(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110945837 B
(45)授权公告日 2022.11.01

(21)申请号 201780093345.8
(22)申请日 2017.06.01
(65)同一申请的已公布的文献号
申请公布号 CN 110945837 A
(43)申请公布日 2020.03.31
(85)PCT国际申请进入国家阶段日
2020.01.19
(86)PCT国际申请的申请数据
PCT/IN2017/050216 2017.06.01
(87)PCT国际申请的公布数据
WO2018/220638 EN 2018.12.06
(73)专利权人 瑞典爱立信有限公司
地址 瑞典斯德哥尔摩
(72)发明人 F·克 V·K·茨赫
R·D·塔利科蒂
(74)专利代理机构 北京市路盛律师事务所 11326
专利代理师 李宓 陈静
(51)Int.Cl.
H04L 43/026(2022.01)
H04L 43/0852(2022.01)
H04L 43/0876(2022.01)
H04L 43/10(2022.01)
H04L 43/12(2022.01)
H04L 43/16(2022.01)
G06F 15/16(2006.01)

(56)对比文件
CN 103250392 A,2013.08.14
CN 105993149 A,2016.10.05
US 2017142034 A1,2017.05.18
US 2013010600 A1,2013.01.10
US 2017149640 A1,2017.05.25
US 2016285729 A1,2016.09.29
US 2017041209 A1,2017.02.09

审查员 陈斌

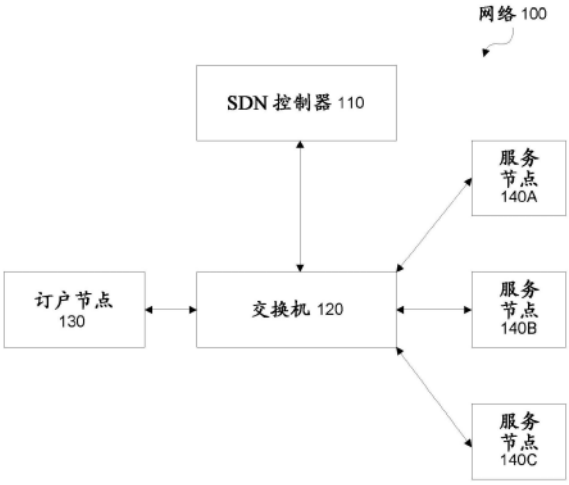
权利要求书3页 说明书20页 附图12页

(54)发明名称

优化SDN中的服务节点监视

(57)摘要

一种由软件定义网络(SDN)网络中的交换机实现的用于监视通信地耦合到所述交换机的服务节点的方法,所述方法包括:生成与从所述服务节点接收的分组相匹配的第一流条目;生成与从所述服务节点接收的分组相匹配的第二流条目,其中,所述第二流条目的优先级低于所述第一流条目的优先级;响应于确定所述第一流条目已超时,删除所述第一流条目并向SDN控制器发送流删除消息;维持与所述第二流条目相关联的统计;以及响应于确定与所述第二流条目相关联的所述统计超过阈值,向所述SDN控制器发送统计触发事件消息。



CN 110945837 B

1.一种由软件定义网络SDN网络中的交换机实现的用于监视通信地耦合到所述交换机的服务节点的方法,所述方法包括:

生成(710)与从所述服务节点接收的分组相匹配的第一流条目,其中所述第一流条目是故障检测流条目;

生成(720)与从所述服务节点接收的分组相匹配的第二流条目,其中,所述第二流条目是恢复检测流条目且所述第二流条目的优先级低于所述第一流条目的优先级;

响应于确定所述第一流条目已超时,删除(740)所述第一流条目并向SDN控制器发送(750)流删除消息,其中向所述SDN控制器发送所述流删除消息使得所述SDN控制器开始显式服务节点监视,其中,所述显式服务节点监视包括将显式服务节点监视流量注入所述SDN网络的数据平面中;

维持(755)与所述第二流条目相关联的统计;以及

响应于确定与所述第二流条目相关联的所述统计超过阈值,向所述SDN控制器发送(770)统计触发事件消息。

2.根据权利要求1所述的方法,其中,向所述SDN控制器发送所述统计触发事件消息使得所述SDN控制器停止显式服务节点监视。

3.根据权利要求1所述的方法,还包括:

在向所述SDN控制器发送所述统计触发事件消息之后,从所述SDN控制器接收用于重新生成所述第一流条目的指令;以及

根据从所述SDN控制器接收的所述指令,重新生成(780)所述第一流条目。

4.根据权利要求1所述的方法,其中,与所述第二流条目相关联的所述统计是与所述第二流条目相匹配的分组的分组计数。

5.根据权利要求1所述的方法,还包括:

生成与从所述服务节点接收的显式服务节点监视分组相匹配的第三流条目,其中,所述第三流条目的优先级高于所述第二流条目的优先级。

6.根据权利要求1所述的方法,其中,所述第一流条目和所述第二流条目包括用于将匹配的分组引导到服务节点流量处理流水线的指令。

7.根据权利要求1所述的方法,其中,所述第一流条目和所述第二流条目是在分组处理流水线的最前流表中生成的。

8.根据权利要求1所述的方法,其中,所述第一流条目包括空闲超时值的指示。

9.根据权利要求1所述的方法,其中,所述第二流条目包括用于所述统计的所述阈值的指示。

10.根据权利要求9所述的方法,其中,所述第二流条目包括所述统计触发事件消息将要在与所述第二流条目相关联的所述统计超过所述阈值的倍数时向所述SDN控制器发送的指示。

11.一种网络设备(904),被配置为充当软件定义网络SDN网络中的交换机以监视通信地耦合到所述交换机的服务节点,所述网络设备包括:

一组一个或多个处理器(942);以及

非暂时性机器可读存储介质(948),其中存储了服务节点监视组件(963),所述服务节点监视组件在由所述一组一个或多个处理器执行时使得所述网络设备:生成与从所述服务

节点接收的分组相匹配的第一流条目，其中所述第一流条目是故障检测流条目；生成与从所述服务节点接收的分组相匹配的第二流条目，其中，所述第二流条目是恢复检测流条目且所述第二流条目的优先级低于所述第一流条目的优先级；响应于确定所述第一流条目已超时，删除所述第一流条目并向SDN控制器发送流删除消息，其中向所述SDN控制器发送所述流删除消息使得所述SDN控制器开始显式服务节点监视，其中，所述显式服务节点监视包括将显式服务节点监视流量注入所述SDN网络的数据平面中；维持与所述第二流条目相关联的统计；以及响应于确定与所述第二流条目相关联的所述统计超过阈值，向所述SDN控制器发送统计触发事件消息。

12.根据权利要求11所述的网络设备，其中，向所述SDN控制器发送所述统计触发事件消息使得所述SDN控制器停止显式服务节点监视。

13.根据权利要求11所述的网络设备，其中，所述服务节点监视组件在由所述一组一个或多个处理器执行时还使得所述网络设备：在向所述SDN控制器发送所述统计触发事件消息之后，从所述SDN控制器接收用于重新生成所述第一流条目的指令；以及根据从所述SDN控制器接收的所述指令，重新生成所述第一流条目。

14.根据权利要求11所述的网络设备，其中，与所述第二流条目相关联的所述统计是与所述第二流条目相匹配的分组的分组计数。

15.一种非暂时性机器可读介质，在其中存储有计算机代码，所述计算机代码在由作为软件定义网络SDN网络中的交换机的网络设备的一组一个或多个处理器执行时使得所述网络设备执行用于监视通信地耦合到所述交换机的服务节点的操作，所述操作包括：

生成(710)与从所述服务节点接收的分组相匹配的第一流条目，其中所述第一流条目是故障检测流条目；

生成(720)与从所述服务节点接收的分组相匹配的第二流条目，其中，所述第二流条目是恢复检测流条目且所述第二流条目的优先级低于所述第一流条目的优先级；

响应于确定所述第一流条目已超时，删除(740)所述第一流条目并向SDN控制器发送(750)流删除消息，其中向所述SDN控制器发送所述流删除消息使得所述SDN控制器开始显式服务节点监视，其中，所述显式服务节点监视包括将显式服务节点监视流量注入所述SDN网络的数据平面中；

维持(755)与所述第二流条目相关联的统计；以及

响应于确定与所述第二流条目相关联的所述统计超过阈值，向所述SDN控制器发送(770)统计触发事件消息。

16.根据权利要求15所述的非暂时性机器可读介质，其中，所述计算机代码在由所述网络设备的所述一组一个或多个处理器执行时使得所述网络设备执行进一步操作，所述操作包括：

在向所述SDN控制器发送所述统计触发事件消息之后，从所述SDN控制器接收用于重新生成所述第一流条目的指令；以及

根据从所述SDN控制器接收的所述指令，重新生成(780)所述第一流条目。

17.根据权利要求15所述的非暂时性机器可读介质，其中，与所述第二流条目相关联的所述统计是与所述第二流条目相匹配的分组的分组计数。

18.根据权利要求15所述的非暂时性机器可读介质，其中，所述计算机代码在由所述网

络设备的所述一组一个或多个处理器执行时使得所述网络设备执行进一步操作,所述操作包括:

生成与从所述服务节点接收的显式服务节点监视分组相匹配的第三流条目,其中,所述第三流条目的优先级高于所述第二流条目的优先级。

优化SDN中的服务节点监视

技术领域

[0001] 本发明的实施例涉及计算机网络领域，并且更具体地，涉及优化软件定义网络(SDN)网络中的服务节点的监视。

背景技术

[0002] 软件定义网络(SDN)是一种计算机联网的方法，该方法采用其中转发平面(有时称为数据平面)与控制平面解耦的分离架构网络。使用分离架构网络通过将网络的智能转移到监督交换机的一个或多个控制器中而简化了实现转发平面的网络设备(例如交换机)。SDN通过提供可编程的网络基础设施而促进了在网络层的快速和开放式创新。

[0003] OpenFlow是使SDN网络中的控制器和交换机能够彼此通信的协议。OpenFlow使得能够对网络中的流控制策略进行动态编程。OpenFlow交换机包括可编程分组处理流水线(有时称为OpenFlow流水线)。OpenFlow流水线包括一个或多个流表，其中，每个流表包括一个或多个流条目。OpenFlow流水线的流表从零开始顺序编号。流水线处理通常在第一流表(例如流表0)处开始。当由流表处理时，将分组与流表的流条目进行匹配以选择流条目。如果找到流条目，则执行该流条目中包括的指令集。

[0004] 网络运营商可以利用SDN技术来提供服务链。通过服务链，订户流量被引导通过服务节点的预定义序列。服务节点提供各种网络服务，例如内容缓存、内容过滤以及与安全相关的服务。服务链使网络运营商能够以高度精细的方式基于预定义的策略(例如基于订户简档和应用简档)来引导流量。它还允许网络运营商方便快速地为订户引入新服务。

[0005] 在基于SDN的服务链的情况下，网络可以分成传输域和服务域。传输域通常包括转发流量的多个交换机(或其他类型的数据平面节点)，服务域通常包括应用服务的多个服务节点。服务节点通常经由虚拟局域网(VLAN)或类似机制来连接到传输域。为了避免长时间的服务中断，重要的是能够监视传输域中的交换机和服务域中的服务节点。

[0006] 可以使用诸如链路层发现协议(LLDP)或双向转发检测(BFD)之类的协议来监视交换机。但是，与交换机不同，服务节点更类似于服务器(并且通常不实现任何交换功能)。因此，用于监视交换机(例如使用LLDP或BFD)的相同技术可能不适用于监视服务节点。例如，LLDP可能不用于监视服务节点，因为服务节点通常不响应LLDP消息。由于这种差异，需要不同的技术来监视服务节点。

[0007] 一种用于监视服务节点的常规技术使用地址解析协议(ARP)。借助该技术，SDN控制器将交换机配置为将从SDN控制器接收的ARP请求消息转发到该交换机所连接的服务节点。SDN控制器还将交换机配置为将其从服务节点接收的任何ARP响应消息递送(punt)到SDN控制器。SDN控制器定期为服务节点生成ARP请求消息，并将ARP请求消息发送到交换机，以便交换机可以将ARP请求消息转发到服务节点。当交换机从服务节点接收到ARP响应消息时，交换机将ARP响应消息递送到SDN控制器。如果SDN控制器在特定时间间隔内未接收到与先前发送的ARP请求消息对应的ARP响应，则SDN控制器可以得出服务节点未运行的结论并相应地更新服务节点的状态。用于监视服务节点的另一种常规技术采用应用级监视，其中，

使用诸如超文本传输协议(HTTP)之类的协议来监视服务节点的活动性(liveliness)。借助这种技术,SDN控制器将HTTP请求发送到服务节点,并基于它是否从服务节点接收到响应来确定服务节点是否运行。但是,如上所述的常规技术具有以下缺点:它们不知道网络中的当前流量负载,使用恒定带宽,并且容易受到网络延迟的影响。

发明内容

[0008] 一种由软件定义网络(SDN)网络中的交换机实现的用于监视通信地耦合到所述交换机的服务节点的方法,所述方法包括:生成与从所述服务节点接收的分组相匹配的第一流条目;生成与从所述服务节点接收的分组相匹配的第二流条目,其中,所述第二流条目的优先级低于所述第一流条目的优先级;响应于确定所述第一流条目已超时,删除所述第一流条目并向SDN控制器发送流删除消息;维持与所述第二流条目相关联的统计;以及响应于确定与所述第二流条目相关联的所述统计超过阈值,向所述SDN控制器发送统计触发事件消息。

[0009] 一种被配置为充当软件定义网络(SDN)网络中的交换机以监视通信地耦合到所述交换机的服务节点的网络设备。所述网络设备包括一组一个或多个处理器以及其中存储有服务节点监视组件的非暂时性机器可读存储介质。所述服务节点监视组件在由所述一组一个或多个处理器执行时使得所述网络设备:生成与从所述服务节点接收的分组相匹配的第一流条目;生成与从所述服务节点接收的分组相匹配的第二流条目,其中,所述第二流条目的优先级低于所述第一流条目的优先级;响应于确定所述第一流条目已超时,删除所述第一流条目并向SDN控制器发送流删除消息;维持与所述第二流条目相关联的统计;以及响应于确定与所述第二流条目相关联的所述统计超过阈值,向所述SDN控制器发送统计触发事件消息。

[0010] 一种非暂时性机器可读介质,在其中存储有计算机代码,所述计算机代码在由作为软件定义网络(SDN)网络中的交换机的网络设备的一组一个或多个处理器执行时使得所述网络设备执行用于监视通信地耦合到所述交换机的服务节点的操作。所述操作包括:生成与从所述服务节点接收的分组相匹配的第一流条目;生成与从所述服务节点接收的分组相匹配的第二流条目,其中,所述第二流条目的优先级低于所述第一流条目的优先级;响应于确定所述第一流条目已超时,删除所述第一流条目并向SDN控制器发送流删除消息;维持与所述第二流条目相关联的统计;以及响应于确定与所述第二流条目相关联的所述统计超过阈值,向所述SDN控制器发送统计触发事件消息。

附图说明

[0011] 通过参考以下说明书和用于例示本发明实施例的附图,可以最好地理解本发明。在附图中:

[0012] 图1是根据一些实施例的可以在其中实现服务节点监视的网络的图;

[0013] 图2是示出根据一些实施例的将服务链应用于订户流量的交换机的图;

[0014] 图3是示出根据一些实施例的显式服务节点监视技术的图;

[0015] 图4是示出根据一些实施例的由SDN控制器配置交换机以监视服务节点的操作的图;

[0016] 图5是示出根据一些实施例的当故障检测流条目超时时由SDN控制器和交换机进行的操作的图；

[0017] 图6是示出根据一些实施例的当与恢复检测流条目相关联的统计超过阈值时由SDN控制器和交换机进行的操作的图；

[0018] 图7是根据一些实施例的用于监视SDN网络中的服务节点的过程的流程图；

[0019] 图8是根据一些实施例的用于监视SDN网络中的服务节点的过程的流程图；

[0020] 图9A示出了根据本发明一些实施例的示例性网络内的网络设备(ND)之间的连接性以及ND的三个示例性实现；

[0021] 图9B示出了根据本发明一些实施例的实现专用网络设备的示例性方式；

[0022] 图9C示出了根据本发明一些实施例的在其中可以耦合虚拟网元(VNE)的各种示例性方式；

[0023] 图9D示出了根据本发明一些实施例的在每个ND上具有单个网元(NE)的网络，并且在该直接方法内，将传统的分布式方法(通常由传统路由器使用)与用于维护可达性和转发信息(也称为网络控制)的集中式方法进行了对比；

[0024] 图9E示出了根据本发明一些实施例的其中每个ND都实现单个NE但是集中式控制平面已经将不同ND中的多个NE抽象成(表示)虚拟网络之一中的单个NE的简单情况；

[0025] 图9F示出了根据本发明一些实施例的其中多个VNE在不同ND上实现并相互耦合并且其中集中式控制平面已将这些多个VNE抽象化以使得它们在虚拟网络之一中表现为单个VNE的情况；

[0026] 图10示出了根据本发明一些实施例的具有集中式控制平面(CCP)软件1050的通用控制平面设备。

具体实施方式

[0027] 以下说明书描述了用于监视软件定义网络(SDN)网络中的服务节点的方法和装置。在下面的说明书中，阐述了许多特定的细节(例如逻辑实现、操作码、指定操作数的手段、资源分区/共享/复制实现、系统组件的类型和相互关系以及逻辑分区/集成选择)以提供对本发明的更透彻理解。然而，本领域的技术人员将理解，可以在没有这种具体细节的情况下实践本发明。在其他情况下，未详细示出控制结构、门级电路和完整的软件指令序列以免模糊本发明。本领域普通技术人员借助所包括的说明书将能够实现适当的功能而无需过度的实验。

[0028] 在说明书中对“一个实施例”、“实施例”、“示例实施例”等的引用指示所描述的实施例可以包括特定的特征、结构或特性，但是不是每个实施例都一定包括该特定的特征、结构或特性。而且，这样的短语不一定指代同一实施例。此外，当结合实施例描述特定的特征、结构或特性时，可以认为结合其他实施例(无论是否明确描述)实现这样的特征、结构或特性是本领域技术人员公知的。

[0029] 本文可以使用带有虚线边框(例如大虚线、小虚线、点划线和点)的方括号内的文本和块来示出向本发明的实施例添加附加特征的可选操作。然而，这种标示不应被认为意味着这些是唯一的选项或可选的操作，和/或在本发明的特定实施例中，带有实线边框的块不是可选的。

[0030] 在下面的说明书和权利要求书中，可以使用术语“耦合”和“连接”及其派生词。应该理解的是，这些术语并不旨在彼此等同。“耦合”用于指示两个或多个元件可以相互协作或交互，这两个或多个元件可以或可以不直接物理或电气接触。“连接”用于指示在彼此耦合的两个或多个元件之间建立通信。

[0031] 电子设备使用诸如机器可读存储介质(例如磁盘、光盘、固态驱动器、只读存储器(ROM)、闪存设备、相变存储器)和机器可读传输介质(也称为载波)(例如电、光、无线电、声音或其他形式的传播信号，例如载波、红外信号)的机器可读介质(也称为计算机可读介质)存储和传输(在内部和/或通过网络与其他电子设备)代码(其由软件指令组成并且有时称为计算机程序代码或计算机程序)和/或数据。因此，电子设备(例如计算机)包括硬件和软件，例如耦合到一个或多个机器可读存储介质以存储代码和/或存储数据的一组一个或多个处理器(例如，其中，处理器是微处理器、控制器、微控制器、中央处理单元、数字信号处理器、专用集成电路、现场可编程门阵列、其他电子电路、一个或多个以上的组合)，所述代码用于在该组处理器上执行。例如，电子设备可以包括包含代码的非易失性存储器，因为即使当电子设备被关闭(当电源被切断)时，该非易失性存储器也可以保持代码/数据，并且当电子设备被开启时，该电子设备的处理器要执行的那部分代码通常从较慢的非易失性存储器复制到该电子设备的易失性存储器(例如动态随机存取存储器(DRAM)、静态随机存取存储器(SRAM))中。典型的电子设备还包括一组一个或多个物理网络接口(NI)以建立与其他电子设备的网络连接(以使用传播信号来发送和/或接收代码和/或数据)。例如，该组物理NI(或该组物理NI与执行代码的该组处理器相结合)可以执行任何格式化、编码或转换，以允许电子设备在有线和/或无线连接上发送和接收数据。在一些实施例中，物理NI可以包括能够在无线连接上从其他电子设备接收数据和/或经由无线连接向其他设备发送数据的无线电电路。该无线电电路可以包括适合于射频通信的发射机、接收机和/或收发机。无线电电路可以将数字数据转换为具有适当参数(例如频率、定时、信道、带宽等)的无线电信号。然后可以经由天线将无线电信号发送到适当的接收者。在一些实施例中，该组物理NI可以包括网络接口控制器(NIC)(也称为网络接口卡)、网络适配器或局域网(LAN)适配器。NIC可以有助于将电子设备连接到其他电子设备，从而允许它们通过将电缆插入连接到NIC的物理端口来经由电线进行通信。可以使用软件、固件和/或硬件的不同组合来实现本发明的实施例的一个或多个部分。

[0032] 网络设备(ND)是通信地互连网络上的其他电子设备(例如其他网络设备、最终用户设备)的电子设备。一些网络设备是提供对多种联网功能(例如路由、桥接、交换、第2层聚合、会话边界控制、服务质量和/或订户管理)的支持和/或提供对多个应用服务(例如数据、语音和视频)的支持的“多种服务网络设备”。

[0033] 如上所述，用于监视SDN网络中的服务节点的常规技术具有以下缺点：它们不知道网络中的当前流量负载。随着网络的数据平面上的流量负载增加，控制平面上的流量负载也增加。这导致交换机消耗自己更多的计算资源，从而使交换机更加难于为较低优先级的任务(例如监视)保留计算资源。这导致在最需要的时候监视速度变慢(当流量负载较高时，故障检测的任何延迟都比流量负载较低时的消耗更大)。常规技术还具有使用恒定带宽的缺点。用于监视目的的带宽量通常是取决于所需的故障检测速度和网络的大小的恒定量。这意味着所使用的带宽量与网络中的当前流量负载无关。常规技术还具有易受网络延迟影

响的缺点。路径监视流量通常以与普通数据流量相同的服务类别发送。随着网络上流量负载的增加,路径监视流量开始经历排队延迟增加,这会使路径故障检测延迟。

[0034] 本文公开的实施例通过重用在交换机和服务节点之间发送的现有流量以监视服务节点并且仅当一段时间在交换机和服务节点之间没有流量流过时才诉诸显式服务节点监视技术,克服了传统技术的一些缺点。如本文所用,显式服务节点监视技术是指依赖于将监视流量注入网络的监视技术,其中,监视流量的唯一或主要目的是监视服务节点(而不是用于携带用户数据)(例如上述基于ARP的监视技术中的ARP消息)。注入到网络中的这种流量在本文中可以被称为显式服务节点监视流量(其可以包括显式服务节点监视分组)。实施例基于以下观察:当流量在交换机和服务节点之间流动时,无需将显式服务节点监视流量注入网络中,因为可以从交换机和服务节点之间流动的现有流量推断出服务节点可操作。如下面将更详细描述的,实施例可以通过在交换机处对一组特定流条目进行编程来实现此目的。从本文提供的公开中将显而易见的是,本文公开的实施例的一个优点在于,它们仅在交换机和服务节点之间没有流量发送时才诉诸于显式服务节点监视技术(其依赖于将显式服务节点监视流量注入网络中),并且因此可以减少了交换机处的带宽使用和处理负载。

[0035] 图1是根据一些实施例的可以在其中实现服务节点监视的网络的图。网络100(其是SDN网络)包括通信地耦合到交换机120的SDN控制器110。在一个实施例中,SDN控制器110和交换机120使用诸如OpenFlow(例如OpenFlow1.5)的南向(southbound)通信协议或类似的南向协议彼此通信。SDN控制器110可以使用OpenFlow或类似的南向协议来配置和管理交换机120的转发行为。为了清楚和容易理解,主要在SDN控制器110和交换机120将OpenFlow实现为南向通信协议的上下文中描述实施例。但是,应该理解,SDN控制器110和交换机120可以实现其他类型的南向通信协议,并且可以在SDN控制器110和交换机120实现其他类型的南向通信协议的上下文中实现本文公开的服务节点监视技术而不背离本公开的精神和范围。为了例示,网络100被示为包括管理单个交换机120的单个SDN控制器110。但是,应该理解,网络100可以包括一个以上的SDN控制器110,并且给定的SDN控制器110可以管理多个交换机120。

[0036] 交换机120可以包括分组处理流水线,该分组处理流水线包括一组流表。每个流表可以包括一组流条目,其中,每个流条目包括分组匹配标准(例如在匹配字段中携带的)和当分组与分组匹配标准相匹配时要执行的一组对应的指令。如果分组与流条目的分组匹配标准相匹配,则称该分组与该流条目相匹配。在一个实施例中,当交换机120接收到数据平面中的分组时,交换机120最初将分组与分组处理流水线的最前流表中的流条目进行匹配。交换机120然后可以继续将分组与分组处理流水线的后续流表中的流条目进行匹配。如果分组与流条目相匹配,则交换机120执行该流条目的一组对应指令。在一个实施例中,为流条目分配优先级。如果分组与流表中多个流条目的分组匹配标准相匹配,则仅选择具有最高优先级的流条目(并且执行其指令)。在流条目中指定的一组指令可以包括例如用于修改分组、将分组引导到分组处理流水线中的另一流表和/或将分组输出到指定端口的指令。

[0037] 交换机120还通信地耦合到订户节点130和多个服务节点140(例如服务节点140A-C)。订户节点130可以是访问网络100的订户设备,例如膝上型计算机、移动电话、智能电话、平板电脑、平板电话、互联网协议语音(VoIP)电话、用户设备、终端、便携式媒体播放器、游戏系统、机顶盒、或其任何组合。交换机120可以将由订户节点130生成的流量(在本文中可

以称为“订户流量”)转发到服务节点140中的一个或多个以将服务应用于该订户的流量。例如,每个服务节点140可以应用不同的服务,例如内容缓存、内容过滤和与安全有关的服务(例如深度分组检查(DPI))。服务节点140与交换机120的不同之处在于,它们更类似于服务器,因为它们通常不实现任何交换功能。在一个实施例中,服务节点140可以被实现为物理节点或虚拟节点(例如使用网络功能虚拟化(NFV))。在一个实施例中,订户节点130通过虚拟可扩展局域网(VxLAN)或类似机制连接到交换机120。在一个实施例中,交换机120通过虚拟局域网(VLAN)或类似机制连接到每个服务节点140。

[0038] 图2是示出根据一些实施例的将服务链应用于订户流量的交换机的图。如图所示,交换机120通过引导订户流量210A通过服务节点140的预定义序列而将服务链应用于订户流量210A。在该示例中,交换机120引导订户流量210A通过服务节点140A、服务节点140B和服务节点140C(按此顺序)。相比之下,交换机120不对订户流量210B应用服务链,因此订户流量210B不被引导通过服务节点140。

[0039] 图3是示出根据一些实施例的显式服务节点监视技术的图。显式服务节点监视技术是一种使用地址解析协议(ARP)以用于监视服务节点140的常规技术。利用该技术,SDN控制器110将交换机120配置为将从SDN控制器110接收的ARP请求消息转发到服务节点140。SDN控制器110还将交换机120配置为将其从服务节点140接收的任何ARP响应消息递送到SDN控制器110。随后,在操作1中,SDN控制器110将ARP请求消息发送给交换机120。在操作2中,交换机120将ARP请求消息转发到服务节点140。在操作3中,交换机120从服务节点140接收ARP响应消息。在操作4中,交换机120将ARP响应消息递送到SDN控制器110。如果SDN控制器110接收到ARP响应消息,则SDN控制器110可以推断服务节点140正在运行。然而,如果SDN控制器110在特定时间间隔内没有接收到与先前发送的ARP请求消息相对应的ARP响应消息,则SDN控制器110可以推断服务节点140未在运行,并相应地更新服务节点140的状态。诸如上述的传统技术具有以下缺点:它们不知道网络中的当前流量负载,使用恒定带宽,并且容易受到网络延迟的影响。

[0040] 图4是示出根据一些实施例的由SDN控制器配置交换机以监视服务节点的操作的图。如图所示,SDN控制器110将交换机120配置/编程为将流量转发到服务节点140并从服务节点140接收流量(例如作为服务链的一部分),而无需显式服务节点监视。在一个实施例中,SDN控制器110对交换机120中的一组特定流条目进行编程,以处理从服务节点140接收的流量并监视服务节点140。这些流条目可以包括用于检测服务节点140(或交换机120与服务节点140之间的通信路径)何时可能已发生故障的故障检测流条目以及用于检测服务节点140(或交换机120与服务节点140之间的通信路径)何时可能已从故障中恢复的恢复检测流条目。故障检测流条目和恢复检测流条目在下文中更详细地描述。

[0041] 故障检测流条目检测服务节点140何时可能已发生故障。在一个实施例中,故障检测流条目包括与从服务节点140接收的分组相匹配的分组匹配标准,以及用于将匹配的分组转发到服务节点流量处理流水线(用于正常处理来自服务节点140的流量,这可能例如涉及将分组引导到负责处理来自服务节点140的流量的另一流表)的指令。在一个实施例中,SDN控制器110利用超时机制对故障检测流条目进行编程,以使得如果在指定的时间段(例如等于或略小于显式服务节点监视技术发送显式服务节点监视流量的时间间隔)内没有分组匹配故障检测流条目,则故障检测流条目超时。在一个实施例中,如果故障检测流条目超

时，则交换机120删除故障检测流条目，并且向SDN控制器110发送指示故障检测流条目已经被删除的消息(该消息在本文中可以称为流删除消息)。

[0042] 表I中示出了故障检测流条目的示例。

[0043]

分组匹配标准	优先级	空闲超时	指令
服务节点端口、本地 IP、远程 IP	50	1 秒	向服务节点流量处理流水线转发

[0044] 表I

[0045] 表I中所示的故障检测流条目的分组匹配标准被设置为匹配从服务节点140接收的分组(由服务节点端口、本地互联网协议(IP)地址(例如服务节点140的IP地址)和远程IP地址(例如交换机120的IP地址)标识)。故障检测流条目的优先级被设置为50。故障检测流条目的空闲超时被设置为1秒，使得如果至少一秒没有分组与故障检测流条目相匹配，则该故障检测流条目超时。故障检测流条目的指令包括用于将匹配的分组转发到服务节点流量处理流水线的指令，应该理解，表1所示的故障检测流条目通过示例方式提供并且并非旨在限制。例如，在其他实施例中，分组匹配标准可以使用比所示的更多或更少的字段来匹配从服务节点140接收的分组。

[0046] 图5是示出根据一些实施例的当故障检测流条目超时时由SDN控制器和交换机进行的操作的图。当交换机120与服务节点140之间的通信路径可运行时，交换机120接收来自服务节点140的流量，并且属于该流量的分组将与交换机120中的故障检测流条目相匹配。结果，分组将被转发到服务节点流量处理流水线，并被相应地处理。但是，如图5所示，当交换机120与服务节点140之间的通信路径出现故障时，交换机120不会接收到来自服务节点140的流量。在这种情况下，如果通信路径未在指定的时间段内(例如在故障检测流条目中编程的空闲超时值内)恢复，则故障检测流条目超时。这导致交换机120删除故障检测流条目。结果，在操作1，交换机120将流删除消息发送到SDN控制器110。基于接收到流删除消息，SDN控制器110可以确定交换机120在至少指定的时间段内未从服务节点140接收到流量。作为响应，在操作2处，SDN控制器110可以开始显式服务节点监视，以确定服务节点140(或交换机120与服务节点140之间的通信路径)是否已发生故障，或者是否只是不存在来自服务节点140的流量。

[0047] 应该注意，当故障检测流条目超时时，它可以指示或者(1)服务节点140以及交换机120与服务节点140之间的通信路径可运行，但是服务节点140没有要发送到交换机120的流量或者(2)服务节点140或交换机120与服务节点140之间的通信路径发生故障。当故障检测流条目超时时，可以使用显式服务节点监视技术来确认故障检测流条目超时的原因。在由于服务节点140没有任何流量要发送到交换机120而导致故障检测流条目超时的情况下，可以允许显式服务节点监视继续，直到服务节点140有一些流量要发送到交换机120为止，这可以由恢复检测流条目来检测。

[0048] 恢复检测流条目检测服务节点140和/或交换机120与服务节点140之间的通信路径何时可能已从故障中恢复。在一个实施例中，恢复检测流条目包括与故障检测流条目相同的分组匹配标准和相同的指令。即，恢复检测流条目包括与从服务节点140接收的分组相匹配的分组匹配标准以及用于将匹配的分组转发到服务节点流量处理流水线的指令。另

外，在一个实施例中，恢复检测流条目包括在本文中称为统计触发指令(例如OpenFlow中的OFPIT_STAT_TRIGGER指令)的指令。统计触发指令指示交换机120维持与恢复检测流条目相关联的统计并且在与恢复检测流条目相关联的统计超过阈值时向SDN控制器110发送消息(该消息在本文可以称为统计触发事件消息)。在一个实施例中，恢复检测流条目包括阈值的指示。例如，阈值可以被指示为分组的数量，在这种情况下，交换机120跟踪与恢复检测流条目相匹配的分组的数量，并且当与恢复检测流条目相匹配的分组数量超过了指定的分组数量时将统计触发事件消息发送到SDN控制器110。作为另一示例，该阈值可以被指示为字节计数，在这种情况下，交换机120跟踪与恢复检测流条目匹配的分组的累积字节计数，并且当与恢复检测流条目相匹配的分组的累积字节计数超过指定的字节计数时向SDN控制器110发送统计触发事件消息。在一个实施例中，统计触发指令包括当与恢复检测流条目相关联的统计超过阈值的任何倍数时(例如OpenFlow中的OSPSTF_PERIODIC标志设置)将向SDN控制器110发送统计触发事件消息的指示。

[0049] 在一个实施例中，恢复检测流条目的优先级低于对应的故障检测流条目的优先级。这样，当交换机120包括故障检测流条目和恢复检测流条目两者时，从服务节点140接收的分组可以匹配故障检测流条目和恢复检测流条目两者的分组匹配标准，但是交换机120仅执行故障检测流条目的指令，因为它具有更高的优先级。然而，在故障检测流条目被删除之后(例如由于故障检测流条目超时)，从服务节点140接收的分组与恢复检测流条目相匹配，并且交换机120执行恢复检测流条目的指令(包括统计触发指令)。

[0050] 表II中示出了恢复检测流条目(与表I中所示的故障检测流条目相对应)的示例。

[0051]

分组匹配标准	优先级	空闲超时	指令
服务节点端口、本地 IP、远程 IP	5		向服务节点流量处理流水线转发； STAT_TRIGGER （阈值＝1 个分组）

[0052] 表II

[0053] 表II中所示的恢复检测流条目的分组匹配标准被设置为匹配从服务节点140接收的分组(由服务节点端口、本地互联网协议(IP)地址和远程IP地址来标识)。该分组匹配标准通常被设置为与对应的故障检测流条目的分组匹配标准相同。恢复检测流条目的优先级被设置为5，其低于对应的故障检测流条目的优先级(其被设置为50)。在本示例中，较高的数字指示较高的优先级(而较低的数字指示较低的优先级)，但是应当理解，可以使用不同的约定。恢复检测流条目的指令包括用于将匹配的分组转发到服务节点流量处理流水线的指令。另外，这些指令包括指示如果至少一个分组与恢复检测流条目匹配(阈值＝1个分组)则交换机120向SDN控制器110发送统计触发事件消息的统计触发指令(STAT_TRIGGER)。在此示例中，恢复检测流条目不具有空闲超时(它没有超时)。应当理解，表II中所示的恢复检测流条目是通过示例方式提供的并且并非旨在进行限制。例如，在其他实施例中，分组匹配标准可以使用比所示的更多或更少的字段来匹配从服务节点140接收的分组。

[0054] 在一个实施例中，统计触发指令使用以下结构和字段：

[0055]
```
/* Instruction structure for OFPIT_STAT_TRIGGER */
struct ofp_instruction_stat_trigger {
        uint16_t type; /* OFPIT_STAT_TRIGGER */
        uint16_t len; /* Length is padded to 64 bits. */
        uint32_t flags; /* Bitmap of OFPSTF_* flags. */
        struct ofp_stats thresholds; /* Threshold list. Variable size. */
};
OFP_ASSERT(sizeof(struct ofp_instruction_stat_trigger) == 16);
```

[0056] Flags(标志)字段是定义统计触发的行为的比特图(bitmap)。它可以包括以下flags的组合：

[0057]
```
enum ofp_stat_trigger_flags {
        OFPSTF_PERIODIC = 1 << 0, /* Trigger for all multiples of thresholds. */
        OFPSTF_ONLY_FIRST = 1 << 1, /* Trigger on only first reach threshold. */
};
```

[0058] 当设置了OFPSTF_PERIODIC标志时，触发不仅将应用于阈值字段中的值，还将应用于这些值的所有倍数。例如，允许在流的生命周期中每100个分组有一个触发。当设置了OFPSTF_ONLY_FIRST标志时，仅考虑被越过的第一阈值，而忽略其他阈值。

[0059] 图6是示出根据一些实施例的当与恢复检测流条目相关联的统计超过阈值时由SDN控制器和交换机进行的操作的图。继续参考图5描述的示例，当服务节点140或交换机120与服务节点140之间的通信路径从故障中恢复并且交换机120再次开始从服务节点140接收流量时，属于从服务节点140接收的流量的分组将与交换机120中的恢复检测流条目相匹配(因为故障检测流条目已经被删除)。结果，分组将被转发到服务节点流量处理流水线，并被相应地处理。另外，如图6所示，如果与恢复检测流条目相关联的统计超过阈值(或其倍数)，则在操作1中，交换机120将统计触发事件消息发送到SDN控制器110。基于接收到统计触发事件消息，SDN控制器110可以确定交换机120再次从服务节点140接收流量，并且服务节点140或交换机120与服务节点140之间的通信路径已经从故障中恢复。作为响应，在操作2，SDN控制器110可以停止显式服务节点监视，并且在操作3，对交换机120中的故障检测流条目进行重新编程(以检测将来的故障)。

[0060] 在一个实施例中，SDN控制器110对交换机120中的流条目进行编程以处理显式服务节点监视流量。这样的流条目在本文中可以被称为显式服务节点监视流条目。在一个实施例中，显式服务节点监视流条目包括与从服务节点140接收的显式服务节点监视分组(例如ARP消息)相匹配的分组匹配标准，以及用于将匹配的分组转发到服务节点监视流水线(以用于正常的显式服务节点监视处理)的指令。在一个实施例中，该流条目的优先级高于恢复检测流条目的优先级，以使显式服务节点监视分组不触发交换机120向SDN控制器110发送统计触发事件消息(例如基于执行恢复检测流条目的指令)。

[0061] 表III中示出了显式服务节点监视流条目的示例。

[0062]

分组匹配标准	优先级	空闲超时	指令
服务节点端口、本地IP、远程IP、监视协议类型	10		向服务节点监视流水线转发

[0063] 表III

[0064] 表III中所示的显式服务节点监视流条目的分组匹配标准被设置为匹配从服务节点140接收的具有显式服务节点监视协议类型(例如ARP消息)的分组(由服务节点端口、本地互联网协议(IP)地址和远程IP地址来标识)。显式服务节点监视流条目的优先级被设置为10,该优先级高于相应的恢复检测流条目的优先级(表II中所示)。显式服务节点监视流条目的指令包括用于将匹配的分组转发到服务节点监视流水线的指令。在此示例中,显式服务节点监视流条目不具有空闲超时(它没有超时)。应该理解,表III中所示的服务节点监视流条目是通过示例的方式提供的并且并非旨在进行限制。例如,在其他实施例中,分组匹配标准可以使用比所示的更多或更少的字段来匹配从服务节点140接收的分组。

[0065] 当执行显式服务节点监视时,从服务节点140接收的显式服务节点监视分组(例如ARP消息)可以与显式服务节点监视流条目和恢复检测流条目两者的分组匹配标准相匹配,但是交换机120仅执行显式服务节点监视流条目的指令,因为它具有更高的优先级。这允许交换机120处理显式服务节点监视分组而无需触发统计触发事件消息。

[0066] 图7是根据一些实施例的用于监视SDN网络中的服务节点的过程的流程图。在一个实施例中,该过程由充当SDN网络100中的交换机120的网络设备执行。将参考其他附图的示例性实施例来描述该流程图和其他流程图中的操作。然而,应当理解,流程图的操作可以由本发明的实施例而不是参考其他附图所讨论的实施例执行,并且参考这些其他附图所讨论的本发明的实施例可以执行与参考所述流程图讨论的那些操作不同的操作。

[0067] 在一个实施例中,该过程由交换机120生成第一流条目(例如故障检测流条目)(框710)和第二流条目(例如恢复检测流条目)(框720)发起。第一流条目和第二流条目都可以具有与从服务节点140接收的分组相匹配的分组匹配标准,但是第二流条目的优先级低于第一流条目的优先级。这样,如果交换机120接收到与两个流条目的分组匹配标准都匹配的分组,则交换机120执行第一流条目的指令。在一个实施例中,第一流条目和第二流条目在分组处理流水线的最前流表(例如OpenFlow中的表0)中生成。在一个实施例中,第一流条目和第二流条目包括用于将匹配的分组引导到服务节点流量处理流水线的指令。在一个实施例中,如果在给定的时间段内没有分组与第一流条目相匹配,则第一流条目超时。在一个实施例中,第一流条目包括空闲超时值的指示(例如指示第一流条目超时会花费多长时间的值)。如果交换机120确定第一流条目已经超时(决策框730),则交换机120删除第一流条目(框740),并将流删除消息发送到SDN控制器110(框750)以指示第一流条目已被删除。这可以向SDN控制器110指示服务节点140或交换机120与服务节点140之间的通信路径可能已发生故障并且促使SDN控制器110开始显式服务节点监视,其中,显式服务节点监视涉及将显式服务节点监视流量注入SDN网络100的数据平面中。交换机120然后可以执行显式服务节点监视(例如通过将从SDN控制器110接收的ARP消息转发到服务节点140,并且将从SDN控制器110接收的ARP消息转发到SDN控制器110)(框752)。返回到决策框730,如果第一流条目未超时,则交换机120继续正常的分组处理,直到第一流条目超时为止。

[0068] 一旦第一流条目已被删除，从服务节点140接收的分组将与第二流条目进行匹配。交换机120维持与第二流条目相关联的统计(例如基于在第二流条目中包括的统计触发指令)(框755)。交换机120确定与第二流条目相关联的统计是否超过阈值(决策框760)。在一个实施例中，与第二流条目相关联的统计是与第二流条目相匹配的分组的分组计数。在另一个实施例中，与第二流条目相关联的统计是与第二流条目相匹配的分组的累积字节计数。在一个实施例中，第二流条目包括阈值的指示。如果与第二流条目相关联的统计未超过阈值，则交换机120继续正常的分组处理，直到与第二流条目相关联的统计超过阈值为止。如果交换机120确定与第二流条目相关联的统计已经超过阈值，则交换机120将统计触发事件消息发送到SDN控制器110(框770)。这可以向SDN控制器110指示服务节点140或交换机120与服务节点140之间的通信路径可能已经从故障中恢复并且促使SDN控制器110停止显式服务节点监视，这也促使交换机120停止执行显式服务节点监视(框775)。在一个实施例中，第二流条目包括统计触发事件消息将要在与第二流条目相关联的统计超过阈值的倍数(例如OpenFlow中的OFPSTF_PERIODIC标志设置)时向SDN控制器110发送的指示。在该实施例中，每当与第二流条目相关联的统计超过阈值的倍数(例如每个分组或每10,000字节)时，交换机120就向SDN控制器110发送统计触发事件消息。在一个实施例中，在交换机120将统计触发事件消息发送到SDN控制器110之后，交换机120从SDN控制器110接收用于重新生成第一流条目的指令。交换机120然后可以重新生成第一流条目(框780)(例如根据从SDN控制器110接收的指令)。

[0069] 在一个实施例中，交换机120生成与从服务节点140接收的显式服务节点监视分组(例如ARP消息)相匹配的第三流条目(例如显式服务节点监视流条目)。在一个实施例中，第三流条目的优先级高于第二流条目的优先级。这允许交换机120处理显式服务节点监视分组而无需触发统计触发事件消息。

[0070] 图8是根据一些实施例的用于监视SDN网络中的服务节点的过程的流程图。在一个实施例中，该过程由充当SDN网络100中的SDN控制器110的网络设备执行。

[0071] 在一个实施例中，该过程由SDN控制器110向交换机120发送用于生成与从服务节点140接收的分组相匹配的第一流条目的指令而发起(框810)。SDN控制器110还向交换机120发送用于生成与从服务节点140接收的分组相匹配的第二流条目的指令，其中，第二流条目的优先级低于第一流条目的优先级(框820)。在一个实施例中，第一流条目是故障检测流条目，第二流条目是对应的恢复检测流条目。SDN控制器110可随后从交换机120接收指示第一流条目已被删除的流删除消息(框830)。这可以指示服务节点140或交换机120与服务节点140之间的通信路径可能已发生了故障。作为响应，SDN控制器110可以开始显式服务节点监视(框840)。

[0072] 随后，SDN控制器110可以从交换机120接收指示与第二流条目相关联的统计超过阈值的统计触发事件消息(框850)。这可以指示服务节点140或交换机120与服务节点140之间的通信路径可能已经从故障中恢复。作为响应，SDN控制器110可以停止显式服务节点监视(框860)，并且向交换机120发送用于重新生成第一流条目的指令(框870)。

[0073] 本文公开的实施例的优点在于，它们仅在没有从服务节点140接收到流量时才诉诸显式服务节点监视技术，因此能够减少交换机120处的带宽使用和处理负载。随着服务节点140的数量增加，该优点变得更加明显。本文公开的实施例的另一个优点是它们避免了通

常由控制器驱动的监视所涉及的许多延迟，其中，控制器->交换机->服务节点->交换机->控制器的路径易于发生延迟。由于控制器->交换机和交换机->控制器的路径位于网络的控制平面上，因此它们构成了延迟的主要部分。本文公开的实施例的另一优点是它们不需要任何专有扩展(至少对于本文公开的OpenFlow实现而言)。根据本文提供的公开内容，其他优点对于本领域普通技术人员将是显而易见的。

[0074] 图9A示出了根据本发明的一些实施例的示例性网络内的网络设备(ND)之间的连接性以及ND的三个示例性实现。图9A示出了ND 900A-H及其在900A-900B、900B-900C、900C-900D、900D-900E、900E-900F、900F-900G和900A-900G之间以及在900H与900A、900C、900D和900G中的每一个之间的线路连接。这些ND是物理设备，并且这些ND之间的连接可以是无线的也可以是有线的(通常称为链路)。从ND 900A、900E和900F延伸的附加线路例示了这些ND充当网络的入口点和出口点(并且因此，这些ND有时称为边缘ND;而其他ND可以称为核心ND)。

[0075] 图9A中的两个示例性ND实现是:1)使用定制的专用集成电路(ASIC)和专用操作系统(OS)的专用网络设备902;以及2)使用通用现成(COTS)处理器和标准OS的通用网络设备904。

[0076] 专用网络设备902包括联网硬件910，联网硬件910包括一组一个或多个处理器912、转发资源914(其通常包括一个或多个ASIC和/或网络处理器)和物理网络接口(NI)916(通过其进行网络连接，例如ND 900A-H之间的连接所示)、以及在其中存储网络软件920的非暂时性机器可读存储介质918。在运行期间，联网软件920可以由联网硬件910执行以实例化一组一个或多个网络软件实例922。每个联网软件实例922以及联网硬件910中执行该网络软件实例的部分(是专用于该联网软件实例的硬件和/或该联网软件实例与其他联网软件实例922在时间上共享的硬件的时间片)形成单独的虚拟网元930A-R。每个虚拟网元(VNE)930A-R均包括控制通信和配置模块932A-R(有时称为本地控制模块或控制通信模块)和转发表934A-R，使得给定的虚拟网元(例如930A)包括该控制通信和配置模块(例如932A)、一组一个或多个转发表(例如934A)以及联网硬件910的执行虚拟网元(例如930A)的部分。

[0077] 在一个实施例中，软件920包括诸如服务节点监视组件925之类的代码，该代码当由联网硬件910执行时使专用网络设备902执行作为联网软件实例922的一部分的本发明的一个或多个实施例的操作。

[0078] 专用网络设备902通常在物理和/或逻辑上被认为包括:1)ND控制平面924(有时称为控制平面)，其包括执行控制通信和配置模块932A-R的处理器912;以及2)ND转发平面926(有时称为转发平面、数据平面或媒体平面)，其包括利用转发表934A-R和物理NI 916的转发资源914。举例来说，在ND是路由器(或正在实现路由功能)的情况下，ND控制平面924(执行控制通信和配置模块932A-R的处理器912)通常负责参与控制如何路由数据(例如分组)(例如数据的下一跳和该数据的出站(outgoing)物理NI)并将该路由信息存储在转发表934A-R中，ND转发平面926负责在物理NI 916上接收该数据并基于转发表934A-R来将该数据转发出适当的物理NI 916。

[0079] 图9B示出了根据本发明的一些实施例的实现专用网络设备902的示例性方式。图9B示出了包括卡938(通常是可热插拔的)的专用网络设备。尽管在一些实施例中，卡938具

有两种类型(一种或多种用作ND转发平面926(有时称为线卡),一种或多种用来实现ND控制平面924(有时称为控制卡)),都是备选实施例可以将功能组合到单个卡上和/或包括附加卡类型(例如一种附加类型的卡称为服务卡、资源卡或多应用卡)。服务卡可以提供专门的处理(例如第4层到第7层服务(例如防火墙、互联网协议安全性(IPsec)、安全套接字层(SSL)/传输层安全性(TLS)、入侵检测系统(IDS)、点对点(P2P)、IP语音(VoIP)会话边界控制器、移动无线网关(网关通用分组无线电服务(GPRS)支持节点(GGSN)、演进分组核心(EPC)网关)。通过示例,可以使用服务卡来终止IPsec隧道并执行伴随的认证和加密算法。这些卡通过一个或多个互连机制(如背板936所示)耦合在一起(例如耦合线卡的第一全网状、耦合所有卡的第二全网状)。

[0080] 返回图9A,通用网络设备904包括硬件940,硬件940包括一组一个或多个处理器942(通常是COTS处理器)和物理NI 946、以及在其中存储有软件950的非暂时性机器可读存储948。在运行期间,处理器942执行软件950以实例化一组或多组一个或多个应用964A-R。尽管一个实施例不实现虚拟化,但是备选实施例可以使用不同形式的虚拟化。例如,在一个这样的备选实施例中,虚拟化层954表示操作系统的内核(或在基本操作系统上执行的填充程序(shim)),其允许创建称为软件容器的多个实例962A-R,每个实例可用于执行一组(或多组)应用964A-R;其中多个软件容器(也称为虚拟化引擎、虚拟专用服务器或jail)是用户空间(通常是虚拟内存空间),这些用户空间彼此分离并与运行操作系统的内核空间分离;并且其中,除非明确允许,否则在给定用户空间中运行的一组应用无法访问其他进程的内存。在另一个这样的备选实施例中,虚拟化层954表示系统管理程序(hypervisor)(有时称为虚拟机监视器(VMM))或在主机操作系统之上执行的系统管理程序,并且多组应用964A-R中的每一者在称为在管理程序之上运行的虚拟机(其在某些情况下可被视为软件容器的紧密隔离的形式)的实例962A-R内的来宾操作系统(guest operating system)之上运行,来宾操作系统和应用可能不知道它们正在虚拟机上运行而不是在“裸机”主机电子设备上运行,或者通过半虚拟化,操作系统和/或应用可能出于优化目的而意识到存在虚拟化。在其他备选实施例中,一个、一些或所有应用被实现为一个或多个单内核,单内核可以通过用应用仅直接编译一组有限的库(例如来自包括OS服务的驱动程序/库的库操作系统(LibOS))来生成,该组有限的库提供该应用所需的特定OS服务。由于可以将单内核实现为直接在硬件940上运行、直接在系统管理程序上运行(在这种情况下,有时将单内核描述为在LibOS虚拟机中运行)、或者在软件容器中运行,所以实施例可以完全通过在由虚拟化层954表示的系统管理程序上直接运行的单内核、通过在实例962A-R表示的软件容器内运行的单内核来实现,或实现为单内核和上述技术的组合(例如都直接在系统管理程序上运行的单内核和虚拟机,在不同软件容器中运行的单内核和多组应用)。

[0081] 一组或多组一个或多个应用964A-R的实例化以及虚拟化(如果实现的话)统称为软件实例952。每组应用964A-R、对应的虚拟化构造(例如实例962A-R)(如果实现的话)、以及硬件940的执行它们的那部分(假设是专用于该执行的硬件和/或时间上共享的硬件的时间片)形成单独的虚拟网元960A-R。

[0082] 虚拟网元960A-R执行与虚拟网元930A-R类似的功能,例如,类似于控制通信和配置模块932A以及转发表934A(硬件940的这种虚拟化有时被称为网络功能虚拟化(NFV)。因此,NFV可用于将许多网络设备类型整合到行业标准的大容量服务器硬件、物理交换机和物

理存储设备上，大容量服务器硬件、物理交换机和物理存储设备可以位于数据中心、ND和客户驻地设备(CPE)中。尽管通过与一个VNE 960A-R对应的每个实例962A-R例示了本发明的实施例，但是备选实施例可以以更精细级别的粒度实现此对应关系(例如线卡虚拟机虚拟化线卡，控制卡虚拟机虚拟化控制卡等)；应当理解，本文中参考实例962A-R到VNE的对应关系所描述的技术也适用于使用这种更精细级别的粒度和/或单内核的实施例。

[0083] 在特定实施例中，虚拟化层954包括提供与物理以太网交换机类似的转发服务的虚拟交换机。具体地，该虚拟交换机在实例962A-R与物理NI 946之间以及可选地在实例962A-R之间转发流量。另外，该虚拟交换机可以在按照策略不被允许彼此通信(例如通过遵守虚拟局域网(VLAN))的VNE 960A-R之间实施网络隔离。

[0084] 在一个实施例中，软件950包括诸如服务节点监视组件963之类的代码，该代码当由处理器942执行时使通用网络设备904执行本发明的一个或多个实施例的操作，作为软件实例962A-R的一部分。

[0085] 图9A中的第三示例性ND实现是混合网络设备906，其在单个ND或ND中的单个卡中包括定制ASIC/专用OS和COTS处理器/标准OS两者。在这样的混合网络设备的特定实施例中，平台VM(即，实现专用网络设备902的功能的VM)可以为混合网络设备906中存在的联网硬件提供半虚拟化。

[0086] 不管ND的上述示例性实现如何，当考虑由ND实现的多个VNE中的单个VNE(例如VNE中仅一个VNE是给定虚拟网络的一部分)时，或者仅单个VNE当前由ND实现，简称网元(NE)有时用于指代该VNE。同样在所有上述示例性实现中，每个VNE(例如VNE 930A-R、VNE 960A-R和混合网络设备906中的那些VNE)在物理NI(例如916、946)上接收数据，然后将该数据转发出适当的物理NI(例如916、946)。例如，实现IP路由器功能的VNE基于IP分组中的某些IP头信息来转发IP分组；其中IP头信息包括源IP地址、目的地IP地址、源端口、目的地端口(其中“源端口”和“目的地端口”在本文中是指协议端口，与ND的物理端口相对)、传输协议(例如用户数据报协议(UDP)、传输控制协议(TCP)和差分服务代码点(DSCP)值)。

[0087] 图9C示出了根据本发明的一些实施例的在其中可以耦合VNE的各种示例性方式。图9C示出了在ND 900A中实现的VNE 970A.1-970A.P(以及可选地VNE 970A.Q-970A.R)和在ND 900H中实现的VNE 970H.1。在图9C中，VNE 970A.1-P彼此分离，因为它们可以接收来自ND 900A外部的分组并将分组转发到ND 900A外部；VNE 970A.1与VNE 970H.1耦合，并且因此它们在各自的ND之间传送分组；VNE 970A.2-970A.3可以可选地在它们自身之间转发分组，而无需将分组转发到ND 900A之外；以及VNE 970A.P可以可选地是包括VNE 970A.Q(后跟VNE 970A.R)的VNE链中的第一个(这有时称为动态服务链，其中，VNE系列中的每个VNE都提供不同的服务，例如一个或多个第4-7层网络服务)。尽管图9C示出了VNE之间的各种示例性关系，但是备选实施例可以支持其他关系(例如更多/更少的VNE、更多/更少的动态服务链、具有一些公共VNE和一些不同VNE的多个不同的动态服务链)。

[0088] 例如，图9A的ND可以形成互联网或专用网络的一部分；其他电子设备(未示出；例如最终用户设备，包括工作站、笔记本电脑、上网本、平板电脑、掌上电脑、手机、智能电话、平板手机、多媒体电话、互联网协议语音(VOIP)电话、终端、便携式媒体播放器、GPS单元、可穿戴设备、游戏系统、机顶盒、支持互联网的家用电器)可以耦合到网络(直接或通过诸如接入网络的其他网络)以在网络(例如，互联网或覆盖(例如通过隧道化)互联网的虚拟专用网

络(VPN))上彼此通信(直接或通过服务器)和/或访问内容和/或服务。这样的内容和/或服务通常由属于服务/内容提供商的一个或多个服务器(未示出)或参与点到点(P2P)服务的一个或多个最终用户设备(未示出)提供,并且可以包括例如公共网页(例如免费内容、店面、搜索服务)、私有网页(例如提供电子邮件服务的用户名/密码访问的网页)和/或VPN上的公司网络。例如,最终用户设备可以被耦合(例如通过耦合到接入网络(有线或无线地)的客户驻地设备)到边缘ND,边缘ND被耦合(例如通过一个或多个核心ND)到其他边缘ND,其他边缘ND被耦合到充当服务器的电子设备。但是,通过计算和存储虚拟化,在图9A中作为ND运行的一个或多个电子设备也可以托管一个或多个这样的服务器(例如在通用网络设备904的情况下,一个或多个软件实例962A-R可以充当服务器;混合网络设备906也是如此;在专用网络设备902的情况下,一个或多个这样的服务器也可以在由处理器912执行的虚拟化层上运行);在这种情况下,认为服务器与该ND的VNE共址(co-located)。

[0089] 虚拟网络是提供网络服务(例如L2和/或L3服务)的物理网络(例如图9A中的物理网络)的逻辑抽象。虚拟网络可以被实现为覆盖网络(有时称为网络虚拟化覆盖),该覆盖网络在底层网络(例如L3网络,诸如使用隧道(例如通用路由封装(GRE)、第2层隧道协议(L2TP)、IPSec)创建覆盖网络的互联网协议(IP)网络)上提供网络服务(例如第2层(L2,数据链路层)和/或第3层(L3,网络层)服务)。

[0090] 网络虚拟化边缘(NVE)位于底层网络的边缘,并参与实现网络虚拟化;NVE的面向网络侧使用底层网络在其他NVE之间来回隧道传输帧;NVE的朝外的一侧与网络外部的系统之间来回收发数据。虚拟网络实例(VNI)是NVE上虚拟网络的特定实例(例如ND上的NE/VNE、ND上NE/VNE的一部分,其中,NE/VNE通过仿真被分为多个VNE);可以在NVE上实例化一个或多个VNI(例如作为ND上的不同VNE)。虚拟接入点(VAP)是NVE上用于将外部系统连接到虚拟网络的逻辑连接点;VAP可以是通过逻辑接口标识符(例如VLAN ID)标识的物理或虚拟端口。

[0091] 网络服务的示例包括:1)以太网LAN仿真服务(类似于互联网工程任务组(IETF)多协议标签交换(MPLS)或以太网VPN(EVPN)服务的基于以太网的多点服务),其中,外部系统通过LAN环境在底层网络上跨网络互连(例如NVE为不同的此类虚拟网络提供单独的L2 VNI(虚拟交换实例),并跨底层网络提供L3(例如IP/MPLS)隧道封装);以及2)虚拟化IP转发服务(从服务定义的角度来看类似于IETF IP VPN(例如边界网关协议(BGP)/MPLS IPVPN)),其中,外部系统在底层网络上通过L3环境跨网络互连(例如NVE为不同的此类虚拟网络提供单独的L3 VNI(转发和路由实例),并跨底层网络提供L3(例如IP/MPLS)隧道封装)。网络服务还可以包括服务质量功能(例如流量分类标记、流量调节和调度)、安全功能(例如用于保护客户驻地免受网络发起的攻击以避免格式错误的路由公告的过滤器)以及管理功能(例如完整的检测和处理)。

[0092] 图9D示出了根据本发明的一些实施例在图9A的每个ND上具有单个网元的网络,并且在该直接转发方法内,将传统的分布式方法(通常由传统路由器使用)与用于维护可达性和转发信息(也称为网络控制)的集中式方法进行了对比。具体地,图9D示出了具有与图9A的ND 900A-H相同的连接性的网元(NE)970A-H。

[0093] 图9D示出了分布式方法972跨NE 970A-H分布用于生成可达性和转发信息的责任;换句话说,邻居发现和拓扑发现的过程是分布式的。

[0094] 例如,如果使用专用网络设备902,ND控制平面924的控制通信和配置模块932A-R通常包括可达性和转发信息模块以实现与其他NE通信以交换路由的一个或多个路由协议(例如外部网关协议(诸如边界网关协议(BGP))、内部网关协议(IGP)(例如开放式最短路径优先(OSPF)、中间系统到中间系统(IS-IS)、路由信息协议(RIP)、标签分发协议(LDP)、资源保留协议(RSVP)(包括RSVP流量工程(TE):用于LSP隧道的RSVP扩展和通用多协议标签交换(GMPLS)信令RSVP-TE)),然后基于一个或多个路由度量选择这些路由。因此,NE 970A-H(例如执行控制通信和配置模块932A-R的处理器912)通过分布式地确定网络内的可达性并计算其各自的转发信息来完成它们参与控制如何路由数据(例如分组)(例如数据的下一跳和该数据的传出物理NI)的责任。路由和邻接被存储在ND控制平面924上的一个或多个路由结构(例如路由信息库(RIB)、标签信息库(LIB)、一个或多个邻接结构)中。ND控制平面924基于路由结构来使用信息(例如邻接和路由信息)对ND转发平面926进行编程。例如,ND控制平面924将邻接和路由信息编程到ND转发平面926上的一个或多个转发表934A-R(例如转发信息库(FIB)、标签转发信息库(LFIB)和一个或多个邻接结构)中。对于第2层转发,ND可以存储一个或多个桥接表,桥接表用于基于数据中的第2层信息来转发该数据。尽管上面的示例使用了专用网络设备902,但是可以在通用网络设备904和混合网络设备906上实现相同的分布式方法972。

[0095] 图9D示出了一种使系统解偶的集中式方法974(也称为软件定义网络(SDN)),该系统对从将流量转发到所选目的地的底层系统发送流量的位置进行决策。所示的集中式方法974负责在集中式控制平面976(有时称为SDN控制模块、控制器、网络控制器、OpenFlow控制器、SDN控制器、控制平面节点、网络虚拟化机构或管理控制实体)中生成可达性和转发信息,并且因此邻居发现和拓扑发现的过程是集中式的。集中式控制平面976具有与包括NE 970A-H(有时称为交换机、转发元件、数据平面元件或节点)的数据平面980(有时称为基础设施层、网络转发平面或转发平面(其不应与ND转发平面混淆))的南向接口982。集中式控制平面976包括网络控制器978,网络控制器978包括集中式可达性和转发信息模块979,集中式可达性和转发信息模块979确定网络内的可达性并将转发信息在南向接口982(其可以使用OpenFlow协议)上分发给数据平面980的NE 970A-H。因此,网络智能被集中在通常与ND分离的电子设备上执行的集中式控制平面976中。在一个实施例中,网络控制器978包括服务节点监视组件981,服务节点监视组件981在由网络控制器978执行时使网络控制器978执行本发明的一个或多个实施例的操作。

[0096] 例如,在数据平面980中使用专用网络设备902的情况下,ND控制平面924的每个控制通信和配置模块932A-R通常包括提供南向接口982的VNE侧的控制代理。在这种情况下,ND控制平面924(执行控制通信和配置模块932A-R的处理器912)通过与集中式控制平面976通信以从集中式可达性和转发信息模块979接收转发信息(以及在某些情况下,可达性信息)的控制代理来执行其参与控制如何路由数据(例如分组)(例如数据的下一跳和该数据的出站物理NI)的责任(应当理解,在本发明的一些实施例中,除了与集中式控制平面976通信之外,控制通信和配置模块932A-R还可以在确定可达性和/或计算转发信息方面起某些作用,尽管比分布式方法的情况下的作用要少;这样的实施例通常被认为属于集中式方法974,但是也可以被认为是混合方法。)

[0097] 尽管以上示例使用专用网络设备902,但是可以用通用网络设备904实现相同的集

中式方法974(例如,每个VNE 960A-R通过与集中式控制平面976进行通信以从集中式可达性和转发信息模块979接收转发信息(以及在某些情况下可达性信息)来完成其控制如何路由数据(例如分组)(例如数据的下一跳和该数据的出站物理NI)的责任);应当理解,在本发明的一些实施例中,除了与集中式控制平面976通信外,VNE 960A-R还可以在确定可达性和/或计算转发信息中起一定作用,尽管比分布式方法的情况下的作用要少)和混合网络设备906。实际上,SDN技术的使用能够增强通常在通用网络设备904或混合网络设备906实现中使用的NFV技术,因为NFV能够通过提供可在其上运行SDN软件的基础设施来支持SDN,并且NFV和SDN都旨在利用商品服务器硬件和物理交换机。

[0098] 图9D还示出了集中式控制平面976具有到其中驻留有应用988的应用层986的北向接口984。集中式控制平面976具有形成用于应用988的虚拟网络992(有时被称为逻辑转发平面、网络服务或覆盖网络(具有作为底层网络的数据平面980的NE 970A-H)的能力。因此,集中式控制平面976维护所有ND和所配置的NE/VNE的全局视图,并有效地将虚拟网络映射到底层ND(包括在物理网络通过硬件(ND、链路或ND组件)故障、添加或删除而变化时维护这些映射)。

[0099] 尽管图9D示出了与集中式方法974分离的分布式方法972,但是在本发明的某些实施例中,网络控制的工作可以不同地分布,或将两者结合。例如:1)实施例通常可以使用集中式方法(SDN)974,但是具有委托给NE的特定功能(例如分布式方法可以用于实现故障监视、性能监视、保护切换和用于邻居和/或拓扑发现的原语(primitives)中的一者或多者);或2)本发明的实施例可以经由集中式控制平面和分布式协议二者执行邻居发现和拓扑发现,并且比较结果以在它们不一致时引发异常。这样的实施例通常被认为属于集中式方法974,但是也可以被认为是混合方法。

[0100] 尽管图9D示出了每个ND 900A-H实现单个NE 970A-H的简单情况,但是应当理解,参考图9D所描述的网络控制方法也适用于其中一个或多个ND 900A-H实现多个VNE(例如VNE 930A-R、VNE 960A-R、混合网络设备906中的那些VNE)的网络。备选地或附加地,网络控制器978也可以在单个ND中仿真多个VNE的实现。具体地,代替(或除了)在单个ND中实现多个VNE之外,网络控制器978还可以将单个ND中的VNE/NE的实现呈现为虚拟网络992中的多个VNE(全部在同一虚拟网络992中、每个都在不同的虚拟网络992中、或某种组合)。例如,网络控制器978可以使ND在底层网络中实现单个VNE(NE),然后在集中式控制平面976内在逻辑上划分该NE的资源以在虚拟网络992中呈现不同的VNE(其中覆盖网络中的这些不同的VNE正在共享底层网络中的ND上的单VNE/NE实现的资源)。

[0101] 另一方面,图9E和9F分别示出了网络控制器978可以作为不同虚拟网络992的一部分而呈现的NE和VNE的示例性抽象。图9E示出了根据本发明的一些实施例的其中每个ND 900A-H实现单个NE 970A-H(参见图9D)但是集中式控制平面976已将不同ND中的多个NE(NE 970A-C和G-H)抽象成(表示为)图9D的虚拟网络992之一中的单个NE 970I的简单情况。图9E示出了在该虚拟网络中,NE 970I耦合至NE 970D和970F,NE 970D和970F两者仍然耦合至NE 970E。

[0102] 图9F示出了根据本发明的一些实施例的其中多个VNE(VNE 970A.1和VNE 970H.1)在不同的ND(ND 900A和ND 900H)上实现并且彼此耦合并且其中集中式控制平面976已经抽象这多个VNE以使得它们在图9D的虚拟网络992之一中表现为单个VNE 970T的情况。因此,

NE或VNE的抽象可以跨越多个ND。

[0103] 尽管本发明的一些实施例将集中式控制平面976实现为单个实体(例如在单个电子设备上运行的软件的单个实例),但是备选实施例可以将功能分散在多个实体上以实现冗余和/或可伸缩性目的(例如在不同电子设备上运行的软件的多个实例)。

[0104] 类似于网络设备实现,运行集中式控制平面976的电子设备以及因此包括集中式可达性和转发信息模块979的网络控制器978可以通过多种方式(例如专用设备、通用(例如COTS)设备或混合设备)实现。这些电子设备将类似地包括处理器、一组一个或多个物理NI以及在其上存储了集中式控制平面软件的非暂时性机器可读存储介质。例如,图10示出了通用控制平面设备1004,其包括硬件1040,硬件1040包括一组一个或多个处理器1042(其通常是COTS处理器)和物理NI 1046,以及其中存储有集中式控制平面(CCP)软件1050和服务节点监视组件1051的非暂时性机器可读存储介质1048。

[0105] 在使用计算虚拟化的实施例中,处理器1042通常执行软件以实例化虚拟化层1054(例如在一个实施例中,虚拟化层1054表示操作系统的内核(或在基本操作系统上执行的填充程序(shim)),其允许创建多个实例1062A-R,实例1062A-R称为均可用于执行一组一个或多个应用的软件容器(表示单独的用户空间,也称为虚拟化引擎、虚拟专用服务器或jail);在另一个实施例中,虚拟化层1054表示系统管理程序(有时称为虚拟机监视器(VMM))或在主机操作系统之上执行的系统管理程序,并且应用在称为由系统管理程序运行的虚拟机(其在某些情况下可被视为软件容器的紧密隔离的形式)的实例1062A-R内的来宾操作系统之上运行;在另一个实施例中,应用被实现为单内核,单内核可以通过用应用直接编译仅一组有限的库(例如来自包括OS服务的驱动程序/库的库操作系统(LibOS))来生成,该组有限的库提供该应用所需的特定OS服务,并且该单内核可以直接在硬件1040上运行、在由虚拟化层1054表示的系统管理程序上直接运行(在这种情况下,有时将单内核描述为在LibOS虚拟机中运行)、或者在由实例1062A-R之一表示的软件容器中运行。再次地,在使用计算虚拟化的实施例中,在操作期间,在虚拟化层1054上执行(例如在实例1062A内)CCP软件1050的实例(图示为CCP实例1076A)。在不使用计算虚拟化的实施例中,CCP实例1076A在“裸机”通用控制平面设备1004上作为单内核或在主机操作系统之上执行。CCP实例1076A以及虚拟化层1054和实例1062A-R(如果已实现的话)的实例化统称为软件实例1052。

[0106] 在一些实施例中,CCP实例1076A包括网络控制器实例1078。网络控制器实例1078包括集中式可达性和转发信息模块实例1079(其是向操作系统提供网络控制器978的上下文并与各种NE进行通信的中间件层),以及中间件层(提供各种网络操作所需的智能,例如协议、网络态势感知和用户界面)之上的CCP应用层1080(有时称为应用层)。在更抽象的级别,集中式控制平面976内的CCP应用层1080与虚拟网络视图(网络的逻辑视图)一起工作,并且中间件层提供从虚拟网络到物理视图的转换。

[0107] 服务节点监视组件1051可以由硬件1040执行以执行作为软件实例1052的一部分的本发明的一个或多个实施例的操作。

[0108] 集中式控制平面976基于CCP应用层1080计算和针对每个流的中间件层映射,将相关消息发送到数据平面980。流可以被定义为其报头与给定的比特模式相匹配的一组分组;从这个意义上讲,传统的IP转发也是基于流的转发,其中,流由例如目的地IP地址来定义;然而,在其他实现中,用于流定义的给定比特模式可以在分组报头中包括更多字段(例如10

个或更多)。数据平面980的不同ND/NE/VNE可以接收不同的消息以及因此不同的转发信息。数据平面980处理这些消息并在适当NE/VNE的转发表(有时称为流表)中编程适当的流信息和对应的动作,然后NE/VNE将入站分组映射到在转发表中表示的流,并基于转发表中的匹配来转发分组。

[0109] 诸如OpenFlow之类的标准定义了用于消息的协议以及用于处理分组的模型。用于处理分组的模型包括报头解析、分组分类和做出转发决策。报头解析描述了如何基于一组众所周知的协议来解释分组。一些协议字段用于构建将在分组分类中使用的匹配结构(或键)(例如第一键字段可以是源媒体访问控制(MAC)地址,第二键字段可以是目的地MAC地址)。

[0110] 分组分类涉及通过基于转发表条目的匹配结构或键来确定转发表中的哪个条目(也称为转发表条目或流条目)与分组最匹配,从而在存储器中执行查找以对分组进行分类。转发表条目中表示的许多流可能与一个分组相对应/匹配;在这种情况下,系统通常被配置为根据定义的方案(例如选择匹配的第一转发表条目)从多个转发表条目中确定一个转发表条目。转发表条目包括一组特定匹配标准(一组值或通配符、或应该将分组的哪些部分与特定值/多个值/通配符进行比较的指示,如由匹配功能所定义的,针对分组报头中的特定字段,或用于某些其他分组内容)以及供数据平面在接收到匹配的分组时采取的一组一个或多个动作。例如,对于使用特定端口的分组,动作可以是将报头推送到该分组上、对该分组进行洪泛或简单地丢弃该分组。因此,用于具有特定传输控制协议(TCP)目的地端口的IPv4/IPv6分组的转发表条目可以包含指定应丢弃这些分组的动作。

[0111] 基于分组分类期间标识的转发表条目,通过执行在分组上的匹配的转发表条目中标识的一组动作,来做出转发决策并执行动作。

[0112] 然而,当未知分组(例如在OpenFlow用语中使用的“未命中分组”或“匹配未命中”)到达数据平面980时,该分组(或分组报头和内容的子集)通常被转发到集中式控制平面976。集中式控制平面976然后将转发表条目编程到数据平面980中,以容纳属于未知分组的流的分组。一旦特定的转发表条目已经由集中式控制平面976编程到数据平面980中,则具有匹配证书的下一个分组将匹配该转发表条目并采取与该匹配条目相关联的一组动作。

[0113] 网络接口(NI)可以是物理的或虚拟的;在IP的上下文中,接口地址是分配给NI(无论是物理NI还是虚拟NI)的IP地址。虚拟NI可以与物理NI相关联,与另一个虚拟接口相关联,或者可以独立存在(例如环回接口、点对点协议接口)。NI(物理或虚拟)可以被编号(带有IP地址的NI)或不编号(没有IP地址的NI)。环回接口(及其环回地址)是经常用于管理目的的NE/VNE(物理或虚拟)的特定类型的虚拟NI(和IP地址);其中这样的IP地址称为节点环回地址。分配给ND的NI的IP地址称为该ND的IP地址;在更细粒度的级别上,分配给NI(其被分配给在ND上实现的NE/VNE)的IP地址可以称为该NE/VNE的IP地址。

[0114] 已经根据计算机存储器内的数据比特上的事务的算法和符号表示来呈现了前述详细描述的某些部分。这些算法描述和表示是数据处理领域技术人员用来最有效地向本领域其他技术人员传达其工作实质的方式。在此,算法通常被认为是导致期望结果的事务的自洽序列。这些事务是需要对物理量进行物理操纵的事务。通常,尽管不是必须的,这些量采取能够被存储、传输、组合、比较和以其他方式操纵的电或磁信号的形式。主要出于通用目的,已经证明有时将这些信号称为比特、值、元素、符号、字符、术语、数字等是方便的。

[0115] 然而，应当牢记，所有这些和类似术语均应与适当的物理量相关联，并且仅仅是应用于这些量的方便标签。除非从上面的讨论明显另有说明，否则应理解，在整个描述中，利用诸如“处理”或“运算”或“计算”或“确定”或“显示”等术语的讨论是指计算机系统或类似电子计算设备的动作和过程，该计算机系统或类似电子计算设备将表示为计算机系统的寄存器和内存内的物理(电子)量的数据转换为其他类似表示为计算机系统内存或寄存器或其他此类信息存储、传输或显示设备内的物理量的数据。

[0116] 本文提出的算法和显示并不固有地与任何特定计算机或其他装置有关。各种通用系统可以与根据本文的教导的程序一起使用，或者可以证明构造更专用的装置以执行所需的方法事务很方便。从上面的描述中将得出各种这些系统所需的结构。另外，没有参考任何特定的编程语言来描述本发明的实施例。要理解，可以使用多种编程语言来实现如本文所述的本发明的实施例的教导。

[0117] 本发明的实施例可以是一种制品，其中，非暂时性机器可读介质(例如微电子存储器)在其上存储了对一个或多个数据处理组件(这里通常称为“处理器”)进行编程来执行上述操作的指令。在其他实施例中，这些操作中的一些可以由包含硬连线逻辑的特定硬件组件(例如专用数字滤波器块和状态机)执行。这些操作可以备选地通过编程的数据处理组件和固定的硬连线电路组件的任意组合来执行。

[0118] 在前述说明书中，已经参考本发明的特定示例性实施例描述了本发明的实施例。显而易见的是，在不脱离所附权利要求书所阐述的本发明的更广泛精神和范围的情况下，可以对其进行各种修改。因此，说明书和附图应被认为是说明性而不是限制性的。

[0119] 在整个说明书中，已经通过流程图示出了本发明的实施例。将理解的是，在这些流程图中描述的事务和事务的顺序仅旨在用于说明的目的，而并非旨在限制本发明。本领域普通技术人员将认识到，可以在不脱离如所附权利要求书所阐述的本发明的更广泛精神和范围的情况下对流程图进行改变。

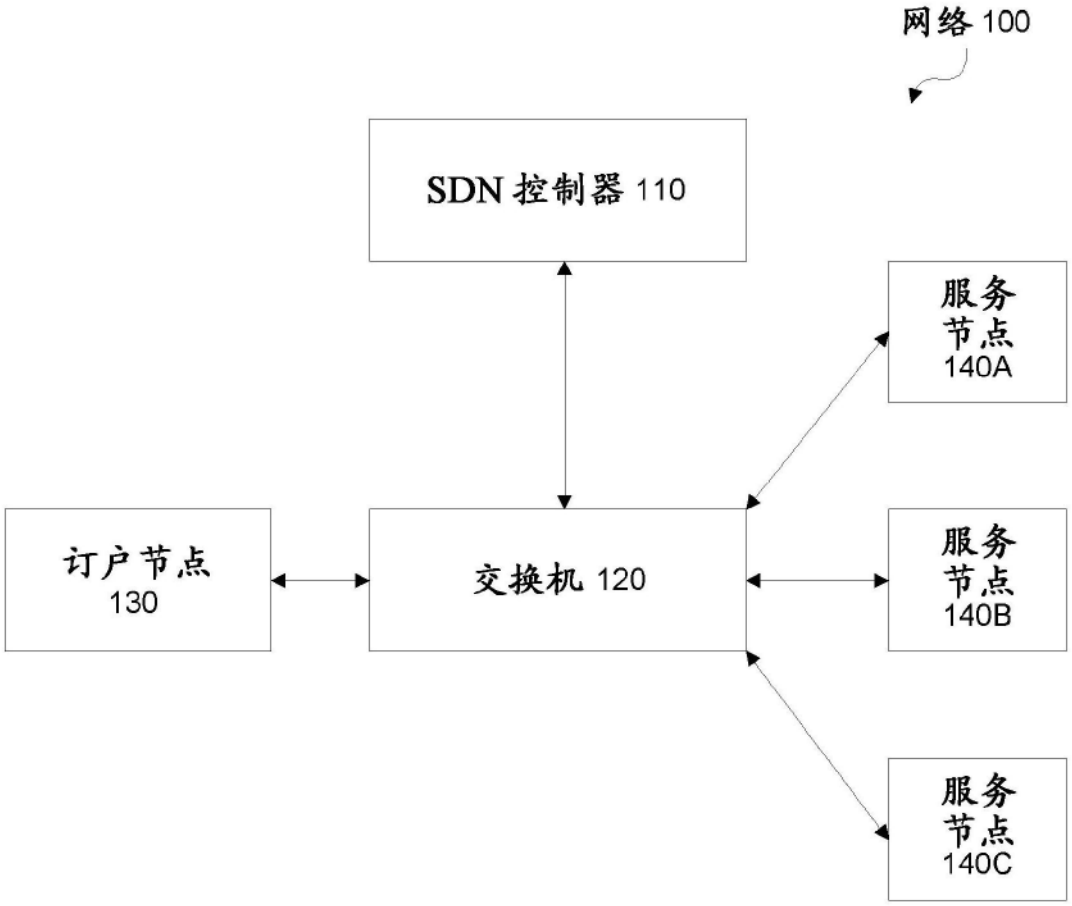

网络 100
SDN 控制器 110
服务节点 140A
订户节点 130
交换机 120
服务节点 140B
服务节点 140C

图1

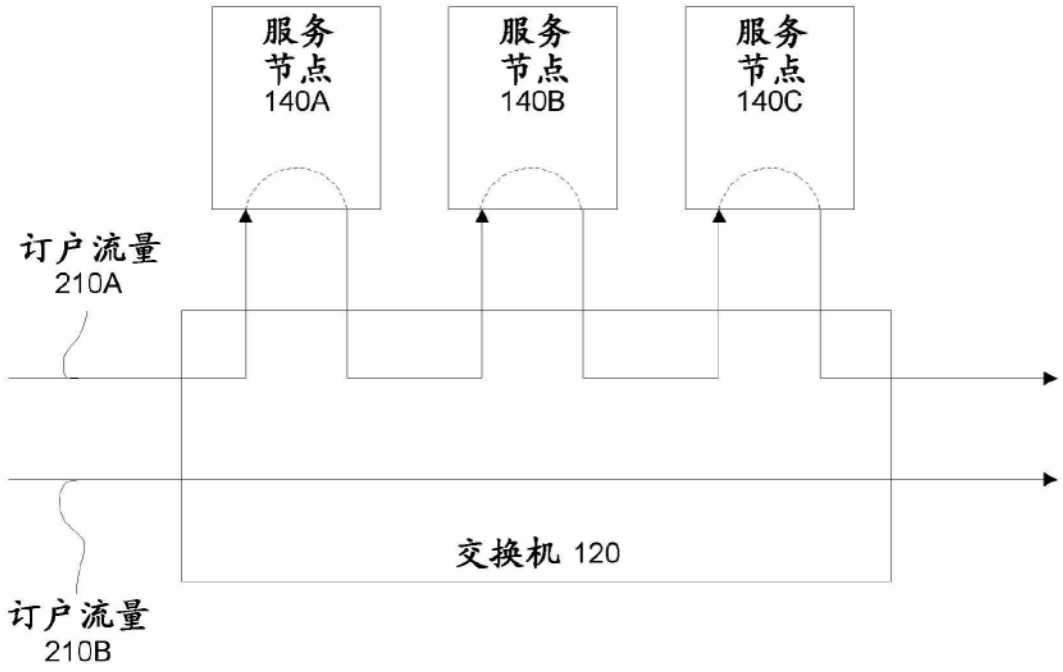

服务节点 140A
服务节点 140B
服务节点 140C
订户流量 210A
交换机 120
订户流量 210B

图2

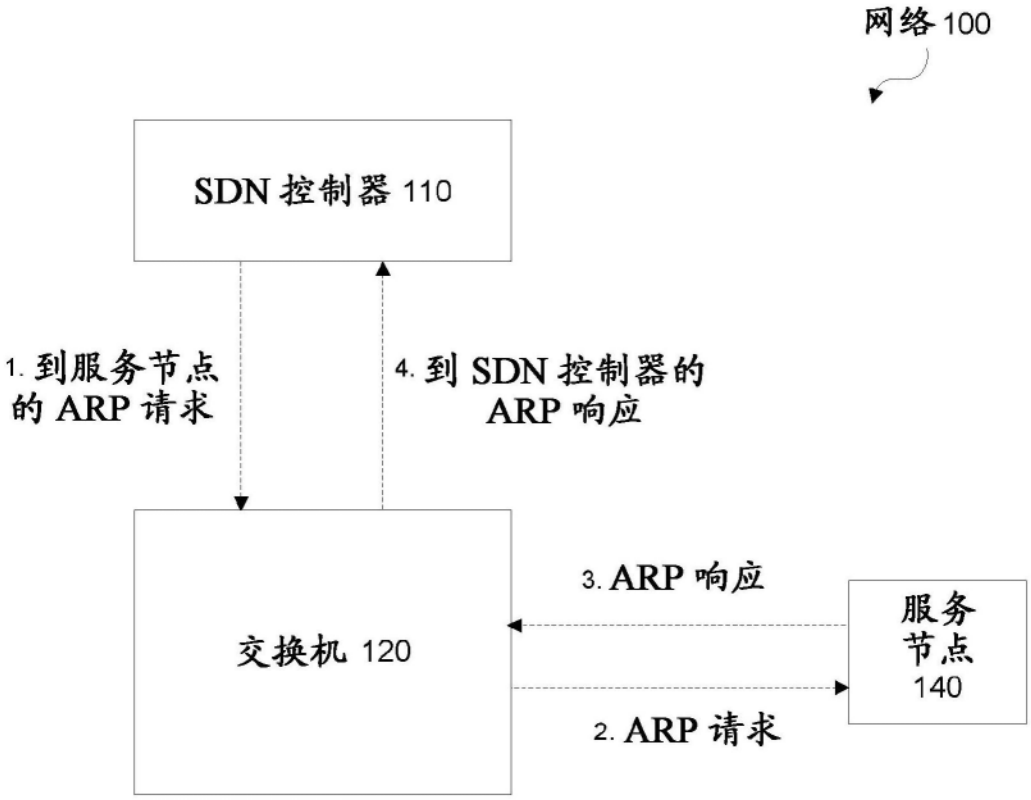

网络 100
SDN 控制器 110
1. 到服务节点的 ARP 请求
4. 到 SDN 控制器的 ARP 响应
交换机 120
3. ARP 响应
2. ARP 请求
服务节点 140

图3

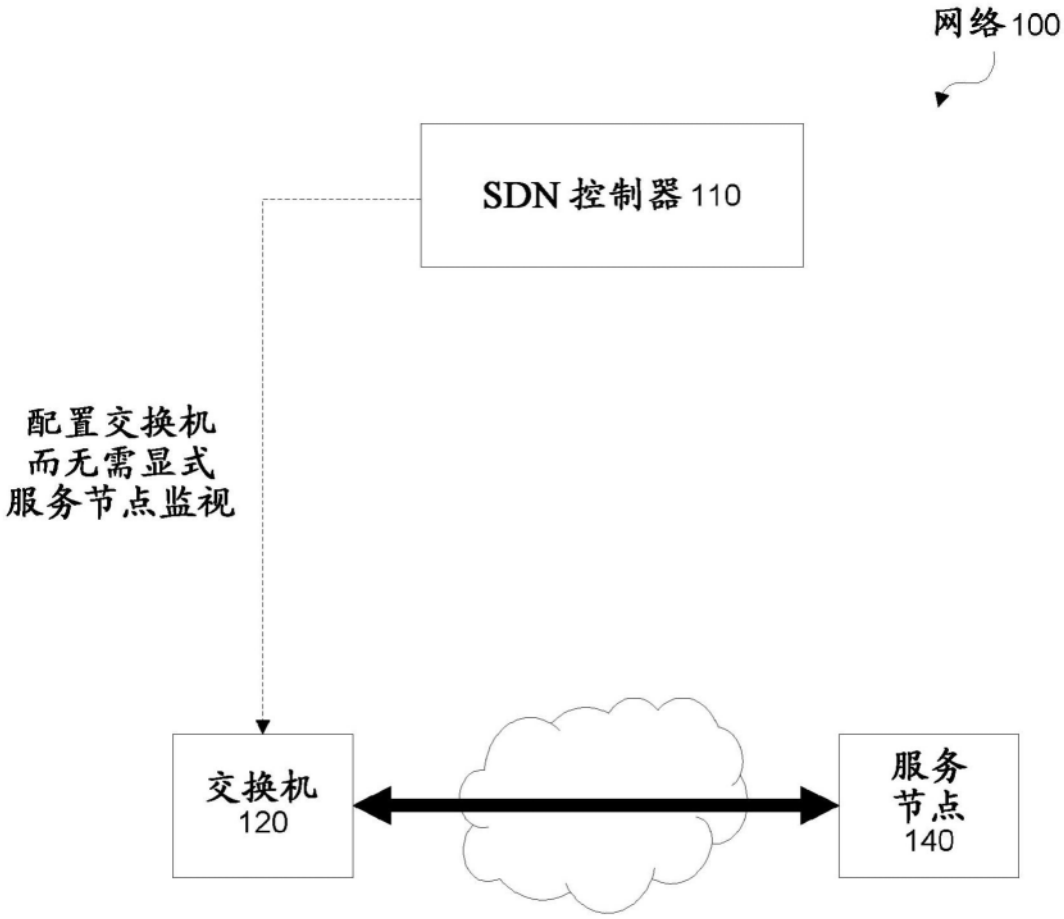


图4

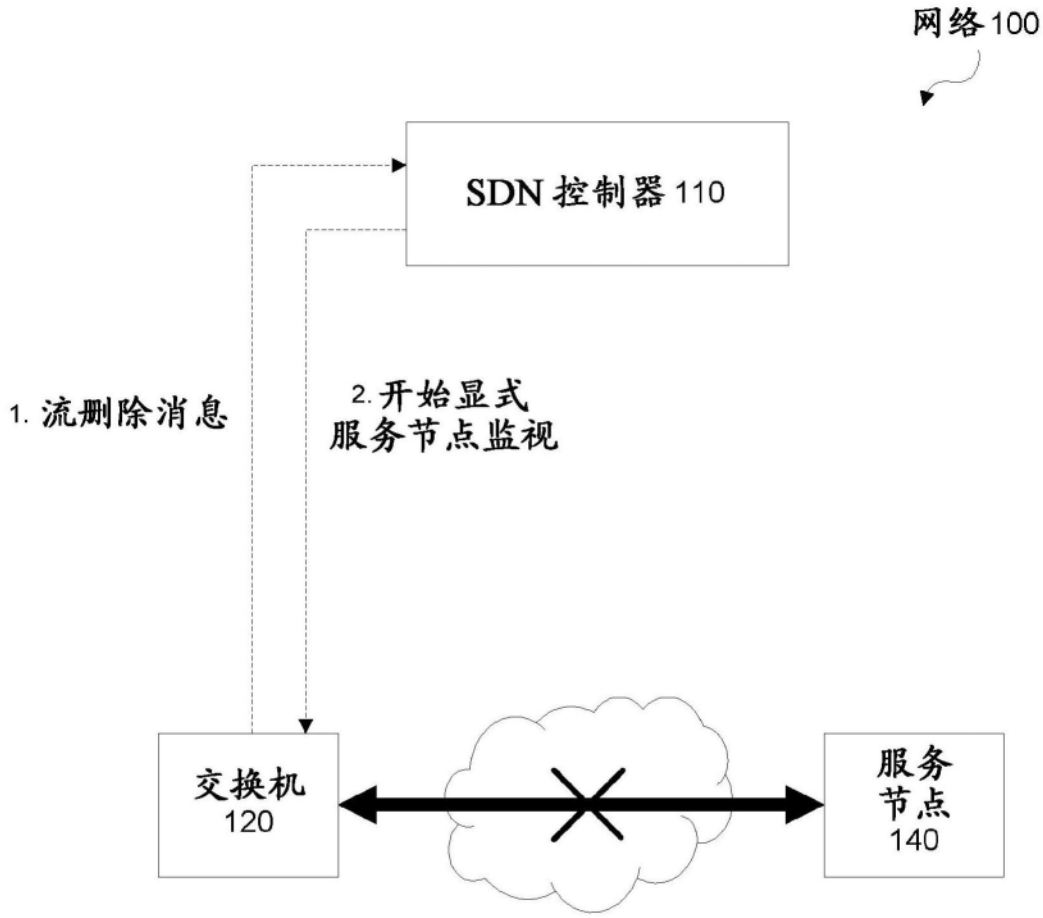


图5

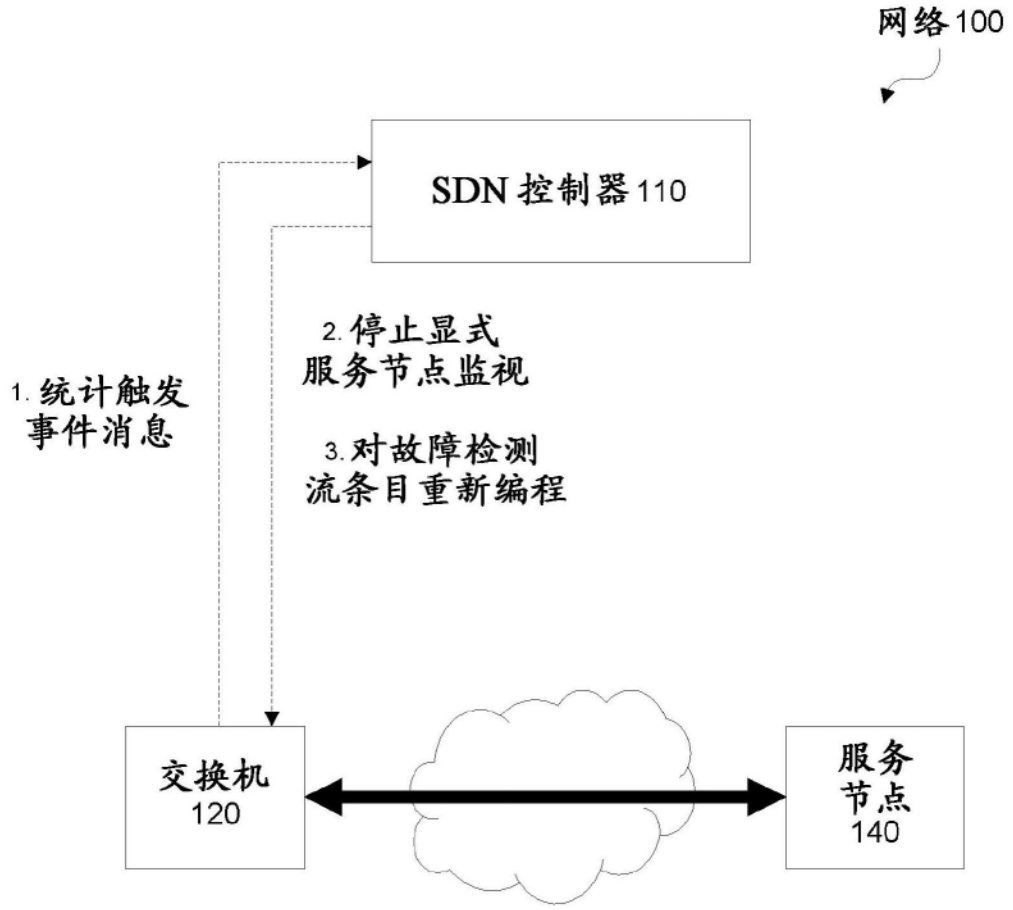

网络100
SDN 控制器110
2. 停止显式
服务节点监视
1. 统计触发
事件消息
3. 对故障检测
流条目重新编程
交换机
120
服务
节点
140

图6

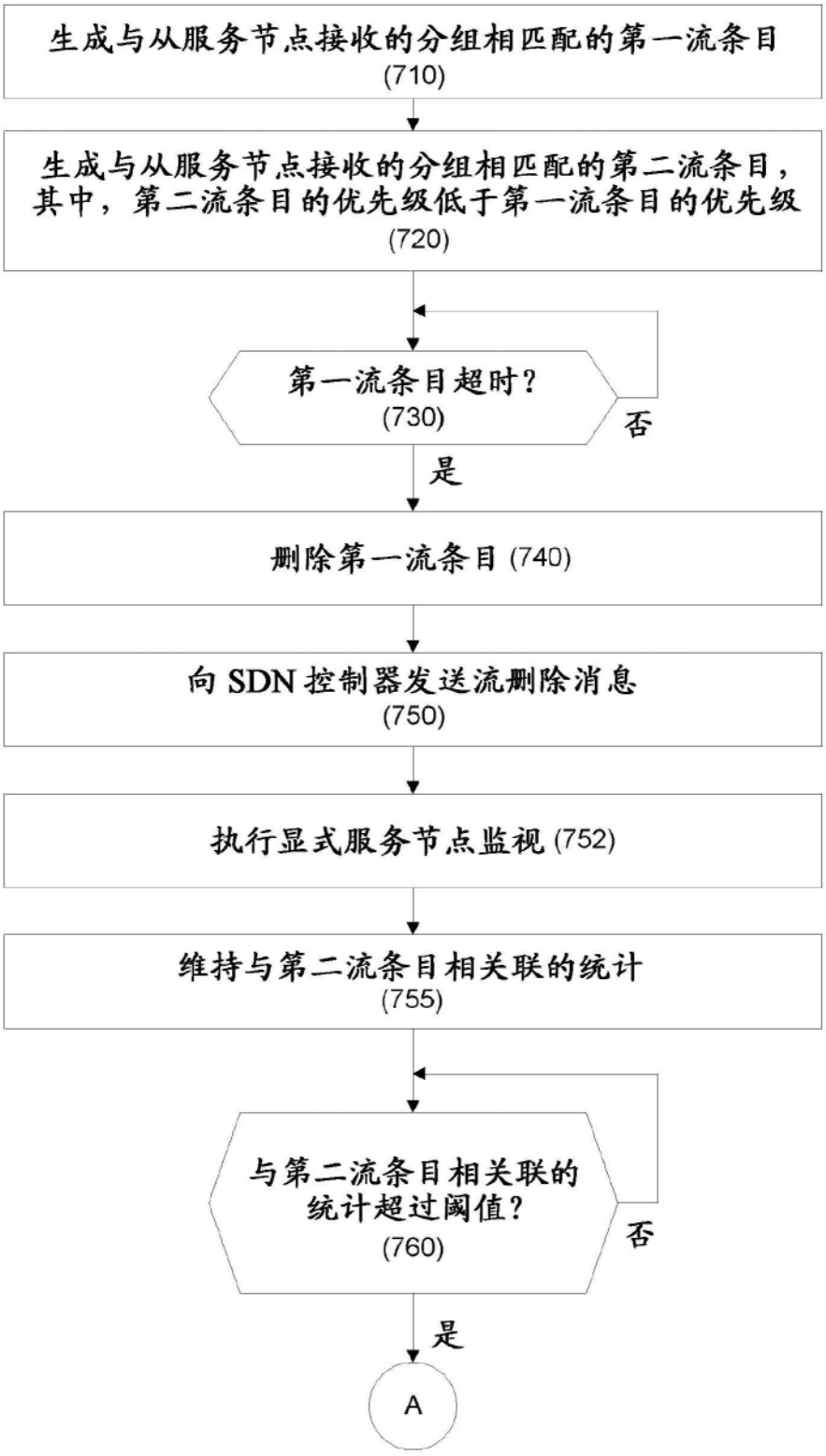

生成与从服务节点接收的分组相匹配的第一流条目
(710)
生成与从服务节点接收的分组相匹配的第二流条目，其中，第二流条目的优先级低于第一流条目的优先级
(720)
第一流条目超时？
(730)
否
是
删除第一流条目 (740)
向 SDN 控制器发送流删除消息
(750)
执行显式服务节点监视 (752)
维持与第二流条目相关联的统计
(755)
与第二流条目相关联的统计超过阈值？
(760)
否
是
A

图7

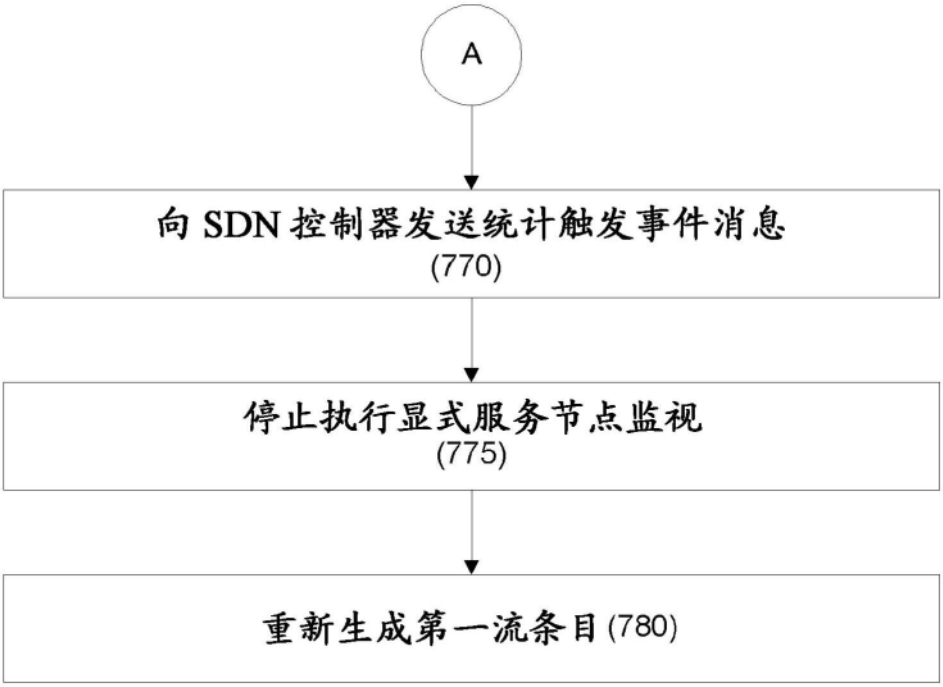


图7 (续)

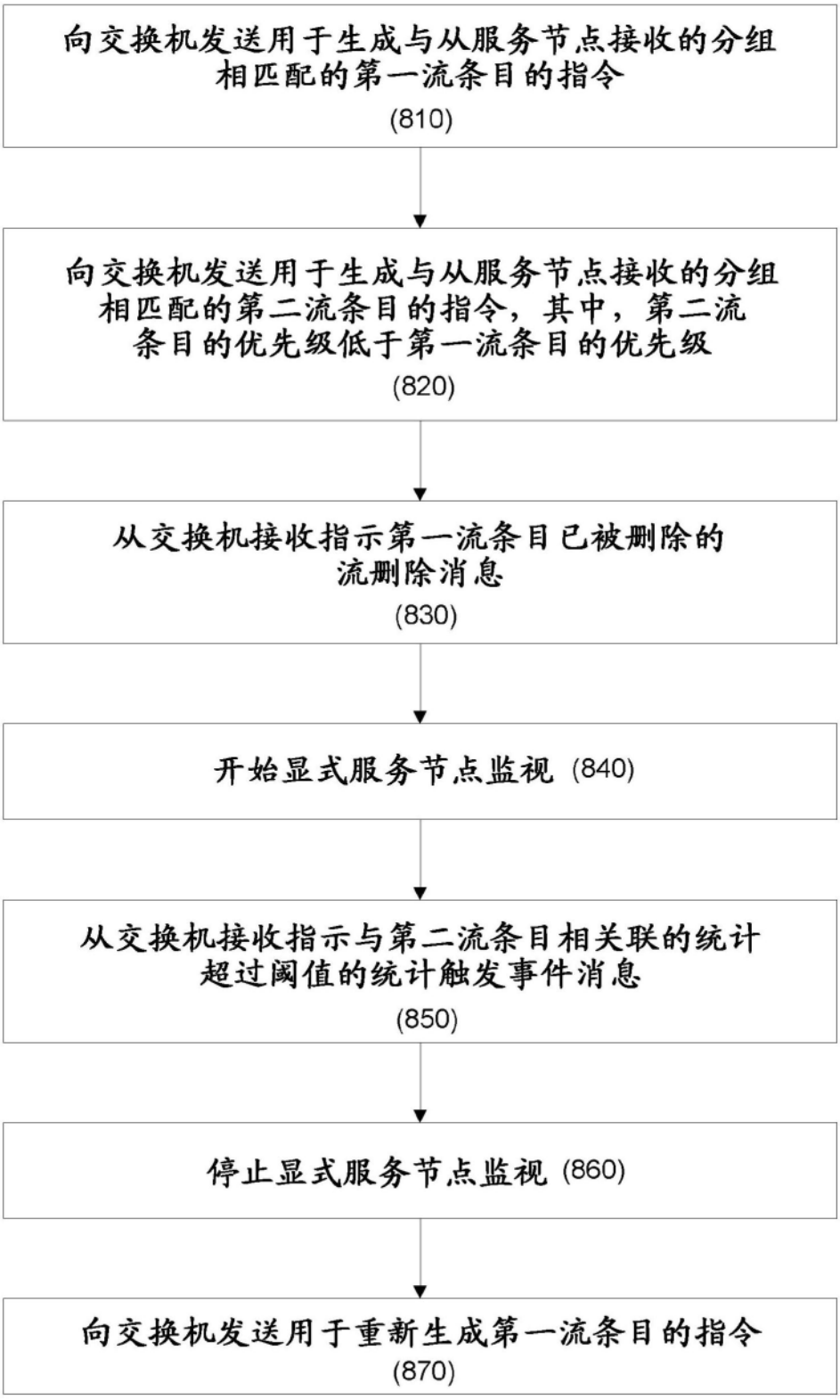

向交换机发送用于生成与从服务节点接收的分组相匹配的第一流条目的指令
(810)
向交换机发送用于生成与从服务节点接收的分组相匹配的第二流条目的指令，其中，第二流条目的优先级低于第一流条目的优先级
(820)
从交换机接收指示第一流条目已被删除的流删除消息
(830)
开始显式服务节点监视 (840)
从交换机接收指示与第二流条目相关联的统计超过阈值的统计触发事件消息
(850)
停止显式服务节点监视 (860)
向交换机发送用于重新生成第一流条目的指令
(870)

图8

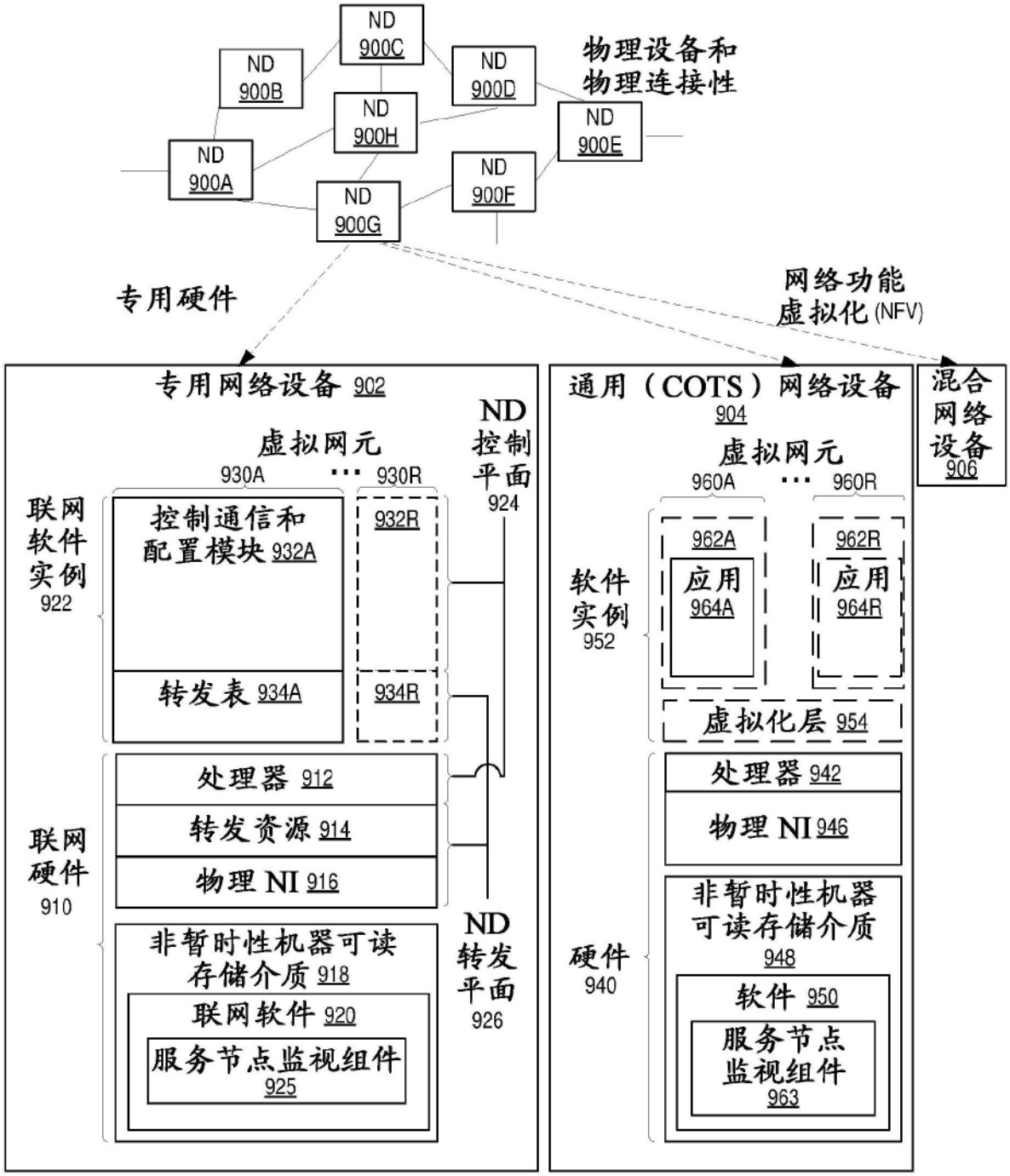

ND 900A
ND 900B
ND 900C
ND 900D
ND 900E
ND 900F
ND 900G
ND 900H
物理设备和物理连接性
专用硬件
网络功能虚拟化 (NFV)
专用网络设备 902
虚拟网元
930A ··· 930R
控制通信和配置模块 932A
932R
转发表 934A
934R
处理器 912
转发资源 914
物理 NI 916
非暂时性机器可读存储介质 918
联网软件 920
服务节点监视组件 925
联网软件实例 922
联网硬件 910
ND 控制平面 924
ND 转发平面 926
通用（COTS）网络设备 904
虚拟网元
960A ··· 960R
962A
应用 964A
962R
应用 964R
虚拟化层 954
处理器 942
物理 NI 946
非暂时性机器可读存储介质 948
软件 950
服务节点监视组件 963
软件实例 952
硬件 940
混合网络设备 906

图9A


卡 938
...
背板 936

图9B

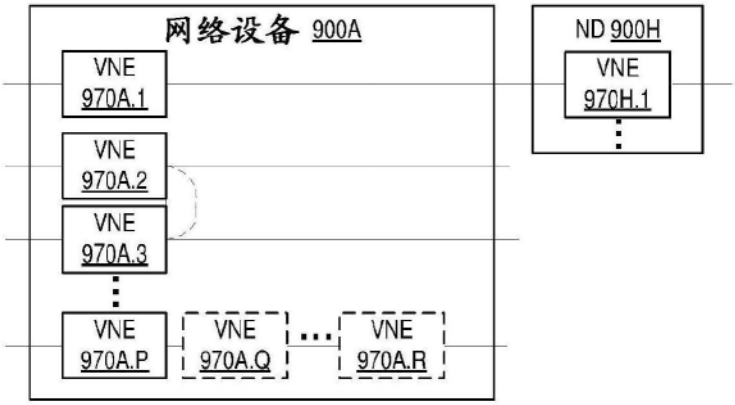

网络设备 900A
VNE 970A.1
VNE 970A.2
VNE 970A.3
VNE 970A.P
VNE 970A.Q
VNE 970A.R
ND 900H
VNE 970H.1

图9C

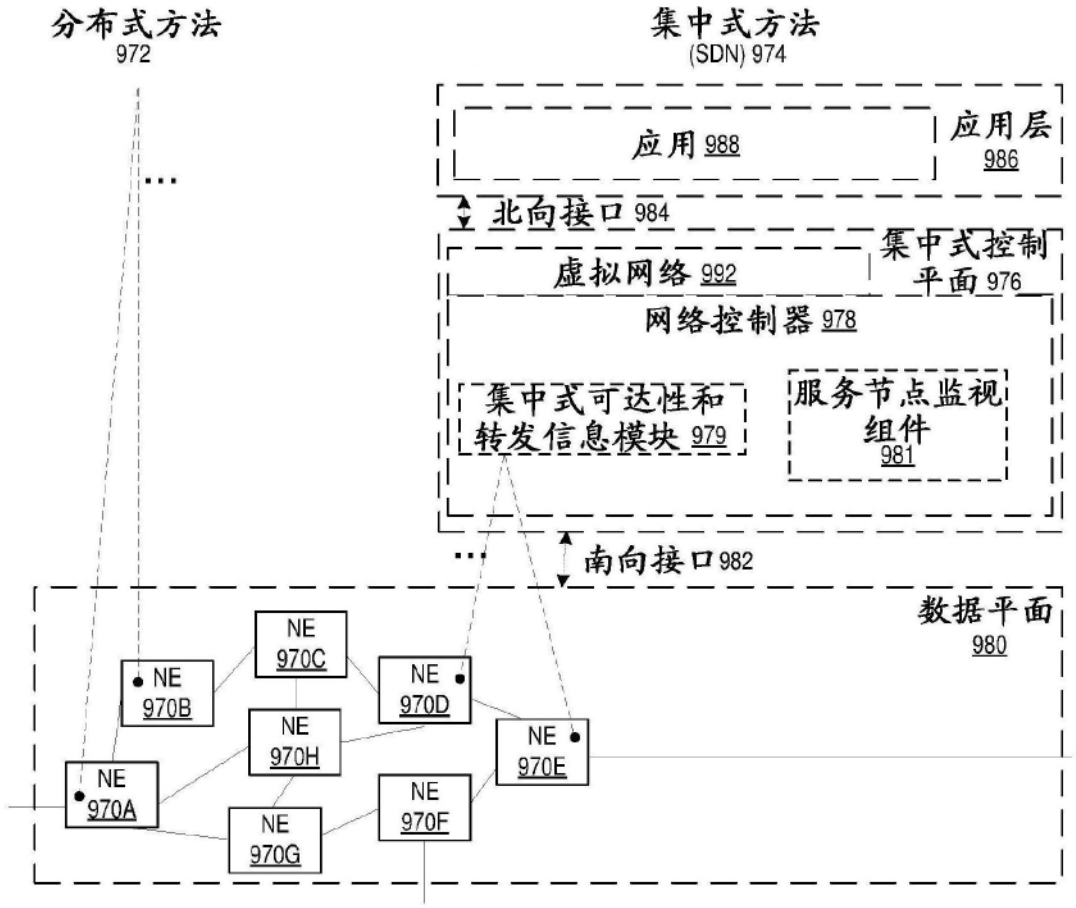

分布式方法
972
集中式方法
(SDN) 974
应用 988
应用层
986
北向接口 984
虚拟网络 992
集中式控制
平面 976
网络控制器 978
集中式可达性和
转发信息模块 979
服务节点监视
组件
981
南向接口 982
数据平面
980
NE
970A
NE
970B
NE
970C
NE
970D
NE
970E
NE
970F
NE
970G
NE
970H

图9D

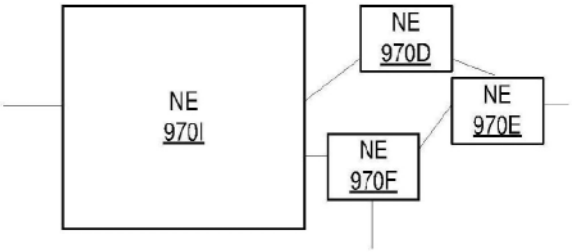

NE
970I
NE
970D
NE
970E
NE
970F

图9E

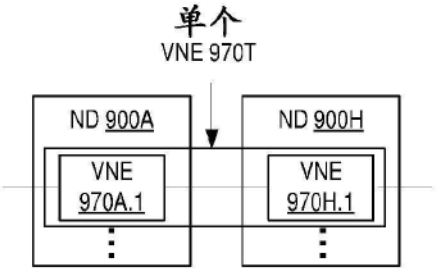

单个
VNE 970T
ND 900A
VNE
970A.1
ND 900H
VNE
970H.1

图9F

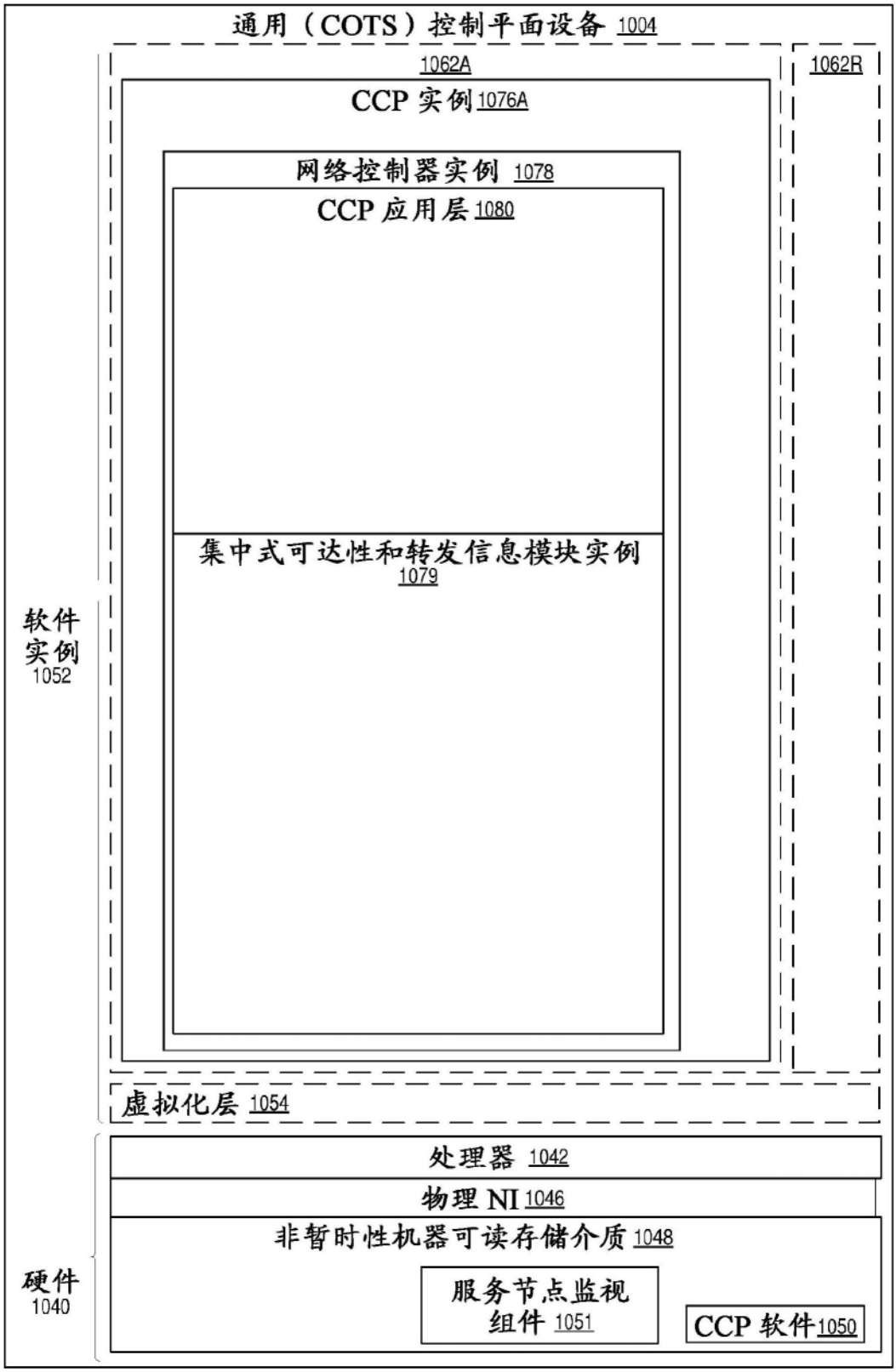

通用（COTS）控制平面设备 1004
1062A
1062R
CCP 实例 1076A
网络控制器实例 1078
CCP 应用层 1080
集中式可达性和转发信息模块实例 1079
软件实例 1052
虚拟化层 1054
处理器 1042
物理 NI 1046
非暂时性机器可读存储介质 1048
服务节点监视组件 1051
CCP 软件 1050
硬件 1040

图10