(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 101676857 B

(45)授权公告日 2013.04.24

(21)申请号 200910204748.2

(22)申请日 2006.06.29

(30)优先权数据

2005-275018 2005.09.22 JP

(62)分案原申请数据

200610094311.4 2006.06.29

(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 藤林昭 藤本和久 中村崇仁

(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)

G06F 1/32(2006.01)

(56)对比文件

US 6314503 B1,2001.11.06, 全文.

CN 1391672 A,2003.01.15, 全文.

US 2004/0054939 A1,2004.03.18, 全文.

审查员 李玉坤

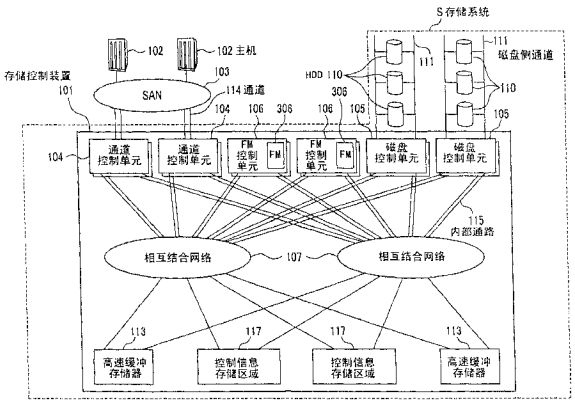
权利要求书3页 说明书13页 附图15页

(54)发明名称

存储系统

(57)摘要

本发明涉及一种存储系统。该存储系统包括：多个磁盘设备，存储从主机计算机接收的数据；多个非易失性设备，存储从主机计算机接收的数据；以及处理单元，控制从主机计算机到与多个磁盘设备和多个非易失性设备相对应的多个逻辑卷的读/写请求，并且管理与多个逻辑卷相对应的多个磁盘设备和多个非易失性设备的存储层和多个磁盘设备和多个非易失性设备的每一个的属性信息之间的关系，其中处理单元基于用户策略将存储层的第一存储层确定为源存储层，以及将存储层的第二存储层确定为目的存储层，并且基于用户策略将存储在第一存储层的至少一个逻辑卷中的数据转移到第二存储层的至少一个逻辑卷。



CN 101676857 B

1. 一种存储系统，包括：

多个硬盘设备，存储从主机计算机接收的数据；

多个非易失性设备，存储从主机计算机接收的数据；以及

处理单元，控制从主机计算机到与所述多个硬盘设备和所述多个非易失性设备相对应的多个逻辑卷的读/写请求，并且管理与所述多个逻辑卷相对应的所述多个硬盘设备和所述多个非易失性设备的存储层与所述多个硬盘设备和所述多个非易失性设备的每一个的属性信息之间的关系，

其中处理单元基于用户策略将所述存储层的第一存储层确定为源存储层，以及将所述存储层的第二存储层确定为目的存储层，并且基于用户策略将存储在所述第一存储层的至少一个逻辑卷中的数据转移到所述第二存储层的至少一个逻辑卷。

2. 根据权利要求 1 所述的存储系统，其中，

用户策略是用于优化存储系统的性能的策略。

3. 根据权利要求 1 所述的存储系统，其中，

用户策略是从主机计算机到与所述多个硬盘设备或所述多个非易失性设备相关的至少一个逻辑卷的实时访问模式。

4. 根据权利要求 1 所述的存储系统，其中，

处理单元管理使用状态管理表，所述使用状态管理表存储与所述多个非易失性设备的每一个的耐久性相关的使用状态信息以及耐久性阈值，

处理单元管理读/写周期管理表，所述读/写周期管理表存储所述多个逻辑卷的每一个的读周期和写周期，

如果处理单元请求优化存储系统的性能，处理单元基于所述使用状态管理表和所述读/写周期管理表来将所述第一存储层确定为所述源存储层，以及将所述第二存储层确定为所述目的存储层。

5. 根据权利要求 4 所述的存储系统，其中，

处理单元还管理设备类型管理表，所述设备类型管理表存储至少一个类型的硬盘设备和一个类型的非易失性设备在一定时间周期内功率消耗量的信息，并且基于所述设备类型管理表来计数激活的所述多个硬盘设备的功率消耗和激活的所述多个非易失性设备的功率消耗的总量，

如果处理单元请求优化存储系统的性能，处理单元基于所述使用状态管理表和所述读/写周期管理表来将所述第一存储层确定为所述源存储层，以及将所述第二存储层确定为所述目的存储层，并且基于包括所述非易失性设备的功率消耗和耐久性的条件的优先级将存储在所述第一存储层的至少一个逻辑卷中的数据转移到所述第二存储层的至少一个逻辑卷。

6. 根据权利要求 5 所述的存储系统，其中，

如果处理单元请求将所述功率消耗的总量优化为较高的优先级，处理单元基于所述读/写周期管理表检测其中读周期与写周期的比率超过预定阈值的由硬盘设备配置的第一逻辑卷，并且处理单元将所述硬盘设备确定为所述第一存储层并且将所述非易失性设备确定为所述第二存储层。

7. 根据权利要求 5 所述的存储系统，其中，

如果不存在所述非易失性设备的空闲容量，处理单元通过将数据从某个非易失性设备转移到在功率消耗的总量的较高优先级的条件的范围内的一个硬盘设备来生成所述多个非易失性设备的所述某个非易失性设备的空闲容量。

8. 根据权利要求 5 所述的存储系统，其中，

如果不存在所述非易失性设备的空闲容量，并且不存在基于所述使用状态管理表满足所述非易失性设备的耐久性的条件的非易失性设备，处理单元基于包括所述非易失性设备的功率消耗的耐久性的条件的优先级确定不将数据从所述第一存储层转移到所述第二存储层。

9. 根据权利要求 5 所述的存储系统，其中，

如果处理单元请求将所述非易失性设备的耐久性优化为较高的优先级，如果处理单元基于所述使用状态管理表检测到所述多个非易失性设备的一个的使用状态超过了耐久性阈值，处理单元将所述非易失性设备确定为所述第一存储层并且将所述硬盘设备确定为所述第二存储层。

10. 根据权利要求 5 所述的存储系统，其中，

如果所述多个磁盘单元的类型是 ATA 先进技术附加磁盘，处理单元管理使用状态管理表，所述使用状态管理表存储与所述多个 ATA 先进技术附加磁盘的每一个的耐久性相关的使用状态信息和耐久性阈值。

11. 根据权利要求 5 所述的存储系统，其中，

通过从主机计算机接收读 / 写请求或通过从主机计算机接收指示以将某个逻辑卷指定为 WORM 单写多读，处理单元请求周期地优化所述非易失性设备的功率消耗的总量或耐久性。

12. 根据权利要求 5 所述的存储系统，其中，

每个非易失性设备与硬盘设备相比具有更少的可重写周期，

所述使用状态管理表包括重写周期作为每个非易失性设备的使用状态信息，并且包括所述重写周期的阈值作为耐久性阈值，以及

如果处理单元请求将所述非易失性设备的耐久性优化为较高的优先级，如果处理单元基于所述使用状态管理表检测到所述多个非易失性设备的一个的重写周期的使用状态超过了所述重写周期的阈值，处理单元将所述非易失性设备确定为所述第一存储层以及将所述硬盘设备确定为所述第二存储层。

13. 根据权利要求 5 所述的存储系统，其中，

所述硬盘设备与每个非易失性设备相比具有较低的读取响应速度，

所述使用状态管理表包括读取数据的使用周期作为每个非易失性设备的使用状态信息，并且包括所述读取数据的使用周期的阈值作为耐久性阈值，以及

如果处理单元请求将功率消耗的总量优化为较高的优先级，如果处理单元基于所述使用状态管理表检测到所述多个非易失性设备的一个的读取数据的使用周期的使用状态超过了所述读取数据的使用周期的阈值，处理单元将所述非易失性设备确定为所述第二存储层以及将所述硬盘设备确定为所述第一存储层。

14. 根据权利要求 5 所述的存储系统，其中，

所述使用状态管理表包括故障区域的数目作为每个非易失性设备的使用状态信息，并

且包括所述故障区域的数目的阈值作为耐久性阈值，以及

如果处理单元请求将非易失性设备的耐久性优化为较高的优先级，并且如果处理设备基于所述使用状态管理表检测到所述多个非易失性设备的一个的故障区域的数目的使用状态超过了所述故障区域的数目的阈值，处理单元将所述非易失性设备确定为所述第一存储层，并且将所述硬盘设备确定为所述第二存储层。

15. 根据权利要求 5 所述的存储系统，其中，

如果所述第二存储层是非易失性设备，处理单元将所述硬盘设备的某个区域确定作为临时存储区域，并且将要被转移的数据连续地一部分一部分地复制到所述硬盘设备的所述临时存储区域，并且

当在数据转移操作过程中从主机计算机接收部分数据的读请求时，处理单元将所述第一存储层的硬盘设备中存储的对应数据重新发送到主机计算机，并且异步地处理数据转移以将数据连续地一部分一部分地转移到所述第二存储层的非易失性设备。

存储系统

[0001] 本中请为 2006 年 6 月 29 日递交的、申请号 200610094311.4、发明名称为“存储控制装置、数据管理系统及数据管理方法”的发明专利申请的分案申请。

技术领域

[0002] 本发明涉及用于减少将数据存储在至少一个硬盘设备以及其它存储介质的存储控制装置中的功率消耗的技术。

背景技术

[0003] 近来，在诸如数据中心之类的信息商务的现场，减少存储系统（存储控制装置和硬盘设备等的总称）中的 TCO(Total Cost of Ownership，整体拥有成本）一直被认为更为重要。此外，需要长期且安全地存储数据。例如，依照诸如个人信息保护法之类的法律规定，日本的金融机构和医疗机构必须积累它们的文件数据而不删除。

[0004] 在这种背景基础上，存储系统应高度可靠且大容量。一般，在使用硬盘驱动器（在下文称为 HDD）的具有大容量的存储系统中，功率消耗与存储介质容量成比例增长。具体地，具有大容量存储系统会导致包括电费在内的 TCO 的增长。

[0005] 鉴于上述缺点，当需要积累大量的信息，如有必要，需要长期保存该数据信息时，考虑到对存储的数据的各种要求，通过恰当地将数据存储在一个合适的存储位置，可以有效地减少整个存储系统的 TCO，尤其是由功率消耗量代表的运行成本。

[0006] 另外，近来，作为非易失性存储器，闪存器一直受到关注。一般地，与 HDD 相比，闪存器减小功率消耗至几十分之一，并可以提供高速读取。由于没有 HDD 所需的驱动机构，闪存器可以小型化，并一般具有高的抗故障性。

[0007] 然而，由于用于保存信息的单元的物理特性，闪存器有写入次数的限制。为了应对该限制，通过采用所谓的平均读写 (WARE LEVELING) 调整，其中提供一种控制来保持单元位置和上位设备表示的地址的对应，并使写入各个单元的次数平均，来提高了闪存器的可重写次数。

[0008] 在下文中，作为保持信息的元件而使用的闪存器、将包括用于进行平均读写调整和对于上位设备的协议处理的机构的闪存器称为“闪存器设备”。

[0009] 尽管在元件级的存储区域的分布方面，上述技术在对闪存器设备的写入次数的限制方面提高了效率，但是，闪存器设备的写入次数的限制依然存在。闪存器有另一缺点，即当需要在写入新数据前删除数据时，写入速度降至与 HDD 相同的速度。

[0010] 为了将具有上述特性的闪存器应用于存储系统，作为传统技术中对所存储的数据的各种要求，从写入性能以及可靠性或成本的观点出发，在专利文献 1（见以下专利文献 1）中公开了在适当的存储位置存储数据的技术。

[0011] 为了实现较低功率消耗的存储系统，介绍了与 MAID(Massive Array ofIdleDisks，大规模非活动磁盘阵列）相关的技术，如在美国专利申请公开号为 2004/0054939 和专利文献 2（见以下的专利文献 2）中所公开的技术。

[0012] 然而，因为专利文献 1 中公开的技术未考虑到各个存储介质间的重写次数、写入速度和功率消耗的差异，所以对整个存储系统进行适当的控制存在很多困难。在美国专利申请公开号为 2004/0054939 和专利文献 2 中的存储系统中，MAID 的可应用目的地受到限制，因此，在更低系统功率消耗和系统的高性能的维持两者间保持折衷存在很多困难。

[0013] 专 利 文 献 1 :John Wilkes, Richard Golding, Carl Staelin, and Tim Sullivan, " The HP AutoRAID hierarchical storage system ", Transactions on ComputerSystems(TOCS)archive, America, ACM(Association for Computing Machinery), February 1996, Volume 14, Issue 1, ISSN :0734-2071, Pages :108-136

[0014] 专 利 文 献 2 :Dennis Colarelli, Dirk Grunwald, and Michael Neufeld, " TheCase for Massive Arrays of Idle Disks(MAID)", [online], January 7, 2002, USENIX(U.S.A.), (Searched on August 5, 2005)

[0015] <URL:http://www.usenix.org/publications/library/proceedings/fast02/wips/colarelli.pdf>

发明内容

[0016] 为了解决上述问题，本发明的目的在于，在存储系统的低功率消耗和高性能的维持之间实现成功的平衡。

[0017] 本发明涉及一种存储系统。该存储系统包括 :多个磁盘设备，存储从主机计算机接收的数据 ;

[0018] 多个非易失性设备，存储从主机计算机接收的数据 ;以及

[0019] 处理单元，控制从主机计算机到与多个磁盘设备和多个非易失性设备相对应的多个逻辑卷的读 / 写请求，并且管理与多个逻辑卷相对应的多个磁盘设备和多个非易失性设备的存储层和多个磁盘设备和多个非易失性设备的每一个的属性信息之间的关系，

[0020] 其中处理单元基于用户策略将存储层的第一存储层确定为源存储层，以及将存储层的第二存储层确定为目的存储层，并且基于用户策略将存储在第一存储层的至少一个逻辑卷中的数据转移到第二存储层的至少一个逻辑卷。

[0021] 在结合附图阅读下面的说明书和权利要求的基础上，本发明的其它方面、特征以及优点将会变得更为明白。

附图说明

[0022] 图 1 是表示根据包括存储系统 S 的本发明的一个实施方式的整体构成的框图。

[0023] 图 2 是表示通道控制单元 104 的结构的框图。

[0024] 图 3 是表示 FM 控制单元 106 的结构的框图。

[0025] 图 4 是表示 FM 控制单元 106 的另一结构的框图。

[0026] 图 5 是表示 FM 控制单元 106 的另一结构的框图。

[0027] 图 6 是表示存在多个存储控制装置 101 的整体构成的一个说明例中的信息流的图。

[0028] 图 7 是表示逻辑卷管理表 700 的一个例子的图。

[0029] 图 8 是表示使用状态管理表 800 的一个例子的图。

[0030] 图 9 是表示读 / 写次数管理表 900 的一个例子的图。

[0031] 图 10 是表示设备类型管理表 1000 的一个例子的图。

[0032] 图 11A 是表示 MP 单元 220 在 HDD110 和 FM 控制单元 106 的 FM 306 之间进行数据转移的判定处理的一个例子的流程图。

[0033] 图 11B 是表示用于在 HDD110 和 FM 控制单元 106 中的 FM 306 之间进行数据转移的判定处理的一个例子的流程图。

[0034] 图 12 是表示当在数据转移操作期间从主机 102 收到读请求时，MP 单元 220 的处理的一个例子的流程图。

[0035] 图 13 时表示当在数据转移操作期间，从主机 102 接收到写请求时，MP 单元 220 的处理一个例子的流程图。

[0036] 图 14 是表示在通过使用图 6 中所示的管理终端 601，对整个存储系统 S 进行功率消耗管理的情况下所使用的整体信息表 1400 的一个例子的图。

具体实施方式

[0037] 下面，参照附图，对如下根据本发明的一个实施方式的存储系统 S 进行详细说明。

[0038] 图 1 是表示包括一个存储系统的本发明的一个实施方式的整体结构的轮廓的框图。存储系统 S 包括存储控制装置 101 和 HDD（硬盘驱动器）110。存储控制装置 101 通过通道 114，经过由 SAN 开关等构成的 SAN(Storage AreaNetwork 存储区域网络）103，连接到一个或多个主机 102（图中为两个）。

[0039] 存储控制装置 101 还经由磁盘侧通道 111 与存储数据的多个 HDD 10 相连。存储控制装置 101 包括 :多个通道控制单元（连接单元）104，多个高速缓冲存储器（存储器单元）113，控制信息存储区（存储器单元）117，多个磁盘控制单元（连接单元）105，多个 FM 控制单元 106，以及经内部通路 115 将上述组件互相连接的相互接合网络 107。

[0040] 通道控制单元 104，通过通道 114 从主机 102 接收 I/O 请求，并解释该 I/O 请求的请求类型（例如，用于将数据读写到 HDD 110 上的读 / 写请求）或其目标地址，以便对该请求进行适当的处理。

[0041] 高速缓冲存储器 113 临时存储将要存储在 HDD110 和 FM 控制单元 106 内的闪存器 306（非易失性存储介质 :后述）上的数据，或将要发送至主机 102 的数据。控制信息存储区 117 是在存储系统 S 上存储控制信息的区域，其包括存储器如 RAM(Random Access Memory，随机存储器）。

[0042] 通过根据要存储的信息类型准备物理上不同的存储器，或通过向其分配逻辑上不同的区域，高速缓冲存储器 113 和控制信息存储区 117 可以独立使用。

[0043] 磁盘控制单元 105，响应来自通道控制单元 104 的请求，通过磁盘侧通道 111 对 HDD 110 进行控制，实施由主机 102 请求的数据获取或数据存储。此时，磁盘控制单元 105 根据 RAID(Redundant Arrays of Inexpensive Disks(廉价磁盘冗余阵列）:管理多个 HDD 的技术）对 HDD 100 进行控制，以便提高存储系统 S 的可靠性、可用性和性能。

[0044] FM 控制单元 106 对闪存器 306（在图 3 中进一步详述）或存储系统 S 内的闪存器设备（图 4 和图 5 中进一步详述）进行控制。FM 控制单元 106 响应来自通道控制单元 104 等的请求，在闪存器 306 或闪存器设备中，进行由主机 102 请求的数据获取或数据存储。此

时，FM 控制单元 106 可对闪存器设备进行 RAID 控制，以便提高存储系统 S 的可靠性、可用性和性能。

[0045] 接着，参照图 2，对通道控制单元 104 的结构进行说明（如有必要，参照图 1）。图 2 是表示通道控制单元 104 的结构的框图。通道控制单元 104 包括 :MP（微处理器）单元（处理单元）220、多个通道协议处理单元 204 和内部网络接口单元 205。MP 单元 220 包括多个处理器 201、存储器模块 202 和外围处理单元 203。

[0046] 每个处理器 201 通过诸如总线之类的连接介质与外围处理单元 203 相连。外围处理单元 203 与存储器模块 202 相连，以便对存储器模块 202 进行控制。外围处理单元 203 还通过通信系统的总线 215 与通道协议处理单元 204 以及内部网络接口单元 205 相连。

[0047] 外围处理单元 203，从与其相连的处理器 201、通道协议处理单元 204 以及内部网络接口单元 205 接收信息包（数据）。若接收到的信息包所指示的传输目的地地址存储在存储器模块 202 中，则进行对应于该地址的适当的处理，如有必要，将数据返回到内部网络接口单元 205。若传输目的地地址是存储器模块 202 之外的位置地址，则外围处理单元 203 进行数据转发（数据传输）。通过诸如 LAN(Local Area Network) 之类的内部通信网络 221，外围处理单元 203 与存储控制装置 101 的另一单元，诸如磁盘控制单元 105 相连（图 6 中进一步详述）。

[0048] 存储器模块 202 具有邮箱 213，用于与外围处理单元 203 相连的处理器 201 之间的通信。处理器 201 通过外围处理单元 203 访问存储器模块 202，根据存储在存储器模块 202 中的控制程序 212 进行适当的处理。

[0049] 存储器模块 202 存储着当进行 DMA(Direct Memory access（直接内存访问）:不通过处理器 201 进行数据传输的技术）时通道协议处理单元 204 使用的传输列表 214。通道协议处理单元 204 对通道 114 进行协议控制，将来自主机 102 的数据转换为协议格式，从而可以在存储系统 S 内对其进行处理。具体地，当通过通道 114 从主机 102 接收 I/O 请求时，通道协议处理单元 204 向处理器 201 通知主机号码、LUN（逻辑单元号码 :HDD 110 的逻辑单元的识别符，与图 7 中逻辑卷 ID 701 对应的存储区）或用于 I/O 请求的访问目的地地址。

[0050] 响应来自通道协议处理单元 204 的通知，处理器 201 访问控制信息存储区 117 上的目录信息、或已在存储器模块 202 上复制并展开的目录信息。如果目录信息中存在 I/O 请求的数据必须被存储到的地址，或 I/O 请求的数据本身，则处理器 201 在存储器模块 202 中创建传输列表 214。根据传输列表 214，通道协议处理单元 204 进行数据传输。

[0051] 如果读请求的数据不存在于高速缓冲存储器 113 上，而是存储在 HDD 110 上，则处理器 201 指示磁盘控制单元 105 以将数据存储在高速缓冲存储器 113 中（此操作称为“升级 (staging)”），然后根据传输列表 214 传输数据。

[0052] 如果由主机 102 读请求的数据存储在闪存器上（如闪存器 306 :在图 3 至图 5 中后述），则处理器 201 在传输列表 214 中设置闪存器的地址。传输列表 214 是高速缓冲存储器 113 或闪存器上的地址的列表。

[0053] 当由主机 102 接收到写请求时，通道协议处理单元 204 通过内部网络接口单元 205，经由内部通路 115 将主机 102 请求的数据写入对应于传输列表 214 中的地址的位置。当接收到来自主机 102 的读请求时，通道协议处理单元 204 从传输列表 214 中记载的对应

地址读取数据，然后将数据返回主机 102。

[0054] 内部网络接口单元 205，用作在通道控制单元 104 和另一存储系统 S 之间经由内部通路 115 进行内部通信时的接口。

[0055] 尽管磁盘控制单元 105 和 104 具有大致相同的结构，但是，磁盘控制单元 105，具有与控制程序 212 对应的部分不同的部分，具有用于和 HDD110 进行通信的、与通道协议处理单元 204 对应的部分。

[0056] 需要注意的是，通道 114 和磁盘侧通道 111 可以具有彼此不同的协议。然而，在进行磁盘侧通道 111 上的协议处理，以使该处理可以在存储系统 S 内进行这一点上，磁盘控制单元 105 中与通道协议处理单元 204 对应的部分与通道控制单元 104 中的通道协议处理单元 204 相似。

[0057] 如果在高速缓冲存储器 113 中存在数据，则处理器 201 根据来自通道控制单元 104 的请求或以一定的时间间隔，将高速缓冲存储器 113 中的数据写入 HDD 110。如果高速缓冲存储器 113 中没有数据，则处理器 201 接收来自通道控制单元 104 的指示，然后根据该指示，从 HDD 110 读取数据，并将该数据写入高速缓冲存储器 113。

[0058] 处理器 201 访问存储在控制信息存储区 117 上的目录信息，以便检索由主机 102 请求的数据将被读出或被存储入的高速缓冲存储器 113 的存储地址。

[0059] 当所请求的数据不在高速缓冲存储器 113 中时，或当为了创建空闲区域而将现有的数据存储在 HDD 110 中（该操作被称为“降级 (destaging)”）时，磁盘控制单元 105 通过磁盘侧通道 111 控制 HDD 110。此时，磁盘控制单元 105 可以对 HDD110 组进行 RAID 控制，从而可以提高整个 HDD 110 的可用性和性能。

[0060] 参照图 3，对 FM 控制单元 106 的结构进行说明（如有必要，参照图 1 和图 2）。图 3 是表示 FM 控制单元 106 的结构的框图。FM 控制单元 106 包括：内部网络接口单元 301、进行 DMA 控制的 DMA 控制器 302、非易失性存储器的存储器模块 304、控制存储器模块 304 的存储器控制器 303、作为非易失性存储元件的闪存器 (FM)306、控制 FM 306 的存储器控制器 305。

[0061] 内部网络接口单元 301 用作经由内部通路 115 在 FM 控制单元 106 和另一存储控制装置 101 之间进行内部通信时的接口。

[0062] 存储器模块 304 具有用于在 FM 控制单元 106 中进行 DMA 的传输列表 308。

[0063] FM 控制单元 106 中包含的 DMA 控制器 302，为了在高速缓冲存储器 113 中创建空闲区，例如，当处理来自主机 102 的写请求时，根据通道控制单元 104 的处理器 201 设置的传输列表 214，进行从高速缓冲存储器 113 到 FM 306 的数据传输。

[0064] 存储器控制器 305，根据经内部通路 115 来自通道控制单元 104 的读请求和 DMA 控制器 302 的写请求，对 FM 306 进行控制并存取 (handle) 数据。存储器控制器 305 在其存储区 307 存储关于 FM 306 的使用的信息。存储在存储区 307 的信息被用来创建使用状态管理表 800（在图 8 中后述）。

[0065] 代替使用 FM306，也可以使用其它存储器，比如：铁电体存储器（使用铁电的材料的非易失性存储器：FeRAM(Ferroelectric Random Access Memory，铁电随机存储器）或相变存储器（通过使用非结晶状态的变化（即，相变化）来存储数据的非易失性存储器，比如 OUM(Ovonic Unified Memory，奥弗辛斯基电效应统一存储器））。

[0066] 参照图 4，对图 3 中的 FM 控制单元 106 的另一结构进行说明（见图 1 至图 3）。图 4 是表示 FM 控制单元 106 的另一结构的框图。该 FM 控制单元 106 使用闪存器 (FM) 设备 409。关于与图 3 相同的结构，使用相同的符号，将省略说明。

[0067] FM 设备 409 是一存储系统，其包括：类似于图 3 中的 FM 306 的一个或多个闪存器、一个或多个存储器控制 303、用于和 FM 协议处理单元 407 进行通信的通信装置（图中未示出）。

[0068] 如图 4 所示，FM 设备 409 存在于每个 FM 控制单元 106 中，其经连接器 408 和 FM 控制单元 106 可拆卸地连接。因此，当 FM 设备 409 发生故障时易于更换 FM 设备 409。为了便于更换 FM 设备 409，通道控制单元 104 的处理器 201 可以设置传输列表 214，以使冗余结构在各个 FM 设备 409 间共享。

[0069] 上述结构也使 FM 设备 409 本身可以用具有更大容量的另一设备替换。通过诸如 FFS(Fast File System，快速文件系统）之类的通用协议，与 FM 设备 409 进行通信。因此，FM 协议处理单元 407 对用于与 FM 设备 409 进行通信的格式进行转换，使其在存储控制装置 101 内可用。

[0070] FM 协议处理单元 407 将关于 FM 设备 409 的使用的信息，存储在存储区 410，存储在存储区 410 的信息被用来创建使用状态管理表 800（在后面的图 8 中描述）。

[0071] 下面，参照图 5，对不同于图 3 和图 4 所示的 FM 控制单元 106 的另一结构进行说明（见图 1 至图 4）。图 5 是表示 FM 控制单元 106 的另一结构的框图。MF 侧通道 510 建立与 FM 设备 409 的连接。对于与图 3 或图 4 中使用相同附图标记的相同的结构，将省略其说明。

[0072] FM 协议处理单元 507 与 FM 协议处理单元 407 类似地工作，在其存储区 508 内，存储关于与其连接的多个 FM 设备 409 的使用状态的信息。在存储区 508 中存储的信息用于创建使用状态管理表 800（在后面的图 8 中描述）。

[0073] 这种结构不仅带来图 4 中所述的 FM 控制单元 106 的特征，而且使得更多 FM 设备 409 与 FM 控制单元 106 连接，从而可以实现大容量的存储系统 S。

[0074] 作为图 3 中的 FM306 的一种安装方式，可以将其直接配置在基板上。在这种情况下，不需要连接器 408、FM 协议处理单元 407 和 FM 侧通道 510，从而可以实现更小型的存储系统 S。

[0075] 此外，在图 3 中，每个存储器控制器 305 可以对其 FM306 进行平均读写调整。存储器控制器 305 可以收集有关在访问 FM306 时所发生的擦除错误发生次数和坏块数的统计信息，并将该信息存储在 FM306 的一部分上，或存储在另外提供的另一 FM306 上，其中可以设置任何单元以根据来自处理器的请求发送该统计信息。

[0076] 接着，参照图 6，对存储控制装置 101 以及其它部件中的信息流进行说明（如有必要，见图 1）。图 6 表示存在多个存储控制装置 101 的情况下的整体结构的一说明性例子中的信息流。对于与图 1 中使用相同附图标记的相同的结构，省略其说明。

[0077] 多个（图 6 中是 2 个）存储控制装置 101 经网络 602 与管理终端 601 相连。

[0078] 管理终端 601 可以是普通的服务器，用于从一台或多台存储控制装置 101 收集内部信息，集中管理该内部信息，如：性能和功率消耗量（可以是根据存储控制装置 101 中各部件的基本功率消耗量，由运行中的部件的数量计算而得到的换算值），或是关于故障信息的统计信息（在图 8 至图 10 中进一步详细说明）。

[0079] 具体地，管理终端 601 具有诸如 CPU(Central Processing Unit) 之类的管理终端处理单元（图中未示出）以及诸如硬盘驱动器之类的管理终端存储单元（图中未示出）。管理终端 601，作为收集信息的一个单元，经由一般的通信网络 602，通过各存储控制装置 101 中的管理单元 603，访问存储在存储控制装置 101 中的信息，例如，存储在控制信息存储区 117 中的图 8 中的使用状态管理表 800 等设备内部信息 604，或者，通过管理单元 603 获取该信息 604。管理终端 601，根据从各个存储系统 101 收集到的设备内部信息 604，生成信息 608（例如，图 14 中的整体信息表 1400）。

[0080] 在存储控制系统 101 内，在各通道控制单元 104 中有 MP 单元 220，在各磁盘控制单元 105 中有 MP 1051。还有用于在 MP 单元 220 和 1051 之间、或 MP 单元 220、1051 和管理单元 603 之间进行通信的内部通信网络 221。

[0081] 管理单元 603 可以直接或间接访问由 MP 单元 220 和 1051 收集到的信息，或者在各个高速缓冲存储器 113 等上累积的构成信息和统计信息等设备内部信息 604。

[0082] 管理终端 601，以预定的标准（阈值）比较所收集到的信息，从而可以根据信息的内容进行恰当的操作。下面，对存储控制装置 101 中的控制的一例进行说明后，对管理终端 601 中的动作和处理进行进一步说明。

[0083] 参照图 7，将对 HDD 110（下文中也称为设备）的设备管理表进行说明（如有必要，参照图 1）。图 7 表示设备管理表 700 的一例。具体地，其说明如何通过使用逻辑卷管理表 700 对存储控制装置 101 中的各存储设备（如控制信息存储区 117）进行管理，以及说明作为主机 102 可见的访问对象的一例的逻辑卷（相当于前面所述的逻辑单元）。

[0084] 尽管在此作为一例说明主机 102 对各个块的访问，但是主机 102 对各个文件的访问也可以同样方式考虑，因为在最低存储层次的控制，由包括各种存储介质 (HDD110 和闪存器等）的设备实现。

[0085] 具体地，代替 SAN 103，也可以使用由 NAS(Network Attached Storage（网络连接存储）：与网络直接相连而使用的文件服务器专用计算机）所使用的另一网络，如因特网。

[0086] 当主机 102 对 HDD 进行写操作时，写访问不对作为 HDD 110 的物理识别符的设备 ID 进行，而是对作为逻辑卷的识别符的逻辑卷 ID 进行。

[0087] 如图 7 所示，在逻辑卷管理表 700 中，逻辑卷 ID 701 是逻辑卷的识别符，其与作为存储控制装置 101 中的 HDD110 的逻辑设备的识别符的逻辑设备 ID702 相关联。逻辑设备 ID 702 还与作为虚拟设备的识别符的虚拟设备 ID 703 相关联。

[0088] 虚拟设备 ID 703 与作为实际的（物理的）设备的识别符的设备 ID 704、表示设备内的空间区域的设备范围 705 相关联。设备范围 705 是可以存储一定数据量的区域的管理单元。根据本发明的实施例，该数据量的具体值与本发明的实质无关，因此该数据量没有限制。

[0089] 虚拟设备 ID 703 定义为由一个或多个设备 ID 704 所表示的设备的存储区域的集合体。

[0090] 需要注意，逻辑卷管理表 700 配置在所有处理器 (MP 单元 220、1051) 可以直接或间接查询该管理表 700 的这样的位置（例如，控制信息存储区 117）。

[0091] 接着，参照图 8，对管理各设备的使用状态管理表进行说明。图 8 表示使用状态管理表的一例。

[0092] 使用状态管理表800，对于与其自己的设备ID 801（对应于图7中的设备ID 704）关联的各个设备，管理其与特性、属性、使用状态有关的信息。现在，假设对闪存器等存储介质进行管理，从使用开始的总写入次数802和总擦除次数803、坏块数量804、坏块增长率805以及平均擦除时间806与设备ID801相关。尽管图中未示出，但在使用状态管理表800中，或在另外的专门用于管理阈值的表中，对每一项信息的各阈值进行管理。

[0093] 各个信息及其阈值被用于管理设备。例如，当设备的平均擦除时间806变长时，可以判定该设备（如FM 306）的使用期限即将结束。

[0094] 除所述信息以外，也可以管理擦除操作错误次数（擦除错误次数）等其它信息，来帮助判断该设备的使用期限。

[0095] 需要注意，当初始化设备或改变设备的配置时，有必要从使用开始持续地存储总写入次数802和总擦除次数803。

[0096] 使用状态管理表800也可对除FM 306以外的、应用使用ATA(ATAttachment：AT附加设备）磁盘驱动器等MAID技术的其它设备和存储设备进行管理，其中，使用次数（总写入次数和总擦除次数）、总加速/减速次数和错误产生率是管理设备使用期限必要的。

[0097] 例如，如果HDD 110由SCSI(Small Computer System Interface：小型计算机系统接口）磁盘和ATA磁盘驱动器组成，或由ATA磁盘驱动器单独组成，则磁盘控制单元105通过使用MAID技术可以控制该ATA磁盘驱动器的旋转和停止。

[0098] 可以与磁盘控制单元105相连的不仅是HDD 110，而且还有装备有HDD110和控制单元（未图示）的磁盘阵列设备（未图示）。在这种情况下，通过磁盘控制单元105的指示以及装备在控制单元（未图示）的MAID，可以控制ATA磁盘驱动器的旋转和停止。即，ATA磁盘驱动器在硬件基础上是脆弱的，因此，和管理FM306相同地，最好在使用状态管理表800上在使用期限方面管理ATA磁盘驱动器。

[0099] 作为管理闪存器的一例，对使用状态管理表800进行说明。然而，在ATA磁盘驱动器、或应用了使用ATA磁盘驱动器MAID的驱动器的装置情况下，使用状态管理表800可包括总加速/减速次数、总旋转时间和错误发生率。具体地，使用状态管理表800，根据其状况，具有与存储介质和设备的使用期限有关的各种信息（总使用时间、错误发生率、坏块增长率、访问次数等）。

[0100] 使用状态管理表800，和逻辑卷表700的情况相同，被配置在所有处理器(MP单元220、1051)可以直接或间接对其进行查询的位置。

[0101] 参照图9，对读/写次数管理表900进行说明（如有必要，见图1）。图9表示读/写次数管理表900的一例。

[0102] 读/写次数管理表900与逻辑卷ID 901（对应于逻辑卷ID 701）、逻辑设备ID 902（对应于逻辑设备ID 702）、RD次数（总读取次数）903、和WR次数（总写入次数）904关联。读/写次数管理表900，每当访问逻辑卷ID 901或逻辑设备ID 902时，作为统计信息更新。

[0103] 读/写次数管理表900，和逻辑卷管理表700的情况相同地，设置在所有处理器(MP单元220、1051)能够查询表900的位置。

[0104] 读/写次数管理表900用在具有与写入的频率相比读取的频率更高的大量数据被从每一定时间段的功率消耗高的HDD 110向每一定时间段的功率消耗低的闪存器转移的

处理中（从后面的图 11 到图 13 中描述）。

[0105] 类似地，尽管与读 / 写次数管理表 900 无关，存储在被指定为 WORM(WriteOnce Read Many，单写多读）、只读、或长期的保持期间的逻辑卷中的数据可以被转移（从 HDD 110 到 FM 306）。

[0106] 接着，参照图 10，对设备类型管理表进行说明（如有必要，见图 1）。图 10 表示设备类型管理表的一个例子。

[0107] 设备类型管理表 1000，与表示设备类型的存储层设备类型 1001、每一定时间段的功率消耗量 1002、表示运行中的设备的数量的活动设备数 1003、以及空闲的低功率消耗设备池的数量（低功率消耗的存储介质的空闲容量）1400 关联。

[0108] 在存储层设备类型 1001 中，“Flash”表示 FM 306，“ATA”表示作为在 MAID 技术中使用的 HDD110 的整个或一部分的 ATA 磁盘驱动器、应用 MAID 技术的磁盘阵列设备在磁盘阵列设备与磁盘控制单元 105 相连的情况下所包含的 ATA 磁盘驱动器、或者包含上述磁盘阵列设备所具有的 ATA 磁盘驱动器的逻辑设备（逻辑卷）。“SCSI”表示作为 HDD110 的整体或者一部分的 SCSI 磁盘驱动器。“Cache”表示作为高速缓冲存储器 113 使用的 RAM。

[0109] 空闲低功率消耗设备池的数量 1004 的单位为兆字节 (MB) 和十亿字节 (GB) 等。

[0110] 存储控制装置 101 的总功率消耗量，通过在其电源单元（未图示）设有测量实际功率消耗的设备（例如功率计）来进行测量和记录的方法可以得到。为了去除测量单元如功率计，通过使用设备类型管理表 1000 来计算总功率消耗量的粗略估计值。

[0111] 换言之，因为存储控制装置 101 监视各部件的操作状态，只要将设备类型管理表 1000 中的信息作为统计信息来管理，就可以得到总功率消耗量的粗略估计值。在图 11A 至图 13 中后述的处理中，可以使用设备类型管理表 1000 以便根据各部件的状态及其功率消耗量，将数据转移至较低功率消耗的设备，以便减少各部件的功率消耗。

[0112] 需要注意，设备类型管理表 1000，和逻辑卷表 700 的情况相同地，设置在所有处理器 (MP 单元 220、1051) 能够直接或间接查询该表 1000 的位置。

[0113] 此外，参照图 11A 至图 13，对存储系统的操作进行说明（如有必要，参见图 6 等）。

[0114] 图 11A 是表示 MP 单元 220 在 HDD 110 和 FM 控制单元 106 之间进行数据转移时的判定处理的一个例子的流程图。

[0115] MP 220 执行该数据转移处理时的时序，例如，当从主机 102 接收周期性读 / 写请求时，或当从主机 102 接收特定某逻辑卷为 WORM 的指示时。

[0116] 在图 3 至图 5 中显示了 FM 控制单元 106 的结构的 3 个例子。在此，对图 3 中的例子进行说明。

[0117] MP 单元 220，当执行该数据转移的判定处理时，对所收集到的信息（图 8 至图 10 中的各个表）进行确认 (S111)。

[0118] 接着，MP 单元 220 根据所收集到的信息和判定条件 (n)，对状态进行分析 (S112)。判定条件 (n) 包括，例如 FM 306 的使用期限管理（通过查询总写入次数 802，判定总写入次数是否超过预定值）、功率消耗的减少（查询设备类型管理表 1000)、因来自主机 102 的 WORM 的指示而进行的数据转移、RD/RW 次数比（通过查询 RD 次数 903 和 WR 次数 904，若 RD 次数 903 对 WR 次数 904 的比值超过预定值，则判定将数据转移至 FM 306 的可能性）。这些判定条件 (n) 按照优先顺序彼此结合使用。

[0119] MP 单元 220，根据在步骤 S112 的分析，判定是否存在超过阈值的状态 (S113)。步骤 S113 用于判定 FM 306 的总写入次数是否超过阈值，若其超过阈值，则将其数据转移至没有写入次数限制的 HDD110。

[0120] 如果没有超过阈值的状态（在 S113 中为“NO”），则 MP 单元 220 结束处理。

[0121] 如果存在超过阈值的状态（在 S113 中为“YES”），则 MP220 选择数据转移的源和目的地 (S114)，并通过查询空闲低功率消耗设备池的数量 1004 等，判断是否存在数据转移的目的地 (S115)。

[0122] 若没有数据转移的目的地（在 S115 中为“NO”），则 MP 单元 220 判断是否可能在判定条件中较高优先级的某条件范围内，生成一个数据转移目的地 (S117)。

[0123] 此外，若不可能生成数据转移的目的地（在 S117 中为“NO”），则 MP 单元 220 结束处理。

[0124] 如果可能生成数据转移的目的地（在 S117 中为“YES”），则 MP 单元 220 生成数据转移的目的地 (S118)，并开始数据转移处理 (S119)。

[0125] 具体地，如果试图从 HDD 110 向任何一个 FM 306 转移数据，但在 FM 306 中没有用于数据转移的目的地（即，没有空闲容量），例如，在判定条件 (n) 中优先级高的某条件（例如与功率消耗相关的条件）的范围内，将 FM306 上的部分数据转移到 HDD 110，以便在 FM306 上生成一些可以向其中转移数据的空闲容量。

[0126] 若存在数据转移目的地（在 S115 中的“YES”），则 MP 单元 220 判断是否由于数据转移，状态超过判定条件 (n) 中优先级较高的某条件 (S116)。

[0127] 若状态超过判定条件 (n) 中优先级较高的条件（在 S116 中的“NO”），则 MP 单元 220 结束处理。例如，尽管由于数据转移，FM 306 在其使用期限内没有问题，但是如果不满足优先级较高的条件，则 MP 单元 220 不进行数据转移。

[0128] 如果状态未超过判定条件中优先级较高的条件（在 S116 中的“YES”），则 MP 单元 220 开始数据转移的处理 (S119)。

[0129] 如上所述，在存储系统 S 内，可以进行数据转移以便优化整体功率消耗和 FM306 的使用期限管理。具体地，通过根据设备的写 / 读频率，进行向具有较低功率消耗的 FM 306 的数据转移，从而可以降低总体功率消耗。

[0130] 接着，参照图 11B，对数据转移处理进行说明（如有必要，见图 6 等）。图 11B 是表示用于使 FM 控制单元 106 在 HDD110 和 FM306 之间进行数据转移的判定处理的一个例子的流程图。需要注意，该步骤在图 11A 的步骤 S119 后执行。

[0131] 首先，MP 单元 220 判断目的地区域是否在 FM306 中（步骤 S1101）。

[0132] 如果目的地区域在 FM306 中（在 S1101 中的“YES”），由于 FM 306 对写入次数和性能敏感，所以，在 HDD 110 中确保暂时存储区 (S1102)。与此同时，MP 单元 220 在控制信息存储区 117 中创建暂时存储区管理表。

[0133] 如果目的地区域不在 FM306 中（在 S1101 中的“NO”），则没有必要确保暂时存储区，从而前进到步骤 S1103。

[0134] 然后，MP 单元 220 创建用于以一定单位大小管理数据转移目的地的存储区的管理表（未图示）(S1103)。该管理表，例如，可以是管理 64KB 单位大小的数据的位图表。只要可以对数据转移的进程提供管理，除位图表以外的其它任何表都可以用于该管理表。

[0135] 接着上述步骤，MP 单元 220 通过向 DMA 控制器 302 发送指示，将源数据复制到数据转移的目的地 (S1104)。

[0136] MP 单元 220，根据上述管理表（即，位图表），检查结束了数据转移的区域，例如，通过将位从“0”变更为“1”(S1105)。

[0137] MP 单元 220，根据管理表判断是否已复制了转移对象的全部数据区域，即，数据转移的进程是否已达到了 100％ (S1106)。

[0138] 如果转移对象的整个数据区域还没有复制完，则 MP 单元 220 重复在步骤 S1104 和步骤 S1105 的处理。

[0139] 如果转移对象的整个数据区域的复制进程已结束（在 S1106 中的“YES”），则 MP 单元 220 前进到步骤 S1107。

[0140] 此外，在步骤 S1101 到步骤 S1106 的数据转移处理期间，存储控制装置 101 偶尔接收数据的读和写请求。以后在图 12 和图 13 中对这种情况下的适当处理进行说明。

[0141] 接着，在步骤 S1107，MP 单元 220 查询在步骤 S1102 所创建的暂时存储区域管理表，并检查在暂时存储区域中是否存在存储的数据。

[0142] 如果在暂时存储区域中存在任何存储的数据（在 S1107 中的”YES”），则 MP 单元 220 将暂时存储区域中所存储的数据反映到转移目的地区域 (S1108)，并重复在步骤 S1108 的处理，直到全部数据的转移已结束（在 S1109 中的“YES “）。

[0143] 如果暂时存储区域内没有存储的数据（在 S1107 中的“NO“），则 MP 单元 220 前进到步骤 S1110。

[0144] MP 单元 220，根据数据转移，重写逻辑卷管理表 700(S1110)。在这种情况下，例如，若在表 700 中的设备 ID704 的适当单元，将 ID 名从转移源设备 ID 变更为转移目的地设备 ID，则没必要重写在主机中的虚拟设备（对应于虚拟设备 ID703)。

[0145] MP 单元 220 将在步骤 S1110 中变得不必要的转移源设备释放，以便于以其它目的对其进行再利用 (S1111)。

[0146] 如上所述，在存储系统 S 中，可以根据图 11A 中所述的有关转移的判定结果进行数据转移。

[0147] 另外，还有另一方法，其在步骤 S1104 至 S1106 实现数据转移的实际操作之前，在步骤 S1111 进行逻辑执行。该方法，不仅可以应用于数据转移操作，而且可以应用于数据复制功能中的高速快照访问功能（一般称为卷镜像分隔 (volume mirror-split) 或快照）。在使用该方法的情况下，可以根据各个目的，在后面图 12 和图 13 中所描述的处理（对于读 / 写请求）中进行适当的改变。

[0148] 接着，参照图 12，对数据转移操作期间从主机 102 接收到读请求时 MP 单元 220 的处理进行说明（图 11B 中的 S1101 至 S1106)（如有需要，见图 6)。图 12 是表示这种情况下 MP 单元 220 的处理的一个例子的流程图。

[0149] 首先，MP 单元 220，从主机 102 接收数据转移操作期间对数据的读请求 (S1201)，然后，在管理表（在图 11 的步骤 S1103 中创建）上，对最新的数据存储位置（在转移源设备或暂时存储区域中）进行确认 (S1202)。

[0150] 接着，MP 单元 220 通过从其适当的存储位置发送数据，向主机 102 返回读响应 (S1203)。

[0151] 然后，如果读对象的数据转移还未结束，MP 单元 220 将转移源数据复制到转移目的地 (S1204)。

[0152] MP 单元 220 将执行了数据转移操作的区域内的操作的完成反映到管理表上 (S1205)。

[0153] 需要注意，步骤 S1203 和步骤 S1204 的处理，可以在步骤 S1202 的处理前进行。

[0154] 如上所述，存储系统 S，即使在数据转移进行中，也可以满足来自主机 102 的读请求。

[0155] 参照图 13，对当在数据转移操作期间从主机 102 接收到写请求时 MP 单元 220 的处理进行说明（图 11B 中的 S1101 至 S1106)（如有必要，见图 6 等）。图 13 是表示在这种情况下 MP 单元 220 的处理的一个例子的流程图。

[0156] 首先，MP 单元 220 从主机 102 接收对转移操作期间的数据的写请求 (S1301)，然后，将对应于其访问地址而编址的、被写请求的数据写入暂时存储区域 (S1302)。

[0157] 然后，MP 单元 220，对在图 11B 中的步骤 S1102 创建的暂时存储区域表进行更新，以便反映最新的数据位于暂时存储区域内 (S1303)。

[0158] 如上所述，即使在进行数据转移时，存储系统 S 也可以满足来自主机 102 的写请求。

[0159] 此外，被主机 102 写请求的数据暂时存储在暂时存储区中，然后在后面被反映，使得即使转移目的地是像 FM306 那样的通过诸如追加、指针替换和块擦除之类的复杂的过程来实施写入的设备，或是像使用 ATA 磁盘驱动器来仅处理用于重复加速和减速的磁盘设备的设备，也可以实现平稳的数据转移操作。

[0160] 尽管描述 MP 单元 220 进行图 11A 至图 13 中的各个处理，但是，其它功能或装置也可用于各处理，如由 MP 单元 220 和 MP 单元 1051 所指示的 DMA 功能。

[0161] 在此，参照图 14，对本方面的另一实施例进行说明（如有必要，见图 6 等）。图 14 表示在图 6 中所描述的管理终端 601 对整个存储系统 S 提供功率消耗管理的情况下，使用的整体信息表 1400 的一个例子。

[0162] 在整体信息表 1400 中，作为存储控制装置 101 的识别符的控制设备 ID1401、各个存储控制装置 101 中的每一定时间段的功率消耗量 1402、表示诸如 FM306 之类的运行中的低功率消耗介质的数量的活动设备 1403、表示运行中的低功率消耗介质的总容量的活动设备总容量 1404、表示诸如 HDD110 之类的普通功率消耗介质的数量的活动设备 1405、表示运行中的普通功率消耗介质的总容量的活动设备总容量 1406、表示空闲低功率消耗设备的数量的空闲低功率消耗设备池 1407、各自关联。

[0163] 即使在某存储控制装置 101 中，不可能将数据从具有普通功率消耗的设备，转移到具有低功率消耗的设备，该整体信息表 1400 使数据可以从该普通功率消耗设备转移至另一存储控制装置 101 中的低功率消耗设备，由此，可以减少整个存储系统 S 的功率消耗。

[0164] 需要注意，在整体信息表 1400 或专用于阈值管理的不同表（未图示）中，对整个存储系统 S 的功率消耗的阈值进行存储和管理。

[0165] 具体的数据转移处理，与根据图 11A 至图 13 中的流程图的处理相似。例如，图 11B 中所示的同样的单个存储控制装置 101 内的设备间的数据转移处理，可以在不同存储控制装置 101 中的设备间进行。然后，主机 102 可以识别访问目的地设备的变更。或者，可以由

主机 102 上的卷虚拟化软件或虚拟化开关实现，以识别访问目的地设备的变化。

[0166] 根据本发明的存储系统 S 可实现这样的系统，该系统具有低功率消耗，没有关于必要的数据的性能的恶化，可大规模构成，并能够将数据存储在最佳的存储介质中。对于每个存储介质具有的特性（例如重写次数的使用期限和容错性），存储系统 S 还可以提高整个系统的可靠性和可用性。

[0167] 上面对根据本发明的实施例进行了说明，但是本发明的实施例不限于这些说明，本领域技术人员明确本发明的必要特征，在不超出权利要求的主旨和范围的情况下，可以对本发明进行多种修改和变更，来使其适用于各种用途和条件。

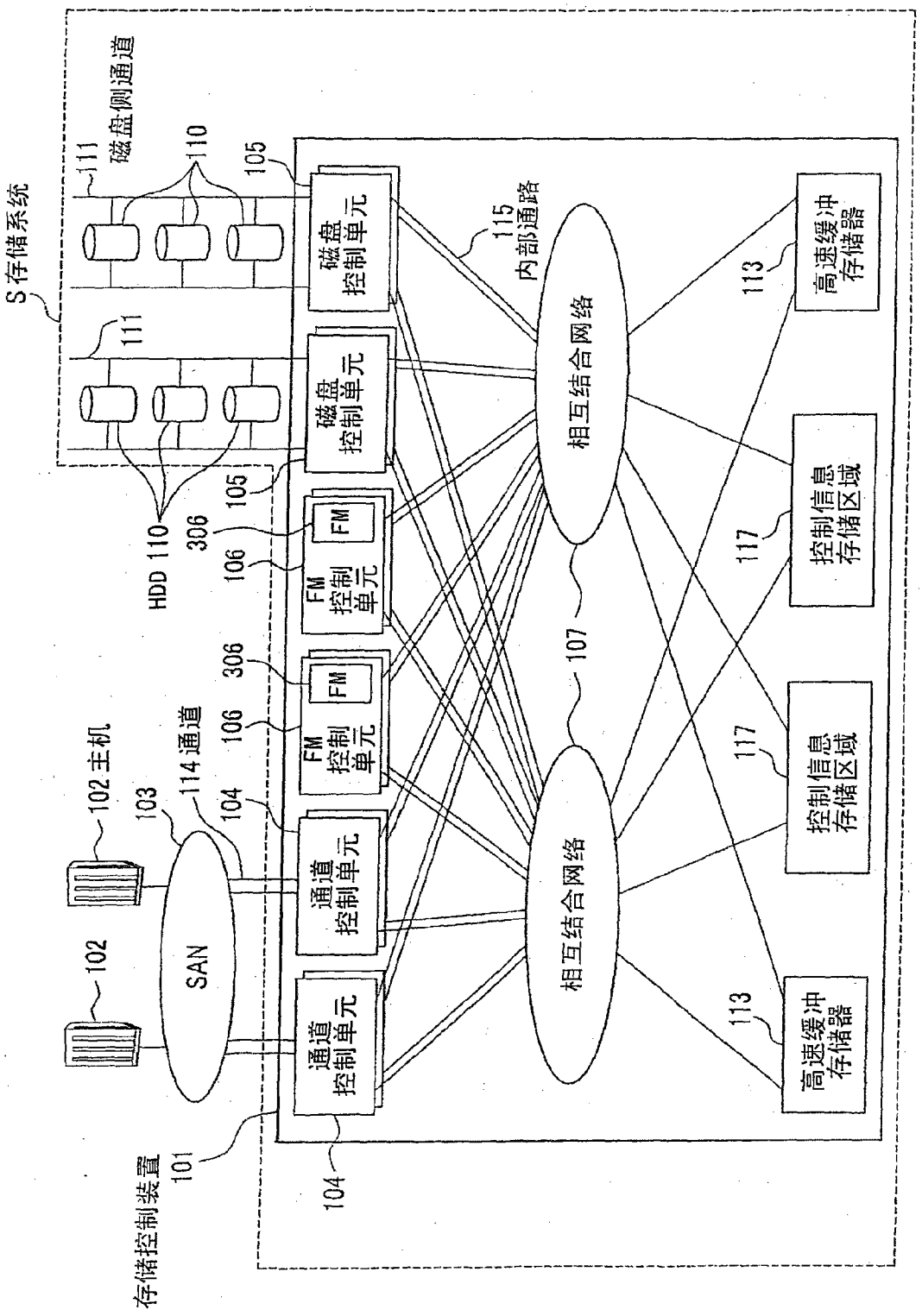

S 存储系统
磁盘侧通道
111
110
105
磁盘控制单元
磁盘控制单元
HDD 110
306
FM
FM控制单元
106
FM控制单元
115
内部通路
相互结合网络
107
相互结合网络
113
高速缓冲存储器
117
控制信息存储区域
控制信息存储区域
高速缓冲存储器
102 主机
103
114 通道
104
通道控制单元
通道控制单元
SAN
102
存储控制装置
101
104

图 1

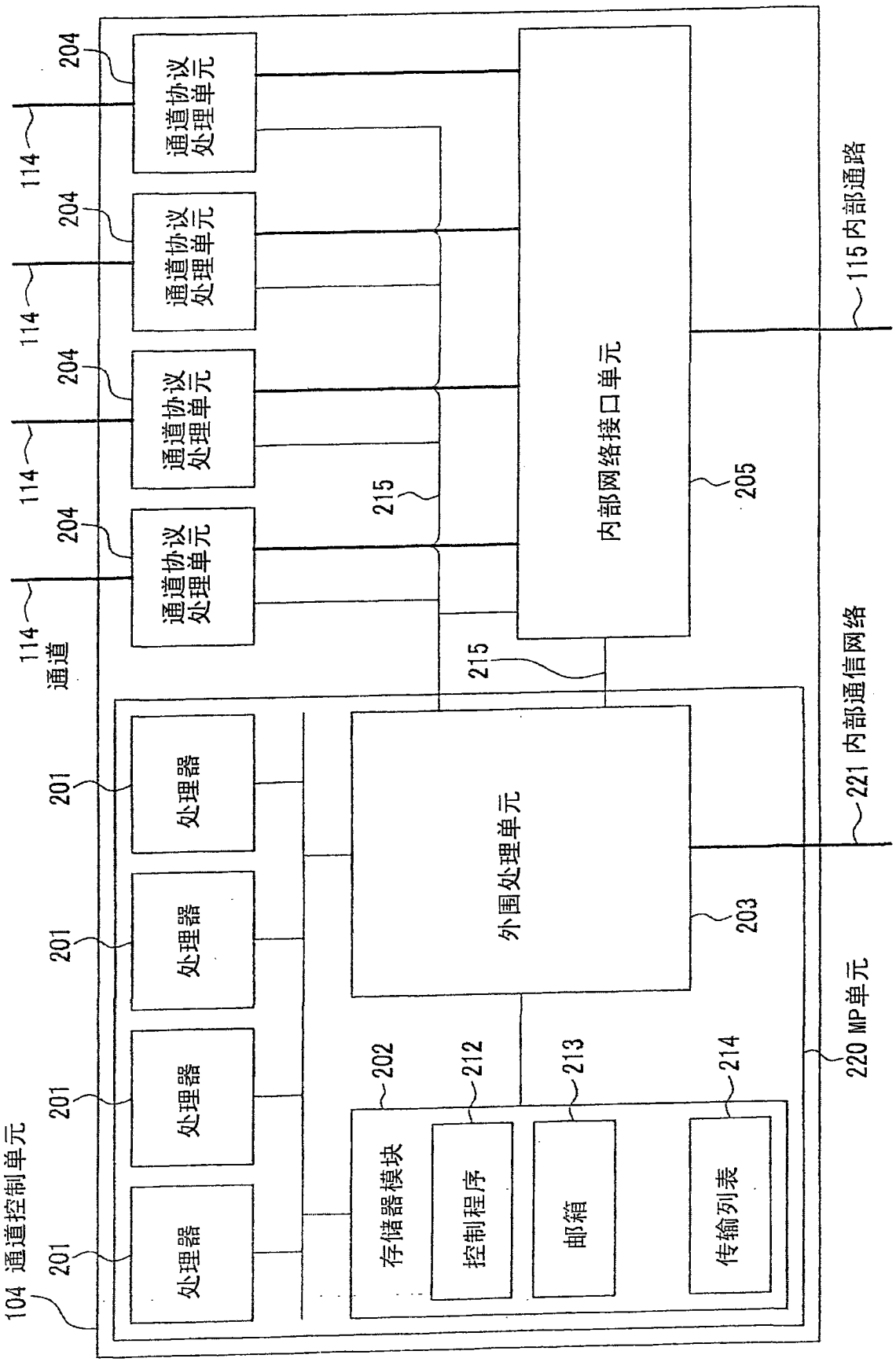

104 通道控制单元
201
处理器
处理器
处理器
处理器
202
存储器模块
212
控制程序
213
邮箱
214
传输列表
203
外围处理单元
215
205
内部网络接口单元
115 内部通路
221 内部通信网络
220 MP单元
114
通道
204
通道协议处理单元
通道协议处理单元
通道协议处理单元
通道协议处理单元

图 2

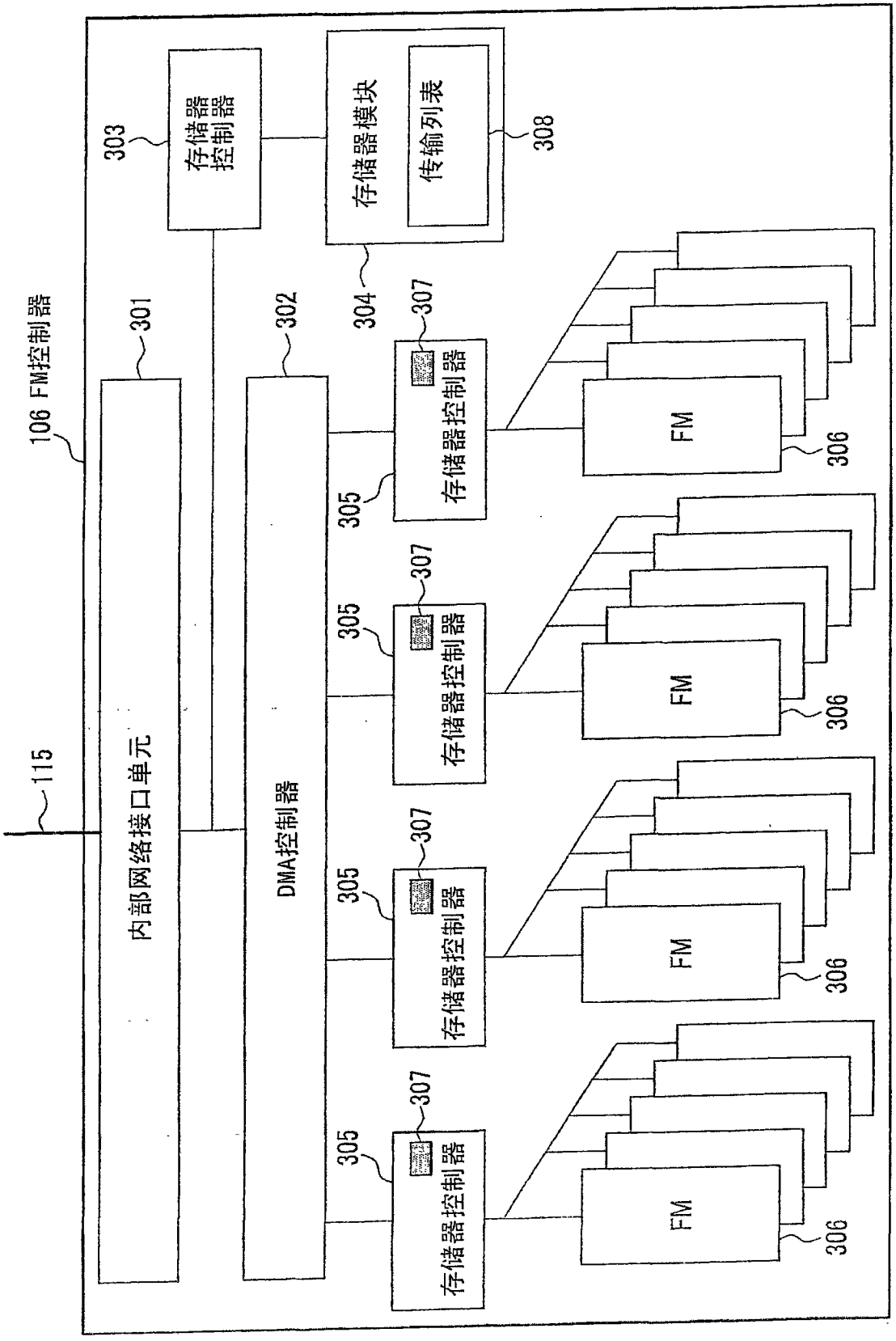

106 FM控制器
115
内部网络接口单元
301
DMA控制器
302
303
存储器控制器
304
存储器模块
传输列表
308
305
存储器控制器
307
FM
306

图 3

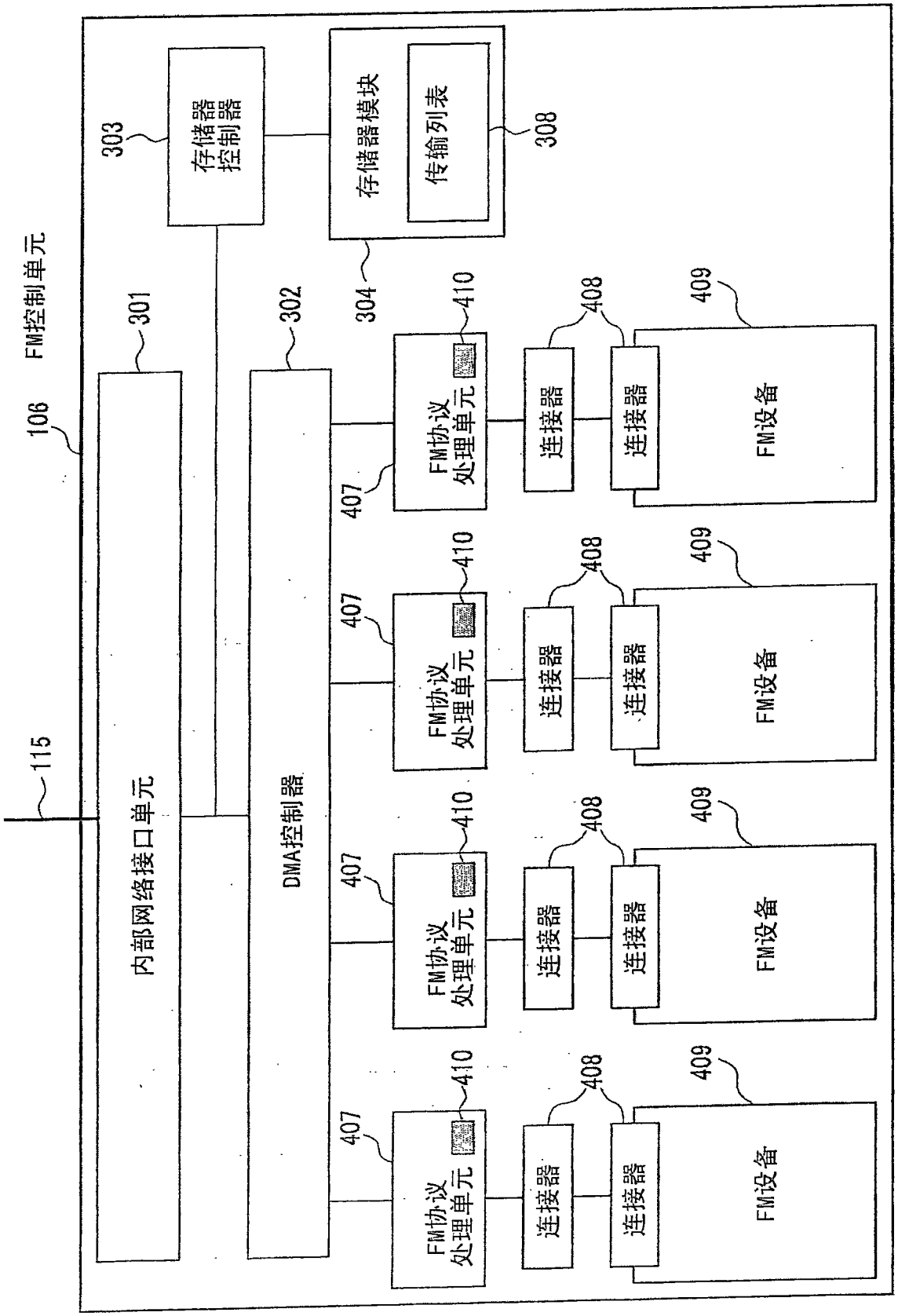

FM控制单元
106
115
内部网络接口单元
301
DMA控制器
302
存储器控制器
303
存储器模块
304
传输列表
308
FM协议处理单元
407
410
连接器
408
FM设备
409

图 4

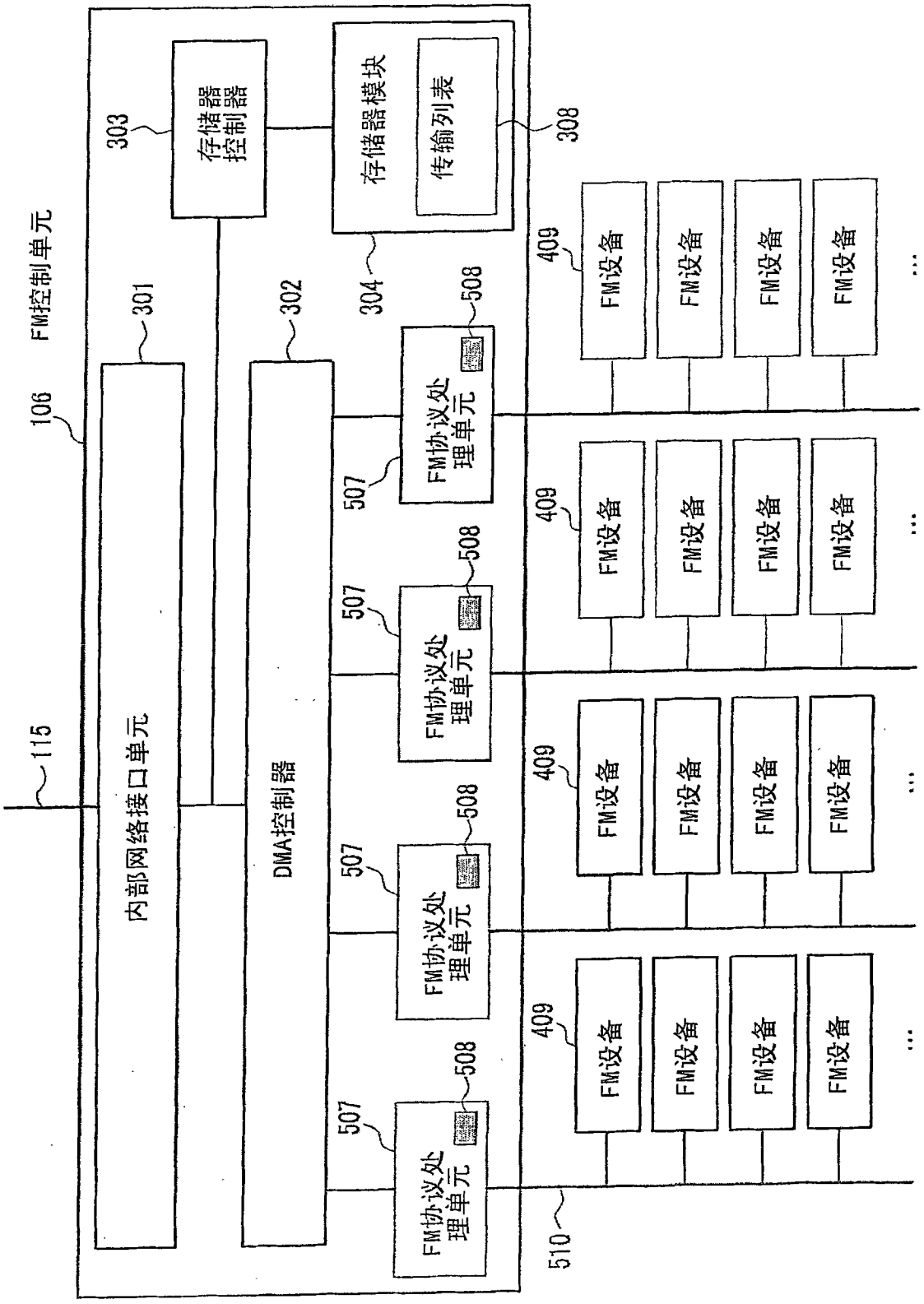

FM控制单元
106
115
内部网络接口单元
301
303
存储器控制器
DMA控制器
302
存储器模块
传输列表
304
308
FM协议处理单元
507
508
409
FM设备
510

图 5

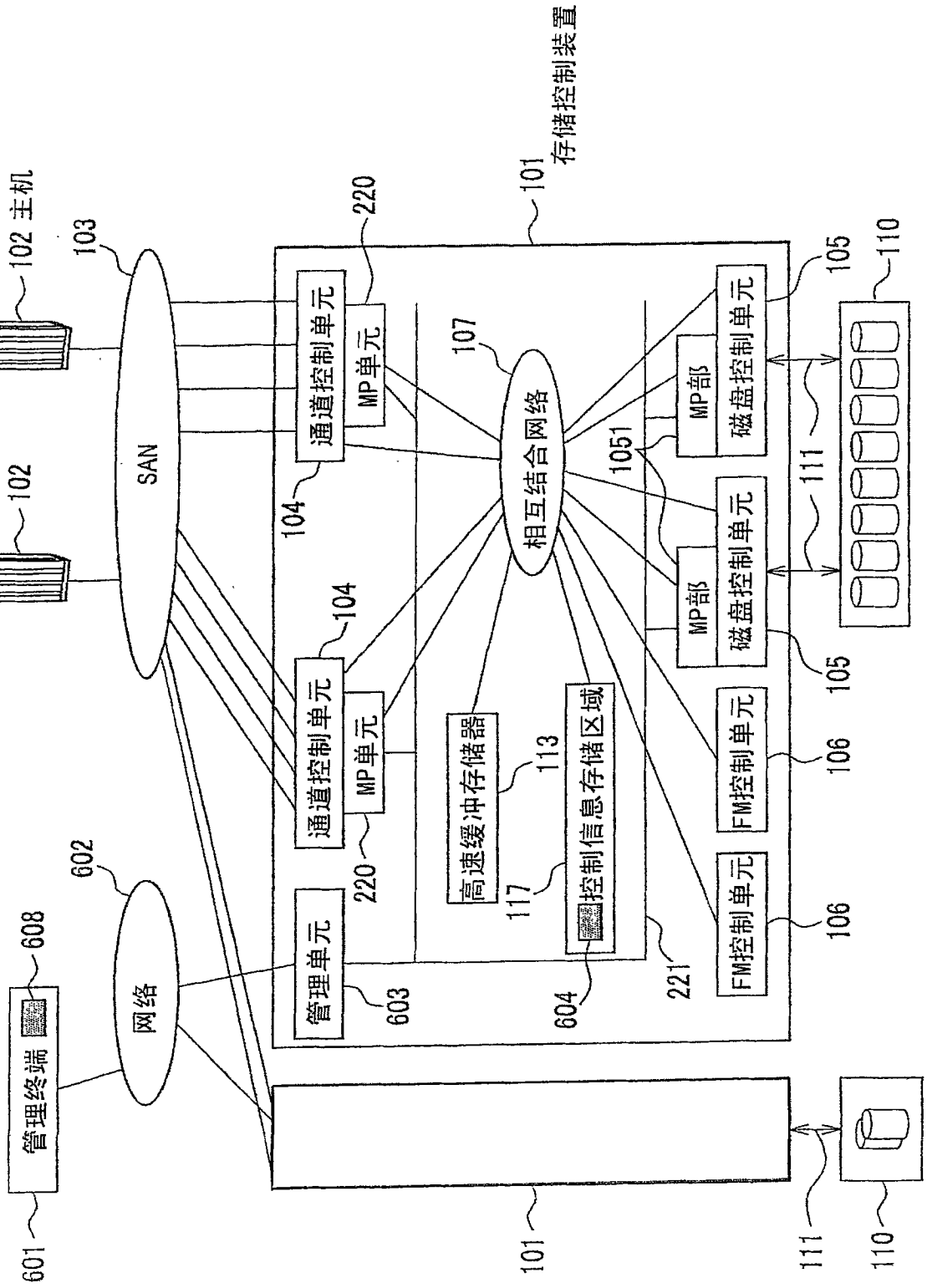

存储控制装置
101
102 主机
102
103
SAN
220
通道控制单元
MP单元
104
107
相互结合网络
1051
MP部
磁盘控制单元
105
110
111
高速缓冲存储器
113
117
控制信息存储区域
FM控制单元
106
602
网络
608
管理终端
601
管理单元
603
604
221

图 6

700 逻辑卷管理表

#	逻辑卷ID	逻辑设备ID	虚拟设备ID	设备ID	设备范围
0	2	001	001	1001	0h,8FFFFh
1	3	002	005	1003	0h,4FFFFh
2	3	002	004	1004	0h,3FFFFh
3	3	002	003	1005	0h,5FFFFh
4	4	004	002	004	0h,3FFFFh
5	5	006	009	004	40000h,8FFFFh
6	5	006	008	005	0h,5FFFFh
7	5	006	007	007	0h,6FFFFh
●	●	●	●	●	●
●	●	●	●	●	●
●	●	●	●	●	●

701 702 703 704 705

图 7

800 使用状态管理表

#	设备ID	总写入次数	总擦除次数	坏块数	坏块增长率	平均擦除时间
0	1000	4000	2000	2	0.001	3ms
1	101	16000	2000	2	0.001	3ms
2	102	30000	3000	2	0.0005	3ms
3	103	3000	1000	1	0.001	3ms
4	104	8000	2000	2	0.001	3ms
5	1001	10000	20000	10	0.007	5ms
6	1006	5000	1000	1	0.001	3ms
●	●	●	●	●	●	●
●	●	●	●	●	●	●
●	●	●	●	●	●	●

801 802 803 804 805 806

图 8

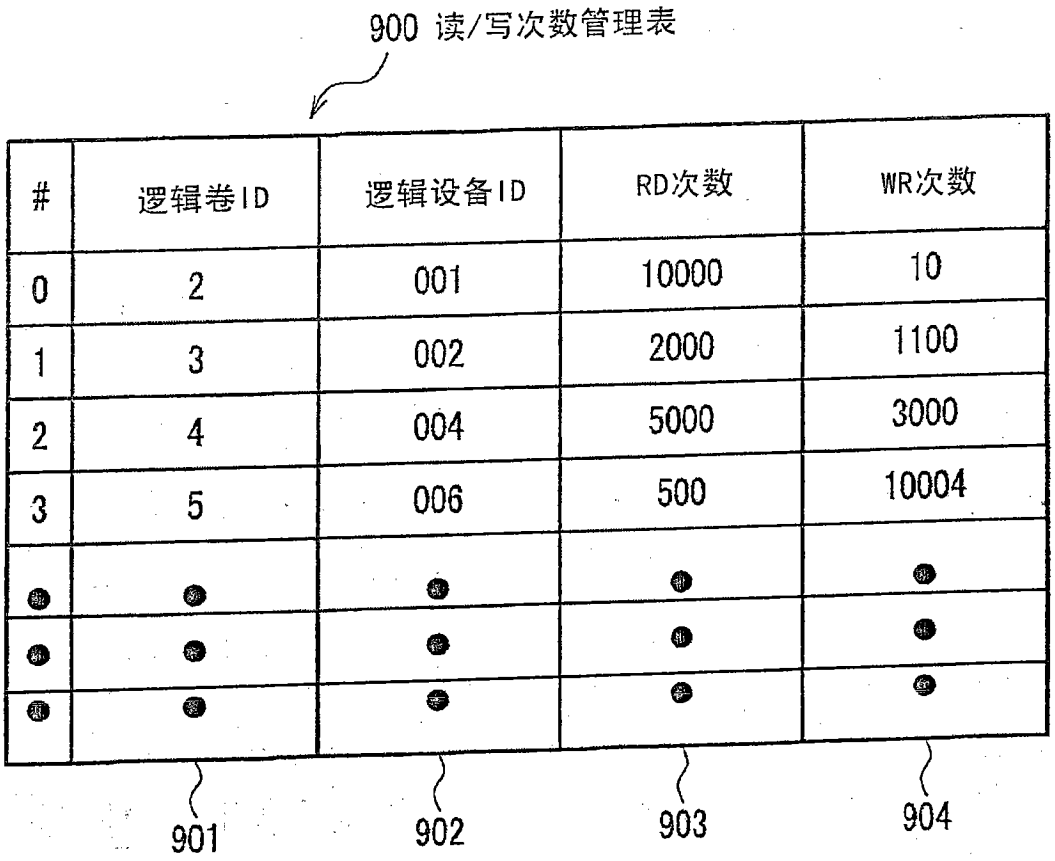

900 读/写次数管理表
逻辑卷ID 逻辑设备ID RD次数 WR次数
0 2 001 10000 10
1 3 002 2000 1100
2 4 004 5000 3000
3 5 006 500 10004
901
902
903
904

图 9

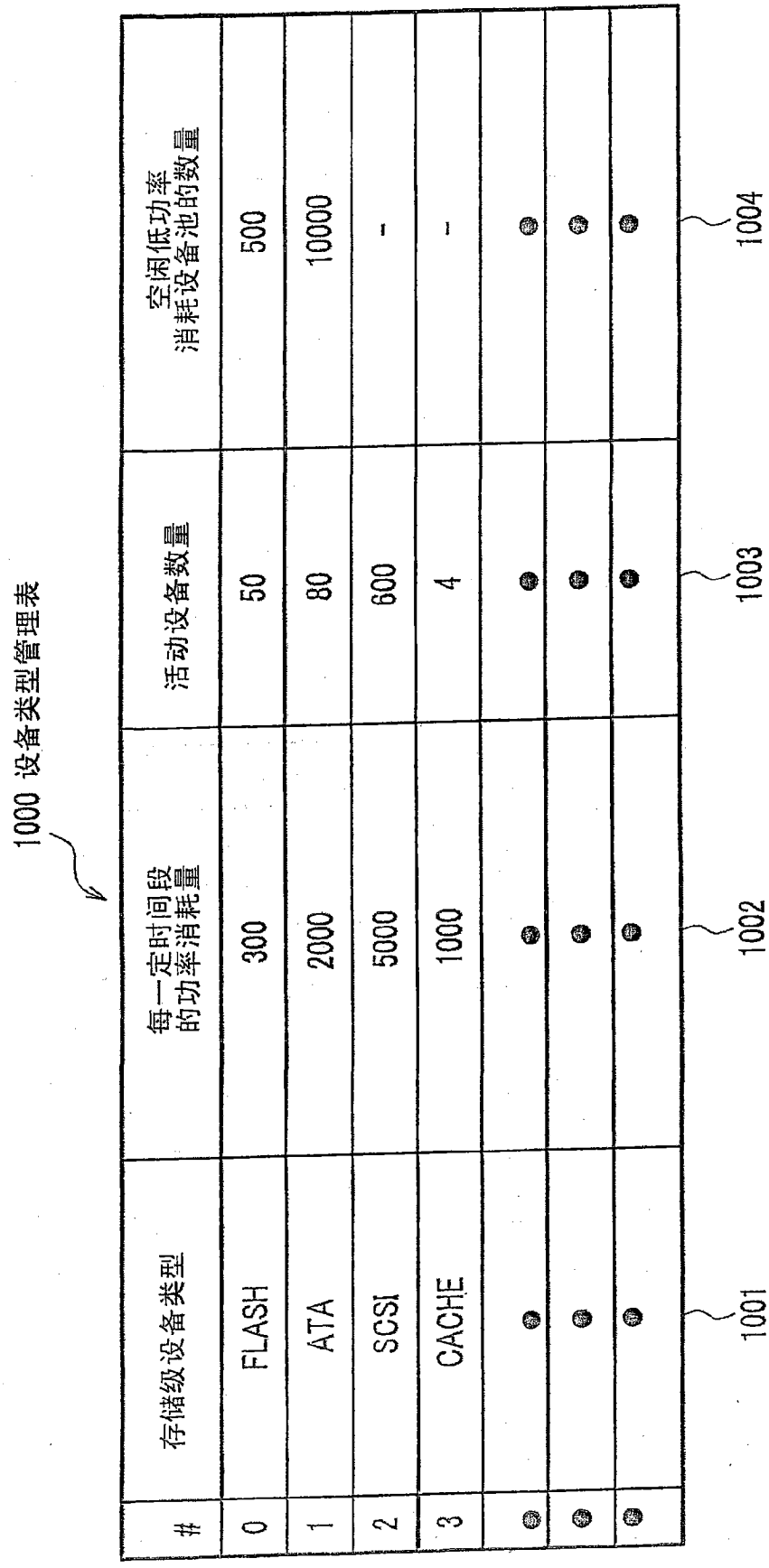

1000 设备类型管理表
#
存储级设备类型
每一定时间段的功率消耗量
活动设备数量
空闲低功率消耗设备池的数量
0
FLASH
300
50
500
1
ATA
2000
80
10000
2
SCSI
5000
600
-
3
CACHE
1000
4
-
1001
1002
1003
1004

图 10

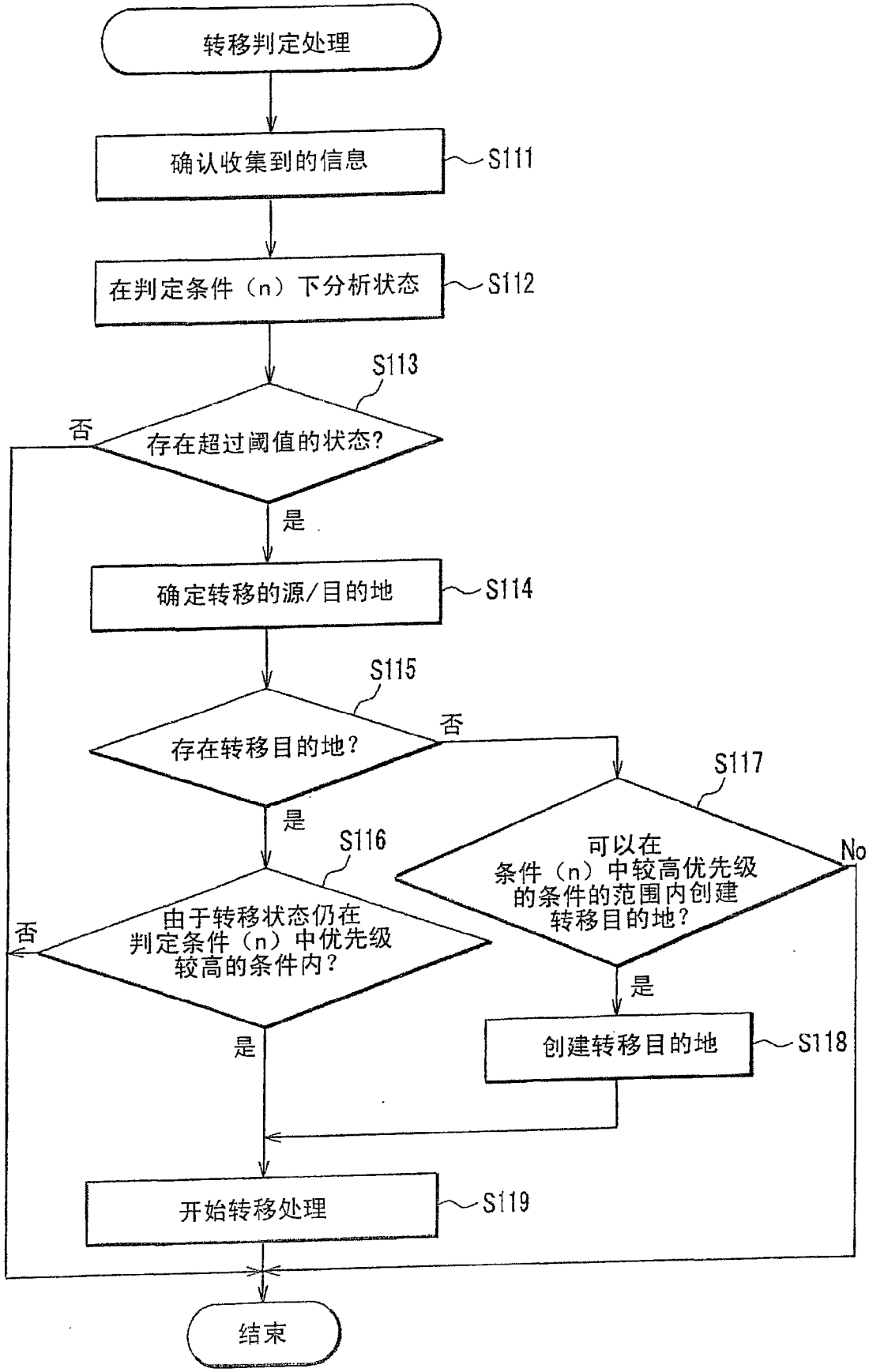

转移判定处理
确认收集到的信息
S111
在判定条件（n）下分析状态
S112
S113
存在超过阈值的状态?
否
是
确定转移的源/目的地
S114
S115
存在转移目的地?
否
是
S116
由于转移状态仍在判定条件（n）中优先级较高的条件内?
否
是
S117
可以在条件（n）中较高优先级的条件的范围内创建转移目的地?
No
是
创建转移目的地
S118
开始转移处理
S119
结束

图 11A

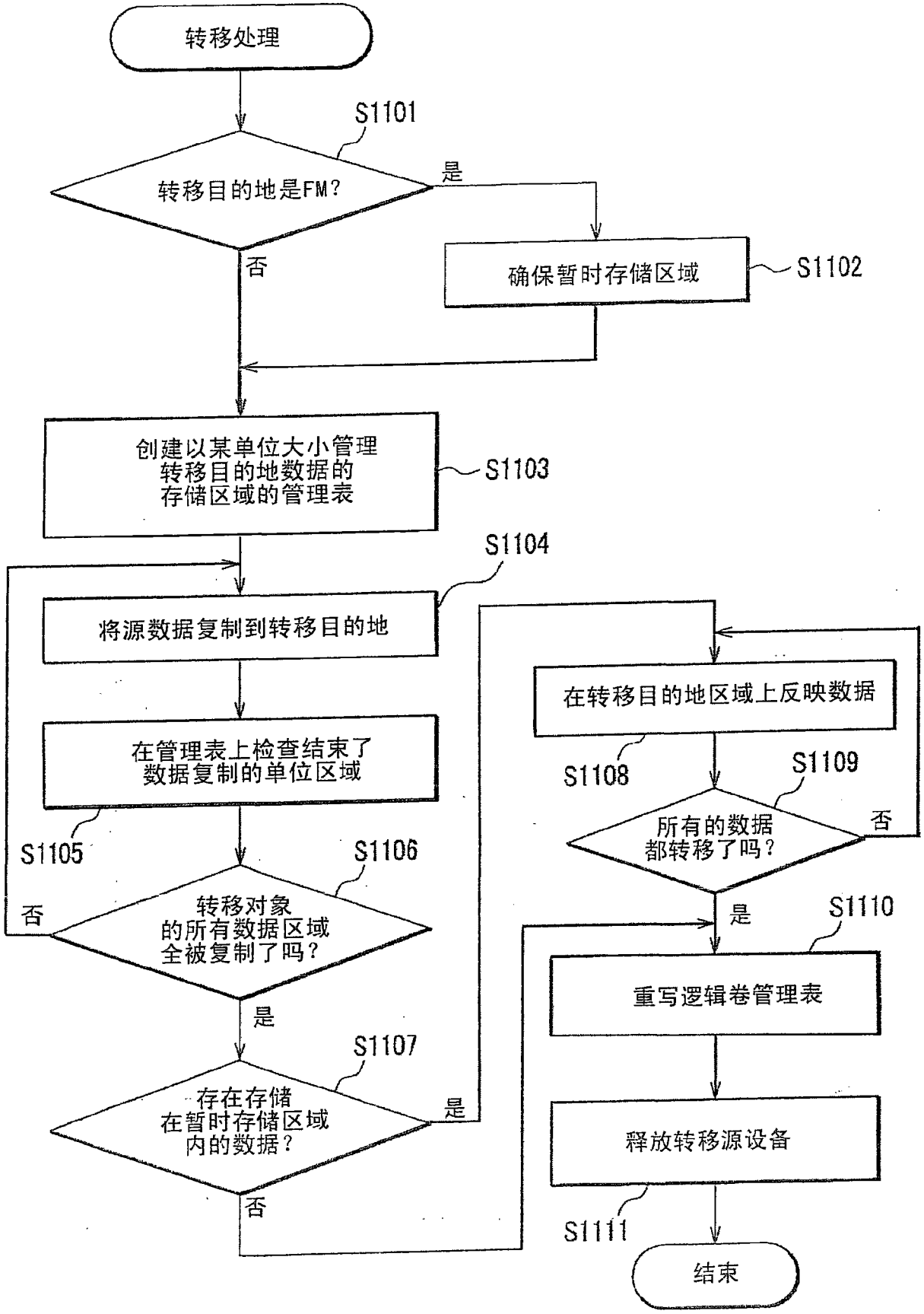

转移处理
S1101
转移目的地是FM?
是
否
确保暂时存储区域
S1102
创建以某单位大小管理
转移目的地数据的
存储区域的管理表
S1103
S1104
将源数据复制到转移目的地
在管理表上检查结束了
数据复制的单位区域
S1105
S1106
否
转移对象
的所有数据区域
全被复制了吗?
是
S1107
存在存储
在暂时存储区域
内的数据?
是
否
在转移目的地区域上反映数据
S1108
S1109
所有的数据
都转移了吗?
否
是
S1110
重写逻辑卷管理表
释放转移源设备
S1111
结束

图 11B

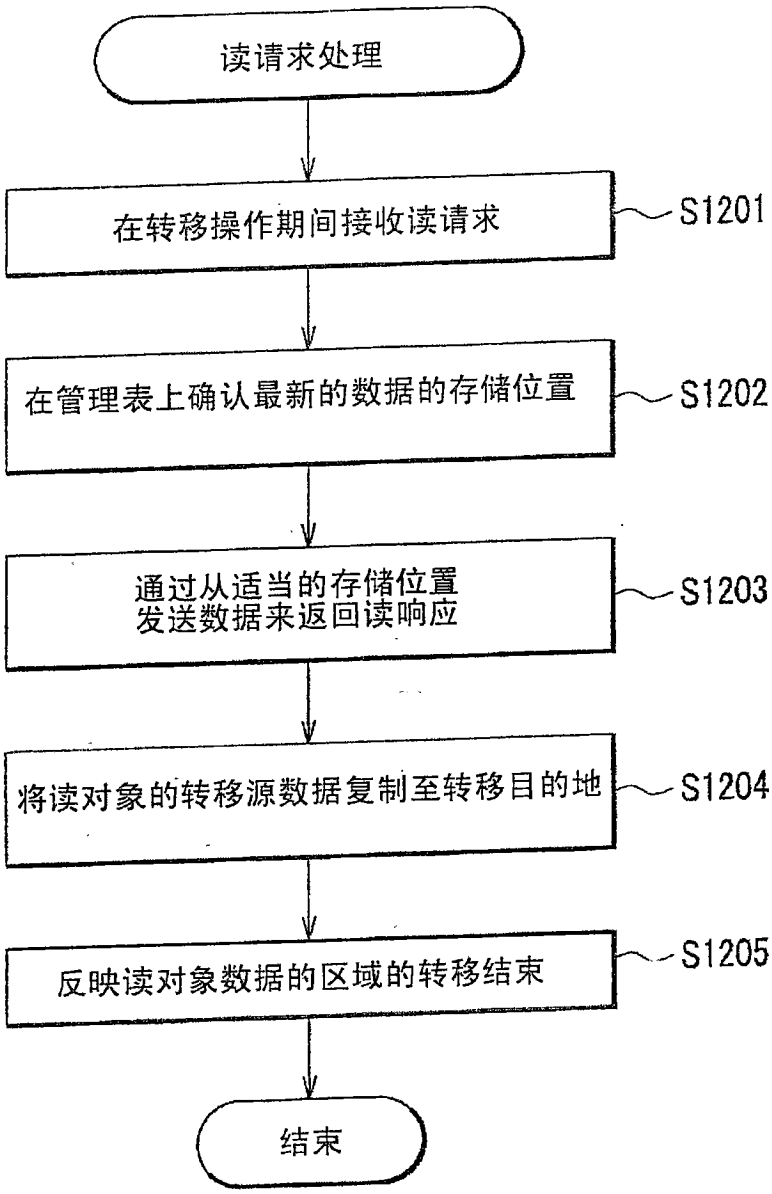

读请求处理
在转移操作期间接收读请求
S1201
在管理表上确认最新的数据的存储位置
S1202
通过从适当的存储位置
发送数据来返回读响应
S1203
将读对象的转移源数据复制至转移目的地
S1204
反映读对象数据的区域的转移结束
S1205
结束

图 12

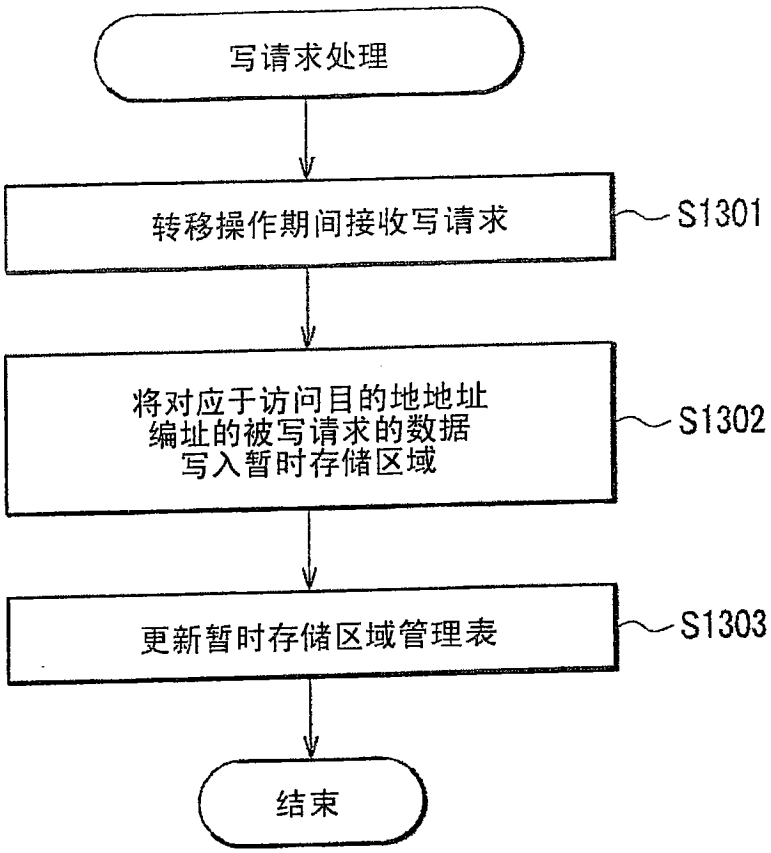

写请求处理
转移操作期间接收写请求
S1301
将对应于访问目的地地址
编址的被写请求的数据
写入暂时存储区域
S1302
更新暂时存储区域管理表
S1303
结束

图 13

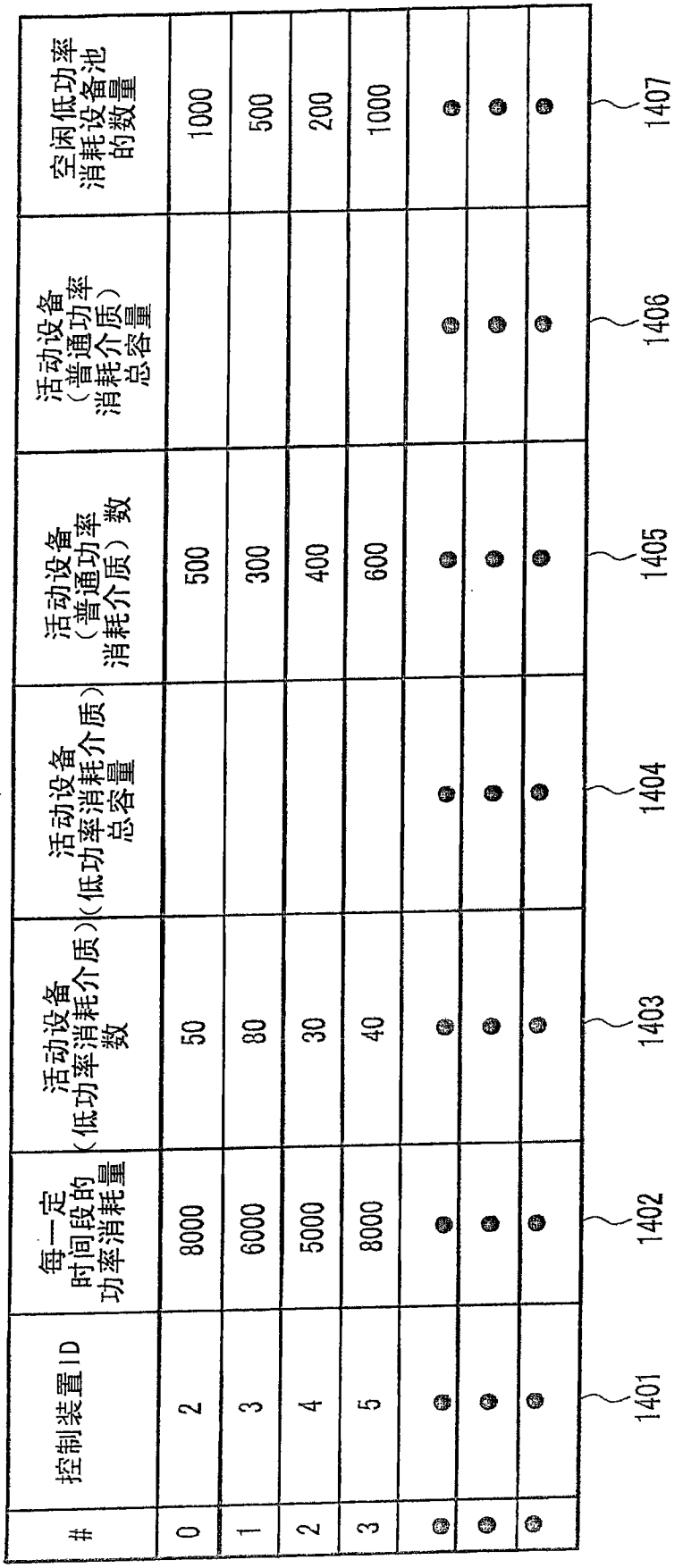

1400 整体信息表
#
控制装置ID
每一定时间段的功率消耗量
活动设备（低功率消耗介质）数
活动设备（低功率消耗介质）总容量
活动设备（普通功率消耗介质）数
活动设备（普通功率消耗介质）总容量
空闲低功率消耗设备池的数量
0 2 8000 50 500 1000
1 3 6000 80 300 500
2 4 5000 30 400 200
3 5 8000 40 600 1000
1401
1402
1403
1404
1405
1406
1407

图 14