



(12)发明专利

(10)授权公告号 CN 105683953 B

(45)授权公告日 2017. 11. 10

(21)申请号 201480058389.3

(22)申请日 2014.10.23

(65)同一申请的已公布的文献号

申请公布号 CN 105683953 A

(43)申请公布日 2016.06.15

(30)优先权数据

61/895,263 2013.10.24 US

14/133,302 2013.12.18 US

(85)PCT国际申请进入国家阶段日

2016.04.22

(86)PCT国际申请的申请数据

PCT/US2014/062066 2014.10.23

(87)PCT国际申请的公布数据

W02015/061629 EN 2015.04.30

(73)专利权人 西部数据技术公司

地址 美国加利福尼亚

(72)发明人 D·S·奥布霍夫 M·A·邦贝

(74)专利代理机构 北京纪凯知识产权代理有限公司 11245

代理人 赵蓉民

(51)Int.Cl.

G06F 17/30(2006.01)

审查员 吴黄飞

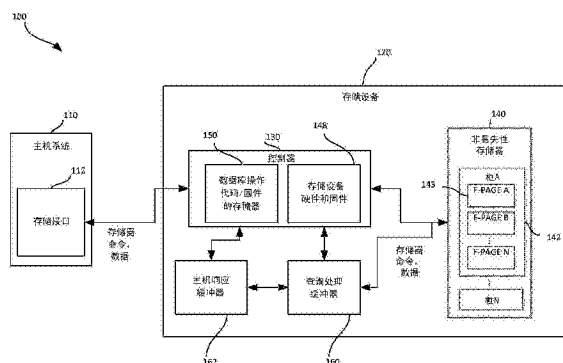
权利要求书2页 说明书5页 附图8页

(54)发明名称

支持加速数据库操作的数据存储设备

(57)摘要

本文公开了用于加速数据库操作的数据存储设备实施例以及相关的方法。在一个实施例中,数据存储设备包括:控制器;一个或多个固态存储器存储设备的阵列;第一存储器,用于存储与数据库操作相关联的处理器可执行指令;以及第二存储器,用于存储与数据库操作有关的数据;其中,控制器被配置为执行指令以进行以下操作:使数据从固态存储器存储设备被读取到第二存储器中;确定数据是否与由指令指定的查询相匹配;以及基于查询匹配的确定来执行数据库操作。



1. 一种非易失性数据存储设备,包括:
控制器;
一个或多个固态存储器存储设备的阵列;
其中,所述一个或多个固态存储器存储设备的阵列被配置为:
存储多个逻辑页,所述逻辑页的大小被配置为与数据库的元组的大小相对应;以及
存储能够基于与所述多个逻辑页相对应的多个逻辑地址来索引的数据;
第一存储器,用于存储与数据库操作相关联的处理器可执行指令;以及
第二存储器,用于存储与所述数据库操作有关的数据;
其中,所述控制器被配置为执行所述指令,所述指令使得所述控制器进行以下操作:
使数据从所述固态存储器存储设备被读取到所述第二存储器中;
确定所述数据是否与由所述指令指定的查询相匹配;以及
基于所述查询匹配的确定来执行数据库操作。
2. 根据权利要求1所述的非易失性数据存储设备,其中,所执行的所述数据库操作包括:将与所述查询相匹配的数据返回到耦合到所述数据存储设备的主机系统。
3. 根据权利要求1所述的非易失性数据存储设备,其中,所执行的所述数据库操作包括:当所述数据与所述查询相匹配时,将指示添加到索引。
4. 根据权利要求1所述的非易失性数据存储设备,其中,所执行的所述数据库操作包括:修改与所述查询相匹配的所述数据并且将所修改后的数据写回到所述固态存储器存储设备。
5. 根据权利要求1所述的非易失性数据存储设备,其中,所述第一存储器被配置为存储能够由耦合到所述数据存储设备的主机系统配置的数据库操作。
6. 根据权利要求1所述的非易失性数据存储设备,其中,所述控制器还被配置为通过选择性地略过对所述逻辑页的读取来执行查询有关的数据库操作。
7. 根据权利要求6所述的非易失性数据存储设备,其中,所述选择性地略过是基于与用户数据访问特权相关联的逻辑的。
8. 根据权利要求6所述的非易失性数据存储设备,其中,所述选择性地略过是基于与查询相关联的逻辑的。
9. 根据权利要求6所述的非易失性数据存储设备,其中,所述选择性地略过是通过对与
所述逻辑页中的一个或多个相关联的逻辑地址的模操作来执行的。
10. 根据权利要求1所述的非易失性数据存储设备,其中,所述控制器还被配置为在执行所述指令时启动看门狗定时器以防止挂起状况。
11. 根据权利要求1所述的非易失性数据存储设备,其中,所述控制器还被配置为对所述第二存储器中的所述数据执行一个或多个逻辑操作。
12. 根据权利要求1所述的非易失性数据存储设备,其中,所述控制器包括专用于执行所述指令的处理器。
13. 根据权利要求1所述的非易失性数据存储设备,其中,所述指令是从耦合到所述数据存储设备的主机系统被接收的。
14. 一种在非易失性数据存储设备中执行数据操作的方法,所述非易失性数据存储设备包括一个或多个固态存储器存储设备的阵列、第一存储器、以及第二存储器,所述方法包

括：

将多个逻辑页存储在一个或多个固态存储器存储设备中，所述逻辑页的大小被配置为与数据库的元组的大小相对应；

将数据存储在一个或多个固态存储器存储设备中，其中，所述数据能够基于与所述多个逻辑页相对应的多个逻辑地址来索引；

使所述数据从所述非易失性数据存储设备的所述固态存储器存储设备读取到所述第二存储器中；

确定所述数据是否与指定的查询相匹配；以及

基于所述查询匹配的确定来执行数据库操作，

其中：

所述第一存储器被配置为存储与数据库操作相关联的处理器可执行指令并且所述查询是由所述指令指定的；以及

所述第二存储器被配置为存储与所述数据库操作有关的数据。

15. 根据权利要求14所述的方法，其中，所执行的所述数据库操作包括：将与所述查询相匹配的数据返回到耦合到所述数据存储设备的主机系统。

16. 根据权利要求14所述的方法，其中，所执行的所述数据库操作包括：当所述数据与所述查询相匹配时，将指示添加到索引。

17. 根据权利要求14所述的方法，其中，所执行的所述数据库操作包括：修改与所述查询相匹配的所述数据并且将所修改后的数据写回到所述固态存储器存储设备。

18. 根据权利要求14所述的方法，其中，所述第一存储器被配置为存储能够由耦合到所述数据存储设备的主机系统配置的数据库操作。

19. 根据权利要求14所述的方法，还包括：通过选择性地略过对所述逻辑页的读取来执行查询有关的数据库操作。

20. 根据权利要求19所述的方法，其中，所述选择性地略过是基于与用户数据访问特权相关联的逻辑的。

21. 根据权利要求19所述的方法，其中，所述选择性地略过是基于与查询相关联的逻辑的。

22. 根据权利要求19所述的方法，其中，所述选择性地略过是通过对与所述逻辑页中的一个或多个相关联的逻辑地址的模操作来执行的。

23. 根据权利要求14所述的方法，还包括：在执行所述指令时启动看门狗定时器以防止挂起状况。

24. 根据权利要求14所述的方法，还包括：对所述第二存储器中的所述数据执行一个或多个逻辑操作。

25. 根据权利要求14所述的方法，其中，所述指令是在专用处理器中执行的。

26. 根据权利要求14所述的方法，其中，所述指令是从耦合到所述数据存储设备的主机系统被接收的。

支持加速数据库操作的数据存储设备

技术领域

[0001] 本公开涉及非易失性数据存储设备以及用于加速在这样的设备中的数据操作的方法。

背景技术

[0002] 经常在执行速度非常重要的环境中执行数据库操作。诸如返回查询结果和进行索引等的常见操作通常是I/O密集的并且在主机系统(例如,计算设备)与这样的操作在其上执行的数据存储设备之间消耗大的数据带宽。

附图说明

[0003] 现在将参照下面的附图来描述实现本发明的各种实施例的各种特征的系统和方法,在附图中:

[0004] 图1A示出了根据本发明的一个实施例的示例数据存储设备。

[0005] 图1B示出了根据本发明的一个实施例的示例数据库操作加速方法。

[0006] 图2示出了根据一个实施例的数据存储设备的内部数据布局。

[0007] 图3A和3B是示出了根据一个实施例的数据库元素的示例布局的框图。

[0008] 图4A和4B是示出了根据一个实施例的可以如何执行过滤读(filtered read)操作的流程图。

[0009] 图5是示出了根据一个实施例的可以如何执行索引操作的流程图。

具体实施方式

[0010] 本公开的一些实施例涉及被配置为加速数据库操作的数据存储设备(例如,固态驱动器(SSD))。在实施例中,数据存储设备支持可变大小的逻辑页,该逻辑页的大小可以被定制以与数据库数据结构内的各个数据单元(例如,元组)相匹配。其结果是,一些数据库操作可以被加速,这是因为可以略过一些数据的逻辑地址范围以减少从存储介质读出并且传送到主机的数据量,这实现了更加高效的数据库操作。

[0011] 尽管描述了本公开的一些实施例,但这些实施例仅以示例的方式给出,而不是要限制本公开的范围。事实上,本文所描述的新颖的方法和系统可以以各种其它形式来实现。此外,可以在本文所描述的方法和系统的形式上做出各种省略、替代和变化,而不偏离本公开的精神。

[0012] 数据存储系统概述

[0013] 图1A示出了根据本发明的一个实施例的示例数据存储设备120。如图所示,数据存储设备120(例如,固态驱动器、混合驱动器等)包括控制器130和非易失性固态存储器140,非易失性固态存储器140包括一个或多个存储单元,例如存储块。图1A示出了块被标识为块“A”142到块“N”的示例。尽管为方便起见,示出了单个的非易失性固态存储器140,但存储设备可以包括多个这样的存储器。非易失性固态存储器140的每个块包括多个闪速页(F页)。

例如,图1A的块A 142包括多个F页,被标识为F页A 143、B、到N。在一些实施例中,每个“F页”是非易失性固态存储器140中能够在单个的操作中被编程或者被作为单元被编程的存储器单元的最小分组。替代非易失性固态存储器140或者除了非易失性固态存储器140之外,可以使用磁旋转介质和/或诸如MRAM和/或相变存储器之类的其它非易失性存储器。

[0014] 控制器130可以从主机系统110中的存储接口112(例如,设备驱动器)接收数据和/或存储访问命令。由存储接口112传送的存储访问命令可以包括由主机系统110发出的写命令和读命令。命令可以指定数据存储设备120中的逻辑块地址,并且控制器130可以在非易失性固态存储器140中执行接收到的命令。在混合硬盘驱动器中,除了非易失性固态存储器140以外,数据可以存储在磁介质存储部件(未在图1A中示出)中。

[0015] 数据存储设备120可以存储从主机系统110接收到的数据,以使得数据存储设备120可以充当主机系统110的存储器存储设备。为便于该功能,控制器130可以实现逻辑接口。逻辑接口可以将存储设备的存储器向主机系统110呈现为可以在其中存储数据的逻辑地址的集合(例如,连续地址)。在内部,控制器130可以将逻辑地址映射到非易失性固态存储器140和/或其它(多个)存储器模块中的各种物理存储器地址。

[0016] 在一个实施例中,控制器130包括存储设备硬件和固件148以及用于数据库操作代码/固件的存储器150。存储设备硬件和固件148用于控制数据存储设备内的数据操作。在一个实施例中,存储器150中的数据库操作代码/固件能够由主机配置,并且可以在其自有的专用的处理器(未在图1A中示出)中执行。对存储在非易失性固态存储器140中的数据执行那些查询。与查询有关的数据被临时地存储在查询处理缓冲器160中并且结果经由主机响应缓冲器162被返回到主机系统110。下面提供了与部件如何交互有关的额外的细节。注意的是,在各种实施例中,这些部件可以被不同地布置。这些部件可以被省略、组合或者分离成另外的子部件。此外,部件148、150、160、和162可以集成到单个的处理器封装中或者被实现为相互通信的各种分立部件。

[0017] 图1B示出了根据一个实施例的示例数据库操作加速方法。在一个实施例中,控制器130被配置为执行图1B中所示的流程。在框168中,控制器被配置为将指令/代码加载到存储器150中。指令/代码可以来自主机系统。在框170中,控制器被配置为执行存储器150中的指令。在一个实施例中,可以使用专用处理器来处理这些执行以提供进一步的加速。在框172中,控制器被配置为使得数据从固态存储器140被读取到查询处理缓冲器160中。在框174中,控制器被配置为确定数据是否与由指令指定的数据库查询相匹配。在框176中,控制器被配置为基于查询匹配的确定来执行数据库操作。在一个实施例中,框172-176中的一个或多个动作可以作为执行指令的结果被触发。如下面进一步说明的,所执行的数据库操作可以例如包括以下中的一个或多个:(1)将与查询相匹配的数据返回到主机系统110;(2)当数据与查询匹配时,将指示添加到索引;(3)修改与查询相匹配的数据并且将修改后的数据写回到固态存储器。

[0018] 对数据库操作的加速

[0019] 图2示出了根据一个实施例的数据存储设备的内部数据布局。如图所示,可变尺寸的逻辑页(L页)212存储跨各种错误校正页(E页210),这些错误校正页本身是物理闪存F页208的物理细分。在一些实施例中,每个F页存在一个E页,即,F页不被细分。L页可以跨过E页、F页的底层物理边界,以及非易失性固态存储器140内的管芯/块/单元的边界(如边界

217所示)。例如,如图所示,L页可以分布跨多个E页。在示例实现中,E页210可以具有由ECC部分216保护的数据部分214。在一些实现中,可以使用压缩来进一步改变被写到非易失性固态存储器中的L页的大小。

[0020] 在一个实施例中,逻辑页的大小被配置为等于数据库的元组的大小,或者是数据库的元组的整数倍。由于逻辑页大小的这种灵活性,当设计/配置数据库时,数据库管理员可以创建在存储的数据与访问索引之间的匹配相关。例如,如图3A所示,如果一个元组占用2个逻辑页,则为了读取元组7,读逻辑块地址(LBA) 14和15将被读取。基于逻辑地址来对数据进行索引提供了许多优点,包括消除了由主机系统的操作系统(OS)进行划分的开销并且允许使用所有可用的存储。

[0021] 此外,逻辑页和数据库数据对齐可以允许在查询操作中(例如,在图1B中的框172中的动作的执行期间)选择性地略过某些逻辑页范围。例如,在图3B中,建立数据记录以使得逻辑页边界与数据库记录的各个字段对齐。例如,如果用户仅对访问名称和地址字段感兴趣,则可以执行目标读取来读取L页0和L页1。能够认识到,图3B所示的示例示出了一个记录并且相同的原理能够被扩展到读取多个不同记录时的情况。进一步按照该示例,由于字段和逻辑地址对齐,因此对某些字段的索引可以通过对逻辑地址的模操作来访问。通过允许略过预先配置的逻辑页布置中的一些逻辑地址,数据库性能可以相对于读取全部数据并且随后过滤和提供匹配的结果的常规方式有大的改进。此外,逻辑页地址可以基于公式和/或条件来访问。例如,不同的数据库用户可以具有不同的访问特权。一个用户可能仅能够访问名称、地址、和电话字段。因此,对于该用户,逻辑可以用公式表示,以使得他的查询将被阻止访问L页3、 $N+3$ 、 $2N+3$ 等,以及L页4、 $N+4$ 、 $2N+4$ 等,其中N是记录中的字段数量。具有更高访问特权的另一个用户可以访问诸如社会安全号和账号等的额外字段,并且基于不同逻辑的公式可以用于允许该访问。当字段与逻辑页边界对齐时,作为不同访问特权的结果的不同查询被高效地处理,允许数据存储设备在存储设备的逻辑地址层执行许多数据库操作共有的过滤。

[0022] 在一个实施例中,数据存储设备包括用于查询处理的专用缓冲器,例如图1所示的查询处理缓冲器160,或者“QPB”。在一个实施例中,QPB是数据路径的一部分并且能够保存一个逻辑页。

[0023] 此外,在一个实施例中,数据存储设备包括用于保存对主机查询的响应的缓冲器,例如,图1所示的主机响应缓冲器162,或“HRB”。在一个实施例中缓冲器的大小是逻辑页大小的整数倍,但取决于配置,可能是不同的。

[0024] 此外,在一个实施例中,数据存储设备包括用于执行主机提供的代码(异代码(xenocode)或XC)的专用处理器。在一个实施例中,异代码应当至少具有对查询处理缓冲器160的读访问权和对主机响应缓冲器162的读/写访问权。在一个实施例中,数据存储设备包括针对异代码的代码存储器(XC存储器或XCM),如图1中的元件150所示。XCM的大小可以被设置为足够大以能够执行查询。此外,在一个实施例中,数据存储设备具有控制寄存器(XCR)的集合,XCR允许存储设备的硬件和固件(例如,图1中的148)与异代码通信并且提供对异代码进行重置或反重置的硬件机制。在一个实施例中,提供“看门狗(watchdog)”功能,以使得针对挂起或超时状况来监视异代码的执行,以使得设备的硬件/固件可以对异代码的执行进行重置并且防止整个存储设备挂起和超时。

[0025] 在一个实施例中,数据存储设备被配置为向主机提供关于XC类型、XCM和HRB的大小、XCR配置、以及执行时间的信息。该信息可以通过电子方式来提供或者在产品文档中提供。

[0026] 查询执行流程

[0027] 在一个实施例中,根据上文描述,数据存储设备可以被填充具有关系数据库元组。图4A和4B示出了可以如何执行过滤读。图4A示出了发生在预期查询执行的初始化中的一些。在框400中,主机系统请求来自数据存储设备的配置信息。该信息可以与包括异代码建立信息在内的各种设置有关。在框402中,数据存储设备利用被请求的配置信息来进行响应,该配置信息包括诸如XC类型、XCM大小、QPB和HRB映射、执行时间等的细节。在框404中,主机系统发送命令(例如,供应商专用命令(VSC))以加载用于执行的异代码。异代码可能之前已经由主机系统发送以用于存储在数据存储设备的XCM或固态存储器中。这允许主机根据需要来配置/改变异代码的最大灵活性,同时提供了与数据存储设备本地化的优化的查询执行的益处。在一个实施例中,可以使用硬件来进一步加速这些操作。在框406中,数据存储设备接收命令,开始准备异代码的执行,并且向主机系统确认准备就绪。准备工作可以包括:利用给定图像来填充XCM,清除HRB等。在框408中,主机系统发送命令(例如,VSC)以启动操作(例如,在该情况下,对逻辑页集合的XC过滤读)。

[0028] 图4B示出了过滤读操作的流程的一部分。图4B示出了对集合中的每个逻辑页执行的动作。在框420中,数据存储设备将下一逻辑页读取到查询处理缓冲器中。在框422中,数据存储设备发布对异代码的重置。该重置与看门狗监视器有关以确保代码正确执行并且不会导致挂起状况。然后,在框424中,执行异代码存储器中的代码。在框426中,如果查询结果为成功,则异代码将指示写到异代码寄存器(例如,XCR.Good=1)。在一个实施例中,当异代码执行完成时,其将完成指示写到异代码寄存器中(例如,XCR.Done=1)。这使得数据存储设备将逻辑页发送到主机系统(例如,经由主机响应缓冲器)并且重置异代码(框428)。在一个实施例中,如果异代码花费过长时间来执行,则看门狗重置该异代码。逻辑页被视为与查询不匹配。作为图4A和4B执行的结果,数据存储设备可以在内部读取所有的数据库元组,但是仅匹配的记录被传送到主机。这样的本地化处理减少了数据传送量并且因此提高了数据吞吐量。处理可以进一步与存储设备的硬件加速相耦合和/或由存储设备的硬件加速来补充,以实现甚至更好的改进。在一个实现中,除了如之前所描述的基于逻辑地址的过滤读操作外,可以执行该过滤读操作。

[0029] 在一个实施例中,索引操作可如下发生。很类似图4A,主机系统和数据存储设备可以执行一些初始化操作。一旦初始化完成,则在一个实施例中,主机系统发送命令(例如,VSC)以利用逻辑页集合来创建异代码辅助的逻辑页子集。针对集合中的每个逻辑页的处理流程在图5中示出。在框500中,数据存储设备将下一逻辑页读取到查询处理缓冲器。在框502中,数据存储设备发布对异代码的重置。然后,在框504中,异代码存储器中的代码被执行。在框506中,如果查询结果为成功,则异代码将逻辑页号(例如,XCR.Page)写到主机响应缓冲器。在一个实施例中,当异代码执行完成时,其将完成指示写到异代码寄存器(例如,XCR.Done=1)。这使得数据存储设备重置异代码(框508)。逻辑页还可以可选地被发送到主机系统。在一个实施例中,如果异代码花费过长的时间来执行,则看门狗重置异代码。在这样的情况下,逻辑页被视为与查询不匹配。在设置完成后,数据存储设备将HRB的内容发送

到主机系统,将索引操作的结果给予主机系统。在一个实施例中,不是返回匹配的页的集合,或者除此之外,异代码可以提供更精细的操作,例如计算均值,或者进行其它统计分析。

[0030] 在一个实施例中,数据存储设备可以提供可配置的看门狗时序以更好地将异代码执行与期望的业务相匹配。

[0031] 在一个实施例中,数据存储设备可以超出对数据内容的只读访问。例如,异代码可以在必要时提供读-修改-写操作。数据存储设备可以通过支持异代码对查询处理缓冲器的写访问以及提供将修改后的逻辑页写回到固态存储器的能力来实现该功能。异代码例如可以被配置为读出与某个值相匹配的页、执行某个操作、以及将修改后的页写回到固态存储器。这可以在没有主机系统与数据存储设备之间的数据传送的情况下进行,并且不需要主机系统的处理能力来执行这样的操作。

[0032] 其它变型

[0033] 本领域技术人员将理解,在一些实施例中,可以使用其它的方式和方法。例如,非易失性固态存储器阵列可以使用NAND闪速存储器设备来实现。可以可替换地使用其它类型的固态存储器设备,例如闪速集成电路阵列、硫族化合物RAM (C-RAM)、相变存储器 (PC-RAM或PRAM)、可编程金属化单元RAM (PMC-RAM或PMCm)、双向统一存储器 (OUM)、电阻RAM (RRAM)、NOR存储器、EEPROM、铁电存储器 (FeRAM)、磁阻RAM (MRAM)、其它分立的NVM (非易失性固态存储器) 芯片、或其任何组合。在一个实施例中,非易失性固态存储器阵列优选地包括具有能够存储一个以上的单个信息位的多级单元的多级单元 (MLC) 设备,尽管可以使用单级单元 (SLC) 存储器设备或SLC和MLC设备的组合。在一个实施例中,数据存储设备120可以包括其它存储器模块,例如一个或多个磁存储器模块。此外,本公开的系统和方法还可以用于更加常规的硬盘驱动器以及包括固态和硬盘驱动器部件的混合驱动器。例如,采用磁记录技术的一些硬盘驱动器可以采用上述的数据寻址和处理方案。

[0034] 尽管已经描述了本公开的一些实施例,这些实施例仅以示例的方式给出,并且不是要限制本公开的范围。事实上,本文所描述的新颖的方法和系统可以以各种其他形式实现。此外,可以在本文所描述的方法和系统的形式上做出各种省略、替代和变化,而不偏离本公开的精神。所附的权利要求及其等价物旨在覆盖将落入本公开的范围和精神内的这种形式或修改。例如,所描述的各部件可以被实现为处理器、ASIC/FPGA、或专用硬件上的软件和/或固件。例如,本领域技术人员将理解的是,在一些实施例中,在一些实施例的过程中采取的实际步骤可以不同于图中所示的那些。取决于实施例,在上述的示例中所描述的步骤中的一些可以被移除,其它的步骤可以被添加,并且步骤的序列可以被改变和/或并行执行。此外,上文公开的具体实施例的特征和属性可以以不同的方式来组合以形成额外的实施例,所有这些实施例落入本公开的范围。尽管本公开提供了一些优选的实施例和应用,但是对于本领域普通技术人员而言显而易见的其它实施例,包括不提供本文阐述的所有特征和优点的实施例,也在本公开的范围。因此,本公开的范围旨在仅通过参照所附的权利要求来定义。

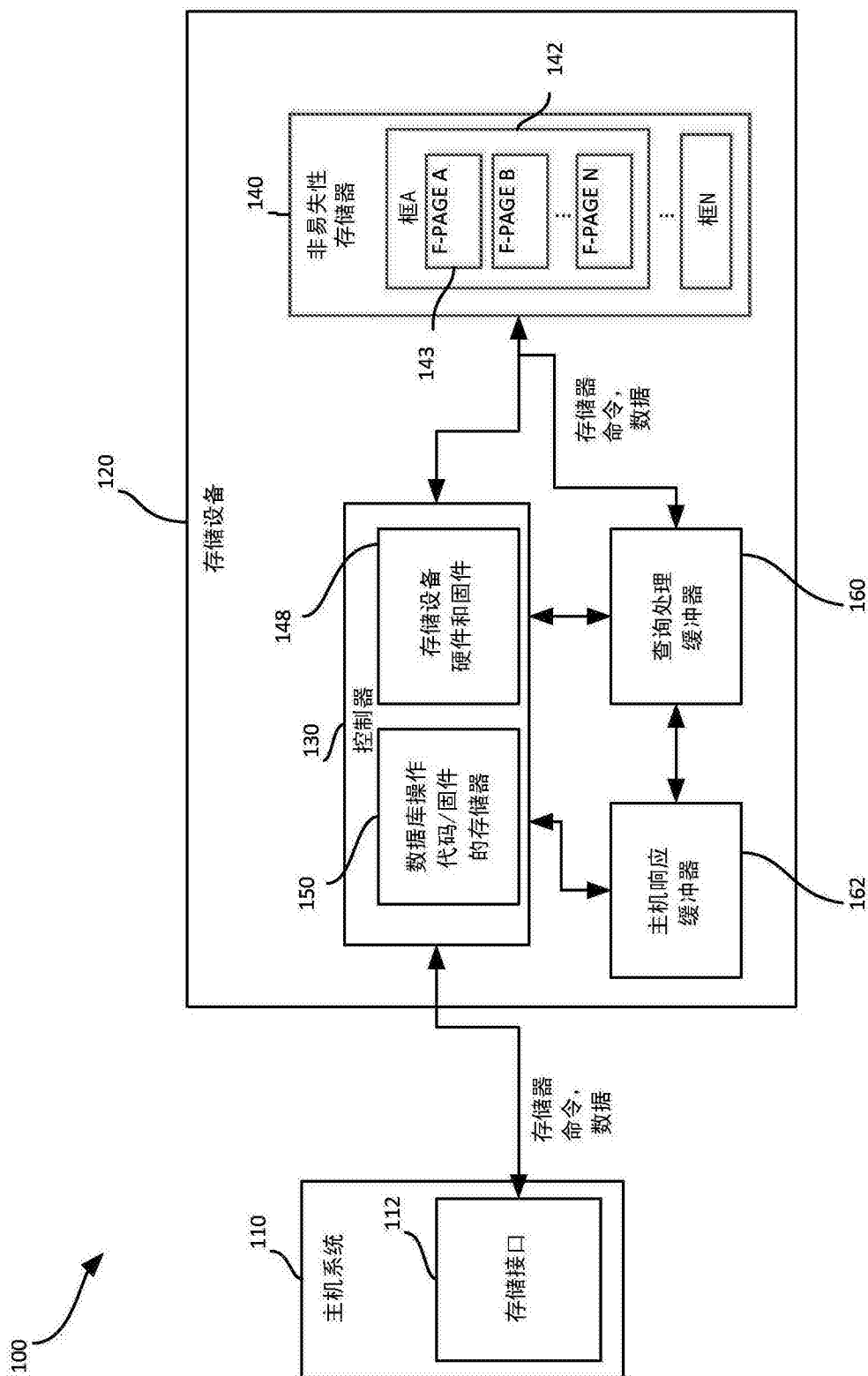


图1A

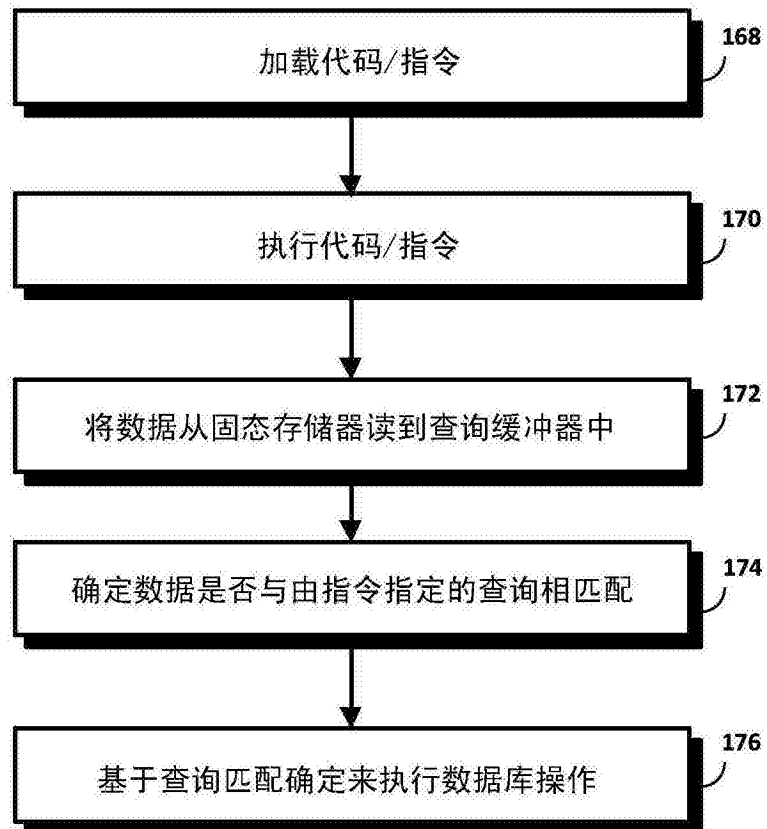


图1B

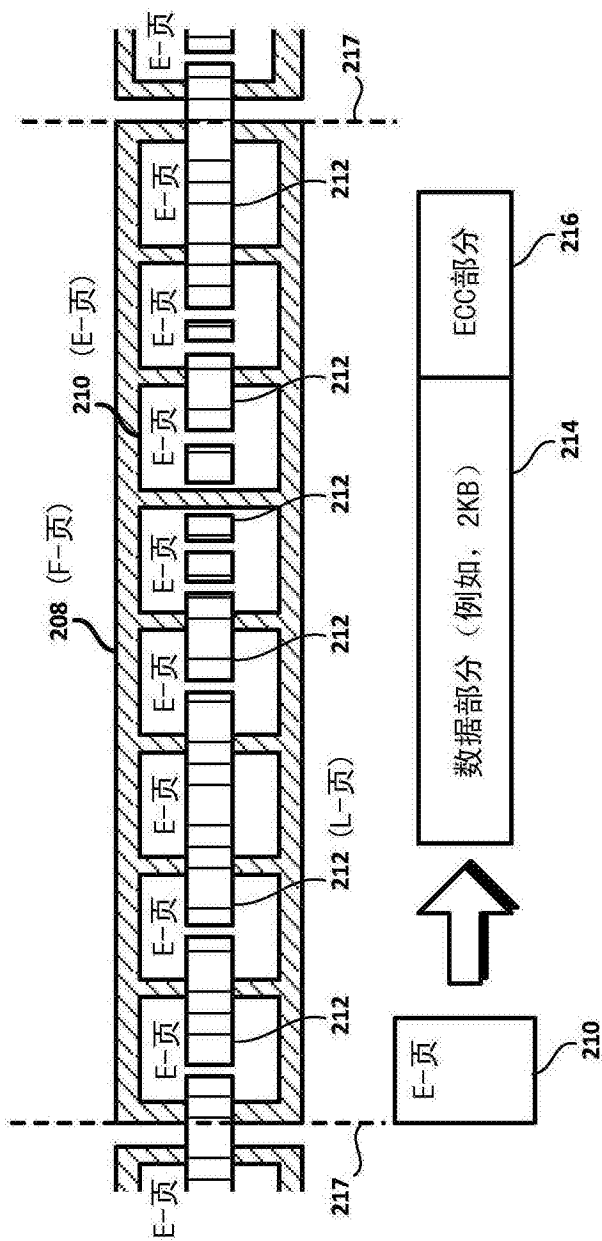


图2

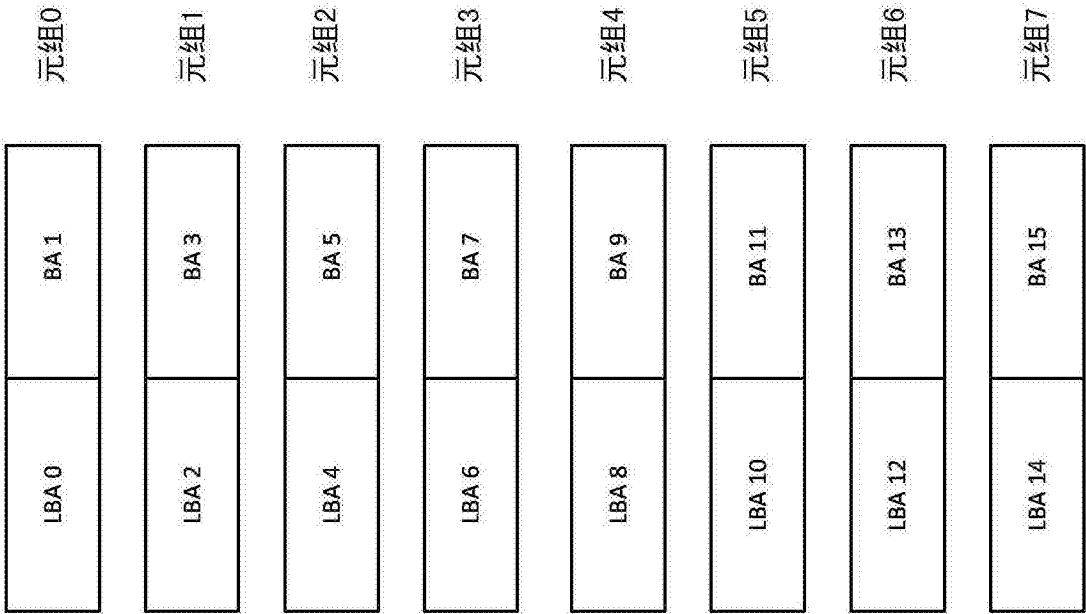


图3A

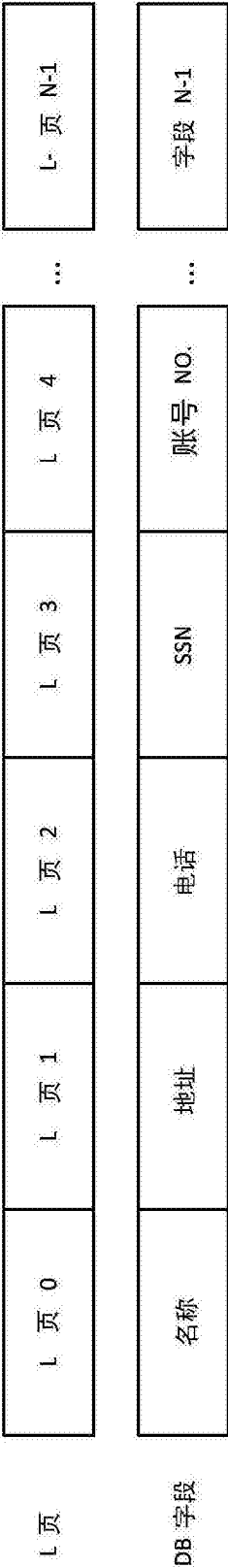


图3B

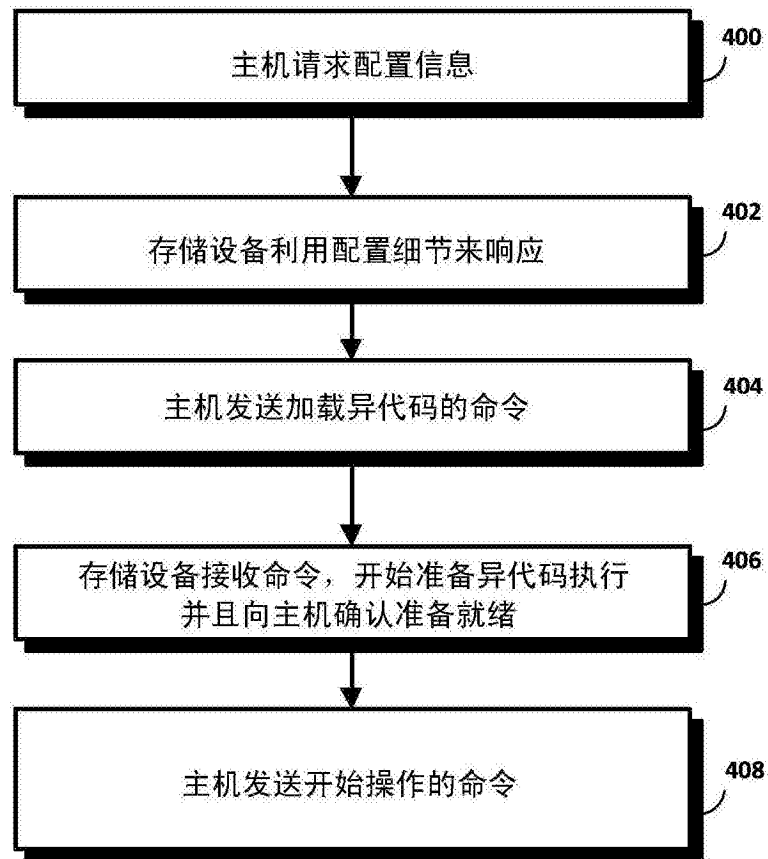


图4A

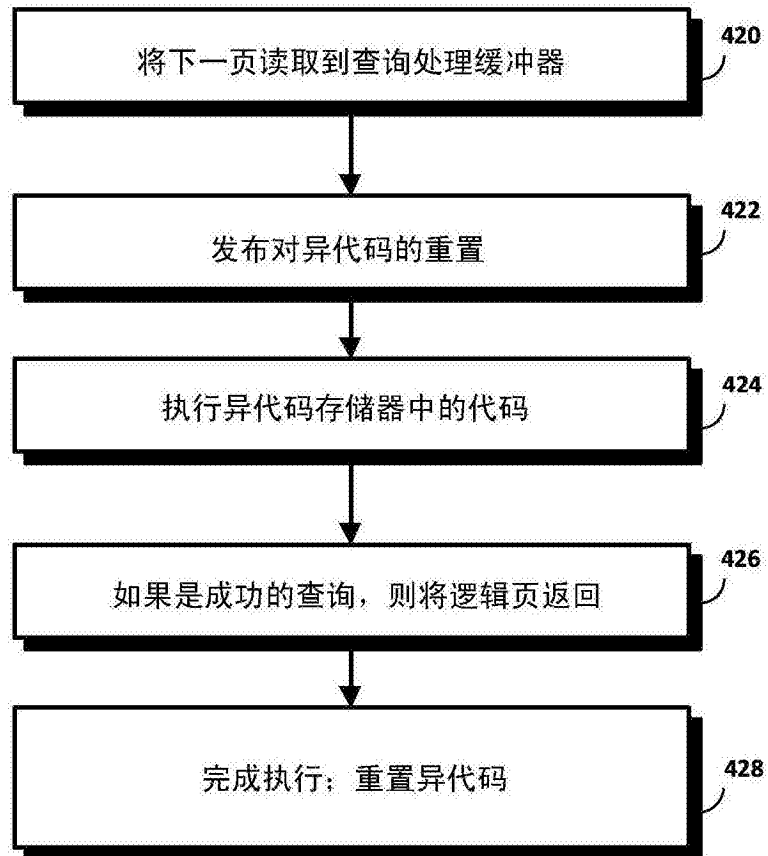
集合中的每个逻辑页的循环

图4B

集合中每个逻辑页的循环

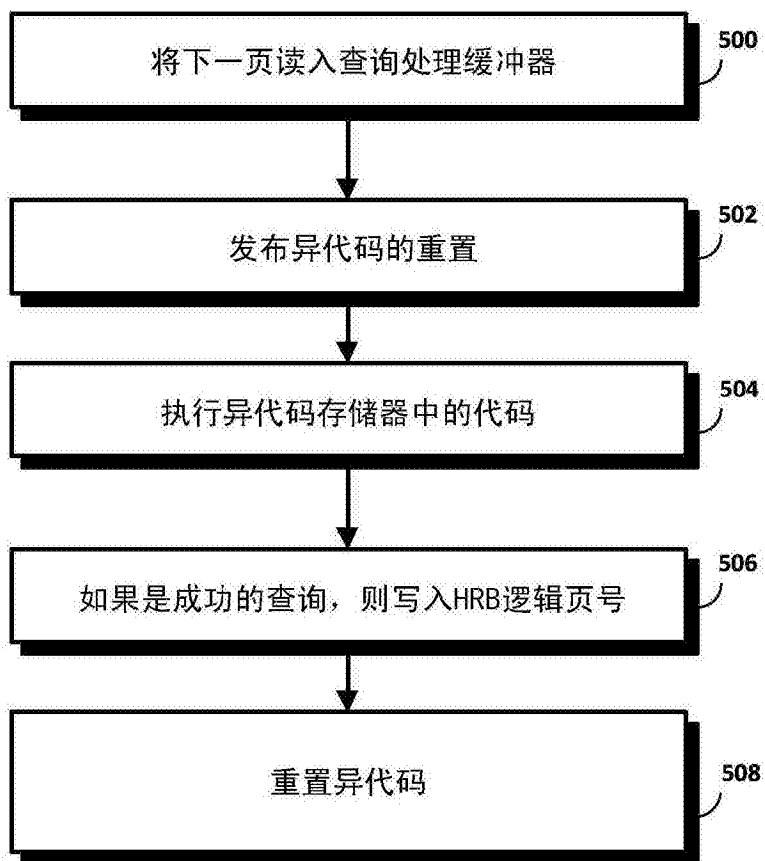


图5