(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113098752 B
(45)授权公告日 2022.07.22

(21)申请号 202110398723.1
(22)申请日 2016.10.29
(65)同一申请的已公布的文献号
申请公布号 CN 113098752 A
(43)申请公布日 2021.07.09
(30)优先权数据
14/956,716 2015.12.02 US
14/956,736 2015.12.02 US
14/956,756 2015.12.02 US
14/956,775 2015.12.02 US
(62)分案原申请数据
201680070479.3 2016.10.29
(73)专利权人 NICIRA股份有限公司
地址 美国加利福尼亚
(72)发明人 沈建军 A•泰丝莫 M•海拉 P•萨卡尔 王华
(74)专利代理机构 中国贸促会专利商标事务所有限公司 11038
专利代理师 鲍进
(51)Int.Cl.
H04L 12/46(2006.01)
H04L 47/125(2022.01)
H04L 101/622(2022.01)
(56)对比文件
CN 103457818 A,2013.12.18
US 2014059111 A1,2014.02.27
US 2015109923 A1,2015.04.23
US 2013058350 A1,2013.03.07
CN 104995880 A,2015.10.21
CN 104272668 A,2015.01.07
US 2015009995 A1,2015.01.08
US 2014269702 A1,2014.09.18
US 2015124822 A1,2015.05.07
审查员 王璐

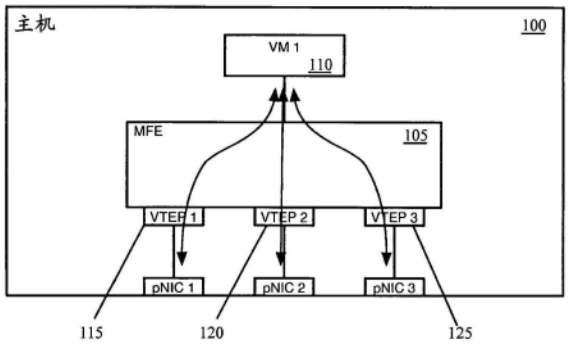
权利要求书3页 说明书22页 附图20页

(54)发明名称

实现多个隧道端点上负载平衡的方法、设备、系统和介质

(57)摘要

本公开涉及多个隧道端点上的负载平衡。一些实施例提供了一种用于受管理转发元件(MFE)的方法。该方法从MFE为其执行第一跳处理的数据计算节点接收分组。数据计算节点与MFE的多个隧道端点相关联。该方法确定用于分组的目的地隧道端点。该方法使用负载平衡算法来选择MFE的所述多个隧道端点之一作为用于分组的源隧道端点。该方法使用源隧道端点和目的地隧道端点封装隧道中的分组。



CN 113098752 B

1.一种用于在第一主机计算机上执行的受管理转发元件MFE的方法,包括:

分别从在与多个隧道端点相关联的第二主机计算机上的相应的第一和第二隧道端点接收第一和第二分组,所述第一和第二分组分别源自于在所述第二主机计算机上执行的第一数据计算节点DCN和第二DCN;

基于从所述第一和第二隧道端点接收到所述第一和第二分组,通过存储所述第一隧道端点与所述第一DCN的第一关联和所述第二隧道端点与所述第二DCN的第二关联,分别为所述第一和第二DCN选择所述第一和第二隧道端点作为用于从所述MFE到所述第一和第二DCN的后续分组的目的地隧道端点;以及

对于在MFE处接收的并且将第一DCN或第二DCN作为目的地的后续分组,使用所存储的第一关联或第二关联从多个隧道端点中选择第一隧道端点或第二隧道端点,并且使用识别为目的地隧道端点的所选隧道端点封装所述后续分组。

2.如权利要求1所述的方法,其中:

所存储的第一关联将第一隧道端点映射到以下之一:(i)第一DCN的MAC地址,和(ii)以第一DCN作为目的地的分组的连接5元组;

所存储的第二关联将所述第二隧道端点映射到以下之一:(i)所述第二DCN的MAC地址,和(ii)以所述第二DCN作为目的地的分组的连接5元组;以及

每个连接5元组包括源和目的地网络层地址、源和目的地传输层端口号以及传输层协议。

3.如权利要求1所述的方法,还包括将所述第一分组递送到在所述第一主机计算机上执行并且连接到所述MFE的第三DCN。

4.如权利要求3所述的方法,其中在所述MFE的第三隧道端点处接收所述第一分组,其中存储所述第一关联包括存储识别所述第一和第三DCN以及所述第一和第三隧道端点的记录,所述方法还包括:

从所述第三DCN接收寻址到所述第一DCN的第三分组;

使用所述记录来选择所述MFE的第三隧道端点作为源隧道端点;以及

利用作为源隧道端点的第三隧道端点和作为目的地隧道端点的第一隧道端点封装所述第三分组。

5.如权利要求1所述的方法,其中所述第一分组包括一组外部报头,所述一组外部报头具有作为源地址的所述第一隧道端点的地址。

6.如权利要求1所述的方法,其中所述多个隧道端点是第一多个隧道端点,所述方法还包括:

在所述MFE处,从与第二多个隧道端点相关联的第三隧道端点接收第三分组,所述第三分组源自于第三DCN;

基于所述第三分组,存储所述第三隧道端点与所述第三数据计算节点的关联;以及

使用所存储的第三隧道端点与第三数据计算节点的关联从第二多个隧道端点中选择第三隧道端点并且利用第三隧道端点作为目的地隧道端点来封装后续分组,所述后续分组在MFE处接收并且将第三DCN作为目的地地址。

7.如权利要求6所述的方法,其中所述第一、第二和第三DCN在同一第二主机计算机上操作。

8.一种用于网络控制器的方法,包括:

识别用于与受管理转发元件MFE一起在主机计算机上执行的数据计算节点DCN,所述MFE具有多个隧道端点,所述多个隧道端点用于与其他主机计算机上的其他MFE建立隧道以建立逻辑网络;以及

向所述MFE分发(i)与多个隧道端点中的一组至少两个隧道端点相关联的隧道端点组标识符,所述至少两个隧道端点是要用于DCN的候选隧道端点,以及(ii)组定义,所述组定义包括标识所述隧道端点组中的隧道端点的至少两个隧道端点标识符,

其中所述MFE(i)使用所述组标识符来标识所述组定义,(ii)从所述组中选择所述隧道端点标识符中的一个,以及(iii)使用与所选隧道端点标识符相关联的隧道端点来将封装后的分组从所述DCN发送到其他MFE。

9.如权利要求8所述的方法,其中:

识别所述DCN包括从本地控制器接收所述DCN已经附接到所述MFE的消息;

本地控制器管理所述MFE;以及

本地控制器和MFE以主机计算机的虚拟化软件操作。

10.如权利要求8所述的方法,还包括:

接收逻辑网络配置;以及

使用所接收的逻辑网络配置来识别所述逻辑网络的其他DCN。

11.如权利要求8所述的方法,还包括:

为多个DCN中的每个DCN在第一表中存储映射,所述映射将DCN映射到隧道端点组标识符;以及

为每个隧道端点组标识符在第二表中存储组定义,所述组定义包括至少两个隧道端点标识符,所述至少两个隧道端点标识符标识在由隧道端点组标识符标识的隧道端点组中的至少两个隧道端点。

12.如权利要求11所述的方法,其中属于第一逻辑网络的第一DCN和属于第二逻辑网络的第二DCN映射到同一隧道端点组标识符。

13.如权利要求8所述的方法,其中将所述隧道端点组标识符和组定义分发到所述MFE包括:将所述隧道端点组标识符和组定义分发到管理所述MFE并且在所述主机计算机上执行的本地控制器。

14.如权利要求13所述的方法,其中所述本地控制器将所接收的隧道端点组标识符和组定义转换为所述MFE可读的配置数据。

15.一种用于第一受管理转发元件MFE的方法,包括:

从连接到所述MFE的第一数据计算节点DCN接收分组,所述分组具有与远程网络中的第二DCN对应的目的地地址;

(i)将目的地地址映射到一组至少两个MFE的组标识符,所述至少两个MFE形成用于向远程网络发送分组的桥接器集群,并且(ii)将所述桥接器集群的组标识符映射到该组MFE的多个隧道端点,其中桥接器集群中的每个MFE具有多个隧道端点中的至少一个,并作为单独的转发元件操作,所述转发元件既连接到远程网络又连接到第一MFE的本地网络;

选择多个隧道端点之一作为所述分组的目的地隧道端点;和

利用与第一MFE相关联的源隧道端点和所选目的地隧道端点封装所述分组。

16.如权利要求15所述的方法,其中:

该组MFE中的形成桥接器集群的MFE被配置为将分组从本地网络桥接到远程网络;和

本地网络是第一层2网络,并且远程网络是第二层2网络。

17.如权利要求16所述的方法,其中:

第一层2网络是逻辑覆盖网络,并且第二层2网络是虚拟局域网(VLAN);和

封装分组包括使用隧道协议进行封装。

18.如权利要求15所述的方法,其中选择所述隧道端点之一包括:

计算所述分组的分组报头值集合的散列,所述报头值集合至少包括分组的目的地地址;和

基于所述散列,将分组报头值集合分配给多个隧道端点中的特定隧道端点。

19.如权利要求18所述的方法,其中所述分组报头值集合包括所述分组的源和目的地网络层地址、源和目的地传输层端口号以及传输层协议。

20.如权利要求15所述的方法,其中:

多个隧道端点是第一多个隧道端点;

第二多个隧道端点与第一MFE相关联;和

所述方法还包括从第二多个隧道端点中选择源隧道端点。

21.如权利要求15所述的方法,还包括:将封装后的分组发送到所述源隧道端点与所述目的地隧道端点之间的物理网络上。

22.一种存储程序的机器可读介质,所述程序在由至少一个处理单元实现时实现如权利要求1-21中任一项所述的方法。

23.一种电子设备,包括:

一组处理单元;和

存储程序的机器可读介质,当由所述处理单元中的至少一个实现时,所述程序实现如权利要求1-21中任一项所述的方法。

24.一种系统,包括用于实现如权利要求1-21中任一项所述的方法的装置。

实现多个隧道端点上负载平衡的方法、设备、系统和介质

[0001] 本申请是申请日为2016年10月29日、申请号为 201680070479.3、发明名称为“多个隧道端点上的负载平衡”的发明专利申请的分案申请。

技术领域

[0002] 本公开通常涉及多个隧道端点上的负载平衡。

背景技术

[0003] 在数据中心内，覆盖网络常常被用于属于同一逻辑网络的虚拟机、服务器等等之间的通信。为了实现这些覆盖，通常在数据中心中的转发元件(例如，虚拟交换机、顶架式(TOR)交换机等等)之间建立IP上的MAC(MAC over IP)隧道。当这些转发元件各自具有单个网络接口时，它们各自被指派单个隧道端点IP地址，并且两个转发元件之间的分组的封装使用两个转发元件的隧道端点IP地址。

[0004] 但是，越来越多地，转发元件可以具有多个这样的隧道端点地址 (例如，以充分利用多个物理接口或L2/L3链路)。因此，管理这些转发元件的控制器以及转发元件本身应当具有用于处理每个转发元件的多个隧道端点地址的方案。

发明内容

[0005] 一些实施例提供了一种用于在源受管理转发元件(MFE)处封装隧道中的分组以传送到目的地受管理转发元件的方法。源MFE (例如，从本地数据计算节点)接收数据分组并且识别该分组需要经由覆盖网络传输到目的地MFE(例如，基于分组的一个或多个目的地地址，诸如MAC和/或IP地址)。源MFE识别用于将分组隧道传输到目的地MFE的源和目的地隧道端点，这可能涉及在用于源和/ 或目的地MFE的多于一个可能的隧道端点之间进行选择。

[0006] 例如，如果源MFE具有可以用于来自源地址的数据分组的多个隧道端点，那么一些实施例的源MFE执行负载平衡技术，以选择源隧道端点中的一个。例如，不同的实施例可以使用源MAC和/或IP 地址的散列、标准连接5元组(源和目的地IP地址，源和目的地传输层端口以及传输协议)的散列或者评估不同隧道端点上的当前负载 (例如，队列的充满度等)以选择源隧道端点的机制。

[0007] 此外，如果目的地MFE具有可以使用的多个隧道端点，那么源 MFE使用类似的机制(例如，使用目的地MAC地址或连接5元组的散列)来选择目的地隧道端点。利用所选择的源和目的地隧道端点， MFE可以封装分组并将分组发送到两个端点之间的物理网络上。

[0008] 为了执行这种选择，在一些实施例中，每个MFE(为直接连接到MFE的本地数据计算节点以及处于与本地数据计算节点相同的逻辑网络上的地址)存储数据计算节点地址到隧道端点组标签的映射以及隧道端点组标签到隧道端点列表的映射。在一些实施例中，可以经由收集这些映射并将其存储在表中的网络控制系统将这种映射信息分发到MFE。

[0009] 使用这种技术，两个数据计算节点之间的分组可以从MFE上的一个隧道端点发出，但是在MFE的不同隧道端点处接收。但是，为了确保用于特定数据计算节点(或特定流)的分

组都通过相同的隧道端点发送和接收，一些实施例的MFE学习数据计算节点地址到特定隧道端点的映射。为了允许每个MFE负载平衡其自己的流量，当第一MFE在其本地数据计算节点之一和远程数据计算节点之间向第二 MFE发送初始分组时，第一MFE将其选择的源隧道端点存储为用于其本地数据计算节点的地址的隧道端点。

[0010] 但是，第一MFE不存储它对用于远程数据计算地址的目的地隧道端点的选择，因为这样做将允许第一MFE执行对第二MFE的负载平衡。相反，当第二MFE接收到初始分组时，它还将由第一 MFE选择的源隧道端点存储为到源数据计算节点地址的映射。当其本地数据计算节点发送回复分组时，第二MFE执行其自己的负载平衡，以选择其源隧道端点，这可以与在其上接收初始分组的隧道端点不同。第二MFE然后存储其本地数据计算节点和选择的隧道端点之间的这种映射。对于这个回复分组的目的地隧道端点，第二MFE使用它从初始分组确定的映射。当第一MFE接收到回复分组时，它存储源隧道端点和远程数据计算节点之间的映射。在这个时候，达到平衡，两个MFE在同一对隧道端点之间发送分组。

[0011] 除了通过隧道将分组发送到具有多个隧道端点的MFE以便到达远程数据计算节点之外，在一些情况下，MFE还从其本地数据计算节点之一接收需要桥接到另一个远程网络(例如，另一个逻辑网络、VLAN等等)的分组。一些实施例使用桥接器集群来执行这种桥接，其可以包括多个MFE(其中的一些或全部可以具有多个隧道端点)。与桥接器集群相关联的所有隧道端点可以被分组，并且源MFE执行类似的选择过程(例如，对目的地地址或连接5元组进行散列)，以选择桥接器集群组中要向其发送分组的隧道端点。

[0012] 前面的发明内容旨在用作针对本发明的一些实施例的简要介绍。它并不意味着是本文档中所公开的所有发明主题的介绍或概述。以下的具体实施方式和参考具体实施方式的附图将进一步描述在发明内容以及其它实施例中所描述的实施例。因此，为了理解本文档所描述的所有实施例，需要对发明内容、具体实施方式和附图进行全面地审阅。此外，所要求保护的主题不受在发明内容、具体实施方式和附图中的说明性细节的限制，而是要由所附权利要求来定义，这是因为所要求保护的主题可以在不脱离本主题的精神的情况下以其它特定形式来体现。

附图说明

[0013] 本发明的新颖特征在所附权利要求中阐述。但是，为了解释的目的，本发明的几种实施例在以下图中阐述。

[0014] 图1概念性地图示了包括具有多个隧道端点的MFE的一些实施例的主机机器。

[0015] 图2概念性地图示了一些实施例的主机机器，其托管连接到在该主机机器上操作并具有两个VTEP的MFE的两个VM。

[0016] 图3概念性地图示了通过物理网络(例如，数据中心的物理网络) 连接的一对主机机器。

[0017] 图4-8图示了通过VM连接到的MFE对两个VM之间的分组的处理以及通过这些MFE的隧道端点映射的学习。

[0018] 图9概念性地图示了由MFE执行的一些实施例的处理，该处理在封装分组并将那个分组输出到物理网络上之前选择源和目的地隧道端点。

[0019] 图10概念性地图示了由MFE执行的一些实施例的处理，该处理在其一个(也可能是

多个)VTEP处接收封装的分组。

[0020] 图11图示了通过物理网络连接并且具有基于每个流进行负载平衡的MFE的一对主机机器。

[0021] 图12图示了第三主机机器,其中第三VM和第三MFE连接到图11的相同物理数据中心网络。

[0022] 图13概念性地图示了使用负载评估来指派用于流的源VTEP。

[0023] 图14概念性地图示了由MFE执行的一些实施例的处理,该处理选择要发送到远程网络的分组的源和目的地隧道端点。

[0024] 图15概念性地图示了包括提供第一VM所位于的第一网络与第二网络之间的桥接连接性的桥接器集群的数据中心网络的示例。

[0025] 图16概念性地图示了从图15的VM向桥接网络中的MAC地址 (MAC1)发送分组。

[0026] 图17概念性地图示了从图15的VM向桥接网络中的不同MAC 地址(MAC2)发送分组。

[0027] 图18概念性地图示了一些实施例的网络控制系统以及这个网络控制系统内的数据传送,因为它涉及在MFE上配置的映射表。

[0028] 图19概念性地图示了由一些实施例的中央控制平面执行的处理,该处理搜集并分发关于隧道端点到数据计算节点地址到受管理转发元件的映射的数据。

[0029] 图20概念性地图示了利用其实现本发明的一些实施例的电子系统。

具体实施方式

[0030] 在本发明的以下详细描述中,阐述和描述了本发明的许多细节、示例和实施例。但是,对本领域技术人员而言清楚明显的是,本发明不限于所阐述的实施例,并且本发明可以在没有所讨论的一些具体细节和示例的情况下实践。

[0031] 一些实施例提供了一种用于在源受管理转发元件(MFE)处封装隧道中的分组以传送到目的地受管理转发元件的方法。源MFE (例如,从本地数据计算节点)接收数据分组并且识别该分组需要经由覆盖网络传输到目的地MFE(例如,基于分组的一个或多个目的地地址,诸如MAC和/或IP地址)。源MFE识别用于将分组隧道传输到目的地MFE的源和目的地隧道端点,这可能涉及在用于源和/ 或目的地MFE的多于一个可能的隧道端点之间进行选择。

[0032] 例如,如果源MFE具有可以用于来自源地址的数据分组的多个隧道端点,那么一些实施例的源MFE执行负载平衡技术,以选择源隧道端点中的一个。例如,不同的实施例可以使用源MAC和/或IP 地址的散列、标准连接5元组(源和目的地IP地址,源和目的地传输层端口以及传输协议)的散列或者评估不同隧道端点上的当前负载 (例如,队列的充满度等)以选择源隧道端点的机制。

[0033] 此外,如果目的地MFE具有可以使用的多个隧道端点,那么源MFE使用类似的机制(例如,使用目的地MAC地址或连接5元组的散列)来选择目的地隧道端点。利用所选择的源和目的地隧道端点, MFE可以封装分组并将分组发送到两个端点之间的物理网络上。

[0034] 为了执行这种选择,在一些实施例中,每个MFE(为直接连接到MFE的本地数据计算节点以及处于与本地数据计算节点相同的逻辑网络上的地址)存储数据计算节点地址到隧道端点组标签的映射以及隧道端点组标签到隧道端点列表的映射。在一些实施例中,可以经由收集这些映射并将其存储在表中的网络控制系统将这种映射信息分发到MFE。

[0035] 使用这种技术,两个数据计算节点之间的分组可以从MFE上的一个隧道端点发出,但是在MFE的不同隧道端点处接收。但是,为了确保用于特定数据计算节点(或特定流)的分组都通过相同的隧道端点发送和接收,一些实施例的MFE学习数据计算节点地址到特定隧道端点的映射。为了允许每个MFE负载平衡其自己的流量,当第一MFE在其本地数据计算节点之一和远程数据计算节点之间向第二 MFE发送初始分组时,第一MFE将其选择的源隧道端点存储为用于其本地数据计算节点的地址的隧道端点。

[0036] 但是,第一MFE不存储它对用于远程数据计算地址的目的地隧道端点的选择,因为这样做将允许第一MFE执行对第二MFE的负载平衡。相反,当第二MFE接收到初始分组时,它还将由第一 MFE选择的源隧道端点存储为到源数据计算节点地址的映射。当其本地数据计算节点发送回复分组时,第二MFE执行其自己的负载平衡,以选择其源隧道端点,这可以与在其上接收初始分组的隧道端点不同。第二MFE然后存储其本地数据计算节点和选择的隧道端点之间的这种映射。对于这个回复分组的目的地隧道端点,第二MFE使用它从初始分组确定的映射。当第一MFE接收到回复分组时,它存储源隧道端点和远程数据计算节点之间的映射。在这个时候,达到平衡,两个MFE在同一对隧道端点之间发送分组。

[0037] 除了通过隧道将分组发送到具有多个隧道端点的MFE以便到达远程数据计算节点之外,在一些情况下,MFE还从其本地数据计算节点之一接收需要桥接到另一个远程网络(例如,另一个逻辑网络、VLAN等等)的分组。一些实施例使用桥接器集群来执行这种桥接,其可以包括多个MFE(其中的一些或全部可以具有多个隧道端点)。与桥接器集群相关联的所有隧道端点可以被分组,并且源MFE执行类似的选择过程(例如,对目的地地址或连接5元组进行散列),以选择桥接器集群组中要向其发送分组的隧道端点。

[0038] 以上介绍了受管理转发元件针对特定地址在多个隧道端点之间进行选择的概念。在下文中,第I部分重点在于具有多个隧道端点的 MFE之间的通信,而第II部分描述具有多个MFE(MFE具有多个隧道端点)的桥接器集群的示例。然后第III部分描述通过网络控制系统对MFE的配置。最后,第IV部分描述利用其实现本发明的一些实施例的电子系统。

[0039] I.具有多个隧道端点的MFE

[0040] 如上面所提到的,一些实施例提供了用于为具有多于一个隧道端点的受管理转发元件(MFE)选择源和目的地隧道端点的技术。图1 概念性地图示了一些实施例的主机机器100,其包括具有多个隧道端点的MFE 105。在这个示例中,主机机器100可以是数据中心(诸如企业数据中心、多租户数据中心等等)中的许多主机机器之一。在这个示例中,该图图示了虚拟机(VM)110,但应当理解的是,典型的主机机器将托管许多(例如,数十、数百等等)数据计算节点 (诸如VM、容器等等)。在一些实施例中,VM 110属于在数据中心内(或跨越多个数据中心)实现的逻辑网络,并且数据中心中的 MFE使用覆盖网络进行不同逻辑网络的通信。这些覆盖网络使用隧道端点的地址跨居间物理网络发送分组,其中封装还包括逻辑网络标识符(LNI)。LNI指示分组属于哪个覆盖网络(例如,哪个逻辑网络或逻辑转发元件),并且允许接收MFE根据由LNI指定的逻辑网络正确地处理分组。

[0041] MFE 105具有三个虚拟隧道端点(VTEP)115-125,其有效地充当MFE的面向外的端口。在一些实施例中,MFE 105是在主机 100的虚拟化软件(例如,虚拟交换机(诸如Open vSwitch、ESX 等等))中实现的软件转发元件。在一些实施例中,VTEP是虚拟交换机的一部分,而在其它实施例中,VTEP是单独的软件实体,但可以被认为是虚拟交换机的扩展(例如,

虚拟交换机与物理网络接口之间的I/O链的一部分，通过该物理网络接口，主机连接到数据中心的物理网络)。在任一种情况下，VTEP在本文中都将被称为MFE的一部分，其中MFE执行源和目的地VTEP的选择。

[0042] 三个VTEP 115-125与三个分离的物理网络接口控制器(pNIC) 对应，它们是连接到数据中心的物理网络的主机机器100的物理接口。从VM 110发送的分组通过这些pNIC输出到这个物理网络上，并且通过这些pNIC接收发送到VM 110的分组。但是，在其它实施例中，隧道端点可以不具有与物理接口的1:1相关性(例如，将多个 VTEP与单个pNIC相关联)。

[0043] 如图1中所示，由VM 110发送的分组将由MFE 105接收(例如，通过VM 110连接到的MFE 105的虚拟端口)，并且可以经由 VTEP 115-125中的任何一个从MFE发出。如下面更详细描述的，MFE可以使用各种标准来确定三个VTEP 115-125中的哪一个用作源VTEP以在将分组发送到物理网络上之前封装这种分组，诸如不同分组特点的散列或不同VTEP的当前负载的评估(例如，查看分组缓冲区或队列)。类似地，可以基于由这种分组的源使用的各种可能标准通过VTEP 115-125中的任何一个来接收用于VM 110的分组。在这种情况下，MFE对分组进行解封装，从LNI中识别该分组属于 VM 110附连到的逻辑网络，并且根据用于逻辑网络的转发表将分组递送到VM。

[0044] 如上面所提到的，一些实施例的主机机器通常将托管多个数据计算节点，诸如多个虚拟机、多个容器或其组合。在一些实施例中，这些数据计算节点可以共享多个隧道端点的使用。图2概念性地图示了托管两个VM 205和210的主机机器200。这些VM 205和210连接到在主机机器200上操作并且具有两个VTEP 220和225的MFE 215。在这种情况下，如图中所示出的，用于VM1 205的流量可以通过两个VTEP 220和225中的任何一个流入和流出，并且用于VM2 210的流量也可以通过这两个VTEP 220和225中的任何一个流入和流出。

[0045] 在不同的负载平衡技术下，由两个VM 205和210对分组使用两个VTEP 220和225将是不同的。例如，如果MFE基于其MAC地址的散列将每个VM指派给VTEP，那么或者VM 205和210两者将共享单个VTEP，或者它们将各自使用两个VTEP中的一个。另一方面，如果指派是基于流的(例如，基于连接5元组的散列)，那么最有可能每个VM将其流量在VTEP之间拆分。类似地，当MFE 215 评估VTEP上的负载并基于这个负载将流指派到不同的VTEP时，用于每个VM的流量将可能在两个VTEP 220和225之间以一种方式或另一种方式拆分。此外，两个VM和两个VTEP之间的流量分布可以随时间而改变。

[0046] 但是，在一些情况下，不是所有的隧道端点都可用于主机上的所有数据计算节点。在一些情况下，具有多个隧道端点的主机可以属于物理数据中心网络的多于一个指定的区，其中其一些隧道端点连接到主机上的第一MFE，并且其它隧道端点连接到主机上的第二MFE，并且那两个(或更多个)MFE指派给不同的区。在这种情况下，数据计算节点可以分别连接到被指派给数据计算节点的逻辑网络所属的区的特定MFE。

[0047] A.MFE之间的通信

[0048] 虽然先前的示例说明了具有单个MFE(该MFE具有多个隧道端点)的单个主机机器，但实际上经由覆盖网络发送的分组具有源和目的地隧道端点。在一些情况下，用于特定分组的源和目的地MFE (即，连接到分组源的MFE，其充当用于处理分组的第一跳MFE，以及第一跳MFE向其隧道传输分组的MFE)具有多个隧道端点，在这种情况下，源MFE需要选择其自己的隧道端点之一以及可能的目的地隧道端点之一。

[0049] 如上面所提到的，一些实施例的源MFE执行负载平衡技术，以选择源隧道端点之一。例如，不同的实施例可以使用源MAC和/或 IP地址的散列、标准连接5元组的散列，或者评估不同隧道端点上的当前负载(例如，队列的充满度等等)的机制，以选择源隧道端点。此外，如果目的地MFE具有可以使用的多个隧道端点，那么源 MFE使用类似的机制(例如，计算某些分组特点的散列值)来选择目的地隧道端点。利用所选择的源和目的地隧道端点两者，MFE可以封装分组并将分组发送到两个端点之间的物理网络上。

[0050] 为了执行这种选择，在一些实施例中，每个MFE为直接连接到 MFE的本地数据计算节点以及处于与本地数据计算节点相同的逻辑网络上的地址二者存储数据计算节点地址到隧道端点组标签的映射，以及存储隧道端点组标签到隧道端点列表的映射。

[0051] 为了确保用于特定数据计算节点(或特定流)的分组都通过相同的隧道端点发送和接收，一些实施例的MFE学习数据计算节点地址到特定隧道端点的映射。为了允许每个MFE对其自己的流量进行负载平衡，当第一MFE在其本地数据计算节点之一和远程数据计算节点之间向第二MFE发送初始分组时，第一MFE将其选择的源隧道端点存储为用于其本地数据计算节点的地址的隧道端点。

[0052] 但是，第一MFE不存储它对用于远程数据计算地址的目的地隧道端点的选择。相反，当第二MFE接收到初始分组时，第二MFE 还将由第一MFE选择的源隧道端点存储为到源数据计算节点地址的映射。当其本地数据计算节点发送回复分组时，第二MFE执行其自己的负载平衡，以选择其源隧道端点，这可以与接收初始分组的隧道端点不同。第二MFE然后存储其本地数据计算节点和选择的隧道端点之间的这种映射。对于这个回复分组的目的地隧道端点，第二 MFE使用它从初始分组确定的映射。当第一MFE接收到回复分组时，它存储源隧道端点和远程数据计算节点之间的映射。在这个时候，达到平衡，两个MFE在同一对隧道端点之间发送分组。

[0053] 图3概念性地图示了通过物理网络335(例如，数据中心的物理网络)连接的一对主机机器300和350。主机机器300包括连接到也在主机机器300上运行的MFE 310的VM 305(最有可能除了多个其它数据计算节点之外还包括该VM 305)。MFE 310具有两个 VTEP 315和320。类似地，第二主机机器350包括连接到也在主机机器350上运行的MFE 360的VM 355。这个MFE 360具有两个 VTEP 365和370。

[0054] 此外，每个MFE存储用于转发分组和执行各种其它分组处理操作(包括VTEP的选择)的各种表。在这种情况下，附图图示了第一MFE 310存储将VTEP组映射到VTEP列表的第一表325和将数据计算节点地址(在这种情况下为VM)映射到VTEP组和那些组内的各个VTEP的第二表330。第一表325示出了与MFE 310(具有 VTEP1和VTEP2)对应的第一组以及与MFE 360(具有VTEP3和 VTEP4)对应的第二组。第二表330示出了映射到用于其相应的 MFE 310和360的VTEP组的两个VM 305和310。在一些实施例中，这些表由从中央控制器传递(并由本地控制器安装在MFE上)的信息填充，如下面第III部分中更详细示出的。在这个时候，第二MFE 360的表375和380包含相同的信息。

[0055] 这个物理网络335可以包括在两个主机机器之间转发分组的各种交换机和路由器。但是，物理网络中的这些转发元件不处理内部(逻辑网络)分组，而是仅根据外部报头来执行交换和/或路由。即，这些各种物理交换机和路由器负责根据源MFE添加的隧道端点地址在隧道端点之间传输分组。应当理解的是，典型的数据中心网络将包括多得多的主机，每

个主机具有多个数据计算节点。照此，由MFE在每个主机上存储的表将更加复杂，并且可能不完全相同(例如，第一 MFE上的一些记录在第二MFE上可能不需要)。

[0056] 下面的图4-8图示了由MFE 310和360对VM 305和355之间的分组的处理，以及由这些MFE对隧道端点映射的学习。图4图示了通过两个阶段405和410从第一主机机器300上的VM 305发送的第一分组400。如本文档中所使用的，分组是指跨网络发送的特定格式的位的集合。应当理解的是，术语分组可以在本文中用于指可以跨网络发送的各种格式化位集合，诸如以太网帧、IP分组、TCP分段、UDP数据报等等。虽然以下示例涉及分组，但应当理解的是，本发明不应当限于任何特定格式或类型的数据消息。

[0057] 这个分组400具有VM1的源地址和VM2的目的地地址。在这种情况下，源和目的地地址是层2(MAC)地址。应当理解的是，典型的分组还将包括源和目的地IP地址，但是在这里没有示出。事实上，如果两个VM 305和355连接到不同的逻辑交换机(例如，各自连接到同一个逻辑路由器)，那么由VM 305发送的目的地MAC 地址将不是VM 355的目的地MAC地址，而是逻辑路由器端口的目的地MAC地址。一旦MFE 310完成其逻辑网络处理，它就将已经执行了作为逻辑路由处理的一部分的MAC地址替换(以及可能的地址解析协议(ARP))，并且因此预封装的分组将具有逻辑路由器的源MAC地址以及目的地MAC地址，如这里图中所示。

[0058] MFE 310接收这个分组，执行各种处理，并使用表330来识别封装所需的源和目的地VTEP组。具体而言，基于源MAC地址 VM1，MFE 310识别出源VTEP组为MFE1，并且基于目的地MAC 地址VM2，识别出目的地VTEP组为MFE2。表325指示用于这些组中每一个的可能VTEP列表。

[0059] 在这个示例中，与基于流的负载平衡相反，MFE 310和360通过将特定的MAC地址指派给特定的VTEP在隧道端点上执行负载平衡。因此，如阶段405中所示，MFE计算源MAC地址VM1的散列。MFE计算散列值(H)，然后执行H模N的计算，其中N是该组中 VTEP的数量，以得出与VTEP之一对应的值。其它实施例可以使用类似的算法，这些算法对于出现故障(go down)的VTEP之一是有抵抗力的，而无需重新布置通过所有其它未受影响的VTEP发送的流量。因此，MAC地址VM1的散列提供VTEP2 320作为源隧道端点，并且MAC地址VM2的散列提供VTEP3 365作为分组400的目的地隧道端点。

[0060] 在第二阶段410中，MFE封装分组400并将封装的分组415发送到物理网络335上。封装的分组415包括具有接口VTEP2(作为源)和VTEP3(作为目的地)的源和目的地地址的外部报头。应当理解的是，在一些实施例中，L2和L3报头都作为封装的一部分被添加，其中包括源和目的地MAC和IP地址。因此，对于源VTEP2，其IP地址和MAC地址都被使用，并且对于VTEP3，至少其IP地址在外部报头中使用。对于目的地MAC地址，将使用下一跳物理路由器的MAC地址，除非VTEP2和VTEP3连接到物理网络中的相同交换机。但是，出于隧道传输的目的，外部IP地址是所选择的隧道端点的IP地址。

[0061] 此外，MFE 310存储其利用源MAC地址VM1对用于分组的源隧道端点的选择。在一些实施例中，MFE不是对每个分组执行散列计算，而是将其选择存储在其将地址映射到隧道端点的表中，使得通过参考该表进行将来的隧道端点选择。但是，对于目的地MAC地址 VM2，没有存储信息。这是因为MFE 310允许MFE 360自己做出关于它自己的哪个隧道端点用于地址VM2的决定，使得MFE 360可以考虑到它自己的任何负载平衡问题。

[0062] 图5图示了在两个阶段505和510上由MFE 360对这个第一封装的分组415的接收。

如第一阶段505中所示，由于网络335的物理交换机/路由器已经将分组转发到与这个VTEP对应的主机机器350 的接口，因此MFE 360通过VTEP3 365接收封装的分组415。在这个时候，分组仍然包括用于VTEP2和VTEP3的隧道端点IP地址。

[0063] 在第二阶段510中，MFE 360已经解封装分组415并执行了任何附加处理，以将内部逻辑分组400递送到VM 355。在一些实施例中，MFE 360检查存储在封装的分组中(例如，在隧道报头内)的逻辑网络信息和/或目的地MAC地址VM2，以便将分组400递送到 VM 355。

[0064] MFE 360还在其地址:组:VTEP映射表380中存储关于源隧道端点的信息。如图所示，与由第一MFE 310为地址VM1存储的信息相同的信息由MFE 360存储-即，MAC地址VM1映射到隧道端点VTEP2。通过存储这个信息，MFE将能够在向VM 305发送将来的分组时使用该映射，所述将来的分组或者从VM 355发送，或者从在主机350上操作的其它数据计算节点发送(因为在这种情况下，映射是基于MAC地址，而不是传输层连接5元组)。但是，在这个时候，MFE 360不存储关于将VM2地址映射到隧道端点的任何信息。

[0065] 图6图示了通过两个阶段605和610从第二主机机器350上的 VM 355发送第一返回分组600。这个分组具有VM2的源地址和 VM1的目的地地址。如上所述，如果两个VM不在同一个逻辑交换机上，那么VM 355可以不发送具有VM1的目的地MAC地址的分组，但是，一旦MFE已经执行了它的第一跳逻辑网络处理(并且在分组400的封装之前)，目的地MAC地址就将被修改为VM2的目的地MAC地址。

[0066] MFE 360接收这个分组，执行其各种处理操作(例如，逻辑网络处理等等)，并识别用于封装分组的源和目的地VTEP。对于目的地VTEP，MFE仅需要表380，因为这个表直接将目的地MAC地址 VM1映射到目的地VTEP 320。不仅MFE 360不需要计算散列和后续的模运算，而且分组600将被发送到通过其发出先前封装的分组 415的相同隧道端点，使得两个方向上的所有通信都将通过那个隧道端点。

[0067] 但是，对于源VTEP，没有信息被保存，因此MFE 360计算源 MAC地址VM2的散列并且使用它(如上所述)来映射到源隧道端点。在这种情况下，所确定的源隧道端点是VTEP4 370，它不是通过其接收第一分组415的隧道端点。如果两个MFE使用不同的散列计算，或者如果MFE具有不同数量的隧道端点并因此在计算出的散列值上使用不同的模数，那么可能出现这种情况。

[0068] 在第二阶段610中，MFE已经封装了分组600并将封装后的分组615发送到物理网络335上。在这种情况下，外部报头源IP地址是VTEP4的IP地址，并且目的地IP地址是VTEP2的IP地址。与第一分组一样，如果VTEP不在同一个物理网络交换机上，那么外部报头源MAC地址将是VTEP4的MAC地址，而目的地MAC地址可以是物理网络335中的居间路由器的MAC地址。

[0069] 此外，MFE 360在表380中存储其利用源MAC地址VM2对用于分组的源隧道端点(VTEP4)的选择。因此，对于从VM 355向 VM 305发送的将来分组，MFE 360将不必执行任何散列或模计算，因为这个信息将被完全存储在其表中。

[0070] 图7图示了通过两个阶段705和710由MFE 310接收封装的分组615。如第一阶段705中所示，当网络335的物理交换机/路由器已经将分组转发到与这个VTEP对应的主机机器300的接口时，MFE 310通过VTEP2 320接收封装的分组615。在这个时候，分组仍然包括用于VTEP4和VTEP2的隧道端点IP地址。

[0071] 在第二阶段710中，MFE 310已经解封装分组615并执行了任何附加处理，以将内部

逻辑分组600递送到VM 305。如上面所指出的，在一些实施例中，MFE 310检查存储在封装的分组中(例如，在隧道报头内)的逻辑网络信息和/或目的地MAC地址VM1，以便将分组递送到VM 305。

[0072] MFE 310还在其地址:组:VTEP映射表330中存储关于源隧道端点的信息。如图所示，与由发送MFE 360为地址VM2存储的信息相同的信息由MFE 310存储-即，MAC地址VM2映射到隧道端点VTEP4。通过存储这个信息，MFE将能够在向VM 355发送将来的分组时使用该映射，该将来的分组或者在MFE处从VM 305 接收，或者从在主机300上操作的其它数据计算节点接收。因此，MFE 310现在包括用于其本地VM 305和远程VM 355两者的VTEP 映射，并且因此将不需要对在这两个VM之间发送的分组执行任何散列或模计算，如下图中所示。

[0073] 图8图示了通过两个阶段805和810从VM 305发送第二分组 800。如第一阶段805中所示，与分组400一样，这个分组800具有源地址VM1和目的地地址VM2。但是，当MFE 310完成其其它处理并确定利用其封装分组的隧道端点时，只需要表330，因为这个表现在包含进行隧道端点选择所需的所有信息。具体而言，该表指示源地址VM1映射到隧道端点VTEP2，并且目的地地址VM2映射到隧道端点VTEP4。

[0074] 因此，在第二阶段810中，MFE已经封装了分组800，并将封装的分组815发送到物理网络335上。封装的分组815将具有与分组 415相似的外部报头，但是目的地IP地址是VTEP4的IP地址，而不是VTEP3的IP地址。外部目的地MAC地址将相同，除非两个 VTEP在同一个物理网络交换机上，在这种情况下，MAC地址也将是VTEP4的MAC地址。在这个时候，连接达到平衡，VM1和VM2之间的分组在VTEP2和VTEP4之间被隧道传输。除非VTEP2 或VTEP4出现故障，或者出于某种负载平衡目的，一个或两个 MFE改变要用于其本地VM地址的VTEP，否则这种平衡将继续。

[0075] B.MFE隧道端点选择处理

[0076] 图9概念性地图示了由MFE执行的一些实施例的处理900，该处理在封装分组并将那个分组输出到物理网络(例如，数据中心内的网络)上之前选择源和目的地隧道端点。取决于执行封装处理的 MFE的类型，可以以不同的方式执行这个处理。

[0077] 例如，在一些实施例的ESX主机中，虚拟交换机执行逻辑交换，而分布式虚拟路由器执行任何必要的逻辑路由。在一些实施例中，每个VTEP与虚拟交换机的虚拟端口相关联，并且在该端口上实现函数调用的堆栈。在一些实施例中，被称为I/O链的这个函数调用堆栈包括用于执行隧道端点选择操作的操作。

[0078] 对于基于内核的虚拟机(KVM)主机，一些实施例使用基于流的MFE，诸如Open vSwitch。在这种情况下，MFE在从控制器接收到的流条目中实现隧道选择，例如，利用匹配源或目的地地址并且绑定允许MFE选择隧道端点之一用于该地址映射到的组的动作的流条目。可以使用学习动作来实现学习方面，该学习动作在由MFE选择源隧道端点或通过隧道接收分组时，创建新的流条目。

[0079] 其它类型的MFE可以包括基于DPDK的MFE，其将封装信息存储在它们的配置数据库中并使用类似于KVM主机的学习动作以及顶架式(TOR)交换机。对于具有硬件VTEP的这种TOR交换机，在一些实施例中，中央控制器利用或者MAC:VTEP绑定(在这种情况下不执行负载平衡)或者MAC地址到组中所有VTEP的绑定来更新TOR中的数据库(例如，OVSDB)。在前一种情况下，中央控制器将尽可能多地执行负载平衡，并为此提供各种不同的选项。例如，控

制器可以使用不同的目的地VTEP用于跨不同TOR的MAC地址，或者跨主机上的逻辑交换机的不同目的地VTEP(例如，将用于第一逻辑交换机的流量发送到VTEP1，并将用于第二逻辑交换机的流量发送到VTEP2)。当用于同一逻辑交换机的多个数据计算节点在主机上运行时，中央控制器可以将TOR配置为将流量发送到用于不同 MAC地址的不同VTEP，即使它们位于同一个逻辑交换机上。

[0080] 如图所示，处理900通过接收(在905)从本地数据计算节点发送的需要经由覆盖网络传送的分组开始。在各种实施例中，本地数据计算节点可以是直接(例如，通过VNIC)连接到执行处理900的 MFE的VM或容器、在直接连接到MFE的VM上操作的容器、连接到TOR交换机的物理服务器。在一些实施例中，MFE是将逻辑网络连接到外部网络的网关，并且分组实际上从外部网络而不是本地数据计算节点接收。不管分组的来源如何，MFE在这个时候已经确定分组需要经由覆盖网络进行传输，并且已经假设地识别出要使用的隧道协议的类型以及要存储在封装中的逻辑网络上下文信息。

[0081] 然后该处理确定(在910)分组特点是否映射到用于源VTEP的特定本地VTEP。例如，在上面图4-8的示例中，MFE存储将MAC 地址映射到隧道端点的表。如果接收到的分组不是通过MFE从源 MAC地址发送的第一分组，那么MFE可以具有先前存储的将分组的源MAC地址映射到其VTEP之一的信息(例如，表条目、流条目等等)。在其它实施例中，映射可以是是特定于连接的而不是特定于 MAC的，并且MFE使用连接5元组(源IP地址、目的地IP地址、源传输端口号、目的地传输端口号，及协议)作为分组特点映射到其特定的一个VTEP。如果MFE只有单个VTEP，那么当然所有的流量都会映射到那个VTEP。

[0082] 当分组特点映射到特定的本地VTEP时(例如，因为分组不是由MFE处理的其类型的第一个)，那么处理选择(在915)分组特点映射到的特定本地VTEP作为用于该分组的源VTEP，并且前进到 930。另一方面，当分组特点不映射到特定本地VTEP时，处理使用 (在920)负载平衡算法来选择一组本地VTEP中的一个本地VTEP 作为用于分组的源VTEP。在一些实施例中，当MFE仅具有用于所有逻辑网络的单个相关联的VTEP组时，MFE从这个组中选择 VTEP。另一方面，当MFE具有多个不同的VTEP组(例如，因为 VTEP连接到与不同逻辑网络相关联的物理数据中心网络的不同区) 时，那么MFE为源地址和逻辑网络选择适当的组，然后选择这个组中的一个VTEP用于分组。

[0083] 如所提到的，负载平衡算法可以是基于散列的(例如，使用源 MAC地址、连接5元组或另一分组特点集合)，或者基于VTEP的当前负载(例如，基于VTEP的队列或缓冲区或VTEP所对应的物理NIC)。虽然前一子部分A中的示例说明了源MAC地址的散列以选择VTEP，但下面的子部分C说明了可以使用的其它技术。

[0084] 接下来，处理900存储(在925)分组特点到所选择的本地 VTEP的映射。例如，基于流的MFE可能会使用学习动作来创建新的流条目，以便具有相同特点集的后续分组将与新创建的流条目匹配，而不是与具有学习动作的流条目匹配。其它类型的MFE可以将映射存储在表中，如前面的示例中所示。

[0085] 然后处理900确定(在930)分组特点是否映射到用于目的地 VTEP的特定远程VTEP。例如，在上面图4-8的示例中，MFE存储将MAC地址映射到隧道端点的表。如果先前已经从当前分组的目的地地址接收到分组，那么MFE将具有先前存储的将当前分组的目的地MAC地址映射到用于具有那个地址的数据计算节点的MFE处的特定VTEP的信息(例如，表条目、

流条目等等)。在其它实施例中,映射可以是特定于连接的而不是特定于MAC的,并且MFE使用连接5元组作为分组特点来映射到用于分组的特定目的地VTEP。

[0086] 当分组特点映射到特定的远程VTEP时,该处理选择(在935) 分组特点映射到的特定VTEP作为用于分组的目的地VTEP。另一方面,当分组特点不映射到用于目的地隧道端点的特定VTEP时, MFE将分组的目的地地址映射(在940)到与远程MFE相关联的一组VTEP。如上面的示例中所示,在一些实施例中,每个MAC地址映射到特定的VTEP组,通常是具有该MAC地址的数据计算节点的主机上的VTEP集合,并且因此与该主机上的MFE相关联。在一些实施例中,这个映射信息基于来自中央控制器的更新(其可以被传递到用于MFE的本地控制器,其进而更新MFE)在MFE处被配置。

[0087] 然后处理900选择(在945)这个组中的一个VTEP作为用于分组的目的地VTEP。与源VTEP一样,目的地VTEP的选择可以基于目的地MAC地址的散列或其它因素(诸如连接5元组)。一些实施例可以将当前VTEP使用信息发送到其它MFE(例如,作为在 VTEP之间发送的保持存活消息的一部分),在这种情况下,MFE 可以使用这个信息来选择目的地VTEP。但是,一般不会使用这种技术,因为它需要在MFE之间发送大量状态信息。

[0088] 在选择了源和目的地隧道端点之后,处理900使用所选择的源和目的地VTEP来封装(在950)分组。在一些实施例中,封装包括所选择的VTEP的源和目的地网络地址(例如,IP地址),以及所选择的VTEP的源MAC地址和适当的目的地MAC地址(如果在与源 VTEP相同的物理网络交换机上,那么是目的地VTEP的MAC地址,或者是用于VTEP的默认网关端口的MAC地址)。除了可以取决于所使用的特定隧道传输协议(例如,VXLAN、STT、Geneve等等) 的其它信息之外,封装还可以包括逻辑网络上下文信息(例如,所确定的逻辑转发元件的逻辑出口端口、逻辑转发元件或逻辑网络标识符等等)。然后处理将封装的分组朝着目的地VTEP发送(在955)到物理网络上,并且结束。

[0089] 虽然处理900由发送MFE执行,但图10概念性地图示了由在其(可能是多个)VTEP中的一个VTEP处接收封装的分组的MFE 执行的一些实施例的处理1000。如同处理900一样,取决于接收分组的MFE的类型(例如,基于流的MFE,诸如OVS、ESX主机、 TOR交换机等等),处理1000可以以不同的方式执行。

[0090] 如图所示,处理1000开始于在位于隧道的目的地VTEP处通过隧道从源VTEP接收(在1005)由特定数据计算节点发送的分组。目的地VTEP是MFE的端口(或连接到MFE的端口),并且具有与接收到的分组的外部报头的目的地网络地址匹配的网络地址。在不同的实施例中,分组可以使用不同的隧道传输协议(例如,VXLAN、 STT、Geneve等等)来封装。

[0091] 然后处理确定(在1010)源VTEP到分组特点的映射是否被 MFE存储。如图4-8中所示,接收MFE存储将源MAC地址到用于封装的分组的源VTEP的映射,使得它可以在向那个地址发送分组时使用这个信息。在其它实施例中,接收MFE存储其它分组特点 (诸如连接5元组)到源VTEP的映射。当映射尚未被MFE存储时,该处理存储(在1015)分组特点到源VTEP的这个映射,以用于向发送接收到的分组的特定数据计算节点发送分组。

[0092] 在解封装分组后,该处理识别(在1020)分组的目的地。然后该处理将分组递送(在1025)到这个识别出的目的地。这个目的地可以是VM、容器、物理服务器(如果MFE是TOR的话)、外部网络(如果MFE是网关的话)等等。

[0093] C.可替换的负载平衡技术

[0094] 在以上子部分A的示例中，分别基于源和目的地MAC地址来确定源和目的地隧道端点两者的选择。对于附连有大量数据计算节点的MFE，假设MAC地址是随机分布的并且因此以接近相等的数量指派给不同的隧道端点，那么这将常常在MFE的隧道端点之间相对好地平衡流量。但是，如果数据计算节点中的一个或两个的流量远远高于其它数据计算节点的流量，那么那个数据计算节点的地址被指派给的隧道端点将承担不成比例的负担。

[0095] 因此，一些实施例使用其它负载平衡技术。例如，图11和12概念性地图示了使用连接5元组来将每个流分开指派给隧道端点，以获得更精细的负载平衡粒度。图11图示了通过两个阶段1102和1107 通过物理网络1135连接的一对主机机器1100和1150。与图3-8的主机机器类似，主机机器1100包括VM 1105，其连接到也在主机机器1100上运行的MFE 1110。MFE 1110具有两个VTEP 1115和1120。第二主机机器1150包括连接到同样在主机机器1150上运行的 MFE 1160的VM 1155，MFE 1160具有两个VTEP 1165和1170。

[0096] MFE 1100还存储类似于图3中所示的表，具有将VTEP组映射到VTEP列表的第一表1125以及将MAC地址映射到VTEP组的第二表1130。如第一阶段1102中所示，VM 1105向MFE发送分组 1140，该分组具有(一个或多个)源地址VM1和(一个或多个)目的地地址VM2。如在前面的示例中那样，MFE对分组执行逻辑网络处理，并且(使用表1130)确定目的地地址映射到VTEP的MFE2 组，并且(使用表1125)确定VTEP的这个组包括VTEP3和 VTEP4。

[0097] 但是，不是散列源MAC地址以确定源VTEP以及散列目的地 MAC地址以确定目的地VTEP，而是MFE 1110散列源和目的地IP 地址(例如，VM 1105和VM 1110的IP地址)、源和目的地传输层端口号和传输层协议(例如，在分组的IP报头中指定的TCP、UDP 等等)的连接5元组。基于这个散列(并且例如使用模函数从每个组中选择VTEP)，MFE 1110确定，对于这个流中的分组，源VTEP 是VTEP2 112。此外，对于至少第一分组，MFE将使用VTEP4 1170作为目的地VTEP。

[0098] 照此，在第二阶段1107中，MFE已经封装了分组1140并且将封装的分组1145发送到物理网络1135上。封装的分组1145包括具有接口VTEP2(作为源)和VTEP4(作为目的地)的源和目的地网络地址的外部报头。此外，MFE 1110将其对用于分组1140所属的流的源隧道端点的选择存储在表1175中。在这种情况下，表1175存储用于正在进行的流的5元组到源和目的地VTEP的映射，以用于那个流。但是，如上面基于MAC地址的示例一样，MFE 1110在这个阶段仅存储源VTEP。接收MFE 1160还将这个VTEP存储为用于反向分组的目的地隧道端点，并且当它处理流的第一反向分组时执行其自己的负载平衡，以选择源VTEP。在这下一个分组之后，将实现稳态，因为位于隧道任一端的MFE 1110和1160将具有为所选择的两个VTEP存储的信息。

[0099] 图12图示了第三主机机器1200，其中第三VM 1205和第三 MFE 1210连接到相同的物理数据中心网络1135。MFE 1210具有三个VTEP 1215-1225。这个图图示了通过两个阶段1202和1207从 VM 1105向VM 1205发送分组。在第一阶段中，MFE 1110从VM 1105接收具有源MAC地址VM1和目的地地址VM3的分组1230。 MFE对分组执行逻辑网络处理，并且(使用表1130)确定目的地地址映射到VTEP的MFE3组，并且(使用表1125)确定VTEP的这个组包括VTEP5、VTEP6和VTEP7。

[0100] 此外，由于这是新流，因此即使先前由VM 1105发送的(一个或多个)分组已经由MFE 1110处理(例如，分组1140)，也不会在表1175中存在用于该分组的5元组的条目。因此，

MFE 1110散列这个新分组的连接5元组，并且确定(例如，通过使用模函数从每个组中选择VTEP)，对于这个流中的分组，源VTEP是VTEP1 1115，并且目的地VTEP是VTEP7 1225。因此，即使分组1140和分组1230的源MAC地址相同，散列连接5元组也导致用于封装这些分组的不同源VTEP。

[0101] 在第二阶段中，MFE 1110已经封装了分组1230并且将封装的分组1235发送到物理网络1135上。封装的分组1235包括具有接口 VTEP1(作为源)和VTEP7(作为目的地)的源和目的地网络地址的外部报头。此外，MFE 1110将其对用于分组1230所属的流的源隧道端点的选择存储在表1175中。如在先前图11的示例中那样， MFE 1110仅存储源VTEP，并且在MFE 1210发送返回分组时学习目的地VTEP。

[0102] 图13概念性地图示了通过两个阶段1302和1307使用负载评估来为流指派源VTEP。该图示出了通过物理网络1335连接的一对主机机器1300和1350。类似于图3-8中的主机机器，主机机器1300 包括连接到也在主机机器1300上操作的MFE 1310的VM 1305。MFE 1310具有两个VTEP 1315和1320。第二主机机器1350包括连接到也在主机机器1350上操作的MFE 1360的VM 1355，MFE 1360 具有两个VTEP 1365和1370。MFE 1300还存储与图3中所示的类似的表，其中第一表1325将VTEP组映射到VTEP列表，并且第二表1330将MAC地址映射到VTEP组。

[0103] 如第一阶段1302中所示，VM 1305向MFE发送分组1340，该分组具有(一个或多个)源地址VM1和(一个或多个)目的地地址 VM2。如在前面的示例中那样，MFE对这个分组执行逻辑网络处理，并且(使用表1130)确定目的地地址映射到VTEP的MFE2组并且 (使用表1125)确定VTEP的这个组包括VTEP3和VTEP4。在这种情况下，MAC地址和连接5元组都不直接映射到特定的VTEP。

[0104] 但是，MFE不是散列任何分组特点来确定源VTEP，而是评估不同VTEP上的负载。在一些实施例中，MFE评估每个VTEP的缓冲区或队列(或与VTEP相关联的物理NIC)。如黑条所示，第一 VTEP1 1315中的缓冲区比第二VTEP2 1320中的缓冲区更满，因此 MFE 1310将VTEP2指派为分组1340的源VTEP。对于目的地 VTEP，没有负载来评估(因为MFE 1310不知道这个信息)，因此 MFE计算连接5元组(或MAC地址)的散列以确定要使用的目的地VTEP(在这种情况下为VTEP4 1370)。

[0105] 照此，在第二阶段1307中，MFE 1310已经封装了分组1340，并将封装的分组1345发送到物理网络1335上。封装的分组1345包括具有接口VTEP2(作为源)和VTEP4(作为目的地)的源和目的地网络地址的外部报头。此外，MFE 1310将其对用于分组1340所属的流的源隧道端点的选择存储在表1375中，该表存储用于正在进行的流的5元组到源和目的地VTEP的映射，以用于那个流。如在前面的示例中那样，MFE仅存储源VTEP，从而允许MFE 1360选择其VTEP之一(例如，基于类似的负载平衡技术、5元组的散列等等)。在其它实施例中，由于MFE以每个分组为基础执行负载平衡，因此不存储针对源VTEP或目的地VTEP的信息。在这种情况下，做出的决定是对于经过相同VTEP的流，更优化的负载平衡胜过具有传入和传出分组的好处。

[0106] D.处理来自物理网络的重复流量

[0107] 如上面的示例中所示，一些实施例的MFE通过多个隧道端点连接到物理网络交换机。因此，在一些情况下，MFE可以接收到相同分组的多个副本。例如，当相同MFE的多个VTEP

附连到的物理交换机接收到带有VTEP地址的分组时,交换机通常会知道向其发送分组的端口。但是,对于第一个分组,这个交换机的表中可能没有目的地MAC地址,因此它会泛洪其所有端口,包括MFE上的所有 VTEP。但是,所有这些VTEP都是同一MFE上的端口,并且让 MFE处理多个此类分组是无益的。因此,MFE不必处理同一个分组的多个副本,一些实施例对外部报头执行检查,以便过滤掉除一个物理NIC(因为它是实际连接到物理网络交换机的物理NIC)之外的所有物理NIC上接收到的流量。

[0108] 如果这个流量是去往连接到MFE的数据计算节点,那么过滤 (例如,使用反向路径转发)防止向数据计算节点递送同一个分组的多个副本,当这么做时可能会对数据计算节点与分组的源之间的通信产生不利影响。通常,MFE将通过对照在其上接收分组的物理NIC检查分组的内部报头来执行过滤。当数据计算节点静态映射到仅单个 VTEP(即使MFE上存在多个VTEP)时,发送到其它物理NIC的副本不会以任何方式被递送到数据计算节点。但是,在本文的示例中,数据计算节点没有被静态地指派给特定的VTEP或物理NIC,而是用于特定数据计算节点的分组可以到达VTEP和物理NIC中的任何一个。因此,一些实施例基于外部报头执行过滤检查(例如,反向路径转发)以避免重复,这也防止在解封装之后查看内部报头以执行过滤检查的需要。

[0109] E.处理VTEP或pNIC功能的丢失

[0110] 如上面所提到的,一些实施例将每个VTEP绑定到不同的物理 NIC。因为物理网络交换机将每个物理NIC视为单独的端口,所以在这种情况下,VTEP应当仅被指派给一个物理NIC(因为具有特定 VTEP的源MAC地址的分组只应当被发送到物理网络交换机的一个端口)。

[0111] 使用数据计算节点地址到VTEP组绑定,当物理NIC出现故障时,VTEP不需要切换到不同的上行链路;相反,在一些实施例中, VTEP从与VTEP组相关联的VTEP列表中移除(至少暂时地),并且来自本地数据计算节点的分组不再从那个VTEP发出。此外,中央网络控制器可以基于由用于那个MFE的本地控制器或由其它 MFE进行的自我报告而获知VTEP出现故障。当VTEP出现故障时,其它VTEP将检测到这种情况,因为隧道端点之间的保持存活消息将不再从出现故障的VTEP接收到。在一些实施例中,具有其它 VTEP的MFE检测到出现故障的VTEP并将这个信息报告给中央控制器。

[0112] 对于VTEP选择,不同的实施例可以以不同方式重新指派被指派给出现故障的VTEP的流和/或MAC地址。例如,一些实施例简单地重新散列在流中的下一个分组上或者来自MAC地址的源地址,并且使用不同的模来考虑到不同数量的VTEP。已经指派给保持工作的VTEP的流或MAC地址不需要,但可以,移动,并且新分组应当平均散布在剩余的VTEP当中。

[0113] II.具有多个VTEP的桥接器集群

[0114] 除了通过隧道将分组发送到具有多个隧道端点的MFE以便到达远程数据计算节点之外,在一些情况下,MFE还从其本地数据计算节点之一接收需要桥接到另一个远程网络(例如,另一个逻辑网络、VLAN等等)的分组。一些实施例使用桥接器集群来执行这种桥接,其可以包括多个MFE(其中的一些或全部可以具有多个隧道端点)。与桥接器集群相关联的所有隧道端点可以被分组,并且源MFE执行类似的选择处理(例如,散列目的地地址或连接5元组),以选择桥接器集群组中要向其发送分组的隧道端点。

[0115] 图14概念性地图示了由MFE执行的一些实施例的处理1400,该处理选择要发送到

远程网络(例如,远程层2网络)的分组的源和目的地隧道端点。如同图9的处理900一样,取决于执行封装处理的 MFE的类型(例如,MFE是否使用流条目、查找表等等),可以以不同的方式执行这个处理1400。

[0116] 如图所示,处理1400通过接收从本地数据计算节点(即,MFE 为其第一跳MFE的数据计算节点)发送的分组而开始(在1405)。数据计算节点可以是在与MFE相同的主机上运行的VM或容器、连接到TOR交换机的物理服务器等等。

[0117] 然后该处理确定(在1410)分组的目的地地址与远程网络中需要桥接的数据计算节点对应。在一些实施例中,第一层2网络(例如,与覆盖(诸如VXLAN、STT、Geneve等等覆盖)连接的网络)中的第一数据计算节点集合可以与第二层2网络(例如,另一个覆盖网络、物理VLAN网络等等)中的第二数据计算节点集合在同一逻辑交换机上。这两个层2网络可能不位于同一数据中心或数据中心的同一物理区中,因此它们被视为需要网桥的分离的层2网络。因此,当第一层2网络中的MFE接收到具有第二层2网络中的目的地地址的分组时,MFE表将MAC地址识别为需要桥接。

[0118] 接下来,处理1400选择(在1415)源隧道端点。这个选择可以用上一部分中描述的任何方式执行。即,MFE可能基于之前确定并存储的源MAC地址到隧道端点的映射来选择源隧道端点。在不同的实施例中,MFE还可以基于源MAC地址的散列、连接5元组的散列、各种隧道端点上的负载的评估或者其它负载平衡技术来从一组隧道端点中选择隧道端点。

[0119] 该处理还将接收到的分组的目的地地址映射(在1420)到与桥接器集群相关联的一组隧道端点,用于将分组桥接到远程网络。在一些实施例中,隧道端点可以一般性地与组相关联。在其它实施例中,隧道端点既有MFE标签又有MFE组标签。桥接器集群可以包括若干MFE,每个MFE具有一个或多个隧道端点。因此,对于需要被桥接的分组,将目的地地址映射到形成桥接器集群的MFE组(即,具有将分组桥接到目的地层2网络的能力的MFE)。这个组本身被映射到隧道端点集合。该隧道端点集合可以包括MFE组中的MFE 的所有隧道端点,或者如果一些隧道端点不具有与执行处理1400的 MFE相同的数据中心的区的连接性,那么仅包括MFE中的一些或全部的子集。

[0120] 接下来,该处理基于分组特点选择(在1425)该组的隧道端点之一作为目的地隧道端点。例如,不同的实施例可以使用散列目的地 MAC地址或连接5元组的负载平衡技术,并且基于这个散列结果 (例如,使用如上所述的模函数)来选择组中的隧道端点之一。此外,当接收到从远程网络中的数据计算节点返回的分组时,一些实施例可以存储目的地MAC地址(或流)到桥接器集群VTEP的映射。一旦选择了源和目的地隧道端点,该处理就利用所选择的隧道端点封装 (在1430)分组并且将分组发送(在1435)到物理网络上,并且结束。

[0121] 图15概念性地图示了包括提供第一VM所位于的第一网络与第二网络之间的桥接连接性的桥接器集群的数据中心网络的示例。如图所示,主机机器1500包括VM 1505和VM 1505连接到的MFE 1510, MFE 1510具有两个VTEP 1515和1520。MFE 1500存储将VTEP 组标识符映射到VTEP列表的第一表1525以及将MAC地址映射到 VTEP组标识符的第二表1530。组标识符包括用于VM 1505的MAC地址VM1的MFE1,以及用于两个MAC地址MAC1和 MAC2的桥接器集群Bridge1。

[0122] 该图还图示了在三个主机1550-1560上执行到物理网络VLAN1 的桥接的三个MFE 1535-1545。这是具有MAC地址MAC1和 MAC2的数据计算节点所位于的网络。这些MFE 1535-

1545各自具有不同数量的VTEP,所有这些VTEP都通过物理网络1565连接到 MFE 1510的VTEP 1515和1520。具体而言,MFE 1535具有两个 VTEP 1570和1575,MFE 1540具有单个VTEP 1580,并且MFE 1545具有三个VTEP 1585-1595。虽然未示出,但是这些主机机器1550-1560中的一些或全部可以托管虚拟机或其它数据计算节点(例如,在与VM1完全分离的逻辑网络上)并且用作用于这些数据计算节点的第一跳MFE。

[0123] 图16概念性地图示了从VM 1505向桥接网络VLAN1中的 MAC地址(MAC1)发送分组1600。如图所示,VM 1505向MFE 1510发送具有源地址VM1和目的地地址MAC1的分组1600。MFE 1510处理分组并且确定MAC地址MAC1映射到两个层2网络之间的桥接器。使用表1130,MFE 1510识别用于该地址的VTEP组 Bridge1,并且使用表1125识别用于这个VTEP组的六个可能的 VTEP。在这种情况下,MFE 1510使用基于MAC的散列来确定源和目的地VTEP,并且因此散列地址VM1以获得VTEP1的源隧道端点并且散列地址MAC1以获得VTEP7的目的地隧道端点。因此, MFE 1510在外部封装报头中用这些源和目的地地址封装分组1600,并且将封装的分组1605发送到物理网络1565上。MFE 1545将接收这个封装的分组,移除封装,并执行将分组桥接到VLAN1上的必要的附加操作。

[0124] 图17概念性地图示了从VM 1505向桥接网络VLAN1中的不同 MAC地址(MAC2)发送分组1700。如图所示,VM 1505向MFE 1510发送具有源地址VM1和目的地地址MAC2的分组1700。MFE 1510以与先前分组1600相同的方式处理该分组,以确定MAC地址 MAC2映射到两个层2网络之间的桥接器。使用表1130,MFE 1510 识别用于地址MAC2的VTEP组Bridge1,并使用表1125识别用于这个组的六个可能的VTEP。在这种情况下,MFE 1510基于从先前分组1600的处理存储的信息将源VTEP识别为VTEP1。但是,没有存储将MAC2映射到特定VTEP的信息,因此MFE计算散列以从桥接器集群VTEP组中选择VTEP5 1580。

[0125] 这不仅与用于桥接先前分组1700的VTEP不同,而且实际上是完全不同的MFE 1540上的VTEP。但是,所有这三个MFE 1535- 1545都具有将分组桥接到同一个VLAN的能力,因此将用于VLAN 的分组发送到集群中的不同MFE没有问题。使用所选择的VTEP, MFE 1510在外部封装报头中用这些VTEP的源和目的地地址封装分组1700,并且将封装的分组1705发送到物理网络1565上。MFE 1540将接收这个封装的分组,移除封装,并执行将分组桥接到 VLAN1上的必要的附加操作。

[0126] III.中央控制器操作

[0127] 在以上所示的各种示例中,MFE被填充有将MAC地址映射到隧道端点组标识符(例如,特定MFE或一组MFE的隧道端点)的表以及将隧道端点组标识符映射到隧道端点集合的表。使用这些表,发送MFE可以将目的地MAC地址映射到一组潜在的隧道端点,然后以多种不同的方式选择这些端点之一。在一些实施例中,MFE由网络控制系统填充,该网络控制系统可以包括中央控制器(或多个中央控制器)以及充当中央控制器与MFE之间的中间体(intermediary) 和翻译器的本地控制器。

[0128] 图18概念性地图示了一些实施例的网络控制系统1800以及这个网络控制系统内的数据传送,因为它涉及在MFE上配置的映射表。如图所示,网络控制系统包括管理平面1805、中央控制平面1810和在主机机器1820上操作的本地控制器1815。在一些实施例中,管理平面1805和控制平面1810都是单个网络控制器机器上的模块或应用。它们也可以是分布式的,因为管理平面和中央控制平面都在众多的网络控制器上运行,不同的控制器处理不

同逻辑网络的配置。在其它实施例中，管理平面1805和中央控制平面1810在分离的物理机器中实现。例如，一些实施例的网络控制系统可以具有管理平面在其上操作的许多机器以及中央控制平面在其上操作的许多其它机器。

[0129] 一些实施例的管理平面1805负责通过API(例如，从云管理系统的用户)接收逻辑网络配置，并将那个网络配置数据转换为逻辑网络的内部表示(例如，作为数据库表的集合)。例如，在一些实施例中，管理平面接收彼此连接的交换机和路由器的逻辑网络，并且定义用于逻辑交换机、逻辑交换机端口、逻辑路由器、逻辑路由器端口等的构造。在一些实施例中，每个逻辑交换机具有指派给其逻辑端口的 MAC和/或IP地址集合，每个MAC地址与将在逻辑网络中实现的不同数据计算节点对应。如图所示，在一些实施例中，管理平面 1805将这个逻辑网络配置数据1825传递到中央控制平面1810。除了逻辑转发元件彼此的连接之外，该信息还可以包括用于每个逻辑交换机的数据计算节点地址(例如，MAC地址)以及逻辑网络、逻辑转发元件和/或逻辑端口标识符的列表。这些标识符可以在一些实施例的封装中使用，以便在用于逻辑网络的MFE之间发送的覆盖分组中编码逻辑网络上下文信息。

[0130] 在一些实施例中，管理平面1805还配置主机机器1820上的各种 MFE 1830上的VTEP，以及TOR交换机上的VTEP、桥接器集群 MFE、网关等等。因为管理平面1805知道用于每个VTEP的信息以及在哪个MFE上供给VTEP，所以管理平面可以将VTEP指派给组 (例如，通过为各个MFE指派MFE标签以及为桥接器集群中的MFE组指派组标签)。因此，管理平面1805附加地将利用它们的组标签在网络中配置的VTEP的列表1835传递到中央控制平面。

[0131] 一些实施例的中央控制平面1810将组标签的绑定表1837存储到 VTEP。在一些实施例中，中央控制平面1810按组来组织这个表，其中用于组的每行具有该组中的所有VTEP的列表(类似于上面的示例中针对各种MFE所示的表)。

[0132] 在一些实施例中，中央控制平面1810还接收来自主机机器1820 的信息。如图所示，每个主机机器包括管理MFE 1830的本地控制器 1815以及通过MFE通信的一个或多个数据计算节点(例如，VM)。在一些实施例中，本地控制器连同MFE 1830一起被集成到主机机器1820的虚拟化软件中。本地控制器从管理平面1805和/或中央控制平面1810接收配置数据，并使用这个数据来配置它们相应的MFE 1830。在一些实施例中，本地控制器经由中央控制平面使用的特定协议(诸如netcpa)接收这个数据，并将这个数据转换为MFE 1830可读的配置数据。例如，当MFE是基于流的MFE时，本地控制器 1815接收描述配置的抽象数据元组并且将这些数据元组转换成用于配置MFE的流条目。当MFE是诸如ESX之类的基于代码的MFE时，本地控制器1815接收抽象数据元组并将它们转换成用于MFE 的适当类型的表条目，然后将这些表条目分发到MFE的适当部分 (诸如虚拟交换机、VTEP、I/O链、上行链路等等)(例如，通过在适当的表中安装每行)。

[0133] 此外，本地控制器1815向中央控制平面1810报告关于它们相应的MFE的各种数据。这可以包括关于失败端口的信息或其它运行时数据。此外，当新的数据计算节点安装在主机机器1820上时，数据计算节点连接到该主机上的MFE 1830，并且本地控制器报告这个数据，包括用于数据计算节点的地址和其它信息(诸如数据计算节点所属的逻辑网络)。如图所示，本地控制器将数据计算节点位置1840 发送到中央控制平面，中央控制平面识别数据计算节点与MFE之间的绑定。在一些实施例中，本地控制器1815还用这个数据报告用于 MFE的VTEP组标签，使得中央控制平面存储具有数据计算节点地址与VTEP组之间的绑定的表

1845。这种表中的每一行可以包括地址和VTEP组标识符，但是一些实施例还包括逻辑网络标识符(从而允许不同的逻辑网络重复使用地址而不会混淆)。

[0134] 网络拓扑评估器1850或中央控制平面中的其它模块使用从管理平面1850以及从本地控制器1815接收的数据向本地控制器1815提供数据以用于它们的MFE 1830。网络拓扑评估器1850使用地址:组绑定，以确定向每个本地控制器发送哪个逻辑网络信息。即，如果特定MFE 1830具有属于与其绑定的特定逻辑网络的数据计算节点，那么中央控制平面将向管理该特定MFE的本地控制器1815发送用于该特定逻辑网络的逻辑网络配置数据1855。中央控制平面1810还将必需的组:VTEP绑定和数据计算节点地址:组绑定1860发送到本地控制器1815。如果数据计算节点存在于主机机器1820上，那么那个主机机器上的MFE 1830将需要接收附连到与数据计算节点相同的逻辑网络的所有数据计算节点的地址:组绑定。此外，MFE 1830 还需要接收用于逻辑网络的其它数据计算节点所位于的每个组(即，每个MFE)的组:VTEP列表绑定，使得如果有必要的话，那么 MFE可以通过覆盖网络向这些MFE中的任何一个发送分组。

[0135] 图19概念性地图示了由一些实施例的中央控制面执行的处理 1900，该处理收集并向受管理转发元件分发关于隧道端点到数据计算节点地址的映射的数据。虽然处理1900被示为线性处理，但应当理解的是，在一些实施例中，中央控制平面将定期接收更新并将更新后的信息分发到MFE，而不是简单地一次接收所有数据并且仅一次向 MFE分发那个数据。此外，在一些情况下，第一中央控制器从管理平面接收逻辑网络配置数据和/或隧道端点信息并且与集群中的其它中央控制器共享这个信息，而集群中的第二网络控制器向本地控制器分发用于MFE的数据。

[0136] 如图所示，处理1900通过接收VTEP组标签而开始(在1905)。该处理还接收(在1910)具有用于每个VTEP的组标签的VTEP列表。在一些实施例中，这个信息是作为单个数据集(例如，在单个事务中)从管理平面接收的。例如，一些实施例不单独向中央控制平面提供组标签列表，而是仅向控制平面提供VTEP列表和指派给它们的组标签。对于每个VTEP，组数据可以包括MFE标签以及集群标签。例如，如果VTEP位于桥接器集群中涉及的MFE上并且还充当用于其主机上的数据计算节点的第一跳MFE，那么VTEP将需要这两种类型的标签，因为发送到数据计算节点的分组将需要去往那个特定的MFE，但发送到远程桥接网络的分组可以被发送到桥接器集群中的任何MFE。

[0137] 虽然一些实施例从管理平面接收这个VTEP数据，如图18所示，但是在其它实施例中，每当在主机上创建/配置VTEP时，中央控制器从本地控制器接收VTEP信息。即，当管理平面在特定的MFE上创建新的VTEP时，用于MFE的本地控制器会检测这个信息并向中央控制器发送向中央控制器通知这个新VTEP的消息。这些消息还提供MFE标签和/或集群标签。

[0138] 在接收到VTEP组信息之后，该处理存储(在1915)将组标签映射到用于每个标签的VTEP列表的表。在一些实施例中，这个表包括用于每个标签的行，具有与该标签相关联的VTEP列表。通过为每个标签确定所有的VTEP，中央控制器基于接收到的指定用于每个VTEP的标签的信息生成这个表。在一些实施例中，如果VTEP 既属于用于特定MFE的一组VTEP又属于用于跨多个MFE的桥接器集群的一组VTEP，那么该VTEP可以在多行中列出。

[0139] 该处理还接收(在1920)用于逻辑网络的逻辑网络配置和标识符。如上面所解释的，在一些实施例中，管理平面通过API从管理员接收用于逻辑网络的配置，并将其翻译成

描述用于中央控制平面的逻辑网络的表格数据。在一些实施例中，管理平面还为每个逻辑交换机、逻辑路由器、逻辑交换机端口、逻辑路由器端口等等指派逻辑转发元件和逻辑端口标识符(例如，UUID)，并且将这个信息作为配置的一部分提供。逻辑网络配置还包括用于附连到逻辑网络中的每个逻辑交换机的数据计算节点的地址(例如，MAC地址)列表，并且在一些实施例中指定数据计算节点的类型(例如，VM、容器、物理服务器等等)。由于管理员可以创建新网络、移除网络或修改配置 (例如，通过向网络添加或移除数据计算节点)，因此可以在不同的时间为不同的逻辑网络接收这种逻辑网络信息。

[0140] 处理1900还接收(在1925)网络中的数据计算节点的位置。在一些实施例中，每当数据计算节点被指派给新位置(例如，新主机机器)时，从本地控制器接收这种信息。在一些实施例中，分离的计算管理系统负责在主机上创建新的数据计算节点。当在MFE上创建新的虚拟接口用于连接到新的数据计算节点时，用于那个MFE的一些实施例的本地控制器检测新接口(例如，新的虚拟端口)并将端口信息报告给中央控制器。这种端口信息指定MAC地址以及其它信息 (诸如用于新数据计算节点的相关联的逻辑网络)。与其它数据一样，这种信息的接收将不会在单个事务中发生，因为不同的数据计算节点将在不同的时间创建，并且本地控制器将在它们检测到新接口时报告这些事件。

[0141] 接下来，处理1900基于数据计算节点的网络中的位置存储(在 1930)将数据计算节点地址映射到组标签的表。即，当本地控制器在其MFE处报告新数据计算节点的存在时，控制器在其表中为该数据计算节点存储新行。在一些实施例中，这个行指定用于数据计算节点的MFE组标签，以及用于数据计算节点的逻辑网络或逻辑交换机标识符。

[0142] 使用逻辑网络配置数据以及数据计算节点到位置的映射，该处理确定(在1935)实现每个逻辑网络的MFE集合(或者相反，由每个 MFE实现的逻辑网络集合)。对于连接到特定MFE的每个数据计算节点，中央控制器确定数据计算节点的逻辑网络，并且确定特定的 MFE应当接收用于那个逻辑网络的配置信息。在一些实施例中，如果数据计算节点在逻辑上连接到第一逻辑交换机，那么MFE将需要接收用于那个逻辑交换机以及逻辑网络的其它逻辑转发元件(例如，该逻辑交换机连接到的到逻辑路由器、连接到那个逻辑路由器的其它逻辑交换机、连接到那个逻辑路由器的其它逻辑路由器以及它们连接的逻辑交换机等等)的配置数据。

[0143] 对于每个MFE，然后处理1900向分发到该MFE的组标签记录分发(在1940)(i)用于由MFE实现的每个逻辑网络的逻辑网络 (或逻辑转发元件)配置，(ii)对于连接到任何逻辑网络中分发到 MFE的任何逻辑交换机(例如，用于其的配置被分发到MFE的任何逻辑交换机)的每个数据计算节点地址，将数据计算节点地址映射到组标签的表的记录，以及(iii)对于与数据计算节点地址相关联的每个组标签，将组标签映射到VTEP列表的表的记录。这是允许本地控制器配置MFE以实现逻辑网络并通过隧道将逻辑网络分组发送到适当VTEP的信息。即，MFE上的每个数据计算节点都属于逻辑网络。基于那个逻辑网络关联，MFE确定MFE所需的逻辑网络 (逻辑转发元件)数据。这个逻辑网络具有附连的其它数据计算节点，因此如果其本地数据计算节点向那些其它数据计算节点发送分组，则 MFE需要该信息以到达这些其它数据计算节点。这种信息包括这些其它数据计算节点的位置(VTEP组标签)以及用于每个这种位置的 VTEP列表。

[0144] IV.电子系统

[0145] 许多上述特征和应用被实现为软件过程，这些软件过程被指定为记录在计算机可读存储介质(也被称为计算机可读介质)上的指令集合。当这些指令被一个或多个处理单元(例如，一个或多个处理器、处理器核心、或其它处理单元)执行时，它们使得该(一个或多个) 处理单元执行在指令中指示的动作。计算机可读介质的示例包括，但不限于，CD-ROM、闪存驱动器、RAM芯片、硬盘驱动器、EPROM等等。计算机可读介质不包括无线地传递或通过有线连接传递的载波和电子信号。

[0146] 在本说明书中，术语“软件”是指包括驻留在只读存储器中的固件或者可以被读入到存储器中以供处理器处理的存储在磁存储中的应用。此外，在一些实施例中，多个软件发明可以被实现为更大程序的子部分，同时保持明显的软件发明。在一些实施例中，多个软件发明也可以被实现为单独的程序。最后，一起实现本文所描述的软件发明的单独程序的任意组合在本发明的范围内。在一些实施例中，当软件程序被安装，以在一个或多个电子系统上操作时，软件程序定义运行且执行软件程序的操作的一个或多个特定的机器实现。

[0147] 图20概念性地图示了实现本发明的一些实施例的电子系统2000。电子系统2000可以用于执行上述任何控制、虚拟化或操作系统应用。电子系统2000可以是计算机(例如，台式计算机、个人计算机、平板计算机、服务器计算机、大型机、刀片计算机等等)、电话、PDA或任何其它类型的电子设备。这样的电子系统包括用于各种其它类型的计算机可读介质的各种类型的计算机可读介质和接口。电子系统 2000包括总线2005、(一个或多个)处理单元2010、系统存储器 2025、只读存储器2030、永久性存储设备2035、输入设备2040和输出设备2045。

[0148] 总线2005统一地表示可通信地连接电子系统2000的许多内部设备的所有系统、外围设备和芯片组总线。例如，总线2005将(一个或多个)处理单元2010与只读存储器2030、系统存储器2025和永久性存储设备2035可通信地连接。

[0149] (一个或多个)处理单元2010从这些各种存储器单元中检索要执行的指令和要处理的数据，以便执行本发明的过程。(一个或多个) 处理单元在不同实施例中可以是单个处理器或多核处理器。

[0150] 只读存储器(ROM)2030存储由(一个或多个)处理单元2010 和电子系统的其它模块所需的静态数据和指令。另一方面，永久性存储设备2035是读写存储器设备。这个设备是即使当电子系统2000关闭时也存储指令和数据的非易失性存储器单元。本发明的一些实施例使用大容量存储设备(诸如磁盘或光盘及其对应的盘驱动器)作为永久性存储设备2035。

[0151] 其它实施例使用可移除存储设备(诸如软盘、闪存驱动器等等) 作为永久性存储设备。与永久性存储设备2035一样，系统存储器 2025是读写存储器设备。但是，与存储设备2035不同的是，系统存储器是易失性读写存储器，诸如随机存取存储器。系统存储器存储处理器在运行时所需的一些指令和数据。在一些实施例中，本发明的过程被存储在系统存储器2025、永久性存储设备2035和/或只读存储器 2030中。(一个或多个)处理单元2010从这些各种存储器单元中检索要执行的指令和要处理的数据，以便执行一些实施例的过程。

[0152] 总线2005还连接到输入和输出设备2040和2045。输入设备使用户能够向电子系统传达信息和选择给电子系统的命令。输入设备 2040包括字母数字键盘和定点设备(也称为“光标控制设备”)。输出设备2045显示由电子系统生成的图像。输出设备包括打印机和显示设备，诸如阴极射线管(CRT)或液晶显示器(LCD)。一些实施例包括用作输入和输出设备两

者的设备，诸如触摸屏。

[0153] 最后，如图20所示，总线2005还通过网络适配器(未示出)将电子系统2000耦合到网络2065。以这种方式，计算机可以是计算机的网络(诸如局域网(“LAN”)、广域网(“WAN”)、或内联网、或网络的网络，诸如互联网)的一部分。电子系统2000的任何或全部组件可以与本发明结合使用。

[0154] 一些实施例包括电子组件，诸如微处理器、在机器可读或计算机可读的介质(可替代地称为计算机可读存储介质、机器可读介质或机器可读存储介质)中存储计算机程序指令的存储设备和存储器。这种计算机可读介质的一些示例包括RAM、ROM、只读压缩盘(CD-ROM)、可记录压缩盘(CD-R)、可重写压缩盘(CD-RW)、只读数字多功能盘(例如，DVD-ROM，双层DVD-ROM)、各种可记录/可重写DVD(例如，DVD-RAM、DVD-RW、DVD+RW等等)、闪存存储器(例如，SD卡、小型SD卡、微型SD卡等等)、磁和/ 或固态硬盘驱动器、只读和可记录Blu-Ray®盘、超密度光盘、任何其它光或磁介质、以及软盘。计算机可读介质可以存储可由至少一个处理单元执行的并且包括用于执行各种操作的指令集的计算机程序。计算机程序或计算机代码的示例包括诸如由编译器产生的机器代码，以及包括由计算机、电子组件、或利用解释器的微处理器执行的更高级代码的文件。

[0155] 虽然以上讨论主要指执行软件的微处理器或多核处理器，但是一些实施例通过一个或多个集成电路，诸如专用集成电路(ASIC)或现场可编程门阵列(FPGA)，来执行。在一些实施例中，这种集成电路执行在该电路自身上存储的指令。

[0156] 如在本说明书中所使用的，术语“计算机”、“服务器”、“处理器”、以及“存储器”都是指电子或其它技术设备。这些术语不包括人或人群。为了本说明书的目的，术语显示或正在显示意味着在电子设备上显示。如本说明书中所使用的，术语“计算机可读介质”、“多个计算机可读介质”和“机器可读介质”被完全限制为以由计算机可读的形式存储信息的、有形的、物理的对象。这些术语不包括任何无线信号、有线下载信号、以及任何其它短暂信号。

[0157] 贯穿本说明书提到包括虚拟机(VM)的计算和网络环境。但是，虚拟机只是数据计算节点(DCN)或数据计算端节点(也被称为可寻址节点)的一个示例。DCN可以包括非虚拟化物理主机、虚拟机、在主机操作系统之上运行而不需要管理程序或单独的操作系统的容器、以及管理程序内核网络接口模块。

[0158] 在一些实施例中，VM使用由虚拟化软件(例如，管理程序、虚拟机监视器等等)虚拟化的主机的资源与在主机上其自己的客户操作系统一起操作。租户(即VM的所有者)可以选择在客户操作系统之上要操作哪些应用。另一方面，一些容器是在主机操作系统之上运行而不需要管理程序或单独的客户操作系统的构造。在一些实施例中，主机操作系统使用命名空间来将容器彼此隔离，并因此提供在不同容器内操作的不同应用组的操作系统级分离。这种分离类似于在虚拟化系统硬件的管理程序虚拟化环境中提供的VM分离，并且因此可以被视为隔离在不同容器中操作的不同应用组的一种虚拟化形式。这种容器比VM更轻巧。

[0159] 在一些实施例中，管理程序内核网络接口模块是包括具有管理程序内核网络接口和接收/传送线程的网络堆栈的非-VM DCN。管理程序内核网络接口模块的一个示例是作为VMware公司的ESXi™管理程序的一部分的vmknic模块。

[0160] 应当理解的是，虽然本说明书提到VM，但是给出的示例可以是任何类型的DCN，包

括物理主机、VM、非-VM容器和管理程序内核网络接口模块。事实上，在一些实施例中，示例网络可以包括不同类型的DCN的组合。

[0161] 虽然本发明已经参考许多特定细节进行了描述，但是本领域普通技术人员将认识到，在不脱离本发明的精神的情况下，本发明可以以其它特定形式体现。此外，多个图(包括图9、10、14和19)概念性地示出了过程。这些过程的特定操作可能没有以所示出和描述的确切顺序执行。特定操作可能没有在一系列连续的操作中执行，并且不同的特定操作可能在不同的实施例中执行。此外，过程可以利用若干子过程来实现，或者作为较大的宏过程的一部分来实现。因此，本领域普通技术人员将理解，本发明不受上述说明性细节的限制，而是由所附权利要求来限定。

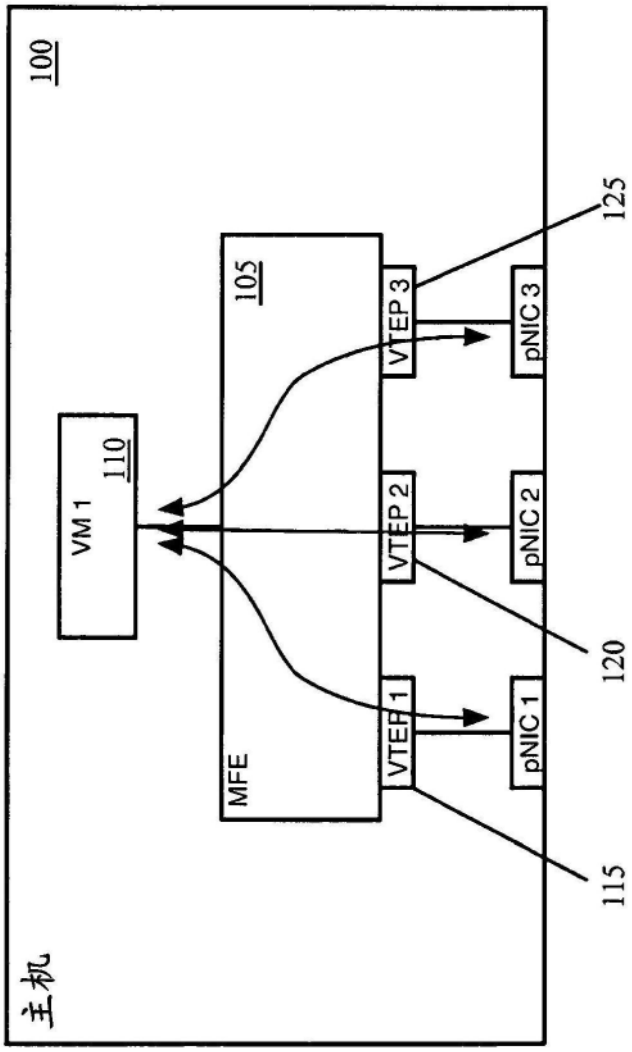

100
105
VM 1
110
MFE
VTEP 3
VTEP 2
VTEP 1
pNIC 3
pNIC 2
pNIC 1
125
120
115
主机

图1

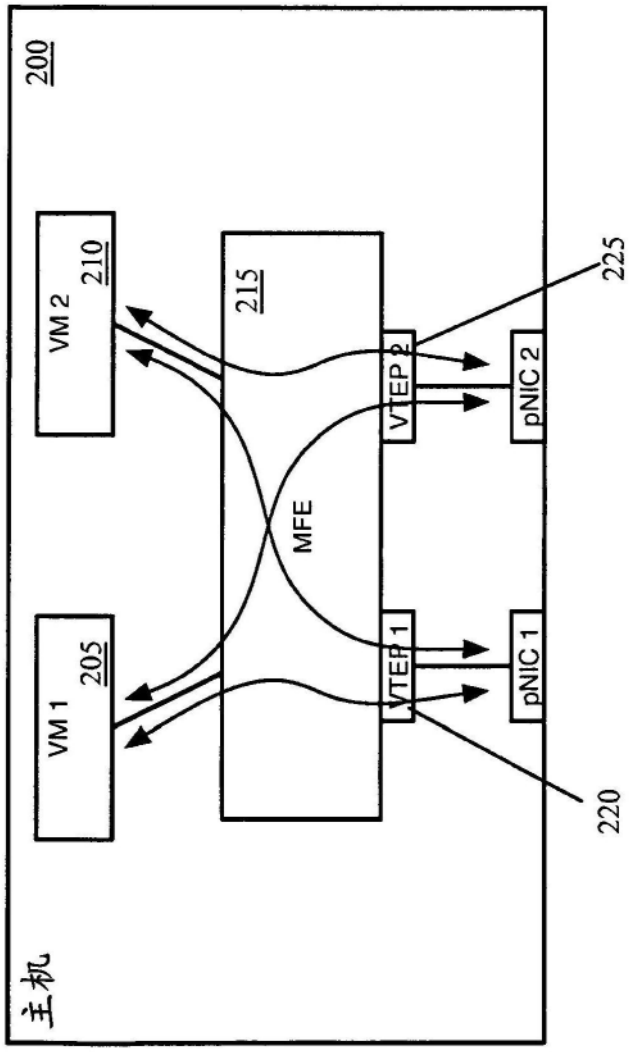


图2

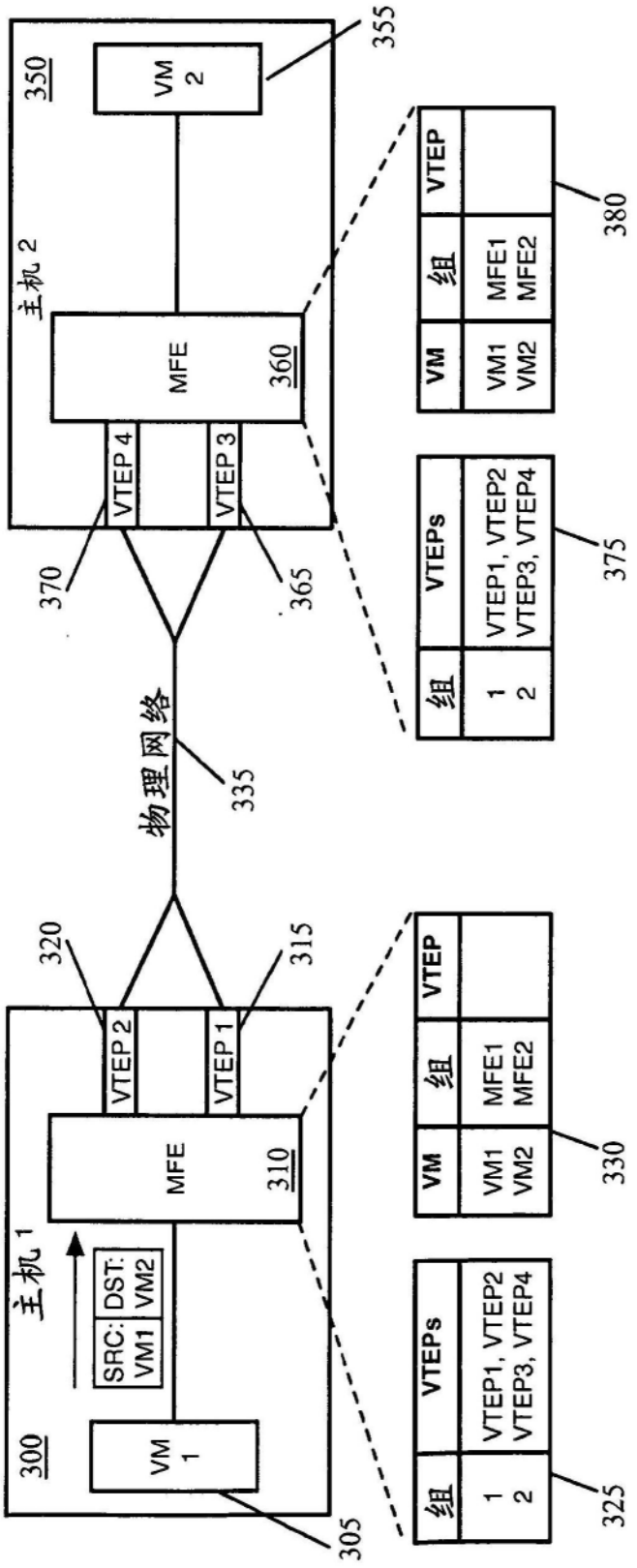

300
主机 1
VM 1
SRC: VM1
DST: VM2
MFE
310
VTEP 2
VTEP 1
320
315
305
物理网络
335
350
主机 2
VM 2
355
MFE
360
VTEP 4
VTEP 3
370
365
组
VTEPs
1
2
VTEP1, VTEP2
VTEP3, VTEP4
325
VM
组
VTEP
VM1
VM2
MFE1
MFE2
330
375
380

图3

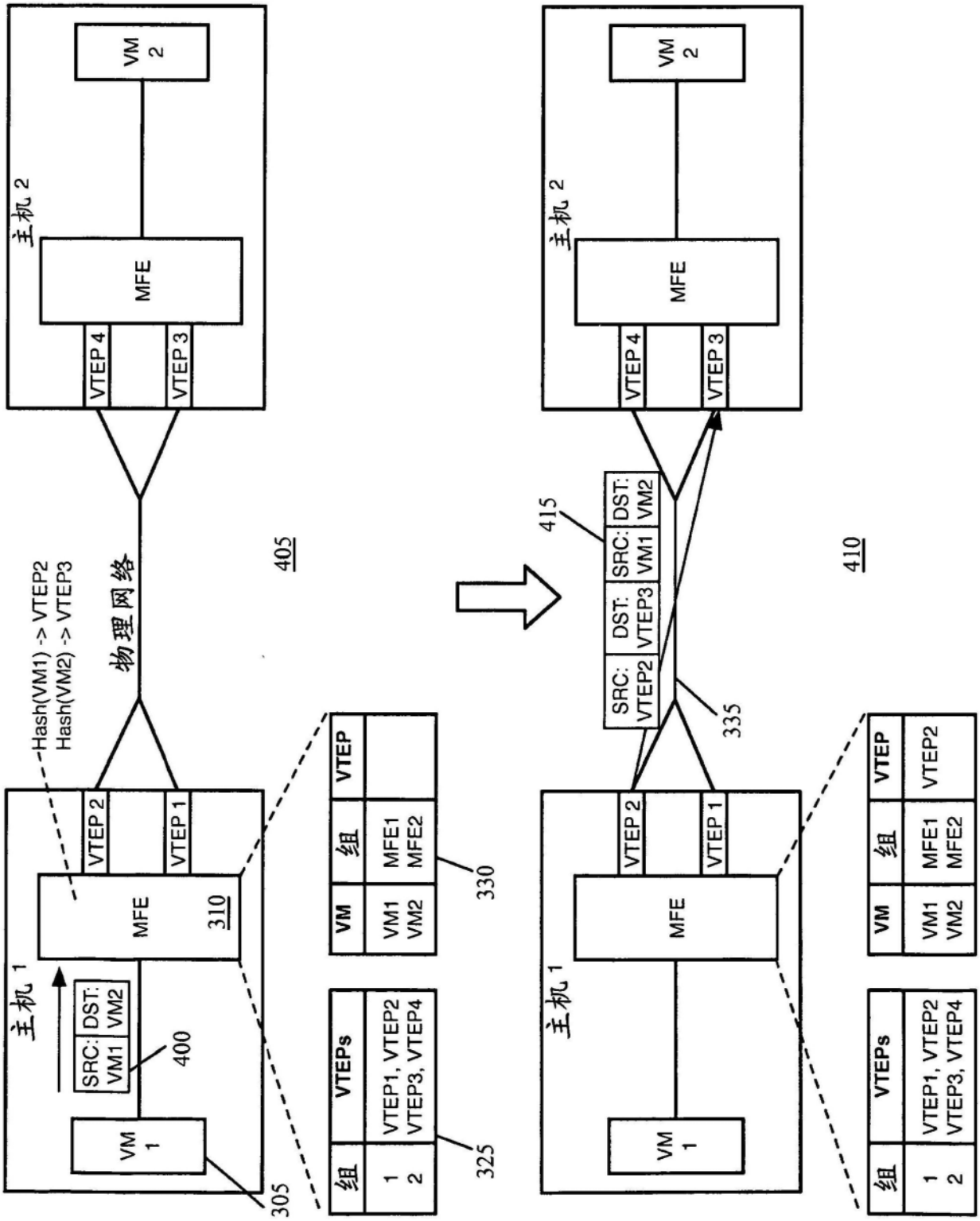

主机 1
主机 2
VM 1
VM 2
MFE
VTEP 1
VTEP 2
VTEP 3
VTEP 4
SRC: VM1
DST: VM2
Hash(VM1) -> VTEP2
Hash(VM2) -> VTEP3
物理网络
400
405
410
415
305
310
325
330
335
组
VTEPs
VM
VTEP
VTEP1, VTEP2
VTEP3, VTEP4
MFE1
MFE2
SRC: VTEP2
DST: VTEP3

图4

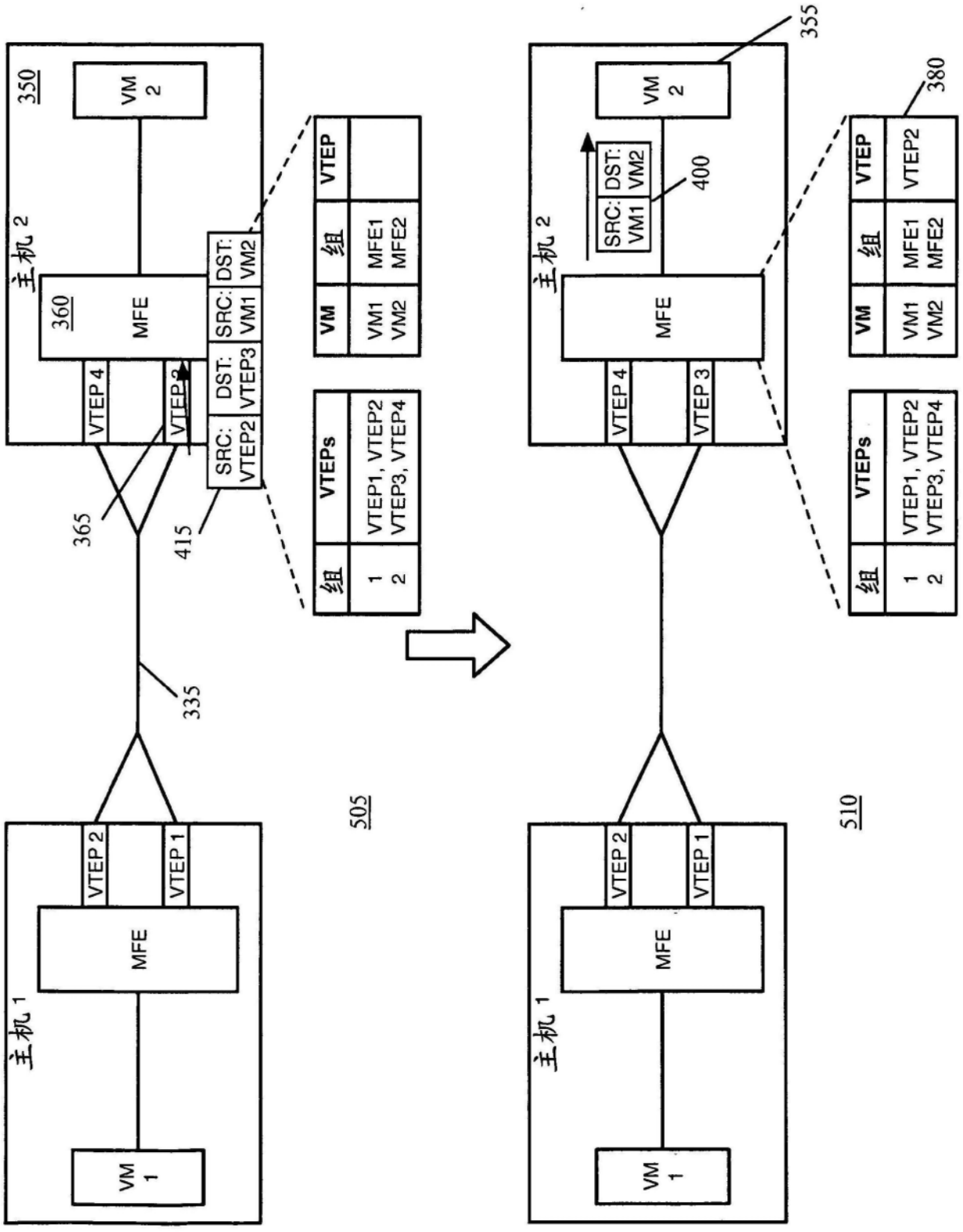

主机 1
VM 1
MFE
VTEP 1
VTEP 2
主机 2
350
VM 2
360
MFE
VTEP 3
VTEP 4
SRC: VTEP2
DST: VTEP3
SRC: VM1
DST: VM2
365
415
335
355
400
380
505
510
组
VTEPs
VTEP1, VTEP2
VTEP3, VTEP4
VM
VM1
VM2
MFE1
MFE2
VTEP

图5

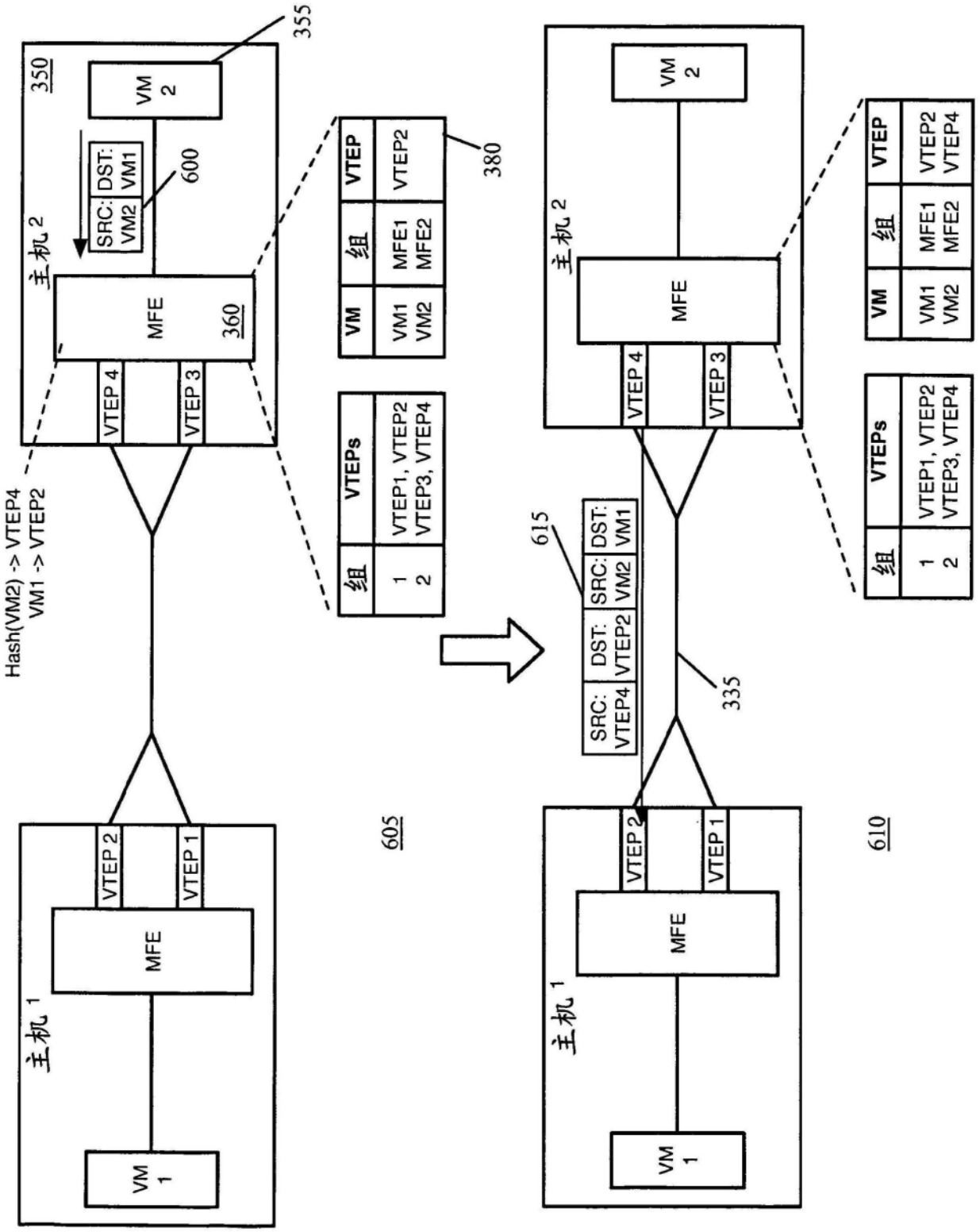

350
355
VM 2
SRC: VM2
DST: VM1
600
主机 2
MFE
360
VTEP 4
VTEP 3
Hash(VM2) -> VTEP4
VM1 -> VTEP2
VM
组
VTEP
VM1 VM2
MFE1 MFE2
VTEP2
380
组
VTEPs
1
2
VTEP1, VTEP2
VTEP3, VTEP4
605
615
SRC: VTEP4
DST: VTEP2
SRC: VM2
DST: VM1
335
610
VTEP 2
VTEP 1
主机 1
VM 1

图6

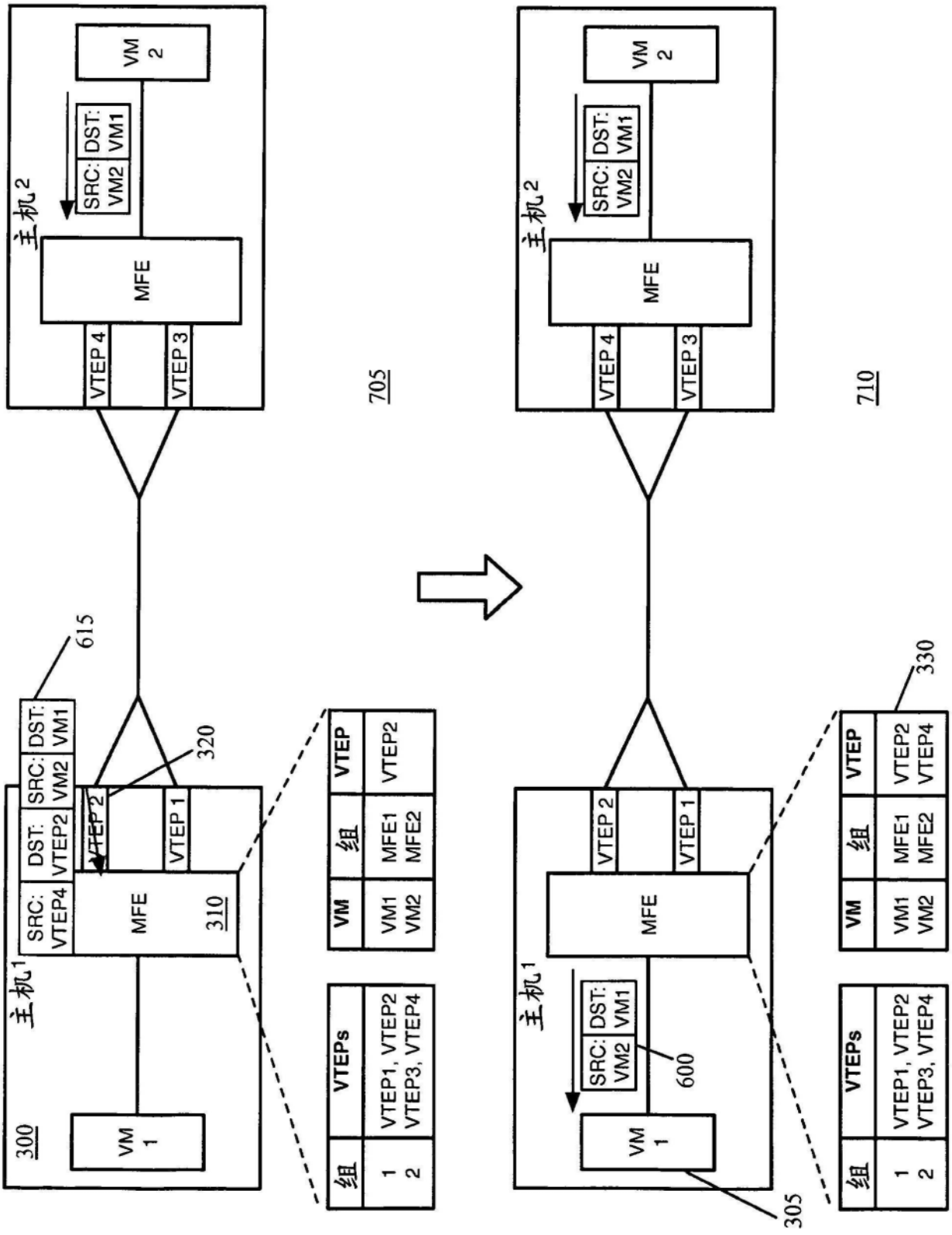

主机2
VM 2
SRC: VM2
DST: VM1
MFE
VTEP 4
VTEP 3
705
710
主机1
300
VM 1
SRC: VTEP4
DST: VTEP2
SRC: VM2
DST: VM1
615
320
VTEP 2
VTEP 1
310
MFE
600
305
330
组
VTEPs
1
2
VTEP1, VTEP2
VTEP3, VTEP4
VM
VTEP
VM1
VM2
MFE1
MFE2
VTEP2
VTEP4

图7

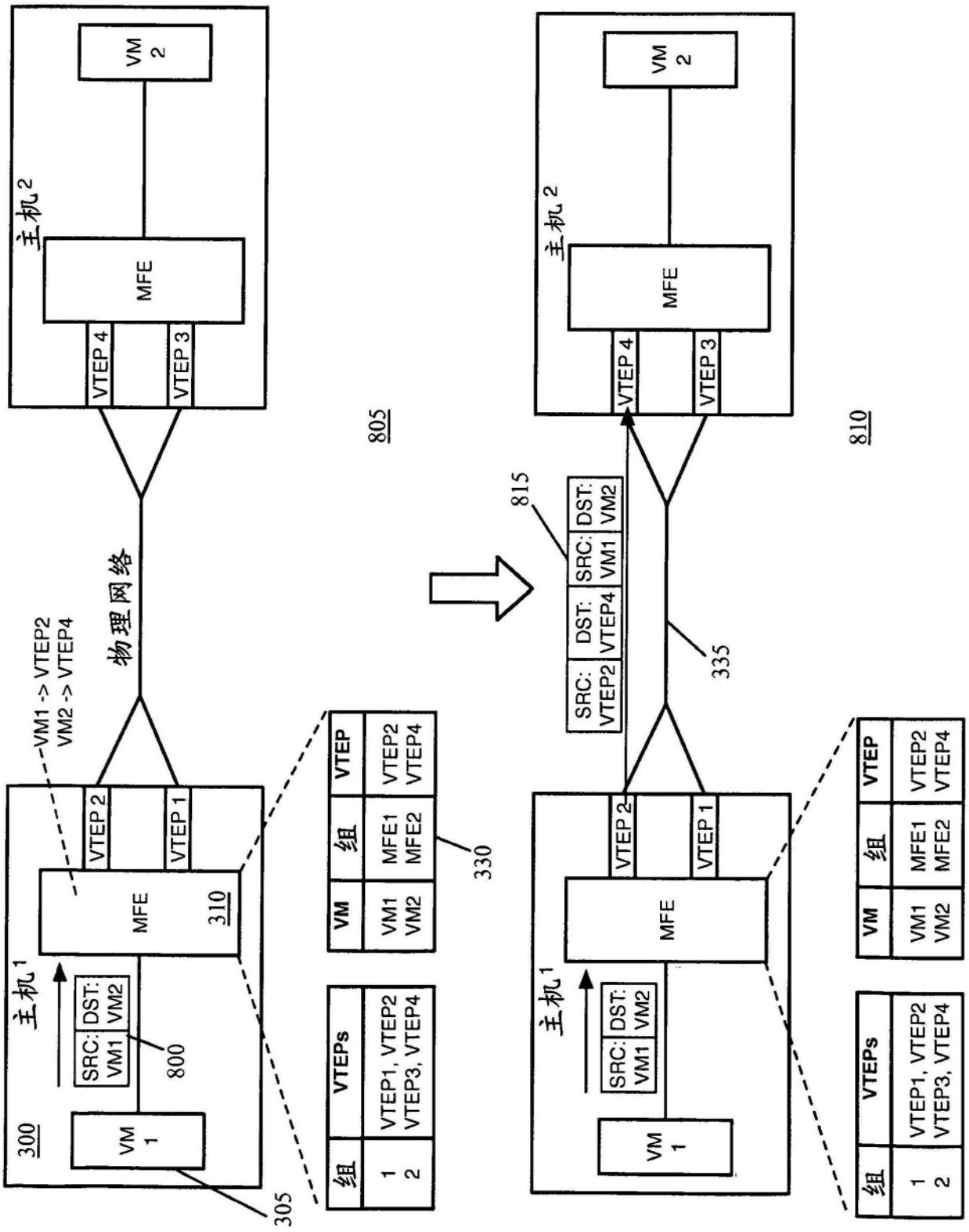

主机2
VM 2
MFE
VTEP 4
VTEP 3
物理网络
VM1 -> VTEP2
VM2 -> VTEP4
主机1
VTEP 2
VTEP 1
MFE
310
SRC: VM1
DST: VM2
800
VM 1
300
305
805
815
SRC: VTEP2
DST: VTEP4
SRC: VM1
DST: VM2
335
810
VM
组
VTEP
VM1 VM2
MFE1 MFE2
VTEP2 VTEP4
330
组
VTEPs
1
2
VTEP1, VTEP2
VTEP3, VTEP4

图8

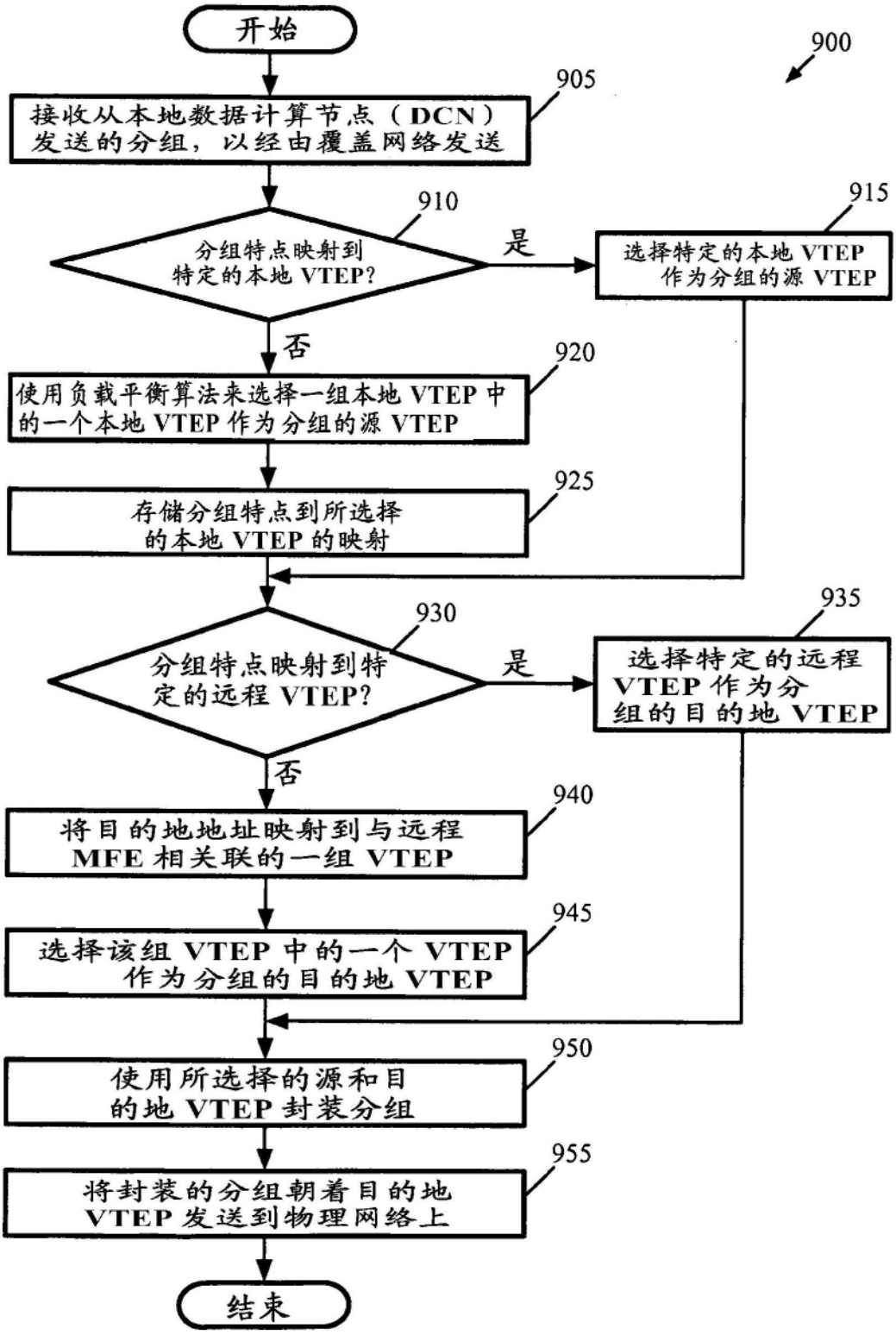

开始
900
905
接收从本地数据计算节点（DCN）发送的分组，以经由覆盖网络发送
910
分组特点映射到特定的本地 VTEP?
是
915
选择特定的本地 VTEP 作为分组的源 VTEP
否
920
使用负载平衡算法来选择一组本地 VTEP 中的一个本地 VTEP 作为分组的源 VTEP
925
存储分组特点到所选择的本地 VTEP 的映射
930
分组特点映射到特定的远程 VTEP?
是
935
选择特定的远程 VTEP 作为分组的目的地 VTEP
否
940
将目的地地址映射到与远程 MFE 相关联的一组 VTEP
945
选择该组 VTEP 中的一个 VTEP 作为分组的目的地 VTEP
950
使用所选择的源和目的地 VTEP 封装分组
955
将封装的分组朝着目的地 VTEP 发送到物理网络上
结束

图9

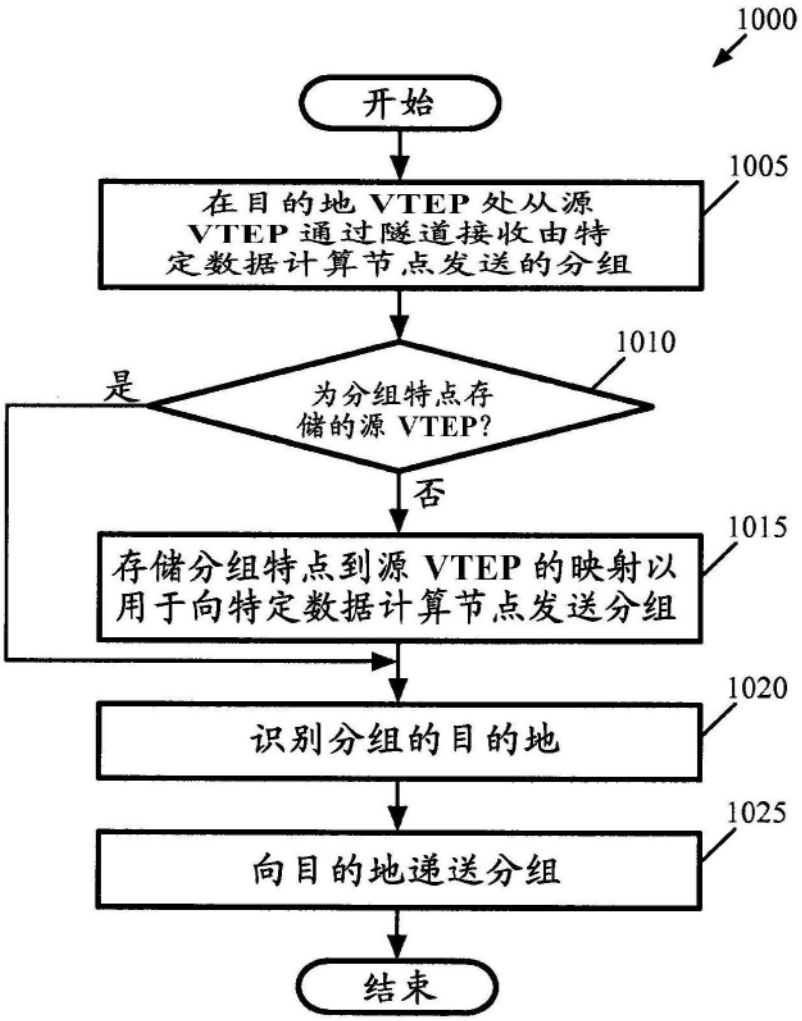

1000
开始
1005
在目的地 VTEP 处从源 VTEP 通过隧道接收由特定数据计算节点发送的分组
1010
为分组特点存储的源 VTEP?
是
否
1015
存储分组特点到源 VTEP 的映射以用于向特定数据计算节点发送分组
1020
识别分组的目的地
1025
向目的地递送分组
结束

图10

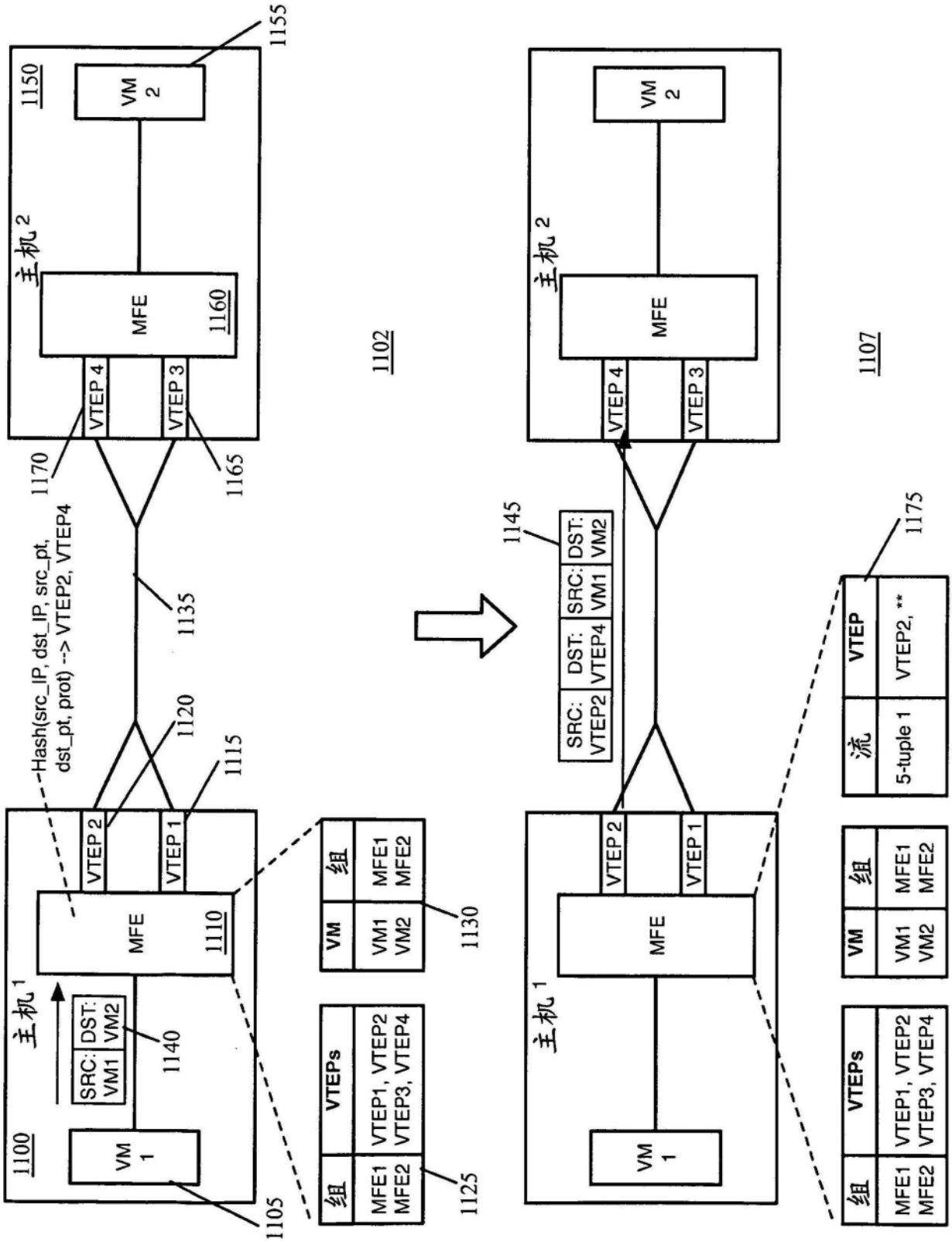

1155
1150
VM 2
主机 2
MFE
1160
VTEP 4
VTEP 3
1170
1165
1102
1107
Hash(src_IP, dst_IP, src_pt, dst_pt, prot) --> VTEP2, VTEP4
1135
1120
1115
1145
SRC: VTEP2
DST: VTEP4
SRC: VM1
DST: VM2
1175
流
VTEP
5-tuple 1
VTEP2, **
VTEP 2
VTEP 1
MFE
1110
主机 1
SRC: VM1
DST: VM2
1140
VM 1
1100
1105
组
VM
MFE1 MFE2
VM1 VM2
1130
VTEPs
VTEP1, VTEP2 VTEP3, VTEP4
1125

图11

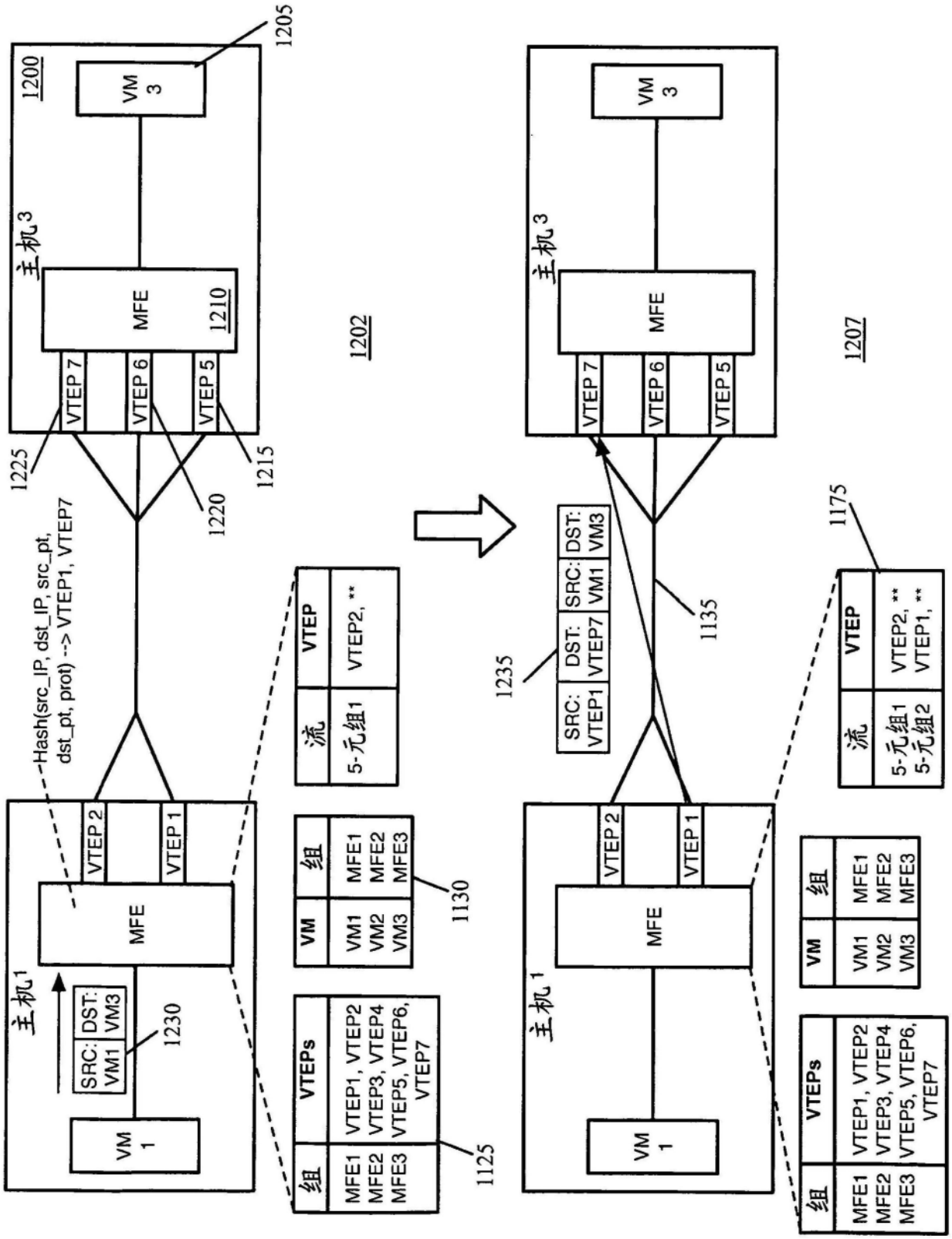

1200
主机3
VM 3
MFE
1210
VTEP 7
VTEP 6
VTEP 5
1205
1225
1220
1215
Hash(src_IP, dst_IP, src_pt, dst_pt, prot) --> VTEP1, VTEP7
1202
主机1
VTEP 2
VTEP 1
MFE
SRC: VM1
DST: VM3
1230
VM 1
流
5-元组1
VTEP
VTEP2, **
VM
组
VM1 VM2 VM3
MFE1 MFE2 MFE3
1130
VTEPs
组
MFE1 MFE2 MFE3
VTEP1, VTEP2 VTEP3, VTEP4 VTEP5, VTEP6, VTEP7
1125
1207
SRC: VTEP1
DST: VTEP7
SRC: VM1
DST: VM3
1235
1135
1175
5-元组1
5-元组2
VTEP2, **
VTEP1, **

图12

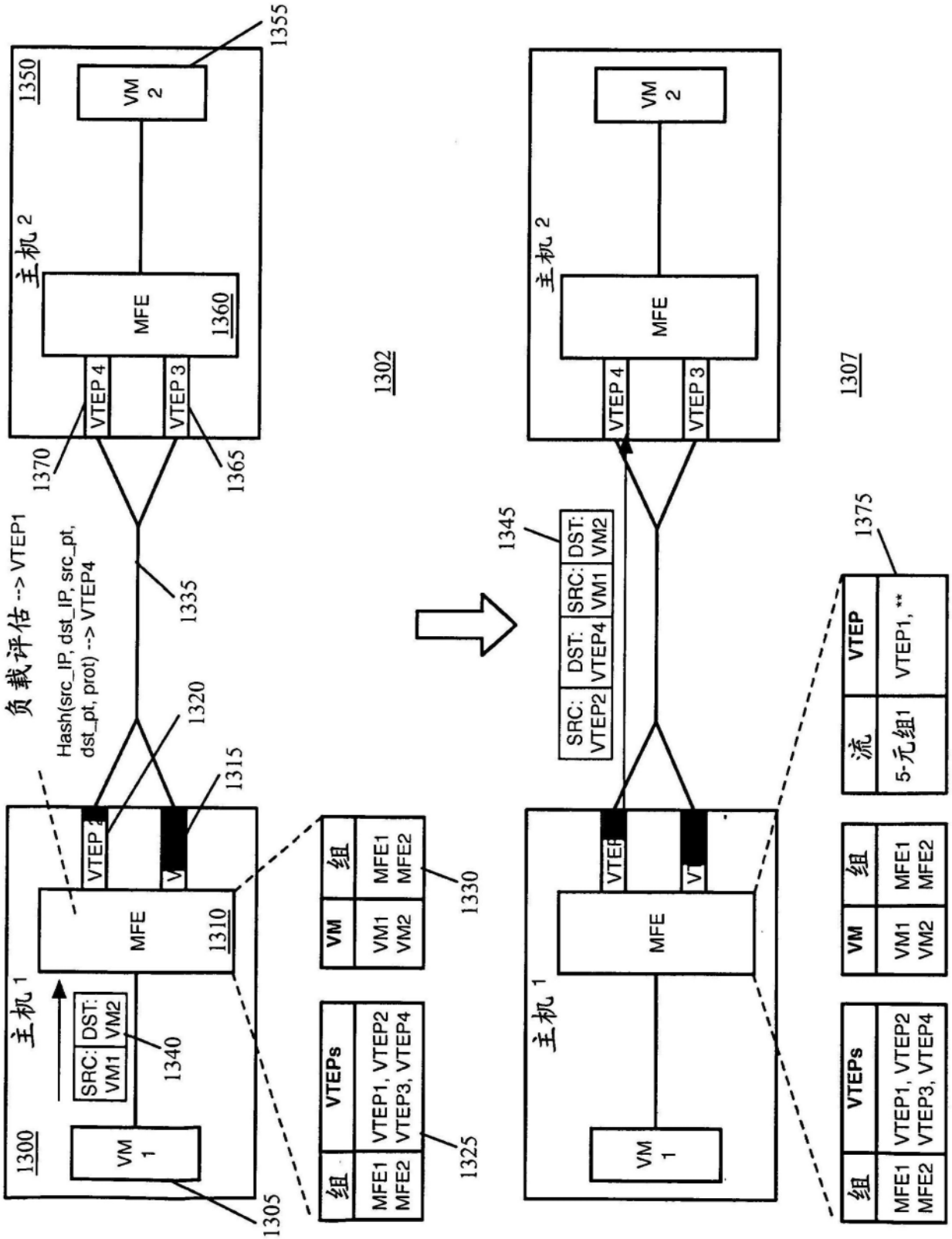

1355
1350
VM 2
主机 2
MFE
1360
VTEP 4
VTEP 3
1370
1365
1335
负载评估 --> VTEP1
Hash(src_IP, dst_IP, src_pt, dst_pt, prot) --> VTEP4
1320
1315
1310
1300
主机 1
SRC: VM1
DST: VM2
1340
VM 1
1305
VTEPs
组
MFE1 MFE2
VTEP1, VTEP2
VTEP3, VTEP4
1325
VM
组
VM1 VM2
MFE1 MFE2
1330
1302
1345
SRC: VTEP2
DST: VTEP4
SRC: VM1
DST: VM2
1307
1375
流
5-元组1
VTEP
VTEP1, **

图13

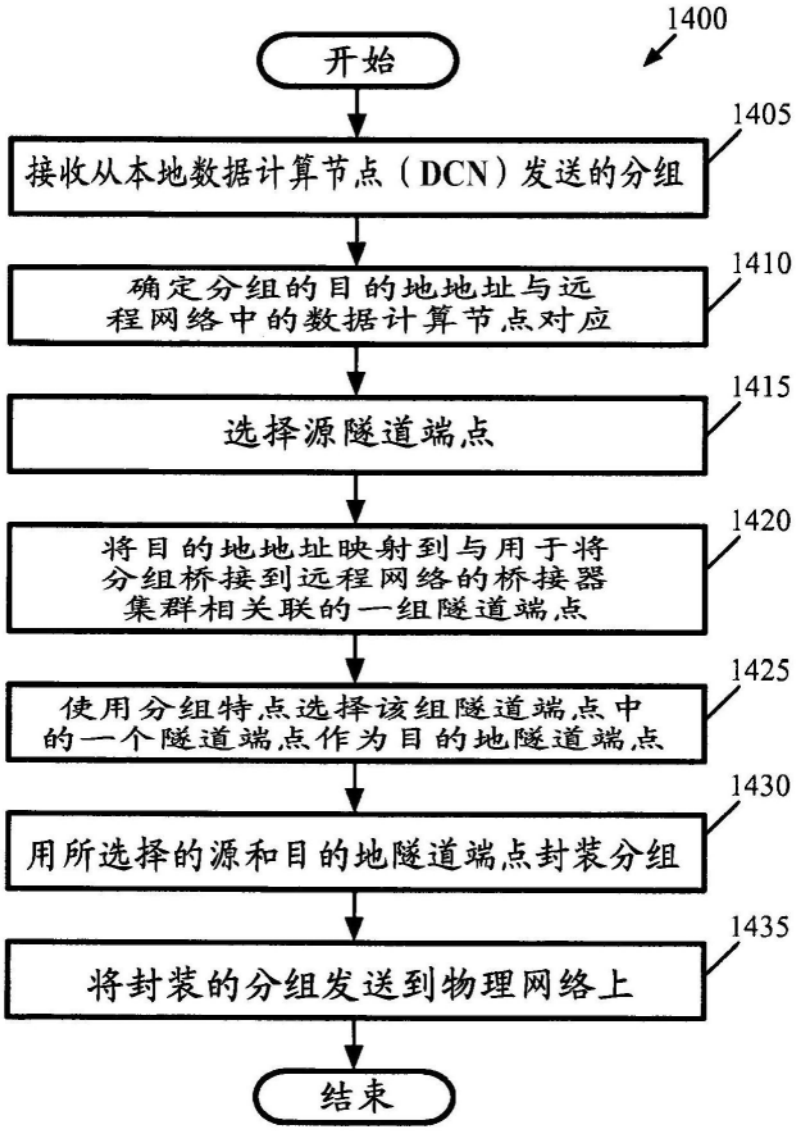

1400
开始
1405
接收从本地数据计算节点（DCN）发送的分组
1410
确定分组的目的地地址与远程网络中的数据计算节点对应
1415
选择源隧道端点
1420
将目的地地址映射到与用于将分组桥接到远程网络的桥接器集群相关联的一组隧道端点
1425
使用分组特点选择该组隧道端点中的一个隧道端点作为目的地隧道端点
1430
用所选择的源和目的地隧道端点封装分组
1435
将封装的分组发送到物理网络上
结束

图14

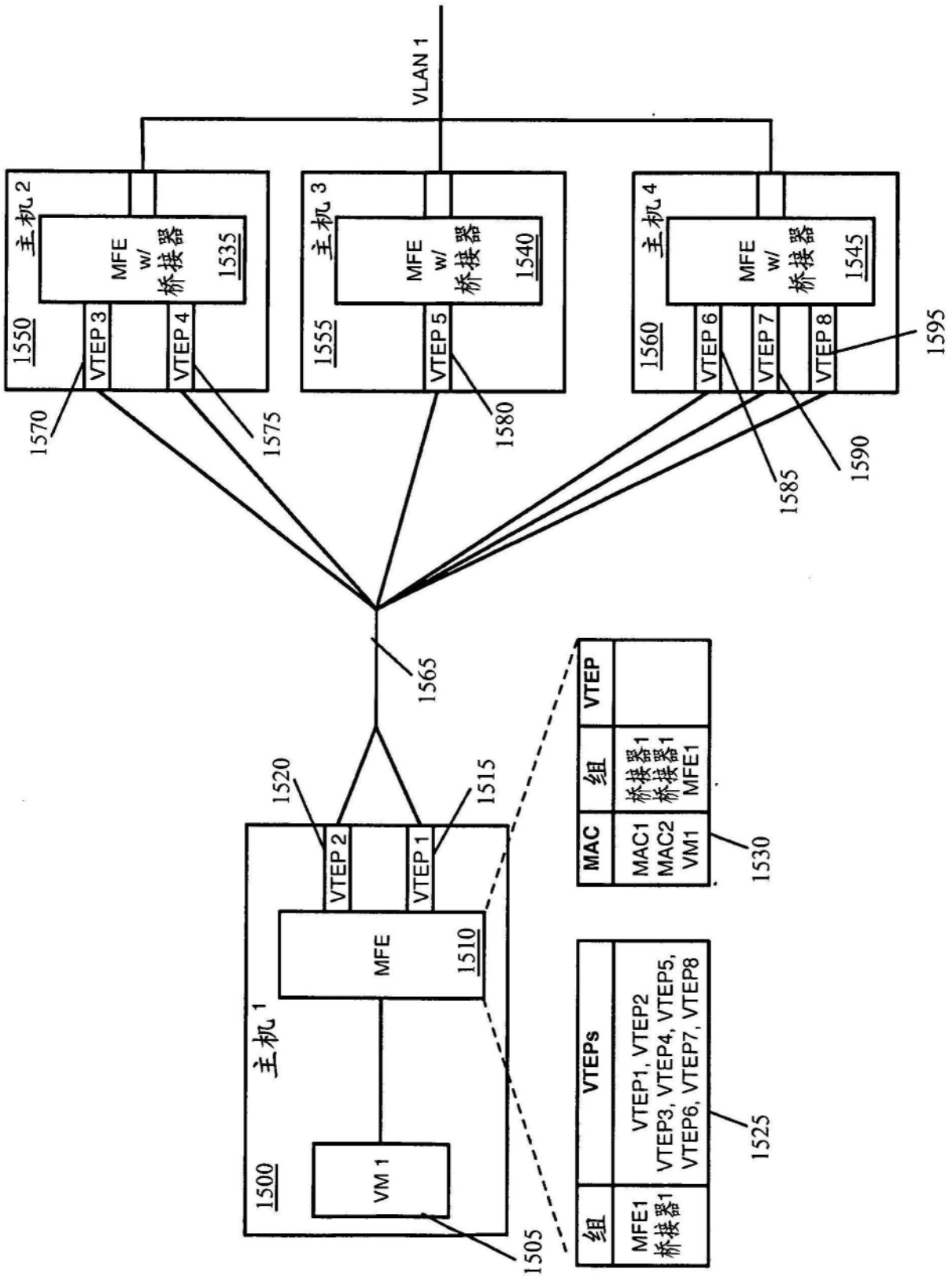

VLAN 1
主机 2
MFE w/ 桥接器
1535
1550
VTEP 3
VTEP 4
主机 3
MFE w/ 桥接器
1540
1555
VTEP 5
主机 4
MFE w/ 桥接器
1545
1560
VTEP 6
VTEP 7
VTEP 8
1595
1570
1575
1580
1585
1590
1565
1520
1515
VTEP 2
VTEP 1
MFE
1510
主机 1
VM 1
1500
1505
MAC
组
VTEP
MAC1
MAC2
VM1
桥接器1
桥接器1
MFE1
1530
VTEPs
VTEP1, VTEP2
VTEP3, VTEP4, VTEP5,
VTEP6, VTEP7, VTEP8
MFE1
桥接器1
1525

图15

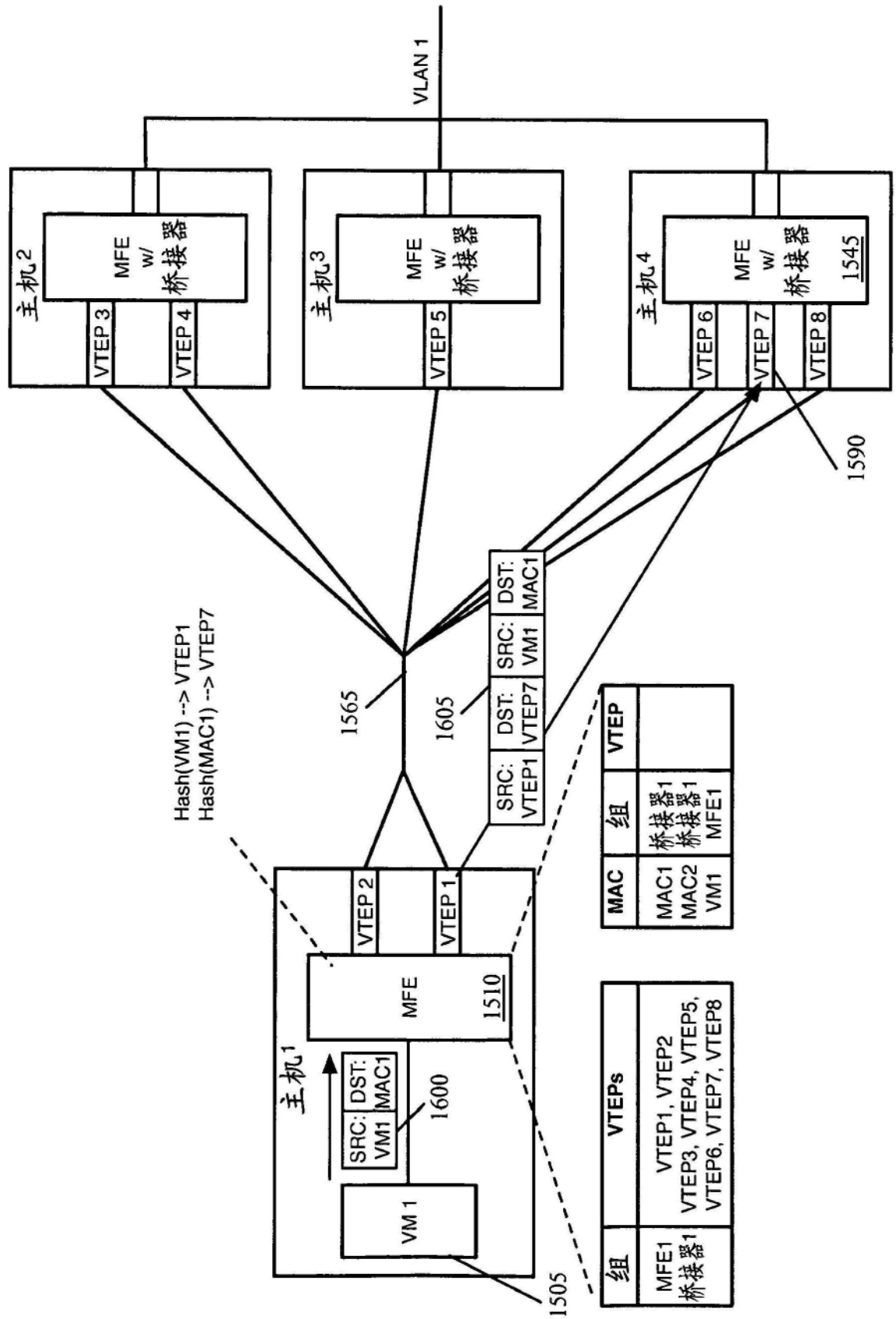

VLAN 1
主机2
MFE w/ 桥接器
VTEP 3
VTEP 4
主机3
MFE w/ 桥接器
VTEP 5
主机4
MFE w/ 桥接器
1545
VTEP 6
VTEP 7
VTEP 8
1590
Hash(VM1) --> VTEP1
Hash(MAC1) --> VTEP7
1565
1605
SRC: VTEP1
DST: VTEP7
SRC: VM1
DST: MAC1
主机1
VTEP 2
VTEP 1
MFE
1510
SRC: VM1
DST: MAC1
1600
VM 1
1505
MAC
组
VTEP
MAC1
MAC2
VM1
桥接器1
桥接器1
MFE1
组
VTEPs
MFE1
桥接器1
VTEP1, VTEP2
VTEP3, VTEP4, VTEP5,
VTEP6, VTEP7, VTEP8

图16

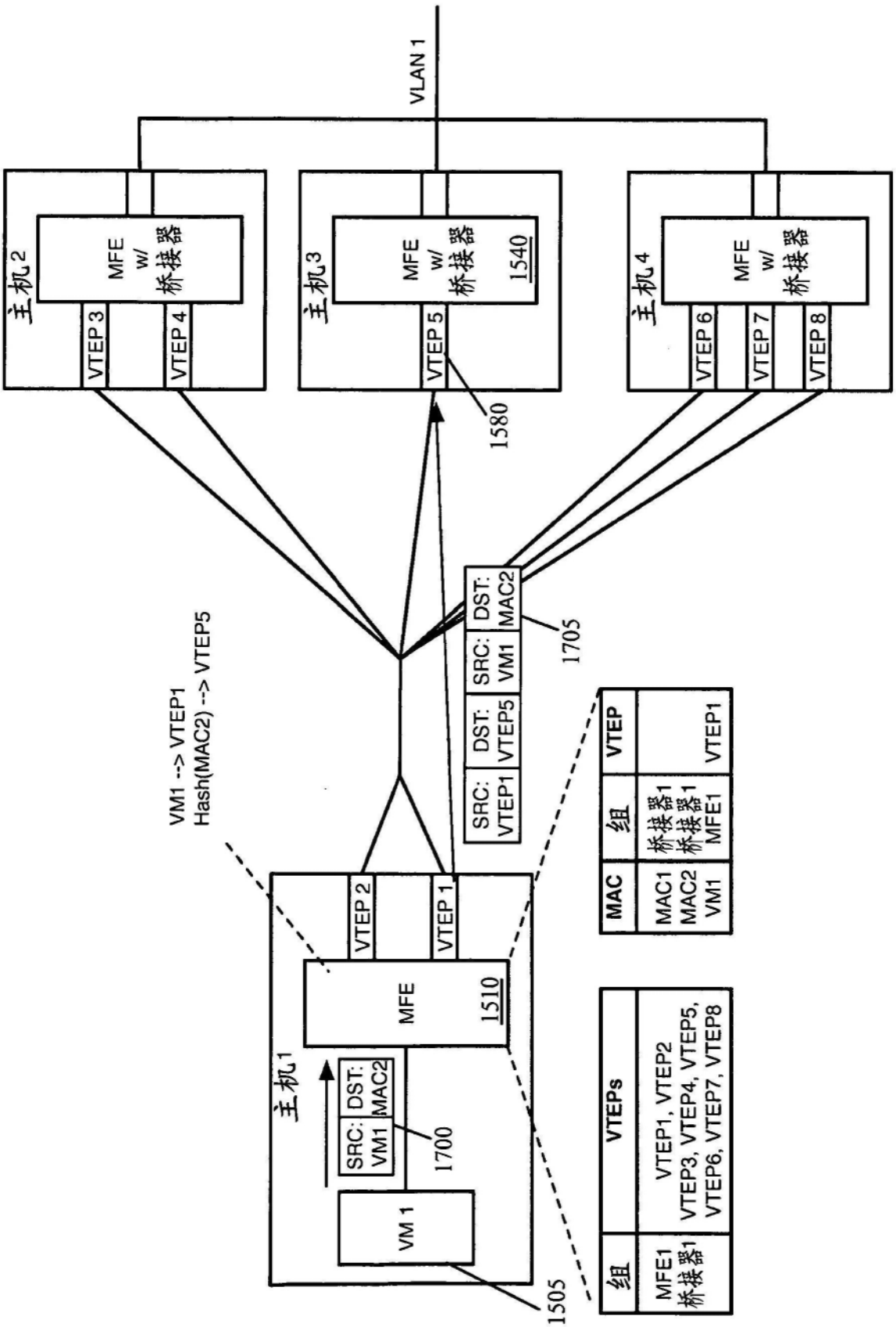

VLAN 1
主机2
MFE w/ 桥接器
VTEP 3
VTEP 4
主机3
MFE w/ 桥接器
1540
VTEP 5
1580
主机4
MFE w/ 桥接器
VTEP 6
VTEP 7
VTEP 8
VM1 --> VTEP1
Hash(MAC2) --> VTEP5
SRC: VTEP1
DST: VTEP5
SRC: VM1
DST: MAC2
1705
VTEP 2
VTEP 1
MFE
1510
主机1
SRC: VM1
DST: MAC2
1700
VM 1
1505
MAC
MAC1
MAC2
VM1
组
桥接器1
桥接器1
MFE1
VTEP
VTEP1
组
MFE1
桥接器1
VTEPs
VTEP1, VTEP2
VTEP3, VTEP4, VTEP5,
VTEP6, VTEP7, VTEP8

图17

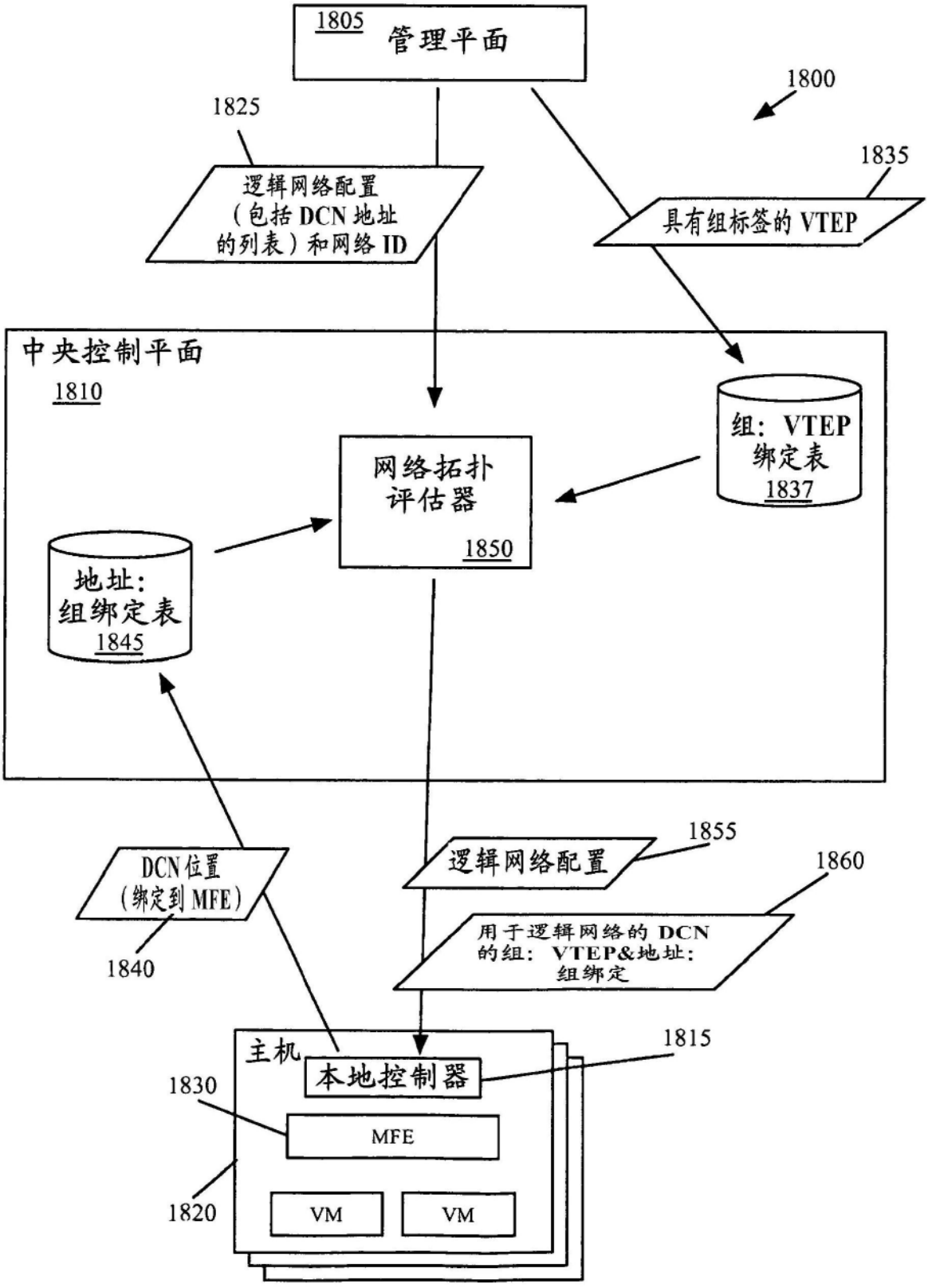

1805
管理平面
1800
1825
逻辑网络配置
（包括 DCN 地址
的列表）和网络 ID
1835
具有组标签的 VTEP
中央控制平面
1810
组：VTEP
绑定表
1837
网络拓扑
评估器
1850
地址：
组绑定表
1845
DCN位置
（绑定到 MFE）
1840
1855
逻辑网络配置
1860
用于逻辑网络的 DCN
的组：VTEP&地址：
组绑定
主机
本地控制器
1815
1830
MFE
1820
VM
VM

图18

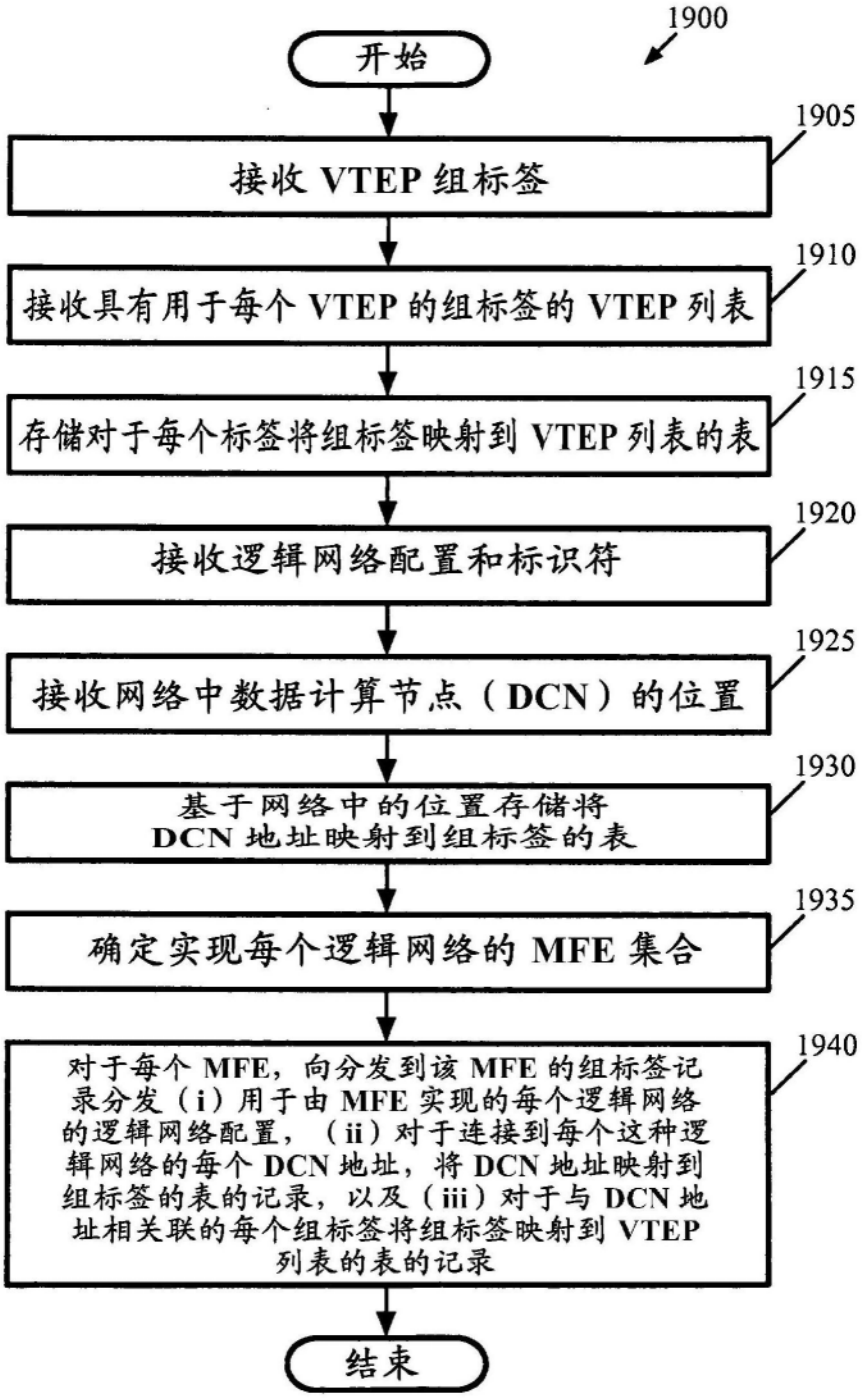

1900
开始
1905
接收 VTEP 组标签
1910
接收具有用于每个 VTEP 的组标签的 VTEP 列表
1915
存储对于每个标签将组标签映射到 VTEP 列表的表
1920
接收逻辑网络配置和标识符
1925
接收网络中数据计算节点（DCN）的位置
1930
基于网络中的位置存储将 DCN 地址映射到组标签的表
1935
确定实现每个逻辑网络的 MFE 集合
1940
对于每个 MFE，向分发到该 MFE 的组标签记录分发（i）用于由 MFE 实现的每个逻辑网络的逻辑网络配置，（ii）对于连接到每个这种逻辑网络的每个 DCN 地址，将 DCN 地址映射到组标签的表的记录，以及（iii）对于与 DCN 地址相关联的每个组标签将组标签映射到 VTEP 列表的表的记录
结束

图19

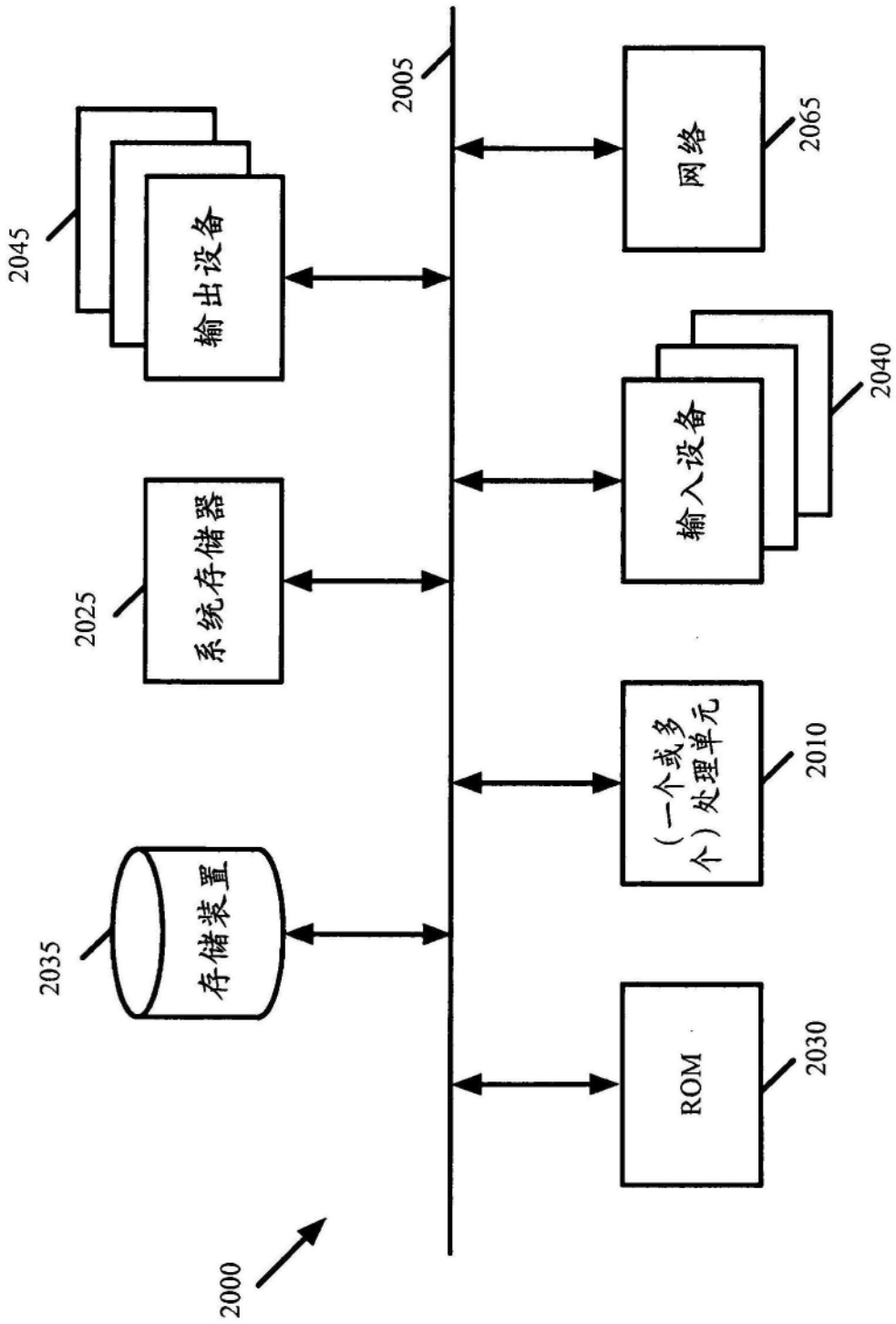

2000
2005
2045
输出设备
2025
系统存储器
2035
存储装置
网络
2065
输入设备
2040
（一个或多个）处理单元
2010
ROM
2030

图20