



(12)发明专利

(10)授权公告号 CN 104813290 B

(45)授权公告日 2018.09.21

(21)申请号 201380059018.2

(73)专利权人 康佩伦特科技公司

(22)申请日 2013.12.05

地址 美国明尼苏达州

(65)同一申请的已公布的文献号

(72)发明人 A·J·弗勒德 D·J·安德森

申请公布号 CN 104813290 A

(74)专利代理机构 北京润平知识产权代理有限公司 11283

(43)申请公布日 2015.07.29

代理人 孙向民 肖冰滨

(30)优先权数据

(51)Int.CI.

13/706,553 2012.12.06 US

G06F 11/00(2006.01)

(85)PCT国际申请进入国家阶段日

审查员 许莎莎

2015.05.12

(86)PCT国际申请的申请数据

PCT/US2013/073347 2013.12.05

(87)PCT国际申请的公布数据

W02014/089311 EN 2014.06.12

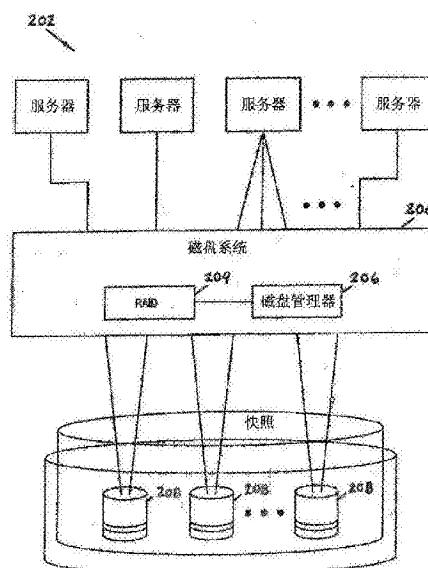
权利要求书2页 说明书8页 附图2页

(54)发明名称

RAID调查器

(57)摘要

一种在数据存储子系统的将要失效的磁盘于失效之前调查数据存储子系统的潜在错误并恢复重建将要失效的磁盘的数据所使用的不可读数据的方法。该方法包括：确定数据存储子系统的多个磁盘中的磁盘达到被识别为将要失效的磁盘的阈值，并在将要失效的磁盘于失效之前调查多个磁盘中的剩余磁盘上的至少一部分数据以识别具有潜在错误的数据存储区域。可以至少部分地利用存储在将要失效的磁盘上的数据来重建已识别的数据存储区域。



1. 一种在数据存储子系统的将要失效的磁盘于失效之前调查所述数据存储子系统的潜在错误并恢复重建所述将要失效的磁盘的数据所使用的不可读数据的方法,所述数据存储子系统提供多个磁盘的存储摘要,该方法包括:

确定所述多个磁盘中的磁盘达到被识别为将要失效的磁盘的阈值;

在识别所述磁盘为将要失效的磁盘以及在所述将要失效的磁盘于失效之前,所述将要失效的磁盘将失效时识别需要重建存储在所述将要失效的磁盘上的所有数据的多个磁盘中的剩余磁盘的每一个数据存储区域,包括识别所述将要失效的磁盘上的每一个RAID扩展区以及识别RAID条中所述多个磁盘中的剩余磁盘的与在所述将要失效的磁盘上识别的每一个RAID扩展区相对应的每一个数据存储区域;

调查所识别的数据存储区域的至少一部分数据以识别具有潜在错误的数据存储区域;以及

至少部分地利用存储在所述将要失效的磁盘上的数据来重建被识别为具有潜在错误的数据存储区域。

2. 根据权利要求1所述的方法,其中,调查所识别的数据存储区域的至少一部分数据以识别具有潜在错误的数据存储区域包括执行读取操作以识别所述潜在错误。

3. 根据权利要求2所述的方法,其中,所述将要失效的磁盘将要失效时识别需要重建存储在所述将要失效的磁盘上的所有数据的多个磁盘中的剩余磁盘的每一个数据存储区域为基于将要失效的磁盘的识别而被自动发起。

4. 根据权利要求2所述的方法,其中,所述将要失效的磁盘将要失效时识别需要重建存储在所述将要失效的磁盘上的所有数据的多个磁盘中的剩余磁盘的每一个数据存储区域为基于用户请求而被发起。

5. 根据权利要求2所述的方法,其中,确定所述多个磁盘中的磁盘达到被识别为将要失效的磁盘的阈值是基于与至少其中一个所述磁盘有关的信息和所述磁盘的使用。

6. 根据权利要求3所述的方法,其中,确定所述多个磁盘中的磁盘达到被识别为将要失效的磁盘的阈值是自动执行的。

7. 根据权利要求1所述的方法,其中,确定所述多个磁盘中的磁盘达到被识别为将要失效的磁盘的阈值包括:

根据预定标准来跟踪所述磁盘的错误;以及

在已跟踪错误的数量达到阈值数量的错误时将所述磁盘识别为将要失效的。

8. 根据权利要求2所述的方法,其中,确定所述多个磁盘中的磁盘达到被识别为将要失效的磁盘的阈值包括:

根据预定标准来跟踪所述磁盘的错误;以及

在已跟踪错误的数量达到阈值数量的错误时将所述磁盘识别为将要失效的。

9. 根据权利要求1所述的方法,该方法还包括移除所述将要失效的磁盘并用替换磁盘来替换所述将要失效的磁盘。

10. 根据权利要求9所述的方法,该方法还包括利用来自所述多个磁盘中的剩余磁盘的数据,在所述替换磁盘上重构所述将要失效的磁盘的数据。

11. 根据权利要求2所述的方法,该方法还包括再次调查重构的数据存储区域以确认成功校正了所述潜在错误。

12.一种信息处理系统,包括:

数据存储子系统,用于提供多个磁盘的存储摘要;以及

磁盘控制器,该磁盘控制器具有访问所述数据存储子系统的权限并被配置成:

确定所述多个磁盘中的磁盘是否达到被识别为将要失效的磁盘的阈值;

在识别所述磁盘为将要失效的磁盘以及在所述将要失效的磁盘于失效之前,所述将要失效的磁盘将失效时识别需要重建存储在所述将要失效的磁盘上的所有数据的多个磁盘中的剩余磁盘的每一个数据存储区域,包括识别所述将要失效的磁盘上的RAID扩展区以及识别RAID条中所述多个磁盘中的剩余磁盘的与在所述将要失效的磁盘上识别的每一个RAID扩展区相对应的每一个数据存储区域;

调查所识别的数据存储区域的至少一部分数据以识别具有潜在错误的数据存储区域;以及

至少部分地利用存储在所述将要失效的磁盘上的数据来重建被识别为具有潜在错误的数据存储区域。

13.根据权利要求12所述的系统,其中,所述磁盘控制器被配置成基于与至少其中一个所述磁盘有关的信息和所述磁盘的使用来确定所述多个磁盘中的磁盘是否达到被识别为将要失效的磁盘的阈值。

14.根据权利要求13所述的系统,其中,与所述磁盘的使用有关的信息包括根据预定标准跟踪的所述磁盘的错误。

15.一种在数据存储子系统的将要失效的磁盘于失效之前调查所述数据存储子系统的潜在错误并恢复重建所述将要失效的磁盘的数据所使用的不可读数据的方法,该方法包括:

根据预定标准来自动跟踪所述数据存储子系统的多个磁盘的错误并在磁盘的已跟踪错误达到预定阈值时将所述磁盘识别为将要失效的磁盘;

在识别所述磁盘为将要失效的磁盘以及在所述将要失效的磁盘完全于失效之前,识别包括所述将要失效的磁盘上的RAID扩展区的每一个RAID条中所述多个磁盘中的剩余磁盘上的每一个数据存储区域;

调查已识别数据存储区域的潜在错误;以及

至少部分地利用存储在所述将要失效的磁盘上的数据来重建被识别为具有潜在错误的所述数据存储区域。

16.根据权利要求15所述的方法,该方法还包括再次调查重构的数据存储区域以确认成功校正了所述潜在错误。

17.根据权利要求15所述的方法,该方法还包括:

移除所述将要失效的磁盘并用替换磁盘来替换所述将要失效的磁盘;以及

利用来自所述多个磁盘中的剩余磁盘的数据,在所述替换磁盘上重构所述将要失效的磁盘的数据。

## RAID调查器

### 技术领域

[0001] 本公开一般涉及调查数据存储系统的潜在错误的系统和方法。特别地，本公开涉及在磁盘失效之前调查数据存储子系统或其他信息处理系统的潜在错误的系统和方法，从而改进了容错性。

### 背景技术

[0002] 随着信息价值和使用的持续增长，个人和商务都在探寻更多的处理和存储信息的方法。用户可用的一种选择是信息处理系统。信息处理系统通常处理、编译、存储和/或传送用于商务、个人或其他用途的信息或数据，从而允许用户利用该信息的价值。由于技术和信息处理需求和要求因不同用户或应用而变化，所以信息处理系统还可以随着以下情况而变化：处理什么样的信息，如何处理信息，处理、存储或传递多少信息，以及可以多快和多高效地处理、存储或传递信息。信息处理系统的多样性允许信息处理系统可以是一般的信息处理系统，也可以是针对特定用户或特定用途而配置的信息处理系统，所述特定用途例如是金融业务处理、航班预约、企业数据存储或全球通信。另外，信息处理系统可以包括各种被设置用来处理、存储和传递信息的硬件和软件部件，且可以包括一个或多个计算机系统、数据存储系统和联网系统。

[0003] 许多信息处理系统尤其是数据存储系统中持续关注的问题是数据可靠性。当然，已经开发了许多解决方案来增加数据可靠性，包括例如利用RAID（独立磁盘冗余阵列）系统，RAID系统通常基于期望或需求的容量、冗余和性能水平来将多个磁盘组合到逻辑单元中，其中数据被分布在称为RAID级（RAID level）的若干路径之一中的磁盘之间。请见1988年加利福尼亚伯克利大学的David A. Patterson、Garth Gibson和Randy H. Katz的文章“*A Case for Redundant Arrays of Inexpensive Disks (RAID)*”。RAID技术已经一般性地增加了数据可靠性。

[0004] 尽管如此，仍然存在着要失效的磁盘能够使得用户数据处于不可恢复状态的若干状况。例如，在一个单冗余RAID例子中，特定的磁盘可能累积太多的错误恢复尝试并因此触发重构以将每个RAID扩展区（extent）从要失效的磁盘迁移到备份磁盘。同时，在扩展区正在重构时，重构条（rebuilding stripe）中另一磁盘的数据会因潜在错误（即，非显而易见的错误）而变得不可读，因为被写入区块中的数据不是可读的。然而，对该数据的读取是需要的，以便重建正在被重构的磁盘的数据，因此重构不能继续，使得用户的数据处于不可恢复状态中。

[0005] 已经引入数据清理（data scrubbing）作为RAID控制器周期性地读取并检查RAID阵列中的所有区块以在这些区块被使用之前检测坏区块的方式。然而，传统的RAID清理并没有足够快地检测潜在错误以便明显地改进数据可靠性。传统RAID清理操作在某一时间对单个RAID设备起作用并对RAID逻辑区块地址起作用，而非“垂直地”也即概念地说对磁盘或磁盘扩展区起作用。作为通过RAID设备上的条进行的清理过程，其向与RAID设备相关联的所有磁盘发送输入/输出（I/O）。在特定磁盘受到怀疑的情况下，其读取RAID设备的所有其

他磁盘,这在可疑磁盘处于即将失效风险中时浪费了宝贵的时间。另外,在具有多个磁盘层级(tier)的系统中,传统RAID清理操作没有针对磁盘类型(诸如具有更高失效趋势的那些磁盘)划分优先顺序。例如,如果更低、不太昂贵存储层级中的磁盘相比于其他相对更昂贵存储层级中的磁盘而言相对更经常地被怀疑失效,则花费时间清理更高、相对更昂贵存储层级中的磁盘在本质上是浪费的。

[0006] 鉴于前面所述,如果怀疑磁盘处于失效风险中,则在那个磁盘性能恶化以需要替换之前获悉那个磁盘上所有扩展区的关联RAID条都是可读的以便重建驻留在要失效磁盘上的所有数据或尽可能多的数据会是非常有用的。通过采用传统RAID清理操作,通常没有办法来快速而高效地确定与存储层级中的所有磁盘相关联的所有RAID设备上的该缺席的启动清理。然而,在与存储层级中的所有磁盘相关联的所有RAID设备上启动清理是太慢并消耗太多的资源。图1中提供了该问题的特定示例,其中图1示出了显示了10个分离磁盘的示例性数据存储系统100,其中简化起见在附图中垂直地示出了仅“磁盘X”102,磁盘X 102被完全示出并被标记以便于讨论。如图1所示,数据分布在三个RAID配置下的10个所示磁盘中:RAID 5遍布5个扩展区;RAID 10遍布2个扩展区;以及RAID 6遍布6个扩展区。如本领域技术人员将意识到的,扩展区和RAID条数据的实际物理配置和布局将典型地依赖于若干因素;因此,图1仅是出于讨论目的的概念性示例。假设磁盘X将失效或者将以其他方式返回太多的明显错误。为了确定磁盘X上的所有数据是否能够利用传统清理操作进行重建,全部RAID 5、RAID 10和RAID6设备将需要被清理。然而,如果存在着一种系统和方法来读取或调查该图中所示的仅水平条中包含的信息,则该决定能够更高效地做出。现在,假设数据存储系统100包括明显比仅所示10个磁盘更多数量的磁盘,例如90个额外的磁盘,并且数据也是类似进行分布;这种新颖系统和方法的效率将明显增加。

[0007] 因此,传统RAID清理操作不足以确定与要失效磁盘上的所有扩展区的相关联RAID条有关的期望信息。本领域中需要一种改进的方法来确定磁盘或磁盘扩展区级处的该信息。更一般地,本领域中需要一种系统和方法来调查数据存储系统的潜在错误,并特别地,需要一种系统和方法来在磁盘失效之前调查数据存储子系统或其他信息处理系统的潜在错误,从而改进容错性。

## 发明内容

[0008] 在一个实施方式中,本公开涉及一种在数据存储子系统的将要失效的磁盘于失效之前调查该数据存储子系统的潜在错误并恢复重建该将要失效的磁盘的数据所使用的不可读数据的方法,其中该数据存储子系统提供多个磁盘的存储摘要。该方法包括:确定多个磁盘中的磁盘达到被识别为将要失效的磁盘的阈值,并在将要失效的磁盘于失效之前调查多个磁盘中的其他磁盘上的至少一部分数据以识别具有潜在错误的数据存储区域。该方法还包括至少部分地利用存储在将要失效的磁盘上的数据来重建已识别的数据存储区域。在一些实施方式中,调查可以包括识别将要失效的磁盘上的RAID扩展区,识别RAID条中多个磁盘中的剩余磁盘上与在将要失效的磁盘上识别的RAID扩展区相对应的数据存储区域,并执行读取操作以识别具有潜在错误的数据存储区域。调查可以依赖于将要失效的磁盘的识别而自动发起,或者可以基于用户请求而被发起。在一些实施方式中,确定磁盘将将要失效的可以基于与磁盘有关的信息和/或磁盘的使用来执行。另外,确定磁盘将将要失效的可以

自动执行。在进一步的实施方式中，确定磁盘将要失效的可以包括根据预定标准来跟踪磁盘的错误，并在已跟踪错误的数量达到阈值数量的错误时将磁盘识别为将要失效的。将要失效的磁盘可以被移除和替换；将要失效的磁盘的数据可以之后利用来自多个磁盘中的剩余磁盘的数据被重建在替换磁盘上。在一些实施方式中，重建的数据存储区域可以被再次调查，以确认成功校正了潜在错误。

[0009] 在另一实施方式中，本公开涉及信息处理系统。该信息处理系统可以包括磁盘控制器和用于提供多个磁盘的存储摘要的数据存储子系统。磁盘控制器可以具有访问数据存储子系统的权限，并且被配置成：确定多个磁盘中的磁盘是否达到被识别为将要失效的磁盘的阈值；在将要失效的磁盘于失效之前，调查多个磁盘中的剩余磁盘上的至少一部分数据以识别具有潜在错误的数据存储区域；以及至少部分地利用将要失效的磁盘上存储的数据来重建已识别的数据存储区域。

[0010] 在另一实施方式中，本公开涉及一种在数据存储子系统的将要失效的磁盘于失效之前调查该数据存储子系统的潜在错误并恢复重建该将要失效的磁盘的数据所使用的不可读数据。该方法可以包括根据预定标准来自动跟踪数据存储子系统的多个磁盘的错误并在该磁盘的已跟踪错误达到预定阈值时将磁盘识别为将要失效的磁盘。在将要失效的磁盘完全于失效之前，条中多个磁盘中的剩余磁盘上与将要失效的磁盘上的扩展区相对应的数据存储区域可以被识别并被调查潜在错误。具有潜在错误的数据存储区域可以至少部分地利用存储在将要失效的磁盘上的数据进行重建。

[0011] 虽然公开了多个实施方式，但是根据示出并描述了本发明示例性实施方式的具体实施方式部分，本公开的其他实施方式对于本领域技术人员而言将是显而易见的。如将意识到的，本公开的各种实施方式能够在各种显而易见的方面中进行修改，并且都不背离本公开的精神和范围。因此，附图和详细的描述实际上被认为是示例说明，而不认为是限制性的。

## 附图说明

[0012] 虽然本说明书以特别指出并明显要求保护被认为形成本公开各种实施方式的主题结束，但是应当相信，本发明将从以下结合附图的描述中得到更好的理解，其中：

[0013] 图1是示意和概念示例性数据存储系统，其具有采用三个RAID配置而被分布在10个磁盘上的数据，其示出了利用传统数据清理来增加容错性的普遍问题。

[0014] 图2是适用于本公开各种实施方式的磁盘系统的示意图。

[0015] 图3是根据本公开一个实施方式的相对简单数据存储系统的示意图，其示出了调查数据存储系统的潜在错误的方法。

## 具体实施方式

[0016] 本公开涉及一种用于调查数据存储系统的潜在错误的新颖且有利的系统和方法。特别地，本公开涉及在磁盘失效之前调查数据存储子系统或其他信息处理系统的潜在错误的新颖且有利的系统和方法，从而改进了容错性。

[0017] 出于本公开的目的，信息处理系统可以包括任何手段或手段的集合，所述手段或手段的集合可操作以计算、估算、确定、分类、处理、传送、接收、检索、创建、切换、存储、显

示、传递、表明、检测、记录、复制、操作或利用任何形式的用于商务、科学、控制或其他目的信息、智能或数据。例如，信息处理系统可以是个人计算机（例如，台式或膝上型计算机）、平板计算机、移动设备（例如，个人数字助理（PDA）或智能电话）、服务器（例如，刀片服务器（blade server）或机架服务器（rack server））、网络存储设备或任意其他合适的设备，并且其尺寸、形状、性能、功能和价格可以变化。信息处理系统可以包括随机存取存储器（RAM）、一个或多个处理资源（例如中央处理单元（CPU）或硬件或软件控制逻辑）、ROM、和/或其它类型的非易失性存储器。信息处理系统的其他部件还可以包括一个或多个磁盘、一个或多个用于与外部设备通信的网络端口、以及各种输入和输出（I/O）设备（例如键盘、鼠标、触摸屏和/或视频显示器）。信息处理系统还可以包括一条或多条总线，所述总线可操作以在各种硬件部件之间传送通信。

[0018] 虽然各种实施方式并不局限于任何特定类型的信息处理系统，但是本公开的系统和方法在磁盘系统或虚拟磁盘系统（诸如2009年11月3日发布的发明名称为“Virtual Disk Drive System and Method”的美国专利No.7,613,945中所描述的，其全部内容通过引用都合并到本公开中）中会是特别有用的。这种磁盘系统通过基于例如RAID至磁盘的映射来在存储器的页池、或磁盘存储区块矩阵或多个磁盘之间动态地分配用户数据，来允许数据的高效存储。通常，动态分配向用户服务器呈现了虚拟磁盘设备或卷。对于服务器而言，卷（volume）与传统存储器（诸如磁盘）起的作用相同，但是提供了多个存储设备（例如RAID设备）的存储摘要以创建动态大小的存储设备。依赖于例如但不局限于数据类型或数据的访问模式，可以在这种磁盘系统中利用数据分级存储（data progression）来逐渐地将数据移动到具有恰当的整体数据成本的存储空间。通常，数据分级存储可以通过考虑例如物理存储设备的货币成本、物理存储设备的效率和/或逻辑存储设备的RAID级别来确定磁盘磁盘系统中的存储成本。基于这些确定，数据分级存储可以相应地移动数据，以便数据被存储在可用的最恰当成本的存储器上。另外，这种磁盘系统可以通过以例如预定时间间隔、用户配置的动态时间戳（诸如每隔几分钟或几小时等）或以服务器指定的时间自动生成并存储该系统或磁盘存储区块矩阵的快照或时间点副本，来避免数据遭受系统故障或病毒攻击。这些带有时间戳的快照允许在系统失效之前从之前时间点恢复数据，从而将系统恢复为那个时间时所呈现的系统。这些快照或时间点副本还可以在主存储器能够保持可操作的同时被系统或系统用户用于其它用途，诸如但不局限于测试。通常，通过使用快照功能，用户可以查看存储系统在之前时间点时所呈现的状态。

[0019] 图2示出了信息处理系统环境202（诸如美国专利No.7,613,945中所公开的并适用于本公开的各种实施方式）中的磁盘或数据存储系统200的一个实施方式。如图2所示，磁盘系统200可以包括数据存储子系统204（如本领域技术人员将意识到的，其可以包括RAID子系统）和磁盘管理器206（其具有至少一个磁盘存储系统控制器）。数据存储子系统204和磁盘管理器206能够基于例如RAID至磁盘的映射或其他存储映射技术来将数据动态地分配在多个磁盘208的磁盘空间上。

[0020] 如上所述，已经引入数据清理作为周期性地读取并检查RAID系统中的所有区块以在这些区块被使用之前检测坏区块的方式。然而，传统RAID清理操作在某一时间对单个RAID设备起作用并对RAID逻辑区块地址起作用，而非“垂直地”也即概念地说对磁盘或磁盘扩展区起作用。另外，在具有多个磁盘层级（tier）的系统中，传统RAID清理操作没有针对磁

盘类型(诸如具有更高失效趋势的那些磁盘)划分优先顺序。因此,通过采用传统RAID清理操作,通常没有方法来快速且高效地确定磁盘是否被怀疑将要失效,以便那个磁盘上所有扩展区的关联RAID条能够被读取以重建驻留在该要失效磁盘上的所有数据或尽可能多的数据。在与存储层级中的所有磁盘相关联的所有RAID设备上启动清理是太慢并消耗太多的资源。

[0021] 本公开对数据存储系统或其他信息处理系统(诸如但不局限于美国专利No.7,613,945中描述的数据存储系统的类型)中的传统数据清理和容错性进行了改进。特别地,本公开涉及(本文称为但不受该名字限制的)针对数据存储子系统或信息处理系统的数据调查器或RAID调查器(或简称为“调查器”(Surveyor))。所公开的实施方式能够提供针对过多错误的改进的容错性以及改进的后端永续性(resiliency)。

[0022] 通常,这里描述的调查器的各种实施方式的功能是在磁盘发生故障之前调查数据存储系统的潜在错误并对重建失效磁盘所需的任何不可读数据进行恢复。也就是说,通常,调查器可以读取磁盘或RAID扩展区并在重构操作之前对任何潜在错误或不可读数据区域进行写校正。如上面所讨论的,该调查器可以将操作瞄准RAID扩展区级而非传统数据清理RAID设备级。也就是说,给定了目标要失效磁盘,该调查器可以确定驻留在那个磁盘上的所有扩展区。对于每个已识别的扩展区,调查器可以读取所有关联的条数据,可选地排除目标磁盘上的条数据,因为那个数据在重建目标磁盘的数据时不需要,并且因为大概目标磁盘将要失效或者具有其他问题,所以向目标磁盘发送不需要的I/O会是特别不期望的。调查器之后可以在重构目标磁盘的操作之前对任何潜在错误或不可读数据区域进行写校正。实际上,一旦所有周围潜在错误被校正,要失效磁盘就能够被替换并能够相对容易地被重构。

[0023] 在更具体的程度上,本公开的实施方式参照图3所示的相对简单的数据存储系统示例302进行描述,其中图3示出了在三个条312、314和316中在区块中分布的数据遍布了4个磁盘或磁盘扩展区304、306、308、310(分别被标记为“磁盘0”、“磁盘1”、“磁盘2”和“磁盘3”)。当前,没有数据被写入条318中。虽然本公开的各种实施方式能够利用并使用任何RAID配置和级别,但是在图3中,为了易于说明,选择了遍布四个磁盘或磁盘扩展区的RAID 5配置。在RAID 5范例的情况下,奇偶校验信息(parity information)连同数据一起被分布在磁盘或磁盘扩展区上。除了其中一个磁盘之外的所有磁盘必须恰当操作,并因此RAID阵列的操作将不受到单个磁盘故障的破坏。更特别地,一旦单个磁盘失效,对该故障磁盘的任何读取都能够根据RAID阵列中正在操作的磁盘上的其余分布数据和奇偶校验信息来计算。尽管如此,单个磁盘失效典型地导致整个RAID阵列的性能下降,直到故障磁盘被替换且数据被重构为止。

[0024] 假设图3中的磁盘2发生故障。磁盘2需要被替换,并且存储在磁盘2的数据区块中的数据需要在替换磁盘上进行重构。虽然在图3中磁盘0-3被标记为“磁盘”,但是应当意识到,可以利用任何适当的存储扩展区、存储设备或存储设备类型的组合,包括但不限于磁带和/或固态磁盘。在许多情况中,对数据进行重构能够是相当平静无事的。例如,数据区块320的数据可以根据数据区块322、324和326中分布在磁盘0、1和3上的其余数据和奇偶校验信息来高效地进行计算。

[0025] 然而,假设其中一个可操作磁盘或磁盘扩展区上存在潜在错误(诸如在磁盘0的数据区块328处示出的),致使数据区块328的数据不可读。一旦磁盘2出现相同故障并被替换,

为了在替换磁盘上重构磁盘2的数据区块330,将需要分别来自磁盘0的数据区块328、磁盘1的数据区块332和磁盘3的数据区块334的数据。然而,由于磁盘2已经失效且数据区块320不可读,所以存在着双重失效,并且致使数据区块330的数据不可恢复。因此,期望在磁盘失效之前对这种潜在错误进行校正。

[0026] 因此,在一个实施方式中,在磁盘失效之前,可以首先确定任何给定磁盘是被怀疑将要失效还是以其他方式在返回准确数据方面具有增加的困难。根据本公开,可以使用确定磁盘何时将要失效或即将失效的任何适当方法。在一个实施方式中,这可以至少部分地基于与磁盘相关的信息或该磁盘的公共使用信息(诸如但不局限于该磁盘已经服务了多长时间、该磁盘的类型和质量、该磁盘的大小、在该磁盘上执行的I/O操作的数量、该磁盘所处的存储层级或者任何其他适当的信息)进行确定。这种确定可以手动执行或自动执行。在额外的或可替换的实施方式中,可以通过自动跟踪在执行针对磁盘的I/O时返回的明显错误的数量来检测磁盘故障。当已经达到某阈值数量的错误时,该磁盘可以被标志或以其他方式被识别为被怀疑即将失效。这里使用的“即将失效”不意味着局限于仅实际即将失效风险中的磁盘,而是也意味着被怀疑失效或以其他方式被标志为要失效磁盘,而不必考虑它们是否实际上将要失效或处于实际即将失效风险中。实际上,依赖于确定这种状态的因素和方法,甚至良好操作的磁盘也可以被识别为被怀疑即将失效。而且,什么被确定为是“明显”错误可以因系统不同而异,但可以通常是针对任何给定系统的预定标准。在一个实施方式中,明显错误可以包括被确定是不准确的和/或是不可读的针对磁盘的读取操作的任何结果。当然,那只是什么可以被确定为“明显”错误的一个示例,而且不是确定明显错误的唯一方式。另外,被用来确定磁盘正接近即将失效的阈值数量的错误可以因系统不同、存储层级不同、磁盘类型不同等而异,并且可以是由系统、管理员或其他用户确定的、所期望的或需要的任何适当的阈值。另外,在其他实施方式中,该阈值不需要是特定数量的错误,而是可以是与一些参考数相比的错误百分比,可以是在指定时间段内或错误被接收的其他速率出现的特定数量的错误,或者可以是由用于识别限度或范围的任何其他适当方法所确定的阈值。

[0027] 在一个实施方式中,当磁盘失效被预料到或以其他方式被感测到时,调查器可以运行以一般性地确定在重构要失效磁盘上存储的数据所需的数据存储区域中是否存在任何潜在错误。如前所述,调查器可以依据要失效磁盘的检测而自动运行,或者可以应用户或管理器请求而运行。在其他实施方式中,调查器运行的时刻不需要基于磁盘故障是否将要发生的确定,而是能够基于任何其他因素或以任何适当定时间隔(包括但不局限于周期性地、随机地、基于用户请求、持续地等等)手动或自动地运行。

[0028] 通常,如上所述,调查器可以确定驻留在那个磁盘上的所有扩展区,并且对于每一个已识别扩展区而言,其可以读取所有关联的条数据,可选地排除目标磁盘上的条数据,因为那个数据在重建目标磁盘的数据时不需要。如果任何读取操作揭示了潜在错误,则调查器可以之后对不可读数据区域进行写校正。因此,潜在错误可以在完全失效之前被校正,并重构针对目标磁盘的操作。

[0029] 特别参考图2的数据存储系统,例如,可以确定仅磁盘2已经返回或者正在返回太多的明显错误以达到或超过用于识别磁盘被怀疑即将失效的预定阈值。一旦做出这种检测,调查器可以确定驻留在那个磁盘上的所有扩展区,并且对于每一个已识别扩展区而言,

其可以读取排除磁盘2上的条数据后的所有关联的条数据,因为那个数据在重建磁盘2的数据时不需要。因此,在图3的简化示例中,调查器可以将数据区块320、330和336识别为驻留在要失效的磁盘2上。对于已识别数据区块320、330和336中的每一者,调查器可以在磁盘2完全失效之前识别关联条312、314和316中需要被调查潜在错误的数据区块。在该示例中,调查器将识别并因此调查或尝试读取条312的数据区块322、324和326、条314的数据区块328、332和334以及条316的数据区块338、340和342。调查器将不需要调查条318中的任何数据区块,因为磁盘2没有数据被写入那个条中。当对条314的数据区块进行调查时,调查器将识别与数据区块328相关联的潜在错误。为了校正该潜在错误,调查器可以分别利用磁盘1的数据区块332、磁盘2的数据区块330和磁盘3的数据区块334来重建数据区块328的数据。调查器校正数据区块328中的潜在错误是可能的,因为在磁盘2实际失效之前磁盘2的被怀疑失效已经被识别了,因此数据区块330中的数据依然可用于重建数据区块328。调查器将针对每个识别的潜在错误执行相同动作。然而,由于在图3的示例中将没有发现其他潜在错误,所以磁盘2能够之后被移除并用新的磁盘替换,并且磁盘2的数据能够容易地根据磁盘0、1和3的数据进行准确重建。

[0030] 在另一实施方式中,调查器可以再次调查之前被识别为具有潜在错误的扇区,以确定条的读取是否已经得到改进,并因此确定潜在错误实际被移除了。然而,在其他情况下,在对潜在错误进行写校正时可以假定潜在错误是固定的。

[0031] 虽然上面参照图2的相对简单的采用了RAID 5的数据存储系统进行了讨论,但是更复杂的数据存储系统(包括具有更少或更多磁盘或磁盘扩展区的数据存储系统,以及存储空间以任何组织或未组织的方式被划分成多个逻辑区块)被认为位于本公开的精神和范围内。类似地,RAID不需要被采用,但是如果被采用,则在数据存储系统中可以使用任何RAID级别或RAID级别的组合。图2的示例仅出于易于说明的目的,并且不意欲局限于这种简单的数据存储系统。例如,图1提供了具有更复杂RAID配置的示例性数据存储系统,如上面指出的,其将受益于本公开的各种实施方式。通常,这里描述的针对数据调查器或RAID调查器的系统和方法可以应用于利用某类冗余方案的任何数据存储系统或其他信息处理系统上,其中要失效的存储扇区的数据可以根据存储在一个或多个其他可操作存储扇区处的数据进行重构,无论冗余是否涉及现在存在的或以后开发的镜像、RAID或其他存储技术。这里描述的针对数据调查器或RAID调查器的系统和方法的各种实施方式有助于保持冗余数据的准确性并因此增加容错性。

[0032] 另外,在具有额外冗余的实施方式中(诸如双冗余RAID设备,例如RAID 6和RAID 10),调查器的约束可以因这种系统的升高的冗余以及内在增加的容错性而放松。例如,调查器可以容忍或忽略每个数据条的单个读取失败或潜在错误,因为即使具有一个读取失败,在这些系统上仍然存在着足够的内在冗余以便在没有要失效磁盘的情况下重建数据。然而,这里描述的调查器能够进一步改进这种系统的效率和容错性,特别在不止一个潜在错误存在于数据条上的情况下。

[0033] 另外,虽然参照调查器进行了公开,但是应当意识到,调查器能够包括一个或多个此种调查器并且调查器可以包括单独软件和硬件部件的任何组合和/或可以在数据存储子系统的一个或多个磁盘控制器上被操作。实际上,调查器可以包括多个软件部件,每个软件部件执行特定功能,其中任何一个软件部件操作一个或多个互连的或不连接的硬件部件。

[0034] 如上面讨论的,本公开的涉及用于在磁盘失效之前调查数据存储子系统或其他信息处理系统的潜在错误的系统和方法的各种实施方式提供了与传统数据清理方法相比的明显优势。例如,本公开的各种实施方式可以通过帮助避免多个读取失败(无论是源自于磁盘失效还是潜在错误)来增加容错性,从而导致不可恢复的数据并导致在完全磁盘故障之前快速且高效地检测潜在错误。因此,本公开的系统和方法能够用于避免不可恢复的场景并用于通过消除重构期间利用传统RAID清理操作所需要的非生产性工作(non-productive work)来改进重构时间。

[0035] 在前面的描述中,已经出于说明和描述的目的呈现了本公开的各种实施方式。它们不意欲是排他性的或者将本发明局限于所公开的精确形式。鉴于上面的教导,各种修改和改变都是可能的。已经选择和描述了各种实施方式以提供对本公开的原理及其实际应用的最佳说明,以及使得本领域普通技术人员能够将具有各种修改的各种实施方式用于所设想的特定用途。当根据所附权利要求书合理、合法且公正地赋予的广度进行解释时,所有这些修改和改变都位于本公开的由所附权利要求书所确定的范围内。

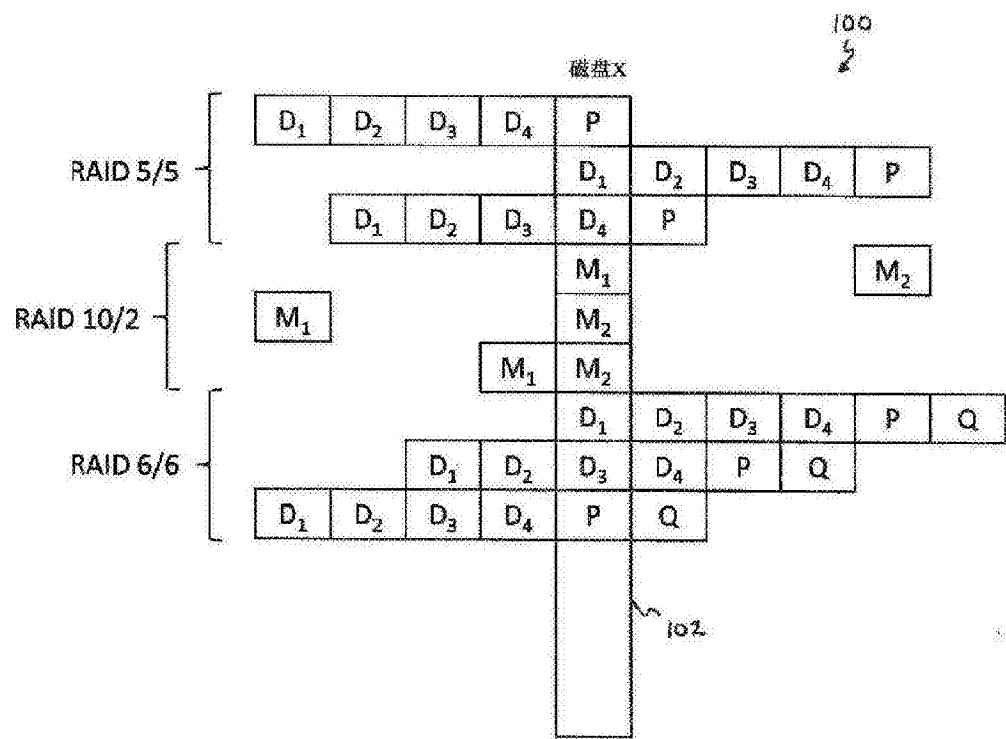


图1

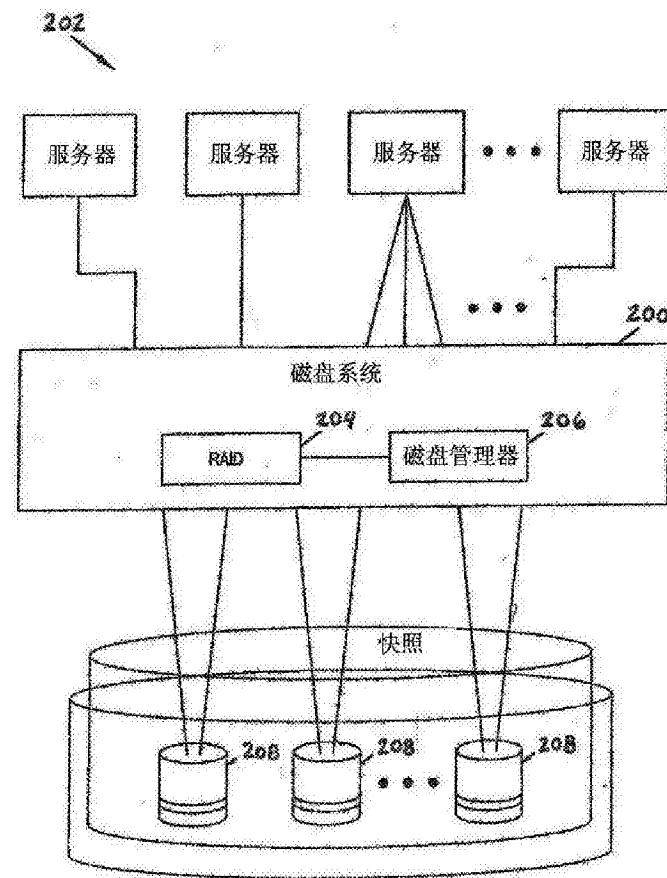


图2

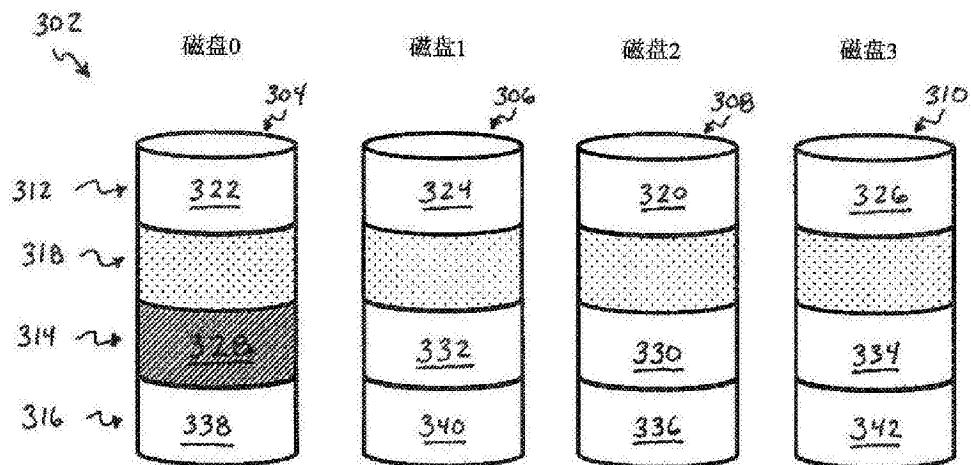


图3