(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111886833 B
(45)授权公告日 2023.07.11

(21)申请号 201880091055.4
(22)申请日 2018.01.12
(65)同一申请的已公布的文献号
申请公布号 CN 111886833 A
(43)申请公布日 2020.11.03
(85)PCT国际申请进入国家阶段日
2020.09.10
(86)PCT国际申请的申请数据
PCT/IN2018/050018 2018.01.12
(87)PCT国际申请的公布数据
WO2019/138415 EN 2019.07.18
(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
(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 41/0663*(2022.01)
*H04L 41/0659*(2022.01)
*H04L 41/0695*(2022.01)
*H04L 45/28*(2022.01)
*H04L 45/64*(2022.01)
*H04L 49/253*(2022.01)

(56)对比文件
WO 2016037443 A1,2016.03.17
US 2017118066 A1,2017.04.27
US 2017099217 A1,2017.04.06
US 2015207724 A1,2015.07.23
US 2016112328 A1,2016.04.21
US 2015304205 A1,2015.10.22

审查员 付苗

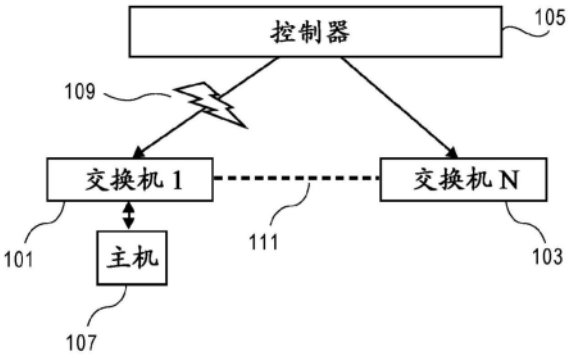
权利要求书2页 说明书21页 附图10页

(54)发明名称

# 重定向控制信道消息的方法和用于实现该方法的设备

(57)摘要

一种用于响应于与在软件定义网络SDN网络中的目标交换机的控制信道故障而自动重定向控制信道消息的方法和系统，自动重定向在SDN网络的拓扑响应于控制信道故障而被更新之前发生。该方法包括：确定控制信道消息是否要被发送给SDN控制器；检查控制信道的可用性；响应于控制信道不可用，选择用于到达SDN控制器的后备端口；以及经由后备端口向SDN控制器转发控制信道消息。




CN 111886833 B

1.一种用于响应于与在软件定义网络SDN网络中的目标交换机的控制信道故障而自动重定向控制信道消息的方法,所述自动重定向在所述SDN网络的拓扑响应于所述控制信道故障而被更新之前发生,所述方法包括:

确定(351)控制信道消息是否要被发送给SDN控制器;

检查(353)所述控制信道的可用性;

响应于所述控制信道由于沿到所述SDN控制器的路径的故障而不可用,选择(357)用于到达所述SDN控制器的后备端口;

封装(359)所述控制信道消息以包括所述目标交换机的源互联网协议地址;以及

经由所述后备端口向所述SDN控制器转发(361)所述控制信道消息。

2.根据权利要求1所述的方法,还包括:

经由所述后备端口从所述SDN控制器接收控制信道消息。

3.根据权利要求1所述的方法,还包括:

从所述SDN控制器接收用于选择不同的后备端口以改进控制信道消息到所述SDN控制器的路由的配置。

4.一种网络设备,用于实现用于响应于与在软件定义网络SDN网络中的目标交换机的控制信道故障而自动重定向控制信道消息的方法,所述自动重定向在所述SDN网络的拓扑响应于所述控制信道故障而被更新之前发生,所述网络设备包括:

非暂时性计算机可读介质(818),其中存储有控制信道重定向器;以及

处理器(812),其被耦接到所述非暂时性计算机可读介质,所述处理器被配置为执行所述控制信道重定向器,所述控制信道重定向器用于:确定控制信道消息是否要被发送给SDN控制器;检查所述控制信道的可用性;响应于所述控制信道由于沿到所述SDN控制器的路径的故障而不可用,选择用于到达所述SDN控制器的后备端口;封装所述控制信道消息以包括所述目标交换机的源互联网协议地址;以及经由所述后备端口向所述SDN控制器转发所述控制信道消息。

5.根据权利要求4所述的网络设备,其中,所述控制信道重定向器还被配置为:经由所述后备端口从所述SDN控制器接收控制信道消息。

6.根据权利要求4所述的网络设备,其中,所述控制信道重定向器还被配置为:从所述SDN控制器接收用于选择不同的后备端口以改进控制信道消息到所述SDN控制器的路由的配置。

7.一种计算设备,用于实现多个虚拟机,所述多个虚拟机实现网络功能虚拟化NFV,其中,所述多个虚拟机中的至少一个虚拟机实现用于响应于与在软件定义网络SDN网络中的目标交换机的控制信道故障而自动重定向控制信道消息的方法,所述自动重定向在所述SDN网络的拓扑响应于所述控制信道故障而被更新之前发生,所述计算设备包括:

非暂时性计算机可读介质(848),其中存储有控制信道重定向器;以及

处理器(842),其被耦接到所述非暂时性计算机可读介质,所述处理器被配置给所述至少一个虚拟机,所述至少一个虚拟机执行所述控制信道重定向器,所述控制信道重定向器用于:确定控制信道消息是否要被发送给SDN控制器;检查所述控制信道的可用性;响应于所述控制信道由于沿到所述SDN控制器的路径的故障而不可用,选择用于到达所述SDN控制器的后备端口;封装所述控制信道消息以包括所述目标交换机的源互联网协议地址;以及

经由所述后备端口向所述SDN控制器转发所述控制信道消息。

8.根据权利要求7所述的计算设备,其中,所述控制信道重定向器还被配置为:经由所述后备端口从所述SDN控制器接收控制信道消息。

9.根据权利要求7所述的计算设备,其中,所述控制信道重定向器还被配置为:从所述SDN控制器接收用于选择不同的后备端口以改进控制信道消息到所述SDN控制器的路由的配置。

10.一种控制平面设备,其与在软件定义网络SDN网络中的多个数据平面节点通信,所述控制平面设备实现用于响应于与在软件定义网络SDN网络中的目标交换机的控制信道故障而自动重定向控制信道消息的方法,所述自动重定向在所述SDN网络的拓扑响应于所述控制信道故障而被更新之前发生,所述控制设备包括:

非暂时性计算机可读介质(948),其中存储有控制信道重定向器;以及

处理器(942),其被耦接到所述非暂时性计算机可读介质,所述处理器被配置为执行所述控制信道重定向器,所述控制信道重定向器用于:将所述多个节点配置为具有一组后备端口,所述一组后备端口用于响应于所检测到的由于沿到所述SDN控制器的路径的故障而与所述控制平面设备的主控制信道的故障来转发控制信道消息;将所述多个节点配置为通过经由后备交换机的控制端口向所述控制平面设备转发所接收的去往所述控制平面设备的控制信道消息来处理后备交换机角色;以及将所述多个数据平面节点配置为通过向目的地数据平面节点转发来自所述控制平面设备的入站控制信道消息来处理所述入站控制信道消息,

其中,所述控制信道重定向器还被配置为:确定控制信道消息是否要被发送给SDN控制器;检查所述控制信道的可用性;响应于所述控制信道由于沿到所述SDN控制器的路径的故障而不可用,选择用于到达所述SDN控制器的后备端口;封装所述控制信道消息以包括所述目标交换机的源互联网协议地址;以及经由所述后备端口向所述SDN控制器转发所述控制信道消息。

11.根据权利要求10所述的控制平面设备,其中,所述控制信道重定向器还被配置为:响应于接收到控制信道重定向的指示而导出控制信道重定向路径。

12.根据权利要求10所述的控制平面设备,其中,所述控制信道重定向器还被配置为:响应于对所导出的控制信道重定向路径信息的分析而更新所述多个数据平面节点中控制信道消息的重定向。

13.根据权利要求10所述的控制平面设备,其中,所述控制信道重定向器还被配置为:经由所述后备端口从所述SDN控制器接收控制信道消息。

14.根据权利要求10所述的控制平面设备,其中,所述控制信道重定向器还被配置为:从所述SDN控制器接收用于选择不同的后备端口以改进控制信道消息到所述SDN控制器的路由的配置。

# 重定向控制信道消息的方法和用于实现该方法的设备

## 技术领域

[0001] 本发明的实施例涉及软件定义网络(SDN)领域;更具体地,涉及在SDN控制器与SDN数据平面节点之间传送控制消息。

## 背景技术

[0002] 一个软件定义网络(SDN)通过提供可编程的网络基础设施,促进了在网络层处的快速和开放式创新。SDN将控制平面功能与数据平面功能分离,从而使得SDN网络中的节点能够由用于SDN网络的一个或一组控制器来配置。SDN网络中的节点实现数据平面并由控制器来配置,该控制器管理SDN网络的控制平面。以这种方式,可以更新网络基础设施,而不必单独替换或更新网络中的每个节点。这也使节点的维护成本效益更高,因为它们具有较低的资源要求并且不太复杂,从而使节点的制造和维护成本更低。由此,网络中的复杂性被限于该组控制器,该组控制器比SDN网络的数据平面的众多节点在升级方面更具成本效益。SDN网络可以为网络编程提供标准接口,以及还可以为网络编程提供某些定义的语义。

[0003] 在SDN网络中使用流控制协议来管理SDN网络中的控制平面通信和数据平面节点的配置。可以利用任何流控制协议。OpenFlow交换规范是流控制协议的示例。OpenFlow使得能够在SDN中对流控制策略进行动态编程。流控制协议使用流表定义了数据平面节点的操作的基本组织,这些流表被编程为实现流控制策略,这些流控制策略实现SDN网络中的分组转发。

[0004] 数据平面中的节点与SDN控制器进行通信,以将信息发送给SDN控制器进行处理,并经由控制信道来接收配置、命令和已处理的数据。但是,在某些情况下,数据平面节点可能由于SDN网络中或实现给定节点与SDN控制器之间的通信的任何路径中的链路或节点故障而变得与SDN控制器断开连接。在SDN控制器丢失的情况下,数据平面节点可能无法完成某些功能或解决问题,并且结果,断开连接的数据平面节点的操作可能中断。

## 发明内容

[0005] 在一个实施例中,提供了一种由网络设备实现的用于响应于与在软件定义网络SDN网络中的目标交换机的控制信道故障而自动重定向控制信道消息的方法,所述自动重定向在所述SDN网络的拓扑响应于所述控制信道故障而被更新之前发生。所述方法包括:确定控制信道消息是否要被发送给SDN控制器;检查所述控制信道的可用性;响应于所述控制信道不可用,选择用于到达所述SDN控制器的后备端口;以及经由所述后备端口向所述SDN控制器转发所述控制信道消息。

[0006] 在另一实施例中,提供了一种实现用于响应于与在SDN网络中的目标交换机的控制信道故障而自动重定向控制信道消息的方法的网络设备,其中,所述自动重定向在所述SDN网络的拓扑响应于所述控制信道故障而被更新之前发生。所述网络设备包括其中存储有控制信道重定向器的非暂时性计算机可读介质,以及耦接至所述非暂时性计算机可读介质的处理器。所述处理器被配置为执行所述控制信道重定向器,所述控制信道重定向器用

于:确定控制信道消息是否要被发送给SDN控制器;检查所述控制信道的可用性;响应于所述控制信道不可用,选择用于到达所述SDN控制器的后备端口;以及经由所述后备端口向所述SDN控制器转发所述控制信道消息。

[0007] 在另一实施例中,一种计算设备实现多个虚拟机,所述多个虚拟机实现网络功能虚拟化(NFV),其中,所述多个虚拟机中的至少一个虚拟机实现用于响应于与在SDN网络中的目标交换机的控制信道故障而自动重定向控制信道消息的方法,其中,所述自动重定向在所述SDN网络的拓扑响应于所述控制信道故障而被更新之前发生。所述计算设备包括其中存储有控制信道重定向器的非暂时性计算机可读介质,以及耦接至所述非暂时性计算机可读介质的处理器,所述处理器被配置给至少一个虚拟机,所述至少一个虚拟机执行所述控制信道重定向器,所述控制信道重定向器用于:确定控制信道消息是否要被发送给SDN控制器;检查所述控制信道的可用性;响应于所述控制信道不可用,选择用于到达所述SDN控制器的后备端口;以及经由所述后备端口向所述SDN控制器转发所述控制信道消息。

[0008] 在一个实施例中,一种控制平面设备与在SDN网络中的多个数据平面节点通信,其中,所述控制平面设备实现用于响应于与在SDN网络中的目标交换机的控制信道故障而自动重定向控制信道消息的方法,所述自动重定向在所述SDN网络的拓扑响应于所述控制信道故障而被更新之前发生。所述控制设备包括其中存储有控制信道重定向器的非暂时性计算机可读介质,以及耦接至所述非暂时性计算机可读介质的处理器。所述处理器被配置为执行所述控制信道重定向器,所述控制信道重定向器用于:将所述多个节点配置为具有一组后备端口,所述一组后备端口用于响应于所检测到的与所述控制平面设备的主控制信道的故障而转发控制信道消息;将所述多个节点配置为通过经由后备交换机的控制端口向所述控制平面设备转发所接收的去往所述控制平面设备的控制信道消息来处理后备交换机角色;以及将所述多个数据平面节点配置为通过向目的地数据平面节点转发来自所述控制平面设备的入站控制信道消息来处理所述入站控制信道消息。

## 附图说明

[0009] 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:

[0010] 图1是SDN网络的一个示例实施例的图,其示出了使交换机1与控制器断开连接的事件;

[0011] 图2是经由代理交换机与控制器通信的已断开连接的交换机的一个实施例的图;

[0012] 图3A是SDN控制器配置控制信道消息的重定向的流程图;

[0013] 图3B是在控制信道故障的情况下在始发交换机处重定向出站控制信道消息的过程的一个实施例的流程图;

[0014] 图3C是由后备SDN交换机实现的用于重定向的过程的一个实施例的流程图;

[0015] 图4是具有SDN控制器的SDN网络的图,该SDN控制器使用控制信道消息和LLDP来确定SDN网络拓扑;

[0016] 图5是SDN网络的一个示例实施例的图,其示出了使交换机1与控制器断开连接的事件;

[0017] 图6是链路状态监视过程的一个实施例的图;

[0018] 图7A是用于探测SDN网络中的链路的初始控制平面步骤的过程的一个实施例的流程图；

[0019] 图7B是响应于SDN网络中给定交换机的丢失控制信道而实现的过程的概括描述的流程图；

[0020] 图8A示出了根据本发明一些实施例的在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8B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8C示出了根据本发明一些实施例的在其中可以耦接虚拟网元(VNE)的各种示例性方式；

[0023] 图8D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8E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在不同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8F示出了根据本发明一些实施例的其中多个VNE在不同ND上实现并相互耦接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9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950的通用控制平面设备。

## 具体实施方式

[0027] 以下说明描述了用于管理软件定义网络(SDN)控制器与在该SDN网络中的节点或“交换机”之间的通信的方法和装置。特别地，实施例提供了一种用于处理导致SDN交换机与控制器之间的通信中断的事件的方法和系统。实施例提供了使SDN交换机能够经由SDN网络中的另一个SDN交换机与SDN控制器通信的方法和系统。该中间交换机可以用作在SDN控制器与SDN交换机之间中继通信的代理，以确保SDN交换机的连续功能，直到流控制通信会话在SDN交换机与SDN控制器之间被重新建立为止。

[0028] 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。

[0029] 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。

[0030] 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。

[0031] 在下面的说明书和权利要求书中，可以使用术语“耦接”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。“耦接”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”用于指示在彼此耦接的两个或多个元件之间建立通信。

[0032] 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。

[0033] 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”。

[0034] 概述

[0035] 如果在SDN控制器与SDN交换机之间的连接断开，则在没有由本文的实施例提供的过程和系统的情况下，关键控制信道消息无法从SDN交换机到达SDN控制器或控制信道消息无法从SDN交换机到达SDN控制器。例如，在这种情况下，即使SDN控制器仍在执行，但是由于

没有控制信道可用于SDN交换机与SDN控制器进行通信，从SDN交换机到SDN控制器的对于数据路径功能很关键的一些关键控制信道消息(例如，分组入(PACKET_IN)消息)也会被错过或丢失。类似地，将存在可能响应于SDN交换机发送的控制信道消息(例如，分组入(PACKET-In)消息)而被发送的从SDN控制器到SDN交换机的关键控制信道消息(例如，分组出(PACKET-OUT)消息)或其他消息，它们在这种情况下会丢失。

[0036] 因此，现有技术在处理SDN控制器与SDN交换机之间的通信问题方面存在缺点。当控制信道出现故障时，现有技术没有提供将控制信道消息(例如，分组入/分组出(packet_in/packet-out)消息)发送给SDN控制器的过程。即使SDN控制器和SDN交换机均正常工作，但是由于缺少SDN控制器和SDN交换机彼此联系的能力，在这种情况下，由SDN交换机处理的数据业务也会中断。

[0037] 图1是具有控制信道问题的基本SDN网络的一个实施例的图。在简化示例中，第一SDN交换机1 101和一组其他SDN交换机(包括SDN交换机N)103被连接到SDN控制器105。SDN交换机101、103可以是任何数量或类型的联网设备。类似地，SDN控制器可以是一组SDN控制器中的一个SDN控制器，并且可以管理SDN网络(例如，集群)的任何子集或整个SDN网络。SDN控制器105可以类似地是任何类型的计算或联网设备。

[0038] SDN交换机1 101可能丢失它的与SDN控制器105的控制信道109。在这种情况下，无论经由其他SDN交换机103与SDN控制器105的可能连接111如何，该控制信道都将丢失。利用其他连接111需要花费时间来解析，并且在这样的到SDN控制器105的新路径的解析期间，SDN交换机1101无法完成所有数据业务处理功能。例如，由于缺乏连接109，被连接到SDN交换机1 101的主机107可能无法跨SDN网络发送数据，从而导致SDN交换机1 101无法处理主机107的数据业务。

[0039] 实施例提供了一种克服现有技术的这些限制的过程和系统。实施例将一组后备交换机配置为用于控制信道消息的代理。类似地，例如使用OpenFlow配置协议来配置用于控制器端口的一组后备端口，它们将是被导向后备或代理交换机的端口。该组后备端口可以是指向一组相应的后备交换机的逻辑端口。可以以SELECT方式(即，通过流控制选择过程)或通过FAST FAILOVER(即，交换机借以在检测到沿主路径的故障时立即重新配置以使用后备端口/路径而无需解决由故障引起的网络拓扑变化的过程)来使用后备端口。实施例可以配置在主或“源”交换机和代理交换机上的流表中的流控制条目(例如，OpenFlow流表条目)以实现对控制信道消息(例如，PACKET-IN和PACKET-OUT)重定向的处理。实施例可以在重定向后的控制信道消息(例如，PACKET-In)被转发给SDN控制器之前将一些元数据设置到重定向后的控制信道消息，以指示控制信道消息已经采用的从源交换机到代理交换机的路径。在一些实施例中，SDN控制器可以基于传入控制信道消息的性质来改变主/辅后备交换机。实施例还包括将流控制消息嵌入控制信道消息内(例如，在packet_out消息中)的方法。以这种方式，实施例使能每当在SDN控制器与源SDN交换机之间的控制信道出现故障时，通过一组后备交换机来重定向去往和来自SDN控制器的控制信道消息(例如，packet_in/packet_out消息)。实施例还使得能够通过该组后备交换机从SDN控制器向初始的源交换机发送控制信道消息(例如，packet_out消息)。

[0040] 图2是具有被建立以用于在发生连接问题时使能控制信道通信的代理的SDN网络的一个实施例的示意图。在图1的简单示例的基础上，图2说明了在源SDN交换机101与SDN控

制器105之间的主控制信道连接出现故障的情况下，使用一组代理SDN交换机103将控制信道消息重定向到SDN控制器105。实施例建立了将在主控制信道路径发生故障的情况下使用的一组后备端口和后备交换机。因此，后备路径和端口的建立在任何故障之前被确定，并且这使得源交换机能够迅速改变为使用沿路径111的后备端口和后备交换机，从而能够在没有源SDN交换机1 101的功能的显著中断的情况下到达SDN控制器105。

[0041] 因此，这些实施例提供了优于现有技术的优点，因为它们提供了具有用于SDN交换机的稳健的控制信道消息处理(例如Packet_in/Packet_out消息处理)机制的过程和系统。即使在特定交换机与SDN控制器之间的主控制信道出现故障，SDN控制器也可以提供控制信道消息(例如packet_in/packet_out消息)。实施例提供了区分控制信道故障和SDN交换机关闭的稳健方式。另外，实施例基于传入的控制信道消息而提供在SDN控制器处的针对SDN交换机恢复的增强决策制定。

[0042] 后备路径和控制信道重定向的配置

[0043] 实施例提供了当控制信道发生故障时将控制信道消息(例如，packet_in消息)从源SDN交换机重定向到在另一后备路径上的SDN控制器的机制。实施例提供了用于响应于沿着后备路径接收到控制信道消息而从SDN控制器向源SDN交换机进行通信(例如，以提供从源SDN交换机接收的Packet_in消息)的机制。

[0044] 为了实现此系统和过程，SDN控制器可以将一组后备端口配置为在SDN网络中每个交换机上的控制器端口，以用作将控制信道消息发送给SDN控制器的后备或代理(例如，用于发送packet_in消息)。用于配置的流控制协议可以是OpenFlow协议，或例如ovsdb的类似配置协议，它们可用于针对交换机上的控制器端口配置一组后备端口。“控制器端口”是用于将控制信道消息发送给SDN控制器的出站端口。后备端口将是面向所配置的代理交换机或用于给定源交换机的一组代理的下一跳的端口。在源SDN交换机与SDN控制器之间的主控制信道连接出现故障的情况下，代理交换机负责在来自源交换机的控制信道消息(例如，Packet_in消息)向SDN控制器前进时重定向这些控制信道消息。如果代理交换机的控制器端口不活跃(即，下一个代理交换机没有正在工作的控制信道，因为它不是“活动的(LIVE)”)，则可以向代理交换机的该组后备端口咨询一个或多个间接级别(levels of indirection)。这样，实施例提供了用于SDN中的控制信道的M:N冗余模型。该组后备端口可以在SELECT过程或FAST FAILOVER过程中使用，类似于流控制(例如OpenFlow)组构造。这意味着可以基于负载平衡算法或基于活动性(liveness)机制从该组后备端口中选择一个端口。

[0045] 实施例包括检测控制器端口活动性并在SDN交换机处启用packet-in重定向。每当数据分组命中在流控制表中的“output:CONTROLLER”动作时，实际分组将被包装(wrap)在控制信道消息(例如，OpenFlow packet-in消息)中，然后在当前逻辑端口上被朝向SDN控制器输出。如果由于任何原因(例如，心跳超时或类似指示或控制信道中断)，控制器端口被检测为处于非活动(non-live)(即，非活跃(inactive))状态，则SDN交换机可以从该组所配置的后备端口中选择一个后备端口。然后，如下面在本文中进一步讨论的，通过设置分组字段，控制信道消息(例如，packet-in消息)可以被转发给后备交换机。

[0046] 在实施例中，该方法可以独立地或在流控制管道之外(例如，在OpenFlow管道之外)实现。每当交换机执行“输出到控制器”动作时，交换机内部模块(例如“重定向器”)就可

以检测该动作并查找该组后备端口。在其他实施例中，如果可以针对流控制管道内的控制器端口实现活动性检测，则也可以通过流控制(例如，OpenFlow)管道来实现相同的功能。在这种情况下，可以创建FAST FAILOVER组，其中CONTROLLER端口作为桶(bucket)中的第一个端口，而该组后备端口作为其余的桶。每当CONTROLLER端口处于NON-LIVE时，将从该桶中选择下一个LIVE端口。

[0047] 在一个实施例中，该过程实现控制入站信道消息(例如，packet-out消息)重定向。就像用于控制器(例如，packet-in)重定向的出站控制信道消息一样，如果像路由协议情况那样需要独立接收的控制信道消息(例如，packet-out)，则这可以通过入站控制信道重定向来实现。例如，如果SDN控制器检测到朝向SDN交换机的控制信道出现故障，并且如果存在要被发送给该特定SDN交换机的packet-out，则控制器可以检查用于此SDN交换机的任何所配置的后备交换机，并且入站(例如，packet-out)控制信道消息可以在该信道上被朝向后备交换机发送。所需的流控制条目将在后备交换机上被配置以认识到这是朝向另一个SDN交换机的控制信道消息，并且后备交换机将该消息原样转发给目的地交换机。

[0048] 将参考其他附图的示例性实施例描述流程图中的操作。然而，应当理解，流程图的操作可以由除参考其他附图所讨论的实施例之外的本发明的实施例执行，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流程图讨论的操作不同的操作。

[0049] 图3A是SDN控制器配置控制信道消息的重定向的流程图。如上所述，在SDN控制器中的SDN交换机首先被配置为具有一组后备端口，该组后备端口用于响应于与SDN控制器的主控制信道的故障而向SDN控制器转发控制信道消息(框301)。这可以通过创建由负载平衡算法或活动性机制选择的后备端口列表来配置。这些后备端口选择过程可以是常规重定向器机制的一部分。

[0050] 实施例包括配置在源或始发SDN交换机和中间后备SDN交换机上的流控制(例如OpenFlow)条目，以用于控制信道(例如OpenFlow)消息重定向。在后备SDN交换机上，流控制(例如，OpenFlow)条目被配置用于将来自源或始发SDN交换机的所有接收到的控制信道(例如，packet-in消息)分组转发给该控制器(框303)。接收到的发往SDN控制器的控制信道消息被配置为在相应的控制器端口上被转发给该控制器。例如，在后备交换机处的流表可被配置为:匹配OpenFlow端口:6633，src_IP＝始发交换机IP，dest_IP＝控制器IP，动作＝输出到CONTROLLER。

[0051] 该过程还配置SDN交换机以处理来自SDN控制器的返回控制信道分组(框305)。在始发交换机上，一组控制信道消息(例如，使用OpenFlow或ovsdb)处理条目被配置为处理在NORMAL(正常)模式下从代理交换机的端口上接收到的控制信道消息。例如，可以创建条目以:匹配OpenFlow端口:6633，src_IP＝控制器IP，dest_IP＝交换机IP，动作＝NORMAL，或匹配ovsdb端口:6640，src_IP＝控制器IP，dest_IP＝交换机IP，动作＝NORMAL。

[0052] 在一些实施例中，SDN控制器可以导出控制信道(例如，packet-in)重定向路径，以使得SDN控制器可以确定正在具有控制信道中断的SDN交换机(框307)。这是使得SDN控制器能够为具有控制信道中断问题的SDN交换机采用智能恢复机制的可选步骤。在一个示例实施例中，在重定向的每一跳处，后备SDN交换机可以由SDN控制器配置为向SDN控制器输出重定向后的控制信道分组已被处理的指示(例如，CONTROLLER动作将被执行)。例如，可以生成packet-in并且packet-in封装所接收的控制信道分组。每当packet-in或类似的控制信道

消息封装发生时，始发交换机都将它的IP地址添加为packet-in或控制信道消息的源IP地址，然后再将其转发给后备交换机。在后备交换机处，控制信道消息(例如，packet-in消息)被从源始发SDN交换机重定向，并将被再次转发或“输出到CONTROLLER”，这又导致了(例如，packet-in的)与外部封装报头中的当前后备交换器IP位址的再一次封装。在SDN控制器处，当封装后的控制信道消息被提取时，该封装的每个级别处的源IP将提供控制信道所经过的各种SDN交换机的指示。如果需要数据路径节点id的更清晰指示，则在向SDN控制器转发消息之前，在每个交换机处，相应的数据路径节点ID可以被存储在该封装中(例如，在packet-in Cookie字段中设置)。

[0053] 在一些实施例中，SDN控制器可以基于对接收到的重定向后的控制信道消息的导出数据路径的分析，使用该数据路径来更新主和辅后备SDN交换机(309)。当分析传入的重定向后的控制信道消息(例如，packet-in消息)时，SDN控制器可以检查对于源SDN交换机，重定向所花费的跳数是否超过预期或所需的跳数(例如，通过与除控制信道路径之外的最佳路径进行比较)。如果发生低效的重定向，则SDN控制器可以决定更改主后备交换机，即重定向packet-in消息以实现更佳的重定向。

[0054] 在一些实施例中，SDN控制器可以使得其他类型的流控制消息(例如，OpenFlow/ovsdb消息或配置)能够在控制信道故障的情况下通过后备交换机被转发以更新主/辅交换机(框311)。每当控制信道出现故障时，在后备路径需要更新的情况下，需要经由后备交换机在SDN控制器与始发交换机之间的完整控制路径。如上所述，如果在始发和后备交换机上配置了与重定向相关的流条目，则各种流控制(例如，ovsdb)配置消息也可以到达其控制信道出现故障的始发交换机。这样，后备路径可以基于反馈机制被更新以变得更佳。

[0055] 图3B是在控制信道故障的情况下在始发交换机处重定向出站控制信道消息的过程的一个实施例的流程图。在一个实施例中，该过程在数据分组的处理中被模仿，并且流控制过程确定控制信道消息要被发送给SDN控制器(例如，packet-in消息)(框351)。然后进行检查以确定控制信道是否可用(框353)。可以利用任何过程或活动性检测机制来确定控制信道的可用性。在控制信道可用的情况下，则该过程通常通过生成控制信道消息并将其转发给SDN控制器而正常地继续(框355)。如果控制信道不可用，则该过程从由SDN控制器配置的该组可用后备端口中选择一个后备端口(框357)。后备端口是预先配置的，并且可以使用包括任何快速故障转移(failover)机制或负载平衡算法的任何选择机制。后备端口的选择不需要SDN交换机或SDN控制器重新计算SDN网络的拓扑和路由以到达SDN控制器。

[0056] 然后，该过程封装包括始发交换机的IP地址的控制信道消息作为该消息的源地址(框359)。然后，封装后的控制信道消息在后备端口上被转发(框361)。示例配置和算法在下面以伪代码呈现：

[0057] 算法

[0058] 进入交换机1的分组导致PACKET_IN

[0059] 如果(controller_port活动)

[0060] 执行正常处理

[0061] 否则

[0062] 设置src IP＝交换机IP，dest IP＝控制器IP，从该组所配置的后备端口中的一个后备端口上输出PACKET_IN消息。

[0063] 图3C是由后备SDN交换机实现的用于重定向的过程的一个实施例的流程图。当后备SDN交换机直接或间接从始发SDN交换机接收到封装后的控制信道分组时，发起后备交换机过程(框371)。分析接收到的分组以确定它是否是重定向后的控制信道消息(框373)。如果该分组不是重定向后的控制信道消息，则后备SDN交换机正常地处理接收到的数据分组(框375)。如果接收到的分组是控制信道消息，则在被在控制信道端口上向SDN控制器输出之前，该分组可以被重新封装或者可以封装可以被更新以包括后备SDN交换机的源地址(框377)。后备SDN交换机可以与如上所述的始发SDN交换机类似地确定控制信道的可用性。

[0064] 在一个示例实施例中，重定向过程可以用以下伪代码来描述：

[0065] 分组到达后备交换机

[0066] 如果(packet_type＝OpenFlow&&src IP＝交换机IP,destip＝控制器IP)

[0067] 将分组输出到控制器

[0068] 否则控制器接收并处理分组

[0069] 一旦分组到达SDN控制器，控制器就使用该后备SDN交换机的适当端口向后备或代理SDN交换机发送控制信道响应消息(例如，作为嵌在packet_out消息中的OpenFlow消息)，该后备SDN交换机又将控制信道响应转发给始发SDN交换机。

[0070] 在始发或源SDN交换机处，在接收到控制信道响应消息时，始发节点检查接收到的分组是否是控制信道响应消息。此检查可以用以下伪代码来描述：

[0071] (如果(packet_type＝OpenFlow&&src IP＝控制器IP,destip＝交换机IP)

[0072] 则执行NORMAL动作，并执行OpenFlow消息

[0073] 因此，实施例由此提供了用于SDN交换机的稳健的控制信道消息处理(例如，用于packet_in/packet_out)机制。即使在特定SDN交换机与SDN控制器之间的控制信道出现故障，SDN控制器也能够提供控制信道消息(例如packet_in/packet_out消息)。

[0074] 链路监视

[0075] 如上所述，SDN是一种将控制平面(其做出路由和交换检测)与数据平面(其执行转发功能)分离的方法。数据平面是由SDN控制器控制的交换机(SDN交换机)的集群或网络，并且它们经由适当的链路(例如L2/L3链路)被互连以形成网络拓扑。SDN控制器的功能是在任何给定时间维护数据平面网络拓扑的正确视图。SDN控制器依靠对数据平面的定期监视(使用诸如链路层发现协议(LLDP)之类的协议)来实时维护网络拓扑视图。在基于流控制协议链路OpenFlow的SDN和基于LLDP的内部链路监视中，SDN控制器在两个方向上定期在每个链路上发送LLDP分组，并且此过程涉及以下步骤，对于每个链路，SDN控制器构造LLDP消息并发送该LLDP消息作为到源SDN交换机的控制信道消息(例如，packet-out消息)。源SDN交换机提取LLDP消息，并将其沿指定的数据链路转发到目的地SDN交换机。LLDP消息由目的地SDN交换机接收，并被转发给SDN控制器。SDN控制器接收控制信道消息(例如，packet-in消息)，提取LLDP消息，将该LLDP消息与它发起的LLDP消息相关联，以及建立已被遍历的链路的链路状态，从而累积链路状态日期以确定SDN网络的当前拓扑。

[0076] 图4是具有SDN控制器的SDN网络的图，该SDN控制器使用控制信道消息和LLDP来确定SDN网络拓扑。在该示例中，SDN控制器向SDN交换机1发送封装了LLDP消息的控制信道消息。在SDN交换机1处，LLDP消息被提取并被发送给SDN交换机2。LLDP消息被接收并被封装在控制信道消息中以发送回SDN控制器，从而提供针对在交换机1与交换机2之间的链路的链

路状态信息。

[0077] 然而，与上述实施例一样，如图5所示，当SDN控制器与SDN交换机之间的控制信道出现故障时，此过程也会遇到问题。上述链路监视功能涉及LLDP分组遍历控制信道(例如，作为packet-out消息)，然后是作为LLDP消息的数据链路，然后又是控制信道(例如，作为packet-in消息)。在这种情况下，如果在SDN控制器与交换机1或交换机2之间的控制信道会话中断，如图5所示，则上面说明的监视流将被中断，从而导致SDN控制器的错误数据链路故障确定，并从而导致SDN控制器的SDN网络拓扑视图与SDN网络的实际拓扑不同步(out-of-sync)。这是由于以下事实：当SDN控制器丢失与SDN交换机之一的控制信道会话时，它将无法确定这只是控制平面中断还是SDN交换机自身重新引导/崩溃/隔离导致数据平面中断。

[0078] 本文提供的实施例因此被进一步扩展，使得当SDN控制器丢失与SDN交换机(例如，图5中的交换机1)的控制信道会话时，SDN控制器停止对与SDN交换机1相关联的所有链路的定期监视。基于现有拓扑，SDN控制器向SDN交换机2发送探测控制信道消息(例如，作为packet-out消息)。在接收到探测控制信道消息后，SDN交换机2在朝向SDN交换机1的链路上发送探测分组。如果SDN交换机1活动并且活跃，则SDN交换机1通过适当地修改传入的探测数据分组并将其发送回交换机2或类似地用探测响应消息响应SDN交换机2，来响应SDN交换机2的探测数据分组。

[0079] SDN交换机2接收探测响应消息并发送具有从SDN交换机1封装的探测响应的控制信道消息(例如，packet-in消息)。SDN控制器将探测响应消息与SDN交换机1相关联并确定在SDN交换机1与2之间的链路连接链路的链路状态为活动，SDN交换机1为活动，从而更新SDN网络的拓扑视图。

[0080] 但是，如果SDN交换机1出现故障，则SDN控制器或SDN交换机2将不会接收到探测响应消息，并且由此SDN控制器可以确定链路已出现故障并相应地更新拓扑视图。这样，即使当SDN控制器丢失与SDN网络中的某些SDN交换机的控制信道连接时，实施例也有助于在SDN控制器处维持SDN网络拓扑视图。在一些实施例中，该过程可以在不修改或扩展流控制协议(例如，OpenFlow协议)的情况下被实现。

[0081] 图6是链路状态监视过程的一个实施例的图。实施例包括在使用用于L3/L2链路的探测消息来确定链路状态的过程中的变化。关于图6描述了每个示例过程。图6示出了简化的SDN网络配置，其中，交换机1已经丢失了它的与SDN控制器的控制信道连接。但是，交换机1能够经由交换机2到达SDN控制器。在一些实施例中，交换机2还充当如上所述的后备或代理交换机。

[0082] 图7A是用于在SDN网络中探测链路的初始控制平面步骤的过程的一个实施例的流程图。该流程图适用于第2层或第3层实现。第2层是网络的传输层，而第3层是SDN网络的网络层并且可以使用诸如互联网协议(IP)之类的协议进行通信。该过程涉及确定要被发送给已丢失其与SDN控制器的控制信道会话的交换机的探测消息的标签(label)或标记(tag)(框701)。然后，控制器将探测响应配置安装在目标交换机(交换机1)的流表中，以处理探测响应的发送(框703)。控制器将交换机1配置为处理对探测消息(例如，来自交换机2)的接收(框705)。控制器将中间交换机(例如，后备交换机)配置为处理由控制器发送的探测消息(框707)。

[0083] 在L3链路实现的示例中，初始控制平面步骤确定要被用于探测消息的探测消息多

协议标签交换(MPLS)标签,该标签可以是可配置的。控制器在目标交换机(即,图6的示例中的交换机1)上安装探测响应出口组。控制器使用以下动作在流控制(例如OpenFlow)组表中添加具有单个桶的条目:将源IP地址(Src_IP)设置为交换机1IP,将目的地(Dest_IP)地址设置为交换机2IP,将源媒体访问控制地址(Src MAC)设置为交换机1MAC地址,以及将Dest MAC设置为朝向交换机2(和控制器)的下一跳交换机的MAC地址。设置朝向下一跳交换机的输出端口。

[0084] 控制器还在交换机1中安装探测请求入口流。作为该过程的一部分,它在OpenFlow表0中添加具有以下匹配字段的条目:探测MPLS标签、将传入分组标识为来自交换机2的探测分组的源IP(交换机2的IP地址)以及Goto probe response egress group(转到探测响应出口组)指令。

[0085] 控制器类似地在交换机2中安装探测响应入口流。控制器在OpenFlow表0中添加具有以下匹配字段的条目:MPLS探测标签、源IP(交换机1的IP地址)、以及转发给控制器的指令。

[0086] 相对于图6的简化SDN网络,给出了控制平面配置的该示例。本领域技术人员将理解,具有该上下文的第3层控制平面配置是作为示例而非限制给出的,并且这些原理、过程和结构适用于其他上下文。

[0087] 控制器针对网络的第二层实现了类似的控制平面配置。探测L2链路涉及以下步骤来建立控制平面。控制器确定要用于探测的提供商骨干网桥(PBB)探测服务标识符(ISID)标记,该标记可以是可配置的。控制器在交换机1中安装探测响应出口组。控制器使用以下动作在流控制(例如OpenFlow)组表中添加具有单个桶的条目:将Src MAC设置为交换机1MAC,将Dest MAC设置为交换机2的MAC地址,以及在该端口上向交换机2输出动作端口。

[0088] 控制器还在交换机1中安装探测请求入口流。控制器在流控制(例如OpenFlow)表0中添加具有以下匹配字段的条目,这些匹配字段为:PBB探测ISID标记;将传入分组标识为来自交换机2的探测分组的源MAC(交换机2的MAC地址);以及Goto probe response egress group(转到探测响应出口组)指令。

[0089] 控制器在交换机2中安装探测响应入口流。这涉及在流控制(例如OpenFlow)表0中添加具有以下匹配字段的条目:PBB探测ISID标记;源MAC(交换机1的MAC地址);以及向控制器转发该分组的指令。

[0090] 相对于图6的简化SDN网络,给出了控制平面配置的该示例。本领域技术人员将理解,具有该上下文的第2层控制平面配置是作为示例而非限制给出的,并且这些原理、过程和结构适用于其他上下文。

[0091] 图7B是响应于SDN网络中给定交换机的丢失控制信道而实施的过程的概括描述的流程图。概括说明了该过程,并且以下详细介绍了特定的L3/L2实现。该过程始于控制器检测与SDN网络中的给定交换机的控制信道故障(框751)。SDN控制器可以使用任何机制(包括如上所述的活动性检查或经由代理接收控制信道通信)来确定控制信道故障。这生成要被发送给已经丢失与其的控制信道通信的目标SDN交换机的探测请求消息(框753)。然后,该过程将探测请求消息封装在诸如packet_out消息之类的控制信道消息内(框755)。该控制信道消息可以例如关于其元数据被进一步修改,然后经由替代路径/交换机被转发给目标交换机(框757)。

[0092] 中间交换机(例如,图6的交换机2)接收控制信道消息,并将其转发给目标交换机(例如,图6的交换机1)(框759)。然后,目标交换机接收具有所封装的探测请求消息的控制信道消息(框761)。该探测请求被分析,并且探测答复消息被生成并被发送回中间交换机(框763)。中间交换机接收探测响应消息(框765),并将探测响应消息封装为要被发送给控制器的控制信道消息(框767)。

[0093] 控制器接收控制信道消息并分析探测响应信息(框769)。然后,该探测响应信息可被用于确定和更新SDN网络的网络拓扑(框771)。接收到该探测响应表明目标交换机活跃,并且目标交换机和中间交换机之间的链路正在工作。

[0094] 对于第3层,当与目标交换机(例如,交换机1)的控制信道会话丢失时,可以实现以下步骤。控制器生成packet-out消息,并将其发送给中间交换机(例如,交换机2)。具体地说,控制器创建具有交换机2IP地址的src_IP和dest_IP作为交换机1IP地址的“假(dummy)”L3分组。控制器将假L3分组封装在packet-out消息中。在packet-out消息中指定以下动作:将MPLS报头推送到该假分组,将Tag Probe MPLS标签设置到假分组,以及向交换机1输出用于该端口的端口动作。然后,控制器将packet-out消息发送给交换机2,交换机2向交换机1发送该假分组。

[0095] 在中间(交换机2)交换机处,该交换机接收该假分组并向交换机1转发。在交换机1处,该假分组(即,该探测请求)被处理。基于在交换机1中的流表0条目,对探测MPLS标签和src_IP(即交换机2IP地址)进行匹配。响应于该匹配,交换机1将分组发送给探测响应出口组。作为探测响应出口组的一部分,该过程按照动作中的规定来设置源和目的地IP地址、源和目的地MAC地址,并在适当的端口上将它们发送出,以将分组发送给朝向中间交换机2的下一跳。

[0096] 在中间交换机处,该交换机被配置为接收并处理探测响应。基于在交换机1中的表0条目,对探测MPLS标签和源IP地址(例如,交换机1IP地址)进行匹配或比较。如果值匹配,则分组被转发给控制器。

[0097] 在控制器处,在接收到作为packet-in的探测响应时,控制器检查探测MPLS标签、源IP和目的地IP,以确定在交换机1与2之间的链路正在工作。然后,控制器将拓扑链路更新为交换机1和2之间的链路处于UP(工作)状态。如果没有接收到针对被发送以测试该链路和目标交换机的packet-out的packet-in,则控制器将拓扑链路更新为交换机1和2之间的链路发生故障(DOWN)。

[0098] 相对于图6的简化SDN网络,给出了数据平面操作的该示例。本领域技术人员将理解,具有该上下文的第3层数据平面操作是作为示例而非限制给出的,并且这些原理、过程和结构适用于其他上下文。

[0099] 对于第2层,当与目标交换机(例如,交换机1)的控制信道会话丢失时,使用以下过程。控制器生成packet-out或类似的控制信道消息,并将其发送给交换机2。具体地说,控制器创建具有源MAC作为交换机2以及目的地MAC地址作为交换机1的假L2帧。控制器将假L2帧封装在packet-out消息或类似控制信道消息中。控制器为packet-out消息或类似的控制信道消息指定以下动作:将PBB报头推送到假帧,将标记探测ISID标记设置到假帧,以及向交换机1输出用于端口的端口动作。控制器将此控制信道消息(packet-out消息)发送给交换机2,交换机2朝向交换机1发送该假帧。

[0100] 然后,转发的帧由目标交换机(交换机1)接收和处理。假帧(即,探测请求)由目标交换机的流控制管道处理。基于交换机1中的流表0条目,对PBB探测ISID标签和源MAC(交换机2MAC地址)进行匹配。如果找到匹配,则目标交换机将帧发送给探测响应出口组,从而创建探测响应。作为探测响应出口组的一部分,该过程按照动作中的规定来设置该帧的源和目的地MAC地址,并在适当的端口上发出该帧,以将该帧发送给中间交换机(交换机2)。

[0101] 中间交换机接收并处理探测响应。接收到的帧(探测响应)在流控制管道中基于交换机2中的表0条目被处理,该条目被与PBB Probe ISID标记和交换机1的源MAC相匹配。如果找到匹配,则该帧被转发给控制器作为控制信道消息(例如,packet-in消息)。

[0102] 在控制器处,所接收的控制信道消息(例如,packet-in消息)被处理。接收到该帧时,将检查PBB探测ISID标记、源MAC和目的地MAC,以确定目标交换机1与中间交换机2之间的链路是否工作以及目标交换机是否活动。如果链路和交换机正在工作,则控制器将拓扑链路状态更新为在交换机1与交换机2之间为工作(UP)。如果未接收到对给定packet-out的响应帧(即,未接收到packet-in),则控制器将拓扑链路更新为在交换机1与交换机2之间发生故障(DOWN)。

[0103] 相对于图6的简化SDN网络,给出了数据平面操作的该示例。本领域技术人员将理解,具有该上下文的第2层数据平面操作是作为示例而非限制给出的,并且这些原理、过程和结构适用于其他上下文。

[0104] 因此,链路监视实施例提供了一种过程和系统,当交换机与控制器之间的控制连接丢失时,针对终止于该交换机的每个链路,在连接到目标交换机的每个链路的另一端处发送来自该交换机的探测。如果目标交换机活动,则远程交换机获得被发送给控制器的探测响应。控制器可以基于接收到的packet-in来确定链路状态。

[0105] 图8A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8A示出了ND 800A-H及其在800A-800B、800B-800C、800C-800D、800D-800E、800E-800F、800F-800G和800A-800G之间以及在800H与800A、800C、800D和800G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 800A、800E和800F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)。

[0106] 图8A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802;以及2)使用通用现成(COTS)处理器和标准OS的通用网络设备804。

[0107] 专用网络设备802包括联网硬件810,联网硬件810包括一组一个或多个处理器812、转发资源814(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)816(通过其进行网络连接,例如ND 800A-H之间的连接所示)、以及在其中存储网络软件820的非暂时性机器可读存储介质818。在运行期间,联网软件820可以由联网硬件810执行以实例化一组一个或多个网络软件实例822。每个联网软件实例822以及联网硬件810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822在时间上共享的硬件的时间片)形成单独的虚拟网元830A-R。每个虚拟网元(VNE)830A-R均包括控制通信和配置模块832A-R(有时称为本地控制模块或控制通信模块)

和转发表834A-R，使得给定的虚拟网元(例如830A)包括该控制通信和配置模块(例如832A)、一组一个或多个转发表(例如834A)以及联网硬件810的执行虚拟网元(例如830A)的部分。

[0108] 专用网络设备802通常在物理和/或逻辑上被认为包括：1)ND控制平面824(有时称为控制平面)，其包括执行控制通信和配置模块832A-R的处理器812；以及2)ND转发平面826(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834A-R和物理NI 816的转发资源814。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824(执行控制通信和配置模块832A-R的处理器812)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834A-R中，ND转发平面826负责在物理NI 816上接收该数据并基于转发表834A-R来将该数据转发出适当的物理NI 816。

[0109] 图8B示出了根据本发明的一些实施例的实现专用网络设备802的示例性方式。图8B示出了包括卡838(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838具有两种类型(一种或多种用作ND转发平面826(有时称为线卡)，一种或多种用来实现ND控制平面824(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836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)。

[0110] 返回图8A，通用网络设备804包括硬件840，硬件840包括一组一个或多个处理器842(通常是COTS处理器)和物理NI 846、以及在其中存储有软件850的非暂时性机器可读存储848。在运行期间，处理器842执行软件850以实例化一组或多组一个或多个应用864A-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854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862A-R，每个实例可用于执行一组(或多组)应用864A-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854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864A-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862A-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840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854表示的系统管理程序上直接运行的单内核、通过在实例862A-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)。

[0111] 一个或多个组的一个或多个应用864A-R的实例化以及虚拟化(如果实现的话)被统称为软件实例852。每组应用864A-R、对应的虚拟化构造(例如实例862A-R)(如果实现的话)以及执行它们的那部分硬件840(无论是专用于该执行的硬件和/或临时共享的硬件的时间片)形成单独的虚拟网元860A-R。在实施例中，应用864A-R可以包括控制信道重定向器864A-R，其实现本文以上针对网络设备处的控制信道重定向和探测过程的实施例所描述的功能的任何组合或集合。

[0112] 虚拟网元860A-R执行与虚拟网元830A-R类似的功能，例如，类似于控制通信和配置模块832A以及转发表834A(硬件840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 860A-R对应的每个实例862A-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862A-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。

[0113] 在特定实施例中，虚拟化层854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862A-R与物理NI 846之间以及可选地在实例862A-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 860A-R之间实施网络隔离。

[0114] 图8A中的第三示例性ND实现是混合网络设备806，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802的功能的VM)可以为混合网络设备806中存在的联网硬件提供半虚拟化。

[0115] 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 830A-R、VNE 860A-R和混合网络设备806中的那些VNE)在物理NI(例如816、846)上接收数据，然后将该数据转发出适当的物理NI(例如816、846)。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)。

[0116] 图8C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8C示出了在ND 800A中实现的VNE 870A.1-870A.P(以及可选地VNE 870A.Q-870A.R)和在ND 800H中实现的VNE 870H.1。在图8C中，VNE 870A.1-P彼此分离，因为它们可以接收来自

ND 800A外部的分组并将分组转发到ND 800A外部;VNE 870A.1与VNE 870H.1耦合,并且因此它们在各自的ND之间传送分组;VNE 870A.2-870A.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 800A之外;以及VNE 870A.P可以可选地是包括VNE 870A.Q(后跟VNE870A.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8C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)。

[0117] 例如,图8A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8A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804的情况下,一个或多个软件实例862A-R可以充当服务器;混合网络设备806也是如此;在专用网络设备802的情况下,一个或多个这样的服务器也可以在由处理器812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)。

[0118] 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8A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)。

[0119] 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。

[0120] 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)。

[0121] 图8D示出了根据本发明的一些实施例在图8A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8D示出了具有与图8A的ND 800A-H相同的连接性的网元(NE)870A-H。

[0122] 图8D示出了分布式方法872跨NE 870A-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。

[0123] 例如,如果使用专用网络设备802,ND控制平面824的控制通信和配置模块832A-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 870A-H(例如执行控制通信和配置模块832A-R的处理器812)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824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824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826进行编程。例如,ND控制平面824将邻接和路由信息编程到ND转发平面826上的一个或多个转发表834A-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802,但是可以在通用网络设备804和混合网络设备806上实现相同的分布式方法872。

[0124] 图8D示出了一种使系统解偶的集中式方法874(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874负责在集中式控制平面876(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876具有与包括NE 870A-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880(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882。集中式控制平面876包括网络控制器878,网络控制器878包括集中式可达性和转发信息模块879,集中式可达性和转发信息模块879确定网络内的可达性并将转发信息在南向接口882(其可以使用OpenFlow协议)上分发给数据平面880的NE 870A-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876中。

[0125] 例如,在数据平面880中使用专用网络设备802的情况下,ND控制平面824的每个控制通信和配置模块832A-R通常包括提供南向接口882的VNE侧的控制代理。在这种情况下,ND控制平面824(执行控制通信和配置模块832A-R的处理器812)通过与集中式控制平面876通信以从集中式可达性和转发信息模块879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876通信之外,控制通信和配置模块832A-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874,但是也可以被认为是混合方法。)

[0126] 尽管以上示例使用专用网络设备802,但是可以用通用网络设备804实现相同的集中式方法874(例如,每个VNE 860A-R通过与集中式控制平面876进行通信以从集中式可达性和转发信息模块879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876通信外,VNE 860A-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806。实际上,SDN技术的使用能够增强通常在通用网络设备804或混合网络设备806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。

[0127] 图8D还示出了集中式控制平面876具有到其中驻留有应用888的应用层886的北向接口884。集中式控制平面876具有形成用于应用888的虚拟网络892(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880的NE 870A-H)的能力。因此,集中式控制平面876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)。在实施例中,应用888可以包括控制信道重定向器881,其实现本文以上针对网络设备和控制器处的控制信道重定向和探测过程的实施例所描述的功能的任何组合或集合。

[0128] 尽管图8D示出了与集中式方法874分离的分布式方法872,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)874,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874,但是也可以被认为是混合方法。

[0129] 尽管图8D示出了每个ND 800A-H实现单个NE 870A-H的简单情况,但是应当理解,参考图8D所描述的网络控制方法也适用于其中一个或多个ND 800A-H实现多个VNE(例如VNE 830A-R、VNE 860A-R、混合网络设备806中的那些VNE)的网络。备选地或附加地,网络控制器878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878还可以将单个ND中的VNE/NE的实现呈现为虚拟网络892中的多个VNE(全部在同一虚拟网络892中、每个都在不同的虚拟网络892中、或某种组合)。例如,网络控制器878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876内在逻

辑上划分该NE的资源以在虚拟网络892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)。

[0130] 另一方面,图8E和8F分别示出了网络控制器878可以作为不同虚拟网络892的一部分而呈现的NE和VNE的示例性抽象。图8E示出了根据本发明的一些实施例的其中每个ND 800A-H实现单个NE 870A-H(参见图8D)但是集中式控制平面876已将不同ND中的多个NE(NE 870A-C和G-H)抽象成(表示为)图8D的虚拟网络892之一中的单个NE 870I的简单情况。图8E示出了在该虚拟网络中,NE 870I耦合至NE 870D和870F,NE 870D和870F两者仍然耦合至NE 870E。

[0131] 图8F示出了根据本发明的一些实施例的其中多个VNE(VNE 870A.1和VNE 870H.1)在不同的ND(ND 800A和ND 800H)上实现并且彼此耦合并且其中集中式控制平面876已经抽象这多个VNE以使得它们在图8D的虚拟网络892之一中表现为单个VNE 870T的情况。因此,NE或VNE的抽象可以跨越多个ND。

[0132] 尽管本发明的一些实施例将集中式控制平面876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)。

[0133] 类似于网络设备实现,运行集中式控制平面876的电子设备以及因此包括集中式可达性和转发信息模块879的网络控制器878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9示出了通用控制平面设备904,其包括硬件940,硬件940包括一组一个或多个处理器942(其通常是COTS处理器)和物理NI 946,以及其中存储有集中式控制平面(CCP)软件950的非暂时性机器可读存储介质948。

[0134] 在使用计算虚拟化的实施例中,处理器942通常执行软件以实例化虚拟化层954(例如在一个实施例中,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)),其允许创建多个实例962A-R,实例962A-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954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962A-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940上运行、在由虚拟化层954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962A-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954上执行(例如在实例962A内)CCP软件950的实例(图示为CCP实例976A)。在不使用计算虚拟化的实施例中,CCP实例976A在“裸机”通用控制平面设备904上作为单内核或在主机操作系统之上执行。CCP实例976A以及虚拟化层954和实例962A-R(如果已实现的话)的实例化统称为软件实例952。

[0135] 在一些实施例中,CCP实例976A包括网络控制器实例978。网络控制器实例978包括

集中式可达性和转发信息模块实例979(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980(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980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。在实施例中,应用层980可以包括控制信道重定向器981,其实现本文以上针对网络设备和控制器处的控制信道重定向和探测过程的实施例所描述的功能的任何组合或集合。

[0136] 集中式控制平面876基于CCP应用层980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880。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880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880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。

[0137] 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)。

[0138] 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。

[0139] 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880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876。集中式控制平面876然后将转发表条目编程到数据平面880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876编程到数据平面880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。

[0140] 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。

[0141] 例如,尽管附图中的流程图示出了由本发明的某些实施例执行的操作的特定顺序,但是应当理解,这种顺序是示例性的(例如,替代实施例可以以不同的顺序执行操作、合并某些操作、重叠某些操作等)。

[0142] 虽然已经根据若干实施例描述了本发明,但是本领域技术人员将认识到,本发明不限于所描述的实施例,而是可以在所附权利要求的精神和范围内通过修改和变更来实践。因此,该描述被认为是说明性的而非限制性的。

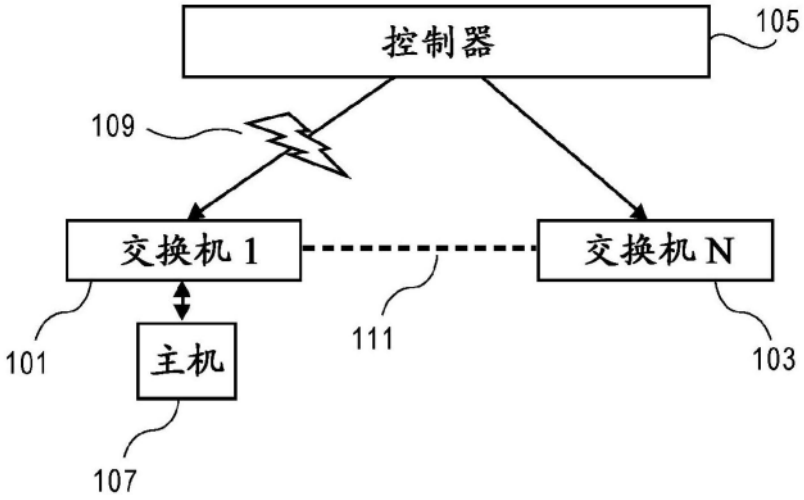

控制器
105
109
交换机 1
交换机 N
101
主机
111
103
107

图1

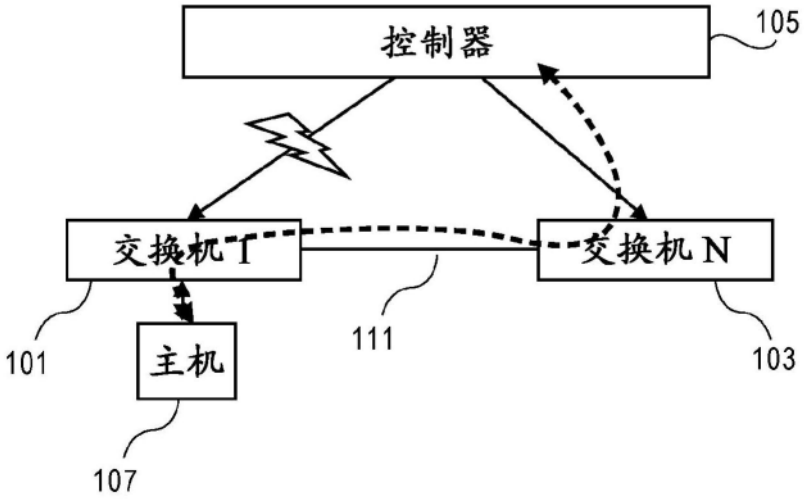

控制器
105
交换机 1
交换机 N
101
主机
111
103
107

图2

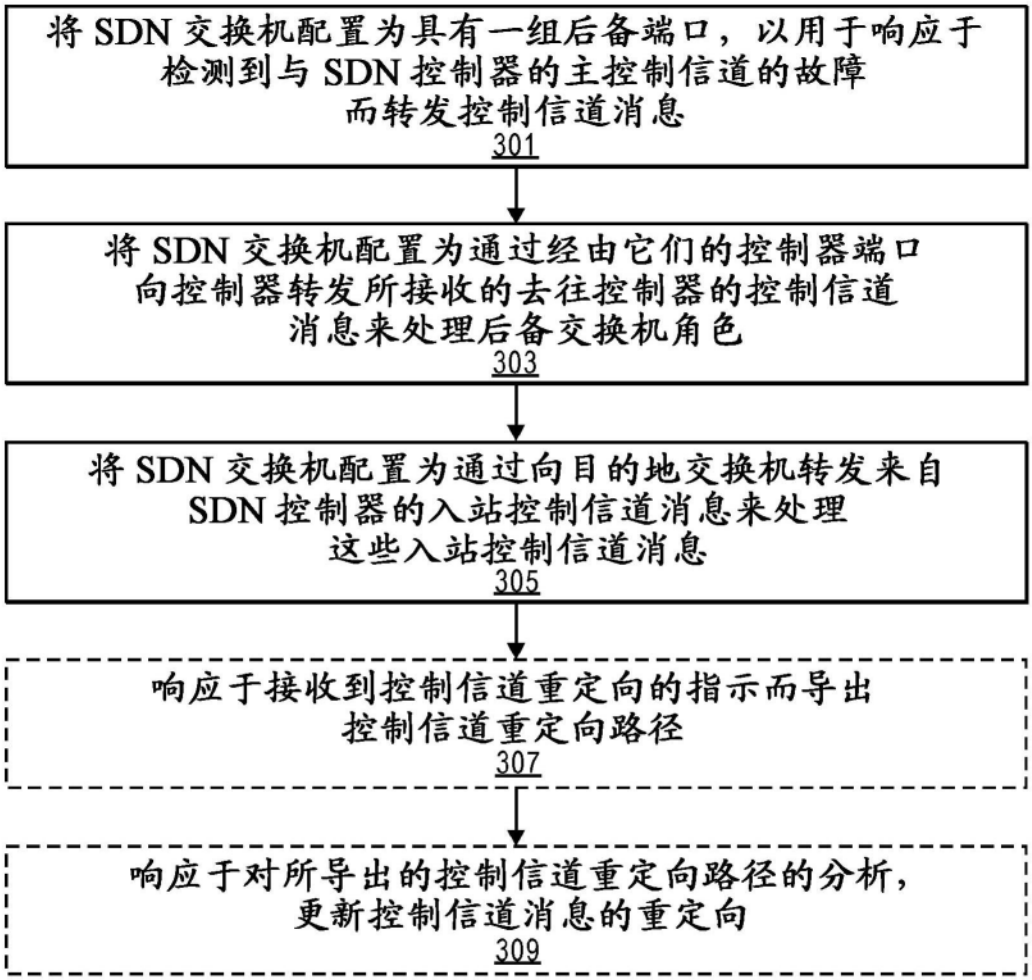

将SDN交换机配置为具有一组后备端口，以用于响应于检测到与SDN控制器的主控制信道的故障而转发控制信道消息
301
将SDN交换机配置为通过经由它们的控制器端口向控制器转发所接收的去往控制器的控制信道消息来处理后备交换机角色
303
将SDN交换机配置为通过向目的地交换机转发来自SDN控制器的入站控制信道消息来处理这些入站控制信道消息
305
响应于接收到控制信道重定向的指示而导出控制信道重定向路径
307
响应于对所导出的控制信道重定向路径的分析，更新控制信道消息的重定向
309

图3A

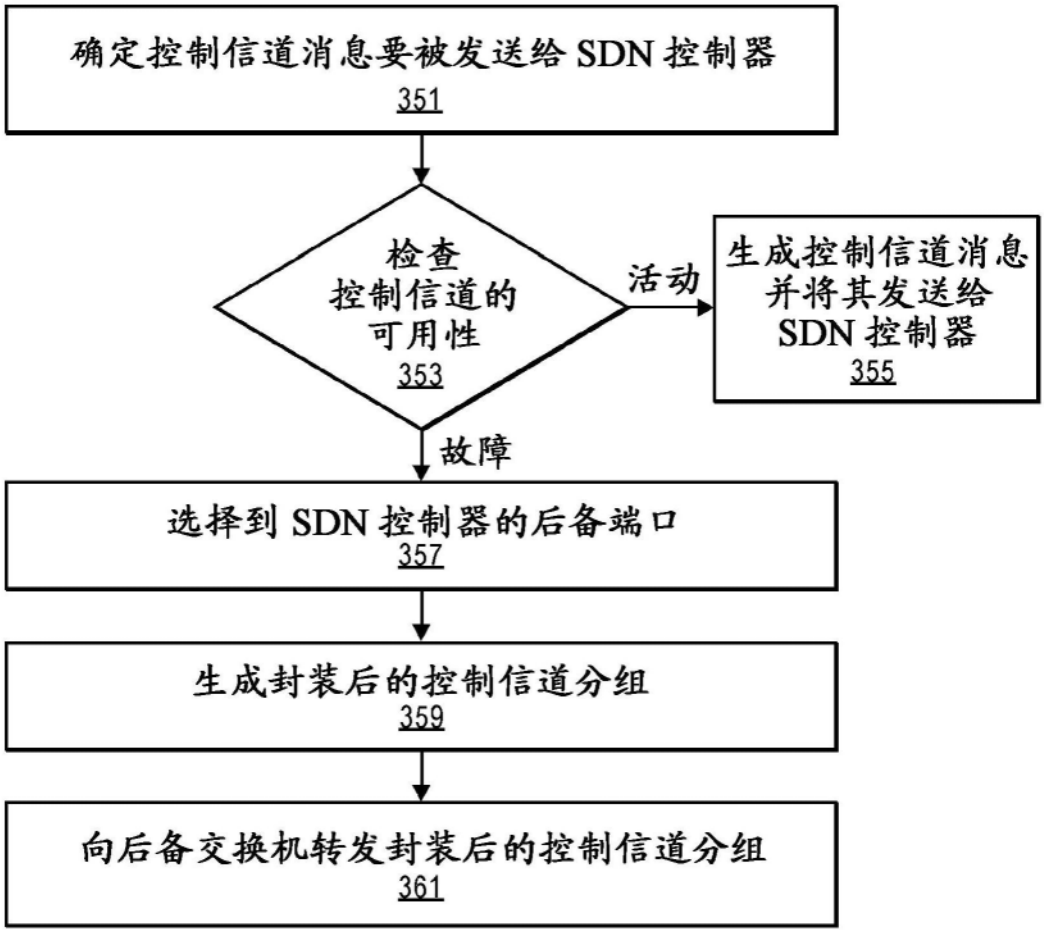

确定控制信道消息要被发送给 SDN 控制器
351
检查
控制信道的
可用性
353
活动
生成控制信道消息
并将其发送给
SDN 控制器
355
故障
选择到 SDN 控制器的后备端口
357
生成封装后的控制信道分组
359
向后备交换机转发封装后的控制信道分组
361

图3B

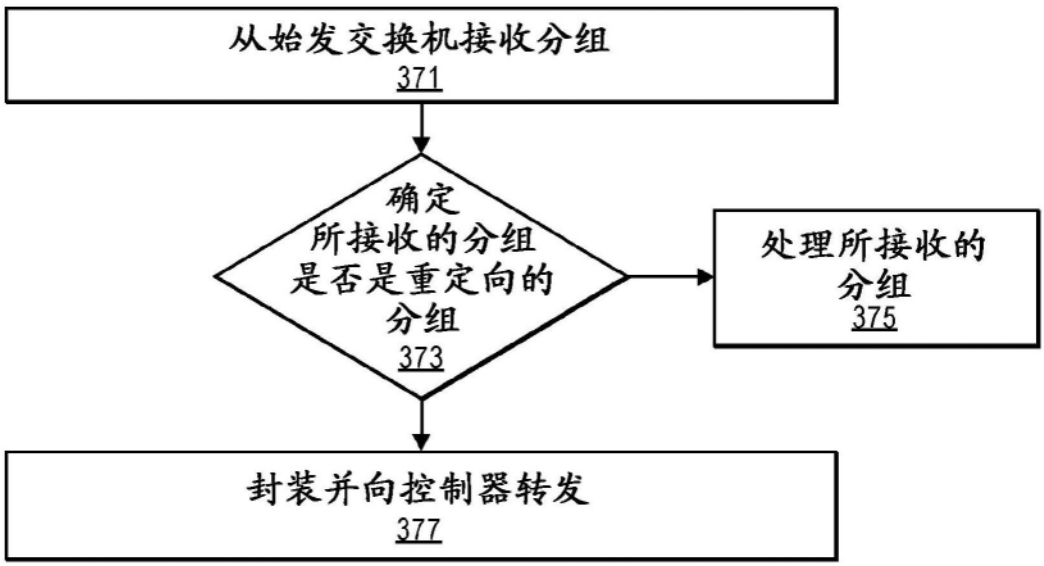

从始发交换机接收分组
371
确定
所接收的分组
是否是重定向的
分组
373
处理所接收的
分组
375
封装并向控制器转发
377

图3C

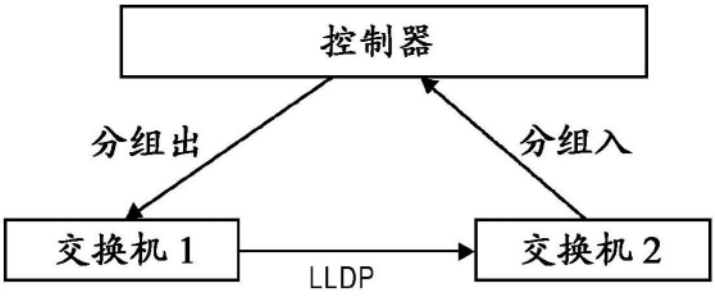

控制器
分组出
分组入
交换机 1
交换机 2
LLDP

图4

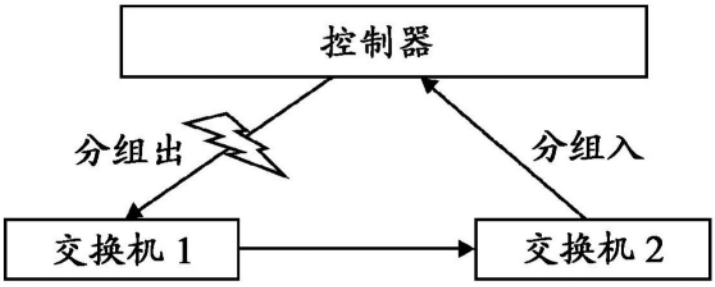

控制器
分组出
分组入
交换机 1
交换机 2

图5

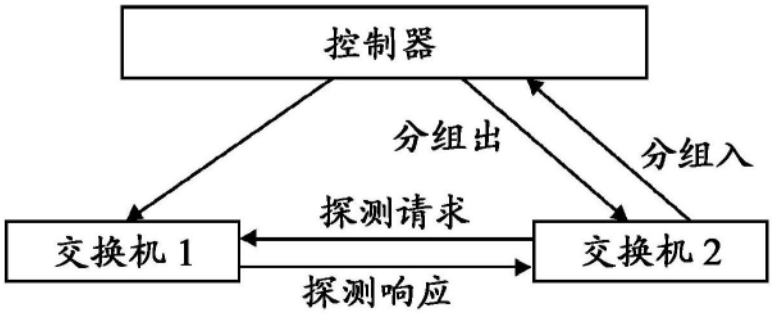

控制器
分组出
分组入
探测请求
交换机 1
交换机 2
探测响应

图6

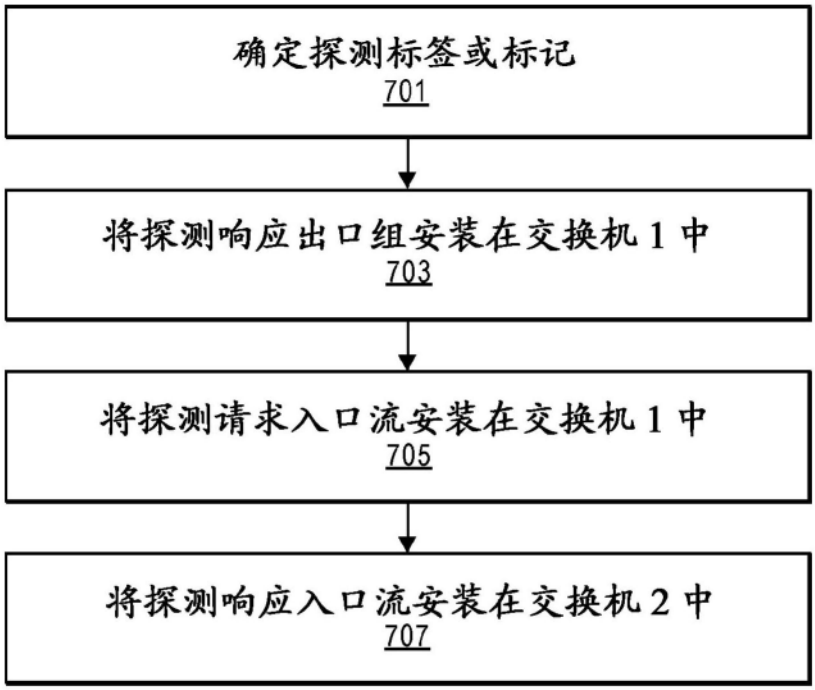

确定探测标签或标记
701
将探测响应出口组安装在交换机1中
703
将探测请求入口流安装在交换机1中
705
将探测响应入口流安装在交换机2中
707

图7A

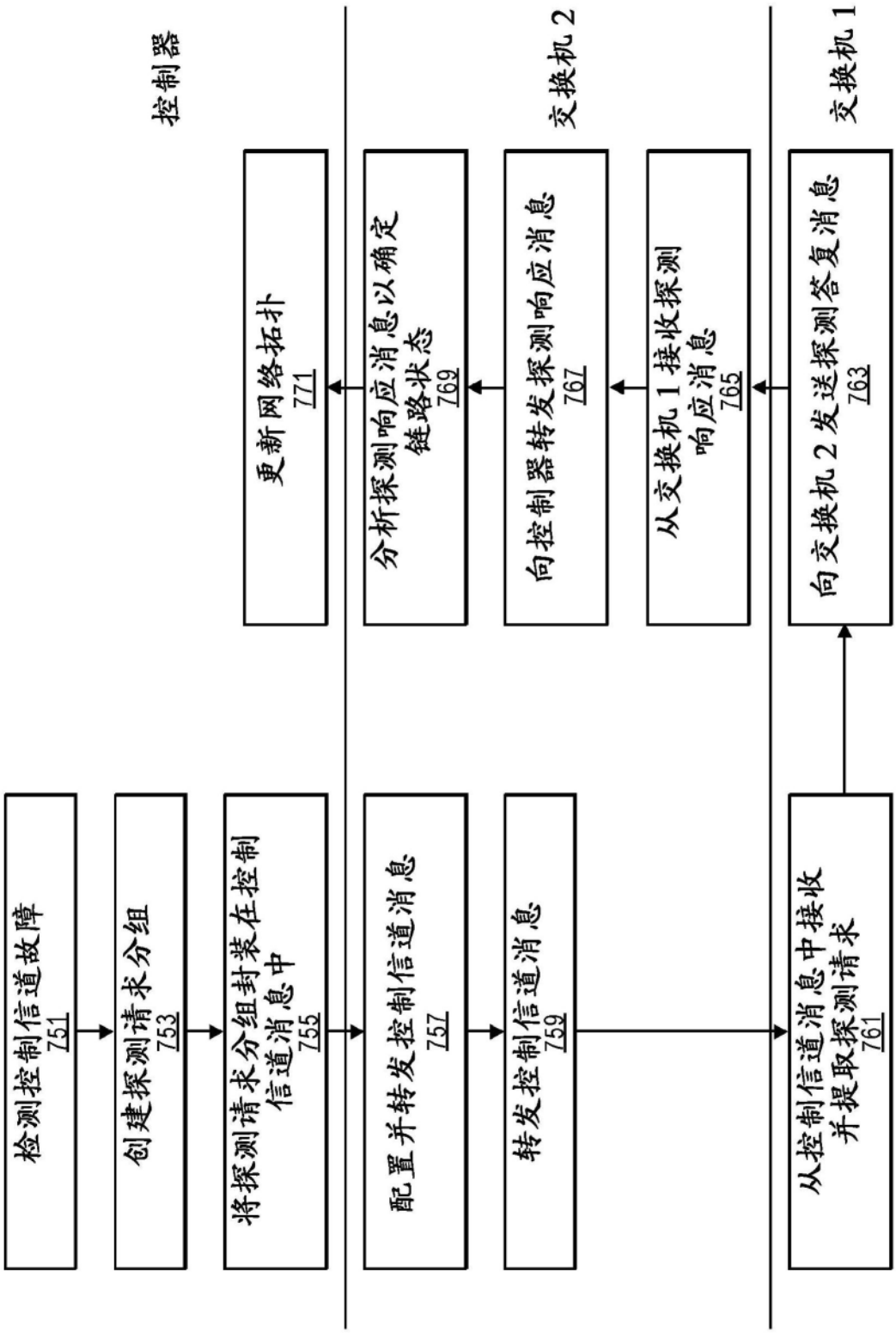

控制器
交换机 2
交换机 1
检测控制信道故障
751
创建探测请求分组
753
将探测请求分组封装在控制信道消息中
755
配置并转发控制信道消息
757
转发控制信道消息
759
从控制信道消息中接收并提取探测请求
761
更新网络拓扑
771
分析探测响应消息以确定链路状态
769
向控制器转发探测响应消息
767
从交换机 1 接收探测响应消息
765
向交换机 2 发送探测答复消息
763

图7B

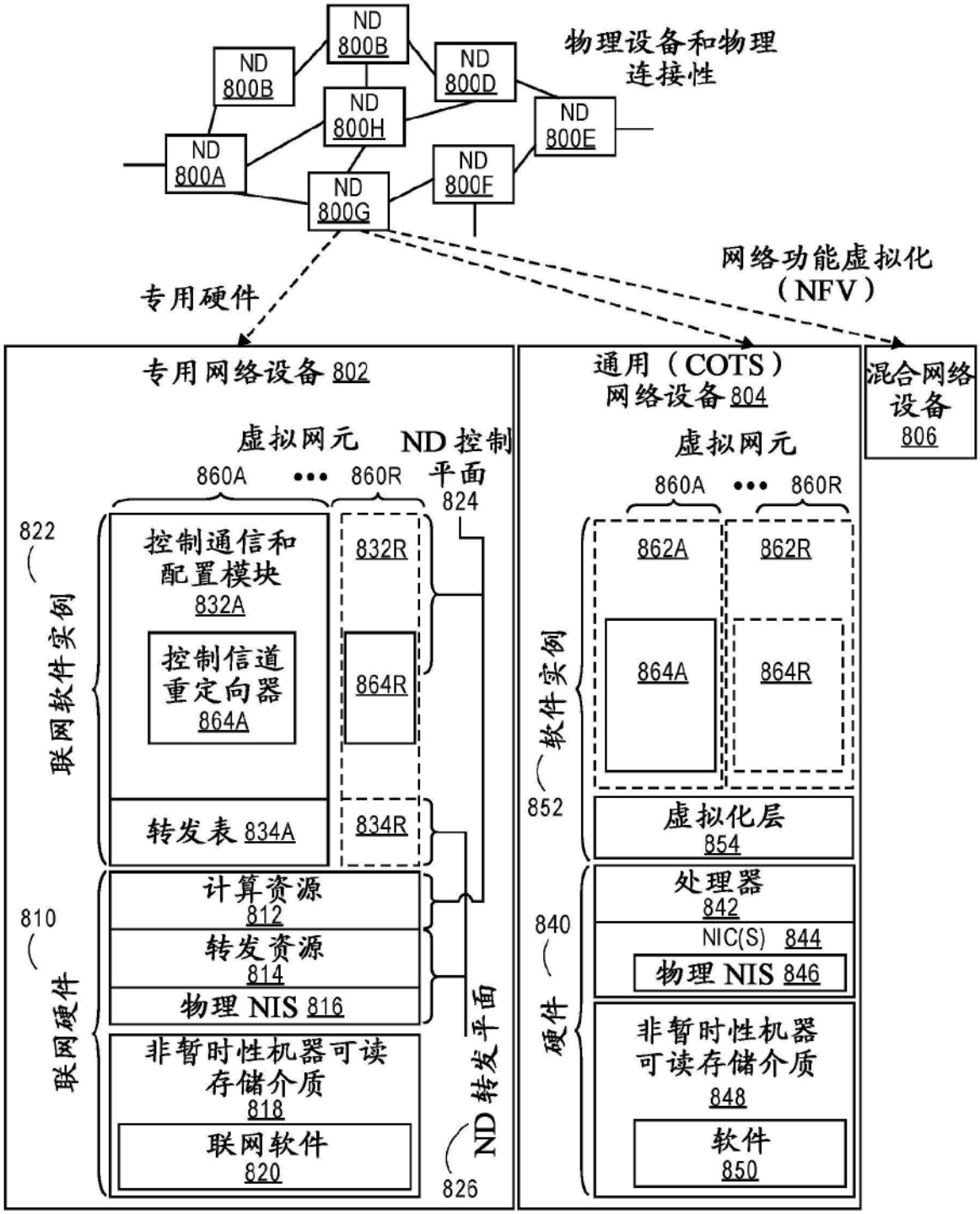

ND 800A
ND 800B
ND 800B
ND 800D
ND 800E
ND 800F
ND 800G
ND 800H
物理设备和物理连接性
专用硬件
网络功能虚拟化（NFV）
专用网络设备 802
虚拟网元
ND 控制平面
860A
860R
824
822
联网软件实例
控制通信和配置模块 832A
控制信道重定向器 864A
832R
864R
转发表 834A
834R
计算资源 812
转发资源 814
物理 NIS 816
810
联网硬件
非暂时性机器可读存储介质 818
联网软件 820
ND 转发平面
826
通用（COTS）网络设备 804
虚拟网元
860A
860R
862A
862R
864A
864R
852
软件实例
虚拟化层 854
处理器 842
840
NIC(S) 844
物理 NIS 846
硬件
非暂时性机器可读存储介质 848
软件 850
混合网络设备 806

图8A


卡838
背板836

图8B

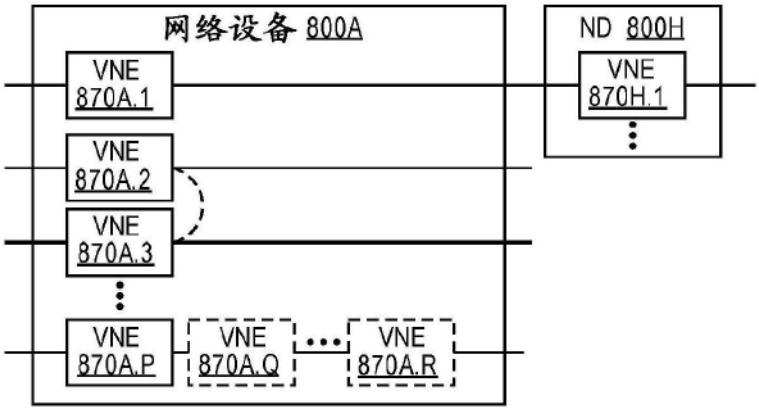

网络设备800A
ND 800H
VNE 870A.1
VNE 870A.2
VNE 870A.3
VNE 870A.P
VNE 870A.Q
VNE 870A.R
VNE 870H.1

图8C

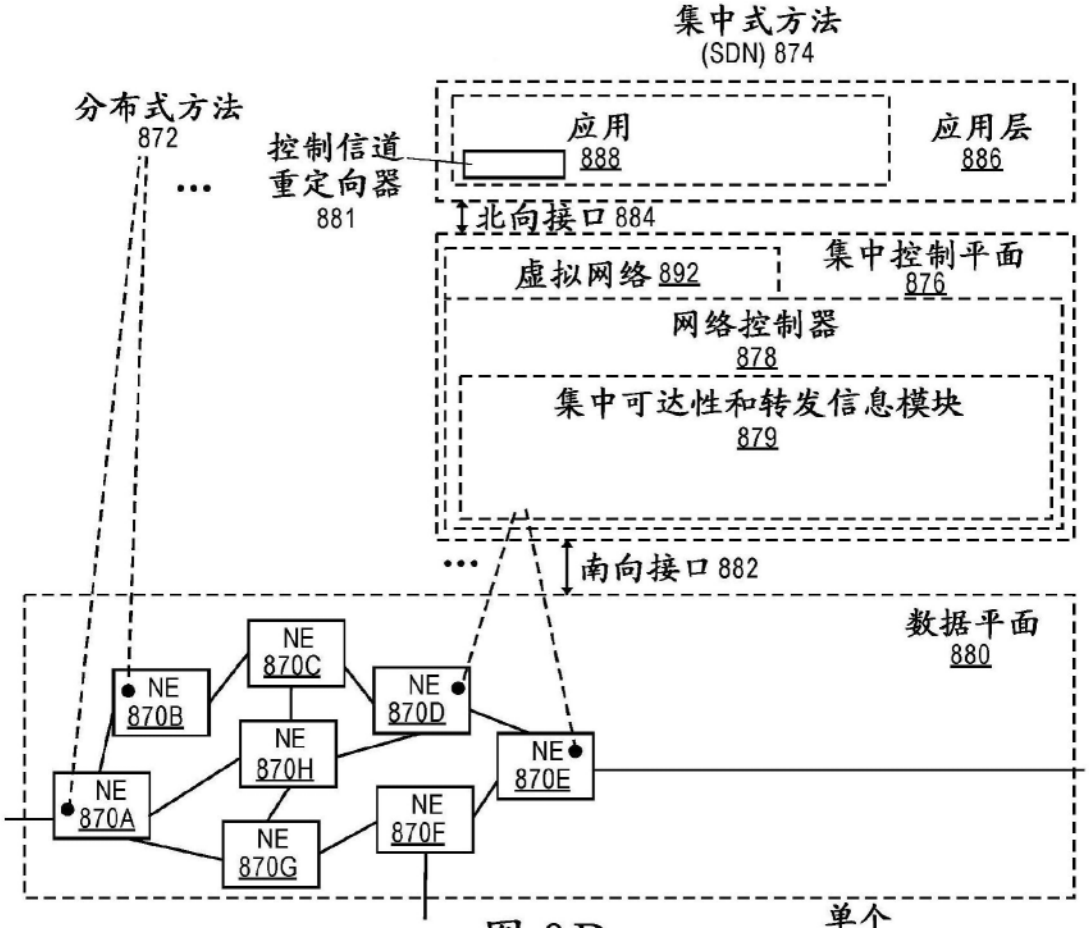

集中式方法
(SDN) 874
分布式方法
872
控制信道
重定向器
881
应用
888
应用层
886
北向接口 884
虚拟网络 892
集中控制平面
876
网络控制器
878
集中可达性和转发信息模块
879
南向接口 882
数据平面
880
NE
870A
NE
870B
NE
870C
NE
870D
NE
870E
NE
870F
NE
870G
NE
870H

图 8D

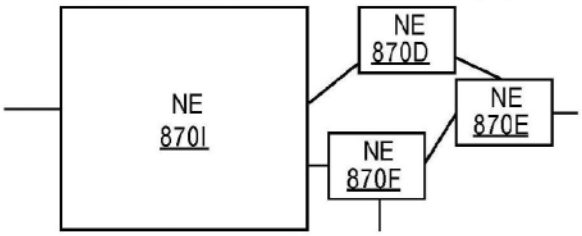

NE
870I
NE
870D
NE
870E
NE
870F

图 8E

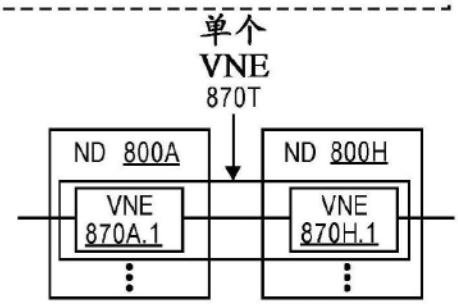

单个
VNE
870T
ND 800A
VNE
870A.1
ND 800H
VNE
870H.1

图 8F

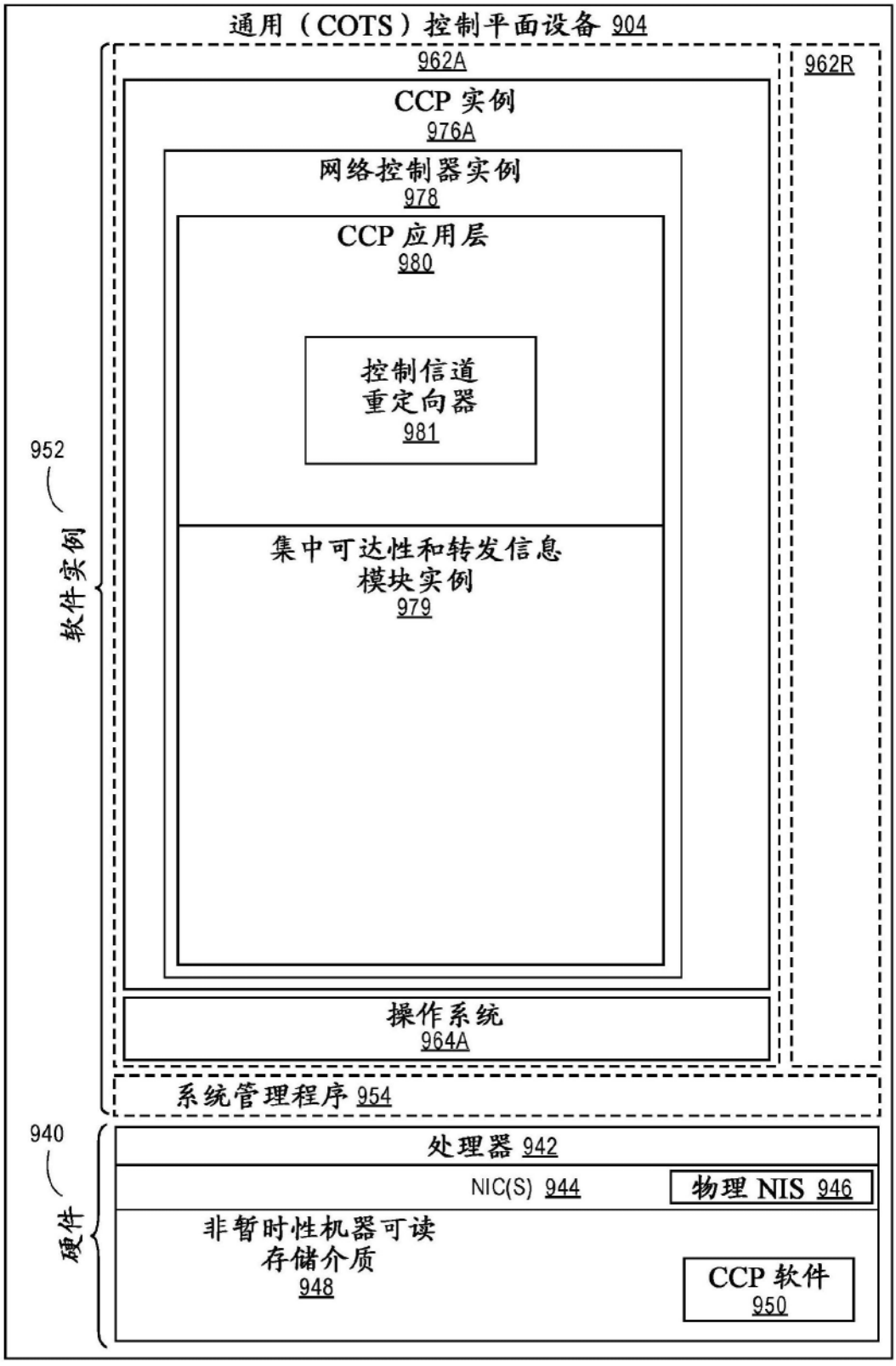

通用（COTS）控制平面设备 904
962A
962R
CCP 实例
976A
网络控制器实例
978
CCP 应用层
980
控制信道
重定向器
981
集中可达性和转发信息
模块实例
979
952
软件实例
操作系统
964A
系统管理程序 954
940
硬件
处理器 942
NIC(S) 944
物理 NIS 946
非暂时性机器可读
存储介质
948
CCP 软件
950

图9