(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 101414245 B
(45)授权公告日 2011.10.26

(21)申请号 200810108912.5

(22)申请日 2008.06.06

(30)优先权数据
2007-272900 2007.10.19 JP

(73)专利权人 株式会社日立制作所
地址 日本东京都

(72)发明人 武藤淳一 黑川勇 绪方兰
神道一惠

(74)专利代理机构 北京银龙知识产权代理有限公司 11243
代理人 许静

(51)Int.Cl.
G06F 3/06(2006.01)

(56)对比文件
CN 1523504 A,2004.08.25, 全文.
US 6442659 B1,2002.08.27, 全文.
WO 2006/020774 A2,2006.02.23, 全文.
US 5696934 A,1997.12.09, 全文.

审查员 王毅

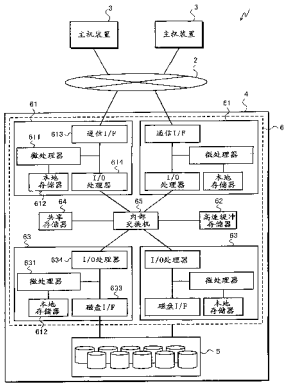
权利要求书 3 页 说明书 15 页 附图 22 页

(54)发明名称

存储装置以及使用该存储装置的数据存储方法

(57)摘要

本发明提供存储装置以及使用该存储装置的数据存储方法。本发明的课题是提供一种可以在预定的RAID结构下高效地存储基于写入数据的压缩数据的存储装置。本发明是具备磁盘设备和对其进行控制的磁盘控制器的存储装置。磁盘控制器通过卷容量虚拟化功能提供由实际卷和虚拟卷构成的数据卷。虚拟卷与用于存储数据实体的池卷相关联,将数据实体存储在池卷中。磁盘控制器依照来自主机装置的写入请求,在RAID5控制下压缩写入数据,将该压缩数据存储在实际卷的存储区域中,并且,在无法完全存储在该存储区域中时,将压缩数据中未存储完的部分存储在虚拟卷中。



CN 101414245 B

1. 一种存储装置，其具备：具有存储数据的存储介质的磁盘设备、控制所述磁盘设备的磁盘控制器，该存储装置的特征在于，

所述磁盘控制器具备：

用于与主机装置连接的通道适配器；

与所述磁盘设备相连的磁盘适配器；以及

对所述通道适配器和所述磁盘适配器之间交换的数据进行暂时存储的高速缓冲存储器，

所述磁盘适配器，

控制所述磁盘设备以便形成数据卷和池卷，所述数据卷包含被定义了比所述存储介质的固有存储容量大的存储容量的、与所述存储介质的存储区域相对应的实际卷，以及分配了所述实际卷以外的存储区域的虚拟卷，所述池卷被分配了用于存储与分配给所述虚拟卷的存储区域相对应的数据的存储区域，

所述磁盘适配器，

在 RAID3、RAID4、RAID5 或 RAID6 结构的控制下，基于依照从所述主机装置发送的写入请求的写入数据，生成该写入数据所属的校验组内的校验数据，

压缩所述校验组内的每一个所述写入数据以及每一个所述生成的校验数据，生成压缩数据以及压缩校验数据，

判断所述校验组内的每一个所述压缩数据以及每一个所述压缩校验数据是否超过规定大小，

将所述压缩数据以及所述压缩校验数据中未超过所述规定大小的部分的数据存储在所述实际卷的存储区域中，将所述压缩数据以及所述压缩校验数据中超过所述规定大小的部分的数据存储在所述虚拟卷所对应的所述池卷的存储区域中。

2. 根据权利要求 1 所述的存储装置，其特征在于，

所述磁盘适配器，

根据构成所述写入数据的数据块以及所述校验组内的其它数据块，生成所述校验数据的数据块，

根据压缩数据列是否由超过规定数量的压缩数据块构成，来判断是否超过了所述规定大小，所述压缩数据列是通过对所述校验组内的每一个所述数据块进行压缩而得到的，

将未超过所述规定数量的所述压缩数据块存储在所述实际卷的存储区域中，将超过所述规定数量的所述压缩数据块存储在所述虚拟卷所对应的所述池卷的存储区域中。

3. 根据权利要求 2 所述的存储装置，其特征在于，

所述磁盘适配器，

将所述校验组内的每一个所述压缩数据块分散存储在构成所述数据卷的、物理上不同系列的磁盘设备中。

4. 根据权利要求 3 所述的存储装置，其特征在于，

所述磁盘适配器，

将构成所述压缩数据列的所述压缩数据块存储在构成所述数据卷的同一系列的磁盘设备中所形成的实际卷以及虚拟卷中。

5. 根据权利要求 2 所述的存储装置，其特征在于，

存储所述校验组内的每一个所述数据块的、在所述高速缓冲存储器上形成的存储区域的大小，大于所述实际卷的单位存储区域的大小。

6. 根据权利要求 2 所述的存储装置，其特征在于，

所述磁盘适配器，

判断除所述校验数据的压缩数据块以外的压缩数据块的数量，是否少于所述校验数据的压缩数据块的数量，

当判断为除所述校验数据的压缩数据块以外的压缩数据块的数量，少于所述校验数据的压缩数据块的数量时，将未超过所述规定数量的所述压缩数据块存储在所述实际卷的存储区域中，将超过所述规定数量的所述压缩数据块存储在所述虚拟卷所对应的所述池卷的存储区域中。

7. 根据权利要求 6 所述的存储装置，其特征在于，

所述磁盘适配器，

当判断为除所述校验数据的压缩数据块以外的压缩数据块的数量，多于所述校验数据的压缩数据块的数量时，将除所述校验数据的压缩数据块以外的每一个压缩数据块二重化，

将所述二重化后的每一个压缩数据块存储在构成所述数据卷的、物理上不同系列的磁盘设备中。

8. 根据权利要求 7 所述的存储装置，其特征在于，

所述磁盘适配器，

分别将所述校验组内的所述二重化后的压缩数据块中的一部分压缩数据块存储在所述数据卷的实际卷中，将另一部分压缩数据块存储在所述数据卷的虚拟卷中。

9. 根据权利要求 1 所述的存储装置，其特征在于，

所述磁盘适配器，

存储针对依照所述写入请求的写入数据的、基于 RAID 结构的数据存储方式，按照该存储的数据存储方式压缩所述写入数据，存储在所述数据卷中。

10. 一种存储装置中的数据存储方法，该存储装置具备：具有存储数据的存储介质的磁盘设备；和控制所述磁盘设备的磁盘控制器，所述数据存储方法的特征在于，具有以下步骤：

所述磁盘控制器控制所述磁盘设备以便形成数据卷和池卷，所述数据卷包含被定义了比所述存储介质的固有存储容量大的存储容量的、与所述存储介质的存储区域相对应的实际卷，以及分配了所述实际卷以外的存储区域的虚拟卷，所述池卷被分配了用于存储与分配给所述虚拟卷的存储区域相对应的数据的存储区域；

所述磁盘控制器具备用于与主机装置连接的通道适配器，所述磁盘控制器按照 RAID3、RAID4、RAID5 或 RAID6 结构，基于依照从所述主机装置发送的写入请求的写入数据，生成该写入数据所属的校验组内的校验数据；

所述磁盘控制器压缩所述校验组内的每一个所述写入数据以及每一个所述生成的校验数据，生成压缩数据以及压缩校验数据；

所述磁盘控制器判断所述校验组内的每一个所述压缩数据以及每一个所述压缩校验数据是否超过规定大小；

所述磁盘控制器，将所述压缩数据以及所述压缩校验数据中未超过所述规定大小的部分的数据存储在所述实际卷的存储区域中，并且将所述压缩数据以及所述压缩校验数据中超过所述规定大小的部分的数据存储在所述虚拟卷所对应的所述池卷的存储区域中。

存储装置以及使用该存储装置的数据存储方法

技术领域

[0001] 本发明涉及存储装置以及使用该存储装置的数据存储方法，尤其涉及在 RAID 控制下压缩写入数据来高效地进行存储的存储技术。

背景技术

[0002] 已知为了有效利用存储装置有限的存储资源而压缩数据来进行存储的技术。压缩数据来进行存储的情况下，压缩后的数据的大小不固定，很大程度上取决于压缩前的数据的内容。

[0003] 下述专利文献 1 公开了一种存储装置，其在磁盘驱动器中压缩数据来进行存储时，将无法完全存储在规定存储区域中的压缩数据的一部分存储在溢出 (over flow) 用的存储区域中。具体而言，在专利文献 1 中，当将来自主处理装置的数据写入磁盘驱动器时，通过数据压缩单元压缩该数据，根据写入目的地地址，判定能否将压缩数据存储在磁盘驱动器内的固定大小存储区域的相应块中，当判定为无法完全存储时，将该压缩数据的一部分存储在相应的块中，同时将该压缩数据的剩余部分存储在溢出存储区域中。

[0004] 另外，近年来，从大容量和高可靠性等观点出发，采用 RAID(RedundantArrays of Independent Disks) 技术的存储装置逐渐占据主流。在 RAID 中，根据硬盘驱动器的结构定义了 RAID 级别 (RAID0 ～ 6)。例如，RAID1 是在多台硬盘驱动器中同时写入相同内容的数据（镜像）的方式。另外，RAID5 是将块单位的数据和纠错码数据（校验数据）一起分散写入到多台硬盘驱动器的方式。

[0005] 【专利文献 1】特开平 5-189157 号公报

[0006] RAID 结构中，在实用中总体性能优秀的 RAID5 的结构是主流。如上所述，在 RAID5 中从耐故障性的观点出发而存在校验数据。因此，在 RAID5 的结构下，在压缩写入数据来进行存储的情况下也压缩其校验数据来进行存储。与写入数据相比，校验数据的数据位的随机性一般较高，无法期待高压缩效率。 因此，很可能由于压缩效率的低下而导致不断发生数据的溢出，在仅仅准备了以往的溢出用存储区域的情况下，有可能立即耗尽。另外，在预先准备足够大小的溢出用存储区域的情况下，运转成本升高，很不经济。

发明内容

[0007] 因此，本发明的目的在于，提出一种可以在预定的 RAID 结构下高效地存储基于写入数据的压缩数据的存储装置。

[0008] 具体而言，本发明提出一种存储装置，其在无法将压缩数据完全存储在规定的存储区域中时，将该压缩数据中未存储完的部分（溢出部分）存储在可以根据使用实际情况而灵活扩展的存储区域（扩展存储区域）中。

[0009] 另外，本发明提出一种存储装置，其当在某 RAID 结构下无法期待压缩数据的高效存储时，通过可以期待高效存储的其它 RAID 结构来存储该压缩数据。即，本发明提出一种根据针对校验数据的压缩效率，来选择存储压缩数据的最佳 RAID 结构的存储装置。

[0010] 为了解决上述问题，本发明的存储装置包含以下技术特征而构成。

[0011] 即，遵从第一观点的本发明是具备：具有存储数据的存储介质的磁盘设备、和控制所述磁盘设备的磁盘控制器的存储装置。所述磁盘控制器具备：用于与主机装置连接的通道适配器；与所述磁盘设备相连的磁盘适配器；以及对所述通道适配器和所述磁盘适配器之间交换的数据进行暂时存储的高速缓冲存储器。另外，所述磁盘适配器控制所述磁盘设备以便形成数据卷和池卷，所述数据卷包含被定义了比所述存储介质的固有存储容量大的存储容量的、与所述存储介质的存储区域相对应的实际卷，以及分配了所述实际卷以外的存储区域的虚拟卷，所述池卷被分配了用于存储与分配给所述虚拟卷的存储区域相对应的数据的存储区域。

[0012] 所述磁盘适配器，在预定的 RAID 结构的控制下，基于依照从所述主机装置发送的写入请求的写入数据，生成该写入数据所属的校验组内的校验数据，压缩所述校验组内的每一个所述写入数据以及每一个所述生成的校验数据，生成压缩数据以及压缩校验数据。另外，所述磁盘适配器判断所述校验组内的每一个所述压缩数据以及每一个所述压缩校验数据是否超过规定大小，将所述压缩数据以及所述压缩校验数据中未超过所述规定大小的部分的数据存储在所 述实际卷的存储区域中，将所述压缩数据以及所述压缩校验数据中超过所述规定大小的部分的数据存储在所述虚拟卷所对应的所述池卷的存储区域中。

[0013] 另外，本发明可以作为方法发明来把握。即，遵从第二观点的本发明，是具备以下各部的存储装置中的数据存储方法：具有存储数据的存储介质的磁盘设备；和控制所述磁盘设备的磁盘控制器。

[0014] 所述数据存储方法包含以下步骤：所述磁盘控制器控制所述磁盘设备以便形成数据卷和池卷，所述数据卷包含被定义了比所述存储介质的固有存储容量大的存储容量的、与所述存储介质的存储区域相对应的实际卷，以及分配了所述实际卷以外的存储区域的虚拟卷，所述池卷被分配了用于存储与分配给所述虚拟卷的存储区域相对应的数据的存储区域；所述磁盘控制器按照预定的 RAID 结构，基于依照从所述主机装置发送的写入请求的写入数据，生成该写入数据所属的校验组内的校验数据；所述磁盘控制器压缩所述校验组内的每一个所述写入数据以及每一个所述生成的校验数据，生成压缩数据以及压缩校验数据；所述磁盘控制器判断所述校验组内的每一个所述压缩数据以及每一个所述压缩校验数据是否超过规定大小；所述磁盘控制器，将所述压缩数据以及所述压缩校验数据中未超过所述规定大小的部分的数据存储在所述实际卷的存储区域中，并且将所述压缩数据以及所述压缩校验数据中超过所述规定大小的部分的数据存储在所述虚拟卷所对应的所述池卷的存储区域中。

[0015] 根据本发明，可以在预定的 RAID 结构下高效地存储基于写入压缩数据的压缩数据。

[0016] 另外，根据本发明，在无法将压缩数据完全存储在规定的存储区域中时，将其存储在扩展存储区域中，因此，可以采用与存储区域的使用实际情况对应的灵活的磁盘设备的结构。

[0017] 而且，根据本发明，根据针对包含校验位的数据的压缩效率来选择最佳 RAID 结构，因此可以进行压缩数据的高效存储。

[0018] 附图说明

[0019] 图 1 是用于说明本发明的一个实施方式的存储系统的结构的图。

[0020] 图 2 是用于从功能上说明本发明的一个实施方式的存储装置中的磁盘控制器的磁盘适配器的图。

[0021] 图 3 是表示本发明的一个实施方式的存储装置中的共享存储器的内容的图。

[0022] 图 4 是表示本发明的一个实施方式的存储装置中的区段管理表的一例的图。

[0023] 图 5 是表示本发明的一个实施方式的存储装置的共享存储器中存储的地址管理表的一例的图。

[0024] 图 6 是用于说明本发明的一个实施方式的存储装置的共享存储器中存储的数据卷的图。

[0025] 图 7 是用于从概念上说明本发明的一个实施方式的存储装置中的存储区域的动态分配的图。

[0026] 图 8 是表示本发明的一个实施方式的存储装置中的共享存储器中存储的数据块管理表的一例的图。

[0027] 图 9 是表示本发明的一个实施方式的存储装置中的共享存储器中存储的磁盘设备管理表的一例的图。

[0028] 图 10 是用于说明本发明的一个实施方式的存储装置中的数据写入机制的图。

[0029] 图 11 是用于说明本发明的一个实施方式的存储装置中的数据写入机制的图。

[0030] 图 12 是用于说明本发明的一个实施方式的存储装置中的数据写入处理的流程图。

[0031] 图 13 是用于说明针对本发明的一个实施方式的存储装置的虚拟卷的数据写入处理的流程图。

[0032] 图 14 是用于说明本发明的一个实施方式的存储装置中的数据写入机制的图。

[0033] 图 15 是用于说明本发明的一个实施方式的存储装置中的数据写入机制的图。

[0034] 图 16 是表示本发明的一个实施方式的存储装置中的共享存储器中存储的磁盘设备管理表的一例的图。

[0035] 图 17 是用于说明本发明的一个实施方式的存储装置中的数据写入处理的 流程图。

[0036] 图 18 是用于说明本发明的一个实施方式的存储装置中的数据写入机制的图。

[0037] 图 19 是用于说明本发明的一个实施方式的存储装置中的数据写入机制的图。

[0038] 图 20 是表示本发明的一个实施方式的存储装置中的共享存储器中存储的磁盘设备管理表的一例的图。

[0039] 图 21 是用于说明本发明的一个实施方式的存储装置中的数据写入处理的流程图。

[0040] 图 22 是用于说明本发明的一个实施方式的存储装置中的数据写入处理的流程图。

[0041] 图 23 是用于说明本发明的一个实施方式的存储装置中的数据写入处理的流程图。

[0042] 图 24 是表示本发明的一个实施方式的存储装置中的共享存储器中存储的磁盘设备管理表的一例的图。

[0043] 图 25 是用于说明本发明的一个实施方式的存储装置中的数据写入处理的流程图。

[0044] 符号说明

[0045] 1 计算机系统、2 网络、3 主机装置、4 存储装置、5 磁盘设备、6 磁盘控制器、61 通道适配器、62 高速缓冲存储器、63 磁盘适配器、64 共享存储器、65 内部交换机

具体实施方式

[0046] 接下来，参照附图说明本发明的实施方式。

[0047] （第一实施方式）

[0048] 在第一实施方式中说明如下构成的存储装置：在 RAID5 的结构下压缩写入数据，当无法将该压缩后的数据完全存储在数据卷上的实际存储区域中时，将该未存储完的部分存储在通过卷容量虚拟化功能提供的扩展存储区域中。

[0049] 图 1 是用于说明本发明的一个实施方式的存储系统的结构的图。如该图所示，计算机系统 1 包含经由网络 2 与主机装置 3 相连的存储装置 4。

[0050] 网络 2 是例如 LAN、因特网或 SAN(Storage Area Network)，包含网络交换机和集线器等而构成。在本实施方式中，网络 2 由使用光纤通道协议的 SAN(FC-SAN) 构成。

[0051] 主机装置 3 例如是，作为银行的业务系统或飞机的座席预约业务系统等的核心的计算机。主机装置 3 具备处理器、主存储器、通信接口和本地输入输出装置等硬件资源，另外，具备设备驱动程序和操作系统 (OS)、应用程序等软件资源（未图示）。由此，主机装置 3 在处理器的控制下执行各种程序，通过与硬件资源的协同工作来实现希望的处理。例如，主机装置 3 通过在处理器的控制下执行业务应用程序，访问以下详细描述的存储装置 4，实现希望的业务系统。

[0052] 存储装置 4 是向主机装置 3 提供数据存储服务的辅助存储装置。存储装置 4 具备：存储数据的磁盘设备 5；和管理磁盘设备 5 的结构、控制对磁盘设备 5 的写入或读出的 I/O 处理的磁盘控制器 6。存储装置 4 还可以包含用于管理存储装置 4 全体的管理装置乃至服务处理器（未图示）。或者也可以构成包含管理装置的功能的磁盘控制器 6。管理装置例如是安装了管理程序的通用计算机。

[0053] 本实施方式的存储装置 4 如后所述，在 RAID 控制下压缩写入数据，将该压缩数据存储在磁盘设备 5 的规定的存储区域。当压缩数据超过规定的块大小时，存储装置 4 将该超过的部分存储在扩展存储区域（池卷）中。如后所述，池卷 (Pool Volume) 是与通过卷容量虚拟化功能提供的虚拟卷相对应的、存储数据实体的逻辑卷。

[0054] 磁盘设备 5 例如是包含多个硬盘驱动器 (HDD) 或半导体存储器等存储介质而构成的物理设备 (PDEV)。有时也将多个硬盘的集合体称为磁盘阵列 (DiskArray)。在磁盘设备 5 中，在磁盘控制器 6 的控制下形成用于向主机装置 3 提供的一个以上的逻辑设备 (LDEV)。可以对应于将若干硬盘驱动器虚拟地汇总为一个设备而得到的虚拟设备 VDEV，来形成逻辑设备。

[0055] 逻辑设备是主机装置 3 能够识别的逻辑存储装置。可以对逻辑设备分配逻辑单元 (LU)。在这种情况下，将各逻辑设备分配给后述的通道适配器 61 中设置的各端口，由此，主机装置 3 将逻辑设备识别为逻辑单元。对各逻辑单元 赋予逻辑单元号码 (LUN)。另外，将

逻辑单元分割为 I/O 访问的最小单位、即块，对各块分配逻辑块地址 (LBA)。由此，主机装置 3 通过将由 LUN 和 LBA 构成的逻辑地址赋予存储装置 4，可以对特定逻辑单元中的任意块中存储的数据进行访问。逻辑设备以及逻辑单元可以通过应用存储装置 4 的系统环境而区别，但在本发明中可以将二者取相同含义。

[0056] 在逻辑设备中定义与属性对应的逻辑卷 (LVOL 或 VOL)。但习惯上有时也将二者取相同含义。在本实施方式中，作为逻辑卷而定义数据卷和池卷。如后所述，数据卷包含实际卷以及虚拟卷。

[0057] 磁盘控制器 6 具备 :通道适配器 (CHA)61、高速缓冲存储器 (CM)62、磁盘适配器 (DKA)63 以及共享存储器 (SM)64，这些部件 (component) 经由内部交换器 (SW)65 相互连接。从高速性、耐故障性的观点出发，理想的是将这些部件冗余化。

[0058] 磁盘控制器 6 是控制磁盘设备 5，实现数据存储服务的存储装置 4 的主系统板 (Main System Board)。本实施方式的磁盘控制器 6 包含实现除了基于来自主机装置 3 的访问请求的通常的 I/O 处理功能以外，还实现数据压缩 / 解压功能、RAID 功能以及卷容量虚拟化功能的结构。

[0059] 通道适配器 61 是用于经由网络 2 将存储装置 4 与主机装置 3 可通信地连接的部件。在本例中搭载了两个通道适配器 61。各通道适配器 61 具备 :微处理器 611、本地存储器 (local memory)612、通信接口 (I/F)613 和 I/O 处理器 614。

[0060] 微处理器 611 是通过执行存储在本地存储器 612 中的微程序，总体地控制通道适配器 61 的动作的芯片电路。通信 I/F613 从网络 2 取得通信数据包 (packet) 乃至帧 (frame)，或者向网络 2 送出通信数据包。通信 I/F613 具备用于与网络 2 连接的多个端口 (未图示)。I/O 处理器 614 控制与经由内部交换机 65 连接的磁盘控制器 6 内的其它部件 (例如高速缓冲存储器 62 或共享存储器 64) 之间的数据收发。

[0061] 具体而言，微处理器 611，当经由通信 I/F613 接收到数据包时，进行与规定的协议对应的协议变换处理，取出内部数据。接着，微处理器 611 解释该内部数据并控制 I/O 处理器 614，若该内部数据是写入命令，则将该写入命令写 入共享存储器 64，将其写入数据 (即用户数据) 写入到高速缓冲存储器 62 中。I/O 处理器 614 接受该控制，将写入命令写入共享存储器 64，并且将其写入数据写入到高速缓冲存储器 62 中。

[0062] 另外，微处理器 611 按照读出命令，通过 I/O 处理器 614 从高速缓冲存储器 62 取出读出数据，进行与规定的协议对应的协议变换处理来生成数据包，并且经由通信 I/F613 发送至网络 2。

[0063] 高速缓冲存储器 62 为了对主机装置 3 提供高系统性能，对主机装置 3 和磁盘设备 5(逻辑卷) 之间交换的用户数据进行暂时存储 (高速缓冲存储)。即，高速缓冲存储器 62 是用于通道适配器 61 和磁盘适配器 63 之间的数据转发的存储器。高速缓冲存储器 62 例如由 DRAM 等易失性存储器构成。或者，高速缓冲存储器 62 也可以由闪速存储器等非易失性存储器构成。

[0064] 高速缓冲存储器 62 如后所述，还用作对压缩后的写入数据 (压缩数据) 进行暂时存储的存储器。例如通过磁盘适配器 63 来进行数据压缩 / 解压处理。

[0065] 磁盘适配器 63 是对经由磁盘通道连接的磁盘设备 5 进行 I/O 访问的部件。在本例中搭载了两个磁盘适配器 63。各磁盘适配器 63 具备 :微处理器 631、本地存储器 632、磁盘

接口 (I/F)633 和 I/O 处理器 634。在本实施方式中，磁盘适配器 63 安装了 I/O 控制功能、数据压缩 / 解压功能、RAID 功能以及卷容量虚拟化功能。这些功能例如作为图 2 所示的固件 (firmware) 而实现。

[0066] 微处理器 631 是通过执行存储在本地存储器 632 中的微程序，总体控制磁盘适配器 63 的动作的芯片电路。在本实施方式中，微处理器 631 在 RAID 控制下对高速缓冲存储器 62 中存储的写入数据进行按照规定的压缩算法的压缩处理来生成压缩数据，将其写入磁盘设备 5 的规定的存储区域。这里所说的规定的存储区域是指实际卷上的存储区域（实际存储区域）。将压缩数据中无法完全存储在该实际存储区域中的部分的数据，存储在通过卷容量虚拟化功能而管理的扩展存储区域（池卷）中。压缩 / 解压算法，例如可以应用 LZW 方式等已知算法。

[0067] 磁盘 I/F633 是用于对作为物理设备的磁盘设备 5 进行 I/O 访问的接口。I/O 处理器 634 控制与经由内部交换机 65 连接的其它部件（例如高速缓冲存储器 62 或共享存储器 64）之间的数据收发。

[0068] 更具体而言，微处理器 631 定期地（例如数十毫秒）或不定期地通过 I/O 处理器 634 参照共享存储器 64。微处理器 631 当发现共享存储器 64 上的未处理的写入命令时，从高速缓冲存储器 62 取出对应的写入数据，在 RAID 控制下进行压缩处理，生成压缩数据。即，生成基于写入数据的校验数据，生成与包含该校验数据的写入数据对应的压缩数据。将压缩数据再次存储在高速缓冲存储器 62 上的压缩数据用块中。然后，微处理器 631 在 RAID 控制下将生成的压缩数据存储在形成了数据卷的磁盘设备 5 上的规定的存储区域（块）中（降级）。例如，若在 RAID5(3D+1P) 的结构中，将压缩数据与其校验数据一起分散存储在物理上不同系列的磁盘设备 5 中。另外，此时，微处理器 631 判断能否将生成的压缩数据完全存储在数据卷中规定的实际存储区域中，当判断为无法完全存储时，将该压缩数据中未存储完的部分的数据存储在通过卷容量虚拟化功能而管理的扩展存储区域中。

[0069] 另外，微处理器 631 按照共享存储器 64 上的读出命令，在 RAID 控制下从规定的逻辑卷读出压缩数据，将其暂时存储在高速缓冲存储器 62 中。然后，微处理器 631 进行解压处理，将所得到的解压数据（原始数据）再次写入高速缓冲存储器 62（升级）。当将压缩数据的一部分存储在扩展存储区域中时，微处理器 631 适宜从扩展存储区域读出该部分压缩数据，合并成一个压缩数据。

[0070] 共享存储器 64 对存储装置 4 内的各部件应参照的各种信息进行存储。共享存储器 64 由例如 DRAM 等易失性存储器构成。

[0071] 内部交换机 65 是由交叉开关 (Crossbar Switch) 等构成的交换设备。内部交换机 65 协调所输入的数据信号的冲突，切换数据信号的路径，构建发送源模块和发送目的地模块的路径。此外，内部交换机 65 也可以是数据包交换方式的交换设备。

[0072] 图 3 是表示本发明的一个实施方式的存储装置 4 中的共享存储器 64 的内容的图。如该图所示，共享存储器 64 存储例如系统结构信息、高速缓冲存储管理表、数据块管理表、物理设备管理表等各种信息。

[0073] 系统结构信息是磁盘设备 5 的结构信息或存储装置 4 内的各部件上运行的微程序的版本信息等、与系统结构相关的信息。作为磁盘设备 5 的结构信息，包含例如 RAID 结构信息、物理设备 / 逻辑设备（卷）的结构信息、卷容量虚 拟化结构信息。物理设备 / 逻辑

设备结构信息，例如被定义为区段（segment）管理表以及地址管理表。另外，卷容量虚拟化结构信息例如包含池卷管理表。

[0074] 通过系统管理者操作的管理装置，来设定·管理系统结构信息。例如，系统管理者可以操作管理装置来设定 RAID5(3D+1P) 的结构。另外，系统管理者，与硬盘驱动器的增设相配合地操作管理装置，来设定卷容量虚拟化结构信息。

[0075] 高速缓冲存储管理表是管理在高速缓冲存储器 62 中高速缓冲存储的数据的目录信息的表。高速缓冲存储管理表管理如下内容：构成写入数据的每一个数据块对应于磁盘设备 5（逻辑卷）上的哪个块。

[0076] 如上所述，在本实施方式中，在逻辑卷（数据卷）中不存储写入数据本身，而是存储其压缩数据。因此，典型地考虑平均压缩率，将数据卷上规定的存储区域的块大小规定得小于在高速缓冲存储器 62 上规定的块大小。另外，磁盘适配器 63，将从高速缓冲存储器 62 的规定的块中读出并压缩的数据，再次存储在高速缓冲存储器 62 的其它块（压缩数据用块）中，然后进行降级。因此，高速缓冲存储管理表，将高速缓冲存储器 62 上的 RAW 状态的写入数据以及压缩数据的各逻辑地址，与数据卷上的逻辑地址对应起来。

[0077] 数据块管理表是用于管理通过 RAID 控制的每个校验组的数据块的表。所谓校验组 (parity group)，在本说明书中是指同时并行地被写入阵列磁盘的数据块组。例如，若是 RAID5(3D+1P)，则由 3 个数据块和一个校验块构成的数据块组属于同一校验组。

[0078] 磁盘设备管理表，是用于管理在 RAID 控制下形成数据卷的、具有条带 (stripe) 结构的（即作为物理上不同系列而构成的）“磁盘设备”的地址空间的使用状况的表。准备条带结构数量的磁盘设备管理表。例如，若是 RAID5(3D+1P)，则准备 4 个磁盘设备管理表。此外，如按照惯用的 RAID 的说明，则可以认为条带构成的各个磁盘设备等同于各个物理设备。然而，在本实施方式中，以包含通过卷容量虚拟化功能而不具有物理存储区域的虚拟卷的概念，说明条带构成的“磁盘设备”。因此，在本说明书中，将在 RAID 控制下条带构成的各个磁盘设备称为“伪磁盘设备”。若是 RAID5(3D+1P)，则在 4 个伪磁盘设备上形成某个数据卷。

[0079] 图 4 表示本发明的一个实施方式的存储装置 4 中的区段管理表的一例。区段管理表，是用于管理存储装置 4 所提供的逻辑卷上的存储区域空间的表。如该图所示，区段管理表包含：磁盘 ID(Disk ID)、区段号码 (Segment#)、LBA 开始地址 (LBA_start)、LBA 结束地址 (LBA_end)、区段的使用状况 (In_use)。

[0080] 磁盘 ID 是用于唯一识别构成磁盘设备 5 的物理设备的识别符。区段号码是用于唯一识别区段的号码。区段是磁盘控制器 6 管理的磁盘设备 5 上的存储区域。LBA 开始地址以及 LBA 结束地址是表示磁盘设备 5 上的区段的物理的开始位置以及结束位置的地址。根据 LBA 开始地址以及 LBA 结束地址求得区段的大小。使用状况表示区段是否已被使用。在区段处于使用中的情况下设定“1”的值，在未使用的情况下设定“0”的值。

[0081] 图 5 表示本发明的一个实施方式的存储装置 4 的共享存储器 64 中存储的地址管理表的一例。地址管理表是将存储装置 4 提供的逻辑卷上的存储区域（即逻辑地址空间）和磁盘设备 5 上的实际的存储区域（即物理地址空间）对应起来的表。磁盘适配器 63 通过参照地址管理表，将作为数据的存储目的地而指定的逻辑地址变换为磁盘设备 5 上的物理地址，按照该物理地址来进行访问。另外，地址管理表对逻辑卷中作为池卷而被使用的区

段进行定义。

[0082] 即，如该图所示，地址管理表包含逻辑单元号码 (LUN)、区段号码 (Segment#)、LBA 开始地址 (LBA_start)、LBA 结束地址 (LBA_end) 以及池分配状况。逻辑单元号码是用于唯一识别存储装置 4 向主机装置 3 提供的逻辑卷的号码。区段号码是用于唯一识别区段管理表中所管理的存储区域的号码。逻辑卷 LU 由多个区段构成。LBA 开始地址以及 LBA 结束地址表示区段的逻辑的开始地址和结束地址。池分配状况表示该区段是否作为池卷被使用。当区段作为池卷被使用时，写入“1”。

[0083] 图 6 是用于说明本发明的一个实施方式的存储装置 4 的共享存储器 64 中存储的数据卷的图。如该图所示，数据卷 DVOL0 具有 0 ～ 99999 的逻辑地址空间，分配了第一磁盘设备以及第二磁盘设备中的各个区段。第一磁盘设备 5 如图 4 所示那样，是分配了磁盘 ID “0”的物理设备，第二磁盘设备 5 是分配了磁盘 ID “1”的物理设备。另外，如图 5 所示那样，区段号码“1”所表示的区段提供数据卷 DVOL0 上的 0 ～ 1299 的逻辑地址空间。另外，区段号码“3” 所表示的区段提供 1300 ～ 1999 的逻辑地址空间。于是，数据卷 DVOL0 像一个磁盘设备那样被处理。

[0084] 数据卷 DVOL0 的逻辑地址空间 2000-99999，未被分配作为物理设备的磁盘设备。磁盘控制器 6 表面上也对数据卷 DVOL0 的该空间存储数据。为此，存储装置 4 还具有被称为池卷 PVOL 的逻辑卷，将应该对数据卷 DVOL0 内的存储区域中未被分配磁盘设备 5 上的区段的存储区域（扩展存储区域）存储的数据，存储在该池卷 PVOL 上。即，在数据卷 DVOL0 的逻辑地址空间 2000 ～ 49999 中存储的数据实体，被存储在池卷 PVOL 中。因此，在内部按照池卷管理表对池卷 PVOL 来进行针对数据卷的扩展存储区域的访问。应该存储在扩展存储区域中的数据，是无法完全存储在通过物理设备而实际提供的存储区域（实际存储区域）中的压缩后的数据块。

[0085] 该图简略地说明了应用了卷容量虚拟化功能的数据卷 DVOL。在本实施方式中，磁盘设备 5 根据 RAID 结构而条带构成。因此，实际通过物理上不同系列的 4 个伪磁盘设备来提供数据卷 DVOL。

[0086] 图 7 是用于在概念上说明本发明的一个实施方式的存储装置 4 中的存储区域的动态分配的图。

[0087] 如上所述，在本实施方式的存储装置 4 的磁盘设备 5 中形成数据卷 DVOL 和池卷 PVOL。数据卷 DVOL 是提供给主机装置 3 的逻辑卷。数据卷 DVOL 依存于构成实际安装的磁盘设备 5 的物理设备的固有存储容量，分为实际卷和虚拟卷。即，实际卷包含在由构成磁盘设备 5 的物理设备实际提供的区段上形成的物理存储区域。与之相对，虚拟卷对应于实际提供的区段以外的虚拟的存储区域（扩展存储区域）。因此，在虚拟卷中不存储数据实体。池卷 PVOL 是提供在将来追加安装磁盘设备、确保物理存储区域并且定义新数据卷 DVOL 之前的期间，对应该存储在虚拟卷中的数据实体进行暂时存储的存储区域的逻辑卷。

[0088] 针对数据卷 VOL 中的实际卷上的存储区域，存储装置 4 将主机装置 3 识别的逻辑地址 (LUN 和 LBA) 与磁盘设备 5 中的存储区域上的物理地址一对一地关联起来。由此，主机装置 3 可以通过指定规定的逻辑地址来访问磁盘设备 5 上所希望的存储区域。

[0089] 另一方面，关于池卷 PVOL 上的存储区域，不将主机装置 3 识别的逻辑地址与实际用于访问数据的磁盘设备 5 上的物理地址直接关联起来。磁盘适配器 63 按照指定了数据

卷 DVOL 中的虚拟卷上的存储区域的数据写入命令，将该存储区域动态地分配给池卷 PVOL。另外，磁盘适配器 63 按照从数据卷 DVOL 中的虚拟卷上的存储区域的读出命令，从池卷 PVOL 上的该存储区域读出数据。

[0090] 这样，存储装置 4 保持池卷管理表，管理虚拟卷和池卷之间的动态的分配。池卷管理表例如被保持在磁盘控制器 6 的共享存储器 64 中。存储装置 4 当无法将压缩数据完全存储在实际卷中时，将该未存储完的部分存储在池卷 PVOL 中，同时将虚拟卷上的逻辑地址和表示池卷地址的指针关联起来，登录在池卷管理表中。

[0091] 图 8 是表示本发明的一个实施方式的存储装置 4 中的共享存储器 64 中存储的数据块管理表的一例的图。如该图所示，数据块管理表将槽 ID(slot ID) 和数据卷内的地址对应起来。槽 ID 是用于识别校验组内的数据块的识别符。由此，数据块管理表管理如下内容：校验组内的数据块存储在哪个物理设备的哪个地址。

[0092] 图 9 是表示本发明的一个实施方式的存储装置 4 中的共享存储器 64 中存储的磁盘设备管理表的一例的图。如该图所示，磁盘设备管理表将各伪磁盘设备内的各个地址所表示的存储区域、与其中存储的数据块（槽识别符）对应起来。另外，磁盘设备管理表管理着扩展存储区域中的使用状况。在本例中表示了与伪磁盘设备 (0) 对应的磁盘设备管理表。即，该图的磁盘设备管理表表示在以地址“xxxx”表示的存储区域中存储“D1-1”、在扩展存储区域 1 中存储“D1-2”的情况。另外，还表示可以使用扩展存储区域 2。而且，还表示扩展存储区域 n 尚未被分配实际的存储区域。

[0093] 图 10 和图 11 是用于说明本发明的一个实施方式的存储装置 4 中的数据写入机制 (mechanism) 的图。具体而言，这些图表示某个写入数据在被压缩后，在 RAID5(3D+1P) 下被存储在不同系列的伪磁盘设备中的情况。

[0094] 针对数据卷的写入数据，在存储装置 4 的内部作为被分块化的数据（数据块）来处理。单位块的大小例如是 64KB。在本实施方式中，将写入数据压缩 后存储，因此在数据卷上规定小于该大小的（例如 16KB) 的单位存储区域。高速缓冲存储器 62 上的 64KB 的各块，与数据卷上的 16KB 的各块相对应。因此，在将 64KB 的数据压缩时，若其压缩数据未超过 16KB，则可以容纳在与高速缓冲存储器 62 上的 64KB 块对应的数据卷上的 16KB 块（基本块）中，但是若其压缩数据超过了 16KB，则需要该 16KB 块以外的备用块。即，当 64KB×1 数据块的压缩数据由 16KB×3 数据块构成时，将超出的 2 数据块存储在备用块中。

[0095] 现在，假设存在对某个数据卷的数据 D1 的写入请求。为了简化说明，假设写入数据 D1 是块大小为 64KB 的单一的数据块。为了在 RAID 控制下将写入数据 D1 存储在数据卷中，需要与相同校验组内的数据块 D2 和 D3 一起计算校验位 (parity) 并将它们存储（图 10(A))。

[0096] 数据块 D1 ～ D3 在磁盘适配器 63 的控制下被压缩，得到图 10(B) 所示的压缩数据块。压缩数据块的单位大小为 16KB。即，压缩前的数据块 D1 成为压缩数据块 D1-1 和 D1-2；压缩前的数据块 D2 成为压缩数据块 D2-1 和 D2-2，并且压缩前的数据块 D3 成为压缩数据块 D3-1。另外，将这些压缩前数据块 D1 ～ D3 的校验数据 P 变换为压缩数据块 P1-1 ～ P1-4。即，压缩数据块 D1-2、D2-2 以及 P2-2 ～ P2-4 是无法完全存储在数据卷上的基本块中的数据块。

[0097] 在本实施方式中，高速缓冲存储器 62 上的 64KB 块，与数据卷上的 16KB 基本块相

关联。基本块是数据卷的实际卷中的存储区域（实际存储区域）。另外，根据数据的内容，有时无法将其压缩数据完全存储在实际存储区域中，因此，将该无法完全存储的部分的数据块存储在虚拟卷的块（扩展存储区域）中。

[0098] 即，如图 11 所示，压缩数据块 D1-1、D2-1、D3-1 以及 P-1，在 RAID 控制下分别被存储在构成数据卷的 RAID 组中的伪磁盘设备 (0) ～伪磁盘设备 (3) 上的实际卷中的块（实际存储区域）中。另外，压缩数据块 D1-2、D2-2 以及 P-2 ～ P-4，分别被存储在该 RAID 组中的对应的伪磁盘设备 (0) ～伪磁盘设备 (4) 上的虚拟卷中的块（扩展存储区域）中。但是，如上所述，虚拟卷未被分配存储数据实体的物理设备的区段。因此，虚拟卷中存储的数据实体被存储在池卷中。

[0099] 图 12 是用于说明本发明的一个实施方式的存储装置 4 中的数据写入处理 的流程图。具体而言，该图说明了数据写入处理时的磁盘适配器 63 的动作。

[0100] 即，磁盘适配器 63 参照共享存储器 64，当发现未处理的写入命令时，确定高速缓冲存储器 62 上的与之对应的写入数据。磁盘适配器 63 根据该写入数据生成遵从 RAID 结构的校验数据 (STEP1201)。即，在本例中，磁盘适配器 63 遵从 RAID5(3D+1P) 的结构，针对每 3 个数据块生成 1 个校验数据块。通过这 3 个数据块和 1 个校验数据块形成 1 个校验组。因此，即使是 1 个数据块的写入，也需要与校验组内的其余两个数据块一起生成新的校验数据块。一般将其称为“写惩罚”(write penalty)。在该阶段，各数据块为 64KB 的块大小。

[0101] 接着，磁盘适配器 63 对该校验组中的各数据块（包含校验数据块）进行压缩处理，分别生成压缩数据列 (STEP1202)。压缩数据列由 1 个以上的数据块（压缩数据块）构成。磁盘适配器 63 将各个压缩所得的数据块写入高速缓冲存储器 62 中的压缩数据用块中，更新高速缓冲存储管理表。压缩数据用块例如是 16KB 的块大小。

[0102] 磁盘适配器 63 接着为了将压缩数据块分散存储在不同系列的伪磁盘设备中，而进行以下处理。

[0103] 具体而言，磁盘适配器 63，首先将校验周期 (parity cycle) 变量 I 设定为 0(STEP1203)。接着，磁盘适配器 63 在数据块管理表中登录该压缩数据块的条目 (STEP1204)。即，磁盘适配器 63，对于构成应该存储在伪磁盘设备 (I) 中的压缩数据列的先头的压缩数据块，分配伪磁盘设备 (I) 中的实际存储区域的地址。即，当压缩数据列由压缩数据块 D1-1 以及 D1-2 构成时，对该压缩数据块 D1-1 首先分配伪磁盘设备 (I) 中的实际存储区域的地址。接着，磁盘适配器 63 按照该地址在该实际存储区域中存储压缩数据块 (STEP1205)。

[0104] 然后，磁盘适配器 63 判断该压缩数据列是否包含超过部分的压缩数据块 (STEP1206)。即，判断压缩数据列是否由两个以上的压缩数据块构成。磁盘适配器 63，当判断为没有超过部分的压缩数据块时 (STEP1206，否)，将校验周期变量 I 的值增加 1(STEP1209)，判断校验周期变量 I 是否在规定的校验周期数（在本例中为 4) 以下 (STEP1210)。当校验周期变量 I 在规定的校验周期数以下时 (STEP1210，否)，由于校验组内有未处理的压缩数据块，因此 返回 STEP1204 的处理。

[0105] 磁盘适配器 63，当判断为存在超过部分的压缩数据块时 (STEP1206，是)，在数据块管理表中登录该超过部分的压缩数据块的条目 (STEP1207)。即，磁盘适配器 63 对该超过

部分压缩数据块分配伪磁盘设备（I）中的扩展存储区域的地址。

[0106] 接着，磁盘适配器 63 进行用于将超过部分的压缩数据块存储在虚拟卷中的处理(STEP1208)。在后面详细描述针对虚拟卷的存储处理。

[0107] 然后，磁盘适配器 63 将校验周期变量 I 的值增加 1(STEP1209)，判断校验周期变量 I 是否在规定的校验周期数（在本例中为 4）以下 (STEP1210)。当校验周期变量 I 在规定的校验周期数以下时 (STEP1201，否)，由于校验组内有未处理的压缩数据块，因此返回 STEP1204 的处理。

[0108] 当校验周期变量 I 达到规定的校验周期数时 (STEP1210，是)，磁盘适配器 63 结束该数据写入处理。

[0109] 图 13 是用于说明针对本发明的一个实施方式的存储装置 4 中的虚拟卷的数据写入处理的流程图。具体而言，该图是用于说明图 12 所示的 STEP1208 的详细情况的流程图。

[0110] 即，如该图所示，磁盘适配器 63 首先判断是否已经对虚拟卷分配了池卷上的存储区域 (STEP1301)。磁盘适配器 63，当判断为未对虚拟卷分配池卷上的存储区域时(STEP1301，否)，在池卷管理表中追加关于超过部分的压缩数据块的条目 (STEP1302)。若以上述例子而言，则关于压缩数据块 D1-2，将虚拟卷的地址和池卷的地址对应起来。

[0111] 接着，磁盘适配器 63 在磁盘设备管理表的扩展存储区域中追加数据块(STEP1303)。然后，磁盘适配器 63 将超过部分的压缩数据块存储在虚拟卷中 (STEP1304)。在这种情况下，实际上在磁盘适配器 63 的控制下，将该超过部分的压缩数据块存储在提供池卷的磁盘设备 5 中。

[0112] 如上所述，根据本实施方式，在采用为了压缩存储写入数据而将高速缓冲存储器 62 上的块固定地分配给磁盘设备 5 上的更小的块的结构的存储装置中，即使在无法将压缩数据完全存储在磁盘设备 5 上的该块中时，也可以将该压缩数据中未存储完的部分的数据存储在扩展存储区域中。该扩展存储区域，作为 与之相关联的池卷而被管理，因此可以采用与存储区域的使用实际相对应的灵活的磁盘设备的结构。

[0113] （第二实施方式）

[0114] 在第二实施方式中，说明在 RAID 结构下对写入数据的压缩数据进行镜像(mirroring) 来存储时，将复制的压缩数据存储在虚拟卷中而构成的存储装置 4。

[0115] 图 14 和图 15 是用于说明本发明的一个实施方式的存储装置 4 中的数据写入机制的图。具体而言，这些图说明了在将某写入数据 D4 ～ D6 压缩后，在 RAID1(4D) 下存储在伪磁盘设备中的情况。

[0116] 考虑对某数据卷存储写入数据 D4 ～ D6 的情况。为了简化说明，设写入数据 D4 ～ D6 是块大小为 64KB 的单一的数据块。数据块 D4 ～ D6 在磁盘适配器 63 的控制下被压缩，得到压缩数据块 D4-1 ～ D6-1。另外，以压缩数据块 D4-1’ ～ D6-1’ 表示压缩数据块 D4-1 ～ D6-1 的复制数据（镜像数据）。压缩数据块的大小是 16KB。因此，存储装置 4 对于 64KB×3 数据块存储 16KB×6 数据块（图 14）。

[0117] 在本实施方式中如图 15 所示，将这些压缩数据块分散存储在每个伪磁盘设备的实际存储区域中，同时将镜像后的压缩数据块分散存储在扩展存储区域中。但是，如本例所示，当数据周期 (data cycle) 是 3、伪磁盘设备是 4 个时，也可以将一个镜像后的压缩数据块存储在第四伪磁盘设备的实际存储区域中。

[0118] 图 16 是表示在本发明的一个实施方式的存储装置 4 中的共享存储器 64 中存储的磁盘设备管理表的一例。如该图所示，本实施方式的磁盘设备管理表，针对每个槽 ID，将第一地址和第二地址对应起来。第二地址是存储镜像后的压缩数据块的数据卷的地址。

[0119] 图 17 是用于说明本发明的一个实施方式的存储装置 4 中的数据写入处理的流程图。具体而言，该图说明了数据写入处理时的磁盘适配器 63 的动作。此外，在本例中，RAID 组是 4D。

[0120] 即，如该图所示，磁盘适配器 63 参照共享存储器 64，当发现未处理的写入命令时，确定高速缓冲存储器 62 上与之对应的写入数据。磁盘适配器 63 遵从 RAID1，对构成该写入数据的数据块进行压缩处理，生成压缩数据块 (STEP1701)。而且将该压缩数据块二重化 (STEP1702)。将压缩数据块及其二重化后的数据块存储在高速缓冲存储器 62 中的压缩数据用块中。在本例中，将 3 个压缩数据块作为 1 个数据周期，在其后接着存储二重化后的压缩数据块。

[0121] 然后，磁盘适配器 63 为了将压缩数据块分散存储在不同系列的伪磁盘设备中而进行以下处理。即，磁盘适配器 63 首先将 1 个数据周期的量的压缩数据块分散存储在不同的伪磁盘设备 (0) ～伪磁盘设备 (2) 中 (STEP1703)。然后，磁盘适配器 63 将数据周期变量 I 设置为 0(STEP1704)。

[0122] 磁盘适配器 63 判断数据周期变量 I 是否为 0(STEP1705)。即，磁盘适配器 63 选择在伪磁盘设备 (0) 中存储的压缩数据块 (例如 D4-1) 所对应的二重化后的压缩数据块 (例如 D4-1’)。

[0123] 磁盘适配器 63，当判断为数据周期变量 I 是 0 时 (STEP1705，是)，将该二重化后的压缩数据块存储在伪磁盘设备 (3) 的实际存储区域中 (STEP1706)，在数据块管理表中登录该二重化后的压缩数据块的条目 (STEP1707)。

[0124] 与之相对，当数据周期变量 I 是 0 以外时，磁盘适配器 63 将该二重化后的压缩数据块存储在伪磁盘设备 (I-1) 的扩展存储区域中 (STEP1708)，在数据块管理表中登录该二重化后的压缩数据块的条目 (STEP1709)。即，将伪磁盘设备 (0) 以外的伪磁盘设备中存储的压缩数据块所对应的二重化后的压缩数据块，存储在循环改变 (rotate shift) 了一次的伪磁盘设备中。

[0125] 然后，磁盘适配器 63 将数据周期变量 I 的值增加 1(STEP1710)，判断数据周期变量 I 是否到达 1 个数据周期 (STEP1711)。当数据周期变量 I 未到达 1 个数据周期时 (STEP1711，否)，磁盘适配器 63 返回 STEP1705 的处理。另一方面，当数据周期变量 I 到达了 1 个数据周期时 (STEP1711，是)，结束该写入处理。

[0126] 如上所述，根据本实施方式，在 RAID1 的结构下将写入数据的压缩数据进行镜像来存储时，将复制的压缩数据存储在虚拟卷中，因此可以实现高效的数据存储。

[0127] (第三实施方式)

[0128] 第三实施方式是上述第一实施方式中表示的技术特征和上述第二实施方式中表示的技术特征的组合。具体而言，在本实施方式中，存储装置 4 在数据 写入处理时，针对每个校验组生成校验数据块，对各数据块进行压缩处理。存储装置 4 基于压缩处理的结果，判断使用校验位 (parity) 的数据存储方式和基于镜像的数据存储方式中的哪种方式是高效的数据存储，按照被认为高效的数据存储方式，将压缩数据块存储在数据卷中。在这种情况

下，如上所述，适宜将压缩数据块存储在虚拟卷中。

[0129] 图 18 和图 19 是用于说明本发明的一个实施方式的存储装置 4 中的数据写入机制的图。具体而言，这些图表示了在将某写入数据压缩后，在 RAID5(3D+1P) 或 RAID1(4D) 下存储在伪磁盘设备中的情况。

[0130] 现在，假定有针对某个数据卷的数据 D1 的写入请求，接着有数据 D5 的写入请求。为了简化说明，设写入数据是块大小为 64KB 的单一的数据块。为了在 RAID5 控制下将写入数据（数据块）存储在数据卷中，需要与各个相同的校验组内的数据块一起计算校验位并将它们存储。即，关于数据块 D1，与数据块 D2 和 D3 一起计算校验数据块 P1，另外，关于数据块 D5，与数据块 D4 和 D6 一起计算校验数据块 P2。它们是在高速缓冲存储器 62 上的 64KB 块中存储的、压缩前的数据块。

[0131] 假定数据块 D1 ～ D3 和 P1 通过磁盘适配器 63 被压缩，得到图 18(A) 所示的数据块。它们是被存储在高速缓冲存储器 62 上的 16KB 块中的压缩数据块。如上所述，压缩数据块 D1-2、D2-2 以及 P1-2 ～ P1-4，是无法完全存储在数据卷的实际卷的实际存储区域中的数据块。

[0132] 另一方面，假定数据块 D4 ～ D6 以及 P2 通过磁盘适配器 63 被压缩，得到图 18(B) 所示的压缩数据块。即，将数据块 D4 ～ D6 变换为，压缩到可以存储在数据卷的实际卷中的一个实际存储区域中而得的压缩数据块 D4-1 ～ D6-1。另外，数据块 P2 的压缩效率低（压缩率高），被变换为压缩数据块 P2-1 ～ P2-4。此外，在本说明书中，压缩率是指将压缩后的数据大小除以压缩前的数据大小而得到的值。因此，所谓压缩率低，意味着压缩后的数据大小较小、压缩效率高。

[0133] 因此，根据数据块 D4 ～ D6 和 P2 生成 7 个压缩数据块。与之相对，在将数据块 D4-1 ～ D6-1 二重化来存储在数据卷中时，若存储 6 个压缩数据块便足够。因此，本实施方式的存储装置 4 在压缩写入数据后，检查包含校验位的压 缩数据块的数量，选择存储更少数量的压缩数据块便足够的方式。

[0134] 因此，如图 19 所示，将压缩数据块 D1-1、D2-1、D3-1 以及 P-1 在 RAID5 控制下分别存储在构成数据卷的 RAID 组中的伪磁盘设备 (0) ～伪磁盘设备 (3) 的实际卷中。另外，将压缩数据块 D1-2、D2-2 以及 P-2 ～ P-4 分别存储在该 RAID 组中的对应的伪磁盘设备 (0) ～伪磁盘设备 (4) 的虚拟卷中。

[0135] 另外，压缩数据块 D4-1 ～ D6-1 在 RAID1 控制下被镜像，利用扩展存储区域存储在伪磁盘设备 (0) ～伪磁盘设备 (3) 中。

[0136] 图 20 是表示本发明的一个实施方式的存储装置 4 中的共享存储器 64 中存储的磁盘设备管理表的一例的图。具体而言，该图表示了按照上述例子通过使用校验位的数据存储方式存储了写入数据时的磁盘设备管理表。通过基于镜像的数据存储方式存储了写入数据时的磁盘设备管理表，与图 16 所示的磁盘设备管理表相同。

[0137] 图 21 ～图 23 是用于说明本发明的一个实施方式的存储装置 4 中的数据写入处理的流程图。具体而言，该图说明了数据写入处理时的磁盘适配器 63 的动作。

[0138] 即，磁盘适配器 63 参照共享存储器 64，当发现未处理的写入命令时，确定高速缓冲存储器 62 上与之对应的写入数据。磁盘适配器 63 根据该写入数据，生成遵从 RAID 结构的校验数据 (STEP2101)。即，在本例中，磁盘适配器 63 遵从 RAID5(3D+1P) 的结构，针对每

3 个数据块形成由 1 个校验数据块构成的校验组。在此阶段，各数据块是 64KB 的块大小。

[0139] 接着，磁盘适配器 63 对形成的校验组内的各数据块进行压缩处理，分别生成压缩数据列 (STEP2102)。磁盘适配器 63 将构成各压缩数据列的每一个压缩数据块，写入高速缓冲存储器 62 中的压缩数据用块中，更新高速缓冲存储管理表。压缩数据用块例如是 16KB 的块大小。

[0140] 接着，磁盘适配器 63 判断校验组内的压缩数据块（不包含压缩校验数据块）数是否比压缩校验数据块数少 (STEP2103)。即，磁盘适配器 63 判断使用校验位的数据存储方式和基于镜像的数据存储方式中的哪种方式可以高效地存储写入数据。

[0141] 磁盘适配器 63，当判断为校验组内的压缩数据块数比压缩校验数据块数 少时，进行图 22 所示的处理。图 22 所示的处理是使用校验位的数据存储处理，与图 12 所示的 STEP1203 以后的处理相同，因此省略说明。

[0142] 与之相对，磁盘适配器 63 在判断为校验组内的压缩数据块数比压缩后校验数据块数多时，进行图 23 所示的处理。图 23 所示的处理是基于镜像的数据存储处理，与图 17 所示的 STEP1702 以后的处理相同，因此省略说明。

[0143] 如上所述，根据本实施方式，当压缩写入数据来进行存储时，根据包含校验位的压缩数据的大小、和不使用校验位而进行二重化所得到的压缩数据的大小，选择最佳的 RAID 结构，因此可以实现压缩数据的高效存储。

[0144] （第四实施方式）

[0145] 第四实施方式是第三实施方式的变形，其特征在于，存储了针对数据写入请求的前一次的数据存储方式，当存在针对该数据的其它写入请求时，利用该存储的数据存储方式。

[0146] 图 24(A) 和 (B) 是表示本发明的一个实施方式的存储装置 4 中的共享存储器 64 中存储的磁盘设备管理表的一例的图。如该图所示，本实施方式的磁盘设备管理表与上述磁盘设备管理表大体相同，但在保持了存储前一次的数据存储方式的学习信息这点不同。即，该图 (A) 表示将 RAID5 作为学习信息而保持的状态，该图 (B) 表示没有学习信息的状态。

[0147] 图 25 是用于说明本发明的一个实施方式的存储装置 4 中的数据写入处理的流程图。

[0148] 如该图所示，磁盘适配器 63 参照共享存储器 64，当发现未处理的写入命令时，确定高速缓冲存储器 62 上的与之对应的写入数据。磁盘适配器 63 参照针对写入数据的数据块管理表，取得当前的学习信息 (STEP2501)，判断当前的学习信息是否是“RAID1”(STEP2502)。

[0149] 磁盘适配器 63，当判断为当前的学习信息不是“RAID1”时 (STEP2502，否)，在 RAID5 的控制下进行压缩数据的存储处理。

[0150] 即，磁盘适配器 63 如前所述那样生成遵从 RAID 结构的校验数据块 (STEP2503)，接着，压缩校验组内的各数据块 (STEP2504)。

[0151] 然后，磁盘适配器 63 判断校验组内的压缩数据块（不包含压缩校验数据块）数是否比压缩校验数据块数少 (STEP2505)。即，磁盘适配器 63 判断使 用校验位的数据存储方式和使用镜像的数据存储方式中的哪种方式可以高效地存储写入数据。

[0152] 磁盘适配器 63，当判断为校验组内的压缩数据块数比压缩校验数据块数多时，进

行基于镜像的数据存储处理 (STEP2509)。基于镜像的数据存储处理与图 23 所示的处理相同。

[0153] 与之相对，磁盘适配器 63，当判断为校验组内的压缩数据块数比压缩校验数据块数少时，进行使用校验位的数据存储处理 (STEP2506)。使用校验位的数据存储处理与图 22 所示的处理相同。

[0154] 然后，为了下一次的数据写入处理，磁盘适配器 63 将数据块管理表的学习信息设定为“RAID5”(STEP2507)。

[0155] 在 STEP2502 中，当判断为当前的学习信息是“RAID1”时 (STEP2502，是)，在 RAID1 控制下进行压缩数据的存储处理。

[0156] 即，磁盘适配器 63 首先压缩数据块 (STEP2508)，接着进行基于镜像的数据存储处理 (STEP2509)。

[0157] 在结束基于镜像的数据存储处理后，磁盘适配器 63 为了判断是否更新学习信息，而生成遵从 RAID 结构的校验数据块 (STEP2510)，接着，压缩校验组内的各数据块 (STEP2511)。磁盘适配器 63 判断校验组内的压缩数据块 (不包含压缩后校验数据块) 数是否比压缩校验数据块数少 (STEP2512)。

[0158] 磁盘适配器 63，当判断为校验组内的压缩数据块数比压缩校验数据块数少时 (STEP2512，是)，将数据块管理表的学习信息设定为“RAID5”(STEP2507)。

[0159] 与之相对，磁盘适配器 63 当判断为校验组内的压缩数据块数比压缩后校验数据块数多时 (STEP2512，否)，将数据块管理表的学习信息设定为“RAID1”(STEP2513)。

[0160] 通过以上所述，根据本实施方式，由于存储了前一次的数据存储方式，因此可以进行更高效的 RAID 结构的选择。

[0161] 上述实施方式是用于说明本发明的例子，本发明不限定于上述实施方式。本发明在不超出其主旨的范围内可以通过各种形态来实施。

[0162] 例如，在上述实施方式中，在流程图中按顺序地构成了磁盘适配器的动作， 但只要不在动作结果中产生矛盾，也可以改变处理的顺序或者并行进行处理。

[0163] 另外，在上述实施方式中采用了 RAID5，但也可以是作为使用校验位的其它 RAID 结构而已知的 RAID3、RAID4 或 RAID6。

[0164] 产业上的可利用性

[0165] 本发明可以广泛应用于采用 RAID 结构的存储装置。

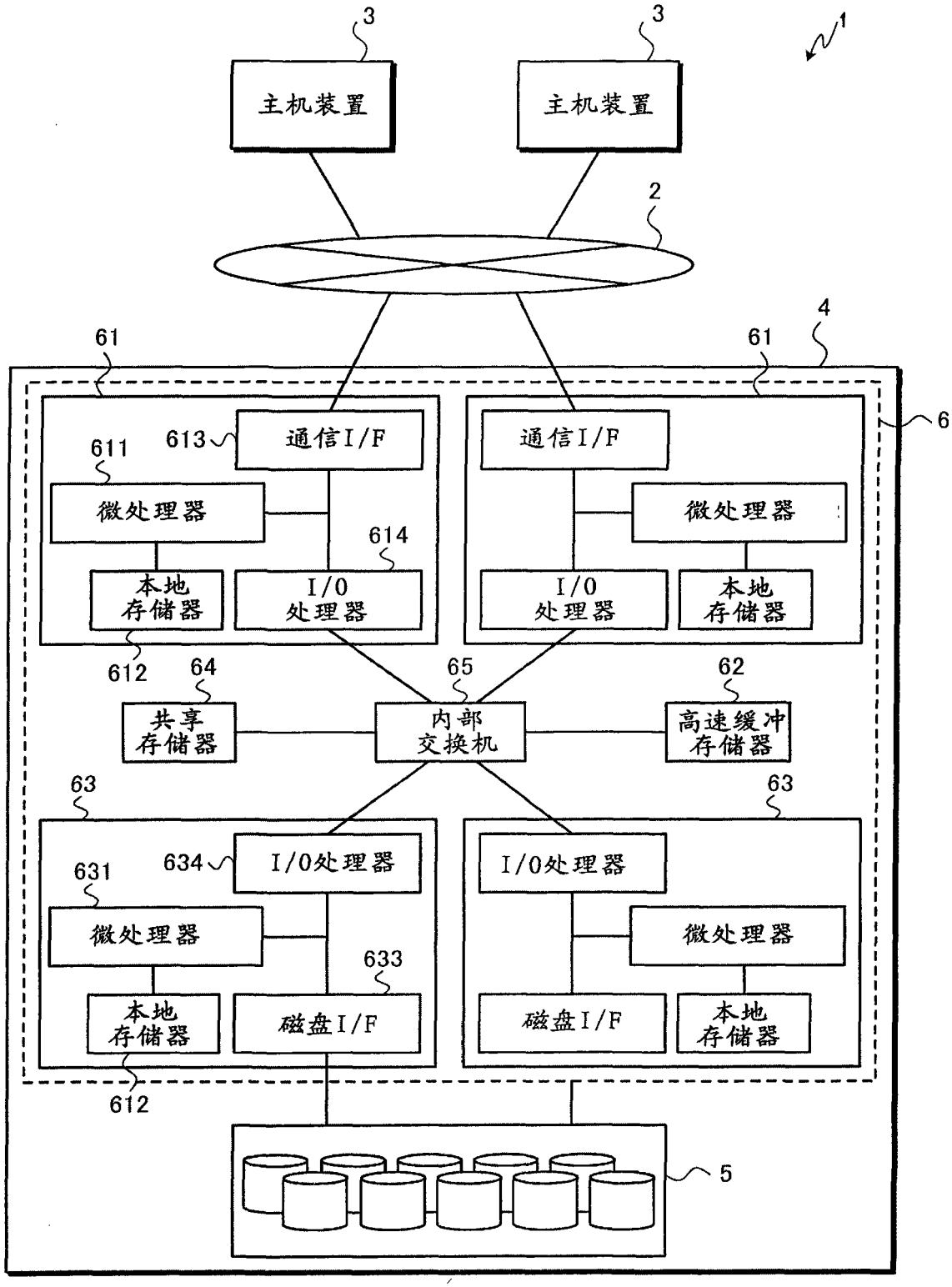

1
2
3
3
主机装置
主机装置
4
5
6
61
61
611
613
通信I/F
通信I/F
微处理器
微处理器
614
本地存储器
I/O处理器
I/O处理器
本地存储器
612
64
65
62
共享存储器
内部交换机
高速缓冲存储器
63
63
631
634
I/O处理器
I/O处理器
微处理器
微处理器
633
本地存储器
磁盘I/F
磁盘I/F
本地存储器
612

图 1

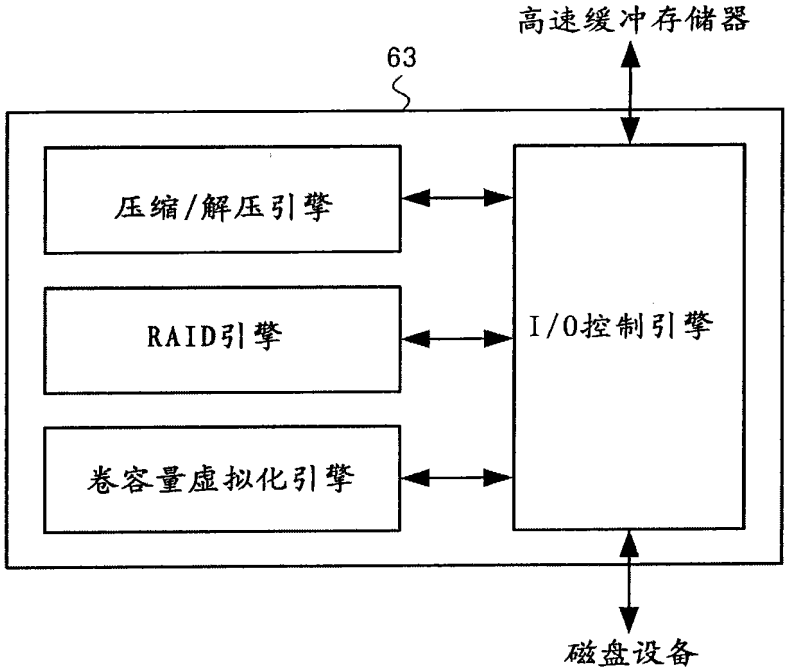


图 2

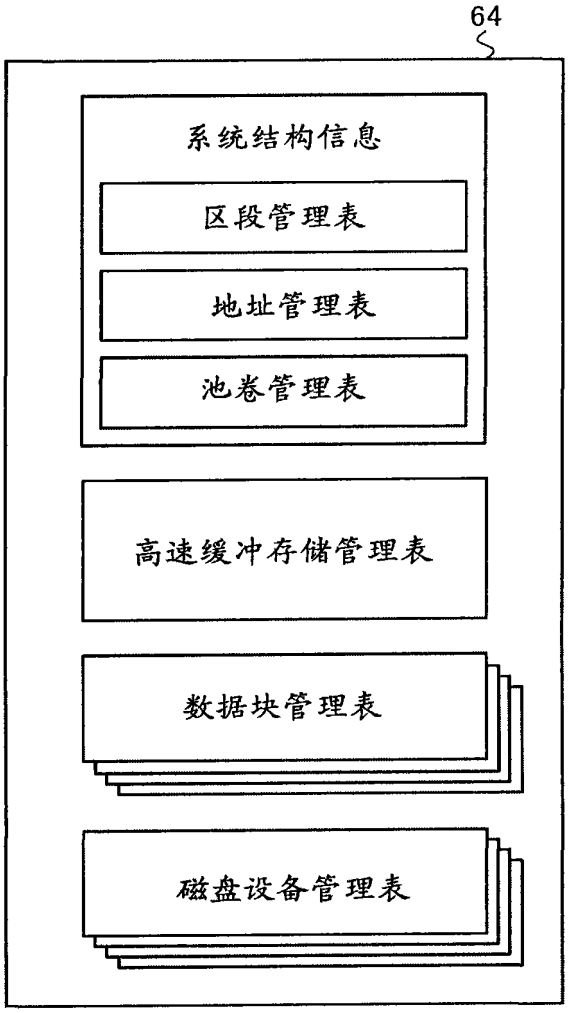


图 3

区段管理表

Disk ID	Segment #	LBA_start	LBA_end	In_use
0	0	0	499	1
0	1	500	1799	1
0	2	1800	2999	0
0	3	1800	2999	1
:	:	:	:	:
1	6	0	699	0
1	7	700	2999	1
:	:	:	:	:

图 4

地址管理表

LUN	Segment #	LBA_start	LBA_end	Pool
0	1	0	1299	0
0	6	1300	1999	0
1	2	0	1199	0
2	8	0	299	0
:	:	:	:	:
9	8	19000	19999	1
:	:	:	:	:

图 5

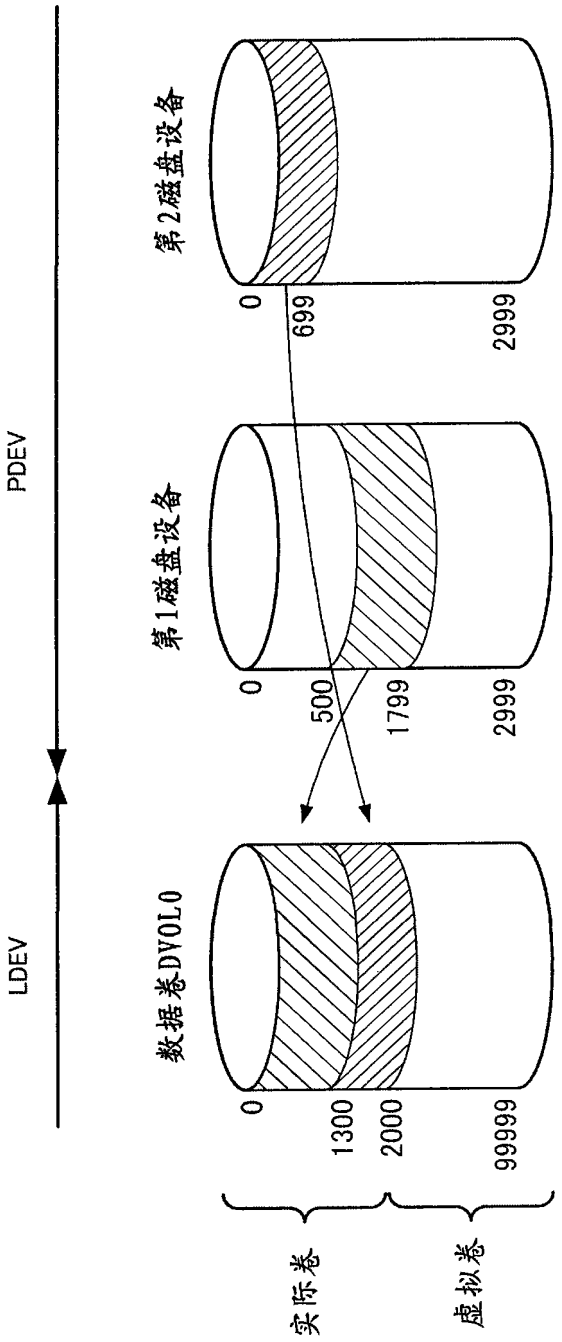

PDEV
LDEV
第2磁盘设备
第1磁盘设备
数据卷DVOL0
0
699
2999
0
500
1799
2999
0
1300
2000
99999
实际卷
虚拟卷

图 6

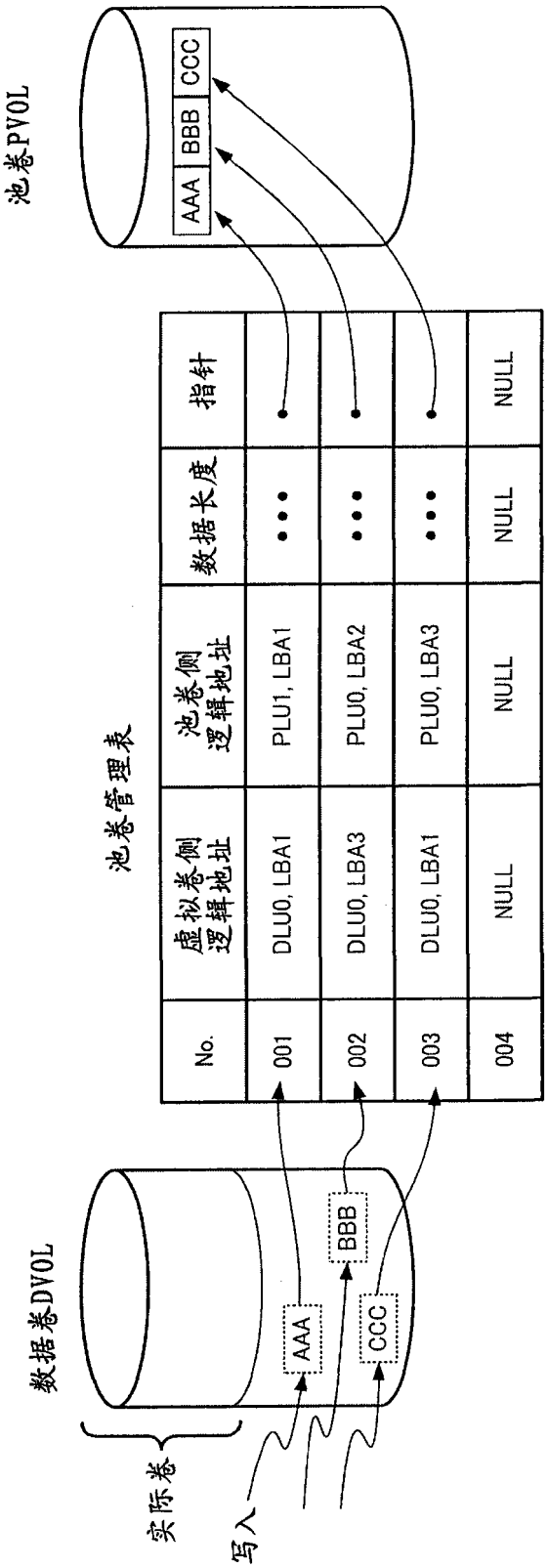

池卷PV0L
AAA
BBB
CCC
池卷管理表
No.
虚拟卷侧逻辑地址
池卷侧逻辑地址
数据长度
指针
001
DLU0, LBA1
PLU1, LBA1
002
DLU0, LBA3
PLU0, LBA2
003
DLU0, LBA1
PLU0, LBA3
004
NULL
NULL
NULL
NULL
数据卷DV0L
AAA
BBB
CCC
实际卷
写入

图 7

数据块管理表

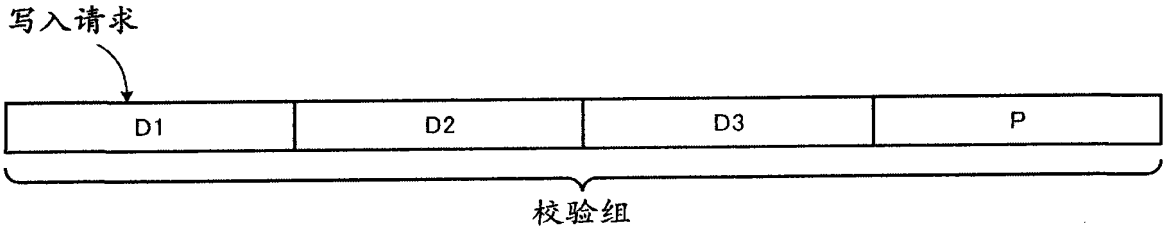
槽ID	地址
D1-1	DEVICE(0): xxxx
D2-1	DEVICE(1): xxxx
D3-1	DEVICE(2): xxxx
P1-3	DEVICE(3): xxxx
D1-2	DEVICE(0): xxxx
D2-2	DEVICE(1): xxxx
D3-2	UNUSED
P1-2	DEVICE(3): xxxx
D1-3	UNUSED
D2-3	UNUSED
D3-3	UNUSED
P1-3	DEVICE(3): xxxx
D1-4	UNUSED
D2-4	UNUSED
D3-4	UNUSED
P1-4	DEVICE(3): xxxx

图 8

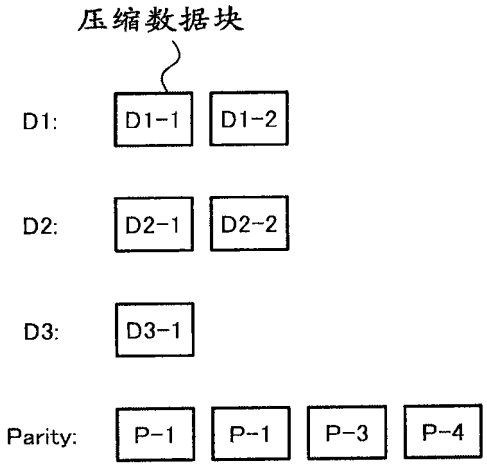
磁盘设备管理表

	地址	使用状况
实际卷	xxxx	D1-1
	:	:
	:	:
	:	:
	:	:
虚拟卷	扩展1	D2-1
	扩展2	Free
	扩展3	Free
	:	:
	扩展n	Null

图 9


写入请求
D1
D2
D3
P
校验组

（A）


压缩数据块
D1:
D1-1
D1-2
D2:
D2-1
D2-2
D3:
D3-1
Parity:
P-1
P-1
P-3
P-4

（B）

图 10

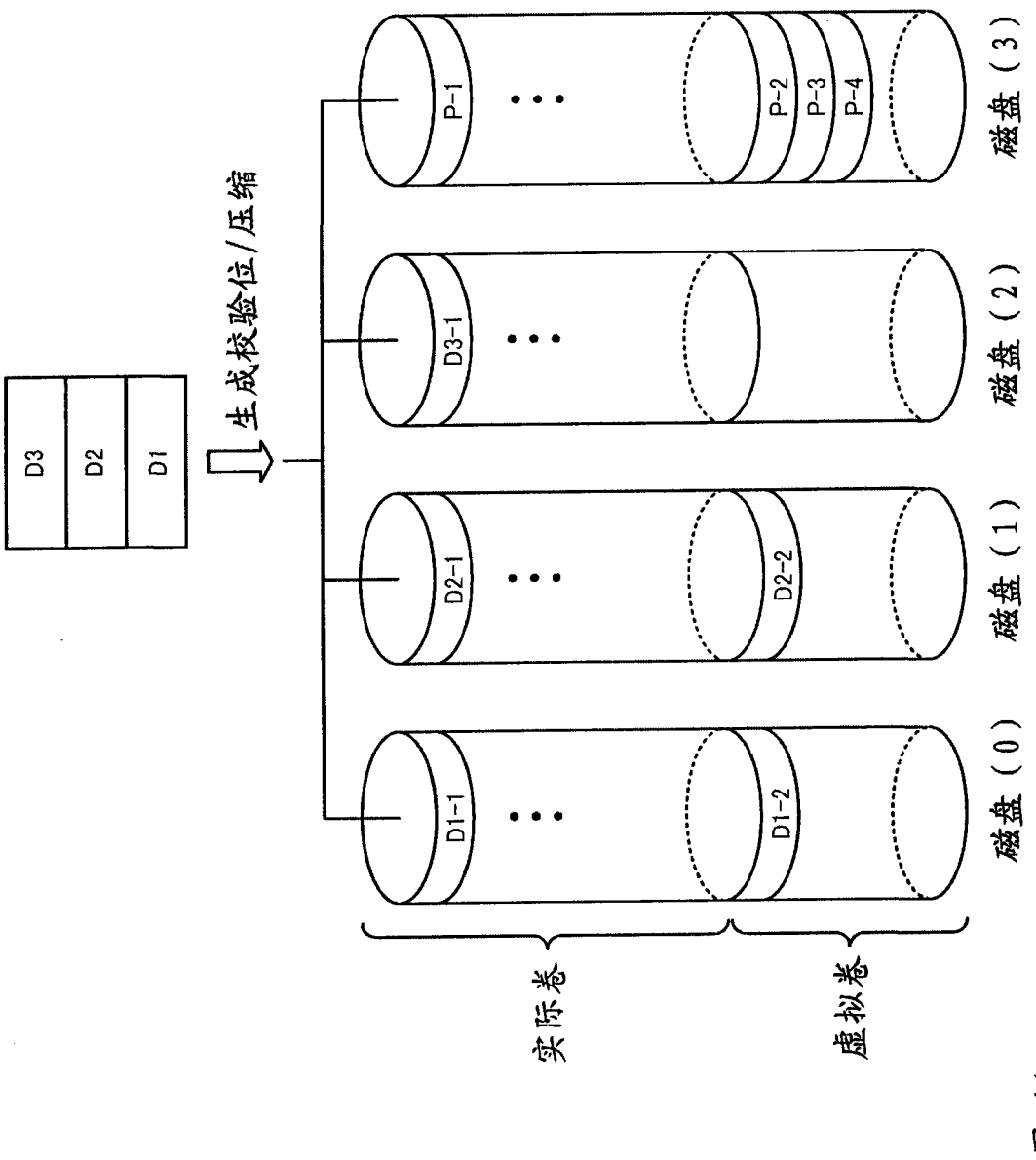

D3
D2
D1
生成校验位/压缩
D1-1
D1-2
D2-1
D2-2
D3-1
P-1
P-2
P-3
P-4
磁盘（0）
磁盘（1）
磁盘（2）
磁盘（3）
实际卷
虚拟卷

图 11

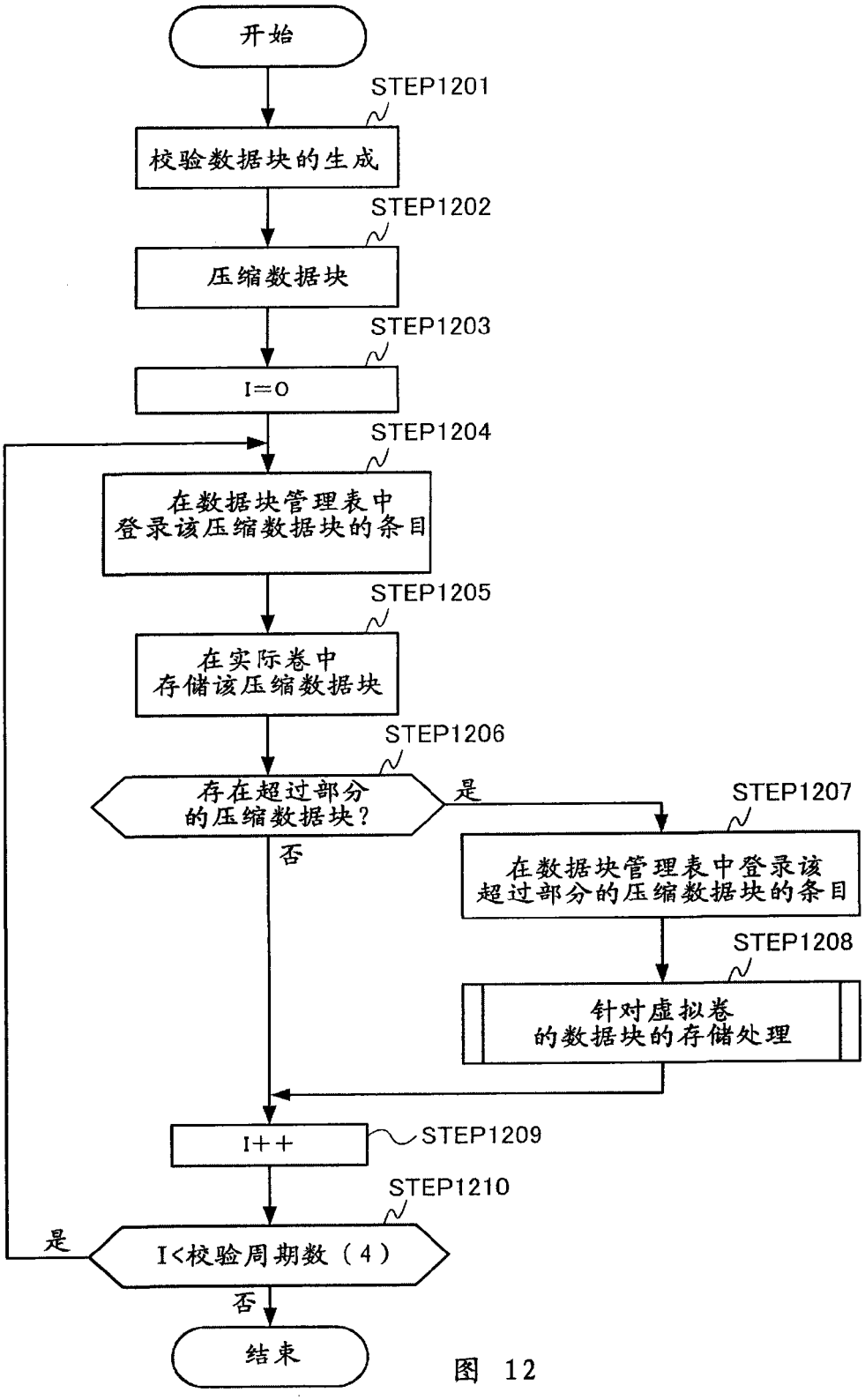

开始
STEP1201
校验数据块的生成
STEP1202
压缩数据块
STEP1203
I=0
STEP1204
在数据块管理表中
登录该压缩数据块的条目
STEP1205
在实际卷中
存储该压缩数据块
STEP1206
存在超过部分
的压缩数据块？
是
STEP1207
在数据块管理表中登录该
超过部分的压缩数据块的条目
STEP1208
针对虚拟卷
的数据块的存储处理
否
I++
STEP1209
STEP1210
是
I<校验周期数（4）
否
结束

图 12

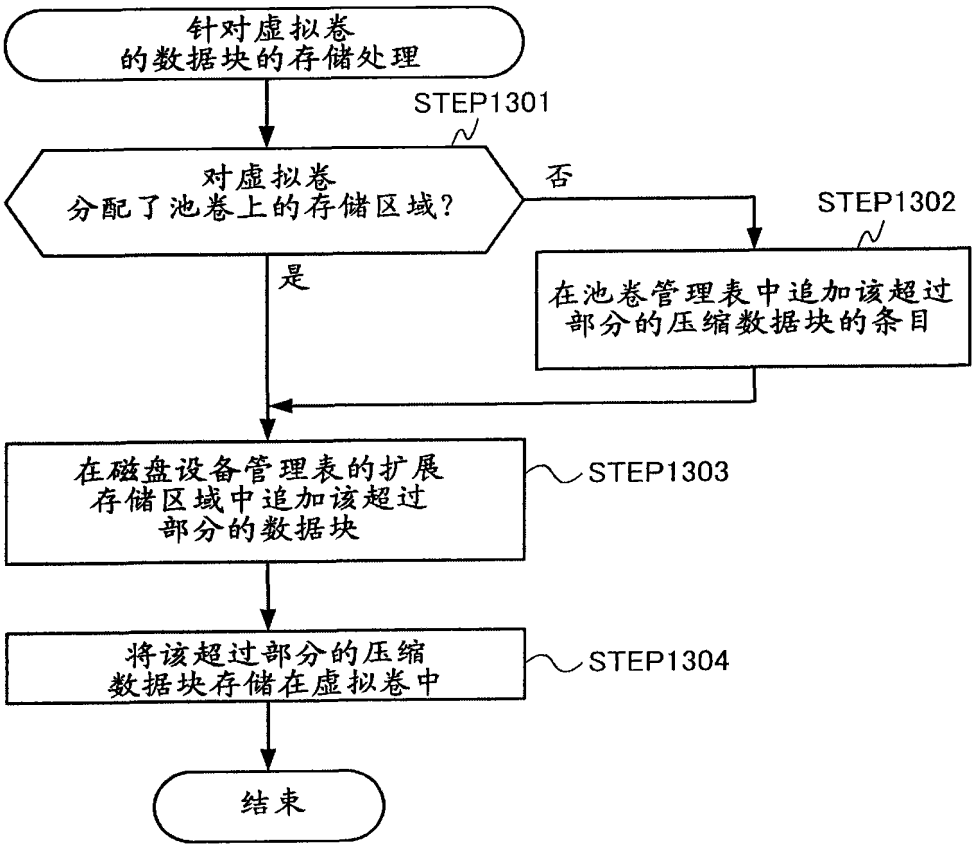

针对虚拟卷
的数据块的存储处理
STEP1301
对虚拟卷
分配了池卷上的存储区域?
否
是
STEP1302
在池卷管理表中追加该超过
部分的压缩数据块的条目
在磁盘设备管理表的扩展
存储区域中追加该超过
部分的数据块
STEP1303
将该超过部分的压缩
数据块存储在虚拟卷中
STEP1304
结束

图 13

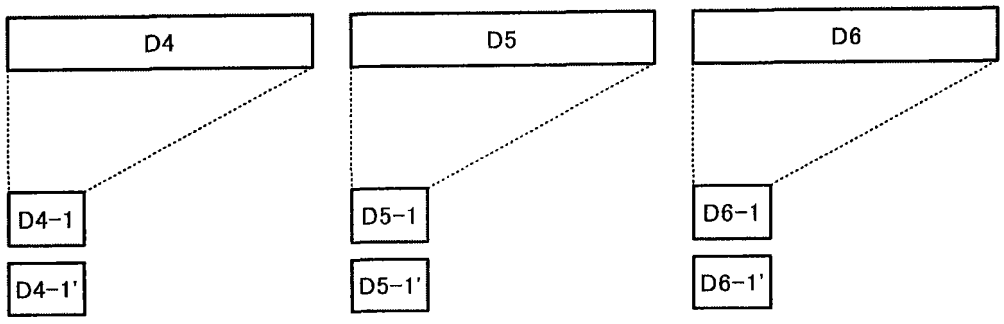

D4
D5
D6
D4-1
D4-1'
D5-1
D5-1'
D6-1
D6-1'

图 14

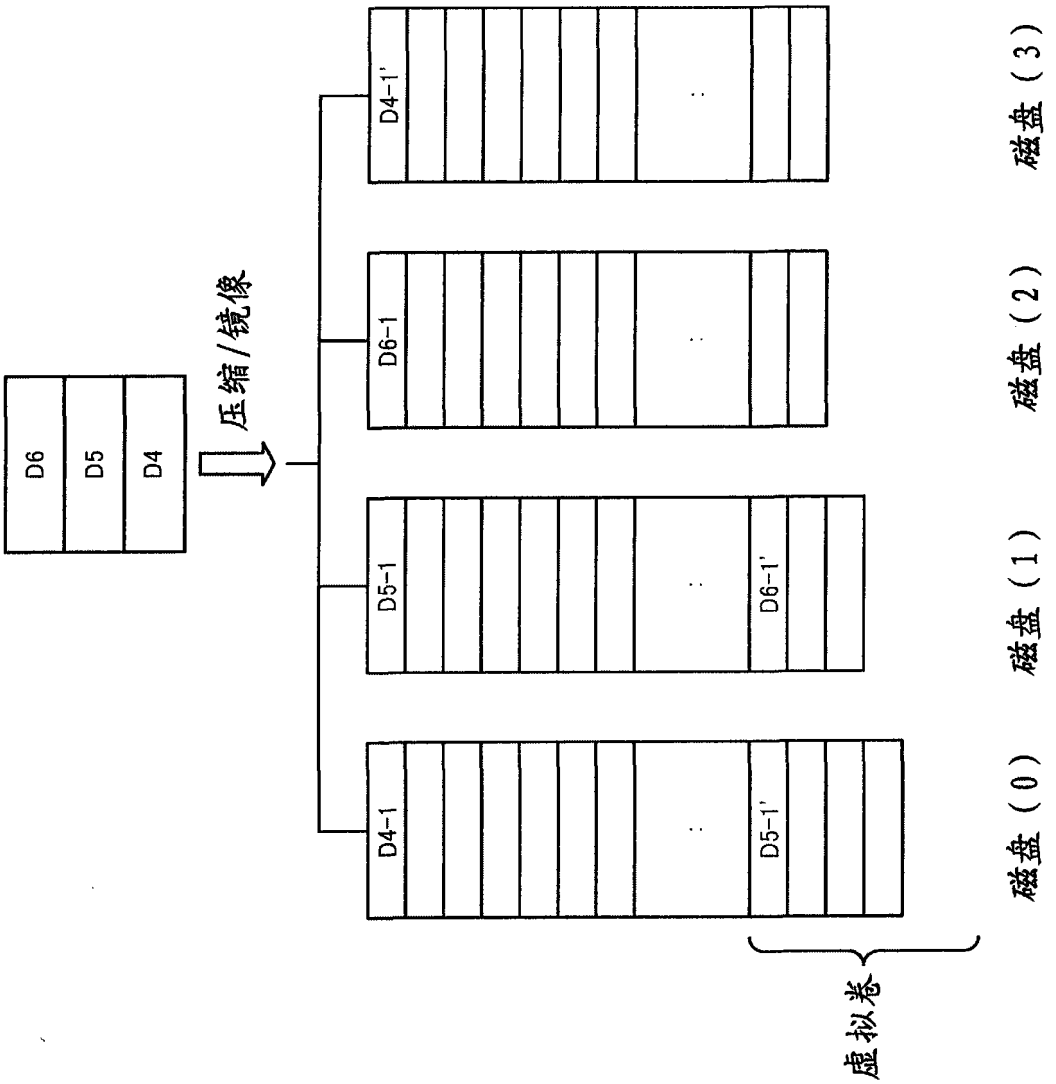


图 15

数据块管理表

RAID级别	RAID 1	
槽ID	地址1	地址2
D1-1	DEVICE(0): xxxx	DEVICE(3): xxxx
D2-1	DEVICE(1): xxxx	DEVICE(0): xxxx
D3-1	DEVICE(2): xxxx	DEVICE(1): xxxx
P1-3	UNUSED	UNUSED
D1-2	UNUSED	UNUSED
D2-2	UNUSED	UNUSED
D3-2	UNUSED	UNUSED
P1-2	UNUSED	UNUSED
D1-3	UNUSED	UNUSED
D2-3	UNUSED	UNUSED
D3-3	UNUSED	UNUSED
P1-3	UNUSED	UNUSED
D1-4	UNUSED	UNUSED
D2-4	UNUSED	UNUSED
D3-4	UNUSED	UNUSED
P1-4	UNUSED	UNUSED

图 16

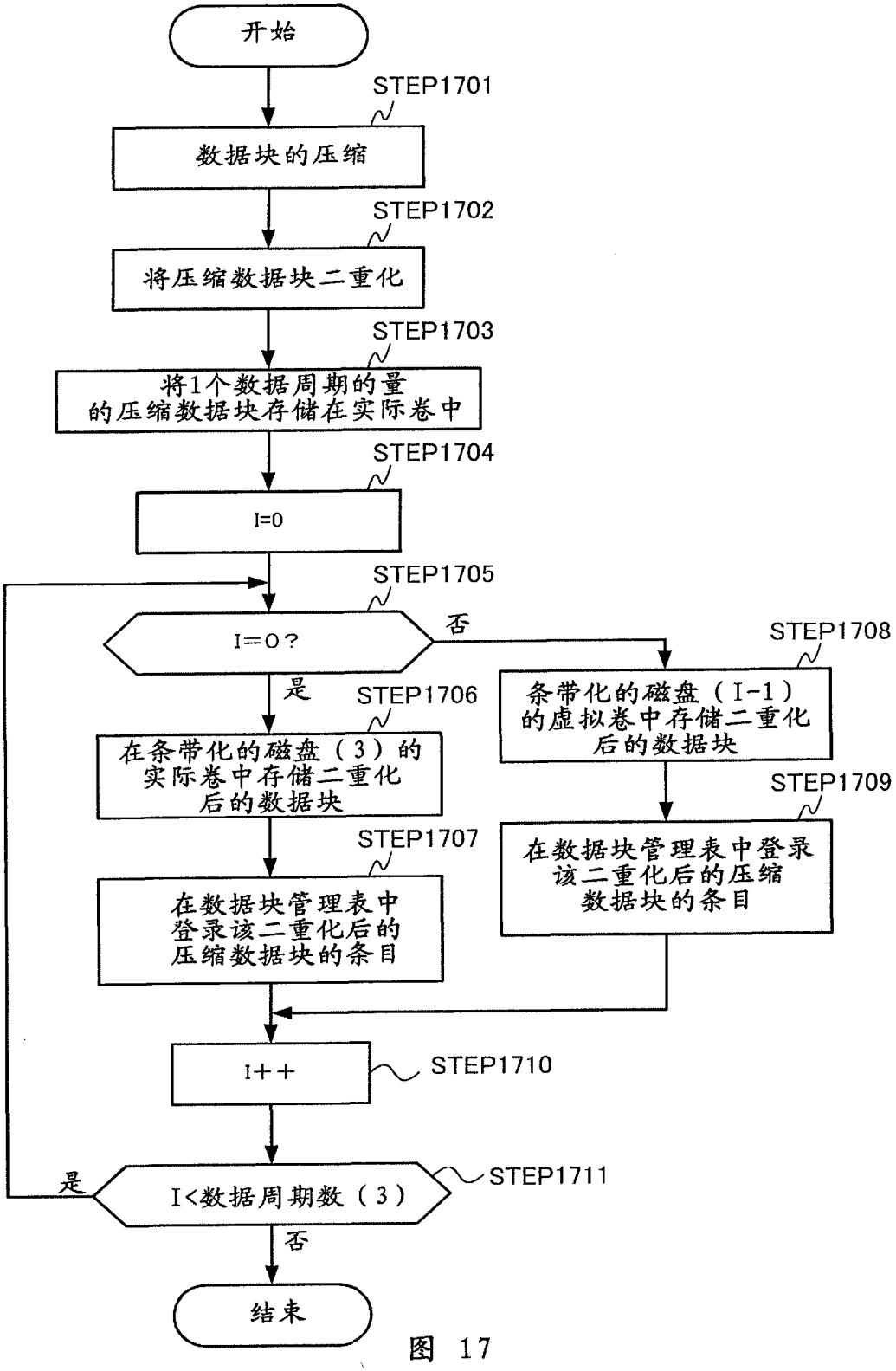
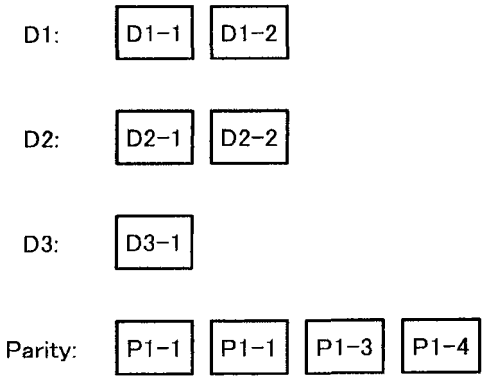
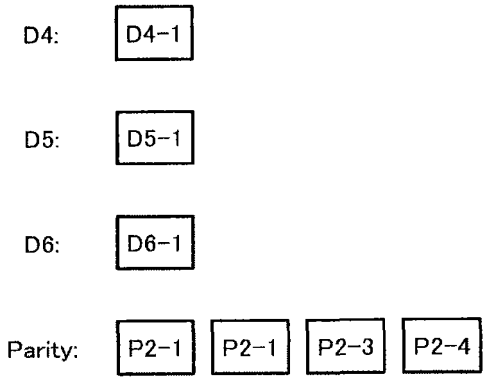

开始
STEP1701
数据块的压缩
STEP1702
将压缩数据块二重化
STEP1703
将1个数据周期的量的压缩数据块存储在实际卷中
STEP1704
I=0
STEP1705
I=0？
否
是
STEP1706
在条带化的磁盘（3）的实际卷中存储二重化后的数据块
STEP1707
在数据块管理表中登录该二重化后的压缩数据块的条目
STEP1708
条带化的磁盘（I-1）的虚拟卷中存储二重化后的数据块
STEP1709
在数据块管理表中登录该二重化后的压缩数据块的条目
I++
STEP1710
I<数据周期数（3）
STEP1711
是
否
结束

图 17


D1:
D1-1
D1-2
D2:
D2-1
D2-2
D3:
D3-1
Parity:
P1-1
P1-1
P1-3
P1-4

(A)


D4:
D4-1
D5:
D5-1
D6:
D6-1
Parity:
P2-1
P2-1
P2-3
P2-4

(B)

图 18

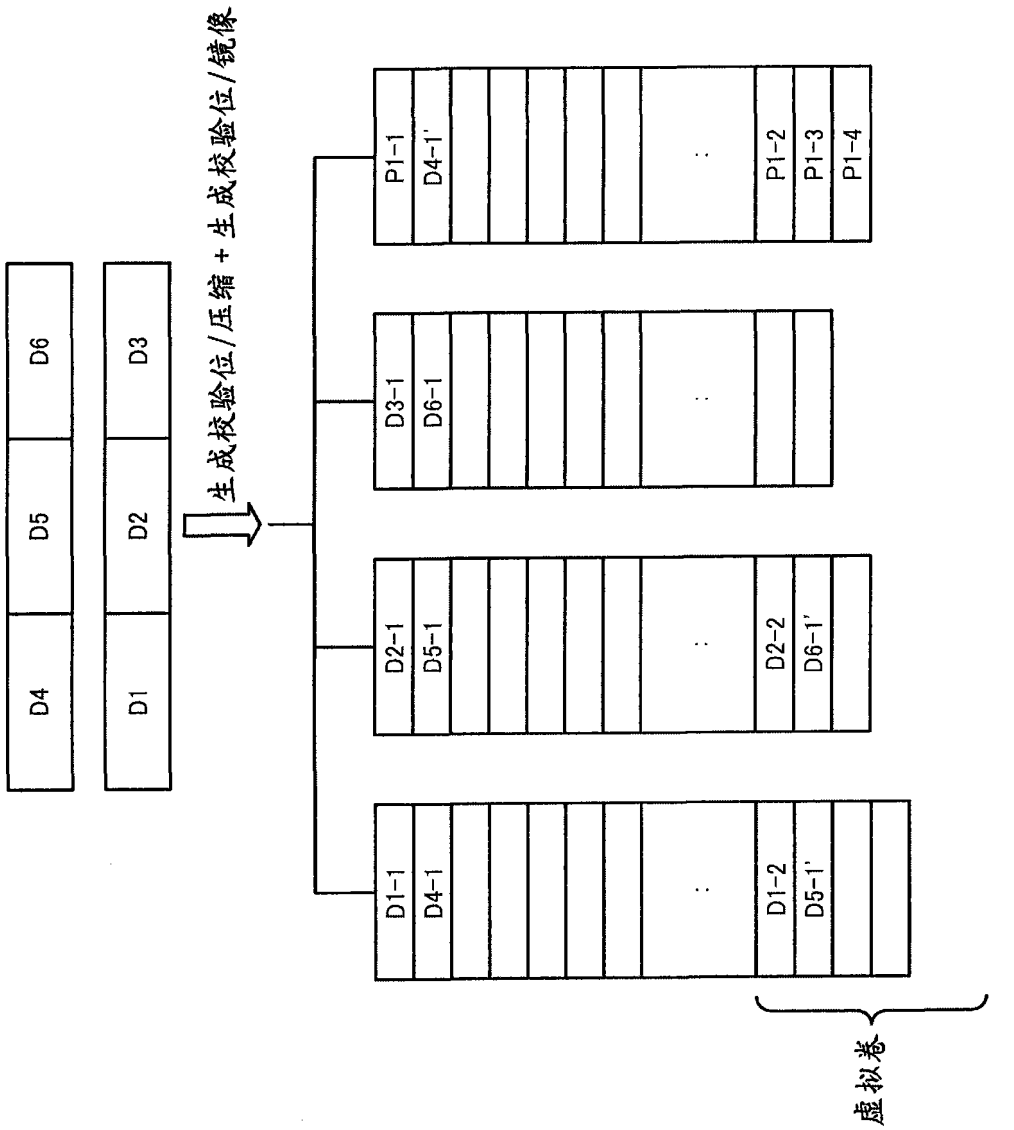

D4
D5
D6
D1
D2
D3
生成校验位/压缩 + 生成校验位/镜像
D1-1
D4-1
D1-2
D5-1'
D2-1
D5-1
D2-2
D6-1'
D3-1
D6-1
P1-1
D4-1'
P1-2
P1-3
P1-4
虚拟卷
磁盘（0）
磁盘（1）
磁盘（2）
磁盘（3）

图 19

数据块管理表

RAID 级别	RAID 5	
槽ID	地址1	地址2
D1-1	DEVICE(0): xxxx	UNUSED
D2-1	DEVICE(1): xxxx	UNUSED
D3-1	DEVICE(2): xxxx	UNUSED
P1-3	DEVICE(3): xxxx	UNUSED
D1-2	DEVICE(0): xxxx	UNUSED
D2-2	DEVICE(1): xxxx	UNUSED
D3-2	UNUSED	UNUSED
P1-2	DEVICE(3): xxxx	UNUSED
D1-3	UNUSED	UNUSED
D2-3	UNUSED	UNUSED
D3-3	UNUSED	UNUSED
P1-3	DEVICE(3): xxxx	UNUSED
D1-4	UNUSED	UNUSED
D2-4	UNUSED	UNUSED
D3-4	UNUSED	UNUSED
P1-4	DEVICE(3): xxxx	UNUSED

图 20

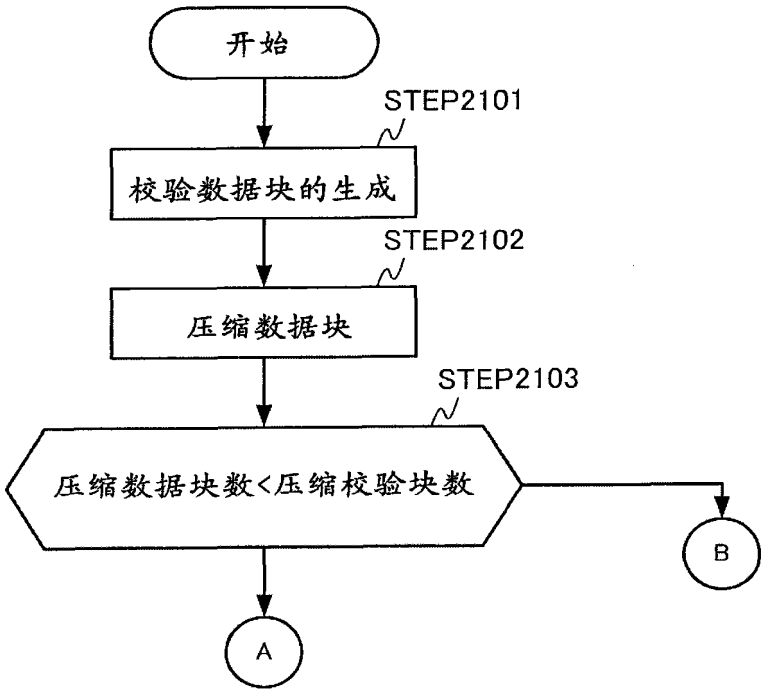

开始
STEP2101
校验数据块的生成
STEP2102
压缩数据块
STEP2103
压缩数据块数<压缩校验块数
B
A

图 21

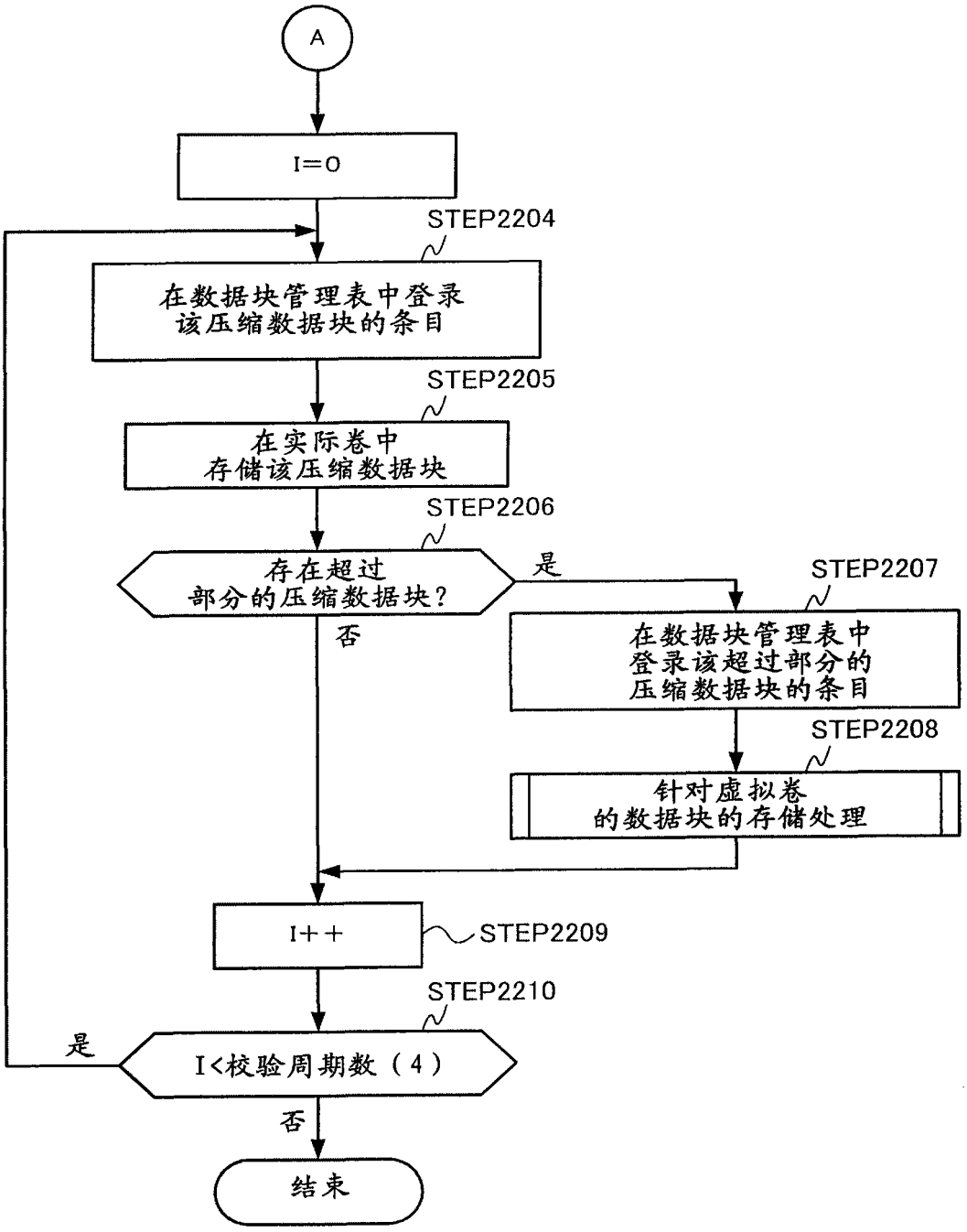

A
I=0
STEP2204
在数据块管理表中登录
该压缩数据块的条目
STEP2205
在实际卷中
存储该压缩数据块
STEP2206
存在超过
部分的压缩数据块？
是
否
STEP2207
在数据块管理表中
登录该超过部分的
压缩数据块的条目
STEP2208
针对虚拟卷
的数据块的存储处理
I++
STEP2209
STEP2210
是
I<校验周期数（4）
否
结束

图 22

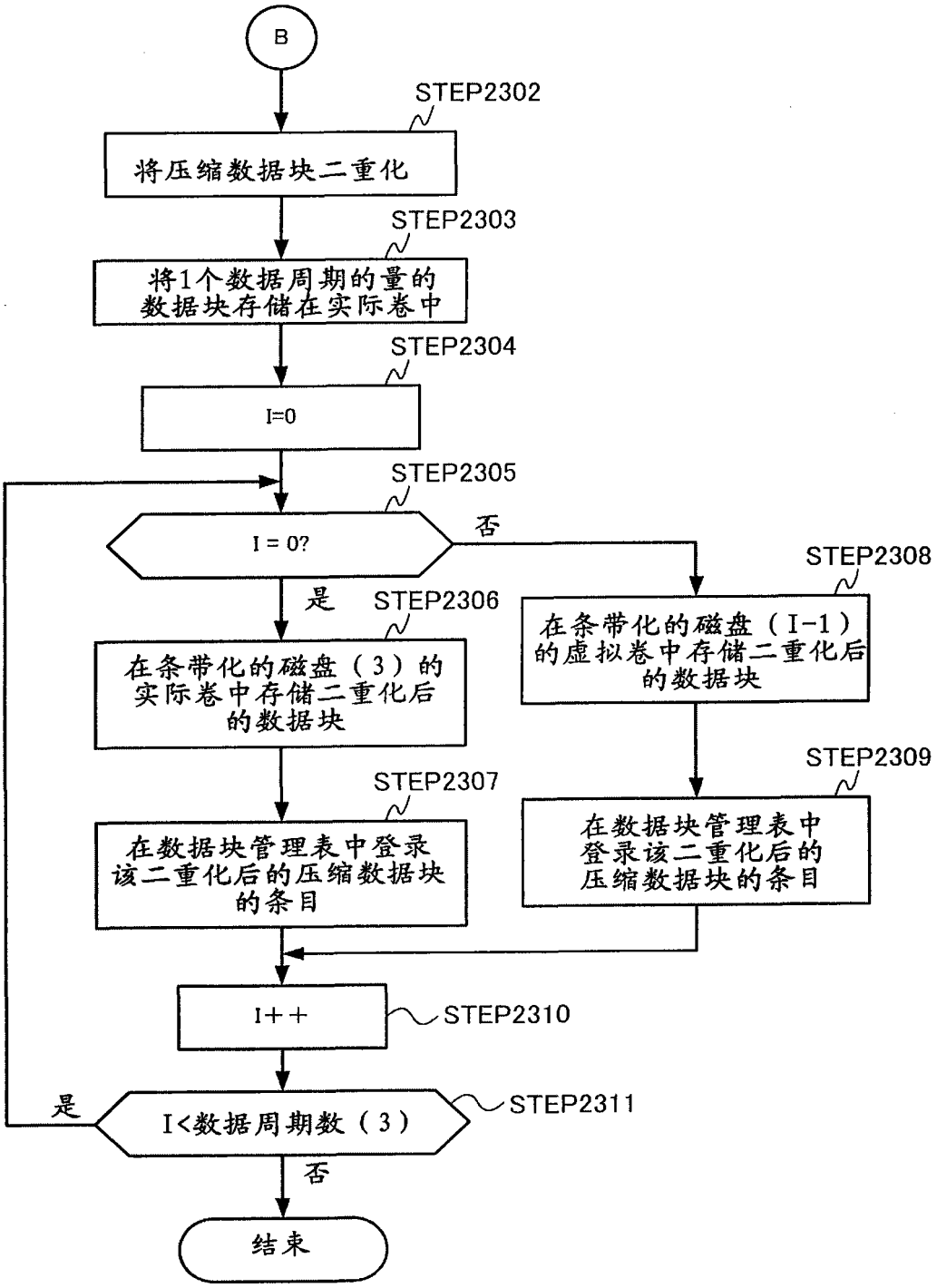

B
STEP2302
将压缩数据块二重化
STEP2303
将1个数据周期的量的数据块存储在实际卷中
STEP2304
I=0
STEP2305
I = 0?
否
是
STEP2306
在条带化的磁盘（3）的实际卷中存储二重化后的数据块
STEP2307
在数据块管理表中登录该二重化后的压缩数据块的条目
STEP2308
在条带化的磁盘（I-1）的虚拟卷中存储二重化后的数据块
STEP2309
在数据块管理表中登录该二重化后的压缩数据块的条目
I++
STEP2310
是
I<数据周期数（3）
STEP2311
否
结束

图 23

数据块管理表

RAID 级别	RAID 5	
学习信息	RAID 5	
槽ID	地址1	地址2
D1-1	DEVICE(0): xxxx	UNUSED
D2-1	DEVICE(1): xxxx	UNUSED
D3-1	DEVICE(2): xxxx	UNUSED
P1-3	DEVICE(3): xxxx	UNUSED
D1-2	DEVICE(0): xxxx	UNUSED
D2-2	DEVICE(1): xxxx	UNUSED
D3-2	UNUSED	UNUSED
P1-2	DEVICE(3): xxxx	UNUSED
D1-3	UNUSED	UNUSED
D2-3	UNUSED	UNUSED
D3-3	UNUSED	UNUSED
P1-3	DEVICE(3): xxxx	UNUSED
D1-4	UNUSED	UNUSED
D2-4	UNUSED	UNUSED
D3-4	UNUSED	UNUSED
P1-4	DEVICE(3): xxxx	UNUSED

(A)

数据块管理表

RAID 级别	RAID 1	
学习信息	-	
槽ID	地址1	地址2
D1-1	DEVICE(0): xxxx	DEVICE(3): xxxx
D2-1	DEVICE(1): xxxx	DEVICE(0): xxxx
D3-1	DEVICE(2): xxxx	DEVICE(1): xxxx
P1-3	UNUSED	UNUSED
D1-2	UNUSED	UNUSED
D2-2	UNUSED	UNUSED
D3-2	UNUSED	UNUSED
P1-2	UNUSED	UNUSED
D1-3	UNUSED	UNUSED
D2-3	UNUSED	UNUSED
D3-3	UNUSED	UNUSED
P1-3	UNUSED	UNUSED
D1-4	UNUSED	UNUSED
D2-4	UNUSED	UNUSED
D3-4	UNUSED	UNUSED
P1-4	UNUSED	UNUSED

(B)

图 24

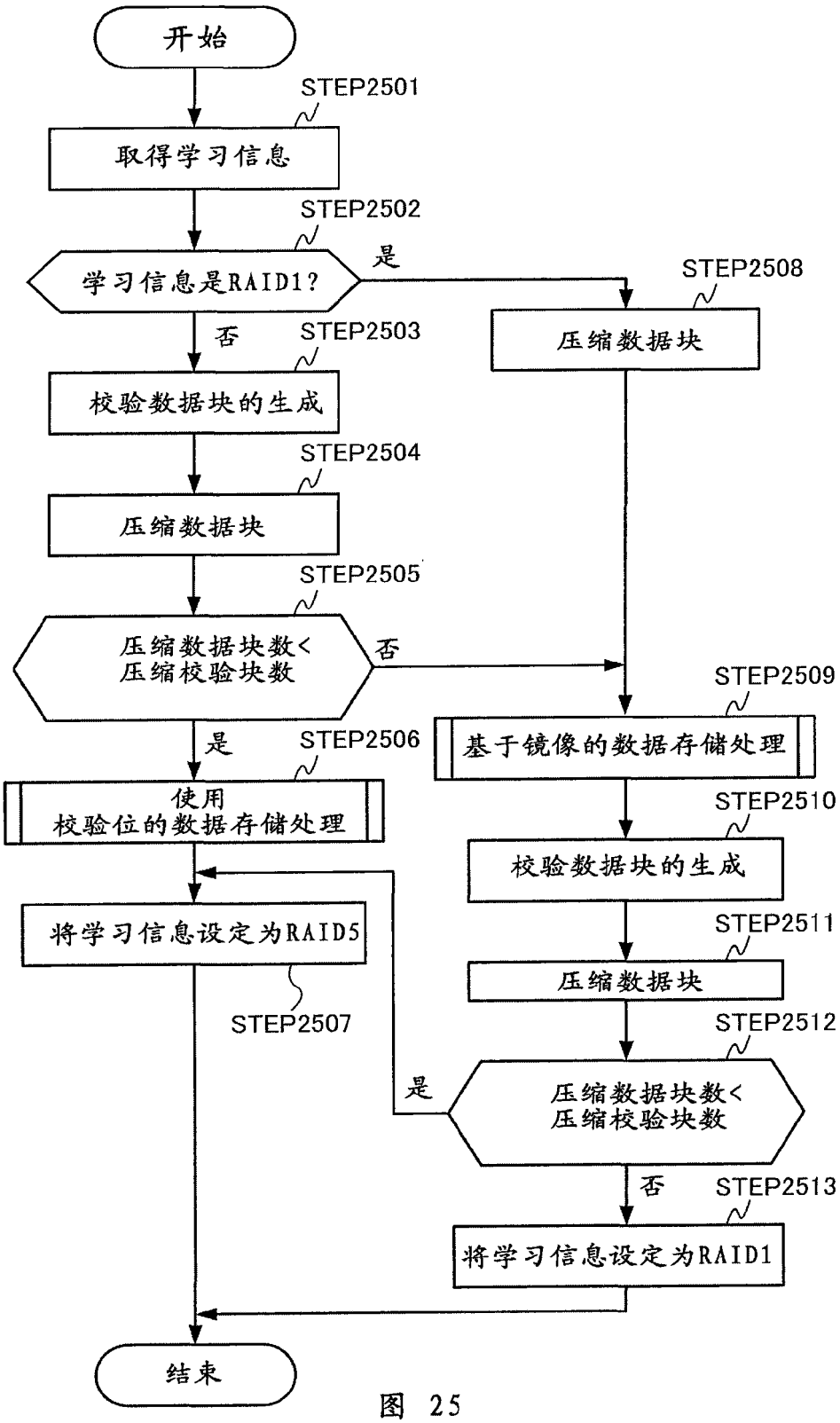

开始
STEP2501
取得学习信息
STEP2502
学习信息是RAID1?
是
否
STEP2508
压缩数据块
STEP2503
校验数据块的生成
STEP2504
压缩数据块
STEP2505
压缩数据块数<
压缩校验块数
否
是
STEP2506
使用
校验位的数据存储处理
STEP2509
基于镜像的数据存储处理
STEP2510
校验数据块的生成
STEP2511
压缩数据块
STEP2512
压缩数据块数<
压缩校验块数
是
否
STEP2513
将学习信息设定为RAID1
将学习信息设定为RAID5
STEP2507
结束

图 25