(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 105468293 B
(45)授权公告日 2019.02.12

(21)申请号 201510629154.1

(22)申请日 2015.09.29

(65)同一申请的已公布的文献号
申请公布号 CN 105468293 A

(43)申请公布日 2016.04.06

(30)优先权数据
14/501,917 2014.09.30 US

(73)专利权人 EMC公司
地址 美国马萨诸塞州

(72)发明人 H·塔巴雷茨 R·阿加瓦尔
J·P·费雷拉 J·S·邦威克
M·W·夏皮罗

(74)专利代理机构 北京英赛嘉华知识产权代理有限责任公司 11204
代理人 王达佐 王艳春

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

(56)对比文件
US 2011/0173484 A1,2011.07.14,
US 2011/0173484 A1,2011.07.14,
CN 102150140 A,2011.08.10,
US 8189379 B2,2012.05.29,
JP 特开2010-79486 A,2010.04.08,
CN 103902234 A,2014.07.02,
JP 特表2014-515537 A,2014.06.30,

审查员 陈国耀

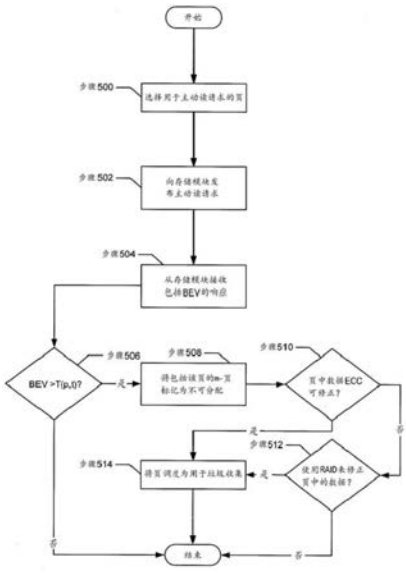
权利要求书3页 说明书10页 附图12页

(54)发明名称

通过预测损坏的m-页提高闪存利用率的方法和系统

(57)摘要

本发明公开了一种通过预测损坏的m-页提高闪存利用率的方法和系统,其涉及一种用于管理持久性存储器的方法。该方法包括选择用于主动读请求的页,其中,该页位于持久性存储器中。该方法还包括:向该页发布主动读请求,响应于该主动读请求而接收用于存储在该页上的数据的位错误值(BEV),获得用于该页的BEV阈值(T),其中,使用与该页相关联的编程/擦除循环值和存储在该页上的数据的保持时间来确定T,进行BEV大于T的第一确定,基于该第一确定:识别m-页,其中,m-页是一组页,其中,该页在该一组页中,以及将该m-页设定为对未来操作不可分配。



CN 105468293 B

1.一种用于管理持久性存储装置的方法,所述方法包括:

选择用于主动读请求的页,其中,所述页位于所述持久性存储装置中;

由控制模块向所述页发布所述主动读请求;

响应于所述主动读请求而接收用于存储在所述页上的数据的位错误值;

获得用于所述页的位错误值阈值T,其中,使用与所述页相关联的编程/擦除循环值和存储在所述页上的所述数据的保持时间来确定T;

进行所述位错误值大于T的第一确定;

基于所述第一确定:

识别m-页,其中,所述m-页是一组页,其中,所述页在所述一组页中;以及

将所述m-页设定为对未来操作不可分配,并且

其中,在将所述m-页设定为不可分配之后,所述页保持可读。

2.根据权利要求1所述的方法,还包括:

进行所述页上的所述数据为能够使用纠错码修正的第二确定;

基于所述第二确定:

将所述页上的所述数据调度为被写入到所述持久性存储装置中的新的可分配页。

3.根据权利要求2所述的方法,还包括:

在所述调度之后:

将所述页上的所述数据作为垃圾收集操作的一部分写入到所述新的可分配页。

4.根据权利要求1所述的方法,还包括:

进行所述页上的所述数据为不能够使用纠错码修正的第二确定;

基于所述第二确定,进行使用独立磁盘冗余阵列修正机制来主动地修正所述页上的所述数据的第三确定。

5.根据权利要求4所述的方法,其中,所述第三确定将持久性存储装置中的至少一个其它页的位错误值考虑在内,其中,所述页和所述至少一个其它页是独立磁盘冗余阵列条带的一部分。

6.根据权利要求1所述的方法,其中,未来操作包括选自由写操作和垃圾收集操作组成的组的至少一个。

7.根据权利要求1所述的方法,其中,所述位错误值指定所述页中不正确的位的百分比。

8.根据权利要求1所述的方法,其中,所述位错误值指定所述页中不正确的位的数目。

9.根据权利要求1所述的方法,其中,使用纠错码来确定所述位错误值。

10.根据权利要求1所述的方法,其中,所述页位于所述持久性存储装置中的固态存储器模块上,其中,所述固态存储器模块包括多层单元。

11.根据权利要求1所述的方法,其中,使用数据被写入到所述页的第一时间和与所述主动读请求相关联的第二时间来确定所述保持时间。

12.根据权利要求11所述的方法,其中,从存储器中的存储器内数据结构获得所述第一时间,其中,所述存储器位于控制模块中。

13.根据权利要求1所述的方法,其中,所述编程/擦除循环值是编程/擦除循环值范围。

14.一种持久性存储装置管理系统,包括:

存储模块，所述存储模块包括存储模块控制器和持久性存储装置；以及

控制模块，所述控制模块被操作连接到所述存储模块和客户端，其中，所述控制模块执行以下步骤：

选择用于主动读请求的页，其中，所述页位于所述持久性存储装置中；

由控制模块向所述页发布所述主动读请求；

响应于所述主动读请求而接收用于存储在所述页上的数据的位错误值；

获得用于所述页的位错误值阈值T，其中，使用与所述页相关联的编程/擦除循环值和存储在所述页上的所述数据的保持时间来确定T；

进行所述位错误值大于T的第一确定；

基于所述第一确定：

识别m-页，其中，所述m-页是一组页，其中，所述页在所述一组页中；以及

将所述m-页设定为对未来操作不可分配，并且

其中，在将所述m-页设定为不可分配之后，所述页保持可读。

15.根据权利要求14所述的系统，其中，所述持久性存储装置包括闪存，并且其中，所述页位于所述闪存中。

16.根据权利要求14所述的系统，

其中，所述控制模块包括存储器，其中，所述存储器包括存储器内数据结构，所述存储器内数据结构包括用于存储在所述持久性存储装置中的所有数据的产生时间和多个条目，其中，所述多个条目中的每一个包括多个位错误值阈值中的一个、多个编程/擦除循环值中的一个以及多个保持时间中的一个；

其中，获得所述位错误值包括在所述存储器内数据结构中执行查找，

其中，使用用于存储在来自所述存储器的所述页上的数据的产生时间来确定用于存储在所述页上的所述数据的保持时间。

17.根据权利要求14所述的系统，其中，所述控制模块还执行以下步骤：

进行所述页上的所述数据为能够使用纠错码修正的第二确定；

基于所述第二确定：

将所述页上的所述数据调度为被写入到所述持久性存储装置中的新的可分配页；

在所述调度之后，将所述页上的所述数据作为垃圾收集操作的一部分写入到所述新的可分配页。

18.根据权利要求14所述的系统，其中，所述控制模块还执行以下步骤：

进行所述页上的所述数据为不能够使用纠错码修正的第二确定；以及

基于所述第二确定，进行使用独立磁盘冗余阵列修正机制来主动地修正所述页上的所述数据的第三确定。

19.根据权利要求18所述的系统，其中，所述第三确定将所述持久性存储装置中的至少一个其它页的位错误值考虑在内，其中，所述页和所述至少一个其它页是独立磁盘冗余阵列条带的一部分。

20.一种包括计算机可读程序代码的非临时计算机可读介质，所述计算机可读程序代码在被计算机处理器执行时使得所述计算机处理器能够：

选择用于主动读请求的页，其中，所述页位于持久性存储装置中；

由控制模块向所述页发布所述主动读请求；

响应于所述主动读请求而接收用于存储在所述页上的数据的位错误值；

获得用于所述页的位错误值阈值T，其中，使用与所述页相关联的编程/擦除循环值和存储在所述页上的所述数据的保持时间来确定T；

进行所述位错误值大于T的第一确定；

基于所述第一确定：

识别m-页，其中，所述m-页是一组页，其中，所述页在所述一组页中；以及

将所述m-页设定为对未来操作不可分配，并且

其中，在将所述m-页设定为不可分配之后，所述页保持可读。

21.一种用于管理持久性存储装置的方法，所述方法包括：

选择用于主动读请求的页，其中，所述页位于所述持久性存储装置中；

由控制模块向所述页发布所述主动读请求；

响应于所述主动读请求而接收用于存储在所述页上的数据的位错误值；

获得用于所述页的位错误值阈值T，其中，使用与所述页相关联的编程/擦除循环值和存储在所述页上的所述数据的保持时间来确定T，其中，获得T包括在存储器中的存储器内数据结构中执行查找，其中，所述存储器位于可操作地连接到所述持久性存储装置的控制模块中，其中，所述存储器内数据结构包括用于存储在所述持久性存储装置中的所有数据的产生时间和多个条目，其中，所述多个条目中的每一个包括多个T中的一个，多个编程/擦除循环值中的一个和多个保持时间中的一个；

进行所述位错误值大于T的第一确定；以及

基于所述第一确定：

识别m-页，其中，所述m-页是一组页，其中，所述页在所述一组页中；

将所述m-页设定为对未来操作不可分配，并且

其中，在将所述m-页设定为不可分配之后，所述页保持可读。

通过预测损坏的m-页提高闪存利用率的方法和系统

技术领域

[0001] 本文公开的实施方式总体上涉及用于提高闪存利用率的方法和系统。更具体地，本文公开的实施方式涉及一种用于管理持久性存储器的方法以及与该方法相关联的一种系统和一种非临时计算机可读介质。

背景技术

[0002] 用于存储系统的一个重要性能度量是与检索存储在存储系统中的数据有关的延迟。存储系统的性能随着读取延迟的减小而改善。如果存储系统能够可靠地从存储介质中检索无错误数据，则可以减小用于存储系统的读取延迟。当未检索到无错误数据时，存储系统可执行附加动作以便从检索数据去除错误。例如，存储系统可使用诸如纠错码(ECC)和/或RAID之类的纠错机制来从检索数据去除错误或者另外生成无错误数据。纠错机制的使用导致读取延迟的增加，其伴随有相应的性能下降。

发明内容

[0003] 总体上，在一方面，本发明涉及一种用于管理持久性存储器的方法，该方法包括：选择用于主动读请求的页，其中，该页位于持久性存储器中；向该页发布主动读请求；响应于该主动读请求而接收用于存储在该页上的数据的位错误值(BEV)；获得用于该页的BEV阈值(T)，其中，使用与该页相关联的编程/擦除循环值和存储在该页上的数据的保持时间来确定T；进行BEV大于T的第一确定；基于该第一确定：识别m-页，其中，m-页是一组页，其中，该页在该一组页中；将该m-页设定为对未来操作不可分配。

[0004] 总体上，在一方面，本发明涉及一种系统，该系统包括：包括存储模块控制器和持久性存储器的存储模块；以及被操作连接到存储模块和客户端的控制模块，其中，该控制模块执行至少以下步骤：选择用于主动读请求的页，其中，该页位于持久性存储器中；向该页发布主动读请求；响应于该主动读请求而接收用于存储在该页上的数据的位错误值(BEV)；获得用于该页的BEV阈值(T)，其中，使用与该页相关联的编程/擦除循环值和存储在该页上的数据的保持时间来确定T；进行BEV大于T的第一确定；基于该第一确定：识别m-页，其中，m-页是一组页，其中，该页在该一组页中；以及将该m-页设定为对未来操作不可分配。

[0005] 总体上，在一方面，本发明涉及一种包括计算机可读程序代码的非临时计算机可读介质，该计算机可读程序代码在被计算机处理器执行时使得计算机处理器能够：选择用于主动读请求的页，其中，该页位于持久性存储器中；向该页发布主动读请求；响应于该主动读请求而接收用于存储在该页上的数据的位错误值(BEV)；获得用于该页的BEV阈值(T)，其中，使用与该页相关联的编程/擦除循环值和存储在该页上的数据的保持时间来确定T；进行BEV大于T的第一确定；基于该第一确定：识别m-页，其中，m-页是一组页，其中，该页在该一组页中；以及将该m-页设定为对未来操作不可分配。

[0006] 根据以下描述和所附权利要求，本发明的其它方面将是显而易见的。

附图说明

[0007] 图1A—1C示出了根据本发明的一个或多个实施方式的系统。

[0008] 图2示出了根据本发明的一个或多个实施方式的存储设备。

[0009] 图3示出了根据本发明的一个或多个实施方式的存储模块。

[0010] 图4示出了根据本发明的一个或多个实施方式的各种部件之间的关系。

[0011] 图5示出了根据本发明的一个或多个实施方式的用于处理主动读请求的方法。

[0012] 图6示出了根据本发明的一个或多个实施方式的用于处理主动读请求的方法。

[0013] 图7A—D示出了根据本发明的一个或多个实施方式的示例。

具体实施方式

[0014] 现在将参考附图来详细地描述本发明的特定实施方式。在本发明的实施方式的以下详细描述中，阐述了许多特定细节以便提供本发明的更透彻理解。然而，对于本领域的技术人员而言将显而易见的是可在没有这些特定细节的情况下实施本发明。在其它情况下，并未详细地描述众所周知的特征以避免不必要地使本描述复杂化。

[0015] 在图1—7D的以下描述中，在本发明的各种实施方式中，相对于附图描述的任何部件可等价于相对于任何其它图描述的一个或多个类似名称的部件。为了简便起见，将不会对每个图的这些部件重复描述。因此，通过引用而结合每个图的部件的每个实施方式并假定为可选地存在于具有一个或多个类似名称的部件的每个图内。另外，根据本发明的各种实施方式，图的部件的任何描述将被解释为除相对于任何其它图中的相应类似名称部件所述的实施方式之外、与之相结合或作为其替代可实现的可选实施方式。

[0016] 一般地，本发明的实施方式涉及通过主动地识别固态存储器中的页来增加固态储存器的利用率，对于该页而言，存在先前存储的数据在随后被请求时将不可检索的高可能性。换言之，本发明的实施方式主动地识别将来可能出故障的页，其中，此类故障很可能触发诸如RAID(廉价磁盘冗余陈列)方案之类的纠错机制的执行。本发明的实施方式基于用于主动读取页上的数据的位错误值(BEV)是否大于阈值(T)来主动地识别很可能出故障的页，其中，T取决于页的P/E循环值(p)和主动读取页上的数据的保持时间(t)。主动地识别在将来具有高故障可能性的页的能力导致限制此类页存储数据，并且因此限制了调用纠错机制以读取已存储数据的需要。由于RAID(或其它纠错机制)被调用的可能性降低，所以系统的性能增加(即，存在用于对读请求提供服务的较低读延迟)。

[0017] 以下对用于实现本发明的一个或多个实施方式的一个或多个系统和方法进行描述。

[0018] 图1A—1C示出了根据本发明的一个或多个实施方式的系统。参考图1A，该系统包括被操作连接到存储设备102的一个或多个客户端(客户端A 100A、客户端M 100M)。

[0019] 在本发明的一个实施方式中，客户端100A、100M对应于包括向存储设备102发布读请求和/或向存储设备102发布写请求的功能的任何物理系统。虽然在图1A中未示出，但客户端100A、100M中的每一个可包括客户端处理器(未示出)、客户端存储器(未示出)以及实现本发明的一个或多个实施方式所需的任何其它软件和/或硬件。

[0020] 在本发明的一个实施方式中，客户端100A—100M被配置为执行包括文件系统的操

作系统(OS)。文件系统提供了用于从存储设备102进行文件的存储和检索的机制。更具体地,文件系统包括执行所需动作以向存储设备发布读请求和写请求的功能。文件系统还提供了编程界面以使得能够创建和删除文件、读和写文件、在文件内执行查找、创建和删除目录、管理目录内容等。另外,文件系统还提供了管理界面以创建和删除文件系统。在本发明的一个实施方式中,为了访问文件,操作系统(经由文件系统)通常提供文件操作界面以打开、关闭、读和写每个文件内的数据和/或操作相应的元数据。

[0021] 继续图1A的讨论,在本发明的一个实施方式中,客户端100A、100M被配置为使用以下协议中的一个或多个与存储设备102通信:外围部件互连(PCI)、快速PCI(PCIe)、扩展PCI(PCI-X)、快速非易失性存储器(NVMe)、快速PCI结构上的快速非易失性存储器(NVMe)、以太网结构上的快速非易失性存储器(NVMe)以及无限带宽结构上的快速非易失性存储器(NVMe)。本领域的技术人员将认识到本发明不限于上述协议。

[0022] 在本发明的一个实施方式中,存储设备102是一种包括易失性和持久性存储器且被配置成为来自一个或多个客户端100A、100M的读请求和/或写请求提供服务的系统。下面在图2中描述存储设备102的各种实施方式。

[0023] 参考图1B,图1B示出了其中将客户端100A、100M连接到以网状配置(在图1B中表示为存储设备网格104)布置的多个存储设备104A、104B、104C、104D的系统。如图1B中所示,以完全连接的网状配置示出了存储设备网格104—亦即,存储设备网格104中的每个存储设备104A、104B、104C、104D被直接地连接到存储设备网格104中的每个其它存储设备104A、104B、104C、104D。在本发明的一个实施方式中,可将客户端100A、100M中的每一个直接地连接到存储设备网格104中的一个或多个存储设备104A、104B、104C、104D。本领域的技术人员将认识到在不脱离本发明的情况下可使用其它网状配置(例如,部分连接网格)来实现存储设备网格。

[0024] 参考图1C,图1C示出了其中将客户端100A、100M连接到以扇出配置布置的多个存储设备104A、104B、104C、104D的系统。在此配置中,每个客户端100A、100M被连接到存储设备104A、104B、104C、104D中的一个或多个;然而,在单独存储设备104A、104B、104C、104D之间不存在通信。

[0025] 本领域的技术人员将认识到虽然图1A—1C示出了被连接到有限数目的客户端的存储设备,但在不脱离本发明的情况下可将存储设备连接到任何数目的客户端。本领域的技术人员将认识到虽然图1A—1C示出了各种系统配置,但本发明不限于上述系统配置。此外,本领域的技术人员将认识到在不脱离本发明的情况下可使用任何其它物理连接将客户端(无论系统的配置如何)连接到(一个或多个)存储设备。

[0026] 图2示出了根据本发明的一个或多个实施方式的存储设备的实施方式。该存储设备包括控制模块200和存储模块组202。下面描述这些部件中的每一个。一般地,控制模块200被配置成管理来自一个或多个客户端的读和写请求的服务。特别地,控制模块被配置成经由IOM(下面讨论)从一个或多个客户端接收请求、处理请求(其可包括向存储模块发送请求)、并在该请求已被提供服务之后向客户端提供响应。另外,控制模块200包括生成并发布主动读请求且还执行各种垃圾收集操作的功能。下面参考图5和6来描述控制模块关于为读请求提供服务的操作。下面包括关于控制模块中的部件的附加细节。

[0027] 继续图2的讨论,在本发明的一个实施方式中,控制模块200包括输入/输出模块

(IOM)204、处理器208、存储器210以及可选地现场可编程门阵列(FPGA)212。在本发明的一个实施方式中，IOM 204是客户端(例如，图1A—1C中的100A、100M)与存储设备中的其它部件之间的物理接口。IOM支持以下协议中的一个或多个：PCI、PCIe、PCI-X、以太网(包括但不限于在IEEE802.3a-802.3bj下限定的多个标准)、无限带宽(Infiniband)以及融合以太网上的远程直接存储器访问(RDMA)(RoCE)。本领域的技术人员将认识到在不脱离本发明的情况下可使用除上文所列的那些之外的协议来实现IOM。

[0028] 继续图2，处理器208是具有被配置成执行指令的单核或多核的一组电子电路。在本发明的一个实施方式中，可使用复杂指令集(CISC)架构或精简指令集(RISC)架构来实现处理器208。在本发明的一个或多个实施方式中，处理器208包括根复合体(由PCIe协议定义)。在本发明的一个实施方式中，如果控制模块200包括根复合体(可将其集成到处理器208中)，则存储器210经由根复合体而连接到处理器208。替换地，存储器210使用另一点到点连接机制而直接地连接到处理器208。在本发明的一个实施方式中，存储器210对应于任何易失性存储器，包括但不限于动态随机存取存储器(DRAM)、同步DRAM、SDR SDRAM以及DDR SDRAM。

[0029] 在本发明的一个实施方式中，处理器208被配置成创建并更新存储器内数据结构(未示出)，其中，该存储器内数据结构被存储在存储器210中。在本发明的一个实施方式中，该存储器内数据结构包括在图4中描述的信息。

[0030] 在本发明的一个实施方式中，处理器被配置成将各种类型的处理卸载到FPGA 212。在本发明的一个实施方式中，FPGA 212包括计算用于被写入到(一个或多个)存储模块的数据和/或从(一个或多个)存储模块读取的数据的校验和的功能。此外，FPGA 212可包括出于使用RAID方案(例如，RAID 2—RAID 6)在(一个或多个)存储模块中存储数据的目的而计算P和/或Q奇偶信息的功能和/或执行恢复使用RAID方案(例如，RAID 2—RAID 6)存储的已损坏数据所需的各种计算的功能。在本发明的一个实施方式中，存储模块组202包括每个被配置成存储数据的一个或多个存储模块214A、214N。下面在图3中描述存储模块的一个实施方式。

[0031] 图3示出了根据本发明的一个或多个实施方式的存储模块。存储模块300包括存储模块控制器302、存储器(未示出)以及一个或多个固态存储器模块304A、304N。下面描述这些部件中的每一个。

[0032] 在本发明的一个实施方式中，存储模块控制器300被配置成接收从一个或多个控制模块读取数据和/或向其写入数据的请求。此外，存储模块控制器300被配置成使用存储器(未示出)和/或固态存储器模块304A、304N来服务读和写请求。

[0033] 在本发明的一个实施方式中，存储器(未示出)对应于任何易失性存储器，包括但不限于动态随机存取存储器(DRAM)、同步DRAM、SDR SDRAM以及DDR SDRAM。

[0034] 在本发明的一个实施方式中，固态存储器模块对应于使用固态存储器来存储持久性数据的任何数据存储器件。在本发明的一个实施方式中，固态存储器可包括但不限于NAND闪存和NOR闪存。此外，NAND闪存和NOR闪存可包括单层单元(SLC)、多层单元(MLC)或三层单元(TLC)。本领域的技术人员将认识到本发明的实施方式不限于存储类存储器。

[0035] 图4示出了根据本发明的一个或多个实施方式的各种部件之间的关系。更具体地，图4示出了存储在控制模块的存储器中的各种类型的信息。此外，控制模块包括更新存储在

控制模块的存储器中的信息的功能。可将下面描述的信息存储在一个或多个存储器内数据结构中。此外，如果(一个或多个)数据结构类型保持信息之间的关系(如下所述)，可使用任何数据结构类型(例如，阵列、链表、散列表等)来组织(一个或多个)存储器内数据结构内的以下信息。

[0036] 存储器包括逻辑地址400到物理地址402的映射。在本发明的一个实施方式中，逻辑地址400是从客户端(例如，图1A中的100A、100M)的角度看数据看起来常驻在该处的地址。换言之，逻辑地址400对应于当向存储设备发布读请求时被客户端上的文件系统使用的地址。

[0037] 在本发明的一个实施方式中，逻辑地址是(或包括)通过把散列函数(例如，SHA-1、MD-5等)应用到n元组而生成的散列值，其中，n元组是＜对象ID，偏移ID＞。在本发明的一个实施方式中，对象ID定义文件，并且偏移ID定义相对于文件的起始地址的位置。在本发明的另一实施方式中，n元组是＜对象ID，偏移ID，产生时间＞，其中，产生时间对应于创建该文件(使用对象ID来识别)时的时间。替换地，逻辑地址可包括逻辑对象ID和逻辑字节地址或者逻辑对象ID和逻辑地址偏移。在本发明的另一实施方式中，逻辑地址包括对象ID和偏移ID。本领域的技术人员将认识到可将多个逻辑地址映射到单个物理地址，并且逻辑地址内容和/或格式不限于上述实施方式。

[0038] 在本发明的一个实施方式中，物理地址402对应于图3中的固态存储器模块304A、304N中的物理位置。在本发明的一个实施方式中，可将物理地址定义为以下n元组：＜存储模块，通道，芯片使能，LUN，平面，块，页号，字节＞。

[0039] 在本发明的一个实施方式中，每个物理地址402与编程/擦除(P/E)循环值404相关联。P/E循环值可表示：(i)已经在由物理地址定义的物理位置上执行的P/E循环的数目，或(ii)P/E循环范围(例如，5,000—9,999次P/E循环)，其中，在由物理地址定义的物理位置上执行的P/E循环的数目在P/E循环范围内。在本发明的一个实施方式中，P/E循环是数据到擦除块(即，用于擦除操作的最小可寻址单元，通常为一组的多个页)中的一个或多个页的写入和该块的擦除，任一顺序均可。

[0040] 可基于每个页、基于每个块、基于每组块和/或以任何其它水平的粒度来存储P/E循环值。控制模块包括在数据被写入到固态存储模块(和/或从其擦除)时适当地更新P/E循环值402的功能。

[0041] 在本发明的一个实施方式中，所有数据(即，客户端上的文件系统已经请求被写入到固态存储模块的数据)406与产生时间408相关联。产生时间408可对应于：(i)数据被写入到固态存储模块中的物理位置的时间；(ii)客户端发布用以将数据写入到固态存储模块的写请求的时间；或者(iii)对应于(i)或(ii)中的写事件的无单位值(即，序号)。

[0042] 在本发明的一个实施方式中，存储器内数据结构包括用于已被作为读请求的一部分或者作为主动读请求的一部分读取的任何页的至少一个位错误值(BEV)。该BEV指定不正确的检索数据(即，响应于读请求或主动读请求而从页读取的数据)中的位数。可替换地将BEV表示为不正确的给定页中的位的百分比。可使用纠错码(ECC)来确定用于给定页的BEV，其中，还将用于存储在给定页上的数据的ECC存储在该页上。换言之，页可包括数据和用于数据的ECC。用于页的BEV可由存储模块控制器(参见图3,302)确定。控制模块中的存储器(图2,210)可存储从给定页获得的最后BEV和/或可存储用于给定页的多个BEV。

[0043] 可使用存储的BEV值作为图5中的步骤512和图6中的步骤618中的确定的一部分。下面在图5和6中描述关于存储的BEV的使用的附加细节。

[0044] 在本发明的一个实施方式中，存储器内数据结构包括＜保持时间，P/E循环值＞到BEV阈值416的映射。在本发明的一个实施方式中，保持时间对应于在数据到固态存储模块的写入与正在从固态存储模块读取数据的时间之间所经历的时间。可以用时间单位(秒、天、月等)来表示保持时间或者可表示为无单位值(例如，当将产生时间表示为无单位值时)。在本发明的一个实施方式中，可将＜保持时间，P/E循环值＞中的P/E循环值表示为P/E循环或P/E循环范围。

[0045] 在本发明的一个实施方式中，通过执行实验以确定针对保持时间和P/E循环值的给定组合预测在时间t+1的页的故障的在时间t的BEV来确定BEV阈值416。优化BEV阈值416以便能够在不必将持久性存储器中的m-页标记为不可分配的同时成功地从固态存储器模块读取数据。

[0046] 通过基于保持时间和P/E循环值来修改(一个或多个)BEV阈值，存储设备将可改变在给定保持时间和P/E循环值下的给定页的故障可能性的各种变量考虑在内。通过基于上述变量来理解页如何随时间推移而出故障，可使用适当的BEV阈值以主动地确定给定页在将来是否可能出故障。

[0047] 在本发明的一个实施方式中，可如下在实验上确定用于给定＜保持时间(t)，P/E循环值(p)＞的BEV阈值416：(i)针对P/E循环值(p)在时间t+1(例如，处于保持时间两个月)确定用于一组页的BEV；(ii)识别将触发RAID(或另一纠错机制)的使用的所有页(即，对于其而言可不使用ECC来修正检索数据中的错误的所有页)；(iii)针对P/E循环值(p)(即，与在(i)中使用的相同P/E循环值)在时间t(例如，处于保持时间一个月)确定用于在(ii)中识别的所有页的BEV；(iv)通过减小在(ii)中识别的页的数目直至在时间t+1达到纠错机制激活极限(例如，仅1％的读取应在t+1触发纠错机制的使用)为止来识别BEV阈值(T(p,t))。

[0048] 更具体地，在(iv)中，从具有最高BEV的(ii)中的页开始，连续地从在(ii)中识别的该组页中去除在(ii)中识别的页。针对从(ii)去除的每个页，还去除已去除页是其一部分的m-页中的其它页(例如，如果从(ii)中去除了页A且m-页具有四个页，则也去除作为页A所属的m-页的一部分的其它三个页)。在时间t去除上述页的结果导致这些页在时间t+1不存储任何数据，并且因此这些页不能在t+1触发纠错机制的激活。在图7A—7C中描述了确定BEV阈值的示例。

[0049] 在本发明的一个实施方式中，控制模块(图2，200)使用上述信息(参见图4)来执行以下各项中的一个或多个：(i)服务客户端读请求；(ii)服务写请求；(iii)服务主动读请求；以及(iv)垃圾收集操作。控制模块可并行地执行以下请求和/或操作中的一个或多个。

[0050] 在本发明的一个实施方式中，由客户端(例如，图1A，100A)发布客户端读请求，其中，该读请求包括逻辑地址。对该读请求的响应是：(i)从持久性存储器检索的数据以及可选地用于从持久性存储器检索的数据的BEV或者(ii)指示数据已损坏的通知(即，不能从持久性存储器检索数据和/或不能使用由存储模块控制器和/或控制模块实现的纠错机制来修正或重构数据)。在以下图6中包括关于为读请求提供服务的附加细节。

[0051] 在本发明的一个实施方式中，由控制模块(例如，图2，200)发布主动读请求，其中，该主动读请求包括物理地址。对主动读请求的响应是：用于从持久性存储器检索的数据的

BEV以及可选地从持久性存储器检索的数据。在以下图5中包括关于服务读请求的附加细节。

[0052] 在本发明的一个实施方式中，由客户端(例如，图1A，100A)发布写请求，其中，该写请求包括要存储在持久性存储器中的数据或对该数据的引用。在接收到写请求时，控制模块确定要用来将数据存储在持久性存储器中的一个或多个页。从持久性存储器中的所述一组可分配页中识别被选择用来存储数据的页，其中，控制模块保持可分配页的列表。当可作为写请求的一部分或者作为垃圾收集操作(下面描述)的一部分将数据写入到页时，认为页是可分配页(下面在图5和6中讨论)。

[0053] 在本发明的一个实施方式中，由控制模块实现作为垃圾收集过程的一部分执行的垃圾收集操作。垃圾收集过程的目的是回收死页(即，不再包括活动数据(即，正在被控制模块和/或在客户端上执行的一个或多个应用程序使用的数据)的页)。这可通过以下各项来实现：(i)识别包括活动页和死页的组合的持久性存储器中的块；以及(ii)将活动数据移动至仅包括活动页的持久性存储器中的另外(一个或多个)块中的一个或多个页。可仅将被作为垃圾收集操作的一部分重写到(一个或多个)新页的数据写入到作为可分配页(下面在图5和6中讨论)的(一个或多个)页。

[0054] 转到流程图，虽然连续地提出并描述流程图中的各步骤，但本领域的技术人员将认识到可按照不同的顺序执行某些或所有步骤，可将其组合或省略，并且可并行地执行某些或所有步骤。

[0055] 图5示出了根据本发明的一个或多个实施方式的用于由存储设备处理客户端读请求的方法。

[0056] 在步骤500中，针对主动读请求选择持久性存储器中的页。该页可以是持久性存储器中的任何活动页(即，包括活动数据的任何页)。可由控制模块保持/管理活动页的列表并将其存储在控制模块内的存储器中。

[0057] 在步骤502中，由控制模块向存储模块发布主动读请求，其中，该存储模块是包括(在步骤500中选择的)页位于其上面的固态存储器模块的存储模块。读请求的格式可以是存储模块控制器所支持的任何格式。主动读请求可包括(在步骤500中选择的)页的物理地址以及标志(或其它内容)，其指示请求是主动读请求，而不是例如客户端读请求。

[0058] 在步骤504中，从存储模块接收包括至少用于从页(即，在步骤500中选择的页)读取的数据的BEV的响应。

[0059] 在步骤506中，进行关于对于给定保持时间和P/E循环值而言BEV是否大于BEV阈值(T(t,p))的确定。在本发明的一个实施方式中，针对存储在物理地址处的数据确定保持时间(t)。可使用数据的产生时间(参见图4，408)和主动读请求的时间(例如，控制模块发布主动读请求的时间)来确定保持时间。从控制模块的存储器(参见图2，210)获得数据的产生时间。通过确定主动读请求的时间与产生时间之间的差来计算保持时间。在本发明的一个实施方式中，可通过使用页的物理地址作为密钥在存储器内数据结构(位于控制模块的存储器中)中执行查找来确定P/E循环值。查询的结果可以是与物理地址相关联的实际P/E循环值(例如，与对应于物理地址的物理位置位于其中的块相关联的P/E循环值)或者可以是P/E循环值范围(例如，5,000-9,999次P/E循环)，其中，与物理地址相关联的实际P/E循环值位于P/E循环值范围内。使用以下密钥＜保持时间，P/E循环值＞从存储器内数据结构(参见图

4)获得BEV阈值(T(t,p))。

[0060] 如果针对给定保持时间和P/E循环值BEV小于BEV阈值(T(t,p)),则该过程结束;否则该过程前进至步骤508。

[0061] 在步骤508中,将包括(在步骤500中选择的)所述页的m-页标记为不可分配。更具体地,将作为与(在步骤500中选择的)所述页相同的m-页的一部分的每个页标记为不可分配。一旦将页标记为不可分配,则不使用该页来存储任何未来活动数据作为写操作或垃圾收集操作的一部分。在本发明的一个实施方式中,m-页是一个或多个页。可在单个原子处理中将m-页中的页写入到持久性存储器中。例如,m-页可以是在单个原子事务中被写入到持久性存储器的四个页。如果数据到m-页中的单个页的写入失败,则整个事务(即,数据到构成m-页的四个页的写入)失败。

[0062] 继续图5的讨论,在步骤510中,进行关于(在步骤500中选择的)页中的数据是否是ECC可修正(即,存储模块控制器是否可以仅使用页的ECC来修正数据中的错误)的确定。如果页中的数据是ECC可修正的,则过程前进至步骤514;否则,过程前进至步骤512。

[0063] 在步骤512中,进行关于是否调用RAID方案或另一纠错机制来重构(在步骤500中选择的)页中的已损坏数据作为垃圾收集过程的一部分的确定。是否调用RAID方案或另一纠错机制以重构已损坏数据的确定可基于RAID条带(所述页是其一部分)中的其它页的状态。

[0064] 例如,如果在RAID条带中存在六个页(四个数据页、一个P奇偶页、一个Q奇偶页)且仅一个页已损坏,则可进行不调用RAID方案或另一纠错机制的确定,因为上述RAID条带仍可具有足够的未损坏页以重构RAID条带内的所有数据。换言之,如果在RAID条带中存在至少四个未损坏页,上述RAID条带可能能够重构RAID条带内的所有数据。由于当前在RAID条带中存在五个未损坏页,所以RAID条带中的一个附加页可被损坏而不影响恢复数据的能力。然而,如果上述RAID条带包括两个已损坏页(即,ECC不可修正的页),则进行重构(在步骤500中选择的)该页上的数据以及其它已损坏页中的数据的确定,因为RAID条带中的一个附加已损坏页(即,三个已损坏页)将导致不能重构RAID条带中的任何已损坏数据。

[0065] 在本发明的一个实施方式中,控制模块跟踪RAID条带成员身份(即,哪些页是RAID条带的一部分)和RAID条带几何结构(即,奇偶页的数目、每个奇偶页中的奇偶值的类型(例如,P奇偶值、Q奇偶值等))。控制模块可使用BEV 412来确定给定RAID条带中的哪些页是ECC可修正和ECC不可修正的。

[0066] 本领域的技术人员将认识到在不脱离本发明的情况下可使用其它策略来确定是否调用RAID方案或另一纠错机制。

[0067] 继续图5的讨论,在步骤514中,如果页中的数据是ECC可修正的,则将该页调度用于垃圾收集。将该页(或者该页位于其中的块)调度用于垃圾收集可包括将该页调度作为将作为垃圾收集操作的一部分来处理的下一页(即,从该页读取活动数据并重写(一旦被ECC修正)到持久性存储器中的新的可分配页)。

[0068] 继续步骤514,如果该页上的数据是ECC不可修正的,则必须使用RAID方案来重构该页上的数据。更具体地,将该页调度用于垃圾收集。将该页(或者该页位于其中的块)调度用于垃圾收集可包括将该页调度作为将作为垃圾收集操作的一部分来处理的下一页(即,重构用于该页的数据,并将已重构数据写入到持久性存储器中的新的可分配页)。该页上的

数据的重构可包括从RAID条带中的多个其它页读取数据，并且然后由控制模块执行一个或多个操作以便重构该页上的数据。该过程然后结束。

[0069] 在本发明的一个实施方式中，在持久性存储器中的所有活动页上周期性地执行图5中所示的过程。可使用在控制模块中执行的低优先级线程来实现图5中所示的过程。

[0070] 图6示出了根据本发明的一个或多个实施方式的用于由存储设备处理客户端读请求的方法。

[0071] 在步骤600中，由控制模块从客户端接收客户端读请求，其中，客户端读请求包括逻辑地址。在步骤602中，根据逻辑地址来确定物理地址(其包括页号)。如上文所讨论的，控制模块中的存储器包括逻辑地址到物理地址的映射(参见图4的讨论，400、402)。在本发明的一个实施方式中，通过使用逻辑地址到物理地址的映射以及在步骤600中从客户端请求获得的逻辑地址来执行查找(或查询)而确定物理地址。

[0072] 在步骤604中，使用物理地址生成控制模块读请求。控制模块读请求的格式可以是存储模块控制器所支持的任何格式。

[0073] 在步骤606中，从存储模块接收响应，其包括用于从页(即，在步骤500中选择的页)读取的数据的BEV及(i)来自该页的数据或(ii)数据被损坏的指示(即，页上的数据是ECC不可修正的)。

[0074] 在步骤608中，进行关于在步骤606中接收到的响应是否包括数据的确定。如果在步骤606中接收到的响应包括数据，则过程前进至步骤612；否则，该过程前进至步骤610。

[0075] 在步骤610中，当在步骤606中接收到的响应不包括数据时，控制模块继续使用例如RAID方案或另一纠错机制来重构页上的数据。

[0076] 在步骤612中，向客户端提供数据(或已重构数据)。在步骤614中，进行关于对于给定保持时间和P/E循环值而言BEV是否大于BEV阈值(T(t,p))的确定。在本发明的一个实施方式中，针对存储在物理地址处的数据确定保持时间(t)。如果针对给定的保持时间和P/E循环值BEV小于BEV阈值(T(t,p))，则该过程结束；否则该过程前进至步骤616。

[0077] 在步骤616中，将包括(在步骤500中选择的)所述页的m-页标记为不可分配。更具体地，将作为与(在602中在物理地址中指定的)所述页相同的m-页的一部分的每个页标记为不可分配。

[0078] 在步骤618中，进行关于是否主动地修正数据的确定。如果数据是ECC可修正的，则可根据上述步骤510进行确定。如果数据是ECC不可修正的，则可根据步骤512来进行关于是否主动地进行修正的确定。如果进行主动地修正数据的确定，则过程前进至步骤620；否则，该过程结束。在步骤620中，根据上文在步骤514中的讨论将(一个或多个)页调度为用于垃圾收集。

[0079] 图7A—7D示出了根据本发明的一个或多个实施方式的示例。以下示例并不意图限制本发明的范围。

[0080] 关于图7A—7C，图7A—7C图示出用p的P/E循环值来确定用于保持时间t-1的BEV阈值的一个实施方式，其中，目标是在保持时间t使将修正已损坏数据的RAID的使用局限于小于数据读取的1％。图7A示出了在保持时间t的用于各页的BEV的分布。在不实现本发明的实施方式的情况下，读请求的3.25％将要求使用RAID方案来重构已损坏数据(即，页中的3.25％具有大于50位错误的BEV，在本示例中其为可使用ECC来修正的位错误的最大数目)。

[0081] 图7B示出了在保持时间t-1的用于相同页的BEV的分布。图7C示出了用于在保持时间t被损坏的页(即,在保持时间t具有大于50的BEV的页)的在保持时间t-1的BEV的分布。连续地去除图7C中所示的页(从最高BEV开始连同作为相应m-页的一部分的相关页一起)直至具有大于50的BEV的在保持时间t的页的数目对应于小于所有页的1%为止。在本示例中,被去除的页数对应于在保持时间t-1具有大于43的BEV的页。在本示例中,在时间t的已去除页的百分比是7.69%(即,持久性存储器中的页的7.69%是不可分配的)。

[0082] 参考图7D,图7D示出了实现本发明的一个或多个实施方式的假定性能益处。具体地,针对p的P/E循环值和对读请求的不超过1%调用RAID的要求,本发明的实施方式针对读请求的不超过1%调用RAID,而未实现本发明的实施方式的存储设备以保持时间增量5对大于10%的读请求调用RAID。换言之,在未实现本发明的实施方式的存储设备中更频繁地10倍地调用RAID,从而与实现本发明的一个或多个实施方式的存储设备相比导致用于此类存储设备的高读延迟。

[0083] 可使用由系统中的一个或多个处理器执行的指令来实现本发明的一个或多个实施方式。此外,此类指令可对应于存储在一个或多个非临时计算机可读介质上的计算机可读指令。

[0084] 虽然已针对有限数目的实施方式描述了本发明,但受益于本公开的本领域的技术人员将认识到可以设计不脱离如在这里公开的本发明的范围的其它实施方式。因此,应仅由所附权利要求来限制本发明的范围。

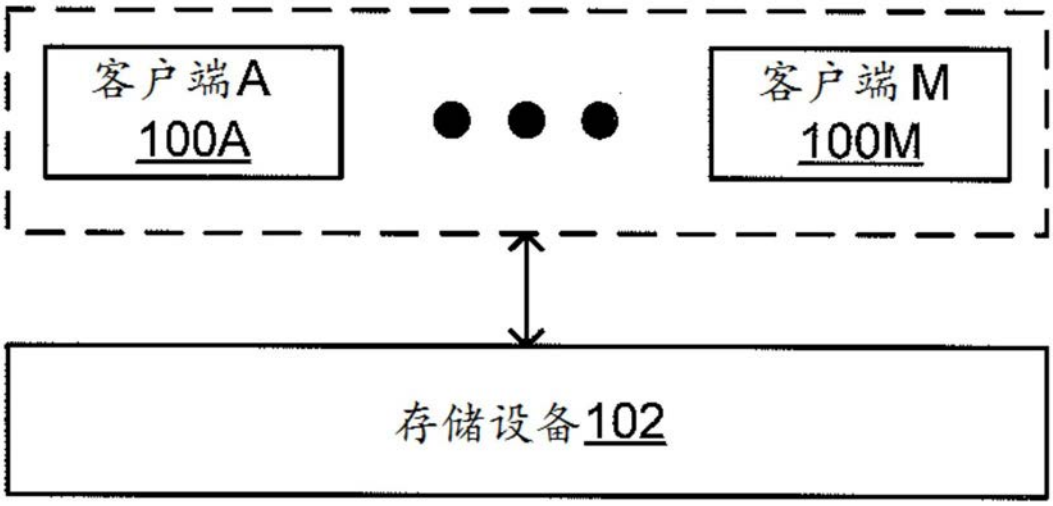


图1A

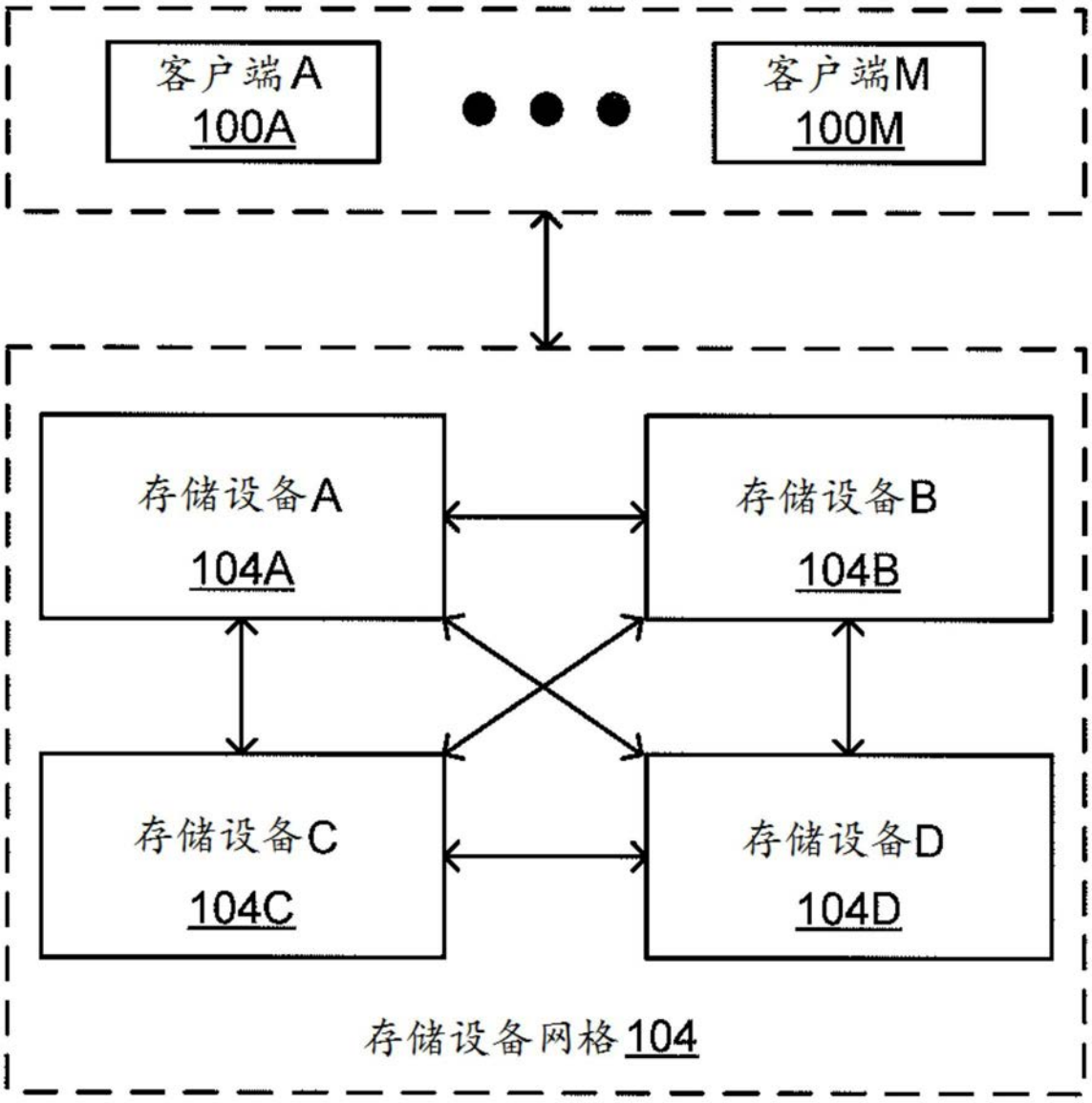

客户端A
100A
客户端M
100M
存储设备A
104A
存储设备B
104B
存储设备C
104C
存储设备D
104D
存储设备网格104

图1B

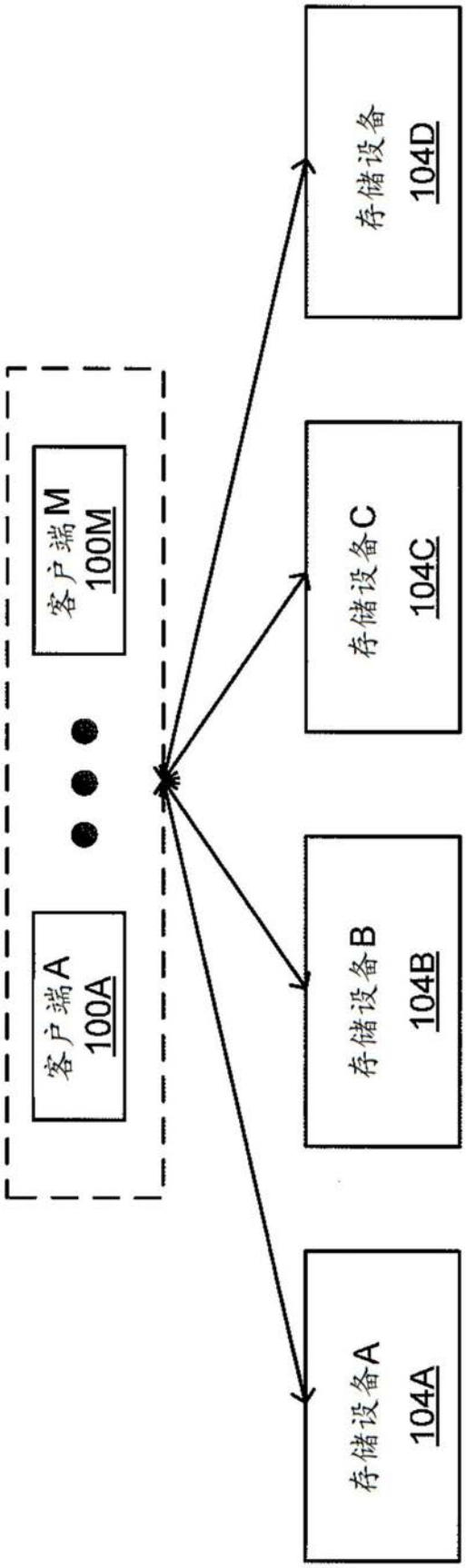

客户端A
100A
客户端M
100M
存储设备A
104A
存储设备B
104B
存储设备C
104C
存储设备
104D

图1C

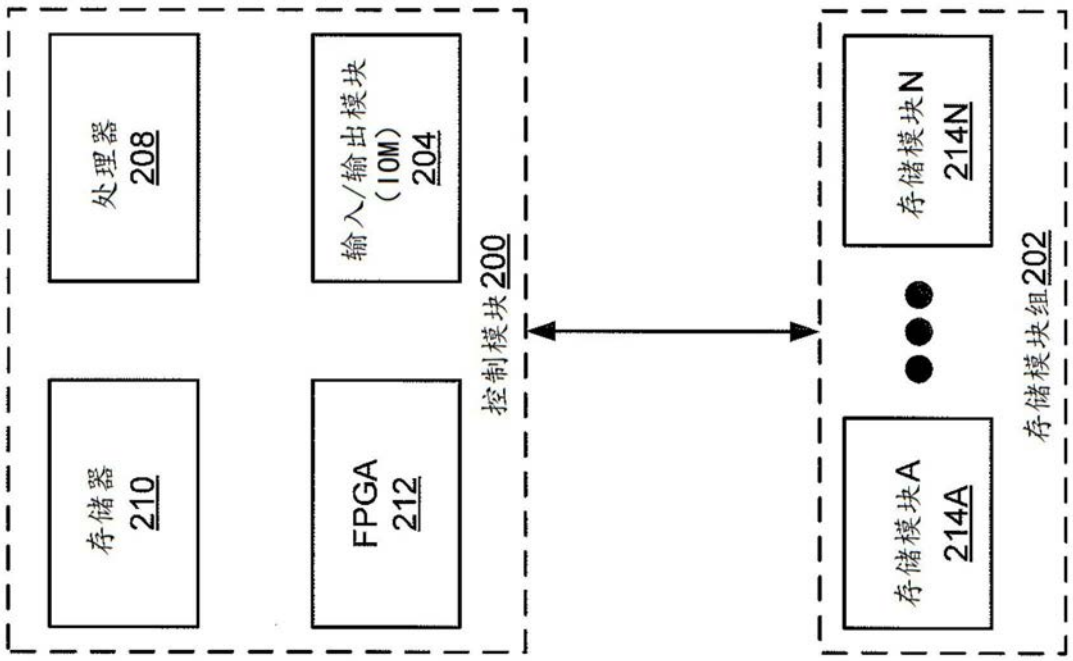

处理器
208
输入/输出模块
(IOM)
204
存储器
210
FPGA
212
控制模块200
存储模块N
214N
存储模块A
214A
存储模块组202

图2

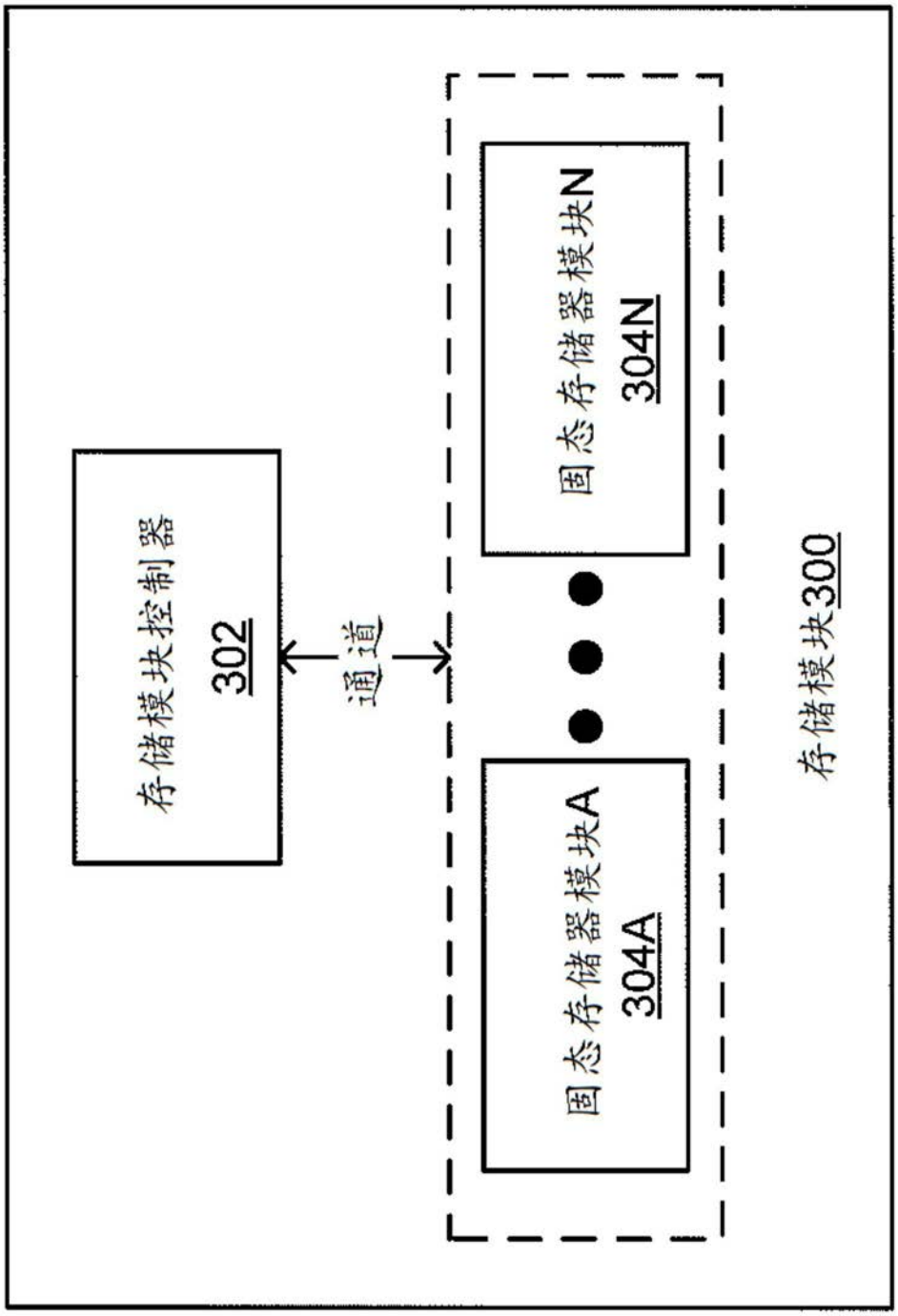


图3

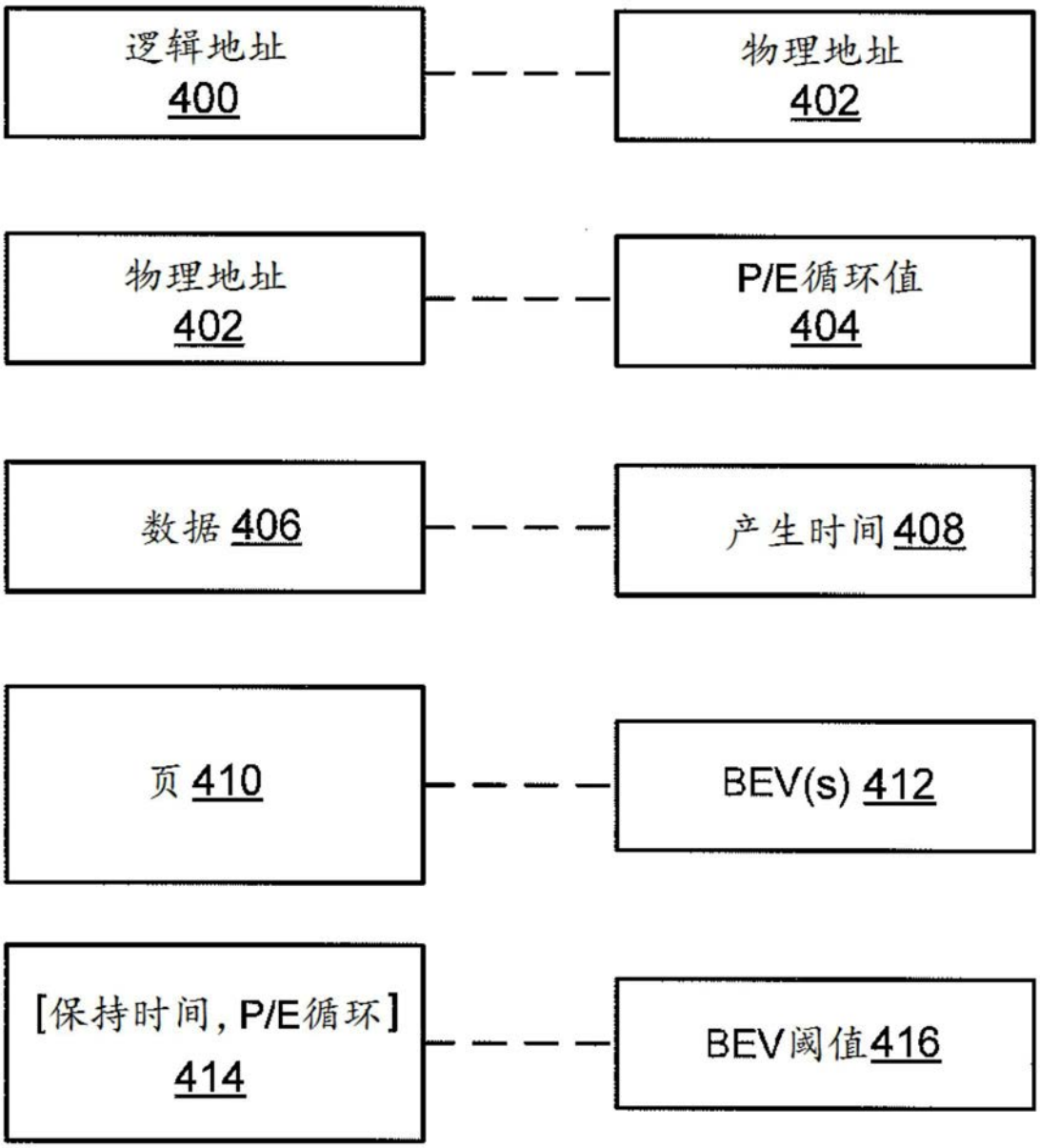

逻辑地址
400
物理地址
402
物理地址
402
P/E循环值
404
数据 406
产生时间 408
页 410
BEV(s) 412
[保持时间, P/E循环]
414
BEV阈值 416

图4

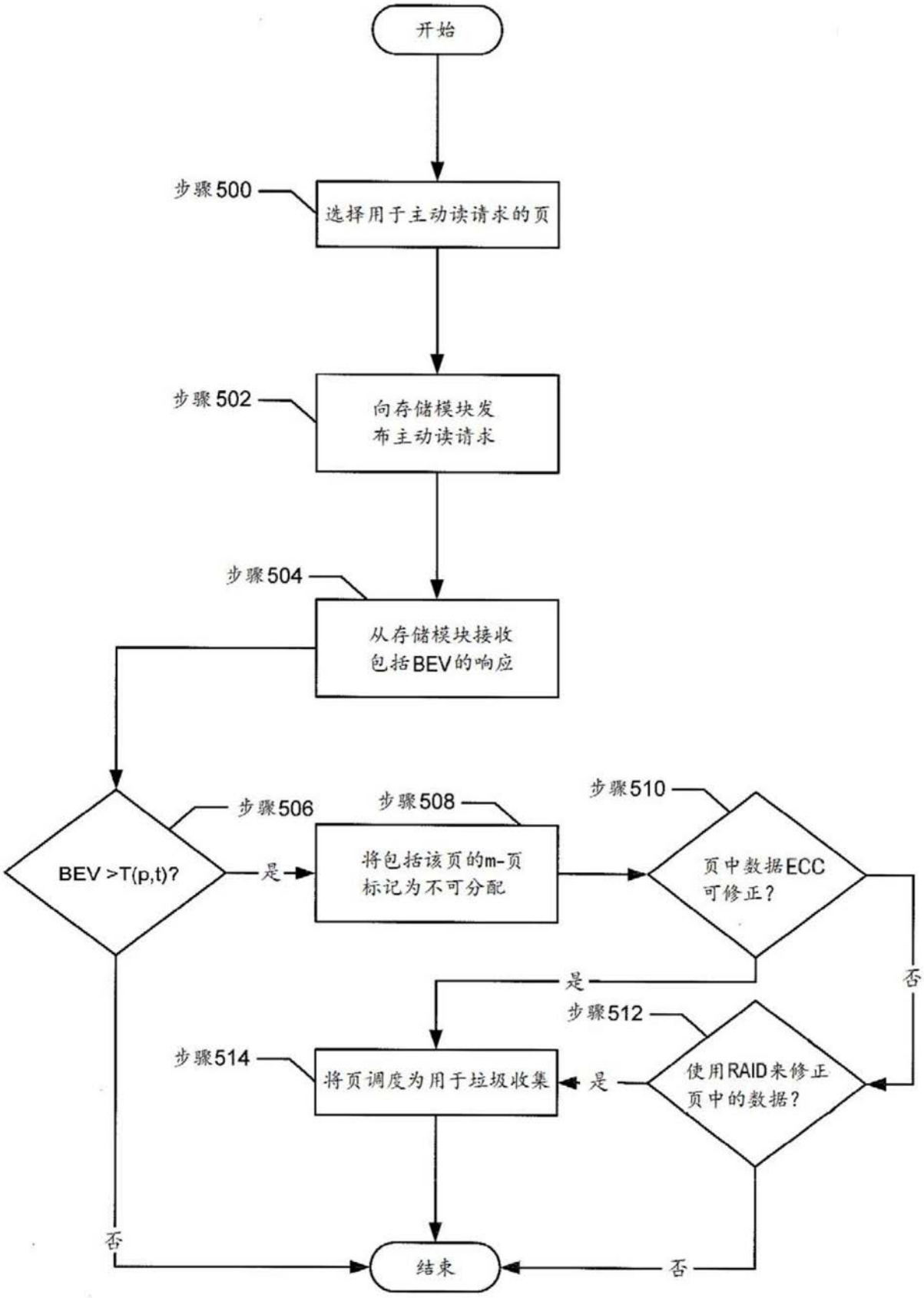

开始
步骤500
选择用于主动读请求的页
步骤502
向存储模块发布主动读请求
步骤504
从存储模块接收包括BEV的响应
步骤506
BEV >T(p,t)?
是
步骤508
将包括该页的m-页标记为不可分配
步骤510
页中数据ECC可修正?
否
是
步骤512
使用RAID来修正页中的数据?
是
步骤514
将页调度为用于垃圾收集
否
否
结束

图5

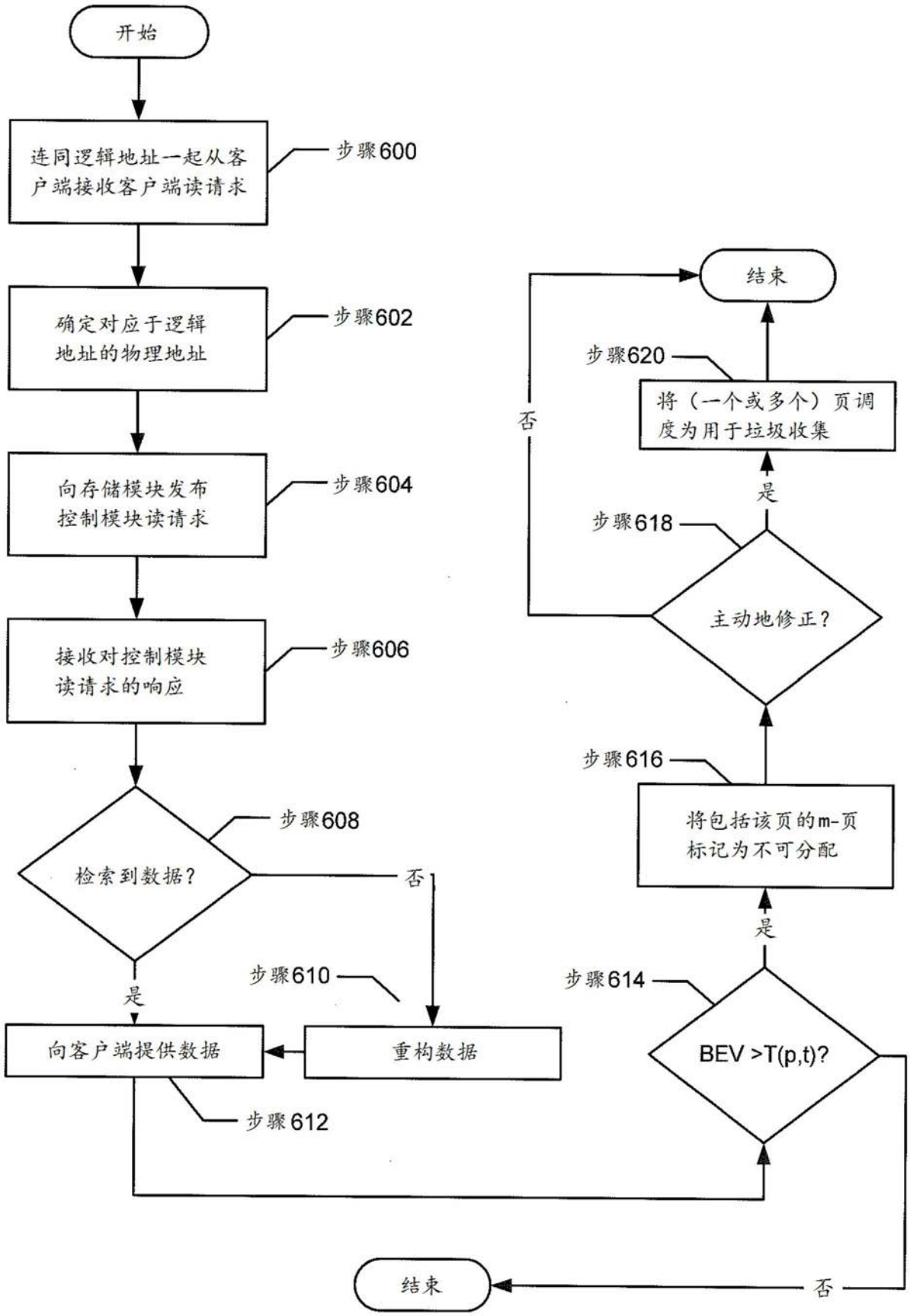

开始
连同逻辑地址一起从客户端接收客户端读请求
步骤600
确定对应于逻辑地址的物理地址
步骤602
向存储模块发布控制模块读请求
步骤604
接收对控制模块读请求的响应
步骤606
检索到数据?
步骤608
否
是
步骤610
向客户端提供数据
重构数据
步骤612
结束
BEV >T(p,t)?
步骤614
否
是
将包括该页的m-页标记为不可分配
步骤616
主动地修正?
步骤618
否
是
将（一个或多个）页调度为用于垃圾收集
步骤620
结束

图6

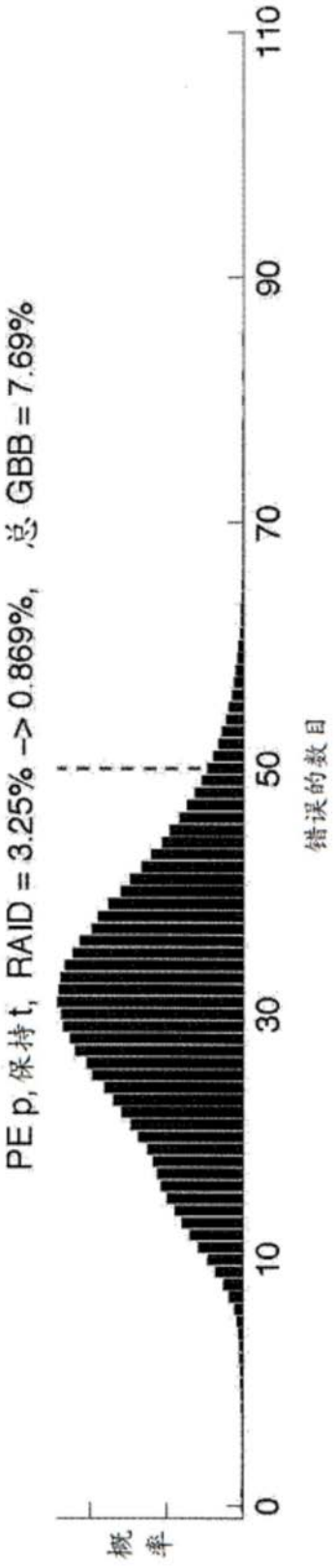


图7A

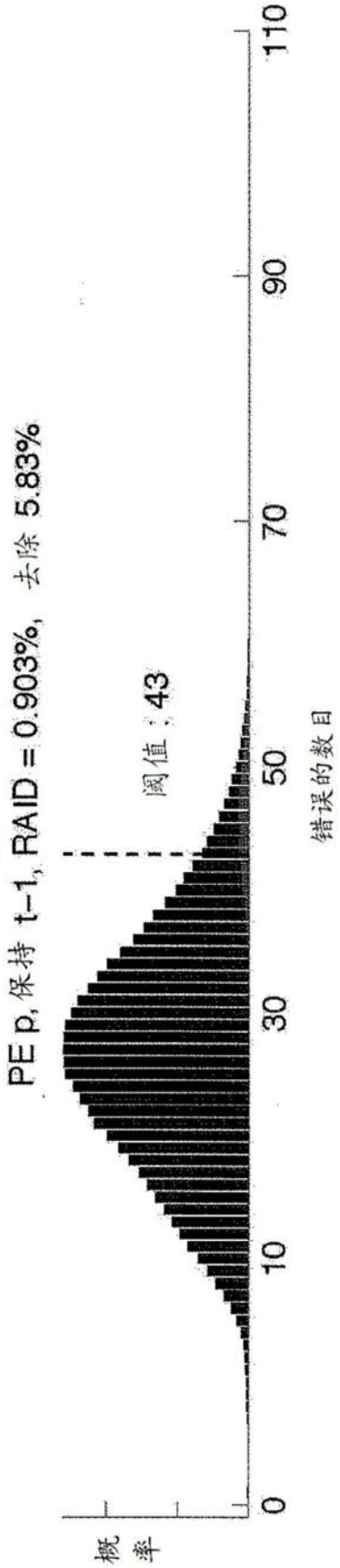


图7B

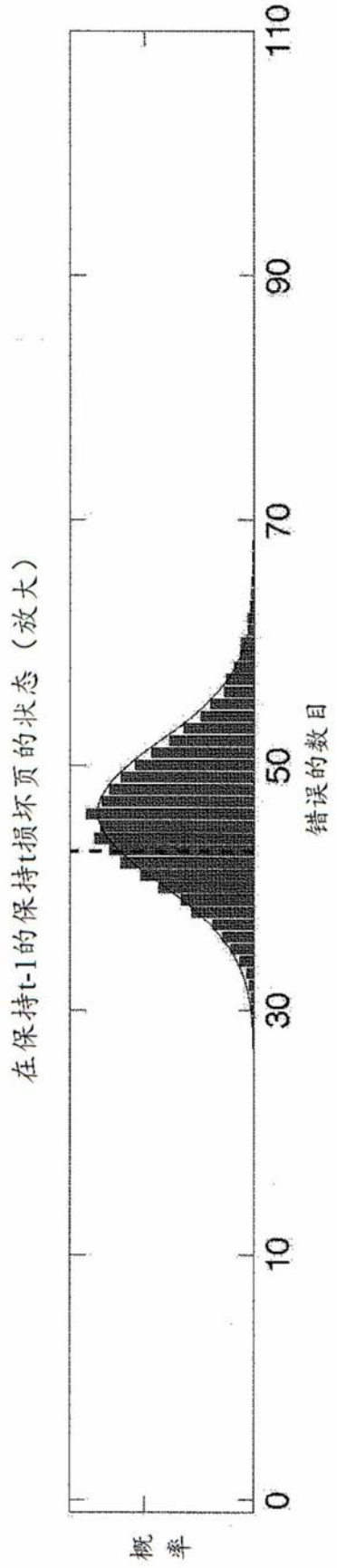


图7C

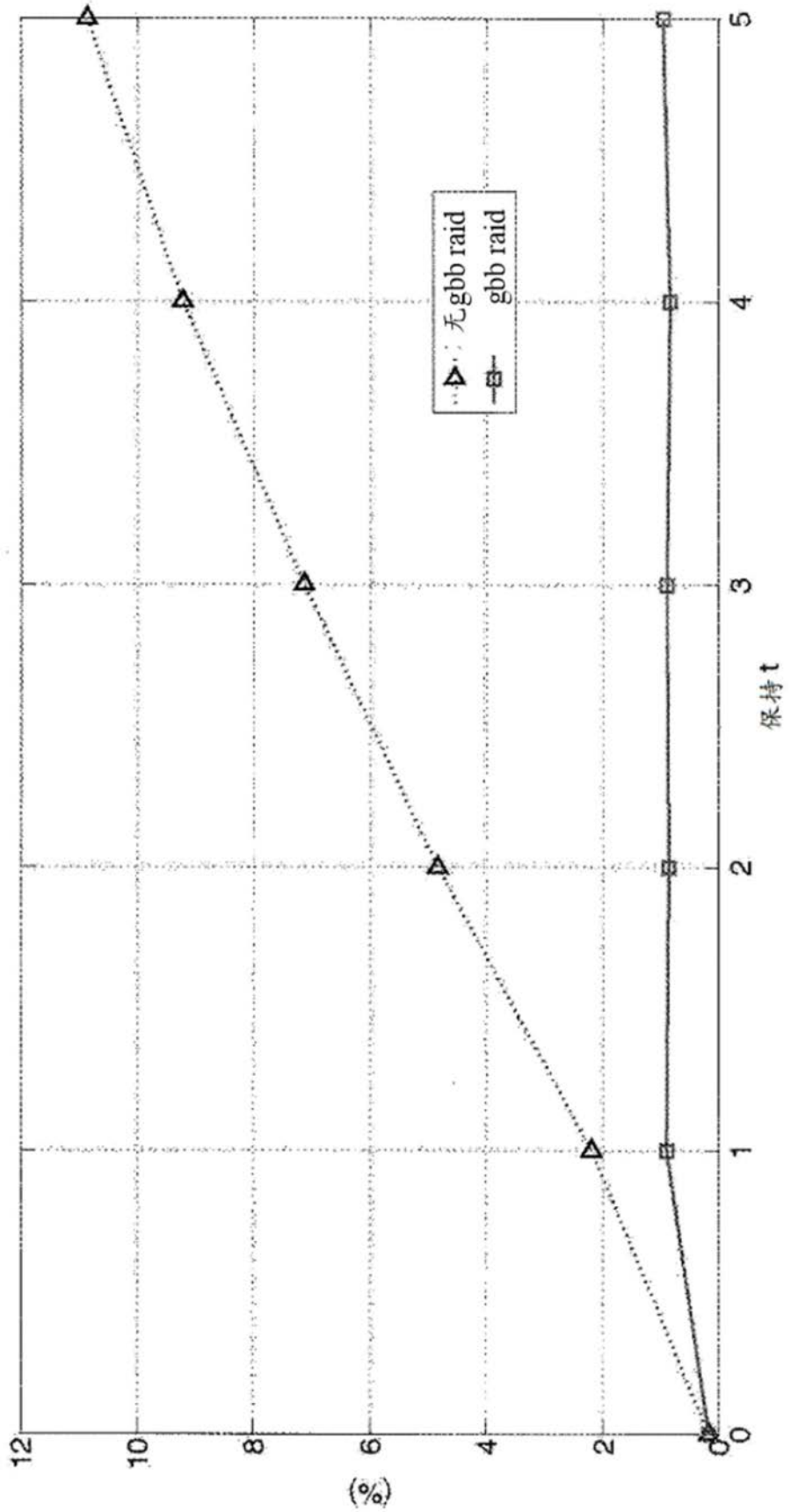


图7D