(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113924546 B
(45)授权公告日 2024.12.13

(21)申请号 202080040193.7
(22)申请日 2020.05.28
(65)同一申请的已公布的文献号
申请公布号 CN 113924546 A
(43)申请公布日 2022.01.11
(30)优先权数据
16/439,122 2019.06.12 US
(85)PCT国际申请进入国家阶段日
2021.11.30
(86)PCT国际申请的申请数据
PCT/IB2020/055061 2020.05.28
(87)PCT国际申请的公布数据
WO2020/250071 EN 2020.12.17
(73)专利权人 国际商业机器公司
地址 美国纽约
(72)发明人 S·托米克 R·A·普莱卡
N·约安努 N·帕潘德尔欧
A·D·弗赖伊 T·费希尔
R·I·斯托伊卡
(74)专利代理机构 中国贸促会专利商标事务所有限公司 11038
专利代理师 程晨
(51)Int.Cl.
G06F 3/06(2006.01)
(56)对比文件
US 2011271043 A1,2011.11.03
US 2012240012 A1,2012.09.20
US 2013124787 A1,2013.05.16
审查员 郭海洋

权利要求书3页 说明书11页 附图7页

(54)发明名称

非易失性存储器中的磨损感知块模式转换

(57)摘要

一种非易失性存储器包括多个物理存储块，每个物理存储块包括相应的多个单元，其中，多个单元中的每个单元能够单独地存储多位数据。控制器将多个物理块之中的物理块分配到含有在第一(例如,QLC)模式下操作的物理块的第一池以用于存储每单元较大数量的位，并且将多个物理块之中的其他物理块分配到含有在第二(例如,SLC)模式下操作的物理块的第二池以用于存储每单元较小数量的位。控制器至少基于为所转移的物理块测量的误码率在第一池和第二池之间转移物理块。



CN 113924546 B

1.一种操作包括各自包括相应的多个单元的用于存储的多个物理块的非易失性存储器的方法,其中,所述多个单元中的每个单元能够单独地存储多位数据,所述方法包括:

控制器将所述多个物理块之中的物理块分配到第一池,第一池包含在第一模式下操作的用于存储每单元较大数量的位的物理块,并且将所述多个物理块之中的其他物理块分配到第二池,第二池包含在第二模式下操作的用于存储每单元较小数量的位的物理块;

对于第一池和第二池中的每一个:

跟踪被分配来存储用于不同写入热度的多个写入流中的每个写入流的数据的物理块的数量和所述多个写入流中的每个写入流中的编程/擦除(P/E)周期的数量;和

对于所述多个写入流中的每个写入流,基于分配给该写入流的物理块的数量和分配给该写入流的物理块的P/E周期的总数来计算相应的流磨损度量;和

控制器在第一池和第二池之间转移物理块,其中所述转移包括基于以下各项来选择用于转移的物理块:(1)所述多个写入流中的选择的写入流,(2)所选择的写入流的流磨损度量,和(3)被分配来存储所选择的写入流的数据的物理块的块健康。

2.根据权利要求1所述的方法,其中,所述转移包括:

将具有较高误码率的物理块从第一池转移到第二池;和

将具有较低误码率的物理块从第二池转移到第一池。

3.根据权利要求1所述的方法,并且进一步包括控制器,其根据在对物理块进行编程之后不久测量的误码率来确定所述多个写入流的所有写入流中的物理块的块健康。

4.根据权利要求1所述的方法,其中,所述选择包括:

基于所选择的写入流具有反映所述多个写入流的所有写入流中的最大磨损的流磨损度量,从分配给所选择的写入流的物理块当中选择最健康的物理块用于转移;和

基于所选择的写入流具有反映所述多个写入流的所有写入流中的最小磨损的流磨损度量,从分配给所选择的写入流的物理块当中选择最不健康的物理块用于转移。

5.根据权利要求1所述的方法,其中:

第一模式是四级单元(QLC)模式;并且

第二模式是单级单元(SLC)模式。

6.根据权利要求1所述的方法,其中,所述转移包括基于第一池和第二池当中的给定池中的未编程的物理块的可用数量不满足阈值来转移物理块。

7.一种数据存储系统,包括:

用于非易失性存储器的控制器,所述非易失性存储器包括各自包括相应的多个单元的用于存储的多个物理块,其中,所述多个单元中的每个单元能够单独地存储多位数据,其中,控制器被配置为执行:

将所述多个物理块之中的物理块分配到第一池,第一池包含在第一模式下操作的用于存储每单元较大数量的位的物理块,并且将所述多个物理块之中的其他物理块分配到第二池,第二池包含在第二模式下操作的用于存储每单元较小数量的位的物理块;

对于第一和第二池中的每一个:

跟踪被分配来存储用于不同写入热度的多个写入流中的每个写入流的数据的物理块的数量和所述多个写入流中的每个写入流中的编程/擦除(P/E)周期的数量;和

对于所述多个写入流中的每个写入流,基于分配给该写入流的物理块的数量和分配给

该写入流的物理块的P/E周期的总数来计算相应的流磨损度量;和

在第一池和第二池之间转移物理块,其中所述转移包括基于以下各项来选择用于转移的物理块:(1)所述多个写入流中的选择的写入流,(2)所选择的写入流的流磨损度量,和(3)被分配来存储所选择的写入流的数据的物理块的块健康。

8.根据权利要求7所述的数据存储系统,其中,所述转移包括:

将具有较高误码率的物理块从第一池转移到第二池;和

将具有较低误码率的物理块从第二池转移到第一池。

9.根据权利要求7所述的数据存储系统,其中,控制器进一步被配置为执行:

根据在对物理块进行编程之后不久测量的误码率来确定所述多个写入流的所有写入流中的物理块的块健康。

10.根据权利要求7所述的数据存储系统,其中,所述选择包括:

基于所选择的写入流具有反映所述多个写入流的所有写入流中的最大磨损的流磨损度量,从分配给所选择的写入流的物理块当中选择最健康的物理块用于转移;和

基于所选择的写入流具有反映所述多个写入流的所有写入流中的最小磨损的流磨损度量,从分配给所选择的写入流的物理块当中选择最不健康的物理块用于转移。

11.根据权利要求7所述的数据存储系统,其中:

第一模式是四级单元(QLC)模式;并且

第二模式是单级单元(SLC)模式。

12.根据权利要求7所述的数据存储系统,其中,所述转移包括基于第一池和第二池当中的给定池中的未编程的物理块的可用数量不满足阈值来转移物理块。

13.一种计算机程序产品,所述计算机程序产品包括具有体现在其中的程序指令的计算机可读存储介质,所述程序指令可由包括各自包括相应的多个单元的用于存储的多个物理块的非易失性存储器的控制器执行,其中,所述程序指令当被执行时,使控制器执行:

将所述多个物理块之中的物理块分配到第一池,第一池包含在第一模式下操作的用于存储每单元较大数量的位的物理块,并且将所述多个物理块之中的其他物理块分配到第二池,第二池包含在第二模式下操作的用于存储每单元较小数量的位的物理块;

对于第一和第二池中的每一个:

跟踪被分配来存储用于不同写入热度的多个写入流中的每个写入流的数据的物理块的数量和所述多个写入流中的每个写入流中的编程/擦除(P/E)周期的数量;和

对于所述多个写入流中的每个写入流,基于分配给该写入流的物理块的数量和分配给该写入流的物理块的P/E周期的总数来计算相应的流磨损度量;和

在第一池和第二池之间转移物理块,其中所述转移包括基于以下各项来选择用于转移的物理块:(1)所述多个写入流中的选择的写入流,(2)所选择的写入流的流磨损度量,和(3)被分配来存储所选择的写入流的数据的物理块的块健康。

14.根据权利要求13所述的程序产品,其中,所述转移包括:

将具有较高误码率的物理块从第一池转移到第二池;和

将具有较低误码率的物理块从第二池转移到第一池。

15.根据权利要求13所述的程序产品,其中,所述程序指令进一步使控制器执行:

根据在对物理块进行编程之后不久测量的误码率来确定所述多个写入流的所有写入

流中的物理块的块健康。

16.根据权利要求13所述的程序产品,其中,所述选择包括:

基于所选择的写入流具有反映所述多个写入流的所有写入流中的最大磨损的流磨损度量,从分配给所选择的写入流的物理块当中选择最健康的物理块用于转移;和

基于所选择的写入流具有反映所述多个写入流的所有写入流中的最小磨损的流磨损度量,从分配给所选择的写入流的物理块当中选择最不健康的物理块用于转移。

17.根据权利要求13所述的程序产品,其中,所述转移包括基于第一池和第二池当中的给定池中的未编程的物理块的可用数量不满足阈值来转移物理块。

非易失性存储器中的磨损感知块模式转换

技术领域

[0001] 本发明涉及数据处理和数据存储,具体涉及其中块能够在提供不同的每单元位存储容量的多个块模式下操作的非易失性存储器系统。更具体地,本发明涉及非易失性存储器系统中的物理存储的块在不同块模式之间的磨损感知转换。

背景技术

[0002] NAND闪存是电可编程且可擦除的非易失性存储器技术,该技术将每个存储器单元的一个或多个数据位存储为晶体管或类似电荷陷阱结构的浮置栅极上的电荷。浮置栅极上的电荷量调制晶体管的阈值电压。通过施加适当的读取电压并且测量电流量,可以确定存储器单元的编程阈值电压,从而可以检测所存储的信息。在典型的实现方式中,NAND闪存阵列是按物理存储器的块(亦称“擦除块”)组织的,每个块包括多个物理页,每个物理页依次包含多个存储器单元。借助于用于访问存储器单元的字和位线的安排,闪存阵列通常以页为基础进行编程,但以块为基础进行擦除。

[0003] 在过去的十年中,闪存存储技术的位存储密度有了惊人的增长,这是由降低闪存的每千兆字节的价格的目标驱动的。该趋势的关键驱动器之一是每个闪存单元所存储的位数的稳定增加。这一趋势的关键驱动因素之一是每个闪存单元存储的位数稳步增加。例如,每个闪存单元可以存储的位数已经从一个(单级单元(SLC))稳定地增加到两个(多级单元(MLC))到三个(三级单元(TLC)),最近增加到四个(四级单元(QLC))。每个闪存单元可以存储的位数的增加是以耐久性和性能随着每个单元的位数增加而降低为代价的。

[0004] 企业级闪存控制器采用各种技术来改善新一代NAND闪存的耐久性和性能限制。例如,闪存控制器可以采用一个或多个闪存管理技术,包括实施高级纠错码,调整电压电平以降低误码率,通过基于写入热度隔离数据来减少内部垃圾收集开销,基于块来管理块(即,健康分级),以及执行健康感知磨损均衡和数据放置。尽管这些高级闪存管理技术延长了使用寿命并提高了性能,但随着位密度的持续增长,仅靠现有的闪存管理技术无法保证可接受的设备寿命(例如,在繁重的写工作负载下,企业级SSD的7年或更长的保修期)。例如,与TLC NAND闪存技术相比,QLC NAND闪存技术的存储密度提高了33%,但其额定使用寿命比TLC NAND闪存短几倍(例如,每块仅数千编程/擦除(P/E)周期)。

[0005] 提高器件耐久性的一种有希望的途径是利用NAND闪存以多种模式操作的能力。例如,一些TLC NAND闪存支持SLC模式,其中某些选定块每个单元仅存储单个位,因此以牺牲存储密度为代价来提高选定块的性能和耐久性。在SLC模式下配置的块可以维持多一至两个数量级的P/E周期,但是仅提供在TLC模式下配置的块的存储容量的三分之一。在这样的系统中,频繁的写数据可以存储在SLC块中,而不频繁的写数据可以存储在在原生TLC模式下操作的块上。

[0006] 支持多个操作模式的现有闪存控制器通常采用分层架构,其中,进入的用户写首先被存储在较低存储密度层(例如,SLC层)的块中。当较低存储密度层中的存储用尽(或达到某个其他阈值填充水平)时,闪存控制器发起转储过程,通过将数据从较低存储密度层转

移到较高存储密度层(例如,TLC层)来释放较低存储密度层中的空间—优选地从未来最可能被覆写的数据开始。闪存控制器然后以常规方式管理驻留在较高存储密度层中的数据,例如通过定期地重新定位被选择用于垃圾收集的块中的有效数据,擦除垃圾收集的块,然后再次用从较低存储密度层转储的数据来写入被擦除的块。通常,在这种类型的常规分层存储架构中,块被静态地分配给较低密度存储层或较高存储密度层。

发明内容

[0007] 本发明认识到,块到采用不同位存储密度的块的多层存储器存储系统的各个层的完全静态分配,可能导致给定层过早地耗尽存储容量或达到其耐久性极限。为了解决这个问题,可以在操作模式之间选择性地转换非易失性存储器的块。

[0008] 在本发明的一个方面,提供了操作包括各自包括相应的多个单元的多个物理存储块的非易失性存储器的方法,其中,所述多个单元中的每个单元能够单独地存储多位数据,所述方法包括:控制器将所述多个物理块之中的物理块分配到第一池,第一池包含在第一模式下(例如,QLC模式)操作的用于存储每单元较大数量的位的物理块,并且将所述多个物理块之中的其他物理块分配到第二池,第二池包含在第二模式下(例如,SLC模式)操作的用于存储每单元较小数量的位的物理块。

[0009] 在至少一个实施例中,控制器将具有较高误码率的物理块从第一池转移到第二池,并将具有较低误码率的物理块从第二池转移到第一池。

[0010] 在至少一个实施例中,对于第一和第二池中的每一个,控制器跟踪被分配来存储用于不同写入热度的多个写入流中的每个写入流的数据的物理块的数量和所述多个写入流中的每个写入流中的编程/擦除(P/E)周期的数量。另外,对于所述多个写入流中的每个写入流,控制器基于分配给该写入流的物理块的数量和分配给该写入流的物理块的P/E周期的总数来计算相应的流磨损度量。控制器然后可基于以下各项来选择用于转移的物理块:(1)所述多个写入流中的选择的写入流,(2)所选择的写入流的流磨损度量,和(3)被分配来存储所选择的写入流的数据的物理块的块健康。例如,可以根据在对物理块进行编程之后不久测量的误码率来确定所述多个写入流的所有写入流中的物理块的块健康。

[0011] 在至少一个实施例中,控制器这样来选择用于转移的物理块,即,如果所选择的写入流具有反映所述多个写入流的所有写入流中的最大磨损的流磨损度量,控制器从分配给所选择的写入流的物理块当中选择最健康的物理块用于转移,并且,如果所选择的写入流具有反映所述多个写入流的所有写入流中的最小磨损的流磨损度量,控制器从分配给所选择的写入流的物理块当中选择最不健康的物理块用于转移。

[0012] 在至少一个实施例中,控制器基于第一池和第二池当中的给定池中的未编程物理块的可用数量不满足阈值,在第一池和第二池之间转移物理块。

附图说明

[0013] 图1A是根据一个实施例的数据处理环境的高级框图;

[0014] 图1B是图1A的数据存储系统的闪存卡的更详细的框图;

[0015] 图2示出了根据一个实施例的示范性NAND闪存模块;

[0016] 图3是根据一个实施例的在闪存管理中采用的闪存管理功能和数据结构的高级流

程图;

[0017] 图4示出了示范性分层存储架构,其中物理块池被配置成在不同模式下操作,从而每单元存储不同数量的位;

[0018] 图5是示范性分层存储架构的写入放大率与设备利用率的关系图;

[0019] 图6是根据一个实施例的用于配置非易失性存储器的示范性过程的高级逻辑流程图;以及

[0020] 图7是根据一个实施例的用于在块池之间动态传送块的示范性过程的高级逻辑流程图。

具体实施方式

[0021] 参考附图,特别是图1A,示出了示范性数据处理环境100的高级框图,该示范性数据处理环境100包括具有如本文中进一步描述的非易失性存储器阵列的数据存储系统120。如图所示,数据处理环境100包括一个或多个主机,诸如具有一个或多个处理指令和数据的处理器104的处理器系统102。处理器系统102可以另外包括本地存储106(例如,动态随机存取存储器(DRAM)或磁盘),其可以存储由处理器104执行的处理的程序代码、操作数和/或执行结果。在各种实施例中,处理器系统102可以是例如移动计算设备(诸如智能电话或平板计算机)、膝上型或台式个人计算机系统、服务器计算机系统(诸如可从国际商业机器公司获得的POWER系列服务器之一)或大型计算机系统。处理器系统102还可以是使用不同处理器(诸如ARM、POWER、英特尔x86)的嵌入式处理器系统,或与存储器高速缓存、存储器控制器、本地存储、I/O总线集线器等中的任何组合的任何其他处理器。

[0022] 每个处理器系统102进一步包括经由I/O通道110直接(即,没有任何中间设备)或间接(即,通过至少一个中间设备)耦合到数据存储系统120的输入/输出(I/O)适配器108。在一些实施例中,数据存储系统120可以集成到处理器系统102中。在各种实施例中,I/O通道110可以采用已知或未来开发的通信协议中的任何一个或其组合,包括例如光纤通道(FC)、以太网FC(FCoE)、因特网小型计算机系统接口(iSCSI)、InfiniBand、转移控制协议/因特网协议(TCP/IP)、外围组件互连高速(PCIe)等。经由I/O通道110通信的I/O请求包括处理器系统102从数据存储系统120请求数据的读取请求和处理器系统102请求将数据存储在数据存储系统120中的写入请求。

[0023] 尽管不是必需的,但是在图示的实施例中,数据存储系统120包括多个接口卡122,数据存储系统120通过这些接口卡122经由I/O通道110接收和响应主机的I/O请求。每个接口卡122耦合至多个廉价磁盘冗余阵列(RAID)控制器124中的每一个,以便促进容错和负载平衡。每个RAID控制器124依次耦合(例如,通过PCIe总线)到在所示示例中包括承载NAND闪存的多个闪存卡126的非易失性存储介质。在其他实施例中,可以采用替代的和/或附加的非易失性存储设备。

[0024] 在所示实施例中,数据存储系统120的操作由耦接到接口卡122和RAID控制器124的冗余系统管理控制器(SMC)123管理。在各种实施例中,系统管理控制器123可以利用硬件或执行固件和/或软件的硬件来实现。

[0025] 图1B示出图1A的数据存储系统120的闪存卡126的示范性实施例的更详细的框图。闪存卡126包括网关130,该网关130充当闪存卡126与RAID控制器124之间的接口。网关130

耦合到通用处理器(GPP)132,其可以被配置(例如,通过程序代码)为对网关130接收的请求执行预处理和/或调度闪速卡126对请求的服务。GPP 132耦合到GPP存储器134(例如,动态随机存取存储器(DRAM)),其可便利地缓冲由GGPP 132在其处理期间创建、引用和/或修改的数据或者去往闪存控制器140中的一个或多个的流经网关130的数据。

[0026] 网关130还耦合到多个闪存控制器140,每个闪存控制器控制相应的NAND闪存系统150。闪存控制器140可以例如由专用集成电路(ASIC)和/或现场可编程门阵列(FPGA)和/或微处理器来实现,并且每个具有相关联的闪存控制器存储器142(例如,DRAM)。在利用FPGA实现闪存控制器140的实施例中,GPP132可在数据存储系统120的启动期间编程和配置闪存控制器140。在启动之后,一般而言,闪存控制器140从网关130接收请求读取存储在NAND闪存系统150中的数据和/或将数据存储在NAND闪存系统150中的读和写请求。闪存控制器140例如通过访问NAND闪存系统150以从NAND闪存系统150读取所请求的数据或将所请求的数据写入NAND闪存系统150或通过访问与NAND闪存系统150相关联的存储器高速缓存(未示出)来服务于这些请求。

[0027] 闪存控制器140实现闪存转换层(FTL),该闪存转换层提供逻辑到物理地址转换以实现对NAND闪存系统150内的特定存储器位置的访问。通常,由闪存控制器140从主机设备(诸如处理器系统102)接收的请求包含要访问(读取或写入)数据的逻辑块地址(LBA),并且,如果是写入请求,则包含要存储到数据存储系统120的写入数据。该请求还可指定要访问的数据的量(或大小)。取决于数据存储系统120所支持的协议和特征,还可传送其他信息。闪存转换层将从RAID控制器124接收的LBA转换为被分配给NAND闪存系统150中的对应物理位置的物理地址。闪存控制器140可以执行地址转换和/或将逻辑地址和物理地址之间的映射存储在逻辑到物理转换数据结构(诸如逻辑到物理转换表(LPT))中,其可以方便地存储在闪存控制器存储器142中。

[0028] 在各种实施例中,NAND闪存系统150可以采取许多形式。在图1B所示的实施例中,每个NAND闪存系统150包括多个(例如,32个)可单独寻址的NAND闪存存储设备152。在所示示例中,闪存存储设备152采取板装闪存模块的形式,例如,多级单元(MLC)、三级单元(TLC)或四级单元(QLC)NAND闪存模块。由闪存存储设备152提供的有效存储容量,可以通过例如由闪存控制器140和/或高级控制器(诸如GPP 132、RAID控制器124或SMC123)实施数据压缩来增加。

[0029] 现在参见图2,描绘了可以用于实现图1B的NAND闪存存储设备152中的任何一个的示范性闪存模块200的框图。闪存模块200包括一个或多个存储器芯片,每个存储器芯片实现由NAND闪存单元的二维或三维阵列形成的至少一个存储器阵列202。如图2所示,存储器阵列202内的存储器单元物理地布置在多个物理块204中,每个物理块又包括多个物理页206。这些页可以按页组管理,页组可以各自例如由耦合到公共字线的所有页、由3D NAND闪存中的一个或多个层中的所有页、由一个或多个层中的一组页、或通常由具有相似特征的页形成。

[0030] 如本领域技术人员已知的,NAND闪存(诸如在存储器阵列202中采用的NAND闪存)必须在被编程之前被擦除。进一步,NAND闪存可能(但不是必须)受其构造约束,使得可被擦除的最小存储粒度是块204,并且可被读取或写入请求访问的最小存储粒度固定在单个物理页206的大小。在这方面应当理解,由主机设备提供的LBA对应于逻辑地址空间内的逻辑

页面,其中每个逻辑页面通常具有4千字节(kB)的大小。相比之下,物理页206通常具有较大的大小,例如16kB,且因此可托管多个逻辑页。

[0031] 闪存模块200还包括行解码器210和列解码器212,存储器阵列202的字线可以通过行解码器210寻址,存储器阵列202的位线可以通过列解码器212寻址。此外,闪存模块200包括使物理页206的存储器单元能够被并行编程或读取的读/写电路214。闪存控制器200还包括控制电路205,其提供对存储器阵列202的操作—包括对存储器阵列202中的物理页206进行的读和写访问、块204的擦除—以及对施加到存储器阵列202的相关电压的幅度、持续时间和极性的芯片级控制。

[0032] 已经描述了数据存储系统120的一个示范性实施例的一般物理结构,现在参见图3描述数据存储系统120的某些操作方面,图3是根据一个实施例由GPP 132和/或闪存控制器140采用的闪存管理功能和数据结构的高级流程图。

[0033] 数据存储系统120一般不允许外部设备(例如,主机)直接寻址和/或访问NAND闪存系统150内的物理存储器位置。相反,数据存储系统120通常被配置为向主机设备呈现一个或多个逻辑卷,每个逻辑卷具有连续的逻辑地址空间,因此允许主机设备向逻辑地址空间内的逻辑块地址(LBA)读取数据和从逻辑地址空间内的逻辑块地址(LBA)写入数据,同时允许不同级别的控制器(例如,系统管理控制器123、RAID控制器124、闪存控制器140和GPP 132)中的一个或多个控制与不同LBA相关联的数据实际驻留在包括NAND闪存系统150的物理存储器位置中的何处。以此方式,可以智能地管理和优化NAND闪存系统150的性能和寿命。

[0034] 在所示出的实施例中,每个闪存控制器140使用逻辑到物理地址转换数据结构(诸如逻辑到物理转换(LPT)表300)来为相关联的LBA集合进行逻辑到物理地址转换,所述逻辑到物理地址转换数据结构可以存储在例如相关联的闪存控制器存储器142中。应当注意,提供给闪存控制器140的逻辑地址可以不同于最初提供给数据存储系统120的逻辑地址,因为数据存储系统120内的各种组件可以执行外部设备与闪存控制器140之间的地址转换操作。

[0035] 在GPP 132上运行的闪存管理代码跟踪NAND闪存系统150的在随时待用(RTU)队列306中的随时待用的被擦除块,该队列306可以存储在例如GPP存储器134中。在所示实施例中,在GPP 132上运行的闪存管理代码维持每个通道或平面(即,每个数据总线)一个的RTU队列306,并且将要被重用的每个被擦除块的标识符在对应于其通道或平面的RTU队列306中排队。此外,可以针对不同的健康等级、热仓/写入流或块操作模式(即,SLC或QLC操作模式)维护单独的RTU队列306。由在GPP 132上运行的闪存管理代码执行的构建块条带功能320为存储来自在RTU队列306中排队的被擦除的块的数据和相关的奇偶校验信息而构建新的块条带。新的块条带然后在闪存控制器140中排队,以用于数据放置。块条带优选地由驻留在不同通道中的块形成,这意味着构建块条带功能320可以通过从不同的RTU队列306中拉来新块条带的每个块以方便地构建块条带。通常,构建块条带功能320试图用近似相等的健康(即,预期的剩余使用寿命)的块来构建条。

[0036] 响应于从诸如处理器系统102的主机接收的写入请求,闪存控制器140的数据放置功能310通过参考LPT表300,确定写入请求中指示的目标LBA当前是否被映射至NAND闪存系统150中的物理存储器页,并且,如果是,则改变当前与目标LBA相关联的每个数据页的状态,以指示相关联的数据不再有效。此外,如果必要,数据放置功能310还分配页条带,用以

存储写入请求的写入数据以及—如果存在的话—作为写入请求目标的现有页条带的任何未更新的数据(即,对于小于逻辑页的写入请求,来自先前写入到未被覆写且必须以读取-修改-写入方式处理的相同逻辑地址的剩余有效数据),和/或存储写入请求的写入数据以及—如果存在的话—作为写入请求目标的现有页条带的任何未更新(即,仍有效)的数据至已经分配的剩余有空闲空间的页条带。可以从已经分配用于保存数据的块条带或者从由构建块条带功能320构建的新的块条带分配页条带。在优选实施例中,页条带分配可以基于可用于分配的块的健康度和写入数据的LBA的“热”度(即,估计或测量的写入访问频率)。数据放置功能310然后将写入数据、相关联的元数据(例如,循环冗余码(CRC)和纠错码(ECC)值)以及页条带的奇偶校验信息写入所分配的页条带。闪存控制器140还更新LPT表300以将用于存储写入数据的物理页与主机设备指示的LBA相关联。此后,闪存控制器140可以通过参考LPT表300来访问数据以服务主机读取请求,如图3中进一步所示的那样。

[0037] 一旦块条带中的所有页都已经被写入,闪存控制器140将块条带放置到被占用块队列302的其中之一中,运行在GPP 132上的闪存管理代码利用该队列来方便垃圾收集。如上所述,通过写入处理,页被无效,因此NAND闪存系统150的部分变为未使用。相关联的闪存控制器140(和/或GPP 132)最终需要通过由垃圾收集器312执行的垃圾收集来回收该空间。垃圾收集器312基于许多因素选择用于垃圾收集的特定块条带,所述因素例如包括,块条带内的块204的健康,以及擦除块204内的多少数据是无效的。在所示示例中,在整个块条带上执行垃圾收集,并且,GPP 132上运行的闪存管理代码将准备好被再循环的块条带记录在重定位队列304中,该重定位队列304可以方便地在相关联的闪存控制器存储器142或GPP存储器134中实现。

[0038] 由GPP 132或闪存控制器140执行的闪存管理功能另外包括重定位功能314,该重定位功能314重定位在重定位队列304中排队的块条带中保存的数据。为了重定位这样的数据,重定位功能314向数据放置功能310发出重定位写入请求,以请求将旧块条带的有效数据写入NAND闪存系统150中的新块条带。此外,重定位功能314更新LPT表300,以解除数据的逻辑地址和物理地址之间的当前关联。一旦所有剩余的有效数据都已经从旧块条带转移,则将旧块条带传递到分解块条带功能316,其将旧块条带分解成其构成块204,从而解除块204的关联。然后在相关闪存模块200的闪存控制器140和/或控制电路205的指引下,擦除先前形成分解的块条带的块204中的每个块,并递增每个被擦除的块的对应编程/擦除(P/E)周期计数334。基于每个被擦除块204的健康度量332(例如,误码率(BER)度量、不可纠正的错误、P/E周期计数等),每个被擦除块204或者被GPP 132上执行的闪存管理功能之中的块退出功能318退出(即,退出使用)中撤回),或者可替代地,通过将块204的标识符置于相关联的GPP存储器134中的适当的随时待用(RTU)队列306上而被准备再用。

[0039] 如图3中进一步所示,在GPP 132和/或闪存控制器140上执行的闪存管理功能额外地包括后台健康检查器330。独立于主机(诸如处理器系统102)的读取和写入请求而操作的后台健康检查器330连续地为属于在被占用块队列302中记录的块条带的块确定一个或多个块健康度量332。基于块健康度量332中的一个或多个,后台健康检查器330将块条带置于重定位队列304上供重定位功能314处理。优选地由后台健康检查器监测和记录的关键的块健康度量332,涉及针对有效块和物理页观察到的误码率(BER)度量,并且可以包括例如每个块的最差页BER、每个块的平均页BER、每个块的最差页BER和平均页BER的变化速率等。为

了获得尽可能最准确的健康估计,可以通过分析有效和无效数据来确定健康,由此确保完全分析包含几乎全部无效数据的块,以确定总体块健康。

[0040] GPP 132和/或闪存控制器140的一个可以结合在后台健康检查器330内或者可以单独实现的功能,是用于对存储在存储器阵列202的存储器单元内的数据位进行解码的读取电压阈值336的周期性适配(校准)。这些读取电压阈值336可以被单独地定义为任何期望的粒度级别(例如,每页、块内的每页组、每块等),优选地被选择以改进一个或多个误码率度量。如在图3中进一步描述的,GPP 132和/或闪存控制器140优选地跟踪额外的统计以促进读取电压阈值336的智能适配。这些额外的统计可以例如包括标识每个块已经经历的编程/擦除周期的数量的P/E周期计数334,以及指示给定存储器单元(例如,页、页组和/或块)自从被编程以来已经被读取的次数的读取计数338。此外,这些统计可以包括校准触发器340,诸如指示自对该存储器单元的读取电压阈值进行适配以来所经过的时间量的每存储器单元的计数器和/或指示自对该块的读取电压阈值上次进行适配以来每个块所经历的编程/擦除周期数的每块的计数器。

[0041] 现在参见图4,示出了可以在非易失性存储器设备中实现的示范性存储架构。在该示例中,GPP 132和/或闪存控制器140(在本文中统称“控制器”)配置NAND闪存系统150中的闪存存储设备152中的一些或全部,以实现物理块204的多个池,其中,不同块池中的物理块204被配置为在每个单元存储不同数量的位的不同模式下操作。因此,在该示例中,虽然存储器阵列202的全部包括例如能够每单元存储四位的QLC闪存,但是,控制器将闪存设备152的存储器阵列202中的一些物理块204分配给其中完全编程的块每单元存储四位的QLC池400,并将存储器阵列202的其他物理块204分配给其中完全编程的块每单元仅存储单个位的SLC池402。例如,可以在GPP存储器124和/或闪存控制器存储器142中存储池成员资格信息,作为块模式元数据335。如上所述,如果频繁写入的LBA被映射到SLC池402内的物理块204,则以这种方式以多种不同模式操作块可以提供改进的性能和/或耐久性。

[0042] 尽管不是必需的,在一些实施例中,控制器以分层的布置来组织QLC池400和SLC池402,其中,主机写入请求中包含的数据首先被写入从SLC池402选择的物理块204中,如箭头404所示。当SLC池402通过使用而填充并且在SLC池402中需要和/或期望更多的容量时,控制器通过重定位写入将数据从SLC池402转储到QLC池400,如箭头406所示。为了在块池400和402二者中确保当块经受磨损时物理块204的足够供应,控制器根据需要和/或期望在池400和402之间另外转移(重新分配)块204以提供存储容量和/或耐久性。物理块204(与数据相反)在块池400与402之间的转移在图4中由箭头408描绘。

[0043] 本公开认识到,在块池400与402之间转移块的过程如果没有仔细设计,可能会不利地导致性能降低(由于错误恢复活动增加),或者甚至在错误恢复不成功的情况下(例如,如果将具有高误码率的块从块池402转移到块池400)导致数据丢失。一般而言,控制器通过从块池400选择较不健康/最不健康的块转移到块池402并且通过从块池402选择比较健康/最健康的块转移到块池400来在块池400和402之间转移块。

[0044] 本公开认识到,这种模式转换策略虽然适合一些系统,但在采用热隔离或具有非常高的设备利用率的系统中是次优的,这可直接导致加速磨损、甚至提前的设备退出。这些负面影响的一个原因是,QLC池400中的块204的磨损(例如按平均P/E周期来测量)在某些条件下可能通常远高于SLC池402中的块204的磨损,这部分地归因于写入放大率。图5是如图4

中所示的示范性分层存储架构的写入放大率与设备利用率的关系图。如图所示,在高设备利用率下(例如,当超过90%的闪存设备被写入数据时),QLC池400的写入放大率会呈指数增长,从而导致QLC池400中的块的磨损越来越快。如下文更详细讨论的,因此,模式转换策略在选择在块池之间转移的块时最好考虑到块磨损。

[0045] 现在参见图6,描绘了根据一个实施例的用于配置非易失性存储器的示范性过程的高级逻辑流程图。所示过程例如可以由硬件、固件、软件或其组合中的控制器(例如,GPP 132和/或闪存控制器140)在数据存储系统120启动时执行。与在此呈现的其他逻辑流程图一样,操作是按逻辑而不是严格按时间顺序呈现的,并且在一些实施例中,可以按与所呈现的顺序不同的顺序或同时执行操作。

[0046] 图6的配置过程开始于框600,然后前进到框602,其示出的闪存设备152的控制器配置块池的期望数量。在图4的示例中,控制器在框602处实现两个块池,即QLC块池400和SLC块池402。在其他实施例中,可以实现更大数量的块池和/或以不同块模式操作的块池(例如,TLC或MLC)。控制器还向每个块池分配初始数量的物理块204,并基于每个物理块204的块池成员资格,在块模式元数据335中记录每个物理块204的对应块模式。

[0047] 该过程从框602前进到框604,其示出控制器处理在框602处实现的每个块池的处理循环。如果控制器在框604处确定不是所有块池都已经被处理,则该过程转到框610,其示出控制器选择下一个待处理的块池(例如,QLC池400或SLC池402)。在框612,控制器然后为属于所选择的块池的块204配置健康等级的期望数量。在各种实施例中,健康等级的数目可以变化,但是已经实验性地发现在两个和八个(含)之间的健康等级的数目是有效的。例如,在图4中描绘的一个实施例中,控制器在框612处配置四个健康等级410a-410d或412a-412d。这里采用的健康等级被具体地定义为基于除了仅仅编程/擦除(P/E)周期的数量之外的至少一个标准,例如,物理存储器的每个块(和/或每个块内的物理页)的一个或多个误码率度量。例如,在一个实施例中,不同健康等级由在编程之后的少量读访问内确定的最差页BER的相应范围定义,因为该误码率度量准确地反映对块的磨损的永久性影响。基于所测量的最差页BER,最健康的块是在编程之后不久具有最低最差页BER的块,最不健康的块是在编程之后不久具有最高最差页BER的块。

[0048] 在方框614处,控制器还为由闪存储存设备152服务的LBA配置热仓的期望数量。每个热仓对应于LBA的写入热度(即,写入频率)的相应范围。在各种实施例中,热仓的数目可以变化,但是已经实验性地发现热仓数目在两个和八个(含)之间是有效的。例如,在一个实施例中,控制器在框614处配置四个热仓,分别对应于最频繁被写入的热LBA、较不频繁被写入的暖LBA、更不频繁被写入的凉LBA以及最不频繁被写入的冷LBA。在本文中将目标为给定热仓内的LBA的一组主机写入请求称为“写入流”。在框614处配置的热仓(且因此写入流)的数目可以等于、但无须等于在框612处配置的健康等级的数目。

[0049] 在框616处,控制器建立在框612处配置的热仓与在框614处配置的健康等级之间的映射。在一个特别优选的实施例中,将热和健康与直接相关,其中最热的LBA被映射到最健康等级,最冷的LBA被映射到较不健康等级。当然,在其他实施例中,可以采用写入热度与健康等级之间的其他映射。例如,在一个采用数字写入热度和健康等级的实施例中,在框616处建立的映射,是由经选择以从彼此映射的健康等级和写入热度产生近似恒定结果(例如,C≈f(热,健康))的数学函数来确定。在另一实施例中,从写入热度到健康等级的映射将

每个热仓映射到健康等级的排序的优先级列表。在从RTU队列306中的块建立块条带时,在优选RTU队列306中可能不存在具有给定健康等级的块。排序的优先级列表允许分配来自健康等级尽可能接近优选RTU队列的另一RTU队列306的块。

[0050] 在框616之后,图6中给出的配置过程返回到框604。如果在框604作出否定性确定,从而表明仍然有一个或多个其它块池有待处理,则该过程返回至已经描述的框610-616。然而,如果在框604确定在框602配置的所有块池已经被处理,则图6的配置过程在框606结束。

[0051] 现在参见图7,示出了根据一个实施例的用于在块池之间动态地转移块的示范性过程的高级逻辑流程图。所示过程例如可以由硬件、固件、软件或其组合中的控制器(例如,GPP 132和/或闪存控制器140)在数据存储系统120启动时执行。

[0052] 图7的过程开始于700,然后前进到框702,其示出了控制器确定是否要重新调整块池400和402之一的大小。例如,控制器可以基于一个或多个标准来确定是否重新调整给定块池的大小,该一个或多个标准包括:属于一个或多个RTU队列306内的给定块池的块的填充水平未能满足阈值;属于一个或多个RTU队列306内的另一块池的块的填充水平;给定块池内的块的P/E周期计数;给定块池内的被分配一个或多个健康等级的块的数量,等等。在框702做出的确定,可以例如由后台健康检查器330、垃圾收集器312、块退出功能318和/或构建块条带功能320中的一个或多个触发。响应于控制器在框702确定不要调整块池的大小,过程在框702重复。响应于在框702确定要重新调整块池的大小,该过程行进至框704。

[0053] 框704示出控制器选择物理块204将被转移到的块池。另外,在框706,控制器选择物理块204将被添加到给定块池的热仓或写流(例如,热、温、凉或冷)。控制器然后在框708确定当前被分配来存储与所选择的热仓中的LBA相关联的数据的块204的流磨损度量。在一个优选实施例中,控制器跟踪分配给每个热仓/写入流的块204,并且(例如,从P/E周期计数334)确定那些块204已经经历的P/E周期的总数。根据这些数目,控制器可将所选择的热仓的流磨损度量计算为在热仓中存储数据的块204的总数除以那些块204已经经历的P/E周期的总数。

[0054] 基于在框708处确定的流磨损度量,控制器然后从其他块池中选择要转移到所选择的块池的具有适合于所选择的热仓的单个块健康的至少一个未被编程的块204(框710)。此选择确保最健康的块204将经历最高数量的未来P/E周期,最不健康的块204将经历最少数量的未来P/E周期。结果,所有类型的工作负荷和所有的设备利用点的设备耐久性都得到改善。在框712,控制器将在框710选择的块204转移到所选择的块池,用新操作模式(例如,QLC或SLC)更新块模式元数据335,并且,如果必要的话,更新块健康等级。此后,图7的过程返回至已经描述的框702。

[0055] 如已经描述的,在至少一个实施例中,非易失性存储器包括多个物理存储块,每个物理存储块包括相应的多个单元,其中多个单元中的每个单元能够单独地存储多位数据。控制器将多个物理块之中的物理块分配到含有在第一(例如,QLC)模式下操作的物理块的第一池以用于存储每单元较大数量的位,并且将多个物理块之中的其他物理块分配到含有在第二(例如,SLC)模式下操作的物理块的第二池以用于存储每单元较小数量的位。控制器至少基于为所转移的物理块测量的误码率在第一池和第二池之间转移物理块。

[0056] 本发明可以是系统、方法和/或计算机程序产品。计算机程序产品可包括其上具有用于使处理器执行本发明的各方面的计算机可读程序指令的计算机可读存储介质。

[0057] 计算机可读存储介质可以是可保留和存储供指令执行设备使用的指令的有形设备。计算机可读存储介质可以是,例如但不限于,电子存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备、或者上述的任意合适的组合。计算机可读存储介质的更具体示例的非穷尽列表包括以下各项:便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式紧凑盘只读存储器(CD-ROM)、数字通用盘(DVD)、记忆棒、软盘、具有记录在其上的指令的诸如穿孔卡或槽中的凸出结构之类的机械编码设备、以及上述各项的任何合适的组合。如本文所使用的计算机可读存储介质不应被解释为暂时性信号本身,例如无线电波或其他自由传播的电磁波、通过波导或其他转移介质传播的电磁波(例如,穿过光纤电缆的光脉冲)或通过电线发射的电信号。

[0058] 本文中所描述的计算机可读程序指令可以经由网络(例如,因特网、局域网、广域网和/或无线网络)从计算机可读存储介质下载到相应的计算/处理设备,或者下载到外部计算机或外部存储设备。网络可以包括铜转移电缆、光转移纤维、无线转移、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配器卡或网络接口接收来自网络的计算机可读程序指令,并转发计算机可读程序指令以存储在相应计算/处理设备内的计算机可读存储介质中。

[0059] 用于执行本发明的操作的计算机可读程序指令可以是汇编指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或以一种或多种程序设计语言的任何组合编写的源代码或目标代码,这些程序设计语言包括面向对象的程序设计语言(诸如Smalltalk、C++等)、以及常规的过程式程序设计语言(诸如“C”程序设计语言或类似程序设计语言)。计算机可读程序指令可以完全地在用户计算机上执行、部分在用户计算机上执行、作为独立软件包执行、部分在用户计算机上部分在远程计算机上执行或者完全在远程计算机或服务器上执行。在后一种情况下,远程计算机可通过任何类型的网络(包括局域网(LAN)或广域网(WAN))连接至用户计算机,或者可连接至外部计算机(例如,使用因特网服务提供商通过因特网)。在一些实施例中,包括例如可编程逻辑电路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA)的电子电路可以通过利用计算机可读程序指令的状态信息来使电子电路个性化来执行计算机可读程序指令,以便执行本发明的各方面。

[0060] 参照根据本发明实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述了本发明。应当理解,流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合,都可以由计算机可读程序指令实现。

[0061] 这些计算机可读程序指令可被提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器以产生机器,使得经由计算机或其他可编程数据处理装置的处理器执行的指令创建用于实现在流程图和/或框图的或多个框中指定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中,这些指令使得计算机、可编程数据处理装置、和/或其他设备以特定方式工作,从而其中存储有指令的计算机可读存储介质包括包含实现流程图和/或框图中的或多个方框中规定的功能/动作的方面的指令的制造品。

[0062] 也可以把计算机可读程序指令加载到计算机、其他可编程数据处理装置、或其他设备上,使得在计算机、其他可编程装置或其他设备上执行一系列操作步骤,以产生计算机

实现的处理,使得在计算机、其他可编程装置或其他设备上执行的指令实现流程图和/或框图中的或多个方框中规定的功能/动作。

[0063] 附图中的流程图和框图示出了根据本发明的不同实施例的系统、方法和计算机程序产品的可能实现方式的架构、功能和操作。对此,流程图或框图中的每个框可表示指令的模块、段或部分,其包括用于实现指定的逻辑功能的一个或多个可执行指令。在一些备选实现中,框中标注的功能可以不按照图中标注的顺序发生。例如,取决于所涉及的功能,连续示出的两个框实际上可以基本上同时执行,或者这些框有时可以以相反的顺序执行。也要注意的是,框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合,可以用执行规定的功能或动作或执行专用硬件与计算机指令的组合的专用的基于硬件的系统来实现。

[0064] 虽然已经参照一个或多个优选实施例具体地示出本发明,但是本领域技术人员将理解,在不脱离由所附权利要求限定的本发明的范围的情况下,可以在其中进行形式和细节上的各种改变。例如,尽管已经关于包括指引某些功能的闪存控制器的数据存储系统描述了各方面,但是应当理解,本发明可以替代地实现为包括存储程序代码的存储设备的程序产品,所述程序代码可以由处理器处理以执行这样的功能或使得执行这样的功能。如本文中使用的,“存储设备”被具体限定为仅包括法定制品并且排除信号介质本身、瞬态传播信号本身和能量本身。

[0065] 此外,尽管已经描述了包括NAND闪存的使用的实施例,但是应当理解,本发明的实施例也可以与任何其他类型的非易失性随机存取存储器(NVRAM)一起使用。

[0066] 上述附图以及特定结构和功能的书面描述并非是要限制申请人已发明的范围或所附权利要求的范围。相反,所提供附图和书面说明是用以教导本领域的技术人员制造和使用请求专利保护的本发明。本领域技术人员将理解,为了清楚和理解,并非描述或显示了本发明的商业实施例的所有特征。本领域技术人员还将认识到,结合本发明各方面的实际商业实施例的开发将需要许多具体实施的决策来实现开发者对商业实施例的最终目标。此类具体实施的决策可包括并且可能不限于与系统相关的、商业相关的、政府相关的和其他的约束的合规性,这些约束可随具体实施、位置和时间而变化。虽然开发者的努力在绝对意义上可能是复杂且耗时的,但是对于受益于本公开的本领域技术人员而言,这样的努力仍然是常规任务。必须理解,本文公开和教导的发明容易受到众多和各种修改和替代形式的影响。最后,使用单数术语(诸如但不限于“一个”),并不意味着限制项目数量。

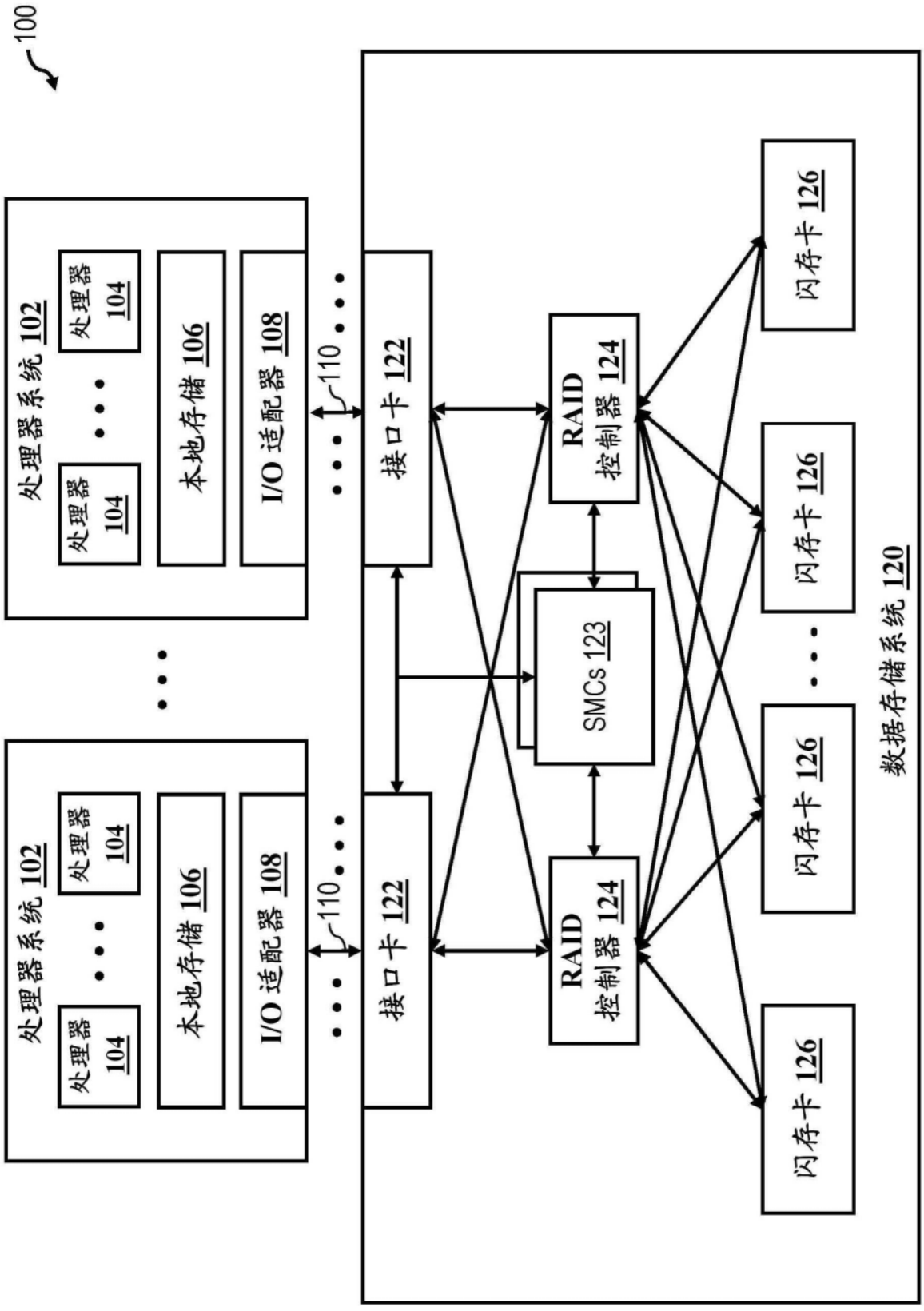
100
处理器系统 102
处理器 104
处理器 104
本地存储 106
I/O 适配器 108
110
处理器系统 102
处理器 104
处理器 104
本地存储 106
I/O 适配器 108
110
接口卡 122
接口卡 122
RAID 控制器 124
RAID 控制器 124
SMCs 123
闪存卡 126
闪存卡 126
闪存卡 126
闪存卡 126
闪存卡 126
数据存储系统 120

图1A

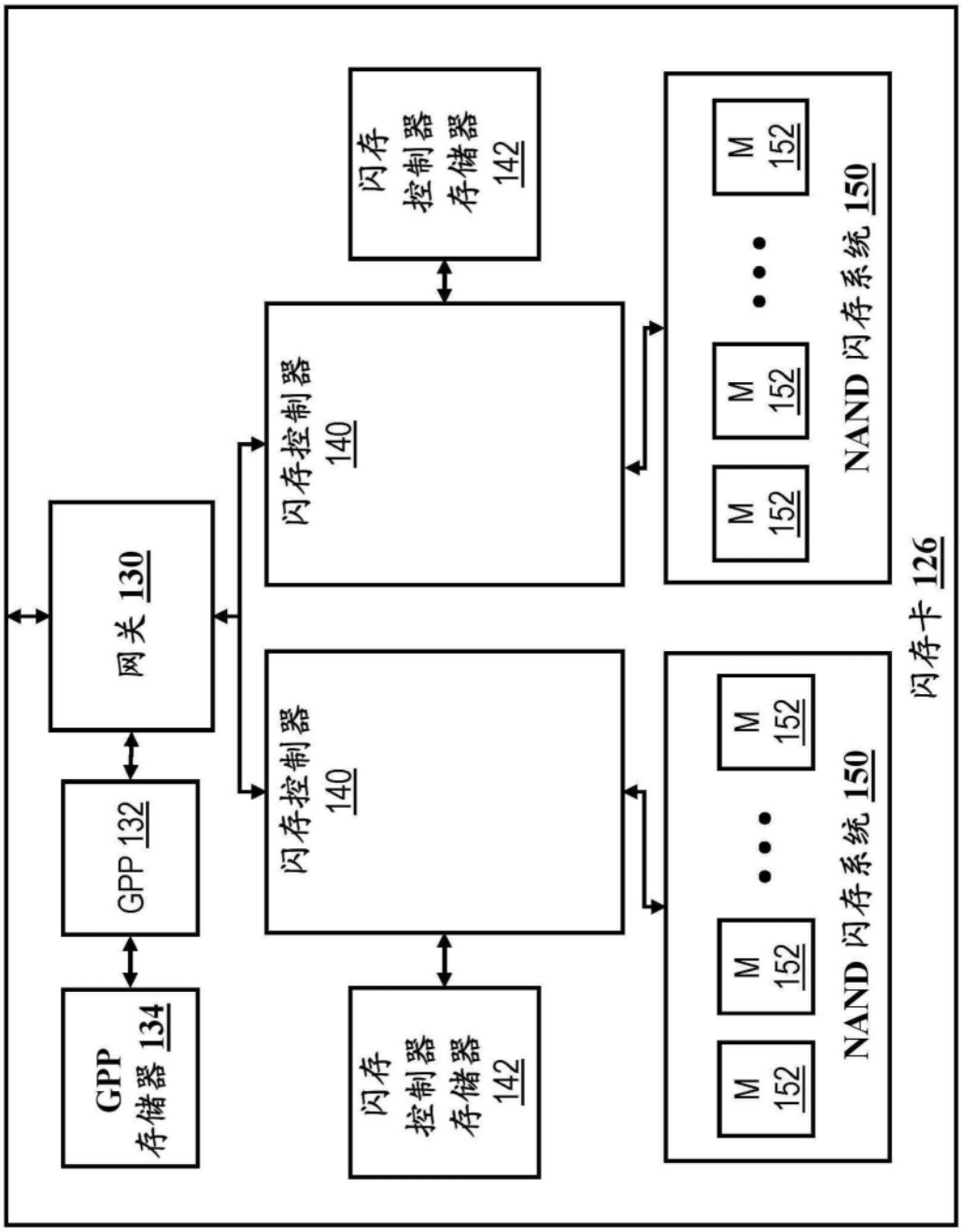
闪存卡 126
网关 130
GPP 132
GPP 存储器 134
闪存控制器 140
闪存控制器 140
闪存控制器存储器 142
闪存控制器存储器 142
NAND 闪存系统 150
NAND 闪存系统 150
M 152
M 152
M 152
M 152
M 152
M 152

图1B

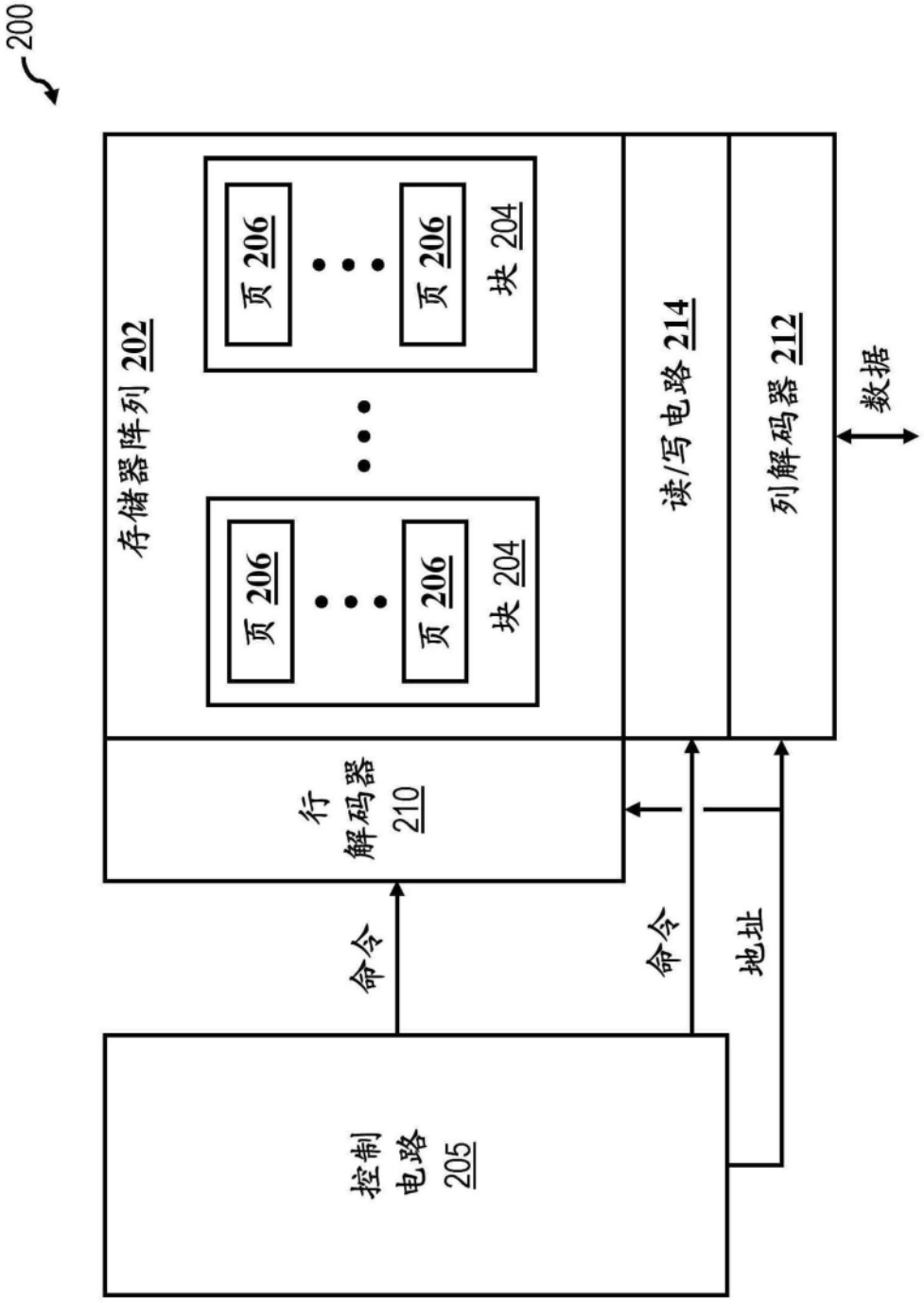
200
存储器阵列 202
页 206
页 206
块 204
页 206
页 206
块 204
行解码器 210
读/写电路 214
列解码器 212
数据
命令
命令
地址
控制电路 205

图2

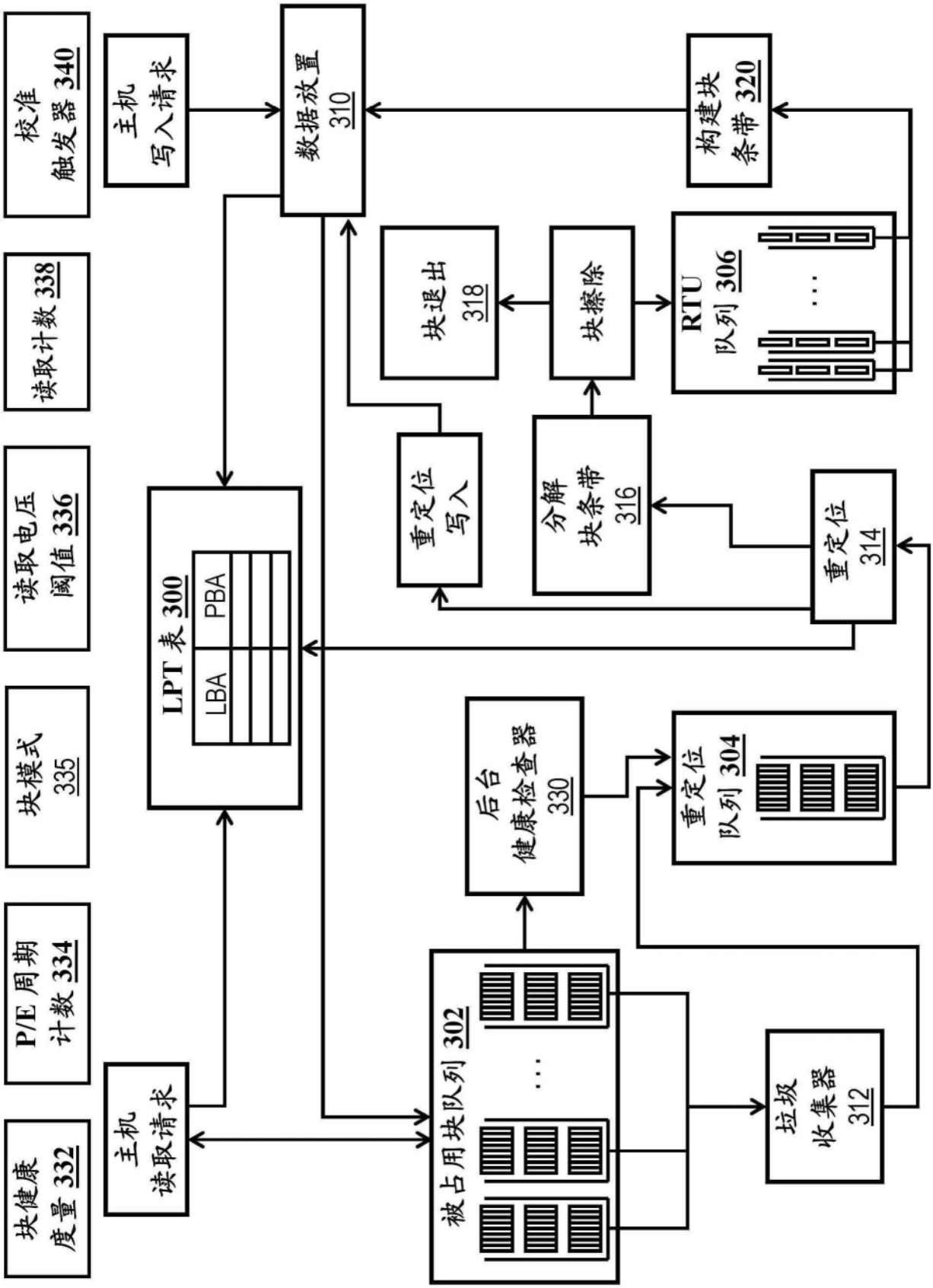
校准
触发器 340
主机
写入请求
数据放置
310
构建块
条带 320
读取计数 338
块退出
318
块擦除
RTU
队列 306
读取电压
阈值 336
重定位
写入
分解
块条带
316
重定位
314
LPT 表 300
LBA
PBA
块模式
335
后台
健康检查器
330
重定位
队列 304
P/E 周期
计数 334
被占用块队列 302
垃圾
收集器
312
块健康
度量 332
主机
读取请求

图3

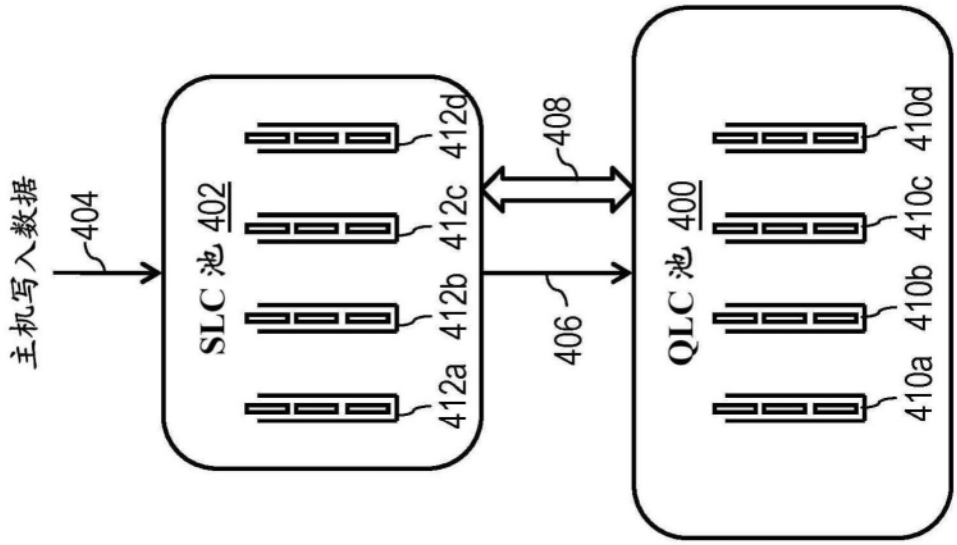
主机写入数据
404
SLC 池 402
412a
412b
412c
412d
406
408
QLC 池 400
410a
410b
410c
410d

图4

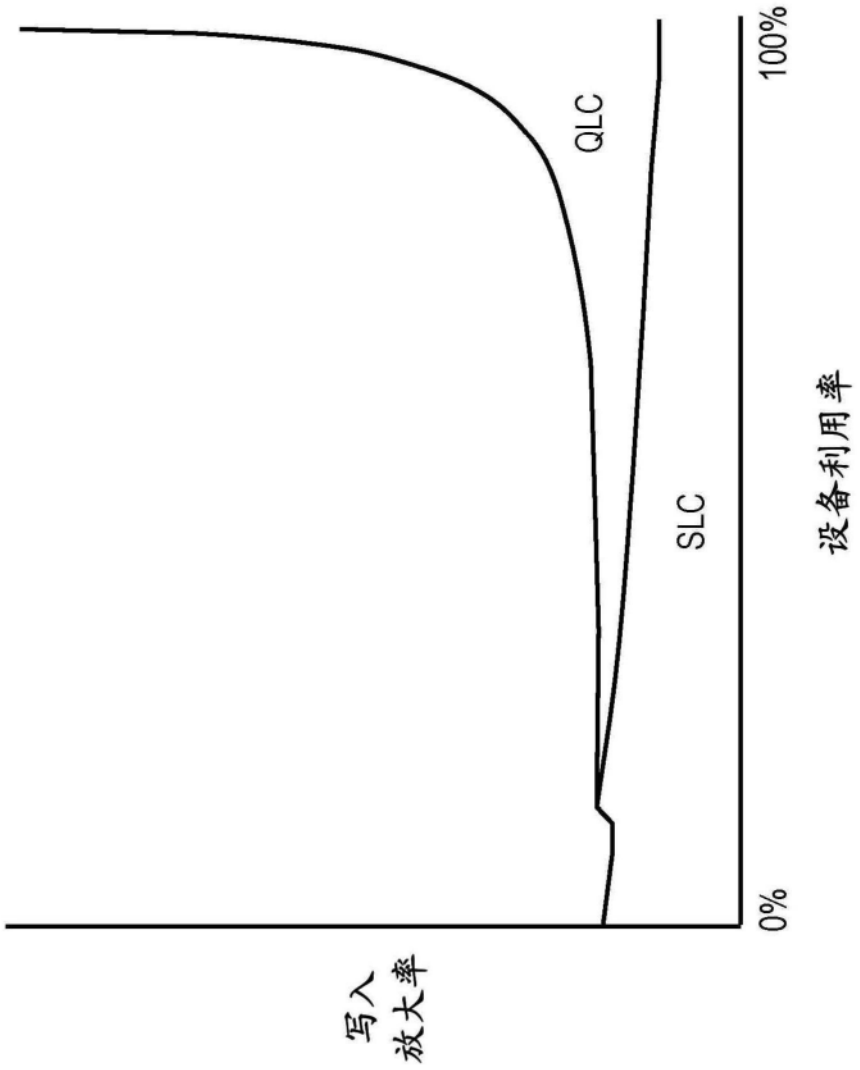

图5

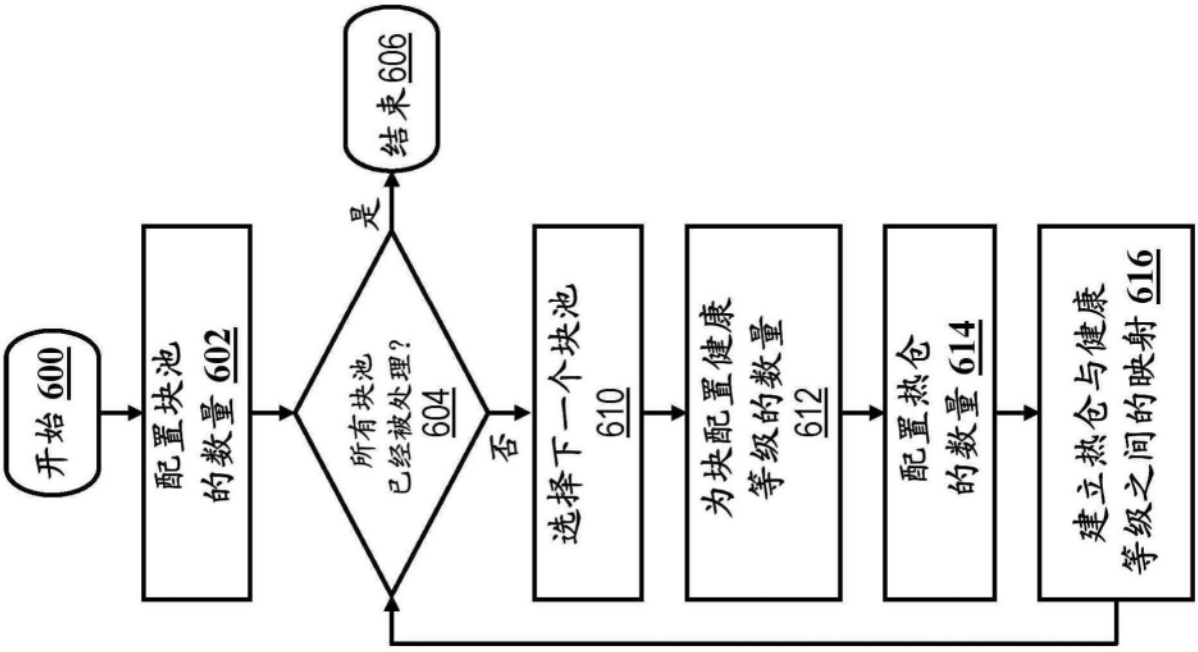
开始 600
配置块池的数量 602
所有块池已经被处理? 604
是
结束 606
否
选择下一个块池 610
为块配置健康等级的数量 612
配置热仓的数量 614
建立热仓与健康等级之间的映射 616

图6

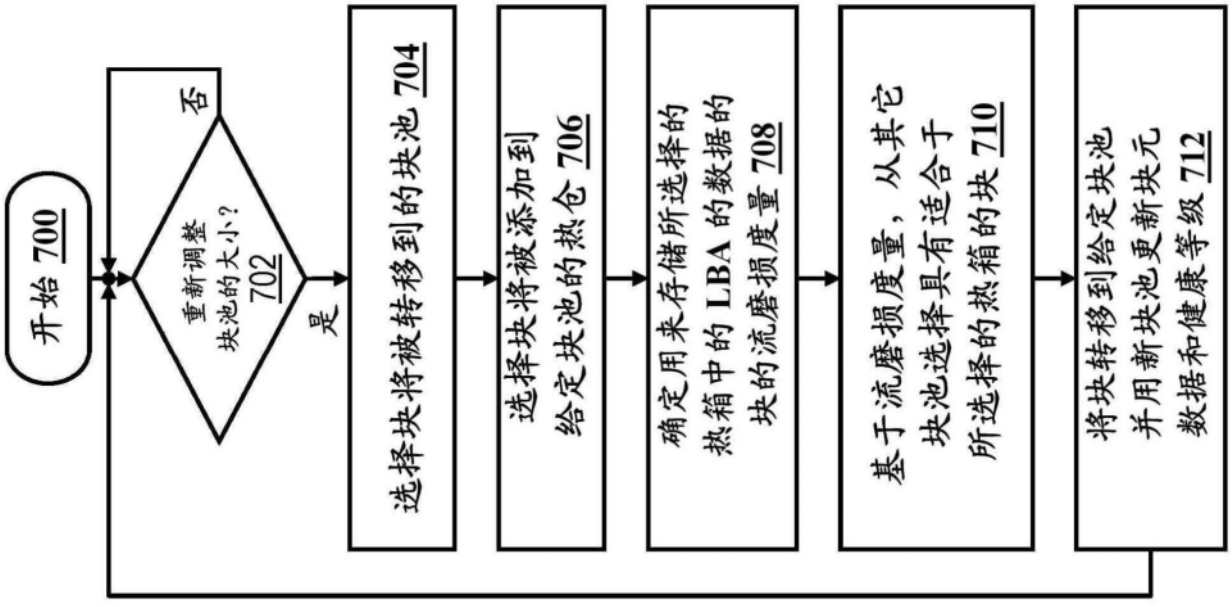
开始 700
重新调整块池的大小? 702
否
是
选择块将被转移到的块池 704
选择块将被添加到给定块池的热仓 706
确定用来存储所选择的热箱中的 LBA 的数据的块的流磨损度量 708
基于流磨损度量，从其它块池选择具有适合于所选择的热箱的块 710
将块转移到给定块池并用新块池更新块元数据和健康等级 712

图7