(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117561705 A
(43)申请公布日 2024.02.13

(21)申请号 202280045223.2
(22)申请日 2022.06.23
(30)优先权数据
63/215,264 2021.06.25 US
17/734,833 2022.05.02 US
(85)PCT国际申请进入国家阶段日
2023.12.25
(86)PCT国际申请的申请数据
PCT/US2022/034788 2022.06.23
(87)PCT国际申请的公布数据
WO2022/271990 EN 2022.12.29
(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 D·D·贝克尔 L·苏莱克 J·R·尤克尔
(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 45/24(2006.01)

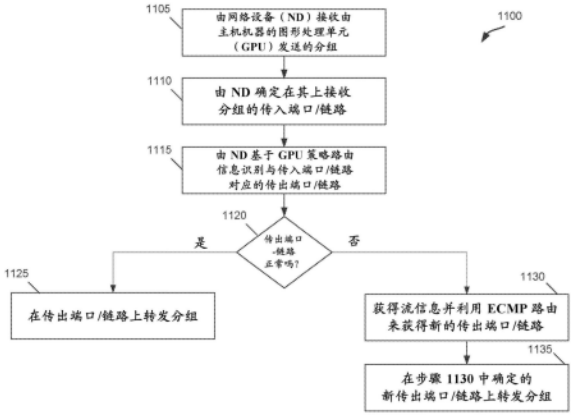
权利要求书2页 说明书39页 附图16页

(54)发明名称

用于图形处理单元的路由策略

(57)摘要

本文讨论的是用于托管在云环境中的几个主机机器上的图形处理单元(GPU)的路由机制。对于由主机机器的GPU传输并由网络设备接收的分组,网络设备确定在其上接收到分组的网络设备的传入端口链路。网络设备基于GPU路由策略识别与传入端口-链路对应的网络设备的传出端口-链路。GPU路由策略在接收分组之前预先配置并建立网络设备的每个传入端口-链路到网络设备的唯一传出端口-链路的映射。在网络设备的传出端口-链路上转发分组。



CN 117561705 A

1.一种方法,包括:

对于由主机机器的图形处理单元(GPU)传输并由网络设备接收的分组,

由网络设备确定在其上接收到所述分组的网络设备的传入端口-链路;

由网络设备基于GPU路由策略识别与所述传入端口-链路对应的传出端口-链路,其中所述GPU路由策略是在接收所述分组之前预先配置的并建立网络设备的每个传入端口-链路到网络设备的唯一传出端口-链路的映射;以及

由网络设备在网络设备的所述传出端口-链路上转发所述分组。

2.如权利要求1所述的方法,其中转发步骤还包括:

由网络设备核实与网络设备的所述传出端口-链路相关联的条件;以及

响应于满足所述条件,在网络设备的所述传出端口-链路上转发分组。

3.如权利要求2所述的方法,其中所述条件与确定网络设备的传出端口-链路是否处于活动状态对应。

4.如权利要求2所述的方法,还包括:

响应于不满足所述条件

由网络设备获得与所述分组相关联的流信息;

由网络设备基于所述流信息执行等价多路径算法,以获得网络设备的新的传出端口-链路;以及

由网络设备在网络设备的所述新的传出端口-链路上转发分组。

5.如权利要求4所述的方法,其中与所述分组相关联的流信息至少包括与源端口、目的地端口、源IP地址、目的地IP地址相关的信息。

6.如权利要求1所述的方法,其中网络设备是架顶式TOR交换机。

7.如权利要求6所述的方法,其中TOR和主机机器包括在机架中,所述主机机器经由多个链路与所述TOR通信地耦合,每个链路与网络接口卡相关联。

8.如权利要求6所述的方法,其中网络设备的所述传入端口-链路是将所述主机机器连接到所述TOR的第一链路,并且网络设备的所述传出端口-链路是将所述TOR连接到主干交换机的第二链路。

9.如权利要求1所述的方法,其中所述分组属于GPU工作负载。

10.一种网络设备,包括:

处理器;以及

存储器,包括指令,所述指令在利用处理器执行时使所述网络设备至少:

对于由主机机器的图形处理单元(GPU)传输并由网络设备接收的分组,

确定在其上接收到所述分组的网络设备的传入端口-链路;

基于GPU路由策略识别与所述传入端口-链路对应的传出端口-链路,其中所述GPU路由策略是在接收所述分组之前预先配置的并建立网络设备的每个传入端口-链路到网络设备的唯一传出端口-链路的映射;以及

在网络设备的所述传出端口-链路上转发所述分组。

11.如权利要求10所述的网络设备,还被配置为:

核实与网络设备的所述传出端口-链路相关联的条件;以及

响应于满足所述条件,在网络设备的所述传出端口-链路上转发所述分组。

12.如权利要求11所述的网络设备,其中所述条件与确定网络设备的传出端口-链路是否处于活动状态对应。

13.如权利要求11所述的网络设备,还被配置为:

响应于不满足所述条件

获得与所述分组相关联的流信息;

基于所述流信息执行等价多路径算法,以获得网络设备的新的传出端口-链路;以及

在网络设备的所述新的传出端口-链路上转发分组。

14.如权利要求13所述的网络设备,其中与所述分组相关联的流信息至少包括与源端口、目的地端口、源IP地址、目的地IP地址相关的信息。

15.如权利要求10所述的网络设备,其中网络设备是架顶式TOR交换机。

16.如权利要求15所述的网络设备,其中所述TOR和所述主机机器包括在机架中,所述主机机器经由多个链路与所述TOR通信地耦合,每个链路与网络接口卡相关联。

17.如权利要求15所述的网络设备,其中网络设备的所述传入端口-链路是将所述主机机器连接到所述TOR的第一链路,并且网络设备的所述传出端口-链路是将所述TOR连接到主干交换机的第二链路。

18.如权利要求10所述的网络设备,其中所述分组属于GPU工作负载。

19.一种存储特定计算机可执行指令的非暂态计算机可读介质,指令在由处理器执行时使计算机系统执行包括以下的操作:

对于由主机机器的图形处理单元(GPU)传输并由网络设备接收的分组,

由网络设备确定在其上接收到所述分组的网络设备的传入端口-链路;

由网络设备基于GPU路由策略识别与所述传入端口-链路对应的传出端口-链路,其中所述GPU路由策略是在接收所述分组之前预先配置的并建立网络设备的每个传入端口-链路到网络设备的唯一传出端口-链路的映射;以及

由网络设备在网络设备的所述传出端口-链路上转发分组。

20.如权利要求19所述的存储特定计算机可执行指令的非暂态计算机可读介质,其中转发步骤还包括:

由网络设备核实与网络设备的所述传出端口-链路相关联的条件;以及

响应于满足所述条件,在网络设备的所述传出端口-链路上转发所述分组。

用于图形处理单元的路由策略

[0001] 相关申请的交叉引用

[0002] 本申请根据35U.S.C.119(e)要求于2021年6月25日提交的美国临时申请No.63/215,264和2022年5月2日提交的美国非临时申请No.17/734,833的利益和优先权。出于所有目的,前述申请的全部内容通过引用整体并入本文。

技术领域

[0003] 本公开涉及用于托管在云环境中的几个主机机器上的图形处理单元(GPU)的框架和路由机制。

背景技术

[0004] 组织不断将业务应用和数据库迁移到云端,以降低购买、更新和维护内部部署的硬件和软件的成本。高性能计算(HPC)应用始终消耗100%的可用计算能力来实现特定成果或结果。HPC应用要求专用网络性能、快速存储、高计算能力和大量存储器,而这些资源在构成当今商品云的虚拟化基础设施中是短缺的。

[0005] 云基础设施服务提供商提供更新更快的CPU和图形处理单元(GPU)来解决HPC应用的要求。通常,构造虚拟拓扑是为了供给托管在几个主机机器上的各种GPU以便彼此通信。在实践中,使用环形拓扑来连接各个GPU。但是,环形网络本质上总是阻塞的。因此,整个系统的性能降级。本文讨论的实施例解决了与跨越几个主机机器的GPU的连接性相关的这些和其它问题。

发明内容

[0006] 本公开一般而言涉及用于托管在云环境中的几个主机机器上的图形处理单元(GPU)的路由机制。本文描述了各种实施例,包括方法、系统、存储可由一个或多个处理器执行的程序、代码或指令的非暂态计算机可读存储介质等。提及这些说明性实施例并不是为了限制或定义本公开,而是为了提供示例以帮助其理解。在详细描述部分中讨论了附加的实施例,并且其中提供了进一步的描述。

[0007] 本公开的一个实施例针对一种方法,包括:对于由主机机器的图形处理单元(GPU)传输并由网络设备接收的分组,由网络设备确定在其上接收分组的网络设备的传入端口-链路;由网络设备基于GPU路由策略识别与传入端口-链路对应的传出端口-链路,其中GPU路由策略是在接收分组之前预先配置的并且建立网络设备的每个传入端口-链路到网络设备的唯一传出端口-链路的映射;以及由网络设备在网络设备的传出端口-链路上转发分组。

[0008] 本公开的一方面提供了一种系统,该系统包括一个或多个数据处理器以及包含指令的非暂态计算机可读存储介质,指令在一个或多个数据处理器上执行时,使得该一个或多个数据处理器执行本文公开的一种或多种方法的部分或全部。

[0009] 本公开的另一方面提供了一种有形地实施在非暂态机器可读存储介质中的计算

机程序产品,包括被配置为使一个或多个数据处理器执行本文公开的一种或多种方法的部分或全部的指令。

[0010] 当参考以下说明书、权利要求书和附图时,前述以及其它特征和实施例将变得更加明显。

附图说明

[0011] 当参考附图阅读以下具体实施方式时,可以更好地理解本公开的特征、实施例和优点。

[0012] 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。

[0013] 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。

[0014] 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)。

[0015] 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)。

[0016] 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。

[0017] 图6描绘了根据某些实施例的结合CLOS网络布置的云基础设施的简化框图。

[0018] 图7描绘了根据某些实施例的描绘图6的云基础设施中的流冲突的示例性场景。

[0019] 图8描绘了根据某些实施例的在云基础设施中实现的基于策略的路由机制。

[0020] 图9描绘了根据某些实施例的云基础设施的框图,其图示了云基础设施中的不同类型的连接。

[0021] 图10图示了根据某些实施例的包括在云基础设施中的机架的示例性构造。

[0022] 图11A图示了描绘根据某些实施例的由网络设备在路由分组时执行的步骤的流程图。

[0023] 图11B图示了描绘根据某些实施例的由网络设备在路由分组时执行的步骤的另一个流程图。

[0024] 图12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。

[0025] 图13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。

[0026] 图14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。

[0027] 图15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。

[0028] 图16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

[0029] 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。

[0030] 云基础设施的示例体系架构

[0031] 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CSP的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。

[0032] 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。

[0033] 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。

[0034] 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。

[0035] 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。虚拟或覆盖网络可以包括一个或多个虚拟云网络(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,SMARTNIC)执行的功能、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。

[0036] 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、负载平衡和聚类等)。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平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。

[0037] 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。

[0038] 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。

[0039] 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。

[0040] 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虚拟网络中的路由。

[0041] 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)与虚拟网络中的实体相关联。物理IP地址和覆盖IP地址都是真实IP地

址的类型。它们与虚拟IP地址分离,虚拟IP地址映射到多个真实IP地址。虚拟IP地址提供了虚拟IP地址与多个真实IP地址之间的一对多映射。

[0042] 云基础设施或CSPI在物理上托管在世界上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)、负载平衡资源、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。

[0043] 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。

[0044] 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。

[0045] 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。

[0046] 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。

[0047] 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)。

[0048] 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。

[0049] 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、负载平衡器、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。

[0050] 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该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。

[0051] 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。

[0052] 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。

[0053] 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。

[0054] 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。

[0055] 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。

[0056] 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。

[0057] 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。

[0058] 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。

[0059] 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地址与和子网相关联的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。

[0060] 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。

[0061] 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性组或安全性列表来实现。网络安全性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。

[0062] 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内的子网和相关联的信息、与VCN相关联的一个或多个VR、VCN中的计算实例和相关联的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。

[0063] 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12、13、14和15中描绘(参见标号1216、1316、1416和1516)并在下面进行描述。

[0064] 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。

[0065] 图1、2、3、4、5、12、13、14和15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。

[0066] 如图1中描绘的示例中所示,分布式环境100包括提供服务和资源的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、负载平衡器等。

[0067] 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。

[0068] 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。

[0069] 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。

[0070] VCN A 104还可以包括一个或多个负载平衡器。例如,可以为子网提供负载平衡器,并且可以将负载平衡器配置为跨子网上的多个计算实例负载平衡流量。还可以提供负载平衡器以在VCN中跨子网负载平衡流量。

[0071] 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。

[0072] 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。

[0073] 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VCN VR以及VNIC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。

[0074] 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。

[0075] 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。

[0076] 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12、13、14和15中描绘(例如,通过附图标记1234、1236、1238、1334、1336、1338、1434、1436、1438、1534、1536和1538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。

[0077] 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MicrosoftAzure云、亚马逊AWS云等)通信。

[0078] 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 1120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。

[0079] 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。

[0080] 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VPN Connect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。

[0081] 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。

[0082] 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具有用于它建立的每个对等体的单独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。

[0083] 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。

[0084] 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)。

[0085] 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)。

[0086] 通过允许流量流经FastConnect/IPsec链路和客户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。

[0087] 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。

[0088] 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。

[0089] 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路来提供私有访问模型。

[0090] 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。

[0091] 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。

[0092] 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。

[0093] 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。

[0094] 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。

[0095] 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。

[0096] 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。

[0097] 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促进分组或帧往返于计算实例的通信。VNIC在创建计算实例时与计算实例相关联。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。

[0098] 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。

[0099] 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)。

[0100] 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。

[0101] 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。

[0102] 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。

[0103] 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。

[0104] 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。

[0105] 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。

[0106] 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。

[0107] 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。

[0108] 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。

[0109] 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。

[0110] 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。

[0111] 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。

[0112] 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。

[0113] 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cis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。

[0114] 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组管线,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。

[0115] 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12、13、14和15中描绘(参见标号1216、1316、1416和1516)并在下文中描述。VCN数据平面的示例在图12、13、14和15中描绘(参见标号1218、1318、1418和1518)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。

[0116] 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)。

[0117] 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。

[0118] 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。

[0119] 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。

[0120] 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。

[0121] 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。

[0122] 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对应的功能性(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标签。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)。

[0123] 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。

[0124] 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。

[0125] 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)。

[0126] 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VM2408。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。

[0127] 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 A416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。

[0128] 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。

[0129] 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为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缩放(scale)。

[0130] 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,更多叶交换机和主干交换机)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。

[0131] 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:

[0132] ocid1.<RESOURCE TYPE>.<REALM>.[REGION][.FUTURE USE].<UNIQUE ID>

[0133] 其中,

[0134] ocid1:指示CID的版本的文字串;

[0135] resource type:资源的类型(例如,实例、卷、VCN、子网、用户、组等);

[0136] 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;

[0137] 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;

[0138] future use:保留以备将来使用。

[0139] uique ID:ID的唯一部分。格式可以因资源或服务的类型而异。

[0140] 图6描绘了根据某些实施例的结合CLOS网络布置的云基础设施600的框图。云基础设施600包括多个机架(例如,机架1 610…机架M 620)。每个机架包括多个主机机器(本文也称为主机)。例如,机架1 610包括多个主机(例如,K个主机机器)主机1-A、612至主机1-K 614,并且机架M包括K个主机机器,即,主机M-A 622至主机M-K 624。应该认识到的是,图6的图示(即,每个机架包括相同数量的主机机器,例如K个主机机器)旨在是说明性和非限制性的。例如,与机架1 610中包括的主机机器的数量相比,机架M 620可以具有更多或更少数量的主机机器。

[0141] 每个主机机器包括多个图形处理单元(GPU)。例如,如图6中所示,主机机器1-A 612包括N个GPU,例如,GPU 1 613。而且,应该认识到的是,图6中的图示(使每个主机机器包括相同数量的GPU,即N个GPU)旨在是说明性和非限制性的,即每个主机机器可以包括不同数量的GPU。每个机架都包括架顶(TOR)交换机,该交换机与主机机器上托管的GPU通信地耦合。例如,机架1 610包括通信地耦合到主机1-A 612和主机1-K 614的TOR交换机(即,TOR 1)616,并且机架M 620包括通信地耦合到主机M-A 622和主机M-K 624的TOR交换机(即,TOR M)626。应该认识到的是,图6中描绘的TOR交换机(即,TOR 1 616和TOR M 626)各自包括N个端口,这些端口被用于将TOR交换机通信地耦合到包括在机架中的每个主机机器上托管的N个GPU。如图6中所描绘的TOR交换机与GPU的耦合旨在是说明性而非限制性的。例如,在一些实施例中,TOR交换机可以具有多个端口,每个端口与每个主机机器上的GPU对应,即主机机器上的GPU可以经由通信链路连接到TOR的唯一端口。而且,由网络设备(例如,TOR 1 616)接收的流量在本文被表征为在网络设备的特定传入端口-链路上接收的流量。例如,当主机1-A 612的GPU 1 613将数据分组传输到TOR 1 616(使用链路617)时,在TOR 1交换机的端口619上接收该数据分组。TOR 1 616将这个数据分组表征为在TOR的第一传入端口-链路上接收到的信息。应该认识到的是,类似的概念可以应用于TOR的所有传出端口-链路。

[0142] 来自每个机架的TOR交换机通信地耦合到多个主干交换机,例如主干交换机1 630和主干交换机P 640。如图6中所示,TOR 1 616分别经由两个链路连接到主干交换机1 630,并经由另外两个链路连接到主干交换机P 640。从特定TOR交换机传输到主干交换机的信息在本文中被称为经由上行链路进行的通信,而从主干交换机传输到TOR交换机的信息在本文中被称为经由下行链路进行的通信。根据一些实施例,TOR交换机和主干交换机在CLOS网络布置(例如,多级交换网络)中连接,其中每个TOR交换机形成CLOS网络中的"叶"节点。

[0143] 根据一些实施例,主机机器中包括的GPU执行与机器学习相关的任务。在这种设置中,单个任务可以跨大量GPU(例如,64个)执行/分布,而这些GPU进而可以跨多个主机机器并跨多个机架分布。由于所有这些GPU都在处理相同的任务(即,工作负载),因此它们都需要以时间同步的方式彼此通信。此外,在任何给定时刻,GPU处于计算模式或通信模式之一,即GPU在大致相同的时刻彼此通信。工作负载的速度由最慢GPU的速度确定。

[0144] 通常,为了将分组从源GPU(例如,主机1-A 612的GPU 1 613)路由到目的地GPU(例如,主机M-A 622的GPU 1 623),利用等价多路径(ECMP)路由。在ECMP路由中,当存在可用于将流量从发送方路由到接收方的多条等价路径时,使用选择技术来选择特定路径。因而,在接收流量的网络设备(例如,TOR交换机或主干交换机)处,使用选择算法来选择传出链路,用于将流量从下一跳的网络设备转发到另一个网络设备。这种传出链路选择发生在从发送方到接收方的路径中的每个网络设备处。基于散列的选择算法是广泛使用的ECMP选择技术,其中散列可以基于分组的4元组(例如,源端口、目的地端口、源IP、目的地IP)。

[0145] ECMP路由是一种流感知路由技术,其中每个流(即,数据分组的流)在流的持续时间内被散列到特定路径。因此,流中的分组是使用特定传出端口/链路从网络设备转发的。这样做通常是为了确保流中的分组按次序到达,即不要求对分组的重新排序。但是,ECMP路由不知道带宽(或吞吐量)。换句话说,TOR和主干交换机对并行链路上的流执行统计流感知(吞吐量不感知)ECMP负载平衡。

[0146] 在标准ECMP路由(即仅流感知路由)中,问题在于由网络设备通过两个单独的传入链路接收的流可以被散列到同一传出链路,从而导致流冲突。例如,考虑这样一种情况:两个流通过两个单独的传入100G链路进入,并且每个流都被散列到同一100G传出链路。这种情况导致拥塞(即流冲突)并导致分组被丢弃,因为传入带宽是200G,但传出带宽是100G。例如,下面描述的图7图示了流冲突700的示例性场景。

[0147] 如图7中所示,存在两个流:流1 710,其从主机机器1-A 612的第一GPU指向TOR交换机616(以实线描绘);以及流2 720,其从同一主机机器612上的另一个GPU指向TOR交换机616(由虚线描绘)。注意的是,这两个流在单独的链路(即,TOR的单独的传入端口-链路)上被指向TOR交换机616。假设图7中描绘的所有链路都具有100G的容量(即,带宽)。在TOR交换机616执行ECMP路由算法的情况下,两个流可能被散列以使用TOR的同一传出端口-链路,例如连接到链路730的TOR端口,链路730将TOR交换机616连接到主干交换机630。在这种情况下,两个流之间存在冲突(由"X"标记表示),这导致分组被丢弃。

[0148] 无论协议如何,这种冲突场景对于所有类型的流量来说一般都是有问题的。例如,TCP是智能的,因为当分组被丢弃并且发送方没有得到该被丢弃分组的确认时,该分组被重传。但是,对于远程直接存储器访问(RDMA)类型的流量,情况变得更糟。RDMA网络出于多种原因不使用TCP(例如,TCP具有不利于低时延和高性能的复杂逻辑)。RDMA网络使用诸如

RDMA over Infiniband或RDMA over融合以太网(RoCE)之类的协议。在RoCE中,存在拥塞控制算法,其中当发送方识别出拥塞或被丢弃分组的发生时,发送方减慢分组的传输。对于被丢弃的分组,不仅被丢弃的分组,而且被丢弃的分组周围的几个分组都被重传,这进一步消耗了可用带宽并导致性能差。

[0149] 下面描述的是克服上述流冲突问题的技术。应该认识到的是,流冲突问题影响来自CPU和GPU的流量。但是,由于严格的时间同步要求,流冲突问题对于GPU来说是个更大的问题。此外,应认识到的是,标准ECMP路由机制将由于其以统计带宽不感知方式路由信息的固有特性而无论网络是订阅过量还是订阅不足都引发流冲突场景。没有订阅过量的网络是指设备(例如,TOR、主干交换机)的传入链路的带宽与传出链路的带宽相同的网络。注意的是,如果所有链路具有相同的带宽容量,那么传入链路的数量与传出链路的数量相同。

[0150] 根据一些实施例,克服上述流冲突问题的技术包括基于GPU的策略路由机制(本文也称为基于GPU的流量工程机制)和经修改的ECMP路由机制。下面将进一步详细描述这些技术中的每一种。

[0151] 图8描绘了根据某些实施例的在图6的云基础设施的网络设备中实现的基于策略的路由机制。具体而言,云基础设施800包括多个机架,例如机架1 810至机架M 820。每个机架包括主机机器,该主机机器包括多个GPU。例如,机架1 810包括主机机器,即主机1-A812,并且机架M 820包括主机机器,即主机M-A 822。每个机架包括TOR交换机,例如机架1 810包括TOR 1交换机814并且机架M820包括TOR M交换机824。每个机架中的主机机器通信地耦合到机架中的相应TOR交换机。TOR交换机,即TOR交换机1 814和TOR交换机M 824进而通信地耦合到主干交换机,即主干交换机830和主干交换机840。为了描述基于策略的路由并且为了说明起见,云基础设施800被描绘为每个机架包括单个主机。但是,应认识到的是,基础设施中的每个机架可以具有多于一台主机机器。

[0152] 根据一些实施例,从发送方到接收方的数据分组在网络中逐跳地路由。在将传入端口-链路绑定到传出端口-链路的每个网络设备处配置路由策略。网络设备可以是TOR交换机或主干交换机。参考图8,描绘了两个流:来自主机机器812上的GPU 1的流1,其预期目的地是主机机器822上的GPU 1,以及来自主机机器812上的GPU N的流2,其预期目的地是主机机器822上的GPU N。网络设备,即TOR 1 814、主干1 830、TOR M 824和主干P 840,被配置为将传入端口-链路绑定(或匹配)到传出端口-链路。在每个网络设备处维护传入端口-链路与传出端口-链路的匹配(例如,在策略表中)。

[0153] 参考图8,可以观察到,关于流1(即由实线描绘的流),当TOR 1 814在链路850上接收分组时,TOR被配置为在传出链路855上转发接收到的分组。类似地,当主干交换机830经由链路855接收分组时,它被配置为在传出链路860上转发分组。最终,当TOR M824在链路860上接收分组时,它被配置为在传出链路865上转发该分组,以便将该分组传输到其预期目的地,即主机机器822上的GPU 1。类似地,关于流2(即由虚线描绘的流),当TOR 1 814在链路870上接收分组时,TOR 1被配置为在传出链路875上将接收到的分组转发到主干P 840。当主干交换机840经由链路875接收分组时,它被配置为在传出链路880上转发分组。最终,当TOR M824在链路880上接收分组时,它被配置为在传出链路885上转发该分组,以便将该分组传输到其预期目的地,即主机机器822上的GPU N。

[0154] 以这种方式,根据基于GPU策略的路由机制,每个网络设备被配置为将传入端口/

链路绑定到传出端口/链路以避免冲突。例如,考虑流1和流2的第一跳流,TOR 1 814在链路850上接收第一数据分组(与流1对应),并在链路870上接收第二数据分组(与流2对应)。由于TOR 1 814被配置为将在传入链路/端口850上接收到的数据分组转发到传出链路/端口855,并且将在传入链路/端口870上接收到的数据分组转发到传出链路/端口875,因此确保了第一分组和第二分组不会冲突。应认识到的是,在云基础设施内的每个网络设备处,传入端口-链路与传出端口-链路之间存在1-1对应关系,即独立于流和/或流执行的协议执行传入端口-链路与传出端口-链路之间的映射。此外,根据一些实施例,在特定网络设备的传出链路发生故障的情况下,网络设备被配置为将其路由策略从基于GPU策略的路由切换到标准ECMP路由以获得新的可用输出链路(来自各种可用的输出链路)并将流发送到新的输出链路。要注意的是,在这种情况下,会发生流冲突,从而导致拥塞。

[0155] 现在转向图9,描绘了根据某些实施例的图示不同类型的连接的云基础设施900的框图。基础设施900包括多个机架,例如机架1910、机架D 920和机架M 930。机架910和930包括多个主机机器。例如,机架1 910包括多个主机(例如,K个主机机器)主机1-A912至主机1-K 914,并且机架M包括K个主机机器,即主机M-A932至主机M-K 934。机架D 920包括一个或多个主机机器922,每个主机机器922包括多个CPU,即主机机器922是非GPU主机机器。机架910、920和930中的每一个包括TOR交换机,即分别通信地耦合到相应机架中的主机机器的TOR 1、916、TOR D 926和TOR M 936。另外,TOR交换机916、926和936通信地耦合到多个主干交换机,即主干交换机940和950。

[0156] 如图9中所示,存在从GPU主机(即主机1-A 912)到另一个GPU主机(即主机M-A 932)的第一连接(即,由虚线描绘的连接1)以及从GPU主机(即主机1-K 914)到非GPU主机(即主机D922)的第二连接(即,由点线描绘的连接2)。关于连接1,通过基于基于GPU的策略路由机制配置每个中间网络设备,对与流相关联的数据分组进行逐跳路由,如图8中先前描述的。具体而言,连接1的数据分组沿着图9中描绘的虚线链路进行路由,即从主机1-A到TOR 1,从TOR 1到主干交换机1,从主干交换机1到TOR M,并最终从TOR M到主机M-A。每个网络设备被配置为将传入链路-端口绑定到网络设备的传出链路-端口。

[0157] 相比之下,连接2是从基于GPU的主机(即主机1-K 914)到非基于GPU的主机(即,主机D 922)的。在这种情况下,TOR 1916被配置为将传入链路-端口即端口和链路971(在其上从基于GPU的主机接收数据分组的)绑定到传出链路-端口即输出端口和和连接到输出端口的链路972。因此,TOR 1使用传出链路972将数据分组转发到主干交换机1 940。由于数据分组的目的地是非基于GPU的主机,因此主干交换机1 940不利用基于策略的路由机制将分组转发到TOR D。更确切地说,主干交换机1 940利用ECMP路由机制来选择可用链路980之一以将分组转发到TOR D,TOR D然后将数据分组转发到主机D 922。

[0158] 在一些实施例中,通过实现ECMP路由的修改版本,本文描述的云基础设施的网络设备避免了先前参考图7描述的流冲突。在这种情况下,ECMP散列算法被修改,使得来自网络设备的特定传入端口-链路的流量被散列(并指向)到网络设备的同一传出端口-链路。具体而言,每个网络设备实现经修改的ECMP算法以确定要用于转发分组的输出端口-链路,其中根据经修改的ECMP算法,通过特定传入端口-链路接收的任何分组总是散列到同一传出端口-链路。例如,考虑以下情况:在网络设备的第一传入端口-链路上接收第一分组和第二分组,而在网络设备的第二传入端口-链路上接收第三分组和第四分组。在这种情况下,网

络设备被配置为实现经修改的ECMP算法,其中网络设备在网络设备的第一传出端口-链路上传输第一分组和第二分组,并且在网络设备的第二传出端口-链路上传输第三分组和第四分组,其中网络设备的第一入端口-链路与网络设备的第二入端口-链路不同,并且网络设备的第一传出端口-链路与网络设备的第二输出端口-链路不同。在某些实施方式中,可以修改存储在转发信息数据库中的信息(例如,转发表、ECMP表等)以启用上述功能性。

[0159] 现在转向图10,图示了根据某些实施例的机架1000的示例性构造。如图10中所示,机架1000包括两个主机机器,即主机机器1010和主机机器1020。应认识到的是,虽然机架1000被描绘为包括两个主机机器,但是机架1000可以包括更多数量的主机机器。每个主机机器包括多个GPU和多个CPU。例如,主机机器1010包括多个CPU 1012和多个GPU 1014,而主机机器1020包括多个CPU 1022和多个GPU 1024。

[0160] 主机机器经由不同的TOR交换机通信地耦合到不同的网络架构。例如,主机机器1010和1020经由TOR交换机(即,TOR 1交换机1050)通信地耦合到网络架构(本文中称为机架1000的前端网络)。前端网络可以与外部网络对应。更具体而言,主机机器1010经由网络接口卡(NIC)1030和耦合到TOR 1交换机1050的网络虚拟化设备(NVD)1035连接到前端网络。主机机器1020经由NIC 1040和耦合到TOR 1交换机1050的NVD 1045连接到前端网络。因此,通过一些实施例,每个主机机器的CPU可以经由NIC、NVD和TOR交换机与前端网络通信。例如,主机机器1010的CPU 1012可以经由NIC 1030、NVD 1035和TOR 1交换机1050与前端网络通信。

[0161] 主机机器1010和1020连接到另一侧的启用服务质量(QoS)的后端网络。启用(QoS)的后端网络在本文中被称为与如图6中所示的GPU集群网络对应的后端网络。主机机器1010经由另一个NIC 1065连接到将主机机器1010通信地耦合到后端网络的TOR 2交换机1060。类似地,主机机器1020经由NIC 1080连接到将主机机器1020通信地耦合到后端网络的TOR 2交换机1060。因此,每个主机机器的多个GPU可以经由NIC和TOR交换机与后端网络通信。以这种方式,多个CPU利用单独的NIC集合(与GPU利用的NIC集合相比)以便分别与前端网络进和后端网络进行通信。

[0162] 图11A描绘了根据某些实施例的示例性流程图1100,其描绘了由网络设备在路由分组时执行的步骤。图11A中描绘的处理可以以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)、硬件或其组合来实现。软件可以存储在非暂态存储介质上(例如,存储器设备上)。图11A中呈现的和下面描述的方法旨在是说明性和非限制性的。虽然图11A描绘了以特定顺序或次序发生的各种处理步骤,但这并不旨在是限制性的。在某些替代实施例中,这些步骤可以以某种不同的次序执行,或者一些步骤也可以并行执行。

[0163] 该过程开始于步骤1105,其中网络设备接收由主机机器的图形处理单元(GPU)传输的分组。在步骤1110中,网络设备确定在其上接收分组的传入端口/链路。在步骤1115中,网络设备基于策略路由信息来识别与传入端口/链路(在其上接收到分组)对应的传出端口/链路。根据一些实施例,策略路由信息与用于网络设备的预先配置的GPU路由表对应,该预先配置的GPU路由表将网络设备的每个传入端口-链路绑定到网络设备的唯一传出链路端口。

[0164] 该过程然后移至步骤1120,其中执行查询以确定传出端口-链路是否处于运作状

态,例如传出链路是否处于活动状态。如果对查询的响应是肯定的(即链路处于活动状态),那么该过程移至步骤1125,否则,如果对查询的响应是否定的(即链路处于故障状态/非活动状态),那么过程移至步骤1130。在步骤1125中,网络设备利用传出端口-链路(在步骤1115中识别出的)将接收到的数据分组转发到另一个网络设备。在步骤1130中,网络设备获得数据分组的流信息,例如流信息可以与和分组相关联的四元组(即,源端口、目的地端口、源IP地址、目的地IP地址)对应。基于获得的流信息,网络设备利用ECMP路由来识别新的传出端口-链路,即可用的传出端口-链路。该过程然后移至步骤1135,其中网络设备利用新获得的传出端口-链路来转发在步骤1105中接收到的数据分组。

[0165] 图11B描绘了根据某些实施例的另一个示例性流程图1150,其描绘了由网络设备在路由分组时执行的步骤。图11B中描绘的处理可以以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)、硬件或其组合来实现。软件可以存储在非暂态存储介质上(例如,存储器设备上)。图11B中呈现的和下面描述的方法旨在是说明性和非限制性的。虽然图11B描绘了以特定顺序或次序发生的各种处理步骤,但这并不旨在是限制性的。在某些替代实施例中,这些步骤可以以某种不同的次序执行,或者一些步骤也可以并行执行。

[0166] 该过程开始于步骤1155,其中网络设备接收由主机机器的图形处理单元(GPU)传输的分组。在步骤1160中,网络设备确定接收到的分组的流信息。在一些实施方式中,流信息可以与和分组相关联的4元组(即,源端口、目的地端口、源IP地址、目的地IP地址)对应。在步骤1165中,网络设备通过实现ECMP路由的修改版本来计算传出端口-链路。根据经修改的ECMP算法,通过特定传入端口-链路接收的任何分组始终经过散列处理,以便在同一传出端口-链路上被传输。

[0167] 该过程然后移至步骤1170,其中执行查询以确定传出端口-链路(在步骤1165中确定的)是否处于运作状态,例如传出链路处于活动状态。如果对查询的响应是肯定的(即链路处于活动状态),那么过程移至步骤1175,否则,如果对查询的响应是否定的(即链路处于故障状态/非活动状态),那么过程移至步骤1175。该过程移至步骤1180。在步骤1175中,网络设备利用传出端口-链路(在步骤1165中识别出的)将接收到的数据分组转发到另一个网络设备。在(步骤1165的)识别出的输出端口-链路被确定为处于非活动状态的情况下,过程移至步骤1180。在步骤1180中,网络设备实现ECMP路由(即,标准ECMP路由)以识别新的传出端口-链路。该过程然后移至步骤1185,其中网络设备利用新计算的输出端口-链路来转发在步骤1155中接收到的数据分组。

[0168] 要注意的是,上述路由源自主机机器GPU的数据分组的技术导致对于较小集群的吞吐量增加20%,而对于较大集群的吞吐量增加70%(即,比标准ECMP路由算法提高3倍)。

[0169] 示例云基础设施实施例

[0170] 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、安全性、负载平衡和聚类等)。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平衡,以维持应用的可用性和

性能。

[0171] 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。

[0172] 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。

[0173] 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)。

[0174] 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。

[0175] 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。

[0176] 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安全性组规则以定义如何设置网络的安全性以及一个或多个虚拟机(VM)。也可以供给其它基础设施元素,诸如负载平衡器、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。

[0177] 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。

[0178] 图12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1200。服务运营商1202可以通信地耦合到可以包括虚拟云网络(VCN)1206和安全主机子网1208的安全主机租赁1204。在一些示例中,服务运营商1202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 1206和/或互联网的网络进行通信的个人消息传递设备。

[0179] VCN 1206可以包括本地对等网关(LPG)1210,其可以经由包含在SSH VCN 1212中的LPG 1210通信地耦合到安全壳(SSH)VCN 1212。SSH VCN 1212可以包括SSH子网1214,并且SSH VCN 1212可以经由包含在控制平面VCN 1216中的LPG 1210通信地耦合到控制平面VCN 1216。此外,SSH VCN 1212可以经由LPG 1210通信地耦合到数据平面VCN 1218。控制平面VCN 1216和数据平面VCN 1218可以包含在可以由IaaS提供商拥有和/或操作的服务租赁1219中。

[0180] 控制平面VCN 1216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1220。基于DMZ的服务器可以承担有限责任并有助于控制安全性漏洞。此外,DMZ层1220可以包括一个或多个负载平衡器(LB)子网1222、可以包括(一个或多个)应用子网1226的控制平面应用层1224、可以包括(一个或多个)数据库(DB)子网1230(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1228。包含在控制平面DMZ层1220中的(一个或多个)LB子网1222可以通信地耦合到包含在控制平面应用层1224中的(一个或多个)应用子网1226和可以包含在控制平面VCN 1216中的互联网网关1234,并且(一个或多个)应用子网1226可以通信地耦合到包含在控制平面数据层1228中的(一个或多个)DB子网1230以及服务网关1236和网络地址转换(NAT)网关1238。控制平面VCN 1216可以包括服务网关1236和NAT网关1238。

[0181] 控制平面VCN 1216可以包括数据平面镜像应用层1240,其可以包括(一个或多个)应用子网1226。包含在数据平面镜像应用层1240中的(一个或多个)应用子网1226可以包括可以执行计算实例1244的虚拟网络接口控制器(VNIC)1242。计算实例1244可以将数据平面镜像应用层1240的(一个或多个)应用子网1226通信地耦合到可以包含在数据平面应用层1246中的(一个或多个)应用子网1226。

[0182] 数据平面VCN 1218可以包括数据平面应用层1246、数据平面DMZ层1248和数据平面数据层1250。数据平面DMZ层1248可以包括(一个或多个)LB子网1222,其可以通信地耦合到数据平面应用层1246的(一个或多个)应用子网1226和数据平面VCN 1218的互联网网关1234。(一个或多个)应用子网1226可以通信地耦合到数据平面VCN 1218的服务网关1236和数据平面VCN 1218的NAT网关1238。数据平面数据层1250还可以包括可以通信地耦合到数据平面应用层1246的(一个或多个)应用子网1226的(一个或多个)DB子网1230。

[0183] 控制平面VCN 1216和数据平面VCN 1218的互联网网关1234可以通信地耦合到元

数据管理服务1252,元数据管理服务1252可以通信地耦合到公共互联网1254。公共互联网1254可以通信地耦合到控制平面VCN 1216和数据平面VCN 1218的NAT网关1238。控制平面VCN 1216和数据平面VCN 1218的服务网关1236可以通信地耦合到云服务1256。

[0184] 在一些示例中,控制平面VCN 1216或数据平面VCN 1218的服务网关1236可以对云服务1256进行应用编程接口(API)调用,而无需通过公共互联网1254。从服务网关1236到云服务1256的API调用可以是单向的:服务网关1236可以对云服务1256进行API调用,并且云服务1256可以将请求的数据发送到服务网关1236。但是,云服务1256可以不发起对服务网关1236的API调用。

[0185] 在一些示例中,安全主机租赁1204可以直接连接到服务租赁1219,服务租赁1219否则可以被隔离。安全主机子网1208可以通过LPG 1210与SSH子网1214通信,LPG 1210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1208连接到SSH子网1214可以使安全主机子网1208访问服务租赁1219内的其它实体。

[0186] 控制平面VCN 1216可以允许服务租赁1219的用户设置或以其它方式供给期望的资源。在控制平面VCN 1216中供给的期望资源可以在数据平面VCN 1218中部署或以其它方式使用。在一些示例中,控制平面VCN 1216可以与数据平面VCN 1218隔离,并且控制平面VCN 1216的数据平面镜像应用层1240可以经由VNIC 1242与数据平面VCN 1218的数据平面应用层1246通信,VNIC 1242可以包含在数据平面镜像应用层1240和数据平面应用层1246中。

[0187] 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1252的公共互联网1254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1252可以通过互联网网关1234将请求传送到控制平面VCN 1216。请求可以被包含在控制平面DMZ层1220中的(一个或多个)LB子网1222接收。(一个或多个)LB子网1222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1222可以将请求传输到包含在控制平面应用层1224中的(一个或多个)应用子网1226。如果请求被验证并且需要对公共互联网1254的调用,那么对公共互联网1254的调用可以被传输到可以对公共互联网1254进行调用的NAT网关1238。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1230中。

[0188] 在一些示例中,数据平面镜像应用层1240可以促进控制平面VCN 1216和数据平面VCN 1218之间的直接通信。例如,可能期望对包含在数据平面VCN 1218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 1242,控制平面VCN 1216可以直接与包含在数据平面VCN 1218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。

[0189] 在一些实施例中,控制平面VCN 1216和数据平面VCN 1218可以包含在服务租赁1219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 1216或数据平面VCN 1218。替代地,IaaS提供商可以拥有或操作控制平面VCN 1216和数据平面VCN 1218,这两者平面都可以包含在服务租赁1219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1254进行存储。

[0190] 在其它实施例中,包含在控制平面VCN 1216中的(一个或多个)LB子网1222可以被配置为从服务网关1236接收信号。在这个实施例中,控制平面VCN 1216和数据平面VCN

1218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1254。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1219上,服务租赁1219可能与公共互联网1254隔离。

[0191] 图13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1300。服务运营商1302(例如,图12的服务运营商1202)可以通信地耦合到安全主机租赁1304(例如,图12的安全主机租赁1204),该安全主机租赁1304可以包括虚拟云网络(VCN)1306(例如,图12的VCN 1206)和安全主机子网1308(例如,图12的安全主机子网1208)。VCN 1306可以包括本地对等网关(LPG)1310(例如,图12的LPG 1210),其可以经由包含在SSH VCN 1312中的LPG 1210通信地耦合到安全壳(SSH)VCN 1312(例如,图12的SSH VCN 1212)。SSH VCN 1312可以包括SSH子网1314(例如,图12的SSH子网1214),并且SSH VCN 1312可以经由包含在控制平面VCN 1316中的LPG 1310通信地耦合到控制平面VCN 1316(例如,图12的控制平面VCN 1216)。控制平面VCN 1316可以包含在服务租赁1319(例如,图12的服务租赁1219)中,并且数据平面VCN 1318(例如,图12的数据平面VCN 1218)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1321中。

[0192] 控制平面VCN 1316可以包括控制平面DMZ层1320(例如,图12的控制平面DMZ层1220),其可以包括(一个或多个)LB子网1322(例如,图12的(一个或多个)LB子网1222)、可以包括(一个或多个)应用子网1326(例如,图12的(一个或多个)应用子网1226)的控制平面应用层1324(例如,图12的控制平面应用层1224)、可以包括(一个或多个)数据库(DB)子网1330(例如,类似于图12的(一个或多个)DB子网1230)的控制平面数据层1328(例如,图12的控制平面数据层1228)。包含在控制平面DMZ层1320中的(一个或多个)LB子网1322可以通信地耦合到包含在控制平面应用层1324中的(一个或多个)应用子网1326和可以包含在控制平面VCN 1316中的互联网网关1334(例如,图12的互联网网关1234),并且(一个或多个)应用子网1326可以通信地耦合到包含在控制平面数据层1328中的(一个或多个)DB子网1330以及服务网关1336(例如,图12的服务网关)和网络地址转换(NAT)网关1338(例如,图12的NAT网关1238)。控制平面VCN 1316可以包括服务网关1336和NAT网关1338。

[0193] 控制平面VCN 1316可以包括可以包括(一个或多个)应用子网1326的数据平面镜像应用层1340(例如,图12的数据平面镜像应用层1240)。包含在数据平面镜像应用层1340中的(一个或多个)应用子网1326可以包括可以执行计算实例1344(例如,类似于图12的计算实例1244)的虚拟网络接口控制器(VNIC)1342(例如,1242的VNIC)。计算实例1344可以促进数据平面镜像应用层1340的(一个或多个)应用子网1326和可以包含在数据平面应用层1346(例如,图12的数据平面应用层1246)中的(一个或多个)应用子网1326之间经由包含在数据平面镜像应用层1340中的VNIC 1342和包含在数据平面应用层1346中的VNIC 1342的通信。

[0194] 包含在控制平面VCN 1316中的互联网网关1334可以通信地耦合到元数据管理服务1352(例如,图12的元数据管理服务1252),该元数据管理服务1352可以通信地耦合到公共互联网1354(例如,图12的公共互联网1254)。公共互联网1354可以通信地耦合到包含在控制平面VCN 1316中的NAT网关1338。包含在控制平面VCN 1316中的服务网关1336可以通信地耦合到云服务1356(例如,图12的云服务1256)。

[0195] 在一些示例中,数据平面VCN 1318可以包含在客户租赁1321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 1316,并且IaaS提供商可以为每个客户设置包含在服务租赁1319中的唯一计算实例1344。每个计算实例1344可以允许包含在服务租赁1319中的控制平面VCN 1316和包含在客户租赁1321中的数据平面VCN 1318之间的通信。计算实例1344可以允许在包含在服务租赁1319中的控制平面VCN 1316中供给的资源被部署或以其它方式在包含在客户租赁1321中的数据平面VCN 1318中使用。

[0196] 在其它示例中,IaaS提供商的客户可以具有存在于客户租赁1321中的数据库。在这个示例中,控制平面VCN 1316可以包括数据平面镜像应用层1340,其可以包括(一个或多个)应用子网1326。数据平面镜像应用层1340可以驻留在数据平面VCN 1318中,但数据平面镜像应用层1340可能不在数据平面VCN 1318中。即,数据平面镜像应用层1340可以访问客户租赁1321,但是数据平面镜像应用层1340可能不存在于数据平面VCN 1318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1340可以被配置为对数据平面VCN 1318进行调用,但可以不被配置为对包含在控制平面VCN 1316中的任何实体进行调用。客户可能期望在数据平面VCN 1318中部署或以其它方式使用在控制平面VCN 1316中供给的资源,并且数据平面镜像应用层1340可以促进客户的期望部署或资源的其它使用。

[0197] 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 1318。在这个实施例中,客户可以确定数据平面VCN 1318可以访问什么,并且客户可以限制从数据平面VCN 1318对公共互联网1354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 1318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1321中的数据平面VCN 1318上可以帮助将数据平面VCN 1318与其它客户和公共互联网1354隔离开。

[0198] 在一些实施例中,云服务1356可以由服务网关1336调用以访问公共互联网1354、控制平面VCN 1316或数据平面VCN 1318上可能不存在的服务。云服务1356与控制平面VCN 1316或数据平面VCN 1318之间的连接可以不是实时的或连续的。云服务1356可以存在于由IaaS提供商拥有或操作的不同网络上。云服务1356可以被配置为接收来自服务网关1336的调用并且可以被配置为不接收来自公共互联网1354的调用。一些云服务1356可以与其它云服务1356隔离,并且控制平面VCN 1316可以与可能与控制平面VCN 1316不在同一区域的云服务1356隔离。例如,控制平面VCN 1316可能位于“区域1”,并且云服务“部署12”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 1316中的服务网关1336对部署12进行调用,那么该调用可以被传输到区域1中的部署12。在这个示例中,控制平面VCN 1316或区域1中的部署12可能不与区域2中的部署12通信地耦合或以其它方式通信。

[0199] 图14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1400。服务运营商1402(例如,图12的服务运营商1202)可以通信地耦合到安全主机租赁1404(例如,图12的安全主机租赁1204),该安全主机租赁1404可以包括虚拟云网络(VCN)1406(例如,图12的VCN 1206)和安全主机子网1408(例如,图12的安全主机子网1208)。VCN 1406可以包括LPG 1410(例如,图12的LPG 1210),其可以经由包含在SSH VCN 1412中的LPG 1410通信地耦合到SSH VCN 1412(例如,图12的SSH VCN 1212)。SSH VCN 1412可以包括SSH子网1414(例如,图12的SSH子网1214),并且SSH VCN 1412可以经由包含在控制平面VCN 1416中的LPG 1410通信地耦合到控制平面VCN 1416(例如,图12的控制平面VCN 1216)并且经由包含在数据平面VCN 1418中的LPG 1410耦合到数据平面VCN 1418(例如,图12的数据平面

1218)。控制平面VCN 1416和数据平面VCN 1418可以包含在服务租赁1419(例如,图12的服务租赁1219)中。

[0200] 控制平面VCN 1416可以包括可以包括(一个或多个)负载平衡器(LB)子网1422(例如,图12的(一个或多个)LB子网1222)的控制平面DMZ层1420(例如,图12的控制平面DMZ层1220)、可以包括(一个或多个)应用子网1426(例如,类似于图12的(一个或多个)应用子网1226)的控制平面应用层1424(例如,图12的控制平面应用层1224)、可以包括(一个或多个)DB子网1430的控制平面数据层1428(例如,图12的控制平面数据层1228)。包含在控制平面DMZ层1420中的(一个或多个)LB子网1422可以通信地耦合到包含在控制平面应用层1424中的(一个或多个)应用子网1426和可以包含在控制平面VCN 1416中的互联网网关1434(例如,图12的互联网网关1234),并且(一个或多个)应用子网1426可以通信地耦合到包含在控制平面数据层1428中的(一个或多个)DB子网1430以及服务网关1436(例如,图12的服务网关)和网络地址转换(NAT)网关1438(例如,图12的NAT网关1238)。控制平面VCN 1416可以包括服务网关1436和NAT网关1438。

[0201] 数据平面VCN 1418可以包括数据平面应用层1446(例如,图12的数据平面应用层1246)、数据平面DMZ层1448(例如,图12的数据平面DMZ层1248),以及数据平面数据层1450(例如,图12的数据平面数据层1250)。数据平面DMZ层1448可以包括可以通信地耦合到数据平面应用层1446的(一个或多个)可信应用子网1460和(一个或多个)不可信应用子网1462以及包含在数据平面VCN 1418中的互联网网关1434的(一个或多个)LB子网1422。

[0202] (一个或多个)可信应用子网1460可以通信地耦合到包含在数据平面VCN 1418中的服务网关1436、包含在数据平面VCN 1418中的NAT网关1438以及包含在数据平面数据层1450中的(一个或多个)DB子网1430。(一个或多个)不可信应用子网1462可以通信地耦合到包含在数据平面VCN 1418中的服务网关1436和包含在数据平面数据层1450中的(一个或多个)DB子网1430。数据平面数据层1450可以包括可以通信地耦合到包含在数据平面VCN 1418中的服务网关1436的(一个或多个)DB子网1430。

[0203] (一个或多个)不可信应用子网1462可以包括可以通信地耦合到租户虚拟机(VM)1466(1)-(N)的一个或多个主VNIC 1464(1)-(N)。每个租户VM 1466(1)-(N)可以通信地耦合到可以包含在相应容器出口VCN 1468(1)-(N)中的相应应用子网1467(1)-(N),相应容器出口VCN 1468(1)-(N)可以包含在相应客户租赁1470(1)-(N)中。相应的辅助VNIC 1472(1)-(N)可以促进数据平面VCN 1418中包含的(一个或多个)不可信应用子网1462与容器出口VCN 1468(1)-(N)中包含的应用子网之间的通信。每个容器出口VCN 1468(1)-(N)可以包括NAT网关1438,该NAT网关1438可以通信地耦合到公共互联网1454(例如,图12的公共互联网1254)。

[0204] 包含在控制平面VCN 1416中并且包含在数据平面VCN 1418中的互联网网关1434可以通信地耦合到元数据管理服务1452(例如,图12的元数据管理系统1252),该元数据管理服务1452可以通信地耦合到公共互联网1454。公共互联网1454可以通信地耦合到包含在控制平面VCN 1416中并且包含在数据平面VCN 1418中的NAT网关1438。包含在控制平面VCN 1416中和包含在数据平面VCN 1418中的服务网关1436可以通信地耦合到云服务1456。

[0205] 在一些实施例中,数据平面VCN 1418可以与客户租赁1470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。

[0206] 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层应用1446的功能。运行该功能的代码可以在VM 1466(1)-(N)中执行,并且该代码可以不被配置为在数据平面VCN 1418上的其它任何地方运行。每个VM 1466(1)-(N)可以连接到一个客户租赁1470。包含在VM 1466(1)-(N)中的相应容器1471(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1471(1)-(N)运行代码,其中容器1471(1)-(N)可能至少包含在(一个或多个)不可信应用子网1462中包含的VM 1466(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1471(1)-(N)可以通信地耦合到客户租赁1470并且可以被配置为传输或接收来自客户租赁1470的数据。容器1471(1)-(N)可以不被配置为从数据平面VCN 1418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1471(1)-(N)。

[0207] 在一些实施例中,(一个或多个)可信应用子网1460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1460可以通信地耦合到(一个或多个)DB子网1430并且被配置为在(一个或多个)DB子网1430中执行CRUD操作。(一个或多个)不可信应用子网1462可以通信地耦合到(一个或多个)DB子网1430,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1430中执行读取操作。可以包含在每个客户的VM 1466(1)-(N)中并且可以运行来自客户的代码的容器1471(1)-(N)可以不与(一个或多个)DB子网1430通信地耦合。

[0208] 在其它实施例中,控制平面VCN 1416和数据平面VCN 1418可以不直接通信地耦合。在这个实施例中,控制平面VCN 1416和数据平面VCN 1418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 1410可以由IaaS提供商建立,其可以促进控制平面VCN 1416和数据平面VCN 1418之间的通信。在另一个示例中,控制平面VCN 1416或数据平面VCN 1418可以经由服务网关1436调用云服务1456。例如,从控制平面VCN 1416对云服务1456的调用可以包括对可以与数据平面VCN 1418通信的服务的请求。

[0209] 图15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1500。服务运营商1502(例如,图12的服务运营商1202)可以通信地耦合到安全主机租赁1504(例如,图12的安全主机租赁1204),该安全主机租赁1504可以包括虚拟云网络(VCN)1506(例如,图12的VCN 1206)和安全主机子网1508(例如,图12的安全主机子网1208)。VCN 1506可以包括LPG 1510(例如,图12的LPG 1210),该LPG 1510可以经由包含在SSH VCN 1512(例如,图12的SSH VCN 1212)中的LPG 1510通信地耦合到SSH VCN 1512。SSH VCN 1512可以包括SSH子网1514(例如,图12的SSH子网1214),并且SSH VCN 1512可以经由包含在控制平面VCN 1516中的LPG 1510通信地耦合到控制平面VCN 1516(例如,图12的控制平面VCN 1216)并且经由包含在数据平面VCN 1518中的LPG 1510耦合到数据平面VCN 1518(例如,图12的数据平面1218)。控制平面VCN 1516和数据平面VCN 1518可以包含在服务租赁1519(例如,图12的服务租赁1219)中。

[0210] 控制平面VCN 1516可以包括可以包括(一个或多个)LB子网1522(例如,图12的(一

个或多个)LB子网1222)的控制平面DMZ层1520(例如,图12的控制平面DMZ层1220)、可以包括(一个或多个)应用子网1526(例如,图12的(一个或多个)应用子网1226)的控制平面应用层1524(例如,图12的控制平面应用层1224)、可以包括(一个或多个)DB子网1530(例如,图14的(一个或多个)DB子网1430)的控制平面数据层1528(例如,图12的控制平面数据层1228)。包含在控制平面DMZ层1520中的(一个或多个)LB子网1522可以通信地耦合到包含在控制平面应用层1512中的(一个或多个)应用子网1526和可以包含在控制平面VCN 1516中的互联网网关1534(例如,图12的互联网网关1234),并且(一个或多个)应用子网1526可以通信地耦合到包含在控制平面数据层1528中的(一个或多个)DB子网1530以及服务网关1536(例如,图12的服务网关)和网络地址转换(NAT)网关1538(例如,图12的NAT网关1238)。控制平面VCN 1516可以包括服务网关1536和NAT网关1538。

[0211] 数据平面VCN 1518可以包括数据平面应用层1546(例如,图12的数据平面应用层1246)、数据平面DMZ层1548(例如,图12的数据平面DMZ层1248)),以及数据平面数据层1550(例如,图12的数据平面数据层1250)。数据平面DMZ层1548可以包括可以通信地耦合到数据平面应用层1546的(一个或多个)可信应用子网1560(例如,图14的(一个或多个)可信应用子网1460)和(一个或多个)不可信应用子网1562(例如,图14的(一个或多个)不可信应用子网1462)以及包含在数据平面VCN 1518中的互联网网关1534的(一个或多个)LB子网1522。(一个或多个)可信应用子网1560可以通信地耦合到包含在数据平面VCN 1518中的服务网关1536、包含在数据平面VCN 1518中的NAT网关1538以及包含在数据平面数据层1550中的(一个或多个)DB子网1530。(一个或多个)不可信应用子网1562可以通信地耦合到包含在数据平面VCN 1518中的服务网关1536和包含在数据平面数据层1550中的(一个或多个)DB子网1530。数据平面数据层1550可以包括可以通信地耦合到包含在数据平面VCN 1518中的服务网关1536的(一个或多个)DB子网1530。

[0212] (一个或多个)不可信应用子网1562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1562内的租户虚拟机(VM)1566(1)-(N)的主VNIC 1564(1)-(N)。每个租户VM 1566(1)-(N)可以在相应的容器1567(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 1568中包含的数据平面应用层1546中的应用子网1526。相应的辅助VNIC 1572(1)-(N)可以促进包含在数据平面VCN 1518中的(一个或多个)不可信应用子网1562和包含在容器出口VCN 1568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1554(例如,图12的公共互联网1254)的NAT网关1538。

[0213] 包含在控制平面VCN 1516中和包含在数据平面VCN 1518中的互联网网关1534可以通信地耦合到元数据管理服务1552(例如,图12的元数据管理系统1252),该元数据管理服务1552可以通信地耦合到公共互联网1554。公共互联网1554可以通信地耦合到包含在控制平面VCN 1516中并且包含在数据平面VCN 1518中的NAT网关1538。包含在控制平面VCN 1516中并且包含在数据平面VCN 1518中的服务网关1536可以通信地耦合到云服务1556。

[0214] 在一些示例中,图15的框图1500的体系架构所示的模式可以被认为是图14的框图1400的体系架构所示的模式的例外,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 1566(1)-(N)中包含的相应容器1567(1)-(N)。容器1567(1)-(N)可以被配置为对包含在数据平面应用层1546的(一个或多个)应用子网1526中的相应辅助VNIC 1572(1)-(N)

进行调用,该数据平面应用层1546可以包含在容器出口VCN 1568中。辅助VNIC 1572(1)-(N)可以将调用传输到NAT网关1538,NAT网关1538可以将调用传输到公共互联网1554。在这个示例中,可以由客户实时访问的容器1567(1)-(N)可以与控制平面VCN 1516隔离,并且可以与数据平面VCN 1518中包含的其它实体隔离。容器1567(1)-(N)也可以与来自其它客户的资源隔离。

[0215] 在其它示例中,客户可以使用容器1567(1)-(N)来调用云服务1556。在这个示例中,客户可以运行容器1567(1)-(N)中从云服务1556请求服务的代码。容器1567(1)-(N)可以将该请求传输到辅助VNIC 1572(1)-(N),辅助VNIC 1572(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1554。公共互联网1554可以经由互联网网关1534将请求传输到包含在控制平面VCN 1516中的(一个或多个)LB子网1522。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1526,该(一个或多个)应用子网1526可以经由服务网关1536将请求传输到云服务1556。

[0216] 应当认识到的是,各图中描绘的IaaS体系架构1200、1300、1400、1500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。

[0217] 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)。

[0218] 图16图示了其中可以实现各种实施例的示例计算机系统1600。系统1600可以用于实现上述任何计算机系统。如图所示,计算机系统1600包括经由总线子系统1602与多个外围子系统通信的处理单元1604。这些外围子系统可以包括处理加速单元1606、I/O子系统1608、存储子系统1618和通信子系统1624。存储子系统1618包括有形计算机可读存储介质1622和系统存储器1610。

[0219] 总线子系统1602提供用于让计算机系统1600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1602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1602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。

[0220] 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1604控制计算机系统1600的操作。一个或多个处理器可以被包括在处理单元1604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1604可以被实现为一个或多个独立的处理单元1632和/或1634,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1604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。

[0221] 在各种实施例中,处理单元1604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1604中和/或存储子系统1618中。通过适当的编程,(一个或多个)处理器1604可以提供上述各种功能。计算机系统1600可以附加地包括处理加速单元1606,其可以包括数字信号处理器(DSP)、专用处理器,等等。

[0222] I/O子系统1608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。

[0223] 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。

[0224] 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1600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。

[0225] 计算机系统1600可以包括包含软件元件、被示为当前位于系统存储器1610中的存储子系统1618。系统存储器1610可以存储可加载并且可在处理单元1604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。

[0226] 取决于计算机系统1600的配置和类型,系统存储器1610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1604立即访问和/或目前正被处理单元1604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1610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1600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1610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1612,程序数据1614,以及操作系统1616。作为示例,操作系统1616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® 10OS和Palm®OS操作系统的移动操作系统。

[0227] 存储子系统1618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1618中。这些软件模块或指令可以被处理单元1604执行。存储子系统1618也可以提供用于存储根据本公开被使用的数据的储存库。

[0228] 存储子系统1600也可以包括可被进一步连接到计算机可读存储介质1622的计算机可读存储介质读取器1620。与系统存储器1610一起并且,可选地,与其相结合,计算机可读存储介质1622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。

[0229] 包含代码或代码的部分的计算机可读存储介质1622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)或其它光学储存器、磁带盒、磁带、磁盘储存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1600访问的任何其它介质。

[0230] 举例来说,计算机可读存储介质1622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1622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1622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的组合的混合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1600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。

[0231] 通信子系统1624提供到其它计算机系统和网络的接口。通信子系统1624用作用于从其它系统接收数据和从计算机系统1600向其它系统发送数据的接口。例如,通信子系统1624可以使计算机系统1600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1624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1624可以提供有线网络连接(例如,以太网)。

[0232] 在一些实施例中,通信子系统1624也可以代表可以使用计算机系统1600的一个或多个用户接收结构化和/或非结构化数据馈送1626、事件流1628、事件更新1630等形式的输入通信。

[0233] 举例来说,通信子系统1624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1626,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。

[0234] 此外,通信子系统1624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1628和/或事件更新1630。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。

[0235] 通信子系统1624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1626、事件流1628、事件更新1630,等等,这一个或多个数据库可以与耦合到计算机系统1600的一个或多个流式数据源计算机通信。

[0236] 计算机系统1600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。

[0237] 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1600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。

[0238] 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内。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。

[0239] 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内。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。

[0240] 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。

[0241] 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。

[0242] 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。

[0243] 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。

[0244] 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。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和应用中被使用。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。

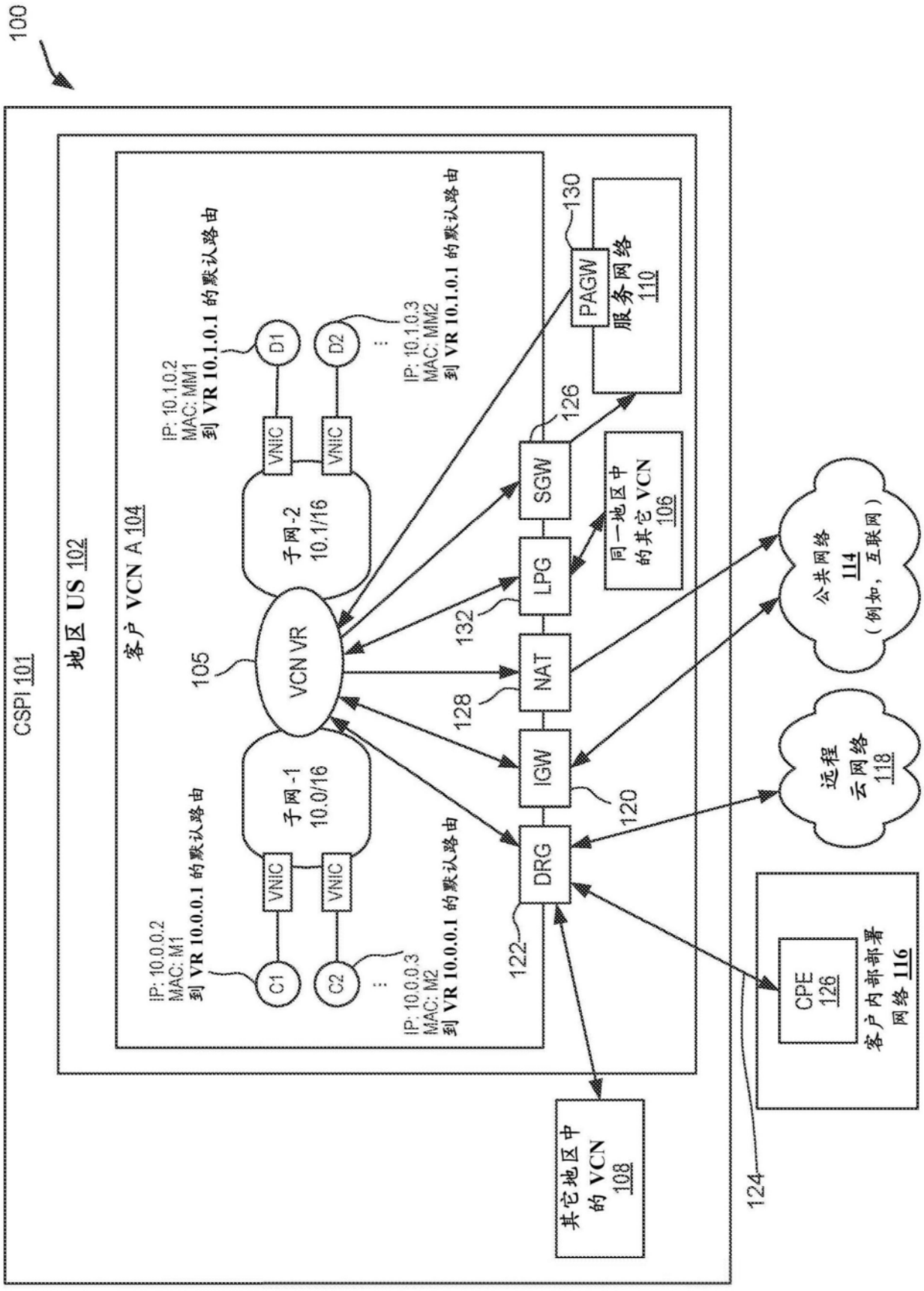

100
CSPI 101
地区 US 102
客户 VCN A 104
105
VCN VR
子网-1
10.0/16
子网-2
10.1/16
VNIC
VNIC
VNIC
VNIC
C1
C2
D1
D2
IP: 10.0.0.2
MAC: M1
到 VR 10.0.0.1 的默认路由
IP: 10.0.0.3
MAC: M2
到 VR 10.0.0.1 的默认路由
IP: 10.1.0.2
MAC: MM1
到 VR 10.1.0.1 的默认路由
IP: 10.1.0.3
MAC: MM2
到 VR 10.1.0.1 的默认路由
DRG
IGW
NAT
LPG
SGW
PAGW
122
120
128
132
126
130
其它地区中
的 VCN
108
同一地区中
的其它 VCN
106
服务网络
110
公共网络
114
(例如，互联网)
远程
云网络
118
CPE
126
客户内部部署
网络 116
124

图1

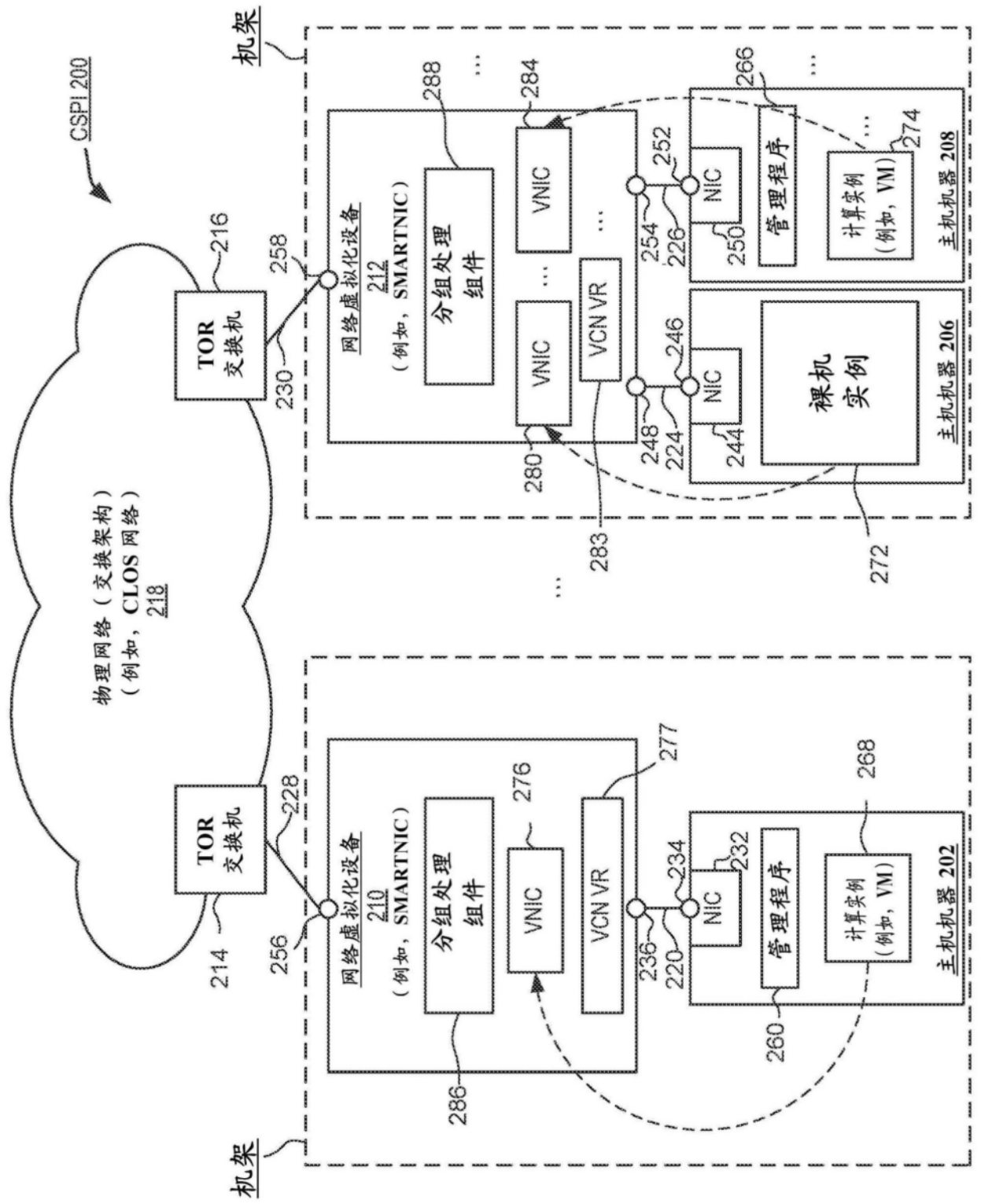

CSPI 200
物理网络（交换架构）
（例如，CLOS 网络）
218
TOR
交换机
216
214
TOR
交换机
258
230
228
256
网络虚拟化设备
212
（例如，SMARTNIC）
网络虚拟化设备
210
（例如，SMARTNIC）
分组处理
组件
288
分组处理
组件
286
VNIC
284
VNIC
280
VNIC
276
VCN VR
283
VCN VR
277
NIC
252
NIC
246
NIC
234
NIC
232
254
226
248
224
236
220
250
244
管理程序
266
管理程序
260
裸机
实例
272
计算实例
（例如，VM）
274
计算实例
（例如，VM）
268
主机机器 208
主机机器 206
主机机器 202
机架
机架

图2

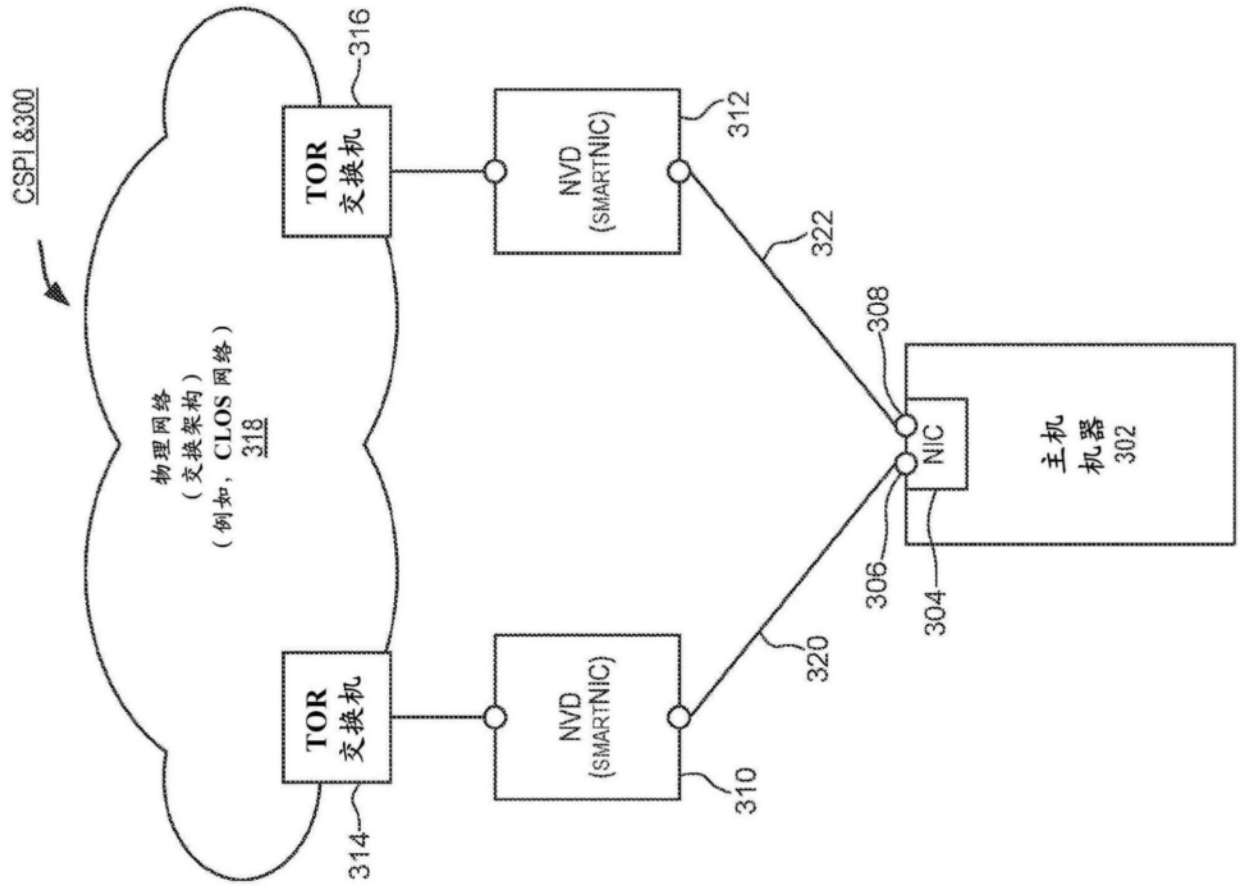

CSP1 &300
物理网络
（交换架构）
（例如，CLOS 网络）
318
TOR
交换机
316
TOR
交换机
314
NVD
(SMARTNIC)
312
NVD
(SMARTNIC)
310
322
320
308
NIC
306
304
主机
机器
302

图3

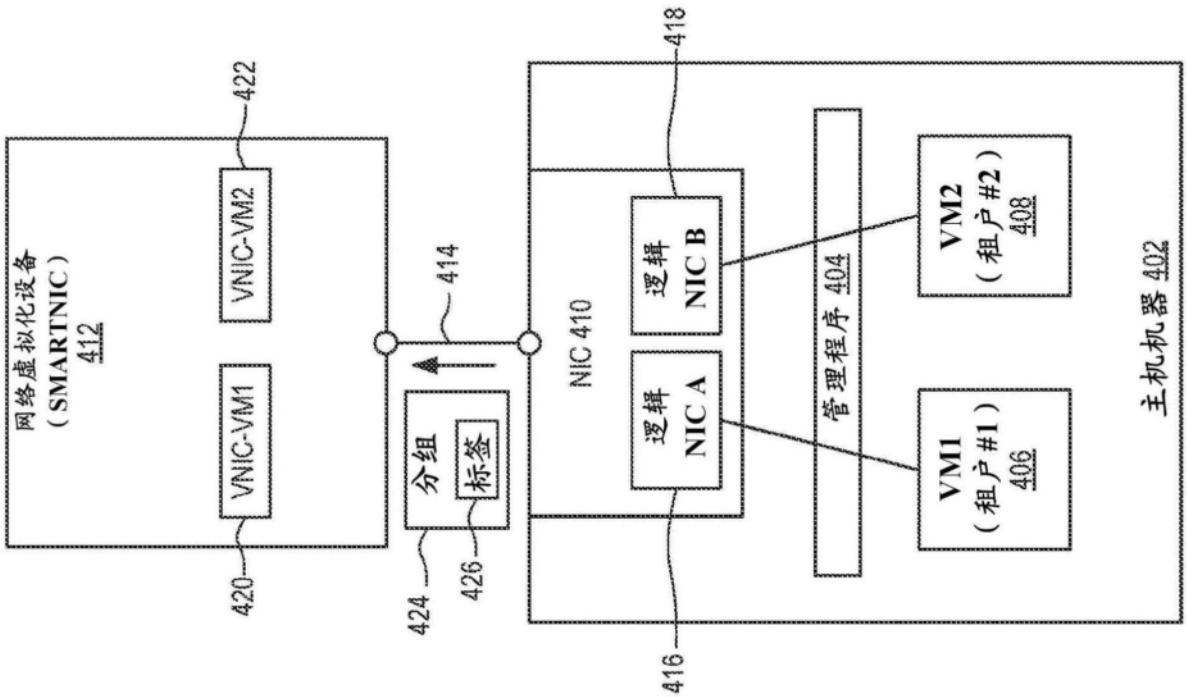

网络虚拟化设备
（SMARTNIC）
412
VNIC-VM1
VNIC-VM2
422
420
分组
标签
414
424
426
NIC 410
逻辑
NIC A
逻辑
NIC B
418
416
管理程序 404
VM1
（租户#1）
406
VM2
（租户#2）
408
主机机器 402

图4

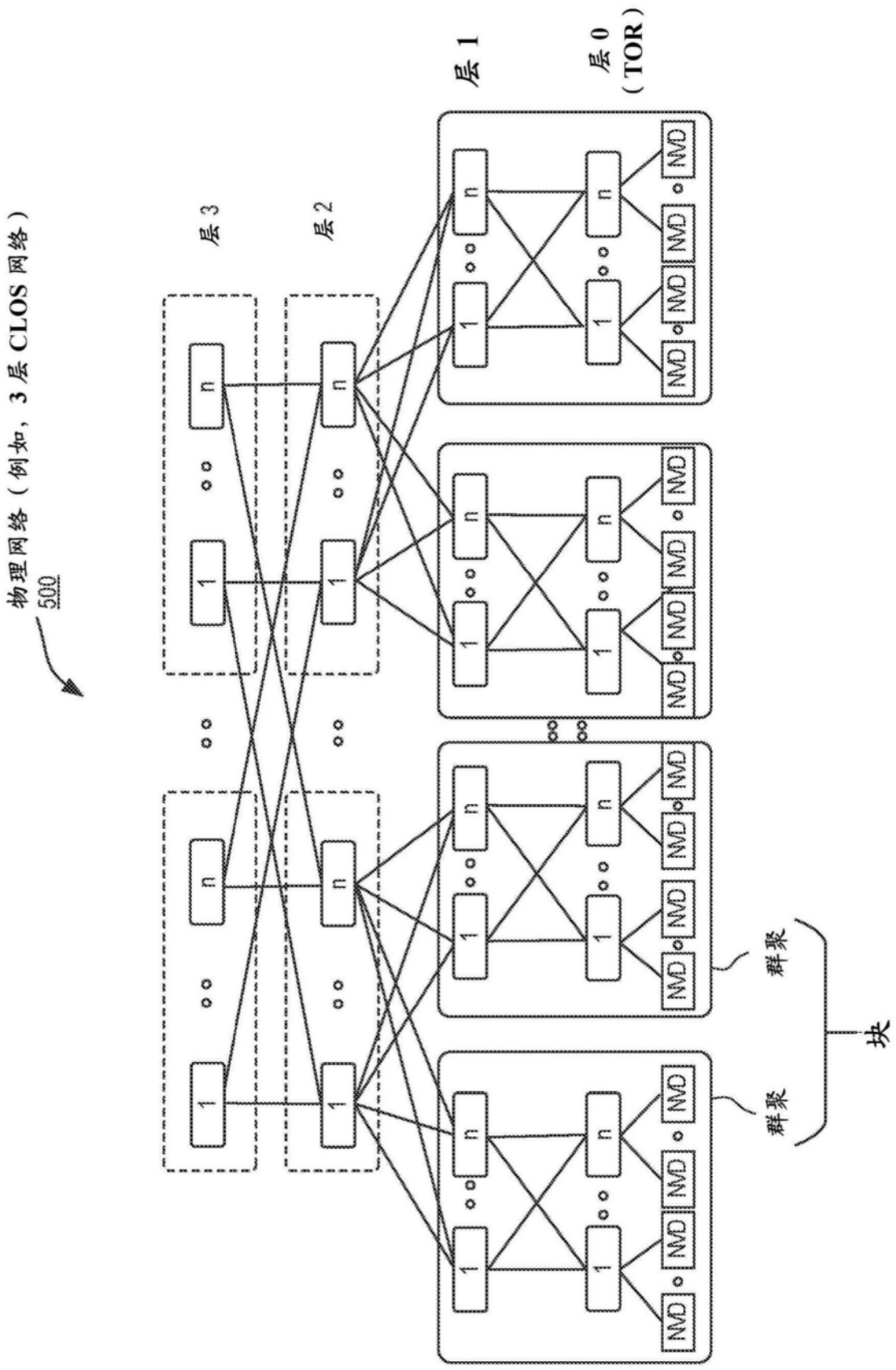


图5

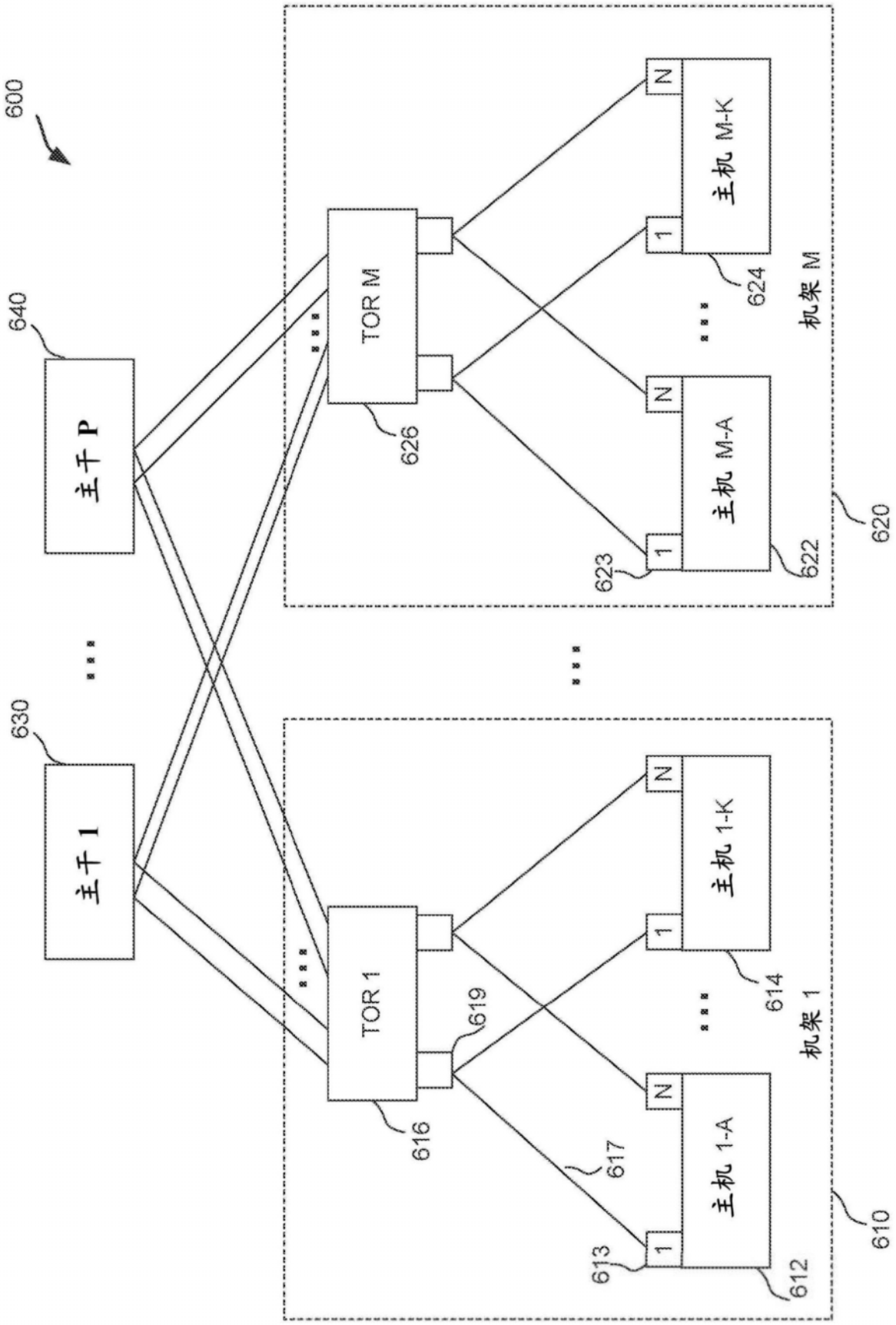

600
640
主干 P
630
主干 1
TOR M
TOR 1
626
616
619
617
623
613
N
1
主机 M-K
主机 M-A
主机 1-K
主机 1-A
624
622
614
612
机架 M
机架 1
620
610

图6

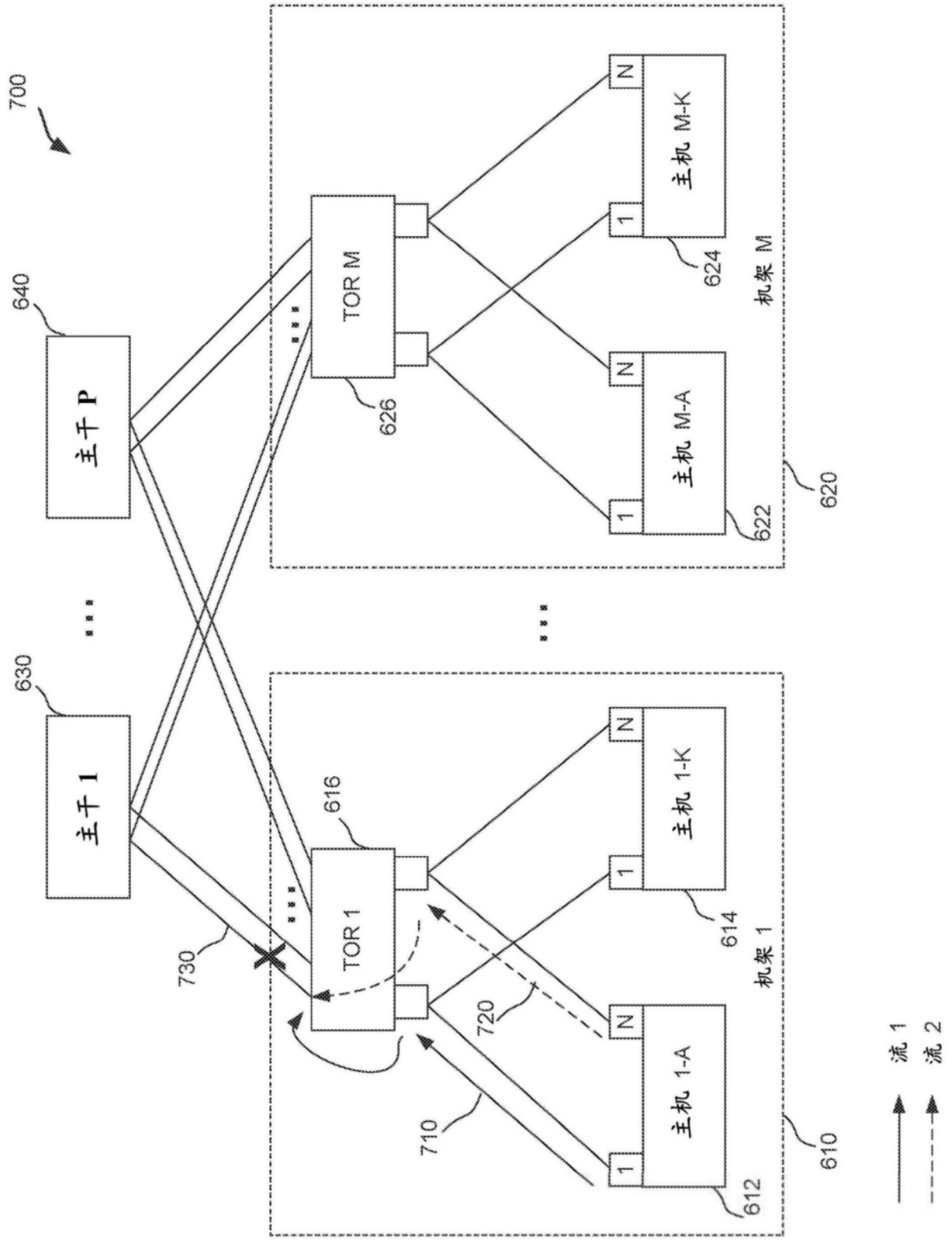

700
640
主干 P
630
主干 1
TOR M
626
TOR 1
616
N
1
主机 M-K
624
主机 M-A
622
620
机架 M
主机 1-K
614
主机 1-A
612
610
机架 1
730
710
720
流 1
流 2

图7

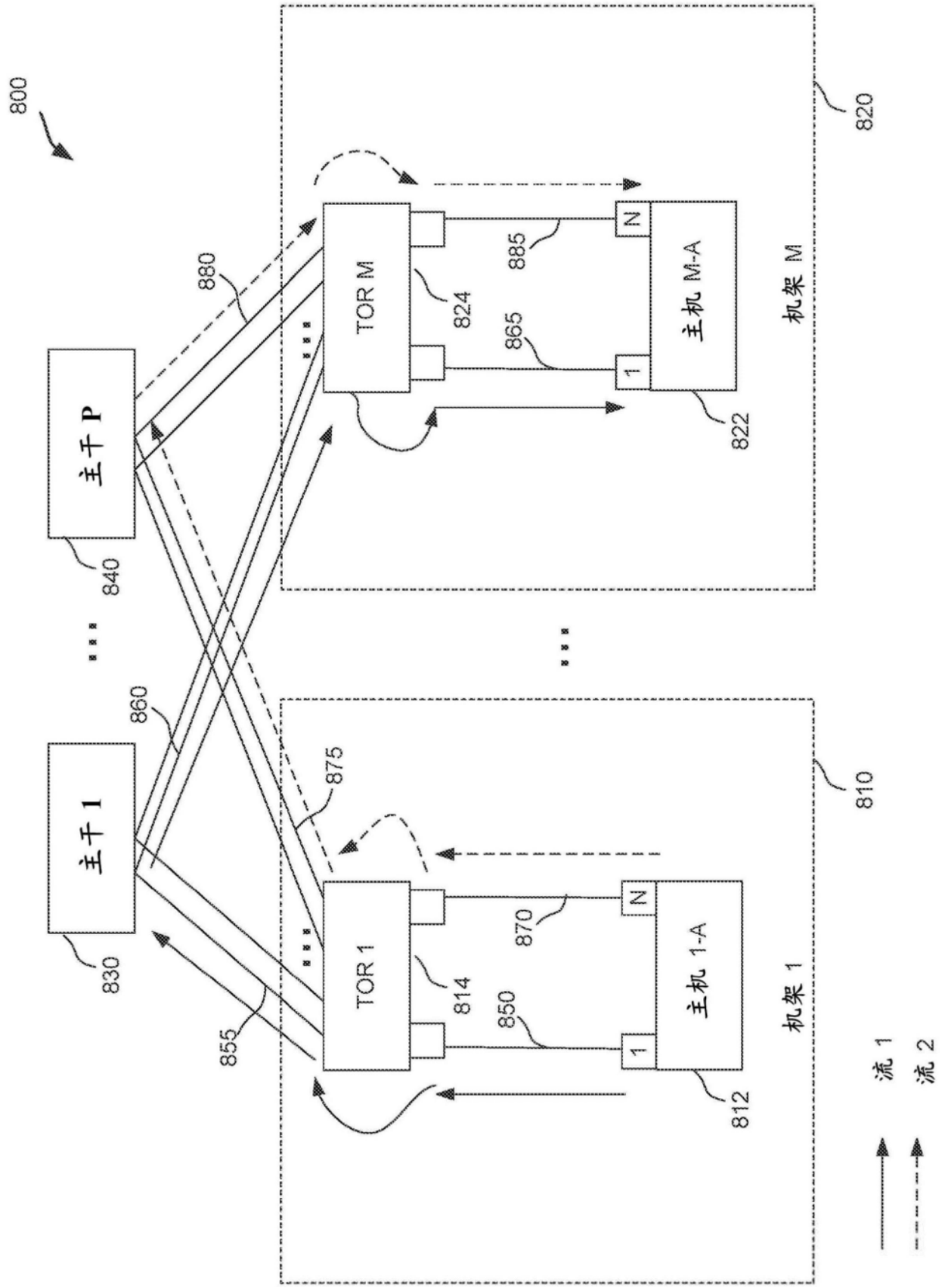

800
820
机架 M
TOR M
主机 M-A
N
1
885
824
865
822
880
主干 P
840
860
875
810
机架 1
TOR 1
主机 1-A
870
814
850
812
830
855
流 1
流 2

图8

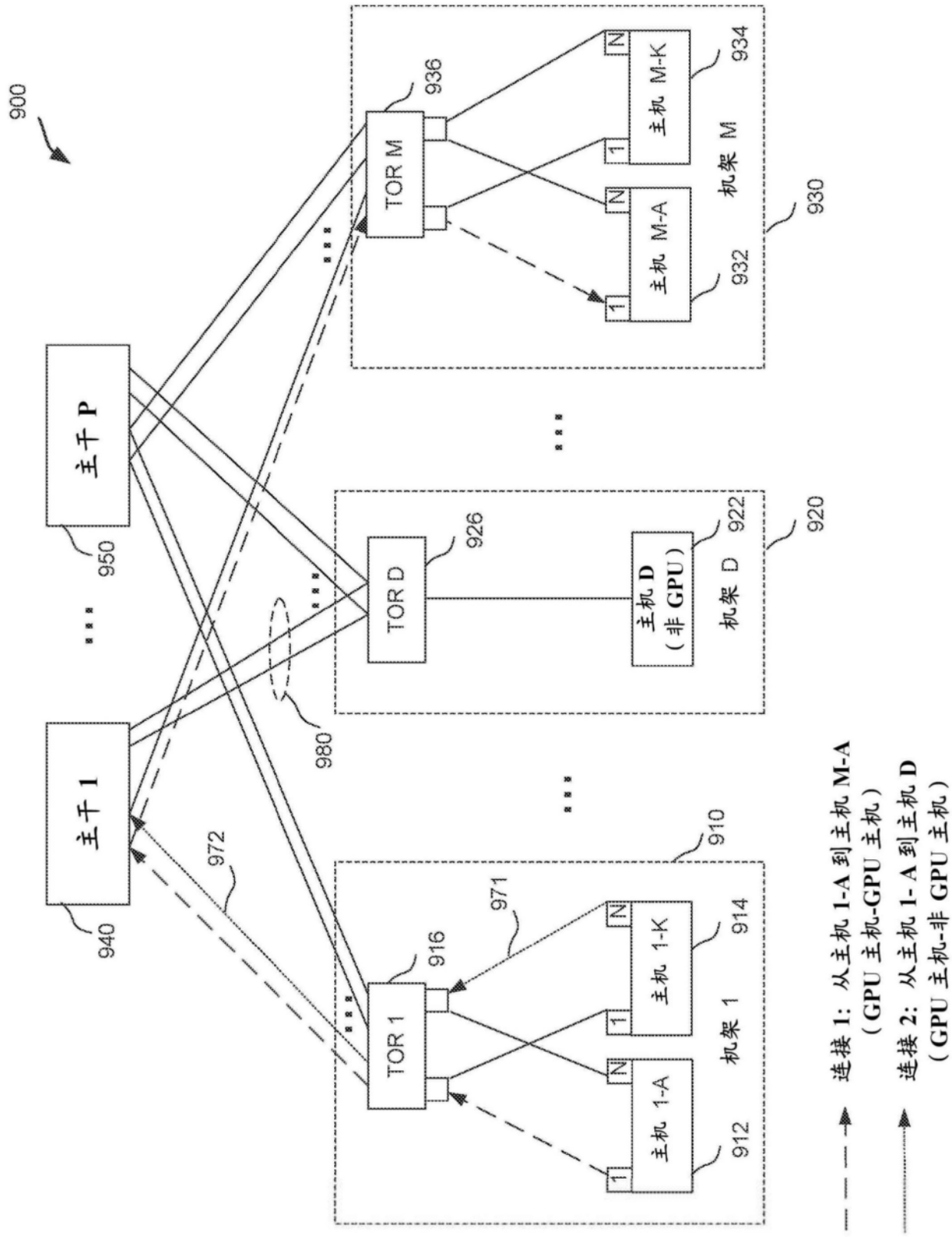

900
主干 P
主干 1
950
940
980
972
971
TOR M
936
TOR D
926
TOR 1
916
主机 M-K
934
主机 M-A
932
机架 M
930
主机 D
（非 GPU）
922
机架 D
920
主机 1-K
914
主机 1-A
912
机架 1
910
N
1
连接 1：从主机 1-A 到主机 M-A
（GPU 主机-GPU 主机）
连接 2：从主机 1- A 到主机 D
（GPU 主机-非 GPU 主机）

图9

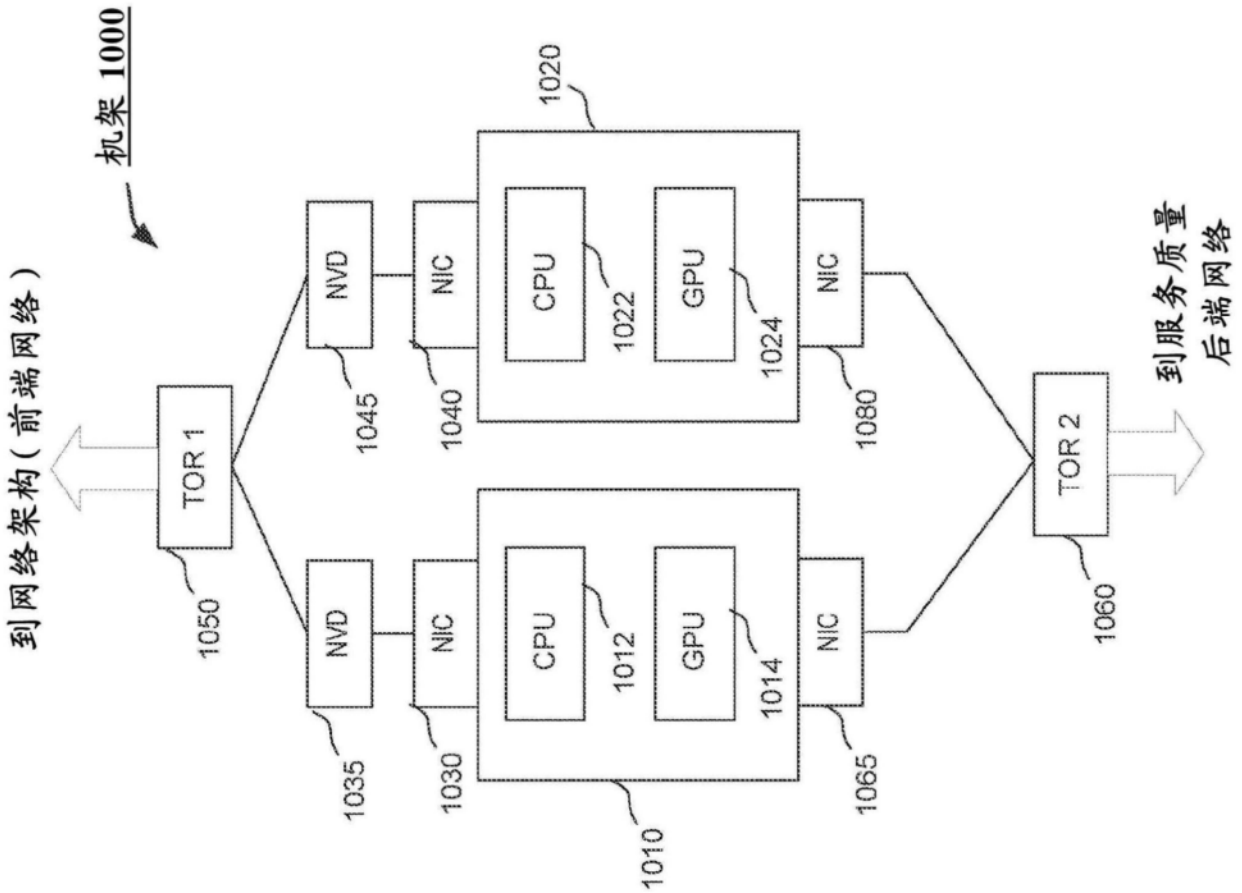

机架 1000
到网络架构（前端网络）
TOR 1
1050
NVD
1045
NIC
1040
CPU
1022
GPU
1024
NIC
1080
1020
NVD
1035
NIC
1030
CPU
1012
GPU
1014
NIC
1065
1010
TOR 2
1060
到服务质量后端网络

图10

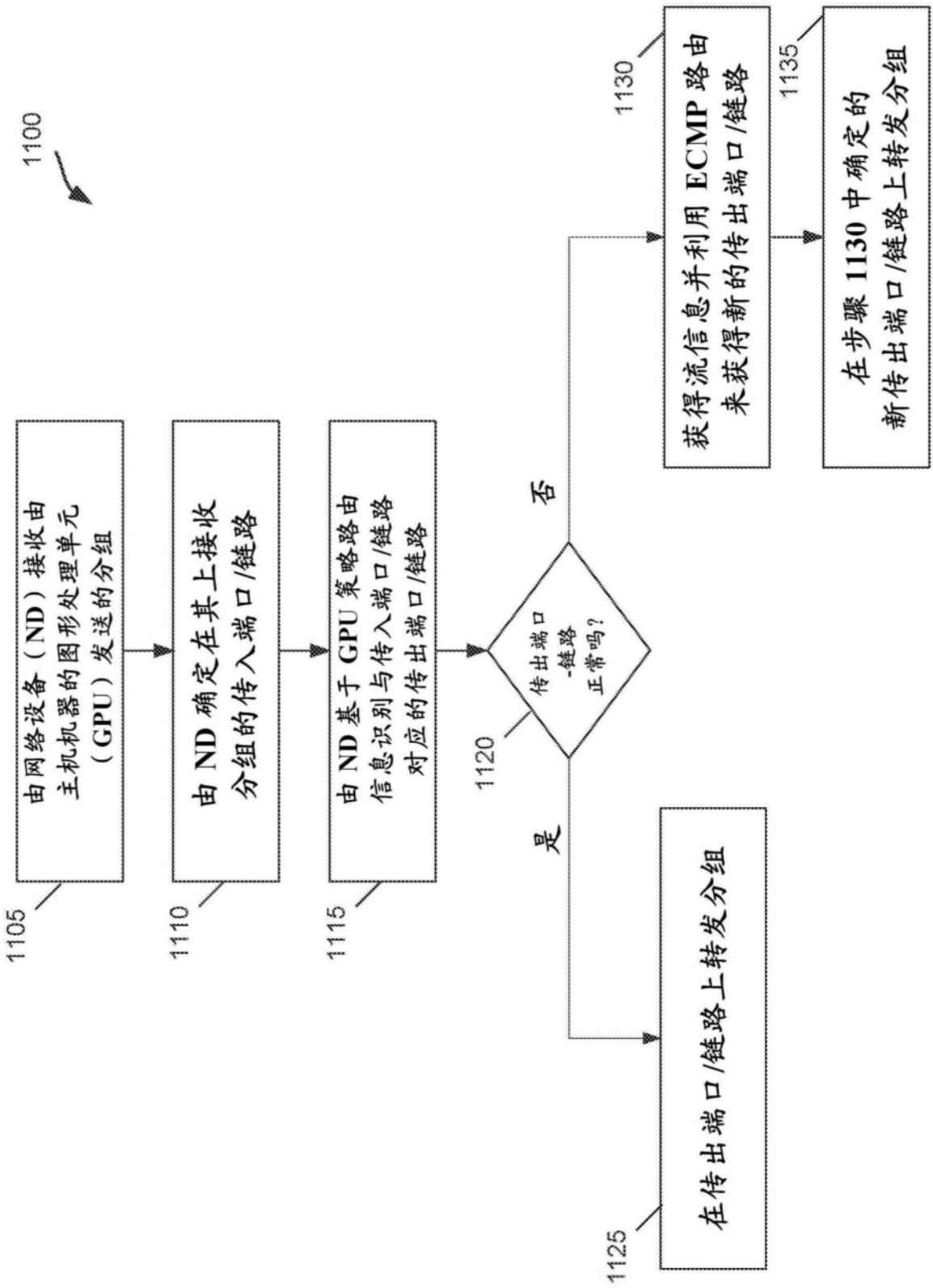

1100
由网络设备（ND）接收由主机机器的图形处理单元（GPU）发送的分组
1105
由 ND 确定在其上接收分组的传入端口/链路
1110
由 ND 基于 GPU 策略路由信息识别与传入端口/链路对应的传出端口/链路
1115
传出端口-链路正常吗?
1120
否
获得流信息并利用 ECMP 路由来获得新的传出端口/链路
1130
在步骤 1130 中确定的新传出端口/链路上转发分组
1135
是
在传出端口/链路上转发分组
1125

图11A

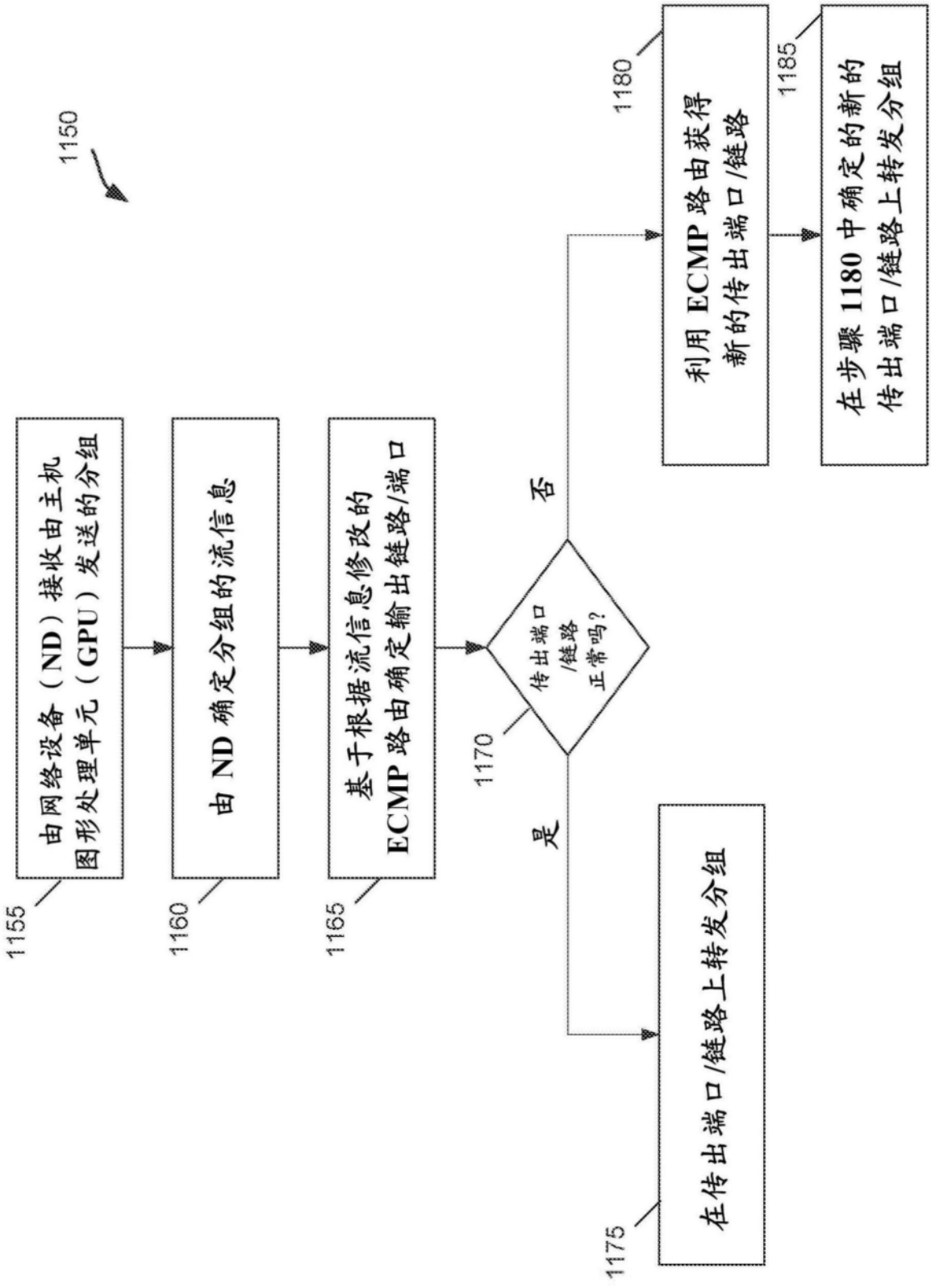

1150
由网络设备（ND）接收由主机图形处理单元（GPU）发送的分组
1155
由ND确定分组的流信息
1160
基于根据流信息修改的ECMP路由确定输出链路/端口
1165
传出端口/链路正常吗？
1170
否
是
利用ECMP路由获得新的传出端口/链路
1180
在步骤1180中确定的新的传出端口/链路上转发分组
1185
在传出端口/链路上转发分组
1175

图11B

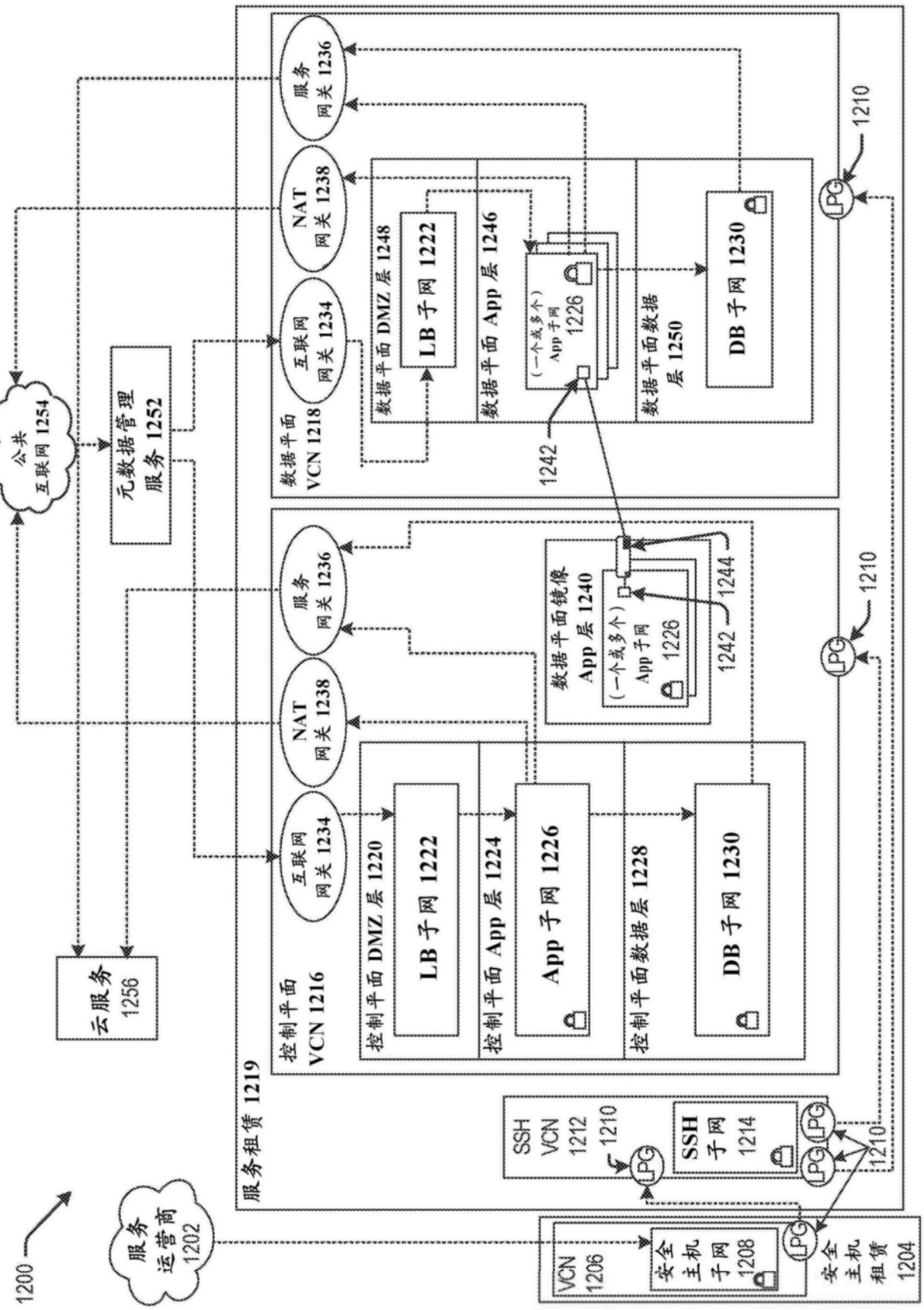

1200
服务运营商 1202
服务租赁 1219
VCN 1206
安全主机子网 1208
安全主机租赁 1204
SSH VCN 1212
SSH 子网 1214
1210
LPG
控制平面 VCN 1216
控制平面 DMZ 层 1220
LB 子网 1222
控制平面 App 层 1224
App 子网 1226
控制平面数据层 1228
DB 子网 1230
互联网网关 1234
NAT 网关 1238
服务网关 1236
数据平面镜像 App 层 1240
(一个或多个) App 子网 1226
1242
1244
数据平面 VCN 1218
数据平面 DMZ 层 1248
数据平面 App 层 1246
数据平面数据层 1250
云服务 1256
元数据管理服务 1252
公共互联网 1254

图12

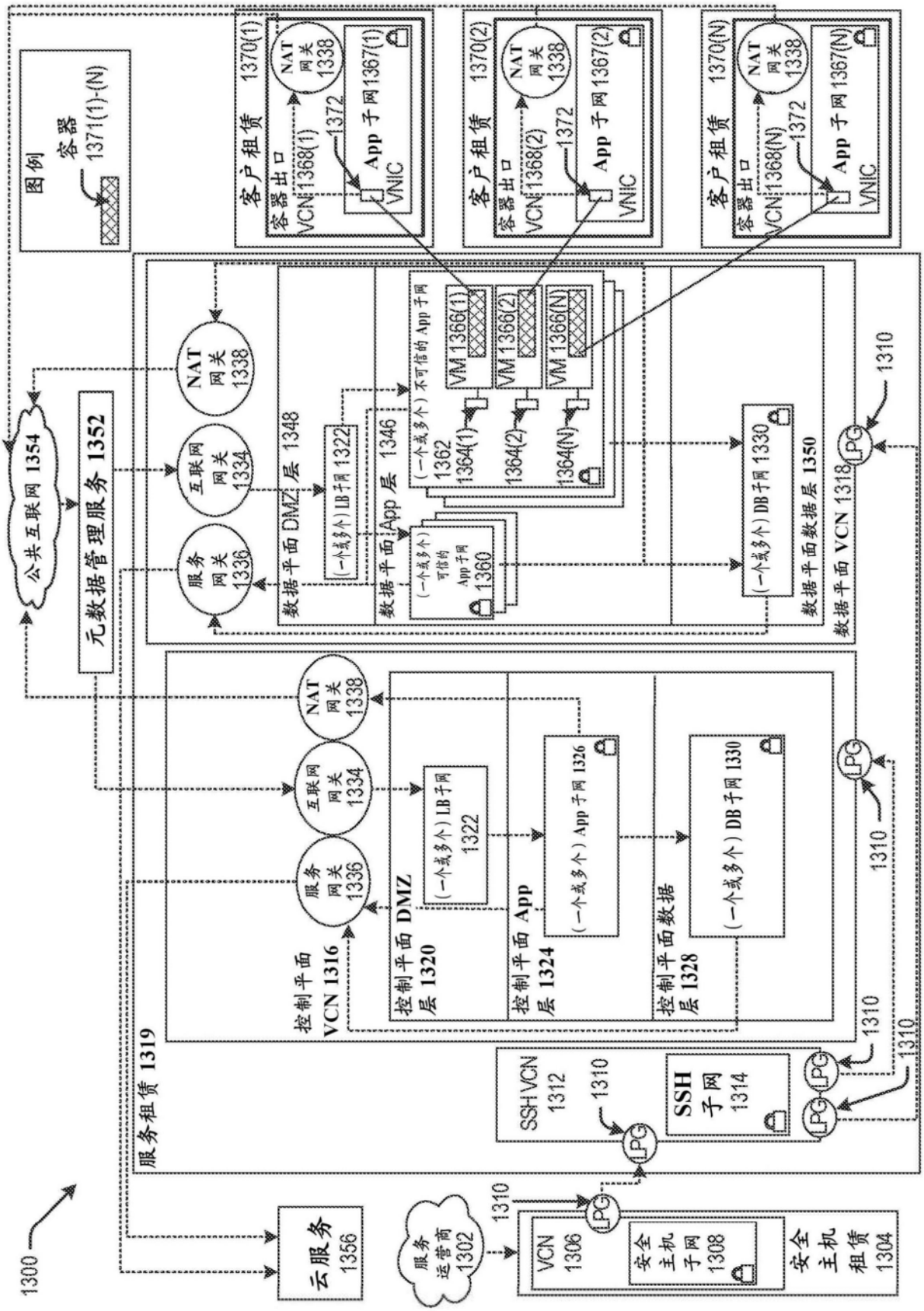

1300
云服务 1356
服务运营商 1302
VCN 1306
安全主机子网 1308
安全主机租赁 1304
SSH VCN 1312
SSH 子网 1314
服务租赁 1319
控制平面 VCN 1316
控制平面层 1320 DMZ
控制平面层 1324 App
控制平面层 1328 数据
(一个或多个) LB子网 1322
(一个或多个) App子网 1326
(一个或多个) DB子网 1330
服务网关 1336
互联网网关 1334
NAT网关 1338
LPG
1310
公共互联网 1354
元数据管理服务 1352
数据平面 DMZ 层 1348
数据平面 App 层 1346
数据平面数据层 1350
数据平面 VCN 1318
(一个或多个) 可信的 App 子网 1360
(一个或多个) 不可信的 App 子网 1362
VM 1366(1)
VM 1366(2)
VM 1366(N)
1364(1)
1364(2)
1364(N)
客户租赁 1370(1)
客户租赁 1370(2)
客户租赁 1370(N)
容器出口 VCN 1368(1)
容器出口 VCN 1368(2)
容器出口 VCN 1368(N)
1372
App 子网 1367(1) VNIC
App 子网 1367(2) VNIC
App 子网 1367(N) VNIC
图例
容器 1371(1)-(N)

图13

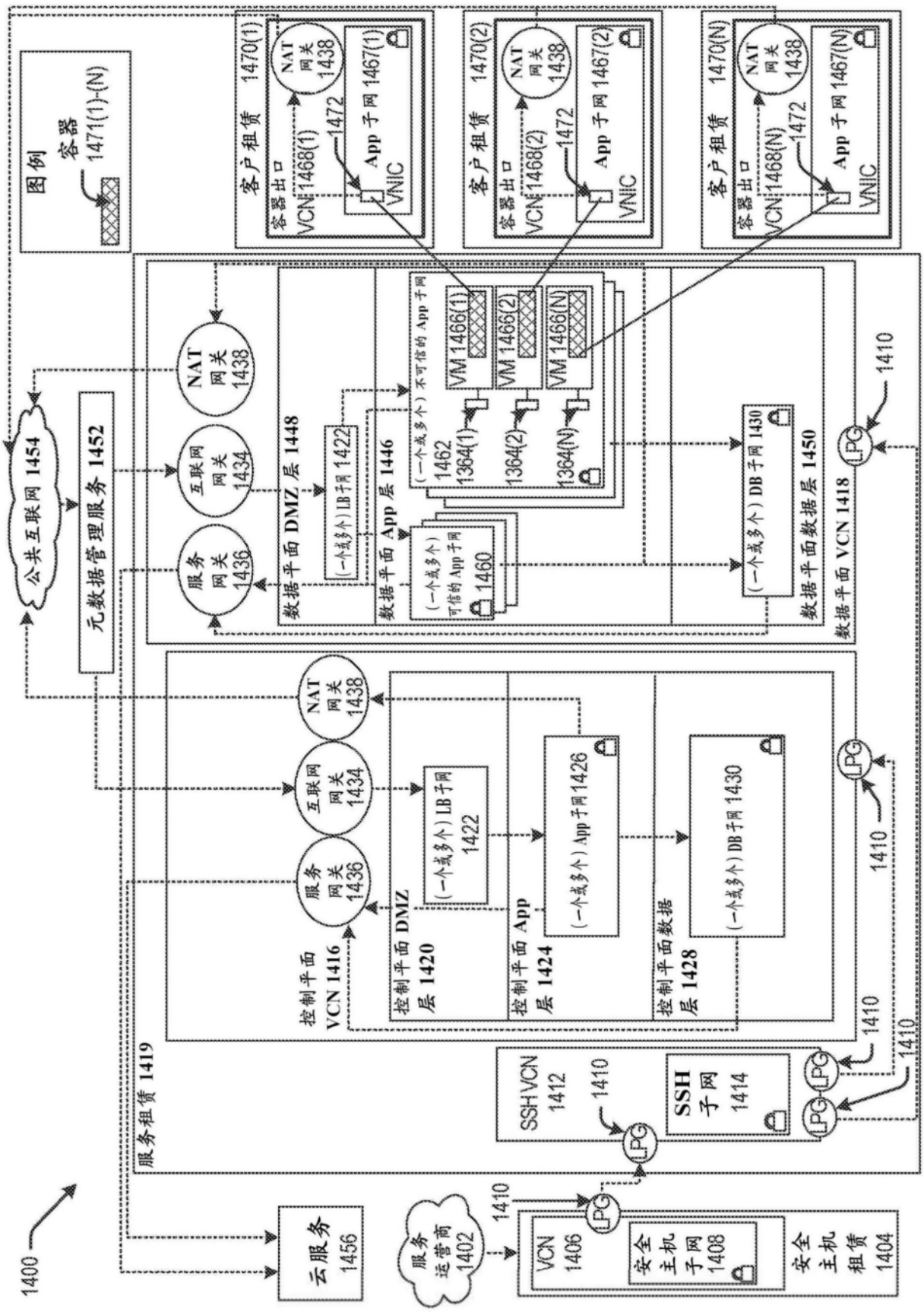

图例
容器
1471(1)-(N)
客户租赁 1470(1)
容器出口
VCN 1468(1)
NAT 网关 1438
1472
App 子网 1467(1)
VNIC
客户租赁 1470(2)
VCN 1468(2)
App 子网 1467(2)
客户租赁 1470(N)
VCN 1468(N)
App 子网 1467(N)
公共互联网 1454
元数据管理服务 1452
服务网关 1436
互联网网关 1434
数据平面 DMZ 层 1448
(一个或多个) LB子网 1422
数据平面 App 层 1446
(一个或多个) 不可信的 App 子网
VM 1466(1)
VM 1466(2)
VM 1466(N)
1462
1364(1)
1364(2)
1364(N)
(一个或多个) 可信的 App 子网 1460
(一个或多个) DB 子网 1430
数据平面数据层 1450
数据平面 VCN 1418
1410
服务租赁 1419
控制平面 VCN 1416
控制平面 DMZ 层 1420
控制平面 App 层 1424
(一个或多个) App子网 1426
控制平面数据层 1428
SSH VCN 1412
SSH 子网 1414
LPG
1400
云服务 1456
服务运营商 1402
VCN 1406
安全主机子网 1408
安全主机租赁 1404

图14

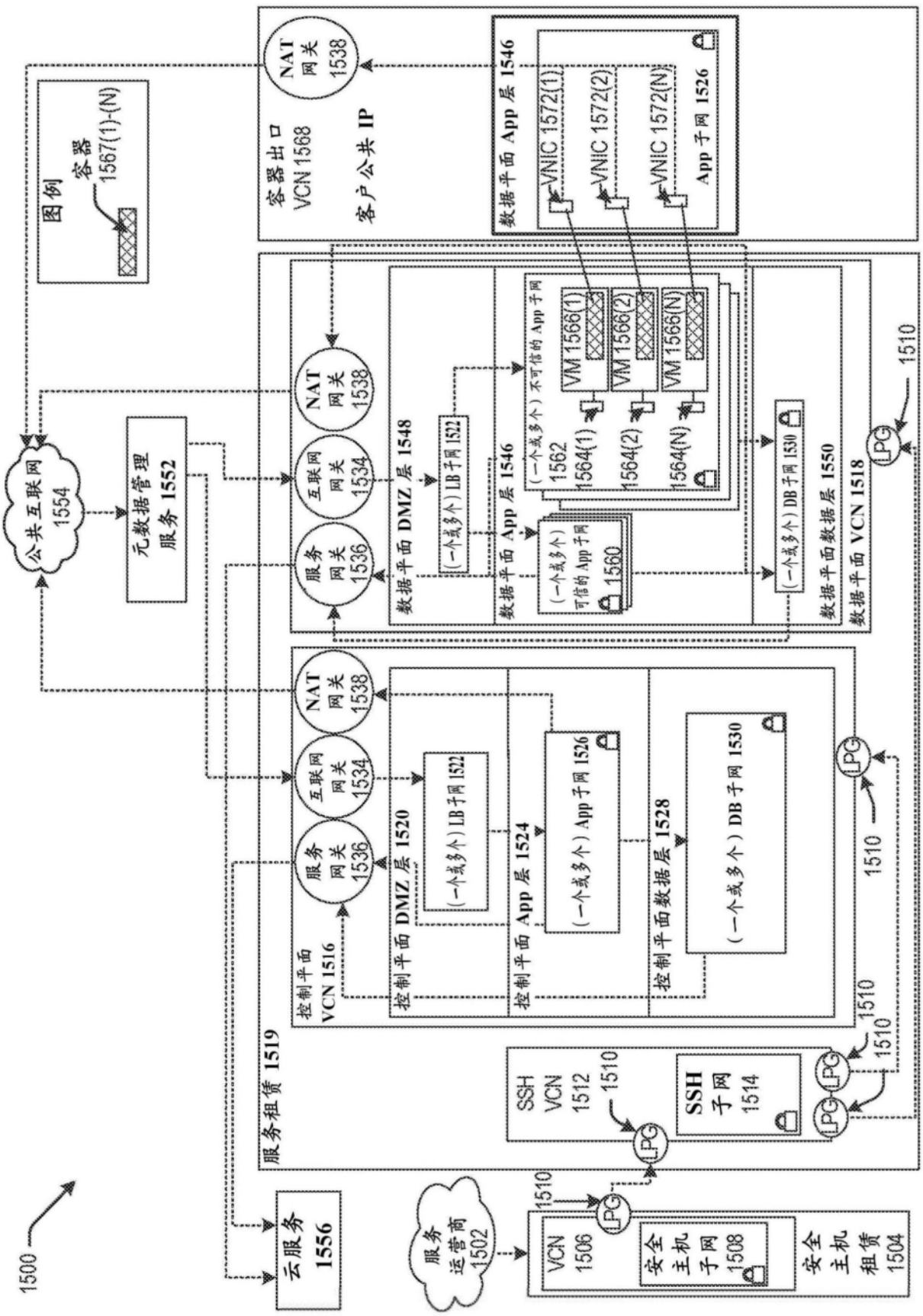

1500
图例
容器
1567(1)-(N)
容器出口
VCN 1568
客户公共 IP
NAT 网关 1538
数据平面 App 层 1546
VNIC 1572(1)
VNIC 1572(2)
VNIC 1572(N)
App 子网 1526
公共互联网 1554
元数据管理服务 1552
服务网关 1536
互联网网关 1534
数据平面 DMZ 层 1548
(一个或多个) LB 子网 1522
(一个或多个) 不可信的 App 子网
VM 1566(1)
VM 1566(2)
VM 1566(N)
1562
1564(1)
1564(2)
1564(N)
(一个或多个) 可信的 App 子网
1560
(一个或多个) DB 子网 1530
数据平面数据层 1550
数据平面 VCN 1518
1510
服务租赁 1519
控制平面 VCN 1516
控制平面 DMZ 层 1520
控制平面 App 层 1524
(一个或多个) App 子网 1526
控制平面数据层 1528
SSH VCN 1512
SSH 子网 1514
LPG
云服务 1556
服务运营商 1502
VCN 1506
安全主机子网 1508
安全主机租赁 1504

图15

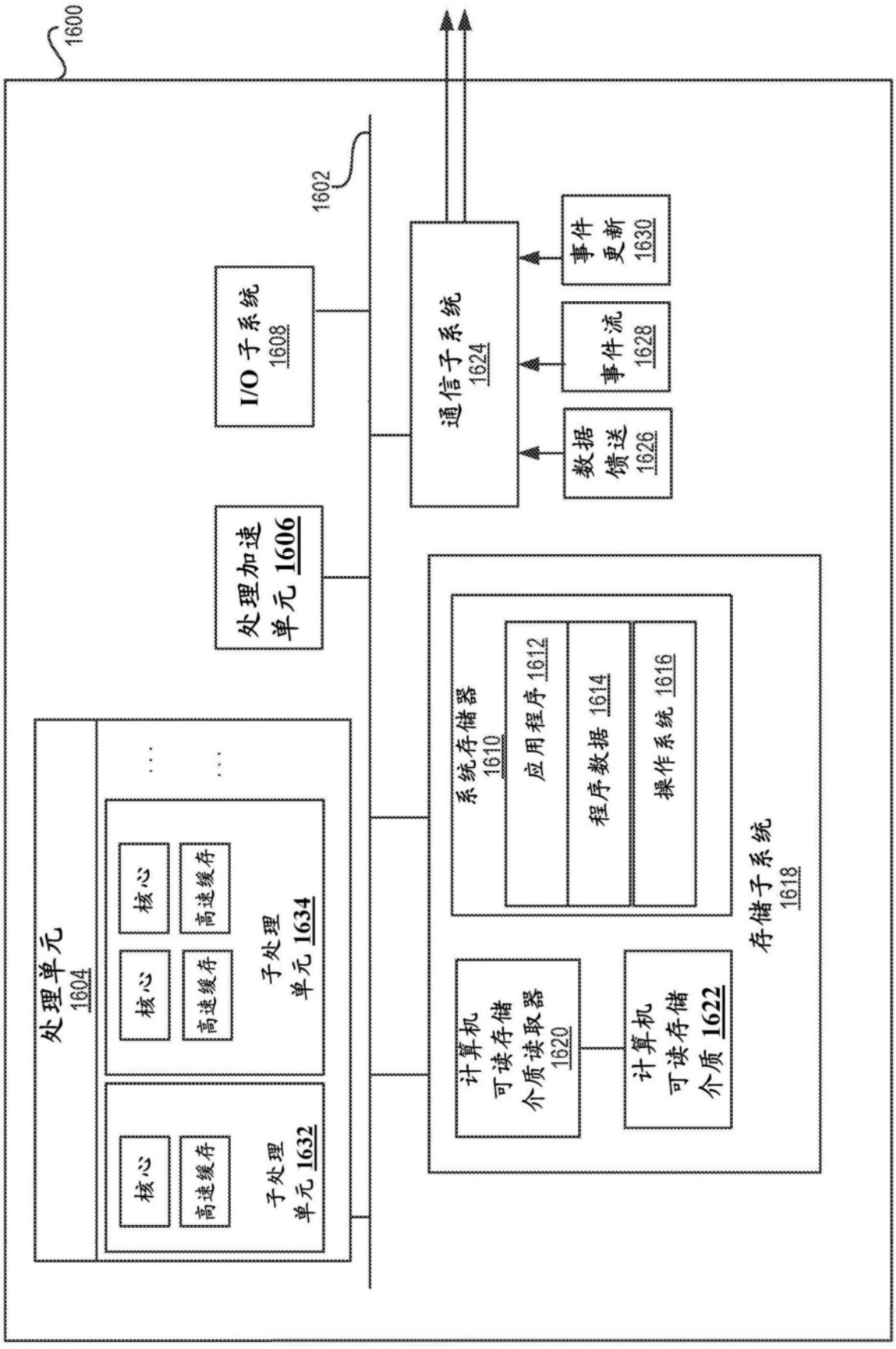

1600
1602
处理单元
1604
核心
高速缓存
子处理单元 1632
核心
高速缓存
核心
高速缓存
子处理单元 1634
I/O 子系统
1608
处理加速单元 1606
通信子系统
1624
数据馈送
1626
事件流
1628
事件更新
1630
系统存储器
1610
应用程序 1612
程序数据 1614
操作系统 1616
计算机可读存储介质读取器
1620
计算机可读存储介质 1622
存储子系统
1618

图16