(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 104813617 A
(43)申请公布日 2015.07.29

(21)申请号 201380058724.5
(22)申请日 2013.11.06
(30)优先权数据
13/674,315 2012.11.12 US
(85)PCT国际申请进入国家阶段日
2015.05.11
(86)PCT国际申请的申请数据
PCT/US2013/068641 2013.11.06
(87)PCT国际申请的公布数据
WO2014/074546 EN 2014.05.15
(71)申请人 阿尔卡特朗讯公司
地址 法国布洛涅-比扬古
(72)发明人 S·C·汉卡 R·H·J·达 席尔瓦
S·吴 A·C·常
(74)专利代理机构 北京市中咨律师事务所 11247
代理人 杨晓光 于静
(51)Int.Cl.
H04L 12/46(2006.01)
H04L 12/24(2006.01)
H04L 12/931(2006.01)
H04L 29/12(2006.01)
H04L 12/26(2006.01)

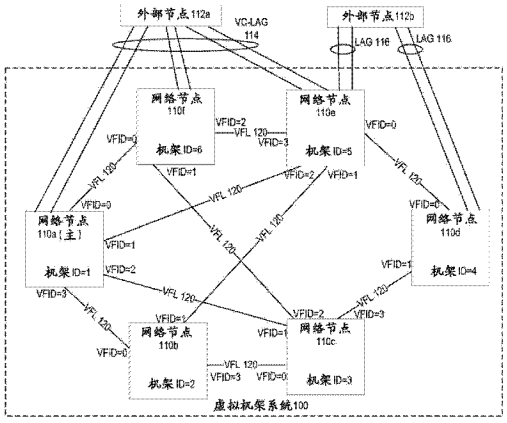
权利要求书2页 说明书23页 附图16页

(54)发明名称

在用于虚拟机架系统的可操作节点中确定是否发布管理动作触发虚拟机架分裂告警的网络节点和方法

(57)摘要

虚拟机架系统包括配置有主虚拟机架地址的多个网络节点。网络节点经由虚拟组织链路(VFL)连接，所述虚拟组织链路提供在网络节点之间交换分组的连接。虚拟机架系统可操作以提供告警来帮助防止响应于管理动作的虚拟机架分裂事件。虚拟机架系统的拓扑被分析以确定一个或更多的管理动作的可能影响。基于该分析，可能直接或间接地导致虚拟机架分裂的管理动作被请求时生成告警。



CN 104813617 A

1.一种在虚拟机架系统中可操作的网络节点,包括 :

一个或多个网络接口模块,可操作地耦合到多个虚拟组织链路 (VFL),其中所述 VFL 可操作地耦合到所述虚拟机架系统中的多个网络节点 ;

至少一个管理模块,可操作用于 :

接收第一管理动作,其中所述第一管理动作包括所述虚拟机架系统中的多个网络节点中的至少一个的机架标识符 ;

访问一个或多个重置列表,其中所述一个或多个重置列表包括所述虚拟机架系统中的多个网络节点中的一个或多个的机架标识符以及重置是否触发虚拟机架分裂的相应指示符 ;以及

从所述一个或多个重置列表和所述第一管理动作确定是否发布管理动作触发虚拟机架分裂的告警。

2.如权利要求 1 所述的网络节点,其中所述至少一个管理模块可进一步操作用于 :

当所述一个或多个重置列表指示发布管理动作触发虚拟机架分裂的告警时,从所述一个或多个重置列表确定可被重置以避免虚拟机架分裂的一个或多个其他网络节点的机架标识符是否被列入。

3.如权利要求 2 所述的网络节点,其中所述管理动作包括以下中的至少一个 :

用于重置由所述机架标识符标识的所述多个网络节点中的至少一个的管理命令 ;

用于重启由所述机架标识符标识的所述多个网络节点中的至少一个的管理命令 ;

用于将由所述机架标识符标识的所述多个网络节点中的至少一个断电的管理命令 ;

用于设置由所述机架标识符标识的所述多个网络节点中的至少一个的关闭状态的管理命令 ;和

用于设置由所述机架标识符标识的所述多个网络节点中的至少一个的非服务状态的管理命令。

4.如权利要求 1 所述的网络节点,其中所述至少一个管理模块可进一步操作用于 :

访问虚拟机架系统的拓扑信息的拓扑数据库,其中所述拓扑信息包括所述多个网络节点的机架标识符以及所述网络接口模块和耦合到所述网络节点中的多个虚拟组织链路 (VFL) 的 VFL 成员端口接口的标识 ;和

基于所述拓扑数据库生成所述一个或多个重置列表,其中所述一个或多个重置列表包括下列中的一个或多个 :

所述虚拟机架系统中所述多个网络节点中的一个或多个的机架标识符以及机架标识符的重置是否触发虚拟机架分裂的相应指示符 ;

所述虚拟机架系统中所述多个网络节点中的一个或多个的一个或多个网络接口模块 NIM 的 NIM 标识符以及一个或多个 NIM 标识符的重置是否触发虚拟机架分裂的相应指示符 ;和

所述虚拟机架系统中所述多个网络节点中的一个或多个的 VFL 成员端口标识符以及 VFL 成员端口的重置是否触发虚拟机架分裂的相应指示符。

5.一种在虚拟机架系统中可操作节点中的方法,所述虚拟机架系统包括可操作地耦合到多个网络节点的多个虚拟组织链路 (VFL),所述方法包括 :

接收第一管理动作,其中所述第一管理动作包括所述虚拟机架系统中的多个网络节点

中的至少一个的机架标识符；

访问一个或多个重置列表，其中所述一个或多个重置列表包括所述虚拟机架系统中多个网络节点中的一个或多个的机架标识符以及所述多个网络节点中的一个或多个的重置是否触发虚拟机架分裂的相应指示符；以及

从所述一个或多个重置列表和所述第一管理动作确定是否发布管理动作触发虚拟机架分裂的告警。

6. 如权利要求 5 所述的方法，进一步包括：

当所述一个或多个重置列表指示发布管理动作触发虚拟机架分裂的告警时，从所述一个或多个重置列表确定可被重置以避免虚拟机架分裂的一个或多个其他网络节点的机架标识符是否被列入。

7. 如权利要求 6 所述的方法，其中所述管理动作包括以下中的至少一个：

用于重置由所述机架标识符标识的所述多个网络节点中的至少一个的管理命令；

用于重启由所述机架标识符标识的所述多个网络节点中的至少一个的管理命令；

用于将由所述机架标识符标识的所述多个网络节点中的至少一个断电的管理命令；

用于设置由所述机架标识符标识的所述多个网络节点中的至少一个的关闭状态的管理命令；和

用于设置由所述机架标识符标识的所述多个网络节点中的至少一个的非服务状态的管理命令。

8. 如权利要求 7 所述的方法，进一步包括：

访问虚拟机架系统的拓扑信息的拓扑数据库，其中所述拓扑信息包括所述多个网络节点的机架标识符以及所述网络接口模块和耦合到所述网络节点中的多个虚拟组织链路 (VFL) 的 VFL 成员端口接口的标识；和

基于所述拓扑数据库生成所述一个或多个重置列表，其中所述一个或多个重置列表包括下列中的一个或多个：

所述虚拟机架系统中所述多个网络节点中的一个或多个的机架标识符以及机架标识符的重置是否触发虚拟机架分裂的相应指示符；

所述虚拟机架系统中所述多个网络节点中的一个或多个的一个或多个网络接口模块 NIM 的 NIM 标识符以及一个或多个 NIM 标识符的重置是否触发虚拟机架分裂的相应指示符；和

所述虚拟机架系统中所述多个网络节点中的一个或多个的 VFL 成员端口标识符以及 VFL 成员端口的重置是否触发虚拟机架分裂的相应指示符。

9. 如权利要求 8 所述的方法，进一步包括：

接收第二管理动作，其中所述第二管理动作包括所述多个网络节点之一中的网络接口模块 (NIM) 之一的第一 NIM 标识符；和

访问一个或多个重置列表以确定是否发布管理动作触发虚拟机架分裂的告警。

10. 如权利要求 9 所述的方法，进一步包括：

接收第三管理动作，其中所述第三管理动作包括第一 VFL 成员端口接口标识符；和

访问一个或多个重置列表以确定是否发布管理动作触发虚拟机架分裂的告警。

在用于虚拟机架系统的可操作节点中确定是否发布管理动作触发虚拟机架分裂告警的网络节点和方法

技术领域

[0001] 本发明一般涉及数据网络，特别地涉及在一个或多个数据网络的节点之间提供拓扑冗余和弹性的系统和方法。

背景技术

[0002] 数据网络包括各种计算设备，例如彼此通信和/或与连接到网络的各种其他网络单元或远程服务器进行通信的网络个人计算机、IP 电话设备或服务器。例如，数据网络可以包括但不限于城域以太网或企业以太网网络，这些数据网络网络支持多个应用，包括例如 IP 语音 (VoIP)、数据和视频应用。这种网络有规律地包括互连的节点，通常称为交换机或路由器，用来经过网络路由业务量。

[0003] 数据网络面临的关键挑战之一是需要网络弹性，即不管可能的组件故障、链路故障或类似的情况而维持高可用性的能力，高可用性是提供令人满意的网络性能的关键。通过拓扑冗余可以部分地获得网络弹性，即通过提供冗余节点（和冗余节点内的组件）以及节点之间的多个物理路径来防止单点故障，并且部分地通过 L2/L3 协议在出现故障时利用所述冗余来聚合于替代路径以交换/路由通过网络的业务流。可以理解的是，检测和聚合时间必须在网络中快速发生（有利地，小于一秒），以实现到备用路径的无缝转换。各种类型的网络拓扑在网络内被实现以提供网络单元之间的冗余，例如环形网络、部分网状网络、全网状网络、集线器网络等。网络单元之间的聚合时间和冗余通常根据网络中实现的网络类型的不同而变化。

[0004] 网络单元的架构也是可变的并影响网络的弹性。例如，各种节点架构包括单个交换单元、可堆叠交换单元、基于多插槽机架的网络单元等。通常，根据成本和网络需求，这些节点架构的类型中的一个被选择并实施为一种网络拓扑类型。然而，一旦实现，有时难以升级或从一种网络拓扑类型转变到另一种网络拓扑类型。在网络拓扑内，也难以从一种节点架构的类型转换到另一种节点架构的类型或者在一个网络内结合各种节点架构的类型。

[0005] 因此，需要一种系统和方法，用于在一个或多个不同类型的网络拓扑内的具有一个或多个不同类型的节点架构之间提供弹性。

附图说明

[0006] 图 1a-c 示出了根据本发明的虚拟机架系统的实施例的示意性框图；

[0007] 图 2 示出了根据本发明的虚拟机架系统中的网络拓扑发现过程的实施例的逻辑流程图；

[0008] 图 3 示出了根据本发明的虚拟机架系统的网络节点中的拓扑数据库的实施例的示意性框图；

[0009] 图 4 示出了根据本发明的虚拟机架系统中的网络节点的实施例的示意性框图；

[0010] 图 5 示出了根据本发明的虚拟机架系统的网络节点的网络接口模块的实施例的

示意性框图；

[0011] 图6示出了根据本发明的虚拟机架系统中的分组的前挂(prepended)报头的实施例的示意性框图；

[0012] 图7示出了根据本发明的虚拟机架系统中流经网络节点的分组的实施例的示意性框图；

[0013] 图8示出了根据本发明的虚拟机架管理应用的实施例的示意性框图；

[0014] 图9示出了根据本发明的虚拟机架系统中主(master)地址保留(retention)的实施例的示意性框图；

[0015] 图10示出了根据本发明的虚拟机架系统中主地址释放的实施例的示意性框图；

[0016] 图11示出了根据本发明的虚拟机架系统中主网络节点故障的实施例的示意性框图；

[0017] 图12示出了根据本发明的虚拟机架系统中VFL故障的实施例的示意性框图；

[0018] 图13示出了根据本发明的虚拟机架系统中从主网络节点的故障恢复的方法的实施例的逻辑流程图；

[0019] 图14示出了根据本发明的虚拟机架系统中生成一个或多个重置列表的方法的实施例的逻辑流程图；

[0020] 图15示出了根据本发明的虚拟机架系统中生成重置列表的实施例的示意性框图；和

[0021] 图16示出了根据本发明的虚拟机架系统中生成网络节点重置列表的方法的实施例的逻辑流程图；

[0022] 图17示出了根据本发明的虚拟机架系统中生成重置列表的另一个实施例的示意性框图；

[0023] 图18示出了根据本发明的虚拟机架系统中生成网络接口模块重置列表和VFL成员端口重置列表的方法的实施例的逻辑流程图；

[0024] 图19示出了根据本发明的虚拟机架系统中用于处理管理动作以有助于防止虚拟机架分裂的方法的实施例的逻辑流程图；以及

[0025] 图20示出了根据本发明的虚拟机架系统中处理用于一个或多个参数的配置的管理动作以有助于防止虚拟机架分裂的方法的实施例的逻辑流程图。

具体实施方式

[0026] 本申请中涉及到如下标准并在此引入作为参考：1)链路聚合控制协议(LACP)，之前由IEEE802.3ad任务组在2000年3月将其添加于IEEE802.3标准的条款43，并且目前被结合在2008年11月3日的IEEE802.1AX-2008中；以及2)IEEE标准802.1Q，虚拟桥接局域网，2003版。

[0027] 图1a示出包括通过专用链路集合群可操作地耦合的多个网络节点110的虚拟机架系统100的实施例，所述专用链路集合群用于传送控制和寻址信息，被称为虚拟组织(fabric)链路(VFL)120。VFL 120及其操作详细描述于2011年1月20日提交的美国专利申请No.13/010,168，题为“SYSTEM AND METHOD FOR MULTI-CHASSIS LINK AGGREGATION”，出于所有目的，该未决申请在此引入作为参考并作为本美国实用专利申请的一部分。VFL

120 提供网络节点 110 之间用于交换信息的连接，所述信息涉及流量转发、MAC 寻址、组播流、地址解析协议 (ARP) 表、第 2 层控制协议（如生成树、以太网环路保护、逻辑链路检测协议）、路由协议（如 RIP、OSPF、BGP) 以及网络节点和外部链路的状态。

[0028] 在实施例中，多个网络节点 110 作为具有统一管理能力的单个虚拟网络节点来工作。例如网络节点 110a 的主网络节点被选择，并且主网络节点 110 的本地 MAC 地址被其他网络节点 110 采纳用作虚拟机架系统 100 的主 MAC 地址。外部节点 112 使用主 MAC 地址以寻址虚拟机架系统 100 中的网络节点 110。同样地，网络节点 110 与外部节点 112 透明地操作并且被外部节点 112 视为单一逻辑设备。

[0029] 外部节点 112 使用单一干线 (trunk) 或链路、链路聚合群 (LAG)116 或虚拟机架链路聚合群 (VC-LAG)114，可操作地耦合到虚拟机架系统 100 中的一个或多个网络节点 110。为了提供增强的弹性和移除单点或甚至两点故障，VC-LAG 114 可操作地将外部节点耦合到虚拟机架系统 100 中的两个或更多网络节点 110。外部节点可以使用负载均衡技术来通过可用的 VC-LAG 的链路 114 分配流量。例如，物理链路 VC-LAG114 的物理链路之一被外部节点选择以基于负载均衡算法（通常包括作用在源和目的地因特网协议 (IP) 或媒体接入控制 (MAC) 地址信息上的散列函数）传输分组，以便更有效地使用带宽。

[0030] 在正常操作期间，虚拟机架系统内的网络节点 110 共享主 MAC 地址用作许多种层 2 和层 3 协议的系统标识。例如，生成树协议和 LACP 协议使用主 MAC 地址作为虚拟机架系统 110 的标识符。因特网协议 (IP) 路由也利用主 MAC 地址来向网络中的外部网络单元标识虚拟机架系统 100，例如对端 (peer) 使用 MAC 地址作为发往虚拟机架系统 100 的分组的以太网目的地地址。同样地，虚拟机架系统 100 中的网络节点 110 被外部网络节点 112 视为单个逻辑节点。此外，虚拟机架系统 100 中的网络节点 110 被作为具有的统一管理、操作和维护管理系统的单个节点管理。

[0031] 由于虚拟机架系统 100 中的网络节点 110 被外部节点 112 视为单一逻辑设备，外部节点 112 可操作地主动地转发 VC-LAG114 的所有链路上的流量。该特征使得外部节点 112 向网络节点 110 多导向 (multiple homing) 成为可能而无需外部节点和网络节点之间的生成树协议，同时还促进了边缘上行链路故障以及网络节点 110 故障的载体 - 等级检测和聚合时间。对于虚拟机架系统 100 所有 VC-LAG 114 上行链路的主动转发模式的另一个优点是 VC-LAG114 链路带宽使用效率的增加。

[0032] 在虚拟机架系统 100 中，为网络节点 110 分配被称为机架标识符或机架 ID 的全局唯一标识符。在虚拟机架系统 100 内，网络节点 110 分配内部 VFL 标识符 (VFID) 至它配置的每一个 VFL 120。由于 VFL 的 VFID 被用于 VFL 120 的内部标识和配置，网络节点 110 可以向 VFL 120 分配与另一网络节点 110 向该 VFL 120 分配的 VFID 相同或不同的 VFID。VFL120 提供用于在网络节点 110 之间交换信息的连接，所述信息涉及流量转发、MAC 寻址、组播流、地址解析协议 (ARP) 表、第 2 层控制协议（如生成树，以太网环路保护、逻辑链路检测协议）、路由协议（如 RIP、OSPF、BGP)，这详细地描述于 2011 年 1 月 20 日提交的美国专利申请 No. 13/010,168“SYSTEM AND METHOD FOR MULTI-CHASSIS LINK AGGREGATION”。在实施例中，网络节点 110 之间诸如媒体访问控制 (MAC) 地址表的第 2 层地址表的同步由通过第 2 层上的分组流经由 VFL120 驱动以及由周期性保活机制驱动，通过所述周期性保活机制，拥有给定 MAC 地址的网络节点 110 洪泛 (flood) 携带有这样的 MAC 地址作为源地址的特

定分组。该同步机制还需要实现标准 MAC 冲洗 (flushing) 机制来处理网络节点 110 或其一些组件出问题 (go down) 的情况。通过未知目的地 MAC 地址的洪泛，在 VFL120 上的 MAC 地址源学习得以实现。在源学习过程中，网络节点 110 在 VFL 120 上交换具有前挂报头的分组，所述前挂报头包括源 MAC 地址和相关的硬件设备信息，诸如源机架 ID、源网络接口标识符和源端口标识符信息。网络节点 110 使用该信息来维护同步的 MAC 地址表，所述同步的 MAC 地址表使用基于最小消息传送的 MAC 表同步。利用同步的 MAC 地址表，网络节点 110 可操作以处理和转发在虚拟机架系统 100 中的网络节点 110 之间的分组。

[0033] 图 1a 示出了网络节点 110 耦合在部分网状网络拓扑中。然而，虚拟机架系统 100 中的网络节点 110 可以耦合在多种类型网络拓扑的任意一种中而不影响虚拟机架系统 100 的工作。图 1b 示出了具有被配置在环形网络拓扑中通过 VFL120 耦合的多个网络节点 110 的虚拟机架系统 100。图 1c 示出了具有被配置在集线器和轮辐 (spoke) 或星型网络拓扑中的多个网络节点 110 的虚拟机架系统 100。这里没有描述虚拟机架系统 100 还支持的例如线性、树、全网状网络、混合式等的其他网络拓扑。为了支持多种不同类型的网络拓扑，虚拟机架系统 100 中的网络节点 110 可操作用于执行网络拓扑发现过程。

[0034] 图 2 示出了虚拟机架系统 100 中网络拓扑发现过程 130 实施例的逻辑流程图。该过程由虚拟机架系统 100 中的活动网络节点 110 在启动、重启、网络中状态改变的指示或在预定的时间段执行。在步骤 132 中，网络节点 110 检测到它正在以虚拟机架模式操作。例如，网络节点 110 的一个或多个参数被配置成指示虚拟机架的操作模式。网络节点 110 检测到所述参数指示虚拟机架模式操作（而不是例如单机模式或多机架模式）。接着在步骤 134 中，网络节点 110 执行一个或多个控制协议来发现虚拟机架系统 100 中的其他网络节点 110 并且交换拓扑和配置信息。网络节点 110 使用该信息来构建虚拟机架系统 100 的拓扑数据库。所述拓扑数据库包括：其他网络节点 110 的标识信息（例如本地 MAC 地址、机架标识符）、托管活动 VFL120（或其他活动的交换机间链路）的网络接口的标识信息、VFL 120 的标识信息以及在网络节点 110 上关联的成员端口。这样，网络节点 110 了解到虚拟机架系统 100 中网络节点 110 之间的活动连接以及其他网络节点 110 的配置信息。下面的表 1 是在发现阶段之后网络节点 110a 的拓扑数据库的示例，在本例中例如机架 ID = 1。表 1 包括存储在拓扑数据库中的示例性信息，但其他未示出的信息和数据也可包括在拓扑数据库中。此外，拓扑数据库可以存储在单独的数据库或表中或者与网络节点 110 中的其他表或数据库结合。

[0035]

拓扑数据库 - 机架 1			
本地机架数据	邻居 [1]	邻居[2]	邻居[3]
机架 ID = 1 正常运行时间 = 4 分 50 秒 优先级 = 100 机架 MAC = A 机架群 = 0 主 CMM = CMM-A 机架类型 = OS10K 作用 = 未指定 状态 = 未指定	机架 ID = 2 正常运行时间 = 5 分 10 秒 优先级 = 100 机架 MAC = B 机架群 = 0 主 CMM = CMM-A 机架类型 = OS10K 作用 = 未指定 状态 = 未指定	机架 ID = 4 正常运行时间 = 5 分 5 秒 优先级 = 100 机架 MAC = D 机架群 = 0 主 CMM = CMM-B 机架类型 = OS10K 作用 = 未指定 状态 = 未指定	机架 ID = 3 正常运行时间 = 5 分 1 秒 优先级 = 100 机架 MAC = C 机架群 = 0 主 CMM = CMM-A 机架类型 = OS10K 作用 = 未指定 状态 = 未指定

[0036] 表 1

[0037] 在图 2 的步骤 136 中，主网络节点被选择以执行虚拟机架系统 100 的管理和其他任务。然后主网络节点的本地 MAC 地址被其他网络节点 110 采用。下面的表 2 是具有机架 ID ＝ 1 的所选主网络节点 110 的拓扑数据库的示例。如表 2 所示，在拓扑数据库中，机架 ID ＝ 1 的网络节点被表示为具有主角色并且其他节点被表示为具有从角色。

[0038]

拓扑数据库 - 机架 1			
本地 机架数据	邻居 [1]	邻居 [2]	邻居 [3]
机架 ID = 1	机架 ID = 2	机架 ID = 4	机架 ID = 3

[0039]

正常运行时间 = 5 分 50 秒 优先级 = 100 机架 MAC = A 机架群 = 0 主 CMM = CMM-A 机架类型 = OS10K 作用 = 主 状态 = 运行	正常运行时间 = 6 分 10 秒 优先级 = 100 机架 MAC = B 机架群 = 0 主 CMM = CMM-A 机架类型 = OS10K 作用 =从 状态 = 运行	正常运行时间 = 6 分 5 秒 优先级 = 100 机架 MAC = D 机架群 = 0 主 CMM = CMM-B 机架类型 = OS10K 作用 =从 状态 = 运行	正常运行时间 = 6 分 1 秒 优先级 = 100 机架 MAC = C 机架群 = 0 主 CMM = CMM-A 机架类型 = OS10K 作用 =从 状态 = 运行

[0040] 表 2

[0041] 主网络节点 110 的选择基于包括机架优先级、正常运行时间 (up time)、机架 ID 和机架 MAC 地址的参数的优先级排序的列表。正常运行时间参数可以将优先级给予操作时间

较长的网络节点 110。机架优先级参数是用户配置的优先级，其定义主网络节点 110 的用户偏好而不考虑机架 ID 或正常运行时间。各种不同参数的使用增加了主网络节点 110 的选择的灵活性。在拓扑数据库中所示的机架群参数标识了虚拟机架系统 100。具有不同机架群标识的一个或多个附加虚拟机架系统 100 中也可以在网络中操作。拓扑数据库还标识网络节点 110 中的活动或者主控制管理器模块 (CMM) 以及网络节点 110 的机架类型。

[0042] 在网络拓扑发现过程 130 的步骤 138 中，网络节点 110 执行一个或多个协议来监视虚拟机架系统 100 中的连接和网络节点 110 的状态或状况。拓扑数据库中维护网络节点 110 的当前状态。检测到的虚拟机架系统 100 中的网络节点 110 的状态改变可以发起路由的改变、主节点的改变等。通过拓扑自发现和网络节点 110 的监控，虚拟机架系统 100 可操作用于以最小的预配置和干预来支持多个不同类型的网络拓扑。

[0043] 图 3 示出了在选择主网络节点 110 之后，虚拟机架系统 100 中的网络节点 110 的拓扑数据库 144 的例子。在这个例子中，网络节点 110a 被采用作为主网络节点，网络节点 110b 和 110c 是从节点。网络节点 110a 的本地 MAC 地址 (例如，主 MAC 地址 = A) 被网络节点 110a-c 采用作为虚拟机架 MAC 地址。另外，主 MAC 地址 (MAC = A) 被采用作为用于管理应用的应用 MAC 地址。

[0044] 虚拟机架系统 100 也可操作用于包括具有一个或多个不同类型的节点架构的网络节点 110，例如单个模块、可堆叠、或基于多插槽机架的架构。图 4 示出了在具有不同类型的节点架构的虚拟机架系统 100 中的网络节点 110 实施例的示意性框图。在这个例子中，网络节点 110a 具有基于多插槽机架 (multi-slot chassis-based) 的架构，该架构具有多个网络接口模块 152a-n。通常，基于多插槽机架的架构共有一个外壳、控制管理模块 (CMM) 150a-b 和公共电源，所述公共电源具有例如线路卡或端口模块的一个或多个网络接口模块 (NIM) 152a-n。网络接口模块 152n 包括排队模块 212 和交换模块 210 并且这些模块通过集成到该机架的底板 (backplane) 的组织交换 (fabric switch) 214 连接。

[0045] 在本例中的网络节点 110b 具有可堆叠节点架构并包括通过底板连接 142 耦合的多个网络单元 140a-n。每个网络单元 140a-n 可操作为独立节点，并且包含它自己的外壳、控制管理模块 (CMM) 150、交换模块 210、排队模块 212 和电源。在一些堆叠架构中，一种网络单元 (在该例中的网络单元 140a) 被指定为堆栈的主要或主单元以用于管理目的。

[0046] 网络节点 110c 具有单个模块节点架构，例如单独的可堆叠单元 140 或者可选地，具有单个网络接口模块 152 的基于多插槽机架的架构。

[0047] 网络节点 110a-c 对应于图 1a-c 中的虚拟机架系统 100 中的网络单元 110 中的一个或多个。例如，虚拟机架系统 100 可操作以包括仅仅具有基于多插槽机架的节点架构的网络节点 110 或包括仅仅具有可堆叠节点架构的网络节点 110 或包括具有两种或更多类型的节点架构的网络节点 110 的组合，这里的两种或更多类型例如基于多插槽机架架构、可堆叠节点架构和单个模块节点架构。虽然未示出，虚拟机架系统 100 还可以包括由其他类型的节点架构和配置构成的网络节点 110。

[0048] 网络节点 110a 和网络节点 110b 通过 VFL120a 可操作地耦合。网络节点 110a 和 110b 以内部 VFL 的标识符 (VFID) 标出 (designate) VFL120a，例如如图 3 所示对于网络节点 110a，VFID = 3，以及对于网络节点 110b，VFID = 0。网络节点 110a 和网络节点 110c 通过 VFL120b 可操作地耦合。网络节点 110a 和 110c 以内部 VFL 标识符 (VFID) 标出 VFL120b，

例如如图 3 所示对于网络节点 110a，VFID ＝ 2 以及对于网络节点 110c，VFID ＝ 1。此外，如图 1a-c 所示，网络节点 110a-c 还通过附加的 VFL120 可操作地耦合到一个或多个其他网络节点 110。网络节点 110a 和 110b 之间的 VFL120a 被描述为虚拟机架系统 100 中的不同网络节点 110 之间的 VFL 120 的操作和配置的概括 (generalization)。

[0049] 网络节点 110a 和网络节点 110b 之间的 VFL 120a 可操作地耦合到一个或多个交换模块 210 中的一个或多个 VFL 成员端口。为了在一个或多个端口、链路或模块发生故障情况下的冗余，VFL 120a 可操作地包括多个聚合链路，所述聚合链路使用网络节点 110a 和 110b 的不同交换模块 210 之间的 LACP 或类似聚合协议生成。例如在图 4 中，VFL120a 包括网络节点 110a 的 NIM 152a 和网络节点 110b 的堆叠式网络单元 140a 之间的物理链路的第一子集 A 以及网络节点 110a 的 NIM 152b 和网络节点 110b 的堆叠式网络单元 140b 之间的物理链路的第二子集 B。

[0050] 网络节点 110 在虚拟机架系统 100 中被分配唯一的机架标识符。用于每个网络节点 110 的机架 ID 是唯一和全局的，并且通过拓扑发现，网络节点 110 可以意识到虚拟机架系统 100 中其对等网络节点 110 的机架 ID 的。另外，诸如网络节点 110 中的交换模块 210 和端口接口的各种组件的唯一硬件设备标识符或模块标识符 (MID) 被生成以允许用于本地和远程管理的目的。在实施例中，交换模块 210 的硬件设备标识符 MID 在虚拟机架系统内具有全局意义，而对于其他组件——例如排队模块 212——的 MID 可以只具有本地意义。例如，分配给交换模块 210 的硬件设备标识符可以被其他网络节点 110 所知，而其他设备的硬件设备标识符被限制到本地网络节点 110 并且对于其他网络节点 110 而言没有什么意义。例如，交换模块 210 的端口接口被分配全局唯一硬件设备标识符，该标识符包括机架 ID、交换模块 ID 和端口接口 ID。在实施例中，虚拟机架系统中的网络节点 110 在前挂报头模式操作来通过 VFL 120 交换数据和控制分组。

[0051] 图 5 更详细地示出了在前挂报头模式中操作的网络接口模块 (NIM) 152 实施例的示意性框图。尽管示出了网络接口模块 152，堆叠式网络单元 140 或单模块网络单元可操作用于执行类似功能以在前挂报头模式中操作。交换模块 210 包括从虚拟机架系统 100 连接到外部节点 112 的多个外部端口 240。外部端口 240 中的一个或多个可以包括用于 VC-LAG 114、LAG116、单一干线或其他干线群、固定链路等的成员端口。外部端口 240 可具有相同的物理接口类型，例如铜端口 (CAT-5E/CAT-6)、多模光纤端口 (SX) 或单模光纤端口 (LX)。在另一个实施例中，外部端口 240 可以具有一个或多个不同的物理接口类型。

[0052] 外部端口 240 被分配有例如设备端口值的外部端口接口标识符 (端口 ID)，例如与交换模块 210 相关的 gport 和 dport 值。在实施例中，网络节点 110 的机架 ID、交换模块 210 的 MID、以及外部端口接口标识符 (端口 ID) 被用作虚拟机架系统 100 的网络节点 110 中的物理外部端口接口 240 的全局唯一标识符。在另一个实施例中，全局唯一模块标识符 (MID) 基于机架标识符被分配给虚拟机架系统的网络节点的交换模块 210。例如，交换 MID 0-31 被分配给机架 ID ＝ 1，交换 MID 32-63 被分配到机架 ID ＝ 2 等。在这种情况下，全局唯一交换 MID 和外部端口标识符 (端口 ID) 被用作虚拟机架系统 100 的网络节点 110 中的物理外部端口接口 240 的全局唯一标识符。

[0053] 当外部端口 240 上接收到分组时，交换模块 210 将该分组传送给前挂报头接口 (PPHI)246，该接口添加前挂报头 (或以其他方式修改分组报头) 以包括与所述分组相关的

源和/或目的地MAC地址相关联的硬件设备信息(HDI)。在实施例中,前挂报头可以包括例如分组优先级和负载均衡标识符的其他信息。为了获得与分组的MAC地址相关联的HDI信息,PPHI在MAC/HDI转发表250中执行查找过程。存储在地址表存储器248中的MAC/HDI转发表250包括MAC地址的列表和相关联的硬件设备信息。硬件设备信息唯一地标识网络节点110、交换模块210和/或用于路由分组的端口接口240。硬件设备信息包括例如交换模块210的机架ID、MID和/或与目的地MAC地址相关联的端口240的端口接口ID。MAC/HDI转发表250可以包括一个或多个表,如源干线映射、干线位图表、干线群表、VLAN映射表等。在实施例中,MAC/HDI转发表250或其部分可以位于NIM152的排队模块或其他模块。

[0054] 基于拓扑数据库144,在网络节点110生成VFL路由配置表254以确定单播流量的路由。VFL路由配置表254包括机架ID和相关联的VFL ID(VFID)。与机架ID相关联的VFID标识虚拟机架系统100中的VFL 120以用于向由目的地机架ID标识的网络节点110路由分组。在另一个实施例中,当全局唯一模块标识符(MID)分配给虚拟机架系统100中的网络节点110的交换模块210时,VFL路由配置表254包括全局唯一MID和相关联的VFL ID(VFID)。在实施例中,VFL路由配置表254是使用最短路径算法、基于流量的算法或其他类型的路由算法而生成的。用于图1a所示的虚拟机架系统100的VFL路由配置表254的例子在如下的表3中示出。

[0055]

VFL 路由	
机架1上的配置	
目的地机架 ID/MID	外出 VFL ID
1 (MID=0-31)	N/A (本地)
2 (MID=32-63)	3
3 (MID=64)	2
4 (MID=65-97)	2 或 1
5 (MID=98)	1
6 (MID=99-115)	0

VFL 路由	
机架2上的配置	
目的地机架 ID/MID	外出 VFL ID
1 (MID=0-31)	0
2 (MID=32-63)	N/A (本地)
3 (MID=64)	3
4 (MID=65-97)	3 或 1
5 (MID=98)	1
6 (MID=99-115)	0 or 1

[0056]

VFL 路由 机架 3 上的配置	
目的地机架 ID/MID	外出 VFL ID
1 (MID=0-31)	1
2 (MID=32-63)	0
3 (MID=64)	N/A (本地)
4 (MID=65-97)	3
5 (MID=98)	3 或 2
6 (MID=99-115)	2

VFL 路由 机架 4 上的配置	
目的地机架 ID/MID	外出 VFL ID
1 (MID=0-31)	0 或 1
2 (MID=32-63)	0 或 1
3 (MID=64)	1
4 (MID=65-97)	N/A (本地)
5 (MID=98)	0
6 (MID=99-115)	0 或 1

VFL 路由 机架 5 上的配置	
目的地机架 ID/MID	出 VFL ID
1 (MID=0-31)	2
2 (MID=32-63)	1
3 (MID=64)	1 或 0
4 (MID=65-97)	0
5 (MID=98)	N/A (本地)
6 (MID=99-115)	1

VFL 路由 机架 6 上的配置	
目的地机架 ID/MID	出 VFL ID
1 (MID=0-31)	0
2 (MID=32-63)	0 或 1
3 (MID=64)	1
4 (MID=65-97)	1 或 2
5 (MID=98)	2
6 (MID=99-115)	N/A (本地)

[0057] 表 3

[0058] 尽管 MAC/HDI 转发表 250 和 VFL 路由表 254 被示为地址表存储器 248 中的独立表，这些表可以被组合或来自表之一的数据可以被包括到另一个表或者这些表可以被分成一个或更多其他表。

[0059] 在实施例中，分组的前挂报头中的硬件设备信息 HDI 包括与目的地机架 ID 相关的 VFL 端口 252 的外出 VFID，如表 3 所示。该前挂报头还包括与接收该分组的源端口相关的硬件设备信息 HDI，例如端口接口 ID、交换模块 210 的 MID 和机架 ID。在实施例中，诸如 VLAN ID、分组类型（多播、单播、广播）、分组优先级和负载均衡标识符的附加信息也被添加到前挂报头中。

[0060] 具有前挂报头的分组然后被发送到排队模块 212，用于在组织交换 214 上路由。基于 VFL 路由配置表 254，排队模块 212 将带有前挂报头的分组路由到连接到外出 VFL 120 的交换模块 210。

[0061] 排队模块 212 包括分组缓冲器 260、用于提供流量和缓冲器管理的队列管理 262 以及全局 HDI 地址表 264。全局 HDI 地址表 264 将外出 VFL ID 映射到其他 NIM15 中的一个或

多个的排队模块 212 中的适当队列。例如，排队模块 212 将分组交换到 VFL 端口接口 252 中的一个或多个的适当出口队列以用于在外出 VFL 120 上传输。在实施例中，对应于特定 VFL 端口接口的外出队列的确定可操作地基于前挂报头中的负载均衡标识符。

[0062] 尽管交换模块 210 和排队模块 212 被示为单独的集成电路或模块，这些模块的一个或多个功能或组件可被包括在其他模块或者被组合成可选模块或者在一个或多个集成电路中被实现。

[0063] 图 6 示出了虚拟机架系统 100 中的分组的前挂报头的实施例的的示意性框图。前挂报头 300 包括源 HDI 302、目的地 HDI 304、VLAN ID 306、分组类型 308、源 MAC 地址 310 和目的地 MAC 地址 312 的字段。在实施例中，前挂报头还可以包括负载均衡标识符 314 和分组优先级 316。目的地 HDI 304 包括例如端口标识符（设备端口 (dport) 和 / 或全局端口值 (GPV)）、交换模块 210 的 MID 和 / 或与目的地 MAC 地址相关联的目的地网络节点 110 的机架 ID。源 HDI 302 包括例如端口标识符（设备端口 (dport) 和 / 或全局端口值 (GPV)）、交换模块 210 的 MID 和 / 或与源 MAC 地址相关联的源网络节点的机架 ID。负载均衡标识符 314 由排队模块 212 使用来确定 VFL 成员端口以用于在外出 VFL120 上分组的传输。分组优先级 316 由排队模块 212 使用来确定特定优先级队列。

[0064] 图 7 示出了在虚拟机架系统 100 中流经网络节点 110a 至另一网络节点 11b 的分组的实施例的示意性框图。在这一例子中，来自具有源 MAC 地址“MAC1”的虚拟机架系统 100 的外部设备 300 发送具有目的地 MAC 地址“MAC2”的分组。在该示例中具有机架 ID = 1 的网络节点 110a 在交换模块 210n 上的外部端口接口 240 接收分组，所述交换模块 210n 例如具有 MID = 31，所述外部端口接口 240 例如具有端口 ID = 2。交换模块 210n 提取目的地 MAC 地址 MAC2 并且在 MAC/HDI 转发表 250 中执行地址表查找来确定与目的地 MAC 地址 MAC2 相关的硬件设备信息 (HDI)。目的地 HDI 可以包括，例如目的地机架 ID 和与目的地 MAC 地址相关的设备模块标识符 (MID) 和端口标识符。目的地 HDI 还可以包括到与目的地 MAC 地址相关联的目的地设备的路径中的一个或多个其他网络节点或硬件模块的标识符。当目的地 MAC 地址与另一个网络节点相关时，例如目的地机架 ID 不是本地机架 ID，则交换模块 210 确定与目的地机架 ID 相关的外出 VFL ID。该外出 VFL ID 可以被添加到前挂报头中的目的地 HDI 中。对于图 5 所示的例子，VFL 路由表 254 指示目的地机架 ID = 2 与具有 VFID = 3 的 VFL 120 相关联。

[0065] 交换模块 210n 还在前挂报头中包括与初始 (originating) 外部端口接口相关的源硬件设备信息 (HDI)，例如端口 ID = 2。源 HDI 可以包括一个或多个硬件设备标识符，例如初始交换模块 210 的 MID、源端口标识符、源 NIM 152 的 MID、源机架 ID 等。此外，在实施例中，前挂报头包括基于从初始分组（源 MAC 地址、目的地 MAC 地址、源 IP 地址、目的地 IP 地址）检索出的参数确定的分组优先级和负载均衡标识符。

[0066] 具有前挂报头的分组被发送给排队模块 212n，排队模块 212n 然后确定网络节点 110 上的 NIM 152 以基于目的地 HDI 发送分组。当目的地 HDI 指示网络节点上的本地外部端口接口（例如基于前挂报头中包含的目的地 MID) 时，排队模块将分组放置到出口队列以传输给本地外部端口接口的对应 NIM 152。在图 5 所示的另一个例子中，当目的地 HDI 指示该分组需要在 VFL 120 上发送到虚拟机架系统 100 中的另一个网络节点 110 时，排队模块从 VFL ID 确定外出 NIM 152 来发送该分组。在该例子中，排队模块确定 VFID = 3 可操

作地耦合到 NIM 152a 并且经由组织交换 214 发送具有前挂报头的分组至 NIM 152a。在负载均衡方法中，当多个交换模块 210 可操作地耦合到外出 VFL 120 时，要传输的流量可以分布在多个交换模块 210 之间。另外，交换模块 210 上 VFL 成员端口（高优先级队列，低优先级等）的选择可操作地基于前挂报头携带的负载均衡标识符参数。NIM152a 上的排队模块 212a 接收具有前挂报头的分组并且将分组排队用于在具有 VFID ＝ 3 的 VFL120 上传输。然后，交换模块 210a 在具有 VFID ＝ 3 的 VFL120 上发送具有包括源和 / 或目的地 HDI 的前挂报头的分组至机架 ID ＝ 2 的网络节点 110b。

[0067] 在实施例中，交换模块 210a 可以在在 VFL 120 上传输之前改变前挂报头。例如，交换模块 210a 可以将具有本地意义的目的地 HDI（例如，gport 值或本地硬件设备标识符）转换成具有全局意义的 HDI 或者从前挂报头中移除外出 VFID。

[0068] 在实施例中，NIM 152 中的 MAC/HDI 转发表 250 被填充或更新，以响应流经虚拟机架系统 100 的层 2 分组。由于前挂报头包括源 MAC 地址和源 HDI 信息，NIMS 152，例如实施例中的特定交换模块 210，可以用该信息填充 MAC/HDI 转发表 250。通过以前挂的报头模式操作来在 VFL120 上交换具有源 MAC 地址和源 HDI 的层 2 分组，交换模块 210 能够在虚拟机架系统 100 中的网络模块 110 之间同步 MAC/HDI 转发表 250。尽管 MAC/HDI 转发表 250 和 VFL 路由表 254 被描述为位于交换模块 210 内，可选地或附加地，这些表可以被包括在排队模块 212n 或网络节点 110 的其它模块内。在另一个实施例中，CMM 150（主和次）还可以包括 MAC/HDI 转发表 250 和 VFL 路由表 254。

[0069] 图 8 示出了在虚拟机架系统 100 的网络节点 110 中可操作的虚拟机架管理器应用或模块 400 的实施例的示意性框图。在具有基于多插槽机架节点架构的网络节点 110 的实施例中，虚拟机架管理器模块 400 包括在网络节点 110 的中央管理模块 (CMM)150（被称为 VCM-CMM 402）和该网络节点的指定网络接口模块 (NIM)152 中的处理模块 266（被称为 VCM-NIM 404）之间分布的功能。在可堆叠节点架构中，指定的或主可堆叠网络单元 140 操作 VCM-NIM 404。指定的 NIM 152 或可堆叠单元 140 的使用避免仅在 CMM 150 上集中 VCM 模块 400 的功能。虚拟机架管理器模块 400 的功能的分布的例子见表 4 所示。

[0070]

VCM-CMM 402	VCM-NIM 404
• 至虚拟机架功能的单元和网络管理接口 • 虚拟机架操作和来自网络节点概览的状态的协调	• 控制协议状态机 • 与其他软件组件的服务接口，即，由 VCM 模块 400 使用以向其他软件组件提供服务或者从其他软件组件请求服务的接口 • 底层交换模块设备的规划（programming）：全局模块标识符 (MID)、环路防止（loop prevention）、虚拟机架进程间通信基础设施、VFL 成员端口规划等

[0071] 表 4

[0072] 在实施例中，VCM-CMM 402 包括虚拟机架管理器模块 400 与单元和 / 或网络管理器模块 406 之间的接口以及到注册到在网络节点 110 上可操作的 VCM 模块 400 的其他应用 408 的接口。虚拟机架管理器模块 400 通知已注册的应用 408 何时以虚拟机架模式工作。更一般地，虚拟机架管理器模块 400 提供较宽范围的通知，以向感兴趣的应用通知本地节点环境中和虚拟机架系统 100 的其他网络节点 110 中的虚拟机架系统的状态。一些状态信息由管理配置驱动，而其他状态信息由运行时的判决触发，所述判决在控制数据交换、协商和达成协议时，在虚拟机架系统中由网络节点单独或由多个网络节点 110 作出。虚拟机架管理器模块 400 还与 VLAN 管理应用模块 410、生成树协议 (STP) 应用模块 412、源学习应用模块 414、链路聚合应用模块 416 和端口管理器应用模块 418 相连接，用于从这些系统组件请求服务。例如，VCM 400 可以请求 VLAN 管理器来配置 VFL 成员端口作为控制 VLAN 的成员，以便允许虚拟机架系统 100 中的网络节点 110 之间的进程间通信信道的设置。

[0073] VCM-NIM 404 执行硬件模块的模块标识配置 (例如 MID)。VCM-NIM 404 还与在排队模块 212 中的队列管理 262 相连接以执行硬件设备 / 队列映射功能和机架间环路避免功能。VCM-NIM 404 还包括用于 VFL 120 的控制和管理的虚拟机架状态功能。虚拟组织链路控制管理和配置 VFL 120 并且与端口管理器应用模块 418 相连接以监视和 / 或控制 VFL120 的状态及其相应的成员端口。它还跟踪并更新 VFL 120 的状态。VCM-NIM 404 使用标准 LACP 协议或其他类似的协议跟踪每个 VFL 成员端口的状态以及物理层的链路状态。除了 LACP 协议之外，虚拟机架状态协议执行周期性的“保活”检查 (hello 协议)，以便检查在两个虚拟机架交换机上的指定的 NIM 上运行的组件的状态和 / 或可操作性。所有虚拟机架协议分组在系统中必须被分配高优先级以避免错误 / 过早的故障检测，因为这样的过早故障检测在系统中可能具有很严重的破坏性影响。通过在主指定 NIM 152 上运行虚拟机架状态协议，备份的指定 NIM 模块能够在发生故障时承担状态协议处理的控制。

[0074] VCM-CMM 402 和 VCM-NIM 404 向端口管理器应用模块 418 注册，以接收关于 VFL 120 的成员端口和链路的端口状态和状态链路事件。在另一个实施例中，虚拟机架管理器模块 400 可以包括端口管理器应用模块以监视 VFL 120 的端口和链路状态。虚拟机架管理器模块 400 跟踪 VFL 120 的操作状态和有关 VFL 状态的处理事件，即聚合创建 / 删除聚合 / 向上聚合 / 向下聚合。端口管理应用模块 418 提供链路状态通知到 VCM-CMM 402 和 VCM-NIM 404 两者。

[0075] 在实施例中，在虚拟机架系统 100 中实现传输控制协议以传输网络节点 110 的指定 NIM 152 或堆叠式网络单元 140 之间的控制协议分组。传输控制协议在具有不同节点架构的网络节点 110 中是可操作的。对于基于多插槽机架的节点架构，具有指定处理模块 266 的指定 NIM 152 操作传输控制协议，例如作为 VCM-NIM 404 的一部分。在可堆叠节点架构中，指定或主堆叠网络单元 140 操作传输控制协议。

[0076] 机架监督模块 420 提供至网络节点 110 的硬件的接口，并且控制对各种应用模块的监视和启动或重启、控制软件重新加载和软件升级 (例如服务中软件升级 ISSU)、为单元管理器模块 406 提供命令行界面 (CLI)、以及控制对网络节点 110 的系统的状态或镜像文件 (image file) 的访问。在虚拟机架模式期间，机架监督模块 420 控制启动序列、控制软件重新加载和 ISSU，并且提供了用于访问虚拟机架参数的接口。

[0077] 配置管理器模块422可操作用于将网络节点110的操作从虚拟机架模式转换为独立模式(standalone mode)或将网络节点110从独立模式转换为虚拟机架模式。配置管理器模块还可操作于配置虚拟机架管理器模块400和多机架管理器模块424。配置管理器模块422的操作和网络节点110的操作状态在下文中将更详细地进行描述。

[0078] 虚拟机架系统100中的网络节点110可以以多个操作模式操作,包括虚拟机架模式、独立模式和多机架(MC-LAG)模式。根据操作模式,各种参数和配置被修改。表5示出了取决于操作模式的机架ID向网络节点110的分配。

[0079]

操作模式	最少机架ID	最大机架ID
独立	0	0
多机架(MCLAG)	1	2
虚拟机架	1	N

[0080] 表5

[0081] 在独立模式中,网络节点110作为单个节点被操作,并使用其配置的本地MAC地址而不是全局虚拟机架MAC地址。如详细地描述于2011年1月20日提交的题为“SYSTEM AND METHOD FOR MULTI-CHASSIS LINK AGGREGATION”的美国专利申请No.13/010,168,在多机架模式中,两个网络节点被配置为虚拟节点,它们的MAC转发表和ARP表被同步,然而它们仍然充当单独的桥接器和路由器,它们中的每一个都使用它们自己的本地机架MAC地址。在如本文所述的虚拟机架模式中,N个网络节点被配置为虚拟机架系统100中虚拟机架节点。从1到N的全局唯一的机架ID被分配到虚拟机架系统100中的多个网络节点中的每一个。

[0082] 当网络节点110以独立模式操作时,端口标识符和配置遵循如下格式:0/<插槽>/<端口>,其中机架ID等于“零”,插槽标识多插槽结构或可堆叠网络单元140中的每一个网络接口模块(NIM)152,并且端口是端口接口标识符。当网络节点110以多机架模式操作时,端口配置遵循如下格式:<机架>/<插槽>/<端口>,其中机架ID等于1或2并且表示操作/当前的/运行的机架ID。当网络节点110以虚拟机架模式操作时,端口配置遵循如下格式:<机架>/<插槽>/<端口>,其中机架ID是范围1,2……N内的数字并且表示操作/当前的/运行的机架ID。

[0083] 在虚拟机架系统100中,当检测到导致与主网络节点110的通信丢失的故障时,在虚拟机架系统100中发生分裂或断裂。当虚拟机架系统中的网络节点110被分裂成两个或更多子集时,虚拟机架系统100的拓扑以及因此它所提供的服务会受到严重的影响。这种情况被称为虚拟机架分裂或断裂。当虚拟机架拓扑被分裂时,虚拟机架系统100面临由于多个节点重置事件所产生的多个问题:重复的MAC地址、重复的可配置资源(例如IP接口)、连接性丢失、管理访问丢失、以及不稳定性。

[0084] 虚拟机架分裂可以典型地由节点110中的一个或多个的故障触发,例如电源故障、硬件/软件故障、服务中软件升级等。虚拟机架分裂也可以由耦合网络节点110的VFL 120中一个或多个变得不可操作而触发,例如,VFL 120被在物理上折断、移除、管理上的垮

塌 (brought down)、或由于与托管这样的链路的模块或链路自身相关的硬件 / 软件故障造成的垮塌。在实施例中，当虚拟机架网络 100 的两个子集之间发生虚拟机架分裂时，网络节点 110 的第一子集不再能与第二子集中的主设备网络节点通信。第一子集中剩余的活动网络节点选择新的主网络节点。在实施例中，剩余活动网络节点保留发生故障的主网络节点的主 MAC 地址。在另一个实施例中，剩余活动网络节点采用新选择的主网络节点的本地 MAC 地址作为虚拟机架系统 100 的新虚拟机架 MAC 地址。

[0085] 图 9 示出了在虚拟机架系统 100 中主地址保留的实施例的示意性框图。由于功能损坏或计划断电维护或不能操作的 VFL 120 链路或其他故障，主网络节点 110a 不能与虚拟机架系统 100 中的剩余节点 110b、110c 通信。剩余网络节点 110b 和 110c 选择新的主网络节点，在该示例中为网络节点 110b。在该实施例中，剩余网络节点 110b 和 110c 保留之前的主网络节点 110a 的 MAC 地址作为虚拟机架系统 100 的虚拟机架 MAC 地址。之前的主网络节点 110a 被从剩余活动网络节点 110b 和 110c 的拓扑数据库 144 和 MAC 矩阵中移除。该实施例因为由剩余活动网络节点 110 保留了之前的主 MAC 地址作为虚拟机架 MAC 地址而被称为主 MAC 地址保留。

[0086] 图 10 示出虚拟机架系统 100 中主地址释放的实施例的示意性框图。由于功能损坏或计划断电维护或不能操作的 VFL 120 链路或其他故障，主网络节点 110a 不能与虚拟机架系统 100 中的剩余节点通信。剩余网络节点 110b 和 110c 选择新的主网络节点，即该示例中的网络节点 110b。在该实施例中，剩余网络节点 110b 和 110c 释放作为虚拟机架 MAC 地址的之前的主网络节点 110a MAC 地址。剩余活动网络节点 110b 和 110c 采用新选择的主网络节点 110b 的本地 MAC 地址作为虚拟机架系统 100 的虚拟机架 MAC 地址。之前的主网络节点 110a 从剩余活动网络节点 110b 和 110c 的拓扑数据库和 MAC 矩阵中被移除。由于作为虚拟机架 MAC 地址的不活动的之前的主 MAC 地址的释放，该实施例被称为主 MAC 释放。

[0087] 剩余网络节点 110 基于一个或多个因素确定保留或释放不活动的主网络单元的 MAC 地址。例如，一个因素是，是否 MAC 保留功能在管理上是已启用的。另一个因素是主网络节点状态的改变是否在系统中引起虚拟机架分裂，例如主网络节点和 / 或一个或多个其他节点仍然使用故障的之前的主网络节点的 MAC 地址进行操作。发现或监控协议或其他类型的控制协议被用于确定主网络节点的故障前后的拓扑以确定虚拟机架系统中是否发生了分裂。当虚拟机架系统中的分裂已经发生时，例如新选择的主网络节点确定之前的主网络节点和 / 或一个或多个其他节点仍然在工作，其释放作为虚拟机架 MAC 地址的之前的主 MAC 地址。新选择的主网络节点还可以将用户端口也转变到阻塞状态以防止作为虚拟机架 MAC 地址的两个 MAC 地址的重复操作。

[0088] 图 11 示出了虚拟机架系统 100 中的主网络节点故障的实施例的示意性框图。在这个例子中，主网络节点 110a 发生故障并且不能操作。新选择的主网络节点 110b 通过执行一个或多个协议 (Hello 协议、ping 等) 尝试确定之前的主网络节点 110a 的状态，或者可以从单元管理器模块 406 请求网络节点 110a 的状态更新。新选择的主网络节点 110b 尝试在 VFL 链路 120a 的故障或之前的主网络节点 110a 故障之间进行区分。当新选择的主网络节点 110b 确定之前的主网络节点 110a 已经发生故障时，例如它不再能操作，新选择的主网络节点 110b 保留之前的主网络节点 110a 的 MAC 地址作为虚拟机架 MAC 地址，如图 9 所

示。当之前的主网络节点 110a 被从活动拓扑数据库 144 中移除时，新选择的主网络节点 110b 保留之前的主节点的 MAC 地址，机架监督模块 420 启动 MAC 保留计时器。该 MAC 保留计时器是可配置的，并为之前的主网络节点 110a 设置预定时间段以重置的和变活动 (become active)。一旦耗尽预定时间段，如果之前的主网络节点 110a 仍然不能操作，则由新选择的主网络节点 110b 生成告警消息。虚拟机架系统管理器可以确定以发布用户命令来释放保留的 MAC 地址并采用新选择的主网络节点 110b 的本地 MAC 地址作为虚拟机架系统 100 的虚拟机架 MAC 地址。

[0089] 图 12 示出了虚拟机架系统 100 中 VFL 故障的实施例的示意性框图。在本例中，耦合到主网络节点 110a 的 VFL 120a 发生故障而之前的主网络节点 110a 保持可操作。新选择的主网络节点 110b 尝试通过执行一个或多个协议 (Hello 协议、ping 等）确定之前的主网络节点 110a 的状态，或者可以从单元管理器模块 406 请求网络节点 110a 的状态更新。新选择的主网络节点 110b 尝试在 VFL 链路 120a 的故障或之前的主网络节点 110a 的故障之间做出区分。当新选择的主网络节点 110b 确定 VFL120a 已经发生故障，但之前的主网络节点 110a 可操作时，新选择的主网络节点 110b 释放之前的主网络节点 110a 的 MAC 地址，如图 10 所示。剩余活动网络节点 110b 和 110c 采用新选择的主网络节点 110b 的本地 MAC 地址作为虚拟机架系统 100 的虚拟机架 MAC 地址。另外，新选择的主网络节点 110b 将用户端口到转换到阻塞状态以防止作为虚拟机架 MAC 地址的两个 MAC 地址的重复操作。之前的主网络节点 110a 的 MAC 地址的释放也会影响其他的层 2 和层 3 服务。例如，响应于 MAC 地址改变，生成树协议和 LACP 可能需要重新配置和 / 或重新启动而层 3 分组可能需要被发送到相邻的节点。

[0090] 图 13 示出了在虚拟机架系统 100 中从主网络节点的故障恢复的方法 600 的实施例的逻辑流程图。在步骤 602，由于功能损坏或计划断电维护或不能操作的 VFL 120 链路或其他故障，在虚拟机架系统 100 中检测到与主网络节点的通信丢失。虚拟机架系统中的剩余网络节点在步骤 604 中选择新的主网络节点。在步骤 606，新选择的主网络节点确定是否启用 MAC 保留功能。如果启用，在步骤 608，新选择的主网络节点确定故障是否引起虚拟 - 机架系统的分裂，例如主网络节点和 / 或一个或多个其他节点是不能操作的还是仍然使用之前的主 MAC 地址工作。当虚拟机架分裂未发生时，在步骤 610 剩余网络节点保留之前的主网络节点 110a 的 MAC 地址作为虚拟机架系统 100 的虚拟机架 MAC 地址。在步骤 612，MAC 保留计时器开始对预定时间段进行计时。在步骤 614 中，预定时间段期满时，如果之前的主网络节点 110a 仍然不能操作，则由新选择的主网络节点生成告警消息。

[0091] 当新选择的主网络节点确定虚拟机架分裂已经发生，例如步骤 608 中之前的主网络节点仍然在工作，或者在步骤 606 中 MAC 保留功能被禁用，那么在步骤 616 新选择的主网络节点释放作为虚拟机架 MAC 地址的之前的主网络节点的 MAC 地址。在步骤 618 中，剩余活动网络节点采用新选择的主网络节点的本地 MAC 地址作为虚拟机架系统 100 的虚拟机架 MAC 地址。

[0092] 当虚拟机架分裂发生时，虚拟机架系统的拓扑以及因此它所提供给终端用户的服务会受到严重的影响。当虚拟机架的拓扑被分裂时，虚拟机架系统 100 面临由于多个交换机重置事件所产生的多个问题：重复的 MAC 地址、重复的可配置资源（例如 IP 接口）、连接性丢失、管理访问的丢失、以及不稳定性。虚拟机架分裂可以典型地由节点 110 中的一个或

多个的故障触发，例如电源故障、硬件 / 软件故障、服务中软件升级等。虚拟机架分裂也可以由 VFL120 中的一个或多个变得不可操作而触发，例如，VFL 120 被在物理上折断、移除、管理上垮塌、或由于与托管这样的链路的模块或者链路自身相关的硬件 / 软件故障造成的垮塌。虚拟机架分裂的一个原因包括管理动作，例如发布重置或关闭虚拟机架系统的一部分导致拓扑分裂的管理命令。可能触发虚拟机架分裂的其他管理动作包括参数的不一致配置，例如 a) 在虚拟机架拓扑内不同的网络节点上配置不同的控制 VLAN ；b) 在虚拟机架拓扑内不同的网络节点上配置不同的问候 (hello) 间隔 ；c) 在虚拟机架的拓扑内的不同网络节点上配置不同的机架群。由于虚拟机架系统 100 支持多个网络拓扑并且网络节点可以在地理上分开，难以预见管理动作的影响。

[0093] 在实施例中，在网络节点 110 中的方法和装置提供了告警来帮助防止虚拟机架系统 100 中的管理动作直接或间接地导致虚拟机架分裂事件。例如，可以直接或间接地导致虚拟机架分裂的管理动作特别地包括 ：重置网络节点、重置托管一个或多个 VFL 120 的网络接口模块 152、停止 (bring down)VFL 120、设置网络节点 110 进入关闭 / 非服务状态以及在不同的网络节点 110 上配置不一致参数。另外，可能直接或间接地导致虚拟机架分裂的管理动作所产生的间接事件特别地包括 ：响应于 ISSU(服务中软件升级) 操作交换机被重置和在系统执行 ISSU 操作时托管一个或多个 VFL 120 的网络接口模块 NIM 152 被停止 (bring down)。在实施例中，虚拟机架系统 100 的当前拓扑被分析以确定一个或多个的管理动作的可能的影响。根据该分析，当可能直接或间接地导致虚拟机架分裂的管理动作被请求时生成告警。

[0094] 图 14 示出了虚拟机架系统 100 中生成一个或多个重置列表的方法 700 的实施例的逻辑流程图。在步骤 702 中，网络节点使用一个或多个控制协议以发现虚拟机架系统 100 中的其他网络节点 110，并交换拓扑和配置信息。网络节点 110 使用拓扑信息来构建本文描述的虚拟机架系统 100 的拓扑数据库 144。拓扑数据库包括，例如以下类型的拓扑信息 ：其他网络节点 110 的标识信息 (例如，本地 MAC 地址、机架标识符)、托管 VFL 120(或其他活动的交换机间链路) 的网络接口模块 NIM 152 的标识信息、用于 VFL 120 及其在托管网络接口模块 NIM 152 上的相关成员端口的标识信息。拓扑数据库 144 在网络节点 110 的 CMM 150 中被维护，并且可被复制到可操作地耦合至虚拟机架系统 100 中的网络节点 110 的单元管理器模块 406 之中。该拓扑数据库 144 的描述包括了示例性信息，然而本文没有描述的其他信息和数据也可以被包括在拓扑数据库中。此外，拓扑数据库 144 可被存储在单独的数据库或表中或者与网络节点 110 中的其他表或数据库相组合。

[0095] 在步骤 704 中，拓扑信息被分析以确定一个或多个重置列表或结构，所述重置列表或结构表现了一个或多个管理动作对虚拟机架系统 100 的影响。例如，生成的网络节点重置列表包括如果被单独重置将触发虚拟机架分裂的网络节点的列表。其还可以包括需要在相同的时间被重置以防止系统中的虚拟机架分裂的网络节点的列表。生成的网络接口模块重置列表包括托管活动的 VFL 120 的网络接口模块 152 的列表以及指示重置网络接口模块 152 是否将使得虚拟机架分裂的状态。还可以生成本文未具体描述的表示一个或多个管理动作对虚拟机架系统 100 的影响的其他重置列表和结构。

[0096] 在实施例中，响应于一个或多管理动作来访问重置列表以提供管理动作的可能影响的信息。例如，在管理动作的处理过程中针对受管理动作影响的设备 (网络节点、NIM 或

端口接口）分析重置列表，所述管理动作例如重新装载或重新设定网络节点 110、重新装载或重新设定网络接口模块 NIM 152、禁用网络接口模块 152、关闭网络节点、执行 ISSU、禁用端口接口、禁用 VFL 120 等的管理命令。

[0097] 当重置列表指示管理动作不会引起虚拟机架分裂时，管理动作的处理继续进行。然而，当重置列表指示管理动作可能引起虚拟机架分裂时，则显示告警。告警包括例如响应于该管理动作，虚拟机架分裂可能发生的指示。该告警还可以包括一个或多个建议以避免虚拟机架分裂，例如，用重置列表中指定的一个或多个其他设备（例如网络节点、NIM 或端口接口）来重置期望的设备 (intended device)（例如网络节点、NIM 或端口接口）。在另一个实施例中，阻止处理管理动作。包括不执行管理动作的通知的告警或另一消息被发布。

[0098] 图 15 示出了虚拟机架系统 100 中重置列表的生成的实施例的示意性框图。在该实施例中，网络节点重置列表被生成，该网络节点重置列表提供了网络节点的机架标识符的列表以及网络节点的重置是否会使得虚拟机架系统中的虚拟机架分裂的相应指示符。网络节点的重置包括关机、重启、重置、断电、非服务状态或以其他方式响应于管理动作而变得不能操作。例如，管理动作可包括这样的管理命令：重置网络节点、重启网络节点、断开网络节点电源、将网络节点 110 的状态设定为关闭或非服务状态或其他导致网络节点变得不能操作的管理动作。拓扑数据库 144 中的拓扑信息包括虚拟机架系统 100 中的网络节点 110 之间的可用路径。在图 15 的例子中，网络节点 110a、110b 和 110c 具有线性拓扑，拓扑数据库 144 包括表 6 中用于网络节点 110a、110b 和 110c 之间的可用路径的以下示例信息。

[0099]

源网络节点至 目的地网络节点	可用路径
机架 ID =1 至机架 ID=2	机架 ID=1, VFL ID=0 → 机架 ID=2, VFL=1
机架 ID =1 至 机架 ID=3	机架 ID=1, VFL ID=0 → 机架 ID=2, VFL ID=1→ 机架 ID=2, VFL ID=2 → 机架 ID=3, VFL ID=3
机架 ID =2 至 机架 ID=1	机架 ID=2, VFL ID=1 → 机架 ID=1, VFL ID=0
机架 ID =2 至 机架 ID=3	机架 ID=2, VFL ID=2 → 机架 ID=3, VFL ID=2
机架 ID =3 至机架 ID=1	机架 ID=3, VFL ID=2 → 机架 ID=2, VFL ID=2 → 机架 ID=2, VFL ID=1 → 机架 ID=1, VFL ID=0
机架 ID =3 至 机架 ID=2	机架 ID=3, VFL ID=2 → 机架 ID=2, VFL ID=2

[0100] 表 6

[0101] 对于图 15 的虚拟机架系统 100 中的拓扑信息的这一示例，网络节点 110a（具有机架 ID = 1）的重置不会引起系统中的虚拟机架分裂。类似地，网络节点 110c（具有机架 ID = 3）的重置不会引起系统中的虚拟机架分裂。然而，网络节点 110b（具有机架 ID = 2）的重置将引起系统中的虚拟机架分裂，因为节点的第一子集（网络节点 110a）将被分裂或分离或不能与系统中节点的第二子集（网络节点 110c）通信。在该示例实施例中，为了防止虚拟机架分裂，网络节点重置列表指示网络节点 110b 和 110c 的组合应该同时被重置以防止虚拟机架分裂。下面的表 7 示出了用于虚拟机架系统 100 的该实施例的示例性网络节点

重置列表。

[0102]

重置网络节点	虚拟机架分裂的指示符	重置建议
机架 ID ＝ 1	假	N/A
机架 ID ＝ 2	真	机架 ID ＝ 2,3
机架 ID ＝ 3	假	N/A

[0103] 表 7

[0104] 该重置列表被存储于虚拟机架系统 100 中的一个或多个网络节点 110 的 CMM150 中的重置列表表格 710 之中和 / 或被存储于可操作地耦合到虚拟机架系统 100 的单元管理器模块 406 之中。在实施例中，重置列表 710 由单元管理器模块 406 生成或存储。当管理动作输入到单元管理器模块时，单元管理器模块 406 执行重置列表 710 的分析，并确定是否发布对管理动作的告警。单元管理器模块 406 包括到一个或多个网络节点 110 的本地耦合的设备或远程设备。在另一个实施例中，虚拟机架系统 100 的一个或多个网络节点 110 的 CMM 150 生成或存储重置列表 710。当网络节点 110 的 CMM 150 接收到管理动作时，CMM 150 访问重置列表 710，并确定是否发布对管理动作的告警。

[0105] 图 16 示出了用于生成虚拟机架系统 100 中的网络节点重置列表的方法 750 的实施例的逻辑流程图。针对虚拟机架系统 100 中的多个网络节点 110 执行分析。在步骤 752 中，多个网络节点 110 中的网络节点 110 中的一个（例如，源网络节点）被选择用于分析。在步骤 754 中，拓扑数据库 144 中的拓扑信息被访问以根据源网络节点确定第一目的地节点。在步骤 756 中，根据拓扑数据库中的拓扑信息 144 确定源节点和目的地节点之间的路径数量。在步骤 758 中，当路径的数量大于 1 时，则在步骤 762 中，对于重置 ID ＝目的地网络节点来说，虚拟机架分裂的告警状态为假。由于存在多条路径，重置目的地节点不会将源网络节点从虚拟机架系统 100 中分裂或者隔离。当在步骤 758 中路径的数量等于 1 并且在步骤 760 中目的地节点是沿着路径的最后一跳时，则在步骤 762 中，对于重置 ID ＝目的地网络节点来说，虚拟机架分裂的告警状态为假。当在步骤 758 中路径数量等于 1 并且在步骤 760 中目的地节点不是沿路径的最后一跳时，则在步骤 764 中，对于重置 ID ＝目的地网络节点来说，虚拟机架分裂告警指示器被设置为真。在步骤 766 中，分析确定了源网络节点和目的地网络节点之间的路径上的其他网络节点。目的地网络节点和所述其他网络节点被列出在网络节点重置列表上作为同时进行重置的建议以避免虚拟机架分裂。在步骤 768 中，确定是否需要分析其他目的地节点。如果是，则过程进入到步骤 754。如果没有另外的目的地节点需要分析，则在步骤 770 中，该源网络节点的网络节点重置列表被存储。

[0106] 图 17 示出了虚拟机架系统 100 中生成重置列表的另一个实施例的示意性框图。当托管 VFL120 的一个或多个 NIM 152 被重置时可能引起虚拟机架分裂。NIM 152 的重置包括关机、重启、重置、断电、非服务状态、或以其他方式响应于管理动作而变得不可操作。例如，在网络节点 110b 中，NIM 152c 托管可操作地耦合至网络节点 110a 和 110b 的 VFL 120a。当 NIM 152 响应管理动作被重置时，VFL120a 将不能操作从而引起系统中的虚拟机架分裂。节点的第一子集（网络节点 110a）将与系统中的节点的第二子集（网络节点 110b 和 110c）

分裂，例如节点的第一子集（网络节点 110a）将不能与系统中的节点的第二子集（网络节点 110b 和 110c）通信。在实施例中，为了防止虚拟机架分裂，网络接口模块重置列表被生成，其包括为了避免虚拟机架分裂不应当响应于管理动作而变得不可操作的 NIM 的列表。下面的表 8 示出了用于图 16 中虚拟机架系统 100 的实施例的示例性网络接口模块重置列表。网络接口模块是由它的网络节点 110 的机架 ID 和插槽号码来标识的。

[0107]

重置 NIM（机架 ID，NIM ID）	虚拟机架分裂的指示符
机架 ID ＝ 1，NIM ID ＝插槽 1	假
机架 ID ＝ 1，NIM ID ＝插槽 2	假
机架 ID ＝ 1，NIM ID ＝插槽 3	假
机架 ID ＝ 1，NIM ID ＝插槽 2，3	真
机架 ID ＝ 2，NIM ID ＝插槽 1	真
机架 ID ＝ 2，NIM ID ＝插槽 2	假
机架 ID ＝ 2，NIM ID ＝插槽 3	真
机架 ID ＝ 3，NIM ID ＝插槽 1	假
机架 ID ＝ 3，NIM ID ＝插槽 2	真
机架 ID ＝ 3，NIM ID ＝插槽 3	假

[0108] 表 8

[0109] 在实施例中，告警可包括在重置会引起虚拟机架分裂的 NIM 152 之前，将 VFL 120 重新配置到另一个 NIM 152 的建议。

[0110] 在另一实施例中，重置 VFL 120 的活动成员端口的端口接口 240 可能会导致虚拟机架分裂。例如，VFL 成员端口的重置包括关机、重启、重置、断电、非服务状态、阻塞模式或以其他方式响应于管理动作使得所述端口接口不能操作。例如，响应于以下管理动作可能导致虚拟机架分裂：将端口接口置为阻塞模式、重置端口接口、重启端口接口或将所述端口接口置为关机或非服务状态或以其它方式使得端口不能操作。在实施例中，为了防止虚拟机架分裂，VFL 成员端口重置列表被生成，其包括托管 VFL 链路的端口接口 240 的列表以及当相应的端口接口 240 被重置时是否应当发布告警以避免虚拟机架分裂的指示符。下面的表 9 示出了用于图 16 中的虚拟机架系统 100 的实施例的示例性 VFL 成员端口重置列表。端口接口由其网络节点 110 的机架 ID、其 NIM 152 的插槽号以及端口 ID 来标识。

[0111]

重置端口（机架 ID，插槽，端口 ID）	虚拟机架分裂的指示符

机架 ID = 1,NIM ID = Slot 2, 端口 ID = 1	假
机架 ID = 1,NIM ID = Slot 3, 端口 ID = 1	假
机架 ID = 2,NIM ID = Slot 1, 端口 ID = 2	真
机架 ID = 2,NIM ID = Slot 3, 端口 ID = 1	假
机架 ID = 2,NIM ID = Slot 3, 端口 ID = 2	假
机架 ID = 2,NIM ID = Slot 3, 端口 ID = 1,2	真
机架 ID = 3,NIM ID = Slot 2, 端口 ID = 1	真

[0112] 表 9

[0113] 在实施例中，告警可以包括在重置将引起虚拟机架分裂的端口接口之前，将 VFL 120 的成员端口接口重新配置到另一端口接口 240 的建议。

[0114] 图 18 示出了在虚拟机架系统 100 中生成网络接口模块重置列表和 / 或 VFL 成员端口重置列表的方法 800 的实施例的逻辑流程图。针对虚拟机架系统 100 中的多个网络节点 110 的网络节点 110 的 NIM 152（由插槽 ID 标识）执行分析。在步骤 802 中，多个网络节点 110 的网络节点 110 之一（例如，源网络节点）的第一 NIM 152 被选择用于分析。在步骤 804 中，拓扑数据库中的拓扑信息 144 被访问以确定 VFL 成员端口是否被包括在所选择的 NIM 上。在步骤 804 中，当在 NIM 上没有 VFL 成员端口时，在步骤 808 中，对于重置 ID = 机架 ID、NIM 的插槽 ID，虚拟机架分裂的告警状态为假。在步骤 804 中，当 NIM 上有 VFL 成员端口时，则拓扑数据库 144 中的拓扑信息被访问以确定 NIM 是否托管了 VFL 的所有 VFL 成员端口（例如，源网络节点上的另一个 NIM 是不是也托管了 VFL 的 VFL 成员端口）。如果不是，在步骤 808 中，对于重置 ID = 机架 ID、NIM 的插槽 ID，虚拟机架分裂的告警状态为假。在步骤 806 中，如果 NIM 托管了 VFL 的所有成员端口，那么在步骤 810 中，拓扑数据库 144 中的拓扑信息被访问以确定源网络节点和目的地节点之间的路径（例如，VFL）的数量。在步骤 812 中，当其他路径或 VFL 连接源与目的地节点时，那么在步骤 808 中，对于重置 ID = 机架 ID、NIM 的插槽 ID，虚拟机架分裂的告警状态为假。在步骤 812 中，当 NIM 托管作为源网络节点和目的地网络节点之间的唯一路径或连接的 VFL 的成员端口时，在步骤 814 中，拓扑数据库 144 中的拓扑信息被访问以确定 VFL 是否为到目的地网络节点的路径中的下一跳。如果不是，则在步骤 808 中，对于重置 ID = 机架 ID、NIM 的插槽 ID，虚拟机架分裂的告警状态为假。在步骤 814 中，如果 NIM 托管到目的网络节点的路径中的下一跳的 VFL 的唯一成员端口，则在步骤 816 中，对于重置 ID = 机架 ID、NIM 的插槽 ID，虚拟机架分裂的告警状态为真。

[0115] 在步骤 818 中，继续分析以生成 VFL 成员端口重置列表。对于 NIM 的每个 VFL 成员端口，在步骤 818 执行分析。在步骤 818 中，拓扑数据库 144 中的拓扑信息被访问以确定 NIM 上的 VFL 的成员端口是否为 VFL 的唯一成员端口。如果不是，则在步骤 820 中，对于重置 ID = 机架 ID、插槽 ID、VFL 成员端口的端口 ID，虚拟机架分裂的告警指示符为假。在步骤 818 中，当 NIM 上的 VFL 的成员端口是 VFL 的唯一成员端口时，则在步骤 822 中，对于重

置 ID ＝机架 ID、插槽 ID、VFL 成员端口的端口 ID，虚拟机架分裂的告警指示符为真。图 18 中的分析被执行以生成虚拟机架系统 100 中的网络接口模块重置列表和 / 或 VFL 成员端口重置列表。

[0116] 图 16 和图 18 示出了生成用于本文所述的示例性拓扑的重置列表的示例性过程。类似的，可以执行附加的或替换的过程或分析以确定用于这些拓扑或其他拓扑的重置列表。

[0117] 图 19 示出了在虚拟机架系统 100 中用于处理管理动作以有助于防止虚拟机架分裂的方法 850 的实施例的逻辑流程图。在实施例中，在步骤 852 中接收管理动作。管理动作包括管理命令和网络节点 110、NIM 152 或端口接口 240 的标识符。响应于管理动作，在步骤 854 中访问重置列表中的一个或多个。例如，在管理动作的处理过程中，针对受管理的动作的影响的所标识的设备（网络节点，NIM 或端口接口）来访问重置列表中的一个或多个，所述管理动作例如重新装载或重置网络节点 110、重新装载或重置 NIM 152、禁用 NIM 152、关闭网络节点 110、执行 ISSU、禁用端口接口 240、禁用 VFL120 等的管理命令。例如，网络节点重置列表、NIM 重置列表和 / 或 VFL 成员端口接口重置列表被访问。其他或附加的列表也可以被访问以帮助确定是否发布有关虚拟机架分裂的告警。在步骤 856 中，确定是否响应于管理动作而发布虚拟机架分裂的告警。

[0118] 当确定不发布告警时，例如重置列表中的一个或多个指示管理动作不会引起虚拟机架分裂，则管理动作的处理进行到步骤 858。然而，当确定要发布告警时，例如重置列表指示管理动作可能引起虚拟机架分裂，则发布告警并传输到用户设备以用于显示。所述告警包括例如响应于管理动作可能发生虚拟机架分裂的指示。告警还可以包括一个或多个建议以避免虚拟机架分裂，例如，用重置列表中指定的一个或多个其他设备重置期望的设备（诸如重置一个或多个其他网络节点）。该告警还可以包括在执行管理动作之前将 VFL 成员端口接口 240 重新配置到一个或多个其他 NIM 152 的建议。

[0119] 在实施例中，告警由单元管理器模块在图形用户界面中显示，所述图形用户界面例如命令行界面 (CLI)、Webview 或可替换的管理应用。在实施例中，提供了一个或多个选项，例如中止管理动作、继续进行管理的动作而不管告警，以及使用告警建议继续进行管理动作，例如重新启动期望的设备加上建议的重置列表中的其他设备。在另一个实施例中，禁止处理管理动作。包括不执行管理动作的通知的告警或另一消息被发布。

[0120] 重置列表有助于在虚拟机架系统中响应于管理动作对虚拟机架分裂进行告警。这种预防减少了由于虚拟机架分裂而出现的有害事件，带来了更加稳定和稳健的系统，所述有害事件例如重复的 MAC 地址、重复的可配置资源（例如，IP 接口）、连接性丢失、管理访问的丢失、由于多次交换机重置事件引起的不稳定等。

[0121] 图 20 示出了在虚拟机架系统 100 中用于处理一个或多个参数的配置的管理动作以帮助防止虚拟机架分裂的方法 900 的实施例的逻辑流程图。例如，可能触发虚拟机架分裂的管理动作包括在网络节点 110 上不一致的参数配置，例如 a) 在虚拟机架拓扑内不同的网络节点 110 上配置不同的控制 VLAN ；b) 在虚拟机架拓扑内不同的网络节点 110 上配置健康监视消息（例如问候或保活消息）的不同间隔 ；c) 在虚拟机架拓扑内的不同网络节点 110 上配置不同的机架群（虚拟机架系统标识符）。例如，当虚拟机架系统中的网络节点 110 被配置了不同的虚拟机架系统标识符时，那么网络节点 110 可以发起虚拟机架分裂变成不

同的虚拟机架系统。另外，为控制分组配置不同的控制 VLAN 可能引起控制分组被丢弃和未处理并导致虚拟机架分裂。而且，配置用于健康监视的不同间候间隔可能引起对节点或节点模块的故障的不正确结论。本文未公开的其他参数的错误配置也可能引起会导致虚拟机架分裂的故障或模块功能损坏。

[0122] 在步骤 902 中，接收管理动作以配置一个或多个网络节点 110 上的一个或多个参数。在步骤 904 中，所述配置被分析以确定所述一个或多个参数在虚拟机架系统中的网络节点 110 之间是否冲突或者是否将以其他的方式引起故障或导致虚拟机架分裂。在步骤 906 中，如果所述配置可能导致故障，则在步骤 910 中对该管理动作发布告警。所述告警包括例如，网络节点之间的参数冲突的指示和 / 或所述配置可能导致故障或虚拟机架分裂的指示。告警还可以包括一个或多个建议以避免故障或虚拟机架分裂，例如，不同的配置参数。在另一个实施例中，阻止处理管理动作。包括不执行管理动作的通知的告警或另一消息被发布。

[0123] 当在步骤 906 中确定参数的配置不会导致故障或虚拟机架分裂时，管理动作的处理进行到步骤 908。

[0124] 本文所使用的术语“可操作地耦合到”、“耦合到”、和 / 或“耦合”包括物品之间的直接耦合和 / 或物品之间经由中间物品（例如，物品包括但不限于组件、单元、电路、和 / 或模块）的间接耦合，其中对于间接耦合来说，中间物品不修改信号的信息但可能调整其电流水平、电压水平和 / 或功率水平。本文中还使用的，推断的耦合 (inferred coupling)（也就是说一个单元通过推断耦合到另一个单元）包括两个物品之间以与“耦合到”相同方式的直接和间接耦合。

[0125] 本文进一步使用的术语“可操作用于”或“可操作地耦合到”表示包括一个或多个电力连接、（多个）输入、（多个）输出等的物品，当被激活时物品执行一个或多个其对应的功能，并且还可以包括到一个或多个其他物品的推断耦合。本文还可进一步使用的术语“与…相关”包括单独物品的直接和 / 或间接的耦合和 / 或一个物品被嵌入在另一物品内、或一个物品被配置为由或者被另一物品使用。在此可以使用的术语“有利地比较”，表明两个或更多物品、信号等之间的比较提供了期望的关系。例如，当期望的关系是信号 1 具有比信号 2 更大的幅值时，当信号 1 的幅度大于信号 2 的幅度时或者当信号 2 的幅度小于信号 1 的幅度时有利的比较可以被实现。

[0126] 还可以在本文中使用的术语“处理模块”、“处理电路”和 / 或“处理单元”可以是单个处理设备或多个处理设备。这样的处理设备可以是微处理器、微控制器、数字信号处理器、微型计算机、中央处理单元、现场可编程门阵列、可编程逻辑器件、状态机、逻辑电路、模拟电路、数字电路、和 / 或任何基于电路的硬编码和 / 或操作指令操作信号（模拟和 / 或数字）的设备。处理模块、模块、处理电路和 / 或处理单元可以是或者可以进一步包括存储器和 / 或集成的存储器单元，其可以是单个存储设备、多个存储器设备、和 / 或另一个处理模块、模块、处理电路、和 / 或处理单元的嵌入电路。这样的存储设备可以是只读存储器、随机存取存储器、易失性存储器、非易失性存储器、静态存储器、动态存储器、闪存、高速缓冲存储器和 / 或存储数字信息的任何设备。注意，如果处理模块、模块、处理电路、和 / 或处理单元包括一个以上的处理设备，所述处理设备可以集中设置（例如，经由有线和 / 或无线总线结构直接地耦合在一起）或者可以是分布式地设置（例如，经由局域网和 / 或广域网通过

间接耦合的云计算）。进一步注意到，如果处理模块、模块、处理电路、和／或处理单元经由状态机、模拟电路、数字电路、和／或逻辑电路实现其功能中的一个或多个，存储相应操作指令的存储器和／或存储单元可以嵌入在包括状态机、模拟电路、数字电路、和／或逻辑电路的电路的内部或外部。更要进一步注意，存储器单元可以存储，并且处理模块、模块、处理电路、和／或处理单元执行，对应于一个或多个附图中所示的至少一些步骤和／或功能的硬编码和／或操作指令。这样的存储器设备或存储器单元可以被包含在制造产品内。

[0127] 上面已在示出了特定功能的性能及其关系的方法步骤的帮助下描述了本发明。这些功能构造块和方法步骤的边界和顺序在此处为了描述的方便做了任意的定义。只要特定的功能和关系被适当地执行，可以定义替代的边界和顺序。任何这样的替代边界或序列都落入要求保护的发明的范围和精神内。此外，为了描述的方便对这些功能构造块的边界已经做了任意的定义。只要某些重要的功能被适当地执行，可以定义替代的边界。类似地，为了说明某些重要的功能，流程图块也在此做了任意的定义。为了使用的扩展，流程图块的边界和顺序可以以其它方式定义并且仍执行某些重要的功能。这样的功能构造块和流程图块以及序列的替代定义因此仍落在要求保护的发明的范围和精神内。本领域技术人员还将认识到，功能示意框图，和这里的其他示意性块、模块和组件，可按所示意的实现或组合或分离为离散组件、专用集成电路、执行适当软件的处理器等或其任意组合。

[0128] 在本文中，本发明至少部分地就一个或多个实施例来进行了描述。本文描述实施例以说明本发明、其中一个方面、其中一种特征、其中一种概念，和／或其中一个例子。体现了本发明的装置、制造产品、机器和／或过程的物理实施例可以包括本文讨论的参照一个或多个实施例描述的方面、特征、概念、例子等中的一个或多个。此外，在图和图之间，实施例可以包含相同或具有类似名称的功能、步骤、模块等，它们可以使用相同的或不同的附图标记，并且这样的功能、步骤、模块等可以是相同或类似的功能、步骤、模块等，或者是不同的功能、步骤、模块等。

[0129] 除非特别声明，本文给出的附图中的去往、来自和／或单元之间的信号可以是模拟的或数字的、连续时间方式或离散时间方式、以及单端的或差分的。例如，如果信号路径被显示为单端路径，它也代表差分信号路径。类似地，如果信号路径被显示为差分路径，它也代表单端信号路径。虽然本文描述了一个或多个特定的架构，使用一个或多个未明确示出的数据总线、单元之间的直接连接性，和／或在其他单元之间间接耦合的其他架构同样可实现。

[0130] 本发明的各种实施例的描述中使用了术语“模块”。模块包括可操作以执行如本文所描述的一个或多个功能的处理模块（如上所述）、功能块、硬件和／或存储在存储器的软件。注意，如果模块是经由硬件实现，硬件可独立地操作和／或结合软件和／或固件操作。当模块被实现为存储在存储器中的软件时，该模块能够操作以使用处理模块或其他硬件来执行存储在模块的存储器中的软件以执行本文所描述的功能。本文所述的模块可以包括一个或多个子模块，它们中的每一个可以是一个或多个模块，可以被包含在一个或多个其他模块中或包括一个或多个其他模块。

[0131] 虽然本文清楚地描述了本发明的各种功能和特征的特定组合，这些特征和功能的其他组合同样是可能的。在这里所描述的实施例并不受限于所描述的特定实施例，并且可以包括其他组合和实施例。

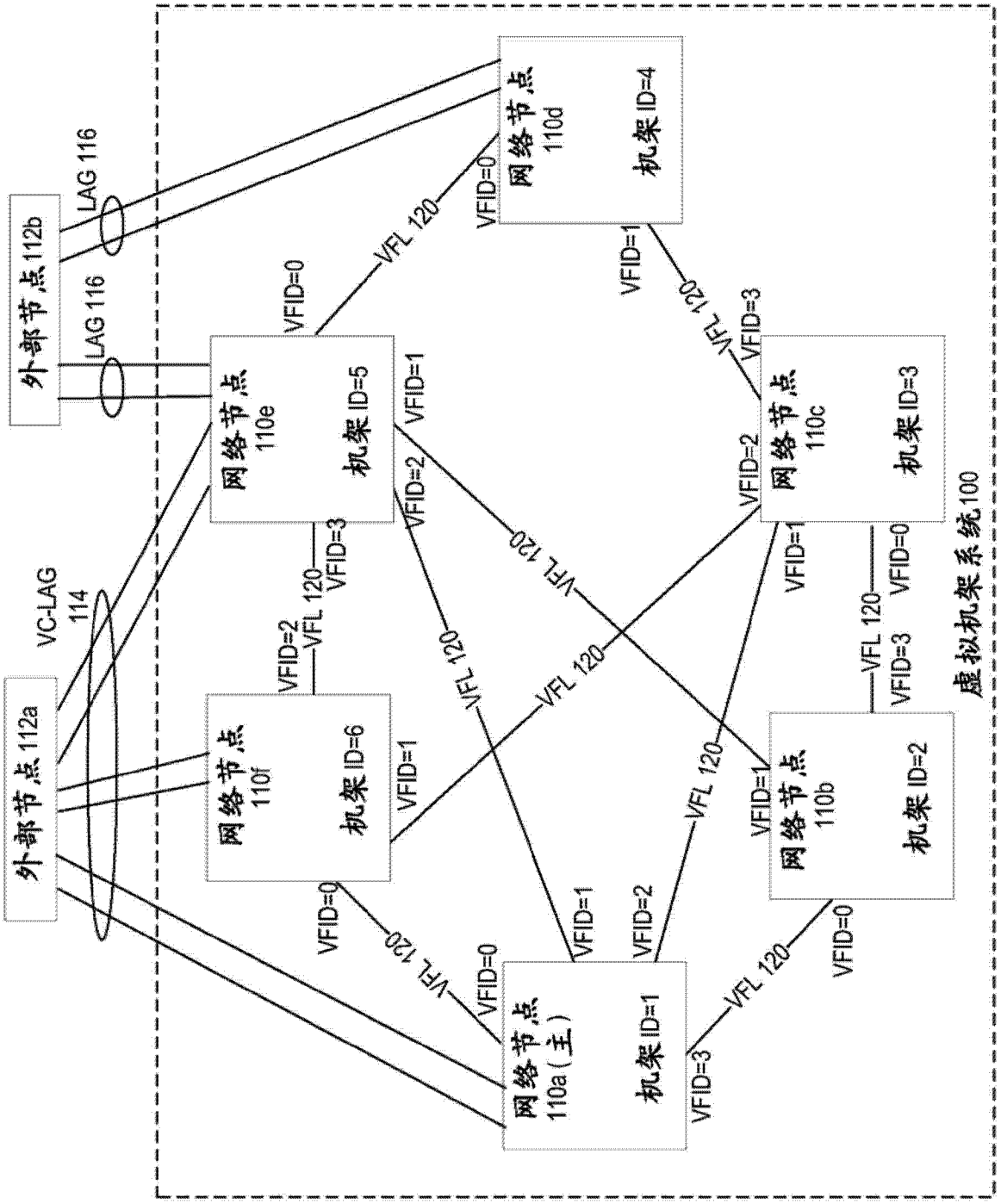

虚拟机架系统100
外部节点112a
外部节点112b
VC-LAG 114
LAG 116
LAG 116
网络节点 110a（主）
机架 ID=1
网络节点 110b
机架 ID=2
网络节点 110c
机架 ID=3
网络节点 110d
机架 ID=4
网络节点 110e
机架 ID=5
网络节点 110f
机架 ID=6
VFL 120
VFID=0
VFID=1
VFID=2
VFID=3

图 1a

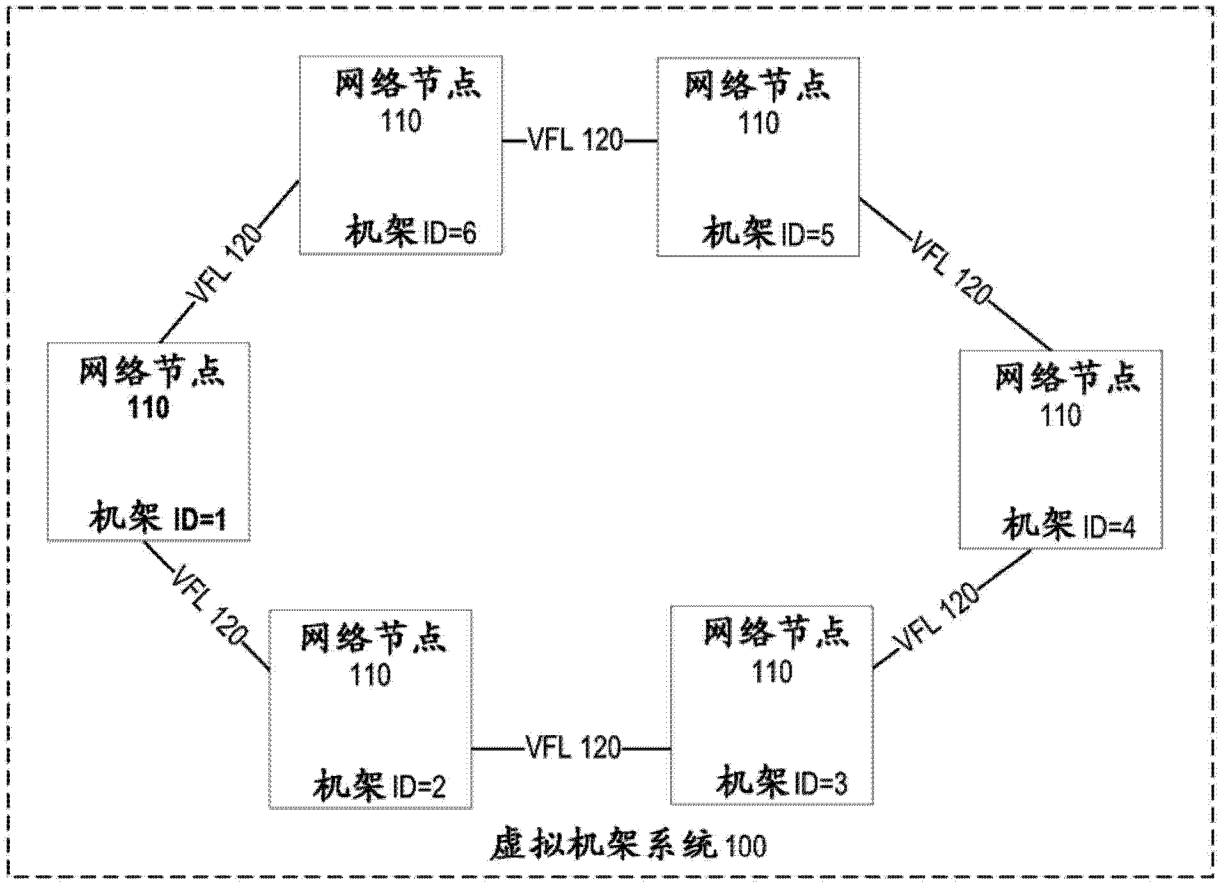

网络节点
110
机架ID=6
VFL 120
网络节点
110
机架ID=5
VFL 120
网络节点
110
机架ID=4
VFL 120
网络节点
110
机架ID=3
VFL 120
网络节点
110
机架ID=2
VFL 120
网络节点
110
机架 ID=1
VFL 120
虚拟机架系统100

图 1b

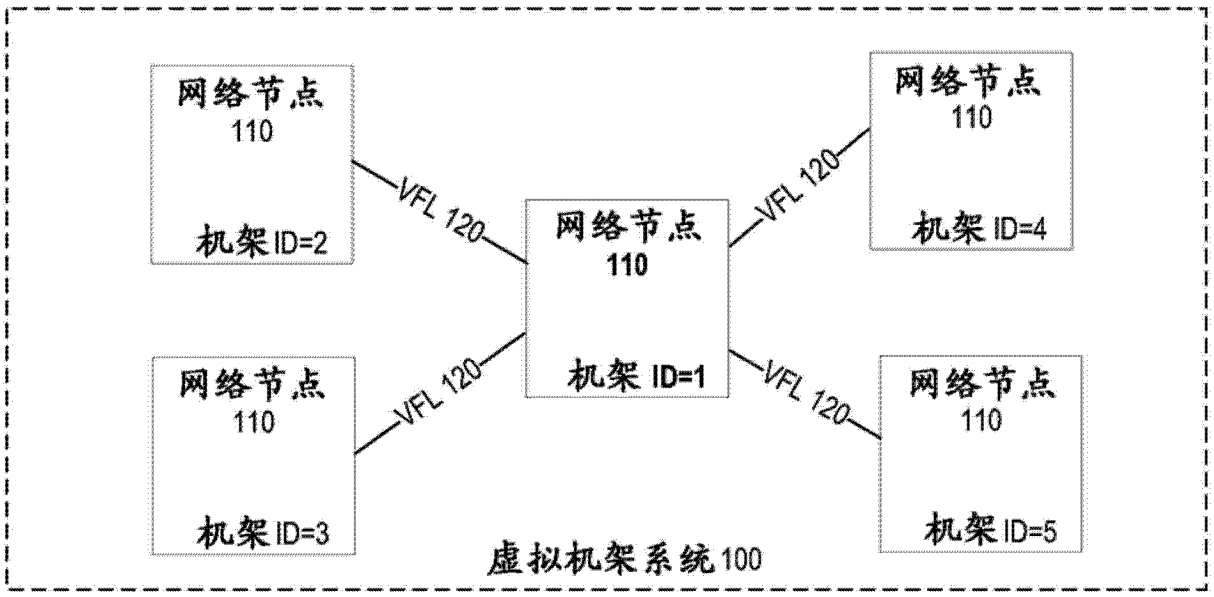

网络节点
110
机架ID=2
VFL 120
网络节点
110
机架ID=4
VFL 120
网络节点
110
机架 ID=1
VFL 120
网络节点
110
机架ID=3
VFL 120
网络节点
110
机架ID=5
虚拟机架系统100

图 1c

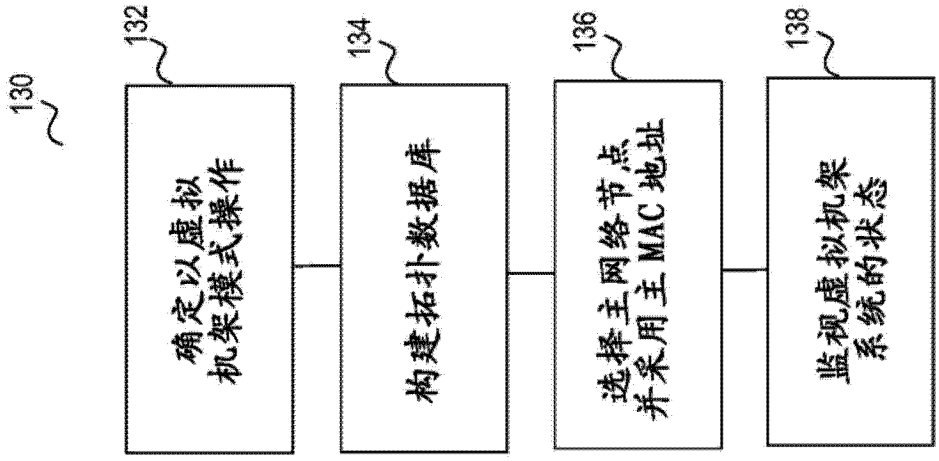

130
132
确定以虚拟机架模式操作
134
构建拓扑数据库
136
选择主网络节点并采用主 MAC 地址
138
监视虚拟机架系统的状态

图 2

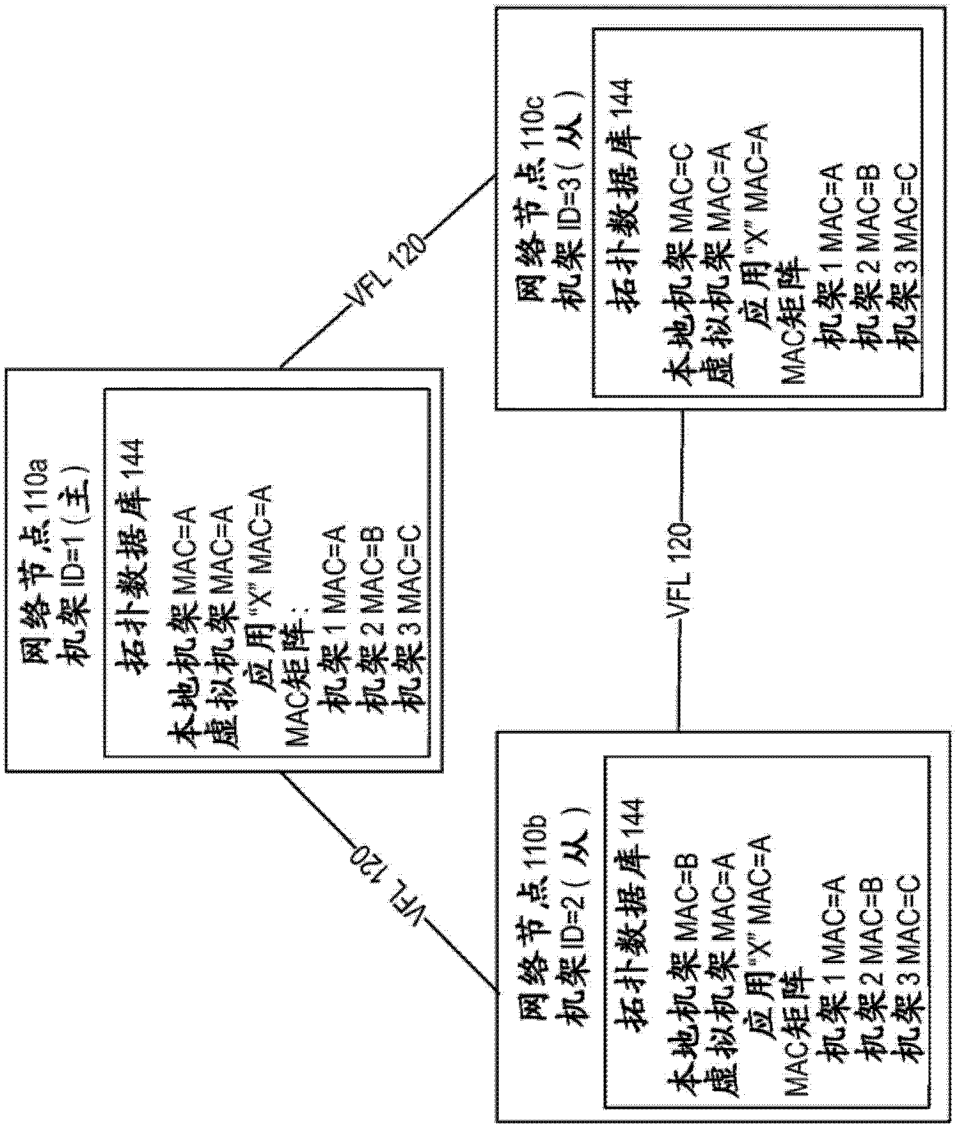

网络节点 110a
机架 ID=1（主）
拓扑数据库 144
本地机架 MAC=A
虚拟机架 MAC=A
应用 "X" MAC=A
MAC矩阵：
机架 1 MAC=A
机架 2 MAC=B
机架 3 MAC=C
网络节点 110c
机架 ID=3（从）
拓扑数据库 144
本地机架 MAC=C
虚拟机架 MAC=A
应用 "X" MAC=A
MAC矩阵
机架 1 MAC=A
机架 2 MAC=B
机架 3 MAC=C
网络节点 110b
机架 ID=2（从）
拓扑数据库 144
本地机架 MAC=B
虚拟机架 MAC=A
应用 "X" MAC=A
MAC矩阵
机架 1 MAC=A
机架 2 MAC=B
机架 3 MAC=C
VFL 120
VFL 120
VFL 120

图 3

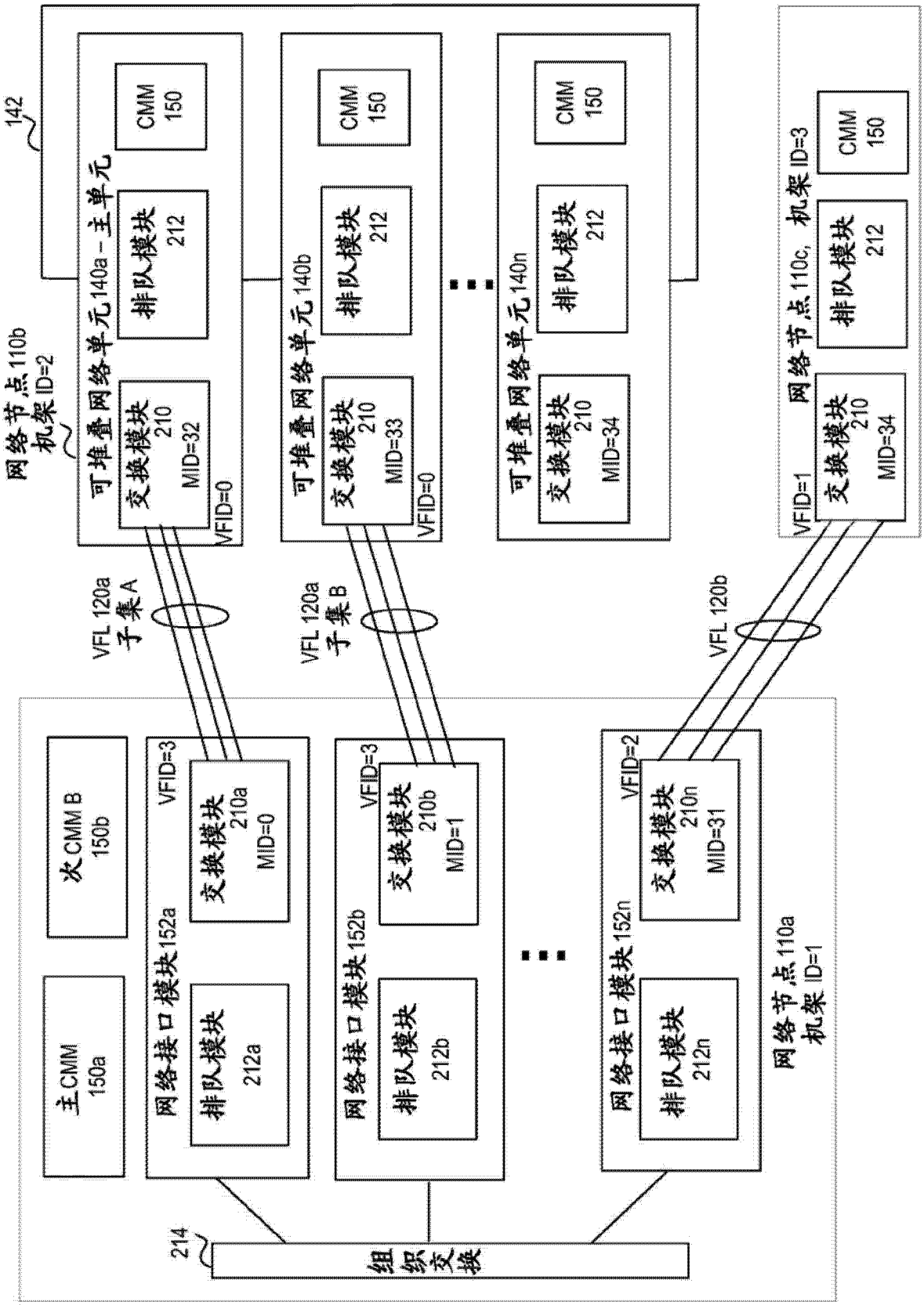

142
网络节点110b
机架ID=2
可堆叠网络单元140a – 主单元
CMM 150
排队模块 212
交换模块 210 MID=32
VFID=0
可堆叠网络单元140b
CMM 150
排队模块 212
交换模块 210 MID=33
VFID=0
可堆叠网络单元140n
CMM 150
排队模块 212
交换模块 210 MID=34
网络节点110c, 机架ID=3
CMM 150
排队模块 212
交换模块 210 MID=34
VFID=1
VFL 120a 子集A
VFL 120a 子集B
VFL 120b
次CMM B 150b
主CMM 150a
网络接口模块152a
VFID=3
交换模块 210a MID=0
排队模块 212a
网络接口模块152b
VFID=3
交换模块 210b MID=1
排队模块 212b
网络接口模块152n
VFID=2
交换模块 210n MID=31
排队模块 212n
网络节点110a
机架ID=1
214
组织交换

图 4

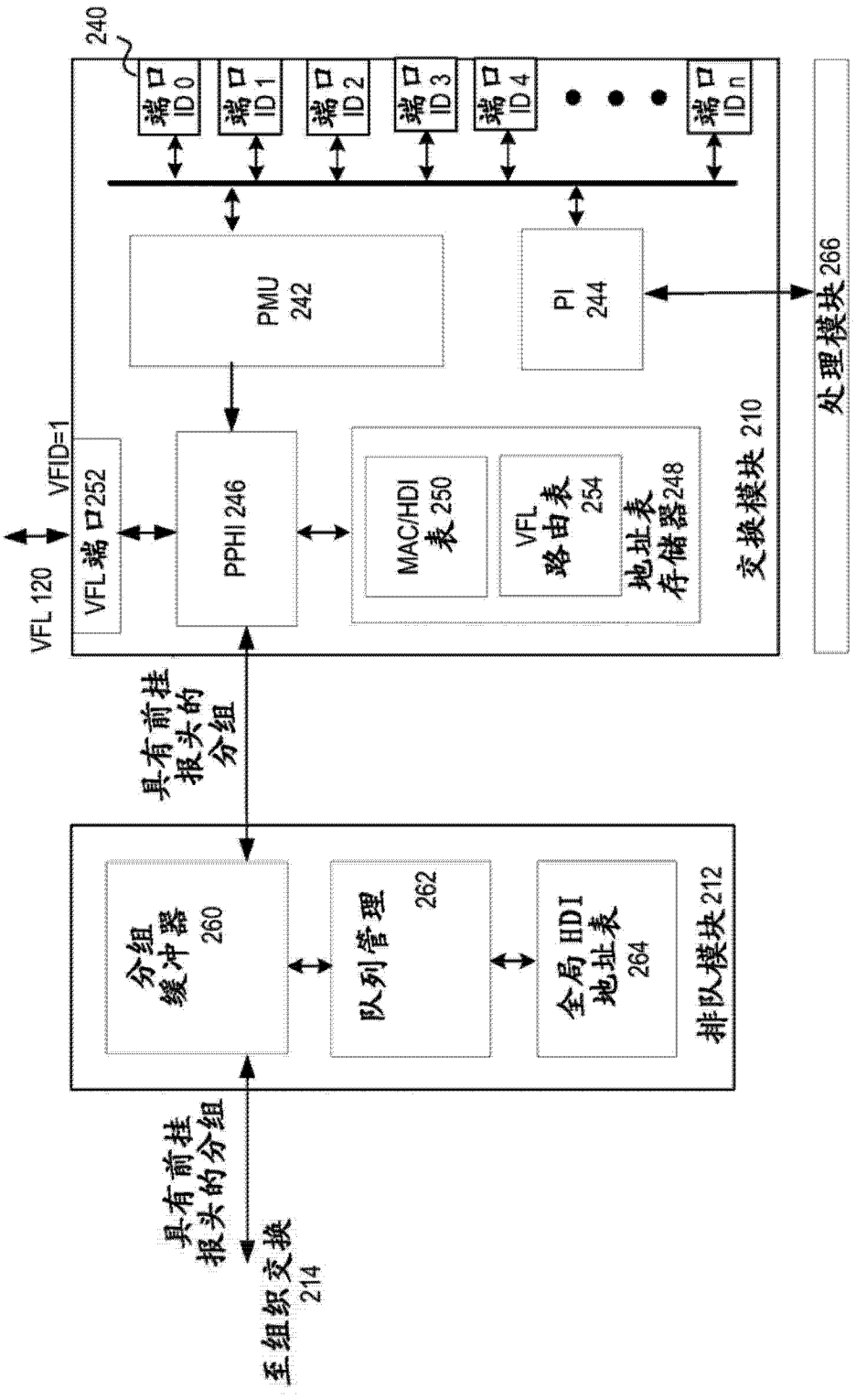

240
端口 ID 0
端口 ID 1
端口 ID 2
端口 ID 3
端口 ID 4
端口 ID n
PMU 242
PI 244
处理模块 266
VFID=1
VFL 120
VFL 端口 252
PPHI 246
MAC/HDI 表 250
VFL 路由表 254
地址表存储器 248
交换模块 210
具有前挂报头的分组
分组缓冲器 260
队列管理 262
全局 HDI 地址表 264
排队模块 212
具有前挂报头的分组
至组织交换 214

图 5


前挂报头 300
源 HDI 302
目的地 HDI 304
VLAN ID 306
分组类型 308
源 MAC 地址 310
目的地 MAC 地址 312
负载均衡指标 314
分组优先级 316

图 6

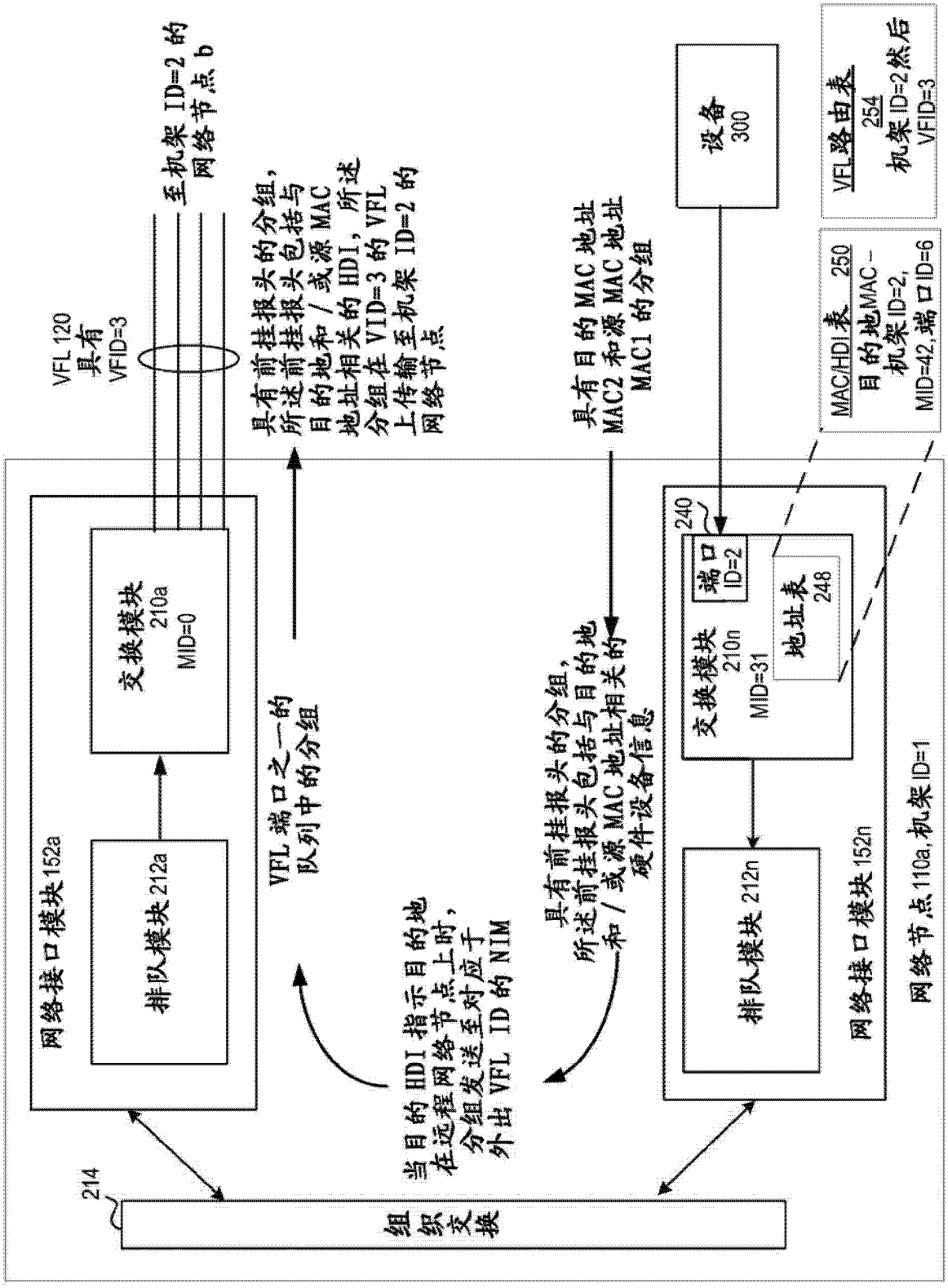

VFL 120
具有
VFID=3
至机架 ID=2 的
网络节点 b
具有前挂报头的分组，
所述前挂报头包括与
目的地和 / 或源 MAC
地址相关的 HDI，所述
分组在 VID=3 的 VFL
上传输至机架 ID=2 的
网络节点
具有目的 MAC 地址
MAC2 和源 MAC 地址
MAC1 的分组
设备
300
VFL 路由表
254
机架 ID=2 然后
VFID=3
MAC/HDI 表 250
目的地 MAC－
机架 ID=2,
MID=42, 端口 ID=6
网络接口模块 152a
排队模块 212a
交换模块
210a
MID=0
VFL 端口之一的
队列中的分组
当目的 HDI 指示目的地
在远程网络节点上时，
分组发送至对应于
外出 VFL ID 的 NIM
具有前挂报头的分组，
所述前挂报头包括与目的地
和 / 或源 MAC 地址相关的
硬件设备信息
240
端口
ID=2
交换模块
210n
MID=31
地址表
248
排队模块 212n
网络接口模块 152n
网络节点 110a, 机架 ID=1
214
组织交换

图 7

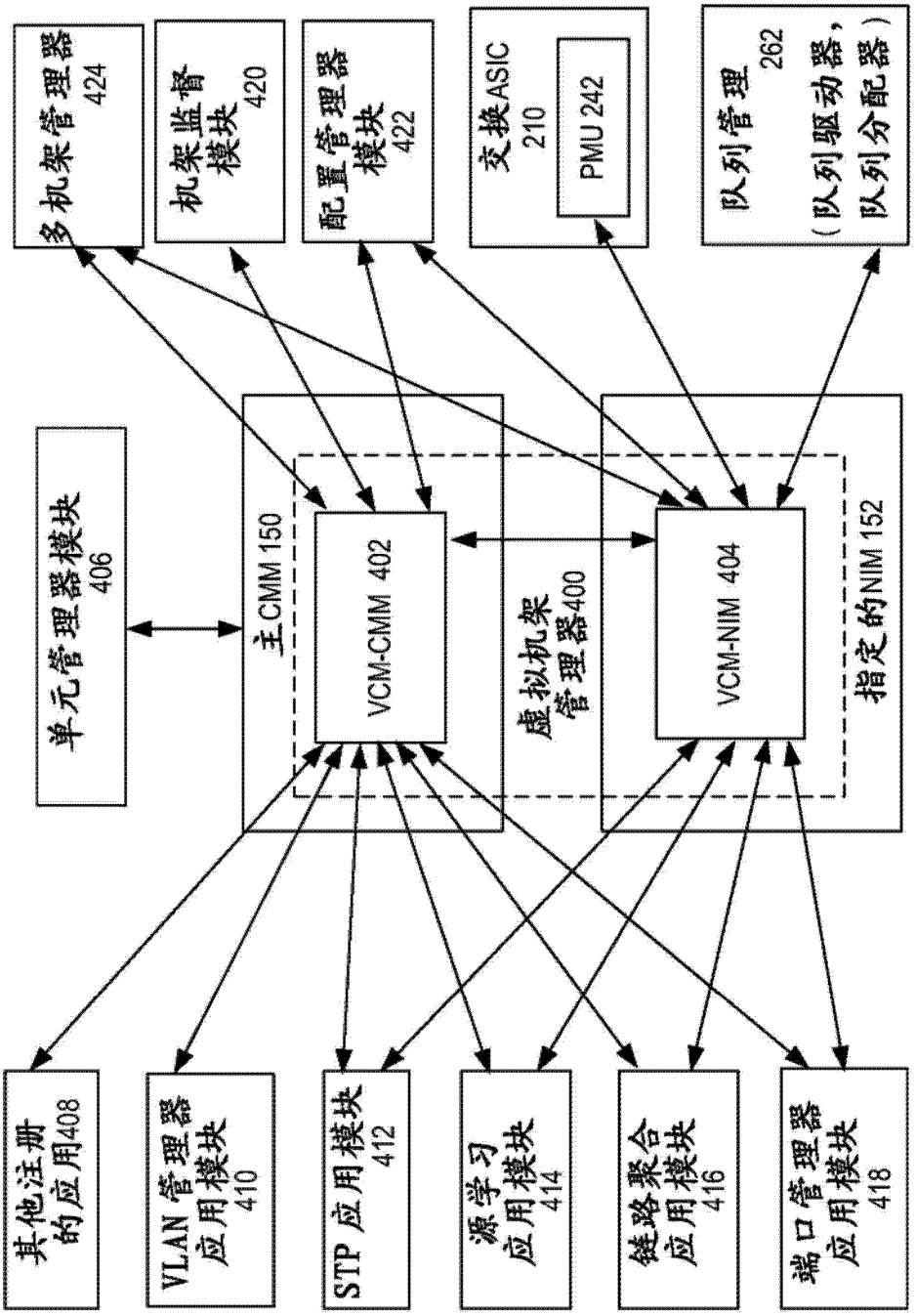

多机架管理器 424
机架监督模块 420
配置管理器模块 422
交换ASIC 210
PMU 242
队列管理 262
（队列驱动器，队列分配器）
单元管理器模块 406
主CMM 150
VCM-CMM 402
虚拟机架管理器400
VCM-NIM 404
指定的NIM 152
其他注册的应用408
VLAN 管理器应用模块 410
STP 应用模块 412
源学习应用模块 414
链路聚合应用模块 416
端口管理器应用模块 418

图 8

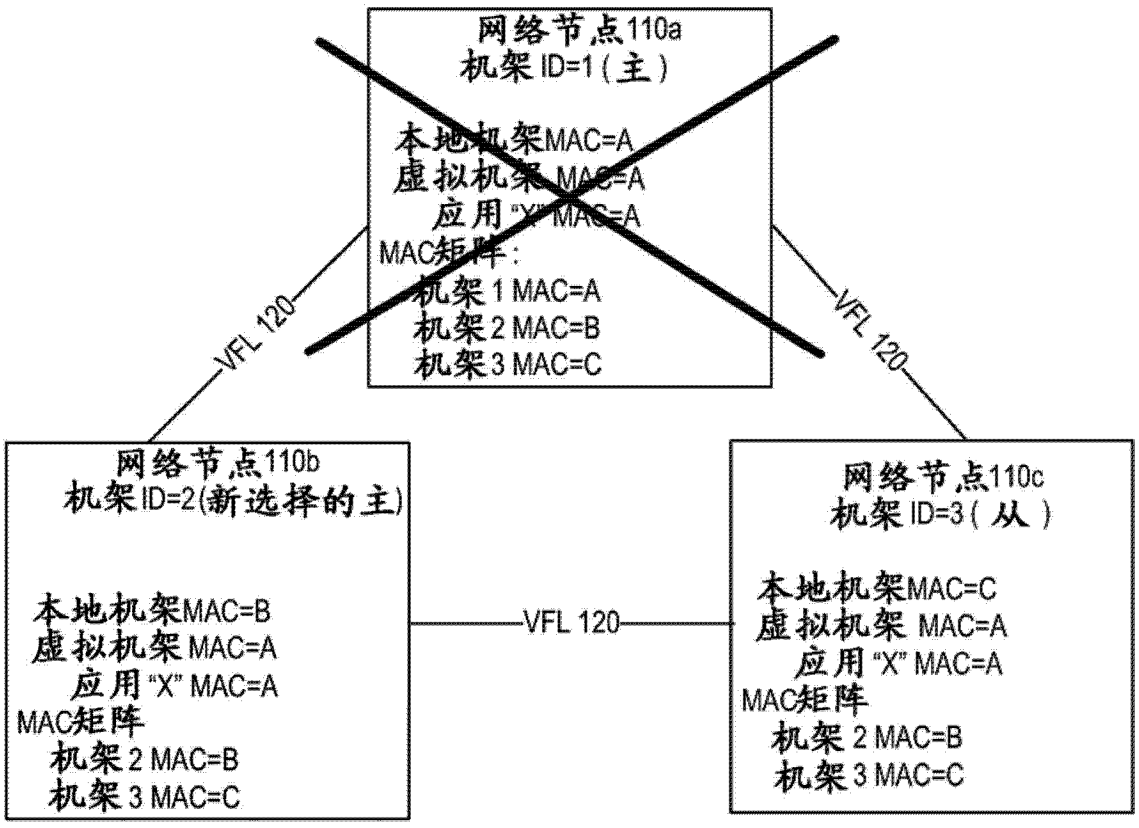

网络节点110a
机架ID=1(主)
本地机架MAC=A
虚拟机架 MAC=A
应用"X" MAC=A
MAC矩阵:
机架1 MAC=A
机架2 MAC=B
机架3 MAC=C
VFL 120
VFL 120
网络节点110b
机架ID=2(新选择的主)
本地机架MAC=B
虚拟机架 MAC=A
应用"X" MAC=A
MAC矩阵
机架2 MAC=B
机架3 MAC=C
VFL 120
网络节点110c
机架ID=3(从)
本地机架MAC=C
虚拟机架 MAC=A
应用"X" MAC=A
MAC矩阵
机架2 MAC=B
机架3 MAC=C

图 9

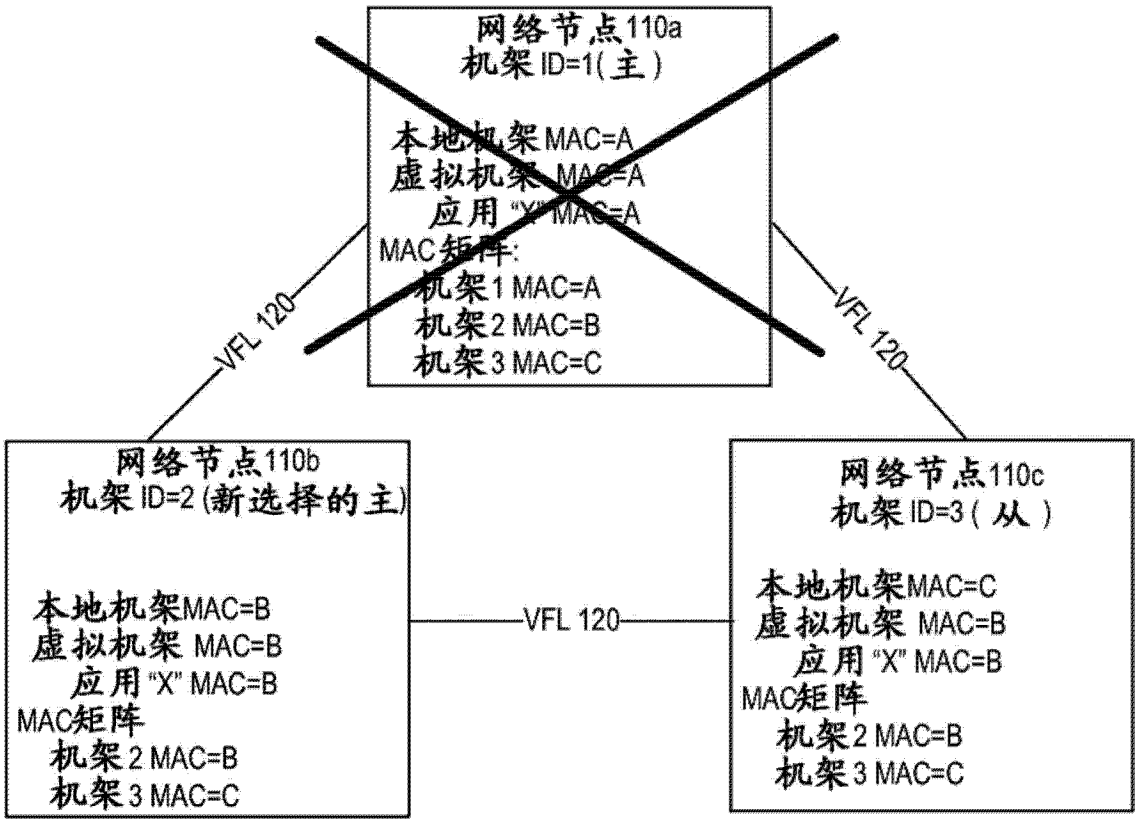

网络节点110a
机架ID=1(主)
本地机架MAC=A
虚拟机架 MAC=A
应用"X" MAC=A
MAC矩阵:
机架1 MAC=A
机架2 MAC=B
机架3 MAC=C
VFL 120
VFL 120
网络节点110b
机架ID=2(新选择的主)
本地机架MAC=B
虚拟机架 MAC=B
应用"X" MAC=B
MAC矩阵
机架2 MAC=B
机架3 MAC=C
VFL 120
网络节点110c
机架ID=3(从)
本地机架MAC=C
虚拟机架 MAC=B
应用"X" MAC=B
MAC矩阵
机架2 MAC=B
机架3 MAC=C

图 10

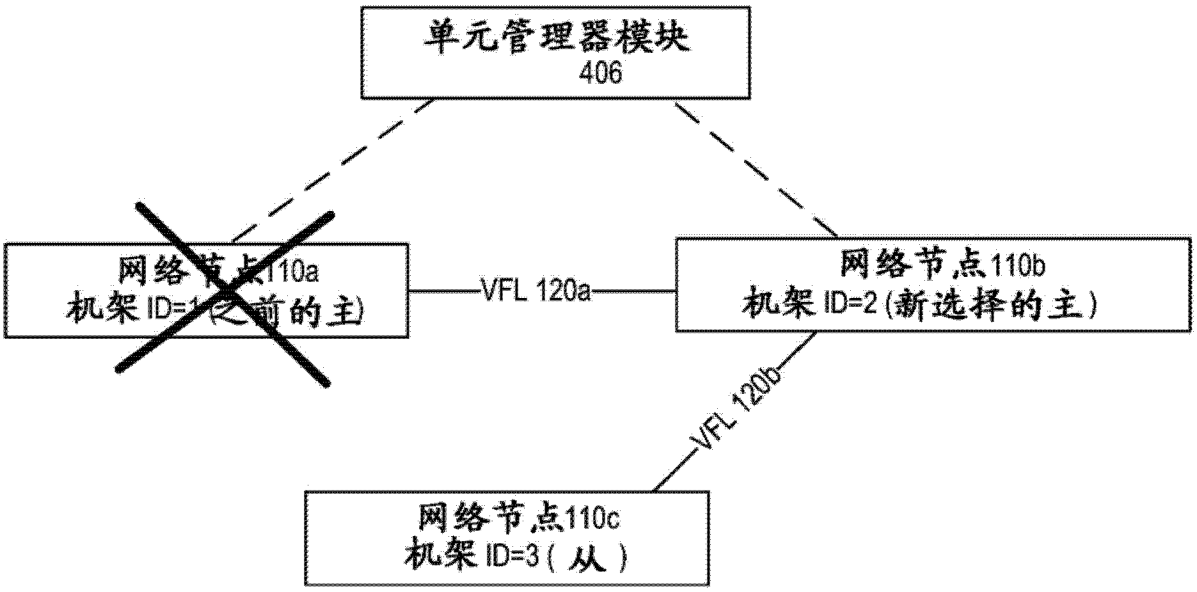

单元管理器模块
406
网络节点110a
机架ID=1(之前的主)
VFL 120a
网络节点110b
机架ID=2(新选择的主)
VFL 120b
网络节点110c
机架ID=3(从)

图 11

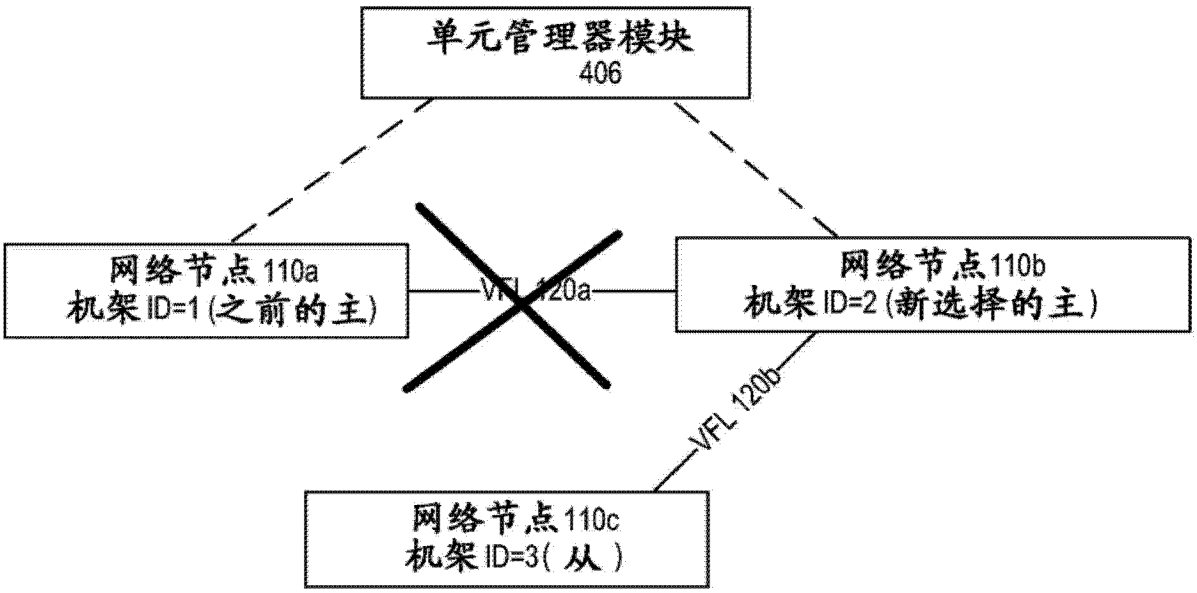

单元管理器模块
406
网络节点110a
机架ID=1(之前的主)
VFL 120a
网络节点110b
机架ID=2(新选择的主)
VFL 120b
网络节点110c
机架ID=3(从)

图 12

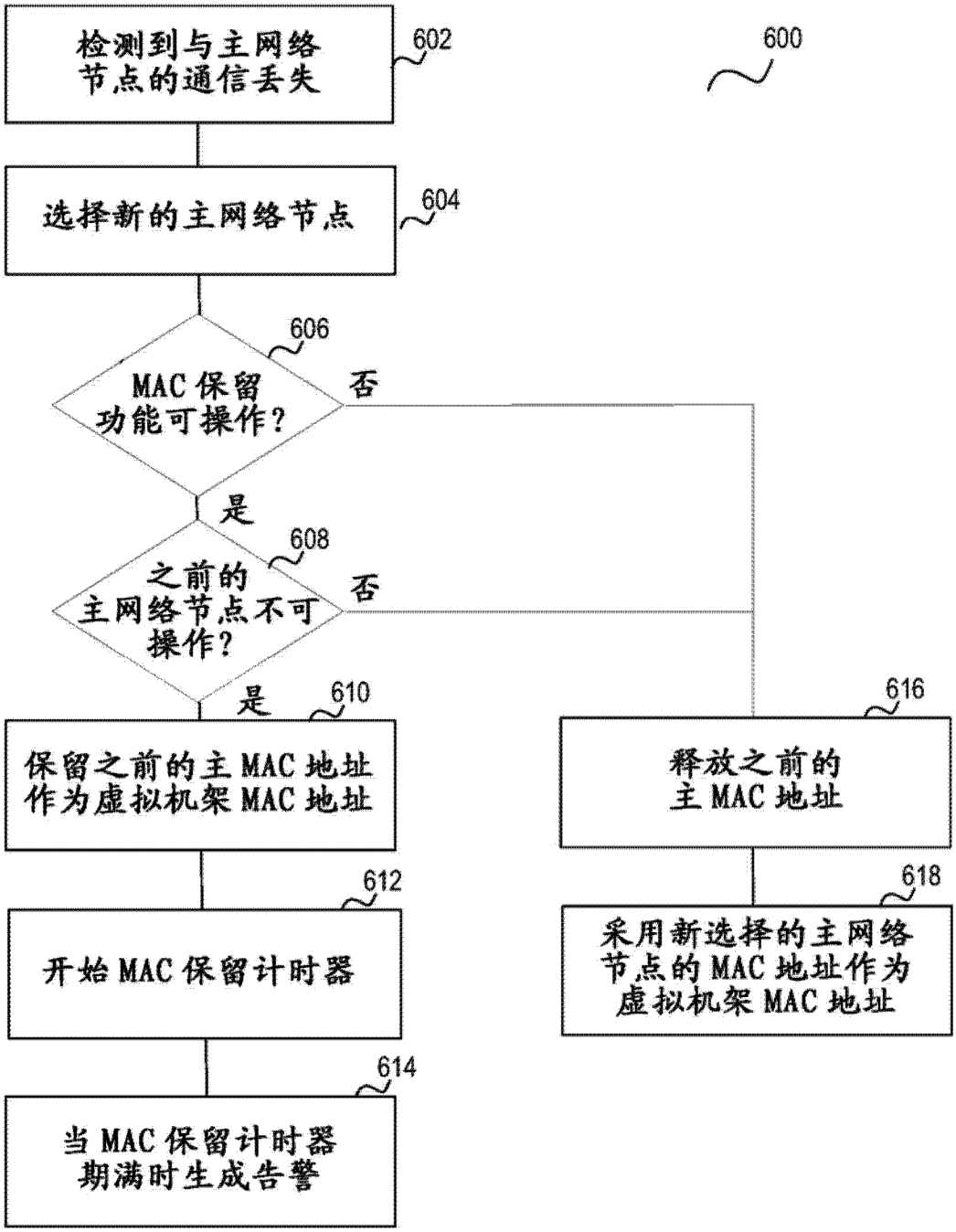

检测到与主网络节点的通信丢失
602
600
选择新的主网络节点
604
606
MAC 保留功能可操作？
否
是
608
之前的主网络节点不可操作？
否
是
610
保留之前的主 MAC 地址作为虚拟机架 MAC 地址
616
释放之前的主 MAC 地址
612
开始 MAC 保留计时器
618
采用新选择的主网络节点的 MAC 地址作为虚拟机架 MAC 地址
614
当 MAC 保留计时器期满时生成告警

图 13

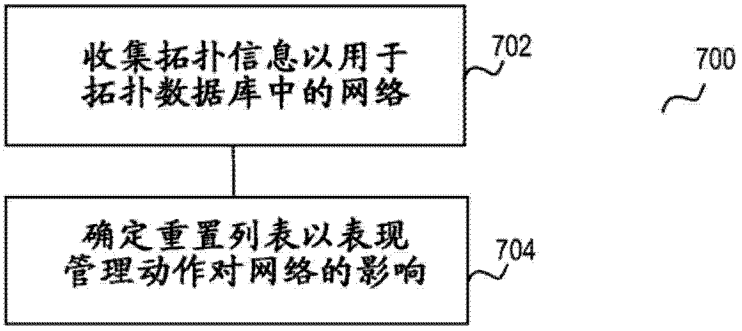

收集拓扑信息以用于
拓扑数据库中的网络
702
700
确定重置列表以表现
管理动作对网络的影响
704

图 14

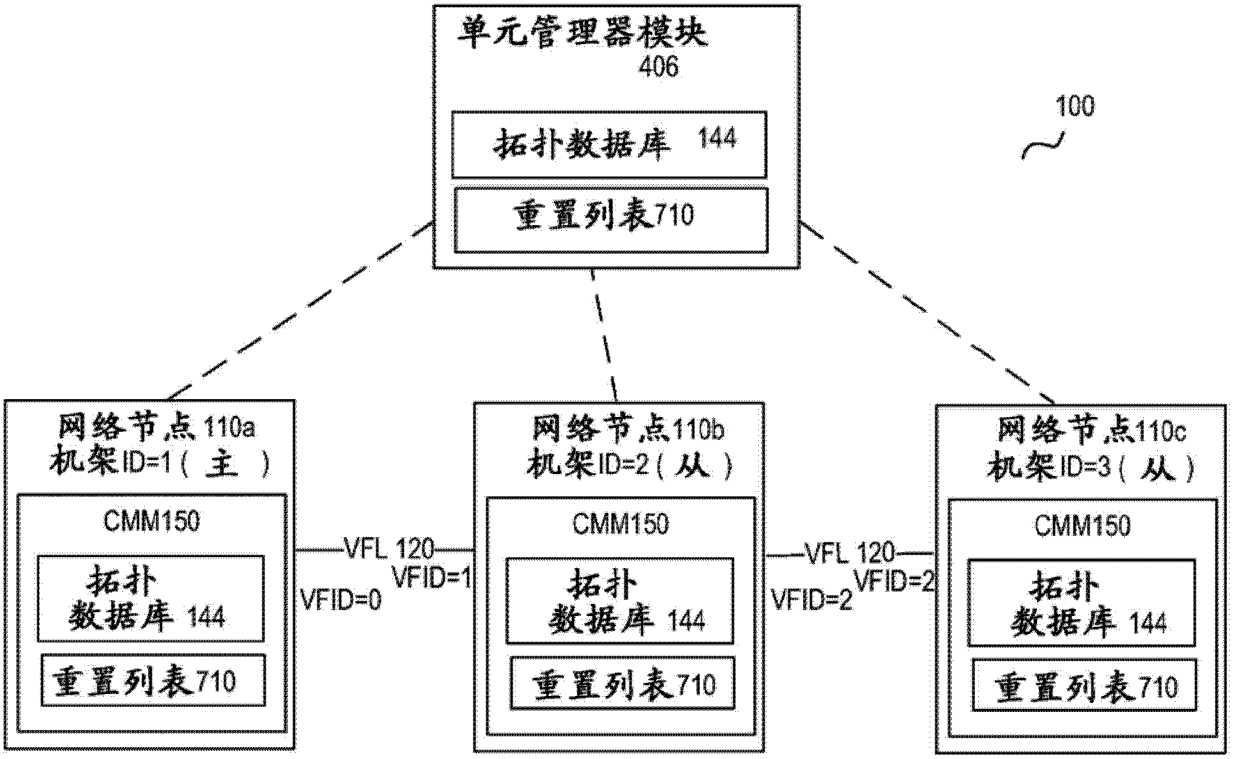

单元管理器模块
406
拓扑数据库 144
重置列表710
100
网络节点 110a
机架ID=1（主）
CMM150
拓扑
数据库 144
重置列表710
VFL 120
VFID=0
VFID=1
网络节点110b
机架ID=2（从）
CMM150
拓扑
数据库 144
重置列表710
VFL 120
VFID=2
VFID=2
网络节点110c
机架ID=3（从）
CMM150
拓扑
数据库 144
重置列表710

图 15

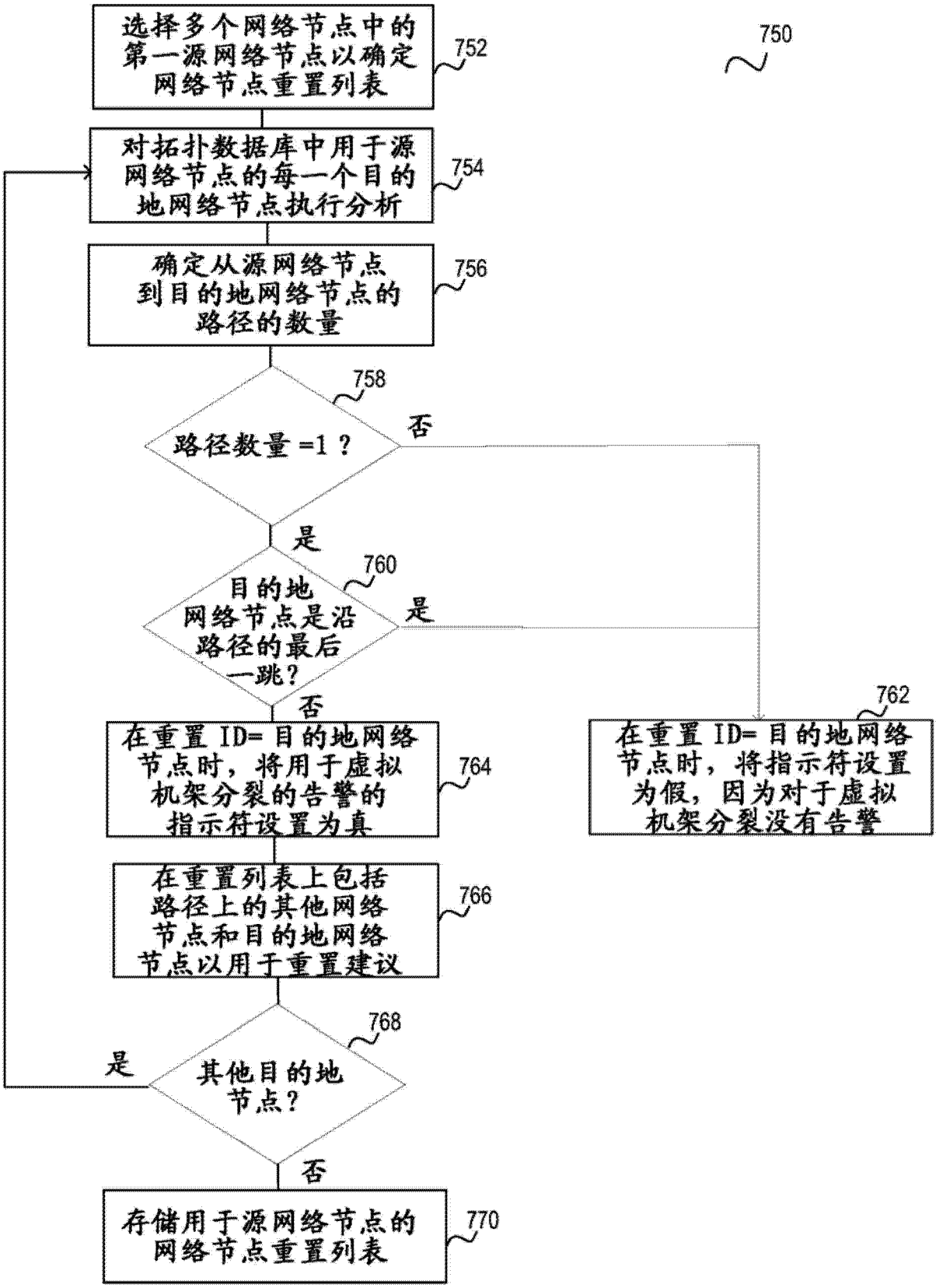

选择多个网络节点中的第一源网络节点以确定网络节点重置列表
752
750
对拓扑数据库中用于源网络节点的每一个目的地网络节点执行分析
754
确定从源网络节点到目的地网络节点的路径的数量
756
758
路径数量 =1 ？
否
是
760
目的地网络节点是沿路径的最后一跳？
是
否
762
在重置 ID= 目的地网络节点时，将指示符设置为假，因为对于虚拟机架分裂没有告警
在重置 ID= 目的地网络节点时，将用于虚拟机架分裂的告警的指示符设置为真
764
在重置列表上包括路径上的其他网络节点和目的地网络节点以用于重置建议
766
768
其他目的地节点？
是
否
存储用于源网络节点的网络节点重置列表
770

图 16

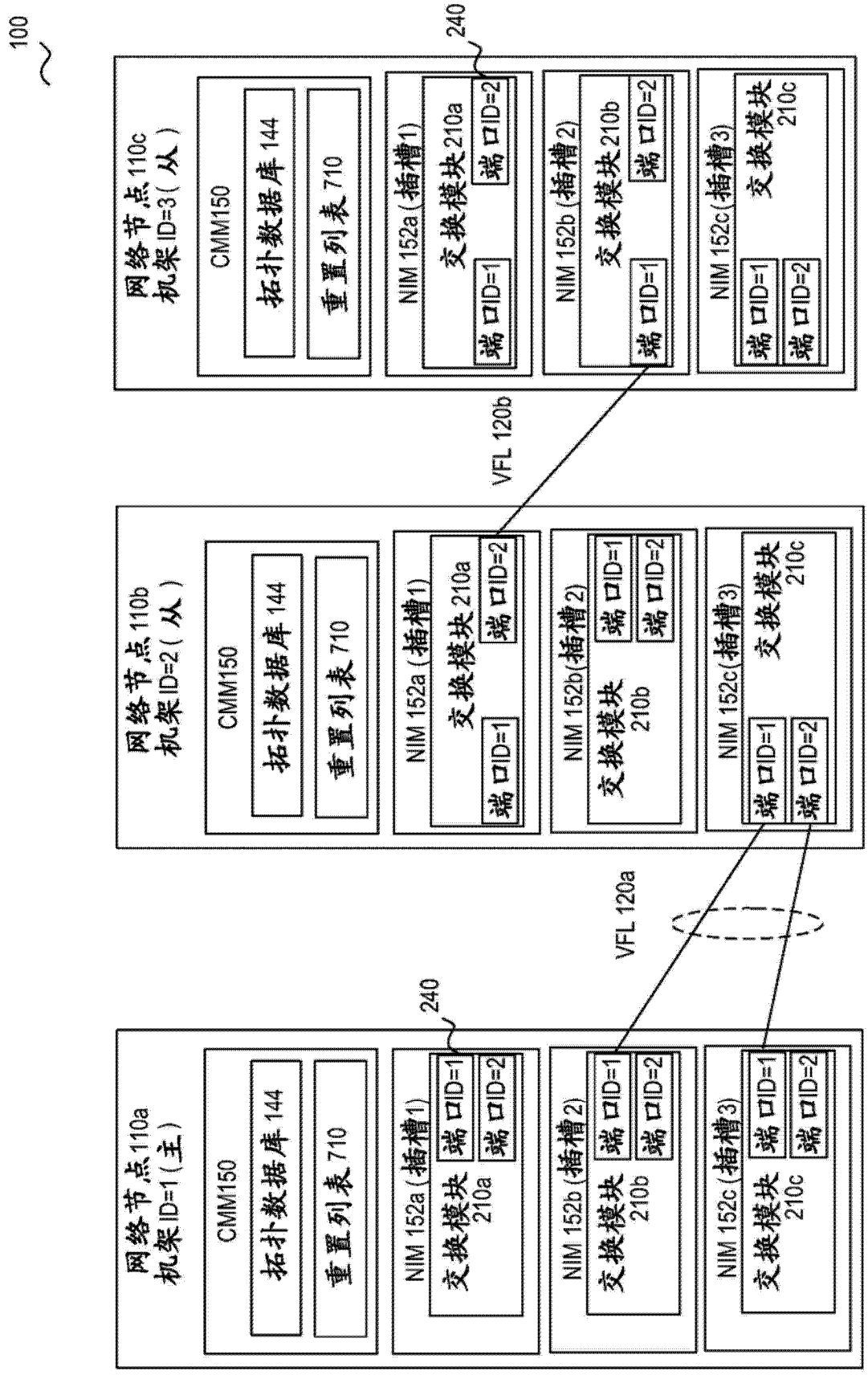

100
网络节点110c
机架ID=3(从)
CMM150
拓扑数据库144
重置列表710
NIM 152a(插槽1)
交换模块210a
端口ID=1
端口ID=2
240
NIM 152b(插槽2)
交换模块210b
NIM 152c(插槽3)
交换模块210c
VFL 120b
网络节点110b
机架ID=2(从)
VFL 120a
网络节点110a
机架ID=1(主)

图 17

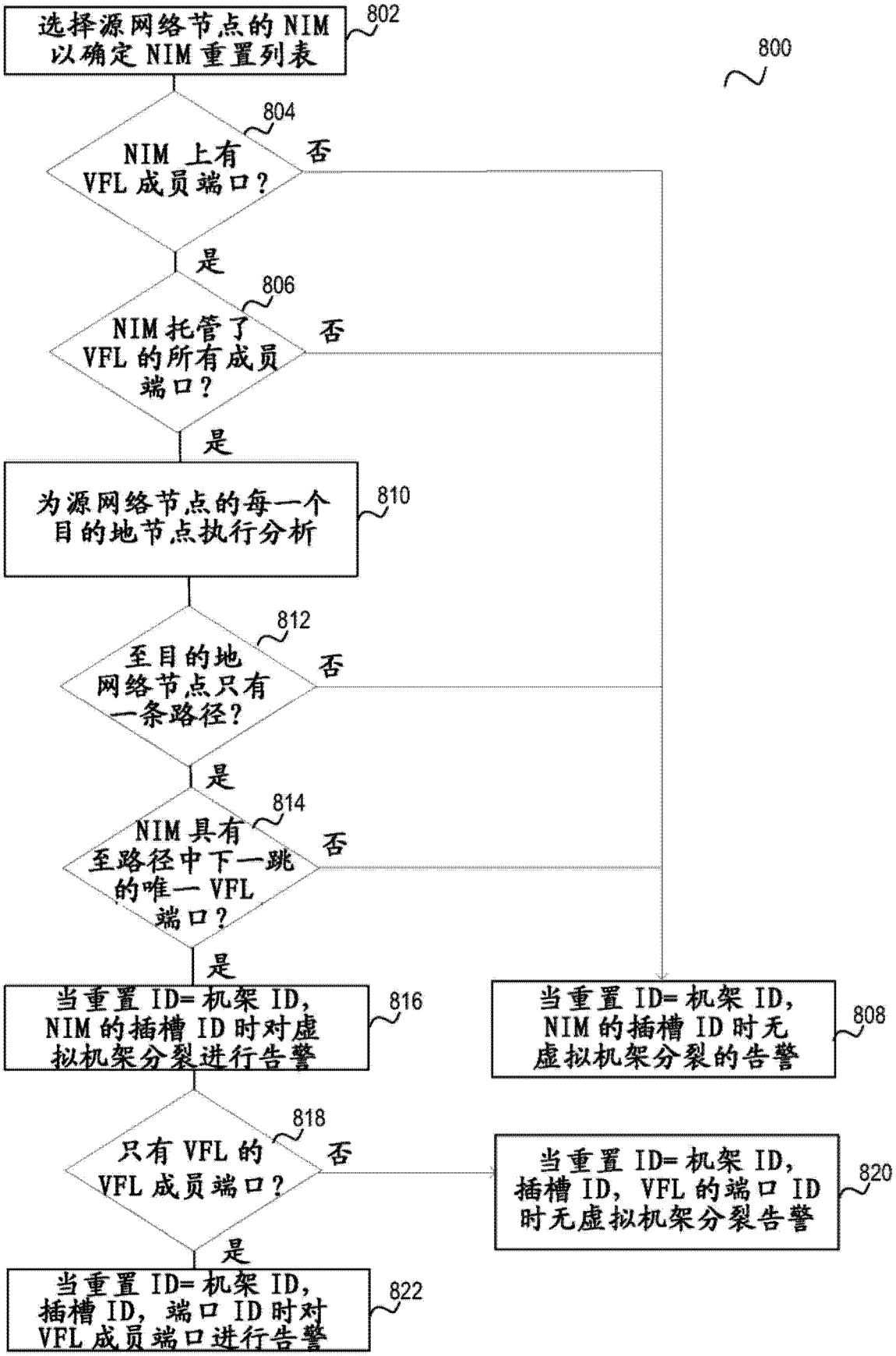

选择源网络节点的 NIM 以确定 NIM 重置列表
802
800
NIM 上有 VFL 成员端口？
804
否
是
NIM 托管了 VFL 的所有成员端口？
806
否
是
为源网络节点的每一个目的地节点执行分析
810
至目的地网络节点只有一条路径？
812
否
是
NIM 具有至路径中下一跳的唯一 VFL 端口？
814
否
是
当重置 ID= 机架 ID，NIM 的插槽 ID 时对虚拟机架分裂进行告警
816
当重置 ID= 机架 ID，NIM 的插槽 ID 时无虚拟机架分裂的告警
808
只有 VFL 的 VFL 成员端口？
818
否
当重置 ID= 机架 ID，插槽 ID，VFL 的端口 ID 时无虚拟机架分裂告警
820
是
当重置 ID= 机架 ID，插槽 ID，端口 ID 时对 VFL 成员端口进行告警
822

图 18

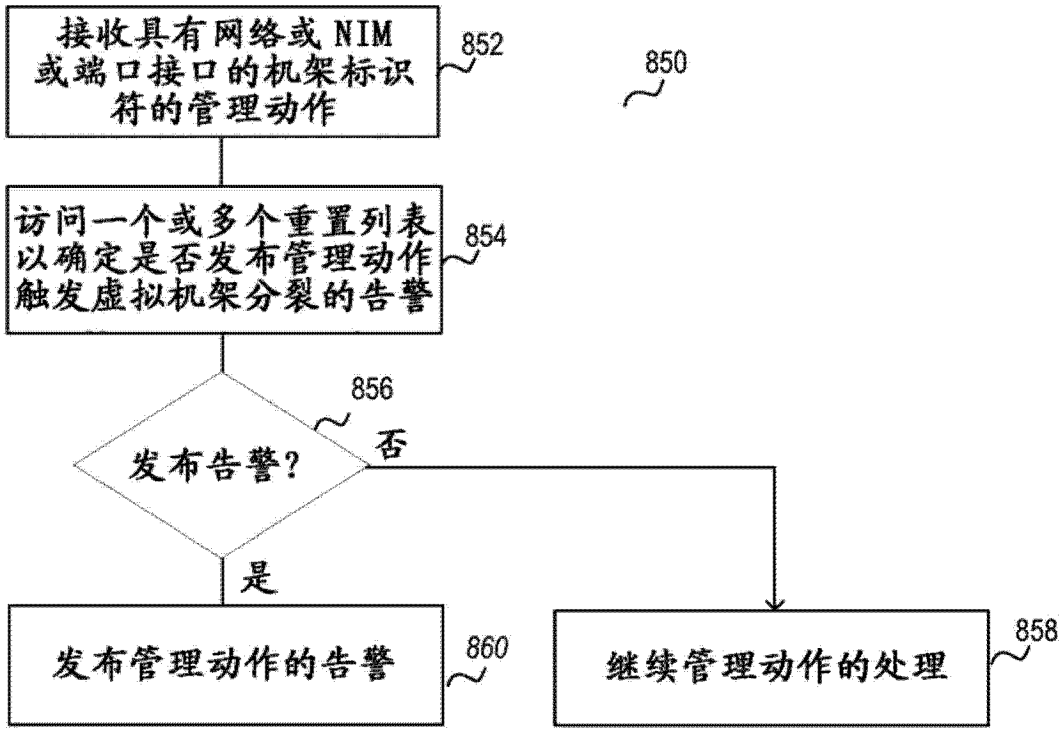

接收具有网络或 NIM 或端口接口的机架标识符的管理动作
852
850
访问一个或多个重置列表以确定是否发布管理动作触发虚拟机架分裂的告警
854
856
发布告警?
否
是
发布管理动作的告警
860
继续管理动作的处理
858

图 19

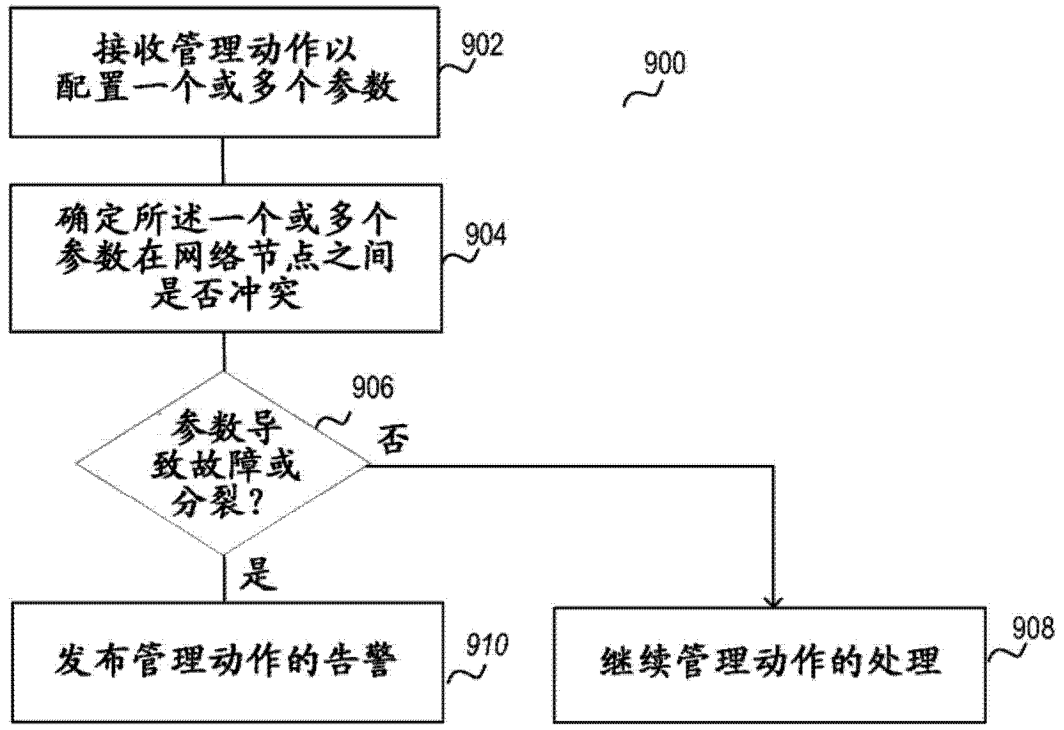

接收管理动作以配置一个或多个参数
902
900
确定所述一个或多个参数在网络节点之间是否冲突
904
906
参数导致故障或分裂?
否
是
发布管理动作的告警
910
继续管理动作的处理
908

图 20