



[12] 发明专利说明书

专利号 ZL 03820109.7

[45] 授权公告日 2008年4月2日

[11] 授权公告号 CN 100378672C

[22] 申请日 2003.6.24 [21] 申请号 03820109.7

[30] 优先权

[32] 2002.6.24 [33] US [31] 10/179,465

[86] 国际申请 PCT/US2003/020119 2003.6.24

[87] 国际公布 WO2004/001600 英 2003.12.31

[85] 进入国家阶段日期 2005.2.24

[73] 专利权人 网络装置公司

地址 美国加利福尼亚州

[72] 发明人 S·R·克莱曼 S·H·斯特朗格

[56] 参考文献

US5948110A 1999.9.7

US5657439A 1997.8.12

US5657468A 1997.8.12

US5313585A 1994.5.17

CN1350674A 2002.5.22

审查员 姜玲玲

[74] 专利代理机构 中国专利代理(香港)有限公司

代理人 李亚非 王勇

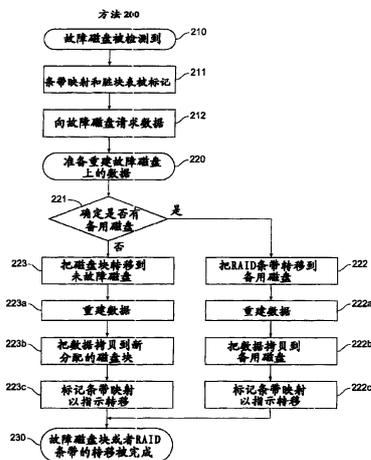
权利要求书2页 说明书15页 附图4页

[54] 发明名称

在 RAID 数据重建和转移中使用文件系统信息

[57] 摘要

当磁盘故障(210)时,存储系统只转移包括已分配数据的那些磁盘块,并在可能的情况下把未分配的磁盘块视为逻辑上为零。当没有备用磁盘时,源磁盘块被逻辑上置零并且重新计算与源磁盘块相关联的 RAID 条带的奇偶性(223)。当存在备用时,备用上的未分配块在迁移时逻辑上或物理上被置零(222)。故障磁盘的写入操作被重定向到其它未故障磁盘,并且保存该使用中的磁盘块已经被如此“转移”到那些其它的未故障磁盘的记录。未使用磁盘块被提前主动置零。使用关于已分配磁盘块的信息,并通过拷贝那些包括已分配数据或奇偶性的数据块,以及通过在镜像清除那些不包括任何已分配块的数据块,目标镜像拷贝被创建。



1. 一种在廉价磁盘冗余阵列数据重建和转移中使用文件系统信息的方法，包括：

确定第一群数据块中的故障数据块，所述第一群数据块包括保存与所述第一群数据块相关联的冗余信息的块；

只有当所述故障数据块存储分配的数据时，重建所述分配的数据并把所述分配的数据发送到与第二群相关联的目标数据块，所述第二群至少具有一个未故障的数据块；

保存关于所述故障数据块的信息，所述信息指示所述故障数据块已经被从第一磁盘迁移到多个第二磁盘之一；

逻辑上把所述第一群中的所述故障数据块设置为选定值，以及为所述第一群重新计算所述冗余信息。

2. 权利要求1的方法，包括：

在第一磁盘的全部故障数据块已经被迁移到多个第二磁盘时，逻辑上从存储系统除去所述的第一磁盘。

3. 权利要求1的方法，包括：

识别访问所述故障数据块的请求；以及

把所述访问请求重定向到所述目标数据块。

4. 一种在廉价磁盘冗余阵列数据重建和转移中使用文件系统信息的方法，包括：

响应于廉价磁盘冗余阵列组的第一磁盘的失败事件，对于所述第一磁盘的每个磁盘块，

只有当所述磁盘块存储分配的数据时，执行以下操作：

将所述磁盘块迁移到多个第二磁盘之一的目标磁盘块；

保存指示所述磁盘块已经被迁移到所述目标磁盘块的信息；

逻辑上把所述磁盘块设置为零，以及

重新计算对应于所述磁盘块的奇偶块。

5. 权利要求4的方法，进一步包括逻辑上从所述廉价磁盘冗余阵列组除去所述第一磁盘。

6. 权利要求4的方法，进一步包括把访问所述第一磁盘的磁盘块的请求重定向到对应于所述磁盘块的目标磁盘块。

7. 权利要求 4 的方法，其中指示所述磁盘块已经被迁移到所述目标磁盘块的信息被保存在条带映射中。

在RAID数据重建和转移中使用文件系统信息

发明背景

1. 发明领域

本发明涉及在RAID数据重建和转移(migration)中使用文件系统信息。

2. 相关技术

在计算机系统中保存大量存储的一个已知方法被称作RAID(“廉价磁盘冗余阵列”)。RAID存储器包括一个磁盘群,其中至少部分被用来储存数据和至少部分被用来存储与那些数据相关的奇偶信息,其中数据和奇偶性被分配到多个磁盘。如果那个储存数据或奇偶性的磁盘块故障或整个磁盘都出现故障,则允许RAID存储系统(从剩余的存储数据和奇偶信息)重建原始数据。在被称作RAID条带的一组磁盘块中,数据和奇偶性是彼此相关的。

RAID存储器的一个已知问题是,如果驱动器出现故障,则RAID存储系统不再具有相同的冗余度,而如果希望保存相同的冗余度,则必须对其做出调整。即,希望RAID存储系统允许迟些的第二磁盘故障而不造成数据丢失,但是重建RAID条带中两个分离的故障磁盘的数据是不可能的。为了保存相同的冗余度,RAID存储系统具有从故障磁盘重建数据到其上的备用磁盘,或(如果存在足够的可用空间),则从故障磁盘把数据转移到其它磁盘。在后一种情况中,与每个RAID条带相关的奇偶性必须响应于所分配数据的移动而被调整。然而,数据重建和奇偶性重算包括实质性的磁盘驱动器读写操作,并且最小化磁盘访问量和在磁盘和其余的计算机系统之间来回传送的信息量将是有利的。

因此,有利的是提供一个存储系统(其可能是RAID存储系统或另一类型的存储系统),其保存数据的可靠性、允许逻辑上删除故障磁盘、和最小化计算量、执行数据转移或重建的磁盘存取量,或奇偶性调整量。这些和其它优点在本发明实施例中提供,其中,文件系统保存的信息(包括磁盘块实际上用于记录所存储的数据的相关信息)被用于和移动,以及奇偶性的重算。本发明最小化计算和磁盘存取的优点也可以用于存储系

统的一般操作，以及用于把存储卷从源存储系统镜像到目标存储系统。

发明内容

本发明提供一种方法和系统，它保持数据可靠性、允许逻辑删除故障磁盘、和最小化执行数据转移或重建或奇偶性重算所需的计算量和磁盘存取量。在优选实施例中，当执行数据重建和转移，以及奇偶性重算时，存储系统使用实质上只和所分配的磁盘块有关的信息，并忽略未分配的磁盘块，未分配的磁盘块已经被设置为一个预定值，例如零。当磁盘故障时，存储系统只转移故障磁盘上已分配数据的那些磁盘块，而在可能的情况下把未分配的磁盘块视为逻辑上为零。

在本发明的一方面中，文件系统保存的信息识别在故障磁盘上实际被分配的磁盘块。当有备用磁盘或没有备用磁盘时，RAID群中的磁盘都可能出现故障。当磁盘块将从故障磁盘被转移而没有备用磁盘时，那些磁盘块从故障磁盘被转移到RAID群中的其它未故障磁盘，因而减少了RAID群中的磁盘数量。当那些磁盘块被转移时（即，数据被重建和拷贝到目标磁盘块上），源磁盘块逻辑上被设置为零并且为与源磁盘块相关的RAID条带重新计算奇偶性。一旦所有已分配磁盘块都已经从故障磁盘被转移，故障磁盘就可以逻辑上从RAID群中除去。当磁盘块将从故障磁盘被转移到备用磁盘时，那些磁盘块用类似方法被转移，不同在于，当转移的时候，备用磁盘上未分配的存储块逻辑上或物理上被设置成选定值（零）。

在本发明的一方面中，通过重定向写入操作（预定用于故障磁盘上的磁盘块）到其它未故障磁盘，文件系统主动地协助RAID存储系统。文件系统还保持一个记录，即哪些使用中的磁盘块已经被这样“转移”到那些其它的未故障磁盘，因而提供了一个指示符，其指示最初保存在故障磁盘上的存储数据已经完全被转移。当写入磁盘时，文件系统提前主动把未使用磁盘块设置为选定值（零），从而在写入以及迟些写入的时候最小化奇偶性的重算量。在优选实施例中，至少在下列情况中，文件系统把未使用磁盘块设置为零：（1）当新磁盘被添加到RAID群时，（2）当包括未使用磁盘块的RAID条带被写入时。在优选实施例中，文件系统保存“一致性点”，其中已存储的数据被确保自相一致，并在将一致性点写入磁盘时把未使用磁盘块设置为零。

在本发明的一方面中，RAID存储系统可能使用一种技术（比如RAID级

别5)，其中，奇偶性存储块可能在不同的磁盘上被找到。当那些磁盘之一故障时，文件系统识别与故障磁盘上奇偶信息相关的RAID条带，并把整组磁盘块从那些条带转移到其它磁盘，以便把那些全部条带的奇偶信息保存在未故障磁盘上。

在本发明的一方面中，存储系统使用已分配磁盘块的相关信息来改进RAID存储系统执行的操作。当写入任何一个RAID条带时，存储系统响应于磁盘块被分配而确定通过减去还是重新计算来计算奇偶性更加有效。

在本发明的一方面中，通过拷贝(到镜像拷贝)那些包括已分配数据和奇偶性的那些存储块，和通过清除(在镜像拷贝)那些不包括任何已分配数据或奇偶性的存储块，存储卷的目标镜像拷贝可以使用与已分配磁盘块相关的文件系统信息而被创建。因此，目标存储卷可以响应于源存储卷而被构造，而不需要任何奇偶性重算。

本发明通常适用于文件系统和存储系统，其中，数据被保存在多个装置上，和其中，那些多个装置上的至少一部分记录信息可以被丢弃(例如不是文件系统的一部分或可以被逐出高速缓存的高速缓存对象)。这些应用没有被具体限制到保存一致性点的文件系统，也没有被具体限制到RAID存储系统，它们也不局限于与在此公开的具体应用相关的系统。

附图简述

图1示出文件系统和RAID存储系统的框图，包括在RAID数据重建中使用文件系统信息。

图2示出包括在RAID数据重建中使用文件系统信息的方法的处理流程图。

图3示出一个系统框图，该系统不重新计算奇偶性就能够镜像。

图4示出不用重新计算奇偶性的镜像方法的处理流程图。

优选实施例具体描述

在此处的说明书中，本发明的优选实施例被描述，包括优选的处理步骤和数据结构。所属领域技术人员在细读此申请之后将会了解，本发明的那些实施例可以使用没有具体描述的多种其它技术来实现，而不用过度的试验或进一步发明，而且这类其它技术将在本发明的范围和精神

内。

词汇

下列术语涉及或者是指本发明的方面或其实施例。其中每个术语的一般意义被定为说明性的而绝不是限制性的。

一致性点 - 通常，这指的是由自相一致的文件系统保存的可识别数据集，因此它可以被保存在磁盘上而不用担心一致性点内的数据引用会导致文件系统错误。词组“一致性点”被定为足够宽以覆盖有其自相一致性的文件系统以及那些原子地自相一致的文件系统，该自相一致性由用于近期改变的一组日志项来保证。

磁盘块 - 通常，这指的是存储数据的大容量存储系统中可分配的部分。词组“磁盘块”被定为足够宽以覆盖磁盘上量化或未量化的可分配空间。

文件系统 - 通常，这指的是直接在存储系统上管理磁盘块的任何应用程序，包括保存命名方案和在大容量存储器上保存的文件数据之间的关联的系统。词组“文件系统”被定为足够宽以覆盖文件系统的变形，包括那些直接对磁盘读写的系统和那些允许不同的子系统或操作系统其它部分来读写磁盘的系统。

存储系统 - 通常是用于储存数据的系统，例如在RAID阵列中排列的一组一个或多个磁盘。存储系统可以被分成一个或多个卷，每个卷起一个存储系统的作用。术语“存储系统”和“卷”或“存储卷”可以被可交换地使用。

磁盘块的转移 - 通常，这指的是用于从已被丢失的磁盘块把数据拷贝或重建数据到大容量存储系统不同部分上的任何技术，比如没有故障因而可用的磁盘。

奇偶性重算 - 通常，这指的是用于在丢失之后响应于已存储的数据重建奇偶信息或其它冗余信息的任何技术，无论是否与其它冗余信息结合。

RAID群 - 通常，这指的是RAID存储系统内包括的和RAID存储系统用来保存数据冗余性的磁盘组。一部分系统可能包括RAID存储系统，RAID存储系统把它们的磁盘分成多于一个的RAID群。

RAID存储系统 - 通常，这指的是用于在大容量存储系统上保存数据

的任何技术，它包括一组冗余信息(比如可能的奇偶信息、汉明码、或类似实际数据拷贝的其它冗余形式)，并响应于丢失提供重建数据的可能性。

RAID条带 - 通常，这指的是磁盘块和冗余信息之间的任何关联，其中，磁盘块和冗余信息是相互依赖的因此至少一部分可以在丢失之后被重建。

重建 - 通常，这指的是用于在丢失之后响应于冗余信息来重建数据的任何技术，无论是否与其它已存储的数据结合。

零值映射(zeromap) - 通常，这指的是具有一组指示磁盘块逻辑上或物理上已经被设置为零的项目的表。

条带映射(stripmap) - 通常，这指的是具有一组指示哪些RAID条带已经从故障磁盘被转移的项目的表。

本发明的范围和精神不受限于任何这些定义，不受限于其中提及的具体例子，而是打算包括这些和其它术语具体化的最广泛的概念。

系统元件

图1示出文件系统和RAID存储系统的框图，包括在RAID数据重建中使用文件系统信息。

系统100包括文件系统120、存储系统140和上面两者之间的通信链路160。系统还包括用于运行文件系统120的处理器(未示出)、程序和数据存储器(未示出)。

文件系统120包括一组文件系统信息表121，其指示关于存储系统140中磁盘块(不管其是个别的还是成群的)的信息。文件系统信息表121可以被记录在存储器中，或者可以被记录在存储系统140被选择的部分中，或者其它方式(比如在非易失性存储器或其它辅助存储器装置中)，只要即使存储系统140丢失了数据，那些文件系统信息表121仍然是文件系统120可访问的。文件系统信息表121中的特定信息进一步被说明如下。

存储系统140包括一组磁盘141，其中每个包括一组磁盘块142，磁盘块142包括数据(存储在至少一些磁盘块142内)，在RAID条带组中排列的磁盘块142包括至少一些冗余信息(存储在至少一些磁盘块142内)。RAID条带143被处理，从而使得在任何个别磁盘141上的任何个别磁盘块142上的信息与在其它磁盘141上的其它磁盘块142上的信息相关。如果任何磁

盘块142乃至整个磁盘141丢失，这允许任何个别磁盘块142上的信息被重建或者重新计算。RAID存储系统在大容量存储系统的技术领域是已知的。

通信链路160连接文件系统120和存储系统140，并包括文件系统120和存储系统140可用来交换磁盘块142的信息和系统100操作信息的任何技术。在优选实施例中，通信链路160包括连接到文件系统120和存储系统140的总线。

文件系统信息表121包括块映射(blockmap)表122，块映射表122具有一组块映射项目123，其指示哪个磁盘块142被已分配的文件系统数据所使用。从而，块映射表122为每个磁盘141上的每个磁盘块142指示该磁盘块142是被文件系统120使用(例如由逻辑“1”值指示)还是不由文件系统120使用(例如由逻辑“0”值指示)。说明书中所用的“已分配的文件系统数据”不同于可能由文件系统120保存的任何临时数据结构，临时数据结构可能包括磁盘上分量。已分配的文件系统数据也不同于可能由文件系统120保存的文件系统数据的备份拷贝，比如在关于优选“WAFL”文件系统120的公开说明中描述的“瞬态(snapshot)”数据。

文件系统信息表121包括零值映射表124，其具有一组零值映射项目125，它指示哪个磁盘块142已经逻辑上或物理上被置零。

文件系统信息表121包括条带映射表126，其具有一组条带映射项目127，它指示哪个RAID条带143已经从故障磁盘被转移。在本发明的一方面中，其中，当那些RAID条带143包括故障磁盘上的奇偶信息时，RAID条带143被转移，每个条带映射项目127指示来自整个RAID条带143的磁盘块142是否已经被转移(因为奇偶信息本身不被转移)。

已分配数据的转移

如果磁盘块142中的数据被丢失(例如磁盘块142被损坏)，或磁盘141上的所有数据都丢失(例如磁盘故障)，存储系统140能够重建丢失数据。然而，对于任何丢失数据并在块映射表122中被指示没有任何已分配数据的磁盘块142来说，不必为该磁盘块142重建数据。从而，对于个别的磁盘块142，如果丢失但没有使用，则存储系统140无须重建任何数据。对于整个磁盘141，如果丢失，则存储系统140只须重建当时正在被使用的磁盘块142的数据。

如果没有备用磁盘，则文件系统120指示存储系统140重建已分配磁

盘块142的数据(即被块映射表122指示为已分配的磁盘块142),并把重建数据拷贝到存储系统140中相同RAID群中的其它未故障磁盘141上的磁盘块142上。在替换实施例中,被重建的数据可以被拷贝到存储系统140中不同RAID群中的其它未故障磁盘上。故障磁盘141的磁盘块142然后不再被写入,因此文件系统120将它们视为可用VBN(虚拟块号)空间中的“洞”。例如,如果故障磁盘是磁盘#1、#2、#3、#4和#5(奇偶性)中的磁盘#3,则磁盘#1、#2、#3、#4的VBN将仍然有效,而磁盘#3的VBN则无效。当所有来自故障磁盘141的数据都已经被除去时,故障磁盘141逻辑上或物理上从RAID群被除去,从而把RAID群中的磁盘141的数量减1。

所属领域技术人员将在细读本申请之后认识到,使用块映射表122来减少数据重建量允许存储系统140最小化工作量,同时又保存了相同的故障容许程度。所属领域技术人员在细读本申请之后也将认识到,本技术可以不用过度试验或新发明而被应用到已经容许多个故障的RAID系统和类似RAID的系统,比如科贝特(Corbett)或“偶奇的(EVENODD)”那些系统。更一般地说,所属领域技术人员在细读本申请之后也将认识到,本技术可以不用过度试验或新发明而被应用到没有其中必须放置数据或元数据的空间的所有存储系统。

当故障磁盘141上的磁盘块142的数据被转移时,存储系统140逻辑上清除了(置零)该磁盘块142的数据。文件系统120设置对应RAID条带143的零值映射项目125。存储系统140重新计算对应RAID条带143的奇偶性,并把重新计算的奇偶性写入对应RAID条带143中的奇偶性磁盘141。

如果存在备用磁盘,则文件系统120用同样的方式指示存储系统140转移来自自己分配磁盘块142的数据。对于未分配的磁盘块142,存储系统140写入另一未故障磁盘141上的对应磁盘块142,而不尝试重建来自于未分配的磁盘块142的数据。相反地,存储系统140物理上清除(置零)作为未分配磁盘块142转移目标的磁盘块142。在优选实施例中,文件系统120指示存储系统140使用SCSI“写入相同的(write same)”命令来指导具有作为转移目标的磁盘块142的磁盘141;这节省了磁盘141的活动和文件系统120与存储系统140之间的带宽。

条带映射表126包括每个RAID条带143的一个条带映射项目127,为RAID条带143指示RAID条带143中的磁盘块142(来自故障磁盘141)是否已经被转移到另一未故障磁盘141。当它的对应条带映射项目127指示特定

磁盘块142已经从故障磁盘141转移到另一未故障磁盘141时，存储系统140能够在未来的奇偶性重算期间考虑除去该特定的磁盘块142。更确切地说，在写入特定条带时，存储系统140通过假定指示的所有磁盘块142一律为零来重新计算奇偶性。当条带映射表126指示所有来自故障磁盘141的磁盘块142已经被转移到其它未故障磁盘141时，文件系统120和存储系统140可以逻辑上全面删除故障磁盘141。

访问请求的重定向

文件系统120在重定向请求中帮助存储系统140访问故障磁盘141上的磁盘块142。在优选实施例中，文件系统120包括写入时拷贝 (copy-on-write) 技术，其中，所有对磁盘块142的写入操作 (在任何磁盘141上，不包括刚才的故障磁盘141) 都通过从作为写操作目标的磁盘块142拷贝数据、修改该拷贝和把到目标磁盘块142的指针调整为指向新修改过的拷贝。在优选实施例中，当产生存储卷的一致性点时，这些修改被集合到一起；然而，对于这类集合没有特殊的要求。

如果磁盘141故障或磁盘块142被损坏，则文件系统120标记它的文件系统信息表 (包括具有脏块项目的脏块表) 以指示每个被丢失的磁盘块142被标记为脏。第一结果是任何对脏磁盘块142的写入尝试将导致写入时拷贝操作将被执行。第二结果是文件系统120将不迟于下一个一致性点，作为向磁盘写入一致性点的一部分而产生磁盘块142的拷贝。

因为用户对丢失磁盘块142的请求被重定向，并且因为文件系统120将不迟于下一个一致性点产生磁盘块142拷贝，所以存储系统140可以等候来自于文件系统120的指令而不需要通过积极地重建丢失的磁盘块142来响应。

磁盘块的提前主动置零

在某些情况下，文件系统120提前主动指示存储系统140把整个磁盘块142置零，从而允许容易地重算 (或不重算) 与包含磁盘块142的条带相关的奇偶信息。在优选实施例中，文件系统120指示存储系统140使用 SCSI “写入相同的” 命令，如上所述。

文件系统120在至少下列情况中把整个磁盘块设置为零：

- 当新的磁盘141被添加到RAID群中时，每个RAID条带143从而被加宽了一个磁盘块142，并且每个这类条带的奇偶性从而响应于新的磁盘块142中的数据。文件系统提前主动指示存储系统140把新的磁盘141中的所

有磁盘块142设置为零，而不是重新计算任何奇偶信息，从而使奇偶信息保持不变。

奇偶性优选地作为RAID条带143中磁盘块142的所有数据的模2和来计算，模2和也被称作异或运算(“XOR”)。因此，插入全是零的新磁盘块142没有改变RAID条带143中的奇偶性。

- 当包括未分配的磁盘块142(即，没有被文件系统120标记为“使用中”的磁盘块142)的RAID条带被写入时，文件系统120提前主动指示存储系统140把那些磁盘块中的数据设置为零。根据实现，这允许文件系统120或存储系统140，或者系统100的诸如fly-by XOR(快速XOR)子系统之类的其它组件重新计算那些条带的奇偶信息，而不用读取未使用的磁盘块142。无须读取未使用磁盘块142减少了存储系统140执行的读取操作量，并减少了文件系统120和存储系统140之间所用的通信带宽量。

- 当文件系统120准备向磁盘写入一致性点时，它通常写入磁盘141上的相对大量的磁盘块142。文件系统120尝试把那些磁盘块142集合到整个RAID条带143中，因此写入操作可以尽可能的有效，并且因此奇偶性计算可以被最小化(被看作已分配数据的每个磁盘块142的开销操作)。当写入RAID条带143时，文件系统120指示存储系统140清除RAID条带中不是一致性点一部分的那些磁盘块142(并且因此将被文件系统120标记为未分配的)。这允许奇偶性计算无须读取那些磁盘块142就开始进行。

奇偶性的有效计算

在本发明的一方面中，RAID存储系统执行的操作响应于目标RAID条带143中磁盘块142的零值映射表125。响应于那些零值映射表125，文件系统120可以对目标RAID条带143中的非零磁盘块的数量进行计数；这允许文件系统120或存储系统140在写入任何个别RAID条带143时确定，通过减法来计算奇偶性或通过重算来计算奇偶性是否更有效。

存储系统140可以通过减法来计算奇偶性，即当向磁盘141写入磁盘块142时，存储系统140可以从相关联的奇偶性(用于那些RAID条带143)减去磁盘块142中的旧数据并把将写入磁盘块142的新数据加到相关联的奇偶性上。该减法和加法都是逐位模二进行。通过减法来计算奇偶性在RAID存储系统的技术领域是已知的。可替换地，存储系统140可以通过加上(模二)那些RAID条带143的所有磁盘块142来重新计算奇偶信息。

当向磁盘141写入一组磁盘块142时，文件系统120确定通过减法计算

奇偶性是否将需要更少的磁盘操作，或从整个RAID条带143重新计算奇偶性是否将需要更少的磁盘操作。文件系统120可以从零值映射表124来确定这个；它可以确定RAID条带中是否有足够磁盘块是零，并且可以因此彻底地省去奇偶信息的计算。则RAID系统也可以简单地把条带143中的未分配块置零，如果它们没有被预先置零(由零值映射表指示)。这对于非WAFL文件系统来说是特别有用的，其中，更可能在条带中存在未分配块。

奇偶性丢失时的转移

在RAID级别4系统中，奇偶性被保存在存储系统140中的单个磁盘141上；即，所有的RAID条带都在相同的磁盘141上具有它们的奇偶性，其可能因此被称为“奇偶性磁盘”。在RAID级别5系统中，奇偶性分布在存储系统140中的多个磁盘141上；即，每个RAID条带可能在不同的磁盘141上具有其奇偶性，因此没有单个的“奇偶性磁盘”。

如果RAID级别5系统中的一个磁盘141故障，则那些将故障磁盘141用于其奇偶性的RAID条带不再具有奇偶性块，并且如果它们剩下的磁盘块142中的一个或多个丢失数据，则因此受到丢失信息的影响。文件系统120将那些RAID条带的磁盘块142标记为“脏”，因此下次一致性点被写入磁盘时，那些磁盘块142被写入具有有效奇偶性块的RAID条带143中的相同或其它磁盘上的不同单元。结果，RAID条带143中奇偶性丢失的磁盘块142被写入具有可用奇偶性块的其它RAID条带143中的磁盘块142(不一定都写入相同的RAID条带143)。

操作方法

图2示出包括在RAID数据重建中使用文件系统信息的方法的处理流程图。

方法200由系统100执行。尽管方法200被连续地描述，然而方法200的流程点和步骤可以用流水线或其它方法由串联或并联分离元件来执行，不管异步或同步。除非其中明确地指示，否则没有方法20必须与说明书列出的流程点或步骤相同的顺序来执行的特定要求。

已分配数据的转移

在流程点210，系统100已经检测到故障磁盘141。

在步骤211，文件系统120标记条带映射表126以指示还没有来自任何RAID条带143的磁盘块142已经从故障磁盘141被转移。文件系统120还标记脏块表以指示故障磁盘141上的所有磁盘块142都将被当做脏块。在方

法200中，第一RAID条带143指的是故障磁盘141上的RAID条带。第二RAID条带143指的是备用或其它未故障磁盘上的RAID条带。

在步骤212，文件系统120接收访问故障磁盘141上的磁盘块142之一的用户请求。因为用户请求只适用于包括文件数据或者元数据的已分配数据，所以方法200在流程点220开始从磁盘块142重建数据。

在流程点220，方法200准备从故障磁盘141上的磁盘块142重建数据。

在步骤221，存储系统140确定是否存在备用磁盘141。如果有，则方法200进行到步骤222。否则，方法200进行到步骤223。

在步骤222(存在备用磁盘141)，存储系统140把与磁盘块142相关的第一RAID条带143转移到备用磁盘141，并且方法200进行到流程点230。为了执行这个步骤，存储系统140执行下列的子步骤：

- 在子步骤222a，存储系统140从磁盘块142重建数据。存储系统140从第一RAID条带143使用另一个磁盘块142和奇偶性块。然而，存储系统140可以忽略第一RAID条带中的那些磁盘块142，对于该第一RAID条带，相关联的零值映射项目125指示磁盘块142一律地是零。

- 在子步骤222b，存储系统140从RAID条带143把数据拷贝到备用磁盘141上的第二目标RAID条带143。然而，存储系统140不拷贝未分配的那些磁盘块142，而是使用它们相关联的块映射项目123来确定拷贝哪一些。通过使用SCSI“写入相同的”命令，存储系统140提前主动清除第二目标RAID条带143中的那些未分配的块映射项目123的磁盘块142。存储系统140响应于它对那些磁盘块142的抢先的清除来重新计算奇偶性。

- 在子步骤222c，文件系统120标记对应的条带映射项目127以指示第一RAID条带143被完全转移到备用磁盘141。

在步骤223(没有备用磁盘141)，存储系统140把磁盘块142中的数据从故障磁盘141转移到另一个未故障磁盘141，并且方法200进行到流程点230。为了执行这个步骤，存储系统140执行下列的子步骤：

- 在子步骤223a，存储系统140从磁盘块142重建数据。存储系统140使用来自其它磁盘块142和第一RAID条带143的数据和奇偶性块。然而，存储系统140可以忽略第一RAID条带143中的那些磁盘块142，对于该第一RAID条带143，相关联的零值映射项目125指示磁盘块142一律地是零。

- 在子步骤223b，存储系统140把数据拷贝到新分配的磁盘块142中。存储系统140重新计算原始第一RAID条带143的奇偶性，假定故障磁盘块

142现在逻辑上是零。

- 在子步骤223c, 文件系统120标记与磁盘块142相关的第一RAID条带143的对应条带映射项目127, 从而指示来自磁盘块142的数据被转移到未故障磁盘141。文件系统120标记故障磁盘块142的对应零值映射项目125以指示磁盘块142现在逻辑上是零。

在流程点230, 故障磁盘块142或整个第一RAID条带143的转移已经完成。系统100重复转移直到所有已分配的磁盘块142都已经从故障141被转移到备用磁盘141或其它未故障磁盘141。

无须重新计算奇偶性的镜像

图3示出一个系统框图, 该系统不重新计算奇偶性就能够镜像。

如上所述, 当重新建立(或最初建立)储存卷的镜像拷贝(“卷”有时在此可用作“系统”的同义词)时, 通过当包括那些未分配块的条带被写入时保证目标存储卷上的未分配块被设置为零, 目标存储卷(也称作镜像存储卷)可以使用来自源存储卷的计算出的奇偶性。

无须再计算奇偶性就能够镜像的系统300包括含有源存储系统305的源系统301、类似于参考图1描述的系统100和存储系统140, 并且目标(或镜像)系统310包括目标存储系统320, 也类似于参考图1描述的系统100和存储系统140。源存储系统305包括一组含有磁盘块331的源RAID条带330; 目标存储系统320类似地包括一组含有磁盘块341的目标RAID条带。目标RAID条带340类似于源RAID条带330, 优选地, 在逻辑上和源RAID条带330一致。

源系统301和目标系统310使用通信链路350连接。在优选实施例中, 通信链路350包括光纤信道或SAN(存储区域网)。在其它实施例中, 通信链路350可以包括LAN(局域网)、WAN(广域网)、或其组合, 例如互联网连接。所属领域技术人员将认识到, 通信链路350可能包括用于把数据从源系统301发送到目标系统310的任何技术, 并且决不局限于在此描述的具体实施例。

源RAID条带330包括一组已分配块334、至少一个未分配块332和一组奇偶性块333。

当重新建立(或最初建立)源存储系统305和目标存储系统320之间的镜像关系时, 源系统301从源RAID条带330选择一组已分配块334以发送到目标系统310。在优选实施例中, 这些已分配块334在源系统301从文件系

统信息导出，并且只包括那些由源系统301处的文件系统指示为已分配的磁盘块。

源系统301把已选择的已分配块334连同与RAID条带330相关的奇偶性块333一起发送到目标系统310。此外，源系统301置零任何没有被预置零的未分配块332。目标系统310从这些已分配块334接收数据并将它们存储在其目标RAID条带340中的已分配块344的对应单元内。类似地，目标系统310接收相关联的奇偶性块333并把它们存储在其目标RAID条带340中的奇偶性块343的对应单元内。

因为目标系统310具有逻辑上和源RAID条带330一致的目标RAID条带340，所以目标系统310可以确定它自己所有的未分配块342在源存储系统305中是未分配的。目标存储系统320因此可以使用SCSI“写入相同的”命令把所有那些未分配块342设置为零。

结果，在磁盘块被发送、接收和存储之后，目标存储系统320和源存储系统305基本一致；从而，源存储系统305和目标存储系统320之间的镜像关系被重新建立。

镜像方法

图4示出无须重新计算奇偶性的镜像方法的处理流程图。

方法400由源系统301和目标(镜像)系统来执行。源系统包括源存储系统305和类似于文件系统120的源文件系统(未示出)。目标系统类似于目标系统310。与方法200相似，尽管方法400被连续地描述，然而方法400的流程点和步骤可以采用流水线或其它方法由串联或并联的分离元件来执行，不管异步或同步。同样类似于方法200，除非其中明确指示，否则方法400没有必须用与说明书列出流程点或步骤相同的顺序来执行的特定要求。

在流程点410，源系统301和目标系统准备重新建立镜像关系。在优选实施例中，源系统301和目标系统已经通信以便各自达到镜像关系即将被重新建立的状态。此外，在优选实施例中，源系统301和目标系统已经确定将从源系统301被发送到目标系统的磁盘块的最小集以实现镜像的重新建立。一个用于确定磁盘块最小集的方法在WO 02/29572A(网络仪器公司)中被进一步描述，名称为“Recovery of File System Data in File Servers Mirrored File System Volumes”，于2002年4月11日公布。

在步骤411，源系统301选择将被发送到目标系统的一组信息。如上

所述，只有已分配的磁盘块需要被发送。在这个步骤，被选择发送的信息包括(a)条带信息和(b)奇偶信息。条带信息描述被发送的磁盘块在RAID条带中怎样组织。奇偶信息包括为那些RAID条带计算的奇偶性。对于有关被定义RAID条带中哪个块是未分配块的信息而言，源系统301明确地发送那些信息，或者目标系统响应于条带信息和它接收的那些磁盘块的标识来确定该信息。

在步骤412，源系统301发送(和目标系统接收)已分配的磁盘块、在步骤411中描述的条带信息，和在步骤411中描述的奇偶信息。在这个步骤，源系统301也可以把RAID条带中没有预置零的未分配块置零。

在步骤413，目标系统把数据从已分配的磁盘块写入它们在指定磁盘驱动器及其存储卷的指定RAID条带上的指定位置。

在步骤414，目标系统把来自奇偶信息的数据写入其存储卷的指定RAID条带。

在步骤415，目标系统把零值写入其存储卷的指定RAID条带中的未分配块。在优选实施例中，目标系统使用SCSI“写入相同的”命令来把相同数据字节的拷贝(即，零)写入每个未分配块的每个单元；如上所述，这比向实际的磁盘块写零要快一些，并且占用的文件系统和存储卷之间的通信带宽较少。

因为在源系统301的未分配块逻辑上或物理上为零，所以当未分配块被假定为零时源系统301发送到目标系统的奇偶信息是正确的。因此，目标系统可以安全地把未分配块实际上设置为零，同时使用相同的奇偶信息而无须再计算奇偶性。

在流程点420，目标系统是源系统301的物理和逻辑拷贝。源系统301和目标系统之间的任何文档记载操作被完成，并且它们之间的镜像关系被重新建立。

方法400可以在重新建立或最初建立源存储卷和目标存储卷之间的镜像关系的任何时候被执行，只要在源存储卷和目标存储卷都使用了等效RAID条带。

本发明的通用性

本发明通常适用于这样的文件系统和存储系统，其中，数据被保存在多个装置上，并且那些多个装置上记录的至少一些信息可以被丢弃(例如不是文件系统的一部分或可以被逐出高速缓存的高速缓存对象)。这些

应用没有被具体限制到保存一致性点的文件系统，也没有被具体限制到RAID存储系统，它们也不一定与在此公开的具体应用相关。

在细读本申请之后，所属领域技术人员将会清楚，本发明在其大多数一般形式中的其它和进一步的应用。无须过度试验或进一步发明，本发明可以用于这类其它和进一步的应用。尽管优选实施例在此被公开，然而许多保持在本发明概念、范围和精神之内的许多变化是可能的；所属领域技术人员在细读本申请之后将很清楚这些变化。

- 本发明适用于其中数据和元数据没有在大容量存储器上被分配固定单元的任何存储系统。这可以包括文件服务器、数据库或网络高速缓存、或另一类型的存储装置。尽管在优选实施例中，本发明主要是用于使用RAID存储系统的文件服务器，然而此外也没有特定要求来限制本发明的适用性。

- 本发明适用于其中数据可以从冗余信息被重建的任何系统；这可以包括任何类型的存储系统，乃至使用至少一些冗余信息的通信系统。尽管在优选实施例中，本发明主要是用于使用多个磁盘驱动器和奇偶性的存储系统，然而此外也没有特定要求来限制本发明的适用性。

- 虽然术语“磁盘块”已经遍及本公开内容被使用，然而本发明同样适用于大容量存储系统中的其它类型的数据块，比如用于磁带、光驱动器的数据块等等。

所属领域技术人员在细读本申请之后将认识到，这些替换实施例是说明性并且绝不是限制性的。

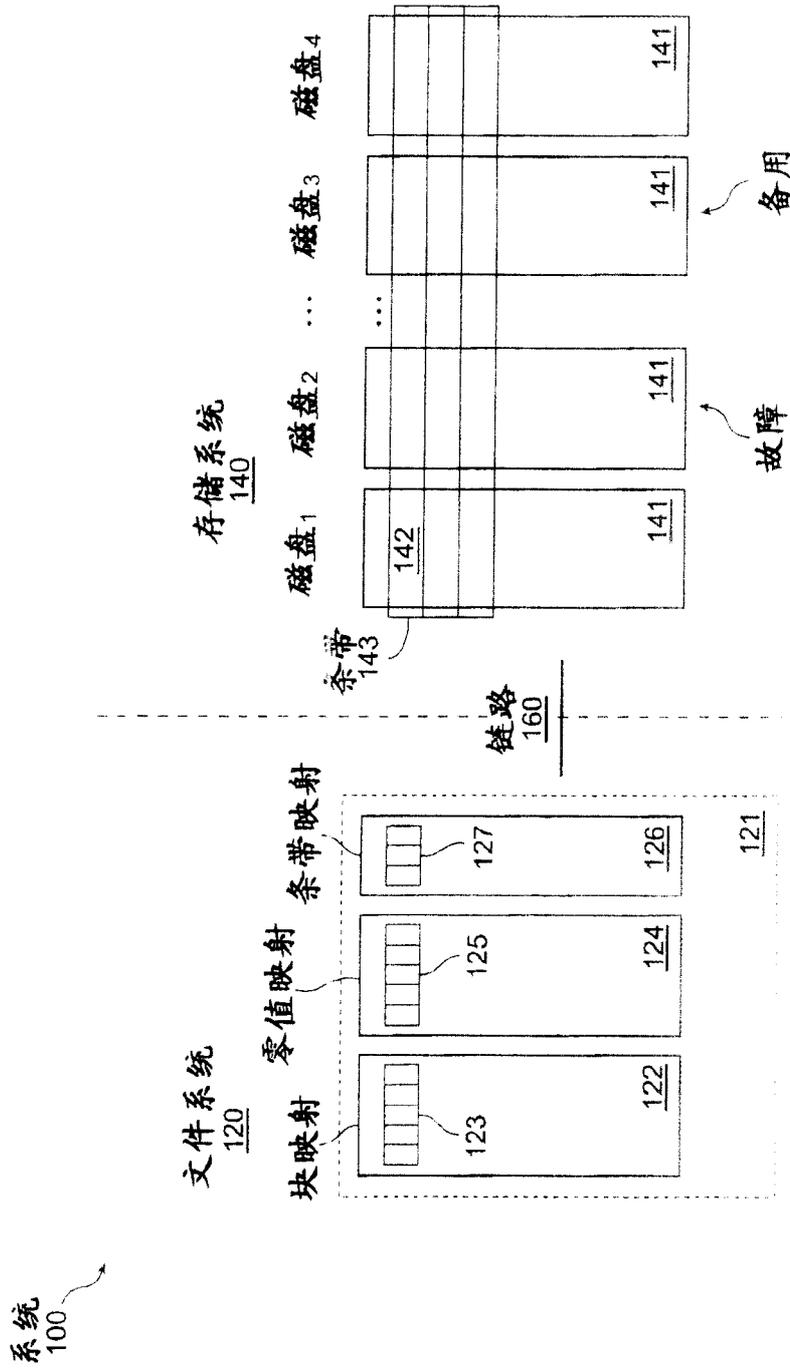


图 1

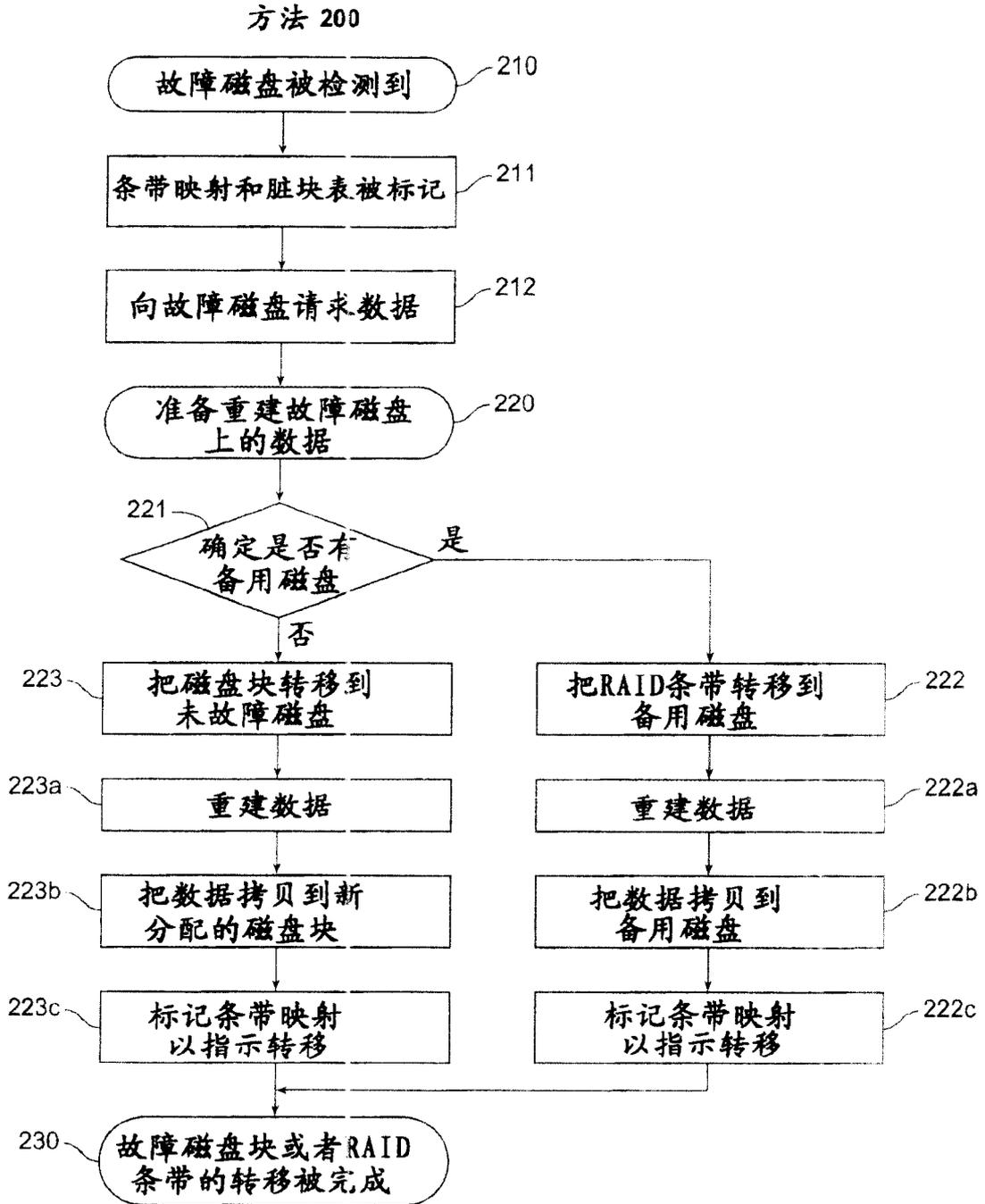


图 2

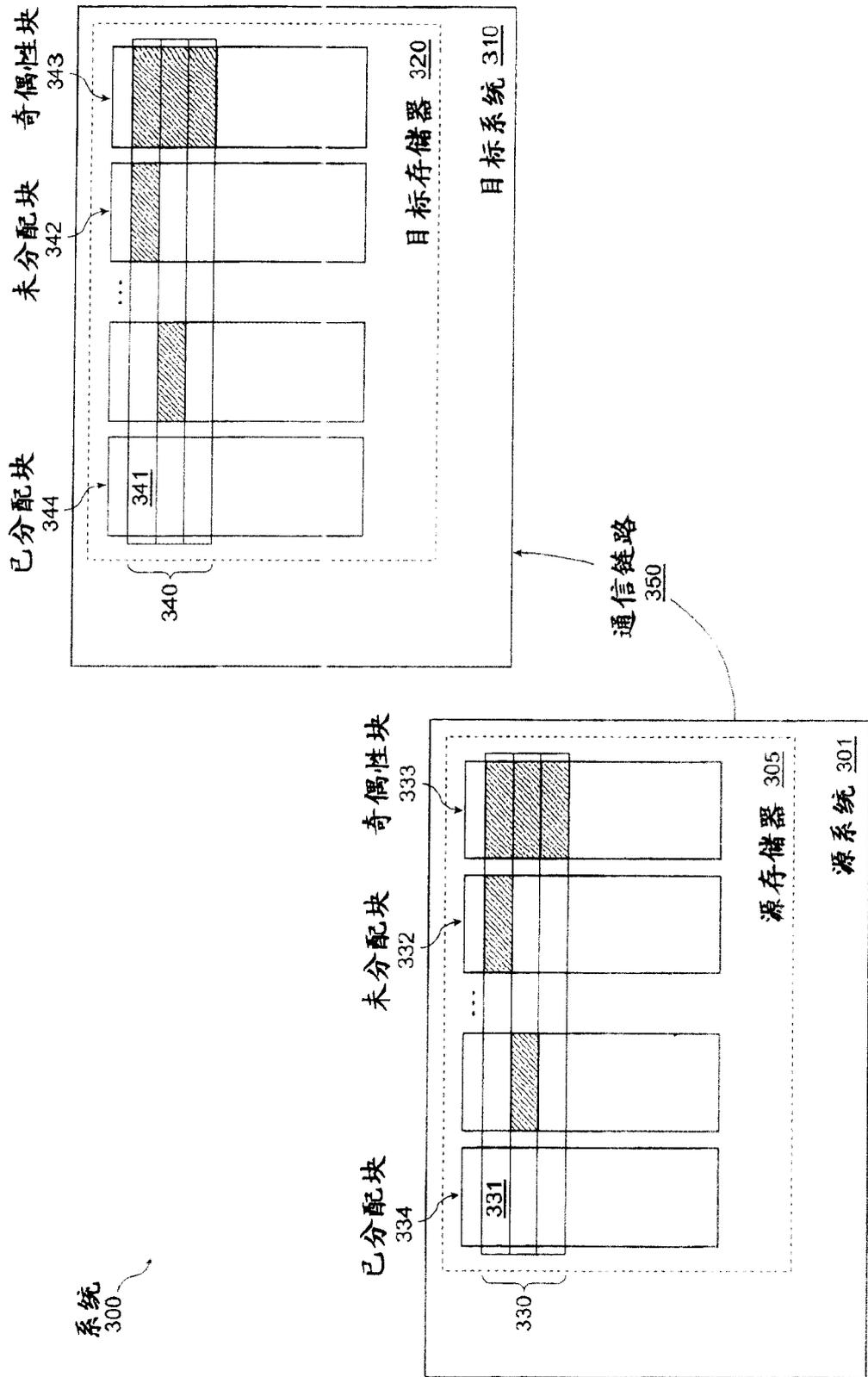


图 3

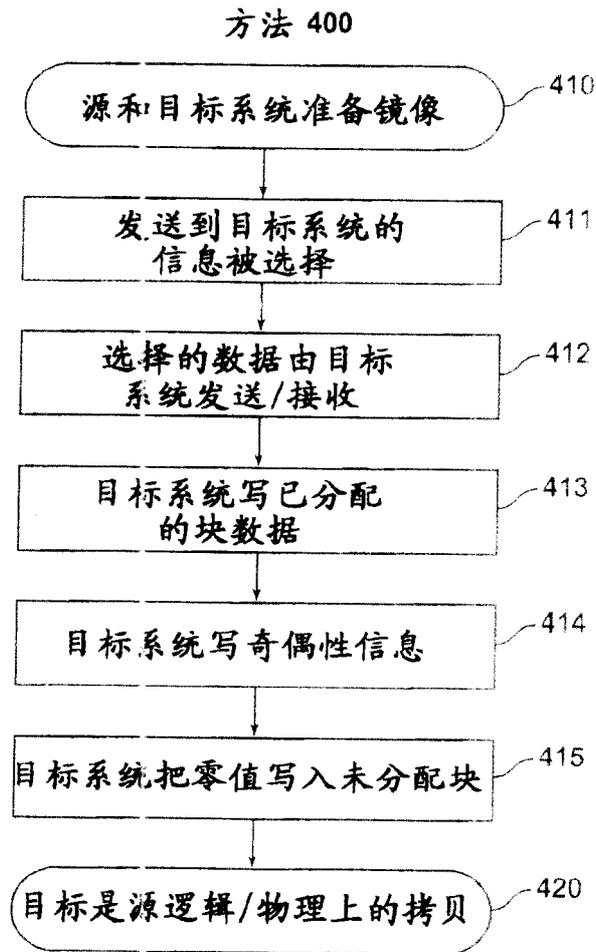


图 4