(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118051472 A
(43)申请公布日 2024.05.17

(21)申请号 202311518667.6
(22)申请日 2023.11.14
(30)优先权数据
63/425,937 2022.11.16 US
18/154,755 2023.01.13 US
(71)申请人 三星电子株式会社
地址 韩国京畿道
(72)发明人 D•L•赫尔米克
C-C•C-J•A•吴
(74)专利代理机构 北京市柳沈律师事务所 11105
专利代理师 巫资青
(51)Int.Cl.
G06F 16/11(2019.01)
G06F 3/06(2006.01)
G06F 16/172(2019.01)

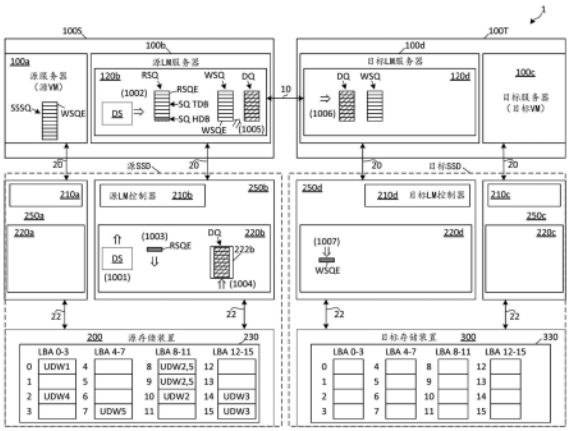
权利要求书2页 说明书16页 附图8页

(54)发明名称

存储数据实时迁移的在存储设备处的预填充命令

(57)摘要

提供了一种用于管理数据迁移操作的方法，包括:由存储设备创建指示要从存储设备的源存储装置复制到目标存储装置的数据在源存储装置处的位置的读取提交队列条目,该读取提交队列条目包括包含元数据的字段,该元数据包括用于从源存储装置读取数据的信息。



CN 118051472 A

1.一种用于管理数据迁移操作的方法,该方法包括:由存储设备创建指示要从所述存储设备的源存储装置复制到目标存储装置的数据在所述源存储装置处的位置的读取提交队列条目,所述读取提交队列条目包括包含元数据的字段,所述元数据包括用于从所述源存储装置读取所述数据的信息。

2.根据权利要求1所述的方法,还包括:

由所述存储设备向主机发送所述读取提交队列条目;

由所述存储设备接收基于所述读取提交队列条目的命令;以及

由所述存储设备基于所述命令从所述源存储装置读取所述数据。

3.根据权利要求1所述的方法,还包括基于所述读取提交队列条目存储在所述存储设备上而从所述源存储装置读取所述数据。

4.根据权利要求1所述的方法,其中:

所述读取提交队列条目的字段是所述读取提交队列条目的多个条目中的一个条目;以及

所述多个条目包括用于从所述源存储装置读取所述数据的信息。

5.根据权利要求1所述的方法,其中,所述元数据包括数据指针或命令标识符。

6.根据权利要求1所述的方法,其中,使用来自与用户数据写入相对应的写入提交队列条目的信息来创建所述读取提交队列条目。

7.根据权利要求1所述的方法,还包括基于在所述源存储装置中接收的用户数据写入来生成分散收集列表或位图,

其中,基于所述分散收集列表或所述位图来创建所述读取提交队列条目。

8.根据权利要求1所述的方法,还包括:

由所述存储设备将所述数据从所述源存储装置复制到所述存储设备的高速缓存中;或者

致使所述数据存储在主机上。

9.根据权利要求1所述的方法,还包括由所述存储设备创建写入提交队列条目,以供目标实时迁移服务器或目标实时迁移存储设备处理。

10.一种用于管理数据迁移操作的存储设备,所述存储设备被配置为创建指示要从所述存储设备的源存储装置复制到目标存储装置的数据在所述源存储装置处的位置的读取提交队列条目,所述读取提交队列条目包括包含元数据的字段,所述元数据包括用于从所述源存储装置读取所述数据的信息。

11.根据权利要求10所述的存储设备,其中:

所述读取提交队列条目的字段是所述读取提交队列条目的多个条目中的一个条目;以及

所述多个条目包括用于从所述源存储装置读取所述数据的信息。

12.根据权利要求10所述的存储设备,其中,所述元数据包括数据指针或命令标识符。

13.根据权利要求10所述的存储设备,其中,使用来自与用户数据写入相对应的写入提交队列条目的信息来创建所述读取提交队列条目。

14.根据权利要求10所述的存储设备,其中,所述存储设备被配置为:

基于在所述源存储装置中接收的用户数据写入来生成分散收集列表或位图;以及

基于所述分散收集列表或所述位图来创建所述读取提交队列条目。

15.根据权利要求10所述的存储设备,被配置为创建写入提交队列条目,以供目标实时迁移服务器或目标实时迁移存储设备处理。

16.一种用于管理数据迁移操作的系统,所述系统包括:

主机;以及

存储设备,

其中,所述存储设备被配置为创建指示要从所述存储设备的源存储装置复制到目标存储装置的数据在所述源存储装置处的位置的读取提交队列条目,所述读取提交队列条目包括包含元数据的字段,所述元数据包括用于从所述源存储装置读取所述数据的信息。

17.根据权利要求16所述的系统,其中,所述存储设备被配置为:

向所述主机发送所述读取提交队列条目;

接收基于所述读取提交队列条目的命令;以及

基于所述命令从所述源存储装置读取所述数据。

18.根据权利要求16所述的系统,其中,所述存储设备被配置为基于所述读取提交队列条目存储在所述存储设备上而从所述源存储装置读取所述数据。

19.根据权利要求16所述的系统,其中,使用来自与用户数据写入相对应的写入提交队列条目的信息来创建所述读取提交队列条目。

20.根据权利要求16所述的系统,其中,所述存储设备被配置为创建写入提交队列条目,以供目标实时迁移服务器或目标实时迁移存储设备处理。

存储数据实时迁移的在存储设备处的预填充命令

技术领域

[0001] 根据本公开的一个或多个实施例的一个或多个方面涉及用于数据存储的系统和方法。

背景技术

[0002] 本背景技术章节旨在仅提供上下文,并且本章节中任何实施例或概念的公开内容不构成承认所述实施例或概念是现有技术。

[0003] 在数据存储的领域中,实时迁移(live migration)操作(或实时迁移过程)可以是指当源存储装置可能潜在地从源服务器接收用户数据读取或用户数据写入(例如,源存储装置是实时的)时将数据(例如,存储数据)从一个存储装置复制到另一存储装置(例如,从源存储装置复制到目标存储装置)的迁移操作或迁移过程。如本文所使用的,“实时迁移”组件(例如,“实时迁移服务器”或“实时迁移存储设备”等)是指可能涉及从源存储装置到目标存储装置的数据传递并且可能比系统的其他组件具有更高/附加特权(例如,用以访问系统内的数据的更高/附加特权)的组件。

[0004] 数据结构可以用于跟踪源存储装置内的要复制到目标存储装置的数据的位置。尽管与存储设备的存储容量相比,与数据结构相关联的元数据相对较小,但是处理数据结构的操作可能消耗相对大量的主机处理资源。例如,主机可以处理数据结构的元数据以创建用于从源存储装置中提取出数据以进行实时迁移的命令(例如,提交队列条目(submission queue entry,SQE))。

[0005] 相应地,可能存在适合于改进用于管理或执行数据迁移(例如,实时迁移)的元数据的通信的方法、设备和系统。

发明内容

[0006] 本公开的一个或多个实施例的各方面涉及计算机存储系统,并且提供了对用于处理提交队列条目以管理数据迁移的系统和方法的改进。

[0007] 根据本公开的一个或多个实施例,提供了一种用于管理数据迁移操作的方法,包括由存储设备创建指示要从存储设备的源存储装置复制到目标存储装置的数据在源存储装置处的位置的读取提交队列条目,该读取提交队列条目包括包含元数据的字段,该元数据包括用于从源存储装置读取数据的信息。

[0008] 该方法还可以包括:由存储设备向主机发送读取提交队列条目,由存储设备接收基于读取提交队列条目的命令,以及由存储设备基于该命令从源存储装置读取数据。

[0009] 该方法还可以包括基于读取提交队列条目存储在存储设备上而从源存储装置读取数据。

[0010] 读取提交队列条目的字段可以是读取提交队列条目的多个条目中的一个条目,并且该多个条目可以包括用于从源存储装置读取数据的信息。

[0011] 元数据可以包括数据指针或命令标识符。

[0012] 可以使用来自与用户数据写入相对应的写入提交队列条目的信息来创建读取提交队列条目。

[0013] 该方法还可以包括基于在源存储装置中接收的用户数据写入来生成分散收集列表或位图,其中,基于分散收集列表或位图来创建读取提交队列条目。

[0014] 该方法还可以包括:由存储设备将数据从源存储装置复制到存储设备的高速缓存,或者致使数据存储在主机上。

[0015] 该方法还可以包括:由存储设备创建写入提交队列条目,以供目标实时迁移服务器或目标实时迁移存储设备处理。

[0016] 根据本公开的一个或多个其他实施例,提供了一种用于管理数据迁移操作的存储设备,该存储设备被配置为:创建指示要从存储设备的源存储装置复制到目标存储装置的数据在源存储装置处的位置的读取提交队列条目,该读取提交队列条目包括包含元数据的字段,该元数据包括用于从源存储装置读取数据的信息。

[0017] 读取提交队列条目的字段可以是读取提交队列条目的多个条目中的一个条目,并且该多个条目可以包括用于从源存储装置读取数据的信息。

[0018] 元数据可以包括数据指针或命令标识符。

[0019] 可以使用来自与用户数据写入相对应的写入提交队列条目的信息来创建读取提交队列条目。

[0020] 存储设备可以被配置为:基于在源存储装置中接收的用户数据写入来生成分散收集列表或位图,并且基于分散收集列表或位图来创建读取提交队列条目。

[0021] 存储设备可以被配置为:创建写入提交队列条目,以供目标实时迁移服务器或目标实时迁移存储设备处理。

[0022] 根据本公开的一个或多个其他实施例,提供了一种用于管理数据迁移操作的系统,包括主机和存储设备,其中,存储设备被配置为创建指示要从存储设备的源存储装置复制到目标存储装置的数据在源存储装置处的位置的读取提交队列条目,该读取提交队列条目包括包含元数据的字段,该元数据包括用于从源存储装置读取数据的信息。

[0023] 存储设备可以被配置为:向主机发送读取提交队列条目,接收基于读取提交队列条目的命令,以及基于该命令从源存储装置读取数据。

[0024] 存储设备可被配置为基于读取提交队列条目存储在存储设备上而从源存储装置读取数据。

[0025] 可以使用来自与用户数据写入相对应的写入提交队列条目的信息来创建读取提交队列条目。

[0026] 存储设备可以被配置为:创建写入提交队列条目,以供目标实时迁移服务器或目标实时迁移存储设备处理。

附图说明

[0027] 参考以下附图描述了本公开的非限制性和非穷尽性实施例,其中,除非另有说明,否则在各个视图中,相同的附图标记指代相同的部件。

[0028] 图1是描绘根据本公开的一个或多个实施例的用于使用实时迁移服务器创建提交队列条目来管理数据迁移的系统的系统图。

[0029] 图2是描绘根据本公开的一个或多个实施例的用于使用实时迁移存储设备创建提交队列条目来管理数据迁移的系统的系统图。

[0030] 图3A是描绘根据本公开的一个或多个实施例的作为位图的数据结构的图。

[0031] 图3B是描绘根据本公开的一个或多个实施例的作为分散收集列表的数据结构的图。

[0032] 图3C是描绘根据本公开的一个或多个实施例的读取提交队列条目的字段的图。

[0033] 图3D是描绘根据本公开的一个或多个实施例的写入提交队列条目的字段的图。

[0034] 图4是根据本公开的一个或多个实施例的示例数据队列。

[0035] 图5是描绘根据本公开的一个或多个实施例的用于使用实时迁移存储设备创建提交队列条目来管理数据迁移的方法的示例操作的流程图。

[0036] 贯穿附图的几个视图,相对应的附图标记指示相对应的组件。技术人员将会理解,附图中的元件是为了简单和清楚而示出的,并不一定是按比例绘制的。例如,附图中的一些元件和区域的尺寸可能相对于其他元件和区域被夸大,以帮助提高对各种实施例的清楚和理解。此外,可能没有示出与实施例的描述不相关的常见但公知的元件和部件,以便于更清楚地理解这些各种实施例并使描述清楚。

具体实施方式

[0037] 通过参考一个或多个实施例的详细描述和附图,可以更容易地理解本公开的各方面和实现本公开的方法。在下文中,将参考附图更详细地描述实施例。然而,所描述的实施例可以以各种不同的形式体现,并且不应被解释为仅限于本文所示的实施例。相反,这些实施例是作为示例提供的,使得本公开将是彻底和完整的,并且将向本领域技术人员充分传达本公开的各方面。相应地,对于本领域普通技术人员来说,对完整理解本公开的各方面和特征不必要的过程、元素和技术的描述可以省略。

[0038] 除非另有说明,否则在整个附图和书面描述中,相同的附图标记、字符或其组合表示相同的元件,因此,将不再重复其描述。此外,为了使描述清楚,可能没有示出与实施例的描述无关的部分。在附图中,为了清楚起见,可能夸大了元件和区域的相对尺寸。

[0039] 在详细描述中,出于解释的目的,阐述了许多具体细节以提供对各种实施例的透彻理解。然而,显而易见的是,各种实施例可以在没有这些具体细节的情况下或者利用一个或多个等效布置来实践。

[0040] 将会理解,尽管术语“第零”、“第一”、“第二”、“第三”等可以在本文中用于描述各种元件、组件、区域和/或区段,但是这些元件、组件、区域和/或区段不应受这些术语的限制。这些术语用于将一个元件、组件、区域或区段与另一元件、组件、区域或区段区分。因此,在不脱离本公开的精神和范围的情况下,下面描述的第一元件、组件、区域或区段可以称为第二元件、组件、区域或区段。

[0041] 应当理解,当一个元件或组件被称为“在另一元件或组件上”、“连接到另一元件或组件”或者“耦合到另一元件或组件”时,它可以直接在另一元件或组件上、直接连接到另一元件或组件或者直接耦合到另一元件或组件,或者可以存在一个或多个中介元件或组件。然而,“直接连接/直接耦合”是指一个组件直接连接或耦合另一组件而无需中间组件。与此同时,描述组件之间关系的其他表述(诸如“在……之间”、“紧接在……之间”或“邻近于”和

“直接邻近于”)可以被类似地解释。此外,还应理解,当一个元件或组件被称为在两个元件或组件“之间”时,它可以是两个元件或组件之间的唯一元件或组件,或者也可以存在一个或多个中介元件或组件。

[0042] 本文使用的术语仅仅是为了描述特定实施例的目的,而不是旨在限制本发明。如本文所使用的,单数形式“一”和“一个”也旨在包括复数形式,除非上下文另有清楚指示。还应理解,当在本说明书中使用时,术语“包括”、“包含”、“具有”、“带有”、“涵盖”和“含有”指定所陈述的特征、整数、步骤、操作、元件和/或组件的存在,但是不排除一个或多个其他特征、整数、步骤、操作、元件、组件和/或其组合的存在或添加。如本文所使用的,术语“或”和“和/或”中的每一个包括一个或多个相关列出项目的任何和所有组合。

[0043] 出于本公开的目的,诸如“……中的至少一个”的表述在元素列表之前时修饰整个元素列表,而不修饰列表中的单个元素。例如,“X、Y和Z中的至少一个”和“从由X、Y和Z组成的组中选择的至少一个”可以被解释为仅X、仅Y、仅Z、或者X、Y和Z中的两个或更多个的任何组合,诸如XYZ、XYY、YZ和ZZ。

[0044] 如本文所使用的,术语“实质上”、“大约”、“近似”和类似术语用作近似术语,而不是程度术语,并且旨在说明本领域普通技术人员将认识到的测量值或计算值的固有偏差。如本文所使用的,“大约”或“近似”包括所陈述的值,并且意味着在考虑到所讨论的测量和与特定量的测量相关联的误差(即,测量系统的限制)的情况下,本领域普通技术人员所确定的特定值在可接受的偏差范围内。例如,“大约”可以意味着在一个或多个标准偏差内,或者在所陈述的值的±30%、20%、10%、5%内。此外,当描述本公开的实施例时使用“可以”是指“本公开的一个或多个实施例”。

[0045] 当一个或多个实施例可以不同地实施时,特定过程次序可以不同于所描述的次序来执行。例如,两个连续描述的过程可以实质上同时执行,或者以与所描述的次序相反的次序执行。

[0046] (例如,在本文包括的任何系统图中)所描述的任何组件或任何组件组合可以用于执行本文包括的任何流程图的一个或多个操作。此外,(i)操作仅仅是示例,并且可以设计未明确覆盖的各种附加操作,并且(ii)操作的时间次序可以变化。

[0047] 可以使用任何合适的硬件、固件(例如,专用集成电路)、软件或者软件、固件和硬件的组合来实施根据本文描述的本公开的实施例的电子或电气设备和/或任何其他相关设备或组件。例如,这些设备的各种组件可以形成在一个集成电路(integrated circuit,IC)芯片上或分离的IC芯片上。此外,这些设备的各种组件可以在柔性印刷电路膜、载带封装(tape carrier package,TCP)、印刷电路板(printed circuit board,PCB)上实施,或者形成在一个衬底上。

[0048] 此外,这些设备的各种组件可以是运行在一个或多个计算设备中的一个或多个处理器上的进程或线程,其执行计算机程序指令并与其他系统组件交互以执行本文描述的各种功能。计算机程序指令被存储在存储器中,例如,存储器可以在使用标准存储器设备(诸如随机存取存储器(random-access memory,RAM))的计算设备中实施。例如,计算机程序指令也可以被存储在其他非暂时性计算机可读介质(诸如CD-ROM、闪存驱动器等)中。此外,本领域技术人员应当认识到,在不脱离本公开的实施例的精神和范围的情况下,各种计算设备的功能可以被组合或集成到单个计算设备中,或者特定计算设备的功能可以被分布在一

个或多个其他计算设备上。

[0049] 除非另外定义,否则本文使用的所有术语(包括技术术语和科学术语)具有与本发明构思所属领域的普通技术人员通常理解的含义相同的含义。还应理解,术语(诸如在常用词典中定义的那些)应当被解释为具有与它们在相关领域和/或本说明书的上下文中的含义一致的含义,并且不应以理想化或过于正式的意义来解释,除非本文如此明确定义。

[0050] 如上所述,在数据存储的领域中,实时迁移操作或实时迁移过程可以是指在源存储装置可能潜在地从源服务器(例如,从可能被指派了比管理程序(或实时迁移服务器)更低的监管特权的主机)接收用户数据读取或用户数据写入时(例如,在源存储装置是实时的时)将数据(例如,存储数据)从源存储装置复制到目标存储装置的操作或过程(其可以称为迁移过程)。

[0051] 例如,数据中心所有者、运营商和/或销售计算资源的软件平台运营商可以实施实时迁移策略,其中,当源存储装置处的数据可能被改变时(例如,当源存储装置处的数据可能基于用户数据写入的接收而被更新时),可以将数据从源存储装置移动到新的目标存储装置。实时迁移策略可以涉及将命名空间(例如,存储装置中的一组对象)从源控制器/存储装置移动到目标控制器/存储装置。

[0052] 诸如位图和/或分散收集列表(例如,向量列表,其中每个向量给出整体读取或写入请求中的一个分段的位置和长度)之类的数据结构可以用于跟踪源存储装置内要复制到目标存储装置的数据的位置。例如,可能具有大存储容量(例如,8太字节(TB))的整个源存储驱动器可以在一个或多个数据结构中具有由显著较小尺寸的跟踪元数据表示的每个逻辑块地址/LBA(例如,4千字节(KB)),以指示相对应的LBA是否包括应当从源存储装置复制到目标存储装置的数据。

[0053] 如下面将参考附图进一步详细讨论的,实时迁移控制器(例如,源实时迁移控制器)可以通过创建跟踪元数据来协助实时迁移服务器(例如,源实时迁移服务器)管理实时迁移操作。例如,创建跟踪元数据可以包括创建作为临时通信机制的数据结构(例如,位图或分散收集列表),该数据结构用于标识在将数据从源存储装置复制到目标存储装置的实例的第一次迭代发生的时间段期间在源存储装置内写入的用户数据写入的位置。也就是说,在包含先前从源存储装置复制到目标存储装置的数据的源存储装置的位置处接收到的用户数据写入可能致使源存储装置数据改变。

[0054] 例如,源存储装置中与致使源存储装置数据改变的用户数据写入相对应的位置(例如,LBA)可以称为包含“脏数据”的“脏页”。实时迁移控制器可以跟踪具有数据结构的脏页的位置,并且可以将数据结构发送到实时迁移服务器,以在下一次复制迭代中将脏数据从源存储装置复制到目标存储装置(例如,用源存储装置中的当前/更近的数据来更新目标存储装置中的数据)。

[0055] 尽管与存储设备的存储容量相比,与数据结构相关联的元数据相对较小,但是处理数据结构可能消耗大量的主机处理资源。例如,主机可以处理数据结构的元数据以创建命令(例如,SQE)。SQE可以包括与要发送到实时迁移存储设备的读取命令相关联的读取SQE,以从源存储装置中提取数据以进行实时迁移。SQE还可以包括与要发送到目标存储装置的写入命令相关联的写入SQE,以用于将从源存储装置中提取的数据写入目标存储装置。在处理数据结构以创建SQE时,实时迁移服务器可能执行解析成数据结构的冗余步骤,然后

可以退出数据结构。

[0056] 根据本公开的用于管理实时迁移的系统可以使用源实时迁移存储设备来创建和预填充SQE(例如,读取命令、写入命令等),从而通过从源实时迁移服务器移除与创建和填充SQE相关联的工作负荷以及通过将工作负荷转移到源实时迁移存储设备来提高整体系统性能。通过在源实时迁移存储设备处创建和预填充SQE的一个或多个字段,存储系统可以实现实时迁移数据的更快的读取性能,从而导致更快的整体数据迁移。

[0057] 在一个或多个实施例中,源实时迁移存储设备可以基于从源实时迁移服务器接收到由源实时迁移存储设备创建和预填充的SQE而从源存储装置读取数据。例如,源实时迁移服务器可以(i)完成填充源实时迁移存储设备在创建SQE时未填充的SQE的字段,和/或(ii)覆写源实时迁移存储设备在创建SQE时预填充的SQE的一些字段。

[0058] 在一个或多个其他实施例中,代替源实时迁移存储设备创建SQE并将其发送到源实时迁移服务器、然后等待从源实时迁移服务器接收命令,源实时迁移存储设备可以具有来自SQE的足够信息以自动从源存储装置检索数据(例如,无需等待从源实时迁移服务器接收命令)。

[0059] 在一个或多个实施例中,代替源实时迁移存储设备以位图的形式或以分散收集列表的形式创建数据结构,源实时迁移存储设备可以使用与源服务器提交队列相对应的写入SQE(和相对应的用户数据写入)来生成SQE。例如,源实时迁移存储设备可以被配置为将与脏页相对应的写入SQE转换成读取SQE。例如,源实时迁移存储设备可以被配置为从写入SQE的一个或多个字段中提取元数据,以预填充相对应的读取SQE的一个或多个字段。

[0060] 在一个或多个实施例中,源实时迁移存储设备可以将从源存储装置复制的脏数据存储在源实时迁移存储设备的高速缓存中。例如,源实时迁移存储设备可以在存储设备上创建数据队列,并且可以稍后(例如,异步地)将数据队列发送到源实时迁移服务器。

[0061] 在一个或多个实施例中,源实时迁移存储设备可以存储从源存储装置复制到源实时迁移服务器上的缓冲区的脏数据。例如,源实时迁移存储设备可以在源实时迁移服务器上创建数据队列。在一个或多个实施例中,可以在来自源的读取SQE和针对目标的写入SQE中都使用(例如,标识)数据队列的位置(或多个位置)。如果填写(例如,预填充)写入SQE,则目标实时迁移服务器可以建立匹配的存储器地址范围,以用于在源实时迁移服务器与目标实时迁移服务器之间复制数据队列。

[0062] 在一个或多个实施例中,源实时迁移存储设备还可以创建写入SQE,以供目标实时迁移服务器或目标实时迁移存储设备在将数据写入目标存储装置时处理。

[0063] 图1是描绘根据本公开的一个或多个实施例的用于使用源实时迁移服务器100b创建提交队列条目(SQE)来管理数据迁移的系统1的系统图。

[0064] 参考图1,系统1可以包括源主机系统100S和目标主机系统100T。源主机系统100S可以连接到源存储装置200。目标主机系统100T可以连接到目标存储装置300。源主机系统100S和目标主机系统100T可以与中央处理单元(central processing unit,CPU)相关联,或者可以包括CPU。源存储装置200和目标存储装置300可以包括逻辑块地址(logical block addres,LBA)(例如,LBA0至LBA 15,如图1中作为简化示例的源存储装置200内所描绘的)。LBA可以与源存储装置200和目标存储装置300中用于存储数据(例如,用户数据)的物理位置相关联。

[0065] 目标主机系统100T可以包括目标服务器100c(例如,目标虚拟机(virtual machine,VM))。目标服务器100c可以经由通信链路20连接到目标存储设备250c。通信链路20可以包括各种技术或者可以通过各种技术来实施(例如,高速外围组件互连(peripheral component interconnect express,PCIe)、基于PCIe的高速非易失性存储器(nonvolatile memory express,NVMe)、基于交换结构的NVMe(NVMe over Fabrics,NVMe-oF)等)。

[0066] 目标存储设备250c可以包括目标存储控制器210c和目标控制器存储器220c。目标控制器存储器220c可以包括或者可以是RAM(例如,静态随机存取存储器(static random-access memory,SRAM)或动态随机存取存储器(dynamic random-access memory,DRAM))。目标存储控制器210c可以包括嵌入式逻辑,或者可以通过嵌入式逻辑来实施。嵌入式逻辑可以嵌入在目标存储设备250c内。嵌入式逻辑可以使目标存储控制器210c能够处理来自服务器(例如,目标服务器100c)的用以将数据复制/写入目标存储装置300的请求。

[0067] 目标存储设备250c可以包括目标存储装置300(或目标存储装置300的一部分)。例如,目标存储装置300可以包括目标存储器330。目标存储器330可以包括或者可以是长期存储器。例如,目标存储器330可以包括非易失性存储器和/或存储器层(包括易失性和非易失性存储器),并且可以对应于一个或多个目标存储设备250c的长期存储器设备。例如,目标存储装置300可以是指由分布式存储系统(例如,虚拟化分布式存储系统)的一个或多个目标存储设备250c组成的长期存储器。目标存储设备250c可以是包括一个或多个底层目标存储设备250c(其可以是虚拟的或物理的)的固态驱动器(solid-state drive,SSD)。目标存储设备250c可以经由存储接口22连接到目标存储装置300。存储接口22可以包括或者可以是闪存接口(例如,NAND闪存接口)或远程接口(例如,通过以太网来实施的接口)。

[0068] 源主机系统100S可以包括源服务器100a。源服务器100a可以是主机,或者可以是源VM。源服务器100a可以服务来自用户应用的读取和写入请求。例如,用户可以使用应用(例如,文字处理应用)向源存储装置200发送用户数据写入UDW(例如,用户应用数据写入请求)或用户数据读取UDR(例如,用户应用数据读取请求)。读取/写入请求可以经由源存储设备250a发送到源存储装置200。在被发送到源存储装置200之前,来自源服务器100a的写入请求可以在源服务器100a处排队,以在源服务器提交队列SSSQ中包括一个或多个写入提交队列条目WSQE。类似地,在一个或多个实施例中,在被发送到源存储装置200之前,来自源服务器100a的读取请求可以在源服务器100a处排队,以在源服务器提交队列SSSQ中包括一个或多个读取提交队列条目RSQE。

[0069] 源服务器100a可以经由通信链路20连接到源存储设备250a。通信链路20可以包括各种技术或者可以通过各种技术来实施(例如,PCIe、基于PCIe的NVMe、NVMe-oF等)。

[0070] 源存储设备250a可以包括源存储控制器210a和源控制器存储器220a。源控制器存储器220a可以包括或者可以是RAM(例如,SRAM或DRAM)。源存储控制器210a可以包括用于处理来自服务器(例如,源服务器100a)的用以向源存储装置200复制/写入数据的请求的嵌入式逻辑(例如,嵌入在源存储设备250a内的软件或固件),或者可以通过该嵌入式逻辑来实施。源存储设备250a可以包括源存储装置200或其一部分。

[0071] 例如,源存储装置200可以包括源存储器230。源存储器230可以包括或者可以是长期存储器。例如,源存储器230可以包括非易失性存储器和/或存储器层(包括易失性和非易失性存储器)。源存储器230可以对应于一个或多个源存储设备250a的长期存储器设备。例

如,源存储装置200可以是指由分布式存储系统(诸如虚拟化分布式存储系统)的一个或多个源存储设备250a组成的长期存储器。源存储设备250a可以是包括一个或多个底层源存储设备250a的固态驱动器(SSD)。底层源存储设备可以是虚拟的或物理的。源存储设备250a可以经由存储接口22连接到源存储装置200。存储接口22可以包括或者可以是闪存接口(例如,NAND闪存接口)或远程接口(其可以通过以太网来实施)。

[0072] 源主机系统100S可以包括源实时迁移服务器100b。源实时迁移服务器100b可以是与源服务器100a或目标服务器100c相比被指派了更高/附加特权的服务器。特权可以对应于系统1内的数据访问。源实时迁移服务器100b可以被指派来管理用于将数据从源存储装置200复制到目标存储装置300的实时迁移操作(例如,实时迁移过程)。源实时迁移服务器100b可以包括源实时迁移服务器存储器120b。

[0073] 源实时迁移服务器存储器120b可以包括或者可以是RAM(例如,SRAM或DRAM)。源实时迁移服务器存储器120b可以用于存储用于管理(例如,监控和/或指导)实时迁移操作的元数据。源实时迁移服务器100b可以经由通信链路20连接到源实时迁移存储设备250b。通信链路20可以包括各种技术或者可以通过各种技术来实施(例如,PCIe、基于PCIe的NVMe、NVMe-oF等)。

[0074] 源实时迁移存储设备250b可以是与源存储设备250a或目标存储设备250c相比被指派了更高/附加特权的存储设备。所指派的特权可以对应于系统1内的数据访问。源实时迁移存储设备250b可以被指派来通过从源存储装置200读取数据来协助实时迁移操作。

[0075] 源实时迁移存储设备250b可以包括源实时迁移控制器存储器220b和源实时迁移控制器210b。

[0076] 源实时迁移控制器存储器220b可以包括或者可以是RAM(例如,SRAM或DRAM)。源实时迁移控制器存储器220b可以用于存储用于管理(例如,监控和/或指导)实时迁移操作的元数据。源实时迁移控制器存储器220b还可以包括用于存储(例如,临时存储)来自源存储装置200的数据的实时迁移控制器高速缓存222b。

[0077] 源实时迁移控制器210b可以包括嵌入式逻辑,或者可以通过嵌入式逻辑来实施,嵌入式逻辑可以是源实时迁移存储设备250b的一部分,用于协助源实时迁移服务器100b将数据从源存储装置200复制到目标存储装置300。

[0078] 源实时迁移存储设备250b可以经由存储接口22连接到源存储装置200。存储接口22可以包括或者可以是闪存接口(例如,NAND闪存接口)或远程接口(例如,通过以太网来实施)。

[0079] 源实时迁移服务器100b可以经由“实时迁移服务器至目标”链路10连接或链接到目标主机系统100T。取决于系统1是经由硬件(HW)、软件(SW)还是HW和SW的组合来实施,“实时迁移服务器至目标”链路10可以通过各种技术来实施。

[0080] 除了上述目标服务器100c之外,目标主机系统100T可以包括目标实时迁移服务器100d。目标实时迁移服务器100d可以是与源服务器100a或目标服务器100c相比被指派了更高/附加特权的服务器。特权可以对应于系统1内的数据访问。目标实时迁移服务器100d可以被指派来协助源实时迁移服务器100b管理实时迁移操作。目标实时迁移服务器100d可以包括目标实时迁移服务器存储器120d。

[0081] 目标实时迁移服务器存储器120d可以包括或者可以是RAM(例如,SRAM或DRAM)。目

标实时迁移服务器存储器120d可以用于存储用于管理(例如,监控和/或指导)实时迁移操作的元数据。目标实时迁移服务器100d可以经由通信链路20连接到目标实时迁移存储设备250d。通信链路20可以包括各种技术或者可以通过各种技术来实施(例如,PCIe、基于PCIe的NVMe、NVMe-oF等)。

[0082] 目标实时迁移存储设备250d可以是与源存储设备250a或目标存储设备250c相比被指派了更高/附加特权的存储设备。所指派的特权可以对应于系统1内的数据访问。目标实时迁移存储设备250d可以被指派来通过向目标存储装置300写入数据(例如,源自源存储装置200的数据)来协助实时迁移操作。

[0083] 目标实时迁移存储设备250d可以包括目标实时迁移控制器存储器220d。目标实时迁移控制器存储器220d可以包括或者可以是RAM(例如,SRAM或DRAM)。目标实时迁移控制器存储器220d可以用于存储用于管理(例如,监控和/或指导)实时迁移操作的元数据。

[0084] 目标实时迁移存储设备250d还可以包括目标实时迁移控制器210d。目标实时迁移控制器210d可以包括嵌入式逻辑,或者可以通过嵌入式逻辑来实施,嵌入式逻辑可以是目标实时迁移存储设备250d的一部分,用于协助目标实时迁移服务器100d向目标存储装置300写入数据。目标实时迁移存储设备250d可以经由存储接口22连接到目标存储装置300。存储接口22可以包括或者可以是闪存接口(例如,NAND闪存接口)或远程接口(例如,通过以太网来实施)。

[0085] 系统1的服务器和存储设备可以经由HW、SW或HW和SW的组合来实施。例如,在一些实施例中,“实时迁移服务器至目标”链路10可以是物理网络连接(例如,以太网)。在一些实施例中,一个或多个服务器(例如,源服务器100a、源实时迁移服务器100b、目标实时迁移服务器100d或目标服务器100c)可以是软件实体。例如,一个或多个服务器可以是由与一个或多个中央处理单元(CPU)相关联的管理程序所管理的虚拟机(VM)。

[0086] 同样,存储设备250a、250b、250c、250d和/或源存储装置200和/或目标存储装置300中的一个或多个可以通过HW和/或SW技术来虚拟化和实施。例如,存储设备250a、250b、250c、250d和/或源存储装置200和/或目标存储装置300中的一个或多个可以由物理存储设备的任何组合来提供。在一个或多个实施例中,如图1中的虚线所指示,源存储设备250a、源实时迁移存储设备250b和源存储装置200可以是源SSD的组件。类似地,目标存储设备250c、目标实时迁移存储设备250d和目标存储装置300可以是目标SSD的组件。在一个或多个其他实施例中,迁移可以是从第一虚拟机到第二虚拟机,注意,两个虚拟机可以由一个物理存储设备支持。在一个或多个实施例中,源实时迁移服务器100b和目标实时迁移服务器100d可以是相同的。例如,在同一物理存储设备内,迁移可以是从源实时迁移控制器到目标实时迁移控制器,或者从源命名空间到目标命名空间。应当理解,可以实施例如物理设备与虚拟设备之间的各种其他组合,并且不脱离本公开的精神和范围。

[0087] 源实时迁移服务器100b和源实时迁移存储设备250b,连同目标实时迁移服务器100d和目标实时迁移存储设备250d,可以通过传达指示与源存储装置200中的LBA相对应的数据状态的元数据来协调实时迁移过程的管理。

[0088] 例如,在实时迁移过程期间,在第一次复制迭代期间,源存储装置的区域可以从源存储装置200逐渐复制到目标存储装置300。源存储装置200的区域可以对应于具有要在第一次复制迭代期间复制的数据的一个或多个存储位置(例如,LBA或“页”)。例如,在第一次

复制迭代期间,存储位置可以是包含要从源存储装置200复制到目标存储装置300的“映射数据”的“映射页”。源存储装置200的区域可以接收用户数据写入UDW,这可以致使对已经从源存储装置200复制到目标存储装置300以进行实时迁移的一个或多个存储位置(例如,LBA或“页”)处的数据的改变。如上所述,这种存储位置可以称为包含“脏数据”的“脏页”。在第二次复制迭代中,可以将脏数据从源存储装置200复制到目标存储装置300,以保持目标存储装置300是更新的。

[0089] 仍然参考图1,在第一次复制迭代期间,源存储装置200可以按照以下次序接收对以下LBA的用户数据写入UDW:在第一时间的LBA 0(描绘为UDW1);在第二时间的LBA 8-10(描绘为UDW 2);在第三时间的LBA 14和15(描绘为UDW 3);在第四时间的LBA 2(描绘为UDW 4);以及在第五时间的LBA7-9(描绘为UDW5,并且被描绘为UDW2,5,用于在第二时间和第五时间重叠对LBA8和9的用户数据写入)。

[0090] 为了跟踪要在实时迁移过程的第二次复制迭代期间复制的脏页(例如,LBA0、2、7、8-10、14和15),实时迁移存储设备250b可以使用第一格式的数据结构DS来跟踪脏页位置(操作1001)。例如,数据结构DS可以以位图的形式(参见图3A)或以分散收集列表的形式(参见图3B)来提供。来自数据结构DS的信息可以用于生成读取提交队列条目RSQE,以使系统1能够从源存储装置200读取数据。来自数据结构DS的信息也可以用于生成写入提交队列条目WSQE,以使系统1能够向目标存储装置300写入数据。数据结构DS可以指示在实时迁移过程的第一次复制迭代期间映射页的位置。数据结构DS可以指示在实时迁移过程的第二次或以后的复制迭代期间脏页的位置。

[0091] 在一些系统1中,源实时迁移存储设备250b可以向源实时迁移服务器100b发送数据结构DS,以解析数据结构DS。源实时迁移服务器100b可以创建相对应的读取提交队列条目RSQE(操作1002)。在一些系统1中,源实时迁移服务器100b还可以创建包括写入提交队列条目WSQE的写入提交队列WSQ,写入提交队列条目WSQE用以指示目标实时迁移服务器100d和目标实时迁移存储设备250d如何将数据从源存储装置200写入目标存储装置300。

[0092] 可以将读取提交队列条目RSQE从源实时迁移服务器100b发送到源实时迁移存储设备250b(操作1003)。读取提交队列条目RSQE可以包括具有元数据的各种字段(参见图3C)(操作1004-1007),该元数据包括用以指示源实时迁移存储设备250b如何从源存储装置200读取数据以将数据从源存储装置200复制到目标存储装置300的信息。例如,取决于从源存储装置200到源实时迁移存储设备250b的复制迭代(操作1004),源实时迁移存储设备250b可以使用读取提交队列条目RSQE来读取脏数据或映射数据。

[0093] 在一些实施例中,源实时迁移存储设备250b可以在源实时迁移存储设备250b处创建数据队列DQ(操作1004)。数据队列DQ可以被存储(例如,临时存储)在源实时迁移存储设备高速缓存222b中(操作1004)。源实时迁移服务器100b可以从源实时迁移存储设备250b接收数据队列DQ(操作1005)。源实时迁移服务器100b可以向目标实时迁移服务器100d发送数据队列DQ和写入提交队列WSQ(操作1006)。

[0094] 目标实时迁移服务器100d可以向目标实时迁移存储设备250d发送写入提交队列条目WSQE,以在向目标存储装置300写入数据时使用(操作1007)。写入提交队列条目WSQE可以包括具有元数据的各种字段(参见图3D),该元数据包括用以指示目标实时迁移存储设备250d如何将数据从源存储装置200写入目标存储装置300的信息。读取提交队列条目RSQE的

字段可以很大程度上与写入提交队列条目WSQE的字段平行(例如,比较图3C与图3D)。

[0095] 图2是描绘根据本公开的一个或多个实施例的用于使用实时迁移存储设备创建提交队列条目来管理数据迁移的系统的系统图。

[0096] 参考图2,在一个或多个实施例中,源实时迁移存储设备250b可以创建SQE(例如,RSQE和/或WSQE),而不是源实时迁移服务器100b创建SQE(操作2001)。另外,实时迁移存储设备250b可以预填充SQE的一个或多个字段(例如,可以在其中生成元数据),以从源实时迁移服务器100b移除相关联的处理负担,并且加速实时迁移过程。

[0097] 在一个或多个实施例中,源实时迁移存储设备250b可以创建读取提交队列条目RSQE,可以预填充读取提交队列条目RSQE的一些或全部字段,并且可以向源实时迁移服务器100b发送读取提交队列条目RSQE。源实时迁移服务器100b可以:(i)完成填充读取提交队列条目RSQE的字段,并且(ii)将读取提交队列条目RSQE添加到读取提交队列RSQ(操作2002)。源实时迁移服务器100b可以通过读取提交队列条目RSQE向源实时迁移存储设备250b发送读取命令(操作2003)。源实时迁移存储设备250b可以使用读取提交队列条目RSQE从源存储装置200读取相对应的数据(操作2004)。

[0098] 替代地,在一个或多个实施例中,源实时迁移存储设备250b可以基于由源实时迁移存储设备250b创建的读取提交队列条目RSQE而从源存储装置200读取数据(操作2003)。例如,代替向源实时迁移服务器100b发送读取提交队列条目RSQE、然后等待从源实时迁移服务器100b接收命令(操作2001、2002、2003和2004),源实时迁移存储设备250b可以具有来自源实时迁移存储设备250b所创建的读取提交队列条目RSQE的足够的信息,以自动从源存储装置200中检索数据(操作2001、2003和2004)。

[0099] 在一个或多个实施例中,代替源实时迁移存储设备250b以位图形式或分散收集列表形式创建数据结构DS,源实时迁移存储设备250b可以使用与源服务器提交队列SSSQ相对应的写入提交队列条目WSQE(和相对应的用户数据写入UDW)来生成读取提交队列条目RSQE。例如,源实时迁移存储设备250b可以被配置为将与脏页相对应的写入提交队列条目WSQE转换成读取提交队列条目RSQE。换句话说,源实时迁移存储设备250b可以从来自源服务器100a的写入提交队列条目WSQE的一个或多个字段中提取元数据,并且基于所提取的元数据来创建读取提交队列条目RSQE。

[0100] 在一个或多个实施例中,源实时迁移存储设备250b可以将从源存储装置200复制的脏数据或映射数据存储在源实时迁移存储设备高速缓存222b中。例如,源实时迁移存储设备250b可以在存储设备250b上创建数据队列DQ,并且可以在稍后的时间将数据队列DQ发送到源实时迁移服务器100b(例如,异步地)。

[0101] 在一个或多个实施例中,源实时迁移存储设备250b可以存储从源存储装置200复制到源实时迁移服务器100b上的缓冲区的脏数据或映射数据。例如,源实时迁移存储设备250b可以将数据队列DQ发送到源实时迁移服务器100b上的缓冲区。

[0102] 在一个或多个实施例中,源实时迁移存储设备250b还可以创建写入提交队列条目WSQE,以供目标实时迁移服务器100d或者目标实时迁移存储设备250d在将源自源存储装置200的数据写入目标存储装置300时处理(操作2004-2007)。

[0103] 图3A是描绘根据本公开的一个或多个实施例的作为位图的数据结构的图。

[0104] 参考图3A,图1和图2的数据结构DS可以以位图的形式创建,其中每个LBA对应于位

图中的位位置。取决于复制的迭代,脏页或映射页可以由“1”位来标识。例如,对于脏页,上述用户数据写入UDW可以由与LBA 0、2、7-10、14和15相对应的1来描绘。对于映射页,映射页可以由与LBA 0、2、7-10、14和15相对应的1来描绘。在一个或多个实施例中,基于该简单示例,源实时迁移存储设备250b可以创建四个读取提交队列条目RSQE。例如:(i)与LBA 0相对应的第一读取提交队列条目RSQE1可以包括起始LBA(starting LBA,SLBA)字段条目0(例如,SLBA=0)和LBA数量(number of LBAs,NLBA)字段条目0(例如,NLBA=0)(参见图3C和3D);(ii)与LBA2相对应的第二读取提交队列条目RSQE2可以包括SLBA=2和NLBA=0;(iii)与LBA 7-10相对应的第三读取提交队列条目RSQE3可以包括SLBA=7和NLBA=3;以及(iv)与LBA 14和15相对应的第四读取提交队列条目RSQE4可以包括SLBA=14和NLBA=1。

[0105] 相应地,源实时迁移存储设备250b可以被配置为将作为位图的数据结构DS转换成一个或多个预填充的SQE(例如,一个或多个读取提交队列条目(read submission queue entry,RSQE)或者一个或多个写入提交队列条目(one or more write submission queue entry,WSQE))。在一个或多个实施例中,以类似于可以如何将读取数据或位图数据返回到源实时迁移服务器100b的方式,可以在数据缓冲区中将SQE提供给源实时迁移服务器100b。可选地,SQE可以被插入到由源实时迁移服务器100b选择的提交队列中。

[0106] 图3B是描绘根据本公开的一个或多个实施例的作为分散收集列表的数据结构的图。

[0107] 参考图3B,图1和图2的数据结构DS可以以分散收集列表的形式创建,其中取决于复制的迭代,每个用户数据写入UDW或映射页被指示为列表中的条目(例如,在日志中)。例如,分散收集列表可以包括与用户数据写入UDW相对应的五个日志条目或者与LBA 0、2、7-10、14和15相对应的上述映射页。如上所述,对于脏页,用户数据写入UDW可能按照以下顺序发生:在第一时间的LBA 0(描绘为UDW 1);在第二时间的LBA 8-10(描绘为UDW 2);在第三时间的LBA 14和15;在第四时间的LBA 2;以及在第五时间的LBA 7-9。

[0108] 在一个或多个实施例中,基于该简单示例,源实时迁移存储设备250b可以创建五个读取提交队列条目RSQE。例如:(i)与LBA 0相对应的第一读取提交队列条目RSQE1可以包括起始LBA(SLBA)字段条目0(例如,SLBA=0)和LBA数量(NLBA)字段条目0(例如,NLBA=0)(参见图3C和图3D);(ii)与LBA 8-10相对应的第二读取提交队列条目RSQE2可以包括SLBA=8和NLBA=2;(iii)与LBA 14和15相对应的第三读取提交队列条目RSQE3可以包括SLBA=14和NLBA=1;(iv)与LBA 2相对应的第四读取提交队列条目RSQE4可以包括SLBA=2和NLBA=0;以及(iv)与LBA7-9相对应的第五读取提交队列条目RSQE5可以包括SLBA=7和NLBA=2。

[0109] 相应地,源实时迁移存储设备250b可以被配置为将作为分散收集列表的数据结构DS转换成一个或多个预填充的SQE(例如,一个或多个读取提交队列条目(RSQE)或者一个或多个写入提交队列条目(WSQE))。

[0110] 在一个或多个实施例中,如果启用重叠检测,则源实时迁移存储设备250b可以被配置为组合第二读取提交队列条目RSQE2和第五读取提交队列条目RSQE5。例如,如果启用重叠检测,则源实时迁移存储设备250b可以被配置为创建包括与LBA 7-10相对应的SLBA=7和NLBA=3的读取提交队列条目。

[0111] 在一个或多个实施例中,可以在数据缓冲区中将提交队列条目SQE提供给源实时

迁移服务器100b,像命令数据一样。替代地,在一个或多个实施例中,提交队列条目SQE可以被放置在日志中,而不是日志条目中。替代地,在一个或多个实施例中,提交队列条目SQE可以被插入到由源实时迁移服务器100b选择的提交队列中。例如,源实时迁移服务器100b可以选择专门为这种提交队列条目指定的提交队列。例如,源实时迁移服务器100b可以选择针对读取SQE的读取提交队列和针对写入SQE的写入提交队列,使得读取SQE和写入SQE被组织在分离的提交队列中。在一个或多个实施例中,源实时迁移服务器100b可以选择检查和/或更改提交队列条目。在一个或多个实施例中,源实时迁移服务器100b可以选择何时提交门铃更新(例如,给源实时迁移存储设备250b的通知,其指示新的SQE已经被添加到提交队列以供处理)。

[0112] 图3C是描绘根据本公开的一个或多个实施例的读取提交队列条目的字段的图,图3D是描绘根据本公开的一个或多个实施例的写入提交队列条目的字段的图。

[0113] 参考图3C和图3D,如上所述,源实时迁移存储设备250b可以整体或部分地为源实时迁移服务器100b构造(例如,创建)针对读取命令(或写入命令)的SQE。可以根据NVMe协议来格式化SQE。操作码(OpCode)对于读取可以是02h,而对于写入可以是01h。因此,OpCode字段可以基于命令的类型来预填充。如果源实时迁移服务器100b提供缓冲空间,则源实时迁移存储设备250b可以填写元数据指针MD字段、物理区域页条目1(PRP1)字段或分散收集列表条目1(SGL1)字段、以及物理区域页条目2(PRP2)字段或分散收集列表条目2(SGL2)字段的数据指针。由源实时迁移存储设备250b执行的这些预填充操作可以针对读取、写入、不可校正写入、写入零和复制命令来执行。换句话说,预填充可以不仅限于读取和写入。

[0114] 在一个或多个实施例中,以下字段以及其他未列出的字段可以被设置为“关”,或者替代地,源实时迁移服务器100b可以初始化这些字段的偏好:强制单元访问(Force Unit Access,FUA);有限重试(Limited Retry,LR);融合;以及数据集管理。

[0115] 在一个或多个实施例中,可以根据写入命令来复制以下字段和其他未列出的字段,或者主机可以初始化这些字段的偏好:命名空间标识符(Namespace Identifier,NSID);PRP/SGL;预期逻辑块存储标签(Expected Logical Block Storage Tag,ELBST)/预期初始逻辑块参考标签(Expected Initial Logical Block Reference Tag,EILBRT);预期逻辑块应用标签(Expected Logical Block Application Tag,ELBAT);预期逻辑块应用标记掩码(Expected Logical Block Application Tag Mask,ELBATM);存储标签检查(storage tag check,STC);以及保护信息字段(Protection Information Field,PRINFO)。

[0116] 虽然上面讨论了特定字段的示例,但是本公开不限于此。例如,包括上述字段和/或其他字段的不同字段组合可以被设置为“关”,被初始化为偏好,或者根据写入命令来复制。

[0117] 在一个或多个实施例中,命令标识符(Cmd ID)字段可以留给源实时迁移服务器100b来填充。替代地,源实时迁移服务器100b可以分配命令标识符池以供源实时迁移存储设备250b使用。

[0118] 图4是根据本公开的一个或多个实施例的示例数据队列。

[0119] 参考图4,在一个或多个实施例中,来自图3C和图3D的被称为PRP1、PRP2、SGL1、SGL2和MD的SQE字段可以留给源实时迁移服务器100b来填充。替代地,源实时迁移服务器

100b可以在源实时迁移服务器存储器120b中分配数据队列DQ。在一个或多个实施例中,源实时迁移存储设备250b可以在数据队列DQ中分配内存空间。用于实时迁移过程的读取操作的数据可以被放置在数据队列DQ中,并且描述在哪里找到该数据的元数据可以在用于读取的读取SQE的PRP、SGL和MD字段中提供。如果用于读取的数据大于可以在读取SQE中描述的数据,则源实时迁移存储设备250b(参见图2)可以创建扩展的PRP和SGL结构(例如,附加的PRP和SGL结构)。用于读取的读取SQE的PRP和SGL区域可以用指向数据队列DQ上的扩展的PRP和SGL区域的元数据来填充。因此,扩展的PRP和SGL区域以及所有PRP和SGL区域的相关数据都可以从由源实时迁移服务器100b分配的数据队列DQ区域中分配出。图4描绘了源实时迁移存储设备250b以循环先进先出(first-in first-out,FIFO)方式管理数据队列DQ区域的一种实施方法。随着更多的现有内存空间被使用,源实时迁移存储设备250b可以在数据队列DQ的末端放置(例如,分配)更多的内存空间。当数据队列DQ的区域不再被使用时,源实时迁移服务器100b可以负责与源实时迁移存储设备250b进行通信。例如,源实时迁移服务器100b可以在将数据队列DQ释放回源实时迁移存储设备250b之前将数据队列DQ从源实时迁移服务器100b复制到目标实时迁移服务器100d。例如,数据队列DQ的第一区域406可以包括被填写并由要发送到目标实时迁移服务器100d的下一个SQE使用的PRP和MD;数据队列DQ的第二区域408可以被源实时迁移服务器100b用于复制到目标实时迁移服务器100d;并且与数据队列DQ的先进部分404和数据队列402的更新缓冲区释放部分402相对应的第三区域410可以被标识为空闲空间,从而由源实时迁移存储设备250b分配给数据队列DQ的末端。在一个或多个实施例中,源实时迁移服务器100b可以通过PCIe寄存器写入或管理命令将数据队列DQ上的内存空间返回给源实时迁移存储设备250b。

[0120] 类似地,由源实时迁移存储设备250b消耗的命令标识符可以由源实时迁移存储设备250b重用于新的SQE形成。例如,源实时迁移存储设备250b可以从源实时迁移服务器100b接收与读取SQE相对应的命令标识符。实时迁移存储设备250b可以从由读取SQE标识的源存储装置200读取数据。实时迁移存储设备250b可以通过将与读取SQE相对应的数据写入数据队列DQ中来完成读取SQE,并且使用与读取SQE相对应的命令标识符来形成写入SQE。相应地,可以通过实时迁移存储设备250b重用命令标识符创建写入SQE来减少履行读取命令和形成相对应的写入SQE的完成时间(例如,完成时间可以是零)。

[0121] 在一个或多个实施例中,与上述各种队列相对应的源实时迁移服务器存储器120b(参见图2)的一个或多个内存位置可以是连续的。例如,读取提交队列RSQ可以包括布置在连续内存位置中的RSQE;写入提交队列WSQ可以包括布置在连续内存位置中的WSQE;和/或数据队列DQ可以包括布置在连续内存位置中的数据队列条目(data queue entry,DQE)。替代地,在一个或多个实施例中,与上述各种队列(例如,RSQ、WSQ和DQ)相对应的源实时迁移服务器存储器120b(参见图2)的一个或多个内存位置可以是非连续的(例如,可以被中介内存位置分隔开)。例如,代替在33-52的连续范围内分配20个内存位置,可以在20-29和34-44的非连续范围内分配20个内存位置。20个内存位置也可以无序地布置。例如,10个内存位置35-44可以布置在剩余的10个内存位置20-29之前。在这种实施例中,源实时迁移服务器100b可以通过向源实时迁移存储设备250b发送完成通信来传达哪些数据队列DQ区域已经完成。实时迁移存储设备250b可以基于完成通信来跟踪数据队列DQ区域之间的间隙。

[0122] 图5是描述根据本公开的一个或多个实施例的用于使用实时迁移存储设备创建提

交队列条目来管理数据迁移的方法的示例操作的流程图。

[0123] 参考图5,方法5000可以包括一个或多个以下操作。源实时迁移存储设备250b(参见图2)可以基于数据结构DS来创建读取提交队列条目RSQE,数据结构DS是位图、分散收集列表或写入提交队列条目,并且指示要从源存储装置200复制到目标存储装置300的数据在源存储装置200处的位置(例如,LBA)(操作5001)。源实时迁移存储设备250b可以在读取提交队列条目的字段中创建(例如,生成)元数据,该元数据包括用于从源存储装置读取数据的信息(操作5002)。源实时迁移存储设备250b可以将数据从源存储装置200复制到源实时迁移存储设备250b的高速缓存222b,或者可以致使数据存储在源实时迁移服务器100b(例如,主机)上(操作5003)。源实时迁移存储设备250b可以创建写入提交队列条目WSQE,以供目标实时迁移服务器100d或目标实时迁移存储设备250d处理(操作5004)。应当理解,取决于复制的迭代,要从源存储装置200复制到目标存储装置300的数据可以是映射数据或脏数据。例如,映射数据将在实时迁移过程中的第一次(或初始的)复制迭代中被复制,而脏数据将在实时迁移过程中的第二次(或每次随后的)复制迭代中被复制。

[0124] 如本文所使用的,当“数据结构”被描述为“指示”数据的位置时,它可以意味着数据结构的元素(例如,数据结构是或包括位图的情况下的位图的位)指向或指示数据的位置。

[0125] 相应地,本公开的实施例为用于管理(例如,执行)数据迁移(例如,实时迁移)的元数据的通信提供了改进和优点。通过预填充命令,可以通过在存储设备(例如,实时迁移存储设备)处执行原本可能由主机(例如,实时迁移服务器)执行的任务来加速实时迁移过程。

[0126] 本公开的示例实施例可以扩展到以下陈述,但不限于此:

[0127] 陈述1:一种示例方法包括由存储设备创建指示要从存储设备的源存储装置复制到目标存储装置的数据在源存储装置处的位置的读取提交队列条目,该读取提交队列条目包括包含元数据的字段,该元数据包括用于从源存储装置读取数据的信息。

[0128] 陈述2:一种示例方法包括根据陈述1所述的方法,还包括:由存储设备向主机发送读取提交队列条目,由存储设备接收基于读取提交队列条目的命令,以及由存储设备基于该命令从源存储装置读取数据。

[0129] 陈述3:一种示例方法包括根据陈述1和2中任一项所述的方法,还包括基于读取提交队列条目存储在存储设备上而从源存储装置读取数据。

[0130] 陈述4:一种示例方法包括根据陈述1-3中任一项所述的方法,其中,读取提交队列条目是读取提交队列条目的多个条目中的一个条目,并且该条目包括用于从源存储装置读取数据的信息。

[0131] 陈述5:一种示例方法包括根据陈述1-4中任一项所述的方法,其中,元数据包括数据指针或命令标识符。

[0132] 陈述6:一种示例方法包括根据陈述1-5中任一项所述的方法,其中,使用来自与用户数据写入相对应的写入提交队列条目的信息来创建读取提交队列条目。

[0133] 陈述7:一种示例方法包括根据陈述1-5中任一项所述的方法,还包括基于在源存储装置中接收的用户数据写入来生成分散收集列表或位图,其中,基于分散收集列表或位图来创建读取提交队列条目。

[0134] 陈述8:一种示例方法包括根据陈述1-7中任一项所述的方法,还包括:由存储设备

将数据从源存储装置复制到存储设备的高速缓存,或者致使数据存储在主机上。

[0135] 陈述9:一种示例方法包括根据陈述1-8中任一项所述的方法,还包括:由存储设备创建写入提交队列条目,以供目标实时迁移服务器或目标实时迁移存储设备处理。

[0136] 陈述10:一种用于执行根据陈述1-9中任一项所述的方法的示例设备包括控制器和存储器。

[0137] 陈述11:一种用于执行根据陈述1-9中任一项所述的方法的示例系统包括主机和存储设备。

[0138] 虽然已经参考本文描述的实施例具体示出和描述了本公开的实施例,但是本领域普通技术人员将理解,在不脱离如在所附权利要求及其等同物中阐述的本公开的精神和范围的情况下,可以在形式和细节上进行各种改变。

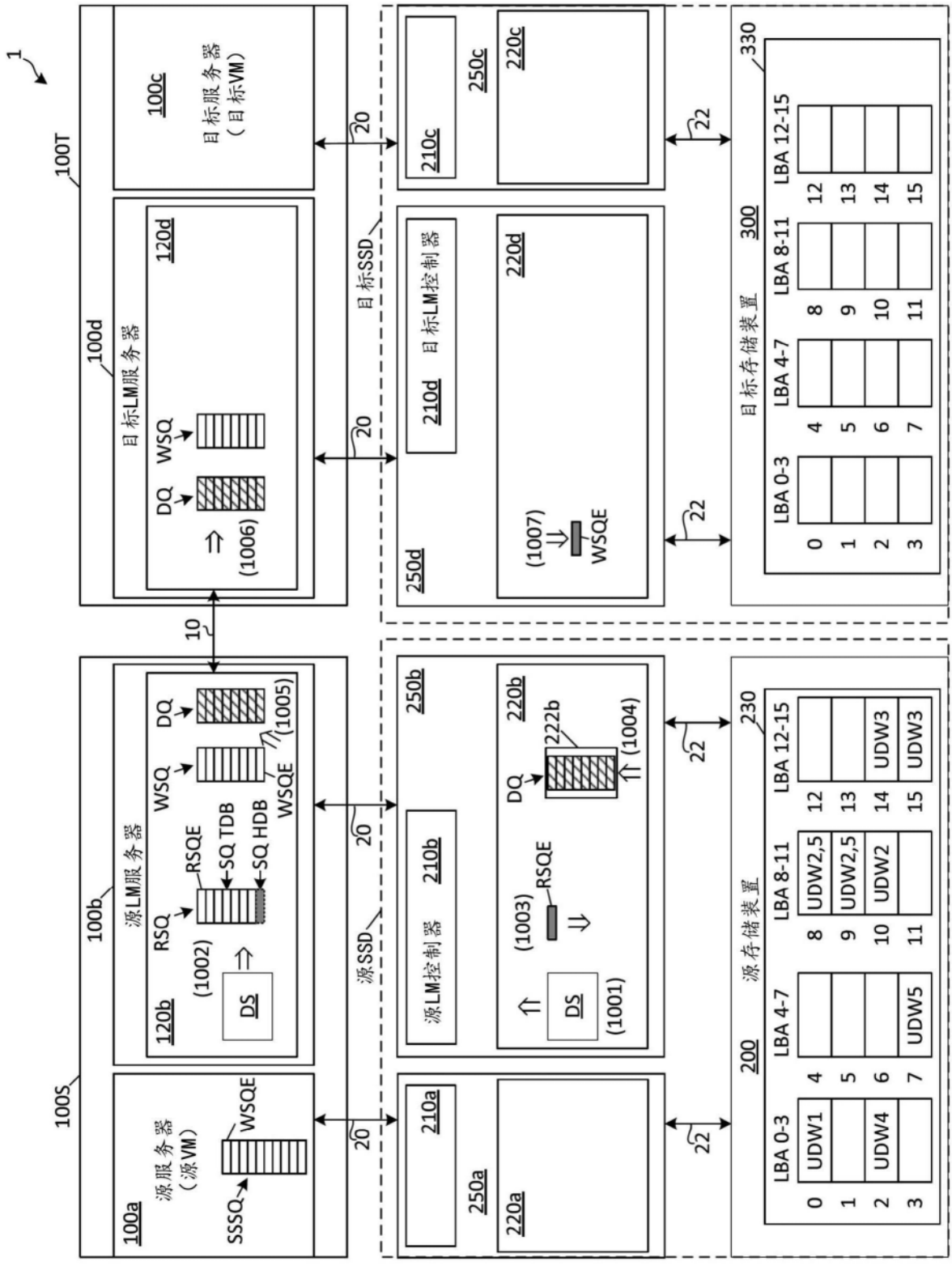

1
100T
100c
目标服务器（目标VM）
100d
目标LM服务器
120d
WSQ
DQ
(1006)
目标SSD
210c
250c
220c
210d
目标LM控制器
250d
220d
(1007)
WSQE
目标存储装置
300
330
LBA 0-3
LBA 4-7
LBA 8-11
LBA 12-15
10
20
22
100S
100b
源LM服务器
120b
RSQ
(1002)
DS
RSQE
SQ TDB
SQ HDB
WSQE
(1005)
源SSD
250b
源LM控制器
210b
220b
DQ
222b
(1004)
(1003)
(1001)
100a
源服务器（源VM）
SSSQ
210a
250a
220a
200
源存储装置
230
UDW1
UDW4
UDW5
UDW2,5
UDW2
UDW3

图1

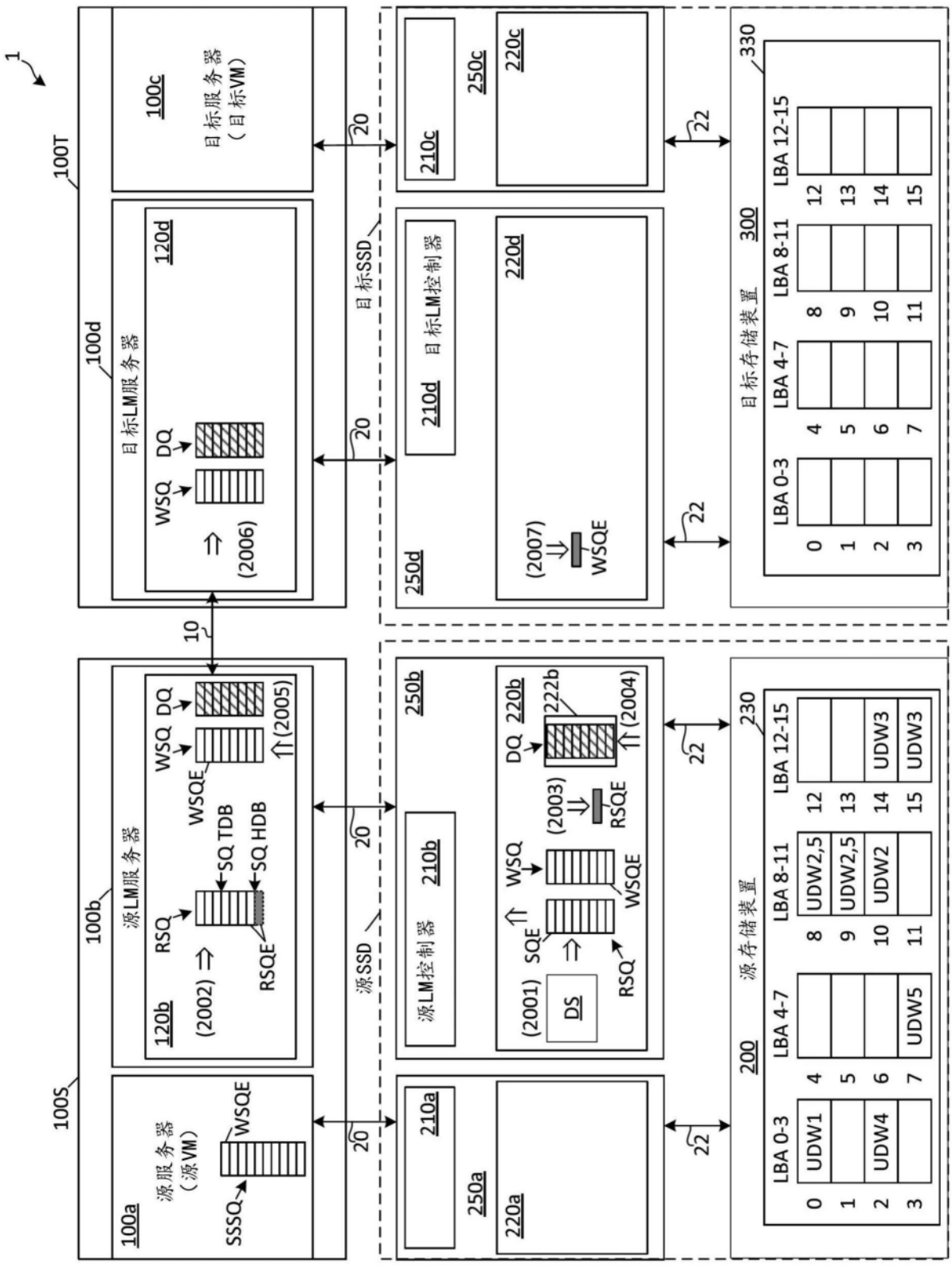

1
100T
100c
目标服务器（目标VM）
100d
目标LM服务器
120d
DQ
WSQ
(2006)
20
目标SSD
210c
250c
220c
210d
目标LM控制器
250d
220d
(2007)
WSQE
22
330
目标存储装置
300
LBA 0-3
LBA 4-7
LBA 8-11
LBA 12-15
10
100b
源LM服务器
120b
RSQ
(2002)
RSQE
SQ TDB
SQ HDB
WSQE
(2005)
源SSD
源LM控制器
210b
250b
220b
222b
(2004)
(2003)
RSQE
SQE
(2001)
DS
230
源存储装置
200
100S
100a
源服务器（源VM）
SSSQ
WSQE
210a
250a
220a
UDW1
UDW4
UDW5
UDW2,5
UDW2
UDW3

图2

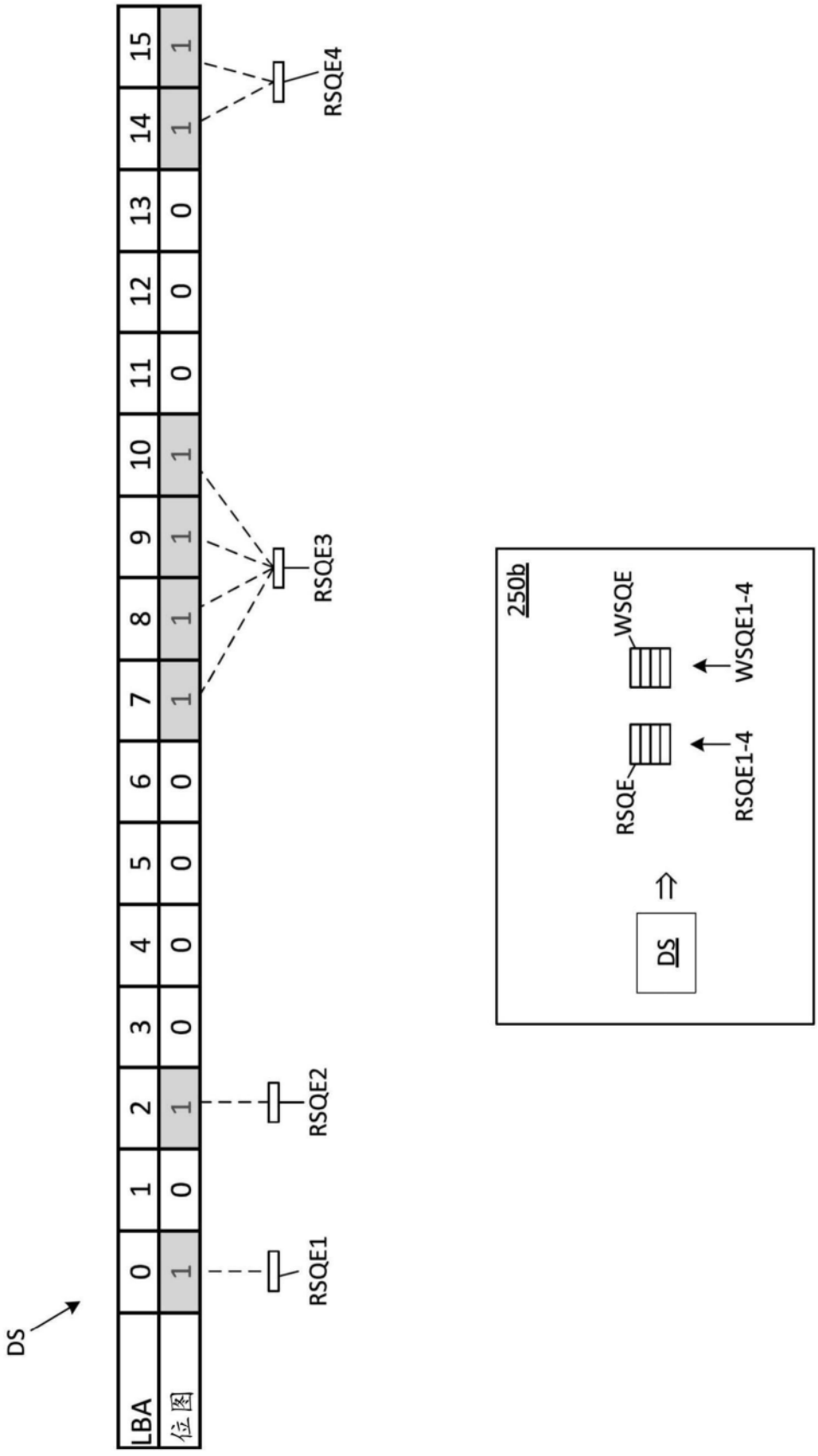

LBA
位图
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
1
0
1
0
0
0
0
1
1
1
1
0
0
0
1
1
DS
RSQE1
RSQE2
RSQE3
RSQE4
250b
DS
RSQE
RSQE1-4
WSQE
WSQE1-4

图3A

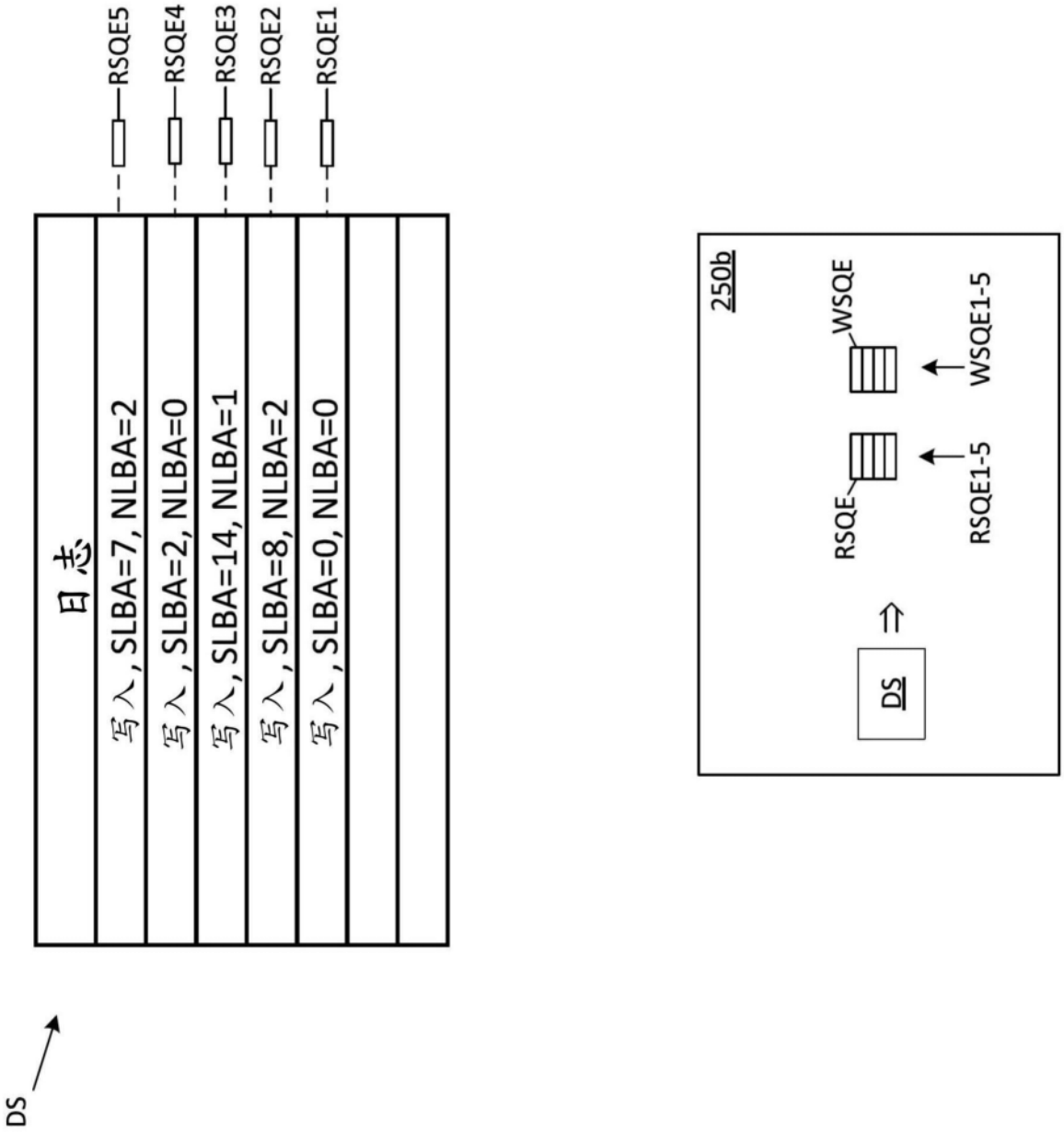

RSQE5
RSQE4
RSQE3
RSQE2
RSQE1
日志
写入, SLBA=7, NLBA=2
写入 , SLBA=2, NLBA=0
写入, SLBA=14, NLBA=1
写入, SLBA=8, NLBA=2
写入, SLBA=0, NLBA=0
DS
250b
WSQE
WSQE1-5
RSQE
RSQE1-5
DS

图3B

RSQE

读取命令																
	位															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
DWORD 0	OpCode								融合						PRP/SGL	
	Cmd ID															
DWORD 1	NS ID															
DWORD 2	ELBST/EILBRT															
DWORD 3																
DWORD 4	MD 元数据指针															
DWORD 5																
DWORD 6	PRP1 或 SGL1															
DWORD 7	PRP条目1（或SGL部分1）															
DWORD 8	PRP2 或 SGL2															
DWORD 9	PRP条目2（或SGL部分2）															
DWORD 10	SLBA 起始LBA															
DWORD 11																
DWORD 12	NLBA LBA数量															
									STC			PRINFO			FUA	LR
DWORD 13	数据集管理															
DWORD 14	ELBST/EILBRT															
DWORD 15	ELBAT															
	ELBATM															

图3C

WSQE

写入命令																
	位															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
DWORD 0	OpCode								融合						PRP/SGL	
	Cmd ID															
DWORD 1	NS ID															
DWORD 2	LBST/ILBRT															
DWORD 3																
DWORD 4	MD 元数据指针															
DWORD 5																
DWORD 6	PRP1 或 SGL1 PRP条目1（或SGL部分1）															
DWORD 7																
DWORD 8	PRP2 或 SGL2 PRP条目2（或SGL部分2）															
DWORD 9																
DWORD 10	SLBA 起始LBA															
DWORD 11																
DWORD 12	NLBA LBA数量															
	指导类型								STC			PRINFO			FUA	LR
DWORD 13	数据集管理															
	指导特定															
DWORD 14	LBST/ILBRT															
DWORD 15	LBAT															
	LBATM															

图3D

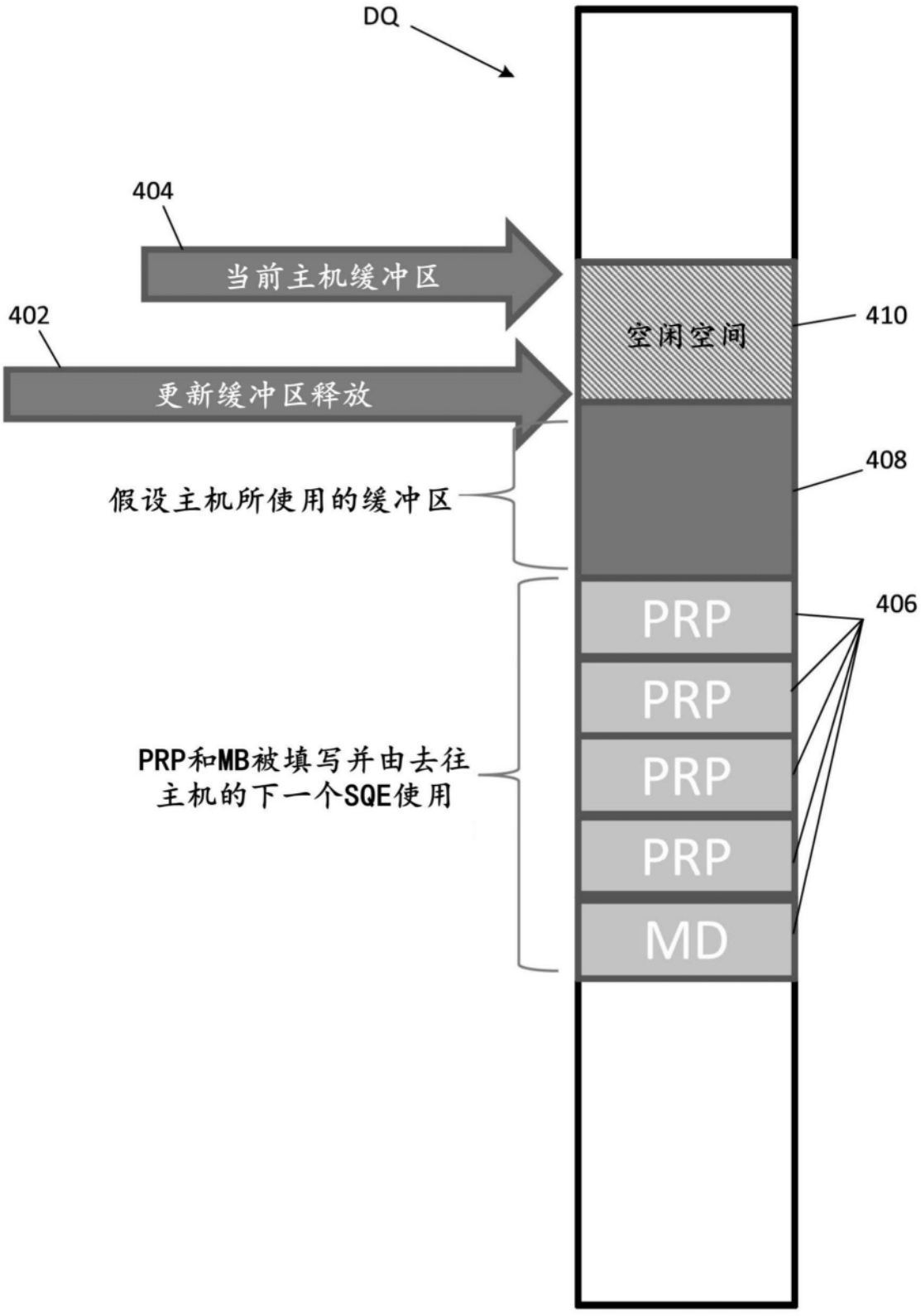

DQ
404
当前主机缓冲区
402
更新缓冲区释放
空闲空间
410
408
假设主机所使用的缓冲区
406
PRP
PRP
PRP
PRP
MD
PRP和MB被填写并由去往
主机的下一个SQE使用

图4

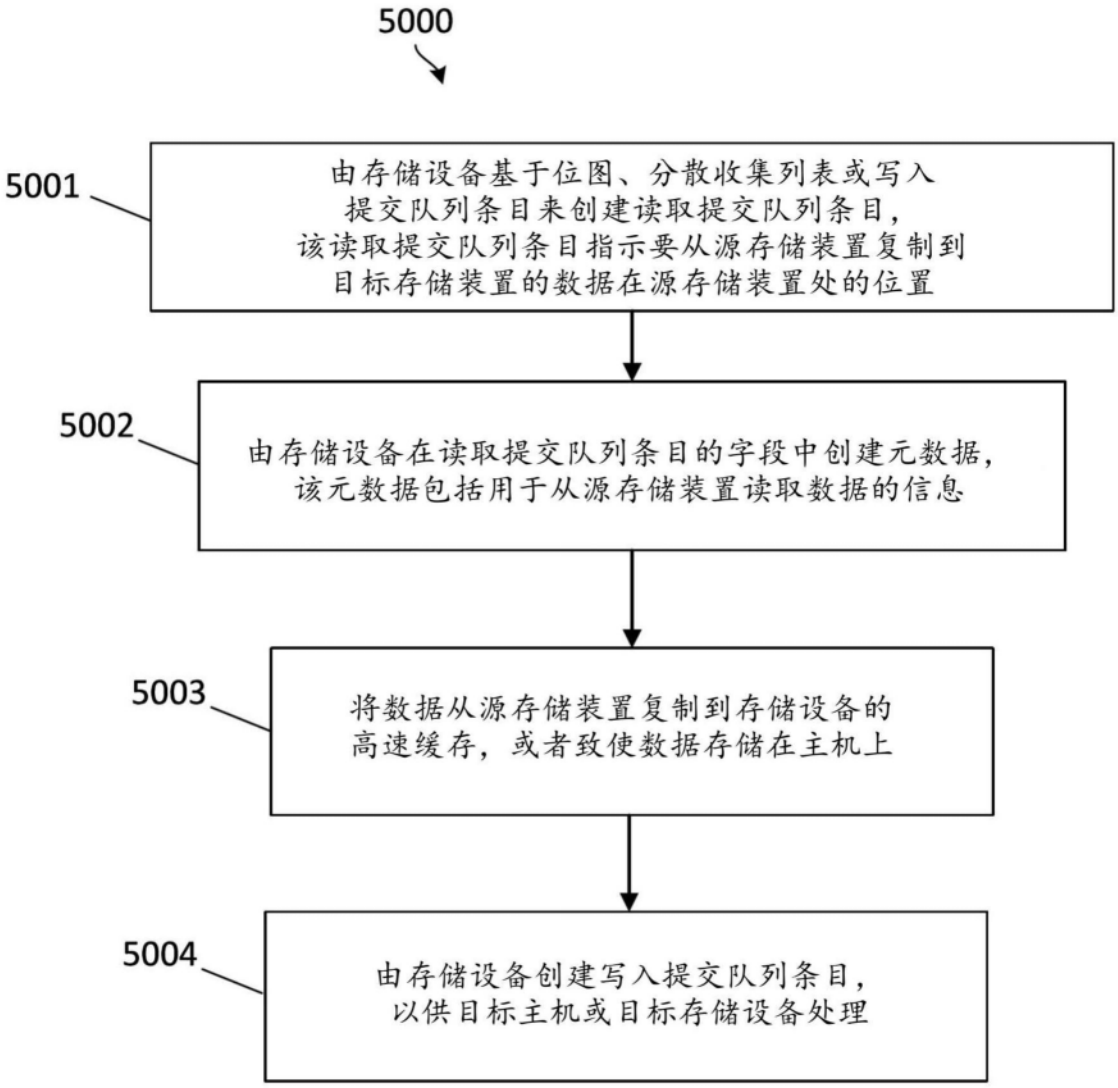

5000
5001
由存储设备基于位图、分散收集列表或写入
提交队列条目来创建读取提交队列条目，
该读取提交队列条目指示要从源存储装置复制到
目标存储装置的数据在源存储装置处的位置
5002
由存储设备在读取提交队列条目的字段中创建元数据，
该元数据包括用于从源存储装置读取数据的信息
5003
将数据从源存储装置复制到存储设备的
高速缓存，或者致使数据存储在主机上
5004
由存储设备创建写入提交队列条目，
以供目标主机或目标存储设备处理

图5