(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 103645997 B
(45)授权公告日 2016.12.07

(21)申请号 201310733189.0

(22)申请日 2013.12.26

(65)同一申请的已公布的文献号
申请公布号 CN 103645997 A

(43)申请公布日 2014.03.19

(73)专利权人 深圳市迪菲特科技股份有限公司
地址 518000 广东省深圳市南山区南头关口二路智恒战略性新兴产业园30栋5楼

(72)发明人 陈学伟

(74)专利代理机构 深圳华奇信诺专利代理事务所(普通合伙) 44328
代理人 曲卫涛

(51)Int.Cl.
G06F 12/14(2006.01)

(56)对比文件
EP 1597674 B1,2008.04.09,
CN 102981930 A,2013.03.20,
CN 102799533 A,2012.11.28,
CN 103064804 A,2013.04.24,
CN 103389918 A,2013.11.13,

审查员 李江

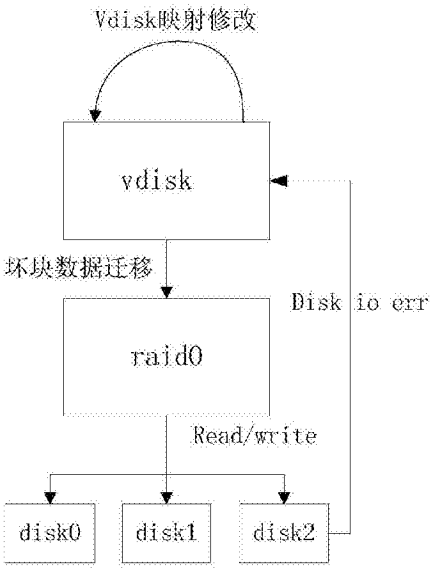
权利要求书1页 说明书5页 附图3页

(54)发明名称

一种数据保护的方法与系统

(57)摘要

本发明公开了一种数据保护的方法与系统，其中，所述方法包括以下步骤：S1，对冗余磁盘阵列中发生磁盘读写错误进行捕捉，获取介质读写错误的区域信息；S2，根据所述区域信息，将其中的数据迁移到介质正常的区域；S3，在逻辑层卷重新进行映射。采用上述方案，重建时间快，无需热备盘，及时发现介质读写错误的问题，实现了冗余磁盘阵列的快速数据恢复，并且减少了替换之后磁头的移动，降低了坏块替换的性能影响，提升了冗余磁盘阵列的鲁棒性，具有很高的市场应用价值。



CN 103645997 B

1.一种数据保护的方法，其特征在于，包括以下步骤：

S1，对冗余磁盘阵列中发生磁盘读写错误进行捕捉，获取介质读写错误的区域信息；

S2，根据所述区域信息，将其中的数据迁移到介质正常的区域；

S3，在逻辑层卷重新进行映射；

其中，步骤S1具体包括以下步骤：在冗余磁盘阵列中出现磁盘读写错误时进行重试，判断磁盘读写错误是否可修复，否则上报介质读写错误及其区域信息；步骤S2中，所述数据迁移按冗余磁盘阵列条带对齐；步骤S2具体包括以下步骤：根据所述区域信息，按冗余磁盘阵列条带对齐，读取并根据预配置的策略，将出现磁盘读写错误所在的错误冗余磁盘阵列条带的数据，迁移到正常冗余磁盘阵列条带上。

2.根据权利要求1所述方法，其特征在于，所述区域信息包括逻辑区块地址及长度。

3.根据权利要求2所述方法，其特征在于，步骤S3具体包括以下步骤：在逻辑层卷修改所述错误冗余磁盘阵列条带所在物理地址的逻辑映射。

4.根据权利要求3所述方法，其特征在于，步骤S3之后，还执行以下步骤：将所述逻辑映射的修改写入元数据内。

5.根据权利要求4所述方法，其特征在于，步骤S1之前，还执行步骤S0：预配置所述策略。

6.根据权利要求5所述方法，其特征在于，步骤S0中，所述策略包括在每个虚拟磁盘之后设置一替换区域，用于对应虚拟磁盘的坏块替换。

7.一种数据保护的系统，包括逻辑盘卷管理层，其特征在于，所述逻辑盘卷管理层设置策略配置模块、磁盘读写错误捕捉模块、数据迁移模块与修改模块；

所述策略配置模块与所述数据迁移模块连接，用于预配置策略并存储；

所述磁盘读写错误捕捉模块与所述数据迁移模块连接，用于在冗余磁盘阵列中出现磁盘读写错误时进行重试，判断磁盘读写错误重复发生时，上报介质读写错误及其区域信息；

所述数据迁移模块还与所述修改模块连接，用于根据所述区域信息，按冗余磁盘阵列条带对齐，读取并根据预配置的策略，将出现磁盘读写错误所在的错误冗余磁盘阵列条带的数据，迁移到正常冗余磁盘阵列条带上；

所述修改模块用于在逻辑层卷修改所述错误冗余磁盘阵列条带所在物理地址的逻辑映射。

一种数据保护的方法与系统

技术领域

[0001] 本发明涉及冗余磁盘组的数据保护，尤其涉及的是，一种数据保护的方法与系统。

背景技术

[0002] 现有的存储系统，对数据的可靠性具有极高的要求，而数据的存储介质却不可能是百分之百可靠的，一旦介质损坏，那么将产生难以估量的损失。虽然现在有冗余磁盘组(RAID,Redundant Arrays of Inexpensive Disks)来降低介质损坏造成损失的可能性，但是现有冗余阵列数据恢复的功能却有以下局限性：

[0003] 一、重建时间比较长，现在重建都是按一块盘为单位进行重建的，随着现在盘的容量越来越大，重建时间越来越长，系统的可靠性得不到保证；

[0004] 二、需要一块热备盘，否则无法实现重建，这样就需要浪费一定数量的存储空间。

发明内容

[0005] 本发明所要解决的技术问题是提供一种新的数据保护的方法与系统。

[0006] 本发明的技术方案如下：一种数据保护的方法，其包括以下步骤：S1，对冗余磁盘阵列中发生磁盘读写错误进行捕捉，获取介质读写错误的区域信息；S2，根据所述区域信息，将其中的数据迁移到介质正常的区域；S3，在逻辑层卷重新进行映射。

[0007] 优选的，所述方法中，步骤S1具体包括以下步骤：在冗余磁盘阵列中出现磁盘读写错误时进行重试，判断磁盘读写错误是否可修复，否则上报介质读写错误及其区域信息。

[0008] 优选的，所述方法中，步骤S2中，所述数据迁移按冗余磁盘阵列条带对齐。

[0009] 优选的，所述方法中，步骤S2具体包括以下步骤：根据所述区域信息，按冗余磁盘阵列条带对齐，读取并根据预配置的策略，将出现磁盘读写错误所在的错误冗余磁盘阵列条带的数据，迁移到正常冗余磁盘阵列条带上。

[0010] 优选的，所述方法中，所述区域信息包括逻辑区块地址及长度。

[0011] 优选的，所述方法中，步骤S3具体包括以下步骤：在逻辑层卷修改所述错误冗余磁盘阵列条带所在物理地址的逻辑映射。

[0012] 优选的，所述方法中，步骤S3之后，还执行以下步骤：将所述逻辑映射的修改写入元数据内。

[0013] 优选的，所述方法中，步骤S1之前，还执行步骤S0：预配置所述策略。

[0014] 优选的，所述方法中，步骤S0中，所述策略包括在每个虚拟磁盘之后设置一替换区域，用于对应虚拟磁盘的坏块替换。

[0015] 本发明的又一技术方案如下：一种数据保护的系统，包括逻辑盘卷管理层，其中，所述逻辑盘卷管理层设置策略配置模块、磁盘读写错误捕捉模块、数据迁移模块与修改模块；所述策略配置模块与所述数据迁移模块连接，用于预配置策略并存储；所述磁盘读写错误捕捉模块与所述数据迁移模块连接，用于在冗余磁盘阵列中出现磁盘读写错误时进行重试，判断磁盘读写错误重复发生时，上报介质读写错误及其区域信息；所述数据迁移模块还

与所述修改模块连接，用于根据所述区域信息，按冗余磁盘阵列条带对齐，读取并根据预配置的策略，将出现磁盘读写错误所在的错误冗余磁盘阵列条带的数据，迁移到正常冗余磁盘阵列条带上；所述修改模块用于在逻辑层卷修改所述错误冗余磁盘阵列条带所在物理地址的逻辑映射。

[0016] 采用上述方案，本发明及时发现介质错误，触发主动数据迁移行为，具有很高的市场应用价值。

附图说明

[0017] 图1为本发明的一个实施例的功能实现示意图；

[0018] 图2为本发明的一个实施例的替换区域设置示意图；

[0019] 图3为本发明的一个实施例的LVM监控示意图；

[0020] 图4为现有技术的映射结构示意图；

[0021] 图5为本发明的一个实施例的映射结构示意图。

具体实施方式

[0022] 为了便于理解本发明，下面结合附图和具体实施例，对本发明进行更详细的说明。附图中给出了本发明的较佳的实施例。但是，本发明可以以许多不同的形式来实现，并不限于本说明书所描述的实施例。相反地，提供这些实施例的目的是使对本发明的公开内容的理解更加透彻全面。

[0023] 需要说明的是，当元件被称为“固定于”另一个元件，它可以直接在另一个元件上或者也可以存在居中的元件。当一个元件被认为是“连接”另一个元件，它可以是直接连接到另一个元件或者可能同时存在居中元件。本说明书所使用的术语“垂直的”、“水平的”、“左”、“右”以及类似的表述只是为了说明的目的。

[0024] 除非另有定义，本说明书所使用的所有的技术和科学术语与属于本发明的技术领域的技术人员通常理解的含义相同。本说明书中在本发明的说明书中所使用的术语只是为了描述具体的实施例的目的，不是用于限制本发明。本说明书所使用的术语“和/或”包括一个或多个相关的所列项目的任意的和所有的组合。

[0025] 本发明的一个实施例是，一种数据保护的方法，其包括以下步骤：S1，对冗余磁盘阵列中发生磁盘读写错误进行捕捉，获取介质读写错误的区域信息；S2，根据所述区域信息，将其中的数据迁移到介质正常的区域；S3，在逻辑层卷重新进行映射。优选的，所述方法中，步骤S2中，所述数据迁移按冗余磁盘阵列条带对齐。其中，采用LVM(Logical Volume Manager，逻辑盘卷管理)方式管理所述冗余磁盘阵列；LVM是建立在硬盘和分区之上的一个逻辑层，来提高磁盘分区管理的灵活性。LVM把几个底层设备做成物理卷、物理卷组，从而能格式化、能进行扩展和收缩、用于储存数据。其过程是PD(块设备)→PV(物理卷)→VG(物理卷组)→LV(逻辑分区)。通过使用现有的LVM模式，所有物理磁盘和分区，无论它们的大小和分布方式如何，都被抽象为单一存储(single storage)源。

[0026] 例如，LVM可以将分区和磁盘聚合成一个虚拟磁盘(VD，Virtual Disk)，从而用小的存储空间组成一个统一的大空间。这个虚拟磁盘称为卷组(volume group)。因此可以建立比最大的磁盘还大的文件系统，还可以在磁盘池中添加磁盘和分区，对现有的文件系统

进行在线扩展，又如，用一个160GB磁盘替换两个80GB磁盘，而不需要让系统离线，也不需要在磁盘之间手工转移数据；当存储空间超过所需的空间量时，从池中去除磁盘，从而缩小文件系统。

[0027] 例如，所述方法包括以下步骤：S1，在冗余磁盘阵列中出现磁盘读写错误时进行重试，判断磁盘读写错误是否可修复，否则上报介质读写错误及其区域信息；其中，重试的目的是判断此错误区域是否不可修复，因为有的介质区域重试写或重试读后能够恢复正常，不影响后续的使用；例如，步骤S1中判断磁盘读写错误是否可修复，是则不作处理，或者仅上报上层或管理员。S2，根据所述区域信息，将其中的数据迁移到介质正常的区域；S3，在逻辑层卷重新进行映射。例如，步骤S1中实时获取磁盘读写错误信息或者实时获取磁盘读写信息并判断是否发生磁盘读写错误，是则重试以判断磁盘读写错误是否重复发生。又如，采用同一文件或者同样大小文件进行重试，以判断磁盘读写错误是否重复发生。又如。采用定时机制判断是否发生磁盘读写错误。

[0028] 又如，所述方法包括以下步骤：S1，在冗余磁盘阵列中出现磁盘读写错误时进行重试，判断磁盘读写错误是否重复发生，是则上报介质读写错误及其区域信息；S2，根据所述区域信息，按冗余磁盘阵列条带对齐，读取并根据预配置的策略，将出现磁盘读写错误所在的错误冗余磁盘阵列条带的数据，迁移到正常冗余磁盘阵列条带上；S3，在逻辑层卷重新进行映射。优选的，所述方法中，所述区域信息包括逻辑区块地址及长度。根据上报的介质读写错误的IO及其中的LBA(Logical Block Address/Addressing，逻辑区块地址/寻址)和LENGTH(长度)，进行按阵列条带对齐，根据策略，将出现磁盘读写错误所在RAID条带(strip RAID)的数据，按配置的策略，迁移到其他正常条带所在的RAID条带上。采用RAID条带技术，把数据存放在两个或更多的硬盘的分区上，这些数据有一半是在一个硬盘上，另一半在另一个硬盘上，这样数据是从两个硬盘上同时读出来的，从而可以提高硬盘读写的速度。例如，在逻辑层卷一层中重新进行映射；又如，在重新映射之后，发送上报通知。

[0029] 优选的，所述方法中，步骤S3具体包括以下步骤：在逻辑层卷修改所述错误冗余磁盘阵列条带所在物理地址的逻辑映射。优选的，所述方法中，步骤S3之后，还执行以下步骤：将所述逻辑映射的修改写入元数据(Metadata)内，元数据是用来描述数据的数据，作为最小的一种数据单位，元数据可以为数据说明其元素或属性，例如名称、大小、数据类型等，或其结构，例如长度、字段、数据列等，或其相关数据，例如位于何处、如何联系、拥有者等。这样有助于记录错误信息，在系统重启之后，新的映射仍然生效。优选的，步骤S3中，还包括步骤：重建RAID。优选的，步骤S3中，还包括步骤：根据元数据中的修改后的所述逻辑映射，完成RAID的重建。

[0030] 优选的，所述方法中，步骤S1之前，还执行步骤S0：预配置所述策略。例如，策略根据数据保护的对象设置，又如，策略根据数据保护的级别设置。优选的，所述策略设置替换区域；优选的，所述方法中，步骤S0中，所述策略包括在每个虚拟磁盘之后设置一替换区域，用于对应虚拟磁盘的坏块替换。例如，请参考图2，在虚拟磁盘V0之后设置一替换区域，然后再设虚拟磁盘V1；又如，在每个虚拟磁盘之后设置一替换区域，其容量为前一虚拟磁盘容量的0.1%-0.2%，或者，替换区域容量为前一虚拟磁盘容量与后一虚拟磁盘容量之和的0.05%-0.08%；又如，该替换区域的大小动态调整，根据前一虚拟磁盘的剩余容量或者根据系统启动时LVM分配后的虚拟磁盘的剩余容量调整，例如，动态调整为剩余容量的1%-2%。优选的，

所述策略还包括分配或调整冗余磁盘阵列条带。这样，通过在对冗余阵列中发生磁盘读写错误时，及时进行捕捉，从而知道出现介质错误的区域所在，然后触发主动数据迁移行为，将出现磁盘读写错误的区域数据迁移到新的介质正常区域，数据迁移完成后，在逻辑卷层重新进行映射。这样，以后的数据读取写入都是使用新的正常条带，从而实现了冗余磁盘阵列的快速恢复。

[0031] 结合应用于上述任一相关实施例，优选的，本发明还提供了一种数据保护的系统，包括逻辑盘卷管理层，其中，所述逻辑盘卷管理层设置策略配置模块、磁盘读写错误捕捉模块、数据迁移模块与修改模块。即，在现有的LVM中增加如下几个功能模块：策略配置模块、磁盘读写错误捕捉模块、数据迁移模块以及逻辑卷映射的修改模块。其实现如图1所示，raid0对disk0、disk1和disk2进行读写操作，发现磁盘发生读写错误(Disk io err)时，进行vdisk(VD)映射修改，然后执行坏块数据迁移。

[0032] 所述策略配置模块与所述数据迁移模块连接，用于预配置策略并存储；所述磁盘读写错误捕捉模块与所述数据迁移模块连接，用于在冗余磁盘阵列中出现磁盘读写错误时进行重试，判断磁盘读写错误重复发生时，上报介质读写错误及其区域信息；所述数据迁移模块还与所述修改模块连接，用于根据所述区域信息，按冗余磁盘阵列条带对齐，读取并根据预配置的策略，将出现磁盘读写错误所在的错误冗余磁盘阵列条带的数据，迁移到正常冗余磁盘阵列条带上；所述修改模块用于在逻辑层卷修改所述错误冗余磁盘阵列条带所在物理地址的逻辑映射。

[0033] 策略配置模块，用于配置替换的策略。为了优化性能，在坏块替换之后减少磁盘磁头的移动，在每个VD之后保留一定的空间，如图2所示，此空间专用于VD内的坏块替换。替换区域的大小可以根据配置策略算出。比如：按比例计算，一个VD大小为1T，那么保留VD大小的1/1000空间用于替换，即1G。

[0034] 磁盘读写错误捕捉模块，在出现磁盘读写错误之后，进行重试，如果仍然为磁盘读写错误，那么将此介质读写错误的读写IO(input/output，输入输出，也称“读写”)信息给上报给LVM。优选的，磁盘读写错误捕捉模块及时发现磁盘读写错误，然后进行重试。

[0035] LVM监控模块，即数据迁移模块，根据上报的介质读写错误的IO，根据IO的LBA和LENGTH，进行按阵列条带对齐，根据策略，将出现磁盘读写错误所在RAID条带的数据，按配置的策略，迁移到其他正常条带所在的RAID条带上。如图3所示，LVM监控模块获知DISK介质读写错误IO的lba和len，然后根据策略到RAID寻找一个替换条带，并在数据迁移过程中LOCK住此条带，然后修改映射地址。

[0036] 修改模块在数据迁移完成之后，修改该条带所在物理地址的逻辑映射。比如原来的映射为一个extent(扩展区，又称为“片区”)对应一块物理区域，如图4所示。现在很多文件系统都采用了extent替代block(块，又称为“块区”，通常磁头一次可以读取一个block)来管理磁盘。Extent就是一些连续的block，一个extent由起始的block加上长度进行定义，例如，片区是由逻辑块偏移量(logical offset)、长度(length)和物理地址(physical address)组成的三元组来描述。其中由逻辑块偏移量和长度可能计算出物理地址。物理卷(PV，physical volume)是指硬盘分区或从逻辑上与磁盘分区具有同样功能的设备，如RAID，是LVM的基本存储逻辑块，但与基本的物理存储介质不同，物理存储介质如分区、磁盘等，PV包含有与LVM相关的管理参数。在映射修改之后如图5所示，原本指向被替换区域的

extent1修改为指向替换区域。在映射修改完成后，将此映射的修改写入元数据内，以便重启之后，新的映射仍然生效。

[0037] 优选的，该系统还设置错误上传模块，其与磁盘读写错误捕捉模块连接，用于在发生磁盘读写错误时，上报介质读写错误及其区域信息给上层，例如，管理员。优选的，该系统还设置重建模块，用于在映射修改之前或者之后，重建RAID。

[0038] 进一步地，本发明的实施例还包括，上述各实施例的各技术特征，相互组合形成的数据保护的方法和系统，通过在VD逻辑层实现了坏块替换，可以和RAID的重建等相互结合，进一步提升RAID的健壮性(robustness，又称鲁棒性)；若磁盘不足，也可以单独使用来保护数据；按条带对齐替换区域，从而替换之后不影响性能；还通过配置策略，预留好替换所需空间，并且预留空间都在每个VD之后，从而减少替换之后磁头的移动，减低坏块替换的性能影响。

[0039] 上述各实施例的技术效果包括重建时间快，无需热备盘，及时发现介质读写错误的问题，实现了冗余磁盘阵列的快速数据恢复，并且减少了替换之后磁头的移动，降低了坏块替换的性能影响，提升了冗余磁盘阵列的鲁棒性，具有很高的市场应用价值。

[0040] 需要说明的是，上述各技术特征继续相互组合，形成未在上面列举的各种实施例，均视为本发明说明书记载的范围；并且，对本领域普通技术人员来说，可以根据上述说明加以改进或变换，而所有这些改进和变换都应属于本发明所附权利要求的保护范围。

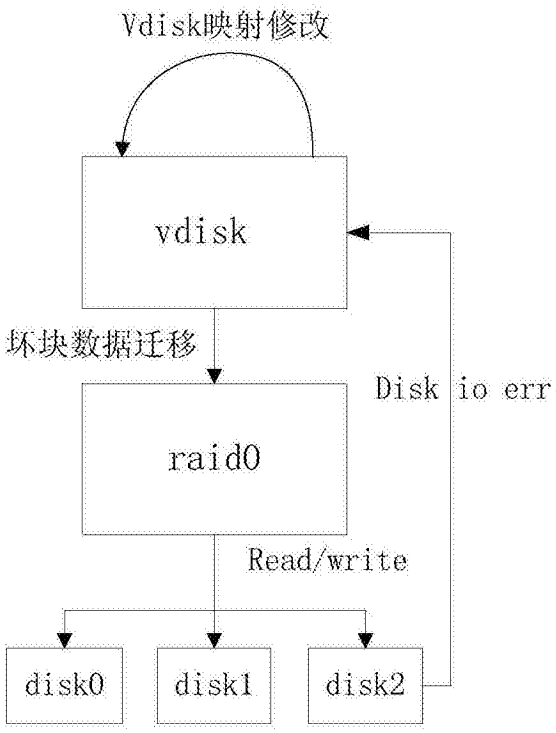

Vdisk映射修改
vdisk
坏块数据迁移
Disk io err
raid0
Read/write
disk0
disk1
disk2

图1

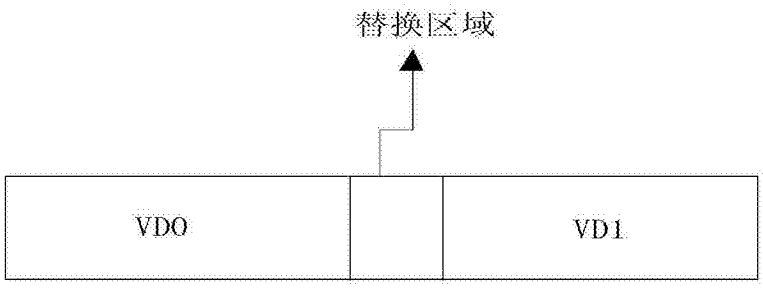

替换区域
VD0
VD1

图2

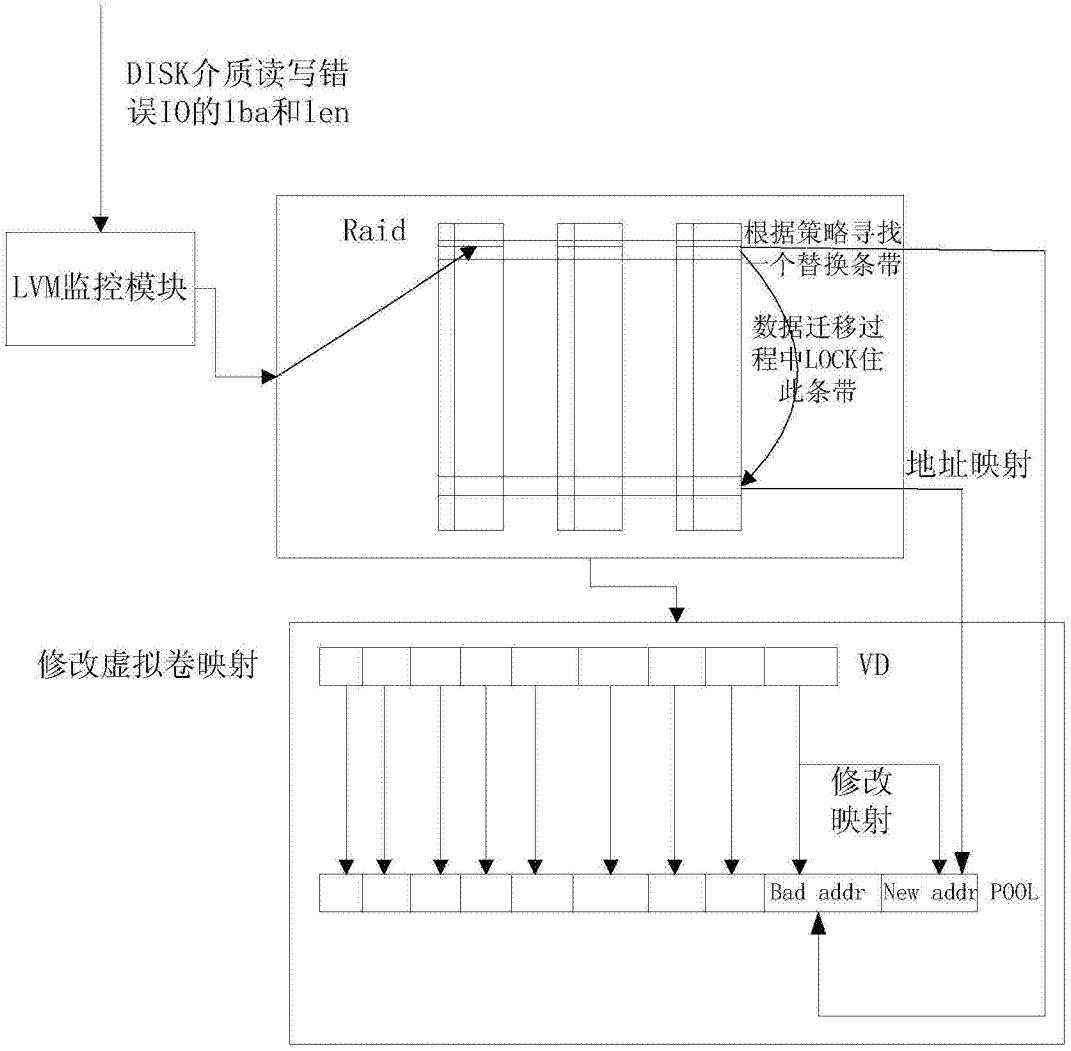

DISK介质读写错
误IO的lba和len
LVM监控模块
Raid
根据策略寻找
一个替换条带
数据迁移过
程中LOCK住
此条带
地址映射
修改虚拟卷映射
VD
修改
映射
Bad addr
New addr
POOL

图3

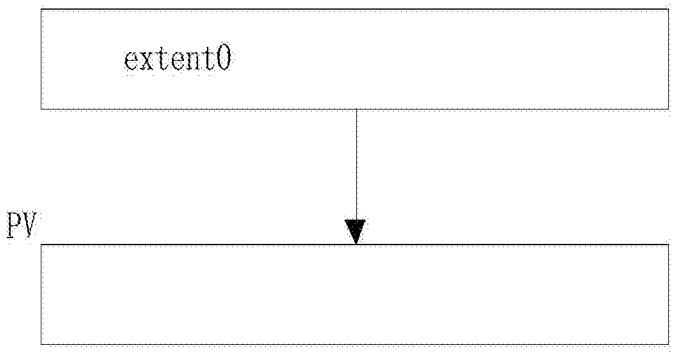

extent0
PV

图4

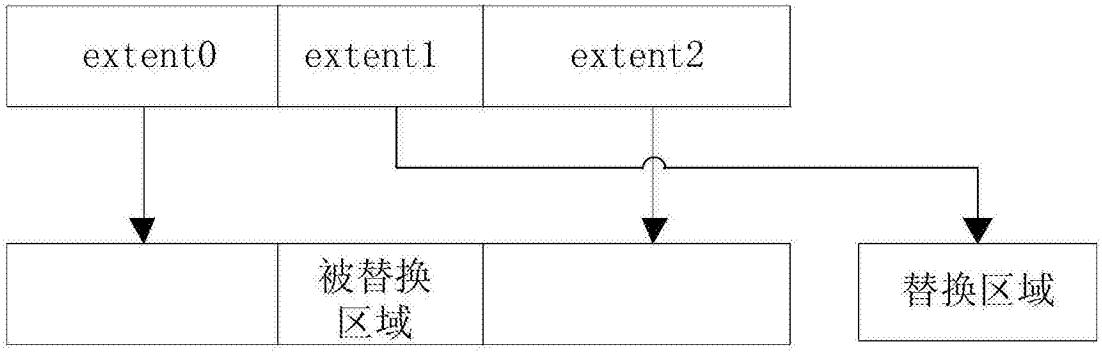


图5