(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 102986176 A
(43)申请公布日 2013.03.20

(21)申请号 201180024665.0

(22)申请日 2011.05.19

(30)优先权数据
61/346,434 2010.05.19 US
13/110,309 2011.05.18 US

(85)PCT申请进入国家阶段日
2012.11.19

(86)PCT申请的申请数据
PCT/US2011/037099 2011.05.19

(87)PCT申请的公布数据
WO2011/146686 EN 2011.11.24

(71)申请人 阿尔卡特朗讯公司
地址 法国巴黎

(72)发明人 F·巴鲁斯 W·亨德里克斯

(74)专利代理机构 北京市中咨律师事务所 11247
代理人 张潇 杨晓光

(51)Int.Cl.
H04L 12/723(2013.01)
H04L 12/46(2006.01)
H04L 12/761(2013.01)
H04L 29/12(2006.01)

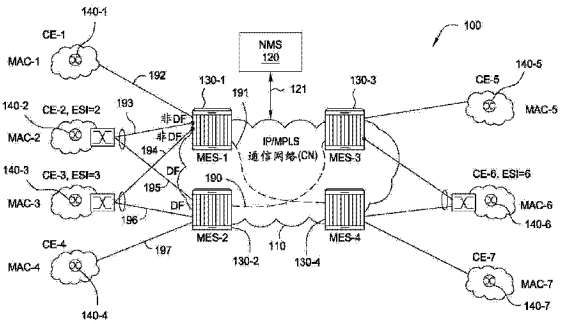
权利要求书 2 页 说明书 10 页 附图 7 页

(54)发明名称

用于BGP MAC-VPN的MPLS标签分配的方法和装置

(57)摘要

本发明包括一种用于在支持边界网关协议(BGP)媒体接入控制(MAC)虚拟专用网络(VPN)的多协议标签交换(MPLS)基础设置中分发泛洪标签的方法和装置。



CN 102986176 A

1. 一种用于在支持边界网关协议(BGP)媒体接入控制(MAC)虚拟专用网络(VPN)的多协议标签交换(MPLS)基础设施中分发泛洪标签的方法,所述方法包括:

在目的地提供商边缘(PE)路由器处,为每个进行通告的 MAC-VPN 实例(MVI)生成通用泛洪标签(GFL);

在目的地 PE 路由器处,为每个进行通告的指定转发器(DF)以太网段标识符(ESI)生成多归属泛洪标签(MHFLx);

使用包括路由标识(RD)和 ESI 的 MAC-VPN 网络层可达信息(NLRI)来向源 PE 路由器分发每个生成的 GFL 和 MHFLx 标签。

2. 根据权利要求 1 所述的方法,进一步包括:

在源 PE 路由器处,根据由目的地 PE 通告的相应 GFL,复制和转发所有通过任何接入电路(AC)接收到的广播 / 未知单播 / 多播(BUM)业务到目的地 PE 路由器。

3. 根据权利要求 2 所述的方法,进一步包括:

在源 PE 路由器处,将在 ESIx 的非 DF AC 上进入的 BUM 业务标记为由目的地 PE 所分发的相应 MHFLx。

4. 根据权利要求 1 所述的方法,进一步包括:

在目的地 PE 路由器处,将在 P2MP LSP 上接收到的任何数据包泛洪到所有本地 MVI 端点。

5. 根据权利要求 4 所述的方法,其中通过非 DF AC 接收到的数据包不被泛洪本地 MVI 端点。

6. 权利要求 1 所述的方法,进一步包括:

在源 PE 路由器处,为每个使用 NLRI 进行通告的 MVI,使用包含多播 VLAN RT 格式来生成 NLRI。

7. 权利要求 6 所述的方法,其中所述包含多播 VLAN RT 格式包括:MVI 的路由标识(RD)、ESI、以太网标签以及源路由器 IP 地址。

8. 权利要求 1 所述的方法,进一步包括:

在源 PE 路由器处,为每个使用 NLRI 进行通告的非 DF ESI 生成相应的 MHFL,其中所述 NLRI 包括:包含路由标识(RD)的以太网段 RT 格式、特定 ESIx、相应的 MHFLx 以及源路由器 IP 地址。

9. 一种用于在支持边界网关协议(BGP)媒体接入控制(MAC)虚拟专用网络(VPN)的多协议标签交换(MPLS)基础设施分发泛洪标签的装置,所述方法包括:

用于在目的地提供商边缘(PE)路由器处,为每个进行通告的 MAC-VPN 实例(MVI)生成通用泛洪标签(GFL)的部件;

用于在目的地 PE 路由器处,为每个进行通告的指定转发器(DF)以太网段标识符(ESI)生成多归属泛洪标签(MHFLx)的部件;

用于使用包括路由标识(RD)和 ESI 的 MAC-VPN 网络层可达信息(NLRI)来向源 PE 路由器分发每个生成的 GFL 和 MHFLx 标签的部件。

10. 一种计算机程序产品,其中当计算机处理计算机指令时,改编所述计算机的操作以提供一种用于在支持边界网关协议(BGP)媒体接入控制(MAC)虚拟专用网络(VPN)的多协议标签交换(MPLS)基础设置中分发泛洪标签的方法,该方法包括:

在目的地提供商边缘(PE)路由器处,为每个进行通告的 MAC-VPN 实例(MVI)生成通用泛洪标签(GFL);

在目的地 PE 路由器处,为每个进行通告的指定转发器(DF)以太网段标识符(ESI)生成多归属泛洪标签(MHFLx);

使用包括路由标识(RD)和 ESI 的 MAC-VPN 网络层可达信息(NLRI)来向源 PE 路由器分发每个生成的 GFL 和 MHFLx 标签。

用于 BGP MAC-VPN 的 MPLS 标签分配的方法和装置

[0001] 相关申请的交叉引用

[0002] 本申请要求将2010年5月19日提交的、名称为“MPLS LABELDISTRIBUTION SCHEME FOR BGP MAC-VPNs”(用于 BGPMAC-VPN 的 MPLS 标签分发方案)的序号为 61/184,205 的临时专利申请的权益,此处通过参考的方式将其引入。

技术领域

[0003] 本发明涉及通信网络领域,并且更具体地,涉及多协议标签交换(MPLS)网络。

背景技术

[0004] 多协议标签交换(MPLS)实现了各种不同的端到端服务的高效递送。MPLS 通过使用标签交换路径(LSP)来支持这种服务的递送。基于不同的因素,在给定的 MPLS 网络中可以提供数百或者数千的 LSP。随着网络条件的改变,给定的 MPLS 网络中所提供的 LSP 通常也需要改变。

[0005] 边界网关协议(BGP)媒体接入控制(MAC)虚拟专用网络(VPN)支持虚拟专用 LAN 服务(VPLS)中的基于 BGP 的 MAC 地址分发。

[0006] 不幸地,由于缺少用于标签分配的可行机制或处理方法,因此,在基于 MPLS 标签的架构环境中还没有针对提供 BGP MAC-VPN 这一问题的可行的解决方案。

发明内容

[0007] 通过本发明的一种用于在支持 BGP MAC-VPN 的 MPLS 架构中分配标签的方法和装置,克服了现有技术的各种缺点。

[0008] 一个实施方式是,一种用于在支持边界网关协议(BGP)媒体接入控制(MAC)虚拟专用网络(VPN)的多协议标签交换(MPLS)基础设施中分发泛洪(flooding)标签的方法,其中该方法包括:在目的地提供商边缘(PE)路由器处,为每个进行通告(Advertising)的 MAC-VPN 实例(MVI)生成通用泛洪标签(GFL);在目的地 PE 路由器处,为每个进行通告的指定转发器(DF,designated forwarder)以太网段标识符(ESI)生成多归属泛洪标签(MHFLx,Multi-Homing Flooding Label);以及,使用包括路由标识(RD,Route-Distinguisher)和 ESI 的 MAC-VPN 网络层可达信息(NLRI,Network Layer Reachability Information)来向源 PE 路由器分发每个生成的 GFL 和 MHFLx 标签。

附图说明

[0009] 通过考虑下面的详细说明,并结合附图,可更加容易理解本发明的教导,其中:

[0010] 图 1 描述了通信网络架构的高级框图;

[0011] 图 2 描述了根据一个实施方式的下行流(downstream)标签分配方法的流程图;

[0012] 图 3 描述了根据一个实施方式的上行流(upstream)标签分配方法的流程图;

[0013] 图 4 描述了适于在此处介绍的各种实施方式中使用的计算机架构和优选交换结

构；

[0014] 图 5-7 描述了根据各种实施方式进行操作的通信网络架构的高级框图。

[0015] 为了便于理解，尽可能的使用相同的参考数字来表示附图中通通用的相同元件。

具体实施方式

[0016] 此处，将在支持边界网关协议（BGP）媒体接入控制（MAC）虚拟专用网络（VPN）的多协议标签交换（MPLS）架构的环境中主要地描述和介绍本发明。所描述的 BGP MAC-VPN 在虚拟专用 LAN 服务（VPLS）转发信息库（FIB）中提供基于 BGP 的 MAC 地址分发，由此消除了多协议标签交换（MPLS）核心网上 MAC 的学习（learning）和泛洪。进一步的，所述系统能够为第 2 层多点到多点 VPN 服务提供多路径或激活 / 激活访问弹性（active access resiliency）。

[0017] 此处所提供的标签分配方案应对各种挑战，包括：（1）数据包复制，例如远端客户边缘（或客户设备）CE 接收到相同数据包的两个复制；（2）环路预防，例如起始于特定 CE 的数据包返回到该特定 CE；（3）MAC 表不稳定性，例如 MAC M1 在位于不同链路的目的地 CE 处表现不同，因而产生了重新排序问题和 MAC 表不稳定性。

[0018] 图 1 描述了根据一个实施方式的通信网络架构的高级框图。特别地，图 1 的架构 100 提供了支持媒体接入控制（MAC）虚拟专用网络（VPN）或 MAC-VPN 的边界网关协议（BGP）多协议标签交换（MPLS）网络（BGPMPLS 网络）。

[0019] 架构 100 包括 IP/MPLS 通信网络（CN）110、网络管理系统（NMS）120、多个提供商边缘（PE）路由器（或者 MPLS 边缘交换机（MES））130-1 至 130-4（其共同构成 PE 路由器 130）以及多个客户边缘（CE）路由器 140-1 至 140-7（其共同构成 CE 路由器 140）。

[0020] 所述 PE 路由器 130 通过由 CN 110 的 MPLS 基础设施内大量的路由器或交换机元件（未示出）所实现的 MPLS 标签交换路径（LSP）隧道的全网络（full mesh）连接在一起。

[0021] 各种 CE 路由器 140-1 至 140-7 的每一个都关联到相应的媒体接入控制（MAC）并且被连接到一个或多个 PE 路由器 130。例如，在图 1 的示例性实施方式中，PE 路由器 130-1 连接到 CE 路由器 140-1 至 140-3，PE 路由器 130-2 连接到 CE 路由器 140-2 至 140-4，PE 路由器 130-3 连接到 CE 路由器 140-5 和 140-6，以及 PE 路由器 130-4 连接到 CE 路由器 140-6 和 140-7。应当理解的是，可将更多或更少的 CE 路由器 140 连接到各种 PE 路由器 130，并且此处仅出于示意性的目的来提供特定的组合 / 连接。

[0022] 根据基于每个服务的入口（ingress）和出口（egress）虚拟连接（VC）标签来对数据分组或数据报进行路由。所述 PE 路由器 130 利用所述 VC 标签对同一组 LSP 隧道上从不同服务到来的业务进行解多路复用。

[0023] PE 路由器学习到达它们的访问和网络端口上的业务的源媒体接入控制（MAC）地址。每个 PE 路由器 130 维持用于每个 VPLS 服务实例（instance）的转发信息库（FIB），并且将所学习的 MAC 地址填充到服务的 FIB 表中。使用 LSP 隧道，基于 MAC 地址来交换所有的业务并且在所有的参与 PE 路由器之间转发所有的业务。

[0024] 将用于所述服务（例如，泛洪到 PE 路由器）的未知数据包（即，目的地 MAC 地址尚未被学习的数据包）在所有的 LSP 上转发到参与的 PE 路由器上，直到适当的目的地或者目标站响应为止，使得与所述服务相关联的 PE 路由器学习 MAC 地址。

[0025] NMS 120 是网络管理系统，其适于执行此处所描述的各种管理功能。NMS 120 适于

与 CN 110 的节点进行通信。NMS 120 还可适于与其它操作支持系统(例如,元件管理系统(EMS)、拓扑管理系统(TMS)以及类似系统或其各种组合)进行通信。

[0026] 可在网络节点、网络操作中心(NOC)或任何其它能够与 CN 110 以及各种相关元件进行通信的位置处实施 NMS 120。NMS 120 可支持用户接口能量,以使得一个或多个用户能够执行各种网络管理、配置、供应或者与控制相关的功能(例如,输入信息、核查信息、启动此处所描述的各种方法的执行等)。参照各种实施方式,NMS 120 的各种实施方式适于执行此处所讨论的功能。

[0027] 为了简化关于各种实施方式的操作的讨论,在图 1 中具体地引用多个路径。特别地,路径 190 用于在 MES-2 (130-2) 和网络 110 之间传送数据,路径 191 在 MES-1 (130-1) 和网络 110 之间传送数据,路径 192 在 MES-1 (130-1)和 CE1 (140-1)之间传送数据,路径 193 在 MES-1 (130-1) 和 CE2 (130-2) 之间传送数据,路径 194 在 MES-1 (130-1) 和 CE3 (130-3)之间传送数据,路径 195 在 MES-2 (130-2)和 CE2 (130-2)之间传送数据,路径 196 在 MES-2 (130-2) 和 CE3 (130-3) 之间传送数据,以及路径 197 在 MES-2 (130-2) 和 CE4 (130-4) 之间传送数据。如图 1 所示,存在其它路径。

[0028] 基于 BGP MPLS 的 MAC-VPN

[0029] 以上描述的通信网络实现支持媒体接入控制(MAC) 虚拟专用网络(VPN) 或 MAC-VPN 的边界网关协议(BGP) 多协议标签交换(MPLS) 网络(BGP MPLS 网络)。现在将描述各种实现方式细节。

[0030] 如前所述,MAC-VPN 网络包括 :与被设置在 MPLS 基础设施边缘的 PE 或者 MPLS 边缘交换机(MES) 连接的 CE。CE 可以是主机、路由器或交换机。MPLS 边缘交换机提供 CE 之间的第 2 层虚拟桥接连通性。提供商网络中可存在多个 MAC-VPN。MES 上的 MAC-VPN 的实例可被称为 MAC-VPN 实例(MVI)。MES 通过 MPLS LSP 基础设施进行连接。

[0031] 通过控制窗格(control pane)学习(learn)MAC

[0032] MES 之间的学习发生在控制平面(control plane)中,特别是 BGP 控制平面。该控制平面学习有利地实现负载平衡,允许 CE 连接到附属的多个激活节点并且改善在某些网络故障事件中的收敛时间(convergencetime)。MES 和 CE 之间的学习发生在数据平面中,例如根据 IEEE 802.1x、802.1aq、LLDP 或其它协议。

[0033] MES 上的第 2 层转发表可包含所述控制平面已知的所有 MAC 目的地或者利用基于高速缓冲存储器的方案所选择的已知 MAC 目的地的子集。例如,可仅利用传递特定 MES 的激活数据流的 MAC 目的地来填充特定 MES 的转发表。

[0034] MAC-VPN 的策略属性类似于 IP-VPN 的策略属性。MAC-VPN 实例需要路由标识(RD),并且 MAC-VPN 需要一个或多个路由目标(RT)。CE 附着于 VLAN 上的特定 MVI 内 MES 上的 MAC-VPN 或者简单地附着于以太网接口。当附着点是 VLAN 时,特定 MAC-VPN 中可能存在一个或多个 VLAN。一些部署方案确保 MAC-VPN 上 VLAN 的唯一性 :给定的 MAC-VPN 的所有附着点使用相同的 VLAN,并且其它任何 MAC-VPN 都不使用该 VLAN。这被称为“单个 VLAN MAC-VPN”。

[0035] 以太网段标识符

[0036] 如果 CE 多归属于两个或多个 MES,则该组附着电路构成以太网段。以太网段可以按链路聚集组(Link Aggregation Group)的形式呈现给 CE。以太网段具有被表示为以太

网段标识符(ESI)的标识符。将单归属的 CE 视为附着于具有 ESI 0 的以太网段;其它情况下,以太网段具有唯一的非零 ESI。

[0037] 可使用各种机制来分派 ESI:(1)可配置 ESI;(2)如果在作为主机的 CE 和 MES 之间使用了链路聚集控制协议(LACP),则可由 LACP 来确定 ESI;(3)如果在作为主机的 CE 和 MES 之间使用了链路标签分发协议(LLDP),则可以通过 LLDP 来确定 ESI;以及(4)在非直接连接的主机以及主机和 MES 之间的桥接 LAN 的情况下,基于第 2 层桥接协议来确定 ESI,其中,通过监听以太网段上的 BPDU 来取得 ESI 的值(MES 通过监听 BPDU 来学习交换机 ID、MSTP ID 以及根网桥 ID)。

[0038] 确定单播 MAC 地址的可达性

[0039] MES 转发 MES 基于目的地 MAC 地址接收的数据包。因此,MES 必须能够学习如何到达给定的目的地单播 MAC 地址。存在两种 MAC 地址学习方式,即“本地学习”和“远端学习”。

[0040] 本地学习是特定 MES 学习连接到 MES 的 CE 的 MAC 地址。即,特定 MAC-VPN 中的 MES 支持本地数据平面,其中通过标准以太网学习过程从相连接的 CE 学习。当 MES 从 CE 网络接收到数据包时,例如 DHCP 请求、用于其自己 MAC 的免费 ARP 请求,用于对等体(peer)的 ARP 请求等,MES 在数据平面学习 MAC 地址。可替换地,如果 CE 是主机,则 MES 可使用运行在 MES 和主机之间的诸如 LLDP 的协议的扩展,在控制平面中学习主机的 MAC 地址。在 CE 是主机或者连接到主机的交换网络的情况中,通过给定的 MES 可获得的 MAC 地址可移动,使得其变为可经由另一 MES 获得。这被称为 MAC 移动。

[0041] 远端学习是特定 MES 学习远端 CE 的 MAC 地址;即,“在后面”的或通过其它 MES 连接的 CE,或者“在后面”的或通过远端 CE 连接的 CE 或主机。在控制平面上执行 MAC 地址的远端学习。为了达到远端学习,每个 MES 在控制平面中通告其从本地附着的 CE 学习的 MAC 地址。

[0042] MES 控制平面通告

[0043] 通过使用 BGP 扩展,将每个 MES 在控制平面上通告的其学习的 MAC 地址提供给在 MAC-VPN 中的其它 MES。特别地,将 BGP 扩展以使用表示为 MAC-VPN-NLRI 的网络层可达性信息(NLRI)来通告这些 MAC 地址。该扩展包括 MAC-VPN 中新的地址族标识符(AFI)以及新的子地址族识别号子序列(SAFI,Subsequent Address Family Identifier)。

[0044] 当用于 BGP MAC VPN 时,MAC-VPN-NLRI 对多个信息元素或者字段进行编码,例如路由类型(RT)、长度字段和值字段。

[0045] 路由类型(RT)用于识别以下值字段的格式。可定义多个路由类型码点。长度字段用于在以下值字段的八位字节中指示长度。所述值字段 - 携带每个 RT 专用的信息。

[0046] 针对这种讨论的目的,将使用以下 RT,也可使用其它的 RT:

[0047] (a)以太网标记(tag)自动发现 - 允许指定转发器(DF)选择和负载平衡功能。可用于快速 MAC 回收;

[0048] (b)MAC 通告 - 用于在 MES 之间的 MAC 地址通告;

[0049] (c)包含多播的 VLAN- 提供一种指示在源 MES 对某些分组进行泛洪的机制。通常,在入口 MES 处对 BUM 业务(BUM = 广播,未知的单播、多播业务)进行泛洪。由于只有 DF 对被标记为泛洪到到 MH CE 的数据包进行转发,因此,这保证了只有泛洪数据包的一个拷贝被

传递到多归属(MH)CE;和/或

[0050] (d)以太网段路由-提供环路避免;使用MH ESI专用的标签对来自非DF附着(attachment)电路的MH CE的进入的业务进行标记。在包含用于MH ESI的DF的接收MES处,所述标记被用于阻塞数据包以防止其被转发回相同的MH CE。

[0051] 现在将讨论携带MAC通告RT的NLRI的示例性结构。该结构包括下述内容:

[0052] (1)通告NLRI的MAC-VPN实例的路由标识(RD)。具体地,在MES上为每个MAC-VPN实例分派唯一的RD,例如通过使用类型1的RD。值字段可包括MES的IP地址(例如,环回地址),接着是对于MES唯一的数字。可通过MES生成该数字,或者该数字可为VLAN ID的所有或一部分(例如,在单NLAN MAC-VPN的情况下)。

[0053] (2)VLAN ID,如果通过VLAN从CE处学习了MAC地址(否则设置为0);

[0054] (3)以太网段标识符(ESI);

[0055] (4)MAC地址;

[0056] (5)可选地,一个或多个与学习的MAC地址相关联的IP地址;

[0057] (6)MAC-VPN MPLS标签,其中MES使用MAC-VPN MPLS标签转发从远端MES接收到的数据包。MES可在给定的MAC-VPN实例(被表示为Per-MVI标签分派)中向所有MAC地址通告相同的MAC-VPN标签,或者为每个MAC地址通告唯一的MAC-VPN标签。Per-MVI标签分派需要最少数量的MAC-VPN标签,但需要在出口MES处除MPLS查找之外的MAC查找以用于转发。唯一的MAC地址标签分派允许在仅执行MPLS标签查找之后(例如,无MAC查找),由出口MES转发数据包(例如,从另一个MES到所连接的CE所接收的)。

[0058] (7)一个或多个路由目标(RT)属性,可通过从与通告相关联的VLANID自动获得或配置(例如,在IP VAN中)所述属性。可通过将RT的全局管理者字段设置为MES的IP地址。从与通告相关联的VLAN ID自动获得路由目标(RT)属性。对于MES上的所有MAC-VPN实例,该IP地址应当是通用的,例如MES的环回地址。如果MAC-VPN包含多个VLAN,可为MAC-VPN中的每个VLAN使用不同的RT,且从用于所述MAC-VPN的VLAN获得仅包括一个VLAN的用于MAC-VPN的RT。

[0059] (8)可选地,与MAC地址相关联的IP地址,例如当IP地址的数量大于1且不能在NLRI中编码时。

[0060] 标 签分配的数据平面影响

[0061] 示例性地,可通过per-Mac、per-ESI或者per-VMI来提供标签分配。存在需要考虑多个权衡(tradeoff),包括以下权衡。如果通过per-MAC来提供标签分配,其结果是非常大的标签数,具有可选MAC查找的出口转发以及对ETREE的支持。如果通过per-ESI提供标签分配,其结果是中等的标签数,具有可选MAC查找的出口转发以及对ETREE的支持。如果通过per-VMI提供标签分配,其结果是较低的标签数,具有需要MAC查找的出口转发以及不支持ETREE。

[0062] 多归属CE的指定转发器(DF)选择

[0063] 如果作为主机或路由器的CE直接地多归属到MAC-VPN中的多于一个的MES,那么只有其中的一个MES负责某些动作。具体地,只有一个MES将向CE发送多播、广播和未知单播业务(例如,MES不知道其目的地MAC地址的业务)。典型地,CE使用单个链路来发送数据包。如果CE是主机,则主机CE将用于到达MES的多个链路视为链路聚集族(LAG)或束

(bundle)。

[0064] 如果桥接网络通过交换机被多归属于 MAC-VPN 中多于一个的 MES，那么只有其中的一个 MES 负责某些动作。具体地，多归属桥接网络中只有一个 MES 会(1)向多归属桥接网络之外的其它 MES 转发数据包；(2)向桥接网络发送多播、广播以及未知的单播业务。

[0065] 将特定的一个 MES 称为用于以太网段的指定转发器(DF)MES，其中通过所述以太网段将所述 CE 多归属于两个或多个 MES。所述以太网段可以是链路束，例如其中主机或路由器直接连接到 MES，或者桥接 LAN 网络，例如其中 CE 为交换机。

[0066] 对以太网段或以太网段和 VLAN 的组合，MES 使用 BGP 来执行指定转发器(DF)选择。为了执行 DF 选择，针对 MAC-VPN 中的每个以太网段，每个 MES 使用 MAC-VPN-NLRI 来在 BGP 中通告以太网标记自动发现路由类型。

[0067] 典型地，每个以太网标记自动发现 NLRI 包含以下信息元素或字段：

[0068] (1)通告 NLRI 的 MAC-VPN 实例的路由标识(RD)。

[0069] (2)以太网段标识符。

[0070] (3)可选地，可设置为 0 的 VLAN ID。

[0071] (4)被称为“ESI 标签”上行流分派的 MPLS 标签。

[0072] (5)P 隧道属性，例如在 VPLS-MCAST 中指定的。

[0073] (6)一个或多个路由目标(RT)属性。

[0074] 通过构建 MES 候选列表并从候选列表选择 DF，进行特定 ESI 和 VLAN 的组合的 DF 选择。

[0075] 在 MES 或 NMS 处构建候选列表，并且所述候选列表包括具有特定{ESI，VLAN}元组的所有路由，其中 MES 引入 MAC-VPN 实例，如果有的话，包括由 MES 自身所产生的路由。

[0076] 之后，由引入以太网标记自动发现路由类型的这些 MES 来从该候选列表中选择或选出 DF MES。在一个实施方式中，所选择的 DF 是候选列表中所有 MES 中具有最高 IP 地址的 MES。通过这种方式，每个 MES 将为给定的 ESI 和 VLAN 的组合(除了在路由瞬态期间)选择相同的 DFMES。

[0077] BGP MAC VPN 问题

[0078] 以上描述的机制有助于应对与 BGP MAC-VPN 相关联的各种挑战，涉及：(1)数据包复制，其中远端 CE 接收相同数据包的两个拷贝；(2)环路预防，其中将源自 CE1 的数据包返回到 CE1 (例如，永久环路和 / 或暂时性环路，在 ETH 数据包中无 TTL 等)；以及(3)MAC 表不稳定性，其中 MAC 表 M1 在不同链路的目的地 CE2 上呈现不同(使得在链路之间需要经常移动，进而造成重新排序问题和 MAC 表不稳定性)。

[0079] 在一个实施方式中，通过 BGP MAC VPN 机制来解决 MAC 表格的不稳定性，其中，在 CE 处使用链路聚集族(LAG)，使得在多个链路上出现的相同 MAC 不会造成 MAC 移动 /MAC 表不稳定性。在这种实施方式中，CE 处的 MAC 学习被停止并且被替代为 CE<-PE MAC 协议。这种方式是 IEEE 802.1aq 规范中描述的方法的修改版本。

[0080] 由每个目的地 PE 路由器为每个泛洪域生成通用泛洪标签，并且使用下行流标签配置(图 2)进行分配，或者由源 PE 生成通用泛洪标签并使用上行流标签配置(图 3)进行分配。源 PE 路由器相应地根据其目的地 MAC 地址(如果已知)来路由数据包并为 MAC 地址添加相关联的单播标签。如果 MAC 地址未知或者其为一个组 MAC (多播 / 广播)，则在数据包

中添加适当的泛洪标签以指示数据包在 BGP MAC VPN 的源处被洪泛。可在数据包中添加另外的点到点通道标签(下行流标签分配情况下)或者点到多点通道标签(上行流标签分配情况)以将其在 MPLS 网络 110 中进行传输。

[0081] 图 2 描述了根据一个实施方式的下行流标签分配方法的流程图。特别地,图 2 描述了适于在提供点到点(P2P)标签交换路径(LSPs)的 BGPMAC VPN 中进行泛洪标签分配的泛洪标签分配方法 200。

[0082] 为 MAC-VPN 实例(MVI)内的每个泛洪域提供一个泛洪标签,并为与多归属 CE 相关的每一个以太网分片标识符(ESI)提供一个泛洪标签。所产生的泛洪标签通过此前描述的标识为 MAC-VPN-NLRI 的 BGP 网络层可达信息(NLRI)通告给其它 PEs。

[0083] 在步骤 210 中,在目的地提供商边缘(PE)路由器处,对每一个使用 NLRI 通告的 MVI,相应的 PE 路由器生成通用泛洪标签(GFL)并在 NLRI 中包含所包括的多播 VLAN RT 格式 :MVI 的路由区分器(RD)、ESI、以太网标记以及源路由器 IP 地址。即 :NLRI :RF+ESI+ 以太网标记 + 路由器 IP。GFL 包含在 P 通道属性中,其中通道类型为点到点。

[0084] 在步骤 220 中,在目的地 PE 路由器处,对每一个使用 NLRI 通告的 DF ESI,相应的 PE 路由器同样生成各自的的多归属泛洪标签(MHFL),并在 NLRI 中包含所包括的以太网分段 RT 格式 :路由区分器(RD)、特定的 ESIx、相应的 MHFLx 以及源路由器 IP 地址。即 :NLRI :RD+ESIx+MHFLx+PE IP。

[0085] 在步骤 230 中,在源提供商设备(PE)路由器处,所有的引入到任何接入电路(AC,attachment circuit)中的广播 / 未知单播 / 多播(BUM)业务都被复制并发送到所有的目的地 PEs,所述目的地 PEs 为使用被每个目的地 PE 通告的 GFL 的 MAC VPN 成员。

[0086] 在步骤 240 中,在源提供商设备(PE)路由器处,除了 GFL,进入 ESIx 的非 DF AC 的 BUM 业务被标记为 MHFLx,该 MHFLx 由具有相应 ESIx 的目的地 PE 分配。即,与特定 ESI 相关的多归属泛洪标签仅为起始于与特定 ESI 相关的非 DF AC 的 BUM 业务保留。

[0087] 在步骤 250 中,在目的地 PE 路由器处,除了非 DF ACs 外,在 P2MPLSP 上接收的任何数据包都被洪泛到所有的本地 MVI 端节点上。当具有 MHFLx 时,所述数据包还不会在用于 ESIx 的 DF AC 上发送。

[0088] 图 3 描述了根据一个实施方式的上行流标签分配方法的流程图。特别地,图 3 描述了适用于在使用点到多点(P2MP)标签交换路径(LSPs)的 BGP MAC-VPN 中进行泛洪标签分配的泛洪标签分配方法 300。

[0089] 为每一个 MAC-VPN 实例(MVI)提供一个 P2MP LSP 标签,并且为每一个在以太网分段标识符(ESI)上的非指定转发器(non-DF)提供一个泛洪标签。使用此前描述的标记为 MAC-VPN-NLRI 的网络层可达信息(NLRI)来向其他 PEs 传播或通告所生成的标签。

[0090] 在步骤 310 中,在源提供商边缘(PE)路由器处,对每个使用 NLRI 进行通告的 MVI,相应的 PE 使用内含多播 VLAN RT 格式来生成 NLRI :MVI 的路由区分器(RD),ESI,以太网标记以及源路由器 IP 地址。即,NLRI :RD+ESI+ 以太网标记 + 路由器 IP,所述路由器 IP 具有 P2MP 通道类型的 P 通道(PMSI 通道)属性,ACs 上引入的任何 BUM 业务不使用 GFL,因为与 P2MP LSP 相关的标签已经指示所述业务在源 PE 处洪泛。

[0091] 在步骤 320 中,在源 PE 路由器处,对于正在使用 NLRI 进行通告的每个非 DF ESI,相应的 PE 还会生成各自的 MHFL 并且在 NLRI 中包含以太网分段 RT 格式 :路由区分器

(RD)、特定的 ESIx、相应的 MHFLx 以及具有 P 通道属性的源路由器 IP 地址。即，NLRI：RD+ESIx+MHFLx+PE IP，以及具有 P- 通道属性。所述 MHFLx 用于除了 P2MP LSP 标签之外的在用于 ESIx 的非 DF AC 上引入的任何 BUM 业务。

[0092] 在步骤 330 中，在目的地 PE 路由器处，除了非 DF ACs 外，在 P2MPLSP 上接收的任何数据包都被洪泛到所有本地 MVI 端节点上。当具有 MHFLx 时，所述数据包还不会在用于 ESIx 的 DF AC 上发送。

[0093] 图 2-3 的方法计算了通用泛洪标签(GFL)以及多归属泛洪标签(MHFLx)的分配和使用。

[0094] 在这种方式中，提供了环回避免机制来防止起源于源 PE 上的非 DFxAC 的业务被通过 DFx 转发回 CEx。该机制还在当洪泛到 DFx 和非 DFx 两者端节点上的相同数据包被转发到 CEx 时防止数据包复制。

[0095] 例如，从图 1 可以看出，CE-2 和 CE-3 两者都多连接到 MES-1 和 MES-2，其中 MES-2 被选择作为 DF。从 CE-2 或 CE-3 到 MES-1 的业务流将被 MES-1 抛弃，而从 CE-2 或 CE-3 到 MES-2 的业务流将或者被 MES-2 转发到正确的目的地 MES（MES-2 已知的目的地 MAC 地址）或者被洪泛到此处所描述的其它 MESs（MES-2 未知的目的地 MAC 地址）。在这种方式中，无需将业务洪泛回传输的起源处。

[0096] 以上描述的方法和技术提供了一种 BGP MAC VPN 解决方案，其适用于当没有使用汇聚树，或者 GFL 和 MHFLx 时，使用数据包中的第三 MPLS 标签或用于 P2MP LSP/MP2MP LSP 的第二 MPLS 标签阻止环路和数据包复制。

[0097] 通常来说，在 BUM 业务的情况下，在数据包中加入第三标签。如果没有使用汇聚树，则对 P2MP LSP/MP2MP LSPs(或者 GFL 或者 MHFLx 标签)使用第二标签。在源 MES 处，使用(上行流 / 下行流)通用泛洪标签(GFL)，GFL=0（代表空标签）来标记 BUM 业务。在不同的实施方式中，所有 MES（s）上的 GFL 标签都相同，对于本地泛洪，所述数据包被发送到所有本地 AC（s)上，其为 SH 和 DF MH AC（s)，以及在远端 MES，所述泛洪传输仅在 SH&DF ACs 上使用 GFL 进行标记。

[0098] 计算机硬件 / 软件实施方式

[0099] 图 4 描述了适用于此处所描述的多种实施方式的计算机架构和可选的交换机制。所述计算机架构可被适配以执行此处所描述的特定功能，包括标签生成、标签分配、数据包路由、报文路由、传输路由、控制面处理功能、数据面处理功能等等。

[0100] 所述计算机架构示例性的包括处理器元件 410（例如中央处理单元(CPU)和 / 或其它合适的处理器)、存储器 420（例如随即访问存储器(RAM)、只读存储器(ROM)等等)、BGP MAC-VPN 模块 / 处理器 425（其可包含在存储器 420 中）以及各种输入 / 输出设备 430。

[0101] 存储器 420 被描述为包含控制程序 422、数据存储器 424 以及支持程序 426。存储器 420 的这些不同程序和数据存储部分可用于存储用于执行此处所描述的算法的程序、用于支持各种算法的程序、数据库、路由表以及支持各种算法的其它数据结构、报告功能 / 程序等等。

[0102] 不同的输入 / 输出设备 430 可包括用户输入设备，例如键盘、键面、鼠标等等；用户输出设备例如显示器、扬声器等等；输入通信端口、输出通信端口；接收器 / 发射器(例如网络连接或其它合适类型的接收器 / 发射器)；存储设备(例如硬盘驱动、致密磁盘驱动、光盘

驱动等等)。

[0103] 可选交换机制 490 包括交换构造 492 和入口 / 出口端口 494。具体地,可选交换机制 490 被描述为通过第一组多个输入 / 输出端口 494A 来与第一组其它路由 / 交换设备通信,以及通过第二组多个输入 / 输出端口 494B 来与第二组其它路由 / 交换设备通信。所述可选交换机制 490 被描述为相对普通配置。所述可选交换机制 490 的其它相关配置对本领域技术人员来说是易于理解的,并且发明人主张将其包含在本实施方式的范围之内。

[0104] 在一个实施方式中,用于执行各种实施方式的所述方法相关的计算机软件代码可被下载到存储器上并通过处理器来执行用以实现以上所讨论的功能。在一个实施方式中,用于执行各种实施方式的所述方法相关的计算机软件代码可被存储在计算机可读存储媒介上,例如 RAM 存储器、磁性或光学驱动或磁盘等等。计算机适于作为此处所描述的任何网络元件而使用,包括但不限于客户边缘(CE)路由器、提供商边缘(PE)路由器、MPLS 边缘交换机(MESs)以及此处所描述的其它网络元件。

[0105] 应当注意,此处所描述的功能可通过软件和 / 或软硬件结合的方式来执行,例如使用通用目的计算机、一个或多个应用特定集成电路(ASIC)和 / 或任何其它硬件等价物。

[0106] 应当注意,此处所讨论的作为软件方法的一些步骤可在硬件中实施,例如,作为与处理器协作的电路来执行各种方法步骤。此处描述的部分功能 / 元件可被实现为计算机程序产品,其中当被计算机处理时,计算机指令可适配计算机的操作从而执行或提供此处所描述的方法和 / 或技术。用于执行本发明方法的指令可存储在切实固定或可移动的媒体上,通过切实的或不切实的广播或其它信号承载媒质来发送,和 / 或存储在根据指令运行的计算设备的存储器内。

[0107] 尽管此处实施例中主要描述的 BGP MAC-VPN 功能被用于特定协议,然而 BGP MAC-VPN 功能的原理可被适配用于任何其它合适的协议中。

[0108] 尽管此处实施例中主要描述的 BGP MAC-VPN 功能被用于特定类型网络(示例性的,IP/MPLS 网络),然而 BGP MAC-VPN 功能的原理可适配用于任何其它合适的网络中。

[0109] 通常来说,此处讨论的通用架构的计算机硬件、软件和 / 或固件的可在与网络相关的多个节点、网络元件或网络管理单元的每个上进行复制和使用。此外,位于不同位置、节点、网络元件或者网络管理系统元件的这些计算机硬件、软件和 / 或固件可操作地相互通信以实现此处所设计的各种步骤、协议、交互等。

[0110] 图 5-7 描述了根据各种实施方式进行操作的通信网络架构的高级框图。具体地,根据各种实施方式,图 5-7 描述了具有沿指示泛洪行为的参考路径 190-197 的业务流指示箭头的图 1 的架构。

[0111] 图 5 示出了响应于在非 DF MH PE 路由器处接收到 BUM 业务的 PE 路由器泛洪行为的实例。具体地,CE-2 通过路径 193 转发 BUM 业务到 MES-1。MES-1 是关于 CE-2 的非 DF PE 路由器。

[0112] 响应地,作为关于 CE-2 的非 DF PE 路由器的 MES-1 通过路径 191 将 BUM 业务泛洪到所有其它的 ME 130 以及泛洪到任意 CE,其中 ME-1 作为对于所述 CE 的 DF MH 路由器(在这个例子中,CE-1 通过路径 192)。注意的是,BUM 业务不是通过路径 194 从 MES-1 泛洪到 CE-3,这是因为 ME-1 不是对应于 CE-3 的 DF 路由器。

[0113] 响应地,通过路径 190 从 MES-1 接收泛洪的 BUM 业务的 MES-2,将 BUM 业务泛洪到

所有其归属的或本地的 CE，除了具有 BUM 业务相同的 ESI 的本地 CE 之外；即，CE-2。在这种方式中，起始于 CE-2 的 BUM 业务不会被泛洪到或者路由回 CE-2。

[0114] 图 5 中还描述了在 NMS 120 的控制下，与经过 IP/MPLS 核心 110 传递的业务相关联的各种标签。堆栈的标签包括：与入口复制栈 510 相关联的三个标签（其中，将第三标签表示为 LBL ＝ 2+16）、与 P2MP LSP/MP2MPLSP 520 相关联的两个标签以及与 P2MP LSP/MP2MP LSP+ 汇聚树 530 相关联的三个标签（其中，第三标签被表示为 LBL ＝ 2+16）。

[0115] 图 6 示出了响应于在 DF MH PE 路由器处接收到 BUM 业务的 PE 路由器泛洪行为的实例。具体地，CE-2 通过路径 195 转发 BUM 业务到 MES-2。MES-2 是关于 CE-2 的 DF PE 路由器。

[0116] 响应地，作为关于 CE-2 的 DF MH 路由器的 MES-2，通过路径 190 将 BUM 业务泛洪到所有其它的 ME 130 以及泛洪到任意 CE，其中 ME-3 作为对于所述 CE 的 DF MH 路由器（在这个例子中，CE-3 通过路径 196，并且 CE-4 通过路径 197）。注意的是，BUM 业务不从 MES-2 通过路径 195 泛洪返回到 CE-2。

[0117] 响应地，通过路径 191 从 MES-2 处接收泛洪的 BUM 业务的 MES-1，将 BUM 业务泛洪到所有其归属的本地 CE；即，CE-1。MES-1 不转发 BUM 业务到任何连接的非 DF CE，例如在这个例子中的 CE-2 和 CE-3。在这种方式中，起始于 CE-2 的 BUM 业务不会被泛洪或者路由回到 CE-2。

[0118] 图 6 中还描述了在 NMS 120 的控制下，与经过 IP/MPLS 核心 110 传递的业务相关联的各种标签。堆栈的标签包括：与入口复制栈 610 相关联的三个标签（其中，第三标签被表示为 LBL ＝ 2+16）、与 P2MP LSP/MP2MPLSP 620 相关联的两个标签，以及与 P2MP LSP/MP2MP LSP+ 汇聚树 630 相关联的三个标签（其中，第三标签被表示为 LBL ＝ 2+16）。注意的是，由于 DF MH 站点对 BUM 业务的处理，因此被表示为 ALU5 的第三实体与 P2MP LSP/MP2MP LSP 栈 620 相关联。

[0119] 图 7 示出了响应于在 SH MH PE 路由器处接收到的 BUM 业务的 PE 路由器泛洪行为的实例。具体地，CE-1 通过路径 192 转发 BUM 业务到 MES-1。MES-1 是关于 CE-1 的 DF PE 路由器。

[0120] 响应地，作为关于 CE-1 的 DF MH 路由器的 MES-1，通过路径 191 将 BUM 业务泛洪到所有其它的 ME 130 以及泛洪到任意 CE，其中 ME-1 作为对于所述 CE 的 DF MH 路由器（在这个例子中没有）。注意的是，BUM 业务不从 MES-1 通过路径 192 泛洪返回到 CE-1。

[0121] 响应地，通过路径 190 从 MES-1 处接收泛洪的 BUM 业务的 MES-2，将 BUM 业务泛洪到所有其 SH 和 MH CE，即，CE-2、CE-3 和 CE-4。在这种方式中，起始于 CE-1 的 BUM 传输不会被泛洪或路由回到 CE-1。

[0122] 图 7 中还描述了在 NMS 120 的控制下，与经过 IP/MPLS 核心网 110 传递的业务相关联的各种标签。堆栈的标签包括：与入口复制栈 710 相关联的三个标签（其中第三标签被表示为 LBL ＝ 0）、与 P2MP LSP/MP2MPLSP 720 相关联的两个标签以及与 P2MP LSP/MP2MP LSP+ 汇聚树 730 相关联的三个标签（其中第三标签被表示为 LBL ＝ 0）。

[0123] 尽管此处已经详细示出和描述了包含本发明技术的各种实施方式，然而，对本领域技术人员来说，可容易得到包含这些教导的其它各种实施方式。

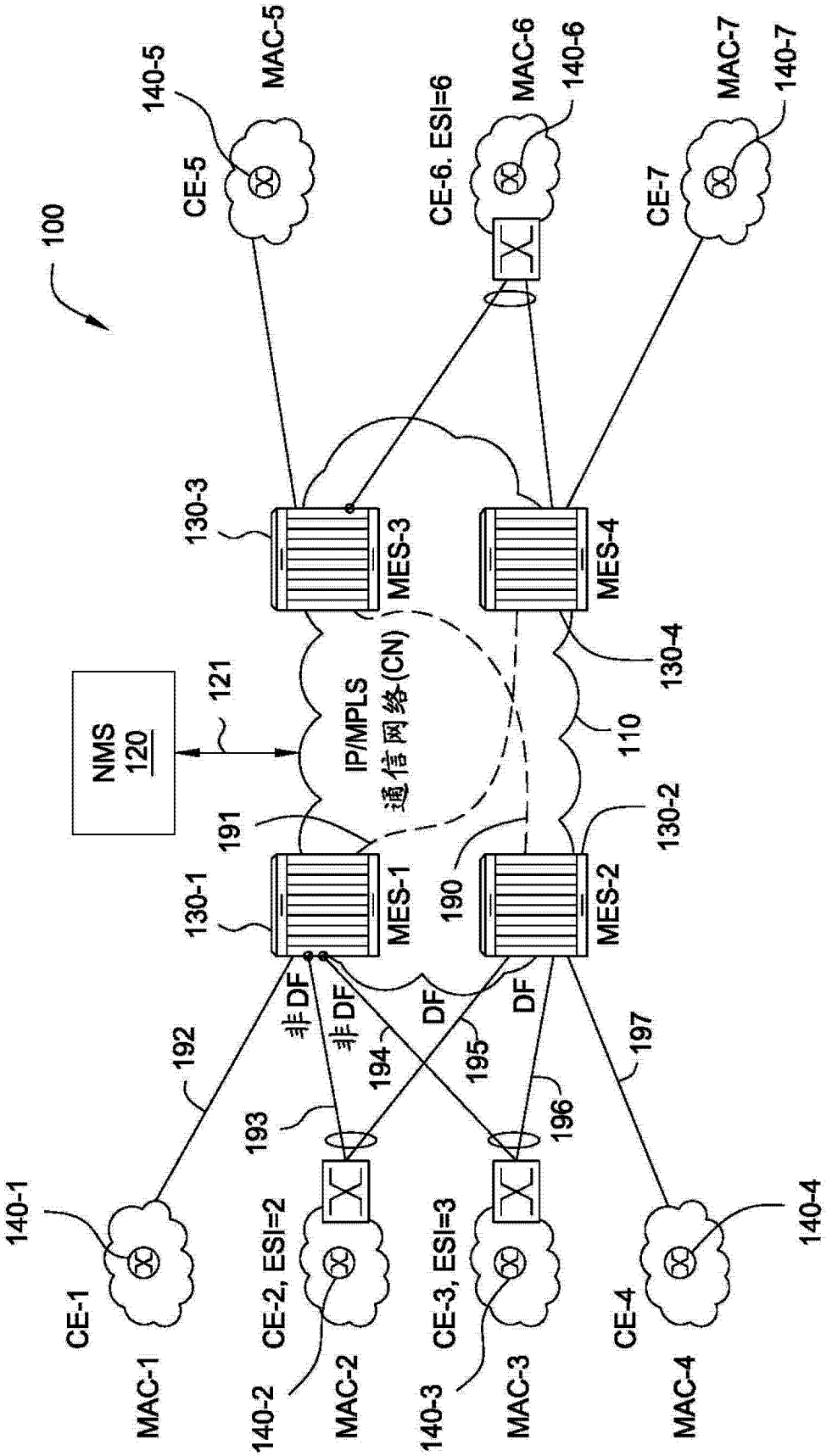

100
NMS
120
121
IP/MPLS
通信网络(CN)
110
MES-1
MES-2
MES-3
MES-4
130-1
130-2
130-3
130-4
191
190
CE-1
CE-2, ESI=2
CE-3, ESI=3
CE-4
CE-5
CE-6. ESI=6
CE-7
MAC-1
MAC-2
MAC-3
MAC-4
MAC-5
MAC-6
MAC-7
140-1
140-2
140-3
140-4
140-5
140-6
140-7
192
193
194
195
196
197
非DF
非DF
DF
DF

图 1

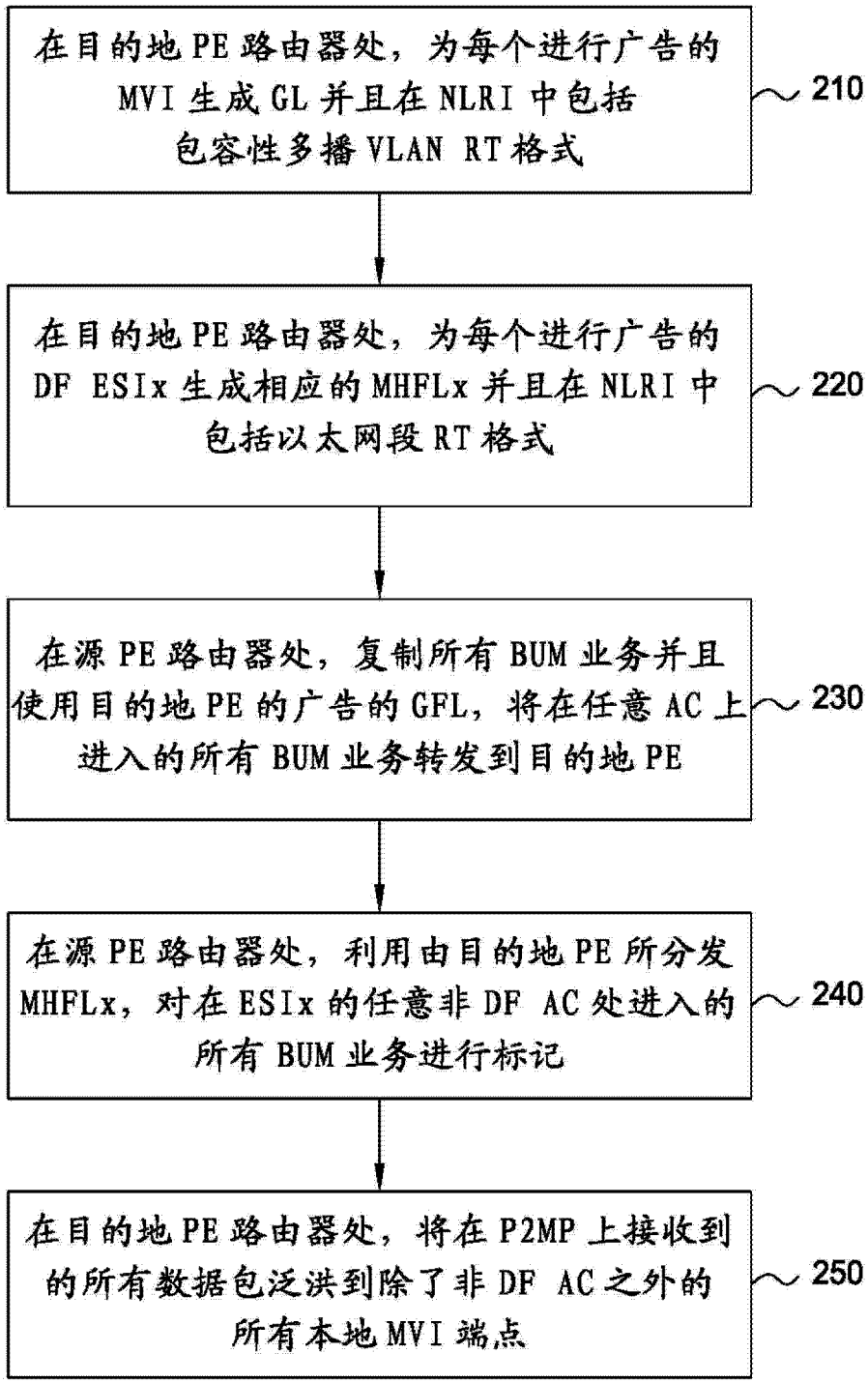

在目的地 PE 路由器处，为每个进行广告的 MVI 生成 GL 并且在 NLRI 中包括包容性多播 VLAN RT 格式
210
在目的地 PE 路由器处，为每个进行广告的 DF ESIx 生成相应的 MHFLx 并且在 NLRI 中包括以太网段 RT 格式
220
在源 PE 路由器处，复制所有 BUM 业务并且使用目的地 PE 的广告的 GFL，将在任意 AC 上进入的所有 BUM 业务转发到目的地 PE
230
在源 PE 路由器处，利用由目的地 PE 所分发 MHFLx，对在 ESIx 的任意非 DF AC 处进入的所有 BUM 业务进行标记
240
在目的地 PE 路由器处，将在 P2MP 上接收到的所有数据包泛洪到除了非 DF AC 之外的所有本地 MVI 端点
250

图 2

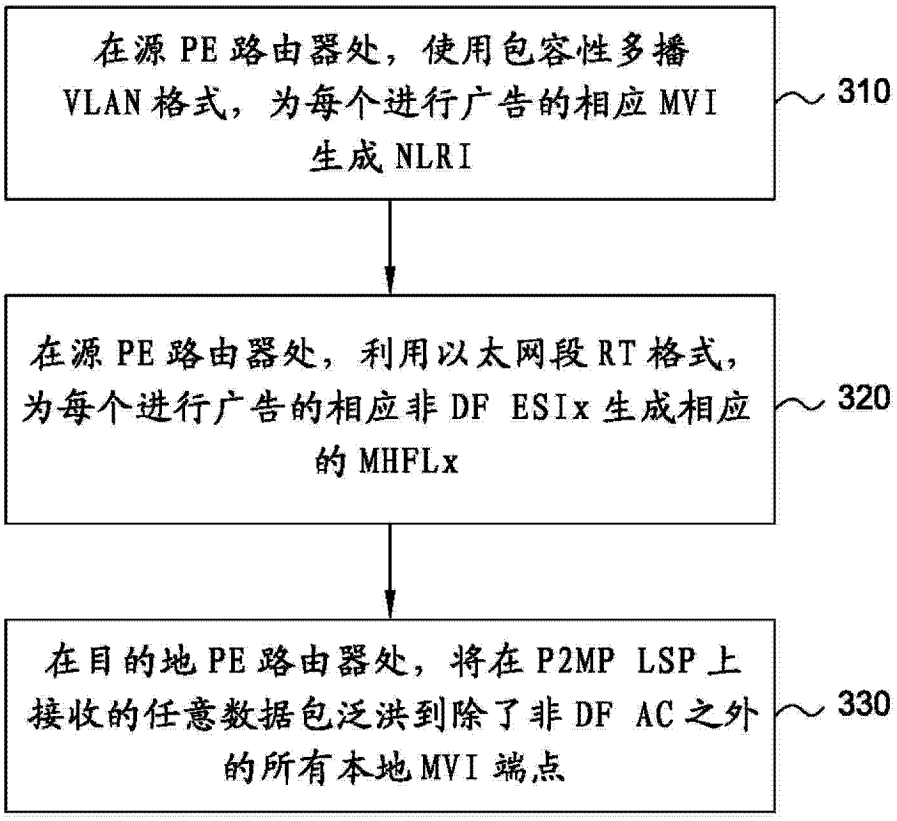

在源 PE 路由器处，使用包容性多播 VLAN 格式，为每个进行广告的相应 MVI 生成 NLRI
310
在源 PE 路由器处，利用以太网段 RT 格式，为每个进行广告的相应非 DF ESIx 生成相应的 MHFLx
320
在目的地 PE 路由器处，将在 P2MP LSP 上接收的任意数据包泛洪到除了非 DF AC 之外的所有本地 MVI 端点
330

图 3

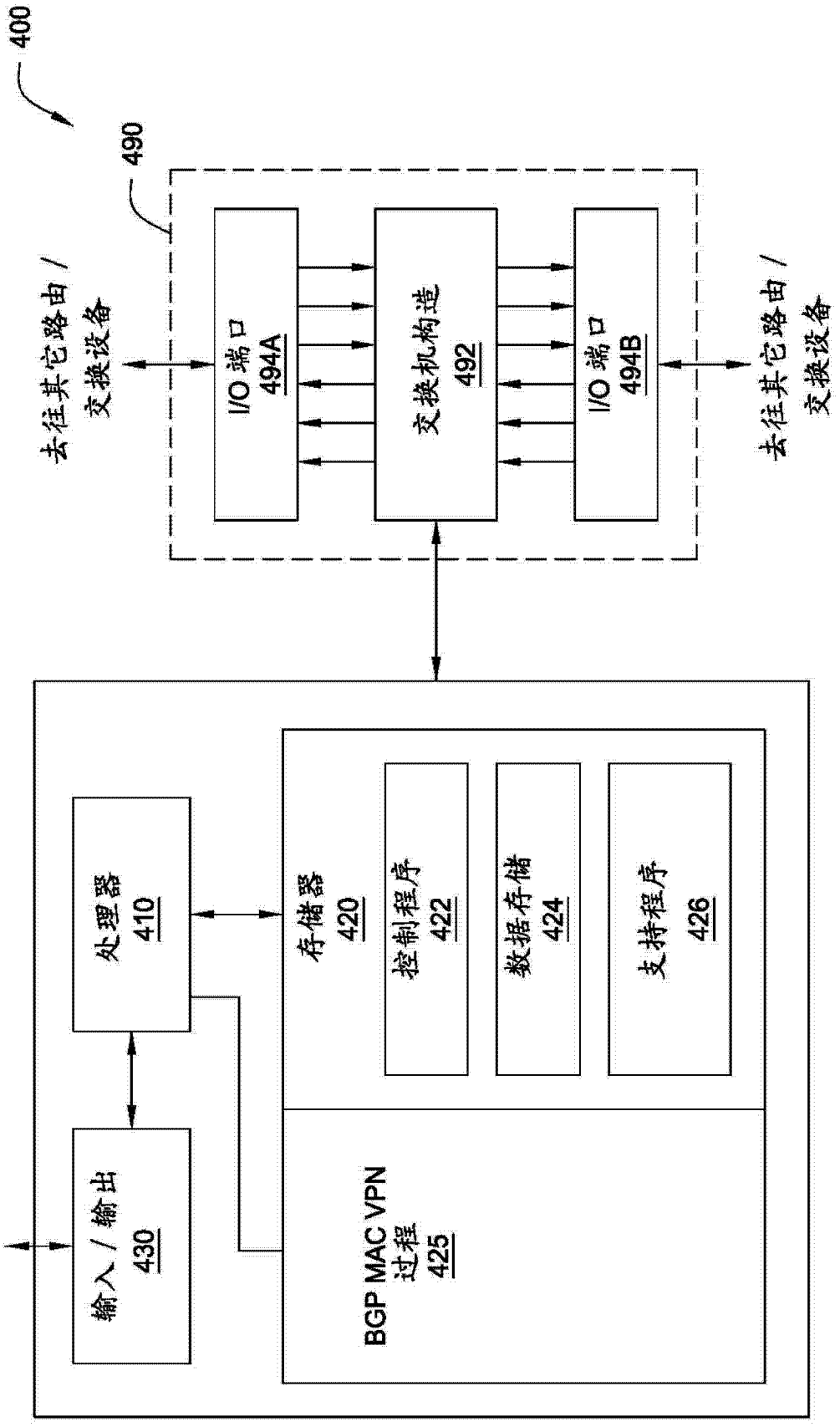

400
490
去往其它路由 / 交换设备
I/O 端口 494A
交换机构造 492
I/O 端口 494B
去往其它路由 / 交换设备
处理器 410
存储器 420
控制程序 422
数据存储 424
支持程序 426
输入 / 输出 430
BGP MAC VPN 过程 425

图 4

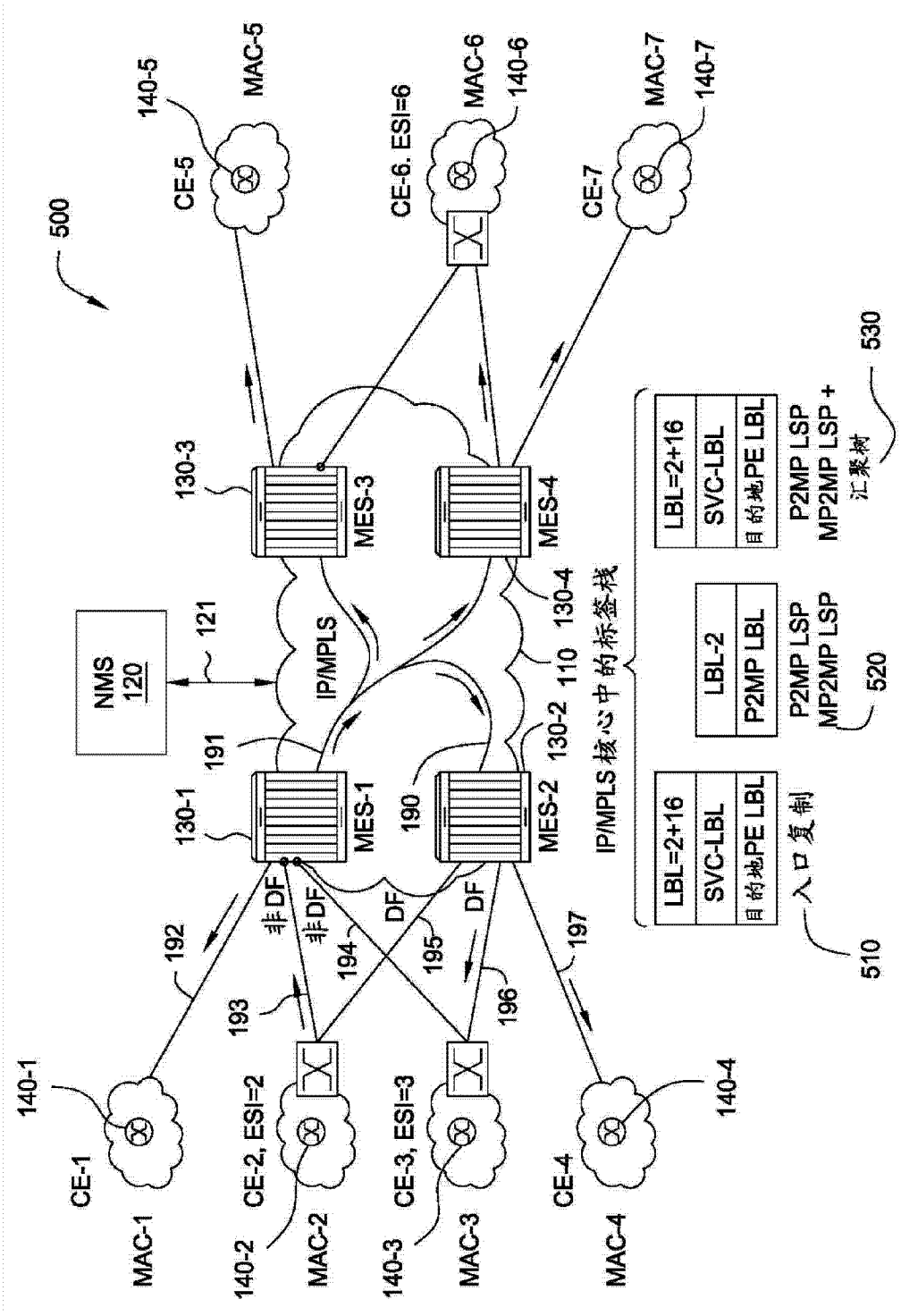

500
NMS 120
121
IP/MPLS
110
MES-1
MES-2
MES-3
MES-4
130-1
130-2
130-3
130-4
191
190
192
193
194
195
196
197
非DF
非DF
DF
DF
CE-1
CE-2, ESI=2
CE-3, ESI=3
CE-4
CE-5
CE-6. ESI=6
CE-7
MAC-1
MAC-2
MAC-3
MAC-4
MAC-5
MAC-6
MAC-7
140-1
140-2
140-3
140-4
140-5
140-6
140-7
IP/MPLS核心中的标签栈
LBL=2+16
SVC-LBL
目的地PE LBL
LBL-2
P2MP LBL
P2MP LSP
MP2MP LSP
P2MP LSP
MP2MP LSP +
汇聚树
入口复制
510
520
530

图 5

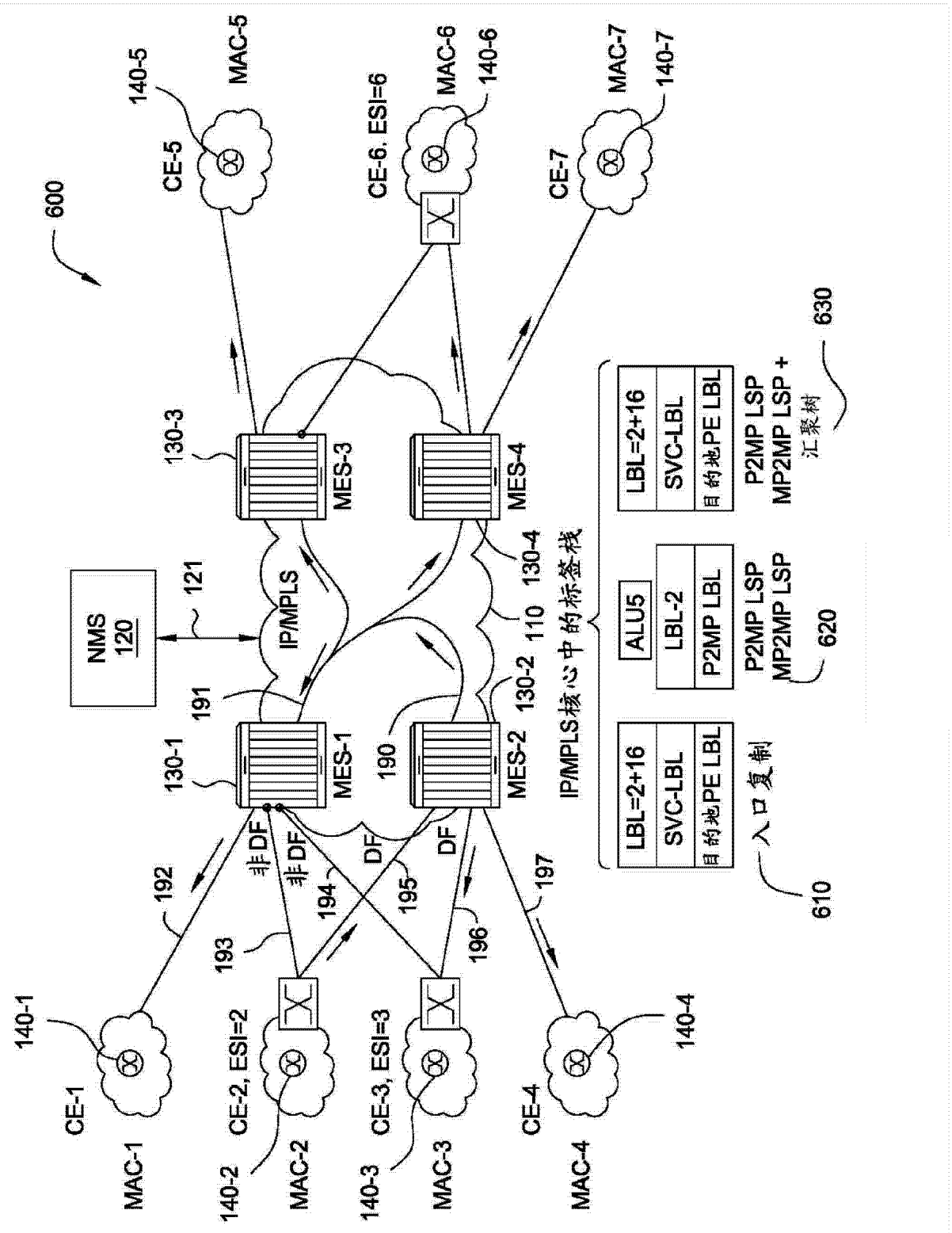

600
NMS 120
121
IP/MPLS
MES-1
MES-2
MES-3
MES-4
130-1
130-2
130-3
130-4
110
191
190
CE-1
CE-2, ESI=2
CE-3, ESI=3
CE-4
CE-5
CE-6, ESI=6
CE-7
MAC-1
MAC-2
MAC-3
MAC-4
MAC-5
MAC-6
MAC-7
140-1
140-2
140-3
140-4
140-5
140-6
140-7
非DF
非DF
DF
DF
192
193
194
195
196
197
IP/MPLS核心中的标签栈
LBL=2+16
SVC-LBL
目的地PE LBL
ALU5
LBL-2
P2MP LBL
P2MP LSP
MP2MP LSP
620
P2MP LSP
MP2MP LSP +
汇聚树
630
入口复制
610

图 6

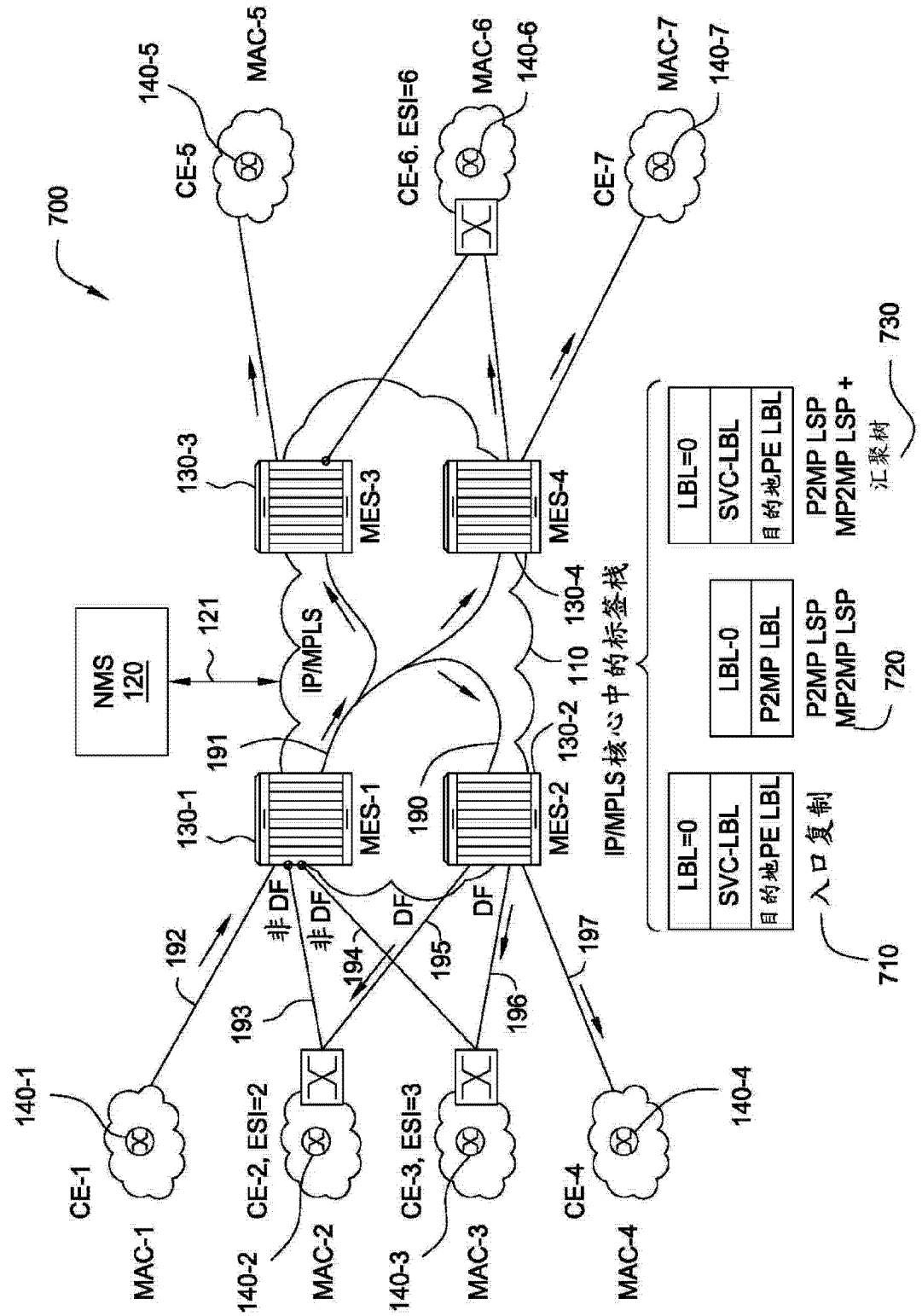

700
CE-5
MAC-5
140-5
CE-6, ESI=6
MAC-6
140-6
CE-7
MAC-7
140-7
130-3
MES-3
MES-4
130-4
NMS
120
121
IP/MPLS
110
191
130-1
MES-1
190
MES-2
130-2
IP/MPLS核心中的标签栈
LBL=0
SVC-LBL
目的地PE LBL
LBL-0
P2MP LBL
P2MP LSP
MP2MP LSP
P2MP LSP +
MP2MP LSP +
汇聚树
730
720
710
入口复制
非 DF
非 DF
DF
DF
192
193
194
195
196
197
CE-1
MAC-1
140-1
CE-2, ESI=2
MAC-2
140-2
CE-3, ESI=3
MAC-3
140-3
CE-4
MAC-4
140-4

图 7