



(12) 发明专利申请

(10) 申请公布号 CN 116210204 A

(43) 申请公布日 2023.06.02

(21) 申请号 202180061266.5

(22) 申请日 2021.07.14

(30) 优先权数据

63/051,728 2020.07.14 US

63/132,377 2020.12.30 US

(85) PCT国际申请进入国家阶段日

2023.01.13

(86) PCT国际申请的申请数据

PCT/US2021/041678 2021.07.14

(87) PCT国际申请的公布数据

W02022/015883 EN 2022.01.20

(71) 申请人 甲骨文国际公司

地址 美国加利福尼亚

(72) 发明人 J·S·布拉尔

L·M·克雷格-斯蒂克勒斯

S·贝克 B·E·博克曼

P·C·琼斯

(74) 专利代理机构 中国贸促会专利商标事务所

有限公司 11038

专利代理师 边海梅

(51) Int.Cl.

H04L 12/28 (2006.01)

H04L 12/46 (2006.01)

H04L 45/02 (2022.01)

H04L 45/586 (2022.01)

H04L 45/745 (2022.01)

H04L 47/125 (2022.01)

H04L 47/783 (2022.01)

H04L 61/103 (2022.01)

G06F 9/455 (2018.01)

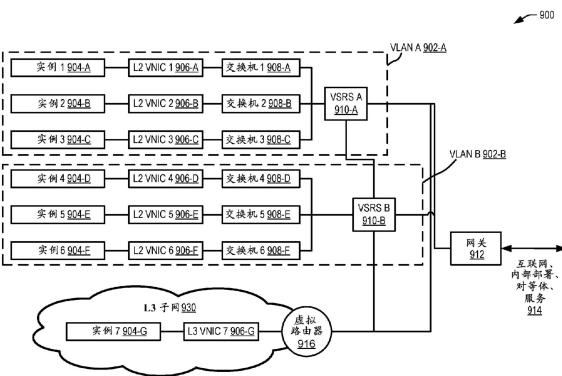
权利要求书2页 说明书67页 附图28页

(54) 发明名称

用于VLAN交换和路由服务的系统和方法

(57) 摘要

本文公开了用于VLAN交换和路由服务(VSRS)的系统和方法。一种方法可以包括为VSRS的实例生成表,该VSRS将第一虚拟层2网络(VLAN)与第二网络耦合。该表可以包含识别虚拟层2网络中的实例的IP地址、MAC地址和虚拟接口标识符的信息。该方法可以包括用VSRS接收来自第一实例的分组,该分组被指定用于递送到虚拟层2网络内的第二实例,基于与分组一起接收的信息和包含在表内的信息来用VSRS识别虚拟层2网络内的第二实例以递送分组,以及将分组递送到识别出的第二实例。



1. 一种方法,包括:

为VLAN交换和路由服务 (VSRS) 的实例生成表,该VSRS将第一虚拟层2网络 (VLAN) 与第二网络耦合,其中该表包含识别虚拟层2网络内的实例的IP地址、MAC地址和虚拟接口标识符的信息;

用VSRS接收来自第一实例的分组,该分组被指定用于递送到虚拟层2网络内的第二实例;

基于与分组一起接收的信息和包含在表内的信息,用VSRS识别虚拟层2网络内的第二实例以递送分组;以及

将分组递送到识别出的第二实例。

2. 如权利要求1所述的方法,其中第一虚拟层2网络包括多个实例。

3. 如权利要求2所述的方法,其中VLAN包括多个L2虚拟网络接口卡 (VNIC) 和多个交换机,其中所述多个实例中的每个实例与包括唯一L2虚拟网络接口卡 (VNIC) 和唯一交换机的对通信耦合。

4. 如权利要求1至3中的任一项所述的方法,其中基于与分组一起接收的信息和包含在表内的信息,用VSRS识别虚拟层2网络内的第二实例以递送分组包括:

用VSRS确定该表不包括用于第二实例的映射信息;

用VSRS挂起分组的递送;

用VSRS向VLAN中的VNIC广播ARP请求,该ARP请求包含第二实例的IP地址;以及

用VSRS从第二实例的VNIC接收ARP响应。

5. 如权利要求4所述的方法,还包括基于接收到的ARP响应来更新表。

6. 如权利要求2或3所述的方法,其中第一实例在虚拟层2网络外部并且在第二网络中。

7. 如权利要求6所述的方法,其中第二网络包括L3网络。

8. 如权利要求6所述的方法,其中第二网络包括第二虚拟层2网络。

9. 如权利要求2或3所述的方法,其中表是基于由VSRS接收的通信而生成的。

10. 如权利要求9所述的方法,还包括将VSRS实例化为多个硬件节点上的服务。

11. 如权利要求10所述的方法,还包括跨硬件节点分布表。

12. 如权利要求11所述的方法,其中跨硬件节点分布的表可由另一个VSRS实例化访问。

13. 如权利要求1至3中的任一项所述的方法,其中第一实例在第一虚拟层2网络内部。

14. 如权利要求1至13中的任一项所述的方法,还包括:用VSRS从虚拟层2网络内部的第三实例接收分组,其中该分组被指定用于递送到虚拟层2网络外部的第四实例并将该分组转发到第四实例。

15. 如权利要求1至13中的任一项所述的方法,还包括:用VSRS从虚拟层2网络内部的第三实例接收分组,其中该分组被指定用于递送到由虚拟层2网络内部的第三实例使用的服务。

16. 如权利要求15所述的方法,其中服务包括以下至少之一:DHCP;NTP;以及DNS。

17. 如权利要求1至13中的任一项所述的方法,还包括:用VSRS从一个虚拟层2网络内部的第三实例接收分组,其中该分组被指定用于递送到第二虚拟层2网络中的第四实例。

18. 如权利要求1至9中的任一项所述的方法,还包括:跨服务节点的机群分发用于具有层2和层3网络信息的VSRS的实例的表以提供VSRS的高度可靠和高度可扩展的实例化。

19. 如权利要求1所述的方法,还包括:

用VSRS从第一虚拟层2网络内部的第三实例接收分组;以及
用VSRS学习第三实例的映射。

20. 一种系统,包括:

物理网络包括:

至少一个处理器;以及

网络虚拟化设备,

其中所述至少一个处理器被配置为:

实例化VLAN交换和路由服务(VSRS)的实例,VSRS将第一虚拟层2网络与第二网络耦合;

为VSRS的实例生成表,其中该表包含识别虚拟层2网络内的实例的IP地址、MAC地址和虚拟接口标识符的信息;

用VSRS接收来自第一实例的分组,该分组被指定用于递送到虚拟层2网络内的第二实例;

基于与分组一起接收的信息和包含在表内的信息,用VSRS识别虚拟层2网络内的第二实例以递送分组;以及

将分组递送到识别出的第二实例。

21. 一种非暂态计算机可读存储介质,存储可由一个或多个处理器执行的多个指令,所述多个指令在由所述一个或多个处理器执行时使所述一个或多个处理器:

实例化VLAN交换和路由服务(VSRS)的实例,VSRS将第一虚拟层2网络与第二网络耦合;

为VSRS的实例生成表,其中该表包含识别虚拟层2网络内的实例的IP地址、MAC地址和虚拟接口标识符的信息;

用VSRS接收来自第一实例的分组,该分组被指定用于递送到虚拟层2网络内的第二实例;

基于与分组一起接收的信息和包含在表内的信息,用VSRS识别虚拟层2网络内的第二实例以递送分组;以及

将分组递送到识别出的第二实例。

用于VLAN交换和路由服务的系统和方法

[0001] 相关申请的交叉引用

[0002] 本申请要求以下申请的权益：

[0003] (1) 美国临时申请No.63/051,728,于2020年7月14日提交且标题为“VLAN Switching And Routing Service And Layer-2Networking In A Virtualized Cloud Environment”,以及

[0004] (2) 美国临时申请No.63/132,377,于2020年12月30日提交且标题为“Layer-2 Networking In A Virtualized Cloud Environment”。出于所有目的,上述临时申请的全部内容通过引用并入本文。

[0005] 本申请还与2021年7月14日提交的标题为“VIRTUAL LAYER-2NETWORK”的美国申请No._____ (代理人案卷No.088325-1203134-276500US) 相关,并且本申请还与2021年7月14日提交的标题为“INTERFACE-BASED ACLS IN A LAYER-2NETWORK”的美国申请No._____ (代理人案卷Attorney Docket No.088325-1256549-276520US) 相关,每个相关申请的全部内容都出于所有目的通过引用并入本文。

背景技术

[0006] 云计算提供计算资源的按需可用性。云计算可以基于用户可通过互联网访问的数据中心。云计算可以提供基础设施即服务(IaaS)。可以创建虚拟网络以供用户使用。但是,这些虚拟网络具有限制其功能性和价值的局限性。因而,期望进一步的改进。

发明内容

[0007] 本公开的一个方面涉及一种计算机实现的方法。该方法包括在虚拟化的云环境中提供虚拟层3网络并且在虚拟化的云环境中提供虚拟层2网络,虚拟层3网络由底层的物理网络托管,虚拟层2网络由底层的物理网络托管。

[0008] 在一些实施例中,虚拟层2网络可以是虚拟局域网(VLAN)。在一些实施例中,VLAN包括多个端点。在一些实施例中,多个端点可以是多个计算实例。在一些实施例中,VLAN包括多个L2虚拟网络接口卡(L2 VNIC)和多个交换机。

[0009] 在一些实施例中,多个计算实例中的每一个与包括唯一的L2虚拟网络接口卡(L2 VNIC)和唯一交换机的对通信地耦合。在一些实施例中,多个交换机一起可以形成分布式交换机。在一些实施例中,多个交换机中的每一个根据从与交换机配对的L2 VNIC接收的映射表来路由出站流量。在一些实施例中,映射表为VLAN内的端点识别接口到MAC地址映射。

[0010] 在一些实施例中,该方法还包括在网络虚拟化设备(NVD)上实例化包括唯一L2 VNIC和唯一交换机的对。在一些实施例中,该方法包括从VLAN内的另一个端点在多个计算实例之一的唯一L2VNIC处接收寻址到多个计算实例之一的分组,并且用多个计算实例之一的唯一L2 VNIC学习另一个端点的映射。在一些实施例中,另一个端点的映射包括另一个端点的接口到MAC地址映射。

[0011] 在一些实施例中,该方法包括用多个计算实例之一的唯一L2VNIC解封装接收到的

分组,并将解封装的分组转发到多个计算实例之一。在一些实施例中,该方法包括用多个计算实例之一学习另一个端点的IP地址到MAC地址映射。

[0012] 在一些实施例中,该方法包括从VLAN中的第一计算实例发送IP分组,该IP分组包括VLAN中的第二计算实例的目的地IP地址,在与第一计算实例相关联的第一L2 VNIC处接收IP分组,在第一L2VNIC处封装IP分组,并经由第一交换机将IP分组转发到第二计算实例。在一些实施例中,第一交换机和第一L2 VNIC一起用于与第一计算实例通信地耦合的对。在一些实施例中,该方法还包括在第二VNIC处接收IP分组,该第二VNIC与第二计算实例相关联,在第二VNIC处解封装IP分组,以及将IP分组从第二VNIC转发到第二计算实例。

[0013] 在一些实施例中,虚拟层2网络包括多个虚拟局域网(VLAN)。在一些实施例中,多个VLAN中的每一个包括多个端点。在一些实施例中,多个VLAN包括第一VLAN和第二VLAN。在一些实施例中,第一VLAN包括多个第一端点,并且第二VLAN包括多个第二端点。在一些实施例中,多个VLAN中的每一个具有唯一标识符。在一些实施例中,第一VLAN中的多个第一端点之一与第二VLAN中的多个第二端点之一通信。

[0014] 本公开的一个方面涉及一种包括物理网络的系统。物理网络包括至少一个主机机器及至少一个网络虚拟化设备。物理网络可以在虚拟化的云环境中提供虚拟层3网络并且在虚拟化的云环境中提供虚拟层2网络,该虚拟层3网络由底层的物理网络托管,该虚拟层2网络由底层的物理网络托管。

[0015] 本公开的一个方面涉及一种非暂态计算机可读存储介质,其存储可由一个或多个处理器执行的多个指令。当由一个或多个处理器执行时,多个指令使一个或多个处理器在虚拟化的云环境中提供虚拟层3网络并在虚拟化的云环境中提供虚拟层2网络,该虚拟层3网络由底层的物理网络托管,该虚拟层2网络由底层的物理网络托管。

[0016] 本公开的一个方面涉及一种方法,包括为VLAN交换和路由服务(VSRS)的实例生成表,VSRS将第一虚拟层2网络与第二网络耦合。在一些实施例中,该表包含识别IP地址、MAC地址和第一虚拟层2网络内的实例的虚拟接口标识符的信息。该方法包括用VSRS从第一实例接收指定用于递送到第一虚拟层2网络中的第二实例的分组,用VSRS识别第一虚拟层2网络中的第二实例以基于随分组接收的信息和表中包含的信息递送分组,并将分组递送到识别出的第二实例。

[0017] 在一些实施例中,第一虚拟层2网络包括多个实例。在一些实施例中,第一虚拟2层网络包括多个L2虚拟网络接口卡(L2 VNIC)和多个交换机。在一些实施例中,多个实例中的每一个与包括唯一L2虚拟网络接口卡(L2 VNIC)和唯一交换机的对通信地耦合。

[0018] 在一些实施例中,基于随分组接收的信息和包含在表中的信息用VSRS识别第一虚拟层2网络内的用于递送分组的第二实例包括用VSRS确定该表不包括用于第二实例的映射信息,用VSRS暂停分组的递送,用VSRS向第一虚拟层2网络中的L2 VNIC广播ARP请求,该ARP请求包含第二实例的IP地址,并用VSRS接收来自第二实例的L2 VNIC的ARP响应。

[0019] 在一些实施例中,该方法还包括基于接收到的ARP响应来更新表。在一些实施例中,第一实例在第一虚拟层2网络外部并且在第二网络中。在一些实施例中,第二网络可以是L3网络。在一些实施例中,第二网络可以是第二虚拟层2网络。在一些实施例中,表是基于由VSRS接收的通信生成的。

[0020] 在一些实施例中,该方法包括将VSRS实例化为多个硬件节点上的服务。在一些实

施例中,该方法包括跨硬件节点分布表。在一些实施例中,跨硬件节点分布的表可由另一个VSRS实例化访问。在一些实施例中,第一实例在第一虚拟层2网络内部。

[0021] 在一些实施例中,该方法包括用VSRS从第一虚拟层2网络内部的第三实例接收分组。在一些实施例中,该分组被指定用于递送到第一虚拟层2网络外部的第四实例并且将分组转发到第四实例。在一些实施例中,该方法包括用VSRS从第一虚拟层2网络内部的第三实例接收分组。在一些实施例中,该分组被指定用于递送到由第一虚拟层2网络内的第三实例使用的服务。在一些实施例中,该服务可以是以下至少之一:DHCP;NTP;以及DNS。

[0022] 在一些实施例中,该方法包括用VSRS从第一虚拟层2网络内部的第三实例接收分组。在一些实施例中,该分组被指定用于递送到第二虚拟层2网络中的第四实例。在一些实施例中,该方法包括跨一组服务节点分发具有层2和层3网络信息的VSRS的实例的表以提供VSRS的高度可靠和高度可扩展的实例化。在一些实施例中,该方法包括用VSRS从第一虚拟层2网络内的第三实例接收分组,并且用VSRS学习第三实例的映射。

[0023] 本公开的一个方面涉及一种系统。该系统包括物理网络。物理网络包括至少一个处理器和网络虚拟化设备。至少一个处理器可以实例化VLAN交换和路由服务(VSRS)的实例,VSRS将第一虚拟层2网络与第二网络耦合,并且为VSRS的实例生成表。在一些实施例中,该表包含识别IP地址、MAC地址和第一虚拟层2网络内的实例的虚拟接口标识符的信息。至少一个处理器可以用VSRS从第一实例接收被指定用于递送到第一虚拟层2网络内的第二实例的分组,用VSRS识别第一虚拟层2网络内的第二实例以基于随分组接收的信息和表中包含的信息递送分组,并将分组递送到识别出的第二实例。

[0024] 本公开的一个方面涉及一种非暂态计算机可读存储介质,其存储可由一个或多个处理器执行的多个指令。多个指令在由一个或多个处理器执行时使一个或多个处理器:实例化VLAN交换和路由服务(VSRS)的实例,该VSRS将第一虚拟层2网络与第二网络耦合,并为VSRS的实例生成表。在一些实施例中,该表包含识别IP地址、MAC地址和用于第一虚拟层2网络内的实例的虚拟接口标识符的信息。当由一个或多个处理器执行时,多个指令使一个或多个处理器:用VSRS从第一实例接收被指定用于递送到第一虚拟层2网络内的第二实例的分组,用VSRS识别第一虚拟层2网络内的第二实例以基于随分组接收的信息和表中包含的信息来递送分组,并将分组递送到识别出的第二实例。

[0025] 本公开的一个方面涉及一种方法。该方法包括:经由第一虚拟层2网络中的目的地L2虚拟网络接口卡(目的地L2 VNIC)将分组从虚拟网络中的源计算实例发送到目的地计算实例,对于具有源虚拟网络接口卡(源VNIC)的分组评估访问控制列表(ACL),在分组中嵌入与分组相关的ACL信息,将封装的分组转发到虚拟交换和路由服务(VSRS),该VSRS将第一虚拟层2网络(VLAN)与第二网络耦合,用VSRS识别第一虚拟层2网络内的目的地L2 VNIC以基于随分组接收的信息和映射表中包含的映射信息来递送分组,用VSRS访问来自分组的ACL信息,并将访问的ACL信息应用到分组。

[0026] 在一些实施例中,分组包括IP分组。在一些实施例中,源计算实例位于虚拟L3网络中。在一些实施例中,源计算实例位于第二虚拟层2网络中。

[0027] 在一些实施例中,该方法包括用源VNIC封装分组。在一些实施例中,该方法包括用VSRS接收和解封装分组。在一些实施例中,基于随分组接收的信息和映射表中包含的映射信息用VSRS识别第一虚拟层2网络内的目的地L2 VNIC以递送分组包括用VSRS确定映射表

不包括用于目的地计算实例的映射信息,暂停分组的VSRS转发,用VSRS向第一虚拟层2网络中的L2 VNIC广播ARP请求,该ARP请求包含目的地计算实例的IP地址,以及用VSRS从目的地计算实例的L2 VNIC接收ARP响应。在一些实施例中,L2 VNIC之一是目的地计算实例的L2 VNIC。

[0028] 在一些实施例中,该方法包括基于接收到的ARP响应来更新表。在一些实施例中,基于随分组接收的信息和映射表中包含的映射信息用VSRS识别第一虚拟层2网络内的目的地L2 VNIC以递送分组包括确定映射表包括用于目的地计算实例的映射信息,并基于映射表中包含的映射信息识别目的地L2 VNIC。在一些实施例中,在分组中嵌入与分组相关的ACL信息包括将ACL信息作为元数据存储存在分组中。在一些实施例中,用VSRS访问来自分组的ACL信息包括提取分组中包含ACL信息的元数据。

[0029] 在一些实施例中,将访问的ACL信息应用于分组包括确定ACL信息与目的地L2 VNIC无关。在一些实施例中,将访问的ACL信息应用于分组还包括经由目的地L2 VNIC将分组转发到目的地计算实例。在一些实施例中,将访问的ACL信息应用于分组包括用VSRS确定ACL信息与目的地L2 VNIC相关。在一些实施例中,将访问的ACL信息应用于分组还包括:用VSRS确定目的地L2 VNIC符合ACL信息;以及用VSRS将分组经由目的地L2 VNIC转发到目的地计算实例。

[0030] 在一些实施例中,将访问的ACL信息应用于分组还包括:用VSRS确定目的地L2 VNIC不符合ACL信息;以及VSRS丢弃分组。在一些实施例中,将访问的ACL信息应用于分组还包括用VSRS向源计算实例发送指示分组的丢弃的响应。

[0031] 本公开的一个方面涉及一种包括物理网络的系统。该物理网络包括至少一个第一处理器、网络虚拟化设备和至少一个第二处理器。至少一个处理器可以从在物理网络上实例化的虚拟网络中的源计算实例经由在物理网络上实例化的第一虚拟层2网络内的目的地L2虚拟网络接口卡(目的地L2 VNIC)向目的地计算实例发送分组。网络虚拟化设备可以实例化源VNIC。源VNIC可以评估用于分组的访问控制列表(ACL),在分组中嵌入与分组相关的ACL信息,并将分组转发到虚拟交换和路由服务(VSRS),该VSRS将第一虚拟层2网络(VLAN)与第二网络耦合。至少一个第二处理器可以实例化VSRS。VSRS可以基于随分组接收的信息和映射表中包含的映射信息来识别目的地L2 VNIC以递送分组,从分组访问ACL信息,并将访问的ACL信息应用于分组。

[0032] 在一些实施例中,将访问的ACL信息应用于分组包括确定ACL信息与目的地L2 VNIC相关,确定目的地L2 VNIC符合ACL信息,以及用VSRS经由目的地L2 VNIC将分组转发到目的地计算实例。

[0033] 本公开的一个方面涉及一种非暂态计算机可读存储介质,其存储可由一个或多个处理器执行的多个指令。当由一个或多个处理器执行时,多个指令使一个或多个处理器经由第一虚拟层2网络中的目的地L2虚拟网络接口卡(目的地L2 VNIC)将分组从虚拟网络中的源计算实例发送到目的地计算实例,用源虚拟网络接口卡(源VNIC)评估用于分组的访问控制列表(ACL),在分组中嵌入与分组相关的ACL信息,将分组转发到虚拟交换和路由服务(VSRS),该VSRS将第一虚拟层2网络(VLAN)与第二网络耦合,基于随分组接收的信息和映射表中包含的映射信息用VSRS识别第一虚拟层2网络内的目的地L2 VNIC以递送分组,用VSRS访问来自分组的ACL信息;并将访问的ACL信息应用于分组。

附图说明

[0034] 图1是分布式环境的高级图,示出了根据某些实施例的由云服务提供者基础设施托管的虚拟或覆盖云网络。

[0035] 图2描绘了根据某些实施例的CSPI内的物理网络中的物理组件的简化体系架构图。

[0036] 图3示出了根据某些实施例的CSPI内的示例布置,其中主机机器连接到多个网络虚拟化设备(NVD)。

[0037] 图4描绘了根据某些实施例的主机机器和NVD之间的连接性,用于提供I/O虚拟化以支持多租赁(tenancy)。

[0038] 图5描绘了根据某些实施例的由CSPI提供的物理网络的简化框图。

[0039] 图6是计算网络的一个实施例的示意图。

[0040] 图7是虚拟局域网(VLAN)的逻辑和硬件示意图。

[0041] 图8是多个连接的L2 VLAN的逻辑示意图。

[0042] 图9是多个连接的L2 VLAN和子网的逻辑示意图。

[0043] 图10是VLAN内的VLAN内通信和学习的一个实施例的示意图。

[0044] 图11是VLAN的实施方式视图的实施例的示意图。

[0045] 图12是图示用于VLAN内通信的过程的一个实施例的流程图。

[0046] 图13是用于VLAN内通信的过程的示意图。

[0047] 图14是图示用于虚拟L2网络中的VLAN间通信的过程的一个实施例的流程图。

[0048] 图15是用于VLAN间通信的过程的示意图。

[0049] 图16是图示用于入口分组流的过程的一个实施例的流程图。

[0050] 图17是用于入口通信的过程的示意图。

[0051] 图18是图示用于来自VLAN的出口分组流的过程的一个实施例的流程图。

[0052] 图19是用于出口分组流的过程的示意图。

[0053] 图20是图示用于延迟的访问控制列表(ACL)分类的过程的一个实施例的流程图。

[0054] 图21是图示用于ACL的早期分类的过程的一个实施例的流程图。

[0055] 图22是图示用于基于发送者的下一跳路由的过程的一个实施例的流程图。

[0056] 图23是图示用于延迟的下一跳路由的过程的一个实施例的流程图。

[0057] 图24是图示根据至少一个实施例的用于将云基础设施实现为服务系统的一种模式的框图。

[0058] 图25是图示根据至少一个实施例的用于将云基础设施实现为服务系统的另一种模式的框图。

[0059] 图26是图示根据至少一个实施例的用于将云基础设施实现为服务系统的另一种模式的框图。

[0060] 图27是图示根据至少一个实施例的用于将云基础设施实现为服务系统的另一种模式的框图。

[0061] 图28是图示根据至少一个实施例的示例计算机系统的框图。

具体实施方式

[0062] 在以下描述中,出于解释的目的,阐述了具体细节以便提供对某些实施例的透彻理解。但是,显然可以在没有这些具体细节的情况下实践各种实施例。附图和描述并非旨在限制。词语“示例性”在本文用于表示“用作示例、实例或图示”。本文描述为“示例性”的任何实施例或设计不一定被解释为优选或优于其它实施例或设计。

[0063] 示例虚拟联网体系架构

[0064] 术语云服务一般用于指由云服务提供者(CSP)使用由CSP提供的系统和基础设施(云基础设施)按需(例如,经由订阅模型)向用户或客户提供的服务。通常,构成CS的P基础设施的服务器和系统与客户自己的内部部署服务器和系统是分开的。因此,客户可以利用由CSP提供的云服务,而无需为服务购买单独的硬件和软件资源。云服务被设计为为订阅客户提供对应用和计算资源的简单、可扩展的访问,而无需客户投资购买用于提供服务的基础设施。

[0065] 有几个云服务提供者提供各种类型的云服务。有各种不同类型或模型的云服务,包括软件即服务(SaaS)、平台即服务(PaaS)、基础设施即服务(IaaS)等。

[0066] 客户可以订阅由CSP提供的一项或多项云服务。客户可以是任何实体,诸如个人、组织、企业等。当客户订阅或注册由CSP提供的服务时,将为该客户创建租赁或账户。然后,客户可以经由这个账户访问与该账户相关联的订阅的一个或多个云资源。

[0067] 如上所述,基础设施即服务(IaaS)是一种特定类型的云计算服务。在IaaS模型中,CSP提供基础设施(称为云服务提供者基础设施或CSPI),这些基础设施可以被客户使用来构建他们自己的可定制网络并部署客户资源。因此,客户的资源和网络由CSP提供的基础设施托管在分布式环境中。这与传统计算不同,在传统计算中,客户的资源和网络由客户提供的基础设施托管。

[0068] CSPI可以包括形成物理网络的互连的高性能计算资源,包括各种主机机器、存储器资源和网络资源,该物理网络也被称为基板网络或底层网络。CSPI中的资源可以散布在一个或多个数据中心上,这些数据中心可以在地理上散布在一个或多个地理地区上。虚拟化软件可以由这些物理资源执行以提供虚拟化的分布式环境。虚拟化在物理网络上创建覆盖网络(也称为基于软件的网络、软件定义的网络或虚拟网络)。CSPI物理网络为在物理网络之上创建一个或多个覆盖或虚拟网络提供了底层基础。物理网络(或基板网络或底层网络)包括物理网络设备,诸如物理交换机、路由器、计算机和主机机器等。覆盖网络是运行在物理基板网络之上的逻辑(或虚拟)网络。给定的物理网络可以支持一个或多个覆盖网络。覆盖网络通常使用封装技术来区分属于不同覆盖网络的流量。虚拟或覆盖网络也称为虚拟云网络(VCN)。虚拟网络是使用软件虚拟化技术实现的(例如,管理程序、由网络虚拟化设备(NVD)(例如,智能网卡)实现的虚拟化功能、架顶式(TOR)交换机、实现由NVD执行的一个或多个功能的智能TOR,以及其它机制)以创建可以在物理网络之上运行的网络抽象的层。虚拟网络可以采用多种形式,包括对等网络、IP网络等。虚拟网络通常是或者层3IP网络或者层2VLAN。这种虚拟或覆盖联网的方法常常被称为虚拟或覆盖层3网络。为虚拟网络开发的协议的示例包括IP-in-IP(或通用路由封装(GRE))、虚拟可扩展局域网(VXLAN—IETF RFC 7348)、虚拟专用网络(VPN)(例如,MPLS层3虚拟专用网络网络(RFC 4364))、VMware的NSX、GENEVE(通用网络虚拟化封装)等。

[0069] 对于IaaS,由CSP提供的基础设施(CSPI)可以被配置为通过公共网络(例如,互联网)提供虚拟化计算资源。在IaaS模型中,云计算服务提供者可以托管基础设施组件(例如,服务器、存储设备、网络节点(例如,硬件)、部署软件、平台虚拟化(例如,管理程序层)等)。在一些情况下,IaaS提供者还可以供应各种服务来伴随那些基础设施组件(例如,计费、监视、日志记录、安全性、负载均衡和聚类等)。因此,由于这些服务可以是策略驱动的,因此IaaS用户可以能够实现策略来驱动负载均衡以维持应用可用性和性能。CSPI提供基础设施和互补云服务的集合,这些使客户能够在高度可用的被托管的分布式环境中构建和运行广泛的应用和服务。CSPI在可从各种联网位置(诸如从客户的内部部署网络)安全访问的灵活的虚拟网络中提供高性能计算资源和能力以及存储容量。当客户订阅或注册由CSP提供的IaaS服务时,为该客户创建的租赁是CSPI内的安全且隔离的分区,客户可以在其中创建、组织和管理他们的云资源。

[0070] 客户可以使用由CSPI提供的计算、存储器和联网资源构建他们自己的虚拟网络。可以在这些虚拟网络上部署一个或多个客户资源或工作负载,诸如计算实例。例如,客户可以使用由CSPI提供的资源来构建一个或多个可定制且私有的虚拟网络,称为虚拟云网络(VCN)。客户可以在客户VCN上部署一个或多个客户资源,诸如计算实例。计算实例可以采用虚拟机、裸机实例等形式。因此,CSPI提供基础设施和互补云服务的集合,这些使客户能够在高度可用的虚拟被托管环境中构建和运行广泛的应用和服务。客户不管理或控制由CSPI提供的底层物理资源,但可以控制操作系统、存储装置和已部署的应用;并且可能对所选择的联网组件(例如,防火墙)进行有限的控制。

[0071] CSP可以提供控制台,其使客户和网络管理员能够使用CSPI资源来配置、访问和管理部署在云中的资源。在某些实施例中,控制台提供可以被用于访问和管理CSPI的基于web的用户界面。在某些实施方式中,控制台是由CSP提供的基于web的应用。

[0072] CSPI可以支持单租赁或多租赁体系架构。在单租赁体系架构中,软件(例如,应用、数据库)或硬件组件(例如,主机机器或服务器)服务于单个客户或租户。在多租赁体系架构中,软件或硬件组件服务于多个客户或租户。因此,在多租赁体系架构中,CSPI资源在多个客户或租户之间共享。在多租赁情况下,在CSPI中采取预防措施并实施保护措施,以确保每个租户的数据都是隔离的并且对其他租户保持不可见。

[0073] 在物理网络中,网络端点(“端点”,“endpoint”)是指连接到物理网络并与其所连接的网络来回通信的计算设备或系统。物理网络中的网络端点可以连接到局域网(LAN)、广域网(WAN)或其它类型的物理网络。物理网络中传统端点的示例包括调制解调器、集线器、桥接器、交换机、路由器和其它网络设备、物理计算机(或主机机器)等。物理网络中的每个物理设备都有可以被用于与设备通信的固定网络地址。这个固定网络地址可以是层2地址(例如,MAC地址)、固定的层3地址(例如,IP地址)等。在虚拟化的环境或虚拟网络中,端点可以包括各种虚拟端点,诸如由物理网络的组件托管(例如,由物理主机机器托管)的虚拟机。虚拟网络中的这些端点由覆盖地址寻址,诸如覆盖层2地址(例如,覆盖MAC地址)和覆盖层3地址(例如,覆盖IP地址)。网络覆盖通过允许网络管理者使用软件管理(例如,经由实现用于虚拟网络的控制平面的软件)在与网络端点相关联的覆盖地址周围移动来实现灵活性。因而,与物理网络不同,在虚拟网络中,可以使用网络管理软件将覆盖地址(例如,覆盖IP地址)从一个端点移动到另一个端点。由于虚拟网络建立在物理网络之上,因此虚拟网络中组

件之间的通信涉及虚拟网络和底层物理网络。为了促进此类通信,CSPI的组件被配置为学习和存储将虚拟网络中的覆盖地址映射到基板网络中的实际物理地址的映射,反之亦然。然后使用这些映射来促进通信。客户流量被封装以促进在虚拟网络中的路由。

[0074] 因而,物理地址(例如,物理IP地址)与物理网络中的组件相关联,并且覆盖地址(例如,覆盖IP地址)与虚拟或覆盖网络中的实体相关联。物理IP地址是与基板或物理网络中的物理设备(例如,网络设备)相关联的IP地址。例如,每个NVD都有相关联的物理IP地址。覆盖IP地址是与覆盖网络中的实体相关联(诸如与客户的虚拟云网络(VCN)中的计算实例相关联)的覆盖地址。各自具有其自己的私有VCN的两个不同的客户或租户可以潜在地在彼此不知情的情况下在其VCN中使用相同的覆盖IP地址。物理IP地址和覆盖IP地址都是真实IP地址的类型。这些与虚拟IP地址是分开的。虚拟IP地址通常是表示或映射到多个真实IP地址的单个IP地址。虚拟IP地址提供虚拟IP地址与多个真实IP地址之间的一对多映射。例如,负载均衡器可以使用VIP来映射或表示多个服务器,每个服务器都有其自己的真实IP地址。

[0075] 云基础设施或CSPI在物理上托管在世界上的一个或多个地区的一个或多个数据中心中。CSPI可以包括物理或基板网络中的组件和位于建立在物理网络组件之上的虚拟网络中的虚拟化的组件(例如,虚拟网络、计算实例、虚拟机等)。在某些实施例中,CSPI在领域、地区和可用性域中被组织和托管。地区通常是包含一个或多个数据中心的局部化的地理区域。地区一般彼此独立并且可以相隔很远,例如,跨越国家甚至大陆。例如,第一个地区可以在澳大利亚,另一个在日本,另一个在印度,等等。CSPI资源在地区之间划分,使得每个地区都有自己独立的CSPI资源子集。每个地区可以提供核心基础设施服务和资源的集合,诸如计算资源(例如,裸机服务器、虚拟机、容器和相关基础设施等);存储资源(例如,块卷存储装置、文件存储装置、对象存储装置、归档存储装置);网络资源(例如,虚拟云网络(VCN)、负载均衡资源、与内部部署网络的连接)、数据库资源;边缘联网资源(例如,DNS);以及访问管理和监视资源等。每个地区一般都有多条路径将其连接到领域中的其它地区。

[0076] 一般而言,应用被部署在它被最频繁使用的地区中(即,部署在与该地区相关联的基础设施上),因为使用附近的资源比使用远处的资源更快。应用也可以出于各种原因而部署在不同的地区,诸如冗余以减轻地区范围内事件(诸如大型天气系统或地震)的风险,以满足法律管辖区、税域和其它商业或社会标准等的不同要求。

[0077] 地区内的数据中心可以被进一步组织和细分为可用性域(AD)。可用性域可以与位于地区内的一个或多个数据中心对应。地区可以由一个或多个可用性域组成。在这种分布式环境中,CSPI资源或者是特定于地区的,诸如虚拟云网络(VCN),或者是特定于可用性域的,诸如计算实例。

[0078] 地区内的AD彼此隔离,具有容错能力,并且被配置为使得它们极不可能同时发生故障。这是通过AD不共享关键基础设施资源(诸如联网、物理电缆、电缆路径、电缆入口点等)来实现的,使得地区内的一个AD处的故障不太可能影响同一地区内其它AD的可用性。同一地区内的AD可以通过低时延、高带宽网络彼此连接,这使得为其它网络(例如,互联网、客户的内部部署网络等)提供高可用性连接并且在多个AD中构建复制系统以实现高可用性和灾难恢复成为可能。云服务使用多个AD来确保高可用性并防止资源故障。随着IaaS提供者提供的基础设施的增长,可以添加更多地区和AD以及附加的容量。可用性域之间的流量通

常是加密的。

[0079] 在某些实施例中,地区被分组为领域。领域是地区的逻辑集合。领域彼此隔离并且不共享任何数据。同一领域中的地区可以彼此通信,但不同领域中的地区不能。客户与CSP的租赁或账户存在于单个领域中,并且可以散布在属于该领域的一个或多个地区中。通常,当客户订阅IaaS服务时,会在领域内客户指定的地区(称为“主”地区)中为该客户创建租赁或账户。客户可以将客户的租赁扩展到领域内的一个或多个其它地区。客户无法访问不在客户的租赁所在领域中的地区。

[0080] IaaS提供者可以提供多个领域,每个领域迎合客户或用户的特定集合。例如,可以为商业客户提供商业领域。作为另一个示例,可以为特定国家为那个国家内的客户提供领域。作为又一个示例,可以为政府等提供政府领域。例如,政府领域可以迎合特定政府并且可以具有比商业领域更高的安全性级别。例如,Oracle云基础设施(Oracle Cloud Infrastructure,OCI)目前为商业地区提供领域,并且为政府云地区提供两个领域(例如,FedRAMP授权的和IL5授权的)。

[0081] 在某些实施例中,AD可以被细分成一个或多个故障域。故障域是AD内的基础设施资源的分组,以提供反亲和性。故障域允许计算实例的分布,使得这些实例不位于单个AD内的相同物理硬件上。这被称为反亲和性。故障域是指共享单点故障的硬件组件(计算机、交换机等)的集合。计算池在逻辑上被划分为故障域。因此,影响一个故障域的硬件故障或计算硬件维护事件不影响其它故障域中的实例。取决于实施例,用于每个AD的故障域的数量可以有所不同。例如,在某些实施例中,每个AD包含三个故障域。故障域充当AD内的逻辑数据中心。

[0082] 当客户订阅IaaS服务时,来自CSPI的资源被供应给客户并与客户的租赁相关联。客户可以使用这些供应的资源来构建私有网络并在这些网络上部署资源。由CSPI在云中托管的客户网络被称为虚拟云网络(VCN)。客户可以使用为客户分配的CSPI资源来设置一个或多个虚拟云网络(VCN)。VCN是虚拟或软件定义的私有网络。部署在客户的VCN中的客户资源可以包括计算实例(例如,虚拟机、裸机实例)和其它资源。这些计算实例可以表示各种客户工作负载,诸如应用、负载均衡器、数据库等。部署在VCN上的计算实例可以通过公共网络(诸如互联网)与公共可访问端点(“公共端点”)、与同一VCN或其它VCN(例如,客户的其它VCN或不属于该客户的VCN)中的其它实例、与客户的内部部署数据中心或网络以及与服务端点和其它类型的端点通信。

[0083] CSP可以使用CSPI来提供各种服务。在一些情况下,CSPI的客户本身可以像服务提供者一样起作用并使用CSPI资源提供服务。服务提供者可以暴露服务端点,该服务端点以标识信息(例如,IP地址、DNS名称和端口)为特征。客户的资源(例如,计算实例)可以通过访问由服务为特定服务暴露的服务端点来使用该特定服务。这些服务端点一般是用户可经由公共通信网络(诸如互联网)使用与端点相关联的公共IP地址公开访问的端点。可公开访问的网络端点有时也称为公共端点。

[0084] 在某些实施例中,服务提供者可以经由用于服务的端点(有时称为服务端点)来暴露服务。服务的客户然后可以使用这个服务端点来访问服务。在某些实施方式中,为服务提供的服务端点可以由打算消费该服务的多个客户访问。在其它实施方式中,可以为客户提供专用服务端点,使得只有该客户可以使用该专用服务端点来访问服务。

[0085] 在某些实施例中,当创建VCN时,它与私有覆盖无类域间路由(CIDR)地址空间相关联,该地址空间是指派给VCN的一系列私有覆盖IP地址(例如,10.0/16)。VCN包括相关联的子网、路由表和网关。VCN驻留在单个地区内,但可以跨越该区域的可用性域中的一个或多个或所有。网关是为VCN配置的虚拟接口并且启用VCN与VCN外部的一个或多个端点之间的流量的通信。可以为VCN配置一种或多种不同类型的网关,以启用去往和来自不同类型端点的通信。

[0086] VCN可以被细分为一个或多个子网络,诸如一个或多个子网。因此,子网是可以在VCN内创建的配置的单元或细分。VCN可以有一个或多个子网。VCN内的每个子网都与连续范围的覆盖IP地址(例如,10.0.0.0/24和10.0.1.0/24)相关联,这些IP地址不与该VCN中的其它子网重叠并且表示VCN的地址空间内的地址空间子集。

[0087] 每个计算实例都与虚拟网络接口卡(VNIC)相关联,这使得计算实例能够参与VCN的子网。VNIC是物理网络接口卡(NIC)的逻辑表示。一般而言,VNIC是实体(例如,计算实例、服务)与虚拟网络之间的接口。VNIC存在于子网中,具有一个或多个相关联的IP地址,以及相关联的安全性规则或策略。VNIC相当于交换机上的层2端口。VNIC附接到计算实例和VCN内的子网。与计算实例相关联的VNIC使计算实例成为VCN的子网的一部分,并且使计算实例能够与和计算实例位于同一子网上的端点、与VCN中的不同子网中的端点或与VCN外部的端点通信(例如,发送和接收分组)。因此,与计算实例相关联的VNIC确定计算实例如何与VCN内部和外部的端点连接。当创建计算实例并将其添加到VCN内的子网时,创建用于计算实例的VNIC并将其与该计算实例相关联。对于包括计算实例的集合的子网,该子网包含与计算实例的该集合对应的VNIC,每个VNIC附接到计算机实例的该集合内的计算实例。

[0088] 经由与计算实例相关联的VNIC为每个计算实例指派私有覆盖IP地址。这个私有覆盖网络IP地址在创建计算实例时指派给与计算实例相关联的VNIC并被用于路由去往和来自计算实例的流量。给定子网中的所有VNIC都使用相同的路由表、安全性列表和DHCP选项。如上所述,VCN内的每个子网都与连续范围的覆盖IP地址(例如,10.0.0.0/24和10.0.1.0/24)相关联,这些IP地址不与该VCN中的其它子网重叠并且表示VCN的地址空间内的地址空间子集。对于VCN的特定子网上的VNIC,指派给该VNIC的私有覆盖IP地址是来自为子网分配的连续范围的覆盖IP地址的地址。

[0089] 在某些实施例中,除了私有覆盖IP地址之外,计算实例还可以可选地被指派附加的覆盖IP地址,诸如例如如果在公共子网中,那么是一个或多个公共IP地址。这多个地址或者在同一个VNIC上或者在与计算实例相关联的多个VNIC上指派。但是,每个实例具有主VNIC,它是在实例启动期间创建的并且与指派给该实例的覆盖私有IP地址相关联—这个主VNIC无法被删除。称为辅助VNIC的附加VNIC可以添加到与主VNIC相同的可用性域中的现有实例。所有VNIC都在与该实例相同的可用性域中。辅助VNIC可以位于与主VNIC相同的VCN中的子网中,或者位于同一VCN或不同VCN中的不同子网中。

[0090] 如果计算实例在公共子网中,那么可以可选地为其指派公共IP地址。在创建子网时,可以将子网指定为或者公共子网或者私有子网。私有子网意味着子网中的资源(例如,计算实例)和相关联的VNIC不能具有公共覆盖IP地址。公共子网意味着子网中的资源和相关联的VNIC可以具有公共IP地址。客户可以指定子网存在于单个可用性域中或跨地区或领域中的多个可用性域。

[0091] 如上所述,VCN可以被细分成一个或多个子网。在某些实施例中,为VCN配置的虚拟路由器(VR)(称为VCN VR或简称为VR)使得VCN的子网之间能够通信。对于VCN内的子网,VR表示用于该子网的逻辑网关,它使该子网(即,该子网上的计算实例)能够与VCN内其它子网上的端点以及VCN外的其它端点通信。VCN VR是逻辑实体,其被配置为在VCN中的VNIC和与VCN相关联的虚拟网关(“网关”)之间路由流量。下文关于图1进一步描述网关。VCN VR是层3/IP层概念。在一个实施例中,对于VCN存在一个VCN VR,其中VCN VR具有潜在不受限制数量的由IP地址寻址的端口,VCN的每个子网有一个端口。以这种方式,VCN VR对于VCN VR所附接到的VCN中的每个子网具有不同的IP地址。VR还连接到为VCN配置的各种网关。在某些实施例中,用于子网的覆盖IP地址范围中的特定覆盖IP地址被保留用于该子网的VCN VR的端口。例如,考虑VCN有两个子网,相关联的地址范围分别是10.0/16和10.1/16。对于地址范围为10.0/16的VCN中的第一个子网,这个范围内的地址被保留用于该子网的VCN VR的端口。在一些情况下,范围内的第一个IP地址可以被保留用于VCN VR。例如,对于覆盖IP地址范围为10.0/16的子网,可以为该子网的VCN VR的端口保留IP地址10.0.0.1。对于地址范围为10.1/16的同一VCN中的第二个子网,VCN VR可以具有用于IP地址为10.1.0.1的第二个子网的端口。VCN VR对于VCN中的每个子网具有不同的IP地址。

[0092] 在一些其它实施例中,VCN内的每个子网可以具有其自己的相关联的VR,该VR可由子网使用与VR相关联的保留或默认IP地址寻址。例如,保留或默认IP地址可以是与该子网相关联的IP地址范围中的第一个IP地址。子网中的VNIC可以使用这个默认或保留IP地址与与子网相关联的VR通信(例如,发送和接收分组)。在这种实施例中,VR是该子网的入口/出口点。与VCN内的子网相关联的VR可以与和VCN内的其它子网相关联的其它VR通信。VR还可以与和VCN相关联的网关通信。子网的VR功能在为子网中的VNIC执行VNIC功能的一个或多个NVD上运行或由其执行。

[0093] 可以为VCN配置路由表、安全性规则和DHCP选项。路由表是用于VCN的虚拟路由表并且包括通过网关或特殊配置的实例将流量从VCN内的子网路由到VCN外的目的地的规则。可以定制VCN的路由表以控制分组如何转发/路由到VCN和从VCN路由。DHCP选项是指在实例启动时自动提供给实例的配置信息。

[0094] 为VCN配置的安全性规则表示用于VCN的覆盖防火墙规则。安全性规则可以包括入口和出口规则,并指定允许进出VCN实例的流量的类型(例如,基于协议和端口)。客户可以选择给定规则是有状态的还是无状态的。例如,客户可以通过用源CIDR 0.0.0.0/0和目的地TCP端口22设置有状态入口规则来允许来自任何地方的传入SSH流量到实例的集合。安全性规则可以使用网络安全组或安全性列表来实现。网络安全组由仅适用于该组中的资源的安全性规则的集合组成。另一方面,安全性列表包括适用于使用该安全性列表的任何子网中的所有资源的规则。可以为VCN提供带有默认安全性规则的默认安全性列表。为VCN配置的DHCP选项提供在实例启动时自动提供给VCN中的实例的配置信息。

[0095] 在某些实施例中,用于VCN的配置信息由VCN控制平面确定和存储。例如,用于VCN的配置信息可以包括关于以下的信息:与VCN相关联的地址范围、VCN内的子网和相关信息、与VCN相关联的一个或多个VR、VCN中的计算实例和相关信息、VNIC、执行与VCN相关联的各种虚拟化网络功能的NVD(例如,VNIC、VR、网关)、用于VCN的状态信息以及其它VCN相关信息。在某些实施例中,VCN分发服务向NVD发布由VCN控制平面或其部分存储的配置信息。

分发的信息可以被用于更新由NVD存储和使用的信息(例如,转发表、路由表等),以将分组转发到VCN中的计算实例或从计算实例转发分组。

[0096] 在某些实施例中,VCN和子网的创建由VCN控制平面(CP)处置并且计算实例的启动由计算控制平面处置。计算控制平面负责为计算实例分配物理资源,然后调用VCN控制平面来创建VNIC并将其附接到计算实例。VCN CP还将VCN数据映射发送到被配置为执行分组转发和路由功能的VCN数据平面。在某些实施例中,VCN CP提供负责向VCN数据平面提供更新的分发服务。VCN控制平面的示例也在图24、图25、图26和图27中描绘(参见参考标记24116、2516、2616和2716)并在下面进行描述。

[0097] 客户可以使用由CSPI托管的资源创建一个或多个VCN。部署在客户VCN上的计算实例可以与不同的端点通信。这些端点可以包括由CSPI托管的端点和CSPI外部的端点。

[0098] 图1、图2、图3、图4、图5、图24、图25、图26和图28中描绘了用于使用CSPI实现基于云的服务的各种不同体系架构,并且在下文中进行描述。图1是分布式环境100的高级图,示出了根据某些实施例的由CSPI托管的覆盖或客户VCN。图1中描绘的分布式环境包括覆盖网络中的多个组件。图1中描绘的分布式环境100仅仅是个示例,并且不旨在不当地限制要求保护的实施例的范围。许多变化、替代和修改是可能的。例如,在一些实施方式中,图1中描绘的分布式环境可以具有比图1中所示的那些更多或更少的系统或组件,可以组合两个或更多个系统,或者可以具有不同的系统配置或布置。

[0099] 如图1中描绘的示例中所示,分布式环境100包括提供服务 and 资源的CSPI 101,客户可以订阅并使用这些服务和资源来构建他们的虚拟云网络(VCN)。在某些实施例中,CSPI 101向订阅客户提供IaaS服务。CSPI 101内的数据中心可以被组织到一个或多个地区。图1中示出了一个示例地区“地区US”102。客户已经为地区102配置了客户VCN 104。客户可以在VCN 104上部署各种计算实例,其中计算实例可以包括虚拟机或裸机实例。实例的示例包括应用、数据库、负载均衡器等。

[0100] 在图1中描绘的实施例中,客户VCN 104包括两个子网,即,“子网-1”和“子网-2”,每个子网具有其自己的CIDR IP地址范围。在图1中,子网-1的覆盖IP地址范围是10.0/16,并且子网-2的地址范围是10.1/16。VCN虚拟路由器105表示用于VCN的逻辑网关,它启用VCN 104的子网之间以及与VCN外部的其它端点的通信。VCN VR 105被配置为在VCN 104中的VNIC和与VCN 104相关联的网关之间路由流量。VCN VR 105为VCN 104的每个子网提供端口。例如,VR 105可以为子网-1提供具有IP地址10.0.0.1的端口并为子网-2提供具有IP地址10.1.0.1的端口。

[0101] 可以在每个子网上部署多个计算实例,其中计算实例可以是虚拟机实例和/或裸机实例。子网中的计算实例可以由CSPI 101内的一个或多个主机机器托管。计算实例经由与计算实例相关联的VNIC参与子网。例如,如图1中所示,计算实例C1经由与计算实例相关联的VNIC成为子网-1的一部分。同样,计算实例C2经由与C2相关联的VNIC成为子网-1的一部分。以类似的方式,多个计算实例(可以是虚拟机实例或裸机实例)可以是子网-1的一部分。经由其相关联的VNIC,每个计算实例都被指派私有覆盖IP地址和介质访问控制地址(MAC地址)。例如,在图1中,计算实例C1的覆盖IP地址为10.0.0.2,MAC地址为M1,而计算实例C2的私有覆盖IP地址为10.0.0.3,MAC地址为M2。子网-1中的每个计算实例(包括计算实例C1和C2)都有到使用IP地址10.0.0.1的VCN VR 105的默认路由,该地址是用于子网-1的

VCN VR 105的端口的IP地址。

[0102] 子网-2上可以部署多个计算实例,包括虚拟机实例和/或裸机实例。例如,如图1中所示,计算实例D1和D2经由与相应计算实例相关联的VNIC成为子网-2的一部分。在图1中所示的实施例中,计算实例D1的覆盖IP地址为10.1.0.2,MAC地址为MM1,而计算实例D2的私有覆盖IP地址为10.1.0.3,MAC地址为MM2。子网-2中的每个计算实例(包括计算实例D1和D2)都有到使用IP地址10.1.0.1的VCN VR 105的默认路由,该地址是用于子网-2的VCN VR 105的端口的IP地址。

[0103] VCN A 104还可以包括一个或多个负载均衡器。例如,可以为子网提供负载均衡器,并且可以将负载均衡器配置为跨子网上的多个计算实例负载均衡流量。还可以提供负载均衡器以在VCN中跨子网负载均衡流量。

[0104] 部署在VCN 104上的特定计算实例可以与各种不同的端点通信。这些端点可以包括由CSPI 200托管的端点和CSPI 200外部的端点。由CSPI 101托管的端点可以包括:与特定计算实例位于同一子网上的端点(例如,子网-1中两个计算实例之间的通信);位于不同子网上但位于同一VCN内的端点(例如,子网-1中的计算实例与子网-2中的计算实例之间的通信);同一地区中不同VCN中的端点(例如,子网-1中的计算实例与同一地区106或110中的VCN中的端点之间的通信,子网-1中的计算实例与同一地区中服务网点110中的端点之间的通信);或不同地区中的VCN中的端点(例如,子网-1中的计算实例与不同地区108中的VCN中的端点之间的通信)。由CSPI 101托管的子网中的计算实例也可以与不由CSPI 101托管(即,在CSPI 101外部)的端点通信。这些外部端点包括客户的内部部署网络116中的端点、其它远程云托管网络118中的端点、可经由公共网络(诸如互联网)访问的公共端点114及其它端点。

[0105] 使用与源计算实例和目的地计算实例相关联的VNIC促进同一子网上的计算实例之间的通信。例如,子网-1中的计算实例C1可能想要将分组发送到子网-1中的计算实例C2。对于源自源计算实例且其目的地是同一子网中的另一个计算实例的分组,该分组首先由与源计算实例相关联的VNIC处理。由与源计算实例相关联的VNIC执行的处理可以包括从分组报头确定分组的目的地信息、识别为与源计算实例相关联的VNIC配置的任何策略(例如,安全性列表)、确定用于分组的下一跳、根据需要执行任何分组封装/解封装功能,然后将分组转发/路由到下一跳,目的是促进分组到其预期目的地的通信。当目的地计算实例与源计算实例位于同一子网中时,与源计算实例相关联的VNIC被配置为识别与目的地计算实例相关联的VNIC并将分组转发到该VNIC进行处理。然后执行与目的地计算实例相关联的VNIC并将分组转发到目的地计算实例。

[0106] 对于要从子网中的计算实例传送到同一VCN中不同子网中的端点的分组,通过与源和目的地计算实例相关联的VNIC以及VCN VR来促进通信。例如,如果图1中子网-1中的计算实例C1想要将分组发送到子网-2中的计算实例D1,那么该分组首先由与计算实例C1相关联的VNIC处理。与计算实例C1相关联的VNIC被配置为使用默认路由或VCN VR的端口10.0.0.1将分组路由到VCN VR 105。VCN VR 105被配置为使用端口10.1.0.1将分组路由到子网-2。然后,与D1相关联的VNIC接收并处理分组并且VNIC将分组转发到计算实例D1。

[0107] 对于要从VCN 104中的计算实例传送到VCN 104外部的端点的分组,通信由与源计算实例相关联的VNIC、VCN VR 105和与VCN 104相关联的网关促进。一种或多种类型的网关

可以与VCN 104相关联。网关是VCN和另一个端点之间的接口,其中该另一个端点在VCN外部。网关是层3/IP层概念并且使VCN能够与VCN外部的端点通信。因此,网关促进VCN与其它VCN或网络之间的流量流动。可以为VCN配置各种不同类型的网关以促进与不同类型的端点的不同类型的通信。取决于网关,通信可以通过公共网络(例如,互联网)或通过私有网络进行。各种通信协议可以用于这些通信。

[0108] 例如,计算实例C1可能想要与VCN 104外部的端点通信。分组可以首先由与源计算实例C1相关联的VNIC处理。VNIC处理确定分组的目的地在C1的子网-1外部。与C1相关联的VNIC可以将分组转发到用于VCN 104的VCN VR 105。VCN VR 105然后处理分组,并且作为处理的一部分,基于分组的目的地,确定与VCN 104相关联的特定网关作为用于分组的下一跳。然后,VCN VR 105可以将分组转发到特定的识别出的网关。例如,如果目的地是客户的内部部署网络内的端点,那么分组可以由VCN VR 105转发到为VCN 104配置的动态路由网关(DRG)网关122。分组然后可以从网关转发到下一跳,以促进分组到其最终预期目的地的传送。

[0109] 可以为VCN配置各种不同类型的网关。可以为VCN配置的网关的示例在图1中描绘并且在下文进行描述。与VCN相关联的网关的示例也在图24、图25、图26和图27中描绘(例如,通过附图标记2434、2436、2438、2534、2536、2538、2634、2636、2638、2734、2736和2738引用的网关)并且如下所述。如图1中描绘的实施例中所示,动态路由网关(DRG) 122可以添加到客户VCN 104或与客户VCN 104相关联,并为客户VCN 104与另一个端点之间的私有网络流量通信提供路径,其中该另一个端点可以是客户的内部部署网络116、CSPI 101的不同地区中的VCN 108,或者不是由CSPI 101托管的其它远程云网络118。客户内部部署网络116可以是使用客户的资源构建的客户网络或客户数据中心。对客户内部部署网络116的访问一般是非常受限的。对于同时具有客户内部部署网络116和由CSPI 101在云中部署或托管的一个或多个VCN 104的客户,客户可能想要他们的内部部署网络116和他们的基于云的VCN 104能够彼此沟通。这使客户能够构建扩展的混合环境,包括由CSPI 101托管的客户的VCN 104和他们的内部部署网络116。DRG 122启用这种此通信。为了启用此类通信,设置通信通道124,其中信道的一个端点位于客户内部部署网络116中,而另一个端点位于CSPI 101中并连接到客户VCN 104。通信信道124可以通过公共通信网络(诸如互联网)或私有通信网络。可以使用各种不同的通信协议,诸如公共通信网络(诸如互联网)上的IPsec VPN技术、使用私有网络而不是公共网络的Oracle的FastConnect技术等。客户内部部署网络116中形成通信信道124的一个端点的设备或装备被称为客户现场装备(CPE),诸如图1中描绘的CPE 126。在CSPI 101侧,端点可以是执行DRG 122的主机机器。

[0110] 在某些实施例中,可以将远程对等连接(RPC)添加到DRG,这允许客户将一个VCN与不同地区中的另一个VCN对等。使用这种RPC,客户VCN 104可以使用DRG 122与另一个地区中的VCN 108连接。DRG 122还可以被用于与不由CSPI 101托管的其它远程云网络118(诸如Microsoft Azure云、亚马逊AWS云等)通信。

[0111] 如图1中所示,可以为客户VCN 104配置互联网网关(IGW) 120,其使VCN 104上的计算实例能够与可通过公共网络(诸如互联网)访问的公共端点114通信。IGW 120是将VCN连接到诸如互联网之类的公共网络的网关。IGW 120使VCN(诸如VCN 104)内的公共子网(其中公共子网中的资源具有公共覆盖IP地址)能够直接访问公共网络114(诸如互联网)上的公

共端点112。使用IGW 120,可以从VCN 104内的子网或从互联网发起连接。

[0112] 可以为客户的VCN 104配置网络地址转换(NAT)网关128,并使客户的VCN中没有专用公共覆盖IP地址的云资源能够访问互联网,并且在不暴露那些资源的情况下这样做以指引传入的互联网连接(例如,L4-L7连接)。这使VCN内的私有子网(诸如VCN 104中的私有子网-1)能够私有访问互联网上的公共端点。在NAT网关中,只能从私有子网发起到公共互联网的连接,而不能从互联网发起到私有子网的连接。

[0113] 在某些实施例中,可以为客户VCN 104配置服务网关(SGW)126,并为VCN 104和服务网络110中支持的服务端点之间的私有网络流量提供路径。在某些实施例中,服务网络110可以由CSP提供并且可以提供各种服务。这种服务网络的示例是Oracle的服务网络,它提供可供客户使用的各种服务。例如,客户VCN 104的私有子网中的计算实例(例如,数据库系统)可以将数据备份到服务端点(例如,对象存储装置,Object Storage),而无需公共IP地址或访问互联网。在某些实施例中,VCN可以只有一个SGW,并且连接只能从VCN内的子网发起,而不能从服务网络110发起。如果VCN与另一个对等,那么另一个VCN中的资源通常不能访问SGW。用FastConnect或VPNConnect连接到VCN的内部部署网络中的资源也可以使用为该VCN配置的服务网关。

[0114] 在某些实施方式中,SGW 126使用服务无类域间路由(CIDR)标签的概念,该标签是表示用于感兴趣的服务或服务组的所有地区性公共IP地址范围的字符串。客户在他们配置SGW和相关路由规则以控制到服务的流量时使用服务CIDR标签。如果服务的公共IP地址在未来发生改变,那么客户可以在配置安全性规则时可选地使用它,而无需调整它们。

[0115] 本地对等网关(LPG)132是可以添加到客户VCN 104并使VCN 104能够与同一地区中的另一个VCN对等的网关。对等是指VCN使用私有IP地址进行通信,流量无需穿过公共网络(诸如互联网)或无需通过客户的内部部署网络116路由流量。在优选实施例中,VCN对于它建立的每个对等体具有单独LPG。本地对等或VCN对等是用于在不同应用或基础设施管理功能之间建立网络连接性的常见做法。

[0116] 服务提供者(诸如服务网络110中的服务的提供者)可以使用不同的访问模型提供对服务的访问。根据公共访问模型,服务可以作为可由客户VCN中的计算实例经由公共网络(诸如互联网)公开访问的公共端点暴露,和/或可以经由SGW 126私有访问。根据特定的私有访问模型,服务可作为客户的VCN中的私有子网中的私有IP端点进行访问。这被称为私有端点(PE)访问并且使服务提供者能够将其服务作为客户的私有网络中的实例暴露。私有端点资源表示客户的VCN内的服务。每个PE在客户的VCN中由客户选择的子网中表现为VNIC(称为PE-VNIC,具有一个或多个私有IP)。因此,PE提供了使用VNIC在私有客户VCN子网中呈现服务的方式。由于端点作为VNIC暴露,因此与VNIC相关联的所有特征(诸如路由规则、安全性列表等)现在都可以用于PE VNIC。

[0117] 服务提供者可以注册他们的服务以启用通过PE的访问。提供者可以将策略与服务相关联,这限制了服务对客户租赁的可见性。提供者可以在单个虚拟IP地址(VIP)下注册多个服务,尤其是对于多租户服务。可以有表示相同服务的多个这样的私有端点(在多个VCN中)。

[0118] 私有子网中的计算实例然后可以使用PE VNIC的私有IP地址或服务DNS名称来访问服务。客户VCN中的计算实例可以通过将流量发送到客户VCN中的PE的私有IP地址来访问

该服务。私有访问网关 (PAGW) 130是可以附接到服务提供者VCN (例如,服务网络110中的VCN) 的网关资源,该服务提供者VCN充当来自/去往客户子网私有端点的所有流量的入口/出口点。PAGW 130使提供者能够在不利用其内部IP地址资源的情况下扩展PE连接的数量。提供者只需要为在单个VCN中注册的任何数量的服务配置一个PAGW。提供者可以将服务表示为一个或多个客户的多个VCN中的私有端点。从客户的角度来看,PE VNIC不是附接到客户的实例,而是看起来附接到客户希望与之交互的服务。去往私有端点的流量经由PAGW 130路由到服务。这些被称为客户到服务私有连接 (C2S连接)。

[0119] 通过允许流量流经FastConnect/IPsec链接和客户VCN中的私有端点,PE概念还可以被用于将用于服务的私有访问扩展到客户的内部署网络和数据中心。通过允许流量在LPG 132和客户的VCN中的PE之间流动,还可以将服务的私有访问扩展到客户的对等VCN。

[0120] 客户可以在子网级别控制VCN中的路由,因此客户可以指定客户的VCN (诸如VCN 104) 中的哪些子网使用每个网关。VCN的路由表被用于决定是否允许流量通过特定网关离开VCN。例如,在特定实例中,用于客户VCN 104内的公共子网的路由表可以通过IGW 120发送非本地流量。用于同一客户VCN 104内的私有子网的路由表可以通过SGW 126发送以CSP服务为目的地的流量。所有剩余流量可以经由NAT网关128发送。路由表仅控制流出VCN的流量。

[0121] 与VCN相关联的安全性列表被用于控制经由入站连接经由网关进入VCN的流量。子网中的所有资源都使用相同的路由表和安全性列表。安全性列表可以被用于控制允许进出VCN的子网中的实例的特定类型的流量。安全性列表规则可以包括入口 (入站) 和出口 (出站) 规则。例如,入口规则可以指定允许的源地址范围,而出口规则可以指定允许的目的地地址范围。安全性规则可以指定特定的协议 (例如,TCP、ICMP)、特定的端口 (例如,用于SSH的22、用于Windows RDP的3389) 等。在某些实施方式中,实例的操作系统可以强制执行自己的符合安全性列表规则的防火墙规则。规则可以是有状态的 (例如,跟踪连接并自动允许响应,而没有针对响应流量的明确安全性列表规则) 或无状态的。

[0122] 来自客户VCN的访问 (即,通过部署在VCN 104上的资源或计算实例) 可以被分类为公开访问、私有访问或专用访问。公开访问是指使用公共IP地址或NAT来访问公共端点的访问模型。私有访问使VCN 104中具有私有IP地址的客户工作负载 (例如,私有子网中的资源) 能够访问服务,而无需穿越诸如互联网之类的公共网络。在某些实施例中,CSPI 101使具有私有IP地址的客户VCN工作负载能够使用服务网关访问服务 (的公共服务端点)。因此,服务网关通过在客户的VCN和驻留在客户的私有网络外部的服务的公共端点之间建立虚拟链接来提供私有访问模型。

[0123] 此外,CSPI可以使用诸如FastConnect公开对等之类的技术提供专用公开访问,其中客户内部部署实例可以使用FastConnect连接访问客户VCN中的一个或多个服务,而无需遍历诸如互联网之类的公共网络。CSPI还可以使用FastConnect私有对等提供专用私有访问,其中具有私有IP地址的客户内部部署实例可以使用FastConnect连接访问客户的VCN工作负载。FastConnect是对使用公共互联网将客户的内部部署网络连接到CSPI及其服务的网络连接性替代方案。与基于互联网的连接相比,FastConnect提供了简单、灵活且经济的方式来创建具有更高带宽选项的专用和私有连接以及更可靠和一致的网络体验。

[0124] 图1和上面的伴随描述描述了示例虚拟网络中的各种虚拟化的组件。如上所述,虚

拟网络建立在底层物理或基板网络上。图2描绘了根据某些实施例的为虚拟网络提供底层的CSPI 200内的物理网络中的物理组件的简化体系架构图。如图所示,CSPI 200提供分布式环境,包括由云服务提供者(CSP)提供的组件和资源(例如,计算、存储器和网络资源)。这些组件和资源被用于向订阅客户(即,已订阅由CSP提供的一个或多个服务的客户)提供云服务(例如,IaaS服务)。基于客户订阅的服务,为客户供应CSPI 200的资源的子集(例如,计算、存储器和网络资源)。然后,客户可以使用由CSPI 200提供的物理计算、存储器和联网资源构建其自己的基于云的(即,CSPI托管的)可定制且私有虚拟网络。如前面所指示的,这些客户网络被称为虚拟云网络(VCN)。客户可以在这些客户VCN上部署一个或多个客户资源,诸如计算实例。计算实例可以是虚拟机、裸机实例等形式。CSPI 200提供基础设施和互补云服务的集合,这使客户能够在高度可用的托管环境中构建和运行广泛的应用和服务。

[0125] 在图2中描绘的示例实施例中,CSPI 200的物理组件包括一个或多个物理主机机器或物理服务器(例如,202、206、208)、网络虚拟化设备(NVD)(例如,210、212)、架顶式(TOR)交换机(例如,214、216)和物理网络(例如,218),以及物理网络218中的交换机。物理主机机器或服务器可以托管和执行参与VCN的一个或多个子网的各种计算实例。计算实例可以包括虚拟机实例和裸机实例。例如,图1中描绘的各种计算实例可以由图2中描绘的物理主机机器托管。VCN中的虚拟机计算实例可以由一个主机机器或多个不同的主机机器执行。物理主机机器也可以托管虚拟主机机器、基于容器的主机或功能等。图1中描绘的VNIC和VCN VR可以由图2中描绘的NVD执行。图1中描绘的网关可以由图2中描述的主机机器和/或NVD执行。

[0126] 主机机器或服务器可以执行在主机机器上创建和启用虚拟化的环境的管理程序(也称为虚拟机监视器或VMM)。虚拟化或虚拟化的环境促进了基于云的计算。一个或多个计算实例可以由主机机器上的管理程序在该主机机器上创建、执行和管理。主机机器上的管理程序使主机机器的物理计算资源(例如,计算、存储器和网络资源)能够在由主机机器执行的各种计算实例之间共享。

[0127] 例如,如图2中所描绘的,主机机器202和208分别执行管理程序260和266。这些管理程序可以使用软件、固件或硬件或其组合来实现。通常,管理程序是位于主机机器的操作系统(OS)之上的进程或软件层,OS进而在主机机器的硬件处理器上执行。管理程序通过使主机机器的物理计算资源(例如,诸如处理器/核心之类的处理资源、存储器资源、网络资源)能够在由主机机器执行的各种虚拟机计算实例之间共享来提供虚拟化的环境。例如,在图2中,管理程序260可以位于主机机器202的OS之上并使主机机器202的计算资源(例如,处理、存储器和网络资源)能够在由主机机器202执行的计算实例(例如,虚拟机)之间共享。虚拟机可以具有其自己的操作系统(称为访客操作系统),其可以与主机机器的OS相同或不同。由主机机器执行的虚拟机的操作系统可以与由同一主机机器执行的另一个虚拟机的操作系统相同或不同。因此,管理程序使多个操作系统能够被同时执行,同时共享主机机器的相同计算资源。图2中描绘的主机机器可以具有相同或不同类型的管理程序。

[0128] 计算实例可以是虚拟机实例或裸机实例。在图2中,主机机器202上的计算实例268和主机机器208上的计算实例274是虚拟机实例的示例。主机机器206是提供给客户的裸机实例的示例。

[0129] 在某些情况下,可以向单个客户供应整个主机机器,并且由该主机机器托管的一

个或多个计算实例(或者虚拟机或者裸机实例)所有都属于同一客户。在其它情况下,主机机器可以在多个客户(即,多个租户)之间共享。在这种多租赁场景中,主机机器可以托管属于不同客户的虚拟机计算实例。这些计算实例可以是不同客户的不同VCN的成员。在某些实施例中,裸机计算实例由没有管理程序的裸机服务器托管。当供应裸机计算实例时,单个客户或租户维护对托管该裸机实例的主机机器的物理CPU、存储器和网络接口的控制,并且主机机器不与其他客户或租户共享。

[0130] 如前所述,作为VCN的一部分的每个计算实例都与使该计算实例能够成为VCN的子网的成员的VNIC相关联。与计算实例相关联的VNIC促进去往和来自计算实例的分组或帧的通信。VNIC在创建计算实例时与该计算实例相关联。在某些实施例中,对于由主机机器执行的计算实例,与该计算实例相关联的VNIC由连接到主机机器的NVD执行。例如,在图2中,主机机器202执行与VNIC 276相关联的虚拟机计算实例268,并且VNIC 276由连接到主机机器202的NVD 210执行。作为另一个示例,由主机机器206托管的裸机实例272与由连接到主机机器206的NVD 212执行的VNIC 280相关联。作为又一个示例,VNIC 284与由主机机器208执行的计算实例274相关联,并且VNIC 284由连接到主机机器208的NVD 212执行。

[0131] 对于由主机机器托管的计算实例,连接到该主机机器的NVD还执行与计算实例是其成员的VCN对应的VCN VR。例如,在图2中描绘的实施例中,NVD 210执行与计算实例268是其成员的VCN对应的VCN VR 277。NVD 212还可以执行与VCN对应的一个或多个VCN VR 283,该VCN与由主机机器206和208托管的计算实例对应。

[0132] 主机机器可以包括使主机机器能够连接到其它设备的一个或多个网络接口卡(NIC)。主机机器上的NIC可以提供使主机机器能够通信连接到另一个设备的一个或多个端口(或接口)。例如,主机机器可以使用在主机机器上和NVD上提供的一个或多个端口(或接口)连接到NVD。主机机器也可以连接到其它设备(诸如另一个主机机器)。

[0133] 例如,在图2中,主机机器202使用链路220连接到NVD 210,链路220在由主机机器202的NIC 232提供的端口234与NVD 210的端口236之间延伸。主机机器206使用链路224连接到NVD 212,链路224在由主机机器206的NIC 244提供的端口246与NVD 212的端口248之间延伸。主机机器208使用链路226连接到NVD 212,链路226在由主机机器208的NIC 250提供的端口252与NVD 212的端口254之间延伸。

[0134] NVD进而经由通信链路连接到架顶式(TOR)交换机,这些交换机连接到物理网络218(也称为交换机架构)。在某些实施例中,主机机器和NVD之间以及NVD和TOR交换机之间的链路是以太网链路。例如,在图2中,NVD 210和212分别使用链路228和230连接到TOR交换机214和216。在某些实施例中,链路220、224、226、228和230是以太网链路。连接到TOR的主机机器和NVD的集合有时称为机架(rack)。

[0135] 物理网络218提供使TOR交换机能够彼此通信的通信架构。物理网络218可以是多层网络。在某些实施方式中,物理网络218是交换机的多层Clos网络,其中TOR交换机214和216表示多层和多节点物理交换网络218的叶级节点。不同的Clos网络配置是可能的,包括但不限于2层网络、3层网络、4层网络、5层网络和一般的“n”层网络。Clos网络的示例在图5中描绘并在下面进行描述。

[0136] 主机机器和NVD之间可能存在各种不同的连接配置,诸如一对一配置、多对一配置、一对多配置等。在一对一配置实施方式中,每个主机机器连接到自己单独的NVD。例如,

在图2中,主机机器202经由主机机器202的NIC 232连接到NVD 210。在多对一配置中,多个主机机器连接到一个NVD。例如,在图2中,主机机器206和208分别经由NIC 244和250连接到同一个NVD 212。

[0137] 在一对多配置中,一个主机机器连接到多个NVD。图3示出了CSPI 300内的示例,其中主机机器连接到多个NVD。如图3中所示,主机机器302包括网络接口卡(NIC) 304,其包括多个端口306和308。主机机器300经由端口306和链路320连接到第一NVD 310,并经由端口308和链路322连接到第二NVD 312。端口306和308可以是以以太网端口并且主机机器302与NVD 310和312之间的链路320和322可以是以以太网链路。NVD 310进而连接到第一TOR交换机314并且NVD 312连接到第二TOR交换机316。NVD 310和312与TOR交换机314和316之间的链路可以是以以太网链路。TOR交换机314和316表示多层物理网络318中的层0交换设备。

[0138] 图3中描绘的布置提供了两条单独的物理网络路径,从物理交换机网络318到主机机器302:第一条路径穿过TOR交换机314到NVD 310再到主机机器302,第二条路径穿过TOR交换机316到NVD 312再到主机机器302。单独的路径提供了主机机器302的增强的可用性(称为高可用性)。如果路径(例如,路径之一中的链路断开)或设备(例如,特定NVD未运行)之一出现问题,那么另一条路径可以被用于与主机机器302之间的通信。

[0139] 在图3中描绘的配置中,主机机器使用由主机机器的NIC提供的两个不同端口连接到两个不同的NVD。在其它实施例中,主机机器可以包括使主机机器能够连接到多个NVD的多个NIC。

[0140] 返回去参考图2,NVD是执行一个或多个网络和/或存储虚拟化功能的物理设备或组件。NVD可以是具有一个或多个处理单元(例如,CPU、网络处理单元(NPU)、FPGA、分组处理管线等)、存储器(包括高速缓存)和端口的任何设备。各种虚拟化功能可以由NVD的一个或多个处理单元执行的软件/固件来执行。

[0141] NVD可以以各种不同的形式实现。例如,在某些实施例中,NVD被实现为称为smartNIC的接口卡或具有板载嵌入式处理器的智能NIC。smartNIC是独立于主机机器上的NIC的设备。在图2中,NVD 210和212可以被实现为分别连接到主机机器202以及主机机器206和208的smartNIC。

[0142] 但是,smartNIC只是NVD实施方式的一个示例。各种其它实施方式是可能的。例如,在一些其它实施方式中,NVD或由NVD执行的一个或多个功能可以结合到一个或多个主机机器、一个或多个TOR交换机以及CSPI 200的其它组件中或由其执行。例如,NVD可以在主机机器中实施,其中由NVD执行的功能由主机机器执行。作为另一个示例,NVD可以是TOR交换机的一部分,或者TOR交换机可以被配置为执行由NVD执行的功能,这使TOR交换机能够执行用于公共云的各种复杂分组转换。执行NVD的功能的TOR有时被称为智能TOR。在向客户提供虚拟机(VM)实例而非裸机(BM)实例的其它实施方式中,由NVD执行的功能可以在主机机器的管理程序内部实现。在一些其它实施方式中,NVD的功能中的一些可以卸载到在一组主机机器上运行的集中式服务。

[0143] 在某些实施例中,诸如当实现为如图2中所示的smartNIC时,NVD可以包括使其能够连接到一个或多个主机机器以及一个或多个TOR交换机的多个物理端口。NVD上的端口可以被分类为面向主机的端口(也称为“南端口”)或面向网络或面向TOR的端口(也称为“北端口”)。NVD的面向主机的端口是用于将NVD连接到主机机器的端口。图2中面向主机的端口的

示例包括NVD 210上的端口236以及NVD 212上的端口248和254。NVD的面向网络的端口是用于将NVD连接到TOR交换机的端口。图2中面向网络的端口的示例包括NVD 210上的端口256和NVD 212上的端口258。如图2中所示,NVD 210使用从NVD 210的端口256延伸到TOR交换机214的链路228连接到TOR交换机214。类似地,NVD 212使用从NVD 212的端口258延伸到TOR交换机216的链路230连接到TOR交换机216。

[0144] NVD经由面向主机的端口从主机机器接收分组和帧(例如,由主机机器托管的计算实例生成的分组和帧),并且在执行必要的分组处理之后,可以经由NVD的面向网络的端口将分组和帧转发到TOR交换机。NVD可以经由NVD的面向网络的端口从TOR交换机接收分组和帧,并且在执行必要的分组处理之后,可以经由NVD的面向主机的端口将分组和帧转发到主机机器。

[0145] 在某些实施例中,在NVD和TOR交换机之间可以有多个端口和相关联的链路。可以聚合这些端口和链路以形成多个端口或链路的链路聚合体组(称为LAG)。链路聚合允许将两个端点之间(例如,NVD和TOR交换机之间)的多条物理链路视为单条逻辑链路。给定LAG中的所有物理链路都可以以相同的速度在全双工模式下操作。LAG有助于增加两个端点之间的连接的带宽和可靠性。如果LAG中的物理链路之一发生故障,那么流量将动态且透明地重新指派到LAG中的其它物理链路之一。聚合的物理链路比每条单独的链路递送更高的带宽。与LAG相关联的多个端口被视为单个逻辑端口。流量可以在LAG的多条物理链路之间进行负载平衡。可以在两个端点之间配置一个或多个LAG。这两个端点可以位于NVD和TOR交换机之间、主机机器和NVD之间,等等。

[0146] NVD实现或执行网络虚拟化功能。这些功能由NVD执行的软件/固件执行。网络虚拟化功能的示例包括但不限于:分组封装和解封装功能;用于创建VCN网络的功能;用于实现网络策略的功能,诸如VCN安全性列表(防火墙)功能性;促进分组路由和转发到VCN中的计算实例和从该计算实例转发的功能;等等。在某些实施例中,在接收到分组后,NVD被配置为执行分组处理流水线以处理分组并确定如何转发或路由分组。作为这个分组处理管线的一部分,NVD可以执行与覆盖网络相关联的一个或多个虚拟功能,诸如执行与VCN中的计算实例相关联的VNIC、执行与VCN相关联的虚拟路由器(VR)、分组的封装和解封装以促进虚拟网络中的转发或路由、某些网关(例如,本地对等网关)的执行、安全性列表的实施方式、网络安全组、网络地址转换(NAT)功能性(例如,逐个主机地将公共IP转换到私有IP)、节流功能和其它功能。

[0147] 在某些实施例中,NVD中的分组处理数据路径可以包括多个分組管线,每个管线由一系列分组变换级组成。在某些实施方式中,在接收到分组后,分组被解析并分类到单个管线。然后以线性方式处理分组,一个级接一个级,直到分组或者被丢弃或者通过NVD的接口发送出去。这些级提供基本的功能性分组处理构建块(例如,验证报头、强制执行节流、插入新的层2报头、强制执行L4防火墙、VCN封装/解封装等),以便可以通过组合现有级来构建新的管线,并且可以通过创建新级并将它们插入现有管线来添加新的功能性。

[0148] NVD可以执行与VCN的控制平面和数据平面对应的控制平面和数据平面功能。VCN控制平面的示例也在图24、图25、图26和图27中描绘(参见参考标记2416、2516、2616和2716)并在下文中描述。VCN数据平面的示例在图24、图25、图26和图27中描绘(参见参考标记2418、2518、2618和2718)并在下文中描述。控制平面功能包括用于配置控制数据如何被

转发的网络的功能(例如,设置路由和路由表、配置VNIC等)。在某些实施例中,提供了VCN控制平面,它集中计算所有覆盖到基板映射并将它们发布到NVD和虚拟网络边缘设备(诸如各种网关,诸如DRG、SGW、IGW等)。防火墙规则也可以使用相同的机制发布。在某些实施例中,NVD仅获得与该NVD相关的映射。数据平面功能包括基于使用控制平面设置的配置来实际路由/转发分组的功能。VCN数据平面是通过在客户的网络分组穿过基板网络之前对其进行封装来实现的。封装/解封装功能性在NVD上实现。在某些实施例中,NVD被配置为拦截进出主机机器的所有网络分组并执行网络虚拟化功能。

[0149] 如上面所指示的,NVD执行各种虚拟化功能,包括VNIC和VCN VR。NVD可以执行与连接到VNIC的一个或多个主机机器托管的计算实例相关联的VNIC。例如,如图2中所描绘的,NVD 210执行与由连接到NVD 210的主机机器202托管的计算实例268相关联的VNIC 276的功能性。作为另一个示例,NVD 212执行与由主机机器206托管的裸机计算实例272相关联的VNIC 280,并执行与由主机机器208托管的计算实例274相关联的VNIC 284。主机机器可以托管属于不同VCN(属于不同客户)的计算实例,并且连接到主机机器的NVD可以执行与计算实例对应的VNIC(即,执行VNIC相关功能性)。

[0150] NVD还执行与计算实例的VCN对应的VCN虚拟路由器。例如,在图2中描绘的实施例中,NVD 210执行与计算实例268所属的VCN对应的VCN VR 277。NVD 212执行与由主机机器206和208托管的计算实例所属的一个或多个VCN对应的一个或多个VCN VR 283。在某些实施例中,与该VCN对应的VCN VR由连接到托管属于该VCN的至少一个计算实例的主机机器的所有NVD执行。如果主机机器托管属于不同VCN的计算实例,那么连接到该主机机器的NVD可以执行与那些不同VCN对应的VCN VR。

[0151] 除了VNIC和VCN VR之外,NVD还可以执行各种软件(例如,守护进程)并且包括一个或多个硬件组件,这些组件促进由NVD执行的各种网络虚拟化功能。为了简单起见,这些各种组件被分组在一起作为图2中所示的“分组处理组件”。例如,NVD 210包括分组处理组件286并且NVD 212包括分组处理组件288。例如,用于NVD的分组处理组件可以包括分组处理器,该分组处理器被配置为与NVD的端口和硬件接口交互以监视由NVD接收和使用NVD传送的所有分组并存储网络信息。网络信息例如可以包括识别由NVD处置的不同网络流的网络流信息和每个流信息(例如,每个流统计信息)。在某些实施例中,网络流信息可以以每个VNIC为基础来存储。分组处理器可以执行逐分组的操纵以及实现有状态的NAT和L4防火墙(FW)。作为另一个示例,分组处理组件可以包括被配置为将由NVD存储的信息复制到一个或多个不同的复制目标存储库的复制代理。作为又一个示例,分组处理组件可以包括被配置为执行NVD的日志记录功能的日志记录代理。分组处理组件还可以包括用于监视NVD的性能和健康状况并且还可能监视连接到NVD的其它组件的状态和健康状况的软件。

[0152] 图1示出了示例虚拟或覆盖网络的组件,包括VCN、VCN内的子网、部署在子网上的计算实例、与计算实例相关联的VNIC、用于VCN的VR以及为VCN配置的网关的集合。图1中描绘的覆盖组件可以由图2中描绘的物理组件中的一个或多个执行或托管。例如,VCN中的计算实例可以由图2中描绘的一个或多个主机机器执行或托管。对于由主机机器托管的计算实例,与该计算实例相关联的VNIC通常由连接到该主机机器的NVD执行(即,VNIC功能性由连接到该主机机器的NVD提供)。用于VCN的VCN VR功能由连接到托管或执行作为该VCN的一部分的计算实例的主机机器的所有NVD执行。与VCN相关联的网关可以由一种或多种不同类

型的NVD执行。例如,某些网关可以由smartNIC执行,而其它网关可以由一个或多个主机机器或NVD的其它实施方式执行。

[0153] 如上所述,客户VCN中的计算实例可以与各种不同的端点通信,其中端点可以与源计算实例在同一子网内、在不同子网中但与源计算实例在同一VCN内,或者与位于源计算实例的VCN外部的端点通信。使用与计算实例相关联的VNIC、VCN VR以及与VCN相关联的网关来促进这些通信。

[0154] 对于VCN中同一子网上的两个计算实例之间的通信,使用与源和目的地计算实例相关联的VNIC促进通信。源和目的地计算实例可以由同一主机机器或不同的主机机器托管。源自源计算实例的分组可以从托管源计算实例的主机机器转发到连接到该主机机器的NVD。在NVD上,使用分组处理管线处理分组,这可以包括与源计算实例相关联的VNIC的执行。由于用于分组的目的地端点位于同一子网内,因此与源计算实例相关联的VNIC的执行导致分组被转发到执行与目的地计算实例相关联的VNIC的NVD,然后NVD处理分组并将其转发到目的地计算实例。与源和目的地计算实例相关联的VNIC可以在同一个NVD上执行(例如,当源和目的地计算实例由同一个主机机器托管时)或在不同的NVD上执行(例如,当源和目的地计算实例由连接到不同NVD的不同主机机器托管时)。VNIC可以使用由NVD存储的路由/转发表来确定用于分组的下一跳。

[0155] 对于要从子网中的计算实例传送到同一VCN中不同子网中的端点的分组,源自源计算实例的分组从托管源计算实例的主机机器传送到连接到该主机机器的NVD。在NVD上,使用分组处理管线处理分组,这可以包括一个或多个VNIC以及与VCN相关联的VR的执行。例如,作为分组处理管线的一部分,NVD执行或调用与和源计算实例相关联的VNIC的功能性(也称为执行VNIC)。由VNIC执行的功能性可以包括查看分组上的VLAN标识符。由于分组的目的地在子网外部,因此接下来由NVD调用和执行VCN VR功能性。然后,VCN VR将分组路由到执行与目的地计算实例相关联的VNIC的NVD。与目的地计算实例相关联的VNIC然后处理分组并将分组转发到目的地计算实例。与源和目的地计算实例相关联的VNIC可以在同一个NVD上执行(例如,当源和目的地计算实例由同一个主机机器托管时)或在不同的NVD上执行(例如,当源和目的地计算实例由连接到不同NVD的不同主机机器托管时)。

[0156] 如果用于分组的目的地在源计算实例的VCN外部,那么源自源计算实例的分组从托管源计算实例的主机机器传送到连接到该主机机器的NVD。NVD执行与源计算实例相关联的VNIC。由于分组的目的地端点在VCN外部,因此分组随后由用于该VCN的VCN VR处理。NVD调用VCN VR功能性,这会导致分组被转发到执行与VCN相关联的适当网关的NVD。例如,如果目的地是客户的内部部署网络内的端点,那么分组可以由VCN VR转发到执行为VCN配置的DRG网关的NVD。VCN VR可以在与执行与源计算实例相关联的VNIC的NVD相同的NVD上执行,或者由不同的NVD执行。网关可以由NVD执行,NVD可以是smartNIC、主机机器或其它NVD实施方式。然后,分组由网关处理并转发到下一跳,该下一跳促进将分组传送到其预期的目的地端点。例如,在图2中描绘的实施例中,源自计算实例268的分组可以通过链路220(使用NIC 232)从主机机器202传送到NVD 210。在NVD 210上,调用VNIC 276,因为它是与源计算实例268相关联的VNIC。VNIC 276被配置为检查分组中封装的信息,并确定用于转发分组的下一跳,目的是促进将分组传送到其预期的目的地端点,然后将分组转发到确定的下一跳。

[0157] 部署在VCN上的计算实例可以与各种不同端点通信。这些端点可以包括由CSPI

200托管的端点和CSPI 200外部的端点。由CSPI 200托管的端点可以包括同一VCN或其它VCN中的实例,这些VCN可以是客户的VCN,或者是不属于客户的VCN。由CSPI 200托管的端点之间的通信可以通过物理网络218执行。计算实例还可以与不由CSPI 200托管或在CSPI 200外部的端点通信。这些端点的示例包括客户的内部部署网络内的端点或数据中心,或可通过公共网络(诸如互联网)访问的公共端点。与CSPI 200外部的端点的通信可以使用各种通信协议通过公共网络(例如,互联网)(图2中未示出)或专用网络(图2中未示出)来执行。

[0158] 图2中描绘的CSPI 200的体系架构仅仅是示例并且不旨在限制。在替代实施例中,变化、替代方案和修改是可能的。例如,在一些实施方式中,CSPI 200可以具有比图2中所示的系统或组件更多或更少的系统或组件,可以组合两个或更多个系统,或者可以具有不同的系统配置或布置。图2中描绘的系统、子系统和其它组件可以在由相应系统的一个或多个处理单元(例如,处理器、核心)执行的软件(例如,代码、指令、程序)中、使用硬件或其组合实现。软件可以存储在非暂态存储介质上(例如,存储器设备上)。

[0159] 图4描绘了根据某些实施例的主机机器和NVD之间的连接,用于提供I/O虚拟化以支持多租赁。如图4中所描绘的,主机机器402执行提供虚拟化的环境的管理程序404。主机机器402执行两个虚拟机实例,属于客户/租户#1的VM1 406和属于客户/租户#2的VM2 408。主机机器402包括经由链路414连接到NVD 412的物理NIC 410。每个计算实例都附接到由NVD 412执行的VNIC。在图4中的实施例中,VM1 406附接到VNIC-VM1 420并且VM2 408附接到VNIC-VM2 422。

[0160] 如图4中所示,NIC 410包括两个逻辑NIC,逻辑NIC A 416和逻辑NIC B 418。每个虚拟机附接到其自己的逻辑NIC并被配置为与其自己的逻辑NIC一起工作。例如,VM1 406附接到逻辑NIC A 416并且VM2 408附接到逻辑NIC B 418。虽然主机机器402仅包括由多个租户共享的一个物理NIC 410,但由于逻辑NIC,因此每个租户的虚拟机相信他们有自己的主机机器和网卡。

[0161] 在某些实施例中,每个逻辑NIC被指派有其自己的VLAN ID。因此,特定的VLAN ID被指派给用于租户#1的逻辑NIC A 416,并且分离的VLAN ID被指派给用于租户#2的逻辑NIC B 418。当分组从VM1 406传送时,由管理程序将指派给租户#1的标签附加到分组,然后分组通过链路414从主机机器402传送到NVD 412。以类似的方式,当分组从VM2 408被传送时,指派给租户#2的标签由管理程序附加到分组,然后分组通过链路414从主机机器402传送到NVD 412。因而,从主机机器402传送到NVD 412的分组424具有识别特定租户和相关联的VM的相关联的标签426。在NVD上,对于从主机机器402接收到的分组424,与该分组相关联的标签426被用于确定分组是由VNIC-VM1 420还是由VNIC-VM2 422处理。分组随后由对应的VNIC处理。图4中描述的配置使每个租户的计算实例能够相信他们拥有自己的主机机器和NIC。图4中描述的设置提供了I/O虚拟化以支持多租赁。

[0162] 图5描绘了根据某些实施例的物理网络500的简化框图。图5中描绘的实施例被结构化Clos网络。Clos网络是特定类型的网络拓扑,被设计为提供连接冗余性,同时维持高对分带宽和最大资源利用率。Clos网络是一种无阻塞、多级或多层交换网络,其中级数或层数可以是二、三、四、五等。图5中描绘的实施例是3层网络,包括层1、层2和层3。TOR交换机504表示Clos网络中的层0交换机。一个或多个NVD连接到TOR交换机。层0交换机也被称为物理网络的边缘设备。层0交换机连接到也称为叶交换机的层1交换机。在图5中描绘的实施例

中,“n”个层0TOR交换机的集合连接到“n”个层1交换机的集合并形成群聚 (pod)。群聚中的每个层0交换机都互连到该群聚中的所有层1交换机,但群聚之间没有交换机的连接性。在某些实施方式中,两个群聚被称为块。每个块由“n”个层2交换机 (有时称为主干交换机) 的集合提供服务或连接到这些交换机。物理网络拓扑中可以有几个块。层2交换机进而连接到“n”个层3交换机 (有时称为超级主干交换机)。物理网络500之上的分组的通信通常使用一种或多种层3通信协议来执行。通常,物理网络的所有层 (TOR层除外) 都是n向冗余的,因此允许高可用性。可以为群聚和块指定策略,以控制交换机在物理网络中彼此的可见性,从而启用物理网络的扩展 (scale)。

[0163] Clos网络的特征是从一个层0交换机到另一个层0交换机 (或从连接到层0交换机的NVD到连接到层0交换机的另一个NVD) 的最大跳数是固定的。例如,在3层Clos网络中,分组从一个NVD到达另一个NVD最多需要七跳,其中源和目标NVD连接到Clos网络的叶层。同样,在4层Clos网络中,分组从一个NVD到达另一个NVD最多需要九跳,其中源和目标NVD连接到Clos网络的叶层。因此,Clos网络体系架构在整个网络中维持一致的时延,这对于数据中心内部和数据中心之间的通信是重要的。Clos拓扑可水平扩展且具有成本效益。通过在各个层添加更多交换机 (例如,更多叶交换机和骨干交换机) 以及通过增加相邻层的交换机之间的链路数量,可以轻松增加网络的带宽/吞吐量容量。

[0164] 在某些实施例中,CSPI内的每个资源都被指派称为云标识符 (CID) 的唯一标识符。这个标识符作为资源的信息的一部分包括在内并且可以被用于管理资源,例如,经由控制台或通过API。用于CID的示例语法是:

[0165] `ocid1.<RESOURCE TYPE>.<REALM>.[REGION][.FUTURE USE].<UNIQUE ID>`

[0166] 其中,

[0167] `ocid1`:指示CID的版本的文字串;

[0168] `resource type`:资源的类型 (例如,实例、卷、VCN、子网、用户、组等);

[0169] `realm`:资源所在的领域。示例值为用于商业领域的“c1”、用于政府云领域的“c2”或用于联邦政府云领域的“c3”等。每个领域可以有其自己的域名;

[0170] `region`:资源所在的地区。如果该地区不适用于该资源,那么这个部分可能为空;

[0171] `future use`:保留以备将来使用。

[0172] `unique ID`:ID的唯一部分。格式可以因资源或服务的类型而异。

[0173] L2虚拟网络

[0174] 将他们的内部部署应用过渡到由云服务提供者 (CSP) 提供的云环境的企业客户的数量继续快速增加。但是,这些客户中的许多很快意识到,过渡到云环境的道路可能非常坎坷,要求客户重新构建和重新设计他们现有的应用以使其在云环境中可工作。这是因为内部部署环境编写的应用常常取决于物理网络的用于监视、可用性和扩展的功能。因此,这些内部部署应用需要重新构建和重新设计,然后才能在云环境中工作。

[0175] 内部部署应用无法轻松过渡到云环境的原因有几个。主要原因之一是当前的云虚拟网络在OSI模型的层3处操作,例如在IP层处,并且不提供应用所需的层2能力。基于层3的路由或转发包括基于分组的层3报头中包含的信息 (例如,基于分组的层3报头中包含的目的IP地址) 确定分组要发送到哪里 (例如,发送到哪个客户实例)。为了促进这一点,IP地址在虚拟化的云网络中的位置是通过集中控制和编排系统或控制器来确定的。这些可以包

括例如与虚拟化的云环境中的客户实体或资源相相关联的IP地址。

[0176] 许多客户在他们的内部部署环境中运行应用,这些环境对层2联网功能有严格的要求,当前的云产品和IaaS服务提供者目前没有解决这些问题。例如,当前云产品的流量使用使用层3报头的层3协议进行路由,并且不支持应用所需的层2特征。这些层2特征可以包括诸如地址解析协议(ARP)处理、介质访问控制(MAC)地址学习和层2广播能力、层2(基于MAC的)转发、层2联网构造及其它。通过在虚拟化的云网络中提供虚拟化的层2联网功能性,如本公开中所述,客户现在可以将其遗留应用无缝地迁移到云环境,而无需任何实质性的重新构建或重新设计。例如,本文描述的虚拟化的层2联网能力使此类应用(例如,VMware vSphere、vCenter、vSAN和NSX-T组件)能够像在内部部署环境中一样在层2进行通信。这些应用能够在公共云中运行相同的版本和配置,从而使客户能够使用其遗留的内部部署应用,包括与遗留应用相关联的现有知识、工具和过程。客户还能够从他们的应用访问原生云服务(例如,使用VMware软件定义的数据中心(SDDC))。

[0177] 作为另一个示例,有几个遗留的内部部署应用(例如,企业聚类软件应用、网络虚拟器具)要求层2广播支持以进行故障转移。示例应用包括Fortinet FortiGate、IBM QRadar、Palo Alto防火墙、Cisco ASA、Juniper SRX和Oracle RAC(真正的应用聚类)。通过如本公开中所述在虚拟化的公共云中提供虚拟化的层2网络,这些应用现在能够在虚拟化的公共云环境中不加改变地运行。如本文所述,提供了与内部部署相当的虚拟化的层2联网功能性。本公开中描述的虚拟化的层2联网功能性支持传统的层2网络。这包括对客户定义的VLAN以及单播、广播和多播层2流量功能的支持。基于层2的分组路由和转发包括使用层2协议和使用分组的层2报头中包含的信息,例如,基于层2报头中包含的目的地MAC地址来路由或转发分组。由企业应用(例如,聚类软件应用)使用的协议(诸如ARP、无偿地址解析协议(GARP)和反向地址解析协议(RARP))现在也可以在云环境中工作。

[0178] 传统虚拟化的云基础设施支持虚拟化的层3联网而不支持层2联网有几个原因。层2网络通常无法像层3网络那样扩展。层2网络控制协议不具备扩展所期望的复杂级别。例如,层3网络不必担心层2网络必须解决的分组循环问题。IP分组(即,层3分组)具有生存时间(TTL)的概念,而层2分组则没有。包含在层3分组中的IP地址具有拓扑概念,诸如子网、CIDR范围等,而层2地址(例如,MAC地址)则没有。层3IP网络有促进排除故障的内置工具,诸如用于查找路径信息的ping、traceroute等。此类工具对于层2不可用。层3网络支持多路径,这在层2网络不可用。由于缺乏专门用于网络中实体之间交换信息的复杂控制协议(例如,边界网关协议(BGP)和开放最短路径优先(OSPF)),层2网络必须依赖广播和多播才能按顺序学习网络,这会对网络性能产生不利影响。随着网络的变化,用于层2的学习过程必须重复,而层3则不需要。出于这些原因和其它原因,云IaaS服务提供者更期望提供在层3而不是在层2操作的基础设施。

[0179] 但是,虽然有多个缺点,但是许多内部部署应用仍需要层2功能性。例如,假设虚拟化的云配置,其中客户(客户1)在虚拟网络“V”中有两个实例,具有IP1的实例A和具有IP2的实例B,其中实例可以是计算实例(例如,裸机、虚拟机或容器)或服务实例(诸如负载均衡器、nfs挂载点或其它服务实例)。虚拟网络V是与其它虚拟网络和底层物理网络隔离的独特地址空间。例如,可以使用包括分组封装或NAT在内的各种技术来实现这种隔离。出于这个原因,用于客户的虚拟网络中的实例的IP地址不同于托管它的物理网络中的地址。提供集

中式SDN(软件定义的网络)控制平面,它知道所有虚拟IP地址的物理IP和虚拟接口。当分组从实例A发送到虚拟网络V中的IP2的目的地时,虚拟网络SDN堆栈需要知道IP2所在的位置。它必须提前知道这一点,以便它可以将分组发送到托管用于V的虚拟IP地址IP2的物理网络中的IP。可以在云中修改虚拟IP地址的位置,从而改变物理IP与虚拟IP地址之间的关系。每当要移动虚拟IP地址(例如,将与虚拟机相关联的IP地址移动到另一个虚拟机或将虚拟机迁移到新的物理主机)时,必须对SDN控制平面进行API调用,从而让控制器知道IP正在移动,以便它可以更新SDN堆栈中的所有参与者,包括分组处理器(数据平面)。但是,有些应用的类不会进行此类API调用。示例包括各种内部部署应用、由各种虚拟化软件供应商(诸如VMware等)提供的应用。在虚拟化的云环境中促进虚拟层2网络的价值使得能够支持未编程为进行此类API调用的应用或依赖其它层2联网特征的应用,诸如支持非IP层3和MAC学习。

[0180] 虚拟层2网络创建广播域,其中学习由广播域的成员执行。在虚拟层2域中,在这个层2域内任何主机上的任何MAC上都可以有任何IP,并且系统将学习使用标准层2联网协议,并且系统将虚拟化这些联网原语,而无需由中央控制器明确告知MAC和IP在该虚拟层2网络中的位置。这使得需要低时延故障转移的应用能够运行,需要支持到多个节点的广播或多播协议的应用,以及不知道如何对SDN控制平面或API端点进行API调用以确定IP和MAC地址的位置的遗留应用。因此,需要在虚拟化的云环境中提供层2联网能力,以便能够支持IP层3级别处不可用的功能性。

[0181] 在虚拟化的云环境中提供虚拟层2的另一个技术优势是它使得能够支持各种不同的层3协议(诸如IPV4、IPV6),包括非IP协议。例如,可以支持各种非IP协议,诸如IPX、AppleTalk等。因为现有的云IaaS提供者不在其虚拟化的云网络中提供层2功能性,所以它们无法支持这些非IP协议。通过提供本公开中描述的层2联网功能性,可以为层3处的协议以及为需要和依赖层2级别功能性的可用性的应用提供支持。

[0182] 使用本公开中描述的技术,在虚拟化的云基础设施中提供层3和层2功能性两者。如前所述,基于层3的联网提供了某些效率,尤其适合扩展,这是层2联网没有提供的。在层3功能性之外提供层2功能性允许充分利用由层3提供的此类效率(例如,提供更可扩展的解决方案),同时以更可扩展的方式提供层2功能性。例如,虚拟化的层3避免使用广播用于学习目的。通过提供层3以提高效率,同时提供虚拟化的层2以启用那些需要它的应用和那些没有层2功能性就无法运行的应用,并支持非IP协议等,为客户提供了虚拟化的云环境中的完全灵活性。

[0183] 客户本身具有混合环境,其中层2环境与层3环境一起存在,并且虚拟化的云环境现在可以支持这两种环境。客户可以拥有子网之类的层3网络和/或诸如VLAN之类的层2网络,并且这两个环境可以在虚拟化的云环境中彼此通信。

[0184] 虚拟化的云环境也需要支持多租赁。多租赁使得在同一虚拟化的云环境中供应层3功能性和层2功能性两者在技术上变得困难和复杂。例如,层2广播域必须跨云提供者的基础设施中的许多不同客户被管理。本公开中描述的实施例克服了这些技术问题。

[0185] 对于虚拟化提供者(例如,VMware),模拟物理层2网络的虚拟化的层2网络允许工作负载不加改变地运行。由这种虚拟化提供者提供的应用然后可以在由云基础设施提供的虚拟化的层2网络上运行。例如,此类应用可以包括需要在层2网络上运行的实例的集合。当客户想要将这种应用从其内部部署环境提升并转移到虚拟化的云环境时,他们不能只获取

该应用并在云中运行它,因为那些应用依赖于底层的层2网络(例如,层2网络特征被用于执行虚拟机的迁移,或移动MAC和IP地址所在的位置),这是当前虚拟化的云提供者未提供的。出于这些原因,此类应用无法在虚拟化的云环境中原生运行。使用本文描述的技术,云提供者除了提供虚拟化的层3网络之外还提供虚拟化的层2网络。现在,这种应用堆栈可以不加改变地在云环境中运行,并且可以在云环境中运行嵌套的虚拟化。客户现在可以在云中运行其自己的层2应用并对其进行管理。应用提供者不必对其软件进行任何改变来促进这一点。此类遗留应用或工作负载(例如,遗留负载均衡器、遗留应用、KVM、Openstack、聚类软件)现在可以在虚拟化的云环境中不加改变地运行。

[0186] 通过提供本文所述的虚拟化的层2功能性,虚拟化的云环境现在可以支持各种层3协议,包括非IP协议。以以太网为例,可以支持各种不同的EtherType(层2报头中的一个字段,它告诉正在发送的层3分组的类型;告诉在层3处期望什么协议),包括各种非IP协议。EtherType是以太网帧中的两个八位字节字段。它被用于指示哪种协议被封装在帧的有效载荷中并在接收端由数据链路层用于确定如何处理有效载荷。EtherType还用作802.1QVLAN标记、封装来自VLAN的分组以便通过以太网中继与其它VLAN流量多路复用传输的基础。EtherType的示例包括IPV4、IPv6、地址解析协议(ARP)、AppleTalk、IPX等。支持层2协议的云网络在层3处可以支持任何协议。以类似的方式,当云基础设施提供对层3协议的支持时,它可以在层4支持各种协议,诸如TCP、UDP、ICMP等。当在层3处提供虚拟化时,网络可以不受层4协议的影响。类似地,当在层2提供虚拟化时,网络可以不受层3协议的影响。这种技术可以被扩展以支持任何层2网络类型,包括FDDI、Infiniband等。

[0187] 因而,许多为物理网络编写的应用(尤其是与共享广播域的计算机节点的集群一起工作的应用)使用L3虚拟网络不支持的层2特征。以下六个示例突出了不提供层2联网能力可能导致的复杂情况:

[0188] (1) 在没有预先API调用的情况下指派MAC和IP。网络器具和管理程序(诸如VMware)不是为云虚拟网络构建的。他们假设他们能够使用MAC,只要它是唯一的即可,并且或者从DHCP服务器获取动态地址或者使用指派给集群的任何IP。常常没有任何机制可以将它们配置为通知控制平面有关这些层2和层3地址的指派。如果不知道MAC和IP的位置,那么层3虚拟网络不知道将流量发送到哪里。

[0189] (2) 用于高可用性和实况迁移的MAC和IP的低时延迟重新分配。许多内部部署应用使用ARP来重新指派IP和MAC以实现高可用性—当集群或HA对中的实例停止响应时,新活动实例将发送无偿ARP(GARP)以将服务IP重新指派给其MAC或发送反向ARP(RARP)以将服务MAC重新指派给其接口。这在管理程序上实况迁移实例时也是重要的:新主机必须在访客迁移后发送RARP,以便将访客流量发送到新主机。指派不仅无需API调用即可完成,而且还需要极低的时延(亚毫秒)。这无法通过对REST端点的HTTPS调用来完成。

[0190] (3) 通过MAC地址进行接口多路复用。当管理程序在单个主机上托管多个虚拟机时,所有这些虚拟机都在同一个网络上,访客接口通过它们的MAC来区分。这要求在同一虚拟接口上支持多个MAC。

[0191] (4) VLAN支持。单个物理虚拟机主机将需要位于多个广播域上,如使用VLAN标记所指示的。例如,VMware ESX使用VLAN进行流量分离(例如,访客虚拟机可以在一个VLAN上通信,在另一个VLAN上存储,而在再另一个VLAN上托管虚拟机)。

[0192] (5) 广播和多播流量的使用。ARP要求L2广播,并且存在内部部署应用将广播和多播流量用于集群和HA应用的示例。

[0193] (6) 支持非IP流量。由于L3网络要求IPv4或IPv6报头进行通信,因此使用IP以外的任何L3协议都不起作用。L2虚拟化意味着VLAN内的网络可以与L3协议无关—L3报头可以是IPv4、IPv6、IPX或其它任何东西—甚至完全不存在。

[0194] 如本文所公开的,可以在云网络内创建层2 (L2) 网络。这个虚拟L2网络包括一个或几个虚拟化的L2 VLAN (本文简称VLAN)。每个VLAN可以包括多个计算实例,每个计算实例可以与至少一个L2虚拟接口 (例如,L2 VNIC) 和本地交换机相关联。在一些实施例中,每对L2 VNIC和交换机托管在NVD上。NVD可以托管多个这样的对,其中每个对都与不同的计算实例相关联。本地交换机的集合表示VLAN的模拟的单个交换机。L2 VNIC表示模拟的单个交换机上的端口的集合。VLAN可以经由VLAN交换和路由服务 (VSRS) (本文也称为真实虚拟路由器 (RVR) 或L2VSRS) 连接到其它VLAN、层3 (L3) 网络、内部部署网络和/或其它网络。

[0195] 现在参考图6,示出了计算网络的一个实施例的示意图。VCN 602驻留在CSPI 601中。VCN 602包括将VCN 602连接到其它网络的多个网关。这些网关包括DRG 604,其可以将VCN 602连接到例如内部部署网络 (诸如内部部署数据中心606)。网关还可以包括网关600,其可以包括例如用于将VCN 602与另一个VCN连接的LPG,和/或用于将VCN 602连接到互联网的IGW和/或NAT网关。VCN602的网关还可以包括服务网关610,其可以将VCN 602与服务网络612连接。服务网络612可以包括一个或几个数据库和/或存储库,包括例如自主数据库614和/或对象存储库616。服务网络可以包括概念网络,该概念网络包括IP范围的聚合,例如可以是公共IP范围。在一些实施例中,这些IP范围可以覆盖由CSPI 601提供者提供的公共服务中的一些或全部。例如,可以通过互联网网关或NAT网关访问这些服务。在一些实施例中,服务网络为服务网络中的服务提供一种方式,以便通过用于该目的的专用网关 (服务网关) 从本地区域访问服务网络中的服务。在一些实施例中,这些服务的后端可以在例如它们自己的私有网络中实现。在一些实施例中,服务网络612可以包括另外的附加数据库。

[0196] VCN 602可以包括多个虚拟网络。这些网络各自可以包括一个或几个计算实例,这些计算实例可以在它们各自的网络内、网络之间或在VCN 602外部进行通信。VCN 602的虚拟网络之一是L3子网620。L3子网620是位在VCN 602内创建的配置的单元细分。子网620可以包括VCN 602的虚拟化的云环境中的虚拟层3网络,该VCN 602托管在CPSI 601的底层物理网络上。虽然图6描绘了单个子网620,但是VCN 602可以具有一个或多个子网。VCN 602内的每个子网可以与不与该VCN中的其它子网重叠并且表示该VCN的地址空间内的地址空间子集的覆盖IP地址的连续范围 (例如,10.0.0.0/24和10.0.1.0/24) 相关联。在一些实施例中,这个IP地址空间可以与和CPSI 601相关联的地址空间隔离。

[0197] 子网620包括一个或多个计算实例,并且具体地包括第一计算实例622-A和第二计算实例622-B。计算实例622-A、622-B可以在子网620内彼此通信,或者它们可以与子网620外部的其它实例、设备和/或网络通信。子网620外部的通信由虚拟路由器 (VR) 624启用。VR 624启用于子网620与VCN 602的其它网络之间的通信。对于子网620,VR 624表示使子网620 (即,计算实例622-A、622-B) 能够与VCN 602内的其它网络上的端点以及与VCN 602外的其它端点通信的逻辑网关。

[0198] VCN 602还可以包括附加网络,并且具体地可以包括一个或几个L2 VLAN (本文称

为VLAN)，它们是虚拟L2网络的示例。这一个或几个VLAN可以各自包括位于VCN 602的云环境中/或由CPSI 601的底层物理网络托管的虚拟层2网络。在图6的实施例中，VCN 602包括VLAN A 630和VLAN B 640。VCN 602内的每个VLAN 630、640可以与覆盖IP地址的连续范围（例如，10.0.0.0/24和10.0.1.0/24）相关联，这些IP地址不与该VCN中的其它网络（诸如该VCN中的其它子网或VLAN）重叠并且表示VCN的地址空间内的地址空间子集。在一些实施例中，VLAN的这个IP地址空间可以与和CPSI 601相关联的地址空间隔离。

[0199] VLAN 630、640中的每一个可以包括一个或几个计算实例，并且具体而言，VLAN A 630可以包括例如第一计算实例632-A和第二计算实例632-B。在一些实施例中，VLAN A 630可以包括附加的计算实例。VLAN B 640可以包括例如第一计算实例642-A和第二计算实例642-B。计算实例632-A、632-B、642-A、642-B中的每一个都可以具有IP地址和MAC地址。这些地址可以以任何期望的方式指派或生成。在一些实施例中，这些地址可以在计算实例的VLAN的CIDR内，并且在一些实施例中，这些地址可以是任何地址。在VLAN的计算实例与VLAN外部的端点通信的实施例中，这些地址中的一个或两个来自VLAN CIDR，而当所有通信都是VLAN内时，这些地址不限于VLAN CIDR内的地址。与地址由控制平面指派的网络不同，VLAN中计算实例的IP和/或MAC地址可以由该VLAN的用户/客户指派，并且这些IP和/或MAC地址可以然后根据下面讨论的用于学习的过程由VLAN中的计算实例发现和/或学习。

[0200] 每个VLAN可以包括VLAN交换和路由服务（VSRS），并且具体而言，VLAN A 630包括VSRS A 634并且VLAN B 640包括VSRS B 644。每个VSRS 634、644参与VLAN内的层2交换和本地学习并且还执行所有必要的层3网络功能，包括ARP、NDP和路由。VSRS执行ARP（其是层2协议），因为VSRS必须将IP映射到MAC。

[0201] 在这些基于云的VLAN中，每个虚拟接口或虚拟网关可以与一个或多个介质访问控制（MAC）地址相关联，这些地址可以是虚拟MAC地址。在VLAN内，一个或几个计算实例632-A、632-B、642-A、642-B（例如，其可以是裸机、VM或容器和/或一个或几个服务实例）可以经由虚拟交换机直接彼此通信。经由VSRS 634、644启用VLAN外部的通信，诸如与其它VLAN或与L3网络的通信。VSRS 634、644是分布式服务，为VLAN网络提供层3功能，诸如IP路由。在一些实施例中，VSRS 634、644是可水平扩展、高度可用的路由服务，其可以位于IP网络和L2网络的交叉点处并参与基于云的L2域内的IP路由和L2学习。

[0202] VSRS 634、644可以跨基础设施内的多个节点分布，并且VSRS 634、644功能可以是可扩展的，尤其是可以水平可扩展的。在一些实施例中，实现VSRS 634、644的功能的每个节点彼此共享和复制路由器和/或交换机的功能。此外，这些节点可以将它们自己作为单个VSRS 634、644呈现给VLAN 630、640中的所有实例。VSRS 634、644可以在CPSI 601内的任何虚拟化设备上实现，并且具体地在虚拟网络内。因此，在一些实施例中，VSRS 634、644可以在任何虚拟网络虚拟化设备（包括NIC、SmartNIC、交换机、智能交换机或通用计算主机）上实现。

[0203] VSRS 634、644可以是驻留在支持云网络的一个或几个硬件节点上的服务，硬件节点诸如例如一个或几个服务器，诸如例如一个或几个x86服务器，或一个或几个支持云网络的联网设备，诸如一个或几个NIC并且具体地是一个或几个SmartNIC。在一些实施例中，VSRS 634、644可以在服务器机群上实现。因此，VSRS 634、644可以是跨节点的机群分布的服务，该机群可以是集中管理的机群或者可以分布到参与并共享L2和L3学习以及评估路由

和安全性策略的虚拟联网执行器的边缘。在一些实施例中,每个VSRS实例都可以用新的映射信息更新其它VSRS实例,因为这个新的映射信息是由VSRS实例学习的。例如,当VSRS实例学习其VLAN中一个或多个CI的IP、接口和/或MAC映射时,VSRS实例可以将该更新后的信息提供给VCN内的其它VSRS实例。经由这种交叉更新,与第一VLAN相关联的VSRS实例可以知道用于其它VLAN中的CI(在一些实施例中,用于VCN 602内的其它VLAN中的CI)的映射,包括IP、接口和/或MAC映射。当VSRS驻留在服务器机群上和/或跨节点的机群分布时,这些更新可以大大加快。

[0204] 在一些实施例中,VSRS 634、644还可以托管联网所必需的一个或几个更高级别的服务,包括但不限于:DHCP中继;DHCP(托管);DHCPv6;邻居发现协议(诸如IPv6邻居发现协议);DNS;托管DNSv6;用于IPv6的SLAAC;NTP;元数据服务;以及块存储(blockstore)挂载点。在一些实施例中,VSRS可以支持一个或几个网络地址转换(NAT)功能以在网络地址空间之间进行转换。在一些实施例中,VSRS可以结合反欺骗、反MAC欺骗、用于IPv4的ARP高速缓存中毒保护、IPv6路由通告(RA)保护、DHCP保护、使用访问控制列表(ACL)的分组过滤;和/或反向路径转发检查。VSRS可以实现的功能包括例如ARP、GARP、分组过滤器(ACL)、DHCP中继和/或IP路由协议。例如,VSRS 634、644可以学习MAC地址、使过期的MAC地址失效、处置MAC地址的移动、审查(vet)MAC地址信息、处置MAC信息的泛滥、风暴控制的处置、环路预防、经由例如云中的协议(诸如IGMP)的层2多播、包括日志在内的统计信息搜集、使用SNMP的统计信息、监视,和/或搜集和使用用于广播、总流量、位、跨越树分组等的统计信息。

[0205] 在虚拟网络内,VSRS 634、644可以表现为不同的实例化。在一些实施例中,VSRS的这些实例化中的每一个都可以与VLAN 630、640相关联,并且在一些实施例中,每个VLAN 630、640可以具有VSRS 634、644的实例化。在一些实施例中,VSRS 634、644的每个实例化可以具有与VSRS 634、644的每个实例化与之相关联的VLAN 630、640的对应的一个或几个唯一表。VSRS 634、644的每个实例化可以生成和/或策划与VSRS 634、644的那个实例化相关联的唯一表。因此,虽然单个服务可以为一个或几个云网络提供VSRS 634、644功能性,但云网络内VSRS 634、644的各个实例化可以具有唯一的层2和层3转发表,而多个此类客户网络可以具有重叠的层2和层3转发表。

[0206] 在一些实施例中,VSRS 634、644可以支持跨多个租户的冲突VLAN和IP空间。这可以包括在同一VSRS 634、644上拥有多个租户。在一些实施例中,这些租户中的一些或全部可以选择使用以下一些或全部:相同的IP地址空间、相同的MAC空间和相同的VLAN空间。这可以为用户选择地址提供极大的灵活性。在一些实施例中,经由为每个租户提供不同的虚拟网络来支持这种多租赁,该虚拟网络是云网络内的私有网络。每个虚拟网络都被赋予唯一标识符。类似地,在一些实施例中,每个主机可以具有唯一标识符,和/或每个虚拟接口或虚拟网关可以具有唯一标识符。在一些实施例中,这些唯一标识符,并且具体而言是租户的虚拟网络的唯一标识符,可以在每次通信中被编码。通过为每个虚拟网络提供唯一标识符并将其包括在通信中,VSRS 634、644的单个实例化可以为具有重叠地址和/或名称空间的多个租户提供服务。

[0207] VSRS 634、644可以执行这些交换和/或路由功能以促进和/或启用VLAN 630、640内的L2网络的创建和/或与其的通信。这个VLAN 630、640可以在云计算环境内,并且更具体而言在该云计算环境中的虚拟网络内,找到。

[0208] 例如,VLAN 630、640中的每一个包括多个计算实例632-A、632-B、642-A、642-B。VSRS 634、644启用一个VLAN 630、640中的计算实例与另一个VLAN 630、640中或子网620中的计算实例之间的通信。在一些实施例中,VSRS 634、644启用一个VLAN 630、640中的计算实例与另一个VCN、该VCN外部的另一个网络(包括互联网、内部部署数据中心等)之间的通信。在这种实施例中,例如,计算实例(诸如计算实例632-A)可以向VLAN外部的端点(在这个实例中,L2 VLAN A 630外部的端点)发送通信。计算实例(632-A)可以向VSRS A 634发送通信,VSRS A 634可以将通信指向与期望端点通信耦合的路由器624、644或网关604、608、610。与期望端点通信耦合的路由器624、644或网关604、608、610可以从计算实例(632-A)接收通信并且可以将通信指向期望端点。

[0209] 现在参考图7,示出了VLAN 700的逻辑和硬件示意图。如看到的,VLAN 700包括多个端点,具体而言包括多个计算实例和VSRS。多个计算实例(CI)在一个或几个主机机器上实例化。在一些实施例中,这可以是一对一的关系,使得每个CI在唯一的主机机器上被实例化,和/或在一些实施例中,这可以是多对一的关系,使得多个CI在单个共用的主机机器上被实例化。在各种实施例中,通过被配置为使用L2协议彼此通信,CI可以是层2CI。图7描绘了一个场景,其中一些CI在唯一的主机机器上被实例化,并且其中一些CI共享共用的主机机器。如图7中所看到的,实例1(CI1) 704-A在主机机器1 702-A上被实例化,实例2(CI2) 704-B在主机机器2 702-B上被实例化,并且实例3(CI3) 704-A C和实例4(CI4) 704-D在共用的主机机器702-C上被实例化。

[0210] CI 704-A、704-B、704-C、704-D中的每一个都与VLAN 700中的其它CI 704-A、704-B、704-C、704-D并且与VSRS 714通信地耦合。具体而言,CI 704-A、704-B、704-C、704-D中的每一个都经由L2 VNIC和交换机连接到VLAN 700中的其它CI 704-A、704-B、704-C、704-D并且连接到VSRS 714。每个CI 704-A、704-B、704-C、704-D都与唯一的L2 VNIC和交换机相关联。交换机可以是本地的L2虚拟交换机,并且与L2 VNIC唯一相关联并为L2 VNIC部署。具体而言,CI1 704-A与L2 VNIC 1 708-A和交换机1 710-A相关联,CI2 704-B与L2 VNIC 2 708-B和交换机710-B相关联,CI3 704-C与L2 VNIC 3 708-C和交换机3 710-C相关联,并且CI4 704-D与L2VNIC 4 708-D和交换机4 710-D相关联。

[0211] 在一些实施例中,每个L2 VNIC 708及其相相关联的交换机710可以在NVD 706上被实例化。这种实例化可以是一对一的关系,使得单个L2 VNIC 708及其相相关联的交换机710在唯一的NVD 706上被实例化,或者这个实例化可以是多对一的关系,使得多个L2 VNIC 708及其相相关联的交换机710在单个共用的NVD 706上被实例化。具体而言,L2 VNIC 1 708-A和交换机1 710-A在NVD 1 706-A上被实例化,L2 VNIC 2 708-B和交换机2 710-B在NVD 2 上被实例化,并且L2 VNIC 3 708-C和交换机3 710-C以及L2 VNIC 4 708-D和开关710-D都在共用的NVD(即,NVD 706-C)上被实例化。

[0212] 在一些实施例中,VSRS 714可以支持跨多个租户的冲突VLAN和IP空间。这可以包括在同一VSRS 714上具有多个租户。在一些实施例中,这些租户中的一些或全部可以选择使用以下一些或全部:相同的IP地址空间、相同的MAC空间和相同的VLAN空间。这可以为用户选择地址提供极大的灵活性。在一些实施例中,通过为每个租户提供不同的虚拟网络来支持这种多租赁,该虚拟网络是云网络内的私有网络。每个虚拟网络(例如,每个VLAN或VCN)都被赋予唯一标识符,诸如可以是VLAN标识符的VCN标识符。这个唯一标识符可以由例

如控制平面选择,并且具体而言是由CSPI的控制平面选择。在一些实施例中,这个唯一标识符可以包括一个或几个位,其可以被包括在和/或用在分组封装中。

[0213] 类似地,在一些实施例中,每个主机可以具有唯一标识符,和/或每个虚拟接口或虚拟网关可以具有唯一标识符。在一些实施例中,这些唯一标识符,并且具体而言是租户的虚拟网络的唯一标识符,可以在每次通信中被编码。通过为每个虚拟网络提供唯一标识符并将其包含在通信中,VSRS的单个实例化可以为具有重叠地址和/或名称空间的多个租户提供服务。

[0214] 在一些实施例中,VSRS 714可以基于与通信相相关联的VCN标识符和/或VLAN标识符并且具体地在通信的VCN报头内部来确定分组属于哪个租户。在本文公开的实施例中,离开或进入VLAN的通信可以具有可以包括VLAN标识符的VCN报头。基于包含VLAN标识符的VCN报头,VSRS可以确定租赁,或者换句话说,接收方VSRS可以确定将通信发送到哪个VLAN和/或哪个租户。

[0215] 此外,属于VLAN的每个计算实例(例如,L2计算实例)被赋予唯一的接口标识符,该标识符识别与计算实例相相关联的L2 VNIC。接口标识符可以包括在来自和/或去往计算机实例的流量中(例如,通过包括在帧的报头中)并且可以由NVD用来识别与计算实例相相关联的L2 VNIC。换句话说,接口标识符可以唯一地指示计算实例和/或其相关联的L2 VNIC。如图7中所指示的,交换机710-A、710-B、710-C、710-D可以一起形成L2分布式交换机712,本文也称为分布式交换机712。从客户的角度来看,L2分布式交换机712中的每个交换机710-A、710-B、710-C、710-D是连接到VLAN中的所有CI的单个交换机。但是,模拟单个交换机的用户体验的分布式交换机是无限可扩展的并且包括本地交换机的集合(例如,在图7的说明性示例中,交换机710-A、710-B、710-C、710-D)。如图7中所示,每个CI在连接到NVD的主机机器上执行。对于连接到NVD的主机上的每个CI,NVD托管层2VNIC和与计算实例相关联的本地交换机(例如,L2虚拟交换机,在NVD本地,与层2VNIC相关联,并且是L2分布式交换机712的一个成员或组件)。层2VNIC表示层2VLAN上的计算实例的端口。本地交换机将L2 VNIC连接到与层2VLAN的其它计算实例相关联的其它L2 VNIC(例如,其它端口)。

[0216] CI 704-A、704-B、704-C、704-D中的每一个可以与VLAN 700中的CI 704-A、704-B、704-C、704-D中的其它通信,或者与VSRS 714通信。CI 704-A、704-B、704-C、704-D之一通过将分组发送到CI 704-A、704-B、704-C、704-D中的接收方的MAC地址和接口或发送到VSRS 714,来将分组发送到CI 704-A、704-B、704-C、704-D中的另一个或发送到VSRS 714。MAC地址和接口标识符可以包括在分组的报头中。如上文所解释的,接口标识符可以指示CI 704-A、704-B、704-C、704-D的接收方或VSRS 714的L2 VNIC。

[0217] 在一个实施例中,CI1 704-A可以是源CI,L2 VNIC 708-A可以是源VNIC,并且交换机710-A可以是源交换机。在这个实施例中,CI3 704-C可以是目的地CI,并且L2 VNIC 3 708-C可以是目的地VNIC。源CI可以发送带有源MAC地址和目的地MAC地址的分组。这个分组可以被NVD 706-A拦截,从而实例化源VNIC和源交换机。

[0218] 对于VLAN 700,L2 VNIC 708-A、708-B、708-C、708-D可以各自学习MAC地址到L2 VNIC的接口标识符的映射。可以基于从VLAN 700内接收的分组和/或通信来学习这个映射。基于这个先前确定的映射,源VNIC可以确定与VLAN内的目的地CI相关联的目的地接口的接口标识符,并且可以封装分组。在一些实施例中,这个封装可以包括GENEVE封装,并且具体

而言是L2 GENEVE封装,该封装包括被封装的分组的L2(以太网)报头。封装的分组可以识别目的地MAC、目的地接口标识符、源MAC和源接口标识符。

[0219] 源VNIC可以将封装的分组传递到源交换机,该源交换机可以将分组指向目的地VNIC。在接收到分组后,目的地VNIC可以解封装分组,然后可以将分组提供给目的地CI。

[0220] 现在参考图8,示出了多个连接的L2 VLAN 800的逻辑示意图。在图8中描绘的具体实施例中,两个VLAN都位于同一个VCN中。如所看到的,多个连接的L2 VLAN 800可以包括第一VLAN(VLAN A 802-A)和第二VLAN(VLAN B 802-B)。这些VLAN 802-A、802-B中的每一个都可以包括一个或几个CI,每个CI可以具有相关联的L2 VNIC和相关联的L2虚拟交换机。另外,这些VLAN 802-A、802-B中的每一个可以包括VSRS。

[0221] 具体而言,VLAN A 802-A可以包括连接到L2 VNIC 1 806-A和交换机1 808-A的实例1 804-A,连接到L2 VNIC 2 806-B和交换机808-B的实例2 804-B,以及连接到L2 VNIC 3 806-C和交换机3 808-C的实例3 804-C。VLAN B 802-B可以包括连接到L2 VNIC 4 806-D和交换机4 808-D的实例4 804-D,连接到L2 VNIC 5 806-E和交换机808-E的实例5 804-E,以及连接到L2 VNIC 6 806-F和交换机3 808-F的实例6 804-F。VLAN A 802-A还可以包括VSRS A 810-A,VLAN B 802-B可以包括VSRS B 810-B。VLAN A 802-A的CI 804-A、804-B、804-C中的每一个可以通信地耦合到VSRS A 810-A,以及VLAN B 802-B的CI 804-D、804-E、804-F中的每一个可以通信地耦合到VSRS B 810-B。

[0222] VLAN A 802-A可以经由它们各自的VSRS 810-A、810-B通信地耦合到VLAN B 802-B。每个VSRS同样可以耦合到网关812,网关812可以为每个VLAN 802-A、802-B中的CI 804-A、804-B、804-C、804-D、804-E、804-F提供对VLAN 802-A、802-B所在的VCN外部的其它网络的访问。在一些实施例中,这些网络可以包括例如一个或几个内部部署网络、另一个VCN、服务网络、诸如互联网之类的公共网络等。

[0223] VLAN A 802-A中的CI 804-A、804-B、804-C中的每一个可以经由每个VLAN 802-A、802-B的VSRS 810A、810-B与VLAN B 802-B中的CI 804-D、804-E、804-F通信。例如,VLAN 802-A、802-B之一中的CI 804-A、804-B、804-C、804-D、804-E、804-F之一可以向VLAN 802-A、802-B中的另一个中的CI 804-A、804-B、804-C、804-D、804-E、804-F发送分组。这个分组可以经由源VLAN的VSRS离开源VLAN并且可以进入目的地VLAN,并且经由目的地VSRS路由到目的地CI。

[0224] 在一个实施例中,CI 1 804-A可以是源CI,L2 VNIC 806-A可以是源VNIC,并且交换机808-A可以是源交换机。在这个实施例中,CI 5 804-E可以是目的地CI,并且L2 VNIC 5 806-E可以是目的地VNIC。VSRS A 810-A可以是被识别为SVSRS的源VSRS,并且VSRS B 810-B可以是被识别为DVSRS的目的地VSRS。

[0225] 源CI可以发送具有MAC地址的分组。这个分组可以被实例化源VNIC和源交换机的NVD拦截。源VNIC封装该分组。在一些实施例中,这个封装可以包括Geneve封装,并且具体而言是L2 Geneve封装。封装的分组可以识别目的地CI的目的地地址。在一些实施例中,这个目的地地址还可以包括目的地VSRS的目的地地址。目的地CI的目的地地址可以包括目的地IP地址、目的地CI的目的地MAC和/或目的地CI的目的地VNIC的目的地接口标识符。目的地VSRS的目的地地址可以包括目的地VSRS的IP地址、与目的地VSRS相关联的目的地VNIC的接口标识符,和/或目的地VSRS的MAC地址。

[0226] 源VSRS可以从源交换机接收分组,可以从分组的目的地地址查找VNIC映射,该目的地地址可以是目的地IP地址,并且可以将分组转发到目的地VSRS。目的地VSRS可以接收分组。基于分组中包含的目的地地址,目的地VSRS可以将分组转发到目的地VNIC。目的地VNIC可以接收和解封装分组,然后可以将分组提供给目的地CI。

[0227] 现在参考图9,示出了多个连接的L2 VLAN和子网900的逻辑示意图。在图9中所示的具体实施例中,VLAN和子网都位于同一个VCN中。这被指示为VLAN和子网的虚拟路由器和VSRS直接连接,而不是通过网关连接。

[0228] 如所看到的,这可以包括第一VLAN (VLAN A 902-A)、第二VLAN (VLAN B 902-B) 和子网930。这些VLAN 902-A、902-B中的每一个可以包括一个或几个CI,其中每一个都可以具有相关联的L2 VNIC和相关联的L2交换机。另外,这些VLAN 902-A、902-B中的每一个可以包括VSRS。同样,可以是L3子网的子网930可以包括一个或几个CI,每个CI可以具有相关联的L3 VNIC,并且L3子网930可以包括虚拟路由器916。

[0229] 具体而言,VLAN A 902-A可以包括连接到L2 VNIC 1 906-A和交换机1 908-A的实例1 904-A,连接到L2 VNIC 2 906-B和交换机908-B的实例2 904-B,以及连接到L2 VNIC 3 906-C和交换机3 908-C的实例3 904-C。VLAN B 902-B可以包括连接到L2 VNIC 4 906-D和交换机4 908-D的实例4 904-D,连接到L2 VNIC 5 906-E和交换机908-E的实例5 904-E,以及连接到L2 VNIC 6 906-F和交换机3 908-F的实例6 904-F。VLAN A 902-A还可以包括VSRS A 910-A,并且VLAN B 902-B可以包括VSRS B 910-B。VLAN A 902-A的CI 904-A、904-B、904-C中的每一个可以通信地耦合到VSRS A 910-A,并且VLAN B 902-B的CI 904-D、904-E、904-F中的每一个可以通信地耦合到VSRS B 910-B。L3子网930可以包括一个或几个CI,并且具体而言可以包括实例7 904-G,其通信地耦合到L3 VNIC 7 906-G。L3子网930可以包括虚拟路由器916。

[0230] VLAN A 902-A可以经由它们各自的VSRS 910-A、910-B通信地耦合到VLAN B 902-B。L3子网930可以经由虚拟路由器916与VLAN A 902-A和VLAN B 902-B通信耦合。每个虚拟路由器916和VSRS实例910-A、910-B同样可以耦合到网关912,网关912可以为每个VLAN 902-A、902-B中和子网930中的CI 904-A、904-B、904-C、904-D、904-E、904-F、904-G提供对VLAN 902-A、902-B和子网930所在的VCN外部的其它网络的访问。在一些实施例中,这些网络可以包括例如一个或几个内部部署网络、另一个VCN、服务网络、诸如互联网之类的公共网络等。

[0231] 每个VSRS实例910-A、910-B可以为离开相关联的VLAN 902-A、902-B的分组提供出口路径,并为进入相关联的VLAN 902-A、902-B的分组提供入口路径。从VLAN 902-A、902-B的VSRS实例910-A、910-B,分组可以被发送到任何期望的端点,包括L2端点(诸如在同一VCN上或不同VCN或网络上的另一个VLAN中的L2CI)和/或L3端点(诸如在同一VCN上或不同VCN或网络上的子网中的L3 CI)。

[0232] 在一个实施例中,CI 1 904-A可以是源CI,L2 VNIC 906-A可以是源VNIC,并且交换机908-A可以是源交换机。在这个实施例中,CI 7 904-G可以是目的地CI,并且VNIC 7 906-G可以是目的地VNIC。VSRS A 910-A可以是被识别为SVSRS的源VSRS,而虚拟路由器(VR)916可以是目的地VR。

[0233] 源CI可以发送具有MAC地址的分组。这个分组可以被实例化源VNIC和源交换机的

NVD拦截。源VNIC封装该分组。在一些实施例中,这个封装可以包括Geneve封装,并且具体而言是L2 Geneve封装。封装的分组可以识别目的地CI的目的地地址。在一些实施例中,这个目的地地址还可以包括源CI的VLAN的VSRS的目的地地址。目的地CI的目的地地址可以包括目的地IP地址、目的地CI的目的地MAC和/或目的地CI的目的地VNIC的目的地接口标识。

[0234] 源VSRS可以从源交换机接收分组,可以从分组的目的地地址查找VNIC映射,该目的地地址可以是目的地IP地址,并且可以将分组转发到目的地VR。目的地VR可以接收分组。基于分组中包含的目的地地址,目的地VR可以将分组转发到目的地VNIC。目的地VNIC可以接收和解封装分组,然后可以将分组提供给目的地CI。

[0235] 在虚拟L2网络内学习

[0236] 现在参考图10,示出了VLAN 1000内的VLAN内通信和学习的一个实施例的示意图。此处的学习特定于L2 VNIC、VSRS VNIC和/或L2虚拟交换机如何学习MAC地址与L2 VNIC/VRSR VNIC之间的关联(更具体而言,与L2计算实例或VSRS相关联的MAC地址与和这些L2计算实例的L2 VNICS相关联或与VSRS VNIC相关联的标识符之间的关联)。一般而言,学习是基于入口流量的。对于接口到MAC地址学习的一方面,这种学习不同于L2计算实例可以实现以学习目的地MAC地址的学习过程(例如,ARP过程)。这两个学习过程(例如,L2 VNIC/L2虚拟交换机的学习过程和L2计算实例的学习过程)在图12中被图示为联合实现。

[0237] 如所看到的,VLAN 1000包括与NVD 1 1001-A通信耦合的计算实例1 1000-A,NVD 1 1001-A实例化L2 VNIC 1 1002-A和L2交换机1 1004-A。VLAN 1000还包括与NVD 2 1001-B通信耦合的计算实例2 1000-B,NVD 2 1001-B实例化L2 VNIC 2 1002-B和L2交换机2 1004-A。VLAN 1000还包括在服务器机群上运行的VSRS 1010,并且其包括VSRS VNIC 1002-C和VSRS交换机1004-C。所有交换机1004-A、1004-B、1004-C一起形成分布式交换机。VSRS 1010与端点1008通信耦合,端点1008可以包括网关,并且具体而言可以包括例如另一个VSRS形式的L2/L3路由器,或者例如虚拟路由器形式的L3路由器。

[0238] 托管VLAN 1000的VCN的控制平面1001维护识别VLAN 1000上的每个L2 VNIC和L2 VNIC的网络放置的信息。例如,对于L2VNIC,这个信息可以包括与L2 VNIC相关联的接口标识符和/或托管L2 VNIC的NVD的物理IP地址。控制平面1001用这个信息更新VLAN 1000中的接口,例如周期性地更新或按需更新。因此,VLAN中的每个L2 VNIC 1002-A、1002-B、1002-C从控制平面1001接收识别VLAN中的接口的信息,并用这个信息填充表格。由L2 VNIC填充的表格可以本地存储到托管L2 VNIC的NVD。在L2 VNIC 1002-A、1002-B、1002-C已经包括当前表的情况下,L2 VNIC 1002-A、1002-B、1002-C可以确定L2 VNIC的1002-A、1002-B、1002-C当前表与从控制平面1001接收的信息/表之间的任何差异。在一些实施例中,L2 VNIC 1002-A、1002-B、1002-C可以更新其表以匹配从控制平面1001接收的信息。

[0239] 如图10中所看到的,分组经由L2交换机1004-A、1004-B、1004-C被发送,并由接收方L2 VNIC 1002-A、1002-B、1002-C接收。当分组由L2 VNIC 1002-A、1002-B、1002-C接收时,该VNIC学习源接口(源VNIC)与该分组的源MAC地址的映射。基于其从控制平面1010接收的信息的表,VNIC可以将源MAC地址(来自接收到的分组,在本文中也称为帧)映射到源VNIC的接口标识符和VNIC的IP地址和/或托管VNIC的NVD的IP地址(其中接口标识符和(一个或多个)IP地址可从表中获得)。因此,L2 VNIC 1002-A、1002-B、1002-C基于接收到的通信和/或分组学习接口标识符到MAC地址的映射,L2 VNIC 1002-A、1002-B、1002-C可以用这个学

习到的映射信息更新它的表,L2 FWD表1006-A、1006-B、1006-C。在一些实施例中,L2转发表包括MAC地址并将其与接口标识符或物理IP地址中的至少一个相关联。在此类实施例中,MAC地址是指派给L2计算实例的地址并且可以与与L2计算实例相关联的L2 VNIC模拟的端口对应。接口标识符可以唯一地识别L2 VNIC和/或L2计算实例。虚拟IP地址可以是L2 VNIC的地址。并且物理IP地址可以是托管L2 VNIC的NVD的IP地址。由L2 VNIC更新的L2转发可以本地存储在托管L2 VNIC的NVD上,并由与L2 VNIC相关联的L2虚拟交换机用于指引帧。在一些实施例中,共用VLAN内的L2VNIC可以彼此共享其映射表的全部或部分。

[0240] 鉴于上述网络体系架构,接下来描述流量流。为了解释清楚起见,结合计算实例2 1000-B、L2 VNIC 2 10002-B、L2交换机2 1004-B和NVD 2 1001-B来描述流量流。该描述等效地适用于流向和/或来自其它计算实例的流量。

[0241] 如上文所解释的,VLAN在VCN内被实现为L3物理网络之上的覆盖L2网络。VLAN的L2计算实例可以发送或接收包括覆盖MAC地址(也称为虚拟MAC地址)作为源和目的地MAC地址的L2帧。L2帧还可以封装包括覆盖IP地址(也称为虚拟IP地址)作为源和目的地IP地址的分组。在一些实施例中,计算实例的覆盖IP地址可以属于VLAN的CIDR范围。另一个覆盖网络IP地址可以属于该CIDR范围(在这种情况下,L2帧在VLAN内)或在CIDR范围之外(在这种情况下,L2帧发往另一个网络或从另一个网络接收)流动。L2帧还可以包括唯一识别VLAN的VLAN标记和可以被用于区分同一NVD上的多个L2 VNIC的VLAN标记。L2帧可以由NVD经由隧道在封装的分组中从计算实例的主机机器、从另一个NVD或从托管VSRS的服务器机群接收。在这些不同情况下,封装的分组可以是在物理网络上发送的L3分组,其中源和目的地IP地址是物理IP地址。不同类型的封装是可能的,包括Geneve封装。NVD可以解封装接收到的分组以提取L2帧。类似地,为了发送L2帧,NVD可以将其封装在L3分组中并在物理基板上发送它。

[0242] 对于来自实例2 1000-B的VLAN内出口流量,NVD 2 1001-B通过以太网链路从实例2 1000-B的主机机器接收帧。该帧包括识别L2VNIC 2 1000-B的接口标识符。该帧包括实例2 1000-B的覆盖MAC地址(例如,M.2)作为源MAC地址,以及实例1 1000-A的覆盖MAC地址(例如,M.1)作为目的地MAC地址。给定接口标识符,NVD 2 1001-B将帧传递到L2 VNIC 2 1002-B以供进一步处理。L2VNIC 2 1002-B将帧转发到L2交换机2 1004-B。基于L2转发表1006-B,L2交换机2 1004-B确定目的地MAC地址是否已知(例如,与L2转发表1006-B中的条目匹配)。

[0243] 如果已知,那么L2交换机2 1004-B确定L2 VNIC 1 1002-A是相关的隧道端点并将帧转发到L2 VNIC 1 1002-A。转发可以包括帧在分组中的封装和分组的解封装(例如, Geneve封装和解封装),其中分组包括帧、NVD 1 1001-A的物理IP地址(例如,IP.1)作为目的地地址,并且NVD 2 1001-B的物理IP地址(例如,IP.2)作为源地址。

[0244] 如果未知,那么L2交换机2 1004-B将帧广播到VLAN的各个L2 VNIC(例如,包括L2 VNIC 1 1002-A和VLAN的任何其它L2VNIC),在那里广播的帧在相关NVD之间被处理(例如,封装、发送、解封装)。在一些实施例中,可以在物理网络处执行,或更具体地说模拟,这个广播,从而将帧分别封装到每个L2 VNIC,包括VLAN中的VSRS。因此,广播是经由物理网络处的一系列复制的单播分组来模拟的。进而,每个L2 VNIC接收帧并学习L2 VNIC 21002-B的接口标识符与源MAC地址(例如,M.2)和源物理IP地址(例如,IP.2)之间的关联。

[0245] 对于从计算实例1 1000-A到计算实例2 1000-B的VLAN内入口流量,NVD 2 1001-B从NVD 1接收分组。该分组具有IP.1作为源地址和帧,其中帧包括M.2作为目的地MAC地址和

M.1作为源MAC地址。帧还包括L2 VNIC 1 1002-A的网络标识符。在解封装后,L2VNIC 2接收帧并学习到这个接口标识符与M.1和/或IP.1相关联,并且如果之前未知,那么将学习到的信息存储在交换机2处的L2转发表1006-B中以用于后续出口流量。可替代地,在解封装后,L2VNIC 2接收帧并学习到这个接口标识符与M.1和/或IP.1相关联,并且如果这个信息已知,那么刷新到期时间。

[0246] 对于从VLAN 1000中的实例2 1000-B发送到另一个VLAN中的实例的出口流量,除了使用VSRS VNIC和VSRS交换机之外,可以存在与上述出口流量类似的流。特别地,目的地MAC地址不在VLAN 1000的L2广播内(它在另一个L2 VLAN内)。因而,目的地实例的覆盖目的地IP地址(例如,IP.A)用于这个出口流量。例如,L2 VNIC 2 1002-B确定IP.A在VLAN 1000的CIDR范围之外。因而,L2 VNIC 2 1002-B将目的地MAC地址设置为默认网关MAC地址(例如,M.DG)。基于M.DG,L2交换机2 1004-B将出口流量发送到VSRS VNIC(例如,经由隧道,用适当的端到端封装)。VSRS VNIC将出口流量转发到VSRS交换机。进而,VSRS交换机执行路由功能,其中,基于覆盖目的地IP地址(例如,IP.A),VLAN 1000的VSRS交换机将出口流量发送到另一个VLAN的VSRS交换机(例如,经由这两个VLAN之间的虚拟路由器,也用适当的端到端封装)。接下来,另一个VLAN的VSRS交换机通过确定IP.A在这个VLAN的CIDR范围内来执行交换功能,并基于IP.A执行其ARP高速缓存的查找以确定与IP.A相关联的目的地MAC地址。如果ARP高速缓存中不存在匹配项,那么ARP请求被发送到另一个VLAN的不同L2 VNIC以确定目的地MAC地址。否则,VSRS交换机将出口流量发送到相关的VNIC(例如,经由隧道,用适当的封装)。

[0247] 对于从另一个VLAN中的实例到VLAN 1000中的实例的入口流量,流量流与上述类似,只是方向相反。对于从VLAN 1000中的实例到L3网络的出口流量,流量流与上述类似,只是VLAN 1000的VSRS交换机经由虚拟路由器将分组直接路由到虚拟L3网络中的目的地VNIC(例如,无需通过另一个VSRS交换机路由分组)。对于从虚拟L3网络到VLAN 1000中的实例的入口流量,流量与上述类似,只是分组由VLAN 1000A的VSRS交换机接收,该VSRS交换机在VLAN内将其作为帧发送。对于VLAN 1000与其它网络之间的流量(出口或入口),类似地使用VSRS交换机,其中其路由功能被用于出口以经由适当的网关(例如,IGW、NGW、DRG、SGW、LPG)发送分组,并且其中其交换功能用于入口以在VLAN 1000内发送帧。

[0248] 现在参考图11,示出了VLAN 1100(例如,基于云的虚拟L2网络)的实施例的示意图,并且具体而言示出了VLAN的实施方式视图。

[0249] 如上所述,VLAN可以包括“n”个计算实例1102-A、1102-B、1102-N,每个计算实例在主机机器上执行。如前面所讨论的,计算实例与主机机器之间可以存在一对一关联,或者多个计算实例与单个主机机器之间可以存在多对一关联。每个计算实例1102-A、1102-B、1102-N都可以是L2计算实例,在这种情况下,它与至少一个虚拟接口(例如,L2 VNIC)1104-A、1104-B、1104-N和交换机1106-A、1106-B、1106-N相关联。交换机1106-A、1106-B、1106-N是L2虚拟交换机并且一起形成L2分布式交换机1107。

[0250] 与主机机器上的计算实例1102-A、1102-B、1102-N相关联的L2VNIC 1104-A、1104-B、1104-N和交换机1106-A、1106-B、1106-N的对是连接到主机机器的NVD 1108-A、1108-B、1108-N上的一对软件模块。每个L2 VNIC 1104-A、1104-B、1104-N表示客户感知到的单个交换机(本文称为vswitch)的L2端口。一般而言,主机机器“i”执行计算实例“i”并连接到NVD

“i”。进而,NVD“i”执行L2 VNIC“i”和“交换机”i。L2 VNIC“i”表示vswitch的L2端口“i”。“i”是1和“n”之间的正整数。在此,同样,虽然描述了一对一的关联,但是其它类型的关联也是可能的。例如,单个NVD可以连接到多个主机,每个主机执行属于VLAN的一个或多个计算实例。如果是这样,那么NVD托管多对L2 VNIC和交换机,每对与计算实例之一对应。

[0251] VLAN可以包括VSRS 1110的实例。VSRS 1110执行交换和路由功能并且包括VSRS VNIC 1112和VSRS交换机1114的实例。VSRS VNIC 1112表示vswitch上的端口,其中这个端口经由虚拟路由器将vswitch连接到其它网络。如图所示,VSRS 1110可以在服务器机群1116上被实例化。

[0252] 控制平面1118可以跟踪识别L2 VNIC 1104-A、1104-B、1104-N及其在VLAN中的放置的信息。控制平面1110还可以将这个信息提供给VLAN中的接口1104-A、1104-B、1104-N。

[0253] 如图11中所示,VLAN可以是基于云的虚拟L2网络,其可以构建在物理网络1120之上。在一些实施例中,这个物理网络1120可以包括NVD 1108-A、1108-B、1108-N。

[0254] 一般而言,VLAN的第一L2计算实例(例如,计算实例1 1102-A)可以使用L2协议与VLAN的第二计算实例(例如,计算实例2 1102-B)通信。例如,可以通过VLAN在这两个L2计算实例之间发送帧。不过,帧可以被封装、隧道化、路由和/或进行其它处理,使得帧可以通过底层物理网络1120被发送。

[0255] 例如,计算实例1 1102-A发送以计算实例2 1102-B为目的地的帧。取决于主机机器1与NVD 1、NVD1和物理网络1120、物理网络1120NVD 2、NVD 2和主机机器2之间的网络连接(例如,TCP/IP连接、以太网连接、隧道连接等),可以对帧应用不同类型的处理。例如,帧由NVD 1接收并封装,依此类推,直到帧到达计算实例2。假设这种处理使得帧可以在底层物理资源之间发送,并且为了目的为了简洁明了,在描述VLAN和相关的L2操作时省略了它的描述。

[0256] 虚拟L2网络通信

[0257] 多种形式的通信可以发生在虚拟L2网络内或与虚拟L2网络一起发生。这些可以包括VLAN内通信。在此类实施例中,源计算实例可以将分组发送到与源计算实例(CI)处于同一VLAN中的目的地计算实例。通信还可以包括将分组发送到源CI的VLAN外部的端点。这可以包括例如第一VLAN中的源CI到第二VLAN中的目的地CI之间的通信、第一VLAN中的源CI到L3子网中的目的地CI之间的通信,和/或从第一VLAN中的源CI到包含源CI的VLAN的VCN外部的目的地CI的通信。这个通信还可以包括例如在目的地CI处从目的地CI的VLAN外部的源CI接收通信。这个源CI可以位于另一个VLAN中、L3子网中或位于包含源CI的VLAN的VCN外部。

[0258] VLAN内的每个CI可以在流量流中发挥积极作用。这包括学习接口标识符到MAC地址(本文也称为接口到MAC地址)、VLAN内实例的映射以维护VLAN内的L2转发表,以及通信分组的发送和/或接收。VSRS可以在VLAN内的通信中以及与VLAN外部的源或目的地CI的通信中发挥积极作用。VSRS可以在L2网络中和L3网络中维持存在以启用出口和入口通信。

[0259] VLAN内通信

[0260] 现在参考图12,示出了图示用于VLAN内通信的过程1200的一个实施例的流程图。在一些实施例中,过程1200可以由共用VLAN内的计算实例执行。在源CI向VLAN内的目的地CI发送分组,但不知道那个目的地CI的IP到MAC地址映射的情况下,可以具体地执行该过程。例如,当源CI向具有VLAN中IP地址的目的地CI发送分组,但源CI不知道那个IP地址的

MAC地址时,会发生这种情况。在这种情况下,可以执行ARP过程以学习目的地MAC地址和IP到MAC地址映射。

[0261] 在源CI知道IP到MAC地址映射的情况下,源CI可以将分组直接发送到目的地CI,并且不需要执行ARP过程。在一些实施例中,这个分组可以被源VNIC拦截,VLAN内通信中的源VNIC是L2VNIC。如果源VNIC知道用于目的地MAC地址的接口到MAC地址映射,那么源VNIC可以封装分组,例如在L2封装中,并且可以将封装的分组转发到目的地VNIC,VLAN内通信中的目的地VNIC是L2 VNIC,用于目的地MAC地址。

[0262] 如果源VNIC不知道用于MAC地址的接口到MAC地址映射,那么源VNIC可以执行接口到MAC地址学习过程的一方面。这可以包括源VNIC将分组发送到VLAN中的所有接口。在一些实施例中,这个分组可以经由广播被发送到VLAN内的所有接口。在一些实施例中,这种广播可以以串行单播的形式在物理网络处实现。这个分组可以包括目的地MAC和IP地址,以及源VNIC的接口、MAC地址和IP地址。VLAN中的每个VNIC都可以接收这个分组并且可以学习源VNIC的接口到MAC地址映射。

[0263] 每个接收方VNIC还可以解封装分组并将解封装的分组转发到它们相关联的CI。每个CI可以包括网络接口,该网络接口可以评估转发的分组。如果网络接口确定已接收到转发的分组的CI与目的地MAC和/或IP地址不匹配,那么丢弃该分组。如果网络接口确定已接收到转发的分组的CI与目的地MAC和/或IP地址匹配,那么该分组由CI接收。在一些实施例中,具有与分组的目的地MAC和/或IP地址相匹配的MAC和/或IP地址的CI可以向源CI发送响应,由此源VNIC可以学习目的地CI的接口到MAC地址映射,并且由此源CI可以学习目的地CI的IP到MAC地址映射。

[0264] 当源CI不知道IP到MAC地址映射时,或者当源CI的目的地CI的IP到MAC地址映射是陈旧的时,可以执行过程1200。

[0265] 因此,当IP到MAC地址映射已知时,源CI可以发送分组。当IP到MAC地址映射未知时,可以执行过程1200。当接口到MAC地址映射未知时,可以执行上面概述的接口到MAC地址学习过程。当接口到MAC地址映射已知时,VNIC可以将分组发送到目的地CI。

[0266] 过程1200开始于方框1202,其中源CI确定对于源CI来说目的地CI的IP到MAC地址映射是未知的。在一些实施例中,这可以包括源CI确定用于分组的目的地IP地址,并且确定该目的地IP地址不与存储在源CI的映射表中的MAC地址相关联。可替代地,源CI可以确定用于目的地CI的IP到MAC地址映射是陈旧的。在一些实施例中,如果映射在某个时间限制内未被更新和/或验证,那么该映射可以是陈旧的。在确定对于源CI来说目的地CI的IP到MAC地址映射是未知的和/或陈旧的后,源CI发起对目的地IP的ARP请求并发送ARP请求以进行以太网广播。

[0267] 在方框1204处,源VNIC(本文也称为源接口)从源CI接收ARP请求。源接口识别VLAN上的所有接口,并向VLAN广播域上的所有接口发送ARP请求。如前面所提到的,由于控制平面知道VLAN上的所有接口并将该信息提供给与VLAN的接口,源接口同样知道VLAN中的所有接口并且能够将ARP请求发送到VLAN中的每个接口。为了这样做,源接口复制ARP请求并针对VLAN上的每个接口封装ARP请求之一。每个封装的ARP请求包括源CI接口标识符和源CI MAC和IP地址、目标IP地址和目的地CI接口标识符。源CI接口通过将复制和封装的ARP请求作为串行单播发送到VLAN中的每个接口来复制以太网广播。

[0268] 在方框1206处,VLAN广播域中的所有接口接收并解封装分组。VLAN广播域中接收分组的每个接口学习源CI的源VNIC的接口到MAC地址映射(例如,源CI的源接口的接口标识符到MAC地址)作为识别源CI MAC和IP地址以及源CI接口标识符的分组。作为学习用于源CI的接口到MAC地址映射的一部分,每个接口可以更新它们的映射表(例如,它的L2转发表),并且可以向它的相关联的交换机和/或CI提供更新后的映射。除VSRS外,每个接收方接口都可以将解封装的分组转发到其相关联的CI。转发的解封装的分组的CI接收方并且具体而言是该CI的网络接口可以确定分组的目标IP地址是否与CI的IP地址匹配。如果与该接口相关联的CI的IP地址与接收到的分组中指定的目的地CI IP地址不匹配,那么在一些实施例中,该分组被CI丢弃,并且不采取进一步的动作。在VSRS的情况下,VSRS可以确定分组的目标IP地址是否与VSRS的IP地址相匹配。如果VSRS的IP地址与接收到的分组中指定的目标IP地址不匹配,那么在一些实施例中,该分组被VSRS丢弃并且不采取进一步的动作。

[0269] 如果确定接收到的分组中指定的目的地CI IP地址匹配与接收方接口相关联的CI(目的地CI)的IP地址,那么,并且如方框1208中所指示的,目的地CI发送响应,该响应可以是对源接口的单播ARP响应。这个响应包括目的地CI MAC地址和目的地CI IP地址,以及源CI IP和MAC地址。如下面将讨论的,如果VSRS确定目的地IP地址与VSRS IP地址相匹配,那么VSRS可以发送ARP响应。

[0270] 这个响应由封装单播ARP响应的目的地接口接收,如方框1210中所指示的。在一些实施例中,这个封装可以包括Geneve封装。目的地接口可以将封装的分组经由目的地交换机转发到源接口。封装的分组包括目的地CI MAC和IP地址和目的地CI接口标识符,以及源CI MAC和IP地址以及源CI接口标识符。

[0271] 在方框1212处,源接口接收并解封装ARP响应。源接口可以基于封装和/或封装的分组中包含的信息进一步学习目的地CI的接口到MAC地址映射。在一些实施例中,源接口可以将ARP响应转发到源CI。

[0272] 在方框1214处,源CI接收ARP响应。在一些实施例中,源CI可以基于包含在ARP响应中的信息来更新映射表,并且具体而言是基于目的地CI的MAC和IP地址更新映射表以反映IP到MAC地址映射。随后,源CI然后可以发送分组,该分组可以是包括IP分组的任何分组,并且具体而言是到目的地CI的IPv4或IPv6分组。这个分组可以包括源CI的MAC地址和IP地址作为分组的源MAC地址和源IP地址,以及目的地CI的MAC地址和IP地址作为目的地MAC地址和目的地IP地址。

[0273] 在方框1216处,源接口可以从源CI接收分组。源接口可以封装分组,并且在一些实施例中,可以用Geneve封装来封装分组。源接口可以将封装的分组转发到目的地CI,并且具体而言是转发到目的地接口。封装的分组可以包括源CI的IP和MAC地址以及接口标识符作为源MAC地址、源IP地址和源接口标识符,以及目的地CI的MAC地址、IP地址和接口标识符作为目的地MAC地址、IP地址和目的地接口标识符。

[0274] 在方框1218处,目的地接口从源接口接收分组。目的地接口可以解封装分组,然后将分组转发到目的地CI。在方框1220处,目的地CI从目的地接口接收分组。

[0275] 现在参考图13,示出了用于VLAN内通信的过程1300的示意图1300。如所看到的,VLAN A 1302的VLAN CIDR为10.0.3.0/24。VLAN A 1302包括VSRS VNIC(VRVI) 1304,其可以在一个或几个硬件上被实例化,并且具体而言可以在服务器机群1306上被实例化。VRVI

1304可以具有10.0.3.1的IP地址。VLAN可以包括具有IP地址10.0.3.2并且与可以实例化L2 VNIC 1(VI1) 1314和L2交换机1的NVD 1(SN1) 1312通信耦合的计算实例1(CI1) 1310。VLAN可以包括具有IP地址10.0.3.3并且与可以实例化L2 VNIC 2(VI2) 1324和L2交换机2的NVD 2(SN2) 1322通信耦合的计算实例2(CI2) 1320。VLAN可以包括具有IP地址10.0.3.4并且与可以实例化L2 VNIC 3(VI3) 1334和L2交换机3的NVD 3(SN3) 1332通信耦合的计算实例3(CI3) 1330。

[0276] 在图13的示例中,并且应用图12的方法1200,CI3 1330是源CI并且VI3 1334是源接口。另外,CI2 1320是目的地CI并且VI2 1324是目的地接口。CI3确定它没有用于目的地IP地址(10.0.3.3)的IP到MAC映射,导致CI3 1330发送ARP请求。ARP请求可以针对已知地址,并且具体而言可以针对CI2 1320的已知IP地址。因此,在一些实施例中,ARP请求可以针对10.0.3.3。

[0277] 这个ARP请求由SN3 1332和VI3 1334接收。VI3 1334复制ARP请求以创建用于VLAN 1302中的每个CI的ARP请求。VI3 1334封装每个ARP请求并向VLAN中的每个接口发送ARP请求。这些封装的ARP请求可以包括识别源CI(CI3)的MAC地址、源接口(VI3)的接口标识符和目的地MAC地址的信息。这些请求可以被发送到VLAN中的每个接口,如箭头1350所示。在VLAN中,这些请求可以是广播ARP请求。

[0278] VLAN中的接口接收封装的ARP请求,并且解封封装ARP请求。基于包含在ARP请求中和/或与ARP请求相关联的信息,VLAN中的接口更新它们的映射。具体而言,例如,VI1 1314、VI2 1324和VRVI 1304中的每一个接收来自CI 3 1330的ARP请求,解封封装ARP请求,并学习源CI的接口到MAC地址的映射,接口标识符和MAC地址都包括在封装的分组中。VRVI 1304可以基于封装的分组中包含的信息进一步更新用于源CI的IP到MAC地址映射。

[0279] 具有带所请求IP地址的CI的VLAN中的接口可以如箭头1352所指示的那样向CI 3 1330发送ARP响应。具体而言,如图13中所示,VI2 1324是CI2 1320的接口,因此可以发送ARP响应。来自CI2的ARP响应可以由VI2接收,可以被封装,并且可以作为ARP单播被发送到发出请求的接口,具体而言是VI3。如本申请中早先所指示的,这个ARP响应的发送可以包括向相关联的交换机提供封装的ARP响应,该交换机可以将封装的ARP响应发送到VI3。

[0280] ARP响应可以由VI3接收并且可以被解封封装。VI3可以基于接收到的ARP响应学习CI2的接口到MAC映射,并且可以将更新后的学习到的映射提供给VI3的相关联的交换机。可以将解封封装的分组提供给CI3,CI3可以基于解封封装的分组学习CI2 1320的IP到MAC地址映射。CI3可以向CI2发送分组,该分组可以是IP分组,诸如IPv4或IPv6分组。这个分组可以将CI3的MAC和IP地址作为源地址,并且可以将CI2的IP和MAC地址作为目的地地址。

[0281] 由CI3发送的分组可以由VI3接收,VI3可以封装分组并将分组转发到接口VI2。VI2可以接收分组,可以解封封装分组,并且可以将分组转发到CI2。

[0282] VLAN间通信

[0283] 现在参考图14,示出了图示用于虚拟L2网络中的VLAN间通信的过程1400的一个实施例的流程图。过程1400可以由两个连接的VLAN的全部或部分(诸如图8中所示的多个连接的L2 VLAN 800)执行。在一些实施例中,过程1400可以在第一VLAN(源VLAN)中的计算实例(源CI)将分组发送到第二VLAN(目的地VLAN)中的目的地计算实例(目的地CI)时执行。在一些实施例中,源CI可以基于目的地CI的IP地址确定目的地CI在源VLAN外部。例如,源CI可以

确定目的地IP在源VLAN CIDR外部。在这种情况下,源CI可以确定经由源VLAN的VSRS将IP分组发送到目的地CI。如果源CI已经知道第一VLAN(源VSRS)中VSRS的映射,那么源CI可以将分组直接指向源VSRS。如果源CI不知道源VSRS的映射,那么源CI及其相关联的VNIC(源接口或源VNIC)首先学习源VSRS的映射。在VLAN间通信的实施例,源和目的地VNIC都是L2 VNIC。过程1400的第一步(步骤1402至1410)涉及由源VNIC和源CI对源VSRS映射的学习。

[0284] 过程1400开始于方框1402,其中具有目的地IP地址的源CI发起ARP请求。ARP请求被用于确定源VSRS的IP->MAC地址映射。ARP请求由源CI发送到以太网广播。ARP请求包括源CI的IP地址作为源IP以及MAC地址。

[0285] 在方框1404处,源VNIC接收ARP请求并复制ARP请求。具体而言,源VNIC接收来自源CI的ARP请求,识别VLAN上的所有接口,并将ARP请求发送到VLAN广播域上的所有接口。如前面所提到的,由于控制平面知道VLAN上的所有接口并将该信息提供给与VLAN的接口,因此源接口同样知道VLAN中的所有接口并且能够将ARP请求发送到VLAN中的每个接口。为了这样做,源接口复制ARP请求并针对VLAN上的每个接口封装ARP请求之一。每个封装的ARP请求包括源CI接口标识符和源CI IP和MAC地址作为源地址,以及目的地CI接口标识符和IP地址作为目的地地址。源CI接口通过经由串行单播向VLAN中的每个接口发送复制和封装的ARP请求来实现以太网广播。在一些实施例中,源VNIC可以用Geneve封装来封装ARP请求。

[0286] 在方框1406处,VLAN广播域中的所有接口都接收并解封装分组。VLAN广播域中接收分组的每个接口学习源CI的源VNIC的接口->MAC地址映射,因为分组识别出源CI MAC和IP地址以及源接口标识符。作为学习用于源CI的接口->MAC地址映射的一部分,每个接口可以更新它们的映射表,并且可以向它们相关联的交换机和/或CI提供更新后的映射表。除VSRS外,每个接收方接口都将解封装的分组转发到其相关联的CI。转发的解封装的分组的CI接收方(具体而言是该CI的网络接口)可以确定分组的目标IP地址是否与CI的IP地址匹配。如果与该接口相关联的CI的IP地址与接收到的分组中指定的目的地CI IP地址不匹配,那么不采取进一步的动作。

[0287] 源VSRS确定目标IP地址与源VSRS的IP地址匹配,并且如步骤1408中所指示的,源VSRS封装并发送响应,该响应可以是对源接口的单播ARP响应。这个响应包括源CI MAC地址、IP地址和源CI接口标识符作为目的地地址。该响应还包括源VSRS MAC地址、IP地址和VSRS接口标识符作为源地址。在一些实施例中,ARP响应的封装可以包括Geneve封装。

[0288] 在方框1410处,源接口接收并解封装ARP响应。源接口还可以基于封装中和/或封装的分组中包含的信息学习用于源VSRS的接口到MAC地址映射。在一些实施例中,源接口可以将ARP响应转发到源CI。

[0289] 在方框1412处,源CI接收ARP响应。在一些实施例中,源CI可以基于包含在ARP响应中的信息并且具体而言是基于源VSRS的MAC地址和IP地址更新映射表。在一些实施例中,例如,源CI可以更新其映射表以反映源VSRS的IP到MAC地址映射。然后,源CI可以向源VSRS发送分组,该分组可以是任何分组,包括IP分组,具体而言是IPv4或IPv6分组。在一些实施例中,这可以包括发送具有目的地CI的IP地址作为目的地地址的IP分组。在一些实施例中,目的地CI的IP地址可以包含在分组的报头中,诸如在例如分组的L3报头中。报头还可以在分组的另一个报头中包括源VSRS的MAC地址,诸如例如在分组的L2报头中。分组还可以包括源CI的MAC地址和IP地址作为源MAC地址和源IP地址。

[0290] 在方框1414处,源接口可以从源CI接收分组。源接口可以封装分组。源接口可以将封装的分组转发到源VSRS,并且具体而言是转发到源VSRS VNIC。封装的分组除了分组的地址外还可以包括源CI的MAC地址和接口标识符作为源MAC地址和源接口标识符,以及源VSRS的MAC地址和接口标识符作为目的地MAC地址和目的地接口标识符。

[0291] 在方框1416处,源VSRS接收封装的分组。源VSRS解封装分组并剥离分组中与源VSRS相关的任何地址信息,包括例如源VSRS IP地址、MAC地址和/或源VSRS接口标识符。源VSRS识别分组的目的地CI。在一些实施例中,源VSRS基于包括在分组中的目的地CI的IP地址来识别分组的目的地CI。源VSRS查找用于分组的目的地IP地址的映射。如果IP地址在VCN的IP地址空间内,那么源VSRS在用于VCN的IP地址空间中查找用于分组的目的地IP地址的映射。源VSRS然后用L3封装重新封装分组。在一些实施例中,这种L3封装可以包括例如MPLSoUDPL3封装。然后源VSRS将分组转发到目的地VLAN的VSRS(目的地VSRS)。在一些实施例中,源VSRS可以将分组转发到目的地VSRS,使得目的地VSRS是分组的隧道端点(TEP)。L3封装的分组包括源CI IP地址和MAC地址,以及目的地CI的IP地址。

[0292] 在方框1418处,目的地VSRS接收分组并解封装分组。在一些实施例中,这种解封装可以包括移除L3封装。如果目的地VSRS知道用于目的地CI的IP到MAC和MAC到接口映射,那么目的地VSRS识别目的地VLAN内目的地CI的接口和MAC地址。可替代地,如果目的地VSRS不知道目的地CI映射,那么可以执行过程1600的步骤1612至1622。

[0293] 在方框1420处,目的地VSRS将分组转发到目的地CI的目的地接口。在一些实施例中,这可以包括用L2封装来封装分组。在一些实施例中,分组可以包括源CI的IP地址和MAC地址和/或目的地VSRS的MAC地址和接口标识符。在一些实施例中,分组还可以包括目的地CI的MAC地址和接口标识符。

[0294] 在方框1422处,目的地接口接收并解封装分组。具体而言,目的地VNIC移除L2封装。在一些实施例中,目的地VNIC将分组转发到目的地CI。在方框1424处,目的地CI接收分组。

[0295] 现在参考图15,示出了用于VLAN间通信的过程1400的示意图1500。如所看到的,VLAN A 1502-A的VLAN CIDR为10.0.3.0/24。VLAN A 1502-A包括VSRS VNIC A (VRVI A) 1504-A,其可以在一个或多个硬件上被实例化,并且具体而言可以在服务器机群1506上被实例化。VRVI A 1504-A可以具有IP地址10.0.3.1。VLAN可以包括具有IP地址10.0.3.2并且与可以实例化L2 VNIC 1 (VI1) 1514和L2交换机1的NVD 1 (SN1) 1512通信耦合的计算实例1 (CI1) 1510。VLAN可以包括具有IP地址10.0.3.3并且与可以实例化L2VNIC 2 (VI2) 1524和L2交换机2的NVD 2 (SN2) 1522通信耦合的计算实例2 (CI2) 1520。VLAN可以包括具有IP地址10.0.3.4并且与可以实例化L2 VNIC 3 (VI3) 1514和L2交换机3的NVD 3 (SN3) 1532通信耦合的计算实例3 (CI3) 1530。

[0296] VLAN B 1502-B具有10.0.34.0/24的VLAN CIDR。VLAN B 1502-B包括VSRS VNIC B (VRVI B) 1504-B,其可以在一个或多个硬件上被实例化,并且具体而言可以在服务器机群1506上被实例化。VRVI B 1504-B的IP地址可以是10.0.4.1。VLAN可以包括具有IP地址10.0.4.2并且与可以实例化L2 VNIC 4 (VI3) 1544和L2交换机4的NVD 4 (SN4) 1542通信耦合的计算实例4 (CI4) 1540。

[0297] 在图15的示例中,并且应用图14的方法1400,CI3 1530是源CI并且VI3 1534是源

接口。另外,CI4 1540是目的地CI并且VI41544是目的地接口。CI3确定它没有用于VRVI A 1504-A的IP到MAC映射,导致CI3 1530发送ARP请求。ARP请求可以针对地址,并且具体而言可以针对VRVI A 1504-A的已知IP地址。因此,在一些实施例中,ARP请求可以针对10.0.3.1。

[0298] 这个ARP请求由SN3 1532和VI3 1534接收。VI3 1534复制ARP请求以便为VLAN 1502-A中的每个CI创建ARP请求。VI3 1534用L2封装来封装每个ARP请求并将ARP请求发送到VLAN中的每个接口。这些封装的ARP请求可以包括识别源CI(CI3)的MAC和IP地址、源接口VI3的接口标识符和目标IP地址的信息。这些请求可以如箭头1550所示被发送到VLAN中的每个接口,并且具体而言可以作为串行单播被发送,使得VLAN中的每个接口接收ARP请求。

[0299] VLAN中的接口接收封装的ARP请求,并且解封装ARP请求。基于包含在ARP请求中和/或与ARP请求相关联的信息,VLAN中的接口更新它们的映射。具体而言,例如,VI1 1514、VI2 1524和VRVI A 1504-A中的每一个接收来自CI 3 1530的单播ARP请求,解封装该单播ARP请求,并学习源CI的接口到MAC地址的映射,接口标识符和MAC地址都包括在封装的分组中。

[0300] VRVI A 1504-A确定其IP地址与所请求的IP地址匹配,并如箭头1552所指示的那样向CI 3 1530发送ARP响应。来自VRVI A 1504-A的ARP响应可以由VRVI A 1504-A用例如L2封装来封装,并且可以作为ARP单播发送到发出请求的接口,具体而言是VI3。如本申请中早先所指示的,这个ARP响应的发送可以包括向VRVI A 1504-A的相关联的交换机提供封装的ARP响应,该交换机可以将封装的ARP响应发送到VI3。

[0301] ARP响应可以由VI3接收并且可以被解封装。VI3可以基于接收到的ARP响应学习VRVI A 1504-A的接口到MAC的映射,并且可以向VI3的相关联的交换机提供更新后的学习到的映射。可以将解封装的分组提供给CI3。CI3可以学习VRVI A 1504-A的IP地址到MAC地址映射,可以发送分组,该分组可以是IP分组,诸如IPv4或Ipv6分组。这个分组可以具有CI3的MAC地址和IP地址作为源地址。这个分组还可以包括目的地CI(CI4 1540)的IP地址作为目的地地址,并且可以具有VRVI A 1504-A MAC地址和IP地址。

[0302] 由CI3发送的分组可以由VI3接收,VI3可以封装该分组并将该分组转发到VRVI A 1504-A。VRVI A 1504-A可以接收分组,可以解封装分组,并且可以查找分组的目的地IP地址(CI4的IP地址)的映射。在一些实施例中,分组的解封装可以包括从分组中剥离L2报头,或者换句话说,从报头中剥离与VLAN A 1502-A相关的信息。这个被剥离的信息可以包括例如CI3的MAC地址、VI3的接口的接口标识符和/或VRVI A 1504-A的MAC地址和接口标识符。在一些实施例中,VSRs以及因此VRVI A 1504-A可以驻留在L2网络和L3网络两者中。在L2网络中,因此在VLAN内,VSRs可以使用L2通信协议,而VSRs在与L3网络通信时可以使用L3通信协议。与在VLAN中执行的学习相反,VSRs可以从控制平面学习到L3网络中的端点的映射。在一些实施例中,例如,控制平面可以提供映射L3网络中的实例的IP地址和/或MAC地址的信息。

[0303] VRVI A 1504-A可以查找分组中包含的IP目的地地址的映射,或者换句话说,可以查找针对目的地CI的IP地址的映射。在一些实施例中,查找这个映射可以包括将VRVI B 1504-B识别为VLAN B 1502-B的VSRs。VRVI A 1504-A可以封装分组并将封装的分组转发到VRVI B 1504-B,VRVI B 1504-B可以是用于目的地VLAN和/或目的地接口的隧道端点

(TEP)。这个封装的分组的转发由方框1556指示。这个转发的分组可以包括作为源地址的源CI的IP地址CI3,以及作为目的地地址的目的地CI(CI4)的IP地址。

[0304] VRVI B 1504-B可以接收和解封装分组。VRVI B 1504-B还可以识别与目的地IP地址对应的VLAN B 1502-B内的接口。VRVI B1504-B可以封装分组并添加L2报头以在VLAN B 1504-B内建立隧道。VRVI B 1504-B然后将分组转发到目的地CI。目的地接口VI4 1544可以接收和解封装分组,然后可以将以太网帧或分组转发到目的地CI 1540。

[0305] 入口分组流

[0306] 现在参考图16,示出了图示用于入口分组流的过程1600的一个实施例的流程图。具体而言,图16示出了用于来自子网的入口分组的过程1600的一个实施例。该过程可以由系统600的全部或部分执行,并且具体而言可以由VLAN的实体和一些(对VLAN来说)外部源CI执行,该源CI可以驻留在L3子网上。

[0307] 过程1600开始于方框1602,其中源L3 CI确定将分组发送到目的地CI,并且具体而言是发送到VLAN内的目的地IP地址。源L3CI不知道该映射,因此发送对包含源L3 CI的子网的虚拟路由器的MAC地址的ARP请求。在一些实施例中,这个ARP请求的发送可以包括将ARP请求发送到以太网广播。在步骤1604处,源L3 CI的源接口用VR的MAC地址回复ARP请求。在一些实施例中,源接口可以基于由源接口访问的映射信息确定VR的MAC地址。

[0308] 在从源接口接收到ARP回复后,源L3 CI将IP分组发送到VR,如方框1606中所指示的。在一些实施例中,这可以包括源L3 CI发送分组,并且源L3接口接收、封装和转发这个分组。在一些实施例中,源L3接口可以用L3封装来封装分组,L3封装包括从L3报头开始的原始分组。源L3接口可以将分组转发到VR。这个IP分组可以被发送到VR MAC地址和VR接口,并且可以包括源L3 CI MAC地址作为MAC地址和源L3 CI接口标识符作为源接口标识符。

[0309] VR接收并解封装分组。然后VR可以查找用于分组目的地的IP地址的VNIC映射。VR可以确定分组在VLAN CIDR中,然后可以封装分组并将封装的分组转发到包含目的地CI的VLAN的VSRS。在一些实施例中,VR可以用L3封装来封装分组。封装的分组可以包括目的地IP地址、VSRS的MAC地址和接口标识符,以及源IP地址。

[0310] VSRS接收并解封装分组,如方框1610中所指示的。当VSRS知道目的地CI的映射,并且具体地是知道接收到的分组中的目的地IP地址的映射时,VSRS将IP分组转发到与目的地IP地址对应的CI的TEP。这可以包括生成和封装具有与目的地CI的MAC地址对应的目的地MAC和与目的地接口对应的目的地接口标识符的L2分组。在入口分组流的实施例中,目的地接口是L2 VNIC。

[0311] 如果VSRS不知道目的地CI映射,那么过程1600在方框1612处继续,其中VSRS接收并解封装IP分组。

[0312] 当VSRS不知道目的地MAC地址到VNIC映射时,VSRS可以执行接口到MAC地址学习过程。这可以包括VSRS将分组发送到VLAN内的所有接口。在一些实施例中,这个分组可以经由广播被发送到VLAN内的所有接口。这个分组可以包括目的地MAC和IP地址,以及VSRS的接口标识符和MAC地址以及源CI的IP地址。VLAN中的每个VNIC都可以接收这个分组并且可以学习VSRS的接口到MAC地址映射。

[0313] 每个接收方VNIC还可以解封装分组并将解封装的分组转发到它们相关联的CI。每个CI可以包括网络接口,该网络接口可以评估转发的分组。如果网络接口确定已接收到转

发的分组的CI与目的地MAC和/或IP地址不匹配,那么丢弃该分组。如果网络接口确定已接收到转发的分组的CI与目的地MAC和/或IP地址匹配,那么该分组被CI接收。在一些实施例中,具有与分组的目的地MAC和/或IP地址匹配的MAC和/或IP地址的CI可以向VSRS发送响应,由此VSRS可以学习目的地CI的接口到MAC地址映射和IP到MAC地址映射。

[0314] 可替代地,当VSRS不知道目的地CI映射,并且具体而言不知道目的地IP地址到MAC地址的映射时,VSRS挂起该IP分组。然后VSRS对VSRS的VLAN广播域中的所有接口生成对目的地IP地址的ARP请求。这个ARP请求包括VSRS MAC作为源MAC,VSRS接口的接口标识符作为源接口,以及目的地IP地址作为目标IP地址。在一些实施例中,这个ARP请求可以广播到VLAN中的所有接口。

[0315] 在方框1614处,VLAN广播域中的所有接口都接收并解封装分组。接收分组的VLAN广播域中的每个接口都学习VSRS的接口→MAC地址映射,因为分组识别出VSRS MAC地址和接口标识符。作为学习用于VSRS的接口→MAC地址映射的一部分,每个接口都可以更新它们的映射表,并且可以向它们相关联的交换机和/或CI提供更新后的映射表。

[0316] 每个接收方接口可以将解封装的分组转发到它们相关联的CI。转发的解封装的分组的CI接收方并且具体而言是该CI的网络接口可以确定分组的目的地IP地址是否与CI的IP地址相匹配。如果与该接口相关联的CI的IP地址与接收到的分组中指定的目的地CI IP地址不匹配,那么不采取进一步的动作。

[0317] 如果确定接收到的分组中指定的目的地CI IP地址匹配与接收方接口相关联的CI(目的地CI)的IP地址,那么,并且如方框1616中所指示的,目的地CI发送响应,该响应可以是对源接口的单播ARP响应。这个响应包括目的地CI MAC和IP地址,以及VSRS MAC和IP地址。这个响应由封装单播ARP响应的目的地接口接收,如方框1618中所指示的。目的地接口可以经由目的地交换机将封装的分组转发到VSRS。封装的分组包括目的地CI MAC和IP地址和目的地CI接口标识符,以及VSRS MAC和IP地址和VSRS接口标识符。

[0318] 在方框1620处,VSRS并且具体而言是VSRS接口接收并解封装ARP响应。VSRS并且具体而言是VSRS接口可以基于封装中和/或封装的分组中包含的信息进一步学习用于目的地CI的接口到MAC地址映射。

[0319] 在方框1622处,VSRS然后可以封装L2报头并将其添加到之前挂起的IP分组,然后将之前挂起的IP分组转发到目的地CI,并且具体而言是转发到目的地接口。目的地接口可以解封装分组并将解封装的分组提供给目的地CI。由VSRS转发的这个分组可以将VSRS的MAC地址和接口标识符作为源MAC地址和源接口标识符,将目的地CI的MAC地址和接口标识符作为目的地MAC地址和目的地接口标识符。这个分组还可以包括目的地CI的IP地址和源CI的IP地址。

[0320] 目的地接口从VSRS接收分组,然后解封装分组。这种解封装可以包括移除由VSRS添加的报头,这些报头可以是VCN报头。然后目的地接口可以将分组转发到目的地CI,并且目的地CI可以从目标接口接收分组。

[0321] 现在参考图17,示出了用于入口通信的过程1600的示意图1700。如所看到的,VLAN A 1502-A的VLAN CIDR为10.0.3.0/24。VLAN A 1502-A包括VSRS VNIC A(VRVI A) 1504-A,其可以在一个或多个硬件上被实例化,并且具体而言可以在服务器机群1506上被实例化。VRVI A 1504-A可以具有IP地址10.0.3.1。VLAN可以包括具有IP地址10.0.3.2并且与可以

实例化L2 VNIC 1 (VI1) 1514和L2交换机1的NVD 1 (SN1) 1512通信耦合的计算实例1 (CI1) 1510。VLAN可以包括具有IP地址10.0.3.3并且与可以实例化L2 VNIC 2 (VI2) 1524和L2交换机2的NVD 2 (SN2) 1522通信耦合的计算实例2 (CI2) 1520。VLAN可以包括具有IP地址10.0.3.4并且与可以实例化L2 VNIC 3 (VI3) 1514和L2交换机3的NVD 3 (SN3) 1532通信耦合的计算实例3 (CI3) 1530。

[0322] 可以是L3计算实例4 (CI4) 1744的计算实例可以驻留在VLAN A 1702-A外部的子网1739上。CI4可以具有IP地址10.0.4.4,并且可以与可以实例化L3 VNIC 4 (VI4) 1744的NVD 4 (SN4) 1742通信耦合。

[0323] 在图17的示例中,并且应用图16的方法1600,CI4 1740是源CI并且VI4 1744是源接口。另外,CI3 1730是目的地CI并且VI3 1734是目的地接口。CI4确定它不知道向CI 3发送分组的映射。因此,CI4发送对子网虚拟路由器的IP地址的ARP请求。响应于这个请求,在一些实施例中可以驻留在包含子网虚拟路由器实例的NVD上的VI4直接用VR IP地址回复CI4。CI4学习VR IP地址,并将分组发送到VR。VR接收分组,然后基于映射信息封装分组并将分组转发到VSRS,如箭头1748所指示的。

[0324] VSRS接收分组,并且如果VSRS不知道映射,并且具体而言是不知道到目的地CI的IP到MAC地址映射,那么VSRS挂起分组,并向VLAN A 1704-A中的所有接口发送ARP请求。这个ARP请求包括VSRS MAC地址和接口标识符,并且VLAN A 1704-A中的每个接收方接口基于ARP请求学习VSRS的映射。每个接口同样解封装分组并将分组发送到其CI。在接收到解封装的分组后,CI3确定它是分组中识别出的CI,并且CI3生成用其MAC地址回答的ARP回复。VI3接收并封装ARP回复,并将封装的ARP回复转发到VSRS。封装的ARP回复包括VSRS的MAC地址和接口标识符,以及CI3的接口的MAC地址和接口标识符。

[0325] VSRS接收ARP回复并学习用于CI3的IP地址到MAC地址和MAC地址到接口映射。然后VSRS将先前挂起的分组转发到CI3,该分组由VI3接收、解封装并转发到CI3。

[0326] 出口分组流

[0327] 现在参考图18,示出了图示用于来自VLAN的出口分组流的过程1800的一个实施例的流程图。在一些实施例中,分组可以从VLAN流出以流向另一个VLAN、子网或另一个网络。过程1800可以由系统600的全部或部分执行,并且具体而言可以由VLAN的实体执行。在一些实施例中,该过程的部分可以由一些(对于源VLAN来说)外部目的地CI执行,该目的地CI可以驻留在L3子网上。

[0328] 在一些实施例中,当VLAN(源VLAN)中的计算实例(源CI)向VLAN外部的目的地计算实例(目的地CI)发送分组时,可以执行过程1800。如果源CI已经知道第一VLAN中VSRS(源VSRS)的映射,那么源CI可以将分组直接指向源VSRS。如果源CI不知道源VSRS的映射,那么源CI及其相关联的VNIC(源接口或源VNIC)首先学习源VSRS的映射,具体而言是IP到MAC地址映射。在来自L2 VLAN的出口分组流的实施例中,源VNIC是L2 VNIC。过程1800的第一步(步骤1802至1810)涉及由源VNIC和源CI对源VSRS映射的学习。

[0329] 过程1800开始于方框1802,其中源CI发起ARP请求。ARP请求被用于确定源VSRS的IP到MAC地址映射。ARP请求由源CI发送到以太网广播。ARP请求包括源CI MAC和IP地址以及源CI的接口标识符作为源地址和源接口标识符。ARP请求还包括源VSRS的IP地址。

[0330] 在方框1804处,源VNIC接收ARP请求并复制ARP请求。具体而言,源VNIC从源CI接收

ARP请求,识别出VLAN上的所有接口,并将ARP请求发送到VLAN广播域上的所有接口。如前面所提到的,由于控制平面知道VLAN上的所有接口并将该信息提供给与VLAN的接口,因此源接口同样知道VLAN中的所有接口并且能够将ARP请求发送到VLAN中的每个接口。为了这样做,源接口复制ARP请求并为VLAN上的每个接口封装ARP请求之一。每个封装的ARP请求包括作为源地址的源CI接口标识符和源CI MAC和/或IP地址,以及作为目的地地址的目的地CI接口标识符。源CI接口通过经由串行单播发送复制和封装的ARP请求来复制以太网广播。在一些实施例中,源VNIC可以用Geneve封装来封装ARP请求。

[0331] 在方框1806处,VLAN广播域中的所有接口都接收并解封装分组。接收分组的VLAN广播域中的每个接口都学习源CI的接口到MAC地址映射,因为分组识别出源CI MAC地址和接口标识符。除此之外,VSRS还学习源CI的IP到MAC地址映射。作为学习用于源CI的接口到MAC地址映射的一部分,每个接口都可以更新它们的映射表,并且可以向它们相关联的交换机和/或CI提供更新后的映射表。除VSRS外,每个接收方接口都可以将解封装的分组转发到其相关联的CI。转发的解封装的分组的CI接收方并且具体而言是该CI的网络接口可以确定分组的目的地IP地址是否与CI的IP地址相匹配。如果与该接口相关联的CI的IP地址与接收到的分组中指定的目的地CI IP地址不匹配,那么不采取进一步的动作。

[0332] 源VSRS确定目的地IP地址与源VSRS的IP地址匹配,并且如步骤1808中所指示的,源VSRS封装并发送响应,该响应可以是对源接口的单播ARP响应。这个响应包括作为目的地地址的源CI MAC地址和IP地址,以及作为目的地接口标识符的源CI接口标识符。响应还包括作为源地址的源VSRS MAC地址和IP地址,以及作为源接口标识符的源VSRS接口标识符。

[0333] 在方框1810处,源接口接收并解封装ARP响应。源接口还可以基于封装中和/或封装的分组中包含的信息来学习源VSRS的接口到MAC地址映射。在一些实施例中,源接口可以将ARP响应转发到源CI。

[0334] 在方框1812处,源CI接收ARP响应。在一些实施例中,源CI可以基于包含在ARP响应中的信息并且具体而言是基于源VSRS的MAC地址和源VSRS的IP地址来更新映射表。然后,源CI可以向源VSRS发送分组,该分组可以是任何分组,包括IP分组,并且具体而言是IPv4或IPv6分组。在一些实施例中,这可以包括发送具有源VSRS MAC地址和源VSRS IP地址的目的地地址的IP分组。分组还可以包括源CI的MAC地址和IP地址作为源地址。

[0335] 在方框1814处,源接口从源CI接收分组。源接口封装分组。源接口可以将封装的分组转发到源VSRS,并且具体而言是转发到源VSRS VNIC。封装的分组可以包括源CI的MAC地址和接口标识符作为源MAC地址和源接口标识符,以及源VSRS的MAC地址和接口标识符作为目的地MAC地址和目的地接口标识符。封装的分组还可以包括源CI的IP地址和源VSRS的IP地址。

[0336] 在方框1816处,源VSRS接收封装的分组。源VSRS识别分组的目的地CI。在一些实施例中,源VSRS基于包括在分组中的目的地CI的IP地址来识别分组的目的地CI。源VSRS查找用于分组的目的地IP地址的映射。如果IP地址在VCN的IP地址空间内,那么源VSRS在VCN的IP地址空间中查找用于分组的目的地IP地址的映射。源VSRS然后用L3封装来重新封装分组。

[0337] 然后,源VSRS将分组转发到目的地CI。在一些实施例中,这可以包括将封装的分组转发到与包含目的地CI的子网相关联的VR,和/或将分组转发到网关以允许分组退出VCN。

在一些实施例中,将分组转发到目的地CI可以包括将分组转发到与目的地CI相关联的TEP。L3封装的分组包括源CI的IP地址作为源地址,以及目的地CI的IP地址作为目的地地址。

[0338] 在方框1820处,目的地接口接收并解封装分组。在一些实施例中,目的地VNIC将分组转发到目的地CI。在方框1822处,目的地CI接收分组。

[0339] 现在参考图19,示出了用于出口分组流的过程1800的示意图1900。如图所示,VLAN A 1502-A的VLAN CIDR为10.0.3.0/24。VLAN A 1502-A包括VSRS VNIC A (VRVI A) 1504-A,其可以在一个或多个硬件上被实例化,具体而言可以在服务器机群1506上被实例化。VRVI A 1504-A可以具有IP地址10.0.3.1。VLAN可以包括具有IP地址10.0.3.2并且与可以实例化L2 VNIC 1 (VI1) 1514和L2交换机1的NVD 1 (SN1) 1512通信耦合的计算实例1 (CI1) 1510。VLAN可以包括具有IP地址10.0.3.3并且与可以实例化L2 VNIC 2 (VI2) 1524和L2交换机2的NVD 2 (SN2) 1522通信耦合的计算实例2 (CI2) 1520。VLAN可以包括具有IP地址10.0.3.4并且与可以实例化L2 VNIC 3 (VI3) 1514和L2交换机3的NVD 3 (SN3) 1532通信耦合的计算实例3 (CI3) 1530。

[0340] 可以是L3计算实例4 (CI4) 1944的计算实例可以驻留在VLAN A 1902-A外部的子网1939上。CI4可以具有IP地址10.0.4.4,并且可以与可以实例化L3 VNIC 4 (VI4) 1944的NVD 4 (SN4) 1942通信耦合。子网1939可以包括虚拟路由器 (VR) 1948。VR 1948可以具有IP地址10.0.4.1。VR 1948可以在例如SmartNIC、服务器、服务器机群等上被实例化。

[0341] 在图19的示例中,并且应用图18的方法1800,CI3 1930是源CI并且VI3 1934是源接口。另外,CI4 1940是目的地CI并且VI4 1944是目的地接口。CI3确定它没有用于VRVI A 1904-A的IP到MAC映射,导致CI3 1930发送ARP请求。ARP请求可以针对已知地址,并且具体而言可以针对VRVI A 1904-A的已知IP地址。因此,在一些实施例中,ARP请求可以针对10.0.3.1。

[0342] 这个ARP请求由SN3 1932和VI3 1934接收。VI3 1934复制ARP请求以便为VLAN 1902-A中的每个CI创建ARP请求。VI3 1934用L2封装来封装每个ARP请求并将ARP请求发送到VLAN中的每个接口。这些封装的ARP请求可以包括识别源CI (CI3) 的MAC地址、源接口标识符VI3和目标IP地址的信息。这些请求可以如箭头1950所示发送到VLAN中的每个接口,并且具体而言可以广播,使得VLAN中的每个接口接收ARP请求。

[0343] VLAN中的接口接收封装的ARP请求,并解封装ARP请求。基于包含在ARP请求中和/或与ARP请求相关联的信息,VLAN中的接口更新它们的映射。具体而言,例如,VI1 1914、VI2 1924、VRVI A 1904-A中的每一个接收来自CI 3 1930的单播ARP请求,解封装该单播ARP请求,并且学习源CI的接口到MAC地址的映射,接口标识符和MAC地址都包括在封装的分组中。

[0344] VRVI A 1904-A确定其IP地址与所请求的IP地址匹配,并如箭头1952所指示的向CI 3 1930发送ARP响应。来自VRVI A 1904-A的ARP响应可以由VRVI A 1904-A封装,并且可以作为ARP单播发送到发出请求的接口,并且具体而言是VI3。如本申请中早先所指示的,这个ARP响应的发送可以包括向VRVI A 1904-A的相关联的交换机提供封装的ARP响应,该交换机可以将封装的ARP响应发送到VI3。

[0345] ARP响应可以由VI3接收并且可以被解封装。VI3可以基于接收到的ARP响应学习VRVI A 1904-A的接口到MAC的映射,并且可以将更新后的学习到的映射提供给VI3的相关联的交换机。可以将解封装的分组提供给CI3。CI3可以基于接收到的分组学习IP-MAC地址

映射。CI3可以发送分组,该分组可以是IP分组,诸如IPv4或IPv6分组。这个分组可以具有CI3的MAC和IP地址作为源地址,并且可以具有目的地CI(CI4 1940)的IP地址作为目的地地址。在一些实施例中,目的地地址还可以包括VRVI A的1904-A MAC地址。

[0346] pause由CI3发送的分组可以由VI3接收,VI3可以封装该分组并将该分组转发到VRVI A 1904-A。VRVI A 1904-A可以接收分组、解封装分组,并查找用于分组的目的地IP地址(CI4的IP地址)的映射。在一些实施例中,VSRs以及因此VRVI A 1904-A可以驻留在L2网络和L3网络两者中。在L2网络中,因此在VLAN内,VSRs可以使用L2通信协议,而VSRs在与L3网络通信时可以使用L3通信协议。与在VLAN中执行的学习相反,VSRs可以从L3控制平面学习到L3网络中的端点的映射。在一些实施例中,例如,L3控制平面可以提供映射L3网络中的实例的IP地址、MAC地址和/或接口标识符的信息。

[0347] VRVI A 1904-A可以查找用于分组中包含的IP目的地地址的映射,或者换句话说,可以查找用于目的地CI的IP地址的映射。在一些实施例中,查找这个映射可以包括识别CI4 1940所在的子网1939和/或识别与CI4所在的子网1939相关联的VR。VRVI A 1904-A可以封装分组并将封装的分组转发到VR 1948,VR 1948可以是用于子网1939和/或目的地CI 1940的隧道端点(TEP)。这个封装的分组的转发由方框1956指示。

[0348] VR 1948可以接收并解封装分组。VR 1948还可以识别与目的地IP地址对应的子网1939内的接口。在一些实施例中,VR 1948可以位于与目的地接口VI4 1944相同的NVD上。因此,VR 1948可以将分组直接转发到目的地CI,并且具体而言是转发到CI4 1940,如方框1958中所指示的。

[0349] 基于接口的访问控制列表过滤

[0350] VSRs可以提供基于接口的访问控制列表(ACL)过滤。这可以包括评估用于VLAN的入口安全策略。这还可以包括在VSRs确定将接收到的分组发送到何处时基于学习到的接口到VLAN中MAC和IP地址的映射为VSRs处的发送方评估出口安全策略。这会导致ACL的延迟的分类。

[0351] 在一些实施例中,ACL可以包括与系统内的对象相关联的许可的列表。这些对象可以包括物理网络内的硬件,这些硬件可以包括例如一个或几个服务器、SmartNIC、主机机器等。这些对象可以包括虚拟网络内的一个或几个虚拟对象。这些虚拟对象可以包括例如一个或几个接口、计算实例、地址(诸如IP地址和/或MAC地址)等。在一些实施例中,ACL可以指定哪些用户、对象和/或系统进程被授予对对象的访问和/或允许对给定对象执行哪些操作。

[0352] ACL可以特定于一个或几个CI。因此,在一些实施例中,一些或所有CI可以具有唯一的ACL和/或可以维护唯一的ACL。在一些实施例中,CI ACL可以指定CI可以向哪些接口和/或地址(或者MAC或者IP)发送分组、CI不能向哪些接口和/或地址(或者MAC或者IP)发送分组、可以允许CI发送到一个或几个接口和/或地址的一种或几种类型的分组,和/或可以禁止CI发送到一个或几个接口的一种或几种类型的分组。在一些实施例中,CI的ACL可以存储在网络内其它实体(包括诸如可以强制执行ACL的一个或几个VR或VSRs之类的其它实体)可访问的位置。

[0353] 例如,当从IP网络为VLAN内的一个或几个预期接收方接收通信时,VSRs可以基于该通信的发送方的ACL来确定并应用过滤和/或限制通信的递送。在一些实施例中,这可以

例如通过以下操作来实现：(1) VSRS基于访问发送方的ACL的副本做出通信决定；或者(2) VSRS基于在该通信的分组元数据中编码的信息对VSRS接收到的通信做出通信决定。

[0354] 现在参考图20，示出了图示用于延迟的访问控制列表 (ACL) 分类的过程2000的一个实施例的流程图。过程2000可以由系统600的全部或部分执行，并且具体而言可以由VSRS 624、634执行。

[0355] 过程2000开始于方框2002，其中源CI将分组发送到目的地MAC或IP地址。在一些实施例中，源CI可以在包含分组被发送到的目的地MAC地址或IP地址的VLAN外部。

[0356] 在方框2006处，包含目的地MAC或IP地址的VLAN的VSRS接收分组。在一些实施例中，分组可以用L2封装来封装，并且在一些实施例中，分组可以用L3封装来封装。VSRS可以解封封装分组并识别源CI，如方框2008中所指示的。在识别出源CI后，VSRS可以访问用于源CI的ACL。在一些实施例中，例如，用于源CI的ACL可以存储在VSRS可访问的位置。在一些实施例中，访问用于源CI的ACL可以包括在源CI的ACL中检索信息，例如，该信息可以识别对分组递送的一个或几个限制。在一些实施例中，这个信息可以包括基于一个或几个IP地址、MAC地址、TCP和/或UDP源和目的地端口、EtherType等的一个或几个规则。

[0357] 在方框2010处，VSRS识别用于分组的目的地接口。在VSRS没有用于目的地地址的映射的实施例中，VSRS可以确定映射信息，如上文关于图16的过程1600的步骤1612至1620所讨论的那样。在先前学习到映射，或者经由执行过程1600的步骤1612至1620中的一些或全部学习到的实施例中，VSRS可以基于VSRS通过与VSRS的VLAN内的接口通信而学习到的映射来识别目的地接口。在一些实施例中，识别目的地接口可以包括基于目的地地址并且具体而言是分组的目的地IP地址和/或MAC地址查找目的地接口。

[0358] 在方框2012处，VSRS将源CI ACL应用于目的地接口。这可以包括确定源CI ACL的任何部分是否与目的地接口相关，如果是，那么应用源CI ACL的该部分。在方框2014处，如果目的地接口符合源CI ACL，或者换句话说，如果源CI ACL允许从源CI向目的地接口发送分组，那么VSRS将分组转发到目的地接口。在一些实施例中，分组的这种转发可以包括封装该分组。在一些实施例中，可以根据L2封装（诸如例如L2 Geneve封装）来封装分组。VSRS可以将分组转发到目的地接口，并且更具体而言是转发到将目的地接口作为其TEP的目的地CI。目的地接口可以接收分组，解封封装分组，并将分组转发到目的地CI。

[0359] 可替代地，在方框2016处，如果目的地接口不符合源CI ACL，或者换句话说，如果源CI ACL不允许从源CI向目的地接口发送分组，那么VSRS丢弃该分组。在一些实施例中，VSRS可以响应源CI指示丢弃分组和/或指示丢弃分组的原因。在一些实施例中，VSRS可以更新与发送接口相关联的操作度量/统计数据以反映ACL决定。

[0360] 现在参考图21，示出了图示用于ACL的早期分类和将该分类结合在元数据中的过程2100的一个实施例的流程图。过程2100可以由系统600的全部或部分执行，并且具体而言可以由源CI和VSRS 624、634执行。

[0361] 在方框2102处，源CI确定将分组发送到目的地CI。在一些实施例中，这可以包括源CI确定将分组发送到目的地CI的MAC地址和/或IP地址。如方框2104中所指示的，源VNIC可以发送分组。源VNIC可以接收分组，可以评估用于分组的ACL，并且可以将与分组相关的ACL信息嵌入在分组的一部分中。在一些实施例中，这个信息可以包括基于一个或几个IP地址、MAC地址、TCP和/或UDP源和目的地端口、EtherType等的一个或几个规则。源VNIC还可以封

装分组。在一些实施例中,ACL可以作为元数据存储封装的分组中,并且具体而言可以存储在分组的报头中。在一些实施例中,源CI可以在包含分组被发送到的目的地CI的VLAN外部。

[0362] pause在方框2106处,包含目的地MAC或IP地址的VLAN的VSRS接收分组。在一些实施例中,分组可以用L2封装来封装,并且在一些实施例中,分组可以用L3封装来封装。VSRS可以解封装分组。在一些实施例中,VSRS可以从分组中提取识别分组的目的地并且具体而言是识别目的地IP地址的信息。

[0363] 在方框2108处,VSRS识别用于分组的目的地接口。在VSRS没有用于目的地地址的映射的实施例中,VSRS可以确定映射信息,如上文关于图16的过程1600的步骤1612至1620所讨论的那样。在先前学习到映射,或者通过执行过程1600的步骤1612至1620中的一些或全部学习到的实施例中,VSRS可以基于由VSRS通过与VSRS的VLAN内的接口通信学习到的映射来识别目的地接口。因此,在一些实施例中,VSRS可以确定它具有用于目的地地址的映射信息,并且可以基于这个映射信息识别目的地接口。在一些实施例中,识别目的地接口可以包括基于目的地地址并且具体而言是分组的目的地IP地址和/或MAC地址来查找目的地接口。

[0364] 在方框2110处,VSRS访问包含在分组的一部分中的ACL信息。在一些实施例中,这可以包括从分组中提取元数据,并且具体而言是从分组报头中提取元数据。在一些实施例中,这可以包括解码在分组报头中的分组元数据中编码的信息。

[0365] 在方框2112处,VSRS应用从分组的该部分检索到的ACL信息。具体而言,这可以包括将编码在中的安全信息应用于目的地接口。这可以包括确定ACL信息的任何部分是否与目的地接口相关,如果是,那么应用ACL信息的该部分。在方框2114处,如果目的地接口符合ACL信息和/或符合ACL信息的一个或多个规则,或者换句话说,如果ACL信息允许从源CI向目的地接口发送分组,那么VSRS将分组转发到目的地接口。在一些实施例中,分组的这种转发可以包括封装该分组。在一些实施例中,可以根据L2封装(诸如例如L2Geneve封装)来封装分组。VSRS可以将分组转发到目的地接口,并且更具体而言是转发到将目的地接口作为其TEP的目的地CI。目的地接口可以接收分组、解封装分组,并将分组转发到目的地CI。

[0366] 可替代地,在方框2116处,如果目的地接口不符合ACL信息,或者换句话说,如果ACL信息不允许从源CI向目的地接口发送分组,那么VSRS丢弃该分组。在一些实施例中,VSRS可以响应源CI指示丢弃分组和/或指示丢弃分组的原因。

[0367] 下一跳路由

[0368] VSRS的一些实施例可以促进下一跳路由,并且具体而言可以延迟下一跳评估直到在VSRS处接收到通信。因为VLAN外部的通信的发送方可能不知道用于VLAN中实例的更新后的虚拟IP地址,所以VLAN外部的发送方无法准确指定下一跳路由。

[0369] 在一些实施例中,VSRS可以促进下一跳路由指定。在一些实施例中,这可以例如通过以下操作实现:(1)发送方为通信制定下一跳路由的初始指定,并且VSRS在接收到通信后重新评估该下一跳指定;或者(2)延迟下一跳指定,直到VSRS接收到通信。

[0370] 现在参考图22,示出了图示用于基于发送方的下一跳路由的过程2200的一个实施例的流程图。在一些实施例中,这可以包括下一跳路由的评估与下一跳路由的指定的分离。这会导致源CI评估路由策略并在通信的虚拟分组报头的分组元数据内编码下一跳路由。在

此类实施例中,分组可以包括下一跳目的地的预期虚拟IP地址,并且下一跳路由可以编码在分组元数据中。

[0371] 这个通信由VSRS接收,VSRS然后使用来自分组元数据的编码的下一跳路由来确定VLAN内接收通信的实例,以及该实例的目的地虚拟IP地址。实例的这种确定可以使用由VSRS生成和/或策划的表,并且具体而言可以使用链接虚拟IP地址、MAC地址和/或虚拟接口ID的表。使用这些表,VSRS可以识别与预期下一跳目的地的MAC地址和/或虚拟接口ID对应的虚拟IP地址。在一些实施例中,VSRS基于分组元数据中包含的编码的下一跳路由对接收方实例的这种识别可以导致VSRS将通信发送到作为由分组中的虚拟IP地址指示并由分组的发送方指定的下一跳目的地的不同目的地虚拟IP地址。

[0372] 过程2200可以由系统600的全部或部分执行,并且具体而言可以由源CI和VSRS 624、634执行。

[0373] 该过程开始于方框2202,其中源CI可以发送分组。在一些实施例中,源CI可以在包含分组被发送到的目的地MAC地址或IP地址的VLAN外部。

[0374] 在方框2204处,源VNIC可以基于一个或几个路由规则做出路由决定。在一些实施例中,这可以包括源VNIC检索和/或访问一个或几个路由规则,然后基于那些路由规则做出路由决定。在一些实施例中,这个路由决定可以是用于分组的下一跳路由决定。在一些实施例中,这些路由规则可以存储在例如源CI的网络中的路由表中。在一些实施例中,这可以包括例如子网路由表。

[0375] 在方框2206处,源VNIC可以将路由决定嵌入在分组的一部分中。在一些实施例中,这可以包括封装分组,并将路由决定嵌入在分组的元数据中,该元数据可以例如被编码在分组的报头中。在封装分组并将路由决定嵌入在封装的分组中后,源VNIC可以将封装的分组转发到VSRS。

[0376] 在方框2208处,包含目的地MAC或IP地址的VLAN的VSRS接收分组。在一些实施例中,分组可以用L2封装来封装,并且在一些实施例中,分组可以用L3封装来封装。在一些实施例中,由VSRS接收分组可以包括对分组的解封装。在一些实施例中,由VSRS接收分组可以包括提取嵌入在分组的一部分中的路由决定。在一些实施例中,这可以包括对包含路由决定的编码的元数据进行解码。

[0377] 在方框2210处,VSRS将路由决定应用于VSRS路由信息以确定分组的目的地接口。在一些实施例中,这可以包括将解码的路由决定应用于VSRS路由信息以确定VLAN中与路由信息对应的目的地CI。在方框2212处,VSRS将分组发送到VLAN中确定的CI。在一些实施例中,这可以包括VSRS将分组转发到目的地接口。在一些实施例中,分组的这种转发可以包括封装分组,该封装可以根据L2封装(诸如例如L2 Geneve封装)。VSRS可以将分组转发到目的地接口,更具体而言是转发到将目的地接口作为其TEP的目的地CI。目的地接口可以接收分组、解封装分组,并将分组转发到目的地CI。

[0378] 现在参考图23,示出了图示用于延迟的下一跳路由的过程2300的一个实施例的流程图。过程2300可以由系统600的全部或部分执行,并且具体而言可以由源VNIC和VSRS 624、634执行。

[0379] 在一些实施例中,例如,源VNIC可以基于可以包含在路由表中的路由规则来为通过VLAN的通信做出路由决定。这种通信可以由VSRS接收,该VSRS可以重新评估下一跳指定,

并且可以查询同一路由表的副本以识别路由规则。基于这个路由规则,VSRS确定VLAN内与该路由规则对应的实例、该实例的虚拟IP地址,并将通信发送到该实例。确定与路由规则对应的实例可以包括由VSRS从链接虚拟IP地址、MAC地址和/或虚拟接口ID的表中检索信息,以及确定与作为预期的下一跳目的地的虚拟接口和/或实例相关联的虚拟IP地址。

[0380] 过程2300开始于方框2302,其中源CI可以发送分组。在一些实施例中,这可以包括源CI将分组发送到目的地CI的MAC地址和/或IP地址。在一些实施例中,源CI可以在包含分组被发送到的目的地MAC地址或IP地址的VLAN外部。

[0381] 在方框2204处,源VNIC可以接收分组。然后,源VNIC可以基于一个或几个路由规则做出路由决定。在一些实施例中,这可以包括源VNIC检索和/或访问一个或几个路由规则,然后基于那些路由规则做出路由决定。在一些实施例中,这个路由决定可以是用于分组的下一跳路由决定。在一些实施例中,这些路由规则可以存储在例如源CI的网络中的路由表中。在一些实施例中,这可以包括例如子网路由表。然后源VNIC可以封装分组,并且可以将分组发送到VSRS。

[0382] 在方框2206处,包含目的地MAC或IP地址的VLAN的VSRS接收分组。在一些实施例中,分组可以用L2封装来封装,并且在一些实施例中,分组可以用L3封装来封装。在一些实施例中,由VSRS接收分组可以包括对分组的解封装。

[0383] 在方框2308处,VSRS检索与分组相关的路由信息。在一些实施例中,这可以包括检索与接收到的分组相关的路由表和/或路由表的部分。在一些实施例中,可以从例如控制平面接收这个路由表。在接收到路由信息后,VSRS识别与接收到的分组相关的路由规则。

[0384] 在方框2310处,VSRS将路由规则应用于VSRS路由信息以确定VLAN中用于分组的目的地接口。在一些实施例中,这可以包括将路由规则应用于VSRS路由信息以确定VLAN中与路由信息对应的目的地CI。在方框2312处,VSRS将分组发送到VLAN中确定的CI。在一些实施例中,这可以包括VSRS将分组转发到目的地接口。在一些实施例中,分组的这种转发可以包括封装分组,该封装可以根据L2封装(诸如例如L2Geneve封装)。VSRS可以将分组转发到目的地接口,并且更具体而言是转发到将目的地接口作为其TEP的目的地CI。目的地接口可以接收分组、解封装分组,并将分组转发到目的地CI。

[0385] 示例实施方式

[0386] 如以上所指出的,基础设施即服务(IaaS)是一种特定类型的云计算。IaaS可以被配置为通过公共网络(例如,互联网)提供虚拟化计算资源。在IaaS模型中,云计算提供商可以托管基础设施组件(例如,服务器、存储设备、网络节点(例如,硬件)、部署软件、平台虚拟化(例如,管理程序层)等)。在一些情况下,IaaS提供商还可以提供各种服务来伴随这些基础设施组件(例如,计费、监视、记载、负载均衡和聚类等)。因此,由于这些服务可能是策略驱动的,因此IaaS用户可以能够实现策略来驱动负载均衡,以维持应用的可用性和性能。

[0387] 在一些情况下,IaaS客户可以通过诸如互联网之类的广域网(WAN)访问资源和服务,并且可以使用云提供商的服务来安装应用栈的剩余元素。例如,用户可以登录到IaaS平台以创建虚拟机(VM)、在每个VM上安装操作系统(OS)、部署诸如数据库之类的中间件、为工作负载和备份创建存储桶,甚至将企业软件安装到那个VM中。然后,客户可以使用提供商的服务来执行各种功能,包括平衡网络流量、解决应用问题、监视性能、管理灾难恢复等。

[0388] 在大多数情况下,云计算模型将需要云提供商的参与。云提供商可以但不一定是

专门提供(例如,供应、出租、销售)IaaS的第三方服务。实体也可能选择部署私有云,从而成为其自己的基础设施服务提供商。

[0389] 在一些示例中,IaaS部署是将新应用或应用的新版本放置到准备好的应用服务器等上的处理。它还可以包括准备服务器(例如,安装库、守护进程等)的处理。这通常由云提供商管理,位于管理程序层之下(例如,服务器、存储装置、网络硬件和虚拟化)。因此,客户可以负责处理(OS)、中间件和/或应用部署(例如,在(例如,可以按需启动的)自助服务虚拟机等上)。

[0390] 在一些示例中,IaaS供给可以指获取计算机或虚拟主机以供使用,甚至在它们上安装所需的库或服务。大多数情况下,部署不包括供给,并且供给可能需要被首先执行。

[0391] 在一些情况下,IaaS供给存在两个不同的挑战。首先,在任何东西运行之前供给初始基础设施集存在最初的挑战。其次,一旦所有东西已被供给,就存在演进现有基础设施(例如,添加新服务、更改服务、移除服务等)的挑战。在一些情况下,可以通过启用以声明方式定义基础设施的配置来解决这两个挑战。换句话说,基础设施(例如,需要哪些组件以及它们如何交互)可以由一个或多个配置文件来定义。因此,基础设施的总体拓扑(例如,哪些资源依赖于哪些资源,以及它们如何协同工作)可以以声明的方式描述。在一些情况下,一旦定义了拓扑,就可以生成创建和/或管理配置文件中描述的不同组件的工作流。

[0392] 在一些示例中,基础设施可以具有许多互连的元件。例如,可能存在一个或多个虚拟私有云(VPC)(例如,可配置和/或共享计算资源的潜在按需池),也称为核心网络。在一些示例中,还可以供给一个或多个入站/出站流量组规则以定义如何设置网络的入站/出站流量以及一个或多个虚拟机(VM)。也可以供给其它基础设施元素,诸如负载均衡器、数据库等。随着期望和/或添加越来越多的基础设施元素,基础设施可以逐步演进。

[0393] 在一些情况下,可以采用连续部署技术来使得能够跨各种虚拟计算环境部署基础设施代码。此外,所描述的技术可以使得能够在这些环境内进行基础设施管理。在一些示例中,服务团队可以编写期望部署到一个或多个但通常是许多不同的生产环境(例如,跨各种不同的地理位置,有时跨越整个世界)的代码。但是,在一些示例中,必须首先设置将在其上部署代码的基础设施。在一些情况下,供给可以手动完成,可以利用供给工具供给资源,和/或一旦供给基础设施就可以利用部署工具部署代码。

[0394] 图24是图示根据至少一个实施例的IaaS体系架构的示例模式的框图2400。服务运营商2402可以通信地耦合到可以包括虚拟云网络(VCN)2406和安全主机子网2408的安全主机租赁2404。在一些示例中,服务运营商2402可以使用一个或多个客户端计算设备,其可以是便携式手持设备(例如,**iPhone®**、蜂窝电话、**iPad®**、计算平板、个人数字助理(PDA))或可穿戴设备(例如,Google **Glass®**头戴式显示器)、运行软件(诸如Microsoft Windows **Mobile®**)和/或各种移动操作系统(诸如iOS、Windows Phone、Android、BlackBerry 8、Palm OS等),并且支持互联网、电子邮件、短消息服务(SMS)、**Blackberry®**或其它通信协议。替代地,客户端计算设备可以是通用个人计算机,包括例如运行各种版本的Microsoft **Windows®**、Apple **Macintosh®**和/或Linux操作系统的个人计算机和/或膝上型计算机。客户端计算设备可以是运行各种商业上可获得的**UNIX®**或类UNIX操作系统,包括但不限于各种GNU/Linux操作系统(诸如例如Google

Chrome OS)中的任何一种的工作站计算机。替代地或附加地,客户端计算设备可以是任何其它电子设备,诸如瘦客户端计算机、支持互联网的游戏系统(例如,具有或不具有**Kinect®**手势输入设备的Microsoft Xbox游戏控制台),和/或能够通过可以访问VCN 2406和/或互联网的网络进行通信的个人消息传递设备。

[0395] VCN 2406可以包括本地对等网关(LPG) 2410,其可以经由包含在SSH VCN 2412中的LPG 2410通信地耦合到安全壳(SSH) VCN 2412。SSH VCN 2412可以包括SSH子网2414,并且SSH VCN 2412可以经由包含在控制平面VCN 2416中的LPG 2410通信地耦合到控制平面VCN 2416。此外,SSH VCN 2412可以经由LPG 2410通信地耦合到数据平面VCN 2418。控制平面VCN 2416和数据平面VCN 2418可以包含在可以由IaaS提供商拥有和/或操作的服务租赁2419中。

[0396] 控制平面VCN 2416可以包括充当外围网络(例如,公司内部网和外部网络之间的公司网络的部分)的控制平面非军事区(DMZ)层2420。基于DMZ的服务器可以承担有限责任并有助于控制漏洞。此外,DMZ层2420可以包括一个或多个负载平衡器(LB)子网2422、可以包括(一个或多个)应用子网2426的控制平面应用层2424、可以包括(一个或多个)数据库(DB)子网2430(例如,(一个或多个)前端DB子网和/或(一个或多个)后端DB子网)的控制平面数据层2428。包含在控制平面DMZ层2420中的(一个或多个)LB子网2422可以通信地耦合到包含在控制平面应用层2424中的(一个或多个)应用子网2426和可以包含在控制平面VCN 2416中的互联网网关2434,并且(一个或多个)应用子网2426可以通信地耦合到包含在控制平面数据层2428中的(一个或多个)DB子网2430以及服务网关2436和网络地址转换(NAT)网关2438。控制平面VCN 2416可以包括服务网关2436和NAT网关2438。

[0397] 控制平面VCN 2416可以包括数据平面镜像应用层2440,其可以包括(一个或多个)应用子网2426。包含在数据平面镜像应用层2440中的(一个或多个)应用子网2426可以包括可以执行计算实例2444的虚拟网络接口控制器(VNIC) 2442。计算实例2444可以将数据平面镜像应用层2440的(一个或多个)应用子网2426通信地耦合到可以包含在数据平面应用层2446中的(一个或多个)应用子网2426。

[0398] 数据平面VCN 2418可以包括数据平面应用层2446、数据平面DMZ层2448和数据平面数据层2450。数据平面DMZ层2448可以包括(一个或多个)LB子网2422,其可以通信地耦合到数据平面应用层2446的(一个或多个)应用子网2426和数据平面VCN 2418的互联网网关2434。(一个或多个)应用子网2426可以通信地耦合到数据平面VCN 2418的服务网关2436和数据平面VCN 2418的NAT网关2438。数据平面数据层2450还可以包括可以通信地耦合到数据平面应用层2446的(一个或多个)应用子网2426的(一个或多个)DB子网2430。

[0399] 控制平面VCN 2416和数据平面VCN 2418的互联网网关2434可以通信地耦合到元数据管理服务2452,元数据管理服务2452可以通信地耦合到公共互联网2454。公共互联网2454可以通信地耦合到控制平面VCN 2416和数据平面VCN 2418的NAT网关2438。控制平面VCN 2416和数据平面VCN 2418的服务网关2436可以通信地耦合到云服务2456。

[0400] 在一些示例中,控制平面VCN 2416或数据平面VCN 2418的服务网关2436可以对云服务2456进行应用编程接口(API)调用,而无需通过公共互联网2454。从服务网关2436到云服务2456的API调用可以是单向的:服务网关2436可以对云服务2456进行API调用,并且云服务2456可以将请求的数据发送到服务网关2436。但是,云服务2456可以不发起对服务网

关2436的API调用。

[0401] 在一些示例中,安全主机租赁2404可以直接连接到服务租赁2419,服务租赁2419否则可以被隔离。安全主机子网2408可以通过LPG 2410与SSH子网2414通信,LPG 2410可以使得能够在否则隔离的系统上进行双向通信。将安全主机子网2408连接到SSH子网2414可以使安全主机子网2408访问服务租赁2419内的其它实体。

[0402] 控制平面VCN 2416可以允许服务租赁2419的用户设置或以其它方式供给期望的资源。在控制平面VCN 2416中供给的期望资源可以在数据平面VCN 2418中部署或以其它方式使用。在一些示例中,控制平面VCN 2416可以与数据平面VCN 2418隔离,并且控制平面VCN 2416的数据平面镜像应用层2440可以经由VNIC 2442与数据平面VCN 2418的数据平面应用层2446通信,VNIC 2442可以包含在数据平面镜像应用层2440和数据平面应用层2446中。

[0403] 在一些示例中,系统的用户或客户可以通过可以将请求传送到元数据管理服务2452的公共互联网2454来做出请求,例如创建、读取、更新或删除 (CRUD) 操作。元数据管理服务2452可以通过互联网网关2434将请求传送到控制平面VCN 2416。请求可以被包含在控制平面DMZ层2420中的(一个或多个)LB子网2422接收。(一个或多个)LB子网2422可以确定请求是有效的,并且响应于该确定,(一个或多个)LB子网2422可以将请求传输到包含在控制平面应用层2424中的(一个或多个)应用子网2426。如果请求被验证并且需要对公共互联网2454的调用,那么对公共互联网2454的调用可以被传输到可以对公共互联网2454进行调用的NAT网关2438。请求可能期望存储的存储器可以存储在(一个或多个)DB子网2430中。

[0404] 在一些示例中,数据平面镜像应用层2440可以促进控制平面VCN 2416和数据平面VCN 2418之间的直接通信。例如,可能期望对包含在数据平面VCN 2418中的资源应用对配置的更改、更新或其它适当的修改。经由VNIC 2442,控制平面VCN 2416可以直接与包含在数据平面VCN 2418中的资源通信,并且从而可以执行对配置的更改、更新或其它适当的修改。

[0405] 在一些实施例中,控制平面VCN 2416和数据平面VCN 2418可以包含在服务租赁2419中。在这种情况下,系统的用户或客户可能不拥有或操作控制平面VCN 2416或数据平面VCN 2418。替代地,IaaS提供商可以拥有或操作控制平面VCN 2416和数据平面VCN 2418,这两者平面都可以包含在服务租赁2419中。该实施例可以使得能够隔离可能阻止用户或客户与其它用户或其它客户的资源交互的网络。此外,该实施例可以允许系统的用户或客户私自存储数据库,而无需依赖可能不具有期望威胁预防级别的公共互联网2454进行存储。

[0406] 在其它实施例中,包含在控制平面VCN 2416中的(一个或多个)LB子网2422可以被配置为从服务网关2436接收信号。在这个实施例中,控制平面VCN 2416和数据平面VCN 2418可以被配置为由IaaS提供商的客户调用而无需调用公共互联网2454。IaaS提供商的客户可能期望这个实施例,因为客户使用的(一个或多个)数据库可以由IaaS提供商控制并且可以存储在服务租赁2419上,服务租赁2419可能与公共互联网2454隔离。

[0407] 图25是图示根据至少一个实施例的IaaS体系架构的另一个示例模式的框图2500。服务运营商2502(例如,图24的服务运营商2402)可以通信地耦合到安全主机租赁2504(例如,图24的安全主机租赁2404),该安全主机租赁2504可以包括虚拟云网络 (VCN) 2506(例如,图24的VCN 2406) 和安全主机子网2508(例如,图24的安全主机子网2408)。VCN 2506可

以包括本地对等网关(LPG)2510(例如,图24的LPG 2410),其可以经由包含在SSH VCN 2512中的LPG 2410通信地耦合到安全壳(SSH)VCN 2512(例如,图24的SSH VCN 2412)。SSH VCN 2512可以包括SSH子网2514(例如,图24的SSH子网2414),并且SSH VCN 2512可以经由包含在控制平面VCN 2516中的LPG 2510通信地耦合到控制平面VCN 2516(例如,图24的控制平面VCN 2416)。控制平面VCN 2516可以包含在服务租赁2519(例如,图24的服务租赁2419)中,并且数据平面VCN 2518(例如,图24的数据平面VCN 2418)可以包含在可能由系统的用户或客户拥有或操作的客户租赁2521中。

[0408] 控制平面VCN 2516可以包括控制平面DMZ层2520(例如,图24的控制平面DMZ层2420),其可以包括(一个或多个)LB子网2522(例如,图24的(一个或多个)LB子网2422)、可以包括(一个或多个)应用子网2526(例如,图24的(一个或多个)应用子网2426)的控制平面应用层2524(例如,图24的控制平面应用层2424)、可以包括(一个或多个)数据库(DB)子网2530(例如,类似于图24的(一个或多个)DB子网2430)的控制平面数据层2528(例如,图24的控制平面数据层2428)。包含在控制平面DMZ层2520中的(一个或多个)LB子网2522可以通信地耦合到包含在控制平面应用层2524中的(一个或多个)应用子网2526和可以包含在控制平面VCN 2516中的互联网网关2534(例如,图24的互联网网关2434),并且(一个或多个)应用子网2526可以通信地耦合到包含在控制平面数据层2528中的(一个或多个)DB子网2530以及服务网关2536(例如,图24的服务网关)和网络地址转换(NAT)网关2538(例如,图24的NAT网关2438)。控制平面VCN 2516可以包括服务网关2536和NAT网关2538。

[0409] 控制平面VCN 2516可以包括可以包括(一个或多个)应用子网2526的数据平面镜像应用层2540(例如,图24的数据平面镜像应用层2440)。包含在数据平面镜像应用层2540中的(一个或多个)应用子网2526可以包括可以执行计算实例2544(例如,类似于图24的计算实例2444)的虚拟网络接口控制器(VNIC)2542(例如,2442的VNIC)。计算实例2544可以促进数据平面镜像应用层2540的(一个或多个)应用子网2526和可以包含在数据平面应用层2546(例如,图24的数据平面应用层2446)中的(一个或多个)应用子网2526之间经由包含在数据平面镜像应用层2540中的VNIC 2542和包含在数据平面应用层2546中的VNIC 2542的通信。

[0410] 包含在控制平面VCN 2516中的互联网网关2534可以通信地耦合到元数据管理服务2552(例如,图24的元数据管理服务2452),该元数据管理服务2552可以通信地耦合到公共互联网2554(例如,图24的公共互联网2454)。公共互联网2554可以通信地耦合到包含在控制平面VCN 2516中的NAT网关2538。包含在控制平面VCN 2516中的服务网关2536可以通信地耦合到云服务2556(例如,图24的云服务2456)。

[0411] 在一些示例中,数据平面VCN 2518可以包含在客户租赁2521中。在这种情况下,IaaS提供商可以为每个客户提供控制平面VCN 2516,并且IaaS提供商可以为每个客户设置包含在服务租赁2519中的唯一计算实例2544。每个计算实例2544可以允许包含在服务租赁2519中的控制平面VCN 2516和包含在客户租赁2521中的数据平面VCN 2518之间的通信。计算实例2544可以允许在包含在服务租赁2519中的控制平面VCN 2516中供给的资源被部署或以其它方式在包含在客户租赁2521中的数据平面VCN 2518中使用。

[0412] 在其它示例中,IaaS提供商的客户可以具有存在于客户租赁2521中的数据库。在这个示例中,控制平面VCN 2516可以包括数据平面镜像应用层2540,其可以包括(一个或多

个)应用子网2526。数据平面镜像应用层2540可以驻留在数据平面VCN 2518中,但数据平面镜像应用层2540可能不在数据平面VCN 2518中。即,数据平面镜像应用层2540可以访问客户租赁2521,但是数据平面镜像应用层2540可能不存在于数据平面VCN 2518中或者由IaaS提供商的客户拥有或操作。数据平面镜像应用层2540可以被配置为对数据平面VCN 2518进行调用,但可以被配置为对包含在控制平面VCN 2516中的任何实体进行调用。客户可能期望在数据平面VCN 2518中部署或以其它方式使用在控制平面VCN 2516中供给的资源,并且数据平面镜像应用层2540可以促进客户的期望部署或资源的其它使用。

[0413] 在一些实施例中,IaaS提供商的客户可以将过滤器应用到数据平面VCN 2518。在这个实施例中,客户可以确定数据平面VCN 2518可以访问什么,并且客户可以限制从数据平面VCN 2518对公共互联网2554的访问。IaaS提供商可能无法应用过滤器或以其它方式控制数据平面VCN 2518对任何外部网络或数据库的访问。客户将过滤器和控制应用到包含在客户租赁2521中的数据平面VCN 2518上可以帮助将数据平面VCN 2518与其它客户和公共互联网2554隔离开。

[0414] 在一些实施例中,云服务2556可以由服务网关2536调用以访问公共互联网2554、控制平面VCN 2516或数据平面VCN 2518上可能不存在的服务。云服务2556与控制平面VCN 2516或数据平面VCN 2518之间的连接可以不是实时的或连续的。云服务2556可以存在于由IaaS提供商拥有或操作的不同网络上。云服务2556可以被配置为接收来自服务网关2536的调用并且可以被配置为不接收来自公共互联网2554的调用。一些云服务2556可以与其它云服务2556隔离,并且控制平面VCN 2516可以与可能与控制平面VCN 2516不在同一区域的云服务2556隔离。例如,控制平面VCN 2516可能位于“区域1”,并且云服务“部署24”可能位于区域1和“区域2”。如果包含在位于区域1中的控制平面VCN 2516中的服务网关2536对部署24进行调用,那么该调用可以被传输到区域1中的部署24。在这个示例中,控制平面VCN 2516或区域1中的部署24可能不与区域2中的部署24通信地耦合或以其它方式通信。

[0415] 图26是图示根据至少一个实施例的IaaS体系架构的另一个示例模式的框图2600。服务运营商2602(例如,图24的服务运营商2402)可以通信地耦合到安全主机租赁2604(例如,图24的安全主机租赁2404),该安全主机租赁2604可以包括虚拟云网络(VCN) 2606(例如,图24的VCN 2406)和安全主机子网2608(例如,图24的安全主机子网2408)。VCN 2606可以包括LPG 2610(例如,图24的LPG 2410),其可以经由包含在SSH VCN 2612中的LPG 2610通信地耦合到SSH VCN 2612(例如,图24的SSH VCN 2412)。SSH VCN 2612可以包括SSH子网2614(例如,图24的SSH子网2414),并且SSH VCN 2612可以经由包含在控制平面VCN 2616中的LPG 2610通信地耦合到控制平面VCN 2616(例如,图24的控制平面VCN 2416)并且经由包含在数据平面VCN 2618中的LPG 2610耦合到数据平面VCN 2618(例如,图24的数据平面2418)。控制平面VCN 2616和数据平面VCN 2618可以包含在服务租赁2619(例如,图24的服务租赁2419)中。

[0416] 控制平面VCN 2616可以包括能包括(一个或多个)负载平衡器(LB)子网2622(例如,图24的(一个或多个)LB子网2422)的控制平面DMZ层2620(例如,图24的控制平面DMZ层2420)、可以包括(一个或多个)应用子网2626(例如,类似于图24的(一个或多个)应用子网2426)的控制平面应用层2624(例如,图24的控制平面应用层2424)、可以包括(一个或多个)DB子网2630的控制平面数据层2628(例如,图24的控制平面数据层2428)。包含在控制平面

DMZ层2620中的(一个或多个)LB子网2622可以通信地耦合到包含在控制平面应用层2624中的(一个或多个)应用子网2626和可以包含在控制平面VCN 2616中的互联网网关2634(例如,图24的互联网网关2434),并且(一个或多个)应用子网2626可以通信地耦合到包含在控制平面数据层2628中的(一个或多个)DB子网2630以及服务网关2636(例如,图24的服务网关)和网络地址转换(NAT)网关2638(例如,图24的NAT网关2438)。控制平面VCN 2616可以包括服务网关2636和NAT网关2638。

[0417] 数据平面VCN 2618可以包括数据平面应用层2646(例如,图24的数据平面应用层2446)、数据平面DMZ层2648(例如,图24的数据平面DMZ层2448),以及数据平面数据层2650(例如,图24的数据平面数据层2450)。数据平面DMZ层2648可以包括可以通信地耦合到数据平面应用层2646的(一个或多个)可信应用子网2660和(一个或多个)不可信应用子网2662以及包含在数据平面VCN 2618中的互联网网关2634的(一个或多个)LB子网2622。(一个或多个)可信应用子网2660可以通信地耦合到包含在数据平面VCN 2618中的服务网关2636、包含在数据平面VCN 2618中的NAT网关2638以及包含在数据平面数据层2650中的(一个或多个)DB子网2630。(一个或多个)不可信应用子网2662可以通信地耦合到包含在数据平面VCN 2618中的服务网关2636和包含在数据平面数据层2650中的(一个或多个)DB子网2630。数据平面数据层2650可以包括可以通信地耦合到包含在数据平面VCN 2618中的服务网关2636的(一个或多个)DB子网2630。

[0418] (一个或多个)不可信应用子网2662可以包括可以通信地耦合到租户虚拟机(VM) 2666(1) - (N)的一个或多个主VNIC 2664(1) - (N)。每个租户VM 2666(1) - (N)可以通信地耦合到可以包含在相应容器出口VCN 2668(1) - (N)中的相应应用子网2667(1) - (N),相应容器出口VCN 2668(1) - (N)可以包含在相应客户租赁2670(1) - (N)中。相应的辅助VNIC 2672(1) - (N)可以促进数据平面VCN 2618中包含的(一个或多个)不可信应用子网2662与容器出口VCN 2668(1) - (N)中包含的应用子网之间的通信。每个容器出口VCN 2668(1) - (N)可以包括NAT网关2638,该NAT网关2638可以通信地耦合到公共互联网2654(例如,图24的公共互联网2454)。

[0419] 包含在控制平面VCN 2616中并且包含在数据平面VCN 2618中的互联网网关2634可以通信地耦合到元数据管理服务2652(例如,图24的元数据管理系统2452),该元数据管理服务2652可以通信地耦合到公共互联网2654。公共互联网2654可以通信地耦合到包含在控制平面VCN 2616中并且包含在数据平面VCN 2618中的NAT网关2638。包含在控制平面VCN 2616中和包含在数据平面VCN 2618中的服务网关2636可以通信地耦合到云服务2656。

[0420] 在一些实施例中,数据平面VCN 2618可以与客户租赁2670集成。在一些情况下,诸如在执行代码时可能期望支持的情况下,这种集成对于IaaS提供商的客户可能是有用的或期望的。客户可能提供可能具有破坏性、可能与其它客户资源通信或可能以其它方式导致非期望效果的代码来运行。作为对此的响应,IaaS提供商可以确定是否运行由客户给与IaaS提供商的代码。

[0421] 在一些示例中,IaaS提供商的客户可以向IaaS提供商授予临时网络访问,并请求附加到数据平面应用层2646的功能。运行该功能的代码可以在VM 2666(1) - (N)中执行,并且该代码可以不被配置为在数据平面VCN 2618上的其它任何地方运行。每个VM 2666(1) - (N)可以连接到一个客户租赁2670。包含在VM 2666(1) - (N)中的相应容器2671(1) - (N)可以

被配置为运行代码。在这种情况下,可以存在双重隔离(例如,容器2671(1)-(N)运行代码,其中容器2671(1)-(N)可能至少包含在(一个或多个)不可信应用子网2662中包含的VM 2666(1)-(N)中),这可以帮助防止不正确的或以其它方式非期望的代码损坏IaaS提供商的网络或损坏不同客户的网络。容器2671(1)-(N)可以通信地耦合到客户租赁2670并且可以被配置为传输或接收来自客户租赁2670的数据。容器2671(1)-(N)可以不被配置为从数据平面VCN 2618中的任何其它实体传输或接收数据。在运行代码完成后,IaaS提供商可以终止或以其它方式处置容器2671(1)-(N)。

[0422] 在一些实施例中,(一个或多个)可信应用子网2660可以运行可以由IaaS提供商拥有或操作的代码。在这个实施例中,(一个或多个)可信应用子网2660可以通信地耦合到(一个或多个)DB子网2630并且被配置为在(一个或多个)DB子网2630中执行CRUD操作。(一个或多个)不可信应用子网2662可以通信地耦合到(一个或多个)DB子网2630,但是在这个实施例中,(一个或多个)不可信应用子网可以被配置为在(一个或多个)DB子网2630中执行读取操作。可以包含在每个客户的VM 2666(1)-(N)中并且可以运行来自客户的代码的容器2671(1)-(N)可以不与(一个或多个)DB子网2630通信地耦合。

[0423] 在其它实施例中,控制平面VCN 2616和数据平面VCN 2618可以不直接通信地耦合。在这个实施例中,控制平面VCN 2616和数据平面VCN 2618之间可能没有直接通信。但是,通信可以通过至少一种方法间接发生。LPG 2610可以由IaaS提供商建立,其可以促进控制平面VCN 2616和数据平面VCN 2618之间的通信。在另一个示例中,控制平面VCN 2616或数据平面VCN 2618可以经由服务网关2636调用云服务2656。例如,从控制平面VCN 2616对云服务2656的调用可以包括对可以与数据平面VCN 2618通信的服务的请求。

[0424] 图27是图示根据至少一个实施例的IaaS体系架构的另一个示例模式的框图2700。服务运营商2702(例如,图24的服务运营商2402)可以通信地耦合到安全主机租赁2704(例如,图24的安全主机租赁2404),该安全主机租赁2704可以包括虚拟云网络(VCN) 2706(例如,图24的VCN 2406)和安全主机子网2708(例如,图24的安全主机子网2408)。VCN 2706可以包括LPG 2710(例如,图24的LPG 2410),该LPG 2710可以经由包含在SSH VCN 2712(例如,图24的SSH VCN 2412)中的LPG 2710通信地耦合到SSH VCN 2712。SSH VCN 2712可以包括SSH子网2714(例如,图24的SSH子网2414),并且SSH VCN 2712可以经由包含在控制平面VCN 2716中的LPG 2710通信地耦合到控制平面VCN 2716(例如,图24的控制平面VCN 2416)并且经由包含在数据平面VCN 2718中的LPG 2710耦合到数据平面VCN 2718(例如,图24的数据平面2418)。控制平面VCN 2716和数据平面VCN 2718可以包含在服务租赁2719(例如,图24的服务租赁2419)中。

[0425] 控制平面VCN 2716可以包括能包括(一个或多个)LB子网2722(例如,图24的(一个或多个)LB子网2422)的控制平面DMZ层2720(例如,图24的控制平面DMZ层2420)、可以包括(一个或多个)应用子网2726(例如,图24的(一个或多个)应用子网2426)的控制平面应用层2724(例如,图24的控制平面应用层2424)、可以包括(一个或多个)DB子网2730(例如,图26的(一个或多个)DB子网2630)的控制平面数据层2728(例如,图24的控制平面数据层2428)。包含在控制平面DMZ层2720中的(一个或多个)LB子网2722可以通信地耦合到包含在控制平面应用层2724中的(一个或多个)应用子网2726和可以包含在控制平面VCN 2716中的互联网网关2734(例如,图24的互联网网关2434),并且(一个或多个)应用子网2726可以通信地

耦合到包含在控制平面数据层2728中的(一个或多个)DB子网2730以及服务网关2736(例如,图24的服务网关)和网络地址转换(NAT)网关2738(例如,图24的NAT网关2438)。控制平面VCN 2716可以包括服务网关2736和NAT网关2738。

[0426] 数据平面VCN 2718可以包括数据平面应用层2746(例如,图24的数据平面应用层2446)、数据平面DMZ层2748(例如,图24的数据平面DMZ层2448),以及数据平面数据层2750(例如,图24的数据平面数据层2450)。数据平面DMZ层2748可以包括可以通信地耦合到数据平面应用层2746的(一个或多个)可信应用子网2760(例如,图26的(一个或多个)可信应用子网2660)和(一个或多个)不可信应用子网2762(例如,图26的(一个或多个)不可信应用子网2662)以及包含在数据平面VCN 2718中的互联网网关2734的(一个或多个)LB子网2722。(一个或多个)可信应用子网2760可以通信地耦合到包含在数据平面VCN 2718中的服务网关2736、包含在数据平面VCN 2718中的NAT网关2738以及包含在数据平面数据层2750中的(一个或多个)DB子网2730。(一个或多个)不可信应用子网2762可以通信地耦合到包含在数据平面VCN 2718中的服务网关2736和包含在数据平面数据层2750中的(一个或多个)DB子网2730。数据平面数据层2750可以包括可以通信地耦合到包含在数据平面VCN 2718中的服务网关2736的(一个或多个)DB子网2730。

[0427] (一个或多个)不可信应用子网2762可以包括可以通信地耦合到驻留在(一个或多个)不可信应用子网2762内的租户虚拟机(VM)2766(1)-(N)的主VNIC 2764(1)-(N)。每个租户VM 2766(1)-(N)可以在相应的容器2767(1)-(N)中运行代码,并且可通信地耦合到可以包含在容器出口VCN 2768中包含的数据平面应用层2746中的应用子网2726。相应的辅助VNIC 2772(1)-(N)可以促进包含在数据平面VCN 2718中的(一个或多个)不可信应用子网2762和包含在容器出口VCN 2768中的应用子网之间的通信。容器出口VCN可以包括可以通信地耦合到公共互联网2754(例如,图24的公共互联网2454)的NAT网关2738。

[0428] 包含在控制平面VCN 2716中和包含在数据平面VCN 2718中的互联网网关2734可以通信地耦合到元数据管理服务2752(例如,图24的元数据管理系统2452),该元数据管理服务2752可以通信地耦合到公共互联网2754。公共互联网2754可以通信地耦合到包含在控制平面VCN 2716中并且包含在数据平面VCN 2718中的NAT网关2738。包含在控制平面VCN 2716中并且包含在数据平面VCN 2718中的服务网关2736可以通信地耦合到云服务2756。

[0429] 在一些示例中,图27的框图2700的体系架构所示的模式可以被认为是图10的框图1000的体系架构所示的模式例外,并且如果IaaS提供商不能直接与客户通信(例如,断开连接的区域),那么这种模式可能是IaaS提供商的客户所期望的。客户可以实时访问每个客户的VM 2766(1)-(N)中包含的相应容器2767(1)-(N)。容器2767(1)-(N)可以被配置为对包含在数据平面应用层2746的(一个或多个)应用子网2726中的相应辅助VNIC 2772(1)-(N)进行调用,该数据平面应用层2746可以包含在容器出口VCN 2768中。辅助VNIC 2772(1)-(N)可以将调用传输到NAT网关2738,NAT网关2738可以将调用传输到公共互联网2754。在这个示例中,可以由客户实时访问的容器2767(1)-(N)可以与控制平面VCN 2716隔离,并且可以与数据平面VCN 2718中包含的其它实体隔离。容器2767(1)-(N)也可以与来自其它客户的资源隔离。

[0430] 在其它示例中,客户可以使用容器2767(1)-(N)来调用云服务2756。在这个示例中,客户可以运行容器2767(1)-(N)中从云服务2756请求服务的代码。容器2767(1)-(N)可

以将该请求传输到辅助VNIC 2772(1) - (N), 辅助VNIC 2772(1) - (N)可以将请求传输到NAT网关, 该NAT网关可以将请求传输到公共互联网2754。公共互联网2754可以经由互联网网关2734将请求传输到包含在控制平面VCN 2716中的一个或多个)LB子网2722。响应于确定请求有效, (一个或多个)LB子网可以将请求传输到(一个或多个)应用子网2726, 该(一个或多个)应用子网2726可以经由服务网关2736将请求传输到云服务2756。

[0431] 应该认识到的是, 各图中描绘的IaaS体系架构2400、2500、2600、2700可以具有除所描绘的那些之外的其它组件。另外, 各图中所示的实施例仅仅是可以结合本公开的实施例的云基础设施系统的一些示例。在一些其它实施例中, IaaS系统可以具有比各图中所示更多或更少的组件、可以组合两个或更多个组件, 或者可以具有不同的配置或组件布置。

[0432] 在某些实施例中, 本文描述的IaaS系统可以包括以自助服务、基于订阅、弹性可扩展、可靠、高度可用和安全的方式交付给客户的应用套件、中间件和数据库服务产品。此类IaaS系统的示例是本受让人提供的Oracle云基础设施(OCI)。

[0433] 图28图示了其中可以实现各种实施例的示例计算机系统2800。系统2800可以用于实现上述任何计算机系统。如图所示, 计算机系统2800包括经由总线子系统2802与多个外围子系统通信的处理单元2804。这些外围子系统可以包括处理加速单元2806、I/O子系统2808、存储子系统2818和通信子系统2824。存储子系统2818包括有形计算机可读存储介质2822和系统存储器2810。

[0434] 总线子系统2802提供用于让计算机系统2800的各种部件和子系统按意图彼此通信的机制。虽然总线子系统2802被示意性地示出为单条总线, 但是总线子系统的替代实施例可以利用多条总线。总线子系统2802可以是若干种类型的总线结构中的任何一种, 包括存储器总线或存储器控制器、外围总线、以及使用任何各种总线体系架构的局部总线。例如, 这种体系架构可以包括工业标准体系架构(ISA)总线、微通道体系架构(MCA)总线、增强型ISA(EISA)总线、视频电子标准协会(VESA)局部总线和外围部件互连(PCI)总线, 其可以被实现为按IEEE P1386.1标准制造的Mezzanine总线。

[0435] 可以被实现为一个或多个集成电路(例如, 常规微处理器或微控制器)的处理单元2804控制计算机系统2800的操作。一个或多个处理器可以被包括在处理单元2804中。这些处理器可以包括单核或多核处理器。在某些实施例中, 处理单元2804可以被实现为一个或多个独立的处理单元2832和/或2834, 其中在每个处理单元中包括单核或多核处理器。在其它实施例中, 处理单元2804也可以被实现为通过将两个双核处理器集成到单个芯片中形成的四核处理单元。

[0436] 在各种实施例中, 处理单元2804可以响应于程序代码执行各种程序并且可以维护多个并发执行的程序或进程。在任何给定的时间, 要被执行的程序代码中的一些或全部代码可以驻留在(一个或多个)处理器2804中和/或存储子系统2818中。通过适当的编程, (一个或多个)处理器2804可以提供上述各种功能。计算机系统2800可以附加地包括处理加速单元2806, 其可以包括数字信号处理器(DSP)、专用处理器, 等等。

[0437] I/O子系统2808可以包括用户接口输入设备和用户接口输出设备。用户接口输入设备可以包括键盘、诸如鼠标或轨迹球的定点设备、结合到显示器中的触摸板或触摸屏、滚动轮、点击轮、拨盘、按钮、开关、键盘、具有语音命令识别系统的音频输入设备、麦克风以及其它类型的输入设备。用户接口输入设备可以包括, 例如, 运动感测和/或手势识别设备, 诸

如的Microsoft **Kinect®**运动传感器,其使得用户能够使用手势和语音命令通过自然用户接口来控制诸如的Microsoft **Xbox®** 360游戏控制器的输入设备并与之交互。用户接口输入设备也可以包括眼睛姿势识别设备,诸如从用户检测眼睛活动(例如,当拍摄照片和/或做出菜单选择时的“眨眼”)并且将眼睛姿势转换为到输入设备(例如,Google **Glass®**)中的输入的Google **Glass®**眨眼检测器。此外,用户接口输入设备可以包括使用户能够通过语音命令与语音识别系统(例如,**Siri®**导航器)交互的语音识别感测设备。

[0438] 用户接口输入设备也可以包括但不限于三维(3D)鼠标、操纵杆或指向棒、游戏面板和绘图板,以及音频/视频设备,诸如扬声器、数码相机、数码摄像机、便携式媒体播放器、网络摄像头、图像扫描仪、指纹扫描仪、条形码阅读器3D扫描仪、3D打印机、激光测距仪和视线跟踪设备。此外,用户接口输入设备可以包括,例如,医学成像输入设备,诸如计算机断层扫描、磁共振成像、正电子发射断层摄影术、医疗超声设备。用户接口输入设备也可以包括,例如,诸如MIDI键盘、数字乐器等的音频输入设备。

[0439] 用户接口输出设备可以包括显示子系统、指示灯,或者诸如音频输出设备的非可视显示器,等等。显示子系统可以是阴极射线管(CRT)、诸如使用液晶显示器(LCD)或等离子显示器的平板设备、投影设备、触摸屏,等等。一般而言,术语“输出设备”的使用意在包括用于从计算机系统2800向用户或其它计算机输出信息的所有可能类型的设备和机制。例如,用户接口输出设备可以包括,但不限于,可视地传达文本、图形和音频/视频信息的各种显示设备,诸如监视器、打印机、扬声器、耳机、汽车导航系统、绘图仪、语音输出设备,以及调制解调器。

[0440] 计算机系统2800可以包括包含软件元件、被示为当前位于系统存储器2810中的存储子系统2818。系统存储器2810可以存储可加载并且可在处理单元2804上执行的程序指令,以及在这些程序的执行期间所产生的数据。

[0441] 取决于计算机系统2800的配置和类型,系统存储器2810可以是易失性的(诸如随机存取存储器(RAM))和/或非易失性的(诸如只读存储器(ROM)、闪存存储器,等等)。RAM通常包含可被处理单元2804立即访问和/或目前正被处理单元2804操作和执行的数据和/或程序模块。在一些实现中,系统存储器2810可以包括多种不同类型的存储器,例如静态随机存取存储器(SRAM)或动态随机存取存储器(DRAM)。在一些实现中,诸如包含有助于在启动期间在计算机系统2800的元件之间传送信息的基本例程的基本输入/输出系统(BIOS),通常可以被存储在ROM中。作为示例,但不是限制,系统存储器2810也示出了可以包括客户端应用、web浏览器、中间层应用、关系数据库管理系统(RDBMS)等的应用程序2812,程序数据2814,以及操作系统2816。作为示例,操作系统2816可以包括各种版本的Microsoft **Windows®**、Apple **Macintosh®**和/或Linux操作系统、各种可商业获得的**UNIX®**或类UNIX操作系统(包括但不限于各种GNU/Linux操作系统、Google **Chrome®**操作系统等)和/或诸如iOS、**Windows®** Phone、**Android®** OS、**BlackBerry®** 280S和**Palm®** OS操作系统的移动操作系统。

[0442] 存储子系统2818也可以提供用于存储提供一些实施例的功能的基本编程和数据结构的有形计算机可读存储介质。当被处理器执行时提供上述功能的软件(程序、代码模块、指令)可以被存储在存储子系统2818中。这些软件模块或指令可以被处理单元2804执

行。存储子系统2818也可以提供用于存储根据本公开被使用的数据的储存库。

[0443] 存储子系统2800也可以包括可被进一步连接到计算机可读存储介质2822的计算机可读存储介质读取器2820。与系统存储器2810一起并且,可选地,与其相结合,计算机可读存储介质2822可以全面地表示用于临时和/或更持久地包含、存储、发送和检索计算机可读信息的远程、本地、固定和/或可移除存储设备加存储介质。

[0444] 包含代码或代码的部分的计算机可读存储介质2822也可以包括本领域已知或使用的任何适当的介质,包括存储介质和通信介质,诸如但不限于,以用于信息的存储和/或传输的任何方法或技术实现的易失性和非易失性、可移除和不可移除介质。这可以包括有形的计算机可读存储介质,诸如RAM、ROM、电可擦除可编程ROM(EEPROM)、闪存存储器或其它存储器技术、CD-ROM、数字多功能盘(DVD)或其它光学存储器、磁带盒、磁带、磁盘存储器或其它磁存储设备,或者其它有形的计算机可读介质。这也可以包括非有形的计算机可读介质,诸如数据信号、数据传输,或者可以被用来发送期望信息并且可以被计算系统2800访问的任何其它介质。

[0445] 举例来说,计算机可读存储介质2822可以包括从不可移除的非易失性磁介质读取或写到其的硬盘驱动器、从可移除的非易失性磁盘读取或写到其的磁盘驱动器、以及从可移除的非易失性光盘(诸如CD ROM、DVD和**Blu-Ray®**盘或其它光学介质)读取或写到其的光盘驱动器。计算机可读存储介质2822可以包括,但不限于,**Zip®**驱动器、闪存卡、通用串行总线(USB)闪存驱动器、安全数字(SD)卡、DVD盘、数字音频带,等等。计算机可读存储介质2822也可以包括基于非易失性存储器的固态驱动器(SSD)(诸如基于闪存存储器的SSD、企业闪存驱动器、固态ROM等)、基于易失性存储器的SSD(诸如固态RAM、动态RAM、静态RAM)、基于DRAM的SSD,磁阻RAM(MRAM) SSD,以及使用基于DRAM和闪存存储器的SSD的組合的混合SSD。盘驱动器及其关联的计算机可读介质可以为计算机系统2800提供计算机可读指令、数据结构、程序模块及其它数据的非易失性存储。

[0446] 通信子系统2824提供到其它计算机系统和网络的接口。通信子系统2824用作用于从其它系统接收数据和从计算机系统2800向其它系统发送数据的接口。例如,通信子系统2824可以使计算机系统2800能够经由互联网连接到一个或多个设备。在一些实施例中,通信子系统2824可以包括用于访问无线语音和/或数据网络的射频(RF)收发器部件(例如,使用蜂窝电话技术,诸如3G、4G或EDGE(用于全球演进的增强型数据速率)的先进数据网络技术,WiFi(IEEE 802.11系列标准),或其它移动通信技术,或其任意组合)、全球定位系统(GPS)接收器部件和/或其它部件。在一些实施例中,作为无线接口的附加或者替代,通信子系统2824可以提供有线网络连接(例如,以太网)。

[0447] 在一些实施例中,通信子系统2824也可以代表可以使用计算机系统2800的一个或多个用户接收结构化和/或非结构化数据馈送2826、事件流2828、事件更新2830等形式的输入通信。

[0448] 举例来说,通信子系统2824可以被配置为实时地从社交网络和/或其它通信服务的用户接收数据馈送2826,诸如**Twitter®**馈送、**Facebook®**更新、诸如丰富站点摘要(RSS)馈送的web馈送和/或来自一个或多个第三方信息源的实时更新。

[0449] 此外,通信子系统2824也可被配置为接收连续数据流形式的数据,这可以包括本

质上可以是连续的或无界的没有明确终止的实时事件的事件流2828和/或事件更新2830。产生连续数据的应用的示例可以包括,例如,传感器数据应用、金融报价机、网络性能测量工具(例如,网络监视和流量管理应用)、点击流分析工具、汽车流量监视,等等。

[0450] 通信子系统2824也可被配置为向一个或多个数据库输出结构化和/或非结构化数据馈送2826、事件流2828、事件更新2830,等等,这一个或多个数据库可以与耦合到计算机系统2800的一个或多个流式数据源计算机通信。

[0451] 计算机系统2800可以是各种类型之一,包括手持便携式设备(例如,**iPhone®**蜂窝电话、**iPad®**计算平板电脑、PDA)、可穿戴设备(例如,**Google®** Glass头戴式显示器)、PC、工作站、大型机、信息站、服务器机架、或任何其它数据处理系统。

[0452] 由于计算机和网络的不断变化的本质,在图中绘出的计算机系统2800的描述仅仅要作为具体的示例。具有比图中绘出的系统更多或更少部件的许多其它配置是可能的。例如,定制的硬件也可以被使用和/或特定的元素可以用硬件、固件、软件(包括applets)或其组合来实现。另外,也可以采用到诸如网络输入/输出设备之类的其它计算设备的连接。基于本文提供的公开内容和示教,本领域普通技术人员将认识到实现各种实施例的其它方式和/或方法。

[0453] 虽然已经描述了具体实施例,但是各种修改、变更、替代构造和等效形式也包含在本公开的范围。实施例不限于在某些特定数据处理环境内操作,而是可以在多个数据处理环境内自由操作。此外,虽然已经使用特定系列的事务和步骤描述了实施例,但是本领域技术人员应该清楚本公开的范围不限于所描述系列的事务和步骤。上述实施例的各种特征和方面可以单独或联合使用。

[0454] 另外,虽然已经使用硬件和软件的特定组合描述了实施例,但是应当认识到硬件和软件的其它组合也在本公开的范围。实施例可以仅用硬件、或仅用软件、或使用它们的组合来实现。本文描述的各种处理可以以任何组合在相同的处理器或在不同的处理器上实现。因而,在组件或模块被描述为被配置为执行某些操作的情况下,可以通过例如设计电子电路来执行操作、通过对可编程电子电路(诸如微处理器)进行编程来执行操作,或其任何组合来完成这样的配置。处理可以使用多种技术进行通信,包括但不限于用于处理间通信的常规技术,并且不同的处理对可以使用不同的技术,或者同一对处理可以在不同时间使用不同的技术。

[0455] 因而,说明书和附图被认为是说明性的而不是限制性的。但是,显然可以对其进行添加、减少、删除和其它修改和改变而不背离权利要求中阐述的更广泛的精神和范围。因此,虽然已经描述了具体的公开实施例,但这些并不旨在进行限制。各种修改和等效形式都在以下权利要求的范围内。

[0456] 在描述所公开的实施例的上下文中(尤其在以下权利要求的上下文中)使用术语“一”和“一个”和“该”以及类似的指称要被解释为涵盖单数和复数,除非本文另有指示或与上下文明显矛盾。除非另有说明,否则术语“包括”、“具有”、“包含(including)”和“包含(containing)”要被解释为开放式术语(即,意思是“包括但不限于”)。术语“连接”应被解释为部分或全部包含在、附加到或连接在一起,即使中间存在一些东西。除非本文另有指示,否则本文中值范围的列举仅旨在用作个别引用落入该范围内的每个单独值的速记方法,并且每个单独值被并入说明书中,就好像它在本文中个别列举一样。除非本文另有指示或与

上下文明显矛盾,否则本文所述的所有方法都可以以任何合适的顺序执行。本文提供的任何和所有示例或示例性语言(例如,“诸如”)的使用仅旨在更好地阐明实施例并且不对本公开的范围构成限制,除非另有声明。说明书中的任何语言都不应被解释为指示任何未要求保护的元素对于本公开的实践是必不可少的。

[0457] 析取语言,诸如短语“X、Y或Z中的至少一个”,除非另有明确说明,否则旨在在一般用于表示项目、术语等的上下文中理解,可以是X、Y或Z,或它们的任何组合(例如,X、Y和/或Z)。因此,这种析取语言通常不旨在也不应暗示某些实施例需要X中的至少一个、Y中的至少一个或Z中的至少一个各自存在。

[0458] 本文描述了本公开的优选实施例,包括已知用于实施本公开的最佳模式。那些优选实施例的变型对于本领域普通技术人员在阅读上述描述后会变得显而易见。普通技术人员应该能够适当地采用这样的变型并且可以以不同于本文具体描述的方式来实践本公开。因而,本公开包括在适用法律允许的情况下对所附权利要求中记载的主题的所有修改和等效形式。此外,除非在本文中另有指示,否则本公开包括在其所有可能的变化中的上述元素的任何组合。

[0459] 本文引用的所有参考文献,包括出版物、专利申请和专利,均以相同的程度通过引用并入本文,就好像每个参考文献个别且具体地指示通过引用并入并在本文中全文阐述一样。

[0460] 在前述的说明书中,本公开的各方面参考其具体实施例进行了描述,但本领域技术人员将认识到的是,本公开不限于此。上述公开的各个特征和方面可以被单独或联合使用。此外,在不脱离本说明书的更广泛精神和范围的情况下,实施例可以在除本文所述的那些之外的任何数量的环境 and 应用中被使用。因而,本说明书和附图应当被认为是说明性而不是限制性的。

100

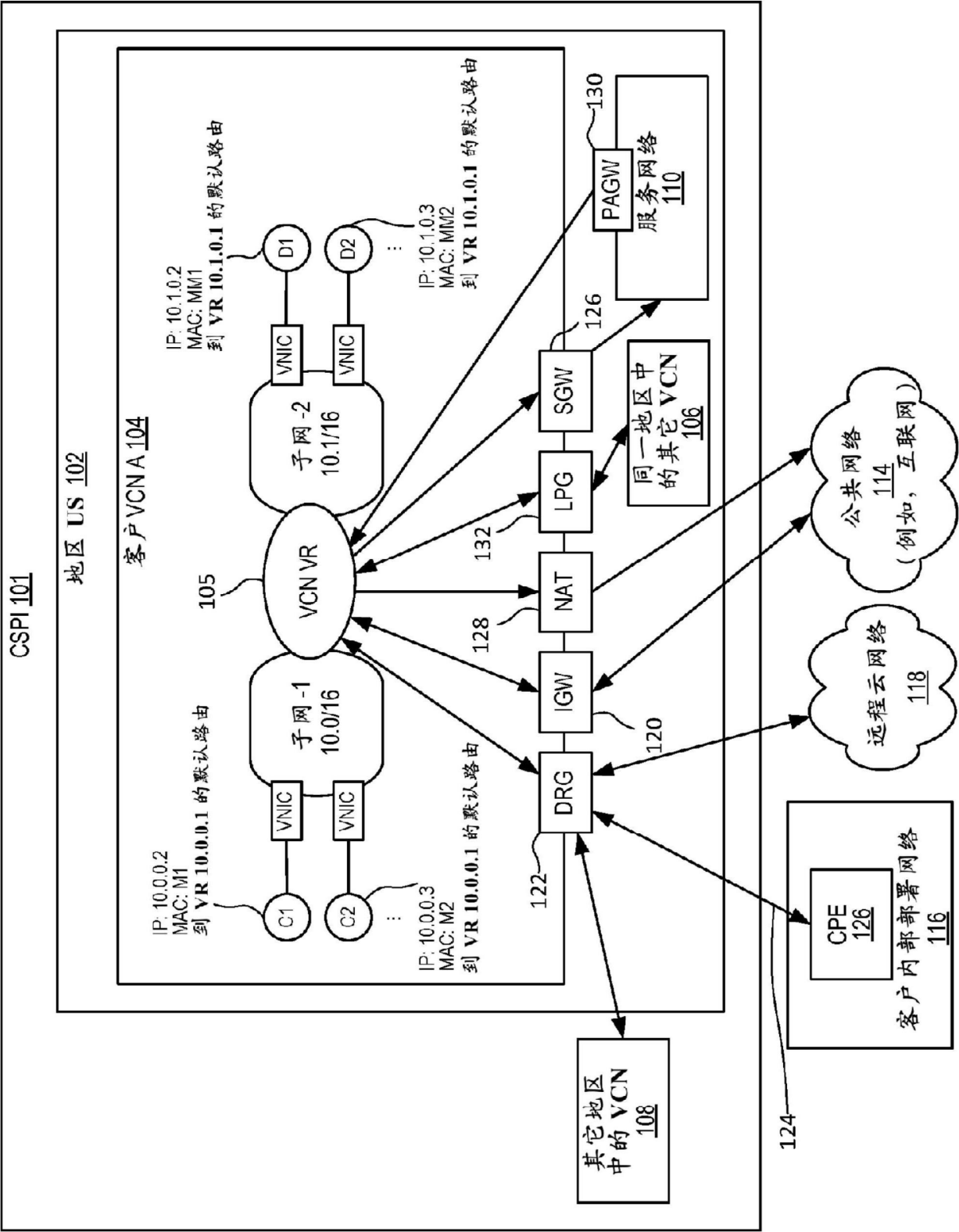


图1

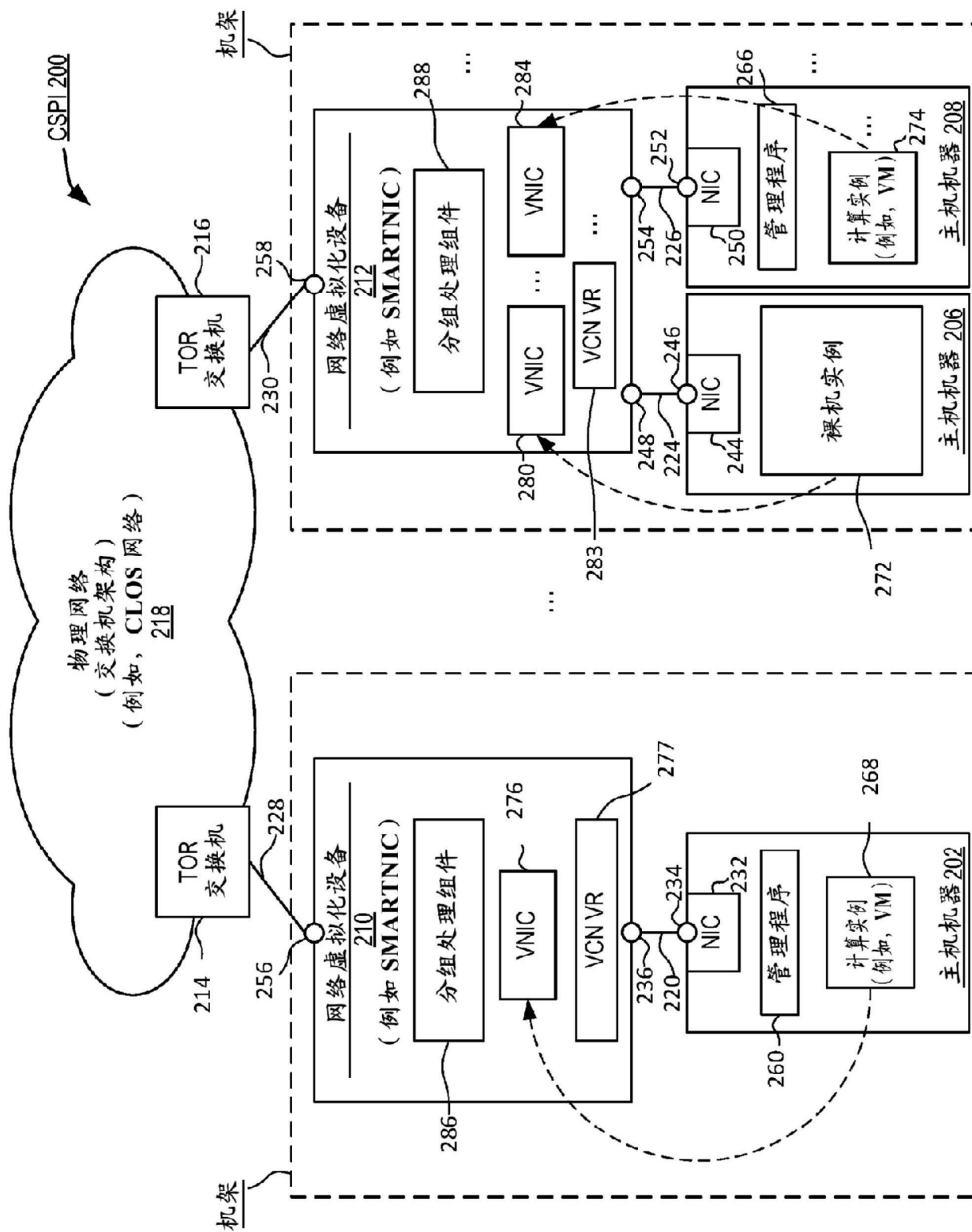


图2

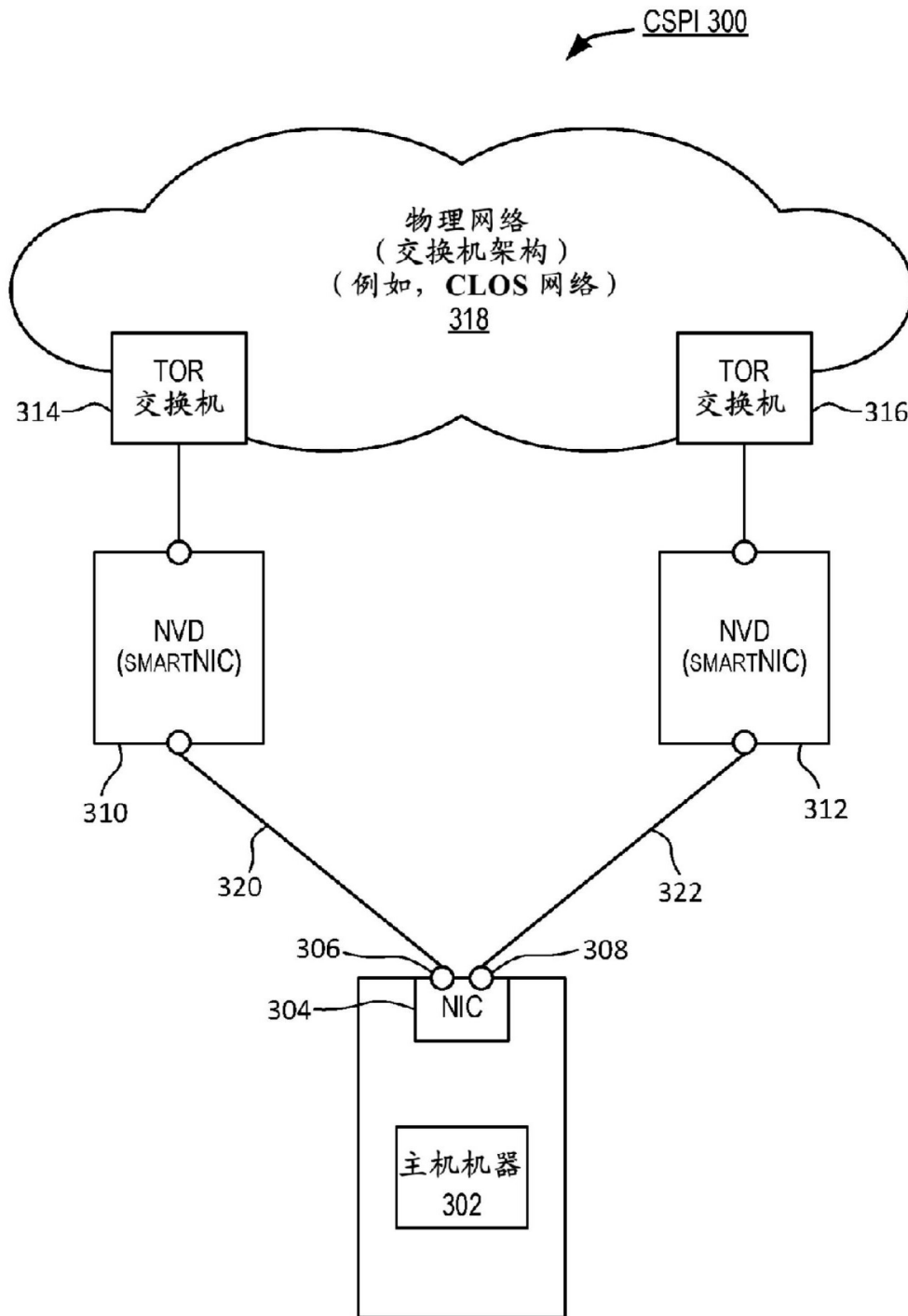


图3

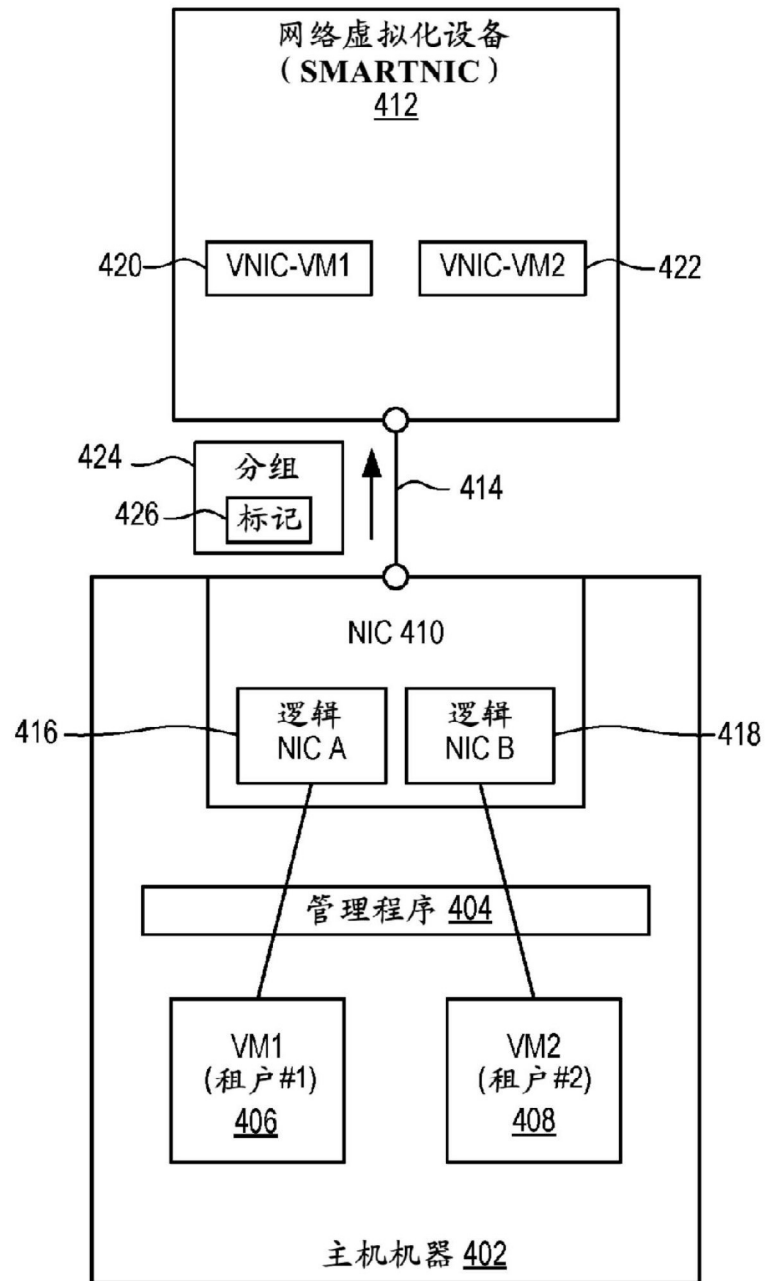


图4

物理网络 (例如, 3 层 CLOS 网络)

500

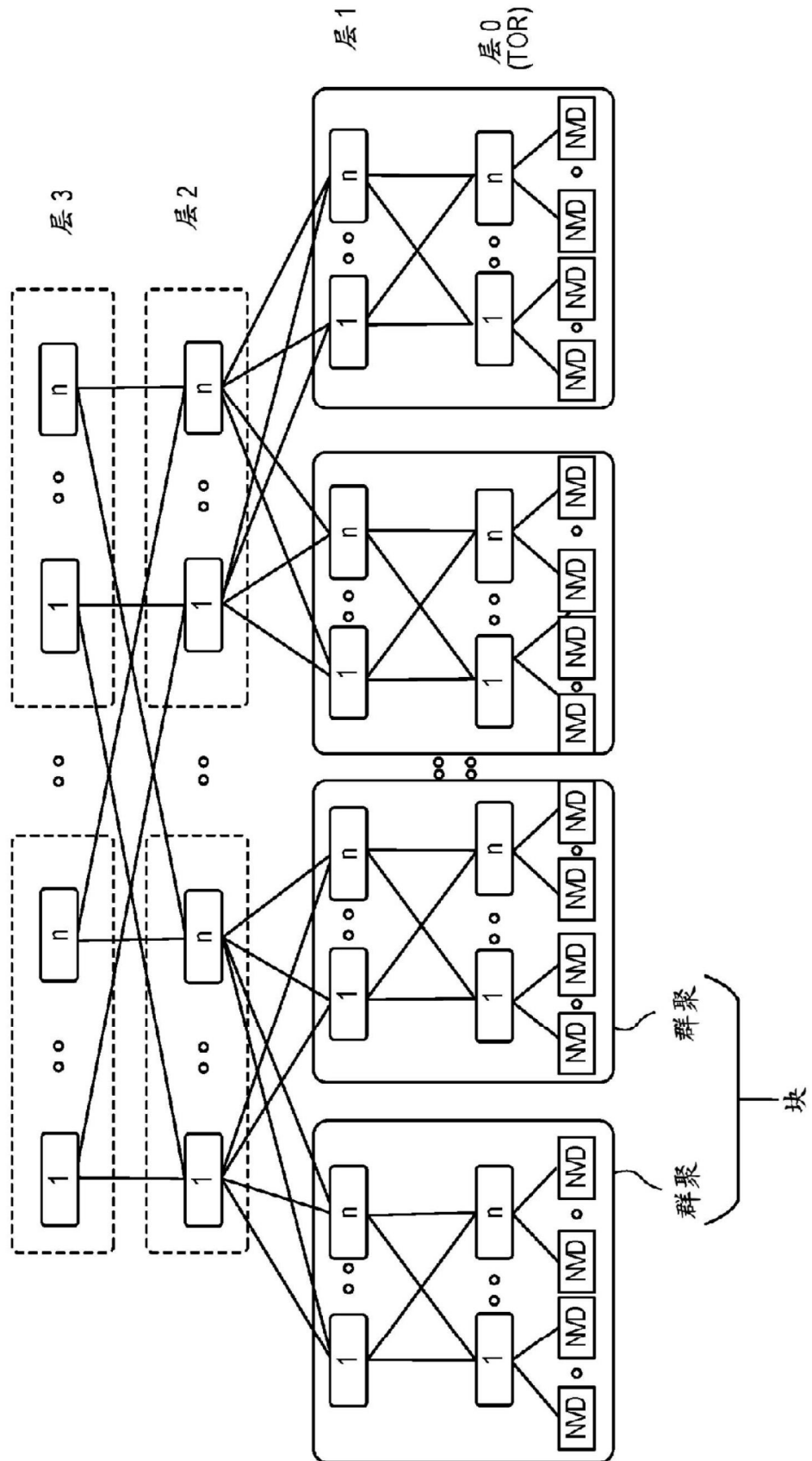


图5

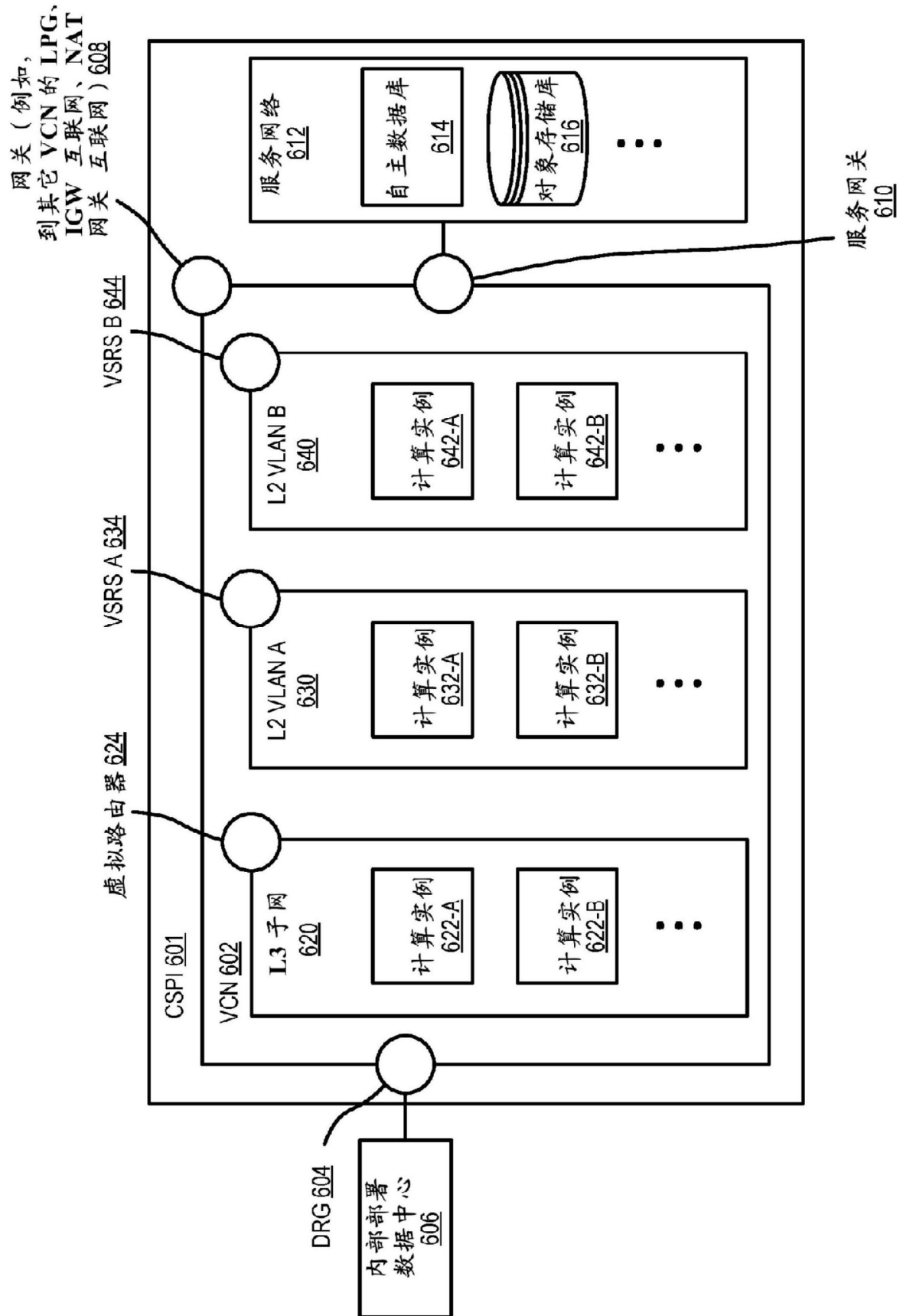


图6

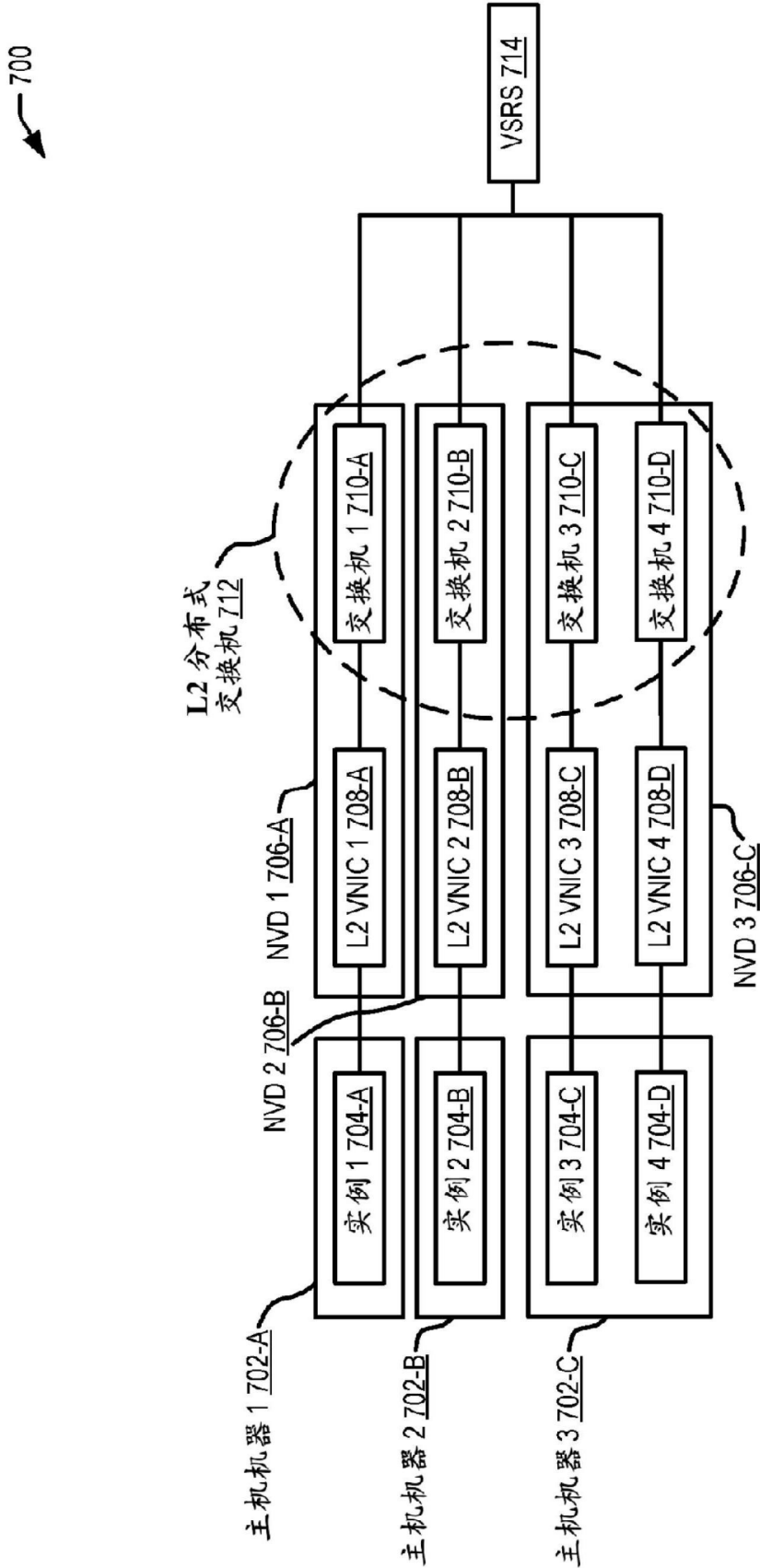


图7

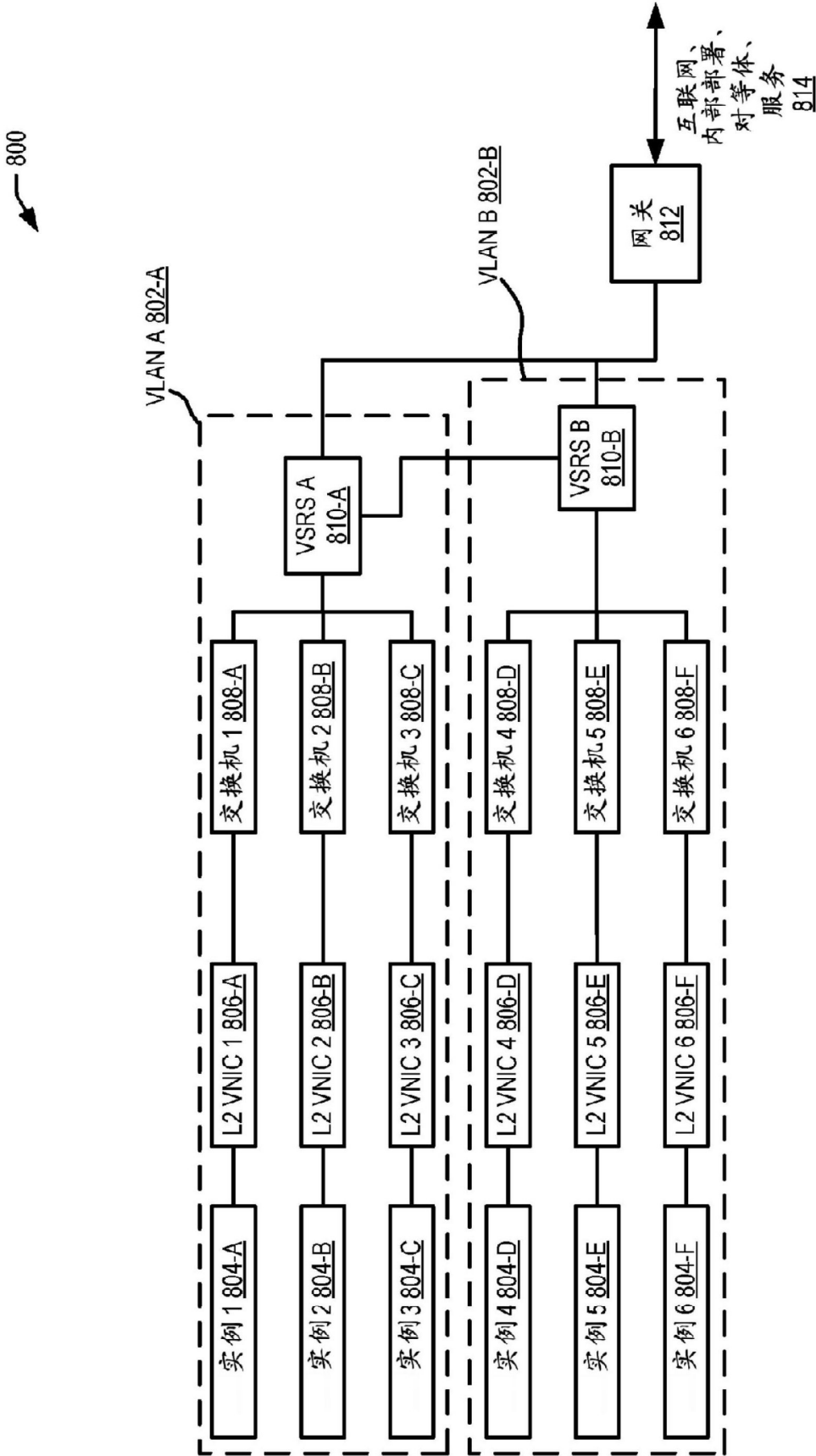


图8

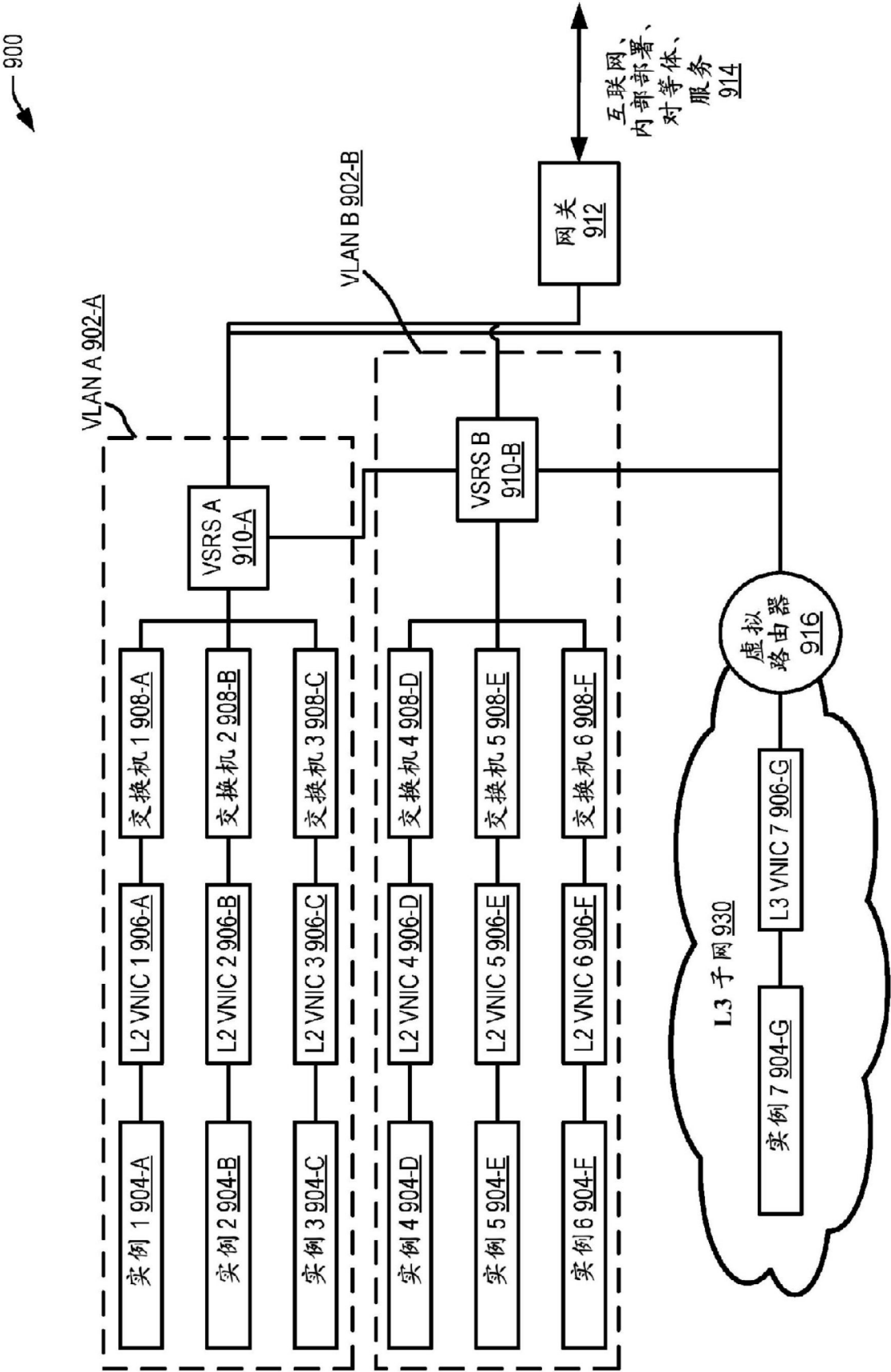


图9

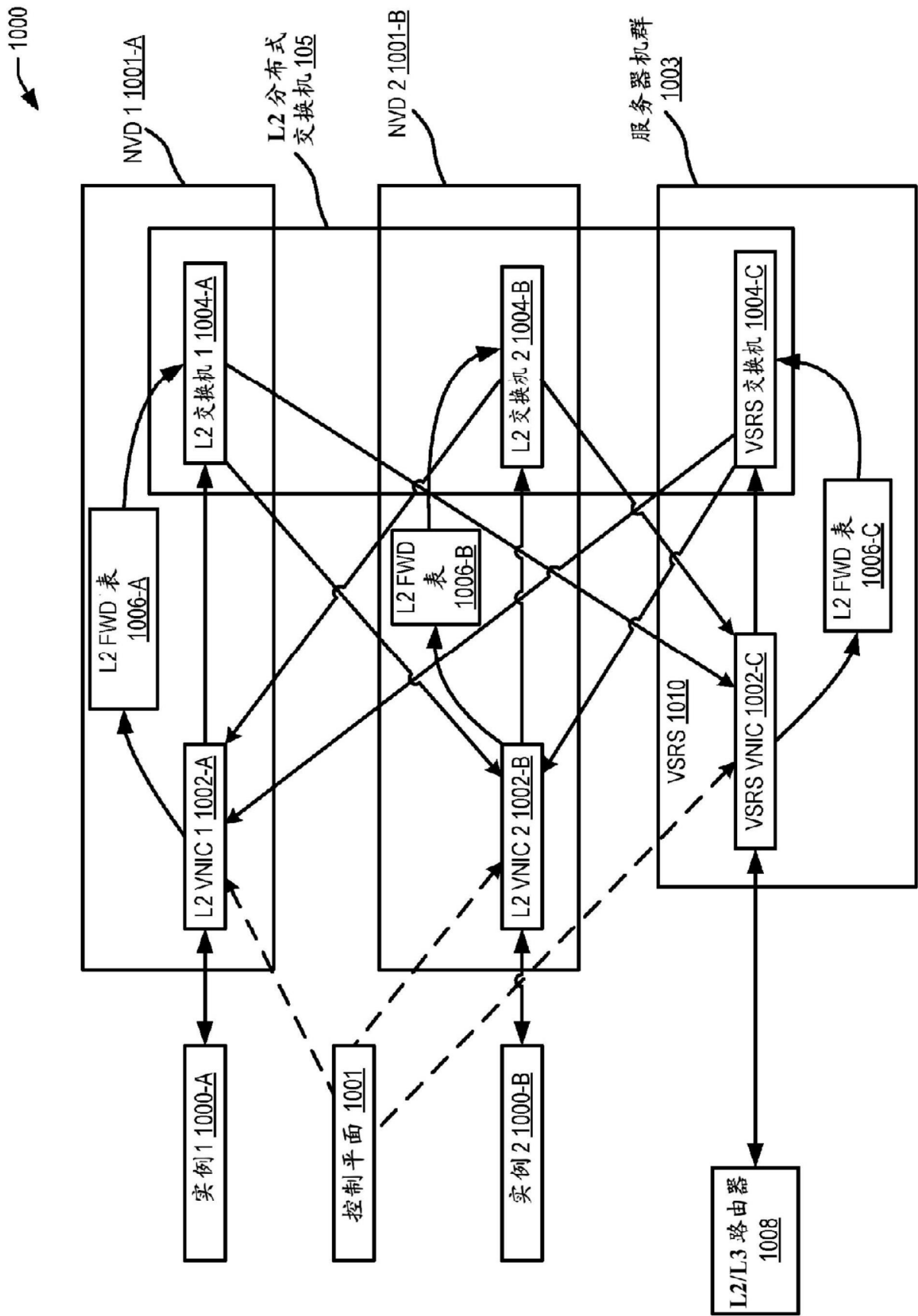


图10

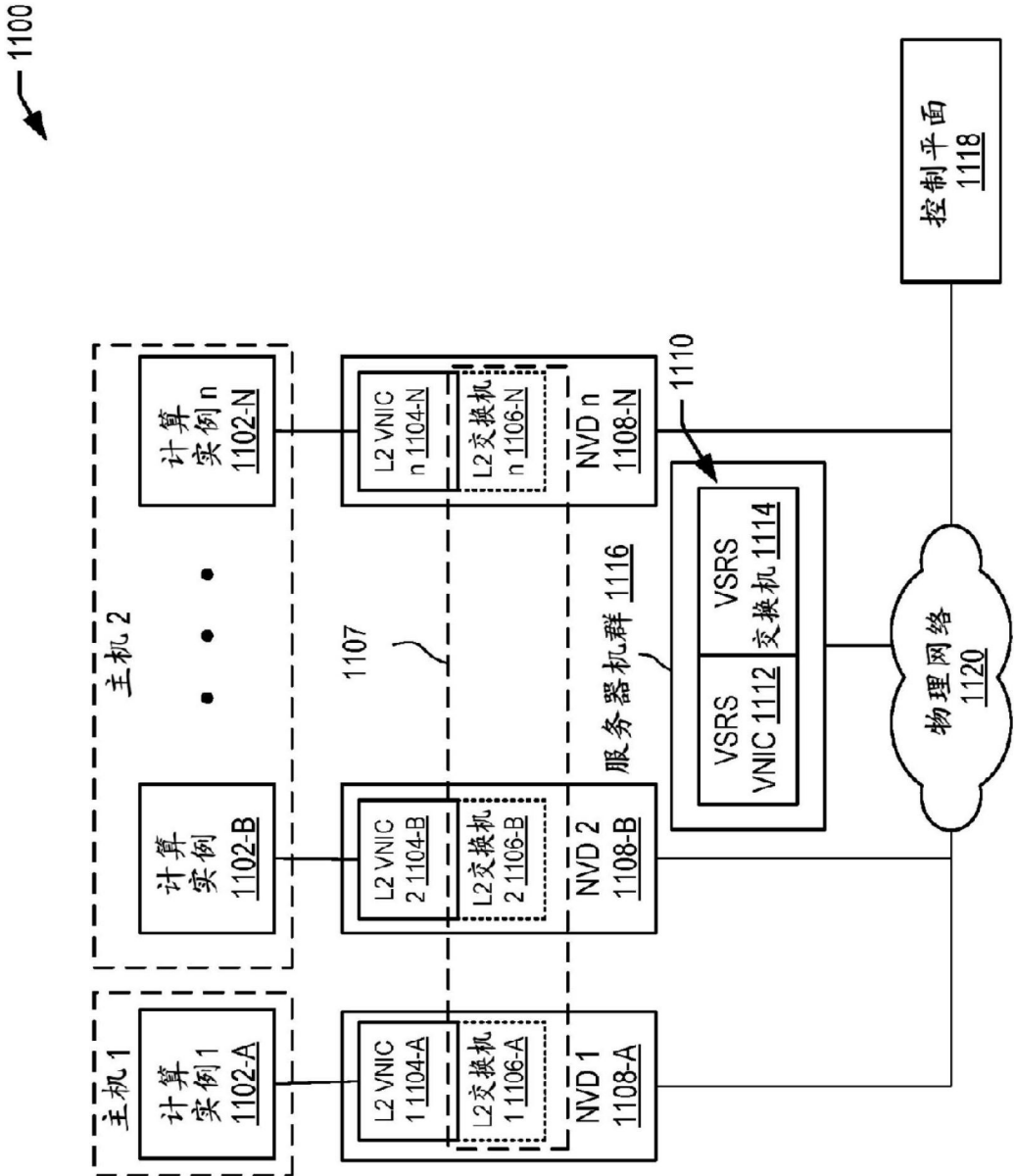


图11

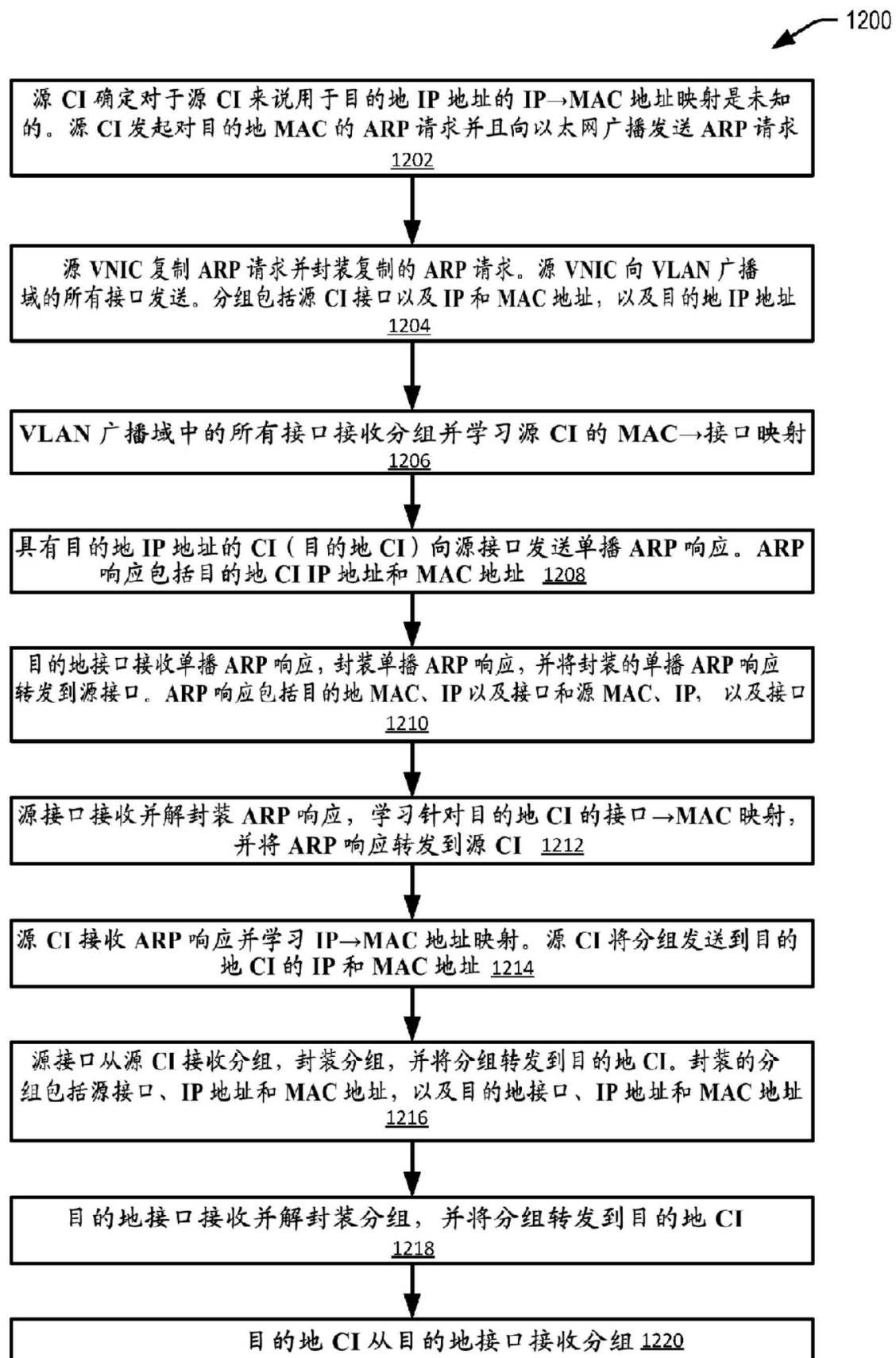


图12

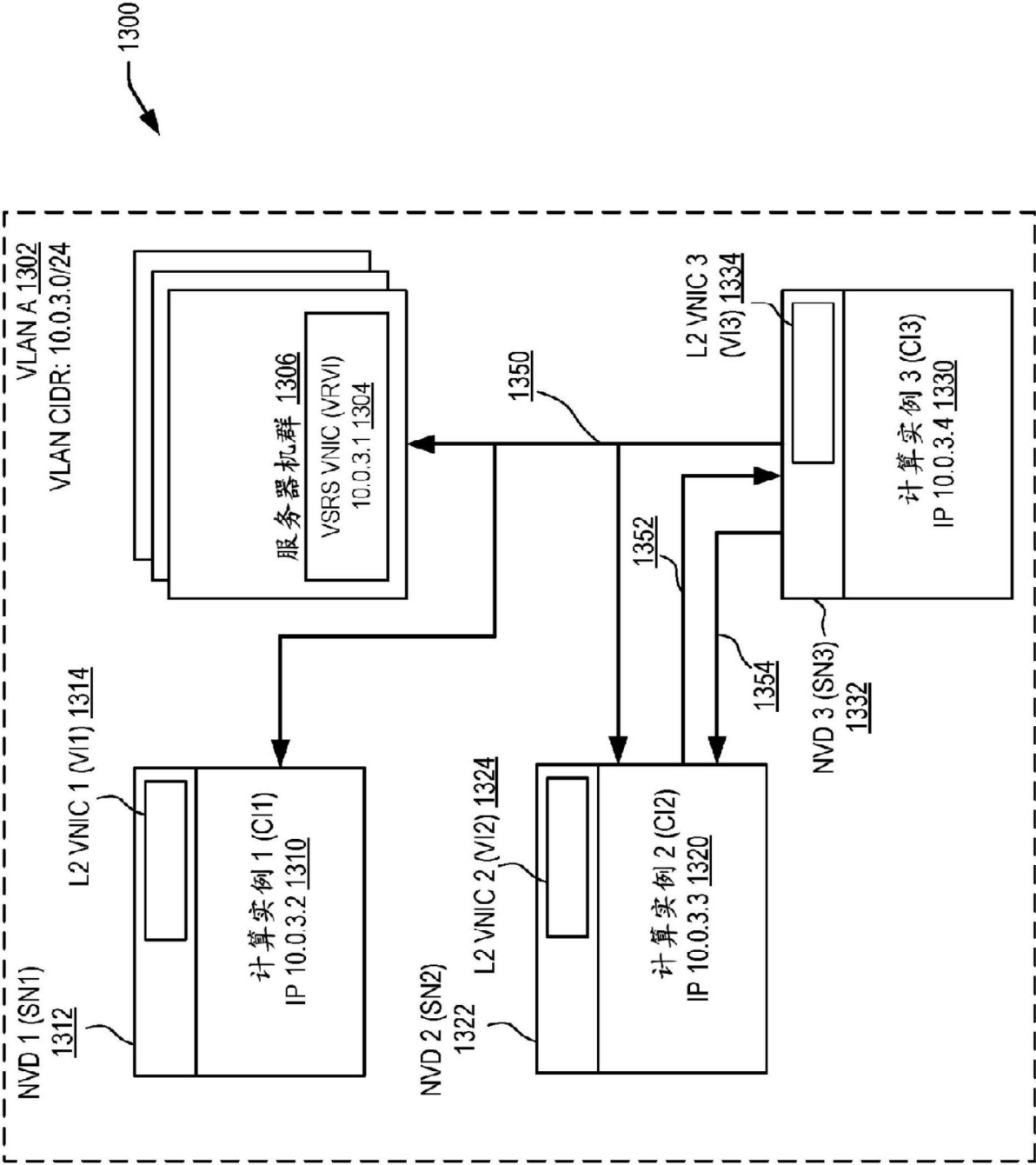


图13

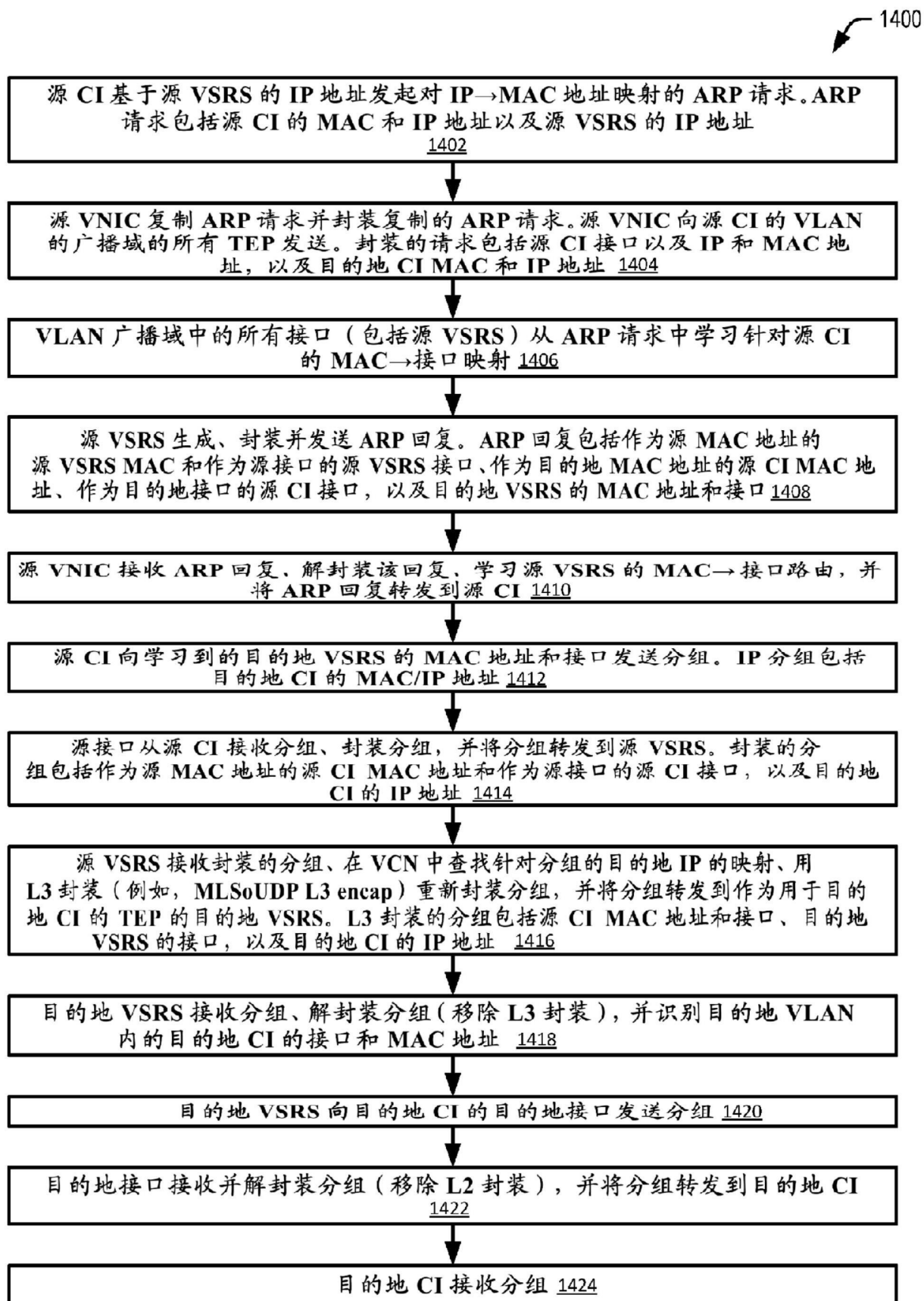


图14

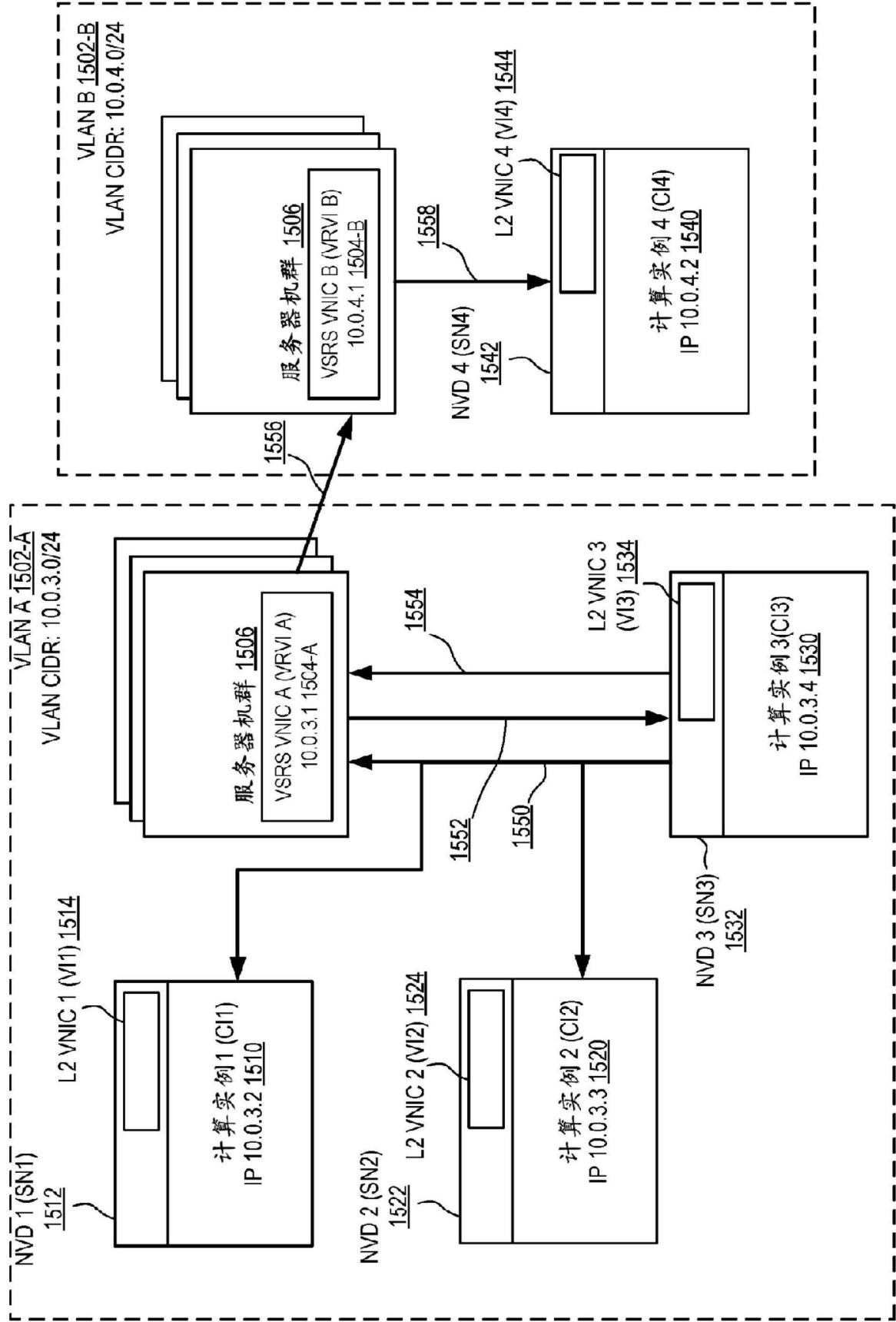


图15

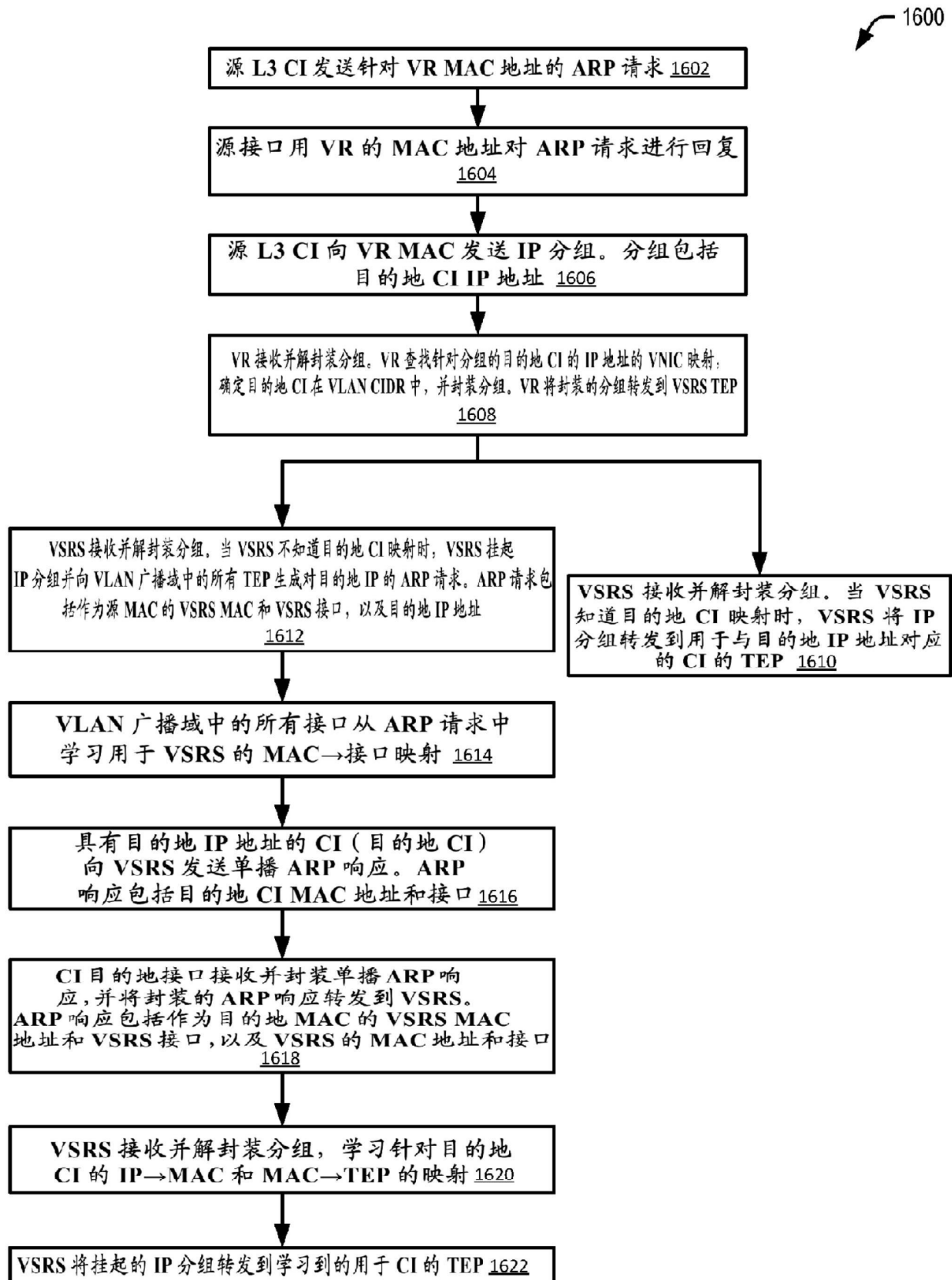


图16

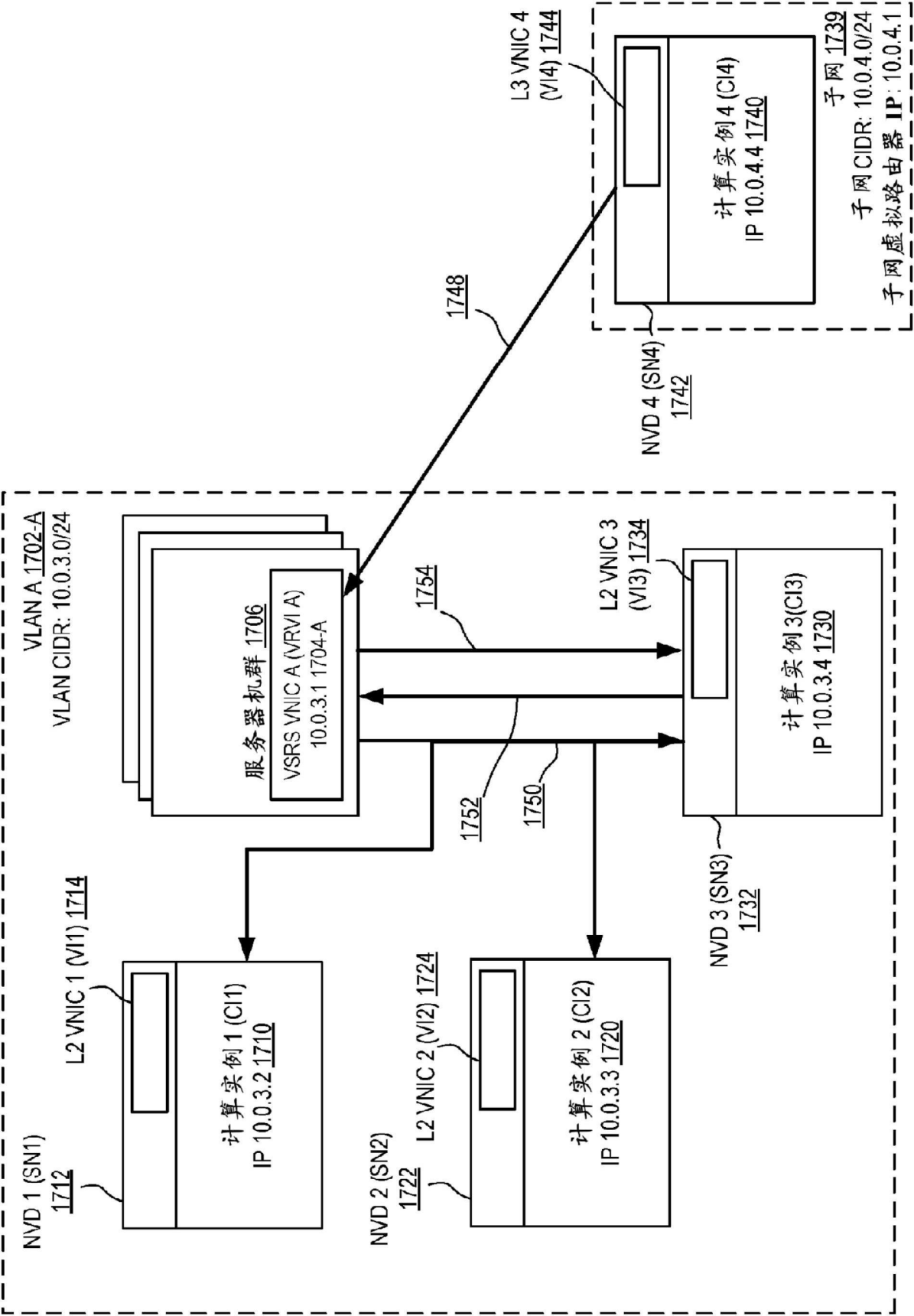


图17

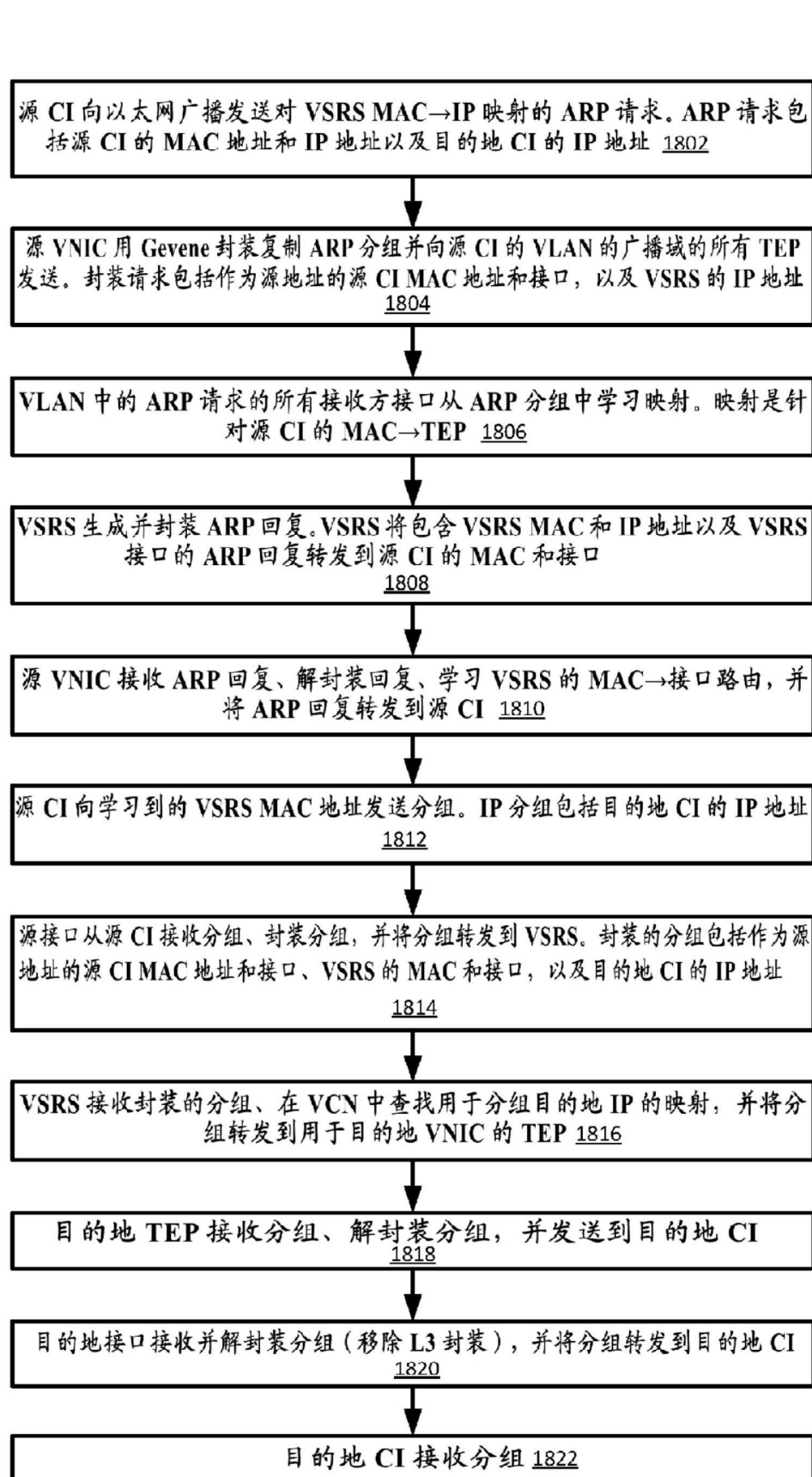


图18

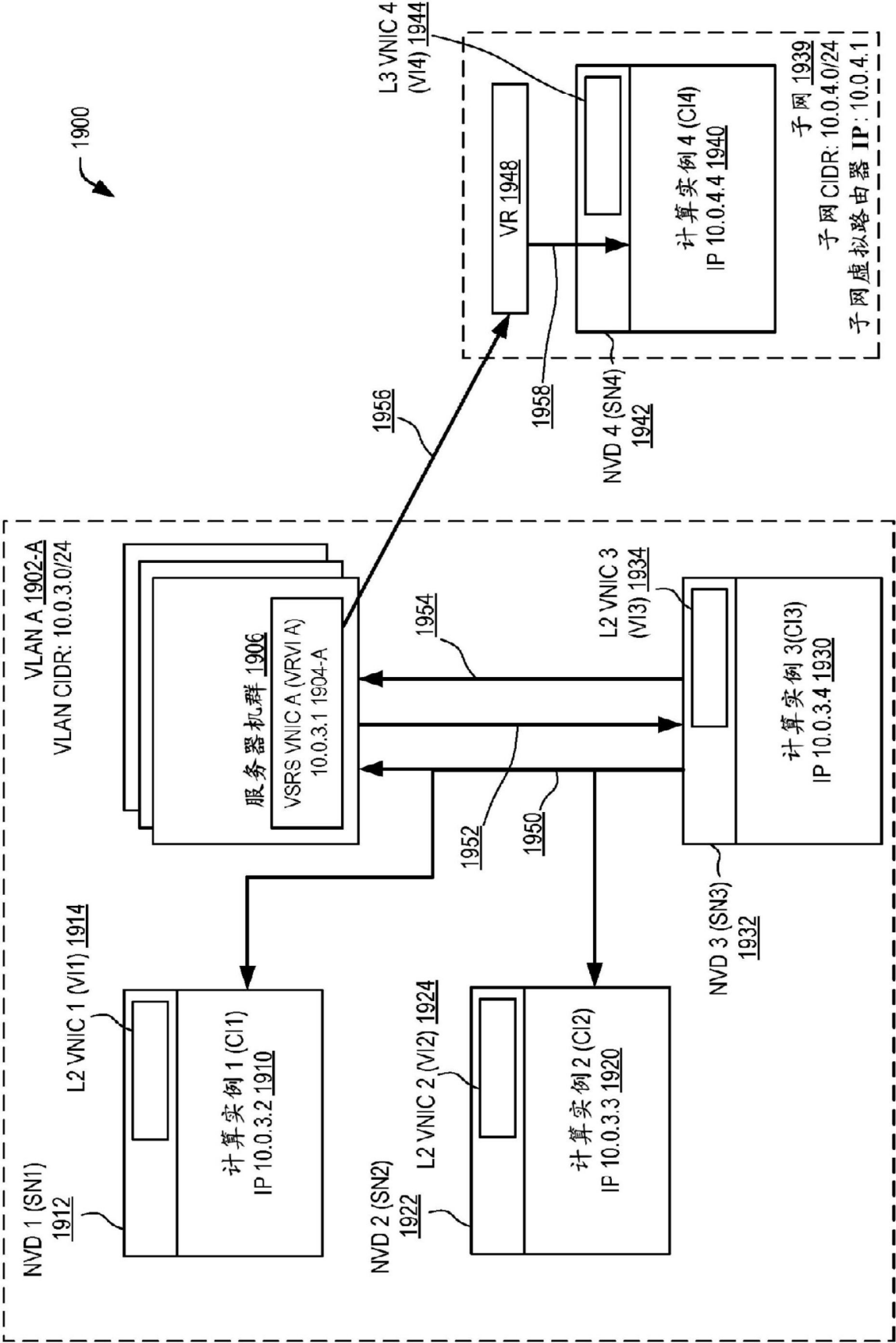


图19

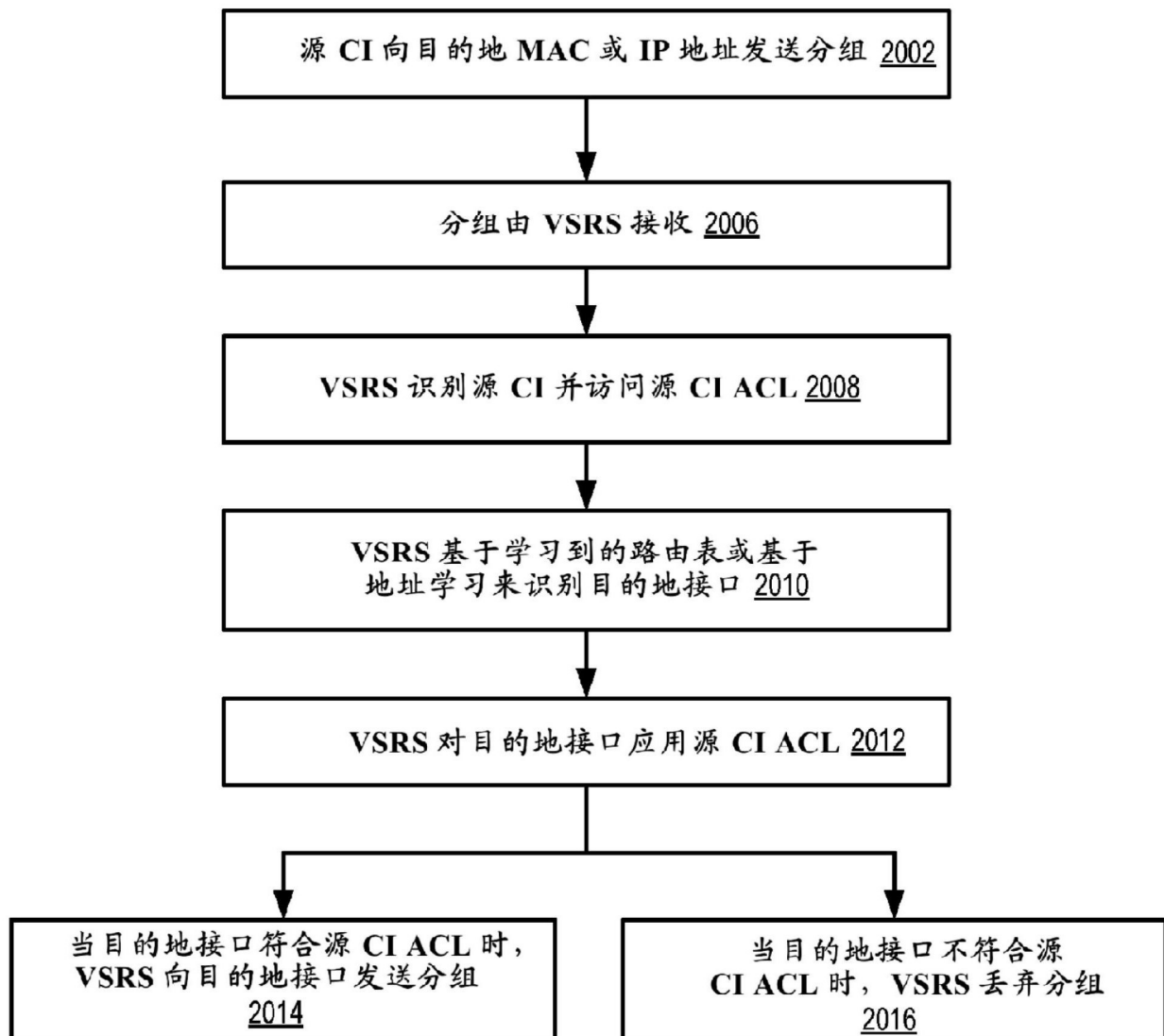
2000
↙

图20

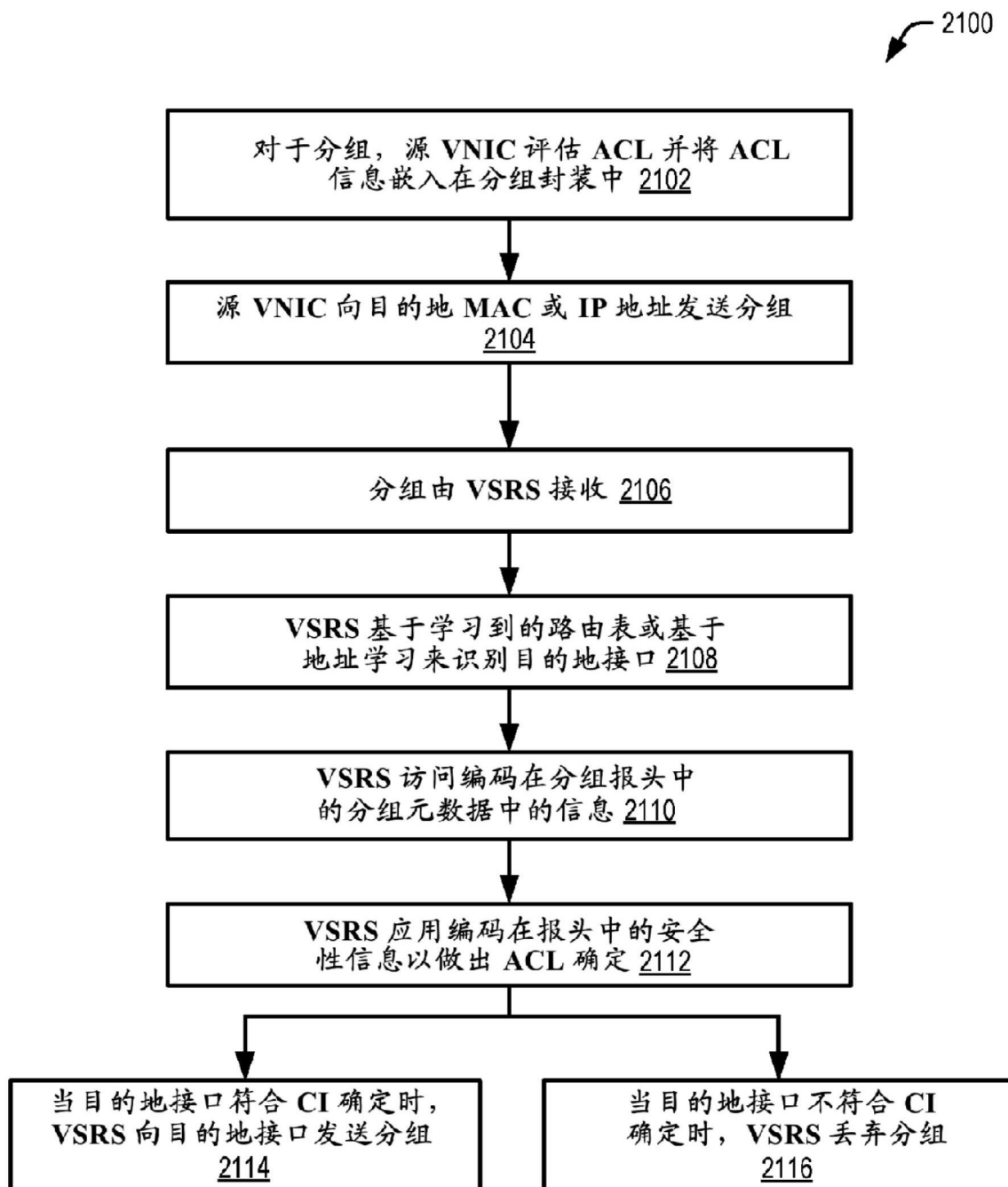


图21

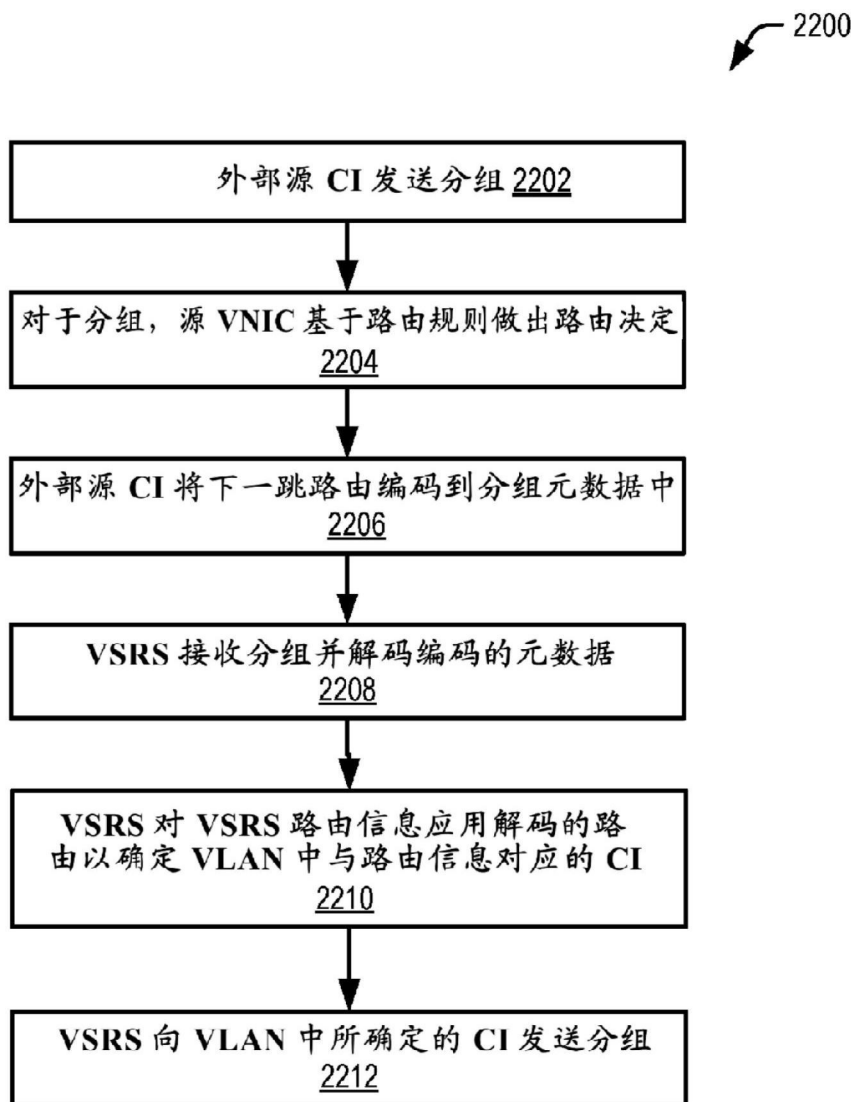


图22

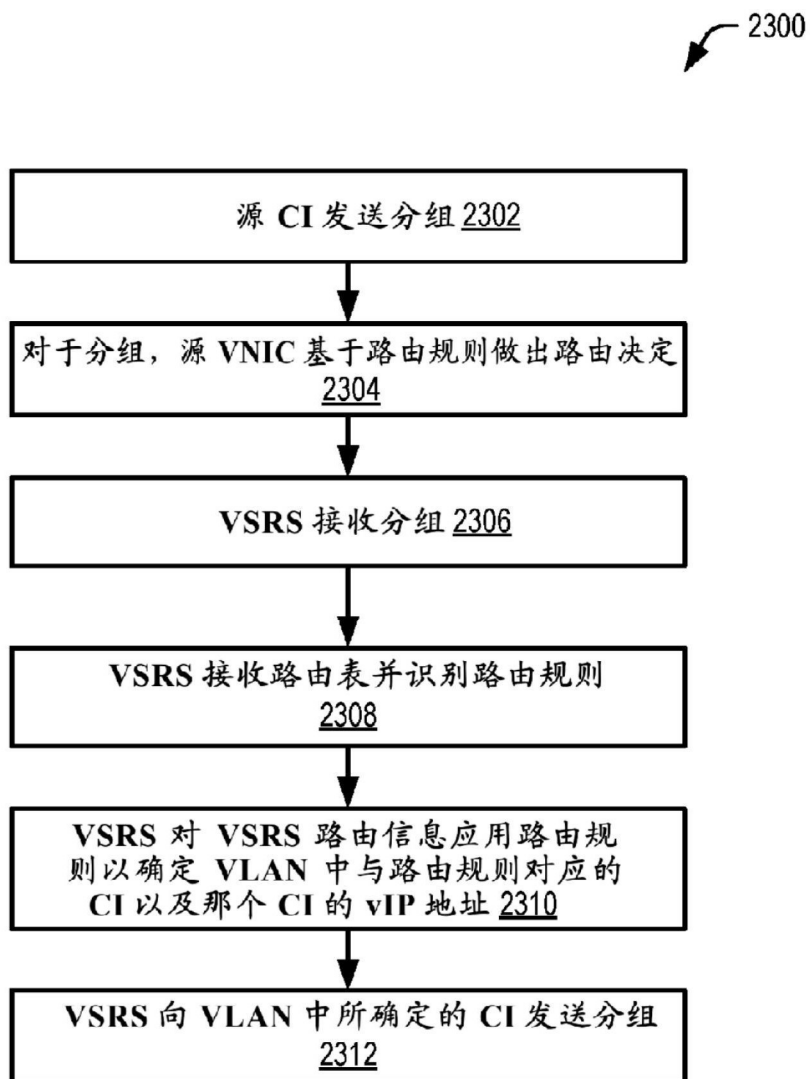


图23

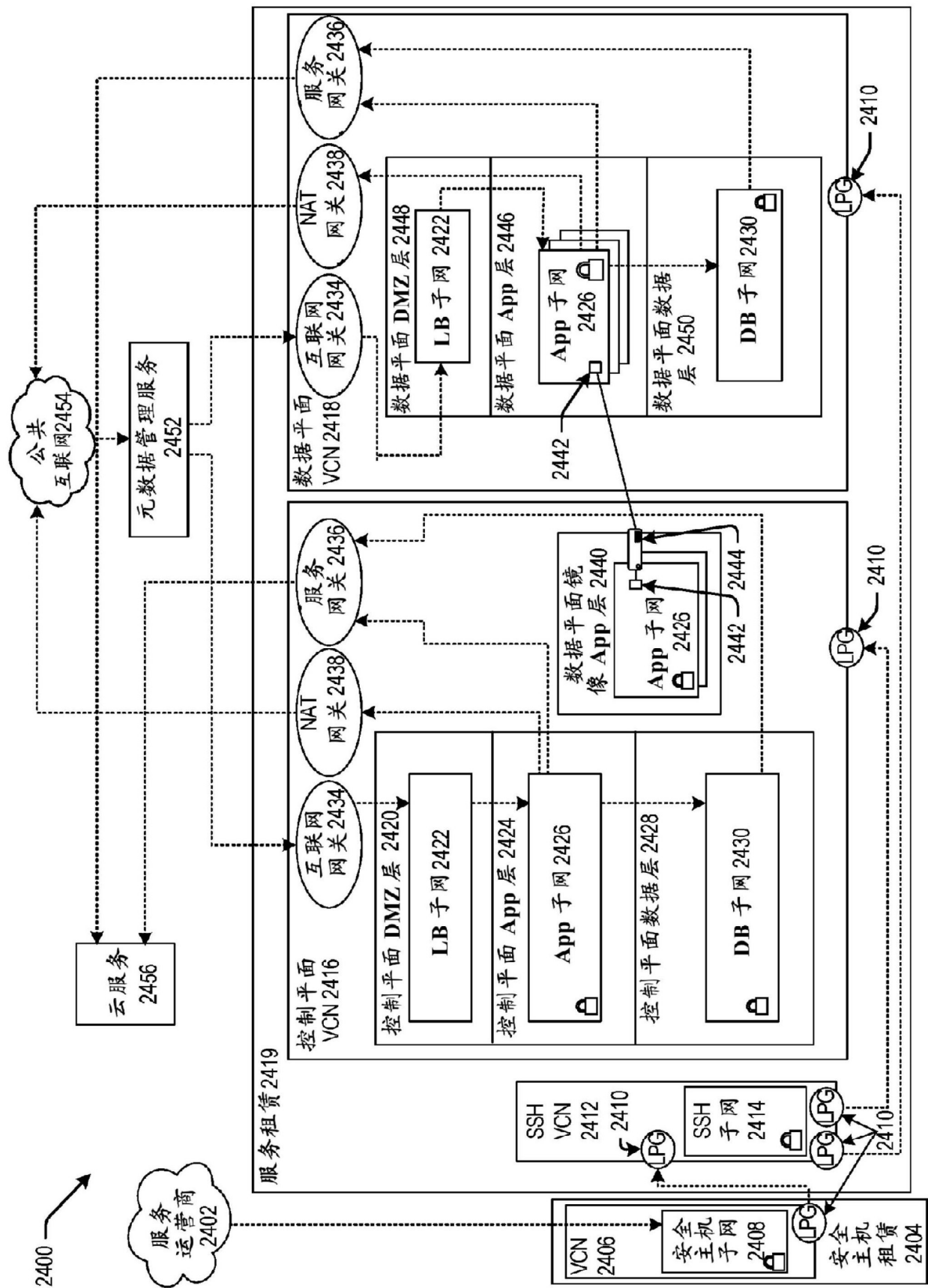


图24

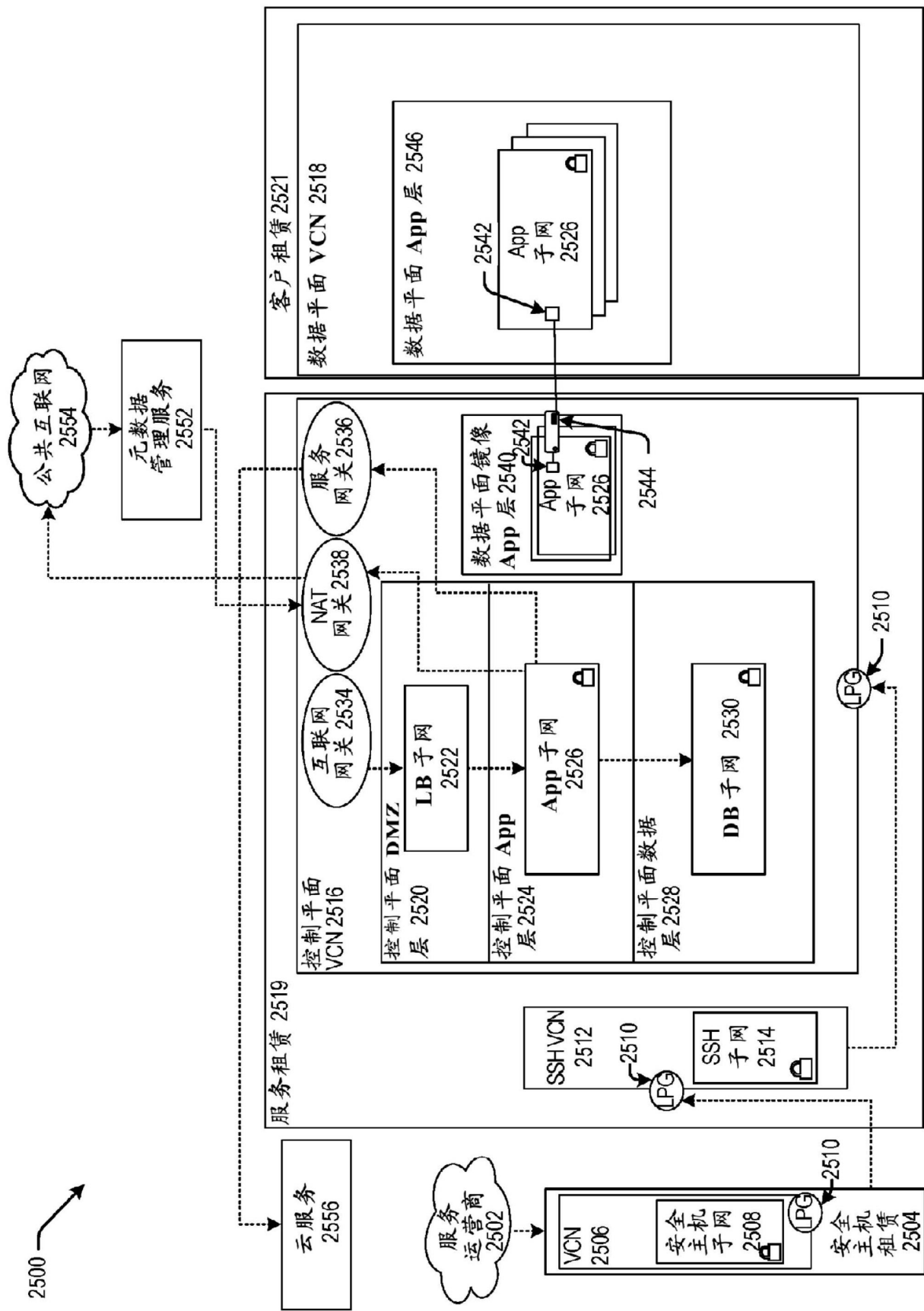


图25

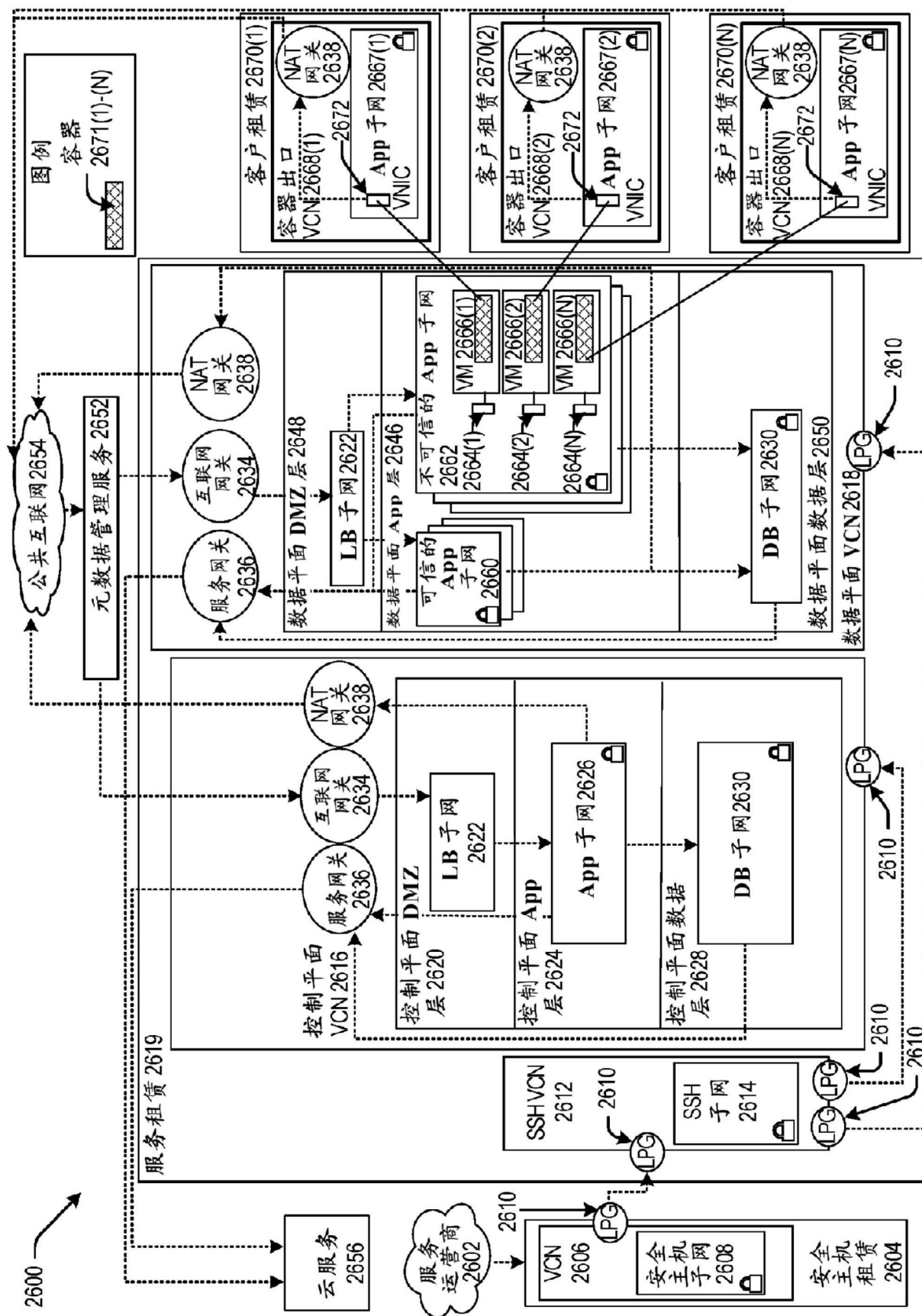


图26

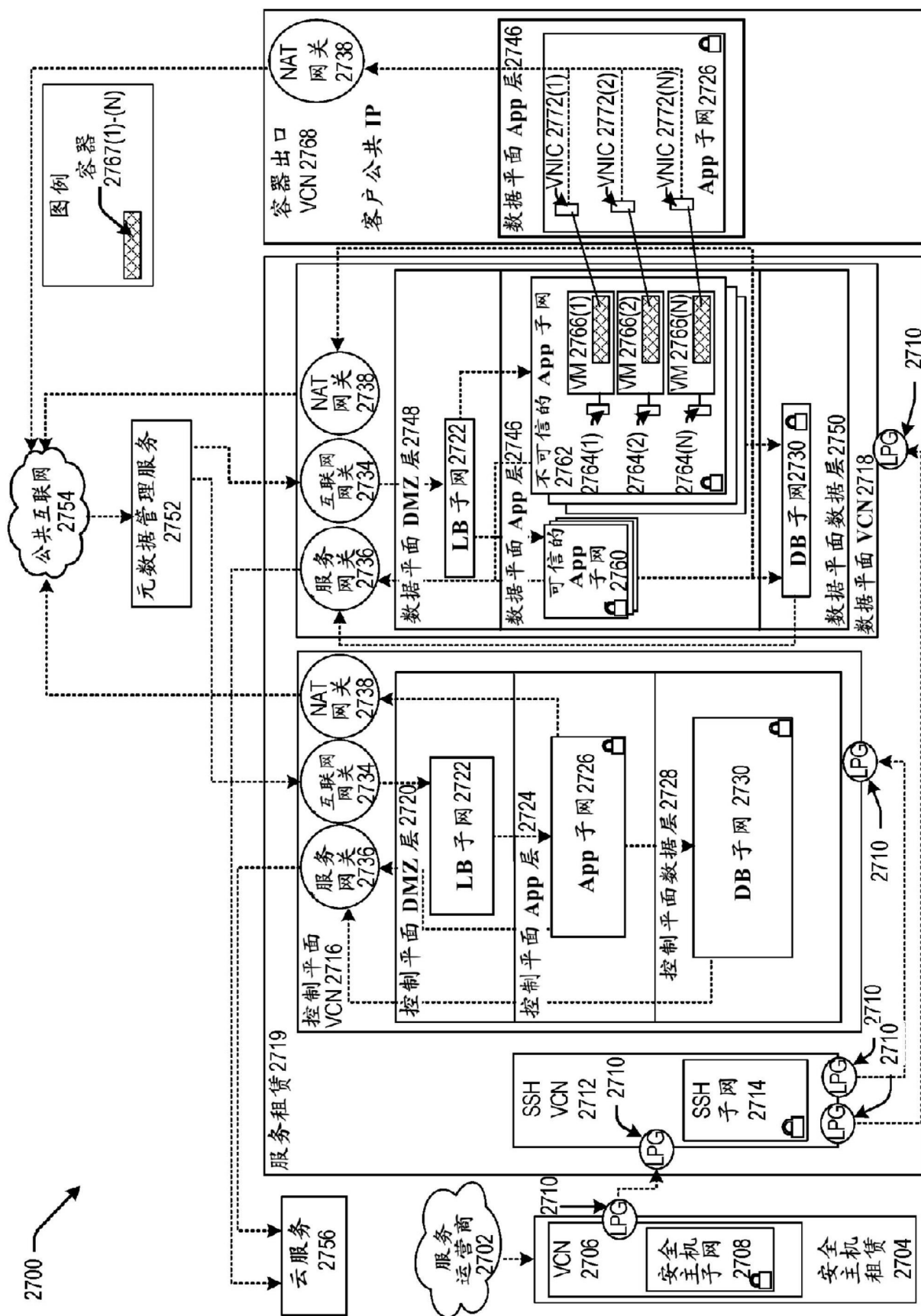


图27

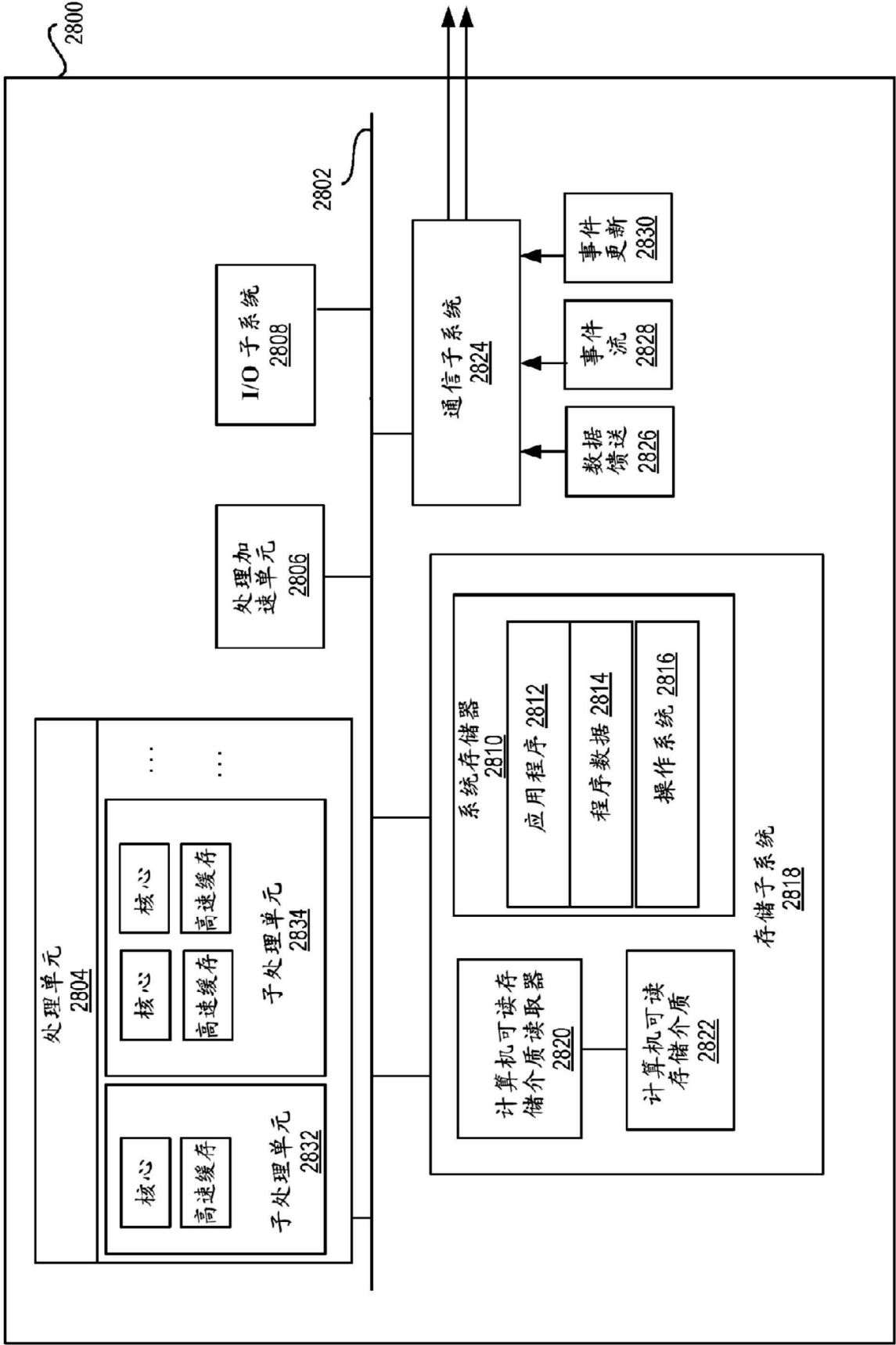


图28