(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 107078956 A
(43)申请公布日 2017.08.18

(21)申请号 201580056868.6

(22)申请日 2015.12.11

(30)优先权数据
62/090,627 2014.12.11 US

(85)PCT国际申请进入国家阶段日
2017.04.20

(86)PCT国际申请的申请数据
PCT/US2015/065290 2015.12.11

(87)PCT国际申请的公布数据
WO2016/094825 EN 2016.06.16

(71)申请人 博科通迅系统有限公司
地址 美国加利福尼亚

(72)发明人 R·贝斯

(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 边海梅

(51)Int.Cl.
H04L 12/717(2013.01)
H04L 12/771(2013.01)

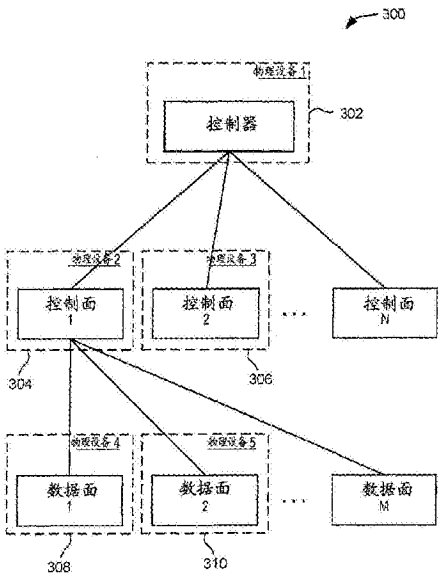
权利要求书4页 说明书17页 附图7页

(54)发明名称

多层分布式路由器体系结构

(57)摘要

一种分布式多层网络路由体系结构包括多个层，这多个层包含包括控制器的控制器层、包括一个或多个控制面子系统的控制面层以及包括一个或多个数据面子系统的数据面层。控制器可以耦合到一个或多个控制面子系统。控制面子系统继而可以耦合到一个或多个数据面子系统，数据面子系统可以包括一个或多个软件数据面子系统和/或硬件数据面子系统。在某些实施例中，分布式路由器的各个子系统的位置可以分布在网络中的各种设备之间。



CN 107078956 A

1.一种路由器系统,包括:

位于第一设备上的控制面子系统;以及

被配置为转发数据报文的多个数据面子系统,所述多个数据面子系统包括第一数据面子系统和第二数据面子系统,第一数据面子系统位于第二设备上并且第二数据面子系统位于第三设备上;

其中所述控制面子系统被配置为:

接收指令;

基于所述指令确定动作;

将与所述动作对应的消息传送到第一数据面子系统;以及

将与所述动作对应的消息传送到第二数据面子系统;

其中第一数据面子系统被配置为执行与所述动作对应的处理;以及

其中第二数据面子系统被配置为执行与所述动作对应的处理。

2.根据权利要求1所述的路由器系统,其中:

第一数据面子系统是由第二设备执行的软件数据面子系;以及

第二数据面子系统是位于第三设备上的硬件数据面子系统,第二数据面子系统包括转发硬件。

3.根据权利要求1或者权利要求2所述的路由器系统,其中所述控制面子系统使用控制信道将所述消息传送到第一数据面子系统和第二数据面子系统。

4.根据权利要求2或者权利要求3所述的路由器系统,其中所述转发硬件是商用硅芯片、网络接口卡(NIC)或者现场可编程门阵列(FPGA)。

5.根据权利要求1、3或者4所述的路由器系统,其中:

第二数据面子系统是位于第三设备上的硬件数据面子系统,第二数据面子系统包括转发硬件;

所述消息被转换成与所述转发硬件对应的一个或多个应用编程接口API的集合;以及

所述转发硬件被配置为执行API的集合。

6.根据权利要求1-5中的任何一项所述的路由器系统,还包括:

由第四设备执行的控制器,所述控制器被配置为与多个控制面子系统通信,所述多个控制面子系统包括所述控制面子系统;

其中所述控制面子系统被配置为从所述控制器接收所述指令。

7.根据权利要求1-6中的任何一项所述的路由器系统,其中第一数据面子系统被配置为:

经由输入端口接收数据报文;

确定用于转发所述数据报文的输出端口;以及

使用所确定的输出端口将所述数据报文转发到下一跳。

8.根据权利要求1-6中的任何一项所述的路由器系统,其中第二数据面子系统被配置为:

经由输入端口接收数据报文;

将所述数据报文转发到第一数据面子系统,以用于执行用于所述报文的第一服务;

在第一数据面子系统已经执行第一服务之后,从第一数据面子系统接收所述数据报

文；

确定用于转发所述数据报文的输出端口；以及

使用所确定的输出端口转发所述数据报文。

9.根据权利要求1-6中的任何一项所述的路由器系统，其中：

第一数据面子系统是由第二设备执行的软件数据面子系统；

第二数据面子系统是位于第三设备上的硬件数据面子系统，第二数据面子系统包括转发硬件；以及

第一服务是用于所述数据报文的防火墙服务、服务质量(QoS)、网络地址转换(NAT)或者安全服务当中的一个。

10.根据权利要求1-6中的任何一项所述的路由器系统，其中：

第一数据面子系统被配置为：

经由输入端口接收数据报文；

执行与用于所述数据报文的第一服务对应的处理；以及

将所述数据报文转发到第二数据面子系统；以及

第二数据面子系统被配置为：

从第一数据面子系统接收所述数据报文；

执行与用于数据报文的第二服务对应的处理；

确定用于转发所述数据报文的输出端口；以及

使用所确定的输出端口将所述数据报文转发到下一跳。

11.一种方法，包括：

由位于第一设备上的控制面子系统接收与网络策略有关的指令；

由所述控制面子系统基于所述指令确定动作；

由所述控制面子系统将与所述动作对应的消息传送到位于与第一设备不同的第二设备上的第一数据面子系统，第一数据面子系统被配置为转发数据报文；

由所述控制面子系统将与所述动作对应的消息传送到位于与第一设备不同的第三设备上的第二数据面子系统，第二数据面子系统被配置为转发数据报文；

响应于从所述控制面子系统接收到所述消息，由第一数据面子系统执行处理；以及

响应于从所述控制面子系统接收到所述消息，由第二数据面子系统执行处理。

12.根据权利要求11所述的方法，其中：

第一数据面子系统是由第二设备执行的软件数据面子系统；

第二数据面子系统是位于第三设备上的硬件数据面子系统，第二数据面子系统包括转发硬件；

将与所述动作对应的所述消息传送到第一数据面子系统包括由所述控制面子系统使用控制信道将所述消息传送到第一数据面子系统；以及

将与所述动作对应的所述消息传送到第二数据面子系统包括由所述控制面子系统使用控制信道将所述消息传送到第二数据面子系统。

13.根据权利要求11或者权利要求12所述的方法，其中：

第二数据面子系统是位于第三设备上的硬件数据面子系统，第二数据面子系统包括转发硬件；

所述方法还包括：

将所述消息转换成与所述转发硬件对应的一个或多个应用编程接口API的集合；以及

由所述转发硬件执行API的集合。

14.根据权利要求11、12或者13所述的方法，还包括：

由第四设备执行控制器，所述控制器被配置为与多个控制面子系统通信，所述多个控制面子系统包括所述控制面子系统；以及

将所述指令从所述控制器传送到所述控制面子系统。

15.根据权利要求11-14中的任何一项所述的方法，还包括：

由第一数据面子系统经由输入端口接收数据报文；

由第一数据面子系统确定用于转发所述数据报文的输出端口；以及

由第一数据面子系统使用所确定的输出端口将所述数据报文转发到下一跳。

16.根据权利要求11-14中的任何一项所述的方法，还包括：

由第一数据面子系统经由输入端口接收数据报文；

由第一数据面子系统将所述数据报文转发到第二数据面子系统，以用于执行用于所述报文的第一服务；

由第二数据面子系统执行第一服务；

在第二数据面子系统已经执行第一服务之后，由第一数据面子系统从第二数据面子系统接收所述数据报文；

由第一数据面子系统确定用于转发所述数据报文的输出端口；以及

由第一数据面子系统使用所确定的输出端口转发所述数据报文。

17.根据权利要求11-14中的任何一项所述的方法，还包括：

由第一数据面子系统经由输入端口接收数据报文；

由第一数据面子系统执行与用于所述数据报文的第一服务对应的处理；

由第一数据面子系统将所述数据报文转发到第二数据面子系统；

由第二数据面子系统执行与用于所述数据报文的第二服务对应的处理；

由第二数据面子系统确定用于转发所述数据报文的输出端口；以及

由第二数据面子系统使用所确定的输出端口将所述数据报文转发到下一跳。

18.一种系统，包括：

由第一设备执行的控制器；

由控制器控制的多个控制面子系统，所述多个控制面子系统包括由第二设备执行的第一控制面子系统；以及

由第一控制面子系统控制的多个数据面子系统，所述多个数据面子系统中的每个数据面子系统被配置为转发数据报文，所述多个数据面子系统包括由第三设备执行的软件数据面子系统和位于第四设备上的硬件数据面子系统，所述软件数据面子系统被配置为使用第一转发信息转发由所述软件数据面子系统接收的数据报文，并且所述硬件数据面子系统被配置为使用第二转发信息转发由所述硬件数据面子系统接收的数据报文；以及

其中：

第一控制面子系统被配置为：

从所述控制器接收与网络策略有关的指令；

确定将要执行以用于实现所述指令的动作;以及

将所述动作传送到所述软件数据面子系统;以及

所述软件数据面子系统被配置为执行所述动作,从而导致第一转发信息被更新。

19.根据权利要求18所述的系统,其中:

第一控制面子系统被配置为将所述动作传送到所述硬件数据面子系统;以及

所述硬件数据面子系统被配置为执行使得第二转发信息被更新的处理。

20.根据权利要求18或者权利要求19所述的系统,其中:

所述软件数据面子系统被配置为:

接收第一数据报文;

执行与用于第一数据报文的第一服务有关的处理;

基于第一转发信息确定用于第一数据报文的第一输出端口;以及

使用第一输出端口转发第一数据报文;以及

所述硬件数据面子系统被配置为:

接收第二数据报文;

执行与用于第一数据报文的第二服务有关的处理,第二服务与第一服务不同;

基于第二转发信息确定用于第一数据报文的第二输出端口;以及

使用第二输出端口转发第二数据报文。

21.一种装置,包括:

用于在第一设备处接收与网络策略有关的指令的构件;

用于在第一设备处基于所述指令确定动作的构件;

用于将与所述动作对应的消息从第一设备传送到与第一设备不同的第二设备的构件,第二设备被配置为转发数据报文;

用于将与所述动作对应的消息从第一设备传送到与第一设备不同的第三设备的构件,第二设备被配置为转发数据报文;

用于响应于从第一设备接收到所述消息而在第二设备处执行处理的构件;以及

用于响应于从第一设备接收到所述消息而在第三设备处执行处理的构件。

多层分布式路由器体系结构

[0001] 相关申请的交叉引用

[0002] 本申请是于2014年12月11日提交的、标题为“MULTI-LAYER ACCELERATED NETWORK ARCHITECTURE”的美国临时申请No.62/090,627的非临时申请，并且依据35U.S.C.119(e)要求该临时申请的权益和优先权，所述临时申请的全部内容通过引用合并在此，用于所有目的。

背景技术

[0003] 本公开涉及数据转发，并且更具体地，涉及多层分布式路由器体系结构。

[0004] 路由器是被配置为在网络之间转发数据(例如，数据报文)以便于将数据从它的源递送到它的预期目的地的专用网络设备。路由器被配置为确定通过互联网络拓扑的最优路由或者数据路径，并且使用这些最优路径传输数据报文。路由器还提供诸如防火墙、服务质量(QoS)等的各种其他服务。

[0005] 常规路由器典型地包括全部封装(house)在单个机盒或者机箱内的控制面子系统以及一个或多个数据面。在常规系统中可以实现为线卡的数据面被配置为经由数据面的一个或多个端口接收并且转发数据报文。转发基于由数据面存储的转发信息而执行。数据面用来转发数据报文的转发信息典型地由路由器的控制面子系统编程，该控制面子系统可以包括一个或多个管理卡。控制面子系统被配置为控制由路由器执行的与联网有关的功能，包括，例如，维护路由信息(例如，路由信息表)、对基于路由信息而转发信息的数据转发面进行编程、处置各种联网控制协议、处置在控制面处终止的报文(例如，控制报文和/或数据报文)的处理、处理访问控制列表(ACL)、服务质量(QoS)、管理功能等。

[0006] 如先前所指示的，在常规路由器中，路由器的控制面子系统以及由控制面子系统控制的数据面全部都物理地共同位于路由器的相同物理网络设备机盒或者机箱内。因此，在路由器的控制面子系统与一个或多个数据面子系统之间存在静态关系。这使得在可能包括成千上万个这种路由器的大型网络(例如，互联网)中对这种路由器进行编程是复杂并且不灵活的。当不得不对这种网络做出改变(例如，数据路径的添加或者重新配置、网络设备的添加或者移除)时，不得不对受改变所影响的每个个体路由器机盒进行单独地编程或者重新编程。这种以路由器为中心的管理要求使得这种网络的管理非常复杂并且耗时。因此，使用常规路由器的网络缺乏现今(以及未来)的网络应用(诸如云计算、移动计算、实时和按需视频流量等)所期望的灵活性和动态可配置性。

发明内容

[0007] 本公开涉及分布式联网，并且更具体而言涉及分布式多层网络路由体系结构。

[0008] 在某些实施例中，提供了一种多层分布式路由器体系结构。路由器的多个层可以包含包括控制器的控制器层、包括一个或多个控制面子系统的控制面层、以及包括一个或多个数据面子系统的数据面层。控制器可以耦合到一个或多个控制面子系统。控制面子系统又可以耦合到一个或多个数据面子系统，数据面子系统可以包括一个或多个软件数据面

子系统和/或硬件数据面子系统。

[0009] 在某些实施例中，分布式路由器的各种子系统的位置可以分布在网络中的各种设备之间。例如，控制器可以位于第一设备上，而由控制器控制的控制面子系统可以位于与第一设备不同的物理设备上。例如，第一控制面子系统可以位于第二设备上并且第二控制面子系统可以在第三设备上执行，其中第一设备、第二设备和第三设备是不同的设备并且可以位于不同的地理位置处并且可以经由一个或多个网络彼此耦合。由控制面子系统控制的数据面子系统也可以位于与控制这些数据面子系统的控制面子系统所位于的设备不同的物理设备上。在上面的示例中，其中第一控制面子系统位于第二设备上，由第一控制面子系统控制的第一数据面子系统可以位于第四设备上，并且也由第一控制面子系统控制的第二数据面子系统可以位于第五设备上，其中第二设备、第四设备和第五设备是不同的设备并且可以位于不同的地理位置处并且经由一个或多个网络彼此耦合。

[0010] 分布式路由器的不同层位于不同物理设备上的能力使得控制器与控制面子系统之间的关系以及控制面子系统与数据面子系统之间的关系能够是动态的。例如，关于控制器和控制面子系统，可以相对控制器远程地动态添加或者移除控制面子系统，并且控制器能够在具有动态改变的情况下操作。由控制器控制的控制面子系统的位置可以逻辑地和物理地分散在不同的物理设备中。关于控制面子系统和数据面子系统，数据面子系统可以动态地添加或者移除。例如，提供增强功能性和服务的软件数据面可以动态地添加到分布式路由器。由控制面子系统控制的数据面子系统的位置可以逻辑地和物理地分散在不同的物理设备中。

[0011] 例如，在一个实施例中，分布式路由器可以包括由第一设备执行的控制器以及由控制器控制的多个控制面子系统。多个控制面子系统可以包括由与第一设备不同的第二设备执行的第一控制面子系统。分布式路由器还可以包括由第一控制面子系统控制的多个数据面子系统。每个这种数据面子系统可以被配置为转发数据报文。多个数据面子系统可以包括由第三设备执行的软件数据面子系统以及位于第四设备上的硬件数据面子系统。软件数据面子系统可以被配置为使用第一转发信息转发由软件数据面子系统接收的数据报文，并且硬件数据面子系统可以被配置为使用第二转发信息转发由硬件数据面子系统接收的数据报文。在一个实施例中，第一控制面子系统可以从控制器接收与网络策略有关的指令。第一控制面子系统然后可以确定将要执行的用于实现指令的动作，并且将动作传送到软件数据面子系统。软件数据面子系统然后可以执行该动作，这可以导致第一转发信息被更新和/或关于由软件数据面子系统提供的一个或多个转发服务(例如，防火墙、QoS、NAT等)的其他信息被更新。第一控制面子系统也可以将该动作传送到硬件数据面子系统。硬件数据面子系统然后可以执行处理，从而使得第二转发信息和/或关于由硬件数据面子系统提供的一个或多个转发服务的信息被更新。

[0012] 在某些实施例中，软件数据面子系统可以被配置为使用软件数据面子系统可访问的第一转发信息和/或通过应用由软件数据面子系统提供的各种转发服务来转发报文。例如，当接收到数据报文时，软件数据面子系统可以执行与用于该数据报文的第一服务有关的处理，基于第一转发信息确定用于该数据报文的输出端口，以及使用所确定的输出端口转发该数据报文。

[0013] 在某些实施例中，硬件数据面子系统可以被配置为使用硬件数据面子系统可访问

的第二转发信息和/或通过应用由硬件数据面子系统提供的各种转发服务来转发报文。当接收到数据报文时，硬件数据面子系统可以执行与用于该数据报文的第二服务有关的处理，基于第二转发信息确定用于该数据报文的输出端口，以及使用所确定的输出端口转发该数据报文。由软件数据面子系统提供的转发服务可以与由硬件数据面子系统提供的转发服务相同或者不同。

[0014] 在某些实施例中，分布式路由器系统可以包括位于第一设备上的控制面子系统以及被配置为转发数据报文的多个数据面子系统。多个数据面子系统可以包括第一数据面子系统和第二数据面子系统，其中第一数据面子系统位于第二设备上并且第二数据面子系统位于第三设备上。在一个场景中，控制面子系统可以接收指令并且然后基于该指令确定动作。控制面子系统可以将与该动作对应的消息传送到第一数据面子系统和传送到第二数据面子系统。第一数据面子系统和第二数据面子系统然后可以执行与该动作对应的处理。

[0015] 在某些实施例中，上面描述的第一数据面子系统可以是由第二设备执行的软件数据面子系统，并且第二数据面子系统可以是位于第三设备上的硬件数据面子系统，第二数据面子系统包括转发硬件。控制面子系统可以使用控制信道与第一数据面子系统和第二数据面子系统通信。可以使用相同的控制信道，而不管目标数据面子系统是软件数据面子系统还是硬件数据面子系统。转发硬件可以是商用硅芯片、网络接口卡(NIC)、现场可编程门阵列(FPGA)等。

[0016] 在某些实施例中，可以为硬件数据面子系统上的转发硬件提供硬件抽象层(HAL)。HAL被配置为接收由控制面子系统传送的消息并且将该消息转换成与转发硬件对应的一个或多个应用编程接口(API)的集合。所确定的API的集合然后可以由转发硬件执行。

[0017] 在某些实施例中，分布式路由器可以包括控制器，该控制器可以由与执行控制面子系统和数据面子系统的设备不同的设备执行。用户可以使用控制器来配置网络范围的策略。与这种策略对应的指令可以从控制器传送到由该控制器控制的一个或多个控制面子系统。

[0018] 数据面子系统(包括软件数据面子系统和/或硬件数据面子系统)表示分布式路由器的数据转发层。在一个实施例中，数据面子系统可以经由输入端口接收数据报文。数据面子系统然后可以确定用于转发数据报文的输出端口。可以基于对数据面子系统可用的转发信息和/或与由数据面子系统提供的转发服务有关的其他信息做出该确定。转发信息和其他信息可以在接收数据报文之前已经从控制该数据面子系统的控制面子系统接收。数据面子系统然后可以使用所确定的输出端口将数据报文转发到下一跳(hop)。

[0019] 在另一个实施例中，当经由输入端口接收到数据报文时，数据面子系统可以将数据报文转发到第二数据面子系统，以执行用于报文的第一服务。在第二数据面子系统已经执行第一服务之后，初始数据面子系统然后可以接收从第二数据面子系统返回的数据报文。初始数据面子系统然后可以确定用于转发数据报文的输出端口，并且然后使用所确定的输出端口转发数据报文。初始数据面子系统可以是软件数据面子系统或者硬件数据面子系统。第一服务可以是诸如防火墙服务、QoS、网络地址转换(NAT)、安全服务等的服务。

[0020] 在又另一个实施例中，当经由输入端口接收到数据报文时，第一数据面子系统可以执行与用于数据报文的第一服务对应的处理，并且然后将数据报文转发到第二数据面子系统。第二数据面子系统然后可以执行与用于数据报文的第二服务对应的处理，确定用于

转发数据报文的输出端口，并且然后使用所确定的输出端口将数据报文转发到下一跳。

[0021] 在一个实施例中，提供了一种装置，包括：用于在第一设备处接收与网络策略有关的指令的构件；用于在第一设备处基于指令确定动作的构件；用于将与动作对应的消息从第一设备传送到与第一设备不同的第二设备的构件，第二设备被配置为转发数据报文；用于将与动作对应的消息从第一设备传送到与第一设备不同的第三设备的构件，第二设备被配置为转发数据报文；用于响应于从第一设备接收到消息而在第二设备处执行处理的构件；以及用于响应于从第一设备接收到消息而在第三设备处执行处理的构件。

[0022] 当参考下面的说明书、权利要求书和附图时，上述内容连同其他特征和实施例将变得更加明显。

附图说明

[0023] 图1是例示根据本发明的实施例的分布式多层路由器体系结构100的简化框图。

[0024] 图2描绘了根据本发明的实施例的具有硬件抽象层(HAL)的示例硬件数据面子系统。

[0025] 图3是例示根据本发明的实施例的体系结构的分布式性质的分布式路由器的简化框图。

[0026] 图4描绘了示出根据本发明的实施例的当接收到数据报文时由分布式路由器执行的处理的简化流程图。

[0027] 图5例示了根据本发明的实施例的分布式路由器的示例。

[0028] 图6是根据一个或多个方面的可以被用来执行分布式路由器的各种组件的计算系统或设备的简化框图。

[0029] 图7描绘了示出根据本发明的实施例的被执行以用于对分布式路由器进行编程的处理的简化流程图。

具体实施方式

[0030] 在下面的描述中，为了解释的目的，陈述了具体细节以便提供对本发明的实施例的透彻理解。然而，将清楚的是，可以在没有这些具体细节的情况下实践各种实施例。图和描述不旨在是限制性的。单词“示例性”在本文中被用来表示“充当示例、实例或者例示”。本文中描述为“示例性”的任何实施例或者设计不一定解释为相对其他实施例或者设计是优选的或者有利的。

[0031] 本公开涉及数据转发，并且更具体而言涉及分布式多层网络路由器体系结构。图1是例示根据本发明的实施例的分布式多层路由器体系结构100的简化框图。图1中描绘的实施例仅是示例并且不旨在过度地限制本发明的要求保护的实施例。本领域普通技术人员将认识到许多变型、替代和修改。例如，在一些其他实施例中，路由器100可以具有比图1中所示出的更多或者更少的组件、可以组合两个或多个组件、或者可以具有组件的不同配置或布置。

[0032] 图1中描绘的分布式路由器100包括多个层，这多个层包含包括控制器102的控制器层、包括一个或多个控制面子系统104-1、104-2的控制面层、以及包括一个或多个数据面子系统106的数据面层。数据面子系统106可以包括软件数据面子系统108-1、108-2等以及

硬件数据面子系统110-1、110-2等当中的一个或多个。分布式路由器100被配置为提供(OSI模型的)层2和层3路由功能性。分布式路由器100也可以提供开放系统互连(OSI)模型的层4-层7处的功能性。

[0033] 控制器102负责总体网络配置和管理。在某些实施例中，控制器102可以被用来设置全局网络范围策略以及它的编排，这可以跨多个网络设备完成。路由器100的用户可以使用控制器102来创建跨网络环境中不同组件和设备的端到端网络服务。控制器102可以被配置为对分布式体系结构中各个组件之间的端到端路径进行编排。控制器102负责设置服务路径并且提供对网络流量的可见性。控制器102也可以被配置为执行各种分析。

[0034] 控制器102经由一个或多个通信信道在通信上耦合到一个或多个控制面子系统104。控制器102可以使用不同的协议与控制面子系统通信。被控制器102用来与第一控制面子系统104-1通信的通信协议可以与被控制器102用来与第二控制面子系统104-2或者其他控制面子系统通信的通信协议相同或者不同。

[0035] 在一些实施例中，被控制器102用来与控制面子系统通信的通信协议的类型取决于控制面子系统是“薄边界(thin edge)”网络设备还是“厚边界(thick edge)”网络设备的一部分。在某些实施例中，如果网络设备可以维持关于流的状态并且可以跨重启维持配置状态，那么它被分类为厚边界设备。厚边界设备的示例包括路由器、防火墙和负载均衡器。厚边界设备能够独立于中央控制器确定路径可达性。典型地，中央控制器将通过对设备的配置告知厚边界设备如何确定路径可达性。这可以是通过策略(诸如ACL、NAT等)以及路由协议(诸如边界网关协议(BGP)、中间服务到中间服务协议(IS-IS)、开放最短路径优先(OSPF)等)的组合而进行的。一旦被通知了策略，厚边界设备就能够独立于第三方控制器确定转发路径。厚边界设备典型地也能够维持关于流的较高级别的状态。维持关于流的状态的示例将是路由器中的有状态的ACL规则，该有状态的ACL规则将报文分类为属于相同的TCP会话并且然后基于为其定义的策略路由这些报文。路由器包含足够的逻辑和存储器来维护两个端点之间的TCP会话的历史的记录，并且能够确定流中的各个报文是相关的，即使这些报文可能不具有相同的格式或者元数据。厚边界设备维护充足的逻辑以维持关于流的状态并且可以在没有控制器102的情况下操作。

[0036] 在某些实施例中，控制器102使用网络配置协议(NETCONF)与厚边界设备中的控制面子系统接口，NETCONF是由IETF开发并且标准化的网络管理协议。NETCONF提供安装、操纵和删除网络设备的配置的机制。它的操作在简单远程过程调用(RPC)层之上实现。NETCONF协议对于配置数据以及对于协议消息使用基于可扩展标记语言(XML)的数据编码。协议消息在安全传输协议之上互换。厚边界设备的示例包括但不限于由加利福尼亚州San Jose的Brocade通信系统公司提供的各种产品，诸如vRouter产品系列(例如，5400vRouter、5600vRouter)、vADC(虚拟应用递送控制器)、当使用NETCONF时的MLX系列、当使用NETCONF时的VDX系列等。

[0037] 如果网络设备使用一系列匹配动作表格来确定转发并且因此不能维持关于流的状态，那么它被分类为薄边界设备。薄边界设备的示例是交换机。薄边界设备典型地基于流条目进行交换并且典型地在物理交换机TCAM的抽象表示上建模。OpenFlow是代表薄边界设备的一种协议。在使用OpenFlow的情况下，中央控制器负责确定网络可达性。它然后对具有明确可达性的薄边界设备进行编程。薄边界设备典型地不具有独立于控制器确定转发路径

的能力，它们也不能维持足够的关于流量的状态以做出复杂的报文/会话关联。在某些实施例中，控制器102使用OpenFlow通信协议与薄边界设备中的控制面子系统通信。OpenFlow标准由开放网络基金会(ONF)管理，ONF是致力于软件定义网络(SDN)的推广和采用的组织。薄边界设备的示例包括但不限于由加利福尼亚州San Jose的Brocade通信系统公司提供的各种产品，诸如当使用OpenFlow时的MLX、当使用OpenFlow时的VDX等。

[0038] 控制器102也可以使用其他协议与控制面子系统通信。在某些实例中，控制器102可以使用YANG或者REST与控制面子系统通信。YANG是用来对由NETCONF远程过程调用和NETCONF通知操纵的配置和状态数据进行建模的数据建模语言。YANG可以被用来对用于在控制器102与控制面子系统之间的通信的配置和状态数据进行建模。在一些实施例中，REST(表述性状态转移)应用编程接口(API)可以由控制器102用来与控制面子系统通信。在替代实施例中，其他通信协议可以用于控制器102与各个控制面子系统之间的通信，这些其他通信协议诸如PCEP(路径计算元件通信协议)、BGP-LS(边界网关协议-链路状态)等等。

[0039] 在某些实施例中，控制器102被实现为由一个或多个处理实体执行的一组指令(代码或者程序)，这些处理实体诸如是由**Intel®**或者**AMD®**提供的处理器。例如，控制器102可以由诸如在图6中描绘并且下面描述的计算系统执行。在多核处理器环境中，控制器102可以由一个或多个处理器的一个或多个核执行。在包括由一个或多个处理器执行的一个或多个虚拟机的虚拟化环境中，控制器102可以在虚拟机内执行或者由虚拟机托管，或者甚至由管理程序(hypervisor)或者网络操作系统托管。

[0040] 如图1中描绘的，控制器102可以控制并且管理一个或多个(一对多关系)控制面子系统，诸如控制面子系统104-1和104-2。单个控制面子系统104能够控制多个数据面子系统(一对多关系)。由控制面子系统104控制的数据面子系统106可以是不同类型，包括软件数据面子系统108和/或硬件数据面子系统110。控制面子系统可以使用控制信道112并行地或者并发地与多个数据面子系统通信。

[0041] 在某些实施例中，控制器配置有它负责的一组控制面子系统。它然后基于网络策略配置这些实体。当配置厚边界设备时，它可以使用NETCONF这样做。当配置薄边界设备时，它可以使用OpenFlow这样做。

[0042] 控制面子系统104被配置为实现从控制器102接收的网络策略。在一个实施例中，控制面子系统104被配置为将策略转换成一组动作，这组动作将要由控制面子系统104控制的数据面106采取以实现策略。控制面子系统104因此负责采取从控制器102接收的网络策略并且将它们转换成用于数据面子系统106的动作。当数据面子系统在它们初始化时向控制面子系统注册它们自己时，控制面子系统获知它控制哪些数据面子系统。在某些实施例中，控制面子系统能使用诸如NETCONF或者OpenFlow的通信协议在外部编程，并且可以使用REST API进行编程。

[0043] 控制面子系统104可以使用各种不同的通信技术与数据面子系统106通信。在某些实施例中，如在图1中描绘的，控制面子系统104-1使用控制信道112与软件数据面108和硬件数据面110二者进行通信。在一个实施例中，经控制信道112的消息传递可以使用具有JSON(JavaScript开放符号)封装的ZeroMQ传输机制。ZeroMQ(也拼写为ØMQ或者0MQ或者ZMQ)是目标在于用在可扩展的分布式或者并发应用中的高性能异步消息传递库。它提供可以在没有专用消息代理(broker)的情况下运行的消息队列。该库被设计为具有熟悉的套接

字风格的API。它提供承载跨各种传输(像进程内、进程间、TCP和多播)的原子消息的套接字。它的异步I/O模型使得能够实现可扩展的多核应用。JSON提供用于提供分布式网络系统的分级视图的机制。从控制面子系统的视角,它在与软件数据面或者硬件数据面通信时使用相同的控制信道112和相同的通信方案。在这个意义上,控制面子系统不区分它正在与软件数据面通讯还是与硬件数据面通讯。

[0044] 例如,控制面子系统104可以从控制器102接收由用户配置的防火墙策略。作为响应,控制面子系统104可以创建用于防火墙策略的数据模型表示,并且将策略有关的信息存储在本地数据结构中。控制面子系统104然后可以将数据模型表示转换成JSON表示,该JSON表示然后被封装到ZeroMQ消息中并且使用ZeroMQ总线经控制信道112被传送到各个数据面子系统。

[0045] 控制面子系统也可以经由控制信道112接收来自一个或多个数据面子系统的消息。这些消息可以例如将关于数据面子系统的状况和状态信息传送到控制面子系统。例如,如果在数据面子系统中存在改变(例如,接口状态改变),那么该信息由数据面子系统封装到JSON表示中,该JSON表示继而被封装到ZeroMQ消息总线中并且使用控制信道112被发送到控制面子系统104。控制面子系统104然后可以采取动作以对状态改变做出响应(例如,改变状态信息、改变路由等)。

[0046] 在某些实施例中,每个控制面子系统104被实现为由一个或多个处理实体执行的一组指令(代码或者程序),这些处理实体诸如由**Intel®**或者**AMD®**提供的处理器。例如,控制面子系统104-1和104-2可以由诸如图6中描绘并且下面描述的计算系统执行。在多核处理器环境中,控制面子系统可以由一个或多个处理器的一个或多个核执行。在包括由一个或多个处理器执行的一个或多个虚拟机的虚拟化环境中,控制面子系统可以在虚拟机内执行或者由虚拟机托管,或者甚至由管理程序(例如,KVM管理程序)或者网络操作系统托管。控制面子系统可以部署在裸金属上、虚拟机中,等等。

[0047] 控制面子系统104负责对数据面子系统106的配置和管理。例如,控制面子系统被配置为维护用于分布式网络的路由和拓扑信息。该信息然后由控制面子系统使用来对一个或多个数据面子系统进行编程,使得数据面能够转发数据报文以便于将数据报文从它们的源递送到它们的预期目的地。在某些实施例中,控制面子系统被配置为将转发信息下载到数据面子系统。该转发信息然后由数据面子系统使用来转发由数据面子系统接收的数据报文。控制面子系统也可以将其他与策略有关的(例如,与防火墙策略有关的)信息下载到数据面子系统。

[0048] 存在控制面子系统可以接收网络和拓扑信息的各种方法。控制面子系统可以执行处理以支持在网络设备之间互换网络和拓扑信息的各种联网协议。控制面子系统可以生成并且处理去往分布式网络系统的其他组件的控制面报文或者由分布式网络系统的其他组件源起的控制面报文。例如,控制面子系统104可以彼此互换拓扑信息或者与其他网络设备或系统互换拓扑信息,以及使用诸如路由信息协议(RIP)、开放最短路径优先(OSPF)或者边界网关协议(BGP)的路由协议来构造/维护路由表。

[0049] 在某些实施例中,为了高可用性或者故障转移(failover)的目的,控制面子系统可以与另一个控制面子系统通信。在某些实施例中,能够使用控制信道启用两个控制面子系统之间的通信。

[0050] 数据面子系统106表示分布式路由器100的数据转发层，并且可以包括一个或多个软件数据面子系统108-1、108-2等(总称为“108”)和硬件数据面子系统110-1、110-2等(总称为“110”)。例如，分布式路由器实现可以仅具有一个或多个软件数据面子系统并且没有硬件数据面子系统。在另一种实现中，所有数据面可以是硬件数据面子系统而没有软件数据面子系统。在再一种实现中，数据面层可以包括软件数据面子系统和硬件数据面子系统的组合。

[0051] 总的来说，数据面子系统106负责接收数据报文并且基于从控制面子系统104接收的转发和配置信息将数据报文转发到网络中的下一跳。每个数据面子系统被编程有从一个或多个控制面子系统接收的转发信息。数据面子系统使用转发信息来转发数据报文。在某些实现中，数据面子系统支持(OSI模型的)层2和层3路由功能性。在某些实施例中，数据面子系统还可以提供OSI层4-层7功能性。数据面子系统106可以提供与路由器有关的功能性，诸如有状态的防火墙和无状态的防火墙、应用流、应用逻辑网关、网络访问设备(NAD)能力，以及在一些实例中，层4路由和更高级别数据报文检查能力的部分或者全部实现、与MPLS有关的处理、网络地址转换(NAT)、有状态的网络附属存储(NAS)、分级QoS(服务质量)、Q-in-Q、对互联网协议安全(IPsec)的支持、通用路由封装(GRE)、虚拟可扩展LAN(VXLAN)、用于数据流的分析和管理的各种工具、深度报文检查(DPI)、最长报文匹配(LPM)报文查找、虚拟私有网络(VPN)以及其他功能性。

[0052] 数据面子系统106可以经由数据面子系统的端口接收数据报文。端口可以是逻辑端口或者物理端口。数据面子系统然后确定将如何转发接收到的报文。作为该确定的一部分，数据面子系统可以从接收到的数据报文中提取信息(例如，提取报头信息)，根据所提取的信息中确定数据报文的源和目的地信息，以及基于所提取的信息、由数据面子系统从控制面子系统接收和由数据面子系统存储的转发信息以及其他可适用的策略有关的信息确定将如何转发数据报文。在某些实施例中，如果数据面子系统不能确定将如何转发接收到的报文，那么丢弃该报文。否则，数据面子系统确定将被用于把报文转发到下一跳的输出端口，以便于将报文传送到它的预期目的地。数据面子系统然后使用所确定的输出端口转发数据报文。

[0053] 数据面子系统也可以经由子系统的端口彼此通信。软件数据面子系统可以与另一个软件数据面子系统或者与另一个硬件数据面子系统通信。硬件数据面子系统可以与另一个硬件数据面子系统或者与另一个软件数据面子系统通信。在某些实施例中，外围组件互连快速路(PCIe)接口被用于各个数据面之间的通信。在某些实施例中，软件数据面子系统108可以使用标准化接口与硬件数据面子系统108通信，其中这些标准化接口使用REST、YANG、ZeroMQ、JSON表示或者任何其他适当协议实现。在数据面子系统的每个之间提供标准化接口允许数据面子系统的换入和换出，而不需要担心数据面子系统将如何彼此通信。一般地，在分布式路由器的各个层之间提供标准化通信接口能够实现在不同层处换入和换出各种组件的灵活性而不影响路由器功能性。这也使得能够以容易并且灵活的方式对路由器进行改变，诸如添加另外的功能性。

[0054] 如上面所指示的，数据面子系统可以是软件数据面子系统108或者硬件数据面子系统110。软件数据面子系统108实现为由一个或多个处理实体执行的一组指令(或者代码)，这些处理实体是诸如由**Intel®、AMD®**、Power、ARM提供的处理器、专用集成电路

(ASIC)或者现场可编程门阵列(FPGA)或者网络处理单元(NPU)。例如,图1中描绘的软件数据面子系统108-1和108-2可以由诸如图6中描绘并且在下面描述的计算系统执行。在多核处理器环境中,软件数据面子系统可以由一个或多个处理器的一个或多个核执行。在包括由一个或多个处理器执行的虚拟机的虚拟化环境中,软件数据面子系统可以在虚拟机内执行或者由虚拟机托管,或者甚至由管理程序(例如,KVM管理程序)或者网络操作系统托管。

[0055] 如上面所指示的,软件数据面子系统也可以由一个或多个NPU执行。NPU一般是专门设计并且优化以处置网络有关的功能的可编程多核微处理器。网络处理器典型地是软件可编程设备并且将具有与通用中央处理单元类似的一般特性,并且针对执行诸如图案匹配、键查找、计算、位操纵、缓冲器的分配和管理等与联网有关的功能而优化。软件数据面子系统可以部署在裸金属上、虚拟机中,等等。

[0056] 软件数据面子系统108被配置为提供高速数据报文转发能力。在一个实施例中,软件数据面子系统可以使用基于Linux的转发流水线(pipeline)。在另一个实施例中,为了更加快速的处理,使用MCEE(多核执行环境)体系结构,MCEE体系结构使得报文能够以高速并行地处理。使用MCEE模型的软件数据面子系统被配置为将接收到的数据报文分散(spray)到专用于处理这些报文的各个核。每个核处理它所接收的报文并且将报文转发到出站端口。在使用MCEE体系结构的情况下,报文处理被显著地加速,因为它使得能够进行报文的并行处理以及将报文分布到专用于处理报文的许多处理资源。

[0057] 在某些实施例中,实现软件数据面子系统的软件被设计并且构建为使得它是便携式的并且能够在不同的处理器体系结构(例如Intel体系结构、ARM设备、PowerPC设备等)上执行。在不同的处理器体系结构上运行的不同软件数据面子系统可以由相同的控制面子系统104控制。

[0058] 在某些实施例中,软件数据面子系统108被配置为提供由传统基于线卡的转发机制提供的所有功能性。可以由软件数据面子系统提供的服务或者功能性的示例包括但不局限于,L2交换、L3转发Ipv4/6、MPLS处理、无状态的和有状态的防火墙功能性、网络地址转换(NAT)、有状态的网络附属存储(NAS)、分级QoS(服务质量)、Q-in-Q、对互联网协议安全(IPsec)的支持、通用路由封装(GRE)、虚拟可扩展LAN(VXLAN)、用于数据流的分析和管理的各种工具、深度报文检查(DPI)、最长报文匹配(LPM)报文查找、虚拟私有网络(VPN)等。由于软件数据面子系统被实现为软件,所以可以通过开发新的软件数据面子系统或者使用附加的功能性更新现有的软件数据面子系统来容易地添加附加的功能性,并且这些新的或者更新后的软件数据面子系统然后可以插入到分布式路由器中。

[0059] 可以使用通用CPU、网络接口卡(NIC)、为联网应用而设计的商用硅和/或FPGA来实现硬件数据面子系统110。可以使用的通用CPU的示例包括但不限于来自**Intel®**的x86处理器、**AMD®**处理器、**TI®**处理器等。可以使用的NIC的示例包括但不限于来自**Intel®**的**Fortville®**网络适配器和控制器、来自**Broadcom®**的**Cumulus®**、来自**Cavium®**的**LiquidIO®**或者其他。可以使用的商用硅的示例包括但不限于来自**Intel®**的Red Rock**Canyon®**(RRC)交换技术、来自Broadcom的**Tomahawk®**等等。**Tomahawk®**是可以用于实现诸如基于PCIe的虚拟私有互换(vPE)的各种转发功能或者支持多协议标签交换(MPLS)等等的高级交换技术的示例。

[0060] 在一些实施例中，硬件数据面子系统110可以提供与由软件数据面子系统提供的那些功能性类似的功能性。在一些其他的实施例中，硬件数据面子系统可以提供对由软件数据面子系统提供的功能性进行补充的功能性。在再有的其他的实施例中，硬件数据面子系统可以提供没有由可用的软件数据面子系统提供的一个或多个功能。在一些实例中，硬件数据面子系统可以从软件数据面子系统卸载某些功能性，并且对于较不处理密集(processing-intensive)的、重复的和/或频繁的功能提供硬件速度切换和处理。在一些实例中，硬件数据面子系统110可以卸载在软件数据面子系统108中较不频繁存在的复杂任务。

[0061] 使用软件数据面子系统108和硬件数据面子系统110实现数据面层为由数据面层执行的处理提供期望的灵活性。在一个实施例中，硬件数据面子系统可以用于使用特征特定的硬件(feature specific hardware)为重复的并且较不密集的联网功能提供硬件速度切换，同时维持在诸如通用处理器的便宜商用硬件组件上执行的软件数据面子系统中执行更加复杂并且较不频繁的功能性的通用性(versatility)和成本效益。

[0062] 如上面所描述的，用来实现硬件数据面子系统的CPU、NPU、商用硅和NIC(统称为“转发硬件”)可以由不同的供应商提供。控制面子系统因此可以同时控制来自多个供应商的硬件数据面子系统。从控制面子系统的视角，它在与软件数据面子系统或者硬件数据面子系统通信时不做区分；控制面子系统使用控制信道(例如，使用JSON编码的ZeroMQ传输)与数据面子系统通信。

[0063] 提供转发硬件的供应商可以提供各种供应商特定的API，用于控制转发硬件以及与转发硬件交互。例如，Broadcom提供用于控制Broadcom的Tomahawk商用硅并且与其交互的API。因此，由硬件数据面子系统从控制信道接收的消息需要转换成转发硬件供应商特定的API。因此关于每个转发硬件提供硬件抽象层(HAL)，其中HAL被配置为将经由控制信道接收的消息转换成转发硬件供应商特定的API。

[0064] 图2描绘根据本发明的实施例的具有HAL的示例硬件数据面子系统200。如图2中所示，HAL 202充当控制信道112与用来实现硬件数据面子系统的转发硬件204之间的接口。HAL层202负责将来自控制面的控制信道的配置、策略和状态信息转换到硬件数据面。HAL 202被配置为接收来自控制信道112的消息并且将它们转换成由转发硬件204的供应商提供的API。在相反的方向上，转发硬件204可以使用供应商特定的API提供信息(例如，网络状况)到HAL 202，并且HAL 202然后可以被配置为将这些转换成由控制信道112使用的格式的消息。

[0065] HAL 202可以被配置为提供各种不同的消息传递格式与各种供应商特定的API之间的转换。例如，HAL 202可以支持基于各种消息传递或者总线协议(诸如REST、YANG、ZeroMQ和/或JSON等等)的接口。

[0066] HAL 202因此提供到底层转发硬件的标准化接口，而不管转发硬件的类型以及提供转发硬件的供应商。在一些实例中，HAL 202可以提供存贮(inventory)或者发现底层转发硬件204的API和功能的能力。HAL的使用增加了分布式路由器体系结构的灵活性和互操作性。控制面子系统可以与可能来自不同供应商的不同硬件数据面子系统一起操作。由一个供应商(例如，Broadcom)提供的硬件数据面子系统可以切换到由另一个供应商(例如，Cavium)提供的硬件数据面子系统，而不需要对控制信道112或者对控制面子系统104做出

任何改变。HAL 202可以获取来自通用组件(例如,来自控制面控制信道)的输入并且将它们转换成转发硬件供应商特定的API。

[0067] 图7描绘示出根据本发明的实施例的被执行以用于对分布式路由器进行编程的处理的简化流程图700。在图7中描绘的特定系列的处理步骤不旨在是限制性的。应当领会到可以按照与图7中描绘的不同的次序执行处理步骤,并且不是图7中描绘的所有步骤都需要被执行。

[0068] 在702处,分布式路由器的用户可以使用分布式路由器的控制器提供指令。在一个实施例中,控制器可以提供可由用户使用以提供指令的接口。指令也可以经由控制器所提供的命令行接口(CLI)提供。指令可以对应于用户想要针对一个或多个网络配置的网络策略。例如,网络策略可以与防火墙、QoS、NAT等有关。

[0069] 在704处,指令从控制器传送到一个或多个控制面子系统。可以使用各种不同的协议将指令从控制器传送到控制面子系统。例如,如果控制面子系统是厚边界设备的一部分,那么诸如NETCONF的协议可以被用于将指令从控制器传送到控制面子系统。如果控制面子系统是薄边界设备的一部分,那么诸如OpenFlow的协议可以被用来将指令从控制器传送到控制面子系统。也可以使用各种其他的协议。

[0070] 在706处,接收指令的控制面子系统可以构建用于该指令的数据模型并且将指令信息存储在本地数据结构中并且将指令转换成将要在数据面层处执行的一个或多个动作。动作可以与将要在数据面层处针对从控制器接收的网络策略指令实现或者更新的一个或多个策略有关。动作可以与将要在数据面层处针对从控制器接收的网络策略指令执行的对转发信息或者配置信息的更新有关。在某些实施例中,控制面子系统可以存储用来将从控制器接收的指令映射到将要在数据面层处执行的一个或多个动作和/或策略的信息。控制面的配置可以包括路由协议配置、防火墙规则、诸如SNMP和NetFlow的管理服务等。

[0071] 在708处,控制面子系统生成与一个或多个动作对应的一个或多个消息并且使用控制信道将一个或多个消息传送到一个或多个数据面子系统。例如,在一个实施例中,控制面子系统可以生成与一个或多个动作对应的一个或多个JSON编码的消息,并且使用ZeroMQ传输总线将消息传送到一个或多个数据面子系统。

[0072] 708中的消息可以被发送到一个或多个软件数据面子系统和/或一个或多个硬件数据面子系统。当经由控制信道接收到一个或多个消息时,在710处,软件数据面子系统被配置为执行与一个或多个动作对应的处理。动作可以对软件数据面进行编程。例如,动作可以用来以编程的方式更新由软件数据面用来转发数据报文的转发信息和/或配置信息。动作也可以编程或者更新由软件数据面子系统提供的服务(例如,防火墙,DPI等)。例如,动作可以使得由数据面子系统存储的关于数据面子系统所提供的服务的信息被存储和/或更新。

[0073] 如果一个或多个消息由硬件数据面子系统接收,则在712处,与硬件数据面子系统的转发硬件对应的HAL层可以将由消息指示的一个或多个动作转换成一个或多个转发硬件供应商特定的API调用。在714处,在712中确定的API调用然后被执行。执行API对硬件数据面子系统进行编程。例如,动作可以用来以编程的方式更新由硬件数据面用于转发数据报文的转发信息和/或配置信息。动作也可以编程或者更新由硬件数据面子系统提供的服务(例如,MPLS转发)。例如,动作可以使得关于由硬件数据面子系统提供的服务的信息被存储

和/或更新。

[0074] 下面的用例帮助例示图7中描绘并且在上面描述的处理。在该示例中，用户可能想要将端口eth0上的IP地址设置为1.1.1.1/24。用户可以使用控制器102提供适当的指令到控制面子系统104。这可以使用NETCONF、命令行接口等完成。接下来，作为响应，控制面子系统104可以构建数据模型和本地数据结构，并且可以将指令转换成将要传送到一个或多个数据面的一个或多个动作。控制面子系统104然后可以使用具有JSON编码的ZeroMQ传输机制(例如，具有ZeroMQ报头和JSON接口的报文)经由控制信道将与动作对应的消息发送到一个或多个数据面子系统。硬件数据面子系统可以接收由控制面子系统104发送的消息。该消息报文将来到硬件数据面子系统的HAL层。HAL然后对0mQ/JSON消息进行解构和反序列化，并且调用转发硬件供应商特定的API来设置接口的IP地址(例如，对于Broadcom转发硬件，可以调用Broadcom API set_interface(1.1.1.1))。以这种方式，指令从控制器流动到数据面子系统，在那里相应的动作被执行。

[0075] 图3是分布式路由器300的简化框图，该简化框图例示了根据本发明的实施例的体系结构的分布式性质。如图3中描绘的，分布式路由器300包括控制器302，控制器302与多个控制面子系统(诸如子系统“控制面1”、“控制面2”等)通信。因此在控制器与控制面子系统之间可以存在一对多关系。每个控制面子系统继而可以控制一个或多个数据面子系统，这一个或多个数据面子系统可以包括软件数据面子系统和硬件数据面子系统中的一个或多个。因此，在控制面子系统与数据面子系统之间可以存在一对多关系。

[0076] 在某些实施例中，路由器300的各个子系统可以分布在网络中的各种设备之间。例如，在图3中描绘的实施例中，控制器可以在第一设备302上执行，控制面子系统_1可以在第二设备304上执行，并且控制面子系统_2可以在第三设备306上执行，其中第一设备302、第二设备304和第三设备306是不同的设备并且可以位于不同的地理位置处。设备302可以经由一个或多个网络与设备304和设备306在通信上耦合。控制面子系统因此可以位于托管控制器的计算系统或者设备远程的计算系统或者设备上。

[0077] 如图3中所示，控制面子系统1控制包括数据面子系统1和数据面子系统2的多个数据面子系统。数据面子系统1(可以是软件数据面子系统)可以由第四设备308执行。数据面子系统2(可以是硬件数据面子系统)可以位于第五设备310上，其中第二设备304、第四设备308和第五设备310是不同的设备并且可以位于不同的地理位置处。设备304可以经由一个或多个网络在通信上到达设备308和设备310。控制面子系统因此可以位于托管或包括数据面子系统的计算系统或者设备远程的计算系统或设备上。

[0078] 因此，在某些实施例中，路由器300的各个子系统或者组件可以跨彼此处于远程的多个设备而分布。然而，不要求路由器300的各个子系统必须跨越多个设备或者计算系统而分布。在某些实施例中，路由器300的子系统中的一些可以共同位于相同的物理设备上，而其他可以分布在多个远程的物理设备或者计算系统上。在一些实施例中，路由器300的用户提供有配置路由器300的各个子系统的位置的能力，包括控制器的位置、各个控制面子系统的位置以及各个数据面子系统的位置。

[0079] 分布式路由器的不同层位于不同物理设备上的能力使得控制器与控制面子系统之间以及控制面子系统与数据面子系统之间的关系能够是动态的。例如，关于控制器和控制面子系统，可以相对控制器远程地动态添加或者移除控制面子系统，并且控制器能够在

具有动态改变的情况下操作。由控制器控制的控制面子系统的位置可以逻辑地和物理地分散在不同的物理设备中。关于控制面子系统和数据面子系统，数据面子系统可以动态地添加或者移除。例如，提供增强功能性和服务的软件数据面可以动态地添加到分布式路由器。由控制面子系统控制的数据面子系统的位置可以逻辑地和物理地分散在不同的物理设备中。

[0080] 图4描绘示出根据本发明的实施例的当接收到数据报文时由分布式路由器执行的处理的简化流程图400。图4中描绘的处理可以在由一个或多个处理单元(例如，处理器、核)执行的软件(例如，代码、指令、程序)、硬件或者其组合中实现。软件可以存储在存储器中(例如，非临时性计算机可读存储介质上，诸如存储设备上)。在图4中描绘的特定系列的处理步骤不旨在是限制性的。应当领会到可以按照与图4中描绘的不同的次序执行处理步骤，并且不是图4中描绘的所有步骤都需要被执行。在一个实施例中，图4中描绘的处理可以由数据面子系统执行。

[0081] 在402处，数据报文由数据面子系统经由数据面子系统的输入端口接收。数据面子系统可以是软件数据面子系统或者硬件数据面子系统。输入端口可以是逻辑端口(或者说软件端口)或者物理端口。

[0082] 在404处，数据面子系统从接收到的报文中提取内容。例如，数据面子系统可以提取接收到的数据报文的报头。所提取的信息可以包括数据报文的源和目的地IP地址。

[0083] 在406处，数据面子系统确定它是否可以转发数据报文。在一个实施例中，406中的确定可以基于在404中提取的报文的内容、数据面子系统可用的转发信息以及基于应用任何可适用的转发策略(例如，如应用于由所配置的路由协议和设备策略的当前状态定义的转发信息库(Forwarding Information Base))而做出。转发信息可以由数据面子系统在接收数据报文之前接收并且由数据面子系统存储。在某些实施例中，转发信息可以从路由协议信息和网络状态中得出。与由数据面子系统提供的服务有关的信息也可以在406中使用以确定数据面子系统是否能够处理并且转发接收到的数据报文。

[0084] 如果数据面子系统在406中确定它不能转发接收到的数据报文，那么在408处丢弃该数据报文。可能存在数据面子系统不能转发数据报文的各种原因，诸如，路由协议不具有到下一跳的有效路径、路径不可达、由于诸如防火墙规则的网络策略，等等。

[0085] 如果数据面子系统在406中确定它可以处理接收到的数据报文，那么处理继续进行410。在410处，数据面子系统确定用来从数据面子系统转发数据报文的输出端口。输出端口可以是逻辑端口或者物理端口。软件端口可以在与数据面子系统共同驻留的通用CPU上运行，或者它可以在物理上分离的通用计算平台上。

[0086] 在一个实施例中，410中的确定基于在404中提取的报文信息并且基于数据面子系统可用的转发信息。作为410中处理的一部分，数据面子系统可以从在404中提取的报文内容中确定数据报文的目的地IP地址，并且对目的地地址执行最长前缀匹配，并且然后基于数据面子系统可用的转发信息确定用于转发数据报文的最优路径。然后可以基于最优路径确定输出端口。

[0087] 作为410中处理的一部分，数据面子系统也可以应用可以适用于数据报文的各种策略和服务。例如，如果已经指定防火墙策略并且适用于报文，那么可以执行与防火墙服务有关的处理。也可以针对由数据面子系统提供的其他服务定义策略，这些服务诸如网络地

址转换(NAT)、基于策略的路由、服务质量(QoS)、Q-in-Q、网络安全策略、深度报文检查(DPI)、最长报文匹配(LPM)等。

[0088] 在412处,数据报文被转发到在410中确定的输出端口。输出端口可以在接收数据报文的相同数据面子系统上,或者可以在不同的数据面子系统上。如果输出端口在不同的数据面子系统上,那么数据报文从接收数据报文的数据面子系统转发到包含输出端口的数据面子系统。

[0089] 在414处,使用在410中确定的输出端口将数据报文转发到下一跳。在报文去往远程系统的情况下,转发报文的数据面子系统可以将报文封装在传输协议(例如,VXLAN)中以用于通过物理网络运送。

[0090] 在某些实例中,接收数据报文的数据面子系统可能不能够执行为那个报文所请求的全套服务。在该场景中,数据面子系统(硬件的或者软件的)可以由控制面子系统配置以创建服务链,其中创建数据面子系统的链以用于执行多个服务。例如,在一个实施例中,可以创建服务链,其中第一数据面子系统被配置为执行第一服务并且第二数据面子系统被配置为执行用于数据报文的第二服务。因此,在一个场景中,当第一数据面子系统接收数据报文时,它执行第一服务并且然后将报文转发到第二数据面子系统。第二数据面子系统然后执行用于报文的第二服务,确定用于数据报文的输出端口,并且然后经由所确定的输出端口将报文转发到下一跳。服务链可以包括多个数据面子系统。

[0091] 在第二场景中,当第一数据面子系统接收到数据报文时,它执行第一服务并且然后将报文或者第二服务所需要的报文的一部分转发到第二数据面子系统。第二数据面子系统然后执行用于报文的第二服务并且将报文返回到第一数据面(或者将第二服务已经被执行的信号发送到第一数据面子系统)。第一数据面子系统然后确定用于数据报文的输出端口,并且经由所确定的输出端口将报文转发到下一跳。

[0092] 例如,用户可以具有架顶式(TOR,Top-Of-Rack)交换机并且可能想要与TOR交换机组合运行有状态的防火墙。TOR交换机可以包括第一数据面子系统并且有状态的防火墙可以由TOR交换机远程的第二数据面子系统实现。第二数据面子系统可以例如由云中的虚拟机执行。控制面子系统将通过流条目服务链(flow entry service chain)对TOR中的数据面子系统进行编程以将流量中需要有状态的防火墙的一部分发送到远离TOR运行的第二数据面子系统以用于处理。该流量然后可以由第二数据面子系统转发到下一跳。

[0093] 作为另一个示例,使用多核NPU实现的硬件数据面子系统可以能够执行多协议标签交换(MPLS)。因此,如果接收到的数据报文必须使用MPLS进行转发,那么硬件数据面子系统能够处置报文转发并且转发可以在硬件层处发生而不需要来自任何其他数据面子系统的任何帮助。然而,如果接收到的数据报文还要针对有状态的防火墙或者针对深度报文内省(DPI,Deep Packet Introspection)服务而被处理,那么硬件数据面可以将报文发送到被配置为提供这些服务的软件数据面子系统。一旦报文已经由软件数据面子系统针对这些另外的服务处理,数据报文就可以由软件数据面子系统返回到硬件数据面子系统,并且硬件数据面子系统然后可以使用MPLS将报文转发到下一跳。类似的框架也可以用于提供QoS服务,其中硬件数据面子系统将接收到的数据报文转发到软件数据面子系统用于执行与QoS有关的处理,并且然后数据报文由硬件数据面子系统转发。

[0094] 服务链接使得多个服务能够可能由不同的分布式数据面子系统(硬件数据面子系

统或者软件数据面子系统)在数据报文上执行。路由器100的分布式性质使得能够以灵活的方式启用并且执行服务链接。分布式路由器体系结构允许通用性,使得硬件数据面可以利用一个或多个软件数据面子系统或者其他硬件数据面子系统的能力。类似地,软件数据面子系统可以利用一个或多个其他软件数据面子系统或者其他硬件数据面子系统的能力。

[0095] 在某些实施例中,可以预先处理到达与硬件数据面子系统相关联的端口的数据报文以确定该报文或者该报文的某些方面是否应当使用硬件数据面子系统进行处理或者该报文是否应当被传输到软件数据面子系统以用于进一步的处理。如果该报文应当被传输到软件数据面子系统以用于进一步的处理,则数据报文可以被转发到软件数据面子系统,否则该报文可以完全由硬件数据面子系统处理。

[0096] 图5例示根据本发明的实施例的用于分布式路由器的不同组件的处理资源的示例实现和使用。路由器500包括控制面子系统518、软件数据面子系统520和硬件数据面子系统522的实例。图5中描绘的各个子系统可以彼此远程的放置并且经由一个或多个网络彼此耦合。图5中描绘的实施例仅是示例;可以实现若干其他配置而不背离本公开的范围。例如,分配给控制面子系统或者数据面子系统面的核的数量可以是静态的或者可变的。

[0097] 控制面子系统518分配有处理器核C0 502、C1 504和C2 506并由这些处理器核执行。控制面子系统518经由控制信道与软件数据面子系统520和硬件数据面子系统522通信。软件数据面子系统520分配有核C3 508、C4 510和C5 512并由这些核执行。分配给子系统的核可以来自一个或多个处理器。

[0098] 软件数据面子系统520可以使用诸如PCIe互连的互连514耦合到硬件数据面子系统522。在一些实例中,硬件数据面子系统522可以包括用于接收以及转发数据报文的(一个或多个)硬件网络元件516。

[0099] 在一个实施例中,控制面子系统518和/或软件数据面子系统520可以在由路由器500执行的虚拟机内执行。分配给这些虚拟机的处理、存储器和联网资源可以取决于需求动态地改变。另外的控制面子系统或者软件数据面子系统可以例如通过执行用于将要添加的子系统的另外的虚拟机而添加到路由器500。新添加的子系统然后可以与现有的子系统通信。以这种方式,控制面子系统和软件数据面子系统可以动态地添加到路由器500。

[0100] 另外的硬件数据面子系统也可以被添加到路由器500。新添加的硬件数据面子系统然后可以使用控制信道与控制面子系统动态地通信以及使用PCIe互连514与软件数据面子系统动态地通信。以这种方式,可以动态地添加(或者移除)路由器500的各个子系统。

[0101] 图6是根据本发明的实施例的可以用来执行分布式路由器的各个组件或者子系统的计算系统或者设备600的简化框图。在一些实施例中,计算系统600被配置为实现上面描述的方法中的任何方法。例如,像计算机系统600这样的一个或多个计算机系统可以用来执行分布式路由器的一个或多个子系统,诸如控制器102、一个或多个控制面子系统104以及一个或多个软件数据面子系统108。

[0102] 计算机系统600可以是各种类型的,包括但不限于个人计算机、便携式计算机、工作站、网络计算机、大型机、自助服务终端、PDA、蜂窝电话或者任何其他数据处理系统。由于计算机和网络的不断变化的性质,图6中描绘的对计算机系统600的描述仅旨在作为具体示例,以用于例示计算机系统的优选实施例的目的。比图6中描绘的系统具有更多或者更少组件的许多其他配置是可能的。

[0103] 计算机系统600被示为包括可以经由总线605以电气方式耦合的硬件元件。硬件元件可以包括一个或多个处理器610、一个或多个输入设备615、一个或多个输出设备620、通信子系统630和存储子系统640。总线子系统605提供让计算机系统600的各个组件和子系统按照意图彼此通信的机制。虽然总线子系统605示意性地示为单个总线，但是总线子系统的替代实施例可以利用多个总线。

[0104] 处理器610表示计算机系统600的处理资源并且可以包括但不局限于一个或多个通用处理器和/或一个或多个专用处理器(诸如数字信号处理芯片、图形加速处理器等)。处理器610可以包括一个或多个多核处理器。

[0105] 输入设备615可以包括用于提供输入到计算机系统600的一个或多个不同机制，诸如但不限于鼠标、键盘、触摸板、平板等。输出设备620可以包括用于从计算机系统600输出信息的一个或多个不同机制，诸如但不限于显示单元、打印机等。

[0106] 计算机系统600也可以包括通信子系统630，通信子系统630便于往来计算机系统600的通信。通信子系统630可以包括但不限于调制解调器、网卡(无线或者有线)、红外通信设备、无线通信设备和/或芯片集(诸如Bluetooth®设备、802.11设备、WiFi设备、WiMax设备、蜂窝通信设施等)等。通信子系统630可以允许与网络、其他计算机系统和/或在本文描述的任何其他设备互换数据。在某些实施例中，通信子系统630可以包括根据上面的教导用来实现硬件数据面的转发硬件。

[0107] 存储子系统640提供用于存储可以由一个或多个处理器610执行的信息和代码(指令)的非临时性介质。例如，存储子系统640可以被配置为存储提供本发明的实施例的功能性的基本编程和数据构造。根据本发明的实施例，实现本发明的功能性的软件代码指令或者模块可以存储在存储子系统640中。这些软件模块可以由(一个或多个)处理器610执行。存储子系统640也可以提供用于存储根据本发明使用的数据的贮存库。存储子系统640可以包括存储器子系统642以及文件/盘存储子系统644。

[0108] 存储器子系统642可以包括许多存储器，诸如用于在程序执行期间存储指令和数据的主随机存取存储器(RAM)、在其中存储固定指令的只读存储器(ROM)、闪存存储器等。各种软件元件可以位于系统存储器642内，这些软件元件诸如操作系统646、设备驱动器、可执行库和/或其他代码，诸如一个或多个应用程序648，如这里所描述的，这些软件元件可以包括由各种实施例提供的计算机程序，和/或可以被设计为实现由其他实施例提供的方法，和/或配置由其他实施例提供的系统。

[0109] 文件存储子系统644为程序和数据文件提供持久的(非易失性)存储，并且可以包括硬盘驱动器、软盘驱动器连同相关联的可移除介质、压缩盘只读存储器(CD-ROM)驱动器、光学驱动器、可移除介质盒、本地和/或网络可访问的存储装置以及其他类似的存储介质。

[0110] 本文所使用的术语“机器可读介质”和“计算机可读介质”指参与提供使得机器以具体的方式操作的数据的任何非临时性介质。在使用计算机系统600实现的实施例中，各种计算机可读介质可以涉及提供指令/代码到(一个或多个)处理器610以用于执行，和/或可以被用来存储这种指令/代码。计算机可读介质可以采取诸如非易失性介质和易失性介质的许多形式。

[0111] 上面讨论的方法、系统和设备是示例。在适当的时候，各种实施例可以省略、替换或者添加各种过程或者组件。例如，在替代配置中，所描述的方法可以按照与所描述的不同

的次序来执行，和/或可以添加、省略和/或组合各种阶段。关于某些实施例而描述的特征可以在各种其他实施例中组合。可以以类似的方式组合实施例的不同方面和元件。技术在发展，并且因此元件中的许多是示例，这些示例并不将本公开的范围限制于这些具体示例。

[0112] 在描述中给出具体细节以提供对实施例的透彻理解。然而，实施例可以在没有这些具体细节的情况下实践。例如，众所周知的电路、处理、算法、结构和技术在没有不必要的细节的情况下示出，以便避免模糊实施例。本描述仅提供示例实施例，并且不旨在限制本发明的范围、可适用性或者配置。然而，对实施例的前面的描述将向本领域技术人员提供实现本发明的实施例的实现性描述。可以在元件的功能和布置中做出各种改变而不背离本发明的精神和范围。

[0113] 虽然已经描述了本发明的具体实施例，但是各种修改、替代、替代构造和等同物也涵盖在本发明的范围内。本发明的实施例不限于在某些具体的数据处理环境内的操作，而是可以自由地在多个数据处理环境内操作。另外，虽然已经使用特定系列的事务和步骤描述了某些实施例，但是对本领域那些技术人员应当清楚的是，本发明的范围不限于所描述的事务和步骤系列。虽然一些流程图将操作描述为顺序过程，但是这些操作中的许多可以并行地或者并发地执行。另外，可以重新布置操作的次序。过程可以具有没有包括在图中的另外的步骤。

[0114] 此外，虽然已经使用硬件和软件的特定组合描述了某些实施例，但是应当认识到，硬件和软件的其他组合也在本发明的范围内。本发明的某些实施例可以仅在硬件中实现、或者仅在软件(例如，代码程序、固件、中间件、微码等)中实现或者使用它们的组合实现。本文描述的各种过程可以以任何组合在相同的处理器或者不同的处理器上实现。因此，在组件或者模块被描述为被配置为执行某些操作的情况下，这种配置可以例如通过设计电子电路执行操作、通过对可编程电子电路(诸如微处理器)进行编程以执行操作(诸如通过执行计算机指令或者代码)或者其任何组合来完成。进程可以使用各种技术来通信，这些技术包括但不限于用于进程间通信的常规技术，并且不同的进程对可以使用不同的技术，或者相同的进程对可以在不同的时间使用不同的技术。

[0115] 因此，说明书和附图应在例示性而不是限制性的意义上看待。然而，将明显的是，可以在那里进行添加、删减、删除和其他修改和改变而不背离如在权利要求中阐述的更广泛的精神和范围。因此，虽然已经描述了具体的发明实施例，但是这些发明实施例不旨在是限制性的。各种修改和等同物都在下面的权利要求的范围内。

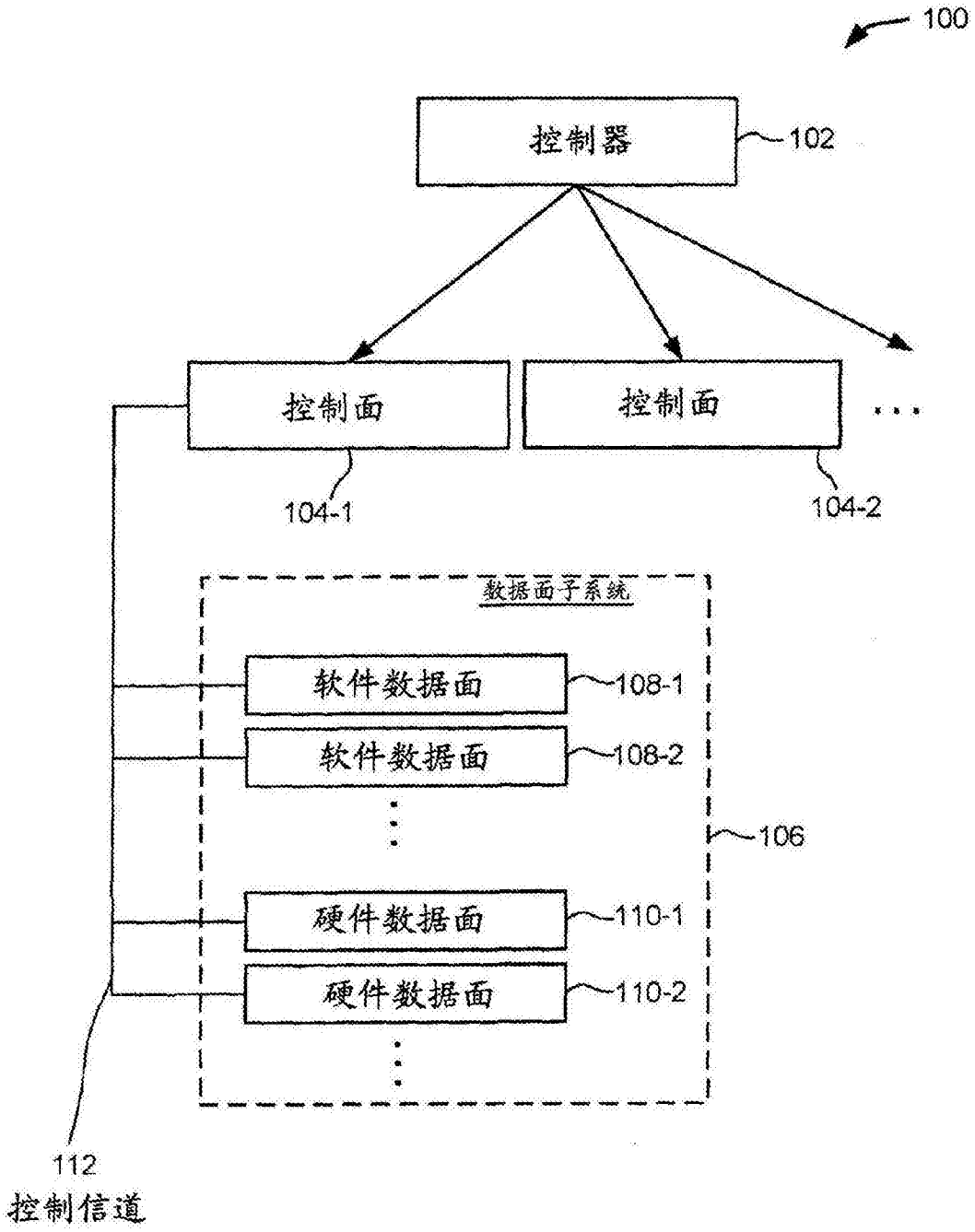

100
控制器
102
控制面
控制面
104-1
104-2
数据面子系统
软件数据面
108-1
软件数据面
108-2
106
硬件数据面
110-1
硬件数据面
110-2
112
控制信道

图1

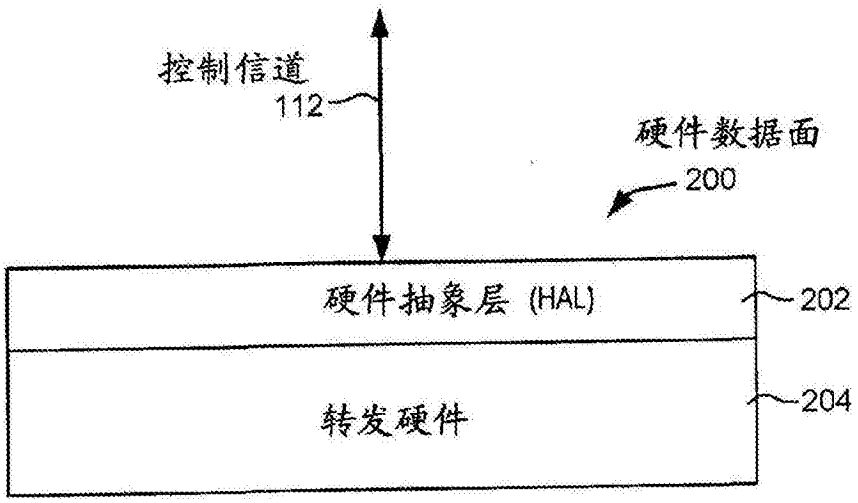


图2

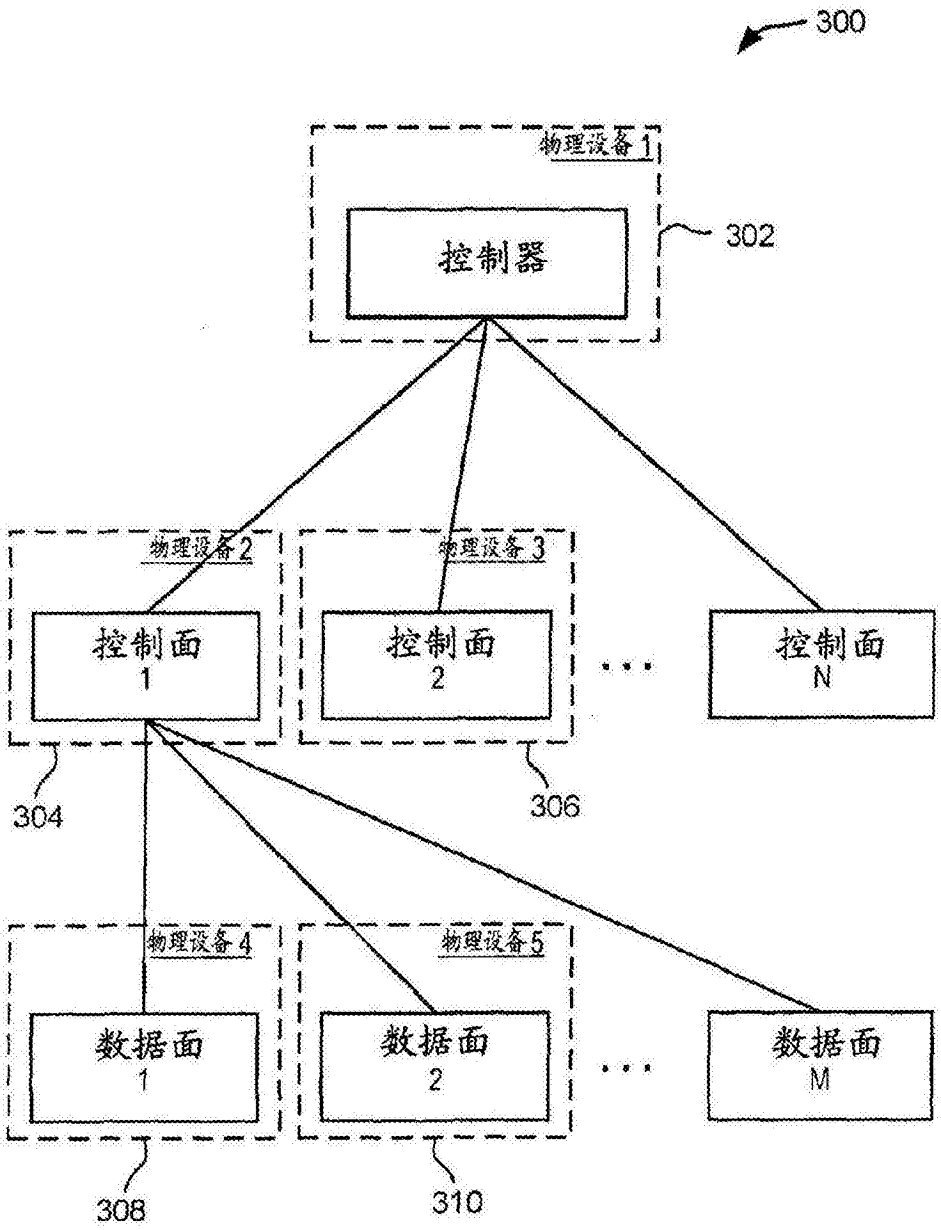

300
物理设备1
控制器
302
物理设备2
控制面
1
物理设备3
控制面
2
控制面
N
304
306
物理设备4
数据面
1
物理设备5
数据面
2
数据面
M
308
310

图3

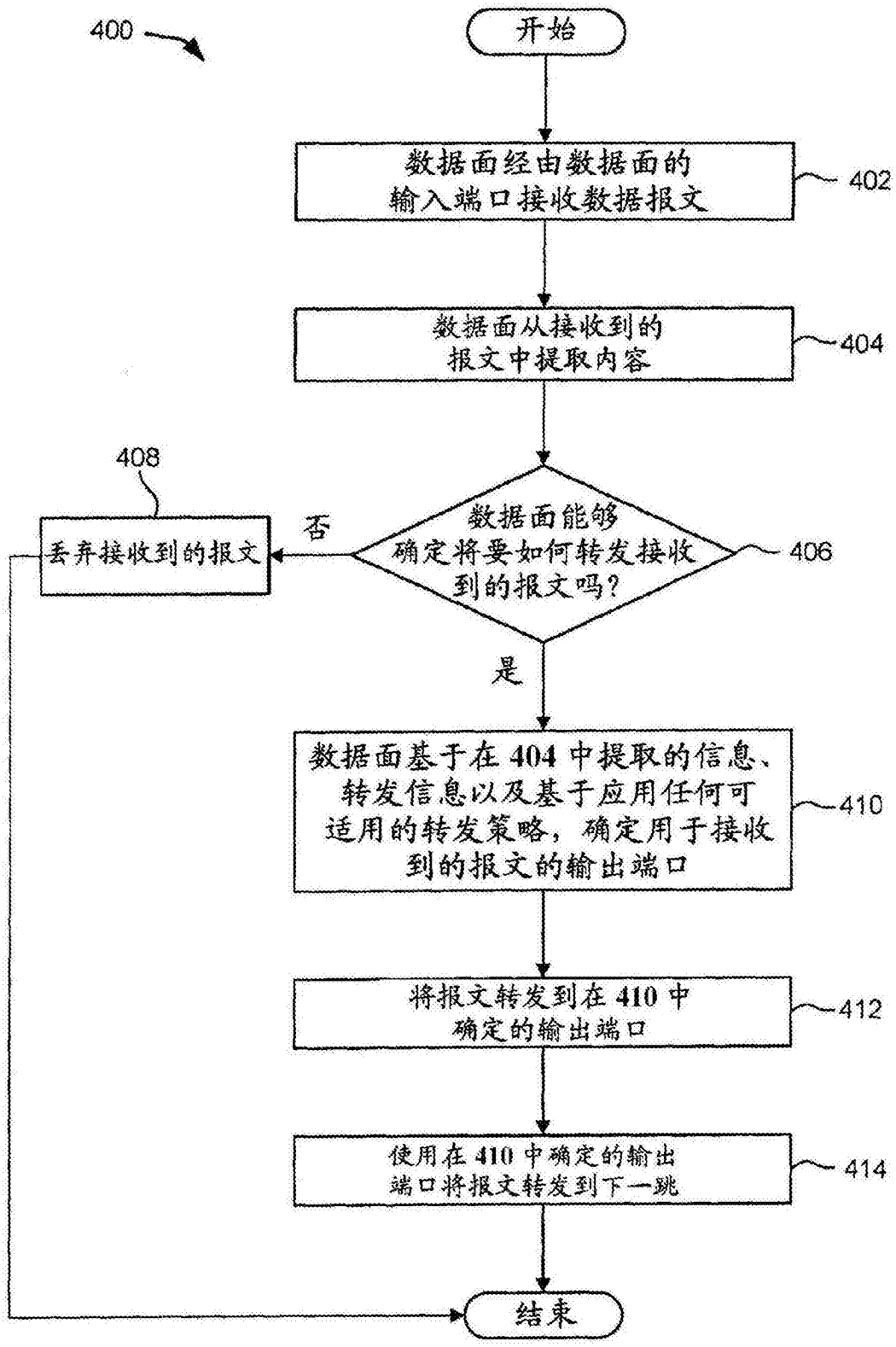

400
开始
数据面经由数据面的输入端口接收数据报文
402
数据面从接收到的报文中提取内容
404
408
丢弃接收到的报文
否
数据面能够确定将要如何转发接收到的报文吗?
406
是
数据面基于在 404 中提取的信息、转发信息以及基于应用任何可适用的转发策略，确定用于接收到的报文的输出端口
410
将报文转发到在 410 中确定的输出端口
412
使用在 410 中确定的输出端口将报文转发到下一跳
414
结束

图4

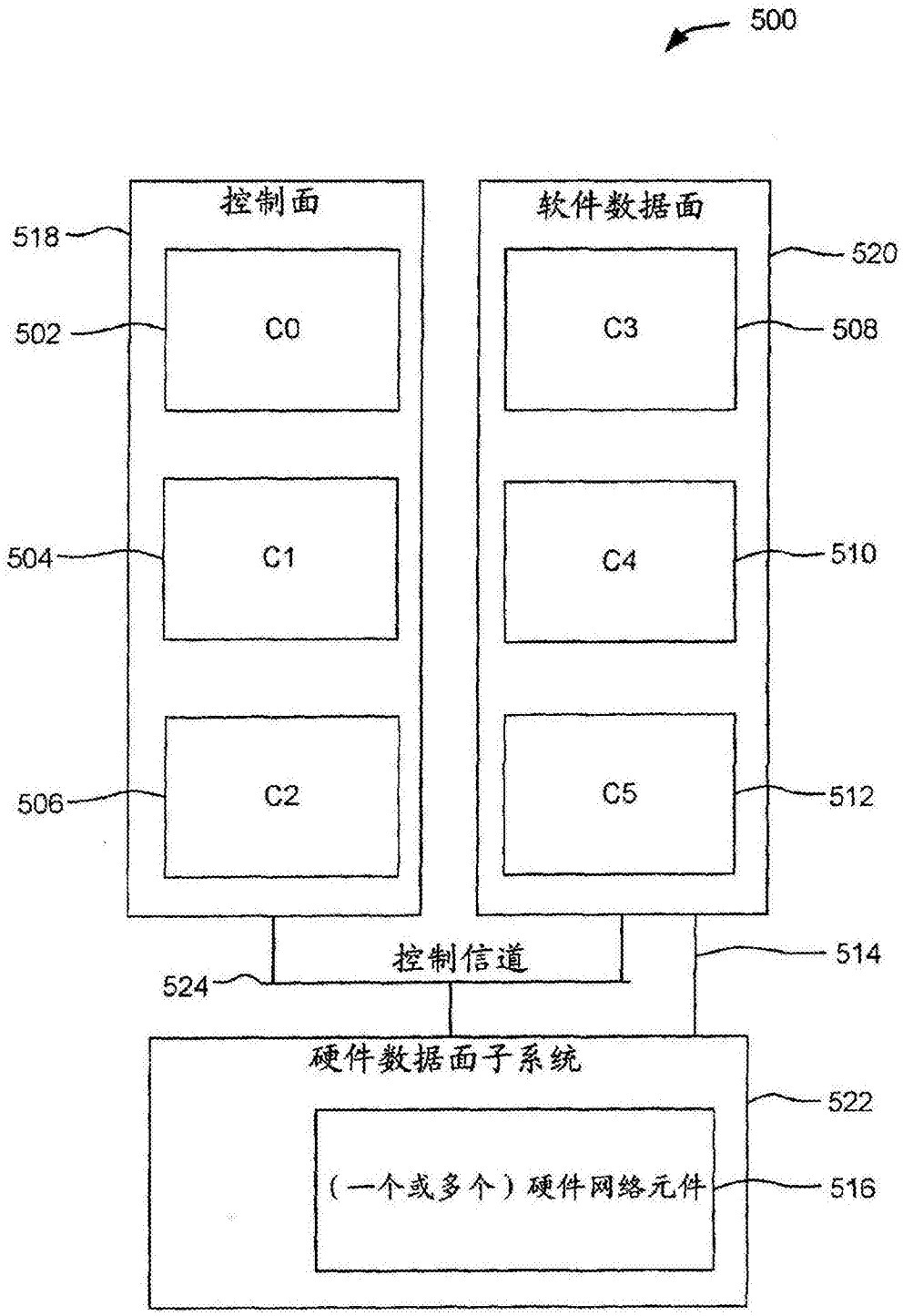

500
控制面
软件数据面
518
520
502
C0
C3
508
504
C1
C4
510
506
C2
C5
512
控制信道
514
524
硬件数据面子系统
522
（一个或多个）硬件网络元件
516

图5

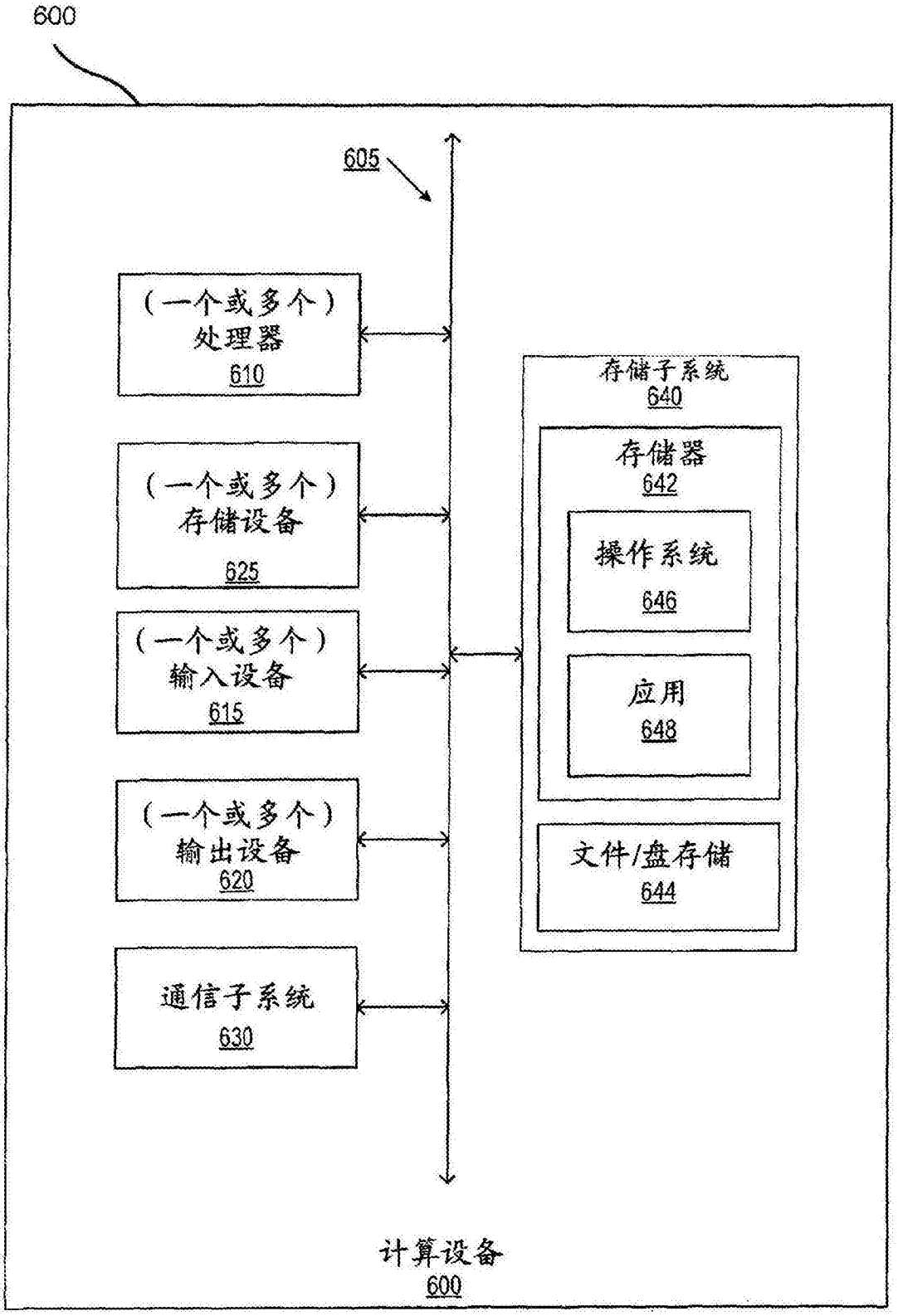

600
605
(一个或多个)
处理器
610
(一个或多个)
存储设备
625
(一个或多个)
输入设备
615
(一个或多个)
输出设备
620
通信子系统
630
存储子系统
640
存储器
642
操作系统
646
应用
648
文件/盘存储
644
计算设备
600

图6

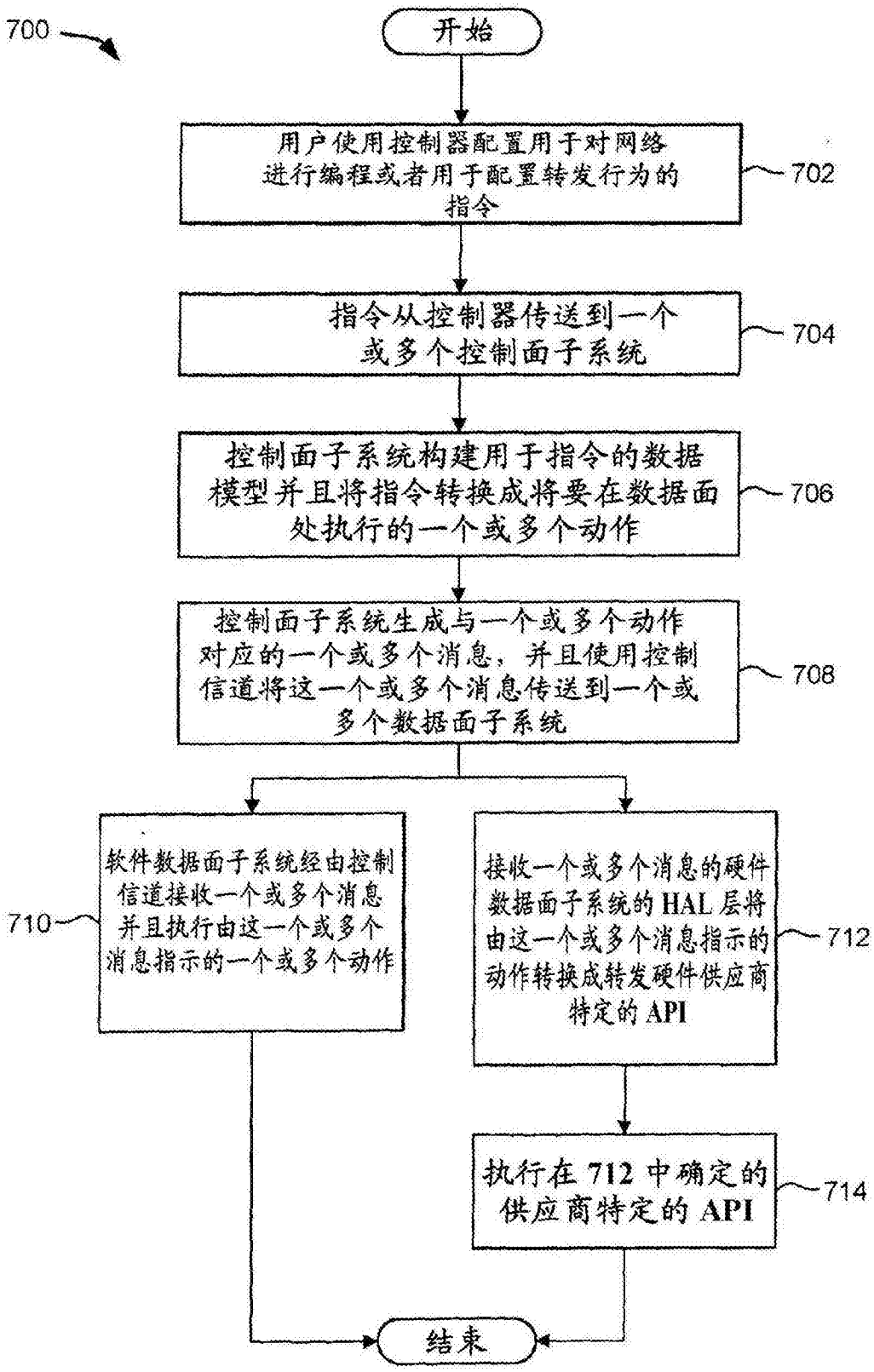

700
开始
用户使用控制器配置用于对网络进行编程或者用于配置转发行为的指令
702
指令从控制器传送到一个或多个控制面子系统
704
控制面子系统构建用于指令的数据模型并且将指令转换成将要在数据面处执行的一个或多个动作
706
控制面子系统生成与一个或多个动作对应的一个或多个消息，并且使用控制信道将这一个或多个消息传送到一个或多个数据面子系统
708
软件数据面子系统经由控制信道接收一个或多个消息并且执行由这一个或多个消息指示的一个或多个动作
710
接收一个或多个消息的硬件数据面子系统的HAL层将由这一个或多个消息指示的动作转换成转发硬件供应商特定的API
712
执行在712中确定的供应商特定的API
714
结束

图7