



(12) 发明专利

(10) 授权公告号 CN 103348326 B

(45) 授权公告日 2016. 03. 30

(21) 申请号 201180053519. 0

(51) Int. Cl.

(22) 申请日 2011. 09. 19

G06F 11/10(2006. 01)

(30) 优先权数据

12/892, 894 2010. 09. 28 US

(56) 对比文件

US 2008155191 A1, 2008. 06. 26,

US 2008155191 A1, 2008. 06. 26,

US 6275898 B1, 2001. 08. 14,

US 2005086429 A1, 2005. 04. 21,

CN 101576833 A, 2009. 11. 11,

(85) PCT国际申请进入国家阶段日

2013. 05. 07

(86) PCT国际申请的申请数据

PCT/US2011/052222 2011. 09. 19

审查员 毕爽君

(87) PCT国际申请的公布数据

W02012/044488 EN 2012. 04. 05

(73) 专利权人 净睿存储股份有限公司

地址 美国加利福尼亚

(72) 发明人 J·科尔格洛夫 J·海斯 洪波

E·米勒

(74) 专利代理机构 中国国际贸易促进委员会专
利商标事务所 11038

代理人 叶勇

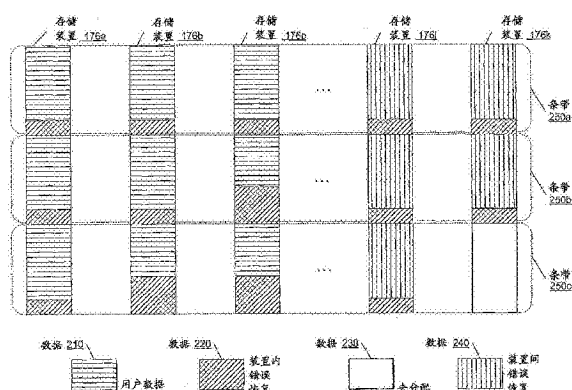
权利要求书3页 说明书19页 附图15页

(54) 发明名称

SSD 环境中的适配 RAID

(57) 摘要

一种用于适配 RAID 几何的系统和方法。计算机系统包括通过网络相互耦合的客户机计算机和数据存储阵列。数据存储阵列利用用于数据存储的固态驱动和快擦写存储器单元。数据存储阵列中的存储控制器被配置为确定用于存储数据的第一 RAID 布局并根据第一 RAID 布局将第一 RAID 条带写入到装置组。响应检测第一条件, 控制器配置为确定与第一 RAID 布局不同的第二 RAID 布局, 并根据第二布局将第二 RAID 条带写入到装置组, 由此, 装置组根据第一 RAID 布局和第二 RAID 布局两者同时存储数据。



1. 一种计算机系统,包括:

被配置为在网络上发送读取和写入请求的客户机计算机;

被配置为接收读取和写入请求的与网络耦合的数据存储子系统,其中,该子系统在包含多个存储装置的装置组上包含多个数据存储位置;

其中,数据存储子系统还包含被配置为完成以下的过程的存储控制器:

确定用于存储数据的第一 RAID 布局和第二 RAID 布局,其中第二 RAID 布局不同于第一 RAID 布局;

根据第一 RAID 布局将第一 RAID 条带写入到所述多个存储装置中的第一子组装置;

根据第二 RAID 布局将第二 RAID 条带写入到所述多个存储装置中的第二子组装置;和

根据第二 RAID 布局将第三 RAID 条带写入到所述多个存储装置中的第三子组装置,其中第三子组装置包括未包含在第二子组中的一个或多个装置;

由此,装置组根据第一 RAID 布局和第二 RAID 布局两者同时存储数据;

其中,响应于检测到第一子组装置中的给定装置的可靠性的变化,所述存储控制器被配置为:

为给定 RAID 条带在给定装置中保持第一量的空间,用于存储装置内冗余数据;以及

同时为所述给定 RAID 条带在不同于所述给定装置的装置中保持第二量的空间,用于存储装置内冗余数据,其中所述第二量不同于所述第一量。

2. 如权利要求 1 所述的计算机系统,其中,存储控制器被配置为基于多个存储装置中的一个或多个的特性确定第一 RAID 布局和第二 RAID 布局。

3. 如权利要求 2 所述的计算机系统,其中,所述特性包含对于多个装置中的一个或多个的访问的数量或速率、存储于多个装置中的一个或多个中的数据的数据的年龄、多个存储装置中的一个或多个中的空闲空间的量、写入到分配的空间中的数据的数据的期望寿命、当前存储于多个存储装置中的一个或多个中的数据的数据的期望寿命、装置年龄、错误率、错误的数量、可恢复错误的数量、不可恢复错误的数量和分配状态中的一个或多个。

4. 如权利要求 1 所述的计算机系统,其中,存储控制器进一步被配置为击碎多个存储装置中的保护数据,由此,存储控制器被配置为至少完成以下的过程:

确定给定数据的保护水平要改变;

识别存储与给定数据对应的保护数据的特定的存储位置;和

去分配特定的存储位置。

5. 如权利要求 4 所述的计算机系统,其中,存储控制器进一步被配置为重新分配特定的存储位置中的一个或多个,以用于在随后的写入中存储非保护数据。

6. 如权利要求 5 所述的计算机系统,其中,重新分配的特定的存储位置不需要对于非保护数据使用与保护数据使用的相同的尺寸或对准。

7. 如权利要求 1 所述的计算机系统,其中,多个存储装置中的存储装置被配置为擦除擦除块尺寸单元中的数据,并且其中,响应检测对于尺寸比擦除块小的数据的一部分的访问失败,存储控制器被配置为:

识别包含失败的数据的部分的数据的特定的擦除块尺寸部分;

重建无法形成数据的重建部分的数据的部分;和

使得存储装置写入包含数据的重建部分的数据的特定的擦除块尺寸部分。

8. 如权利要求 1 所述的计算机系统,其中,响应检测对于数据的一部分的访问失败,存储控制器被配置为:

重建无法形成数据的重建部分的数据的部分;

击碎多个存储装置中的特定的存储装置中的保护数据;和

在特定的存储位置中存储数据的重建部分。

9. 如权利要求 1 所述的计算机系统,其中,存储控制器进一步被配置为向多个存储装置中的存储装置发出使得存储装置重新映射和重写存储于存储装置中的数据的一部分的命令。

10. 一种用于计算机系统的方法,该方法包括:

在数据存储子系统上接收读取和写入请求,其中,子系统在包含多个存储装置的装置组上包含多个数据存储位置;

确定用于存储数据的第一 RAID 布局和第二 RAID 布局,其中第二 RAID 布局不同于第一 RAID 布局;

根据第一 RAID 布局将第一 RAID 条带写入到所述多个存储装置中的第一子组装置;

根据第二 RAID 布局将第二 RAID 条带写入到所述多个存储装置中的第二子组装置;和

根据第二 RAID 布局将第三 RAID 条带写入到所述多个存储装置中的第三子组装置,其中第三子组装置包括未包含在第二子组中的一个或多个装置;

由此,装置组根据第一 RAID 布局和第二 RAID 布局两者同时存储数据;

将第四 RAID 条带写入到所述多个存储装置,其中为所述第四 RAID 条带在第一装置上保持的装置内冗余数据的量不同于为第四 RAID 条带同时在第二装置上保持的装置内冗余数据的量。

11. 如权利要求 10 所述的方法,还包括基于多个存储装置中的一个或多个的特性确定第一 RAID 布局和第二 RAID 布局。

12. 如权利要求 11 所述的方法,其中,所述特性包含对于多个装置中的一个或多个的访问的数量或速率、存储于多个装置中的一个或多个中的数据的数据的年龄、多个存储装置中的一个或多个中的空闲空间的量、要被写入到分配的空间中的数据的数据的期望寿命、当前存储于多个存储装置中的一个或多个中的数据的数据的期望寿命、装置年龄、错误率、错误的数量、可恢复错误的数量、不可恢复错误的数量和分配状态中的一个或多个。

13. 如权利要求 10 所述的方法,还包括击碎多个存储装置中的保护数据,由此,方法包括:

确定给定数据的保护水平要改变;

识别存储与给定数据对应的保护数据的特定的存储位置;和

去分配特定的存储位置。

14. 如权利要求 13 所述的方法,还包括重新分配特定的存储位置中的一个或多个,以用于在随后的写入中存储非保护数据。

15. 如权利要求 14 所述的方法,其中,重新分配的特定的存储位置不需要对于非保护数据使用与保护数据使用的相同的尺寸或对准。

16. 如权利要求 10 所述的方法,其中,多个存储装置中的存储装置被配置为擦除擦除块尺寸单元中的数据,并且其中,响应检测对于尺寸比擦除块小的数据的一部分的访问失

败,方法还包括:

识别包含失败的数据的部分的数据的特定的擦除块尺寸部分;

重建无法形成数据的重建部分的数据的部分;和

使得存储装置写入包含数据的重建部分的数据的特定的擦除块尺寸部分。

17. 一种用于计算机系统的装置,包括:

用于在数据存储子系统上接收读取和写入请求的装置,其中,该子系统在包含多个存储装置的装置组上包含多个数据存储位置;

用于确定用于存储数据的第一 RAID 布局和第二 RAID 布局的装置,其中第二 RAID 布局不同于第一 RAID 布局;

用于根据第一 RAID 布局将第一 RAID 条带写入到所述多个存储装置中的第一子组装置的装置;

用于根据第二 RAID 布局将第二 RAID 条带写入到所述多个存储装置中的第二子组装置的装置;和

用于根据第二 RAID 布局将第三 RAID 条带写入到所述多个存储装置中的第三子组装置的装置,其中第三子组装置包括未包含在第二子组中的一个或多个装置;

由此,装置组根据第一 RAID 布局和第二 RAID 布局两者同时存储数据;

所述用于计算机系统的装置还包括响应于检测到第一子组装置的给定装置的可靠性的变化而将存储控制器配置为完成以下操作的装置:

为给定 RAID 条带在给定装置中保持第一量的空间,用于存储装置内冗余数据;以及

同时为所述给定 RAID 条带在不同于所述给定装置的装置中保持第二量的空间,用于存储装置内冗余数据,其中所述第二量不同于所述第一量。

SSD 环境中的适配 RAID

技术领域

[0001] 本发明涉及计算机网络,更特别地,涉及在多个固态存储装置之间有效地分布数据。

背景技术

[0002] 随着计算机存储器存储和数据带宽的增加,商业日常管理的数据的量和复杂性增加。诸如数据中心的大规模分布式存储系统一般运行许多商业运算。分布式存储系统可与通过一个或多个网络互连的客户机计算机耦合。如果分布式存储系统的任意部分具有不良的性能或者变得不可用,那么公司运算会受损或者完全停止。因此,期望分布式存储系统对于数据可用性保持高的标准并保持高性能功能。如这里使用的那样,由于存储技术中的一些类型不包含盘,因此,存储盘可被称为存储装置。

[0003] 为了防止数据损失,存储装置常常包含错误检测和校正机制。这些机制常采取由装置产生并存储于装置自身的错误校正代码的形式。另外,分布式存储系统还可利用分散算法,以在存储装置的集合之间分布数据。这些算法一般在不依赖于中心目录的情况下将数据映射到存储装置。这种算法的例子包括可缩放散列(RUSH)和受控可缩放散列复制(CRUSH)。当没有中心目录时,分布式存储系统中的多个客户机可同时访问多个服务器上的数据对象。另外,存储的元数据的量会减少。但是,在具有改变的容量、输入输出(I/O)特性和可靠性问题的多个存储盘之间分布数据仍然存在困难的任务。与存储装置自身类似,这些算法也可包括诸如 RAID 类型算法(例如, RAID5 和 RAID6)或 Reed-Solomon 代码的错误检测和校正算法。

[0004] 与选择的存储装置相关的技术和机制确定用于在可动态添加和去除的多个存储装置之间分布数据的方法。例如,以上描述的算法是针对利用硬盘驱动(HDD)的系统开发的。HDD 包括一个或多个旋转的盘,每个盘涂有磁介质。这些盘日常以每分钟几千转的速度旋转几个小时。另外,电磁致动器负责在旋转的盘上定位磁读取/写入装置。这些致动器经受摩擦、磨损、振动和机械不对准,这导致可靠性问题。上述的数据分布算法基于 HDD 的特性和行为。

[0005] 另一种类型的存储介质的一个例子是固态盘(SSD)。固态盘也可称为固态驱动器。SSD 可模仿 HDD 接口,但 SSD 利用用于存储永久数据的固态存储器而不是在 HDD 中见到的电气机械装置。例如,SSD 可包含快擦写存储器的组。SSD 可在没有移动部分或机械延迟的情况下具有比 HDD 低的访问时间和等待时间。但是,SSD 一般具有明显的写入等待时间。除了不同的输入/输出(I/O)特性以外,SSD 经受与 HDD 不同的失效模式。因此,不能在包含用于存储的 SSD、同时利用对于 HDD 开发的分布数据旋转算法的系统中实现高的性能和高的可靠性。

[0006] 鉴于以上的情况,希望用于有效地分布数据并在多个固态存储装置之间检测和校正错误的系统和方法。

发明内容

[0007] 公开了用于有效地在多个固态存储装置之间分布和管理数据的计算机系统和方法的各种实施例。

[0008] 在一个实施例中,计算机系统包括被配置为在网络上向通过网络耦合为接收读取和写入请求的一个或多个数据存储阵列传输读取和写入请求的多个客户机计算机。设想包括多个存储装置上的多个存储位置的数据存储阵列。在各种实施例中,在用于数据存储和保护的独立驱动配置的冗余阵列(RAID)中配置存储装置。数据存储装置可包含用于数据存储的固态存储器技术,诸如快擦写存储器单元。数据存储子系统还包括存储控制器,该存储控制器被配置为:确定用于存储数据的第一 RAID 布局;和根据第一 RAID 布局将第一 RAID 条带写入到装置组。响应检测第一条件,控制器被配置为:确定与第一 RAID 布局不同的第二 RAID 布局;和根据第二布局将第二 RAID 条带写入到装置组,由此,装置组根据第一 RAID 布局和第二 RAID 布局同时存储数据。在各种实施例中,第一条件包含检测要分配用于在装置组中存储数据的空间。另外,设想控制器至少部分地基于多个存储装置中的一个或多个的特性确定使用的 RAID 布局的类型的实施例。

[0009] 考虑以下的描述和附图,这些和其它的实施例将变得十分明显。

附图说明

[0010] 图 1 是示出网络结构的一个实施例的一般化框图。

[0011] 图 2 是动态装置内冗余方案的一个实施例的一般化框图。

[0012] 图 3 是示出用于调整数据存储子系统内的装置内保护的方法的一个实施例的一般化流程图。

[0013] 图 4 是存储子系统的一个实施例的一般化框图。

[0014] 图 5 是装置单元的一个实施例的一般化框图。

[0015] 图 6 是示出状态表的一个实施例的一般化框图。

[0016] 图 7 是示出柔性 RAID 数据布局结构的一个实施例的一般化框图。

[0017] 图 8 是示出柔性 RAID 数据布局结构的另一实施例的一般化框图。

[0018] 图 9 是示出用于动态确定数据存储子系统内的布局的方法的一个实施例的一般化流程图。

[0019] 图 10 是柔性 RAID 数据布局结构的又一实施例的一般化框图。

[0020] 图 11A 示出装置布局的一个实施例。

[0021] 图 11B 示出段的一个实施例。

[0022] 图 11C 是示出不同的页类型内的数据存储配置的一个实施例的一般化框图。

[0023] 图 12 是示出混合 RAID 数据布局的一个实施例的一般化框图。

[0024] 图 13 是示出用于选择数据存储子系统内的交替的 RAID 几何关系的方法的一个实施例的一般化流程图。

[0025] 虽然能够对于本发明提出各种修改和替代形式,但在附图中作为例子表示并且在这里详细描述特定的实施例。但应理解,其附图和详细的描述不是要将本发明限于特定的公开的形式,相反,本发明要覆盖落入由所附的权利要求限定的本发明的精神和范围内的所有的变更方案、等同方案和替代方案。

具体实施方式

[0026] 在以下的描述中,阐述大量的特定的细节以使得能够彻底理解本发明。但是,本领域技术人员可以想到,可以在没有这些特定的细节的情况下实现本发明。在一些情况下,为了避免混淆本发明,没有详细表示公知的电路、结构、信号、计算机程序指令和技术。

[0027] 参照图 1,表示网络结构 100 的一个实施例的一般化框图。如后面进一步描述的那样,网络结构 100 的一个实施例包含通过网络 180 相互连接并与数据存储阵列 120a ~ 120b 连接的客户机计算机系统 110a ~ 110b。网络 180 可通过交换机 140 与第二网络 190 耦合。客户机计算机系统 110c 通过网络 190 与客户机计算机系统 110a ~ 110b 和数据存储阵列 120a ~ 120b 耦合。另外,网络 190 可通过交换机 150 与因特网 160 或其它的外部网络耦合。

[0028] 注意,在替代性实施例中,客户机计算机和服务器、交换机、网络、数据存储阵列和数据存储装置的数量和类型不限于图 1 所示的那些。在各种时间,一个或多个客户机可离线操作。另外,在操作中,当用户连接、断开和重新连接网络结构 100 时,各单个客户机计算机连接类型可改变。简要提供图 1 所示的部件中的每一个的进一步的描述。首先,描述由数据存储阵列 120a ~ 120b 提供的特征中的一些的概要。

[0029] 在网络结构 100 中,数据存储阵列 120a ~ 120b 中的每一个可用于在诸如客户机计算机系统 110a ~ 110c 的不同的服务器和计算机之间共享数据。另外,数据存储阵列 120a ~ 120b 可被用于盘镜像、备份和恢复、存档和存档数据的检索和从一个存储装置到另一个的数据迁移。在替代性实施例中,为了形成群,一个或多个客户机计算机系统 110a ~ 110c 可通过快速局域网络(LAN)相互链接。相互链接的一个或多个节点形成群,该群可共享诸如驻留于数据存储阵列 120a ~ 120b 中的一个内的群共享容量的存储资源。

[0030] 数据存储阵列 120a ~ 120b 中的每一个包含用于数据存储的存储子系统 170。存储子系统 170 可包含多个存储装置 176a ~ 176m。这些存储装置 176a ~ 176m 可向客户机计算机系统 110a ~ 110b 提供数据存储服务。存储装置 176a ~ 176m 中的每一个可被配置为接收读取和写入请求,并包含多个数据存储位置,每个数据存储位置可作为阵列中的行和列被寻址。在一个实施例中,存储装置 176a ~ 176m 内的数据存储位置可被配置成逻辑、冗余存储容器或 RAID 阵列(便宜 / 独立盘的冗余阵列)。但是,存储装置 176a~176m 可以不包括磁盘。在一个实施例中,存储装置 176a ~ 176m 中的每一个可利用与常规的硬盘驱动(HDD)不同的数据存储的技术。例如,存储装置 176a ~ 176m 中的一个或多个可包含由固态存储器构成的存储装置或者进一步与其耦合以存储永久数据。在其它的实施例中,存储装置 176a ~ 176m 中的一个或多个可包含使用旋转扭矩传输技术、磁阻随机存取存储器(MRAM)技术或其它存储技术的存储器,或者进一步与其耦合。这些不同的存储技术可导致存储装置之间的不同的可靠性特性。

[0031] 在存储装置 176a ~ 176m 中的每一个中使用的技术和机构的类型可确定用于数据对象映射和错误检测和校正的算法。在这些算法中使用的逻辑可包含于基本操作系统(OS) 116、文件系统 140、存储子系统控制器 174 内的一个或多个全局 I/O 调度器 178 和存储装置 176a ~ 176m 中的每一个内的控制逻辑中的一个或多个中。

[0032] 在一个实施例中,包含的固态存储器包含固态驱动(SSD)技术。一般地,SSD 技术

利用快擦写存储器单元。在本领域中,众所周知,快擦写存储器单元基于在浮动门中捕获和存储的电子的范围保持二进制值。完全擦除快擦写存储器单元在浮动门中不存储或者存在最少量的电子。特定的二进制值,诸如用于单级单元(SLC)快擦写的二进制值 1,与擦除的快擦写存储器单元相关。多级单元(MLC)快擦写具有与擦除的快擦写存储器单元相关的二进制值 11。在向快擦写存储器单元内的控制门施加比给定的阈值电压高的电压之后,快擦写存储器单元在浮动门中捕获给定范围的电子。因此,另一特定的二进制值,诸如用于 SLC 快擦写的二进制 0,与编程(写入)快擦写存储器单元相关。MLC 快擦写单元可根据向控制门施加的电压具有与编程的存储器单元相关的多个二进制值中的一个。

[0033] 一般而言,SSD 技术提供的读取访问等待时间比 HDD 技术低。但是,SSD 的写入性能明显受 SSD 内的自由可编程块的可用性影响。由于 SSD 的写入性能与 SSD 的读取性能相比明显较慢,因此,期望类似的等待时间的某些功能或操作会出现问题。另外,HDD 技术和 SSD 技术之间的技术和机构的不同导致数据存储装置 176a ~ 176m 的可靠性特性的不同。

[0034] 在各种实施例中,SSD 内的快擦写单元一般在写入新数据之前被擦除。另外,各种快擦写技术中的擦除操作也必须在块的基础上被执行。因此,块(擦除段或擦除块)内的所有快擦写存储器单元被一起擦除。快擦写擦除块可包含多个页。例如,页可具有 4 千字节(KB)的大小,并且,块可包含 64 页或 256KB。与快擦写装置中的读取操作相比,擦除操作可具有相对较高的等待时间—这又会增加相应的写入操作的等待时间。可在比擦除块尺寸低的粒度水平上执行快擦写技术的编程或读取。例如,可在字节、字或其它的尺寸上编程或读取快擦写单元。

[0035] 快擦写单元在重复的擦除和编程操作之后出现磨损。这种情况下的磨损是由于在基板与 MLC 快擦写单元的浮动门之间的电介质氧化物层中注入和捕获的电荷。在一个例子中,MLC 快擦写单元会具有有限的经受擦除和编程操作的次数,该次数诸如处于从 10000 到 100000 循环的范围中。另外,SSD 可经历导致相邻或相近的快擦写单元在另一快擦写单元被擦除或编程时经历偶然的程序干扰错误。并且,SSD 包含读取干扰错误,其中,当另一快擦写单元被读取时,出现附近的快擦写单元的偶然的程序变化。

[0036] 已知一个或多个存储装置 176a ~ 176m 中的每一个的特性会导致更有效的数据对象映射和错误检测和校正。在一个实施例中,存储控制器 174 内的全局 RAID 引擎 178 可对于存储装置 176a ~ 176m 检测以下方面中的至少一个或多个:对于 I/O 请求的不一致响应时间、用于相应的访问的不正确的数据、错误率和存取速率。至少响应这些特性,全局 RAID 引擎 178 可确定哪个 RAID 数据布局结构要利用存储装置 176a ~ 176m 内的相应的一组存储装置。另外,全局 RAID 引擎 178 可基于存储装置 176a ~ 176m 的特性动态地改变装置内冗余方案和装置间 RAID 数据布局。

[0037] 图 1 示出根据一个实施例的能够描述特征的系统的一个例子。以下提供更多的细节。参照图 1,提供网络结构 100 的部件的进一步的描述。

[0038] 网络结构的部件

[0039] 并且,如图所示,网络结构 100 包含通过网络 180 和 190 互连并与数据存储阵列 120a ~ 120b 连接的客户机计算机系统 110a ~ 110c。网络 180 和 190 可包括包含无线连接、直接局域网络(LAN)连接、存储区域网(SAN)、广域网络(WAN)连接的各种技术,诸如因特网和路由器等。网络 180 和 190 可包含一个或多个 LAN,这些 LAN 也可以是无线的。网

络 180 和 190 还可包含远程直接存储器存取 (RDMA) 硬件和 / 或软件、传输控制协议 / 内部协议 (TCP/IP) 硬件和 / 或软件、路由器、中继器、交换机和 / 或网络等。可以在网络 180 和 190 中使用诸如以太网、光纤通道、以太网光纤通道 (FCoE) 和 iSCSI 等的协议。交换机 140 可利用与网络 180 和 190 相关的协议。网络 190 可与用于因特网 160 的诸如传送控制协议 (TCP) 和因特网协议 (IP) 或 TCP/IP 的一组通信协议连接。交换机 150 可以是 TCP/IP 交换机。

[0040] 客户机计算机系统 110a ~ 110b 可代表任意数量的静止或移动计算机, 诸如桌面个人计算机 (PC)、工作站、膝上型计算机、手持计算机、服务器、服务器农场、个人数字助理 (PDA) 和智能电话等。一般而言, 客户机计算机系统 110a ~ 110b 包括包含一个或多个处理器芯的一个或多个处理器。各处理器芯包含用于根据预定的通用指令集执行指令的电路。例如, 可以选择 x86 指令集构架。作为替代方案, 可以选择 **Alpha®**、**PowerPC®**、**SPARC®** 或任何其它通用指令集构架。处理器芯可对于数据和计算机程序指令访问高速缓存存储器子系统。高速缓存子系统可与包含随机存取存储器 (RAM) 和存储装置的存储器层次耦合。

[0041] 客户机计算机系统内的各处理器芯和存储器层次可与网络接口连接。除了硬件部件以外, 客户机计算机系统 110a ~ 110c 中的每一个可包含存储于计算机层次内的基本操作系统 (OS)。基本 OS 可代表各种特定操作系统中的任一种, 诸如, 例如, **MS-DOS®**、**MS-WINDOWS®**、**OS/2®**、**UNIX®**、**Linux®**、**Solaris®** 或其它已知的操作系统。因而, 基本 OS 可操作为向最终用户提供各种服务, 并提供可操作为支持各种程序的执行的软件框架。另外, 客户机计算机系统 110a ~ 110c 中的每一个可包含用于支持高级虚拟机 (VM) 的管理程序。本领域技术人员公知, 可在桌面和服务器的使用虚拟化, 以完全或部分地从系统的硬件将诸如 OS 的软件去耦合。虚拟化可向最终用户提供多个 OS 在同一机器上运行的假象, 每个 OS 具有其自身的资源, 这种逻辑存储实体 (例如, 逻辑单元数量, LUN) 与数据存储阵列 120a ~ 120b 中的每一个内的存储装置 176a ~ 176m 对应。

[0042] 数据存储阵列 120a ~ 120b 中的每一个可被用于在诸如客户机计算机系统 110a ~ 110c 的不同的服务器之间共享数据。数据存储阵列 120a ~ 120b 中的每一个包含用于数据存储的存储子系统 170。存储子系统 170 可包含多个存储装置 176a ~ 176m。这些存储装置 176a ~ 176m 中的每一个可以是 SSD。控制器 174 可包含用于操作接收的读取 / 写入请求的逻辑。例如, 可至少在控制器 174 中执行以上简要描述的算法。随机存取存储器 (RAM) 712 可被用于批处理操作, 诸如接收的写入请求。

[0043] 基本 OS 132、文件系统 134、任何 OS 驱动程序 (未示出) 和存储于存储器介质 130 中的其它软件可提供访问文件的功能和这些功能的管理。基本 OS 134 和 OS 驱动程序可包含存储于计算机介质 130 上并且可通过处理器 122 执行以在与接收的请求对应中的存储子系统 170 中执行一个或多个存储器访问操作的程序指令。

[0044] 数据存储阵列 120a ~ 120b 中的每一个可使用网络接口 124 以与网络 180 连接。与客户机计算机系统 110a ~ 110c 类似, 在一个实施例中, 可在网络适配器卡上包含网络接口 124 的功能。可通过使用硬件和软件实现网络接口 124 的功能。可在网络接口 124 的网络卡实现上包含随机存取存储器 (RAM) 和只读存储器 (ROM)。可以使用一个或多个应用特定集成电路 (ASIC) 以提供网络接口 124 的功能。

[0045] 在一个实施例中,可以开发寻求对于用户数据和相应的错误校正代码(ECC)信息最优化数据布局的数据存储模型。在一个实施例中,模型至少部分地基于存储系统内的存储装置的特性。例如,在利用固态存储技术的存储系统中,可以利用特定的装置的特性以开发用于存储系统的模型,这些模型又用于通知相应的数据存储配置算法。例如,如果使用的特定的存储装置表现随时间的可靠性的变化,那么可在动态改变数据存储配置时考虑这种特性。

[0046] 一般而言,对于计算系统开发的任何模型是不完整的。在真实世界中经常存在太多的需要应对的变量以完全将给定的系统模型化。在一些情况下,可能能够开发不完整但仍然有价值的模型。如下面更完整地讨论的那样,描述基于下面的装置的特性将存储系统模型化的实施例。在各种实施例中,关于系统会如何表现基于某些预测执行数据存储配置的选择。基于装置的特性的理解,某些装置行为比其它行为更容易预测。但是,装置行为会随时间改变,并且,作为响应,选择的数据布局也会改变。如这里使用的那样,装置的特性可指的是装置的总体特性、诸如芯片或其它部件的装置的子部分的特性、擦除块的特性或者与装置有关的任何其它特性。

[0047] 装置内冗余

[0048] 现在转到图 2,表示示出动态装置内冗余方案的一个实施例的一般化框图。本领域技术人员公知,几个装置内冗余方案中的一个可被选择,以减少存储装置中的潜在扇区错误的影响。术语“扇区”一般指的是 HDD 上的存储的基本单元,诸如盘上的给定轨道内的区段。这里,术语“扇区”也可指的是 SSD 上的分配的基本单元。

[0049] SSD 内的分配单元可包含 SSD 内的一个或多个擦除块。参照图 2,用户数据 210 可指的是由最终用户修改和访问的存储数据和装置间错误校正代码(ECC)数据。装置间 ECC 数据可以是保持用户数据的其它存储装置上的一个或多个页产生的奇偶信息。例如,装置间 ECC 数据可以是在 RAID 数据布局结构中使用的奇偶信息。用户数据 210 可被存储于包含于存储装置 176a ~ 176k 中的一个或多个内的一个或多个页内。在一个实施例中,存储装置 176a ~ 176k 中的每一个是 SSD。

[0050] SSD 内的擦除块可包含几个页。如上所述,在一个实施例中,页可包含 4KB 的数据存储空间。擦除块可包含 64 页或 256KB。在其它的实施例中,擦除块可大到 1 兆字节(MB),并包含 256 页。可以以提供足够大尺寸单元和相对较少的单元的方式选择分配单元尺寸,以减少分配单元的经费跟踪。在一个实施例中,一个或多个状态表可保持分配单元的状态(分配、空闲、擦除、错误)、磨损水平和在分配单元内出现的错误(可校正和 / 或不可校正)的数量的计数。在各种实施例中,分配单元的尺寸可被选择,以平衡给定装置可用的分配单元的数量与分配单元的维护经费。例如,在一个实施例中,分配单元的尺寸可被选择为 SSD 的总存储容量的百分之一的约 1/100。页、擦除块和其它单元配置的数据存储空间的其它量是可能的,并且被设想。

[0051] 当不可访问存储装置内的给定扇区或其它存储单元时,出现潜在扇区错误(LSE)。对于给定的扇区,可能不能完成读取或写入操作。另外,可存在不可校正错误校正代码(ECC)错误。LSE 是在访问给定的扇区之前未被检测的错误。因此,事先存储于给定的扇区中的任何数据会丢失。在存储装置失效之后,当在 RAID 重构中遇到时,单个 LSE 可导致数据丢失。对于 SSD,另一 LSE 的出现的可能性的增加可能源自以下统计中的至少一个:装置

年龄、装置尺寸、访问速率、存储致密性和先前的可校正和不可校正错误的出现。为了防止给定存储装置内的 LSE 和数据丢失,可以在给定的存储装置内使用多个装置内冗余方案中的一个。

[0052] 装置内冗余方案利用给定存储装置内的 ECC 信息,诸如奇偶信息。该装置内冗余方案及其 ECC 信息与给定装置对应,并且,可保持于给定装置内,但与可在内部由装置自身产生和保持的 ECC 相异。一般而言,装置的在内部产生和保持的 ECC 对于包含装置的系统是不可见的。包含于给定存储装置内的装置内 ECC 信息可被用于增加给定存储装置内的数据存储可靠性。该装置内 ECC 信息被添加到可包含于另一存储装置内的其它 ECC 信息,诸如在 RAID 数据布局结构中利用的奇偶信息。

[0053] 高度有效的装置内冗余方案可充分地增强给定 RAID 数据布局的可靠性,以导致用于保持奇偶信息的装置的数量的减少。例如,如果存在附加的装置内冗余以保护各装置上的数据,那么双重奇偶 RAID 布局可被单个奇偶 RAID 布局替代。对于固定程度的存储效率,增加装置内冗余方案的冗余性会增加给定存储装置的可靠性。但是,以这种方式增加冗余性也可增加给定存储装置的输入/输出(I/O)性能的损失。

[0054] 在一个实施例中,装置内冗余方案将装置分成用于存储用户数据的多组位置。例如,分割可以是与由条带 250a ~ 250c 表示的 RAID 布局内的条带对应的装置内的一组位置。如数据 210 所示,用户数据或装置间 RAID 冗余信息可被存储于存储装置 176a ~ 176k 中的每一个内的一个或多个页内。在各存储装置内,装置内错误恢复数据 220 可被存储于一个或多个页内。如这里使用的那样,装置内错误恢复数据 220 可被称为装置内冗余数据 220。本领域技术人员公知,可通过在数据 210 内的信息的选择的位上执行功能,获得装置内冗余数据 220。也可使用基于 XOR 的操作以导出奇偶信息以存储于装置内冗余数据中。装置内冗余方案的其它例子包含单奇偶检查(SPC)、最大距离可分离(MDS)擦除代码、交错奇偶检查代码(IPC)、混合 SPC 和 MDS 代码(MDS+SPC)和列对角奇偶(CDP)。方案根据计算数据 220 的方式关于传输可靠性和经费改变。除了上述的冗余信息以外,系统可被配置为计算装置上的区域的校验和值。例如,可在信息被写入到装置时计算校验和。该校验和被系统存储。当从装置读回信息时,系统可重新计算校验和并将其与原存储的值相比较。如果两个校验和不同,那么信息不被适当地读取,并且,系统可使用其它的方案以恢复数据。校验和功能的例子包括循环冗余检查(CRC)、MD5 和 SHA-1。

[0055] 如条带 250a ~ 250c 所示,在存储装置 176a ~ 176k 中的每一个中,用于存储给定条带内的数据 210 的页的宽度或数量可以相同。但是,如条带 250b ~ 250c 所示,在存储装置 176a ~ 176k 中的每一个中,用于存储给定条带内的装置内冗余数据 220 的页的宽度或数量可以不相同。在一个实施例中,改变给定存储装置的特性或行为可至少部分地确定用于存储相应的装置内冗余数据 220 的宽度。例如,如上所述,快擦写单元经程序干扰错误和读取干扰错误,其中,编程或读取页可干扰附近的页并在这些附近页附近导致错误。当存储装置老化并产生更多的错误时,相应的装置内冗余数据 220 的量会增加。例如,在用于条带 250b 的写入操作之前,存储装置 176a ~ 176k 中的每一个的特性可被监视并被用于预测增加的错误比率。预测的存储装置 176c 和 176j 的错误的增加可被检测。作为响应,装置内冗余数据 220 的量对于存储装置 176c 和 176j 会增加。在图 2 的条带 250a 和 250b 的例子中,可对于条带 250a 和 250b 的存储装置 176c 和 176j 看到存储的保护数据的量。例如,

现在,不是用单一奇偶保护存储装置 176c 和 176j,而是可用双重奇偶或三重奇偶保护这些装置。注意,增加装置 176c 和 176j 的装置内保护的量不必相应地增加同一条带的其它装置。而是,条带的数据可如希望的那样在各装置中具有不同的保护水平。

[0056] 在各种实施例中,在选择的基础上,可出现给定的数据保护的水平的增加或减小。例如,在一个实施例中,保护的增加可仅对于被检测为产生更多的错误的存储装置,诸如以上的例子中的存储装置 176c 和 176j 出现。在另一实施例中,当存储装置 176c 和 176j 被检测为产生更多的错误时,保护的增加可对于存储装置 176a ~ 176k 中的每一个出现。在一个实施例中,诸如装置 176k 的奇偶装置上的装置内保护的量的增加可能需要减少在条带内保护的数据的量。例如,增加存储于给定条带的奇偶装置上的装置内数据的量必须对于由条带内的数据减少由该装置存储的奇偶数据的量。如果该奇偶数据的量减少到小于保护条带内的所有数据所需要的量,那么,如果希望连续的奇偶保护,则必须减少带内的数据。作为减少存储于条带内的数据的量的替代方案,可对于存储奇偶数据选择不同的装置。各种选项是可能的,并且被设想。还应注意,虽然这里描述的图 2 和其它的示图可示出相异的奇偶装置(例如,176k),但是,在各种实施例中,奇偶可跨着多个装置分布,而不是存储于单个装置内。因此,为了便于讨论,图中的单独的奇偶装置的示图一般可被视为逻辑示图。

[0057] 现在参照图 3,表示用于调整数据存储子系统内的装置内保护的方法 300 的一个实施例。在上述的网络结构 100 和数据存储阵列 120a ~ 120b 中体现的部件可一般根据方法 300 操作。本实施例中的步骤以依次的方式被表示。但是,一些步骤可以以与所示的次序不同的次序出现,一些步骤可同时被执行,一些步骤可与其它的步骤组合,并且,在另一实施例中可缺少一些步骤。

[0058] 在块 302 中,确定用于在存储装置中存储用户数据的第一空间量。如上面关于数据 210 描述的那样,该用户数据可以是在最终用户应用中使用的数据或在 RAID 结构中使用的装置间奇偶信息。该第一空间量可包含上述的存储装置内的一个或多个页。在一个实施例中,存储控制器 174 内的全局 RAID 引擎 178 从存储装置 176a ~ 176m 中的每一个接收行为统计。对于包含存储装置 176a ~ 176m 中的两个或更多个的给定装置组,全局 RAID 引擎 178 可确定 RAID 数据布局和装置内冗余的初始量,以在两个或更多个存储装置中的每一个内保持。在块 304 中,RAID 引擎 178 可确定用于在存储装置内存储相应的装置内保护数据的第二空间量。该第二空间量可包含存储装置内的一个或多个页。装置内保护数据可与上述的装置内冗余数据 220 对应。

[0059] 在块 306 中,数据被写入包含于给定装置组内的各存储装置内的第一空间量中。在一个实施例中,用户数据和装置间奇偶信息作为单个 RAID 条带跨着包含于给定装置内的多个存储装置被写入。重新参照图 2,用于被写入的相应的数据的宽度在各存储装置内相同。在块 308 中,通过 ECC 算法、基于 XOR 的算法或任何其它适当的算法产生装置内保护数据。另外,系统可产生校验和以帮助识别还没有适当地检索的数据。在块 310 中,产生的装置内保护数据被写入存储装置中的第二空间量中。

[0060] 在块 312 中,RAID 引擎 178 可监视一个或多个存储装置的行为。在一个实施例中,RAID 引擎 178 可包含相应的存储装置的模型,并且从存储装置接收行为统计以输入到模型。模型可通过利用存储装置的已知的特性预测存储装置的行为。例如,模型可预测给定的存储装置的即将到来的增加的错误率。如果 RAID 引擎 178 检测影响可靠性的给定的

存储装置的特性(条件块 314),那么在块 316 中,RAID 引擎可调整用于存储数据和相应装置内冗余数据的第一空间和第二空间量。例如,RAID 引擎可监视上述的统计,诸如至少装置年龄、存取速率和错误率。重新参照图 2,RAID 引擎 178 可检测存储装置 176c 和 176j 具有错误数量的增加。作为替代方案,RAID 引擎可预测存储装置 176c 和 176j 的错误的数量的增加。因此,在写入第二条带 250b 之前,RAID 引擎 178 可调整用于在存储装置 176a ~ 176k 中的每一个中存储数据 210 和数据 220 的页的数量。类似地,RAID 引擎 178 可检测存储装置 176b 具有更低的可可靠性。因此,在写入第三条带 250c 之前,RAID 引擎 178 可重新调整用于在存储装置 176a ~ 176k 中的每一个中存储数据 210 和数据 220 的页的数量。

[0061] 监视存储装置特性

[0062] 现在参照图 4,表示存储子系统的一个实施例的一般化框图。一个或多个存储装置 176a ~ 176m 中的每一个可在一个或多个装置组 173a ~ 173m 中的一个中被分割。也可存在具有其它装置的其它装置组。用于各存储装置的一个或多个相应的操作队列和状态表可包含于装置单元 400a ~ 400w 中的一个中。这些装置单元可存储于 RAM172 中。可对于装置组 173a ~ 173m 中的每一个包括相应的 RAID 引擎 178a ~ 178m。各 RAID 引擎 178 可包含跟踪包含于相应装置组内的存储装置中的每一个的统计的监视器 410。数据布局逻辑 420 可确定在相应的存储装置内对于用户数据、装置间冗余数据和装置内冗余数据分配的空间量。存储控制器 174 可包含用于执行以下的任务中的至少一个的其它控制逻辑 430: 磨损水准测量、废物收集、I/O 调度、用于到来和外出的信息包的去复制和协议转换。

[0063] 现在转到图 5,表示装置单元的一个实施例的一般化框图。装置单元可包含装置队列 510 和表 520。装置队列 510 可包含读取队列 512、写入队列 514 和诸如其它操作队列 516 的一个或多个其它队列。各队列可包含用于存储一个或多个相应的请求 530a ~ 530d 的多个条目。例如,用于相应的 SSD 的装置单元可包含用于至少存储读取请求、写入请求、修整请求和擦除请求等的队列。表 520 可包含一个或多个状态表 522a ~ 522b,每个包含用于存储状态数据或统计的多个条目 530。还应注意,虽然队列和表在该图和其它的图中被示为包括特定数量的条目,但是,条目本身未必相互对应。另外,队列、表和条目的数量可能与图所示的情况不同并且可相互不同。

[0064] 现在转到图 6,表示示出与给定装置对应的状态表的一个实施例的一般化框图。在一个实施例中,这种表可包含与诸如 SSD 的给定存储装置的状态、错误和磨损水平信息对应的数据。相应的 RAID 引擎能够访问该信息,这允许 RAID 引擎动态地改变对于数据存储分配的空间和用于装置间保护和装置内保护的方案。在一个实施例中,信息可至少包含装置年龄 602、错误率 604、在装置 606 上检测的错误的总数、可恢复错误的数量 608、不可恢复错误的数量 610、装置的访问速率 612、存储的数据的年龄 614 和用于分配空间的一个或多个分配状态 616a ~ 616n。分配状态可包含填充、空置和错误等。

[0065] 灵活 RAID 布局

[0066] 现在转到图 7,表示示出灵活 RAID 数据布局结构的一个实施例的一般化框图。RAID 引擎可确定用于存储装置 176a ~ 176k 的保护水平。例如,RAID 引擎可确定对于存储装置 176a ~ 176k 利用 RAID 双重奇偶。装置间冗余数据 240 可代表从相应的用户数据产生的 RAID 双重奇偶值。在一个实施例中,存储装置 176j 和 176k 可存储双重奇偶信息。应当注意,RAID 奇偶保护的其它水平是可能的并被设想。另外,在其它的实施例中,双重奇

偶信息的存储可在存储装置之间旋转,而不是存储于各 RAID 条带的存储装置 176j 和 176k 内。为了便于解释和描述,双重奇偶信息的存储被示为存储于存储装置 176j 和 176k 内。

[0067] 现在参照图 8,表示出灵活 RAID 数据布局结构的另一实施例的一般化框图。与图 7 所示的例子类似,可对于存储装置 176a ~ 176k 使用双重奇偶。虽然在本例子中描述 RAID 双重奇偶,但是,可以选择 RAID 数据布局结构中的任意冗余量。

[0068] 在操作中,RAID 引擎 178 可监视存储装置 176a ~ 176k 的特性并确定装置表现比初始或其它给定可靠性水平高的可靠性水平。作为响应,RAID 引擎 178 可将 RAID 保护从 RAID 双重奇偶变为 RAID 单一奇偶。在其它的 RAID 数据布局结构中,可以使用支持的冗余量的另一种减少。在其它的实施例中,可由存储控制器 174 内的其它的逻辑执行存储装置 176a ~ 176k 的监视和保护水平的改变。

[0069] 继续以上的例子,可以只产生单一奇偶信息,并对于在给定的 RAID 条带上执行的随后的写入操作存储它。例如,在改变支持冗余量的变化之后,可以不在用于写入操作的随后的 RAID 条带中使用存储装置 176k。另外,存储于存储装置 176k 中的数据可被无效化,由此释放存储。然后可对于其它的用途重新分配与存储装置 176k 中的释放数据对应的页。减少奇偶保护的量并释放以前对于存储奇偶保护数据信号的空间可被称为“奇偶击碎”。另外,在存储装置 176k 是 SSD 的实施例中,可在在条带 250a 内重新写入页之前在存储装置 176k 内出现一个或多个擦除操作。

[0070] 继续奇偶击碎的以上的例子,在奇偶击碎之后存储于条带 250a 内的存储装置 176k 的重新分配页内的数据可保持用户数据或不与条带 250a 对应的其它 RAID 条带的相应的 RAID 单一奇偶信息。例如,存储于条带 250a 内的存储装置 176a ~ 176j 中的数据可与在奇偶击碎之前执行的一个或多个写入操作对应。存储于条带 250a 中的存储装置 176k 中的数据可与在奇偶击碎之后执行的一个或多个写入操作对应。类似地,存储于条带 250b 内的存储装置 176a ~ 176j 中的数据可与在奇偶击碎之前执行的一个或多个写入操作对应。条带 250b 内的存储装置 176k 中的页可被释放,在以后被擦除,并且在以后通过与在改变支持冗余量之后执行的一个或多个写入操作对应的数据被重新写入。应当注意,当跨着存储装置旋转冗余信息时,该方案期望甚至会更有效。在这种实施例中,通过击碎释放的空间类似地跨着存储装置分布。

[0071] 重新参照图 8,在条带 250c 内的存储装置 176k 中表示的去分配的页代表在奇偶击碎之前事先存储了 RAID 双重奇偶信息的存储位置。但是,现在,这些页是无效的,并且还没有被重新分配。在以上的例子中,SSD 的特定的特性确定在存储装置 176k 中释放和重新分配页的方式和定时。这些特性的例子至少包括在重新编程(重新写入)一个或多个页之前擦除整个擦除块。从图 8 可以看出,当奇偶被击碎时,不必击碎整个装置。而是,可以如希望的那样对于各单个条带击碎奇偶。类似地,可通过向条带添加存储于附加的装置上的保护数据增加条带的奇偶保护。

[0072] 现在参照图 9,表示用于动态确定 RAID 布局的方法的一个实施例。在上述的网络结构 100 和数据存储阵列 120a ~ 120b 中体现的部件可一般根据方法 900 操作。在图 9 中,表示两个处理 910 和 920。处理中的每一个可同时或者以给定的次序操作。并且,本实施例中的步骤以依次的方式被表示。但是,一些步骤可以以与所示的次序不同的次序出现,一些步骤可同时被执行,一些步骤可与其它步骤组合,并且,在另一实施例中可缺少一些步

骤。块 910 示出存储控制系统监视系统内的存储装置的特性和行为的处理(块 912)。例如,可以观察和 / 或记录诸如图 6 描述的特性。如果检测特定的条件,诸如可靠性的变化(决定块 914),那么可改变用于存储数据的保护的量(块 916)。例如,当给定的装置的年龄相对年轻时,装置的可靠性可能是未知的(例如,装置可能在相对年轻的年龄上遭受“夭折”和失败)。因此,可以对于每个 RAID 条带使用一个或多个额外存储装置,以存储奇偶信息。在以后,当时间证明装置可靠时,该额外保护可被去除。在各种实施例中,可对于装置保持关于错误率的特性。例如,关于可校正和 / 或不可校正错误的特性可被保持和使用,以关于给定装置的可靠性做出决定。基于该信息,存储控制器可动态改变装置或条带的各种保护水平。

[0073] 图 9 的块 920 一般示出在要分配存储的条带或其它部分时(决定块 922)时可进行关于用于数据的布局和保护水平的决定(块 924)的处理。注意,可在此时执行块 910 的处理。作为替代方案,保护水平可能由处理 910 确定并被存储。块 924 的确定可然后基于存储的数据。在一个实施例中,一旦确定了给定的布局,就可从一组装置选择用于布局的特定的装置(块 925)。例如,在一个实施例中,一组 20 个装置可用。如果确定 5+2 的布局,那么可从 20 的组选择使用任何七个装置。另外,注意,具有选择的 5+2 布局的随后的写入不需要使用相同的 7 个装置。在确定布局、保护水平和用于条带的装置之后,条带被写入(块 926)。

[0074] 在各种实施例中,可以利用 RUSH 算法以确定给定条带的数据和冗余信息驻留于哪些装置上。例如,RUSH 算法可被用于选择特定的装置,以对于存储装置 176a ~ 176k 中的给定的条带利用 8+2RAID 布局。一般而言,如这里使用的那样,对于给定的数据条带,M+N 布局可一般描述包括 M 个数据装置和 N 个奇偶装置的布局。另外,如上所述,奇偶可跨着装置分布而不是完全位于特定的装置内。因此,8+2 布局可包含跨着 10 个装置条带化的数据和奇偶—其中的 8 个装置存储数据,2 个装置存储奇偶。在随后的情况下,可以选择 12+2 布局。以这种方式,可在要写入写入(例如,条带)时动态确定希望的布局和保护特性。在一个实施例中,存储装置 176a ~ 176k 可包含多于 10 的存储装置,诸如 30、50 或更多个存储装置。但是,对于具有 8+2 布局的条带,只使用存储装置中的 10 个。注意,可以选择装置中的任意的 10 个,并且,可对于选择用于存储条带的 10 个装置使用任何适当的算法。例如,可以使用 CRUSH 算法,以选择对于给定的 8+2 个 RAID 布局利用存储装置 176a ~ 176k 中的哪 10 个。

[0075] 在存储装置 176a ~ 176k 的选择的 8+2RAID 布局的一个例子中,可以使用存储装置中的 2 个以存储诸如奇偶信息的错误校正代码(ECC)信息。该信息可被用于执行重构读取请求。重新参照图 8,在本例子中,可以选择存储装置 176j 和 176k 以存储 RAID 双重奇偶信息。并且,奇偶信息可以以旋转的方式存储于包含于 RAID 阵列中的存储装置 176a ~ 176k 中的每一个之间而不是一致地存储于相同的存储装置中。为了便于解释和描述,存储装置 176j 和 176k 被描述为存储 RAID 双重奇偶。

[0076] 在块 926 中,在执行写入操作时,可以跨着包含于 RAID 阵列中的多个存储装置中将元数据、用户数据、装置内奇偶信息和装置间奇偶信息写入为 RAID 条带。在块 912 中,RAID 引擎 178 可监视 RAID 阵列中的一个或多个存储装置的行为。在一个实施例中,如图 4 所示,RAID 引擎 178 可包含监视器 410 和数据布局逻辑 420。RAID 引擎 178 可至少监视给定的存储装置的年龄、错误的数量和类型、自数据的最后的分配的检测配置变化、给定数据的年龄和 RAID 阵列中的存储空间当前使用等。

[0077] 由 RAID 引擎 178 监视的数据可被存储于 RAM172 中, 诸如存储于图 4 所示的装置单元 400a ~ 400w 中的一个中。诸如图 5 和图 6 所示的例子, 可以使用表以存储该数据。包含于相应的 RAID 引擎中的逻辑可通过监视存储装置的更新的统计检测和预测存储装置的行为。例如, 模型可预测给定的存储装置的即将到来的增加的错误率。

[0078] 如果检测到存储装置的增加的可靠性(条件块 908), 那么在块 910 中, RAID 引擎可降低系统内的数据保护水平。例如, 在一个实施例中, 可以减少存储于存储子系统内的奇偶信息的量。关于以上的例子, RAID 引擎可对于相应的 8+2RAID 阵列将 RAID 双重奇偶减少到 RAID 单一奇偶, 从而将其转换成 8+1RAID 阵列。在另一例子中, 给定的 RAID 阵列可在块 916 之前在 RAID 结构中利用 N 级冗余量或奇偶。在块 916 中, RAID 引擎可确定利用 (N-m) 级冗余量, 其中, $N > 1$ 且 $1 \leq m \leq N$ 。因此, 在给定的 RAID 条带的随后的写入操作中, 将存在至少 m 个的存储装置被写入到给定的 RAID 条带中。

[0079] 为了减少系统内的数据保护的水平, RAID 引擎(或另一部件)可执行上述的奇偶击碎。随后, 存储控制器 174 可重新分配作为击碎操作的结果释放的那些页以将其用于随后的写入操作中。

[0080] 当存储装置 176a ~ 176k 中的每一个均老化并且填充数据时, 如上所述, 可以从 RAID 阵列去除额外的奇偶信息。保持元数据、用户数据、相应的装置内冗余信息和装置间冗余信息中的一些。关于以上具有 8+2RAID 阵列的例子, 保持存储于存储装置 176a ~ 176j 中的信息。但是, 可从 RAID 阵列去除额外的装置间冗余信息或额外奇偶信息。例如, 可从 RAID 条带去除存储于存储装置 176k 中的额外奇偶信息。

[0081] 保持的信息, 诸如以上的例子中的存储于存储装置 176a ~ 176j 中的信息, 可保持在适当的位置。可对于随后的写入操作重新使用和重新分配存储额外奇偶信息的存储空间, 诸如以上的例子中的存储装置 176k 中的相应的页。在一个实施例中, 各新的分配接收新的虚拟地址。各新分配可具有任意给定的尺寸、任意给定的对准或几何形式, 并且可配合于任何给定的存储空间(虚拟或物理)中。在一个实施例中, 存储装置 176a ~ 176k 中的每一个和存储装置内的各分配页具有包含识别信息的标题。该识别信息可允许在不改变给定的配置的情况下对于释放的额外奇偶信息重新用户存储空间。

[0082] 在存储装置 176a ~ 176k 中的一个或多个是 SSD 的实施例中, 擦除块在将擦除块内的一个或多个页重新编程之前被擦除。因此, 在存储装置 176k 是 SSD 的实施例中, 相应的擦除块在将存储装置 176k 中的释放页重新编程之前被擦除。关于以上的具有原 8+2RAID 阵列的例子, 一个或多个擦除块在用数据 210 将页重新编程之前在条带 250a ~ 250b 内的存储装置 176k 中被擦除。原 8+2RAID 阵列现在是 8+1RAID 阵列, 使得存储装置 176j 对于在奇偶击碎之前写入的 RAID 条带提供单一奇偶信息。

[0083] 本领域技术人员公知, 在给定存储装置的读取或写入失效的过程中, 可从相应的 RAID 条带中的支持的装置间奇偶信息重构数据。重构的数据可被写入到存储装置。但是, 如果重构的数据无法被写入到存储装置, 那么所有的存储于存储装置上的数据会从相应的奇偶信息重建。重建数据可被重新定位于另一位置。通过快擦写存储器, 快擦写翻译层(FTL)重新映射数据的存储位置。另外, 通过快擦写存储器, 数据的重新定位包含在将擦除块内的相应的页重新编程之前擦除整个擦除块。相对于页以擦除块的粒度保持映射表允许重新映射表更紧凑。并且, 在重新定位过程中, 可以使用在奇偶击碎中释放的额外页。

[0084] 偏移奇偶

[0085] 现在转到图 10, 表示出灵活 RAID 数据布局结构的又一实施例的一般化框图。与图 8 所示的一般化框图类似, 可以使用灵活 RAID 数据布局结构。存储装置 176a ~ 176k 包含跨着多个存储装置布局的多个 RAID 条带。虽然存储装置 176a ~ 176k 中的每一个包含多个页, 但是, 为了便于解释, 只标注页 1010 和页 1020。在表示的例子中, 选择双重奇偶 RAID 数据布局, 其中, 存储装置 176j 和 176k 存储双重奇偶信息。

[0086] 存储装置 176a ~ 176k 中的页中的每一个存储特定类型的数据。一些页存储用户数据 210 和相应的产生的装置间奇偶信息 240。其它的页存储相应的产生的装置内奇偶信息 220。又一些页存储元数据 242。元数据 242 可包含页标题信息、RAID 条带识别信息和一个或多个 RAID 条带的记录数据等。除了装置间奇偶保护和装置内奇偶保护以外, 存储装置 176a ~ 176k 中的页中的每一个可包含诸如存储于各给定页内的校验和的附加的保护。在各种实施例中, 各条带开始时的单个元数据页可源自其它的条带标题。作为替代方案, 该页可以在奇偶碎片中处于不同的偏移上, 因此, 可以通过装置间奇偶保护该数据。“碎片”代表装置的一部分。因此, 奇偶碎片指的是装置的存储奇偶数据的部分。

[0087] 物理层

[0088] 在各种实施例中, 这里描述的系统可包括使得系统的其它元件与存储装置通信的物理层。例如, 调度逻辑、RAID 逻辑和其它逻辑可通过包含软件和 / 或硬件的任意适当的组合的物理层与存储装置通信。一般地, 物理层执行包括提供对于永久存储的访问和执行与数据存储的完整性有关的功能的各种功能。

[0089] 图 11A 示出 500GB 装置的假定装置布局的一个实施例。在各种实施例中, 这里描述的存储装置可通过装置的开始处的分区表 1101 和装置终了处的分区表的复制被格式化。另外, 装置标题 1103 可被存储第一块和最后的块中。例如, 在基于快擦写的存储装置中, 装置标题可被存储于第一擦除块和最后的擦除块中。如上所述, 擦除块是一般处于 256KB ~ 1MB 的范围中的快擦写结构。第一擦除块中的附加的未使用空间可被保留(填充 1105)。可对于写入记录和诊断信息 1107 保留各装置中的第二擦除块。其间的擦除块的剩余部分被分成多个擦除块的分配单元(AU)1109。AU 尺寸可被选择, 因此, 为了得到良好的分配粒度, 对于每个装置可存在合理数量的 AU。在一个实施例中, 在装置上可大概存在 10000 个 AU, 以允许在足够大的单元中分配以避免间接费用, 但为了便于跟踪, 单元的数量不能太多。可在 AU 状态表中保持 AU 的状态(分配 / 释放 / 擦除 / 坏)的跟踪。可在磨损水平表中保持 AU 的磨损水平, 并且, 可在 AU 错误表中保持错误的计数。

[0090] 在各种实施例中, 物理层跨着一组装置(可处于不同的节点上)在每个装置包含一个片段碎片的片段中分配空间。图 11B 示出片段和一个可能的片段布局中的该片段的各种可识别部分的一个实施例。还包括了描述存储在装置中的数据类型图例。在表示的实施例中, 单个片段被示为存储于多个器件中。示出的是数据装置 Data0 ~ DataN 和奇偶装置 Parity P 和 Parity Q。在一个实施例中, 各片段碎片在装置上包含一个或多个分配单元, 使得碎片的尺寸在各装置上相等。片段碎片 1123 被调出, 以示出片段碎片。在图 11B 中还示出在一个实施例中与页对应的 I/O 读取尺寸 1127。还示出可包含用于 I/O 碎片的页奇偶的一个或多个页的 I/O 奇偶大块 1129。

[0091] 在一个实施例中, 各片段将具有其自身的可包含以下参数中的一个或多个的几

何：

[0092] (1)RAID 水平—用于与片段中的装置保护相交的 RAID 水平。它可确定镜像、奇偶或 ECC RAID 以及片段碎片包含多少奇偶。

[0093] (2)装置布局 I/O 碎片尺寸—它代表用于在写入中跨着各装置条带的尺寸。它一般处于 256KB ~ 1MB 的范围中,并且可能是各装置上的擦除块尺寸的倍数。出于解释的目的,图 11B 调出 I/O 碎片尺寸 1125。

[0094] (3)I/O 读取尺寸—这是逻辑读取尺寸。各 I/O 碎片可格式化为一系列的逻辑页。各页又可包含标题和页中的数据的校验和。当读取被发出时,它将用于一个或多个逻辑页,并且,可通过校验和确证各页中的数据。

[0095] (4)I/O 碎片 RAID 水平— I/O 碎片具有碎片内奇偶,以处理在重建中发现的潜在错误。该参数确定对于碎片内保护使用什么类型的奇偶并由此确定保持碎片内奇偶的多少复制。

[0096] (5)I/O 奇偶大块—在各种实施例中,存储装置可在页的基础上完成 ECC。因此,如果看到错误,那么可能指示整个物理页的失败。I/O 奇偶大块是片段中的各装置上的物理页尺寸的最小公倍数,并且,通过剥掉 I/O 奇偶大块或 I/O 读取尺寸中的较大者中的 I/O 碎片,计算碎片内奇偶。可包括页奇偶的一个或多个页。在各种实施例中,在失败的校验和确证的情况下,该奇偶可被用于重建数据。

[0097] 在各种实施例中,当写入各新片段时,选择片段的 RAID 几何。RAID 几何的选择可基于诸如当前的活动节点和装置的组和片段中的数据的类型的因素。例如,如果 10 个节点或装置是可用的,那么可选择(8+2) RAID6 几何,并且,跨着节点剥离以耐受两个装置或节点的片段失败。如果节点失败,那么下一片段可切换到(7+2) RAID6 几何。在该片段内,片段碎片中的一些片段将包含数据并且一些包含 ECC (例如,奇偶)。

[0098] 在一个实施例中,存在五种类型的片段。这些片段中的三种与 AU 状态表、AU 错误表和磨损水平表对应。在一些实施例中,出于附加的保护,这三种片段可被镜像化。除了这三种片段,存在也可通过镜像另外保护的元数据片段。最后,存在保持客户机块和记录信息的数据片段。记录信息包含与片段中的客户机块相关的更新信息。如具有奇偶 P 和奇偶 Q 碎片的图 11B 所示,数据片段可能通过 RAID6 被保护。此外,片段表保持为处于在开始被来自所有片段碎片的标题的信息占据的存储器数据结构中。在一些实施例中,表可在所有节点上被完全高速缓冲,因此,任何节点可将存储访问翻译成物理地址。但是,在其它的实施例中,可以使用对象存储模型,这里,各节点可具有可取得逻辑基准并识别存储数据的片段布局节点的片段表。然后,请求会被转送到节点以识别节点上的确切的存储位置。图 11B 还示出识别占据片段中的明显的量的空间的任何(体积、快照)组合的片段尾部数据。当快照被去除时,数据净化器可帮助基于该数据识别用于废物收集的片段。

[0099] 在一个实施例中,在片段中,写入的基本单元是作为装置中的每一个上的一个 I/O 碎片的 *segio*。*segio* 中的各逻辑页通过包含页的校验和(可被称为“中间”校验和)的页标题被格式化,因此,数据的实际页尺寸比一个页稍大。对于片段的奇偶碎片中的页,页标题较小,使得数据页中的页校验和被奇偶页保护。各 I/O 碎片的最后的页是重新具有较小的标题并且针对页失败保护擦除块中的所有校验和和页数据的奇偶页。页尺寸在这里被称为可以是一个或多个物理快擦写页的 I/O 读取尺寸。对于一些片段,可以使用比物理页小的

读取尺寸。对于查找表信息的读取可被指数驱动并且数据的较小的部分可被读取同时仍获得希望的数据的元数据,可能出现这一点。在这种情况下,读取物理页的一半意味着用更少的数据完全占用 I/O 总线(和网络)并确证(例如,校验和)更少的数据。为了支持比物理页小的读取尺寸,实施例可包括擦除块终了处的多个奇偶页,使得所有奇偶页的总尺寸等于快擦写页尺寸。

[0100] 随着擦除块的磨损水平增加,错误的可能性增加。除了跟踪磨损水平,可关于以识别的错误的较高概率在擦除块上看到错误的频率保持观察数据。对于一些擦除块,可决定保持擦除块终了处的双重或三重错误校正奇偶而不是单个 RAID5 奇偶。在这种情况下,segio 的数据负载可因此减少。仅需要减少 segio 内的不良擦除块而不是所有的擦除块。擦除块中的页标题可被用于识别哪些页是奇偶以及哪些是数据。

[0101] 不管什么时候从存储器读取页,都可通过使用页校验和确证内容。如果确证失败,那么可尝试使用内部装置奇偶的数据的重建。如果失败,那么可使用用于片段的交叉装置 ECC 以重建数据。

[0102] 在数据片段中,负载区域可被分成两个区域。存在格式化为可包含与存储的客户机块有关的更新的记录数据的页。负载区域的剩余部分可包含格式化为客户机块的页。客户机块数据可以以压缩的形式被存储。大量的压缩算法是可能的,并且被设想。另外,在各种实施例中,可对于产生校验和使用 **Intel®** 高级加密标准指令。另外,可存在用于驻留于数据的同一页中并包含读取客户机块所需要的信息的客户机块的标题,包括用于压缩数据的算法的识别。废物收集可利用客户机块标题和 segio 中的记录条目。另外,客户机块可具有可以是用于去复制并用于检查解压缩数据的正确性的未压缩数据的校验和的数据散列。

[0103] 在一些实施例中,片段和 segio 可具有用于它们排序的单调增加的 ID 序号。作为写入 segio 的一部分,逻辑层可记录对于先前的刷新的依赖性。在开始,物理层可构建片段和 segio 的排序列表,并且,如果 segio 依赖于另一未完全 segio,那么它可被退回并且不被视为被写入。

[0104] 磨损水平表

[0105] 各装置的磨损水平表(WLT)可被存储于各装置本地的片段中。该信息也可被存储于各片段碎片的标题中。在一个实施例中,磨损信息是代表擦除和重新使用分配单元的次数的整数。由于磨损信息可能不准确,因此,当存在一定量的活动时或者当系统空闲合理的周期时,可以执行对于装置的表的刷新。WLT 在分配新 WLT 片段时可也负责清洗旧的 WLT 片段。为了增加保护的额外层,旧的复制可在释放之前被保持。例如,表管理器可确保随时保持 WLT 目录的先前的擦除块和当前的擦除块。当它分配新的片段时,在它被写入到新的片段的第二擦除块中之前,它不会释放旧的片段。

[0106] AU 状态表

[0107] AU 状态表(AST)跟踪各 AU 的状态。这些状态包含释放、分配、擦除和不良。AST 可被存储于装置上的片段中。将状态变为分配或释放可以是同步的更新,而将状态变为不良或擦除可以是异步的更新。该表一般可足够小,并具有可在 NVRAM 中记录更新的足够的更新。AST 在分配新的 AST 段时可负责清洗旧的 AST 段。由于可通过扫描驱动上的各 AU 的第一块完全恢复 AST,因此,不需要保持 AST 的旧的复制。

[0108] AU 错误表

[0109] AU 错误表(AET)可被用于跟踪各 AU 内的可恢复错误和不可恢复错误的数量。AET 被存储于装置上的片段上,并且,各字段可以是二字节整数。通过每 AU 四个字节,整个表可相对较小。

[0110] 现在参照图 11C,表示出具有不同的页类型的数据存储配置的一个实施例的一般化框图。在表示的实施例中,虽然其它的类型是可能的并且被设想,但表示三个页类型。表示的页类型包括包含元数据 1150 的页 1110、包含用户数据 1160 的页 1120 和包含奇偶信息 1170 (装置间或装置内)的页 1130。页 1110 ~ 1130 中的每一个包含可包含标题和识别信息的元数据 1140。另外,页 1110 ~ 1130 中的每一个可包含页内错误恢复数据 1142,诸如相应的校验和或其它错误检测和 / 或校正代码。该校验和值可对于存储于给定装置组中的存储装置 176a ~ 176k 中的数据提供更多的保护。

[0111] 并且,页 1130 可包含页间错误恢复数据 1144。数据 1144 可以是存储于其它存储装置中的页内数据 1142 得出的 ECC 信息。例如,重新参照图 10,存储装置间奇偶信息 240 的存储装置 176j 内的各页也可存储页间错误恢复数据 1144。数据 1144 可以是存储于存储装置 176a ~ 176i 中的一个或多个中的页内错误恢复数据 1142 产生的奇偶、校验和或其它值。在一个实施例中,数据 1144 是从存储于其它的存储装置中的一个或多个其它校验和值 1142 产生的校验和值。为了使存储装置 176j 中的给定页中的数据 1144 与存储装置 176a ~ 176i 中的一个或多个中的相应的页中的数据 1142 对准,可以向相应的页添加填充 1146。

[0112] 在一个实施例中,最终用户应用在扇区边界上执行 I/O 操作,其中,扇区是用于 HDD 的 512 字节。为了增加额外的保护,可以增加 8 字节校验和以形成 520 字节扇区。在各种实施例中,可在基于快擦写存储器的系统中使用压缩和重新映射,以允许在字节边界上而不是扇区边界上配置最终用户数据。另外,校验和(8 字节、4 字节或其它)可在标题之后或者在用户数据之前被放在页内,它可被压缩。在页 1110 ~ 1130 中的每一个中表示该放置。

[0113] 当最终用户应用读取 512 字节扇区时,在一个实施例中尺寸可以为 2KB ~ 8KB 的相应的页通过页开始处的 8 字节校验和具有额外的保护。在各种实施例中,页可能对于 2 扇区尺寸的非通电不能被格式化。如页 1110 ~ 1120 所示,校验和可以偏移几个字节进入页中。该偏移允许诸如页 1130 的奇偶页存储覆盖奇偶页的校验和和用于保护其它页的校验和的 ECC。

[0114] 对于另一保护水平,当计算校验和值时,可以包括数据位置信息。页 1110 ~ 1130 中的每一个中的数据 1142 可包含该信息。该信息可包含逻辑地址和物理地址。也可在该信息中包括扇区号码、数据大块和偏移号码、轨道号码和面数量等。

[0115] 替代性几何

[0116] 现在转到图 12,表示出混合 RAID 数据布局 1200 的一个实施例的一般化框图。虽然可以选择任何数量的分区,但是,表示三个分区。各分区与诸如图 1 所示的装置组 713a ~ 173b 的单独的装置组对应。各分区包含多个存储装置。在一个实施例中,诸如 CRUSH 算法的算法可被利用,以选择在 RAID 数据布局结构中使用哪些装置以用于数据存储。

[0117] 在表示的例子中,表示 L+1RAID 阵列、M+1RAID 阵列和 N+1RAID 阵列。在各种实施例中,L、M 和 N 可以不同、相同或者为两者的组合。例如,在分区 1 中表示 RAID 阵列 1210。

其它的存储装置 1212 是分区 1 中的其它 RAID 阵列的候选。类似地, RAID 阵列 1220 示出分区 2 中的给定 RAID 阵列。其它的存储装置 1222 是分区 2 中的其它 RAID 阵列的候选。RAID 阵列 1230 示出分区 3 中的给定 RAID 阵列。其它的存储装置 1232 是分区 3 中的其它 RAID 阵列的候选。

[0118] 在 RAID 阵列 1210、1220 和 1230 中的每一个内, 存储装置 P1 在各 RAID 阵列内提供 RAID 单一奇偶保护。存储装置 D1 ~ DN 在各 RAID 阵列内存储用户数据。并且, 用户数据和 RAID 单个奇偶信息的存储可在存储 D1 ~ DN 与 P1 之间旋转。但是, 用户数据的存储被描述为存储于装置 D1 ~ DN 中。类似地, 为了便于解释和描述, RAID 单一奇偶信息的存储被描述为存储于装置 P1 中。

[0119] 三个分区中的每一个之中的一个或多个存储装置可被选择, 以对于一个或多个给定的 RAID 阵列提供附加的支持的冗余量。例如, 分区 3 中的存储装置 Q1 可与 RAID 阵列 1210、1220 和 1230 中的每一个组合。存储装置 Q1 可对于 RAID 阵列 1210、1220 和 1230 中的每一个提供 RAID 双重奇偶信息。当条带被写入 RAID 阵列 1210、1220 或 1230 中的一个时, 产生并存储该附加的奇偶信息。并且, 该附加奇偶信息可覆盖阵列 1210、1220 和 1230 中的每一个中的条带。因此, 存储 RAID 奇偶信息的存储装置的数量与存储装置的总数的比较低。例如, 如果分区的每一个使用 $N+2$ 个 RAID 阵列, 那么存储 RAID 奇偶信息的存储装置的数量与存储装置的总数的比为 $3(2)/3(N+2)$ 或 $2/(N+2)$ 。相反, 混合 RAID 布局 1200 的比为 $(3+1)/3(N+1)$ 或 $4/(3(N+1))$ 。

[0120] 能够通过增加用于存储用户数据的存储装置的数量减小以上的比率。例如, 不是利用存储装置 Q1, 而是分区中的每一个可利用 $3N+2$ RAID 阵列。在这种情况下, 存储 RAID 奇偶信息的存储装置的数量与存储装置的总数的比为 $2/(3N+2)$ 。但是, 在重构读取操作中, $(3N+1)$ 存储装置对于单一装置失败接收重构读取请求。相反, 对于混合 RAID 布局 1200, 只有 N 个存储装置对于单一装置失败接收重构读取请求。

[0121] 注意, 三个分区中的每一个可利用不同的 RAID 数据布局结构。给定 RAID 数据布局结构的选择可基于存储 RAID 奇偶信息的存储装置的数量与存储装置的总数的给定比值。另外, 选择可基于可在重构中接收重构读取请求的存储装置的给定数量。例如, RAID 阵列 1210、1220 和 1230 可分别包含诸如 $L+a$ 、 $M+b$ 和 $N+c$ 的几何。

[0122] 另外, 可基于以上的条件选择诸如存储装置 Q1 的一个或多个存储装置, 以对于分区内的 RAID 阵列中的一个或多个提供附加的支持冗余量。在三个分区包含以上的 RAID 阵列并且数量 Q 的存储装置对于 RAID 阵列中的每一个提供额外的保护的例子中, 存储 RAID 奇偶信息的存储装置的数量与存储装置的总数的比是 $(a+b+c+Q)/(L+a+M+b+N+c+Q)$ 。对于单一装置失败, 对于以上的例子中的分区 1 ~ 3, 用于接收重构读取请求的存储装置的数量是分别是 L 、 M 和 N 。注意, 以上的讨论一般描述图 12 中的 3 个相异的分区。在这种实施例中, 这种将给定的布局限于特定的装置组的“硬”分区的类型可保证一个分区中的重构读取不与另一分区中冲突。但是, 在其它的实施例中, 分区可没有上面描述的那样硬。而是, 给定装置的池, 可从装置的任一个选择布局。例如, 将装置视为一个大的池, 能够配置诸如 $(L+1, M+1, N+1)+1$ 的布局。因此, 存在几何重叠并且重构读取会冲突的机会。如果 L 、 M 和 N 相对于池的尺寸较小, 那么重构读取相对于正常读取的百分比可保持较低。

[0123] 现在参照图 13, 表示用于在数据存储子系统中选择替代性 RAID 几何的方法 1300

的一个实施例中。在上述的网络结构 100 和数据存储阵列 120a ~ 120b 中体现的部件可一般根据方法 1300 操作。本实施例中的步骤以依次的方式被表示。但是,一些步骤可以以与所示的次序不同的次序出现,一些步骤可同时被执行,一些步骤可与其它的步骤组合,并且,在另一实施例中可缺少一些步骤。

[0124] 在块 1302 中,RAID 引擎 178 或存储控制器 174 中的其它逻辑确定使用给定数量的装置以在存储子系统的各分区内的 RAID 阵列中储存用户数据。然后可使用 RUSH 或其它算法以选择使用哪些装置。在一个实施例中,各分区利用相同数量的存储装置。在其它的实施例中,各分区可利用不同的唯一数量的存储装置以存储用户数据。在块 1304 中,存储控制器 174 可确定支持用于存储子系统的各分区内的相应的装置间错误恢复(奇偶)数据的存储装置的数量。并且,各分区可利用相同数量或不同唯一数量的用于存储 RAID 奇偶信息的存储装置。

[0125] 在块 1306 中,存储控制器可确定支持数量 Q 的存储装置跨着子系统的分区存储额外装置间错误恢复(奇偶)数据。在块 1308 中,可在选择的存储装置中写入用户数据和相应的 RAID 奇偶数据。重新参照图 12,当写入诸如分区 1 中的 RAID 阵列 1210 的给定的 RAID 阵列时,可产生并在分区 3 中的存储装置 Q1 中存储奇偶信息的一个或多个位。

[0126] 如果存储控制器 174 检测用于在给定的分区中执行读取重构的条件(条件块 1310),并且如果给定的分区具有足够数量的保持 RAID 奇偶信息的存储装置以处理大量的不可用的存储装置(条件块 1312),那么在块 1314 中,通过给定的分区内的一个或多个相应的存储装置执行重构读取操作。条件可包括:由于装置失败或者装置在低于给定的性能水平下操作,给定的 RAID 阵列内的存储装置不可用。给定的 RAID 阵列能够用给定分区内的存储 RAID 奇偶信息的存储装置的数量处理最大数量的不可用存储装置。例如,如果以上的例子中的分区 1 中的 RAID 阵列 1210 是 L+a RAID 阵列,那么,当 k 个存储装置不可用时,RAID 阵列 1210 能够仅通过使用分区 1 内的存储装置执行读取重构,这里, $1 \leq k \leq a$ 。

[0127] 如果给定的分区不具有足够的保持 RAID 阵列信息的数量的存储以处理大量的不可用的存储装置(条件块 1312),并且,如果存在足够的数量的 Q 存储装置以处理大量的不可用的存储装置(条件块 1316),那么在块 1318 中,通过一个或多个相应的 Q 存储装置执行重构读取操作。可在读取重构中访问存储用户数据的其它分区中的一个或多个存储装置。这些存储装置的选择可基于存储于一个或多个 Q 存储装置中的奇偶信息的导出的方式。例如,重新参照图 12,可在读取重构中访问分区 2 中的存储装置 D2,原因是可以使用该存储装置以产生存储于存储装置 Q1 中的相应的 RAID 奇偶信息。如果不存在足够的数量的 Q 存储装置以处理大量的不可用的存储装置(条件块 1316),那么在块 1320 中,相应的用户数据可从另一来源被读取或者被视为丢失。

[0128] 注意,上述的实施例可包含软件。在这种实施例中,实现方法和/或机构的程序指令可被传输或者被存储于计算机可读介质上。被配置为存储程序指令的大量的类型的介质是可用的,并且包含硬盘、软盘、CD-ROM、DVD、快擦写存储器、可编程 ROM (PROM)、随机存取存储器(RAM)和各种其它形式的易失性或非易失性存储器。

[0129] 在各种实施例中,这里描述的方法和机构的一个或多个部分可形成云计算环境的一部分。在这种实施例中,可根据一个或多个各种模型作为服务在因特网上提供资源。这种模型可包含基础构架即服务(IaaS)、平台即服务(PaaS)和软件即服务(SaaS)。在 IaaS

中,计算机基础构架作为服务被传输。在这种情况下,计算设备一般被服务提供商拥有和操作。在PaaS模型中,开发人员使用以用于开发软件方案的软件工具和下层的设备可作为服务被提供,并被服务提供商托管。SaaS一般包含服务提供商许可服务作为按需服务。服务提供商可托管软件,或者可对于给定的时间周期将软件部署到顾客。以上的模型的大量的组合是可能的,并且被设想。

[0130] 虽然以大量的细节描述了以上的实施例,但是,对于本领域技术人员来说,一旦完全理解以上的公开,大量的变更和修改将变得十分明显。以下的权利要求应被解释为包含所有这些变更和修改。

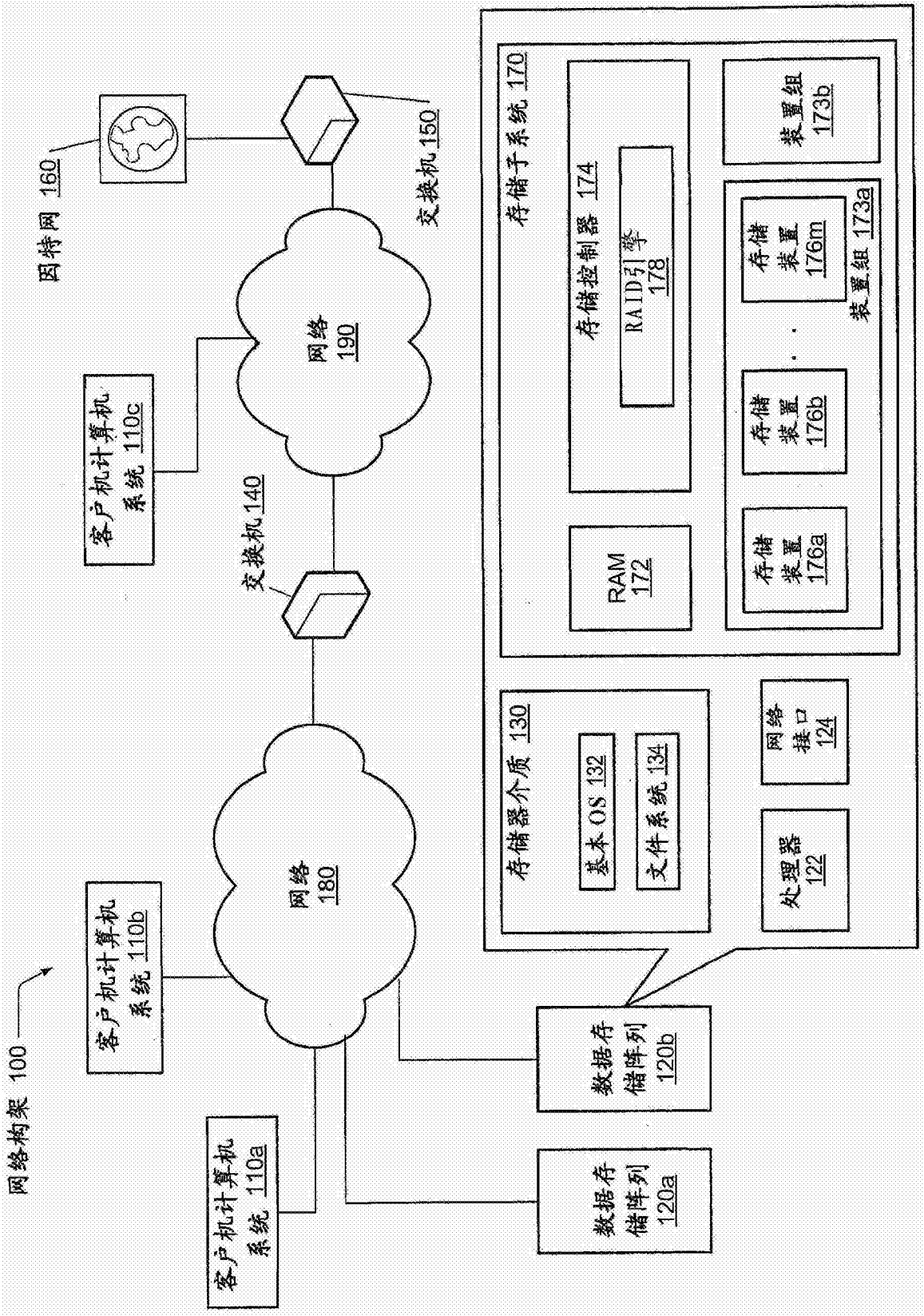


图 1

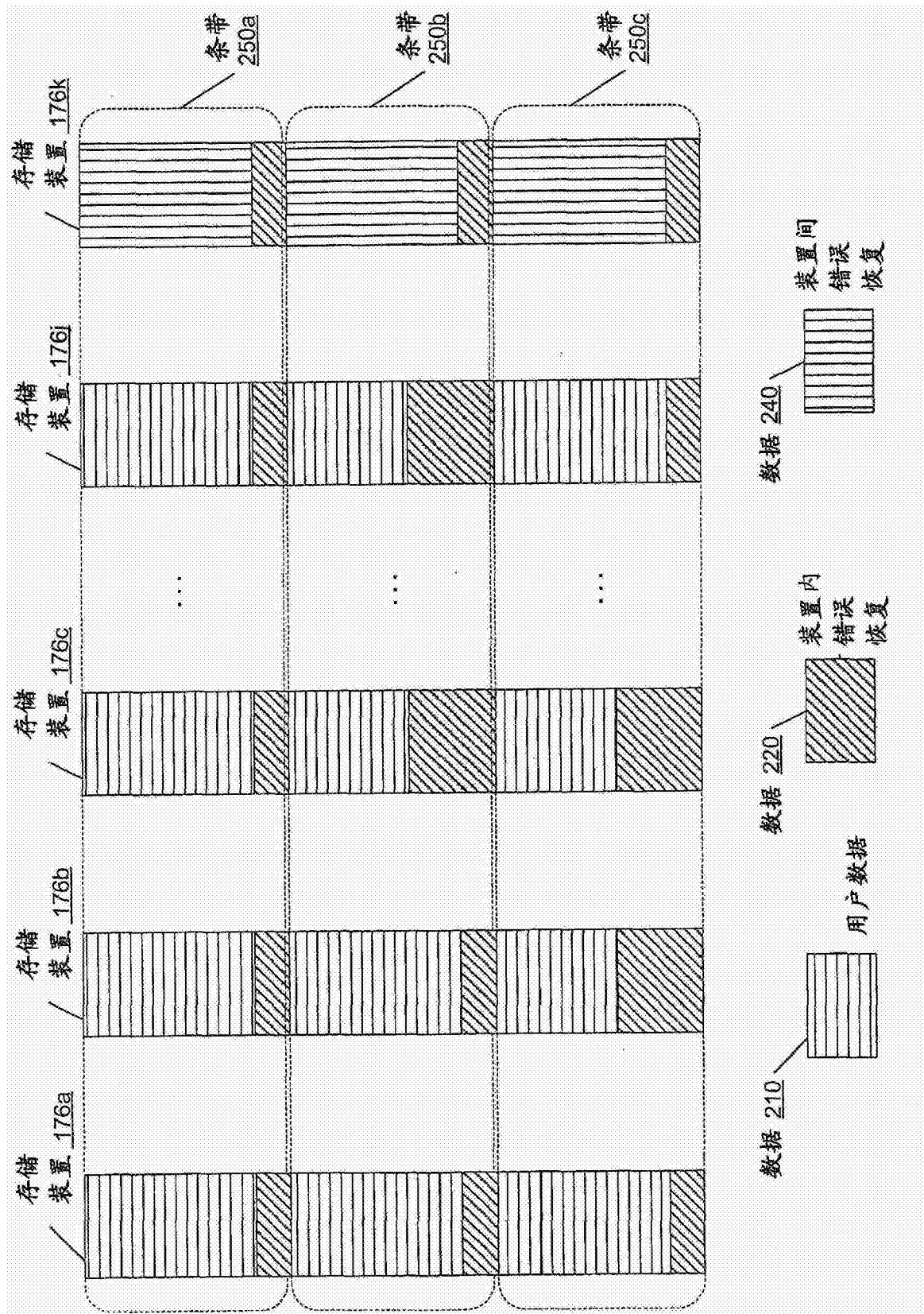


图 2

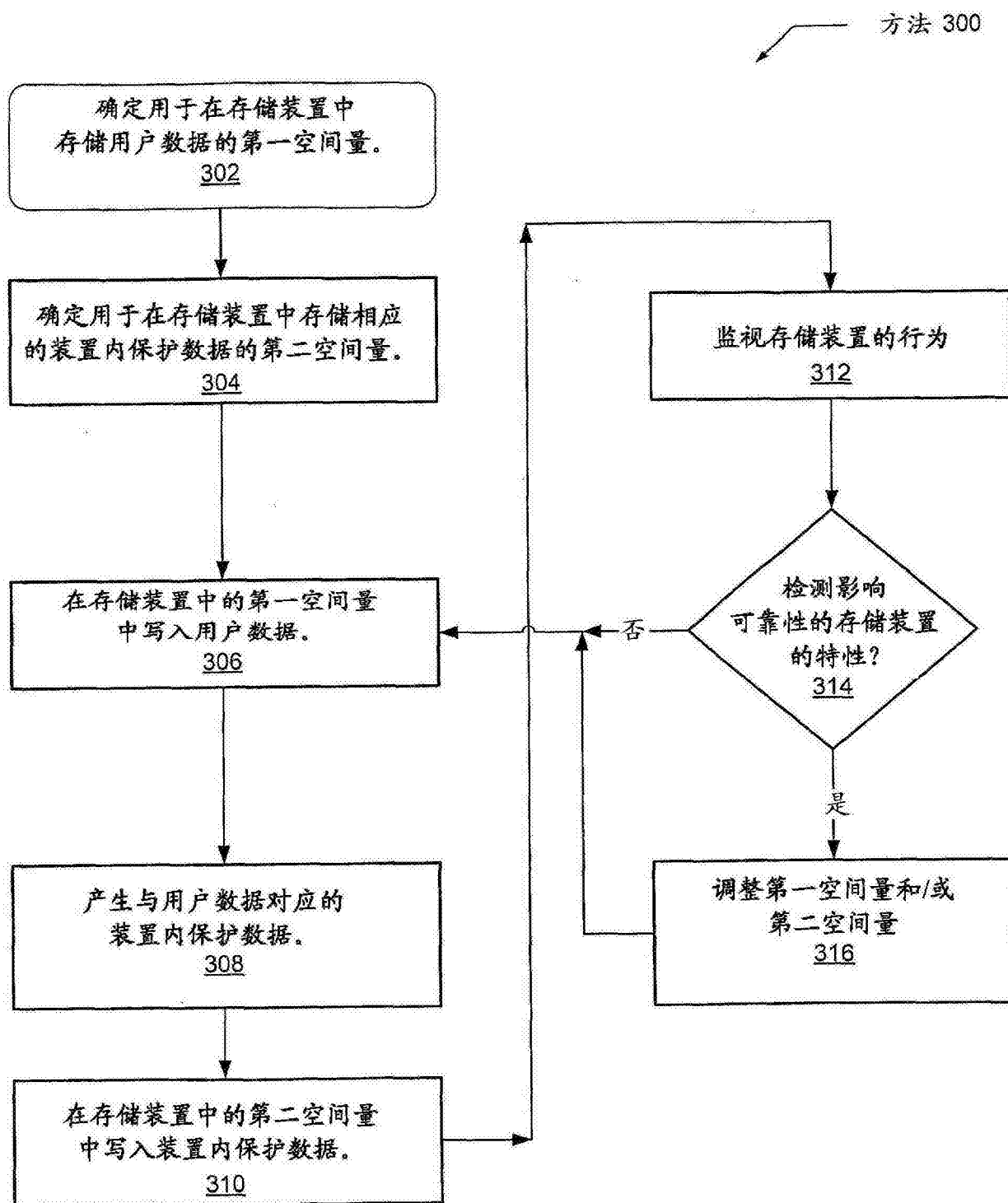


图 3

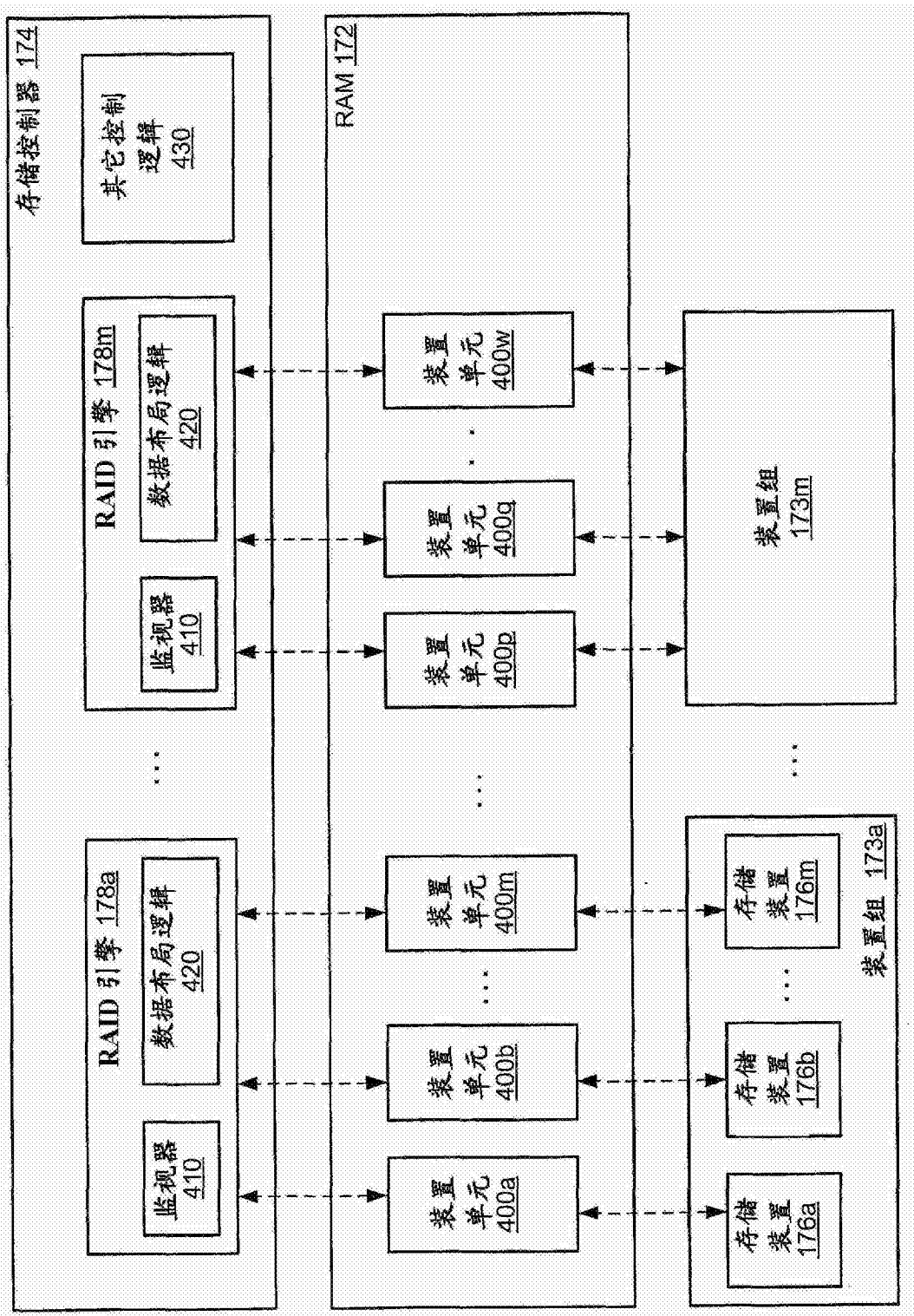


图 4

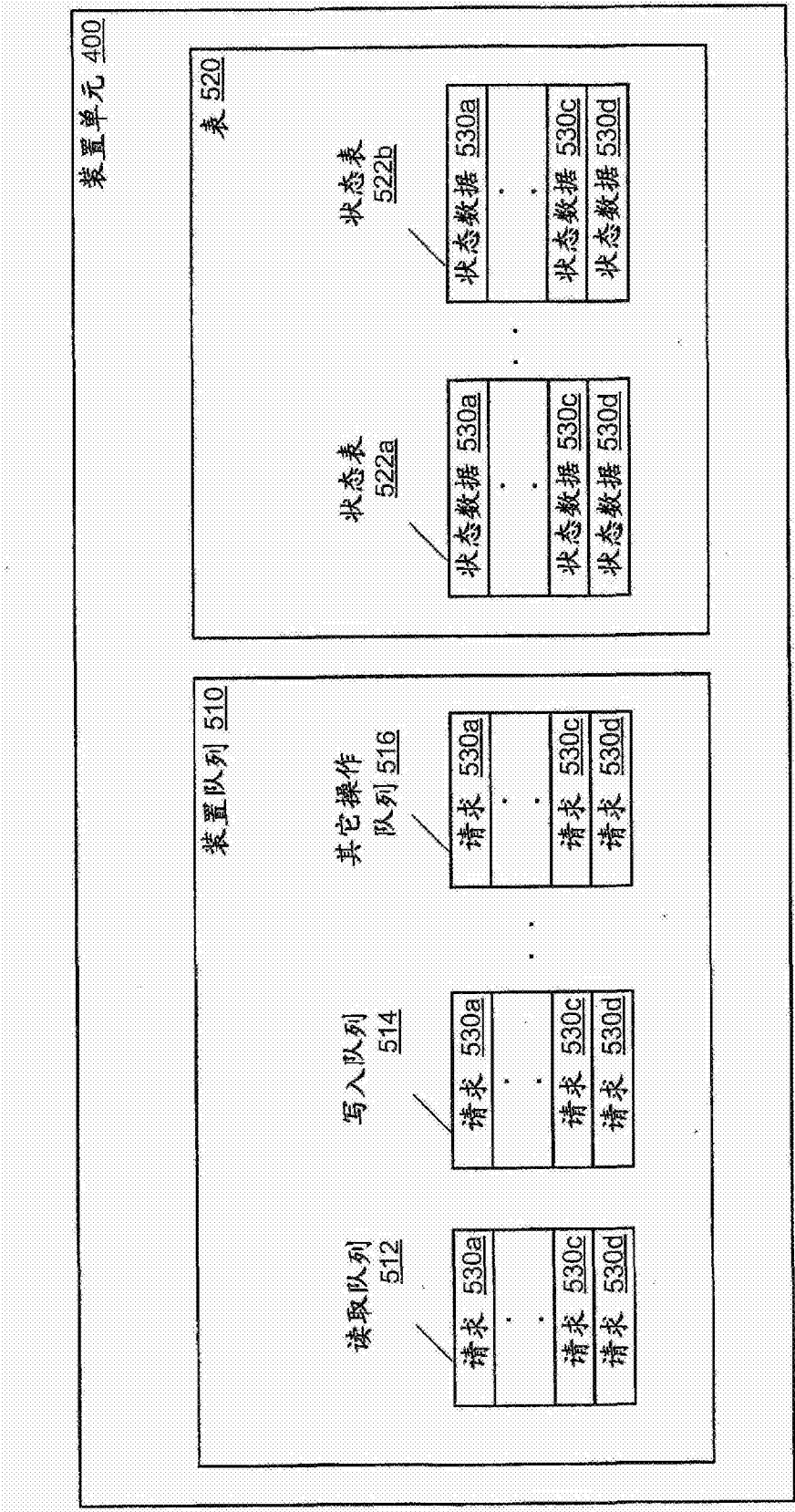
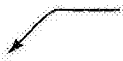


图 5

状态表 522



装置年龄 602
错误率 604
总错误 606
可恢复错误的数量 608
不可恢复错误的数量 610
访问速率 612
数据年龄 614
第一分配空间的分配状态 616a
· · ·
第 n 分配空间的分配状态 616n

图 6

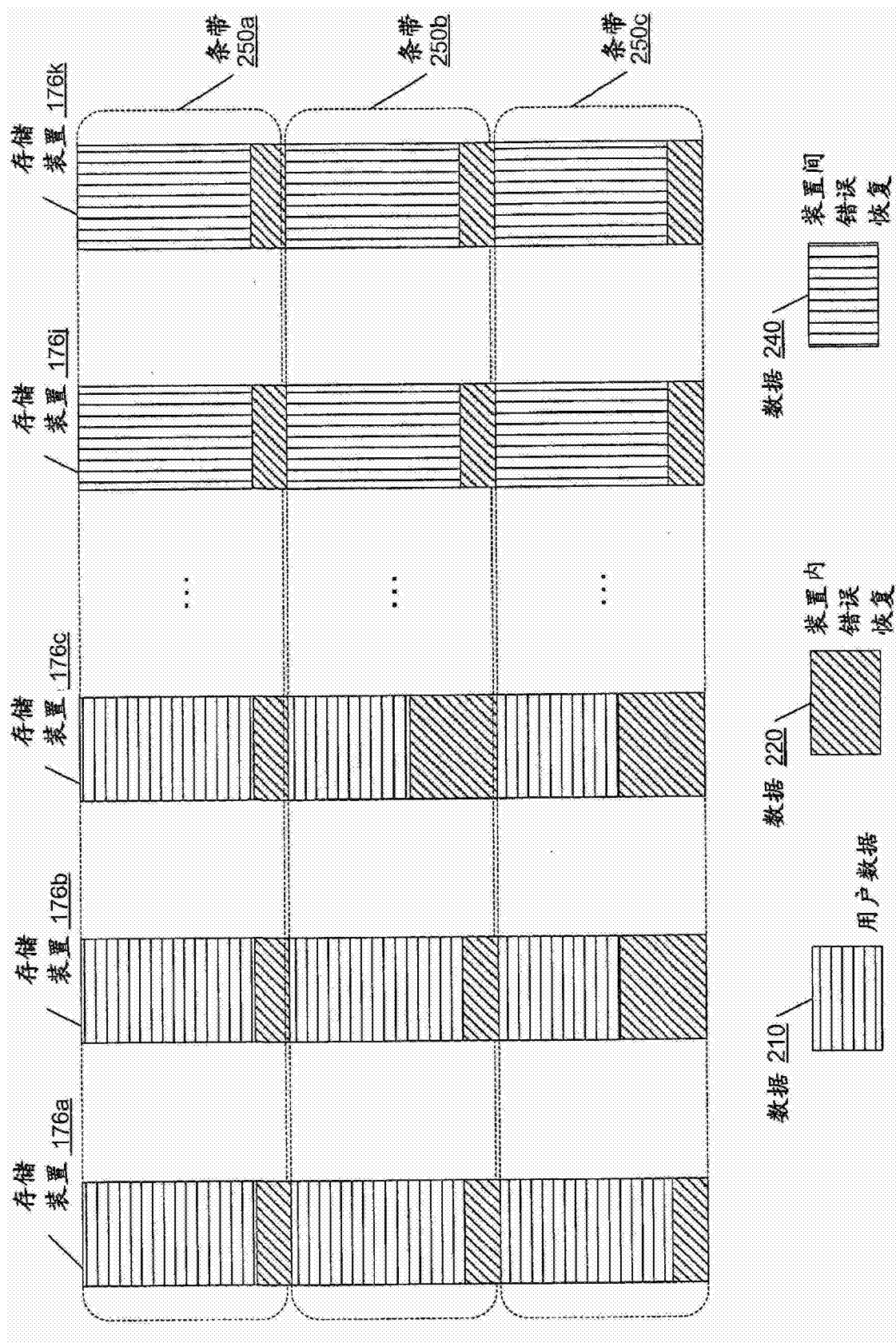


图 7

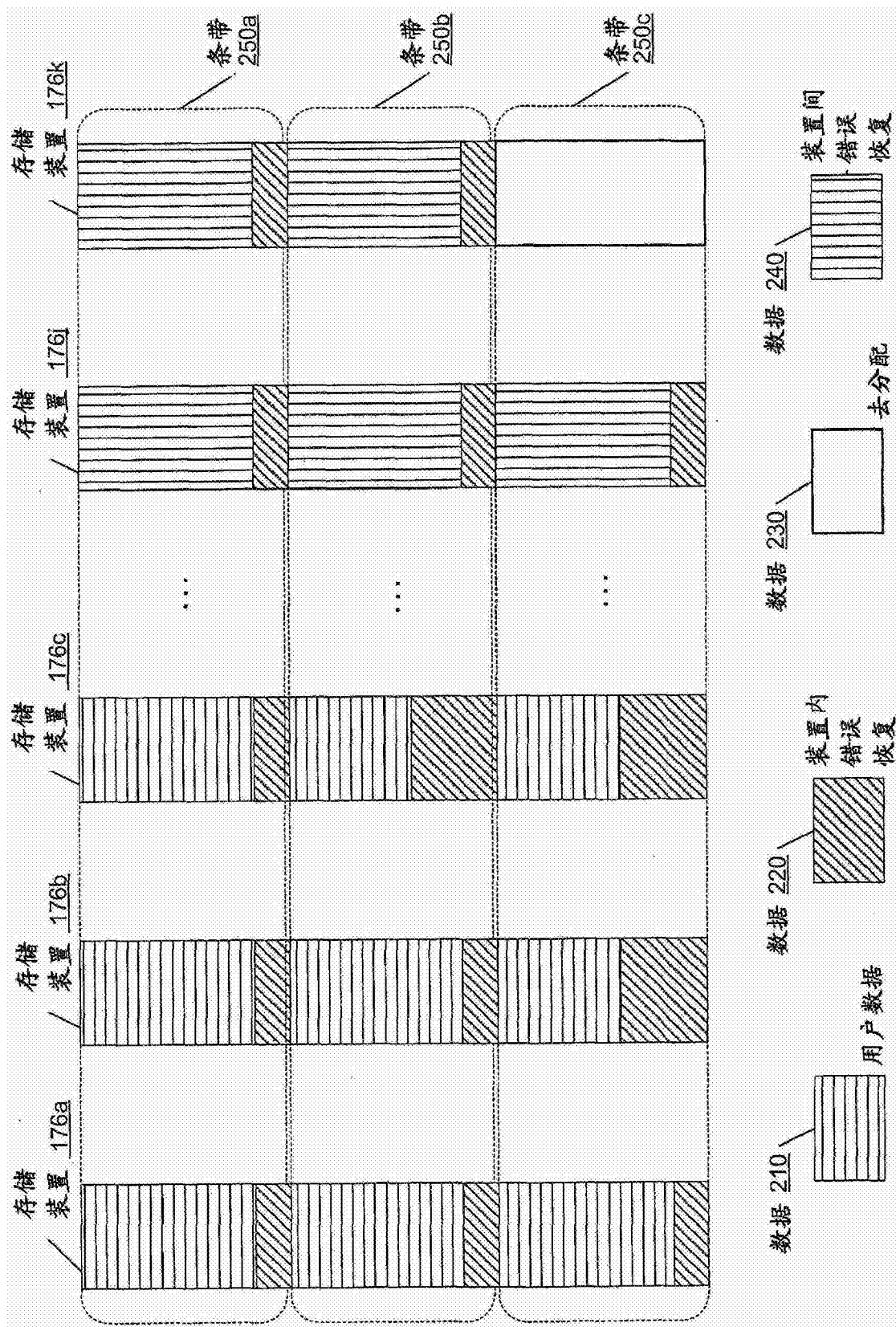


图 8

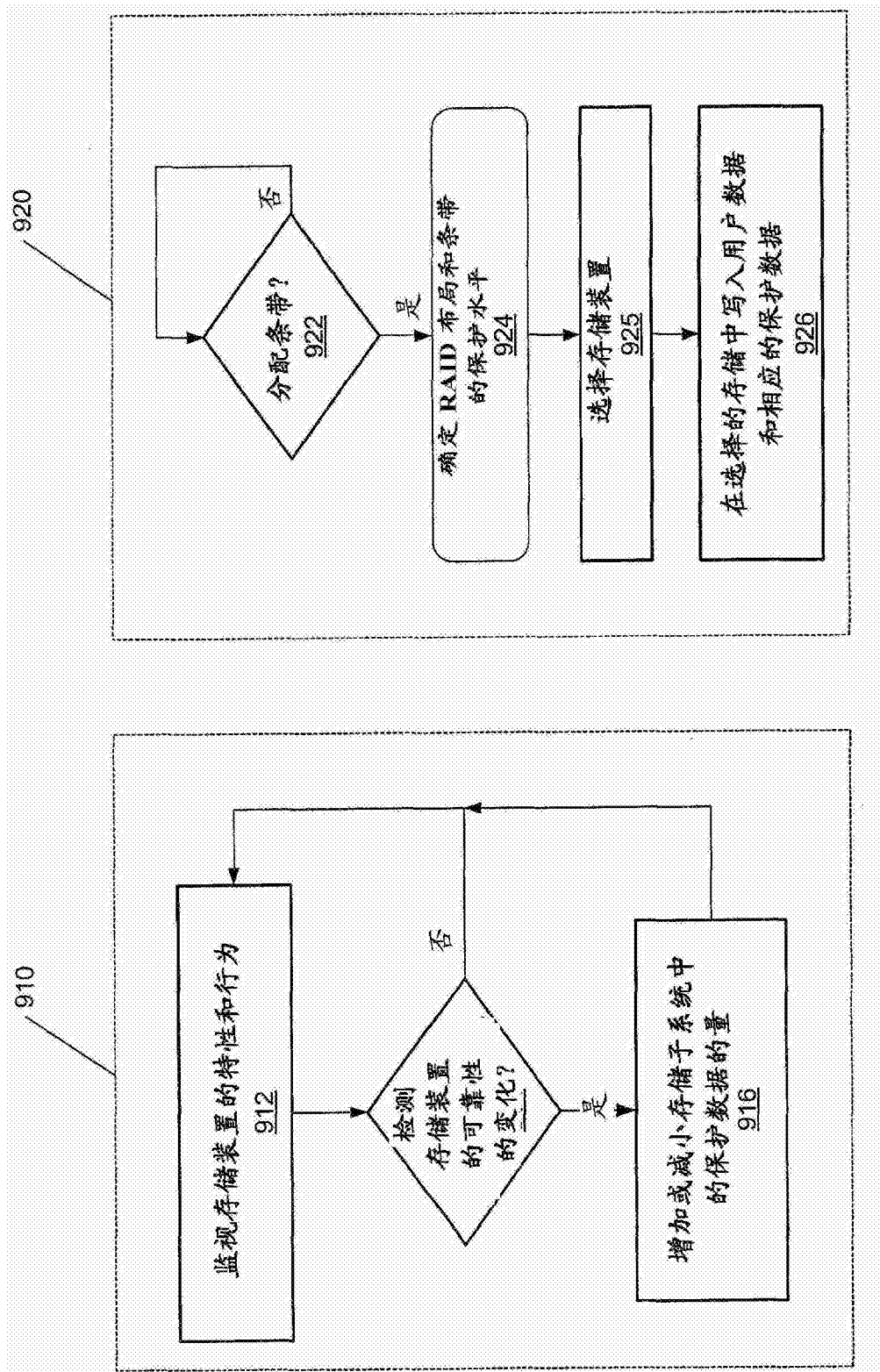


图 9

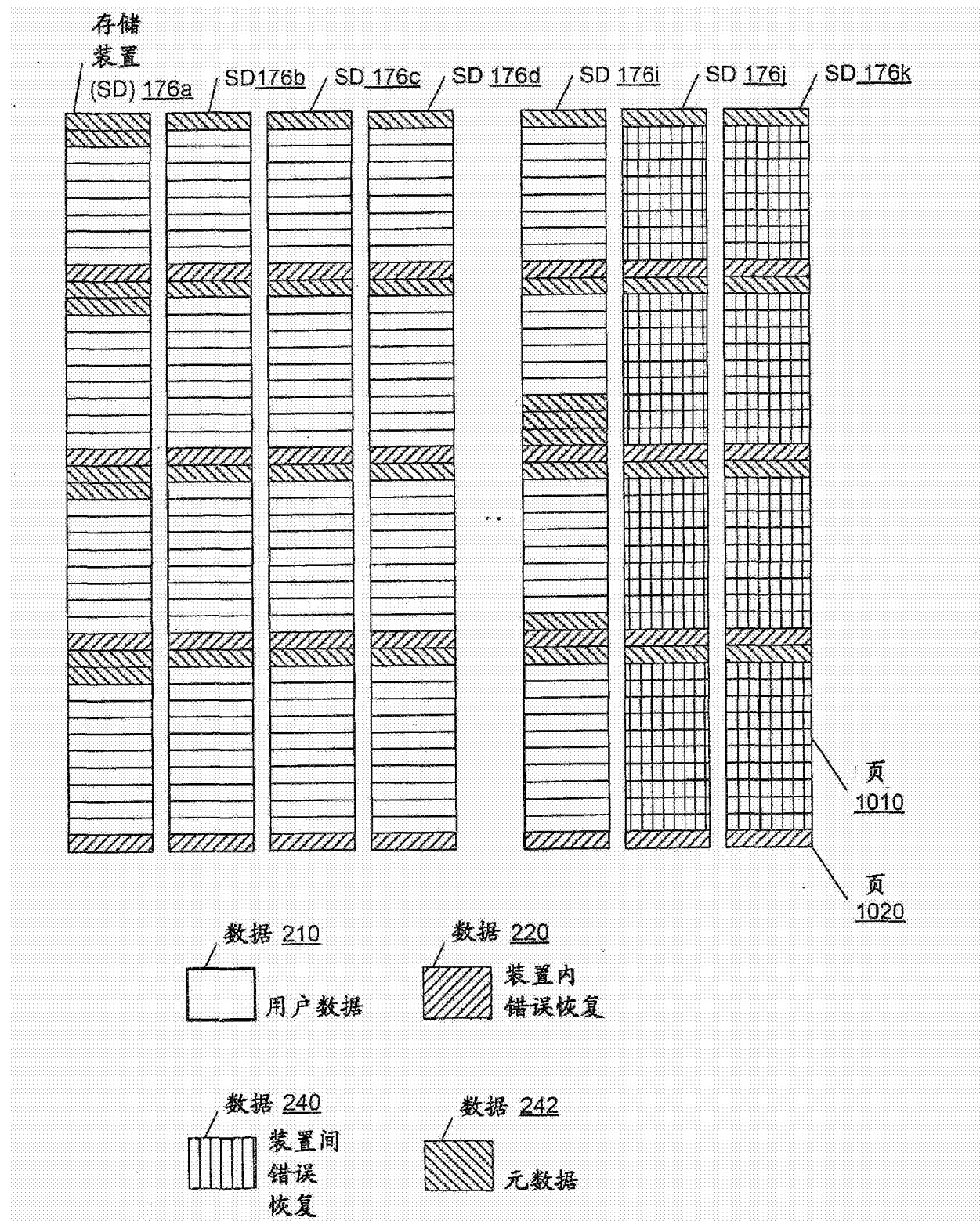


图 10

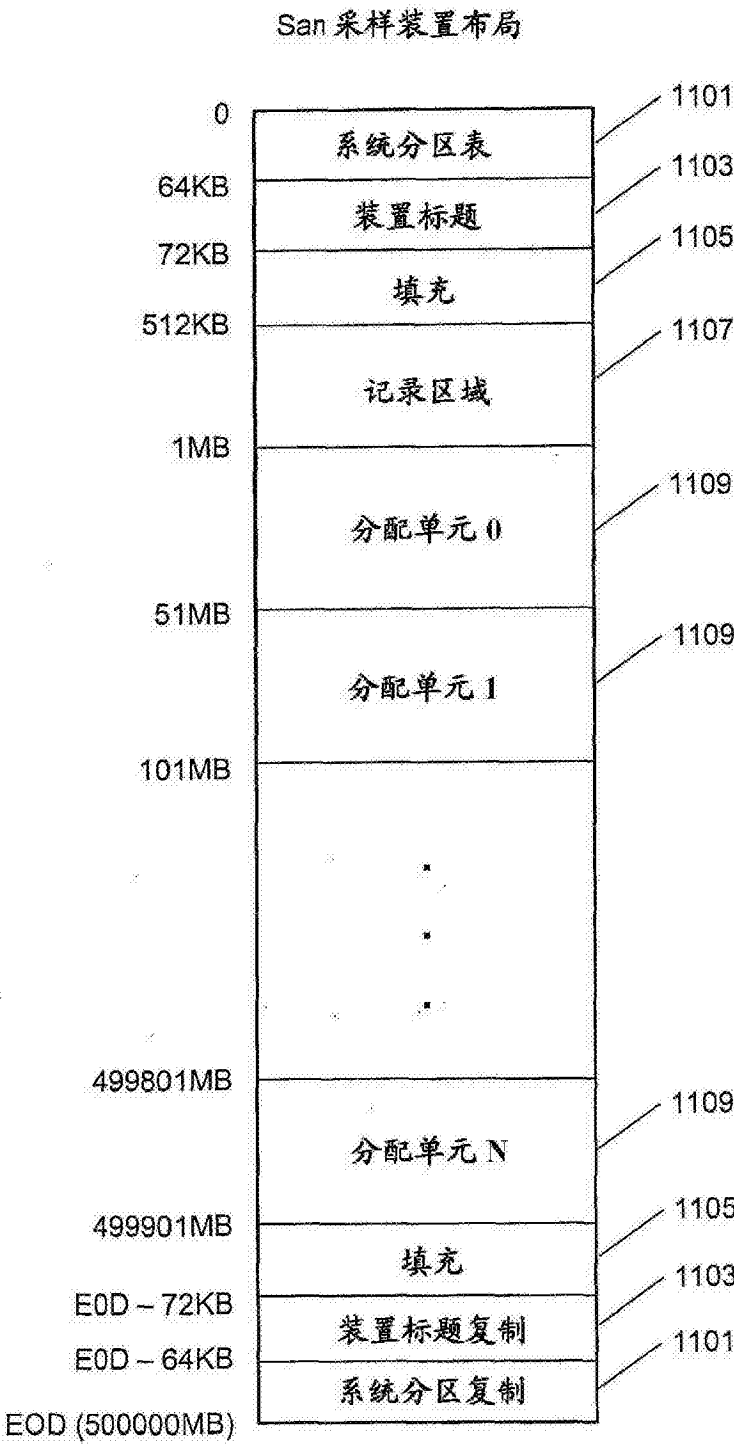


图 11A

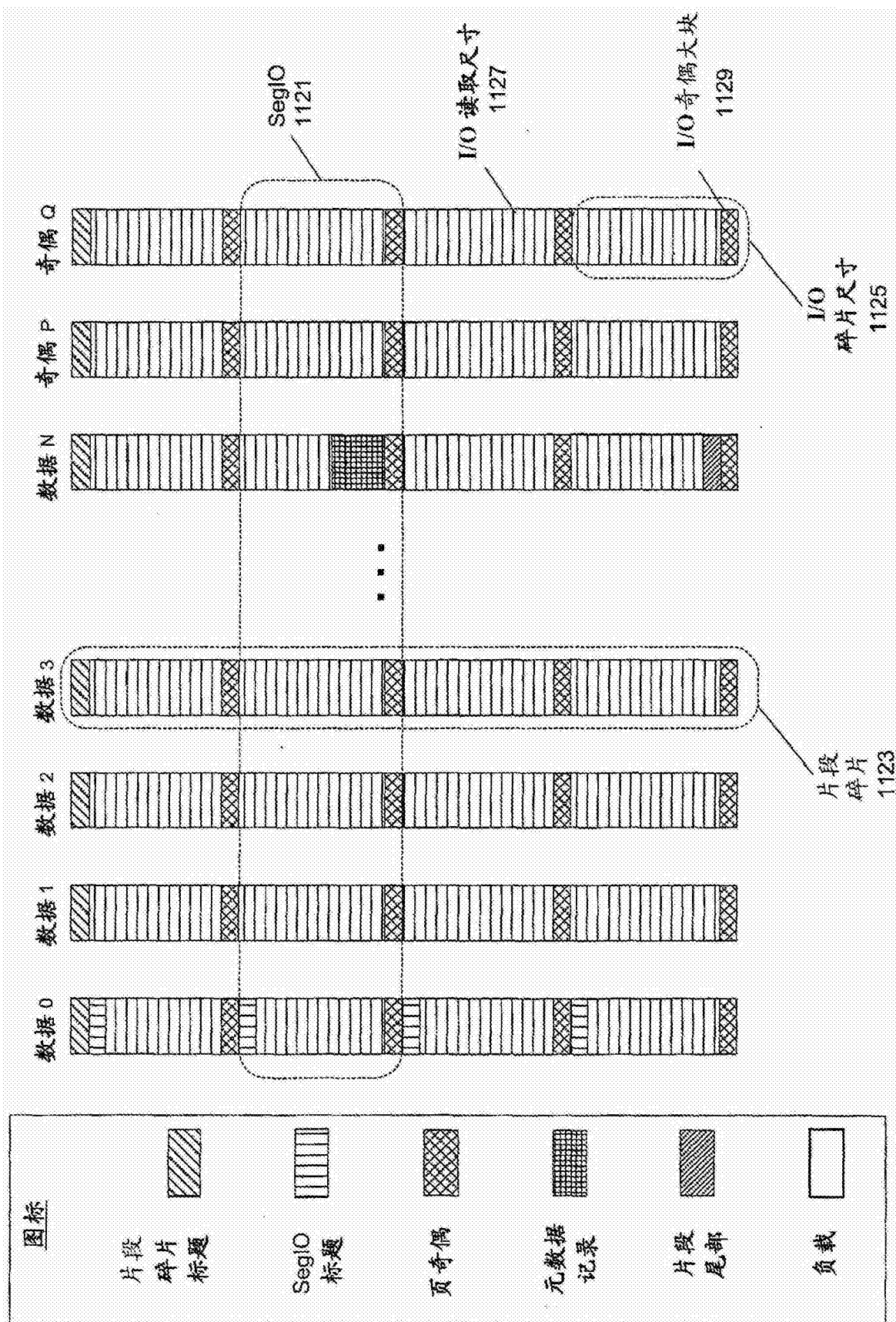


图 11B

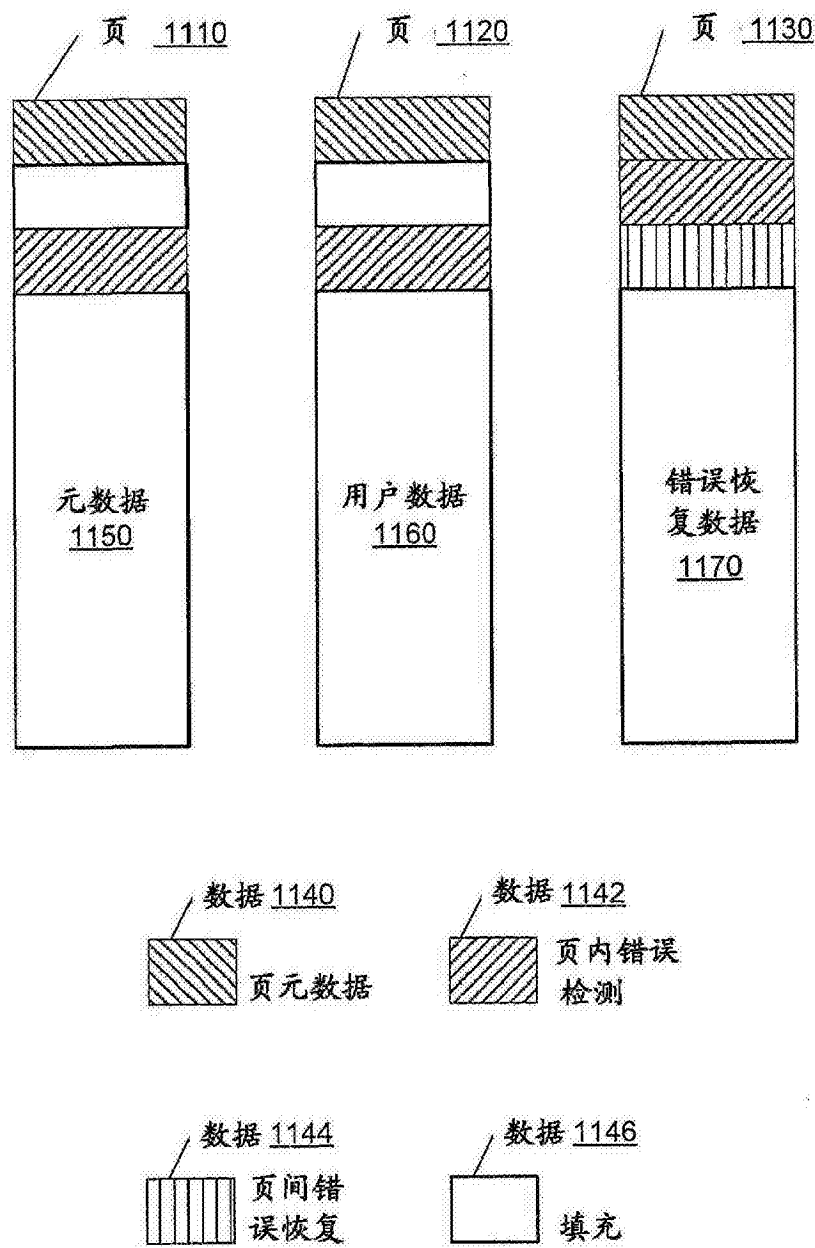


图 11C

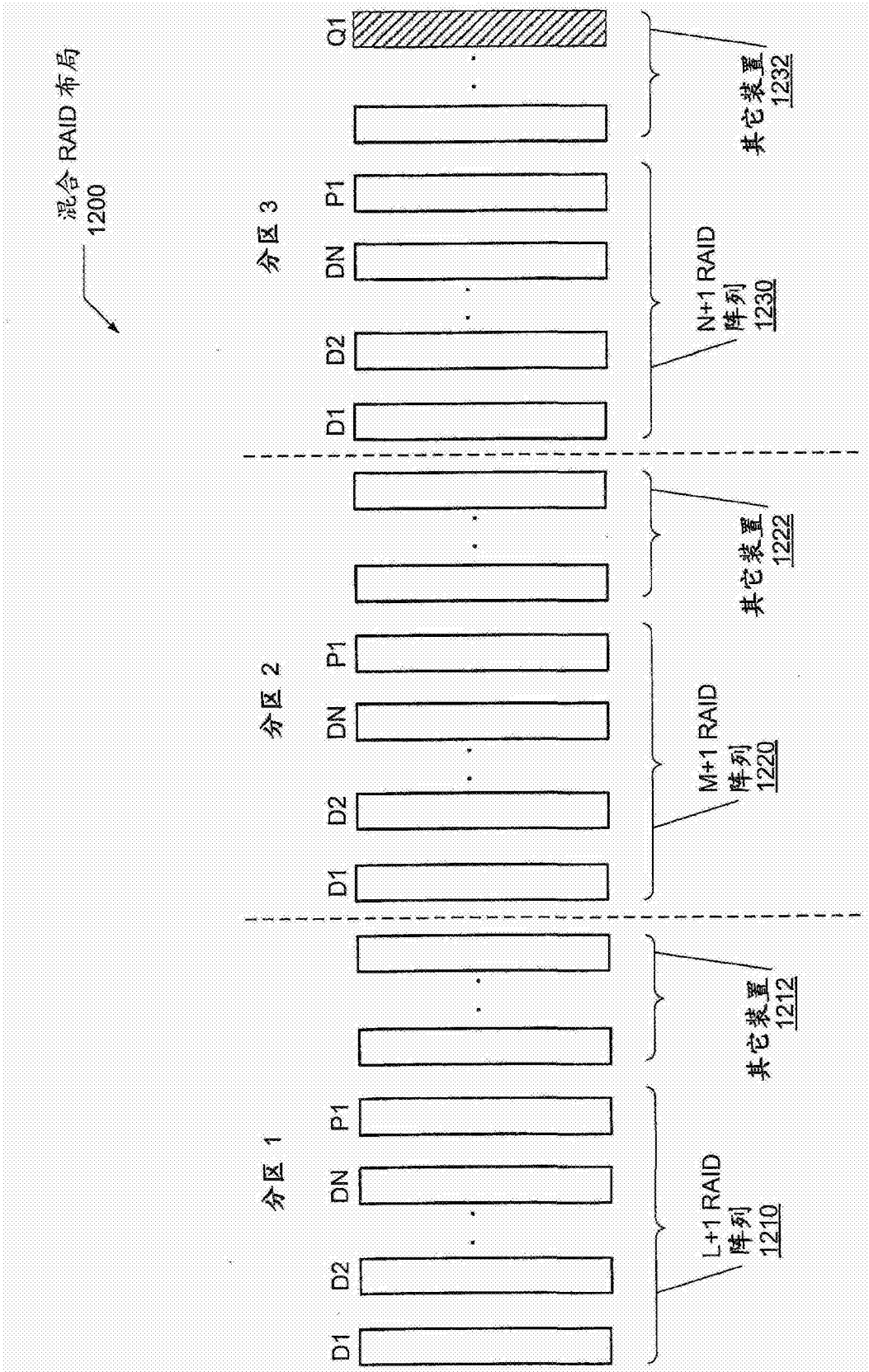


图 12

