(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104272668 A
(43)申请公布日 2015.01.07

(21)申请号 201380022883.X
(22)申请日 2013.05.22
(30)优先权数据
61/650,943 2012.05.23 US
13/899,849 2013.05.22 US
(85)PCT国际申请进入国家阶段日
2014.10.30
(86)PCT国际申请的申请数据
PCT/US2013/042238 2013.05.22
(87)PCT国际申请的公布数据
WO2013/177289 EN 2013.11.28
(71)申请人 博科通讯系统有限公司
地址 美国加利福尼亚州
(72)发明人 M·P·坎彻尔拉
(74)专利代理机构 北京市金杜律师事务所 11256
代理人 王茂华
(51)Int.Cl.
H04L 12/46(2006.01)

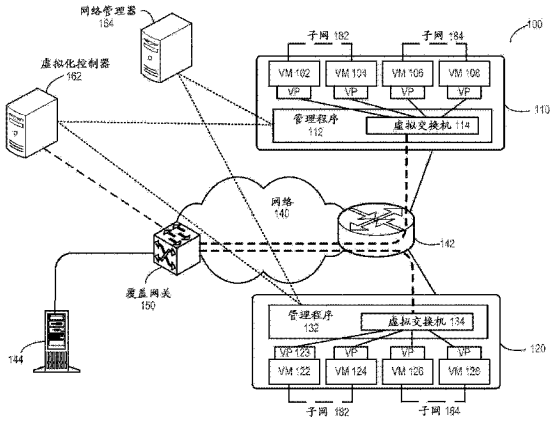
权利要求书3页 说明书11页 附图10页

(54)发明名称

层3覆盖网关

(57)摘要

本发明提供了一种计算系统，该计算系统包括处理器和用于存储指令的计算机可读存储介质。基于这些指令，该处理器操作该计算系统作为覆盖网关150，覆盖网关150与可能不具有隧穿配置的物理服务器144通信。该计算系统发起和终止与虚拟机（例如逻辑子网182上的122）相关联的覆盖隧道。子网通常对应于租户。在操作期间，覆盖网关150维护虚拟机122的IP地址与虚拟交换机134的相对应的虚拟隧道端点地址之间的隧道映射。该隧道映射也能够包括虚拟机122的MAC地址与IP地址之间的映射。该计算系统然后基于虚拟机122的IP地址来为数据分组确定输出端口。该数据分组包括内部分组，并且这个内部分组的目的地地址对应于该虚拟IP地址。在垫补数据平面层的覆盖隧穿机制可以是虚拟可扩展LAN(VXLAN)、通用路由封装(GRE)协议、使用GRE的网络虚拟化(NVGRE)协议、以及openvSwitch GRE协议之一。



CN 104272668 A

1. 一种计算系统，包括：

处理器；

存储指令的计算机可读存储介质，当所述指令由所述处理器执行时促使所述处理器执行一种方法，所述方法包括：

发起或终止与虚拟机相关联的覆盖隧道；

基于从配置系统所接收的信息，将所述虚拟机的虚拟互联网协议（IP）地址映射至被用来终止所述覆盖隧道的第二 IP 地址；以及

基于所述第二 IP 地址来为包括内部分组的数据分组确定输出端口，其中所述内部分组的目的地地址对应于所述虚拟 IP 地址。

2. 根据权利要求 1 所述的计算系统，其中所述映射进一步基于与所述虚拟 IP 地址相对应的虚拟媒体访问控制（MAC）地址。

3. 根据权利要求 1 所述的计算系统，其中所述方法进一步包括更新所述映射，所述映射将所述虚拟机的所述虚拟 IP 地址映射至第三 IP 地址，所述第三 IP 地址被用来为所述数据分组确定所述输出端口。

4. 根据权利要求 1 所述的计算系统，其中所述配置系统是以下各项中的一项或多项：

虚拟化控制器，其将所述虚拟机分配给主机中的管理程序并且将所述虚拟 IP 地址指配给所述虚拟机；

网络管理器，其向所述管理程序通知有关联网信息；以及

垫补设备，其从所述网络管理器获得联网信息。

5. 根据权利要求 4 所述的计算系统，进一步包括：垫补控制平面层，可操作为辨识多个虚拟化控制器，其中相应的虚拟化控制器对应于不同的虚拟化机制。

6. 根据权利要求 1 所述的计算系统，进一步包括：垫补数据平面层，可操作为辨识多种覆盖隧穿机制。

7. 根据权利要求 6 所述的计算系统，其中隧穿机制与以下各项中的一项或多项相关联：

虚拟可扩展局域网（VXLAN）；

通用路由封装（GRE）协议；

使用 GRE 的网络虚拟化（NVGRE）协议；以及

openvSwitch GRE 协议。

8. 根据权利要求 1 所述的计算系统，其中所述方法进一步包括：在数据分组中识别与所述计算系统和远程计算系统相关联的逻辑 IP 地址，其中所述数据分组与所述覆盖隧道相关联。

9. 根据权利要求 8 所述的计算系统，其中所述方法进一步包括：

连同所述远程计算系统来确定所述计算系统的活动状态；以及

响应于检测到所述计算系统不是活动的，阻止所述计算系统处理与所述逻辑 IP 地址相关联的分组。

10. 根据权利要求 9 所述的计算系统，其中所述方法进一步包括：

检测所述远程计算系统的失效；以及

响应于检测到所述失效，处理与所述逻辑 IP 地址相关联的分组。

11. 根据权利要求 8 所述的计算系统，其中所述方法进一步包括：

识别与所述计算系统和远程计算系统相关联的隧道终止 IP 地址，其中所述数据分组与所述覆盖隧道相关联；以及

其中所述隧道终止 IP 地址属于与所述逻辑 IP 地址所属的子网不同的子网。

12. 一种方法，包括：

由计算系统发起或终止与虚拟机相关联的覆盖隧道；

基于从配置系统所接收的信息，将所述虚拟机的虚拟互联网协议 (IP) 地址映射至被用来终止所述覆盖隧道的第二 IP 地址；以及

基于所述第二 IP 地址来为包括内部分组的数据分组确定输出端口，其中所述内部分组的目的地地址对应于所述虚拟 IP 地址。

13. 根据权利要求 12 所述的方法，其中所述映射进一步基于与所述虚拟 IP 地址相对应的虚拟媒体访问控制 (MAC) 地址。

14. 根据权利要求 12 所述的方法，进一步包括更新所述映射，所述映射将所述虚拟机的所述虚拟 IP 地址映射至第三 IP 地址，所述第三 IP 地址被用来为所述数据分组确定所述输出端口。

15. 根据权利要求 12 所述的方法，其中所述配置系统是以下各项中的一项或多项：

虚拟化控制器，其将所述虚拟机分配给主机中的管理程序并且将所述虚拟 IP 地址指配给所述虚拟机；

网络管理器，其向所述管理程序通知有关联网信息；以及

垫补设备，其从所述网络管理器获得联网信息。

16. 根据权利要求 15 所述的方法，进一步包括：辨识多个虚拟化控制器，其中相应的虚拟化控制器对应于不同的虚拟化机制。

17. 根据权利要求 12 所述的方法，进一步包括：辨识多种覆盖隧穿机制。

18. 根据权利要求 17 所述的方法，其中隧穿机制与以下各项中的一项或多项相关联：

虚拟可扩展局域网 (VXLAN)；

通用路由封装 (GRE) 协议；

使用 GRE 的网络虚拟化 (NVGRE) 协议；以及

openvSwitch GRE 协议。

19. 根据权利要求 12 所述的方法，进一步包括：在数据分组中识别与所述计算系统和远程计算系统相关联的逻辑 IP 地址，其中所述数据分组与所述覆盖隧道相关联。

20. 根据权利要求 19 所述的方法，进一步包括：

连同所述远程计算系统来确定所述计算系统的活动状态；以及

响应于检测到所述计算系统不是活动的，阻止所述计算系统处理与所述逻辑 IP 地址相关联的分组。

21. 根据权利要求 20 所述的方法，进一步包括：

检测所述远程计算系统的失效；以及

响应于检测到所述失效，处理与所述逻辑 IP 地址相关联的分组。

22. 根据权利要求 19 所述的方法，进一步包括：

识别与所述计算系统和远程计算系统相关联的隧道终止 IP 地址，其中所述数据分组

与所述覆盖隧道相关联；以及

其中所述隧道终止 IP 地址属于与所述逻辑 IP 地址所属的子网不同的子网。

23. 一种计算装置，包括：

隧穿装置，用于发起或终止与虚拟机相关联的覆盖隧道；

映射装置，用于基于从配置系统所接收的信息而将所述虚拟机的虚拟互联网协议 (IP) 地址映射至被用来终止所述覆盖隧道的第二 IP 地址；以及

转发装置，用于基于所述第二 IP 地址来为包括内部分组的数据分组确定输出端口，其中所述内部分组的目的地地址对应于所述虚拟 IP 地址。

层 3 覆盖网关

技术领域

[0001] 本公开内容涉及网络管理。更具体地说，本公开内容涉及网络中的层 3 覆盖。

背景技术

[0002] 互联网的指数增长已经使它成为用于运行在物理和虚拟设备上的各种各样的应用的一种普及的递送媒介。这些应用随之带来了对带宽的与日俱增的需求。作为结果，装备供应商争相构建具有多功能能力（诸如虚拟机迁移的意识）的更大且更快的交换机，以高效地移动更多流量。然而，交换机的尺寸不能无限地增长。提出几个因素地说，它受限于物理空间、功耗、以及设计复杂性。此外，具有更高能力的交换机通常更加复杂和昂贵。更重要的是，因为过于庞大和复杂的系统通常不提供规模经济，所以归因于增加的每端口成本，简单地增加交换机的尺寸和能力可能证明经济上是不可行的。

[0003] 随着互联网流量变得更加多样化，网络中的虚拟计算作为用于网络架构的有价值的命题而逐渐变得更加重要。虚拟计算的演进已经对网络设置了附加的要求。然而，常规的层 2 网络架构通常不能容易地适应虚拟机的动态性质。例如，在常规的数据中心架构中，主机能够通过形成层 2 广播域的层 2（例如，以太网）互连而被互连。因为层 2 广播域的物理范围限制，所以数据中心通常被分段成不同的层 2 广播域。结果，去往层 2 广播域之外的任何通信通过层 3 网络来承载。因为虚拟机的位置变得更加移动和动态，并且来自虚拟机的数据通信变得更加多样化，所以通常合意的是，网络基础设施能够提供层 3 网络覆盖隧道以辅助数据通信跨越层 2 广播域。

[0004] 尽管覆盖给网络带来许多合意的特征，但是在提供跨越层 2 广播域的逻辑子网中的一些问题仍然未解决。

发明内容

[0005] 本发明的一个实施例提供了一种计算系统。该计算系统包括处理器和用于存储指令的计算机可读存储介质。基于这些指令，该处理器操作该计算系统作为覆盖网关。该计算系统发起和终止与虚拟机相关联的覆盖隧道。在操作期间，该计算系统将该虚拟机的虚拟互联网协议 (IP) 地址映射至第二 IP 地址，该第二 IP 地址被用来基于从配置系统所接收的信息而终止该覆盖隧道。该计算系统然后基于该第二 IP 地址来为数据分组确定输出端口。该数据分组包括内部分组，并且这个内部分组的目的地址对应于该虚拟 IP 地址。

[0006] 在对这个实施例的一种变型中，该映射还基于与该虚拟 IP 地址相对应的虚拟媒体访问控制 (MAC) 地址。

[0007] 在对这个实施例的一种变型中，该计算系统通过将该虚拟机的该虚拟 IP 地址映射至第三 IP 地址来更新该映射，该第三 IP 地址被用来为该数据分组确定输出端口。

[0008] 在对这个实施例的一种变型中，该配置系统是以下各项中的一项或多项：虚拟化控制器、网络管理器、以及垫补设备 (shim device)。该虚拟化控制器将该虚拟机分配给主机中的管理程序并且将这些虚拟 IP 地址指配给该虚拟机。该网络管理器向该管理程序通

知有关联网信息。该垫补设备从该网络管理器获得联网信息。

[0009] 在一种进一步的变型中，该计算系统还包括垫补控制平面层，该垫补控制平面层识别多个虚拟化控制器。相应的虚拟化控制器能够对应于不同的虚拟化机制。

[0010] 在对这个实施例的一种变型中，该计算系统进一步包括垫补数据平面层，该垫补数据平面层识别多个覆盖隧穿机制。

[0011] 在一种进一步的变型中，隧穿机制与以下各项中的一项或多项相关联：虚拟可扩展局域网 (VXLAN)、通用路由封装 (GRE) 协议、使用 GRE 的网络虚拟化 (NVGRE) 协议、以及 openvSwitch GRE 协议。

[0012] 在对这个实施例的一种变型中，该计算系统在数据分组中识别与该计算系统和远程计算系统相关联的逻辑 IP 地址，其中该数据分组与该覆盖隧道相关联。

[0013] 在一种进一步的变型中，该计算系统连同该远程计算系统来确定该计算系统的活动状态。如果该计算系统不是活动的，则该处理器阻止该计算系统处理与该逻辑 IP 地址相关联的分组。

[0014] 在一种进一步的变型中，该计算系统检测到该远程计算系统的失效。一经检测到该失效，该计算系统就开始处理与该逻辑 IP 地址相关联的分组。

[0015] 在一种进一步的变型中，该计算系统识别与该计算系统和远程计算系统相关联的隧道终止 IP 地址，其中该数据分组与该覆盖隧道相关联。这个隧道终止 IP 地址属于与该逻辑 IP 地址所属的子网不同的子网。

附图说明

[0016] 图 1A 图示了根据本发明的一个实施例的具有覆盖网关的示例性虚拟化的网络环境。

[0017] 图 1B 图示了根据本发明的一个实施例的具有辅助覆盖网关的垫补设备的示例性虚拟化的网络环境。

[0018] 图 2 图示了根据本发明的一个实施例的支持多个控制接口和隧穿机制的示例性覆盖网关。

[0019] 图 3 图示了根据本发明的一个实施例的用于常规分组的示例性头部格式以及由覆盖网关提供的它的隧道封装。

[0020] 图 4A 呈现了根据本发明的一个实施例的流程图，该流程图图示了覆盖网关从虚拟化控制器获得隧道映射的过程。

[0021] 图 4B 呈现了根据本发明的一个实施例的流程图，该流程图图示了覆盖网关转发所接收的分组的过程。

[0022] 图 4C 呈现了根据本发明的一个实施例的流程图，该流程图图示了覆盖网关在逻辑子网中转发广播分组、未知单播分组、或者多播分组的过程。

[0023] 图 5A 图示了根据本发明的一个实施例的具有高可用性的示例性覆盖网关。

[0024] 图 5B 图示了根据本发明的一个实施例的对具有高可用性的覆盖网关的多个地址的示例性使用。

[0025] 图 6 图示了根据本发明的一个实施例的作为覆盖网关操作的示例性计算系统。

[0026] 在这些图中，相似的参考标号指代相同的图元素。

具体实施方式

[0027] 以下描述被提出以使得本领域的技术人员能够制造和使用本发明，并且在特定的应用及其要求的背景中被提供。对所公开的各实施例的各种修改对本领域的技术人员将容易是明显的，并且不背离本发明的精神和范围，本文中定义的一般原理可以被应用至其他实施例和应用。因此，本发明不局限于所示出的这些实施例，而是将符合与权利要求相一致的最宽范围。

[0028] 概述

[0029] 在本发明的各实施例中，促进超出物理子网边界的逻辑子网络（子网）的问题通过并入如下的覆盖网关来解决，该覆盖网关提供物理子网之间的虚拟隧穿以形成该逻辑子网。这个逻辑子网逻辑地耦合了属于该逻辑子网但是位于属于不同物理子网的主机中的虚拟机。以这种方式，网络的物理基础设施通常被虚拟化来适应多租赁。网络虚拟化中的挑战之一是将物理网络拓扑与虚拟化的网络子网桥接。

[0030] 例如，数据中心能够包括与客户（或租户）相关联的虚拟机，这些虚拟机运行在位于不同物理主机上的管理程序上。这些虚拟机能够是同一逻辑子网的部分。数据中心的虚拟化控制器通常将相应的虚拟机分配给主机中的管理程序，并且将媒体访问控制 (MAC) 地址和互联网协议 (IP) 地址指配给该虚拟机。通常，管理程序或虚拟交换机使用层 3 虚拟隧穿，以允许属于同一逻辑子网的虚拟机通信。这些管理程序或虚拟交换机能够被称为虚拟隧道端点 (VTEP)。然而，如果虚拟机中的一个虚拟机的主机不具有等同的隧穿配置，则其他虚拟机可能不能经由虚拟隧道而通信。

[0031] 为了解决这个问题，覆盖网关促进了通向相应主机的相应 VTEP（例如，管理程序或虚拟交换机）的虚拟隧穿。该覆盖网关进而与不支持相同隧穿机制的目的地（诸如物理服务器）通信。然而，为了将该虚拟隧道与虚拟机相关联，覆盖网关需要识别用于该虚拟机的 VTEP。为了促进这种识别，覆盖网关维护虚拟机的 MAC 地址与相对应的 VTEP 地址之间的隧道映射。注意，该隧道映射还能够包括 MAC 地址与该虚拟机的 IP 地址之间的映射。

[0032] 在一些实施例中，覆盖网关与虚拟化控制器通信并且获得用于相应虚拟机的隧道映射。无论何时该映射被更新，覆盖网关从虚拟化控制器获得经更新的映射。在一些实施例中，覆盖网关能够包括两个“垫补层”。垫补层操作为两个设备之间的通信接口。一个垫补层操作为控制平面并且与虚拟化控制器对接，以用于获得该映射。另一个垫补层操作为数据平面并且促使去往和来自虚拟机的分组的隧道封装。作为结果，同一覆盖网关能够支持包括多种虚拟化和隧穿机制的多个覆盖网络。

[0033] 在一些实施例中，数据中心中的互连包括以太网结构交换机。在一种以太网结构交换机中，耦合在任意拓扑中的任何数量的交换机可以逻辑地操作为单个交换机。任何新的交换机可以无需任何手动配置而以“即插即用”模式加入或离开该结构交换机。结构交换机对外部设备表现为单个逻辑交换机。在一些进一步的实施例中，该结构交换机是多链路透明互连 (TRILL) 网络并且该结构交换机的相应成员交换机是 TRILL 路由桥 (RBridge)。

[0034] 术语“外部设备”能够指代 VTEP 不能向其直接建立隧道的任何设备。外部设备能够是主机、服务器、常规的层 2 交换机、层 3 路由器、或者任何其他类型的物理或虚拟设备。另外，外部设备能够耦合至更远离网络的其他交换机或者主机。外部设备也能够是用于多

个网络设备进入网络的聚集点。术语“设备”和“机器”可以互换地被使用。

[0035] 术语“管理程序”在通用的意义上被使用，并且能够指代任何虚拟机管理器。创建并且运行虚拟机的任何软件、固件、或硬件能够是“管理程序”。术语“虚拟机”也在通用的意义上被使用，并且能够指代机器或设备的软件实施方式。类似于物理设备的能够执行软件程序的任何虚拟设备能够是“虚拟机”。管理程序在其上运行了一个或多个虚拟机的主机外部设备能够被称为“主机”。

[0036] 术语“隧道”指代一种数据通信，其中使用另一种联网协议来封装一个或多个联网协议。尽管使用基于层 2 协议的层 3 封装的示例来提出本公开内容，但是“隧道”不应当被解释为将本发明的实施例限制于层 2 和层 3 协议。“隧道”能够被建立用于任何网络层、子层、或者网络层的组合。

[0037] 术语“分组”指代能够跨网络一起被传输的一组比特。“分组”不应当被解释为将本发明的实施例限制于层 3 网络。“分组”能够由指代一组比特的其他术语取代，诸如“帧”、“信元”、或“数据报”。

[0038] 术语“交换机”在通用的意义上被使用，并且它能够指代操作在任何网络层中的任何独立交换机或结构交换机。“交换机”不应当被解释为将本发明的实施例限制于层 2 网络。能够将流量转发给外部设备或另一个交换机的任何设备能够被称为“交换机”。“交换机”的示例包括但不限于：层 2 交换机、层 3 路由器、TRILL RBridge、或者包括多个类似的或异构的更小物理交换机的结构交换机。

[0039] 术语“RBridge”指代路由桥，路由桥是实施 TRILL 协议的桥，TRILL 协议如描述在 http://tools.ietf.org/html/rfc6325 处可得到的互联网工程任务组 (IETF) 请求注解 (RFC)“Routing Bridges(RBridge):Base Protocol Specification”中，其通过参考并入本文。本发明的各实施例不限于 RBridges 之中的应用。其他类型的交换机、路由器、以及转发器也能够被使用。

[0040] 术语“交换机标识符”指代能够被用来标识交换机的一组比特。如果交换机是 RBridge，则交换机标识符能够是“RBridge 标识符”。TRILL 标准使用“RBridge ID”来表示被指配给 RBridge 的 48 比特的中间系统到中间系统 (IS-IS) ID，并且使用“RBridge 别名”来表示用作用于该“RBridge ID”的缩写的 16 比特值。在本公开内容中，“交换机标识符”被用作通用术语，不限于任何比特格式，并且能够指代能够标识交换机的任何格式。术语“RBridge 标识符”在通用的意义上被使用，不限于任何比特格式，并且能够指代“RBridge ID”、“RBridge 别名”、或者能够标识 RBridge 的任何其他格式。

[0041] 网络架构

[0042] 图 1A 图示了根据本发明的一个实施例的具有覆盖网关的示例性虚拟化的网络环境。如图 1A 中所图示的，虚拟化的网络环境 100（其能够在数据中心中）包括多个主机 110 和 120，该多个主机 110 和 120 经由一跳或多跳而耦合至网络 140 中的层 3 路由器 142。多个虚拟机 102、104、106 和 108 运行在主机 110 中的管理程序 112 上。相应的虚拟机具有虚拟端口 (VP，或虚拟网络接口卡，VNIC)。运行在管理程序 112 上的相应虚拟机的虚拟端口被逻辑地耦合至虚拟交换机 114，虚拟交换机 114 由管理程序 112 提供。虚拟交换机 114 负责分派虚拟机 102、104、106 和 108 的传出和传入的流量。类似地，多个虚拟机 122、124、126 和 128 运行在主机 120 中的管理程序 132 上。运行在管理程序 132 上的相应虚拟机的虚拟

端口被逻辑地耦合至虚拟交换机 134，虚拟交换机 134 由管理程序 132 提供。在逻辑上，虚拟交换机 114 和 134 用作聚集点并且经由一条或多条链路来耦合路由器 142。

[0043] 还被包括的是虚拟化控制器 162 和网络管理器 164。虚拟化控制器 162，通常基于来自网络管理员的指令，将相应的虚拟机分配给主机中的管理程序，并且将虚拟 MAC 地址和 IP 地址指配给该虚拟机。例如，虚拟化控制器 162 将虚拟机 122 分配给主机 120 中的管理程序 132，并且将虚拟 MAC 地址和 IP 地址指配给虚拟机 122 的虚拟端口 123。由虚拟机 122 生成的以太网帧具有虚拟端口 123 的虚拟 MAC 作为它的源地址。在这个示例中，虚拟机 110 和 120 是网络 140 中的两个不同物理子网的部分。然而，主机 110 中的虚拟机 102 和 104 以及主机 120 中的虚拟机 122 和 124 是逻辑子网 182 的部分。类似地，主机 110 中的虚拟机 106 和 108 以及主机 120 中的虚拟机 126 和 128 是同一逻辑子网 184 的部分。通常，逻辑子网对应于租户。

[0044] 在一些实施例中，虚拟交换机 114 和 134 被逻辑地耦合至网络管理器 164，网络管理器 164 向虚拟交换机 114 和 134 提供彼此通信所需要的联网信息。例如，因为虚拟机 102 和 122 是同一逻辑子网的部分，所以虚拟机 102 能够经由层 2 而与虚拟机 122 通信。然而，这些虚拟机位于不同物理子网中的主机上。因此，虚拟交换机 114 需要知道虚拟机 122 被逻辑地耦合至虚拟交换机 134（例如，虚拟交换机 134 是用于虚拟机 122 的 VTEP）。通过提供这个联网信息，网络管理器 164 使得虚拟交换机 114 和 134 能够分别操作为用于虚拟机 102 和 122 的 VTEP，并且使用层 3 虚拟隧穿来促进这些虚拟机间的通信。然而，因为外部设备（诸如物理服务器 144）可能不具有等同的隧穿配置，所以虚拟机（诸如虚拟机 122）可能不能经由虚拟隧道而与服务器 144 通信。

[0045] 为了与服务器 144 通信，覆盖网关 150 允许相应的 VTEP 经由网络 140 而建立虚拟隧穿。覆盖网关 150 进而与物理服务器 144 通信。在操作期间，虚拟机 122 经由逻辑耦合的虚拟交换机 134 向虚拟服务器 144 发送分组。虚拟交换机 134 将该分组封装在隧道头部中并且将经封装的分组转发给网关 150。一经接收到该经封装的分组，覆盖网关 150 就去除隧道封装，并且基于该分组的目的地地址而将该分组转发给服务器 144。当服务器 144 将分组发回给虚拟机 122 时，覆盖网关 150 接收该分组。然而，为了高效地将这个分组转发给虚拟机 122，覆盖网关 150 需要识别虚拟机 122 被逻辑地耦合至其的虚拟交换机（例如，VTEP）。为了促进该识别，覆盖网关 150 维护虚拟机 122 的 MAC 地址与虚拟交换机 134 的对应 VTEP 地址之间的隧道映射。注意，该隧道映射还能够包括虚拟机 122 的 MAC 地址与 IP 地址之间的映射。

[0046] 例如，覆盖网关 150 能够通过发送用以获得对应 VTEP 地址的带有虚拟机 122 的 IP 地址的广播（例如，地址解析协议 (ARP)）查询，来获得用于虚拟机 122 的这种映射。然而，在具有大量虚拟机的大数据中心中，发送大量广播查询可能是低效的。在一些实施例中，覆盖网关 150 与虚拟化控制器 162 通信并且获得用于相应虚拟机的隧道映射。对于虚拟机 122，这种映射能够包括对主机 120 的标识符（例如，主机 120 的物理网络接口的 MAC 地址）、虚拟端口 123 的 MAC 地址、以及虚拟交换机 134 的对应 VTEP 地址。如果该映射在虚拟化控制器 162 中被更新（例如，由于虚拟机迁移），则覆盖网关 150 从虚拟化控制器 162 获得经更新的隧道映射。

[0047] 基于所获得的隧道映射，覆盖网关 150 将虚拟交换机 134 识别为用于虚拟机 122

的 VTEP，将来自服务器 144 的分组封装在隧道头部中，并且将经封装的分组转发给虚拟交换机 134。一经接收到该经封装的分组，虚拟交换机 134 就去除封装并且将该分组提供给虚拟机 122。假设虚拟化控制器 162 将虚拟机 122 迁移到主机 110。结果，用于虚拟机 122 的隧道映射在虚拟化控制器 162 中被更新。用于虚拟机 122 的经更新的映射包括对主机 110 的标识符和虚拟交换机 114 的对应 VTEP 地址。覆盖网关 150 能够从虚拟化控制 162 接收包括经更新的隧道映射的更新消息。

[0048] 在一些实施例中，覆盖网关 150 能够从网络管理器 164 获得隧道映射。根据本发明的一个实施例，连同图 1A 中的示例，图 1B 图示了具有辅助覆盖网关的垫补设备的示例性虚拟化的网络环境。网络管理器 164 向相应的虚拟交换机提供彼此通信所需要的联网信息。为了从网络管理器 164 获得信息，虚拟化的网络环境 100 包括垫补设备 172，垫补设备 172 运行虚拟交换机 174。这个虚拟交换机 174 被逻辑地耦合至网络管理器 164，网络管理器 164 将虚拟交换机 174 考虑为管理程序中的另一个虚拟交换机。结果，网络管理器 164 向虚拟交换机 174 提供与逻辑地耦合至其他虚拟交换机的虚拟机通信所需要的联网信息。对于虚拟机 122，这种映射能够包括对主机 120 的标识符（例如，主机 120 的物理网络接口的 MAC 地址）、虚拟端口 123 的 MAC 地址、以及虚拟交换机 134 的对应 VTEP 地址。

[0049] 垫补设备 172 能够包括与覆盖网关 150 通信的垫补层 176。覆盖网关 150 经由垫补层 176 来获得联网信息并且构建隧道映射。注意，联网信息可能不包括虚拟机的虚拟 MAC 地址。在这样的场景下，覆盖网关 150 使用广播查询来获得对应的虚拟 MAC 地址，这些广播查询使用虚拟机的虚拟 IP 地址。在一些实施例中，垫补层 176 能够位于网络管理器 164（以虚线表示）上并且将联网信息提供给覆盖 150，由此旁路垫补设备 172。然而，将垫补层 176 与网络管理器 164 集成，在物理硬件中产生了附加的存储器和处理需求，并且可能降低网络管理器 164 的性能。

[0050] 图 2 图示了根据本发明的一个实施例的支持多个控制接口和隧穿机制的示例性覆盖网关。覆盖网关 150 能够包括两个垫补层。一个垫补层操作为控制平面 220 并且与虚拟化控制器 162 对接，用于获得映射。另一个垫补层操作为数据平面 210 并且促进对去往和来自虚拟机的分组的隧道封装。作为结果，覆盖网关 150 能够支持包括多种虚拟化和隧穿机制的多个覆盖网络。

[0051] 控制平面 220 包括多个控制接口 222、224 和 226。相应的控制接口能够与不同的虚拟化管理器通信。控制接口的示例包括但不限于：用于 VMWareNSX 的接口、用于微软系统中心的接口、以及用于 OpenStack 的接口。例如，控制接口 222 能够与 OpenStack 通信，而控制接口 224 能够与微软系统中心通信。数据平面 210 支持多个隧穿机制 212、214 和 216。相应的隧穿机制能够通过促进对应的隧道封装（即，操作为用于不同隧穿机制的 VTEP）来建立不同的覆盖隧道。隧穿机制的示例包括但不限于：虚拟可扩展局域网（VXLAN）、通用路由封装（GRE）、以及它的变型，诸如使用 GRE 的网络虚拟化（NVGRE）和 openvSwitch GRE。例如，隧穿机制 212 能够表示 VXLAN，而隧穿机制 214 能够表示 GRE。

[0052] 具有对不同接口和隧穿机制的支持，如果数据中心包括来自不同销售商的多个虚拟化的网络环境，则同一覆盖网关 150 能够服务这些环境。在图 1A 中的示例中，如果虚拟交换机 114 支持 VXLAN 而虚拟交换机 134 支持 GRE，则网关 150 能够分别使用隧穿机制 212 和 214 来分别向虚拟交换机 114 和 134 提供经隧道封装的覆盖。如果虚拟化控制器 162 运

行 OpenStack，则覆盖网关 150 能够使用接口 222 来获得隧道映射。类似地，如果虚拟化控制器 162 是微软系统中心，则覆盖网关 150 能够使用接口 224 来获得隧道映射。

[0053] 分组格式

[0054] 图 3 图示了根据本发明的一个实施例的用于常规分组的示例性头部格式以及由覆盖网关提供的隧道封装。在这个示例中，常规的以太网分组 300 通常包括有效载荷 308 和以太网头部 310。通常，有效载荷 308 能够包括 IP 分组，该 IP 分组包括 IP 头部 320。IP 头部 320 包括 IP 目的地地址 (DA)312 和 IP 源地址 (SA)314。以太网头部 310 包括 MAC DA 302、MAC SA 304、以及可选地虚拟局域网 (VLAN) 标签 306。

[0055] 假设分组 300 是图 1A 中的从服务器 144 到虚拟机 122 的分组。在一个实施例中，覆盖网关 150 基于隧道映射而将常规分组 300 封装成经封装的分组 350。经封装的分组 350 通常包括封装头部 360，封装头部 360 对应于一种封装机制，如连同图 2 所描述的。封装头部 360 包含封装 DA 352（其对应于虚拟交换机 134 的 VTEP IP 地址）以及封装 SA 354（其对应于覆盖网关 150 的 IP 地址）。在图 1A 中的示例中，经封装的分组 350 基于封装 DA 352 经由网络 140 而被转发。在一些实施例中，封装头部 360 还包括租户标识符 356，租户标识符 356 唯一地标识虚拟化的网络环境 100 中的租户。例如，如果封装头部 360 对应于用于虚拟机 122 的隧道，则租户标识符 356 标识虚拟机 122 所属的租户。以这种方式，网关 150 能够通过对于不同租户使用用于分组的分离隧道封装来维持租户隔离。

[0056] 通常，服务器 144 中的上层应用使用虚拟机 122 的虚拟 IP 地址作为 IP DA 地址 312，并且使用服务器 144 的物理 IP 地址作为 IP SA 地址 314，而生成去往虚拟机 122 的 IP 分组。这个 IP 分组成为有效载荷 308。服务器 144 中的层 2 然后生成以太网头部 310 来封装有效载荷 308。如果服务器 144 和虚拟机 122 位于同一逻辑子网内，则以太网头部 310 的 MAC DA 302 被指配虚拟机 122 的 MAC 地址。以太网头部 310 的 MAC SA 304 是服务器 144 的 MAC 地址。服务器 144 然后经由覆盖网关 150 将以太网分组 300 发送给虚拟机 122。

[0057] 当覆盖网关 150 从服务器 144 接收到以太网分组 300 时，覆盖网关 150 检验以太网头部 310，并且可选地检验 IP 头部 308 和它的有效载荷（例如，层 4 头部）。基于这个信息，覆盖网关 150 确定以太网分组 300 去往同一逻辑子网内的虚拟机 122。随后，覆盖网关 150 组装封装头部 360（对应于一种封装机制）。封装头部 360 的封装 DA 352 被指配虚拟交换机 134 的 VTEP IP 地址的 IP 地址。封装头部 360 的封装 SA 354 是覆盖网关 150 的 IP 地址。注意，覆盖网关 150 的 IP 地址也能够是逻辑 IP 地址。覆盖网关 150 然后附上租户标识符 356 并且将经封装的分组 350 转发给 VTEP 虚拟交换机 134。一经接收到分组 350，虚拟交换机 134 就去除封装头部 360，查验经解除封装的分组 300 中的以太网头部 310，并且将经解除封装的分组 300 提供给虚拟机 122。

[0058] 操作

[0059] 在图 1A 中的示例中，覆盖网关 150 与虚拟化控制器 162 通信以获得隧道映射，并且基于所获得的隧道映射而经由隧道封装来转发所接收的分组。图 4A 呈现了根据本发明的一个实施例的流程图，该流程图图示了覆盖网关从虚拟化控制器获得隧道映射的过程。如连同图 2 所描述的，在操作期间，覆盖网关识别虚拟化控制器（操作 402），并且识别与所识别的虚拟化控制器相对应的本地控制接口（操作 404）。覆盖网关然后经由所识别的控制接口从虚拟化控制器请求信息（操作 406）。作为响应，虚拟控制器发送包括相关隧道映射

的信息消息。

[0060] 覆盖网关接收这个信息消息（操作 408）并且从该信息消息中提取隧道映射（操作 410）。这个隧道映射将相应虚拟机的 MAC 地址映射至对应的 VTEP 地址。注意，隧道映射也能够包括虚拟机的 MAC 地址与 IP 地址之间的映射。覆盖网关然后本地存储所提取的隧道映射（操作 412）。覆盖网关也能够从虚拟化控制器获得用于相应虚拟机的租户信息（操作 414），并且将租户与对应的虚拟机相关联（操作 416）。在一些实施例中，覆盖网关能够获得租户信息作为隧道映射的一部分。

[0061] 图 4B 呈现了根据本发明的一个实施例的流程图，该流程图图示了覆盖网关转发所接收的分组的过程。如连同图 3 所描述的，一经接收到分组（操作 452），覆盖网关就检查该分组是否被封装用于本地 VTEP（即，去往与网关相关联的 VTEP）（操作 454）。如果该分组被封装并且去往与网关相关联的 VTEP，则覆盖网关使用 VTEP IP 地址来将该分组的隧道封装解除封装（操作 456）。如果由覆盖网关接收的分组未被封装（操作 454）或者封装已经利用覆盖网关 VTEP IP 地址而被解除封装（操作 456），则网关检查该分组的目的地是否经由隧道是可到达的（例如，经由隧道去往虚拟机）（操作 458）。当隧道封装已经从该分组被去除（操作 456）并且目的地经由隧道不是可到达的（操作 458），则覆盖网关基于 IP 头部的 IP 地址来执行查找（操作 460）并且基于该查找来转发该分组（操作 462）。注意，如果该分组已经被封装，则该 IP 头部指代内部 IP 头部。

[0062] 如果由覆盖网关接收的分组不被封装用于本地 VTEP（操作 454）或者已经利用覆盖网关 VTEP IP 地址而被解除封装（操作 456），并且目的地经由隧道是可到达的（操作 458），则覆盖网关从隧道映射识别 VTEP 地址以及目的地的租户（操作 470）。覆盖网关能够通过查验该分组的目的地 IP 和 / 或 MAC 地址来识别目的地。覆盖网关然后将该分组封装在隧道封装中确保租户分离（操作 472）。在一些实施例中，覆盖网关针对分离的租户使用分离的隧道并且能够将对租户的标识符包括在封装头部中。覆盖网关将所识别的 VTEP IP 地址指配作为目的地 IP 地址，并且将覆盖网关的 IP 地址指配作为封装头部中的源 IP 地址（操作 474）。注意，如果封装机制基于与层 3 不同的层，则覆盖网关能够使用对应层的 VTEP 和网关地址。覆盖网关然后朝向该 VTEP 转发经封装的分组（操作 476）。

[0063] 通常，广播、未知单播、或多播流量（其能够被称为“BUM”流量）被分发给多个接收者。为了部署的容易性，逻辑交换机通常制造属于这种流量的分组的多个拷贝，并且基于隧道封装来朝向与同一逻辑子网相关联的虚拟机交换机个别地单播这些分组。这通常导致对管理程序处理能力的低效使用，尤其在大规模的部署中。为了解决这个问题，覆盖网关能够促进这种流量的高效分发。虚拟交换机能够简单地将“BUM”分组封装在隧道封装中，并且将该分组转发给覆盖网关。覆盖网关进而在逻辑子网中转发该分组。

[0064] 图 4C 呈现了根据本发明的一个实施例的流程图，该流程图图示了覆盖网关在逻辑子网中转发广播、未知单播、或多播分组的过程。在操作期间，覆盖网关接收属于广播、未知单播、或多播流量的经隧道封装的分组（操作 482）。覆盖网关去除隧道封装（操作 484），并且从隧道映射识别与该分组的逻辑子网相关联的（多个）接口（操作 486）。因为覆盖网关从其接收了该分组的虚拟交换机负责将该分组分发给逻辑子网的成员虚拟机，所以覆盖网关不朝向该虚拟交换机转发该分组。

[0065] 覆盖网关然后制造与所识别的（多个）接口的数量相对应的分组的多个拷贝

(操作 488),并且将该分组的相应拷贝封装在用于相应所识别的接口的隧道封装中(操作 490)。因为如连同图 2 所描述的,覆盖网关支持多种隧穿机制,所以如果与逻辑子网相关联的不同虚拟交换机支持不同的隧穿机制,则覆盖网关仍然能够分发该分组。覆盖网关指配相应的所识别的VTEP IP地址作为目的地IP地址,并且覆盖网关的IP地址作为相应封装头部中的源 IP 地址(操作 492)。注意,如果封装机制基于与层 3 不同的层,则覆盖网关能够使用对应层的 VTEP 和网关地址。覆盖网关然后经由对应的所识别的接口朝向对应的 VTEP 转发经封装的分组的相应拷贝(操作 494)。

[0066] 高可用性

[0067] 在图 1A 中的示例中,如果覆盖网关 150 失效或者遇到链路失效,则覆盖网关 150 不再能够操作为网关。因此,向覆盖网关 150 提供高可用性是至关重要的。图 5A 图示了根据本发明的一个实施例的具有高可用性的示例性覆盖网关。如图 5A 中所图示的,虚拟化的网络环境 500(其能够在数据中心中)包括主机 520,主机 520 经由一跳或多跳而耦合至网络 514 中的交换机 512。多个虚拟机运行在主机 520 中的管理程序 522 上。相应的虚拟机具有虚拟端口。运行在管理程序 522 上的相应虚拟机的虚拟端口被逻辑地耦合至由管理程序 522 提供的虚拟交换机 524。也被包括的是虚拟化控制器 540,其将相应的虚拟机分配给主机中的管理程序,并且向该虚拟机指配虚拟 MAC 地址和 IP 地址。

[0068] 虚拟化的网络环境 500 还包括经由逻辑链路 505 而互相耦合的覆盖网关 502 和 504。逻辑链路 505 能够包括经由层 2 和 / 或层 3 而互连的一个或多个物理链路。在这个示例中,覆盖网关 502 仍然保持活动地可操作,而覆盖网关 504 操作为用于覆盖网关 502 的备用网关。在一些实施例中,覆盖网关 502 与虚拟化控制器 540 通信并且获得用于相应虚拟机的对应隧道映射。在一些实施例中,一经获得该隧道映射,覆盖网关就将包括该隧道映射的信息消息发送给覆盖网关 504。以这种方式,覆盖网关 502 和 504 两者能够具有相同的隧道映射。如果该映射在虚拟化控制器 540 中被更新(例如,由于虚拟机迁移),如连同图 1A 所描述的,则覆盖网关 502 从虚拟化控制器 540 获得经更新的隧道映射,并且将包括经更新的隧道映射的信息消息发送给覆盖网关 504。在一些实施例中,覆盖网关 502 和 504 个别地从虚拟化控制器 540 获得该隧道映射。

[0069] 覆盖网关 502 和 504 能够共享逻辑 IP 地址 510。尽管是可操作的,但是活动的覆盖网关 502 使用逻辑 IP 地址 510 作为 VTEP 地址,而备用的覆盖网关 504 抑制与逻辑 IP 地址 510 相关联的操作(例如,ARP 响应)。作为结果,只有覆盖网关 502 响应对于逻辑 IP 地址 510 的任何 ARP 查询。结果,交换机 512 仅得知覆盖网关 502 的 MAC 地址并且将所有的后续分组转发给覆盖网关 502。

[0070] 在常规操作期间,覆盖网关 502 经由网络 514 而促进通向逻辑交换机 524 的虚拟隧穿,逻辑交换机 524 是主机 520 中用于虚拟机 526 的 VTEP。一经从虚拟机 526 获得分组,虚拟交换机 524 就将该分组封装在隧道头部中并且朝向覆盖网关 502 转发经封装的分组。因为交换机 512 仅得知了覆盖网关 502 的 MAC 地址,所以交换机 512 将该分组转发给覆盖网关 502。一经接收到经封装的分组,覆盖网关 502 就去除隧道封装并且朝向该分组的目的地地址转发该分组。

[0071] 覆盖网关 502 和 504 能够经由链路 505 而交换“保持活着”消息来相互通知关于它们的活动状态。假设失效 530 导致了使覆盖网关 502 不可用的链路或设备失效。覆盖网

关504通过在预定时间段内没有从覆盖网关502接收到该保持活着消息来检测失效530，并且采取与逻辑IP地址510相关联的操作。由于覆盖网关502的失效，交换机512通常清除所得知的网关502的MAC地址。在一些实施例中，覆盖网关504发送不必要的ARP响应消息，该不必要的ARP响应消息允许交换机512得知覆盖网关504的MAC地址并且相应地更新它的转发表。基于经更新的转发表，交换机512将来自虚拟机526的后续分组转发给覆盖网关504。一经接收到经封装的分组，覆盖网关504就去除隧道封装并且朝向该分组的目的地地址转发该分组。

[0072] 图5B图示了根据本发明的一个实施例的具有高可用性的覆盖网关的多个地址的一种示例性使用。在这个示例中，覆盖网关502和504能够具有用于不同目的的不同IP地址。例如，除了逻辑IP地址510之外，覆盖网关502和504还能够具有VTEP IP地址550和网关IP地址552。虚拟机524使用网关IP地址552作为默认网关地址。因此，如果虚拟机526需要在它的逻辑子网之外发送分组，则虚拟机526向网关IP地址552发送ARP请求。在一些实施例中，覆盖网关502和504能够具有用于相应逻辑子网的相应网关IP地址，以操作为用于逻辑子网的默认网关。

[0073] 虚拟交换机524使用逻辑IP地址510作为默认网关地址，并且使用VTEP IP地址550作为默认隧道目的地地址。VTEP IP地址550能够在与主机520中的虚拟机相关联的（多个）逻辑子网之外。为了将经隧道封装的分组发送给VTEP IP地址550，虚拟交换机524向逻辑IP地址510发送ARP请求。因为所有去往VTEP IP地址550的经封装的分组都指向逻辑IP地址510，所以覆盖网关502接收分组并且采取适当的动作。以这种方式，单个VTEP IP地址550（其能够在与相应虚拟机相关联的逻辑子网之外）操作为用于所有逻辑子网的隧道目的地地址。

[0074] 如连同图5A所描述的，一经检测到失效530，覆盖网关504就采取与逻辑IP地址510相关联的操作。因为所有去往VTEP IP地址550的经封装的分组都指向逻辑IP地址510，所以这些经封装的分组指向覆盖网关504。因此，仅向逻辑IP地址510提供高可用性，对于向去往VTEP IP地址550的经隧道封装的分组确保高可用性是足够的。然而，如果VTEP IP地址550在虚拟机526的逻辑子网中，则虚拟机526直接将分组发送给VTEP IP地址550。在这样的场景下，向VTEP IP地址550提供高可用性也是必要的。

[0075] 示例性覆盖网关

[0076] 图6图示了根据本发明的一个实施例的操作为覆盖网关的示例性计算系统。在这个示例中，计算系统600包括通用处理器604、存储器606、多个通信端口602、分组处理器610、隧道管理模块630、转发模块632、控制模块640、高可用性模块620、以及存贮器650。处理器604执行存储在存储器606中的指令来操作计算系统600作为覆盖网关，该覆盖网关发起或终止与虚拟机相关联的覆盖隧道。

[0077] 在操作期间，通信端口602中的一个通信端口从配置系统接收分组。这个配置系统能够是以下各项中的一项或多项：虚拟化控制器、网络管理器、以及垫补设备。分组处理器610，连同控制模块640，从所接收的分组中提取隧道映射。这个隧道映射将虚拟机的虚拟IP地址和/或MAC地址映射至VTEP IP地址。控制模块640将隧道映射存储在存贮器650中并且在操作期间加载在存储器606中。如连同图1A所描述的，当该映射被更新时，控制模块640也更新该映射。隧道管理模块630辨识多种覆盖隧穿机制。当去往虚拟机的数

据分组被接收时，转发模块632从用于虚拟机的映射中获得VTEP IP地址，基于所辨识的隧穿机制来封装该分组，并且基于该VTEP IP地址而为该数据分组在通信端口602之中确定输出端口。

[0078] 如连同图5A所描述的，高可用性模块620将计算系统600与逻辑IP地址相关联，该逻辑IP地址也与一个远程计算系统相关联。高可用性模块620确定计算系统600是活动的覆盖网关还是备用的覆盖网关。如果计算系统600是备用的覆盖网关，则处理器604阻止分组处理器610处理与该逻辑IP地址相关联的分组。当高可用性模块620检测到该远程计算系统的失效时，分组处理器610开始处理与该逻辑IP地址相关联的分组。如连同图5B所描述的，在一些实施例中，高可用性模块620也将计算系统600与VTEP地址相关联，该VTEP地址属于与该逻辑IP地址所属的子网不同的子网。

[0079] 注意，上面所提到的模块能够在硬件中以及在软件中被实施。在一个实施例中，这些模块能够被具体化在存储器中所存储的计算机可执行指令中，该存储器耦合至计算系统600中的一个或多个处理器。当被执行时，这些指令促使该（些）处理器执行前面提到的功能。

[0080] 总之，本发明的各实施例提供了用于促进层3覆盖隧穿的一种计算系统和一种方法。在一个实施例中，该计算系统包括处理器和用于存储指令的计算机可读存储介质。基于这些指令，该处理器操作该计算系统作为覆盖网关。该计算系统发起和终止与虚拟机相关联的覆盖隧道。在操作期间，该计算系统基于从配置系统接收的信息，而将虚拟机的虚拟互联网协议（IP）地址映射至被用来终止该覆盖隧道的第二IP地址。该计算系统然后基于该第二IP地址来为数据分组确定输出端口。该数据分组包括内部分组，并且这个内部分组的目的地地址对应于该虚拟IP地址。

[0081] 本文所描述的这些方法和过程能够被具体化为代码和/或数据，该代码和/或数据能够被存储在计算机可读的非瞬态存储介质中。当计算机系统读取和执行存储在该计算机可读的非瞬态存储介质中的该代码和/或数据时，该计算机系统执行被具体化为数据结构和代码并且被存储在该介质内的这些方法和过程。

[0082] 本文所描述的这些方法和过程能够由硬件模块或装置执行和/或被包括在硬件模块或装置中。这些模块或装置可以包括但不限于：专用集成电路（ASIC）芯片、现场可编程门阵列（FPGA）、在特定时间执行特定软件模块或一段代码的专属或共享的处理器、和/或现在已知的或者以后开发的其他可编程逻辑设备。当这些硬件模块或装置被激活时，它们执行被包括在它们内的这些方法和过程。

[0083] 对本发明的各实施例的前述描述仅为了举例说明和描述的目的而已经被提出。它们不意图为是穷尽的或者限制本公开内容。因此，许多修改和变型对本领域的技术人员将是明显的。本发明的范围由所附权利要求来定义。

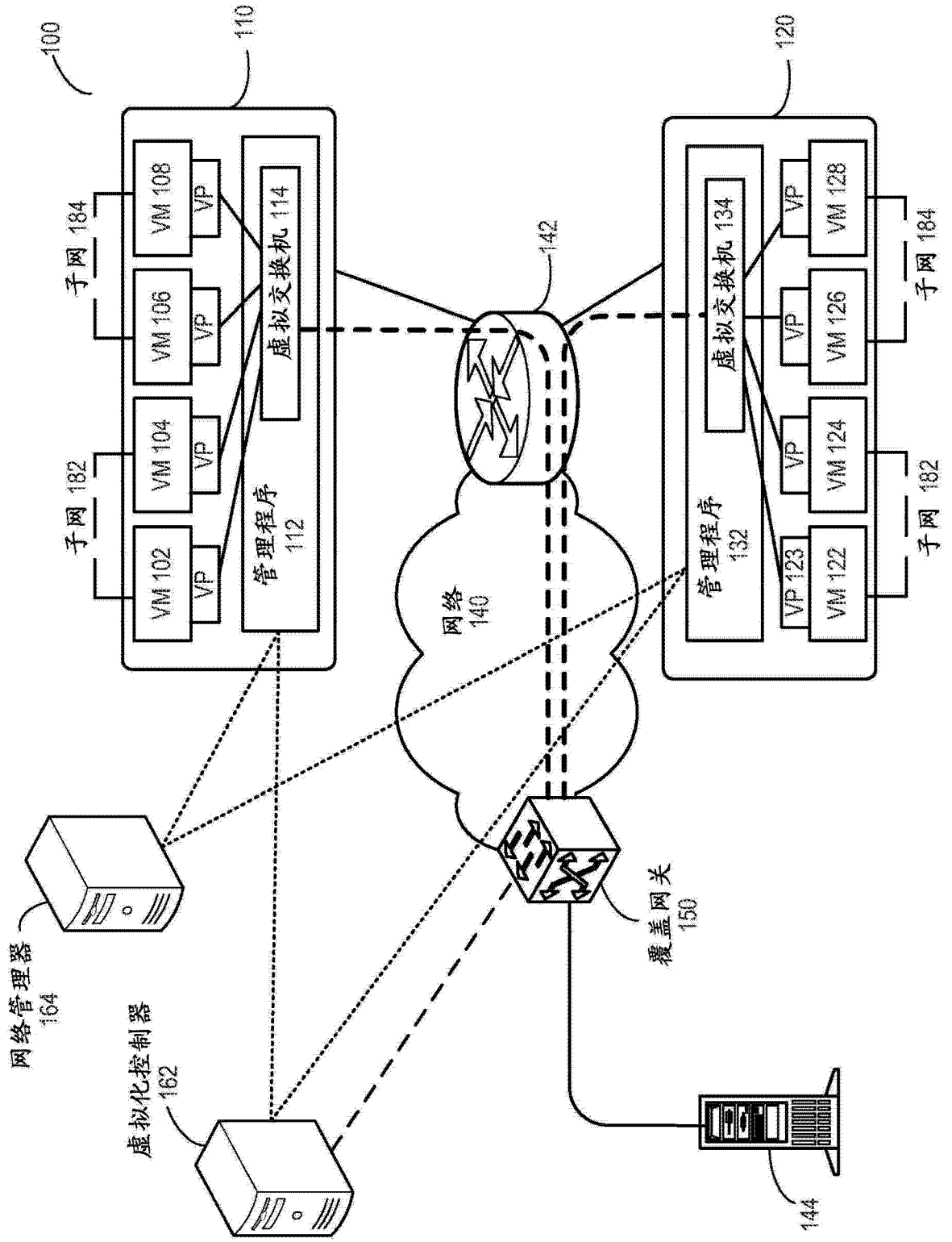

100
110
120
142
子网 184
子网 182
VM 108
VM 106
VM 104
VM 102
VP
虚拟交换机 114
管理程序 112
网络 140
虚拟交换机 134
管理程序 132
VM 128
VM 126
VM 124
VM 122
VP 123
网络管理器 164
虚拟化控制器 162
覆盖网关 150
144

图 1A

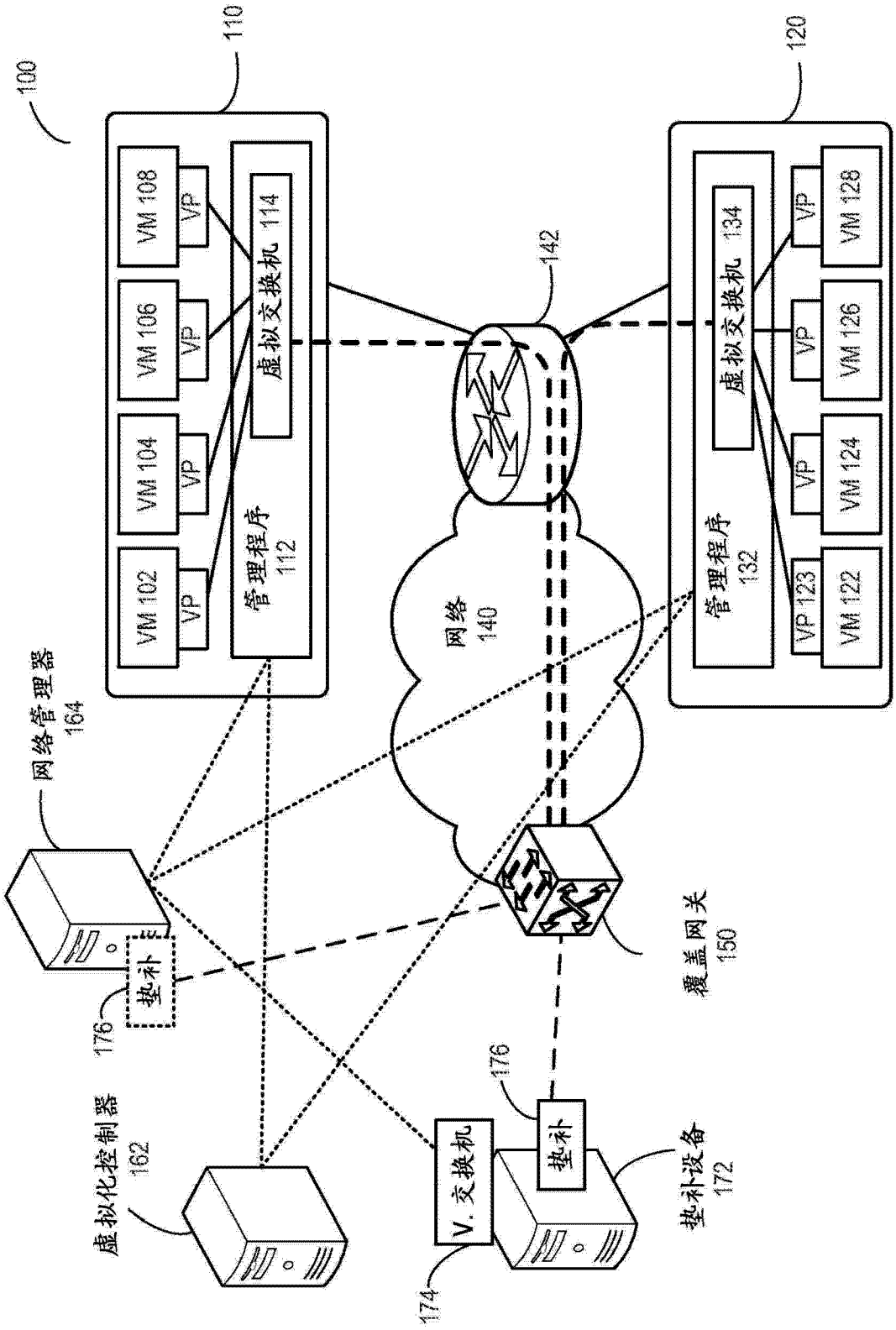

100
110
120
VM 108
VM 106
VM 104
VM 102
VP
虚拟交换机 114
管理程序 112
142
网络 140
虚拟交换机 134
管理程序 132
VM 128
VM 126
VM 124
VM 122
VP 123
网络管理器 164
垫补
176
虚拟化控制器 162
V. 交换机
174
覆盖网关 150
垫补设备 172

图 1B

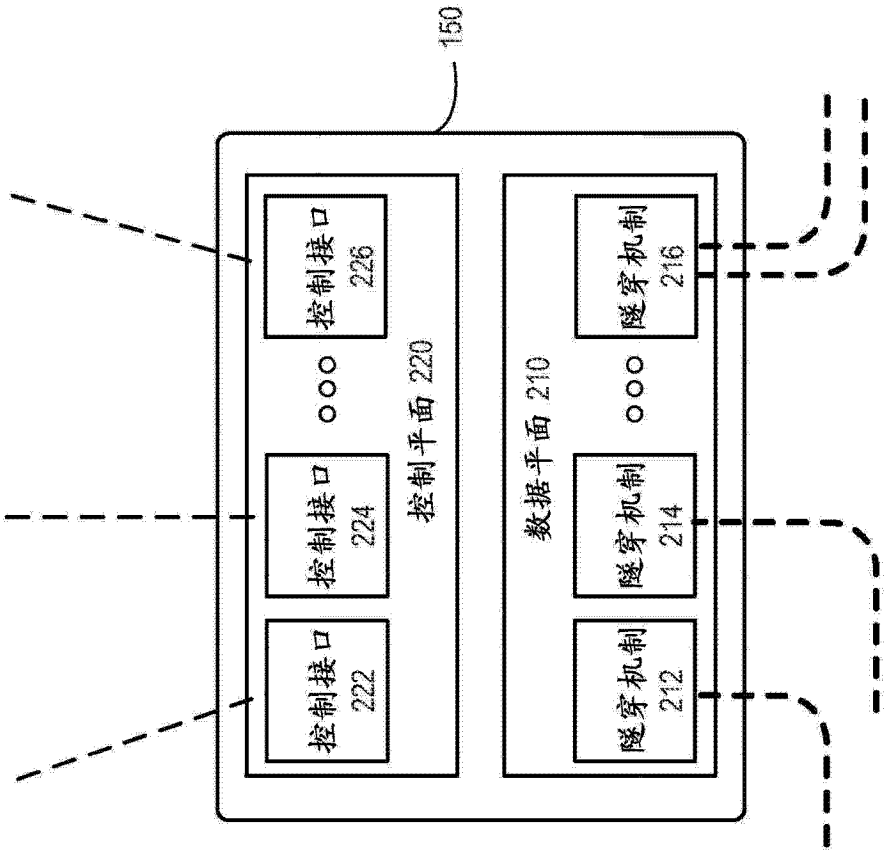

150
控制接口 226
控制接口 224
控制接口 222
控制平面 220
数据平面 210
隧穿机制 216
隧穿机制 214
隧穿机制 212

图 2

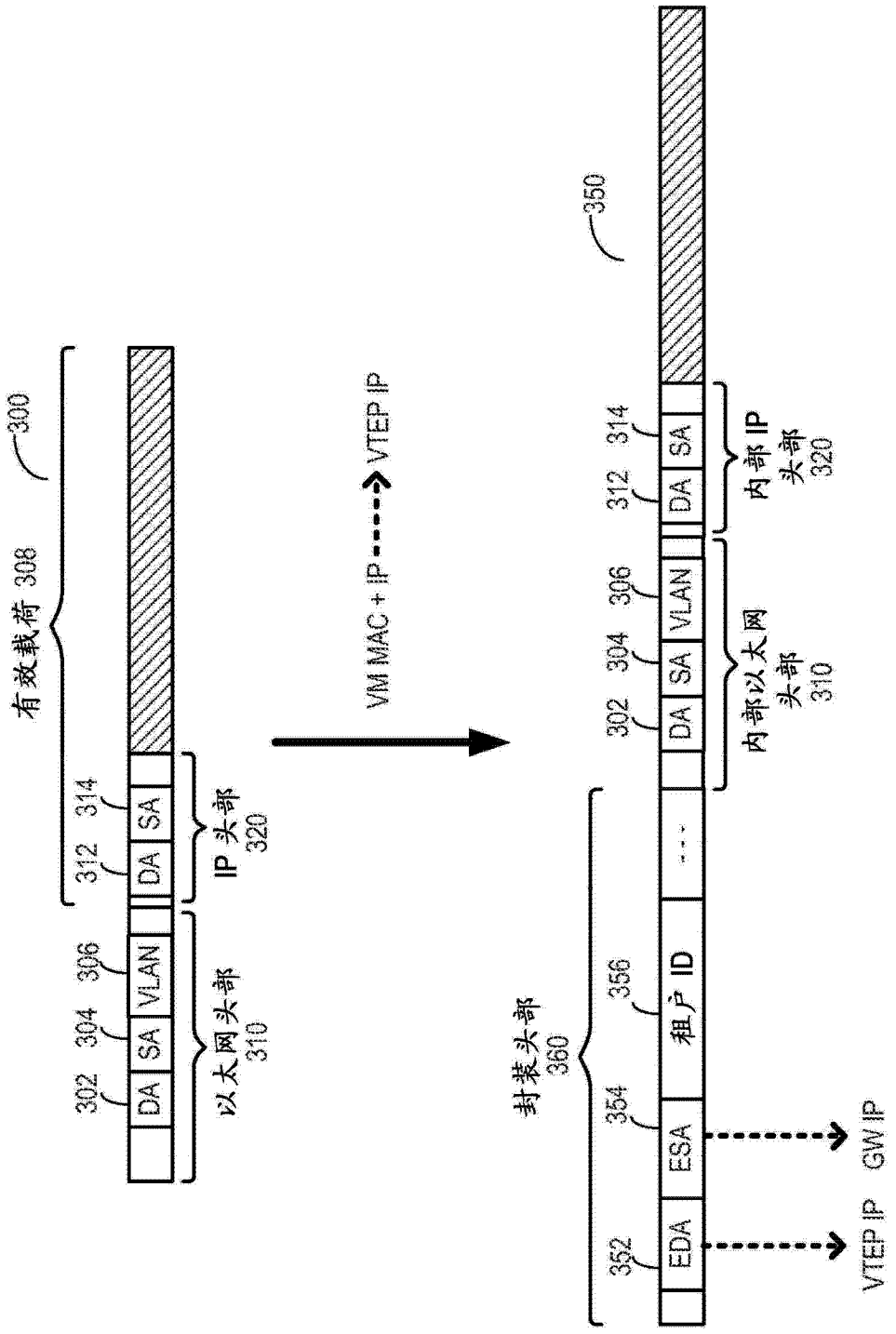

300
有效载荷 308
302
304
306
312
314
DA
SA
VLAN
DA
SA
以太网头部
310
IP 头部
320
VM MAC + IP
VTEP IP
350
封装头部
360
352
354
356
EDA
ESA
租户 ID
302
304
306
312
314
DA
SA
VLAN
DA
SA
内部以太网
头部
310
内部 IP
头部
320
VTEP IP
GW IP

图 3

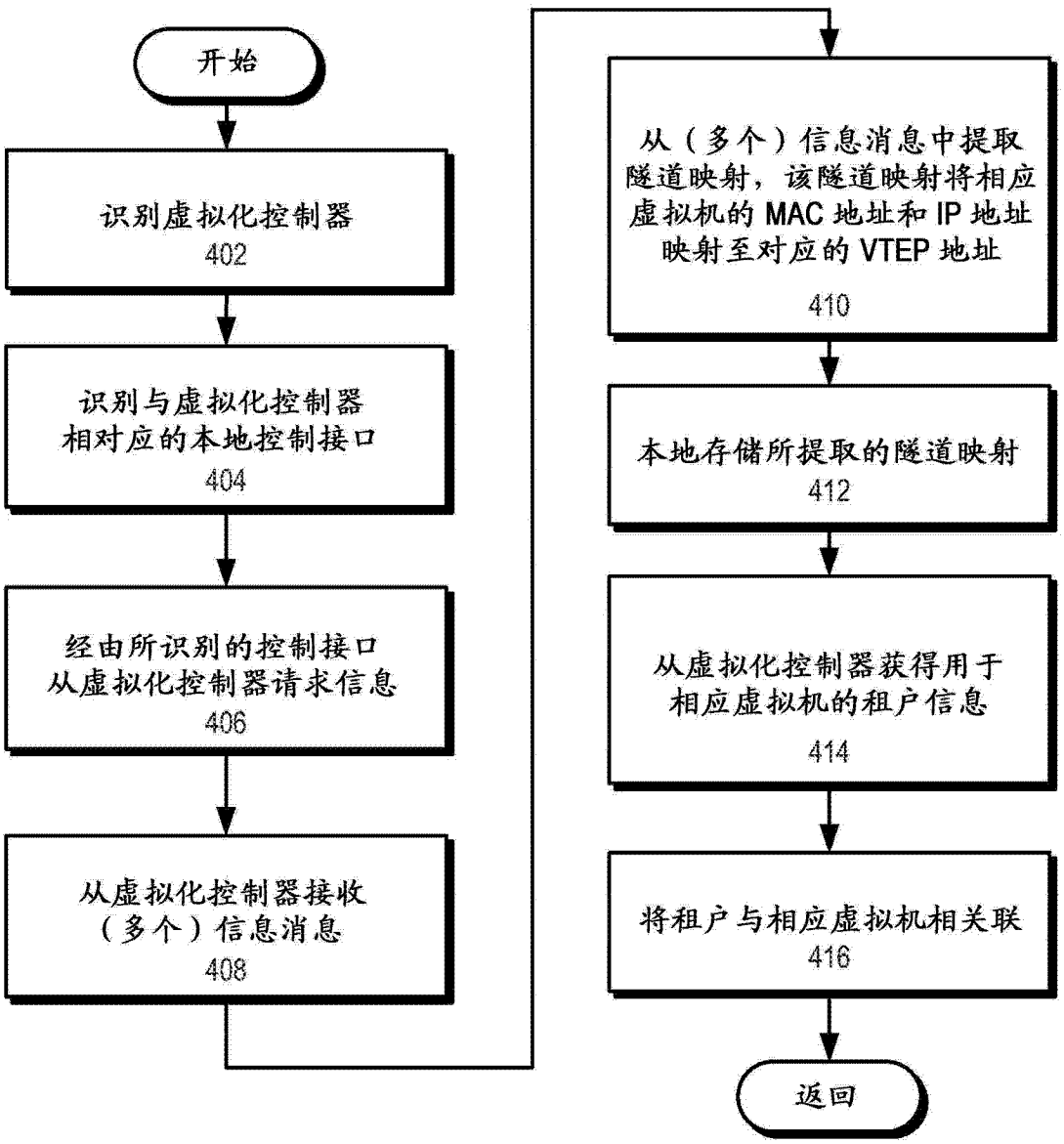

开始
识别虚拟化控制器
402
识别与虚拟化控制器
相对应的本地控制接口
404
经由所识别的控制接口
从虚拟化控制器请求信息
406
从虚拟化控制器接收
（多个）信息消息
408
从（多个）信息消息中提取
隧道映射，该隧道映射将相应
虚拟机的 MAC 地址和 IP 地址
映射至对应的 VTEP 地址
410
本地存储所提取的隧道映射
412
从虚拟化控制器获得用于
相应虚拟机的租户信息
414
将租户与相应虚拟机相关联
416
返回

图 4A

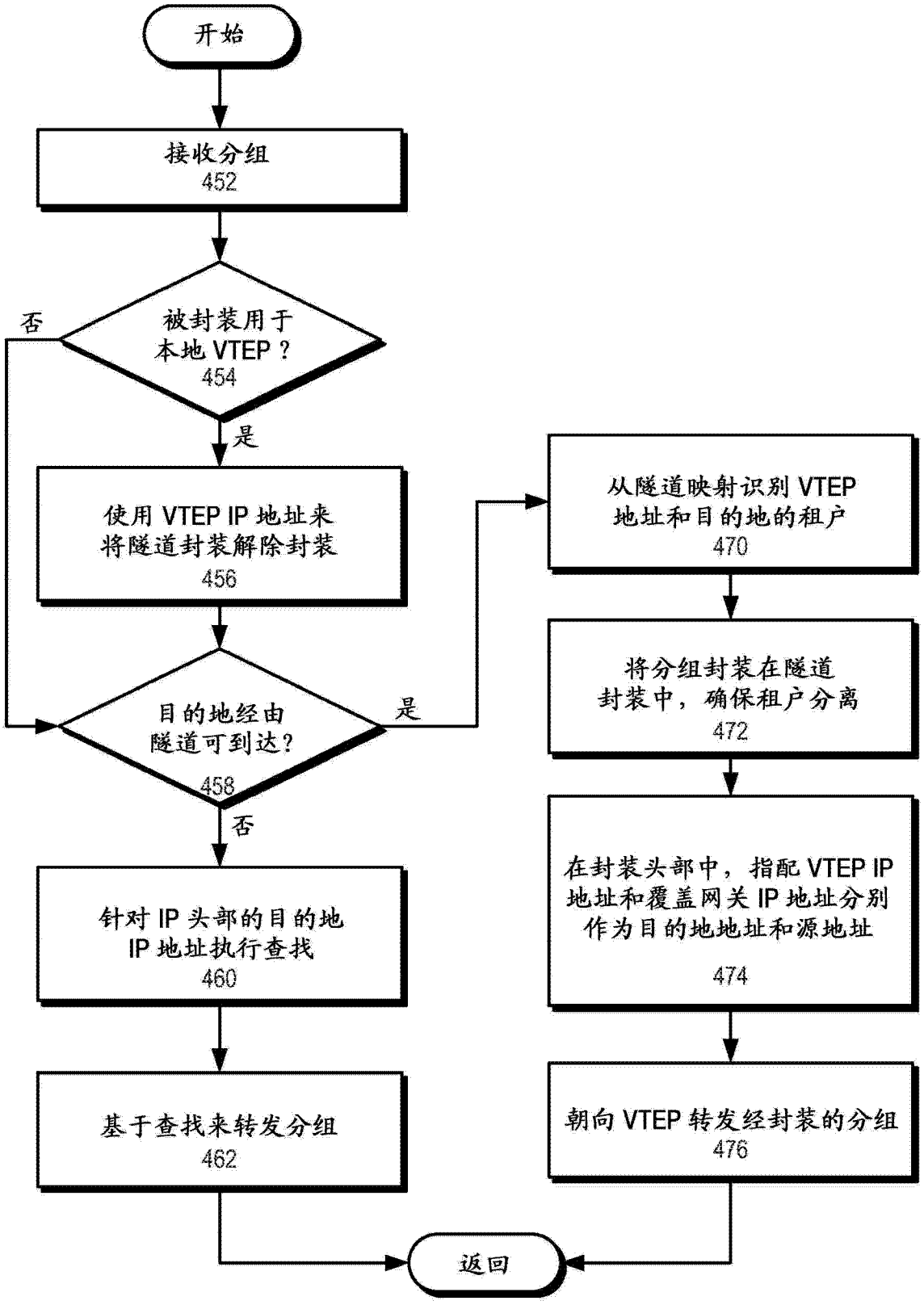

开始
接收分组
452
被封装用于
本地 VTEP ?
454
否
是
使用 VTEP IP 地址来
将隧道封装解除封装
456
目的地经由
隧道可到达?
458
是
否
从隧道映射识别 VTEP
地址和目的地的租户
470
将分组封装在隧道
封装中，确保租户分离
472
在封装头部中，指配 VTEP IP
地址和覆盖网关 IP 地址分别
作为目的地地址和源地址
474
朝向 VTEP 转发经封装的分组
476
针对 IP 头部的目的地
IP 地址执行查找
460
基于查找来转发分组
462
返回

图 4B

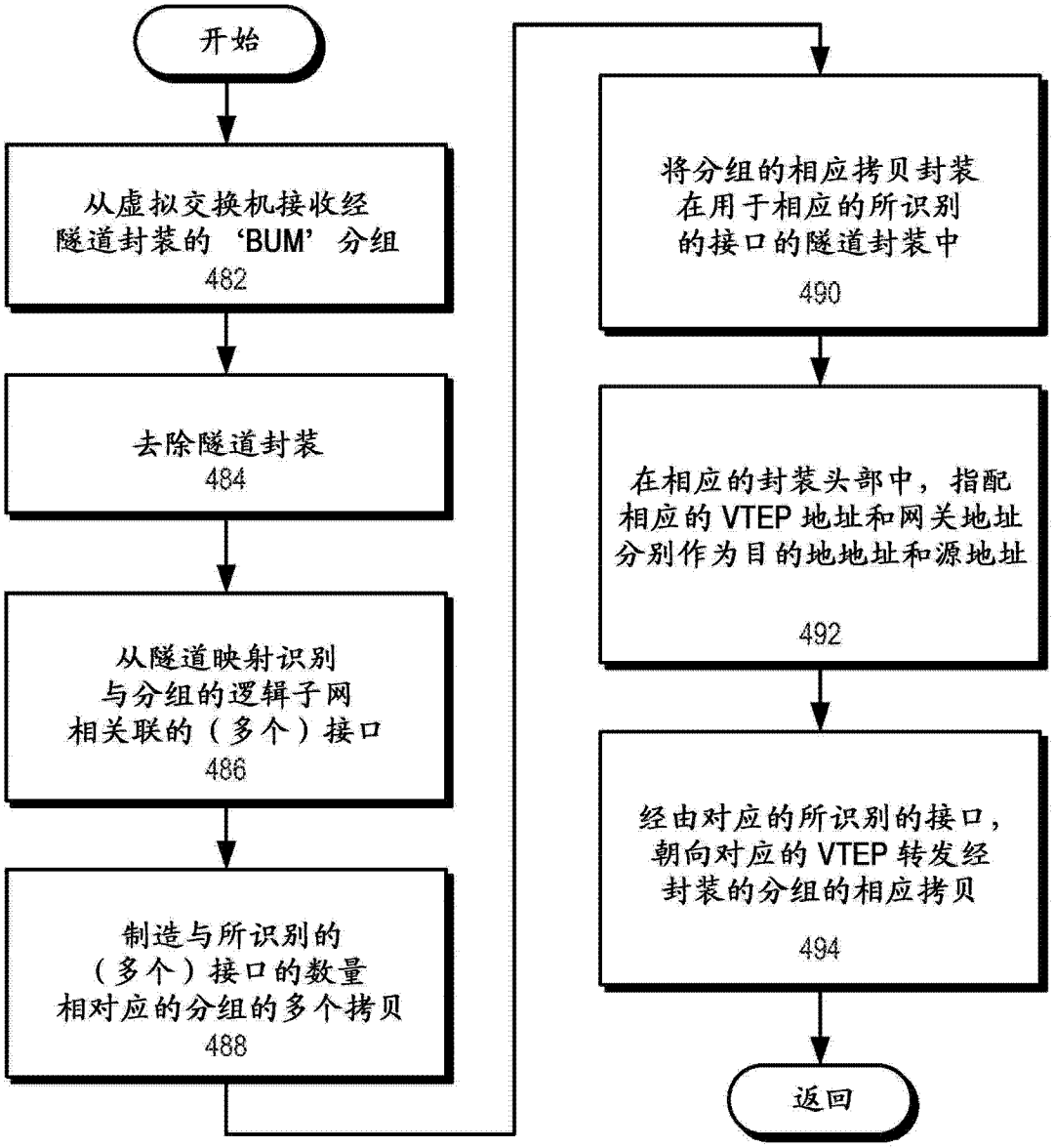

开始
从虚拟交换机接收经隧道封装的‘BUM’分组
482
去除隧道封装
484
从隧道映射识别与分组的逻辑子网相关联的（多个）接口
486
制造与所识别的（多个）接口的数量相对应的分组的多个拷贝
488
将分组的相应拷贝封装在用于相应的所识别的接口的隧道封装中
490
在相应的封装头部中，指配相应的 VTEP 地址和网关地址分别作为目的地地址和源地址
492
经由对应的所识别的接口，朝向对应的 VTEP 转发经封装的分组的相应拷贝
494
返回

图 4C

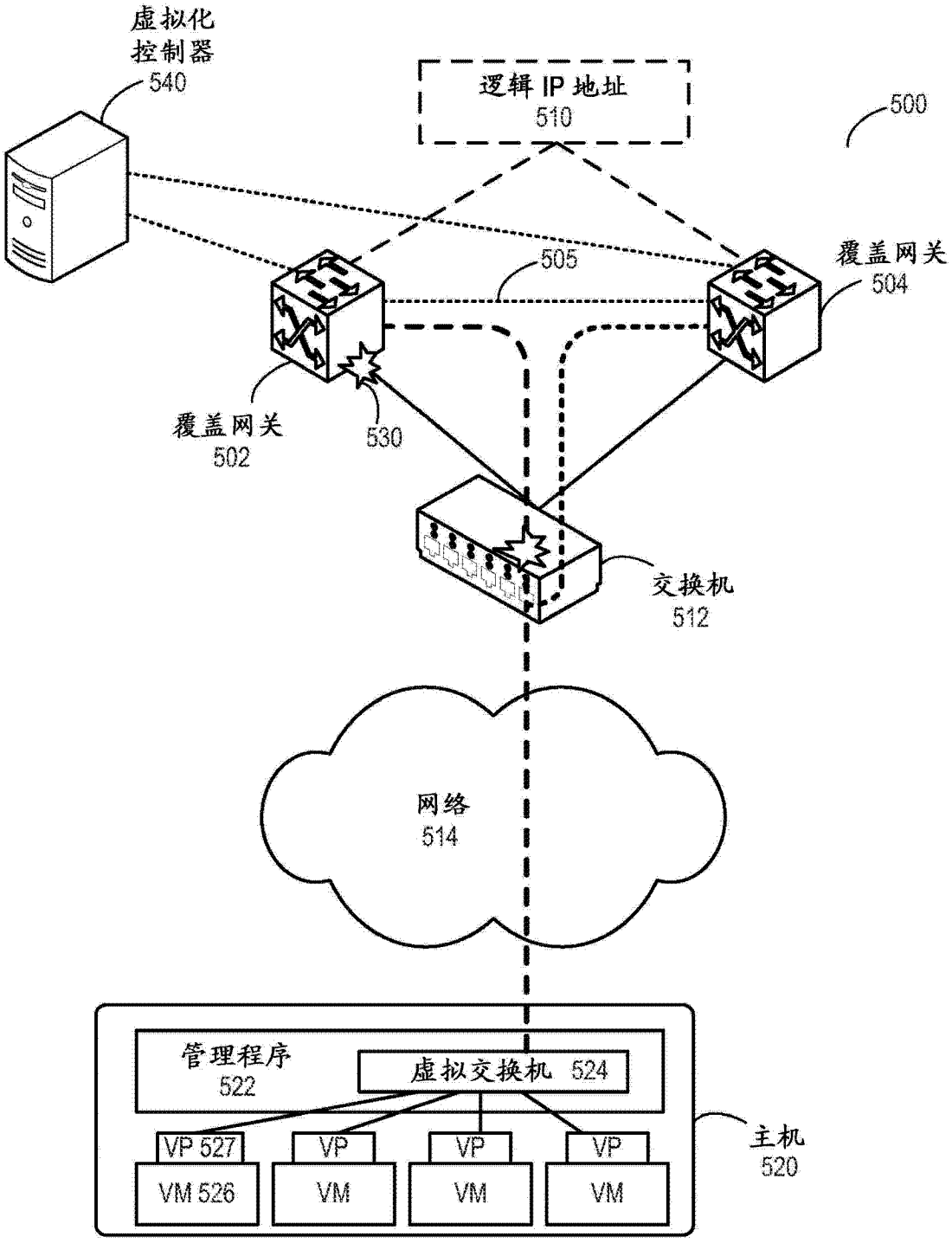

虚拟化
控制器
540
逻辑 IP 地址
510
500
505
覆盖网关
504
覆盖网关
502
530
交换机
512
网络
514
管理程序
522
虚拟交换机 524
主机
520
VP 527
VP
VP
VP
VM 526
VM
VM
VM

图 5A

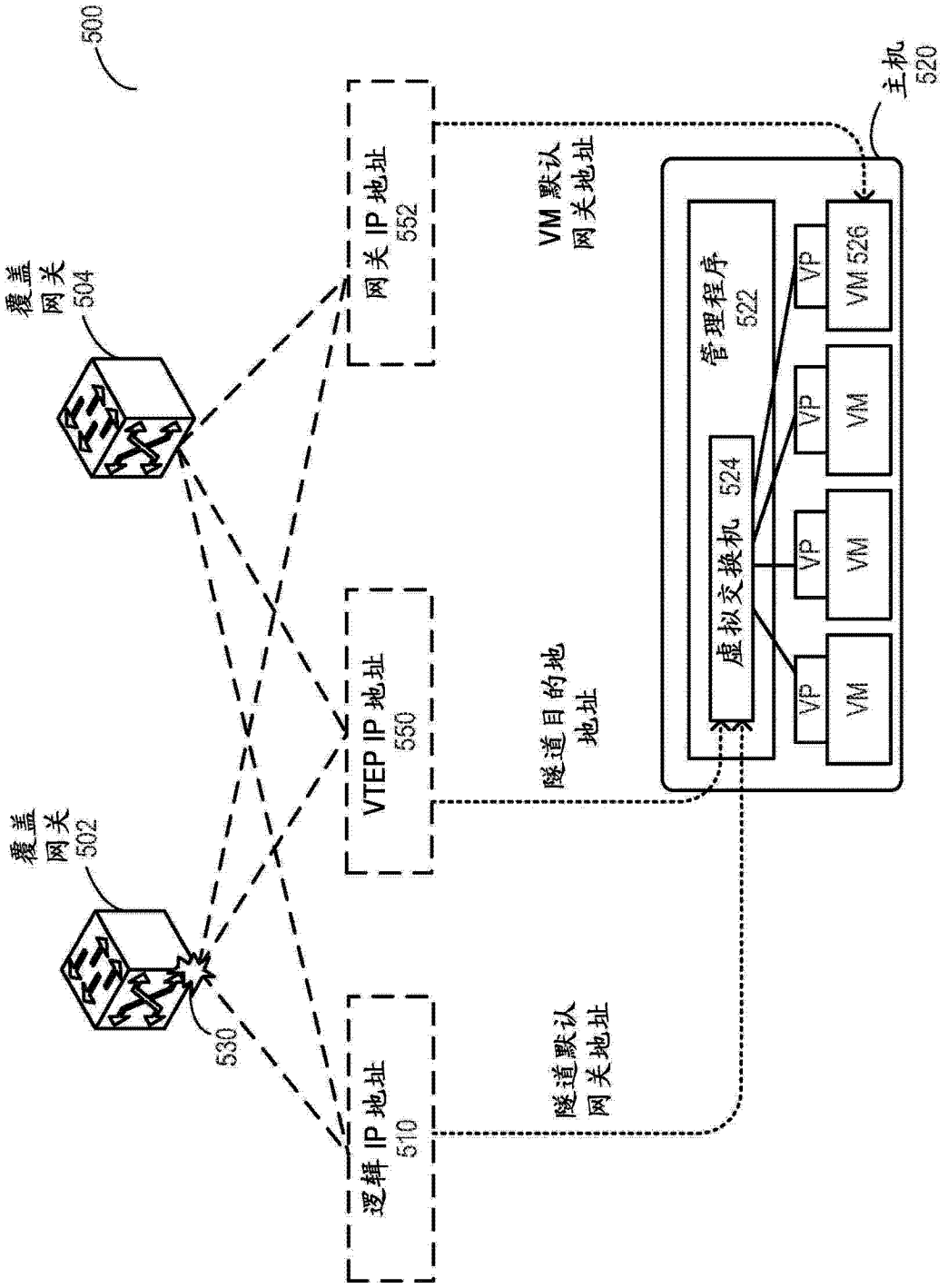

500
主机
520
网关 IP 地址
552
VM 默认
网关地址
覆盖
网关
504
管理程序
522
VP
VM 526
VP
VM
虚拟交换机 524
VP
VM
VP
VM
VTEP IP 地址
550
隧道目的地
地址
覆盖
网关
502
530
逻辑 IP 地址
510
隧道默认
网关地址

图 5B

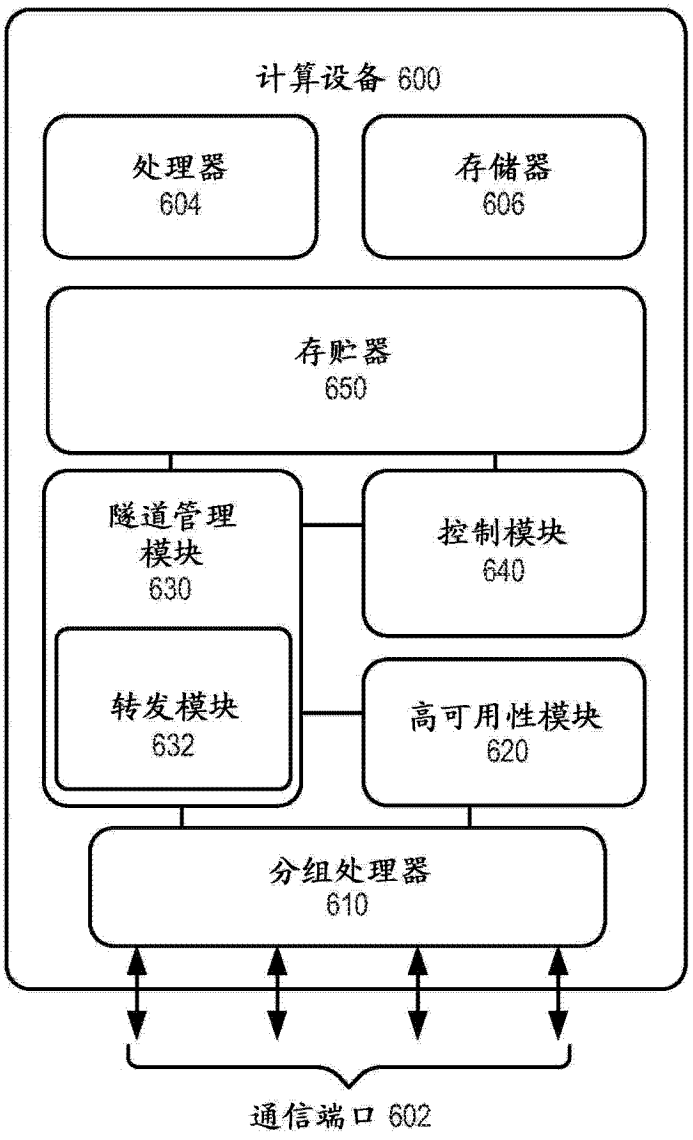

计算设备 600
处理器
604
存储器
606
存贮器
650
隧道管理
模块
630
转发模块
632
控制模块
640
高可用性模块
620
分组处理器
610
通信端口 602

图 6