[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
H04L 29/08 (2006.01)

[12] 发明专利说明书

专利号 ZL 200610139722.0

[45] 授权公告日 2009 年 7 月 8 日

[11] 授权公告号 CN 100512318C

[22] 申请日 2006.9.22

[21] 申请号 200610139722.0

[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 常利民

[56] 参考文献

WO2002/048823A3 2002.6.20

US6856991B1 2005.2.15

CN1835467A 2006.9.20

CN1426211A 2003.6.25

CN1812344A 2006.8.2

审查员 王 澍

[74] 专利代理机构 北京德琦知识产权代理有限公司

代理人 宋志强 麻海明

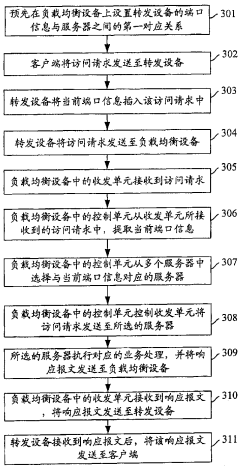
权利要求书 2 页 说明书 11 页 附图 2 页

[54] 发明名称

一种实现负载均衡的方法和系统及负载均衡设备

[57] 摘要

本发明公开了一种实现负载均衡的方法和系统及负载均衡设备。该负载均衡设备包括：控制单元、收发单元和存储单元。该方法包括：设置转发设备的端口号与服务器的第一对应关系；转发设备在当前端口上接收到客户端发来的访问请求，将当前端口的端口号插入该访问请求中发送至负载均衡设备；负载均衡设备根据所设置的第一对应关系，选择与接收到访问请求中的当前端口的端口号对应的服务器，将访问请求发送至所选的服务器。本发明可以实现由一个特定的服务器为特定的客户端提供服务，因此，大大增加了业务实现的灵活性，提高了业务服务质量。



1、一种实现负载均衡的方法，其特征在于，设置转发设备的端口号与服务器的第一对应关系，该方法还包括:

A、转发设备在当前端口上接收到客户端发来的访问请求，将当前端口的端口号插入该访问请求中发送至负载均衡设备;

B、负载均衡设备根据所设置的第一对应关系，选择与接收到访问请求中的当前端口的端口号对应的服务器，将访问请求发送至所选的服务器。

2、根据权利要求 1 所述的方法，其特征在于，在步骤 B 中，所述将访问请求发送至所选的服务器的步骤包括：负载均衡设备直接将插入有当前端口的端口号的访问请求发送至所选的服务器。

3、根据权利要求 2 所述的方法，其特征在于，在步骤 B 之后，进一步包括：所选的服务器将携带有当前端口的端口号的响应报文发送至负载均衡设备；负载均衡设备将该携带有当前端口的端口号的响应报文发送至转发设备；转发设备去除响应报文中的当前端口的端口号，然后将响应报文发送至所述客户端。

4、根据权利要求 1 所述的方法，其特征在于，在步骤 B 中，所述将访问请求发送至所选的服务器的步骤包括：负载均衡设备建立访问请求中的客户端地址与当前端口的端口号之间的第三对应关系，并去除访问请求中的当前端口的端口号，然后将访问请求发送至所选的服务器。

5、根据权利要求 4 所述的方法，其特征在于，在步骤 B 之后，进一步包括：所选的服务器将响应报文发送至负载均衡设备；负载均衡设备根据所建立的第三对应关系和响应报文中携带的客户端地址，将该客户端地址对应的当前端口的端口号插入响应报文中，然后将响应报文发送至转发设备；转发设备去除响应报文中的当前端口的端口号，然后将响应报文发送至所述客户端。

6、根据权利要求 1 至 5 中任意一项所述的方法，其特征在于，所述转发设备的端口号为：所有转发设备的所有端口进行统一编号后，该转发设备端口的端口号；或者，所述转发设备的设备号加该端口在该转发设备上的编

号。

7、一种实现负载均衡的系统，其特征在于，该系统包括：客户端、转发设备、负载均衡设备和多个服务器，其中，

转发设备，用于在当前的端口接收客户端发来的访问请求，将当前端口的端口号插入访问请求中发送至负载均衡设备；

负载均衡设备，用于保存转发设备的端口号与服务器的第一对应关系，在接收到转发设备发来的访问请求后，根据所保存的第一对应关系，选择与接收到访问请求中的当前端口的端口号对应的服务器，将访问请求发送至所选的服务器。

8、根据权利要求 7 所述的系统，其特征在于，所述转发设备为路由器或交换机。

9、一种负载均衡设备，其特征在于，该负载均衡设备包括：控制单元、收发单元和存储单元，其中，

存储单元，用于保存转发设备的端口号与服务器的第一对应关系；

收发单元，用于接收外部转发设备发来的携带当前端口的端口号的访问请求，将访问请求发送至所选的服务器；

控制单元，用于在收发单元接收到携带当前端口的端口号的访问请求后，从该访问请求中提取当前端口的端口号，根据存储单元中保存的第一对应关系，从多个服务器中选择与所提取的当前端口的端口号对应的服务器，控制收发单元将访问请求发送至所选的服务器。

10、根据权利要求 9 所述的负载均衡设备，其特征在于，所述控制单元进一步建立访问请求中的客户端地址与当前端口的端口号之间的第三对应关系，去除访问请求中的当前端口的端口号后，控制收发单元将访问请求发送至所选的服务器，在收发单元接收到所选服务器的响应报文后，根据所建立的第三对应关系，在响应报文中插入与响应报文中的客户端地址对应的当前端口的端口号，然后控制收发单元将响应报文发送至转发设备。

一种实现负载均衡的方法和系统及负载均衡设备

技术领域

本发明涉及网络通信技术，特别是涉及一种实现负载均衡的方法和系统及负载均衡设备。

背景技术

伴随着互联网（Internet）技术的不断发展，网络服务器面对的访问数量大大增加并且更加不可预知。在网络中，如果仅使用一个服务器对客户端提供服务，那么，服务器的处理能力和输入输出能力必然会成为提供服务的瓶颈。

为了解决单台服务器处理能力有限，无法保证为客户端提供服务的缺点，目前出现了负载均衡技术。

负载均衡技术就是在网络侧设置可以为客户端提供服务的多个服务器，并在服务器侧设置一个负载均衡设备。客户端需要进行访问时，通过转发设备即路由器和交换机将访问请求发送至负载均衡设备，负载均衡设备接收到客户端的访问请求后，根据一定的规则从多个服务器中选择一个，由所选的服务器回应客户端，为客户端提供服务。

在负载均衡技术中，由于客户端每次访问服务器时，均需要由负载均衡设备依据一定的规则选择服务器，比如随机选择或根据服务器当前的业务负荷情况选择等，这样，在同一个客户端的不同访问中，所选择出的为客户端提供服务的服务器往往是不同的。

由以上描述可以看出，在现有技术中，由于在同一个客户端的不同访问中，为该客户端提供服务的服务器往往是不同的，这样，则无法限定客户端能访问的实际服务器，也就是说，无法实现由一个特定的服务器为特定的客

户端提供服务。然而，在实际的业务实现中，经常需要由一个特定的服务器针对特定的客户端提供某种特定的业务。比如，根据业务需要，配置服务器1 支持一种增值业务，而虚拟局域网（VLAN）1 内的所有客户端需要使用该种增值业务，但是，由于现有技术无法保证VLAN1 内的所有客户端能够访问到服务器1,因此则无法保证LAN1 内的所有客户端实现该种增值业务。因此，大大限制了业务实现的灵活性，降低了业务服务质量。

发明内容

有鉴于此，本发明的第一目的在于提供一种实现负载均衡的方法，本发明的第二目的在于提供一种实现负载均衡的系统，本发明的第三目的在于提供一种负载均衡设备，以便于实现由一个特定的服务器为特定的客户端提供服务。

为了达到上述目的，本发明的技术方案是这样实现的:

一种实现负载均衡的方法，设置转发设备的端口号与服务器的第一对应关系，该方法还包括:

A、转发设备在当前端口上接收到客户端发来的访问请求，将当前端口的端口号插入该访问请求中发送至负载均衡设备;

B、负载均衡设备根据所设置的第一对应关系，选择与接收到访问请求中的当前端口的端口号对应的服务器，将访问请求发送至所选的服务器。

在步骤 B 中，所述将访问请求发送至所选的服务器的步骤包括: 负载均衡设备直接将插入有当前端口的端口号的访问请求发送至所选的服务器。

在步骤 B 之后，进一步包括: 所选的服务器将携带有当前端口的端口

号的响应报文发送至负载均衡设备；负载均衡设备将该携带有当前端口的端口号的响应报文发送至转发设备；转发设备去除响应报文中的当前端口的端口号，然后将响应报文发送至所述客户端。

在步骤 B 中，所述将访问请求发送至所选的服务器的步骤包括：负载均衡设备建立访问请求中的客户端地址与当前端口的端口号之间的第三对应关系，并去除访问请求中的当前端口的端口号，然后将访问请求发送至所选的服务器。

在步骤 B 之后，进一步包括：所选的服务器将响应报文发送至负载均衡设备；负载均衡设备根据所建立的第三对应关系和响应报文中携带的客户端地址，将该客户端地址对应的当前端口的端口号插入响应报文中，然后将响应报文发送至转发设备；转发设备去除响应报文中的当前端口的端口号，然后将响应报文发送至所述客户端。

所述转发设备的端口号为：所有转发设备的所有端口进行统一编号后，该转发设备端口的端口号；或者，所述转发设备的设备号加该端口在该转发设备上的编号。

一种实现负载均衡的系统，该系统包括：客户端、转发设备、负载均衡设备和多个服务器，其中，

转发设备，用于在当前的端口接收客户端发来的访问请求，将当前端口的端口号插入访问请求中发送至负载均衡设备；

负载均衡设备，用于保存转发设备的端口号与服务器的第一对应关系，

在接收到转发设备发来的访问请求后，根据所保存的第一对应关系，选择与接收到访问请求中的当前端口的端口号对应的服务器，将访问请求发送至所选的服务器。

所述转发设备为路由器或交换机。

一种负载均衡设备，该负载均衡设备包括：控制单元、收发单元和存储单元，其中，

存储单元，用于保存转发设备的端口号与服务器的第一对应关系；

收发单元，用于接收外部转发设备发来的携带当前端口的端口号的访问请求，将访问请求发送至所选的服务器；

控制单元，用于在收发单元接收到携带当前端口的端口号的访问请求后，从该访问请求中提取当前端口的端口号，根据存储单元中保存的第一对应关系，从多个服务器中选择与所提取的当前端口的端口号对应的服务器，控制收发单元将访问请求发送至所选的服务器。

所述控制单元进一步建立访问请求中的客户端地址与当前端口的端口号之间的第三对应关系，去除访问请求中的当前端口的端口号后，控制收发单元将访问请求发送至所选的服务器，在收发单元接收到所选服务器的响应报文后，根据所建立的第三对应关系，在响应报文中插入与响应报文中的客户端地址对应的当前端口的端口号，然后控制收发单元将响应报文发送至转发设备。

由此可见，本发明具有以下优点：

1、在本发明中，利用客户端所连接的转发设备的端口信息来限定为客户端提供服务的服务器，也就是说，可以实现由一个特定的服务器为特定的客户端提供服务，因此，大大增加了业务实现的灵活性，提高了业务服务质量。

2、在本发明中，由于是利用转发设备的端口信息与服务器的对应关系来实现特定服务器为特定客户端服务，因此，在客户端的地址或服务器的地址发生变化后，只需更新该对应关系即可，而无需对客户端和服务器的配置进行任何改动，因此，使得本发明简单且易于实现，增强了本发明的实用性。

附图说明

图 1 是在本发明中实现负载均衡的系统的结构示意图。

图 2 是在本发明中负载均衡设备内部的结构示意图。

图 3 是在本发明实施例中实现负载均衡的流程图。

具体实施方式

目前，由于客户端在接入网络时，其分配的 IP 地址往往是动态的，也就是说，同一个客户端在不同访问中使用的 IP 地址很可能是不相同的，这样，则无法通过客户端的 IP 地址来限定为客户端提供服务的服务器。然而，对客户端与服务器进行通信的过程进行分析可知，同一个客户端接入转发设备即路由器和交换机的端口是相同的，也就是说，同一个客户端不同访问中的所有报文都是通过转发设备上的同一个端口进行传输的，因此，可以利用客户端所连接的转发设备的端口信息来限定为客户端提供服务的服务器。

针对上述特点，本发明提出了一种实现负载均衡的方法，其核心思想是：设置转发设备的端口信息与服务器的第一对应关系；转发设备在当前端口上接收到客户端发来的访问请求，将当前端口信息插入该访问请求中发送至负

载均衡设备；负载均衡设备根据所设置的第一对应关系，选择与接收到访问请求中的当前端口信息对应的服务器，将访问请求发送至所选的服务器。

相应的，本发明提出了一种实现负载均衡的系统。图 1 是在本发明中实现负载均衡的系统的结构示意图。参见图 1，在本发明中，实现负载均衡的系统包括：客户端、转发设备、负载均衡设备和多个服务器，其中，

转发设备，用于在当前的端口接收客户端发来的访问请求，将当前端口信息插入访问请求中发送至负载均衡设备；

负载均衡设备，用于保存转发设备的端口信息与服务器的第一对应关系，在接收到转发设备发来的访问请求后，根据所保存的第一对应关系，选择与接收到访问请求中的当前端口信息对应的服务器，将访问请求发送至所选的服务器。

相应的，本发明还提出了一种负载均衡设备。图 2 是在本发明中负载均衡设备内部的结构示意图。参见图 2，在本发明中，负载均衡设备内部的结构包括：控制单元、收发单元和存储单元，其中，

存储单元，用于保存转发设备的端口信息与服务器的第一对应关系；

收发单元，用于接收外部转发设备发来的携带当前端口信息的访问请求，将访问请求发送至所选的服务器；

控制单元，用于在收发单元接收到携带当前端口信息的访问请求后，从该访问请求中提取当前端口信息，根据存储单元中保存的第一对应关系，从多个服务器中选择与所提取的当前端口信息对应的服务器，控制收发单元将访问请求发送至所选的服务器。

在本发明中，所述的转发设备可以是客户端与负载均衡设备之间的路由器或交换机。

在本发明中，可以通过当前端口的端口号来表示当前端口的信息，或者，也可以通过当前端口对应的 VLAN 的标签来表示当前端口的信息。

为使本发明的目的、技术方案和优点更加清楚，下面结合附图及具体实施例对本发明作进一步地详细描述。

图 3 是在本发明实施例中实现负载均衡的流程图。参见图 1、图 2 和图 3，在本发明中，实现负载均衡的过程包括以下步骤：

步骤 301：预先在负载均衡设备上设置转发设备的端口信息与服务器之间的第一对应关系。

通过本步骤的过程，可以实现将转发设备上的特定端口与特定服务器对应，相应的，也就实现了连接到转发设备上特定端口的客户端与特定服务器的对应。

另外，在本步骤中，可以采用包括但不限于以下两种方式来表示转发设备的端口信息，从而完成设置第一对应关系：

方式一、使用 VLAN 标签表示转发设备上的端口信息。

在实际的业务实现中，由于一个 VLAN 内客户端可使用的业务是相同的，并且一个 VLAN 内所有客户端连接的转发设备的端口是固定的，比如，VLAN1 内的所有客户端均连接到转发设备的第一组端口上，VLAN2 内的所有客户端均连接到转发设备的第二组端口上。因此，在本步骤 301 中，可以采用方式一来间接地表示转发设备上的端口信息。

当采用该方式一时，还需要在转发设备上设置 VLAN 标签与该 VLAN 所连转发设备端口的第二对应关系，即使用 VLAN 的标签表示端口信息。比如，VLAN1 内的客户端连接到转发设备的第一组端口上，那么，则可以在转发设备上设置 VLAN1 与第一组端口之间的第二对应关系，即使用 VLAN1 表示第一组端口；并且，在本步骤中，在负载均衡设备上设置可标识转发设备端口的 VLAN 标签与服务器的第一对应关系，比如，转发设备第一组端口所连的 VLAN1 内的客户端需要访问特定的服务器 1，那么，在本步骤中；则设置可标识第一组端口信息的 VLAN1 与服务器 1 之间的第一对应关系。所设置的第一对应关系和第二对应关系的形式可参见如下表 1 所示。

第一对应关系	第二对应关系
VLAN1～服务器 1	第一组端口～VLAN1

表 1

方式二、使用转发设备的端口号表示转发设备的端口信息。

在该方式二中，直接使用客户端所连的转发设备的端口号表示端口信息，并建立端口号与服务器之间的第一对应关系。

参见图 2，当负载均衡设备内部的结构如图 2 所示时，在本步骤 301 中，是将所述的第一对应关系设置在负载均衡设备的存储单元中。

步骤 302：客户端将访问请求发送至转发设备。

步骤 303：转发设备在当前的一个端口上接收到客户端发来的访问请求，将当前端口信息插入该访问请求中。

当在上述步骤 301 中，采用方式一，即使用 VLAN 标签表示端口信息，并在转发设备中设置了端口与 VLAN 标签的第二对应关系，以及在负载均衡设备中设置了 VLAN 标签与服务器的第一对应关系，那么，在本步骤 302 中，在转发设备接收到访问请求之后，并在将当前端口信息插入访问请求之前，转发设备首先根据所设置的第二对应关系，确定与当前端口对应的 VLAN 标签，将所确定的与当前端口对应的 VLAN 标签作为当前端口信息，这样，所述的转发设备将当前端口信息插入访问请求中的具体实现为：转发设备将所确定的与当前端口对应的 VLAN 标签插入访问请求中。

当在上述步骤 301 中，采用方式二，即在负载均衡设备中设置了端口号与服务器的第一对应关系，那么，在本步骤 302 中，转发设备将当前端口信息插入该访问请求中的过程具体包括：转发设备直接将当前端口的端口号插入访问请求中。

在本步骤 303 中，转发设备可以将当前端口信息插入访问请求中的预留字段上，或插入访问请求中新增的字段上。

步骤 304：转发设备将访问请求发送至负载均衡设备。

步骤 305：负载均衡设备中的收发单元接收到访问请求。

步骤 306：负载均衡设备中的控制单元从收发单元所接收到的访问请求中，提取当前端口信息。

当在上述步骤 301 中，采用方式一，则在本步骤中，负载均衡设备中的控制单元提取的是 VLAN 标签。

当在上述步骤 301 中，采用方式二，则在本步骤中，负载均衡设备中的控制单元提取的是端口号。

步骤 307：负载均衡设备中的控制单元根据存储单元中保存的第一对应关系以及所提取的当前端口信息，从多个服务器中选择与当前端口信息对应的服务器。

当在上述步骤 301 中，采用方式一，则在本步骤中，负载均衡设备中的控制单元根据 VLAN 标签与服务器的对应关系，选择与所提取的 VLAN 标签对应的服务器。

当在上述步骤 301 中，采用方式二，则在本步骤中，负载均衡设备中的控制单元根据端口号与服务器的对应关系，选择与所提取的端口号对应的服务器。

步骤 308：负载均衡设备中的控制单元控制收发单元将访问请求发送至所选的服务器。

在本步骤中，负载均衡设备可以直接将插入有当前端口信息的访问请求发送至所选的服务器；

或者，负载均衡设备中的控制单元也可以首先建立访问请求中的客户端地址与当前端口信息之间的第三对应关系，并去除访问请求中的当前端口信息，然后再控制收发单元将不携带当前端口信息的访问请求发送至所选的服务器。

步骤 309：所选的服务器根据接收到的访问请求执行对应的业务处理，并将响应报文发送至负载均衡设备。

这里，如果在步骤 308 中，负载均衡设备将插入有当前端口信息的访问

请求发送至所选的服务器，那么，在本步骤中，所选服务器返回的响应报文中携带有当前端口信息；

如果在步骤308中，负载均衡设备将未携带当前端口信息的访问请求发送至所选的服务器，那么，在本步骤中，所选服务器返回的响应报文中不携带当前端口信息。

步骤310：负载均衡设备中的收发单元接收到响应报文，将响应报文发送至转发设备。

这里，在负载均衡设备中的收发单元接收到响应报文时，如果响应报文中携带有当前端口信息，则收发单元可以直接将携带有当前端口信息的响应报文发送至转发设备；如果响应报文中不携带当前端口信息，那么，控制单元可以根据所建立的第三对应关系和响应报文中携带的客户端地址，将该客户端地址对应的当前端口信息插入响应报文中，然后控制收发单元将携带有当前端口信息的响应报文发送至转发设备。

步骤311：转发设备接收到响应报文后，将响应报文发送至对应的客户端。

这里，转发设备在接收到响应报文后，首先去除响应报文中的当前端口信息，然后再执行所述的将响应报文发送至对应的客户端。

需要说明的是，在上述步骤308至步骤311中，无论负载均衡设备在将访问请求发送至所选服务器前是否去除了当前端口信息，所选服务器在返回响应报文时，均可以不在响应报文中携带当前端口信息，并且，负载均衡设备也可以不在响应报文中插入当前端口信息，而直接将不携带当前端口信息的响应报文发送至转发设备，转发设备根据响应报文中的客户端地址将该响应报文发送至对应的客户端即可。

还需要说明的是，在上述图3所示的过程中，所述的转发设备可以是路由器或交换机。并且，所述转发设备端口的端口号可以是所有转发设备的所有端口进行统一编号后，该转发设备端口的端口号；或者，所述转发设备端口的端口号也可以是由该转发设备的设备号加该端口在该转发设备上的编

号来构成。

总之，以上所述仅为本发明的较佳实施例而已，并非用于限定本发明的保护范围。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

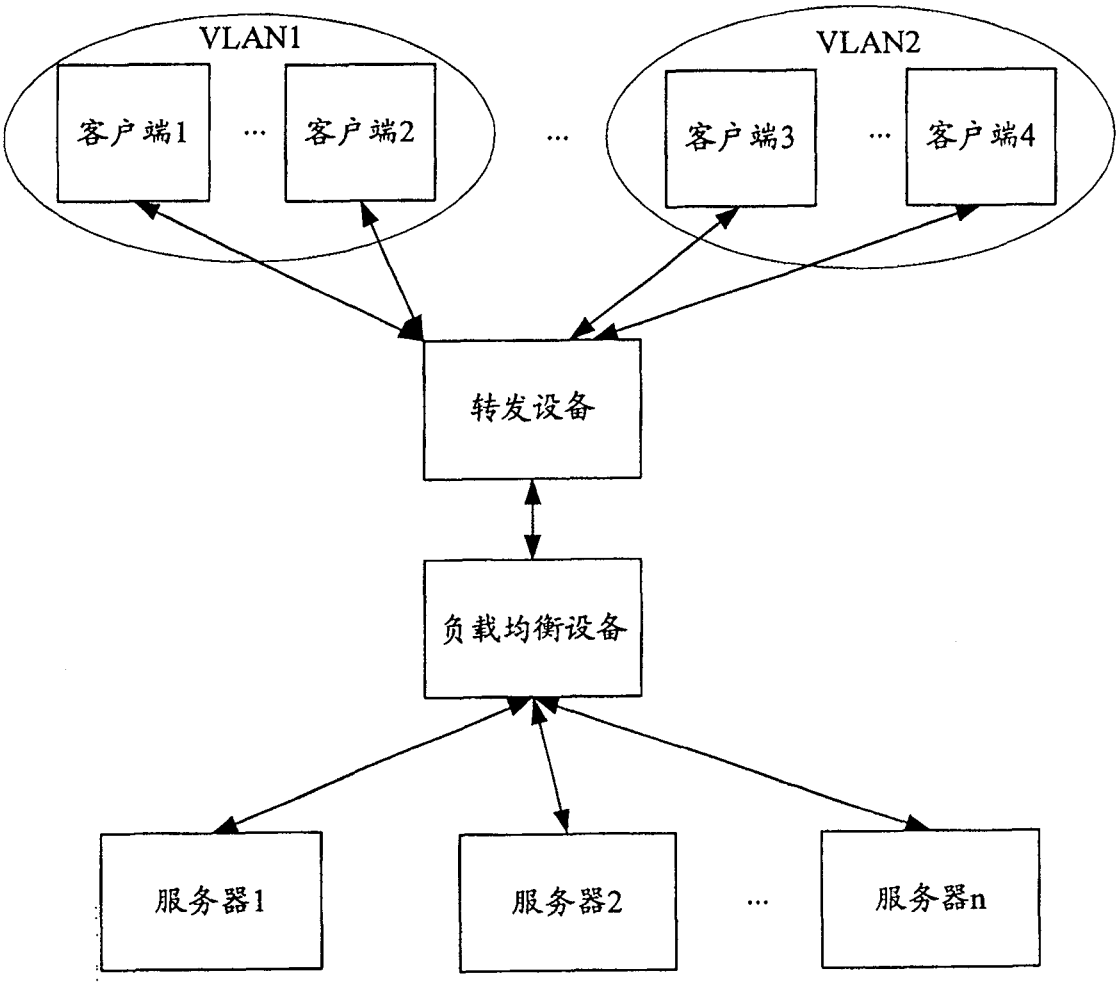

VLAN1
客户端1
客户端2
VLAN2
客户端3
客户端4
转发设备
负载均衡设备
服务器1
服务器2
服务器n

图 1

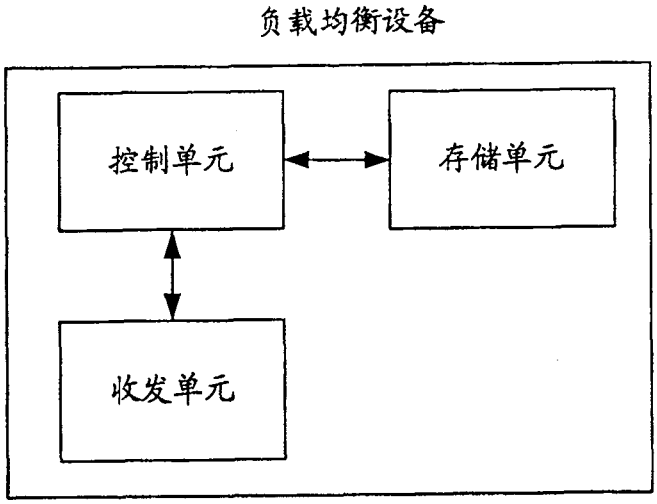

负载均衡设备
控制单元
存储单元
收发单元

图 2

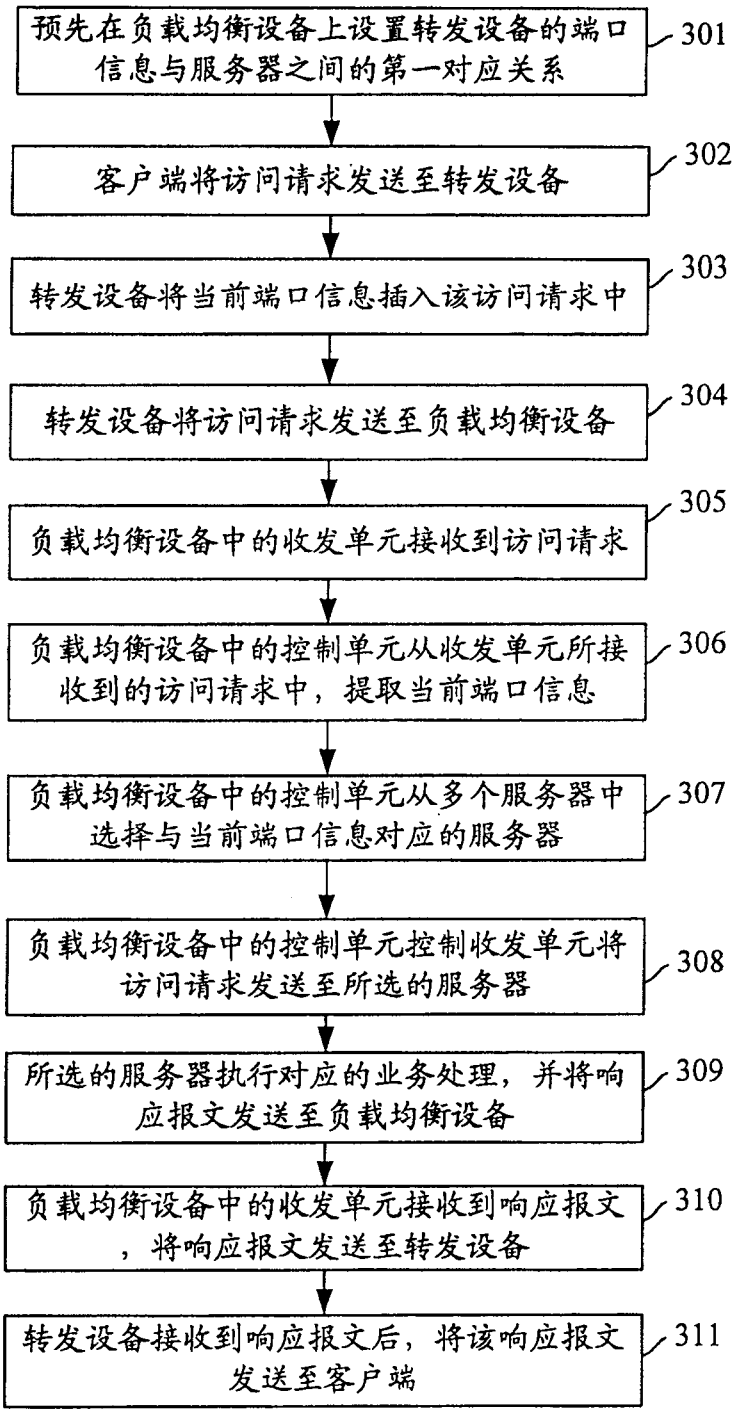

预先在负载均衡设备上设置转发设备的端口信息与服务器之间的第一对应关系
301
客户端将访问请求发送至转发设备
302
转发设备将当前端口信息插入该访问请求中
303
转发设备将访问请求发送至负载均衡设备
304
负载均衡设备中的收发单元接收到访问请求
305
负载均衡设备中的控制单元从收发单元所接收到的访问请求中，提取当前端口信息
306
负载均衡设备中的控制单元从多个服务器中选择与当前端口信息对应的服务器
307
负载均衡设备中的控制单元控制收发单元将访问请求发送至所选的服务器
308
所选的服务器执行对应的业务处理，并将响应报文发送至负载均衡设备
309
负载均衡设备中的收发单元接收到响应报文，将响应报文发送至转发设备
310
转发设备接收到响应报文后，将该响应报文发送至客户端
311

图 3