(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 106063195 B
(45)授权公告日 2019.05.28

(21)申请号 201580010033.7
(22)申请日 2015.02.19
(65)同一申请的已公布的文献号
申请公布号 CN 106063195 A
(43)申请公布日 2016.10.26
(30)优先权数据
14/188,027 2014.02.24 US
(85)PCT国际申请进入国家阶段日
2016.08.23
(86)PCT国际申请的申请数据
PCT/US2015/016658 2015.02.19
(87)PCT国际申请的公布数据
WO2015/127107 EN 2015.08.27
(73)专利权人 第三雷沃通讯有限责任公司
地址 美国科罗拉多州
(72)发明人 威廉姆·托马斯·塞拉
詹姆斯·麦克尔·塞拉
(74)专利代理机构 中科专利商标代理有限责任公司 11021
代理人 倪斌
(51)Int.Cl.
H04L 12/28(2006.01)
(56)对比文件
WO 2012081202 A1,2012.06.21,
WO 2012081202 A1,2012.06.21,
US 2013194914 A1,2013.08.01,
US 2011261723 A1,2011.10.27,
CN 103026669 A,2013.04.03,
审查员 李奇

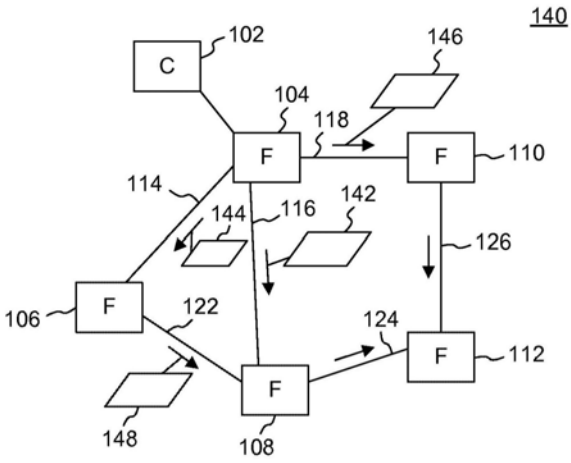
权利要求书4页 说明书11页 附图7页

(54)发明名称

具有单独控制设备和转发设备的网络中的控制设备发现

(57)摘要

软件定义网络(SDN)包括多个转发设备以及与转发设备分开布置的路由控制设备。路由控制设备建立到网络转发设备的路径以及从网络转发设备起始的路径。使用这些路径,转发设备向路由控制设备发送反映网络拓扑结构的信息。本文中所公开的实施例实现了自动发现网络的拓扑结构以及到路由控制设备的路径和从路由控制设备起始的路径。



CN 106063195 B

1.一种用于发现从多个转发设备到路由控制设备的路线的方法,包括:

(a)在所述多个转发设备中的转发设备的数据链路层端口上接收源自于所述路由控制设备的控制分组,其中所述数据链路层端口经由链路将所述转发设备连接到另一转发设备,并且所述控制分组包括指示将分组从所述另一转发设备路由到所述路由控制设备的估计时间的性能度量;

(b)至少部分地基于(i)接收到所述控制分组的数据链路层端口和(ii)所述控制分组中的所述性能度量,识别所述转发设备上的由MAC地址标识的哪个数据链路层端口将数据转发到达所述路由控制设备;

(c)响应于在所述转发设备处接收到新数据流,在(b)中识别的数据链路层端口上将消息发送到所述路由控制设备,所述消息指示所述路由控制设备确定通过所述多个转发设备到目的地的路径并将所述多个转发设备的子集配置为沿所确定的路径转发所述新数据流;以及

(d)根据所述路由控制设备的配置命令,转发所述新数据流。

2.根据权利要求1所述的方法,其中,所述发送(c)包括:通过所述路由控制设备转发所述新数据流的初始分组。

3.根据权利要求1所述的方法,还包括:

(e)在所述转发设备的与在(a)中接收到分组的数据链路层端口不同的其余数据链路层端口上,将控制分组从所述转发设备转发到另一转发设备。

4.根据权利要求3所述的方法,还包括:

(f)在所述转发设备处,将标签发送到所述多个转发设备中的另一转发设备,其中,当所述标签附着到在所述转发设备处接收到的分组时,所述标签指示所述转发设备在(b)中识别的数据链路层端口上朝所述路由控制设备转发所述分组。

5.根据权利要求3所述的方法,其中,所述控制分组包括序列号,并且所述方法还包括:

(f)确定所述转发设备是否之前已经转发了来自所述路由控制设备的另一控制分组,所述另一控制分组具有所述序列号,并指示比所述控制分组中的估计时间少的将分组从所述转发设备路由到所述路由控制设备的另一估计时间,以及

其中,当在(f)中确定所述转发设备之前没有转发所述另一控制分组时,发生转发步骤(e)。

6.根据权利要求1所述的方法,还包括:

(e)在(b)中识别的数据链路层端口上将消息发送到所述路由控制设备,所述消息识别所述转发设备以及在所述多个转发设备中所述转发设备所连接到的其他转发设备,

其中,所述路由控制设备使用所述消息来建立表示网络拓扑结构的数据库,并使用所述数据库来确定路径。

7.根据权利要求1所述的方法,其中,所述控制分组包括所述路由控制设备的标识符,并且所述方法还包括:

(e)在所述转发设备的数据链路层端口上接收包括另一路由控制设备的标识符在内的另一控制分组;

(f)至少部分地基于接收到所述另一控制分组的数据链路层端口,识别所述转发设备上的哪个数据链路层端口将数据转发到达所述另一路由控制设备;以及

(g)响应于在所述转发设备处接收到新数据流,确定所述路由控制设备和所述另一路由控制设备中的哪一个发送消息,所述消息指示所述路径配置为通过所述多个转发设备。

8.一种用于发现到路由控制设备的路线的转发设备,包括:

交换发现模块,被配置为:(i)在转发设备的数据链路层端口上接收源自于所述路由控制设备的控制分组,其中所述数据链路层端口经由链路将所述转发设备连接到另一转发设备,并且所述控制分组包括指示将分组从所述另一转发设备路由到所述路由控制设备的估计时间的性能度量;以及(ii)至少部分地基于(i)接收到所述控制分组的数据链路层端口和(ii)所述控制分组中的所述性能度量来识别所述转发设备上的由MAC地址标识的哪个数据链路层端口将数据转发到达所述路由控制设备;

路径请求模块,被配置为响应于在所述转发设备处接收到新数据流,在所识别的数据链路层端口上将消息发送到所述路由控制设备,所述消息指示所述路由控制设备进行以下操作:(i)确定通过多个转发设备到目的地的路径,以及(ii)将所述多个转发设备的子集配置为沿所确定的路径转发所述新数据流;以及

路由表,位于所述转发设备内,用于使得所述转发设备能够根据所述路由控制设备的配置命令转发所述新数据流。

9.根据权利要求8所述的设备,其中,所述路径请求模块还被配置为:通过所述路由控制设备转发所述新数据流的初始分组。

10.根据权利要求8所述的设备,其中,所述交换发现模块被配置为:在所述转发设备的与接收到控制分组的数据链路层端口不同的其余数据链路层端口上,将控制分组从所述转发设备转发到另一转发设备。

11.根据权利要求10所述的设备,其中,所述交换发现模块被配置为:将标签发送到所述多个转发设备中的另一转发设备,其中,当所述标签附着到在所述转发设备处接收到的分组时,所述标签指示所述转发设备在由所述交换发现模块识别的数据链路层端口上朝所述路由控制设备转发所述分组。

12.根据权利要求10所述的设备,其中,控制分组包括序列号,并且所述转发设备还包括:性能度量模块,被配置为:

确定所述转发设备是否之前已经转发了来自所述路由控制设备的另一控制分组,所述另一控制分组具有所述序列号,并指示比所述控制分组中的性能度量更好的将分组从所述转发设备路由到所述路由控制设备的另一性能度量,以及

其中,当所述路径请求模块确定之前没有转发所述另一控制分组时,发生所述转发。

13.根据权利要求10所述的设备,其中,所述交换发现模块还被配置为:

在所识别的数据链路层端口上将消息发送到所述路由控制设备,所述消息识别所述转发设备以及在所述多个转发设备中所述转发设备所连接到的其他转发设备,

其中,所述路由控制设备被配置为使用所述消息来建立表示网络拓扑结构的数据库,并使用所述数据库来确定路径。

14.根据权利要求10所述的设备,其中,所述控制分组包括所述路由控制设备的标识符,并且所述交换发现模块还被配置为:

在所述转发设备的数据链路层端口上接收包括另一路由控制设备的标识符在内的另一控制分组;

至少部分地基于接收到所述另一控制分组的数据链路层端口，识别所述转发设备上的哪个数据链路层端口将数据转发到达所述另一路由控制设备；以及

响应于在所述转发设备处接收到新数据流，确定所述路由控制设备和所述另一路由控制设备中的哪一个发送消息，所述消息指示所述路径配置为通过所述多个转发设备。

15.一种有形地实现指令程序的程序存储设备，所述指令能够由至少一个机器执行以执行用于发现从多个转发设备到路由控制设备的路线的方法，所述方法包括：

(a)在所述多个转发设备中的转发设备的数据链路层端口上接收源自于所述路由控制设备的控制分组，其中所述数据链路层端口经由链路将所述转发设备连接到另一转发设备，并且所述控制分组包括指示将分组从所述另一转发设备路由到所述路由控制设备的估计时间的性能度量；

(b)至少部分地基于(i)接收到所述控制分组的数据链路层端口和(ii)所述控制分组中的所述性能度量，识别所述转发设备上的由MAC地址标识的哪个数据链路层端口将数据转发到达所述路由控制设备；

(c)响应于在所述转发设备处接收到新数据流，在(b)中识别的数据链路层端口上将消息发送到所述路由控制设备，所述消息指示所述路由控制设备确定通过所述多个转发设备到目的地的路径并将所述多个转发设备的子集配置为沿所确定的路径转发所述新数据流；以及

(d)根据所述路由控制设备的配置命令，转发所述新数据流。

16.根据权利要求15所述的程序存储设备，其中，所述发送(c) 包括：通过所述路由控制设备转发所述新数据流的初始分组。

17.根据权利要求15所述的程序存储设备，所述方法还包括：

(e)在所述转发设备的与在(a)中接收到分组的数据链路层端口不同的其余数据链路层端口上，将控制分组从所述转发设备转发到另一转发设备。

18.根据权利要求17所述的程序存储设备，所述方法还包括：

(f)在所述转发设备处，将标签发送到所述多个转发设备中的另一转发设备，其中，当所述标签附着到在所述转发设备接收到的分组时，所述标签指示所述转发设备在(b)中识别的数据链路层端口上朝所述路由控制设备转发所述分组。

19.根据权利要求17所述的程序存储设备，其中，所述控制分组包括序列号，并且所述方法还包括：

(f)确定所述转发设备是否之前已经转发了来自所述路由控制设备的另一控制分组，所述另一控制分组具有所述序列号，并指示比所述控制分组中的估计时间少的将分组从所述转发设备路由到所述路由控制设备的另一估计时间，以及

其中，当在(f)中确定所述转发设备之前没有转发所述另一控制分组时，发生转发步骤(e)。

20.根据权利要求15所述的程序存储设备，还包括：

(e)在(b)中识别的数据链路层端口上将消息发送到所述路由控制设备，所述消息识别所述转发设备以及在所述多个转发设备中所述转发设备所连接到的其他转发设备，

其中，所述路由控制设备使用所述消息来建立表示网络拓扑结构的数据库，并使用所述数据库来确定路径。

21.根据权利要求15所述的程序存储设备,其中,所述控制分组包括所述路由控制设备的标识符,并且所述方法还包括:

(e)在所述转发设备的数据链路层端口上接收包括另一路由控制设备的标识符在内的另一控制分组;

(f)至少部分地基于接收到所述另一控制分组的数据链路层端口,识别所述转发设备上的哪个数据链路层端口将数据转发到达所述另一路由控制设备;以及

(g)响应于在所述转发设备处接收到新数据流,确定所述路由控制设备和所述另一路由控制设备中的哪一个发送消息,所述消息指示所述路径被配置为通过所述多个转发设备。

具有单独控制设备和转发设备的网络中的控制设备发现

技术领域

[0001] 本领域总体上涉及网络路由。

背景技术

[0002] 通信网络可以例如提供允许在两个地理上的远程位置之间传送数据的网络连接。网络连接可以跨越连接诸如路由器的通信设备的多个链路。网络可以根据链路通过通信设备如何互连而具有不同拓扑结构。在给定特定网络拓扑的情况下，在源和目的地之间可以存在多条路线。根据当前容量和使用情况，一些路线可能比其他路线更可取。

[0003] 传统的路由算法依赖于每个路由器从与其邻近的链路和设备获得的本地信息来路由数据。路由器将这种信息保持在路由表中。并且基于传入分组的目的地址，路由器使用其路由表将分组转发至特定邻近设备。为了开发路由表，每个路由器使用如边界网关协议(BGP)的协议与本地邻近路由器交换路由和可达性信息。以这种方式，每个路由器既转发分组，也进行控制功能以更新其自身的路由表。

[0004] 虽然使用本地信息在一些情况下是所希望的，但是其可能不总是高效地路由数据。为了更高效地路由数据，被称为软件定义网络(SDN)的另一种技术将控制功能和转发功能分为单独的设备。控制设备使用网络拓扑的全局知识，以针对各个数据流确定通过转发设备的网络的路径。以这种方式，路由控制设备可以例如通过网络建立使延迟最小化或使带宽最大化的路径。

发明内容

[0005] 在实施例中，一种计算机实现的方法发现从网络转发设备至路由控制设备的路线。该发现方法包括：在转发设备的端口上接收控制分组。然后，所述方法至少部分地基于接收到控制分组的端口，识别转发设备上的哪个端口将数据转发到达路由控制设备。响应于在转发设备接收到新数据流，所述方法在所识别的端口上将命令发送到路由控制设备。所述命令指示路由控制设备确定通过转发设备到目的地的路径并配置转发设备以沿所确定的路径转发所述新数据流。然后，所述方法根据路由控制设备的配置命令，转发所述新数据流。

[0006] 还公开了系统和计算机程序产品实施例。

[0007] 以下参照附图详细地描述本发明的其他实施例、特征和优点以及各种实施例的结构和操作。

附图说明

[0008] 并入本文中且形成说明书的一部分的附图示出了本公开，并且与描述一起进一步用于解释本公开的原理且使相关领域技术人员能够制造和使用本公开。

[0009] 图1A是示出了路由控制设备将控制分组发送到转发设备的示图。

[0010] 图1B是示出了转发设备在整个网络中分发(flood)控制分组的示图。

[0011] 图1C是示出了控制设备如何收集反映网络拓扑的信息的示图。

[0012] 图2是示出了用于找到从控制设备至转发设备的路径的方法的流程图。

[0013] 图3是示出了用于建立从每个转发设备至控制设备的路径的方法的流程图。

[0014] 图4A是示出了使用从转发设备至控制设备的路径的网络连接的示图。

[0015] 图4B是示出了路由控制设备建立快速路径的示图。

[0016] 图5是示出了具有多个控制设备的网络的示图。

[0017] 图6是示出了控制设备和转发设备的模块的示图。

[0018] 要素首次出现的附图通常由相应附图标记中的最左边的一个或多个数字来指示。在附图中，相似的附图标记可以指示相同或功能相似的要素。

具体实施方式

[0019] 如上所述，SDN路由技术使用网络拓扑的全局知识以高效地路由分组。这些技术使用彼此分离的路由控制设备和转发设备。当转发设备通过网络接收到新的数据流时，转发设备联系路由控制设备以确定用于该数据流的新路径。路由控制设备确定通过网络的转发设备的新路径，并相应地配置转发设备。

[0020] 为了运行，路由控制设备和转发设备需要配置有网络信息。具体地，为了请求通过互连的转发设备创建新路径，转发设备需要知晓如何将数据路由到路由控制设备。并且，路由控制设备需要知晓网络拓扑以确定通过网络的路径。手动地配置这种信息会是耗时且容易出错的。

[0021] 这里公开的实施例使得拓扑信息能够被自动地发现。在实施例中，路由控制设备连接到转发设备并将控制分组发送到转发设备。控制分组在地址字段中包含控制设备的地址，作为其源地址。控制设备还将序列号附着到每个控制分组。

[0022] 转发设备附着性能信息，所述性能信息例如指示网络需要多长时间将信息从转发设备发送到路由控制设备。然后，转发设备在该转发设备的连接到其他转发设备的其他端口上广播或分发(flood)控制分组。并且，如果其他转发设备之前没有转播具有更好性能信息的分组，则这些转发设备更新性能信息并转播该分组。

[0023] 当转发设备接收到控制分组时，它们记得在哪个端口控制分组接收到具有最佳性能信息的分组。转发设备知晓的该端口在至路由控制设备的最快路线上。并且，每个转发设备可以请求在至路由控制设备的该路线上建立标签交换路径。

[0024] 利用所建立的路径，转发设备将关于其邻近网络拓扑和网络条件的信息发送到控制设备。使用来自所有转发设备的这种信息，控制设备可以创建全局网络拓扑和网络条件的数据库。该拓扑数据库使得控制设备能够在网络中建立源转发设备和目的转发设备之间的路径。

[0025] 一旦向控制器建立了该路径，则控制器可以在之后的任何时间，使用可以利用更完整的拓扑和性能信息的另一种方法来确定用于转发设备和控制器之间的通信的更好路径。如果确定了该更好路径，则可以通过控制器重新用信号通知特定的新路径并且随后消除现有路径，来替换现有路径。

[0026] 以下具体实施方式分为五个部分。第一部分参照图1A至图1C描述了网络中的控制设备将控制分组发送到每个转发设备，并从转发设备收集拓扑和网络条件信息。第二部分

参照图2至图3描述了在控制设备和每个转发设备之间建立双向连接。第三部分参照图4A至图4B描述了使用控制设备建立并传送数据而不需要边缘路由器缓冲大量分组。第四部分参照图5描述了使用网络中的多个控制设备。最后的第五部分参照图6描述了控制设备系统及其模块以及转发设备系统及其模块。

[0027] 控制设备和控制分组

[0028] 图1A是示出了通信网络的示图100。通信网络可以是局域网(LAN)、城域网(MAN)或广域网(WAN)。其可以利用任意点到点或多点到多点的网络协议。所使用的网络接入协议可以包括例如:多协议标签交换(MPLS)、以太网、异步传输模式(ATM)、高级数据链路控制(HDLC)或分组中继。

[0029] 通信网络包括通过链路互连的多个转发设备,例如转发设备104、106、108、110和112。转发设备是转发分组的设备,包括位于数据链路层(OSI层2)和网络层(OSI层3)的设备。

[0030] 通信网络还包括路由控制设备102。路由控制设备102可以连接到至少一个转发设备,例如转发设备104。路由控制设备102可以在地理上远离网络中的其他转发设备。

[0031] 在示例中,用户可以将数据从源转发设备(例如转发设备106)发送到目的转发设备(例如110)。数据可以是被划分成多个分组的数据流,并且每个分组可以指定转发设备110或另一下游设备作为其目的地。

[0032] 路由控制设备102提供用于建立网络连接的智能路由。为此,控制设备102需要网络链路和设备的拓扑结构和条件的知识。路由控制设备102直接连接转发设备104。为了控制设备获得网络的知识,控制设备102将控制分组120发送到转发设备104。

[0033] 控制分组120可以包括表明该分组是控制分组的指示、路由控制设备102的标识符(例如其媒体访问控制地址)和该分组的序列号。在实施例中,控制设备102可以按照均匀时间间隔发送新的控制分组。每次控制设备102发送新分组时,控制设备102可以使序列号递增,从而发送具有不同编号的新的控制分组。

[0034] 在说明性示例中,控制分组120可以包括以下信息:

[0035] 类型标记:C(用于控制分组)

[0036] 控制设备的MAC地址:01:23:45:67:89:ab

[0037] 控制分组的序列号:1

[0038] 一旦路由控制设备102将控制分组120发送到相邻转发设备104,则网络转发设备如图1B所示在整个网络中分发该分组。

[0039] 图1B是示出了网络转发设备在整个网络中分发控制分组的示图。在接收到控制分组120之后,转发设备104可以在除了接收到控制分组的端口之外的所有端口上送出该控制分组。

[0040] 如上所述,当路由控制设备102产生控制分组120时,控制设备可以将其地址和序列号包括在控制分组中。在转发设备104接收到控制分组时,并且在转发控制分组之前,转发设备104可以进行两次修改。首先,转发设备104可以将其自身的标识符(例如其MAC地址)添加到控制分组。该地址被添加到控制分组中的地址字段,以跟踪控制分组所经过的所有转发设备。其次,转发设备104可以修改控制分组中的性能度量。性能度量可以包括例如:将控制分组从控制设备102发送到转发设备104的延迟。可以例如由转发设备104使用链路层

发现协议(LLDP)分组交换来收集该延迟信息。该延迟可以被确定为移动平均或以其他方式被平滑化,以防止由于网络对通信量的变化作出反应而导致的不稳定变化。

[0041] 利用这两次修改,转发设备104将该控制分组作为控制分组142、144和146在它的其他端口上进行转发。此时,全部三个控制分组142、144和146可以包含相同信息。继续上述示例并假设转发设备104的MAC地址是02:02:02:02:02:bc,并且控制设备102和转发设备104之间的延迟是5纳秒(ns),则控制分组142、144和146可以包括以下信息:

[0042] 类型标记:C(用于控制分组)

[0043] 控制设备的MAC地址:01:23:45:67:89:ab

[0044] 控制分组的序列号:1

[0045] 中间转发设备的地址:02:02:02:02:02:bc

[0046] 性能度量:5ns

[0047] 转发设备104在链路114上将控制分组144转发到转发设备106;在链路116上将控制分组142转发到转发设备108;以及在链路118上将控制分组146转发到转发设备110。

[0048] 一旦转发设备106接收到控制分组144,则转发设备106将其自身的地址添加到地址字段并更新控制分组的性能度量以产生控制分组148。例如,转发设备106可以添加将数据从转发设备106发送到转发设备104的延迟。继续上述示例并假设转发设备106的MAC地址是03:03:03:03:03:cd,并且控制器102和转发设备104之间的延迟是2ns,则控制分组148可以包括以下信息:

[0049] 类型标记:C(用于控制分组)

[0050] 控制设备的MAC地址:01:23:45:67:89:ab

[0051] 控制分组的序列号:1

[0052] 中间转发设备的地址:02:02:02:02:02:bc,03:03:03:03:03:cd

[0053] 性能度量:7ns

[0054] 转发设备106在链路122上将分组148发送到转发设备108。

[0055] 转发设备108接收控制分组142和144。使用这两个控制分组中的每一个中的延迟信息,转发设备108可以选择建立至控制设备102的具有较小延迟的路径。使用控制分组中的地址序列字段,转发设备108可以知晓为了到达控制设备,其需要与哪个邻近转发设备进行通信。

[0056] 例如,转发设备108可以知晓在链路120上将数据发送到转发设备106的时间是3ns,以及在链路116上将数据发送到转发设备104的时间是9ns。将上述度量添加到在分组148和142中接收到的性能度量,转发设备108可以确定经由转发设备106发送到控制设备102需要10ns,而经由转发设备104的发送需要14ns。由于该原因,转发设备108可以选择经由转发设备106(而不经由设备104)将数据路由到控制设备102,从而采用更快的路径。

[0057] 此外,转发设备108可以再转发控制分组142和148。如果转发设备108在接收到表示较慢路径的分组(在这种情况下是分组142)之前接收到表示较快路径的分组(在这种情况下是分组148),则转发设备108可以不转发较慢的分组。这是因为转发设备108知道前往控制设备102的任何数据应当采用通过转发设备106的更快路径,而不管其源自于转发设备108还是另一转发设备(例如设备112)。

[0058] 在确定了至控制设备102的最佳路径是通过转发设备106的路径之后,转发设备

108可以与转发设备106建立双向路径。在实施例中，为了建立该双向路径，转发设备108在其路由表(也可以称为转发表)中设置规则：为了到达控制设备，转发设备108在特定端口(接收到控制分组144的端口)上并利用特定标签来发送分组。与转发设备106传送该信息，因而这两个转发设备遵循相同的规则。因此，当转发设备106利用所设置的标签并在所设置的端口上接收到来自转发设备108的分组时，其知晓该分组是针对控制设备102的。类似地，转发设备106知晓通过转发设备104转发针对控制设备102的通信量。所以，转发设备106建立与转发设备104的标签，以将数据路由到控制设备102。还可以将标签建立为以类似方式通过转发设备106将数据从控制设备102路由到转发设备104。

[0059] 以这种方式，通过分析并从控制设备再转发控制分组，每个转发设备可以发现如何与控制设备通信并建立至控制设备的标签交换路径。如以下参照图4A至图4B所述，当转发设备接收到新数据流时，转发设备可以将来自新数据流的分组转发到控制设备，直到控制设备在网络中建立用于分组的路径。

[0060] 除了使用标签交换路径将数据转发到控制设备之外，实施例还可以使用标签交换路径来帮助控制设备发现网络的拓扑信息，如图1C所示。

[0061] 图1C是示出了控制设备102收集反映网络拓扑的信息的示图170。

[0062] 每个转发设备将信息发送到路由控制设备，该图示出了路由控制设备连接到哪些转发设备以及周围链路和转发设备的条件如何。周围链路和转发设备的条件可以包括例如通信量和拥塞信息、延迟、分组丢失率、抖动(jitter)等。可以例如使用链路层发现协议(LLDP)来收集该信息。在实施例中，转发设备定期地发送该信息。并且，转发设备使用在它们与控制设备之间建立的路径来发送该信息。

[0063] 在图1C所示的示例中，建立了转发设备108至控制设备102的路径。该路径包括转发设备108、106和104。转发设备108将包括转发设备108的周围拓扑在内的信息发送到控制设备102。在示例中，转发设备108以分组172将该信息发送到路由控制设备102。

[0064] 使用从包括转发设备108在内的所有转发设备接收到的拓扑信息，路由控制设备102创建数据库174，数据库174反映转发设备的网络的拓扑结构。

[0065] 建立至控制设备的路径和从控制设备起始的路径

[0066] 图2是示出了用于找到从控制设备至转发设备的路径的方法200的流程图。

[0067] 在步骤202，控制设备物理地连接到转发设备的网络中的转发设备。在步骤204，控制设备将例如链路层发现协议(LLDP)分组的控制(或发现)分组发送到其所连接到的转发设备。

[0068] 接收到该分组的转发设备在步骤206验证该分组是否源自于控制设备。在示例中，转发设备可以通过例如查看是否出现了特定标记来验证该分组来自于控制设备。

[0069] 控制分组包括性能度量，所述性能度量示出网络链路或网络设备的条件。例如，所述度量可以示出链路延迟。因此，在该示例中，较低的度量示出较低的延迟，由此导致更好的性能。度量还可以是诸如延迟、分组丢失、抖动等的多个条件的函数。

[0070] 接收到控制分组的每个转发设备更新性能度量。例如，更新后的度量示出控制分组从前一或上游转发设备到达转发设备的附加延迟。在更新度量之后，转发设备在该转发设备的除了接收到控制分组的端口之外的所有端口上分发控制分组。

[0071] 转发设备可以再次接收其之前接收到的相同控制分组。这可以发生，因为控制分

组在整个网络中分发。转发设备可以通过被控制设备附着到每个控制分组的唯一序列号来识别重复控制分组。

[0072] 在步骤208，转发设备查看分组的性能度量是否比从控制设备接收到的先前度量更好。如果度量在由转发设备接收到的所有控制分组中是最佳性能度量，则在步骤210，转发设备将分组的传入端口存储为到达控制设备的最佳方式。

[0073] 在步骤212，转发设备检查序列号以确定之前是否转发了该控制分组，从而避免控制分组在网络中无限循环。由于控制分组将唯一序列号添加到其产生的每个控制分组，因此转发设备可以使用控制分组的序列号来确定之前是否转发了该控制分组。或者，如果例如转发设备之前已经转发了具有较大序列号的更近期的控制分组，则转发设备可以使用序列号确定当前控制分组是否过期。可以使用多个其他机制(例如控制分组的至少一部分的散列或校验和)来避免分组在网络中无限循环。

[0074] 在步骤214，转发设备确定是否之前没有转发分组以及该分组是否包括示出了最佳性能的度量。如果控制分组未示出最佳性能度量，或者它之前被转发设备转发过，则转发设备丢弃该分组，并且处理结束。

[0075] 如果控制分组包含最佳性能度量并且之前未被转发过，则处理200在步骤216建立路径。以下在图3的描述中详细地描述步骤216。

[0076] 接着，在步骤218，转发设备更新性能度量，以反映连接上游转发设备和当前转发设备的链路的性能、以及这两个转发设备自身的性能。接着，在步骤220，接收转发设备在它的除了接收到控制分组的端口之外的所有端口上分发分组。以这种方式，更新后的控制分组被发送到其他转发设备。

[0077] 其他转发设备重复从步骤206开始的处理，并以相同方式处理分组。

[0078] 图3是示出了用于建立转发设备104和106之间的路径的方法216的流程图。作为示例，如果控制分组已经从转发设备104行进到转发设备106，则转发设备104是106的上游转发设备。

[0079] 以上继续参照图1A至图1B以及图2，假设转发设备106从上游转发设备104接收到新的控制分组并且该新的控制分组具有到目前为止的最佳性能特征。为此，在图2中的步骤216建立路径。

[0080] 在步骤302，转发设备106将针对至控制设备的路径的请求发送到下一上游转发设备。在步骤306，转发设备104确定用于建立路径的标签。在步骤308，转发设备104将该标签发送到转发设备106。在步骤304，转发设备106在其路由表中相应地建立该标签规则。

[0081] 在步骤310，转发设备104请求至控制设备的路径。为了创建路径，转发设备104可以向另一上游转发设备请求标签，从而重复步骤302、306、308和310。在步骤312，转发设备104在其路由表中建立标签规则。

[0082] 例如，在步骤302接收到来自转发设备106的请求之后，转发设备104选择标签L1用于它们二者之间的路径。在将该标签发送到转发设备106之后，转发设备106将在指定到控制设备的所有数据分组上附上该标签。

[0083] 在该示例中，在通过连接到转发设备106的端口接收到具有标签L1的数据分组时，转发设备104将数据分组转发到控制设备。为了将分组转发到控制设备，转发设备104可以知晓(在步骤312建立了标签规则)在具有特定标签的特定端口上进行转发。以这种方式，建

立从转发设备到控制设备的路径。

[0084] 此外，可以以类似方式建立其他方向上的路径：从控制设备到转发设备。在这种情况下，每个转发设备可以在其路由表中建立标签。路由表中的条目可以包括用于路由数据的相关端口。

[0085] 在网络中的转发设备之间建立了这些路径之后，网络中存在从控制设备到每个转发设备以及从每个转发设备到控制设备的路径。在图3的示例中，在转发设备106、转发设备104和控制设备之间建立了路径之后，转发设备106使用该路径将数据发送到控制设备。反之亦然，控制设备使用该路径将数据发送到转发设备106。

[0086] 建立用于新数据流的快速路径

[0087] 图4A示出了说明从网络用户442寻址到服务器444的数据流的示图400。数据流包括分组402、404、406、408、410和412。当数据流的第一分组(分组402)到达转发设备106时，其被路由到控制设备102。控制设备102确定该分组属于新连接，并开始针对该连接建立不流过控制设备102的快速路径的处理。

[0088] 在示例中，在分组402之后，还发送分组404和406。转发设备106可以缓冲分组404和406，或者可以将这些分组转发到控制设备102，就像其对第一分组402做的那样。

[0089] 在控制设备接收到第一分组之后，其将该分组发送到转发设备104。转发设备104使用包括在该分组中的目的地址和在图2和图3中的步骤216所建立的路径将该分组转发到具有相应标签的相应端口。

[0090] 接着，控制设备计算从源到目的地的路径。该路径包括网络链路和路由器，但是不包括控制设备。例如，使用MPLS协议，控制设备创建用于数据流的标签交换路径(LSP)。为了确定路径，控制设备可以考虑当前或历史带宽使用情况。例如，控制设备可以考虑之前几周(或其他循环时间段)期间的网络使用情况。控制设备还可以创建用于连接的多个路径以实现多个路径上的负载均衡。

[0091] 控制设备还可以考虑延迟、抖动、分组丢失或跨越各种路径的任何其他性能度量、用户的服务水平协议或传输的数据的类型。例如，广播视频数据可能需要大量带宽，但是延迟可能相对不重要。另一方面，IP语音(VoIP)数据可能不需要如此高的带宽，但是延迟可能更加重要。对于广播视频数据，控制设备可以选择高带宽、高延迟路径，对于VoIP数据，控制设备可以选择低带宽、低延迟路径。

[0092] 在另一实施例中，控制设备可以通过特定服务器路由数据。例如，具有特定类型或指向特定目的地的数据可能需要由路线上的特定洗涤服务器(scrubbing server)进行洗涤(scrubbed)。洗涤服务器可以用于扫描恶意内容的数据，监视传入数据，或者对数据执行其他分析。在该实施例中，控制设备可以确定其通过该特定洗涤服务器或服务器组。

[0093] 在确定了快速路径之后，控制设备针对沿该路径的每个转发设备更新路由表。如果针对数据流计算出多个路径，则控制设备根据所有路径创建路由表。在一个实施例中，更新后的路由表可以指示转发设备如何转发具有源/目的地址和源/目的端口的特定组合的分组。在备选实施例中，可以利用标签识别数据流，并且更新后的路由表可以指示如何转发具有该标签的分组。

[0094] 控制设备将更新后的路由表发送到网络路由器。控制设备可以使用在步骤216针对所有转发设备建立的路径来发送路由表。图4B示出了如何配置路由器的示例。

[0095] 图4B示出了说明控制设备102如何配置网络转发设备以建立用户442和服务器444之间的路径的示图420。

[0096] 在图4B中，控制设备102确定用户442和服务器444之间的数据流遵循包括链路120、122、124和转发设备106、108、112和110在内的路径。为了配置转发设备，控制设备102使用配置命令422、424、426和428发送更新后的路由表。这些配置命令配置在将用户442连接至目的地444的快速路径上的所有转发设备。具体地，命令422指示转发设备106将数据流中的分组转发到链路120；命令424指示转发设备108将数据流中的分组转发到链路122；命令426指示转发设备112将数据流中的分组转发到链路124；以及命令428指示转发设备110将数据流中的分组转发到链路124。

[0097] 在实施例中，为了确保分组继续流过控制设备直至路径被完全建立，从出口点(转发设备104)到入口点(转发设备106)配置沿路径的路由器。首先，命令428配置转发设备110。其次，命令426配置转发设备112。第三，命令424配置转发设备108。第四，命令422配置转发设备106。

[0098] 返回参照图4A，直至路径被建立，用户442继续发送数据流的分组-分组402、404和406。如上所述，转发设备106继续将路径上的这些分组导向控制设备102。

[0099] 在接收到这些分组时，控制设备102使用之前在图2中的步骤216确定的默认路径将这些分组路由到它们的目的地。以这种方式，当网络中的路径正被建立(如图4B所示)时，分组继续被路由到它们的目的地，从而避免在边缘路由器中缓冲初始分组的需要。

[0100] 一旦路径被建立(例如，转发设备被配置有它们的新路由表)，则数据沿由控制设备建立的快速路径流动。此时，数据可以以通过转发设备的较大速率和较低的端到端延迟而流动，因为其不再通过控制设备。

[0101] 多个控制设备

[0102] 图5是示出了具有多个控制设备的网络的示图。在图5所示的示例实施例中，控制设备102连接到转发设备104，控制设备504连接到转发设备110，控制设备506连接到转发设备506。

[0103] 在示例实施例中，每个控制设备102、504或506定期地或按不同时间间隔将控制分组发送到与其直接连接的转发设备。并且按照如图2示出的相同处理，每个控制设备建立到网络中的每个转发设备的双向路径。

[0104] 在图5的示例实施例中，每个转发设备将表明其邻近转发设备的拓扑结构的信息和网络条件信息发送到网络中的每个控制设备。在实施例中，转发设备使用在处理200的步骤216所建立的双向路径来发送该信息。在实施例中，转发设备定期地或按不同时间间隔将该信息发送到控制设备。

[0105] 出于冗余目的，单个物理网络可以在具有多个拓扑数据库的多个控制设备上被镜像。这些控制设备中的每个控制设备具有在它们自身与网络中的每个路由器之间建立的路径。在至一个控制设备的默认路径失败的情况下，业务可以选择至不同控制设备的路径。

[0106] 在一个示例实施例中，每个转发设备每次挑选一个控制设备，以转发数据并建立至目的地的快速路径。在该实施例中，转发设备可以基于选择算法挑选这些控制设备中的任何一个控制设备。选择算法可以是例如轮询调度算法(round robin algorithm)。转发设备还可以在用于选择控制设备的算法中考虑控制设备每次建立新连接的数目或网络负载。

另外，转发设备可以向控制器询问建立至目的地的快速路径的最低延迟。轮询调度方法可以与最低延迟方法结合使用。例如，或许通过将度量四舍五入到例如最接近的5，设备可以选择对跨越至控制器的多条路径的通信量进行平衡，由此将性能值3、4、5、6、7视为与5相同的值。在具有最低的四舍五入的度量的这些控制器之中，转发设备可以采用轮询调度算法来选择创建快速路径的控制器。这有助于限制由于网络对真实世界情况作出反应而导致的性能值的轻微修改的影响。

[0107] 在实施例中，网络中的控制设备不需要彼此同步。如果控制设备失败，在一段时间内未从该控制设备接收到控制分组之后，转发设备移除用于保持至该控制设备的路径的标签。转发设备利用诸如轮询调度的相同选择算法继续使用其他控制设备。

[0108] 具有多个控制设备大大提高了网络的恢复能力(resiliency)。如果控制设备失败，或者连接到控制设备的链路失败，则转发设备可以继续使用网络中的其他控制设备。多个控制设备还有助于平衡网络中的负载。通过智能地选择具有较少数目的新连接请求的控制设备，或者通过选择具有最佳路径(例如具有最低延迟的路径)的控制设备，转发设备可以在请求新连接时避免网络的拥塞段。

[0109] 控制设备和转发设备模块

[0110] 图6示出了说明控制设备102及其模块和转发设备104及其模块的示图600。

[0111] 具体地，控制设备102包括控制器发现模块610、路径目的地模块608、数据转发模块612和拓扑数据库606。各个模块可以如以上参照图2中的方法所述进行操作。

[0112] 控制器发现模块610被配置为确定转发设备的网络的拓扑结构。控制器发现模块610将控制分组发送到与控制设备直接连接的转发设备。控制分组包括控制设备的地址和序列号。使用地址和序列号的组合，每个控制分组可以在整个网络中被唯一地识别。

[0113] 随着控制分组穿越网络，接收到控制分组的每个转发设备将其自身的地址添加到控制分组。此外，每个转发设备将表明网络特征的度量更新到每个控制分组。在接收到控制分组时，转发设备从控制分组知晓为了到达该转发设备控制分组所经过的路径。此外，使用性能度量信息，转发设备知晓在多个转发分组所经过的多条路径之中哪条路径是到控制设备的最佳路径，如处理200所述。因此，转发设备可以动态地更新到控制设备的最佳路径。

[0114] 在建立到转发设备的路径之后，控制器发现模块610从网络中的转发设备接收拓扑结构和网络条件信息。使用该信息和来自网络中的其他转发设备的类似信息，控制器发现模块610创建拓扑数据库606。

[0115] 拓扑数据库606存储所发现的网络中的转发设备的拓扑结构。拓扑数据库606反映转发设备和链路如何连接在一起的当前拓扑结构。此外，拓扑数据库606存储表明网络的当前条件的信息。网络的条件可以包括关于网络链路或设备的延迟、分组丢失、拥塞或抖动。

[0116] 路径确定模块608建立用于新网络连接的路径。在源转发设备请求连接时，路径确定模块608确定从源转发设备到目的转发设备的连接路径。为了确定路径，路径确定模块608使用拓扑数据库606。使用存储在拓扑数据库606中的拓扑结构和网络条件信息，路径确定模块608可以找到从源转发设备到目的地的最优路径。最优路径可以是例如最短路径或者具有最小延迟的路径或者分组丢失最少的路径。

[0117] 在确定用于连接的路径之后，路径确定模块608还将配置命令发送到所确定的路径上的每个转发设备。如图4B中的示例实施例所示，路径确定模块创建并发送配置命令

422、424、426和428。配置命令配置相应的转发设备。所配置的转发设备在由路径确定模块608针对连接所确定的路径上对属于该连接的分组进行路由。

[0118] 在路径确定模块确定针对连接的路径时，数据转发模块612运送针对所请求的连接的初始分组，并通过发送配置命令配置转发设备。如图4A所示，通过从源转发设备到控制设备的路径将初始分组402、404、406、408和410转发到控制模块。在连接路径正被建立时，数据转发模块612通过从控制设备到目的转发设备的路径转发初始分组。

[0119] 在图6所示的示例实施例中，转发设备104包括路径请求模块622、交换发现模块624、路由表626和性能度量模块628。

[0120] 在转发设备接收到用于新数据连接的请求之后，路径请求模块622将消息发送到请求建立连接路径的路由控制设备。路径请求模块622使用在处理200的步骤216所建立的到控制设备的路径来发送用于建立连接路径的请求。路径请求指示路由控制设备确定从源转发设备到目的转发设备的路径。然后，路径确定模块608确定连接路径并将转发设备配置为沿所确定的路径转发新数据流。

[0121] 交换发现模块624接收由控制设备的控制器发现模块610发送的控制分组。交换发现模块624确定从转发设备到控制设备的路径，如方法200的步骤216中所示。在实施例中，交换发现模块624基于控制分组中的信息和它接收到控制分组的端口来进行该确定。

[0122] 为了建立从转发设备到控制设备的路径，交换发现模块624部分地基于接收到控制分组的端口来识别转发设备上的哪个端口将数据转发到达路由控制设备。

[0123] 在实施例中，如果从邻近设备到来的控制分组中的性能度量表明从控制设备到转发设备的具有最佳性能的路径，则交换发现模块624确定用于路由到控制设备的标签，如方法300的步骤306所示。

[0124] 在如处理200的步骤214所示，确定接收到的控制分组是否指示最佳性能路径以及接收到的控制分组是否尚未在转发设备的其他端口上被转发之后，交换发现模块624还在它的除了接收到控制分组的端口之外的端口上分发控制分组。

[0125] 转发设备的路由表626包含关于如何在网络中转发数据分组的信息。配置命令422、424、426和428利用关于如何转发分组的信息来更新转发设备中的路由表。此外，在接收到用于建立到控制设备102的路径的标签之后，交换发现模块624基于所确定的标签在路由表626中建立转发规则。

[0126] 性能度量模块628更新传入控制分组的性能度量字段。传入控制分组包括这样的字段，该字段包含从控制设备到先前转发设备的路径的网络性能度量。性能度量模块更新性能度量字段，使得其反映直到当前转发设备的路径的性能。

[0127] 性能度量模块628还确定转发设备是否之前转发了具有更好性能度量的另一控制分组。性能度量模块628向交换发现模块624通知是否在转发设备端口上分发了具有更好度量的之前的相同控制分组。

[0128] 结论

[0129] 本文中使用的术语“用户”可以包含网络连接服务的客户(例如，利用网络连接服务的企业的员工)和服务提供商自身的网络管理员二者。用户还可以处于不同公司或组织。

[0130] 拓扑数据库606可以是包括永久存储器的任何存储类型的结构存储器。在示例中，每个数据库可以被实现为关系数据库或文件系统。

[0131] 图6中的每个设备和模块可以用硬件、软件、固件或其任意组合来实现。

[0132] 图6中的每个设备和模块可以实现在相同或不同的计算设备上。这种计算设备可以包括但不限于：个人计算机、如移动电话的移动设备、工作站、嵌入式系统、游戏机、电视、机顶盒或任何其他计算设备。此外，计算设备可以包括但不限于：具有用于执行和存储指令的处理器和存储器的设备，包括非暂时性存储器。存储器可以有形地体现数据和程序指令。软件可以包括一个或多个应用以及操作系统。硬件可以包括但不限于：处理器、存储器和图形用户界面显示器。计算设备还可以具有多个处理器以及多个共享或单独的存储器组件。例如，计算设备可以是群集或分布式计算环境或服务器群的一部分或整体。

[0133] 对于不同要素或步骤，有时使用诸如“(a)”、“(b)”、“(i)”、“(ii)”等的标识符。为了清楚起见使用这些标识符，并且这些标识符不必指定要素或步骤的顺序。

[0134] 以上已经借助功能建立块描述了本发明，其中，功能建立块示出了指定功能及其关系的实现。本文中为了便于描述，任意地定义这些功能建立块的边界。可以定义替代边界，只要指定功能及其关系被适当执行。

[0135] 特定实施例的前述描述将充分全面地揭示本发明的总体性质，在不脱离本发明的总体构思的情况下，其他人员可以通过应用本领域内的技术知识容易地修改和/或改写这些特定实施例的各种应用，而无需过多实验。因此，基于本文中呈现的教导和引导，这种改写和修改意在处于所公开的实施例的等同物的意义和范围内。应当理解，本文中的措辞或术语是为了描述的目的，而不是限制，从而本说明书中的术语或措辞应当由本领域技术人员考虑教导和引导来进行解释。

[0136] 本发明的宽度和范围不应当受上述任何一个示例性实施例的限制，但是应当仅根据以下权利要求及其等同物来限定。

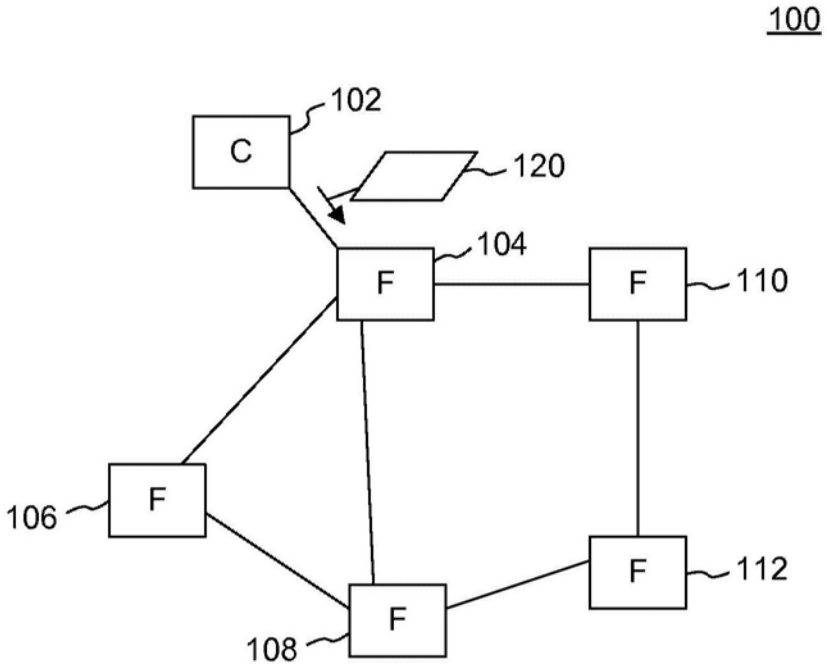

100
102
C
120
104
F
F
110
106
F
F
112
F
108

图1A

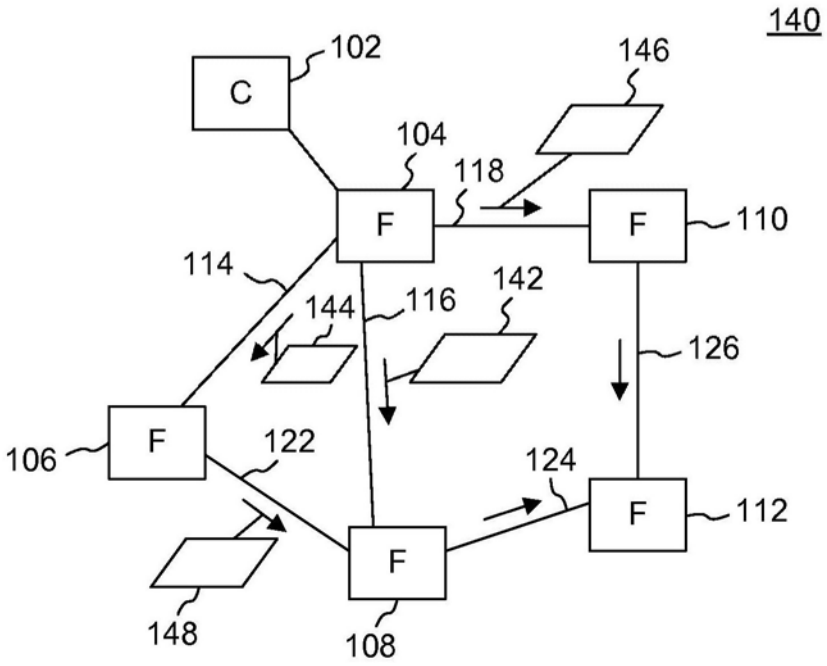

140
102
C
146
104
118
F
F
110
114
142
144
116
126
F
106
122
124
F
112
F
148
108

图1B

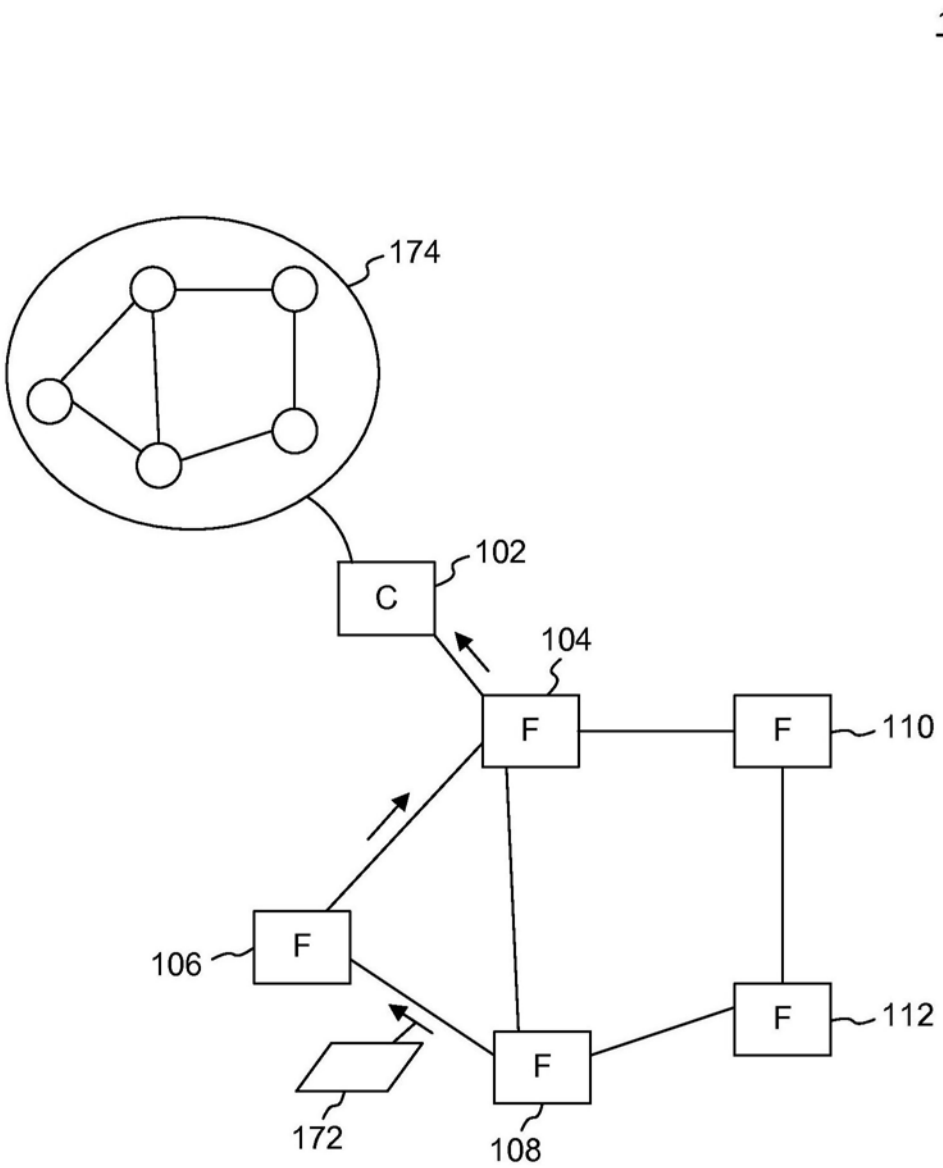


图1C

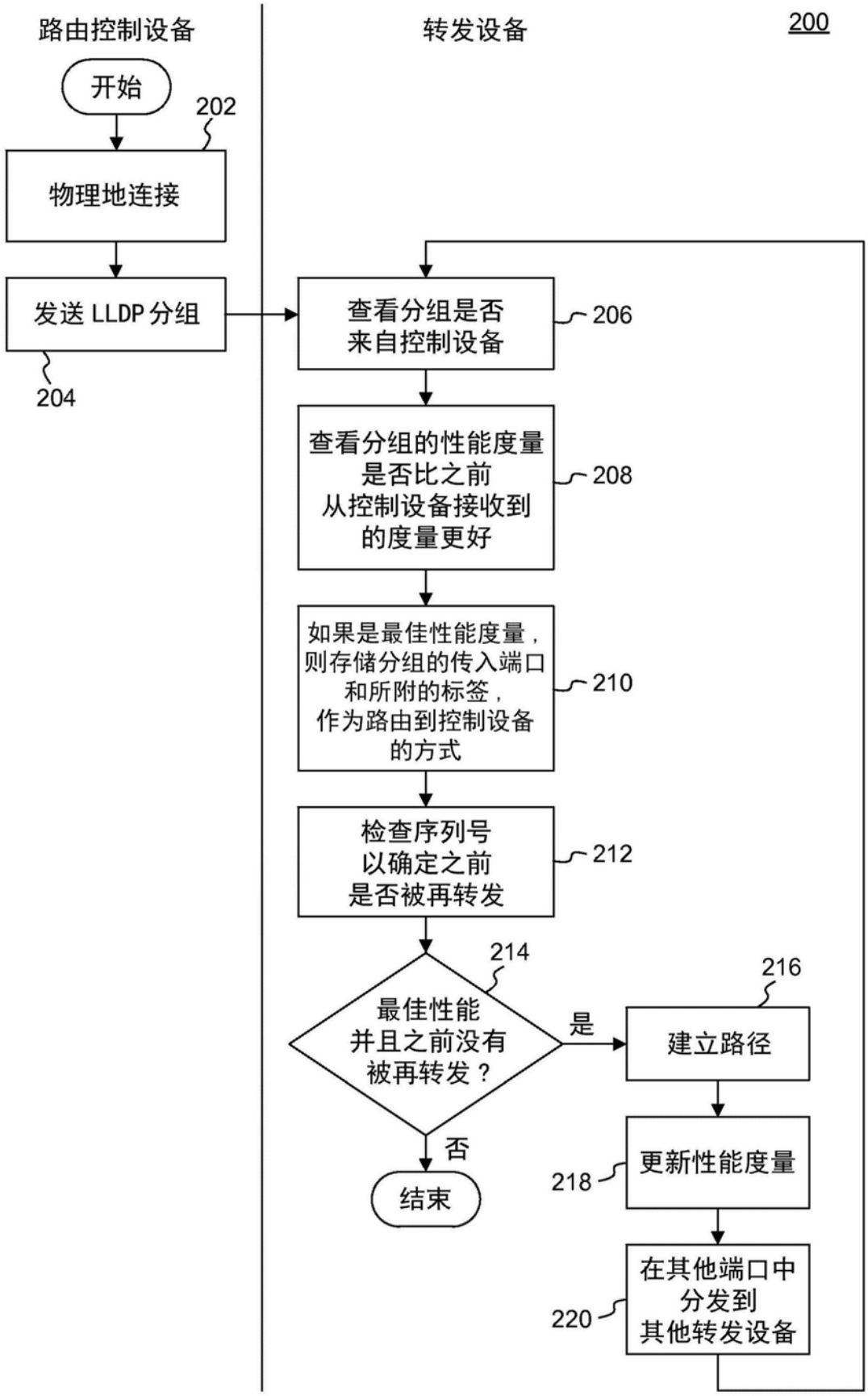

路由控制设备
转发设备
200
开始
202
物理地连接
发送 LLDP 分组
204
查看分组是否
来自控制设备
206
查看分组的性能度量
是否比之前
从控制设备接收到
的度量更好
208
如果是最佳性能度量，
则存储分组的传入端口
和所附的标签，
作为路由到控制设备
的方式
210
检查序列号
以确定之前
是否被再转发
212
214
最佳性能
并且之前没有
被再转发？
是
否
结束
216
建立路径
更新性能度量
218
在其他端口中
分发到
其他转发设备
220

图2

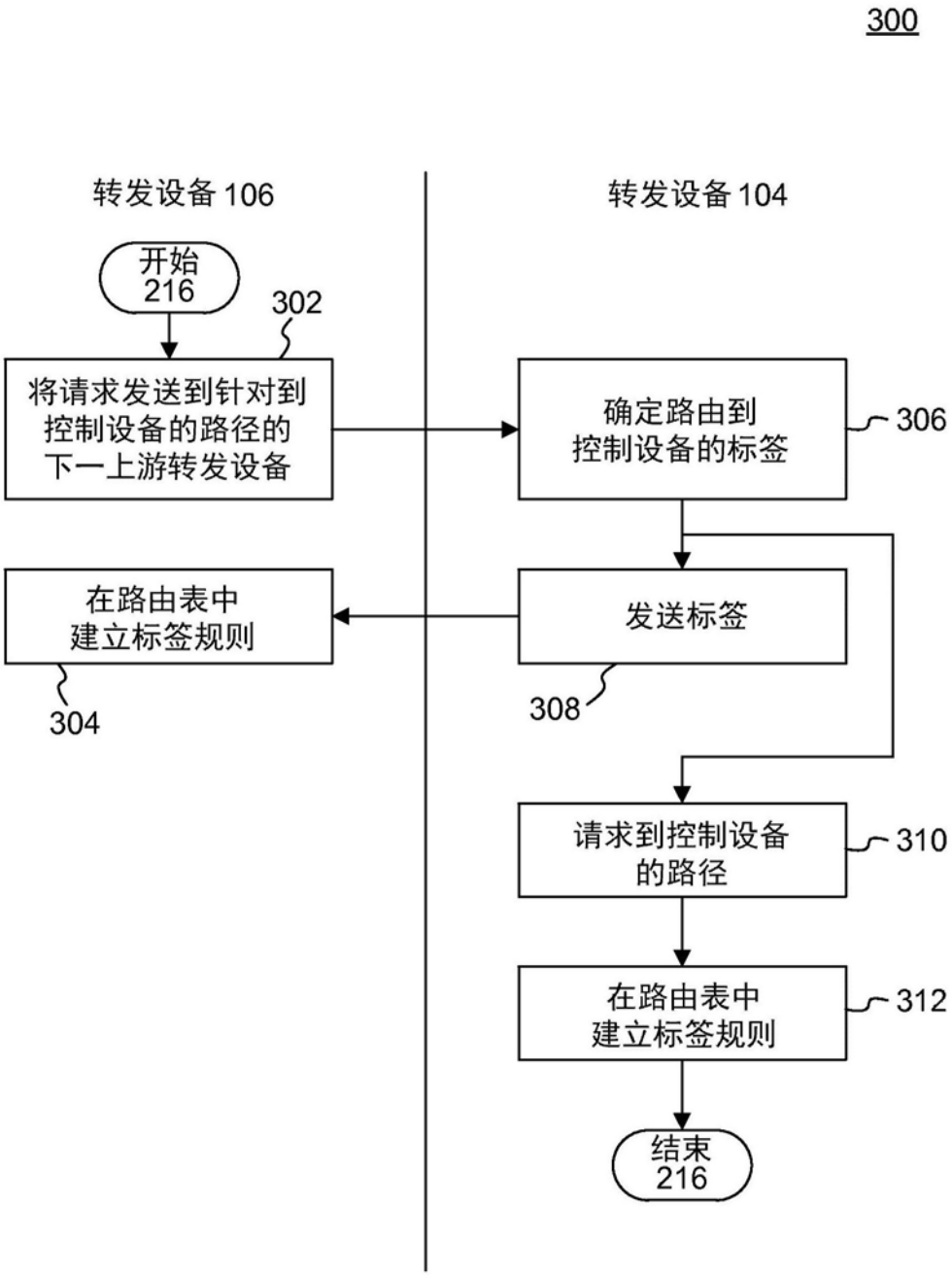

300
转发设备 106
转发设备 104
开始
216
302
将请求发送到针对到
控制设备的路径的
下一上游转发设备
确定路由到
控制设备的标签
306
在路由表中
建立标签规则
304
发送标签
308
请求到控制设备
的路径
310
在路由表中
建立标签规则
312
结束
216

图3

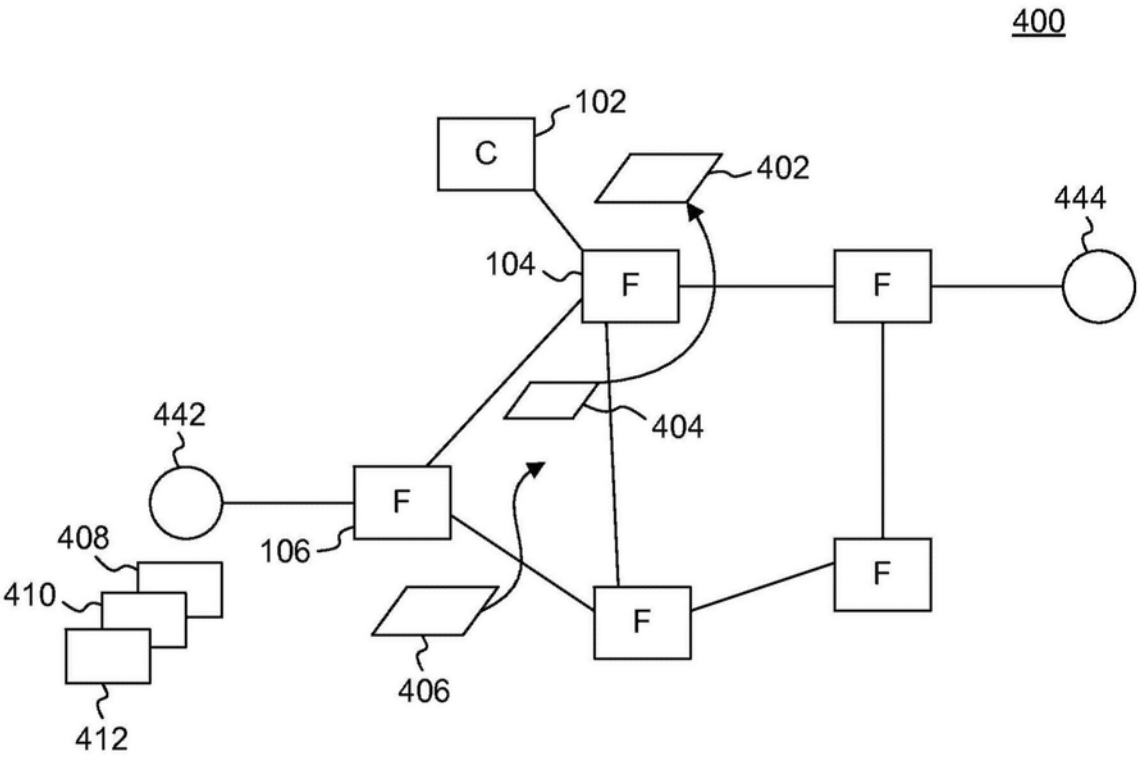

400
102
C
402
444
104
F
F
404
442
F
106
408
410
412
406
F
F

图4A

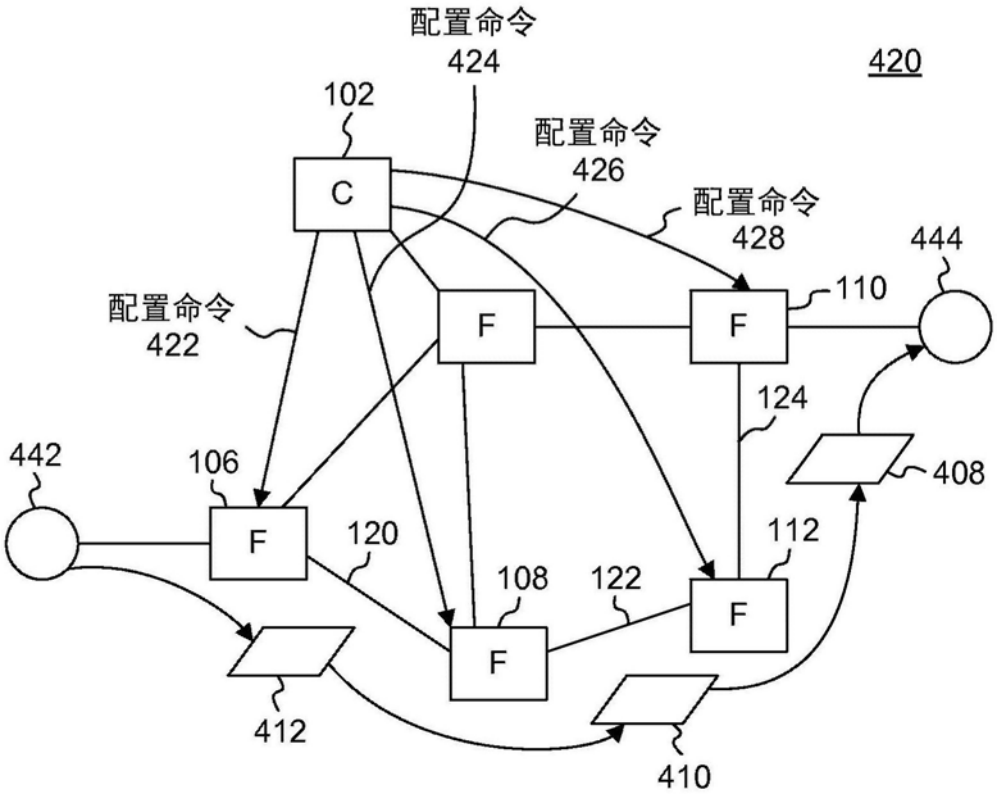

配置命令
424
420
102
配置命令
426
C
配置命令
428
444
110
配置命令
422
F
F
124
442
106
408
F
120
112
108
122
F
F
412
410

图4B

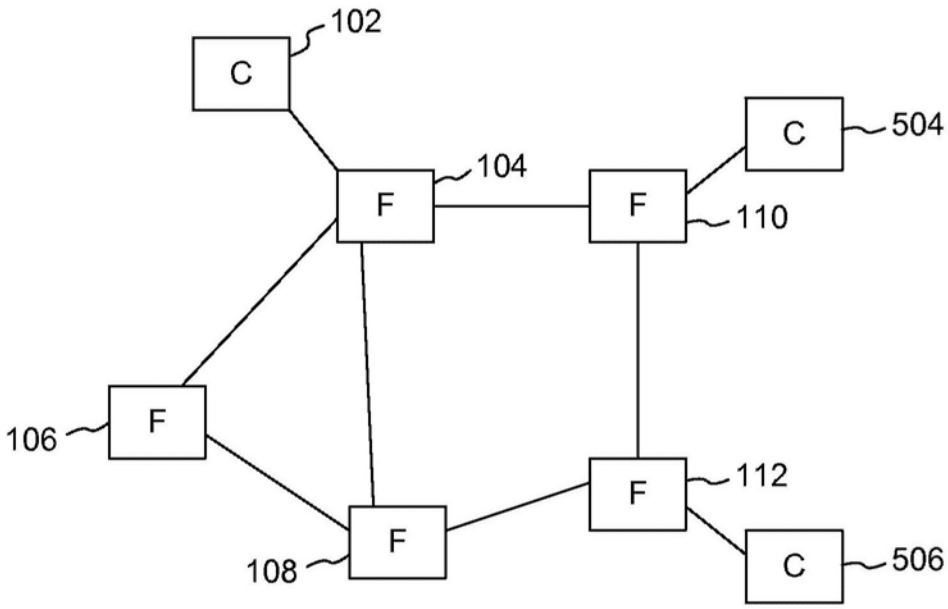


图5

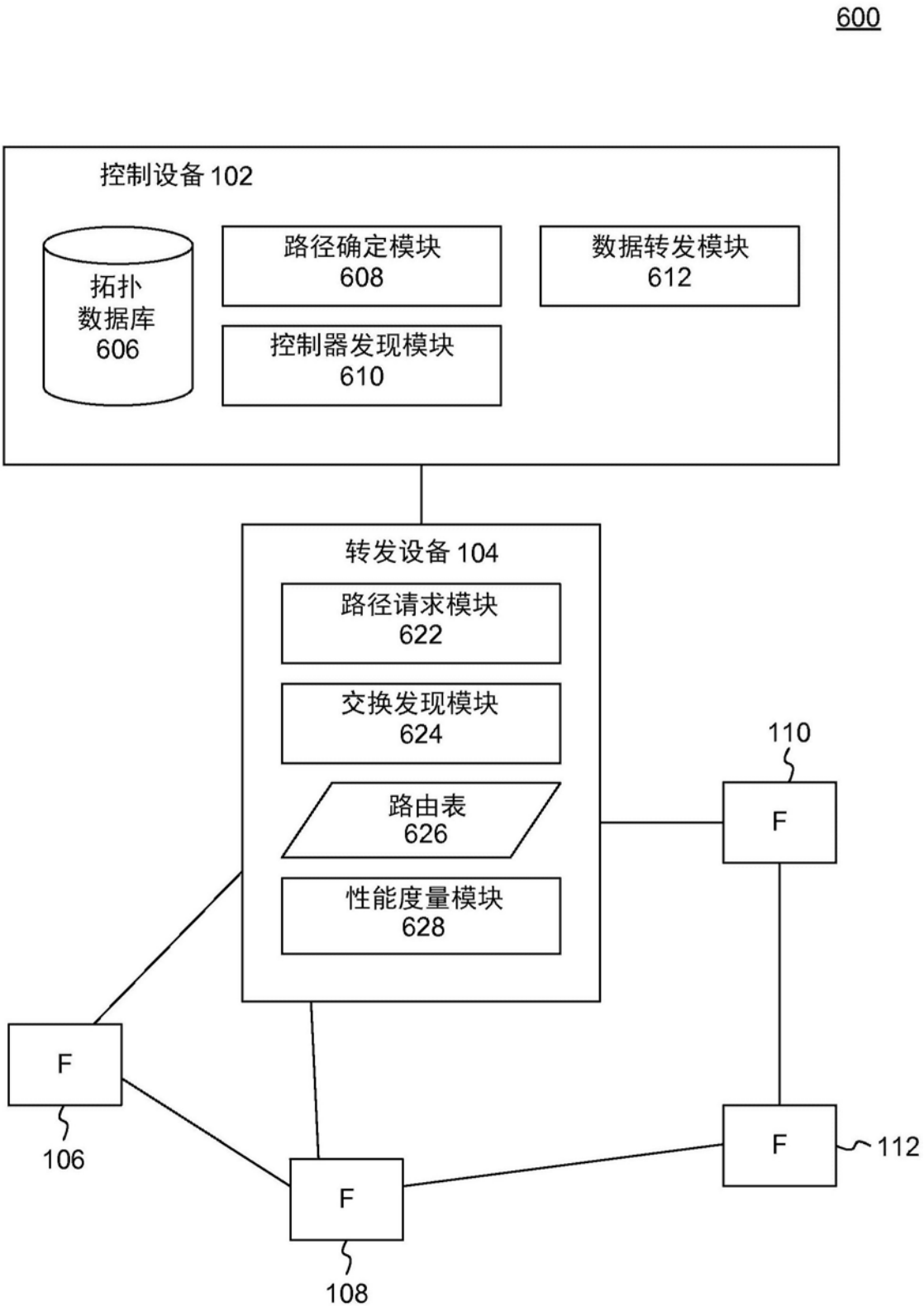

600
控制设备 102
拓扑
数据库
606
路径确定模块
608
数据转发模块
612
控制器发现模块
610
转发设备 104
路径请求模块
622
交换发现模块
624
路由表
626
性能度量模块
628
110
F
F
106
F
108
F
112

图6