(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 101727297 B
(45)授权公告日 2012.06.27

(21)申请号 200910203112.6

(22)申请日 2009.05.27

(30)优先权数据
2008-279101 2008.10.30 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 田中胜也 岛田健太郎

(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 1848066 A,2006.10.18,
US 7290168 B1,2007.10.30,
WO 02/065290 A1,2002.08.22,
EP 1918818 A2,2008.05.07,
US 2006/0075186 A1,2006.04.06,

审查员 梁艳

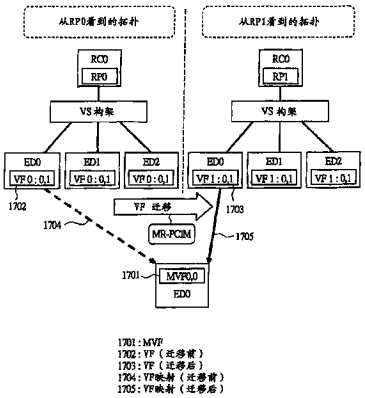
权利要求书 3 页 说明书 19 页 附图 23 页

(54)发明名称

存储装置、存储控制器内部网络的数据路径故障切换方法

(57)摘要

本发明涉及存储装置、存储控制器内部网络的数据路径故障切换方法。对存储控制器内部网络应用了 MR-IOV 的存储装置，可以执行数据路径的故障切换。在存储控制器内部网络中构成为，可以从根端口“RP0”访问各端点设备 (EDO ～ ED2) 的虚拟功能 (VF)“VF 0:0,1”。同样地，构成为可以从根端口“RP1”访问各端点设备的“VF 1:0,1”。正常时，在从“RP0”至“EDO”的第 1 数据路径中，通过 VF 映射将“VF 0:0,1”和“MVF 0,0”连接。当第 1 数据路径上发生故障时，通过由 MR-PCIM 执行 VF 迁移，在从“RP1”至“EDO”的第 2 数据路径中，通过 VF 映射将“VF 1:0,1”和“MVF 0,0”连接。结果，实现向第 2 数据路径的故障切换。



CN 101727297 B

1. 一种存储装置，具有存储控制器以及通过 PCI Express 连接了存储控制器内的组件的内部网络，其特征在于，

所述存储控制器及其内部网络具有：

总体上具有多个根端口的多个根联合体，每个根联合体具有一个以上根端口；

与每个所述根联合体连接的多个处理器；

多个端点设备，其是具备预定的接口、并具有可以从所述多个根端口的各个根端口访问的多个虚拟功能的多根已知设备；以及

交换器，其是将所述多个根端口和所述多个端点设备连接来构成数据路径、并具备多个端口的多根已知交换器，

所述多个端点设备中的至少一个，可以执行变更所述虚拟功能和所述预定的接口之间的映射的虚拟功能迁移，

所述存储控制器及其内部网络，以全部所述多个根端口可以对全部所述多个端点设备内的虚拟功能进行访问的方式，设定所述交换器以及所述端点设备，

所述存储控制器及其内部网络，作为当从所述多个根端口之一经由所述交换器到所述多个端点设备之一的数据路径之一中发生了故障时、执行数据路径的故障切换的处理，在从故障状态的第 1 数据路径向用于代替的第 2 数据路径的切换中，进行以下处理：切换访问源根端口的处理、以及通过所述虚拟功能迁移来切换访问目的地端点设备的虚拟功能和所述预定的接口之间的映射的处理，

在执行数据路径的故障切换的处理时，参照共享存储区中的包含正常时数据路径、代替数据路径、虚拟功能迁移、连累数据路径的故障切换管理表，以确认发生了故障的数据路径的切换条件。

2. 根据权利要求 1 所述的存储装置，其特征在于，

作为所述端点设备，具有：

具备用于连接上位装置的前端接口、并可以执行所述虚拟功能迁移的第 1 端点设备；和

具备用于连接存储装置群的后端接口、并可以执行所述虚拟功能迁移的第 2 端点设备，

所述交换器将所述多个根端口和所述第 1 以及第 2 端点设备连接，

所述存储控制器及其内部网络，以全部所述多个根端口可以对全部所述多个端点设备内的虚拟功能进行访问的方式，设定所述交换器以及所述第 1 以及第 2 端点设备，

所述存储控制器及其内部网络，作为当从所述多个根端口之一到所述第 1 或第 2 端点设备之一的数据路径之一中发生了故障时、执行数据路径的故障切换的处理，在从故障状态的第 1 数据路径向用于代替的第 2 数据路径的切换中，进行以下处理：切换访问源根端口的处理、以及通过所述虚拟功能迁移来切换访问目的地端点设备的虚拟功能和所述预定的接口之间的映射的处理。

3. 根据权利要求 2 所述的存储装置，其特征在于，

作为所述端点设备具有第 3 端点设备，该第 3 端点设备具备用于连接高速缓冲存储器的高速缓冲存储器接口，

所述交换器将所述多个根端口和所述第 1、第 2 以及第 3 端点设备连接，

以全部所述多个根端口可以对全部所述多个端点设备内的虚拟功能进行访问的方式，设定所述交换器以及所述第 1、第 2 以及第 3 端点设备，

作为当从所述多个根端口之一到所述第 1 或第 2 或第 3 端点设备之一的数据路径之一中发生了故障时、执行数据路径的故障切换的处理，在从故障状态的第 1 数据路径向用于代替的第 2 数据路径的切换中，进行以下处理 ：切换访问源根端口的处理、以及通过所述虚拟功能迁移来切换访问目的地端点设备的虚拟功能和所述预定的接口之间的映射的处理。

4. 根据权利要求 3 所述的存储装置，其特征在于，

所述第 3 端点设备具备对应于所述虚拟功能而被映射的预定的接口，并针对所述预定的接口对应附加了所述高速缓冲存储器接口，该第 3 端点设备中具备的预定的接口的数量在访问该第 3 端点设备的根端口的数量以上。

5. 根据权利要求 3 所述的存储装置，其特征在于，

在所述存储控制器及其内部网络中，

把成为所述第 1 端点设备的预定的接口的所述前端接口、和成为所述第 3 端点设备的预定的接口的所述高速缓冲存储器接口设定为可以从相同根端口进行访问，

把成为所述第 2 端点设备的预定的接口的所述后端接口、和成为所述第 3 端点设备的预定的接口的所述高速缓冲存储器接口设定为可以从相同根端口进行访问。

6. 根据权利要求 3 所述的存储装置，其特征在于，

所述多个根联合体分别具有第 1 根端口和第 2 根端口，

在所述存储控制器及其内部网络中，

把各个根联合体的所述第 1 根端口设定为可以对成为所述第 1 端点设备的预定的接口的所述前端接口、以及成为所述第 3 端点设备的预定的接口的所述高速缓冲存储器接口进行访问，

把各个根联合体的所述第 2 根端口设定为可以对成为所述第 2 端点设备的预定的接口的所述后端接口、以及成为所述第 3 端点设备的预定的接口的所述高速缓冲存储器接口进行访问。

7. 根据权利要求 1 所述的存储装置，其特征在于，

所述端点设备具备 ：用于与该存储装置的数据读出或写入相关的数据转发的第 1 功能、可以对与所述第 1 功能相伴的虚拟功能的状态迁移进行控制的第 2 功能。

8. 根据权利要求 7 所述的存储装置，其特征在于，

在所述存储控制器或者与其连接的维护终端装置中，具备多根 PCI 管理器的程序，

在从所述根端口到端点设备的虚拟层次的结构中，

所述端点设备，通过从所述多根 PCI 管理器的程序向所述第 2 功能的、未发生数据路径故障的虚拟层次中的带内通信，使所述第 1 功能的虚拟功能进行状态迁移。

9. 根据权利要求 7 所述的存储装置，其特征在于，

在所述存储控制器或者与其连接的维护终端装置中，具备多根 PCI 管理器的程序，

所述端点设备，通过从所述多根 PCI 管理器的程序向所述第 2 功能的带外通信，使所述第 1 功能的虚拟功能进行状态迁移。

10. 根据权利要求 7 所述的存储装置，其特征在于，

在所述存储控制器中具备单根 PCI 管理器的程序，

在从所述根端口到端点设备的虚拟层次的结构中，

所述端点设备，通过自主功能判定可否进行发生了数据路径故障的虚拟层次中的所述单根 PCI 管理器的程序与所述第 2 功能之间的通信，当不可进行通信时，使所述第 1 功能的虚拟功能进行状态迁移。

11. 根据权利要求 1 所述的存储装置，其特征在于，

所述存储控制器及其内部网络管理以下多个表：

管理所述根端口和端点设备内虚拟功能的对应关系的表、

管理所述虚拟功能和所述预定的接口之间的映射的表、以及

管理所述故障切换时的所述虚拟功能的迁移的表，

所述存储控制器及其内部网络，在执行所述数据路径的故障切换的处理时，参照所述多个表来进行以下处理：切换访问源根端口的处理、和通过所述虚拟功能迁移来切换访问目的地端点设备的虚拟功能和所述预定的接口之间的映射的处理。

12. 一种通过 PCI Express 连接了存储控制器内的组件的存储控制器内部网络的数据路径故障切换方法，其特征在于，

所述存储控制器的内部网络具有：总体上多个根端口；与具有所述根端口的根联合体连接的处理器；具备预定的接口、并具有可以从所述多个根端口的各个根端口访问的多个虚拟功能的多个端点设备；将所述多个根端口和所述多个端点设备连接来构成数据路径、并具备多个端口的交换器，

所述多个端点设备中的至少一个，可以执行变更所述虚拟功能和所述预定的接口之间的映射的虚拟功能迁移，

所述存储控制器的内部网络，以全部所述多个根端口可以对全部所述多个端点设备内的虚拟功能进行访问的方式，设定所述交换器以及所述端点设备，

所述存储控制器的内部网络，作为当从所述多个根端口之一经由所述交换器到所述多个端点设备之一的数据路径之一中发生了故障时、执行数据路径的故障切换的处理步骤，在从故障状态的第 1 数据路径向用于代替的第 2 数据路径的切换中，具有以下处理步骤：切换访问源根端口的处理步骤；通过所述虚拟功能迁移来切换访问目的地端点设备的虚拟功能和所述预定的接口之间的映射的处理步骤，

在执行数据路径的故障切换的处理时，参照共享存储区中的包含正常时数据路径、代替数据路径、虚拟功能迁移、连累数据路径的故障切换管理表，以确认发生了故障的数据路径的切换条件。

13. 根据权利要求 12 所述的存储控制器内部网络的数据路径故障切换方法，其特征在于，

所述端点设备具备：用于与该存储装置的数据读出或写入相关的数据转发的第 1 功能、以及可以对与所述第 1 功能相伴的虚拟功能的状态迁移进行控制的第 2 功能，

具有：所述端点设备根据从单根 PCI 管理器的程序或者多根 PCI 管理器的程序向所述第 2 功能的通信，使所述第 1 功能的虚拟功能进行状态迁移的处理步骤。

存储装置、存储控制器内部网络的数据路径故障切换方法

技术领域

[0001] 本发明涉及存储装置（也称为磁盘阵列装置等），尤其涉及存储控制器内部网络的数据路径（data path）的故障切换（fail-over）的技术。

背景技术

[0002] 存储装置一般具备存储控制器（也简称为控制器）和可以随机存取的非易失性记录介质。该记录介质例如是具备大量硬盘驱动器（HDD）的磁盘阵列。存储控制器具备：用于连接上位装置（主机系统等）的前端接口（front-end interface：简称为 FEI/F）、用于连接磁盘阵列的后端接口（back-end interface：简称为 BEI/F）、临时保存上位装置对磁盘阵列读写的数据的高速缓冲存储器（cache memory 简称为 CM）。存储控制器还具备用于控制上位装置和高速缓冲存储器之间、以及磁盘阵列和高速缓冲存储器之间的数据转发的处理器。

[0003] 在要求高可靠性的存储装置中，将存储控制器内的处理器、FEI/F、BEI/F 以及 CM 相互连接的网络被构成为在数据路径故障发生时可以进行故障切换。另外，在要求高性能的存储装置中，将多个处理器连接在存储控制器内部网络（简称为 CN）上，控制数据传输。

[0004] 关于具备在发生数据路径故障时可以进行故障切换的内部网络（CN）的存储系统，作为与其相关的公知技术，有特开 2008-84066 号公报（专利文献 1）。

[0005] 另外，作为与连接处理器和 FEI/F 或者 BEI/F 等的网络相关的、标准化的标准（基本规范），已知有 PCI Express（简称为 PCIe）。并且，用于由多个处理器共享 I/O 设备的“Multi-Root I/O Virtualization and Sharing Specification”（以下称为 MR-IOV）已被标准化。由于 MR-IOV 是标准规范，因此考虑到将遵从 MR-IOV 的交换器（switch）等组件进行商品化。

[0006] 即，若使用作为标准规范的 MR-IOV 来构成存储控制器内部网络（CN），则可以期待存储控制器的低成本化。

[0007] 此外，“Single-Root I/O Virtualization and Sharing Specification”（以下称为 SR-IOV）也是公知技术。在 SR-IOV 中，多个 SI（System Image）可以从一个根端口（Root Port：简称为 RP）向多个端点设备（Endpoint device：简称为 EPD）进行访问。另外，在 MR-IOV 中，把从一个 RP 观察端点而得到的树状拓扑称为虚拟层次（Virtual Hierarchy：简称为 VH）。在 MR-IOV 中，在网络内构成多个 VH，在各 VH 中，一个以上的 SI 可以从一个 RP 对多个 EPD 进行访问。

[0008] 在现有的 MR-IOV 规范中，关于 RP 和端点设备（EPD）之间的数据路径中发生了故障时的数据路径的故障切换（从故障路径向代替路径的切换），未进行考虑，没有定义具体的方法。

[0009] 因此，在对存储控制器内部网络（CN）应用了 MR-IOV 的存储装置中，存在无法执行数据路径的故障切换的问题。

[0010] 【专利文献 1】JP 特开 2008-84066 号公报

发明内容

[0011] 本发明是鉴于上述问题而提出的发明，其主要目的是，提供一种与对存储控制器内部网络 (CN) 应用了 MR-IOV 的存储装置相关的、可以执行数据路径的故障切换的技术。

[0012] 以下，简单说明本申请中公开的发明中的代表性发明的概要。为了达成所述目的，本发明的代表性的实施方式，是具有通过 PCI Express(PCIe) 连接了存储控制器 (SC) 内的组件的内部网络 (CN) 的存储装置（或系统）的技术，特征在于具有如下所示的结构。

[0013] 本实施方式的存储装置的存储控制器 (SC) 以及其内部网络 (CN) 具备：例如总体上具有多个 RP 的多个根联合体 (Root complex：RC)，每个根联合体具有一个以上（尤其多个）RP；与每个所述 RC 连接的多个处理器；多个 EPD，其是具有预定的接口（任务虚拟功能 (MVF)）、并具有可以从所述多个（全体）RP 的各个 RP 访问的多个虚拟功能 (VF) 的 MRA(Multi-Root Aware) 设备；以及交换器，其是将所述多个 RP 和所述多个 EPD 连接来构成数据路径、并具备多个端口的 MRA 交换器。并且，所述多个 EPD 中的至少一个，可以执行变更（转移）所述 VF 和所述预定的接口之间的映射（对应）的 VF 迁移（具备 VF 迁移功能）。并且，SC 以及 CN，以全部所述多个 RP 可以对全部所述多个 EPD 内的 VF 进行访问的方式，设定所述交换器以及所述 EPD。并且，SC 以及 CN，作为当从所述多个 RP 中的一个经由所述交换器到所述多个 EPD 之一的数据路径之一中发生了故障时、执行数据路径的故障切换的处理，在从该故障状态的第 1 数据路径向用于代替的第 2 数据路径的切换中，进行以下处理：切换该数据路径的访问源 RP 的处理（第 1 切换）、以及通过 VF 迁移来切换该访问目的地 EPD 的 VF 和所述预定的接口之间的映射的处理（第 2 切换）。

[0014] 在本存储装置中，例如通过 PCIe 的 MRA 交换器连接存储控制器 (SC) 内的组件。MRA 设备 (MRA 交换器）是遵从 MR-IOV 的装置。RC 具有一个以上（尤其是多个）RP。从 RC 内的 RP 通过交换器构成访问 EPD 内的 VF 的数据路径，相对于 EPD 内的 VF，作为 MVF 而映射了预定的接口（功能）。并且，在以下这样的 MRA 设备 (EPD) 中构成前端接口 (FEI/F) 以及后端接口 (BEI/F) 等的功能。即，作为 MRA 设备 (EPD)，是具有可以从 SC 内的各个 RP 访问的 VF，并且可以执行 VF 迁移（转移）的 MRA 设备。

[0015] 另外，存储控制器内部网络 (CN)，作为 MRA 设备即 EPD，包括具有 FEI/F 的第 1EPD 以及具有 BEI/F 的第 2EPD。第 1EPD 以及第 2EPD 具有可以从 SC 内的各 RP（多个）访问的多个 VF，换言之，至少具有与 VH（多个）数量相应的可以访问的 VF 资源。并且，该 EPD 被预先设定为可以执行 VF 迁移（例如具有设定 VF 迁移功能的有效 / 无效位的寄存器等，并且被设定为有效）。

[0016] 另外，在具有可以从 SC 内的各个 RP 访问的 VF 的 MRA 设备中构成高速缓冲存储器接口 (CMI/F) 等。存储控制器内部网络 (CN)，作为 MRA 设备即 EPD，包括具有 CMI/F 的第 3EPD。第 3EPD 具有可以从 SC 内的各 RP（多个）访问的多个 VF。

[0017] 并且，以 SC 内的全部 RP 可以对 FEI/F（第 1EPD) 或 BEI/F（第 2EPD) 或 CMI/F（第 3EPD) 等的各 MRA 设备 (EPD) 内的 VF 进行访问的方式，设定所述 MRA 交换器以及 MRA 设备。

[0018] 并且，当从所述 SC 内的 RP 到 EPD 的功能的数据路径中发生了故障时，通过 MR-PCIM 等进行该数据路径的访问源 RP 的切换、和访问目的地 MRA 设备内 VF 和 MVF 之间的映射的基于 VF 迁移的切换，由此执行数据路径的故障切换。

[0019] 在本实施方式中，针对每个 RP 使各个 EPD 具有可以访问的资源 (VF)。例如，针对 4 个 RP，在多个 EPD 的各个 EPD 中具有 4 个 VF。此外，VF 是在 CPU（处理器）侧虚拟可见的存在，MVF 是功能的实体。VF 和 MVF 之间的对应（映射）基于 MR-IOV 的规范。

[0020] 如上述结构那样，第 1 特征在于，在故障切换时不仅变更访问源 RP，而且通过 VF 迁移来变更访问目的地的 VF-MVF 的映射。

[0021] 第 2 特征在于，作为针对 VF 迁移控制单元的、发生数据路径故障时的对策，设置 SR-PCIM 以外的代替单元。使 EPD 具有作为该代替单元的功能。例如存在第 1 方法（基于带内 (inband) 通信的方法）、第 2 方法（基于带外 (outband) 通信的方法）、第 3 方法（基于自主功能的方法）。

[0022] 以下，简单地说明通过本申请所公开的发明中的代表性发明而得到的效果。根据本发明的代表性的实施方式，关于对存储控制器内部网络应用了 MR-IOV 的存储装置，可以执行数据路径的故障切换。在存储控制器（内部网络）中，即使从 RP 到 EPD 的数据路径中发生了故障，也可以故障切换到其它数据路径，因此确保了高可靠性。

附图说明

[0023] 图 1 是表示本发明的实施方式 1 的存储装置的块结构的图。

[0024] 图 2 是表示实施方式 1 中的端点设备 (ED0) 的块结构的图。

[0025] 图 3 是表述实施方式 1 中的端点设备 (ED1) 的块结构的图。

[0026] 图 4 是实施方式 1 中的交换器端口类型管理表的说明图。

[0027] 图 5 是实施方式 1 中的 VH 以及 VF 映射策略管理表的说明图。

[0028] 图 6 是实施方式 1 中的交换器 (SW00) 管理表的说明图。

[0029] 图 7 是实施方式 1 中的交换器 (SW01) 管理表的说明图。

[0030] 图 8 是实施方式 1 中的交换器 (SW10) 管理表的说明图。

[0031] 图 9 是实施方式 1 中的交换器 (SW11) 管理表的说明图。

[0032] 图 10 是实施方式 1 中的交换器 (SW12) 管理表的说明图。

[0033] 图 11 是实施方式 1 中的 ED0 的 VF 映射管理表的说明图。

[0034] 图 12 是实施方式 1 中的 ED1 的 VF 映射管理表的说明图。

[0035] 图 13 是实施方式 1 中的 ED2 的 VF 映射管理表的说明图。

[0036] 图 14 是实施方式 1 中的内部网络的虚拟拓扑的说明图。

[0037] 图 15 是实施方式 1 中的故障切换管理表的说明图。

[0038] 图 16 是实施方式 1 中的故障切换动作（向代替数据路径的切换）的说明图。

[0039] 图 17 是实施方式 1 中的故障切换动作 (VF 迁移）的说明图。

[0040] 图 18 是表示实施方式 1 中的故障切换处理的流程图。

[0041] 图 19 是实施方式 1 中的 ED0 的 VF 映射管理表（之二）的说明图。

[0042] 图 20 是实施方式 1 中的 ED2 的 VF 映射管理表（之二）的说明图。

[0043] 图 21 是表示本发明的实施方式 2 的存储装置中的端点设备的块结构的图。

[0044] 图 22A 是表示实施方式 2 中的第 1 方法中的 VF 迁移处理（之一）的流程图。

[0045] 图 22B 是表示实施方式 2 中的第 1 方法中的 VF 迁移处理（之二）的流程图。

[0046] 图 23 是表示实施方式 2 中的第 2 方法中的 VF 迁移处理的流程图。

[0047] 图 24 是表示实施方式 2 中的第 3 方法中的 VF 迁移处理的流程图。

[0048] 图 25 是与基于 MR-IOV 规范的 VF 迁移相伴的 VF 的状态迁移的说明图。

[0049] 图 26 是表示一个实施方式中的存储装置的利用例（之一）的说明图。

[0050] 图 27 是表示一个实施方式中的存储装置的利用例（之二）的说明图。

[0051] 图 28 是表示一个实施方式中的存储装置的利用例（之三）的说明图。

[0052] 符号说明

[0053] 10 :存储装置 ;15 :磁盘阵列 (DA) ;20 :存储控制器 (SC) 及其内部网络 (CN) ;21 :处理器 (CPU0、CPU1) ;22 :存储器 (MEM0、MEM1) ;23 :高速缓冲存储器 (CM) ;30 :根联合体 (RC)(RC0、RC1) ;40 :交换器 (SW00、SW01、SW10、SW11、SW12) ;50、51 :端点设备 (EPD)(ED0、ED1、ED2) ;102 :前端接口 (FEI/F) ;103、105、107 :组件 ;104 :高速缓冲存储器接口 (CMI/F) ;106 :后端接口 (BEI/F) ;108 :程序 (MR-PCIM 以及 SR-PCIM) ;109 :程序 (MR-PCIM) ;110 :故障切换控制程序 (FP) ;111 :故障切换控制表 (FT) ;190 :维护终端 (SVP) ;191 :管理端口 ;201、301 :端口 ;202、302 :基本功能 (BF) ;203 ;303 :物理功能 (PF) ;204、304 :虚拟功能 (VF) ;205、305 :内部路由控制部 ;206、306 :VF 映射 (VF Mapping) 控制部 ;207、307 :任务虚拟功能 (MVF) ;208、308 :管理端口 ;400 :交换器端口类型管理表 ;500 :VH 及 VF 映射策略管理表 ;600、700、800、900、1000 :交换器管理表 ;1100、1200、1300、1100b、1300b :VF 映射 (VF Mapping) 管理表 ;1401 :根端口 ;1402 :虚拟交换器架构 (VSF) ;1403、1404、1405 :MR-Link ;1406 :VF ;1407 :VF 映射 (VF Mapping) ;1500 :故障切换管理表 ;1701 :MVF ;1702、1703 :VF ;1704、1705 :VF 映射 ;2102 :BF ;2103 :PF ;2104 :VF 状态控制部 ;2105 :管理端口 ;RP0、RP1、RP2、RP3 :根端口 ;P000、P001、P002、P003、P004、P010、P011、P012、P013、P014、P100、P101、P102、P110、P111、P112、P120、P121、P122 :端口 ;N00 :网络 ;C00、D00 :通道 ;VS0、VS 1 :虚拟交换器 (VS) ;K1、K2、K3、K4 :数据路径 ;L1、L2、L3、L4、L5 :链路（交换器间链路）

具体实施方式

[0054] 以下，根据附图详细说明本发明的实施方式。此外，在用于说明实施方式的全部附图中，对于相同部分原则上赋予相同符号，省略其重复的说明。

[0055] （实施方式 1）

[0056] 使用图 1 ～图 20 等，说明本发明的实施方式 1 的存储装置（换言之，存储系统）。在实施方式 1 中，作为概要，根据图 1、图 2 等的结构，如图 16、图 17 等所示，在存储控制器的内部网络中从正常数据路径（故障状态）向代替数据路径的故障切换动作时，变更访问源的根端口（例如 RP0 → RP1），并且通过变更端点设备的虚拟功能 (VF) 和任务虚拟功能 (MVF) 之间的映射 (VF 映射)（例如 ED0 的“VF 0 :0,1”-“MVF 0 :0”→“VF 1 :0,1”-“MVF0 :0”）来实现 VF 迁移的控制。

[0057] < 存储装置 >

[0058] 图 1 表示作为本发明的实施方式的存储装置 10 的块结构。存储装置 10 具备存储控制器 (SC)20 和磁盘阵列 (DA)15。DA15 由多个硬盘驱动器 (HDD) 或半导体存储器驱动器等存储装置群构成。另外，对于 SC20，经由通道 C00 连接外部的上位装置（主机系统）等。另外，对于 SC20，经由网络 N00 连接维护终端 (maintenance terminal :设为 SVP)190。

[0059] 存储控制器 (SC)20 及其内部网络 (CN) 具备 :2 个处理器 21(CPU0、CPU1) ;2 个存储

器 22(MEM0、MEM1) ;2 个根联合体 (RC)30(RC0、RC1) ;3 个端点设备 (EPD)50(ED0、ED1、ED2) ;1 个高速缓冲存储器 (CM)23 ;以及多个 (5 个) 交换器 40(SW00、SW01、SW10、SW11、SW12)。通过上述组件的基于交换器 (switch) 的相互连接 (PCIe 架构) 而构成了内部网络 (CN)。

[0060] 此外,在 SC20 以及 CN 中,将处理器 21、存储器 22 等至少二重化,相同符号组件是左右对称的结构。另外,构成 SC20(以及 CN) 的组件 (交换器等) 的数量也可以不限于图 1 的结构例。例如可以增加 RC30、SW40、EPD50 的数量。

[0061] 作为处理器 21 的 CPU0(CPU1) 通过执行在作为存储器 22 的 MEM0(MEM1) 中存储的程序来进行各种处理。具体而言,CPU0(CPU1) 控制与 SC20 相连的上位装置 (主机系统) 和 DA15 之间的数据转发。作为存储器 22 的 MEM0(MEM1),存储通过作为处理器 21 的 CPU0(CPU1) 执行的程序、或者被参照的表的数据等。

[0062] 在本例中,在各存储器 22 中存储了 :单根 PCI 管理程序 (SR-PCIM) 以及多根 PCI 管理程序 (MR-PCIM) 双方的程序 108、故障切换控制程序 (FP)110、和各种管理信息的表 (后述)。另外,在 CM23 的一部分区域中也存储包含故障切换控制表 (FT)111 的各种管理信息的表。另外,在维护终端 (SVP)190 中也存储 MR-PCIM 的程序 109 等程序或表。

[0063] 各种表的信息可以由管理者 (Administrator) 设定或变更等。管理者例如可以操作维护终端 (SVP)190 来进行包含该设定等的维护 •管理系列的操作。在这种情况下,管理者将与表的设定等相关的信息输入 SVP190。SVP190 将输入的信息经由网络 N00 发送至处理器 21(CPU0(CPU1))。该处理器 21 根据接收到的信息,生成对象表或对其进行变更等。然后,该处理器 21 将该表存储在存储器 22(MEM0(MEM1)) 等中。

[0064] 第 1 根联合体 30(RC0) 与第 1 处理器 21(CPU0) 连接,并且具备 2 个 PCIe 基本规范的根端口 (第 1RP、第 2RP) 即 RP0、RP1。同样地,第 2 根联合体 30(RC1) 与第 2 处理器 21(CPU1) 连接,并且具备 2 个根端口 (第 1RP、第 2RP) 即 RP2、RP3。

[0065] 第 1 根端口 (RP0),经由第 1 交换器 40(SW00) 以及第 3 交换器 40(SW10) 与第 1 端点设备 50(ED0) 相连。另外,RP0 经由 SW00 以及第 4 交换器 40(SW11) 与第 3 端点设备 50(ED1) 相连。而且,RP0 经由 SW00 以及第 5 交换器 40(SW12) 与第 2 端点设备 50(ED2) 相连。

[0066] 第 2 根端口 (RP1),经由第 2 交换器 40(SW01) 以及第 3 交换器 40(SW10) 与第 1 端点设备 50(ED0) 相连。另外,RP1 经由 SW01 以及第 4 交换器 40(SW11) 与第 3 端点设备 50(ED1) 相连。而且,RP1 经由 SW01 以及第 5 交换器 40(SW12) 与第 2 端点设备 50(ED2) 相连。

[0067] 第 3 根端口 30(RP2)(RC1 内的第 1RP),经由第 1 交换器 40(SW00) 以及第 3 交换器 40(SW10) 与第 1 端点设备 50(ED0) 相连。另外,RP2 经由 SW00 以及第 4 交换器 40(SW11) 与第 3 端点设备 50(ED1) 相连。而且,RP2 经由 SW00 以及第 5 交换器 40(SW12) 与第 2 端点设备 50(ED2) 相连。

[0068] 第 4 根端口 (RP3)(RC1 内的第 2RP),经由第 2 交换器 40(SW01) 以及第 3 交换器 40(SW10) 与第 1 端点设备 50(ED0) 相连。另外,RP3 经由 SW01 以及第 4 交换器 40(SW11) 与第 3 端点设备 50(ED1) 相连。而且,RP3 经由 SW01 以及第 5 交换器 40(SW12) 与第 2 端点设备 40(ED2) 相连。

[0069] 交换器 (SW00、SW01、SW10、SW11、SW12) 是遵从 MR-IOV 规范的多根已知 (MRA :

Multi-RootAware) 交换器。另外，端点设备 50(ED0、ED1、ED2) 是 MRA 设备。第 1 和第 2 交换器 (SW00、SW01) 与 RC30 侧相连。第 3 ～第 5 交换器 (SW10、SW11、SW12) 与 EPD50 侧相连。

[0070] 第 1 交换器 40(SW00) 具备用于构成数据路径的端口 P000、P001、P002、P003、P004，还具备可以与网络 N00 连接的管理端口 (management port)191。同样地，第 2 交换器 40(SW01) 具备端口 P010、P011、P012、P013、P014 和管理端口 191。经由网络 N00 将 SVP190 与交换器 40 的管理端口 191 连接，由此，管理者可以设定、管理 PCIe 架构。

[0071] 第 3 交换器 40(SW10) 具备用于构成数据路径的端口 P100、P101、P102。同样地，第 4 交换器 40(SW11) 具备端口 P110、P111、P112。第 5 交换器 40(SW12) 具备端口 P120、P121、P122。3 个交换器 40(SW10、SW11、SW12) 与两个交换器 40(SW00、SW01) 同样经由网络 N00 与 SVP190 相连。

[0072] 第 1 端点设备 50(ED0) 经由通道 C00 与外部的上位装置相连。上位装置例如是在本存储装置 10 内读写数据的计算机。ED0 具备前端接口 (FEI/F)102。FEI/F102 对通道 C00 上与 SC20 内的数据转发协议进行相互变换。此外，SW10 和 ED0 可以集中作为 1 个组件 103 被安装。

[0073] 第 2 端点设备 50(ED2) 经由通道 D00 与 DA15 相连。ED2 具备后端接口 (BEI/F)106。BEI/F106 对通道 D00 上与 SC20 内的数据转发协议进行相互变换。此外，SW12 和 ED2 可以集中作为 1 个组件 107 被安装。

[0074] 第 3 端点设备 50(ED1) 具备用于与高速缓冲存储器 (CM)23 连接的高速缓冲存储器接口 (CMI/F)104。CM23 临时存储 (高速缓冲存储) 从第 1 端点设备 50(ED0) 以及第 2 端点设备 50(ED2) 等接收到的数据 (在与 FEI/F102 或 BEI/F106 之间的转发数据)。CM23 被共用于 SC20 内的数据或信息的存储。另外，在 CM23 内的存储区域的一部分中确保的共享存储区域中，存储控制信息等。此外，SW11 和 ED1 也可以集中作为 1 个组件 105 被安装。

[0075] EPD50(ED0、ED1、ED2) 经由网络 N00 与 SVP190 相连。SVP190 将通过存储装置 10 的管理者输入的设定信息发送至 EPD50(ED0、ED1、ED2)。

[0076] 此外，SC20 可以代替 2 个作为端点设备 50 的 ED0(搭载了 FEI/F102) 以及 ED2(搭载了 BEI/F106)，而具备 1 个端点设备。在这种情况下，该 1 个端点设备进行该两个 EPD(ED0 以及 ED2) 的对应处理。

[0077] MR-PCIM(Multi-Root PCI Manager) 具有设定、管理 PCIe 架构 (内部网络 (CN)) 等功能。另外，与本特征相关，MR-PCIM 具有控制数据路径的故障切换等功能。

[0078] MR-PCIM 的程序，例如被存储在存储器 22 中 (程序 108)，在处理器 21 上运行。或者，MR-PCIM 的程序可以在 SVP190 上运行 (程序 109)，并经由网络 N00 和交换器 40 的管理端口 191 来设定、管理 PCIe 架构。

[0079] SR-PCIM(Single-Root PCI Manager) 在每个根端口的虚拟层次 (VirtualHierarchy :VH) 内运行。SR-PCIM 的程序例如被存储在存储器 22 中 (程序 108)，在处理器 21 上运行。

[0080] 控制 SC20(CN) 中的数据路径的故障切换的故障切换控制程序 (FP)110，例如被存储在存储器 22 中，在处理器 21 上运行。FP110，可以作为 MR-PCIM 的功能而通过内置的方式被安装，或者也可以作为独立的程序被安装。但是，将 FP110 安装为可以直接或间接地与

MR-PCIM 以及 SR-PCIM 通信。

[0081] 在本系统 (SC20) 中，运行 MR-PCIM 和 SR-PCIM 两者。SR-PCIM（在各 VH 内运行），管理 EPD 的 SR-IOV 特性。MR-PCIM，管理 EPD 的 MR-IOV 特性。VF 和 MVF 之间的对应（映射）由 MR-PCIM 管理。MR-PCIM 和 SR-PCIM 协作来控制 VF 迁移。

[0082] MR-PCIM 基于 MR-IOV 规范。在本例中，在两个处理器 21(CPU0、CPU1) 以及 SVP190 这 3 个位置分别配置 MR-PCIM 的程序 (108、109)。根据情况，其中一个活动地 (Active) 运行，其它两个用于备份运行。

[0083] 作为在 CM23 的一部分中设置的共享存储区域中存储的信息之一，有故障切换控制表 (FT)111。从处理器 21（故障切换控制程序 (FP)110) 访问、参照 FT111。在处理器 21 侧的 RP 间无法通信，因此作为可以从双方处理器 21(CPU0、CPU1) 侧的 RP 访问的位置，故障切换控制程序 (FR)110 在本例中被配置在 CM23 内。关于故障切换动作，由故障切换控制程序 (FP)110 参照 FT111 来自动执行。

[0084] <端点设备 ED0>

[0085] 图 2 表示第 1 端点设备 50(ED0) 的块结构。ED0 具备 :可以与 MR-Link(Multi-Root Link)（交换器 40 侧）连接的端口 201；内部路由控制部 (internal rooting controller)205、各功能 (BF202、PH203、VF204) ;VF 映射 (VF Mapping) 控制部 206 ;FEI/F102 的任务虚拟功能 (MVF)207 ;以及可以经由网络 N00 与 SVP190 通信的管理端口 208 等。

[0086] ED0 具备功能号码 0 的一个基本功能 202(“BF 0 :0”)（基本功能 :简称为 BF)。另外，ED0 具备 BF202（“BF 0 :0”）附带的多个 (4 个）物理功能 203（“PF 0 :0”，“PF 1 :0”，“PF 2 :0”，“PF 3 :0”)（物理功能 :简称为 PF)。而且，ED0 针对每个 PF，具备各 PF 附带的多个 (4 个）虚拟功能 204（“VF 0 :0,1”，“VF 1 :0,1”，“VF 2 :0,1”，“VF 3 :0,1”)（虚拟功能 :简称为 VF)。通过这些功能 (BF、PF、VF) 支持 4 个 VH，由此，ED0 可以设定为能够从 SC20 内设置的全部 4 个根端口 (RP0、RP1、RP2、RP3) 进行访问。例如，“PF 0 :0”和“VF 0 :0,1”对应于第 1VH(VH0)。

[0087] BF202 存储了基本的设定。VF204 与 MVF107 被映射。VF204 等是用于与存储装置 10 的数据读出或写入等相关联的数据转发等的功能。BF 是管理与 MRA 设备的 MR-IOV 相关的特性的功能，还保持 VF 和 MVF 的映射信息。PF 是管理与 SR-IOV 规范相关的设备的特性的功能。VF 是 PF 附带的功能，附带的 PF 以及相同 PF 附带的其它 VF 共享 1 个以上的链路等物理资源。在本例的端点设备中，功能号码 0 的 BF、PF、VF 负责端点设备的 FEI/F 等接口功能。

[0088] 多个 (2 个)MVF207（“MVF 0,0”，“MVF 0,1”) 是 FEI/F102 的功能的实体（任务虚拟功能 :简称为 MVF)。ED0 在本例中具备 2 个 MVF207，由此可以同时并行地执行最多 2 个接口 (FEI/F) 功能。

[0089] VF 映射控制部 206，根据在 BF202 中存储的设定，将多个 VF204（“VF0 :0,1”，“VF 1 :0,1”，“VF 2 :0,1”，“VF 3 :0,1”) 中最多 2 个 VF204 与 MVF207（“MVF 0,0”，“MVF 0,1”)1 对 1 地连接（映射）。

[0090] 说明从端口 201 至通道 C00 的数据流。ED0 将端口 201 接收到的数据向内部路由控制部 205 转发。内部路由控制部 205 按照接收到的数据的发送目的地，向各功能进行路径变更。VF 映射控制部 206 从 VF204 向 MVF207 传输数据。MVF207 向通道 C00 转发数据。

从通道 C00 至端口 201 的数据流是上述的反方向。

[0091] 此外，关于第 2 端点设备 50(ED2) 的结构，除了 MVF 的功能是与通道 D00 连接的 BEI/F106 以外，与图 2 的 ED1 的结构大体相同。

[0092] 此外，关于功能的表述，例如在“BF a :b”中，a 是 VH 号码，b 是功能号码。另外，在“PF a :b”中，a 是 VH 号码，b 是功能号码。另外，在“VF a :b，c”中，a 是 VH 号码，b 是功能号码，c 是 VF 号码。另外，在“MVF a，b”中，a 是功能号码，b 是 MVF 号码。

[0093] < 端点设备 ED1>

[0094] 图 3 表示第 3 端点设备 50(ED1) 的块结构。ED1 具备 :可以与 MR-Link(交换器 40 侧) 连接的端口 301、内部路由控制部 305、各功能 (BF302、PF303、VF304)、VF 映射 (VF Mapping) 控制部 306、CMI/F104 的任务虚拟功能 (MVF)307、以及可以经由网络 N00 与 SVP190 通信的管理端口 308 等。

[0095] ED1 具备功能号码 0 的 1 个 BF302(“BF 0 :0”)。另外，ED1 具备 BF302(“BF 0 :0”) 附带的多个 (4 个)PF303(“PF 0 :0”,“PF1 :0”,“PF 2 :0”,“PF3 :0”)。而且，ED1 针对每个 PF 具备各 PF 附带的多个 (4 个)VF304(“VF 0 :0,1”,“VF 1 :0,1”,“VF 2 :0,1”,“VF 3 :0,1”)。通过它们支持 4 个 VH，由此，ED1 可以设定为能够从 SC20 内设置的 4 个根端口 (RP0、RP1、RP2、RP3) 进行访问。

[0096] MVF307(“MVF 0,0”,“MVF 0,1”,“MVF 0,2”,“MVF 0,3”是 CMI/F104 的功能的实体。ED1 具备 4 个 MVF307，由此可以同时并行地执行最多 4 个接口 (CMI/F) 功能。

[0097] VF 映射控制部 306 根据在 BF302 中存储的设定，将多个 VF304(“VF0 :0,1”,“VF 1 :0,1”,“VF 2 :0,1”,“VF 3 :0,1”) 中最多 4 个 VF304 与 MVF307(“MVF 0,0”,“MVF 0,1”,“MVF 0,2”,“MVF 0,3”)1 对 1 地连接。在搭载 CMI/F104 的 ED1 中，可以将全部的 VF304 与 MVF307 映射。

[0098] 说明从端口 301 至 CM23 连接通道的数据流。ED1 将端口 301 接收到的数据转发至内部路由控制部 305。内部路由控制部 305 按照接收到的数据的发送目的地，向各功能进行路径变更。VF 映射控制部 306 从 VF304 向 MVF307 传输数据。MVF307 向 CM23 连接通道转发数据。从 CM23 连接通道至端口 301 的数据流是上述的反方向。

[0099] 作为映射到 VF 的 MVF(对应了 CMI/F104) 的数量 (x)，ED1 具有访问 ED1 的 RP 的数量 (y :在本例中是 4) 以上 (在本例中，x ＝ y ＝ 4)。

[0100] < 初始设定 >

[0101] 接下来，顺次说明存储了构成 SC20 的内部网络 (CN) 的 PCIe 架构的初始设定信息等的各种表。各种表包括 :图 4 的交换器端口类型管理表 400 ;图 5 的 VH 以及 VH 映射策略管理表 500 ;图 6 ～图 10 的每个交换器 (SW00 ～ SW12) 的交换器管理表 600、700、800、900、1000 ;图 11 ～图 13 的每个 EPD50(ED0、ED1、ED2) 的 VF 映射管理表 1100、1200、1300。

[0102] < 交换器端口类型管理表 >

[0103] 图 4 是交换器端口类型管理表 400 的说明图。交换器端口类型管理表 400，作为项目 (列) 而存储 SC20 内的交换器端口 (“Switch Port”)a(所述交换器 400 的端口号码、例如“P000”等) 与其端口类型 (“PCIM Capable”)b(PCIM 能够) 的对应。MR-PCIM 经由端口类型 b 为“Yes”(是) 的交换器端口将 PCIe 架构初始化。仅连接 EPD50 的交换器端口 a 的端口类型 b 为“No”(否)。

[0104] <VH 以及 VF 映射策略管理表 >

[0105] 图 5 是 VH 以及 VF 映射策略管理表 500 的说明图。VH 以及 VF 映射策略管理表 500，作为项目而具有“VH”a、“Authorized(认可)”b、“VH Mapping”(VH 映射)c{“VS in”d、“Inter-SW Link”e、“VH in”f}、“VF Mapping”(VF 映射)g。

[0106] 将各 VH 的 RP(RP 号码) 存储在 a 中。在 a 中存储的 RP 上，若 MR-PCIM 运行，则将“Authorized”b 设为“Yes”。在本实施方式中，为了确保冗余性，设定为 MR-PCIM 可以在全部 RP 上运行。

[0107] c 存储 VH 映射 (VH Mapping) 设定信息。d 存储各 VH 中使用的交换器 40 内虚拟交换器 (VS)。在本例中，在全部 VH 中使用 EPD50 连接侧的 3 个交换器 40(SW10、SW11、SW12) 内的虚拟交换器 (VS)。例如，在 RP0 的 VH 中设定了 SW00 和 3 个交换器 (SW10、SW11、SW12) 的 VS。

[0108] 此外，虚拟交换器 (VS) 是基于 MR-IOV 规范的术语。在交换器设备 40 内对应设置了 1 个以上 VS。例如在 SW00 内具有 VS0、VS1 这两个 VS。

[0109] e 存储在各 VH 中使用的交换器间链路 (inter-switch link) 的一览。在本例中，各 VH 使用 EPD50 连接侧的 3 个交换器 40(SW10、SW11、SW12) 各自具有的 2 条交换器间链路中的 1 条的链路。例如在 RP0 的 VH 中，作为交换器间链路，设定了 P002-P100 的链路 (SW00-SW10 间的路径)、P003-P110 的链路 (SW00-SW11 间的路径)、P004-P120 的链路 (SW00-SW12 间的路径)。

[0110] f 存储属于各 VH 的 EPD50 的一览。在本例中，全部 EPD50(ED0、ED1、ED2) 属于各 VH 内。

[0111] g 存储 VF 映射 (VF Mapping) 设定信息。在本例中，在各 VH 中对每个 EPD50 分配了 1 个 VF。此外，分配给各 VH 的每个 EDP50 的 VF 的个数不限于 1 个，是 1 个以上即可。在本例中，各 EPD50 支持 VH 的个数 (连接的根端口数) 以上的 VF。

[0112] < 每个交换器的交换器管理表 >

[0113] 图 6 ～图 10 是 5 个交换器 40(SW00 ～ SW12) 的各自的交换器管理表的说明图。图 6 是第 1 交换器 40(SW00) 的交换器管理表 600。图 7 是第 2 交换器 40(SW01) 的交换器管理表 700。

[0114] 交换器管理表，作为项目而具有“Slot(时隙)”a、“Root(根)”b、“VS”c、“Bridge(桥)”d、“Enable(使能)”e、“Mapped(映射)”f、“Port(端口)”g、“Port VHN(端口 VH 号码)”h。

[0115] 在本例中，如行 601、603、701、703 所示，4 个 RP(RP0、RP1、RP2、RP3) 是 PCIe 基本规范的 RP，因此各交换器 40 中与 RP 连接的端口的“PortVHN”h(端口 VH 号码) 成为“VH0”。端口 VH 号码是与交换器 40 的各端口连接的链路中使用的 VH 号码 (基于 MR-IOV 的规范的术语)。

[0116] 在本例中，在 SW00 的交换器管理表 600 中，与 RP0 连接的虚拟交换器 VS0 的下游 (downstream) 侧端口的“Port VHN”h 被映射为“VH0”(602)。另外，与 RP2 连接的虚拟交换器 VS 1 的下游侧端口的“Port VHN”h 被映射为“VH1”(604)。

[0117] 在本例中，在 SW01 的交换器管理表 700 中，与 RP1 连接的虚拟交换器 VS0 的下游侧端口的“Port VHN”h 被映射为“VH0”(702)。另外，与 RP3 连接的虚拟交换器 VS1 的下游

侧端口的“Port VHN”h 被映射为“VH1”(704)。

[0118] 此外，关于“Bridge”d，所谓下游侧，在图 1 中是指 SW00、SW01 的端点设备侧，所谓上游 (upstream) 侧，在图 1 中是指 SW00、SW01 的根端口侧。

[0119] 图 8 是第 3 交换器 40(SW10) 的交换器管理表 800。在本例中，在将 EPD50 和交换器 40(SW10，SW11，SW12) 连接的链路中，把与将 RP0 设为根端口的 VH 相对应的 VH 号码设为 VH0。同样地，对应于 RP1 设为 VH1，对应于 RP2 设为 VH2，对应于 RP3 设为 VH3。在本例中，SW10 的下游侧端口 P102，在 RP 为 RP0 时将“Port VHN”h 映射为“VH0”(801)。另外，端口 P102 的“Port VHN”h，在 RP 为 RP1 时被映射为“VH1”(802)，在 RP 为 RP2 时被映射为“VH2”(803)，在 RP 为 RP3 时被映射为“VH3”(804)。

[0120] 图 9 是第 4 交换器 40(SW11) 的交换器管理表 900。在本例中，SW11 的下游侧端口 P112，在 RP 为 RP0 时将“Port VHN”h 映射为“VH0”(901)。另外，端口 P112 的“Port VHN”h，在 RP 为 RP1 时被映射为“VH1”(902)，在 RP 为 RP2 时被映射为“VH2”(903)，在 RP 为 RP3 时被映射为“VH3”(904)。

[0121] 图 10 是第 5 交换器 40(SW12) 的交换器管理表 1000。在本例中，SW12 的下游侧端口 P122，在 RP 为 RP0 时，将“Port VHN”h 映射为“VH0”(1001)。另外，端口 P122 的“Port VHN”h，在 RP 为 RP1 时被映射为“VH1”(1002)，在 RP 为 RP2 时被映射为“VH2”(1003)，在 RP 为 RP3 时被映射为“VH3”(1004)。

[0122] 到图 10 为止的设定，是系统硬件初始设定的值，在本系统的设计、制造阶段已被预先设定。另外，管理者根据具体的系统利用例，设定图 11 以后的表的必要项目的值。

[0123] < 每个 EPD 的 VF 映射管理表 (1)>

[0124] 图 11 ～图 13 是每个 EPD50(ED0、ED1、ED2) 的 VF 映射 (VF Mapping) 管理表 (1100、1200、1300) 的说明图。这些 VF 映射管理表被存储在每个 EPD50 的 BF 内，通过 MR-PCI 被设定、管理。

[0125] 图 11 是第 1EPD50(ED0)(搭载 FEI/F102) 的 VF 映射管理表 1100 的说明图。VF 映射管理表 1100，作为项目而具有“PF”a、“LVF#”b、“VF State(VF 状态)”c、“VF”d、“MVF”e。d 和 e 表示 VF 和 MVF 的映射。“LVF#”b 中的 LVF 是逻辑 VF(基于 MR-IOV 规范的术语)。在 BF 内，VF 作为 LVF 被管理。根据 MR-IOV 规范，“LVF#”b、“VF State(VF 状态)”c、“MVF”e 的对应，作为 LVF 表被存储在 BF 内。LVF 表不保存与 VH 相关的信息。为了进行 VF 迁移控制，虽然 LVF 表的信息充分，但在本系统中通过在 LVF 表中追加“PF”a 以及“VF”d 的项目，将 LVF 表扩展成追加了 VH 信息的 VF 映射管理表。其结果是，本例可以容易地判断 LVF 以及 MVF 可以从哪个根端口来访问。

[0126] 属于 VH0 的 VF(“VF 0 :0,1”)，其“VF State”c 为“Active.Available(A.A)”(活动 . 可用)，被映射到 MVF(“MVF 0,0”)(1101)。属于 VH1 的 VF(“VF1 :0,1”)，其“VF State”c 为“Inactive.Unavailable(I.U)”(不活动 . 不可用)，没有被映射到 MVF(“none”(无))(1102)。属于 VH2 的 VF(“VF 2 :0,1”)，其“VF State”c 为“A.A”，被映射到 MVF(“MVF 0,1”)(1103)。属于 VH3 的 VF(“VF 3 :0,1”)，其“VF State”c 为“I.U”，没有被映射到 MVF(“none”)(1104)。这些 VF 映射的结果是，ED0 在初始状态下可以从 RP0 以及 RP2 进行访问，不可以从 RP1 以及 RP3 进行访问。

[0127] 图 12 是第 3EPD50(ED1)(搭载 CMI/F104) 的 VF 映射管理表 1200 的说明图。属于

VH0 的 VF(“VF 0 :0,1”),其“VF State”c 为“A.A”,被映射到 MVF(MVF 0,0)(1201)。属于 VH1 的 VF(“VF 1 :0,1”),其“VF State”c 为“A.A”,被映射到 MVF(“MVF 0,1”)(1202)。属于 VH2 的 VF(“VF 2 :0,1”),其“VF State”c 为“A.A”,被映射到 MVF(“MVF 0,2”)(1203)。属于 VH3 的 VF(“VF 3 :0,1”),其“VF State”c 为“A.A”,被映射到 MVF(“MVF 0,3”)(1204)。这些 VF 映射的结果是,ED1 在初始状态下可以从全部 4 个 RP(RP0、RP1、RP2、RP3) 进行访问。

[0128] 在搭载 CMI/F104 的 ED1 的表设定中,特别地,VF 状态全部为“A.A”。在 CMI/F104 中与 FEI/F 102 以及 B EI/F106 不同,因为成为具有可以从全部 RP 访问的 VF 资源的结构。

[0129] 图 13 是第 2EPD50(ED2) 的 VF 映射管理表 1300 的说明图。基本上,与 ED1 的情况相同。属于 VH0 的 VF(“VF 0 :0,1”),其“VF State”c 为“I.U”,未被映射到 MVF(1301)。属于 VH1 的 VF(“VF 1 :0,1”),其“VF State”c 为“A.A”,被映射到 MVF(“MVF 0,0”)(1302)。属于 VH2 的 VF(“VF 2 :0,1”),其“VF State”c 为“1.U”,未被映射到 MVF(1303)。属于 VH3 的 VF(“VF3 :0,1”),其“VF State”c 为“A.A”,被映射到 (“MVF 0,1”)(1304)。这些 VF 映射的结果是,ED2 在初始状态下可以从 RP1 以及 RP3 进行访问,不可以从 RP0 以及 RP2 进行访问。

[0130] < 设定以及利用例 >

[0131] 在本实施方式中,根据以上初始设定,通过管理者进行的设定,作为利用例,在 RP0 和 RP2 为 RP 的 VH 中执行 ED0(搭载 FEI/F102) 和 ED1(搭载 CMI/F104) 之间的数据转发,在 RP1 和 RP3 为 RP 的 VH 中执行 ED2(搭载 BEI/F106) 和 ED1(搭载 CMI/F104) 之间的数据转发。图 26 等表示上述那样的利用例。

[0132] 在图 26 的利用例中,在内部网络 (CN) 中,作为数据路径,将 ED0 的 MVF(FEI/F102) 和 ED1 的 MVF(CM1/F104) 设定为可以从相同 RP(例如 RP0) 访问,并且将 ED2 的 MVF(BEI/F106) 和 ED1 的 MVF(CMI/F104) 设定为可以从相同 RP(例如 RP3) 访问。因此,ED1 具备访问 ED1 的根端口数以上的 MVF,ED0 和 ED2 具备访问 ED1 的根端口数的半数以上的 MVF。结果,ED0 的 MVF(FEI/F102) 和 ED1 的 MVF(CMI/F104) 之间,以及 ED2 的 MVF(BEI/F106) 和 ED1 的 MVF(CMI/F104) 之间可以进行数据转发。

[0133] 另外,在内部网络 (CN) 中,分别与处理器 21 连接的 2 个 RC30(RC0、RC1) 分别具有 2 个 RP(第 1RP、第 2RP),作为数据路径,各 RC30 的第 1RP(RP0、RP2) 被设定为可以对 ED0 的 MVF(FEI/F102) 以及 ED1 的 MVF(CMI/F104) 进行访问,并且,各 RC30 的第 2RP(RP1、RP3) 被设定为可以对 ED2 的 MVF(BEI/F106) 以及 ED1 的 MVF(CMI/F104) 进行访问。

[0134] 例如,在与 CPU0 连接的 RC0 中,可以从 RP0 对 FEI/F102(ED0) 和 CMI/F104(ED1) 进行访问,可以从 RP1 对 BE1/F106(ED2) 和 CMI/F104(ED1) 进行访问。

[0135] 通过本利用例,可以同时并行地、高效地处理使用了 FEI/F102 和 CMI/F104 的数据转发 (例如来自上位装置的数据输入输出)、和使用了 BEI/F 106 和 CMI/F104 的数据转发 (例如针对 DA15 的数据输入输出)。

[0136] 在本利用例的存储装置 10 中,经由 CM23 进行 FEI/F102 和 BEI/F106 之间的数据转发。因此,在 FEI/F102 侧和 BEI/F106 侧应用 VH 不同的所述设定例那样的 VH 的分配。另外,CPU0 使用 RC0 的两个 RP(RP0、RP1),另外,CPU1 使用 RC1 的两个 RP(RP2、RP3),由此,各处理器 21 可以控制 FEI/F102 侧和 BEI/F106 侧的双方的数据转发。在负荷分散以及冗余

性方面，也将其应用于存储装置。

[0137] ＜内部网络的拓扑＞

[0138] 图 14 中表示了以上说明的 PCIe 架构的初始设定后的、SC20 内部网络 (CN) 的虚拟拓扑。全部 RP1401 经由虚拟交换器架构 (VS Fabric)1402（对应于交换器 40(VS)）与 EPD50(ED0、ED1、ED2) 内的、分配给各个 VH 的 VF1406（面向 RP0 的 VF、面向 RP1 的 VF、面向 RP2 的 VF、面向 RP3 的 VF) 连接。VSW Fabric1402 和各 EPD50(VF1406) 之间，通过 MR-Link(1403、1404、1405) 连接。VF 1406 通过 VF 映射 (VF Mapping)1407 与各 EPD50 的 MVF 1408（“MVF 0,0”,“MVF 0,1”等）连接。在本例的 PCIe 架构中，针对每个 EPD50 具备分配给各 RP 的（分配给各 VH) 的 VF，各 RP 可以独立地访问各 EPD50 内 VF。但是实际上能够使用 EPD50 的功能的 VF 仅仅是通过 VF 映射与 MVF 连接的 VF。

[0139] ＜故障切换管理表＞

[0140] 图 15 是故障切换管理表 1500（所述 FT111) 的说明图。故障切换管理表 1500 如前所述，由故障切换控制程序 (FP)110 参照，被存储在 CM23 的一部分中设置的共享存储区域中。FP110 可以从各 RP(RP0 ～ RP3) 参照故障切换管理表 1500，使用在本表中设定为可以使用（“E”:Enable) 的数据路径，访问 EPD50(ED0 ～ ED2)。连接存储了故障切换管理表 1500 的 CM23 的 ED1，在本表内存储的数据中发生了变更的情况下，向全部 RP(RP0 ～ RP3) 发送中断。接收到中断的各 RP 确认本表的变更内容，并根据需要变更访问各 EPD50 的数据路径。

[0141] 故障切换管理表 1500，作为项目而具有“正常时数据路径”A(a ～ d)、“代替数据路径”B(e ～ h)、“VF 迁移”C(i)、“连累数据路径（第 1 级链路故障）”D(j ～ i)、“连累数据路径（第 2 级链路故障）”E(m ～ o)。关于“正常时数据路径”A，具有“S”（状态）a、“Root”(RP)b、“ED”(EPD)c、“VF”d。同样地，关于“代替数据路径”B，具有“S”e、“Root”f、“ED”g、“VF”h。关于“连累数据路径（第 1 级链路故障）”D，具有“Root”(RP)j、“ED”(EPD)k、“VF”1。同样地，关于“连累数据路径（第 2 级链路故障）”E，具有“Root”m、“ED”n、“VF”o。

[0142] 在图 15 的设定例中，故障切换前的正常时（通常时候）的数据路径成为图 26（例如数据路径 K1) 所示那样。

[0143] “正常时数据路径”A(a ～ d) 存储各 RP 为了对各 EPD50 进行访问而使用的、正常时（或初始状态）的数据路径（“正常时数据路径”:第 1 数据路径）及其状态。正常时数据路径中的 b 存储访问源 RP，c 存储访问目的地 EPD50，d 存储访问目的地 EPD50 内 VF，a 存储该数据路径的状态。状态 a 是“E”为有效 (Enable)、“D”为无效 (Disable)。

[0144] 在“代替数据路径”B(e ～ h) 存储当 A(a ～ d) 中存储的正常时（或初始状态）的数据路径中发生了故障时，作为该数据路径的替代而使用的数据路径（“代替数据路径”:第 2 数据路径）及其状态。代替数据路径中的 f 存储访问源 RP，g 存储访问目的地 EPD50，h 存储访问目的地 EPD50 内 VF，e 存储该数据路径的状态。

[0145] 与正常数据路径对应的状态 a 和代替数据路径的状态 e，在数据路径相同的情况下存储相同值。

[0146] “VF 迁移”C(i)，存储是否需要用于执行数据路径的故障切换（从第 1 向第 2 数据路径的切换）的 VF 迁移 (VF Migration)(yes/no) 的信息。关于访问目的地 EPD50 为 ED1

的数据路径，将 VF 迁移设定为不需要 (no)。

[0147] 在本实施方式中，作为特征之一，是使用 VF 迁移来执行数据路径的故障切换。例如，ED0 在初始状态下，可以从 RP0 和 RP2 访问（图 26，数据路径 K1、K3）。初始状态的数据路径中发生了故障时，访问源 RP 的变更（以及访问目的地 VF 的变更）和 VF 迁移的执行的结果是，例如以可以从 RP1 或 RP2 向 ED0 进行访问的方式变更数据路径，从而可以回避故障链路（图 16、图 17、图 26、图 27、从 K1 向 K2 的故障切换的情况）。

[0148] 图 27 表示与图 15 的设定对应的各个代替数据路径（省略到 ED1 的路径）。例如与 K1 对应的 K2、与 K3 对应的 K4。在正常时数据路径和代替数据路径中经由不同的链路。

[0149] 另外，ED1 针对来自全部 RP(4 个）的访问，映射了 MVF(4 个），因此，即使在从某 RP 向 ED1 的访问中发生了故障时，仅通过变更访问源 RP（不需要“VF 迁移”），可以从其它 RP 向 ED 1 访问 (ED1 中不需要将 VF 迁移控制功能设为有效）。

[0150] “连累数据路径（第 1 级链路故障）”D(j ～ 1) 以及“连累数据路径（第 2 级链路故障）”E(m ～ o) 存储当 A 中存储的数据路径中发生了故障时，有可能同时发生故障的数据路径（连累数据路径）。例如，考虑到连接 RP0 和 SW00 的链路上发生了故障时，除了从 RP0 向 ED0 访问的数据路径以外，从 RP0 向 ED1、ED2 进行访问的数据路径上也发生故障。

[0151] 在将各 RP 和作为 RP 连接侧交换器的第 1 交换器 40(SW00) 或第 2 交换器 40(SW01) 连接的链路（第 1 级链路、上游侧链路）上发生了故障时（第 1 级链路故障），关于被正常时数据路径连累而发生故障的数据路径（连累数据路径），j 存储访问源 RP，k 存储访问目的地 EPD50，1 存储访问目的地 EPD50 内 VF。

[0152] 在将作为 RP 连接侧交换器的 SW00 或 SW01、作为 EPD50 连接侧交换器的第 3 交换器 40(SW10)、第 4 交换器 40(SW11) 或第 5 交换器 (SW12) 连接的链路（第 2 级链路）上发生了故障时（第 2 级链路故障），关于被正常时数据路径连累而发生故障的数据路径（连累数据路径），m 存储访问源 RP，n 存储访问目的地 EPD50，o 存储访问目的地 EPD50 内 VF。

[0153] 另外，在本例中，具备 FEI/F102 的 ED0 和具备 CMI/F104 的 ED1 通过相同 VH 进行数据转发，并且，具备 BEI/F106 的 ED2 和具备 CMI/F104 的 ED1 通过相同 VH 进行数据转发。存储装置 10 的管理者，以使该转发条件在正常时数据路径以及代替数据路径的双方中成立的方式设定故障切换管理表 1500。

[0154] ＜故障切换动作＞

[0155] 图 16 以及图 17 是故障切换动作的说明图。是根据所述图 1 的结构，从第 1 数据路径 K1 向第 2 数据路径 K2 的故障切换的情况。

[0156] 首先，在图 16 中，作为正常时数据路径，具有例如从 RC0 内的 RP0 通过第 1 交换器 40(SW00) 和第 3 交换器 40(SW10) 对 ED0 进行访问的第 1 数据路径 K1（经由链路 L1、L2、L5 的路径）。并且，考虑第 1 数据路径 K1 上的、例如 SW00 和 SW10 之间的链路 L2 中发生了故障的情况。此时，当参照故障切换管理表 1500 时可知，与发生了故障的数据路径 (K1) 对应的代替数据路径，是从 RC0 内的 RP1 通过第 2 交换器 40(SW01) 和第 3 交换器 40(SW10) 对 ED0 进行访问的第 2 数据路径 K2（经由链路 L3、L4、L5 的路径）。因此，进行从第 1 数据路径 K1 向第 2 数据路径 K2 的故障切换动作。

[0157] 图 17 是关于图 14 中所示的 CN 的虚拟拓扑中的数据路径变更（以及用于实现故障切换的 VF 迁移的动作）的说明图。左侧是从 RP0 看到的拓扑，右侧是从 RP1 看到的拓扑。

[0158] 图 16 的第 1 数据路径 K1（正常→故障状态），相当于图 17 中从 RC0 内的 RP0 经由 VS Fabric(VS) 到 ED0 内的第 1VF1702（“VF0 :0,1”）的数据路径。在正常时，通过 VF 映射 (VF Mapping)1704 将第 1VF 1702（“VF0 :0,1”）和 MVF 1701（“MVF0,0”）连接（映射）。

[0159] 另外，图 16 的第 2 数据路径 K2（代替），相当于图 17 中从 RC0 内的 RP1 经由 VS Fabric(VS) 到 ED0 内的第 2VF 1703（“VF1 :0,1”）的数据路径。

[0160] 如前所述，为了从 K 1 向 K2 进行故障切换，将访问源 RP 从 RP0 变更为 RP1，而且进行 VF 迁移（变更 VF 和 MVF 的映射）。此外，与 VF 迁移联动地变更访问目的地 VF。MR-PCIM 为了将第 2VF1703（“VF1 :0,1”）和 MVF1701（“MVF0,0”）连接（映射）而执行 VF 迁移 (VF Migration)。该 VF 迁移的执行结果是，解除了基于所述 VF 映射 1704 的 MVF1701（“MVF0, 0”）和第 1VF1702（“VF0 :0,1”）的连接，通过新的 VF 映射 1705 将 MVF1701（“MVF0,0”）和第 2VF 1703（“VF1 :0,1”）连接。通过本例的 VF 迁移，访问目的地 VF 被从“VF0 :0, I”变更为“VF 1 :0,1”，从 VF 映射 1704 的状态变更为 VF 映射 1705 的状态。结果，第 2 数据路径 K2 成为可以使用。

[0161] 如上所述，在本实施方式中，通过访问源 RP 变更和 VF 迁移来实现数据路径的故障切换。

[0162] < 数据路径故障切换方法 >

[0163] 图 18 是本实施方式的存储装置 10 中的、数据路径故障切换方法以及处理 (1800) 的流程 (S 表示处理步骤）。作为处理的概要，MR-PCIM 检测数据路径上的错误（故障），FP110 参照 FT111 来确认故障切换的条件，MR-PCIM 以及 SR-PCIM 进行 VF 迁移控制。

[0164] 在 S1801 中，当 MR-PCIM 检测出 RP 和 EPD50 之间、或者 2 个 EPD50 间的数据转发 (EPD50 访问）中的超过预定次数的错误时，将其结果通知给故障切换控制程序 (FP)110。

[0165] 在 S1802 中，FP 110 参照故障切换管理表 (FT)1500(111)，确认发生了故障的数据路径的故障切换条件（所述代替数据路径、是否需要 VF 迁移、连累数据路径等）。

[0166] 在 S1803 中，FP110 参照 FT1500 的“VF 迁移”C 的项目，判定是否需要 VF 迁移。当需要 (yes) 时进入 S1804。

[0167] 在 S1804 中，FP110 将 FT1500 的故障状态的数据路径（“正常时数据路径”A) 的状态 a 设定为“D”(Disable)。

[0168] 在 S1805 中，FP110 对 MR-PCIM 请求执行针对对象 EPD50 内 VF 的、从“VF h :f, s”向“VF a :f, c”的 VF 迁移 (VF Migration)(h, f, s, a, c 等是数值）。

[0169] 在 S1806 中，MR-PCIM 执行 VF 迁移 (VF Migration)，并更新 EPD50 的 VF 映射 (VF Mapping) 管理表。

[0170] 在 S1807 中，FP110 将 FT1500 的代替数据路径（“代替数据路径”B) 的状态 e 设定为“E”(Enable)。

[0171] 在 S1808 中，判定有无未处理的连累数据路径。在有未处理的连累数据路径的情况下，返回 S1802，在没有的情况下结束。例如，在将 RP 和交换器 40 连接的链路（上游侧）中发生了故障时，在向多个 EPD50 的访问中发生故障，需要执行多次的数据路径故障切换。在这种情况下，成为进行数据路径故障切换的对象的连累数据路径，被存储在 FT1500 的“连累数据路径”D、E 中。

[0172] 另一方面，在所述 S1803 中，当 VF 迁移为否 (no) 时进入 S1809。在 S1809 中判定

故障数据路径是否是正在使用的数据路径。关于正在使用的数据路径，FT1500 的状态变为“E”(Enable)。当状态为“E”(Enable) 时进入 S1810，当状态为“D”(Disable) 时进入 S1808。本例中的图 15 的 FT1500 的设定中，不需要数据路径故障切换时的 VF 迁移的、从各 RP 向 ED1（搭载 CMI/F104) 的访问的数据路径的状态为“E”(Enable)。

[0173] 在 S1810 中，FP110 将 FT1500 的故障数据路径的状态设定为“D”(Disable)。

[0174] 在 S1811 中，MR-PCIM 使发生了访问故障的 VF 从“Active.Availabel 状态 (A.A)”迁移到“Active.MigrateOut 状态 (A.M)”（活动．迁出）。

[0175] 在 S1812 中，在故障 VH 内工作的 SR-PCIM 或其代替单元，使发生了访问故障的 VF 从“A.M”迁移到“Inactive.Unavailable 状态 (I.U)”。该代替单元，在 SR-PCIM 和 VF 之间的通信故障时，代替 SR-PCIM 来控制 VF 状态迁移（作为实施方式 2，在后面叙述）。

[0176] <每个 EPD 的 VF 映射管理表 (2)>

[0177] 图 19 表示 EDO（搭载 FEI/F102) 的 VF 映射管理表 (1100b) 的一例（变更后）。在图 19 的状态下，对应于图 11 的 VF 映射管理表 1100 的状态，将 EDO 的 MVF(“MVF 0,1”) 的映射从 VH2(VF “VF 2 :0,1”) 向 VH3(VF “VF 3 :0,1”) 变更（行 1901、1902)。如图 12 所示，ED1（搭载 CMI/F104) 与 EDO 同样地可以通过 VH3 来访问（“VF 状态”为“A.A”)，因此即使是图 19 的映射状态，EDO 和 ED1 也可以在相同 VH3 内工作。但是，如图 13 所示，正在由 ED2 使用 VH3，因此 FE1/F102 侧和 BEI/F106 侧的数据转发在相同 VH3 内发生。

[0178] 图 20 表示 ED2（搭载 BEI/F106) 的 VF 映射管理表 (1300b) 的一例。在图 20 的状态下，对应于图 13 的 VF 映射管理表 1300 的状态，正在将 ED2 的 MVF(“MVF 0,1”) 的映射从 VH3(“VF 3 :0,1”) 向 VH2(“VF 2 :0,1”) 变更（行 2001、2002)。

[0179] 这些 VF 迁移（图 19、图 20) 的结果是，具备 FEI/F 102 的 EDO 和具备 CMI/F104 的 ED1 通过 VHO 和 VH3 进行数据转发，并且，具备 BEI/F106 的 ED2 和具备 CMI/F104 的 ED1 通过 VH1 和 VH2 进行数据转发。如此，将在 FEI/F102 和 BEI/F106 的数据转发中使用的 VH 分离，可以有效地实现处理器负荷以及架构 (fabric)（内部网络 (CN)) 内通信量的分散等。

[0180] 图 28 表示图 19 和图 20 的 VF 映射应用时的数据路径的例子（省略到 ED1 的路径）。

[0181] 此外，不限于故障发生时，未发生故障的数据路径可以与故障数据路径的故障切换同样地，执行从图 13 到图 20 那样的数据路径变更。

[0182] （实施方式 2)

[0183] 接下来，使用图 21 以后说明实施方式 2。实施方式 2 是在实施方式 1 上追加了功能 (VF 状态迁移单元）的结构。

[0184] <VF 状态迁移单元 >

[0185] 说明与基于所述 SR-PCIM 的 VF 状态迁移对应的代替单元 (VF 状态迁移单元）。VF 迁移 (VF Migration) 由 MR-PCIM 执行，但 VF 迁移中必需的 VF 状态迁移的一部分由 SR-PCIM 执行。但是，考虑到当数据路径故障发生时等，SR-PCIM 和 VF 之间的通信变得困难。因此，在实施方式 2 中代替 SR-PCIM 而设置使 VF 进行状态迁移的 VF 状态迁移单元。

[0186] 图 21 表示可以代替 SR-PCIM 使 VF 进行状态迁移的 EPD51（例如 EDO) 的块结构。EPD51，对应于图 2 中所示的 EPD50(EDO) 的结构，首先，追加了功能号码 1 的一个 BF2102(“BF 0,1”)。BF2102(“BF 0 :1”) 的冒号后的功能号码 1，表示用于 VF 的状态控

制。EPD51 还具备 BF2102 附带的 4 个 PF2103(“PF 0 :1”、“PF 1 :1”、“PF 2 :1”、“PF 3 :1”)。另外,追加了 VF 状态 (VF State) 控制部 2104。VF 状态控制部 2104,与 PF2103 以及 VF204 以及管理端口 2105 连接。VF 状态控制部 2104,可以使功能号码 0 的 4 个 VF204(“VF 0 :0,1”、“VF 1 :0,1”、“VF 2 :0,1”、“VF 3 :0,1”)进行状态迁移。

[0187] BF2102 以及 PF2103 是可以控制与 VF204 等关联的 VF 的状态迁移的功能。BF2102 与 EPD51 的 MR-IOV 规范相关,管理与功能号码 1 的功能相关的特性。针对每个 VH 而具备 PF2103,以便可以从各 RP 访问,PF2103 与 VF 状态控制部 2104 相连。各 RP 可以经由分配给各 VH 的功能号码 1 的 PF,与 VF 状态控制部 2104 通信。

[0188] 在实施方式 2 中,作为使用了 EPD51 的 VF 状态迁移以及控制的方法(结构),有以下 3 种。是可以通过设定等来选择性地使用它们的结构。此外,也可以成为仅具备某一种方法的结构。

[0189] 第 1 方法(基于带内 (inband) 通信的方法):VF 状态控制部 2104,通过与 PF2103 连接,通过来自 SC20 内动作的程序(例如,与该 EPD51 正常连接的数据路径中的 RP 上的处理器 21 执行的 FP110) 的指令,可以使 VF204(“VF0 :0,1”、“VF 1 :0,1”、“VF 2 :0,1”、“VF 3 :0,1”)进行状态迁移。

[0190] 当无法进行基于某 VH(RP) 中的 SR-PCIM 的、向 EPD51 内 VF 的访问时,在可以正常访问的其它 VH(RP) 中,经由功能号码 1 的 PF,向 EPD51 内的 VF 状态控制部 2104 进行访问。例如,对于 ED0,不从 RP0、而从 RP1 进行访问(利用 PF 2103(“PF 1 :1”))。此外,在成为仅第 1 方法的结构时,不需要管理端口 2105。

[0191] 第 2 方法(基于带外 (outband) 通信的方法):另外,VF 状态控制部 2104 经由管理端口 2105 和网络 N00 与 SVP190 连接,由此,可以接收基于来自在 PCIe 架构外工作的程序的带外通信的命令,使 VF204(“VF0 :0,1”、“VF 1 :0,1”、“VF2 :0.1 ”、“VF3 :0,1”)进行状态迁移。此外,在采用仅第 2 方法的结构时,不需要 BF2102、PF2103。

[0192] 第 3 方法(基于自主功能的方法):另外,VF 状态控制部 2104,通过自主功能(通过自身进行与 VF 状态迁移相关的判断控制的功能),可以使 VF204(“VF 0 :0,1”、“VF 1 :0,1”、“VF2 :0,1”、“VF3 :0,1”)进行状态迁移。此时,与 BF202(“BF 0 :0”)相连的 VF 状态控制部 2104,更新 BF202(“BF 0 :0”)内的 VF 映射管理表。此外,在采用仅第 3 方法的结构时,不需要管理端口 2105。

[0193] 如上所述,VF 状态控制部 2104,关于功能号码 0 的 VF 的状态控制,有 3 个方法,但关于选择使用哪个方法,例如可以由管理者经由管理端口 2105 针对每个 EPD51 预先设定等。

[0194] <VF 迁移处理 (1)>

[0195] 图 22(A)、(B) 表示使用第 1 方法,在 EPD51 中使用功能号码 1 的功能使功能号码 0 的 VF 进行状态迁移,执行 VF 迁移 (VF Migration) 的处理 (2200) 的流程。

[0196] 在 S2201 中,实施方式 2 中的 FP110 等的高级别软件对 MR-PCIM 请求从作为发生了访问故障的 EPD51 内 VF 的“VF h :f, s”向“VF a :f, c”的 VF 迁移 (VF Migration)。

[0197] 用于执行 VF 迁移的必要条件是:转移目的地 VF 的状态为“Inactive.Unavailable 状态 (I.U)”,并且转移源 VF 的状态为“Active.Available 状态 (A.A)”或者“Dormant.Migration 状态 (D.M)”(休眠 . 迁移)。

[0198] 在S2202中，MR-PCIM判定转移目的地VF“VF a:f,c”是否是“I.U”。在“是”的情况下进入S2203。在“否”的情况下，由于VF迁移的必要条件不成立，因此进行错误结束。

[0199] 在S2203中，MR-PCIM判定转移源VF“VF h:f,s”是否是“A.A”。在“是”的情况下进入S2204。在“否”的情况下进入S2221。

[0200] 在S2204中，MR-PCIM，通过改写VF状态条目(VF State entry)(与VF映射管理表的“VF State”c的条目同值)，使作为发生了访问故障的EPD51内VF的“VF h:f,s”的VF状态(VF State)从“A.A”向“A.M”迁移(在“VHh”中，启动向“A.M”的迁移操作)。在该迁移中将SR VF迁移状态位(SR VFMigration Status bit)进行置位(set)(向图22(B)的S2205接续)。此外，该动作与图25的(1)对应。

[0201] 图22(B)，在S2205中，MR-PCIM等(MR-PCIM或者可以与MR-PCIM通信的软件(FP110等))，选定可以与发生了访问故障的EPD51通信的VH、即“VH n”(n≠h)。

[0202] 在S2206中，MR-PCIM等经由PF2103“PF n:1”与发生了访问故障的EPD51内的VF状态控制部2104通信，使VF状态控制部2104对VF“VF h:f,s”发行FLR(Function Level Reset)(功能等级复位)。此外，该动作与图25的(2)对应。

[0203] 在S2207中，MR-PCIM等经由PF2103“PF n:1”(n≠h)与发生了访问故障的EPD51的VF状态控制部2104通信，使VF“VF h:f,s”的VF状态从“A.M”向“I.U”迁移。在该迁移中将MR VF迁移状态位进行置位，启动向MR-PCIM的中断。此外，该动作与图25的(3)对应。

[0204] 在S2208中，接收了上述中断的MR-RCIM，为了确定中断的原因，浏览发生了访问故障的EPD51的VF映射管理表，查找从“A.M”迁移到“I.U”的VF(“VF h:f,s”)。此外，该动作与图25的(4)对应。

[0205] 在S2224中，MR-PCIM保存与VF“VF h:f,s”关联的VF映射条目(VF Mapentry)的值。另外，该动作与图25的(5)对应。在S2225中，MR-PCIM在与VF“VF h:f,s”关联的VF Map entry中设置表示“empty slot”(空时隙)的0。另外，该动作与图25的(6)对应。在S2226中，MR-PCIM在与VF“VFa:f,c”关联的VF映射条目中设置在S2224中保存的值。另外，该动作与图25的(7)对应。

[0206] 在S2227中，MR-PCIM通过改写与VF“VF a:f,c”关联的VF Map entry，对于VF“VF a:f,c”(“VH a”)启动从“I.U”向“D.M”的迁移操作。在该迁移中将SR VF迁移状态位进行置位。另外，该动作与图25的(8)对应。

[0207] 在S2228中，在与VF“VF a:f,c”关联的“VH a”内运行的SR-PCIM，浏览VF状态阵列(VF State array)来确定迁移到“D.M”的VF“VF a:f,c”，并且对该VF“VF a:f,c”发行FLR。另外，该动作与图25的(9)对应。

[0208] 在S2229中，在“VH a”中运行的SR-PCIM，在开始使用VF“VF a:f,c”前，使VF“VF a:f,c”的VF状态从“D.M”向“A.A”迁移。由此，正常地结束VF转移。另外，该动作与图25的(10)对应。

[0209] 另一方面，在图22(A)的S2221中，判定转移源VF“VF h:f,s”是否是“D.M”。在“是”的情况下进入S2222。在“否”的情况下，由于VF迁移的必要条件不成立，因此结束。

[0210] 在S2222中，MR-PCIM改写与VF“VF h:f,s”关联的VF State entry，撤销VF“VF h:f,s”的迁移操作。其结果是，VF“VF h:f,s”从“D.M”向“I.U”迁移(此后，向图23(B)

的 S2224 接续）。另外，该动作与图 25 的 (11) 对应。

[0211] 此外，在与 MR-IOV 规范中的 S2205 ～ S2207 相当的处理中，首先在作为与 VF “VF h :f, s”关联的 VH 的“VH h”中运行的 SR-PCIM 浏览 VF State array，确定向“A.M”迁移的 VF “VF h :f, s”，并中止 VF “VF h :f, s”的使用，并对 VF “VF h :f, s”发行 FLR。接着，SR-PCIM 使 VF “VF h :f, s”的 VF 状态从“A.M”向“I.U”迁移，并且在该迁移中将 MR VF 迁移状态位进行置位，启动向 MR-PCIM 的中断。

[0212] 如上所述，在实施方式 2（第 1 方法）的故障切换方法以及 VF 迁移处理中，特征在于，在 S2205 ～ S2207 的处理中使用在 EPD51 内设置的 VF 状态控制部 2104 来处理在上述 MR-IOV 规范中应该由 SR-PCIM 处理的步骤。

[0213] <VF 迁移处理 (2)>

[0214] 图 23 是使用第 2 方法，在 EPD51 中 MR-PCIM 使用经由管理端口 2105 和网络 N00 的带外通信，通过功能号码 1 的功能使功能号码 0 的 VF 进行状态迁移，执行 VF 迁移的处理 (2300) 的流程。

[0215] 在第 2 方法中，首先，最初与图 22(A) 的 S2201 ～ S2204 等的处理相同，接着在图 23 中，作为与第 1 方法不同的部分而具有 S2305、S2306。

[0216] 在 S2305 中，MR-PCIM 等 (MR-PCIM 或可以与 MR-PCIM 通信的软件（故障切换控制程序等）），与发生了访问故障的 EPD51 内 VF 状态控制部 2104，进行经由网络 00 的带外通信。使 VF 状态控制部 2104 对 VF（“VF h :f, s”）发行 FLR。

[0217] 在 S2306 中，MR-PCIM 等与发生了访问故障的 EPD51 的 VF 状态控制部 2104 进行经由网络 00 的带外通信，使 VF（“VF h :f, s”）的 VF 状态从“A.M”向“I.U”迁移。在该迁移中，将 MR VF 迁移状态位进行置位，并且启动向 MR-PCIM 的中断。以后的处理与前文所述的相同。

[0218] 如上所述，在实施方式 2（第 2 方法）的故障切换方法以及 VF 迁移处理中，特征在于，在 S2305 ～ S2306 中使用在 EPD51 内设置的 VF 状态控制部 2104 和带外通信，来处理在 MR-IOV 规范中应该由 SR-PCIM 处理的步骤。

[0219] <VF 迁移处理 (3)>

[0220] 图 24 是使用第 3 方法，在 EPD51 中由 VF 状态控制部 2104 的自主控制功能使功能号码 0 的 VF 进行状态迁移，执行 VF 迁移的处理 (2400) 的流程。

[0221] 在第 3 方法中，首先，最初与图 22(A) 的 S2201 ～ S2204 等处理相同，另外，S2221 ～ S2222、图 22(B) 的 S2224 ～ S2229 等处理也相同，接着，在图 24 中，作为与第 1 方法不同的部分而具有 S2405 ～ S2410。

[0222] 在 S2405 中，发生了访问故障的 EPD51 的 VF 状态控制部 2104 确认是否可以与发生了故障的 VH 的 SR-PCIM 通信。例如，在从 RP3 向端点设备内的 VF(VF3 :0, 1) 访问的数据路径上发生了故障的情况下，VF 状态控制部 2104 控制与 VF(VF3 :0, 1) 属于相同 VH 的功能号码 1 的 PF(PF3, 1)。VF 状态控制部 2104 实施从 PF(PF3, 1) 向 RP3 的访问测试，该访问测试与从 RP3 对 VF(VF3 :0, 1) 的访问使用相同的物理数据路径。根据其结果，判定在 VH3 内运行的 SR-PCIM 是否可以与 VF(VF3 :0, 1) 通信。

[0223] 在 S2406 中，判定发生了访问故障的 EPD51 是否可以与故障 VH 内 SR-PCIM 通信。若可以，则进入 S2409。若不可以，则进入 S2407。

[0224] 在 S2407 中，发生了访问故障的 EPD51 的 VF 状态控制部 2104，对于迁移到“A.M”的 VF “VF h :f，s”发行 FLR。

[0225] 在 S2408 中，发生了访问故障的 EPD51 的 VF 状态控制部 2104 使 VF “VFh :f，s”从“A.M”向“I.U”迁移。在该迁移中，将 MR VF 迁移状态位进行置位，启动向 MR-PCIM 的中断。

[0226] 另一方面，在 S2409 中，在与 VF “VF h :f，s”关联的“VH h”中运行的 SR-PCIM 浏览 VF State array 来确定迁移到“A.M”的 VF “VF h :f，s”，中止该 VF “VF h :f，s”的使用，并对 VF “VF h :f，s”发行 FLR。

[0227] 在 S2410 中，SR-PCI M 使 VF “VF h :f，s”从“A.M”迁移到“I.U”。在该迁移中，将 MR VF 迁移状态位进行置位，启动向 MR-PCIM 的中断。

[0228] 在 S2411 中，接收到中断的 MR-PCIM 为了确定中断的原因，查找从“A.M”迁移到“1.U”的 VF（“VF h :f，s”）。以后的处理与前文所述的相同。

[0229] 如上所述，在实施方式 2（第 3 方法）的故障切换方法以及 VF 迁移处理中，特征在于，在 S2405 ～ S2410 中，通过在 EPD51 内设置的 VF 状态控制部 2104 的自主控制，来处理在 MR-IOV 规范中应该由 SR-PCIM 处理的步骤。

[0230] <VF 迁移的状态迁移 >

[0231] 图 25 为了补充而表示与基于 MR-IOV 规范的 VF 迁移有关的状态迁移。VF 具有 4 个状态（“A.A Active.Available 状态”，“A.M(Active.MigrateOut 状态）”，“I.U(Inactive.Unavailable 状态）”，“D.M(Dormant.MigrateIn 状态（休眠．迁入））”）。实线的箭头表示 MR-PCIM 的状态迁移。虚线的箭头表示 SR-PCIM 的状态迁移。“A.A”，“A.M”，是 VF active（完全可由 SR 使用），“I.U”是 VF inavtive（不可由 SR 使用），“D.M”是 VF Dormant（对应于 SR）。

[0232] 在故障切换动作时，作为与 VF 迁移相伴的 VF 状态迁移控制，在 (1) 中，通过 MR-PCIM 的控制，从“A.A”向“A.M”迁移，在 (3) 中，通过 SR-PCIM 的控制，从“A.M”向“U.U”迁移。但是在 (3) 中从“A.M”向“U.U”迁移时，由于数据路径故障等，有时不可能进行 SR-PCIM 的控制 (VF 状态迁移控制）。因此，如前所述设置了 VF 状态迁移控制的代替单元。

[0233] 如上所述，根据各实施方式，在对 SC20 的内部网络 (CN) 应用了 MR-IOV 的存储装置 10 中，通过 MR-PCIM 等，根据该数据路径中的访问源 RP 变更和访问目的地 EPD 内 VF 的 VF 迁移的动作，可以执行从 RP 到 EPD 的数据路径的故障切换。因此，在使用了本存储装置 10 的系统中，即使在数据路径中发生了故障时，也可以故障切换到其它数据路径，确保了高可靠性。

[0234] 以上，根据实施方式具体说明了本发明人做出的发明，但本发明不限定于所述实施方式，在不脱离其主旨的范围内，当然可以进行各种变更。

[0235] 产业上的可利用性

[0236] 本发明不仅可以应用于具有内部网络的存储装置，也可以广泛应用于刀片型服务器 (blade server) 等计算机及其内部网络技术等。

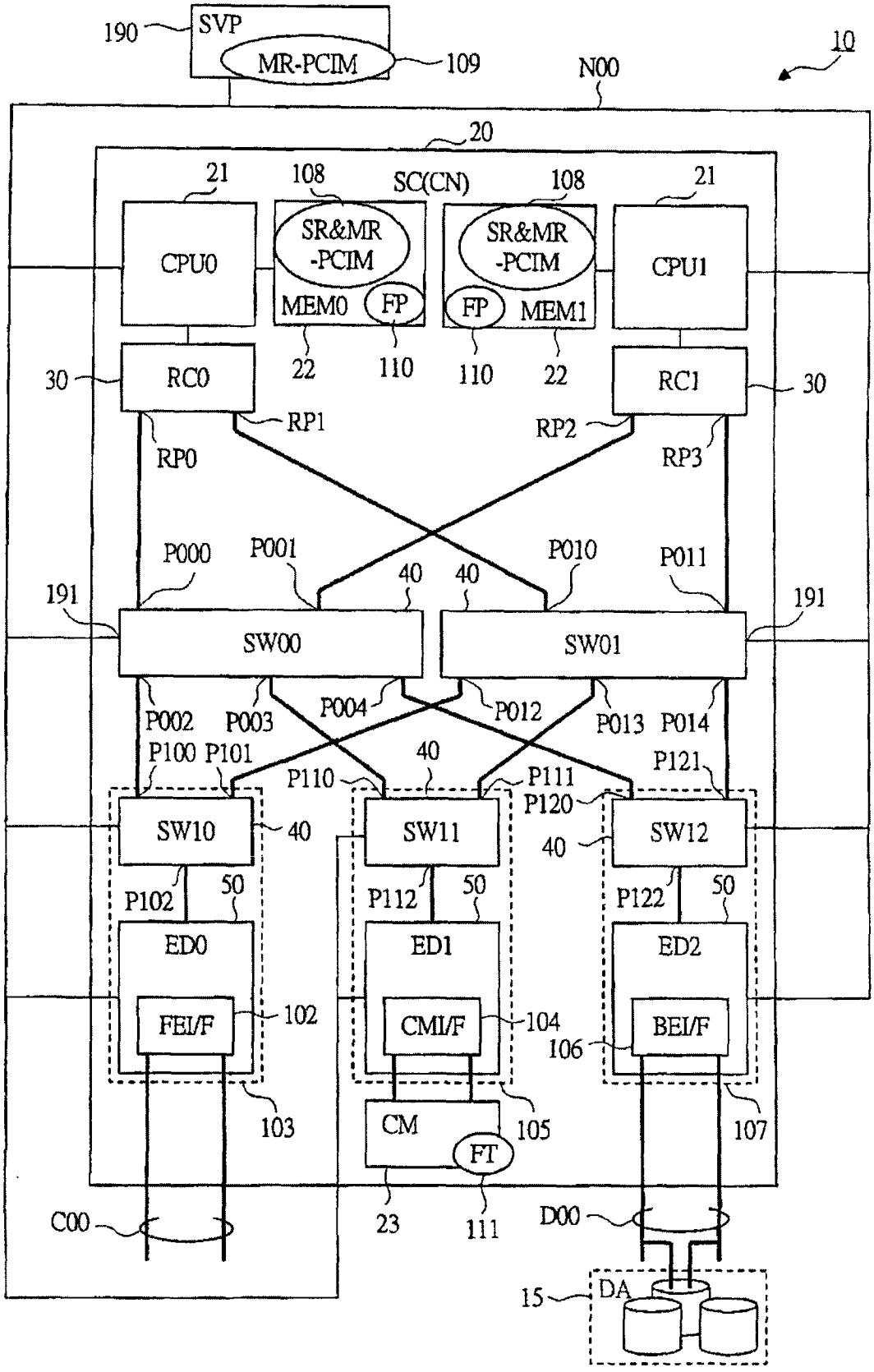

190
SVP
MR-PCIM
109
N00
10
20
21
108
SC(CN)
108
21
CPU0
SR&MR
-PCIM
MEM0
FP
FP
MEM1
SR&MR
-PCIM
CPU1
30
RC0
22
110
110
22
RC1
30
RP0
RP1
RP2
RP3
P000
P001
40
40
P010
P011
191
SW00
SW01
191
P002
P003
P004
P012
P013
P014
P100
P101
40
P110
P111
P120
P121
SW10
40
SW11
40
SW12
P102
50
P112
50
P122
50
ED0
ED1
ED2
FEI/F
102
CMI/F
104
BEI/F
106
103
CM
105
107
FT
C00
23
111
D00
15
DA

图 1

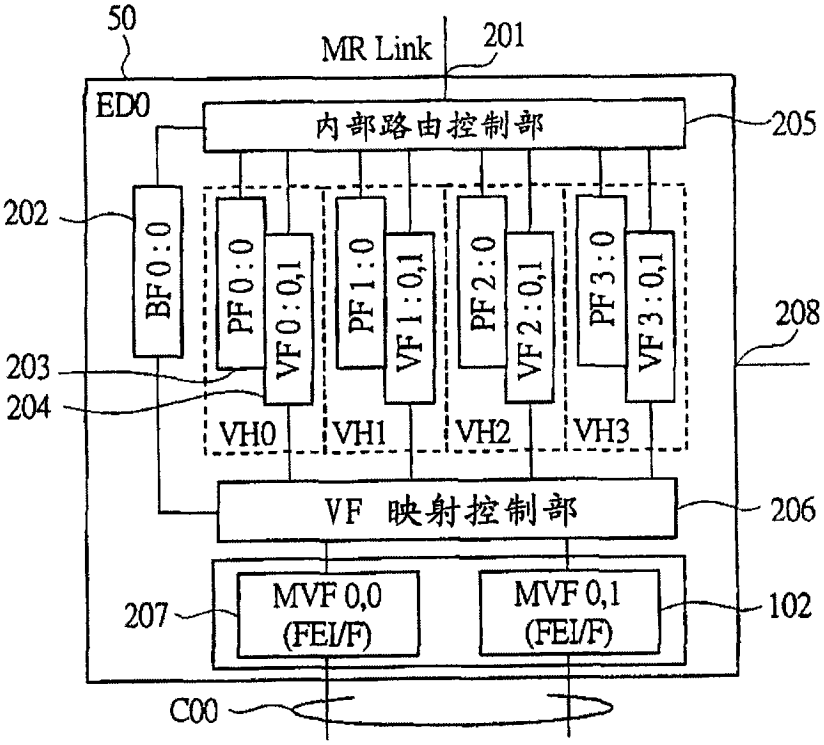

50
MR Link
201
ED0
内部路由控制部
205
202
BF 0 : 0
PF 0 : 0
VF 0 : 0,1
PF 1 : 0
VF 1 : 0,1
PF 2 : 0
VF 2 : 0,1
PF 3 : 0
VF 3 : 0,1
208
203
204
VH0
VH1
VH2
VH3
VF 映射控制部
206
207
MVF 0,0 (FEI/F)
MVF 0,1 (FEI/F)
102
C00

图 2

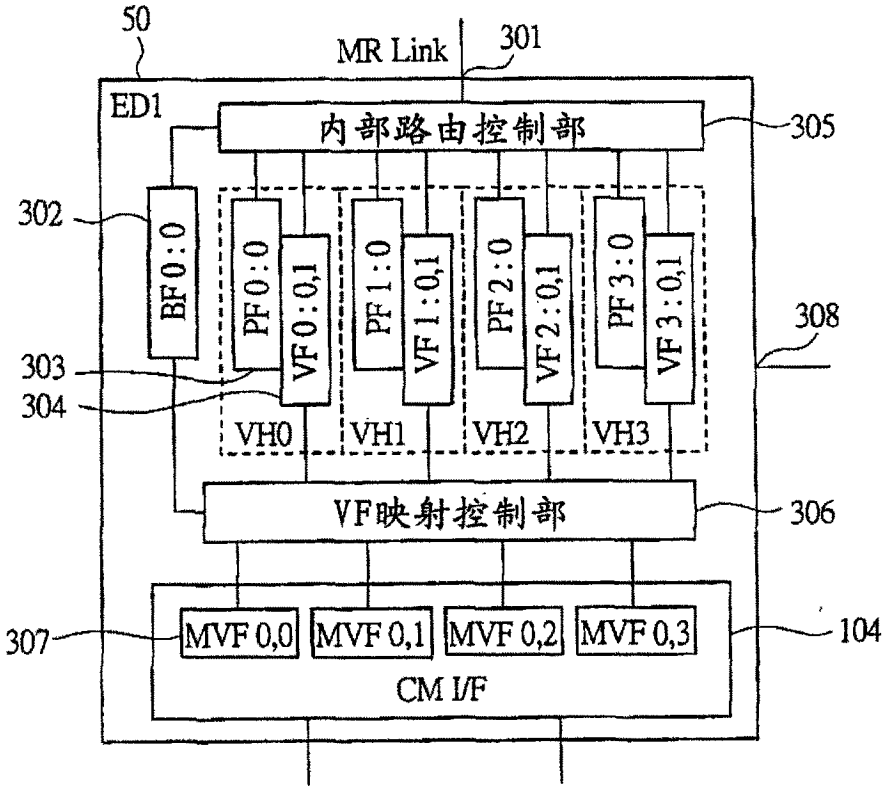

50
MR Link
301
ED1
内部路由控制部
305
302
BF 0 : 0
PF 0 : 0
VF 0 : 0,1
PF 1 : 0
VF 1 : 0,1
PF 2 : 0
VF 2 : 0,1
PF 3 : 0
VF 3 : 0,1
308
303
304
VH0
VH1
VH2
VH3
VF映射控制部
306
307
MVF 0,0
MVF 0,1
MVF 0,2
MVF 0,3
104
CM I/F

图 3

400

a 交换器端口	b PCIM 能够
P000, P001	是
P010, P011	
P002, P003, P004	否
P012, P013, P014	
P100, P110, P120	是
P101, P111, P121	
P102, P112, P122	否

图 4

500

a VH	b 认可	d e f	c VH 映射	g VF 映射
RP0	是	VS in	SW00, SW10, SW11, SW12	ED0 : 1VFs
		Inter-SW Link	P002 to P100, P003 to P110 P004 to P120	ED1 : 1VFs
		VH in	ED0, ED1, ED2	ED2 : 1VFs
RP1	是	VS in	SW01, SW10, SW11, SW12	ED0 : 1VFs
		Inter-SW Link	P012 to P101, P013 to P111 P014 to P121	ED1 : 1VFs
		VH in	ED0, ED1, ED2	ED2 : 1VFs
RP2	是	VS in	SW00, SW10, SW11, SW12	ED0 : 1VFs
		Inter-SW Link	P002 to P100, P003 to P110 P004 to P120	ED1 : 1VFs
		VH in	EP0, ED1, ED2	ED2 : 1VFs
RP3	是	VS in	SW01, SW10, SW11, SW12	ED0 : 1VFs
		Inter-SW Link	P012 to P101, P013 to P111 P014 to P121	ED1 : 1VFs
		VH in	ED0, ED1, ED2	ED2 : 1VFs

图 5

600

a	b	c	d	e	f	g	h	
时隙	根	VS	桥	使能	映射	端口	端口VH号码	
0	RP0	VS0	上游	是	是	P000	VH 0	601
1	RP0	VS0	下游 0	是	是	P002	VH 0	602
2	RP0	VS0	下游 1	是	是	P003	VH 0	
3	RP0	VS0	下游 2	是	是	P004	VH 0	
4		VS0	下游 3	否	否			
5		VS0	下游 4	否	否			
6		VS0	下游 5	否	否			
7		VS0	下游 6	否	否			
8		VS0	下游 7	否	否			
9	RP2	VS1	上游	是	是	P001	VH 0	603
10	RP2	VS1	下游 0	是	是	P002	VH 1	604
11	RP2	VS1	下游 1	是	是	P002	VH 1	
12	RP2	VS1	下游 2	是	是	P004	VH 1	
13		VS1	下游 3	否	否			
14		VS1	下游 4	否	否			
15		VS1	下游 5	否	否			
16		VS1	下游 6	否	否			
17		VS1	下游 7	否	否			

图 6

700

a	b	c	d	e	f	g	h
时隙	根	VS	桥	使能	映射	端口	端口VH号码
0	RP1	VS0	上游	是	是	P010	VH 0
1	RP1	VS0	下游 0	是	是	P012	VH 0
2	RP1	VS0	下游 1	是	是	P013	VH 0
3	RP1	VS0	下游 2	是	是	P014	VH 0
4		VS0	下游 3	否	否		
5		VS0	下游 4	否	否		
6		VS0	下游 5	否	否		
7		VS0	下游 6	否	否		
8		VS0	下游 7	否	否		
9	RP3	VS1	上游	是	是	P011	VH 0
10	RP3	VS1	下游 0	是	是	P012	VH 1
11	RP3	VS1	下游 1	是	是	P013	VH 1
12	RP3	VS1	下游 2	是	是	P014	VH 1
13		VS1	下游 3	否	否		
14		VS1	下游 4	否	否		
15		VS1	下游 5	否	否		
16		VS1	下游 6	否	否		
17		VS1	下游 7	否	否		

701 702 703 704

图 7

800

a 时隙	b 根	c VS	d 桥	e 使能	f 映射	g 端口	h 端口VH号码	
0	RP0	VS0	上游	是	是	P0100	VH 0	
1	RP0	VS0	下游 0	是	是	P0102	VH 0	801
2		VS0	下游 1	否	否			
3	RP1	VS1	上游	是	是	P0101	VH 0	
4	RP1	VS1	下游 0	是	是	P0102	VH 1	802
5		VS1	下游 1	否	否			
6	RP2	VS2	上游	是	是	P0100	VH 1	
7	RP2	VS2	下游 0	是	是	P0102	VH 2	803
8		VS2	下游 1	否	否			
9	RP3	VS3	上游	是	是	P0101	VH 1	
10	RP3	VS3	下游 0	是	是	P0102	VH 3	804
11		VS3	下游 1	否	否			

图 8

900

a 时隙	b 根	c VS	d 桥	e 使能	f 映射	g 端口	h 端口VH号码	
0	RP0	VS0	上游	是	是	P0110	VH 0	
1	RP0	VS0	下游 0	是	是	P0112	VH 0	901
2		VS0	下游 1	否	否			
3	RP1	VS1	上游	是	是	P0111	VH 0	
4	RP1	VS1	下游 0	是	是	P0112	VH 1	902
5		VS1	下游 1	否	否			
6	RP2	VS2	上游	是	是	P0110	VH 1	
7	RP2	VS2	下游 0	是	是	P0112	VH 2	903
8		VS2	下游 1	否	否			
9	RP3	VS3	上游	是	是	P0111	VH 1	
10	RP3	VS3	下游 0	是	是	P0112	VH 3	904
11		VS3	下游 1	否	否			

图 9

1000

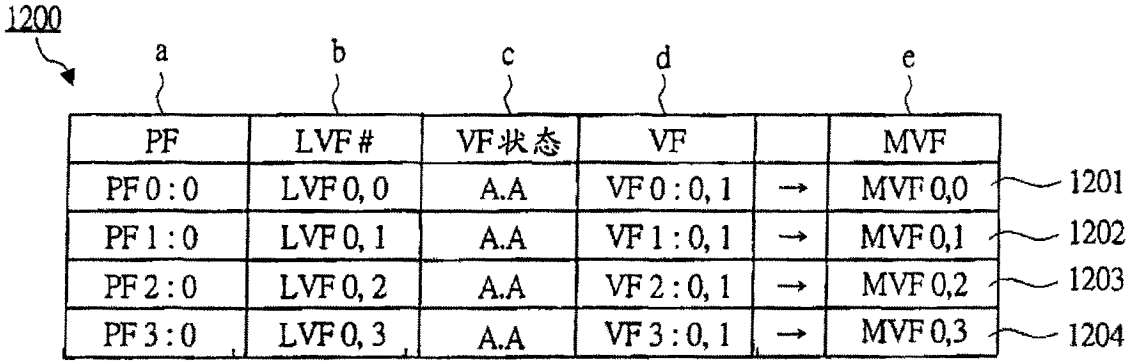
a 时隙	b 根	c VS	d 桥	e 使能	f 映射	g 端口	h 端口VH号码	
0	RP0	VS0	上游	是	是	P0120	VH 0	
1	RP0	VS0	下游 0	是	是	P0122	VH 0	1001
2		VS0	下游 1	否	否			
3	RP1	VS1	上游	是	是	P0121	VH 0	
4	RP1	VS1	下游 0	是	是	P0122	VH 1	1002
5		VS1	下游 1	否	否			
6	RP2	VS2	上游	是	是	P0120	VH 1	
7	RP2	VS2	下游 0	是	是	P0122	VH 2	1003
8		VS2	下游 1	否	否			
9	RP3	VS3	上游	是	是	P0121	VH 1	
10	RP3	VS3	下游 0	是	是	P0122	VH 3	1004
11		VS3	下游 1	否	否			

图 10

1100

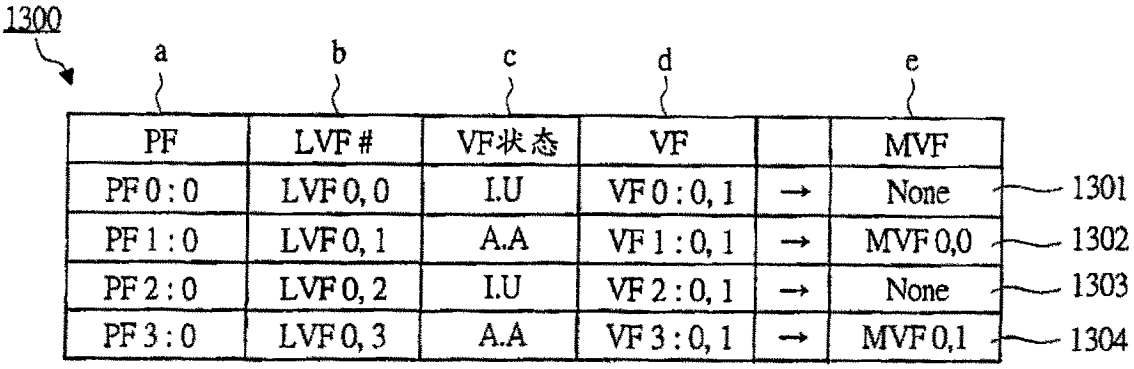
a PF	b LVF #	c VF 状态	d VF		e MVF	
PF 0 : 0	LVF 0, 0	A.A	VF 0 : 0, 1	→	MVF 0, 0	1101
PF 1 : 0	LVF 0, 1	I.U	VF 1 : 0, 1	→	无	1102
PF 2 : 0	LVF 0, 2	A.A	VF 2 : 0, 1	→	MVF 0,1	1103
PF 3 : 0	LVF 0, 3	I.U	VF 3 : 0, 1	→	无	1104

图 11



PF	LVF #	VF状态	VF		MVF	
PF 0 : 0	LVF 0, 0	A.A	VF 0 : 0, 1	→	MVF 0,0	1201
PF 1 : 0	LVF 0, 1	A.A	VF 1 : 0, 1	→	MVF 0,1	1202
PF 2 : 0	LVF 0, 2	A.A	VF 2 : 0, 1	→	MVF 0,2	1203
PF 3 : 0	LVF 0, 3	A.A	VF 3 : 0, 1	→	MVF 0,3	1204

图 12



PF	LVF #	VF状态	VF		MVF	
PF 0 : 0	LVF 0, 0	I.U	VF 0 : 0, 1	→	None	1301
PF 1 : 0	LVF 0, 1	A.A	VF 1 : 0, 1	→	MVF 0,0	1302
PF 2 : 0	LVF 0, 2	I.U	VF 2 : 0, 1	→	None	1303
PF 3 : 0	LVF 0, 3	A.A	VF 3 : 0, 1	→	MVF 0,1	1304

图 13

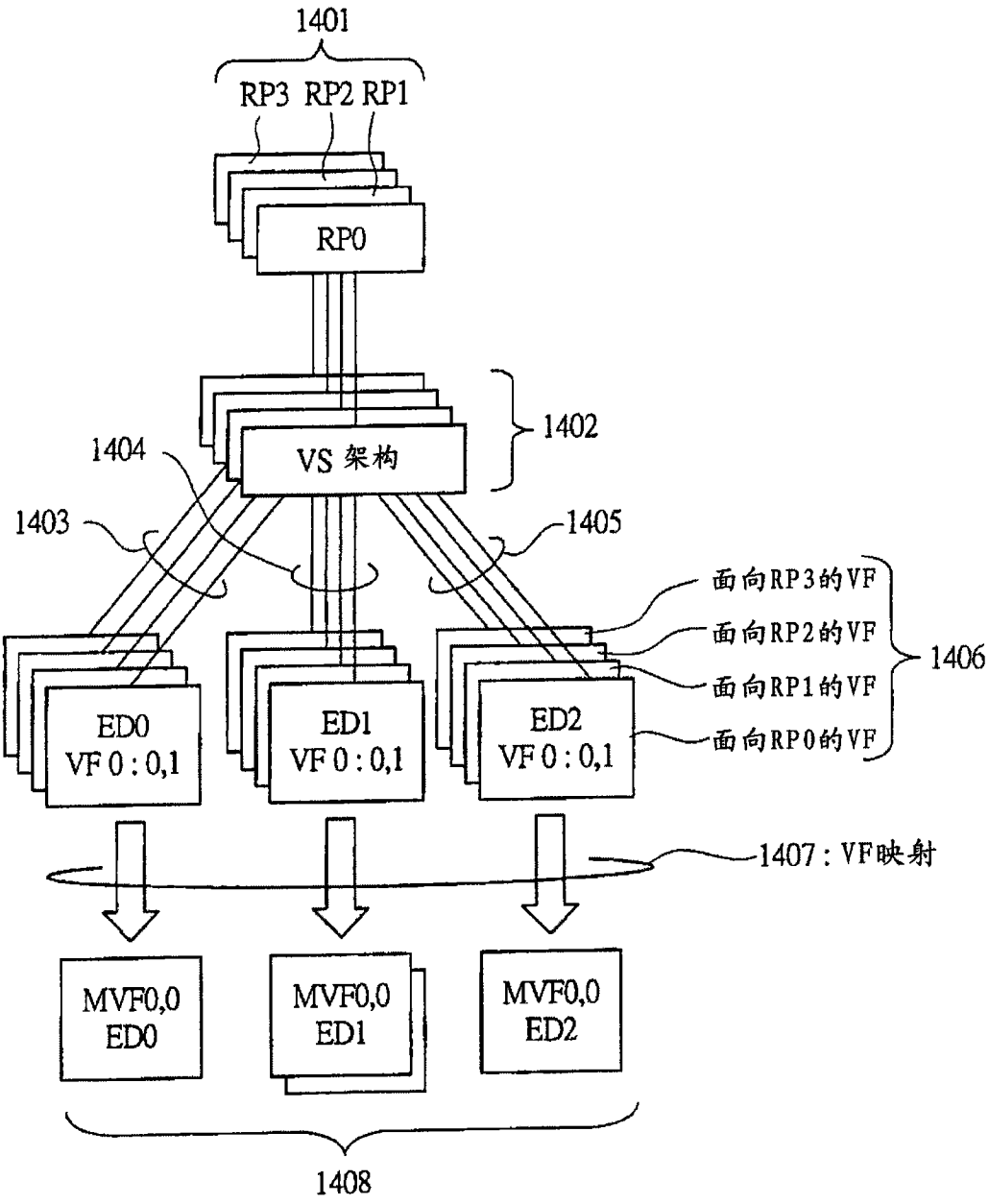

1401
RP3 RP2 RP1
RP0
VS 架构
1402
1404
1403
1405
面向RP3的VF
面向RP2的VF
面向RP1的VF
面向RP0的VF
1406
ED0
VF 0 : 0,1
ED1
VF 0 : 0,1
ED2
VF 0 : 0,1
1407 : VF映射
MVF0,0
ED0
MVF0,0
ED1
MVF0,0
ED2
1408

图 14

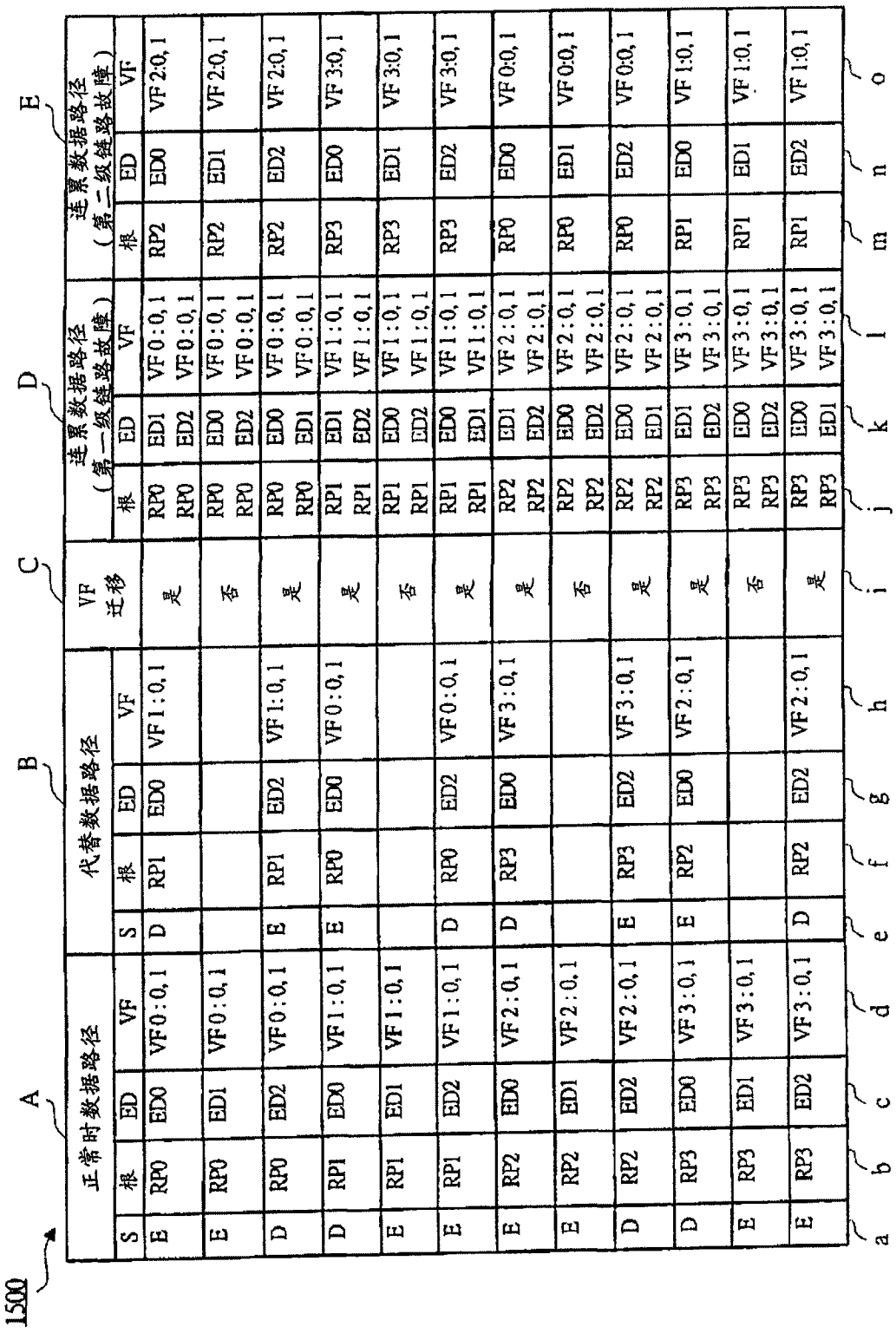

1500
A
B
C
D
E
正常时数据路径
代替数据路径
VF 迁移
连累数据路径（第一级链路故障）
连累数据路径（第二级链路故障）
S 根 ED VF S 根 ED VF 根 ED VF 根 ED VF
E RP0 ED0 VF 0:0,1 D RP1 ED0 VF 1:0,1 是 RP0 ED1 VF 0:0,1 RP2 ED0 VF 2:0,1
RP0 ED2 VF 0:0,1
E RP0 ED1 VF 0:0,1 否 RP0 ED0 VF 0:0,1 RP2 ED1 VF 2:0,1
RP0 ED2 VF 0:0,1
D RP0 ED2 VF 0:0,1 E RP1 ED2 VF 1:0,1 是 RP0 ED0 VF 0:0,1 RP2 ED2 VF 2:0,1
RP0 ED1 VF 0:0,1
D RP1 ED0 VF 1:0,1 E RP0 ED0 VF 0:0,1 是 RP1 ED1 VF 1:0,1 RP3 ED0 VF 3:0,1
RP1 ED2 VF 1:0,1
E RP1 ED1 VF 1:0,1 否 RP1 ED0 VF 1:0,1 RP3 ED1 VF 3:0,1
RP1 ED2 VF 1:0,1
E RP1 ED2 VF 1:0,1 D RP0 ED2 VF 0:0,1 是 RP1 ED0 VF 1:0,1 RP3 ED2 VF 3:0,1
RP1 ED1 VF 1:0,1
E RP2 ED0 VF 2:0,1 D RP3 ED0 VF 3:0,1 是 RP2 ED1 VF 2:0,1 RP0 ED0 VF 0:0,1
RP2 ED2 VF 2:0,1
E RP2 ED1 VF 2:0,1 否 RP2 ED0 VF 2:0,1 RP0 ED1 VF 0:0,1
RP2 ED2 VF 2:0,1
D RP2 ED2 VF 2:0,1 E RP3 ED2 VF 3:0,1 是 RP2 ED0 VF 2:0,1 RP0 ED2 VF 0:0,1
RP2 ED1 VF 2:0,1
D RP3 ED0 VF 3:0,1 E RP2 ED0 VF 2:0,1 是 RP3 ED1 VF 3:0,1 RP1 ED0 VF 1:0,1
RP3 ED2 VF 3:0,1
E RP3 ED1 VF 3:0,1 否 RP3 ED0 VF 3:0,1 RP1 ED1 VF 1:0,1
RP3 ED2 VF 3:0,1
E RP3 ED2 VF 3:0,1 D RP2 ED2 VF 2:0,1 是 RP3 ED0 VF 3:0,1 RP1 ED2 VF 1:0,1
RP3 ED1 VF 3:0,1
a b c d e f g h i j k l m n o

图 15

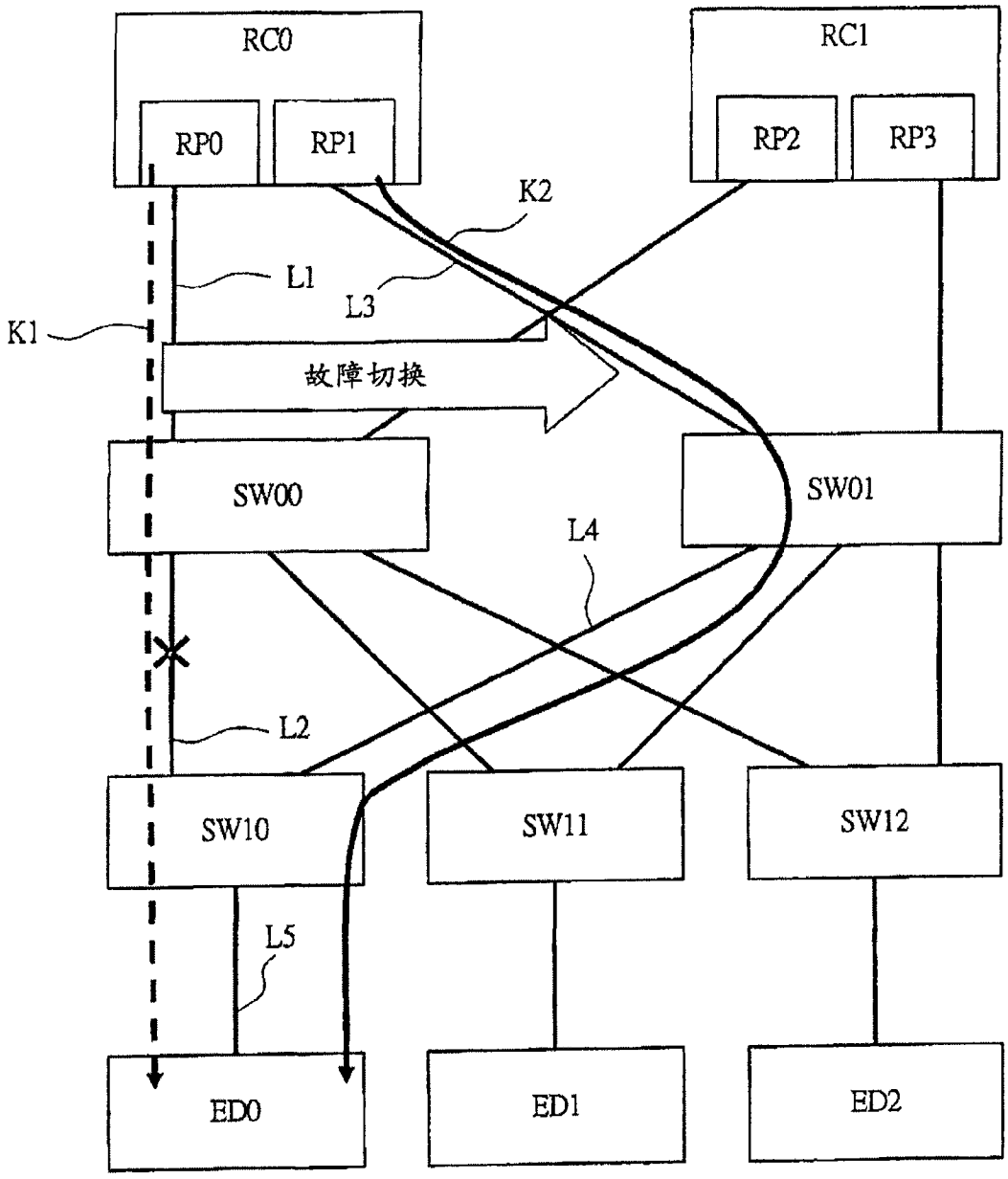
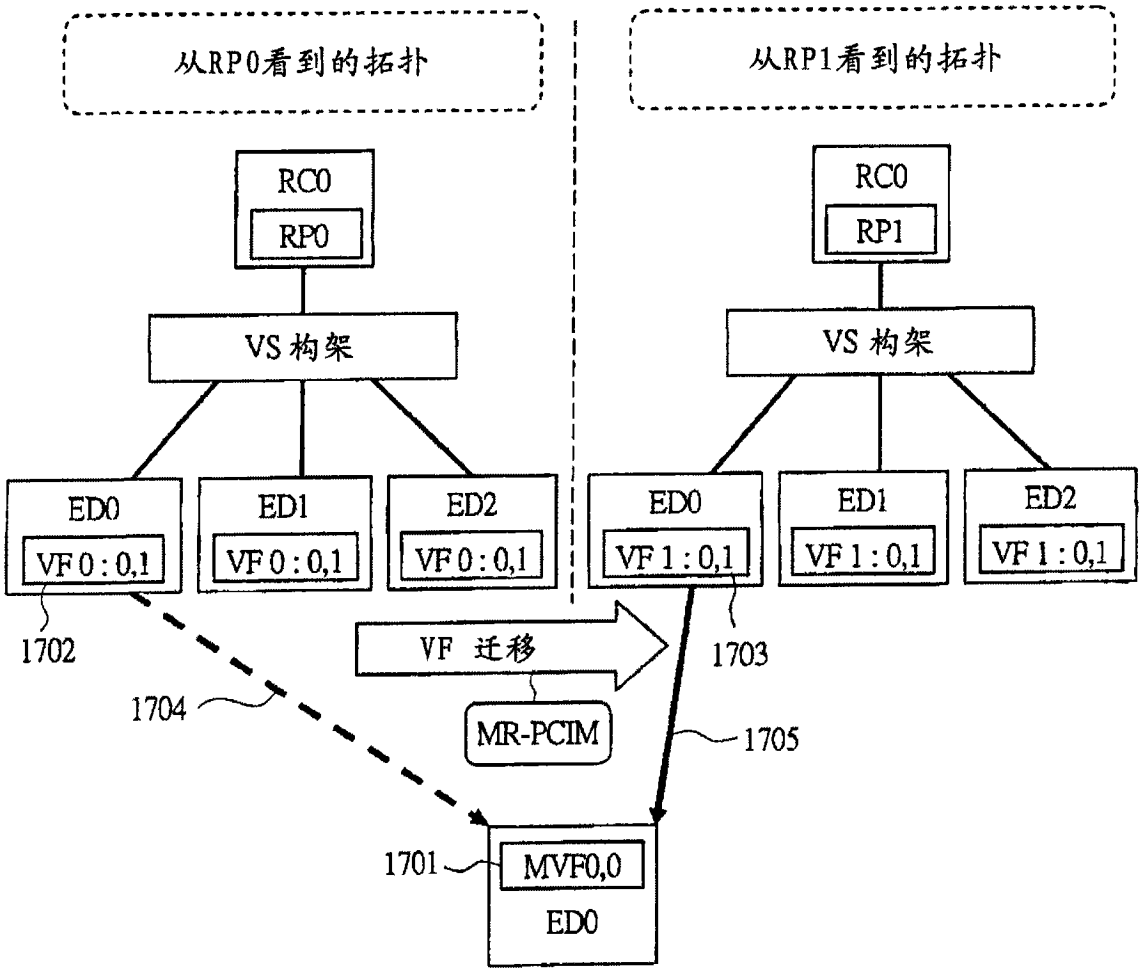

RC0
RC1
RP0
RP1
RP2
RP3
K2
L1
L3
K1
故障切换
SW00
SW01
L4
L2
SW10
SW11
SW12
L5
ED0
ED1
ED2

图 16



1701 : MVF
1702 : VF（迁移前）
1703 : VF（迁移后）
1704 : VF映射（迁移前）
1705 : VF映射（迁移后）

图 17

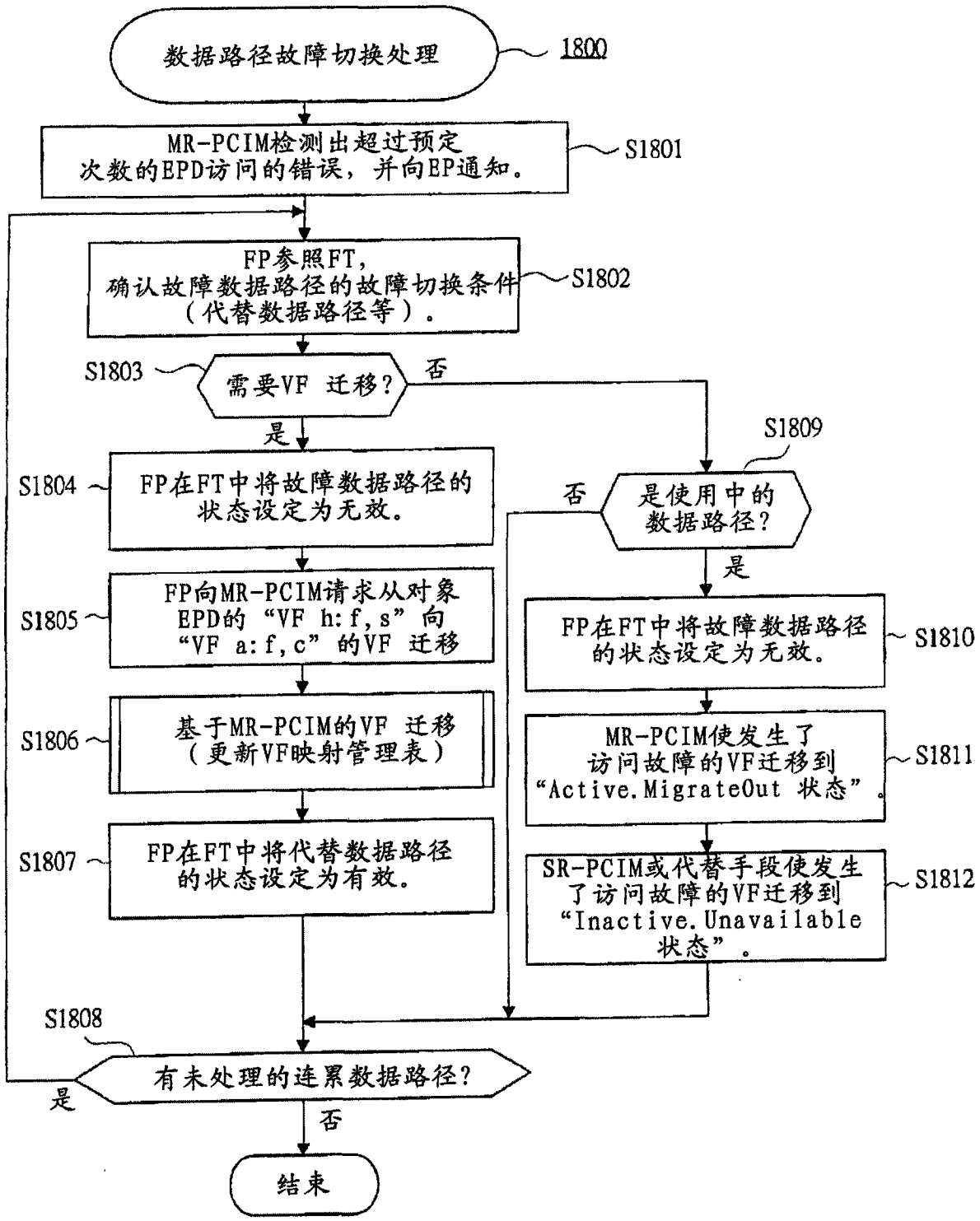
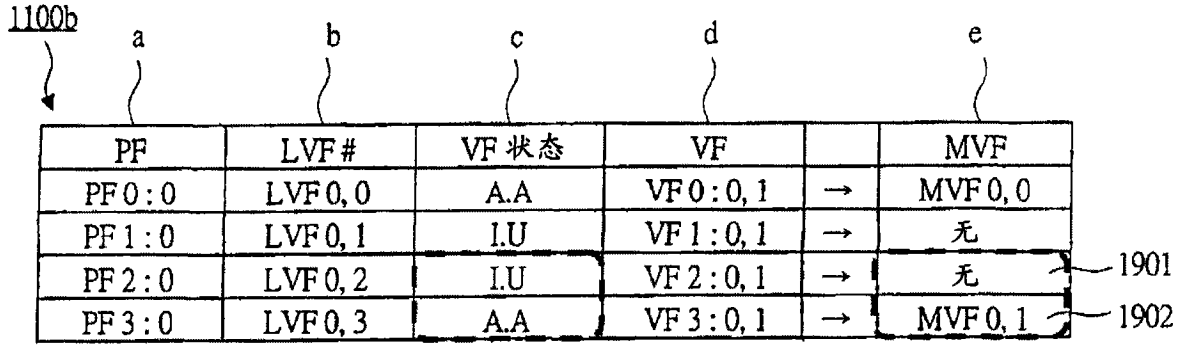
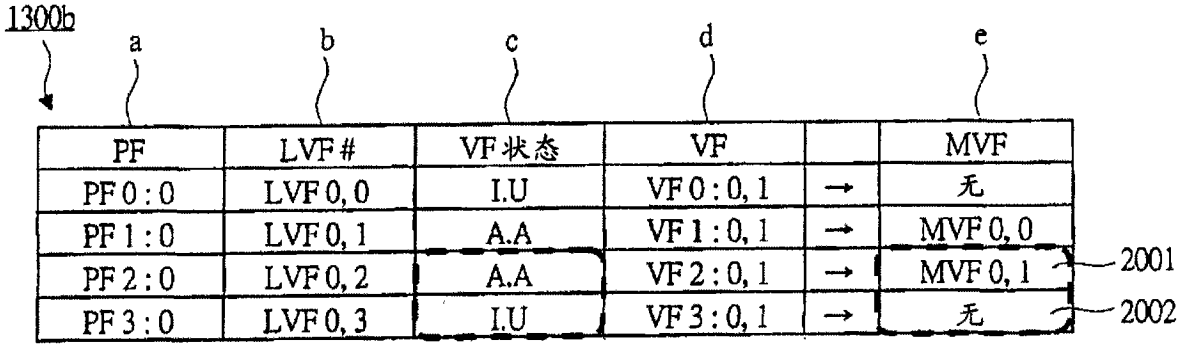

数据路径故障切换处理
1800
MR-PCIM检测出超过预定次数的EPD访问的错误，并向EP通知。
S1801
FP参照FT，确认故障数据路径的故障切换条件（代替数据路径等）。
S1802
S1803
需要VF 迁移？
否
是
S1809
S1804
FP在FT中将故障数据路径的状态设定为无效。
否
是使用中的数据路径？
是
S1805
FP向MR-PCIM请求从对象EPD的“VF h:f,s”向“VF a:f,c”的VF 迁移
FP在FT中将故障数据路径的状态设定为无效。
S1810
S1806
基于MR-PCIM的VF 迁移（更新VF映射管理表）
MR-PCIM使发生了访问故障的VF迁移到“Active.MigrateOut 状态”。
S1811
S1807
FP在FT中将代替数据路径的状态设定为有效。
SR-PCIM或代替手段使发生了访问故障的VF迁移到“Inactive.Unavailable状态”。
S1812
S1808
有未处理的连累数据路径？
是
否
结束

图 18



PF	LVF#	VF 状态	VF		MVF
PF 0 : 0	LVF 0, 0	A.A	VF 0 : 0, 1	→	MVF 0, 0
PF 1 : 0	LVF 0, 1	I.U	VF 1 : 0, 1	→	无
PF 2 : 0	LVF 0, 2	I.U	VF 2 : 0, 1	→	无
PF 3 : 0	LVF 0, 3	A.A	VF 3 : 0, 1	→	MVF 0, 1

图 19



PF	LVF#	VF 状态	VF		MVF
PF 0 : 0	LVF 0, 0	I.U	VF 0 : 0, 1	→	无
PF 1 : 0	LVF 0, 1	A.A	VF 1 : 0, 1	→	MVF 0, 0
PF 2 : 0	LVF 0, 2	A.A	VF 2 : 0, 1	→	MVF 0, 1
PF 3 : 0	LVF 0, 3	I.U	VF 3 : 0, 1	→	无

图 20

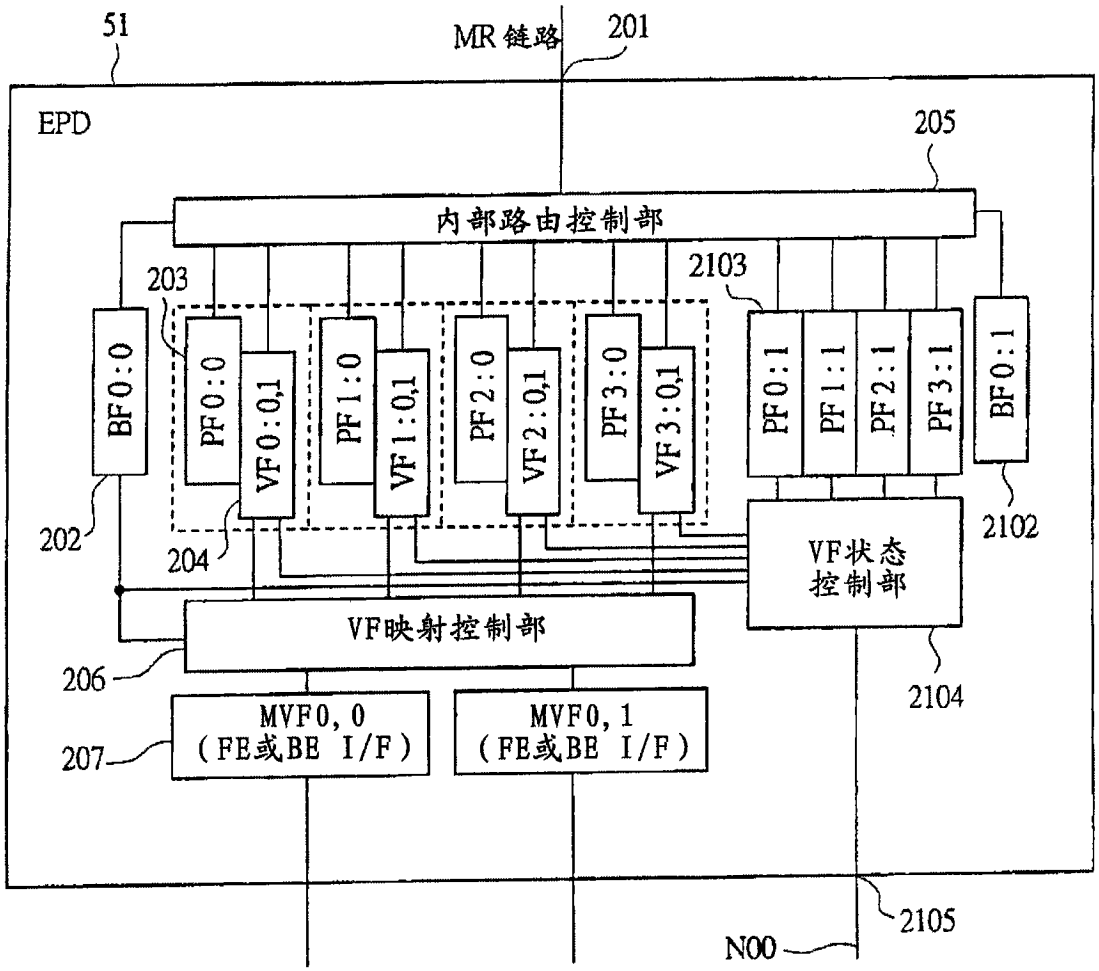

51
MR链路
201
EPD
205
内部路由控制部
2103
203
BF0:0
PF0:0
VF0:0,1
PF1:0
VF1:0,1
PF2:0
VF2:0,1
PF3:0
VF3:0,1
PF0:1
PF1:1
PF2:1
PF3:1
BF0:1
202
204
2102
VF状态
控制部
VF映射控制部
206
2104
MVF0,0
(FE或BE I/F)
MVF0,1
(FE或BE I/F)
207
2105
N00

图 21

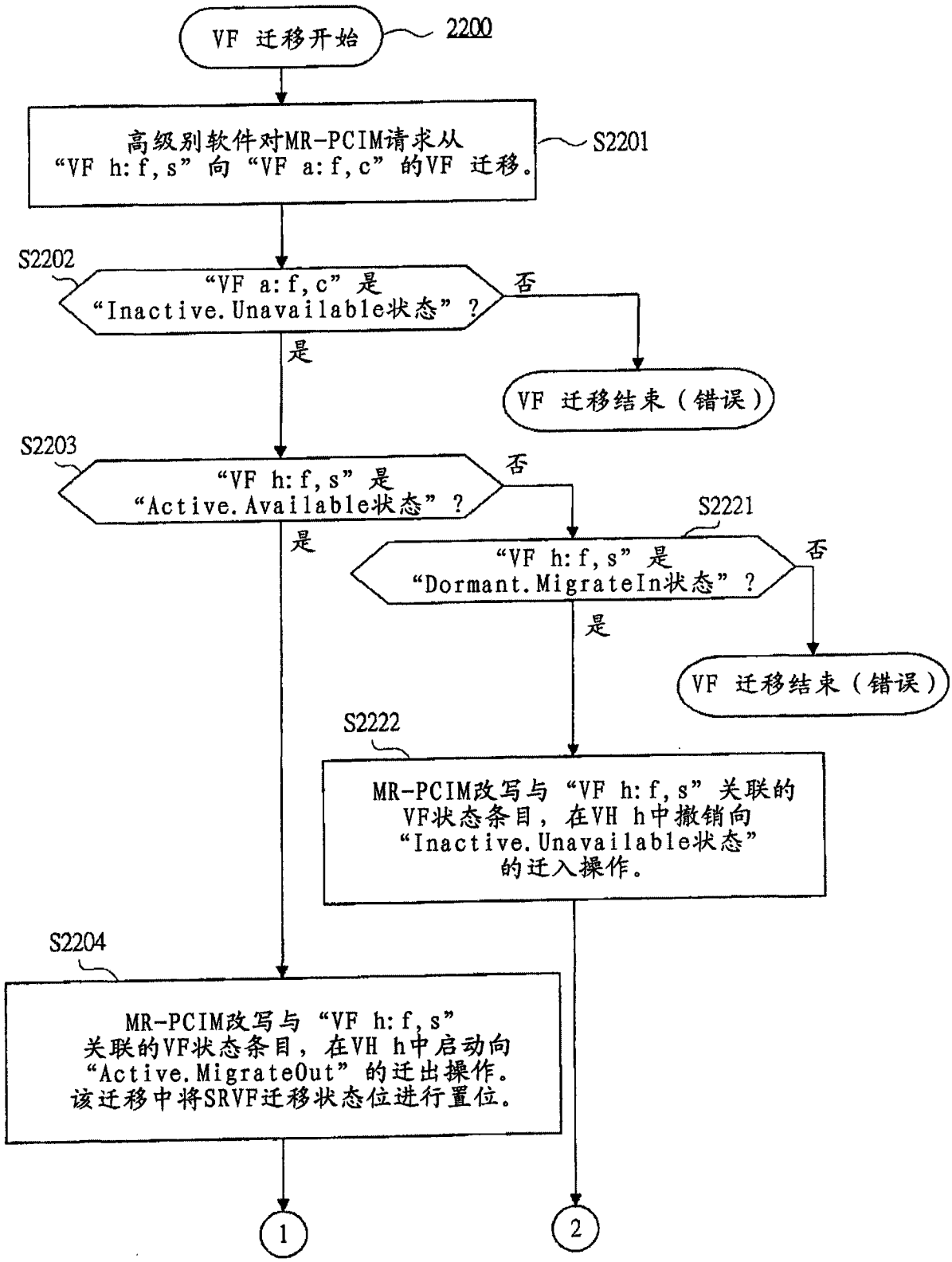

VF 迁移开始
2200
高级别软件对MR-PCIM请求从
“VF h:f,s” 向 “VF a:f,c” 的VF 迁移。
S2201
S2202
“VF a:f,c” 是
“Inactive.Unavailable状态” ?
否
是
VF 迁移结束（错误）
S2203
“VF h:f,s” 是
“Active.Available状态” ?
否
是
S2221
“VF h:f,s” 是
“Dormant.MigrateIn状态” ?
否
是
VF 迁移结束（错误）
S2222
MR-PCIM改写与 “VF h:f,s” 关联的
VF状态条目，在VH h中撤销向
“Inactive.Unavailable状态”
的迁入操作。
S2204
MR-PCIM改写与 “VF h:f,s”
关联的VF状态条目，在VH h中启动向
“Active.MigrateOut” 的迁出操作。
该迁移中将SRVF迁移状态位进行置位。
1
2

图 22(A)

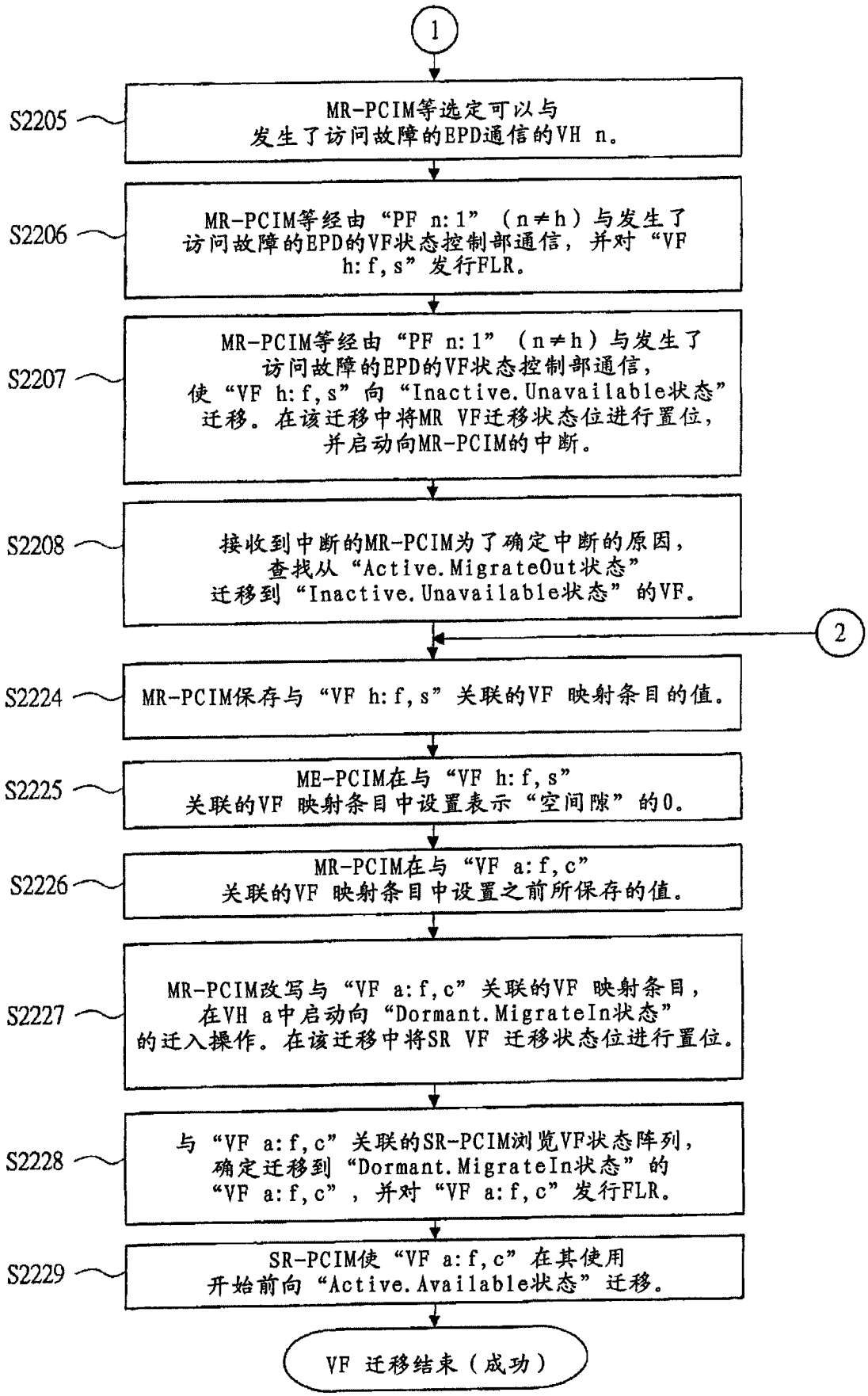

1
S2205
MR-PCIM等选定可以与
发生了访问故障的EPD通信的VH n。
S2206
MR-PCIM等经由“PF n:1”（n≠h）与发生了
访问故障的EPD的VF状态控制部通信，并对“VF
h:f,s”发行FLR。
S2207
MR-PCIM等经由“PF n:1”（n≠h）与发生了
访问故障的EPD的VF状态控制部通信，
使“VF h:f,s”向“Inactive.Unavailable状态”
迁移。在该迁移中将MR VF迁移状态位进行置位，
并启动向MR-PCIM的中断。
S2208
接收到中断的MR-PCIM为了确定中断的原因，
查找从“Active.MigrateOut状态”
迁移到“Inactive.Unavailable状态”的VF。
2
S2224
MR-PCIM保存与“VF h:f,s”关联的VF 映射条目的值。
S2225
ME-PCIM在与“VF h:f,s”
关联的VF 映射条目中设置表示“空间隙”的0。
S2226
MR-PCIM在与“VF a:f,c”
关联的VF 映射条目中设置之前所保存的值。
S2227
MR-PCIM改写与“VF a:f,c”关联的VF 映射条目，
在VH a中启动向“Dormant.MigrateIn状态”
的迁入操作。在该迁移中将SR VF 迁移状态位进行置位。
S2228
与“VF a:f,c”关联的SR-PCIM浏览VF状态阵列，
确定迁移到“Dormant.MigrateIn状态”的
“VF a:f,c”，并对“VF a:f,c”发行FLR。
S2229
SR-PCIM使“VF a:f,c”在其使用
开始前向“Active.Available状态”迁移。
VF 迁移结束（成功）

图 22(B)

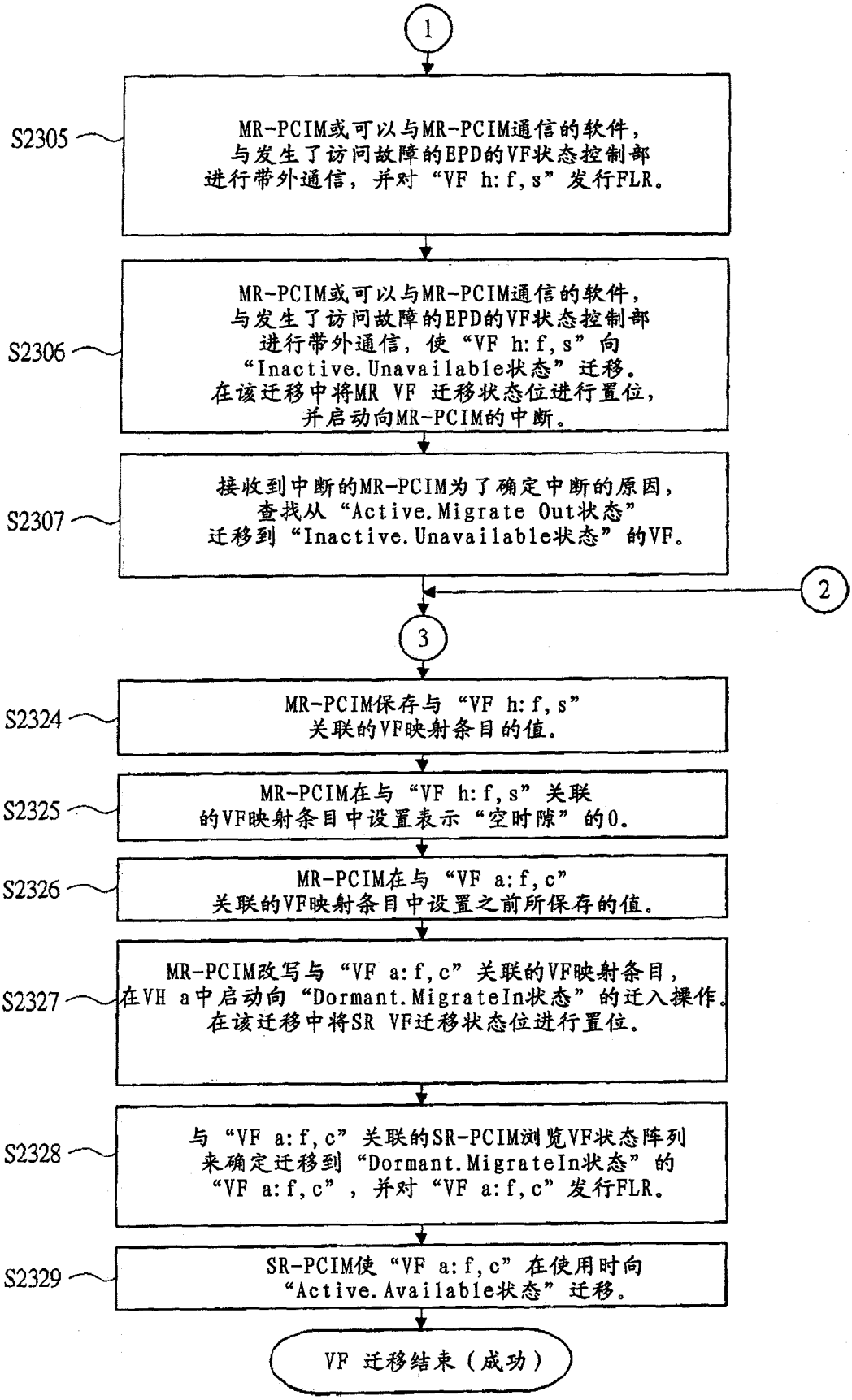

1
S2305
MR-PCIM或可以与MR-PCIM通信的软件，
与发生了访问故障的EPD的VF状态控制部
进行带外通信，并对“VF h:f,s”发行FLR。
S2306
MR-PCIM或可以与MR-PCIM通信的软件，
与发生了访问故障的EPD的VF状态控制部
进行带外通信，使“VF h:f,s”向
“Inactive.Unavailable状态”迁移。
在该迁移中将MR VF 迁移状态位进行置位，
并启动向MR-PCIM的中断。
S2307
接收到中断的MR-PCIM为了确定中断的原因，
查找从“Active.Migrate Out状态”
迁移到“Inactive.Unavailable状态”的VF。
2
3
S2324
MR-PCIM保存与“VF h:f,s”
关联的VF映射条目的值。
S2325
MR-PCIM在与“VF h:f,s”关联
的VF映射条目中设置表示“空时隙”的0。
S2326
MR-PCIM在与“VF a:f,c”
关联的VF映射条目中设置之前所保存的值。
S2327
MR-PCIM改写与“VF a:f,c”关联的VF映射条目，
在VH a中启动向“Dormant.MigrateIn状态”的迁入操作。
在该迁移中将SR VF迁移状态位进行置位。
S2328
与“VF a:f,c”关联的SR-PCIM浏览VF状态阵列
来确定迁移到“Dormant.MigrateIn状态”的
“VF a:f,c”，并对“VF a:f,c”发行FLR。
S2329
SR-PCIM使“VF a:f,c”在使用时向
“Active.Available状态”迁移。
VF 迁移结束（成功）

图 23

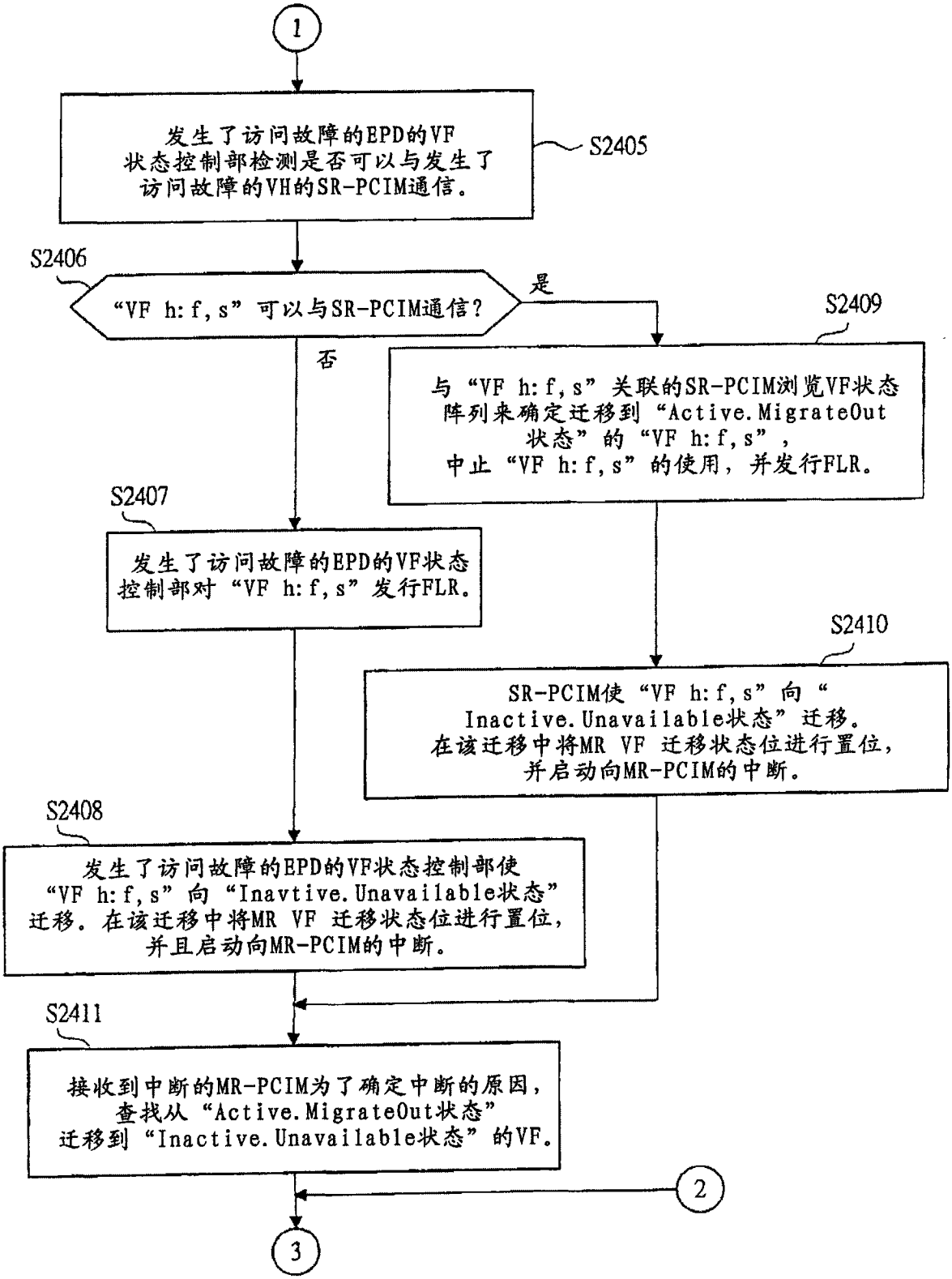

1
发生了访问故障的EPD的VF
状态控制部检测是否可以与发生了
访问故障的VH的SR-PCIM通信。
S2405
S2406
“VF h:f,s”可以与SR-PCIM通信?
是
否
S2409
与“VF h:f,s”关联的SR-PCIM浏览VF状态
阵列来确定迁移到“Active.MigrateOut
状态”的“VF h:f,s”,
中止“VF h:f,s”的使用,并发行FLR。
S2407
发生了访问故障的EPD的VF状态
控制部对“VF h:f,s”发行FLR。
S2410
SR-PCIM使“VF h:f,s”向“
Inactive.Unavailable状态”迁移。
在该迁移中将MR VF 迁移状态位进行置位,
并启动向MR-PCIM的中断。
S2408
发生了访问故障的EPD的VF状态控制部使
“VF h:f,s”向“Inavtive.Unavailable状态”
迁移。在该迁移中将MR VF 迁移状态位进行置位,
并且启动向MR-PCIM的中断。
S2411
接收到中断的MR-PCIM为了确定中断的原因,
查找从“Active.MigrateOut状态”
迁移到“Inactive.Unavailable状态”的VF。
2
3

图 24

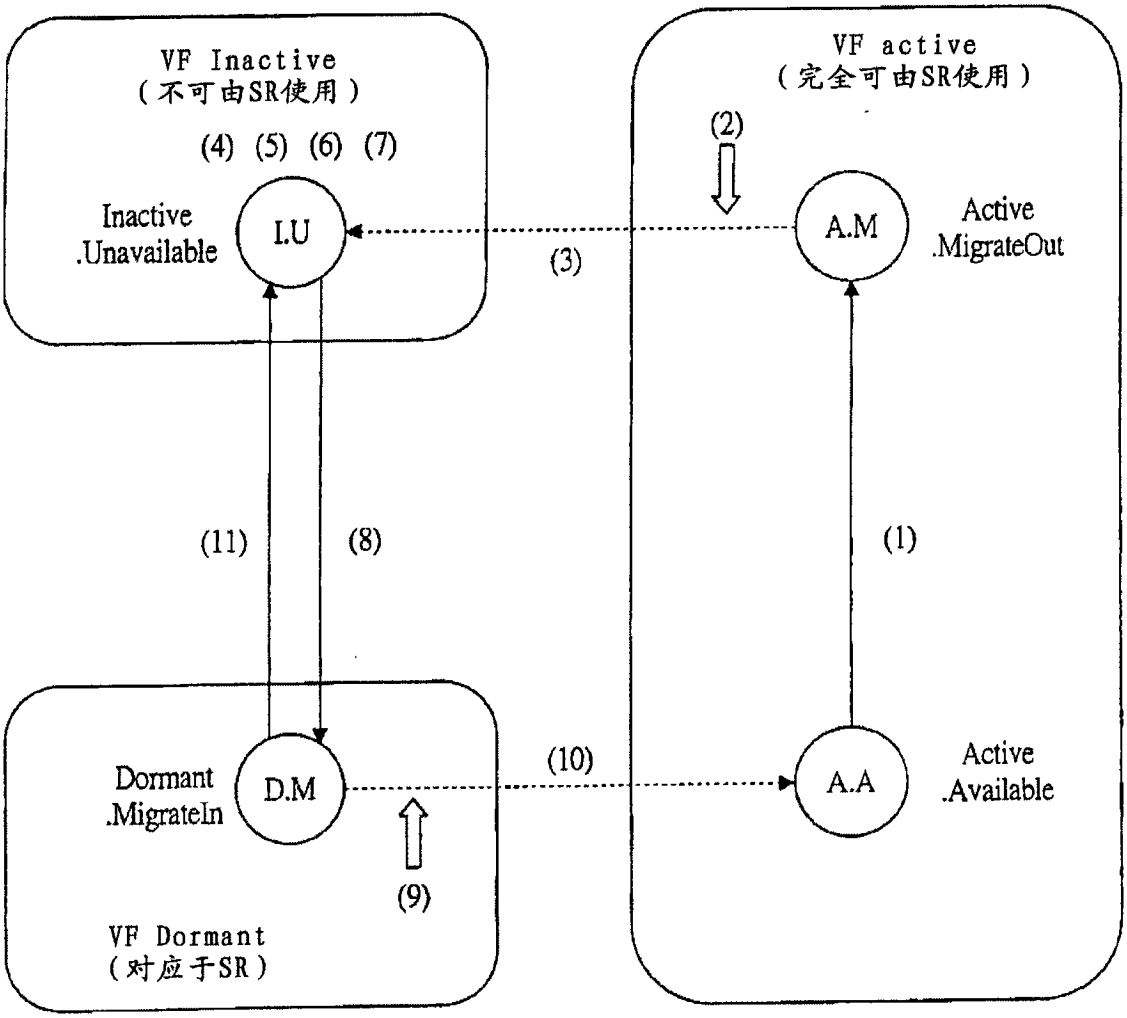

VF Inactive
（不可由SR使用）
(4) (5) (6) (7)
Inactive
.Unavailable
I.U
(3)
VF active
（完全可由SR使用）
(2)
A.M
Active
.MigrateOut
(11)
(8)
(1)
Dormant
.MigrateIn
D.M
(10)
A.A
Active
.Available
(9)
VF Dormant
（对应于SR）

图 25

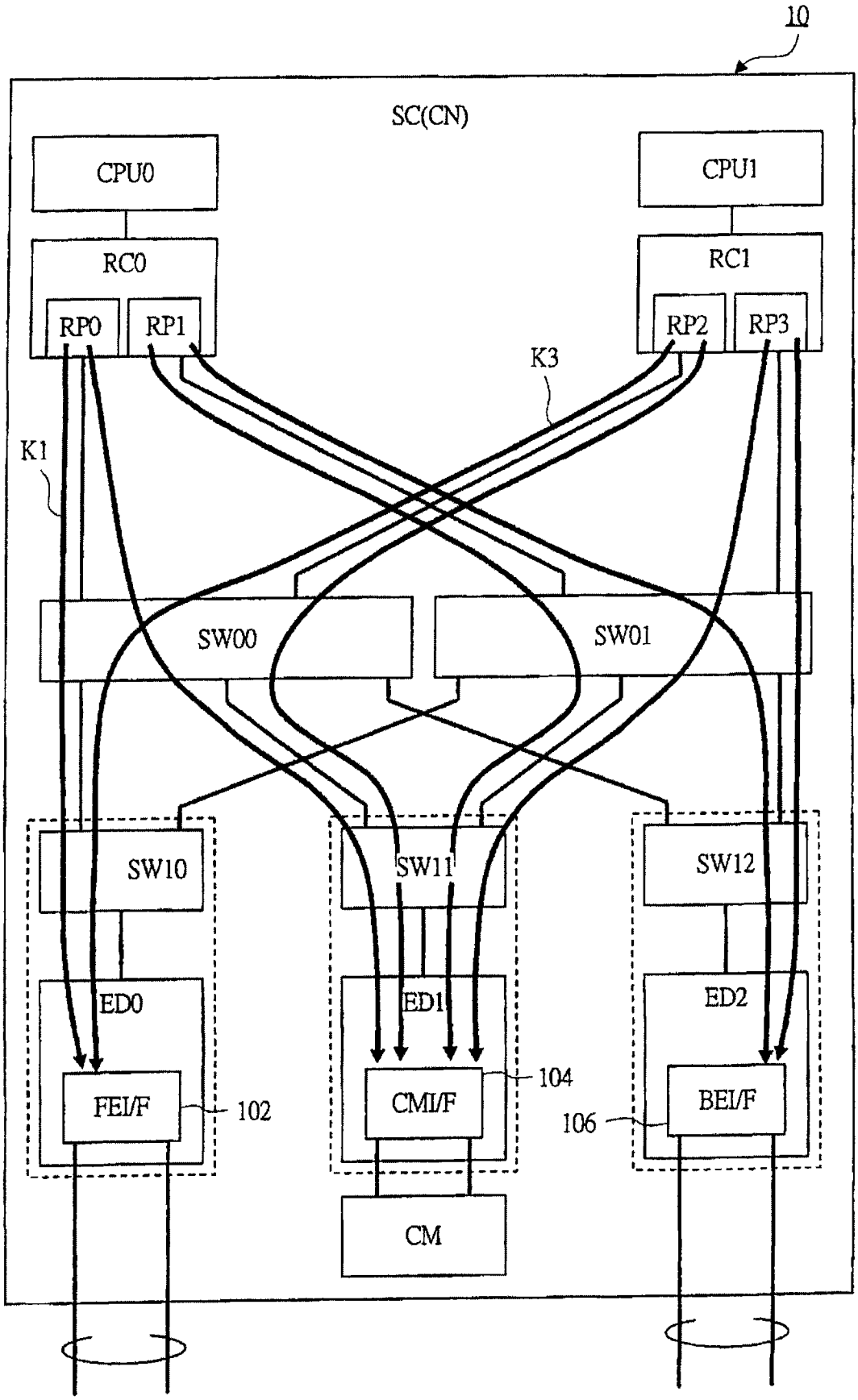

10
SC(CN)
CPU0
CPU1
RC0
RC1
RP0
RP1
RP2
RP3
K3
K1
SW00
SW01
SW10
SW11
SW12
ED0
ED1
ED2
FEI/F
102
CMI/F
104
BEI/F
106
CM

图 26

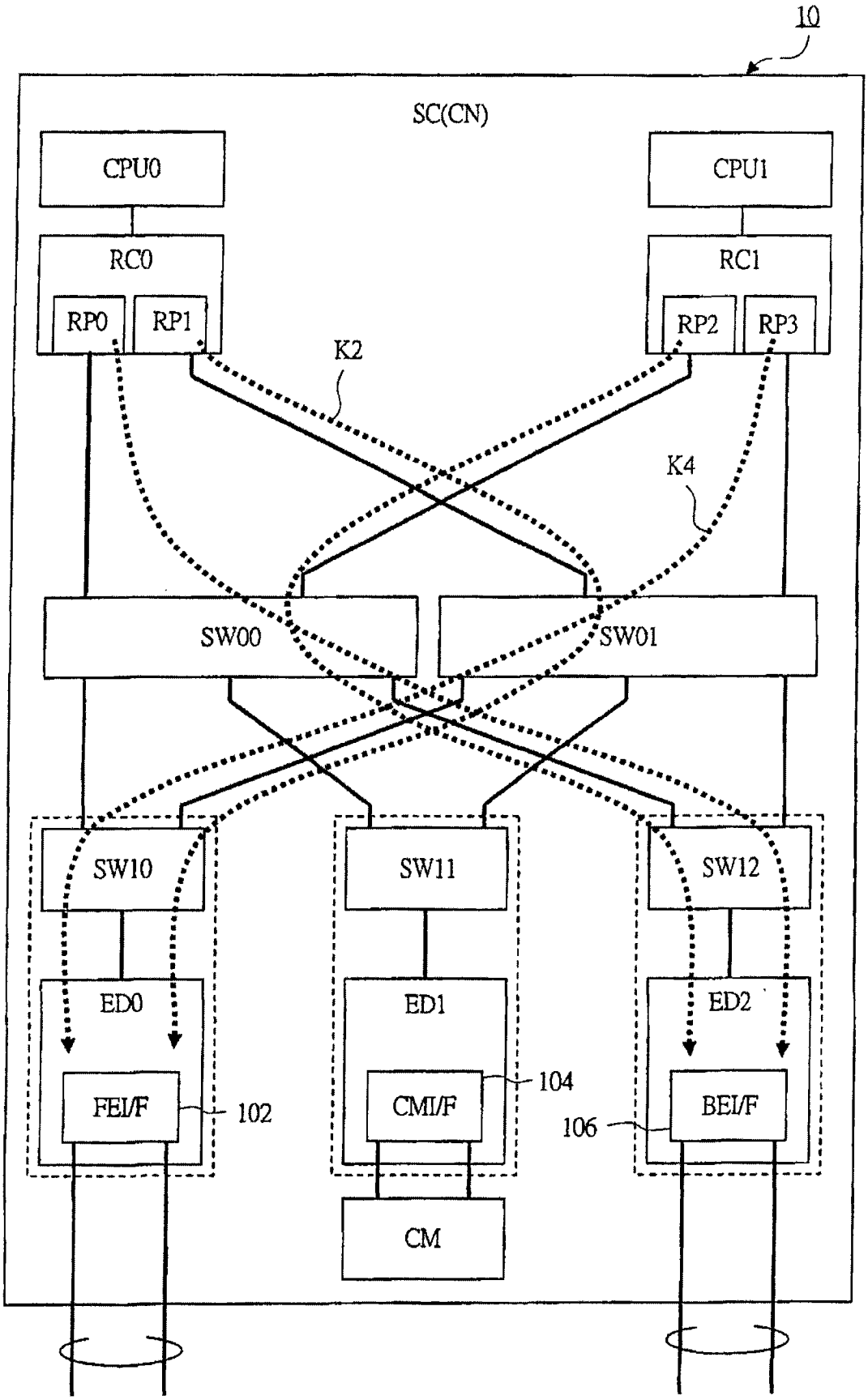

10
SC(CN)
CPU0
CPU1
RC0
RC1
RP0
RP1
RP2
RP3
K2
K4
SW00
SW01
SW10
SW11
SW12
ED0
ED1
ED2
FEI/F
102
CMI/F
104
BEI/F
106
CM

图 27

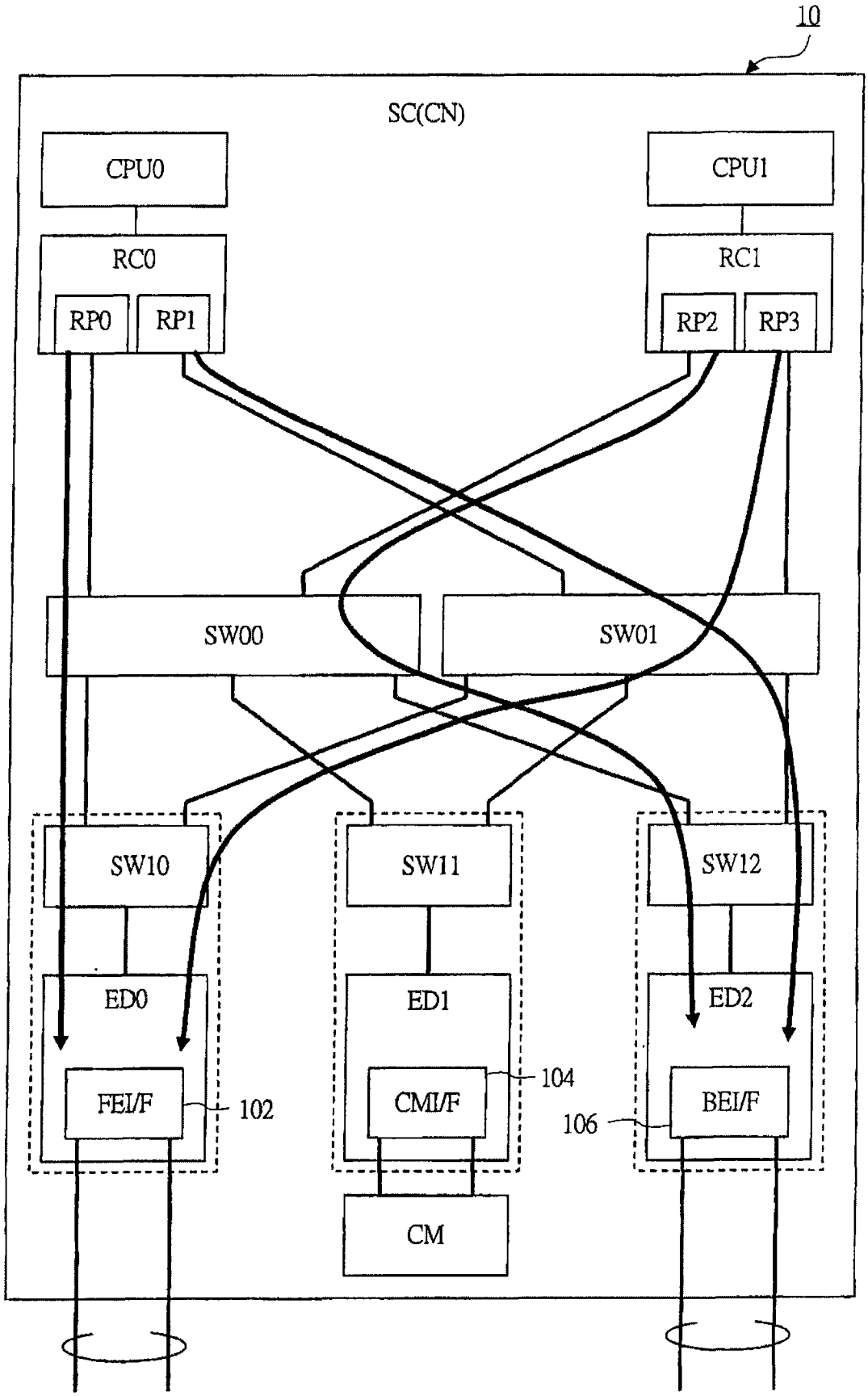

10
SC(CN)
CPU0
CPU1
RC0
RC1
RP0
RP1
RP2
RP3
SW00
SW01
SW10
SW11
SW12
ED0
ED1
ED2
FEI/F
102
CMI/F
104
BEI/F
106
CM

图 28