



(12) 发明专利申请

(10) 申请公布号 CN 103970481 A

(43) 申请公布日 2014. 08. 06

(21) 申请号 201310034811. 9

(22) 申请日 2013. 01. 29

(71) 申请人 国际商业机器公司

地址 美国纽约

(72) 发明人 邹波 李川 钱海波 谢芳

(74) 专利代理机构 北京市中咨律师事务所

11247

代理人 周良玉 于静

(51) Int. Cl.

G06F 3/06 (2006. 01)

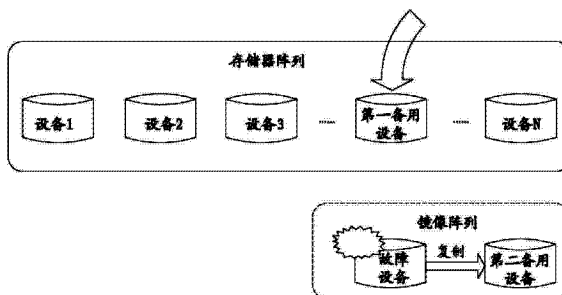
权利要求书2页 说明书11页 附图5页

(54) 发明名称

重建存储器阵列的方法和装置

(57) 摘要

本发明公开了一种重建存储器阵列的方法和装置。所述方法包括：响应于在存储器阵列中出现故障存储器设备，用第一备用存储器设备替换故障存储器设备；利用存储器阵列中除第一备用存储器设备之外的其他存储器设备进行部件重建，从而在第一备用存储器设备中恢复故障存储器设备中的数据；与此并行地，利用第二备用存储器设备进行智能重建，从而将故障存储器设备中的数据复制到第二备用存储器设备；一旦智能重建完成，在存储器阵列中用第二备用存储器设备替换第一备用存储器设备。所述装置与上述方法对应。利用本发明实施例的方法和装置，存储器阵列并行地进行部件重建和智能重建，从而可以快速恢复数据，且避免智能重建不成功带来的影响。



1. 一种重建存储器阵列的方法,所述方法包括:

响应于在所述存储器阵列中出现故障存储器设备,用第一备用存储器设备替换所述故障存储器设备;

利用所述存储器阵列中除第一备用存储器设备之外的其他存储器设备进行部件重建,从而在所述第一备用存储器设备中恢复所述故障存储器设备中的数据;

与上述部件重建并行地,利用所述存储器阵列之外的第二备用存储器设备进行智能重建,从而将所述故障存储器设备中的数据复制到所述第二备用存储器设备;

响应于所述智能重建完成,在所述存储器阵列中用所述第二备用存储器设备替换所述第一备用存储器设备。

2. 根据权利要求1的方法,其中用第一备用存储器设备替换所述故障存储器设备包括,将所述故障存储器设备移出所述存储器阵列,将所述第一备用存储器设备纳入到该存储器阵列,并使得该第一备用存储器设备顶替所述故障存储器设备在存储器阵列中的位置。

3. 根据权利要求1的方法,还包括,检测所述智能重建的执行状态。

4. 根据权利要求1的方法,还包括,响应于所述智能重建失败,终止所述智能重建。

5. 根据权利要求4的方法,其中所述终止智能重建包括,将所述第二备用存储器设备移出与所述故障存储器设备构成的镜像阵列,并将其释放为空闲的备用存储器设备。

6. 根据权利要求1的方法,其中用所述第二备用存储器设备替换所述第一备用存储器设备包括,将所述第一备用存储器设备移出存储器阵列,将所述第二备用存储器设备纳入到该存储器阵列,并使得该第二备用存储器设备顶替第一备用存储器设备在存储器阵列中的位置。

7. 根据权利要求1的方法,还包括,响应于用所述第二备用存储器设备替换所述第一备用存储器设备完成,更新第二备用存储器设备中的数据。

8. 根据权利要求7的方法,其中所述更新第二备用存储器设备中的数据包括,修正所述第二备用存储器设备中的错误数据。

9. 根据权利要求7的方法,其中所述更新第二备用存储器设备中的数据包括,在第二备用存储器设备中恢复在所述智能重建过程中新写入到第一备用存储器设备的数据。

10. 根据权利要求7的方法,其中所述更新第二备用存储器设备中的数据包括,更新所述第二备用存储器设备的位图信息。

11. 一种重建存储器阵列的装置,所述装置包括:

第一替换单元,配置为,响应于在所述存储器阵列中出现故障存储器设备,用第一备用存储器设备替换所述故障存储器设备;

部件重建单元,配置为利用所述存储器阵列中除第一备用存储器设备之外的其他存储器设备进行部件重建,从而在所述第一备用存储器设备中恢复所述故障存储器设备中的数据;

智能重建单元,配置为,与上述部件重建并行地,利用所述存储器阵列之外的第二备用存储器设备进行智能重建,从而将所述故障存储器设备中的数据复制到所述第二备用存储器设备;

第二替换单元,配置为,响应于所述智能重建完成,在所述存储器阵列中用所述第二备

用存储器设备替换所述第一备用存储器设备。

12. 根据权利要求 11 的装置,其中所述第一替换单元配置为,将所述故障存储器设备移出所述存储器阵列,将所述第一备用存储器设备纳入到该存储器阵列,并使得该第一备用存储器设备顶替所述故障存储器设备在存储器阵列中的位置。

13. 根据权利要求 11 的装置,还包括,检测单元,配置为检测所述智能重建的执行状态。

14. 根据权利要求 11 的装置,还包括,终止单元,配置为响应于所述智能重建失败,终止所述智能重建。

15. 根据权利要求 14 的装置,其中所述终止单元配置为,将所述第二备用存储器设备移出与所述故障存储器设备构成的镜像阵列,并将其释放为空闲的备用存储器设备。

16. 根据权利要求 11 的装置,其中所述第二替换单元配置为,将所述第一备用存储器设备移出存储器阵列,将所述第二备用存储器设备纳入到该存储器阵列,并使得该第二备用存储器设备顶替第一备用存储器设备在存储器阵列中的位置。

17. 根据权利要求 11 的装置,还包括更新单元,配置为,响应于所述第二替换单元完成替换,更新第二备用存储器设备中的数据。

18. 根据权利要求 17 的装置,其中所述更新单元进一步配置为,修正所述第二备用存储器设备中的错误数据。

19. 根据权利要求 17 的装置,其中所述更新单元进一步配置为,在第二备用存储器设备中恢复在所述智能重建过程中新写入到第一备用存储器设备的数据。

20. 根据权利要求 17 的装置,其中所述更新单元进一步配置为,更新所述第二备用存储器设备的位图信息。

重建存储器阵列的方法和装置

技术领域

[0001] 本发明涉及存储器阵列,更具体而言,涉及重建存储器阵列的方法和装置。

背景技术

[0002] 随着信息技术的快速发展,需要存储和处理的数据量越来越庞大。为此,在增大单个存储器设备的存储密度和存储容量的同时,往往还采用由多个存储器设备构成的存储器阵列来存储数据。典型地,存储器阵列由多个独立的非易失性存储器设备构成,例如磁盘、SSD 等设备;这些存储器设备共同连接到存储器阵列控制器,在该控制器的控制下执行与数据存储相关的操作。

[0003] 另一方面,为了确保存储数据的安全性,通常在存储器阵列中提供一定的冗余度,从而能够在部分数据出现损坏时能够进行数据修复。这样的存储器阵列又称为冗余磁盘阵列 RAID。现有技术中已经提供了多个级别的 RAID。

[0004] RAID1 又称为磁盘镜像阵列。在这样的阵列中,在主磁盘上存储数据的同时也在镜像磁盘上写同样的数据。当主磁盘损坏时,镜像磁盘则代替主磁盘进行工作。因为有镜像磁盘进行完全的数据备份,所以 RAID1 的数据安全性在所有的 RAID 级别中是最高的。但是,可以理解,RAID1 的磁盘利用率较低。

[0005] RAID2 通过引入错误修正码 ECC 将数据进行编码,然后将编码后的数据分区为独立的比特,写入磁盘中。RAID3 和 RAID4 进一步地利用数据交错(interleaving)存储技术,将编码后的数据进行分区,分别存储在磁盘中,并将同比特的校验数据存储单独磁盘中。

[0006] RAID5 是储存性能、数据安全和存储成本兼顾的存储解决方案。RAID5 通过把数据条带化(striping)分布到不同的存储设备上来提高数据访问的并行性。具体地,在 RAID5 中,将数据和相对应的校验信息存储到组成 RAID5 的各个磁盘上,并且校验信息和相对应的数据分别存储于不同的磁盘上。由于 RAID5 在每个条带中采用一个校验块来存储校验信息,因此 RAID5 能够容忍一个磁盘出现故障。也就是说,当一个磁盘中的数据发生损坏后,可以利用剩下的磁盘中的数据和相应的校验信息来恢复被损坏的数据。由于 RAID5 兼顾了数据安全性和存储成本,因此应用比较广泛。

[0007] RAID6 通过将每个条带中的校验块增加到两个来增加数据安全性。相应地,RAID6 能够容忍两个磁盘同时出现故障。此外,还提供了 RAID10 和 RAID50 等其他级别的冗余磁盘阵列,他们在数据安全性、磁盘利用率、读写速度等不同方面具有各自的特点。

[0008] 如前所述,RAID 阵列由于其冗余度而具有数据恢复能力。恢复 RAID 中故障磁盘中的数据的过程又称为重建(rebuild)。图 1A 示意性示出 RAID5 中的数据块的重建。在具有 N 个存储器设备(例如磁盘)的 RAID5 中,每一条带中均具有 N-1 个数据块和一个校验块。当某个数据块 D_n 发生损坏,可以利用同一条带中的其他数据块 D_i (i 不等于 n) 和相应的校验块 P 计算得到损坏的数据块 D_n 。如果损坏的是校验块,则可以通过再次对同一条带中的数据块进行校验运算而重新获得该校验块。因此,当阵列中任一磁盘出现故障,均可以利用剩下的磁盘中的数据来恢复故障磁盘中的数据。这样的重建过程又称为部件重建。部件

重建一般不影响 RAID 阵列与主机的输入和输出 (IO)。但是,可以理解,由于需要读取各个磁盘中的数据并进行运算,部件重建需要花费较长的时间(通常为几个小时)。为此,还提出了智能重建作为补充,以快速地重建故障磁盘中的数据。

[0009] 图 1B 示出智能重建的示意图。智能重建主要应用于某个存储器设备开始出现故障但仍然能够进行存取的情况。如图 1B 所示,假定在由 N 个存储器设备(例如磁盘)构成的 RAID5 阵列中,磁盘 n 开始出现故障,例如出现介质错误。为了避免部件重建,在磁盘 n 仍然能够进行存取的情况下,在该磁盘 n 和一备用磁盘之间建立镜像关系,也就是使得磁盘 n 和备用磁盘构成 RAID1 阵列,从而将磁盘 n 的数据复制到备用磁盘。此时,磁盘 n 同时属于 RAID5 阵列(原阵列)和 RAID1 阵列(镜像阵列)。尽管图 1B 仅示出了 RAID5 阵列作为例子,但是智能重建也可以类似地应用于其他 RAID 阵列,例如 RAID6、RAID10 等。由于智能重建仅涉及故障磁盘 n 和备用磁盘之间的数据拷贝,因此重建过程要比部件重建快得多。

[0010] 然而,可以理解,在进行智能重建过程中,需要频繁地存取故障磁盘以从中复制数据。这经常会使得已经出现介质错误的故障磁盘加速损坏。因此,常常出现这样的情况,在智能重建尚未完成的时候,故障磁盘进一步损坏,无法从中读取数据,于是智能重建不得不终止。如前所述,在开始进行智能重建时,在故障磁盘和备用磁盘之间建立了镜像关系。镜像关系的建立涉及许多配置数据的写入,包括原 RAID 阵列的元数据、镜像阵列的元数据、各种位图(bitmap)数据等等。相应地,为了终止智能重建,就需要解除在故障磁盘和备用磁盘之间建立的镜像关系,清除以上所述的配置数据。为了避免引入进一步的复杂性,在清除配置数据期间,往往需要暂停(quiesce)原 RAID 阵列与主机的 IO,以确保故障磁盘尽快清理与备用磁盘之间的关联。在故障磁盘损坏严重的情况下,上述清理过程需要较长时间来执行。在此期间,存储器阵列 RAID 与主机的 IO 完全被抑制,从而使得其读写受到严重影响。

[0011] 因此,希望提出更有利的重建方案,能够在恢复 RAID 阵列中损坏数据的过程中,减小对阵列的影响。

发明内容

[0012] 考虑到现有技术中存在的问题,提出本发明,旨在提供更有利的存储器阵列重建方案。

[0013] 根据本发明的一个方面,提供了一种重建存储器阵列的方法,所述方法包括:响应于在所述存储器阵列中出现故障存储器设备,用第一备用存储器设备替换所述故障存储器设备;利用所述存储器阵列中除第一备用存储器设备之外的其他存储器设备进行部件重建,从而在所述第一备用存储器设备中恢复所述故障存储器设备中的数据;与上述部件重建并行地,利用所述存储器阵列之外的第二备用存储器设备进行智能重建,从而将所述故障存储器设备中的数据复制到所述第二备用存储器设备;响应于所述智能重建完成,在所述存储器阵列中用所述第二备用存储器设备替换所述第一备用存储器设备。

[0014] 根据本发明的另一个方面,提供了一种重建存储器阵列的装置,所述装置包括:第一替换单元,配置为响应于在存储器阵列中出现故障存储器设备,用第一备用存储器设备替换所述故障存储器设备;部件重建单元,配置为利用所述存储器阵列中除第一备用存储器设备之外的其他存储器设备进行部件重建,从而在所述第一备用存储器设备中恢复所述

故障存储器设备中的数据；智能重建单元，配置为，与上述部件重建并行地，利用所述存储器阵列之外的第二备用存储器设备进行智能重建，从而将所述故障存储器设备中的数据复制到所述第二备用存储器设备；第二替换单元，配置为，响应于所述智能重建完成，在所述存储器阵列中用所述第二备用存储器设备替换所述第一备用存储器设备。

[0015] 利用本发明实施例的方法和装置，存储器阵列可以并行地在两个备用存储器设备中分别进行部件重建和智能重建。在智能重建成功的情况下，存储器阵列利用智能重建快速恢复故障存储器设备中的数据；即使智能重建不成功，存储器阵列可以常规地进行部件重建，而不会受到终止智能重建所带来的影响。从而，存储器阵列可以更加快速、灵活且安全地进行数据重建和恢复。

附图说明

[0016] 通过结合附图对本公开示例性实施方式进行更详细的描述，本公开的上述以及其它目的、特征和优势将变得更加明显，其中，在本公开示例性实施方式中，相同的参考标号通常代表相同部件。

[0017] 图 1A 示意性示出 RAID5 中的数据块的重建；

[0018] 图 1B 示出智能重建的示意图；

[0019] 图 2 示出了适于用来实现本发明实施方式的示例性计算机系统 / 服务器 12 的框图；

[0020] 图 3 示出根据本发明一个实施例的重建存储器阵列的方法的流程图；

[0021] 图 4 示出根据一个实施例的存储器设备的数据存储结构；

[0022] 图 5 示出重建过程中存储器阵列的结构示意图；

[0023] 图 6A 示出经过部件重建的存储器阵列的结构示意图；

[0024] 图 6B 示出经过智能重建的存储器阵列的结构示意图；以及

[0025] 图 7 示出根据本发明一个实施例的装置的框图。

具体实施方式

[0026] 下面将参照附图更详细地描述本公开的优选实施方式。虽然附图中显示了本公开的优选实施方式，然而应该理解，可以以各种形式实现本公开而不应被这里阐述的实施方式所限制。相反，提供这些实施方式是为了使本公开更加透彻和完整，并且能够将本公开的范围完整地传达给本领域的技术人员。

[0027] 所属技术领域的技术人员知道，本发明可以实现为系统、方法或计算机程序产品。因此，本公开可以具体实现为以下形式，即：可以是完全的硬件、也可以是完全的软件（包括固件、驻留软件、微代码等），还可以是硬件和软件结合的形式，本文一般称为“电路”、“模块”或“系统”。此外，在一些实施例中，本发明还可以实现为在一个或多个计算机可读介质中的计算机程序产品的形式，该计算机可读介质中包含计算机可读的程序代码。

[0028] 可以采用一个或多个计算机可读的介质的任意组合。计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子（非穷举的列表）包括：具有一个或多个导线的电连接、便

便携式计算机磁盘、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)、光纤、便携式紧凑磁盘只读存储器 (CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本文件中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0029] 计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括——但不限于——电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0030] 计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括——但不限于——无线、电线、光缆、RF 等等,或者上述的任意合适的组合。

[0031] 可以以一种或多种程序设计语言或其组合来编写用于执行本发明操作的计算机程序代码,所述程序设计语言包括面向对象的程序设计语言——诸如 Java、Smalltalk、C++,还包括常规的过程式程序设计语言——诸如“C”语言或类似的设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、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。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络——包括局域网 (LAN) 或广域网 (WAN)——连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0032] 下面将参照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,都可以由计算机程序指令实现。这些计算机程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器,从而生产出一种机器,这些计算机程序指令通过计算机或其它可编程数据处理装置执行,产生了实现流程图和/或框图中的方框中规定的功能/操作的装置。

[0033] 也可以把这些计算机程序指令存储在能使得计算机或其它可编程数据处理装置以特定方式工作的计算机可读介质中,这样,存储在计算机可读介质中的指令就产生出一个包括实现流程图和/或框图中的方框中规定的功能/操作的指令装置(instruction means)的制品(manufacture)。

[0034] 也可以把计算机程序指令加载到计算机、其它可编程数据处理装置、或其它设备上,使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤,以产生计算机实现的过程,从而使得在计算机或其它可编程装置上执行的指令能够提供实现流程图和/或框图中的方框中规定的功能/操作的过程。

[0035] 图 2 示出了适于用来实现本发明实施方式的示例性计算机系统/服务器 12 的框图。图 2 显示的计算机系统/服务器 12 仅仅是一个示例,不应对本发明实施例的功能和使用范围带来任何限制。

[0036] 如图 2 所示,计算机系统/服务器 12 以通用计算设备的形式表现。计算机系统/服务器 12 的组件可以包括但不限于:一个或者多个处理器或者处理单元 16,系统存储器 28,连接不同系统组件(包括系统存储器 28 和处理单元 16)的总线 18。

[0037] 总线 18 表示几类总线结构中的一种或多种,包括存储器总线或者存储器控制器,外围总线,图形加速端口,处理器或者使用多种总线结构中的任意总线结构的局域总线。举例来说,这些体系结构包括但不限于工业标准体系结构(ISA)总线,微通道体系结构(MAC)总线,增强型 ISA 总线、视频电子标准协会(VESA)局域总线以及外围组件互连(PCI)总线。

[0038] 计算机系统/服务器 12 典型地包括多种计算机系统可读介质。这些介质可以是任何能够被计算机系统/服务器 12 访问的可用介质,包括易失性和非易失性介质,可移动的和不可移动的介质。

[0039] 系统存储器 28 可以包括易失性存储器形式的计算机系统可读介质,例如随机存取存储器(RAM)30 和/或高速缓存存储器 32。计算机系统/服务器 12 可以进一步包括其它可移动/不可移动的、易失性/非易失性计算机系统存储介质。仅作为举例,存储系统 34 可以用于读写不可移动的、非易失性磁介质(图 3 未显示,通常称为“硬盘驱动器”)。尽管图 2 中未示出,可以提供用于对可移动非易失性磁盘(例如“软盘”)读写的磁盘驱动器,以及对可移动非易失性光盘(例如 CD-ROM, DVD-ROM 或者其它光介质)读写的光盘驱动器。在这些情况下,每个驱动器可以通过一个或者多个数据介质接口与总线 18 相连。存储器 28 可以包括至少一个程序产品,该程序产品具有一组(例如至少一个)程序模块,这些程序模块被配置以执行本发明各实施例的功能。

[0040] 具有一组(至少一个)程序模块 42 的程序/实用工具 40,可以存储在例如存储器 28 中,这样的程序模块 42 包括——但不限于——操作系统、一个或者多个应用程序、其它程序模块以及程序数据,这些示例中的每一个或某种组合中可能包括网络环境的实现。程序模块 42 通常执行本发明所描述的实施例中的功能和/或方法。

[0041] 计算机系统/服务器 12 也可以与一个或多个外部设备 14 (例如键盘、指向设备、显示器 24 等)通信,还可与一个或者多个使得用户能与该计算机系统/服务器 12 交互的设备通信,和/或与使得该计算机系统/服务器 12 能与一个或多个其它计算设备进行通信的任何设备(例如网卡,调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口 22 进行。并且,计算机系统/服务器 12 还可以通过网络适配器 20 与一个或者多个网络(例如局域网(LAN),广域网(WAN)和/或公共网络,例如因特网)通信。如图所示,网络适配器 20 通过总线 18 与计算机系统/服务器 12 的其它模块通信。应当明白,尽管图中未示出,可以结合计算机系统/服务器 12 使用其它硬件和/或软件模块,包括但不限于:微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID 系统、磁带驱动器以及数据备份存储系统等。

[0042] 以下结合附图和具体例子描述本发明的实施方式。根据本发明的实施例,在存储器阵列中出现故障存储器设备的情况下,用第一备用存储器设备替换该故障存储器设备来进行部件重建。与此同时地,在已经脱离存储器阵列的故障存储器设备和第二备用存储器设备之间建立镜像关系,并将故障存储器设备中的数据复制到第二备用存储器设备。换言之,与原存储器阵列的部件重建并行地进行智能重建。一旦智能重建顺利完成,就用第二备用存储器设备替换第一存储器设备。如果智能重建不能顺利完成,那么就终止智能重建的过程。由于故障存储器设备已经被移出原存储器阵列,上述智能重建过程的终止不会影响原存储器阵列的读写;相应地,原存储器阵列继续进行部件重建,直到重建完成。由此,存储器阵列可以利用智能重建来快速恢复故障存储器设备中的数据;即使智能重建不成功,终止智能重建的过程也不会影响存储器阵列,从而实现灵活高效且安全的数据重建。下面

具体描述上述发明构思的实现方式。

[0043] 现在参看图 3, 其示出根据本发明一个实施例的重建存储器阵列的方法的流程图。可以理解, 上述存储器阵列由多个独立的非易失性存储器设备构成, 例如磁盘、SSD 等设备; 并且, 存储器阵列具有一定冗余度, 从而具有数据恢复能力。当存储器阵列中某个存储器设备出现故障时, 可以利用图 3 所示的方法流程来恢复故障存储器设备中的数据, 从而重建该存储器阵列。具体地, 如图 3 所示, 重建方法包括如下步骤: 步骤 30, 响应于在存储器阵列中出现故障存储器设备, 用第一备用存储器设备替换所述故障存储器设备; 步骤 32, 利用存储器阵列中除第一备用存储器设备之外的其他存储器设备进行部件重建, 从而在所述第一备用存储器设备中恢复所述故障存储器设备中的数据; 步骤 34, 与上述步骤 32 并行地, 利用存储器阵列之外的第二备用存储器设备进行智能重建, 从而将故障存储器设备中的数据复制到第二备用存储器设备; 以及步骤 36, 响应于所述智能重建完成, 在所述存储器阵列中用所述第二备用存储器设备替换所述第一备用存储器设备。下面结合具体例子描述上述各个步骤的执行。

[0044] 具体地, 首先, 在步骤 30, 响应于在存储器阵列中出现故障存储器设备, 用第一备用存储器设备替换故障存储器设备。可以理解, 上述故障存储器设备是已经出现一定量的介质错误但仍然能够进行存取的存储器设备。如本领域技术人员所知, 介质错误(Media Error) 一般是指存储器设备的介质(例如磁盘) 读写错误。介质错误可以反映出存储器设备的存储介质的工作情况。出现介质错误并不意味着存储器设备必须进行更换, 因为存储器设备能够对坏块(killsector) 进行屏蔽和迁移。但是一般地, 应根据介质错误对存储器设备进行诊断, 由此判断该存储器设备是否能够继续在存储器阵列中进行工作。相应地, 在一个实施例中, 上述重建方法还包括, 检测存储器阵列中的故障存储器设备。可以利用现有技术中的多种方式来检测上述故障存储器设备。在一个例子中, 存储器设备的制造商可能针对存储器设备能够容忍的介质错误设定有一个报警阈值。当存储器设备上出现的介质错误超过预设的报警阈值, 该存储器设备就会发出警告。相应地, 可以监视存储器阵列中各个存储器设备的运行状态, 将发出警告的存储器设备确定为故障存储器设备, 从而实现故障存储器设备的检测。在另一例子中, 由存储器阵列的控制器统计各个存储器设备上出现的介质错误, 并将介质错误超过一阈值的存储器设备确定为故障存储器设备。此外, 还可以通过现有技术中的其他方式来检测存储器阵列中的故障存储器设备。

[0045] 一旦检测到故障存储器设备, 在步骤 30, 用第一备用存储器设备替换该故障存储器设备。在一个实施例中, 上述替换包括, 将故障存储器设备移出存储器阵列, 将空闲的第一备用存储器设备纳入到该存储器阵列, 并使得该第一备用存储器设备在存储器阵列中发挥故障存储器设备原先所起的作用。为了实现上述替换, 存储器阵列的控制器需要修改控制器中存储的阵列的控制信息。此外, 还需要修改各个存储器设备中所存储的阵列配置信息。

[0046] 图 4 示出根据一个实施例的存储器设备的数据存储结构。该存储器设备可以是存储器阵列中的任一存储器设备。如图 4 所示, 存储器设备被划分为用户区和保留区, 用户区中存储有用户数据和阵列位图数据, 保留区中存储有阵列元数据和底层保留数据。在这些数据中, 阵列位图数据用位图的形式记录存储器设备中的介质使用状态。阵列元数据记录存储器阵列的配置信息, 例如阵列中包含的存储器设备的数目、各个存储器设备的标识

等。由于阵列元数据记录了存储器阵列的配置信息,因此在进行存储器设备的替换时,需要修改该阵列元数据,更具体地,将有关故障存储器设备的信息修改为第一备用存储器设备的信息。尽管图 4 示出了典型的数据存储结构,但是存储器设备还有可能采用其他的结构进行数据存储。例如,存储器设备可以进一步划分配置数据,并将其分别存储于不同位置。相应地,可以与存储结构相对应地修改配置信息,从而实现存储器设备的替换。通过上述替换,故障存储器设备完全脱离原存储器阵列,不再作为存储器阵列中的成员;而第一备用存储器设备顶替故障存储器设备,成为存储器阵列的成员。

[0047] 接着,在步骤 30 进行替换的基础上,并行执行步骤 32 和 34。

[0048] 在步骤 32 中,利用存储器阵列中除第一备用存储器设备之外的其他存储器设备进行部件重建,从而在所述第一备用存储器设备中恢复故障存储器设备中的数据。如前所述,存储器阵列具有一定冗余度,从而具有数据恢复能力。具体地,在存储器阵列中,除新纳入的第一备用存储器设备之外,其他存储器设备均正常地存储有数据块和校验块。通过对这些正常工作的存储器设备中的数据块和校验块进行运算,就可以获得被移除的故障存储器设备中存储的数据。恢复的数据被存储在所述第一备用存储器设备中。可以理解,该过程与现有技术中常规的部件重建相似,因此不再对其进行详细描述。

[0049] 与步骤 32 并行地,在步骤 34 中,利用存储器阵列之外的第二备用存储器设备进行智能重建。更具体而言,步骤 34 的智能重建是指,将故障存储器设备和第二备用存储器设备构成镜像阵列,并将故障存储器设备中的数据复制到第二备用存储器设备中的过程。可以理解,步骤 32 的过程与常规的智能重建的执行过程相似,因此,在下文中仍然将其称为智能重建。但是应理解,与现有技术的智能重建不同的是,步骤 34 的智能重建是在存储器阵列之外进行的,因而与存储器阵列的操作相隔离。

[0050] 图 5 示出重建过程中存储器阵列的结构示意图。如图 5 所示,通过步骤 30,故障存储器设备被移出存储器阵列,并且第一备用存储器设备顶替故障存储器设备的位置,成为存储器阵列的成员。由此,可以在存储器阵列中进行部件重建。另一方面,被移出存储器阵列的故障存储器设备与第二备用存储器设备构成镜像阵列,以执行智能重建。与现有技术中故障存储器设备同时属于原存储器阵列和用于智能重建的镜像阵列不同,在图 5 中,故障存储器设备完全脱离存储器阵列,仅仅位于镜像阵列中。于是,存储器阵列中的部件重建和镜像阵列中的智能重建同时并行地进行,互不影响,互不干扰。存储器阵列甚至并不知晓智能重建的进行。

[0051] 如本领域技术人员所知,在部件重建过程中,存储器阵列仍然可以响应于主机的命令进行数据读写。取决于存储器阵列的配置和与主机之间的 I/O,部件重建通常要花费几个小时的时间。而上述智能重建并不存在与主机之间的 I/O 压力,并且仅涉及数据的复制,因此如果能够成功完成的话,通常只需要几十分钟的时间,比部件重建要快得多。以写入速度为 480MB/s 的 600G 硬盘为例,考虑到 I/O 压力,部件重建通常需要 4-6 小时。如果如图 5 所示独立地进行上述智能重建,在全速写入的理想情况下,仅需要 $600G/480M/60=21$ 分钟的时间。

[0052] 由于在多数情况下,上述智能重建的过程会远远早于部件重建而完成,因此在步骤 32 和 34 并行地开始部件重建和智能重建之后,检测故障存储器设备所进行的智能重建的执行状态。为此,在一个实施例,中,监视故障存储器设备的工作状态。如果故障存储器设

备正在进行向第二备用存储器设备的数据复制,则确定智能重建的过程仍在进行中;如果故障存储器设备已经完成全部数据的复制,则确定智能重建成功完成;如果故障存储器设备由于损坏严重而停止数据的复制,则确定智能重建失败。由此,通过监视故障存储器设备的工作状态来检测智能重建的执行状态。在另一实施例中,故障存储器设备被配置为,在完成数据复制和/或被迫停止数据复制时发出报告。通过接收来自故障存储器设备的报告,可以检测智能重建的执行状态。此外,本领域技术人员可以采用其他方式来检测上述智能重建的执行状态。

[0053] 在一个实施例中,图3的重建方法还包括,响应于上述智能重建失败,也就是响应于从故障存储器设备到第二备用存储器设备的数据复制失败,终止智能重建。在一个实施例中,终止智能重建包括,清除故障存储器设备和第二备用存储器设备之间的镜像关系。在另一实施例中,直接将第二备用存储器设备移出与故障存储器设备构成的镜像阵列,将其释放为空闲的备用存储器设备,而不再进行任何元数据、位图数据等配置数据的操作。不论采用哪种方式,终止智能重建的过程都与原存储器阵列相隔离,不会对原存储器阵列带来任何影响。相应地,原存储器阵列继续进行其部件重建,直到重建完成。

[0054] 图6A示出经过部件重建的存储器阵列的结构示意图。如图所示,在智能重建失败的情况下,存储器阵列通过部件重建进行数据恢复。最终重建的存储器阵列是将故障存储器设备替换为第一备用存储器设备的阵列。相应地,在暂时地使用第二备用存储器设备之后,该第二备用存储器设备重新成为空闲的备用存储器设备。在整个过程中,存储器阵列独立地进行部件重建,完全不知晓智能重建开始和终止的过程,从而避免了由智能重建带来的影响。

[0055] 另一方面,如果智能重建成功完成,那么第二备用存储器设备就已经获得了故障存储器设备中存储的全部数据。由于故障存储器设备中的数据已得到恢复,存储器阵列中的部件重建就没有必要继续进行。因此,在步骤36,一旦智能重建成功完成,也就是从故障存储器设备到第二备用存储器设备的数据复制完成,就终止利用第一备用存储器设备进行的部件重建,并用第二备用存储器设备替换第一备用存储器设备。与步骤30的替换类似地,步骤36包括,将第一备用存储器设备移出存储器阵列,将第二备用存储器设备纳入到该存储器阵列,并使得该第二备用存储器设备顶替第一备用存储器设备在存储器阵列中的位置。因此,在智能重建成功的情况下,在步骤30和步骤36执行两次替换,经过这两次替换,最终由第二备用存储器设备替换了最初的故障存储器设备的位置。相应地,第一备用存储器设备被移出存储器阵列,重新作为空闲的备用存储器设备。

[0056] 图6B示出经过智能重建的存储器阵列的结构示意图。如图所示,在智能重建成功的情况下,最终重建的存储器阵列是将故障存储器设备替换为第二备用存储器设备的阵列。相应地,第一备用存储器设备在暂时地被用于部件重建之后被释放,重新成为空闲的备用存储器设备。整个重建过程与智能重建的时间相当。由此,存储器阵列利用智能重建快速地恢复了故障存储器设备中的数据;同时,由于智能重建与存储器阵列相隔离地进行,确保了存储器阵列不会受到智能重建的影响。

[0057] 在一个实施例中,在将第二备用存储器设备纳入到存储器阵列的基础上,进一步考虑重建过程中可能产生的数据完整性问题。该数据完整性问题可能由于以下因素而引入。由于第二备用存储器设备直接从故障存储器设备复制数据,而故障存储器设备已经存

在一些介质错误,这使得第二备用存储器设备获得的数据中存在一部分错误数据。另一方面,第二备用存储设备在完成智能重建之后才被纳入到存储器阵列。在第二备用存储器进行智能重建期间,存储器阵列一边实施部件重建,一边正常地与主机进行 IO 通信,因此有可能写入了新的数据。这些因素都可能导致第二备用存储器设备中的数据存在完整性的问题。为此,在一个实施例中,图 3 的重建方法还包括,响应于用第二备用存储器设备替换第一备用存储器设备完成,更新第二备用存储器设备中的数据。

[0058] 在一个实施例中,上述更新步骤包括,修正第二备用存储器设备中的错误数据。如之前结合图 4 所述,在存储器阵列的存储器设备中存储有阵列位图数据,该位图数据用位图的形式记录各个存储器设备中的介质使用状态,例如哪些扇区已经使用,哪些扇区尚未使用,哪些扇区出现错误等等。在步骤 34 的复制数据过程中,第二备用存储器设备将故障存储器设备中存储的阵列位图数据也一并复制过来。由此,第二备用存储器设备就包含有这样的阵列位图数据。基于此阵列位图数据,第二备用存储器设备就可以确定复制的数据块中哪些数据块存在错误。相应地,在第二备用存储器设备被纳入到存储器阵列之后,基于如前所述的存储器阵列的数据恢复能力,就可以利用存储器阵列中其他存储器设备中的相应数据块重建上述存在错误的的数据块,从而修正错误数据。可以理解,在其他实施例中,存储器设备还可以采用其他形式来记录介质使用状态,例如可以采用一个坏块表的形式,该坏块表可以包含多个条目,每一条目记录一个坏块的信息。不管采用什么样的形式,存储器设备总是可以对介质使用状态进行记录。基于对故障存储器设备的介质使用状态进行记录的信息,就可以确定存在错误的的数据块,进而修正第二备用存储器设备中的错误数据。

[0059] 在一个实施例中,上述更新步骤还包括,在第二备用存储器设备中恢复在步骤 34 的智能重建过程中新写入到第一备用存储器设备的数据。上述数据的恢复也可以通过参考位图数据来实现。

[0060] 一般地,在利用第一备用存储器设备进行常规的部件重建时,会在第一备用存储器设备中维持一个重建位图和校验位图。在一个例子中,初始地将重建位图中所有位都设定为 1。可以理解,在部件重建过程中,逐个条带地进行数据恢复。每当完成一个条带的的数据重建,就改写重建位图中与该条带对应的位的值,例如将其改写为 0。在此基础上,根据本发明一个实施例,可以在第一备用存储器阵列中进一步创建一个新写入位图。在一个例子中,初始地,该新写入位图的各个位均被设定为“0”。一旦在部件重建过程中从主机新写入了某个条带的的数据,就在该新写入位图中,将与该新写入条带对应的位修改为“1”。在将第一存储器设备移出存储器阵列时,首先从中读取并记录上述新写入位图。基于该新写入位图,就可以确定,在执行部件重建过程中,存储器阵列的哪些条带进行了更新。接着,在将第二备用存储器纳入到存储器阵列之后,就可以利用存储器阵列中的其他存储器设备中同一条带中的数据块来恢复原本写入到第一备用存储器设备中的数据,并将恢复的数据写入到第二备用存储器设备。由此,在第二备用存储器设备中恢复了步骤 34 的数据复制过程中新写入到第一存储器设备的数据。

[0061] 在一个例子中,可以使得第一备用存储器设备将部件重建过程中恢复的数据块和新写入的数据块记录在同一位图文件中。在用第二备用存储器设备替换第一备用存储器设备时,通过对比两个备用存储器设备中的位图数据来确定是否存在新写入的数据块。接着,类似地,在将第一备用存储器设备移出之后,利用其他存储器设备中的数据块,在第二备用

存储器设备中恢复上述新写入的数据块。

[0062] 如前所述,在其他实施例中,也可以采用除位图之外的其他形式来记录存储器设备的介质使用状态,由此确定新写入的数据块的信息。基于确定的新写入的数据块的信息,可以在第二备用存储器设备中恢复新写入的数据。

[0063] 由于以上在第二备用存储器设备中进行了数据修正和数据恢复,因此,在一个实施例中,上述更新步骤还包括,更新第二备用存储器设备的位图信息。在一个具体例子中,在用第二备用存储器设备替换第一备用存储器设备时,将两个备用存储器设备中的位图信息进行组合,以在第二备用存储器设备中形成更新的位图信息。在另一例子中,基于第二备用存储器设备中已更新的数据重建位图信息。

[0064] 经过对第二备用存储器设备的更新,确保了存储器阵列的数据完整性。通常,步骤 34 的智能重建仅花费几十分钟的时间,因此在这段时间内不会有大量数据写入。另一反面,如果智能重建能够成功完成,那么一般来说,故障存储器设备损坏并不严重,其中的错误数据量非常有限。因此,以上更新步骤所花费的时间通常并不会很长。根据现有的典型存储器阵列的配置,即使在执行更新步骤的情况下,重建整个存储器阵列的总时间也不会超过 1 个小时。这比完全进行部件重建所花费的 4-6 小时具有明显的优势。

[0065] 以上结合具体例子描述了本发明实施例的具体执行过程。可以理解,本发明的实施例的方法不仅可以应用于 RAID5 阵列,也可以应用于其他存储器阵列,只要该存储器阵列具有足够的冗余度来恢复损坏的数据块。例如,对于 RAID6 阵列,如果阵列中的一个存储器设备出现故障,那么可以完全参照以上例子,利用两个(第一和第二)备用存储器设备进行阵列的重建。如果 RAID6 中两个存储器设备同时出现故障,那么可以针对每一个故障存储器设备,采用两个备用存储器设备来进行重建,也就是一共采用四个备用存储器设备进行阵列重建。在重建完成之后,释放其中的两个备用存储器设备。对于 RAID10,可以逐一对故障存储器设备实施上述重建方式。对于其他存储器阵列,本领域技术人员在阅读本说明书教导之后,可以采用类似的并行执行部件重建和智能重建的方法来进行数据恢复和阵列重建。

[0066] 基于同一发明构思,本发明的实施例还提供了一种重建存储器阵列的装置。图 7 示出根据本发明一个实施例的装置的框图。如图 7 所示,该装置总体上示出为 700,并包括:第一替换单元 70,配置为,响应于在存储器阵列中出现故障存储器设备,用第一备用存储器设备替换所述故障存储器设备;部件重建单元 72,配置为利用存储器阵列中除第一备用存储器设备之外的其他存储器设备进行部件重建,从而在所述第一备用存储器设备中恢复所述故障存储器设备中的数据;智能重建单元 74,配置为,与部件重建单元 72 的操作并行地,利用存储器阵列之外的第二备用存储器设备进行智能重建,从而将故障存储器设备中的数据复制到所述第二备用存储器设备;第二替换单元 76,配置为,响应于所述智能重建完成,在所述存储器阵列中用所述第二备用存储器设备替换所述第一备用存储器设备。

[0067] 根据一个实施例,第一替换单元 70 配置为,将故障存储器设备移出所述存储器阵列,将第一备用存储器设备纳入到该存储器阵列,并使得该第一备用存储器设备顶替所述故障存储器设备在存储器阵列中的位置。

[0068] 根据一个实施例,装置 700 还包括检测单元(未示出),配置为检测所述智能重建的执行状态。

[0069] 在一个实施例中,装置 700 还包括终止单元 77 (虚线示出),配置为响应于所述智能重建失败,终止所述智能重建。

[0070] 在一个实施例中,终止单元 77 配置为,将所述第二备用存储器设备移出与故障存储器设备构成的镜像阵列,并将其释放为空闲的备用存储器设备。

[0071] 根据一个实施例,第二替换单元 76 配置为,将所述第一备用存储器设备移出存储器阵列,将所述第二备用存储器设备纳入到该存储器阵列,并使得该第二备用存储器设备顶替第一备用存储器设备在存储器阵列中的位置。

[0072] 根据一个实施例,装置 700 还包括更新单元 78 (虚线示出),配置为,响应于第二替换单元替换完成,更新第二备用存储器设备中的数据。

[0073] 在一个实施例中,所述更新单元 78 进一步配置为,修正所述第二备用存储器设备中的错误数据。

[0074] 在一个实施例中,所述更新单元 78 进一步配置为,在第二备用存储器设备中恢复所述智能重建过程中新写入到第一存储器设备的数据。

[0075] 在一个实施例中,所述更新单元 78 进一步配置为,更新所述第二备用存储器设备的位图信息。

[0076] 在一个实施例中,上述装置 700 包含在现有的存储器阵列控制器中。在另一实施例中,上述装置 700 实现为独立的控制工具,与现有的存储器阵列控制器连接和通信。

[0077] 以上装置 700 的具体执行方式可以参照之前结合具体例子对方法的描述,在此不再赘述。

[0078] 根据本发明实施例的方法和装置,存储器阵列可以利用智能重建来快速恢复数据,同时可以避免在智能重建不成功的情况下存储器阵列存取受影响的风险,从而实现灵活高效且安全的数据重建。

[0079] 附图中的流程图和框图显示了根据本发明的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段或代码的一部分,所述模块、程序段或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意,在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个连续的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的,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操作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0080] 以上已经描述了本发明的各实施例,上述说明是示例性的,并非穷尽性的,并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下,对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择,旨在最好地解释各实施例的原理、实际应用或对市场中的技术的技术改进,或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。



图 1A

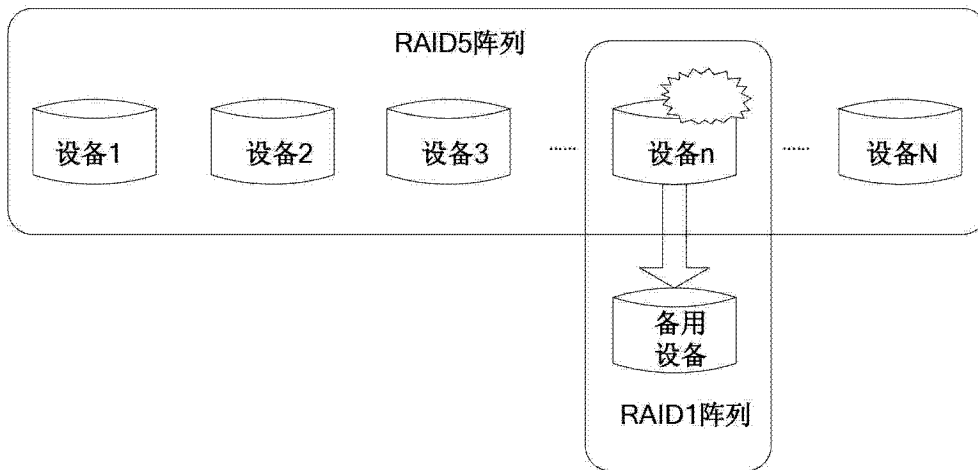


图 1B

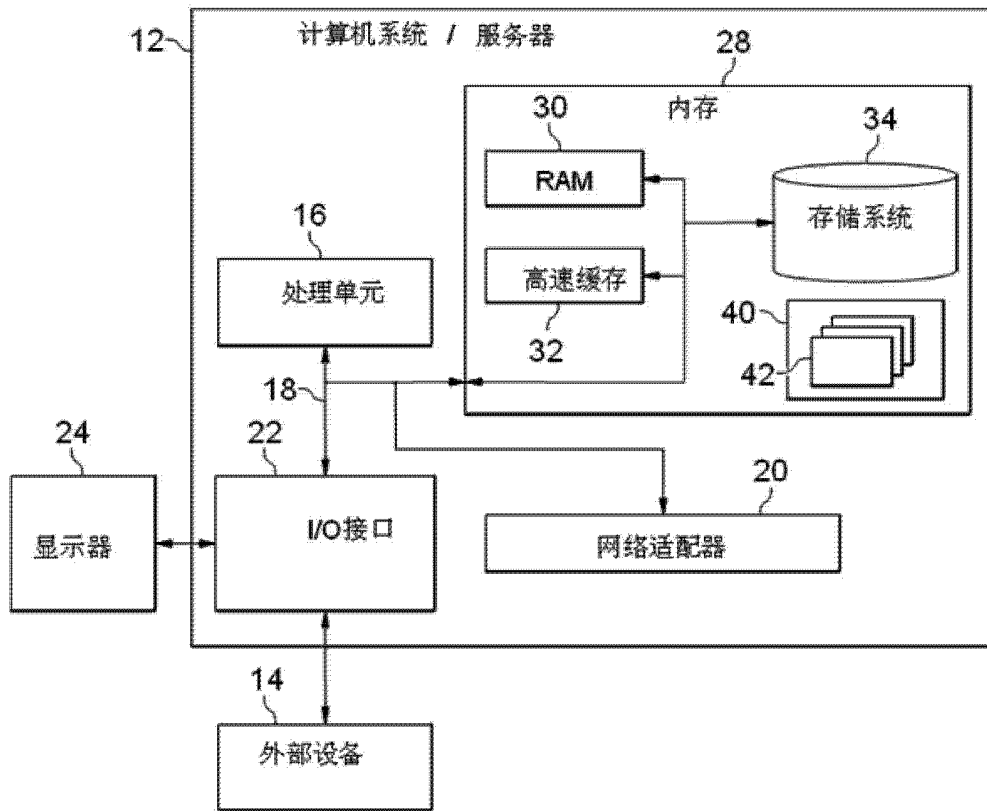


图 2

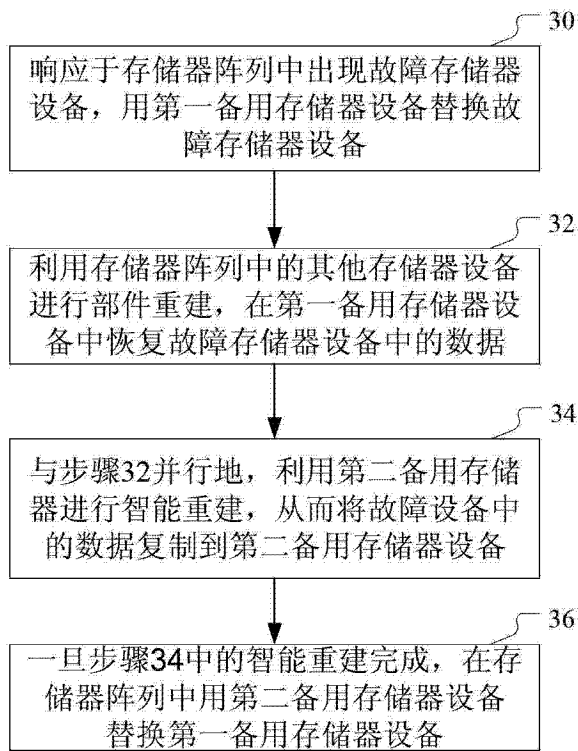


图 3

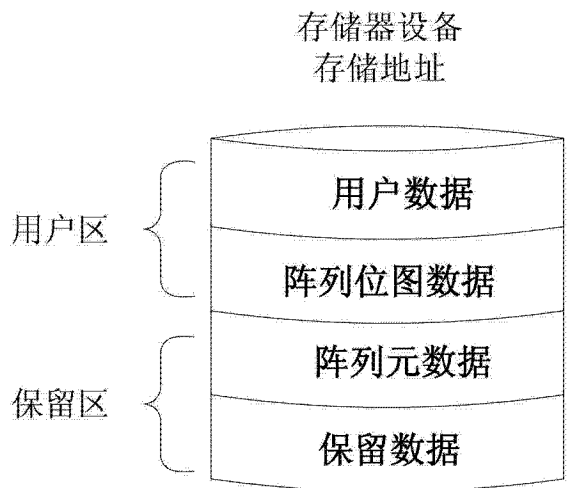


图 4

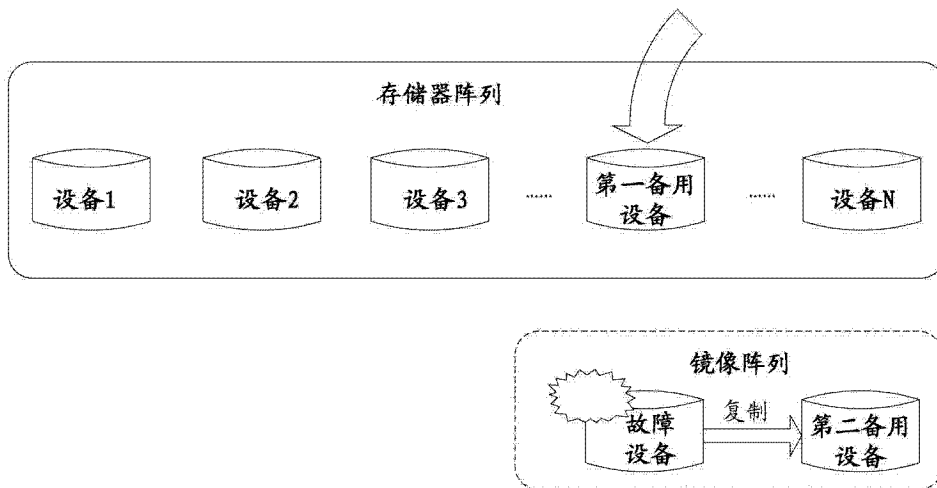


图 5

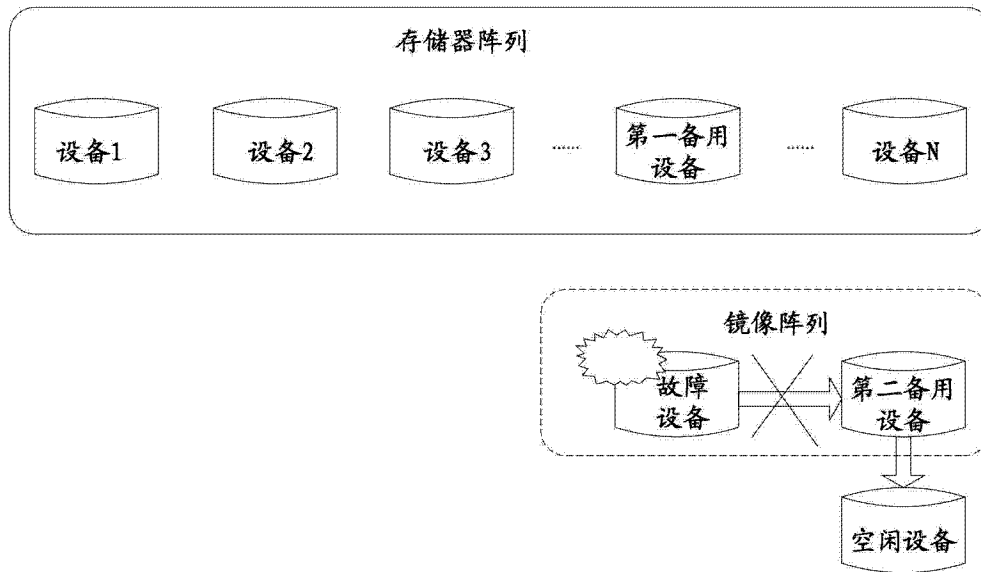


图 6A

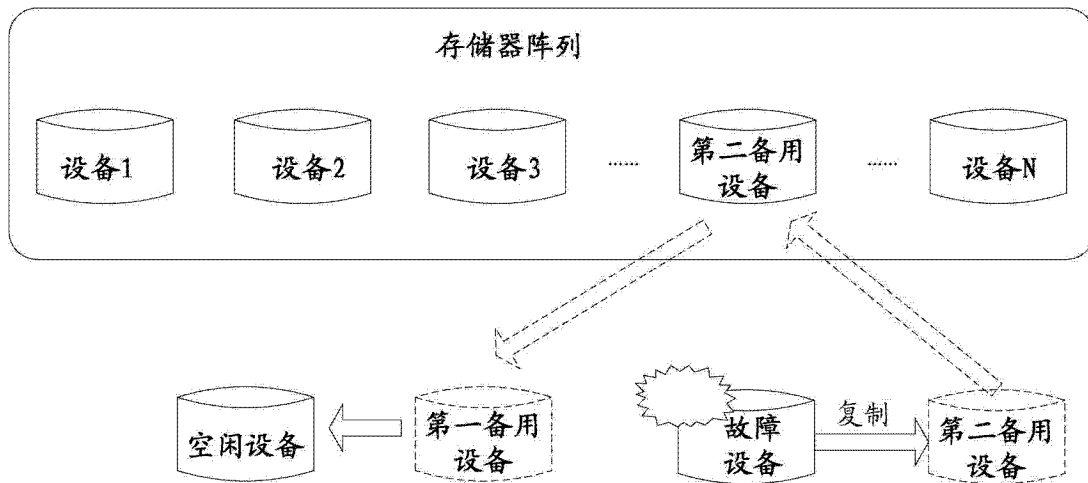


图 6B

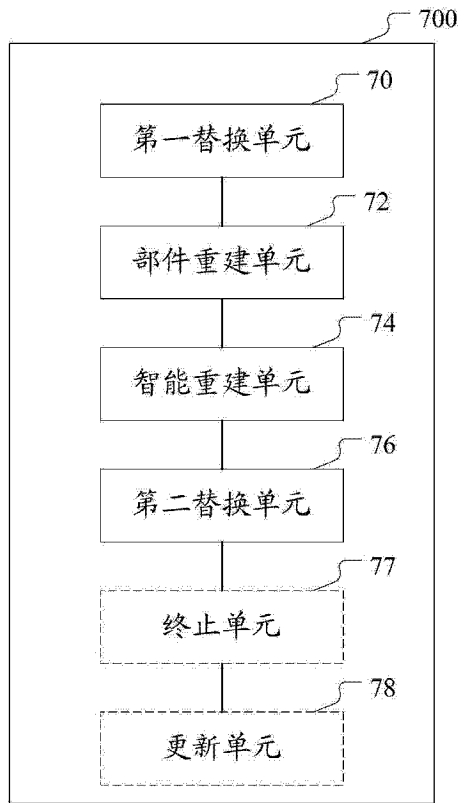


图 7