(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104718536 A
(43)申请公布日 2015.06.17

(21)申请号 201380038949.4
(22)申请日 2013.06.24
(30)优先权数据
13/532,312 2012.06.25 US
(85)PCT国际申请进入国家阶段日
2015.01.22
(86)PCT国际申请的申请数据
PCT/US2013/047335 2013.06.24
(87)PCT国际申请的公布数据
WO2014/004381 EN 2014.01.03
(71)申请人 NETAPP 股份有限公司
地址 美国加利福尼亚
(72)发明人 S·K·埃尔普拉 V·加格
S·C·韦尼
(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 罗银燕
(51)Int.Cl.
G06F 11/20(2006.01)

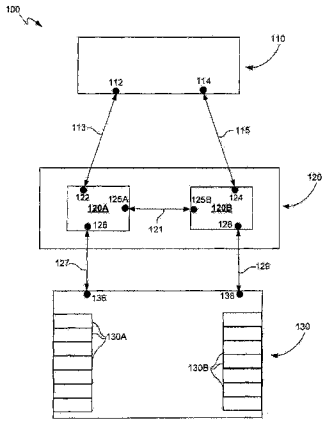
权利要求书3页 说明书10页 附图6页

(54)发明名称

网络存储系统中的非破坏性控制器更换

(57)摘要

一种基于网络的存储系统包括多个系统控制器和布置成集合体的多个物理存储设备，其中每个存储设备具有指示它所属的系统控制器的所有权部分。第一和第二系统控制器与彼此、存储设备和单独的主机服务器通信，并且每个系统控制器可被指定为系统节点，该系统节点控制存储设备的各个集合体，并且基于来自所述单独的主机服务器的命令对存储设备进行读和写。所述第一系统控制器控制第一系统节点，并且可促成利用单独的第三控制器进行第二系统控制器的自动热交换更换，该第二系统控制器最初控制第二系统节点，该第三控制器随后控制该第二系统节点。所述第一系统控制器可在第二控制器的自动热交换更换期间接管第二系统节点的控制。



CN 104718536 A

1. 一种基于网络的存储系统，包括：

多个物理存储设备，所述多个物理存储设备包括至少存储设备的第一和第二集合体，其中，所述存储设备的第一和第二集合体中的每一个存储设备包括其上的被配置为指示它所属的系统控制器的所有权部分；以及

多个系统控制器，所述多个系统控制器包括至少第一和第二系统控制器，所述第一和第二系统控制器分别与彼此、所述多个存储设备和单独的主机服务器通信，所述多个系统控制器中的每一个适于指定为控制存储设备的各个集合体并且用于基于从所述单独的主机服务器接收的命令对所述多个存储设备进行读和写的系统节点，

其中，所述第一系统控制器适于控制第一系统节点，并且被配置为促成利用单独的第三控制器进行所述第二系统控制器的自动热交换更换，该第二系统控制器最初控制第二系统节点，该单独的第三控制器随后控制该第二系统节点，其中，所述第一系统控制器进一步适于在所述第二控制器的自动热交换更换期间接管第二系统节点的控制，并且其中，所述第一系统控制器进一步适于在所述自动热交换更换期间自动地与所述单独的第三控制器互换系统标识符。

2. 根据权利要求 1 所述的基于网络的存储系统，其中，所述多个物理存储设备包括多个磁盘。

3. 根据权利要求 2 所述的基于网络的存储系统，其中，所述多个磁盘被包含在一个或多个文卷内。

4. 根据权利要求 1 所述的基于网络的存储系统，其中，所述第一和第二系统控制器包括高可用性控制器对。

5. 根据权利要求 1 所述的基于网络的存储系统，其中，存储设备的单个集合体内的每个存储设备属于同一个系统控制器。

6. 根据权利要求 1 所述的基于网络的存储系统，其中，所述第一系统控制器被配置为在不使所述第二系统节点进入维护模式的情况下促成自动热交换。

7. 根据权利要求 1 所述的基于网络的存储系统，其中，所述第一系统控制器被配置为在不需要用户手动地将信息输入到所述系统中的情况下促成自动热交换。

8. 根据权利要求 7 所述的基于网络的存储系统，其中，所述第一系统控制器被进一步配置为在所述第一系统控制器已接管第二系统节点的控制时，检测请求所述第二系统节点的控制的单独的查询控制器的标识。

9. 根据权利要求 8 所述的基于网络的存储系统，其中，所述第一系统控制器被进一步配置为当所述第一系统控制器检测到所述查询控制器具有与所述第二系统控制器的标识不匹配的标识时，将所述第二系统控制器的标识返回给所述查询控制器。

10. 根据权利要求 9 所述的基于网络的存储系统，其中，所述查询控制器是所述单独的第三控制器。

11. 根据权利要求 8 所述的基于网络的存储系统，其中，所述第一系统控制器被进一步配置为当所述第一系统控制器检测到所述查询控制器具有与所述第二系统控制器的标识不匹配的标识时，重写由所述第二系统控制器控制的每个存储设备的所有权部分以反映所述查询控制器的所有权。

12. 一种适于在冗余存储阵列环境中操作的第一基于网络的存储系统控制器，所述第

一控制器包括：

第一模块，所述第一模块适于促成第一系统节点的操作，所述第一系统节点基于从单独的主机服务器接收的命令来控制对于布置成第一集合体的第一多个存储设备中的每一个的读和写；

第二模块，所述第二模块适于促成与高可用性对布置中的单独的第二控制器的交互，其中，所述单独的第二控制器操作第二系统节点，所述第二系统节点基于从所述单独的主机服务器接收的命令来控制对于布置成第二集合体的第二多个存储设备中的每一个的读和写；以及

第三模块，所述第三模块适于通过在所述单独的第二控制器的自动热交换更换期间接管第二系统节点的控制来促成利用随后操作该第二系统节点的单独的第三控制器进行所述单独的第二控制器的自动热交换更换，其中，所述第二或第三模块中的至少一个进一步适于在利用单独的第三控制器进行所述单独的第二控制器的自动热交换更换期间自动地与所述单独的第三控制器互换系统标识符。

13. 根据权利要求 12 所述的第一系统控制器，其中，所述第三模块进一步适于在不使所述第二系统节点进入维护模式的情况下促成自动热交换。

14. 根据权利要求 12 所述的第一系统控制器，其中，所述第三模块进一步适于当所述单独的第三控制器在所述第三模块操作第二系统节点的同时准备控制所述第二系统节点时，检测所述单独的第三控制器的标识。

15. 根据权利要求 14 所述的第一系统控制器，其中，所述第三模块进一步适于当所述第三模块检测到所述单独的第三控制器具有与所述单独的第二控制器的标识不匹配的标识时，将所述单独的第二控制器的标识返回给所述单独的第三控制器。

16. 根据权利要求 14 所述的第一系统控制器，其中，所述第三模块进一步适于当所述第三模块检测到所述单独的第三控制器具有与所述单独的第二控制器的标识不匹配的标识时，重写由所述单独的第二控制器控制的每个存储设备的所有权部分以反映所述单独的第三控制器的所有权。

17. 根据权利要求 12 所述的第一系统控制器，进一步包括：

第四模块，所述第四模块适于当所述单独的第三控制器在自动热交换期间发生故障时，促成中止所述自动热交换更换并且使所述第二集合体的所有控制回复到所述第一系统控制器。

18. 一种更换基于网络的存储系统中的控制器的非破坏性方法，所述方法包括：

检测更换控制器上的自动热交换更换过程的存在；

检测原始系统控制器上的自动热交换更换过程的存在，其中，所述原始系统控制器和更换控制器被配置为作为高可用性控制器对进行操作；

更新存储设备的集合体中的每个存储设备上的第一所有权部分，以反映所述存储设备的集合体中的每个存储设备由所述更换控制器拥有；

启动所述更换控制器；以及

更新所述存储设备的集合体中的每个存储设备上的第二所有权部分，以与各个存储设备上的更新的第一所有权部分匹配。

19. 根据权利要求 18 所述的方法，进一步包括步骤：

在所述原始系统控制器上对于伙伴控制器状态进行轮询；

在所述原始系统控制器上检测所述伙伴控制器状态为等待归还；以及

在所述原始系统控制器上从所述伙伴控制器接收该伙伴控制器是更换控制器的标识。

20. 根据权利要求 18 所述的方法，其中，所述更新步骤中的每一个在不使各个系统节点进入维护模式的情况下以及在不要求用户手动地将信息输入到所述系统中的情况下被执行以使得能够进行自动热交换。

网络存储系统中的非破坏性控制器更换

技术领域

[0001] 本公开一般涉及网络存储系统，并且更特别地涉及网络存储系统上的控制器的更换。

背景技术

[0002] 存储区域网络（“SAN”）环境通常包括大量通过使用双控制器模型进行操作的存储设备。在许多情况下，这样的存储设备可包括至少一个磁盘阵列，这些磁盘阵列可被归类为独立磁盘冗余阵列（“RAID”）。在控制器通常被称为高可用性（“HA”）对的这样的双控制器模型布置下，单个的控制器可被分配以作为各种存储设备卷（volume）或阵列的主要控制器或“所有者”来操作系统节点。在故障或者更换控制器的其它原因的情况下，这些控制器还可从它们的替代或配对控制器接管存储设备卷。

[0003] HA 对中的系统控制器的更换或换出一般是众所周知的，并且在一些情况下，通常涉及控制器头（head）、NVRAM 卡和 / 或整个控制器的更换。这样的过程有时被称为“头交换（headswap）”，并且通常导致对于至少 HA 对和分配给该 HA 对的 RAID 的整个操作的显著破坏，否则对于 HA 和 RAID 可能所属的更大的 SAN 的整个操作的显著破坏。例如，头交换的常见方法涉及将受交换影响的系统节点引导到维护模式并且运行磁盘再分配操作。虽然有效，但是这是破坏性的，因为受影响节点所拥有的储存器在该处理期间通常是不可用的。

[0004] 头交换的其它方法可导致较少的破坏。例如，HA 对的一个控制器上的头交换可涉及未被更换的系统控制器接管更换节点。以这种方式，受影响节点所拥有的存储卷和设备被其余的控制器接管，并且至少在头交换处理期间是可用的。若干个步骤被手动执行以利用新安装的控制器换出即将离去（outgoing）的控制器。在旧控制器的这个换出完成之后，然后执行手动磁盘再分配操作，并且提供受影响节点从该其余的系统控制器到新安装的控制器的归还（giveback）。

[0005] 不幸的是，存在可由这样的非破坏性、但是大量手动处理引起的若干问题。HA 配对系统的头交换检测通常取决于检测集合体（aggregate）是外来的 RAID 同化、以及仅在这样的情况下发生的用于匹配磁盘所有权（ownership）的集合体的所有权清理。在一些情况下，这可使头交换检测不可靠。另外，这样的手动头交换过程依赖于用户准确地输入新控制器系统标识符。如果在该手动 ID 输入处理中发生任何错误，则头交换失败，并且更换节点必须整个重启。此外，当同时执行从控制器的归还和存储操作时，可出现问题，当在磁盘存活时重新分配它们时，可发生多磁盘崩溃（panic），并且通常不存在对于多次头交换的支持。

[0006] 尽管用于头交换的许多网络存储系统、设备和方法在过去通常很好地工作，但是总是存在对于改进的期望。特别地，所期望的是能够以克服前述问题的自动化的、非破坏性的而且可靠的方式为系统控制器提供头交换过程的网络存储系统和方法。

发明内容

[0007] 本公开的优点在于提供促成(facilitate)基于网络的存储系统中的更好的头交换的改进系统和方法。这样的改进系统和方法优选地能够以自动化的、非破坏性的而且可靠的方式更换HA对中的控制器头、NVRAM卡和/或全系统控制器。特别地，所公开的实施例涉及更自动化的头交换过程，在该过程中，HA配对控制器在引导或头交换处理中较早地互换它们的系统标识符，使得这些控制器知道头交换处理并且相应地动作。另外，归还涉及迅速地更新磁盘和RAID所有权以反映适当的系统标识符，使得更换控制器能够平稳地启动和整合。

[0008] 在本公开的各种实施例中，基于网络的存储系统可包括多个物理存储设备和多个系统控制器。所述物理存储设备可包括至少存储设备的第一和第二集合体，其中，每个存储设备包括其上的被配置为指示它所属的系统控制器的所有权部分。所述多个系统控制器可包括至少第一和第二系统控制器，所述第一和第二系统控制器分别与彼此、所述多个存储设备和单独的主机服务器通信。系统控制器中的每一个可适于作为控制存储设备的各个集合体并且基于从所述单独的主机服务器接收的命令对所述多个存储设备进行读和写的系统节点的指定。特别地，所述第一系统控制器可控制第一系统节点，并且被配置为促成用单独的第三控制器对所述第二系统控制器进行自动热交换更换，该第二系统控制器最初控制第二系统节点，该单独的第三控制器随后控制该第二系统节点。所述第一系统控制器可在第二控制器的该自动热交换更换期间接管第二系统节点的控制。此外，所述第一系统控制器还可在所述自动热交换更换期间自动地与所述单独的第三控制器互换系统标识符。

[0009] 在各种其它实施例中，适于在冗余存储阵列环境中操作的第一基于网络的存储系统控制器可包括多个模块。第一模块可适于促成第一系统节点的操作，所述第一系统节点基于从单独的主机服务器接收的命令来控制对于布置成第一集合体的第一多个存储设备中的每一个的读和写。第二模块可适于促成与高可用性对布置中的单独的第二控制器的交互，其中，所述单独的第二控制器促成第二系统节点的操作，所述第二系统节点基于从所述单独的主机服务器接收的命令来控制对于布置成第二集合体的第二多个存储设备中的每一个的读和写。第三模块可适于促成用随后操作该第二系统节点的单独的第三控制器对所述单独的第二控制器进行自动热交换更换。这可通过第三模块促成第一系统控制器在所述单独的第二控制器的自动热交换更换期间接管第二系统节点的控制来实现。另外，所述第二或第三模块中的至少一个可进一步适于在自动热交换更换期间自动地与所述单独的第三控制器互换系统标识符。

[0010] 在更进一步的实施例中，提供非破坏性地更换基于网络的存储系统中的控制器的各种方法。这样的方法可涉及：检测更换控制器上的自动热交换更换过程的存在；检测原始系统控制器上的自动热交换更换过程的存在，其中，所述原始系统控制器和更换控制器被配置为作为高可用性控制器对进行操作；更新存储设备的集合体中的每个存储设备上的第一所有权部分，以反映所述存储设备的集合体中的每个存储设备由所述更换控制器拥有；启动所述更换控制器；以及更新所述存储设备的集合体中的每个存储设备上的第二所有权部分，以与各个存储设备上的更新的第一所有权部分匹配。

[0011] 当查阅以下附图和详细描述时，本发明的其它装置、方法、特征和优点对于本领域的技术人员将是清楚的或者将变得清楚。意图在于，所有这样的另外的系统、方法、特征和优点都包括在本说明书内，在本发明的范围内，并且由所附的权利要求保护。

附图说明

[0012] 所包括的附图是用于说明性的目的，并且仅用于提供对于所公开的促成基于网络的存储系统中的非破坏性控制器更换的发明性设备、系统和方法的可能的结构和布置的示例。这些附图绝不限制本领域的技术人员在不背离本发明的精神和范围的情况下对本发明可能进行的形式和细节上的任何改变。

[0013] 图 1 以框图格式示出根据本发明的一个实施例的具有 HA 控制器对和多个存储设备的示例性的基于网络的存储系统。

[0014] 图 2 以框图格式示出根据本发明的一个实施例的具有多个模块的示例性的 HA 控制器。

[0015] 图 3A-3C 以框图和表格格式示出根据本发明的一个实施例的经受 HA 控制器对经历接管和归还过程的、对于存储设备的所有权指定的示例性进程。

[0016] 图 4A-4C 以框图和表格格式示出根据本发明的一个实施例的经受 HA 控制器对经历头交换的、对于存储设备的所有权指定的示例性进程。

[0017] 图 5A-5B 以框图和表格格式示出根据本发明的替代实施例的经受 HA 控制器对经历头交换的、对于存储设备的所有权指定的示例性的替代进程。

[0018] 图 6 提供根据本发明的一个实施例的从受影响的 HA 对中的其余的控制器的角度来讲的更换基于网络的存储系统中的控制器的示例性的非破坏性方法的流程图。

[0019] 图 7 提供根据本发明的一个实施例的从新控制器的角度来讲的更换基于网络的存储系统中的控制器的示例性的非破坏性方法的流程图。

[0020] 图 8 提供根据本发明的一个实施例的更换基于网络的存储系统中的控制器的示例性的非破坏性总体方法的流程图。

具体实施方式

[0021] 在该部分中描述根据本发明的装置和方法的示例性应用。提供这些示例仅仅在于增加背景并且帮助本发明的理解。因此对于本领域的技术人员将清楚的是，可以在没有这些具体细节中的一些或全部的情况下实施本发明。在其它情况下，为了避免不必要地模糊本发明，公知的处理步骤没有被详细描述。其它应用是可能的，使得以下示例不应被当作限制。

[0022] 在以下详细描述中，对附图进行参考，这些附图形成本说明书的一部分，并且在这些附图中，通过图示示出了本发明的具体实施例。尽管充分详细地描述这些实施例以使得本领域的技术人员能够实施本发明，但是理解，这些示例不是限制，使得可以使用其它实施例，并且可以在不背离本发明的精神和范围的情况下进行改变。

[0023] 本公开在各种实施例中涉及促成基于网络的存储系统中的非破坏性控制器更换的设备、系统和方法。这样的设备、系统和方法优选地能够以自动化的、非破坏性的而且可靠的方式更换 HA 对中的控制器头、NVRAM 卡和 / 或全系统控制器。在各种特定实施例中，更自动化的头交换过程涉及 HA 配对控制器在引导或头交换处理中较早地互换或提供它们的系统标识符，使得所有的控制器都知道头交换处理并且相应地动作。另外，归还给新的更换控制器可涉及迅速地更新磁盘和 RAID 所有权以反映适当的系统标识符，使得更换控制

器能够平稳地启动和整合。虽然本文中所公开的各种示例集中于 HA 对内的头交换的特定方面，但是将理解，本文中所公开的各种原理和实施例可视情况应用于基于网络的存储应用和系统中的其它控制器布置。

[0024] 开始于图 1，以框图格式示出了具有 HA 控制器对和多个存储设备的示例性的基于网络的存储系统。系统 100 可包括具有多个端口 112、114 的主机或服务器 110，该端口 112、114 促成沿着链路 113、115 对于多个控制器 120 的通信。控制器 120 可包括 HA 控制器对 120A、120B，该控制器 120A、120B 具有用于促成与主机 100 的通信的端口 122、124 以及用于促成沿着链路 136、138 与存储设备 130 的通信的另外的端口 136、138。内部连接或链路 121 可促成 HA 控制器 120A、120B 的端口 125A、125B 之间的通信。存储设备 130 可被布置成由 HA 控制器对 120 控制的集合体或 RAID，并且可包括若干个存储设备或卷 130A、130B。特别地，HA 控制器 120A 可操作第一系统节点，该第一系统节点可被指定为一组存储设备或卷 130A 的所有者或控制器，而 HA 控制器 120B 可操作第二系统节点，该第二系统节点可被指定为单独的一组存储设备或卷 130B 的所有者或控制器。如将容易意识到的，还可以包括与对于 HA 控制器对的控制器和存储设备布置相关的其它特征和细节。

[0025] 接着继续到图 2，以框图格式类似地示出了具有多个模块的示例性 HA 控制器。控制器 220 可以是 HA 对（诸如以上在图 1 中所阐述的 HA 对）的一部分。多个端口和链路可将控制器 220 耦合到若干个其它系统组件。例如，端口 221 和相关联的链路 225 可耦合到单独的伙伴 (partner) 或配对的 HA 控制器，端口 222 和相关联的链路 213 可耦合到单独的主机，而端口 226 和相关联的链路 227 可耦合到单独的存储阵列或存储设备的集合体。HA 控制器 220 内的多个模块 240、250、260、270 可促成若干个功能，包括与非破坏性的、自动的头交换过程相关联的那些功能。模块 240、250、260、270 中的每一个可适于视情况与其它模块和 / 或各种单独的系统组件进行通信或交互。

[0026] 例如，第一模块 240 可适于促成第一系统节点的操作，该第一系统节点基于从单独的主机服务器接收的命令来控制对布置成第一集合体或卷的第一多个存储设备中的每一个进行读和写。第二模块 250 可适于促成与高可用性对布置中的单独的第二控制器的交互。再次，该单独的第二控制器可适于操作第二系统节点，该第二系统节点基于从该单独的主机服务器接收的命令来控制对布置成第二集合体或卷的第二多个存储设备中的每一个进行读和写。第三模块 260 可适于促成利用单独的第三控制器进行单独的第二控制器的自动热交换更换，该单独的第三控制器随后操作第二系统节点。这至少部分通过第三模块 260 在利用单独的第三控制器进行单独的第二控制器的自动热交换更换期间接管第二系统节点的控制来实现。此外，第二或第三模块中的至少一个还可适于在自动热交换更换期间自动地与单独的第三控制器互换系统标识符。第四模块 270 可适于当单独的第三控制器在自动热交换期间发生故障时，促成中止自动热交换更换并使存储单元的第二集合体或卷的所有控制回复到第一系统控制器。以下更深入地提供模块 240、250、260、270 的各种细节和特征。

[0027] 如将意识到的，典型的具有 HA 控制器对的基于网络的存储系统通常通过在磁盘被分配给由适当的控制器操作的节点时将该控制器的“nvram_system_ID”或另一个合适的系统标识符写入到该磁盘的所有权部分或区域来操作。磁盘的该所有权部分或区域可以任何数量的合适的方式被具体地提及，但是为了讨论的目的在本文中将被称为磁盘的

“SANOWN”区域或部分。该处理针对分配的磁盘进行，这可帮助建立关于哪个节点或控制器拥有磁盘的映射。该相同的信息还被高速缓存在 RAID 标签中以用于识别集合体的所有者。就这点而论，每当 NVRAM 卡、控制器头和 / 或整个控制器被更换（即，头交换）时，这些磁盘上所有权然后也改变。

[0028] 可涉及控制器故障转移（“CFO”）、储存器故障转移（“SFO”）或这两者的用于进行头交换的一种方法将首先把更换节点引导到维护模式。这通常被进行，原因在于用于新控制器的 NVRAM ID 或其它标识符不同于存储在磁盘上的标识符，使得被更换的控制器在普通的操作中不能有效地启动。当更换节点和新控制器处于维护模式并且离线时，然后可运行磁盘再分配操作以提取旧的被更换的控制器所拥有的磁盘的列表并改变每一个中的 SANOWN 以反映新控制器的标识符。但是如以上所提到的，该处理是破坏性的，因为受影响的磁盘的储存器在整个过程期间是不可用的。

[0029] 这样的破坏可例如通过使 HA 对中的其余的控制器接管正被更换的即将离去的控制器所拥有的存储设备或卷来避免。例如，在 HA 控制器对包括控制器 A 和 B 并且控制器 B 将被新控制器 C 更换的情况下，那么，A 可在 B 被 C 更换时执行 B 所拥有的储存器的接管。然后在 A 处于对象储存器的控制下时对 A 执行从 B 到 C 的磁盘再分配，此后，从 A 到 C 执行归还。但是如以上所提到的，该处理可引起它自己的多组问题。

[0030] 例如，SANOWN 可以是负责处理磁盘所有权的主要模块，其中 RAID 也高速缓存磁盘所有权。协议在这两层之间演变以处理所有权不一致。每当磁盘的所有权改变时，SANOWN 就可模拟对于该磁盘的假删除和添加事件，并且还将通知发送到 RAID，RAID 一接收到这样的通知就采取适当的动作。但是 RAID 协议通常要求属于集合体的磁盘由同一个节点拥有。如果这些磁盘中的一个的磁盘所有权改变，则该磁盘将从集合体移除。于是在一些次序中，磁盘再分配可使伙伴磁盘的所有权从 B 变为 C，C 继而对于该磁盘（甚至包括作为集合体的一部分的那些）产生删除和添加事件。这可使 RAID 认为集合体已丢失了它的所有磁盘，并且可导致 A 上的崩溃，这然后可导致比原始的基本处理甚至更多的破坏。

[0031] 虽然该特定问题可通过在磁盘再分配处理期间不发送通知来避免，但是这样的变通方案使系统进入基本上或完全不一致的状态，在这种状态下，SANOWN 和 RAID 所有权值不同。这种不一致的状态可能是不稳定的，因为当然后实现不同的所有权值时，RAID 触发内部再扫描的任何操作然后将导致系统崩溃。另外，前述变通方案还需要大量的人工干预，诸如用户手动地输入新控制器的系统标识符。虽然不方便，但是这样的处理还易于出现新系统标识符的错输入的用户错误，这然后将导致所意图的头交换处理发生故障。无论如何，更自动化的、可靠的而且非破坏性的更好的头交换方法是优选的。

[0032] 这样的改进方法可继续在 RAID 层高速缓存所有权，同时以改进的方式精心安排所有权不一致。这些方法的特征通常在于，在头交换处理中较早地互换系统标识符，当集合体离线时更新所有权，节点处理存在两个伙伴的瞬时状况的能力，以及更可靠的错误恢复机制。作为特定的示例，改进的头交换过程可包括五个总体有序的步骤：

[0033] ●控制器 C 上的头交换检测；

[0034] ●控制器 A 上的头交换检测；

[0035] ● CFO 归还期间的所有权更新；

[0036] ● CFO 归还之后的控制器 C 的引导次序；和

[0037] ● SFO 归还期间的所有权更新。

[0038] 再次，在这些步骤中，反映了这样的系统，在该系统中，控制器“A”和“B”形成 HA 对，并且系统控制器“B”正被新的或更换系统控制器“C”更换，该系统控制器“C”然后将与“A”配对以形成新的 HA 对。

[0039] 关于控制器 C 上的头交换检测，新安装的系统控制器 C 启动并且咨询 SANOWN 以查看是否存在附连到控制器 C 的由控制器 C 拥有的任何磁盘。如果与其附连的磁盘不由控制器 C 拥有，则该新控制器将传统地重启。然而，在改进的头交换过程下，控制器 C 在做出重启决定之前检查其 HA 伙伴控制器（即，控制器 A）的状态。就这点而论，控制器 C 经由互连读取控制器 A 的状态，并且确定控制器 A（即，其 HA 伙伴）是否处于接管模式。如果互连断掉，或者如果控制器 A 不处于接管模式，则控制器 C 按照它通常的那样重启。

[0040] 但是如果控制器 A 被确定为处于接管模式，则控制器 C 进入“等待归还”状态，而不重启。在等待从控制器 A 归还时，新控制器 C 继续经由互连将其系统标识符发送到控制器 A。在控制器 C 等待归还时，控制器 A 可在 A 接管之前将系统标识符发送到控制器 C，该系统标识符指示哪个控制器是对于 A 的 HA 伙伴控制器。如果控制器 A 发送的该系统标识符与控制器 C 的系统标识符不匹配，则控制器 C 能够确定它是换出 (depart) 的控制器 B 的更换控制器并且头交换过程在进行中。

[0041] 关于控制器 A 上的头交换检测，在控制器 A 处于接管模式时，该其余的系统控制器 A 通过互连链路重复地对于其 HA 伙伴控制器的状态进行轮询。就这点而论，控制器 A 经由互连读取控制器 C 的状态，并且确定控制器 C（即，其 HA 伙伴）是否处于“等待归还”模式。如果对于控制器 A 的 HA 伙伴（在这种情况下，即，控制器 C）被确定为处于“等待归还”模式或状态，并且如果 HA 伙伴发送与原始 HA 伙伴（即，旧的控制器 B）的标识符不匹配的系统标识符，则控制器 A 能够确定其当前的 HA 伙伴控制器是更换控制器并且头交换过程在进行中。

[0042] 随后，在新控制器 C 处于其“等待归还”模式或状态时，该其余的控制器 A 发起归还过程。该归还过程被设计为取得最初由换出的控制器 B 拥有、现在被其余的控制器 A 接管的存储设备或卷，并且将这些存储单元给予控制器 C。该归还可分为两个阶段。首先，归还的 CFO 阶段可涉及：在控制器 A 上使 CFO 集合体离线，并然后将这些 CFO 集合体归还给控制器 C，此后控制器 C 进一步启动。在控制器 C 已完全启动之后，随后的归还的 SFO 阶段可涉及：在控制器 A 上使 SFO 集合体离线，并然后将这些 SFO 集合体归还给控制器 C。

[0043] 关于 CFO 归还期间的所有权更新，控制器 A 使 B 拥有的 CFO 集合体离线，并发起归还处理。在该处理中，控制器 A 将 CFO 集合体磁盘的 SANOWN 所有权从 B 变为 C，并且将控制器 A 自身上的 HA 伙伴系统标识符更新为 C。重要的是，控制器 A 保留旧控制器 B 系统标识符的知识 (knowledge)，使得控制器 A 能够处理当前由控制器 B 和控制器 C 两者拥有的磁盘的混合。在其启动期间，控制器 C 变得知道它处于局部头交换状态，并且 CFO 集合体磁盘的 RAID 所有权也被更新为 C。

[0044] 关于控制器 C 在 CFO 归还之后的引导次序，控制器 C 在其 HA 伙伴（即，控制器 A）移出接管模式之后继续寻找具有反映控制器 C 的系统标识符的所有权值的磁盘。控制器 C 继续寻找这样的磁盘，直到没有更多的磁盘被发现为止。在启动期间，在局部头交换处理在进行中时，控制器 C 将 RAID 所有权更新为其系统标识符。在这完成之后，可使受影响的存

储设备或卷重回在线，并且控制器 C 的启动可继续。一旦控制器 C 完全启动，它就准备好接收可以应用的任何 SFO 集合体磁盘。

[0045] 最后，关于 SFO 归还期间的所有权更新，一旦控制器 C 完全启动，SFO 归还就被触发。这可涉及控制器 A 使先前由控制器 B 拥有的 SFO 集合体逐一离线，并且将 SFO 集合体中的所有磁盘的 SANOWN 和 RAID 所有权更新为 C。在控制器 A 已将所有的 SFO 集合体磁盘归还给其新的 HA 伙伴控制器 C 之后，然后 A 可忘记其旧的伙伴控制器 B 的系统标识符。控制器 A 然后清除伙伴头交换信息，并且控制器 C 清除局部头交换信息，并且头交换完成。

[0046] 再次，控制器 A 在许多方面可以是典型的 HA 配对控制器，使得它可包括适于促成控制对布置成第一集合体的第一多个存储设备中的每一个的读和写的第一系统节点的操作的第一模块，并且还包括适于促成与高可用性对布置中的单独的第二控制器的交互的第二模块。如上所述，控制器 A 上的第三模块可适于通过在单独的第二控制器的自动热交换更换期间接管第二系统节点的控制来促成利用单独的第三控制器进行单独的第二控制器的自动热交换更换。

[0047] 该改进的头交换过程的各种优点可被实现。一个优点提供头交换过程的大部分或基本上全部被自动化，这可减少或消除人工用户干预以及由这样的行为可引起的可能错误。这可能至少部分是由于在头交换处理中在 HA 控制器之间系统标识符的较早互换而促成的。另外，当存储目标或集合体短暂离线时，执行磁盘所有权改变，这使错误最小化，同时改进数据可靠性。另一个优点提供旧的和新的控制器系统标识符两者在系统中被持久地记住，这允许改进的错误恢复，因为头交换过程期间任一节点上的故障可被更得体地处理。

[0048] 这样的错误恢复可确保头交换过程期间任一节点上的故障不会导致永久的数据中断。用于这样的错误恢复的合适的协议或规则可涉及，例如，控制器 C 在 CFO 归还之后、但在头交换完成之前发生故障。在这样的情形下，控制器 A 可再次执行受影响的节点的接管，并且使所有的 SANOWN/RAID 所有权回复到一致状态。就这点而论，一致状态将涉及使所有权回到 B，其中接管和数据再次由控制器 A 服务。然后，在某一稍后时间，可安装并且启动控制器 C 或另一个更换控制器，控制器 C 或另一个更换控制器一被安装和启动，新的头交换周期就将发生。关于控制器 A，这样的错误恢复可由例如第四模块控制，该第四模块适于当控制器 C 在自动热交换期间发生故障时，促成中止自动热交换更换并使受影响的集合体的所有控制回复到控制器 A。

[0049] 作为另一个错误恢复示例，控制器 A 可能在 CFO 归还之后、但在头交换完成之前发生故障。在这样的情形下，新控制器 C 然后可接管。由于控制器 C 仅必须应对一个伙伴（即，控制器 A），所以作为接管的一部分，控制器 C 对先前由旧控制器 B 拥有的其余的集合体进行所有权清理（即，将磁盘所有权从 B 变为 C）。

[0050] 接着转到图 3A-3C，以框图和表格格式示出了经受 HA 控制器对经历接管和归还过程的、对于存储设备的所有权指定的示例性进程。图 3A 描绘了接管之前的布置 300，其中，A 和 B 是 HA 对中的两个节点，A 将接管 B。接管之前的 B 的 CFO 和 SFO 集合体的所有权在表中对于所有值反映为“B”。接着，图 3B 描绘了接管之后的布置 302，其中，A 接管了 B 的卷。如所示的，SFO 集合体的 SANOWN 当前的所有者和 RAID 所有者已从 B 变为 A。接着，图 3C 描绘了从 A 到 B 的归还期间的布置 304。这样的归还是两步处理，其中，所有的 CFO 集合体被一起归还，如果适当，之后 SFO 集合体被逐一归还，如果适用的话。如可从该特定的布

置 304 看出的，CFO 集合体的所有权不存在改变，而 SFO 集合体经历从 A 回到 B 的改变。

[0051] 图 4A-4C 以框图和表格格式示出了经受 HA 控制器对经历头交换的、对于存储设备的所有权指定的示例性进程。如将容易意识到的，图 4A-4C 中所示的示图和格式反映传统的处理，在该处理中，即将离去的控制器 B 通过新控制器 C 进行头交换。图 4A 描绘了布置 400，其涉及在控制器 A 上运行“磁盘再分配”命令以将属于 B 的磁盘的所有权变为新控制器 C。图 4B 描绘了布置 402，其涉及在磁盘按照图 4A 被再分配之后发出归还。特别地，正在启动的新控制器可检测 SANOWN 和 RAID 所有权之间的不一致性，并然后可更新 RAID 所有权以与 SANOWN 所有权匹配。就这点而论，控制器 C 上的所有权改变涉及 RAID 所有权值从 B 更新为 C。图 4C 然后描绘了布置 404，其中，在接管控制器更新 SANOWN 和 RAID 所有权以使彼此匹配的情况下，进行 SFO 归还。在这种情况下，SFO 所有权值从 A 变为 C。

[0052] 现在继续图 5A-5B，以框图和表格格式类似地示出了根据替代实施例的经受 HA 控制器对经历头交换的、对于存储设备的所有权指定的示例性替代进程。按照以上五个一般步骤及其细节中所阐述的改进的头交换处理，图 5A 提供了布置 500，其涉及改进的归还过程。在该过程中，SANOWN 和 RAID 所有权在归还期间改变。如所示的，在控制器 A 上的归还期间，SANOWN 原所有者和当前所有者从 B 变为 C。而且，在控制器 C 上的启动期间，RAID 所有者从 B 变为 C。图 5B 然后描绘了布置 502，其中，SFO 集合体在控制器 C 完全启动之后被归还。如所示的，SANOWN 和 RAID 所有权都被改变以反映 C。

[0053] 再次，前述改进的头交换过程是非破坏性的、自动化的、而且更可靠的，所有这些是至少部分由于以下原因而导致的：HA 配对控制器之间系统标识符的较早互换；两个 HA 配对控制器识别头交换在进行中的能力；以及当标识符的不匹配或者另一个问题出现时，控制器简单地拒绝重启。

[0054] 现在转到图 6-8，提供了根据前述改进处理和特征的执行头交换的各种方法。首先，图 6 提供了从受影响的 HA 对中的其余的控制器的角度来讲的更换基于网络的存储系统中的控制器的示例性的非破坏性方法的流程图。特别地，这样的方法可涉及以上详细提供的各种网络、系统、控制器、存储设备及其各种特征中的任何一个的使用。此外，将容易意识到，并非该流程图中所阐述的每一个方法步骤都总是必要的，并且还可以包括本文中没有阐述的进一步的步骤。此外，在一些实施例中，在合适时，可重新布置步骤的顺序。例如，在一些情况下，步骤 610 可能在步骤 612 之后或者与步骤 612 同时发生。

[0055] 开始于开始步骤 600，在处理步骤 602，HA 对中的第一控制器在接管模式下操作。在接着的判定步骤 604，关于是否已从来自第一控制器所属的 HA 对的伙伴控制器接收到系统标识符进行查询。如果否，则方法回复到处理步骤 602，在处理步骤 602，第一控制器继续在接管模式下操作。然而，当在步骤 604 接收到来自配对的控制器的系统标识符时，然后该方法继续到判定步骤 606，在判定步骤 606，利用第一系统控制器关于所接收的系统标识符是否与已经存档 (on file) 的伙伴系统标识符匹配进行查询。如果系统标识符匹配，则该方法移至处理步骤 608，在处理步骤 608，确定没有头交换发生，并且对于旧的 HA 配对控制器的正常归还过程发生。该方法然后从步骤 608 移至结束步骤 624。

[0056] 然而，在判定步骤 606 所接收的系统标识符与已经存档的系统标识符不匹配的情况下，然后该方法继续到处理步骤 610，在处理步骤 610，确认或“检测到”头交换处理或模式在进行中。在处理步骤 612，来自旧的被更换的控制器的系统标识符被发送到新控制器，

此后，在判定步骤614关于归还过程是否已被发起进行查询。如果否，则该方法循环回到步骤614，直到归还实际被发起为止。在归还被发起之后，该方法移至处理步骤616，在处理步骤616，新控制器的系统标识符被写入到第一控制器上的伙伴邮箱或类似的条目，此后，在处理步骤618，第一控制器更新新系统标识符的受影响磁盘所有权。第一控制器上的状态被适当地更新以反映新控制器的存在及其已进行了先前由旧的被更换的控制器拥有的磁盘或卷的归还的状况。在处理步骤622，退出头交换模式，一退出头交换模式，HA控制器对及其各自控制的卷的正常操作就发生。该方法然后在结束步骤624结束。

[0057] 接着，图7提供了从新控制器的角度来讲的更换基于网络的存储系统中的控制器的示例性的非破坏性方法的流程图。再次，该方法可涉及以上详细提供的各种网络、系统、控制器、存储设备及其各种特征中的任何一个的使用，并且将容易意识到，并非所阐述的每一个步骤都总是必要的，还可以包括进一步的步骤，并且在一些实施例中，在合适时，可重新布置步骤的顺序。开始于开始步骤700，在处理步骤702，新的或更换控制器开始在其新位置启动。在引导处理中较早地，在判定步骤704关于是否存在附连到新控制器的、实际由该新控制器拥有的任何磁盘进行查询。如果是，则该方法移至处理步骤706，在处理步骤706，RAID同化和正常引导处理发生。此时没有接管或归还发生，并且该方法然后从步骤706移至结束步骤730。

[0058] 然而，如果在判定步骤704实际不存在由该控制器拥有的磁盘，则该方法继续到判定步骤708，在判定步骤708，关于HA配对的伙伴控制器是否处于接管模式进行查询。如果否，则该方法回复到处理步骤702，并且重复步骤702至704。但是当在判定步骤708确定伙伴控制器处于接管模式时，则该方法移至处理步骤710，在处理步骤710，新控制器将其系统标识符发送到配对的伙伴控制器。然后在步骤712关于作为响应是否从伙伴控制器接收回系统标识符进行查询。如果否，则该方法循环回到处理步骤710，直到实际从伙伴控制器接收回系统标识符为止。再次，来自伙伴控制器的该系统标识符代表在接管模式发生之前与现在处于接管模式的伙伴控制器配对的控制器的系统标识符。

[0059] 在步骤712接收到系统标识符之后，在接着的判定步骤714关于所接收的系统标识符是否与启动控制器的系统标识符匹配进行查询。如果系统标识符确实匹配，则启动控制器认识到它是原始伙伴控制器，没有头交换发生，并且该方法回复到处理步骤702。然而，如果系统标识符不匹配，则该方法继续进行到处理步骤716，在处理步骤716，确认或“检测到”头交换在进行中。然后等待归还。然后在随后的处理步骤718发起归还处理，并且该方法然后继续移至处理步骤720，在处理步骤720，将发现所有的磁盘。在接着的处理步骤722，改变集合体所有权以与磁盘所有权匹配，此后，头交换完成并且在处理步骤724退出头交换模式。该方法然后在结束步骤726结束。

[0060] 最后移至图8，提供了更换基于网络的存储系统中的控制器的示例性的非破坏性总体方法的流程图。特别地，这样的方法可涉及使用或操作上述各种基于网络的存储系统控制器或其它组件中的任何一个。再次，将容易意识到，并非该流程图中所阐述的每一个方法步骤总是必要的，并且还可以包括本文中没有阐述的进一步的步骤。此外，实际的步骤顺序可以根据需要针对各种应用而改变。例如，步骤802和804可颠倒或者同时执行。

[0061] 开始于开始步骤800，在处理步骤802，在新引入的（例如，更换）HA配对控制器上检测头交换或系统控制器热交换的存在。在处理步骤804还在原始HA配对控制器上检测

头交换或热交换的存在，此后，在处理步骤 806，原始控制器对新引入的伙伴控制器针对其状态进行轮询。在接着的处理步骤 808，原始控制器检测到新伙伴控制器的状态为等待归还，此后，在处理步骤 810，原始控制器接收新伙伴控制器的系统标识符。

[0062] 在随后的处理步骤 812，更新存储设备的集合体中的每个存储设备上的第一所有权部分，以反映该集合体中的每个存储设备现在由新引入的或更换控制器拥有。然后在处理步骤 814 启动新引入的控制器，此后，在处理步骤 816 更新存储设备中的每一个上的第二所有权部分，以反映新引入的控制器的所有权。该方法然后继续进行以在结束步骤 818 完成。没有描绘的进一步的步骤可包括，例如，原始控制器将旧的被更换的控制器的标识符发送到新引入的控制器、和 / 或在发起头交换或归还处理之前确定所接收的标识符是否与所存储的标识符匹配。根据需要，其它方法步骤可包括前述来自图 6 和图 7 中所示的方法的步骤中的一个或多个。

[0063] 尽管出于清晰和理解的目的已通过图示和示例详细地描述了前述发明，但是将认识到，在不背离本发明的精神或本质特性的情况下，可以在大量其它具体的变型和实施例中体现上述发明。可以实施各种改变和修改，并且理解，本发明不受前述细节限制，而是相反由权利要求的范围所限定。

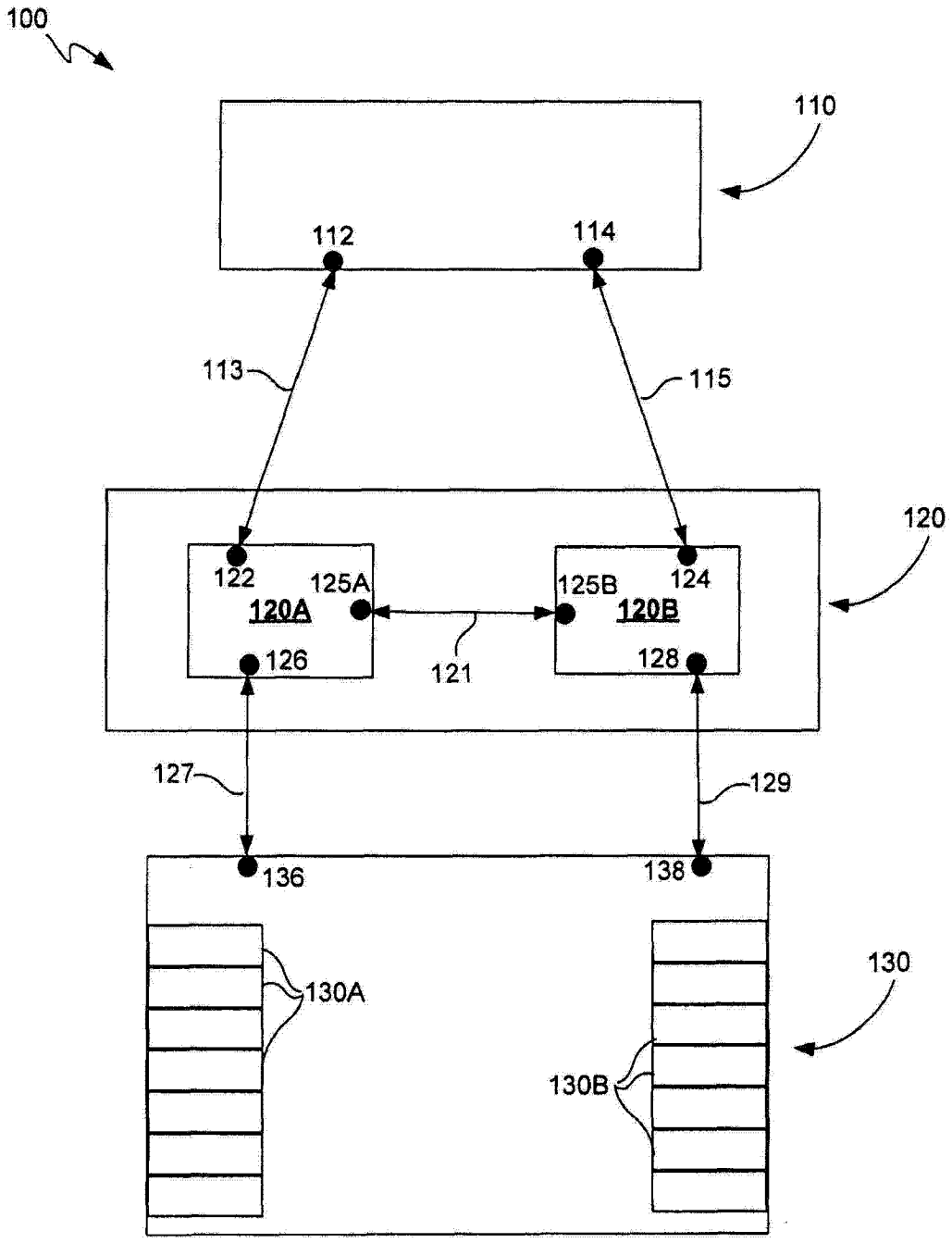

100
110
112
114
113
115
120
122
125A
120A
125B
124
120B
126
121
128
127
129
136
138
130
130A
130B

图 1

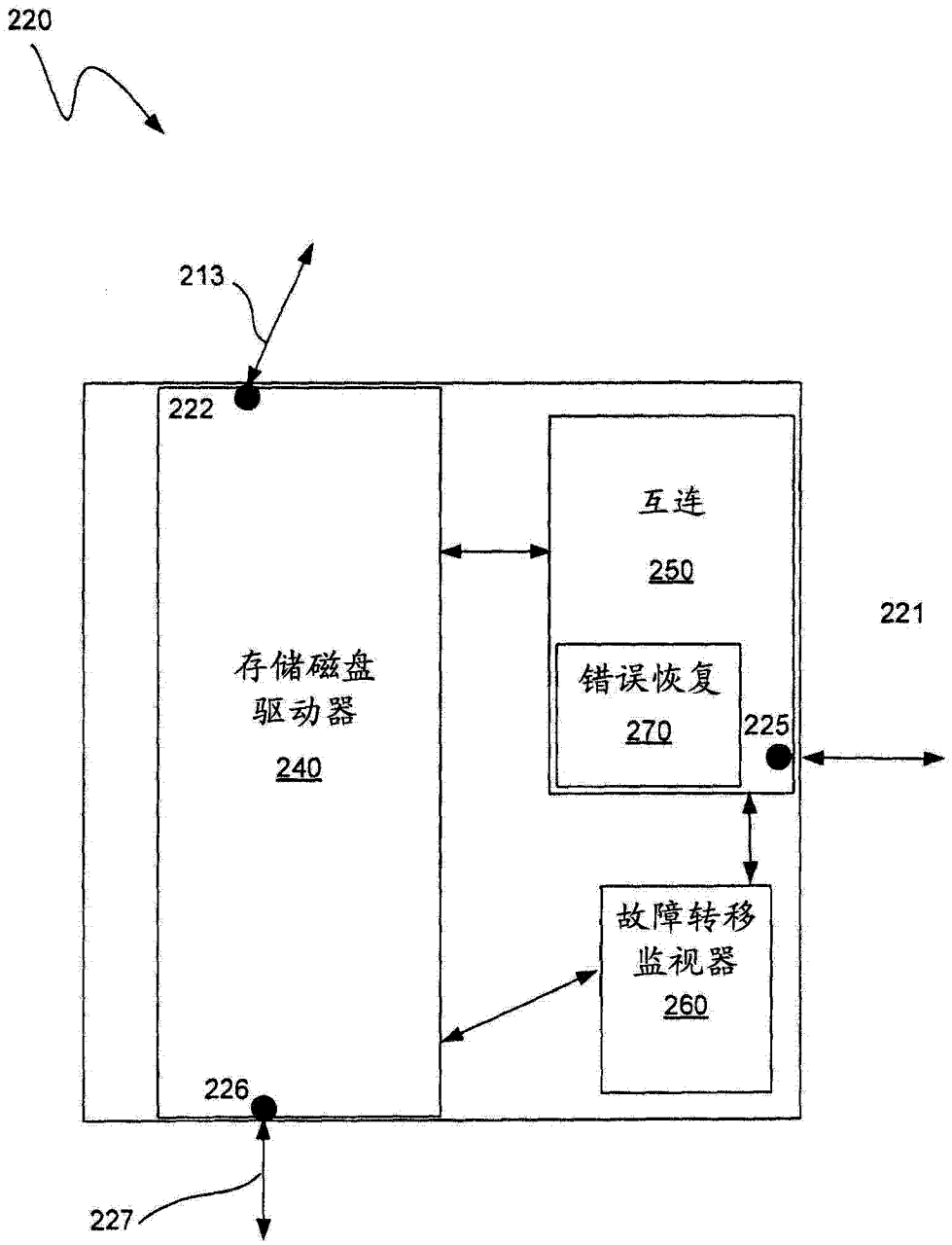

220
213
222
互连
250
221
存储磁盘
驱动器
240
错误恢复
270
225
故障转移
监视器
260
226
227

图 2

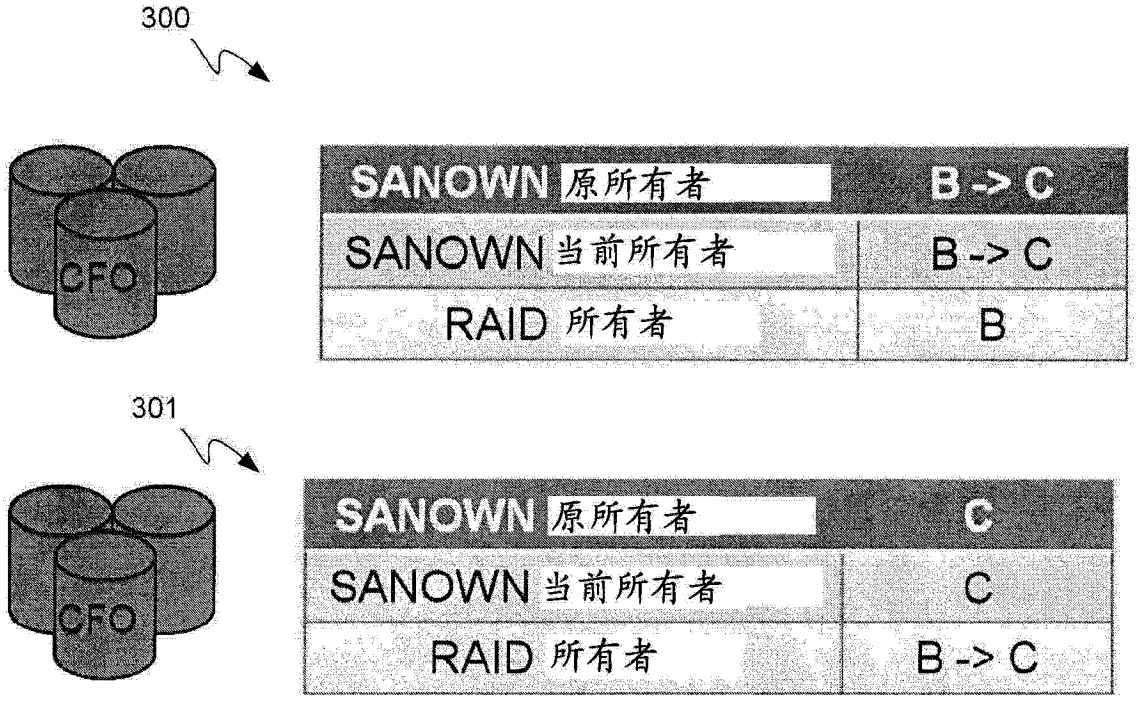

300
CFO
SANOWN 原所有者
B -> C
SANOWN 当前所有者
B -> C
RAID 所有者
B
301
CFO
SANOWN 原所有者
C
SANOWN 当前所有者
C
RAID 所有者
B -> C

图 3A


302
SFO
SANOWN 原所有者
B -> C
SANOWN 当前所有者
A -> C
RAID 所有者
A -> C

图 3B

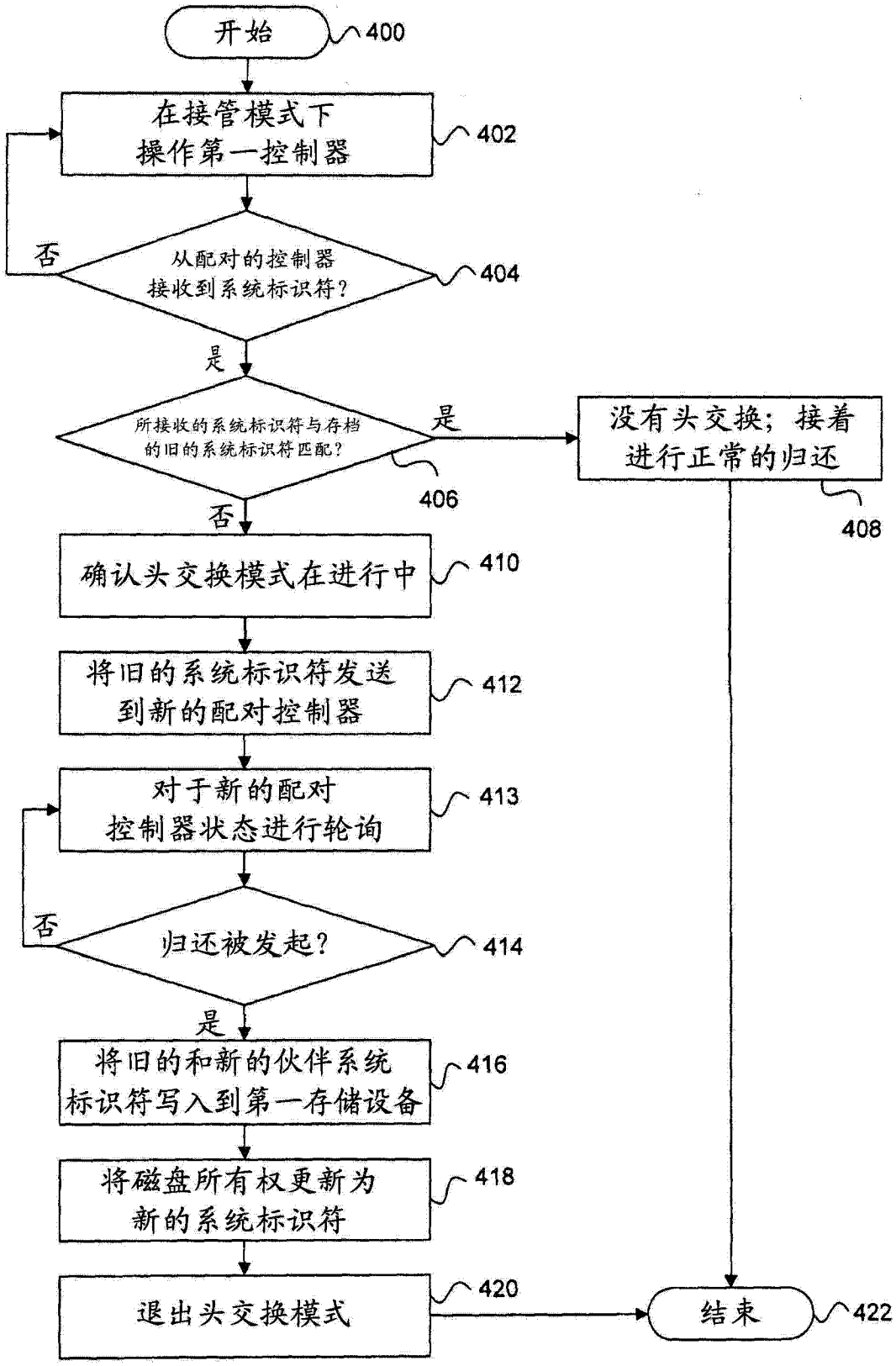

开始
400
在接管模式下
操作第一控制器
402
否
从配对的控制器
接收到系统标识符?
404
是
所接收的系统标识符与存档
的旧的系统标识符匹配?
是
406
没有头交换；接着
进行正常的归还
408
否
确认头交换模式在进行中
410
将旧的系统标识符发送
到新的配对控制器
412
对于新的配对
控制器状态进行轮询
413
否
归还被发起?
414
是
将旧的和新的伙伴系统
标识符写入到第一存储设备
416
将磁盘所有权更新为
新的系统标识符
418
退出头交换模式
420
结束
422

图 4

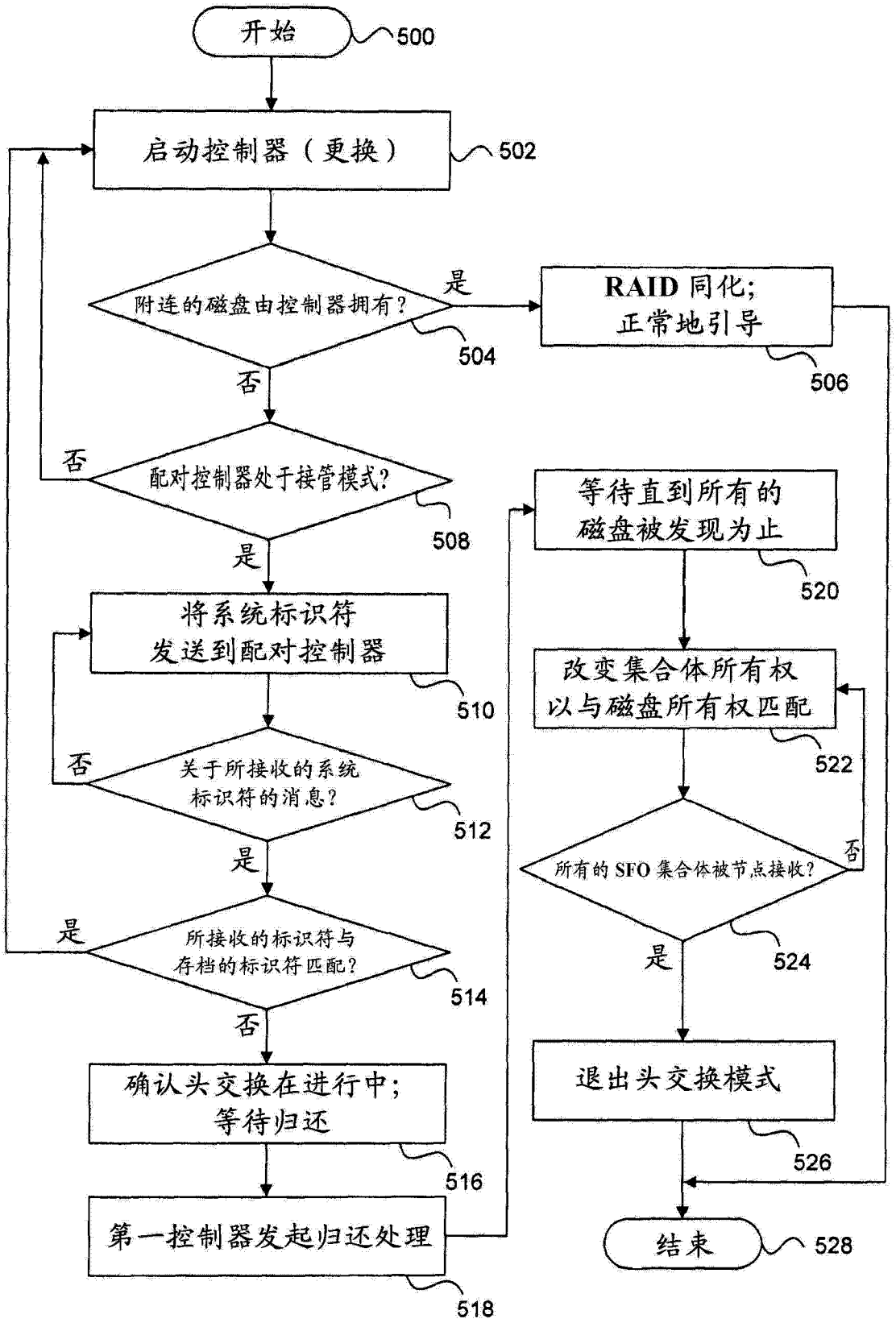

开始
500
启动控制器（更换）
502
附连的磁盘由控制器拥有？
是
504
否
RAID 同化；
正常地引导
506
否
配对控制器处于接管模式？
508
是
将系统标识符
发送到配对控制器
510
否
关于所接收的系统
标识符的消息？
512
是
是
所接收的标识符与
存档的标识符匹配？
514
否
确认头交换在进行中；
等待归还
516
第一控制器发起归还处理
518
等待直到所有的
磁盘被发现为止
520
改变集合体所有权
以与磁盘所有权匹配
522
所有的 SFO 集合体被节点接收？
否
524
是
退出头交换模式
526
结束
528

图 5

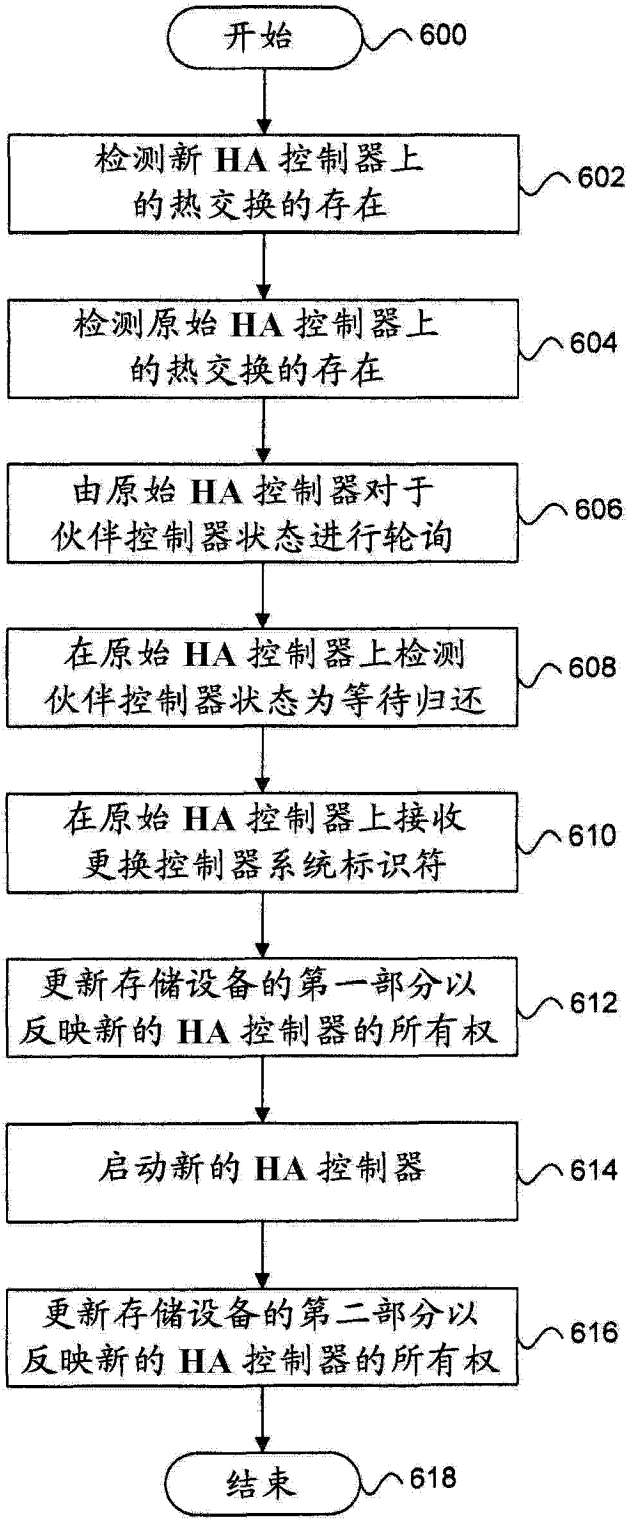

开始
600
检测新 HA 控制器上的热交换的存在
602
检测原始 HA 控制器上的热交换的存在
604
由原始 HA 控制器对于伙伴控制器状态进行轮询
606
在原始 HA 控制器上检测伙伴控制器状态为等待归还
608
在原始 HA 控制器上接收更换控制器系统标识符
610
更新存储设备的第一部分以反映新的 HA 控制器的所有权
612
启动新的 HA 控制器
614
更新存储设备的第二部分以反映新的 HA 控制器的所有权
616
结束
618

图 6