(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 106933747 B
(45)授权公告日 2019.08.20

(21)申请号 201610867445.9

(22)申请日 2016.09.29

(65)同一申请的已公布的文献号
申请公布号 CN 106933747 A

(43)申请公布日 2017.07.07

(30)优先权数据
62/273,323 2015.12.30 US
15/089,237 2016.04.01 US

(73)专利权人 三星电子株式会社
地址 韩国京畿道水原市

(72)发明人 霍囝囝 崔昌皓

(74)专利代理机构 北京铭硕知识产权代理有限公司 11286
代理人 韩明星 王兆赓

(51)Int.Cl.
*G06F 12/02*(2006.01)

(56)对比文件
US 2014208001 A1,2014.07.24,
US 6128630 A,2000.10.03,
US 2014149473 A1,2014.05.29,
US 8949684 B1,2015.02.03,

审查员 吴黄飞

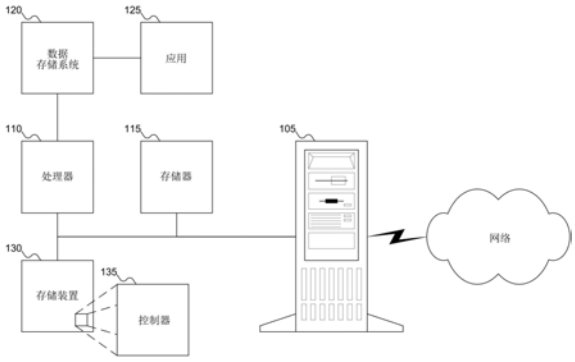
权利要求书2页 说明书11页 附图9页

(54)发明名称

基于多流的数据存储系统和数据存储方法

(57)摘要

提供一种基于多流的数据存储系统和数据存储方法。公开了一种当使用执行二次写的应用和/或数据存储系统时使用具有日志信息和数据的多流的系统和方法。指定日志信息应该被写入一个流的日志写请求可与日志信息一起被发送。指定数据应该被写入另一个流的数据写请求可与数据一起被发送。然后存储装置上的控制器可将日志信息写入与第一流关联的块并将数据写入与第二流关联的块。




CN 106933747 B

1.一种系统,包括:

计算机,包括处理器和存储器;

存储装置;

应用,运行在所述处理器上,所述应用操作为向存储装置发送日志写请求并且向运行在所述处理器上的数据存储系统发送数据写请求,日志写请求包括日志信息并被分配到第一流,数据写请求包括数据并被分配到第二流;

所述数据存储系统被操作为向存储装置发送第二日志写请求并且向存储装置发送第二数据写请求,第二日志写请求包括第二日志信息并被分配到第三流,第二数据写请求包括所述数据并被分配到第二流;以及

存储装置上的控制器,控制器操作为指示存储装置将日志信息写入被分配到第一流的第一块,将所述数据写入被分配到第二流的第二块,并且将第二日志信息写入被分配到第三流的第三块,

其中,第一流、第二流和第三流由数据特性来定义。

2.根据权利要求1所述的系统,其中,控制器被操作为接收无效请求以在数据写请求完成之后删除日志信息。

3.根据权利要求2所述的系统,其中,所述应用被操作为发送无效请求。

4.根据权利要求3所述的系统,其中,所述应用被操作为发送响应于所述应用接收到数据写请求已完成的信号的无效请求。

5.根据权利要求1所述的系统,其中,数据存储系统还被操作为发送第二无效请求。

6.根据权利要求5所述的系统,其中,数据存储系统还被操作为:向存储装置发送第二无效请求,以在第二数据写请求完成之后删除日志信息。

7.根据权利要求1所述的系统,其中,

日志写请求作为直接输入/输出(I/O)请求被发送;

数据写请求作为缓冲的输入/输出(I/O)请求被发送;

其中,日志写请求被用于保证数据写请求中的所述数据被写入存储装置。

8.根据权利要求1所述的系统,其中,包括在第二数据写请求中的所述数据是数据写请求中的所述数据。

9.根据权利要求1所述的系统,其中,日志写请求被用于保证数据写请求中的所述数据被写入存储装置。

10.根据权利要求1所述的系统,其中,第三块是第一块。

11.根据权利要求1所述的系统,其中,

日志写请求通过所述应用被分配到第一流;

数据写请求通过所述应用被分配到第二流。

12.根据权利要求1所述的系统,其中,第一块和第二块具有单个的介质类型。

13.一种方法,包括:

从执行日志写和数据写二者的应用识别待写的数据;

从所述应用向对无效数据执行垃圾回收的存储装置发送日志写请求,日志写请求被分配到第一流,日志写请求作为直接输入/输出(I/O)请求被发送;

从所述应用向数据存储系统发送数据写请求,数据写请求包括所述数据并被分配到第

二流，数据写请求作为缓冲的输入/输出(I/O)请求被发送；

从数据存储系统向存储装置发送第二日志写请求，第二日志写请求被分配到第三流；以及

从数据存储系统向存储装置发送第二数据写请求，第二数据写请求被分配到第二流并包括所述数据，

其中，日志写请求和第二日志写请求被用于保证数据写请求中的所述数据被写入存储装置。

14.根据权利要求13所述的方法，还包括：向存储装置发送无效请求以在数据写请求被写入存储装置之后删除日志信息。

15.根据权利要求14所述的方法，其中，向存储装置发送无效请求的步骤包括：从所述应用向存储装置发送无效请求。

16.根据权利要求15所述的方法，其中，从所述应用向存储装置发送无效请求的步骤包括：在所述应用接收在存储装置上数据写请求已完成的信号。

17.根据权利要求13所述的方法，还包括：向存储装置发送无效请求以在第二数据写请求被写入存储装置之后删除通过第二日志写请求写入的所述数据。

18.一种有形的存储介质，所述有形的存储介质具有存储于其上的非暂时性指令，当由机器执行所述非暂时性指令时，导致：

从执行日志写和数据写二者的应用识别待写的数据；

从所述应用向对无效数据执行垃圾回收的存储装置发送日志写请求，日志写请求被分配到第一流，日志写请求作为直接输入/输出(I/O)请求被发送；

从所述应用向数据存储系统发送数据写请求，数据写请求包括所述数据并被分配到第二流，数据写请求作为缓冲的输入/输出(I/O)请求被发送；

从数据存储系统向存储装置发送第二日志写请求，第二日志写请求被分配到第三流；以及

从数据存储系统向存储装置发送第二数据写请求，第二数据写请求被分配到第二流并包括所述数据，

其中，日志写请求和第二日志写请求被用于保证数据写请求中的所述数据被写入存储装置。

19.根据权利要求18所述的有形的存储介质其中，所述有形的存储介质具有存储于其上的另外的非暂时性指令，当由机器执行所述另外的非暂时性指令时，导致：向存储装置发送无效请求以在数据写请求被写入存储装置之后删除日志信息。

20.根据权利要求19所述的有形的存储介质，其中，向存储装置发送无效请求的步骤包括：从所述应用向存储装置发送无效请求。

21.根据权利要求18所述的有形的存储介质，其中，所述有形的存储介质具有存储于其上的另外的非暂时性指令，当由机器执行所述另外的非暂时性指令时，导致：向存储装置发送无效请求以在第二数据写请求被写入存储装置之后删除通过第二日志写请求写入的所述数据。

# 基于多流的数据存储系统和数据存储方法

[0001] 本申请要求于2015年12月30日提交的第62/273,323号美国临时专利申请及2016年4月1日提交的第15/089,237号美国临时专利申请的权益,所述申请通过引用包含于此。

## 技术领域

[0002] 本发明构思涉及存储装置,更具体地,涉及当使用执行日志记录的应用时利用垃圾回收优化存储装置的使用。

## 背景技术

[0003] 基于NAND闪存的固态驱动器(SSD)已经在企业服务器和数据中心被广泛使用以加快各种各样的数据存储系统。SSD中的闪存具有独特特性,因此直接用SSD来替换传统磁盘不能充分利用装置的全部潜力。一个显著原因是SSD仅写入空闲闪存块,并且使用垃圾回收处理来恢复无效闪存块,以用于重新使用。由于传统的操作系统和应用不在热数据和冷数据之间进行区分,所以具有不同寿命的数据的混合使垃圾回收更难于管理和回收闪存。这既影响SSD的性能也影响SSD的寿命。

[0004] 现今,许多数据存储系统——包括对象存储系统(例如,Ceph)、块存储系统(例如,高速缓存和其他高速缓存系统)和文件存储系统(例如,IBM JFS/JFS2、Linux xfs和Linux ext4)——为了数据持久性和性能的目的而使用日志。这种系统存储数据的两个副本:一个在日志部分,一个在数据部分。当在纯SSD环境下部署这种系统时,由于性能和成本的原因,它们通常在相同的SSD上存储日志和实际数据。当数据被接收待写入时,数据存储系统首先将数据的一个副本存储在被刷新到磁盘的日志中,并将数据的第二个副本存储在存储器中的文件系统页高速缓存中。然后数据存储系统将成功返回到用户应用。最后,在后台的某个时间,数据存储系统将文件系统页高速缓存中的那些数据记录刷新到磁盘中,并移除磁盘上的日志中的相同的数据记录。每次数据写重复这个处理,即使在日志仅用于元数据时也发生这个处理。

[0005] 这种二次写(double-write)方法在使用SSD时有一个问题:每个SSD块内的闪存内部碎片。这种内部碎片问题造成更多的垃圾回收,导致存储系统性能降低、更长的读/写延迟以及更短的SSD寿命。

[0006] 仍然需要一种方式来使用避免(或至少最小化)SSD中的闪存的碎片的使用SSD的二次写方法。

## 发明内容

[0007] 一种系统可包括计算机、存储装置、应用和控制器。一种计算机可包括处理器和存储器。一种应用可在处理器上运行。应用可被操作为向存储装置发送日志写请求和数据写请求二者。日志写请求可包括日志信息并与第一流关联。数据写请求可包括数据并与第二流关联。一种在存储装置上的控制器可被操作为指导存储装置将日志信息写入与第一流关联的第一块并将数据写入与第二流关联的第二块。

[0008] 根据本发明构思的实施例的一方面，一种基于多流的数据存储系统可包括：计算机，包括处理器和存储器；存储装置；应用，运行在处理器上，所述应用可被操作为向存储装置发送日志写请求和数据写请求二者，日志写请求可包括日志信息并与第一流关联，数据写请求可包括数据并与第二流关联；存储装置上的控制器，控制器可被操作为指示存储装置将日志信息写入与第一流关联的第一块并将数据写入与第二流关联的第二块。

[0009] 控制器可被操作为接收无效请求以在数据写请求完成之后删除日志信息。

[0010] 所述应用可被操作为发送无效请求。

[0011] 所述应用可被操作为发送响应于所述应用接收到数据写请求已完成的信号的无效请求。

[0012] 所述应用可被操作为向在处理器上运行的数据存储系统发送数据写请求；数据存储系统可被操作为向存储装置发送第二数据写请求，第二数据写请求包括数据并与第二流关联。

[0013] 数据存储系统还可被操作为向存储装置发送第二日志写请求，第二日志写请求包括第二日志信息并与第三流关联。

[0014] 数据存储系统还可被操作为发送第二无效请求。

[0015] 根据本发明构思的实施例的另一方面，一种基于多流的数据存储方法可包括：从执行日志写和数据写二者的应用识别待写的数据；从所述应用向对无效数据执行垃圾回收的存储装置发送日志写请求，日志写请求被分配到第一流；从所述应用向存储装置发送数据写请求，数据写请求被分配到第二流。

[0016] 所述数据存储方法还可包括：向存储装置发送无效请求以在数据写请求被写入存储装置之后删除日志信息。

[0017] 向存储装置发送无效请求的步骤可包括：从所述应用向存储装置发送无效请求。

[0018] 从所述应用向存储装置发送无效请求的步骤可包括：在所述应用接收在存储装置上数据写请求已完成的信号。

[0019] 从所述应用向存储装置发送数据写请求的步骤可包括：从所述应用向数据存储系统发送数据写请求；从数据存储系统向存储装置发送第二数据写请求。

[0020] 从数据存储系统向存储装置发送第二数据写请求的步骤可包括：从数据存储系统向存储装置发送第二日志写请求。

[0021] 所述数据存储方法还可包括：向存储装置发送无效请求以在第二数据写请求被写入存储装置之后删除通过第二日志写请求写入的数据。

[0022] 其他特征和方面将从以下详细描述、附图和权利要求而清楚。

## 附图说明

[0023] 图1示出根据本发明构思的实施例的能够使用具有日志记录的数据存储系统的服务器。

[0024] 图2示出图1的服务器的另外的细节。

[0025] 图3示出图1的应用与图1的存储装置进行通信以执行日志记录和数据写二者。

[0026] 图4示出使用多流来存储日志和数据的图1的存储装置。

[0027] 图5A至图5B示出使用传统方法与本发明构思的实施例的图1的存储装置的使用的

比较。

[0028] 图6A至图6B示出图1的应用和图1的数据存储系统与图1的存储装置进行通信以执行日志记录和数据写二者。

[0029] 图7A至图7B示出根据本发明构思的实施例的图1的应用和图1的数据存储系统与图1的存储装置进行通信并执行日志记录和数据写二者的示例过程的流程图。

## 具体实施方式

[0030] 现将详细描述本发明构思的实施例，所述实施例的示例在附图中示出。在以下详细描述中，阐述大量具体细节以实现对本发明构思的彻底理解。然而，应该理解，具有本领域普通技术的人员可在没有这些具体细节的情况下实践本发明构思。在其他示例中，未详细描述公知的方法、过程、组件、电路和网络，以免不必要地模糊实施例的各方面。

[0031] 将会理解，虽然术语第一、第二等可在这里使用以描述各种元件，但是这些元件不应被这些术语限制。这些术语仅用于将一个元件与另一元件区分。例如，在不脱离本发明构思的范围的情况下，第一模块可被称为第二模块，类似地，第二模块可被称为第一模块。

[0032] 这里用在本发明构思的描述中的术语仅为描述特定实施例的目的，而并非意图限制本发明构思。如在本发明构思的描述和权利要求中所使用，除非上下文明确地另有指示，否则单数形式也意图包括复数形式。还将理解，如在这里使用的术语“和/或”表示并包含一个或多个相关所列项的任意可能组合。还将理解，当在本说明书中使用术语“包括”和/或“包含”时，指定存在所陈述的特征、整体、步骤、操作、元件和/或组件，但不排除存在或添加一个或多个其他的特征、整体、步骤、操作、元件、组件和/或它们的组。附图的组件和特征不必按比例绘制。

[0033] 图1示出根据本发明构思的实施例的能够使用具有日志记录的数据存储系统的服务器。在图1中，示出服务器105。服务器105可以是任何种类的服务器。服务器105可包括处理器110和存储器115。处理器110可以是任何种类的处理器；存储器115可以是任何种类的存储器。

[0034] 数据存储系统120可运行在处理器110上。数据存储系统120可以是执行二次写(即，日志记录和数据写二者)的任何系统。数据存储系统120不只意图包括对象(和文件)存储系统(诸如Ceph®)，还意图包括在其他操作系统上运行的应用，所述应用在其他操作系统上执行二次写。(Ceph是美国Inktank Storage公司的注册商标。)

[0035] 除了数据存储系统120之外，应用125可在数据存储系统120之上运行。在本发明构思的一些实施例中，为了内部原因，应用125自身可执行二次写。例如，应用125可以是实时模拟程序。这种程序高度依赖执行操作的时间。如果实时模拟程序被缓冲的但未写入的数据打断，则模拟的结果会被浪费。因此，模拟程序会想要保证：数据即使不通过数据存储系统120被写入存储装置130，也能通过日志记录被存储。

[0036] 存储装置130可以是执行无效数据的垃圾回收的任何期望种类的存储装置。作为示例，存储装置130可以是基于闪存的固态驱动器(SSD)。存储装置130可具有负责管理存储装置130的操作的控制器135。例如，除了其他功能之外，控制器135可管理数据读和写，并且可在存储装置130上将逻辑块地址映射到物理块地址。除了其他组件之外，控制器135可包括能够将控制器135(直接或间接地)连接到服务器105的物理接口、控制存储装置130的操

作的处理器、用于提供针对存储在闪存中的数据的错误检测和纠正能力的纠错码电路、用于管理存储装置130内的动态随机存取存储器(DRAM)的DRAM控制器、以及用于管理闪存的一个或多个闪存控制器。控制器135还可包括多流控制器，多流控制器可管理何种数据被写入何种块(如下所述，与不同流关联的块)。在本发明构思的一些实施例中，控制器135可以是利用这些组件的功能适当编程的单个芯片；在本发明构思的其他实施例中，控制器135可包括这些组件中的一些或全部作为单独组件(例如，芯片)。

[0037] 图2示出图1的服务器的另外的细节。参照图2，通常，服务器105包括一个或多个处理器110，处理器110可包括可被用来协调服务器105的组件的操作的存储器控制器205和时钟210。处理器110还可结合到存储器115，存储器115可包括随机存取存储器(RAM)、只读存储器(ROM)或其他状态保存介质作为示例。处理器105还可结合到存储装置130和网络连接器215，例如，网络连接器215可以是以太网连接器。处理器110还可连接到总线220，除了其他组件之外，可使用输入/输出引擎230管理的用户接口225和输入/输出接口端口可附接到总线220。

[0038] 图3示出图1的应用125与图1的存储装置130进行通信以执行日志记录和数据写二者。在图3中，示出应用125在没有图1的数据存储系统120的情况下与存储装置130进行通信。在只有一个软件元件(被表示为图3中的应用125)与存储装置130进行通信的本发明构思的实施例中，该软件元件可以是执行日志记录的任何软件元件。因此，在图3中可使用图1的数据存储系统120在不损失适用性的情况下替换应用125。(以下图6A至图6B描述图1的应用125和数据存储系统120均与存储装置130执行日志记录的本发明构思的实施例。)

[0039] 在图3中，应用125可向存储装置130发送日志写请求310。日志写请求310可包括日志信息305和流标识符315。流标识符315可指定被分配日志信息305的特定的流。如以下参照图4所述，不同的流可与存储装置130上基于一个或多个特性(诸如预期的寿命或任何其他划分标准)划分数据的不同的块或超级块关联。可使用直接输入/输出(I/O)命令来发送日志写请求310，以保证日志信息305被立即写入存储装置130。可选地，日志写请求310可被发送到(即使未满)被立即刷新的缓冲器，以再次保证日志信息305被立即写入存储装置130。

[0040] 应用125还可向存储装置130发送数据写请求320。数据写请求320可包括数据325和流标识符330。流标识符330可与流标识符315识别不同的流，以使数据325被写入与日志信息305不同的流(因此被写入不同的块或超级块)。由于日志写请求310被立即写入存储装置130，因此数据写请求320可作为缓冲的写请求被发送，而数据325可最终但不必须立即地被写入存储装置130。

[0041] 最终，存储装置130可向应用125发送信号335。信号335可指示数据写请求320已经完成并且数据325已被写入存储装置130。在这时，不再需要日志信息305以保证数据被写入存储装置130的某处。在执行垃圾回收的存储装置上，数据通常在可经垃圾回收而被删除之前被无效。因此，应用125可向存储装置130发送无效请求340，请求从存储装置130删除日志信息305。由于日志信息305被写入与数据325不同的块(或超级块)，因此无效请求340将不会在存储装置130内导致无效碎片化的块(或超级块)。

[0042] 图4示出使用多流来存储日志和数据的图1的存储装置130。在图4中，示出存储装置130被划分成多个块，诸如块405、410、415和420，等等。每个块可依次被划分成多个页：例

如，块405被示出为包括页425、430、435和440，块410被示出为包括页445、450、455和460。虽然图4示出块405、410、415和420均具有四个页，但块405、410、415和420可包括任何期望数量的页，示出的四个仅为示例。

[0043] 如SSD可出现的，页可表示可从存储装置130读出或可写入存储装置130的数据的最小单位。相反，在本发明构思的一些实施例中，块可表示执行垃圾回收的数据的最小单位。在本发明构思的其他实施例(图4中未示出)中，存储装置130的块可被组织成被称为超级块的更大的组，超级块可以是执行垃圾回收的数据的最小单位。无论对块还是超级块执行垃圾回收，垃圾回收的最小单位都大于页。这种差异可解释垃圾回收为何可对存储装置130的操作具有负面影响：如果在针对垃圾回收的块的一个或多个页中存在有效数据，则该数据必须在所述块可进行垃圾回收之前被复制到另一个块。例如，如果页425包含有效数据，则该数据必须在块405可进行垃圾回收之前被复制到例如块410中的页445(假设页445空闲)。

[0044] 在本发明构思的一些实施例中，页可被组织为块。但是块可被组织为超级块，对超级块执行垃圾回收，而不是对块执行垃圾回收。然而，虽然超级块的构思可对存储装置130中的垃圾回收的实现具有影响，但是从理论的观点来看，超级块只是对用于垃圾回收的目的的块的大小的重新限定。关于块的任何讨论可被理解为也适用于超级块。

[0045] 如以上参照图3提到的，单独的块可被分配给流。例如，块405和410可被分配给流315，而块415和420可被分配给流330。在“热”数据和“冷”数据可被划分成不同的流的本发明构思的实施例中，流分配可避免能造成垃圾回收操作出问题的日志写和数据写的混合。

[0046] 图5A至图5B示出使用传统方法与本发明构思的实施例的存储装置130的使用的比较。如上所述，在传统系统中日志写和数据写被写入图1的存储装置130中的相同块。在图5A中，示出块505具有包含日志写的页510、515和520和包含数据写的页525、530和540。由于日志写往往具有短的寿命(因为一旦相应的数据写完成，它们会被删除)，因此日志写和数据写的混合造成碎片化的块，如块535所示(块535是在日志写被无效之后的块505)。如果块535随后进行垃圾回收，则页525、530和540必须首先被复制到另一个块。这种复制花费时间，减慢对存储装置130的其他的读和写操作。

[0047] 但是如在本发明构思的实施例中，如果日志写和数据写被发送到不同的流，则擦除日志写不会造成碎片化的块。图5B示出这种情形。在图5B中，日志写被发送到块545；数据写被发送到块550。当日志写510、515和520被无效时，块545不存储任何需要被复制到另一个块的数据；数据写525、530和540被存储在块550中。(虽然以上描述简化了情形，日志写不必在同一时间被全部擦除，但是通常日志写在被写入之后不久(尤其是与数据写的寿命相比)被删除。因此，块545中的所有数据应该在最后的日志写被写入块545之后不久被无效，并且整个块可在不必将任何数据复制到另一个块的情况下进行垃圾回收。)

[0048] 图6A至图6B示出图1的应用125和图1的数据存储系统120与图1的存储装置130进行通信以执行日志记录和数据写二者。与应用125在没有数据存储系统120的情况下与存储装置130进行通信的图3相反，图6A至图6B可在事件的顺序中包括数据存储系统120。

[0049] 在图6A中，应用125如前所述发送具有日志信息305和流标识符315的日志写请求310。由于日志写请求310可以是直接I/O命令，因此图6A示出应用125跳过数据存储系统120向存储装置130发送日志写请求310。但是在本发明构思的其他实施例中，应用125可向数据

存储系统120发送日志写请求310，日志写请求310具有使得数据存储系统120执行直接I/O命令以完成日志写请求310的指令。然而，与图3相反，应用125将具有数据325和流标识符330的数据写请求320发送到数据存储系统120，而不是存储装置130。数据存储系统120随后可负责监督向存储装置130写数据325。

[0050] 由于数据存储系统120自身可针对它自己的数据和/或元数据执行日志记录，因此数据存储系统120可向存储装置130发送日志写请求605。日志写请求605可包括日志信息610和流标识符615。这表明写单个数据单元可涉及多个日志，并因此涉及多个流。

[0051] 在图6B中，数据存储系统120可向存储装置130发送它自己的数据写请求620。数据写请求620可包括数据325和流标识符330。注意，数据写请求620中的数据325和流标识符330可与数据写请求320中的相同：这是合理的，因为相同的数据被写入，只是绕道通过数据存储系统120而已。最终，存储装置130可发送信号625，通知数据存储系统120数据写请求620已完成，之后数据存储系统120可发送无效请求630以删除日志信息610。存储装置130还可将信号335发送回应用125(如果存储装置130知道应用125的存在)，以使得应用125可发送它自己的无效请求340。但是在本发明构思的其他实施例中，数据存储系统120可使用无效请求630来删除日志信息610和日志信息305二者。而在本发明构思的其他实施例中，数据存储系统120可向应用125发送信号335，以通知应用125它可发送无效请求340。

[0052] 在以上讨论的本发明构思的实施例中，应用125和/或数据存储系统120负责擦除日志信息305和/或610。因此，应用125和/或数据存储系统120需要接收信号335和/或625，来知道何时安全地删除日志信息305和/或610。但是在本发明构思的其他实施例中，存储装置130可知道数据315的来源，并且一旦数据写请求320和/或620完成即可自动删除日志信息305和/或610，避免应用125和/或数据存储系统120需要发送无效请求340和/或630。

[0053] 在本发明构思的一些实施例中，数据存储系统120的多个实例可共存于单个存储装置130上。例如，存储装置130可具有多个日志文件系统划分。或者，存储装置130可保持多个对象存储实例。在本发明构思的这种实施例中，数据存储系统120的每个实例可向相同的存储装置130发送它自己的日志写请求605。每个单独的日志写请求605可包括它自己的日志信息610和流标识符615。数据存储系统120的每个实例可向相同的存储装置130发送它自己的数据写请求620。每个单独的数据写请求620可包括它自己的数据325和流标识符330。数据存储系统120的每个实例可将它自己的日志信息610和数据325存储到不同的流，然后存储装置130可将各种日志信息610和数据325存储在不同的块或超级块中。因此，除了针对数据存储系统120的单独的实例的日志信息610和数据325在不同的块或超级块中以外，来自数据存储系统120的不同实例的不同日志信息610也可被存储在不同的块或超级块中，并且来自数据存储系统120的不同实例的不同数据325也可被存储在不同的块或超级块中。

[0054] 图7A至图7B示出根据本发明构思的实施例的图1的应用125和图1的数据存储系统120与图1的存储装置130进行通信并执行日志记录和数据写二者的示例过程的流程图。在图7A中，在块705，图1的应用125可识别将写入图1的存储装置130的图3的数据325。在块710，图1的应用125可将图3的日志写请求310作为直接I/O命令发送到图1的存储装置130，指定图3的流标识符315作为用于图3的日志信息305的流。

[0055] 在这时，操作可沿不同的路径进行。在本发明构思的一些实施例中，在块715，图1的应用125可将图3的数据写请求320作为缓冲的I/O命令发送到图3的存储装置130。然后，

在块720，图1的应用125可从图1的存储装置130接收指示图3的数据写请求320已完成的信号335。最后，在块725，图1的应用125可向图1的存储装置130发送图3的无效请求340，以删除图3的日志信息305。

[0056] 可选地，在本发明构思的其他实施例中，图1的应用125可不向图1的存储装置130发送图3的数据写请求320。相反，在块730，图1的应用125可向图1的数据存储系统120发送图3的数据写请求320。在块735，图1的数据存储系统120可向图1的存储装置130发送图6A的第二日志写请求605。在块740，图1的数据存储系统120可向图1的存储装置130发送图6B的第二数据写请求620。在块745，图1的数据存储系统120可从图1的存储装置130接收指示图6B的数据写请求620已经完成的图6B的信号625。在块750，图1的数据存储系统120可向图1的存储装置130发送图6B的无效请求630，以删除图6A的日志信息610。然后处理可继续图7A的块720。

[0057] 在图7A至图7B中，示出本发明构思的一些实施例。但是本领域的技术人员将认识到，通过改变块的次序、通过省略块或者通过包括附图中未示出的链接，本发明构思的其他实施例也是可行的。无论是否明确描述，流程图的所有这种变化都被认为是本发明构思的实施例。

[0058] 以下讨论意图提供合适的机器的简短、一般的描述，在所述机器中可实现本发明构思的特定方面。所述机器可通过来自传统输入装置(诸如键盘、鼠标等)的输入以及通过从另一机器接收的指令、与虚拟现实(VR)环境的交互、生物反馈、或其他输入信号被至少部分地控制。如这里使用的，术语“机器”意图广泛包含单个机器、虚拟机器、或者一起操作的通信地结合的机器、虚拟机器或者装置的系统。示例性机器包括计算装置(诸如个人计算机、工作站、服务器、便携式计算机、手持装置、电话、平板等)以及诸如私有或公共交通的交通装置(例如，汽车、火车、出租车等)。

[0059] 所述机器可包括嵌入式控制器，诸如，可编程或非可编程逻辑装置或阵列、专用集成电路(ASIC)、嵌入式计算机、智能卡等。所述机器可使用到一个或多个远程机器的一个或多个连接(诸如通过网络接口、调制解调器或其他通信结合)。可通过物理和/或逻辑网络(诸如内联网、互联网、局域网、广域网等)的方式将机器互连。本领域技术人员将理解，网络通信可使用各种有线和/或无线的近程或远程载波和协议，包括射频(RF)、卫星、微波、电气和电子工程师协会(IEEE)802.11、蓝牙®、光学、红外、线缆、激光等。

[0060] 本发明构思的实施例可通过参照或结合包括函数、进程、数据结构、应用程序等的关联数据来描述，当所述关联数据被机器访问时，导致所述机器执行任务或定义抽象数据类型或低层硬件环境。例如，关联数据可被存储在易失性和/或非易失性存储器(例如，RAM、ROM等)中，或者被存储在其他存储装置和它们的关联存储介质(包括硬盘驱动器、软盘、光学存储器、磁带、闪存、存储棒、数字视频盘、生物存储器等)中。关联数据可通过传输环境(包括物理和/或逻辑网络)以包、串行数据、并行数据、传播信号等的形式被传递，并且可以以压缩或加密的格式被使用。关联数据可被用于分布式环境中，并被本地和/或远程存储用于机器访问。

[0061] 本发明构思的实施例可包括有形的、非暂时性的机器可读介质，所述机器可读介质包括可由一个或多个处理器执行的指令，所述指令包括用于执行如这里所述的本发明构思的元素的指令。

[0062] 已参照示出的实施例描述并说明了本发明构思的原理，将认识到，示出的实施例可在不脱离这些原理的情况下在布置和细节上被修改，并可以以任何期望的方式来组合。并且，虽然前述讨论已集中于特定实施例，但可预期其他的配置。具体地，尽管在这里使用诸如“根据本发明构思的实施例”等的表述，但是这些短语是为一般性参考实施例的可能性，而不意图将本发明构思限制于特定实施例配置。如这里使用的，这些术语可参考可组合到其他实施例中的相同或不同的实施例。

[0063] 前述说明性实施例将不被解释为限制其发明构思。虽然已经描述一些实施例，但是本领域技术人员将容易理解，在不实质脱离本公开的新颖性教导和优点的情况下，很多修改对于那些实施例是可行的。因此，所有这种修改意图被包括在如权利要求所限定的本发明构思的范围内。

[0064] 本发明构思的实施例可无限制地延伸到以下声明：

[0065] 声明1、本发明构思的实施例包括一种系统，包括：

[0066] 计算机，包括处理器和存储器；

[0067] 存储装置；

[0068] 应用，运行在处理器上，所述应用被操作为向存储装置发送日志写请求和数据写请求二者，日志写请求包括日志信息并与第一流关联，数据写请求包括数据并与第二流关联；

[0069] 存储装置上的控制器，控制器被操作为指示存储装置将日志信息写入与第一流关联的第一块并将数据写入与第二流关联的第二块。

[0070] 声明2、本发明构思的实施例包括根据声明1所述的系统，其中，存储装置包括固态驱动器(SSD)。

[0071] 声明3、本发明构思的实施例包括根据声明1所述的系统，其中，所述应用使用直接输入/输出(I/O)命令向存储装置被操作为发送日志写请求。

[0072] 声明4、本发明构思的实施例包括根据声明1所述的系统，其中，所述应用使用缓冲的写命令向存储装置被操作为发送数据写请求。

[0073] 声明5、本发明构思的实施例包括根据声明1所述的系统，其中，控制器被操作为接收无效请求以在数据写请求完成之后删除日志信息。

[0074] 声明6、本发明构思的实施例包括根据声明5所述的系统，其中，所述应用被操作为发送无效请求。

[0075] 声明7、本发明构思的实施例包括根据声明6所述的系统，其中，所述应用被操作为发送响应于所述应用接收到数据写请求已完成的信号的无效请求。

[0076] 声明8、本发明构思的实施例包括根据声明7所述的系统，其中，所述应用被操作为发送响应于所述应用从控制器接收到数据写请求已完成的信号的无效请求。

[0077] 声明9、本发明构思的实施例包括根据声明1所述的系统，其中：

[0078] 所述应用向在处理器上运行的数据存储系统被操作为发送数据写请求；

[0079] 数据存储系统向存储装置被操作为发送第二数据写请求，第二数据写请求包括数据并与第二流关联。

[0080] 声明10、本发明构思的实施例包括根据声明9所述的系统，其中，数据存储系统还向存储装置被操作为发送第二日志写请求，第二日志写请求包括第二日志信息并与第三流

关联。

[0081] 声明11、本发明构思的实施例包括根据声明10所述的系统，其中，数据存储系统还被操作为发送第二无效请求。

[0082] 声明12、本发明构思的实施例包括根据声明10所述的系统，其中，数据存储系统被操作为发送响应于数据存储系统接收到第二数据写请求已完成的第二信号的第二无效请求。

[0083] 声明13、本发明构思的实施例包括根据声明12所述的系统，其中，数据存储系统被操作为发送响应于数据存储系统从控制器接收到第二数据写请求已完成的第二信号的第二无效请求。

[0084] 声明14、本发明构思的实施例包括一种方法，包括：

[0085] 从执行日志写和数据写二者的应用识别待写的数据；

[0086] 从所述应用向对无效数据执行垃圾回收的存储装置发送日志写请求，日志写请求被分配到第一流；

[0087] 从所述应用向存储装置发送数据写请求，数据写请求被分配到第二流。

[0088] 声明15、本发明构思的实施例包括根据声明14所述的方法，其中：

[0089] 从所述应用向存储装置发送日志写请求的步骤包括：从所述应用向固态驱动器(SSD)发送日志写请求；

[0090] 从所述应用向存储装置发送数据写请求的步骤包括：从所述应用向SSD发送数据写请求。

[0091] 声明16、本发明构思的实施例包括根据声明14所述的方法，其中：

[0092] 从所述应用向存储装置发送日志写请求的步骤包括：使用直接输入/输出(I/O)命令从所述应用向存储装置送日志写请求；

[0093] 从所述应用向存储装置发送数据写请求的步骤包括：使用缓冲的写命令从所述应用向存储装置发送数据写请求。

[0094] 声明17、本发明构思的实施例包括根据声明14所述的方法，还包括：向存储装置发送无效请求以在数据写请求被写入存储装置之后删除日志信息。

[0095] 声明18、本发明构思的实施例包括根据声明17所述的方法，其中，向存储装置发送无效请求的步骤包括：从所述应用向存储装置发送无效请求。

[0096] 声明19、本发明构思的实施例包括根据声明18所述的方法，其中，从所述应用向存储装置发送无效请求的步骤包括：在所述应用接收在存储装置上数据写请求已完成的信号。

[0097] 声明20、本发明构思的实施例包括根据声明19所述的方法，其中，在所述应用接收在存储装置上数据写请求已完成的信号的步骤包括：在所述应用接收来自存储装置的在存储装置上数据写请求已完成的信号。

[0098] 声明21、本发明构思的实施例包括根据声明14所述的方法，其中，从所述应用向存储装置发送数据写请求的步骤包括：

[0099] 从所述应用向数据存储系统发送数据写请求；

[0100] 从数据存储系统向存储装置发送第二数据写请求。

[0101] 声明22、本发明构思的实施例包括根据声明21所述的方法，其中，从数据存储系统

向存储装置发送第二数据写请求的步骤包括:从数据存储系统向存储装置发送第二日志写请求。

[0102] 声明23、本发明构思的实施例包括根据声明22所述的方法,还包括:向存储装置发送无效请求以在第二数据写请求被写入存储装置之后删除通过第二日志写请求写入的数据。

[0103] 声明24、本发明构思的实施例包括根据声明23所述的方法,其中,向存储装置发送无效请求的步骤包括:从数据存储系统向存储装置发送无效请求。

[0104] 声明25、本发明构思的实施例包括根据声明24所述的方法,其中,从数据存储系统向存储装置发送无效请求的步骤包括:在数据存储系统接收在存储装置上数据写请求已完成的信号。

[0105] 声明26、本发明构思的实施例包括根据声明25所述的方法,其中,从数据存储系统接收在存储装置上数据写请求已完成的信号的步骤包括:在数据存储系统接收来自存储装置的在存储装置上数据写请求已完成的信号。

[0106] 声明27、本发明构思的实施例包括一种物品,包括一种有形的存储介质,所述有形的存储介质具有存储于其上的非暂时性指令,当所述非暂时性指令由机器执行时,导致:

[0107] 从执行日志写和数据写二者的应用识别待写的数据;

[0108] 从所述应用向对无效数据执行垃圾回收的存储装置发送日志写请求,日志写请求被分配到第一流;

[0109] 从所述应用向存储装置发送数据写请求,数据写请求被分配到第二流。

[0110] 声明28、本发明构思的实施例包括根据声明27所述的物品,其中:

[0111] 从所述应用向存储装置发送日志写请求的步骤包括:从所述应用向固态驱动器(SSD)发送日志写请求;

[0112] 从所述应用向存储装置发送数据写请求的步骤包括:从所述应用向SSD发送数据写请求。

[0113] 声明29、本发明构思的实施例包括根据声明27所述的物品,其中:

[0114] 从所述应用向存储装置发送日志写请求的步骤包括:使用直接输入/输出(I/O)命令从所述应用向存储装置送日志写请求;

[0115] 从所述应用向存储装置发送数据写请求的步骤包括:使用缓冲的写命令从所述应用向存储装置发送数据写请求。

[0116] 声明30、本发明构思的实施例包括根据声明27所述的物品,所述有形的存储介质还具有存储于其上的非暂时性指令,当由机器执行所述非暂时性指令时,导致向存储装置发送无效请求以在数据写请求被写入存储装置之后删除日志信息。

[0117] 声明31、本发明构思的实施例包括根据声明30所述的物品,其中,向存储装置发送无效请求的步骤包括:从所述应用向存储装置发送无效请求。

[0118] 声明32、本发明构思的实施例包括根据声明31所述的物品,其中,从所述应用向存储装置发送无效请求的步骤包括:在所述应用接收在存储装置上数据写请求已完成的信号。

[0119] 声明33、本发明构思的实施例包括根据声明32所述的物品,其中,在所述应用接收在存储装置上数据写请求已完成的信号的步骤包括:在所述应用接收来自存储装置的在存

储装置上数据写请求已完成的信号。

[0120] 声明34、本发明构思的实施例包括根据声明27所述的物品，其中，从所述应用向存储装置发送数据写请求的步骤包括：

[0121] 从所述应用向数据存储系统发送数据写请求；

[0122] 从数据存储系统向存储装置发送第二数据写请求。

[0123] 声明35、本发明构思的实施例包括根据声明34所述的物品，其中，从数据存储系统向存储装置发送第二数据写请求的步骤包括：从数据存储系统向存储装置发送第二日志写请求。

[0124] 声明36、本发明构思的实施例包括根据声明35所述的物品，所述有形的存储介质还具有存储于其上的非暂时性指令，当由机器执行所述非暂时性指令时，导致：向存储装置发送无效请求以在第二数据写请求被写入存储装置之后删除通过第二日志写请求写入的数据。

[0125] 声明37、本发明构思的实施例包括根据声明36所述的物品，其中，向存储装置发送无效请求的步骤包括：从数据存储系统向存储装置发送无效请求。

[0126] 声明38、本发明构思的实施例包括根据声明37所述的物品，其中，从数据存储系统向存储装置发送无效请求的步骤包括：在数据存储系统接收在存储装置上数据写请求已完成的信号。

[0127] 声明39、本发明构思的实施例包括根据声明38所述的物品，其中，从数据存储系统接收在存储装置上数据写请求已完成的信号的步骤包括：在数据存储系统接收来自存储装置的在存储装置上数据写请求已完成的信号。

[0128] 因此，考虑到这里描述的实施例的多种置换，本具体实施方式和随附的材料仅意在说明性，而不应该被认为限制本发明构思的范围。因此，本发明构思所要求保护的内容是可以在权利要求及其等同物的范围和精神之内的所有这种修改。

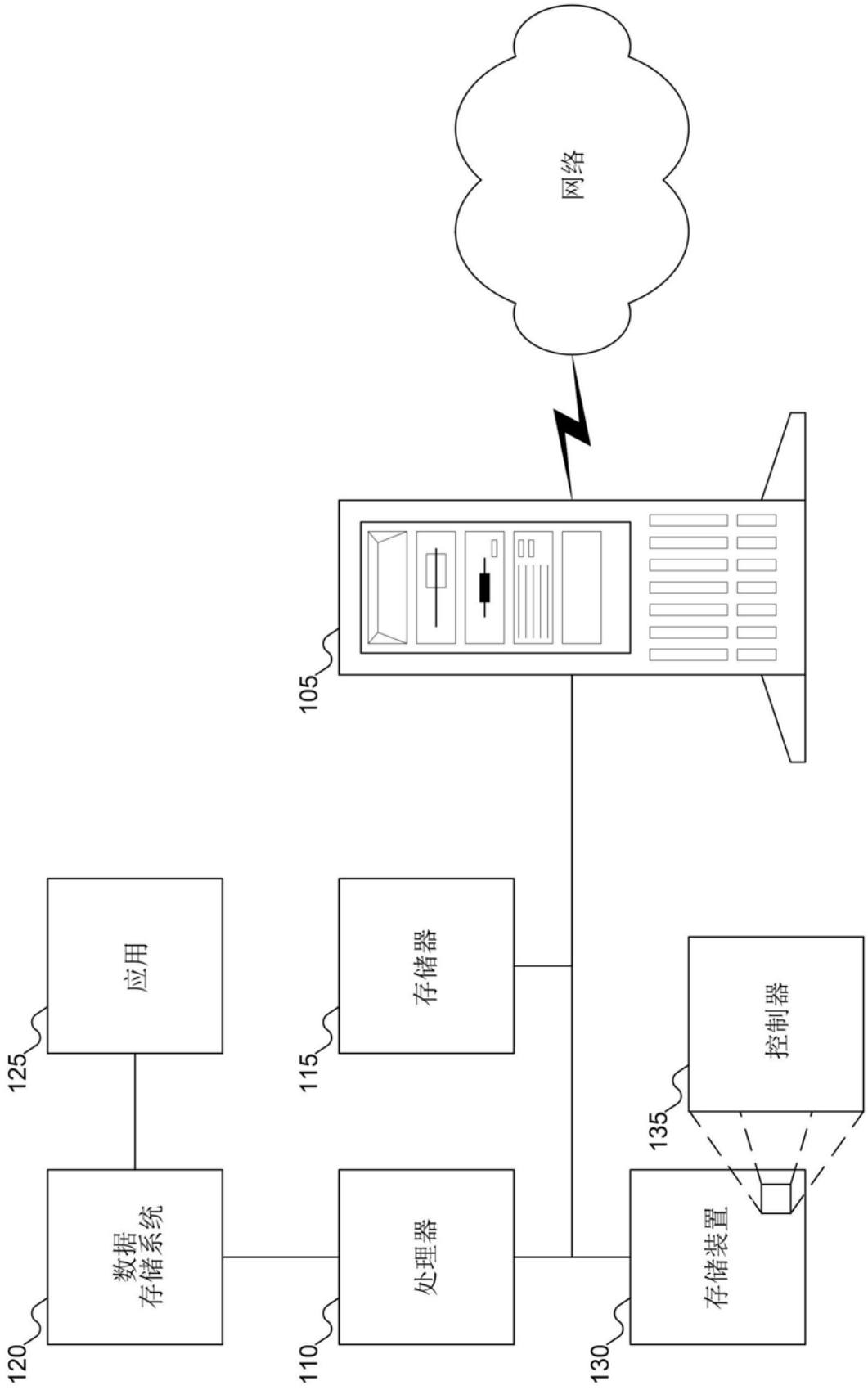



图1

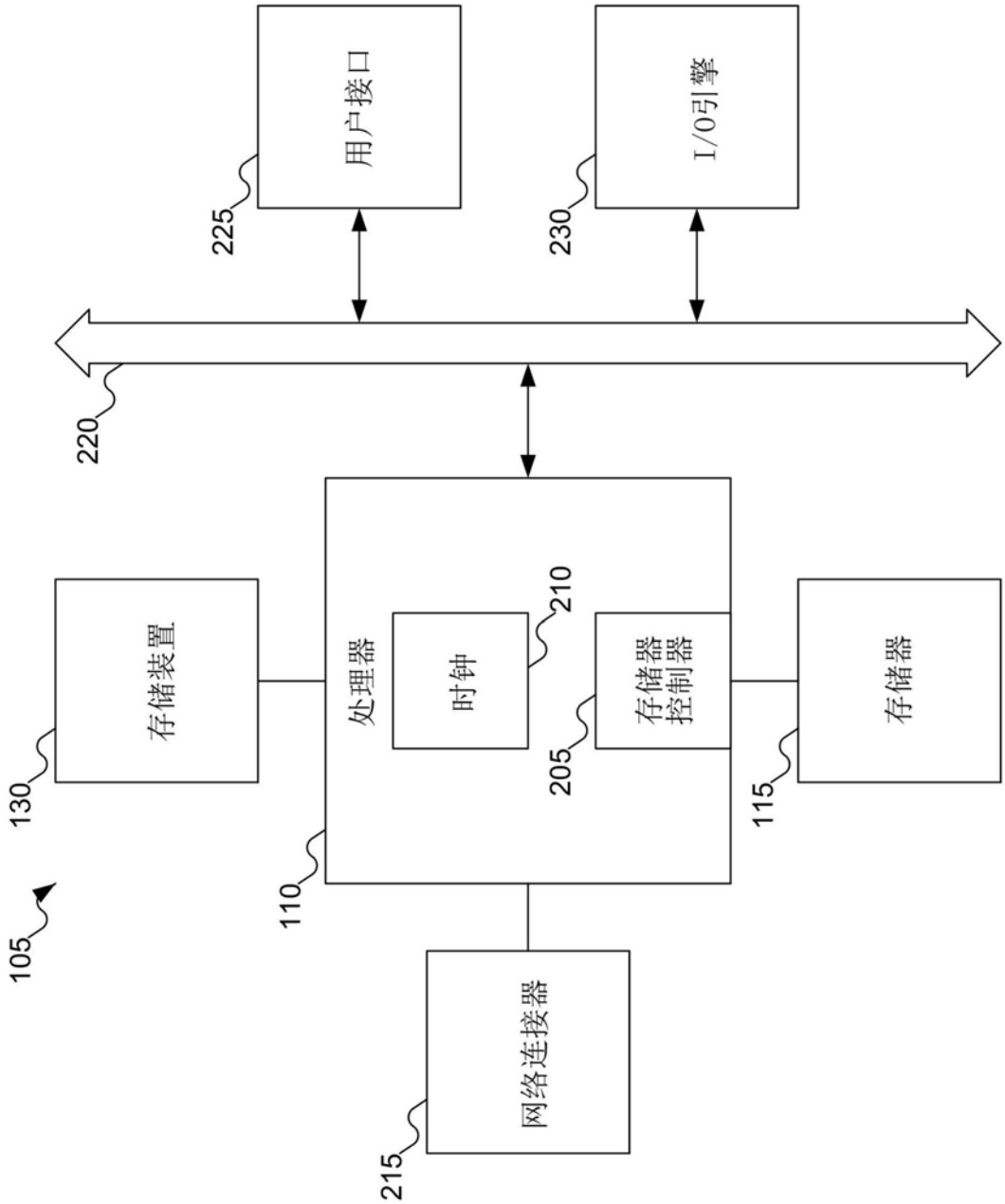

用户接口
225
I/O引擎
230
220
处理器
时钟
210
存储器控制器
205
存储装置
130
存储器
115
110
105
网络连接器
215

图2

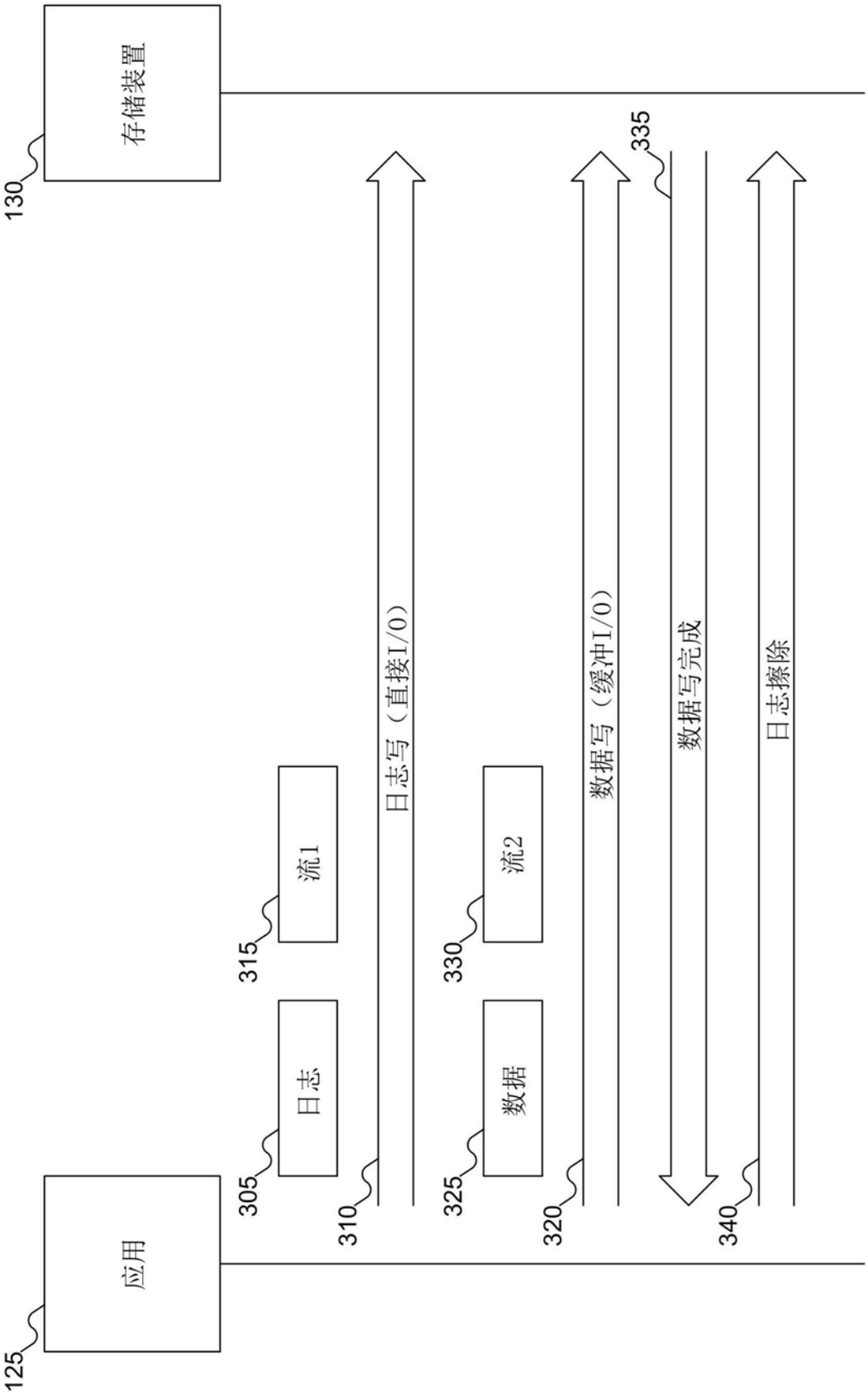

存储装置
130
335
流1
315
日志写（直接I/O）
流2
330
数据写（缓冲I/O）
数据写完成
日志擦除
日志
305
310
数据
325
320
340
应用
125

图3

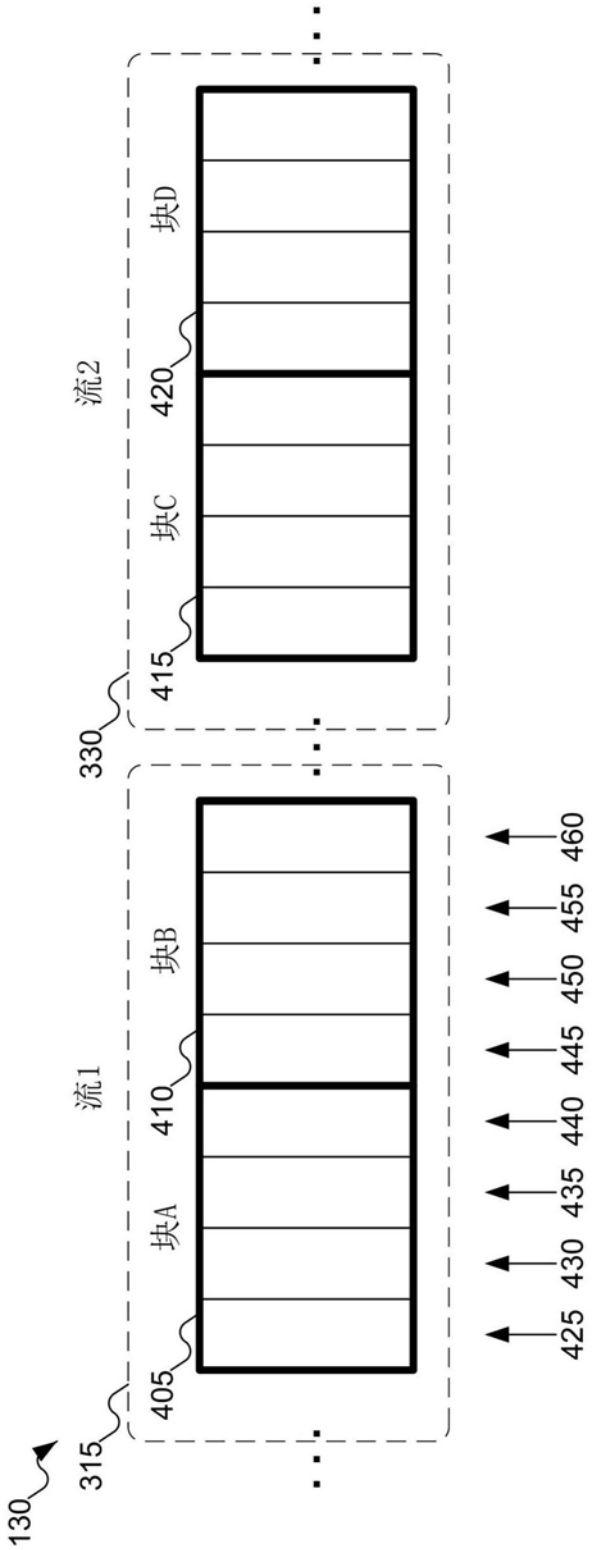

130
315
流1
405
块A
410
块B
330
流2
415
块C
420
块D
425
430
435
440
445
450
455
460

图4

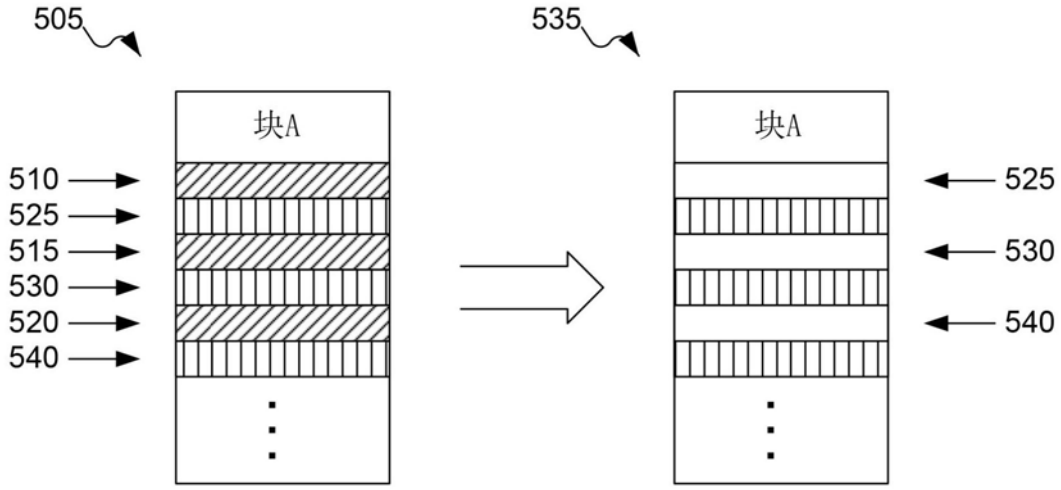

505
535
块A
块A
510
525
515
530
520
540
525
530
540

图5A

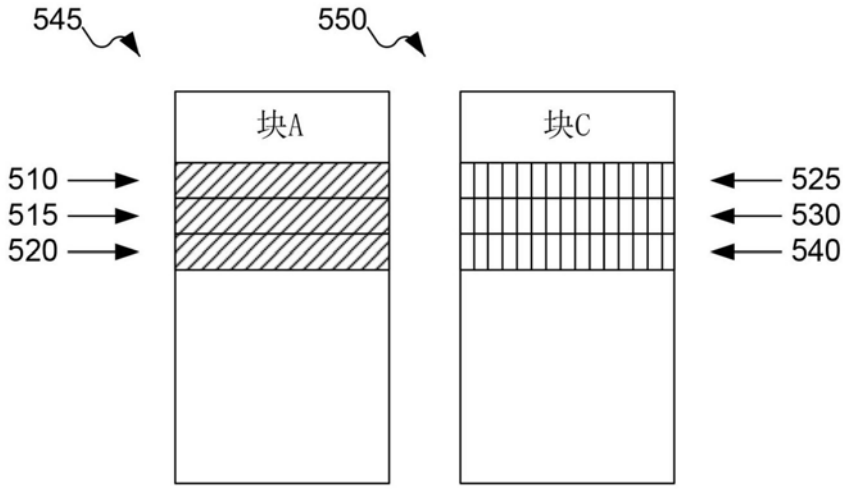

545
550
块A
块C
510
515
520
525
530
540

图5B

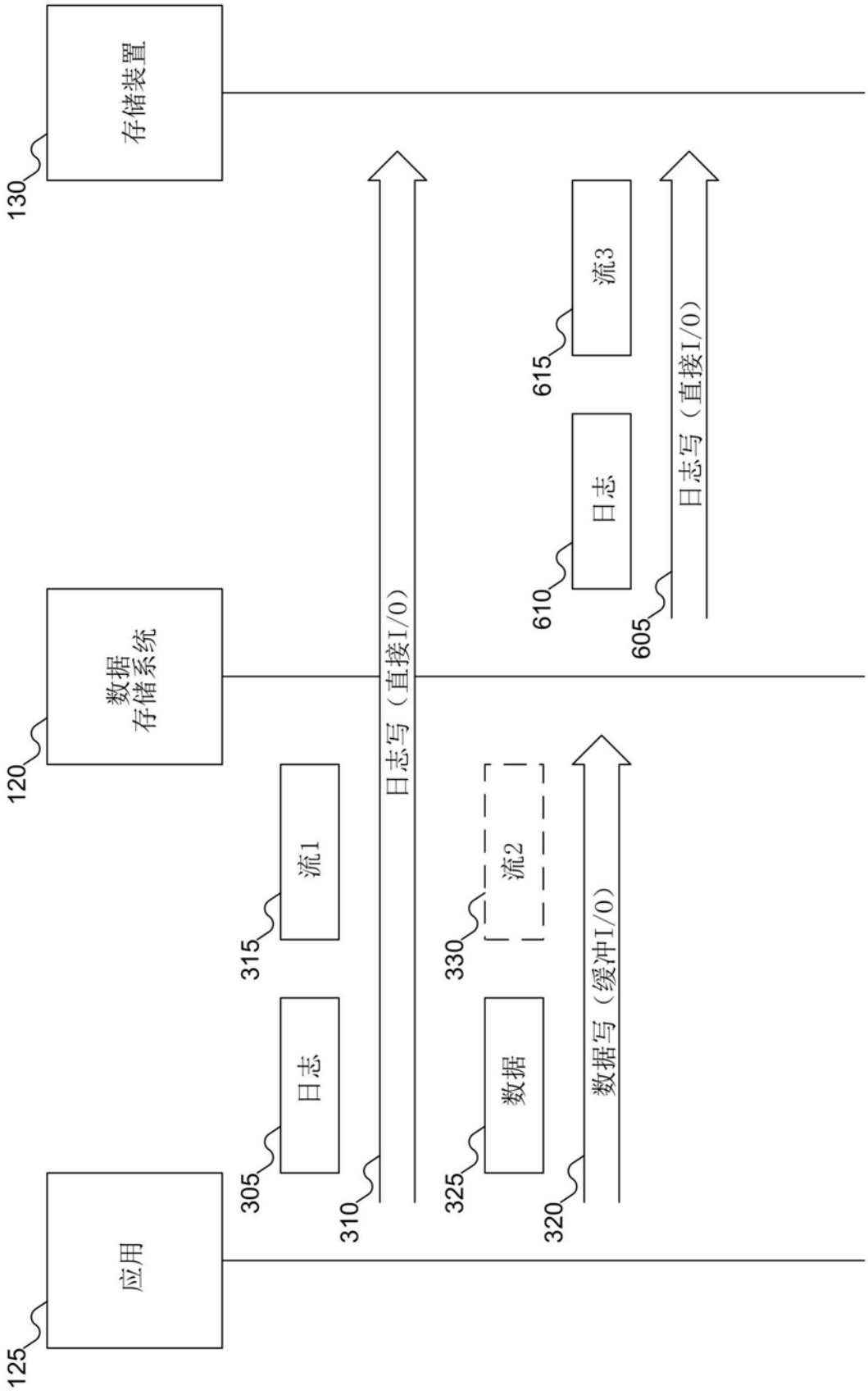

存储装置
130
数据
存储系统
120
应用
125
日志
305
流1
315
日志写（直接I/O）
310
数据
325
流2
330
数据写（缓冲I/O）
320
日志
610
流3
615
日志写（直接I/O）
605

图6A

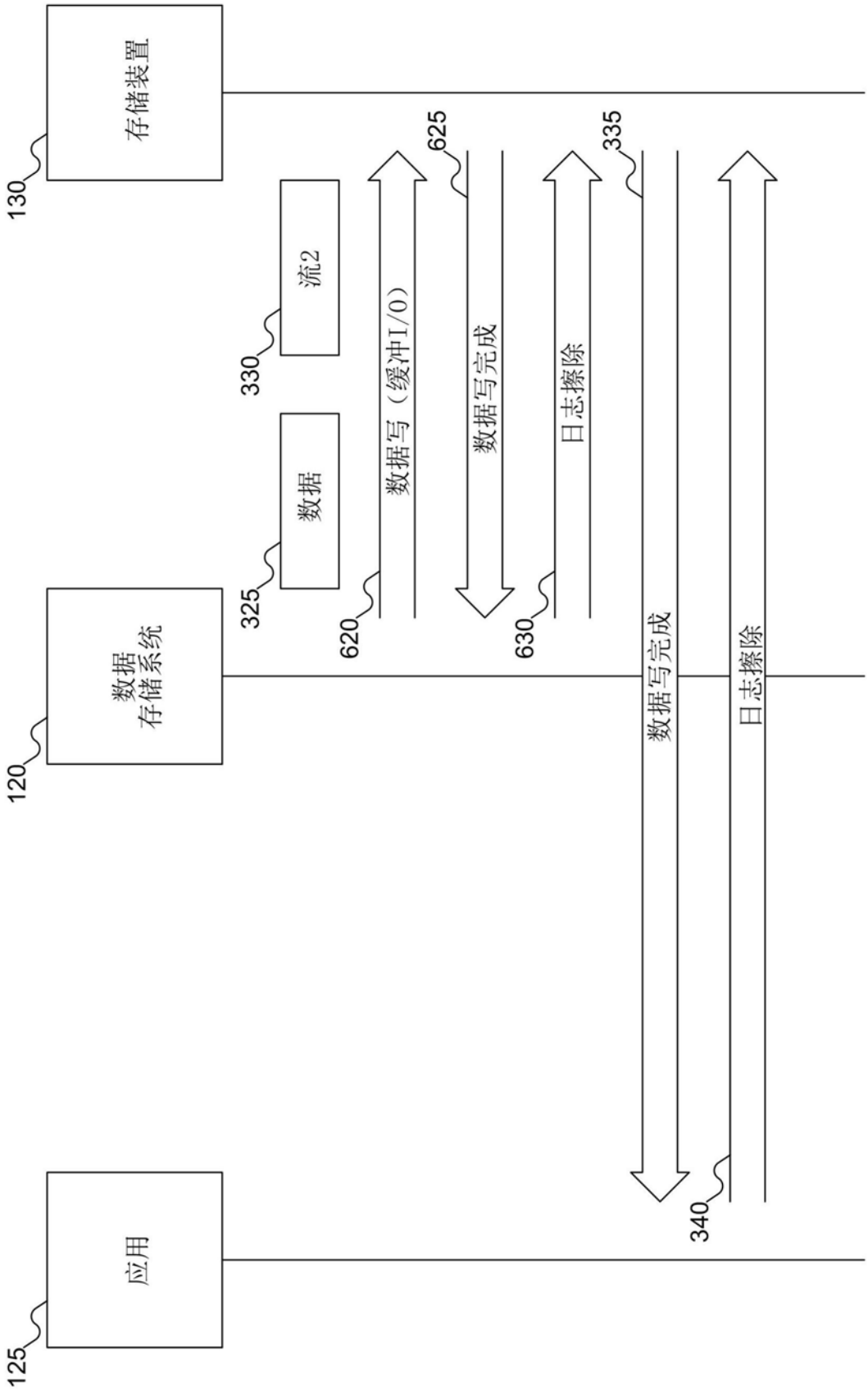

存储装置
130
流2
330
数据
325
数据写（缓冲I/O）
620
625
数据写完成
日志擦除
630
335
数据写完成
日志擦除
340
数据
存储系统
120
应用
125

图6B

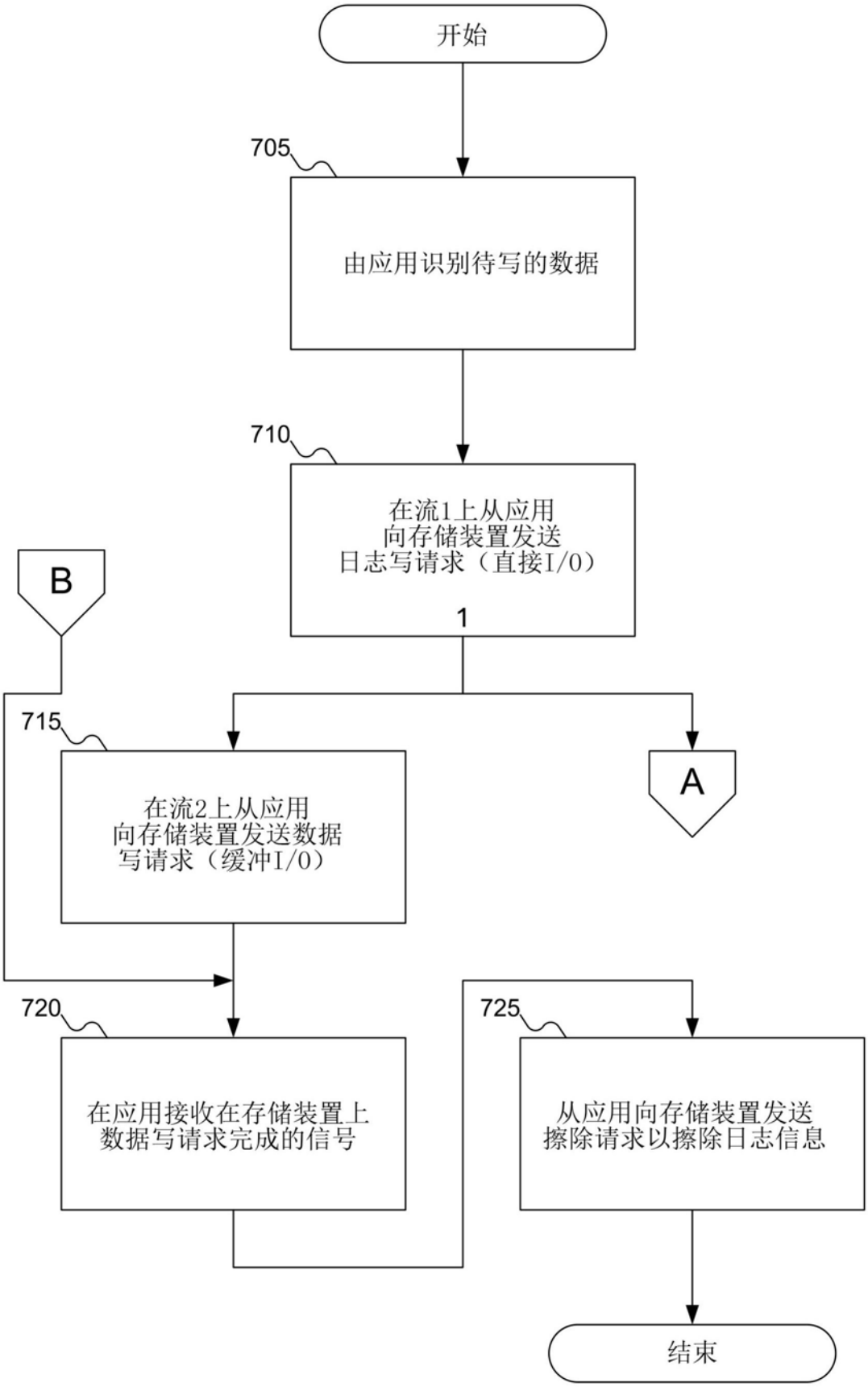

开始
705
由应用识别待写的数据
710
在流1上从应用
向存储装置发送
日志写请求（直接I/O）
1
B
715
在流2上从应用
向存储装置发送数据
写请求（缓冲I/O）
A
720
在应用接收在存储装置上
数据写请求完成的信号
725
从应用向存储装置发送
擦除请求以擦除日志信息
结束

图7A

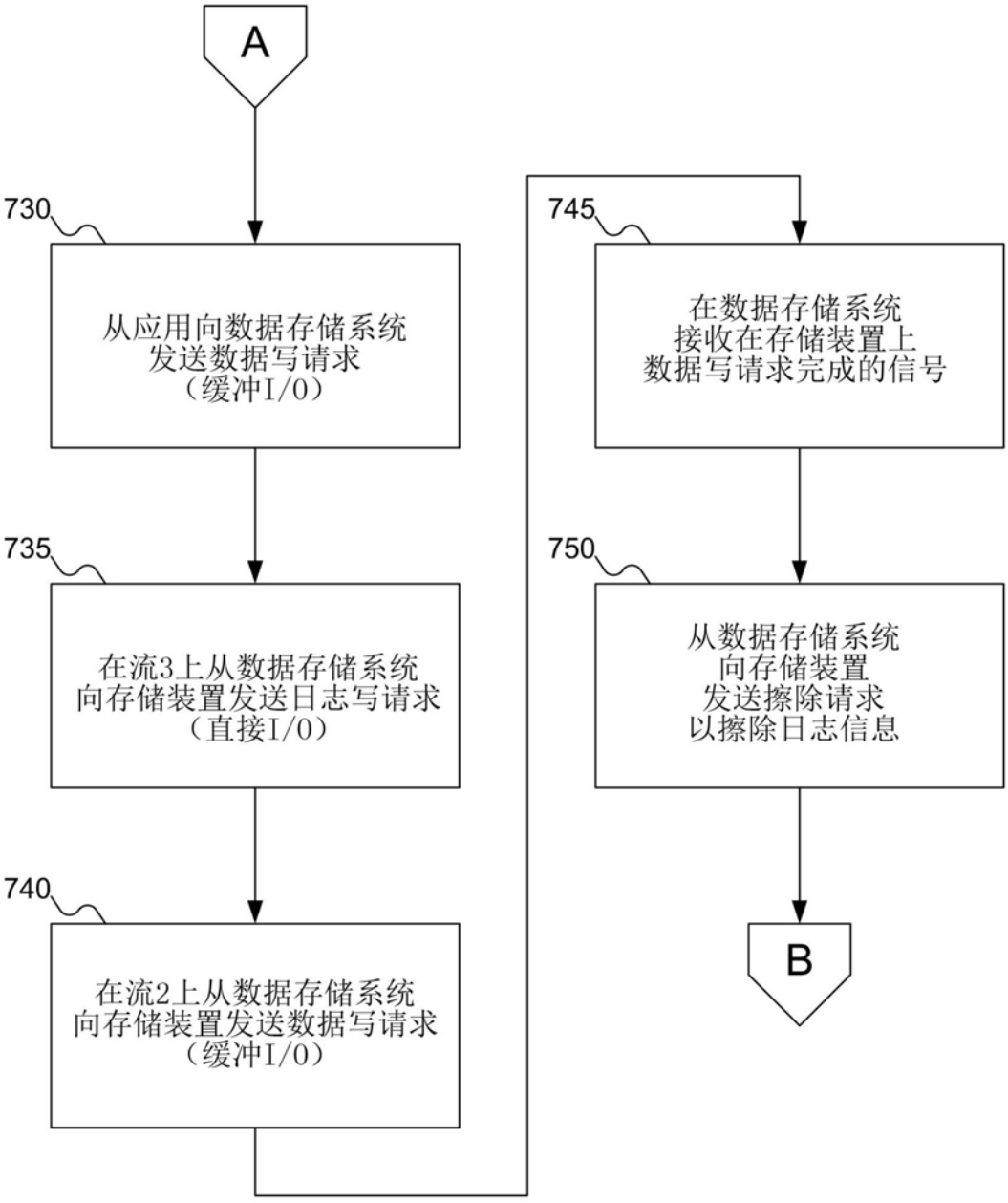

A
730
从应用向数据存储系统
发送数据写请求
（缓冲I/O）
735
在流3上从数据存储系统
向存储装置发送日志写请求
（直接I/O）
740
在流2上从数据存储系统
向存储装置发送数据写请求
（缓冲I/O）
745
在数据存储系统
接收在存储装置上
数据写请求完成的信号
750
从数据存储系统
向存储装置
发送擦除请求
以擦除日志信息
B

图7B