(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 102118319 B
(45)授权公告日 2013.09.18

(21)申请号 201110086357.2

(22)申请日 2011.04.06

(73)专利权人 杭州华三通信技术有限公司
地址 310053 浙江省杭州市高新技术产业开发区之江科技工业园六和路 310 号华为杭州生产基地

(72)发明人 李蔚

(74)专利代理机构 北京德琦知识产权代理有限公司 11018
代理人 谢安昆 宋志强

(51)Int.Cl.
H04L 12/891(2013.01)
H04L 29/12(2006.01)

(56)对比文件
CN 101815007 A,2010.08.25,
CN 101605104 A,2009.12.16,
CN 101977153 A,2011.02.16,
KR 20040074680 A,2004.08.26,
US 2010234042 A1,2010.09.16,

审查员 刘毅

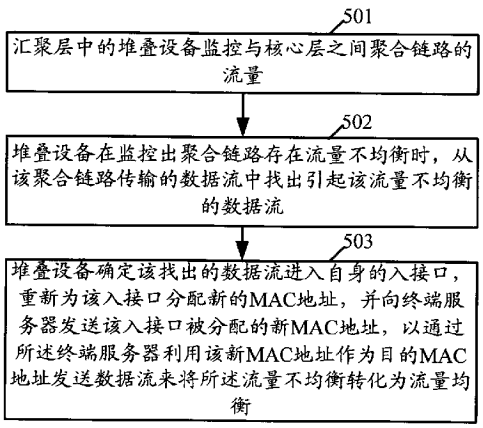
权利要求书2页 说明书7页 附图5页

(54)发明名称

流量负载均衡方法和装置

(57)摘要

本发明提供了流量负载均衡方法和装置。其中，该方法包括：汇聚层中堆叠设备监控其与核心层之间聚合链路的流量；所述堆叠设备在监控出聚合链路存在流量不均衡时，从该聚合链路传输的数据流中找出引起该流量不均衡的数据流；所述堆叠设备确定该找出的数据流进入自身的入接口，重新为该入接口分配新的 MAC 地址，并向终端服务器发送该入接口被分配的新 MAC 地址，以通过所述终端服务器利用该新 MAC 地址作为目的 MAC 地址发送数据流来将所述流量不均衡转化为流量均衡。采用本发明，能够实现汇聚层到核心层之间聚合链路中各成员链路负载均衡。



CN 102118319 B

1. 一种流量负载均衡方法，其特征在于，该方法包括：

汇聚层中堆叠设备监控其与核心层之间聚合链路的流量；

所述堆叠设备在监控出聚合链路存在流量不均衡时，从该聚合链路传输的数据流中找出引起该流量不均衡的数据流；

所述堆叠设备确定该找出的数据流进入自身的入接口，重新为该入接口分配新的 MAC 地址，并向终端服务器发送该入接口被分配的新 MAC 地址，以由所述终端服务器向其接入的接入设备发送目的 MAC 地址为所述新 MAC 地址的数据流；

所述堆叠设备的成员设备接收到所述接入设备转发的目的 MAC 地址为所述新 MAC 地址的数据流时，按照本地转发优先的原则从用于转发该数据流的聚合链路中选择出一条成员链路，通过该确定的成员链路向核心层转发该数据流；其中，所述接入设备通过以下步骤转发目的 MAC 地址为所述新 MAC 地址的数据流至所述堆叠设备的成员设备：所述接入设备将该数据流的目的 MAC 地址和负载均衡算法进行哈希运算，根据哈希运算结果从用于转发该数据流的聚合链路中选择出一条成员链路，通过该确定的成员链路向汇聚层转发该数据流。

2. 根据权利要求 1 所述的方法，其特征在于，所述聚合链路存在流量不均衡包括：

当聚合链路中一条成员链路的使用带宽超过第一设定阈值，且该成员链路与其他一条成员链路的使用带宽之差超过第二设定阈值时，确定该聚合链路存在流量不均衡。

3. 根据权利要求 2 所述的方法，其特征在于，所述从聚合链路传输的数据流中找出引起该流量不均衡的数据流包括：

统计出所述聚合链路中使用带宽超过第一设定阈值，且与其他一条成员链路的使用带宽之差超过第二设定阈值的成员链路传输的数据流；

按照选取大流量数据流的原则从统计出的数据流中选取设定值 N 条数据流，将选取的数据流确定为引起该流量不均衡的数据流。

4. 根据权利要求 1 至 3 任一所述的方法，其特征在于，所述堆叠设备确定该找出的数据流进入自身的入接口包括：

针对找出的每一数据流，所述堆叠设备确定该数据流的源 IP 地址，根据该确定的源 IP 地址通过单播逆向路径转发 URPF 查找到该数据流的入接口；

所述入接口为虚拟局域网 VLAN 虚接口。

5. 根据权利要求 1 至 3 任一所述的方法，其特征在于，所述堆叠设备向终端服务器发送该入接口被分配的新 MAC 地址包括：

所述堆叠设备在被分配了新 MAC 地址的入接口上向接入层发送免费地址解析协议 ARP 报文；接入层中接收到所述免费 ARP 报文的接入设备将所述免费 ARP 报文转发至其接入的终端服务器。

6. 一种流量负载均衡装置，其特征在于，该装置为汇聚层中的堆叠设备，具体包括：监控单元、查找单元、分配单元、以及级联成所述装置的多个成员设备；其中，

监控单元，用于监控所述装置与核心层之间聚合链路的流量；

查找单元，用于在所述监控单元监控出聚合链路存在流量不均衡时，从该聚合链路传输的数据流中找出引起该流量不均衡的数据流；

分配单元，包括：确定子单元、分配子单元和发送子单元，其中，

所述确定子单元，用于确定所述查找单元找出的数据流进入所述装置的入接口；

所述分配子单元，用于为所述确定子单元确定的入接口重新分配新的 MAC 地址；

所述发送子单元，用于向终端服务器发送所述确定子单元确定的入接口被分配的新 MAC 地址，以由所述终端服务器向其接入的接入设备发送目的 MAC 地址为所述新 MAC 地址的数据流；

每一成员设备在接收到接入层中接入设备转发的来自所述终端服务器的数据流时，按照本地转发优先的原则从用于转发该数据流的聚合链路中选择出一条成员链路，通过该确定的成员链路向核心层转发该数据流；

其中，来自所述终端服务器的数据流的目的 MAC 地址为所述新 MAC 地址，所述接入设备通过以下操作转发该数据流至所述成员设备：将作为该数据流目的 MAC 地址的新 MAC 地址和负载均衡算法进行哈希运算，根据哈希运算结果从用于转发该数据流的聚合链路中选择出一条成员链路，利用该确定的成员链路向汇聚层转发该数据流。

7. 根据权利要求 6 所述的装置，其特征在于，所述查找单元包括：

链路不均衡确定子单元，用于当聚合链路中一条成员链路的使用带宽超过第一设定阈值，且该成员链路与其他一条成员链路的使用带宽之差超过第二设定阈值时，确定该聚合链路存在流量不均衡。

8. 根据权利要求 7 所述的装置，其特征在于，所述查找单元还包括：

数据流查找子单元，用于统计出所述聚合链路中使用带宽超过第一设定阈值，且与其他一条成员链路的使用带宽之差超过第二设定阈值的成员链路传输的数据流，按照选取大流量数据流的原则从统计出的数据流中选取设定值 N 条数据流，将选取的数据流确定为引起该流量不均衡的数据流。

9. 根据权利要求 6 至 8 任一所述的装置，其特征在于，所述确定子单元用于针对找出的每一数据流，确定该数据流的源 IP 地址，根据该确定的源 IP 地址通过单播逆向路径转发 URPF 查找到该数据流的入接口；

所述入接口为虚拟局域网 VLAN 虚接口。

10. 根据权利要求 6 至 8 任一所述的装置，其特征在于，所述发送子单元在被分配了新 MAC 地址的入接口上向接入层发送免费地址解析协议 ARP 报文，以使接入层中接收到所述免费 ARP 报文的接入设备将所述免费 ARP 报文转发至其接入的终端服务器。

流量负载均衡方法和装置

技术领域

[0001] 本发明涉及通信技术，特别涉及流量负载均衡方法和装置。

背景技术

[0002] 随着企业对信息访问依赖性的增加，数据中心对企业日常业务影响也越来越大。一旦企业的数据中心出现故障，将对企业日常业务的正常运作造成极大的冲击，给企业带来巨大的损失。总的来看，数据中心出现故障时，企业的损失分为以下几个方面：对企业日常工作的冲击比如员工无法正常工作、直接财产损失，比如订单丢失、企业合作伙伴损失赔偿等、企业声誉的损失比如失去部分客户等。因此，可靠性已经成为衡量一个数据中心优劣的重要方面。

[0003] 目前，为了提高数据中心的可靠性，在数据中心网络中会采用双节点、双链路冗余组网方案。同时，为了提高网络设备、链路的利用率，通常会采用目前最先进的链路聚合技术和堆叠技术。下面分别对链路聚合技术和堆叠技术进行描述：

[0004] 链路聚合技术：

[0005] 链路聚合，也称为以太网链路聚合，其是将多个物理以太网端口聚合在一起构成一个逻辑上的聚合链路，也即，同一聚合链路内的多条物理链路被视为一条逻辑链路。其中，构成聚合链路的多条物理链路被称为成员链路。通过链路聚合，可以实现数据流量在聚合链路中各个成员链路之间分担，以增加带宽，并且，同一聚合链路的各个成员链路之间彼此动态备份，提高了连接可靠性。

[0006] 其中，聚合链路可分为两类：负载分担聚合链路和非负载分担聚合链路。目前，负载分担聚合链路是主流的应用方式。其中，负载分担聚合链路的聚合负载分担模式是可以选择性的被配置，通过改变聚合负载分担的模式可以灵活地实现聚合组流量的负载分担。

[0007] 堆叠技术：

[0008] 所谓堆叠技术，其核心思想是将多台设备通过堆叠口级联在一起，进行必要的配置后，虚拟化成一台“分布式设备”，为了便于描述，这个“虚拟设备”也称为堆叠设备，构成该堆叠设备的各个设备称为成员设备，各成员设备之间的连接称为堆叠链路，具体如图 1 所示。通过堆叠技术，可以实现多台成员设备的协同工作、统一管理和不间断维护。

[0009] 在堆叠技术中存在本地转发优先的概念。所谓本地转发优先具体为：堆叠设备中的成员设备接收的数据流优先从本设备的出接口转发，只有在本设备没有出接口的情况下，才会通过堆叠链路将接收的数据流发送至其他成员设备，由其他成员设备的出接口转发。参见图 2，图 2 为现有本地转发优先示意图。如图 2 所示，2 台交换机即交换机 1 和交换机 2 进行级联，构成堆叠设备。堆叠设备的下行和上行均支持链路聚合。堆叠设备中的成员设备比如交换机 1 接收的数据流会优先从本设备的出接口转发出去，当该交换机 1 没有出接口时，才会通过与另一成员设备比如交换机 2 间的堆叠链路将接收的数据流转发到交换机 2 的出接口，由该出接口转发。通过本地转发优先，能够保证数据流一般不经过堆叠链路，如此，可以要求堆叠设备中的成员设备无需设置比较多的堆叠口，降低了设备成本，

并且，由于数据流经过最少的设备转发，这减少了转发时延。

[0010] 以上对链路聚合技术和堆叠技术进行了描述，下面对采用链路聚合技术和堆叠技术的数据中心组网进行描述。

[0011] 参见图 3，图 3 为现有典型的数据中心组网示意图。图 3 所示的组网分为核心层、汇聚层、接入层三层结构。其中，汇聚层上的设备即设备 C 和设备 D 级联成堆叠设备，核心层上的设备即设备 A 和设备 B 级联成堆叠设备，而接入层没有采用堆叠技术。在图 3 中，各层设备（即接入层的接入设备 E 至接入设备 H、汇聚层的堆叠设备和核心层的堆叠设备）之间通过聚合链路连接（在图 3 以圆圈显示），其中，汇聚层的堆叠设备与核心层的堆叠设备之间的聚合链路至少为 1 条，图 3 为简单描述，仅示出了一条，记为 L11。

[0012] 在图 3 所示的组网中，接入层上的接入设备即接入设备 E 至接入设备 H 分别接入 1 个以上的终端服务器（图 3 未示出），汇聚层上的设备即设备 C 和设备 D 作为网关，汇聚层和核心层之间跑路由。这是目前最常见的组网方式。

[0013] 在该组网方式中，终端服务器发送的数据流先到达接入层。当接入层上的接入设备接收到该数据流时，将配置的负载均衡算法和该数据流的目的 MAC 地址做 Hash 运算，根据哈希运算结果从用于转发该数据流的聚合链路中选择成员链路（实质为选择用于转发该数据流的聚合出端口），之后通过该成员链路向汇聚层转发该数据流。

[0014] 但是，由于不同终端服务器、不同业务应用对应不同流量的数据流，甚至不同数据流的流量相差非常大，这就导致上述聚合链路中各个成员链路上的流量是不均衡的。当该不均衡流量传输至汇聚层中的堆叠设备时，由于堆叠设备的成员设备具有上述的本地转发优先特性，即，当汇聚层中堆叠设备的成员设备接收到数据流时优先通过聚合链路中自身的出接口对应的成员链路转发，这会导致该堆叠设备与核心层之间的聚合链路比如图 3 所示的聚合链路 L11 中成员链路的流量不均衡，比如，聚合链路 L11 中存在图 4 所示的一条成员链路比如设备 C 和设备 A 之间的链路 L_{CA} 满负荷过载丢包，其他成员链路比如设备 D 和设备 B 之间的链路 L_{DB} 轻载的情况，从而使汇聚层到核心层的聚合链路负载均衡成为一句空话，严重影响了用户服务质量，降低了整个系统的转发性能。

发明内容

[0015] 本发明提供了流量负载均衡方法和装置，以实现汇聚层到核心层之间的聚合链路负载均衡。

[0016] 本发明提供的技术方案包括：

[0017] 一种流量负载均衡处理方法，包括：

[0018] 汇聚层中堆叠设备监控其与核心层之间聚合链路的流量；

[0019] 所述堆叠设备在监控出聚合链路存在流量不均衡时，从该聚合链路传输的数据流中找出引起该流量不均衡的数据流；

[0020] 所述堆叠设备确定该找出的数据流进入自身的入接口，重新为该入接口分配新的 MAC 地址，并向终端服务器发送该入接口被分配的新 MAC 地址，以通过所述终端服务器利用该新 MAC 地址作为目的 MAC 地址发送数据流来将所述流量不均衡转化为流量均衡。

[0021] 一种流量负载均衡装置，该装置为汇聚层中的堆叠设备，具体包括：

[0022] 监控单元，用于监控所述装置与核心层之间聚合链路的流量；

[0023] 查找单元，用于在所述监控单元监控出聚合链路存在流量不均衡时，从该聚合链路传输的数据流中找出引起该流量不均衡的数据流；

[0024] 分配单元，包括：确定子单元、分配子单元和发送子单元，其中，

[0025] 所述确定子单元，用于确定所述查找单元找出的数据流进入所述装置的入接口；

[0026] 所述分配子单元，用于为所述确定子单元确定的入接口重新分配新的 MAC 地址；

[0027] 所述发送子单元，用于向终端服务器发送所述确定子单元确定的入接口被分配的新 MAC 地址，以通过所述终端服务器利用该新 MAC 地址作为目的 MAC 地址发送数据流来将所述流量不均衡转化为流量均衡。

[0028] 由以上技术方案可以看出，本发明中，汇聚层中堆叠设备监控与核心层之间聚合链路的流量，当监控出聚合链路存在流量不均衡时，从该聚合链路传输的数据流中找出引起该流量不均衡的数据流，确定该找出的数据流对应的入接口，重新为该入接口分配新的 MAC 地址，并向终端服务器发送该入接口被分配的新 MAC 地址，这样，当终端服务器后续向被分配了新 MAC 地址的入接口发送数据流时，就以该新 MAC 地址作为目的 MAC 地址发送，当接入设备接收到该数据流时，利用该新 MAC 地址从聚合链路中选择一成员链路，通过该选择的成员链路转发该数据流至汇聚层，而汇聚层中堆叠设备的成员设备接收到数据流时，依据本地转发优先原则从聚合链路中选择成员链路，通过该选择的成员链路向核心层转发该数据流。从概率上来说，接入设备依据新 MAC 地址确定的成员链路会与之前依据旧 MAC 地址（分配新 MAC 地址之前使用的 MAC 地址）确定的成员链路不同，这保证了接入设备转发的两种不同的数据流（即用新 MAC 地址作为目的 MAC 地址的数据流与用旧 MAC 地址作为目的 MAC 地址的数据流）到达汇聚层中堆叠设备的不同成员设备，基于堆叠设备中成员设备的本地转发优先特性，这会减轻了汇聚层中堆叠设备与核心层之间聚合链路中成员链路之间极端异常的流量不均衡现象，使该现象向好的方面即聚合链路中成员链路之间流量负载均衡转化。

附图说明

[0029] 图 1 为现有堆叠技术示意图；

[0030] 图 2 为现有本地转发优先示意图；

[0031] 图 3 为现有典型的数据中心组网示意图；

[0032] 图 4 为现有聚合链路中成员链路负载不均衡示意图；

[0033] 图 5 为本发明实施例提供的基本流程图；

[0034] 图 6 为本发明实施例应用示意图；

[0035] 图 7 为本发明实施例提供的装置结构图。

具体实施方式

[0036] 为了使本发明的目的、技术方案和优点更加清楚，下面结合附图和具体实施例对本发明进行详细描述。

[0037] 参见图 5，图 5 为本发明实施例提供的基本流程图。该图 5 所示的流程应用于接入层接入、汇聚层上的设备作为网关，汇聚层和核心层之间跑路由的组网方式。其中，接入层中接入设备根据来自终端服务器的数据流的目的 MAC 地址转发该数据流至汇聚层，汇聚层

中堆叠设备通过聚合链路向核心层转发接收的数据流。

[0038] 基于此，如图 5 所示，该方法包括：

[0039] 步骤 501，汇聚层中的堆叠设备监控与核心层之间聚合链路的流量。

[0040] 在本步骤 501 中，被监控的聚合链路为从所述堆叠设备至核心层之间的聚合链路中指定的至少一条聚合链路，其中，该指定操作可由用户根据经验、或者实际情况执行。

[0041] 步骤 502，堆叠设备在监控出聚合链路存在流量不均衡时，从该聚合链路传输的数据流中找出引起该流量不均衡的数据流。

[0042] 本步骤 502 中，当聚合链路中一条成员链路的使用带宽超过第一设定阈值，且与其他一条成员链路的使用带宽之差超过第二设定阈值时，确定该聚合链路存在流量不均衡。

[0043] 基于此，步骤 502 中的所述从该聚合链路传输的数据流中找出引起该流量不均衡的数据流包括：统计出所述聚合链路中使用带宽超过第一设定阈值，且与其他一条成员链路的使用带宽之差超过第二设定阈值的成员链路传输的数据流，按照选取大流量数据流的原则从统计出的数据流中选取设定值 N 条数据流，将选取的数据流确定为引起该流量不均衡的数据流。比如，堆叠设备监控出其与核心层之间的聚合链路 L1 中存在成员链路 L1a 的使用带宽超过第一设定阈值，且与其他一条成员链路比如 L1b 的使用带宽之差超过第二设定阈值，则，堆叠设备确定聚合链路 L1 存在流量不均衡，并统计出聚合链路 L1 中成员链路 L1a 上传输的数据流，从统计出的数据流中选取 N 条流量比较大的数据流，将选取的该 N 条数据流确定为引起该聚合链路 L1 流量不均衡的数据流。

[0044] 在上面描述中，第一设定阈值与第二设定阈值无关，两者在设定时，均可根据经验或者实际情况设置，本发明实施例并不具体限定。

[0045] 另外，在上面描述中，N 的取值也可根据经验或者实际情况设置，较佳地，本发明建议该 N 取值为 10。

[0046] 步骤 503，堆叠设备确定该找出的数据流进入自身的入接口，重新为该入接口分配新的 MAC 地址，并向终端服务器发送该入接口被分配的新 MAC 地址，以通过所述终端服务器利用该新 MAC 地址作为目的 MAC 地址发送数据流来将所述流量不均衡转化为流量均衡。

[0047] 本步骤 503 中，堆叠设备确定该找出的数据流进入自身的入接口包括：针对找出的每一数据流，所述堆叠设备确定该数据流的源 IP 地址，根据该确定的源 IP 地址通过单播逆向路径转发 (URPF) 查找到该数据流的入接口。

[0048] 其中，上述的入接口具体为 VLAN 虚接口。VLAN 虚接口的 MAC 地址为终端服务器通过接入设备向汇聚层转发数据流的目的 MAC 地址，该 MAC 地址是汇聚层中堆叠设备中的成员设备比如汇聚交换机创建该 VLAN 虚接口时为该 VLAN 虚接口分配的，其是可以修改的。也即，VLAN 虚接口的 MAC 地址是可以被重新分配的。

[0049] 另外，本步骤 503 中，所述堆叠设备向终端服务器发送该入接口被分配的新 MAC 地址是通过接入层中的接入设备实现的，具体包括：所述堆叠设备在被分配了新 MAC 地址的入接口上发送免费 ARP 报文；接入层中接收到所述免费 ARP 报文的接入设备将所述免费 ARP 报文转发至其接入的终端服务器。

[0050] 另外，本步骤 503 中，所述通过终端服务器利用新 MAC 地址作为目的 MAC 地址发送数据流来将所述流量不均衡转化为流量均衡是通过以下方式实现：终端服务器向其接入的

接入设备发送目的 MAC 地址为新 MAC 地址的数据流；所述接入设备接收到终端服务器发送的数据流后，将该数据流的目的 MAC 地址和负载均衡算法进行哈希运算，根据哈希运算结果从用于转发该数据流的聚合链路中选择出一条成员链路（比如，用于转发该数据流的聚合链路中各成员链路被分别编号，从该聚合链路中选择出编号为所述哈希运算结果的成员链路），通过该确定的成员链路向汇聚层转发该数据流；汇聚层中堆叠设备的成员设备接收到数据流时，按照本地转发优先的原则从用于转发该数据流的聚合链路中选择出一条成员链路，通过该确定的成员链路向核心层转发该数据流。

[0051] 从概率上来说，当接入设备依据新 MAC 地址确定成员链路时，会由于与之前依据的旧 MAC 地址（即被分配新 MAC 地址之前所采用的 MAC 地址）不同而导致确定的成员链路与之前确定的成员链路不同，进而保证了本次转发的数据流与之前的数据流到达汇聚层中堆叠设备的不同成员设备。这样，由于汇聚层中堆叠设备的成员设备具有本地转发优先的特性，当与之前不同的成员设备接收到数据流时，其会优先通过本设备的出端口向核心层转发接收的数据流，这改善了聚合链路中成员链路流量负载不均衡的现象，使这种现象向流量负载均衡转化。

[0052] 至此，完成图 5 所示的流程。

[0053] 为使图 5 所示的流程更加容易理解，下面通过图 6 进行详细描述。

[0054] 参见图 6，图 6 为本发明实施例提供的组网示意图。为简单描述，图 6 仅示出了汇聚层中存在一个堆叠设备（由设备 C 和设备 D 级联成的堆叠设备 1）、核心层中存在一个堆叠设备（由设备 A 和设备 B 级联成的堆叠设备 2）、以及接入层中存在 4 个接入设备（即设备 E 至设备 H）。图 6 以接入层中的接入设备不以 IRF 设计为例。

[0055] 另外，图 6 中各层之间的设备通过聚合链路连接，具体如图 6 所示的圆圈。并且，为简单描述，图 6 仅示出了堆叠设备 1 与堆叠设备 2 之间存在一个聚合链路 L，该聚合链路 L 包含两个成员链路，分别为：设备 C 与设备 A 之间的链路 L_{CA}、设备 D 与设备 B 之间的链路 L_{DB}。多个聚合链路的原理与该一个聚合链路的原理类似。

[0056] 基于图 5 所示的流程，则首先，堆叠设备 1 监控聚合链路 L 的流量。如果堆叠设备 1 与堆叠设备 2 之间存在多个聚合链路，则可根据经验或者实际情况从该多个聚合链路中指定其中的若干个。

[0057] 当堆叠设备 1 监控出聚合链路 L 中的一个成员链路（以下以 L_{CA} 为例）的使用带宽（这里可以采用使用带宽的百分比）超过第一设定阈值，且与另一条成员链路 L_{DB} 的使用带宽（这里可以采用使用带宽的百分比）差值超过第二设定阈值，则认为聚合链路 L 存在严重的流量不均衡现象，需要立刻调整。

[0058] 基于此，堆叠设备 1 从 L_{CA} 上找出流量比较大的 N 条数据流。之后，针对每一条数据流，根据该数据流的源地址通过 URPF 查找该数据流进入自身的入接口即 vlan 虚接口，为该数据流对应的 vlan 虚接口重新分配新 MAC 地址，其中，该分配具体实现时可为：从汇聚层对应的 MAC 地址池中随机分配一个 MAC 地址，只要与该 vlan 虚接口之前的 MAC 地址不同即可。如此，即可为该 N 条数据流对应的 vlan 虚接口重新分配 MAC 地址。

[0059] 之后，堆叠设备 1 在被分配了新 MAC 的 vlan 虚接口上发送免费 ARP 报文，接收到该免费 ARP 报文的接入设备从免费 ARP 报文中学习新 MAC 地址，并将所述免费 ARP 报文转发至其接入的终端服务器，终端服务器从免费 ARP 报文中得到新 MAC 地址，并利用得到的新

MAC 地址更新 ARP 表项。当终端服务器再次向被分配了新 MAC 地址的 vlan 虚接口发送数据流时，将该新 MAC 地址作为该数据流的目的 MAC 地址；当接入层中的接入设备接收到终端服务器再次发送的数据流时，重新根据新的目的 MAC 地址和负载均衡算法进行哈希计算，根据该哈希运算结果从用于转发该数据流的聚合链路中选择成员链路。从概率上讲，接入设备会由于目的 MAC 地址不同而导致该确定的成员链路与之前确定的成员链路不同，之后通过该确定的成员链路向汇聚层发送该数据流。

[0060] 当汇聚层中堆叠设备 1 的成员设备接收到该数据流时，优先通过本地的出端口发送。基于上面描述的接入设备会由于目的 MAC 地址不同而导致该确定的成员链路与之前确定的成员链路不同，则可以得到该数据流与之前发送的数据流到达汇聚层中堆叠设备 1 的成员设备不同，比如之前到达堆叠设备 1 的设备 C，本次到达设备 D，如此，设备 D 从聚合链路 L 的成员链路 L_{DB} 上将接收的数据流发送出去，这一方面，减轻了 L_{CA} 的负载，另一方面，使 L_{DB} 与 L_{CA} 之间的负载趋向平衡。

[0061] 可以看出，从概率上来说，终端服务器采用新 MAC 作为目的 MAC 地址发送数据流会打破之前聚合链路中成员链路之间那种极端异常的流量不均衡现象，使该现象向好的方面即聚合链路中成员链路之间流量负载均衡转化。

[0062] 当然，如果堆叠设备 1 后续又监控到与核心层之间聚合链路出现流量不均衡现象时，可以重复上面的方法。也即，本发明提供的方法是一个循环的、不断优化的自动过程。

[0063] 至此，完成本发明实施例提供的方法描述。下面对本发明实施例提供的装置进行描述。

[0064] 参见图 7，图 7 为本发明实施例提供的装置结构图。该装置为汇聚层中的堆叠设备，如图 7 所示，该装置具体包括：监控单元、查找单元和分配单元。

[0065] 其中，监控单元，用于监控所述装置与核心层之间聚合链路的流量；

[0066] 查找单元，用于在所述监控单元监控出聚合链路存在流量不均衡时，从该聚合链路传输的数据流中找出引起该流量不均衡的数据流；

[0067] 分配单元，包括：确定子单元、分配子单元和发送子单元，其中，

[0068] 所述确定子单元，用于确定所述查找单元找出的数据流进入所述装置的入接口；

[0069] 所述分配子单元，用于为所述确定子单元确定的入接口重新分配新的 MAC 地址；

[0070] 所述发送子单元，用于向终端服务器发送所述确定子单元确定的入接口被分配的新 MAC 地址，以通过所述终端服务器利用该新 MAC 地址作为目的 MAC 地址发送数据流来将所述流量不均衡转化为流量均衡。

[0071] 本实施例中，如图 7 所示，所述查找单元包括：链路不均衡确定子单元和数据流查找子单元。

[0072] 其中，链路不均衡确定子单元，用于当聚合链路中一条成员链路的使用带宽超过第一设定阈值，且该成员链路与其他一条成员链路的使用带宽之差超过第二设定阈值时，确定该聚合链路存在流量不均衡。

[0073] 数据流查找子单元，用于统计出所述聚合链路中使用带宽超过第一设定阈值，且与其他一条成员链路的使用带宽之差超过第二设定阈值的成员链路传输的数据流，按照选取大流量数据流的原则从统计出的数据流中选取设定值 N 条数据流，将选取的数据流确定为引起该流量不均衡的数据流。

[0074] 本实施例中，所述确定子单元用于针对找出的每一数据流，确定该数据流的源 IP 地址，根据该确定的源 IP 地址通过单播逆向路径转发 URPF 查找到该数据流的入接口；所述入接口为虚拟局域网 VLAN 虚接口。

[0075] 所述发送子单元在被分配了新 MAC 地址的入接口上向接入层发送免费 ARP 报文，以使接入层中接收到所述免费 ARP 报文的接入设备将所述免费 ARP 报文转发至其接入的终端服务器。

[0076] 需要说明的是，所述装置还包括：级联成所述装置的多个成员设备（在图 7 中未示出）。

[0077] 其中，每一成员设备在接收到接入层中接入设备转发的数据流时，按照本地转发优先的原则从用于转发该数据流的聚合链路中选择出一条成员链路，通过该确定的成员链路向核心层转发该数据流；这里，该数据流的目的 MAC 地址为新 MAC 地址，接入设备通过以下操作将该数据流转发至所述成员设备：将作为该数据流目的 MAC 地址的新 MAC 地址和负载均衡算法进行哈希运算，根据哈希运算结果从用于转发该数据流的聚合链路中选择出一条成员链路，利用该确定的成员链路向汇聚层转发该数据流。

[0078] 至此，完成本发明实施例提供的装置描述。

[0079] 由以上技术方案可以看出，本发明中，接入层中接入设备根据来自终端服务器的数据流的目的 MAC 地址从聚合链路中选择一成员链路，通过该选择的成员链路转发该数据流至汇聚层，汇聚层中堆叠设备依据本地转发优先的原则从聚合链路中选择成员链路转发接收的数据流至核心层，并监控与核心层之间聚合链路的流量，当监控出聚合链路存在流量不均衡时，从该聚合链路传输的数据流中找出引起该流量不均衡的数据流，确定该找出的数据流对应的入接口，重新为该入接口分配新的 MAC 地址，并通过接入设备向终端服务器发送该入接口被分配的新 MAC 地址，这样，当终端服务器后续向被分配了新 MAC 地址的入接口发送数据流时，就以该新 MAC 地址作为目的 MAC 地址向接入设备发送，当接入设备接收到该数据流时，利用该新 MAC 地址从聚合链路中选择一成员链路，通过该选择的成员链路转发该数据流至汇聚层，而汇聚层中堆叠设备的成员设备接收到数据流时，依据本地转发优先原则从聚合链路中选择成员链路，通过该选择的成员链路向核心层转发该数据流。从概率上来说，接入设备依据新 MAC 地址确定的成员链路会与之前依据旧 MAC 地址（分配新 MAC 地址之前使用的 MAC 地址）确定的成员链路不同，这保证了接入设备转发的两种不同的数据流（即用新 MAC 地址作为目的 MAC 地址的数据流与用旧 MAC 地址作为目的 MAC 地址的数据流）到达汇聚层中堆叠设备的不同成员设备，基于本地转发优先原则，进而减轻了汇聚层中堆叠设备与核心层之间聚合链路中成员链路之间极端异常的流量不均衡现象，使该现象向好的方面即聚合链路中成员链路之间流量负载均衡转化。

[0080] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明保护的范围之内。

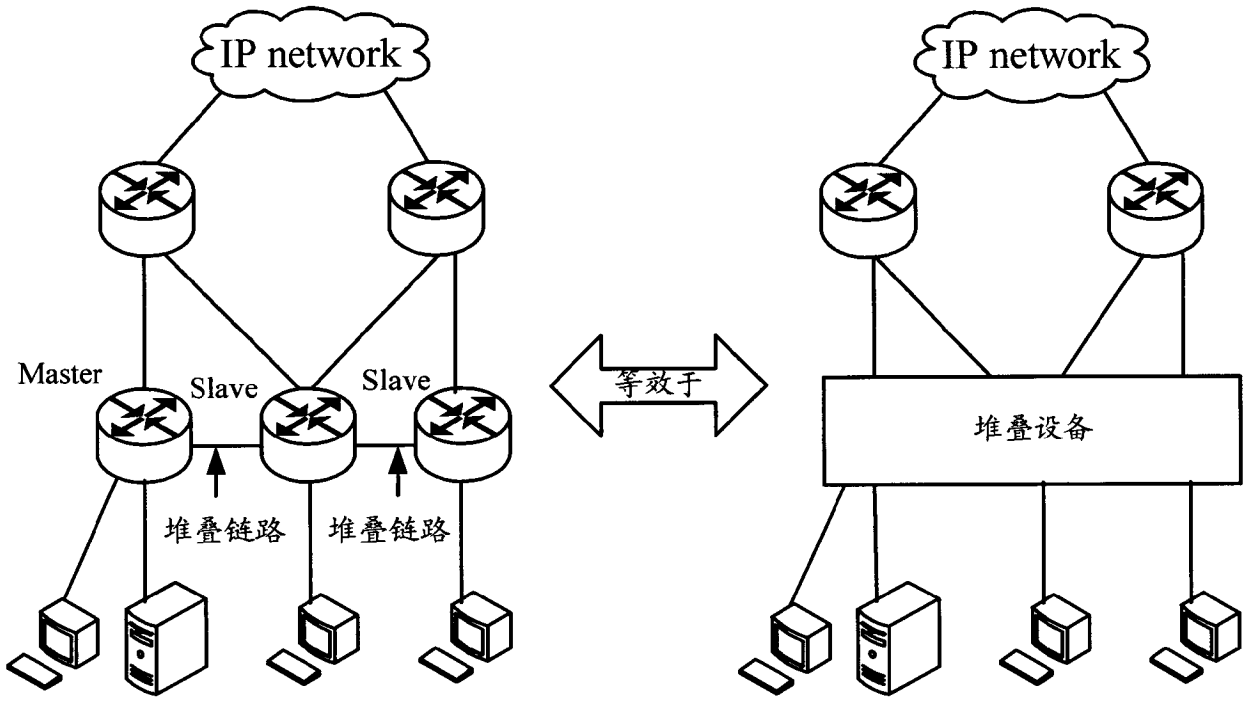

IP network
Master
Slave
Slave
堆叠链路
堆叠链路
等效于
IP network
堆叠设备

图 1

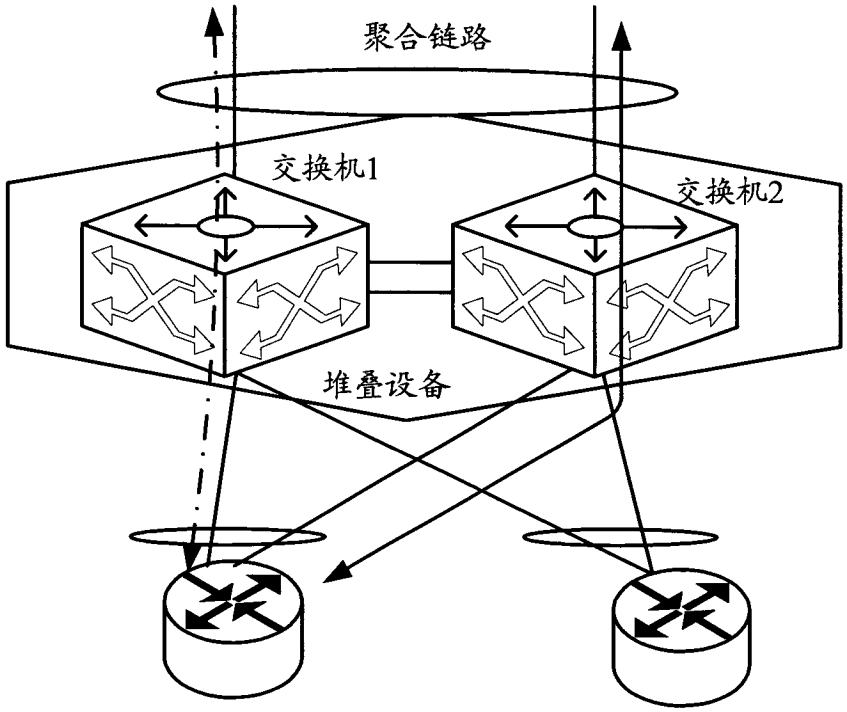

聚合链路
交换机1
交换机2
堆叠设备

图 2

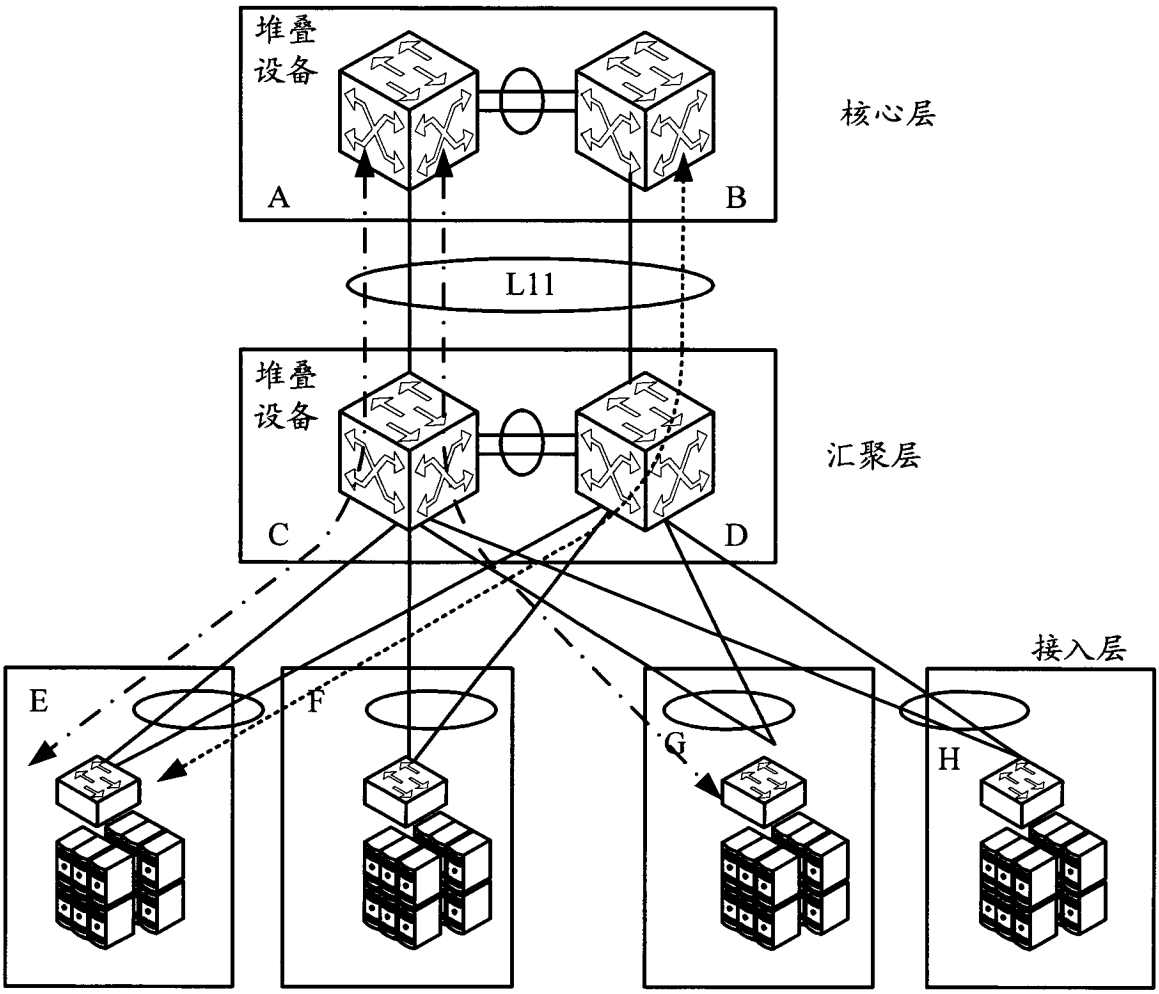


图 3

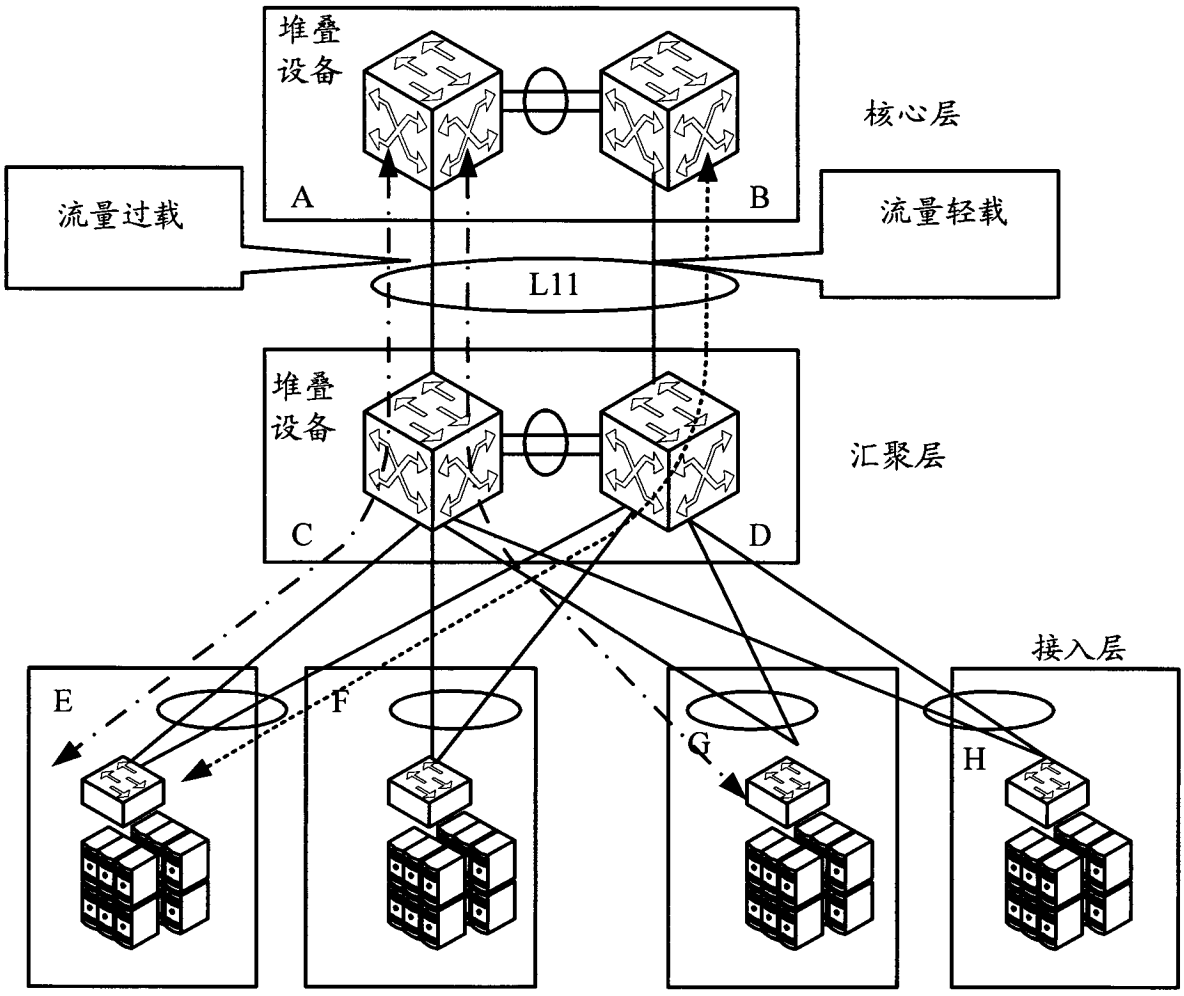


图 4

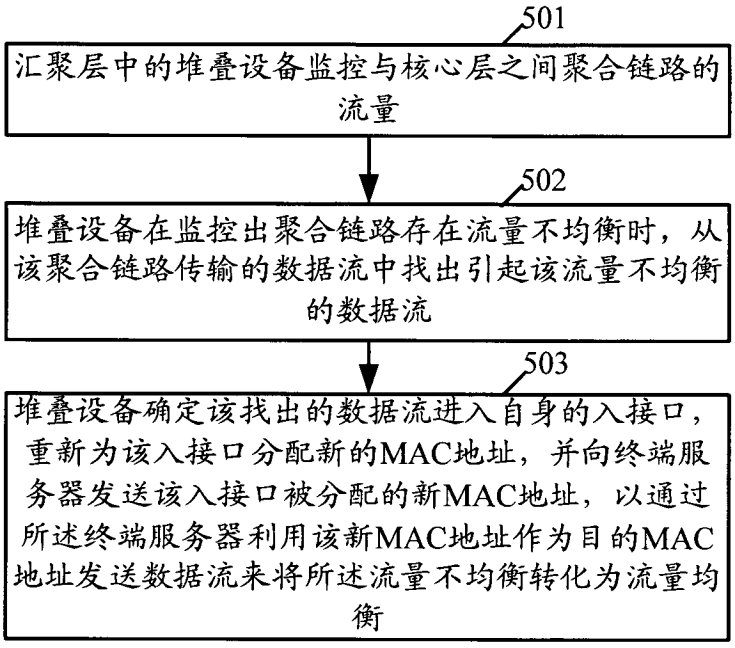

501
汇聚层中的堆叠设备监控与核心层之间聚合链路的流量
502
堆叠设备在监控出聚合链路存在流量不均衡时，从该聚合链路传输的数据流中找出引起该流量不均衡的数据流
503
堆叠设备确定该找出的数据流进入自身的入接口，重新为该入接口分配新的MAC地址，并向终端服务器发送该入接口被分配的新MAC地址，以通过所述终端服务器利用该新MAC地址作为目的MAC地址发送数据流来将所述流量不均衡转化为流量均衡

图 5

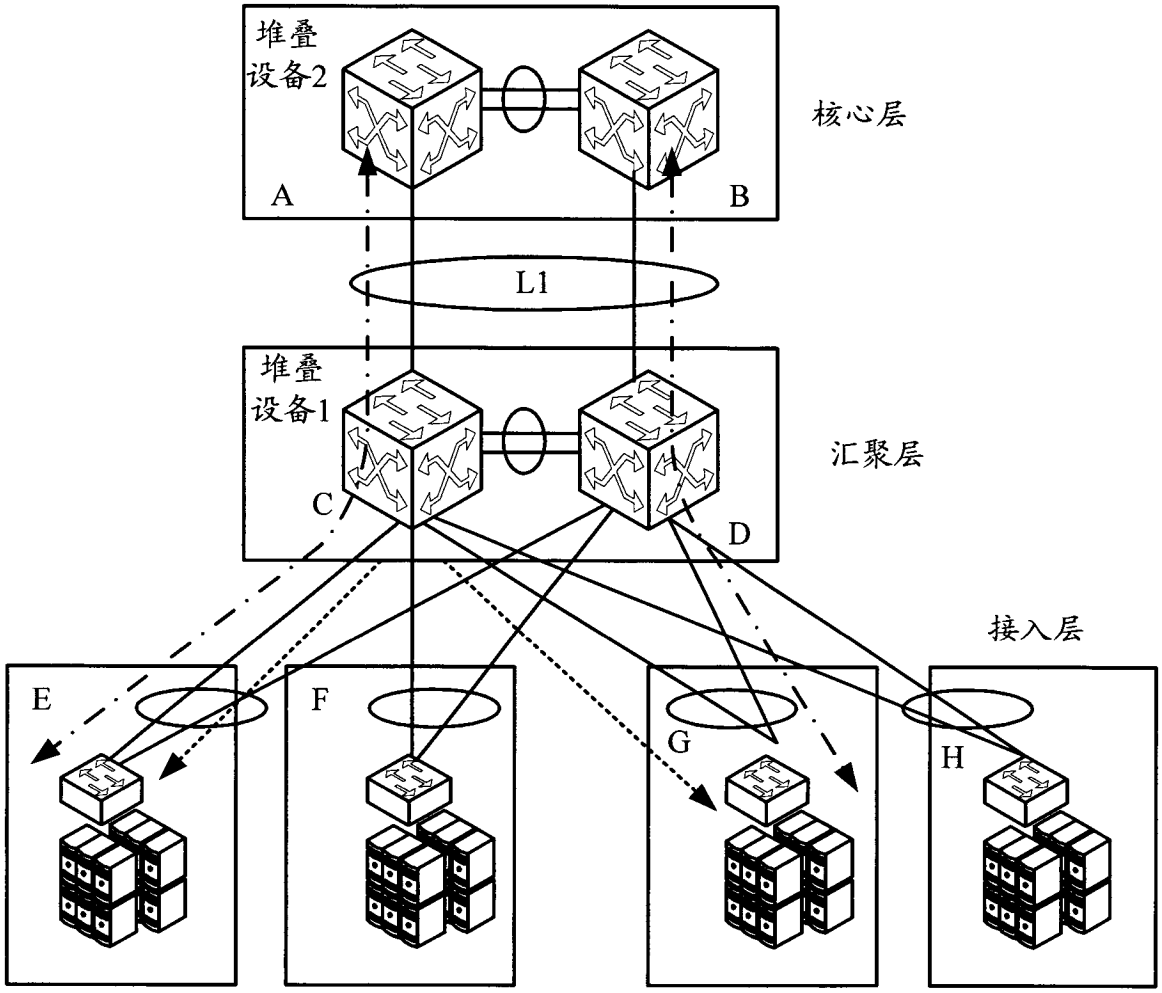

堆叠
设备2
A
B
核心层
L1
堆叠
设备1
C
D
汇聚层
接入层
E
F
G
H

图 6

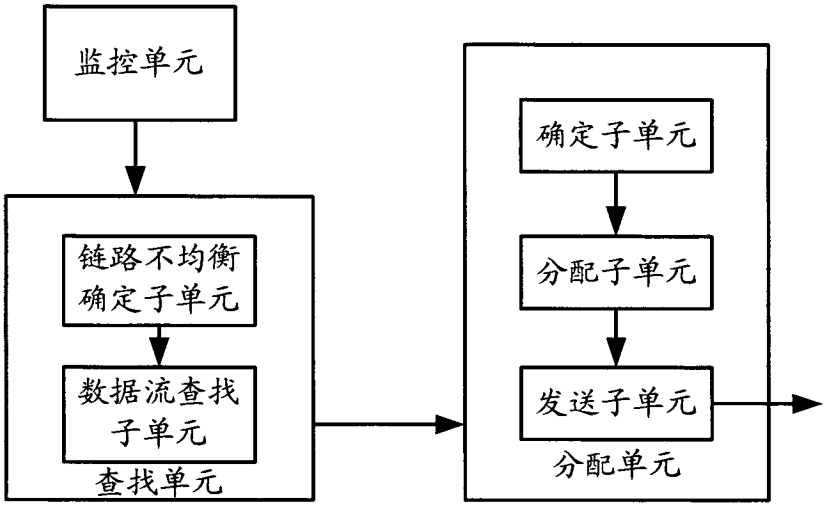

监控单元
链路不均衡
确定子单元
数据流查找
子单元
查找单元
确定子单元
分配子单元
发送子单元
分配单元

图 7