(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107071088 B
(45)授权公告日 2020.06.05

(21)申请号 201710263367.6

(22)申请日 2012.08.17

(65)同一申请的已公布的文献号
申请公布号 CN 107071088 A

(43)申请公布日 2017.08.18

(30)优先权数据
61/524,754 2011.08.17 US
61/643,339 2012.05.06 US
61/654,121 2012.06.01 US
61/666,876 2012.07.01 US

(62)分案原申请数据
201280046542.1 2012.08.17

(73)专利权人 NICIRA股份有限公司
地址 美国加利福尼亚

(72)发明人 T•考珀内恩 张荣华 M•卡萨多 P•萨卡尔 J•E•格鲁斯四世 D•J•温德兰德特 M•马哈杰安 J•皮提特 K•E•埃米顿

(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 张劲松

(51)Int.Cl.
H04L 29/12(2006.01)
H04L 12/803(2013.01)
H04L 12/801(2013.01)
H04L 12/715(2013.01)
H04L 12/741(2013.01)

(56)对比文件
US 2011085559 A1,2011.04.14,
CN 1826771 A,2006.08.30,
WO 2010068618 A1,2010.06.17,
US 2010246443 A1,2010.09.30,
US 2005257256 A1,2005.11.17,
KR 20050083427 A,2005.08.26,

审查员 王维国

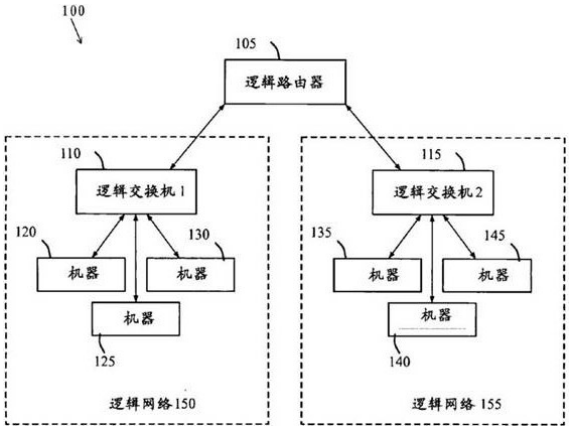
权利要求书2页 说明书74页 附图69页

(54)发明名称

逻辑L3路由

(57)摘要

本公开涉及逻辑L3路由。描述了一种用于在第一逻辑域中的源机器与第二逻辑域中的目的地机器之间逻辑地路由分组的新颖方法。该方法将受管理交换元件配置为第二级受管理交换元件。该方法在包括第二级受管理交换元件的主机中配置路由器。该方法将第二级受管理交换元件与路由器通信地耦合。当路由器从第一逻辑域接收到被定址到第二逻辑域的分组时,该方法使路由器路由分组。



CN 107071088 B

1.一种用于配置多个受管理转发元件MFE以实现逻辑L3路由器和多个逻辑L2交换机的方法,所述方法包括:

对于实现特定的逻辑L2交换机的一组MFE中的每一个MFE,生成第一组数据记录,所述第一组数据记录用于配置MFE以实现用于处理被发送到逻辑地耦合到所述特定的逻辑L2交换机的网络地址的分组的逻辑L3路由器以及特定的逻辑L2交换机;以及

对于所述一组MFE中的每一个MFE,生成第二组数据记录,所述第二组数据记录用于配置MFE以对被发送到逻辑地耦合到所述特定的逻辑L2交换机的网络地址的分组的子集实现负载均衡处理,所述第二组数据记录指定跨越多个机器均衡所述分组的子集,所述多个机器逻辑地耦合到所述特定的逻辑L2交换机并且物理地耦合到多个不同的MFE。

2.根据权利要求1所述的方法,还包括对于所述一组MFE中的每一个MFE,生成第三组数据记录,所述第三组数据记录用于配置MFE以在对从第二逻辑L2交换机发送到所述特定的逻辑L2交换机的分组执行负载均衡处理之前对所述分组实现源网络地址转换NAT处理。

3.根据权利要求1所述的方法,其中,逻辑地耦合到所述特定的逻辑L2交换机的所述多个机器包括提供相同服务的多个虚拟机。

4.根据权利要求1所述的方法,其中,所述一组MFE中的MFE中的每一个在不同的主机上操作,其中,负载均衡守护进程被配置在主机中的每一个上以从逻辑地耦合到所述特定的逻辑L2交换机的所述多个机器中选择机器。

5.根据权利要求4所述的方法,其中,所述负载均衡守护进程基于所述多个机器中的每一个的当前工作负载从所述多个机器中选择机器。

6.根据权利要求4所述的方法,其中,在具有特定的MFE的特定主机中操作的特定的负载均衡守护进程将所选择的机器的地址发送到所述特定的MFE。

7.根据权利要求4所述的方法,其中,用于配置特定的MFE以执行负载均衡处理的所述第二组数据记录对于所述特定的MFE指定从所述负载均衡守护进程请求对逻辑地耦合到所述特定的逻辑L2交换机的机器中的一个的选择。

8.根据权利要求1所述的方法,还包括配置所述一组MFE中的每一个MFE以与所述一组MFE中的其它MFE中的每一个建立隧道。

9.根据权利要求1所述的方法,其中,生成第一组数据记录和第二组数据记录包括生成第一组流条目和第二组流条目。

10.根据权利要求1所述的方法,其中,所述方法由网络控制器执行。

11.一种用于配置多个受管理转发元件MFE以实现逻辑L3路由器和多个逻辑L2交换机的系统,所述系统包括:

用于对于实现特定的逻辑L2交换机的一组MFE中的每一个MFE生成第一组数据记录的部件,所述第一组数据记录用于配置MFE以实现用于处理被发送到逻辑地耦合到所述特定的逻辑L2交换机的网络地址的分组的逻辑L3路由器以及特定的逻辑L2交换机;以及

用于对于所述一组MFE中的每一个MFE生成第二组数据记录的部件,所述第二组数据记录用于配置MFE以对被发送到逻辑地耦合到所述特定的逻辑L2交换机的网络地址的分组的子集实现负载均衡处理,所述第二组数据记录指定跨越多个机器均衡所述分组的子集,所述多个机器逻辑地耦合到所述特定的逻辑L2交换机并且物理地耦合到多个不同的MFE。

12.根据权利要求11所述的系统,还包括用于对于所述一组MFE中的每一个MFE生成第

三组数据记录的部件，所述第三组数据记录用于配置MFE以在对从第二逻辑L2交换机发送到所述特定的逻辑L2交换机的分组执行负载均衡处理之前对所述分组实现源网络地址转换NAT处理。

13.根据权利要求11所述的系统，其中，逻辑地耦合到所述特定的逻辑L2交换机的所述多个机器包括提供相同服务的多个虚拟机。

14.根据权利要求11所述的系统，其中，所述一组MFE中的MFE中的每一个在不同的主机上操作，其中，负载均衡守护进程被配置在主机中的每一个上以从逻辑地耦合到所述特定的逻辑L2交换机的所述多个机器中选择机器。

15.根据权利要求14所述的系统，其中，所述负载均衡守护进程基于所述多个机器中的每一个的当前工作负载从所述多个机器中选择机器。

16.根据权利要求14所述的系统，其中，在具有特定的MFE的特定主机中操作的特定的负载均衡守护进程将所选择的机器的地址发送到所述特定的MFE。

17.根据权利要求14所述的系统，其中，用于配置特定的MFE以执行负载均衡处理的所述第二组数据记录对于所述特定的MFE指定从所述负载均衡守护进程请求对逻辑地耦合到所述特定的逻辑L2交换机的机器中的一个的选择。

18.根据权利要求11所述的系统，还包括用于配置所述一组MFE中的每一个MFE以与所述一组MFE中的其它MFE中的每一个建立隧道的部件。

19.根据权利要求11所述的系统，其中，用于生成第一组数据记录和第二组数据记录的部件包括用于生成第一组流条目和第二组流条目的部件。

20.一种存储控制器应用的机器可读介质，所述控制器应用在被至少一个处理单元执行时实施根据权利要求1-10中的任何一项所述的方法。

逻辑L3路由

[0001] 本申请是基于申请号为201280046542.1、申请日为2012年8月17日、发明名称为“分布式逻辑L3路由”的专利申请的分案申请。

背景技术

[0002] 许多当前的企业具有包括交换机、集线器、路由器、服务器、工作站和其它联网设备的大型且尖端的网络，这些网络支持多种连接、应用和系统。计算机网络的增加的尖端性，包括虚拟机迁移、动态工作负载、多租赁和依客户而定的服务质量和安全配置要求网络控制的更好范式。网络在传统上通过对单独组件的低级别配置来管理。网络配置经常取决于底层网络：例如，利用访问控制列表(“ACL”)条目阻止用户的访问要求知道用户的当前IP地址。更复杂的任务要求更广泛的网络知识：迫使访客用户的端口80流量穿过HTTP代理要求知道当前的网络拓扑和每个访客的位置。此过程在网络交换元件(switching element)跨越多个用户被共享的情况下逐渐困难。

[0003] 作为响应，存在朝着一种被称为软件定义网络(SDN)的新的网络控制范式的日益发展。在SDN范式中，在网络中的一个或多个服务器上运行的网络控制器控制、维护并实现逐个用户地管治共享的网络交换元件的转发行为的控制逻辑。作出网络管理决策经常要求关于网络状态的知识。为了促成管理决策的作出，网络控制器创建并维护网络状态的视图并提供应用编程接口，在该应用编程接口上管理应用可访问网络状态的视图。

[0004] 维护大型网络(包括数据中心和企业网络)二者的一些主要目标是可扩展性(scalability)、移动性和多租赁。用于处理这些目标中的一个的许多方法都导致妨碍其它目标中的至少一个。例如，可以容易地在L2域内为虚拟机提供网络移动性，但L2域不能扩展到大的尺寸。另外，保持用户隔离大大地使移动性复杂化。这样，需要能够满足可扩展性、移动性和多租赁目标的改进方案。

发明内容

[0005] 一些实施例在一些情况下将逻辑路由建模为由实现在L3域中操作的逻辑数据路径集合(LDPS)的逻辑路由器互连在L2域中操作的两个或更多个逻辑数据路径(LDP)集合的行为。从一个逻辑L2域穿越到另一个逻辑L2域的分组在一些实施例中将采取以下四个步骤。下文中按照网络控制系统实现的逻辑处理操作来描述这四个步骤。然而，要理解，这些操作由网络的受管理交换元件基于由网络控制系统产生的物理控制平面数据来执行。

[0006] 第一，将通过发端逻辑L2域的L2表管道来处理分组。该管道将以目的地媒体访问控制(MAC)地址被转发到与逻辑路由器的逻辑端口附接的逻辑端口结束。

[0007] 第二，将通过逻辑路由器的L3数据路径来处理分组，这同样通过经由此路由器的L3表管道发送它来完成。在一些实施例中在路由器的L3数据路径中跳过物理路由器中常见的L2查找阶段，因为逻辑路由器将仅接收要求路由的分组。

[0008] 在一些实施例中，L3转发决策将使用前缀(由逻辑路由器的逻辑控制平面配设的转发信息库(FIB)条目)。在一些实施例中，控制应用用于接收逻辑控制平面数据，并将此数

据转换成逻辑转发平面数据，该逻辑转发平面数据随后被提供给网络控制系统。对于L3转发决策，一些实施例使用前缀FIB条目来实现最长的前缀匹配。

[0009] 结果，L3路由器将把分组转发到与目的地L2 LDPS相“连接”的逻辑端口。在将分组进一步转发到该LDPS之前，L3路由器将把发端MAC地址改变成在其域中所定义的那个，以及将把目的地IP地址解析成目的地MAC地址。该解析在一些实施例中由L3数据管道的最末“IP输出”阶段执行。同一管道将递减TTL并更新检验和(并且如果TTL达到零则以ICMP响应)。

[0010] 应当注意，一些实施例在将经处理的分组馈送到下一LDPS之前改写MAC地址，因为如果没有这个改写，则在下一LDPS处可得到不同的转发决策。还应当注意，即使传统的路由器利用ARP来执行目的地IP地址的解析，但一些实施例在L3逻辑路由器中不将ARP用于此目的，因为只要下一跳是逻辑L2数据路径，这个解析就保持在虚拟化应用的内部。

[0011] 第三，将通过目的地逻辑L2域的L2表管道来处理分组。目的地L2表管道确定其应当发送分组的逻辑出口端口。在未知MAC地址的情况下，这个管道将通过依赖于一些分布式查找机制来解析MAC地址位置。在一些实施例中，受管理交换元件依赖于MAC学习算法，例如，它们洪泛未知分组。在这些或其它实施例中，MAC地址位置信息也可通过其它机制获得，例如通过带外获得。如果这种机制在一些实施例中可用，则最末的逻辑L2表管道使用此机制来获得MAC地址位置。

[0012] 第四，分组被发送到表示逻辑端口附接的附接到物理端口的逻辑端口。在这个阶段，如果端口是点对点媒介(例如，虚拟网络接口，VIF)，则除了将分组发送到该端口以外就没有什么要做的了。然而，如果最末的LDPS是L3路由器并且因此附接是物理L3子网，则附接点在一些实施例中在将分组送出之前通过使用ARP来解析目的地IP地址。在该情况下，源MAC地址在VIF的情况下将依出口而定，而不是逻辑MAC接口地址。在其它实施例中，利用ARP解析目的地IP地址在第二步骤期间由L3逻辑路由器执行。

[0013] 在上述示例中，仅存在互连逻辑L2数据路径的单个逻辑路由器，但没有什么对拓扑进行限制。普通技术人员将认识到，对于更丰富的拓扑可互连更多的LDP集合。

[0014] 在一些实施例中，控制应用允许按照指明逻辑L3管道的一个或多个表来定义L3特定的逻辑状态。管理LDPS管道的相应逻辑控制平面或者可依赖于静态路由配置，或者可通过标准的路由协议与其它LDP集合对等。

[0015] 在一些实施例中，虚拟化应用定义上述的四步L2/L3分组处理成为物理控制平面数据的物理实现，该物理控制平面数据当被受管理交换元件转换成物理转发数据时，实现全部或主要在第一跳受管理边缘交换元件处执行的逻辑管道执行的序列。为了维持物理流量的本地性，第一跳执行该一系列管道(具有所有要求的状态)并直接地将流量向物理网络中的最终出口位置发送。当使用快捷隧道时，虚拟化应用通过将快捷隧道网超出单个LDPS扩展到所有互连的LDP集合的端口的并集来利用逻辑L3数据路径互连逻辑L2数据路径。当所有操作都在第一跳执行时，第一跳元件通常能够访问分组所穿越的逻辑网络的所有状态。

[0016] 以上发明内容部分旨在用作对本发明的一些实施例的简要介绍。其并不意欲成为对本文档中公开的所有发明主题的介绍或概述。接下来的具体实施方式部分和具体实施方式部分中参考的附图将进一步描述发明内容部分中描述的实施例以及其它实施例。因此，为了理解本文档描述的所有实施例，需要全面查阅发明内容部分、具体实施方式部分和附

图。另外，要求保护的主题不受发明内容部分、具体实施方式部分和附图中的说明性细节限制，而应由所附权利要求来限定，因为要求保护的主题可以其它具体的形式实现，而不脱离这些主题的精神。

附图说明

[0017] 本发明的新颖特征在所附权利要求中阐明。然而，为了解释的目的，本发明的若干实施例在以下附图中阐明。

[0018] 图1概念性示出了一些实施例的网络体系结构。

[0019] 图2概念性示出了一些实施例的用于通过逻辑交换机和逻辑路由器处理网络数据的处理管道。

[0020] 图3概念性示出了在单个L3路由器中实现逻辑路由器的网络体系结构。

[0021] 图4概念性示出了在受管理交换元件中实现逻辑路由器的网络体系结构。

[0022] 图5概念性示出了以分布式方式实现路由器以使得若干受管理交换元件中的每一个在L3路由分组的网络体系结构。

[0023] 图6概念性示出了上文参考图2描述的逻辑处理管道的示例实现。

[0024] 图7概念性示出了一些实施例的用于通过逻辑交换机、逻辑路由器和逻辑交换机处理分组的逻辑处理管道。

[0025] 图8概念性示出了一些实施例的实现逻辑路由器和逻辑交换机的示例网络体系结构。

[0026] 图9概念性示出了一些实施例的实现逻辑路由器和逻辑交换机的示例网络体系结构。

[0027] 图10概念性示出了一些实施例的实现逻辑路由器和逻辑交换机的示例网络体系结构。

[0028] 图11概念性示出了一些实施例的包括受管理交换元件和L3的主机的示例体系结构。

[0029] 图12概念性示出了受管理交换元件和L3路由器中的逻辑交换机和逻辑路由器的示例实现。

[0030] 图13A-13C概念性示出了在上文参考图12描述的受管理交换元件和L3路由器中实现的逻辑交换机、逻辑路由器的示例操作。

[0031] 图14概念性示出了一些实施例执行来转发分组以确定向哪个受管理交换元件发送分组的过程。

[0032] 图15概念性示出了上文参考图8描述的主机。

[0033] 图16概念性示出了一过程，一些实施例在第一和第二L3路由器在同一主机中被实现时使用该过程来直接地将分组从第一L3路由器转发到第二L3路由器。

[0034] 图17概念性示出了上文参考图2描述的逻辑处理管道的示例实现。

[0035] 图18概念性示出了一些实施例的用于通过一逻辑交换机、一逻辑路由器和另一逻辑交换机处理分组的逻辑处理管道。

[0036] 图19概念性示出了一些实施例的实现逻辑路由器和逻辑交换机的示例网络体系结构。

[0037] 图20概念性示出了一些实施例的实现逻辑路由器和逻辑交换机的示例网络体系结构。

[0038] 图21概念性示出了一些实施例的实现逻辑路由器和逻辑交换机的示例网络体系结构。

[0039] 图22概念性示出了一些实施例的包括基于流条目实现逻辑路由器的受管理交换元件的主机的示例体系结构。

[0040] 图23概念性示出了受管理交换元件中的逻辑交换机和逻辑路由器的示例实现。

[0041] 图24概念性示出了上文参考图23描述的逻辑交换机、逻辑路由器和受管理交换元件的示例操作。

[0042] 图25概念性示出了上文参考图2描述的逻辑处理管道的示例实现。

[0043] 图26概念性示出了一些实施例的用于通过一逻辑交换机、一逻辑路由器和另一逻辑交换机处理分组的逻辑处理管道。

[0044] 图27概念性示出了一些实施例的实现逻辑路由器和逻辑交换机的示例网络体系结构。

[0045] 图28概念性示出了一些实施例的实现逻辑路由器和逻辑交换机的示例网络体系结构。

[0046] 图29概念性示出了对接收到的分组执行所有的L2和L3处理以转发和路由的第一跳交换元件的示例。

[0047] 图30A-30B概念性示出了上文参考图29描述的逻辑交换机、逻辑路由器和受管理交换元件的示例操作。

[0048] 图31概念性示出了受管理交换元件在其上运行的主机的示例软件体系结构。

[0049] 图32概念性示出了一些实施例执行来转换网络地址的过程。

[0050] 图33概念性示出了一些实施例的执行包括NAT操作在内的整个逻辑处理管道的第一跳交换元件。

[0051] 图34概念性示出了当向受管理交换元件发送返回分组时受管理交换元件不执行逻辑处理管道的示例。

[0052] 图35概念性示出了一些实施例执行来将分组发送到其地址被NAT的目的地机器的过程。

[0053] 图36示出了当VM从第一主机迁移到第二主机时将NAT状态从第一主机迁移到第二主机的示例。

[0054] 图37示出了当VM从第一主机迁移到第二主机时将NAT状态从第一主机迁移到第二主机的另一示例。

[0055] 图38示出了执行负载均衡的逻辑路由器和逻辑交换机的示例物理实现。

[0056] 图39示出了执行负载均衡的逻辑路由器和逻辑交换机的另一示例物理实现。

[0057] 图40示出了执行负载均衡的逻辑路由器和逻辑交换机的还一示例物理实现。

[0058] 图41概念性示出了在共同提供服务(例如web服务)的机器之间均衡负载的负载均衡守护进程。

[0059] 图42示出了为不同用户向不同逻辑网络提供DHCP服务的DHCP守护进程。

[0060] 图43示出了中央DHCP守护进程和若干本地DHCP守护进程。

[0061] 图44概念性示出了在最末跳交换元件处执行一些逻辑处理的示例。

[0062] 图45A-45B概念性示出了上文参考图44描述的逻辑交换机、逻辑路由器和受管理交换元件的示例操作。

[0063] 图46概念性示出了在最末跳交换元件处执行一些逻辑处理的示例。

[0064] 图47A-47B概念性示出了上文参考图46描述的逻辑交换机、逻辑路由器和受管理交换元件的示例操作。

[0065] 图48概念性示出了受管理交换元件在其上运行的主机的示例软件体系结构。

[0066] 图49概念性示出了一些实施例执行来解析网络地址的过程。

[0067] 图50示出了允许各自运行L3守护进程的若干主机(或VM)避免广播ARP请求的映射服务器。

[0068] 图51示出了一些实施例执行来维护包括IP和MAC地址的映射的映射表的过程。

[0069] 图52示出了一些实施例执行来维护包括IP和MAC地址的映射的映射表的过程。

[0070] 图53概念性示出了一些实施例的控制器实例通过利用诸如nLog的表映射处理器(未示出)对表执行表映射操作来生成流。

[0071] 图54示出了示例体系结构和用户接口。

[0072] 图55示出了上文参考图54描述的阶段之前的表。

[0073] 图56示出了在用户提供逻辑端口的识别符、与端口相关联的IP地址和网络掩码以向逻辑路由器添加逻辑端口之后的表。

[0074] 图57示出了一组表映射操作的结果。

[0075] 图58示出了一组表映射操作的结果。

[0076] 图59示出了在上文参考图54描述的阶段之后的表。

[0077] 图60示出了一组表映射操作的结果。

[0078] 图61示出了一组表映射操作的结果。

[0079] 图62示出了在上文参考图61描述的阶段之后添加到一些表的新的行。

[0080] 图63示出了在控制应用通过执行如上文参考图55-62描述的表映射操作来生成逻辑数据之后的体系结构。

[0081] 图64概念性示出了实现本发明的一些实施例的电子系统。

具体实施方式

[0082] 本发明的一些实施例提供了一种网络控制系统,该网络控制系统允许由物理网络的交换元件来实现逻辑数据路径(LDP)集合(例如逻辑网络)。为了实现LDP集合,一些实施例的网络控制系统从逻辑转发平面数据生成物理控制平面数据。物理控制平面数据随后被推送到受管理(managed)交换元件,在这里其通常被转换成允许受管理交换元件执行其转发决策的物理转发平面数据。基于物理转发数据,受管理交换元件可根据在物理控制平面数据中指明(specify)的逻辑处理规则来处理数据分组。

[0083] 单个逻辑数据路径集合提供了交换架构以互连多个逻辑端口,这些逻辑端口可附接到物理或虚拟端点。在一些实施例中,这种LDP集合和逻辑端口的创建和使用提供了与虚拟局域网(VLAN)相对应的逻辑服务模型。这个模型在一些实施例中将网络控制系统的操作限制到仅定义逻辑L2交换能力。然而,其它实施例将网络控制系统的操作延伸到逻辑L2交

换能力和逻辑L3交换能力两者。

[0084] 一些实施例的网络控制系统支持以下逻辑L3交换能力。

[0085] • 逻辑路由。代替对分组只执行L2交换，一些实施例的网络控制系统还定义了物理控制平面数据以引导受管理交换元件在跨越L2广播域(IP子网)时基于因特网协议(IP)地址来转发分组。这种逻辑L3路由解决了L2网络的可扩展性问题。

[0086] • 网关虚拟化(virtualization)。代替利用单纯的L2接口与外部网络接合，一些实施例的网络控制系统可使用IP接口来与外部网络交互。在一些实施例中，即使当存在去往和来自外部网络的多个物理出口(egress)和入口(ingress)点时，网络控制系统也通过定义单个逻辑网关来定义这种IP接口。从而，一些实施例通过使用网关虚拟化来与外部IP网络接合。

[0087] • 网络地址转换。可对整个L3子网进行网络地址转换(NAT'ed)。在一些实施例中，逻辑网络使用私有地址并且对于外部网络仅暴露经网络地址转换的IP地址。另外，在一些实施例中，逻辑网络的子网通过NAT互连或使用目的地NAT来实现细粒度应用级路由决策。

[0088] • 状态过滤。与NAT类似，一些实施例通过使用状态访问控制列表(ACL)来将子网与外部网络隔离。另外，一些实施例将ACL置于逻辑子网之间。

[0089] • 负载均衡。在一些情况下，逻辑网络用于提供服务。对于这些和其它情况，网络控制系统为应用集群(cluster)提供虚拟IP地址。在一些实施例中，网络控制系统指明使得能够在一组逻辑IP地址上散布到来的应用流量的负载均衡操作。

[0090] • DHCP。虽然可以设立虚拟机(VM)来在逻辑网络内提供动态IP地址分配服务，但服务提供商可能更偏好在基础设施级别的动态主机配置协议(DHCP)服务的高效实现。从而，一些实施例的网络控制系统在基础设施级别提供DHCP服务的高效实现。

[0091] 对于这些L3特征中的每一个的设计将在下文中描述。按照实现来说这些特征很大程度上是正交的，因此普通技术人员将会意识到这些特征不是都必须由一些实施例的网络控制系统提供。在进一步描述这些特征之前，应当提及若干假设。这些假设如下。

[0092] • 大型网络。跨越多个L2网络的逻辑L3网络将大于逻辑L2网络。一些实施例利用映射化简(map-reduce)分布式处理技术来为10K个服务器那么大的服务器集群解决逻辑L3问题。

[0093] • 物理流量非本地性(non-locality)。数据中心内的逻辑子网可在数据中心内交换巨大的流量。一些实施例在可能的程度上保留流量本地性。在上文提到的映射化简示例中，流量就端点而言不具有本地性。

[0094] • 逻辑流量本地性。当涉及到在逻辑子网之间交换的流量时，确实存在本地性。换言之，对于上文提到的映射化简集群，不是每一个逻辑网络都具有客户端。

[0095] • 功能的放置。如这里通过引用并入的美国专利申请13/177,535中提到的，受管理交换元件在一些实施例中为：(1)物理网络的边缘交换元件(即，与由物理网络连接的虚拟或物理计算设备具有直接连接的交换元件)，以及(2)插入在受管理交换元件层级中以简化和/或促成所控制的边缘交换元件的操作的非边缘交换元件。如美国专利申请13/177,535中进一步描述的，边缘交换元件在一些实施例中包括：(1)与由网络连接的虚拟或物理计算设备具有直接连接的交换元件，以及(2)将网络的第一受管理部分连接到网络的第二受管理部分(例如，与第一受管理部分不同的物理位置中的部分)或者连接到网络的未管理

部分(例如,连接到企业的内部网络)的集成元件(称为扩展器)。一些实施例理想地在第一受管理边缘交换元件处即在第一跳边缘交换元件处执行逻辑L3路由,这可在也容宿(host)着由物理网络互连的虚拟机的超管理器(hypervisor)中实现。理想情况下,第一跳交换元件执行全部或大部分L3路由,因为一些实施例的网络控制系统可将非边缘交换元件(内部网络)看作只不过是用于互连设备的架构。

[0096] 下文描述的实施例中的一些是在一种由用于管理一个或多个共享的转发元件的一个或多个控制器(下文也称为控制器实例)所形成的新型分布式网络控制系统中实现的。共享的转发元件在一些实施例中可包括虚拟或物理网络交换机、软件交换机(例如Open vSwitch)、路由器和/或其它交换设备,以及在这些交换机、路由器和/或其它交换设备之间建立连接的任何其它网络元件(例如负载均衡器等等)。这种转发元件(例如,物理交换机或路由器)在下文中也被称为交换元件。与现成的交换机不同,软件转发元件在一些实施例中是通过将其(一个或多个)交换表和逻辑存储在独立设备(例如独立的计算机)的存储器中来形成的交换机,而在其它实施例中,它是通过将其(一个或多个)交换表和逻辑存储在也执行超管理器和在该超管理器之上的一个或多个虚拟机的设备(例如计算机)的存储器中来形成的交换机。

[0097] 在一些实施例中,控制器实例允许系统接受来自用户的逻辑数据路径集合并对交换元件进行配置以实现这些逻辑数据路径集合。在一些实施例中,一种类型的控制器实例为执行一个或多个模块的设备(例如通用计算机),这些模块将用户输入从逻辑控制平面转换到逻辑转发平面,然后将逻辑转发平面数据转换成物理控制平面数据。这些模块在一些实施例中包括控制模块和虚拟化模块。控制模块允许用户指明和填充逻辑数据路径集合,而虚拟化模块通过将逻辑数据路径集合映射到物理交换基础设施上来实现指明的逻辑数据路径集合。在一些实施例中,控制和虚拟化应用是两个分开的应用,而在其它实施例中它们是同一应用的一部分。

[0098] 从对于特定逻辑数据路径集合的逻辑转发平面数据,一些实施例的虚拟化模块生成对于实现逻辑数据路径集合的任何受管理交换元件通用的通用物理控制平面(UPCP)数据。在一些实施例中,此虚拟化模块是作为该特定逻辑数据路径集合的主控制器的控制器实例的一部分。此控制器被称为逻辑控制器。

[0099] 在一些实施例中,UPCP数据随后被转换成针对每个特定受管理交换元件的定制物理控制平面(CPCP)数据,该转换由作为该特定受管理交换元件的主物理控制器实例的控制器实例进行,或者由该特定受管理交换元件的机箱控制器(chassis controller)进行,这在同时递交的标题为“Chassis Controller”且代理人案卷号为No.NCRA.P0081的美国专利申请＊＊中被进一步描述;这里通过引用并入该同时递交的美国专利申请。当该机箱控制器生成CPCP数据时,该机箱控制器通过物理控制器从逻辑控制器的虚拟化模块获得UPCP数据。

[0100] 无论是物理控制器还是机箱控制器生成CPCP数据,针对特定受管理交换元件的CPCP数据都需要被传播到该受管理交换元件。在一些实施例中,通过网络信息库(NIB)数据结构来传播CPCP数据,网络信息库数据结构在一些实施例中为面向对象的数据结构。使用NIB数据结构的若干示例在美国专利申请13/177,529和13/177,533中描述,这里通过引用并入这些美国专利申请。如这些申请中所述,NIB数据结构在一些实施例中也用来充当不同

的控制器实例之间的通信媒介，并且存储关于逻辑数据路径集合(例如逻辑交换元件)和/或实现这些逻辑数据路径集合的受管理交换元件的数据。

[0101] 然而，其它实施例不使用NIB数据结构来将CPCP数据从物理控制器或机箱控制器传播到受管理交换元件、在控制器实例之间通信以及存储关于逻辑数据路径集合和/或受管理交换元件的数据。例如，在一些实施例中，物理控制器和/或机箱控制器经由配置协议通过OpenFlow条目和更新与受管理交换元件通信。另外，在一些实施例中，控制器实例使用一个或多个直接通信信道(例如RPC调用)来交换数据。此外，在一些实施例中，控制器实例(例如这些实例的控制和虚拟化模块)按照被写入到关系数据库数据结构中的记录来表达逻辑和/或物理数据。在一些实施例中，此关系数据库数据结构是用于实现控制器实例的一个或多个模块的表映射引擎(称为nLog)的输入和输出表的一部分。

[0102] I.逻辑路由

[0103] 一些实施例在一些情况下将逻辑路由建模为由实现在L3域中操作的LDPS的逻辑路由器互连在L2域中操作的两个或更多个LDP集合的行为。从一个逻辑L2域穿越到另一个逻辑L2域的分组在一些实施例中将采取以下四个步骤。下文中按照网络控制系统实现的逻辑处理操作来描述这四个步骤。然而，要理解，这些操作是由网络的受管理交换元件基于由网络控制系统产生的物理控制平面数据来执行的。

[0104] 第一，将通过发端逻辑L2域的L2表管道来处理分组。该管道将以目的地媒体访问控制(MAC)地址被转发到与逻辑路由器的逻辑端口附接的逻辑端口结束。

[0105] 第二，将通过逻辑路由器的L3数据路径来处理分组，这再次通过通过此路由器的L3表管道发送它来完成。在一些实施例中在路由器的L3数据路径中跳过物理路由器中常见的L2查找阶段，因为逻辑路由器将只接收要求路由的分组。

[0106] 在一些实施例中，L3转发决策将使用由逻辑路由器的逻辑控制平面配设的前缀转发信息库(FIB)条目。在一些实施例中，控制应用用于接收逻辑控制平面数据，并将此数据转换成逻辑转发平面数据，该逻辑转发平面数据随后被提供给网络控制系统。对于L3转发决策，一些实施例使用前缀FIB条目来实现最长的前缀匹配。

[0107] 结果，L3路由器将把分组转发到与目的地L2 LDPS“连接”的逻辑端口。在将分组进一步转发到该LDPS之前，L3路由器将把发端MAC地址改变为在其域中所定义的那个以及把目的地IP地址解析成目的地MAC地址。该解析在一些实施例中由L3数据管道的最末的“IP输出”阶段(stage)执行。同一管道将递减TTL并更新检验和(并且如果TTL达到零则以ICMP响应)。

[0108] 应当注意，一些实施例在将经处理的分组馈送到下一LDPS之前改写MAC地址，因为如果没有这个改写，则在下一LDPS处可导致不同的转发决策。还应当注意，即使传统的路由器利用ARP来执行目的地IP地址的解析，但一些实施例在L3逻辑路由器中不将ARP用于此目的，因为只要下一跳为逻辑L2数据路径，这个解析就保持在虚拟化应用的内部。

[0109] 第三，将通过目的地逻辑L2域的L2表管道来处理分组。目的地L2表管道确定其应当发送分组的逻辑出口端口。在未知MAC地址的情况下，这个管道将通过依赖于一些分布式查找机制来解析MAC地址位置。在一些实施例中，受管理交换元件依赖于MAC学习算法，例如，它们洪泛(flood)未知分组。在这些或其它实施例中，MAC地址位置信息也可通过其它机制获得，例如通过带外获得。如果这种机制在一些实施例中可用，则最末的逻辑L2表管道使

用此机制来获得MAC地址位置。

[0110] 第四，分组被发送到表示逻辑端口附接(attachment)的附接到物理端口的逻辑端口。在这个阶段，如果端口是点对点媒体(例如，虚拟网络接口，VIF)，则除了将分组发送到该端口以外就没有什么要做的了。然而，如果最末的LDPS为L3路由器并且因此附接为物理L3子网，则附接点在一些实施例中在将分组发送出之前通过使用ARP来解析目的地IP地址。在该情况下，源MAC地址在VIF的情况下将是依出口而定(egress specific)的，而不是逻辑MAC接口地址。在其它实施例中，利用ARP解析目的地IP地址是在第二步骤期间由L3逻辑路由器执行的。

[0111] 在上述示例中，只存在互连逻辑L2数据路径的单个逻辑路由器，但没有什么限制拓扑。普通技术人员将会认识到，对于更丰富的拓扑可互连更多的LDP集合。

[0112] 在一些实施例中，控制应用允许按照指明逻辑L3管道的一个或多个表来定义L3特定的逻辑状态。管理LDPS管道的相应逻辑控制平面或者可依赖于静态路由配置，或者可通过标准的路由协议与其它LDP集合对等(peer)。

[0113] 在一些实施例中，虚拟化应用将上述四步L2/L3分组处理的物理实现定义成物理控制平面数据，该物理控制平面数据当被受管理交换元件转换成物理转发数据时，实现了全部或绝大多数在第一跳受管理边缘交换元件处执行的逻辑管道执行的序列。为了维持物理流量的本地性，第一跳执行该一系列管道(具有所有要求的状态)并直接将流量向物理网络中的最终出口位置发送。当使用快捷(cut short)隧道时，虚拟化应用通过将快捷隧道网(mesh)超出单个LDPS扩展到所有互连的LDP集合的端口的并集(union)来利用逻辑L3数据路径互连逻辑L2数据路径。

[0114] 当所有事情都在第一跳执行时，第一跳元件通常能够访问分组所穿越的逻辑网络的所有状态。第一跳交换元件处的逻辑管道的执行的状态的散播(dessemination)(及其扩展含义)在下文进一步描述。

[0115] 图1概念性示出了一些实施例的网络体系结构100。具体而言，此图示出了在两个LDP集合(例如逻辑网络)150和155之间路由分组的逻辑路由器105。如图所示，网络体系结构100包括逻辑路由器105、逻辑交换机110和115以及机器120、125、130、135、140和145。

[0116] 逻辑交换机110为美国专利申请13/177,535中描述的逻辑交换机(或逻辑交换元件)。逻辑交换机110是跨越若干受管理交换元件(未示出)而实现的。逻辑交换机110在L2(第2层)在机器120-130之间路由网络流量。也就是说，逻辑交换机110基于逻辑交换机110具有的一个或多个转发表(未示出)作出交换决策以在机器120-130之间在数据链路层路由网络数据。逻辑交换机110与若干其它逻辑交换机(未示出)一起为逻辑网络150路由网络流量。逻辑交换机115是另一逻辑交换机。逻辑交换机115为逻辑网络155在机器135-145之间路由流量。

[0117] 逻辑路由器在一些实施例中在不同的逻辑网络之间在L3(第3层一网络层)路由流量。具体而言，逻辑路由器基于一组路由表在两个或更多个逻辑交换机之间路由网络流量。在一些实施例中，逻辑路由器在单个受管理交换元件中实现，而在其它实施例中，逻辑路由器以分布式方式在若干不同的受管理交换元件中实现。这些不同实施例的逻辑路由器将在下文中进一步详细描述。逻辑路由器105在逻辑网络150和155之间在L3路由网络流量。具体而言，逻辑路由器105在两个逻辑交换机110和115之间路由网络流量。

[0118] 机器120、125、130、135、140和145是能够交换数据分组的机器。例如，每个机器120、125、130、135、140和145具有网络接口控制器(NIC)，使得在机器120、125、130、135、140和145上执行的应用能够通过逻辑交换机110和115和逻辑路由器105在它们之间交换数据。

[0119] 逻辑网络150和155的不同在于每个网络中的机器使用不同的L3地址。例如，逻辑网络150和155是用于一公司的两个不同部门的不同IP子网。

[0120] 在操作中，逻辑交换机110和115和逻辑路由器105像交换机和路由器那样工作。例如，逻辑交换机110路由源自机器120-130中的一个并前往机器120-130中的另一个的数据分组。当逻辑网络150中的逻辑交换机110接收到以逻辑网络155中的机器135-145中的一个为目的地的数据分组时，逻辑交换机110将该分组发送到逻辑路由器105。逻辑路由器105随后基于分组的头(header)中包括的信息来将该分组路由到逻辑交换机115。逻辑交换机115随后将分组路由到机器135-145中的一个。源自机器135-145中的一个的数据分组被逻辑交换机110和115以及逻辑路由器105以类似的方式路由。

[0121] 图1示出了在两个逻辑网络150和155之间路由数据的单个逻辑路由器。普通技术人员将会认识到，在两个逻辑网络之间可存在多于一个的在路由分组中所涉及的逻辑路由器。

[0122] 图2概念性示出了一些实施例的用于通过逻辑交换机和逻辑路由器处理网络数据的处理管道200。具体而言，处理管道200包括三个阶段205-215，用于分别通过逻辑交换机220、逻辑路由器225、然后是逻辑交换机230来处理数据分组。此图在图的上半部示出了逻辑路由器225以及逻辑交换机220和230，在图的下半部示出了处理管道200。

[0123] 逻辑路由器225与上文通过参考图1描述的逻辑路由器105的类似之处在于逻辑路由器225在逻辑交换机220和220之间路由数据分组。逻辑交换机220和230与逻辑交换机110和115类似。逻辑交换机220和230各自为逻辑网络在L2转发流量。

[0124] 当逻辑交换机220接收到分组时，逻辑交换机220执行逻辑处理管道200的阶段205(L2处理)以便在一个逻辑网络中转发分组。当分组以另一逻辑网络为目的地时，逻辑交换机220将该分组转发到逻辑路由器225。逻辑路由器225随后对分组执行逻辑处理管道200的阶段210(L3处理)以便在L3路由数据。逻辑路由器225将此分组发送到另一逻辑路由器(未示出)，或者，如果逻辑路由器225耦合到逻辑交换机230，则逻辑路由器225将分组发送到逻辑交换机230，逻辑交换机230将把分组直接发送到分组的目的地机器。直接将分组发送到分组的目的地的逻辑交换机230执行逻辑处理管道200的阶段215(L2处理)以便将分组转发到分组的目的地。

[0125] 在一些实施例中，逻辑交换机和逻辑路由器由一组受管理交换元件(未示出)来实现。一些实施例的这些受管理交换元件通过执行例如逻辑处理管道200的逻辑处理管道来实现逻辑交换机和逻辑路由器。一些实施例的受管理交换元件基于受管理交换元件中的流条目(entry)来执行逻辑处理管道。受管理交换元件中的流条目(未示出)由一些实施例的网络控制系统配置。逻辑处理管道200的更多细节将在下文中进一步描述。

[0126] 接下来的三幅图，图3、图4和图5概念性示出了一些实施例的逻辑交换机和逻辑路由器的若干实现。图3和图4示出了集中式(centralized)L3路由的两种不同实现，而图5示出了分布式L3路由。

[0127] 图3概念性示出了网络体系结构300。具体而言，图3示出了逻辑路由器225在单个

L3路由器360(例如硬件路由器或软件路由器)中实现。L3路由器360为不同的逻辑网络路由分组,其中每个逻辑网络包括在若干不同的受管理交换元件中实现的若干逻辑交换机。此图被水平地划分成分别表示逻辑和物理实现的左半部和右半部。此图还被垂直地划分成分别表示第2层和第3层的下半部和上半部。图3示出了网络体系结构300包括L3路由器360和受管理交换元件305、310、315和320。此图还示出了逻辑交换机220和230中的每一个逻辑地耦合到三个VM。

[0128] L3路由器360实现逻辑路由器225。L3路由器360在包括逻辑交换机220和230的不同逻辑网络之间路由分组。L3路由器360根据L3条目335来路由分组,其中L3条目335指明以何种方式在L3路由分组。例如,一些实施例的L3条目为路由表中的条目(例如,路由),这些条目指明具有落入IP地址的特定范围中的目的地IP地址的分组应当通过逻辑路由器225的特定物理逻辑端口被发送出去。在一些实施例中,逻辑路由器225的逻辑端口被映射到L3路由器的端口,并且逻辑路由器225基于映射来生成L3条目。逻辑路由器的端口到实现逻辑路由器的L3路由器的映射将在下文中进一步描述。

[0129] 一些实施例的受管理交换元件305-320以分布式方式实现逻辑交换机。也就是说,这些实施例中的逻辑交换机可跨越受管理交换元件305-320中的一个或多个来实现。例如,逻辑交换机220可跨越受管理交换元件305、310和315来实现,而逻辑交换机230可跨越受管理交换元件305、315和320来实现。逻辑地耦合到逻辑交换机220和230的六个VM 362、364、368、370、372和374如图所示耦合到受管理交换元件310-320。

[0130] 一些实施例的受管理交换元件305-320各自根据指明应当以何种方式在L2转发分组的L2流条目来转发分组。例如,L2流条目可指明具有特定目的地MAC地址的分组应当通过逻辑交换机的特定逻辑端口被发送出去。受管理交换元件305-320中的每一个具有一组L2流条目340(为了简单起见,交换元件305-315的流条目340没有被描绘)。每个受管理交换元件的L2流条目通过控制器集群被配置在该受管理交换元件中。通过配置受管理交换元件的L2流条目来配置受管理交换元件将在下文中进一步详细描述。

[0131] 一些实施例的受管理交换元件305是第二级受管理交换元件。第二级受管理交换元件为受管理非边缘交换元件,与受管理边缘交换元件对比,该受管理非边缘交换元件不直接向机器发送和从机器接收分组。第二级受管理交换元件促成非边缘受管理交换元件和边缘受管理交换元件之间的分组交换。美国专利申请13/177,535中描述的池节点(pool node)和扩展器也是第二级受管理交换元件。一些实施例的受管理交换元件305用作扩展器。也就是说,受管理交换元件305通信地桥接被一个或多个其它网络(未示出)分离的远程受管理网络(未示出)。

[0132] 一些实施例的受管理交换元件305通信地耦合到L3路由器360。当存在需要在L3路由的分组时,受管理交换元件310-320将分组发送到受管理交换元件305,使得L3路由器360在L3路由分组。关于在L3路由器中实现的集中式逻辑路由器的更多细节将在下文中参考图6-16来进一步描述。

[0133] 图4概念性示出了网络体系结构400。具体而言,图4示出了逻辑路由器225在受管理交换元件410中实现。与L3路由器360在L3路由分组的网络体系结构300对比,在网络体系结构400中受管理交换元件410在L3路由分组。此图被水平地划分成分别表示逻辑和物理实现的左半部和右半部。此图还被垂直地划分成分别表示第2层和第3层的下半部和上半部。

[0134] 除了网络体系结构400不包括L3路由器360，网络体系结构400与网络体系结构300类似。受管理交换元件410实现逻辑路由器225。也就是说，受管理交换元件410在包括逻辑交换机220和230的不同逻辑网络之间路由分组。一些实施例的受管理交换元件410根据指明应当以何种方式在L3路由分组的L3条目405来路由分组。然而，与一些实施例的L3条目335对比，L3条目405不是用于路由表的条目。相反，L3条目405是流条目。如美国专利申请13/177,535中所述，流条目包括限定符(qualifier)和动作，而路由表中的条目只是用于找出分组的下一跳的查找表。另外，L3流条目可指明生成路由表中的条目的方式(未示出)。

[0135] 除了实现集中式逻辑路由器以外，一些实施例的受管理交换元件410还实现跨越若干受管理交换元件实现的一个或多个逻辑交换机。受管理交换元件410因此具有其自己的一组L2流条目340(未描绘)。在体系结构400中，受管理交换元件410和310-320以分布式方式一起实现逻辑交换机220和230。

[0136] 一些实施例的受管理交换元件410从而实现集中式逻辑路由器和逻辑交换机二者。在其它实施例中，集中式逻辑路由器和逻辑交换机的实现可被分离到两个或更多个受管理交换元件中。例如，一个受管理交换元件(未示出)可利用流条目实现集中式逻辑路由器，而另一个受管理交换元件(未示出)可以分布式方式基于流条目实现逻辑交换机。关于基于流条目在受管理交换元件中实现的集中式逻辑路由器的更多细节将在下文中参考图17-24进一步描述。

[0137] 图5概念性示出了网络体系结构500。具体而言，图5示出了以分布式方式实现逻辑路由器225以使得若干受管理交换元件中的每一个在L3路由分组。图5示出了网络体系结构500包括四个受管理交换元件505、510、515和520。

[0138] 受管理交换元件505、510、515和520实现用于若干不同逻辑网络的若干逻辑交换机和逻辑路由器。一些实施例的受管理交换元件505、510、515和520中的每一个为边缘交换元件。也就是说，受管理交换元件具有耦合到该受管理交换元件的一个或多个机器。耦合到受管理交换元件的机器还逻辑地耦合到逻辑交换机。耦合到受管理交换元件的机器可以逻辑地耦合到同一逻辑交换机，或者可以不逻辑地耦合到同一逻辑交换机。

[0139] 受管理交换元件505、510、515和520中的每一个实现将会路由和转发去往和来自耦合到受管理交换元件的机器的分组的至少一个逻辑路由器和至少一个逻辑交换机。换言之，当受管理交换元件从耦合到受管理交换元件的机器接收分组时，受管理交换元件作出逻辑转发决策和逻辑路由决策。受管理交换元件505、510、515和520中的每一个根据逻辑流条目550中的L2条目和L3条目来作出逻辑转发和路由决策。逻辑流条目550包括一组L2流条目530和一组L3流条目535。关于分布式逻辑路由器的更多细节将在下文中参考图25-30B来进一步描述。

[0140] 图6-16示出了在路由器中实现的集中式逻辑路由器。图6概念性示出了上文参考图2描述的逻辑处理管道200的示例实现。图6示出了网络体系结构600。在网络体系结构600中，逻辑处理管道200由三个受管理交换元件615、620和625以及L3路由器635执行。具体地，L2处理205和L2处理215以分布式方式在受管理交换元件615、620和625上执行。L3处理210由L3路由器635来执行。图6还示出了源机器610和目的地机器630。

[0141] 受管理交换元件615是直接从耦合到边缘交换元件的机器接收分组的边缘交换元件。受管理交换元件615从源机器610接收分组。当受管理交换元件615接收到来自源机器

610的分组时，受管理交换元件615对分组执行L2处理205的一部分以便逻辑地转发分组。

[0142] 在受管理交换元件615和受管理交换元件620之间可以存在一个或多个受管理交换元件(未示出)。这些受管理交换元件具有网络构造(例如，PIF、VIF等等)，逻辑交换机220(图6中未示出)的逻辑构造(例如逻辑端口)被映射到这些网络构造。

[0143] 当分组要前往另一逻辑网络中的目的地机器630时，分组被转发到受管理交换元件620。受管理交换元件620随后执行L2处理205的剩余部分并将分组发送到L3路由器635，其实现集中式逻辑路由器(未示出)。

[0144] 与上文参考图3描述的L3路由器360类似，L3路由器635是端口被映射到逻辑路由器的端口的硬件路由器或软件路由器。L3路由器635对分组执行L3处理210以便逻辑地路由分组。也就是说，L3路由器635将分组发送到另一逻辑路由器(未示出)或发送到受管理交换元件620。

[0145] 受管理交换元件620是第二级受管理交换元件，其在一些实施例中用作扩展器。受管理交换元件620从L3路由器635接收分组并开始执行逻辑处理管道200的L2处理215。在受管理交换元件620和受管理交换元件625之间可以存在一个或多个受管理交换元件(未示出)。这些受管理交换元件具有网络构造，逻辑交换机230(图6中未示出)的逻辑构造被映射到这些网络构造。

[0146] 受管理交换元件625在该示例中从受管理交换元件620接收分组。受管理交换元件625对分组执行L2处理215的剩余部分以便逻辑地转发分组。在此示例中，受管理交换元件625还是将分组直接发送到目的地机器630的交换元件。然而，在受管理交换元件625和目的地机器630之间可以存在一个或多个受管理交换元件(未示出)。这些受管理交换元件具有网络构造，逻辑交换机230(图6中未示出)的逻辑构造被映射到这些网络构造。

[0147] 虽然在此示例中以分布式方式执行L2处理205和L2处理215，但L2处理205和L2处理215不是必须以分布式方式执行。例如，受管理交换元件615可执行整个L2处理205，并且受管理交换元件625可执行整个L2处理215。在这种情况下，受管理交换元件620将仅在L3路由器和受管理交换元件615和625之间中继分组。

[0148] 图7概念性示出了一些实施例的用于通过逻辑交换机220、逻辑路由器225和逻辑交换机230处理分组的逻辑处理管道200。具体而言，此图示出了当在上文参考图6描述的网络体系结构600中执行时的逻辑处理管道200。如上所述，在网络体系结构600中，L2处理205、L3处理210和L2处理215由受管理交换元件615、620和625以及L3路由器635执行。

[0149] L2处理205在一些实施例中包括八个阶段705-740，用于通过跨越受管理交换元件615和620实现的逻辑网络(未示出)中的逻辑交换机220(图7中未示出)来处理分组。在一些实施例中，接收分组的受管理交换元件615在受管理交换元件615接收到分组时执行L2处理205的一部分。受管理交换元件620随后执行L2处理205的剩余部分。

[0150] 在一些实施例中，分组包括头和有效负荷(payload)。头在一些实施例中包括一组字段(field)，这些字段包含用于通过网络路由分组的信息。逻辑交换机和逻辑路由器可基于头字段中包含的信息来确定交换/路由决策，并且在一些情况下可修改头字段中的一些或全部。

[0151] 在L2处理205的阶段705中，对分组执行入口上下文(context)映射以确定分组的逻辑上下文。在一些实施例中，阶段705在逻辑交换机220接收到分组(例如，分组最初被受

管理交换元件615接收)时执行。逻辑上下文在一些实施例中表示相对于逻辑交换机的分组的状态。逻辑上下文例如可指明分组所属的逻辑交换机、通过其接收分组的逻辑交换机的逻辑端口、通过其发送分组的逻辑交换机的逻辑端口、分组所处的逻辑交换机的逻辑转发平面的阶段,等等。

[0152] 一些实施例基于分组的源MAC地址(即,从其发送分组的机器)来确定分组的逻辑上下文。一些实施例基于分组的源MAC地址和分组的进入端口(例如入口端口)(即,通过其接收分组的受管理交换元件615的端口)来执行逻辑上下文查找。其它实施例可使用分组的头中的其它字段(例如,MPLS头、VLAN id等等)来确定分组的逻辑上下文。

[0153] 在执行第一阶段705之后,一些实施例将表示逻辑上下文的信息存储在分组的头的一个或多个字段中。这些字段也可称为逻辑上下文标签(tag)或逻辑上下文ID。另外,逻辑上下文标签在一些实施例中可与一个或多个已知的头字段(例如,VLAN id字段)一致。这样,这些实施例不以所定义的头字段被使用的方式来利用已知头字段或其伴随特征。可替代地,一些实施例将表示逻辑上下文的信息存储为与分组相关联并与分组一起传递的元数据(而不是存储在分组本身中)。

[0154] 在一些实施例中,第二阶段710是为逻辑交换机220定义的。在一些这样的实施例中,阶段710在分组的逻辑上下文上操作以相对于该逻辑交换机确定分组的入口访问控制。例如,当逻辑交换机接收到分组时,对分组应用入口ACL以控制分组对该逻辑交换机的访问。基于为逻辑交换机定义的入口ACL,可进一步处理分组(例如通过阶段715),或者例如可丢弃分组。

[0155] 在L2处理205的第三阶段715中,在逻辑交换机的上下文中对分组执行L2转发。在一些实施例中,第三阶段715在分组的逻辑上下文上操作以相对于逻辑交换机220处理并转发分组。例如,一些实施例定义用于在第2层处理分组的L2转发表或L2转发条目。

[0156] 另外,当分组的目的地在另一逻辑网络中时(即,当分组的目的地逻辑网络不同于其流量被逻辑交换机220处理的逻辑网络时),逻辑交换机220将分组发送到逻辑路由器225,逻辑路由器225随后将执行L3处理210以便将分组路由到目的地逻辑网络。从而,在第三阶段715,一些实施例的受管理交换元件615确定应当通过逻辑交换机的与逻辑路由器225相关联的逻辑端口(未示出)将分组转发到逻辑路由器225。在其它实施例中,受管理交换元件615不一定要确定分组是否应当被转发到逻辑路由器225。相反,分组将具有逻辑路由器225的端口的地址作为目的地地址并且受管理交换元件615根据该目的地地址来通过逻辑交换机的逻辑端口转发这个分组。

[0157] 在第四阶段720,执行出口上下文映射以识别与分组的逻辑转发的结果相对应的物理结果。例如,分组的逻辑处理可指明,分组要被从逻辑交换机220的一个或多个逻辑端口(例如,逻辑出口端口)发送出去。这样,出口上下文映射操作识别一个或多个受管理交换元件(包括受管理交换元件615和620)的与逻辑交换机220的特定逻辑端口相对应的物理端口。受管理交换元件615确定在前一阶段715确定的逻辑端口所映射到的物理端口(例如VIF)是受管理交换元件620的端口(未示出)。

[0158] L2处理205的第五阶段725基于在第四阶段720执行的出口上下文映射来执行物理映射。在一些实施例中,物理映射确定用于向在第四阶段720中确定的物理端口发送分组的操作。例如,一些实施例的物理映射确定与执行L2处理205的受管理交换元件615的一组端

口(未示出)中的一个或多个端口相关联的一个或多个队列(未示出),通过这一个或多个端口发送分组以便分组到达在第五阶段725中确定的物理端口。这样,受管理交换元件可沿着网络中的正确路径来转发分组以便分组到达所确定的物理端口。

[0159] 如图所示,L2处理205的第六阶段730由受管理交换元件620执行。第六阶段730与第一阶段705类似。阶段730在受管理交换元件620接收到分组时执行。在阶段730,受管理交换元件620查找分组的逻辑上下文并且确定剩下来要执行L2出口访问控制。

[0160] 一些实施例的第七阶段735是为逻辑交换机220定义的。一些这样的实施例的第七阶段735在分组的逻辑上下文上操作以相对于该逻辑交换机确定分组的出口访问控制。例如,可向分组应用出口ACL,以在对分组执行逻辑转发之后控制分组离开逻辑交换机220的访问。基于为逻辑交换机定义的出口ACL,可进一步处理分组(例如,从逻辑交换机的逻辑端口发送出去或发送到调度端口(dipatch port)以便进一步处理)或者例如可丢弃分组。

[0161] 第八阶段740与第五阶段725类似。在第八阶段740,受管理交换元件620确定受管理交换元件620的特定物理端口(未示出),逻辑交换机220的逻辑出口端口被映射到该特定物理端口。

[0162] L3处理210包括六个阶段745-761,用于通过由L3路由器635实现的逻辑交换机220(图7中未示出)来处理分组。如上所述,L3处理涉及执行一组逻辑路由查找以确定通过第3层网络向何处路由分组。

[0163] 第一阶段745在逻辑路由器225接收到分组时(即,当实现逻辑路由器225的L3路由器635接收到分组时)执行逻辑入口ACL查找以确定访问控制。下一阶段746对分组执行网络地址转换(NAT)。特别地,阶段746执行目的地NAT(DNAT)以将分组的目的地地址恢复回向分组的源机器隐藏的目的地机器的真实地址。在能够执行DANT时执行此阶段746。

[0164] 下一阶段750基于分组的L3地址(例如目的地IP地址)和路由表(例如包含L3条目)执行逻辑L3路由以确定一个或多个逻辑端口来通过第3层网络发送分组。由于逻辑路由器225是由L3路由器635实现的,所以路由表在L3路由器635中被配置。

[0165] 在第四阶段755,一些实施例的L3路由器635还对分组执行源NAT(SNAT)。例如,当能够执行源NAT时,L3路由器635将分组的源IP地址替换为不同的IP地址以便隐藏源IP地址。

[0166] 第五阶段760在逻辑路由器225通过在阶段740中确定的端口将分组路由出逻辑路由器225之前执行逻辑L3出口ACL查找以确定访问控制。L3出口ACL查找是基于分组的L3地址(例如源和目的地IP地址)来执行的。

[0167] 第六阶段761执行地址解析以便将目的地L3地址(例如目的地IP地址)转换成目的地L2地址(例如目的地MAC地址)。在一些实施例中,L3路由器635使用标准的地址解析(例如,通过发送出ARP请求或查找ARP缓存)来寻找与目的地IP地址相对应的目的地L2地址。

[0168] 当逻辑路由器225没有耦合到目的地逻辑网络时,逻辑交换机220朝着目的地逻辑网络将分组发送到另一逻辑路由器。当逻辑路由器225耦合到目的地逻辑网络时,逻辑交换机220将分组路由到目的地逻辑网络(即,为目的地逻辑网络转发分组的逻辑交换机)。

[0169] L2处理215在一些实施例中包括八个阶段765、770、775、780、785、790、795和798,用于通过跨越受管理交换元件620和625实现的另一逻辑网络(图7中未示出)中的逻辑交换机230来处理分组。在一些实施例中,接收到分组的受管理网络中的受管理交换元件625在

受管理交换元件625从受管理交换元件620接收到分组时执行L2处理215。除了阶段765、770、775、780、785、790、795和798是由逻辑交换机230执行的(即,由实现逻辑交换机230的受管理交换元件620和625执行),阶段765、770、775、780、785、790、795和798分别与阶段705-740类似。也就是说,阶段765、770、775、780、785、790、795和798被执行以将从L3路由器635接收的分组通过受管理交换元件620和625转发到目的地。

[0170] 图8概念性示出了一些实施例的实现逻辑路由器225和逻辑交换机220和230的示例网络体系结构800。具体而言,网络体系结构800表示实现逻辑网络的物理网络,这些逻辑网络的数据分组被逻辑路由器225和逻辑交换机220和230交换和/或路由。该图在其上半部示出了逻辑路由器225和逻辑交换机220和230。该图在其下半部示出了L3路由器860。下半部还示出了分别在主机890、880和885(例如,由诸如Windows™和Linux™之类的操作系统操作的机器)中运行的第二级受管理交换元件810、受管理交换元件815和820。该图在其上部和下部都示出了VM 1-4。

[0171] 在此示例中,逻辑交换机220在逻辑路由器225、VM 1和VM 2之间转发数据分组。逻辑交换机230在逻辑路由器225、VM 3和VM 4之间转发数据分组。如上所述,逻辑路由器225在逻辑交换机220和230以及可能其它逻辑路由器和交换机(未示出)之间路由数据分组。逻辑交换机220和230和逻辑路由器225通过逻辑端口(未示出)逻辑地耦合并且通过逻辑端口交换分组。这些逻辑端口被映射到L3路由器860和受管理交换元件810、815和820的物理端口。

[0172] 在一些实施例中,逻辑交换机220和230中的每一个跨越受管理交换元件815和820以及可能其它受管理交换元件(未示出)来实现。在一些实施例中,逻辑路由器225是在通信地耦合到受管理交换元件810的L3路由器860中实现的。

[0173] 在此示例中,受管理交换元件810、815和820是分别在主机890、880和885中运行的软件交换元件。受管理交换元件810、815和820具有实现逻辑交换机220和230的流条目。利用这些流条目,受管理交换元件815和820在网络中的耦合到受管理交换元件810、815和820的网络元件之间路由网络数据(例如分组)。例如,受管理交换元件815在VM 1和3以及第二级受管理交换元件810之间路由网络数据。类似地,受管理交换元件820在VM 2和4以及第二级受管理交换元件810之间路由网络数据。如图所示,受管理交换元件815和820各自具有三个端口(描绘为带编号的方形),通过这些端口与耦合到受管理交换元件815和820的网络元件交换数据分组。

[0174] 受管理交换元件810与上文参考图3描述的受管理交换元件305的类似之处在于受管理交换元件810是用作扩展器的第二级受管理交换元件。受管理交换元件810与L3路由器860在同一主机中运行,L3路由器860在此示例中是软件路由器。

[0175] 在一些实施例中,网络控制系统(未示出)建立隧道来促成网络元件之间的通信。例如,受管理交换元件810通过隧道耦合到在主机880中运行的受管理交换元件815,该隧道如图所示端接(terminate)于受管理交换元件815的端口2处。类似地,受管理交换元件810通过端接于受管理交换元件820的端口1处的隧道耦合到受管理交换元件820。

[0176] 在不同实施例中支持不同类型的隧道协议。隧道协议的示例包括无线接入点的控制和配置(CAPWAP)、通用路由封装(GRE)、GRE因特网协议安全性(IPsec),以及其它类型的隧道协议。

[0177] 在此示例中，主机880和885中的每一个如图所示包括受管理交换元件和若干VM。VM 1-4是虚拟机，这些虚拟机每个被指派了一组网络地址(例如，用于L2的MAC地址、用于L3的IP地址，等等)并且可向其它网络元件发送和从其它网络元件接收网络数据。VM由运行在主机880和885上的超管理器(未示出)来管理。

[0178] 现在将描述通过网络体系结构800进行的若干示例数据交换。当耦合到逻辑交换机220的VM 1向也耦合到同一逻辑交换机220的VM 2发送分组时，该分组首先被发送到受管理交换元件815。受管理交换元件815随后对分组执行L2处理205，因为受管理交换元件815是从VM 1接收分组的边缘交换元件。对此分组的L2处理205的结果将指示分组应当被发送到受管理交换元件820以通过受管理交换元件820的端口4到达VM 2。因为VM 1和2在同一逻辑网络中并且因此对于分组的L3路由不是必须的，所以不需要对此分组执行L3处理。然后经由在受管理交换元件815和受管理交换元件820之间桥接的第二级受管理交换元件810将分组发送到受管理交换元件820。分组通过受管理交换元件820的端口4到达VM 2。

[0179] 当耦合到逻辑交换机220的VM 1向耦合到逻辑交换机230的VM 3发送分组时，该分组首先被发送到受管理交换元件815。受管理交换元件815对分组执行L2处理的一部分。然而，因为该分组被从一个逻辑网络发送到另一逻辑网络(即，分组的逻辑L3目的地地址是针对另一逻辑网络的)，所以需要对此分组执行L3处理。

[0180] 受管理交换元件815将分组发送到第二级受管理交换元件810，使得受管理交换元件810对分组执行L2处理的剩余部分以将分组转发到L3路由器860。在L3路由器860处执行的L3处理的结果将指示分组应当被发送回受管理交换元件810。受管理交换元件810随后执行另一L2处理的一部分并且将从L3路由器860接收的分组转发回受管理交换元件815。受管理交换元件815对从受管理交换元件810接收的分组执行L2处理215，并且此L2处理的结果将指示分组应当通过受管理交换元件815的端口5被发送到VM 3。

[0181] 当耦合到逻辑交换机220的VM 1向耦合到逻辑交换机230的VM 4发送分组时，该分组首先被发送到受管理交换元件815。受管理交换元件815对分组执行L2处理205。然而，因为该分组被从一个逻辑网络发送到另一逻辑网络，所以需要执行L3处理。

[0182] 受管理交换元件815经由受管理交换元件810将分组发送到L3路由器860，使得L3路由器860对分组执行L3处理210。在L3路由器860处执行的L3处理210的结果将指示分组应当被发送到受管理交换元件820。受管理交换元件810随后对从受管理交换元件接收的分组执行L2处理的一部分，并且此L2处理的结果将指示分组应当通过受管理交换元件820被发送到VM 4。受管理交换元件820执行L2处理的剩余部分以确定分组应当通过受管理交换元件820的端口5被发送到VM 4。

[0183] 图9概念性示出了一些实施例的实现逻辑路由器225和逻辑交换机220和230的示例网络体系结构900。具体而言，网络体系结构900表示实现逻辑网络的物理网络，这些逻辑网络的数据分组被逻辑路由器225和逻辑交换机220和230交换和/或路由。该图在其上半部示出了逻辑路由器225和逻辑交换机220和230。该图在其下半部示出了L3路由器860。下半部还示出了分别在主机910、890、880和885中运行的第二级受管理交换元件905、第二级受管理交换元件810以及受管理交换元件815和820。该图在其上部和下部都示出了VM 1-4。

[0184] 除了网络体系结构900附加地包括在主机910中运行的受管理交换元件905网络体系结构900与网络体系结构800类似，。一些实施例的受管理交换元件905是用作池节点的第

二级受管理交换元件。

[0185] 在一些实施例中，网络控制系统(未示出)建立隧道来促成网络元件之间的通信。例如，受管理交换元件815在此示例中通过隧道耦合到在主机910中运行的受管理交换元件905，该隧道如图所示端接于受管理交换元件815的端口1处。类似地，受管理交换元件820通过端接于受管理交换元件820的端口2处的隧道耦合到受管理交换元件905。另外，受管理交换元件905和810如图所示通过隧道耦合。

[0186] 如上文参考图8所述，逻辑路由器225和逻辑交换机220和230在L3路由器860以及受管理交换元件810、815和820中实现，除了在数据分组交换中涉及第二级受管理交换元件905。也就是说，受管理交换元件815和810通过受管理交换元件905来交换分组。

[0187] 图10概念性示出了一些实施例的实现逻辑路由器225和逻辑交换机220和230的示例网络体系结构1000。除了存在在受管理交换元件810和受管理交换元件820之间建立的隧道，网络体系结构1000与网络体系结构800类似。此图示出了一些实施例的网络体系结构1000是网络体系结构800和网络体系结构900的混合。也就是说，一些受管理边缘交换元件具有去往与集中式L3路由器耦合的第二级受管理交换元件的隧道，而其它受管理边缘交换元件必须通过用作池节点的第二级受管理交换元件，以便与耦合到集中式L3路由器的第二级受管理交换元件交换分组。

[0188] 图11概念性示出了一些实施例的包括受管理交换元件810和L3路由器860(未示出)的主机890的示例体系结构。具体而言，此图示出了L3路由器860在主机890的命名空间中被配置。主机890在一些实施例中是由能够创建命名空间和虚拟机的操作系统(例如Linux)管理的机器。如图所示，主机890在此示例中包括受管理交换元件810、命名空间1120和NIC 1145。此图还示出了控制器集群1105。

[0189] 控制器集群1105是管理包括受管理交换元件810在内的网络元件的一组网络控制器或控制器实例。受管理交换元件810在此示例中是在主机890中实现的包括用户空间1112和内核1110的软件交换元件。受管理交换元件810包括在用户空间1112中运行的控制守护进程(daemon)1115;以及在内核1110中运行的控制器补丁(patch)1130和网桥1135。用户空间1112和内核1110在一些实施例中具有主机890的操作系统，而在其它实施例中用户空间1112和内核1110具有在主机890上运行的虚拟机。

[0190] 在一些实施例中，控制器集群1105与控制守护进程1115通信(例如通过利用OpenFlow协议或另一通信协议)，控制守护进程1115在一些实施例中是在用户空间1112的后台中运行的应用。控制守护进程1115与控制器集群1105通信以便处理和路由受管理交换元件810接收的分组。具体而言，控制守护进程1115在一些实施例中从控制器集群1105接收配置信息并且配置控制器补丁1130。例如，控制守护进程1115从控制器集群1105接收关于用于处理和路由受管理交换元件810接收的分组的操作的命令。

[0191] 控制守护进程1115还接收用于控制器补丁1130的配置信息以设立连接到在命名空间1120中实现的逻辑路由器(未示出)的端口(未示出)，以使得该逻辑路由器利用适当的条目来填充路由表和其它表。

[0192] 控制器补丁1130是在内核1110中运行的模块。在一些实施例中，控制守护进程1115配置控制器补丁1130。当被配置时，控制器补丁1130包含关于对要接收的分组进行处理和转发的规则(例如流条目)。一些实施例的控制器补丁1130还创建一组端口(例如VIF)

以与命名空间1120交换分组。

[0193] 控制器补丁1130从内核1110的网络堆栈1150或从网桥1135接收分组。控制器补丁1130基于关于处理和路由分组的规则来确定向哪个命名空间发送分组。控制器补丁1130还从命名空间1120接收分组并基于规则将分组发送到网络堆栈1150或网桥1135。关于受管理交换元件的体系结构的更多细节在美国专利申请13/177,535中描述。

[0194] 命名空间1120(例如Linux命名空间)是在主机890中创建的容器。命名空间1120可实现网络堆栈、网络设备、网络地址、路由表、网络地址转换表、网络缓存等等(并非所有这些都在图11中示出)。命名空间1120从而在命名空间被配置为处理具有逻辑源或目的地地址的分组时可实现逻辑路由器。例如可通过配置命名空间的路由表1155来将命名空间1120配置为处理这种分组。在一些实施例中,随着命名空间1120连接到受管理交换元件810并交换分组(即,动态路由),命名空间1120填充路由表1155。在其它实施例中,控制器集群1105可通过利用路由填充路由表1155来直接配置路由表1155。

[0195] 另外,命名空间在一些实施例中还对命名空间路由的分组执行网络地址转换(NAT)。例如,当命名空间将接收到的分组的源网络地址改变成另一网络地址(即,执行源NAT)时。

[0196] 网桥1135在网络堆栈1150和主机外部的网络主机之间路由网络数据(即,通过NIC 1145接收的网络数据)。如图所示,网桥1135在网络堆栈1150和NIC 1145之间以及控制器补丁1130和NIC 1145之间路由网络数据。一些实施例的网桥1135执行标准的L2分组学习和路由。

[0197] 网络堆栈1150可通过NIC 1145从受管理交换元件810外部的网络主机接收分组。网络堆栈1150随后将分组发送到控制器补丁1130。在一些情况下,通过隧道从受管理交换元件外部的网络主机接收分组。在一些实施例中,隧道端接于网络堆栈1150。从而,当网络堆栈1150通过隧道接收分组时,网络堆栈1150拆开隧道头(即,解封出有效负荷)并将拆开的分组发送到控制器补丁1130。

[0198] 现在将描述受管理交换元件810和命名空间1120的示例操作。在此示例中,在受管理交换元件810与主机890外部的受管理交换元件815和820(在图11中未示出)之间建立隧道。也就是说,受管理交换元件810、815和820如图8所示通过隧道连接。隧道端接于网络堆栈1150。

[0199] 受管理交换元件815向受管理交换元件810发送分组,该分组是由VM 1发送到VM 4的。该分组被NIC 1145接收,然后被发送到网桥1135。基于分组头中的信息,网桥1135确定该分组是通过建立的隧道发送的,并且将分组发送到网络堆栈1150。网络堆栈1150拆开隧道头,并将拆开的分组发送到控制器补丁1130。

[0200] 根据控制器补丁1130具有的规则,控制器补丁1130将分组发送到命名空间1120,因为分组是从一个逻辑网络发送到另一逻辑网络的。例如,规则可以称,具有特定目的地MAC地址的分组应当被发送到命名空间1120。在一些情况下,控制器补丁1130在将分组发送到命名空间之前从分组去除逻辑上下文。命名空间1120随后对分组执行L3处理以在两个逻辑网络之间路由分组。

[0201] 通过执行L3处理,命名空间1120确定分组应当被发送到控制器补丁1130,因为目的地网络层地址应当去往属于目的地逻辑网络的逻辑交换机。控制器补丁1130接收分组并

通过网络堆栈1150、网桥1135和NIC 1145经由隧道将分组发送到实现属于目的地逻辑网络的逻辑交换机的受管理交换元件820。

[0202] 如上所述，一些实施例在命名空间1120中实现L3路由器860。然而，其它实施例可在运行在主机890上的VM中实现L3路由器860。

[0203] 图12概念性示出了受管理交换元件和L3路由器中的逻辑交换机和逻辑路由器的示例实现。具体而言，此图示出了在包括第二级受管理交换元件810和L3路由器860的主机890以及受管理交换元件815和820中实现逻辑路由器225以及逻辑交换机220和230。该图在其左半部示出了逻辑路由器225以及逻辑交换机220和230。该图在其右半部示出了第二级受管理交换元件810以及受管理交换元件815和820。该图在其右半部和左半部都示出了VM 1-4。为了简单起见，该图没有示出受管理交换元件的所有组件，例如网络堆栈1150。

[0204] 逻辑交换机220和230以及逻辑路由器225通过逻辑端口逻辑地耦合。如图所示，逻辑交换机220的逻辑端口X耦合到逻辑路由器225的逻辑端口1。类似地，逻辑交换机230的逻辑端口Y耦合到逻辑路由器225的逻辑端口2。逻辑交换机220和230通过这些逻辑端口与逻辑路由器225交换数据分组。另外，在此示例中，逻辑交换机220将逻辑端口X与MAC地址01:01:01:01:01:01相关联，该MAC地址是逻辑路由器225的逻辑端口1的MAC地址。当逻辑交换机220接收到需要L3处理的分组时，逻辑交换机220通过端口X将该分组送出到逻辑路由器225。类似地，逻辑交换机230将逻辑端口Y与MAC地址01:01:01:01:01:02相关联，该MAC地址是逻辑路由器225的逻辑端口2的MAC地址。当逻辑交换机230接收到需要L3处理的分组时，逻辑交换机230通过端口Y将该分组送出到逻辑路由器225。

[0205] 在此示例中，控制器集群1105(图12中未示出)配置受管理交换元件810，以使得受管理交换元件810的端口1与相同的MAC地址01:01:01:01:01:01相关联，而该MAC地址01:01:01:01:01:01与逻辑交换机220的端口X相关联。从而，当受管理交换元件810接收到以此MAC地址作为目的地MAC地址的分组时，受管理交换元件810通过受管理交换元件810的端口1将分组送出到L3路由器860(在命名空间1120中配置)。这样，逻辑交换机220的端口X被映射到受管理交换元件810的端口1。

[0206] 类似地，受管理交换元件810的端口2与相同的MAC地址01:01:01:01:01:02相关联，而该MAC地址01:01:01:01:01:02与逻辑交换机230的端口Y相关联。从而，当受管理交换元件810接收到以此MAC地址作为目的地MAC地址的分组时，受管理交换元件810通过受管理交换元件810的端口2将分组送出到L3路由器860。这样，逻辑交换机230的端口Y被映射到受管理交换元件810的端口2。

[0207] 在此示例中，逻辑路由器225具有逻辑端口1和2以及其它逻辑端口(未示出)。逻辑路由器225的端口1与IP地址1.1.1.1/24相关联，该IP地址表示端口1后面的子网。也就是说，当逻辑路由器225接收到要路由的分组并且该分组具有目的地IP地址例如1.1.1.10时，逻辑路由器225通过端口1将此分组朝着目的地逻辑网络(例如逻辑子网)发送。

[0208] 类似地，逻辑路由器225的端口2在此示例中与IP地址1.1.2.1/24相关联，该IP地址表示端口2后面的子网。逻辑路由器225通过端口2将具有目的地IP地址例如1.1.2.10的分组发送到目的地逻辑网络。

[0209] 在此示例中，L3路由器860通过利用路由填充L3路由器860的路由表(未示出)来实现逻辑路由器225。在一些实施例中，L3路由器860在受管理交换元件810与L3路由器860建

立连接并发送分组时填充其路由表。例如，当L3路由器从受管理交换元件接收到初始分组时，L3路由器860查明以该初始分组的源地址作为目的地地址的分组应当被发送到受管理交换元件810。L3路由器也可执行的标准地址解析(例如通过发送出ARP请求)来查明向何处发送初始分组。L3路由器860将在路由表中存储这些“路由”并且在为L3路由器随后接收的分组作出路由决策时查找这些表。其它L3路由器(未示出)可以类似的方式来填充其路由表。

[0210] 在其它实施例中，控制器集群配置L3路由器860的路由表，使得L3路由器860的端口1与和逻辑路由器225的端口1相关联的相同IP地址相关联。类似地，L3路由器860的端口2与和逻辑路由器225的端口2相关联的相同IP地址相关联。以类似的方式，在受管理交换元件的另一逻辑路由器(未示出)中可实现另一逻辑交换机(未示出)。在这些实施例中的一些中，控制集群可采用一个或多个路由协议来配置L3路由器。

[0211] 图13A-13C概念性示出了上文参考图12描述的受管理交换元件810、815和820以及L3路由器860中实现的逻辑交换机220和230、逻辑路由器225的示例操作。具体而言，图13A-13C示出了从VM 1发送到VM 4的分组如何到达VM 4。

[0212] 当耦合到逻辑交换机220的VM 1向耦合到逻辑交换机230的VM 4发送分组1330时，该分组首先通过受管理交换元件815的端口4被发送到受管理交换元件815。受管理交换元件815对分组执行L2处理。

[0213] 如图13A的上半部所示，受管理交换元件815包括转发表，该转发表包括用于处理和转发分组1330的规则(例如流条目)。当受管理交换元件815通过受管理交换元件815的端口4接收到来自VM 1的分组1330时，受管理交换元件815基于受管理交换元件815的转发表开始处理分组1330。在此示例中，分组1330具有目的地IP地址1.1.2.10，这是VM 4的IP地址。分组1330的源IP地址是1.1.1.10。分组1330还以VM 1的MAC地址作为源MAC地址并且以逻辑路由器225的逻辑端口1的MAC地址(即，01:01:01:01:01:01)作为目的地MAC地址。

[0214] 受管理交换元件815识别转发表中的实现阶段1340的上下文映射的由带圈的1指示的记录(称为“记录1”)。记录1基于进入端口识别分组1330的逻辑上下文，该进入端口是通过其从VM 1接收分组1330的端口4。此外，在一些实施例中，记录1指明受管理交换元件815将分组1330的逻辑上下文存储在分组1330的头的一组字段(例如，VLAN id字段)中。在其它实施例中，受管理交换元件815将逻辑上下文(即，分组所属的逻辑交换机以及该逻辑交换机的逻辑入口端口)存储在交换机的寄存器或元字段中，而不是存储在分组中。记录1还指明通过转发表来进一步处理分组1330(例如通过将分组1330发送到调度端口)。调度端口在美国专利申请13/177,535中描述。

[0215] 基于分组1330的头中存储的逻辑上下文和/或其它字段，受管理交换元件815识别转发表中的实现阶段1342的入口ACL的由带圈的2指示的记录(称为“记录2”)。在此示例中，记录2允许分组1330被进一步处理(即，分组1330可通过逻辑交换机220的入口端口)，从而指明通过转发表来进一步处理分组1330(例如，通过将分组1330发送到调度端口)。此外，记录2指明受管理交换元件815将分组1330的逻辑上下文(即，分组1330已被处理管道1300的第二阶段1342处理)存储在分组1330的头的该组字段中。

[0216] 接下来，受管理交换元件815基于分组1330的头中存储的逻辑上下文和/或其它字段来识别转发表中的实现阶段1344的逻辑L2转发的由带圈的3指示的记录(称为“记录3”)。

记录3指明具有逻辑路由器225的逻辑端口1的MAC地址作为目的地MAC地址的分组要被发送到逻辑交换机220的逻辑端口X。

[0217] 记录3还指明通过转发表来进一步处理分组1330(例如,通过将分组1330发送到调度端口)。另外,记录3指明受管理交换元件815将逻辑上下文(即,分组1330已被处理管道1300的第三阶段1344处理)存储在分组1330的头的该组字段中。

[0218] 基于分组1330的头中存储的逻辑上下文和/或其它字段,受管理交换元件815识别转发表中的实现阶段1346的上下文映射的由带圈的4指示的记录(称为“记录4”)。在此示例中,记录4将L3路由器860的端口1所耦合到的受管理交换元件810的端口1识别为与分组1330要被转发到的逻辑交换机220的逻辑端口X相对应的端口。记录4附加地指明通过转发表来进一步处理分组1330(例如通过将分组1330发送到调度端口)。

[0219] 基于分组1330的头中存储的逻辑上下文和/或其它字段,受管理交换元件815随后识别转发表中的实现阶段1348的物理映射的由带圈的5指示的记录(称为“记录5”)。记录5指明,为了分组1330到达受管理交换元件810,分组1330要通过受管理交换元件815的端口1来发送。在此情况下,受管理交换元件815将把分组1330从与受管理交换元件810耦合的受管理交换元件815的端口1发送出去。

[0220] 如图13A的下半部所示,受管理交换元件810包括转发表,该转发表包括用于处理和路由分组1330的规则(例如流条目)。当受管理交换元件810从受管理交换元件815接收到分组1330时,受管理交换元件810基于受管理交换元件810的转发表开始处理分组1330。受管理交换元件810识别转发表中的实现阶段1350的上下文映射的由带圈的1指示的记录(称为“记录1”)。记录1基于分组1330的头中存储的逻辑上下文来识别分组1330的逻辑上下文。逻辑上下文指明分组1330已被第二和第三阶段1342和1344处理,第二和第三阶段1342和1344由受管理交换元件815执行。这样,记录1指明通过转发表来进一步处理分组1330(例如通过将分组1330发送到调度端口)。

[0221] 接下来,受管理交换元件810基于分组1330的头中存储的逻辑上下文和/或其它字段来识别转发表中的实现阶段1352的出口ACL的由带圈的2指示的记录(称为“记录2”)。在此示例中,记录2允许分组1330被进一步处理(例如,分组1330可通过逻辑交换机220的端口“X”离开逻辑交换机220),并且从而指明通过转发表来进一步处理分组1330(例如,通过将分组1330发送到调度端口)。此外,记录2指明受管理交换元件810将分组1330的逻辑上下文(即,分组1330已被处理管道1300的阶段1352处理)存储在分组1330的头的该组字段中。

[0222] 接下来,受管理交换元件810基于分组1330的头中存储的逻辑上下文和/或其它字段来识别转发表中的实现阶段1354的物理映射的由带圈的3指示的记录(称为“记录3”)。记录3指明为了分组1330到达L3路由器860要通过其发送分组1330的受管理交换元件810的端口。在此情况下,受管理交换元件810将把分组1330从与L3路由器860的端口1耦合的受管理交换元件810的端口1发送出去。在一些实施例中,受管理交换元件810在将分组1330发送到L3路由器860之前从分组1330去除逻辑上下文。

[0223] 如图13B的上半部所示,L3路由器860包括入口ACL表、路由表和出口ACL表,其包括用于处理和路由分组1330的条目。当L3路由器860从受管理交换元件810接收到分组1330时,L3路由器860基于L3路由器860中的这些表开始处理分组1330。L3路由器860识别入口ACL表中的由带圈的1指示的条目(称为“条目1”),该条目通过指明L3路由器860应当基于分

组1330的头中的信息接受分组来实现L3入口ACL。L3路由器860随后识别路由表中的由带圈的2指示的条目(称为“条目2”),该条目通过指明具有其目的地IP地址(即1.1.2.10)的分组1330应当通过逻辑路由器225的端口2被发送到逻辑交换机230来实现L3路由1358。L3路由器860随后识别出口ACL表中的由带圈的3指示的条目(称为“条目3”),该条目通过指明L3路由器860可基于分组1330的头中的信息通过逻辑路由器225的端口2将分组送出来实现L3出口ACL。另外,L3路由器860将分组1330的源MAC地址改写成L3路由器860的端口2的MAC地址(即,01:01:01:01:01:02)。

[0224] L3路由器860随后执行地址解析以将目的地IP地址转换成目的地MAC地址。在此示例中,L3路由器860查找ARP缓存以找到目的地IP地址所映射到的目的地MAC地址。如果ARP缓存对于目的地IP地址不具有相应的MAC地址,则L3路由器860可发送出ARP请求。目的地IP地址将被解析成VM 4的MAC地址。L3路由器860随后利用目的地IP地址被解析成的MAC地址来改写分组1330的目的地MAC。L3路由器860将基于新的目的地MAC地址通过L3路由器860的逻辑端口2将把分组1330发送到逻辑交换机230。

[0225] 如图13B的下半部所示,受管理交换元件810包括转发表,该转发表包括用于处理和转发分组1330的规则(例如流条目)。当受管理交换元件810通过受管理交换元件810的端口2从L3路由器860接收到分组1330时,受管理交换元件810基于受管理交换元件810的转发表开始处理分组1330。受管理交换元件810识别转发表中的实现阶段1362的上下文映射的由带圈的4指示的记录(称为“记录4”)。记录4基于进入端口识别分组1330的逻辑上下文,该进入端口是通过其从L3路由器860接收分组1330的端口2。此外,记录4指明受管理交换元件810把分组1330的逻辑上下文存储在分组1330的头的一组字段(例如,VLAN id字段)中。记录4还指明通过转发表来进一步处理分组1330(例如通过将分组1330发送到调度端口)。

[0226] 基于分组1330的头中存储的逻辑上下文和/或其它字段,受管理交换元件810识别转发表中的实现阶段1364的入口ACL的由带圈的5指示的记录(称为“记录5”)。在此示例中,记录5允许分组1330被进一步处理,并且从而指明通过转发表来进一步处理分组1330(例如,通过将分组1330发送到调度端口)。此外,记录5指明受管理交换元件810将分组1330的逻辑上下文(即,分组1330已被处理管道1300的阶段1362处理)存储在分组1330的头的该组字段中。

[0227] 接下来,受管理交换元件810基于分组1330的头中存储的逻辑上下文和/或其它字段来识别转发表中的实现阶段1366的逻辑L2转发的由带圈的6指示的记录(称为“记录6”)。记录6指明以VM 4的MAC地址为目的地MAC地址的分组应当通过逻辑交换机230的逻辑端口(未示出)来转发。

[0228] 记录6还指明,通过转发表来进一步处理分组1330(例如,通过将分组1330发送到调度端口)。另外,记录6指明受管理交换元件810将逻辑上下文(即,分组1330已被处理管道1300的阶段1366处理)存储在分组1330的头的该组字段中。

[0229] 基于分组1330的头中存储的逻辑上下文和/或其它字段,受管理交换元件810识别转发表中的实现阶段1368的上下文映射的由带圈的7指示的记录(称为“记录7”)。在此示例中,记录7将受管理交换元件820的与VM 4耦合的端口5识别为与分组1330要被转发到的逻辑交换机230的逻辑端口(在阶段1366确定)相对应的端口。记录7附加地指明通过转发表来进一步处理分组1330(例如通过将分组1330发送到调度端口)。

[0230] 基于分组1330的头中存储的逻辑上下文和/或其它字段，受管理交换元件810随后识别转发表中的实现阶段1370的物理映射的由带圈的8指示的记录(称为“记录8”)。记录8指明为了分组1330到达受管理交换元件820要通过其发送分组1330的受管理交换元件810的端口(未示出)。在此情况下，受管理交换元件810将把分组1330从与受管理交换元件820耦合的受管理交换元件810的端口发送出去。

[0231] 如图13C所示，受管理交换元件820包括转发表，该转发表包括用于处理和路由分组1330的规则(例如流条目)。当受管理交换元件820从受管理交换元件810接收到分组1330时，受管理交换元件820基于受管理交换元件820的转发表开始处理分组1330。受管理交换元件820识别转发表中的实现阶段1372的上下文映射的由带圈的4指示的记录(称为“记录4”)。记录4基于分组1330的头中存储的逻辑上下文来识别分组1330的逻辑上下文。逻辑上下文指明分组1330已被阶段1364和1366处理，阶段1364和1366由受管理交换元件810执行。这样，记录4指明通过转发表来进一步处理分组1330(例如通过将分组1330发送到调度端口)。

[0232] 接下来，受管理交换元件820基于分组1330的头中存储的逻辑上下文和/或其它字段，识别转发表中的实现阶段1374的出口ACL的由带圈的5指示的记录(称为“记录5”)。在此示例中，记录5允许分组1330被进一步处理，并且从而指明通过转发表来进一步处理分组1330(例如，通过将分组1330发送到调度端口)。此外，记录5指明受管理交换元件820将分组1330的逻辑上下文(即，分组1330已被处理管道1300的阶段1374处理)存储在分组1330的头的该组字段中。

[0233] 接下来，受管理交换元件820基于分组1330的头中存储的逻辑上下文和/或其它字段来识别转发表中的实现阶段1376的物理映射的由带圈的6指示的记录(称为“记录6”)。记录6指明为了分组1330到达VM 4要通过其发送分组1330的受管理交换元件820的端口5。在此情况下，受管理交换元件820将把分组1330从与VM 4耦合的受管理交换元件820的端口5发送出去。在一些实施例中，受管理交换元件820在将分组1330发送到VM 4之前从分组1330去除逻辑上下文。

[0234] 图14概念性示出了一些实施例执行来转发分组以确定向哪个受管理交换元件发送分组的过程1400。过程1400在一些实施例中由接收分组并将该分组转发到另一受管理交换元件或该分组的目的地机器的受管理边缘交换元件来执行。

[0235] 过程1400开始于从源机器接收分组(在1405)。过程1400随后(在1410)执行L2处理的一部分。随着该过程执行L2处理，过程1400(在1415)确定分组是否需要被发送到第二级受管理交换元件以对分组进行进一步处理。在一些实施例中，该过程基于分组的目的地L2地址来进行此确定。该过程查看目的地L2地址并且通过与目的地L2地址相关联的端口将分组发送出去。例如，当分组的目的地L2地址是L3路由器的L2地址时，该过程将分组从与受管理交换元件相关联的端口送出，该受管理交换元件与L3路由器相关联。当分组的目的地L2地址是目的地机器的L2地址时，该过程将分组发送到与目的地机器直接连接的受管理交换元件或者在路由上更靠近目的地机器的受管理交换元件。

[0236] 当过程1400(在1415)确定分组需要被发送到第二级受管理交换元件时，过程1400(在1420)将分组发送到通信地耦合到实现逻辑路由器的L3路由器的第二级受管理交换元件。否则，过程1400(在1425)将分组发送到目的地机器或另一受管理交换元件。该过程随后

结束。

[0237] 图15概念性示出了上文描述的主机890。具体而言，当受管理交换元件810从一L3路由器接收到分组并且该分组要前往在同一主机890中实现的另一L3路由器时，受管理交换元件810基于流条目直接地桥接这两个L3路由器。

[0238] 如图所示，受管理交换元件810耦合到两个L3路由器1和2。受管理交换元件810包含的流条目在图的右侧示出。流条目指示被定址(address)为从一个L3路由器去往另一L3路由器的流量应当直接去往该另一L3路由器。

[0239] 另外，此图示出了可在主机890中配设额外的路由器，以便在更多的受管理交换元件被配设并且这些受管理交换元件依赖于现有的L3路由器来路由额外的网络流量时提供额外的路由资源。

[0240] 图16概念性示出了过程1600，一些实施例在第一和第二L3路由器实现在同一主机中时使用该过程来直接将分组从第一L3路由器转发到第二L3路由器。过程1600在一些实施例中由与在单个主机中实现的两个或更多个L3路由器交换分组的受管理交换元件(例如上文所述的受管理交换元件810)执行。

[0241] 过程1600开始于从第一L3路由器接收分组(在1605)。过程1600随后(在1610)确定该分组是否被定址到在实现第一L3路由器的同一主机中所实现的第二L3路由器。过程1600通过检查分组的头中的信息(例如目的地MAC地址)来确定这一点。

[0242] 当过程1600(在1610)确定分组要前往第二L3路由器时，过程1600(在1615)将分组发送到第二L3路由器。否则，过程1600(在1620)将分组朝着分组的目的地(例如，另一受管理交换元件或目的地机器)发送。过程1600随后结束。

[0243] 图17-24示出了基于受管理交换元件的流条目在受管理交换元件中实现的集中式逻辑路由器。图17概念性示出了上文参考图2描述的逻辑处理管道200的示例实现。图17示出了网络体系结构1700。在网络体系结构1700中，逻辑处理管道200由三个受管理交换元件1715、1720和1725执行。特别地，L2处理205和L2处理215以分布式方式跨越受管理交换元件1715、1720和1725执行。L3处理210基于受管理交换元件1720的流条目由受管理交换元件1720执行。图17还示出了源机器1710和目的地机器1730。

[0244] 受管理交换元件1715与上文参考图6描述的受管理交换元件615的类似之处在于受管理交换元件1715也是直接从耦合到边缘交换元件的机器接收分组的边缘交换元件。受管理交换元件1715从源机器1710接收分组。当受管理交换元件1715接收到来自源机器1710的分组时，受管理交换元件1715对分组执行L2处理205的一部分以便逻辑地转发分组。当分组要前往在另一逻辑网络中的目的地机器1730时，分组被转发到受管理交换元件1720。

[0245] 在受管理交换元件1715和受管理交换元件1720之间可以存在一个或多个受管理交换元件(未示出)。这些受管理交换元件具有网络构造(例如，PIF、VIF等等)，逻辑交换机220(图17中未示出)的逻辑构造(例如逻辑端口)被映射到这些网络构造。

[0246] 受管理交换元件1720是第二级受管理交换元件，其在一些实施例中用作扩展器。受管理交换元件1720执行L2处理205的剩余部分并且还执行L3处理210。受管理交换元件1720还执行逻辑处理管道200的L2处理215的一部分。受管理交换元件1720随后将分组发送到受管理交换元件1725。

[0247] 在受管理交换元件1720和受管理交换元件1725之间可以存在一个或多个受管理

交换元件(未示出)。这些受管理交换元件具有网络构造,逻辑交换机220(图17中未示出)的逻辑构造被映射到这些网络构造。

[0248] 受管理交换元件1725在该示例中从受管理交换元件1720接收分组。受管理交换元件1725对分组执行L2处理215的剩余部分以便逻辑地转发分组。在此示例中,受管理交换元件1725也是直接向目的地机器1730发送分组的交换元件。然而,在受管理交换元件1725和目的地机器1130之间可以存在一个或多个受管理交换元件(未示出)。这些受管理交换元件具有网络构造,逻辑交换机230(图17中未示出)的逻辑构造被映射到这些网络构造。

[0249] 虽然在此示例中以分布式方式执行L2处理205和L2处理215,但L2处理205和L2处理215不是必须要以分布式方式执行。例如,受管理交换元件1715可执行整个L2处理205,并且受管理交换元件1725可执行整个L2处理215。在这种情况下,受管理交换元件1720将仅执行逻辑处理管道200的L3处理210。

[0250] 图18概念性示出了一些实施例的用于通过逻辑交换机220、逻辑路由器225和逻辑交换机230处理分组的逻辑处理管道200。具体而言,此图示出了当在上文参考图17描述的网络体系结构1700中执行时的逻辑处理管道200。如上所述,在网络体系结构1700中,L2处理205、L3处理210和L2处理215由受管理交换元件1715、1720和1725执行。

[0251] L2处理205在一些实施例中包括七个阶段1805-1835,用于通过跨越受管理交换元件1715和1720实现的逻辑网络(未示出)中的逻辑交换机220(图18中未示出)来处理分组。在一些实施例中,接收分组的受管理交换元件1715在受管理交换元件1715接收到分组时执行L2处理205的一部分。受管理交换元件1720随后执行L2处理205的剩余部分。

[0252] 前五个阶段1805-1825与上文参考图7描述的前五个阶段705-725类似。在L2处理205的阶段1805中,对分组执行入口上下文映射以确定分组的逻辑上下文。在一些实施例中,阶段1805在逻辑交换机220接收到分组(例如,分组最初被受管理交换元件1715接收到)时执行。在执行第一阶段1805之后,一些实施例将表示逻辑上下文的信息存储在分组的头的一个或多个字段中。

[0253] 在一些实施例中,第二阶段1810是为逻辑交换机220定义的。在一些这样的实施例中,阶段1810在分组的逻辑上下文上操作以相对于该逻辑交换机确定分组的入口访问控制。例如,当逻辑交换机接收到分组时,对分组应用入口ACL以控制分组对该逻辑交换机的访问。基于为逻辑交换机定义的入口ACL,可进一步处理分组(例如通过阶段1815),或者例如可丢弃分组。

[0254] 在L2处理205的第三阶段1815中,在逻辑交换机的上下文中对分组执行L2转发。在一些实施例中,第三阶段1815在分组的逻辑上下文上操作以相对于逻辑交换机220处理并转发分组。例如,一些实施例定义L2转发表或L2转发条目以在第2层处理分组。另外,当分组的目的地在另一逻辑网络中时(即,当分组的目的地逻辑网络不同于其流量被逻辑交换机220处理的逻辑网络时),逻辑交换机220将分组发送到逻辑路由器225,逻辑路由器225随后将执行L3处理210以便将分组路由到目的地逻辑网络。从而,在第三阶段1815,受管理交换元件1715确定应当通过与逻辑路由器225相关联的逻辑交换机的逻辑端口(未示出)将分组转发到逻辑路由器225。

[0255] 在第四阶段1820,执行出口上下文映射以识别与分组的逻辑转发的结果相对应的物理结果。例如,分组的逻辑处理可指明分组将要从逻辑交换机220的一个或多个逻辑端口

(例如,逻辑出口端口)发送出去。这样,出口上下文映射操作识别一个或多个受管理交换元件(包括受管理交换元件1715和1720)的与逻辑交换机220的特定逻辑端口相对应的物理端口。受管理交换元件1715确定在前一阶段1815确定的逻辑端口被映射到的物理端口(例如VIF)是受管理交换元件1720的端口(未示出)。

[0256] L2处理205的第五阶段1825基于在第四阶段1820执行的出口上下文映射来执行物理映射。在一些实施例中,物理映射确定用于向在第四阶段1820中确定的物理端口发送分组的操作。例如,一些实施例的物理映射确定与执行L2处理205的受管理交换元件1715的一组端口(未示出)中的一个或多个端口相关联的一个或多个队列(未示出),将通过这一个或多个端口来发送分组以便分组到达在第四阶段1820中确定的物理端口。这样,受管理交换元件可沿着网络中的正确路径来转发分组以便分组到达所确定的物理端口。

[0257] 如图所示,L2处理205的第六阶段1830由受管理交换元件1720执行。第六阶段1830与第一阶段1805类似。阶段1830在受管理交换元件1720接收到分组时执行。在阶段1830,受管理交换元件1720查找分组的逻辑上下文并且确定剩下来要执行L2出口访问控制。

[0258] 一些实施例的第七阶段1835是为逻辑交换机220定义的。一些这样的实施例的第七阶段1835在分组的逻辑上下文上操作以相对于逻辑交换机220确定分组的出口访问控制。例如,可向分组应用出口ACL,以在对分组执行逻辑转发之后控制分组离开逻辑交换机220的访问。基于为逻辑交换机定义的出口ACL,可进一步处理分组(例如,从逻辑交换机的逻辑端口发送出去或发送到调度端口以进一步处理)或者例如可丢弃分组。

[0259] L3处理210包括六个阶段1840-1856,用于基于受管理交换元件1720的L3流条目通过在受管理交换元件1720中实现的逻辑交换机220(图18中未示出)来处理分组。如上所述,L3处理涉及执行一组逻辑路由查找以确定通过第3层网络向何处路由分组。

[0260] 第一阶段1840在逻辑路由器225接收到分组时(即,当实现逻辑路由器225的受管理交换元件1720接收到分组时)执行逻辑入口ACL查找以确定访问控制。下一阶段1841执行DNAT以将分组的目的地地址恢复回对分组的源机器隐藏的目的地机器的真实地址。在能够执行DANT时执行此阶段1841。

[0261] 下一阶段1845基于分组的L3地址(例如目的地IP地址)和路由表(例如包含L3条目)执行逻辑L3路由以确定通过第3层网络向其发送分组的一个或多个端口。由于逻辑路由器225由受管理交换元件1720实现,所以L3流条目在受管理交换元件1720中被配置。

[0262] 在第四阶段1850,一些实施例的受管理交换元件1720还对分组执行SNAT。例如,当能够执行源NAT时,受管理交换元件1720将分组的源IP地址替换为不同的IP地址以便隐藏源IP地址。另外,如下文将进一步描述的,受管理交换元件可使用NAT守护进程来接收流条目以用于转换网络地址。下文将参考图31进一步描述NAT守护进程。

[0263] 第五阶段1855在逻辑路由器225通过在阶段1845中确定的端口将分组路由出逻辑路由器225之前执行逻辑L3出口ACL查找以确定访问控制。L3出口ACL查找是基于分组的L3地址(例如源和目的地IP地址)来执行的。

[0264] 第六阶段1856执行地址解析以便将目的地L3地址(例如目的地IP地址)转换成目的地L2地址(例如目的地MAC地址)。在一些实施例中,受管理交换元件1720使用标准的地址解析(例如,通过发送出ARP请求或查找ARP缓存)来找到与目的地IP地址相对应的目的地L2地址。另外,如下文将进一步描述的,一些实施例的受管理交换元件1720可使用L3守护进程

来接收流条目以用于将L3地址解析成L2地址。下文将参考图48-50来进一步描述L3守护进程。

[0265] 当逻辑路由器225未耦合到目的地逻辑网络时，逻辑交换机220朝着目的地逻辑网络将分组发送到另一逻辑路由器网络。当逻辑路由器225耦合到目的地逻辑网络时，逻辑交换机220将分组路由到目的地逻辑网络(即，对于目的地逻辑网络转发分组的逻辑交换机)。

[0266] L2处理215在一些实施例中包括七个阶段1860-1890，用于通过跨越受管理交换元件1720和1725(未示出)实现的另一逻辑网络(图18中未示出)中的逻辑交换机230来处理分组。除了阶段1860-1890由逻辑交换机230执行(即，由实现逻辑交换机230的受管理交换元件1720和1725执行)，阶段1860-1890分别与阶段1805-1835类似。

[0267] 图19概念性示出了一些实施例的实现逻辑路由器225和逻辑交换机220和230的示例网络体系结构1900。具体而言，网络体系结构1900表示实现逻辑网络的物理网络，这些逻辑网络的数据分组被逻辑路由器225以及逻辑交换机220和230交换和/或路由。该图在其上半部示出了逻辑路由器225以及逻辑交换机220和230。该图在其下半部示出了分别在主机1990、1980和1985(例如，由诸如Windows™和Linux™之类的操作系统操作的机器)中运行的第二级受管理交换元件1910、受管理交换元件1915和1920。该图在其上部和下部都示出了VM 1-4。

[0268] 在此示例中，逻辑交换机220在逻辑路由器225、VM 1和VM 2之间转发数据分组。逻辑交换机230在逻辑路由器225、VM 3和VM 4之间转发数据分组。如上所述，逻辑路由器225在逻辑交换机220和230以及可能其它逻辑路由器和交换机(未示出)之间路由数据分组。逻辑交换机220和230以及逻辑路由器225通过逻辑端口(未示出)逻辑地耦合并且通过逻辑端口交换分组。这些逻辑端口被映射到L3路由器以及受管理交换元件1910、1915和1920的物理端口。

[0269] 在一些实施例中，逻辑交换机220和230中的每一个跨越受管理交换元件1915和1920以及可能其它受管理交换元件(未示出)实现。在一些实施例中，逻辑路由器225在通信地耦合到受管理交换元件1910的L3路由器中实现。

[0270] 在此示例中，受管理交换元件1910、1915和1920是分别在主机1990、1980和1985中运行的软件交换元件。受管理交换元件1910、1915和1920具有实现逻辑交换机220和230的流条目。利用这些流条目，受管理交换元件1915和1920在网络中的耦合到受管理交换元件1910、1915和1920的网络元件之间转发网络数据(例如分组)。例如，受管理交换元件1915在VM 1和3和第二级受管理交换元件1910之间路由网络数据。类似地，受管理交换元件1920在VM 2和4和第二级受管理交换元件1910之间路由网络数据。如图所示，受管理交换元件1915和1920各自具有三个端口(描绘为带编号的方形)，通过这些端口与耦合到受管理交换元件1915和1920的网络元件交换数据分组。

[0271] 受管理交换元件1910与上文参考图4描述的受管理交换元件305的类似之处在于受管理交换元件1910是用作扩展器的第二级受管理交换元件。受管理交换元件1910也基于流条目实现逻辑路由器225。利用这些流条目，受管理交换元件1910在L3路由分组。在此示例中，在受管理交换元件1910中实现的逻辑路由器225在跨越受管理交换元件1910和1915实现的逻辑交换机220和跨越受管理交换元件1910和1920实现的逻辑交换机230之间路由分组。

[0272] 在此示例中，受管理交换元件1910通过隧道耦合到在主机1980中运行的受管理交换元件1915，该隧道如图所示端接于受管理交换元件1915的端口2处。类似地，受管理交换元件1910通过端接于受管理交换元件1920的端口1处的隧道耦合到受管理交换元件1920。

[0273] 在此示例中，主机1980和1985中的每一个如图所示包括受管理交换元件和若干VM。VM 1-4是虚拟机，这些虚拟机每个被指派了一组网络地址(例如，用于L2的MAC地址、用于L3的IP地址，等等)并且能够向其它网络元件发送和从其它网络元件接收网络数据。VM由运行在主机1980和1985上的超管理器(未示出)来管理。

[0274] 现在将描述通过网络体系结构1900进行的若干示例数据交换。当耦合到逻辑交换机220的VM 1向也耦合到同一逻辑交换机220的VM 2发送分组时，该分组首先被发送到受管理交换元件1915。受管理交换元件1915随后对分组执行L2处理205，因为受管理交换元件1915是从VM 1接收分组的边缘交换元件。对此分组的L2处理205的结果将指示分组应当被发送到受管理交换元件1920以通过受管理交换元件1920的端口4到达VM 2。因为VM 1和2在同一逻辑网络中并且因此对于分组的L3路由不是必须的，所以不需要对此分组执行L3处理。分组然后经由在受管理交换元件1915和受管理交换元件1920之间桥接的第二级受管理交换元件1910被发送到受管理交换元件1920。分组通过受管理交换元件1920的端口4到达VM 2。

[0275] 当耦合到逻辑交换机220的VM 1向耦合到逻辑交换机230的VM 3发送分组时，该分组首先被发送到受管理交换元件1915。受管理交换元件1915对分组执行L2处理的一部分。然而，因为该分组被从一个逻辑网络发送到另一逻辑网络(即，分组的逻辑L3目的地地址是针对另一逻辑网络的)，所以需要对此分组执行L3处理。

[0276] 受管理交换元件1915将分组发送到第二级受管理交换元件1910，使得受管理交换元件1910对分组执行L2处理的剩余部分和L3处理210。受管理交换元件1910随后执行另一L2处理的一部分并且将分组转发到受管理交换元件1920。受管理交换元件1915对从受管理交换元件1910接收的分组执行L2处理215，并且此L2处理的结果将指示分组应当通过受管理交换元件1915的端口5被发送到VM 3。

[0277] 当耦合到逻辑交换机220的VM 1向耦合到逻辑交换机230的VM 4发送分组时，该分组首先被发送到受管理交换元件1915。受管理交换元件1915对分组执行L2处理205。然而，因为该分组被从一个逻辑网络发送到另一逻辑网络，所以需要执行L3处理。

[0278] 受管理交换元件1915将分组发送到受管理交换元件1910，使得受管理交换元件1910对分组执行L2处理205的剩余部分和L3处理210。在受管理交换元件1910处执行的L3处理210的结果将指示分组应当被发送到受管理交换元件1915。受管理交换元件1910随后对分组执行L2处理的一部分，并且此L2处理的结果将指示分组应当通过受管理交换元件1920被发送到VM 4。受管理交换元件1920执行L2处理的剩余部分以确定分组应当通过受管理交换元件1920的端口5被发送到VM 4。

[0279] 图20概念性示出了一些实施例的实现逻辑路由器225以及逻辑交换机220和230的示例网络体系结构2000。具体而言，网络体系结构2000表示实现逻辑网络的物理网络，这些逻辑网络的数据分组被逻辑路由器225以及逻辑交换机220和230交换和/或路由。该图在其上半部示出了逻辑路由器225以及逻辑交换机220和230。该图在其下半部示出了分别在主机1990、1980和1985中运行的第二级受管理交换元件1910、受管理交换元件1915和1920。该

图在其上部和下部都示出了VM 1-4。

[0280] 除了网络体系结构2000附加地包括在主机2010中运行的受管理交换元件2005，网络体系结构2000与网络体系结构1900类似，。一些实施例的受管理交换元件2005是用作池节点的第二级受管理交换元件。

[0281] 在一些实施例中，网络控制系统(未示出)建立隧道以促成网络元件之间的通信。例如，受管理交换元件1915在此示例中通过隧道耦合到在主机2010中运行的受管理交换元件2005，该隧道如图所示端接于受管理交换元件1915的端口1处。类似地，受管理交换元件1920通过端接于受管理交换元件1920的端口2处的隧道耦合到受管理交换元件2005。另外，受管理交换元件2005和1910如图所示通过隧道耦合。

[0282] 如上文参考图19所述，逻辑路由器225以及逻辑交换机220和230在受管理交换元件1910、1915和1920中实现，除了在数据分组交换中涉及第二级受管理交换元件2005。也就是说，受管理交换元件1915和1910通过受管理交换元件2005来交换分组。受管理交换元件1920和1910通过受管理交换元件2005来交换分组。受管理交换元件1915和1920通过受管理交换元件2005来交换分组。

[0283] 图21概念性示出了一些实施例的实现逻辑路由器225以及逻辑交换机220和230的示例网络体系结构2100。除了存在受管理交换元件1910和受管理交换元件1920之间建立的隧道，网络体系结构2100与网络体系结构1900类似。此图示出了一些实施例的网络体系结构2100是网络体系结构1900和网络体系结构2000的混合。也就是说，一些受管理边缘交换元件具有去往与集中式L3路由器耦合的第二级受管理交换元件的隧道，而其它受管理边缘交换元件为了与耦合到集中式L3路由器的第二级受管理交换元件交换分组必须通过用作池节点的第二级受管理交换元件。

[0284] 图22概念性示出了一些实施例的包括基于流条目实现逻辑路由器的受管理交换元件1910的主机1990的示例体系结构。主机1990在一些实施例中是由能够创建虚拟机的操作系统(例如Linux)管理的机器。如图所示，主机1990在此示例中包括受管理交换元件1910和NIC 2245。此图还示出了控制器集群2205。

[0285] 控制器集群2205是管理包括受管理交换元件1910在内的网络元件的一组网络控制器或控制器实例。受管理交换元件1910在此示例中是在主机1990中实现的包括用户空间2212和内核2210的软件交换元件。受管理交换元件1910包括在用户空间2212中运行的控制守护进程2215，以及在内核2210中运行的控制器补丁2230和网桥2235。在用户空间2212中运行的还有NAT守护进程2220，下文将对其进行进一步描述。用户空间2212和内核2210在一些实施例中具有主机1990的操作系统，而在其它实施例中用户空间2212和内核2210具有在主机1990上运行的虚拟机。

[0286] 在一些实施例中，控制器集群2205与控制守护进程2215通信(例如利用OpenFlow协议或某种其它通信协议)，控制守护进程2215在一些实施例中是在用户空间2212的后台运行的应用。控制守护进程2215与控制器集群2205通信以便处理和路由受管理交换元件1910接收的分组。具体而言，控制守护进程2215在一些实施例中从控制器集群2205接收配置信息并且配置控制器补丁2230。例如，控制守护进程2215从控制器集群2205接收关于用于处理和路由受管理交换元件1910接收的分组的操作的命令。

[0287] 控制器补丁2230是在内核2210中运行的模块。在一些实施例中，控制守护进程

2215配置控制器补丁2230。当被配置时，控制器补丁2230包含关于对要接收的分组进行处理、转发和路由的规则(例如流条目)。控制器补丁2230实现逻辑交换机和逻辑路由器二者。

[0288] 在一些实施例中，控制器补丁2230使用NAT守护进程来进行网络地址转换。如下文将进一步描述的，NAT守护进程2220生成关于网络地址转换的流条目并将流条目发送回受管理交换元件1910以使用。NAT守护进程将在下文进一步描述。

[0289] 控制器补丁2230从内核2210的网络堆栈2250或从网桥2235接收分组。网桥2235在网络堆栈2250和主机外部的网络主机之间路由网络数据(即，通过NIC 2245接收的网络数据)。如图所示，网桥2235在网络堆栈2250和NIC 2245之间以及网络堆栈2250和NIC 2245之间路由网络数据。一些实施例的网桥2235执行标准的L2分组学习和路由。

[0290] 网络堆栈2250可通过NIC 2245从受管理交换元件1910外部的网络主机接收分组。网络堆栈2250随后将分组发送到控制器补丁2230。在一些情况下，通过隧道从受管理交换元件外部的网络主机接收分组。在一些实施例中，隧道端接于网络堆栈2250。从而，当网络堆栈2250通过隧道接收分组时，网络堆栈2250拆开隧道头(即，解封出有效负荷)并将拆开的分组发送到控制器补丁2230。

[0291] 现在将描述受管理交换元件1910的示例操作。在此示例中，在受管理交换元件1910与主机1990外部的受管理交换元件1915和1920(在图22中未示出)之间建立隧道。也就是说，受管理交换元件1910、1915和1920如图19所示通过隧道连接。隧道端接于网络堆栈2250。

[0292] 受管理交换元件1915向受管理交换元件1910发送分组，该分组是由VM 1发送到VM 4的。该分组被NIC 2245接收，然后被发送到网桥2235。基于分组头中的信息，网桥2235确定该分组是通过建立的隧道发送的，并且将分组发送到网络堆栈2250。网络堆栈2250拆开隧道头，并将拆开的分组发送到控制器补丁2230。

[0293] 根据控制器补丁2230具有的流条目，控制器补丁2230执行L3处理以路由分组，因为分组被从一个逻辑网络发送到另一逻辑网络。通过执行L3处理和一些L2处理，受管理交换元件1910确定分组应当被发送到受管理交换元件1920，因为目的地网络层地址应当去往属于目的地逻辑网络的逻辑交换机。控制器补丁2230通过网络堆栈2250、网桥2235和NIC 2245经由隧道将分组发送到实现属于目的地逻辑网络的逻辑交换机的受管理交换元件1920。

[0294] 图23概念性示出了受管理交换元件中的逻辑交换机和逻辑路由器的示例实现。具体而言，该图示出了逻辑路由器225以及逻辑交换机220和230在第二级受管理交换元件1910以及受管理交换元件1915和1920中的实现。该图在其上半部示出了逻辑路由器225以及逻辑交换机220和230。该图在其下半部示出了受管理交换元件1910-1920。该图在其上半部和下半部都示出了VM 1-4。

[0295] 逻辑交换机220和230以及逻辑路由器225通过逻辑端口逻辑地耦合。逻辑交换机220和230的该特定的配置与上文参考图12描述的示例中示出的配置相同。

[0296] 在图23的示例中，控制器集群2205(图23中未示出)通过向受管理交换元件1910提供流条目来配置受管理交换元件1910，以使得该受管理交换元件基于流条目实现逻辑路由器225。

[0297] 图24概念性示出了上文参考图23描述的逻辑交换机220和230、逻辑路由器225以

及受管理交换元件1910、1915和1920的示例操作。具体而言，图24示出了实现逻辑路由器225的受管理交换元件1910的操作。为了简单起见，受管理交换元件1915和1920执行的逻辑处理管道的部分在图24中没有描绘。逻辑处理管道的这些部分与图13A和图13C的上半部所示的示例中的受管理交换元件815和820执行的逻辑处理的部分类似。也就是说，为了示出图24的示例，图24替换图13A和图13B的下半部。

[0298] 如图24的下半部所示，受管理交换元件1910包括L2条目2405和2415以及L3条目2410。这些条目是控制器集群2205(未示出)提供给受管理交换元件1910的流条目。虽然这些条目被描绘为三个分开的表，但这些表不一定必须是分开的表。也就是说，单个表可包括所有这些流条目。

[0299] 当受管理交换元件1910从受管理交换元件1915接收到从VM 1向VM 4发送的分组2430时，受管理交换元件1910基于受管理交换元件1910的流条目2405开始处理分组2430。受管理交换元件1910识别转发表中的实现阶段1830的上下文映射的由带圈的1指示的记录(称为“记录1”)。记录1基于分组2430的头中存储的逻辑上下文来识别分组2430的逻辑上下文。逻辑上下文指明分组2430已被由受管理交换元件1915执行的逻辑处理的一部分(即，L2入口ACL、L2转发)处理。这样，记录1指明通过转发表来进一步处理分组2430(例如通过将分组2430发送到调度端口)。

[0300] 接下来，受管理交换元件1910基于分组2430的头中存储的逻辑上下文和/或其它字段，识别转发表中的实现阶段1835的出口ACL的由带圈的2指示的记录(称为“记录2”)。在此示例中，记录2允许分组2430被进一步处理(例如，分组2430可通过逻辑交换机220的端口“X”离开逻辑交换机220)，并且从而指明通过受管理交换元件1910的流条目来进一步处理分组2430(例如，通过将分组2430发送到调度端口)。此外，记录2指明受管理交换元件1910将分组2430的逻辑上下文(即，分组2430已被处理管道2400的阶段1835处理)存储在分组2430的头的该组字段中。(要注意，所有记录都指明每当受管理交换元件基于记录执行逻辑处理的某个部分时，执行逻辑处理的受管理交换元件更新该组字段中存储的逻辑上下文。)

[0301] 受管理交换元件1910基于流条目继续处理分组2430。受管理交换元件1910基于分组2430的头中存储的逻辑上下文和/或其它字段来识别L3条目2410中的由带圈的3指示的记录(称为“记录3”)，该记录通过基于分组2430的头中的信息指明受管理交换元件1910应当接受通过逻辑路由器225的逻辑端口1的分组来实现L3入口ACL。

[0302] 受管理交换元件1910随后识别L3条目2410中的由带圈的4指示的流条目(称为“记录4”)，该流条目通过指明具有其目的地IP地址(例如1.1.2.10)的分组2430应当被允许从逻辑路由器225的端口2离开来实现L3路由1845。另外，记录4(或路由表中的另一记录，未示出)指示分组2430的源MAC地址要被改写成逻辑路由器225的端口2的MAC地址(即，01:01:01:01:01:02)。受管理交换元件1910随后识别L3条目2410中的由带圈的5指示的流条目(称为“记录5”)，该流条目通过基于分组2430的头中的信息(例如源IP地址)指明受管理交换元件1910可通过逻辑路由器225的端口2将分组发送出去来实现L3出口ACL。

[0303] 基于分组2430的头中存储的逻辑上下文和/或其它字段，受管理交换元件1910识别L2条目2415中的实现阶段1860的入口ACL的由带圈的6指示的记录(称为“记录6”)。在此示例中，记录6允许分组2430被进一步处理，并且从而指明由受管理交换元件1910来进一步处理分组2430(例如，通过将分组2430发送到调度端口)。此外，记录6指明受管理交换元件

1910将分组2430的逻辑上下文(即,分组2430已被处理管道2400的阶段1860处理)存储在分组2430的头的该组字段中。

[0304] 接下来,受管理交换元件1910基于分组2430的头中存储的逻辑上下文和/或其它字段来识别转发表中的实现阶段1865的逻辑L2转发的由带圈的7指示的记录(称为“记录7”)。记录7指明以VM 4的MAC地址作为目的地MAC地址的分组应当通过连接到VM 4的逻辑交换机230的逻辑端口(未示出)来转发。

[0305] 记录7还指明通过转发表来进一步处理分组2430(例如,通过将分组2430发送到调度端口)。另外,记录7指明受管理交换元件1910将逻辑上下文(即,分组2430已被处理管道2400的阶段1865处理)存储在分组2430的头的该组字段中。

[0306] 基于分组2430的头中存储的逻辑上下文和/或其它字段,受管理交换元件1910识别转发表中的实现阶段1870的上下文映射的由带圈的8指示的记录(称为“记录8”)。在此示例中,记录8将与VM 4耦合的受管理交换元件1920的端口5识别为与分组2430要被转发到的逻辑交换机230的逻辑端口(在阶段1865确定)相对应的端口。记录8附加地指明通过转发表来进一步处理分组2430(例如通过将分组2430发送到调度端口)。

[0307] 基于分组2430的头中存储的逻辑上下文和/或其它字段,受管理交换元件1910随后识别L2条目2415中的实现阶段1875的物理映射的由带圈的9指示的记录(称为“记录9”)。记录9指明为了分组2430到达受管理交换元件1920要通过其发送分组2430的受管理交换元件1910的端口(未示出)。在此情况下,受管理交换元件1910将把分组2430从与受管理交换元件1920耦合的受管理交换元件1910的该端口发送出去。

[0308] 图25-30B示出了基于受管理交换元件的流条目在若干受管理交换元件中实现的分布式逻辑路由器。特别地,图25-30B示出了包括源L2处理、L3路由和目的地L2处理在内的整个逻辑处理管道由第一跳受管理交换元件(即,直接从机器接收分组的交换元件)执行。

[0309] 图25概念性示出了上文参考图2描述的逻辑处理管道200的示例实现。特别地,图25示出了L3处理210可由任何直接从源机器接收分组的受管理交换元件执行。图25示出了网络体系结构2500。在网络体系结构2500中,逻辑处理管道200由受管理交换元件2505执行。在此示例中,L3处理210由受管理交换元件2505基于受管理交换元件2505的流条目执行。图25还示出了源机器2515和目的地机器2520。

[0310] 受管理交换元件2505是直接从耦合到边缘交换元件的机器接收分组的边缘交换元件。受管理交换元件2505从源机器2515接收分组。当受管理交换元件2505接收到来自源机器2515的分组时,受管理交换元件2505在一些实施例中对分组执行整个逻辑处理管道200以便逻辑地转发和路由分组。

[0311] 当接收到的分组要前往在此示例中在另一逻辑网络中的目的地机器2520时,受管理交换元件2505用作:在源机器2515所属的逻辑网络中的逻辑交换机;在目的地机器2520所属的逻辑网络中的逻辑交换机;以及在两个逻辑交换机之间路由分组的逻辑路由器。基于执行逻辑处理管道200的结果,受管理交换元件2505将分组转发到受管理交换元件2510,目的地机器2520通过受管理交换元件2510接收分组。

[0312] 图26概念性示出了一些实施例的用于通过逻辑交换机220、逻辑路由器225和逻辑交换机230处理分组的逻辑处理管道200。具体而言,此图示出了当在上文参考图25描述的网络体系结构2500中执行时的逻辑处理管道200。如上所述,在网络体系结构2500中,L2处

理205、L3处理210和L2处理215由单个受管理交换元件2505执行，受管理交换元件2505是从机器接收分组的边缘交换元件。因此，在这些实施例中，第一跳交换元件执行整个逻辑处理管道200。

[0313] L2处理205在一些实施例中包括四个阶段2605-2620，用于通过逻辑交换机220(图26中未示出)来处理分组。在阶段2605中，对分组执行入口上下文映射以确定分组的逻辑上下文。在一些实施例中，阶段2605在逻辑交换机220接收到分组(例如，分组最初被受管理交换元件2505接收)时执行。

[0314] 在一些实施例中，第二阶段2610是为逻辑交换机220定义的。在一些这样的实施例中，阶段2610在分组的逻辑上下文上操作以相对于该逻辑交换机确定分组的入口访问控制。例如，当逻辑交换机接收到分组时，对分组应用入口ACL以控制分组对该逻辑交换机的访问。基于为逻辑交换机定义的入口ACL，可进一步处理分组(例如通过阶段2615)，或者例如可丢弃分组。

[0315] 在L2处理205的第三阶段2615中，在逻辑交换机的上下文中对分组执行L2转发。在一些实施例中，第三阶段2615在分组的逻辑上下文上操作以相对于逻辑交换机220处理并转发分组。例如，一些实施例定义用于在第2层处理分组的L2转发表或L2转发条目。

[0316] 一些实施例的第四阶段2620是为逻辑交换机220定义的。一些这样的实施例的第四阶段2620在分组的逻辑上下文上操作以相对于该逻辑交换机确定分组的出口访问控制。例如，可向分组应用出口ACL，以在对分组执行逻辑转发之后控制分组离开逻辑交换机220的访问。基于为逻辑交换机定义的出口ACL，可进一步处理分组(例如，从逻辑交换机的逻辑端口发送出去或发送到调度端口以便进一步处理)或者例如可丢弃分组。

[0317] 当分组的目的地在另一逻辑网络中时(即，当分组的目的地逻辑网络不同于其流量被逻辑交换机220处理的逻辑网络时)，逻辑交换机220将分组发送到逻辑路由器225，逻辑路由器225随后在阶段210执行L3处理以便将分组路由到目的地逻辑网络。L3处理210包括六个阶段2635-2651，用于通过由受管理交换元件2505(图26中未示出)实现的逻辑路由器225(图26中未示出)来处理分组。如上所述，L3处理涉及执行一组逻辑路由查找以确定通过第3层网络向何处路由分组。

[0318] 第一阶段2635在逻辑路由器225接收到分组时(即，当实现逻辑路由器225的受管理交换元件2505接收到分组时)执行逻辑入口ACL查找以确定访问控制。在一些实施例中，阶段2635在分组的逻辑上下文上操作以相对于逻辑路由器225确定分组的入口访问控制。下一阶段2636执行DNAT以将分组的目的地地址恢复回对分组的源机器隐藏的目的地机器的真实地址。在能够执行DANT时执行此阶段2636。

[0319] 下一阶段2640基于分组的L3地址(例如目的地IP地址)、包含L3流条目的转发表和分组的逻辑上下文来执行逻辑L3路由以确定一个或多个逻辑端口来通过第3层网络发送分组。

[0320] 一些实施例的第四阶段2645对分组执行SNAT。例如，当能够执行SNAT时，受管理交换元件2505将分组的源IP地址替换为不同的IP地址以便隐藏源IP地址。另外，如下文将进一步描述的，受管理交换元件可使用NAT守护进程来接收用于转换网络地址的流条目。NAT守护进程将在下文参考图31来进一步描述。

[0321] 第五阶段2650在逻辑路由器225通过在阶段2640中确定的端口将分组路由出逻辑

路由器225之前执行逻辑出口ACL查找以确定访问控制。出口ACL查找基于分组的L3地址(例如源和目的地IP地址)来执行。在一些实施例中,阶段2650在分组的逻辑上下文上操作以相对于逻辑路由器225确定分组的出口访问控制。

[0322] 第六阶段2651执行地址解析以便将目的地L3地址(例如目的地IP地址)转换成目的地L2地址(例如目的地MAC地址)。在一些实施例中,受管理交换元件2505使用标准的地址解析(例如,通过发送出ARP请求或查找ARP缓存)来找到与目的地IP地址相对应的目的地L2地址。另外,如下文将进一步描述的,一些实施例的受管理交换元件2505可使用L3守护进程来接收用于将L3地址解析成L2地址的流条目。L3守护进程将在下文参考图48-50来进一步描述。

[0323] 当逻辑路由器225未耦合到目的地逻辑网络时,逻辑交换机220朝着目的地逻辑网络将分组发送到另一逻辑路由器网络。与该另一逻辑路由器的操作相对应的逻辑处理的部分也将在受管理交换元件2505中实现。当逻辑路由器225耦合到目的地逻辑网络时,逻辑交换机220将分组路由到目的地逻辑网络(即,为目的地逻辑网络转发分组的逻辑交换机)。

[0324] L2处理215在一些实施例中包括五个阶段2660-2680,用于通过在逻辑交换机225(图26中未示出)来处理分组。在一些实施例中,在一些实施例中,第一阶段2660是为逻辑交换机225定义的。在一些这样的实施例中,阶段2660在分组的逻辑上下文上操作以相对于逻辑交换机230确定分组的入口访问控制。例如,当逻辑交换机230从逻辑路由器225接收到分组时,对分组应用入口ACL以控制分组对逻辑交换机230的访问。基于为逻辑交换机定义的入口ACL,可进一步处理分组(例如通过阶段2665),或者例如可丢弃分组。

[0325] 在L2处理管道215的第二阶段2665中,在逻辑交换机的上下文中对分组执行L2转发。在一些实施例中,第三阶段2665在分组的逻辑上下文上操作以相对于逻辑交换机220处理并转发分组。例如,一些实施例定义用于在第2层处理分组的L2转发表或L2转发条目。

[0326] 一些实施例的第三阶段2670是为逻辑交换机220定义的。一些这样的实施例的第三阶段2670在分组的逻辑上下文上操作以相对于该逻辑交换机确定分组的出口访问控制。例如,可向分组应用出口ACL,以在对分组执行逻辑转发之后控制分组离开逻辑交换机230的访问。基于为逻辑交换机定义的出口ACL,可进一步处理分组(例如,从逻辑交换机的逻辑端口发送出去或发送到调度端口以进一步处理)或者例如可丢弃分组。

[0327] 在第四阶段2675中,执行出口上下文映射以识别与分组的逻辑转发的结果相对应的物理结果。例如,分组的逻辑处理可指明分组要被从逻辑交换机230的一个或多个逻辑端口(例如,逻辑出口端口)发送出去。这样,出口上下文映射操作识别与逻辑交换机的特定逻辑端口相对应的一个或多个受管理交换元件(包括受管理交换元件2505)的物理端口。

[0328] L2处理215的第五阶段2680基于在第四阶段2675执行的出口上下文映射来执行物理映射。在一些实施例中,物理映射确定用于向在第四阶段2675中确定的物理端口转发分组的操作。例如,一些实施例的物理映射确定与受管理交换元件2505的一组端口(未示出)中的一个或多个端口相关联的一个或多个队列(未示出),通过这一个或多个端口来发送分组以便分组到达在第四阶段2675中确定的物理端口。这样,受管理交换元件可沿着网络中的正确路径来路由分组以便分组到达所确定的(一个或多个)物理端口。另外,一些实施例在第五阶段2680完成之后去除逻辑上下文以便在对分组执行逻辑处理管道200之前将分组返回到其原始状态。

[0329] 图27概念性示出了一些实施例的实现逻辑路由器225以及逻辑交换机220和230的示例网络体系结构2700。具体而言，网络体系结构2700表示实现逻辑网络的物理网络，这些逻辑网络的数据分组通过逻辑路由器225和逻辑交换机220和230交换和/或路由。该图在其上半部示出了逻辑路由器225以及逻辑交换机220和230。该图在其下半部示出了受管理交换元件2505和2510。该图在其上部和下部都示出了VM 1-4。

[0330] 在此示例中，逻辑交换机220在逻辑路由器225、VM 1和VM 2之间转发数据分组。逻辑交换机230在逻辑路由器225、VM 3和VM 4之间转发数据分组。如上所述，逻辑路由器225在逻辑交换机220和230以及其它逻辑路由器和交换机(未示出)之间路由数据分组。逻辑交换机220和230以及逻辑路由器225通过逻辑端口(未示出)逻辑地耦合并且通过逻辑端口交换数据分组。这些逻辑端口被映射或附接到受管理交换元件2505和2510的物理端口。

[0331] 在一些实施例中，逻辑路由器是在受管理网络中的每个受管理交换元件中实现的。当受管理交换元件从耦合到受管理交换元件的机器接收到分组时，受管理交换元件执行逻辑路由。换言之，相对于分组是第一跳交换元件的这些实施例的受管理交换元件执行L3处理210。

[0332] 在此示例中，受管理交换元件2505和2510是分别在主机2525和2530中运行的软件交换元件。受管理交换元件2505和2510具有实现逻辑交换机220和230的流条目以转发和路由受管理交换元件2505和2510从VM 1-4接收的分组。流条目还实现逻辑路由器225。利用这些流条目，受管理交换元件2505和2510可在网络中的耦合到受管理交换元件2505和2510的网络元件之间转发和路由分组。如图所示，受管理交换元件2505和2510每个具有三个端口(例如VIF)，通过这些端口与耦合到受管理交换元件2505和2510的网络元件交换数据分组。在一些情况下，这些实施例中的数据分组将行经在受管理交换元件2505和2510之间建立的隧道(例如，端接于受管理交换元件2505的端口3和受管理交换元件2510的端口3的隧道)。

[0333] 在此示例中，主机2525和2530的每一个如图所示包括受管理交换元件和若干VM。VM 1-4是虚拟机，这些虚拟机的每一个被指派一组网络地址(例如，用于L2的MAC地址、用于网络L3的IP地址等等)并且能够向其它网络元件发送和从其它网络元件接收网络数据。VM由在主机2525和2530上运行的超管理器(未示出)来管理。

[0334] 现在将描述通过网络体系结构2700进行的若干示例数据交换。当耦合到逻辑交换机220的VM 1向也耦合到同一逻辑交换机220的VM 2发送分组时，该分组首先被发送到受管理交换元件2505。受管理交换元件2505随后对分组执行L2处理205。L2处理的结果将指示分组应当通过在受管理交换元件2505和2510之间建立的隧道被发送到受管理交换元件2510并通过受管理交换元件2510的端口4到达VM 2。因为VM 1和2在同一逻辑网络中，所以受管理交换元件2505不执行L3处理210和L2处理215。

[0335] 当耦合到逻辑交换机220的VM 1向耦合到逻辑交换机230的VM 3发送分组时，该分组首先被发送到受管理交换元件2505。受管理交换元件2505对分组执行L2处理205。然而，因为该分组被从一个逻辑网络发送到另一逻辑网络(即，分组的逻辑L3目的地地址是针对另一逻辑网络的)，所以需要执行L3处理210。受管理交换元件2505还执行L2处理215。也就是说，受管理交换元件2505作为接收到分组的第一跳交换元件对分组执行整个逻辑处理管道200。执行逻辑处理管道200的结果将指示分组应当通过受管理交换元件2505的端口5被发送到VM 3。从而，分组不必去往另一受管理交换元件，虽然分组确实经过了两个逻辑交换

机和一逻辑路由器。

[0336] 当耦合到逻辑交换机220的VM 1向耦合到逻辑交换机230的VM 4发送分组时，该分组首先被发送到受管理交换元件2505。受管理交换元件2505作为该分组的第一跳交换元件对该分组执行整个逻辑处理管道200。对此分组执行逻辑处理管道200的结果将指示分组应当通过在受管理交换元件2505和2510之间建立的隧道被发送到受管理交换元件2510并且通过受管理交换元件2510的端口5到达VM 4。

[0337] 图28概念性示出了一些实施例的实现逻辑路由器225以及逻辑交换机220和230的示例网络体系结构2800。具体而言，网络体系结构2800表示实现逻辑网络的物理网络，这些逻辑网络的数据分组通过逻辑路由器225和逻辑交换机220和230交换和/或路由。该图在其上半部示出了逻辑路由器225以及逻辑交换机220和230。该图在其下半部示出了受管理交换元件2505和2510。该图在其上部和下部都示出了VM 1-4。

[0338] 除了网络体系结构2800附加地包括受管理交换元件2805，网络体系结构2800与网络体系结构2700类似，。一些实施例的受管理交换元件2805是用作池节点的第二级受管理交换元件。

[0339] 在一些实施例中，网络控制系统(未示出)建立隧道以促成网络元件之间的通信。例如，受管理交换元件2505在此示例中通过隧道耦合到在主机2810中运行的受管理交换元件2805，该隧道如图所示端接于受管理交换元件2505的端口1处。类似地，受管理交换元件2510通过端接于受管理交换元件2510的端口2处的隧道耦合到受管理交换元件2805。与以上图27所示的示例体系结构2700不同，在受管理交换元件2505和2510之间没有建立隧道。

[0340] 在受管理交换元件2505中实现逻辑路由器225以及逻辑交换机220和230，并且在数据分组交换中涉及第二级受管理交换元件2805。也就是说，受管理交换元件2505和2510通过受管理交换元件2805交换分组。

[0341] 图29概念性示出了对接收到的分组执行所有L2和L3处理以便转发和路由的第一跳交换元件的示例。图29示出了由受管理交换元件2505和2510来实现逻辑路由器225以及逻辑交换机220和230。如图所示，当受管理交换元件2505是第一跳交换元件时，由受管理交换元件2505执行整个逻辑处理管道200。该图在其左半部示出了逻辑路由器225以及逻辑交换机220和230。该图在其右半部示出了受管理交换元件2505和2510。该图在其右半部和左半部都示出了VM 1-4。

[0342] 当耦合到逻辑交换机220的VM 1向也耦合到同一逻辑交换机220的VM 2发送分组时，该分组首先通过受管理交换元件2505的端口4被发送到受管理交换元件2505，因为逻辑交换机220的逻辑端口1被附接或映射到受管理交换元件2505的端口4，分组通过这个逻辑端口1进入逻辑交换机220。

[0343] 受管理交换元件2505随后对分组执行L2处理205。具体而言，受管理交换元件2505首先执行逻辑上下文查找以基于分组的头字段中包括的信息来确定分组的逻辑上下文。在此示例中，分组的源MAC地址是VM 1的MAC地址，并且分组的源IP地址是VM 1的IP地址。分组的目的地MAC地址是VM 2的MAC地址，并且分组的目的地IP地址是VM 2的IP地址。在此示例中，逻辑上下文指明逻辑交换机220是要转发分组的逻辑交换机并且逻辑交换机220的逻辑端口1是通过其接收分组的端口。逻辑上下文还指明逻辑交换机220的端口2是通过其将分组送出到VM 2的端口，因为端口2与VM 2的MAC地址相关联。

[0344] 受管理交换元件2505随后基于所确定的分组的逻辑上下文来执行逻辑转发查找。受管理交换元件2505为分组确定访问控制。例如，受管理交换元件2505确定分组不具有将使得逻辑交换机220拒绝经过逻辑交换机220的端口1的分组的网络地址(例如，源/目的地MAC/IP地址，等等)。受管理交换元件2505还从逻辑上下文识别出逻辑交换机220的端口2是要发送出分组的端口。另外，受管理交换元件2505相对于逻辑交换机220的端口2为分组确定访问控制。例如，受管理交换元件2505确定分组不具有将使得逻辑交换机220不通过逻辑交换机220的端口2发送分组的网络地址。

[0345] 受管理交换元件2505随后执行映射查找以确定逻辑交换机220的逻辑端口2被映射到的物理端口。在此示例中，受管理交换元件2505确定逻辑交换机220的逻辑端口2被映射到受管理交换元件2510的端口4。受管理交换元件2505随后执行物理查找以确定用于将分组转发到物理端口的操作。在此示例中，受管理交换元件2505确定分组应当通过在受管理交换元件2505和2510之间建立的隧道被发送到受管理交换元件2510并通过受管理交换元件2510的端口4到达VM 2。因为VM 1和2在同一逻辑网络中，所以受管理交换元件2505不执行L3处理。受管理交换元件2510不对分组执行任何逻辑处理，而只是通过受管理交换元件2510的端口4将分组转发到VM 2。

[0346] 当耦合到逻辑交换机220的VM 1向耦合到逻辑交换机230的VM 3发送分组时(即，当VM 1和3在不同的逻辑网络中时)，该分组首先通过受管理交换元件2505的端口4被发送到受管理交换元件2505。受管理交换元件2505对分组执行L2处理205。具体而言，受管理交换元件2505首先执行逻辑上下文查找以基于分组的头字段中包括的信息来确定分组的逻辑上下文。在此示例中，分组的源MAC地址是VM 1的MAC地址，并且分组的源IP地址是VM 1的IP地址。因为分组被从VM 1发送到在不同逻辑网络中的VM 3，所以分组具有与端口X相关联的MAC地址作为目的地MAC地址(即，在此示例中为01:01:01:01:01:01)。分组的目的地IP地址是VM 3的IP地址(例如，1.1.2.10)。在此示例中，逻辑上下文指明逻辑交换机220是要转发分组的逻辑交换机并且逻辑交换机220的逻辑端口1是通过其接收分组的端口。逻辑上下文还指明逻辑交换机220的端口X是通过其将分组送出到逻辑路由器225的端口，因为端口X与逻辑路由器225的端口1的MAC地址相关联。

[0347] 受管理交换元件2505随后为分组确定访问控制。例如，受管理交换元件2505确定分组不具有将使得逻辑交换机220拒绝经过逻辑交换机220的端口1的分组的网络地址(例如，源/目的地MAC/IP地址，等等)。受管理交换元件2505还从逻辑上下文识别出逻辑交换机220的端口X是要发送出分组的端口。另外，受管理交换元件2505相对于端口X为分组确定访问控制。例如，受管理交换元件2505确定分组不具有将使得逻辑交换机220不通过端口X发送分组的网络地址。

[0348] 受管理交换元件2505随后对分组执行L3处理210，因为分组的目的地IP地址1.1.2.10是针对另一逻辑网络的(即，当分组的目的地逻辑网络不同于其流量被逻辑交换机220处理的逻辑网络时)。受管理交换元件2505在L3为分组确定访问控制。例如，受管理交换元件2505确定分组不具有将使得逻辑路由器225拒绝经过逻辑路由器225的逻辑端口1的分组的网络地址。受管理交换元件2505还查找L3流条目并确定分组要被发送到逻辑路由器225的逻辑端口2，因为分组的目的地IP地址1.1.2.10属于与逻辑路由器225的逻辑端口2相关联的1.1.2.1/24的子网地址。另外，受管理交换元件2505相对于逻辑路由器225的逻辑端

口2为分组确定访问控制。例如，受管理交换元件2505确定分组不具有将使得逻辑交换机220不通过逻辑端口2发送分组的网络地址。

[0349] 受管理交换元件2505在执行L3处理210时修改分组的逻辑上下文或分组本身。例如，受管理交换元件2505将分组的逻辑源MAC地址修改为逻辑路由器225的逻辑端口2的MAC地址(即，在此示例中为01:01:01:01:01:02)。受管理交换元件2505还将分组的目的地MAC地址修改为VM 3的MAC地址。

[0350] 受管理交换元件2505随后执行L2处理215。具体而言，受管理交换元件2505为分组确定访问控制。例如，受管理交换元件2505确定分组不具有将使得逻辑交换机230拒绝经过逻辑交换机230的端口Y的分组的网络地址(例如，源/目的地MAC/IP地址，等等)。受管理交换元件2505随后确定逻辑交换机230的端口1是通过其将分组送出到目的地VM 3的端口。另外，受管理交换元件2505相对于逻辑交换机230的端口1为分组确定访问控制。例如，受管理交换元件2505确定分组不具有将使得逻辑交换机230不通过逻辑交换机230的端口1发送分组的网络地址。

[0351] 受管理交换元件2505随后执行映射查找以确定逻辑交换机230的逻辑端口1被映射到的物理端口。在此示例中，受管理交换元件2505确定逻辑交换机230的逻辑端口1被映射到受管理交换元件2505的端口5。受管理交换元件2505随后执行物理查找以确定用于将分组转发到物理端口的操作。在此示例中，受管理交换元件2505确定分组应当通过受管理交换元件2505的端口5被发送到VM 3。受管理交换元件2505在此示例中在将分组送出到VM 3之前从分组去除逻辑上下文。从而，分组不必去往另一受管理交换元件，虽然分组确实经过两个逻辑交换机和一逻辑路由器。

[0352] 当耦合到逻辑交换机220的VM 1向耦合到逻辑交换机230的VM 4发送分组时，分组以与从VM 1发送到VM 3的分组被发送到VM 3的方式相类似的方式被发送到VM 4，除了前往VM 4的分组通过在受管理交换元件2505和2510之间建立的隧道被从受管理交换元件2505发送到受管理交换元件2510，并且通过受管理交换元件2510的端口5到达VM 4。

[0353] 图30A-30B概念性示出了上文参考图29描述的逻辑交换机220和230、逻辑路由器225以及受管理交换元件2505和2510的示例操作。具体而言，图30A示出了实现逻辑交换机220和230以及逻辑路由器225的受管理交换元件2505的操作。图30B示出了受管理交换元件2505的操作。

[0354] 如图30A的下半部所示，受管理交换元件2505包括L2条目3005和3015以及L3条目3010。这些条目是控制器集群(未示出)提供给受管理交换元件2505的流条目。虽然这些条目被描绘为三个分开的表，但这些表不一定必须是分开的表。也就是说，单个表可包括所有这些流条目。

[0355] 当耦合到逻辑交换机220的VM 1向耦合到逻辑交换机230的VM 4发送分组3030时，该分组首先通过受管理交换元件2505的端口4被发送到受管理交换元件2505。受管理交换元件2505基于受管理交换元件2505的转发表3005-3015对分组执行L2处理。在此示例中，分组3030具有目的地IP地址1.1.2.10，其为VM 4的IP地址。分组3030的源IP地址是1.1.1.10。分组3030还以VM 1的MAC地址作为源MAC地址并且以逻辑路由器225的逻辑端口1的MAC地址(例如，01:01:01:01:01:01)作为目的地MAC地址。

[0356] 受管理交换元件2505识别转发表中的实现阶段2605的上下文映射的由带圈的1指

示的记录(称为“记录1”)。记录1基于进入端口识别分组3030的逻辑上下文,该进入端口是通过其从VM 1接收分组3030的端口4。此外,记录1指明受管理交换元件2505将分组3030的逻辑上下文存储在分组3030的头的一组字段(例如,VLAN id字段)中。记录1还指明通过转发表来进一步处理分组3030(例如通过将分组3030发送到调度端口)。调度端口在美国专利申请30/177,535中描述。

[0357] 基于分组3030的头中存储的逻辑上下文和/或其它字段,受管理交换元件2505识别转发表中的实现阶段2610的入口ACL的由带圈的2指示的记录(称为“记录2”)。在此示例中,记录2允许分组3030被进一步处理(即,分组3030可通过逻辑交换机220的入口端口),并且从而指明通过转发表来进一步处理分组3030(例如,通过将分组3030发送到调度端口)。此外,记录2指明受管理交换元件2505将分组3030的逻辑上下文(即,分组3030已被处理管道200的第二阶段2610处理)存储在分组3030的头的该组字段中。

[0358] 接下来,受管理交换元件2505基于分组3030的头中存储的逻辑上下文和/或其它字段来识别转发表中的实现阶段2615的逻辑L2转发的由带圈的3指示的记录(称为“记录3”)。记录3指明以逻辑路由器225的逻辑端口1的MAC地址为目的地MAC地址的分组要被发送到逻辑交换机220的逻辑端口X。

[0359] 记录3还指明通过转发表来进一步处理分组3030(例如,通过将分组3030发送到调度端口)。另外,记录3指明受管理交换元件2505将逻辑上下文存储在分组3030的头的该组字段中(即,分组3030已被处理管道200的第三阶段2615处理)。

[0360] 接下来,受管理交换元件2505基于分组3030的头中存储的逻辑上下文和/或其它字段识别转发表中的实现阶段2620的出口ACL的由带圈的4指示的记录(称为“记录4”)。在此示例中,记录4允许分组3030被进一步处理(例如,分组3030可通过逻辑交换机220的端口“X”离开逻辑交换机220),并且从而指明通过受管理交换元件2505的流条目来进一步处理分组3030(例如,通过将分组3030发送到调度端口)。此外,记录4指明受管理交换元件2505将分组3030的逻辑上下文(即,分组3030已被处理管道200的阶段2620处理)存储在分组3030的头的该组字段中。(要注意,所有记录都指明每当受管理交换元件基于记录执行逻辑处理的某个部分时,该受管理交换元件就更新该组字段中存储的逻辑上下文。)

[0361] 受管理交换元件2505基于流条目继续处理分组3030。受管理交换元件2505基于分组3030的头中存储的逻辑上下文和/或其它字段来识别L3条目3010中的由带圈的5指示的记录(称为“记录5”),该记录5通过基于分组3030的头中的信息指明受管理交换元件2505应当接受通过逻辑路由器225的逻辑端口1的分组来实现L3入口ACL。

[0362] 受管理交换元件2505随后识别L3条目3010中的由带圈的6指示的流条目(称为“记录6”),该流条目通过指明具有其目的地IP地址(例如1.1.2.10)的分组3030应当从逻辑路由器225的端口2离开来实现L3路由2640。另外,记录6(或路由表中的另一记录,未示出)指示分组3030的源MAC地址要被改写成逻辑路由器225的端口2的MAC地址(即,01:01:01:01:01:02)。

[0363] 受管理交换元件2505随后识别L3条目3010中的由带圈的7指示的流条目(称为“记录7”),该流条目通过基于分组3030的头中的信息(例如源IP地址)指明受管理交换元件2505允许分组通过逻辑路由器225的端口2离开来实现L3出口ACL。

[0364] 基于分组3030的头中存储的逻辑上下文和/或其它字段,受管理交换元件2505识

别L2条目3015中的实现阶段2660的入口ACL的由带圈的8指示的记录(称为“记录8”)。在此示例中,记录8指明分组3030被受管理交换元件2505进一步处理(例如通过将分组3030发送到调度端口)。此外,记录8指明受管理交换元件2505将分组3030的逻辑上下文(即,分组3030已被处理管道200的阶段2660处理)存储在分组3030的头的该组字段中。

[0365] 接下来,受管理交换元件2505基于分组3030的头中存储的逻辑上下文和/或其它字段来识别L2条目3015中的实现阶段2665的逻辑L2转发的由带圈的9指示的记录(称为“记录9”)。记录9指明以VM 4的MAC地址为目的地MAC地址的分组应当通过连接到VM 4的逻辑交换机230的逻辑端口(未示出)来转发。

[0366] 记录9还指明通过转发表来进一步处理分组3030(例如,通过将分组3030发送到调度端口)。另外,记录9指明受管理交换元件2505将逻辑上下文(即,分组3030已被处理管道200的阶段2665处理)存储在分组3030的头的该组字段中。

[0367] 接下来,受管理交换元件2505基于分组3030的头中存储的逻辑上下文和/或其它字段识别转发表中的实现阶段2670的出口ACL的由带圈的10指示的记录(称为“记录10”)。在此示例中,记录10允许分组3030通过连接到VM 4的逻辑端口(未示出)离开,并且从而指明通过转发表来进一步处理分组3030(例如通过将分组3030发送到调度端口)。此外,记录10指明受管理交换元件2505将分组3030的逻辑上下文(即,分组3030已被处理管道200的阶段2670处理)存储在分组3030的头的该组字段中。

[0368] 基于分组3030的头中存储的逻辑上下文和/或其它字段,受管理交换元件2505识别L2条目3015中的实现阶段2675的上下文映射的由带圈的11指示的记录(称为“记录11”)。在此示例中,记录11将与VM 4耦合的受管理交换元件2510的端口5识别为与分组3030要被转发到的逻辑交换机230的逻辑端口(在阶段2665确定)相对应的端口。记录11附加地指明通过转发表来进一步处理分组3030(例如通过将分组3030发送到调度端口)。

[0369] 基于分组3030的头中存储的逻辑上下文和/或其它字段,受管理交换元件2505随后识别L2条目3015中的实现阶段2680的物理映射的由带圈的12指示的记录(称为“记录12”)。记录12指明受管理交换元件2505的端口3作为通过其发送分组3030的端口以便分组3030到达受管理交换元件2510。在此情况下,受管理交换元件2505将把分组3030从与受管理交换元件2510耦合的受管理交换元件2505的端口3发送出去。

[0370] 如图30B所示,受管理交换元件2510包括转发表,该转发表包括用于处理和路由分组3030的规则(例如流条目)。当受管理交换元件2510从受管理交换元件2505接收到分组3030时,受管理交换元件2510基于受管理交换元件2510的转发表开始处理分组3030。受管理交换元件2510识别转发表中的实现上下文映射的由带圈的1指示的记录(称为“记录1”)。记录1基于分组3030的头中存储的逻辑上下文来识别分组3030的逻辑上下文。逻辑上下文指明分组3030已被由受管理交换元件2505执行的整个逻辑处理200处理。这样,记录4指明通过转发表来进一步处理分组3030(例如通过将分组3030发送到调度端口)。

[0371] 接下来,受管理交换元件2510基于分组3030的头中存储的逻辑上下文和/或其它字段,识别转发表中的实现物理映射的由带圈的2指示的记录(称为“记录2”)。记录2指明为了分组3030到达VM 4通过其发送分组3030的受管理交换元件2510的端口5。在此情况下,受管理交换元件2510将把分组3030从与VM 4耦合的受管理交换元件2510的端口5发送出去。在一些实施例中,受管理交换元件2510在将分组发送到VM 4之前从分组3030去除逻辑上下

文。

[0372] 图31概念性示出了受管理交换元件在其上运行的主机的示例软件体系结构。具体而言，此图示出了运行逻辑处理管道以逻辑地转发和路由分组的受管理交换元件使用NAT守护进程来转换网络地址。此图在其上半部示出了主机3100、受管理交换元件3105、转发表3120、NAT守护进程3110和NAT表3115。此图示出了流条目3125和3130。

[0373] 流条目3125和3130是各自具有限定符和动作的流条目。示为流条目3125和3130的文本可能不是实际的格式。相反，文本只是限定符和动作对的概念图示。在一些实施例中，流条目具有优先级，并且当多于一个流条目的限定符满足时，受管理交换元件采取具有最高优先级的流条目的动作。

[0374] 主机3100在一些实施例中是由能够运行一组软件应用的操作系统(例如，Windows™和Linux™)操作的机器。一些实施例的受管理交换元件3105是在主机3100中执行的软件交换元件(例如Open vSwitch)。如上所述，控制器集群(未示出)通过提供指明受管理交换元件的功能的流条目来配置受管理交换元件。一些实施例的受管理交换元件3105自身不生成流条目。

[0375] 一些实施例的受管理交换元件3105运行上述逻辑处理管道200的全部或一部分。特别地，受管理交换元件3105是执行L3处理210以根据需要基于转发表3120中的流条目对从机器接收的分组进行路由的受管理交换元件(例如，受管理交换元件1720或2505)。在一些实施例中，受管理交换元件3105是从耦合到受管理交换元件的机器(未示出)接收分组的边缘交换元件。在一些这样的实施例中，一个或多个虚拟机(未示出)在主机3100中运行并且耦合到受管理交换元件3105。在其它实施例中，受管理交换元件是第二级受管理交换元件。

[0376] 当受管理交换元件3105被配置为执行网络地址转换(NAT)时，一些实施例的受管理交换元件3105使用NAT守护进程3110来对分组执行NAT。在一些实施例中，受管理交换元件3105不维持查找表来找出从给定地址转换到的地址。反而，受管理交换元件3105向NAT守护进程3110询问地址。

[0377] 一些实施例的NAT守护进程3110是在主机3100上运行的软件应用。NAT守护进程3110维护表3115，该表3115包括地址配对(pairing)，其中每对包括要彼此转换的两个地址。当受管理交换元件3105寻求从给定地址转换到的地址时，NAT守护进程查找表3115以找出该给定地址应当被转换到的地址。

[0378] 不同实施例的受管理交换元件3105和NAT守护进程3110使用不同的技术来寻求和提供地址。例如，一些实施例的受管理交换元件3105向NAT守护进程发送分组，该分组具有原始地址，但不具有转换后的地址。这些实施例的NAT守护进程3110将原始地址转换成转换后的地址。NAT守护进程3110将分组发送回受管理交换元件3105，受管理交换元件3105将执行逻辑转发和/或路由以向目的地机器发送分组。在一些实施例中，受管理交换元件3105最初将元数据与包含要解析的原始地址的分组一起发送给NAT守护进程3110。此元数据包括受管理交换元件3105在其接收到从NAT守护进程3110返回的分组时用来继续执行逻辑处理管道的信息(例如，寄存器值、逻辑管道状态，等等)。

[0379] 在其它实施例中，一些实施例的受管理交换元件3105通过向NAT守护进程3110发送流模板(template)来请求地址，该流模板是不具有地址的实际值的流条目。NAT守护进程

通过查找表3115来找出地址以填写流模板。NAT守护进程3110随后通过将已填写的流模板放入转发表3120中来将填写了实际地址的流模板发送回到受管理交换元件3110。在一些实施例中，NAT守护进程向已填写的流模板指派比未填写的流模板的优先级值高的优先级值。另外，当NAT守护进程3110未能找到转换后的地址时，NAT守护进程将在流模板中指明丢弃分组。

[0380] 现在将按照三个不同的阶段1-3(带圈的1-3)来描述受管理交换元件3105和NAT守护进程3110的示例操作。在此示例中，受管理交换元件3105是从机器(未示出)接收要转发和路由的分组的受管理边缘交换元件。受管理交换元件3105接收分组并基于转发表3120中的流条目来执行L3处理210。

[0381] 在对分组执行L3处理210的同时，受管理交换元件3105(在阶段1)识别流条目3125并执行流条目3125中指明的动作。如图所示，流条目3125指示具有要被转换成X的IP地址1.1.1.10的流模板应当被发送到NAT守护进程3110。在此示例中，流条目3125具有优先级值N，其在一些实施例中是一数字。

[0382] 在阶段2，NAT守护进程3110接收流模板并通过查找NAT表3115查明1.1.1.10要被转换成2.1.1.10。NAT守护进程填写流模板并将已填写的模板(现在为流条目3130)插入转发表3120中。在此示例中，NAT守护进程向已填写的模板指派优先级N+1。

[0383] 在阶段3，受管理交换元件3110使用流条目3130来改变分组的地址。另外，对于受管理交换元件3105随后处理的分组，当分组具有源IP地址1.1.1.10时，受管理交换元件3105使用流条目3130而不是流条目3125。

[0384] 在一些实施例中，NAT守护进程3110和受管理交换元件运行于在主机3100上运行的同一虚拟机中或者在主机3100上运行的不同虚拟机中。NAT守护进程3110和受管理交换元件也可在分开的主机中运行。

[0385] 图32概念性示出了一些实施例执行来转换网络地址的过程3200。在一些实施例中，过程3200由执行L3处理210以在L3路由分组的受管理交换元件(例如受管理交换元件1720、2505或3105)执行。过程3200在一些实施例中在该过程接收到要在L3被逻辑路由的分组时开始。

[0386] 过程3200开始于(在3205)确定分组是否需要网络地址转换(NAT)。在一些实施例中，过程基于流条目来确定分组是否需要NAT。其限定符与分组的头或逻辑上下文中存储的信息匹配的流条目指明分组需要NAT。如上所述，NAT可以是SNAT或DNAT。流条目还将指明要对分组执行哪个NAT。

[0387] 当过程3200(在3205)确定分组不需要NAT时，过程结束。否则，过程3200(在3210)确定过程3200是否需要从NAT守护进程请求将分组的地址(例如源IP地址)转换成的地址。在一些实施例中，过程3200基于流条目来确定过程是否需要询问NAT守护进程。例如，流条目可指明将分组的地址转换成的地址应当通过从NAT守护进程请求该地址来获得。在一些实施例中，当流条目是对于转换后的地址具有空字段或者在该字段中具有指示应当从NAT守护进程获得转换后的地址的某个其它值的流模板时，过程确定NAT守护进程应当提供转换后的地址。

[0388] 当过程(在3210)确定过程不需要向NAT守护进程请求地址时，过程(在3220)从流条目获得转换后的地址。例如，流条目将提供转换后的地址。过程随后前进到3225，下文将

进一步描述3225。当过程(在3210)确定过程需要从NAT守护进程请求地址时,过程3200在3215从NAT守护进程请求并获得转换后的地址。在一些实施例中,过程3200通过向NAT守护进程发送流模板来请求转换后的地址。NAT守护进程将以转换后的地址来填写流模板并且将把已填写的流模板放入过程使用的转发表(未示出)中。

[0389] 接下来,过程3200(在3225)利用转换后的地址来修改分组。在一些实施例中,过程修改分组的头中的地址字段。可替代地或连带地,过程修改逻辑上下文以利用转换后的地址来替换分组的地址。过程随后结束。

[0390] 要注意,本申请中上文和下文使用的MAC地址、IP地址和其它网络地址是用于说明目的的示例,并且可不具有允许范围中的值,除非另有指明。

[0391] II.下一跳虚拟化

[0392] 与外部网络接合(interface)的逻辑网络需要与下一跳路由器交互。不同实施例的虚拟化应用使用不同的模型来使逻辑L3网络通过下一跳路由器与外部网络接合。

[0393] 第一,在固定附接模型中,物理基础设施与一组受管理集成(integration)元件交互,该组受管理集成元件将接收针对给定IP前缀的所有入口流量并且将把所有的出口流量发送回物理网络。在此模型中,对于每给定的一组受管理集成元件的逻辑L3路由器,逻辑抽象可以是单个逻辑上行链路端口。在一些实施例中,可以存在多于单个的集成集群。由控制应用提供的逻辑控制平面负责向上行链路路由出站(outbound)出口流量。在一些实施例中,受管理集成元件的示例包括用作扩展器的第二级受管理交换元件,扩展器在美国专利申请13/177,535中描述。受管理集成元件的示例还包括上文参考图8、图9和图10描述的受管理交换元件。

[0394] 第二,在分布式附接模型中,虚拟化应用遍及其连接的所有受管理边缘交换元件来分配附接。为此,受管理边缘交换元件必须集成到物理路由基础设施。换言之,每个受管理边缘交换元件必须能够与该组受管理交换元件之外的物理路由基础设施通信。在一些实施例中,这些交换元件使用IGP协议(或其它路由协议)来与将分组发送到(由受管理交换元件实现的)逻辑网络中并从逻辑网络接收分组的物理交换元件(例如物理路由器)通信。利用此协议,一些实施例的受管理边缘交换元件可通告主机路由(/32)以将直接入口流量吸引(attract)到其恰当位置。虽然在一些实施例中不存在集中式的流量热点(hotspot),因为入口和出口流量是完全分布在受管理交换元件上的,但逻辑抽象仍然是对于逻辑L3路由器的单个逻辑上行链路端口并且逻辑控制平面负责将流量路由到上行链路。没有什么妨碍为逻辑控制平面暴露(expose)多于单个的上行链路端口,如果这对于控制平面有益的话。然而,上行链路端口的数目在此模型中不必与附接点的数目匹配。

[0395] 第三,在控制平面驱动模型中,逻辑控制平面负责与外部网络集成。以一对一路由集成来暴露控制平面;对于物理网络中的每个附接点,存在一逻辑端口。逻辑控制平面负责在路由协议级与下一跳路由器对等。

[0396] 三个模型全都碰到了不同的设计权衡:固定附接模型意味着非最优的物理流量路由,但要求较少的与物理基础设施的集成。在分布式模型中,在一些实施例中,完全分布式模型扩展性最好,因为逻辑控制平面不负责所有对等流量,对等流量在极端情形中可能是成千上万的对等会话。然而,控制平面驱动模型对于逻辑控制平面给出了最大控制。不过,最大控制要求策略路由,因为如果需要最优物理路由,则出口端口必须依赖于入口端口。

[0397] III.状态分组(stateful packet)操作

[0398] 状态分组操作将NAT放置在被路由的流量的逻辑L3数据路径上。在逻辑管道中,网络地址转换在实际的标准L3管道之前或之后的另外的NAT阶段中进行。换言之,网络地址转换在路由之前或之后命中(hit)分组。在一些实施例中,NAT配置是经由创建实际地址转换条目的流模板进行的。流模板将在下文中进一步描述。

[0399] 放置NAT功能是与在第一跳中执行逻辑分组处理的全部或大部分的方法偏离的一个特征。在第一跳执行大部分或全部操作的基本模型在一些实施例中将对在相对方向上流动的分组的处理放置在不同的第一跳交换元件处:对于给定的传输级流,一个方向上的分组将通过一端的逻辑管道来发送,而相反方向上的分组将通过另一端的管道来发送。不幸的是,每个流NAT状态可能相当丰富(尤其如果NAT支持更高级应用协议的话),并且对于给定的传输流,必须在各方向之间共享状态。

[0400] 因此,一些实施例让逻辑端口的第一跳交换元件接收传输流的开放分组(opening packet)以对两个方向执行逻辑管道。例如,如果VM A向VM B打开TCP连接,则连接到VM A的超管理器的边缘交换元件(其可与超管理器在同一机器上运行)变得负责通过逻辑管道向两个方向发送分组。这允许了完全分布式的NAT功能,以及在逻辑网络拓扑中具有多个NAT。第一跳交换元件将执行所有必要的NAT转换,无论有多少个这样的NAT转换,并且网络地址转换只是变成分组(在该交换元件内)穿越的LDPS管道中的一个另外的步骤。

[0401] 然而,放置通过逻辑管道在相反方向上发送的分组的馈送要求额外的措施;否则,反向分组的第一跳交换元件将执行处理(而不使NAT状态本地可用)。为了此目的,一些实施例允许从(以上VM A的)源边缘交换元件发送到(以上VM B的)目的地边缘交换元件的第一分组建立特殊的“提示(hint)状态”,该提示状态使目的地交换元件不进行处理而直接将该传输流的反向分组发送到源交换元件。源交换元件随后将执行相反方向上的管道并且对于反向分组利用本地NAT状态反转NAT操作。一些实施例使用流模板(下文描述)在目的地交换元件处建立此反向提示状态,从而控制器不需要涉及每个流操作。

[0402] 接下来的两幅图图33和图34示出了放置NAT功能和提示状态。图33概念性示出了一些实施例的执行包括NAT操作2645在内的整个逻辑处理管道200的第一跳交换元件。图33与图29相同,除了逻辑处理管道200包括在L3处理220中描绘的NAT操作2645以指示NAT操作2645被执行。

[0403] 一些实施例的实现逻辑路由器的受管理交换元件在分组被逻辑路由器路由之后对分组执行NAT操作。例如,当耦合到逻辑交换机220的VM 1向耦合到逻辑交换机230的VM 4发送分组时,受管理交换元件2505在将分组送出到受管理交换元件2510之前将分组的源IP地址(例如1.1.1.10)转换成不同的IP地址(例如3.1.1.10)。受管理交换元件2505基于由管理受管理交换元件2505的控制器集群(未示出)在受管理交换元件2505中配置的一组NAT规则(例如流条目)来执行NAT操作2645。

[0404] VM 4接收到的分组以转换后的IP地址3.1.1.10作为分组的源IP地址。从VM 4到VM 1的返回分组将以这个转换后的地址作为分组的目的地IP地址。从而,转换后的IP地址必须被转换回VM 1的IP地址以便此分组到达VM 1。然而,一些实施例的受管理交换元件2510不会执行NAT操作2645来为返回的分组恢复VM 1的IP地址,因为用于执行NAT操作的NAT规则仅在受管理交换元件2505中,而不在受管理交换元件2510中。以这种方式,NAT规则和状态

不需要被所有潜在的受管理边缘交换元件共享。

[0405] 图34概念性示出了这种实施例的示例。具体而言，图34示出了当向受管理交换元件2505发送返回分组时受管理交换元件2510不执行逻辑处理管道。此图还示出了受管理交换元件2505在接收到来自受管理交换元件2510的返回分组时执行逻辑处理管道200，就好像受管理交换元件2505相对于此返回分组为第一跳交换元件那样。图34与图33相同，除了逻辑处理管道是在相对方向上被描绘(箭头指向左)。图34还示出了规则3400和转发表3405。

[0406] 规则3400在一些实施例中是转发表3405中的由管理受管理交换元件2510的控制器集群(未示出)配置的流条目。规则3400指明(或“提示”)当受管理交换元件2510接收到源自受管理交换元件2505的分组时，受管理交换元件2510不应当对去往受管理交换元件2505的返回分组执行逻辑处理管道。

[0407] 当受管理交换元件2510从受管理交换元件2505接收到受管理交换元件2505已对其执行了NAT操作的分组时，受管理交换元件2510基于分组的头中包括的信息(例如逻辑上下文)来找出规则3400。另外，受管理交换元件2510在一些实施例中修改一个或多个其它流条目以指示不应当对来自前往源机器(例如VM 1)的接收分组的目的地机器(例如VM 4)的分组执行逻辑处理管道。

[0408] 受管理交换元件2510随后将此分组转发到目的地机器，例如VM 4。当受管理交换元件2510从VM 4接收到要前往VM 1的返回分组时，受管理交换元件2510将不对此分组执行逻辑处理管道。也就是说，受管理交换元件2510将不执行L2的逻辑转发或L3的逻辑路由。受管理交换元件2510将简单地在此分组的逻辑上下文中指示未对该分组执行逻辑处理。

[0409] 当受管理交换元件2505从受管理交换元件2510接收到此分组时，受管理交换元件2505执行逻辑处理管道200。具体而言，受管理交换元件2505首先执行逻辑上下文查找以基于分组的头字段中包括的信息来确定分组的逻辑上下文。在此示例中，分组的源MAC地址是VM 4的MAC地址，并且分组的源IP地址是VM 4的IP地址。因为分组被从VM 4发送到在不同逻辑网络中的VM 1，所以分组以与逻辑交换机230的端口Y相关联的MAC地址作为目的地MAC地址(即，在此示例中为01:01:01:01:01:02)。分组的目的地IP地址是VM 1的经NAT的IP地址(即，3.1.1.10)。

[0410] 受管理交换元件2505随后相对于逻辑交换机230为分组确定访问控制。例如，受管理交换元件2505确定分组不具有将使得逻辑交换机230拒绝经过逻辑交换机230的端口2的分组的网络地址(例如，源/目的地MAC/IP地址，等等)。受管理交换元件2505还从逻辑上下文识别出逻辑交换机230的端口Y是发送出分组的端口。另外，受管理交换元件2505相对于端口Y为分组确定访问控制。例如，受管理交换元件2505确定分组不具有将使得逻辑交换机230不通过端口Y发送分组的网络地址。

[0411] 接下来，受管理交换元件2505对分组执行NAT操作2645以将目的地IP地址转换回VM 1的IP地址。也就是说，受管理交换元件2505在此示例中基于NAT规则利用1.1.1.10来替换3.1.1.10。受管理交换元件2505随后对分组执行L3处理，因为分组的目的地IP地址(现在为1.1.1.10)是针对另一逻辑网络的。受管理交换元件2505相对于逻辑路由器225的端口2在L3为分组确定入口访问控制。受管理交换元件2505还查找流条目并确定分组要被发送到逻辑路由器225的逻辑端口1，因为分组的目的地IP地址1.1.1.10属于与逻辑路由器225的

逻辑端口1相关联的1.1.1.1/24的子网地址。另外，受管理交换元件2505相对于逻辑路由器225的逻辑端口1为分组确定出口访问控制。受管理交换元件2505还将分组的目的地MAC地址修改为VM 1的MAC地址。

[0412] 受管理交换元件2505随后执行L2处理215。在此示例中，分组的源MAC地址现在是逻辑路由器225的逻辑端口1的MAC地址并且分组的源IP地址仍是VM 4的IP地址。分组的目的地IP地址是VM 1的IP地址(即，1.1.1.10)。在此示例中，逻辑上下文指明逻辑交换机220是要转发分组的逻辑交换机并且逻辑交换机220的逻辑端口X是通过其接收分组的端口。逻辑上下文还指明逻辑交换机220的端口1是通过其将分组送出到目的地VM 1的端口，因为端口1与VM1的MAC地址相关联。

[0413] 受管理交换元件2505随后基于分组的逻辑上下文执行逻辑转发查找，包括分别相对于逻辑交换机220的端口X和端口1确定入口和出口访问控制。受管理交换元件2505执行映射查找来确定逻辑交换机220的逻辑端口1被映射到的物理端口。在此示例中，受管理交换元件2505确定逻辑交换机220的逻辑端口1被映射到受管理交换元件2505的端口4。受管理交换元件2505随后执行物理查找来确定用于将分组转发到物理端口的操作。在此示例中，受管理交换元件2505确定分组应当通过受管理交换元件2505的端口4被发送到VM 1。

[0414] 图35概念性示出了一些实施例执行来将分组发送到其地址被NAT的目的地机器的过程3500。过程3500在一些实施例中由直接从源机器接收分组的受管理边缘交换元件执行。

[0415] 过程3500开始于从源机器接收分组(在3505)。过程随后(在3510)确定分组是否要前往其地址被NAT的目的地机器。在一些实施例中，过程通过查找与分组的头中包括的信息(例如目的地IP地址)匹配的流条目来确定分组是否要前往这样的目的地机器。一个或多个流条目指明当分组被定址到其地址被NAT的目的地机器时，不应当对此分组执行逻辑处理(例如，L2的逻辑转发或L3的逻辑路由)。其它流条目指明当分组被定址到其地址被NAT的目的地机器时应当执行逻辑处理。

[0416] 当过程3500(在3510)确定分组要前往其地址被NAT的目的地机器时，过程3515前进到下文将进一步描述的3520。当过程3500(在3510)确定分组要前往其地址未被NAT的目的地机器时，过程3500对分组执行逻辑处理(例如，L2的逻辑转发和/或L3的逻辑路由)。

[0417] 过程3500随后(在3520)将分组发送到去往目的地机器的路线中的下一跳受管理交换元件。过程3500随后结束。

[0418] 如上所述，在每个分组操作中不涉及控制器。逻辑控制平面仅配设识别什么应当被进行网络地址转换的FIB规则。所有每个流状态都是由数据路径(Open vSwitch)建立的。

[0419] 上述实施例利用了源NAT。然而，一些实施例同时使用目的地NAT(DNAT)。在DNAT的情况下，所有处理都可在源受管理边缘交换元件处完成。

[0420] 另外，在将NAT功能放置在外部网络和逻辑网络之间的情况下，操作与以上所述的没有不同。在此情况下，对于从外部网络进入的流，对于两个方向都将在扩展器(其在此情况下将是第一跳受管理边缘交换元件)保持NAT状态。另一方面，对于朝着外部网络发起的传输流，将在附接到发端主机/VM的受管理边缘交换元件处保持状态。

[0421] 在用于网络地址转换的这个完全分布式方案的情况下，VM移动性支持要求将与VM的建立的NAT状态迁移(migrate)到新的超管理器。如果不迁移NAT状态，则传输连接将断

开。对于这种情况，一些实施例被设计为期望NAT对于发送到已关闭/不存在的TCP流的分组以TCP复位来响应。更高级的实现将与促成NAT状态与VM一起迁移的VM管理系统集成；在此情况下，传输连接不必断开。

[0422] 图36示出了当VM从第一主机迁移到第二主机时将NAT状态从第一主机迁移到第二主机的示例。具体而言，该图示出了使用第一主机的超管理器来迁移VM和与VM相关联的NAT状态。该图示出了两个主机3600和3630。

[0423] 如图所示，主机3600在此示例中是源主机，VM 3625从该源主机迁移到主机3630。在主机3600中，NAT守护进程3610和受管理交换元件3605在运行。NAT守护进程3610与上文参考图31描述的NAT守护进程3110类似。NAT守护进程3610维护包括原始地址和转换后地址的映射的NAT表3115。受管理交换元件3605使用NAT守护进程3610来获得转换后的地址。受管理交换元件在一些实施例中将流模板发送到NAT守护进程3610以如上所述发送原始地址并获得转换后的地址。

[0424] 超管理器3680创建并管理在主机3600中运行的VM。在一些实施例中，超管理器3680将在主机3600中运行的VM迁移出主机3600在该VM迁移到另一主机之前通知给受管理交换元件3605和/或NAT守护进程3610。受管理交换元件3605和/或NAT守护进程3610在一些实施例中通过针对在VM迁移情况下的回调(callback)进行注册来获得这种通知。

[0425] 在一些这样的实施例中，受管理交换元件3605要求NAT守护进程取得与迁移的VM相关联的NAT状态(例如，协议信息和VM的地址映射，等等)并将NAT状态提供给超管理器3680。在一些实施例中，当超管理器3680将迁移直接通知给NAT守护进程3610时，NAT守护进程3610将与迁移的VM相关联的NAT状态提供给超管理器3680。超管理器3680随后将NAT状态与迁移的VM一起迁移到目的地主机。

[0426] 在一些实施例中，NAT守护进程3610将与迁移的VM相关联的NAT状态直接发送到在目的地主机中运行的NAT守护进程。在这些实施例中，NAT守护进程3610和/或受管理交换元件3605将NAT状态的迁移的完成通知给超管理器3680，以便超管理器3680可开始将VM迁移到目的地主机。

[0427] 在一些实施例中，受管理交换元件3605还将与迁移的VM有关的流条目提供给超管理器3680或提供给在目的地主机中运行的受管理交换元件。当超管理器3680被提供这些流条目时，超管理器3680将这些流条目发送给在目的地主机中运行的受管理交换元件的流表。流条目到目的地主机的迁移是可选的，因为单独NAT状态就将使得在目的地主机中运行的受管理交换元件能够获得用于迁移的VM的转换后地址。

[0428] 现在将描述源主机3600的示例操作。当超管理器3680要迁移VM 3625时(例如按照用户输入或来自控制集群的输入)，超管理器3680通知受管理交换元件3605。受管理交换元件3605在此示例中随后要求NAT守护进程3610取得与VM 3625相关联的NAT状态并将所取得的状态发送到超管理器3680。

[0429] 超管理器3680随后通过移动VM的数据来将VM 3625迁移到目的地主机3630。在一些实施例中，超管理器3680能够通过捕获VM 3625的运行状态并将该状态发送到VM 3625来进行实时(live)迁移。超管理器3680还将所取得的NAT状态移动到主机3630的NAT表3645，以便在主机3630中运行的受管理交换元件3635能够为刚被迁移到主机3630中的VM 3625从NAT守护进程3640获得转换后的地址。

[0430] 图37示出了当VM从第一主机迁移到第二主机时将NAT状态从第一主机迁移到第二主机的另一示例。具体而言，该图示出了使用控制集群来要求第一主机的超管理器取得与迁移的VM相关联的NAT状态并将NAT状态发送给第二主机。该图示出了两个主机3600和3630。然而，在主机3600中运行的超管理器3680在此示例中不支持向受管理交换元件或在源主机中运行的NAT守护进程作出通知。

[0431] 因为一些实施例的超管理器3680不将VM迁移到目的地主机通知给受管理交换元件或NAT守护进程，所以与迁移的VM相关联的NAT状态在超管理器3680开始或完成将VM迁移到目的地主机之后被发送到目的地主机。特别地，受管理交换元件3635在一些实施例中将通过例如检测3625的MAC地址来检测VM 3625的迁移，该MAC地址对于受管理交换元件3635而言是新的。受管理交换元件3635将VM 3625的添加(因此用于VM 3625的受管理交换元件3635的新端口)通知给控制集群3705。

[0432] 控制集群3705与上文描述的控制集群1105和2205类似。在接收到来自受管理交换元件3635的关于VM的添加的通知时，控制集群3705要求在源主机3600中运行的超管理器3680取得与被迁移的VM 3625相关联的NAT状态并利用取得的NAT状态来更新NAT表3645。在一些实施例中，控制集群3705附加地要求取得与被迁移的VM 3625相关联的流条目并把这些流条目放入目的地3630的流表3650中。

[0433] 在一些实施例中，控制集群3705可直接要求受管理交换元件和/或NAT守护进程3610将NAT状态和/或流条目发送给NAT守护进程3640和/或受管理交换元件3635以便利用与被迁移的VM 3625相关联的NAT状态和/或流条目来更新NAT表3645和/或3650。

[0434] 现在将描述源主机3600、目的地主机3630和控制集群3705的示例操作。当超管理器3680要迁移VM 3625时(例如，按照用户输入或来自控制集群的输入)，超管理器3680通过将VM 3625的配置数据或运行状态移动到主机3630来迁移VM 3625。现在在主机3630中运行的VM 3625向受管理交换元件3635发送分组。受管理交换元件3635在此示例中通过辨认出分组的源MAC地址对于受管理交换元件3635是新的来检测VM 3625到主机3630的迁移。受管理交换元件3605在此示例中随后将VM 3625的添加(或者对于VM 3625的新端口的创建)通知给控制集群3705。

[0435] 控制集群3705随后要求超管理器3680取得与VM 3625相关联的NAT状态并将该NAT状态发送给目的地主机3630。在目的地主机3630中运行的受管理交换元件3635可为刚迁移到主机3630中的VM 3625从NAT守护进程3640获得转换后的地址。

[0436] IV.负载均衡

[0437] 作为L3管道中的另外步骤，一些实施例实现负载均衡。例如，一些实施例实现基于逻辑捆绑(bundle)的负载均衡步骤，其后是目的地网络地址转换。在一些实施例中，(提供负载均衡服务的)逻辑路由器容宿虚拟IP地址，因此将对发送到该虚拟IP地址(VIP)的ARP请求作出响应。这样，即使流量被从集群成员所在的同一L2域发送到VIP，虚拟IP也会保持可工作。

[0438] 图38示出了执行负载均衡的逻辑路由器和逻辑交换机的示例物理实现。特别地，该图示出了集中式L3路由模型，其中逻辑路由器由L3路由器或受管理交换元件基于流条目实现。该图示出了受管理交换元件3805-3825和VM 3830-3850。该图还示出了包括L2处理3855、DNAT和负载均衡3860、L3路由3865以及L2处理3870和3875的逻辑处理管道。

[0439] 一些实施例的受管理交换元件3805是用作扩展器的第二级受管理交换元件。一些这样的实施例中的受管理交换元件3805与上文描述的受管理交换元件810和1910的相似之处在于受管理交换元件3805基于流条目(未示出)实现逻辑路由器(未示出)或者在实现逻辑路由器的L3路由器在其上运行的同一主机中运行。此外,受管理交换元件3805执行DNAT和负载均衡3860以将目的地地址转换成另一地址并在提供同一服务(例如web服务)的不同机器(例如VM)之间均衡负载。

[0440] 受管理交换元件3805-3825实现与VM 3830-3850连接的逻辑交换机(未示出)。VM 3840和3850在此示例中提供同一服务。也就是说,VM 3840和3850在一些实施例中共同充当提供同一服务的服务器。然而,VM 3840和3850是具有不同的IP地址的分开的VM。受管理交换元件3805或受管理交换元件3805使用的L3路由器(未示出)执行负载均衡以在VM 3840和3850之间分配工作负载。

[0441] 在一些实施例中,负载均衡通过将请求服务的分组的目的地地址转换成提供该服务的VM的不同地址来实现。特别地,受管理交换元件3805或受管理交换元件3805使用的L3路由器(未示出)将请求分组的目的地地址转换成若干VM 3840和3850的地址,以使得这些VM中没有特定的VM获得比其它VM多得多的工作负载。关于找出提供服务的VM的当前工作负载的更多细节将在下文中进一步描述。

[0442] 在一些实施例中,受管理交换元件3805或L3路由器在执行逻辑处理管道的负载均衡3860和DNAT之后执行L3路由3865。因此,受管理交换元件3805或L3路由器在这些实施例中基于转换后的目的地地址将分组路由到不同的受管理交换元件。受管理交换元件3820和3825是边缘交换元件,并且从而直接向VM 3840和3850发送和从VM 3840和3850接收分组。在其它实施例中,受管理交换元件3805或L3路由器在执行逻辑处理管道的负载均衡3860和DNAT之前执行L3路由3865。

[0443] 现在将描述受管理交换元件3805的示例操作。受管理交换元件3810接收请求由VM 3840和3850共同提供的服务的分组。此分组来自VM 3830中的一个,具体而言来自使用特定协议的应用。该分组在此示例中包括识别特定协议的协议号。该分组还包括表示提供服务的服务器的IP地址作为目的地IP地址。为了描述简单起见,对此分组执行源L2处理3855的细节被省略,因为其与上文和下文描述的源L2处理示例类似。

[0444] 在执行源L2处理3855以将分组路由到受管理交换元件3805以执行包括L3路由3865的L3处理之后。在此示例中,受管理交换元件3805对分组执行DNAT和负载均衡3860。也就是说,受管理交换元件3805将分组的目的地IP地址转换成提供服务的VM中的一个的IP地址。在此示例中,受管理交换元件3805选择VM 3840-3850中的在它们之中具有最小工作负载的一个。受管理交换元件3805基于新的目的地IP地址对分组执行L3路由3865(即,路由分组)。

[0445] 受管理交换元件3820接收分组,因为目的地IP地址具有VM 3840中的一个,并且此目的地IP被解析成该VM的MAC地址。受管理交换元件3820将分组转发到VM。此VM将把分组返回给原本请求服务的应用。这些返回的分组将到达受管理交换元件3805,并且受管理交换元件3805将执行NAT并识别出该应用是这些分组的目的地。

[0446] 图39示出了执行负载均衡的逻辑路由器和逻辑交换机的另一示例物理实现。特别地,此图示出了分布式L3路由模型,其中逻辑路由器由也执行源和目的地L2处理的受管理

交换元件实现。也就是说，此受管理交换元件执行整个逻辑处理管道。此图示出了受管理交换元件3905和3820-3825以及VM 3910和3840-3850。此图还示出了包括L2处理3855、DNAT和负载均衡3860、L3路由3865以及L2处理3870-3875的逻辑处理管道。

[0447] 一些实施例的受管理交换元件3905与上文参考图29描述的受管理交换元件2505的相似之处在于受管理交换元件3905实现整个逻辑处理管道。也就是说，受管理交换元件3905实现逻辑路由器和逻辑交换机。此外，受管理交换元件3905执行DNAT和负载均衡3860以将目的地地址转换成另一地址并在提供同一服务(例如web服务)的不同机器(例如VM)之间均衡负载。

[0448] 如上所述，受管理交换元件3905实现与VM 3910和3840-3850连接的逻辑交换机(未示出)。受管理交换元件3905还执行负载均衡以在VM 3840和3850之间分配工作负载。特别地，受管理交换元件3905将请求分组的目的地地址转换成若干VM 3840和3850的地址，以使得这些VM中没有特定的VM得到比其它VM多得多的工作负载。关于找出提供服务的VM的当前工作负载的更多细节将在下文中进一步描述。

[0449] 在一些实施例中，受管理交换元件3905在执行逻辑处理管道的DNAT和负载均衡3860之后执行L3路由3865。因此，受管理交换元件3905基于转换后的目的地地址将分组路由到不同的受管理交换元件。受管理交换元件3820和3825是边缘交换元件，并且从而直接向VM 3840和3850发送和从VM 3840和3850接收分组。在其它实施例中，受管理交换元件3905在执行逻辑处理管道的DNAT和负载均衡3860之前执行L3路由3865。

[0450] 受管理交换元件3905的操作将与上文参考图38描述的示例操作类似，除了受管理交换元件3905执行包括DNAT和负载均衡3860在内的整个逻辑处理管道。

[0451] 图40示出了执行负载均衡的逻辑路由器和逻辑交换机的另一示例物理实现。特别地，此图示出了分布式L3路由模型，其中逻辑路由器由也执行源L2处理的受管理交换元件实现。也就是说，此受管理交换元件作为第一跳受管理交换元件执行源L2处理和L3处理。目的地L2处理由作为最末跳受管理交换元件的另一受管理交换元件执行。此图示出了受管理交换元件4005和3820-3825以及VM 4010和3840-3850。此图还示出了包括L2处理3855、DNAT和负载均衡3860、L3路由3865以及L2处理3870-3875的逻辑处理管道。

[0452] 一些实施例的受管理交换元件4005与上文参考图46描述的受管理交换元件2505的相似之处在于受管理交换元件4005执行逻辑处理管道的源L2处理和L3处理。也就是说，受管理交换元件4005实现逻辑路由器和与源机器相连的逻辑交换机。此外，受管理交换元件4005执行DNAT和负载均衡3860以将目的地地址转换成另一地址并在提供同一服务(例如web服务)的不同机器(例如VM)之间均衡负载。

[0453] 如上所述，受管理交换元件4005实现与VM 4010中的一个或多个连接的逻辑交换机(未示出)。受管理交换元件4005还执行负载均衡以在VM 3840和3850之间分配工作负载。特别地，受管理交换元件4005将请求分组的目的地地址转换成若干VM 3840和3850的地址，以使得这些VM中没有特定的VM得到比其它VM多得多的工作负载。关于找出提供服务的VM的当前工作负载的更多细节将在下文中进一步描述。

[0454] 在一些实施例中，受管理交换元件4005在执行逻辑处理管道的DNAT和负载均衡3860之后执行L3路由3865。因此，受管理交换元件4005基于转换后的目的地地址将分组路由到不同的受管理交换元件。受管理交换元件3820和3825是边缘交换元件，并且从而直接

向VM 3840和3850发送和从VM 3840和3850接收分组。在其它实施例中，受管理交换元件4005在执行逻辑处理管道的DNAT和负载均衡3860之前执行L3路由3865。

[0455] 受管理交换元件4005的操作将与上文参考图38描述的示例操作类似，除了不同的受管理交换元件执行逻辑处理管道的不同部分。

[0456] 图41概念性示出了在共同提供服务(例如web服务)的机器之间均衡负载的负载均衡守护进程。具体而言，此图示出了运行逻辑处理管道以逻辑地转发和路由分组的受管理交换元件使用负载均衡守护进程来在提供服务的机器之间均衡工作负载。此图在其上半部示出了主机4100、受管理交换元件4105、转发表4120、负载均衡守护进程4110和连接表4115。此图示出了流条目4125和4130。

[0457] 流条目4125和4130各自具有限定符和动作。示为流条目4125和4130的文本可能不是实际的格式。相反，文本只是限定符和动作对的概念图示。主机4100在一些实施例中是由能够运行一组软件应用的操作系统(例如，Windows™和Linux™)操作的机器。一些实施例的受管理交换元件4105是在主机4100中执行的软件交换元件(例如Open vSwitch)。如上所述，控制器集群(未示出)通过提供指明受管理交换元件的功能的流条目来配置受管理交换元件。一些实施例的受管理交换元件4105自身不生成流条目。

[0458] 一些实施例的受管理交换元件4105运行上文参考图38-40描述的逻辑处理管道的全部或一部分。特别地，受管理交换元件4105执行L3处理以根据需要基于转发表4120中的流条目对从机器接收的分组进行路由。在一些实施例中，受管理交换元件4105是从耦合到受管理交换元件的机器(未示出)接收分组的边缘交换元件。在一些这样的实施例中，一个或多个虚拟机(未示出)在主机4100中运行并且耦合到受管理交换元件4105。

[0459] 当受管理交换元件4105被配置为执行负载均衡时，一些实施例的受管理交换元件4105使用负载均衡守护进程4110来对分组执行负载均衡。负载均衡守护进程4110与NAT守护进程3110的相似之处在于负载均衡守护进程4110提供转换后的目的地地址(例如，目的地IP地址)。此外，负载均衡守护进程4110基于表4115中包括其IP地址的机器的当前负载来选择要将原始目的地地址转换到的目的地。

[0460] 一些实施例的负载均衡守护进程4110是在主机4100上运行的软件应用。负载均衡守护进程4110维护连接表4115，该连接表4115包括提供服务的机器的可用地址和连接识别符的配对。虽然没有描绘出，但一些实施例的连接表4115还可包括为与地址相关联的机器量化的当前的工作负载。在一些实施例中，负载均衡守护进程4110周期性地与提供服务的VM通信以得到VM的更新状态，其中包括VM的当前工作负载。

[0461] 当受管理交换元件4105寻求基于连接识别符来选择的地址时，负载均衡守护进程在一些实施例中查找表4115以找出给定的目的地地址应当被转换成的地址。在一些实施例中，负载均衡守护进程运行调度方法来识别服务器VM以便在服务器VM之间均衡负载。这种调度算法考虑与地址相关联的机器的当前负载。负载均衡方法的更多细节和示例在美国临时专利申请61/560,279中描述，这里通过引用并入该申请。

[0462] 连接识别符唯一地识别服务的请求者(即，分组的起源或来源)和最后提供所请求的服务的机器之间的连接，以便从机器返回的分组能够被准确地中继回请求者。这些返回的分组的源IP地址将被转换回表示提供服务的服务器的IP地址(称为“虚拟IP地址”)。这些连接识别符之间的映射也将用于随后从该来源发送的分组。在一些实施例中，连接识别符

包括源端口、目的地端口、源IP地址、目的地IP地址、协议识别符，等等。源端口是从其发送分组的端口(例如，TCP端口)。目的地端口是分组被发送到的端口。协议识别符识别用于格式化分组的协议(例如，TCP、UDP等等)的类型。

[0463] 不同实施例的受管理交换元件4105和负载均衡守护进程4110使用不同的技术来寻求和提供地址。例如，一些实施例的受管理交换元件4105将具有原始地址但不具有转换后地址的分组发送给负载均衡守护进程。这些实施例的负载均衡守护进程4110将原始地址转换成转换地址。负载均衡守护进程4110将分组发送回受管理交换元件4105，受管理交换元件4105将执行逻辑转发和/或路由以向目的地机器发送分组。在一些实施例中，受管理交换元件4105最初将元数据与包含要解析的原始地址的分组一起发送给负载均衡守护进程4110。此元数据包括受管理交换元件4105在其接收到从负载均衡守护进程4110返回的分组时用来继续执行逻辑处理管道的信息(例如，寄存器值、逻辑管道状态，等等)。

[0464] 在其它实施例中，一些实施例的受管理交换元件4105通过向负载均衡守护进程4110发送流模板来请求地址，流模板是不具有地址的实际值的流条目。负载均衡守护进程通过查找表4115来查明地址以填写流模板。负载均衡守护进程4110随后通过将已填写的流模板放入转发表4120中来将填入了实际地址的流模板发送回到受管理交换元件4110。在一些实施例中，负载均衡守护进程向已填写的流模板指派比未填写的流模板的优先级值高的优先级值。另外，当负载均衡守护进程4110未能找到转换后的地址时，负载均衡守护进程将在流模板中指明丢弃分组。

[0465] 现在将按照三个不同的阶段1-3(带圈的1-3)来描述受管理交换元件4105和负载均衡守护进程4110的示例操作。在此示例中，受管理交换元件4115是从机器(未示出)接收要转发和路由的分组的受管理边缘交换元件。具体地，分组在此示例中是针对服务的请求。分组具有表示提供所请求的服务的服务器的IP地址。

[0466] 受管理交换元件4105接收此分组并基于转发表4120中的流条目来执行L3处理。在对分组执行L3处理210的同时，受管理交换元件4105(在阶段1)识别流条目4125并执行流条目4125中指明的动作。如图所示，流条目4125指示具有连接识别符的流模板应当被发送到负载均衡守护进程4110以让负载均衡守护进程4110提供新的目的地IP地址。在此示例中，流条目4125具有优先级值N，其在一些实施例中为一数字。

[0467] 在阶段2，负载均衡守护进程4110接收流模板并通过查找连接表4115并运行调度算法来查明具有指明的连接ID的分组的目的地IP地址要被转换成2.1.1.10。负载均衡守护进程填写流模板并将已填写的模板(现在是流条目4130)插入到转发表4130中。在此示例中，负载均衡守护进程向已填写的模板指派优先级N+I。

[0468] 在阶段3，受管理交换元件4110使用流条目4130来改变分组的目的地IP地址。另外，对于受管理交换元件4105随后处理的分组，当分组具有指明的连接识别符时，受管理交换元件4105使用流条目4130而不是流条目4125。

[0469] 在一些实施例中，负载均衡守护进程4110和受管理交换元件运行于在主机4100上运行的同一虚拟机中或者在主机4100上运行的不同虚拟机中。负载均衡守护进程4110和受管理交换元件也可在分开的主机中运行。

[0470] V.DHCP

[0471] 虚拟化应用在一些实施例中定义将DHCP请求路由到在共享主机中运行的DHCP守

护进程的转发规则。对于此功能使用共享主机避免了对每个客户运行一DHCP守护进程的另外的成本。

[0472] 图42示出了为不同用户向不同逻辑网络提供DHCP服务的DHCP守护进程。此图在其左半部示出了分别为两个不同用户A和B实现示例逻辑网络4201和4202。逻辑网络4201和4202的示例物理实现在该图的右半部示出。

[0473] 如该图的左半部所示，逻辑网络4201包括逻辑路由器4205和两个逻辑交换机4210和4215。VM 4220和4225连接到逻辑交换机4210。也就是说，VM 4220和4225发送和接收被逻辑交换机4210转发的分组。VM 4230连接到逻辑交换机4215。逻辑路由器4205在逻辑交换机4210和4215之间路由分组。逻辑路由器4205还连接到DHCP守护进程4206，DHCP守护进程4206向逻辑网络4201中的VM提供DHCP服务，逻辑网络4201中的VM是用户A的VM。

[0474] 用户B的逻辑网络4202包括逻辑路由器4235以及两个逻辑交换机4240和4245。VM 4250和4255连接到逻辑交换机4240。VM 4260连接到逻辑交换机4245。逻辑路由器4235在逻辑交换机4240和4245之间路由分组。逻辑路由器4235还连接到DHCP守护进程4236，DHCP守护进程4236向逻辑网络4202中的VM提供DHCP服务，逻辑网络4202中的VM是用户B的VM。

[0475] 在该图的左半部所示的逻辑实现中，用户的每个逻辑网络具有其自己的DHCP守护进程。在一些实施例中，DHCP守护进程4206和4236可在物理上实现为在不同的主机或VM中运行的分开的DHCP守护进程。也就是说，每个用户将具有仅用于该用户的机器的单独的DHCP守护进程。

[0476] 在其它实施例中，用于不同用户的DHCP守护进程可在物理上实现为向不同用户的VM提供DHCP服务的单个DHCP守护进程。也就是说，不同的用户共享同一DHCP守护进程。DHCP守护进程4270是为用户A和B两者的VM服务的共享DHCP守护进程。如该图的右半部所示，为用户A和B实现逻辑路由器4205和4235以及逻辑交换机4210、4215、4240和4245的受管理交换元件4275-4285使用单个DHCP守护进程4270。因此，用户A和B的VM 4220-4260使用DHCP守护进程4270来动态地获得地址(例如IP地址)。

[0477] 不同实施例的DHCP守护进程4270可在不同主机中运行。例如，一些实施例的DHCP守护进程4270在受管理交换元件4275-4285中的一个在其中运行的同一主机(未示出)中运行。在其它实施例中，DHCP守护进程4270不在受管理交换元件在其上运行的主机中运行，而是在受管理交换元件可访问的单独的主机中运行。

[0478] 图43示出了一中央DHCP守护进程和若干本地DHCP守护进程。中央DHCP守护进程通过本地DHCP守护进程向不同用户的VM提供DHCP服务。每个本地DHCP守护进程维护和管理一批地址以将中央DHCP守护进程的服务卸载(offload)到本地DHCP守护进程。此图示出了包括中央DHCP守护进程4320和两个本地DHCP守护进程4330和4350的示例体系结构。

[0479] 如图所示，中央DHCP守护进程4320在主机4305中运行，受管理交换元件4306也在该主机4305中运行。一些实施例的受管理交换元件4306是对于受管理交换元件4340和4360用作池节点的第二级受管理交换元件。中央DHCP守护进程4320向不同用户的不同VM 4345和4365提供DHCP服务。在一些实施例中，中央DHCP守护进程4320将各批地址中的可用地址(例如IP地址)4325分配到包括本地DHCP守护进程4330和4350在内的不同本地DHCP守护进程以便将DHCP服务卸载到这些本地DHCP守护进程。当本地DHCP守护进程在其自己的那批地址中用完了可指派的可用地址时，中央DHCP守护进程4320向该本地DHCP守护进程提供更多

地址。

[0480] 本地DHCP守护进程4330在主机4310中运行，受管理交换元件4340也在主机4310中运行。受管理交换元件4340是直接向VM 4345发送和从VM 4345接收分组的边缘交换元件。受管理交换元件4340实现不同用户的一个或多个逻辑交换机和逻辑路由器。也就是说，VM 4345可属于不同用户。本地DHCP守护进程4330利用本地DHCP守护进程4330从中央DHCP守护进程4320获得的那批地址4335向VM 4345提供DHCP服务。当本地DHCP守护进程4330在该批地址4335中用完了指派的可用地址时，本地DHCP守护进程4330求助于中央DHCP守护进程4320。在一些实施例中，本地DHCP守护进程4330经由受管理交换元件4340和4306与中央DHCP守护进程4320通信。在一些实施例中受管理交换元件4340和4306具有在它们之间建立的隧道。

[0481] 类似地，本地DHCP守护进程4350在主机4315中运行，受管理交换元件4360也在主机4315中运行。受管理交换元件4360是直接向VM 4365发送和从VM 4365接收分组的边缘交换元件。受管理交换元件4360实现不同用户的一个或多个逻辑交换机和逻辑路由器。本地DHCP守护进程4350利用本地DHCP守护进程4350从中央DHCP守护进程4320获得的那批地址4355向VM 4365提供DHCP服务。在一些实施例中，这批地址4355不包括在分配给在主机4310中运行的本地DHCP守护进程的那批地址4335中的地址。当本地DHCP守护进程4350在该批地址4355中用完可指派的可用地址时，本地DHCP守护进程4350也求助于中央DHCP守护进程4320。在一些实施例中，本地DHCP守护进程4350经由受管理交换元件4360和4306与中央DHCP守护进程4320通信。在一些实施例中受管理交换元件4360和4306具有在它们之间建立的隧道。

[0482] VI.插入服务(interposing service)VM

[0483] 在以上论述中，描述了一些实施例的虚拟化应用提供的各种L3服务。为了使网络控制系统的灵活性最大化，一些实施例插入服务机器，这些服务机器提供与用户现今在物理网络中使用的“中间盒”(middlebox)提供的功能类似的功能。

[0484] 从而，一些实施例的网络控制系统包括至少一个附接到逻辑网络的LDPS的“中间盒”VM。然后，LDP集合的管道状态被(填充逻辑控制平面的)控制应用编程，以使得相关分组被转发到此VM的逻辑端口。在VM处理了该分组之后，分组被发送回逻辑网络，使得其转发通过逻辑网络继续。在一些实施例中，网络控制系统利用许多这种“中间盒”VM。以这种方式插入的中间盒VM可以是非常有状态的，并且实现远超本文档中描述的L3服务的特征。

[0485] VII.可扩展性(scalability)

[0486] 下面讨论一些实施例的逻辑L3交换设计沿着三个维度的可扩展性含义。这三个维度是：(1)逻辑状态，(2)物理隧道状态，以及(3)分布式绑定(binding)查找。逻辑管道处理的大部分发生在第一跳。这意味着所有互连的LDP集合的所有逻辑(表)状态在一些实施例中被散布(disseminate)到网络中可发生管道执行的每个地方。换言之，所有互连的LDP集合的组合逻辑状态在一些实施例中被散布到附接于这些LDP集合中的任何一个的每个受管理边缘交换元件。然而，在一些实施例中，逻辑拓扑的“网格性(meshiness)”不增大逻辑状态的散布负担。

[0487] 为了限制状态散布，一些实施例在源和目的地设备之间均衡管道执行，以使得最末的LDPS管道不是在第一跳执行，而是在最末跳执行。然而，在一些情况下，这可导致没有

散布足够的状态来供每个受管理交换元件进行最末LDPS的逻辑转发决策;没有该状态,源受管理交换元件可能甚至不能够将分组递送到目的地受管理交换元件。从而,一些实施例将约束一般的LDPS模型,以便在源和目的地设备之间均衡管道执行。

[0488] 逻辑状态本身不太可能包含多于最多O(N)个条目(N是互连的LDP集合中的逻辑端口的总数),因为逻辑控制平面在一些实施例中被设计为模仿现今使用的物理控制平面,而物理控制平面受到现有的硬件交换芯片集的能力的限制。因此,散布逻辑状态可能不是系统的主要瓶颈,但随着逻辑控制平面设计的发展,最终其可能会变成一个瓶颈。

[0489] 一些实施例将网络的受管理交换元件划分成由更高级的聚集(aggregation)交换元件互连的团(clique)。代替实现划分以利用“全都在第一跳上(everything on the first-hop)”模型来减少逻辑状态,一些实施例进行划分以减少隧道状态,如下文所讨论的那样。团的示例在上述美国专利申请13/177,535中描述。此申请还描述了在第一跳受管理交换元件处执行逻辑数据处理的全部或大部分的各种实施例。

[0490] 整个系统中维护的物理隧道状态是$O(N^2)$,其中N是互连的LDP集合中的逻辑端口的总数。这是因为任何具有逻辑端口的受管理边缘交换元件都必须能够直接发送流量到目的地受管理边缘交换元件。因此,以有效的方式维护隧道状态而不对任何集中式控制元件施加$O(N^2)$负载变得比纯L2 LDP集合更重要。聚集交换元件在一些实施例中用于将网络切(slice)成团。在这些实施例中的一些中,分组仍然一直在源受管理边缘交换元件中被逻辑地路由,但是代替直接地将其隧道到目的地边缘交换元件,它被发送到基于目的地MAC地址将其朝着目的地路由的池节点。实质上,最末的L2 LDPS横跨多个团,并且池节点用于将该L2域的各部分联接在一起。

[0491] 图44-45B示出了基于受管理交换元件的流条目在若干受管理交换元件中实现的分布式逻辑路由器。特别地,图44-45B示出了目的地L2处理中的一些由最末跳受管理交换元件(即,将分组直接发送到目的地机器的交换元件)执行。

[0492] 图44概念性示出了在最末跳交换元件处执行一些逻辑处理的示例。具体而言,图44示出了耦合到分组的源机器的受管理交换元件2505执行逻辑处理管道200的大部分,并且耦合到目的地机器的受管理交换元件2510执行逻辑处理管道200的一些。该图在其左半部示出了逻辑路由器225以及逻辑交换机220和230。该图在其右半部示出了受管理交换元件2505和2510。该图在其右半部和左半部都示出了VM 1-4。

[0493] 在一些实施例中,受管理交换元件不保持所有信息(例如,查找表中的流条目)来执行整个逻辑处理管道200。例如,这些实施例的受管理交换元件不维护用于相对于目的地逻辑网络的通过其将分组发送到分组的目的地机器的逻辑端口确定访问控制的信息。

[0494] 现在将描述沿着受管理交换元件2505和2510的示例分组流。当耦合到逻辑交换机220的VM 1向耦合到逻辑交换机230的VM 4发送分组时,该分组首先被发送到受管理交换元件2505。受管理交换元件2505随后对分组执行L2处理205和L3处理210。

[0495] 受管理交换元件2505随后执行L2处理215的一部分。具体而言,受管理交换元件2505为分组确定访问控制。例如,受管理交换元件2505确定分组不具有将使得逻辑交换机230拒绝经过逻辑交换机230的端口Y的分组的网络地址(例如,源/目的地MAC/IP地址,等等)。受管理交换元件2505随后确定逻辑交换机230的端口1是通过其将分组送出到目的地VM 4的端口。然而,受管理交换元件2505不相对于逻辑交换机230的端口1为分组确定访问

控制,因为受管理交换元件2505在一些实施例中不具有用来执行出口ACL 2670的信息(例如流条目)。

[0496] 受管理交换元件2505随后执行映射查找以确定逻辑交换机230的逻辑端口1被映射到的物理端口。在此示例中,受管理交换元件2505确定逻辑交换机230的逻辑端口1被映射到受管理交换元件2510的端口5。受管理交换元件2505随后执行物理查找以确定用于将分组转发到物理端口的操作。在此示例中,受管理交换元件2505确定分组应当通过受管理交换元件2505的端口5被发送到VM 4。受管理交换元件2505在此示例中在将逻辑上下文与分组一起送出到VM 4之前修改分组的逻辑上下文。

[0497] 受管理交换元件2505将分组发送到受管理交换元件2510。在一些情况下,受管理交换元件2505通过在受管理交换元件2505和2510之间建立的隧道(例如,端接于受管理交换元件2505的端口3和受管理交换元件2510的端口3的隧道)发送分组。当隧道不可用时,受管理交换元件2505将分组发送到池节点(未示出),以便分组可到达受管理交换元件2510。

[0498] 当受管理交换元件2510接收到分组时,受管理交换元件2510基于分组的逻辑上下文(逻辑上下文将会指示剩下来要对分组执行的是出口ACL 2670)对分组执行出口ACL 2670。例如,受管理交换元件2510确定分组不具有将使得逻辑交换机230不通过逻辑交换机230的端口1发送分组的网络地址。受管理交换元件2510随后如执行L2处理215的受管理交换元件2505所确定的那样通过受管理交换元件2510的端口5将分组发送到VM 4。

[0499] 图45A-45B概念性示出了上文参考图44描述的逻辑交换机220和230、逻辑路由器225以及受管理交换元件2505和2510的示例操作。具体而言,图45A示出了实现逻辑路由器225、逻辑交换机220和逻辑交换机230的一部分的受管理交换元件2505的操作。图45B示出了实现逻辑交换机230的一部分的受管理交换元件2510的操作。

[0500] 如图45A的下半部所示,受管理交换元件2505包括L2条目4505和4515以及L3条目4510。这些条目是控制器集群(未示出)提供给受管理交换元件2505的流条目。虽然这些条目被描绘为三个分开的表,但这些表不一定必须是分开的表。也就是说,单个表可包括所有这些流条目。

[0501] 当耦合到逻辑交换机220的VM 1向耦合到逻辑交换机230的VM 4发送分组4530时,该分组首先通过受管理交换元件2505的端口4被发送到受管理交换元件2505。受管理交换元件2505基于受管理交换元件2505的转发表4505-4515对分组执行L2处理。在此示例中,分组4530具有目的地IP地址1.1.2.10,这是VM 4的IP地址。分组4530的源IP地址是1.1.1.10。分组4530还以VM 1的MAC地址作为源MAC地址并且以逻辑路由器225的逻辑端口1的MAC地址(例如,01:01:01:01:01:01)作为目的地MAC地址。

[0502] 受管理交换元件2505的直到受管理交换元件识别带圈的9并执行L2逻辑处理2665为止的操作与图30A的示例中的受管理交换元件2505的操作类似,除了图45A的示例中的受管理交换元件2505是对分组4530执行的。

[0503] 基于分组4530的头中存储的逻辑上下文和/或其它字段,受管理交换元件2505随后识别L2条目4515中的实现阶段2675的上下文映射的由带圈的10指示的记录(称为“记录10”)。在此示例中,记录10将与VM 4耦合的受管理交换元件2510的端口5识别为与分组4530要被转发到的逻辑交换机230的逻辑端口(在阶段2665确定)相对应的端口。记录10附加地指明通过转发表来进一步处理分组4530(例如通过将分组4530发送到调度端口)。

[0504] 基于分组4530的头中存储的逻辑上下文和/或其它字段，受管理交换元件2505随后识别L2条目4515中的实现阶段2680的物理映射的由带圈的11指示的记录(称为“记录11”)。记录11指明受管理交换元件2505的端口3作为通过其发送分组4530的端口以便分组4530到达受管理交换元件2510。在此情况下，受管理交换元件2505将把分组4530从与受管理交换元件2510耦合的受管理交换元件2505的端口3发送出去。

[0505] 如图45B所示，受管理交换元件2510包括转发表，该转发表包括用于处理和路由分组4530的规则(例如流条目)。当受管理交换元件2510从受管理交换元件2505接收到分组4530时，受管理交换元件2510基于受管理交换元件2510的转发表开始处理分组4530。受管理交换元件2510识别转发表中的实现上下文映射的由带圈的1指示的记录(称为“记录1”)。记录1基于分组4530的头中存储的逻辑上下文来识别分组4530的逻辑上下文。逻辑上下文指明分组4530已被受管理交换元件2505处理到阶段2665。这样，记录1指明通过转发表来进一步处理分组4530(例如通过将分组4530发送到调度端口)。

[0506] 接下来，受管理交换元件2510基于分组4530的头中存储的逻辑上下文和/或其它字段来识别转发表中的实现出口ACL的由带圈的2指示的记录(称为“记录2”)。在此示例中，记录2允许分组4530被进一步处理，并且从而指明通过转发表来进一步处理分组4530(例如，通过将分组4530发送到调度端口)。此外，记录2指明受管理交换元件2510将分组4530的逻辑上下文(即，对于逻辑交换机230的L2出口ACL，分组4530已被处理)存储在分组4530的头的该组字段中。

[0507] 接下来，受管理交换元件2510基于分组4530的头中存储的逻辑上下文和/或其它字段来识别转发表中的实现物理映射的由带圈的3指示的记录(称为“记录3”)。记录3指明为了分组4530到达VM 4要通过其发送分组4530的受管理交换元件2510的端口5。在此情况下，受管理交换元件2510将把分组4530从与VM 4耦合的受管理交换元件2510的端口5发送出去。在一些实施例中，受管理交换元件2510在将分组发送到VM 4之前从分组4530去除逻辑上下文。

[0508] 图46-47B示出了基于受管理交换元件的流条目在若干受管理交换元件中实现的分布式逻辑路由器。特别地，图46-47B示出了源L2处理205和L3处理210由第一跳受管理交换元件(即，直接从源机器接收分组的交换元件)执行，并且整个目的地L2处理215由最末跳受管理交换元件(即，直接向目的地机器发送分组的交换元件)执行。

[0509] 图46概念性示出了在最末跳交换元件处执行一些逻辑处理的示例。图46示出了耦合到分组的源机器的受管理交换元件2505执行L2处理205和L3处理210，并且耦合到目的地机器的受管理交换元件2510执行L2处理215。也就是说，受管理交换元件2505执行针对源逻辑网络的L2转发并且执行L3路由，而针对目的地逻辑网络的L2转发由受管理交换元件2510执行。该图在其左半部示出了逻辑路由器225以及逻辑交换机220和230。该图在其右半部示出了受管理交换元件2505和2510。该图在其右半部和左半部都示出了VM 1-4。

[0510] 在一些实施例中，受管理交换元件不保持所有的信息(例如，查找表中的流条目)来执行整个逻辑处理管道200。例如，这些实施例的受管理交换元件不维护用于对分组执行针对目的地逻辑网络的逻辑转发的信息。

[0511] 现在将描述沿着受管理交换元件2505和2510的示例分组流。当耦合到逻辑交换机220的VM 1向耦合到逻辑交换机230的VM 4发送分组时，该分组首先被发送到受管理交换元

件2505。受管理交换元件2505随后对分组执行L2处理205和L3处理210。

[0512] 受管理交换元件2505将分组发送到受管理交换元件2510。在一些情况下，受管理交换元件2505通过在受管理交换元件2505和2510之间建立的隧道(例如端接于受管理交换元件2505的端口3和受管理交换元件2510的端口3的隧道)发送分组。当隧道不可用时，受管理交换元件2505将分组发送到池节点(未示出)，以便分组可到达受管理交换元件2510。

[0513] 当受管理交换元件2510接收到分组时，受管理交换元件2510基于分组的逻辑上下文(逻辑上下文将指示剩下来要对分组执行的是整个L2处理215)对分组执行L2处理215。受管理交换元件2510随后通过受管理交换元件2510的端口5将分组发送到VM 4。

[0514] 图47A-47B概念性示出了上文参考图46描述的逻辑交换机220和230、逻辑路由器225以及受管理交换元件2505和2510的示例操作。具体而言，图47A示出了实现逻辑交换机220和逻辑路由器225的受管理交换元件2505的操作。图47B示出了实现逻辑交换机230的受管理交换元件2505的操作。

[0515] 如图47A的下半部所示，受管理交换元件2505包括L2条目4705和L3条目4710。这些条目是控制器集群(未示出)提供给受管理交换元件2505的流条目。虽然这些条目被描绘为三个分开的表，但这些表不一定必须是分开的表。也就是说，单个表可包括所有这些流条目。

[0516] 当耦合到逻辑交换机220的VM 1向耦合到逻辑交换机230的VM 4发送分组4730时，该分组首先通过受管理交换元件2505的端口4被发送到受管理交换元件2505。受管理交换元件2505基于受管理交换元件2505的转发表4705-4710对分组执行L2处理。在此示例中，分组4730具有目的地IP地址1.1.2.10，这是VM 4的IP地址。分组4730的源IP地址是1.1.1.10。分组4730还以VM 1的MAC地址作为源MAC地址并且以逻辑路由器225的逻辑端口1的MAC地址(例如，01:01:01:01:01:01)作为目的地MAC地址。

[0517] 受管理交换元件2505的直到受管理交换元件识别带圈的7并执行相对于逻辑路由器225的端口2的L3出口ACL为止的操作与图47A的示例中的受管理交换元件2505的操作类似，除了图47A的示例中的受管理交换元件2505是对分组4730执行的。

[0518] 基于分组4730的头中存储的逻辑上下文和/或其它字段，受管理交换元件2505随后识别L2条目4710中的实现阶段2680的物理映射的由带圈的8指示的记录(称为“记录8”)。记录8指明逻辑交换机230在受管理交换元件2510中实现并且分组应当被发送到受管理交换元件2510。

[0519] 基于分组4730的头中存储的逻辑上下文和/或其它字段，受管理交换元件2505随后识别L3条目4710中的实现阶段2680的物理映射的由带圈的9指示的记录(称为“记录9”)。记录9指明受管理交换元件2505的端口3作为要通过其发送分组4730的端口以便分组4730到达受管理交换元件2510。在此情况下，受管理交换元件2505将把分组4730从与受管理交换元件2510耦合的受管理交换元件2505的端口3发送出去。

[0520] 如图47B所示，受管理交换元件2510包括转发表，该转发表包括用于处理和路由分组4730的规则(例如流条目)。当受管理交换元件2510从受管理交换元件2510接收到分组4730时，受管理交换元件2510基于受管理交换元件2510的转发表开始处理分组4730。受管理交换元件2510识别转发表中的实现上下文映射的由带圈的1指示的记录(称为“记录1”)。记录1基于分组4730的头中存储的逻辑上下文来识别分组4730的逻辑上下文。逻辑上下文

指明受管理交换元件810已对分组4730执行了L2处理205和L3处理210。记录1指明通过转发表来进一步处理分组4730(例如通过将分组4730发送到调度端口)。

[0521] 基于分组4730的头中存储的逻辑上下文和/或其它字段,受管理交换元件2510识别L2转发表中的实现L2入口ACL的由带圈的2指示的记录(称为“记录2”)。在此示例中,记录2允许分组4730通过逻辑交换机230(未示出)的逻辑端口Y,并且从而指明由受管理交换元件2510进一步处理分组4730(例如通过将分组4730发送到调度端口)。此外,记录2指明受管理交换元件2510将分组4730的逻辑上下文(即,分组4730已被处理管道4700的阶段4762处理)存储在分组4730的头的该组字段中。

[0522] 接下来,受管理交换元件2510基于分组4730的头中存储的逻辑上下文和/或其它字段识别L2转发表中的实现逻辑L2转发的由带圈的3指示的记录(称为“记录3”)。记录3指明以VM 4的MAC地址作为目的地MAC地址的分组应当通过与VM 4连接的逻辑交换机230的逻辑端口2转发。

[0523] 记录3还指明通过转发表来进一步处理分组4730(例如,通过将分组4730发送到调度端口)。此外,记录3指明受管理交换元件2510将逻辑上下文(即,分组4730已被处理管道4700的阶段4766处理)存储在分组的该组字段中。

[0524] 接下来,受管理交换元件2510基于分组4730的头中存储的逻辑上下文和/或其它字段识别转发表中的实现出口ACL的由带圈的4指示的记录(称为“记录4”)。在此示例中,记录4允许分组4730被进一步处理,并且从而指明通过转发表来进一步处理分组4730(例如,通过将分组4730发送到调度端口)。此外,记录4指明受管理交换元件2510将分组4730的逻辑上下文(即,对于逻辑交换机230的L2出口ACL,分组4730已被处理)存储在分组4730的头的该组字段中。

[0525] 基于分组4730的头中存储的逻辑上下文和/或其它字段,受管理交换元件2505随后识别L2转发表4715中的实现上下文映射的由带圈的5指示的记录(称为“记录5”)。在此示例中,记录5将与VM4耦合的受管理交换元件2510的端口5识别为与分组4730要被转发到的逻辑交换机230的逻辑端口2相对应的端口。记录5附加地指明通过转发表来进一步处理分组4730(例如通过将分组1330发送到调度端口)。

[0526] 接下来,受管理交换元件2510基于分组4730的头中存储的逻辑上下文和/或其它字段来识别转发表中的实现物理映射的由带圈的6指示的记录(称为“记录6”)。记录6指明为了分组4730到达VM 4要通过其发送分组4730的受管理交换元件2510的端口5。在此情况下,受管理交换元件2510将把分组4730从与VM 4耦合的受管理交换元件2510的端口5发送出去。在一些实施例中,受管理交换元件2510在将分组发送到VM 4之前从分组4730去除逻辑上下文。

[0527] 在分组的逻辑路径上对所有管道的执行提示了分布式查找,即ARP和学习。由于查找现在可由任何具有附接到逻辑网络的逻辑端口的边缘交换元件执行,所以查找的总量将会超过在类似的物理拓扑上执行的查找;即使分组将前往同一端口,不同的发送者也不能共享缓存的查找状态,因为查找将在不同的受管理边缘交换元件上发起。因此,洪泛的问题被逻辑拓扑放大了,并且基于单播映射的查找方案在实践中是优选的。

[0528] 通过向一群映射服务器(例如池或根节点发送特殊查找分组,源边缘交换元件可以进行必要的查找,而不求助于洪泛。在一些实施例中,映射服务器受益于大流量聚集本地

性(并因此受益于客户端侧的良好的缓存命中率)以及仅限数据路径的实现,这种实现导致了良好的吞吐量。

[0529] 图48概念性示出了受管理交换元件在其上运行的主机4800的示例软件体系结构。具体而言,此图示出了主机4800还运行L3守护进程,该L3守护进程为L3守护进程从受管理交换元件接收的分组将L3地址(例如IP地址)解析成L2地址(例如MAC地址)。此图在其上半部示出了主机4800包括受管理交换元件4805、转发表4820、L3守护进程4810和映射表4815。此图还示出了流条目4825和4830。

[0530] 流条目4825和4830各自具有限定符和动作。示为流条目4825和4830的文本可能不是实际的格式。相反,文本只是限定符和动作对的概念图示。在一些实施例中,流条目具有优先级,并且当多于一个流条目的限定符被满足时,受管理交换元件采取具有最高优先级的流条目的动作。

[0531] 主机4800在一些实施例中是由能够运行一组软件应用的操作系统(例如,Windows™和Linux™)操作的机器。一些实施例的受管理交换元件4805是在主机4800中执行的软件交换元件(例如Open vSwitch)。如上所述,控制器集群(未示出)通过提供指明受管理交换元件的功能的流条目来配置受管理交换元件。一些实施例的受管理交换元件4805自身不生成流条目和ARP请求。

[0532] 一些实施例的受管理交换元件4805运行上述逻辑处理管道200的全部或一部分。特别地,受管理交换元件4805是执行L3处理210以根据需要基于转发表4820中的流条目对从机器接收的分组进行路由的受管理交换元件(例如,受管理交换元件1720或2505)。在一些实施例中,受管理交换元件4805是从耦合到受管理交换元件的机器(未示出)接收分组的边缘交换元件。在一些这样的实施例中,一个或多个虚拟机(未示出)在主机4800中运行并且耦合到受管理交换元件4805。在其它实施例中,受管理交换元件是第二级受管理交换元件。

[0533] 当受管理交换元件4805接收到正是被发送到在另一逻辑网络中的目的地机器的第一个分组的分组(或者该分组本身是ARP请求)时,这些实施例的受管理交换元件4805还不知道目的地机器的MAC地址。换言之,受管理交换元件4805不知道下一跳IP地址与目的地MAC地址之间的映射。为了将下一跳IP地址解析成目的地MAC地址,一些实施例的受管理交换元件4805从L3守护进程4810请求分组的目的地MAC地址。

[0534] 一些实施例的L3守护进程4810是在主机4800上运行的软件应用。L3守护进程4810维护表4815,该表4815包括IP和MAC地址的映射。当受管理交换元件4805寻求与下一跳IP地址相对应的目的地MAC地址时,L3守护进程查找映射表4815以找出源IP地址被映射到的目的地MAC地址。(在一些情况下,源IP地址被映射到的目的地MAC地址是下一跳逻辑路由器的MAC地址)。

[0535] 不同实施例的受管理交换元件4805和L3守护进程4810使用不同的技术来寻求和提供地址。例如,一些实施例的受管理交换元件4805向L3守护进程发送分组,该分组具有目的地IP地址,但不具有目的地MAC地址。这些实施例的L3守护进程4810将IP地址解析成目的地MAC地址。L3守护进程4810将分组发送回受管理交换元件4805,受管理交换元件4805将执行逻辑转发和/或路由以向目的地机器发送分组。在一些实施例中,受管理交换元件4805最初将元数据与包含要解析的目的地IP地址的分组一起发送给L3守护进程4810。此元数据包

括受管理交换元件4805在其接收到从L3守护进程4810返回的分组时用来继续执行逻辑处理管道的信息(例如,寄存器值、逻辑管道状态,等等)。

[0536] 在其它实施例中,受管理交换元件4805通过向L3守护进程4810发送流模板来请求目的地地址,该流模板是不具有目的地MAC地址的实际值的流条目。L3守护进程通过查找映射表4815来找出目的地MAC地址以填写流模板。L3守护进程4810随后通过将已填写的流模板放入转发表4820中来将填入了实际目的地MAC地址的流模板发送回到受管理交换元件4810。在一些实施例中,L3守护进程向已填写的流模板指派比未填写的流模板的优先级值高的优先级值。

[0537] 当映射表4815具有针对目的地IP地址的条目并且该条目具有映射到目的地IP地址的目的地MAC地址时,L3守护进程4810使用该目的地MAC地址来写入分组中或填入流模板中。当没有这样的条目时,L3守护进程生成ARP请求并将该ARP分组广播到其它运行L3守护进程的主机或VM。特别地,一些实施例的L3守护进程仅将ARP请求发送到下一跳逻辑L3路由器可附接到的主机或VM。L3守护进程从接收到ARP分组的主机或VM中的一个接收对ARP分组的响应,该响应包含目的地MAC地址。L3守护进程4810将目的地IP地址映射到目的地MAC地址并将此映射添加到映射表4815。在一些实施例中,L3守护进程4810周期性地向响应了ARP请求的另一L3守护进程发送单播分组以检查目的地MAC地址的有效性。以这种方式,L3守护进程4810将IP和MAC地址映射保持为最新。

[0538] 在一些实施例中,当L3守护进程4810在查找了流条目并向其它L3守护进程实例发送ARP请求之后仍未能找到解析出的地址时,L3守护进程将在流模板中指明丢弃分组,或者L3守护进程自身将丢弃分组。

[0539] 当受管理交换元件4805接收到来自另一主机或VM的ARP分组时,一些实施例的受管理交换元件4805不将该ARP分组转发到与该受管理交换元件耦合的机器。这些实施例中的受管理交换元件4800将ARP分组发送到L3守护进程。L3守护进程在映射表4815中维护在本地可用的IP地址与MAC地址(例如,耦合到受管理交换元件4805的机器的IP地址和MAC地址)之间的映射。当映射表4815具有针对接收到的ARP分组的IP地址的条目并且该条目具有耦合到受管理交换元件4805的VM的MAC地址时,L3守护进程响应于ARP分组将MAC地址发送给该ARP分组所源自的主机或VM(即,该主机或VM的L3守护进程)。

[0540] 现在将按照三个不同的阶段1-3(带圈的1-3)来描述受管理交换元件4805和L3守护进程4810的示例操作。在此示例中,受管理交换元件4805是从机器(未示出)接收要转发和路由的分组的受管理边缘交换元件。受管理交换元件4805接收分组并基于转发表4820中的流条目来执行逻辑处理200。

[0541] 当分组恰好是携带目的地机器的IP地址的第一个分组或者分组是来自源机器的ARP请求时,受管理交换元件4820(在阶段1)识别流条目4825并且执行流条目4825中指明的动作。如图所示,流条目4825指示具有要被解析成目的地MAC X的目的地IP地址1.1.2.10的流模板应当被发送到L3守护进程4810。在此示例中,流条目4825具有优先级值N,其在一些实施例中为一数字。

[0542] 在阶段2,L3守护进程4810接收流模板并通过查找映射表4815查明1.1.2.10要被解析成01:01:01:01:01:09。L3守护进程填写流模板并将已填写的模板(现在是流条目4830)插入转发表4830中。在此示例中,L3守护进程向已填写的模板指派优先级N+1。

[0543] 在阶段3,受管理交换元件4810在一些实施例中使用流条目4830来为分组设定目的地MAC地址。另外,对于受管理交换元件4810随后处理的分组,当分组具有目的地IP地址1.1.2.10时,受管理交换元件4805使用流条目4830而不是流条目4825。

[0544] 在一些实施例中,L3守护进程4810和受管理交换元件运行于在主机4800上运行的同一虚拟机中或者在主机4800上运行的不同虚拟机中。在一些实施例中,L3守护进程4810在虚拟机的用户空间中运行。L3守护进程4810和受管理交换元件也可在分开的主机中运行。

[0545] 在一些实施例中,受管理交换元件4805不依赖于L3守护进程4810来解析地址。在一些这样的实施例中,控制集群(图48中未示出)可静态地配置流条目4820,以使得流条目4820包括通过API调用(即,输入)或DHCP获得的IP地址到MAC地址之间的映射。

[0546] 图49概念性示出了一些实施例执行来解析网络地址的过程4900。在一些实施例中,过程4900由执行L3处理210以在L3路由分组的受管理交换元件(例如受管理交换元件1720、2505或3105)执行。过程4900在一些实施例中在该过程接收到要在L3逻辑路由的分组时开始。

[0547] 过程4900开始于(在4905)确定分组是否需要地址解析(例如,将目的地IP地址解析成目的地MAC地址)。在一些实施例中,过程基于流条目来确定分组是否需要L3处理。其限定符与分组的头或逻辑上下文中存储的信息匹配的流条目指明分组需要地址解析。

[0548] 当过程4900(在4905)确定分组不需要地址解析时,过程结束。否则,过程4900(在4910)确定过程4900是否需要从L3守护进程请求将分组的地址(例如目的地IP地址)解析成的地址。在一些实施例中,过程4900基于流条目来确定过程是否需要询问L3守护进程。例如,流条目可指明将分组的地址解析成的地址应当通过从L3守护进程请求解析出的地址来获得。在一些实施例中,当流条目是对于解析出的地址具有空字段或者在该字段中具有指示应当从L3守护进程获得解析出的地址的某个其它值的流模板时,过程确定L3守护进程应当提供解析出的地址。

[0549] 当过程(在4910)确定过程不需要从L3守护进程请求地址时,过程(在4920)从流条目获得解析出的地址。例如,流条目将提供转换后的地址。过程随后前进到4925,下文将进一步描述4925。当过程(在4910)确定过程需要从L3守护进程请求地址时,过程4900在4915从L3守护进程请求并获得解析出的地址。在一些实施例中,过程4900通过向L3守护进程发送流模板来请求解析出的地址。L3守护进程将利用解析出的地址填写流模板并且将把已填写的流模板放入过程使用的转发表(未示出)中。

[0550] 接下来,过程4900利用解析出的地址来修改分组。在一些实施例中,过程修改分组的头中的地址字段。可替代地或连带地,过程修改逻辑上下文利用解析出的地址来替换分组的地址。过程随后结束。

[0551] 图50示出了一些实施例的网络体系结构5000。具体而言,此图示出了允许了各自运行L3守护进程的若干主机(或VM)以避免广播ARP请求的映射服务器。此图示出了一组主机(或VM),包括5005、5010和5015。

[0552] 主机5010和5015与上文参考图48描述的主机4805的相似之处在于主机5010和5015运行L3守护进程、受管理交换元件和一个或多个VM。

[0553] 主机5005运行映射服务器。一些实施例的映射服务器5005维护全局映射表5020,

该全局映射表5020包括在网络中的每一个运行受管理边缘交换元件的主机中运行的L3守护进程所维护的所有映射表的所有条目。在一些实施例中，网络中的L3守护进程发送本地可用的IP地址与MAC地址映射之间的映射的条目。每当耦合到主机的受管理交换元件的机器存在变化时(例如，当VM发生故障或者耦合到受管理交换元件或与受管理交换元件解除耦合时)，主机的L3守护进程就相应更新各自的本地映射表并且还向映射服务器5005发送更新(例如通过发送包含更新的特殊“公布(publish)”分组)，以便映射服务器5005保持全局映射表5005随着变化而更新。

[0554] 在一些实施例中，在运行受管理边缘交换元件的每个主机中运行的L3守护进程在本地映射不具有针对要解析的目的地IP地址的条目时不广播ARP分组。反而，L3守护进程咨询映射服务器5005来将目的地IP地址解析成目的地MAC地址。映射服务器5005通过查找全局映射表5020来将目的地IP地址解析成目的地MAC地址。在映射服务器5005不能解析IP地址的情况下(例如，当全局映射表5020不具有针对该IP地址的条目或者映射服务器5005发生故障时)，L3守护进程将采取广播ARP分组到其它运行受管理边缘交换元件的主机。在一些实施例中，映射服务器5005在实现第二级受管理交换元件(例如池节点)的同一主机或VM中实现。

[0555] 图51示出了一些实施例执行来维护包括IP和MAC地址的映射的映射表的过程5100。在一些实施例中，过程5100由向映射服务器请求解析出的地址的L3守护进程执行。映射服务器在这些实施例中为一组受管理交换元件维护包括IP和MAC地址的映射的全局映射表。过程5100在一些实施例中在过程从受管理交换元件接收到要解析的特定地址时开始。

[0556] 过程开始于(在5105)确定过程对于从受管理交换元件接收的特定地址是否具有解析出的地址。在一些实施例中，过程查找包括IP和MAC地址的映射的本地映射表来确定过程对于特定地址是否具有解析出的地址。

[0557] 当过程5100确定过程具有解析出的地址时，过程前进到5120，下文将进一步描述5120。否则，过程5100从映射服务器请求并获得解析出的地址。过程5100随后(在5115)利用从映射服务器获得的解析出的地址来修改本地映射表。在一些实施例中，过程5100将解析出的地址和该特定地址的新映射插入到本地映射表中。

[0558] 过程5100随后将解析出的地址发送到受管理交换元件。在一些实施例中，过程5100修改具有该特定地址的分组。在其它实施例中，过程5100修改受管理交换元件作为对解析出的地址的请求已发送的流模板。过程随后结束。

[0559] 图52示出了一些实施例执行来维护包括IP和MAC地址的映射的映射表的过程5200。在一些实施例中，过程5200由维护本地映射表并向映射服务器发送更新的L3守护进程执行。映射服务器在这些实施例中为一组受管理交换元件维护包括IP和MAC地址的映射的全局映射表。过程5200在一些实施例中在L3守护进程开始运行时开始。

[0560] 过程5200开始于(在5205)监视一组受管理交换元件。特别地，过程5200监视机器与受管理交换元件的耦合和解除耦合或者耦合到受管理交换元件的机器的任何地址变化。在一些实施例中，该组受管理交换元件包括在L3守护进程在其上运行的同一主机或虚拟机上运行的那些受管理交换元件。

[0561] 接下来，过程5200(在5210)确定对于过程监视的受管理交换元件是否存在这样的变化。当过程(在5210)确定不存在变化时，过程5200循环回5205以继续保持监视该组受管

理交换元件。否则，过程(在5215)修改本地映射表中的相应条目。例如，当VM迁移并耦合到该组中的受管理交换元件中的一个时，过程将迁移的VM的IP地址和MAC地址的映射插入到本地映射表中。

[0562] 过程5200随后向映射服务器发送更新后的映射以便映射服务器可利用IP地址和MAC地址的新的和/或修改后的映射来更新全局映射表。过程随后结束。

[0563] VIII.流生成和流处理

[0564] 如上所述，一些实施例的受管理交换元件基于由一些实施例的控制器集群(一个或多个控制器实例)提供给受管理交换元件的流表来实现逻辑交换机和逻辑路由器。在一些实施例中，控制器集群基于控制器集群检测到的输入或网络事件通过执行表映射操作来生成这些流条目。这些控制器集群及其操作的细节在美国专利申请13/177,533以及以上并入的标题为“Chassis Controller”且代理人案卷号为NCRA.P0081的同时递交的美国专利申请＊＊中描述。

[0565] 如这个同时递交的美国专利申请中提到的，一些实施例中的网络控制系统是包括若干控制器实例的分布式控制系统，这些控制器实例允许系统接受来自用户的逻辑数据路径集合并且配置交换元件来实现这些逻辑数据路径集合。在一些实施例中，一种类型的控制器实例是执行一个或多个模块的设备(例如通用计算机)，这些模块将用户输入从逻辑控制平面转换到逻辑转发平面，然后将逻辑转发平面数据变换到物理控制平面数据。这些模块在一些实施例中包括控制模块和虚拟化模块。控制模块允许用户指明并填充逻辑数据路径集合，而虚拟化模块通过将逻辑数据路径集合映射到物理交换基础设施上来实现指明的逻辑数据路径集合。在一些实施例中，控制和虚拟化模块是两个分开的应用，而在其它实施例中它们是同一应用的一部分。

[0566] 从对于特定逻辑数据路径集合的逻辑转发平面数据，一些实施例的虚拟化模块生成对于实现逻辑数据路径集合的任何受管理交换元件通用的通用物理控制平面(UPCP)数据。在一些实施例中，此虚拟化模块是作为该特定逻辑数据路径集合的主控制器的控制器实例的一部分。此控制器被称为逻辑控制器。

[0567] 在一些实施例中，UPCP数据随后被转换成针对每个特定受管理交换元件的定制物理控制平面(CPCP)数据，该转换由作为该特定受管理交换元件的主物理控制器实例的控制器实例进行，或者由该特定受管理交换元件的机箱控制器(Chassis Controller)进行，这在标题为“Chassis Controller”且代理人案卷号为NCRA.P0081的同时递交的美国专利申请＊＊中被进一步描述。当机箱控制器生成CPCP数据时，机箱控制器通过物理控制器从逻辑控制器的虚拟化模块获得UPCP数据。

[0568] 无论是物理控制器还是机箱控制器生成CPCP数据，针对特定受管理交换元件的CPCP数据都需要被传播到该受管理交换元件。在一些实施例中，通过网络信息库(NIB)数据结构来传播CPCP数据，网络信息库数据结构在一些实施例中是面向对象的数据结构。使用NIB数据结构的若干示例在美国专利申请13/177,529和13/177,533中描述，这里通过引用并入这些美国专利申请。如这些申请中所述，NIB数据结构在一些实施例中也可用来充当不同的控制器实例之间的通信媒介，并且存储关于逻辑数据路径集合(例如逻辑交换元件)和/或实现这些逻辑数据路径集合的受管理交换元件的数据。

[0569] 然而，其它实施例不使用NIB数据结构来将CPCP数据从物理控制器或机箱控制器

传播到受管理交换元件、在控制器实例之间通信以及存储关于逻辑数据路径集合和/或受管理交换元件的数据。例如，在一些实施例中，物理控制器和/或机箱控制器经由配置协议通过OpenFlow条目和更新来与受管理交换元件通信。另外，在一些实施例中，控制器实例使用一个或多个直接通信信道(例如RPC调用)来交换数据。此外，在一些实施例中，控制器实例(例如这些实例的控制和虚拟化模块)按照被写入到关系数据库数据结构中的记录来表达逻辑和/或物理数据。在一些实施例中，此关系数据库数据结构是用于实现控制器实例的一个或多个模块的表映射引擎(称为nLog)的输入和输出表的一部分。

[0570] 图53概念性示出了一些实施例的控制器集群的三个控制器实例。这三个控制器实例包括用于从作为API调用接收的逻辑控制平面(LCP)数据生成UPCP数据的逻辑控制器5300，以及分别用于定制受管理交换元件5320和5325特定的UPCP数据的物理控制器5390和5330。具体而言，一些实施例的逻辑控制器5300通过利用诸如nLog之类的表映射处理器(未示出)对表执行表映射操作来生成通用流。nLog引擎在美国专利申请13/177,533中描述。该图还示出了用户5325以及受管理交换元件5320和5325。

[0571] 如图所示，逻辑控制器5300包括控制应用5305和虚拟化应用5310。在一些实施例中，控制应用5305用于接收逻辑控制平面数据，并且将此数据转换成逻辑转发平面数据，该逻辑转发平面数据随后被提供给虚拟化应用5310。虚拟化应用5310从逻辑转发平面数据生成通用物理控制平面数据。

[0572] 在一些实施例中，逻辑控制平面数据中的一些是从输入转换来的。在一些实施例中，逻辑控制器5300支持一组API调用。逻辑控制器具有将该组API调用转换成LCP数据的输入转换应用(未示出)。利用API调用，用户可配置逻辑交换机和逻辑路由器，就好像用户在配置物理交换元件和路由器那样。

[0573] 物理控制器5390和5330分别是受管理交换元件5320和5325的主控(master)。一些实施例的物理控制器5390和5330从逻辑控制器5300接收UPCP数据并且将UPCP数据分别转换成用于受管理交换元件5320和5325的CPCP数据。物理控制器5390随后将用于受管理交换元件5320的CPCP数据发送到受管理交换元件5320。物理控制器5330将用于受管理交换元件5325的CPCP数据发送给受管理交换元件5325。用于受管理交换元件5320和5325的CPCP数据采用流条目的形式。受管理交换元件5320和5325随后基于流条目执行转发和路由分组。如美国专利申请13/177,533中所述，从LCP数据到LFP数据然后到CPCP数据的这个转换利用nLog引擎来执行。

[0574] 即使图53示出了两个物理控制器对于两个不同的受管理交换元件从UPCP数据生成CPCP数据，普通技术人员将会认识到，在其它实施例中，物理控制器起到简单地将UPCP数据中继到每个交换元件的机箱控制器的作用，机箱控制器进而生成该交换元件的CPCP数据并将此数据推送给其交换元件。

[0575] 图54示出了示例体系结构5400和用户界面5405。具体而言，此图示出了用户向控制器应用发送以便以期望的方式配置逻辑交换机和路由器。此图在其左半部示出了四个阶段5406-5409中的用户界面(UI)5405。此图还在其右半部示出了包括逻辑路由器5425以及两个逻辑交换机5420和5430的体系结构5400。

[0576] UI 5405是示例界面，通过该界面用户可输入一些输入并从控制器实例接收响应以便管理逻辑交换机和路由器。在一些实施例中，UI 5405是作为web应用提供的，并且从而

可利用web浏览器来打开。可替代地或连带地，一些实施例的控制应用可允许用户通过命令行界面输入并接收输入。

[0577] 该图的左半部示出了用户输入一些输入来设置控制器实例管理的网络的一组受管理交换元件要实现的逻辑交换机和逻辑路由器中的逻辑端口。特别地，用户通过(在阶段5406)提供端口的识别符“RP1”、与端口相关联的IP地址“1.1.1.253”以及网络掩码“255.255.255.0”来向逻辑路由器LR添加逻辑端口。用户还通过(在5407)提供端口识别符“SP1”，并且指明该端口要连接到逻辑路由器的逻辑端口RP1来向逻辑交换机LS1添加逻辑端口。用户还通过(在阶段5408)提供端口的识别符“RP2”、与端口相关联的IP地址“1.1.2.253”以及网络掩码“255.255.255.0”来向逻辑路由器LR添加另一逻辑端口。用户还通过(在5409)提供端口识别符“SP2”，并且指明该端口要连接到逻辑路由器的逻辑端口RP2来向逻辑交换机LS2添加另一逻辑端口。该图的右半部示出了添加到逻辑路由器和逻辑交换机的端口。

[0578] 图55-62概念性示出了控制应用5305的示例操作。这些图示出了一组表，控制应用5305使用并修改这组表以便生成要提供给受管理交换元件的流条目。具体而言，受管理交换元件(未示出)基于上文参考图54描述的输入来实现添加到逻辑交换机5420和5430以及逻辑路由器5400的逻辑端口。该图示出了控制应用5305、虚拟化应用5310和物理控制器5330。

[0579] 如图所示的控制应用5305包括输入转换5505、输入表5510、规则引擎5515、输出表5520、导出器(exporter)5525。

[0580] 输入转换5505在一些实施例中与管理工具交互，用户可利用该管理工具来查看和/或修改逻辑网络状态。不同的实施例向用户提供不同的管理工具。例如，输入转换5505在一些实施例中提供图形工具，例如上文参考图54描述的UI 5405。代替图形工具或结合图形工具，其它实施例可向用户提供命令行工具或任何其它类型的管理工具。输入转换5505通过管理工具从用户接收输入并处理接收到的输入以创建、填充和/或修改一个或多个输入表5510。

[0581] 输入表5510与美国专利申请13/288,908中描述的输入表类似，这里通过引用并入该美国专利申请。输入表在一些情况下表示用户管理的逻辑交换机和逻辑路由器的状态。例如，输入表5530是存储与逻辑交换机的逻辑端口相关联的无类域间路由(CIDR)格式的IP地址的表。控制应用利用控制应用通过管理工具接收到的输入或者控制应用检测到的任何网络事件来修改输入表。在控制应用5305修改输入表之后，控制应用5305使用规则引擎5515来处理经修改的输入表。

[0582] 不同实施例的规则引擎5515对不同组输入表执行数据库操作的不同组合以填充和/或修改不同组输出表5520。例如，当输入表5530被改变以指示创建了逻辑路由器的逻辑端口时，规则引擎5515修改表5535以将MAC地址关联到逻辑路由器的逻辑端口。输出表5560包括流条目，这些流条目指明实现逻辑交换机和逻辑路由器的受管理交换元件对被路由/转发的网络数据执行的动作。除了表5530-5560以外，规则引擎5515还可使用其它输入表、常数表和函数表来促成规则引擎5515的表映射操作。

[0583] 输出表也可用作规则引擎5515的输入表。也就是说，输出表中的变化可触发规则引擎5515要执行的另一表映射操作。因此，表5530-5560中的条目可产生于执行表映射操

作，并且也可为另一组表映射操作向规则引擎5515提供输入。这样，输入表和输出表在此图中在单个点线框中示出以指示这些表是输入和/或输出表。

[0584] 表5535用于存储逻辑路由器的逻辑端口和关联的MAC地址的配对。表5540是逻辑路由器在路由分组时使用的逻辑路由表。在一些实施例中，表5540将被发送到实现逻辑路由器的受管理交换元件。表5550用于为逻辑路由器的逻辑端口存储下一跳识别符和IP地址。表5555用于存储逻辑交换机的逻辑端口与逻辑路由器的逻辑端口之间的连接。导出器5525向虚拟化应用5310公布或发送输出表5520中的经修改的输出表。

[0585] 图55示出了上文参考图54描述的阶段5406之前的表5530-5560。表中的条目被描绘为点以指示在这些表中存在一些现有的条目。

[0586] 图56示出了在阶段5406之后的表5530-5560。也就是说，此图示出了在用户提供了逻辑端口的识别符“RP1”、与端口相关联的IP地址“1.1.1.253”和网络掩码“255.255.255.0”以向被识别为“LR”的逻辑路由器5425添加逻辑端口之后的表5530-5560。这里，表5530通过输入转换5505利用新条目来更新。新条目(或行)5601指示被识别为“RP1”的逻辑端口被添加并且与此端口相关联的IP地址由IP地址1.1.1.253、前缀长度24和网络掩码255.255.255.0来指明。

[0587] 规则引擎5515检测对表5530的这个更新并且执行一组表映射操作以更新表5535和5540。图57示出了这组表映射操作的结果。具体而言，此图示出了表5535具有新的行5701，其指示逻辑端口RP1现在与MAC地址01:01:01:01:01:01相关联。在利用其它表或功能(未示出)执行表映射操作时该MAC地址由规则引擎5515生成。

[0588] 图57还示出了表5540具有新的行5702，其是用于逻辑路由器5425的路由表中的条目。逻辑路由器5425(实现逻辑路由器5425的受管理交换元件)将查找此表5540以作出路由决策。行5702指明逻辑端口RP1的下一跳具有唯一识别符“NH1”。行5702还包括路由表中指派给这一行的优先级。此优先级用于确定当路由表中存在多个匹配行时应当使用哪个行来作出路由决策。在一些实施例中，一个条目中对于一行的优先级的值为前缀长度加上基本的优先级值“BP”。

[0589] 规则引擎5515检测对表5540的更新并且执行一组表映射操作来更新表5550。图58示出了这组表映射操作的结果。具体而言，此图示出了表5550具有新的行5801，其指示逻辑路由器5425的逻辑端口RP1的下一跳的IP地址是给定分组的目的地IP地址。(此行中的“0”意味着下一跳的IP是将通过逻辑路由器的RP1路由的给定分组的目的地。)

[0590] 图59示出了在上文参考图54描述的阶段5407之后的表5530-5560。也就是说，此图示出了在用户提供逻辑端口的识别符“SP1”以将该逻辑端口添加到逻辑交换机5420(LS1)并将此端口链接到逻辑路由器5425的逻辑端口RP1之后的表5530-5560。这里，表5555通过输入转换5505利用两个新的行来更新。新的行5901指示(逻辑交换机5420的)被识别为“SP1”的逻辑端口附接到(逻辑路由器5425的)逻辑端口RP1。另外，新的行5902指示逻辑端口RP1附接到逻辑端口SP1。此链接连接上文描述的逻辑处理管道200的L2处理和L3处理部分。

[0591] 规则引擎5515检测表5555的更新并且执行一组表映射操作来更新表5535。图60示出了这组表映射操作的结果。具体而言，此图示出了表5535具有新的行6001，其指示逻辑端口SP1现在与MAC地址01:01:01:01:01:01相关联，因为SP1和RP1现在被链接。

[0592] 规则引擎5515检测表5555的更新并且执行一组表映射操作来更新表5560。图61示出了这组表映射操作的结果。具体而言，此图示出了表5550具有四个新的行(流条目)6101-6104。行6101是指示其目的地MAC地址为01:01:01:01:01:01的分组要被发送到(逻辑交换机5420的)逻辑端口SP 1的流条目。行6102是指示被递送到逻辑端口SP1的任何分组要被发送到逻辑端口RP1的流条目。行6103是指示被递送到逻辑端口RP1的任何分组要被发送到逻辑端口SP1的流条目。行6104是指示具有落在由1.1.1.253/24指明的IP地址的范围内的IP地址的分组应当通过询问L3守护进程来请求MAC地址的流条目。

[0593] 图62示出了在上文描述的阶段5408和5409之后添加到一些表的新的行6201-6209。为了描述简单，省略了通过规则引擎5515的表更新的中间图示。

[0594] 新的行6201指示识别为“RP2”的逻辑端口被添加并且与此端口相关联的IP地址由IP地址1.1.2.253、前缀长度24和网络掩码255.255.255.0来指明。新的行6202指示逻辑端口RP2现在与MAC地址01:01:01:01:01:02相关联。新的行6203指示逻辑端口SP2与MAC地址01:01:01:01:01:02相关联。新的行6204是对于逻辑交换机5430的路由表中的条目。行6204指明逻辑端口RP2的下一跳具有唯一识别符“NH2”。行6204还包括路由表中指派给此行的优先级。

[0595] 新的行6205指示逻辑路由器5425的逻辑端口RP2的下一跳的IP地址是给定分组的目的地IP地址。新的行6206指示(逻辑交换机5430的)被识别为“SP2”的逻辑端口附接到(逻辑路由器5425的)逻辑端口RP2。另外，新的行6207指示逻辑端口RP2附接到逻辑端口SP2。

[0596] 行6208是指示其目的地MAC地址为01:01:01:01:01:02的分组要被发送到(逻辑交换机5430的)逻辑端口SP2的流条目。行6209是指示被递送到逻辑端口SP2的任何分组要被发送到逻辑端口RP2的流条目。行6210是指示被递送到逻辑端口RP2的任何分组要被发送到逻辑端口SP2的流条目。行6211是指示具有落在由1.1.2.253/24指明的IP地址的范围内的IP地址的分组应当通过询问L3守护进程来请求MAC地址的流条目。

[0597] 图62中所示的这些流条目是LFP数据。此LFP数据将被发送到虚拟化应用5310，虚拟化应用5310将从LFP数据生成UPCP数据。然后，UPCP数据将被发送到物理控制器5330，物理控制器5330将为受管理交换元件5325(图62中未示出)定制UPCP数据。最后，物理控制器5330将把CPCP数据发送到受管理交换元件5325。

[0598] 图63示出了在控制应用5305通过执行如上文参考图55-62所述的表映射操作来生成逻辑数据之后的体系结构5400。如图63所示，端口RP1和RP2分别与由1.1.1.253/24和1.1.2.253/24指明的IP地址的范围相关联。另外，端口SP1和SP2分别与MAC地址01:01:01:01:01:01和01:01:01:01:01:02相关联。此图还示出了耦合到逻辑交换机5420的VM 1和耦合到逻辑交换机5430的VM 2。

[0599] 现在将描述逻辑交换机5420和5430、逻辑路由器5425以及VM 1和2的示例操作。此示例假定实现逻辑路由器5425以及逻辑交换机5420和5430的一组受管理交换元件具有所有流条目6101-6104和6208-6211。此示例还假定由控制应用5305产生的逻辑数据被虚拟化应用5310转换成物理控制平面数据并且该物理控制平面数据被受管理交换元件接收并转换成物理转发数据。

[0600] 当VM 1打算向VM 4发送分组时，VM 1首先广播ARP请求以解析逻辑路由器5425的MAC地址。此ARP分组具有VM 1的源IP地址，在此示例中为1.1.1.10，以及VM 4的目的地IP地

址,在此示例中为1.1.2.10。此广播分组具有广播MAC地址“ff:ff:ff:ff:ff:ff”作为目的地MAC地址,并且分组的目标协议地址是1.1.1.253。此广播分组(ARP请求)被复制到受管理交换元件5320的所有端口,包括逻辑端口SP1。然后,基于流条目6102,此分组被发送到逻辑路由器5325的RP1。该分组随后根据流条目6104被发送到L3守护进程(未示出),因为目的地IP地址1.1.2.10落在由1.1.2.253/24指明的IP地址的范围中(即,因为目标协议地址是1.1.1.253)。L3守护进程将目的地IP地址解析成MAC地址01:01:01:01:01:01,这是RP1的MAC地址。L3守护进程将带有此MAC地址的ARP响应发回到VM 1。

[0601] VM 1随后向VM 4发送分组。此分组以VM 1的MAC地址作为源MAC地址,以RP1的MAC地址(01:01:01:01:01:01)作为目的地MAC地址,以VM 1的IP地址(1.1.1.10)作为源IP地址,并且以VM 4的IP地址(1.1.2.10)作为目的地IP地址。

[0602] 逻辑交换机5420随后根据指示具有目的地MAC地址01:01:01:01:01:01的分组要被发送到SP1的流条目6101将分组转发到SP1。当分组到达SP1时,分组随后根据流条目6102被发送到RP1,流条目6102指示被递送到SP1的任何分组要被发送到RP1。

[0603] 此分组随后被发送到逻辑路由器5425的入口ACL阶段,其在此示例中允许分组通过RP1。然后逻辑路由器5425根据条目6204将分组路由到下一跳NH2。此路由决策随后被加载到(实现逻辑路由器5425的受管理交换元件的)寄存器。此分组随后被馈送到下一跳查找过程,该过程使用下一跳的ID NH2来确定下一跳IP地址和分组应当被发送到的端口。在此示例中,基于行6205来确定下一跳,该行6205指示NH2的地址是分组的目的地IP地址并且分组应当被发送到的端口是RP2。

[0604] 分组随后被馈送到MAC解析过程以将目的地IP地址(1.1.2.10)解析成VM 4的MAC地址。L3守护进程解析MAC地址并将新的流条目(例如通过利用解析出的MAC地址填写流模板)放回到实现逻辑路由器5425的受管理交换元件中。根据这个新的流,分组现在具有VM 4的MAC地址作为目的地MAC地址和逻辑路由器5425的RP2的MAC地址(01:01:01:01:01:02)。

[0605] 分组随后通过逻辑路由器5425的出口ACL阶段,其在此示例中允许分组通过RP2离开。分组随后根据流条目6210被发送到SP2,流条目6210指示被递送到RP2的任何分组要被发送到SP2。然后对于逻辑交换机5330的L2处理将把分组发送到VM 4。

[0606] IX.对受管理边缘交换元件实现的修改

[0607] 虽然所有LDPS处理都被推送到受管理边缘交换元件,但仅到实际的附接物理端口集成的接口在一些实施例中处理互操作性问题。这些接口在一些实施例中为主机IP/以太网堆栈实现标准的L2/L3接口。逻辑交换机和逻辑路由器之间的接口保持在虚拟化应用的内部,因此不需要实现与现今的路由器完全相同的协议来交换信息。

[0608] 虚拟化应用在一些实施例中有责任响应发送到第一跳路由器的IP地址的ARP请求。由于逻辑路由器的MAC/IP地址绑定是静态的,所以这没有引入扩展问题。最末跳逻辑路由器在一些实施例中不具有类似的严格要求:只要使虚拟化应用知道附接端口的MAC和IP地址,其就可将它们公布给内部查找服务,内部查找服务对于端点是不暴露的,并且仅被逻辑管道执行使用。不存在向附接端口发送ARP请求的绝对需要。

[0609] 一些实施例将所要求的L3功能实现为紧挨着Open vSwitch运行的外部守护进程。在一些实施例中,该守护进程负责以下操作:

[0610] • 响应ARP请求。在一些实施例中,Open vSwitch向守护进程馈送ARP请求并且守

护进程创建响应。可替代地，一些实施例使用流模板来在受管理边缘交换元件中创建额外的流条目。流模板是使用一组规则来基于接收到的分组动态地生成一系列流条目。在一些这样的实施例中，响应由Open vSwitch自身处理。

[0611] • 建立任何状态性的(NAT、ACL、负载均衡)每流状态。同样，如果流模板足够灵活，则可以移动更多来供Open vSwitch处理。

[0612] • 发起分布式查找。当通过其逻辑管道的序列馈送流量时，根据需要对于映射服务发起分布式查找(例如ARP、学习)。这在一些实施例中将涉及使IP分组排队(queue)。

[0613] 为了在与外部物理网络集成时生成ARP请求，一些实施例假定可利用OpenFlow的LOCAL输出端口将分组丢弃到本地IP堆栈。

[0614] 映射服务本身在一些实施例中通过依赖于Open vSwitch的数据路径功能来实现：受管理边缘交换元件处的守护进程通过向映射服务节点发送特殊的“公布”分组来公布MAC和IP地址绑定，映射服务节点随后将利用流模板来创建流条目。来自受管理边缘交换元件的“查询”分组随后将被这些FIB条目响应，这些FIB条目在将查询分组修改到足以变成响应分组之后将把分组发送到特殊的IN_PORT。

[0615] X.逻辑交换环境

[0616] 上文和下文描述的若干实施例提供了将逻辑转发空间(即，逻辑控制和转发平面)与物理转发空间(即，物理控制和转发平面)完全分离的网络控制系统。这些控制系统通过使用映射引擎将逻辑转发空间数据映射到物理转发空间数据来实现这种分离。通过将逻辑空间与物理空间完全解除耦合，这些实施例的控制系统允许当对物理转发空间作出改变(例如，迁移虚拟机、添加物理交换机或路由器等等)时逻辑转发元件的逻辑视图保持不变。

[0617] 更具体而言，一些实施例的控制系统管理网络，在这些网络上，属于若干不同用户(即，具有被多个不同的相关或无关用户共享的多个容宿计算机和受管理转发元件的私有或公共容宿环境中的若干不同用户)的机器(例如虚拟机)可针对分开的LDP集合交换数据分组。也就是说，属于特定用户的机器可通过对于该用户的LDPS与属于同一用户的其它机器交换数据，而属于不同用户的机器通过在同一物理受管理网络上实现的不同的LDPS与彼此交换数据。在一些实施例中，LDPS(也称为逻辑转发元件(例如，逻辑交换机、逻辑路由器)或者在一些情况下是逻辑网络)是提供互连若干逻辑端口的交换架构的逻辑构造，特定用户的机器(物理的或虚拟的)可附接到这些逻辑端口。

[0618] 在一些实施例中，这种LDP集合和逻辑端口的创建和使用提供了一种逻辑服务模型，该模型在非专业人士看来可能与虚拟局域网(VLAN)的使用相似。然而，存在与用于分割(segment)网络的VLAN服务模型的各种显著区别。在本文描述的逻辑服务模型中，物理网络可变化，而不对用户对网络的逻辑视图有任何影响(例如，添加受管理交换元件或者将VM从一个位置移动到另一位置不会影响用户对逻辑转发元件的视图)。普通技术人员将会认识到，下文描述的所有区别可不适用于特定的受管理网络。一些受管理网络可包括这一节中描述的所有特征，而其它受管理网络将包括这些特征的不同子集。

[0619] 为了使得一些实施例的受管理网络内的受管理转发元件识别分组所属的LDPS，网络控制器集群根据定义LDP集合的用户输入自动地为物理的受管理转发元件生成流条目。当来自特定LDPS上的机器的分组被发送到受管理网络上时，受管理转发元件使用这些流条目来识别分组的逻辑上下文(即，分组所属的LDPS以及分组要前往的逻辑端口)并且根据逻

辑上下文来转发分组。

[0620] 在一些实施例中，分组在没有任何种类的逻辑上下文ID的情况下离开其源机器(以及其源机器的网络接口)。反而，分组仅包含源和目的地机器的地址(例如，MAC地址、IP地址等等)。所有的逻辑上下文信息都在网络的受管理转发元件处添加和去除。当第一受管理转发元件直接从源机器接收到分组时，该转发元件使用分组中的信息以及其接收到分组的物理端口来识别分组的逻辑上下文并将此信息附加到分组。类似地，目的地机器之前的最后一个受管理转发元件在将分组转发到其目的地之前去除逻辑上下文。此外，在一些实施例中，附加到分组的逻辑上下文可被沿途的中间受管理转发元件修改。这样，末端机器(以及末端机器的网络接口)不需要知晓发送分组的逻辑网络。结果，末端机器及其网络接口不需要被配置为适应逻辑网络。反而，网络控制器仅配置受管理转发元件。此外，因为转发处理的大部分在边缘转发元件处执行，所以对于网络的整个转发资源将随着更多机器被添加而自动扩展(scale)(因为每个物理边缘转发元件仅能使这么多的机器附接)。

[0621] 在附加(例如前置)到分组的逻辑上下文中，一些实施例仅包括逻辑出口端口。也就是说，封装分组的逻辑上下文不包括显式用户ID。反而，逻辑上下文捕获在第一跳处作出的逻辑转发决策(即，关于目的地逻辑端口的决策)。由此，在随后的转发元件处可通过检查逻辑出口端口(因为该逻辑出口端口是特定LDPS的一部分)来隐式地确定用户ID(即，分组所属的LDPS)。这导致了一种平的上下文识别符，意味着受管理转发元件不必切开(slice)上下文ID以确定ID内的多条信息。

[0622] 在一些实施例中，出口端口是32比特ID。然而，在一些实施例中对于处理逻辑上下文的受管理转发元件使用软件转发元件使得系统能够在任何时间被修改以改变逻辑上下文的大小(例如改变到64比特或更大)，而硬件转发元件往往更局限于对于上下文识别符使用特定数目的比特。此外，使用例如本文描述的逻辑上下文识别符导致逻辑数据(即，出口上下文ID)与源/目的地地址数据(即，MAC地址)之间的显式分离(explicit separation)。当源和目的地地址被映射到逻辑入口和出口端口时，该信息被分开存储在分组内。从而，在网络内的受管理交换元件处，可完全基于封装分组的逻辑数据(即，逻辑出口信息)来转发分组，而无需对物理地址信息的任何额外查找。

[0623] 在一些实施例中，受管理转发元件内的分组处理涉及重复地向调度端口发送分组，从而有效地将分组重新提交(resubmit)回交换元件中。在一些实施例中，使用软件交换元件提供了执行分组的这种重新提交的能力。硬件转发元件一般涉及固定的管道(这一部分是由于使用ASIC来执行处理)，而一些实施例的软件转发元件则可根据需要扩展分组处理管道，因为没有太多来自执行重新提交的延迟。

[0624] 此外，一些实施例使得能够优化对于单组相关分组(例如单个TCP/UDP流)内的后续分组的多个查找。当第一分组到达时，受管理转发元件执行所有的查找并重新提交以便完全处理该分组。转发元件随后将决策的最终结果(例如，向分组添加出口上下文，以及通过特定隧道离开转发元件的特定端口的下一跳转发决策)与分组的唯一识别符一起缓存(即，TCP/UDP流的唯一识别符)，该唯一识别符将与所有其它的相关分组共享。一些实施例将该缓存的结果推送到转发元件的内核中以进行额外的优化。对于共享唯一识别符的额外分组(即，同一流内的额外分组)，转发元件可使用指明要对分组执行的所有动作的单个缓存查找。一旦分组的流完成(例如，在经过了没有分组匹配识别符的特定量的时间之后)，在

一些实施例中，转发元件清除缓存。对多个查找的使用在一些实施例中涉及将分组从物理空间(例如，物理端口处的MAC地址)映射到逻辑空间(例如，到逻辑交换机的逻辑端口的逻辑转发决策)，然后映射回到物理空间(例如，将逻辑出口上下文映射到交换元件的物理输出端口)。

[0625] 使用封装来提供物理和逻辑地址的显式分离的这种逻辑网络相对于网络虚拟化的其它方法(例如VLAN)提供了显著优势。例如，标签技术(例如，VLAN)使用置于分组上的标签来分割转发表以仅向分组应用与标签相关联的规则。这仅分割了现有的地址空间，而不是引入了新的空间。结果，因为地址被用于虚拟和物理领域中的实体，所以它们必须被暴露给物理转发表。这样，来自于层次化地址映射的聚集的属性不能被利用。此外，因为利用标签技术没有引入新的地址空间，所以所有的虚拟上下文都必须使用相同的地址模型，并且虚拟地址空间限于与物理地址空间相同。标签技术的另一缺点在于不能通过地址重映射来利用移动性。

[0626] XI.电子系统

[0627] 图64概念性示出了实现本发明的一些实施例的电子系统6400。电子系统6400可用于执行上文描述的任何控制、虚拟化或操作系统应用。电子系统6400可以是计算机(例如，桌面计算机、个人计算机、平板计算机、服务器计算机、大型机、刀片计算机等等)、电话、PDA或任何其它种类的电子设备。这种电子系统包括各种类型的计算机可读介质和针对各种其它类型的计算机可读介质的接口。电子系统6400包括总线6405、处理单元6410、系统存储器6425、只读存储器6430、永久性存储设备6435、输入设备6440和输出设备6445。

[0628] 总线6405共同代表通信地连接电子系统6400的大量内部设备的所有系统、外围和芯片集总线。例如，总线6405将处理单元6410与只读存储器6430、系统存储器6425和永久性存储设备6435通信地连接。

[0629] 从这些各种存储器单元，处理单元6410检索指令来执行并检索数据来处理以便执行本发明的过程。处理单元在不同实施例中可以是单个处理器或者多核处理器。

[0630] 只读存储器(ROM)6430存储处理单元6410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，永久性存储设备6435是读写存储器设备。此设备是即使当电子系统6400关断时也存储指令和数据的非易失性存储器单元。本发明的一些实施例使用大容量存储设备(例如磁盘或光盘及其相应的盘驱动器)作为永久性存储设备6435。

[0631] 其它实施例使用可移除存储设备(例如软盘、闪存盘等等)作为永久性存储设备。与永久性存储设备6435一样，系统存储器6425是读写存储器设备。然而，与永久性存储设备6435不同，系统存储器是易失性读写存储器，例如随机访问存储器。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6425、永久性存储设备6435和/或只读存储器6430中。从这些各种存储器单元，处理单元6410检索指令来执行并检索数据来处理以便执行一些实施例的过程。

[0632] 总线6405还连接到输入和输出设备6440、6445。输入设备使得用户能够向电子系统传输信息并选择对电子系统的命令。输入设备6440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6445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，例如阴极射线管(CRT)或液晶显示器(LCD)。一些实施例包括诸如触摸屏的用作输入设备和输出设备二者的设备。

[0633] 最后，如图64所示，总线6405还通过网络适配器(未示出)将电子系统6400耦合到网络6465。以这种方式，计算机可以是计算机的网络(例如局域网(“LAN”)、广域网(“WAN”)或内联网或者网络的网络(例如因特网)的一部分。电子系统6400的任何或所有组件可结合本发明被使用。

[0634] 一些实施例包括电子组件，例如微处理器、存储设备和存储器，它们将计算机程序指令存储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。这种计算机可读介质的一些示例包括RAM、ROM、只读致密盘(CD-ROM)、可记录致密盘(CD-R)、可再写致密盘(CD-RW)、只读数字多功能盘(例如，DVD-ROM、双层DVD-ROM)、多种可记录/可再写DVD(例如，DVD-RAM、DVD-RW、DVD+RW等等)、闪存(例如，SD卡、袖珍SD卡、微型SD卡等等)、磁和/或固态硬盘驱动器、只读和可记录蓝光®盘、超密度光盘、任何其它光或磁介质、以及软盘。计算机可读介质可存储计算机程序，该计算机程序可由至少一个处理单元执行并且包括用于执行各种操作的指令集。计算机程序或计算机代码的示例包括机器代码，例如由编译器产生的那种，以及由计算机、电子组件或微处理器使用解释器来执行的包括更高级代码的文件。

[0635] 虽然以上论述主要涉及执行软件的微处理器或多核处理器，但一些实施例由诸如专用集成电路(ASIC)或现场可编程门阵列(FPGA)的一个或多个集成电路执行。在一些实施例中，这种集成电路执行存储在电路本身上的指令。

[0636] 如本说明书中所使用，术语“计算机”、“服务器”、“处理器”和“存储器”都指的是电子或其它技术设备。这些术语排除人或人的群组。对于本说明书的目的，术语“显示”意味着在电子设备上显示。如本说明书中所使用，术语“计算机可读介质”和“机器可读介质”完全限于以可由计算机读取的形式存储信息的有形物理对象。这些术语排除任何无线信号、有线下载信号和任何其它短暂信号。

[0637] 虽然已参照大量具体的细节描述了本发明，但普通技术人员将认识到，在不脱离本发明的精神的情况下，可以其它具体的形式来实现本发明。此外，多幅附图(包括图14、16、32、35、49、51以及52)概念性示出了过程。这些过程的具体操作可不以所示出和描述的确切顺序执行。具体的操作可不在一个连续的操作系列中执行，并且不同的具体操作可在不同的实施例中执行。另外，过程可利用若干子过程来实现，或者实现为更大的宏过程的一部分。从而，普通技术人员将理解，本发明不受前述说明性细节所限，而应由所附权利要求来限定。

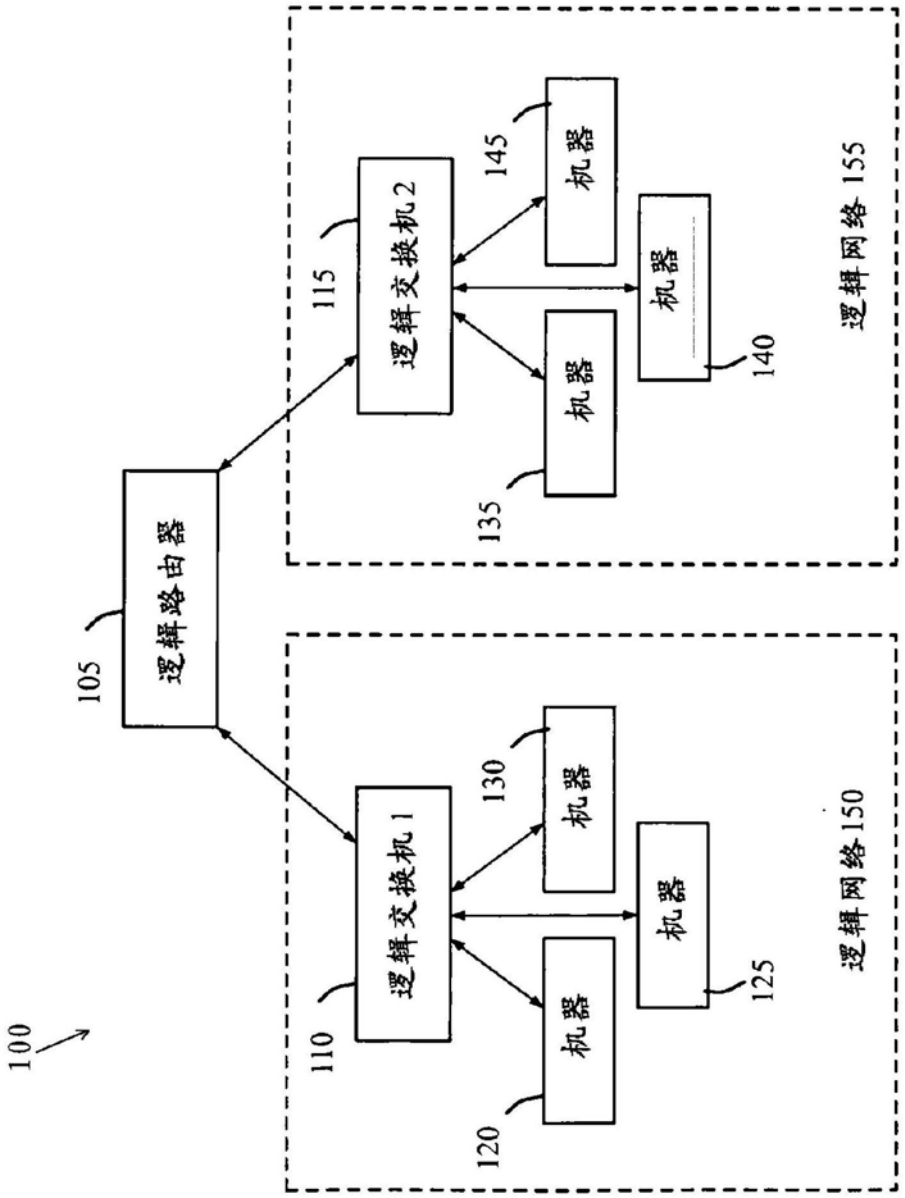

100
105
逻辑路由器
逻辑网络 150
逻辑交换机 1
110
机器
120
机器
125
机器
130
逻辑网络 155
逻辑交换机 2
115
机器
135
机器
140
机器
145

图1

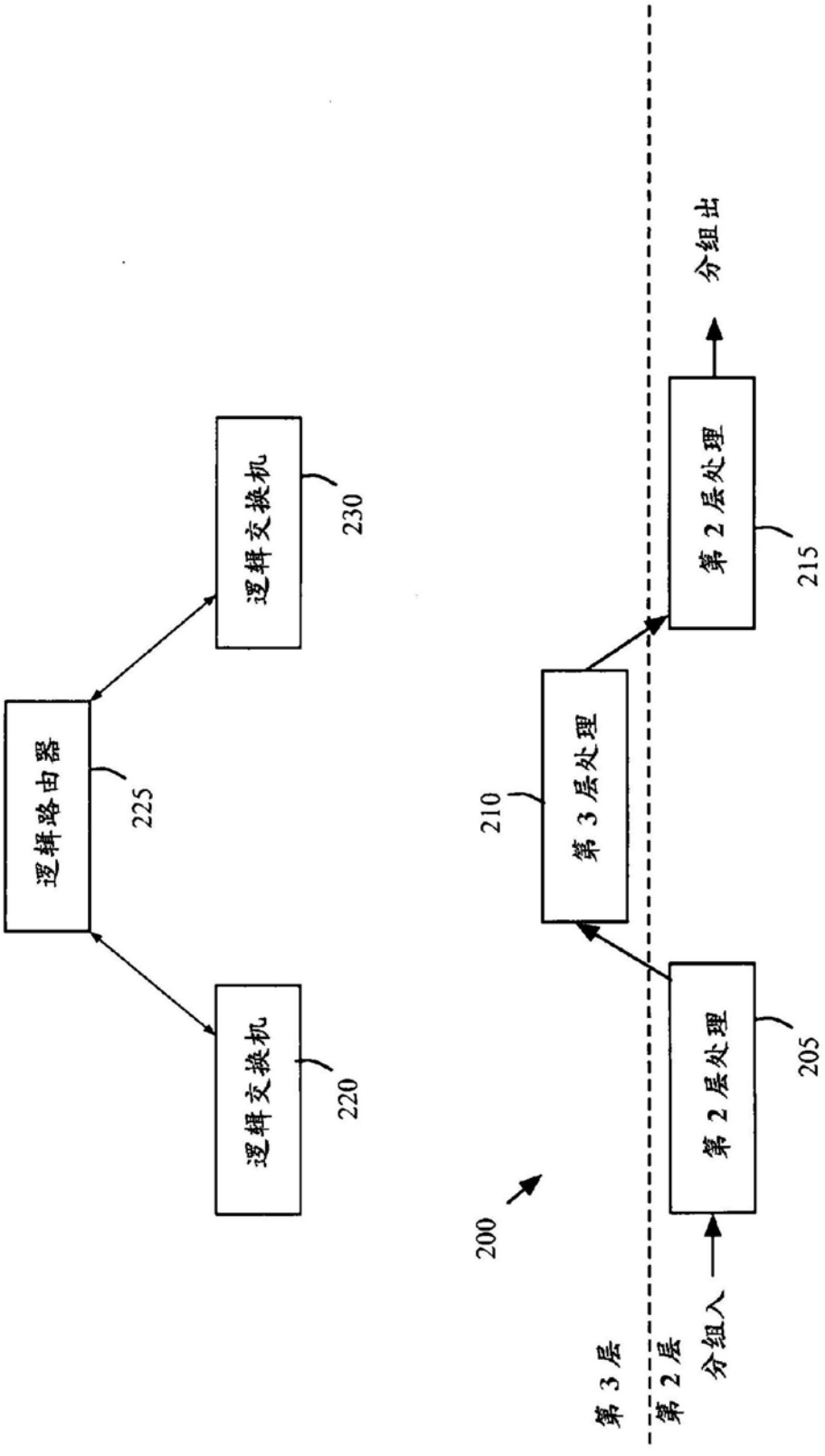

逻辑交换机
230
逻辑路由器
225
逻辑交换机
220
第 3 层处理
210
第 2 层处理
215
分组出
第 2 层处理
205
200
第 3 层
第 2 层
分组入

图2

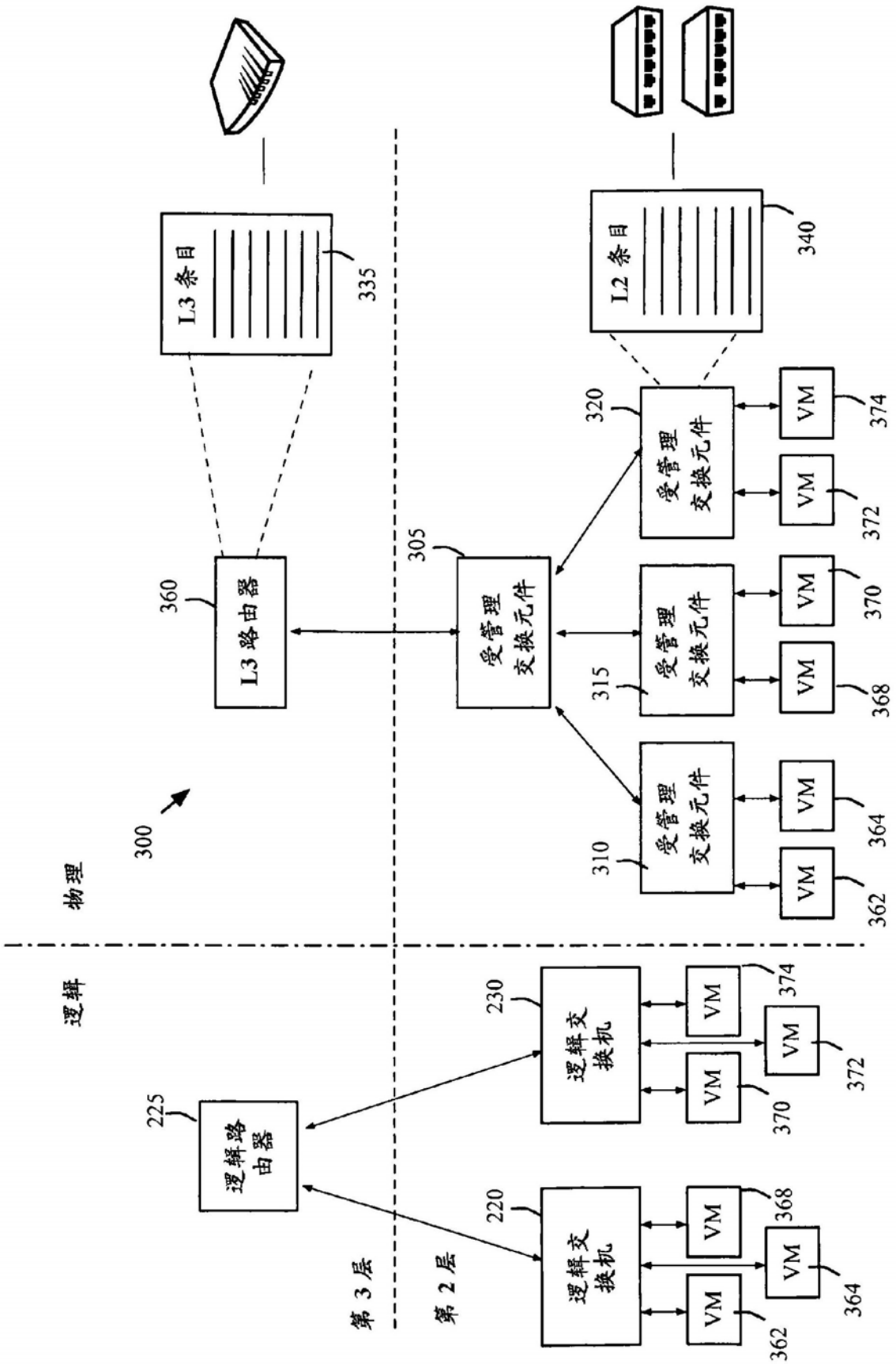

物理
逻辑
300
第 3 层
第 2 层
L3 条目
335
L2 条目
340
L3 路由器
360
受管理交换元件
305
受管理交换元件
320
受管理交换元件
315
受管理交换元件
310
逻辑路由器
225
逻辑交换机
230
逻辑交换机
220
VM
362
364
368
370
372
374

图3

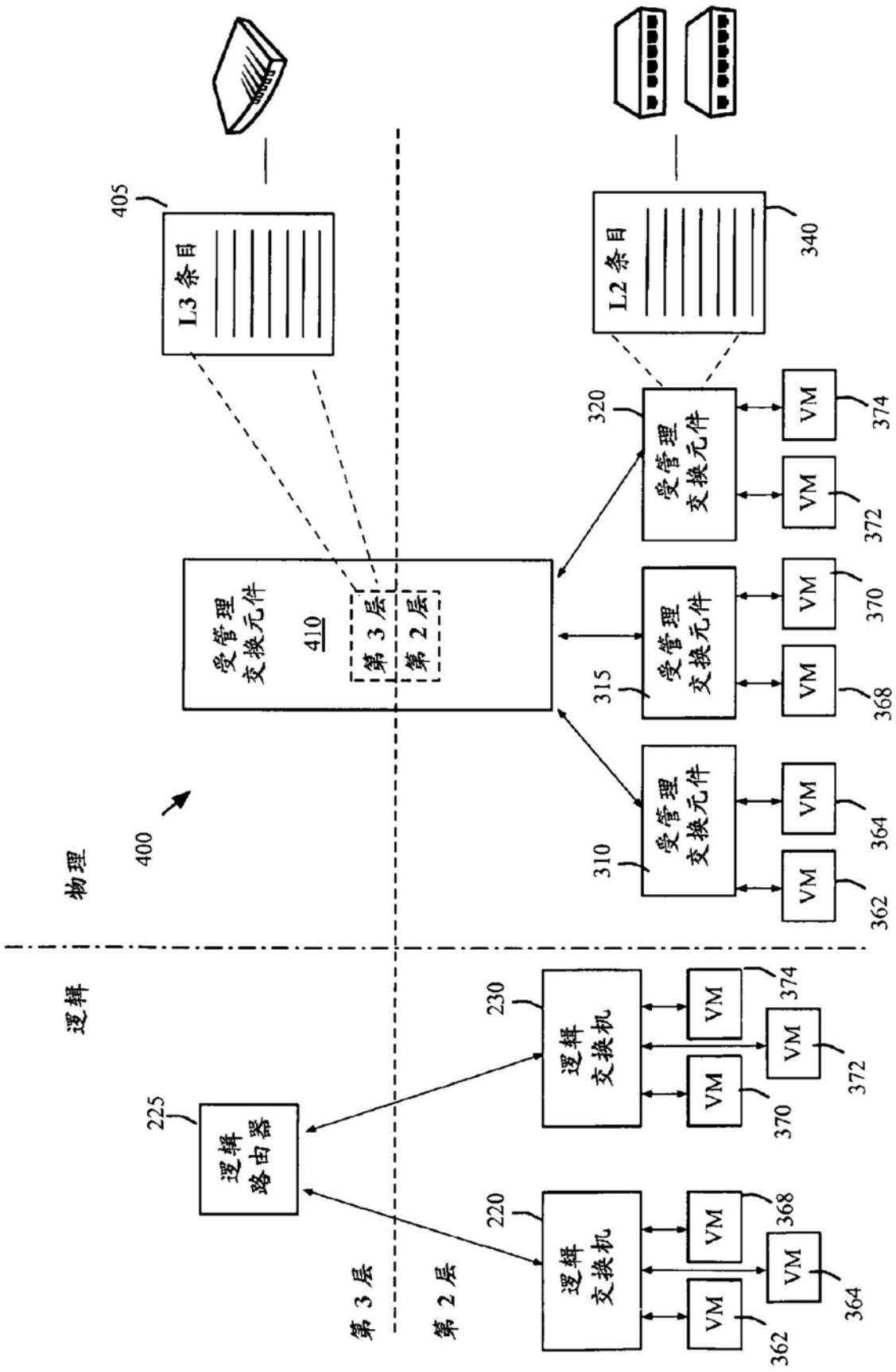

405
L3 条目
340
L2 条目
320
受管理交换元件
315
受管理交换元件
310
受管理交换元件
受管理交换元件
410
第 3 层
第 2 层
400
物理
逻辑
225
逻辑路由器
230
逻辑交换机
220
逻辑交换机
VM
362
364
368
370
372
374
第 3 层
第 2 层

图4

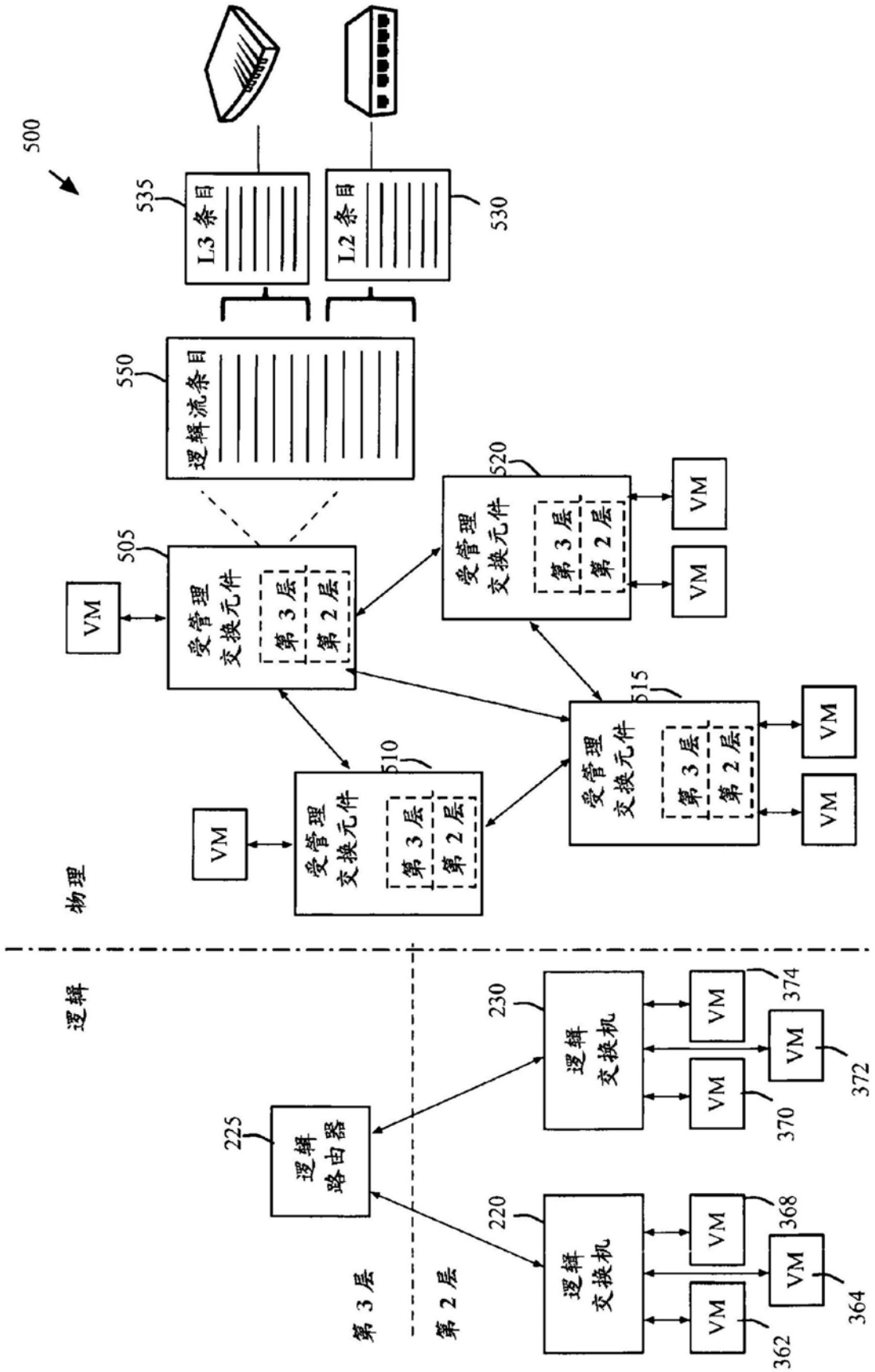

500
535
L3 条目
530
L2 条目
550
逻辑流条目
505
受管理交换元件
第 3 层
第 2 层
520
受管理交换元件
第 3 层
第 2 层
515
受管理交换元件
第 3 层
第 2 层
510
受管理交换元件
第 3 层
第 2 层
VM
物理
逻辑
225
逻辑路由器
230
逻辑交换机
220
逻辑交换机
第 3 层
第 2 层
362
364
368
370
372
374

图5

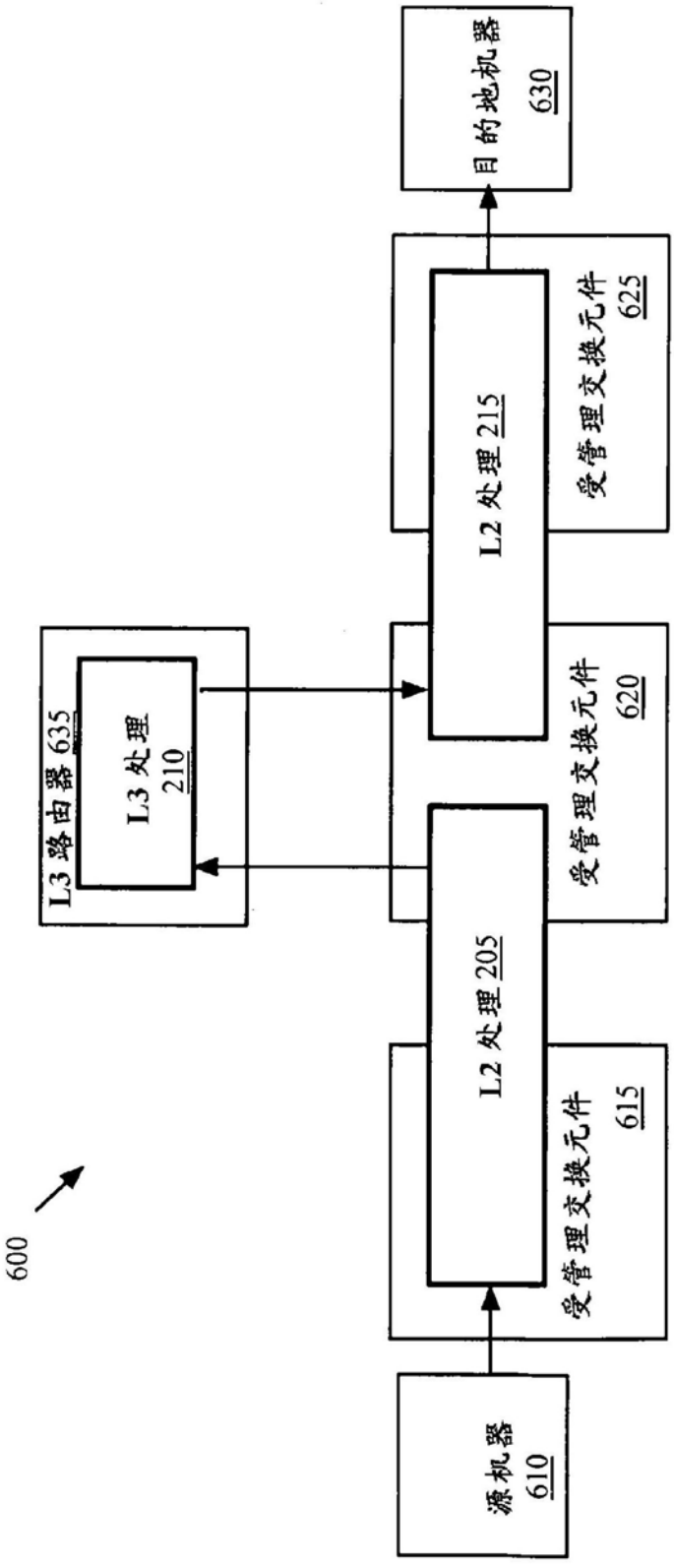

目的地机器
630
受管理交换元件
625
L2 处理 215
受管理交换元件
620
L3 路由器 635
L3 处理
210
L2 处理 205
受管理交换元件
615
源机器
610
600

图6

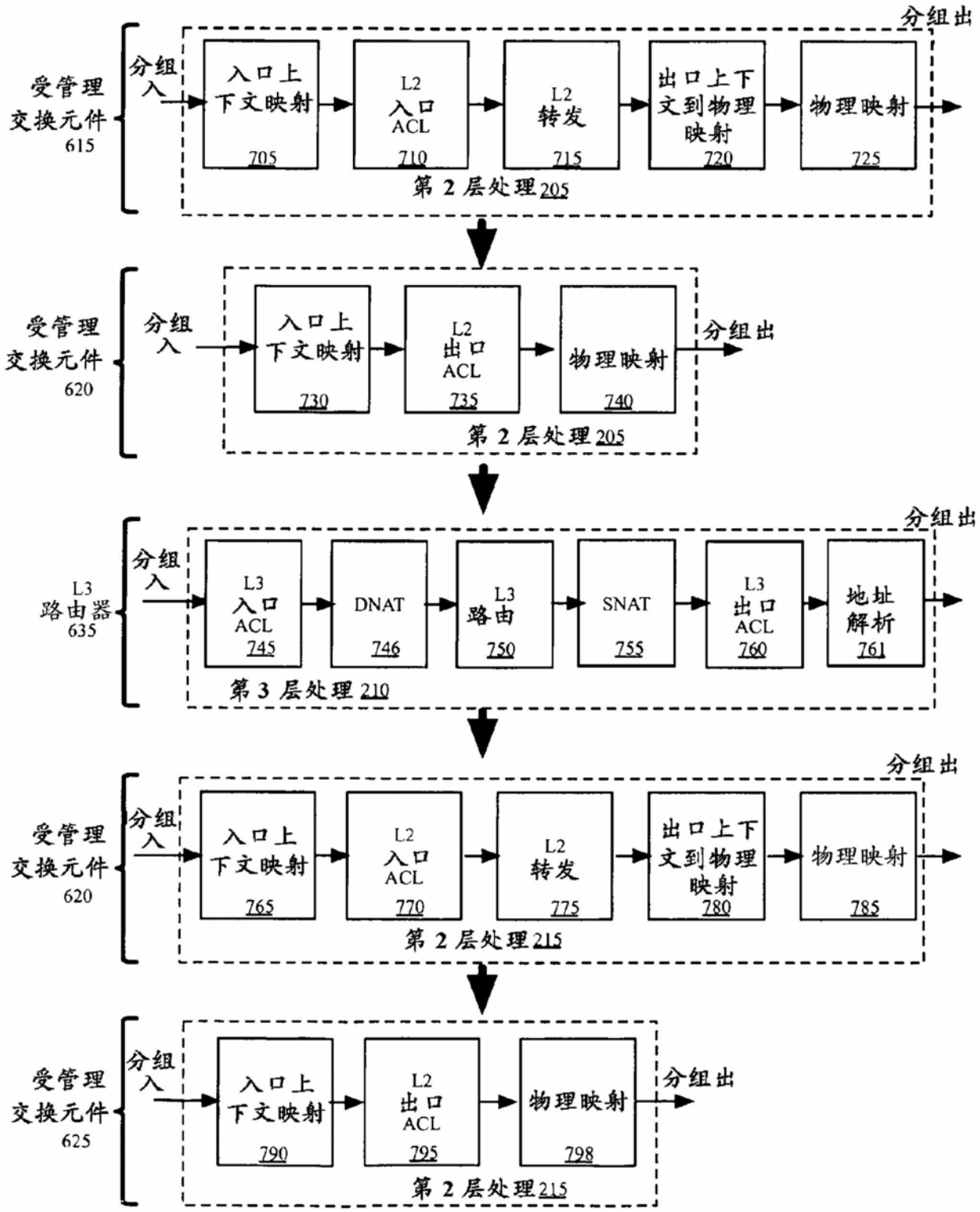

受管理交换元件 615
分组入
入口上下文映射 705
L2 入口 ACL 710
L2 转发 715
出口上下文到物理映射 720
物理映射 725
分组出
第 2 层处理 205
受管理交换元件 620
分组入
入口上下文映射 730
L2 出口 ACL 735
物理映射 740
分组出
第 2 层处理 205
L3 路由器 635
分组入
L3 入口 ACL 745
DNAT 746
L3 路由 750
SNAT 755
L3 出口 ACL 760
地址解析 761
分组出
第 3 层处理 210
受管理交换元件 620
分组入
入口上下文映射 765
L2 入口 ACL 770
L2 转发 775
出口上下文到物理映射 780
物理映射 785
分组出
第 2 层处理 215
受管理交换元件 625
分组入
入口上下文映射 790
L2 出口 ACL 795
物理映射 798
分组出
第 2 层处理 215

图7

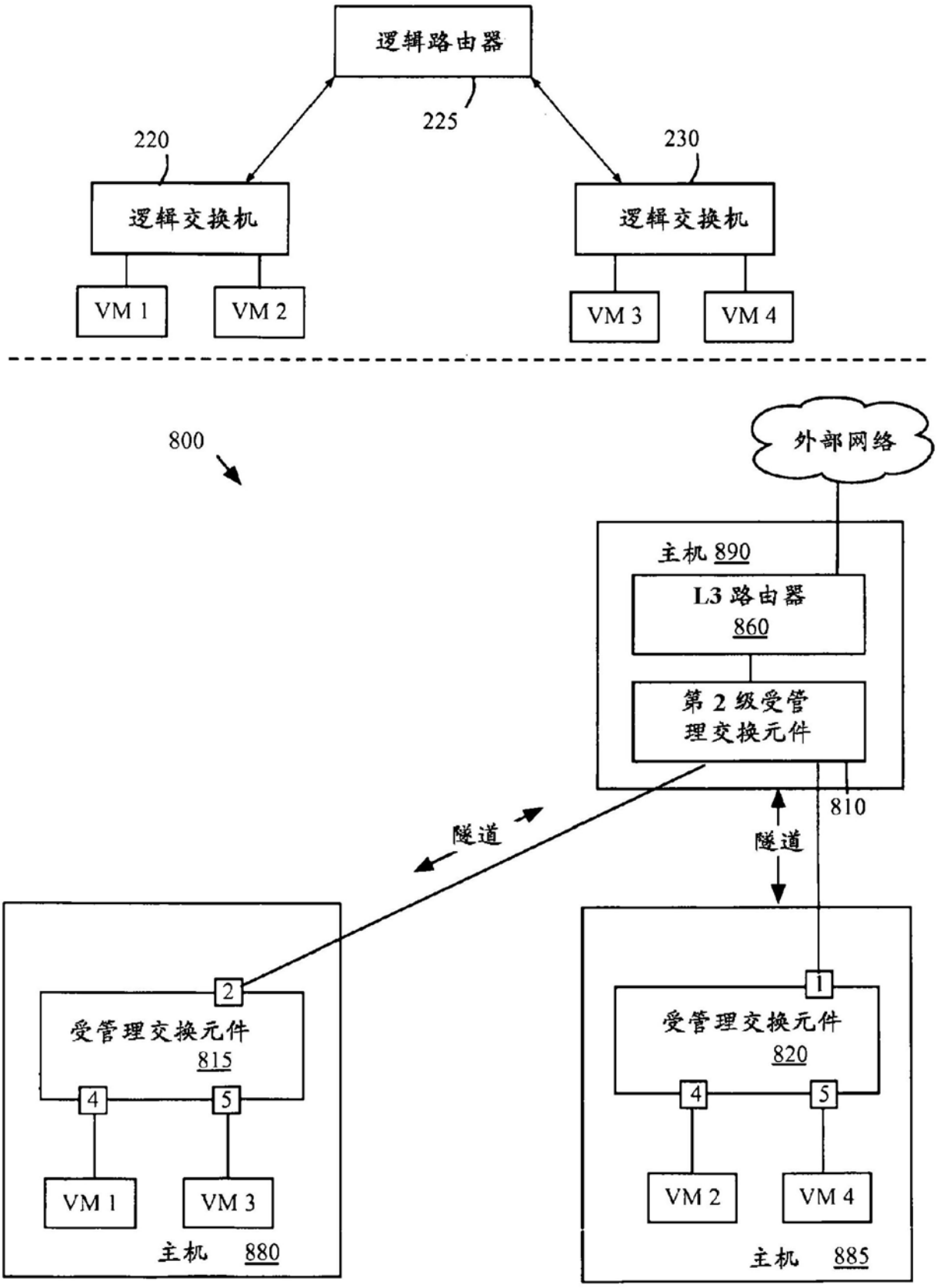

逻辑路由器
225
220
230
逻辑交换机
逻辑交换机
VM 1
VM 2
VM 3
VM 4
800
外部网络
主机 890
L3 路由器
860
第 2 级受管
理交换元件
810
隧道
隧道
2
受管理交换元件
815
4
5
VM 1
VM 3
主机 880
1
受管理交换元件
820
4
5
VM 2
VM 4
主机 885

图8

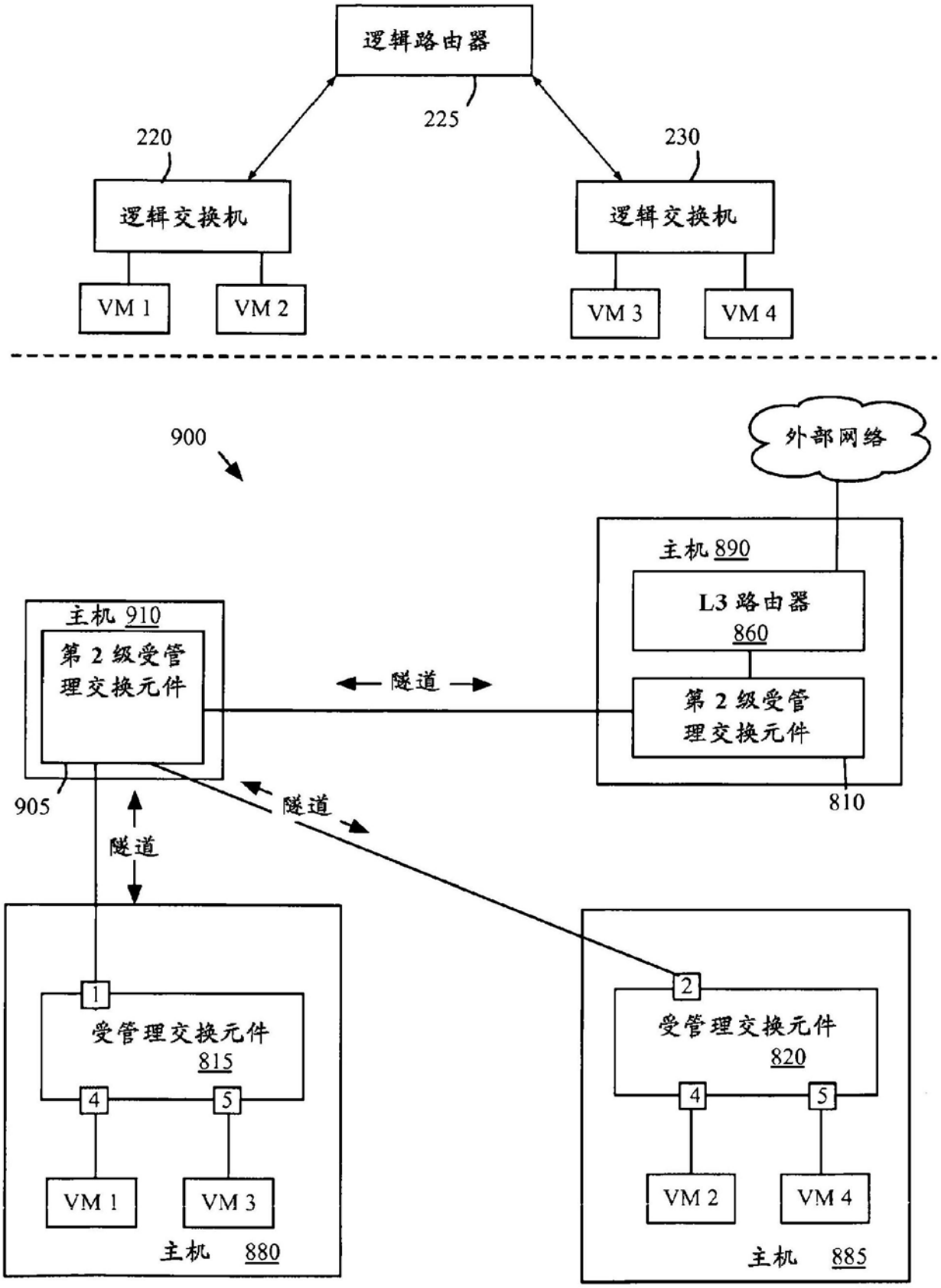

逻辑路由器
225
220
230
逻辑交换机
逻辑交换机
VM 1
VM 2
VM 3
VM 4
900
外部网络
主机 890
L3 路由器
860
主机 910
第 2 级受管理交换元件
隧道
第 2 级受管理交换元件
905
810
隧道
隧道
1
受管理交换元件
815
2
受管理交换元件
820
4
5
4
5
VM 1
VM 3
VM 2
VM 4
主机 880
主机 885

图9

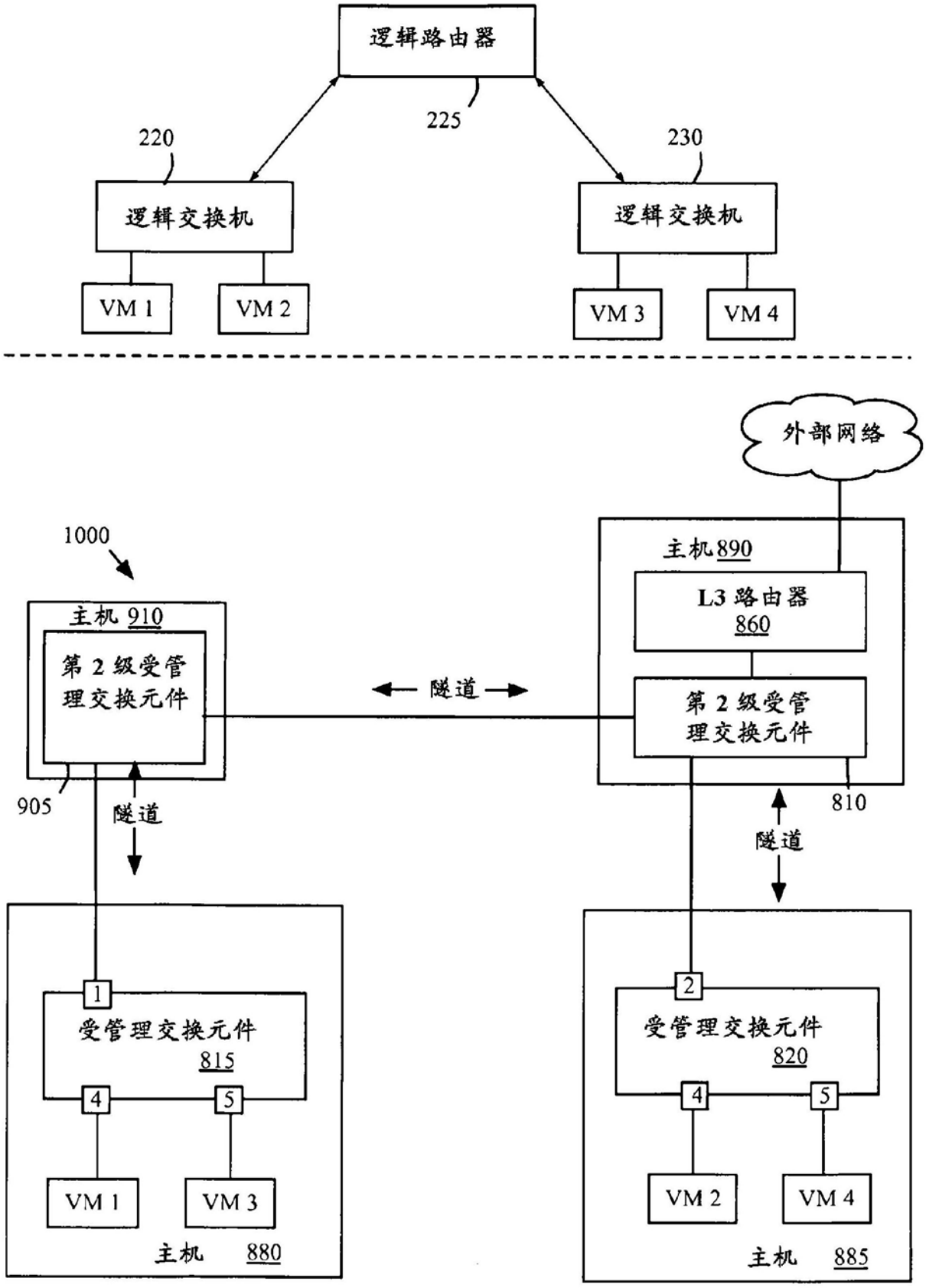

逻辑路由器
225
220
230
逻辑交换机
逻辑交换机
VM 1
VM 2
VM 3
VM 4
外部网络
1000
主机 890
L3 路由器
860
主机 910
第 2 级受管理交换元件
隧道
第 2 级受管理交换元件
905
隧道
810
隧道
1
受管理交换元件
815
2
受管理交换元件
820
4
5
4
5
VM 1
VM 3
VM 2
VM 4
主机 880
主机 885

图10

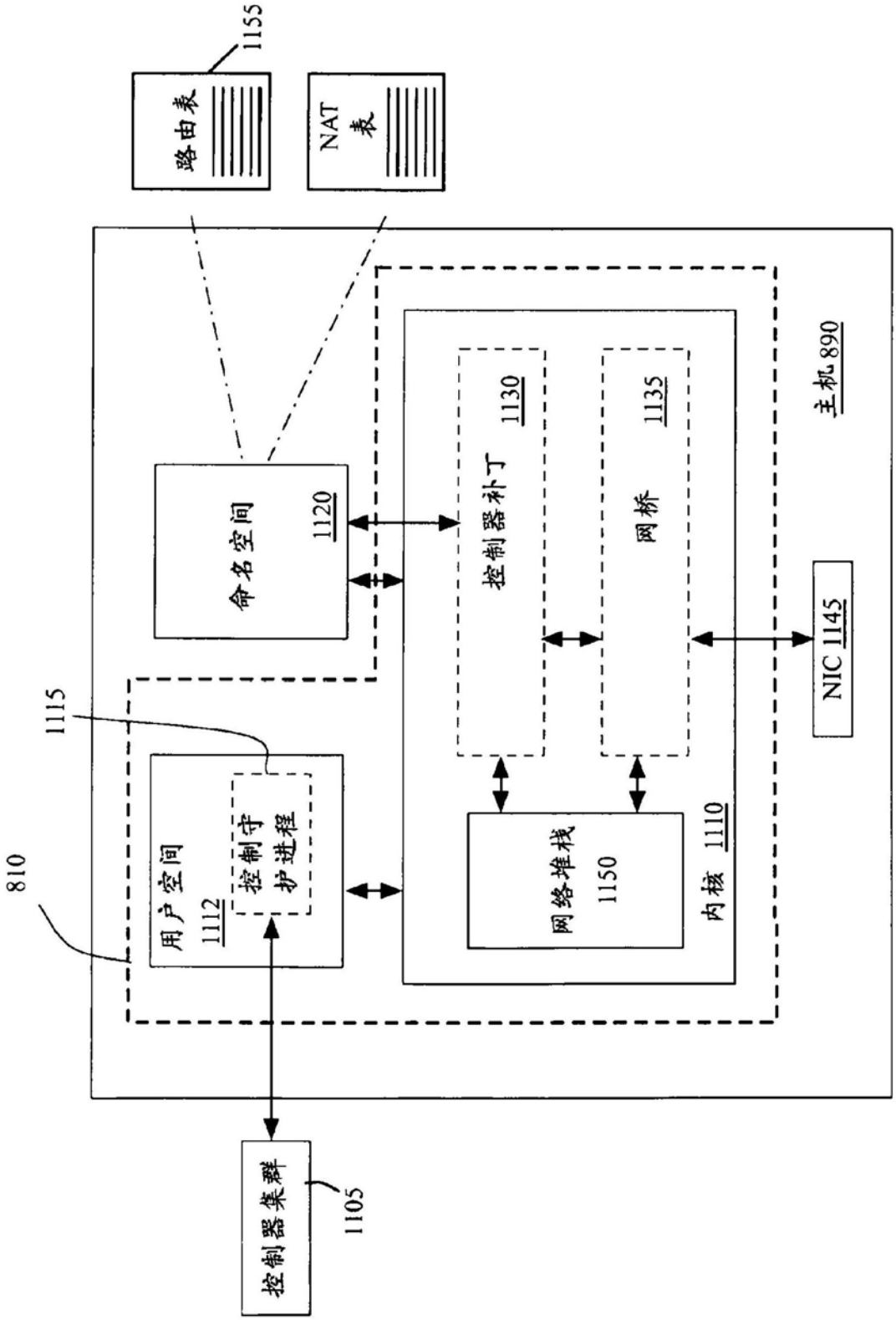

1155
路由表
NAT
表
主机 890
命名空间
1120
控制器补丁
1130
网桥
1135
NIC 1145
1115
用户空间
1112
控制守
护进程
810
网络堆栈
1150
内核 1110
控制器集群
1105

图11

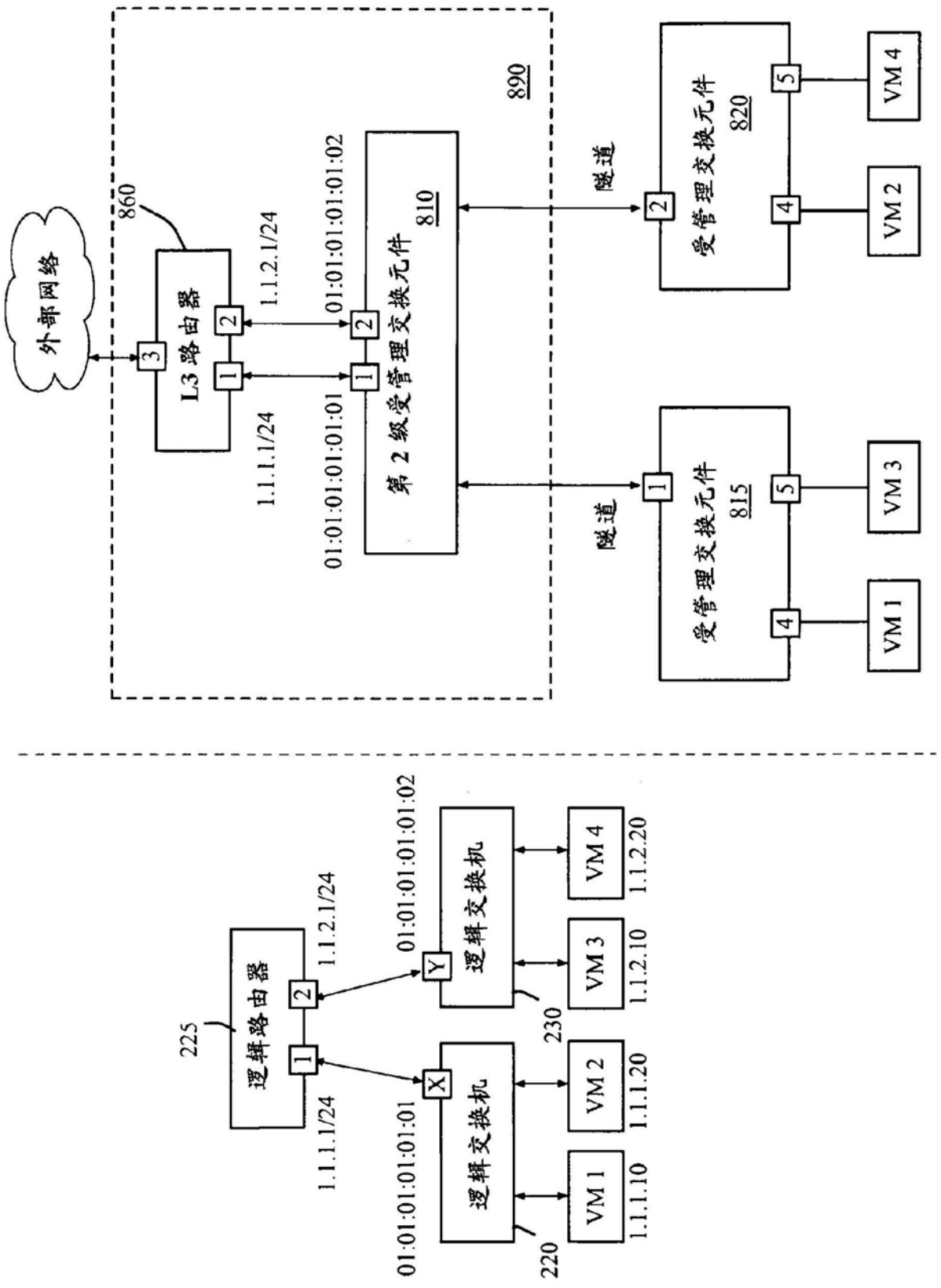

外部网络
860
L3 路由器
1.1.1.1/24
1.1.2.1/24
01:01:01:01:01:01
01:01:01:01:01:02
第 2 级受管理交换元件
810
890
隧道
隧道
受管理交换元件
820
受管理交换元件
815
VM 1
VM 2
VM 3
VM 4
225
逻辑路由器
1.1.1.1/24
1.1.2.1/24
01:01:01:01:01:01
01:01:01:01:01:02
逻辑交换机
逻辑交换机
220
230
VM 1
VM 2
VM 3
VM 4
1.1.1.10
1.1.1.20
1.1.2.10
1.1.2.20

图12

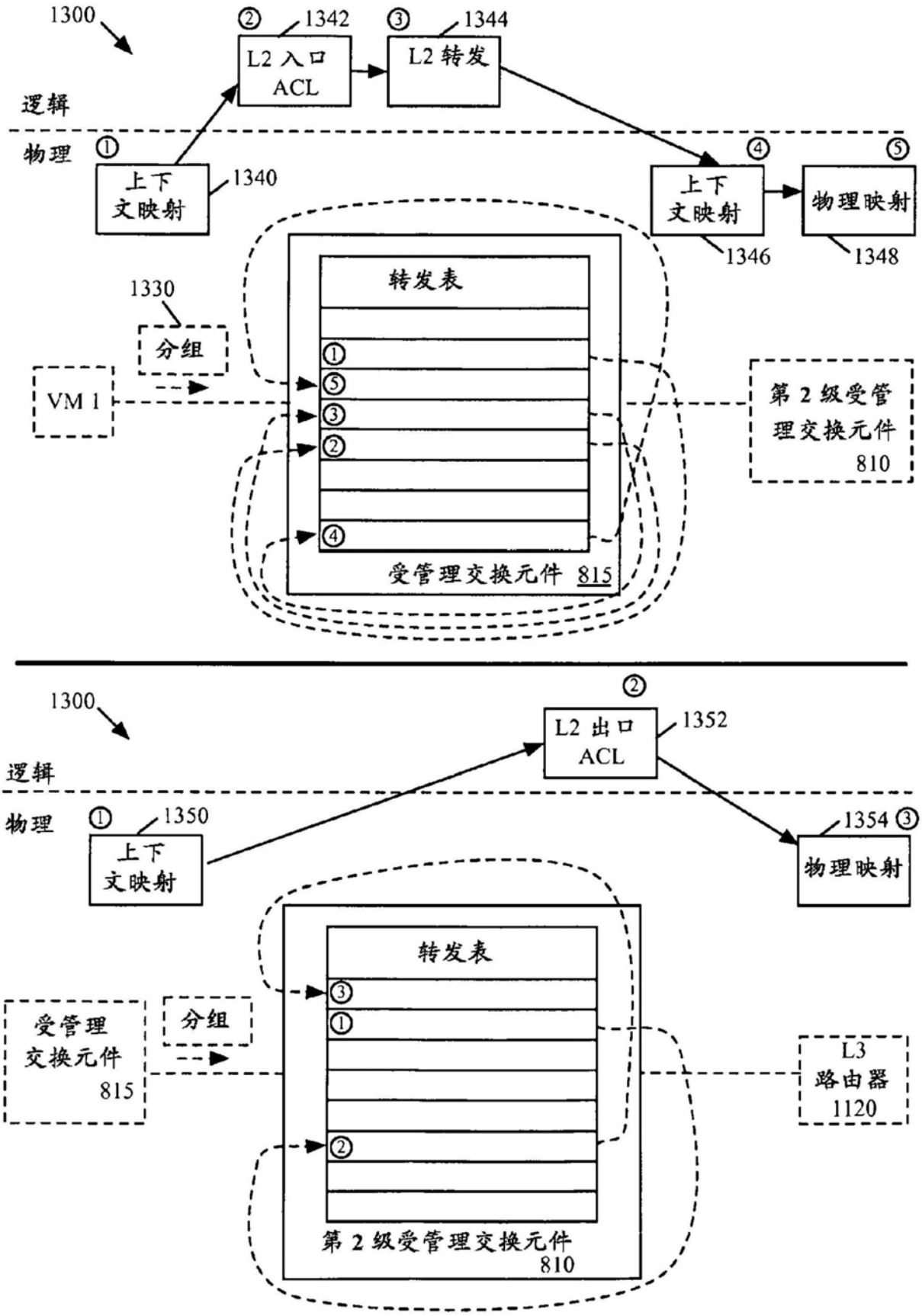

1300
逻辑
物理
L2 入口 ACL
1342
L2 转发
1344
上下文映射
1340
上下文映射
1346
物理映射
1348
转发表
1330
分组
VM 1
第 2 级受管理交换元件 810
受管理交换元件 815
1300
逻辑
物理
L2 出口 ACL
1352
上下文映射
1350
物理映射
1354
转发表
受管理交换元件 815
分组
L3 路由器 1120
第 2 级受管理交换元件 810

图13A

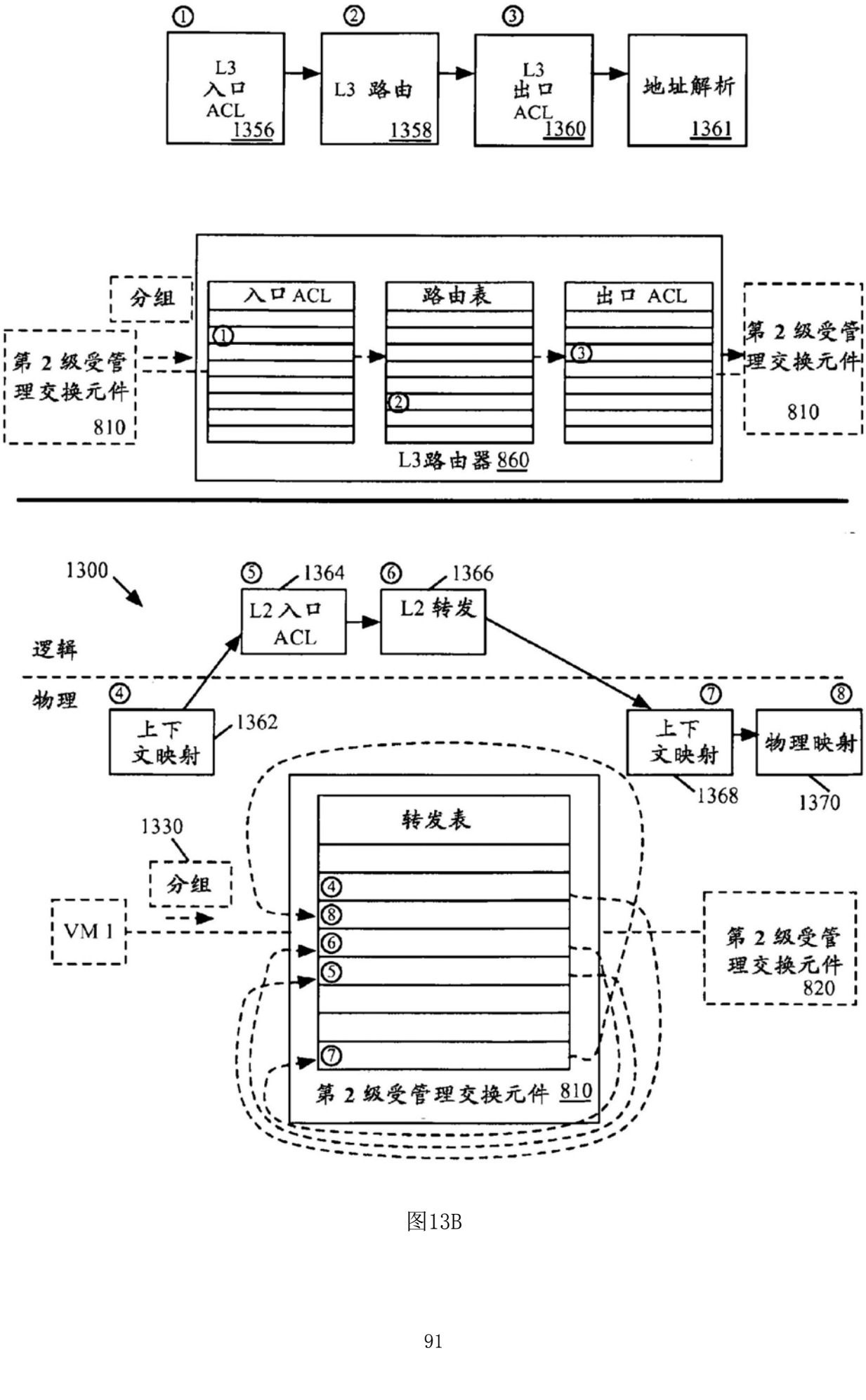

①
L3 入口 ACL 1356
②
L3 路由 1358
③
L3 出口 ACL 1360
地址解析 1361
分组
第 2 级受管理交换元件 810
入口 ACL
路由表
出口 ACL
第 2 级受管理交换元件 810
L3路由器 860
1300
逻辑
物理
⑤ 1364
L2 入口 ACL
⑥ 1366
L2 转发
④
上下文映射 1362
⑦
上下文映射 1368
⑧
物理映射 1370
1330
分组
VM 1
转发表
第 2 级受管理交换元件 820
第 2 级受管理交换元件 810

图13B

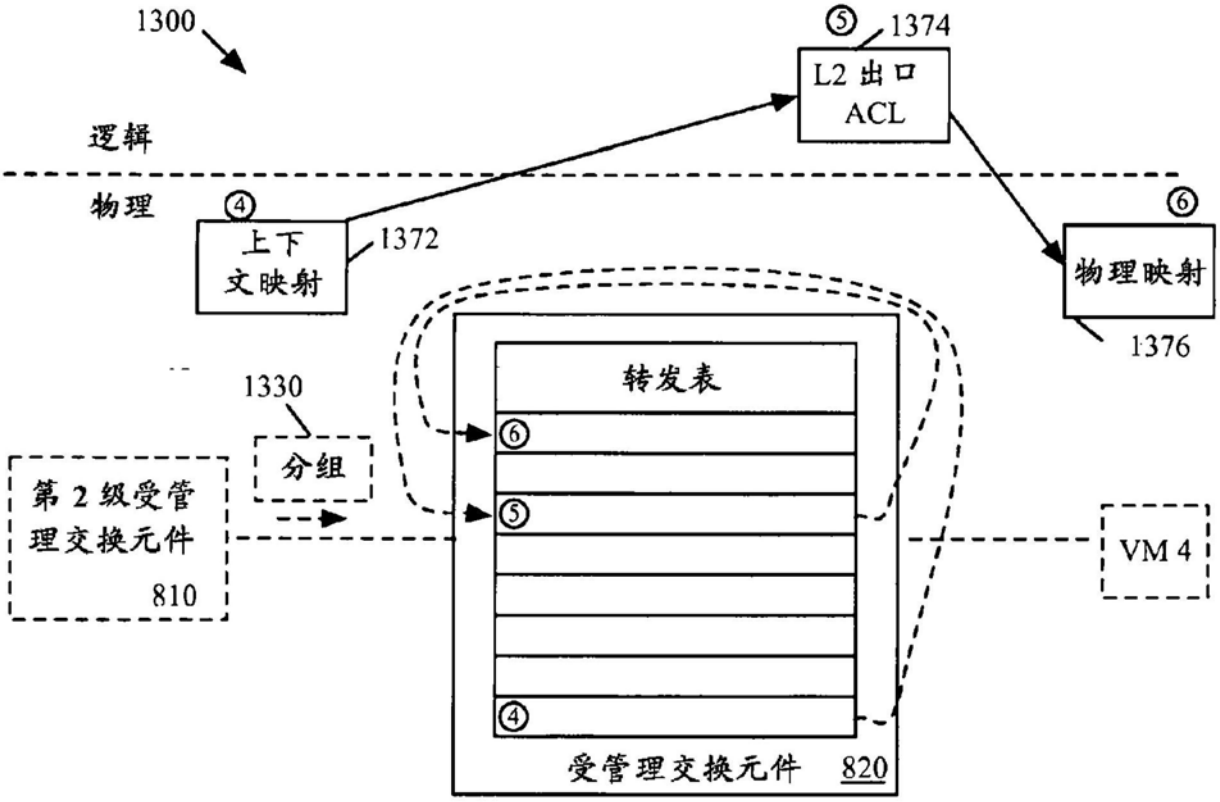

1300
逻辑
物理
⑤ 1374
L2 出口
ACL
④
上下
文映射
1372
⑥
物理映射
1376
1330
分组
第 2 级受管
理交换元件
810
转发表
⑥
⑤
④
受管理交换元件 820
VM 4

图13C

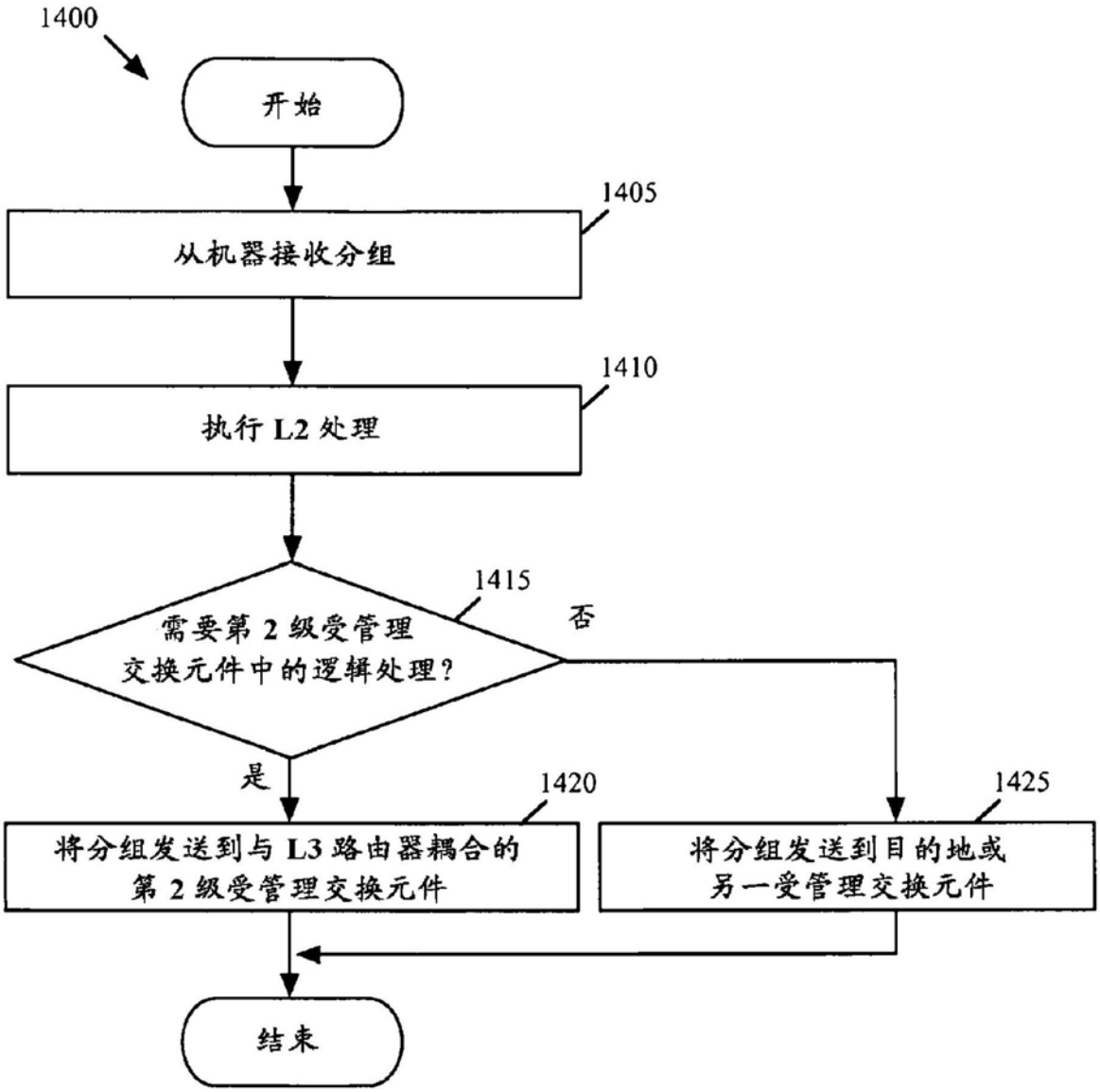

1400
开始
1405
从机器接收分组
1410
执行 L2 处理
1415
需要第 2 级受管理
交换元件中的逻辑处理?
否
是
1420
将分组发送到与 L3 路由器耦合的
第 2 级受管理交换元件
1425
将分组发送到目的地或
另一受管理交换元件
结束

图14

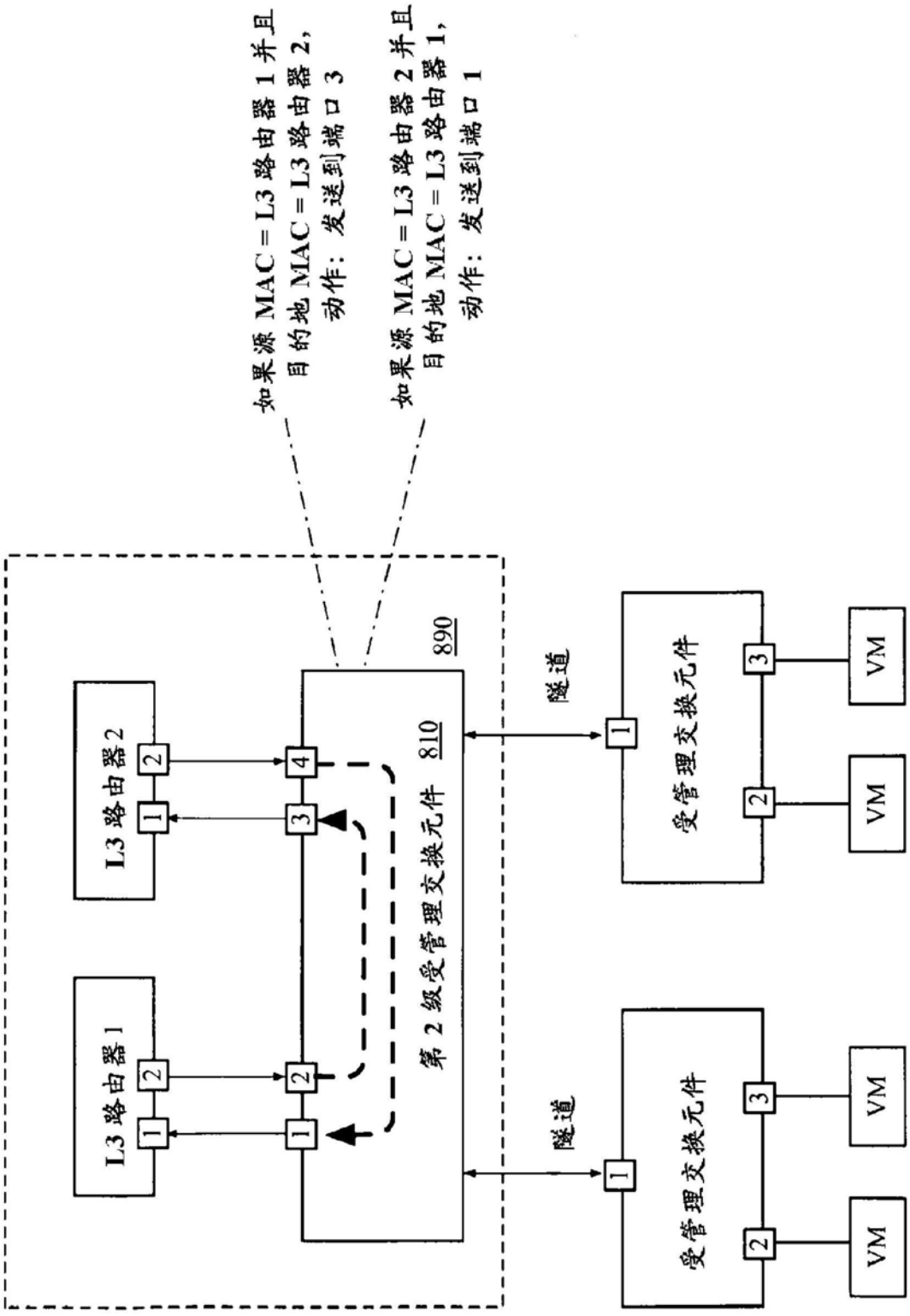

如果源 MAC = L3 路由器 1 并且
目的地 MAC = L3 路由器 2,
动作: 发送到端口 3
如果源 MAC = L3 路由器 2 并且
目的地 MAC = L3 路由器 1,
动作: 发送到端口 1
890
第 2 级受管理交换元件 810
L3 路由器 1
L3 路由器 2
隧道
隧道
受管理交换元件
受管理交换元件
VM
VM
VM
VM

图15

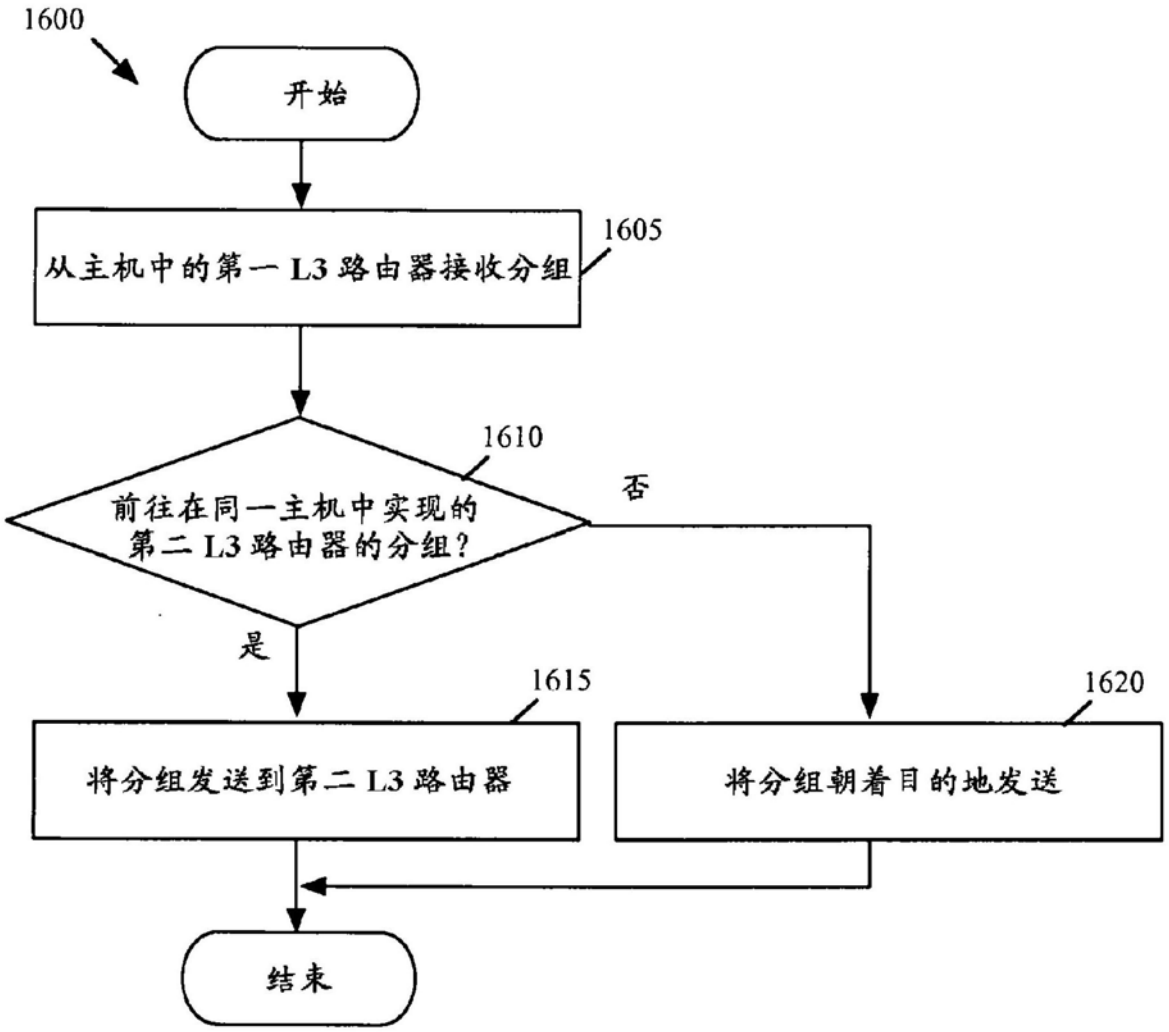

1600
开始
从主机中的第一 L3 路由器接收分组
1605
1610
前往在同一主机中实现的
第二 L3 路由器的分组?
否
是
1615
将分组发送到第二 L3 路由器
1620
将分组朝着目的地发送
结束

图16

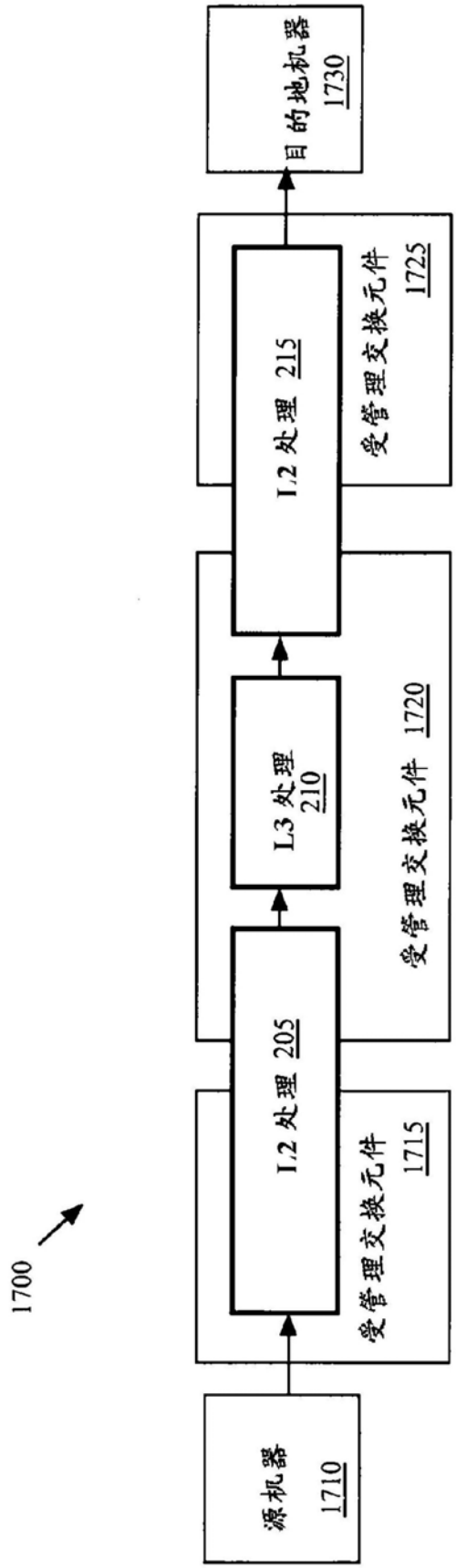

1700
源机器
1710
受管理交换元件
1715
L2 处理 205
L3 处理
210
受管理交换元件 1720
L2 处理 215
受管理交换元件
1725
目的地机器
1730

图17

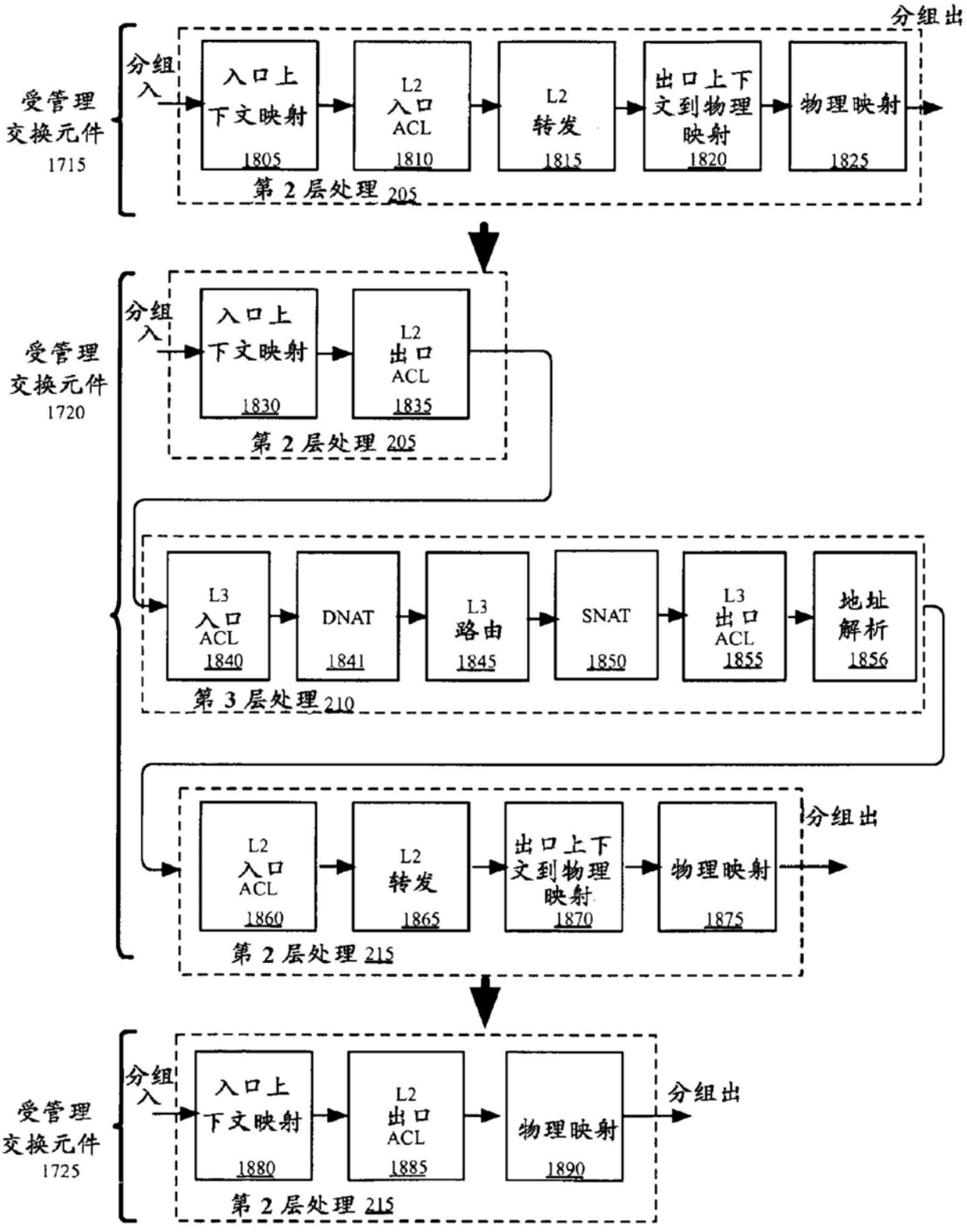

分组出
受管理交换元件 1715
分组入
入口上下文映射 1805
L2 入口 ACL 1810
L2 转发 1815
出口上下文到物理映射 1820
物理映射 1825
第 2 层处理 205
受管理交换元件 1720
分组入
入口上下文映射 1830
L2 出口 ACL 1835
第 2 层处理 205
L3 入口 ACL 1840
DNAT 1841
L3 路由 1845
SNAT 1850
L3 出口 ACL 1855
地址解析 1856
第 3 层处理 210
L2 入口 ACL 1860
L2 转发 1865
出口上下文到物理映射 1870
物理映射 1875
分组出
第 2 层处理 215
受管理交换元件 1725
分组入
入口上下文映射 1880
L2 出口 ACL 1885
物理映射 1890
分组出
第 2 层处理 215

图18

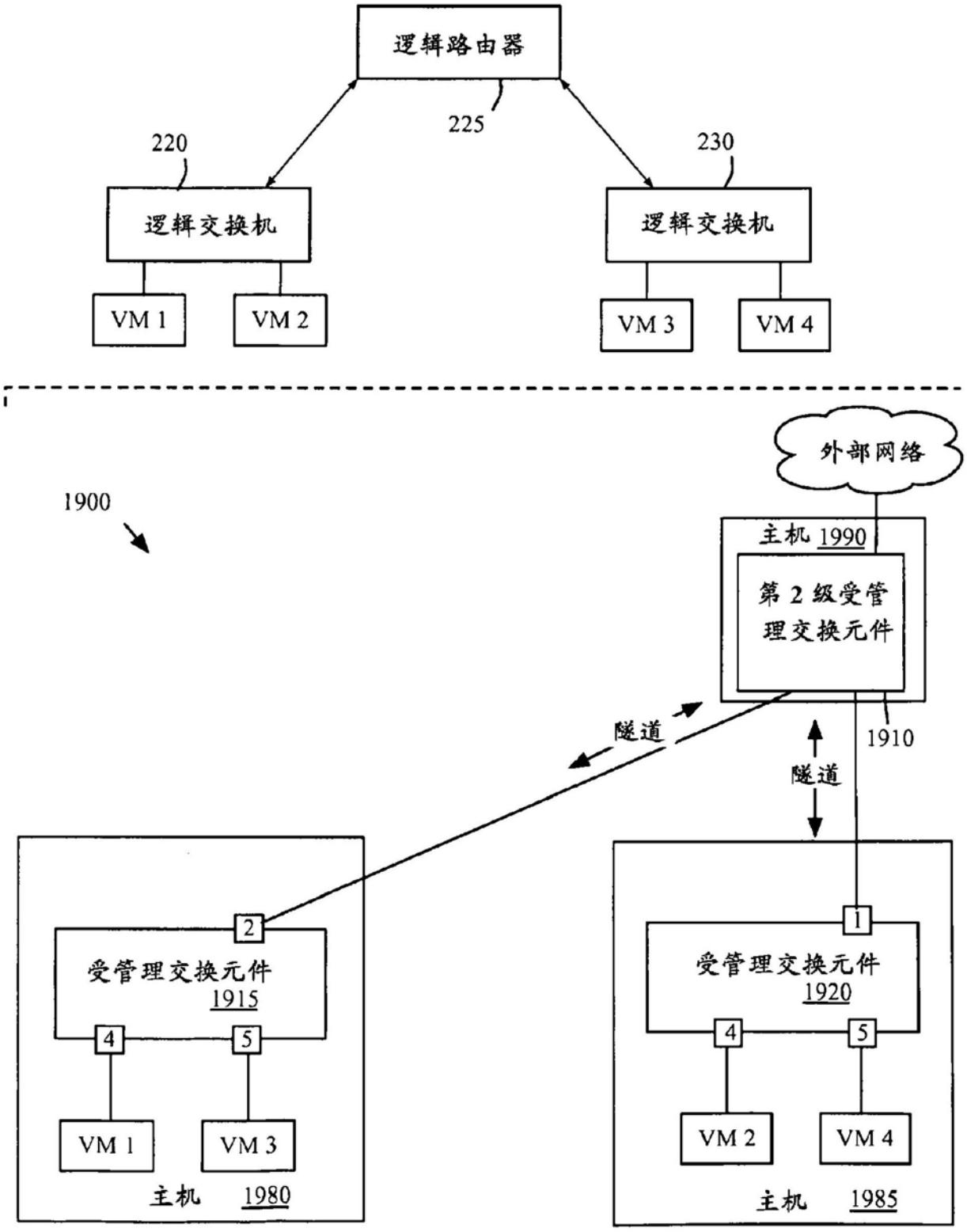

逻辑路由器
225
220
230
逻辑交换机
逻辑交换机
VM 1
VM 2
VM 3
VM 4
外部网络
1900
主机 1990
第 2 级受管
理交换元件
隧道
1910
隧道
2
受管理交换元件
1915
4
5
VM 1
VM 3
主机 1980
1
受管理交换元件
1920
4
5
VM 2
VM 4
主机 1985

图19

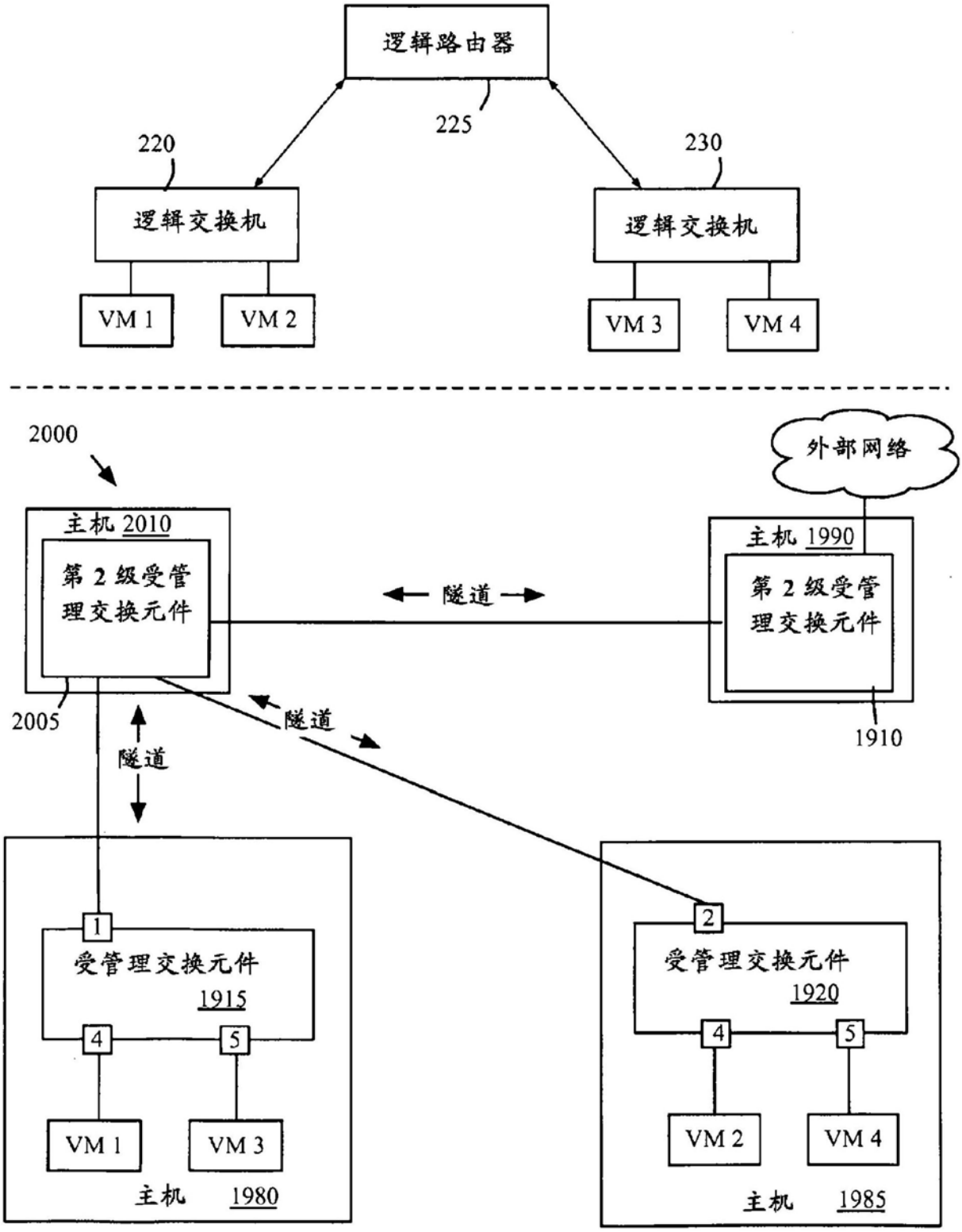

逻辑路由器
225
220
230
逻辑交换机
逻辑交换机
VM 1
VM 2
VM 3
VM 4
2000
外部网络
主机 2010
第 2 级受管理交换元件
隧道
主机 1990
第 2 级受管理交换元件
2005
隧道
隧道
1910
1
受管理交换元件
1915
2
受管理交换元件
1920
4
5
4
5
VM 1
VM 3
VM 2
VM 4
主机 1980
主机 1985

图20

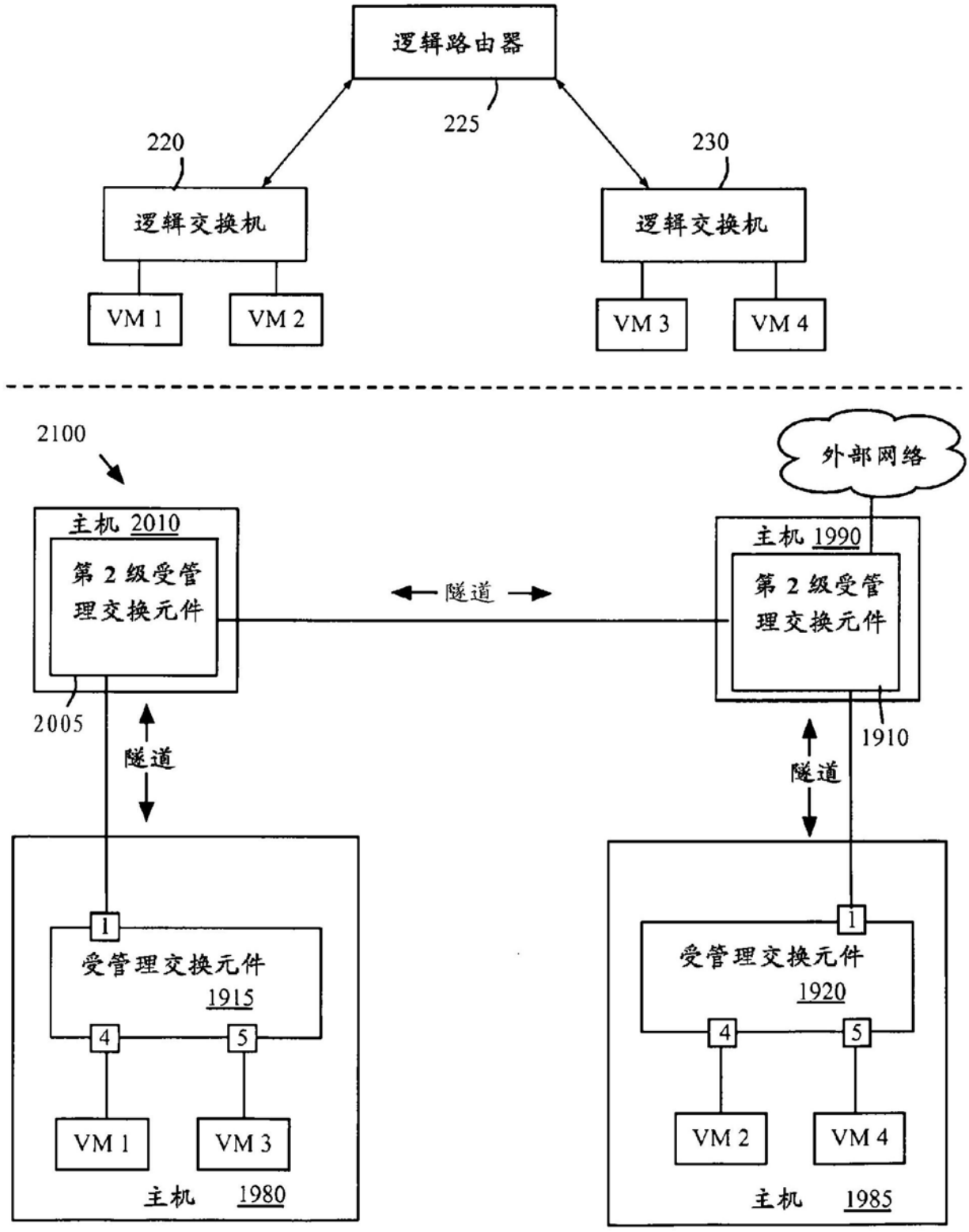

逻辑路由器
225
220
230
逻辑交换机
逻辑交换机
VM 1
VM 2
VM 3
VM 4
2100
外部网络
主机 2010
第 2 级受管理交换元件
隧道
主机 1990
第 2 级受管理交换元件
2005
隧道
1910
隧道
1
受管理交换元件 1915
4
5
VM 1
VM 3
主机 1980
1
受管理交换元件 1920
4
5
VM 2
VM 4
主机 1985

图21

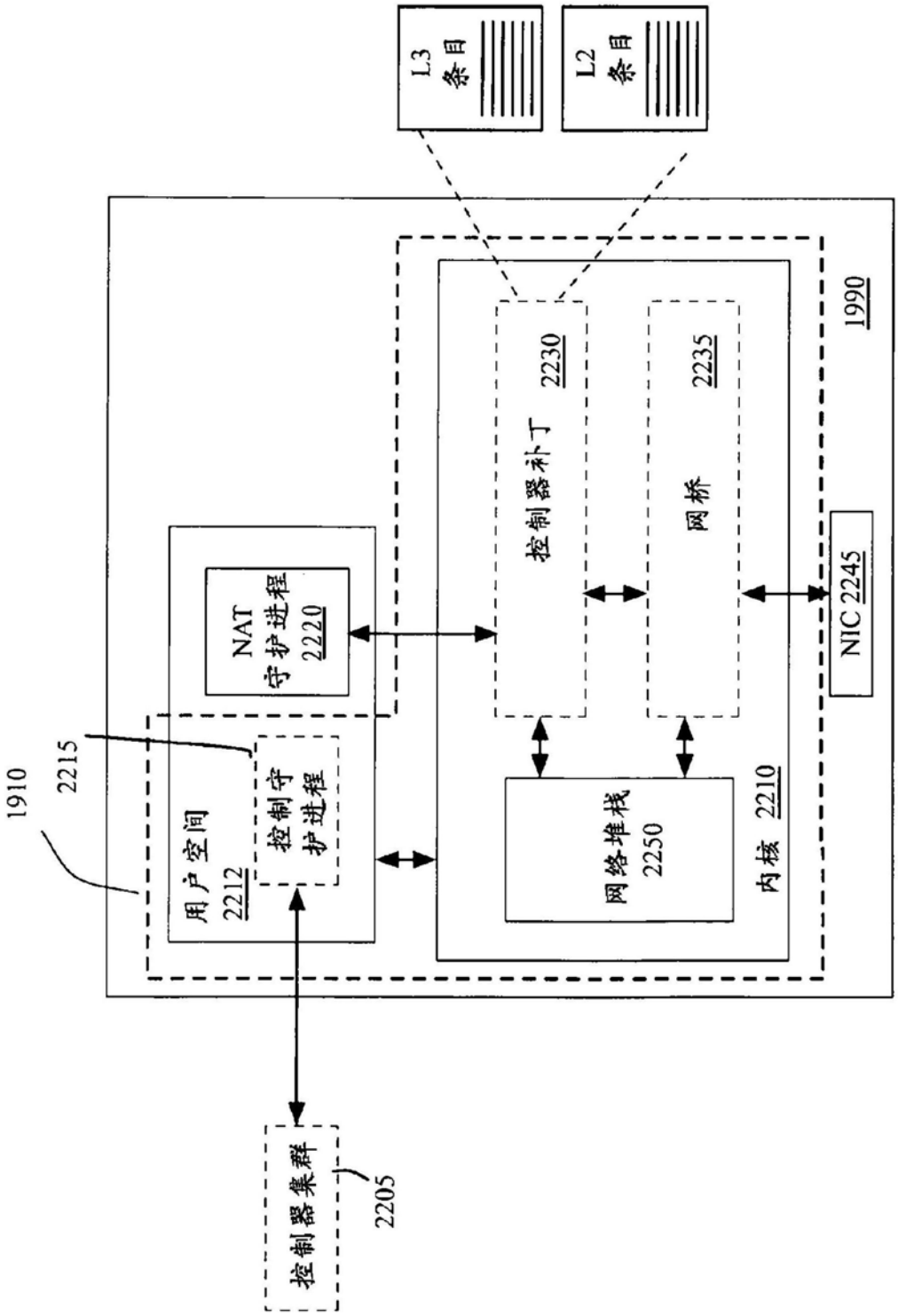

L3 条目
L2 条目
1990
控制器补丁 2230
网桥 2235
NIC 2245
NAT 守护进程 2220
1910
2215
用户空间 2212
控制守护进程
网络堆栈 2250
内核 2210
控制器集群
2205

图22

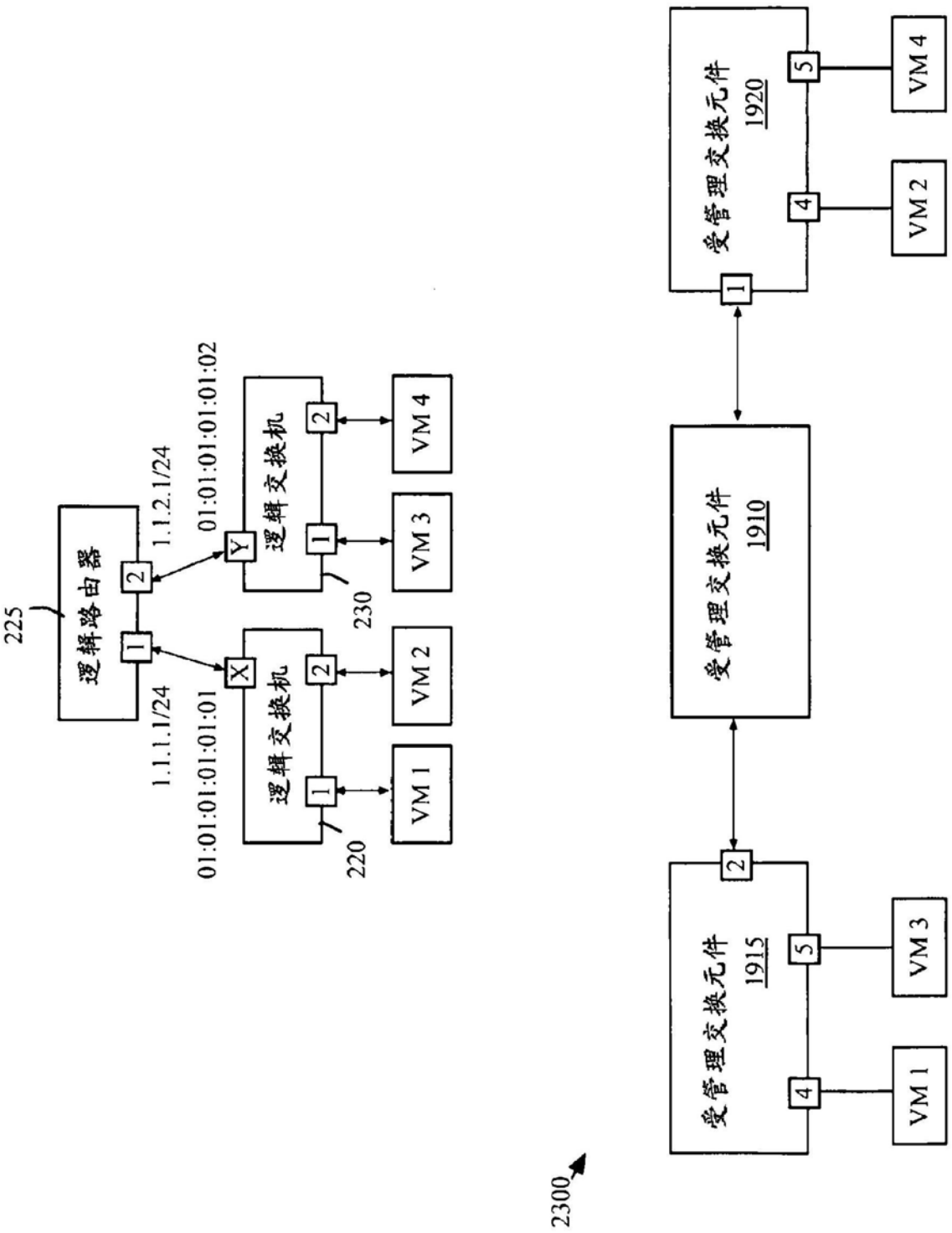

逻辑路由器
225
1
2
1.1.1.1/24
01:01:01:01:01:01
1.1.2.1/24
01:01:01:01:01:02
X
Y
逻辑交换机
逻辑交换机
1
2
1
2
220
230
VM 1
VM 2
VM 3
VM 4
2300
受管理交换元件
1915
受管理交换元件
1910
受管理交换元件
1920
1
2
4
5
4
5
1
2
VM 1
VM 3
VM 2
VM 4

图23

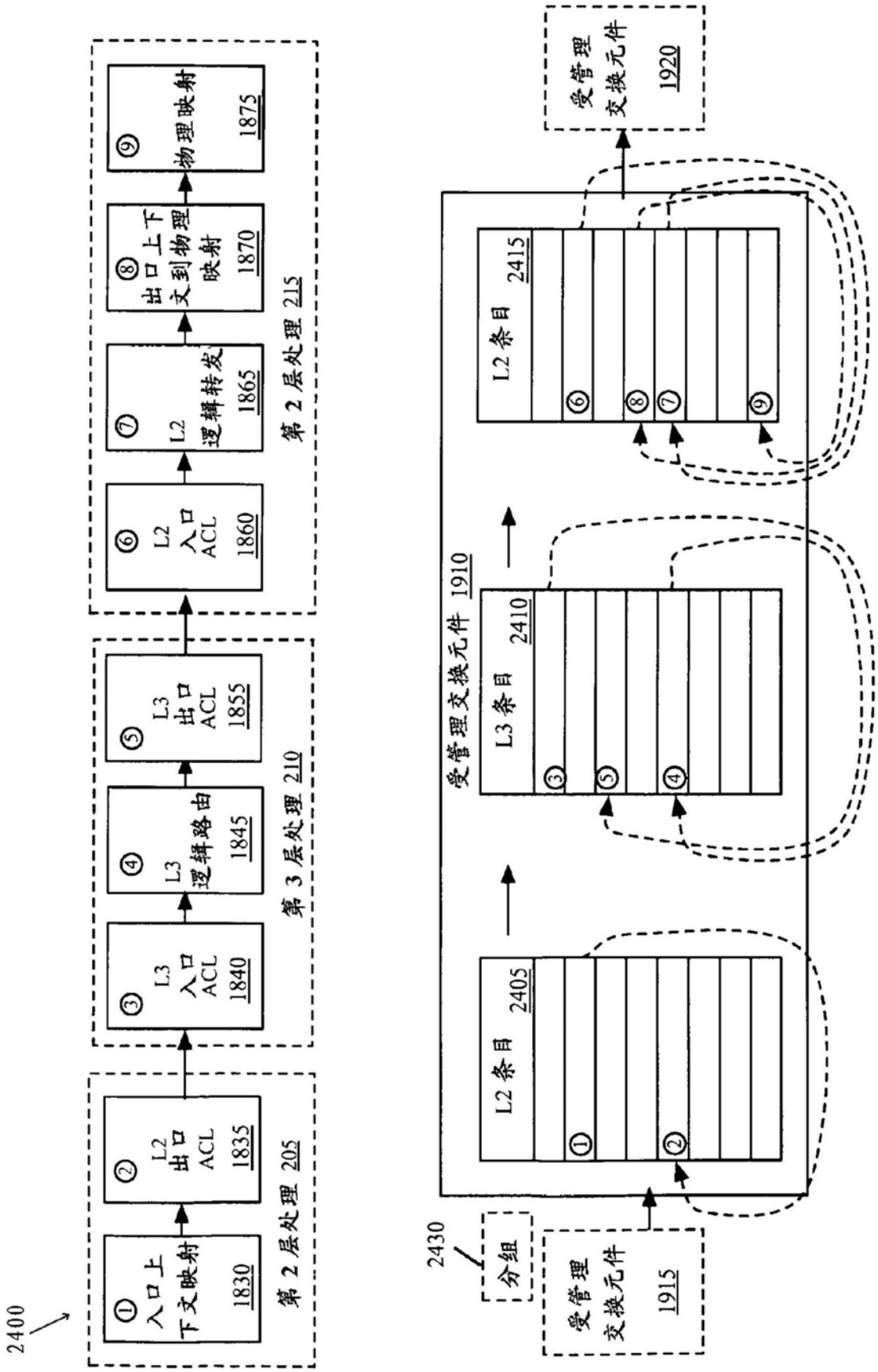

2400
① 入口上下文映射 1830
② L2 出口 ACL 1835
第 2 层处理 205
③ L3 入口 ACL 1840
④ L3 逻辑路由 1845
⑤ L3 出口 ACL 1855
第 3 层处理 210
⑥ L2 入口 ACL 1860
⑦ L2 逻辑转发 1865
⑧ 出口上下文到物理映射 1870
⑨ 物理映射 1875
第 2 层处理 215
2430
分组
受管理交换元件 1915
受管理交换元件 1910
L2 条目 2405
L3 条目 2410
L2 条目 2415
受管理交换元件 1920

图24

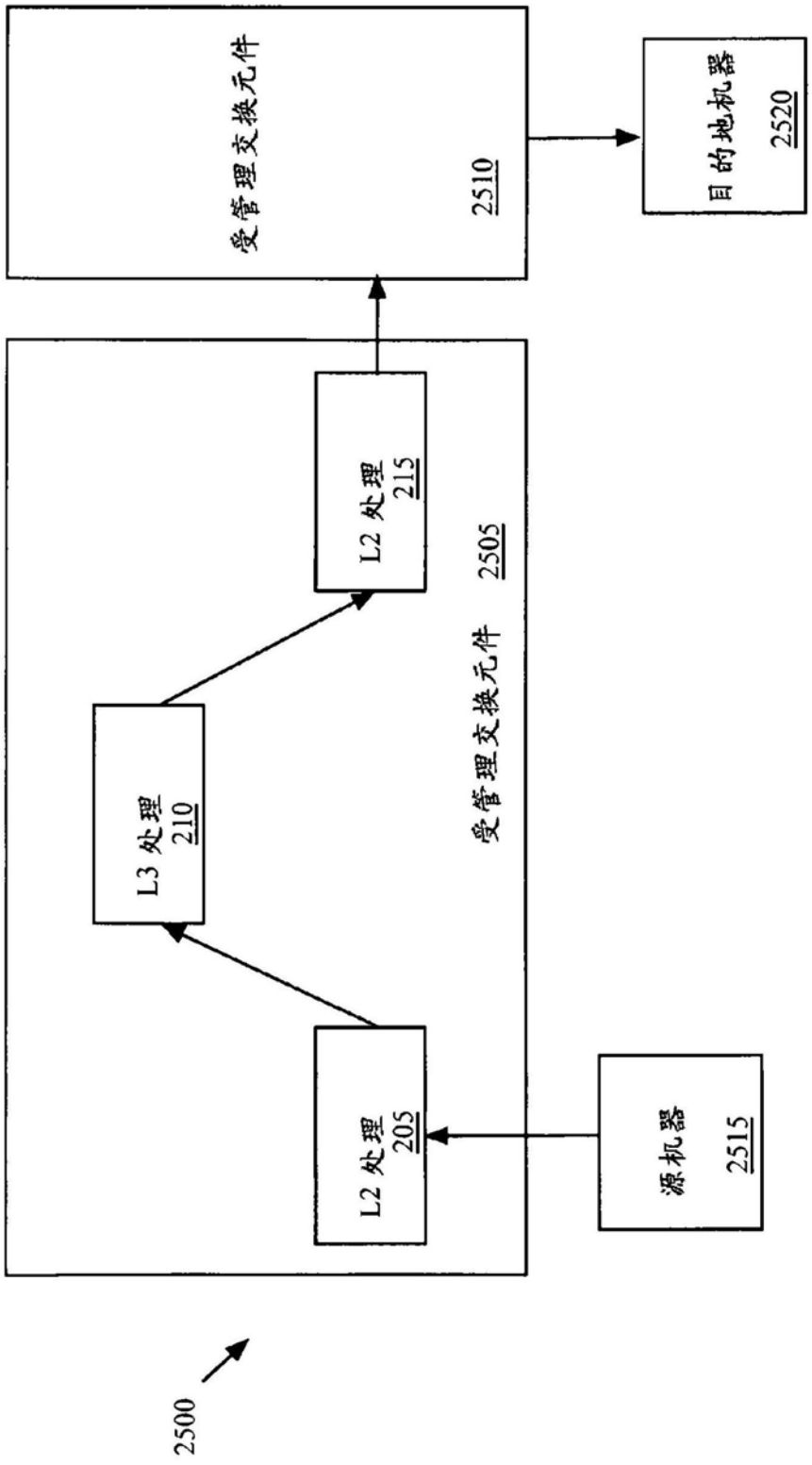

受管理交换元件
2510
目的地机器
2520
L2 处理
215
受管理交换元件
2505
L3 处理
210
L2 处理
205
源机器
2515
2500

图25

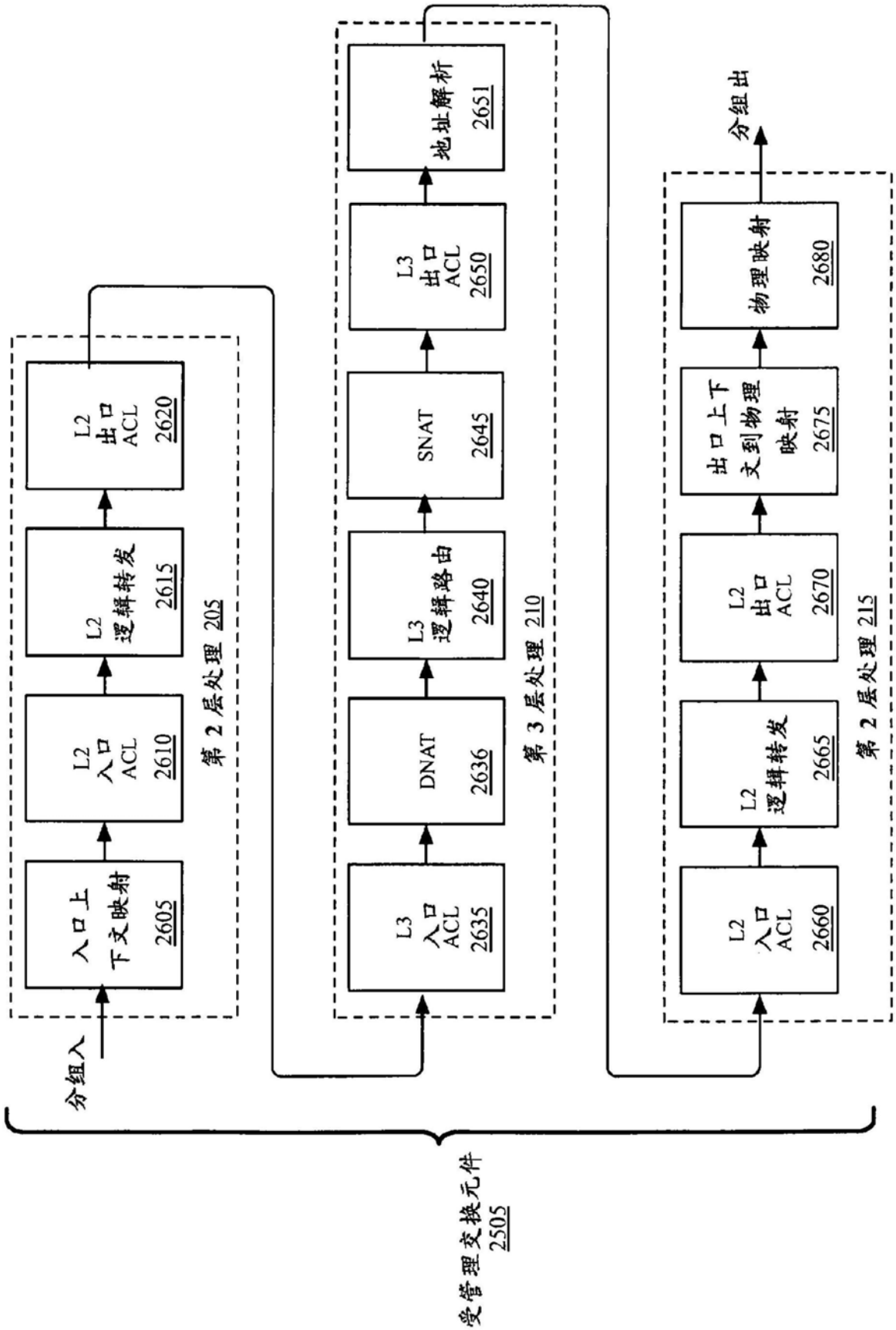

分组入
入口上下文映射 2605
L2 入口 ACL 2610
L2 逻辑转发 2615
L2 出口 ACL 2620
第 2 层处理 205
L3 入口 ACL 2635
DNAT 2636
L3 逻辑路由 2640
SNAT 2645
L3 出口 ACL 2650
地址解析 2651
第 3 层处理 210
L2 入口 ACL 2660
L2 逻辑转发 2665
L2 出口 ACL 2670
出口上下文到物理映射 2675
物理映射 2680
第 2 层处理 215
分组出
受管理交换元件 2505

图26

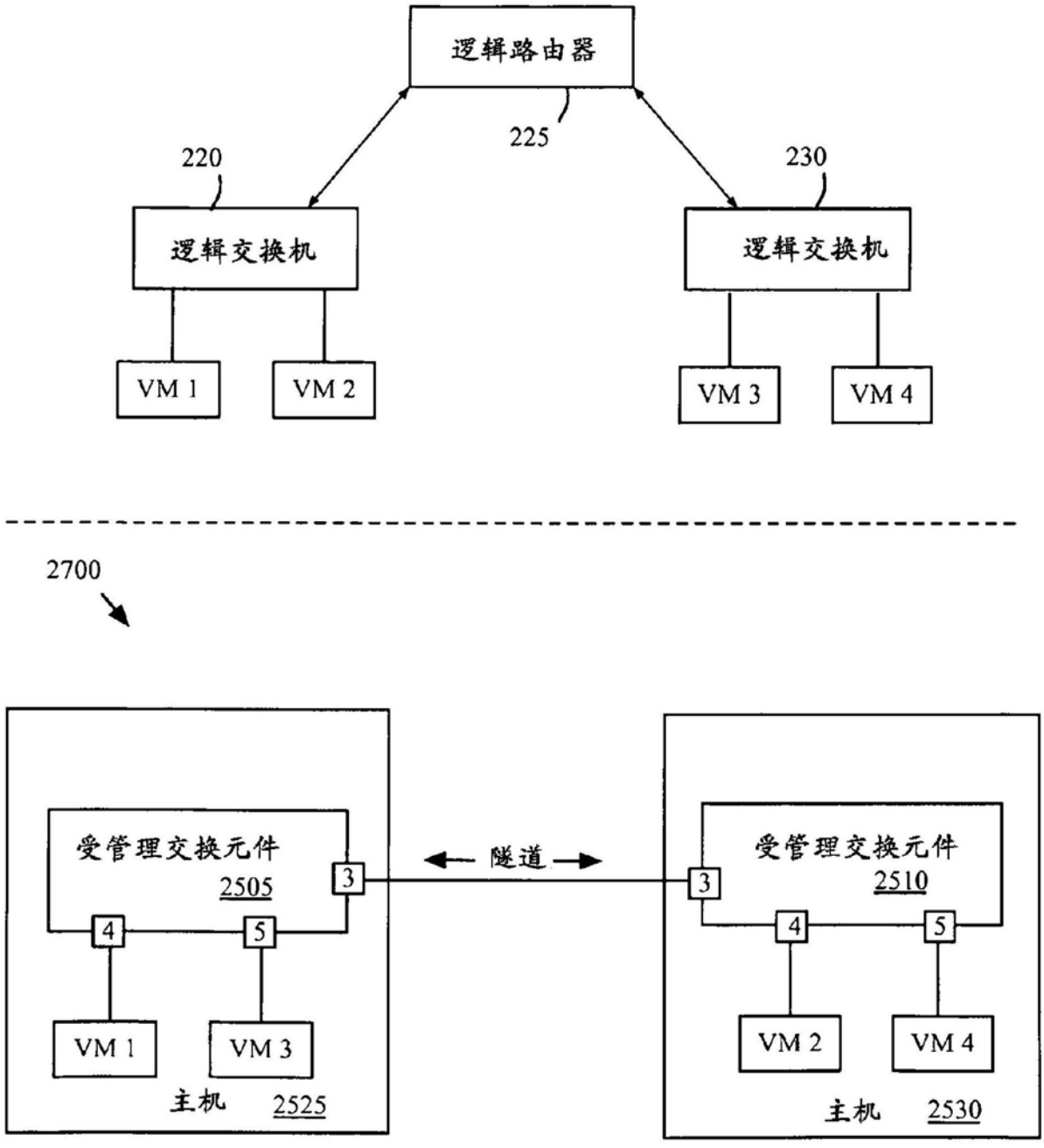

逻辑路由器
225
220
230
逻辑交换机
逻辑交换机
VM 1
VM 2
VM 3
VM 4
2700
受管理交换元件
2505
3
隧道
3
受管理交换元件
2510
4
5
4
5
VM 1
VM 3
VM 2
VM 4
主机 2525
主机 2530

图27

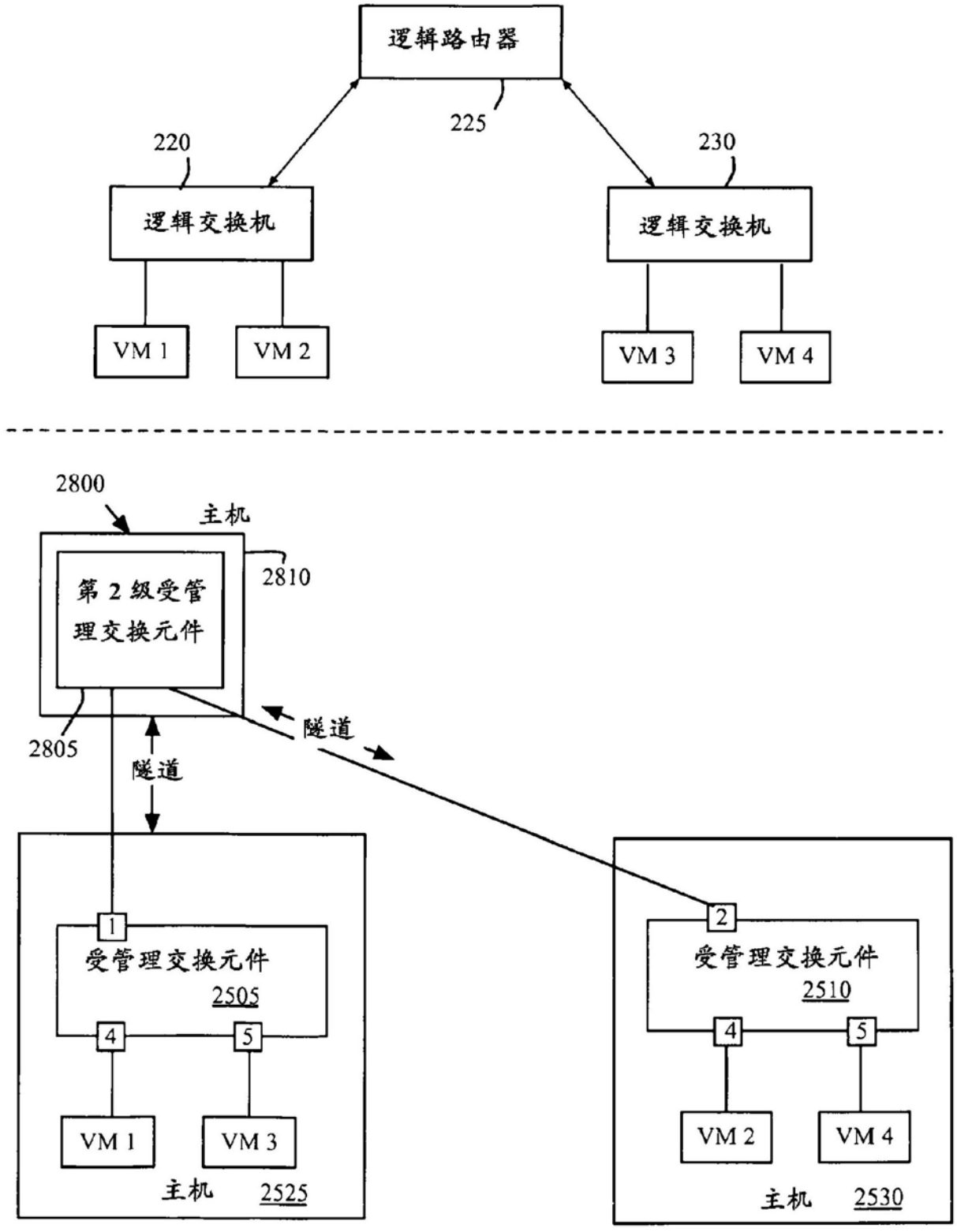

逻辑路由器
225
220
230
逻辑交换机
逻辑交换机
VM 1
VM 2
VM 3
VM 4
2800
主机
2810
第 2 级受管理交换元件
2805
隧道
隧道
1
受管理交换元件
2505
4
5
VM 1
VM 3
主机 2525
2
受管理交换元件
2510
4
5
VM 2
VM 4
主机 2530

图28

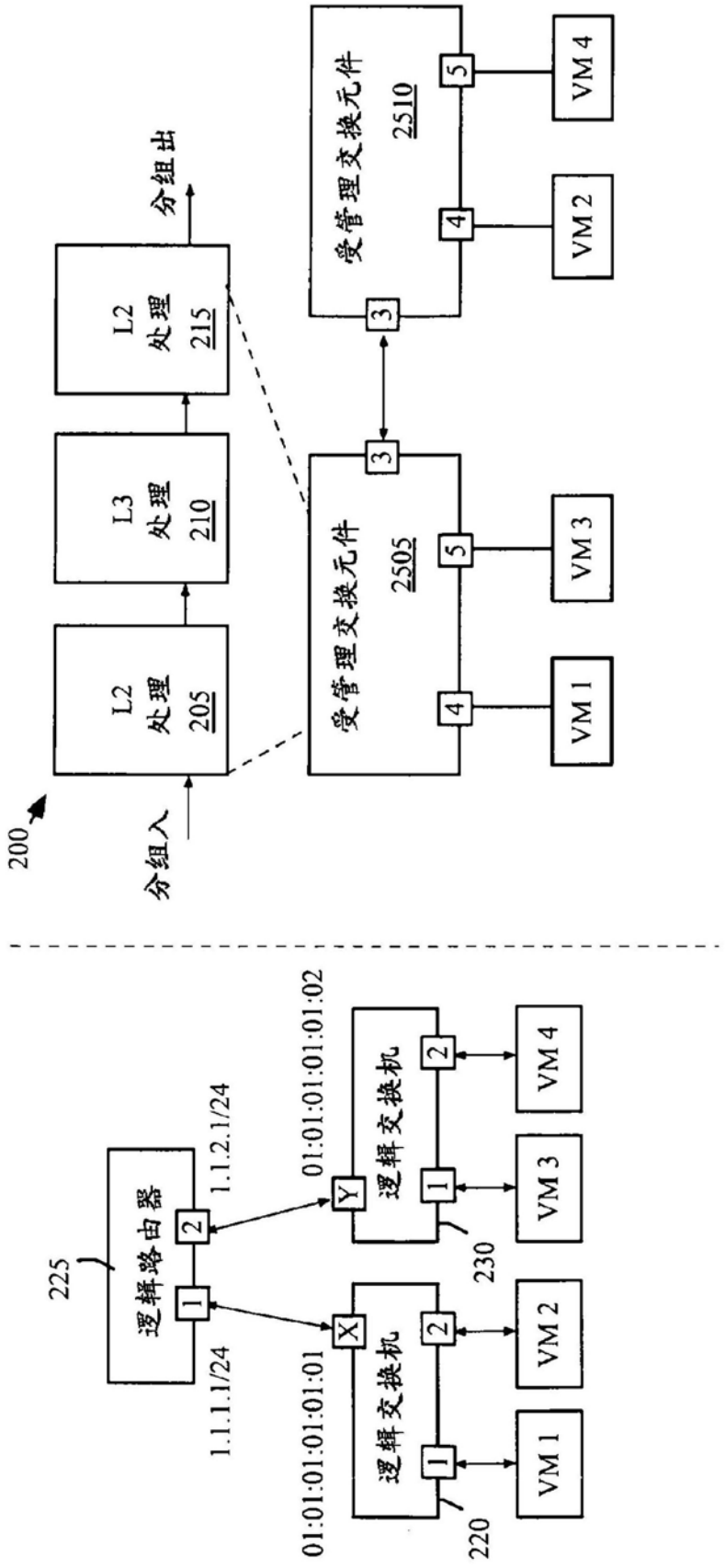

200
分组入
L2 处理 205
L3 处理 210
L2 处理 215
分组出
受管理交换元件 2505
受管理交换元件 2510
VM 1
VM 2
VM 3
VM 4
225
逻辑路由器
1.1.1.1/24
1.1.2.1/24
01:01:01:01:01:01
01:01:01:01:01:02
逻辑交换机
220
230
X
Y

图29

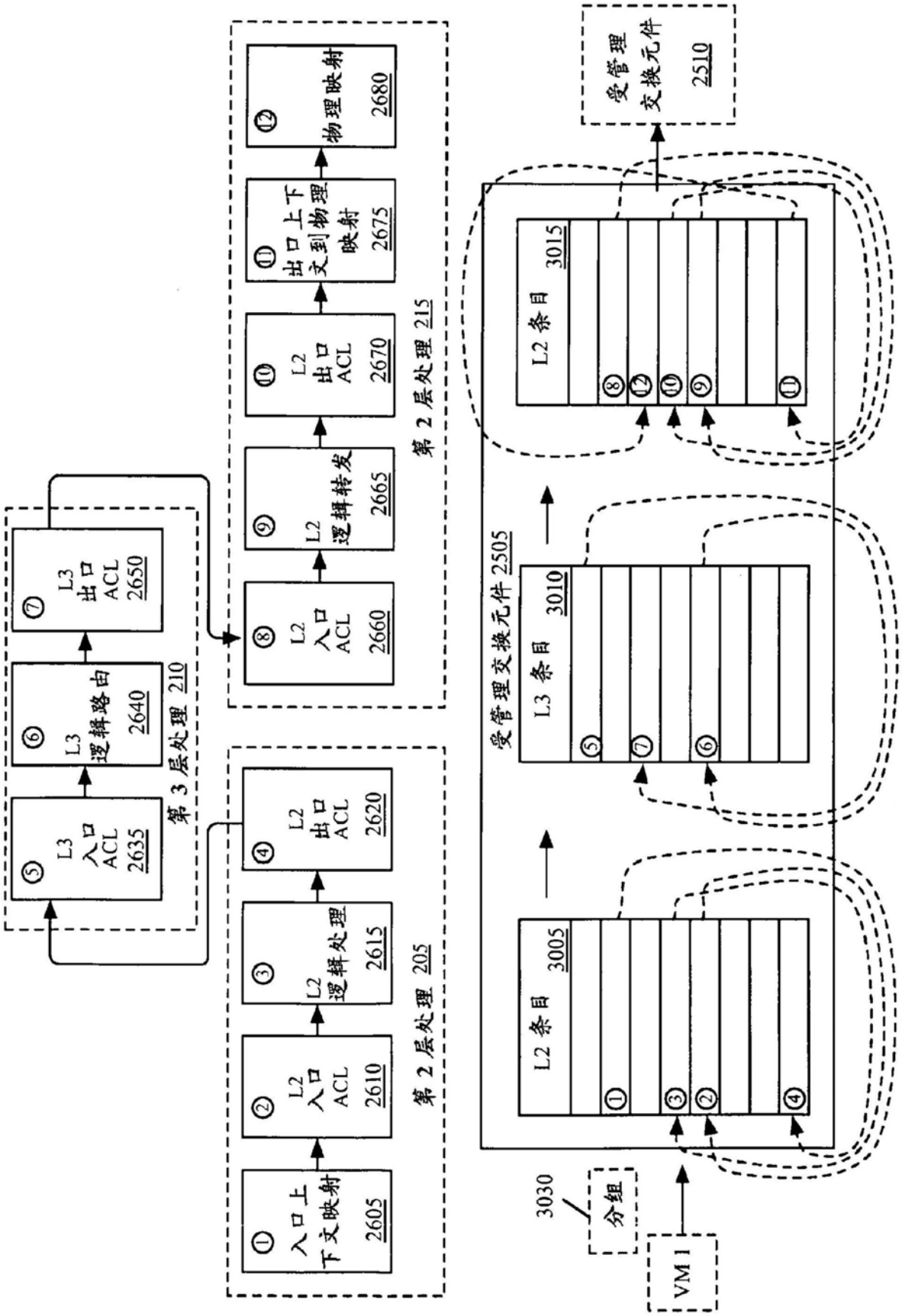

① 入口上下文映射 2605
② L2 入口 ACL 2610
③ L2 逻辑处理 2615
④ L2 出口 ACL 2620
第 2 层处理 205
⑤ L3 入口 ACL 2635
⑥ L3 逻辑路由 2640
⑦ L3 出口 ACL 2650
第 3 层处理 210
⑧ L2 入口 ACL 2660
⑨ L2 逻辑转发 2665
⑩ L2 出口 ACL 2670
⑪ 出口上下文到物理映射 2675
⑫ 物理映射 2680
第 2 层处理 215
受管理交换元件 2505
L2 条目 3005
L3 条目 3010
L2 条目 3015
受管理交换元件 2510
3030
分组
VM 1

图30A

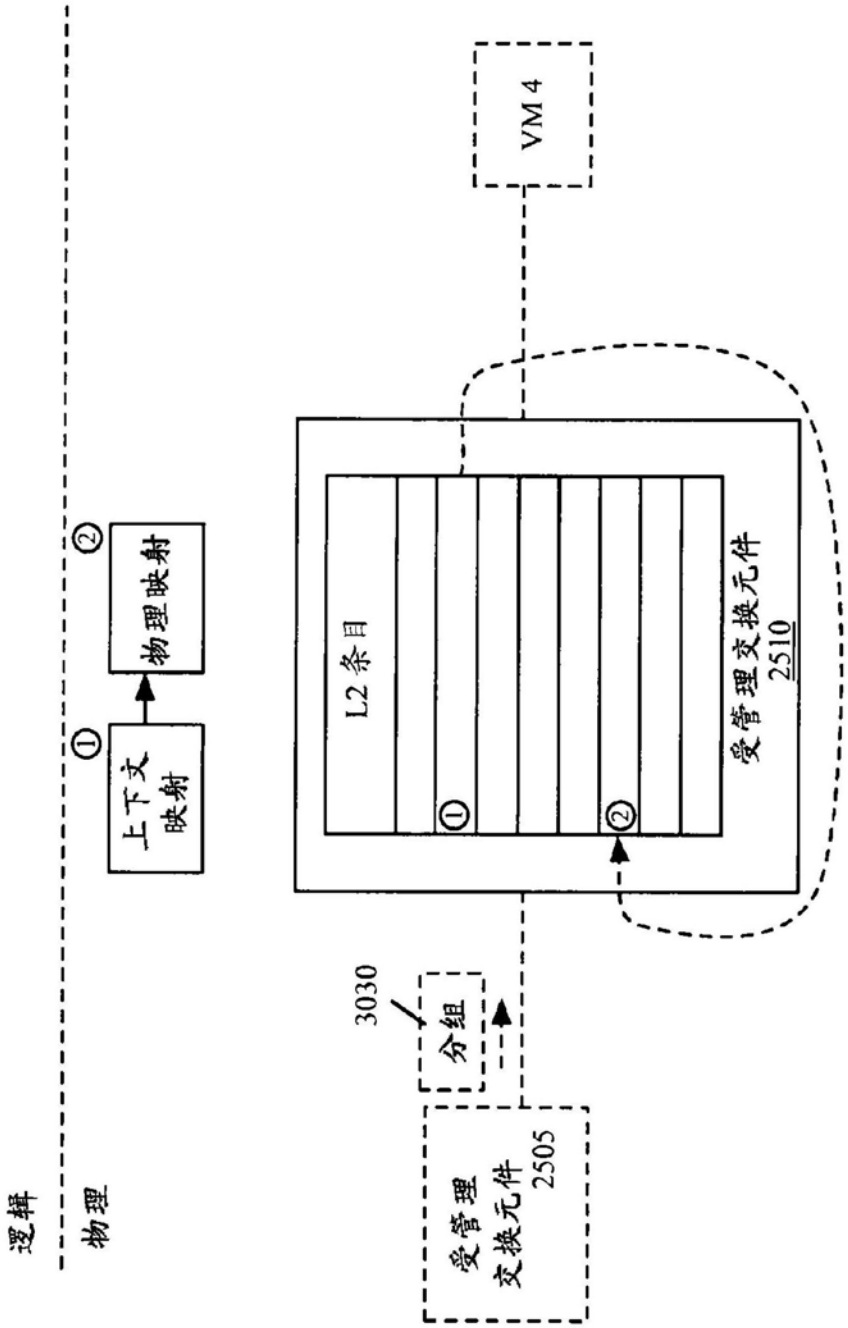

逻辑
物理
①
上下文映射
②
物理映射
3030
分组
受管理交换元件
2505
VM 4
L2 条目
①
②
受管理交换元件
2510

图30B

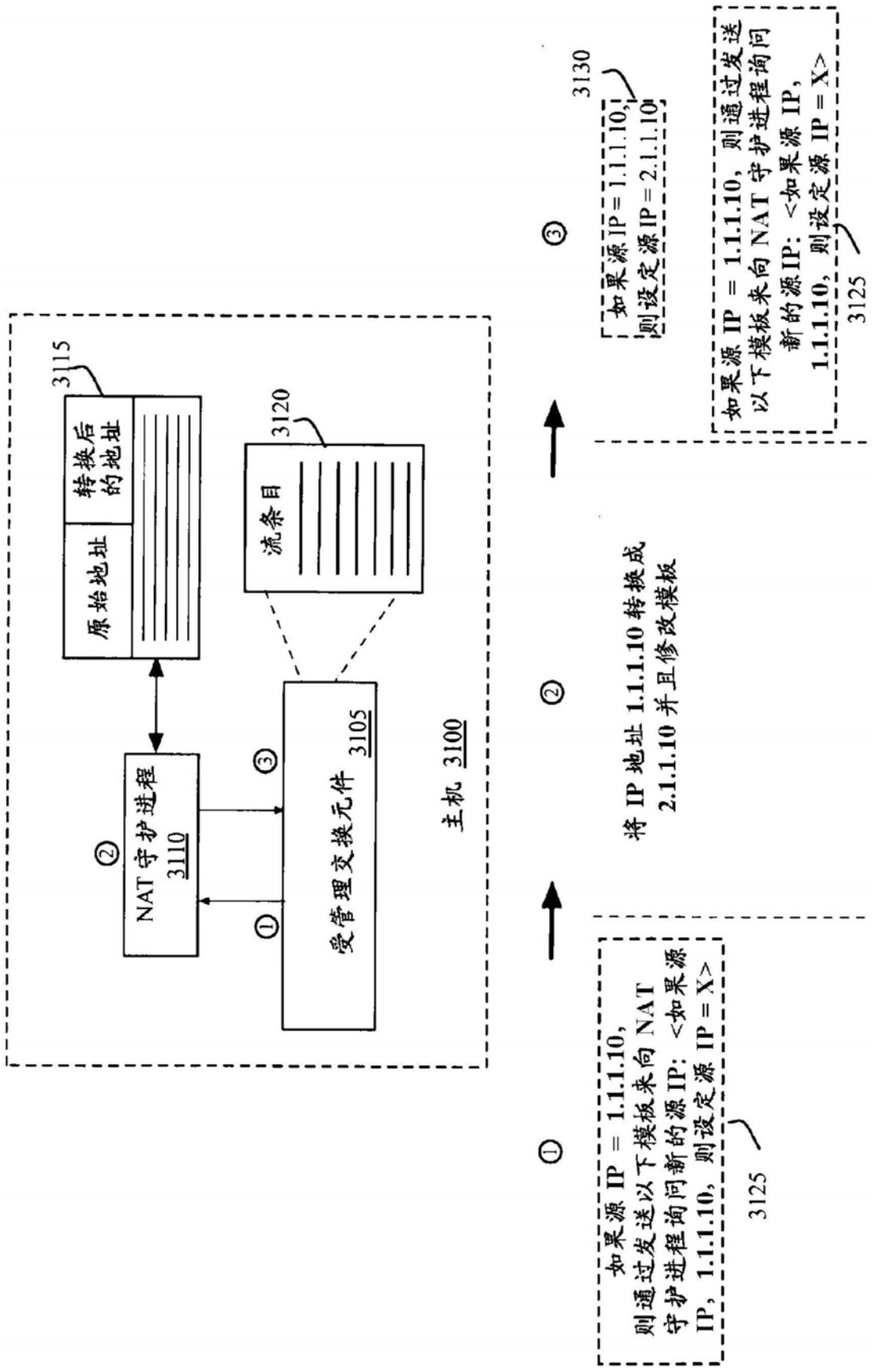

主机 3100
受管理交换元件 3105
NAT 守护进程 3110
流条目 3120
原始地址
转换后的地址
3115
①
②
③
如果源 IP = 1.1.1.10，则通过发送以下模板来向 NAT 守护进程询问 IP，1.1.1.10，则设定源 IP = X>
3125
将 IP 地址 1.1.1.10 转换成 2.1.1.10 并且修改模板
如果源IP = 1.1.1.10，则设定源IP = 2.1.1.10
3130
如果源 IP = 1.1.1.10，则通过发送以下模板来向 NAT 守护进程询问新的源IP：<如果源 IP，1.1.1.10，则设定源 IP = X>
3125

图31

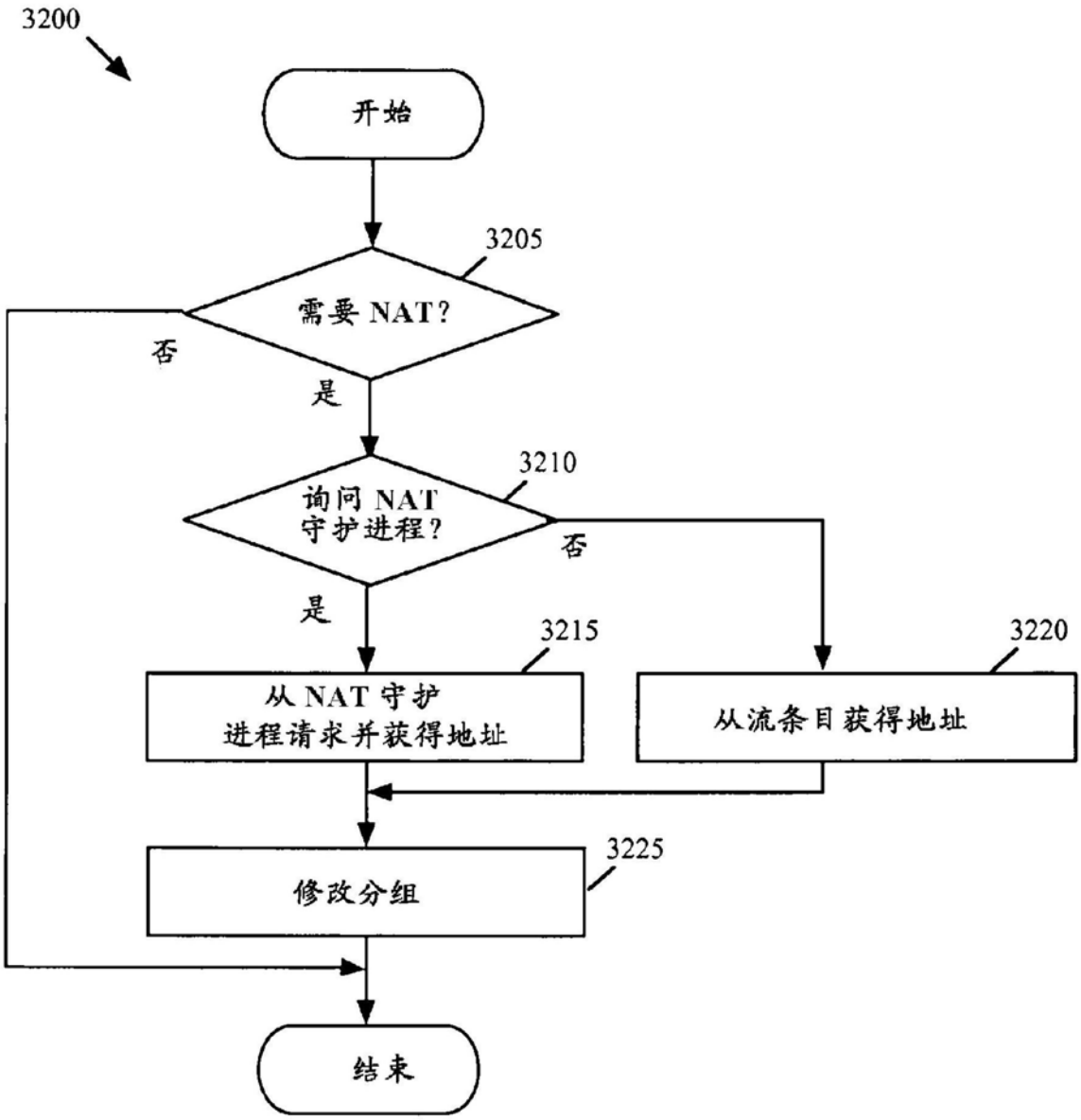

3200
开始
3205
需要 NAT?
否
是
3210
询问 NAT
守护进程?
否
是
3215
从 NAT 守护
进程请求并获得地址
3220
从流条目获得地址
3225
修改分组
结束

图32

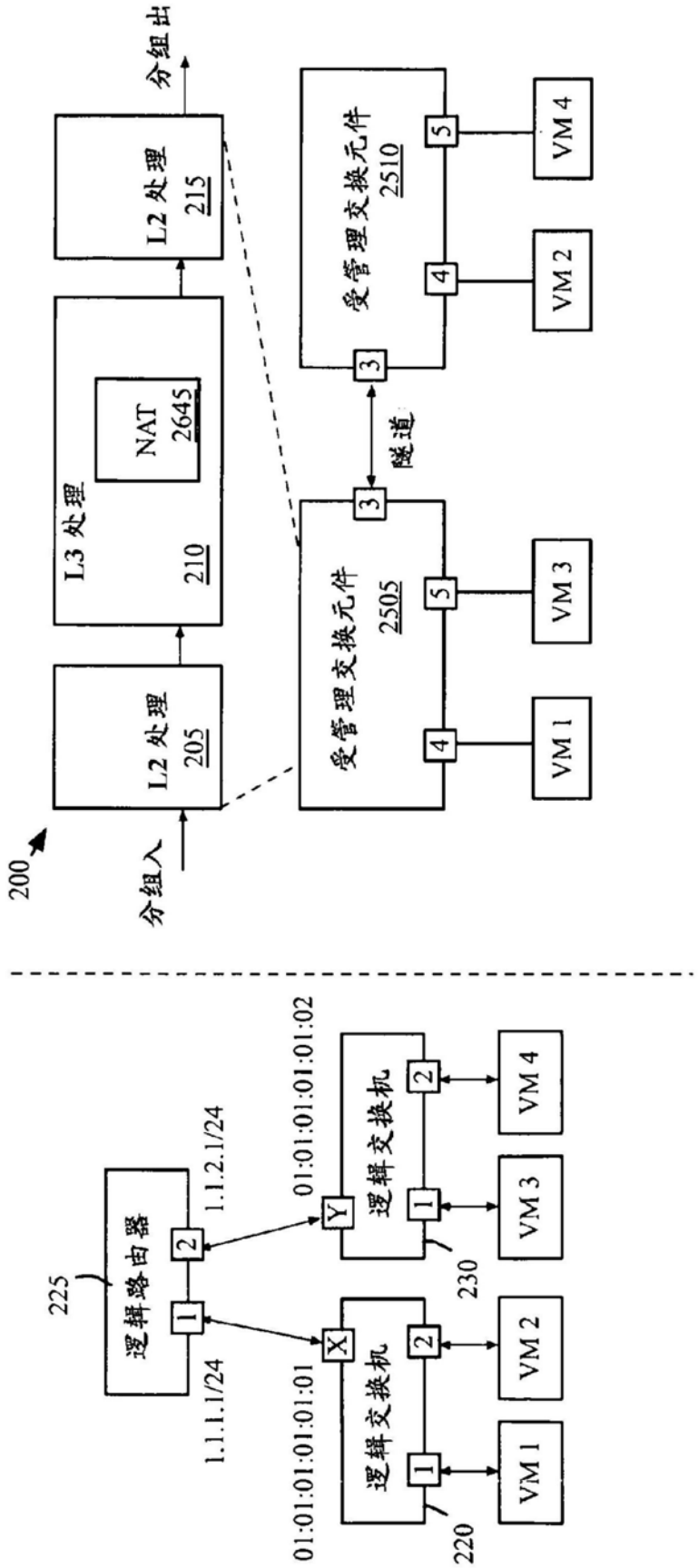

200
分组入
L2 处理
205
L3 处理
210
NAT
2645
L2 处理
215
分组出
受管理交换元件
2505
隧道
受管理交换元件
2510
VM 1
VM 2
VM 3
VM 4
225
逻辑路由器
1.1.1.1/24
1.1.2.1/24
01:01:01:01:01:01
01:01:01:01:01:02
逻辑交换机
220
逻辑交换机
230

图33

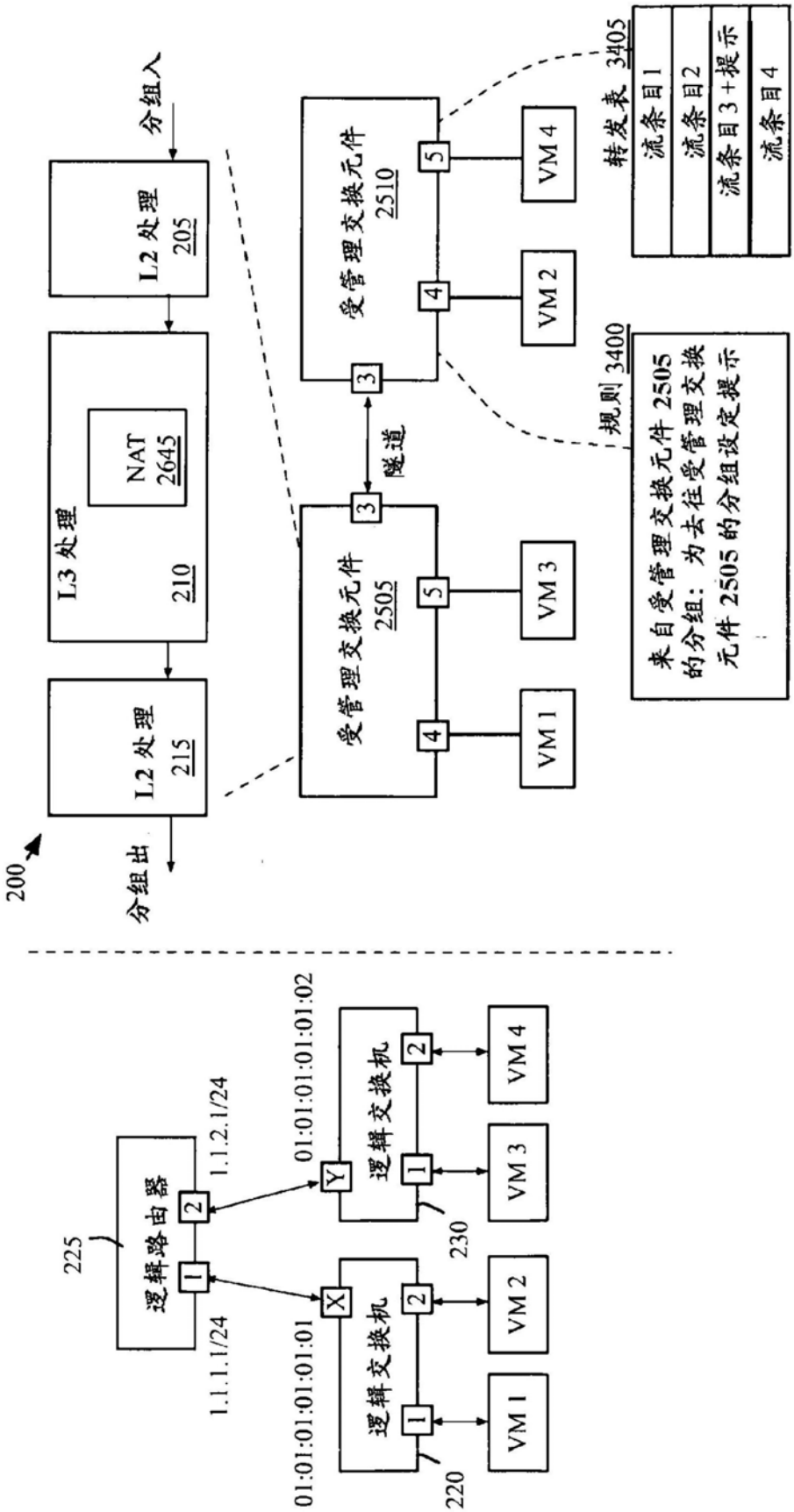

200
分组入
分组出
L2 处理
205
L3 处理
NAT
2645
210
L2 处理
215
受管理交换元件
2505
受管理交换元件
2510
隧道
VM 1
VM 2
VM 3
VM 4
规则 3400
来自受管理交换元件 2505
的分组：为去往受管理交换
元件 2505 的分组设定提示
转发表 3405
流条目 1
流条目 2
流条目 3 +提示
流条目 4
逻辑路由器
225
1.1.1.1/24
1.1.2.1/24
01:01:01:01:01:01
01:01:01:01:01:02
逻辑交换机
220
逻辑交换机
230

图34

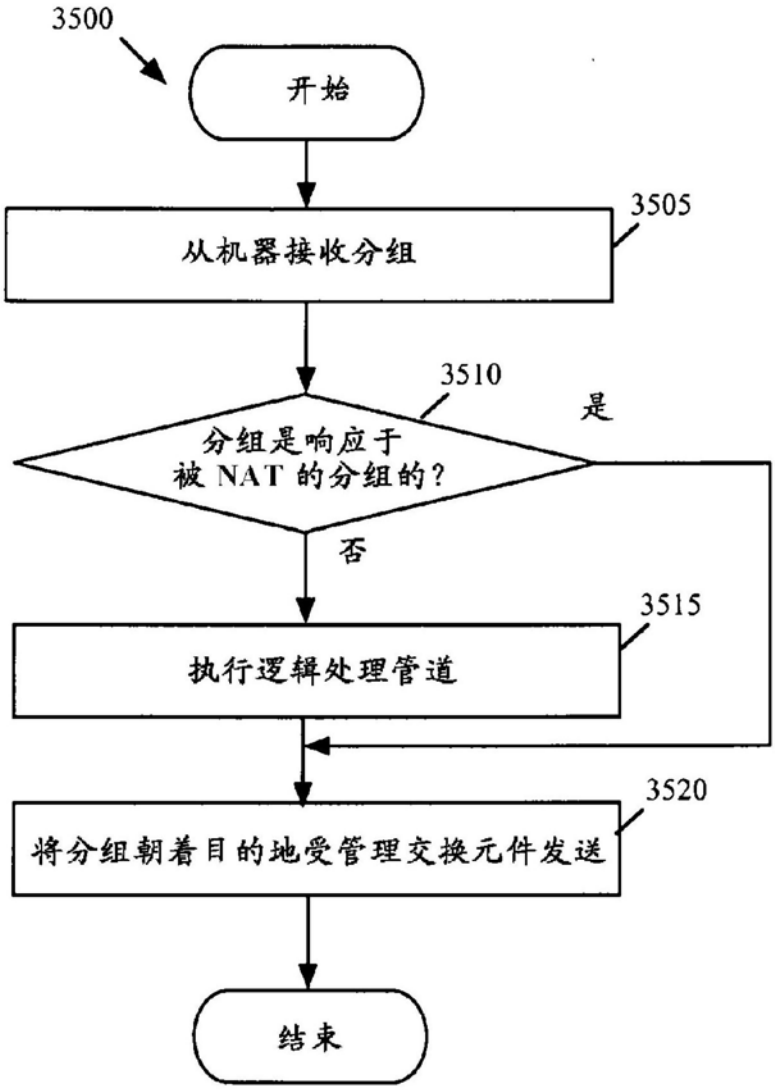

3500
开始
3505
从机器接收分组
3510
分组是响应于
被 NAT 的分组的?
是
否
3515
执行逻辑处理管道
3520
将分组朝着目的地受管理交换元件发送
结束

图35

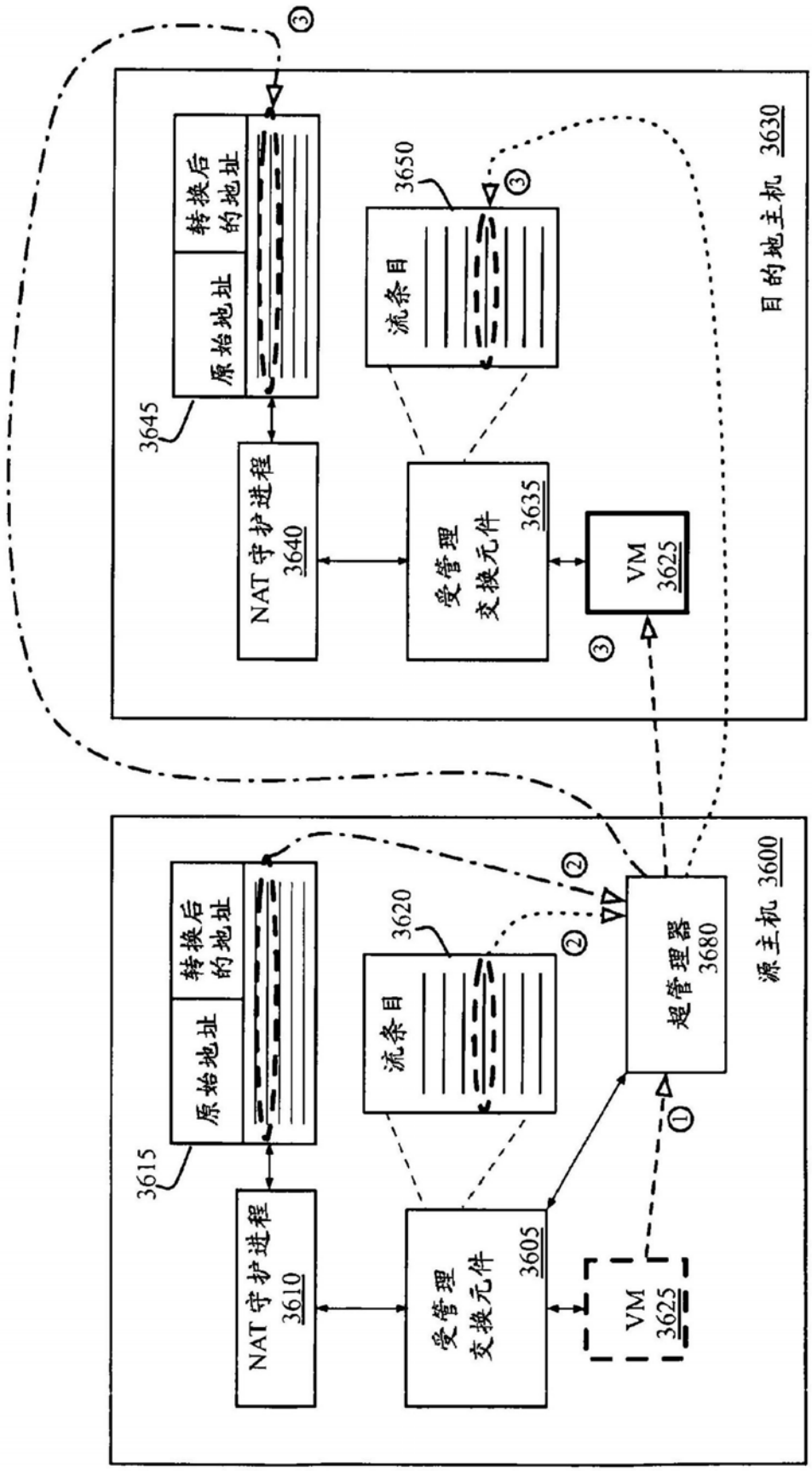

目的地主机 3630
源主机 3600
转换后的地址
原始地址
3645
3650
流条目
NAT 守护进程 3640
受管理交换元件 3635
VM 3625
3615
3620
超管理器 3680
NAT 守护进程 3610
受管理交换元件 3605
VM 3625
①
②
③

图36

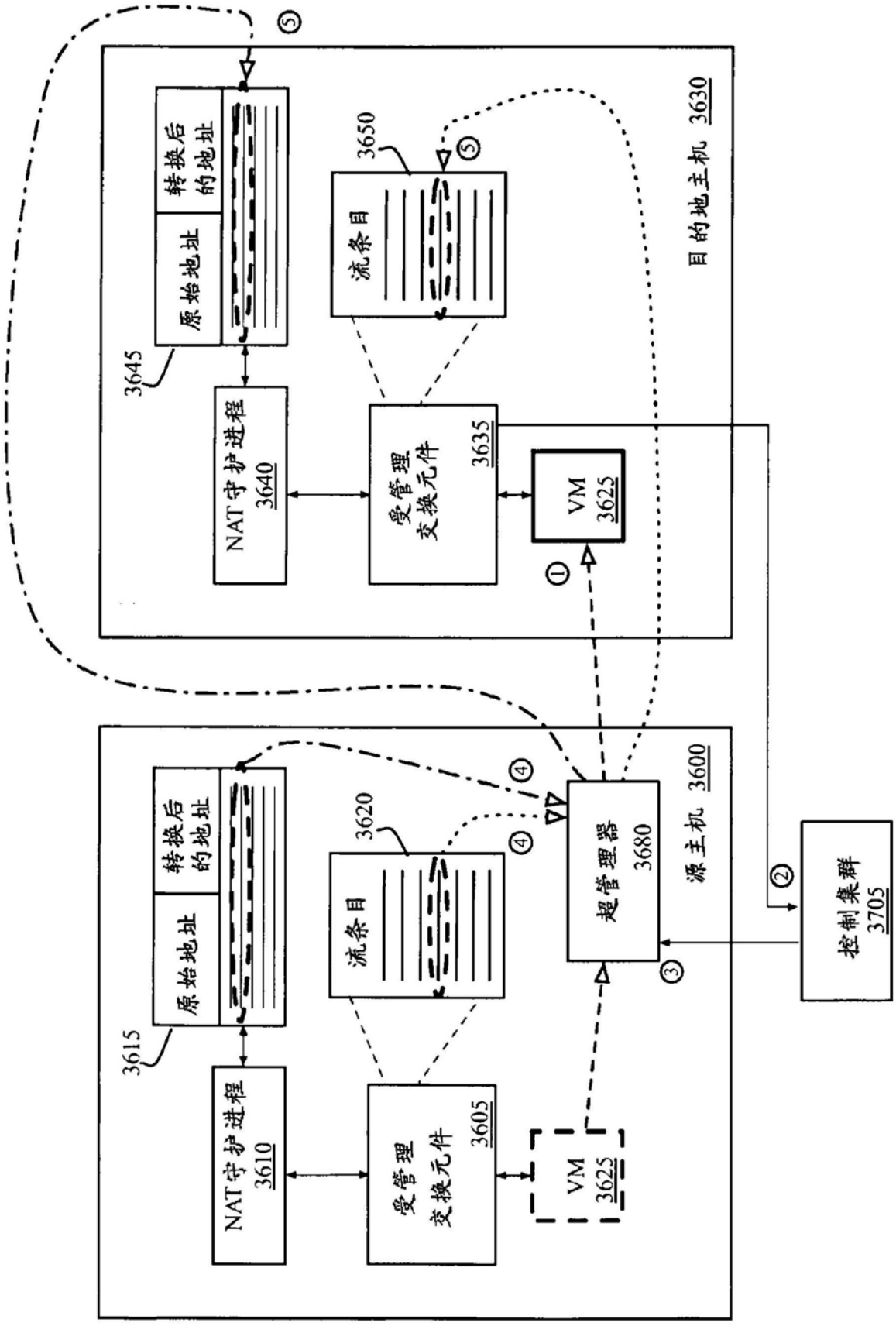

目的地主机 3630
转换后的地址
原始地址
3645
NAT守护进程 3640
流条目
3650
受管理交换元件 3635
VM 3625
源主机 3600
转换后的地址
原始地址
3615
NAT守护进程 3610
流条目
3620
受管理交换元件 3605
VM 3625
超管理器 3680
控制集群 3705
①
②
③
④
⑤

图37

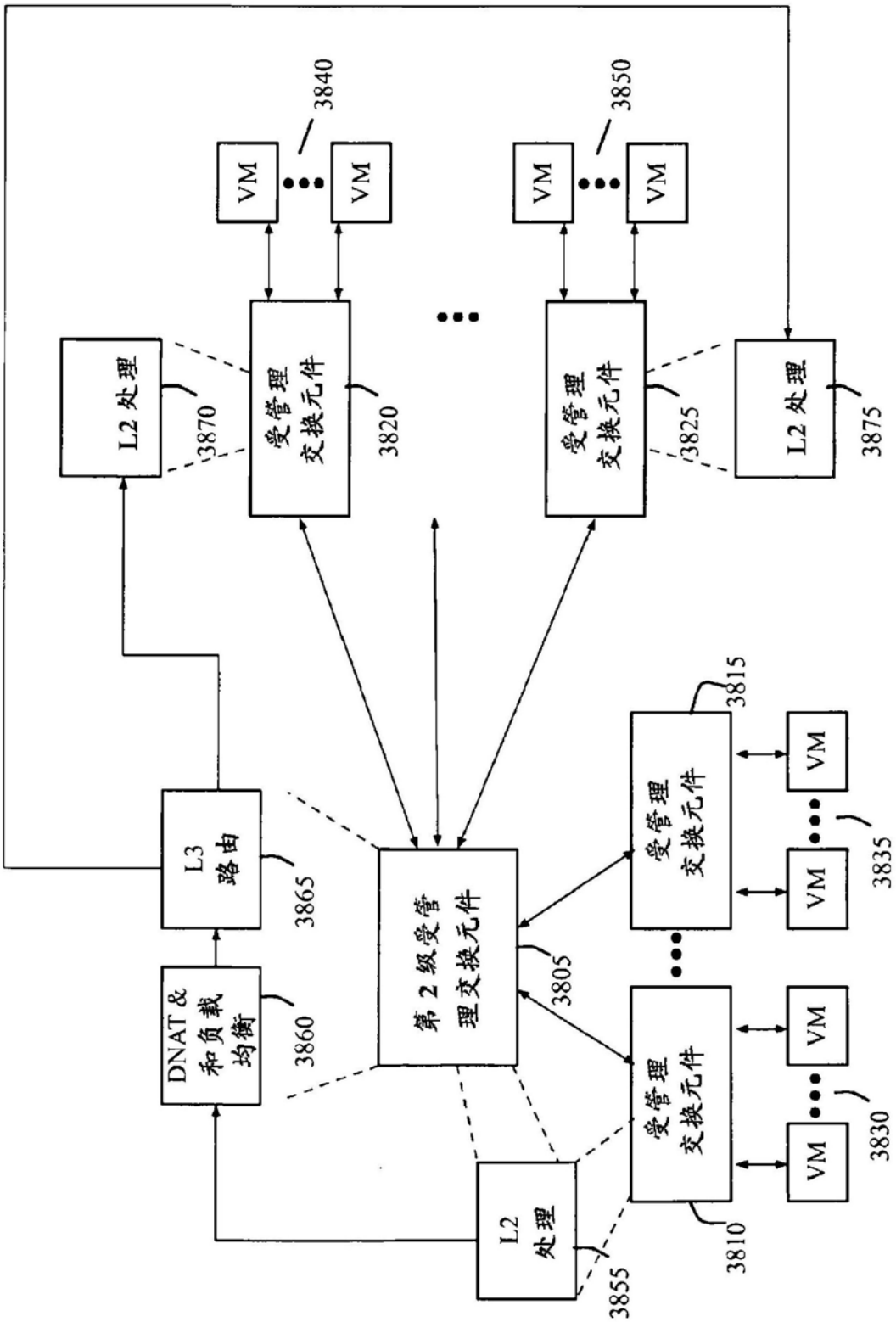

VM
VM
3840
VM
VM
3850
L2 处理
3870
受管理交换元件
3820
受管理交换元件
3825
L2 处理
3875
L3 路由
3865
第 2 级受管理交换元件
3805
受管理交换元件
3815
VM
VM
3835
DNAT & 和负载均衡
3860
受管理交换元件
VM
VM
3830
L2 处理
3855
3810

图38

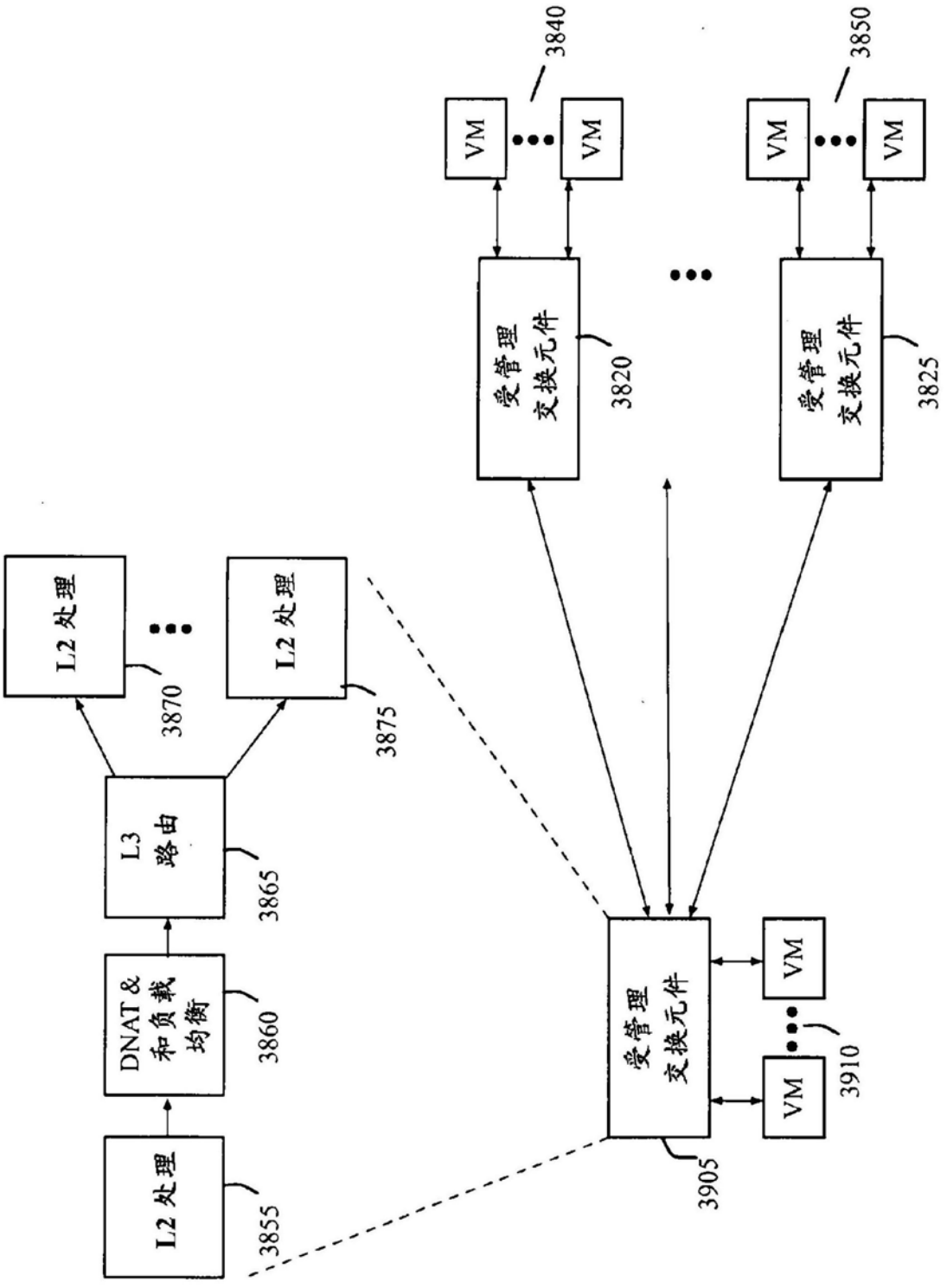

L2 处理
3855
DNAT &
和负载
均衡
3860
L3
路由
3865
L2 处理
3870
L2 处理
3875
受管理
交换元件
3905
VM
VM
3910
受管理
交换元件
3820
VM
VM
3840
受管理
交换元件
3825
VM
VM
3850

图39

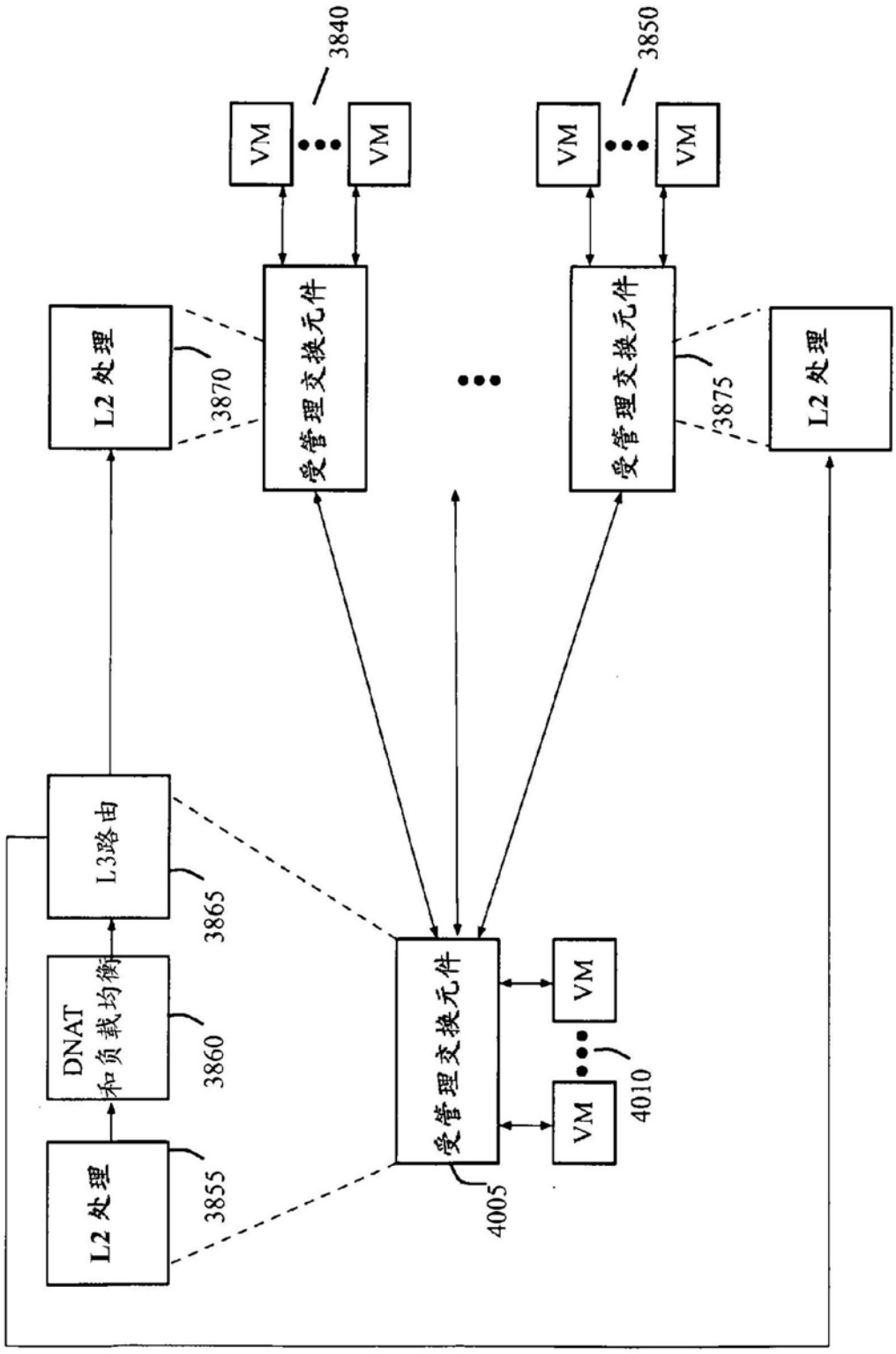

3840
3850
VM
VM
VM
VM
L2 处理
3870
受管理交换元件
受管理交换元件
3875
L2 处理
L3路由
3865
DNAT
和负载均衡
3860
受管理交换元件
VM
VM
4010
L2 处理
3855
4005

图40

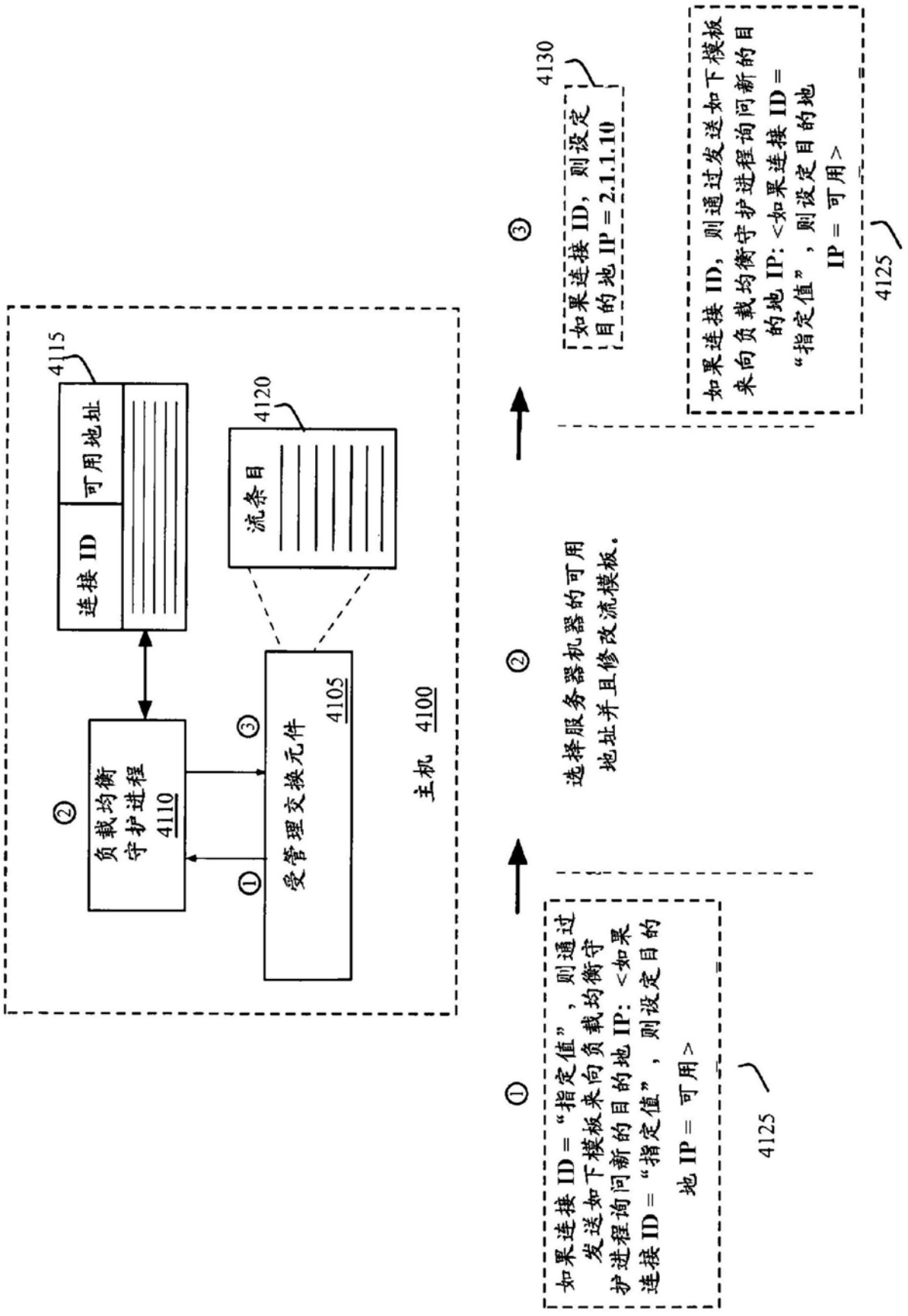

主机 4100
受管理交换元件
4105
负载均衡
守护进程
4110
连接 ID
可用地址
4115
流条目
4120
①
②
③
如果连接 ID = “指定值”，则通过
发送如下模板来向负载均衡守
护进程询问新的目的地 IP: <如果
连接 ID = “指定值”，则设定目的
地 IP = 可用>
4125
选择服务器机器的可用
地址并且修改流模板。
如果连接 ID，则设定
目的地 IP = 2.1.1.10
4130
如果连接 ID，则通过发送如下模板
来向负载均衡守护进程询问新的目
的地 IP: <如果连接 ID =
“指定值”，则设定目的地
IP = 可用>
4125

图41

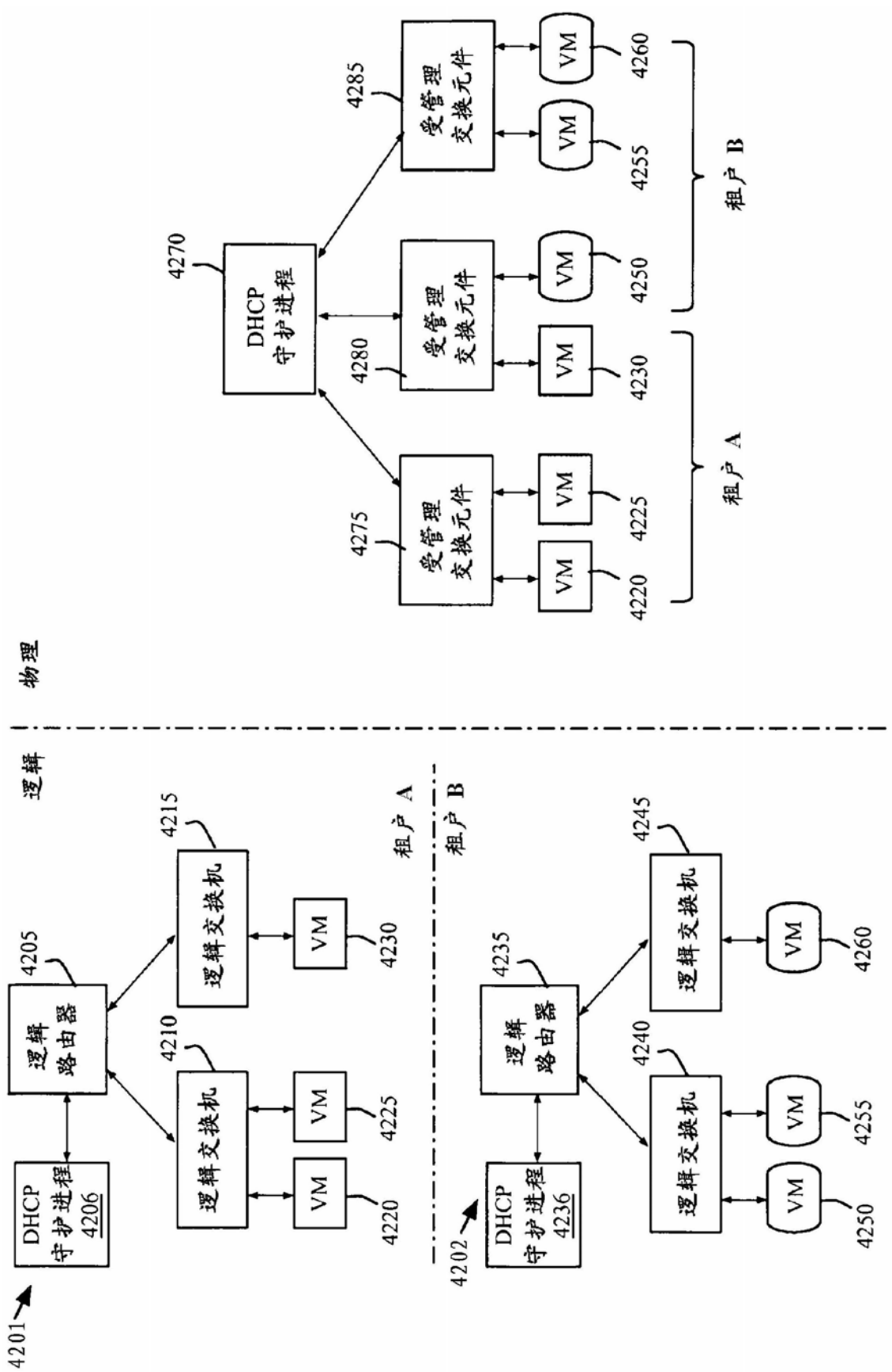

物理
逻辑
4201
DHCP 守护进程 4206
逻辑路由器 4205
逻辑交换机 4210
逻辑交换机 4215
VM 4220
VM 4225
VM 4230
租户 A
租户 B
4202
DHCP 守护进程 4236
逻辑路由器 4235
逻辑交换机 4240
逻辑交换机 4245
VM 4250
VM 4255
VM 4260
DHCP 守护进程 4270
受管理交换元件 4275
受管理交换元件 4280
受管理交换元件 4285
VM 4220
VM 4225
VM 4230
VM 4250
VM 4255
VM 4260
租户 A
租户 B

图42

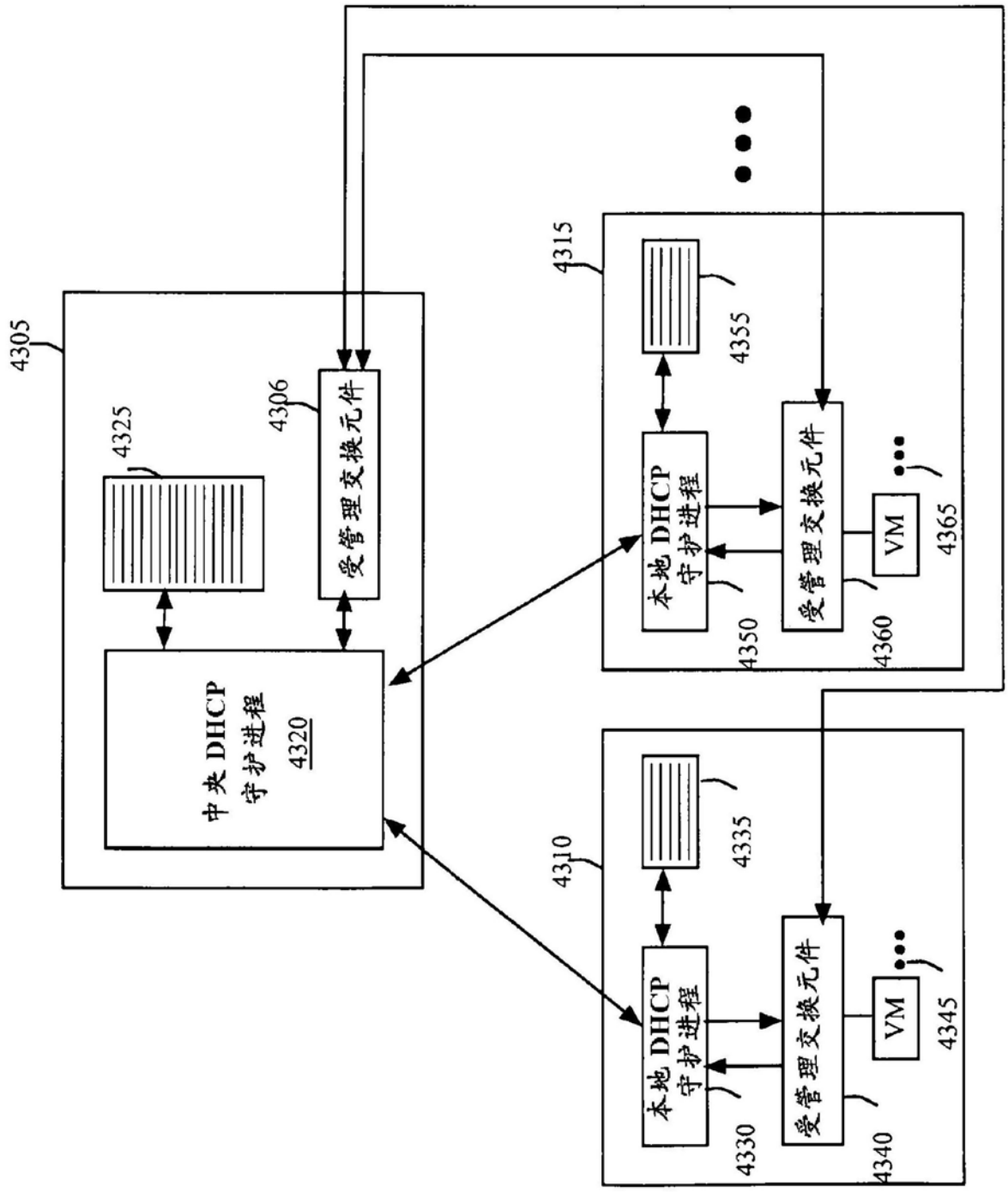

4305
4325
4306
受管理交换元件
中央 DHCP
守护进程
4320
4315
4355
本地 DHCP
守护进程
4350
受管理交换元件
VM
4365
4360
4310
4335
本地 DHCP
守护进程
4330
受管理交换元件
VM
4345
4340

图43

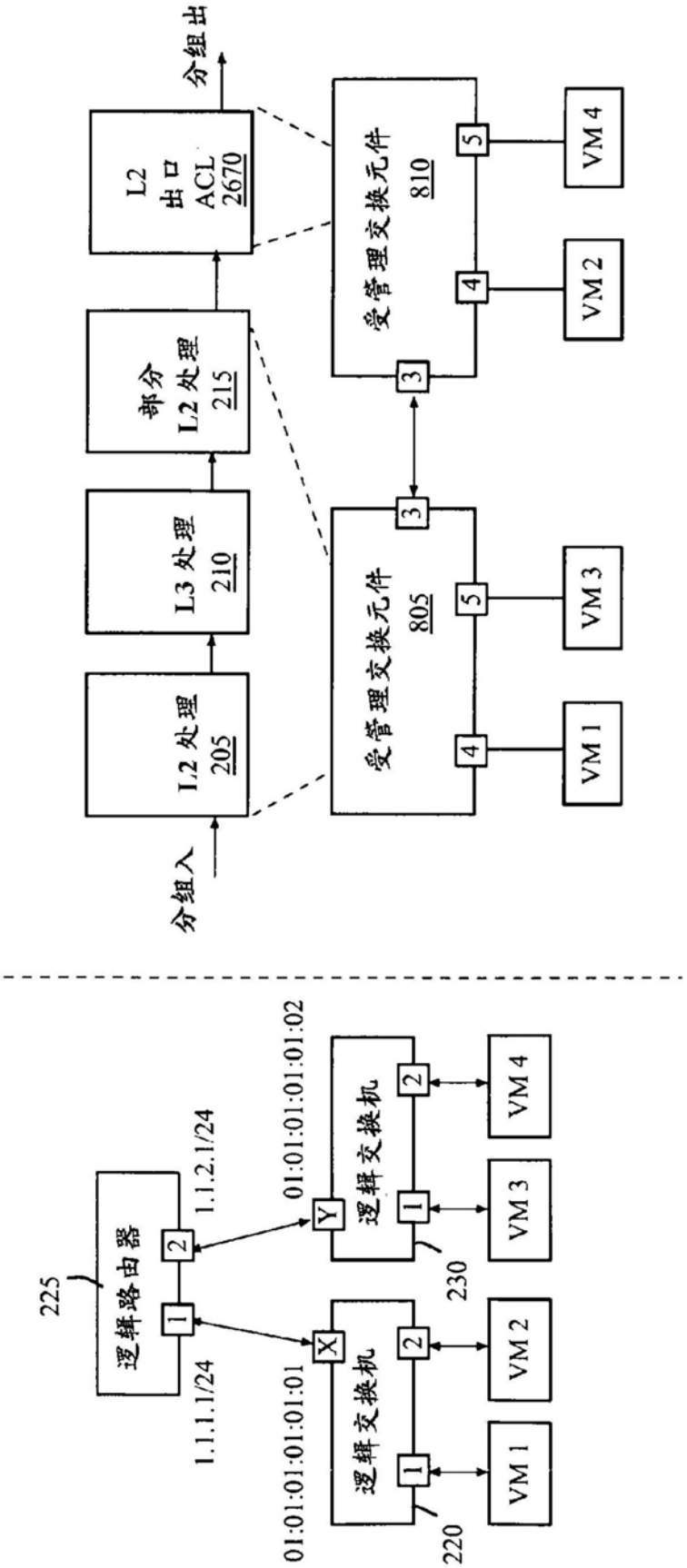

分组出
L2 出口 ACL 2670
部分 L2 处理 215
L3 处理 210
L2 处理 205
分组入
受管理交换元件 810
受管理交换元件 805
VM 4
VM 2
VM 3
VM 1
225
逻辑路由器
1.1.1.1/24
1.1.2.1/24
01:01:01:01:01:01
01:01:01:01:01:02
逻辑交换机
逻辑交换机
220
230

图44

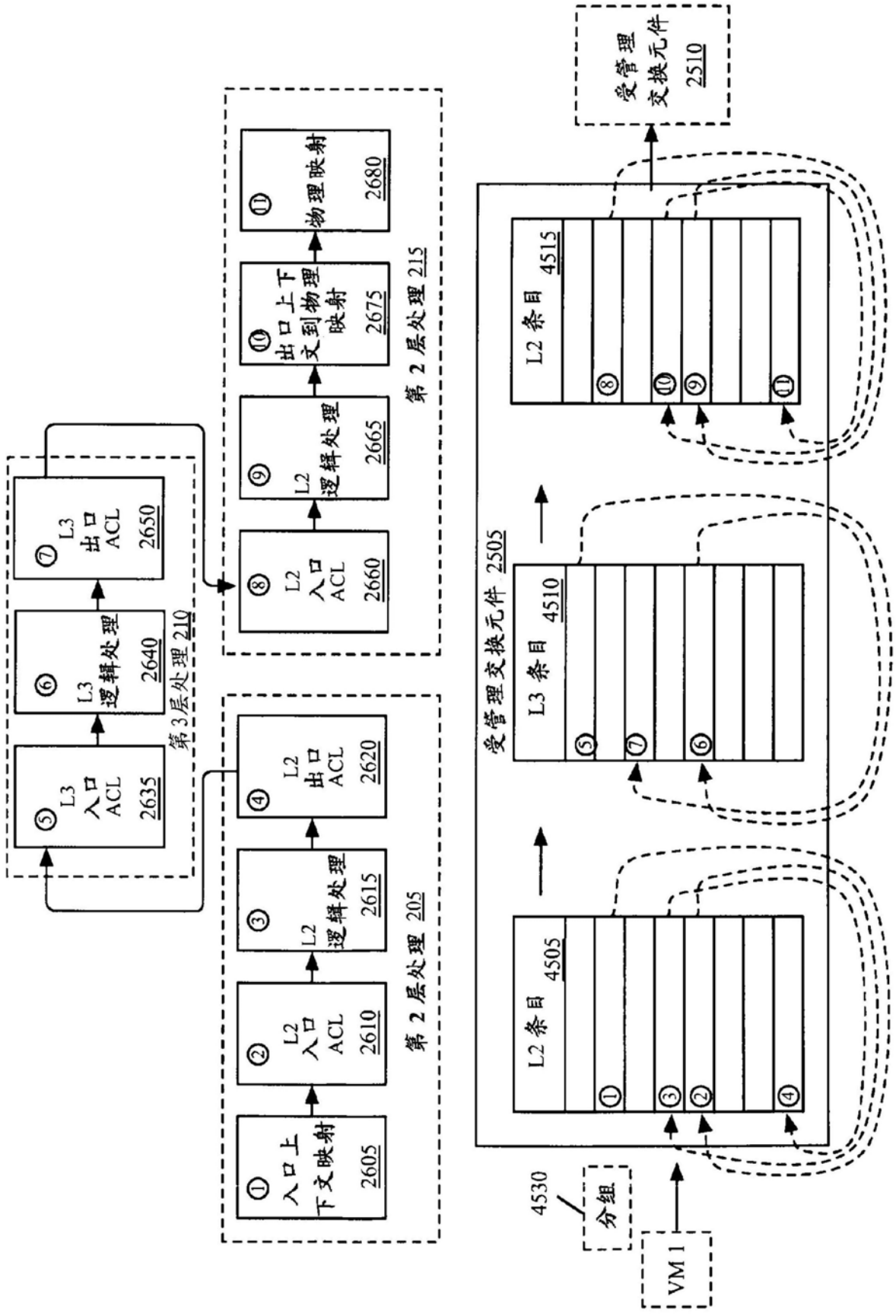

受管理交换元件 2510
L2 条目 4515
L3 条目 4510
L2 条目 4505
受管理交换元件 2505
第 2 层处理 215
① 物理映射 2680
⑩ 出口上下文到物理映射 2675
⑨ L2 逻辑处理 2665
⑧ L2 入口 ACL 2660
第 3 层处理 210
⑦ L3 出口 ACL 2650
⑥ L3 逻辑处理 2640
⑤ L3 入口 ACL 2635
第 2 层处理 205
④ L2 出口 ACL 2620
③ L2 逻辑处理 2615
② L2 入口 ACL 2610
① 入口上下文映射 2605
4530
分组
VM 1

图45A

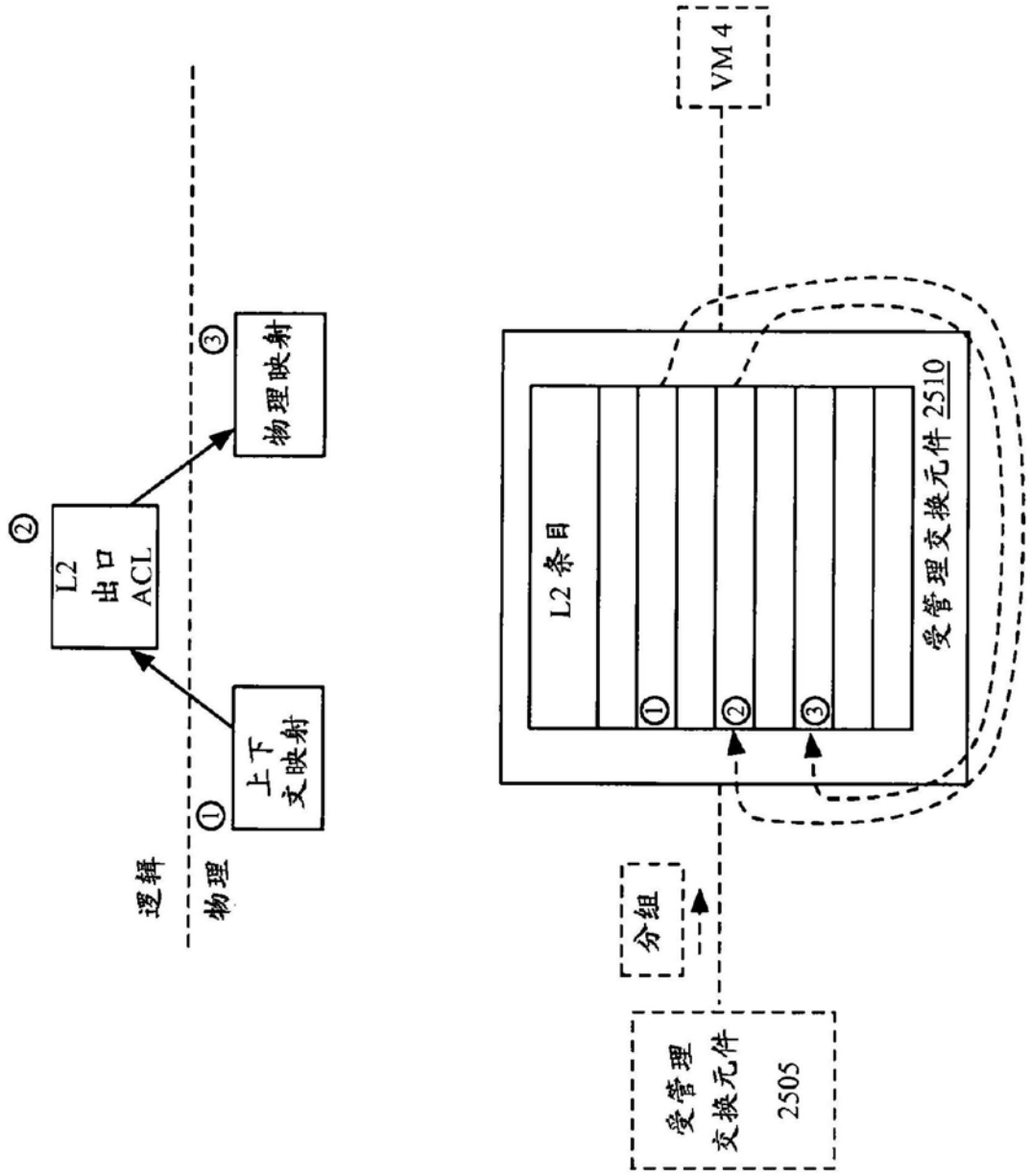

VM 4
受管理交换元件 2510
L2 条目
①
②
③
分组
受管理
交换元件
2505
逻辑
物理
上下
文映射
①
L2
出口
ACL
②
物理映射
③

图45B

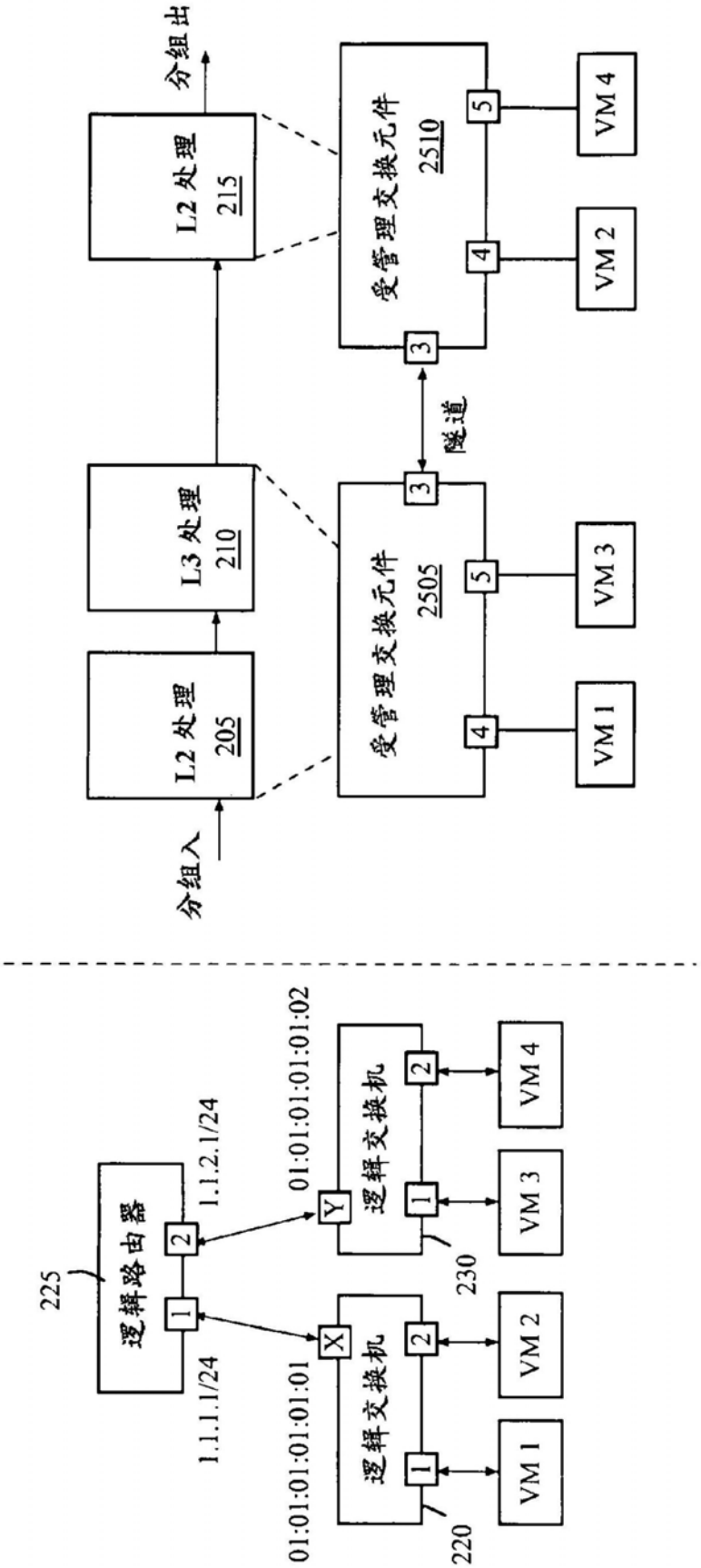

分组出
L2 处理
215
受管理交换元件
2510
VM 4
VM 2
L3 处理
210
隧道
受管理交换元件
2505
VM 3
VM 1
L2 处理
205
分组入
225
逻辑路由器
1.1.2.1/24
01:01:01:01:01:02
逻辑交换机
VM 4
VM 3
230
1.1.1.1/24
01:01:01:01:01:01
逻辑交换机
VM 2
VM 1
220

图46

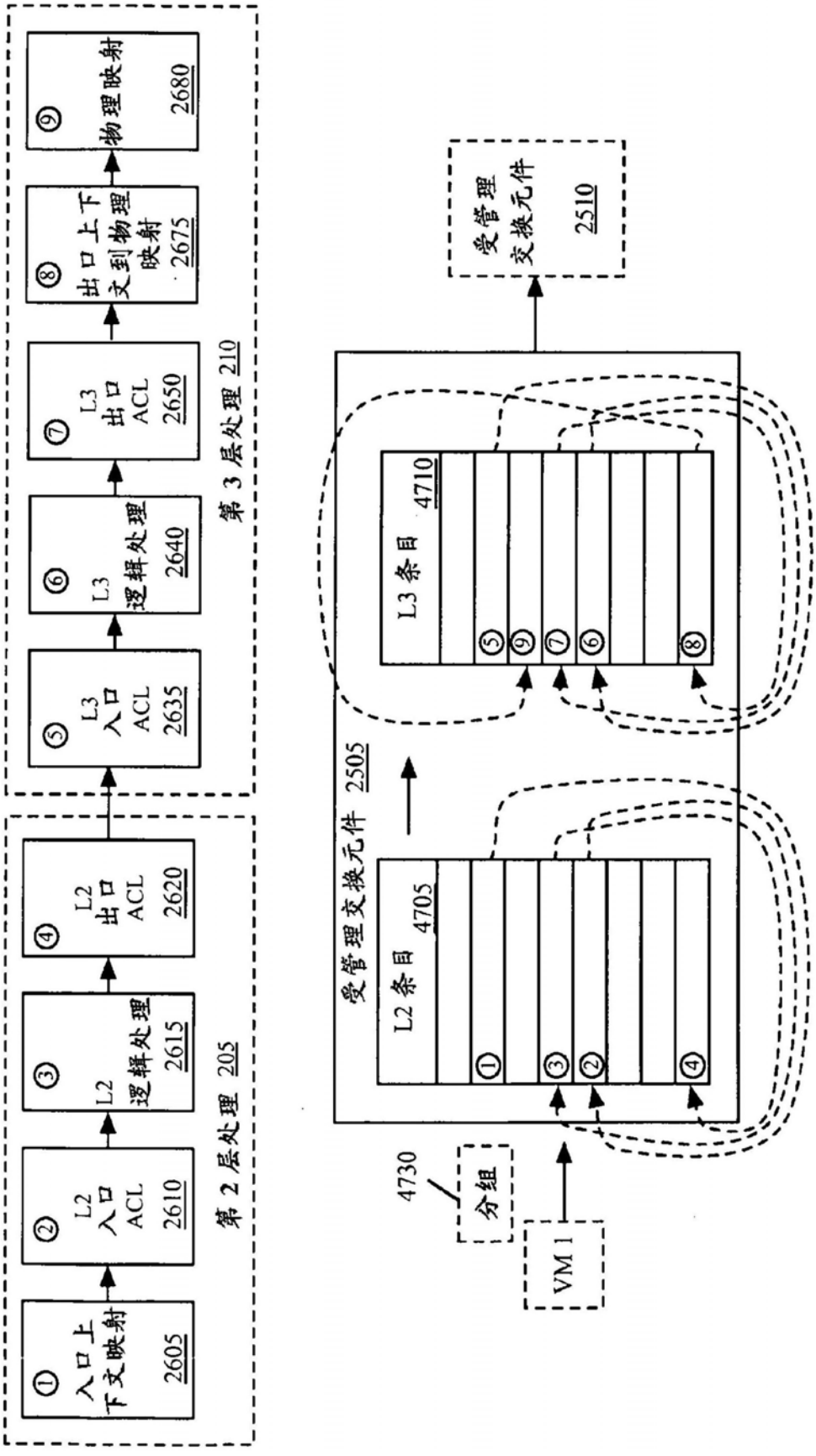

① 入口上下文映射 2605
② L2 入口 ACL 2610
③ L2 逻辑处理 2615
④ L2 出口 ACL 2620
第 2 层处理 205
⑤ L3 入口 ACL 2635
⑥ L3 逻辑处理 2640
⑦ L3 出口 ACL 2650
⑧ 出口上下文到物理映射 2675
⑨ 物理映射 2680
第 3 层处理 210
受管理交换元件 2510
受管理交换元件 2505
L3 条目 4710
L2 条目 4705
4730
分组
VM 1

图47A

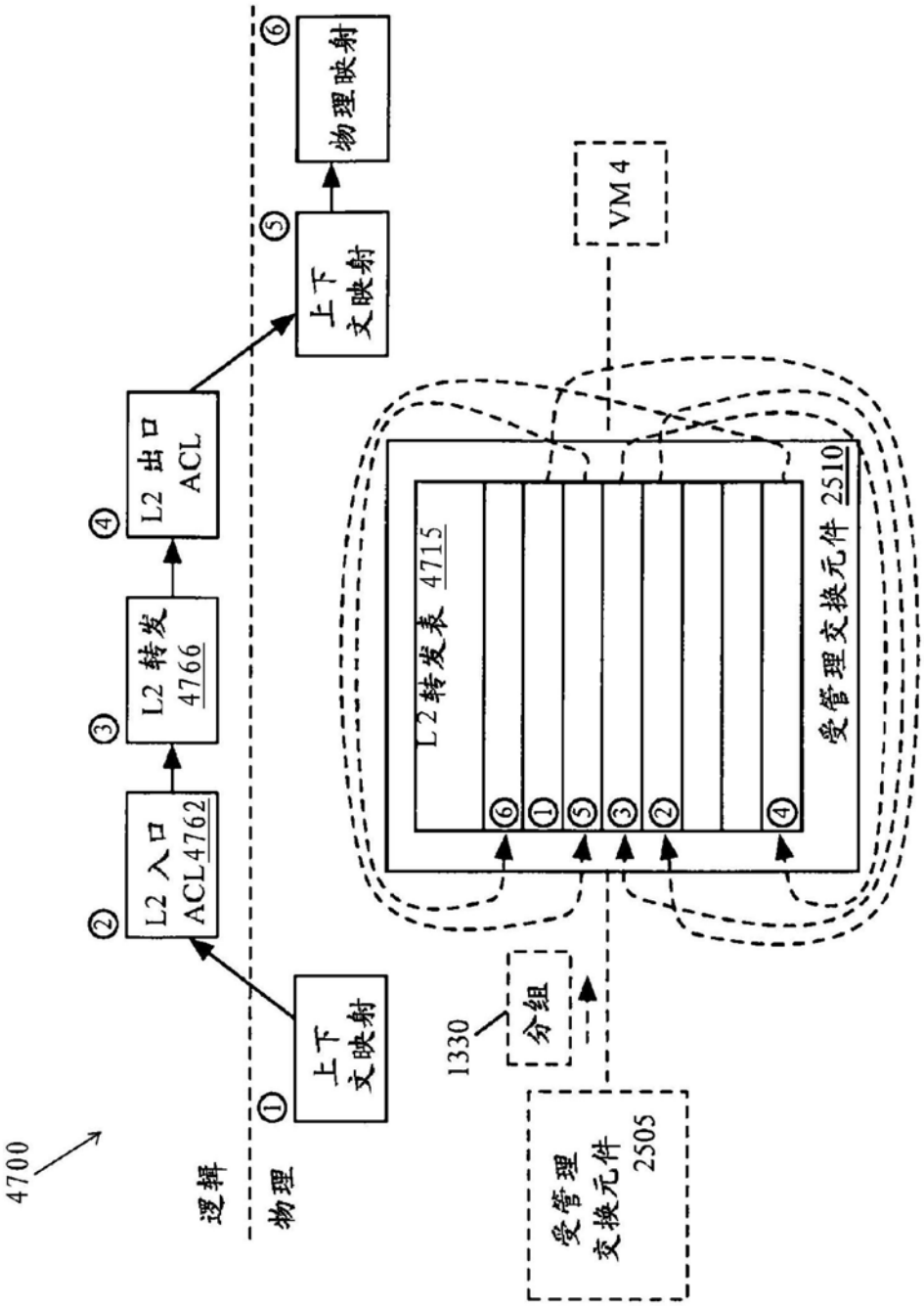

4700
逻辑
物理
① 上下文映射
② L2 入口 ACL4762
③ L2 转发 4766
④ L2 出口 ACL
⑤ 上下文映射
⑥ 物理映射
1330
分组
受管理交换元件 2505
VM 4
受管理交换元件 2510
L2 转发表 4715

图47B

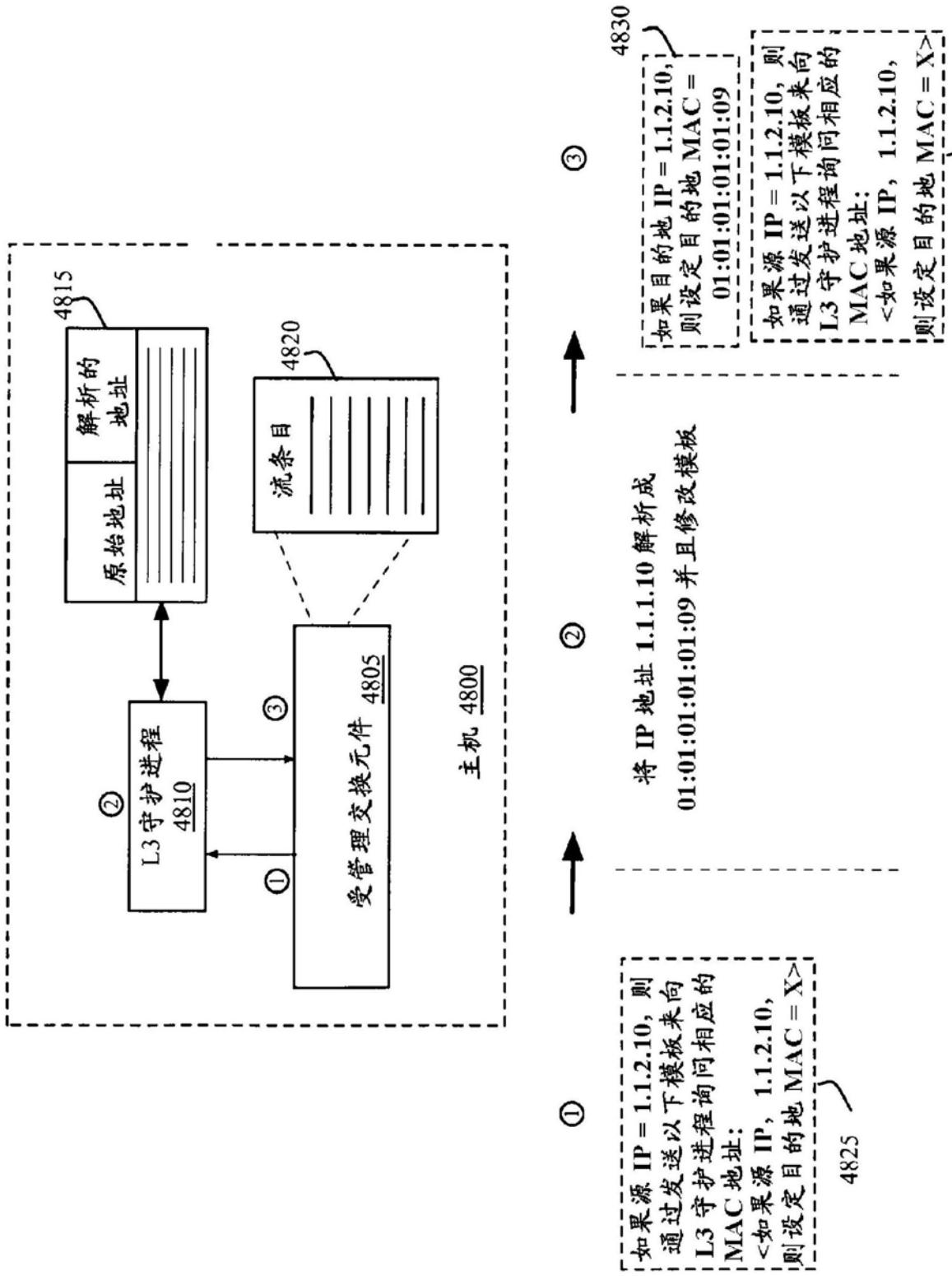

4815
解析的地址
原始地址
4820
流条目
L3守护进程 4810
受管理交换元件 4805
主机 4800
①
②
③
如果源 IP = 1.1.2.10，则通过发送以下模板来向 L3 守护进程询问相应的 MAC 地址：
<如果源 IP，1.1.2.10，则设定目的地 MAC = X>
4825
将 IP 地址 1.1.1.10 解析成 01:01:01:01:01:09 并且修改模板
如果目的地 IP = 1.1.2.10，则设定目的地 MAC = 01:01:01:01:01:09
4830
如果源 IP = 1.1.2.10，则通过发送以下模板来向 L3 守护进程询问相应的 MAC 地址：
<如果源 IP，1.1.2.10，则设定目的地 MAC = X>
4825

图48

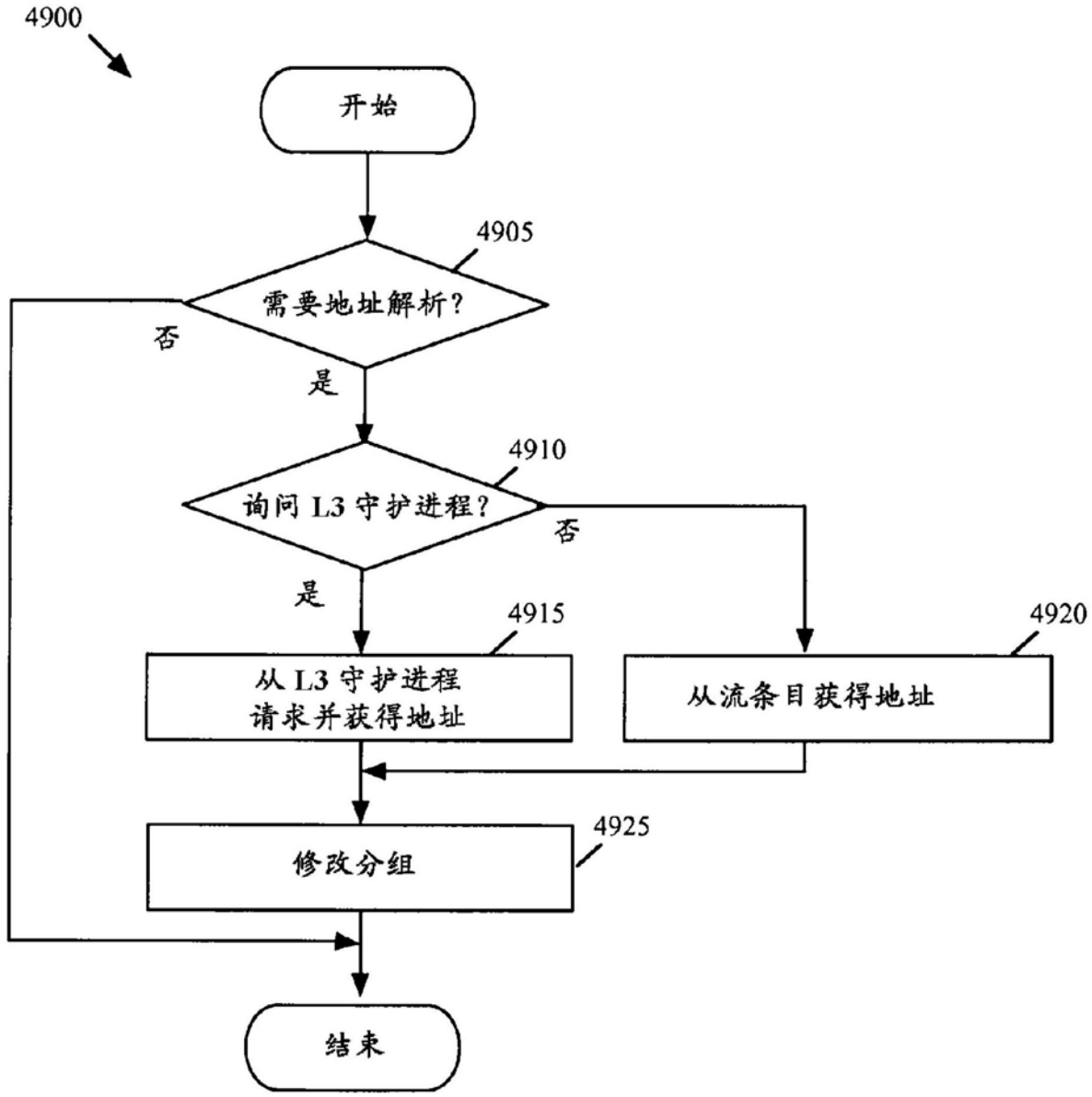

4900
开始
4905
需要地址解析?
否
是
4910
询问 L3 守护进程?
否
是
4915
从 L3 守护进程
请求并获得地址
4920
从流条目获得地址
4925
修改分组
结束

图49

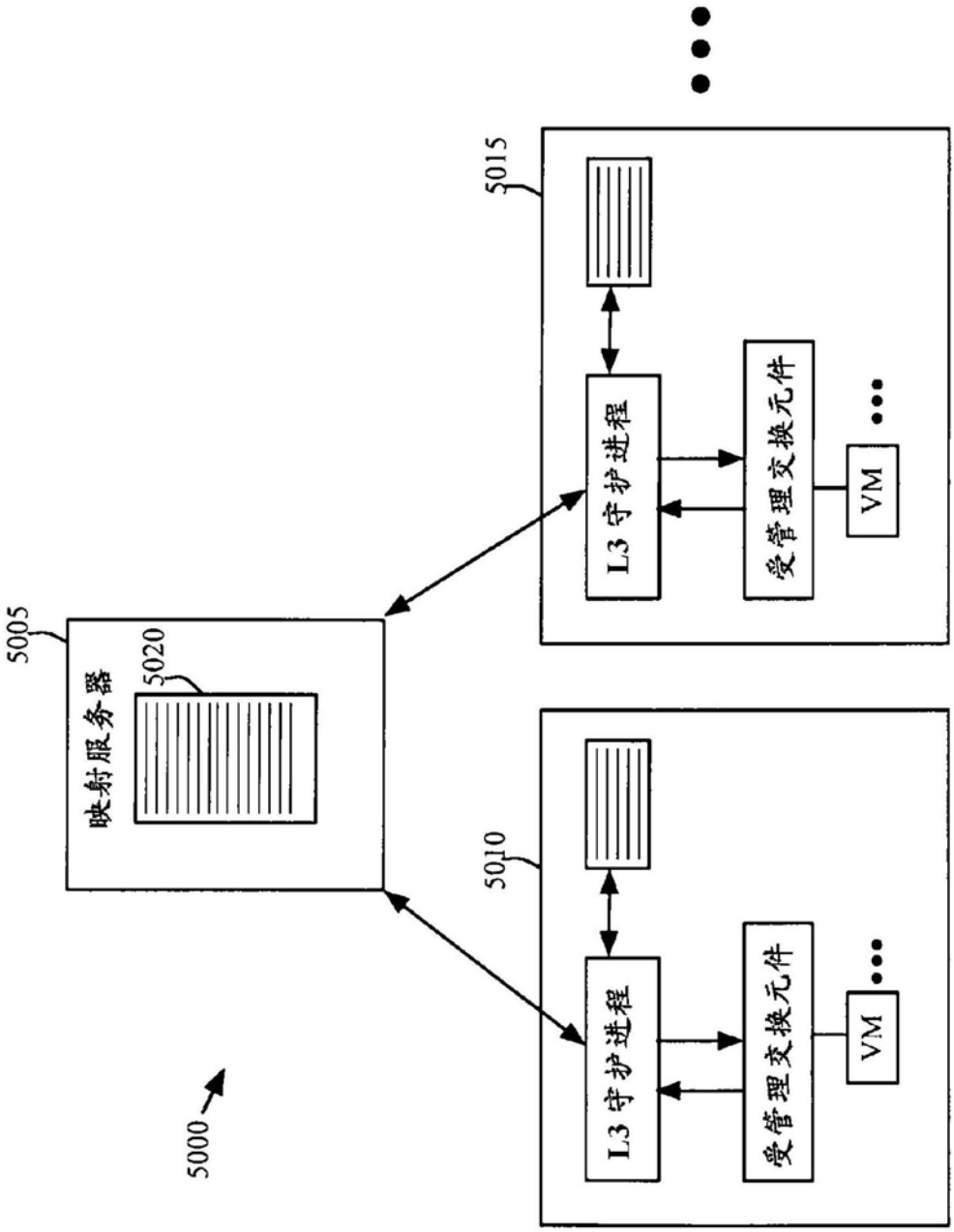

5000
5005
映射服务器
5020
5010
5015
L3 守护进程
受管理交换元件
VM
L3 守护进程
受管理交换元件
VM

图50

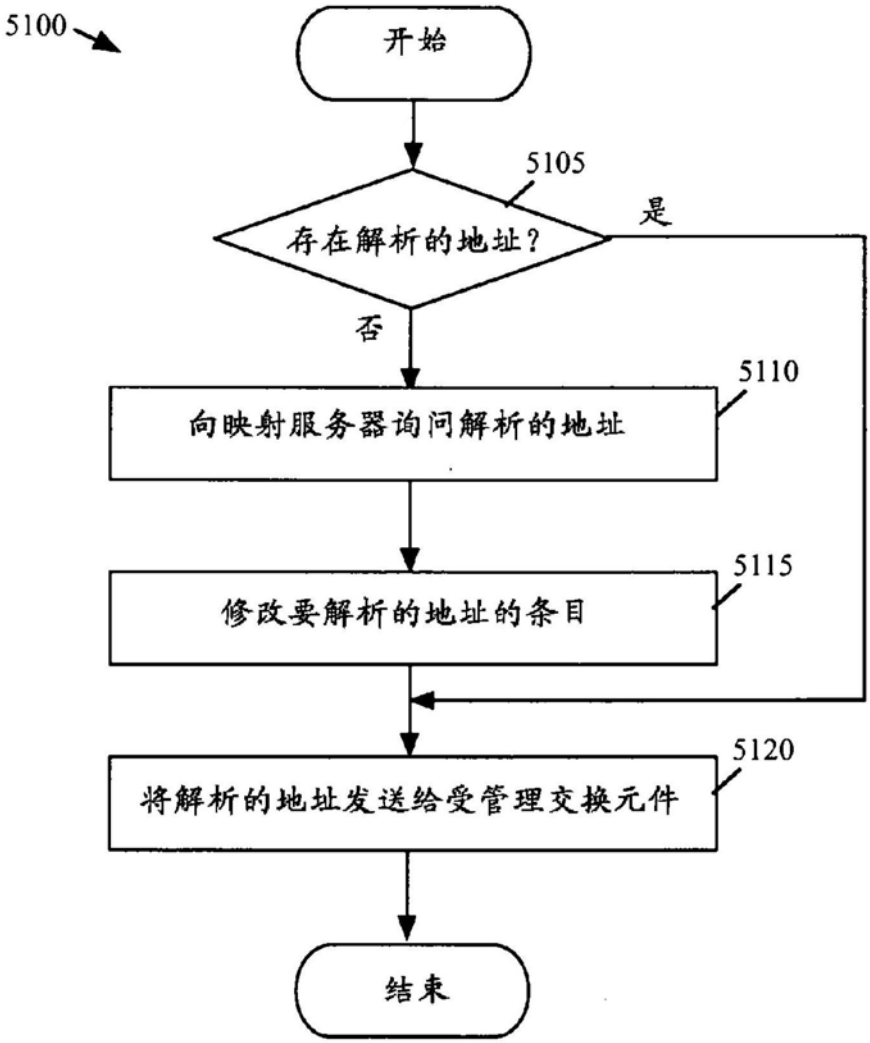

5100
开始
5105
存在解析的地址?
是
否
5110
向映射服务器询问解析的地址
5115
修改要解析的地址的条目
5120
将解析的地址发送给受管理交换元件
结束

图51

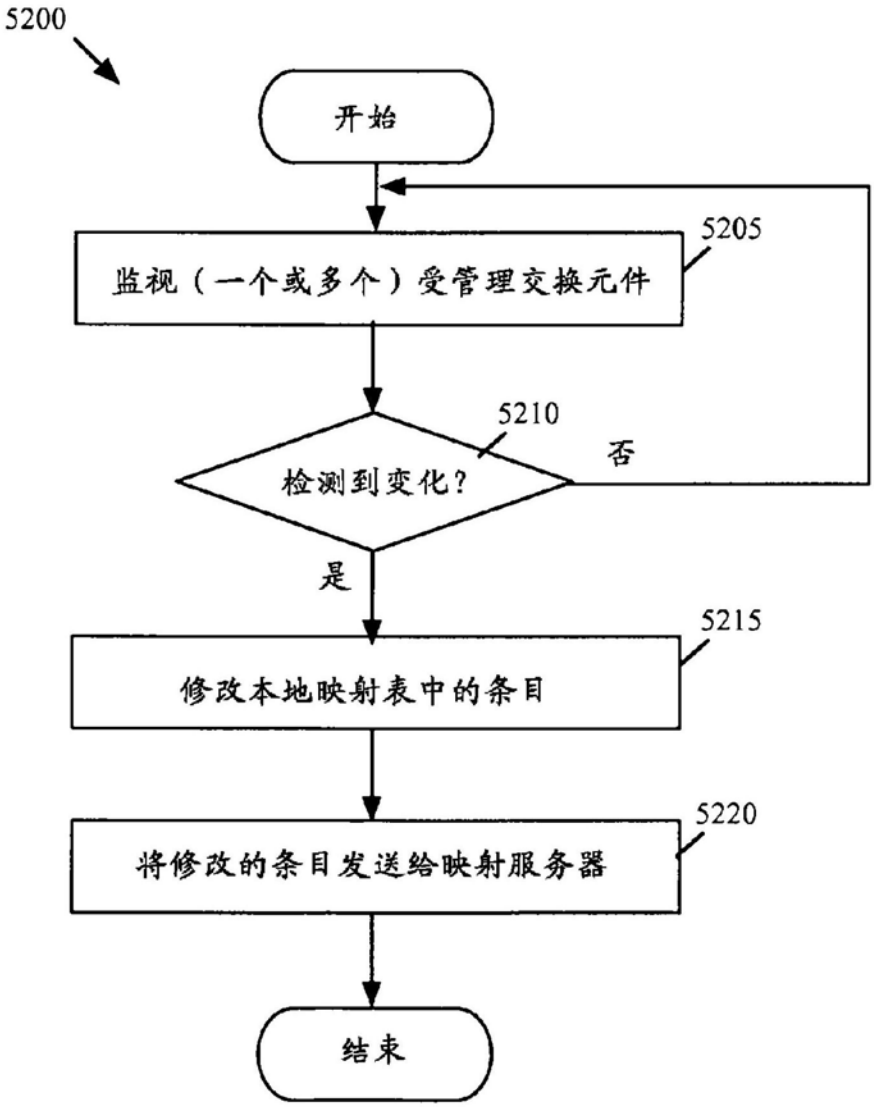

5200
开始
5205
监视（一个或多个）受管理交换元件
5210
检测到变化？
否
是
5215
修改本地映射表中的条目
5220
将修改的条目发送给映射服务器
结束

图52

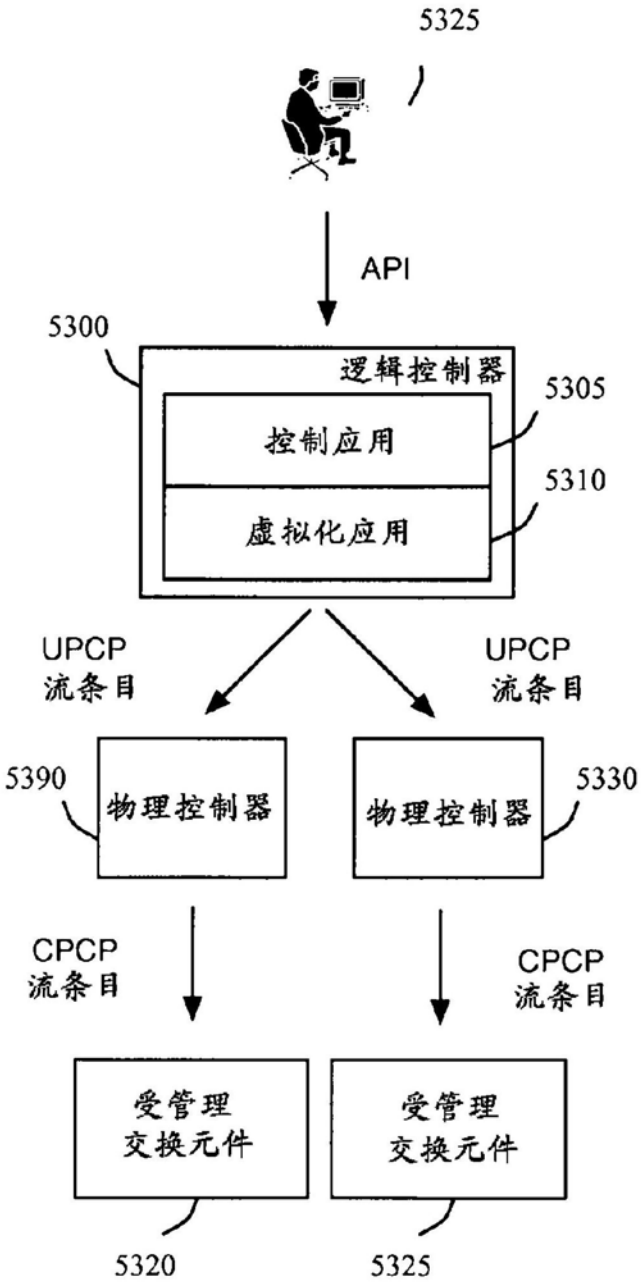

5325
API
5300
逻辑控制器
5305
控制应用
5310
虚拟化应用
UPCP
流条目
UPCP
流条目
5390
物理控制器
物理控制器
5330
CPCP
流条目
CPCP
流条目
受管理
交换元件
受管理
交换元件
5320
5325

图53

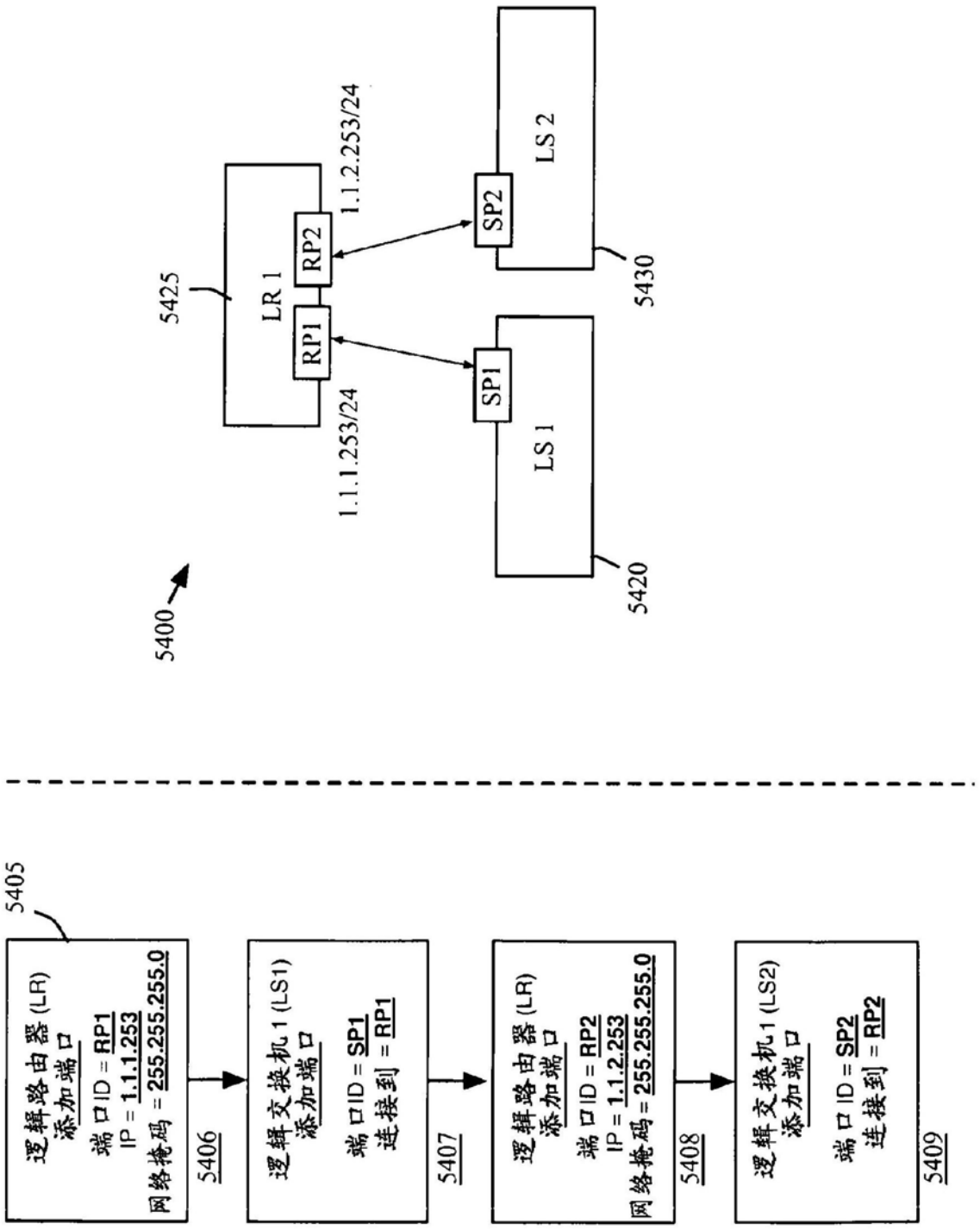

5400
5425
LR 1
RP1
RP2
1.1.1.253/24
1.1.2.253/24
SP1
SP2
LS 1
LS 2
5420
5430
5405
逻辑路由器 (LR)
添加端口
端口 ID = RP1
IP = 1.1.1.253
网络掩码 = 255.255.255.0
5406
逻辑交换机 1 (LS1)
添加端口
端口 ID = SP1
连接到 = RP1
5407
逻辑路由器 (LR)
添加端口
端口 ID = RP2
IP = 1.1.2.253
网络掩码 = 255.255.255.0
5408
逻辑交换机 1 (LS2)
添加端口
端口 ID = SP2
连接到 = RP2
5409

图54

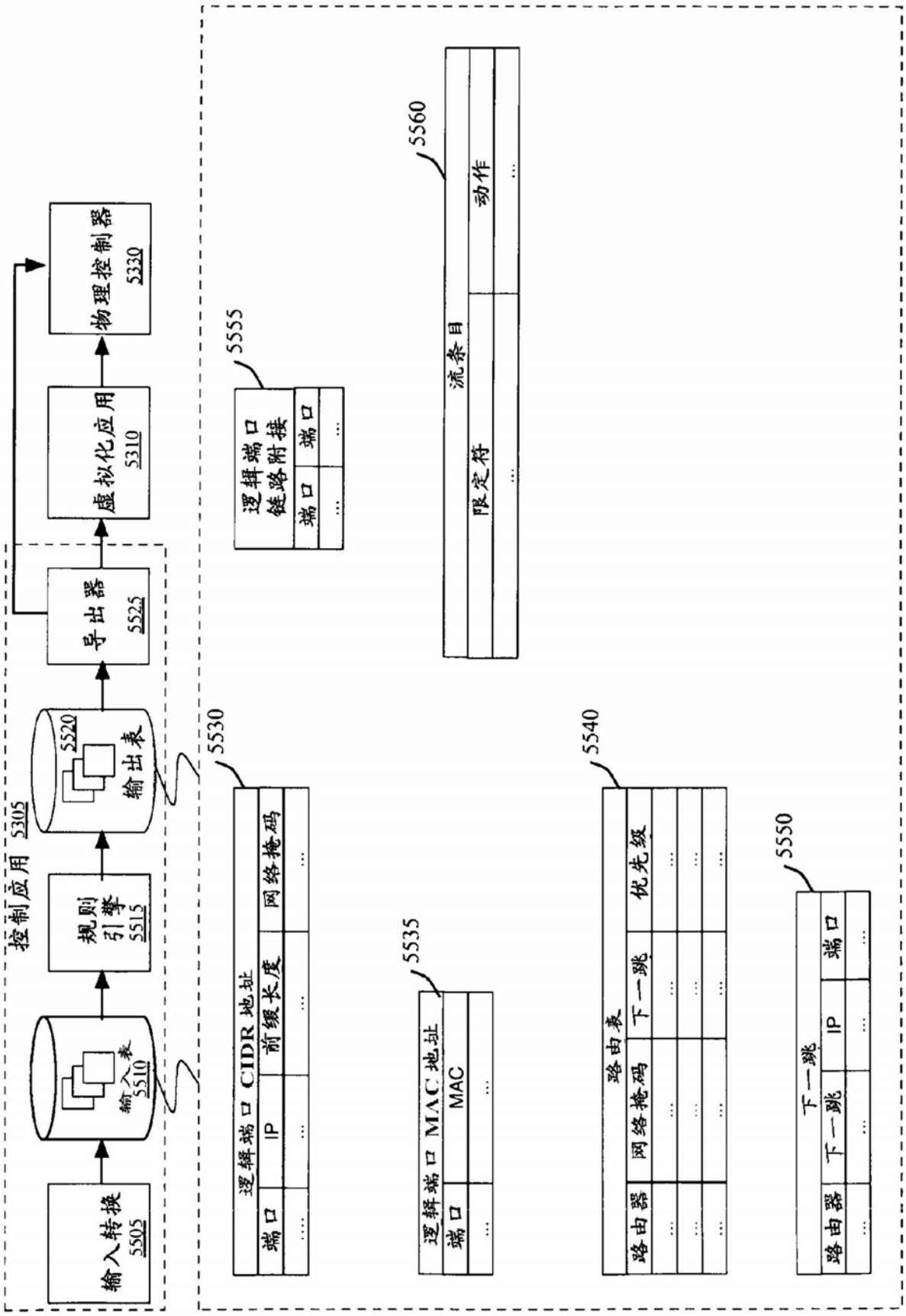

控制应用 5305
输入转换 5505
输入表 5510
规则引擎 5515
输出表 5520
导出器 5525
虚拟化应用 5310
物理控制器 5330
5530
逻辑端口 CIDR 地址
端口
IP
前缀长度
网络掩码
5555
逻辑端口 链路附接
端口
端口
5560
流条目
限定符
动作
5535
逻辑端口 MAC 地址
端口
MAC
5540
路由表
路由器
网络掩码
下一跳
优先级
5550
下一跳
路由器
下一跳
IP
端口

图55

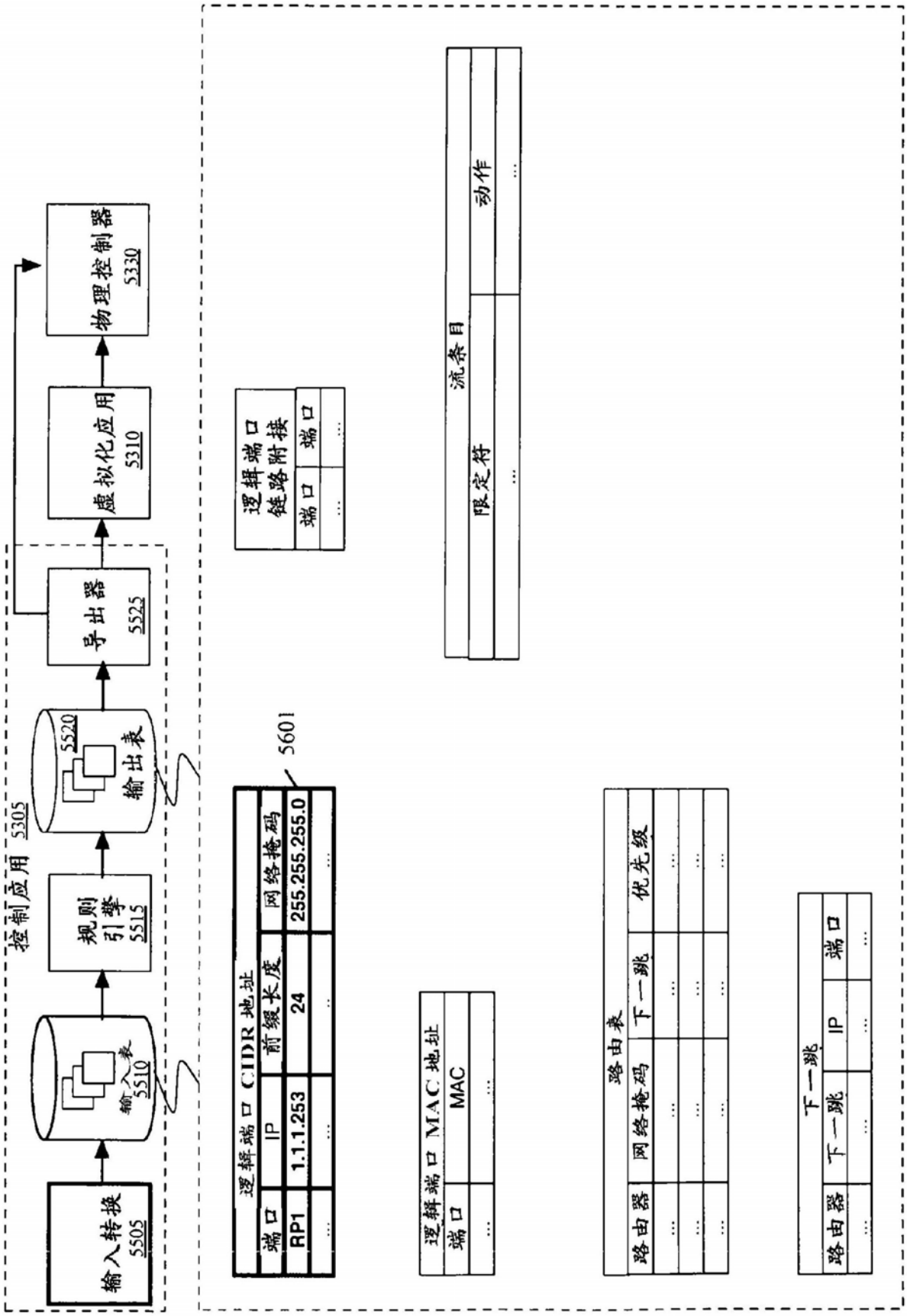

控制应用 5305
输入转换 5505
输入表 5510
规则引擎 5515
输出表 5520
导出器 5525
虚拟化应用 5310
物理控制器 5330
逻辑端口 CIDR 地址
端口
IP
前缀长度
网络掩码
RP1
1.1.1.253
24
255.255.255.0
5601
逻辑端口 MAC 地址
端口
MAC
逻辑端口
链路附接
端口
端口
流条目
限定符
动作
路由表
路由器
网络掩码
下一跳
优先级
下一跳
路由器
下一跳
IP
端口

图56

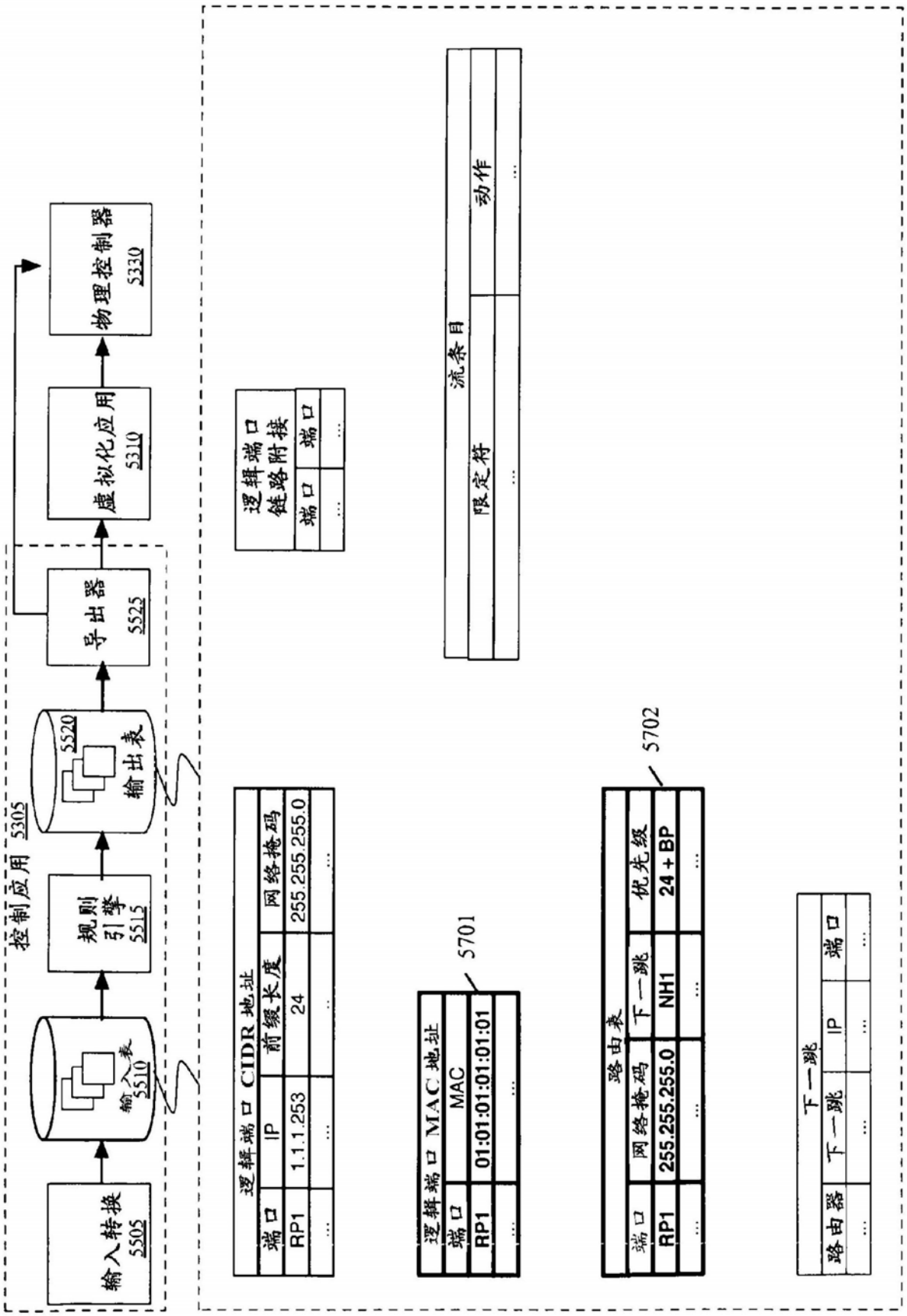

控制应用 5305
输入转换 5505
输入表 5510
规则引擎 5515
输出表 5520
导出器 5525
虚拟化应用 5310
物理控制器 5330
逻辑端口 CIDR 地址
端口 IP 前缀长度 网络掩码
RP1 1.1.1.253 24 255.255.255.0
逻辑端口 MAC 地址
端口 MAC
RP1 01:01:01:01:01:01
5701
逻辑端口
链路附接
端口 端口
流条目
限定符 动作
路由表
端口 网络掩码 下一跳 优先级
RP1 255.255.255.0 NH1 24 + BP
5702
下一跳
路由器 下一跳 IP 端口

图57

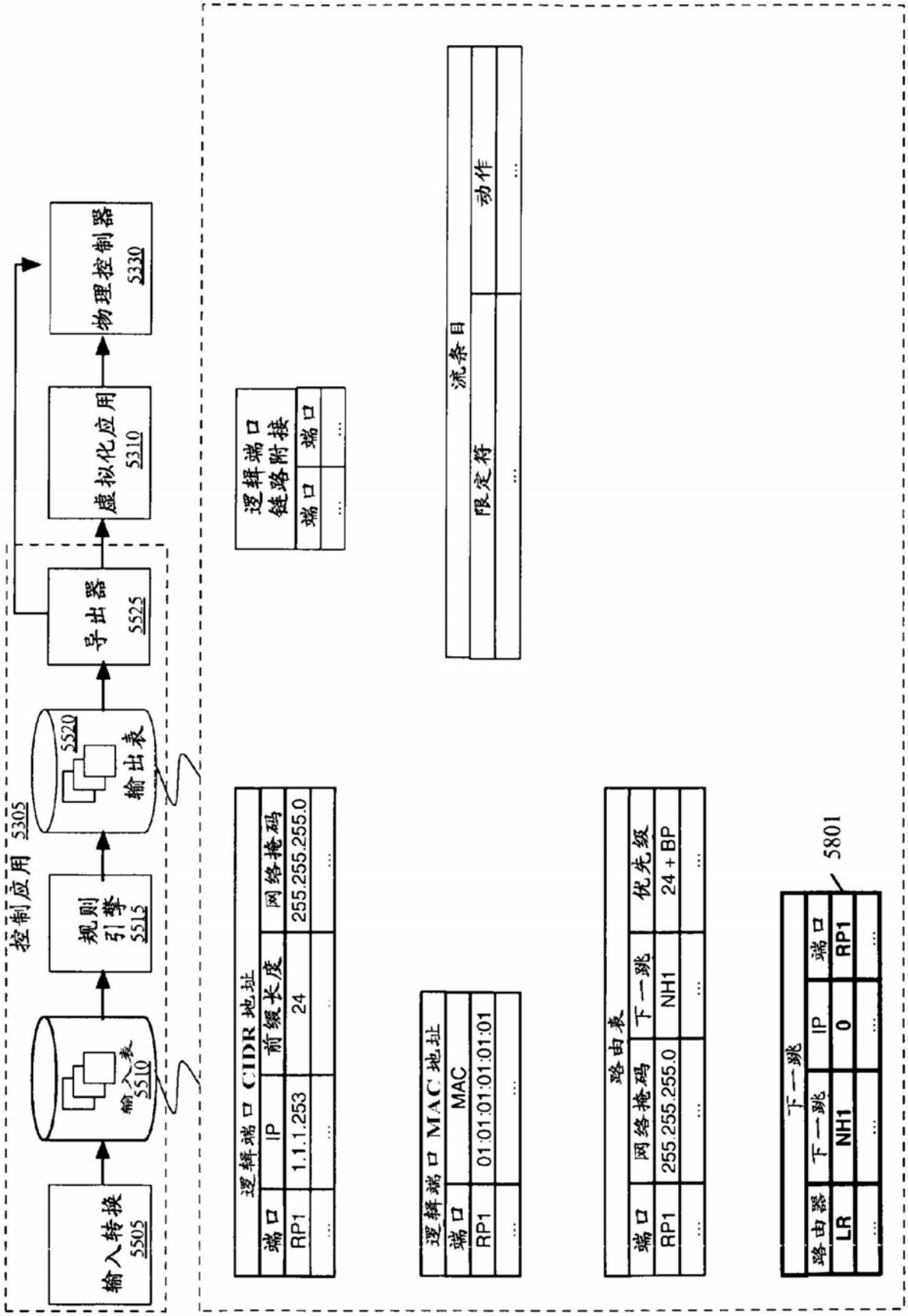

控制应用 5305
输入转换 5505
输入表 5510
规则引擎 5515
输出表 5520
导出器 5525
虚拟化应用 5310
物理控制器 5330
逻辑端口 CIDR 地址
端口 IP 前缀长度 网络掩码
RP1 1.1.1.253 24 255.255.255.0
逻辑端口 MAC 地址
端口 MAC
RP1 01:01:01:01:01:01
逻辑端口链路附接
端口 端口
流条目
限定符 动作
路由表
端口 网络掩码 下一跳 优先级
RP1 255.255.255.0 NH1 24 + BP
下一跳
路由器 下一跳 IP 端口
LR NH1 0 RP1
5801

图58

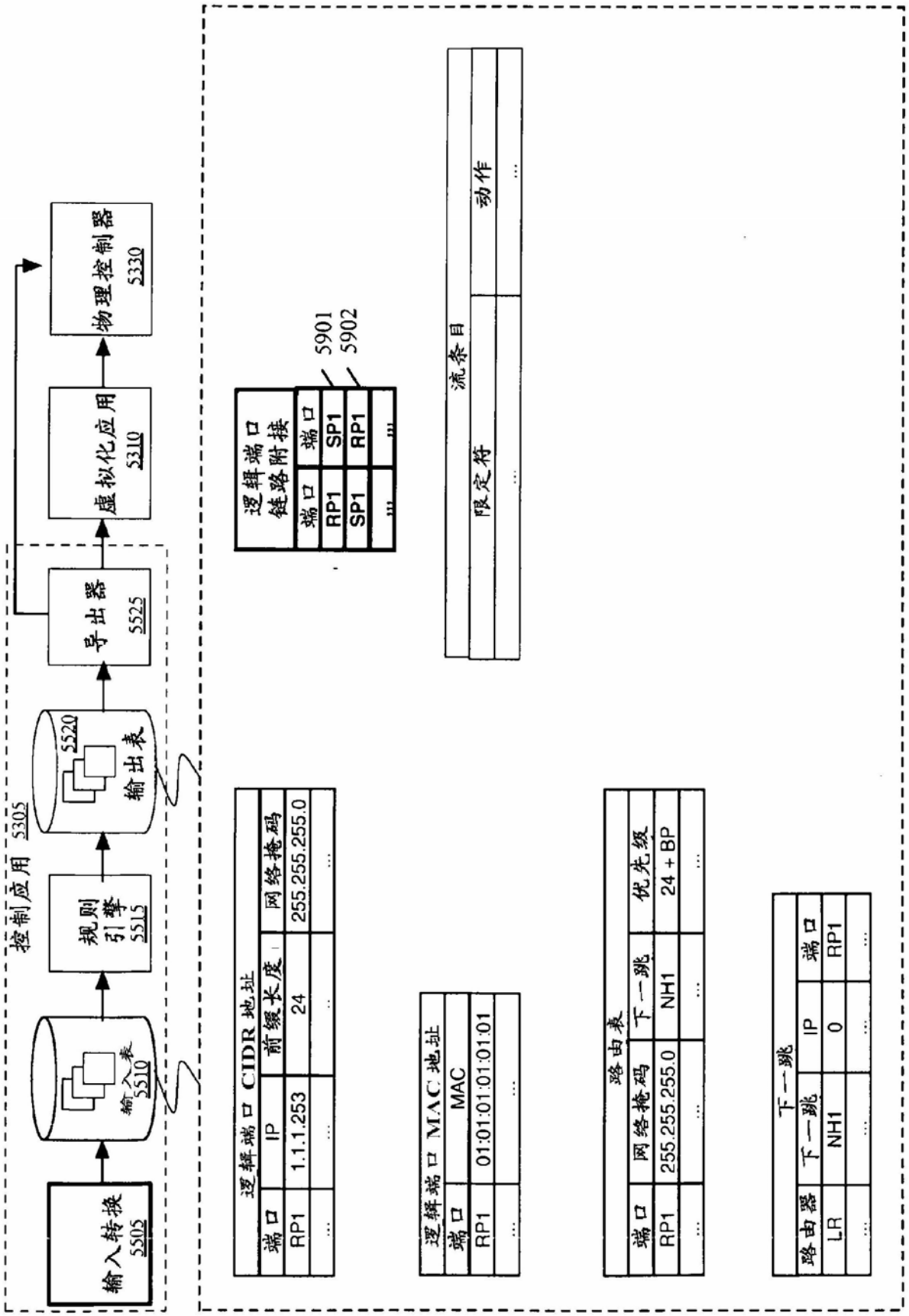

控制应用 5305
输入转换 5505
输入表 5510
规则引擎 5515
输出表 5520
导出器 5525
虚拟化应用 5310
物理控制器 5330
逻辑端口 CIDR 地址
端口 IP 前缀长度 网络掩码
RP1 1.1.1.253 24 255.255.255.0
逻辑端口 MAC 地址
端口 MAC
RP1 01:01:01:01:01:01
路由表
端口 网络掩码 下一跳 优先级
RP1 255.255.255.0 NH1 24 + BP
下一跳
路由器 下一跳 IP 端口
LR NH1 0 RP1
逻辑端口 链路附接
端口 端口
RP1 SP1
SP1 RP1
5901
5902
流条目
限定符 动作

图59

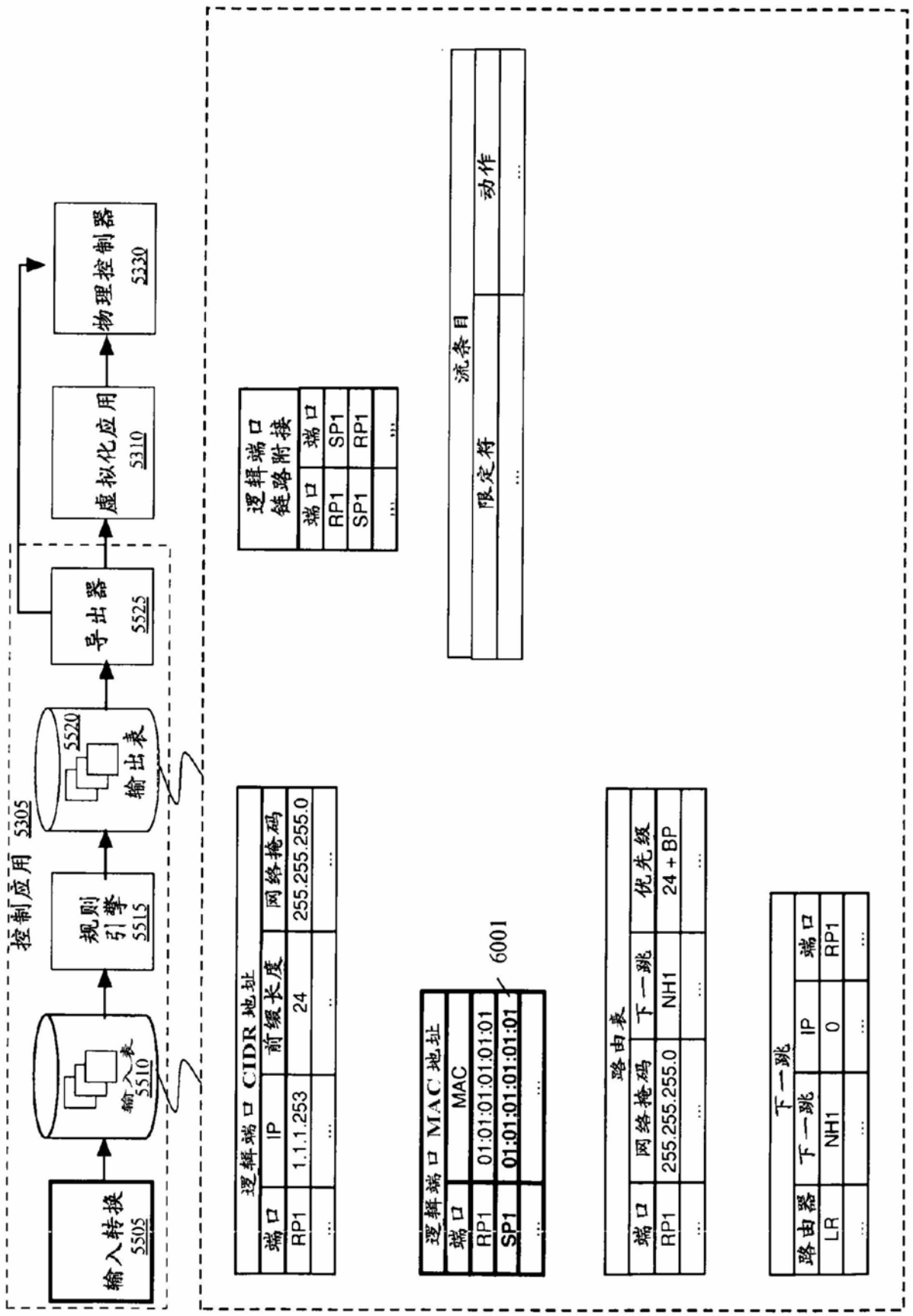

控制应用 5305
输入转换 5505
输入表 5510
规则引擎 5515
输出表 5520
导出器 5525
虚拟化应用 5310
物理控制器 5330
逻辑端口 CIDR 地址
端口 IP 前缀长度 网络掩码
RP1 1.1.1.253 24 255.255.255.0
逻辑端口 MAC 地址
端口 MAC
RP1 01:01:01:01:01:01
SP1 01:01:01:01:01:01
6001
逻辑端口
链路附接
端口 端口
RP1 SP1
SP1 RP1
流条目
限定符 动作
路由表
端口 网络掩码 下一跳 优先级
RP1 255.255.255.0 NH1 24 + BP
下一跳
路由器 下一跳 IP 端口
LR NH1 0 RP1

图60

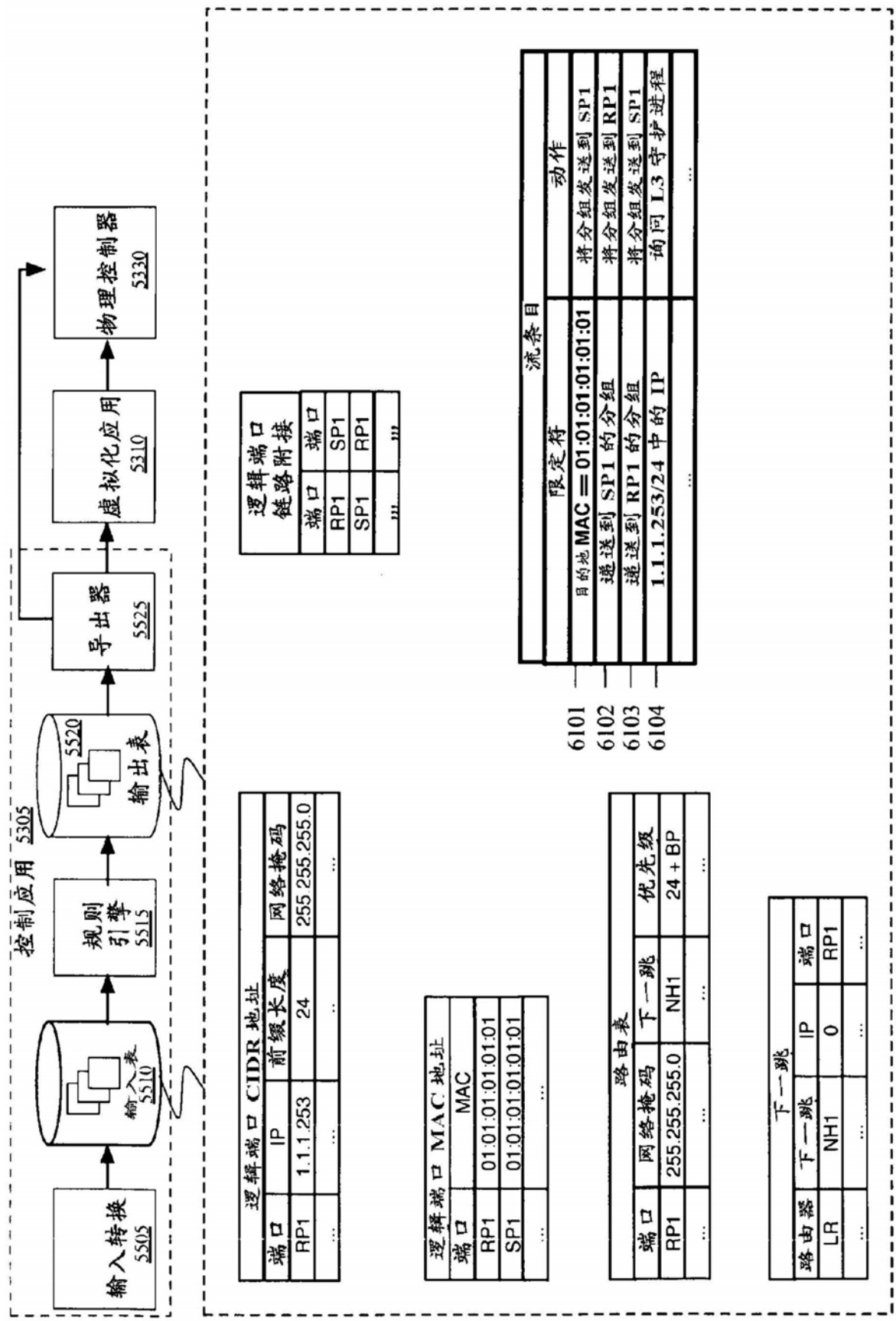

控制应用 5305
输入转换 5505
输入表 5510
规则引擎 5515
输出表 5520
导出器 5525
虚拟化应用 5310
物理控制器 5330
逻辑端口 CIDR 地址
端口 IP 前缀长度 网络掩码
RP1 1.1.1.253 24 255.255.255.0
逻辑端口 MAC 地址
端口 MAC
RP1 01:01:01:01:01:01
SP1 01:01:01:01:01:01
逻辑端口 链路附接
端口 端口
RP1 SP1
SP1 RP1
流条目
限定符 动作
目的地 MAC = 01:01:01:01:01:01 将分组发送到 SP1
递送到 SP1 的分组 将分组发送到 RP1
递送到 RP1 的分组 将分组发送到 SP1
1.1.1.253/24 中的 IP 询问 L3 守护进程
6101
6102
6103
6104
路由表
端口 网络掩码 下一跳 优先级
RP1 255.255.255.0 NH1 24 + BP
下一跳
路由器 下一跳 IP 端口
LR NH1 0 RP1

图61

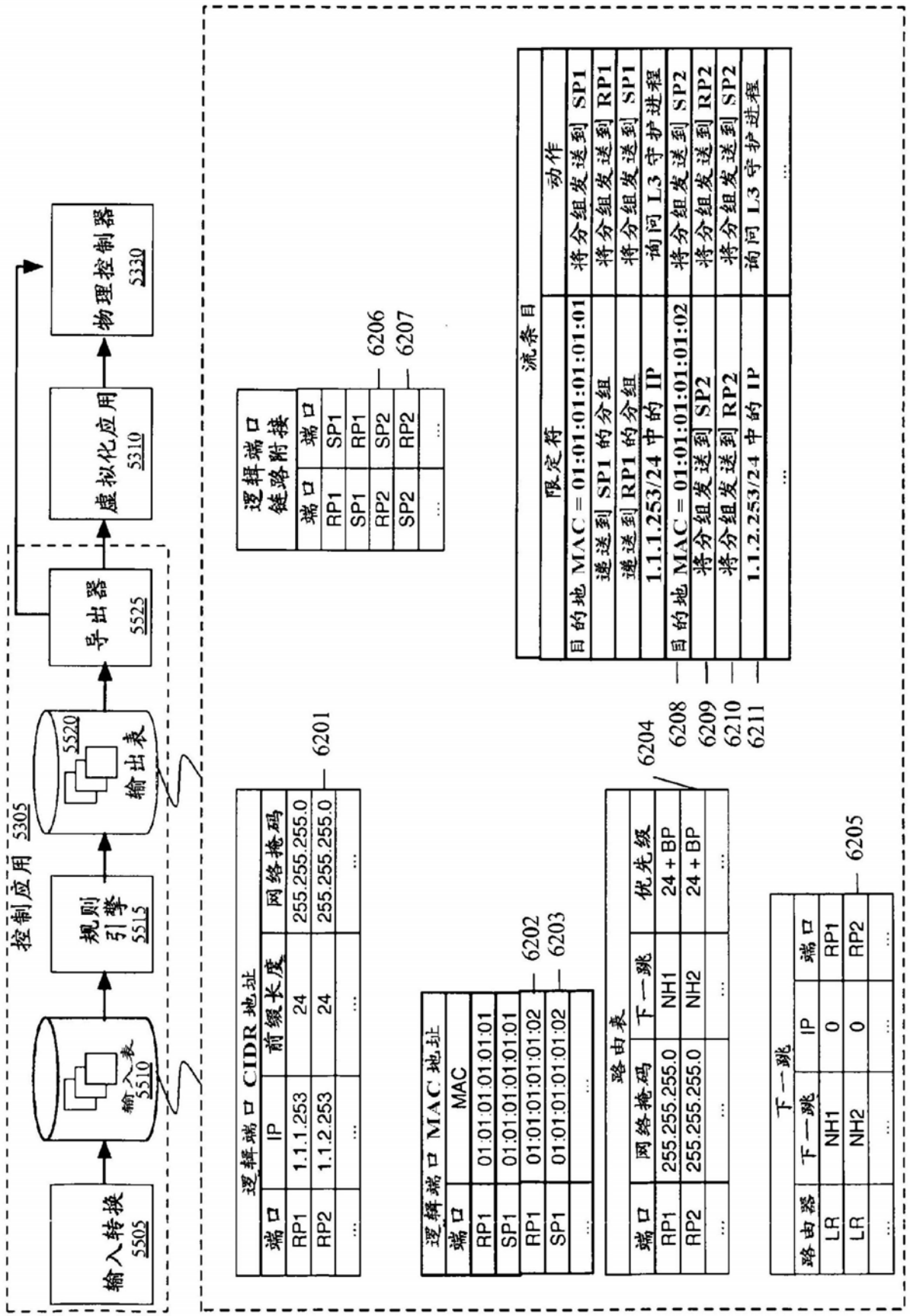

控制应用 5305
输入转换 5505
输入表 5510
规则引擎 5515
输出表 5520
导出器 5525
虚拟化应用 5310
物理控制器 5330
逻辑端口 CIDR 地址
端口 IP 前缀长度 网络掩码
RP1 1.1.1.253 24 255.255.255.0
RP2 1.1.2.253 24 255.255.255.0
6201
逻辑端口 MAC 地址
端口 MAC
RP1 01:01:01:01:01:01
SP1 01:01:01:01:01:01
RP1 01:01:01:01:01:02
SP1 01:01:01:01:01:02
6202
6203
逻辑端口
链路附接
端口 端口
RP1 SP1
SP1 RP1
RP2 SP2
SP2 RP2
6206
6207
路由表
端口 网络掩码 下一跳 优先级
RP1 255.255.255.0 NH1 24 + BP
RP2 255.255.255.0 NH2 24 + BP
6204
下一跳
路由器 下一跳 IP 端口
LR NH1 0 RP1
LR NH2 0 RP2
6205
流条目
限定符 动作
目的地 MAC = 01:01:01:01:01:01 将分组发送到 SP1
递送到 SP1 的分组 将分组发送到 RP1
递送到 RP1 的分组 将分组发送到 SP1
1.1.1.253/24 中的 IP 询问 L3 守护进程
目的地 MAC = 01:01:01:01:01:02 将分组发送到 SP2
将分组发送到 SP2 将分组发送到 RP2
将分组发送到 RP2 将分组发送到 SP2
1.1.2.253/24 中的 IP 询问 L3 守护进程
6208
6209
6210
6211

图62

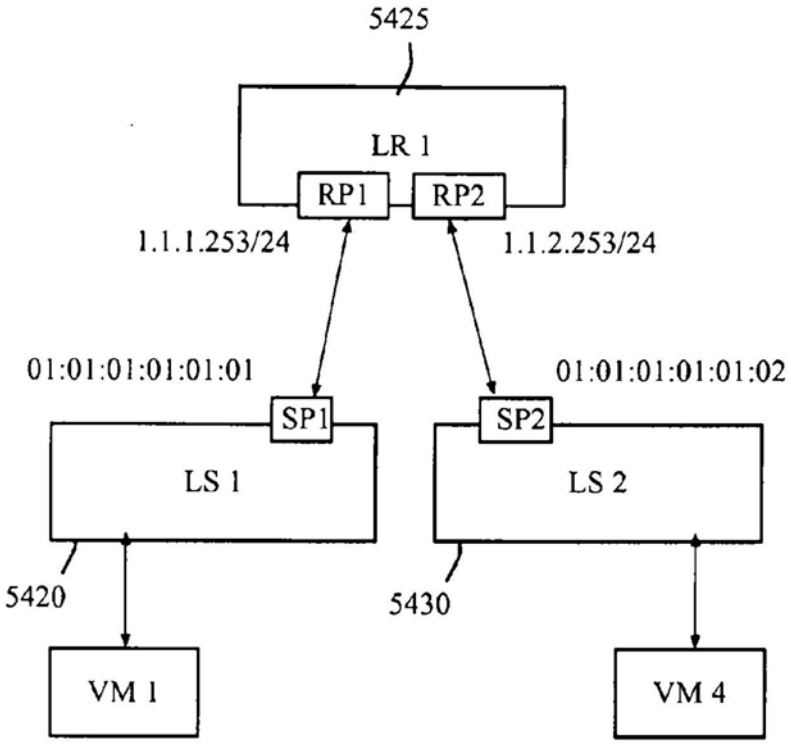

5425
LR 1
RP1
RP2
1.1.1.253/24
1.1.2.253/24
01:01:01:01:01:01
01:01:01:01:01:02
SP1
SP2
LS 1
LS 2
5420
5430
VM 1
VM 4

图63

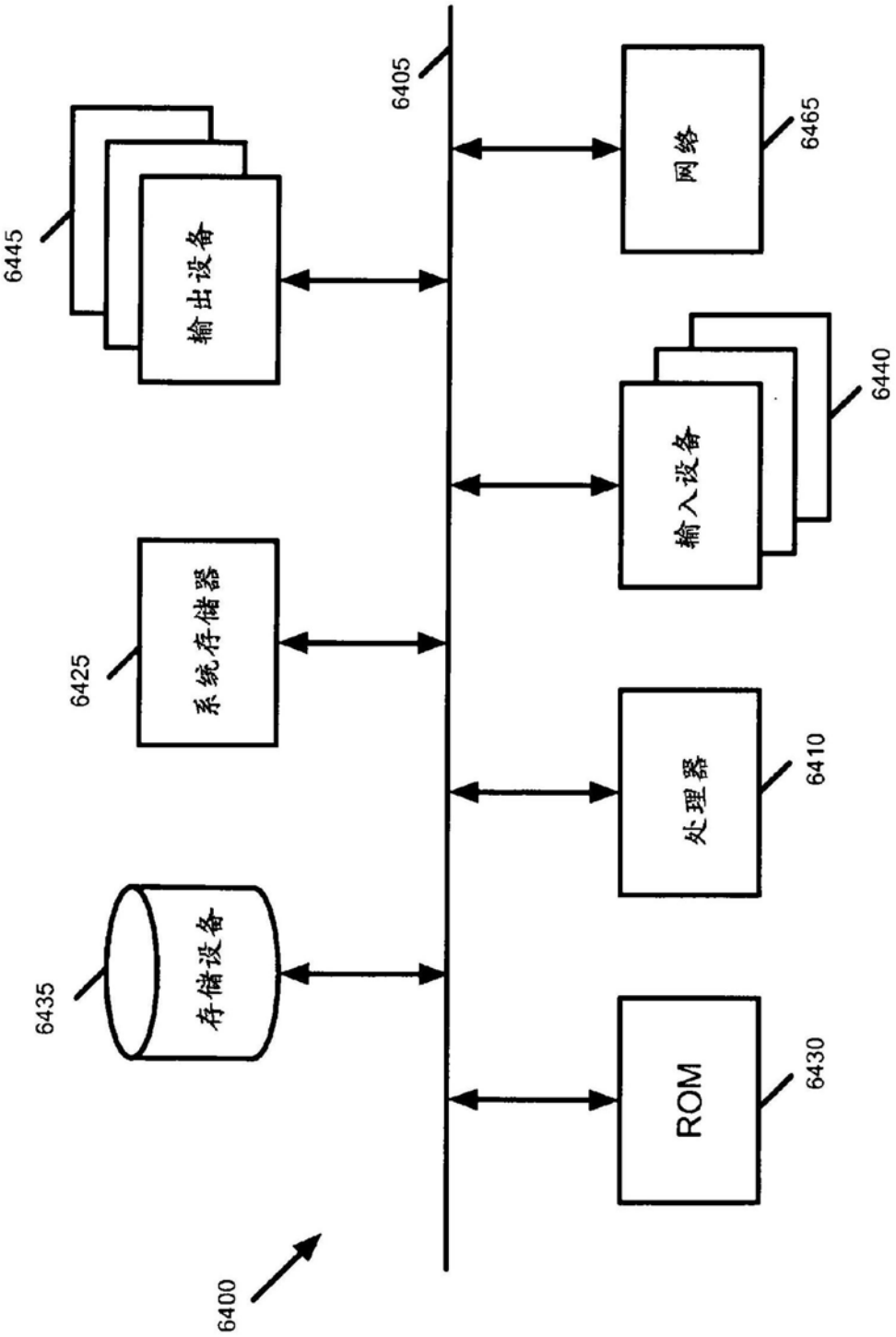

6405
网络
6465
6445
输出设备
6440
输入设备
6425
系统存储器
6410
处理器
6435
存储设备
6430
ROM
6400

图64