(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 103098424 B
(45)授权公告日 2015.09.30

(21)申请号 201180038254.7

(22)申请日 2011.07.19

(30)优先权数据
61/370,622 2010.08.04 US
13/010,168 2011.01.20 US

(85)PCT国际申请进入国家阶段日
2013.02.04

(86)PCT国际申请的申请数据
PCT/US2011/044527 2011.07.19

(87)PCT国际申请的公布数据
WO2012/018521 EN 2012.02.09

(73)专利权人 阿尔卡特朗讯公司
地址 法国巴黎

(72)发明人 R·H·雅各布 达 席尔瓦
C-H·A·张 A·维纳亚加姆
S·K·莫汉达斯 J·达拉夸
J·B·翁

(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 12/703(2013.01)

(56)对比文件
CN 1826769 A,2006.08.30,
US 2010020680 A1,2010.01.28,
WO 2010069382 A1,2010.06.24,

审查员 侯艳兰

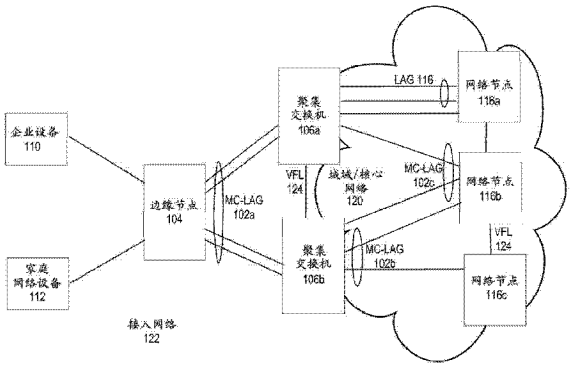
权利要求书2页 说明书17页 附图12页

(54)发明名称

用于多机架链路聚集的系统和方法

(57)摘要

聚集交换机通过多机架链路聚集群组连接至边缘节点,其中聚集交换机通过虚拟组织链路来连接,所述虚拟组织链路提供一种连接,用于在关于MAC寻址的聚集交换机之间交换信息以同步MAC地址表。



CN 103098424 B

1.一种多机架系统中的聚集交换机,包括 :

连接至远程聚集交换机的虚拟组织链路,其中所述远程聚集交换机处于单独物理机架中 ;

第一网络接口模块,能够操作为 :

在外部端口接口上接收进入分组,其中所述分组包括目的地地址 ;

访问 MAC/HDI 转发表,其包括与聚集交换机和远程聚集交换机对应的硬件设备信息 ;

基于进入分组的目的地地址和所述 MAC/HDI 转发表确定与远程聚集交换机上硬件设备对应的目的地硬件设备信息 ;

从进入分组生成具有前挂报头的分组,其中所述前挂报头包括目的地硬件设备信息 ;和

在虚拟组织链路上向远程聚集交换机发送具有前挂报头的分组。

2.如权利要求 1 所述的聚集交换机,还包括 :

具有连接至虚拟组织链路的虚拟组织链路端口的第二网络接口模块 ;

连接至第一和第二网络接口模块的组织电路 ;和

其中第一网络接口模块在所述组织电路上向第二网络接口模块发送具有前挂报头的分组,用于在虚拟组织链路上传输。

3.如权利要求 2 所述的聚集交换机,其中所述第一网络接口模块包括 :

包括多个外部端口接口的交换电路,其中所述第一网络接口模块在多个外部端口接口之一上接收进入分组 ;和

包括队列管理模块的排队电路。

4.如权利要求 3 所述的聚集交换机,其中所述排队电路能够操作为 :

访问全局 HDI 地址表,其包括与聚集交换机和远程聚集交换机对应的硬件设备信息的映射 ;和

基于所述目的地硬件设备信息和所述全局 HDI 地址表确定向连接至虚拟组织链路的第二网络接口模块发送具有前导报头的分组。

5.如权利要求 3 所述的聚集交换机,其中所述交换电路的外部端口接口的一个或多个是连接至边缘节点的多机架链路聚集群组 (MC-LAG) 的成员端口 ;和

其中所述远程聚集交换机包括连接至边缘节点的 MC-LAG 的一个或多个成员端口。

6.一种用于多机架链路聚集的方法,包括 :

在外部端口接口上接收第一分组,其中所述第一分组包括目的地地址 ;

访问 MAC/HDI 转发表,其包括远程聚集交换机的硬件设备信息,并确定与第一分组的目的地地址对应的目的地硬件设备信息 ;

确定与接收第一分组的外部端口接口对应的源硬件设备信息 ;

从第一分组生成具有前挂报头的分组,其中所述前挂报头包括源硬件设备信息和目的地硬件设备信息 ;和

在虚拟组织链路上向远程聚集交换机发送具有前挂报头的分组,其中所述远程聚集交换机处于单独物理机架中。

7.如权利要求 6 所述的方法,还包括 :

访问全局 HDI 地址表,其包括硬件设备信息的映射 ;和

基于所述目的地硬件设备信息和所述全局 HDI 地址表确定在虚拟组织链路上向远程聚集交换机发送具有前导报头的分组。

8. 如权利要求 7 所述的方法，其中所述 MAC/HDI 转发表具有 MAC 地址项目的列表和用于路由分组以到达具有相关 MAC 地址的设备的相关硬件设备信息。

9. 如权利要求 8 所述的方法，还包括：

在虚拟组织链路上从远程聚集交换机接收具有前挂报头的第二分组，其中所述第二分组的前挂报头包括未知源地址和相关硬件设备信息，以及其中所述相关硬件设备信息识别远程聚集交换机上的硬件设备；和

在 MAC/HDI 转发表中存储所述未知源地址和所述相关硬件设备信息。

10. 如权利要求 9 所述的方法，还包括：

确定相关硬件设备信息是否识别远程聚集交换机上的多机架链路聚集群组 (MC-LAG) 的远程成员端口；

当所述相关硬件设备信息识别远程聚集交换机上的 MC-LAG 的远程成员端口时，确定 MC-LAG 的本地成员端口；和

在 MAC/HDI 转发表中存储 MC-LAG 的本地成员端口的未知源地址和硬件设备信息。

用于多机架链路聚集的系统和方法

[0001] 相关专利申请的交叉引用

[0002] 本美国实用新型专利申请根据35U.s.c.§119主张2010年8月4日递交的题为“MULTI-CHASSIS VIRTUAL-FABRIC LINKAGGREGATION SYSTEM”的美国临时申请No.61/370,622的优先权，其通过引用合并于此，并为了所有目的作为当前美国实用新型专利申请的一部分。

技术领域

[0003] 本发明一般地涉及数据网络，更具体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

[0004] 数据网络允许许多不同计算设备(例如个人计算机、IP电话设备或服务器)彼此通信和/或与各个其他网络元件或附连至网络的远程服务器通信。例如，数据网络可不限制地包括城域以太网或企业以太网网络，他们支持包括例如IP语音(VoIP)、数据和视频应用的多个应用。这样的网络常规地包括许多互连的节点，通常已知为交换机或路由器，用于通过网络路由业务。

[0005] 通常，各个节点基于他们在网络的特定区域内的位置(共同地表征为2个或3个“等级”或“层次”，依据网络的大小)来区分。传统上，3层网络包括边缘层、聚集层和核心层(而2层网络仅包括边缘层和核心层)。数据网络的边缘层包括边缘(还称为接入)网络，这典型地提供从企业网络或家庭网络(例如局域网)到城域网络或核心网络的连接。边缘/接入层是网络的入口点，即，名义上与客户网络附连，并且位于边缘层处的交换机已知为边缘节点。不同类型的边缘网络包括数字订户线路、混合光纤同轴电缆(HFC)和光纤到户。边缘节点可执行例如对于附连设备的L2交换功能。边缘节点通常连接至聚集层，后者中止来自多个边缘节点的接入链路。位于聚集层的交换机已知为聚集交换机。聚集交换机可执行例如L2交换和经由从边缘节点的聚集链路接收的业务的L3路由。聚集层连接至城域或核心网络层，后者执行从聚集交换机(3层网络中)或从边缘节点(2层网络中)接收的业务的层3/IP路由。可理解，在网络的每个增加层处的节点典型地具有更大的容量和更快的吞吐量。

[0006] 数据网络面临的主要挑战之一是网络弹性的需求，即，保持高可用性的能力，尽管可能发生组件故障、链路故障等，这对于提供满意的网络性能是关键性的。网络弹性可通过拓扑冗余部分地实现，即，通过提供冗余节点(和节点中的冗余组件)以及节点之间的多个物理路径以防止单点故障，以及可通过L2/L3协议部分地实现，以在故障发生时采用冗余，从而收敛于用于通过网络交换/路由业务流的备用路径。可理解，检测和收敛时间必须快速(有利地，小于1秒)发生，以实现到备用路径的无缝转换。

[0007] 以太网协议是普遍存在于局域网(LAN)(例如家庭和企业网络)中以在计算机和网络之间进行通信的传输技术。然而，在接入和聚集网络以及城域网络中以太网协议技术的

使用不断增加，并且正如企业网络那样改革了边缘网络。作为接入技术，以太网提供了比其他接入技术的明显优点，例如：(i)对于数据、视频和语音应用的适应未来的传输；(ii)对于数据服务的成本有效的架构；和(iii)将确保互操作性的简单的、全局接受的标准。

[0008] 为了使得以太网技术适合于边缘和聚集层网络的运营级服务环境，要解决多个问题，包括对故障的弹性。在一个已知的方案中，生成树协议(STP)被共同地用来检测故障并在以太网网络中发生故障时将业务转移至备用路径。通常，STP依赖于交换机之间的多个物理路径，但是在任一个时间仅一个路径为活动，其他路径位于阻塞模式下(定义“活动/被动”范例)。当发生故障时，备用路径离开阻塞模式进入活动状态，从而重建连接。

[0009] 然而，STP可导致一些网络拓扑中不可接受的收敛时间(例如，多达若干秒)，不限制地包括，数据网络的边缘节点和聚集交换机之间的收敛。此外，STP仅提供活动/被动操作范例，而并非所有链路在相同时间主动转发业务。

[0010] 因此，需要在一个或多个数据网络的节点之间(例如不限制地，在以太网网络的边缘节点和聚集交换机之间)提供弹性的系统和方法。

附图说明

[0011] 图1示出根据本发明的网络架构的实施例的示意性框图；

[0012] 图2示出根据本发明的多机架系统的实施例的示意性框图；

[0013] 图3示出根据本发明的多机架系统中聚集交换机的实施例的示意性框图；

[0014] 图4示出根据本发明的多机架系统中聚集交换机的网络接口模块的实施例的示意性框图；

[0015] 图5示出根据本发明的通过多机架系统中聚集交换机的分组流的实施例的示意性框图；

[0016] 图6示出根据本发明的多机架系统中进行源地址认知的实施例的示意性框图；

[0017] 图7示出根据本发明的多机架系统中进行源地址认知的另一实施例的示意性框图；

[0018] 图8示出根据本发明的多机架系统中聚集交换机的另一实施例的示意性框图；

[0019] 图9示出根据本发明的多机架系统中活动/待机操作模式的实施例的示意性框图；

[0020] 图10示出根据本发明的多机架系统中活动/待机操作模式的另一实施例的示意性框图；

[0021] 图11示出根据本发明的当发生故障时多机架系统中的实施例的示意性框图；和

[0022] 图12示出根据本发明的多机架系统中分组的前挂报头的实施例的示意性框图

具体实施方式

[0023] 图1示出具有多机架链路聚集的弹性网络100的实施例，这提供了更完整地利用网络节点的容量的活动/活动范例(即，所有链路在相同时间主动转发业务)。这里定义以下缩写：

[0024] CMM 机架管理模块

[0025] IGMP 因特网群组管理协议

[0026] IP 因特网协议

[0027] IPMS 因特网协议组播

[0028] LAG 链路聚集

[0029] L2 网络的 OSI 模型的层 2（“数据链路层”）

[0030] L3 网络的 OSI 模型的层 3（“网络层”）

[0031] MAC 媒体访问控制协议

[0032] MC-LAG 多机架链路聚集群组

[0033] MC-VFA 多机架虚拟组织聚集

[0034] NIM 网络接口模块

[0035] STP 生成树协议

[0036] VLAN 虚拟局域网

[0037] VRRP 虚拟路由器冗余协议

[0038] ASIC 专用集成电路

[0039] 在本申请中引用以下标准，并这里通过引用合并于此：1）链路聚集控制协议（LACP），曾经是 IEEE802.3ad 任务组在 2000 年 3 月添加的 IEEE802.3 标准的 43 条，目前合并于 2008 年 11 月 3 日的 IEEE802.1AX-2008 中；和 2）IEEE Std.802.1Q，虚拟桥接局域网，2003 年版。

[0040] LACP 提供一种方法，用于控制两个端节点之间的若干物理链路（称为链路聚集群组（LAG））的绑定以在他们之间形成单逻辑信道。端节点通过交换 LACP 分组来协商物理链路到 LAG 的绑定，或者 LAG 可被手动配置。链路聚集提供了廉价的方式来传送比任一个单端口多的数据，或者链路可单独传递。实施例中，LAG 的端口包括相同物理类型，例如，全铜端口（CAT-5E/CAT-6）、全多模式光纤端口（SX）、或全单模式光纤端口（LX）。另一实施例中，LAG 的端口可具有不同物理类型。

[0041] 为了提供增加的弹性和去除单点故障，如图 1 所示，跨两个设备来划分 LAG，这里称为多机架链路聚集群组（MC-LAG）102。例如，图 1 中，MC-LAG102a 起源于边缘节点 104，并分成两个子集，并连接至两个聚集交换机 106a 和 106b，在每个子集中具有 MC-LAG102a 的一个或多个物理链路。实施例中，边缘节点 104 可使用负载均衡技术，跨 MC-LAG102a 的所有可用链路来分布业务。对于在 MC-LAG102a 上发送的每个分组，基于负载均衡算法（通常涉及在源和目的地因特网协议（IP）或媒体访问控制（MAC）低值信息上运行的散列函数）选择物理链路之一。跨 MC-LAG102 的物理链路的负载均衡导致带宽的更有效使用。

[0042] 如图 1 所示，边缘节点 104 在接入网络 122 上连接至企业网络设备 110，例如在 LAN 中运行的桥、交换机、路由器等，和 / 或他也可连接至家庭网络设备 112，例如 DSL 调制解调器、机顶盒、光线路终端等。边缘节点 104 是交换机或服务器，并且可功能地包括数字订户线路接入多路复用器（DSLAM）、电缆调制解调器终端系统（CMTS）、光线路端子（OLT）等，但是，在实施例中也可包括其他类型的设备。

[0043] 实施例中，聚集交换机 106 与虚拟组织链路（VFL）124 耦合。VFL124 提供用于聚集交换机之间的信息交换的连接，涉及业务转发、MAC 寻址、组播流、地址解析协议（ARP）表、层 2 控制协议（例如生成树（spanning tree）、以太网环保护、逻辑链路检测协议）、路由协议（例如 RIP、OSPF、BGP）和 MC-LAG102A 的状态。聚集交换机 106 对于边缘节点 104 透明地运

行，并被边缘节点看作单逻辑设备。实施例中，边缘节点 104 能够在 MC-LAG102a 上主动地转发业务，而聚集交换机 106 之间的 MAC 地址表和其他转发信息的同步则在 VFL 上随着减少数量的控制消息由 L2 分组流驱动。这个特征使得边缘节点 104 双导向(dual homing)至聚集交换机 106 的对，并提供层 2 多路径内结构以及基础层 3 接入基本结构。此外，实施例中，MC-VFA 特征提供这个功能，而在边缘节点 104 和聚集交换机 106 之间不需要层 2 冗余协议(例如生成树)，同时仍旧促进了运营级检测和对于边缘上行链路故障以及聚集 / 核心交换机故障的收敛时间。许多最近的网络设计(特别针对数据中心)在边缘节点和聚集交换机之间正需要不断增加数目的层 2 邻接。这个趋势推动了生成树协议的限制，例如，环检测功能和收敛时间。在许多当前网络拓扑中，生成树收敛时间可多达若干秒。实施例中，优选地，多机架架构在边缘节点 104 和聚集交换机 106 之间提供了双导向的、层 2 多路径连接，而不需要为了环预防而运行生成树协议操作，同时，在实施例中，在网络拓扑的一些部分中仍旧足以灵活地允许生成树协议操作跟踪多机架功能(例如，在虚拟组织链路上以及将这些设备连接至上流 / 核心交换机的链路上的聚集交换机之间)。

[0044] 一些实施例中，该特征还促进快速故障转移检测和对于接入上行链路故障、虚拟组织链路故障和节点故障的收敛时间。实施例中，MC-VFA 的另一优点是边缘节点 104 的活动 / 活动转发模式，从而可操作的 MC-LAG 上行链路的两个集正在处理业务以增加 MC-LAG 链路的带宽的使用效率。

[0045] 如图 1 所示，实施例中，聚集交换机 106 还使用这里所述的 MC-LAG 功能(作为 M-VFA 架构的一部分)连接至包括一个或多个网络节点 116(例如网络交换机和 / 或路由器)的城域或核心网络 120。例如，聚集交换机 106b 在 MC-LAG102b 上连接至网络节点 116b 和 116c，其中网络节点 116b 和 116c 也在 VFL 上交换状态信息。MC-LAG102b 架构在聚集交换机 106b 和网络节点 116b 和 116c 之间提供双导向的、层 2 多路径连接。实施例中，网络节点 116 也可使用 MC-LAG 功能来连接，如图所示用 MC-LAG102c 和 VFL124。聚集交换机 106 也可使用标准 LAG(例如 LAG118，或其他干线或链路) 连接至网络节点 116。

[0046] 现在，参照图 2 更详细地描述 MC-VFA 架构。边缘节点 104a 通过第一 MC-LAG1102a 连接至聚集交换机 106a 和 106b，而边缘节点 104b 通过第二 MC-LAG2102b 连接至聚集交换机 104a 和 104b。每个 MC-LAG102a 和 102b 包括分成至少两个子集的多个物理链路，其中两个子集的每个包括至少一个物理链路。如图 2 所示，MC-LAG102a 物理链路的第一集在第一聚集交换机 106a 处中止，而 MC-LAG102a 物理链路的第二集在第二聚集交换机 106b 处中止。MC-LAG1 形成逻辑双导向的、层 2 多路径。MC-LAG 成员端口是作为 MC-LAG102 的成员的外部的、用户端口。VFL124 是实施例中跨越多个网络接口模块以用于弹性的端口聚集，并提供机架间业务和控制 / 状态数据传送。多机架系统 140 包括聚集交换机 106、虚拟组织链路 124、MC-LAG102a、MC-LAG102b、以及他们各自附连至下行流边缘设备的 MC-LAG 成员端口。聚集交换机 106a 和 106b 是单独的物理交换机，其每个可操作为单机交换机，并且每个用其自身的单独物理机架来包装。聚集交换机 106a 和 106b 可以在相同地理区域中，例如中心局或数据中心中，或者可以是单独的地理位置，例如不同建筑物或城市，以提供地理多样性。

[0047] 作为附连至聚集交换机的 MC-LAG 客户端运行的边缘节点 104 可使用不同方法向他们的聚集中的链路分配业务，只要对于给定流来说，链路的选择保持固定。这确保在传送

端站点的任意对之间按序传送。实施例中，应该优选地配置相同数目的从边缘设备到每一个 MC-LAG 聚集交换机的上行链路端口。换句话说，如果在边缘交换机和 MC-LAG 聚集交换机之一之间配置两个上行链路，则也应该边缘交换机和其他多机架交换机之间配置两个上行链路。尽管不是强制的，这个布置为多机架交换机和边缘设备之间的流提供了更加异构的业务分布。

[0048] 现在参照图 3 更详细地描述聚集交换机 106 之间的虚拟组织链路（VFL）124。一个实施例中，每个聚集交换机 106 包括至少一个 CMM 模块 150a（主设备）和优选地第二 CMM 模块 150b（备用设备），以及多个网络接口模块（NIM）152，例如线路卡或端口模块。VFL124 是连接至第一和第二聚集交换机 106 中的一个或多个 NIM152 的 VFL 成员端口的聚集。例如，VFL124 包括聚集交换机 106a 的 NIM152a 和聚集交换机 106b 的 NIM152b 之间的物理链路的第一子集 A、以及聚集交换机 106a 和 106b 的 NIM152n 之间的物理链路的第二子集 B。实施例中，VFL 链路连接在位于聚集交换机 106 的 NIM152 中的交换 ASIC210 之间。每个 NIM152 还包括排队 ASIC212，如以下进一步描述。交换组织集成电路（IC）214 提供聚集交换机 106 中各个 NIM152 之间的互连。

[0049] 向多机架系统中的每个聚集交换机 106 分配唯一机架标识符。对于每个聚集交换机 106 的机架 ID 是唯一的和全局的，例如，每个聚集交换机知晓其对端聚集交换机的机架 ID。还生成每个聚集交换机中用于各个组件（例如 IC、NIM、CMM）的唯一硬件设备标识符（MID），用于管理本地和远程对象。实施例中，用于交换 ASIC210 的硬件设备标识符在多机架系统中具有全局意义，而对于其他组件（例如排队 ASIC212）的 MID 可仅具有本地意义。例如，向交换 ASIC210 分配的硬件设备标识符是由两个聚集交换机 106 知晓的，而对于其他设备的硬件设备标识符被限制为本地聚集交换机，并且对于远程聚集交换机没有意义。

[0050] 实施例中，在向其聚集交换机分配的范围内，对交换 ASIC210 分配全局唯一硬件设备标识符（MID），例如，

[0051] 聚集交换机 106a：机架 ID=1 和 MID 值 0-31

[0052] 聚集交换机 106b：机架 ID=2 和 MID 值 32-63

[0053] 图 3 示出向交换 ASIC210 分配的示例性 MID。通过知晓分配的范围，一个模块能够从其 MID 确定交换 ASIC 的位置，如在聚集交换机 106a 或聚集交换机 106b 中。

[0054] 实施例中，交换 ASIC210 在前挂（pre-pended）报头模式下运行，以在聚集交换机 106 之间交换数据和控制分组。图 4 更详细示出网络接口模块（NIM）152 的实施例的示意性框图。交换 ASIC210 包括多个外部端口接口 240，他们连接至外部节点，例如边缘节点 104a 和 104b。一个或多个外部端口接口 240 可包括对于 MC-LAG 物理链路、LAG 或其他干线群组、固定链路等的成员端口。外部端口 240 可具有相同物理接口类型，例如，铜端口（CAT-5E/CAT-6）、多模式光纤端口（SX）、或单模式光纤端口（LX）。另一实施例中，外部端口 240 可具有一个或多个不同物理接口类型。

[0055] 向外部端口 240 分配外部端口接口标识符（端口 ID），例如与交换 ASIC210 相关的设备端口值，如 gport 和 dport 值。实施例中，交换 ASIC210 的 MID 和交换 ASIC210 上外部端口 240 的外部端口接口标识符用来唯一地识别多机架系统中本地或远程聚集交换机上交换 ASIC210 的物理外部端口接口 240。另一实施例中，包括转换模块或其他实体的端口管理器可将交换 ASIC210 的 MID 和外部端口标识符转换成一个整数值，以生成全局端口值

(GPV),例如 MID4 ;设备端口标识符(dport)5 转换成 GPV20。任一个实例中,生成本地和远程聚集交换机两者中 NIM152 的外部端口的唯一外部端口标识符。唯一端口标识符也可分配给交换 ASIC210 的内部端口,例如,从交换 ASIC210 到 NIM152 上的处理模块的内部端口。通过端口标识符和交换 ASIC 的 MID 唯一地识别这些内部端口。

[0056] 交换 ASIC210 还包括分组管理单元(PMU)242,其确定入站分组的目的地址。可将分组交换给交换 ASIC210 的另一外部端口接口 240,给用于向本地或远程聚集交换机上的另一 NIM152 传输的排队 ASIC212,或者给用于向交换 ASIC210 外部或内部的 NIM152 的处理模块 266 传输的处理器接口(PI)244。

[0057] 当分组被发送至本地或远程聚集交换机上的另一 NIM152 时,实施例中,交换 ASIC210 将分组传送至前挂分组报头接口(PPHI),后者增加或修改分组报头以包括硬件设备信息(HDI)。HDI 包括与分组的来源和 / 或目的地相关的硬件设备的标识符。实施例中,前挂报头可包括其他信息,例如分组优先级和负载均衡标识符。为了获得目的地 HDI 信息,PPHI 执行对于 MAC/HDI 转发表 250 的查询处理。存储于地址表存储器 248 中的 MAC/HDI 转发表 250 包括 MAC 地址项目的列表,例如,对于外部设备、节点、模块、连接至聚集交换机 106 的软件或硬件的 MAC 地址。MAC 地址项目包括用在桥接或路由分组以到达具有相关 MAC 地址的设备中的相关硬件设备信息。目的地硬件设备信息包括例如,与目的地 MAC 地址相关的、本地或对端聚集交换机的、交换 ASIC210 的端口标识符和 MID (例如 MID=24,端口 ID=5 或 MID=54,设备端口 =12)。另一实施例中,目的地硬件设备信息可包括与目的地 MAC 地址相关的外部端口接口的全局端口值(GPV)。MAC/HDI 转发表 250 可包括一个或多个表,例如源干线图、干线位图表、干线群组表、VLAN 映射表等。实施例中,MAC/HDI 转发表 250 或其部分也可位于 NIM152 的排队 ASIC 中。

[0058] 实施例中,当交换 ASIC210 包括具有到远程聚集交换机的链路的活动 VFL 成员端口 252 时,MAC/HDI 转发表 250 可包括附加 HDI 信息,例如,将 gport 值关联至交换 ASIC MID 值和设备端口值的表和 / 或具有映射至外部端口接口的逻辑聚集群组标识符的表。

[0059] 实施例中,前挂报头包括与例如外部或内容端口接口的源端口相关的硬件设备信息 HDI,包括交换 ASIC 的硬件设备标识符 MID 和源端口的设备端口标识符。

[0060] 另一实施例中,前挂报头包括与连接至 VFL 端口 124 的交换 ASIC210 相关的 HDI(例如,对于图 3 的聚集交换机 106a,MID=0 或 MID=31)。然后,连接至 VFL 端口的交换 ASIC210 将在 VFL 上发送分组之前转化或转换前挂报头中的 HDI。

[0061] 实施例中,PPHI246 还附加与源端口(例如首先接收分组的外部端口接口 240)相关的源硬件设备信息。源硬件设备信息可包括交换 ASIC210 的 MID 和外部端口接口 240 的端口标识符(例如设备端口)和 / 或全局端口值(GPV)。实施例中,还向前挂报头增加额外信息,例如目的地硬件设备标识符或 MID、目的地设备端口、VLAN ID、分组类型(组播、单播、广播)、分组优先级和负载均衡标识符。实施例中,从例如 MAC/HDI 转发表 250 的地址表 248 提取目的地 HDI。

[0062] 然后,将具有前挂报头的分组发送至排队 ASIC212,用于在组织 IC214 上路由。排队 ASIC212 包括分组缓冲器 260、用于提供业务和缓冲器管理的队列管理 262、和全局 HDI 地址表 264。全局 HDI 地址表 264 将目的地 HDI 映射至一个或多个其他 NIM152 中排队 ASIC212 中的适当队列。例如,映射基于前挂报头中的硬件设备信息提供用于将分组交换到

聚集交换机 106 中其他排队 / 交换 ASIC 中一个或多个外部端口接口的适当出口队列的信息。另一实例中，当目的地 HDI 指示远程聚集交换机(即目的地设备标识符属于远程 / 对端交换机范围)上的目的地时，排队 ASIC212 将分组交换到聚集交换机 106 中一个或多个 VFL 端口接口的适当出口队列，用于在 VFL124 上传输至远程聚集交换机，例如，全局 HDI 地址表 264 指示相关的硬件设备位于远程聚集交换机上。这个情形下，基于在前挂报头中出现的并由交换 ASIC210 先前插入的负载均衡标识符做出与特定 VFL 端口接口对应的出口队列的确定。

[0063] 尽管将交换 ASIC210 和排队 ASIC212 示出为单独集成电路或模块，但是 ASIC 的一个或多个功能或组件可包括在其他 ASIC 上或组合在备用 ASIC 中或实现于一个或多个集成电路中。

[0064] 图 5 示出通过聚集交换机 106a 到 VFL124 的分组流的实施例的示意性框图。这个实例中，具有源 MAC 地址的设备 300 (例如企业设备 110 或家庭网络设备 112) 例如通过边缘节点 104 向具有设备的目的地 MAC 地址的聚集交换机 106a 发送分组，该地址可在远程聚集交换机 106b 的外部端口接口上被访问。在 NIM152n 中例如具有图 5 的 MID=31 的交换 ASIC210n 在例如具有端口 ID=2 的外部端口接口 240 上接收分组。交换 ASIC210n 提取目的地 MAC 地址，并执行地址表查询，以从 MAC/HDI 转发表 250 确定与目的地 MAC 地址相关的硬件设备信息(HDI)。目的地 HDI 可包括例如在到达具有该 MAC 地址的目的地设备(例如，本地聚集交换机 106a 或远程聚集交换机 106b 的 NIM152、排队 ASIC212、交换 ASIC210、外部端口标识符 240、VFL124 的成员端口) 的路径中一个或多个硬件组件的设备模块标识符(MID)。实施例中，目的地 HDI 可包括交换 ASIC210 的 MID 和处理对目的地设备的接入的外部端口接口 240 的端口标识符(例如设备端口)。此外，实施例中，前挂报头包括基于从原始分组提取的参数(源 MAC 地址、目的地 MAC 地址、源 IP 地址、目的地 IP 地址) 确定的分组优先级和负载均衡标识符。另一实例中，HDI 将包括用于外部端口接口 240 的全局端口值(GPV)或提供对目的地设备的接入的 NIM152 的 MID。另一实施例中，当目的地 MAC 地址关联于远程聚集交换机时，HDI 可包括连接至 VFL124 的 NIM152a 或交换 ASIC210 的硬件设备标识符 MID (例如 MID=0)。将目的地 HDI 增加至前挂报头，后者向原始分组报头增加信息(例如层 2，以太网分组报头类型)。交换 ASIC210n 也包括用于与发起外部端口节点相关的一个或多个设备的源硬件设备信息(HDI)，例如端口 ID=2。源 HDI 可包括一个或多个硬件设备标识符，例如发起交换 ASIC210 的 MID、源端口标识符(例如设备端口)、全局端口值、用于源 NIM152 的 MID、机架 ID 等。

[0065] 将具有前挂报头的分组发送至排队 ASIC212n，其随后基于目的地 HDI 确定本地聚集交换机上要发送分组的 NIM152。当目的地 HDI 指示聚集交换机 106a 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 ASIC212n 将分组放置于出口队列中，用于向本地外部端口接口的对应 NIM152 传输。图 5 所示的另一实例中，排队 ASIC212n 确定目的地 HDI 指示远程聚集交换机上的目的地硬件设备，例如，HDI 指示远程聚集交换机上具有 MID=45 的交换 ASIC。为了到达远程聚集交换机，需要在 VFL124 上发送分组。所以，排队 ASIC212n 将具有前挂报头的分组在组织 IC214 上从队列发送至与 VFL124 连接的 NIM152a。基于在前挂报头上承载的负载均衡标识符参数做出 VFL 成员端口的选择。在 NIM152a 上的排队 ASIC212a 接收具有前挂报头的分组，并将分组排队以用于在 VFL124 上传输。然后，交

换 ASIC210a 在 VFL124 上向远程聚集交换机发送具有包含源和 / 或目的地 HDI 的前挂报头的分组。

[0066] 实施例中，交换 ASIC210a 可在 VFL124 上传输之前改变前挂报头。例如，交换 ASIC210a 可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 MID）转换成具有全局意义的 HDI。然后，交换 ASIC210a 在 VFL124 上向远程聚集交换机发送具有包括源和 / 或目的地 HDI 的前挂报头的分组。

[0067] 实施例中，当聚集交换机 106 的多个交换 ASIC210 连接至 VFL124 时，例如图 3 中，交换 ASIC MID=0 和 MID=31，可分布要在 VFL124 上发送的业务。例如，在排队 ASIC212 的全局 HDI 地址表 264 中负载均衡标识符映射表将指示以下分布 ：

[0068]

目的地 MID	出站端口	MID 的设备位置
[0-31]	VFL124	本地
[32-63]	VFL124	远程

[0069] 排队 ASIC212 使用负载均衡标识符或其他负载均衡技术将分组映射至适当 VFL 端口接口。例如，实施例中，在每个聚集交换机上有 8 个 NIM152，每个排队 ASIC212n 具有对于本地聚集交换机中的每个 NIM（模块 ID，端口）配置的 8 个队列的集。实施例中，连接至具有 VFL124 的交换 ASIC210 的排队 ASIC212 具有与每个 VFL 成员端口接口相关的 8 个队列的单独集。将这些队列的每个分配给与连接多机架交换机的内部 VFL 端口相关的 FIFO。实施例中，通过多个虚拟组织链路成员端口，分配队列，使得在托管虚拟组织链路成员端口的排队 ASIC212n 和 212n 之间均等地分布远程机架上的目的地端口。

[0070] 实施例中，将 NIM152 中的 MAC/HDI 转发表进行填充，并随后响应于通过系统的层 2 分组流进行更新。由于前挂报头包括源 MAC 地址和源 HDI 信息，所以实施例中，例如特定的交换 ASIC210 中的 NIMS152 能够用这个信息填充 MAC/HDI 转发表 250。通过在前挂报头模式下操作以在 VFL124 上交换具有源 MAC 地址和源 HDI 的层 2 分组，交换 ASIC210 能够在聚集交换机 106 之间同步 MAC 地址表。尽管描述了在交换 ASIC210 中的 MAC/HDI 转发表，但是，MAC/HDI 转发表可备选地或额外地包含于排队 ASIC212n 或 NIM152 的其他模块中。另一实施例中，CMM150（主设备和次设备）也可包括用于聚集交换机 106 之间的一个或多个类型的链路的 MAC/HDI 转发表。

[0071] 图 6 示出多机架系统的实施例的示意性框图，其展示了源 MAC 认知。每个节点 104 在逻辑聚集群组 LAG1282、多机架逻辑聚集群组 MC-LAG1102a、多机架逻辑聚集群组 MC-LAG2102b 和固定端口链路 280 上连接至聚集交换机 106a 和 106b。实施例中，每个聚集交换机向其他聚集交换机传送对于逻辑聚集群组（例如 LAG1 和其他类型的干线群组）的配置信息，以及与其相关的硬件设备信息。实施例中，硬件设备信息包括与逻辑聚集群组相关的物理端口，例如交换 ASIC 的硬件设备或模块标识符（MID）和与逻辑聚集群组相关的链路的外部端口标识符（设备端口值或 gport 值）。

[0072] 例如，实施例中，聚集交换机 A 通知聚集交换机 B 具有聚集群组标识符 LAG1 的逻辑聚集群组关联于具有硬件设备模块标识符 MID=31 的交换 ASIC 和具有标识符设备端口

=1,2 的外部端口接口。聚集交换机 B 通知聚集交换机 A 具有聚集群组标识符 MC-LAG1 的逻辑聚集群组关联于具有硬件设备模块标识符 MID=45 的交换 ASIC 和具有标识符设备端口=1,2 的外部端口接口。对于交换 ASIC 的 MID 和设备端口值备选地或额外地，可交换与逻辑聚集群组相关的其他硬件设备信息，例如 NIM 的标识符、排队 ASIC 等。聚集交换机 106 为普通聚集和多机架聚集群组提供逻辑聚集群组的配置信息的更新的通知。与任一个聚集交换机的逻辑聚集群组和多机架聚集相关的硬件设备信息包括在两个聚集交换机上的 NIM152 中的一个或多个 MAC/HDI 转发表中。例如，实施例中，两个聚集交换机 106 中的一个或多个 MAC/HDI 转发表包括以下信息：

[0073]

聚集群组的类型	聚集群组标识符	VFL 成员端口的 HDI 列表
LAG	LAG1	(MID=31,端口 ID=1) (MID=31,端口 ID=2)
MC-LAG	MC-LAG1	(MID=31,端口 ID=3) (MID=31,端口 ID=4) (MID=45,端口 ID=1) (MID=45,端口 ID=2)
MC-LAG	MC-LAG2	(MID=31,端口 ID=5)

[0074]

		(MID=45,端口 ID=3)

[0075] 由于逻辑聚集群组(例如 LAG1)的相同聚集群组标识符已知，并由两个聚集交换机 106 利用，所以实施例中，多机架系统将聚集群组标识符的子集分配给每个类型的逻辑群组，并用于每个聚集交换机 106。例如，具有最大 128 个可能聚集群组的实施例中，聚集群组标识符的分配将包括：

[0076]

聚集群组的类型	聚集交换机	范围配置	范围	
			缺省	实例
LAG	机架 1	MIN_LAG_ID_LOCAL MAX_LAG_ID_LOCAL	[0-47]	[0-100]
LAG	机架 2	MIN_LAG_ID_REMOTE MAX_LAG_ID_REMOTE	[48]-[95]	[101-120]
MC-LAG	两个机架	MIN_MC-LAG_ID MAX_MC-LAG_ID	[96-127]	[121-127]

[0077] 聚集交换机 106 基于分配的范围和聚集群组的类型分配聚集群组标识符。由此，通过访问 MAC/HDI 转发表和使用逻辑聚集群组与硬件设备信息之间的映射来执行聚集交换机中的分组转发。典型地，不在前挂报头中传送聚集标识符信息。

[0078] 实施例中，为了促进 LAG 或 MC-LAG 上的负载均衡，当聚集交换机 106 在 VFL124 上接收具有目的地 HDI 信息的（例如 MID，端口 ID）的分组时，聚集交换机 106 通过在包含作为每个 LAG 或 MC-LAG 聚集群组的活动成员的所有端口列表的一个或多个其内部干线表中搜索源 HDI（目的地 MID，目的地端口标识符）识别的端口，来确定目的地 HDI 是否包括在逻辑聚集群组中。当在关联 LAG 或 MC-LAG 中找到目的地端口时，聚集交换机 106 可通过向关联 LAG 的一个或多个不同外部端口接口分配分组来执行负载均衡技术。例如，当连接至远程聚集交换机 106b 中的 VFL 的交换 ASIC210 接收具有 MID=45, 端口 =2 的目的地 HDI 的分组时，交换 ASIC210 从其以下的 MAC/HDI 表确定 MID=45, 端口 =2 是 MC-LAG1 的一部分，如图 6 的实例所示。然后，交换 ASIC 决定执行负载均衡，并通过一个或多个散列算法确定代替地在 MC-LAG1 的 MID=45, 端口 =1 上发送分组。这个特定实例中，随后，交换 ASIC 在从外部端口（MID=45, 端口 =1）发送分组之前去掉前挂报头。

[0079]

聚集交换机 A	
LAG ID	HDI
LAG1	(MID=31,端口 ID=1) (MID=31,端口 ID=2)
MC-LAG1	(MID=31,端口 ID=3) (MID=31,端口 ID=4) (MID=45,端口 ID=1) (MID=45,端口 ID=2)
MC-LAG-2	(MID=31,端口 ID=5) (MID=45,端口 ID=3)

[0080] 再参照图 6，现在描述方法和实现方式的各个实施例，用于认知多机架系统中的源 MAC 地址和相关硬件设备信息（HDI）。首先，实施例中，对于进入聚集交换机之一的配置固定端口上的未知单播分组（例如源自具有源 MAC 地址 =d1 的固定端口 280 的业务），将源 MAC 地址填充至聚集交换机 106a 和 106b 两者上的 MAC/HDI 转发表中，与发起的配置固定端口的硬件设备信息（HDI）关联（例如，交换 ASIC 的 MID 和源端口标识符值或源端口的 gport 值，NIM 标识符，或与源端口相关的其他硬件设备 ID）。由此，实施例中，源 MAC 地址 d1 存储于聚集交换机 A 和聚集交换机 B 两者的一个或多个 MAC/HDI 转发表中，其中 VLAN ID 和 HDI 关联于源端口，例如 MID=45, 端口 ID=4。

[0081] 接着，实施例中，对于进入仅与一个聚集交换机 106 连接的逻辑聚集群组的未知单播业务，例如干线群组或其他类型的 LAG（例如，源自具有源 MAC 地址 =a1 的 LAG1 的业务），将源 MAC 地址填充至聚集交换机 106a 和 106b 两者的 MAC/HDI 转发表中，与发起逻辑聚集群组标识符（例如 LAG1）关联。由此，实施例中，由聚集交换机 A 在 LAG1 上接收的源 MAC 地址 a1 存储于具有 VLAN ID 和逻辑聚集群组标识符 LAG1 的两个聚集交换机 106 的一个或多个 MAC/HDI 转发表中。此外，如这里所述，两个聚集交换机的 MAC/HDI 转发表存储与逻辑聚集群组关联的硬件设备信息（由 CMM150 模块或其他控制面处理通过配置信息的分布认知）。因此，MAC/HDI 转发表包括将 MAC 地址 a1 关联于与 LAG1 相关的干线群组标识符 LAG1 和 HDI 信息的信息。

[0082] 此外，实施例中，对于进入任一个聚集交换机 106 的 MAC-LAG 成员端口上的未知单播业务（例如源自 MC-LAG1 或 MC-LAG2 的业务），将源 MAC 地址填充至 MAC/HDI 转发表中，与 MC-LAG 标识符和 MC-LAG 的本地成员端口的 HDI 信息关联。对于每个聚集交换机 106 上的 MAC/LAG 表来说，MC-LAG 的成员端口的 HDI 信息将是相同的。换句话说，两个聚集交换机完全知晓作为 MC-LAG 聚集群组的活动参与者的成员端口的整个列表，不管成员端口是本地还是远程。

[0083] 通过将 MC-LAG 的成员端口关联于源 MAC 地址，优选地，通过最短路径，经过 MC-LAG

成员端口来转发通过边缘节点 104 之一指向 MAC 地址的业务。这个路径减少了穿越 VFL124 的业务的量。此外，一些特定情形下，他减少了 MAC 移动问题，其中到达和来自边缘节点 104 的业务对于不同的流在 MC-LAG 上采用了不同路径。图 6 的实例中，实施例中，聚集交换机 106 上的一个或多个 MAC/HDI 转发表包括以下信息：

[0084]

聚集交换机 A		
MAC	LAG	LAG ID
a1	是	LAG1
b1	是	MC-LAG1
c1	是	MC-LAG-2
d1	否	-

[0085]

聚集交换机 B		
MAC	LAG	LAG ID
a1	是	LAG1
b1	是	MC-LAG1
c1	是	MC-LAG-2
d1	否	-

[0086]

[0087] 另一实施例中，在节点或网络管理应用中显示的 MAC 地址表可不包括对于逻辑聚集群组的 HDI。用户显示的 MAC 地址表可仅包括对于固定端口的 HDI，因此对于两个聚集交换机 106 类似。

[0088]

聚集交换机 A			
MAC	LAG	LAG ID	HDI
a1	是	LAG1	N/A
b1	是	MC-LAG1	N/A
c1	是	MC-LAG-2	N/A
d1	否	-	(MID=45,端口 ID=4)

[0089]

聚集交换机 B			
MAC	LAG	LAG ID	HDI
a1	是	LAG1	N/A
b1	是	MC-LAG1	N/A
c1	是	MC-LAG-2	N/A
d1	否	-	(MID=45,端口 ID=4)

[0090] 关于与源 MAC 地址相关的 LAG 标识符来同步 MAC/HDI 转发表。此外，与 MAC 地址相关的 VLAN ID 也可被配置，并在两个聚集交换机上被同步。由此，逻辑上，聚集交换机 106 操作为单独桥，用于 MAC 认知。此外，在业务在 VFL124 上流动时，自动发生 MAC 认知，其具有最小层 2/ 控制模块管理软件干预，并且不需要基于处理间通信消息的 MAC 表同步。

[0091] 图 7 更详细示出一种在多机架系统中用于源 MAC 认知的方法的实施例。为了确定对于设备 B 的 MAC 地址，设备 A300a（具有 MAC 地址 =MAC_A）发送 MAC 地址请求（例如，在以太网协议中使用的地址解析分组（ARP）），其具有对于设备 B300b 的目标 IP 地址。例如，MAC 地址请求可包括：

[0092] 源 MAC=MAC_A

[0093] 目的地 MAC=ff:ff:ff:ff:ff:ff（未知）

[0094] 目标 IP=IP_B

[0095] VLAN ID=ID

[0096] 分组类型＝广播（Broadcast）

[0097] 当由边缘节点 104a 接收时，他在 MC-LAG A 上将 MAC 地址请求转发至“逻辑”聚集交换机 106（包括两个物理交换机 106a 和 106b）。依据负载均衡或散列算法，边缘节点 104a 可在 MAC-LAG A 的任一个子集，L_{A1}或 L_{A2}上发送 MAC 地址请求。对于这个实例，假设在连接至聚集交换机 106a 的 L_{A1}上发送 MAC 地址请求。一般地，在以太网交换机中，复制并在与 VLAN ID 相关的每个端口上广播 MAC 地址请求（例如 ARP）。实施例中，当聚集交换机 106a 接收 MAC 地址请求时，他首先将前挂报头附加至具有源逻辑聚集群组标识符（例如 MC-LAG A）和 / 或源 HDI（例如 (MID=12, 端口 ID=1) 的 MAC 地址请求。然后，聚集交换机（例如特定交换 ASIC MID=12）将具有前挂报头的分组的副本广播至具有与 VLANID 相关的外部端口的每个交换 ASIC，这个实例中例如交换 ASICMID=31。然后，在接收具有前挂报头的 MAC 地址请求的聚集交换机 106a（例如 MID=12，MID=31）上的交换 ASIC 认知源 MAC 地址和相关的聚集群组标识符（或者明显地存在于前挂报头中或通过在其干线表中搜索源 HDI 信息，其包含这里所述的 MC-LAG A 的成员端口的完整列表，例如 MID=12, 端口 ID=1,2，和 MID=45, 端口 ID=1,2），并且能够用聚集群组标识符信息填充他们的 MAC/HDI 转发表。例如，依据特定实施例，交换 ASICMID=31 进入源 MAC 地址 MAC_A关联于逻辑聚集群组 MC-LAG A 和 / 或关联于源端口 MID=12, 端口 ID=1 的 HDI 的其 MAC/HDI 转发表中。在从外部端口接口向边缘节点 B 发送 MAC 地址请求之前，聚集交换机 106a（例如具有 MID=31 的交换 ASIC）移除前挂报头，因此保留以太网或 IP 协议报头。

[0098] 聚集交换机 106a 还在 VFL124 上向聚集交换机 106b 发送具有前挂报头的广播分组。聚集交换机 106b 也从具有前挂报头的广播分组认知源 MAC 地址和相关聚集群组标识符和 / 或源 HDI。如上所述，在具有相同 MC-LAG 的对端聚集交换机中关联源自一个聚集交换机中 MC-LAG 本地成员端口并在 VFL 上发送的 MAC 地址，因为两个交换机完全知晓 MC-LAG 成员端口的整个列表。由此，当聚集交换机 106b 接收具有前挂报头的分组时，他存储与源 MAC 地址 MAC_A 相关的 MC-LAG A 的聚集群组标识符。例如，具有 MID=45 的交换 ASIC（和 / 或具有 MID=63 的交换 ASIC）进入源 MAC 地址 MAC_A 与逻辑聚集群组 MC-LAG A 相关的其 MAC/HDI 转发表。

[0099] 尽管通常在与 VLAN ID 相关的每个端口上广播 MAC 地址请求，但是实施例中，循环防止机制防止由聚集交换机 106 接收的分组在本地 MC-LAG 成员端口上的虚拟组织链路 124 上广播。因此，当聚集交换机 106b 在 VFL124 上接收 MAC 地址请求时，他不在本地 MC-LAG A 成员端口 L_{A2} 和本地 MC-LAG B 成员端口 L_{B2} 上广播 MAC 地址请求的副本。这个循环防止机制防止源自聚集交换机 A 的广播分组流通过聚集交换机 B 循环至边缘节点 A 和边缘节点 B。因此，循环防止处理提供了多机架域系统的操作，而在 MC-LAG 成员端口上不需要生成树协议。

[0100] 聚集交换机 106a 和 106b 不生成 MAC 地址请求的响应，因为目的地 IP 地址不对应于在其本地 VLAN 上配置的其本地 IP 接口的任一个。然而，当边缘节点 B 接收 MAC 地址请求（在 L_{B1} 上）时，他将分组广播至设备 B，后者随后将响应。在作为单播分组的响应分组遍历多机架系统到达设备 A 时，在类似的处理中由聚集交换机 106 认知设备 B 的源 MAC 地址。现在，设备 A 和设备 B 能够在多机架链路聚集提供的层 2 多路径架构中与 IP 寻址通信。将 MAC 地址识别为或者与特定端口关联（对于固定端口的情况），或者与聚集群组标识符关联（对于 LAG 或 MC-LAG 的情况）。由于聚集交换机 106 不具有硬件设备标识符 MID 的重叠范围，所以硬件设备标识符在多机架系统 140 中是唯一的。使用全局唯一硬件设备标识符 MID 和外部端口标识符，MAC 地址可关联于固定端口或聚集群组标识符。

[0101] 图 8 示出在多机架系统中维护 MAC/HDI 转发表的实施例的示意性框图。MAC 转发表具有对于项目的缺省的或配置的“老化”时间。当在老化时间期间没有更新 MAC/HDI 转发表中的 MAC 地址时，该项目将从表删除或刷新。然而，在多机架系统中，当分组流对于上游和下游方向具有不同路径时，项目的老化可产生连续溢出的问题。为了维护同步的 MAC 转发表，多机架系统需要跨作为系统一部分的交换机的整个集实现保持实时（keep-alive）机制。保持实时分组是周期性的分组（以等于老化超时参数的常量间隔发送）。这些分组携带保留的组播目的地 MAC 地址，允许分组溢出到多机架系统中所有 NIM152 中的所有交换 ASIC 设备 210。分组的源 MAC 地址等于 MAC 转发表中认知的每个项目的 MAC 地址。作为这个机制的结果，给定 MAC 地址将不老化和删除或刷新，除非他在多机架系统中的聚集交换机的任一个中不再被使用。

[0102] 为了避免持久的 MAC 地址（例如不老化从而不被刷新或删除的地址），向 MAC 项目分配多机架系统中的“所有者”或负责模块。通常，MAC 项目的所有者是特定 NIM152。用各种方式确定 MAC 所有者身份。例如，MAC 所有者身份可取决于首先在上面如下认知的端口的类型。对于与固定端口相关的 MAC 地址，包含接收 MAC 地址业务的外部端口的交换 ASIC 设备 210 是 MAC 项目的所有者，并控制 MAC 地址的老化。其他交换 ASIC210 在接收具有前挂

报头的分组时认知这个 MAC 地址。托管这个交换 ASIC 设备 210 的 NIM152 将不会成为 MAC 项目的所有者。一个设备仅当认知了来自外部端口接口的地址时，成为与固定端口相关的 MAC 项目的所有者。

[0103] 对于在聚集端口(即 LAG 或 MC-LAG)上认知的 MAC 地址，通过对于固定端口所述的类似机制来确定 MAC 地址的所有者。这里的区别在于交换 ASIC210 典型地提供称为远程或本地位的附加特征。仅当项目被建立并且他在 MAC 项目的寿命期间不再改变他的值时，设置这个位。仅当以下情况时，设置本地位(即本地＝ 1 或远程＝ 0)，所述情况包括 :a) 该项目还不存在 ;b) 在前面板端口上接收分组，例如，不存在前导报头。作为这个方法的结果，在设置本地位的系统中始终存在单交换 ASIC 设备 210。托管该交换 ASIC 设备 210 的那个 NIM152 成为这个 MAC 地址的所有者，因此负责保持实时分组。

[0104] NIM152 协调从 MAC/HDI 转发表删除项目。如图 8 所示，在聚集交换机 106 的 CMM150a 和 150b 之间建立逻辑处理间通信连接(IPC) 310。在 NIM152 的任意对之间存在相同逻辑连接。可在 VFL124 上或在 LAN 连接上建立 IPC310。当本地聚集交换机的 NIM152 之一接收 MAC 地址的刷新消息时，他可决定向本地和远程聚集交换机 106a/b 上的其他 NIM152a-n 的每个发送刷新消息。然后，在 NIM152a-n 中交换和 / 或排队 ASIC 中的 MAC/HDI 表刷新对于对应 MAC 地址的项目。是否本地删除项目的决定取决于项目的所有者和认知 MAC 项目的端口的类型。刷新在固定或普通聚集(即 LAG) 上认知的项目(和传播对应的事件)，只要在拥有项目的 NIM152 上接收刷新请求。如果在本地交换机中和远程交换机上都不存在作为聚集的成员的活动 / 可操作端口，仅刷新在 MC-LAG 聚集上认知的项目(和传播刷新事件)。

[0105] CMM150a-b 和 NIM152a-n 知晓多机架系统中 MC-LAG 成员端口和他们的状态(活动 / 失活)的整个列表。当刷新消息包括仅在本地聚集交换机上有效的本地端口标识符(例如 gport 值)时，拥有被删除的该 MAC 地址的 NIM152 将本地端口标识符转换成全局端口标识符(例如 MID 或 modid 和设备端口值)，然后在 IPC 上将刷新消息发送至本地和远程聚集交换机 106a/b 的其他 NIM152a-n。可通过例如端口接口状态通知的不同事件(例如端口关闭)或经由明显的管理请求来触发刷新请求。例如，当 CMM150a 接收“非 mac 认知动态”管理消息或在用户请求时删除静态 MAC，并且满足先前描述的刷新许可需求时，CMM150a 将具有 MAC 地址的刷新消息发送至聚集交换机 106a 的 NIM150a-n 和聚集交换机 106b 的 CMM150b。

[0106] 实施例中，多机架系统中的 MC-LAG 的一个或多个链路可在待机模式下运行。图 9 示出 MC-LAG1102a 的链路 L_{A2} 的第一子集和 MC-LAG2102b 的链路 L_{B1} 的第二子集在待机模式下运行的实施例。除非发生故障，在待机模式下在本地 MC-LAG 成员端口上不发送业务。相反，业务在 VFL 上发送至对端聚集交换机，用于在活动模式下在对应的 MC-LAG 成员端口上传输。例如，当在活动 / 待机模式下运行时，当在聚集交换机 A 处接收具有未知目的地的单播分组时，他不在待机链路 L_{B1} 上广播至边缘节点 C。相反，在 VFL124 上将具有 HDI (例如 MID=31, 端口 ID=1) 的前挂报头发送至聚集交换机 B。通常，在上述活动 / 活动模式下，为了循环防止目的，聚集交换机 B 不在 MC-LAG 本地成员端口上发送在 VFL124 上接收的分组。然而，在活动 / 待机模式下，聚集交换机 B 向在活动模式下的 MC-LAG 本地成员端口发送广播分组。聚集交换机 B 去除前挂报头，并保持内部标准以太网 /IP 分组报头，并在 MC-LAG2 的 L_{B1} 上将分组广播至边缘节点 C。

[0107] 此外，如图 10 所示，指向具有目的地 MAC 地址的边缘节点 104b 的业务在 VFL124 上发送，而并非在处于待机模式下的 MC-LAG 本地成员端口 L_{A2}上发送。活动 / 待机模式可为了维护的原因而配置或实施。由于故障，也可实现 MC-LAG 的本地成员端口的待机模式。多机架系统响应于通过 VFL 重定向业务和 / 或经过实现 MAC 刷新消息以从 MAC/HDI 转发表删除 MAC 表项目的集建立新路径的故障，在收敛上提供快速失败。然后，使用源 MAC 认知经由这里所述的分组流通过多机架系统自动认知路径。

[0108] 图 11 示出当 MC-LAG 链路中发生故障时多机架系统的实施例的示意性框图。实例中，如果由于 NIM 故障、端口故障、链路故障等使得活动 MC-LAG1 成员端口故障，则在聚集交换机 106a 和 106b 上而不在网络节点 116a 和 116b 上，刷新与 MAC/HDI 转发表中的 MC-LAG1 的端口相关的 MAC 地址项目。例如，从聚集交换机 106a 和 106b 中 MAC/HDI 转发表刷新与 MC-LAG1 的端口相关的 MAC 地址，例如 MID=31，端口 ID=3 或 4（例如 MAC 地址＝ b1）。如果在待机模式下，使得 MC-LAG1 成员端口 L_{A2}处于活动模式，然后，在层 2 分组流过多机架系统时，在 MAC/HDI 转发表中重新认知 MAC 地址，与新活动 MC-LAG 本地成员端口关联，例如 MID=45，端口 ID=1 或 2。网络节点 A 和 B 不需要关于 MAC 地址 =b1 刷新其 MAC 项目，但是如之前的故障继续转发至聚集交换机。

[0109] 例如，在操作中，当聚集交换机 B 从网络节点 B 接收具有目的地 MAC 地址 =b1 的分组时，响应于刷新消息，MAC 地址 =b1 已经从其 MAC/HDI 转发表删除。聚集交换机 B 将 MAC 地址看作未知，并广播 MAC 地址请求。聚集交换机 B 将前挂报头增加至分组，并在 VFL 上将其发送至聚集交换机 A，及其其他交换 ASIC MID=45。当交换 ASIC MID=45 接收广播分组时，他将移除前挂报头，并随后在当前活动 MC-LAG1 成员链路 LA2 上广播 MAC 地址请求。当边缘节点 B 接收广播分组时，他也将广播分组，从其具有 MAC 地址 =b1 的附加设备认知 MAC 地址，并向聚集交换机 B 发送应答分组。因此，聚集交换机 B 将认知与 MAC 地址 b1 相关的源 HDI 是 MID=45, 端口 =1，并用更新的项目填充其 MAC/HDI 转发表。由于聚集交换机 A 的 MC-LAG1 成员端口不再可操作，则聚集交换机 B 将 MC-LAG1 看作逻辑聚集群组（LAG），从而现在来自 VFL124 的广播分组溢出至 MC-LAG1 成员端口。

[0110] 因此，在层 2 分组的具有新路径信息的故障流过聚集交换机 106 之后重新填充 MAC/HDI 转发表，而并不使用生成树协议或层 2MAC 表控制消息生成的同步消息。

[0111] 图 12 示出在多机架系统中分组的前挂报头的实施例的示意性框图。前挂报头 300 包括以下字段 ：源 HDI302、目的地 HDI304、VLAN ID306、分组类型 308、源 MAC 地址 310、目的地 MAC 地址 312。实施例中，前挂报头也可包括负载均衡标识符 314 和分组优先级 316。目的地 HDI304 包括例如与目的地 MAC 地址相关的、本地或对端聚集交换机的交换机 ASIC210（例如 MID=24，端口 ID=5 或 MID=54，设备端口 =12）的端口标识符和 MID。另一实施例中，目的地硬件设备信息可包括与目的地 MAC 地址相关的外部端口接口的全局端口值（GPV）。目的地硬件设备信息也可包括与 VFL 连接的交换 ASIC210、NIM152、排队 ASIC 等的 MID。源 HDI302 可包括交换 ASIC210 的 MID 和外部端口接口 240 的端口标识符（例如设备端口）和 / 或全局端口值（GPV）。负载均衡标识符 314 用来帮助排队 ASIC212 决定哪个 VFL 成员端口用作到达对端聚集交换机的中转 / 网关端口。排队 ASIC212 使用分组优先级 316 确定特定优先级队列。

[0112] 网络接口模块 152 包括一个或多个处理设备，例如微处理器、微控制器、数字信号

处理器、微计算机、中央处理单元、场可编程门阵列、可编程逻辑设备、状态机、逻辑电路、模拟电路、数字电路、和/或基于电路的硬编码和/或可操作指令操作信号(模拟和/或数字)的任意设备。NIM152 包括作为内部存储器或外部存储器的存储器。NIM152 的存储器可以是单存储器设备或多个存储器设备。这样的存储器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、高速缓存、和/或存储数字信息的任意设备。NIM152 可经由状态机、模拟电路、数字电路、和/或逻辑电路实现其功能的一个或多个,存储对应可操作指令的存储器可嵌入于包括状态机、模拟电路、数字电路、和/或逻辑电路中,或在其外部。NIM152 可执行由内部存储器和/或外部存储器存储的硬编码的和/或软件和/或可操作的指令,以执行这里所述的步骤和/或功能。NIM152 可实现于单个或一个或多个集成电路中。

[0113] 如这里使用的,术语“基本”和“大概”提供了对于其对应术语和/或项目之间的相对性的行业可接受的容限度。这样的行业可接受的容限度的范围从小于百分之一到百分之五十,并且不限制地对应于组件值、集成电路处理变形、温度变形、上升和下降时间、和/或热噪音。项目之间的这样的相对性的范围从百分之一的差异到大量差异。同样,如这里使用的,术语“耦合至”和/或“耦合”包括在项目之间直接耦合和/或经由中间项目在项目之间间接耦合(例如项目不限制地包括组件、元件、电路、和/或模块),其中,对于间接耦合,中间项目不修改信号的信息,但是可调整其电流电平、电压电平、和/或功率电平。同样,如这里使用的,暗示耦合(即一个元件通过暗示地耦合至另一元件)包括与“耦合至”相同的方式在两个项目之间直接和间接耦合。如这里使用的,术语“能够操作为”指示项目包括处理模块、数据、输入、输出等的一个或多个,以执行所述或必要的对应功能的一个或多个,并且还可包括暗示耦合至一个或多个其他项目以执行所述或必要的对应功能。如这里使用的,术语“连接至”和/或“连接”或“互连”包括在节点/设备之间的直接连接或链路和/或经由中间项目在节点/设备之间的间接连接(例如项目不限制地包括组件、元件、电路、和/或模块)。如这里使用的,暗示连接(即一个元件通过暗示地耦合至另一元件)包括与“连接至”相同的方式在两个项目之间直接和间接耦合。

[0114] 以上借助于显示特定功能的执行及其关系的方法步骤描述了实施例。这里为了便于描述,任意定义了这些功能建立方框和方法步骤的边界和顺序。也可定义备选边界和顺序,只要适当执行特定功能和关系即可。因此,任意这样的备选边界或顺序在主张的本发明的范围和精神内。类似地,这里也任意定义了流程图方框以示出某些明显功能。对此,流程图方框的边界和顺序可能已经被定义,并且仍旧执行某些明显功能。因此,功能建立方框和流程图方框和顺序两者的这样的备选定义在主张的本发明的范围和精神内。本领域技术人员之一也将认识到这里的功能建立方框、和其他图示的方框、模块和组件可被如图所示实现,或通过一个或多个离散元件、网络、系统、数据库、或执行适当软件等的处理模块或其任意组合。

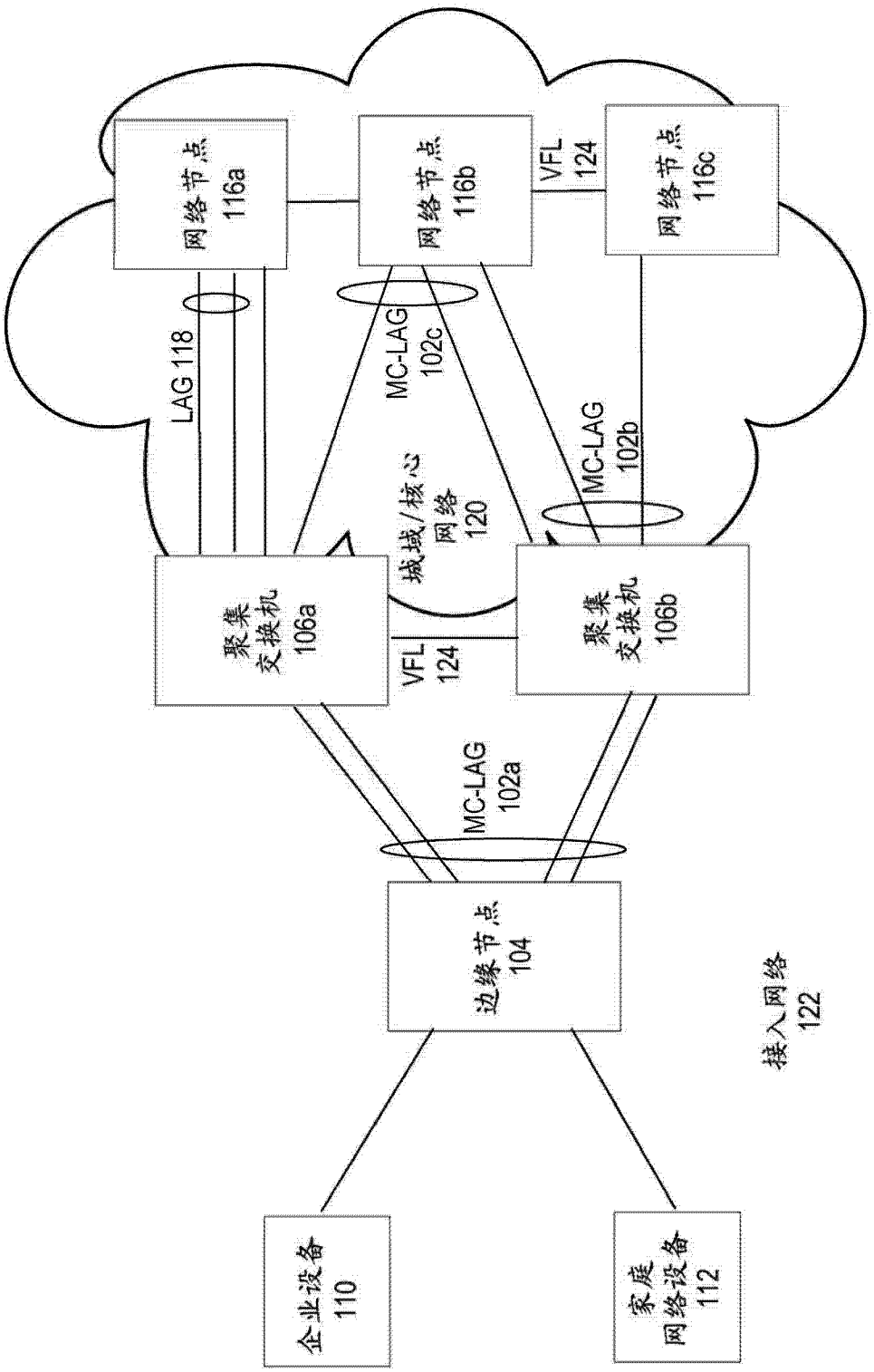

网络节点
116a
网络节点
116b
VFL
124
网络节点
116c
LAG 118
MC-LAG
102c
MC-LAG
102b
城域/核心
网络
120
聚集
交换机
106a
VFL
124
聚集
交换机
106b
MC-LAG
102a
边缘节点
104
接入网络
122
企业设备
110
家庭
网络设备
112

图 1

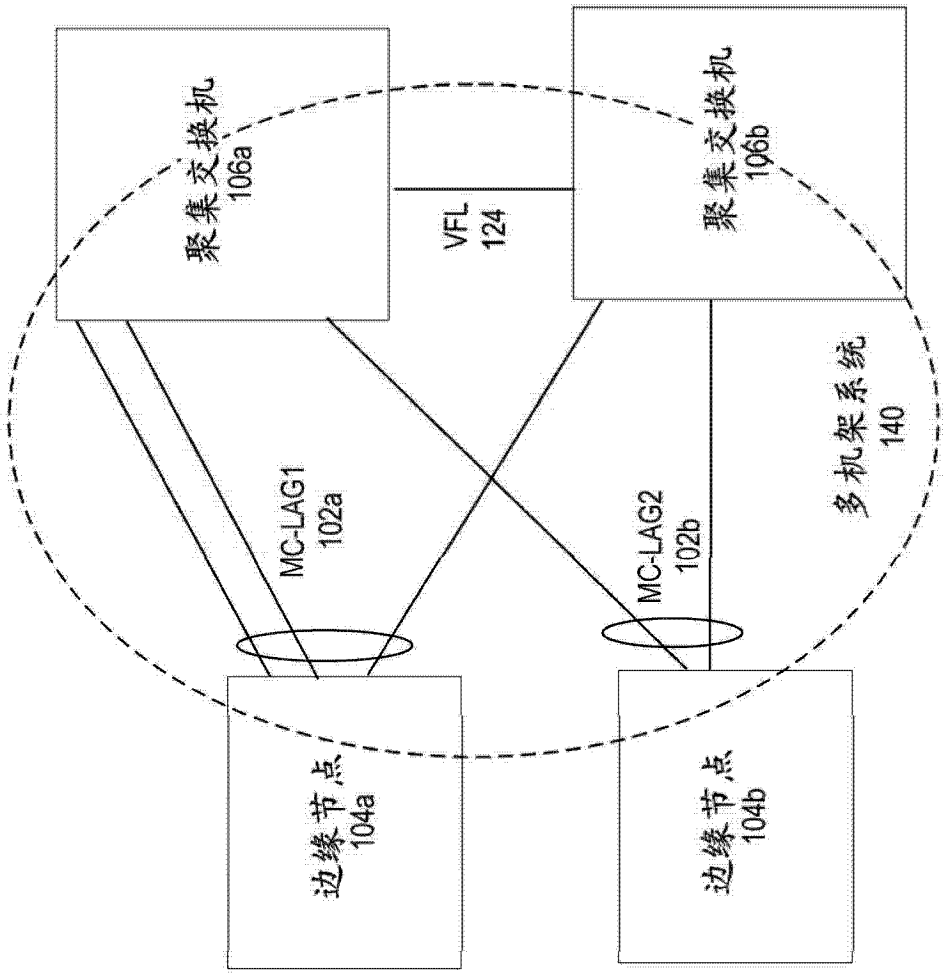

聚集交换机
106a
VFL
124
聚集交换机
106b
MC-LAG1
102a
MC-LAG2
102b
多机架系统
140
边缘节点
104a
边缘节点
104b
VFL/MC-LAG的概况

图 2

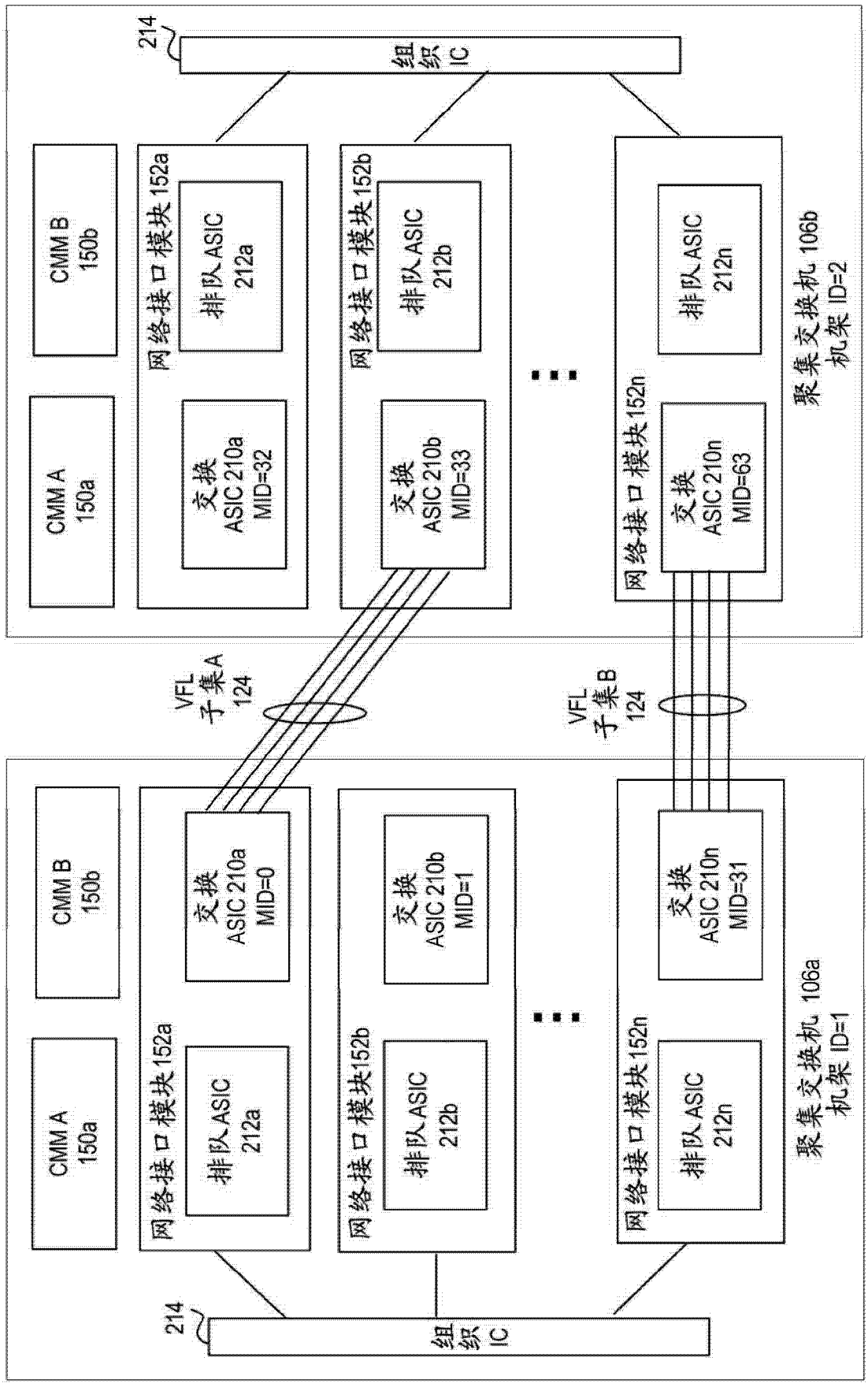

214
组织 IC
CMM B 150b
CMM A 150a
网络接口模块 152a
排队 ASIC 212a
交换 ASIC 210a MID=32
网络接口模块 152b
排队 ASIC 212b
交换 ASIC 210b MID=33
网络接口模块 152n
排队 ASIC 212n
交换 ASIC 210n MID=63
聚集交换机 106b 机架 ID=2
VFL 子集 A 124
VFL 子集 B 124
交换 ASIC 210a MID=0
交换 ASIC 210b MID=1
交换 ASIC 210n MID=31
聚集交换机 106a 机架 ID=1

图 3

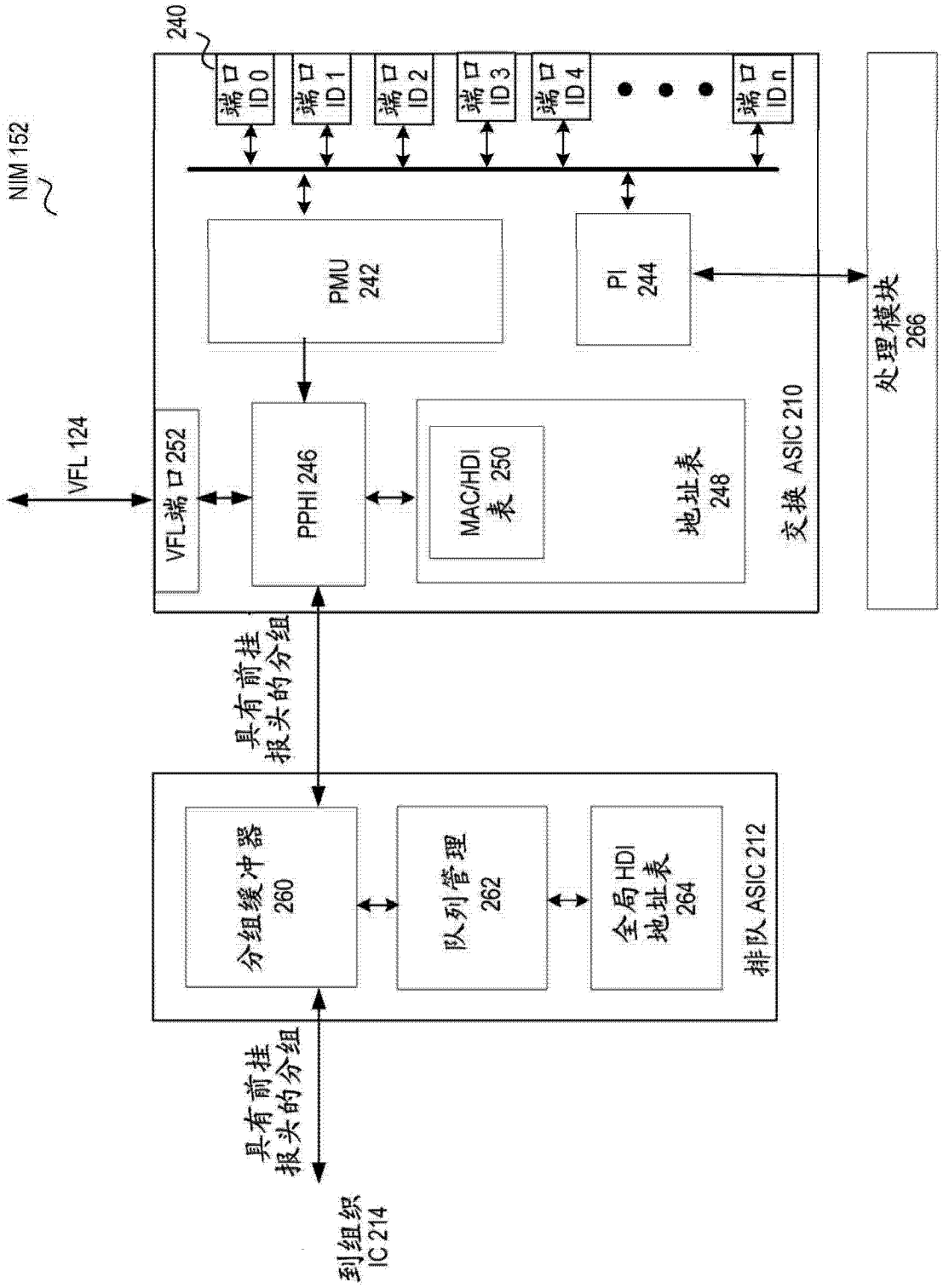

240
端口 ID 0
端口 ID 1
端口 ID 2
端口 ID 3
端口 ID 4
端口 ID n
NIM 152
PMU 242
PI 244
处理模块 266
VFL 124
VFL端口 252
PPHI 246
MAC/HDI 表 250
地址表 248
交换 ASIC 210
具有前挂报头的分组
分组缓冲器 260
队列管理 262
全局 HDI 地址表 264
排队 ASIC 212
具有前挂报头的分组
到组织 IC 214

图 4

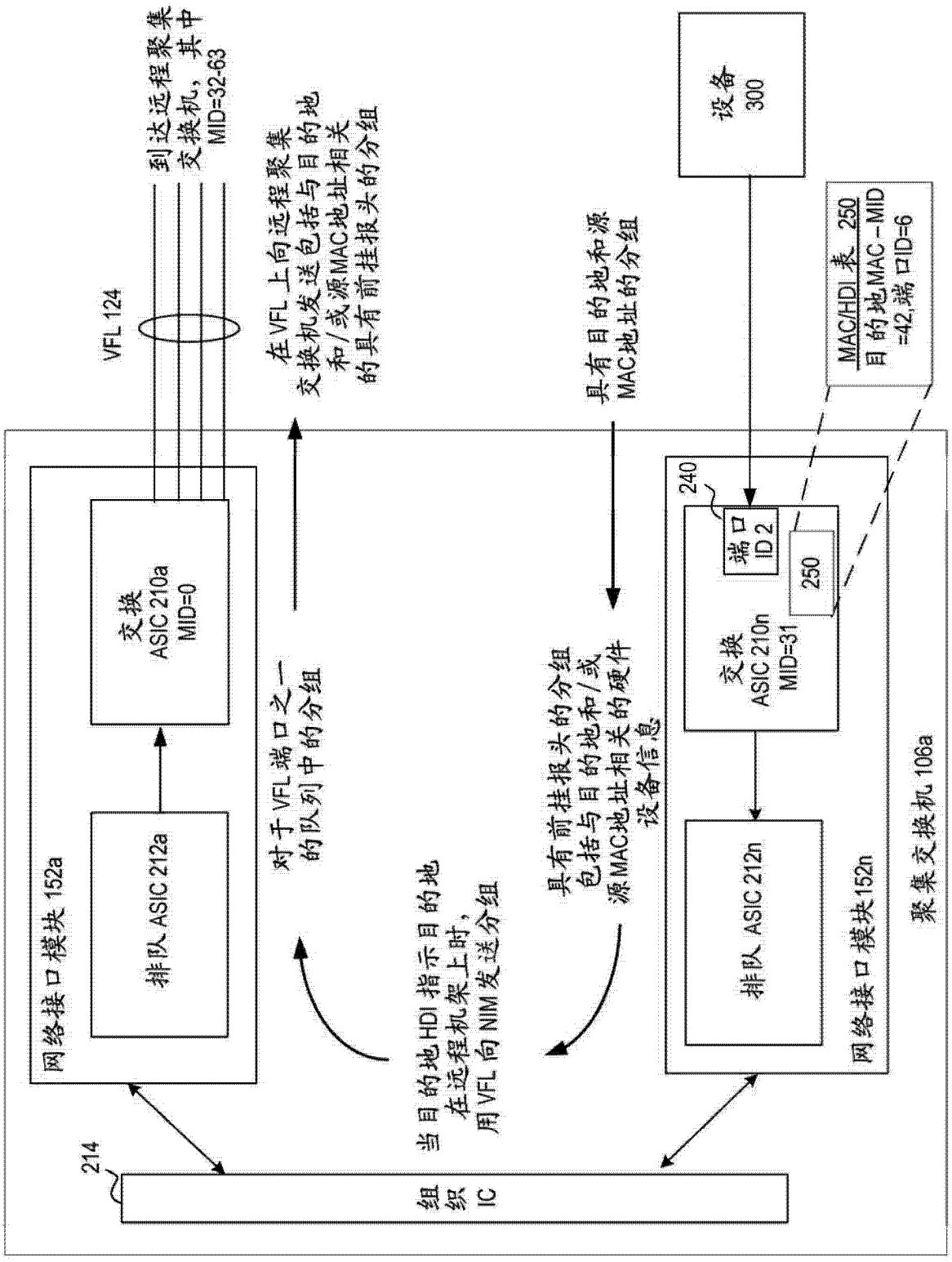

到达远程聚集交换机，其中MID=32-63
VFL 124
在VFL上向远程聚集交换机发送包括与目的地和/或源MAC地址相关的具有前挂报头的分组
具有目的地和源MAC地址的分组
设备 300
MAC/HDI 表 250
目的地MAC－MID =42,端口ID=6
网络接口模块 152a
交换 ASIC 210a MID=0
排队 ASIC 212a
对于VFL端口之一的队列中的分组
当目的地HDI指示目的地在远程机架上时，用VFL向NIM发送分组
具有前挂报头的分组包括与目的地和/或源MAC地址相关的硬件设备信息
240
端口 ID 2
250
交换 ASIC 210n MID=31
排队 ASIC 212n
网络接口模块152n
聚集交换机 106a
214
组织 IC

图 5

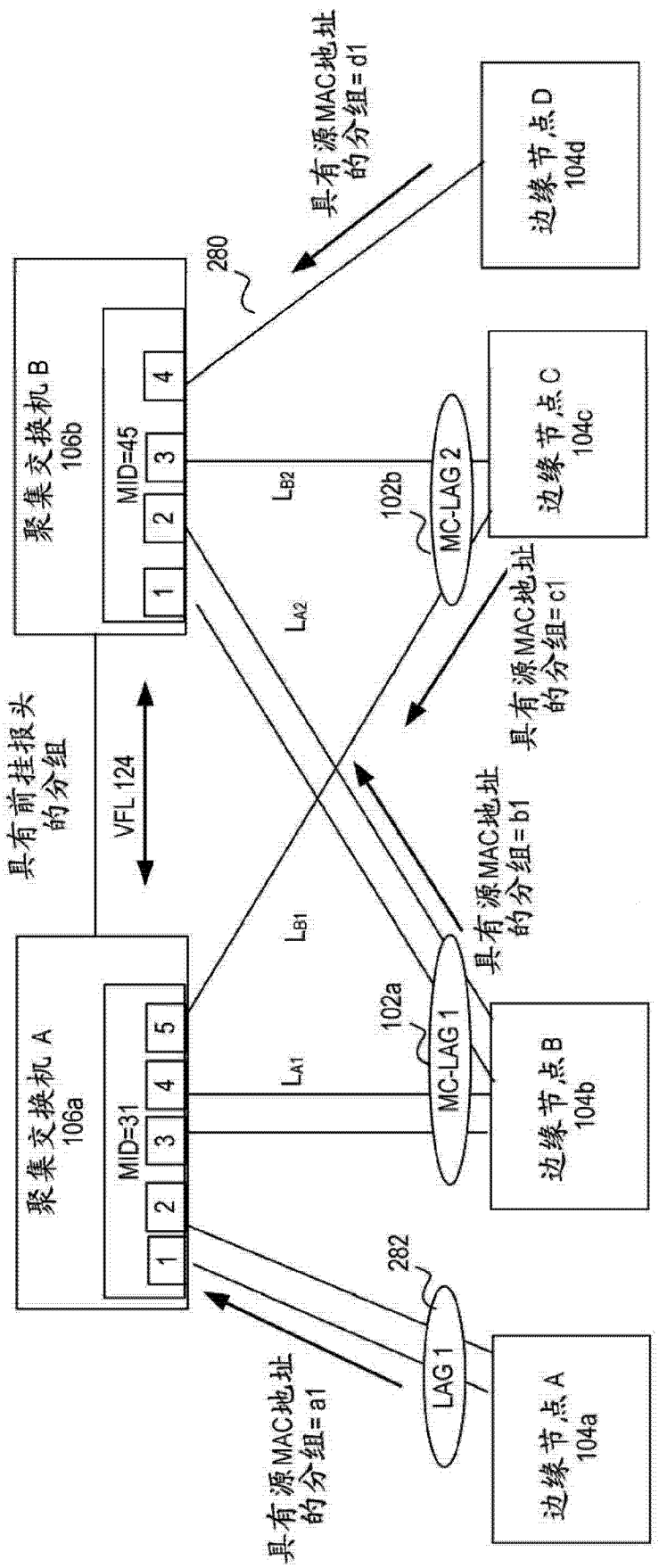

具有源MAC地址的分组= d1
280
聚集交换机 B
106b
MID=45
1
2
3
4
边缘节点 D
104d
边缘节点 C
104c
LB2
102b
MC-LAG 2
LA2
具有源MAC地址的分组= c1
具有前挂报头的分组
VFL 124
具有源MAC地址的分组= b1
LB1
聚集交换机 A
106a
MID=31
1
2
3
4
5
LA1
102a
MC-LAG 1
边缘节点 B
104b
282
LAG 1
具有源MAC地址的分组= a1
边缘节点 A
104a

图 6

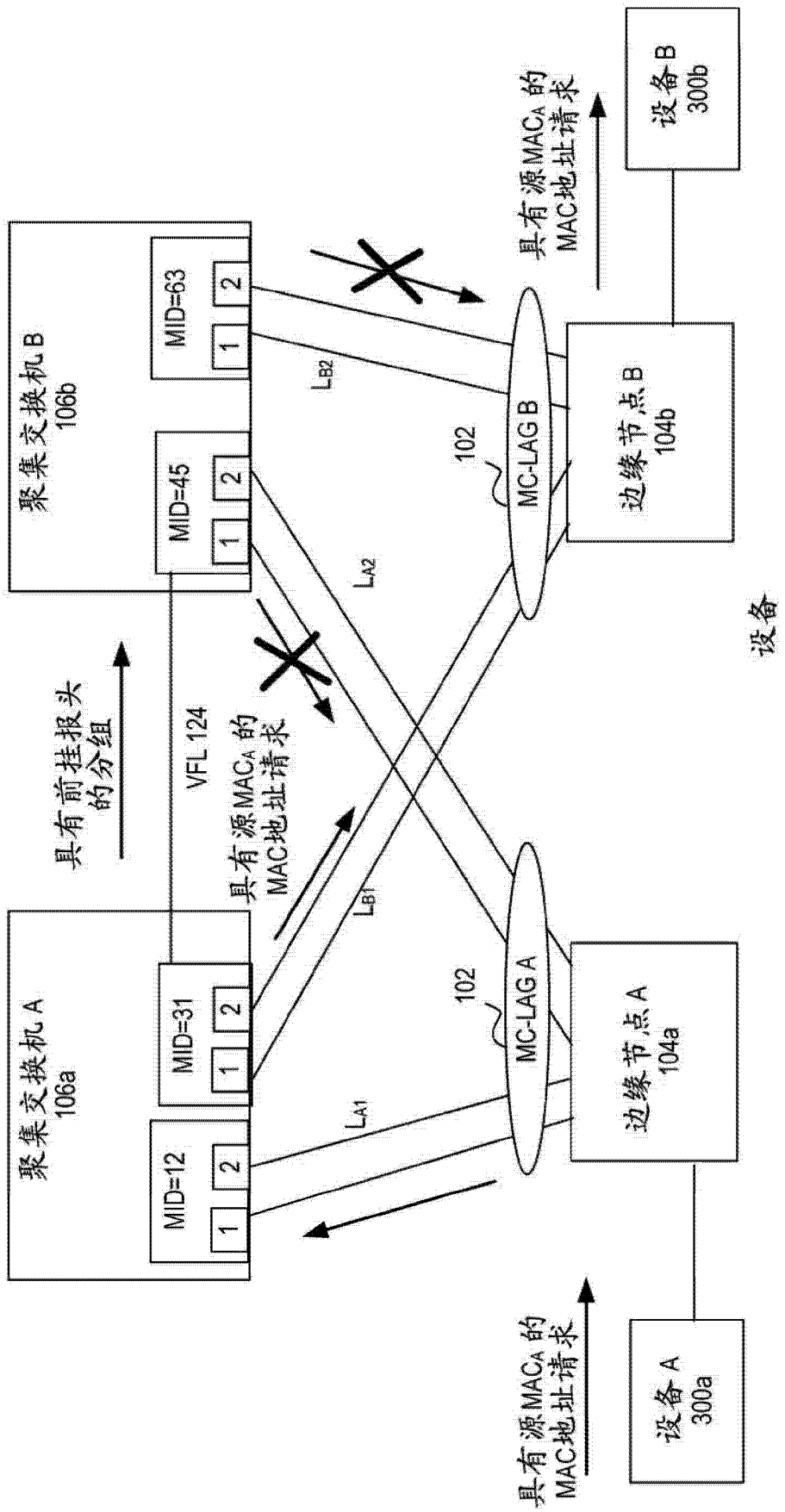

聚集交换机 A
106a
聚集交换机 B
106b
MID=12
MID=31
MID=45
MID=63
VFL 124
具有前挂报头的分组
具有源 MACA 的 MAC地址请求
具有源 MACA 的 MAC地址请求
具有源 MACA 的 MAC地址请求
LA1
LA2
LB1
LB2
MC-LAG A
MC-LAG B
102
102
边缘节点 A
104a
边缘节点 B
104b
设备 A
300a
设备 B
300b
设备

图 7

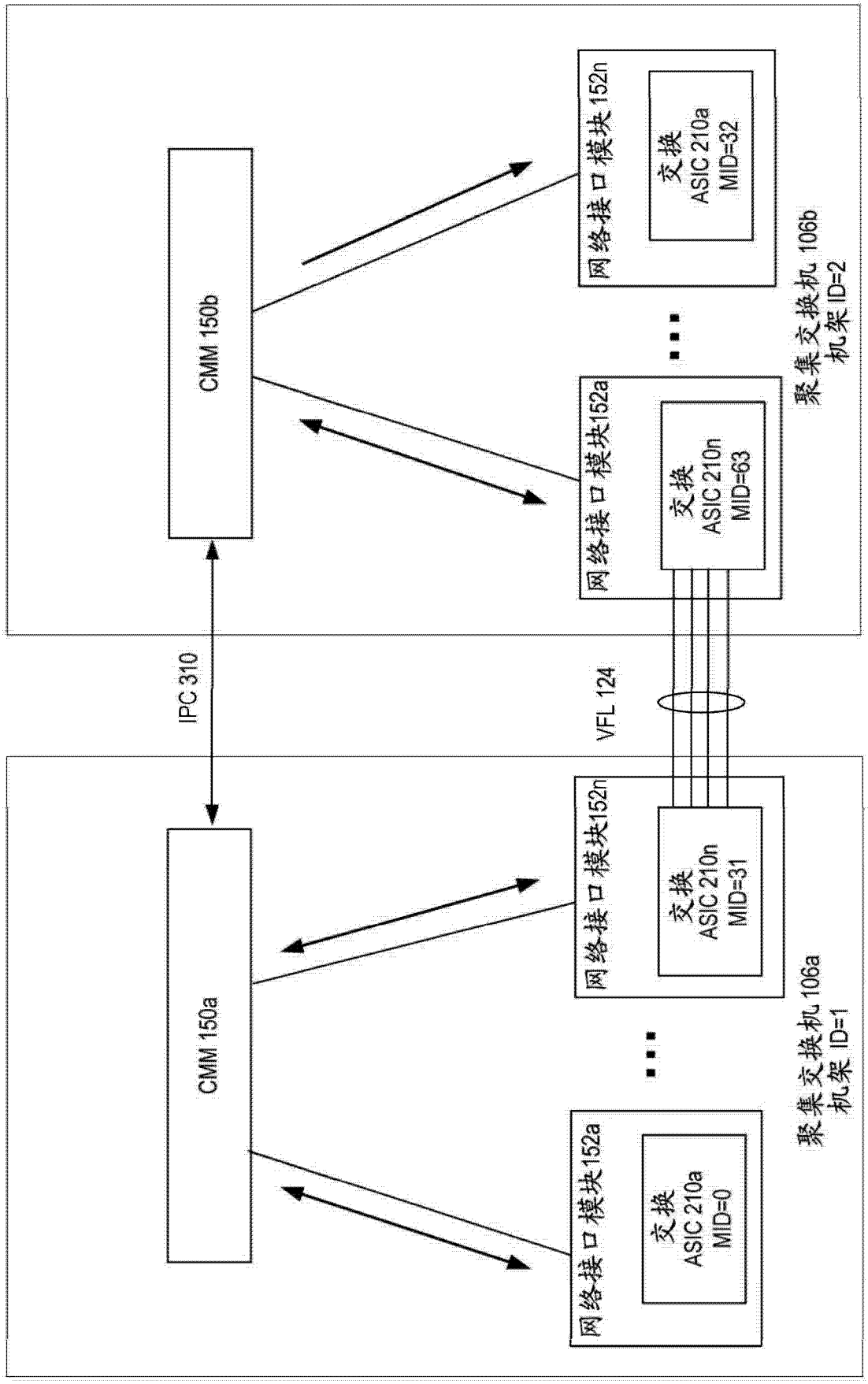

聚集交换机 106b
机架 ID=2
网络接口模块 152n
交换
ASIC 210a
MID=32
网络接口模块152a
交换
ASIC 210n
MID=63
CMM 150b
IPC 310
VFL 124
CMM 150a
网络接口模块152n
交换
ASIC 210n
MID=31
网络接口模块152a
交换
ASIC 210a
MID=0
聚集交换机 106a
机架 ID=1

图 8

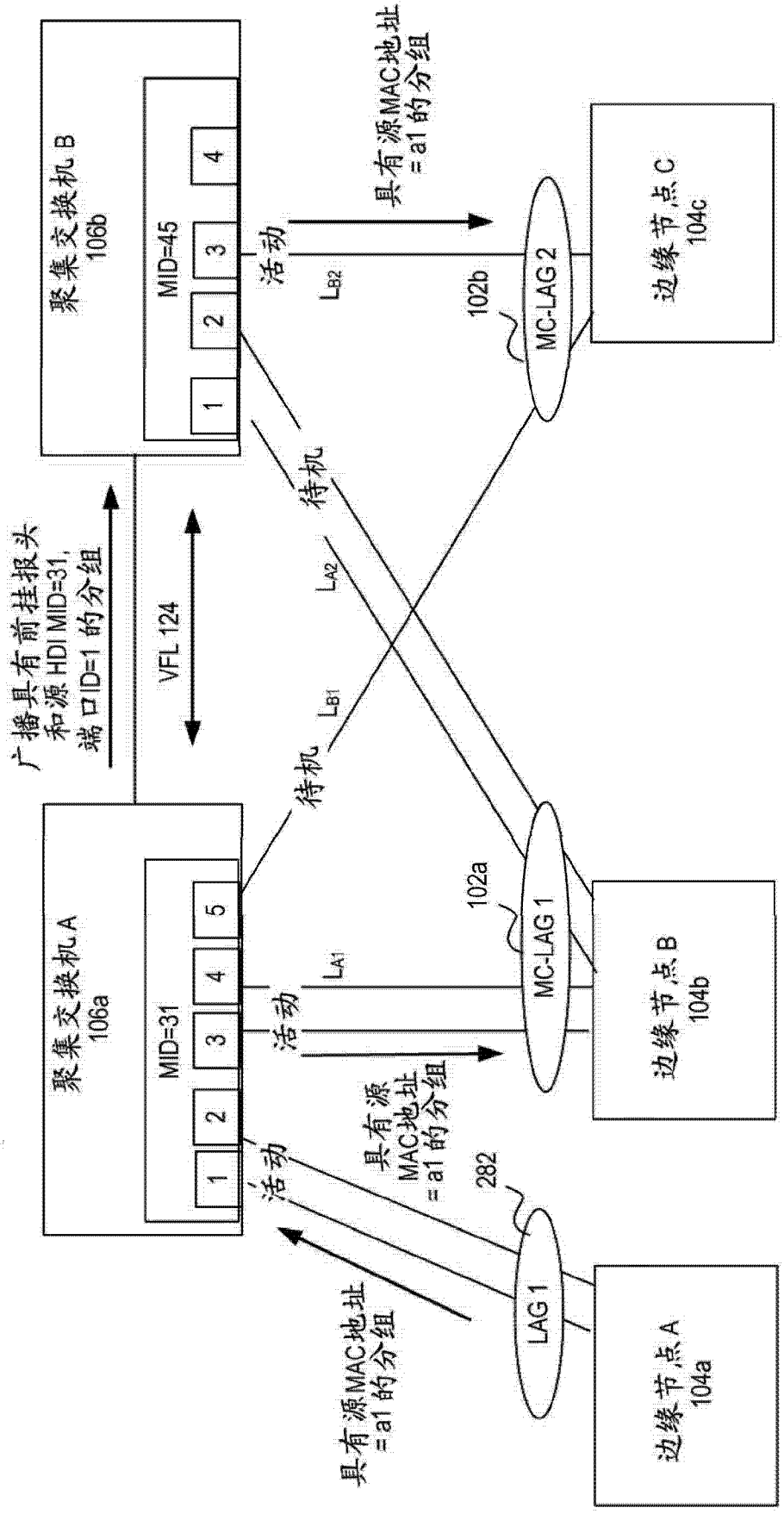

聚集交换机 A
106a
MID=31
1
2
3
4
5
聚集交换机 B
106b
MID=45
1
2
3
4
广播具有前挂报头
和源 HDI MID=31,
端口 ID=1 的分组
VFL 124
活动
待机
L_{A1}
L_{A2}
L_{B1}
L_{B2}
具有源MAC地址
=a1 的分组
具有源
MAC地址
=a1 的分组
具有源MAC地址
=a1 的分组
102a
102b
282
MC-LAG 1
MC-LAG 2
LAG 1
边缘节点 A
104a
边缘节点 B
104b
边缘节点 C
104c

图 9

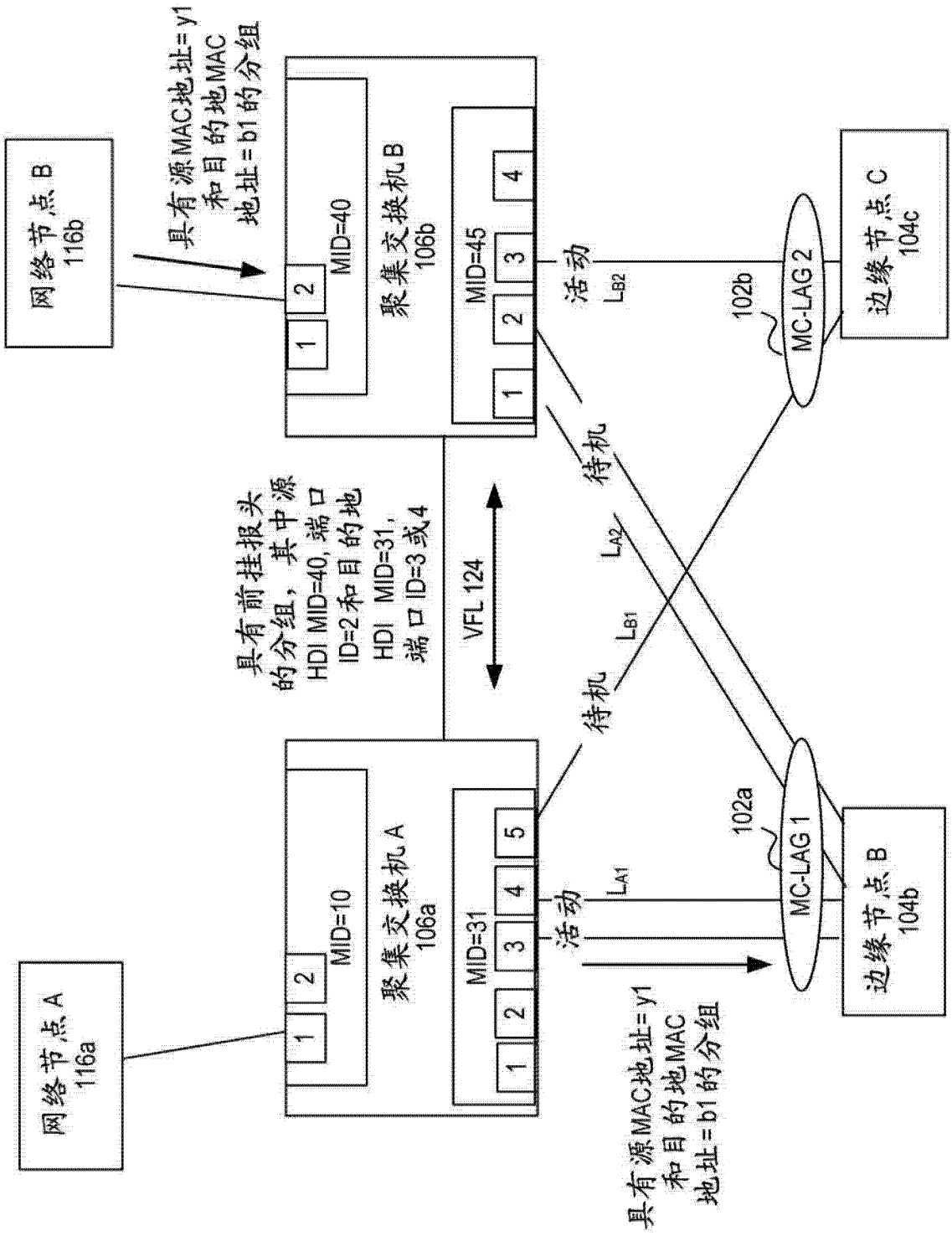

网络节点 A
116a
网络节点 B
116b
具有源MAC地址= y1
和目的地MAC
地址= b1的分组
聚集交换机 A
106a
聚集交换机 B
106b
MID=10
MID=31
MID=40
MID=45
具有前挂报头
的分组，其中源
HDI MID=40, 端口
ID=2和目的地
HDI MID=31,
端口ID=3或4
VFL 124
活动
待机
L_{A1}
L_{A2}
L_{B1}
L_{B2}
MC-LAG 1
MC-LAG 2
102a
102b
边缘节点 B
104b
边缘节点 C
104c
具有源MAC地址= y1
和目的地MAC
地址= b1的分组

图 10

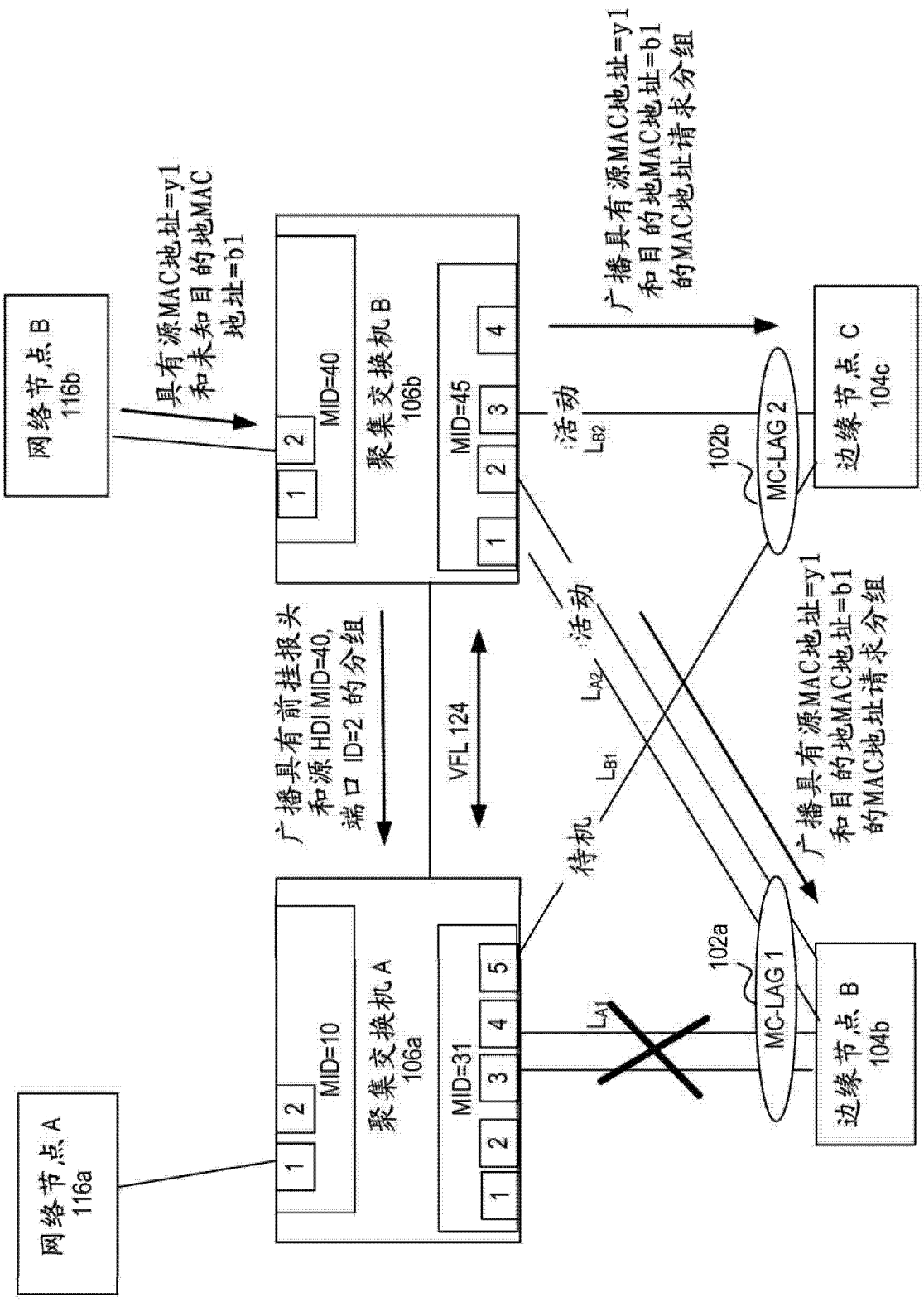

网络节点 B
116b
具有源MAC地址=y1
和未知目的地MAC
地址=b1
聚集交换机 B
106b
MID=40
MID=45
广播具有源MAC地址=y1
和目的地MAC地址=b1
的MAC地址请求分组
活动
L_{B2}
102b
MC-LAG 2
边缘节点 C
104c
广播具有前挂报头
和源 HDI MID=40,
端口 ID=2 的分组
VFL 124
活动
L_{A2}
L_{B1}
待机
广播具有源MAC地址=y1
和目的地MAC地址=b1
的MAC地址请求分组
网络节点 A
116a
聚集交换机 A
106a
MID=10
MID=31
L_{A1}
102a
MC-LAG 1
边缘节点 B
104b

图 11

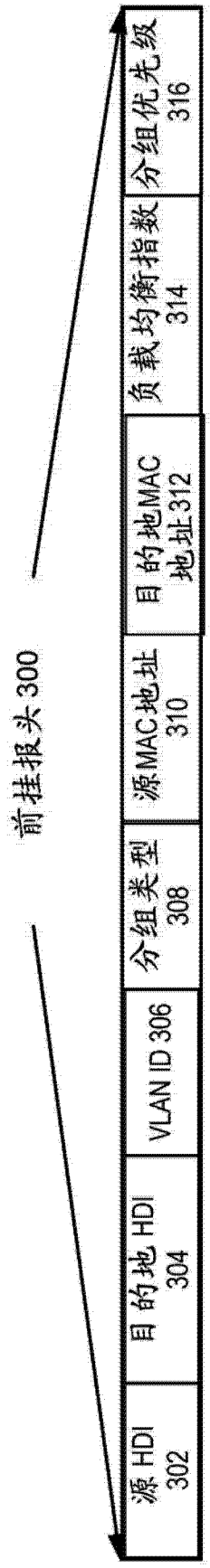

前挂报头 300
源 HDI 302
目的地 HDI 304
VLAN ID 306
分组类型 308
源MAC地址 310
目的地MAC 地址312
负载均衡指数 314
分组优先级 316

图 12