(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109388337 A
(43)申请公布日 2019.02.26

(21)申请号 201710799740.X

(22)申请日 2017.09.07

(30)优先权数据
15/667,109 2017.08.02 US

(71)申请人 东芝存储器株式会社
地址 日本东京都

(72)发明人 Y·克莱因 M·施纳尔奇

(74)专利代理机构 永新专利商标代理有限公司 72002
代理人 刘瑜 王英

(51)Int.Cl.
G06F 3/06(2006.01)

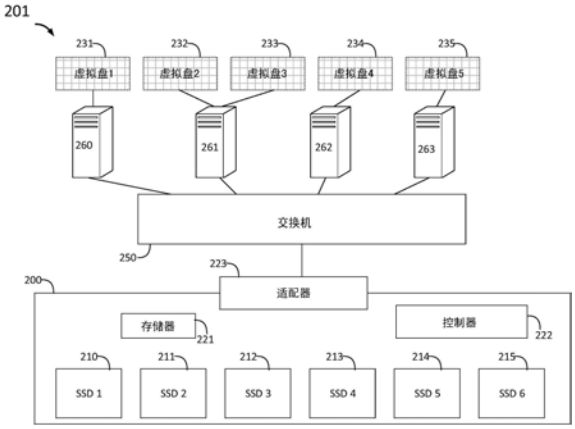
权利要求书3页 说明书11页 附图8页

(54)发明名称

存储设备,以及对存储设备中的多个固态驱动器之间的利用率进行平衡的方法

(57)摘要

存储设备包括第一SSD、第二SSD和控制器。控制器能够计算第一SSD的第一利用率参数和第二SSD的第二利用率参数。如果第一利用率参数小于阈值并且第二利用率参数超过该阈值,则控制器识别存储在第一SSD上的要去除的数据范围。从第一SSD中去除该数据范围导致第一利用率参数超过该阈值。然后,控制器将来自第一SSD的数据范围迁移到第二SSD。



CN 109388337 A

1.一种存储设备,包括:

第一固态驱动器(SSD)和第二SSD,被配置为存储数据;以及

控制器;

其中所述控制器被配置为:

计算所述第一SSD的第一利用率参数和所述第二SSD的第二利用率参数;

如果所述第一利用率参数小于阈值并且所述第二利用率参数超过所述阈值:

识别存储在所述第一SSD上的数据的要去除的数据范围,其中去除来自所述第一SSD的所识别的数据范围导致所述第一利用率参数超过所述阈值,以及

将与来自所述第一SSD的所识别的数据范围对应的数据迁移到所述第二SSD。

2.根据权利要求1所述的存储设备,其中,

所述第一利用率参数是第一有效预留空间值,以及

所述阈值是与目标写入放大值相关联的第二有效预留空间值。

3.根据权利要求1所述的存储设备,所述控制器还被配置为:

在将与来自所述第一SSD的所识别的数据范围对应的数据迁移到所述第二SSD之前,计算包括来自所述第一SSD的所识别的数据范围的所述第二SSD的第三利用率参数;

将所述第三利用率参数与所述阈值进行比较;以及

如果所述第三利用率参数超过所述阈值,则继续将与来自所述第一SSD的所识别的数据范围对应的数据迁移到所述第二SSD。

4.根据权利要求3所述的存储设备,其中如果所述第三利用率参数不超过所述阈值,则所述控制器还被配置为:

确定在保持所述第三利用率参数高于所述阈值的同时所述第二SSD能够接受的第一数据量;

将所确定的第一数据量迁移到所述第二SSD;

选择具有超过所述阈值的第四利用率参数的第三SSD;

确定第二数据量,其小于或等于来自所述第一SSD的所识别的数据范围减去所确定的在保持所述第四利用率参数高于所述阈值的同时所述第三SSD能够接受的第一数据量;以及

将所确定的第二数据量迁移到所述第四SSD。

5.根据权利要求3所述的存储设备,其中,在计算包括与来自所述第一SSD的所识别的数据范围对应的数据的所述第二SSD的所述第三利用率参数之后,所述控制器还被配置为:

计算不包括与所识别的数据范围对应的数据的所述第一SSD的第四利用率参数;

计算包括所述第三利用率参数和所述第四利用率参数的跨多个可用SSD的多个利用率参数的标准差;以及

如果所述标准差低于预定值,则将与来自所述第一SSD的所识别的数据范围对应的数据迁移到所述第二SSD。

6.根据权利要求3所述的存储设备,其中如果所述第三利用率参数小于所述阈值,则所述控制器还被配置为识别具有超过所述阈值的第四利用率参数的第三SSD。

7.根据权利要求1所述的存储设备,还包括适配器,其被配置为允许主机暴露虚拟盘以将数据传送到所述第一SSD和所述第二SSD。

8.根据权利要求1所述的存储设备，还包括被配置为记录所存储的数据的位置的映射表。

9.根据权利要求8所述的存储设备，其中所述控制器还被配置为使用所述映射表来确定与要从所述第一SSD去除的所识别的数据范围对应的数据的位置。

10.根据权利要求9所述的存储设备，其中所述控制器还被配置为通过查询所述映射表来计算所述第一利用率参数和所述第二利用率参数。

11.根据权利要求10所述的存储设备，其中，所述控制器被配置为利用被迁移到所述第二SSD的与所识别的数据范围对应的数据的新位置来更新所述映射表。

12.根据权利要求1所述的存储设备，其中，所述控制器被配置为将与来自所述第一SSD的所识别的数据范围对应的数据迁移到所述第二SSD的未被利用的区域。

13.根据权利要求1所述的存储设备，其中，所述控制器被配置为将与来自所述第一SSD的所识别的数据范围对应的数据迁移到所述第二SSD的未被利用的两个或更多个区域。

14.根据权利要求1所述的存储设备，其中，所述控制器被配置为确定能够迁移到所述第二SSD以导致所述第一利用率参数超过所述阈值的数据的最小数据范围。

15.一种对存储设备中的多个固态驱动器(SSD)之间的利用率进行平衡的方法，所述方法包括：

在控制器处计算至少一个第一SSD的第一利用率参数；

确定所述至少一个第一SSD的所述第一利用率参数小于阈值；

选择具有超过所述阈值的第二利用率参数的第二SSD；

识别存储在所述第一SSD上的数据的数据范围，所识别的数据范围具有的利用率级别使得随着去除来自所述第一SSD的所识别的数据范围而使所述第一SSD的所述第一利用率参数超过所述阈值；以及

将与来自所述第一SSD的所识别的数据范围对应的数据迁移到所述第二SSD。

16.根据权利要求15所述的方法，其中计算所述第一利用率参数包括计算有效预留空间值。

17.根据权利要求16所述的方法，其中计算所述第一利用率参数包括查阅映射表以确定在对应于所述第一SSD的第一逻辑块寻址(LBA)与第二LBA之间的利用率。

18.根据权利要求17所述的方法，其中计算至少一个第一SSD的所述第一利用率参数包括计算多个SSD中的每一个的利用率参数。

19.根据权利要求18所述的方法，所述方法还包括：

将所述多个SSD中的每一个的利用率参数与所述阈值进行比较；

对所述多个SSD中小于所述阈值的SSD子集进行排序；以及

从所述SSD子集中选择具有最小利用率参数的SSD。

20.根据权利要求19所述的方法，所述方法还包括：确定所述第二SSD是否能够接受与所识别的数据范围对应的数据并且继续超过所述阈值。

21.根据权利要求20所述的方法，其中如果确定所述第二SSD不能接受与所识别的数据范围对应的数据并且继续超过所述阈值，则所述方法还包括：

确定在保持所述第二利用率参数高于所述阈值的同时所述第二SSD能够接受的第一数据量；

将所确定的第一数据量迁移到所述第二SSD；

选择具有超过所述阈值的第三利用率参数的第三SSD；

确定第二数据量，其小于或等于来自所述第一SSD的所识别的数据范围减去所确定的在保持所述第三利用率参数高于所述阈值的同时所述第三SSD能够接受的第一数据量；以及

将所确定的第二数据量迁移到所述第三SSD。

存储设备,以及对存储设备中的多个固态驱动器之间的利用率进行平衡的方法

[0001] 相关申请

[0002] 本申请基于并要求于2017年8月2日提交的美国非临时专利申请15/667,109的权益;其全部内容通过引用并入本文。

技术领域

[0003] 本发明一般涉及存储设备,以及对存储器中的多个固态驱动器之间的利用率进行平衡的方法。

背景技术

[0004] 基于闪存的存储介质,诸如在固态驱动器(SSD)中采用的存储器件,包括以平面、块和页面为单位布置的闪存单元。每个平面包含多个块,块是能够被擦除的最小单元。每个块包含多个页面,页面是能够以数据写入的最小单元。在SSD中,存储器控制器负责映射要写入物理闪存页的或要从物理闪存页读取的应用数据的逻辑块地址。

[0005] 存储介质盘具有有限的使用寿命(其在使用寿命期间是可操作的)。硬盘驱动器(HDD)是随着使用而磨损的机械设备,可能会意外地和灾难性地出故障。SSD通常具有HDD两倍以上的寿命,但与HDD相比,它们的磨损方式不同。由于NAND闪存的特点,SSD的使用寿命由NAND闪存能够承受的写操作次数决定。写操作也称为编程/擦除(P/E)周期,因为只有在使用擦除存储器命令去除先前存储器内容之后,才可以使用程序存储器命令写入数据。

[0006] 由于只有在首先擦除闪速存储器中的先前数据之后数据才能被覆盖,所以通常的做法是在SSD中提供额外的工作空间,该额外的工作空间用作闪存块的新擦除的区域,在该区域可以写入新数据。这种做法称为预留空间(OP)。新数据写入后,先前的数据被标记为无效。最终,闪存块中存在足够数量的无效数据页,以使系统开始一个被称为“垃圾收集(GC)”的过程。在GC过程期间,剩余的有效数据页被写入新擦除的块,而原始块(现在其所有页面被标记为无效)被擦除。然后,被擦除的块可以被包括在(可以向其写入数据的)新擦除的块的池中。GC过程涉及到除了对新数据的原始写入之外的额外的数据写入。这种副作用被称为写入放大(WA),其中WA因子是额外的NAND闪存数据写入的平均数与原始写入之比。WA导致SSD的寿命降低,通过有效地将闪存设备的P/E周期限值除以WA因子。例如,WA因子为10的情况下,SSD的寿命将是按P/E周期限值所预测的寿命的十分之一。

[0007] WA与OP成反比,其中减少OP具有增加WA的效果。图1示出了现有技术的绘图100,其示出了OP和WA之间的相关性。绘图100包括表示OP的x轴102和表示WA的y轴104。OP和WA之间的相关性由迹线106表示,其在108处具有拐点。绘图100是OP和WA的相关性的公知模型。在由IBM苏黎世研究实验室的X-Y Hu等人于以色列的海法的SYSTOR2009发表的文章“Write Amplification Analysis in Flash-based Solid State Drives(基于闪存的固态驱动器中的写入放大分析)”一文中,显示和描述了对OP和WA相关性进行建模的类似绘图。

[0008] 从绘图100可以看出,减少OP降低了器件的耐久水平,因为增加WA就增加了每次应

用程序数据写入的NAND闪存写入数量，并降低了具有指定P/E周期限值的器件的寿命。这种关系不是线性的，随着OP在108处降到拐点以下，WA的增长速率变得越来越严重。

[0009] 增加SSD设备容量并减少OP以降低成本的要求导致磁盘寿命缩短，早期磁盘故障风险增加，伴随着对存储数据的成本和可用性/存取的严重影响。因此，长期以来感到需要纠正当今SSD系统中固有的问题。

发明内容

[0010] 本发明的实施例实现了一种存储设备以及对存储器中的多个固态驱动器之间的利用率进行平衡的方法，其能够增加SSD设备的容量并减少OP以降低成本。

[0011] 在一方面，存储设备包括第一SSD、第二SSD和控制器。控制器能够计算第一SSD的第一利用率参数和第二SSD的第二利用率参数。如果第一利用率参数小于阈值并且第二利用率参数超过该阈值，则控制器识别存储在第一SSD上的要去除的数据范围。去除来自第一SSD的该数据范围导致第一利用率参数超过该阈值。然后，控制器将来自第一SSD的该数据范围迁移到第二SSD。

[0012] 在另一方面，一种对存储设备中的多个SSD之间的利用率进行平衡的方法包括：在控制器处计算至少一个第一SSD的第一利用率参数，确定所述至少一个第一SSD的第一利用率参数小于阈值，并且选择具有超过该阈值的第二利用率参数的第二SSD。该方法还包括识别存储在第一SSD上的数据范围，其中数据范围具有的利用率级别使得随着去除来自第一SSD的该数据范围而使第一SSD的第一利用率参数超过该阈值，并且将来自第一SSD的该数据范围迁移到第二SSD。

附图说明

[0013] 图1示出了现有技术中的绘图，说明了预留空间和写入放大之间的相关性；

[0014] 图2示出了根据一个或多个实施例的具有多个SSD和多个虚拟盘的存储设备的框图；

[0015] 图3示出了根据一个或多个实施例的在存储设备中使用的虚拟映射的框图；

[0016] 图4A和图4B示出了根据一个或多个实施例的各种SSD和用于提取SSD的利用率级别的扩展应用编程接口(API)的框图；

[0017] 图5示出了根据一个或多个实施例的说明在整个SSD池范围内(across a pool of SSDs)的预留空间平衡的机制的框图；

[0018] 图6示出了根据一个或多个实施例平衡在SSD池范围内的预留空间的决策流程；

[0019] 图7示出了根据一个或多个实施例的平衡协议的第二阶段的决策流程；和

[0020] 图8示出了根据一个或多个实施例的用于平衡在整个SSD池范围内的利用率参数的方法。

具体实施方式

[0021] 图2示出了包括具有多个SSD和多个虚拟盘的存储设备200的系统201的框图。存储设备200包括多个SSD 210-215，控制器222，存储器221和适配器223。系统201还包括多个主机260-263和多个虚拟盘231-235。主机260-263通过开关250连接到存储设备200。

[0022] 为了清楚起见，图2示出了六个SSD 210-215和五个虚拟盘231-235。然而，应当理解，存储设备200可以包括任何数量的SSD以及任何数量的虚拟盘。类似地，图2示出了四个主机260-263，存储设备200可以包括任何数量的主机设备。术语“SSD”、“SSD驱动器”和“固态设备”可以互换使用。

[0023] 控制器222运行虚拟化软件，该虚拟化软件将多个虚拟盘231-235通过适配器223暴露给多个主机260-263。SSD 210-215用作后端存储器。控制器222是存储设备100上的本地控制器，存储设备100可以是任何标准的SSD存储设备。

[0024] 适配器223向远程主机260-263提供对存储设备200的访问。具有适当连接的远程主机可以通过物理交换机250和诸如iSCSI、光纤通道(FC)、NVME over Fabric、SCSI、或传输协议与经由主机中的驱动程序软件的物理连接的任何组合来访问存储设备200。适配器223可以是以太网端口、光纤通道(FC)端口、SAS、PCIe、InfiniBand(IB)端口或任何其他适当的端口。

[0025] 结果，远程主机(例如，主机260-263中的一个)被提供有被控制器222中的软件暴露的虚拟盘231-235。例如，主机260被提供有虚拟盘231，并且主机261被提供有虚拟盘232和233。主机260和261按照控制器222的软件、通过交换机250和适配器223与具有多个SSD 210-215的存储设备200交互，但主机260-263在设备上“看到”虚拟盘231-2135而不是物理SSD 210-215。每个主机将虚拟盘231-235看作包含数据，但数据实际上由(可被主机260-263经过交换机250和适配器223而访问的)存储设备中的物理SSD 210-215上的控制器222存储和管理。

[0026] 图3示出了说明在存储设备中使用的虚拟映射方案300的框图。图3包括具有第一虚拟映射表328的第一虚拟盘321、具有第二虚拟映射表327的第二虚拟盘322、第一SSD 311、第二SSD 312、第三SSD 313和第四SSD 314。

[0027] 虚拟映射方案300说明了使用第一映射表328将第一虚拟盘321映射到第一SSD 311所示出的第一虚拟化方案，第一映射表328使用虚拟地址到物理地址空间的一对一映射。第一虚拟盘321被映射到第一SSD 311中的第一物理范围331中。第一虚拟盘321中的数据的大小被设置为使得数据可以被连续地放置在第一SSD 311中。来自第一虚拟盘311或来自其他虚拟盘的附加数据321也可以连续放置在第一SSD 311中，直到第一SSD 311没有更多的空间。第一虚拟映射表328存储映射信息、SSD标识、起始逻辑块地址(LBA)的指示符和放置在第一SSD 311中的数据范围的长度。

[0028] 在第二个更复杂的示例中，虚拟映射方案300说明了指向多个物理设备上的多个范围的虚拟地址空间。这通过根据第二映射表328将第二虚拟盘322映射到多个SSD中的范围来示出，多个SSD包括第一SSD 311、第二SSD 312、第三SSD 313和第四SSD 314。如所看到的，第二虚拟盘322位于第一SSD 311的范围332、第二SSD 312中的范围334和335，第三SSD 313中的范围335和第四SSD 314中的范围336中。第二虚拟盘321的第二映射表328是其中存储虚拟盘数据的范围的分散集合列表(SGL)，包括其位置、起始LBA和范围的长度。

[0029] 来自第二虚拟盘321的数据从虚拟空间映射到多个SSD上的物理空间中。来自第二虚拟盘321的数据被分成范围(range)或团块(chunk)，所述范围或团块被放置在存储设备中的多个物理SSD上。

[0030] 如果虚拟盘中的数据太大以至于不能将数据放置在单个物理SSD上，那么更可能

会遇到第二个示例,其包括将来自第二虚拟盘321的虚拟盘数据映射到各种SSD上的范围,数据可能被分割成团块或范围并跨多个SSD放置。将第二虚拟盘321映射到多个SSD上也可以发生在成碎片的虚拟地址空间中,或者在已经使用了一段时间的存储设备中。例如,在存储设备已被使用一段时间并且虚拟盘已被创建和删除之后,需要进行虚拟盘到物理SSD的多范围映射,以在物理存储上分配虚拟盘数据。

[0031] 随着数据被存储在SSD上,SSD的已分配部分以及SSD的利用率级别增加。当SSD具有高的利用率级别时,有效OP可能被最小化并且WA可能很高。而且由于WA与SSD的寿命逆相关,因此高利用率的SSD可能会具有增加的磁盘故障风险。

[0032] 图4A和图4B分别示出SSD和用于提取(extract)SSD的利用率级别的扩展API的存储器表示。图4A示出了SSD 400的和典型的SSD参数的存储器表示401。SSD 400由NAND页面组成,并且通常包含数据的已利用页面422、坏块420和包含在有效OP 421中的空闲的或脏的NAND页。坏块420为了清楚起见而被放大显示,但是在大多数情况下坏块420将占据SSD 400的较小部分。SSD 400具有如图所示的总容量410。然而,只有SSD 400的总容量410中的一部分可供主机使用。SSD 400的可以写入数据的可用部分是导出容量416。为了使闪存转换层(FTL)能够将新数据存储并高效地写入已擦除的页面(具有低WA),并通过在后台独立地执行GC来管理无效数据,总容量410的一部分被保持为OP 415或备用空间。

[0033] 与其中通过从逻辑地址空间到物理地址空间的固定映射维持存储、且物理位置可以被自由重写而无需先擦除以前的数据的HDD相对照,SSD包含NAND页,并且并非所有的物理地址都映射到逻辑地址。在SSD中,逻辑到物理的映射不是固定的,而是映射随数据被重写而改变。SSD随着其被填充的数据的量而被利用,且SSD物理页地址仅在数据写入SSD时被分配。写入SSD的数据量是利用率级别(utilization level),而尚未分配和写入数据的SSD的剩余部分是有效OP区域(effective OP area)。如果SSD没有被充分利用,则未分配的部分要大得多,且有效的OP要高得多。

[0034] 例如,在1TB(太字节)存储磁盘中,利用10%的OP,将有100GB(千兆字节)的OP存储区域对于用户是不可用的,且有900GB的存储可用。900GB的可用存储可被暴露给用户。将100GB的OP区域作为备用页面保留,这就允许SSD在后台执行GC协议,或者以SSD的所需或指定的写入数据传输速度接受新的数据写入命令。虽然用户可以访问900GB,但如果用户仅将400GB的数据写入磁盘,则有效OP区域是剩余的500GB的可访问存储空间加上100GB的预留OP区域。

[0035] OP区域给SSD提供了立即将新数据覆盖为备用的先前擦除的数据块和页面的能力,以便在后台对已被通过覆盖而无效化的旧数据执行GC协议。在没有OP的情况下,GC协议将不得不与数据覆盖同步执行,以使已擦除的页面空间放置新数据,从而减慢SSD的写入表现。虽然OP通常是在SSD的制造期间被确定的并且是已知的,但SSD的有效OP在将数据写入SSD、覆盖现有数据以及去除数据等方面对SSD进行使用期间改变(使用为SSD专门设计和定义的所谓的“修剪(Trim)”命令,其用于对在主机文件系统中在数据被擦除之后释放的LBA进行解除分配)。SSD的当前有效OP在传统存储设备中对于用户或对于存储设备中的本地控制器不是可用的,因为逻辑到物理地址映射的管理以及NAND闪存页面的分配和解除分配保持在SSD的闪存控制器的私有域内。

[0036] 在操作阶段,SSD 400包含有效数据页、无效或脏数据页(例如已被覆盖或修剪的

数据)以及空闲页。执行GC处理后,脏数据页面可能会作为空白页面而变得可用。此外,在操作期间,数据块可能由于超过P/E周期限值或其他NAND故障机制而出故障,并且变得不可操作。这些被表示为坏块420,其随着存储设备的使用随时间而增加,使得存储盘的总容量恶化。

[0037] SSD 400中的OP被定义为SSD 400的未分配或未利用的数据页与总可用数据页之间的比率。在SSD被完全填充以数据(即到其导出容量416)的情况下,未分配的数据页是剩余的备用页面415。在这种情况下,假设SSD 400中没有坏块,则通过以下等式(等式1)将OP表示为比率。

[0038] $$OP = \frac{(\text{总容量}-\text{导出容量})}{\text{导出容量}}$$ (等式1)

[0039] 然而,在大多数情况下,SSD 400将不会被以数据完全填充达到总出口容量416。用户实际上将仅利用(已利用数据422)在SSD 400上的可用于写入的空间的一部分(总容量410减去坏块420或导出容量416)。因此,SSD 400的更实际的参数是由以下等式(等式2)表示的有效OP。

[0040] $$\text{有效}\,OP = \frac{(\text{总容量}-\text{已利用数据}-\text{坏块})}{\text{已利用数据}}$$ (等式2)

[0041] 如等式2所示,有效OP 421由总容量410减去已利用数据422和坏块420来确定。当有效OP低时,SSD被高度地利用,并且可能不稳定以及更容易发生故障。在传统的存储设备中,关于有效OP以及SSD的利用率的信息对于用户来说不是可用的。然而,如果信息可用,存储设备中的软件可以利用这些信息来提供更高的设备稳定性和更长的使用寿命。

[0042] 根据本发明的一个实施例,SSD 400提供扩展API,其以示出已利用数据422的量和坏块420的数量的数据的形式来提供关于SSD 400的信息。API确定用于包括有效OP的SSD 400的利用率数据信息,并且主机或客户端计算机可以使用这样的数据来平衡存储设备中的SSD的利用率。可替代地,或者另外地,主机可以向用户提供信息。通过平衡SSD上的利用率,存储设备的整体使用寿命和操作得到改善,因为对于被过度利用的设备而言,它将WA降低了与对于被利用不足的SSD而言WA的相应增加相比更大的量。

[0043] 在一些实现中,SSD 400提供扩展API,其提供关于SSD 400的地址空间的特定范围430的数据利用率431的信息,如图4B所示。例如,用户可以提供起始LBA和长度,并且SSD通过API返回所提供的范围的数据利用率431。

[0044] 在一些实现中,范围430由用户指定。在一些实现中,范围430由制造商在API中指定。

[0045] OP是通过SSD上的已利用空间和可用空间的比较来提供的,例如由等式1和2所描述的。WA是SSD中的一种现象,其中确实地写入存储器的实际数据量大于打算要写入的数据量(差异是由于有效数据页从部分有效的数据块移动到新的数据块从而得到完全无效的(然后可以被擦除的)数据块而导致的,这是在GC操作中发生的过程)。例如,如果用户打算将1kB(千字节)的数据写入磁盘,并且WA为10,则写入磁盘的数据量可能实际上为10kB。OP和WA可以预测存储磁盘的耐久性和性能特性。在上述示例中,具有的WA为10的磁盘将比具有的WA为1的磁盘快10倍地达到其耐久极限,且将具有恶化的性能,比具有的WA为1的磁盘慢大约10倍。根据有效的OP来推导WA,在Luojie等人的arXiv:1110.4245v1的文献“An Improved Analytic Expression for Write Amplification in NAND Flash”中有详细的

描述，该文献通过引用并入本文。

[0046] WA与OP成反比，并且关系是非线性的，如图1所示的现有技术的绘图100所示。小的OP与非常高的WA相关，而在拐点108之后在约20％OP处，WA随着OP的增加变化很小。在拐点108之前，在例如5％的有效OP处，WA为11。在拐点108之后，在有效OP为35％的情况下，WA为2。磁盘的有效OP对磁盘的WA具有大的影响，并且对磁盘的性能、稳定性和耐久性具有大的影响。在较高的有效OP处，与在较低的有效OP处相比，磁盘的耐久性和性能更优。

[0047] SSD中的OP通常被最小化，以便降低每个可用数据存储单元对于客户的成本。然而，供应商为操作时间段提供保障并优化OP，因此产品在保障的操作时间段之前不会出故障。

[0048] 操作时间段的保障可以基于WA和与P/E计数相关的每天磁盘写入(DWPD)进行近似。由于高利用率或坏块，具有较低的有效OP的SSD可能会在其他SSD之前达到其耐久极限，并且可能会较早出故障。因此，将OP最小化以降低对于客户的成本可能会对产品的使用寿命(operational lifetime)产生不利影响，实际上可能会导致消费者或供应商的成本更高。

[0049] 为了最大化存储设备的使用寿命，可以在SSD的使用寿命期间优化SSD的OP，并且应在每个时间点选择SSD以提供SSD资源的优化。例如，拐点值108或OP的其他值可以被选择为阈值，有效OP应保持在该阈值之上以使WA保持低。在一些实现方式中，阈值用作警报，其通知FTL:SSD中的备用空间量接近最小值，该最小值可能使SSD变为丧失功能状态或只读状态。

[0050] 虽然在存在由存储设备控制器管理的多个单独的SSD的存储设备中，当随着SSD的存储变得更多地被利用而使OP值降低并接近拐点值108时，对于在单个SSD的控制器中的FTL可能存在有限的用来调节OP的范围，但可以通过在SSD之间迁移数据来更灵活地调整和平衡各个SSD的OP。在一些实施例中，存储设备控制器使用由SSD提供的扩展API来访问OP、坏块和数据映射信息。使用该信息，控制器可以将数据从一个SSD迁移到另一个SSD，从而修改源SSD和目的地SSD的有效OP。

[0051] 在一些实施例中，阈值可以是单个阈值。在一些实现方式中，阈值可以是多个阈值。在一些实现方式中，阈值可以是任何其他OP相关阈值。在一些实现方式中，当OP超过特定SSD中的阈值时，存储设备的控制器执行平衡协议以将SSD的OP恢复到阈值内，以优化SSD的操作并防止故障。在一些实现方式中，阈值是对应于2或3的目标WA的有效OP，例如20-40％的有效OP。在WA为2或3的情况下，与较高的WA相比，耐用性显著降低，SSD的性能提高。

[0052] 图5示出了在整个SSD池范围内的OP平衡的框图500，说明了由控制器的扩展API定义的平衡协议。框图包括第一SSD 510、第二SSD 511、第三SSD 512和第四SSD 513。第一SSD 510包括：第一范围530，其包括从虚拟盘写入的数据；以及第二范围531，其也包括从相同或不同的虚拟盘写入的数据。第一SSD 510还具有由箭头的垂直高度指示的第一有效OP 520。第一有效OP可由控制器(如图2中的控制器222)确定，如下所述。类似地，第二SSD 511具有第二有效OP 521，第三SSD 512具有第三有效OP 522，并且第四SSD具有第四有效OP 523，每个均由每个相应箭头的垂直高度指示，并且以同样方式由控制器确定。

[0053] 如上所述，有效OP与SSD的利用率有关，并且预测了SSD的耐久性和性能。与具有更高有效OP的SSD相比，具有非常低有效OP的SSD将更接近其耐久极限并且性能差。通过在SSD

池中在整个SSD范围内平衡利用率，存储设备具有更好的整体性能，并且SSD的故障可能性降低。图5中的每个SSD的有效OP不平衡，如不同高度的垂直箭头所示。通过在控制器上使用扩展API来确定每个SSD的利用率并根据预定的选择启发式来平衡SSD，可以优化存储设备的性能和稳定性。

[0054] 控制器(例如，图2的控制器222)将每个SSD的有效OP确定为SSD的利用率的量度。控制器通过查询表中两个LBA值之间的利用率而根据映射表来确定每个SSD的利用率。控制器可以访问映射表，并且可以通过查询映射表来确定每个SSD上哪些LBA已经被写入以及哪些没有被写入，从而确定利用率级别或有效OP。控制器可以查询映射表以确定SSD或SSD的一部分的利用率。利用率级别或有效OP也可以根据等式2被计算为第一SSD上的已利用空间与总可用空间之间的比率。

[0055] 当特定的SSD的有效OP低于阈值时，控制器执行对另一个SSD中的可用空间的检查。如果控制器识别具有可用空间的SSD，则控制器将一个范围的LBA迁移到具有足够可用(即，未分配的)空间的新的SSD。例如，如果控制器确定第一有效OP 520低于阈值，则控制器检查剩余SSD(即第二SSD 511、第三SSD 512和第四SSD 513)的利用率，以确定是否SSD中任一具有可用空间。在该示例中，控制器确定第三SSD 512具有用于第二范围531的足够的可用空间，并且第三SSD 512具有高于阈值的有效OP。然后，控制器将第二范围531从第一SSD 510迁移到第三SSD 512。第一SSD 510的有效OP 520应该随着将第二范围531从第一SSD向第三SSD 512的资源的迁移而增加。

[0056] 控制器可以通过访问虚拟映射表来确定可以向其迁移数据的目的地SSD。控制器中的扩展API可以访问映射表，并且可以提取LBA到SSD中的物理页面的映射。使用此信息，扩展API会计算SSD的多少区域或页面被映射到虚拟空间，以及多少区域或页面当前未被映射到虚拟空间(因此未分配且可用)。因此，控制器可以确定潜在目的地SSD的各个区域的利用率。

[0057] 尽管图5中的示例示出了单个范围(第二范围531)从第一SSD 510到第三SSD 512的迁移，在其他实现方式中，控制器可以将来自SSD的多个范围迁移到多个其他SSD。控制器还可以通过将单个范围分割成多个较小范围以放置在其他SSD上，而将单个范围(例如第二范围531)迁移到跨多个SSD的多个目标范围。

[0058] 要迁移的范围可以由控制器通过扩展API的用于范围利用率的参数来选择。控制器计算为使有效OP 520超过阈值而必须从第一SSD 510传送的已利用数据量。然后，控制器选择满足所需数据量的至少一个数据范围。在一些实现方式中，如果第一SSD 510上的多个数据范围满足所需的数据量，则控制器选择将满足该要求的最小范围迁移，从而最小化任何迁移惩罚。在一些实现方式中，如果第一SSD 510上的多个数据范围满足所需的数据量，则控制器选择满足要求的最少访问范围。在一些实现方式中，如果第一SSD 510上的多个数据范围满足所需数据量，则控制器选择满足要求的最多访问范围。

[0059] 控制器执行平衡协议，使得SSD池均匀地持续(endure)，并且SSD以低WA操作，以增加SSD的寿命并降低故障的可能性。控制器执行平衡协议以便在整个SSD池范围内均匀分布有效OP，这是通过在迁移后预先计算SSD的最终结果OP并选择SSD上的最佳位置以将一个迁移范围或多个迁移范围在剩余SSD范围内放置。在一些实现方式中，优化是基于对迁移后OP的标准差进行最小化。在一些实现方式中，如果控制器确定特定SSD中的坏块或写错误的数

量已经增加到超过阈值，则控制器可以停止向SSD写入数据，并将剩余数据从该SSD重新平衡到剩余的SSD。在SSD满的情况下，例如当SSD中的坏块的数量使得SSD被阻止运行时，由控制器执行的平衡协议通过利用设备中的其他SSD中的可用空间来防止SSD出故障。

[0060] 使用数据范围的迁移来在整个SSD池范围内对OP进行优化和平衡的SSD池平衡协议，使用经由扩展API提供给控制器的详细数据利用率信息。如上所讨论的，可以确定每个SSD的数据利用率信息，或者可以确定SSD上的每个虚拟映射范围的更详细的数据利用信息。

[0061] 在一些实现方式中，除了对物理SSD上的现有数据执行池平衡协议之外，控制器将不会将虚拟盘映射到SSD上，从而违反操作有效的OP阈值。以这种方式，即使由于耐久性或死机故障而导致的坏块累积，这些设备总是保持在可操作状态。

[0062] 控制器的逻辑可以随着不同的实施情况而变化，使得SSD的利用率被平衡以提高存储设备的性能。控制器可以具有扩展的API，其使用以上讨论的原理来优化性能、耐久性或导出容量。在一些实现方式中，控制器可以具有扩展API，其向外部用户提供与SSD的利用率相关的信息，而不需要在SSD之间实际移动数据。

[0063] 图6示出了用于控制在整个SSD池范围内的OP平衡的决策流程600。控制器(例如图2中的控制器222)可以使用决策流程600来平衡SSD池中的有效OP。

[0064] 在步骤602，控制器被触发以每个预定义的时间段(例如每小时或每天一次)执行平衡协议。控制器可以检测基于时间的触发或由事件的发生来触发。在一些实现方式中，平衡协议被触发的时间段由制造商指定。在一些实现方式中，平衡协议被触发的时间段由用户指定。在一些实现方式中，平衡协议按预定义的时间段触发，并且还由事件(诸如用户的命令)触发。

[0065] 在步骤604中，在平衡协议的触发已被触发之后，控制器检查SSD以检索每个SSD的有效OP级别。控制器可以使用扩展API来通过访问映射表并且从映射表中的LBA确定每个SSD的在第一LBA和第二LBA之间的SSD部分的利用率来确定有效OP。在一些实现方式中，控制器通过使用等式2，根据SSD上的利用率和坏块的量相对于SSD的总容量来计算SSD的有效OP，从而确定有效OP。

[0066] 在一些实现方式中，控制器检索与SSD的利用率相关或与SSD的一部分的利用率相关的不同参数。在步骤606中，控制器将每个SSD的有效OP级别与存储的阈值进行比较，并且确定是否存在具有不满足阈值的有效OP级别的SSD。阈值是由制造商设定的预定有效OP值。在一些实现方式中，阈值等于或低于WA vs.OP的“拐点”值，如图1所示，例如约15％。

[0067] 如果不存在有效OP级别比阈值低的SSD，则控制器不执行任何对SSD的平衡，并且控制器等待路径607上的下一个平衡协议触发。如果控制器确定存在具有低于阈值的有效OP级别的SSD，所有这样的被识别的SSD都经过平衡。

[0068] 例如，参考图5中的图，控制器将确定第一有效OP 520、第二有效OP521、第三有效OP 522和第四有效OP 524，并且将确定第一SSD 510具有低于阈值的有效OP。

[0069] 在图6的步骤608中，控制器确定用于对具有低于阈值的有效OP级别的SSD的OP级别进行平衡的顺序。在一些实现方式中，控制器按有效OP对SSD进行排序，并且确定具有最小有效OP值的SSD将首先被平衡。在一些实现方式中，控制器按有效OP对SSD进行排序，并且基于另一参数(诸如SSD上的数据的访问速率)来确定平衡的顺序。在一些实现方式中，控制

器通过有效OP对SSD进行排序，并且选择具有最小有效OP的SSD进行平衡，并且忽略剩余SSD的顺序。在一些实现方式中，控制器对SSD进行排序并选择具有最小有效OP的SSD，但是还记录在最小有效OP列表的平衡之后未选择用于平衡的剩余SSD的顺序。

[0070] 例如，再次参考图5，控制器确定第三SSD 512可以接受第一SSD 510的一部分数据，同时在迁移之后将有效OP保持在阈值内。

[0071] 在图6的步骤610，控制器确定是否存在可以适应附加利用率而同时保持在有效OP阈值内的SSD。控制器确定哪个SSD可以具有额外的利用率，并选择所确定的SSD中的一个作为平衡协议的目的地SSD。控制器可以在迁移之后预先计算每个可能的目的地SSD的有效OP值，以便确定目的地SSD。预先计算有效OP可能是理想的，因为数据范围的迁移增加了SSD的利用率。控制器可以基于各种参数来选择目的地SSD，例如，诸如选择具有最高有效OP的SSD，选择可以适应附加利用率的第一已识别SSD，从包括最高有效OP的范围内的具有有效OP的一组SSD中随机选择SSD，或者从包括最高有效OP的范围内的具有有效OP的一组SSD中选择一个SSD作为最近最少访问的SSD，等等。

[0072] 如果没有可以在保持有效OP高于阈值的同时适应附加利用率的SSD，则控制器沿着路径611移动到平衡协议的第二阶段，以开始图7中描述的过程。

[0073] 图7示出了决策流程700，用于不存在能够适应附加利用率并且保持超过阈值的有效OP的单个目的地SSD的情况。在控制器已经从来自图6的路径611上启动协议的第二阶段之后，在步骤702开始决策流程700。在步骤704，控制器确定是否存在具有超过阈值的有效OP的SSD。如果没有具有超过阈值的有效OP的SSD，则控制器在步骤706结束程序并返回到等待状态。如果存在具有超过阈值的有效OP的SSD，则控制器进入步骤708。在步骤708，控制器按有效OP对SSD进行排序，并选择具有最大OP的目标SSD。在步骤710，控制器确定可以迁移到目标SSD同时仍然使目标SSD的有效OP高于阈值的数据的大小或量。在步骤712，控制器将所确定的量迁移到目标SSD。

[0074] 在步骤714，控制器确定源SSD的有效OP是否保持低于阈值。如果源SSD的有效OP现在超过阈值，则控制器在步骤716结束程序。如果源SSD的有效OP保持低于阈值，则控制器沿路径718返回以重复该程序，以在步骤704中确定是否存在具有大于阈值的有效OP的SSD作为开始。控制器可以根据流程700循环访问(iterate through)池中的SSD，以在池中的SSD之间分配剩余的附加利用率，使得目的地SSD和源SSD以具有超过阈值的有效OP而结束该过程。

[0075] 再次参考图6，可替代地，如果控制器在图6中的步骤610确定，没有可以适应附加利用率的单个目的地SSD，控制器仍然可以选择具有比源SSD更高的有效OP的目的地SSD，并且向其迁移被计算为使源SSD和目的地SSD的有效OP达到相等值的数据量。在替代实施例中，如果没有可以在保持有效OP高于阈值的同时适应附加利用率的SSD，则控制器结束平衡协议并返回到等待状态。

[0076] 如果在图6的步骤610中，控制器确定存在可以适应附加利用率的SSD，则控制器前进到图6中的步骤612。在选择源SSD和目的地SSD后，控制器将确定哪些数据要从源SSD迁移到目的地SSD。在步骤612，控制器确定源SSD中的将使得有效OP在阈值内的数据范围。控制器在源SSD上确定一个或多个数据范围，所述数据范围具有高利用率并且被设定大小使得从源SSD迁移数据范围将使有效OP在阈值内。如果识别出多于一个的适当的数据范围，控制

器将根据API中指定的参数确定哪个数据范围将被迁移。

[0077] 控制器可以通过使用映射表确定各种数据范围的利用率来确定源SSD上的适当数据范围。例如，控制器访问映射表，并且可以根据映射表中的LBA确定有多少被映射到每个SSD。

[0078] 可以选择选择启发式(selection heruistic)，使得控制器将确定具有最小容量和最高利用率的数据范围或段，以便具有最小的迁移惩罚和最高的影响。然而，在一些实现方式中，其他选择启发式可以是优选的。可以选择满足源SSD和目的地SSD两者的有效OP的阈值要求的任何数据范围。在一些实现方式中，控制器可以基于源SSD上的数据范围的或目的地SSD上的可用空间的访问速率，来确定将迁移哪个数据范围。

[0079] 在一些实现方式中，控制器可以将单个源数据范围分割成多个较小的数据范围以放置在多个目的地SSD上。在一些实现方式中，控制器可以这样做，以便在整个SSD池范围内更均匀地平衡OP。在一些实现方式中，如果没有单个目的地SSD可以接受全部源数据范围同时将有效OP保持在所需阈值内，则控制器可以这样做。

[0080] 在步骤614，控制器将识别的数据范围从源SSD迁移到目的地SSD。数据范围从源SSD到目的地SSD的迁移应当对SSD进行平衡，以使源SSD和目的地SSD都具有满足阈值的有效OP。在迁移数据之前，控制器可以确定存在可以放置数据的目的地SSD的可接受部分。然后，控制器可以更新映射表以反映迁移的数据。

[0081] 在将数据范围从源SSD迁移到目的地SSD之后，控制器可以循环通过需要被平衡的任何剩余的已识别的SSD，执行相同的步骤。继续迁移数据范围以平衡在整个SSD范围内的有效OP，直到池中的所有SSD具有高于阈值的有效OP或直到数据的迁移不再可行。

[0082] 可替代地，在步骤604，控制器可以一次确定一个SSD的有效OP，使得一旦SSD被识别为具有低于阈值的有效OP则控制器就开始平衡协议，而不是在步骤604确定所有SSD的有效OP以及绕过步骤608。

[0083] 作为另一替代方案，控制器可以在步骤610最佳地选择多个目的地和到每个目的地的数据团块的大小，使得迁移之后的所有SSD的有效OP具有最小标准差，同时保持有效OP小于每个SSD的阈值。

[0084] 图8示出了用于平衡在整个SSD池范围内的利用率参数的方法800。方法800示出了控制器(诸如图2中的控制器222)识别用于平衡协议的适当目的地驱动器的操作流程。

[0085] 在步骤802中，计算第一SSD的第一利用率参数和第二SSD的第二利用率参数。第一SSD的第一利用率参数低于第一阈值。第一利用率参数和第二利用率参数可以是有效预留空间的量度，或与SSD的利用率相关的任何其它合适的参数。

[0086] 在步骤804中，将第二利用率参数与第一阈值进行比较。如果第二利用率参数小于第一阈值，则该方法进行到步骤806，其中识别第三SSD，并且为第三SSD计算第三利用率参数。然后在步骤804将第三利用率参数与第一阈值进行比较，并且该方法继续。

[0087] 如果在步骤804第二利用率参数大于第一阈值，则该方法进行到步骤808。在步骤808，识别在第一SSD上的数据的待被去除的数据范围。该数据范围被识别以使得从第一SSD去除该数据范围内的数据将使与第一SSD相关联的利用率参数在第一阈值内。在步骤810，针对包括来自所识别的第一SSD的数据范围的数据的第二SSD计算第四利用率参数。在步骤812，针对不包括第一SSD的所识别数据范围内的数据的第一SSD计算第五利用率参数。步骤

810和812使得控制器能够在将数据从第一SSD迁移到第二SSD之后计算第一SSD和第二SSD的利用率。通过预先计算两个SSD的利用率参数，控制器可以确定所选择的目的地SSD是否为可接受的，以及是否可以对数据的迁移以及可用SSD之间的利用率平衡进行优化。

[0088] 在步骤814，计算跨多个可用SSD的多个利用率参数(包括第四利用率参数和第五利用率参数)的标准差。标准差的计算使得控制器能够确定数据从第一SSD迁移到第二SSD之后跨多个可用SSD的利用率平衡。在步骤816，将标准差与第二阈值进行比较。如果标准差大于第二阈值，则该方法返回到步骤806，并且确定新的目的地SSD。如果标准差小于第二阈值，则该方法进行到步骤818，并将与来自第一SSD的数据范围对应的数据迁移到第二SSD。

[0089] 在一些实现方式中，如果在步骤816标准差大于第二阈值，则不是将该方法返回到步骤806以确定新的目的地SSD，而是可以计算包括来自所识别的数据范围的数据的一部分的第二SSD的利用率参数，并且选择第二目的地SSD作为来自所识别的数据范围的数据的剩余部分的目的地。然后可以重新计算标准差，以确定当该数据范围内的数据的一部分在两个目的地SSD之间分布时，利用率是否平衡。在一些实现方式中，该方法可以针对与第一SSD的所识别的数据范围对应的数据来确定多于两个的目的地SSD。

[0090] 计算SSD池中多个SSD的利用率参数的标准差使得控制器能够优化在整个SSD池范围内的利用率。在整个SSD池范围内的利用率的最佳平衡增加了设备的使用寿命，并最大限度地降低了SSD或设备出故障的风险。

[0091] 本发明的各个方面的其它目的、优点和实施例对于本领域技术人员而言将是显而易见的且在本说明书和附图的范围内。例如但不限于，结构或功能元件可以与本发明一致地重新排列。类似地，根据本发明的原理可以应用于其它示例，尽管这里不详细描述，但是这些示例仍然在本发明的范围内。

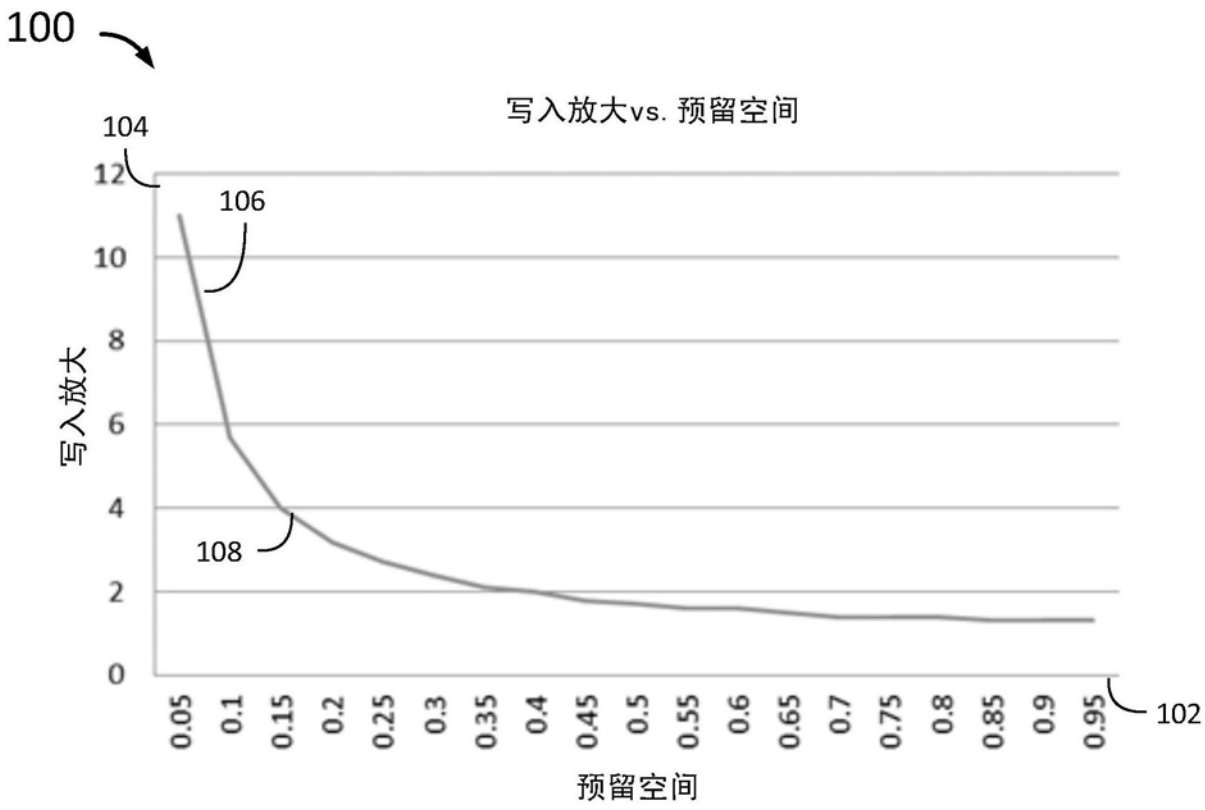

100
写入放大vs. 预留空间
104
106
108
102
写入放大
预留空间
0
2
4
6
8
10
12
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95

图1

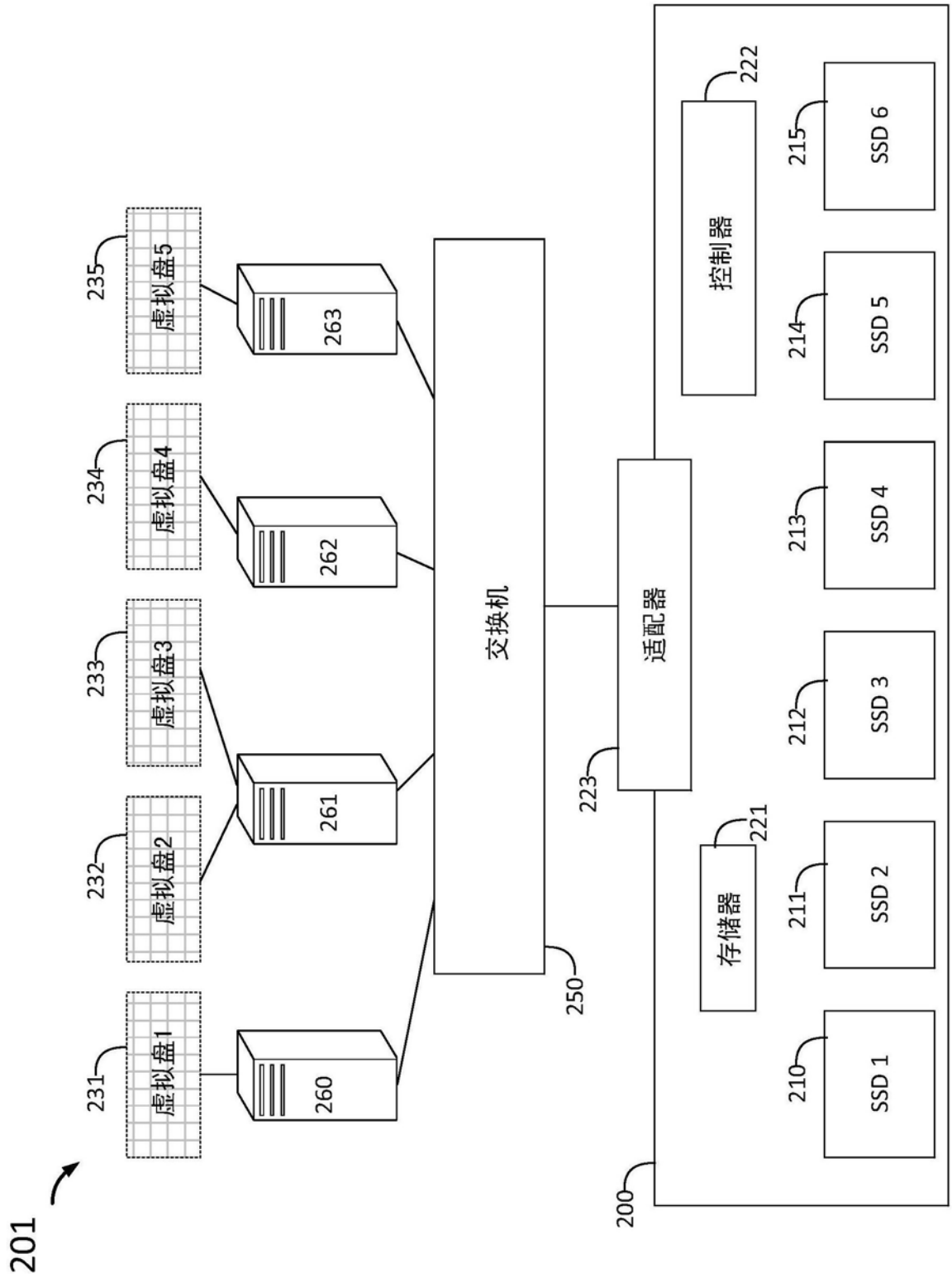

201
200
231
虚拟盘1
232
虚拟盘2
233
虚拟盘3
234
虚拟盘4
235
虚拟盘5
260
261
262
263
250
交换机
223
适配器
210
SSD 1
211
SSD 2
212
SSD 3
213
SSD 4
214
SSD 5
215
SSD 6
221
存储器
222
控制器

图2

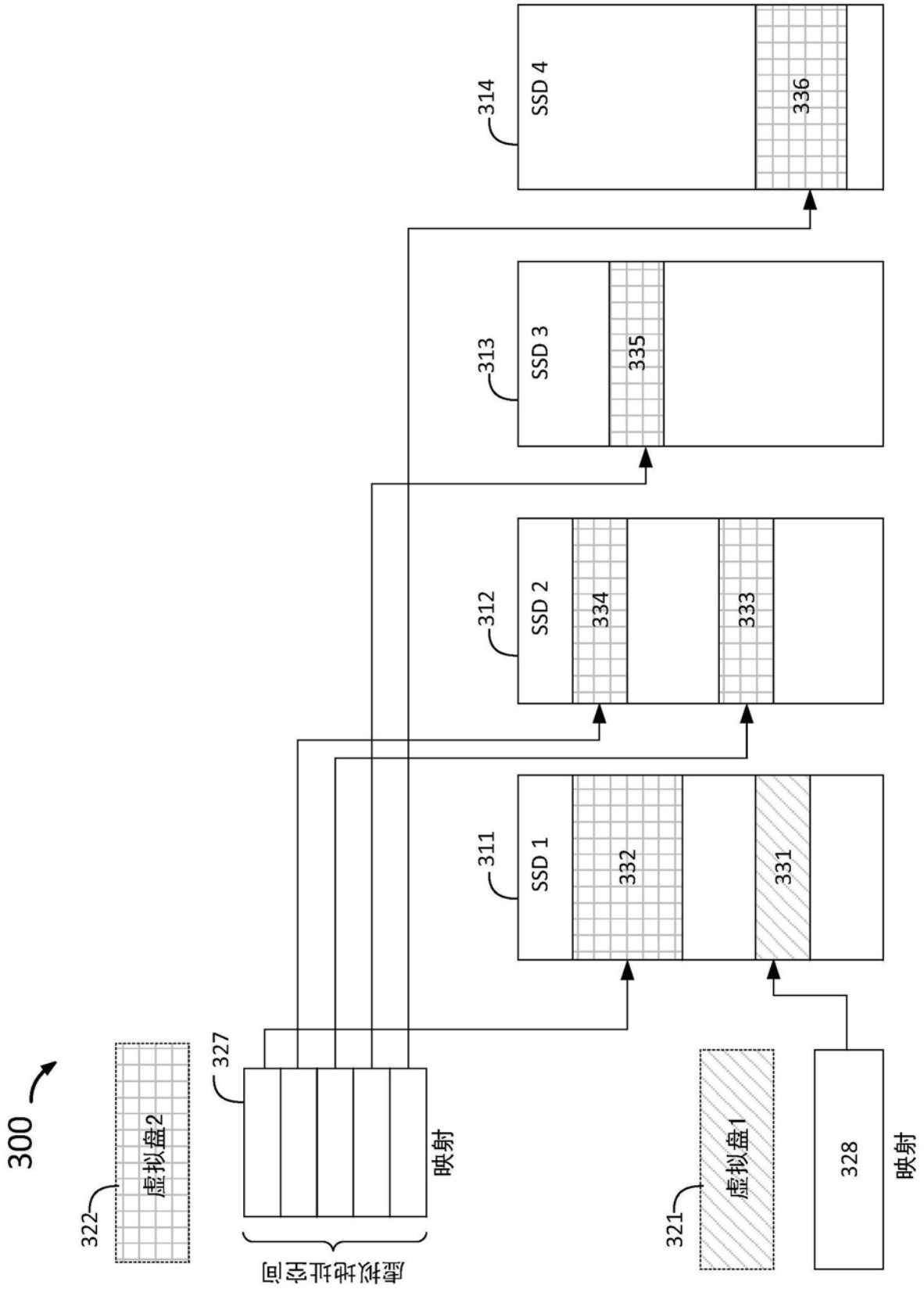

300
虚拟盘2
322
327
虚拟地址空间
映射
虚拟盘1
321
328
映射
311
SSD 1
332
331
312
SSD 2
334
333
313
SSD 3
335
314
SSD 4
336

图3

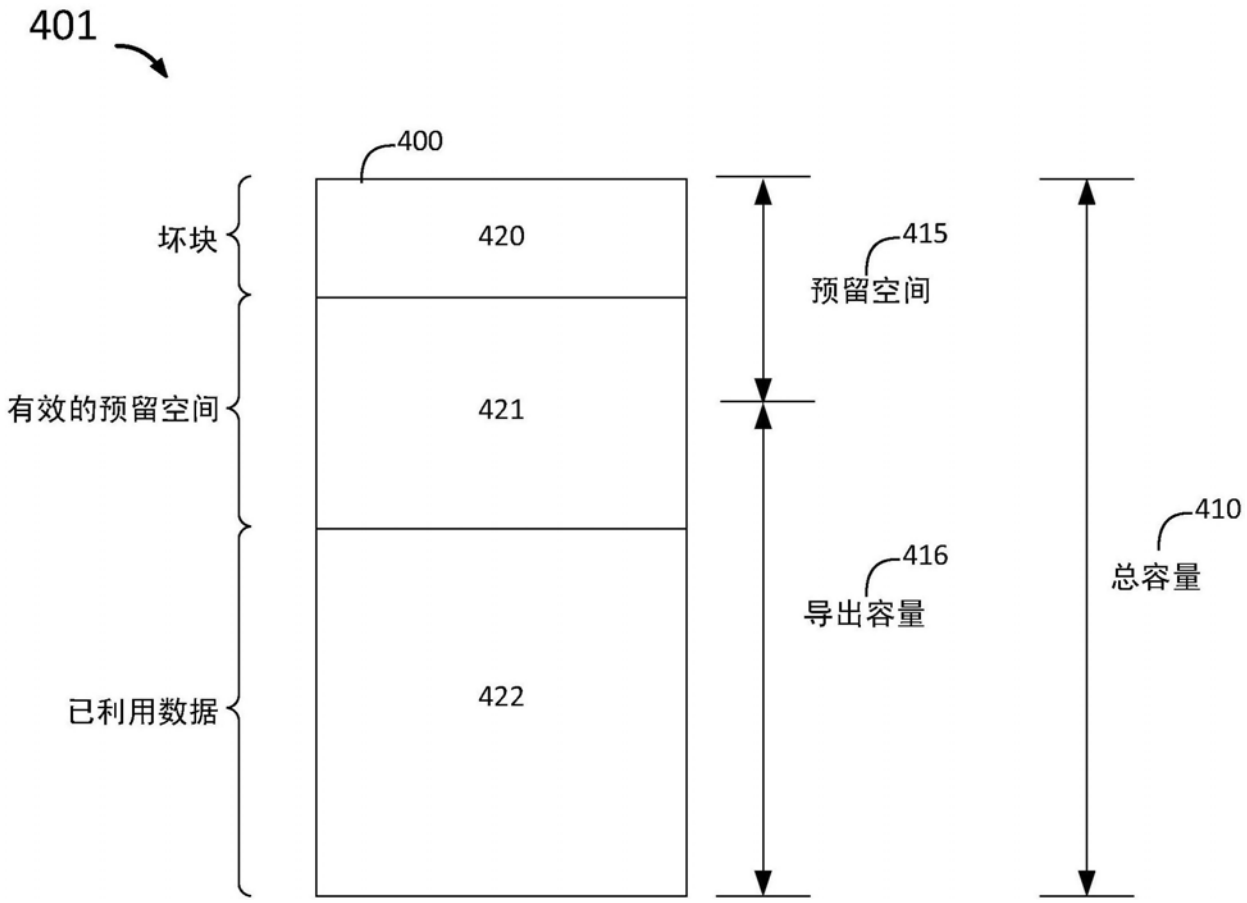

401
400
坏块
420
415
预留空间
有效的预留空间
421
410
总容量
416
导出容量
已利用数据
422

图4A

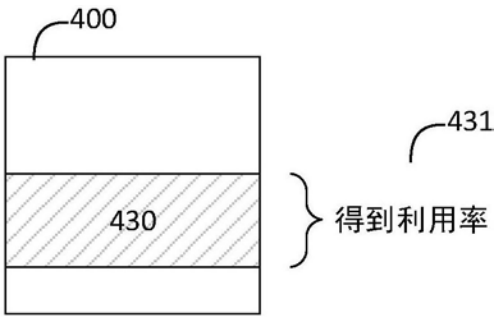

400
431
430
得到利用率

图4B

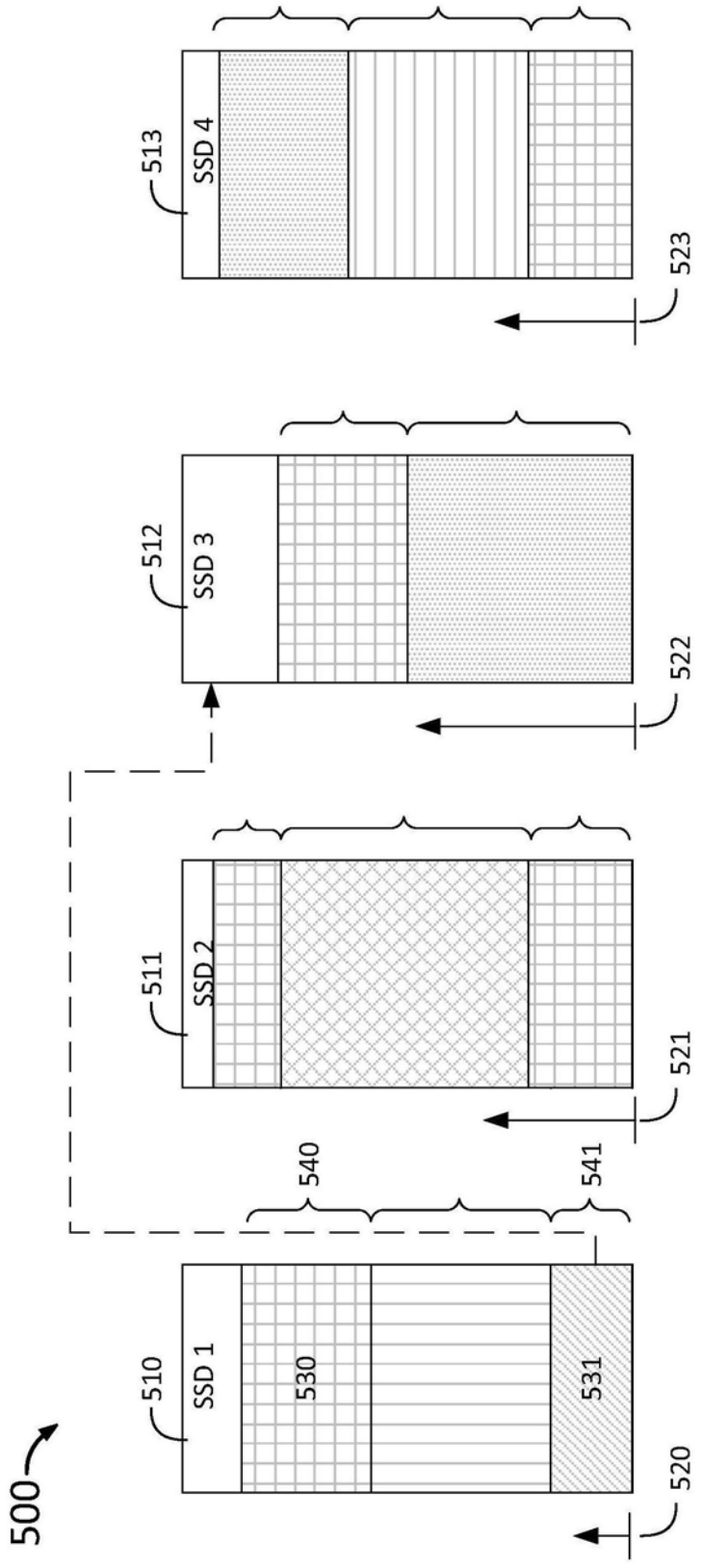

500
510
SSD 1
530
531
520
511
SSD 2
521
512
SSD 3
522
513
SSD 4
523
540
541

图5

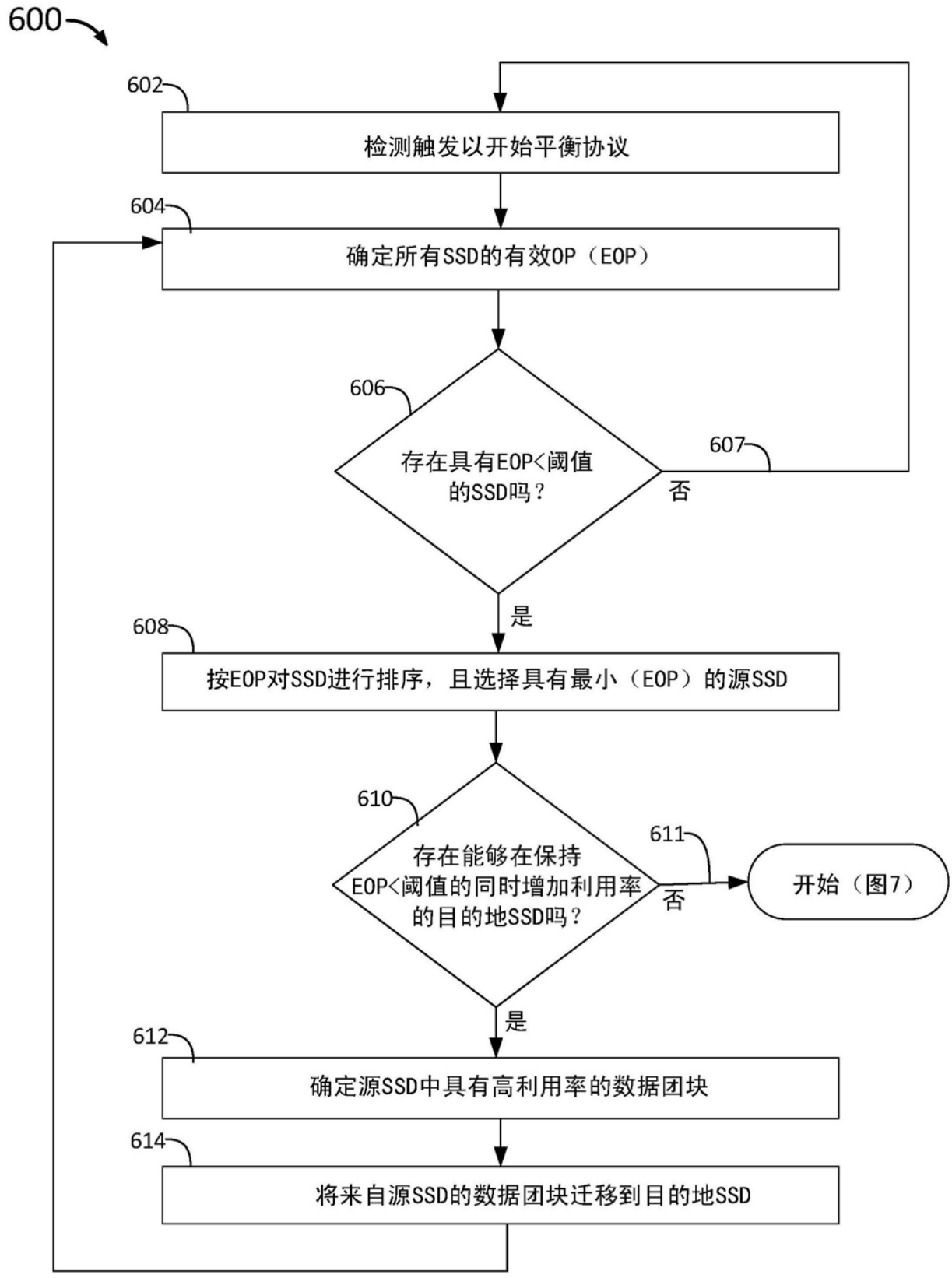

600
602
检测触发以开始平衡协议
604
确定所有SSD的有效OP（EOP）
606
存在具有EOP<阈值的SSD吗？
607
否
是
608
按EOP对SSD进行排序，且选择具有最小（EOP）的源SSD
610
存在能够在保持EOP<阈值的同时增加利用率的目的地SSD吗？
611
否
开始（图7）
是
612
确定源SSD中具有高利用率的数据团块
614
将来自源SSD的数据团块迁移到目的地SSD

图6

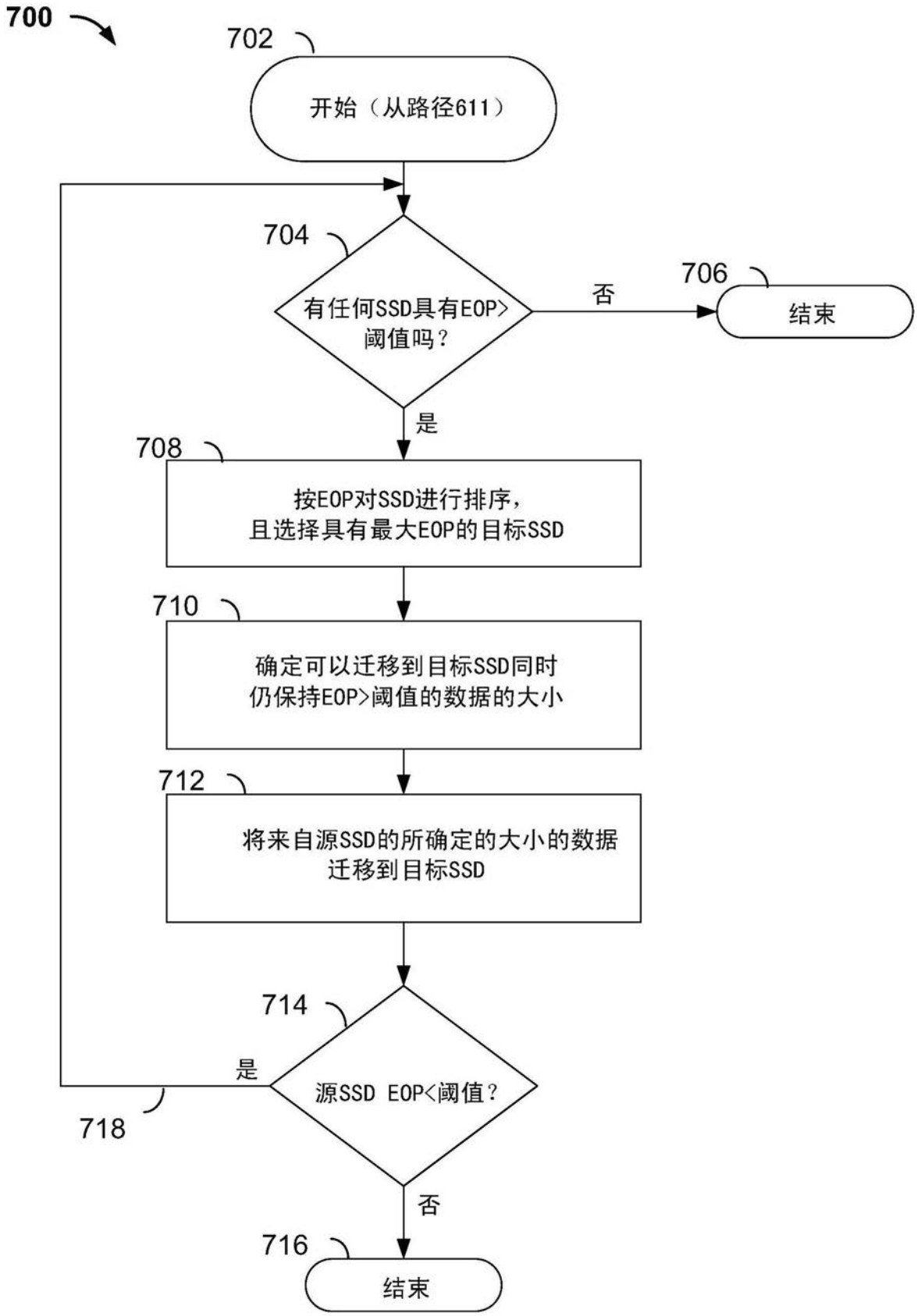

700
702
开始（从路径611）
704
有任何SSD具有EOP>阈值吗？
否
706
结束
是
708
按EOP对SSD进行排序，且选择具有最大EOP的目标SSD
710
确定可以迁移到目标SSD同时仍保持EOP>阈值的数据的大小
712
将来自源SSD的所确定的大小的数据迁移到目标SSD
714
源SSD EOP<阈值？
是
718
否
716
结束

图7

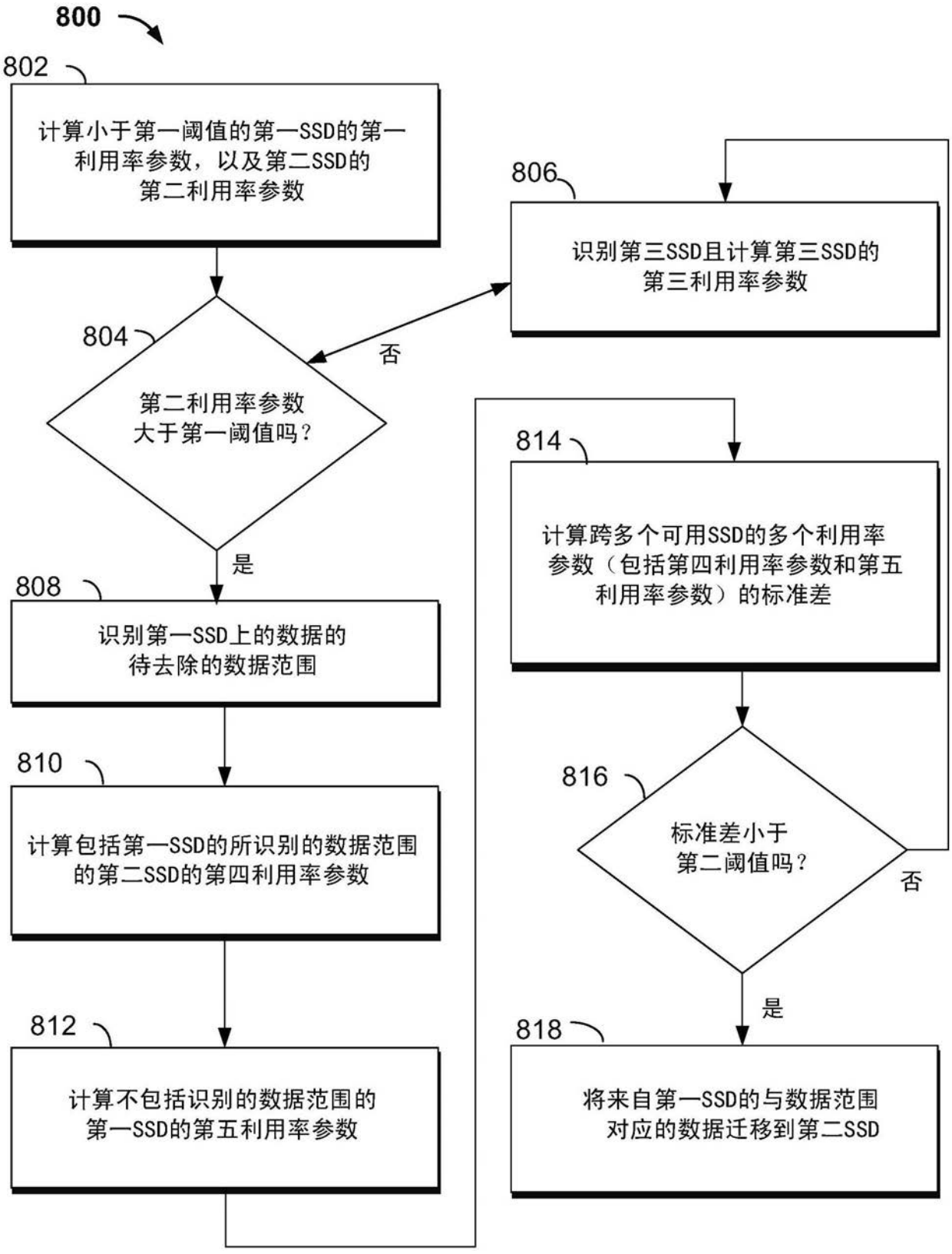

800
802
计算小于第一阈值的第一SSD的第一利用率参数，以及第二SSD的第二利用率参数
804
第二利用率参数大于第一阈值吗？
否
是
806
识别第三SSD且计算第三SSD的第三利用率参数
808
识别第一SSD上的数据的待去除的数据范围
810
计算包括第一SSD的所识别的数据范围的第二SSD的第四利用率参数
812
计算不包括识别的数据范围的第一SSD的第五利用率参数
814
计算跨多个可用SSD的多个利用率参数（包括第四利用率参数和第五利用率参数）的标准差
816
标准差小于第二阈值吗？
否
是
818
将来自第一SSD的与数据范围对应的数据迁移到第二SSD

图8