方面可以在其中实现的示例性系统的框图；

[0009] 图 3 是表示根据此处所描述的主题的各方面的托管数据库的系统的组件的框图；以及

[0010] 图 4 是概括地表示根据此处所描述的主题的各方面的、可在扫描页面以确定阈值且适合时驱逐页面中发生的一些示例性动作的流程图；以及

[0011] 图 5 是概括地表示根据此处所描述的主题的各方面的、可在由 DBMS 接收访问请求且缓冲池已满时发生的一些示例性动作的流程图。

具体实施方式

[0012] 定义

[0013] 如本文所使用的，术语“包括”及其变体被当作开放式术语，表示“包括但不限于”。除非上下文另外清楚地指示出，否则术语“或”被当作“和 / 或”。术语“基于”被当作“至少部分地基于”。术语“一个实施例”和“一实施例”被当作“至少一个实施例”。术语“另一实施例”被当作“至少一个其他实施例”。其他显式或隐式定义可包括在下文中。

[0014] 示例性操作环境

[0015] 图 1 示出可在其上实现本文所描述的主题的各方面的合适的计算系统环境 100 的示例。计算系统环境 100 仅为合适的计算环境的一个示例，并非旨在对本文所描述的主题的各方面的使用范围或功能提出任何限制。也不应该将计算环境 100 解释为对示例性操作环境 100 中示出的任一组件或其组合有任何依赖性或要求。

[0016] 本文所描述的主题的各方面可与众多其他通用或专用计算系统环境或配置一起操作。可适用于这里所述的主题的各方面的已知计算系统、环境或配置的例子包括个人计算机、服务器计算机、手持或膝上型设备、多处理器系统、基于微控制器的系统、机顶盒、可编程消费电子设备、网络 PC、微型计算机、大型计算机、个人数字助理（PDA）、游戏设备、打印机、包括机顶盒，媒体中心或其他家电的家电设备、嵌入汽车或附加到汽车的计算设备、其他移动设备、包括任何上述系统或设备的分布式计算环境等等。

[0017] 本文所描述的主题的各方面可在由计算机执行的诸如程序模块等计算机可执行指令的一般上下文中描述。一般而言，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。本文所描述的主题的各方面也可以在其中任务由通过通信网络链接的远程处理设备执行的分布式计算环境中实现。在分布式计算环境中，程序模块可以位于包括存储器存储设备在内的本地和远程计算机存储介质中。

[0018] 参考图 1，用于实现本文所描述的主题的各方面的示例性系统包括计算机 110 形式的通用计算设备。计算机可包括能够执行指令的任何电子设备。计算机 110 的组件可包括处理单元 120、系统存储器 130 以及将包括系统存储器的各种系统组件耦合至处理单元 120 的系统总线 121。系统总线 121 可以是若干类型的总线结构中的任一种，包括使用各种总线体系结构中的任一种的存储器总线或存储器控制器、外围总线、以及局部总线。作为示例，而非限制，这样的架构包括工业标准架构（ISA）总线、微通道架构（MCA）总线、增强型 ISA（EISA）总线、视频电子技术标准协会（VESA）局部总线、也称为夹层（Mezzanine）总线的外围部件互连（PCI）总线、扩展外围部件互连（PCI-X）总线、高级图形端口（AGP）、以及 PCI Express（PCIe）。

[0019] 计算机 110 通常包括各种计算机可读介质。计算机可读介质可以是能由计算机 110 访问的任何可用介质，并包含易失性和非易失性介质以及可移动、不可移动介质。作为示例而非限制，计算机可读介质可包括计算机存储介质和通信介质。

[0020] 计算机存储介质包括以用于存储诸如计算机可读指令、数据结构、程序模块或其它数据等信息的任何方法或技术来实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括 RAM、ROM、EEPROM、闪存或其它存储器技术 CD-ROM、数字多功能盘（DVD）或其它光盘存储、磁盒、磁带、磁盘存储或其它磁存储设备、或可以用来储存所期望的信息

并可由计算机 110 访问的任一其它介质。

[0021] 通信介质通常以诸如载波或其他传输机制等已调制数据信号来体现计算机可读指令、数据结构、程序模块或其他数据，并包括任何信息传送介质。术语已调制数据信号摂是指具有以在信号中编码信息的方式被设定或改变其一个或多个特征的信号。作为示例而非限制，通信介质包括诸如有线网络或直接线连接之类的有线介质，以及诸如声学、RF、红外及其他无线介质之类的无线介质。上述的任意组合也应包含在计算机可读介质的范围内。

[0022] 系统存储器 130 包括易失性和 / 或非易失性存储器形式的计算机存储介质，如只读存储器（ROM）131 和随机存取存储器（RAM）132。包含诸如在启动期间帮助在计算机 110 内的元件之间传输信息的基本例程的基本输入 / 输出系统 133（BIOS）通常储存储在 ROM 131 中。RAM 132 通常包含处理单元 120 可立即访问和 / 或当前正在操作的数据和 / 或程序模块。作为示例而非限制，图 1 示出了操作系统 134、应用程序 135、其它程序模块 136 和程序数据 137。

[0023] 计算机 110 也可以包括其他可移动 / 不可移动、易失性 / 非易失性计算机存储介质。仅作为示例，图 1 示出了从不可移动非易失性磁介质中读取或向其写入的硬盘驱动器 141，从可移动非易失性磁盘 152 中读取或向其写入的磁盘驱动器 151，以及从诸如 CD ROM 或其它光学介质等可移动非易失性光盘 156 中读取或向其写入的光盘驱动器 155。可以在该示例性操作环境中使用的其他可移动 / 不可移动、易失性 / 非易失性计算机存储介质包括磁带盒、闪存卡、数字多功能盘、其他光盘、数字录像带、固态 RAM、固态 ROM 等等。硬盘驱动器 141 通常通过诸如接口 140 等不可移动存储器接口连接到系统总线 121，而磁盘驱动器 151 和光盘驱动器 155 则通常由诸如接口 150 等可移动存储器接口连接至系统总线 121。

[0024] 以上讨论并在图 1 中示出的驱动器及其相关联的计算机存储介质为计算机 110 提供了对计算机可读指令、数据结构、程序模块和其它数据的存储。例如，在图 1 中，硬盘驱动器 141 被示为存储操作系统 144、应用程序 145、其它程序模块 146 和程序数据 147。注意，这些组件可与操作系统 134、应用程序 135、其他程序模块 136 和程序数据 137 相同，也可与它们不同。操作系统 144、应用程序 145、其他程序模块 146 和程序数据 147 在这里被标注了不同的附图标记是为了说明至少它们是不同的副本。

[0025] 用户可以通过输入设备，如键盘 162 和定点设备 161（通常被称为鼠标、跟踪球或触摸垫）向计算机 110 输入命令和信息。其它输入设备（未示出）可包括话筒、操纵杆、游戏手柄、圆盘式卫星天线、扫描仪、触敏屏、写字板等。这些以及其他输入设备通常通过耦合到系统总线的用户输入接口 160 连接到处理单元 120，但也可通过诸如并行端口、游戏端口或通用串行总线（USB）之类的其他接口和总线结构来连接。

[0026] 监视器 191 或其他类型的显示设备也通过诸如视频接口 190 之类的接口连接至系统总线 121。除了监视器以外，计算机还可包括诸如扬声器 197 和打印机 196 之类的其他外围输出设备，它们可通过输出外围接口 195 来连接。

[0027] 计算机 110 可使用到一个或多个远程计算机（诸如，远程计算机 180）的逻辑连接而在联网环境中操作。远程计算机 180 可以是个人计算机、服务器、路由器、网络 PC、对等设备或其它常见网络节点，且通常包括上文相对于计算机 110 描述的许多或所有元件，但在图 1 中只示出存储器存储设备 181。图 1 中所示的逻辑连接包括局域网（LAN）171 和广域

网(WAN)173,但也可以包括其它网络。这样的联网环境常见于办公室、企业范围计算机网络、内联网和因特网中。

[0028] 当在 LAN 联网环境中使用时,计算机 110 通过网络接口或适配器 170 连接到 LAN 171。当在 WAN 联网环境中使用时,计算机 110 可包括调制解调器 172 或用于通过诸如因特网等的 WAN 173 来建立通信的其它装置。可为内置或可为外置的调制解调器 172 可以经由用户输入接口 160 或其他合适的机构连接至系统总线 121。在联网环境中,相对于计算机 110 所示的程序模块或其部分可被存储在远程存储器存储设备中。作为示例而非限制,图 1 示出了远程应用程序 185 驻留在存储器设备 181 上。应当理解,所示的网络连接是示例性的,并且可使用在计算机之间建立通信链路的其他手段。

[0029] 数据库缓冲池

[0030] 如前所述,I/O 子系统通常成为数据库的瓶颈。图 2 是概括地表示此处所描述的主题的各方面可以在其中实现的示例性系统的框图。系统 200 可以包括一个或多个处理器 202、数据库管理系统 (DBMS)205、旋转介质 210-211,并可以包括其他组件。DBMS 205 可以管理缓冲池 215 中的页面。可以在主存储器 220 和固态存储 225(在此以后有时被称为 SSS 225)中维护缓冲池 215。

[0031] 可以使用诸如计算机 110 的一个或多个计算机来实现系统 200,而系统 200 可以实现一个或多个数据库。该系统的(多个)处理器 202 对应于图 1 的处理单元 120,并且被包括在同一计算机上或可以跨多个计算机分布。处理器 202 执行对应于 DBMS 205 的指令以提供数据库。

[0032] 数据库可包括关系数据库、面向对象的数据库、分层数据库、网络数据库、其他类型的数据库、上述的某种组合或扩展等。可将存储于数据库中的数据组织为表格、记录、对象、其他数据结构等。可将存储于数据库中的数据存储在专用数据库文件、专用硬盘分区、HTML 文件、XML 文件、电子数据表、平面文件、文档文件、配置文件、其他文件等中。

[0033] 可以通过诸如 DBMS 205 的 DBMS 来访问数据库中的数据。DBMS 205 可包括一个或多个程序,其控制数据库的数据的组织、存储、管理和检索。DBMS205 可接收访问数据库中的数据的请求,并可执行提供该访问所需要的操作。此处使用的访问可包括读取数据、写入数据、删除数据、更新数据、以及包括以上两个或更多个的组合等。

[0034] 在描述本文所描述的主题的各方面的时,为简洁起见,本文有时候使用与关系数据库相关联的术语。尽管本文有时候使用关系数据库术语,但也可将本文的技术应用到其它类型的数据库,包括之前已经提到的那些。

[0035] DBMS 205 在操作上管理主存储器和次存储器中的缓冲池中的页面,并基于确定一个页面是否具有对应于至少三种分类的访问来驱逐该页面。这将在下面被更详细地描述,但简言之,如果对该页面的访问位于第一百分比范围内,则该页面可被认作处于第一分类中(例如“火热”)。如果该页面具有位于第二百分比范围内的访问,则该页面可被认作处于第二分类中(例如“热门”)。如果该页面具有位于第三百分比范围内的访问,则该页面可被认作处于第三分类中(例如“冷门”)。页面的分类可被用于将页面驱逐出主存储器 220 和 / 或 SSS 225。

[0036] 处于效率、性能或其他原因,数据库可以将对应于表格、索引的数据或其他数据库数据定位在一个诸如旋转介质 210 的存储设备集上,而将对应于日志的数据定位在诸如旋

转介质 211 的另一存储设备集上。对于某些数据库，仅有一个存储设备集可被用于存储表格和日志。

[0037] 缓冲池 215 包括在(相对于旋转介质 210-211)更快的存储器中的一个或多个数据库的部分。所述一个或多个数据库的部分有时称为页面。缓冲池 215 的存储器(例如主存储器 220)的一些可以是易失的。就是说，当存储器断电时，存储器会丢失存储在其上的任何数据。主存储器 220 可以被实现为 RAM、高速缓存、处理器存储器或其他易失的高速存储器。缓冲池 215 的其他存储器(例如 SSS 225)可以是非易失的。就是说，当非易失的存储器断电时，该存储器会持久保存存储在其上的任何数据。

[0038] 可以动态改变分配给缓冲池 215 的 SSS (或其他存储器)的量。就是说，分配给缓冲池 215 的 SSS(或其他存储器)的字节在 DBMS 205 执行之前、期间或之后，可以被自动地、半自动地或手动地增加、减少或设定为 0。

[0039] DBMS 205 可以维护关于存储在缓冲池 215 中的数据库的页面的元数据。对于每个页面，该元数据可包括，例如，用于存储页面上的数据被访问的最近的一个或两个时间的两个时间戳、指示该页面是否已被修改的脏标记、指示该页面的副本是否被存储在 SSS 225 中的标记、指示该页面已经被访问的次数的计数、指示对该页面的访问的频率的加权值(对于更近的访问具有更多加权)、关于该页面的其他信息等等。该元数据可以被(例如作为页面)存储在缓冲池 215 中或某个其他位置处。

[0040] 在操作中，当 DBMS 205 接收访问数据的请求时，DBMS 205 可以先确定该数据是否在缓冲池 215 中。如果该数据不在缓冲池 215 中，则 DBMS 可以确定在主存储器 220 中是否存在将数据从旋转介质 210 读入该主存储器 220 的空间。如果主存储器 220 中不存在空间，则 DBMS 确定是否可以将一个页面从主存储器 220 中驱逐出以腾出空间来从旋转介质 210 中读取该页面。

[0041] 虽然上述步骤已经以某种顺序方式被描述，在其他实现中，它们可以以不同的顺序或并行地发生。例如，在一个实施例中，DBMS 205 可以尝试维护主存储器 220 中的某个空闲空间部分。为了这么做，周期性地或当 DBMS 205 确定主存储器 220 中的空闲空间减少到阈值之下时，DBMS 205 可以扫描主存储器 220 以确定将哪个页面驱逐出主存储器 220 以释放空间。以此方式，DBMS205 可以保留主存储器 220 中的某些空间以用于在从旋转介质 210 读取和写入页面时使用。

[0042] 在一个实施例中，当存在存储器压力(例如缓冲池中的空闲空间低于预定的、计算的、选择的或其他阈值，进程请求更多的存储器或满足某些其他存储器压力条件)时，DBMS 205 可以执行一个扫描算法。该扫描算法可以确定可被用于将页面分类成不同类别的阈值的数值。例如，在一个实施例中，扫描算法可以基于关于页面而维护的元数据来确定将页面分类成火热、热门和冷门类别的阈值。

[0043] 在一个实施例中，扫描算法可以通过采样选中数目的随机页面的元数据来确定阈值。从所述采样中，可以确定至少两个阈值。阈值可以对应于关于对页面的访问频率的百分比。具有低于两个阈值中最低阈值的访问特性的页面可以被认为是冷门的。具有在最低阈值和较高阈值之间的访问特性的页面可以被认为是热门的。具有高于较高阈值的访问特性的页面可以被认为是火热的。

[0044] 术语“火热”、“热门”和“冷门”并不意味着要对在此所述的主题的各方面施加限

制。这些术语的一个理念在于对页面的访问可以落入百分比中。落入较高百分比范围之内的页面可以被称为是火热的。落入较低和较高百分比之间的页面可以被称为是热门的。落入较低百分比范围之内的页面可以被称为是冷门的。在不背离在此所述的主题的各方面的精神和范围的情况下，可以用其他词、数字、标识符、数据结构等等来取代火热、热门和冷门这些词。而且，虽然仅给出了三种称号，在其他实施例中，可以存在超过三种的称号。

[0045] 一种可以用于分类页面的示范性标准是对页面的访问的频率。另一种可以用于分类页面的示范性标准是对页面的访问历史。例如，访问历史可以包括页面已经被访问的最近一个或两个时间。如果仅最近访问时间被用于分类页面，这种标准有时被称作最近使用(LRU)算法。如果最近访问时间的邻近时间被用于分类页面，这种标准有时被称作(LRU-2)算法。上述例子并不旨在是包括一切的或是穷举的。实际上，基于在此的示教，本领域中的技术人员可以认识到许多其他标准可以被用于分类页面。

[0046] 在一个实施例中，仅当确定较低阈值的采样可以在主存储器 220 和 SSS 225 中的页面上执行时，才在主存储器 220 中的页面上执行确定较高阈值的采样。

[0047] 在确定了这些阈值之后，当 DBMS 205 需要驱逐页面以在主存储器 220 中空出更多空间时，DBMS 205 可以以任意各种顺序，例如包括顺序、循环、随机、最近使用、基于位置、另一顺序等等，来扫描遍历缓冲池中的页面。如果页面在冷门阈值之上且在热门阈值之下，并且 SSS 225 中存在可用的空间，则 DBMS 205 可以将该页面复制到 SSS 225 或(通过向诸如页面队列的数据结构中的要被从主存储器 220 复制到 SSS 225 的页面放置指针)指示只要可行的话该页面就要被复制到 SSS 225 中。在已经将页面复制到 SSS 225 之后，主存储器 220 中释放的空间可以被用于存储另一页面。

[0048] 如果在 SSS 225 中不存在用于热门和冷门页面的足够空间，则冷门的页面将被转储清除(flush)或复制在队列中以转储清除到磁盘(如果是脏的)或标记为可用(如果是干净的)。转储清除到磁盘的页面可以来自主存储器 220 以及 SSS 225。在这种情况下，在一个实施例中，与从 SSS 225 转储清除冷门页面相比，将优先权给予从主存储器 220 转储清除冷门页面。例如，可以为从主存储器 220 和 SSS 225 转储清除冷门页面维护一个或多个数据结构(例如一个或多个队列)。在已经将来自主存储器 220 的冷门页面转储清除到磁盘之后，可以将来自 SSS 225 的冷门页面转储清除到磁盘。在另一个实施例中，从主存储器 220 转储清除冷门页面可以与从 SSS 225 转储清除冷门页面交替进行。如果硬件子系统提供了适合的设备，可以并行执行从主存储器 220 转储清除冷门页面和从 SSS 225 转储清除冷门页面。

[0049] 当合适的硬件(例如直接存储器存取(DMA)硬件等)可用时，将来自 SSS 225 的页面转储清除到磁盘可以在不需要从 SSS 225 读取页面到主存储器 220 的情况下进行。当这样的硬件对于 SSS 225 不可用时，将来自 SSS 225 的页面转储清除到磁盘可以通过读取页面到主存储器 220 并随后将该页面从主存储器 220 复制到磁盘上来执行。

[0050] 当 DBMS 205 需要访问在 SSS 225 中但不在主存储器 220 中的页面时，可以从 SSS 225 将该页面复制到主存储器 220 中。

[0051] 当缓冲页面要被写入 SSS 225 时，如果可能的话，可以将多个写入合并成单个写入。这可以增加 I/O 吞吐量并增加 SSS 225 的寿命期望。

[0052] 如果达到了 SSS 225 的 I/O 阈值，则可将新的 I/O 定向到磁盘。当对 SSS 225 的

访问接近 SSS 225 的读 / 写带宽或某个其他预定带宽时可以达到 I/O 阈值。可以例如依据 I/O 响应时间、I/O 的数目或某个其他因素来确定是否已经达到 I/O 阈值。例如，通过访问的某种模式或频率，传输从 SSS 225 到 RAM 220 的数据的 I/O 可以超过 SSS 225 可用的带宽。在这种情况下，可以将后续 I/O 发送给旋转介质 210 直到 SSS225 可用于更多 I/O。

[0053] 图 3 是表示根据此处所描述的主题的各方面的托管数据库的系统的组件的框图。组件 300 包括元数据 305、主存储器缓冲页面 310、SSS 缓冲页面 315 和其他存储 320。元数据 305 包括关于先前描述的页面的数据。元数据 305 可以被存储在主存储器、高速缓存或某种其他高速存储器中。元数据 305 可以指示页面被存储在主存储器和 / 或 SSS 中。

[0054] 主存储器缓冲页面 310 可以被存储在诸如 RAM 或其他易失存储器之类的主存储器中。这样的存储器可以在不需要机械移动的情况下被访问。换句话说，这样的存储器可以在不需要主存储器的任何组件的物理移动的情况下提供对数据的访问。这样的存储器通常比机械类型的存储要更加快。

[0055] 主存储器缓冲页面 310 可以包括火热页面、热门页面和冷门页面，这取决于有多少空间可用。例如，主存储器缓冲页面 310 可以包括来自其他存储 320 的某些最近被检索的冷门页面。一些主存储器缓冲页面 310 的副本可以被存储在 SSS 缓冲页面 315 中。主存储器缓冲页面 310 中的一些页面可能尚未被复制到 SSS 缓冲页面 315。而且，SSS 缓冲页面 315 可以包括不存在于主存储器缓冲页面 310 中的页面。

[0056] SSS 缓冲页面 315 可以被存储在次存储器中。为了实现性能增益，该次存储器可以比存储 320 执行得更好（例如更快的响应时间、更多的带宽等等）。可以在不需要机械移动的情况下访问某种存储器（例如固态存储），并且该存储器是非易失的。虽然，这种存储器可以比存储 320 执行得更好，但该存储器可能比主存储器更慢。

[0057] 其他存储 320 可以包括诸如硬盘、磁带、其他非易失存储等之类的非易失存储。这种其他存储 320 可以包括组件（例如臂、盘或其他介质等等），这些组件在操作上移动（例如在盘上往返、旋转或其他方式）以提供对存储 320 的存储设备上的存储器的访问。在不昂贵的实现中，存储 320 可以具有比系统 300 的主存储器和 SSS 更少的吞吐量。然而，在一些系统中，通过使用联合工作的许多存储设备（例如磁盘），存储 320 的吞吐量可以接近或超过 SSS 的吞吐量，尽管响应时间相对较大。

[0058] 元数据 305 存储了关于在主存储器缓冲页面 310 和 SSS 缓冲页面 315 中的页面的信息。这样的信息可以包括先前结合图 2 所述的元数据。

[0059] 图 2-3 中示出的组件是示例性的且不意味着包括一切的可能需要或包括的组件。在其他实施例中，结合图 2-3 描述的组件和 / 或功能可被包括在其他组件（示出或未示出）中或者被放置在子组件中而不背离此处所描述的主题的各方面的精神或范围。在某些实施例中，结合图 2-3 所描述的组件和 / 或功能可跨多个设备地分布。

[0060] 图 4-5 是概括地表示根据此处所描述的主题的各方面的可发生的动作的流程图。为解释简明起见，结合图 4-5 来描述的方法被描绘和描述为一系列动作。可以理解和明白，此处所描述的主题的各方面不受所示出的各动作和 / 或各动作次序的限制。在一个实施例中，各动作以如下描述的次序发生。然而，在其它实施例中，各动作可以并行地发生、以另一次序发生、和 / 或与此处未呈现和描述的其它动作一起发生。此外，并非所有示出的动作都是实现根据此处所描述的主题的各方面的方法所必需的。另外，本领域的技术人员将了解

和明白，该方法也可以替代地经由状态图而被表示为一系列相互相关联的状态或者被表示为事件。

[0061] 图 4 是概括地表示根据此处所描述的主题的各方面的、可在扫描页面以确定阈值且适合时驱逐页面中发生的示例性动作的流程图。在框 405，动作开始。

[0062] 在框 410 处，接收扫描页面以确定阈值的请求。例如，参考图 4，响应于存储器压力，DBMS 205 可以向缓冲池组件发送扫描页面的请求。

[0063] 在框 415，访问页面的元数据。例如，参考图 3，访问元数据 305。取代访问所有的元数据，可以采样数据库缓冲池的多个随机选择页面的这些元数据。在一个实施例中，采样页面的元数据可以包括从页面的元数据中为每个经采样的页面获得一个或两个时间戳，其中所述一个或两个时间戳对应于访问经采样的页面的最近的一个或多个时间。将这种元数据提供给一个函数(例如分类函数)，该函数生成对应于对该页面的访问的值。由函数为该采样的元数据所生成的值可以被用于选择阈值。例如，在从函数获得值之后，可以选择对应于这些值的不同预定百分比的阈值。百分比的范围(例如 0-5、5-25、25-100)可以对应于火热、热门和冷门。

[0064] 在框 420，可以从所述元数据中确定冷门阈值。例如，冷门阈值可以对应于具有在 0 到 25 之间的百分比范围的经采样的页面的访问频率。一旦已经确定冷门阈值，如果将上述函数应用于由页面的元数据所表示的访问数据返回小于或等于该冷门阈值的值时，就可以确定该页面是冷门的。

[0065] 在框 425，可以从所述元数据中确定热门阈值。例如，热门阈值可以对应于具有在 25 到 75 之间的百分比范围的经采样的页面的访问频率。一旦已经确定热门阈值，如果将上述函数应用于由页面的元数据所表示的访问数据返回大于冷门阈值且小于或等于热门阈值的值时，就可以确定该页面是热门的。

[0066] 在框 430 处，接收驱逐页面的请求。例如，参考图 2，DBMS 205 的驱逐组件可以接收将一个页面从缓冲池 215 中驱逐出去以为一个或多个其他页面释放空间的请求。

[0067] 在框 435，可以选择缓冲池 215 的页面以考虑用于驱逐。例如，参考图 3，可以选择主存储器缓冲页面 310 中的一个考虑用于驱逐。选择可以使用用于循环遍历页面以尝试找出适于驱逐的页面的算法、队列或其他数据结构等等来做出。

[0068] 在框 440 处，分类页面。这可以通过将上述函数应用于页面的元数据以获得一个值来完成。随后，可以将该值与前述的阈值相比较以确定该页面是火热、热门还是冷门的。

[0069] 在框 445，如果合适，驱逐该页面。例如，如果确定该页面是冷门的，则将该页面驱逐到磁盘。如果该页面是热门的，则将该页面复制到 SSS 中，除非达到 SSS 的吞吐量的阈值，在这种情况下，就将该页面驱逐到磁盘。如果该页面是火热的，则将该页面保留在主存储器中，无需被驱逐。

[0070] 在框 450 处，可以执行其他动作(如果存在)。

[0071] 图 5 是概括地表示根据此处所描述的主题的各方面的、可在由 DBMS 接收访问请求且缓冲池已满时发生的一些示例性动作的流程图。在框 505，动作开始。

[0072] 在框 510，接收访问数据库的数据的请求。例如，参考图 2，DBMS 205 可接收访问数据库中的数据的请求。

[0073] 在框 515，DBMS 确定数据驻留在主存储器之外。

[0074] 例如，参考图 2，DBMS 205 可以确定所请求的数据驻留在 SSS 225 或旋转介质 210 中。在框 520，DBMS 确定主存储器中的缓冲池满了。例如，参考图 2，DBMS 205 确定主存储器 220 被页面填满并且需要驱逐一个页面以为数据腾出空间。

[0075] 在框 525 处，选择用于驱逐的页面。选择用于驱逐的页面可以包括使用根据基于对页面的访问的频率的至少三种分类（例如火热、热门和冷门）中的一个来分类页面的函数（例如先前所述的分类函数）。例如，参考图 2，DBMS 205 可以确定主存储器 220 中用于驱逐的页面。

[0076] 在框 530 处，驱逐页面。例如，参考图 2，DBMS 205 可以将所选的页面驱逐到 SSS 225 或旋转介质 210。将页面驱逐到 SSS 225 可以包括将页面复制到 SSS 225 中。这将把页面保持在缓冲池 215 中，仅仅是不在缓冲池的主存储器 220 中。在这种方式中，SSS 225 可以扩展缓冲池（而不是仅仅作为用于被驱逐出主存储器 220 的页面的单独高速缓存）。

[0077] 在框 535 处，可以执行其他动作（如果存在）。

[0078] 虽然，上述讨论引用了使用固态存储来扩展缓冲池，但在其他实施例中，除了固态存储之外的存储也可被用于扩展缓冲池。其他存储可以是目前现有的或还在开发的。根据在此描述的主题的各方面，其他存储可以具有比旋转存储更好但比主存储器更差的吞吐量、带宽或某个其他特性。

[0079] 如从上述详细描述中可以看见，已经描述了关于数据系统的缓冲池的各方面。尽管本文所描述的主题的各方面易于作出各种修改和替换构造，但其某些说明性实施例在附图中示出并在上面被详细地描述。然而，应当理解，并不旨在将所要求保护主题的各方面限制于所公开的具体形式，而是相反地，目的是要覆盖落入本文所描述的主题的各方面的精神和范围之内的所有修改、替换构造和等效方案。

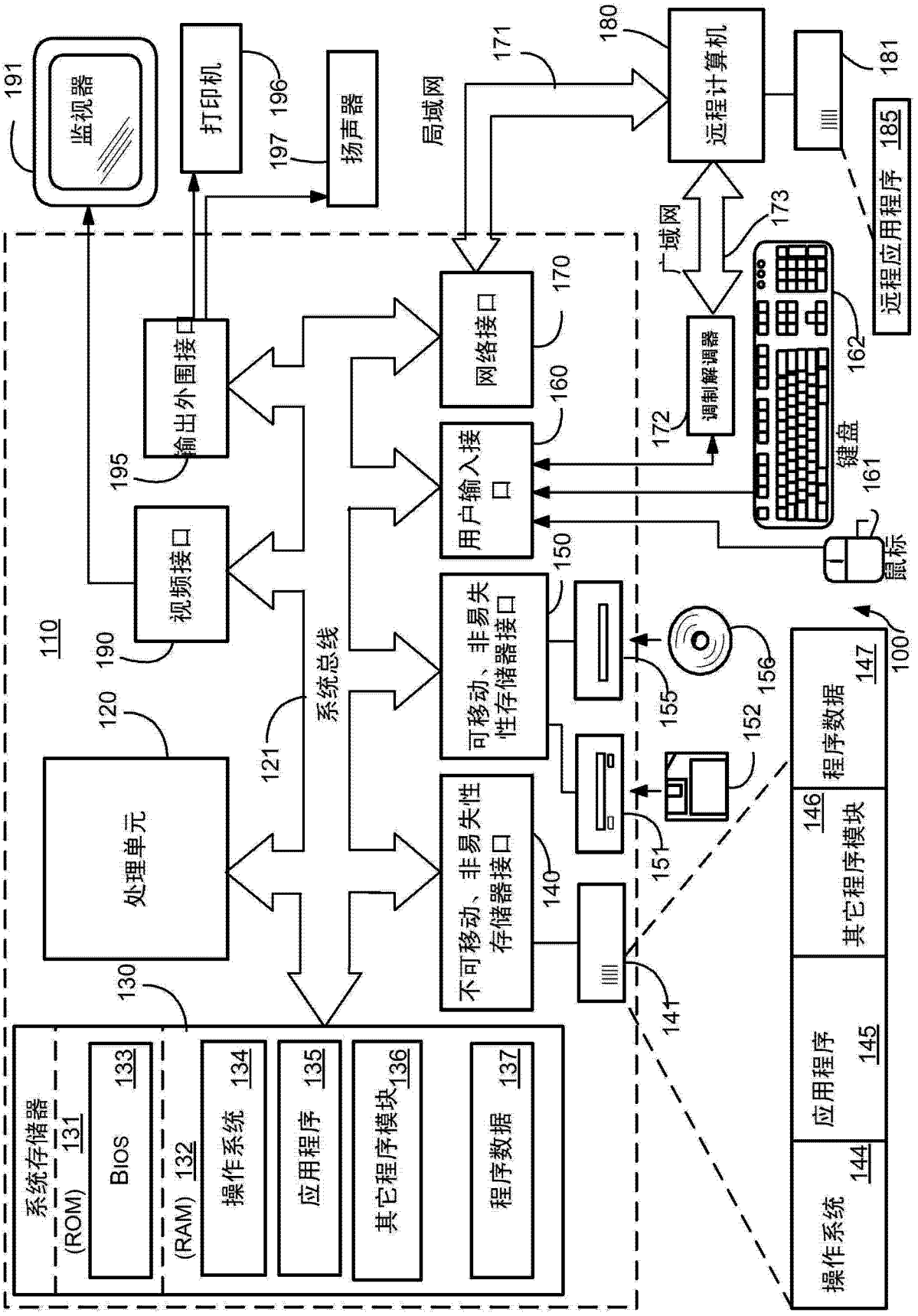

191
监视器
打印机
196
197
扬声器
局域网
171
180
远程计算机
181
185
远程应用程序
广域网
173
162
键盘
161
鼠标
172
调制解调器
网络接口
170
160
输出外围接口
195
用户输入接口
150
视频接口
190
110
120
处理单元
121
系统总线
可移动、非易失性存储器接口
155
156
152
151
不可移动、非易失性存储器接口
140
141
130
147
程序数据
146
其它程序模块
145
应用程序
144
操作系统
100
系统存储器
(ROM) 131
BIOS 133
(RAM) 132
操作系统 134
应用程序 135
其它程序模块 136
程序数据 137

图 1

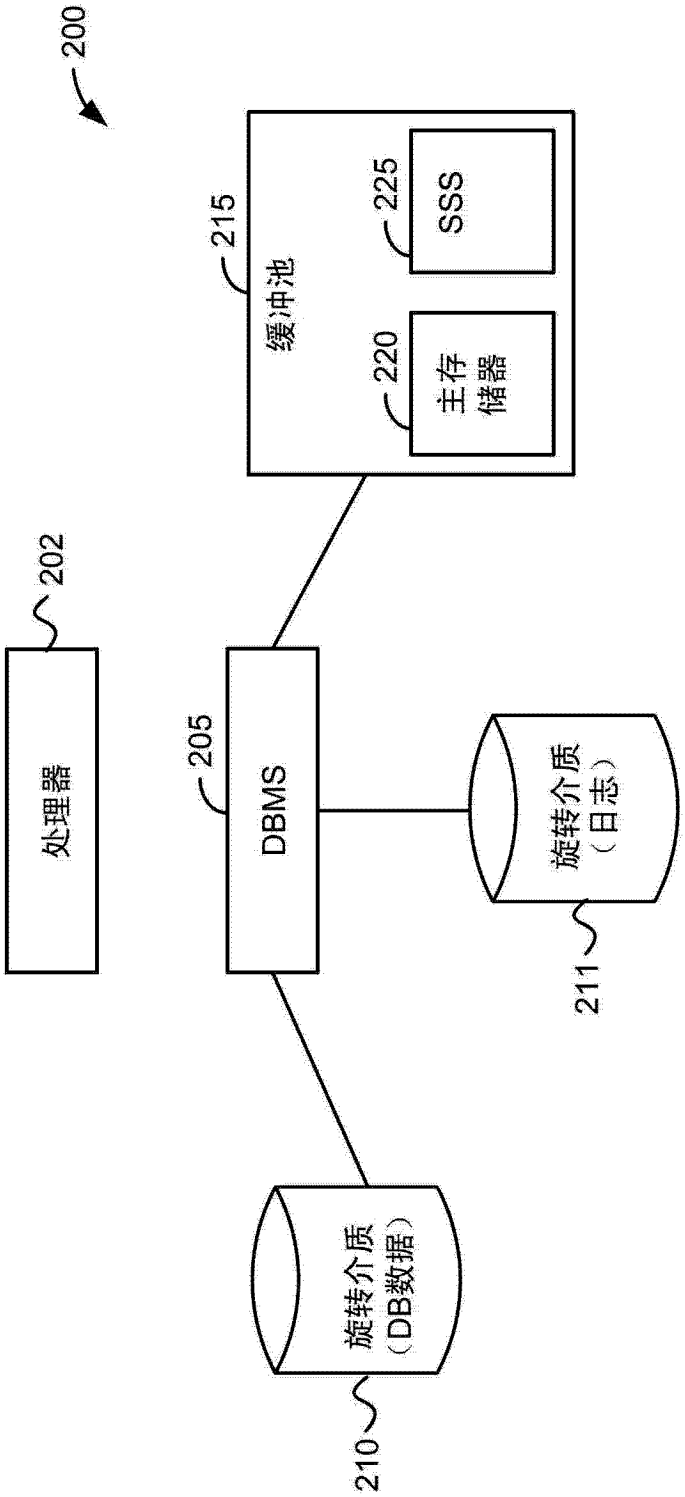

200
215
缓冲池
225
SSS
220
主存储器
202
处理器
205
DBMS
旋转介质（日志）
211
旋转介质（DB数据）
210

图 2

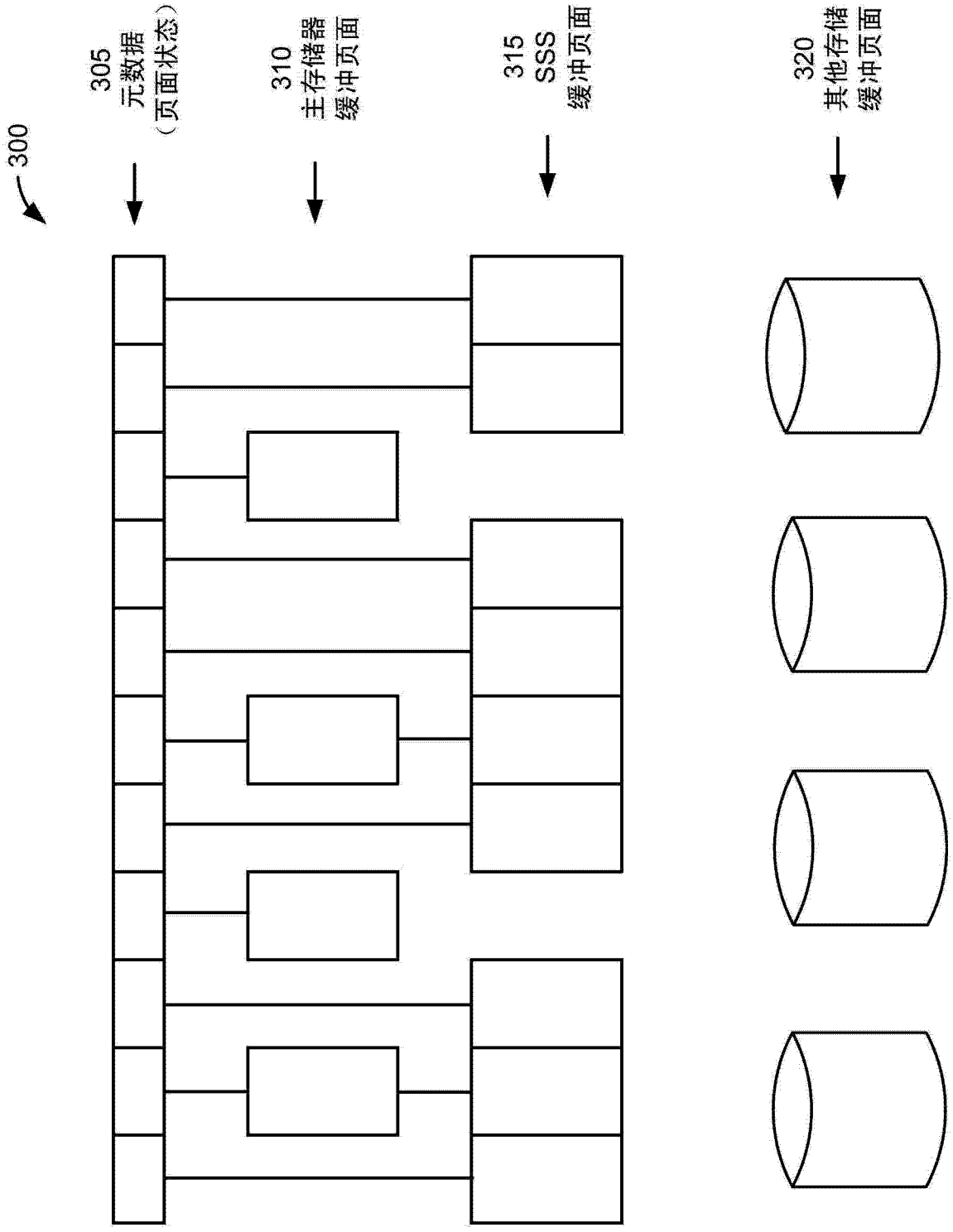


图 3

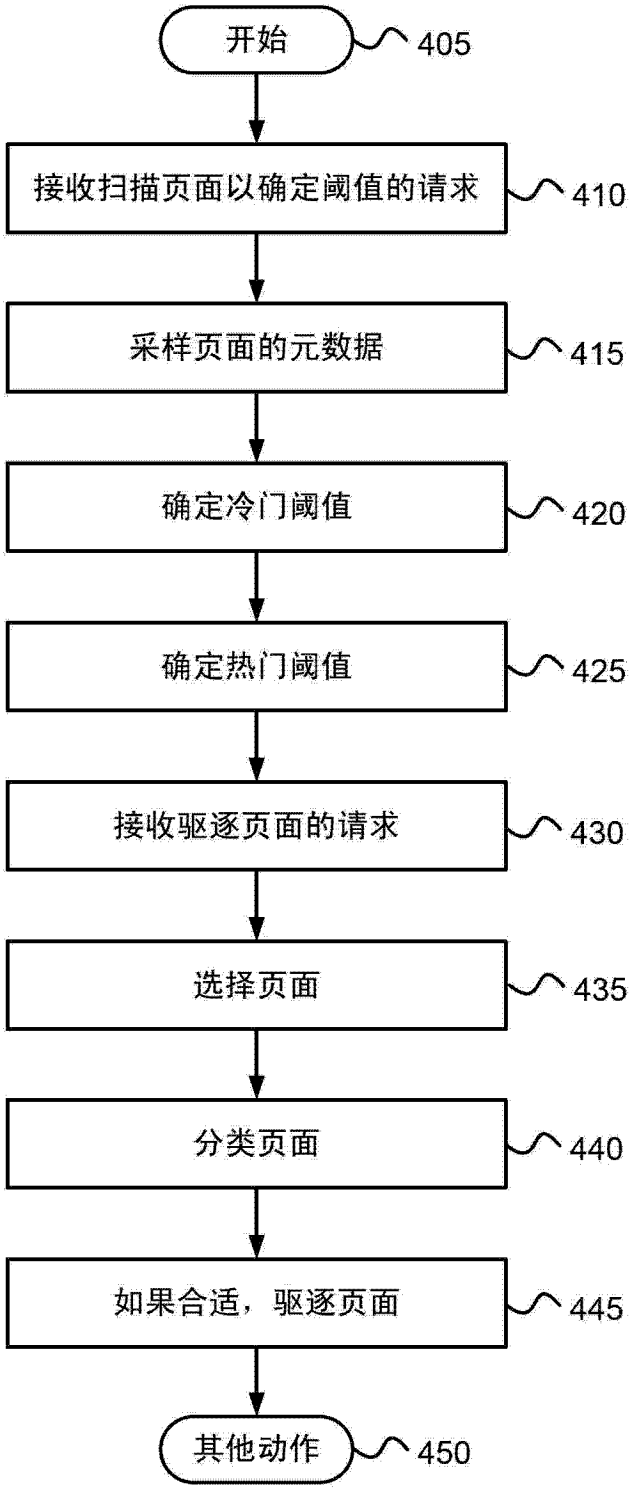

开始
405
接收扫描页面以确定阈值的请求
410
采样页面的元数据
415
确定冷门阈值
420
确定热门阈值
425
接收驱逐页面的请求
430
选择页面
435
分类页面
440
如果合适，驱逐页面
445
其他动作
450

图 4

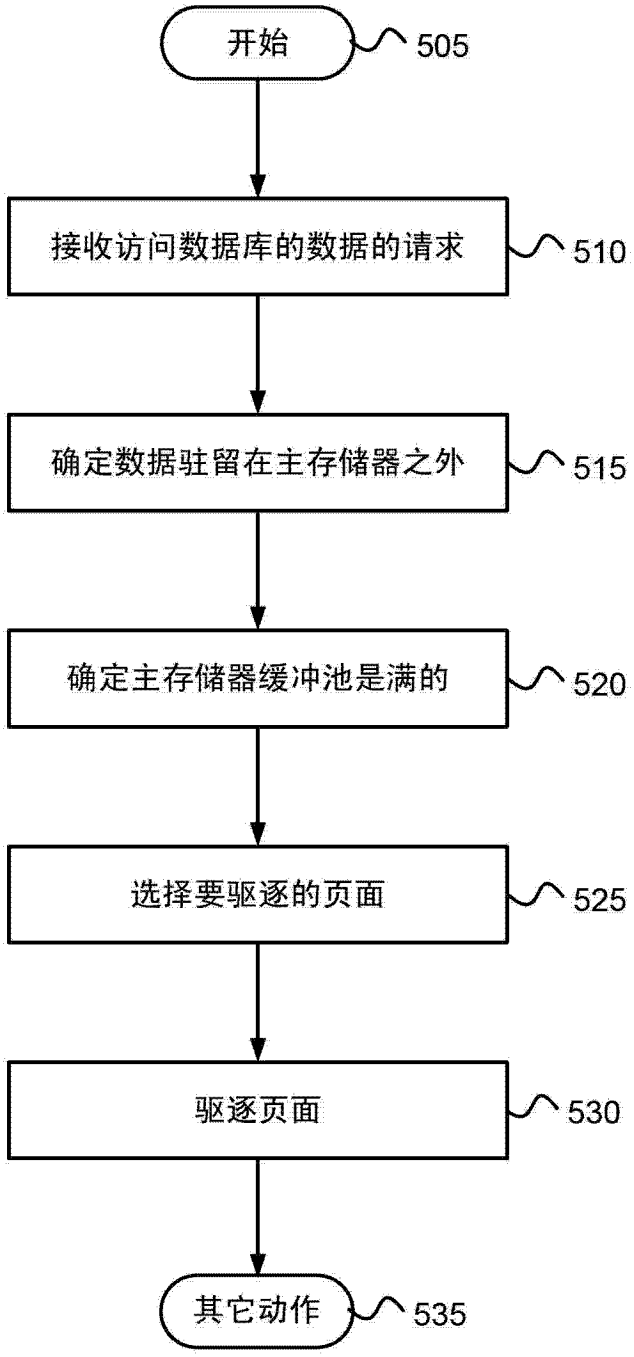

开始
505
接收访问数据库的数据的请求
510
确定数据驻留在主存储器之外
515
确定主存储器缓冲池是满的
520
选择要驱逐的页面
525
驱逐页面
530
其它动作
535

图 5