(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107005439 B
(45)授权公告日 2020.05.05

(21)申请号 201580062959.0

(22)申请日 2015.11.20

(65)同一申请的已公布的文献号
申请公布号 CN 107005439 A

(43)申请公布日 2017.08.01

(30)优先权数据
14/549373 2014.11.20 US

(85)PCT国际申请进入国家阶段日
2017.05.19

(86)PCT国际申请的申请数据
PCT/IB2015/059021 2015.11.20

(87)PCT国际申请的公布数据
WO2016/079722 EN 2016.05.26

(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 张颖 J.哈尔佩恩

(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 姜冰 付曼

(51)Int.Cl.
H04L 12/26(2006.01)
H04B 1/00(2006.01)
H04B 17/00(2015.01)
H04L 12/801(2013.01)

(56)对比文件
CN 103795596 A,2014.05.14,
CN 104092774 A,2014.10.08,
CN 102684940 A,2012.09.19,
US 2014105038 A1,2014.04.17,
薛淼,等."基于SDN的SGi/Gi—LAN Service Chain 关键技术研究".《2014全国无线及移动通信学术大会论文集》.2014,第151-154页.
Q. Wu,D. Wang等."Service Function Chain Control Plane Overview".《Network Working Group Internet-Draft》.2014,第5-15页.

审查员 王淑婷

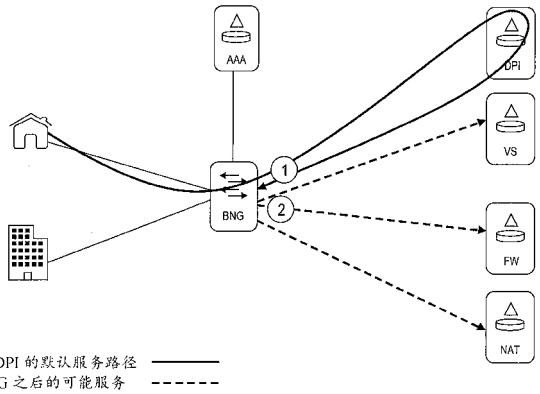
权利要求书2页 说明书24页 附图14页

(54)发明名称

用于在线服务链接的被动性能测量

(57)摘要

一种由计算装置实现的方法,用来监视在线服务链中分组处理的性能。所述计算装置与形成软件定义网络(SDN)和所述在线服务链的多个网络装置通信。所述SDN包含由所述计算装置实现的用来配置所述多个网络装置的控制器。所述多个装置包含对遍历包含至少一个服务的在线服务链的分组进行监视的交换机的集合。



CN 107005439 B

1.一种由计算装置实现的方法,用来监视在线服务链中的分组处理的性能,所述计算装置与形成软件定义网络SDN和所述在线服务链的多个网络装置通信,所述SDN包含由所述计算装置实现的用来配置所述多个网络装置的控制器,所述多个网络装置包含对遍历包含至少一个服务的所述在线服务链的分组进行监视的一组交换机,所述方法包括如下步骤:

检查(705)分组在遍历所述至少一个服务之后是否丢失了;

添加(706)交换机分组丢失记分,其中所述分组丢失了;

添加(707)交换机分组延迟记分,其中所述分组未丢失;

根据对应交换机分组丢失记分,对所述一组交换机的列表进行排序(713),以生成排序的丢失列表;

根据对应交换机延迟记分,对所述一组交换机的列表进行排序(719),以生成排序的延迟列表;以及

根据所述排序的丢失列表和所述排序的延迟列表中的次序,对所述一组交换机的列表进行排序(725)。

2.如权利要求1所述的方法,所述方法进一步包括如下步骤:

检查(715)是否能使所述排序的丢失列表中的交换机诱发延迟或停止数据流;以及

将所述交换机移动(717)到所述交换机被使能所在的所述排序的丢失列表的末端。

3.如权利要求1所述的方法,所述方法进一步包括如下步骤:

检查(719)是否所述排序的延迟列表中的交换机被使能诱发对于数据流的延迟;以及

将所述交换机移动(723)到所述交换机被使能所在的所述排序的延迟列表的末端。

4.如权利要求1所述的方法,其中对于所述多个服务中的每个服务,分组丢失和延迟被记分。

5.如权利要求1所述的方法,其中对于跨给定服务的分组延迟或丢失的每个测量,分组丢失和延迟被记分。

6.一种用来监视在线服务链中的分组处理的性能的计算装置,所述计算装置与形成软件定义网络SDN和所述在线服务链的多个网络装置通信,所述SDN包含由所述计算装置实现的用来配置所述多个网络装置的控制器,所述多个网络装置包含对遍历包含至少一个服务的所述在线服务链的分组进行监视的一组交换机,所述计算装置包括:

非暂态机器可读媒体(1048),用来存储监视模块(1081);以及

处理器(1042),通信耦合到所述非暂态机器可读媒体,所述处理器配置成执行所述监视模块,所述监视模块配置成:检查分组在遍历所述至少一个服务之后是否丢失了;添加交换机分组丢失记分,其中所述分组丢失了;添加交换机分组延迟记分,其中所述分组未丢失;根据对应交换机分组丢失记分,对所述一组交换机的列表进行排序,以生成排序的丢失列表; 根据对应交换机延迟记分,对所述一组交换机的列表进行排序,以生成排序的延迟列表;以及根据所述排序的丢失列表和所述排序的延迟列表中的次序,对所述一组交换机的列表进行排序。

7.如权利要求6所述的计算装置,所述监视模块进一步配置成:检查是否所述排序的丢失列表中的交换机被使能诱发延迟或者停止数据流;以及将所述交换机移动到所述交换机被使能所在的所述排序的丢失列表的末端。

8.如权利要求6所述的计算装置,所述监视模块进一步配置成:检查是否所述排序的延

迟列表中的交换机被使能诱发对于数据流的延迟;以及将所述交换机移动到所述交换机被使能所在的所述排序的延迟列表的末端。

9.如权利要求6所述的计算装置,其中对于所述多个服务中的每个服务,分组丢失和延迟被记分。

10.如权利要求6所述的计算装置,其中对于跨给定服务的分组延迟或丢失的每个测量,分组丢失和延迟被记分。

11.一种实现多个虚拟机以用于实现网络功能虚拟化NFV的计算装置,其中来自所述多个虚拟机的虚拟机配置成监视在线服务链中的分组处理的性能,所述计算装置与形成软件定义网络SDN和所述在线服务链的多个网络装置通信,所述SDN包含由所述计算装置实现的用来配置所述多个网络装置的控制器,所述多个网络装置包含对遍历包含至少一个服务的所述在线服务链的分组进行监视的一组交换机,所述计算装置包括:

非暂态机器可读媒体(1048),用来存储监视模块(1081);以及

处理器(1042),通信耦合到所述非暂态机器可读媒体,所述处理器配置成执行所述虚拟机,所述虚拟机用来实现所述监视模块,所述监视模块配置成:检查分组在遍历所述至少一个服务之后是否丢失了;添加交换机分组丢失记分,其中所述分组丢失了;添加交换机分组延迟记分,其中所述分组未丢失;根据对应交换机分组丢失记分,对所述一组交换机的列表进行排序,以生成排序的丢失列表; 根据对应交换机延迟记分,对所述一组交换机的列表进行排序,以生成排序的延迟列表;以及根据所述排序的丢失列表和所述排序的延迟列表中的次序,对所述一组交换机的列表进行排序。

12.如权利要求11所述的计算装置,所述监视模块进一步配置成:检查是否所述排序的丢失列表中的交换机被使能诱发延迟或者停止数据流;以及将所述交换机移动到所述交换机被使能所在的所述排序的丢失列表的末端。

13.如权利要求11所述的计算装置,所述监视模块进一步配置成:检查是否检查是否所述排序的延迟列表中的交换机被使能诱发对于数据流的延迟;以及将所述交换机移动到所述交换机被使能所在的所述排序的延迟列表的末端。

14.如权利要求11所述的计算装置,其中对于所述多个服务中的每个服务,分组丢失和延迟被记分。

15.如权利要求11所述的计算装置,其中对于跨给定服务的分组延迟或丢失的每个测量,分组丢失和延迟被记分。

用于在线服务链接的被动性能测量

[0001] 对相关申请的交叉引用

[0002] 对由Ying Zhang等人在2014年11月20日提交的“PASSIVE PERFORMANCE MEASUREMENT FOR INLINE SERVICE CHAINING”美国专利申请No.14/549,363的共同未决以及共同拥有的专利申请进行交叉参考。交叉引用的申请通过引用结合于本文中。

技术领域

[0003] 本发明的实施例涉及在线服务链接性能监视的领域。具体地，实施例涉及用于监视软件定义网络(SDN)中在线服务链的性能的方法和系统。

背景技术

[0004] 网络运营商使用它们网络中的不同计算装置(称作为中间盒(middlebox))提供与数据业务和订户管理相关的各种服务。这些服务被称为在线服务。这些服务的示例包含深度分组检测(DPI)、记录/计量/计费/高级计费、防火墙、入侵检测与预防(IDP)、网络地址翻译(NAT)以及管理来自网络运营商订户的数据业务的类似服务。这些服务对吞吐量和分组检测能力具有高的要求。服务对最终用户能够是透明的或者不透明的。在线服务能够托管在专用物理硬件的中间盒中或虚拟机中。

[0005] 服务链接是确立处理数据流的服务序列的过程。如果数据业务需要经历多于一个在线服务，则要求服务链接。而且，如果多于一个的服务链是可能的，则网络运营商需要将连网基础设施配置成通过正确的在线服务链的路径引导数据业务。在本文中使用时，数据业务导引指的是通过正确的在线服务路径指引数据业务。

[0006] 存在若干已被开发成管理如何导引数据业务以提供在线服务链接的机制。这些机制被设计成在给定数据流的端点之间的路径上明示地插入在线服务，或者根据与那个数据流关联的策略通过不同中间盒明示地路由业务。然而，不管使用什么方案导引网络中的业务，都存在如何测试任何给定服务路径的性能的问题。例如，理解遍历服务A、B和C的集合的数据流的延迟和丢失率是合乎期望的。这被称为对于在线服务链接的性能测量。

[0007] 尽管有许多操作经营和管理(OAM)工具来测量通用设置中的可达性，但在线服务链接强加了新挑战。关键挑战是，这些OAM方法主动将分组注入到网络，以测试网络路径的良好性。如果分组被主动注入到服务路径，则分组将被转发到中间盒。中间盒可能不知道如何处置这些注入的分组，并且从而，中间盒可丢弃未知分组。或者，探测分组可混淆中间盒的内部状态。

发明内容

[0008] 一种由计算装置实现的方法，用来监视在线服务链中的分组处理的性能。所述计算装置与形成软件定义网络(SDN)和所述在线服务链的多个网络装置通信。SDN包含由所述计算装置实现的用来配置所述多个网络装置的控制器。所述多个装置包含对遍历包含至少一个服务的所述在线服务链的分组进行监视的交换机的集合。所述方法包含：检查分组在

遍历所述至少一个服务之后是否丢失;添加交换机分组丢失记分,其中所述分组丢失;添加交换机分组延迟记分,其中所述分组未丢失。所述方法进一步包含:根据对应交换机分组丢失记分,对交换机的所述集合的列表进行排序,以生成排序的丢失列表; 根据对应交换机延迟记分,对交换机的所述集合的列表进行排序,以生成排序的延迟列表;以及根据所述排序的丢失列表和所述排序的延迟列表中的次序,对交换机的所述集合的列表进行排序。

[0009] 计算装置监视在线服务链中分组处理的性能。所述计算装置与形成软件定义网络(SDN)和所述在线服务链的多个网络装置通信。所述SDN包含由所述计算装置实现的用来配置所述多个网络装置的控制器。所述多个装置包含对遍历包含至少一个服务的所述在线服务链的分组进行监视的交换机的集合。所述计算装置包括用来存储监视模块的非暂态机器可读媒体以及通信耦合到所述非暂态机器可读媒体的处理器。所述处理器配置成执行所述监视模块。所述监视模块配置成:检查分组在遍历所述至少一个服务之后是否丢失;添加交换机分组丢失记分,其中所述分组丢失;以及添加交换机分组延迟记分,其中所述分组未丢失。所述监视模块进一步配置成:根据对应交换机分组丢失记分,对交换机的所述集合的列表进行排序,以生成排序的丢失列表; 根据对应交换机延迟记分,对交换机的所述集合的列表进行排序,以生成排序的延迟列表;以及根据所述排序的丢失列表和所述排序的延迟列表中的次序,对交换机的所述集合的列表进行排序。

[0010] 一种实现多个虚拟机以用于实现网络功能虚拟化(NFV)的计算装置,其中来自所述多个虚拟机的虚拟机配置成监视在线服务链中分组处理的性能。所述计算装置与形成软件定义网络(SDN)和所述在线服务链的多个网络装置通信。所述SDN包含由所述计算装置实现的用来配置所述多个网络装置的控制器。所述多个装置包含对遍历包含至少一个服务的所述在线服务链的分组进行监视的交换机的集合。所述计算装置包含用来存储监视模块的非暂态机器可读媒体以及通信耦合到所述非暂态机器可读媒体的处理器。所述处理器配置成执行所述虚拟机。所述虚拟机配置成实现所述监视模块。所述监视模块配置成:检查分组在遍历所述至少一个服务之后是否丢失;添加交换机分组丢失记分,其中所述分组丢失;以及添加交换机分组延迟记分,其中所述分组未丢失。所述监视模块进一步配置成:根据对应交换机分组丢失记分,对交换机的所述集合的列表进行排序,以生成排序的丢失列表; 根据对应交换机延迟记分,对交换机的所述集合的列表进行排序,以生成排序的延迟列表;以及根据所述排序的丢失列表和所述排序的延迟列表中的次序,对交换机的所述集合的列表进行排序。

附图说明

[0011] 通过参考用于说明本发明实施例的以下描述和附图可最好地理解本发明。在图中:

[0012] 图1是标准网络配置中的服务链的一个实施例的图解。

[0013] 图2是软件定义网络中的服务链的一个实施例的图解。

[0014] 图3A是用于测量在交换机的延迟和丢失的过程的一个实施例的流程图。

[0015] 图3B是用于测量在控制器的延迟和丢失的过程的一个实施例的流程图。

[0016] 图4是软件定义网络中的测量过程的示例的图解。

[0017] 图5A是用于在交换机进行延迟和丢失的聚合测量的过程的一个实施例的流程图。

[0018] 图5B是用于在控制器进行延迟和丢失的聚合测量的过程的一个实施例的流程图。

[0019] 图6A是用于在控制器生成模板的过程的一个实施例的流程图。

[0020] 图6B是用于在交换机根据配置文件测量延迟和丢失的过程的一个实施例的流程图。

[0021] 图7是用于诊断网络延迟和丢失问题的过程的一个实施例的流程图。

[0022] 图8是实现软件定义网络中的交换机并执行本文上面定义的交换机的任何功能的网络装置的一个实施例的图解。

[0023] 图9A示出了根据本发明一些实施例的示范网络内网络装置(ND)之间的连接性以及ND的三个示范实现。

[0024] 图9B示出了根据本发明的一些实施例用来实现专用网络装置902的示范方式。

[0025] 图9C示出了根据本发明的一些实施例在其中可耦合虚拟网络元件(VNE)的各种示范方式。

[0026] 图9D示出了根据本发明的一些实施例，在图9A的每一个ND上具有单个网络元件(NE)的网络，并且在这个直接的途径内，将传统分布式途径(通常由传统路由器使用)与用于维持可达性和转发信息的集中式途径(也称为网络控制)进行了对比。

[0027] 图9E示出了根据本发明的一些实施例的简单情况，其中每一个ND 900A-H实现单个NE 970A-H(见图9D)，但集中式控制平面976已将不同ND中的多个NE(NE 970A-C和G-H)抽象成(用来表示)图9D的虚拟网络992之一中的单个NE 970I。

[0028] 图9F示出了根据本发明的一些实施例的情况：其中多个VNE(VNE 970A.1和VNE 970H.1)被实现在不同ND(ND 900A和ND 900H)上，并且彼此耦合，并且其中集中式控制平面976已经抽象了这些多个VNE，使得它们看起来好像图9D的虚拟网络992之一内的单个VNE 970T。

[0029] 图10示出了根据本发明的一些实施例包含硬件1040的通用控制平面装置1004，硬件1040包括以下装置的集合：一个或多个处理器1042(其经常是商业化构件(COTS)处理器)和网络接口控制器1044(NIC;也称为网络接口卡)(其包含物理NI 1046)以及其中已存储有集中式控制平面(CCP)软件1050的非暂态机器可读存储媒体1048。

具体实施方式

[0030] 如下描述描述了用于测量包含延迟和丢失的在线服务链中的性能的方法和设备，其中服务链在软件定义网络(SDN)中。在如下描述中，阐述了众多特定细节，诸如逻辑实现、操作码、规定操作数的部件、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，以便提供本发明的更透彻理解。然而，本领域技术人员将领会到，没有此类特定细节也可实践本发明。在其它实例中，控制结构、门级电路以及全软件指令序列尚未详细示出，以免使本发明模糊不清。本领域那些普通技术人员采用所包含的描述将能够实现适当功能性，而无需过度实验。

[0031] 在说明书中对“一个实施例”、“一实施例”、“一示例实施例”等的引用指示所描述的实施例可包含具体特征、结构或特性，但每一个实施例可不一定包含该具体特征、结构或特性。而且，此类短语不一定是指相同实施例。另外，当联系一实施例描述具体特征、结构或特性时，认为它在本领域技术人员的知识范围内，以联系其它实施例(不管是否明确被描

述)来影响此类特征、结构或特性。

[0032] 带括号的文本以及具有虚线边界(例如大虚线、小虚线、点虚线、和点)的框在本文可用于说明向本发明实施例添加附加特征的可选操作。然而,此类记号不应该被采纳为意味着这些是唯一选项或可选的操作,和/或具有实线边界的框在本发明的某些实施例中不是可选的。

[0033] 在以下说明书和权利要求中,可使用术语"耦合"和"连接",连同它们的派生词。应该理解,这些术语不旨在作为彼此的同义词。"耦合"用于指示两个或更多元件与彼此协同操作或交互作用,它们可以或者可以不与彼此直接物理接触或电接触。"连接"用于指示在与彼此耦合的两个或更多元件之间确立通信。

[0034] 概述

[0035] 在本发明的实施例中,通过在相关交换机的转发表上建立规则来利用被动测量途径,交换机收集分组的纲要(digest)以及捕获它们所在的时间戳。结果的概要被发送到控制器。通过将从网络的不同交换机接收的分组进行相关,控制器能够计算给定数据流的延迟和丢失(特别是在这些数据流遍历在线服务链中的服务时)。然而,服务可在飞行中(on the fly)修改分组。从而,实施例提供了根据服务模型来标识跨多个服务的数据流中的分组的不变位(bit)的方法。实施例基于不变位构造散列函数。最后,为了将性能问题与服务的预期行为进行区分,实施例提供了基于服务的模型规范和测量结果标识网络中问题的根本原因的过程。

[0036] 在线服务链接

[0037] 如上面所阐述的,网络运营商需要用来执行灵活业务导引的过程和工具。如果数据业务需要经历多于一个在线服务,则要求服务链接。而且,如果多于一个的服务链是可能的,则网络运营商需要能够将连网基础设施配置成通过正确的在线服务路径来引导正确业务。

[0038] 对于任何解决方案的要求是效率、灵活性、可缩放性和开放性。关于效率,数据业务应该遍历由网络运营商规定的序列中的中间盒,并且不应该不必要地遍历中间盒。如果数据业务能够有选择地通过特定服务被导引,或者被导引远离(旁路)特定服务,则能够取得大的资本支出节省。关于灵活性,任何解决方案的框架应该同时支持订户、应用和运营商特定策略,所有都源于单个控制点。添加或移除新服务应该由网络运营商容易地完成。关于可缩放性,框架应该支持大量规则,并且随着订户/应用数量的增长而缩放。供予(offer)对在线服务的按订户选择的能力能够潜在地导致新供予的创建,以及因此是运营商用来使它们的网络货币化的新方式。关于开放性,该过程应该有可能部署来测量跨网络中的任何类型中间盒的性能,独立于其供应商,以便避免供应商锁住。进一步说,网络运营商应该能够通过在不进行修改的情况下再用它们现有的中间盒来衡平它们的当前投资。

[0039] 当前发明的实施例满足了这些要求(如本文下面所阐述的)。一般而言,网络运营商使用基于策略的路由(PBR)朝正确的服务转发订户业务。网络运营商还可使用访问控制列表(ACL)和虚拟局域网(VLAN)(或其它遂穿技术)向正确的服务和中间盒转发数据分组。

[0040] 在一些情况中,服务链接能够部分由服务本身执行,留给网络运营商对服务路径中的其余跳跃(hop)较少的控制。在此情况中,如果服务中间盒未直接连接到下一跳跃,则服务必须配置成将业务引导到在线服务链中的下一跳跃。

[0041] 图1是描绘实施例所解决的现有技术问题的图解。在此示例中，假定住宅业务将需要DPI和NAT。在此示例中，高端住宅数据业务除了防火墙和统一资源定位符(URL)过滤(未示出的URL过滤)，还将得到与基本住宅数据业务相同的服务。并且最后，企业业务将不会要求NAT，但将需要防火墙和病毒扫描。在此示例中，所有数据业务都经历DPI，并返回到边界网络网关(BNG)，图中的点(1)。从那儿(图中的点2)，BNG必须将业务引导到正确的下一跳跃服务。订户会话由能够定义服务链中第一跳跃服务的认证、授权和记账(AAA)驱动的策略处置；然而，这个订户上下文信息不再与在点(1)来自DPI的返回业务相关联。因此，确定对于特定流的下一服务变得非平凡(non-trivial)。

[0042] 另外，在现有技术中，存在用于处置服务链接的若干其它方法或系统。一个方法是使用单个盒运行多个服务：此途径将所有在线服务合并到单个盒中，并且因此避免了对于跨多个中间盒处理在线服务链接配置的需要。在此途径中，网络运营商通过向其路由器或网关添加附加服务卡来添加新服务。

[0043] 然而，此途径不能满足开放性要求，因为它难以集成现有第三方服务设施。这个解决方案还受制于可缩放性问题，因为服务数量和聚合带宽受路由器的容量限制。机箱中插槽的数量同样受限制。

[0044] 现有技术中的另一途径是使用静态配置的服务链。这种途径配置一个或多个静态服务链，其中每个服务配置成将业务发送到其链中的下一个服务。路由器对进来的数据业务进行归类，并基于归类结果将它转发到在每个链头端的服务。然而，这种途径不支持以集中式方式的策略定义，且反而要求每个服务配置成对业务进行归类，并将业务导引到适当的下一个服务。这种途径要求大量的服务特定配置，并且容易出错。它缺乏灵活性，因为它不支持基于每个订户的业务导引，并且限制能够配置的不同服务链。避开这些限制将要求每个服务上的附加配置以归类和导引业务。

[0045] 另外的途径是基于策略的路由，关于这个途径，每个服务必须配置成在处理数据业务之后将它返回到路由器。路由器在每个服务跳跃之后对业务进行归类，并基于分类结果将它转发到适当服务。然而，此方法受制于缩放性问题，因为在每个服务之后业务被强制通过路由器。路由器必须能够处置N次进来的数据业务线率，以支持具有N-1个服务的服务链。

[0046] 知晓策略的交换层是其中集中通过不同的中间盒序列明示地转发业务的对于数据的知晓策略交换层的途径。此方法满足了效率要求，但无法满足灵活性和可缩放性的要求。每个策略都需要被翻译成所有有关交换机上的低级别转发规则的集合。没有明示的方法来分开地配置应用相关和订户相关的规则。它们需要被人工合并成低级别规则的集合。而且，它要求对于每个新流建立一个规则。因此，难以随订户/应用组合的数量进行缩放。

[0047] 基于SDN的在线服务链接

[0048] 软件定义的连网(SDN)是最近的网络架构，其中控制平面从转发平面(即数据平面)解耦，并且整个路由器被构建为分布式系统。SDN含有网络范围的控制平台，运行在网络中的一个或多个服务器上，监督简单交换机的集合。传统路由器架构遵循集成设计，其中控制平面和数据转发引擎被紧密地耦合在相同盒中，这通常导致过于复杂的控制平面和复杂的网络管理。由于高度复杂性，设备供应商和网络运营商不愿意采用改变，并且网络本身是脆弱的并且难以管理。对新协议和技术开发创建了大负担和高的障碍，这是已知的。

[0049] SDN网络包括多个转发元件,即,操作为彼此互连的交换机的网路装置以及实现指令交换机的转发行为的控制器的小数量的计算装置。

[0050] 转发元件或交换机的主要任务是要根据由远程控制器实现的流表中的规则,将分组从入口端口转发到出口端口。流表含有流条目的集合。每个流条目都含有动作的集合,诸如将分组转发到给定端口,修改分组报头中的某些位,或将分组封装到控制器,或者简单地丢弃分组。对于新数据流中的第一分组,交换机通常将分组转发到控制器以触发对新流条目进行编程。它还能够用于将所有慢路径分组转发到控制器,以便处理诸如互联网控制消息协议(ICMP)分组。流的概念能够广义地定义,例如传送控制协议(TCP)连接,或者来自特定媒体访问控制(MAC)地址或互联网协议(IP)地址的所有业务。

[0051] 集中式SDN控制器添加和移除来自SDN的转发或数据平面中的交换机的流表的流条目。控制器定义数据平面交换机的集合之间的互连和路由。它还处置网络状态分布,诸如从交换机收集信息,以及向它们分布路由指令。控制器还能够编程为支持任何新寻址、路由和复杂分组处理应用。控制器是网络的“大脑”。交换机需要连接到至少一个控制器以正确运作。简单网络拓扑由两个控制器和交换机的集合组成。

[0052] 图2中提供了SDN的示例。在此示例中,当交换机S4接收到新流而不知道将分组发送到哪里时,向控制器转发新接收的数据流的第一接收的分组。在接收到分组时,控制器对新路由条目编程。

[0053] 基于SDN的在线服务链接

[0054] 图2还提供了使用SDN的在线服务链接的示例架构。此示例使用逻辑上集中化的控制器来管理交换机和中间盒。图2中的实线和虚线示出遍历SDN的两个不同服务路径。在此示例中,基于订户、应用和要求的服务次序来设置服务路径。路径是单向的,也就是,对于上游和下游业务规定了不同服务路径。此图中的实线示出了通过病毒扫描、DPI和内容高速缓存的上游业务的服务路径。虚线示出了为SDN所有服务加旁路的服务路径。

[0055] 这个示例架构使用两种不同类型的交换机。周边交换机被放置在服务递送网络的周边上。这些交换机将对进来的业务进行归类,并将其导引向链中的下一个服务。这些是服务或网关节点连接到的交换机。内交换机将通过网络转发业务。这些交换机仅连接到其它交换机。这些交换机可受中央控制器的控制,或可不受其控制。

[0056] 业务导引是两步过程。第一步,对进来的分组归类,并且基于预先定义的订户、应用和排序策略给它们指配服务路径。第二步,沿其所指配的服务路径基于其当前位置向下一服务转发分组。这个两步业务导引过程在任何两个边界路由器(即周边交换机)之间仅需要执行一次,不管连接它们的内交换机的数量如何。

[0057] 服务链接OAM

[0058] 不管使用什么机制来实现服务链接,一个重要的问题是如何验证已正确地建立了路径。目标是要证明给定流的分组已经遍历了预期路径。现有可达性测量过程包含ping 和跟踪路由用来测量从源到目的地的可达性。Ping触发ICMP应答,而跟踪路由触发沿路径的路由器上的ICMP使用期限(TTL)期满消息。两种方法都不要求两端控制。这些函数(ping和跟踪路由)已经被实现了,或者在不同协议层,例如多协议标签交换(MPLS) ping是可用的。

[0059] 然而,如早先所开始的,传统ping/跟踪路由方法不适合于在线服务设置。在传统网络中,ping/跟踪路由分组的丢失指示路径问题。然而,在示例上下文中,ping/跟踪路由

分组可被路径中间中的服务(中间盒)识别,并且从而可被丢弃。类似地,服务可向分组引入附加延迟。从而,我们不能简单地说,丢失的测量分组的症状是由于路径性能问题引起的。因此,我们需要不同的方法来测量用于在线服务链接的路径性能。

[0060] 测量

[0061] 在本发明的实施例中,提供了一种测量在线服务链的丢失和延迟的新方法。代替主动将分组注入到网络,此方法记录了当在每个交换机处看到分组时的时间戳,并且然后将这个信息的紧凑表示转发到控制器。通过将由不同交换机捕获的分组进行相关,控制器能够计算时间戳的差以计算延迟,并使用分组数之差来表示丢失。这个过程涉及一系列步骤。

[0062] 第一,控制器知道每个数据流需要遍历的中间盒的序列。控制器具有服务和网络的拓扑、以及构造的服务链以用于每个流。控制器接收测量用于特定数据流的路径的请求,然后它在被那个数据流遍历的所有有关交换机上建立规则。所述规则在输入端口和流的分组报头的字段上匹配,创建拷贝,并将其发送到控制器。流的字段能够标识数据流的分组,并且输入端口指示分组已遍历了哪个服务。

[0063] 第二,对于每个数据流,该过程对几个时间窗口进行采样。在每个时间窗口中,将要求在路径两端的每一个交换机都记录每个分组到达的时间戳,并在每个时间窗口中维持分组总数的计数器。我们能够通过计算分组计数器的差来计算每个窗口的丢失率,并且通过对记录在不同交换机的相同分组的时间戳的差求和来计算平均延迟。

[0064] 第三,该过程解决中间盒可对分组进行修改使得分组报头上的散列可不总是捕获相同分组的问题。为了解决这个挑战,该过程分析公共中间盒可对分组做的修改的类型。该过程确定基于模型的方法,用以标识不变位并使用它们作为构造散列的密钥。

[0065] 最后,一些中间盒可故意延迟或丢弃分组。这将引入对结果的解释的混淆。该过程将这个结合到中间盒的模型中,并且然后使用此信息帮助指引对测量结果的最可能原因的搜索。

[0066] 该流程图中的操作将参考其它图的示范实施例被描述。然而,应该理解,流程图的操作能够由除了参考其它图论述的那些以外的本发明的实施例执行,并且参考这些其它图论述的本发明的实施例能够执行与参考流程图论述的那些操作所不同的操作。

[0067] SDN中的基本延迟和丢失测量

[0068] 实施例聚焦在被单个集中式控制器控制的由许多交换机组成的网络上。集中式控制器从所有交换机搜集定时信息,并且—实时—演算对于给定路径段的任何交换机对之间的延迟。实施例测量由“实际”数据分组所阅历的延迟——而不是将一些分组注入到网络中,以及测量由这些分组所阅历的延迟。

[0069] 主要思想是要记录在任何交换机对多个分组的到达时间。注意,对于分组在遍历服务之前和之后其也能够在相同交换机上。如果两个交换机上的定时器是同步的,则路径延迟能够被演算为任何交换机对之间的平均时间差。然而,挑战在于,在入口交换机和出口交换机二者记录分组的“相同”集合的到达时间。为了解决这个问题,实施例(1)在测量交换机的两端都记录属于期望流的分组序列(例如200个分组)的到达时间,(2)向这些选择的分组应用散列函数,以在要进行测量所在的任何交换机将完整分组编码成固定数量位。将散列值与时间戳一起存储在时间戳表(TST)中,或备选地使用散列值作为表中的索引来存储

时间戳，(3)将此信息发送到集中式控制器，(4)在控制器将分组时间戳与相同散列值进行比较。路径延迟是所述两个时间戳之间的差。最后，(6)该过程重置时间戳表。

[0070] 当执行延迟演算时，在许多情况中，在第一交换机与第二交换机的选择的分组之间存在某些(但不是完全)交叠。控制器仅基于所述两个交换机的时间戳表条目的公共子集来演算平均路径延迟。当执行丢失演算时，通过计算在第一交换机和第二交换机的分组的差，该过程将获得在这些两个端点之间丢失的分组的总数。

[0071] 图3A是在交换机实现的延迟和丢失测量过程的一个实施例的流程图。这个过程假定，控制器已经将交换机配置成监视具体数据流。如本文下面所进一步论述的，控制器能够将流表配置成通过检查数据流分组的不变字段来标识数据流的分组。在一个实施例中，该过程响应于被测量或监视的数据流的数据分组的接收(框301)。该过程能够在逐个分组的基础上被实行，其中对于数据流接收的每个分组如本文下面所阐述的被处理，或者其中所有分组的子集在它们被接收时被处理。

[0072] 然后能够将散列函数应用于数据分组的序列、或任何组合或序列的子集的每个分组(框303)。散列函数能够操作在数据分组的任何部分上，诸如报头中特别标识的字段，或者包含对于具体数据流已被标识为不变的那些位的位的任何组合。能够利用所得到的散列值来存储对于分组的编组或序列的每个分组的时间戳(框305)。时间戳能够是在交换机接收到分组时的该分组的时间戳，或者对于由交换机一贯捕获的分组到达或处理时间的任何类似指示符。接收的数据流分组的时间戳能够是具有任何大小或格式的整数值或类似数据类型。在一个实施例中，时间戳表由散列值索引，并且接收的数据分组的每个时间戳根据散列值被记录在时间戳表中。在其它实施例中，能够利用其它类型的存储结构，并且散列值能够被用作为密钥，或者类似地与关联的时间戳一起被存储。

[0073] 采用时间戳来标识与记录的数据流的每一个数据分组关联的到达时间或类似时间，该过程准备散列值的集合以及关联的时间戳集合并将它们发送到控制器，以使控制器能够在数据与交换机另一侧上的其它交换机的数据相比较时，确定跨服务的延迟和丢失(框307)。在这个数据已经成功被传递并且由控制器使用任何通信协议接收之后，控制器然后能够用重置时间戳表的命令和确认进行响应(框309)。重置时间戳表清除被传送的数据以释放存储空间来用于正被监视的数据流的连续测量(框311)。当对于数据流接收到附加数据分组时，该过程能够连续操作。交换机能够执行任何数量的类似过程，并维持任何数量的时间戳表以监视和测量任何数量的数据流的性能。

[0074] 图3B是由控制器实现的测量过程的一个示例实施例的流程图。控制器从SDN中的每一个配置的交换机接收散列值和时间戳的集合(框351)。能够在交换机之间比较这些散列值和对应时间戳，以确定分组的延迟和丢失。例如，能够从来自服务的第一交换机下游接收散列值和时间戳的第一集合，而从来自服务的第二交换机上游接收散列值和时间戳的第二集合。比较来自第一交换机和第二交换机的数据能够实现检测与所述两个交换机之间的这个服务相关的分组延迟以及丢失的分组(框353)。

[0075] 该比较检查是否发现从两个交换机接收的是匹配的散列值。在发现匹配的散列值的情况中，则通过比较时间戳能够确定延迟时间。时间戳中的差提供了对于遍历所述两个交换机之间的中间盒或服务的延迟时间。

[0076] 类似地，在其中未发现匹配的散列时，则能够通过标识在第一交换机出现的未发

现被第二交换机报告的那些散列值来确定分组丢失(框355)。在一些情况中,分组被中间盒或服务修改或丢弃,并且这种修改如本文下面所描述的被预见。在演算分组延迟和丢失之后,然后能够将重置命令发送到报告处理的散列值和时间戳的交换机。这使交换机能够释放空间以用于收集随后由交换机在相同数据流中接收的分组的附加时间戳。重置命令清除用于已发送或被控制器处理的那些散列值的时间戳表。

[0077] 图1是在网络入口交换机和出口交换机基于时间戳表的公共条目处理的延迟和丢失测量的一个示例实施例的图。基本方法对于测量对给定时间间隔的小数量流工作良好。在期望处置更高数量的连续信息的情况中,该方法能够被调整成执行在交换机的某些级别的聚合(如本文下面进一步描述的)。

[0078] 代替对于每个分组保持一个时间戳,交换机能够执行对于属于相同流的所有分组的聚合,并对分组的集合维持单个时间戳。在一个实施例中,该方法通过将散列密钥构造为对于相同流的所有分组的公共字段来保持相同流的所有分组的合计。在此情况中,该过程将对于每个流在每个交换机上产生一个条目。在示出的示例中,该过程在图1中第一交换机上保持S1=T1+T2+T3+T4,并且在第二交换机上保持S2=T1'+T2'+T3'+T4',其中S1和S2是聚合的时间戳值,且T1-T4`是单独的分组时间戳值。假定没有分组丢失,因为存在4个条目,则该过程还能够按照(S2-S1)/4来计算平均延迟。此方法减少了在每个交换机上要求的资源量以及在交换机与控制器之间互换的信息量。

[0079] 然而,对基本聚合过程存在两个改进要被做出。第一,如果存在分组丢失怎么办?则S2和S1不再是可比较的。从而,修改的过程需要引入另一计数器,其保持跟踪对于每次合计的分组的数量。在示出的示例中,该过程还能够维持用于S1的计数器,其是C1=4,指示存在4个分组贡献给S1。类似地,该过程还能够在S2上维持另一计数器C2。因此,该过程能够通过简单地比较C2和C1来检测分组丢失。在此实施例中,如果C1和C2是可比较的,则该过程仅使用S1和S2来计算延迟。

[0080] 聚合过程存在的第二个问题是,作为聚合的结果,如果存在任何分组丢失,则合计变得不可用,这在大窗口上能够是相当常见的。从而,该过程能够利用在小窗口大小(例如200ms或每10个分组)上的合计。这样,该过程降低了在每个合计中具有丢失的分组的可能性。它还提供了更细粒度的丢失信息,而不是对于整个流持续期只具有一个总丢失数。

[0081] 第三个问题是,在一些情况中,知道平均延迟值有时是不足够的。从而,除了合计之外,该过程还能够使用散列表(即时间戳表)中的另两个条目来保持最大和最小时间戳。这进一步提供了关于延迟变化的范围的数据。

[0082] 甚至采用交换机上的聚合,仍能够存在有关服务链接环境的严重问题,也就是,服务可在飞行时修改分组。从而,如果过程使用相同散列函数在服务之前和之后匹配分组,则它可能不总是保证该过程能捕获相同分组。对于这些修改进行调整的过程本文在下面相对于图6A和6B进行论述。

[0083] 图5A和5B是用于测量在线服务链中的分组的延迟和丢失的聚合分组处理的一个实施例的示例流程图。在一个实施例中,该过程通过在交换机接收数据分组的序列(框501)在由控制器对交换机进行初始配置之后开始。该序列能够被设置成任何长度,使得计数器跟踪对于具体数据流被接收的分组的集合,并且当它超过指定阈值时,单独或以聚合方式向分组应用散列函数(框503)。利用为该编组生成的散列值,将时间戳的集合存储在时间戳

表中，散列值是表中的索引以标识存储位置，或者使用类似机制(框505)。

[0084] 在预定间隔，或者当分组的每个编组被散列时，将还尚未被控制器报告和/或确认的散列值和时间戳发送到控制器(框507)。该过程能够继续接收和收集分组的聚合编组，并将在从控制器接收到命令时重置时间戳表(框509)。命令的接收指示，分组已被控制器接收和处理，并且交换机能够然后清除并收回时间戳表中被发送的时间戳的空间(框511)。

[0085] 类似地，在图5B中，示出了其在被控制器实行时的过程。控制器从网络中的多个交换机接收聚合的时间戳数据。该过程能够应用于从其接收数据以确定在交换机之间的并且因此在中间的中间盒或服务上的延迟和分组丢失的任何两个交换机。

[0086] 该过程开始于接收对于分组的编组的聚合的散列值和时间戳。这个数据能够从网络中的任何数量的交换机接收，然而，为了方便和清楚起见，示例将假定，从毗连中间盒或服务的入口和出口的两个交换机接收数据。在示例示出的实施例中，具体地从第一交换机和第二交换机接收数据。来自第一交换机的数据然后能够与来自第二交换机的数据相比较。该比较能够注意标识从第一交换机和第二交换机接收的相同散列值，其中在匹配出现时，时间戳或聚合的时间戳合计能够相比较，以确定所述两个交换机之间的延迟时间(框553)。

[0087] 类似地，能够标识分组丢失，其中在第一交换机标识未被第二交换机报告的散列值(框555)。能够存在序列或时间延迟，其用于证实散列值未被接收且在第二交换机刚好未延迟。本文下面进一步论述对于由交换机对分组的可能修改的调整。一旦已经对于散列值的集合演算了分组延迟和丢失，则能够向第一交换机和第二交换机二者发送时间戳表重置命令，这使那些交换机能够释放它们时间戳表中的空间(框557)。

[0088] 用于分组的中间盒处理的建模

[0089] 为了恰当地考虑服务对数据分组的处理，该过程基于可能的修改以及它们对分组的影响的概括而被设计。然后，该过程依赖于基于模型的方法来智能地构造散列以处置数据路径上的此类动态(dynamics)。

[0090] 修改的类型对建模过程是重要的。中间盒修改分组报头，并且为了性能优化，甚至将一个会话映射到另一个。例如，网络地址翻译(NAT)将修改分组的IP报头中的地址字段以翻译网络地址。广域网(WAN)优化器和代理可维持与远程服务器的永久连接，并将若干流聚合成一个大会话。此类修改根本上预防了在服务的两端使用标准散列函数捕获相同分组。例如，在基本设计中，如果该过程基于分组报头的五元组构造散列的密钥，则分组将不会被捕获，或者将在NAT(其中源地址被改变)之后被映射到不同散列表元(bucket)。类似地，负载均衡器也可修改远程服务器的目的地地址。

[0091] 另一方面，一些服务可对分组本身引入性能改变。例如，防火墙可因为违反策略而故意丢弃流的所有分组。这将看起来是严重的分组丢失，因为在服务之后将没有分组被捕获。然而，这不是由于网络问题引起的，而是服务的预期行为。为了克服这个挑战，首先分析的是，什么修改是中间盒能在下表中对分组做出的。

[0092]

动作	示例	不变字段
完全丢弃分组	防火墙，高速缓存服务器(当存在命中时，请求将不被向前发送)	无
修改报头	NAT	不同于 src/dst 地址和 src 端口的字段，例如分组有效载荷
修改有效载荷	冗余消除器(它们消除了未广泛使用的分组上的冗余有效载荷)	分组报头
修改报头和有效载荷二者	HTTP 代理(它基于会话操作，它可修改 HTTP 报头字段)、WAN 优化器(它可将多个用户的请求映射到一个)	报头中的一些字段和有效载荷中的一些字段

[0093] 表1 中间盒修改的分析

[0094] 能够从这个归类中导出分组上的中间盒的影响的模型。根据以上分析，该过程提供了用于每个中间盒的模型。该模型应该包含如下字段：(1)类型：指示中间盒什么类型；

(2)流的终止:是或否:如果存在这个服务将终止流的机会;(3)流的重新映射:是或否,如果服务会将一个进来的流映射到另一个出去的流,包含流聚合;(4)丢弃分组:是或否,如果服务将丢弃分组的子集;(5)延迟分组:是或否,如果服务将向分组故意引入附加延迟,例如速率限制器;(6)修改的字段:它是以(a1,a2)、(a3,a4)…(am, an)形式规定的,其中am是修改字段的起始位,而an是修改字段的结束位。这规定这个服务可修改的所有字段。

[0095] 该过程设想,模型能够通过对中间盒的类型的基本理解来获得,或者由中间盒供应商提供。如果供应商能够提供此类信息,则该过程能够精确地构造不变位。如果由于各种原因,供应商不愿意提供此类信息,则该过程能够提供能够通过理解中间盒的类型而获得的一些粗粒度信息。例如,根据对几个典型开放源中间盒的基本理解,来标识表1中示出的修改类型。而且,随着此类理解的演进,基于模型的途径能够更精确地改进。

[0096] 模型一旦构造了就能够被提供给交换机或控制器来构造散列。散列函数指的是压缩函数,导致输出比输入短。经常,此类函数采取任意或几乎任意长度的输入,备选地,散列函数能够采取具有固定数量例如160位的长度的输入。在密码术的许多部分中使用散列函数,并且存在许多不同类型的散列函数,其具有不同的安全属性。

[0097] 该过程能够使用任何散列函数计算分组的纲要,其将分组映射到能够用于唯一标识分组的字符串。在散列函数的选择上存在几个要求:(1)它能够用硬件实现,因为分组的纲要需要以线速度创建,以避免向交换机引入附加负载;以及(2)它需要在相同流中传送的分组的正常输入下具有低散列冲突。

[0098] 本发明的实施例不聚焦在具体散列函数上。SHA1在本文用作示例。然而,本领域中技术人员将理解到,该过程能够与其它散列函数结合。SHA1是将几乎任意长度的字符串转换成160位的字符串的简单函数:M=SHA1(K,P),其中P是x位的分组,且M是对应纲要。x是可配置参数。

[0099] 在一些实施例中,如下方法构造P。假定,我们需要测量服务S1、S2和S3的服务路径的延迟和丢失率。根据这三个服务的模型,该过程能够搜集的是,S1将修改(x1,y1)位,S2将修改(x2, y2),以及S3将修改(x3, y3)。

[0100] 如果目标是要仅在这个路径的两端进行监视,意味着在S1之前捕获一次分组,并且在S3之后捕获一次分组,则该过程能够将位按如下构造:P=(1,k)-(x1,y1)-(x2,y2)-(x3,y3),假定k是这个流中的分组的最小长度。

[0101] 另一方面,P的选择应该足够长以确保相同流中的任两个分组都不一样。如果在该公式之后P太小,则我们不得不使用更多的监视点来分离服务,使得我们能有足够的位在每步进行监视。

[0102] 图6A和6B提供了建模和配置的示例实现。在一个实施例中,能够发起该过程以给网络中多个服务的每个服务创建模板。模板能够人工创建,或者通过自动识别服务特性以及上面所阐述的归类来创建。一旦已经为每个服务创建了每个模板,则该过程为要被配置的被选择服务生成配置文件(框603)。配置文件被构造成由控制器下载到交换机,以将流表或类似结构配置成对由模板标识的位的规定集合进行散列。模板与要配置的交换机的参数一起应用,以根据交换机的功能性(例如交换机实现OpenFlow或类似流控制协议)生成配置文件。

[0103] 能够执行检查以确定是否已经对于具体交换机配置了所有服务,使得已经对于每

个服务构造了配置文件(框605)。如果尚未处理所有服务来生成关联的配置文件,则该过程能够继续选择要处理的下一服务(框607)。如果所有服务已经被处理来生成配置文件,则该过程能够完成。在一些实施例中,在控制器如此生成的配置文件能够被下载到相应交换机,以使用关联的流控制协议实现在交换机的配置。

[0104] 图6B是在交换机执行的过程的一个实施例的流程图。该过程在交换机通过接收要测量的给定流的数据分组(框651)来触发。在第一次接收到数据流的分组时,对与该数据流关联的每个服务查找对应配置文件(框653)。该过程基于底层模板应用该配置来标识数据流的不变位(框655)。这些不变位被输入到散列函数以生成用于数据流的数据分组的输入序列的散列值的集合(框657)。用于完成延迟和丢失测量的过程(框659)然后如本文以上关于基本或聚合的延迟和丢失测量所论述的继续进行。

[0105] 网络范围根本原因分析

[0106] 假定,引起服务路径的性能衰退的主要问题是交换机和服务的过载,则服务链接OAM的首要目标是要连续监视整个服务网络以确保没有交换机或服务过载。如果服务或交换机过载,则它会将性能衰退引入遍历它的所有数据流。

[0107] 从而,在这部分被解决的问题则是,给定所有测量输入,如何最好地定位最可能引起问题的服务。除了过载问题,分组也能够被服务故意丢弃。从而,有可能使用本文上面所描述的模型来帮助进一步执行根本原因分析。该过程创建列出所有服务的表。每个服务含有指示根本原因可能性的得分。每个服务还具有三个标志:'isdelay'标志指示数据流是否能引起延迟;'isloss'标志指示数据流是否能引起丢失;并且'isstop'标志指示数据流是否将使该流完全停止。

[0108] 接下来,该过程经历来自给定时间窗口的所有流的所有测量。如果存在分组在这个服务之后看不见的指示,则该过程根据测量的丢失率增大服务的得分。对于延迟,我们把总延迟归因于均匀地沿该路径的服务。最后,该过程基于它们的得分、平均延迟以及它们的标志,根据如下规则给服务排名。

[0109] 该过程首先挑选具有最高丢失得分的服务,如果isloss=1且isstop=1,则把它从列表中移除,移动到具有下一最高得分的服务。该过程还通过首先查看延迟值来创建排名,并且如果isdelay=1,则我们把它从列表中移除,移动到具有下一最高得分的服务。最后,该过程组合来自延迟和丢失二者的排名。直觉是,如果服务或交换机被拥塞,则它将看起来具有高分组丢失和较长延迟二者。注意,这里S包含交换机和服务二者。

[0110] 详细算法的一个示例示出如下:

Algorithm 1 Root cause analysis algorithm

procedure *Root_Cause_Analysis(M,S)*

```
for every measurement m ∈ M do
  for every service s that m traverses do
    if packet is lost after traversing s then
      s.loss = s.loss + m.loss
      s.delay = s.delay + m.delay
    end if
  end for
end for
sort S according to s.loss, stored in S_loss
for each sorted S_loss do
  if s.isloss==TRUE OR s.isstop==TRUE then
    move s to the tail
  end if
end for
sort S according to s.delay, stored in S_delay
for each sorted S_delay do
  if s.isdelay==TRUE then
    move s to the tail
  end if
end for
sort S according to their positions in both S_delay and S_loss
```

[0111]

[0112] 根本原因分析算法

[0113] 图7是网络根本原因分析的一个实施例的流程图。在一个实施例中，该过程使用可用的收集的测量信息周期运行。通过选择下一测量来处理，该过程通过这些测量进行迭代(框701)。该过程然后选择与测量关联的下一服务，即，其中该测量遍历该服务(框703)。该过程检查在遍历选择的服务之后分组是否丢失(即，按照上面阐述的过程尚未发现对应的散列)(框705)。如果分组丢失了，则向交换机的分组丢失记分(tally)进行添加(框706)。如果分组未丢失，则向交换机延迟记分添加分组延迟(框707)。在任一情况中，然后检查是否已经处理了所有服务(框709)。如果尚未处理所有服务，则选择下一服务(框703)。如果已经服务于所有服务，则对是否已经处理了所有测量进行检查(框711)。如果尚未处理所有测量，则选择下一测量(框701)。

[0114] 一旦所有服务和测量都已被处理，则该过程开始基于它们关联的记分对它们进行排序，以标识引起最多分组丢失和延迟的那些。首先，交换机列表被创建，并且按交换机分组丢失记分进行排序，由此创建排序的丢失列表(框713)。

[0115] 在已经创建了排序的丢失列表之后，则能够进行检查是否使每个交换机都能够诱发延迟或使数据流停止(框715)。如本文上面所论述的，交换机的配置文件能够指示已经使其能够诱发延迟或使数据流停止。该列表被排序成使得诱发最大延迟的那些交换机被置于列表的头端。如果能使交换机诱发延迟，则交换机能够被移动到列表的末端，由此指示该交换机将不被定为原因，因为已明示地使其能够诱发延迟并停止数据流(即引起丢失)。

[0116] 根据每个交换机的分组延迟记分，生成交换机的另一列表并对其排序，这生成了排序的延迟列表(框719)。检查列表中的每一个交换机以确定是否能使交换机诱发延迟(框721)。如上面所论述的，这由对于该交换机生成的关联配置文件中的标志所指定。如果能使交换机诱发延迟，则交换机能够被移动到列表的末端，该列表被组织其中引起最多延迟的那些交换机被置于列表的头端(框723)。这指示，使能的交换机不被视为网络中的延迟源。

最后，所述两个列表(排序的丢失列表和排序的延迟列表)能够被组合和排序，以创建引起网络中丢失和延迟的最有问题的交换机的总体列表(框725)。用于组合的过程能够在优先顺序上或将丢失或将延迟加权更高，或者它们能够被平均或者类似地组合，以创建对最有问题的交换机进行有效排名的最终列表。

[0117] 本文上面阐述的实施例提供了用于监视在线服务链接的性能的新方法。它既能够用于测试服务链接在被建立之后的其的实现，也能够用于在运行时间期间进行连续监视。该解决方案能够连同任何业务导引机制一起被用在任何类型的网络中，以降低网络中的监视开销。

[0118] 图8是在软件定义网络中实现交换机并执行本文上面所定义的交换机的任何功能的网络装置的一个实施例的图。

[0119] 网络装置(ND)是通信地互连网络上的其它电子装置(例如其它网络装置、末端用户装置)的电子装置。一些网络装置是为多个连网功能(例如路由、桥接、交换、层2聚合、会话边界控制、服务质量和/或订户管理)提供支持和/或为多个应用服务(例如数据、语音和视频)提供支持的“多个服务网络装置”。

[0120] 在一个实施例中，该过程由网络装置801或类似计算装置实现。网络装置801能够具有使其能够接收数据业务并将之朝其目的地进行转发的任何结构。网络装置801能够包含执行网络装置801的功能的网络处理器803或网络处理器的集合。本文所用的“集合”是包含一个项目的任何正整数项目。网络装置801能够执行报告模块807，用来根据由控制器所确定的模型和配置文件，来实现与延迟和丢失的测量(包含聚合的和非聚合的测量)以及配置的实现相关的交换机的功能。

[0121] 网络处理器803能够用分立的硬件、软件模块或其的任何组合实现报告模块807。网络处理器803还能够服务于路由信息库805A以及与数据业务转发和网络拓扑维护相关的类似功能。路由信息库805A能够实现为被利用来转发协议数据单元PDU(即分组)的匹配动作表。报告模块807的功能能够被实现为网络装置内以软件(包含固件)和硬件的任何组合的形式的模块。由网络装置801实现和执行的报告模块807的功能包含本文上面进一步描述的那些。

[0122] 在一个实施例中，网络装置801能够包含线卡817的集合，线卡817通过标识目的地并将数据业务转发到适当的线卡817(具有经由下一跳跃导向或朝向目的地的出口端口)来处理进入数据业务并将之朝向相应目的地节点进行转发。这些线卡817还能够实现转发信息库805B，或其的有关子集。线卡817还能够实现或促进本文上面描述的报告模块807功能。线卡817经由交换机结构(switch fabric)811彼此通信，并使用以太网、光纤或类似通信链路和媒体通过附连网络821与其它节点通信。

[0123] 如本文所描述的，由网络装置801执行的操作可指的是硬件的特定配置(诸如配置成执行某些操作或具有预定功能性的专用集成电路(ASIC))或存储在实施在非暂态计算机可读存储媒体中的存储器中的软件指令。从而，图中示出的技术能够使用在一个或多个电子装置(例如末端站、网络单元)上存储和执行的代码和数据实现。此类电子装置使用计算机可读媒体(诸如非暂态计算机可读存储媒体(例如磁盘、光盘、随机存取存储器、只读存储器、闪速存储器装置、相变存储器))和暂态计算机可读通信媒体(例如电、光、声或其它形式的传播信号—诸如载波、红外信号、数字信号)来存储和传达(内部地和/或通过网络与其它

电子装置)代码与数据。另外,此类电子装置通常包含耦合到一个或多个其它组件诸如一个或多个存储装置(非暂态机器可读存储媒体)、用户输入/输出装置(例如键盘、触摸屏和/或显示器)以及网络连接的一个或多个处理器的集合。处理器集合与其它组件的耦合通常通过一个或多个总线和桥(也叫做总线控制器)进行。从而,给定电子装置的存储装置通常存储用于在那个电子装置的一个或多个处理器的集合上执行的代码和/或数据。本发明实施例的一个或多个部分可使用软件、固件和/或硬件的不同组合来实现。

[0124] 电子装置使用机器可读媒体(也称为计算机可读媒体),诸如机器可读存储媒体(例如磁盘、光盘、只读存储器(ROM)、闪速存储器装置、相变存储器)和机器可读传送媒体(也称为载体)(例如电、光、无线电、声或其它形式的传播信号—诸如载波、红外信号),来存储和传送(内部地和/或通过网络与其它电子装置)代码(其由软件指令构成并且其有时被称作为计算机程序代码或计算机程序)和/或数据。从而,电子装置(例如计算机)包含硬件和软件,诸如耦合到一个或多个机器可读存储媒体的一个或多个处理器的集合,所述一个或多个机器可读存储媒体用来存储用于在所述处理器的集合上执行的代码和/或用来存储数据。比如,电子装置可包含含有代码的非易失性存储器,由于非易失性存储器能够持久存留代码/数据,甚至当电子装置关闭时(当移除电源时),并且在电子装置开启时,要由那个电子装置的处理器执行的代码的那个部分通常从较慢的非易失性存储器拷贝到那个电子装置的易失性存储器(例如动态随机存取存储器(DRAM)、静态随机存取存储器(SRAM))。典型的电子装置还包含一个或多个物理网络接口的集合,用来与其它电子装置确立网络连接(或使用传播信号用来传送和/或接收代码和/或数据)。本发明的实施例的一个或多个部分可使用软件、固件和/或硬件的不同组合来实现。

[0125] 图9A示出了根据本发明的一些实施例的示范网络内网络装置(ND)之间的连接性以及ND的三个示范实现。图9A示出了ND 900A-H以及通过A-B、B-C、C-D、D-E、E-F、F-G和A-G之间以及H与A、C、D和G中每个之间的线形成的连接性。这些ND是物理装置,并且这些ND之间的连接性能够是无线的或有线的(经常称作为链路)。从ND 900A、E和F延伸的附加线示出这些ND充当网络的入口点和出口点(并且从而,这些ND有时被称作为边缘ND;而其它ND可被称为核心ND)。

[0126] 图9A中的两个示范ND实现是:1)使用定制专用集成电路(ASIC)和专有操作系统(OS)的特别用途网络装置902;以及2)使用公共商业化构件(COTS)处理器和标准OS的一般用途网络装置904。

[0127] 特别用途网络装置902包含连网硬件910,连网硬件910包括计算资源(912)(其通常包含一个或多个处理器的集合)、转发资源914(其通常包含一个或多个ASIC和/或网络处理器)和物理网络接口(NI)916(有时称为物理端口)以及其中已存储有连网软件920的非暂态机器可读存储媒体918。物理NI是ND中的硬件,通过其进行网络连接(例如无线地通过无线网络接口控制器(WNIC)或通过插入线缆至连接到网络接口控制器(NIC)的物理端口),诸如由ND 900A-H之间的连接性所示出的那些。在操作期间,连网软件920可由连网硬件910执行以例示一个或多个连网软件实例922的集合。每个连网软件实例922以及执行那个网络软件实例的连网硬件910的那个部分(不论它是专用于那个连网软件实例的硬件和/或由那个连网软件实例与连网软件实例922中的其它连网软件实例所暂时共享的硬件时间片)形成单独的虚拟网络元件930A-R。每个虚拟网络元件(VNE)930A-R包含控制通信和配置模块

932A-R(有时称作为本地控制模块或控制通信模块)以及转发表934A-R,使得给定虚拟网络元件(例如930A)包含控制通信和配置模块(例如932A)、一个或多个转发表(例如934A)的集合以及执行虚拟网络元件(例如930A)的连网硬件910的那个部分。在一些实施例中,控制通信和配置模块932A能够实现报告模块933A,报告模块933A实现用于本文上面描述的丢失和延迟的测量以及配置的交换机功能。

[0128] 特别用途网络装置902经常在物理上和/或逻辑上被视为包含:1)ND控制平面924(有时称作为控制平面),包括执行控制通信和配置模块932A-R的计算资源912;以及2)ND转发平面926(有时称作为转发平面、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。作为示例,其中ND是路由器(或正在实现路由功能性),ND控制平面924(执行控制通信和配置模块932A-R的计算资源912)通常负责参与控制数据(例如分组)要如何被路由(例如用于数据的下一跳跃和用于那个数据的外出物理NI)并将那个路由信息存储在转发表934A-R中,并且ND转发平面926负责在物理NI 916上接收那个数据,并基于转发表934A-R将那个数据转发出物理NI 916中的适当物理NI。

[0129] 电子装置使用机器可读媒体(也称为计算机可读媒体),诸如机器可读存储媒体(例如磁盘、光盘、只读存储器(ROM)、闪速存储器装置、相变存储器)和机器可读传送媒体(也称为载体)(例如电、光、无线电、声或其它形式的传播信号—诸如载波、红外信号),来存储和传送(内部地和/或通过网络与其它电子装置)代码(其由软件指令组成并且其有时被称作为计算机程序代码或计算机程序)和/或数据。从而,电子装置(例如计算机)包含硬件和软件,诸如耦合到一个或多个机器可读存储媒体的一个或多个处理器的集合,所述一个或多个机器可读存储媒体用来存储用于在所述处理器的集合上执行的代码和/或存储数据。比如,电子装置可包含含有代码的非易失性存储器,由于非易失性存储器能够持久存留代码/数据,甚至当电子装置关闭时(当移除电源时),并且当电子装置开启时,要由那个电子装置的处理器执行的代码的那个部分通常从较慢的非易失性存储器拷贝到那个电子装置的易失性存储器(例如动态随机存取存储器(DRAM)、静态随机存取存储器(SRAM))。典型的电子装置还包含一个或多个物理网络接口或集合,用来与其它电子装置确立网络连接(或使用传播信号用来传送和/或接收代码和/或数据)。本发明的实施例的一个或多个部分可使用软件、固件和/或硬件的不同组合来实现。

[0130] 图9B示出了根据本发明的一些实施例实现特别用途网络装置902的示范方式。图9B示出了包含卡938(通常是可热插的)的特别用途网络装置。虽然在一些实施例中卡938具有两种类型(操作为ND转发平面926的一个或多个(有时称为线卡)以及操作以实现ND控制平面924的一个或多个(有时称为控制卡)),但备选实施例可将功能性组合在单个卡上和/或包含另外的卡类型(例如一个另外的卡类型被称为服务卡、资源卡或多应用卡)。服务卡能够提供特定处理(例如,层4到层7服务(例如防火墙、互联网协议安全性(IPsec) (RFC 4301和4309)、安全套接字层(SSL)/传输层安全性(TLS)、入侵检测系统(IDS)、对等(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)、演进的分组核心(EPC)网关))。作为示例,服务卡可用于终止IPsec隧道,并执行伴随的认证和加密算法。这些卡通过示出为背板936的一个或多个互连机制(例如第一全网耦合线卡且第二全网耦合所有卡)被耦合在一起。

[0131] 返回图9A,一般用途网络装置904包含硬件940,硬件940包括以下装置的集合:一

个或多个处理器942(其经常是COTS处理器)和网络接口控制器944(NIC;也称为网络接口卡)(其包含物理NI 946)以及其中已存储有软件950的非暂态机器可读存储媒体948。在操作期间,处理器942执行软件950,以例示监管程序954(有时称作为虚拟机监视器(VMM))以及由监管程序954运行的一个或多个虚拟机962A-R,它们共同被称作为软件实例952。虚拟机是物理机的软件实现,它运行程序就好像它们正在物理非虚拟化机器上执行一样;而且尽管一些系统提供了准虚拟化(其为了优化目的允许操作系统或应用知晓虚拟化的存在),但与运行在"裸机(bare metal)"主机电子装置上相对,应用一般不知道它们正运行在虚拟机上。每一个虚拟机962A-R以及执行那个虚拟机的硬件940的部分(不论它是专用于那个虚拟机的硬件和/或由那个虚拟机与虚拟机962A-R中的其它虚拟机所暂时共享的硬件时间片)形成单独的虚拟网络元件960A-R。

[0132] 虚拟网络元件960A-R执行与虚拟网络元件930A-R类似的功能性。比如,监管程序954能够呈现虚拟操作平台,它对虚拟机962A看起来就像是连网硬件910,并且虚拟机962A可用于实现类似于控制通信和配置模块932A以及转发表934A的功能性(硬件940的这个虚拟化有时被称作为网络功能虚拟化(NFV))。从而,NFV可用于将许多网络设备类型整合在工业标准高容量服务器硬件、物理交换机和物理存储装置上,其能够位于数据中心、ND和客户场所设备(CPE)中。然而,本发明的不同实施例可以不同方式实现虚拟机962A-R中的一个或多个。例如,虽然本发明的实施例采用每个虚拟机962A-R对应于一个VNE 960A-R来示出,但备选实施例可按照更细级别的粒度实现这个对应关系(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);应该理解到,本文中关于虚拟机到VNE的对应关系进行描述的技术还适应于其中使用此类更细级别的粒度的实施例。

[0133] 此外,虚拟网络元件960A-R和虚拟机962A-R能够实现本文上面所描述的报告模块963A-R的功能,其中报告模块963A-R实现用于配置以及延迟和丢失测量的交换机功能。

[0134] 在某些实施例中,监管程序954包含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,这个虚拟交换机在虚拟机与NIC 944之间以及可选地在虚拟机962A-R之间来转发业务;此外,这个虚拟交换机可加强按照策略不被准许彼此通信(例如通过尊重虚拟局域网(VLAN))的VNE 960A-R之间的网络隔离。

[0135] 图9A中的第三示范ND实现是混合网络装置906,其在单个ND中或ND内的单个卡中既包含定制ASCI/专有OS也包含COTS处理器/标准OS。在此类混合网络装置的某些实施例中,平台VM(即,实现特别用途网络装置902的功能性的VM)能够向混合网络装置906中存在的连网硬件提供准虚拟化。

[0136] 不管ND的以上示范实现如何,当考虑由ND实现的多个NVE中的单个NVE时(例如仅一个VNE是给定虚拟网络的部分),或者其中仅单个VNE当前由ND实现,简写术语网络元件(NE)有时用来指的是那个VNE。还有在所有以上示范实现中,每个VNE(例如VNE 930A-R、VNE 960A-R以及混合网络装置906中的那些)在物理NI(例如916、946)上接收数据,并将那个数据转发出物理NI(例如916、946)中的适当物理NI。例如,实现IP路由器功能性的VNE基于IP分组中的一些IP报头信息来转发IP分组;其中IP报头信息包含源IP地址、目的地IP地址、源端口、目的地端口(其中与ND的物理端口相对,"源端口"和"目的地端口"在本文中指的是协议端口)、传输协议(例如用户数据报协议(UDP) (RFC 768、2460、2675、4113和5405)、传送控制协议(TCP) (RFC 793和1180)以及差分服务(DSCP)值(RFC 2474、2475、

2597、2983、3086、3140、3246、3247、3260、4594、5865、3289、3290和3317)。

[0137] 图9C示出根据本发明的一些实施例其中可耦合VNE的各种示范方式。图9C示出ND 900H中的VNE 970H.1和ND 900A中实现的VNE 970A.1-970A.P(且可选地VNE 970A.Q-970A.R)。在图9C中,VNE 970A.1-P它们能从ND 900A外部接收分组并将分组转发到ND 900A外部,在这个意义上来说它们是彼此分开的;VNE 970A.1与VNE 970H.1耦合,并且从而它们在它们的相应ND之间传达分组;VNE 970A.2-970A.3可选地可在它们自身之间转发分组,而无需将它们转发到ND 900A外部;并且VNE 970A.P可选地可以是包含VNE 970A.Q后跟随有VNE 970A.R的VNE链中的第一个(这有时被称作为动态服务链接,其中VNE系列中的每一个VNE提供不同的服务—例如一个或多个层4-7网络服务)。虽然图9C示出了VNE之间的各种示范关系,但备选实施例可支持其它关系(例如更多/更少VNE、更多/更少动态服务链、具有一些公共VNE和一些不同VNE的多个不同动态服务链)。

[0138] 图9A的ND例如可形成互联网或专用网络的部分;并且其它电子装置(未示出;诸如末端用户装置,包含工作站、膝上型计算机、上网本、平板电脑、掌上型计算机、移动电话、智能电话、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、可穿戴装置、游戏系统、机顶盒、互联网使能的家用电器)可耦合到网络(直接或通过其它网络诸如接入网)以通过网络(例如互联网或在互联网上上覆(例如通过互联网遂穿)的虚拟专用网(VPN))彼此通信(直接或通过服务器)和/或访问内容和/或服务。此类内容和/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)或参与对等(P2P)服务的一个或多个末端用户装置(未示出)提供,并且例如可包含公用网页(例如免费内容、店面、搜索服务)、私人网页(例如提供电子邮件服务的用户名/密码访问的网页)和/或通过VPN的企业网络。比如,末端用户装置可耦合(例如通过耦合(有线或无线地)到接入网的客户场所设备)到边缘ND,边缘ND耦合(例如通过一个或多个核心ND)到其它边缘ND,其它边缘ND耦合到充当服务器的电子装置。然而,通过计算和存储虚拟化,在图9A中操作为ND的一个或多个电子装置还可托管一个或多个此类服务器(例如在一般用途网络装置904的情况中,虚拟机962A-R中的一个或多个可操作为服务器;这将同样适应于混合网络装置906;在特别用途网络装置902的情况中,一个或多个此类服务器还能够运行在由计算资源912执行的监管程序上);在此情况中服务器被称为与那个ND的VNE共置。

[0139] 虚拟网络是提供网络服务(例如L2和/或L3服务)的物理网络(诸如图9A中的那个)的逻辑抽象。虚拟网络能够实现为上覆网络(overlay network)(有时称作为网络虚拟化上覆),其通过底层网络(underlay network)(例如L3网络,诸如使用隧道(例如一般路由封装(GRE)、层2遂穿协议(L2TP)、IPSec)来创建上覆网络的互联网协议(IP)网络)提供网络服务(例如层2(L2数据链路层)和/或层3(L3网络层)服务)。

[0140] 网络虚拟化边缘(NVE)坐落在底层网络的边缘处,并且参与实现网络虚拟化;NVE的面向网络侧使用底层网络来往于其它NVE进行帧隧穿;NVE的面向外部侧向网络外部系统发送数据并从中接收数据。虚拟网络实例(VNI)是NVE上虚拟网络的特定实例(例如ND上NE/VNE、ND上NE/NVE的部分(其中那个NE/VNE通过仿真被分成多个VNE));一个或多个VNI能够被例示在NVE上(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能够是通过逻辑接口标识符(例如VLAN ID)所标识的物理端口或虚拟端口。

[0141] 网络服务的示例包含:1)以太网LAN仿真服务(基于以太网的多点服务,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),其中外部系统通过底层网络上的LAN环境跨网络互连(例如,NVE为不同的此类虚拟网络提供单独L2 VNI(虚拟交换实例)以及跨底层网络的L3(例如IP/MPLS)遂穿封装);以及2)虚拟化IP转发服务(从服务定义角度类似于IETF IP VPN(例如边界网关协议(BGP)/MPLS IPVPN RFC 4364)),其中外部系统通过底层网络上的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L3 VNI(转发和路由实例)以及跨底层网络的L3(例如IP/MPLS)遂穿封装)。网络服务还可包含服务能力(例如业务归类标记、业务调节和调度)、安全性能力(例如过滤器用来保护客户场所免于网络发起的攻击,以避免畸形的路由通告)以及管理能力(例如全检测和处理)的质量。

[0142] 图9D示出了根据本发明的一些实施例在图9A的每个ND上具有单个网络元件的网络,并且在这个直接的途径内,将传统分布式途径(通常由传统路由器使用)与用于维持可达性和转发信息(也称为网络控制)的集中式途径进行对照。具体地,图9D示出了与图9A的ND 900A-H具有相同连接性的网络元件(NE) 970A-H。

[0143] 图9D示出了分布式途径972将用于生成可达性和转发信息的责任跨NE 970A-H进行分布;换言之,分布邻居发现和拓扑发现的过程。

[0144] 例如,在使用特殊用途网络装置902的情况中,ND控制平面924的控制通信和配置模块932A-R通常包含可达性和转发信息模块以实现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) (RFC 4271)、内部网关协议(IGP)(例如开放最短路径优先(OSPF) (RFC 2328和5340)、中间系统到中间系统(IS-IS) (RFC 1142)、路由信息协议(RIP)(版本1 RFC 1058、版本2 RFC 2453以及下一代RFC 2080))、标签分布协议(LDP)(RFC 5036)、资源保留协议(RSVP)(RFC 2205、2210、2211、2212以及RSVP业务工程设计(TE):对LSP隧道RFC 3209的RSVP的扩充、通用多协议标签交换(GMPLS)信令RSVP-TE RFC 3473、RFC 3936、4495和4558),它们与其它NE通信以交换路由,并且然后基于一个或多个路由度量选择那些路由。从而,NE 970A-H(例如执行控制通信和配置模块932A-R的计算资源912)通过分布式确定网络内的可达性并演算它们相应的转发信息来执行它们对于参与控制要如何路由数据(例如分组)(例如用于数据的下一跳跃和用于那个数据的外出物理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采用基于路由结构的信息(例如邻接和路由信息)对ND转发平面926编程。例如,ND控制平面924将邻接和路由信息编程为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(RIB)、标签转发信息库(LFIB)和一个或多个邻接结构)中。对于层2转发,ND能够存储用来转发数据(基于那个数据中的层2信息)的一个或多个桥接表。虽然上面的示例使用特殊用途网络装置902,但相同的分布式途径972能够被实现在一般用途网络装置904和混合网络装置906上。

[0145] 图9D示出了集中式途径974(也称为软件定义的连网(SDN)),集中式途径974解耦对关于业务是从底层系统哪里被发送进行判定的系统,其将业务转发到选择的目的地。所示出的集中式途径974具有对于在集中式控制平面976(有时称作为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、网络虚拟化权威机构(authority)或管理控制实体)中生成可达性和转发信息的责任,并且从而邻居发现和拓扑

发现的过程被集中化。集中式控制平面976具有南向接口982，其具有数据平面980(有时称作为基础设施层、网络转发平面或转发平面(其不应该与ND转发平面混淆))，数据平面980包含NE 970A-H(有时称作为交换机、转发元件、数据平面元件或节点)。集中式控制平面976包含网络控制器978，网络控制器978包含集中式可达性和转发信息模块979，模块979确定网络内的可达性并通过南向接口982(其可使用OpenFlow协议)向数据平面980的NE 970A-H分布转发信息。从而，网络智能被集中化于执行在通常与ND分开的电子装置上的集中式控制平面976中。

[0146] 例如，在数据平面980中使用特殊用途网络装置902的情况中，ND控制平面924的每个控制通信和配置模块932A-R通常包含提供南向接口982的VNE侧的控制代理。在此情况中，ND控制平面924(执行控制通信和配置模块932A-R的计算资源912)通过控制代理与集中式控制平面976通信来从集中式可达性和转发信息模块979接收转发信息(并且在一些情况中是可达性信息)来执行其对于参与控制要如何路由数据(例如分组)(例如用于数据的下一跳跃和用于那个数据的外出物理NI)的责任(应该理解到，在本发明的一些实施例中，控制通信和配置模块932A-R除了与集中式控制平面976通信，还可在确定可达性和/或演算转发信息中起到一些作用——尽管比在分布式途径的情况中少有一些;此类实施例一般被视为落在集中式途径974下，但也能够被视为混合途径)。

[0147] 虽然上面示例使用特殊用途网络装置902，但能够采用混合网络装置906以及一般用途网络装置904(例如，每个VNE 960A-R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(并且在一些情况中是可达性信息)来执行其对于控制要如何路由数据(例如分组)(例如用于数据的下一跳跃和用于那个数据的外出物理NI)的责任;应该理解到，在本发明的一些实施例中，VNE 960A-R除了与集中式控制平面976通信外，还可在确定可达性和/或演算转发信息中起到一些作用——尽管比在分布式途径的情况中少一些)实现相同的集中式途径974。事实上，使用SDN技术能够增强通常在一般用途网络装置904或混合网络装置906实现中使用的NFV技术，因为NFV能够通过提供其上能运行SDN软件的基础设施来支持SDN，并且NFV和SDN都目的在于利用商品服务器硬件和物理交换机。

[0148] 图9D还示出了集中式控制平面976具有到应用层986的北向接口984，应用层986中驻留应用988。集中式控制平面976具有形成用于应用988的虚拟网络992(有时称作为逻辑转发平面、网络服务或上覆网络(其中数据平面980的NE 970A-H是底层网络))的本领。从而，集中式控制平面976维持所有ND和配置的NE/NVE的全局视图，并且它将虚拟网络有效地映射到底层ND(包含在物理网络通过硬件(ND、链路或ND组件)故障、添加或移除而改变时，维持这些映射)。

[0149] 虽然图9D示出了分布式途径972与集中式途径974分开，但在本发明的某些实施例中可以不同方式分布网络控制的工作，或将所述二者组合。例如:1)实施例一般可使用集中式途径(SDN) 974，但具有委托给NE的某些功能(例如，分布式途径可用于实现以下中的一个或多个:故障监视、性能监视、保护交换、以及邻居和/或拓扑发现的基元(primitive))；或者2)本发明的实施例可经由集中式控制平面和分布式协议二者来执行邻居发现和拓扑发现，并且比较结果以提出它们所不同意的例外。此类实施例一般被视为落在集中式途径974下，但也能够被视为混合途径。

[0150] 虽然图9D示出了其中每个ND 900A-H实现单个NE 970A-H的简单情况，但应该理解，参考图9D描述的网络控制途径同样适用于其中ND 900A-H中的一个或多个实现多个VNE（例如VNE 930A-R、VNE 960A-R，在混合网络装置906中的那些）的网络。备选地或另外，网络控制器978还可仿真单个ND中多个VNE的实现。具体地，代替在单个ND中实现多个NVE（或除此之外），网络控制器978可将单个ND中的VNE/NE的实现呈现为虚拟网络992中的多个VNE（所有都位于虚拟网络992的同一个中、每个位于虚拟网络992的不同虚拟网络中、或某种组合）。例如，网络控制器978可引起ND实现底层网络中的单个VNE（NE），并且然后逻辑上分割集中式控制平面976内那个NE的资源，以在虚拟网络992中呈现不同的VNE（其中上覆网络中的这些不同VNE正在共享底层网络中ND上的单个VNE/NE实现的资源）。

[0151] 另一方面，图9E和9F分别示出了网络控制器978可将之作为虚拟网络992的不同虚拟网络的部分来呈现的NE和VNE的示范抽象。图9E示出了根据本发明的一些实施例的简单情况，其中每个ND 900A-H实现单个NE 970A-H（见图9D），但集中式控制平面976已将不同ND中的多个NE（NE 970A-C和G-H）抽象成（以表示）图9D的虚拟网络992之一中的单个NE 970I。图9E示出了在这个虚拟网络中，NE 970I耦合到NE 970D和970F，它们二者仍耦合到NE 970E。

[0152] 图9F示出了根据本发明的一些实施例的情况，其中多个VNE（VNE 970A.1和VNE 970H.1）被实现在不同ND（ND 900A和ND 900H）上，并且彼此耦合，以及其中集中式控制平面976已对这些多个VNE进行了抽象，使得它们看起来好像图9D的虚拟网络992之一内的单个VNE 970T。从而，NE或VNE的抽象能够跨越多个ND。

[0153] 虽然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，在单个电子装置上运行的单个软件实例），但备选实施例为了冗余和/或可缩放性目的可将功能性跨多个实体（例如运行在不同电子装置上的多个软件实例）进行分散。

[0154] 类似于网络装置实现，运行集中式控制平面976的电子装置，以及因而包含集中式可达性和转发信息模块979的网络控制器978可以各种方式（例如特殊用途装置、一般用途（例如COTS）装置或混合装置）实现。这些电子装置类似地将包含计算资源、一个或多个物理NIC或其的集合、以及其上已存储有集中式控制平面软件的非暂态机器可读存储媒体。比如，图10示出一般用途控制平面装置1004，其包含硬件1040，硬件1040包括以下装置的集合：一个或多个处理器1042（其经常是COTS处理器）和网络接口控制器1044（NIC；也称为网络接口卡）（其包含物理NI 1046）以及其中已存储有集中式控制平面（CCP）软件1050的非暂态机器可读存储媒体1048。

[0155] 在使用计算虚拟化的实施例中，处理器1042通常执行软件以例示监管程序1054（有时称作为虚拟机监视器（VMM））以及由监管程序1054运行的一个或多个虚拟机1062A-R；它们共同被称作为软件实例1052。虚拟机是物理机的软件实现，其运行程序就好像它们正在物理非虚拟化机器上执行一样；并且尽管一些系统提供了准虚拟化（其为了优化目的而允许操作系统或应用知晓虚拟化的存在），但与运行在“裸机”主机电子装置上相对，应用一般不知晓它们正运行在虚拟机上。此外，在其中使用计算虚拟化的实施例中，在操作期间，在操作系统1064A顶上的CCP软件1050（示出为CCP实例1076A）的实例通常在虚拟机1062A内被执行。在其中不使用计算虚拟化的实施例中，在操作系统1064A顶上的CCP实例1076A在“裸机”一般用途控制平面装置1004上被执行。

[0156] 操作系统1064A提供了基本处理、输入/输出(I/O)和连网能力。在一些实施例中，CCP实例1076A包含网络控制器实例1078。网络控制器实例1078包含集中式可达性和转发信息模块实例1079(其是中间件层，该中间件层向操作系统1064A提供网络控制器978的上下文并与各种NE通信)，以及中间件层(提供对于各种网络操作所要求的智能(intelligence)，诸如协议、网络态势感知和用户界面)上的CCP应用层1080(有时称作为应用层)。在更抽象级别，集中式控制平面976内的这个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。

[0157] 集中式控制平面976基于对于每个流的中间件层映射和CCP应用层1080演算来向数据平面980传送有关消息。流可被定义为分组的集合，其报头匹配给定的位模式；在这个意义上，传统IP转发也是基于流的转发，其中流例如由目的地IP地址定义；然而，在其它实现中，用于流定义的给定位模式可在分组报头中包含更多字段(例如10个或更多)。数据平面980的不同ND/NE/VNE可接收不同消息，以及因而不同的转发信息。数据平面980处理这些消息，并对适当的NE/VNE的转发表(有时称作为流表)中的适当流信息和对应动作进行编程，并且然后NE/VNE将进入的分组映射到在转发表中表示的流，并基于转发表中的匹配转发分组。

[0158] 诸如OpenFlow的标准定义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含报头剖析、分组归类以及进行转发判定。报头剖析描述了如何基于公知的协议集合来解释分组。一些协议字段用于建造将在分组归类中使用的匹配结构(或密钥)(例如，第一密钥字段能够是源媒体访问控制(MAC)地址，并且第二秘钥字段能够是目的地MAC地址)。

[0159] 分组归类牵涉到在存储器中执行查找，以通过基于转发表条目的匹配结构或密钥来确定转发表中的哪个条目(也称作为转发表条目或流条目)最佳地匹配分组来对分组进行归类。有可能在转发表条目中表示的许多流能与分组对应/匹配；在此情况中，系统通常配置成根据定义的方案(例如，选择匹配的第一转发表条目)从许多中确定一个转发表条目。转发表条目既包含特定集合的匹配判据(值或通配符的集合，或对于分组的什么部分应该与具体值/多个值/通配符相比较的指示，如匹配能力所定义的—对于分组报头中的特定字段，或对于某一其它分组内容)以及数据平面在接收到匹配分组时要采取的一个或多个动作的集合。例如，一个动作能够是将报头推到分组上(对于使用具体端口的分组)、充满(flood)分组、或简单地丢弃分组。从而，用于具有具体传送控制协议(TCP)目的地端口的IPv4/IPv6分组的转发表条目能够含有规定应该丢弃这些分组的动作。

[0160] 进行转发判定和执行动作基于在分组归类期间标识的转发表条目，通过执行分组上的匹配的转发表条目中所标识的动作的集合来发生。

[0161] 然而，当未知分组(例如，如OpenFlow用语中所使用的“未中的分组”或“匹配未中”)到达数据平面980时，分组(或者内容和分组报头的子集)通常被转发到集中式控制平面976。集中式控制平面976然后会将转发表条目编程到数据平面980中，以容纳属于未知分组流的分组。一旦特定的转发表条目已被集中式控制平面976编程到数据平面980中，具有匹配证书的下一分组将会匹配那个转发表条目，并采取与那个所匹配的条目相关联的动作的集合。

[0162] 网络接口(NI)可以是物理的或虚拟的；并且在IP上下文中，接口地址是被指配给

NI的IP地址，无论它是物理NI或虚拟NI。虚拟NI可与物理NI关联、与另一虚拟接口关联，或者依靠其自身(例如，环回接口、点对点协议接口)。NI(物理的或虚拟的)可被编号(具有IP地址的NI)或不被编号(没有IP地址的NI)。环回接口(及其环回地址)是经常用于管理用途的NE/VNE(物理的或虚拟的)的特定类型的虚拟NI(以及IP地址)；其中此类IP地址被称作为节点环回地址。指配给ND的NI的IP地址被称作为那个ND的IP地址；在更高粒度级，对指配给在ND上实现的NE/VNE的NI所指配的IP地址能够被称作为那个NE/VNE的IP地址。

[0163] 例如，虽然图中的流程图示出了由本发明的某些实施例所执行的操作的具体顺序，但应该理解，此类顺序是示范性的(例如备选实施例可按不同顺序执行操作，组合某些操作，交叠某些操作等)。

[0164] 虽然本发明已经依据若干实施例进行了描述，但本领域中的那些技术人员将认识到，本发明不限于所描述的实施例，在附带权利要求的精神和范畴内能够采用修改和变更被实践。从而，该描述要被当作是说明性的，而不是限制性的。

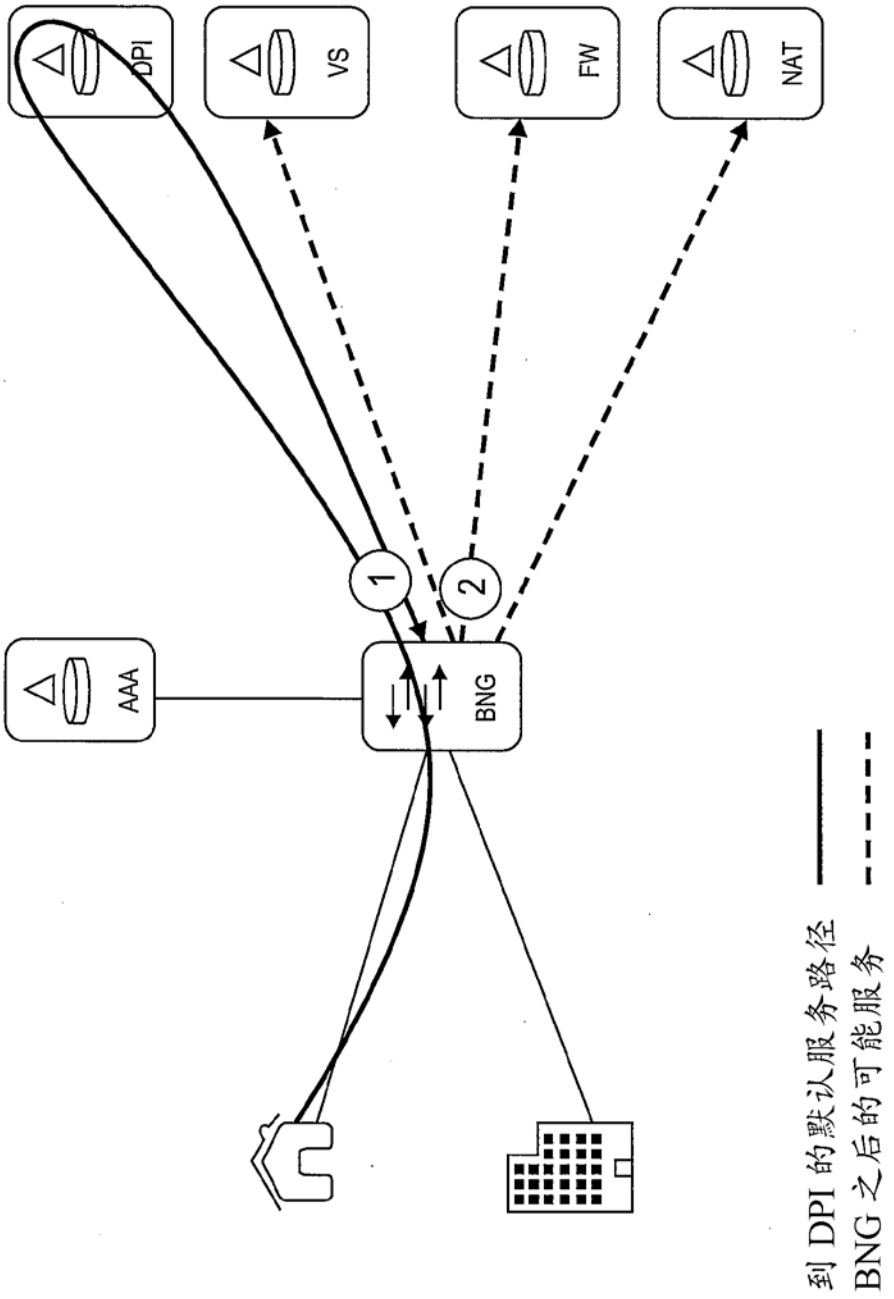


图 1

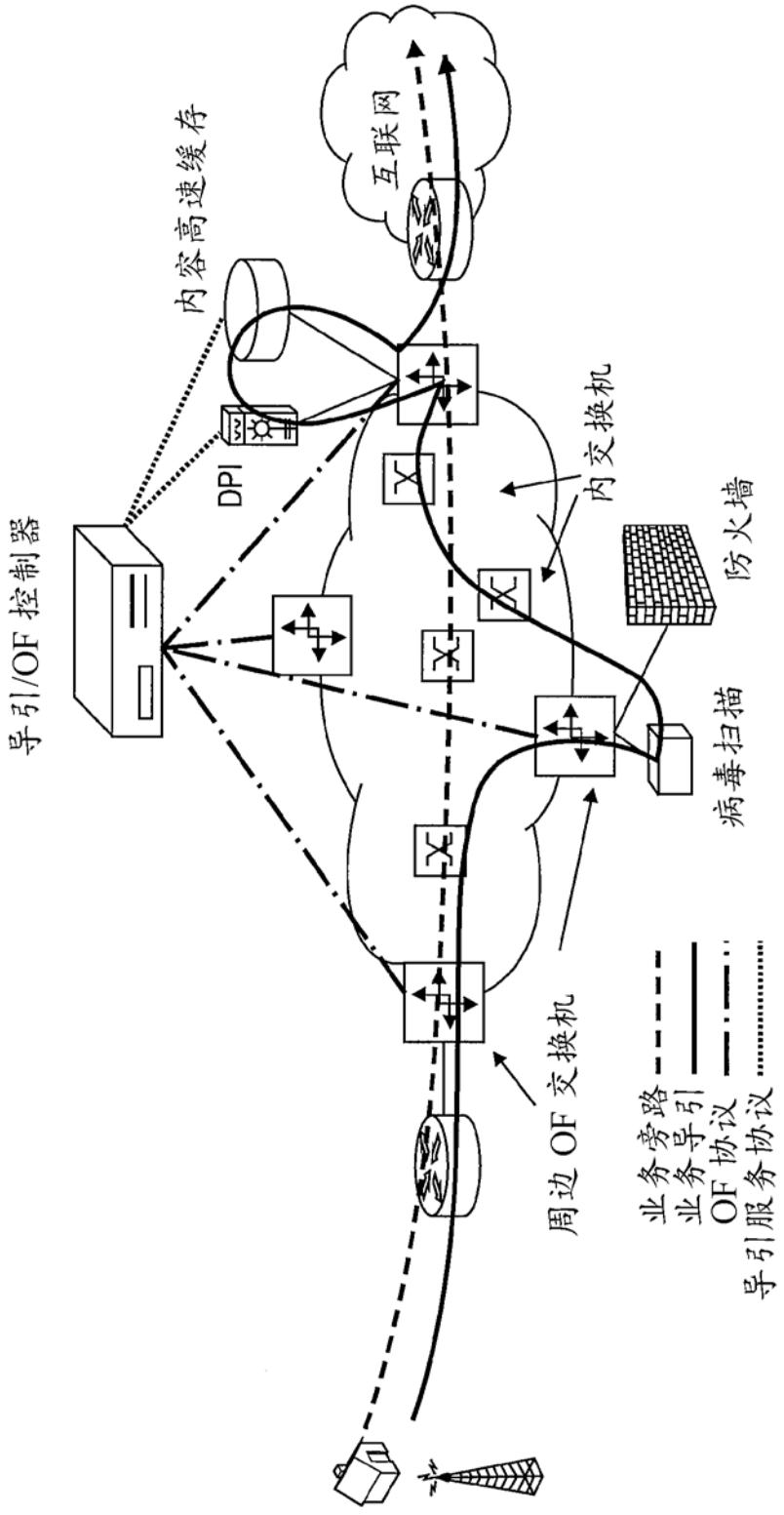

互联网
内容高速缓存
DPI
导引/OF 控制器
内交换机
防火墙
病毒扫描
周边 OF 交换机
业务旁路
业务导引
OF 协议
导引服务协议

图 2

在 SDN 中交换机的基本延迟丢失测量

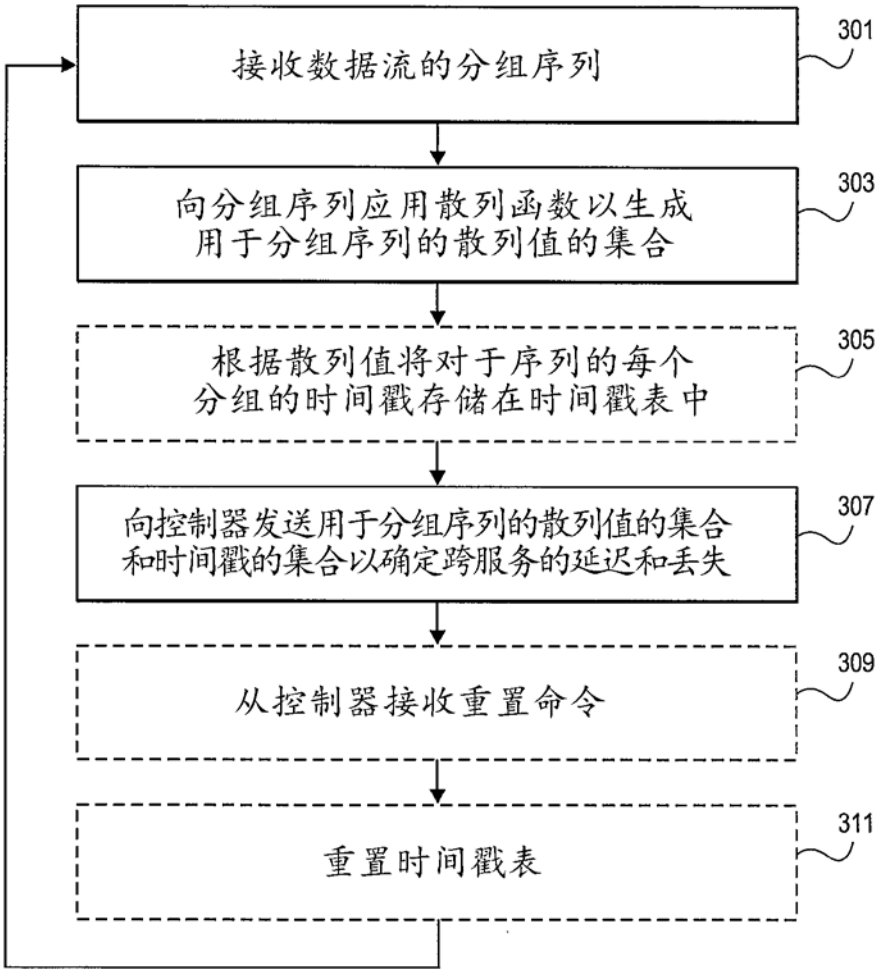


图 3A

在控制器的基本延迟丢失测量

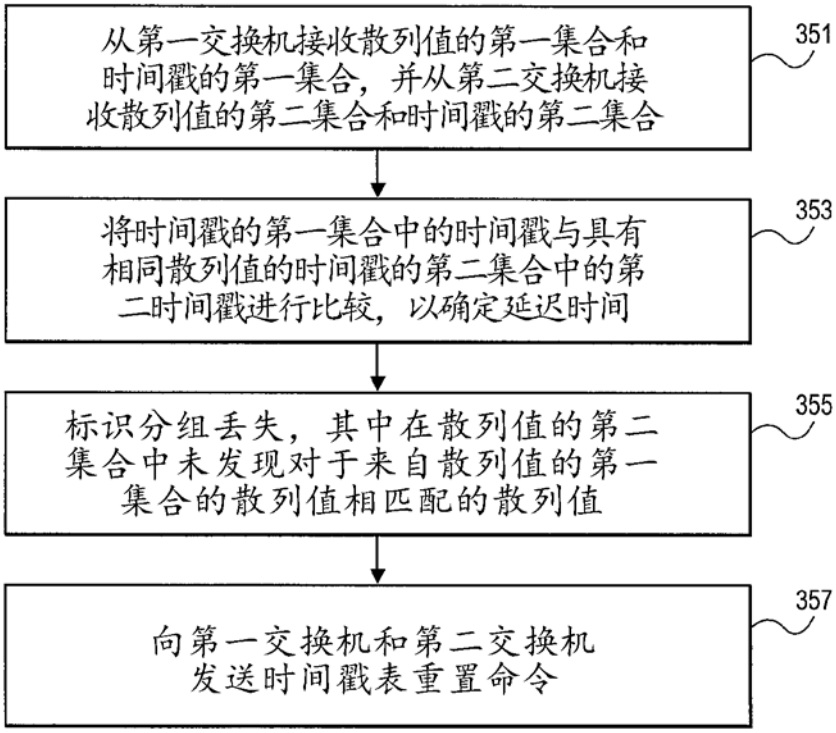


图 3B

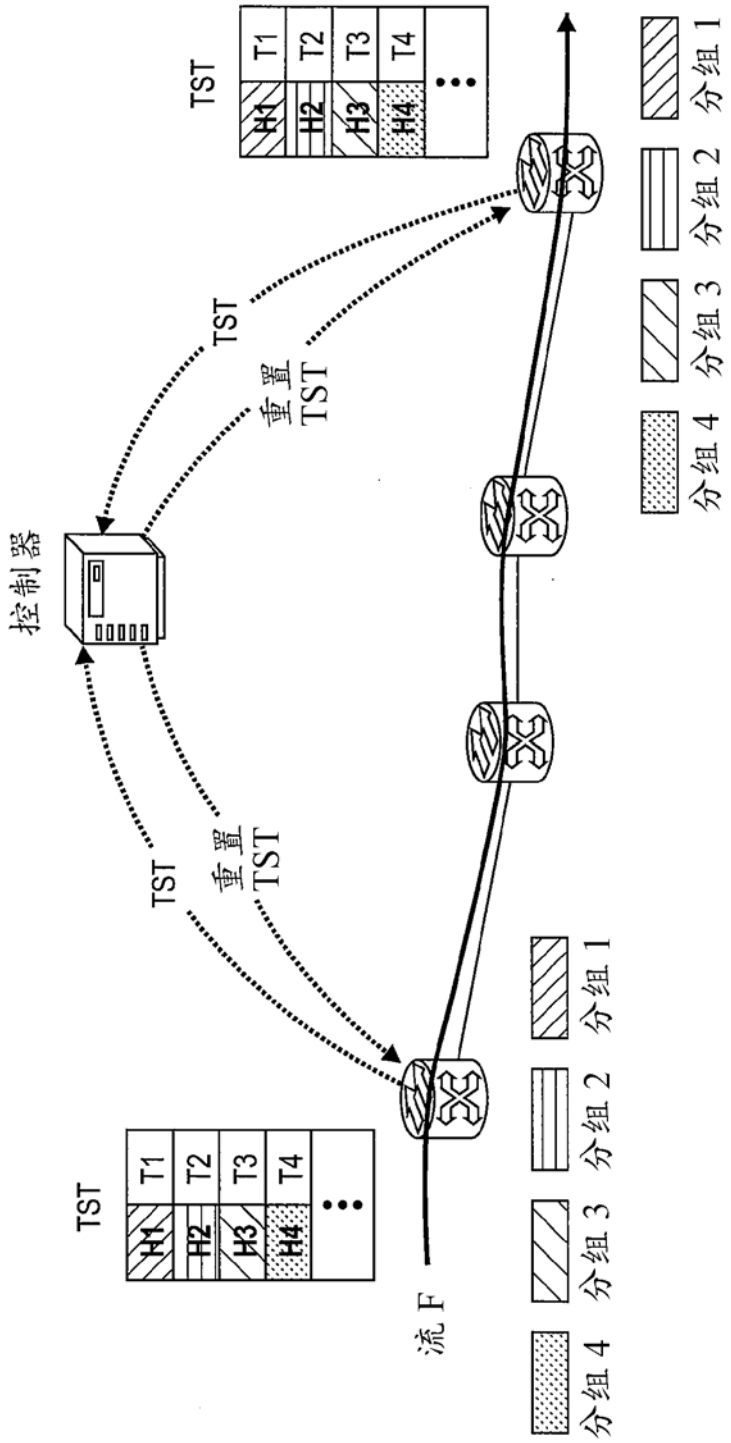

TST
T1
T2
T3
T4
H1
H2
H3
H4
TST
重置TST
控制器
TST
重置TST
TST
T1
T2
T3
T4
H1
H2
H3
H4
流 F
分组 1
分组 2
分组 3
分组 4
分组 1
分组 2
分组 3
分组 4

图 4

在 SDN 中交换机的聚合延迟丢失测量

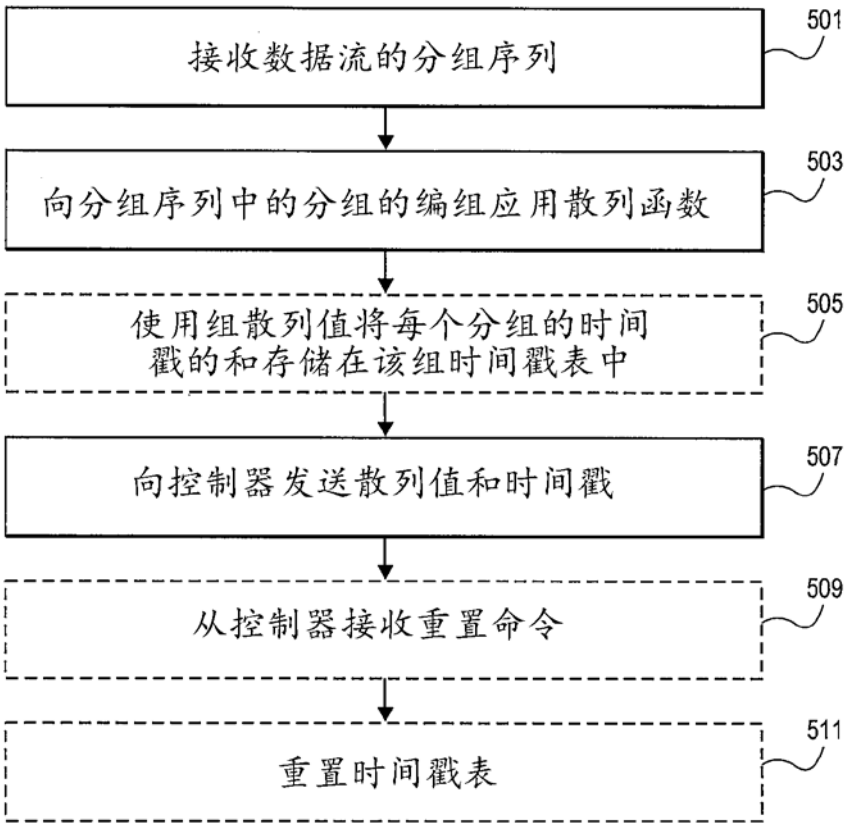


图 5A

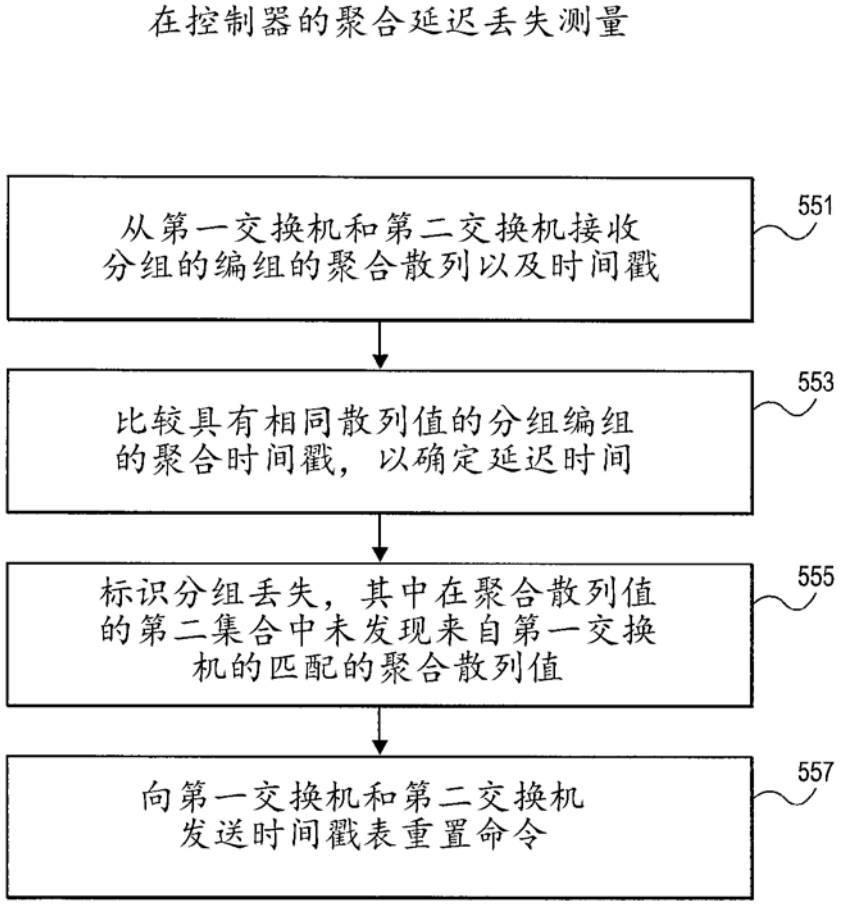

在控制器的聚合延迟丢失测量
从第一交换机和第二交换机接收分组的编组的聚合散列以及时间戳
551
比较具有相同散列值的分组编组的聚合时间戳，以确定延迟时间
553
标识分组丢失，其中在聚合散列值的第二集合中未发现来自第一交换机的匹配的聚合散列值
555
向第一交换机和第二交换机发送时间戳表重置命令
557

图 5B

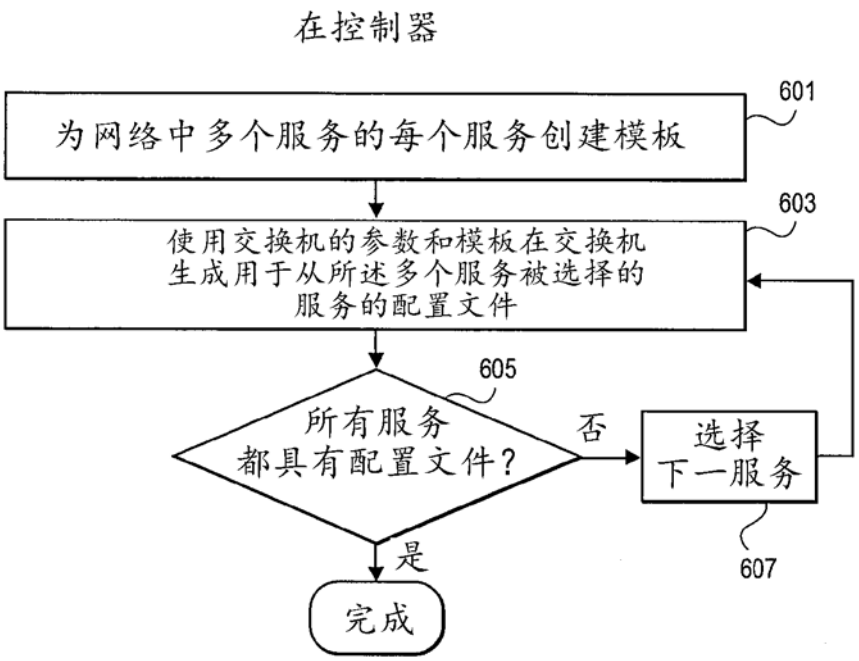

在控制器
为网络中多个服务的每个服务创建模板
601
使用交换机的参数和模板在交换机生成用于从所述多个服务被选择的服务的配置文件
603
所有服务都具有配置文件？
605
否
选择下一服务
607
是
完成

图 6A

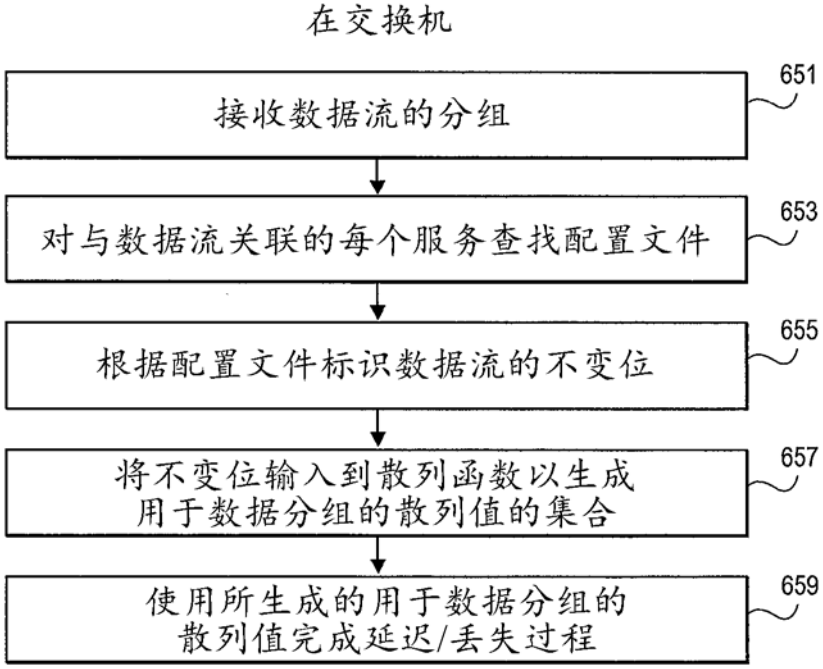

在交换机
接收数据流的分组
651
对与数据流关联的每个服务查找配置文件
653
根据配置文件标识数据流的不变位
655
将不变位输入到散列函数以生成
用于数据分组的散列值的集合
657
使用所生成的用于数据分组的
散列值完成延迟/丢失过程
659

图 6B

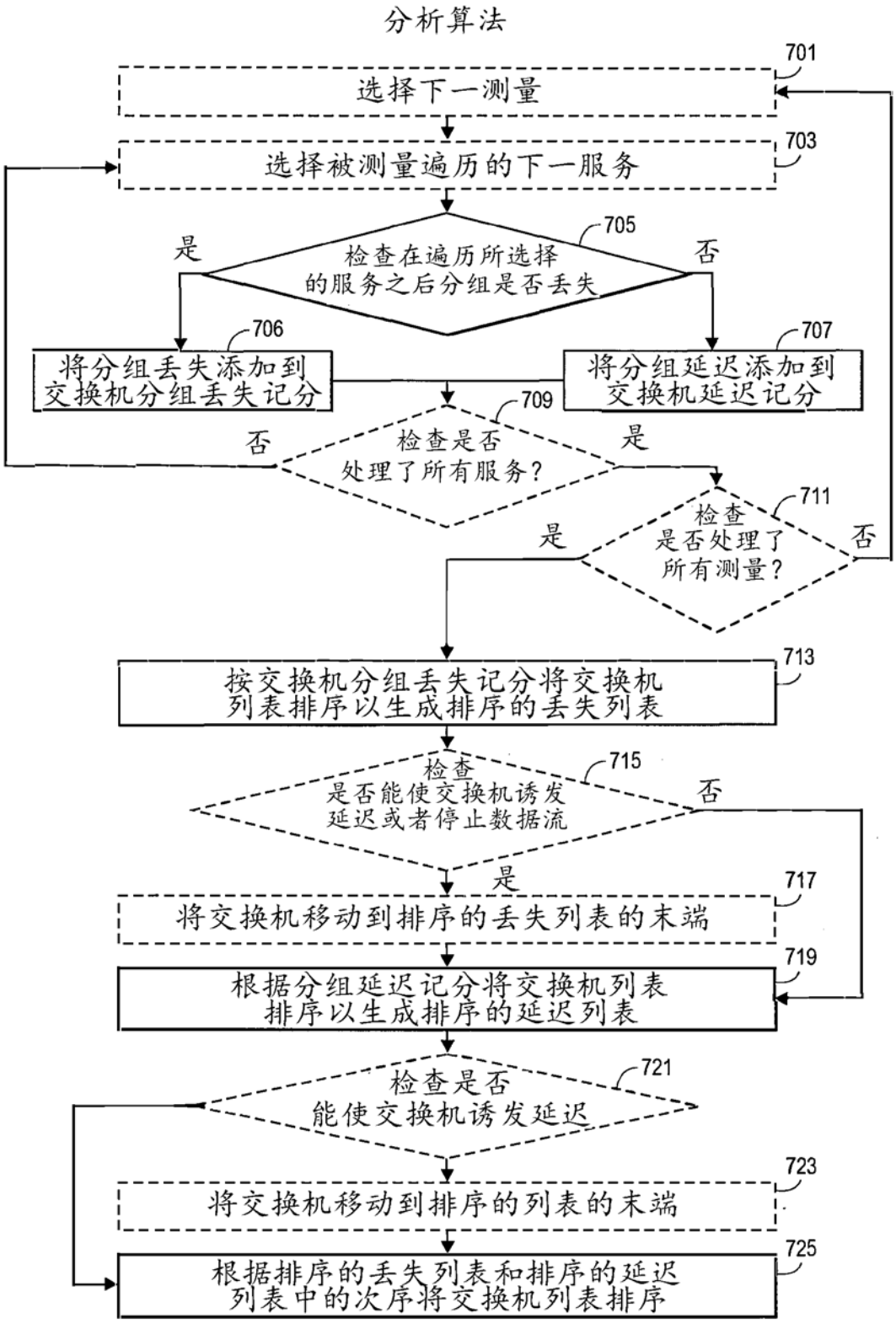

分析算法
701
选择下一测量
703
选择被测量遍历的下一服务
705
检查在遍历所选择的服务之后分组是否丢失
是
否
706
将分组丢失添加到交换机分组丢失记分
707
将分组延迟添加到交换机延迟记分
709
检查是否处理了所有服务?
否
是
711
检查是否处理了所有测量?
是
否
713
按交换机分组丢失记分将交换机列表排序以生成排序的丢失列表
715
检查是否能使交换机诱发延迟或者停止数据流
否
是
717
将交换机移动到排序的丢失列表的末端
719
根据分组延迟记分将交换机列表排序以生成排序的延迟列表
721
检查是否能使交换机诱发延迟
723
将交换机移动到排序的列表的末端
725
根据排序的丢失列表和排序的延迟列表中的次序将交换机列表排序

图 7

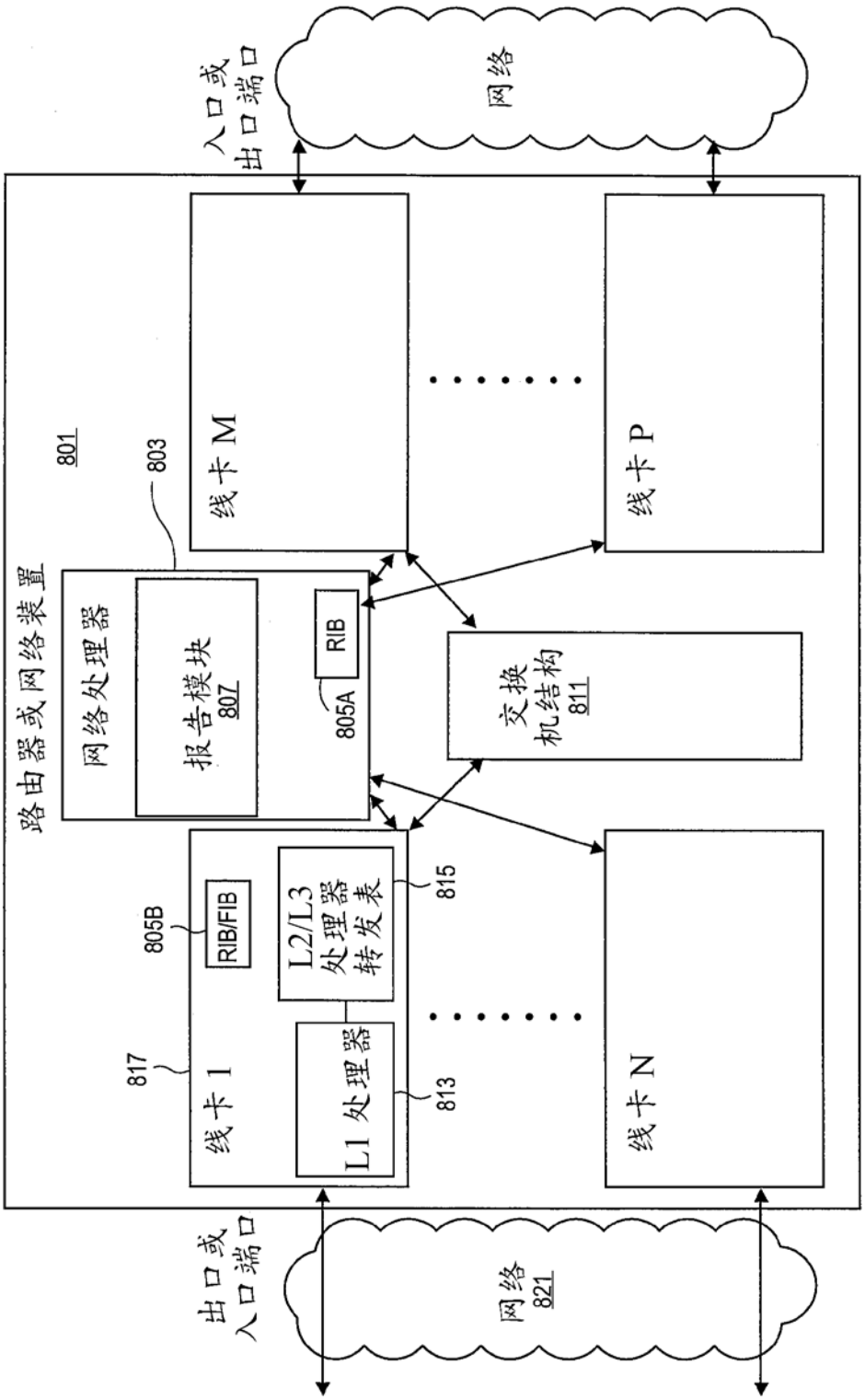

入口或出口端口
网络
801
803
线卡 M
线卡 P
路由器或网络装置
网络处理器
报告模块 807
RIB
805A
交换机结构 811
805B
RIB/FIB
L2/L3 处理器转发表
815
817
线卡 1
L1 处理器
813
线卡 N
出口或入口端口
网络 821

图 8

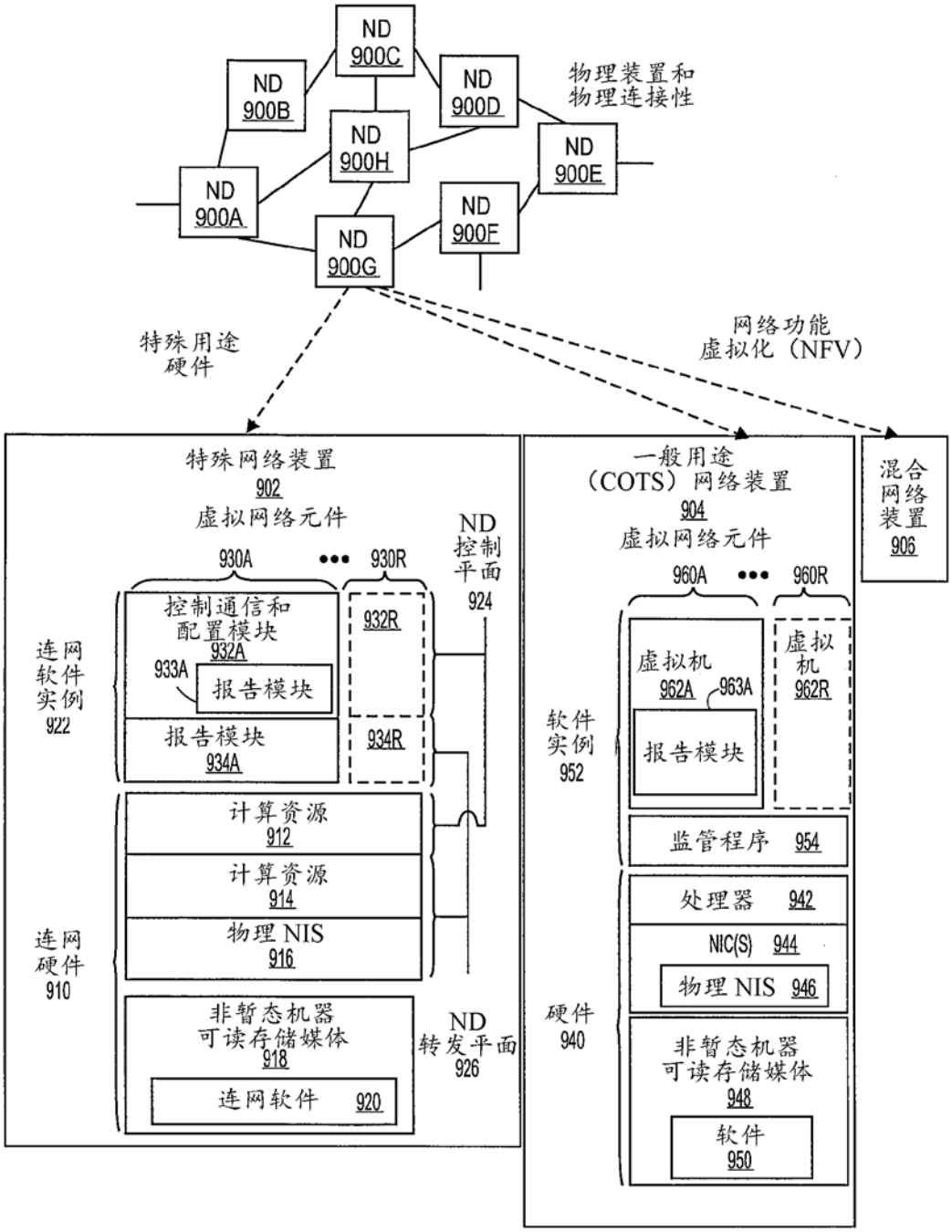

ND 900C
ND 900B
ND 900D
物理装置和物理连接性
ND 900H
ND 900E
ND 900A
ND 900F
ND 900G
特殊用途硬件
网络功能虚拟化（NFV）
特殊网络装置 902
虚拟网络元件
930A
930R
ND 控制平面 924
控制通信和配置模块 932A
933A
报告模块
报告模块 934A
932R
934R
连网软件实例 922
计算资源 912
计算资源 914
物理 NIS 916
连网硬件 910
非暂态机器可读存储媒体 918
连网软件 920
ND 转发平面 926
一般用途（COTS）网络装置 904
虚拟网络元件
960A
960R
虚拟机 962A
963A
报告模块
虚拟机 962R
软件实例 952
监管程序 954
处理器 942
NIC(S) 944
物理 NIS 946
硬件 940
非暂态机器可读存储媒体 948
软件 950
混合网络装置 906

图 9A

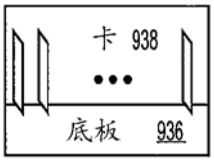

卡 938
底板 936

图 9B

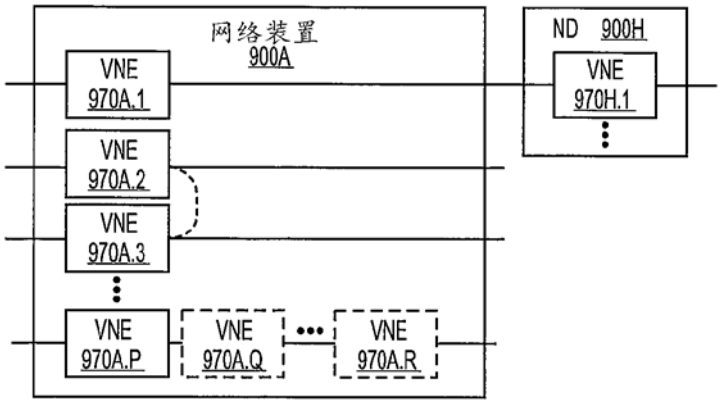

网络装置
900A
VNE
970A.1
VNE
970A.2
VNE
970A.3
VNE
970A.P
VNE
970A.Q
VNE
970A.R
ND 900H
VNE
970H.1

图 9C

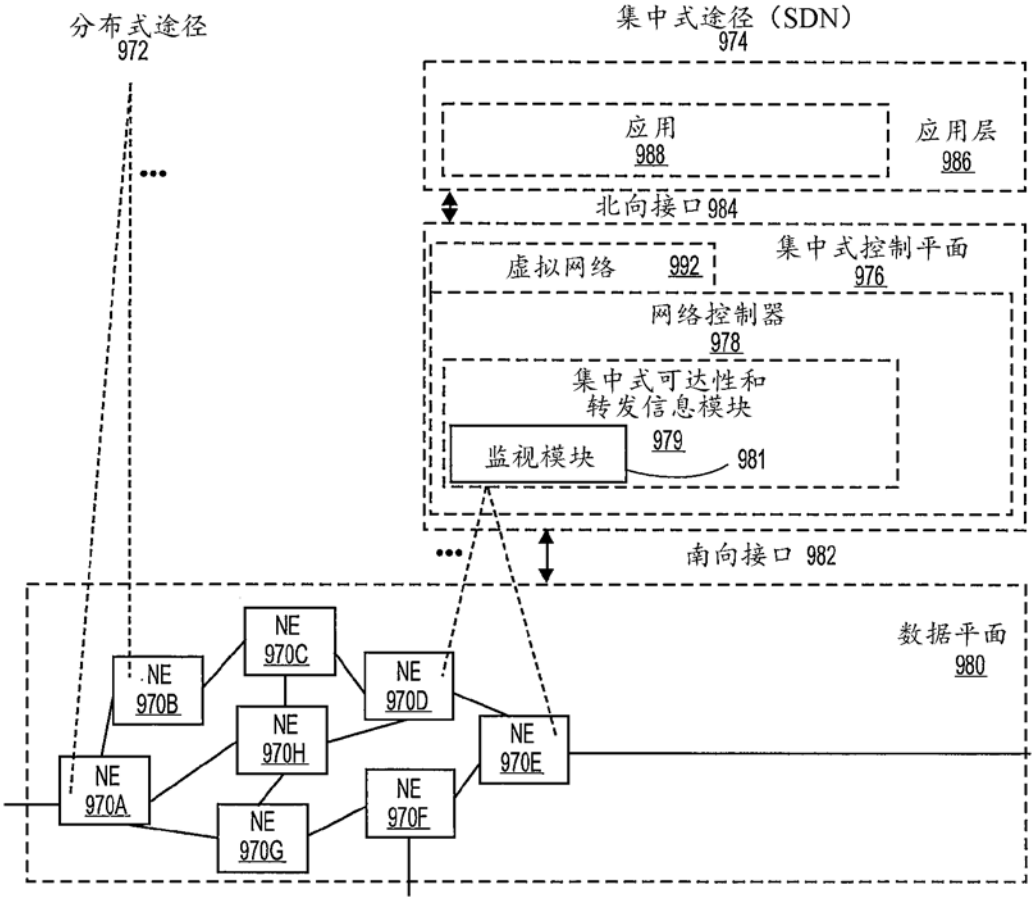

分布式途径
972
集中式途径（SDN）
974
应用
988
应用层
986
北向接口 984
虚拟网络 992
集中式控制平面
976
网络控制器
978
集中式可达性和
转发信息模块
监视模块
979
981
南向接口 982
数据平面
980
NE
970A
NE
970B
NE
970C
NE
970D
NE
970E
NE
970F
NE
970G
NE
970H

图 9D

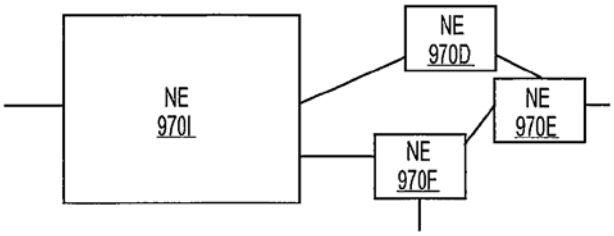

NE
970I
NE
970D
NE
970E
NE
970F

图 9E

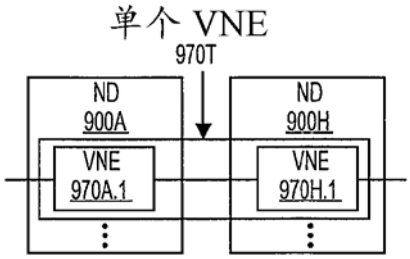

单个 VNE
970T
ND
900A
VNE
970A.1
ND
900H
VNE
970H.1

图 9F

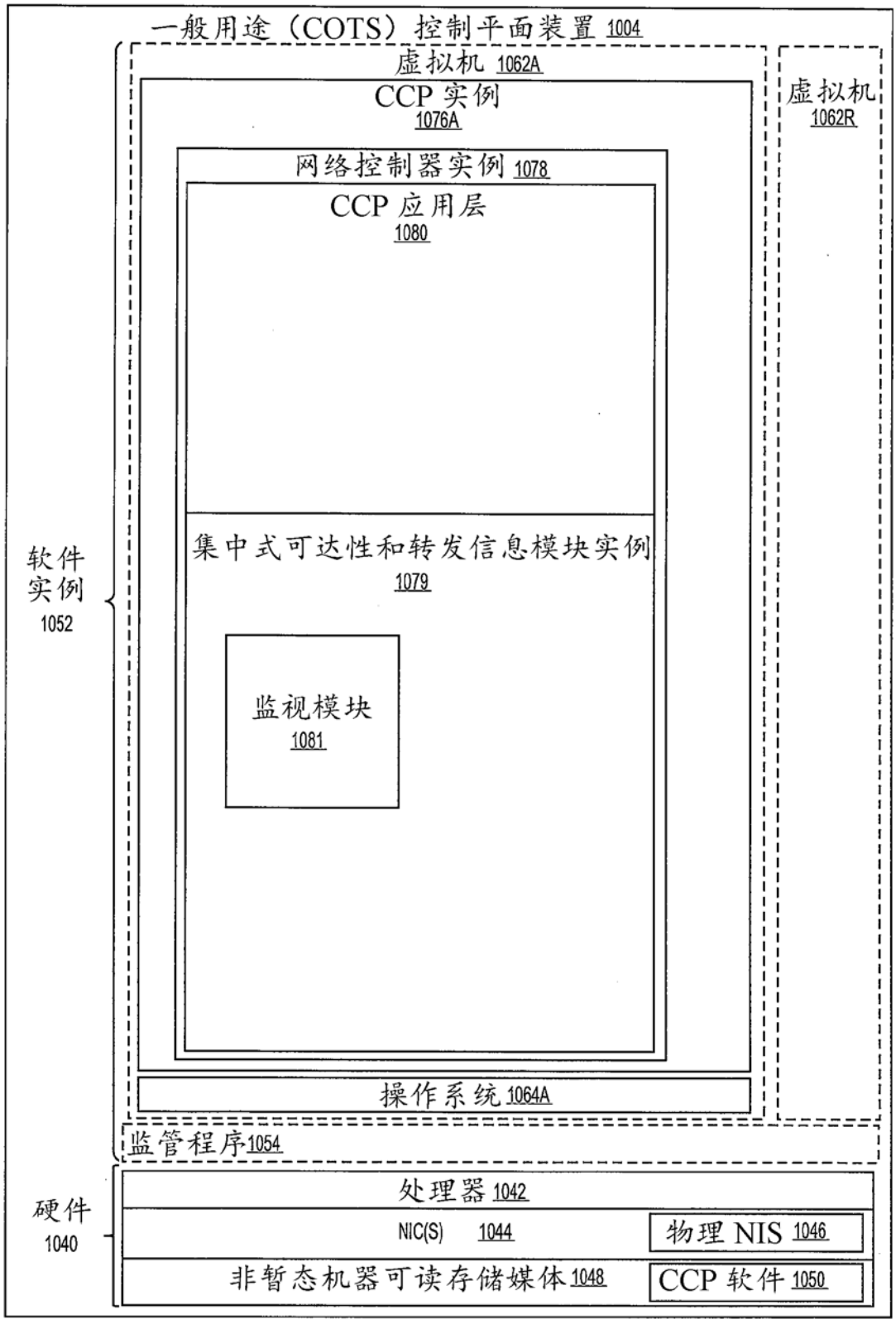

一般用途（COTS）控制平面装置 1004
虚拟机 1062A
CCP 实例
1076A
网络控制器实例 1078
CCP 应用层
1080
集中式可达性和转发信息模块实例
1079
监视模块
1081
虚拟机
1062R
软件
实例
1052
操作系统 1064A
监管程序1054
硬件
1040
处理器 1042
NIC(S) 1044
物理 NIS 1046
非暂态机器可读存储媒体 1048
CCP 软件 1050

图 10