(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114726786 A
(43)申请公布日 2022.07.08

(21)申请号 202210558749.2

(22)申请日 2014.12.30

(30)优先权数据
14/214,553 2014.03.14 US
14/214,561 2014.03.14 US

(62)分案原申请数据
201480077752.6 2014.12.30

(71)申请人 NICIRA股份有限公司
地址 美国加利福尼亚

(72)发明人 A·图巴尔特塞弗 张荣华
B·C·巴斯勒 S·马斯卡里克
R·拉马纳坦 D·J·莱罗伊
S·奈吉哈尔 范凯伟 A·阿泰卡

(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 45/586(2022.01)
H04L 45/02(2022.01)
H04L 41/044(2022.01)
H04L 41/08(2022.01)
H04L 45/58(2022.01)
H04L 45/00(2022.01)
H04L 45/24(2022.01)

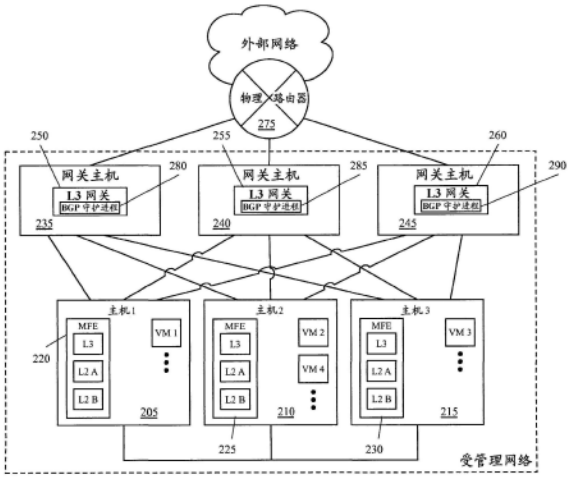
权利要求书3页 说明书36页 附图25页

(54)发明名称

受管理网关的路由通告

(57)摘要

本公开涉及受管理网关的路由通告。一些实施例提供了一种网络系统。该网络系统包括用于托管通过逻辑网络连接到彼此的虚拟机的第一组主机机器。第一组主机机器包括用于在主机机器之间转发数据的受管理转发元件。该网络系统包括用于托管操作为用于在虚拟机和外部网络之间转发数据的网关的虚拟化容器的第二组主机机器。虚拟化容器中的至少一个与外部网络中的至少一个物理路由器对等,以便向物理路由器通告虚拟机的地址。



CN 114726786 A

1.一种用于在受管理网络中的主机计算机上执行的网关的方法,所述主机计算机与所述受管理网络外部的至少一个路由器接口,其中跨所述受管理网络的多个主机计算机实现逻辑网络,所述方法包括:

路由(i)从外部路由器接收的并且被定向到所述逻辑网络中的目的地的分组,以及(ii)从跨其实现所述逻辑网络的主机计算机接收的并且被定向到所述逻辑网络外部的目的地地址的分组;以及

与所述外部路由器对等以便所述逻辑网络向所述外部路由器通告通过所述网关可达的所述逻辑网络中的网络地址集合。

2.如权利要求1所述的方法,其中:

所述逻辑网络包括虚拟机附接到的至少一个逻辑交换机;

所述虚拟机具有网络地址范围;以及

所通告的网络地址集合包括所述网络地址范围。

3.如权利要求1所述的方法,还包括:向所述外部路由器通告所述网关的接口作为所述逻辑网络中的所述网络地址集合的下一跳。

4.如权利要求1所述的方法,其中所述网关由在所述主机计算机上执行的命名空间实现。

5.如权利要求1所述的方法,其中与所述外部路由器对等包括:执行路由协议应用以建立与所述外部路由器的相邻性并与所述外部路由器交换路由信息。

6.如权利要求5所述的方法,其中守护程序在所述主机计算机上执行以接收定义所述路由协议应用的配置的数据库记录,并将所接收的数据库记录转换成所述路由协议应用的配置文件。

7.如权利要求6所述的方法,其中所述守护进程在所述主机计算机的虚拟化软件内执行。

8.如权利要求6所述的方法,其中所述配置包括定义所述路由协议应用的设置的数据、通过所述网关可达的所述网络地址集合、以及定义要与其对等的所述外部路由器的数据。

9.如权利要求8所述的方法,其中定义所述外部路由器的所述数据包括网络地址、自主系统编号、保持活动时间和用于所述外部路由器的抑制定时器。

10.如权利要求1所述的方法,其中:

所述逻辑网络是跨第一多个主机机器实现的第一逻辑网络,并且所述网关是第一网关;

第二逻辑网络是跨所述受管理网络的第二多个主机计算机实现的;并且

在所述主机计算机上执行的第二网关:

(i)路由从所述外部路由器接收的并且被定向到所述第二逻辑网络中的目的地的分组以及(ii)从跨其实现所述第二逻辑网络的主机计算机接收的并且被定向到所述第二逻辑网络外部的目的地地址的分组;以及

与所述外部路由器对等以便所述第二逻辑网络向所述外部路由器通告所述第二逻辑网络中通过所述网关可达的网络地址集合。

11.一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1-10中任一项所述的方法。

12.一种计算设备,包括:

一组处理单元;以及

存储程序的机器可读介质,所述程序在由所述处理单元中的至少一个实现时实现如权利要求1-10中任一项所述的方法。

13.一种系统,包括用于实现如权利要求1-10中任一项所述的方法的装置。

14.一种包括指令的计算机程序产品,所述指令在由计算机执行时使所述计算机执行如权利要求1-10中任一项所述的方法。

15.一种系统,包括:

第一组主机计算机,用于托管包括至少一个子网的逻辑网络的终端机器,其中用于在所述逻辑网络的终端机器之间转发数据消息的受管理转发元件在所述第一组主机计算机上执行;以及

第二组主机计算机,用于托管一组网关,所述网关用于在(i)所述逻辑网络的终端机器与(ii)所述逻辑网络外部的一组网络之间转发数据消息,

其中所述一组网关与一组外部物理路由器对等以将所述逻辑网络的一组子网通告给所述一组外部物理路由器。

16.如权利要求15所述的系统,还包括一组网络控制器,用于通过生成和分发数据以提供(i)所述第一组主机计算机上的受管理转发元件和(ii)所述第二组主机计算机上的网关来管理所述第一组主机计算机和所述第二组主机计算机。

17.如权利要求16所述的系统,其中所述一组网络控制器包括:

第一网络控制器,用于为所述受管理转发元件和所述网关生成配置数据以实现所述逻辑网络;以及

在所述第一组主机计算机和所述第二组主机计算机的所述主机计算机上执行的多个控制器,用于转换所述配置数据并且将所述配置数据提供给所述受管理转发元件和所述网关。

18.如权利要求15所述的系统,其中所述网关在所述第二组主机机器上的虚拟化容器内被实现。

19.如权利要求15所述的系统,其中所述一组网关中的第一网关与所述第一外部物理路由器对等,并且所述一组网关中的第二网关与所述第二外部物理路由器对等。

20.如权利要求19所述的系统,其中所述第一网关将所述逻辑网络的第一子网通告给所述第一外部物理路由器,并且所述第二网关将所述逻辑网络的第二子网通告给所述第二外部物理路由器。

21.如权利要求19所述的系统,其中所述第一网关和所述第二网关各自将所述逻辑网络的一组子网分别通告给所述第一外部物理路由器和所述第二外部物理路由器。

22.如权利要求19所述的系统,其中所述第一网关还与所述第二外部物理路由器对等,并且所述第二网关还与所述第一外部物理路由器对等。

23.如权利要求15所述的系统,其中路由协议应用在所述第二组主机计算机的每个主机计算机上执行,以将在所述主机计算机上托管的网关通告为针对所述逻辑网络子网中的网络地址的下一跳。

24.如权利要求15所述的系统,其中所述逻辑网络是第一逻辑网络,所述一组外部物理

路由器是第一组外部物理路由器，并且所述一组网关是第一组网关，所述系统还包括：

第三组主机计算机，用于托管包括至少一个子网的第二逻辑网络的终端机器，其中用于在所述第三组主机计算机之间转发数据消息的受管理转发元件在所述第三组主机计算机上执行；以及

第四组主机计算机，用于托管用于在(i)所述第二逻辑网络的所述终端机器与(ii)所述第二逻辑网络外部的一组网络之间转发数据消息的第二组网关，

其中所述第二组网关与所述第二组外部物理路由器对等以将所述第二逻辑网络的一组子网通告给所述第二组外部物理路由器。

25.如权利要求24所述的系统，其中所述第一组主机计算机和所述第三组主机计算机具有至少一个共用的主机计算机，并且所述第二组主机计算机和所述第四组主机计算机具有至少一个共用的主机计算机。

受管理网关的路由通告

[0001] 本申请是申请日为2014年12月30日、申请号为201480077752.6、发明名称为“受管理网关的路由通告”的首次提交的发明专利申请的第一代分案申请的分案申请，第一代分案申请的申请日为2014年12月30日，申请号为202010257206.8且发明名称为“受管理网关的路由通告”。

技术领域

[0002] 本发明通常涉及受管理网关的路由通告。

背景技术

[0003] 在诸如互联网的物理L3网络中，路由器利用包括边界网关协议(BGP)的各种路由协议交换路由和可达性信息。BGP的主要功能是允许两个路由器交换通告可用路由或不再可用路由的信息。即，第一路由器可以使用这个协议来通知第二路由器，去往给定IP地址或IP前缀的分组可以被发送到第一路由器。第二路由器然后可以使用这个信息来计算路由。

[0004] 在一些受管理的虚拟化网络内，路由由网络控制器计算并且向下推送到在受管理网络内处理路由的转发元件。由于控制器指示这些转发元件将如何路由分组，因此没有必要在转发元件之间交换路由信息。但是，这些受管理的虚拟化网络可以通过外部网络发送和接收流量。当前，这需要管理员手动向外部网络中的路由器提供路由。

发明内容

[0005] 一些实施例提供了网络控制系统，其使得在由该网络控制系统管理的网络中操作的逻辑网络能够与受管理网络外部的物理路由器对等并且向其通告路由信息。在一些实施例中，逻辑网络包括至少部分地在受管理网关中实现的逻辑路由器，并且这些网关使用路由协议(例如，边界网关协议)与外部物理路由器对等。当多个受管理网关实现逻辑路由器(或逻辑路由器的至少与外部网络接口的部分)时，在一些实施例中，这些多个网关可以单独地将相同路由通告给外部路由器，由此允许外部路由器跨多个网关分发去往通告的目的地的流量。

[0006] 在一些实施例中，逻辑路由器连接虚拟机逻辑上附连到其的一组逻辑交换机。每个逻辑交换机表示特定的一组IP地址(即，子网)，并且在受管理网络中跨虚拟机物理上连接到其(例如，通过虚拟接口)的一组受管理转发元件实现。在一些实施例中，逻辑路由器也由连接到虚拟机的受管理转发元件以分布式方式实现。但是，当逻辑路由器也经由一个或多个端口连接到外部网络时，到外部网络的这些连接通过使用一个或多个网关来实现。在一些实施例中，网关既负责从受管理网络向外部未受管理物理网络发送数据流量又负责处理从外部网络发送到受管理网络中的流量。

[0007] 在一些实施例中，用户(例如，管理员)配置逻辑网络，包括具有连接到外部网络的一个或多个端口的逻辑路由器，用于在受管理网络内实现。此外，用户可以经由这些端口指定逻辑路由器应该与外部网络中的物理路由器对等，以便交换路由信息。在接收到逻辑网

络配置时，负责管理逻辑路由器的网络控制器(或控制器集群)选择一组网关，用于实现到外部网络的连接。在一些实施例中，当逻辑路由器的这些端口已被指定用于与外部路由器对等时，网络控制器将每个这种端口分配给不同的网关。在一些实施例中，这些网关在网络中跨网关的集群散布，使得每个端口被实现在不同的故障域中。

[0008] 选定的网关利用诸如边界网关协议(BGP)的路由协议与外部路由器对等。在一些实施例中，控制器基于逻辑网络配置生成路由协议数据。对于面向外部网络的逻辑路由器的每个端口，控制器识别(i)实现该端口的网关将与其对等的这组外部路由器(即，该网关的邻居)，以及(ii)实现该端口的网关将通告的这组路由。这些路由可以简单地是表示连接到逻辑路由器的逻辑交换机的IP前缀，或者可以附加地包括由用户输入的或由实现逻辑路由器的进程动态生成的其它路由。在一些实施例中，逻辑路由器的不同端口可以将其路由通告给不同的外部网络路由器。一旦网络控制器生成该数据，连同用于网关中逻辑路由器实现的路由表数据，网络控制器就将数据分发到网关(例如，通过网络控制器的层次结构)。

[0009] 在一些实施例中，其上实现逻辑路由器的网关是分组为集群的主机机器，其被分配用于托管逻辑路由器和用于逻辑网络的其它服务。这些网关机器还包括受管理转发元件，其用作发送到VM驻留在其上的受管理转发元件和从这些受管理转发元件发送的分组的隧道端点。一些实施例在诸如命名空间的具有存储路由表能力的虚拟化容器内实现逻辑路由器。此外，一些实施例操作路由协议应用或命名空间中的守护进程(daemon)(例如，BGP守护进程)。在一些情况下，网关主机机器可以具有操作不同逻辑路由器的若干个命名空间，其中的一些或全部包括用于与外部路由器对等的路由协议应用。

[0010] 一个或多个守护进程可以在命名空间之外的网关主机机器上操作(例如，在网关的虚拟化软件中)，以便接收定义路由表和用于特定命名空间的路由协议配置两者的数据元组。这一个或多个守护进程操作来实例化命名空间、向命名空间指配路由表、以及在命名空间中启动路由协议应用。此外，在一些实施例中，(一个或多个)守护进程生成用于路由协议应用的配置文件，并且存储配置文件(例如，在主机机器的文件系统中)以供被路由协议应用访问。

[0011] 一旦安装其配置文件，路由协议应用就开始与外部路由器通信。在一些实施例中，应用以与标准物理路由器将根据其与其邻居交换信息的方式相同的方式行动。例如，一些实施例的BGP守护进程打开与在其配置中识别为邻居的每个路由器的BGP会话、发送如由BGP指定的保持活动(keep-alive)消息、以及经由BGP分组将其路由通告给识别的邻居。在一些实施例中，BGP守护进程还接收由其邻居发送的BGP分组，并且使用这些分组来识别路由。一些实施例的BGP守护进程或者在其本地路由表中(即，在相同的命名空间内)安装路由，或者将路由向上推送到网络控制器使得网络控制器可以为实现逻辑路由器的网关路由表计算新的路由表，或者其组合。但是，在其它实施例中，路由的通告只在一个方向上工作，其中BGP守护进程将路由发送出到其邻居，但不安装从那些邻居接收到的路由。即，BGP守护进程既不将接收到的路由向上推送到网络控制器，也不在本地网关的路由表中安装路由。

[0012] 在一些情况下，实现同一逻辑路由器的多个网关(例如，实现不同端口)可以将(例如，到达特定逻辑交换机上的VM，或者到达由那些VM共享的公共IP的)相同路由通告给同一外部路由器。在这种情况下，外部路由器将把这些多个网关视为用于发送到通告地址的分组的相同成本(equal-cost)的下一跳。因此，在一些实施例中，外部路由器跨通告路由的各

种网关散布发送到那些目的地的分组。外部路由器可以使用各种不同的相同成本的多路径(equal-cost multi-path,ECMP)技术中的任意种来确定分组应该被发送到哪个网关。

[0013] 在上述实施例中,路由协议应用内联(inline)驻留。即,应用在网关上操作,网关是分组通过其被发送的位置。但是,在一些实施例中,(一个或多个)网络控制器为网关充当路由服务器,并且路由协议应用驻留在控制器上。在这种情况下,路由协议配置不被控制器分发到网关,而是用来在控制器上实例化路由协议应用。控制器然后将路由信息通告给外部路由器(并且潜在地从外部路由器接收通告的路由信息)。这种通告的信息向外部路由器通知哪些网关用于哪些路由。如在内联的情况下,外部路由器可以使用ECMP技术来在若干个网关之间分发发送到逻辑网络的分组。

[0014] 前面的发明内容旨在用作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性主题的介绍或概述。以下的具体实施方式和具体实施方式所参考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阅读。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限定,这是因为所要求保护的主题可以在不背离本主题的精神的情况下以其它特定的形式来体现。

附图说明

[0015] 本发明的新颖特征在所附权利要求中阐述。但是,出于解释的目的,本发明的若干种实施例在以下图中阐述。

[0016] 图1概念性地示出了包括逻辑路由器的一些实施例的逻辑网络体系架构。

[0017] 图2概念性地示出了图1的逻辑网络的物理实现。

[0018] 图3概念性地示出了用于指配(provision)受管理转发元件、L3网关和路由协议应用,以便实现逻辑网络和使那些网络的逻辑路由器能够与外部路由器对等的一些实施例的网络控制系统。

[0019] 图4概念性地示出了数据通过一些实施例的分层网络控制系统的传播。

[0020] 图5概念性地示出了用于生成和分发数据,以便实现受管理网络中的逻辑路由器与外部网络之间的一组连接的一些实施例的过程。

[0021] 图6概念性地示出了网关主机的五个单独的集群,以及在那些网关主机上实现的逻辑路由器端口(被称为上行链路)。

[0022] 图7概念性地示出了用于网关主机机器的一些实施例的软件体系架构。

[0023] 图8概念性地示出了用于设置或修改网关主机机器上的L3网关的一些实施例的过程。

[0024] 图9概念性地示出了由一些实施例的路由协议应用(例如,BGP守护进程)执行,以便将路由通告给用于L3网关的外部路由器的一些实施例的过程。

[0025] 图10示出了逻辑网络以及该逻辑网络在受管理网络中的物理实现。

[0026] 图11概念性地示出了由操作来控制受管理网络的控制器集群指配图10的网关主机上的三个命名空间中的BGP守护进程。

[0027] 图12概念性地示出了根据一些实施例的由命名空间中的BGP守护进程发送的BGP

更新分组。

[0028] 图13和14概念性地示出了由进入到图10的受管理网络中的流量采用的路径。

[0029] 图15示出了两个逻辑网络和那些逻辑网络在受管理网络中的物理实现。

[0030] 图16示出了由控制器集群在图15的七个命名空间中指配BGP守护进程。

[0031] 图17概念性地示出了一旦在各个命名空间中运行的守护进程已与路由器建立邻接关系，由图15的各个BGP守护进程发送到外部路由器的BGP更新分组。

[0032] 图18概念性地示出了由进入到图15的受管理网络中的三个分组采用的路径。

[0033] 图19概念性地示出了用于生成用于逻辑网络的BGP配置数据并且然后由在生成该数据的控制器中的BGP服务实现那个配置数据的一些实施例的过程。

[0034] 图20示出了逻辑网络以及那个逻辑网络在其中控制器充当路由服务器的受管理网络中的物理实现。

[0035] 图21概念性地示出了为了使图20的逻辑网络的逻辑路由器实现，由控制器集群发送的数据。

[0036] 图22概念性地示出了由进入图20的受管理网络的若干个分组采用的路径。

[0037] 图23概念性地示出了充当用于逻辑网络的路由服务器的一些实施例的控制器的软件体系架构。

[0038] 图24概念性地示出了其中实现逻辑网络并且其使用单独的网关作为路由服务器的一些实施例的这种受管理网络。

[0039] 图25概念性地示出了本发明的一些实施例利用其实现的电子系统。

具体实施方式

[0040] 在本发明的以下具体实施方式中，阐述和描述了本发明的许多细节、例子和实施例。但是，对于本领域技术人员将清楚和显而易见的是，本发明不限于所阐述的实施例，并且本发明在没有所讨论的一些具体细节和例子的情况下也可以被实践。

[0041] 一些实施例提供了网络控制系统，其使得在由该网络控制系统管理的网络中操作的逻辑网络能够与受管理网络外部的物理路由器对等并且向其通告路由信息。在一些实施例中，逻辑网络包括至少部分地在受管理网关中实现的逻辑路由器，并且这些网关使用路由协议(例如，边界网关协议)与外部物理路由器对等。当多个受管理网关实现逻辑路由器(或逻辑路由器的至少与外部网络接口的部分)时，在一些实施例中，这些多个网关可以单独地将相同路由通告给外部路由器，由此允许外部路由器跨多个网关分发去往被通告的目的地的流量。

[0042] 图1概念性地示出了逻辑网络体系架构100的例子。逻辑网络100包括两个逻辑交换机105和110以及逻辑路由器115。逻辑交换机105和110中的每一个连接若干个虚拟机(在这种情况下，两个虚拟机(VM)由每个逻辑交换机连接，并且逻辑路由器115将两个逻辑交换机(即，逻辑层2域)连接在一起)。此外，逻辑路由器经由三个逻辑端口将逻辑网络连接到外部网络120。虽然在这个例子中，逻辑路由器115具有若干个连接到外部网络的端口(例如，作为上行链路端口)，但是在一些实施例中，逻辑路由器可以只具有单个连接到外部网络的端口。

[0043] 在一些实施例中，逻辑网络是由管理员生成的网络的抽象概念，并且逻辑网络以

虚拟化、分布式的方式在受管理的物理基础设施中(例如,在多租户数据中心中)实现。即,连接到逻辑交换机的虚拟机可以驻留在基础设施内的各种不同的主机机器上,并且在这些主机机器上操作的物理受管理转发元件(例如,软件虚拟交换机)实现逻辑转发元件(逻辑交换机,逻辑路由器等)中的一些或全部。

[0044] 如在这个例子中,逻辑路由器连接虚拟机逻辑上附连到其的一组逻辑交换机。每个逻辑交换机表示特定的一组IP地址(即,子网),并且在受管理网络中跨虚拟机物理上连接到其(例如,通过虚拟接口)的一组受管理转发元件实现。在一些实施例中,逻辑路由器也由连接到虚拟机的受管理转发元件以分布式方式实现。但是,当逻辑路由器也经由一个或多个端口连接到外部网络时,到外部网络的这些连接通过使用一个或多个网关来实现。在一些实施例中,网关既负责从受管理网络向外部未受管理物理网络发送数据流量又负责处理从外部网络发送到受管理网络中的流量。

[0045] 图2概念性地示出了逻辑网络100的这种物理实现。该图示出了包括三个主机机器205-215和三个网关主机机器235-245的受管理网络200。逻辑网络100的VM驻留在主机205-215上,在在主机中操作的虚拟化软件(例如,管理程序、虚拟机监视器等)之上实现。连接到其它逻辑网络的附加虚拟机可以驻留在这些主机以及未在该图示出的受管理网络中的附加主机中的一些或全部上。

[0046] 除了虚拟机之外,主机205-215中的每一个操作受管理转发元件(MFE)220-230。在一些实施例中,该MFE是在主机的虚拟化软件内操作的虚拟交换机(例如,Open vSwitch、或另一个软件转发元件)。在图2所示的例子中,MFE 220-230中每一个实现逻辑交换机105和110两者,以及逻辑路由器115。在一些实施例中,这使得能够进行第一跳逻辑处理,其中对分组的逻辑处理的全部或大部分在接收该分组的第一MFE处执行。因此,从VM 1发送到VM 4的分组将由MFE 220通过逻辑交换机105到逻辑路由器115并且然后到逻辑交换机110被处理。MFE 220将把用于分组的逻辑交换机110的逻辑出口端口识别为VM 4附连到其的端口,并且在主机210处将该出口端口映射为到MFE 230的隧道。

[0047] 在一些实施例中,网络控制器(或控制器集群)通过生成流条目,或MFE转换成流条目的数据元组指配MFE 220-230。这些流条目指定匹配条件(例如,物理入口端口、逻辑入口端口、目的地MAC或IP地址、传输层5元组等)和在匹配条件的分组上采取的动作(例如,将分组分配到逻辑转发元件、分配逻辑出口端口、将数据写入寄存器、在特定的隧道中进行封装等)。因此,为了使MFE通过逻辑网络处理分组,MFE将分组匹配到第一流条目、执行动作(例如,以修改分组或在用于分组的寄存器中存储逻辑上下文数据)、重新提交分组以便匹配另一个流条目等。

[0048] 一些实施例的网关主机机器235-245托管逻辑网络100的用于在外部网络120和逻辑网络100之间实现连接的L3网关250-260(具体而言,逻辑路由器115)。当物理路由器275接收到具有对应于逻辑网络100的虚拟机中的一个的目的地地址,或由逻辑交换机上的VM共享的公共IP的分组时,物理路由器275将分组发送到网关主机235-245中的一个。网关主机235-245还包括MFE,并且在一些实施例中,这些MFE从物理路由器275接收分组并且将分组移交给在其各自主机中的L3网关以供处理。

[0049] 在一些实施例中,用户(例如,管理员)配置逻辑网络100。在接收到具有连接到外部网络的若干个逻辑路由器端口的这种配置时,网络控制器(或控制器集群)选择该组网关

主机机器235-245用于实现这一连接。具体而言，一些实施例选择不同的网关主机机器用于这些逻辑路由器端口中的每一个。在一些实施例中，这些网关在网络中跨网关的集群散布，使得每个端口被实现在不同的故障域中。网络控制器为其一部分在网关主机机器处实现并且其一部分由MFE(例如，MFE 220-230和在网关主机机器235-245上的那些)实现的逻辑路由器计算路由表。

[0050] L3网关250-260为南北流量(即，发送到受管理网络中和从受管理网络中出来的流量)实现逻辑路由器115的路由表的部分。一些实施例只处理入口流量，其中传出流量通过其它手段发送(例如，通过主机机器220-230中的MFE和物理路由器275或外部网络120中的其它网络元件之间的直接连接)。在其它实施例中，L3网关处理入口和出口流量两者。

[0051] 如所示出的，L3网关250-260每一个包括边界网关协议(BGP)守护进程280-290。这些守护进程280-290与外部物理路由器275对等，并且向逻辑路由器115通告到这个路由器的路由。在一些实施例中，BGP守护进程280-290在其与其邻居交换信息方面以与传统物理路由器相同的方式操作。例如，这些BGP守护进程会打开与物理路由器275的BGP会话、发送如由该协议指定的保持活动(keep-alive)消息、以及经由BGP分组将其路由通告给物理路由器275。在一些实施例中，BGP守护进程还接收由物理路由器275发送的BGP分组，并且使用这些分组来识别路由。一些实施例的BGP守护进程或者在其本地路由表中(即，在相同的命名空间内)安装路由，或者将路由向上推送到网络控制器使得网络控制器可以为实现逻辑路由器的所有L3网关计算新的路由表，或者其组合。但是，在其它实施例中，BGP守护进程只在一个方向上有效地工作，即将路由发送出到其邻居(以吸引入口流量)，但不安装从那些邻居接收到的路由。也就是，BGP守护进程既不将接收到的路由向上推送到网络控制器，也不在本地网关的路由表中安装这些路由。

[0052] 在一些实施例中，L3网关与(一个或多个)物理路由器的对等是用户指定的逻辑端口的属性。在一些实施例中，当用户(例如，管理员)为逻辑路由器指定与外部路由器对等时，控制器基于逻辑网络配置生成路由协议数据。对于面向外部网络的逻辑路由器的每个端口，控制器识别(i)实现该端口的网关将与其对等的这组外部路由器(即，该网关的邻居)以及(ii)实现该端口的网关将通告的这组路由。这些路由可以简单地是表示连接到逻辑路由器的逻辑交换机的IP前缀，或者可以附加地包括由用户输入的或由实现逻辑路由器的进程动态生成的其它路由。在一些实施例中，逻辑路由器的不同端口可以将其路由通告给不同的外部网络路由器。一旦网络控制器生成该数据，连同用于L3网关的路由表数据，网络控制器就将数据分发到网关(例如，通过网络控制器的层次结构)。

[0053] 在一些实施例中，L3网关250-260是具有存储路由表能力的虚拟化容器，诸如命名空间。此外，BGP守护进程280-290或其它路由协议应用根据从控制器接收到的数据在这些容器内操作。一个或多个守护进程可以在容器之外的网关主机机器上(例如，在网关的虚拟化软件中)操作，以便从控制器接收定义路由表和用于特定命名空间的BGP配置两者的数据元组。这一个或多个守护进程操作来实例化命名空间、向命名空间指配路由表、以及在命名空间中启动BGP守护进程。此外，在一些实施例中，(一个或多个)守护进程生成用于BGP守护进程的配置文件，并且存储配置文件(例如，在主机机器的文件系统中)以供被路由协议应用访问。一旦安装其配置文件，BGP守护进程就开始与外部路由器邻居通信。

[0054] 在图2所示的例子中，实现同一逻辑路由器115(例如，实现不同端口)的多个网关

235-245将(例如,到达逻辑交换机105和110上的VM的)相同路由通告给同一外部路由器275。在一些实施例中,外部路由器把这些多个L3网关视为用于发送到被通告地址的分组的相同成本的下一跳。因此,在一些实施例中,外部路由器跨通告路由的各种网关散布发送到那些目的地的分组。外部路由器可以使用各种不同的相同成本的多路径(ECMP)技术中的任意种来确定分组应该被发送到哪个网关。

[0055] 在上述实施例中,路由协议应用内联驻留。即,应用在网关上操作,网关是分组通过其被发送的位置。但是,在一些实施例中,(一个或多个)网络控制器为网关充当路由服务器,并且路由协议应用驻留在控制器上。在这种情况下,路由协议配置不被控制器分发到网关,而是用来在控制器上实例化路由协议应用。控制器然后将路由信息通告给外部路由器(并且潜在地从外部路由器接收被通告的路由信息)。这种被通告的信息向外部路由器通知哪些网关用于哪些路由。如在内联的情况下,外部路由器可以使用ECMP技术来在若干个网关之间分发发送到逻辑网络的分组。

[0056] 以上描述介绍了一些实施例的逻辑网络对BGP的使用,但是本领域普通技术人员将认识到,本发明不限于BGP,并且可以使用其它路由协议。下面描述若干个更详细的实施例。首先,第I节描述了由网络控制器指配网关。然后第II节描述了在一些实施例中托管L3网关的主机机器的体系架构。接着,第III节描述了在网关上配置路由协议应用的过程,并且第IV节描述了一些实施例的路由协议应用的操作。然后第V节描述了在一些实施例中使用网络控制器作为路由服务器。最后,第VI节描述了本发明的一些实施例利用其实现的电子系统。

[0057] I.由网络控制器指配网关

[0058] 如所提到的,在一些实施例中,网络控制系统在用于逻辑网络的一个或多个网关中设置和配置逻辑路由器和相关联的路由协议应用。除了选择所述一个或多个网关主机机器用于逻辑路由器之外,网络控制系统中的一个或多个网络控制器还接收由管理员输入的网络配置并且将该信息转换成可以被网关主机机器读取的数据元组。网络控制系统还将数据元组分发到这些主机机器。

[0059] 图3概念性地示出了用于指配受管理转发元件、L3网关和路由协议应用,以便实现逻辑网络和使那些网络的逻辑路由器能够与外部路由器对等的一些实施例的这种网络控制系统300。如所示出的,网络控制系统300包括输入变换控制器305、逻辑控制器310、物理控制器315和320、主机机器325-340、以及两个网关主机机器345和350。如所示出的,主机325-340以及网关主机345和350包括受管理转发元件,其可以实现如在以上图中示出的逻辑转发元件(例如,通过使用流条目)。网关主机345和350还每个包括L3网关,用于处理进入受管理网络和/或从受管理网络外出的分组。这些L3网关附加地包括BGP功能(例如,以BGP守护进程的形式)。本领域普通技术人员将认识到,对于网络控制系统300,各种控制器和主机的许多其它不同的组合是可能的。

[0060] 在一些实施例中,网络控制系统中的每个控制器是具有用作输入变换控制器、逻辑控制器和/或物理控制器的能力的计算机(例如,具有基于x86的处理器)。可替代地,在一些实施例中,给定控制器可以只具有操作为各种类型的控制器中特定一种(例如,只作为物理控制器)的功能。此外,控制器的不同组合可以运行在同一物理机器中。例如,输入变换控制器305和逻辑控制器310可以运行在数据中心管理应用与其交互(或管理员直接与其交

互)的同一计算设备中。

[0061] 一些实施例的输入变换控制器305包括变换从用户接收到的网络配置信息的输入变换应用。虽然在图3中被示为直接从用户接收信息,但是在一些实施例中,用户与数据中心管理应用交互,数据中心管理应用又将网络配置信息传递给输入变换控制器。

[0062] 例如,用户可以指定诸如在图1中示出的网络拓扑结构。对于每个逻辑交换机,用户指定连接到逻辑交换机的机器(即,VM被分配到逻辑交换机的哪些逻辑端口)。用户也可以指定哪些逻辑交换机附连到任何逻辑路由器、用于连接到外部网络的逻辑路由器的一个或多个逻辑端口、以及这些逻辑端口是否与外部物理路由器对等。在一些实施例中,输入变换控制器305将接收到的网络拓扑结构变换为将网络拓扑结构描述为一组数据元组的逻辑控制平面数据。例如,条目可能指出特定的MAC地址A位于特定逻辑交换机的第一逻辑端口X处、逻辑路由器Q位于特定逻辑交换机的第二逻辑端口Y处、或逻辑路由器Q的逻辑端口G是与外部网络接口的上行链路端口。

[0063] 在一些实施例中,每个逻辑网络由特定的逻辑控制器(例如,逻辑控制器310)管理。一些实施例的逻辑控制器310将定义逻辑网络和组成逻辑网络的逻辑转发元件(例如,逻辑路由器、逻辑交换机)的逻辑控制平面数据变换为逻辑转发平面数据,并且将逻辑转发平面数据变换为物理控制平面数据。在一些实施例中,逻辑转发平面数据包括在逻辑层描述的流条目。对于在逻辑端口X处的MAC地址A,逻辑转发平面数据可能包括指定分组的目的地是否匹配MAC A的流条目,以将分组转发到端口X。逻辑路由器Q的端口也将具有MAC地址,并且类似的流条目被创建,用于将具有这个MAC地址的分组转发到逻辑交换机的端口Y。此外,一些实施例的逻辑转发平面数据包括用于将具有未知IP地址的分组发送到例如逻辑端口G的流条目。

[0064] 在一些实施例中,逻辑控制器将逻辑转发平面数据变换为通用物理控制平面数据。通用物理控制平面数据使一些实施例的网络控制系统即使在网络包括大量受管理转发元件(例如,数千个)来实现逻辑转发元件时,和在网络实现大量逻辑网络时也能够缩放。为了表达物理控制平面数据,而无需考虑MFE中的差异和/或MFE的位置细节,通用物理控制平面抽象了不同MFE的共同特点。

[0065] 如上所述,一些实施例的逻辑控制器310将逻辑控制平面数据变换为逻辑转发平面数据(例如,包括在诸如逻辑地址、逻辑入口端口等逻辑网络参数上的匹配的逻辑流条目),然后将逻辑转发平面数据变换为通用物理控制平面数据。在一些实施例中,逻辑控制器应用栈包括用于执行第一变换的控制应用和用于执行第二变换的虚拟化应用。在一些实施例中,这两种应用都使用用于将第一组表映射到第二组表中的规则引擎。即,不同的数据平面被表示为表(例如,nLog表),并且控制器应用使用表映射引擎(例如,nLog引擎)在平面之间进行变换(例如,通过在表上应用连接操作)。在一些实施例中,输入和输出表存储定义不同数据平面的数据元组的集合。

[0066] 物理控制器315和320中的每一个是一个或多个受管理转发元件的主管(例如,位于主机机器内)。在这个例子中,这两个物理控制器中的每一个是位于VM主机机器325-340处的两个受管理转发元件的主管。此外,物理控制器315是两个MFE以及用于特定逻辑网络的L3网关驻留其上的两个网关主机345和350的主管。在一些实施例中,用于逻辑路由器的所有L3网关由同一物理控制器管理(如同在这个图中),但是在其它实施例中,不同的物理

控制器管理用于逻辑网络的不同网关主机。

[0067] 在一些实施例中,物理控制器接收用于逻辑网络的通用物理控制平面数据,并且将该数据变换为用于特定MFE的定制物理控制平面数据,该特定MFE由物理控制器管理并且需要用于特定逻辑网络的数据。在其它实施例中,物理控制器将适当的通用物理控制平面数据传递给具有自己执行这种变换的能力(例如,以在主机机器上运行的机架控制器的形式)的MFE。

[0068] 通用物理控制平面到定制物理控制平面的变换涉及定制流条目中的各种数据。对于上述例子,通用物理控制平面将涉及若干个流条目(即,若干个数据元组)。第一条目指出,如果分组匹配特定的逻辑数据路径集合(例如,基于在特定物理入口端口处接收到的分组)并且目的地地址匹配MAC A,则将分组转发到逻辑端口X。在一些实施例,该条目在通用和定制物理控制平面中将是相同的。附加条目被生成,以将物理入口端口(例如,主机机器的虚拟接口)匹配到逻辑入口端口X(用于从具有MAC A的VM接收到的分组),以及将目的地逻辑端口X匹配到物理MFE的物理出口端口(例如,再次主机机器的虚拟接口)。但是,这些物理入口和出口端口特定于MFE在其上操作的主机机器。因此,通用物理控制平面条目包括抽象物理端口,而定制物理控制平面条目包括附连到特定MFE的实际的物理接口(其在许多情况下是虚拟接口)。

[0069] 在一些实施例中,如所示出的,网关主机还操作受管理转发元件(例如,利用与VM主机325相同的分组处理/虚拟交换软件)。这些MFE还从物理控制器接收使MFE能够实现逻辑转发元件的物理控制平面数据。此外,一些实施例通过分层网络控制系统将路由表数据和路由协议(例如,BGP)配置信息分发到在网关主机中操作的L3网关。管理逻辑网络的逻辑控制器310选择用于逻辑路由器的这组网关主机(例如,利用跨一组主机散布用于各种逻辑路由器的L3网关的负载平衡算法),然后生成数据分配给这些主机。

[0070] 逻辑控制器识别管理这些选定的网关主机中每一个的(一个或多个)物理控制器,并且将路由表和/或路由协议配置数据分发给识别出的物理控制器。在一些实施例中,L3网关配置(例如,路由表、NAT表等)和BGP配置两者都作为一组数据元组分发。例如,一些实施例的BGP配置数据元组指定网关的BGP邻居的IP地址,以及要通告给那些邻居的一组IP地址或前缀。物理控制器然后将这些数据元组分发到网关主机。在一些实施例中,用于特定逻辑路由器的每个网关主机接收相同的路由表和BGP配置。另一方面,在一些实施例中,不同的网关主机可以具有到不同外部物理路由器的连接,并且因此具有不同组的BGP邻居。如在下面详细描述的,网关主机将数据元组转换为(i)用于由在网关主机上作为L3网关操作的容器(例如,VM、命名空间)使用的路由表,和(ii)用于由在容器内操作的BGP模块(例如,守护进程或其它应用)使用的BGP配置文件。

[0071] 以上描述了一些实施例的分层网络控制系统,但是其它实施例的网络控制系统只包括单个控制器(或具有一个活动控制器和一个或多个备用控制器的控制器集群)。图4概念性地示出了数据通过一些实施例的分层网络控制系统的传播。该图的左侧示出了到受管理转发元件的用以实现逻辑网络的逻辑转发元件(例如,逻辑交换机和逻辑路由器)的数据流,而该图的右侧示出了BGP数据向网关主机的传播,以便指配在L3网关内操作的BGP守护进程。

[0072] 在左侧,输入变换控制器305通过API接收被转换为逻辑控制平面数据的网络配

置。该网络配置数据包括诸如在图1中示出的逻辑拓扑结构。在一些实施例中，网络配置指定逻辑交换机到逻辑路由器的附连，其中MAC地址被分配给每个VM和连接到逻辑交换机的每个逻辑路由器端口，以及具有相关联的IP子网的每个逻辑交换机。

[0073] 如所示出的，逻辑控制平面数据由逻辑控制器310(具体而言，由逻辑控制器的控制应用)转换为逻辑转发平面数据，并且然后(由逻辑控制器的虚拟化应用)随后转换为通用物理控制平面数据。在一些实施例中，这些转换在逻辑转发平面处生成流条目(或定义流条目的数据元组)，然后在通用物理控制平面处在逻辑数据路径集(例如，逻辑交换机或路由器)上添加匹配。通用物理控制平面还包括附加的流条目(或数据元组)，用于将一般物理入口端口(即，不特定于任何特定MFE的端口的一般抽象)映射到逻辑入口端口，以及用于将逻辑出口端口映射到一般物理外出端口。例如，对于VM驻留于其处的逻辑交换机的端口，在通用物理控制平面处的流条目将包括当分组的目的地MAC地址与VM的MAC地址匹配时，将分组发送到VM连接到其的逻辑端口的转发决定，以及将逻辑出口端口映射到一般物理(即，虚拟)接口的出口上下文映射条目。对于其它MFE，包括在网关主机处的那些，通用物理控制平面数据包括用于封装在到VM所位于的MFE的隧道中的分组的一般隧道条目。

[0074] 如所示出的，物理控制器315(分层网络控制系统300中的若干个物理控制器之一)将通用物理控制平面数据变换为用于它在主机325、330、345和350处管理的特定MFE的定制物理控制平面数据。这种转换涉及在通用物理控制平面数据中替换用于一般抽象的特定数据(例如，特定的物理端口或隧道封装信息)。例如，在以上段落的例子中，端口集成条目被配置为指定VM附连到其的物理层端口(即，用于实际虚拟接口的标识符)。类似地，用于不同MFE的隧道封装条目将具有不同的隧道封装信息。

[0075] 虽然这个例子将物理控制器315示出为执行通用物理控制平面到定制物理控制平面的变换，但是一些实施例利用主机机器上的机架控制器用于该任务。在这种实施例中，物理控制器不变换物理控制平面数据，而只是用作分发机构，用于将该数据交付给位于网络中的许多主机机器，使得逻辑控制器不必与网络中的每个MFE通信。在这种情况下(未在图中示出)，通用物理控制平面到定制物理控制平面的转换由在主机325和345处的一个模块或元件(即，机架控制器)执行，而在主机325和345处的MFE执行定制物理控制平面到物理转发平面数据的转换。

[0076] 无论物理控制平面数据的定制是由物理控制器还是由主机处的机架控制器执行，在主机325处的MFE(由物理控制器315管理的若干个MFE之一)执行定制物理控制平面数据到物理转发平面数据的变换。在一些实施例中，物理转发平面数据是存储在MFE内(例如，在诸如Open vSwitch的软件虚拟交换机的用户空间和/或内核内)的流条目，针对其MFE实际匹配接收到的分组。此外，在两个网关主机345和350处的MFE执行这种变换，以便在(i)L3网关，(ii)经由隧道在受管理网络内的其它网络实体(例如，VM)，以及(iii)外部网络之间转发分组。

[0077] 图4的右侧示出了传播到网关主机(例如，主机345)，以实现用于L3网关而不是用于MFE的BGP配置的数据。如所示出的，逻辑控制器310将BGP配置转换为定义那个配置的一组数据元组。在一些实施例中，基于由用户(例如，管理员)输入的网络配置，BGP配置由逻辑控制器或输入变换控制器生成。当用户设计逻辑网络时，一些实施例允许用户为逻辑路由器指定到外部网络的连接是否将使用路由协议(或具体地BGP)来与外部路由器对等。在一

些实施例中，用户通过为这些连接选择BGP(或不同的路由协议)为其被自动激活的某种类型的端口(例如，上行链路端口)进行这种指定。此外，在逻辑网络中的每个逻辑交换机将具有相关联的IP子网(或者由用户分配或者由逻辑控制器自动分配)。对于每个端口，或者对于作为整体的逻辑路由器，用户可以指定将把分组发送到端口的外部物理路由器，或者逻辑控制器基于为端口选择的网关生成该数据。

[0078] 基于该信息(即，每个端口连接到其的这组物理路由器、VM/逻辑交换机的IP地址/子网)，逻辑控制器310生成用于BGP配置的这组数据元组。在一些实施例中，这可以由还将逻辑控制平面数据转换为物理控制平面数据的表映射引擎来执行。除了BGP数据元组之外，逻辑控制器生成用于L3网关的逻辑路由器方面的数据元组(例如，路由表)。为了在特定的网关主机上定义容器，一些实施例将每个容器定义为单独的数据元组，其指定容器的存在和在容器上运行的进程，包括BGP。在这个数据元组内，可以启用BGP。此外，该数据元组定义各种BGP选项，诸如路由器ID、是否通告恰当的(graceful)重启能力、以及向所有对等体通告的前缀列表(例如，以无类域间路由(CIDR)的形式)。此外，逻辑控制器创建用于特定L3网关的每个BGP邻居(即，对等的外部路由器)的数据元组。在一些实施例中，这些邻居数据元组指定BGP邻居的地址、指示保持活动(keep-alive)分组之间的时间的保持活动定时器、和网关中的BGP应用通过其与邻居通信的接口、以及其它信息。

[0079] 一旦逻辑控制器310识别用于逻辑路由器的网关主机并且创建数据元组，逻辑控制器然后就识别物理控制器或管理网关主机的控制器。如所提到的，与VM主机325-340一样，每个网关主机都具有分配的主管物理控制器。在图3的例子中，两个网关主机都由物理控制器315管理，因此其它物理控制器320不接收BGP数据元组。

[0080] 为了向网关主机提供逻辑路由器配置数据，一些实施例的逻辑控制器310将数据推送到物理控制器315。在其它实施例中，物理控制器从逻辑控制器请求配置数据(例如，响应于配置数据可用的信号)。

[0081] 物理控制器315将数据传递到网关主机，包括主机345，就像它们传递物理控制平面数据一样。在一些实施例中，BGP数据元组被发送到作为与MFE相关联的软件的部分在主机上运行的数据库，并且被用来配置MFE的某些方面(例如，它的端口信息和其它非流条目配置数据)。

[0082] 在一些实施例中，网关主机345上的进程启动用于L3网关的容器并且将存储在数据库中的BGP数据元组变换为用于在L3网关中操作的应用的BGP配置文件。应用可以加载配置，以便确定其操作配置。

[0083] 以上描述描述了由网络控制系统进行的网络配置到物理控制器传递到主机(例如，经由诸如OpenFlow的协议)的一组物理转发平面流条目的转换。但是，在其它实施例中，用于定义流条目的数据以其它形式被传递，诸如更抽象的数据元组，并且MFE或在具有MFE的主机上运行的过程将这些数据元组转换为用于在处理数据流量中使用的流条目。

[0084] 图5概念性地示出了用于生成和分发数据以便在受管理网络中的逻辑路由器和外部网络之间实现一组连接的一些实施例的过程500。在一些实施例中，过程500由网络控制器(例如，由诸如在图3中示出的网络控制器层次结构中的逻辑控制器)在接收到包括具有已启用的路由器对等的逻辑路由器的网络配置时执行。

[0085] 如所示出的，过程500开始于(在505处)接收创建具有连接到外部网络的一个或多

个端口的逻辑路由器的指令。这些指令可以是网络管理员(例如,通过经控制器API传递逻辑网络配置的云管理应用)设计包括逻辑路由器的逻辑网络的结果。在一些实施例中,创建逻辑路由器的指令具体地指示到外部网络的连接应当利用BGP或另一协议来实现,用于进行路由器对等和路由通告。在其它实施例中,这种能力对具有至少一个到外部网络的连接的所有逻辑路由器被自动启用。

[0086] 接下来,过程(在510处)选择用于连接到逻辑网络的每个端口的网关主机机器。一些实施例将每个端口分配给不同的网关主机,而其它实施例允许多个端口(以及因此托管路由表和BGP服务的多个命名空间)在单个网关主机上被创建。在一些实施例中,网关主机根据集群或故障域进行布置。在一些实施例中,这些集群可以是在受管理网络中物理上位于一起的主机机器的集合,并且因此更可能全部一起失败(例如,由于架顶式交换机的顶部掉落、电源问题等)。不同实施例可以相对于集群不同地将网关分配给主机机器。例如,一些实施例只为特定的逻辑路由器每集群分配一个网关,而其它实施例将用于逻辑路由器的全部网关分配给同一集群。还有的其它实施例可以将网关分配给若干个不同的集群,但是允许两个或更多个网关在单个集群内。

[0087] 此外,在一些实施例中,网关主机机器可以基于那些网关主机被用于的功能被分配给不同的组。例如,在物理受管理网络中,一些实施例使用第一组网关主机用于提供逻辑服务(例如,DHCP、元数据代理)、第二组网关主机用于利用BGP用于路由通告和对于其每个逻辑路由器端口被分配单个网关的L3网关、以及第三组网关主机用于不利用路由通告和对于其每个逻辑路由器端口被分配给多个网关的L3网关。在图5的这种情况下,控制器从第二组中为每个逻辑路由器端口选择网关主机。每个组可以跨越网关主机的若干个集群,从而允许过程500(在510处)从若干个集群(即,故障域)选择第二组内的网关主机机器。

[0088] 一些实施例允许管理员指定控制器向其分配逻辑路由器的每个逻辑端口的集群,并且控制器处理那个集群内实际网关主机的选择。因此,管理员可以指定将两个逻辑端口分配给第一集群中的网关、四个给第二集群中的网关、以及再多两个给第三集群中的网关。然后,控制器将每个逻辑端口分配给它选定的集群中的特定网关主机。对于这种分配,一些实施例使用负载平衡技术,诸如计算逻辑路由器或端口的属性(例如,由控制器分配的UUID)的散列函数对集群中网关主机的数量取模。这有效地随机(尽管算法本身是确定性的)将逻辑路由器端口分配给集群内的网关主机,并且因此长远地跨网关主机使L3网关负载平衡。

[0089] 一些其它实施例可以使用其它技术来跨集群中的主机使L3网关负载平衡。例如,不是利用散列算法在集群中所有网关主机之间进行选择,而是一些实施例仅仅在具有最少量的当前在操作的逻辑路由器的那些网关之间进行选择,并且用网关的这个较小数量对散列函数的结果取模。其它实施例分析每个网关上的逻辑路由器的数量和网关的操作负载(例如,基于经过特定时间帧被处理的分组的数量),以便确定特定的逻辑路由器应当被分配给哪个网关主机。

[0090] 图6概念性地示出了网关主机的五个单独的集群605-625,以及在那些网关主机上实现的逻辑路由器端口(被称为上行链路)。具体而言,第一集群605包括四个网关606-609,第二集群610包括四个网关611-614,第三集群615包括三个网关616-618,第四集群620包括三个网关621-623,并且第五集群625包括五个网关626-630。这个图仅仅示出了用于能够托

管用于使用BGP进行路由通告的端口的L3网关的每个集群的网关主机。在一些实施例中，集群(即，故障域)可以包括分配给不同功能，诸如托管逻辑服务(例如，DHCP、DHCP中继等)，的附加网关主机。

[0091] 在这个例子中，六个不同逻辑路由器(LR1-LR6)中的每一个具有三至五个逻辑端口用于与外部网络连接，这些逻辑端口贯穿这些集群不同地散布。例如，逻辑路由器LR1具有在位于四个不同集群中的主机606、611、613和627中的网关上实现的上行链路端口。逻辑路由器LR3具有在主机608、609、621和623中的网关上实现的上行链路端口，其中在两个不同集群中的每一个当中有两个网关。逻辑路由器LR4的全部三个上行链路端口都在主机627、628、629上在同一集群625中的网关上实现。因此，取决于管理员决定和逻辑网络的需求，用于实现逻辑路由器的上行链路端口的不同配置是可能的。

[0092] 在示出的例子中没有使用明确的负载平衡，其中网关利用例如管理员的明确分配或者散列函数对网关的数量取模而被分配给主机，因此第二集群610和第五集群625两者都分别包括网关主机613和630，其中没有实现网关(尽管这两个集群还包括具有多个网关的主机)。此外，并非所有集群都具有相同数量的网关主机。这可能由于集群仅仅具有不同数量的物理机器，其中一些集群具有被分配给不同任务的不同数量的网关主机、或者其中一些集群具有由于连接或其它问题造成的主机机器离线。在一些实施例中，当在网关上运行的应用识别出网关主机有问题时，该应用通知网络控制器(例如，管理网关主机的物理控制器)该问题，使得网关主机可以被停止使用，直到问题得到解决。

[0093] 不同的实施例不同地处理实现上行链路端口的L3网关的故障切换。在一些实施例中，当具有实现上行链路端口的L3网关的网关主机出现故障时，网络不立即在新主机上取代该L3网关。而是，网络控制器允许传入(并且，在一些实施例中，传出)分组跨其网关仍然处于活动状态的其它上行链路端口分发。另一方面，一些实施例利用用于不同网关主机上(例如，在与发生故障的主机相同的集群中)的上行链路端口的新实现取代该L3网关。

[0094] 图6和以上描述示出了利用单个网关主机来实现每个上行链路端口。但是，为了预防主机故障，一些实施例为连接到外部网络的每个上行链路端口选择活动和备用网关主机两者。即，用于第一网关主机中的逻辑路由器端口的实现之一是活动的，其中MFE被指示向其发送分组，并且其BGP守护进程通告路由。在第二网关中，逻辑路由器端口的其它实现以相同的方式被配置，但是MFE不向其发送流量并且其BGP守护进程不通告路由。如果第一(活动)网关主机发生故障，则MFE将开始向第二(备用)主机发送流量，并且网络控制器将让第二主机知道激活其BGP守护进程。用于高可用性网关的故障切换在标题为“High Availability L3 Gateways for Logical Networks”并且于2014年1月28日提交的美国申请14/166,446中进一步详细描述。美国申请14/166,446通过引用被结合于此。

[0095] 返回到图5，在为连接到外部网络的每个逻辑路由器端口选择网关主机之后，过程500(在515处)为通过逻辑路由器(即，托管附连到逻辑交换机的VM的机器，该逻辑交换机附连到逻辑路由器)发送分组的主机机器上的MFE生成流条目(或定义流条目的数据元组)。除其它功能外，这些流条目还既(i)实现分布式逻辑路由器又(ii)将由逻辑路由器转发到连接到外部网络的逻辑端口之一的分组通过隧道发送到选定的网关主机机器。此外，为VM主机生成的流条目还将包括许多其它条目，诸如为逻辑交换机、入口和出口ACL等实现逻辑转发的那些条目。

[0096] 为了在MFE中实现逻辑路由器，一些实施例生成逻辑转发条目，该逻辑转发条目基于目的地IP地址或地址前缀(以及在用于路由器自己的逻辑管道上)匹配分组，并且基于IP地址识别逻辑路由器的逻辑出口端口。为了生成这些流条目，一些实施例的网络控制器首先生成用于逻辑路由器的路由表。在一些实施例中，这些路由包括用于将分组发送到连接到外部网络的端口之一的缺省路由，以及用于基于与逻辑交换机相关联的IP子网将分组发送到每个附连的逻辑交换机的路由。然后，路由表被嵌入到流条目数据元组中(例如，包括在逻辑路由器管道上的匹配)。此外，用于逻辑路由器的流条目对分组执行MAC地址修改(以将源MAC地址修改为逻辑路由器的逻辑出口端口的地址，并将目的MAC地址修改为匹配目的地IP地址的地址)。这可以包括用于或者执行ARP或者将分组发送到也在主机上操作的ARP守护进程的流条目。在分组的实际路由之外，在一些实施例中，网络控制器根据为逻辑路由器定义的任何策略生成用于L3入口和出口ACL的流条目。

[0097] 不同的实施例使用不同类型的流条目用于选择分组应当被转发到哪些连接到外部网络的逻辑端口。一些实施例通过单个逻辑端口发送所有传出分组，但是，当传出流量的量大时(例如，对于web服务器、流视频应用等)，这个端口在其上实现的网关会变成瓶颈。其它实施例使用类似相同成本的多路径(ECMP)的技术来选择逻辑出口端口用于退出逻辑网络的分组。例如，一些实施例的流条目将端口列为束列，并且然后提供对于给定的分组属性集合(例如，分组属性的散列对端口数量取模)识别向哪些端口发送分组的技术。

[0098] 过程500还(在520处)为网关主机机器上的MFE生成流条目。除其它功能外，这些流条目还将分组转发到在网关主机上实现逻辑路由器端口的容器(例如，命名空间)、将分组转发到连接到外部路由器的NIC，以及将分组通过隧道转发到在VM主机的其它MFE。例如，在一些实施例中，出站分组(即，从VM主机接收到的)由MFE发送到命名空间，用于通过L3网关路由表进行附加路由。在这个路由之后，命名空间将分组返回到MFE(作为新分组)，其中外部路由器被识别为其目的地并且MFE将这个新分组发送到出站NIC。对于传入分组，MFE首先将分组发送到命名空间，用于通过L3网关路由表进行路由，然后接收返回的分组并执行第一跳路由，以识别逻辑路由器的逻辑出口端口(通常逻辑交换机之一附连到其的端口)、识别逻辑交换机的逻辑出口端口(通常VM附连到其的端口)、并且将分组从隧道发送出到适当的MFE。

[0099] 除了流条目(或定义流条目的数据元组)之外，该过程还(在525处)生成用于路由表的数据元组以供处理在实现逻辑端口的每个L3网关处进入(并且，在一些实施例中，外出)的分组。在一些实施例中，这些数据元组由还生成流条目的表映射引擎生成。但是，其它实施例利用单独的路由处理器生成路由条目。如以上所指出的，大多数路由表被实现为发送到MFE的流条目。但是，L3网关的路由表处理被路由到外部网络和从外部网络接收到的分组。因此，路由表(以及可以由L3网关实现的IP网络堆栈的附加方面，诸如NAT表)负责执行任何必要的ARP(例如，到外部网络中)、递减分组TTL(即，作为用于分组的另一跳)，以及负责传出分组选择该分组将被发送到其的外部路由器。

[0100] 在操作515-525处生成的这些流条目和/或数据元组使得L3网关和MFE能够处理数据分组的处理。此外，过程500为在每个L3网关处操作的路由协议应用(例如，BGP守护进程)生成数据。因此，过程(在530处)识别连接到外部网络的每个逻辑端口(即，每个L3网关)与其对等的(一个或多个)外部网络路由器的地址(和其它信息)。在一些实施例中，管理员为

每个逻辑端口输入这种数据，并且处理确保外部路由器被正确地连接到网关主机(或者，例如，网关主机连接到其的架顶式交换机的顶部)。在其它实施例中，网络控制器基于其存储的网络状态信息自动确定每个网关主机连接到其的外部路由器的集合，并且使用这些作为L3网关与其对等的外部网络路由器。

[0101] 基于这些识别出的外部路由器，以及为逻辑网络计算出的路由，该过程(在535处)生成数据元组来定义用于选定的主机机器上的L3网关的路由协议。如以上所指示的，为了定义在特定的网关主机上的L3网关容器，一些实施例将每个容器定义为单独数据元组，其指定容器的存在和在容器上运行的进程，包括BGP或另一个路由协议应用。该数据元组定义各种BGP选项，诸如路由器ID、是否通告恰当的重启能力，以及向所有对等体通告的前缀列表。在一些实施例中，IP地址和/或前缀的这个列表是基于逻辑网络的用户配置(例如，为逻辑网络的逻辑交换机配置的公共IP)。此外，控制器为每个L3网关的每个对等体外部路由器(例如，BGP邻居)生成数据元组。在一些实施例中，这些邻居数据元组指定外部路由器的地址、指示保持活动分组之间的时间的保持活动定时器、和网关中的BGP应用通过其与邻居通信的接口，以及其它信息。

[0102] 利用生成的数据，过程500(在540处)将生成的数据元组和/或流条目分发到各种主机机器。在一些实施例中，两种类型的数据(流条目和路由表/路由协议数据元组)经由不同的协议被分发。一些实施例经由诸如OpenFlow的第一协议将流条目分发到VM主机和网关主机二者，而经由诸如OVSDB的第二协议将路由协议(例如，BGP)信息和路由表分发到网关主机。在一些实施例中使用的OVSDB协议还携带用于MFE的配置信息(用于VM主机和网关主机两者)。

[0103] 以上图5概念性地将过程500示为由控制器执行的单个线性流。但是，本领域普通技术人员将认识到，控制器计算各个流条目和/或数据元组的次序不必按图中示出的次序。例如，控制器可以在生成用于MFE的流条目之前生成用于网关主机的流条目，等等。此外，一些实施例不一直等到所有指示的数据都被计算出来才分发数据，而是可能递增地分发数据。例如，一些实施例分开分发转发数据与路由协议数据，或者一旦用于那个主机的所有数据都已经生成就递增地将转发数据分发到特定主机。

[0104] II.网关主机体系架构

[0105] 以上部分详细描述了由控制器进行的逻辑路由器和路由协议数据的生成以及那种数据到作为L3网关操作的容器(例如，命名空间)驻留在其上的网关主机机器的分发。在一些实施例中，网关主机机器包括负责基于由网络控制系统分发的数据元组创建容器、在容器中建立路由表以及处理来往于命名空间的分组的各种模块(例如，作为用户空间守护进程或内核模块运行)。

[0106] 图7概念性地示出了用于网关主机机器700的一些实施例的软件体系架构。主机机器700是指定用于在命名空间中托管L3网关实现的主机，该L3网关实现可以操作路由协议应用。如所示出的，主机700包括虚拟化软件705以及两个命名空间710和715。在一些实施例中，主机包括命名空间710和715作为容器在其上运行的基础Linux操作系统。在一些实施例中，网关主机机器700是具有基于标准x86的处理器的计算机。

[0107] 虚拟化软件705包括转发元件守护进程725、数据库守护进程730、命名空间守护进程735、高可用性守护进程720和转发元件内核模块740。在一些实施例中，转发元件守护进

程725、数据库守护进程730、命名空间守护进程735和高可用性守护进程720在虚拟化软件705的用户空间中操作，而转发元件内核模块740在虚拟化软件705的内核中操作。在一些实施例中，在主机上使用的转发元件是Open vSwitch(OVS)，并且，除了命名空间守护进程和高可用性守护进程，这些模块还是OVS守护进程、OVSDB守护进程和OVS内核模块。在一些实施例中，命名空间守护进程735和高可用性守护进程720的功能被组合为单个用户空间应用。这个图既示出了用于指配受管理转发元件和命名空间的控制路径连接(被示为虚线)，又示出了用于发送数据分组(包括BGP分组)的数据路径连接(被示为实线)。本领域普通技术人员将认识到，除了示出的涉及虚拟交换机和托管的命名空间的模块之外，一些实施例的虚拟化软件还包括用于执行例如主机机器700的硬件资源(例如，处理器、存储器等)的虚拟化的附加模块。

[0108] 转发元件守护进程725在一些实施例中是与物理网络控制器795通信的应用，以便接收用于处理和转发发送到和来自命名空间710和715的分组(例如，从外部网络进入受管理网络或者离开受管理网络到外部网络的分组)的指令。具体而言，如在上一部分中所描述的，转发元件守护进程725从物理控制器795接收物理控制平面流条目。在一些实施例中，转发元件守护进程通过OpenFlow协议与网络控制器通信，但是其它实施例可以使用不同的通信协议用于将转发数据传送到主机机器。此外，在一些实施例中，转发元件守护进程725在物理控制器795将配置信息发送到数据库守护进程之后从数据库守护进程730检索配置信息。

[0109] 一些实施例的转发元件守护进程725包括流协议模块750和流处理器755。流协议模块750处理与网络控制器795的通信，以便接收用于受管理转发元件的物理控制平面信息(例如，流条目)。如所提到的，在一些实施例中，这种通信使用OpenFlow协议。当流协议模块750接收到这种物理控制平面信息时，它将接收到的信息变换成流处理器755可理解的数据(例如，可用于处理分组的物理转发平面信息)。

[0110] 在一些实施例中，流处理器755管理用于处理和转发(即，交换、路由)分组的规则。例如，流处理器755存储从流协议模块750接收到的规则(例如，在诸如盘驱动器的机器可读存储介质中)。在一些实施例中，这些规则被存储为一组流表(转发表)，其中每个流表包括一组流条目。在一些实施例中，这些流条目包括匹配(即，一组分组特点)和一个或多个动作(即，对匹配这组特点的分组采取的一组动作)。在一些实施例中，流处理器725处理受管理桥760(下面描述)对其不具有匹配规则的分组。在这种情况下，流处理器755对照其存储的规则匹配分组。当分组匹配规则时，流处理器725将匹配的规则和分组发送到受管理桥760，让受管理桥处理。以这种方式，当受管理桥760随后接收到匹配所生成规则的类似分组时，该分组将对照受管理桥中生成的确切匹配规则来匹配并且流处理器755将不必处理该分组。

[0111] 在一些实施例中，数据库守护进程730是还与物理控制器795通信以便配置受管理转发元件(例如，转发元件守护进程725和/或转发元件内核模块740)的应用。例如，数据库守护进程730从物理控制器接收配置信息并且将配置信息存储在一组数据库表745中。这种配置信息可以包括用于创建到其它受管理转发元件的隧道的隧道信息、端口信息等。在一些实施例中，数据库守护进程730通过数据库通信协议(例如，OVSDB)与网络控制器795通信。在一些情况下，数据库守护进程730可以从转发元件守护进程725接收对配置信息的请

求。在这些情况下,数据库守护进程730检索请求的配置信息(例如,从其数据库表745集合中)并且将配置信息发送到转发元件守护进程725。

[0112] 除了转发元件配置(隧道和端口信息等)之外,一些实施例的数据库守护进程730还附加地接收为在命名空间710和715中操作的BGP守护进程定义配置的BGP配置信息。这种信息包括关于BGP守护进程向其对等体通告的路由的信息,以及识别那些对等体的信息。数据库守护进程730可以连同转发元件配置信息一起,或者在与控制器795的单独事务中,接收这种BGP配置信息。

[0113] 如所示出的,数据库守护进程730包括配置检索器765和一组数据库表745(其可以存储在,例如,主机700的硬盘驱动器、易失性存储器或其它储存器上)。配置检索器765负责与物理控制器795通信。在一些实施例中,配置检索器从控制器接收用于受管理转发元件的配置信息。此外,在一些实施例中,配置检索器接收用于配置命名空间710和715的数据元组,以及任何路由表、NAT表、BGP守护进程或者由命名空间提供的其它服务。在一些实施例中,配置检索器765还将这些数据元组转换为数据库表记录,以在数据库表745中存储。

[0114] 具体而言,一些实施例的数据库表745包括容器表,其中数据库中的每条记录定义主机机器上的不同命名空间(或其它容器)。因此,对于主机700,容器表将包括用于两个命名空间710和715中每一个的行。此外,对于每个命名空间,数据库表存储定义路由表的信息(例如,缺省路由、为连接的逻辑交换机定义的任何附加路由,以及任何用户定义的静态路由)。如果路由器执行NAT,则数据库还存储用于逻辑路由器的NAT规则(源NAT和/或目的地NAT)。此外,对于每个命名空间,数据库存储逻辑路由器端口列表,其中具有每个端口的IP地址、MAC地址、网络掩码等。

[0115] 对于命名空间710和715,利用活动BGP守护进程,数据库表记录指示BGP被启用。此外,在一些实施例中,这些记录包含将L3网关的BGP属性指定为对等路由器的附加列。这些属性可以包括本地自主系统编号(在不同的实施例中,其将L3网关所属的逻辑网络或受管理网络识别为一个整体)、路由器标识符(例如,IP地址)、是否通告恰当的重启(用于故障切换目的-在一些实施例中,是仅实现端口的L3网关的命名空间不通告恰当的重启)、以及由BGP守护进程通告的一组地址/前缀。

[0116] 此外,一些实施例为L3网关经由BGP守护进程与其对等的每个外部物理路由器(即,每个BGP邻居)定义数据库表记录(例如,在不同的数据库表中)。在一些实施例中,这些记录指定邻居路由器的一些或全部IP地址、用于路由器的自主系统编号、保持活动定时器(即,发送到邻居以便使BGP会话保持活动的保持活动消息之间的持续时间)、可选的用于MD5认证的密码、抑制(hold-down)定时器持续时间(即,在其之后如果没有接收到保持活动消息则BGP守护进程假定邻居已出故障的持续时间)、以及与BGP邻居的通信通过其被发送的接口。

[0117] 转发元件内核模块740处理并在主机700上运行的命名空间、在主机700外部的网络主机以及在受管理网络中其它主机上操作的转发元件之间转发网络数据(例如,分组)(例如,对于通过(一个或多个)NIC 770或者从命名空间710和715接收到的网络数据分组)。在一些实施例中,转发元件内核模块740实现用于一个或多个逻辑网络(具体而言,命名空间710和715所属于的逻辑网络)的物理控制平面的转发表。为了便于网络数据的处理,转发元件内核模块740与转发元件守护进程725通信(例如,以从流处理器755接收流条目)。

[0118] 图7示出转发元件内核模块740包括受管理桥760。此外，在一些实施例中，虚拟交换机内核模块可以包括附加的桥，诸如物理接口(PIF)桥。一些实施例包括用于主机机器的硬件中每个NIC 770的PIF桥。在这种情况下，在一些实施例中，PIF桥位于受管理桥760和每个NIC 770之间。

[0119] 一些实施例的受管理桥760在命名空间710和715与VM和向命名空间发送流量并从其接收流量的其它主机(包括外部主机)之间执行分组的实际处理和转发。分组，例如，通过隧道端口从在VM主机的MFE，或者经由它们到NIC的连接从外部路由器，在受管理桥760被接收，使得经不同隧道或外部路由器连接到达的分组在桥760的不同接口被接收。对于从其它MFE(例如，在VM主机)接收的分组，受管理桥760基于附加到分组的目的地逻辑端口(或其它信息，诸如目的地MAC或IP地址)通过其与命名空间的(一个或多个)接口将分组发送到适当的命名空间。

[0120] 对于从外部路由器接收的分组，一些实施例的受管理桥760基于例如分组的目的地MAC和/或IP地址将分组发送到适当的命名空间。当外部路由器将分组路由到命名空间时，路由器利用先前发现的ARP信息执行MAC地址替换。在一些实施例中，外部路由器具有命名空间的MAC地址，所述MAC地址与那个命名空间背后的各种IP地址关联，并且因此使用命名空间MAC地址作为用于指向那个网关的分组的目的地地址。在一些实施例中，当进入逻辑网络的分组还不具有附加的逻辑上下文信息时，受管理桥使用这个信息将这些分组指向适当的命名空间。

[0121] 类似地，受管理桥从命名空间710和715接收分组并且基于分组通过其被接收的接口和分组的源和/或目的地地址处理和转发这些分组。在一些实施例中，为了处理分组，受管理桥760存储在流处理器755中存储的规则(和/或从存储在流处理器755中的规则导出的规则)的当前或最近用于处理分组的子集。在这个图中，受管理桥760包括到命名空间710和715当中每一个的两个接口。在一些实施例中，受管理桥包括用于逻辑路由器的每个逻辑端口的单独接口。因此，受管理桥可以通过其一个接口将分组发送到命名空间，并且在通过命名空间路由表路由之后，受管理桥通过不同的接口接收回分组。另一方面，因为命名空间只实现逻辑路由器端口之一，所以一些实施例在命名空间和受管理桥之间只具有单个接口。

[0122] 虽然图7示出了一个受管理桥，但是转发元件内核模块740可以包括多个受管理桥。例如，在一些实施例中，转发元件内核模块740包括用于在主机机器700中实现的每个逻辑网络或者用于驻留在主机中的每个命名空间(其常常与每个逻辑网络相同)的单独的桥。照此，在这个例子中，转发元件内核模块740将包括两个受管理桥，其具有到命名空间710的单独的接口。

[0123] 命名空间710和715当中每一个实现不同的L3网关(即，实现逻辑路由器的不同端口)。在一些实施例中，特定网关主机机器上的所有命名空间都是同一类型(即，利用诸如BGP的路由器对等协议实现单个逻辑路由器端口)。另一方面，一些实施例还允许是几个当中的一个的命名空间为逻辑路由器等效地实现整个路由表或者为具有到外部网络的单个逻辑端口附连的逻辑路由器充当网关。此外，一些实施例还允许命名空间提供除路由之外的其它逻辑服务，诸如DHCP、DHCP中继、元数据代理，等等。

[0124] 如这个图中所指示的，在一些实施例中，为不同的逻辑网络(或者，在一些情况下，为相同的逻辑网络中的同一逻辑路由器或不同逻辑路由器)实现不同L3网关(例如，不同逻

辑端口)的不同命名空间可以驻留在同一主机700上。在这种情况下,命名空间710和715都运行BGP守护进程和路由表。

[0125] 在一些实施例中,命名空间可以提供多个服务。在这种情况下,第一命名空间710包括路由表775、BGP守护进程780和其它服务782。在命名空间710上运行的这些其它服务可能提供ARP功能、网络地址变换(NAT)表,或与路由器相关联的其它特征。第二命名空间715也包括路由表790和BGP守护进程792,连同其它服务794。一些实施例对实现端口并使用路由器对等协议的所有L3网关使用相同的服务集合,而其它实施例允许用户配置所提供的网络堆栈或其它服务。此外,对于其中多个网关同时对逻辑路由器处于活动状态的实现,一些实施例限制使用有状态的服务,诸如NAT。即,网络控制系统防止L3网关利用需要各种网关用于逻辑路由器以共享状态信息的那些服务。

[0126] 一些实施例的命名空间守护进程735管理驻留在主机700上的命名空间710和715以及在那些命名空间中运行的服务(例如,逻辑路由器和L3网关服务)。如所示出的,命名空间守护进程735包括数据库监视器785和BGP配置生成器799。此外,一些实施例包括配置生成器或用于其它服务的类似模块(例如,NAT表生成器、路由表生成器、用于DHCP和可在命名空间中提供的其它服务的配置生成器,等等)。

[0127] 数据库监视器785监听数据库表745以获得对影响实现逻辑路由器的命名空间的特定表的改变。这些改变可以包括新命名空间的建立、命名空间的去除、添加或除去BGP邻居、修改命名空间中的BGP配置或路由表、将新的逻辑交换机附连到逻辑路由器,等等。当数据库监视器785检测到影响命名空间的改变时,它或者使命名空间守护进程在主机上创建用于新逻辑路由器的新命名空间、在现有的命名空间中实例化新过程(例如,对于新启用的服务),或者生成/修改用于命名空间的路由表或其它配置数据。

[0128] 当数据库监视器785检测到新BGP配置数据(或者是具有BGP配置的新命名空间、对现有BGP配置的修改、对用于特定BGP守护进程的邻居集合的修改,等等)时,数据库监视器785将这个数据提供给BGP配置生成器799(或者指示BGP配置生成器799从数据库表745中来检索新数据)。BGP配置生成器使用存储在数据库表745中的数据元组来以守护进程所需的格式为BGP守护进程建立配置文件。在一些实施例中,命名空间守护进程785在主机文件系统783中存储所生成的配置。在一些实施例中,BGP守护进程780和792是可用于Linux或不同操作系统的标准应用。

[0129] 高可用性守护进程720监视网关主机700和/或在主机700上运行的命名空间710和715的健康状况。这个守护进程负责在网关主机700不再健康并且应当停止使用的时候向控制器795报告,由此允许控制器将在主机上操作的命名空间分配给新的网关主机、修改用于在将分组发送到网关主机700上实现的L3网关的VM主机处的隧道封装的流条目。

[0130] 在一些实施例中,高可用性守护进程720包括监视器793和健康状况修改器797。一些实施例的监视器793监视网关主机机器700的各个方面,以确定机器是否应当继续使用或者停止用于托管L3网关(以及用于逻辑网络的其它服务)。监视器793可以监视底层硬件资源(例如,处理器、存储器,等等),以确保这些资源运作良好,足以以必要的速度提供逻辑路由服务。此外,监视器793确保到其它主机机器(例如,向网关主机发送流量的VM主机)的连接是否正常运作。一些实施例通过监视物理NIC以及监视是否从这些主机接收到分组来监视连接。此外,一些实施例的监视器793监视在主机上操作的软件。例如,监视器检查虚拟化

软件705以及命名空间710和715的其它模块，以确保它们没有崩溃或以其它方式发生故障。此外，在一些实施例中，高可用性守护进程720使用双向转发检测(BFD)来直接监视上游路由器(例如，在受管理网络外部的路由器)。

[0131] 当监视器793出于任何理由确定网关主机700应当停止使用时，高可用性守护进程720通知物理控制器795管理网关主机机器700。为了通知控制器，在一些实施例中，健康状况修改器797利用数据库守护进程765(例如，经由配置检索器765)传播到控制器795的信息修改数据库表745。在一些实施例中，健康状况修改器797修改包括用于网关主机700的健康变量的表，以指示网关应当不活动。在一些实施例中，健康状况修改器797修改表745中为每个命名空间创建的行，以指示该命名空间应当被视为不活动。当单个命名空间崩溃时，健康状况修改器797只修改用于崩溃的命名空间的数据。

[0132] 在一些实施例中，配置检索器765检测到数据库表745已被修改并且将更新的数据元组发送到物理控制器795。当控制器795接收到这种指示时，控制器识别具有受影响的逻辑路由器的逻辑控制器，从而使这些控制器(i)在新的网关主机上分配逻辑端口用于实现，以及(ii)为向L3网关发送分组的MFE主机生成新的流条目。

[0133] III.路由协议应用的配置

[0134] 如上一部分中所指出的，在一些实施例中，在网关主机机器上操作的应用(例如，用户空间守护进程)或应用的集合负责接收L3网关配置并在命名空间或网关主机的其它容器中安装那个配置。L3网关配置可以包括路由表、路由协议配置，以及其它数据。除其它功能外，应用还从存储在主机上的数据库表的集合检索信息并使用那个信息来利用其各种功能在主机上的命名空间中设置L3网关。在一些实施例中，这种设置包括指定用于BGP守护进程的各种BGP参数和BGP邻居的配置文件的生成。

[0135] 图8概念性地示出了用于设置或修改网关主机机器上的L3网关的一些实施例的过程800。在一些实施例中，过程800由在网关主机上运行的虚拟化软件中的用户空间守护进程，诸如命名空间守护进程785，执行。如所示出的，过程800开始于(在805处)接收对定义利用BGP实现逻辑路由器端口的L3网关的数据库表的修改，以在主机机器上运行。在一些实施例中，负责在主机上创建L3网关并生成BGP配置文件的应用监听由控制器数据填充的数据库表的集合。当新的行被添加到定义主机上命名空间的表时，或者现有的行被修改时，应用检测这种改变并检索数据。当改变涉及或者利用BGP守护进程创建命名空间或者修改用于现有命名空间的BGP配置时，BGP配置生成器被调用，以便创建或修改用于新的/受影响的BGP守护进程的配置文件。

[0136] 在接收到数据库表后，过程800(在810处)确定用于受影响的L3网关的容器是否已经在主机机器上运行。即，该过程确定对数据库表的修改是用于添加新的网关还是修改现有的网关。在一些实施例中，网关主机上的数据库表首先接收简单地定义新容器的数据元组，并随后接收配置信息，在这种情况下，路由表和/或BGP配置数据将被视为对现有命名空间的修改。

[0137] 当容器尚未在主机机器上操作时，该过程(在815处)创建用于在主机机器上的新L3网关的容器。在一些实施例中，在网关主机机器的虚拟化软件中操作的用户空间应用(例如，命名空间守护进程)负责创建和除去用于主机上的L3网关的容器。如所提到的，在一些实施例中，这个容器是虚拟化容器，诸如在基础操作系统之上运行的命名空间或虚拟机。一

些实施例使用Linux命名空间，因为它使用比典型虚拟机更少的运营资源，并且对于由L3网关执行的操作(例如，包括路由、BGP守护进程的IP堆栈)是足够的。在一些实施例中，每个网关主机机器为众多不同逻辑网络中的众多不同逻辑路由器运行众多(例如，几十个)操作L3网关的命名空间。

[0138] 接下来，过程800(在820处)根据在数据库表中的配置确定是否尚未为受影响的L3网关定义路由表。例如，如果数据库表只定义了新的L3网关而没有提供关于用于该L3网关的命名空间的配置的任何信息，则命名空间守护进程将在主机上创建新的命名空间，但完全不配置该命名空间。此外，如果数据库表只包括对命名空间的其它方面的修改，诸如BGP配置，则命名空间守护进程将不修改特定L3网关的路由表。但是，在一些实施例中，命名空间守护进程确保由BGP守护进程通告的任何路由都也在L3网关的路由表中。照此，如果新的前缀被添加到要在BGP配置中通告的前缀列表，则命名空间守护进程把这些(如果还不存在的话)添加到路由表。

[0139] 但是，当当前安装在容器中的路由表不匹配数据库表中的路由表定义时(或者是因为还没有定义路由表或者是因为路由表定义已被修改)，过程(在825处)为L3网关生成或修改路由表，并且(在830处)在容器中安装该路由表。在一些实施例中，这其实是一个操作，因为命名空间守护进程直接修改命名空间中的IP堆栈。在其它实施例中，命名空间守护进程生成路由表或IP堆栈，然后作为单独的动作将其安装在容器中。

[0140] 接下来，过程800(在835处)确定BGP守护进程是否已在用于L3网关的容器中被起动。例如，如果容器先前未利用配置创建，或者如果容器刚刚在过程800期间创建(即，如果数据库表利用BGP配置定义了新的容器)，则守护进程将还不会在容器中被起动。另一方面，如果对数据库表的修改仅仅是对路由表或BGP配置的更新(例如，添加用于新的逻辑交换机的路由、添加或除去BGP邻居，等等)，则BGP守护进程将已经在用于L3网关的容器中操作。

[0141] 当BGP守护进程尚未被起动时，该过程(在840处)在容器中起动BGP守护进程。在一些实施例中，命名空间守护进程向实现L3网关的命名空间发送指令，以启动BGP守护进程。为了让命名空间真正运行BGP守护进程的实例，在一些实施例中，该软件已缺省地安装在命名空间上。在其它实施例中，或者命名空间检索守护进程(例如，从网关主机上的储存装置)或者命名空间守护进程检索守护进程并将其安装在命名空间上。

[0142] 随着BGP守护进程被起动，过程(在845处)确定BGP守护进程的配置是否匹配在接收到的用于L3网关的数据库表中定义的配置。如果BGP守护进程(在操作840处)刚起动，则守护进程将还没有配置，并且因此很清楚将不匹配在数据库表中定义的配置。此外，数据库表修改可以添加或除去路由，以通告、添加或除去BGP邻居，或者修改用于BGP邻居的数据。但是，如果数据库表修改只影响路由表，则将不需要BGP配置修改。

[0143] 当操作配置不匹配由数据库表定义的配置时，过程(在850处)从数据库表生成配置文件并且在主机机器的文件系统中存储该文件。在一些实施例中，在文件系统的特定目录中，在机器上操作的每个网关被分配用于例如BGP配置文件的子目录，以及用于其它数据(例如，DHCP配置文件，等等)的储存装置。为了生成配置文件，在一些实施例中，命名空间守护进程使用来自数据库表记录的数据元组并将它们变换成由BGP守护进程可读的特定格式。例如，在一些实施例中，配置文件是文本文件。在其它实施例中，命名空间守护进程首先生成中间配置文件(例如，文本文件)，然后将其转换成BGP守护进程可读的二进制快照，并

将这两个文件都存储在用于命名空间中BGP守护进程实例的目录中。在一些实施例中，配置文件定义(i)用于作为路由器的BGP守护进程的自主系统和标识信息，(ii)让BGP守护进程通告的路由的集合，和(iii)关于BGP守护进程的外部路由器对等体的信息。

[0144] 一旦配置文件已经生成，过程800就(在805处)通知BGP守护进程读取配置文件，以便让其配置匹配在数据库表中定义的配置。在一些实施例中，通知经由命名空间守护进程和BGP守护进程之间的网关主机内的TCP连接发生。在一些实施例中，BGP守护进程读取二进制配置文件、计算从其当前操作配置的改变，并应用这些改变。

[0145] IV.在网关中的BGP操作

[0146] 一旦BGP守护进程已被实例化，并且其配置文件已加载，L3网关就可以作为外部路由器的对等体参与路由交换。图9概念性地示出了由一些实施例的路由协议应用(例如，BGP守护进程)执行的一些实施例的过程900，以通告用于L3网关的到外部路由器的路由。过程900表示在初始启动时由BGP守护进程执行的过程。本领域普通技术人员将认识到，在许多情况下，操作将不是以这个图中所示的线性方式被执行。例如，与不同外部路由器的通信可以需要不同长度的设置时间，并且在一些实施例中BGP守护进程将每个对等连接视为单独的过程。

[0147] 如所示出的，过程900开始于(在905处)接收BGP配置。如在前一部分中所描述的，在一些实施例中，主机的虚拟化软件中的用户空间应用(例如，命名空间守护进程)生成BGP配置文件、在用于命名空间的主机文件系统的目录中存储该配置文件，然后向BGP守护进程通知该配置文件。在这个时候，BGP守护进程可以从目录检索BGP配置。

[0148] 接下来，过程900(在910处)安装配置。在一些实施例中，BGP守护进程读取检索出的二进制文件、确定其当前操作配置与在二进制文件中指定的配置之间的差别，并且将这些改变应用到现有的配置，使得新的操作配置匹配配置文件中的配置。如果这是对BGP守护进程的初始设置，则操作配置将没有数据。但是，如果改变局限于添加或除去要通告的路由，或者添加、除去或修改关于邻居物理路由器的信息，则BGP守护进程只修改其配置来实现改变，而不是重新加载整个配置。

[0149] 随着配置被安装，该过程(在915处)识别与其设置连接以便通告路由的BGP邻居(即，对等的物理路由器)。这可以是单个物理路由器(例如，如以下图15中所示)或者网关(利用BGP守护进程)向其通告相同路由的几个物理路由器(例如，如以下在图10中)。在一些实施例中，L3网关在其上操作的网关主机机器和外部物理路由器之间的物理连接是手动设置的并且在网络被配置时由管理员向网络控制器识别，而在其它实施例中，网络控制器识别每个网关自动连接到的路由器，而无需由用户输入这个信息。

[0150] 在一些实施例中，BGP守护进程为每个物理路由器识别路由器的IP地址、路由器所属的自主系统编号、与路由器的BGP会话的保持活动定时器、指定其后如果没有接收到保持活动消息则BGP守护进程应当假设路由器已出故障的持续时间的抑制时间，以及可选地用于认证的密码。单个BGP守护进程与其建立连接的不同物理路由器可以使用不同的BGP设置(例如，不同的保持活动或抑制定时器)并且属于不同的自主系统。

[0151] 接下来，过程(在920处)打开，或尝试打开，与每个识别出的BGP邻居的BGP会话。在一些实施例中，BGP守护进程作为每个连接的标准BGP状态机操作。即，守护进程基本上为与不同物理路由器的每个BGP连接实例化单独的状态机。对于每个连接，守护进程尝试过渡到

已建立(Established)状态,以便能够与物理路由器交换路由更新。即,BGP守护进程尝试发起与对等体的TCP连接、发送打开(Open)消息并接收返回的Open消息,并发送和接收保持活动消息,以便从连接(Connect)状态过渡到打开发送(OpenSent)状态再过渡到打开确认(OpenConfirm)状态并最后过渡到Established状态。当与对等路由器的连接处于Established状态时,BGP守护进程和对等路由器可以交换路由信息。

[0152] 但是,出于各种原因,BGP守护进程可能无法打开与其识别出的一个或多个邻居的会话(也被称为建立相邻性)。例如,如果在用于特定对等路由器的配置文件中提供的自主系统数量不匹配在对等路由器上配置的实际自主系统数量,则相邻性无法建立。过程900假设为每个BGP邻居建立相邻性-在一些实施例中,如果守护进程无法打开与特定路由器的会话,则它继续尝试(例如,尝试建立TCP会话、尝试发送和接收Open消息,等等)。

[0153] 该过程还(在925处)基于配置文件识别向已经与其建立BGP会话的其对等体通告的路由。在一些实施例中,BGP守护进程向它与其对等的每个路由器通告相同地址和前缀。这些可以是单个IP地址(例如,10.1.1.1)或者表示IP地址范围的CIDR前缀(例如,10.1.1/24)。在一些实施例中,BGP守护进程以CIDR斜线表示法(例如,利益/32表示单个IP地址)通告所有路由。

[0154] 利用识别出的前缀和地址,该过程(在930处)为与其已建立相邻性的每个识别出的邻居生成分组。在一些实施例中,这些分组是识别已知的可达的前缀以及分组要到达每个前缀将必需经过的自主系统列表的标准BGP更新(Update)分组。对于到逻辑交换机的路由,BGP分组通告子网(例如,10.1.1/24)以及仅单个(L3网关所属的)自主系统编号,这是因为,一旦到达L3网关,分组就不必为了到达VM主机而被发送到任何其它自主系统。

[0155] 在每次生成分组时,该过程(在935处)将从命名空间生成的分组发送到本地MFE,以便让MFE经外部网络将分组送出到目的地物理路由器邻居。如果BGP守护进程与三个不同的物理路由器建立了相邻性,则守护进程将经由MFE向三个不同的目的地发送相同的BGP Update分组。此外,几个不同的命名空间可以在用于不同逻辑路由器的同一台主机上运行BGP守护进程实例,在这种情况下,同一个路由器可以接收到通告完全不同的路由的几个不同的Update分组。

[0156] 图10-14概念性地示出了在L3网关中使用BGP来向用于逻辑网络的一组三个外部路由器通告路由的一个例子。图10示出了在受管理网络1025中的逻辑网络1000及那个逻辑网络的物理实现。如该图的上半部分中所示,逻辑网络1000类似于图1的逻辑网络100来配置,具有连接两个逻辑交换机1005和1010的单个逻辑路由器1015。第一逻辑交换机1005包括在子网10.0.0/24中的IP地址(有时写成10.0.0.0/24),并且第二逻辑交换机1510包括在子网10.0.1/24中的IP地址(有时写成10.0.1.0/24)。此外,逻辑路由器1015包括连接到已为其激活路由通告(例如,利用BGP)的外部网络1020的三个端口。

[0157] 图10的底部示出了逻辑网络1000的物理实现。在受管理网络中,VM主机机器1030的集合托管附连到逻辑交换机1005和1010的VM。这些VM主机1030可以各自托管来自逻辑网络的单个VM,并且一些可能托管或者来自相同逻辑交换机或者来自不同逻辑交换机的多个VM。VM主机上MFE的转发表各自实现逻辑交换机1005和1010以及逻辑路由器1015。此外,在一些实施例中,这些VM主机1030可以托管来自其它逻辑网络的VM,然后MFE的转发表也将实现这些其它逻辑网络。此外,一些实施例的受管理网络1025包括托管用于其它逻辑网络的

VM但是没有用于逻辑网络1000的VM驻留在其上的附加VM主机。

[0158] 此外，受管理网络1025包括三个网关主机1035-1045。这些网关主机1035-1045中每一个托管实现面向外部网络1020的三个逻辑路由器端口中一个的命名空间。具体而言，第一网关主机1035托管实现第一逻辑路由器端口的第一命名空间1050、第二网关主机1040托管实现第二逻辑路由器端口的第二命名空间1055并且第三网关主机1045托管实现第三逻辑路由器端口的第三命名空间1060。这些命名空间1050-1060中每一个操作BGP守护进程或其它路由协议应用，用于与附连的外部网络路由器交换路由信息。MFE还在网关主机1035-1045中每一个上操作。在一些实施例中，MFE各自实现逻辑交换机1005和1010以及逻辑路由器1015。当从VM传出的分组已经通过逻辑网络的大部分被处理时，这些MFE充当用于传入分组的第一跳MFE，并且在一些实施例中通过逻辑网络处理这些传入的分组。由于网关主机可以实现用于其它逻辑网络的其它命名空间，因此这些MFE也可以实现其它逻辑网络。

[0159] 在这个例子中，三个外部网络路由器1065-1075通过网关主机上的MFE连接到命名空间1050-1060。第一路由器1065连接到仅主机1035上的命名空间1050、第二路由器1070连接到全部三个命名空间1050-1060，而第三路由器1075连接到主机1045上的命名空间1060。这些路由器可以提供到互联网、其它网络等的连接。

[0160] 图11概念性地示出了由操作成控制受管理网络1025的控制器集群1100在网关主机1035-1045上的三个命名字空间1050-1060中指配BGP守护进程。在不同的实施例中，控制器集群1100可以是单个控制器、在主-备用(一个或多个)配置中操作的一对或一组控制器，或者诸如图3中所示的控制器的层次结构。如所示出的，基于进入的定义逻辑网络1000的配置信息，控制器集群1100向三个网关主机1035-1045发送BGP配置数据，以便指配在那些主机上的命名空间中操作的BGP守护进程。除其它信息外，BGP配置数据包括要通告的前缀(其对于每个网关是相同的)和BGP邻居(对等路由器)的列表。

[0161] 在这个例子中，控制器集群将数据1105发送到第一网关主机1035，指示前缀10.0.0/24和10.0.1/24和两个BGP邻居15.1.1.1和16.1.1.1(用于这个网关与其接口的两个路由器的IP地址)。控制器集群将数据1110发送到第二网关主机1040，指示相同的两个前缀和仅一个BGP邻居16.1.1.1。最后，控制器集群将数据1115发送到第三网关主机1045，指示相同的两个前缀和两个BGP邻居16.1.1.1和17.1.1.1。在一些实施例中，控制器集群以与用于网关的其它非流条目配置数据相同的格式发送这个数据(例如，当数据元组利用OVSDB协议被发送时)。从控制器发送的BGP配置数据还可以包括其它数据，诸如自主系统编号(这将跨网关是相同的)、用于网关的路由器标识信息，以及关于对等路由器的附加信息(例如，对等体的自主系统编号)。

[0162] 在从控制器集群1100接收到配置数据之后，在每个网关主机1035-1045上的应用(例如，在虚拟化软件中运行的守护进程)配置在其相应的命名空间上操作的BGP守护进程(例如，通过生成用于BGP守护进程的配置文件)。然后，BGP守护进程开始操作，并尝试设置与它们识别出的对等外部路由器的连接。例如，命名空间1050中的BGP守护进程建立与路由器1065和1070的两个单独TCP连接，然后再通过发送BGP Open和保持活动消息来进一步建立与这些路由器的BGP会话。如果这种消息也是从这些路由器接收的，则BGP守护进程可以向对等路由器发送Update分组。

[0163] 图12概念性地示出了根据一些实施例、由命名空间1050-1060中的BGP守护进程发

送的BGP Update分组。在一些实施例中，这些分组将它们自己识别为BGP Update分组(即，在BGP报头中)、识别源路由器，并识别用于各种前缀的可达性信息。在一些实施例中，这种可达性信息识别(i)CIDR格式中的前缀和(ii)如果被发送到Update分组的源，则分组为了到达由前缀定义的子集中的IP地址而将通过的自主系统的有序集合。例如，在典型的物理网络中，路由器可以识别通过自主系统15、8、6可达的前缀192.10.10.0/24(其中发送路由器位于自主系统15中)。

[0164] 在用于逻辑网络的L3网关的大多数情况下，到附连到逻辑交换机的VM的所有路由都将在其可达性信息中只具有单个自主系统，网关所属的自主系统。一般而言，或者每个逻辑网络是单个自主系统，或者受管理网络作为整体是单个自主系统。但是，在一些情况下，所通告的路由可以有一个以上自主系统(例如，如果受管理网络被划分成分组通过其以便到达逻辑交换机的多个自主系统)。

[0165] 如所示出的，命名空间1050分别向路由器1065和1070发送两个Update分组1205和1210。命名空间1050通过其本地MFE发送这些分组当中每一个，其本地MFE包括到网关主机1035的(一个或多个)NIC的桥。这些分组中每一个是相同的(除了目的地路由器信息之外)，该分组指示两个前缀10.0.0/24和10.0.1/24以及发送命名空间信息。命名空间1055将单个分组1215发送到路由器1070，该分组指示相同的前缀可达性数据但具有不同的自识别信息。最后，第三命名空间1060将两个分组1220和1225发送到路由器1070和1075，该分组也识别具有等效可达性信息的相同的两个前缀，具有其自己的自识别信息。

[0166] 作为接收这些Update分组的结果，外部路由器1065-1075更新自己的路由表。在一些实施例中，路由器将学习到的路由添加到其路由信息库(RIB)，然后重新计算到识别出的目的地的路由，以便在转发信息库(FIB)中使用。在一些实施例中，RIB包括路由器学习到的所有路由(经由连接、路由的手动输入或者或诸如BGP的动态路由协议)，而FIB包括路由器将实际用来转发分组的路由。

[0167] 路由器1065和1075只具有到达前缀10.0.0/24和10.0.1/24的单条途径-分别通过主机1035和1045上的L3网关。但是，路由器1070从全部三个网关主机1035-1045上的命名空间接收路由通告，每个网关主机将其自己指示为到达这些前缀的可能的下一跳。一般而言，当面对RIB中到达特定目的地IP地址或地址范围的多个路由时，物理路由器之一确定哪条路由最佳(例如，基于所遍历的自主系统的数量，或其它数据)并且选择最优路由在FIB中使用。但是，在这个例子中，对于10.0.0/24向路由器1070给出的三条可能路由是等效的。在一些实施例中，路由器1070简单地选择这些路由之一用于其FIB。但是，如果路由器1070能够进行相同成本的多路径(ECMP)转发，则路由器将所有这三条路由(即，到命名空间1050-1060的)都添加到其FIB，作为相同成本的选项。这使得流量跨三个网关散布，从而防止它们当中任何一个对于传入的流量变成单个瓶颈。

[0168] 图13和14概念性地示出了由进入受管理网络1025的流量所采取的路径。首先，图13示出了由从外部源发送到受管理网络中的、具有目的地IP地址10.0.1.1的VM的分组1300所采取的路径。在这个图中，由分组1300所采取的路径被示为粗虚线。分组到达外部路由器1065，它查阅其转发信息库。基于由路由器1065接收到的Update分组1205，其FIB指示具有在范围10.0.1/24内的目的地IP地址的分组应当被发送到命名空间1050。因而，外部路由器1065将分组转发到网关主机1035。

[0169] 分组到达网关主机1035上的MFE，它将分组转发到充当用于特定逻辑网络的网关的命名空间1050。在一些实施例中，外部路由器1065先前已经向网关主机1050发送了请求用于10.0.1.1的MAC地址的ARP请求，并且命名空间1050已利用其MAC地址作出响应。照此，分组1300被寻址到命名空间1050的MAC地址，其使得MFE将分组转发到这个目的地。

[0170] 命名空间1050接收分组、通过其IP网络堆栈(包括其路由表)处理其，并通过与MFE的不同接口将分组返回到MFE。在一些实施例中，在该命名空间中的处理流水线可以包括网络地址变换、防火墙处理和路由当中一些或全部。但是，对于具有多个网关的分布式逻辑路由器，由于状态共享的难度，一些实施例不允许有状态的服务，诸如网络地址变换，在该网关上执行。在一些实施例中，由命名空间执行的路由将目的地IP地址映射到网关附连到的逻辑路由器端口的目的地MAC地址，在一些实施例中。在其它实施例中，路由将目的地IP地址映射到分组被发送到的VM或其它实体的目的地MAC地址。当MFE通过不同的接口接收分组时，这使得MFE能够将分组视为进入逻辑路由器，在这个时候，MFE可以执行逻辑处理，以识别用于分组的逻辑交换机的逻辑出口端口，并且将分组发送VM主机1030中适当的一个。

[0171] 图14示出了通过路由器1070从外部网络发送到分别具有IP地址10.0.1.1和10.0.1.3的VM的两个分组1405和1410。在这种情况下，这两个分组1405都由路由器1070的FIB中的同一条目转发，但是发送到受管理网络1025中的不同网关。当外部路由器1070接收到分组1405时，FIB条目指示路由器使用ECMP技术来选择三个相同成本的目的地1050-1060中的一个。在一些实施例中，路由器1070散列分组属性的集合，以便确定将分组发送到哪个目的地。例如，一些实施例使用源和目的地IP地址，而其它实施例使用源和/或目的地MAC地址、传输连接5元组(源IP地址、目的地IP地址、传输协议、源传输端口号和目的地传输端口号)，或分组属性的其它组合。为了确定如何将散列结果关联到相同成本的目的地中特定的一个，一些实施例简单地计算该散列对列出的目的地数量取模。其它实施例使用诸如一致性散列或最高随机权重的算法，当网关被添加到或从相同成本的目的地列表中被除去时，所述算法对比简单的模N算法更少的流量修改目的地。

[0172] 不考虑所使用的算法(一些实施例甚至可以不使用散列函数，而是代替地使用其它负载平衡技术)，相同路由由用于逻辑路由器的几个活动L3网关通告到相同的外部物理路由器允许该物理路由器使用其ECMP技术在这几个网关之间散布流量。因此，在这种情况下，路由器1070将第一分组1405发送到命名空间1055并将第二分组1410发送到命名空间1060，但是这些分组由路由器中的相同转发条目支配。

[0173] 前面在图10-14中所示的例子示出了在受管理网络1025中实现的单个逻辑网络的例子。图15-18的例子概念性地示出了在网关的集合上实现的两个逻辑网络。在这种情况下，图15的上半部分示出了第一逻辑网络1500和第二逻辑网络1525的体系架构。这些逻辑网络具有相似的体系架构，其中第一逻辑网络1500包括将两个逻辑交换机1505和1510彼此连接并连接到外部网络1520的逻辑路由器1515。第一逻辑交换机1505包括在范围10.0.0/24内的IP地址，并且第二逻辑交换机1510包括在范围10.0.1/24内的IP地址。逻辑路由器1515包括连接到外部网络1520的四个端口。第二逻辑网络1525包括将两个逻辑交换机1530和1535彼此连接并连接到外部网络1520的逻辑路由器1540。第一逻辑交换机1530包括在范围11.0.0/24内的IP地址，并且第二逻辑交换机1535包括在范围11.0.1/24内的IP地址。逻辑路由器1540包括连接到外部网络1520的三个端口。在这种情况下，第一和第二逻辑网络

1500和1525属于不同的租户。

[0174] 图15的底部部分示出了这些网络在受管理网络1550中的物理实现，其类似于图10中所示的逻辑网络1000的物理实现。为简单起见，VM主机1545在这个图中被共同表示为单个框。虽然该图指示网关主机的每个MFE和VM主机1545之间的单个隧道，但是普通技术人员将认识到，在一些实施例中，每个网关主机具有到托管逻辑网络的VM的单独机器的众多单独隧道。

[0175] 受管理网络1550中实现这两个逻辑网络1500和1525的部分包括四个网关主机1555-1570。在这三个网关主机1555、1560、1570上，实现用于逻辑路由器1515和逻辑路由器1540二者的逻辑端口的命名空间操作。即，网关主机1555既托管为逻辑路由器1515实现到外部网络的第一连接的命名空间1557又托管为逻辑路由器1540实现到外部网络的第一连接的命名空间1559。网关主机1560既托管为逻辑路由器1515实现到外部网络的第二连接的命名空间1562又托管为逻辑路由器1540实现到外部网络的第二连接的命名空间1564。网关主机1570既托管为逻辑路由器1515实现到外部网络的第三连接的命名空间1572又托管为逻辑路由器1540实现到外部网络的第三连接的命名空间1574。最后，网关主机1565仅托管为逻辑路由器1515实现到外部网络的第四连接的单个命名空间1567(至少当考虑这两个逻辑网络的实现时-网关主机可以具有用于其它未示出的逻辑网络的命名空间)。因此，在一些实施例中，不同的逻辑路由器可以具有不同数量的面向外部网络的端口，如由管理员配置确定的。此外，网关主机1555-1570中每一个连接到仅单个外部物理路由器1575。

[0176] 图16示出了在七个命名空间1557-1574中由控制器集群1600指配BGP守护进程，类似于图11中所示的指配。但是，在这种情况下，控制器集群为实现用于逻辑路由器1515和1540二者的连接的命名空间生成BGP配置数据。在诸如图3中所示的使用控制器的分层网络的一些实施例中，控制器集群1600包括生成用于两个不同逻辑路由器的BGP配置的两个不同逻辑控制器。然后，这两个不同的逻辑控制器将都把生成的配置数据发送到物理控制器的同一集合，用于分发到网关主机。管理网关主机1555的物理控制器将从两个逻辑控制器接收数据，以分发到网关主机1555。

[0177] 即使同一控制器为两个BGP配置都生成数据，在一些实施例中，控制器也在单独的事务中分发这种数据。因此，与定义命名空间1559和其BGP配置的数据分开，网关主机1555接收定义命名空间1557和其BGP配置的数据。如所示出的，这些配置可以指定同一邻居路由器，但以不同的前缀通告。在一些实施例中，BGP邻居被存储为网关主机上的全局信息，供在主机上的各种命名空间中运行的所有BGP守护进程使用。即，网关主机具有到其的连接的每个外部路由器将是在网关主机上操作的BGP的所有实例的对等体。在其它实施例中，对等是在每个命名空间(每个L3网关)级别上确定的，并且特定主机上的一些BGP守护进程将与路由器对等，而其它的不对等。

[0178] 图17概念性地示出了，一旦在各种命名空间中运行的守护进程已建立与路由器的相邻性，就由各个BGP守护进程发送到外部路由器1575的BGP Update分组。这些分组类似于上面参考图12描述的那些。其结果是，路由器1575将具有用于被发送到在范围10.0.0/24和10.0.1/24内的IP地址的分组的四个相同成本的选项，以及用于被发送到在范围11.0.0/24和11.0.1/24内的IP地址的分组的三个相同成本的选项。

[0179] 图18概念性地示出由进入受管理网络1550的三个分组采取的路径。第一分组1805

和第二分组1810都具有10.0.1.1的目的地IP地址。但是，虽然具有相同的目的地，但这些分组可以具有不同的附加特性(例如，源IP地址、源和目的地传输端口号、传输协议，等等)。照此，利用其ECMP算法，路由器1575将分组发送到不同的命名空间(分组的路径由不同类型的虚线/点线指示)。路由器1575将第一分组1805发送到在网关主机1555中的命名空间1557，同时将第二分组1810发送到在网关主机1565中的命名空间1567。因此，即使被发送到相同IP地址的分组也会被不同地路由到网络中。但是，一些实施例需要外部路由器使用将来自相同传输连接的分组路由到同一网关的算法。使用基于在源/目的IP地址或者连接5元组的计算服务于这一目的。

[0180] 除了被发送到网关主机1555的分组1805，外部路由器1575还将具有目的地IP地址11.0.1.1的分组1815发送到这个网关主机。这第三个分组1815由位于网关主机1555的MFE发送到另一命名空间1559，该另一命名空间将分组路由回MFE，进行逻辑第一跳处理。在一些实施例中，MFE通过目的地MAC地址区分分组，如上所述。

[0181] 这部分提到不同类型的几个分组。术语“分组”在这里以及贯穿本申请被用来指以特定的格式跨网络被发送的位的集合。本领域普通技术人员将认识到，术语“分组”可在本文中被用来指可以跨网络被发送的位的各种格式化集合，诸如以太网帧、TCP段、UDP数据报、IP分组，等等。

[0182] V.作为路由服务器的控制器

[0183] 以上各部分描述了网络控制系统，其中网络控制器生成用于逻辑路由器实现的BGP配置，然后将那个配置发送到既执行用于网络的入口和出口路由又执行到外部网络中一个或多个路由器的路由通告的网关。但是，在一些实施例中，控制器或控制器集群具有到外部路由器的直接连接，并充当路由服务器。即，除了生成配置数据以便让受管理网络实现逻辑网络(例如，BGP配置数据、用于L3网关的路由表、用于MFE的流条目，等等)，控制器还向外部网络中的一个或多个路由器通告路由，由此防止这个流量占用网关MFE的数据路径中的带宽。

[0184] 一些实施例的控制器将BGP更新发送到外部路由器，其不是将分组的源识别为用于通告的前缀的下一跳，而是代替地将实现L3网关的命名空间之一识别为下一跳。此外，在一些实施例中，控制器从路由器接收BGP分组，它可以使用其来补充用于一个或多个逻辑路由器的路由表。

[0185] 图19概念性地示出了用于为逻辑网络生成BGP配置数据、然后通过生成该数据的控制器中的BGP服务实现那个配置数据的一些实施例的过程1900。在一些实施例中，过程1900的部分由控制器中的表映射引擎和/或路由生成引擎执行，而该过程的其它部分由控制器中的BGP应用执行。控制器生成BGP配置，但是随后将其提供给在内部运行的模块，而不是将该配置分发到运行BGP守护进程的网关主机。

[0186] 如所示出的，过程1900开始于(在1905处)接收创建具有连接到外部网络的一个或多个端口的逻辑路由器的指令。这些指令可以是网络管理员设计包括逻辑路由器的逻辑网络(例如，通过云管理应用，该云管理应用通过控制器API经过逻辑网络配置)的结果。在一些实施例中，创建逻辑路由器的指令具体地指示到外部网络的连接应当利用BGP或另一协议来实现，用于路由器对等和路由通告。在其它实施例中，这个能力对具有到外部网络的至少一个连接的所有逻辑路由器自动启用。

[0187] 接下来，过程(在1910处)为连接到逻辑网络的每个端口选择网关主机机器。一些实施例将每个端口分配到不同的网关主机，而其它实施例允许多个端口(以及因此托管路由表的多个命名空间)在单个网关主机上创建。在一些实施例中，网关主机是依据集群或故障域来布置的。在一些实施例中，这些集群可以是受管理网络中物理地一起定位的主机机器的集合，并且因此更有可能全都一起发生故障(例如，由于架顶式交换机的顶部掉落、电源问题等等)。不同实施例可以相对于集群不同地将网关分配给主机机器。例如，一些实施例只为特定的逻辑路由器每集群分配一个网关，而其它实施例将用于逻辑路由器的全部网关分配给同一集群。还有的其它实施例可以将网关分配给若干个不同的集群，但是允许两个或更多个网关在单个集群内。

[0188] 此外，在一些实施例中，网关主机机器可以基于那些网关主机被用于的功能被分配给不同的组。例如，在物理受管理网络中，一些实施例使用第一组网关主机用于提供逻辑服务(例如，DHCP、元数据代理)并且第二组网关主机用于L3网关。每个组可以横跨网关主机的几个集群，由此允许处理器从几个集群(即，故障域)选择在第二组内的网关主机机器。

[0189] 一些实施例允许管理员指定控制器向其分配逻辑路由器的每个逻辑端口的集群，并且控制器处理那个集群内实际网关主机的选择。因此，管理员可以指定将两个逻辑端口分配给第一集群中的网关、在第二集群中的四个网关，并且在第三集群中的再多两个网关。然后，控制器将每个逻辑端口分配给它选定的集群中的特定网关主机。对于这种分配，一些实施例使用负载平衡技术，诸如计算逻辑路由器或端口的属性(例如，由控制器分配的UUID)的散列函数对集群中网关主机的数目取模。这有效地随机(尽管算法本身是确定的)将逻辑路由器端口分配给集群内的网关主机，并且因此从长远来看跨网关主机负载平衡L3网关。

[0190] 一些其它实施例可以使用其它技术来跨集群中的主机负载平衡L3网关。例如，不是使用散列算法在集群中的所有网关主机之间进行选择，而是一些实施例在仅仅具有最少量的当前在操作的逻辑路由器的那些网关之间进行选择，并且用网关的这个较小数目对散列函数的结果取模。其它实施例分析每个网关上的逻辑路由器的数目和网关的操作负载(例如，基于经特定的时间帧被处理的分组的数目)，以便确定特定的逻辑路由器应当被分配给哪个网关主机。

[0191] 接下来，过程1900(在1915处)为VM主机和选定的网关主机机器上的MFE都生成流条目，以便以分布式方式实现逻辑路由器并且在受管理网络中转发分组，以及处理进入或离开网络的分组，并且为路由表生成数据元组，用于处理实现连接到外部网络的每个逻辑端口的L3网关中的分组。这各种流条目和路由表数据元组在上面参考例如图5详细描述过了。

[0192] 然后，该过程(在1920处)将生成的数据元组和/或流条目分发到各个主机机器。在一些实施例中，两种类型的数据(流条目和路由表数据元组)经由不同的协议分发。一些实施例经由诸如OpenFlow的第一协议将流条目分发到VM主机和网关主机二者，而经由诸如OVSDB的第二协议将路由表数据元组分发到网关主机。在一些实施例中使用的OVSDB协议还携带用于MFE的配置信息(用于VM主机和网关主机两者)。

[0193] 除了生成和分发数据用于在网络中分组转发的指配，一些实施例的控制器还负责生成路由协议(例如，BGP)配置并处理与外部路由器的路由信息的交换。照此，过程1900(在

1925处)为连接到外部网络的每个逻辑端口(即,每个L3网关)识别与其对等的(一个或多个)外部网络路由器的地址(以及其它信息)。在一些实施例中,管理员输入用于每个逻辑端口的这种数据,并且处理确保外部路由器被正确连接到网关主机(或者,例如,网关主机连接到其的架顶式交换机的顶部)。在其它实施例中,网络控制器基于其存储的网络状态信息自动地确定每个网关主机连接到其的外部路由器的集合,并且使用这些作为与L3网关对等的外部网络路由器。在一些路由服务器实施例中,管理员还确保控制器能够与外部路由器连接。在各种不同的实施例中,(一个或多个)控制器经由直接连接、通过受管理网络中的其它机器(例如,网关或其它主机机器)等连接到路由器。

[0194] 通过为每个逻辑端口识别出外部路由器,该过程(在1930处)在控制器上利用识别出的外部路由器、逻辑网络配置和选定的主机机器生成并安装BGP配置。在一些实施例中,控制器实例化用于它对其充当路由服务器的每个L3网关的单独BGP过程。因此,如果逻辑路由器被定义为具有面朝外部网络的三个端口,则控制器实例化三个BGP过程(例如,上述BGP守护进程,或者不同的BGP应用),用于通告用于三个网关中每一个的路由。在其它实施例中,控制器实例化为实现用于逻辑路由器的端口的所有网关执行路由通告的单个BGP过程。在一些此类实施例中,单个BGP过程为由控制器管理的所有逻辑路由器(例如,用于多个不同的逻辑网络)处理路由通告。

[0195] 为了生成用于特定逻辑路由器的BGP配置,控制器(例如,控制器中的表映射引擎)识别用于附连到逻辑路由器的逻辑交换机的CIDR前缀,因为这些是作为路由服务器的控制器将通告给外部路由器的前缀(这将对每个网关是相同的)。此外,控制器使用对网关主机机器的选择用于BGP配置,以及为将在网关主机机器上运行的命名空间生成的信息。控制器上的BGP过程将发出通告这些命名空间的分组(而不是其本身)作为用于被通告的路由的实际的(一个或多个)下一跳,并且因此必须能够提供关于命名空间的必需数据(例如,自主系统编号、路由器标识符,等等)。此外,配置需要为每个命名空间识别与其交换路由信息的外部路由器。在一些情况下,命名空间到外部路由器的连接可以类似于图10中的那些(即,不同的L3网关用于连接到不同外部路由器集合的逻辑路由器),在这种情况下,控制器不能简单地向每个外部路由器通告下一跳目的地的相同集合。代替地,控制器存储用于每个下一跳L3网关的的邻居列表,使得其可以将分组发送到通告特定L3网关的这些邻居当中每一个作为到逻辑网络的路由的下一跳。

[0196] 在一些实施例中,控制器生成用于(一个或多个)BGP实例的配置文件,或几个配置文件。这些配置文件可以类似于由上述命名空间守护进程生成的文件。控制器在BGP过程可以访问这些文件并加载它们的配置的位置存储配置文件。在这个时候,控制器可以开始充当路由服务器来联系外部路由器。

[0197] 照此,过程1900(在1935处)打开,或尝试打开,与在配置中识别出的邻居外部路由器的(一个或多个)BGP会话。如在前一部分中所描述的内联模型中那样,几个BGP会话被起动,每个BGP会话作为其自己的独立状态机来操作。例如,如果逻辑网络包括面对外部网络的三个端口(以及因此三个网关),其中每一个连接到两个不同的外部路由器,则在一些实施例中控制器将发起六个单独的BGP会话。在其它实施例中,控制器对每个外部路由器发起仅一个BGP会话,并且发送指定用于通告给外部路由器的路由的几种不同下一跳选项的Update。这个过程1900假设为每个BGP会话建立相邻性-如果BGP过程未能打开与特定路由

器的会话,则在一些实施例中控制器在过渡到操作1940之前继续尝试这样做。

[0198] 利用BGP配置数据,过程(在1940处)为每个建立的BGP会话生成分组。在一些实施例中,这些分组是识别已知的可达前缀、用于那些前缀的下一跳目的地以及分组为到达每个前缀必需经过的自主系统的列表的标准的BGP Update分组。在这种情况下,发送Update分组的控制器不是下一跳-分组代替地识别其中一个L3网关作为下一跳。对于到逻辑交换机的路由,BGP分组通告子网(例如,10.1.1/24)以及仅单个(L3网关所属的)自主系统编号,因为,一旦到达L3网关,分组就不必为了到达VM主机而被发送到任何其它自主系统。

[0199] 对于每个生成的分组,该过程(在1945处)将从控制器生成的分组发送到目的地物理路由器。如上面所提到的,这个连接可被实现为控制器和外部路由器之间的直接连接,或者可以行进通过受管理网络的部分(例如,网关,等等)。如果控制器上的BGP过程建立与用于三个L3网关下一跳的三个不同物理路由器的相邻性,则该过程将把三个不同的BGP Update分组各自发送到三个不同的目的地。此外,控制器可能充当用于几个不同的逻辑网络的路由服务器,在这种情况下,控制器还发送通告完全不同的路由的几个不同的Update分组。

[0200] 图20-22概念性示出了将控制器用作为为逻辑网络通告到外部路由器的路由的路由服务器的例子。图20示出了逻辑网络2000和在受管理网络2025中那个逻辑网络的物理实现。如该图的上半部分中所示,逻辑网络2000类似于前一部分中例子的逻辑网络1000来配置,其中单个逻辑路由器2015连接两个逻辑交换机2005和2010。第一逻辑交换机2005包括在子网12.0.0/24中的IP地址,并且第二逻辑交换机2010包括在子网12.0.1/24中的IP地址。此外,逻辑路由器2015包括连接到外部网络2020的三个端口,对于该外部网络,利用控制器作为路由服务器的路由通告被激活。

[0201] 图20的底部部分示出了逻辑网络2000的物理实现。为简单起见,VM主机2030在这个图中被共同表示为单个框,如在上面的图15中。连接到外部网络2020的逻辑路由器2015的三个端口被实现为分别在网关主机2035-2045上操作的命名空间2050-2060中的L3网关。在这种情况下,三个网关主机各自连接到相同的单个外部路由器2065,以便发送和接收进入和退出逻辑网络的分组。

[0202] 但是,不像先前的例子,命名空间2050-2060不操作BGP守护进程或任何其它路由协议应用,仅运作来处理进入和外出的分组。相反,控制器集群2070操作成(i)向主机机器2030-2045提供指配数据和(ii)作为路由服务器操作,以与外部路由器2065交换路由信息。在这个图中,控制器集群2070和主机机器2030-2045之间的虚线指示控制路径连接,而实线(网关2035-2045和路由器2065之间、网关2035-2045和VM主机2030之间,以及控制器集群2070和路由器2065之间)指示数据路径连接。

[0203] 图21概念性地示出了由控制器集群2070发送的一些控制和数据路径数据,以便实现逻辑路由器2015。如图所示,控制器集群2070向网关主机2035-2045分发逻辑路由器配置数据2105(例如,作为定义用于命名空间的路由表的数据元组、作为用于MFE的流条目,等等)。在一些实施例中,控制器集群在两个通道中发送这种数据,其中用于MFE的流条目经由第一协议(例如,OpenFlow)发送并且定义命名空间和用于命名空间的路由表的数据元组经由第二协议(例如,OVSDB)发送。一些实施例的控制器集群通过控制器的层次分发逻辑路由器配置数据2105,其中单个逻辑控制器生成该数据并将数据分发到管理并直接向三个网关

主机2045提供数据的各种物理控制器。

[0204] 此外，控制器集群2070将三个单独的BGP分组发送到外部网络路由器2065。一些实施例建立与外部路由器2065的三个单独的会话(控制器对其充当路由服务器的每个网关有一个)，而其它实施例作为单个会话的一部分发送三个BGP Update。这些BGP分组各自(i)通告CIDR前缀12.0.0/24和12.0.1/24、(ii)对每个前缀指示用来到达由该前缀定义的范围内的地址的自主系统的有序列表(在大多数情况下，这对于逻辑网络将是单个自主系统)，以及(iii)识别用于被通告的前缀的下一跳。在一些实施例中，仅这个下一跳在三个分组之间变化，因为这识别不同的网关。

[0205] 作为接收到这三个分组的结果，物理路由器2065更新其路由表，以包括用于在所识别出的IP地址范围(12.0.0/24和12.0.1/24)内的分组的三种可能的相同成本的下一跳。假设路由器2065具有ECMP能力，它将在主机2035-2045上的三个L3网关之间散布用于这些IP范围的流量。图22概念性地示出了由进入受管理网络2025的几个分组2205和2210采取的路径。两个分组都由逻辑路由器2065接收，并且由相同的转发信息库条目处理。这个条目声明使用ECMP算法在用于分组的三种可能的下一跳(L3网关)之间作出决定。其结果是，路由器将第一分组2205发送到网关主机2040上的命名空间2055并且将第二分组2210发送到到网关主机2045上的命名空间2060。MFE和命名空间如上面在前一部分中所述的那样处理分组，以便将分组转发到目的地虚拟机。

[0206] 图23概念性地示出了对逻辑网络充当路由服务器的一些实施例的控制器2300的软件体系架构。如图所示，控制器2300包括输入接口2305、表映射状态计算模块2310、主机分配模块2315、分发接口2320、BGP服务2325，以及外部网络接口2330。此外，网络控制器2300包括在一些实施例中存储表映射状态计算模块的输入和/或输出的一个或多个状态存储数据库2335。

[0207] 一些实施例的输入接口2305从一个或多个用户接收输入，以定义逻辑网络(例如，通过逻辑交换机、逻辑路由器、中间件、网关连接到外部网络的VM的集合，等等)。例如，用户可以定义逻辑网络，诸如如上所述在图20中示出的。在一些实施例中，在输入接口接收到的请求依据由用户输入(或选择)的源和目的地MAC地址指定逻辑端口。

[0208] 当输入接口2305接收到逻辑网络的规范时，一些实施例的接口将这个规范变换成定义逻辑网络的逻辑控制平面数据，并将这个数据传递到表映射状态计算模块2310。在一些实施例中，输入接口2305将这个逻辑控制平面数据读到状态计算模块2310的输入表中。一些实施例的表映射状态计算模块2310包括具有输入表和输出表的集合的表映射引擎，并且根据规则集将输入表中的记录映射到输出表中的记录。更具体而言，一些实施例将逻辑控制平面数据变换成逻辑转发平面数据并且随后将逻辑转发平面数据变换成可被向下传递到实现逻辑网络的MFE的通用或定制物理控制平面数据。一些实施例的表映射状态计算模块2310使用nLog，并且在美国公布2013/0058228中更详细地描述，其通过引用被结合于此。

[0209] 除了生成物理控制平面数据，在一些实施例中表映射状态计算模块2310还生成其它数据元组，诸如用于路由表的那些，以及BGP配置数据。如上所述，状态计算模块可以使用由主机分配模块2315为托管网关选择的主机的集合、连接到逻辑网络的VM的IP地址范围以及通过输入接口输入的关于(一个或多个)外部路由器的信息来计算BGP配置数据元组。

[0210] 在一些实施例中，表映射状态计算模块2310在(一个或多个)状态存储数据库2335中存储其输出状态。在一些实施例中，这个数据库2335存储MAC地址到逻辑端口绑定、由表映射状态计算模块2335输出的物理控制平面数据、路由表数据元组、BGP配置信息，以及其它数据。

[0211] 在一些实施例中，主机分配模块2315使用散列函数或其它算法来选择用于逻辑网络的网关主机。基于由状态计算模块2310提供的信息，主机分配模块2315确定网关主机的集合，并将这个选择返回到状态计算模块。例如，在一些实施例中，基于逻辑网络配置输入，状态计算模块2310指定特定的逻辑路由器将具有位于网关主机集群的特定集合中的特定数量的L3网关。状态计算模块2310请求主机分配模块2315选择特定集群中的特定网关主机、状态计算模块在生成状态和BGP配置时所使用的信息。

[0212] 如所示出的，控制器2300通过其MFE接口2320将数据分发到主机机器(VM主机和网关主机)。通过这个接口，控制器在主机机器向MFE、L3网关等分发物理控制平面数据、路由表和配置数据元组等。在一些实施例中，接口是到主机机器的直接连接，而在其它实施例中，控制器2300是将所生成的数据分发到物理控制器的集合的逻辑控制器。此外，在内联模型实施例中，其中BGP服务在网关而不是控制器中操作，控制器使用这个接口来分发BGP配置数据元组。

[0213] 但是，在所示出的实施例中，BGP服务2325在控制器上操作。这个BGP服务从表映射状态计算2310接收并安装配置或配置的集合(例如，作为数据元组的集合)，然后根据这种配置与受管理网络外面的路由器建立BGP会话。在一些实施例中，BGP服务2325结合了命名空间守护进程和BGP守护进程的功能，因为它接收定义配置的数据元组、生成可用于实例化BGP过程的配置文件、读取并安装配置文件，并建立和参与与外部路由器的BGP会话。

[0214] 一些实施例的BGP服务2325通过外部网络接口2330打开并建立与外部路由器2340的BGP会话。这个接口在一些实施例中可以是处理IP分组的NIC，类似于网关和外部路由器之间的连接。通过这个接口，BGP服务2325将针对其建立的每个BGP会话的更新发送到外部路由器2340，使得路由器2340能够经由由控制器2300指配的网关将分组转发到逻辑网络。

[0215] 除了将进入逻辑网络的路由通告到外部路由器，在一些实施例中，作为路由服务器的控制器集群从外部路由器接收BGP分组并使用这些来更新用于逻辑网络的路由表。一般而言，BGP是双向协议，因为对等会话中的每个路由器在会话中向另一个路由器发送其路由信息。照此，一些实施例的(一个或多个)外部路由器将它们的信息发送到控制器集群，指示可达的IP地址和前缀。如在图10中，如果一些L3网关连接到多个路由器，则控制器集群可以对由L3网关通告的各种IP地址确定哪些外部路由器是用于该IP地址的最佳下一跳。然后，控制器集群可以将这个信息添加到其分发到L3网关的路由表。

[0216] 虽然上述部分描述了使用控制器作为路由服务器，但是一些实施例代替地使用与处理对于逻辑网络的入口和出口流量的网关主机分开的一个或多个网关主机机器，作为用于逻辑路由器的路由服务器。图24概念性地示出了逻辑网络(结构与图1或图10的类似)在其中实现并且使用单独的网关作为路由服务器的一些实施例的这种受管理网络2400。为简单起见，这个图没有示出附连到逻辑网络的VM驻留在其上的主机机器。

[0217] 逻辑路由器具有连接到外部网络的三个端口，并且因此这些端口在三个命名空间2420-2430中的三个网关2405-2415上实现。这些命名空间作为L3网关操作，以处理入口和

出口流量，但是不操作路由协议应用，并且因此不与外部网络路由器2435交换数据。代替地，控制器选择第四网关主机2440作为用于逻辑网络的路由服务器操作。命名空间2445在网关主机2440上操作，运行类似于上面在部分II中所示的那些的BGP守护进程。

[0218] 如所示出的，控制器集群2450生成并(i)向三个网关主机2405-2415分发逻辑路由器配置数据，以便配置命名空间2420-2430中的L3网关和(ii)向网关主机2440分发BGP配置数据，以便配置在命名空间2440中操作的BGP守护进程。这使得命名空间2445能够打开与外部路由器2435的一个或多个BGP会话并向外部路由器通告指示三个L3网关作为用于逻辑网络的IP地址的可能下一跳的路由信息。

[0219] VI.电子系统

[0220] 许多上述特征和应用被实现为软件过程，该软件过程被指定为一组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。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、处理器内核、或其它处理单元)执行时，它们使得这(一个或多个)处理单元执行在指令中所指示的动作。计算机可读介质的例子包括，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或通过有线连接传递的载波和电子信号。

[0221] 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用于被处理器处理的存储在磁储存装置中的应用。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是在本发明的范围之内。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，该软件程序定义运行和执行该软件程序的操作的一个或多个特定的机器实现。

[0222] 图25概念性地示出了本发明的一些实施例利用其来实现的电子系统2500。可以使用电子系统2500来执行控制、虚拟化或上述操作系统应用中的任意个。电子系统2500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)、电话、PDA或任何其它种类的电子设备。这种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2500包括总线2505、(一个或多个)处理单元2510、系统存储器2525、只读存储器2530、永久储存装置2535、输入设备2525、以及输出设备2545。

[0223] 总线2505统一地表示通信连接电子系统2500的众多内部设备的所有系统、外设和芯片组总线。例如，总线2505将(一个或多个)处理单元2510与只读存储器2530、系统存储器2525、及永久储存装置2535通信地连接。

[0224] 从这些各种存储器单元中，(一个或多个)处理单元2510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心处理器。

[0225] 只读存储器(ROM)2530存储由(一个或多个)处理单元2510和电子系统的其它模块所需要的静态数据和指令。另一方面，永久储存装置2535是读和写存储器设备。这个设备是即使当电子系统2500关闭时也存储指令和数据的非易失性存储单元。本发明的一些实施例使用大容量存储设备(诸如磁或光盘及其对应的盘驱动器)作为永久储存装置2535。

[0226] 其它实施例使用可移除存储设备(诸如软盘、闪存驱动器等)作为永久存储设备。与永久储存装置2535一样,系统存储器2525是读和写存储器设备。但是,与储存装置2535不同,系统存储器是易失性读和写存储器,例如随机存取存储器。系统存储器存储处理器在运行时需要的一些指令和数据。在一些实施例中,本发明的过程被存储在系统存储器2525、永久储存装置2535和/或只读存储器2530中。从这些各种存储器单元中,(一个或多个)处理单元2510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。

[0227] 总线2505还连接到输入和输出设备2525和2545。输入设备使用户能够传递信息和选择到电子系统的命令。输入设备2540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2545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)。一些实施例包括诸如用作输入和输出设备两者的触摸屏的设备。

[0228] 最后,如在图25中所示,总线2505还通过网络适配器(未示出)将电子系统2500耦合到网络2565。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、或诸如互联网的网络中的网络)的一部分。电子系统2500的任何或所有组件可以与本发明结合使用。

[0229] 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储存装置和存储器。这种计算机可读介质的一些例子包括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)、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例子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别代码的文件。

[0230] 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来执行,诸如专用集成电路(ASIC)或现场可编程门阵列(FPGA)。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。

[0231] 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、以及任何其它短暂的信号。

[0232] 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5、8、9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与所示出和描述的确切顺序来执行。特定操作可能没有在一个连续系列的操作中执行,并且不同的特定操作可能在不同的实施例中执行。此外,过程可以利用若干个子过程来实现,或者作为较大宏过程

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。

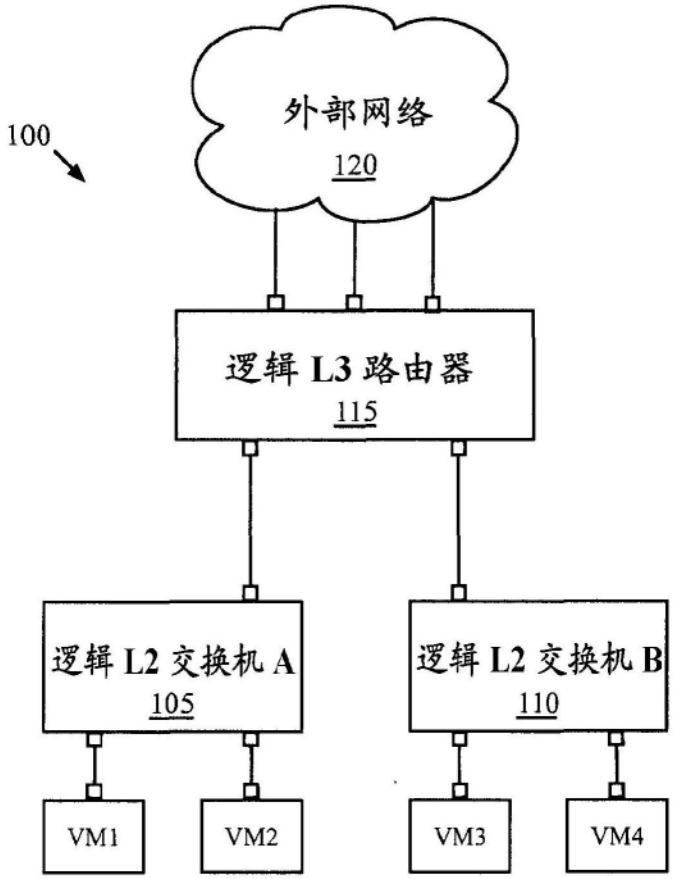

100
外部网络
120
逻辑 L3 路由器
115
逻辑 L2 交换机 A
105
逻辑 L2 交换机 B
110
VM1
VM2
VM3
VM4

图1

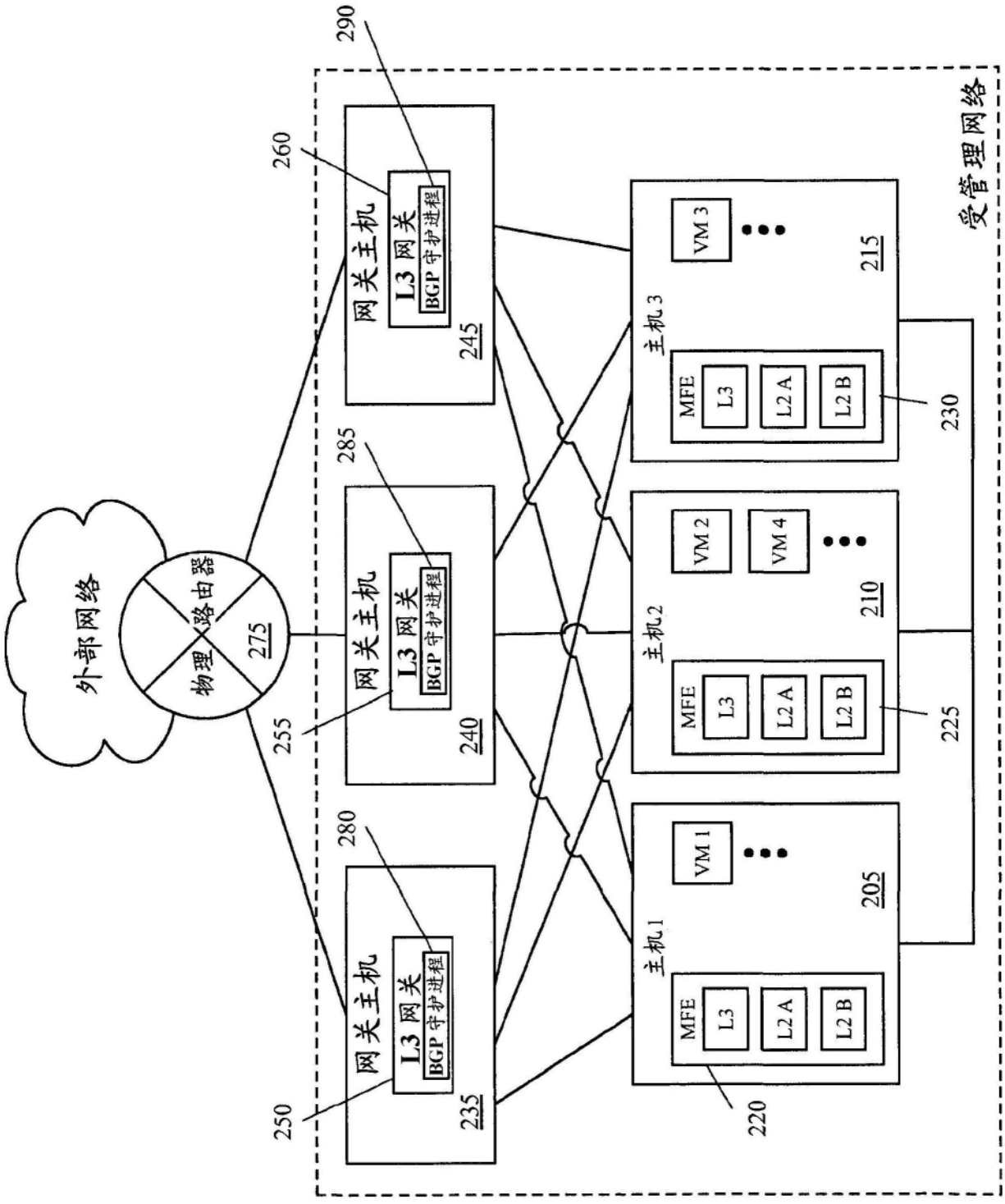

外部网络
物理路由器
275
受管理网络
网关主机
L3 网关
BGP 守护进程
245
240
235
290
285
280
260
255
250
主机1
主机2
主机3
MFE
L3
L2 A
L2 B
VM 1
VM 2
VM 3
VM 4
205
210
215
220
225
230

图2

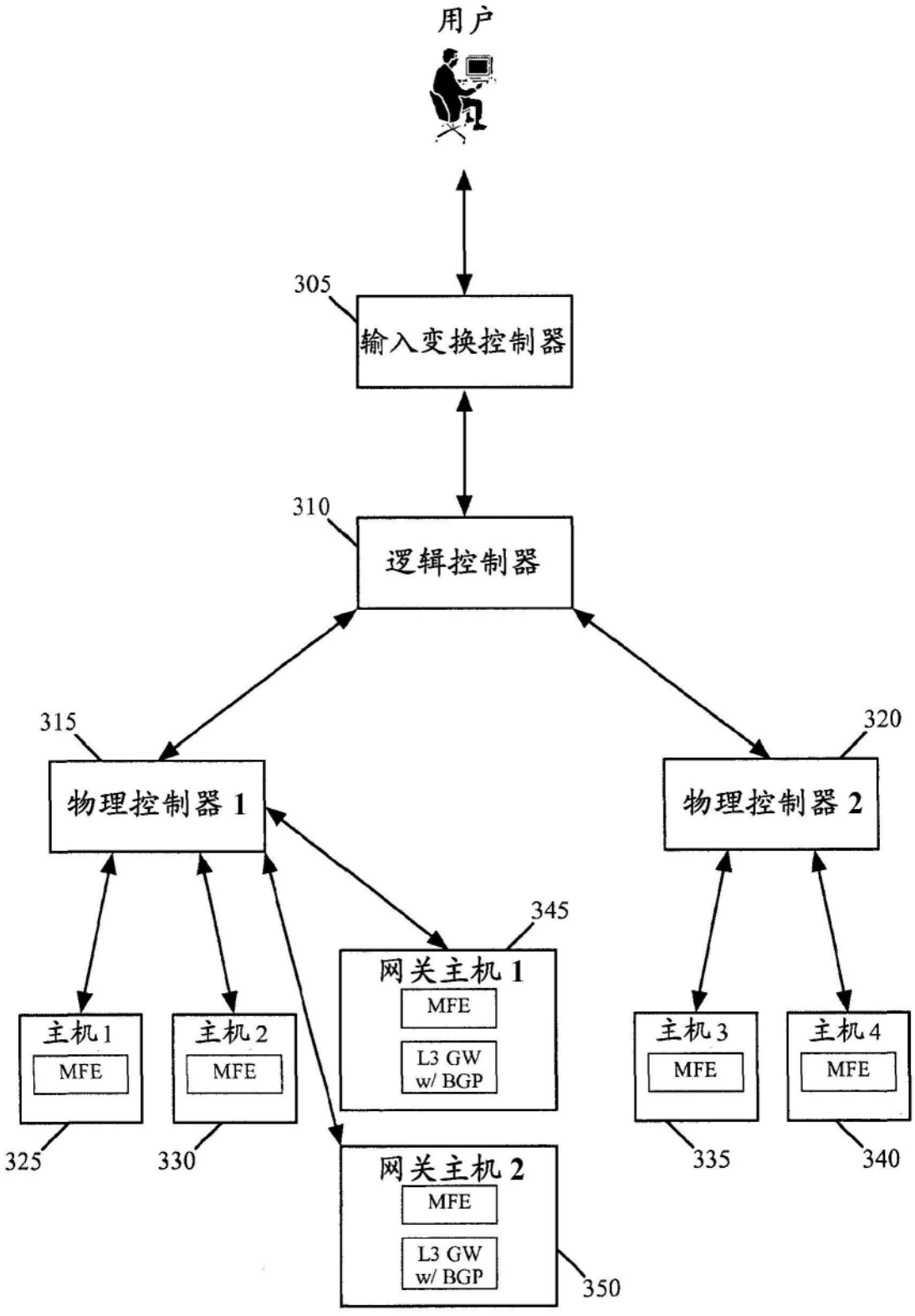

用户
305
输入变换控制器
310
逻辑控制器
315
物理控制器 1
320
物理控制器 2
345
网关主机 1
MFE
L3 GW w/ BGP
主机 1
MFE
主机 2
MFE
主机 3
MFE
主机 4
MFE
网关主机 2
MFE
L3 GW w/ BGP
325
330
335
340
350

图3

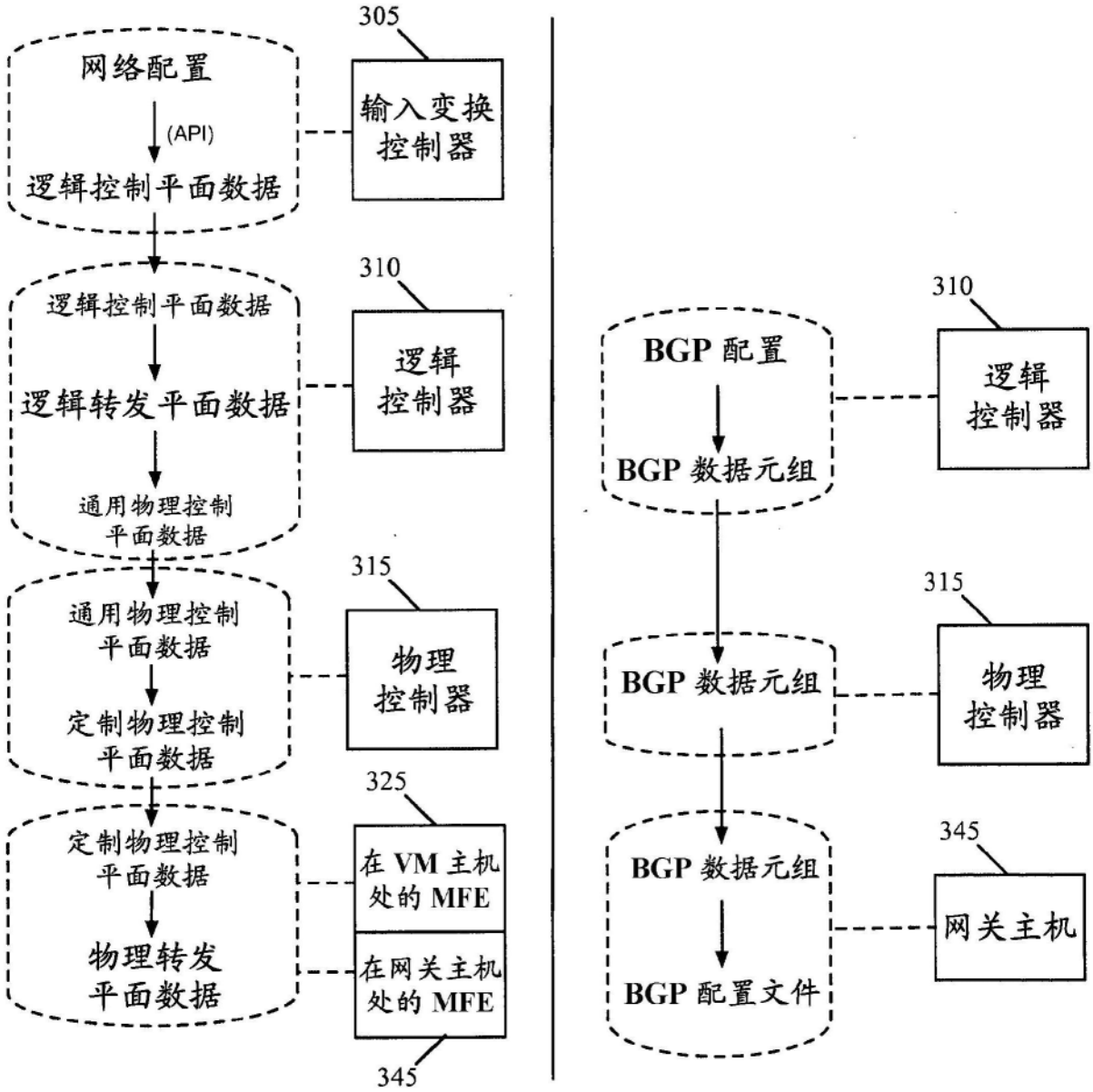

网络配置
(API)
逻辑控制平面数据
305
输入变换控制器
逻辑控制平面数据
逻辑转发平面数据
通用物理控制平面数据
310
逻辑控制器
通用物理控制平面数据
定制物理控制平面数据
315
物理控制器
定制物理控制平面数据
物理转发平面数据
325
在 VM 主机处的 MFE
在网关主机处的 MFE
345
BGP 配置
BGP 数据元组
310
逻辑控制器
BGP 数据元组
315
物理控制器
BGP 数据元组
BGP 配置文件
345
网关主机

图4

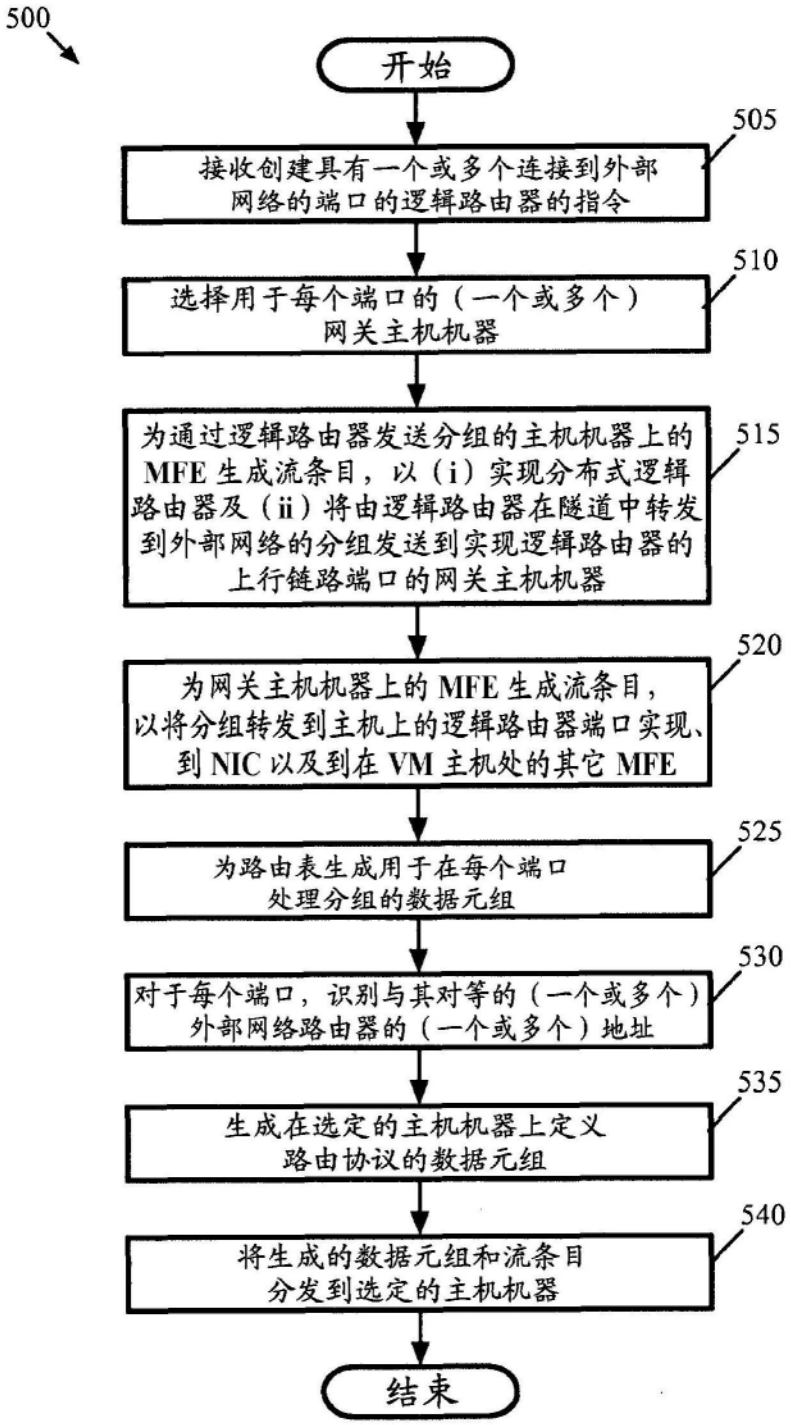

500
开始
505
接收创建具有一个或多个连接到外部网络的端口的逻辑路由器的指令
510
选择用于每个端口的（一个或多个）网关主机机器
515
为通过逻辑路由器发送分组的主机机器上的MFE生成流条目，以（i）实现分布式逻辑路由器及（ii）将由逻辑路由器在隧道中转发到外部网络的分组发送到实现逻辑路由器的上行链路端口的网关主机机器
520
为网关主机机器上的MFE生成流条目，以将分组转发到主机上的逻辑路由器端口实现、到NIC以及到在VM主机处的其它MFE
525
为路由表生成用于在每个端口处理分组的数据元组
530
对于每个端口，识别与其对等的（一个或多个）外部网络路由器的（一个或多个）地址
535
生成在选定的主机机器上定义路由协议的数据元组
540
将生成的数据元组和流条目分发到选定的主机机器
结束

图5

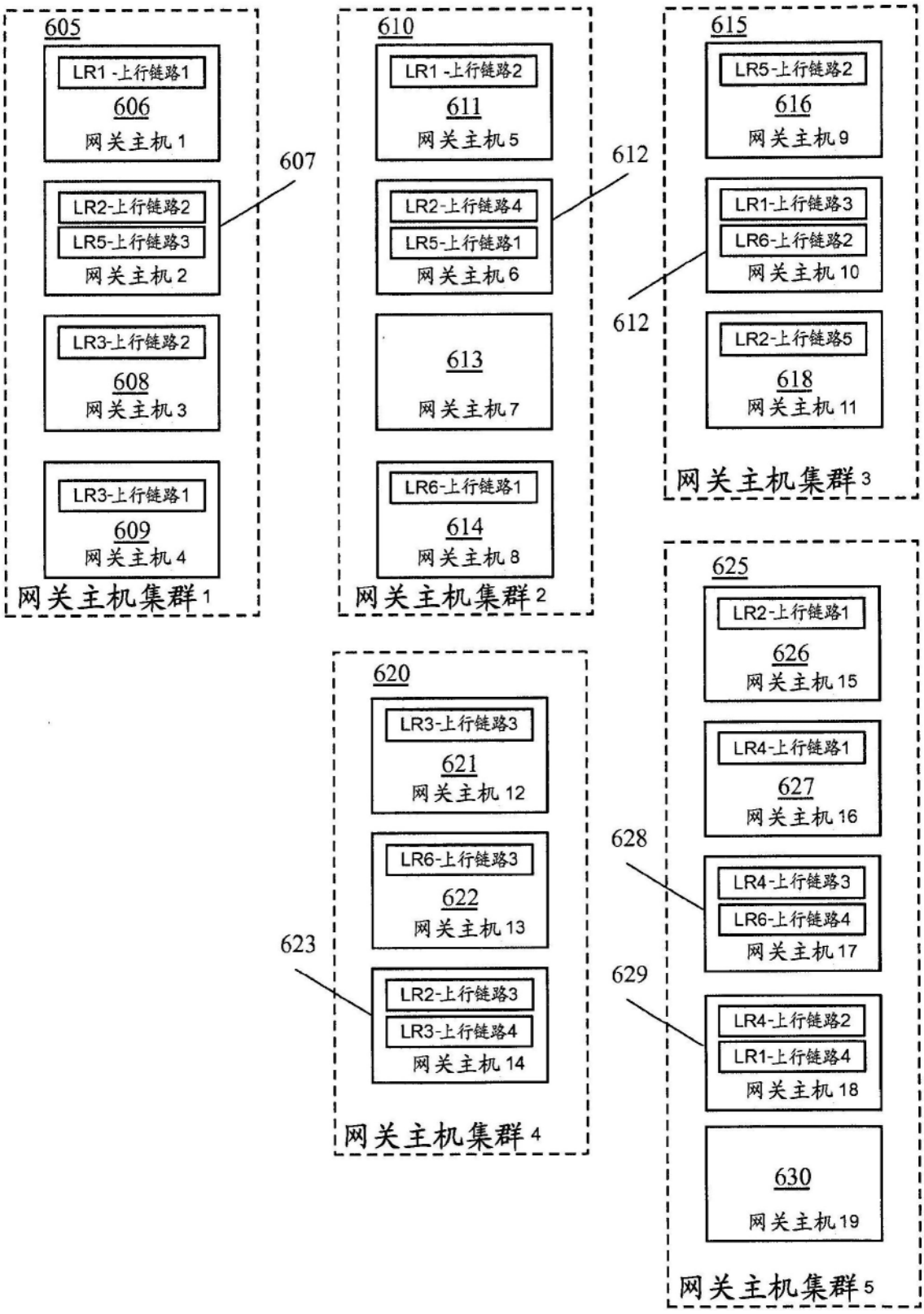

605
LR1 -上行链路1
606
网关主机 1
LR2-上行链路2
LR5-上行链路3
网关主机 2
LR3-上行链路2
608
网关主机 3
LR3-上行链路1
609
网关主机 4
网关主机集群 1
607
610
LR1 -上行链路2
611
网关主机 5
LR2-上行链路4
LR5-上行链路1
网关主机 6
613
网关主机 7
LR6-上行链路1
614
网关主机 8
网关主机集群 2
612
615
LR5-上行链路2
616
网关主机 9
LR1-上行链路3
LR6-上行链路2
网关主机 10
612
LR2-上行链路5
618
网关主机 11
网关主机集群 3
620
LR3-上行链路3
621
网关主机 12
LR6-上行链路3
622
网关主机 13
623
LR2-上行链路3
LR3-上行链路4
网关主机 14
网关主机集群 4
625
LR2-上行链路1
626
网关主机 15
LR4-上行链路1
627
网关主机 16
628
LR4-上行链路3
LR6-上行链路4
网关主机 17
629
LR4-上行链路2
LR1-上行链路4
网关主机 18
630
网关主机 19
网关主机集群 5

图6

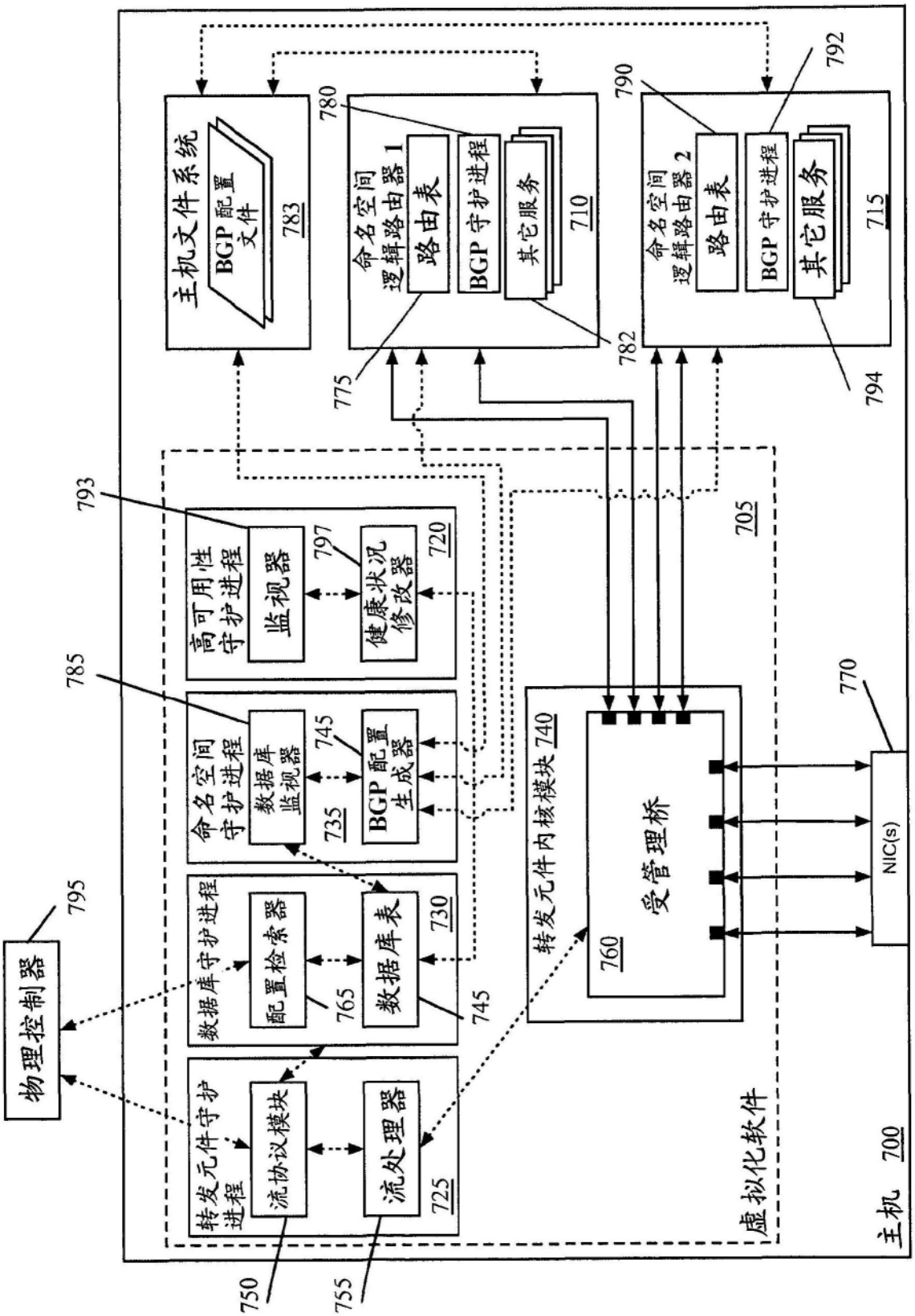

主机文件系统
BGP 配置文件
783
命名空间逻辑路由器 1
路由表
BGP 守护进程
其它服务
710
780
命名空间逻辑路由器 2
路由表
BGP 守护进程
其它服务
715
790
792
775
782
794
793
高可用性守护进程
监视器
健康状况修改器
797
720
705
785
命名空间守护进程
数据库监视器
BGP 配置生成器
735
745
770
转发元件内核模块 740
受管理桥
760
NIC(s)
795
物理控制器
数据库守护进程
配置检索器
数据库表
730
765
745
转发元件守护进程
流协议模块
流处理器
725
750
755
虚拟化软件
主机 700

图7

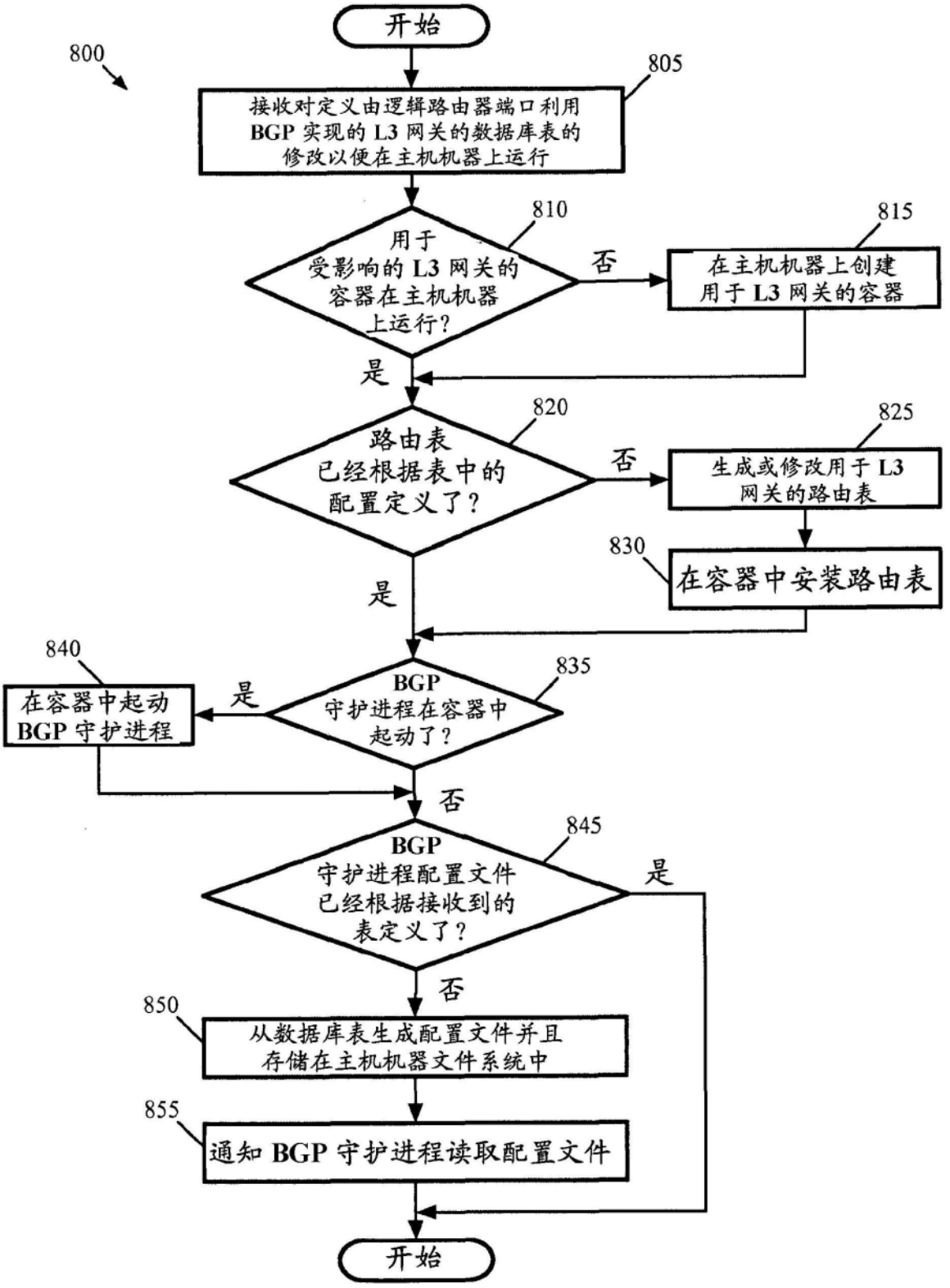

开始
800
805
接收对定义由逻辑路由器端口利用BGP 实现的 L3 网关的数据库表的修改以便在主机机器上运行
810
用于受影响的 L3 网关的容器在主机机器上运行?
否
815
在主机机器上创建用于 L3 网关的容器
是
820
路由表已经根据表中的配置定义了?
否
825
生成或修改用于 L3 网关的路由表
830
在容器中安装路由表
是
840
在容器中起动 BGP 守护进程
是
835
BGP 守护进程在容器中起动了?
否
845
BGP 守护进程配置文件已经根据接收到的表定义了?
是
否
850
从数据库表生成配置文件并且存储在主机机器文件系统中
855
通知 BGP 守护进程读取配置文件
开始

图8

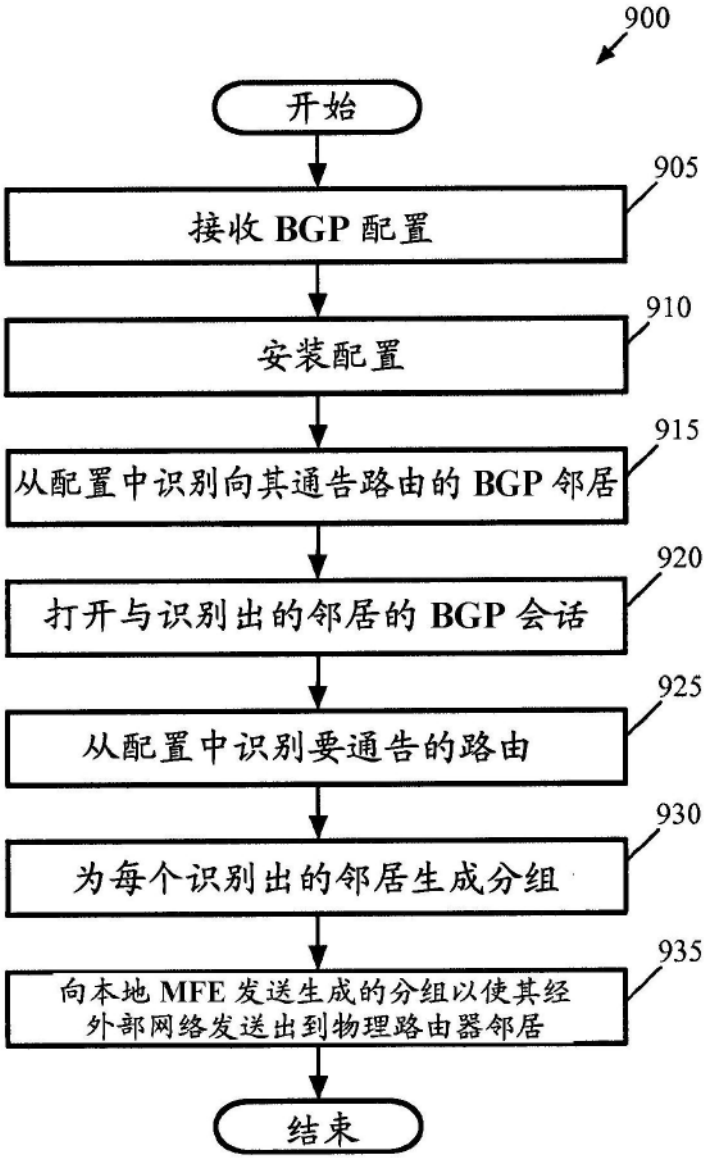

900
开始
905
接收 BGP 配置
910
安装配置
915
从配置中识别向其通告路由的 BGP 邻居
920
打开与识别出的邻居的 BGP 会话
925
从配置中识别要通告的路由
930
为每个识别出的邻居生成分组
935
向本地 MFE 发送生成的分组以使其经外部网络发送出到物理路由器邻居
结束

图9

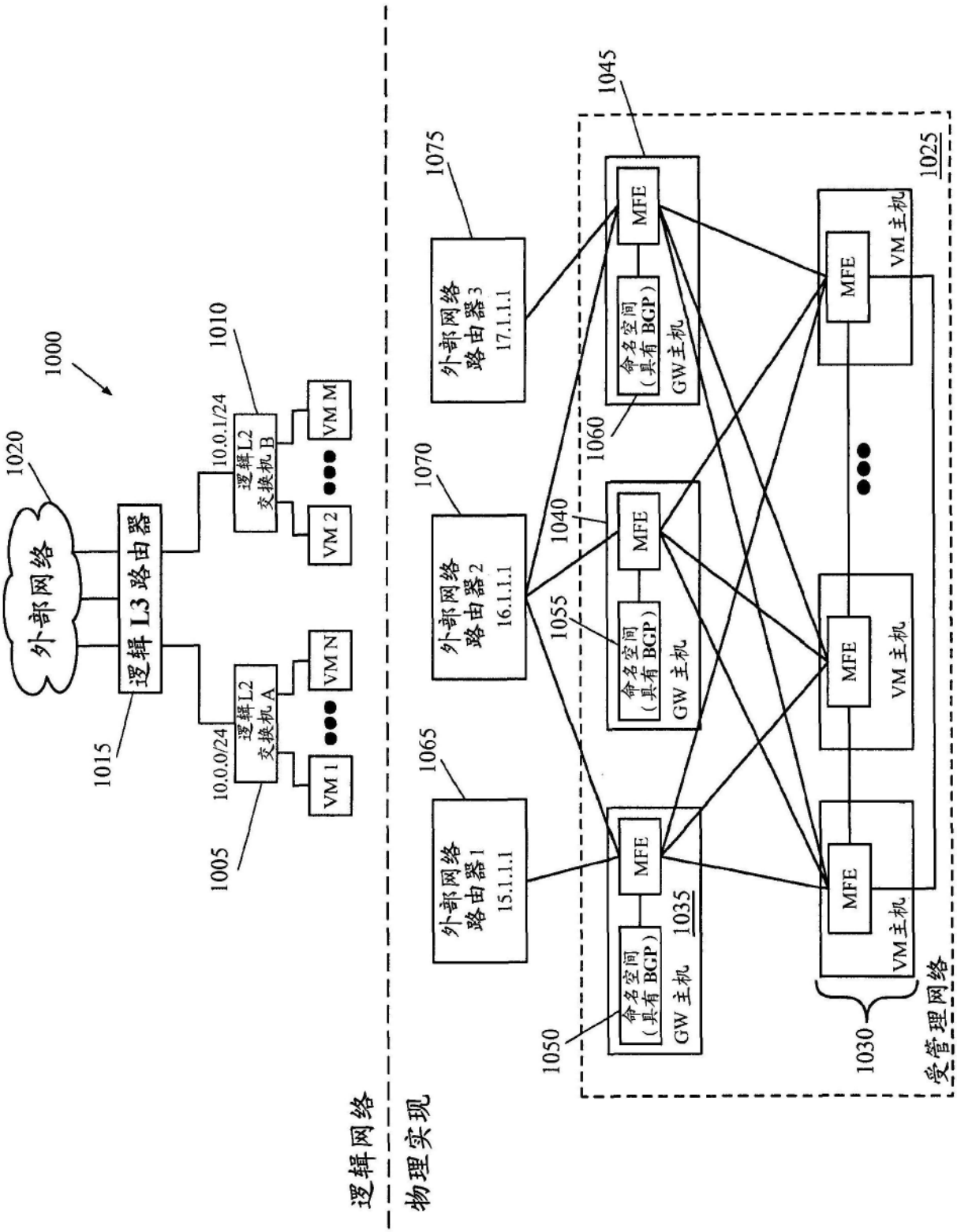

1000
1020
外部网络
逻辑 L3 路由器
1015
10.0.0/24
逻辑L2 交换机 A
VM 1
VM N
1005
10.0.1/24
1010
逻辑L2 交换机 B
VM 2
VM M
逻辑网络
物理实现
外部网络 路由器1 15.1.1.1
1065
外部网络 路由器2 16.1.1.1
1070
外部网络 路由器3 17.1.1.1
1075
1045
1025
MFE
命名空间 (具有 BGP)
GW 主机
1035
1040
1055
1060
1050
VM 主机
1030
受管理网络

图10

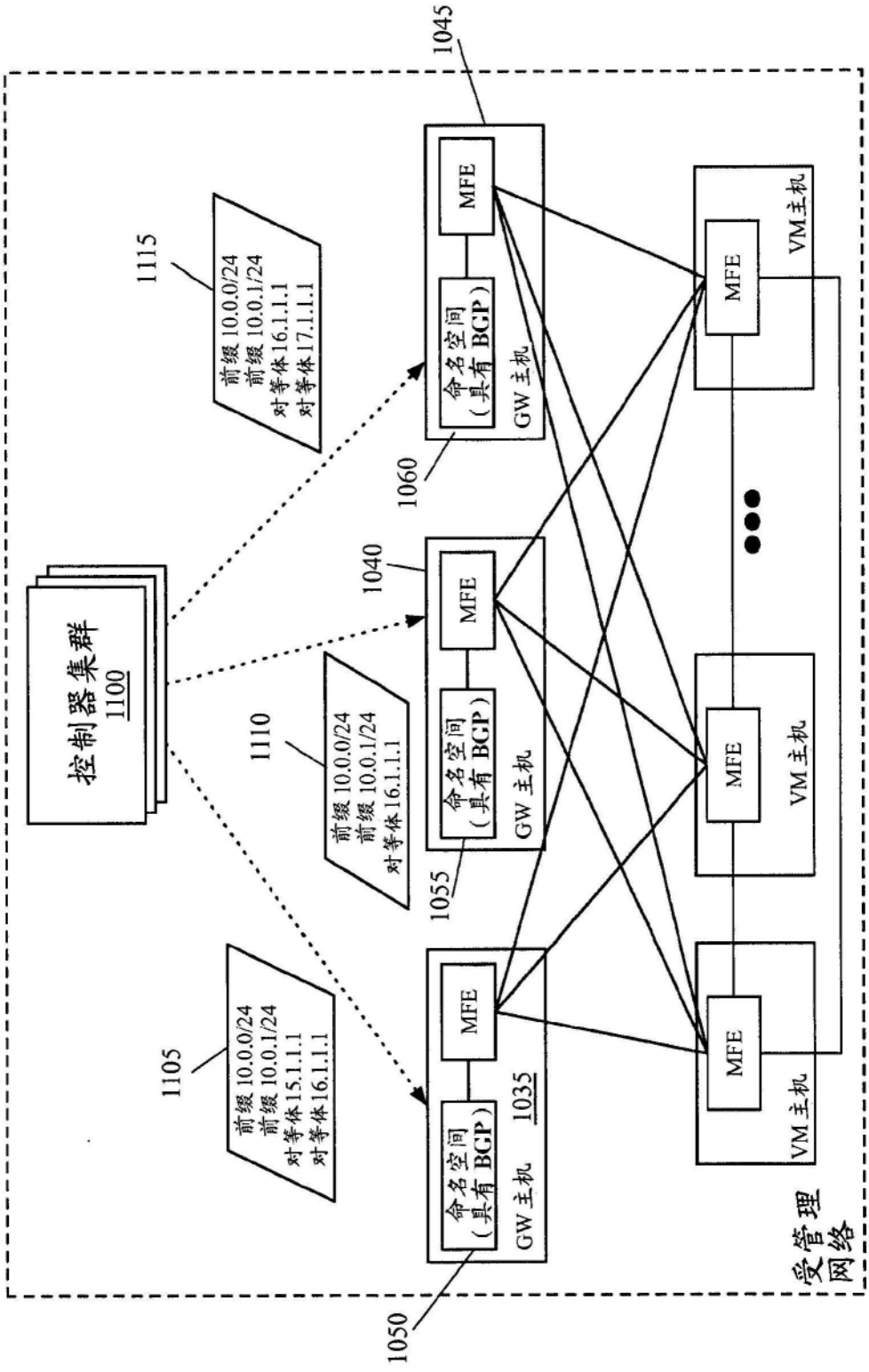

控制器集群
1100
1105
前缀 10.0.0/24
前缀 10.0.1/24
对等体15.1.1.1
对等体16.1.1.1
1110
前缀 10.0.0/24
前缀 10.0.1/24
对等体16.1.1.1
1115
前缀 10.0.0/24
前缀 10.0.1/24
对等体16.1.1.1
对等体17.1.1.1
1045
1040
1035
1060
1055
1050
MFE
命名空间
(具有 BGP)
GW 主机
VM 主机
受管理网络

图11

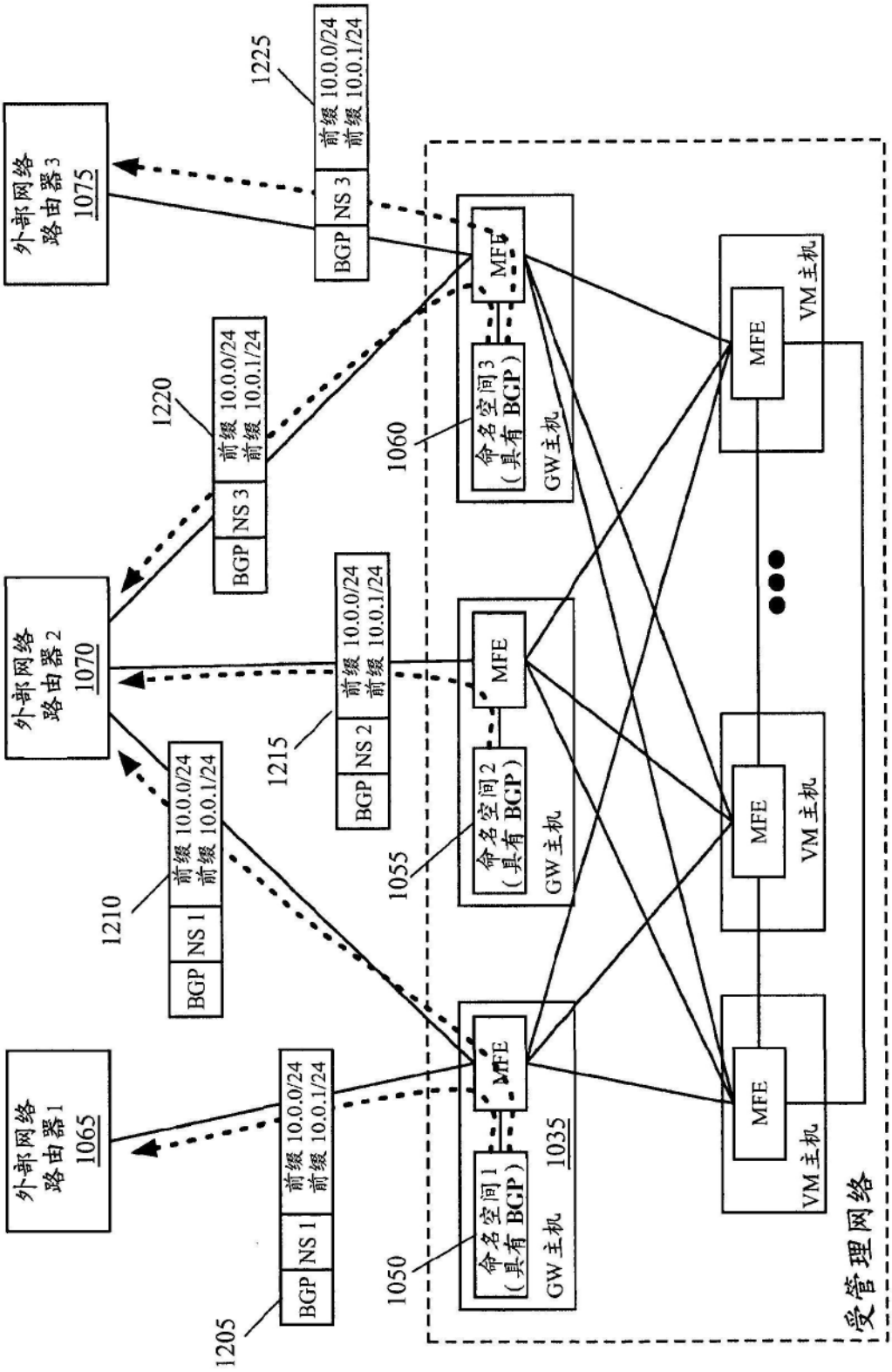

外部网络路由器3 1075
外部网络路由器2 1070
外部网络路由器1 1065
BGP NS 3 前缀 10.0.0/24 前缀 10.0.1/24
BGP NS 2 前缀 10.0.0/24 前缀 10.0.1/24
BGP NS 1 前缀 10.0.0/24 前缀 10.0.1/24
1225
1220
1215
1210
1205
MFE
命名空间3 (具有BGP)
命名空间2 (具有BGP)
命名空间1 (具有BGP)
GW主机
1060
1055
1050
1035
VM主机
受管理网络

图12

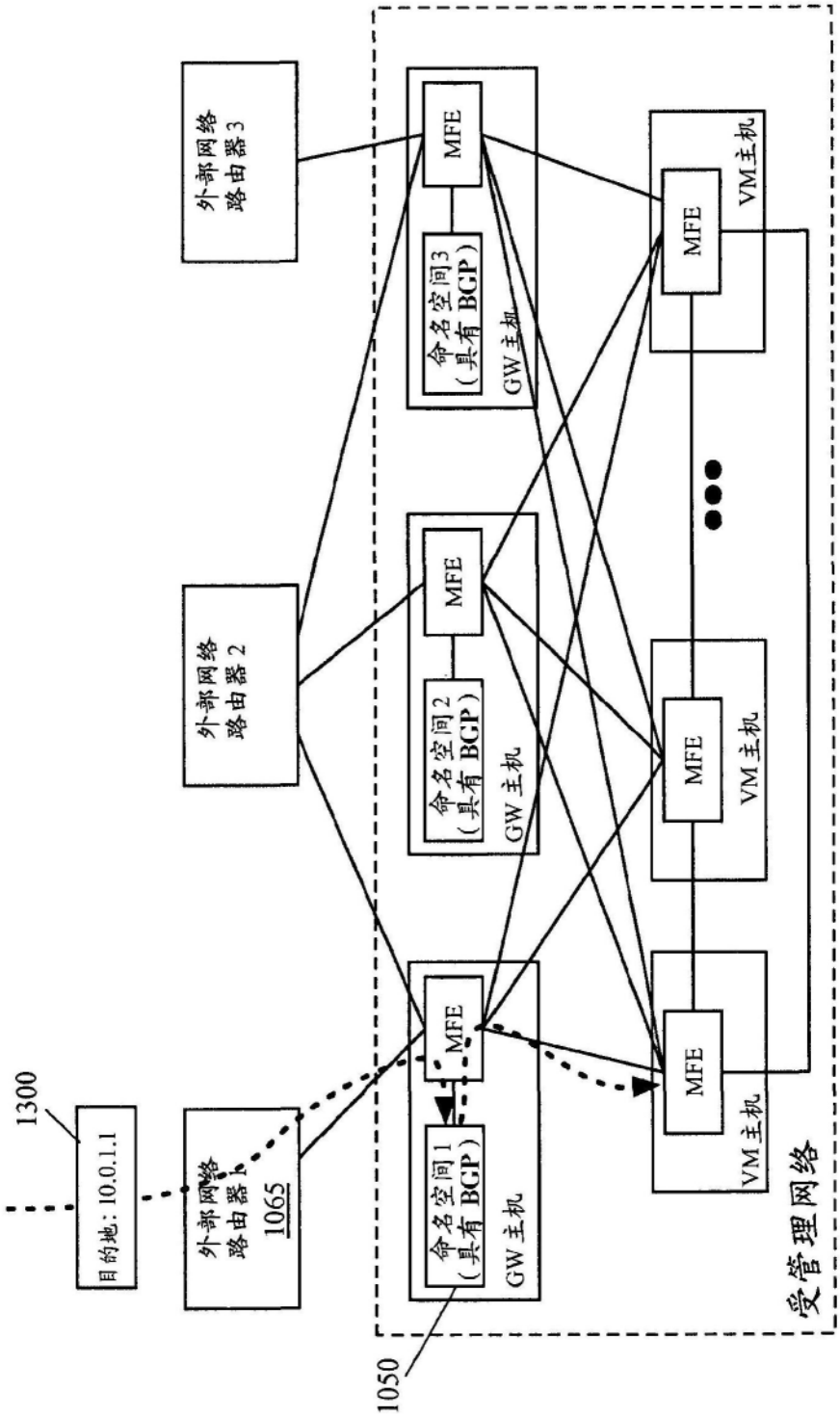

外部网络路由器3
外部网络路由器2
外部网络路由器1
1065
1300
目的地: 10.0.1.1
1050
MFE
命名空间3（具有BGP）
命名空间2（具有BGP）
命名空间1（具有BGP）
GW主机
VM主机
受管理网络

图13

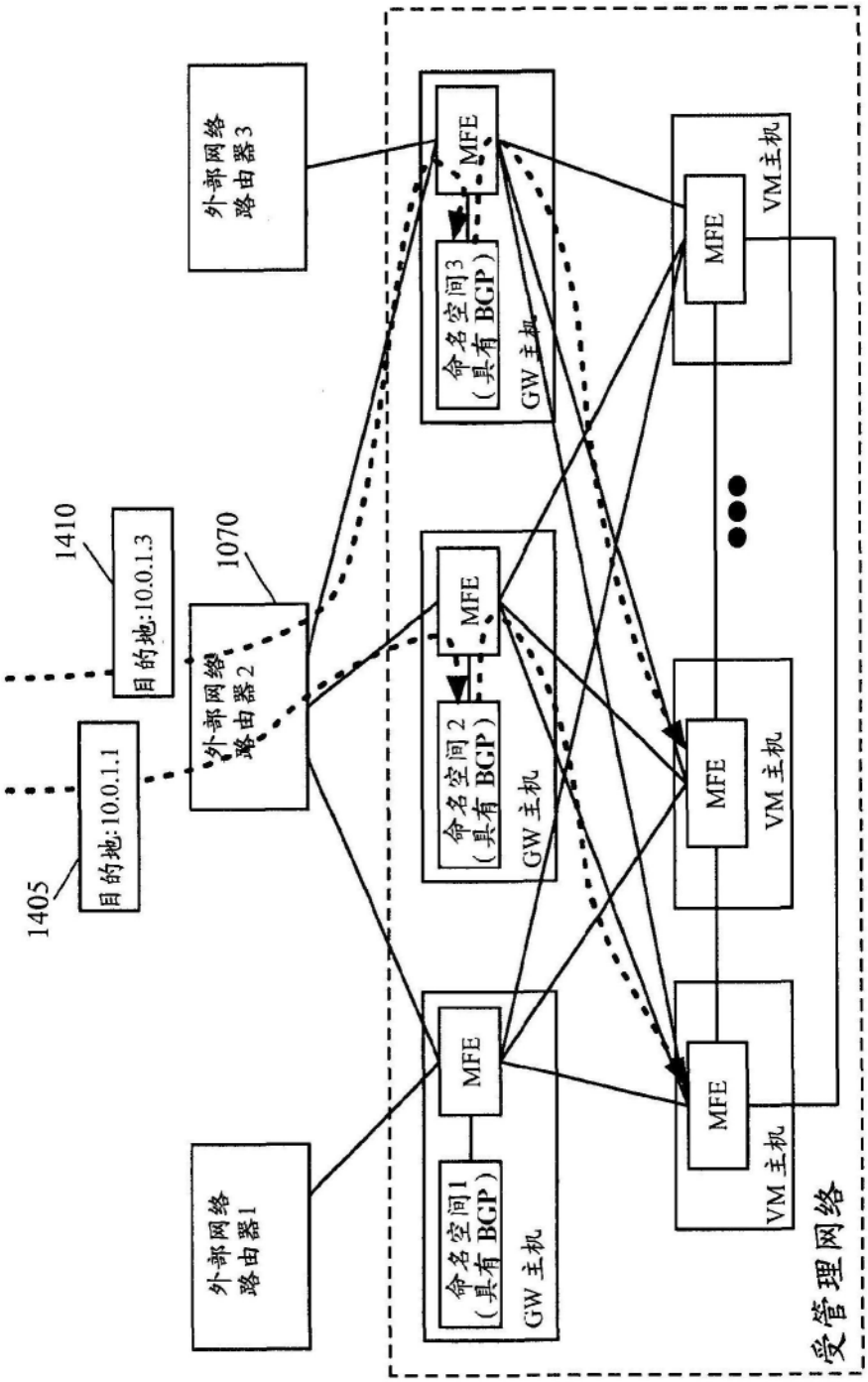

外部网络
路由器3
外部网络
路由器2
外部网络
路由器1
1410
目的地:10.0.1.3
1405
目的地:10.0.1.1
1070
MFE
命名空间3
(具有BGP)
GW主机
MFE
命名空间2
(具有BGP)
GW主机
MFE
命名空间1
(具有BGP)
GW主机
MFE
VM主机
MFE
VM主机
MFE
VM主机
受管理网络

图14

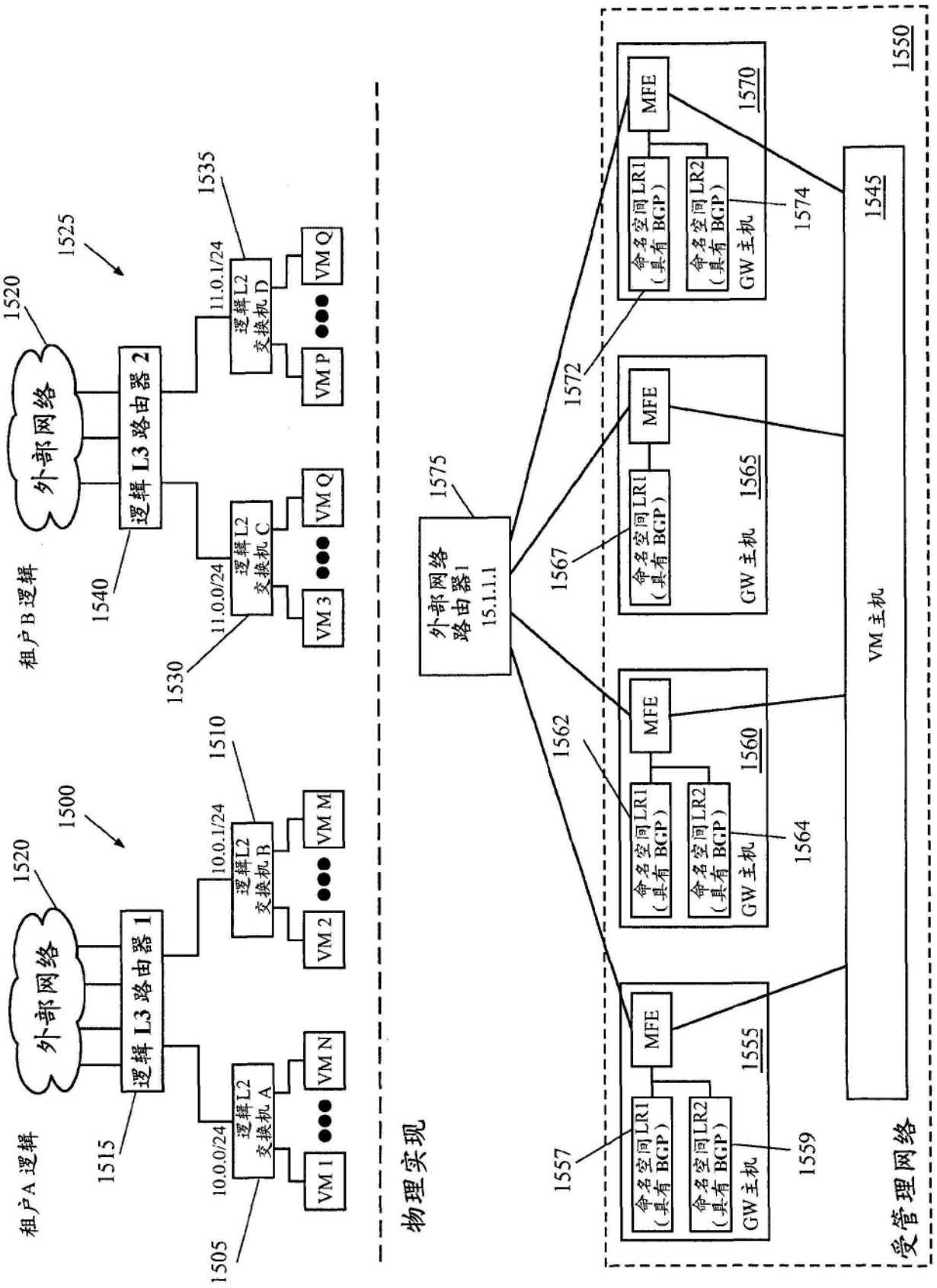

租户A逻辑
外部网络
1520
1500
逻辑L3路由器1
1515
逻辑L2交换机A
10.0.0/24
1505
VM 1
VM N
逻辑L2交换机B
10.0.1/24
1510
VM 2
VM M
租户B逻辑
外部网络
1520
1525
逻辑L3路由器2
1540
逻辑L2交换机C
11.0.0/24
1530
VM 3
VM Q
逻辑L2交换机D
11.0.1/24
1535
VM P
VM Q
物理实现
外部网络路由器1
15.1.1.1
1575
1567
1562
1557
1572
MFE
命名空间LR1（具有BGP）
命名空间LR2（具有BGP）
GW主机
1555
1559
1560
1564
1565
1570
1574
VM主机
1545
1550
受管理网络

图15

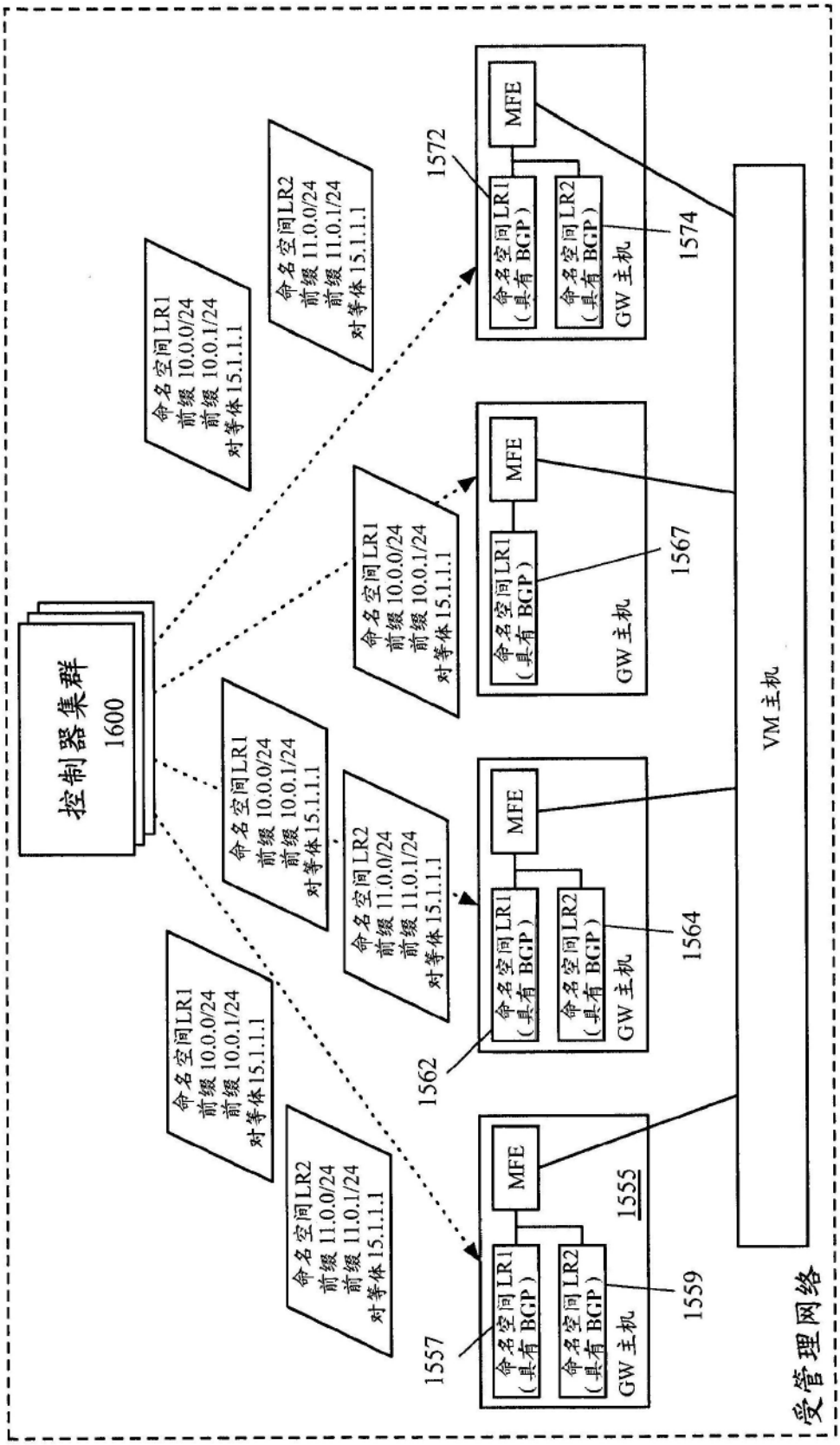

受管理网络
控制器集群
1600
命名空间LR1
前缀 10.0.0/24
前缀 10.0.1/24
对等体15.1.1.1
命名空间LR2
前缀 11.0.0/24
前缀 11.0.1/24
对等体15.1.1.1
MFE
命名空间LR1
(具有 BGP)
命名空间LR2
(具有 BGP)
GW 主机
VM 主机
1557
1559
1555
1562
1564
1567
1572
1574

图16

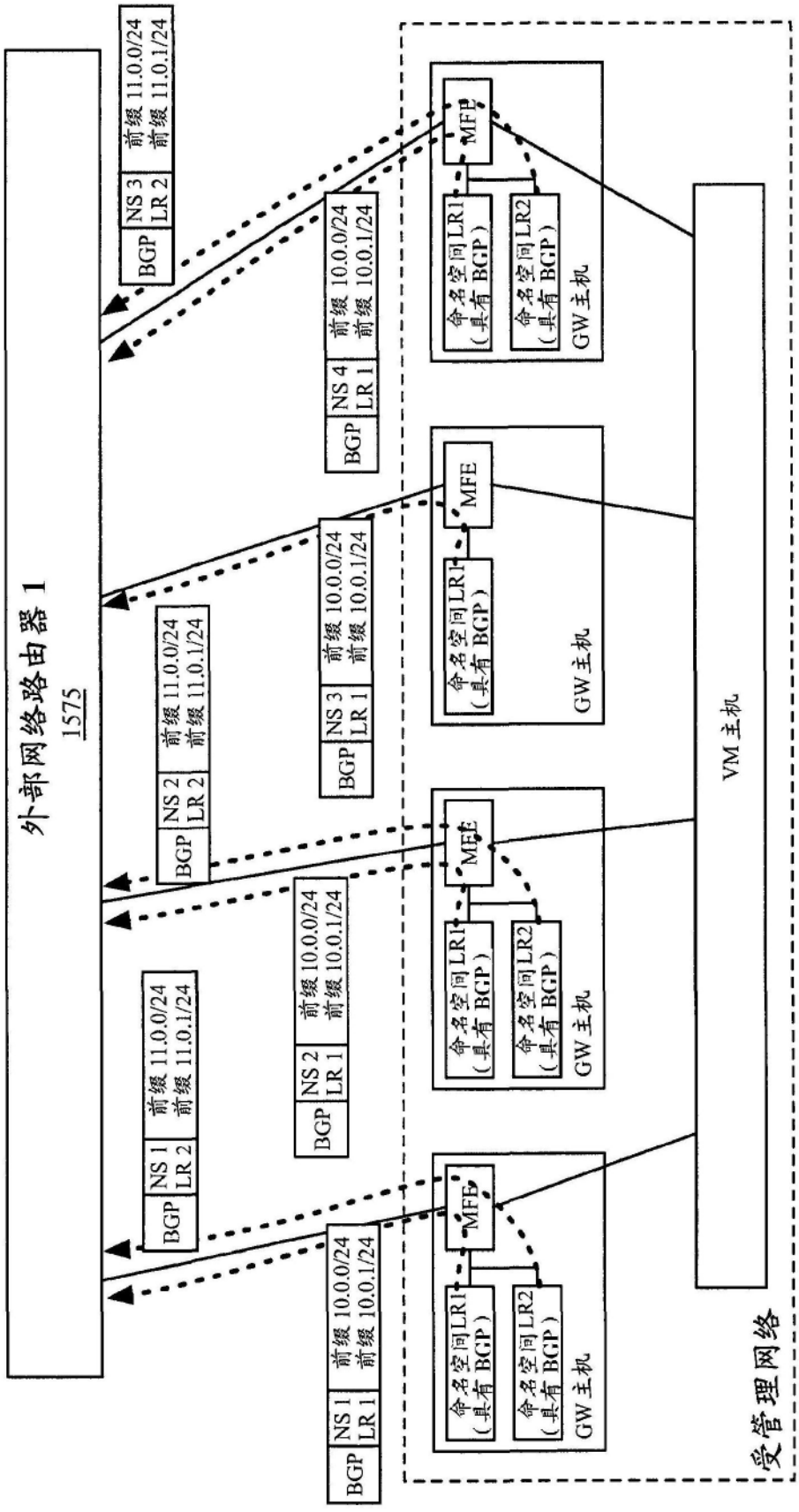

外部网络路由器 1
1575
BGP
NS 1
LR 2
前缀 11.0.0/24
前缀 11.0.1/24
BGP
NS 2
LR 2
前缀 11.0.0/24
前缀 11.0.1/24
BGP
NS 3
LR 2
前缀 11.0.0/24
前缀 11.0.1/24
BGP
NS 1
LR 1
前缀 10.0.0/24
前缀 10.0.1/24
BGP
NS 2
LR 1
前缀 10.0.0/24
前缀 10.0.1/24
BGP
NS 3
LR 1
前缀 10.0.0/24
前缀 10.0.1/24
BGP
NS 4
LR 1
前缀 10.0.0/24
前缀 10.0.1/24
MFE
命名空间LR1（具有 BGP）
命名空间LR2（具有 BGP）
GW 主机
VM 主机
受管理网络

图17

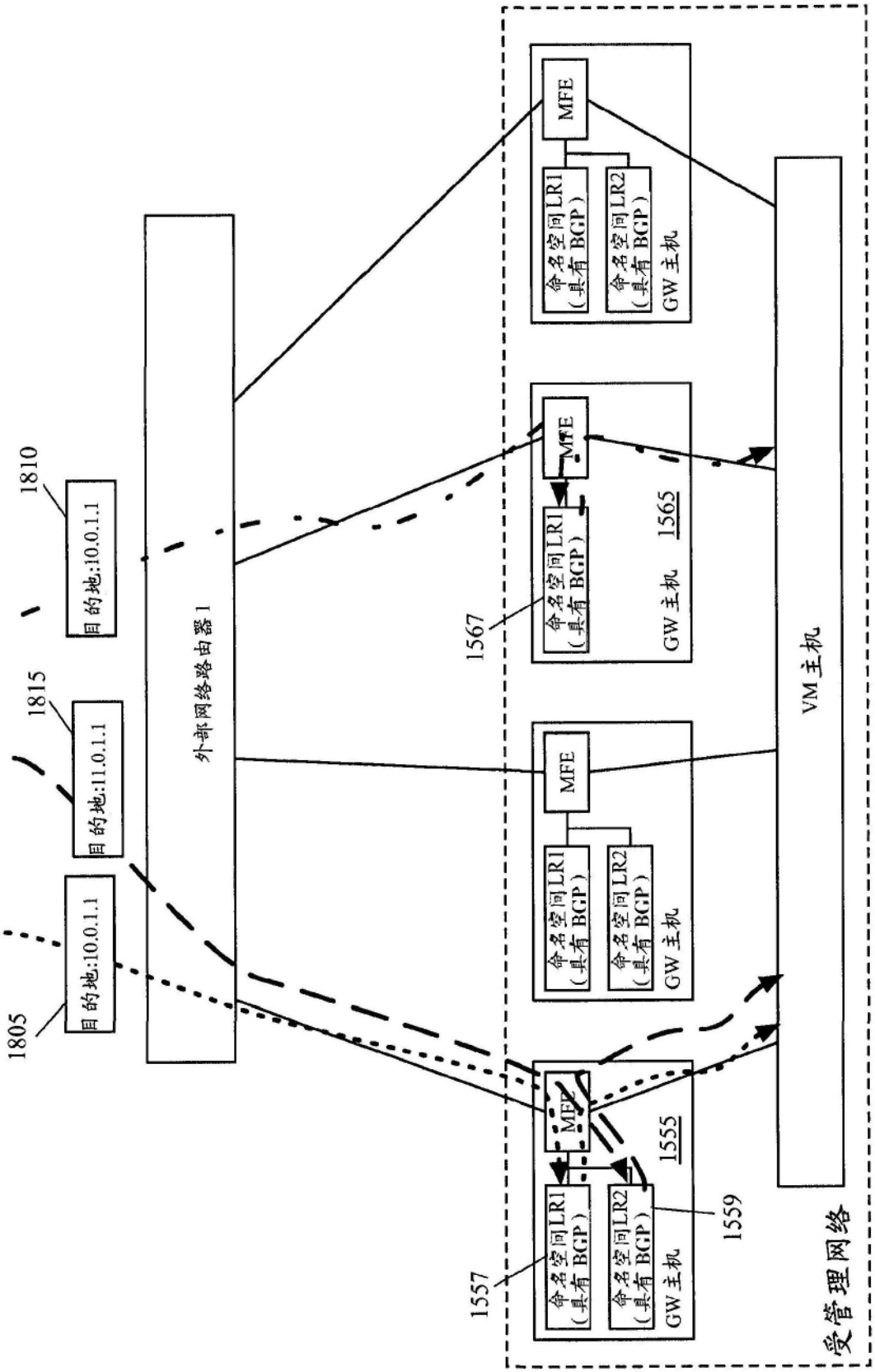

1810
目的地:10.0.1.1
1815
目的地:11.0.1.1
1805
目的地:10.0.1.1
外部网络路由器1
MFE
命名空间LR1（具有BGP）
命名空间LR2（具有BGP）
GW主机
1567
1565
1557
1555
1559
VM主机
受管理网络

图18

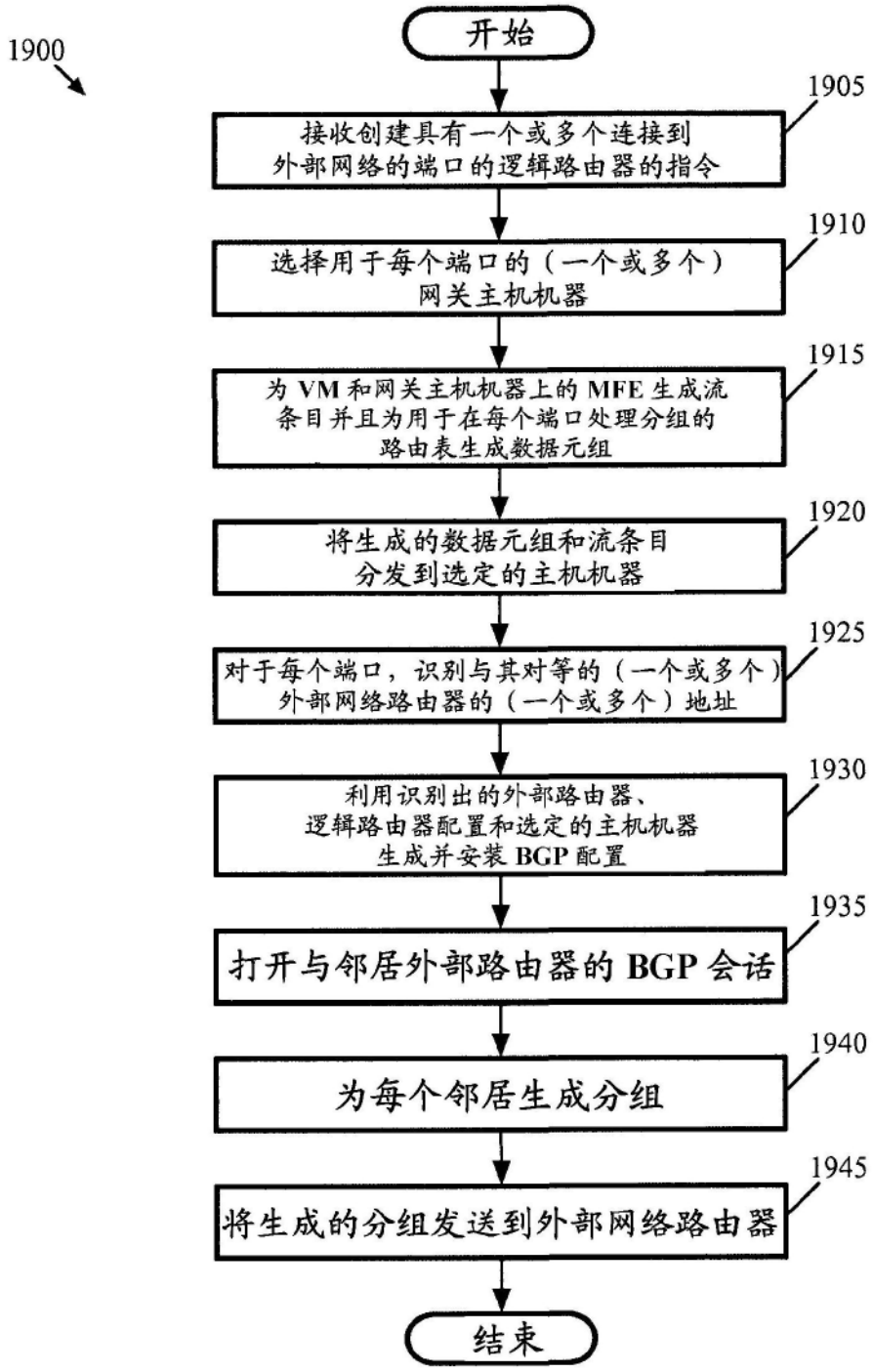

1900
开始
1905
接收创建具有一个或多个连接到外部网络的端口的逻辑路由器的指令
1910
选择用于每个端口的（一个或多个）网关主机机器
1915
为 VM 和网关主机机器上的 MFE 生成流条目并且为用于在每个端口处理分组的路由表生成数据元组
1920
将生成的数据元组和流条目分发到选定的主机机器
1925
对于每个端口，识别与其对等的（一个或多个）外部网络路由器的（一个或多个）地址
1930
利用识别出的外部路由器、逻辑路由器配置和选定的主机机器生成并安装 BGP 配置
1935
打开与邻居外部路由器的 BGP 会话
1940
为每个邻居生成分组
1945
将生成的分组发送到外部网络路由器
结束

图19

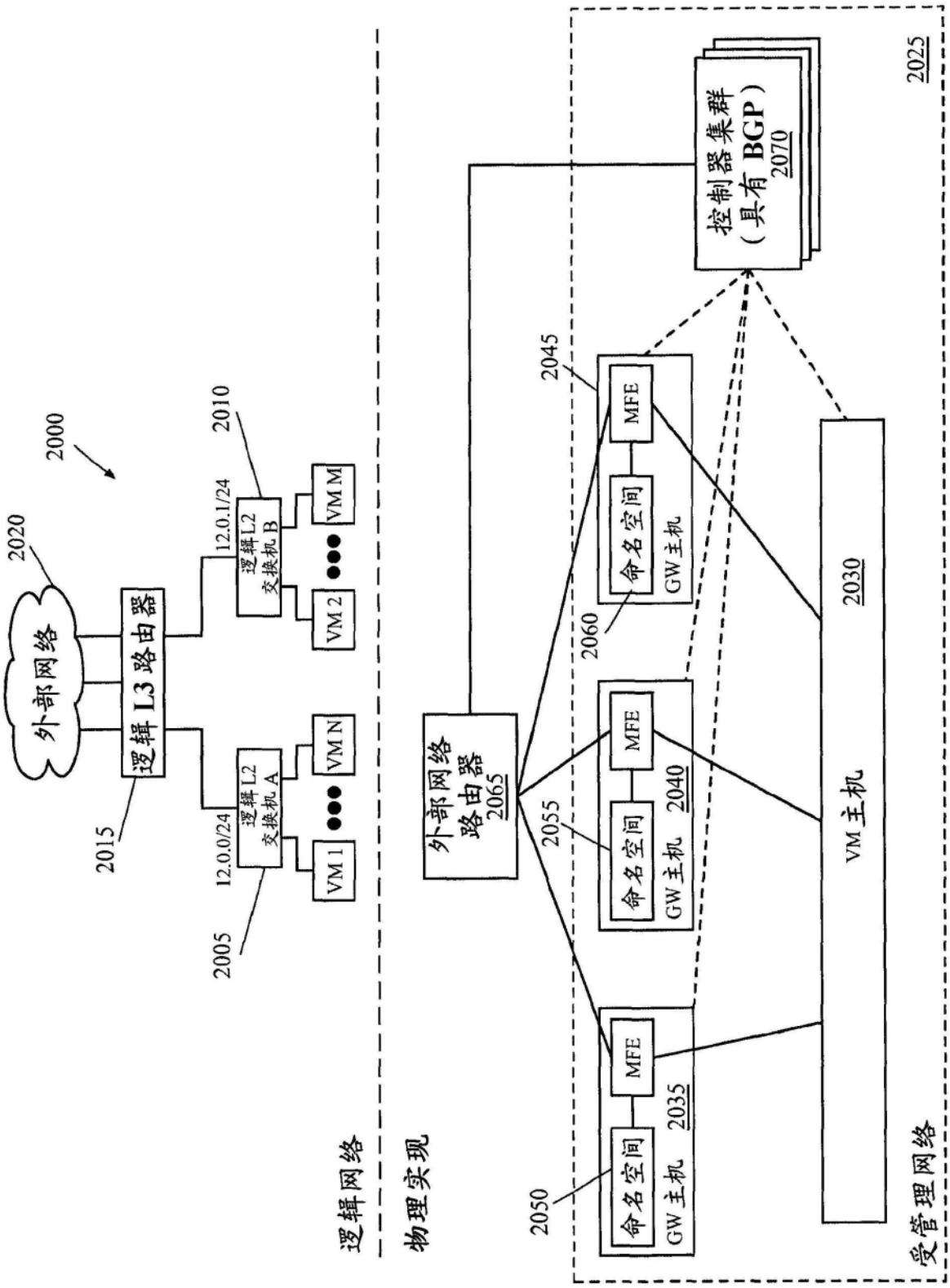

2000
2020
外部网络
逻辑 L3 路由器
2015
12.0.1/24
2010
逻辑 L2 交换机 B
VM M
VM 2
12.0.0/24
逻辑 L2 交换机 A
VM N
VM 1
2005
逻辑网络
物理实现
外部网络路由器 2065
2055
2045
MFE
命名空间
GW 主机
2060
MFE
命名空间
GW 主机 2040
MFE
命名空间
GW 主机 2035
2050
控制器集群（具有 BGP）2070
2025
VM 主机 2030
受管理网络

图20

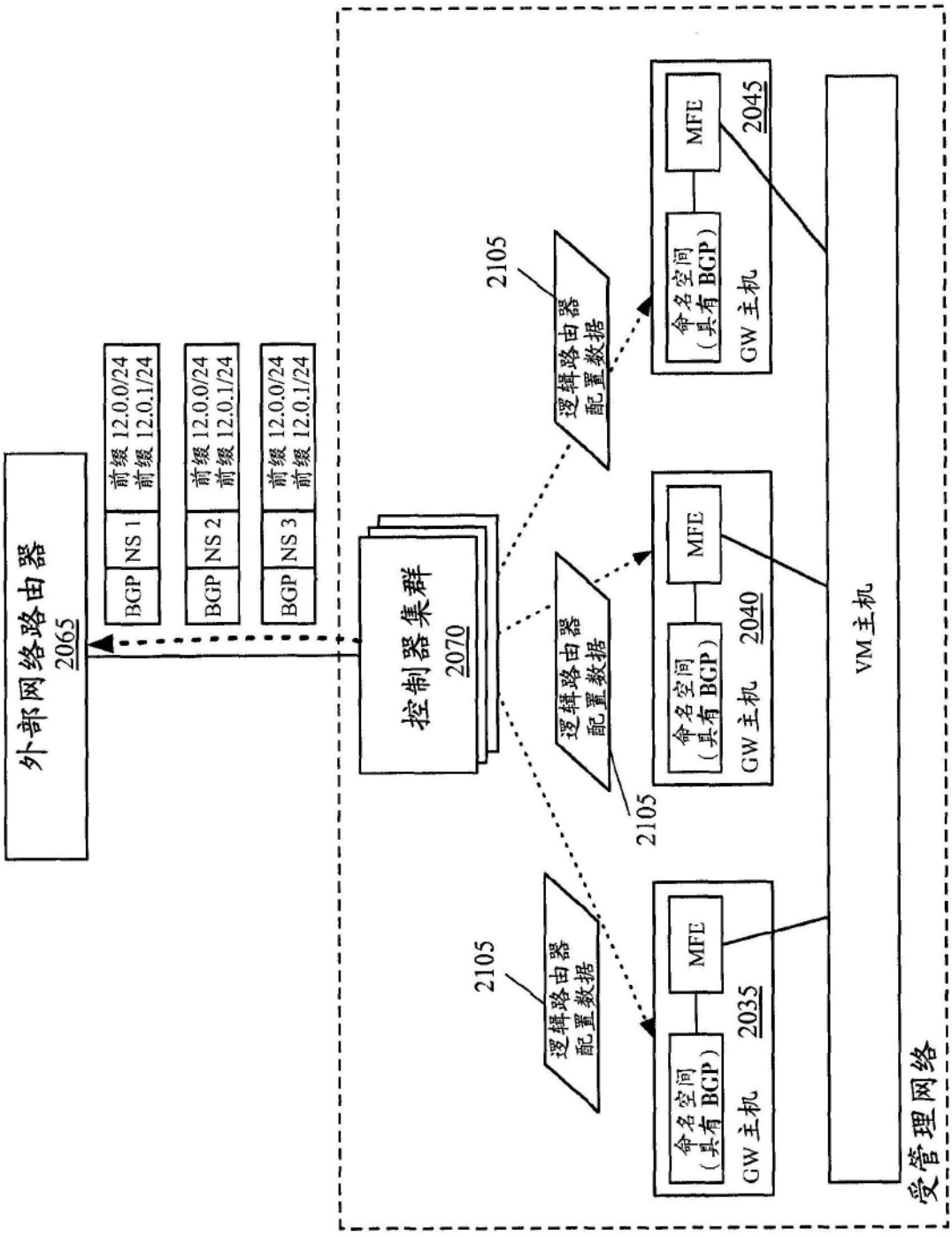

外部网络路由器
2065
BGP
NS 1
前缀 12.0.0/24
前缀 12.0.1/24
BGP
NS 2
前缀 12.0.0/24
前缀 12.0.1/24
BGP
NS 3
前缀 12.0.0/24
前缀 12.0.1/24
控制器集群
2070
2105
逻辑路由器配置数据
2105
逻辑路由器配置数据
2105
逻辑路由器配置数据
命名空间（具有 BGP）
MFE
GW 主机 2045
命名空间（具有 BGP）
MFE
GW 主机 2040
命名空间（具有 BGP）
MFE
GW 主机 2035
VM 主机
受管理网络

图21

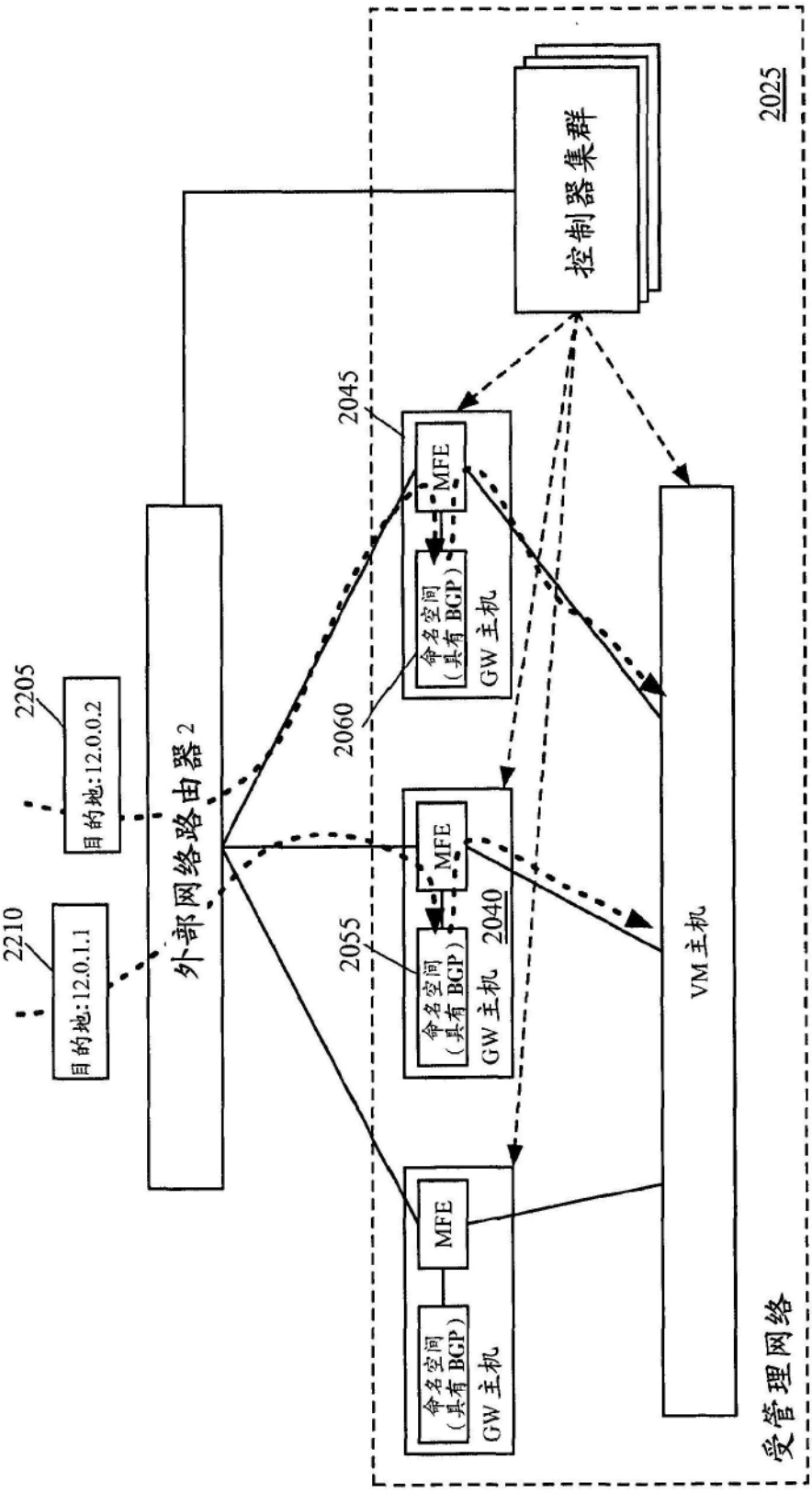

2025
控制器集群
2045
MFE
命名空间（具有BGP）
GW 主机
2060
外部网络路由器 2
2205
目的地:12.0.0.2
2210
目的地:12.0.1.1
2055
2040
VM 主机
受管理网络

图22

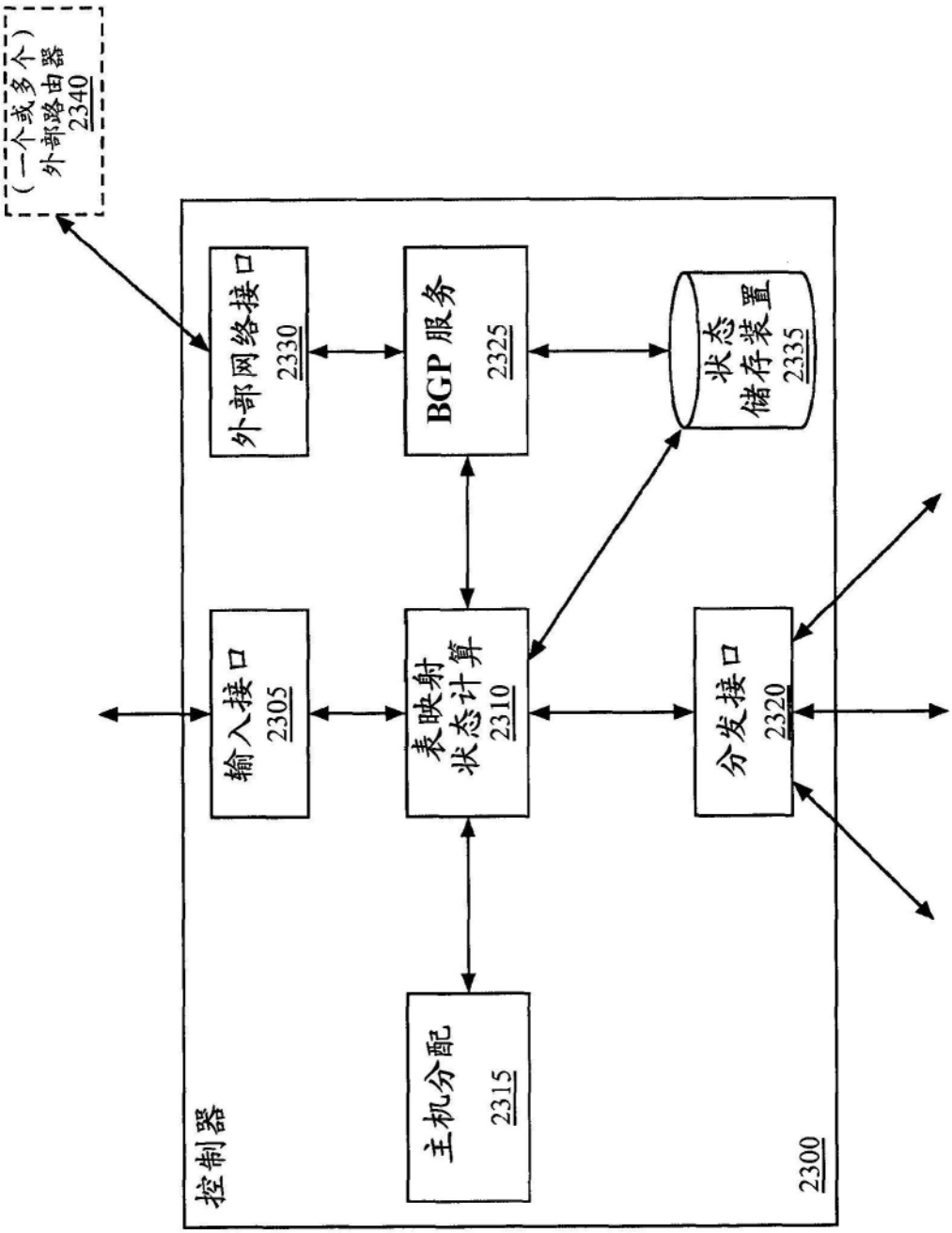

（一个或多个）
外部路由器
2340
外部网络接口
2330
BGP 服务
2325
状态
储存装置
2335
输入接口
2305
表映射
状态计算
2310
分发接口
2320
主机分配
2315
控制器
2300

图23

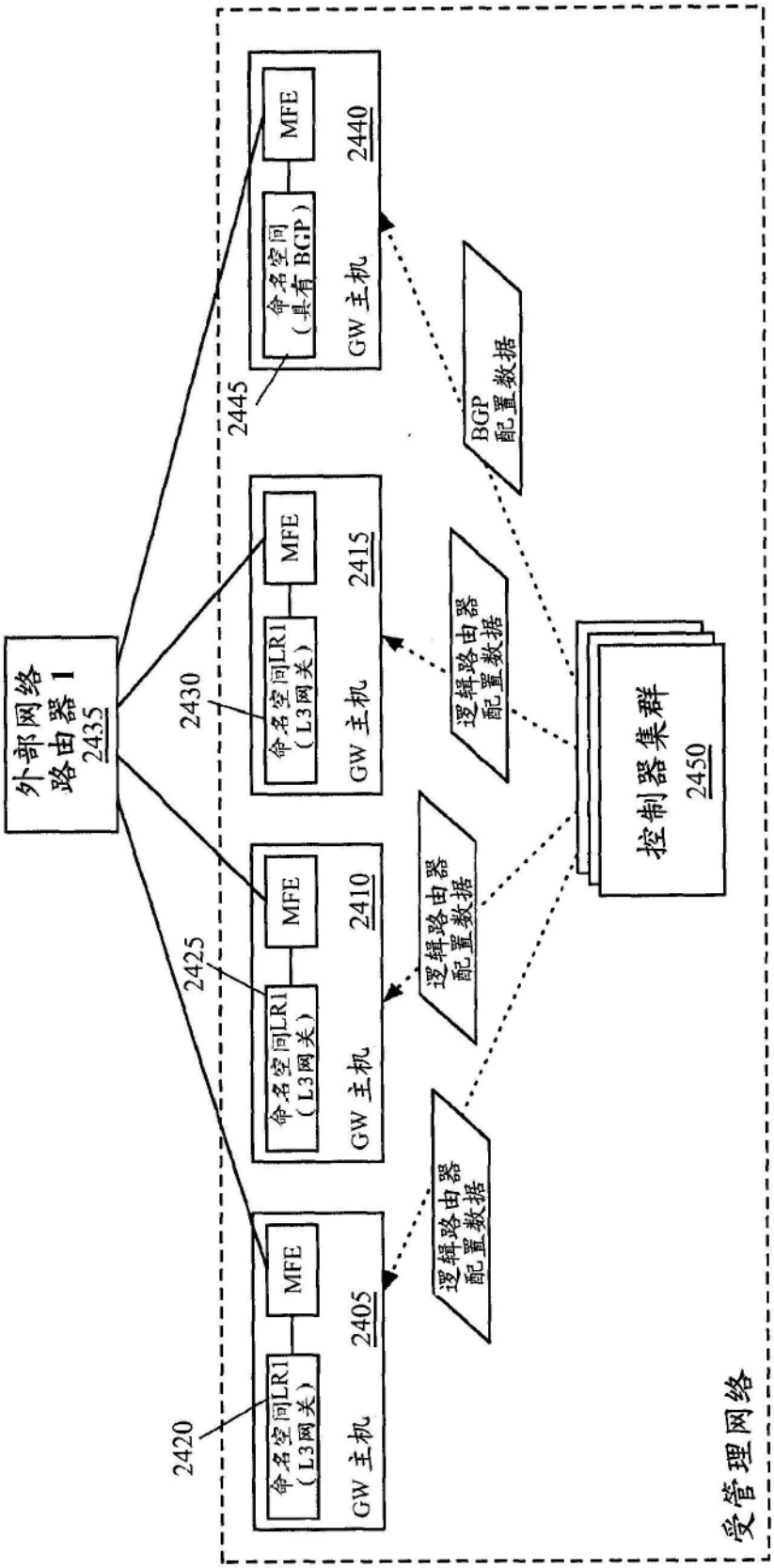

外部网络
路由器 1
2435
受管理网络
2420
2425
2430
2445
命名空间
(L3网关)
LR1
MFE
GW 主机
2405
2410
2415
2440
命名空间
(具有 BGP)
逻辑路由器
配置数据
BGP
配置数据
控制器集群
2450

图24

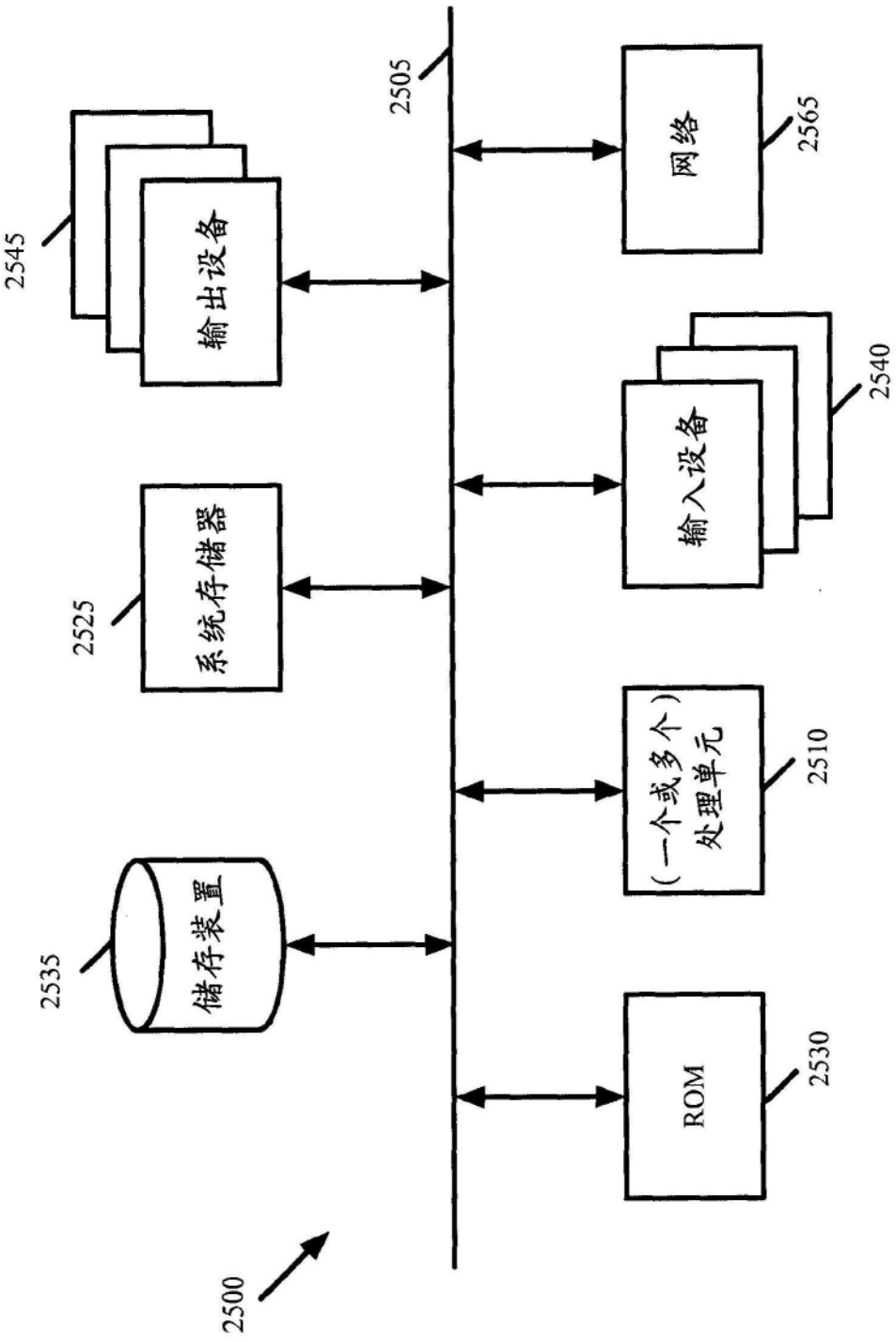

2500
2505
2510
(一个或多个)处理单元
2515
系统存储器
2520
ROM
2525
储存装置
2530
输入设备
2535
输出设备
2540
网络
2545
2565

图25