(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 102449963 A
(43)申请公布日 2012.05.09

(21)申请号 201080023822.1

(22)申请日 2010.05.28

(30)优先权数据
61/182,057 2009.05.28 US
12/605,388 2009.10.26 US

(85)PCT申请进入国家阶段日
2011.11.25

(86)PCT申请的申请数据
PCT/US2010/036757 2010.05.28

(87)PCT申请的公布数据
WO2010/138936 EN 2010.12.02

(71)申请人 微软公司
地址 美国华盛顿州

(72)发明人 P·帕特尔 D·马尔茨 A·格林伯格 袁利华 R·克恩

(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 陈斌

(51)Int.Cl.
H04L 12/56(2006.01)
H04L 12/28(2006.01)
H04L 29/06(2006.01)

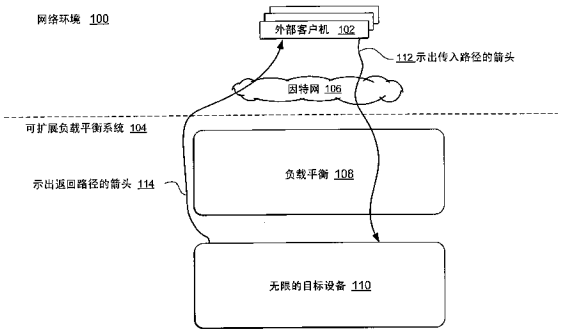
权利要求书 2 页 说明书 13 页 附图 11 页

(54)发明名称

跨层2域的负载平衡

(57)摘要

本申请涉及网络配置，且尤其涉及可扩展负载平衡网络配置。一种实现包括被耦合到可扩展负载平衡系统的外部客户机。可扩展负载平衡系统包括被配置为对来自该外部客户机的分组流的各单独的传入分组进行封装的负载平衡层。该负载平衡层还被配置为将传入分组路由到系统上的目标设备。各目标设备可以跨越多个IP子网。各传入分组可以在到达各单独的目标设备之前穿过负载平衡层的一个或多个负载平衡器。各单独的目标设备可以被配置为将分组流的至少一些传出分组路由到外部客户机而不穿过一个或多个负载平衡器中的任何一个。



CN 102449963 A

1. 一种其上存储有指令的计算机可读存储介质，所述指令在被处理设备执行时执行以下动作：

将网络分组分散在一系列模块之间 (1102)；

在各单独的模块处封装所述各网络分组 (1104)；

使用在所述系列的各模块之间共享的状态来选择将所述网络分组封装到的目标设备 (1106)；以及

从所述一系列模块转发所述网络分组 (1108)。

2. 如权利要求 1 所述的计算机可读存储介质，其特征在于，在所述系列的各模块之间共享的状态是一致散列函数的键空间。

3. 如权利要求 1 所述的计算机可读存储介质，其特征在于，使用等价多路径路由来将各单独的网络分组分散在所述系列的各模块之间。

4. 如权利要求 1 所述的计算机可读存储介质，其特征在于，还包括监视所述目标设备的健康。

5. 如权利要求 1 所述的计算机可读存储介质，其特征在于，响应于所述目标设备的故障来改变在所述系列的各模块之间共享的状态。

6. 如权利要求 1 所述的计算机可读存储介质，其特征在于，在不造成所提供的服务的停机时间的情况下基于负载参数或一个或多个其他参数中的一个或多个来动态地改变所述系列中的多个模块。

7. 如权利要求 1 所述的计算机可读存储介质，其特征在于，所述目标设备是一组目标设备的成员，且在其中接收到所述组的一个或多个现有目标设备将变得不可用或一个或多个新的目标设备可用的指示的情况下，过渡到将与将来通信相关联的网络分组分散到新的一组目标设备的配置，同时继续将与正在进行的通信相关联的网络分组发送给所述一组目标设备。

8. 如权利要求 1 所述的计算机可读存储介质，其特征在于，封装所述分组包括网际协议 (IP)-in-IP 封装。

9. 如权利要求 6 所述的计算机可读存储介质，其特征在于，封装所述分组保留所述分组被发送到的一个或多个虚拟 IP 地址。

10. 一个系统 (104)，包括：

被配置为对来自外部客户机设备 (102) 的分组流的各单独的传入分组进行封装的负载平衡层 (108)，所述负载平衡层 (108) 还被配置为将所述传入分组路由到所述系统 (104) 的目标设备 (110)，其中所述目标设备 (110) 跨一个或多个网际协议 (IP) 子网，且其中所述传入分组在到达各单独的目标设备 (210) 之前穿过所述负载平衡层 (108) 的一个或多个负载平衡器 (208)；以及，

所述各单独的目标设备 (210) 被配置为将所述分组流的至少一些传出分组路由 (222) 到所述外部客户机设备 (102) 而不穿过所述一个或多个负载平衡器 (208) 中的任何一个。

11. 如权利要求 11 所述的系统，其特征在于，所述负载平衡层被配置为利用网际协议 (IP)-in-IP 封装或分组修改及 IP 选项中的一个或两者来封装所述各单独的传入分组。

12. 如权利要求 11 所述的系统，其特征在于，所述负载平衡层包括至少一个动态负载平衡器和至少一个多路复用器，且其中所述至少一个多路复用器被配置为封装所述各单独

的传入分组。

13. 如权利要求 11 所述的系统，其特征在于，所述负载平衡层包括至少一个多路复用器，且其中所述至少一个多路复用器被配置为利用 IP-in-IP 封装来封装所述各单独的传入分组。

14. 如权利要求 11 所述的系统，其特征在于，所述各单独的目标设备包括被配置为解封来自所述负载平衡层的分组的解封组件，或其中各单独的目标设备跨多个虚拟局域网。

15. 如权利要求 11 所述的系统，其特征在于，所述一个或多个负载平衡器包括被配置为提供用于管理所述负载平衡层的虚拟 IP(VIP) 到直接 IP(DIP) 映射的应用程序接口的动态负载平衡器。

跨层 2 域的负载平衡

[0001] 背景

[0002] 负载平衡器可以是可以将一组请求分布在能够处理请求的一组服务器上的网络基础设施的关键部件。常规的负载平衡器可以包括各对设备，每一对设备都是专用硬件。因为使用专用硬件，所以常规的负载平衡器往往花费大量金钱。另一缺点是它们使用按比例扩大策略：单对负载平衡器可以应对受硬件容量限制的多个并发请求。购买包含具有更多容量的硬件的更强大的负载平衡器才能处理附加的请求。在缓解网络中的流量瓶颈方面，直接服务器返回 (DSR) 优化是有用的。然而，常规的负载平衡器的缺点是这种技术是通常限于网络的单个虚拟局域网 (VLAN)。

[0003] 概述

[0004] 本申请涉及网络配置，且尤其涉及可扩展负载平衡网络配置。一种实现包括被耦合到可扩展负载平衡系统的外部客户机。可扩展负载平衡系统包括被配置为对来自该外部客户机的分组流的各单独的传入分组进行封装的负载平衡层。该负载平衡层被进一步配置为将传入分组路由到系统上的各目标设备。各目标设备可以跨多个 IP 子网。各传入分组可以在到达各单独的目标设备之前穿过负载平衡层的一个或多个负载平衡器。各单独的目标设备可以被配置为将分组流的至少一些传出分组路由到外部客户机而不穿过一个或多个负载平衡器中的任何一个。

[0005] 附图简述

[0006] 附图示出了本申请中传达的概念的实现。所示实现的特征可通过参考以下结合附图的描述来更容易地理解。只要可行，各附图中相同的附图标记用来指代相同的元素。此外，每一个附图标记的最左边的数字传达其中首次引入该附图标记的附图及相关联的讨论。

[0007] 图 1- 图 5 示出可以采用根据一些实现的各概念中的一些的网络环境。

[0008] 图 6 示出可以采用根据一些实现的各概念中的一些的可扩展负载平衡体系结构。

[0009] 图 7- 图 8 示出根据各概念的一些实现的在图 1- 图 6 中介绍的一些组件。

[0010] 图 9 示出根据一些实现的与各概念中的一些相一致的散列映射技术。

[0011] 图 10 和图 11 示出根据一些实现的可以实现可扩展负载平衡概念中的一些的流程图。

[0012] 详细描述

[0013] 介绍 / 概览

[0014] 网络负载平衡器可以通过将传入分组分类成会话并将各单独的会话的分组流量分发给所选择的资源（例如，服务器）来帮助增强网络中的资源的利用。为了辅助缓解在负载平衡器处分组流量的瓶颈，可以利用诸如直接服务器返回 (DSR) 等优化技术。DSR 允许来自网络的传出分组流量绕过负载平衡器而不是如同传入分组流量那样穿过它。然而，这种技术通常限于网络的单个虚拟局域网 (VLAN)。相反，图 1 示出各概念中的一些的高级视图。

[0015] 网络示例

[0016] 图 1 示出其中外部客户机 102 可以经由因特网 106 与可扩展负载平衡系统 104 通信的网络环境 100。负载平衡或分散可以被认为是联网设备可以将流量分散在一组有效下一跳的任何合适手段。

[0017] 可扩展负载平衡系统 104 可以包括因为可以支持在 110 处指示的本质上无限量的目标设备而可扩展的负载平衡功能层 108。在这种情况中，术语‘本质上无限’可以一般地意指控制可扩展负载平衡系统 104 的实体所期望的那样多的目标设备。举例来说，目标设备的数量可以是数万个或数十万个或更多个。负载平衡功能层 108 被配置为使得来自外部客户机 102 的通信可以穿过，且被负载平衡功能分发到各单独的目标设备，如箭头 112 所指示。然而，由箭头 114 表示的返回通信不需要在回到外部客户机 102 的路径上穿过负载平衡功能层 108。

[0018] 简言之，一些实现可以利用层 2 域间分组传输技术来实现负载平衡功能 108。在一些情况中，这些层 2 域间分组传输技术可以允许跨越多个 IP 子网使用诸如 DSR 等的负载平衡优化技术，且由此允许使用本质上无限的目标设备 110。出于可扩展性和其他原因，使用网际协议的网络可以将共享它们的 IP 地址中的普通位前缀的主机划分到 IP 子网中。通常，单个子网的范围限于单个 VLAN 的范围。允许使用带有来自不同的子网的网际协议 (IP) 地址的目标设备 110 可以消除用于负载平衡器的先前设计的显著限制。个体 IP 子网可以与可扩展负载平衡系统 104 的若干层 2 域中的一个相关联。在一个或多个实施方式中，可以使用例如 IP-in-IP(在 IP 里面封装 IP) 封装来封装分组流的个体传入分组。这可以例如由负载平衡功能 108 的多路复用器 (MUX 或 Mux) 来完成。

[0019] 通过在到达个体目标设备之前穿过负载平衡功能 108，经封装的传入分组可以被路由到在可扩展负载平衡系统 104 上的资源或目标设备 110。在至少一些实施方式中，负载平衡功能可以使用诸如 DSR 等的优化技术来减少 / 最小化负载平衡功能上的分组流流量。目标设备（例如，服务器）可以与多个 IP 子网或 VLAN 相关联且因而跨越多个 IP 子网或 VLAN。与个体目标设备相关联的组件（例如，软件组件）可以解封所接收的传入分组以便获得 IP 信息。然后，可以将结果（传出分组）从可扩展负载平衡系统 104 路由到外部客户机 102(例如，接收传入分组中的一个或多个的客户机）而不穿过（即穿越）负载平衡功能 108。简要地，可扩展负载平衡系统 104 可以启用新的功能，包括与负载分散和无偿地址解析协议 (G-ARP) 相关联的功能。下面详细叙述这些概念。

[0020] 图 2 示出根据一个或多个实施方式的另一示例网络环境 200。网络环境 200 提供可以实现上面参考图 1 介绍的概念的示例结构或组件。网络环境 200 可以包括经由因特网 106 或其他网络与可扩展负载平衡系统 204 通信的外部客户机 202。可扩展负载平衡系统 204 可以包括一组路由器 206、一组动态负载平衡器 (DLB)208 和一组目标设备 210。在这个实例中，该组路由器 206 表现为路由器 206(1) 和 206(n)。该组 DLB 208 表现为分别包括多路复用器 (即，MUX)212(1) 和 212(n) 的 DLB 208(1) 和 208(n)。该组目标设备 210 表现为应用服务器 214(1) 和 214(n) 以及本地负载平衡器 216(1) 和 216(n)。

[0021] 在 218 处概括示出的虚线箭头示出在可扩展负载平衡系统 204 的各组件之间的潜在通信路径。粗实线箭头 220(1) 和 220(2) 示出通过网络环境 200 从外部客户机 202 到应用服务器 214(1) 的两个潜在的分组流路径。粗实线箭头 222 表示从应用服务器 214(1) 到外部客户机 202 的返回分组流路径。举例来说，粗箭头 220(1) 和 220(2) 可以表示来自外

部客户机 202 的由应用服务器 214(1) 处理的搜索查询。因而，应用服务器 214(1) 可以被称为‘目标设备’。尽管在这里目标设备是应用层应用服务器，但应明白和理解，该示例目标设备可以另外或替代地是另一类型的目标设备，如本地负载平衡器——例如应用层负载平衡器。值得注意的是，尽管传入分组流（即，粗箭头 220(1) 和 220(2)）穿过该组 DLB 208 的成员，但传出返回分组流（即，粗箭头 222）并不必定穿过负载平衡器而是改为绕过 DLB。结果，可以减少在 DLB 中的一个或多个处的分组流流量的瓶颈或使其最小化。在至少一些实施方式中，这可以通过利用 DSR 优化技术来实现。下面描述示例 DSR 优化技术。

[0022] 在至少一些实施方式中，DLB 208(1) 和 208(n) 中的一个或多个上的 MUX 212(1) 和 / 或 212(n) 可以使用 IP-in-IP(IP 中 IP) 封装来将分组流发送给目标设备 210。尽管提供了具体的封装示例，但封装可以是用于对分组进行定址以便沿着路径或路径的一部分传输的任何手段。另外，目标设备上的解封组件 222(1)-222(n) 可以解封传入分组流的一个或多个分组并将结果（即，传出分组流）发送回去给外部客户机 202。在一种情况中，可以在目标设备 210 上将解封组件 222(1)-222(n) 表示成可由目标设备的处理器执行的软件组件。

[0023] 在这种实现中，路由器 206 可以使用等价多路径 (ECMP) 来将分组负载分散在 DLB 208 的 MUX 212(1) 和 212(n) 上。进一步，MUX 可以向发送给目标设备 210 的各分组提供一致散列。在各实现中的一些中，可以在诸如服务器等各单独的设备上实现 DLB 208 和目标设备 210。举例来说，诸如服务器等单个计算设备可以包括带有 MUX 212(1) 和应用服务器 214(1) 的 DLB 208(1)。在其他实现中，DLB 可以在与各目标设备分开的设备上。

[0024] 在操作中，在这一示例中的 DLB 208 中的每一个可以被配置为提供用于管理虚拟 IP- 直接 IP(VIP-DIP) 映射的 VIP 到 DIP 映射（例如，VIP → {槽 $_1$，槽 $_2$，槽 $_3$，...，槽 $_N$}）的应用程序接口 (API)。各单独的槽被分配给 DIP。单个 DIP 可以在这种 VIP 到 DIP 映射中出现多次。这种 VIP 到 DIP 映射可以被称为 VipMap（虚拟 ip 映射）。

[0025] 尽管以上被描述为在单个 VIP 地址和一列 DIP 地址之间的映射，但应理解，每一地址也可以与端口号（例如，诸如端口 80 等传输控制协议 (TCP) 端口）相关联。在这种广义化中，VIP 地址或 VIP 地址和端口号可以被映射到包括作为单独的 DIP 地址或 DIP 地址和端口号的条目的列表。单个 DIP 地址可以出现多次，它可以单独地、或与不同的端口号一起、或与相同的端口号、以任何组合一起出现。也可以存在映射到各 DIP 和各 DIP、端口号组合的相同列表的多个 VIP 或多个 VIP、端口号组合。DLB 208 的各单独的 MUX 212(1)-212(n) 可以各自被配置为对来自各单独的传入分组流分组的首部字段进行散列并将各单独的分组发送给与目标设备 210 相关联的适当的 IP 地址。例如，考虑示例传入分组。DLB 中的一个或两者可以通过计算下式来对示例传入分组散列并选择槽（例如，{槽 $_1$，槽 $_2$，槽 $_3$，...，槽 $_N$}：

[0026] 槽 $_i$ = 散列（分组首部字段）模 N

[0027] 其中 N 是 VIP-DIP 映射中的槽的数量。然后，（各）DLB 的 MUX 可以将示例传入分组发送给槽 $_i$ 中所指示的地址。这种设计的潜在优点是作为同一流（例如，其中所有分组共享 IP 源地址、IP 目的地地址、TCP 源端口、TCP 目的地端口和 IP 协议号的相同 5 元组的 TCP 流）的一部分的各分组可以被转发给相同的目标设备 210，而不管哪个 DLB 208 处理该分组。

[0028] 图3示出提供对以上所描述的网络环境200的替代方案的又一示例网络环境300。简言之，网络环境300类似于网络环境200。然而，在网络环境300中，本地负载平衡器(LLB)可以被认为是在DLB和目标设备之间的居间层。具体地，网络环境300包括经由因特网或其他网络306与可扩展网络平衡系统304通信的外部客户机302。网络平衡系统304包括路由器层308、DLB层310、LLB层312和目标设备层314。在这种情况中，目标设备层314包括应用服务器314(1)-314(n)。LLB层312包括LLB 312(1)-312(n)。

[0029] 解封组件316(1)-316(n)分别驻留在LLB 312(1)-312(n)上。在这种配置中，可以将外部客户机的通信封装在DLB层310处，且当在LLB层312处收到时就解封。然后，可以将该通信转发给适当的应用服务器314(1)-314(n)。到外部客户机302的任何返回通信可以分别绕过DLB层310和LLB层312。绕过DLB层和LLB层可以避免潜在瓶颈和/或为传入的通信保留系统资源。

[0030] 图4示出可扩展负载平衡系统的网络环境400的各组件的又一高级示例。在这个实例中，这些组件包括查询生成器402(1)-402(n)、接入路由器(AR)404(1)-404(n)、层2聚集交换机406(1)-406(n)和架顶式(ToR)交换机408(1)-408(n)。各ToR可以与诸如MUX(M)、健康监视器(H)、服务器(S)、负载平衡器(B)等各种服务器机架组件进行通信。

[0031] 对于提供特定服务的VIP1，各AR 404(1)-404(n)可以被配置为具有N个路线，这些路线中的每一个将其下一跳指向具有相同成本的中间IP(IIP)地址(IIP1至IIPN)。在该AR上，各路线都可以是该VIP的下一跳。因此，该AR可以在N个IIP地址中均匀地分发流量。这些路线可以被配置成该AR上的具有相等规格的静态路线(即，等价静态路线(下面参考图5讨论))。替代地，可以经由与该AR具有适当会话的路由协议(例如，边界网关协议(BGP)或开放最短路径优先(OSPF))发言者(speaker)来动态地建立这些路线。另外，该AR可以被配置为通告该VIP。可以跨各MUX(M)划分各IIP。除了其自己的IP地址(MIP)之外，MUX也可以被配置为带有一个或多个IIP地址，以使得它可以应答对所配置的IIP的ARP请求。因此，单独的MUX可以接收所转发的流量中的一份。一旦接收到分组，该单独的MUX可以运行一致散列算法以便选择一个活动的DLB来转发该流量。

[0032] 基于相同的一组活动DLB，MUX可以使用相同的一致散列算法。因此，可以将分组转发给相同的DLB而不管哪个MUX从AR 404(1)-404(n)接收到它。注意，在向该池添加或从中移除新的DLB时，这可以触发一些本地配置改变；然而，可以保持现有的连接。

[0033] 图5示出用于配置N个等价静态路线的网络环境500和关联的技术。在这种情况中，网络环境500包括接入路由器404(1)(在图4中已介绍)、IIP(1)-IIP(n)、MUX 212(1)-212(n)和DLB 208(1)-208(n)(在图2中已介绍)。网络环境500可为每一VIP配置N个等价静态路线。这些等价静态路线的下一跳指向中间IP(IIP)地址IIP(1)-IIP(n)。可以从独立于VIP池和DIP池的分开的地址池取出这些IIP地址。这种实现也可以开启负载分散以使得流量将被平均分发给这N个IIP地址。

[0034] 在另一实施方式中，可以使用诸如到各路由器的BGP连接等路由协议来将活跃的且取得每一VIP的分组的MUX通知给它们。

[0035] 各种实现可以解决在各MUX模块来来往往时如何保留长期运行连接的问题。在一些实现中所利用的一种方法可以是保留在每一MUX处处理的各单独的流的状态，并在各单独的MUX被添加到可扩展负载平衡系统时将此状态的副本提供给各单独的MUX。为了在不

断开现有连接的情况下处理 MUX 的添加或移除，一个替代方案是在每当由任何 MUX 第一次处理新连接时创建状态信息。可以在各 MUX 之间或者通过对等机制直接地共享这一状态或者通过将这一状态发送给逻辑上集中式的存储来间接地共享这一状态，需要处理连接的各分组的任何 MUX 可以从该逻辑上集中式的存储确定其他 MUX 已经将该连接的各分组发送到的 DIP。

[0036] 一种要求少得多的状态共享且因此更加可扩展的替代实现是，MUX 或者使用 VIP 和 DIP 之间的当前映射（即，VipMap）来转发分组或者处于它们从使用一个 VipMap(V) 来转发分组改变为使用另一 VipMap(V') 来转发分组的过渡时间段中。在这种实施方式中，可以使得各 MUX 就 V、V' 和它们的当前过渡状态达成一致（即是说，它们是处于在 V 和 V' 之间的过渡状态中还是它们都已经对所有分组仅使用 V' 而开始）。在不处于过渡时，所有 MUX 使用当前的 VipMap 来转发所有分组。在处于过渡时，每当它们看到新连接的分组（例如，TCP SYN 分组）时，MUX 创建一段本地状态。在转发与指示新连接的那些分组不同的分组时，MUX 查看它是否拥有该连接的状态。如果它拥有状态，则 MUX 使用新 VipMap V' 来转发该分组，否则它使用旧 VipMap V 来转发该分组。

[0037] 简言之，在至少一些配置中，MUX 212(1)-212(n) 可以具有下列主要组件：(1) 向路由器主张拥有一 IIP 并接收该 IIP 的流量的 IIP 模块，(2) 确定哪个 DLB208(1)-208(n) 转发该流量的一致散列模块，(3) 修改分组的分组重写器，(4) 本地 DLB 监视器。在各种实现中，可以在容易获得的（即，商品）服务器上和 / 或在路由器上实现这些组件中的任何或全部。下面参考图 6- 图 8 更详细地描述各 MUX 组件。

[0038] IIP 模块 (IIP(1)-IIP(n)) 可以负责通过 ARP 协议将 MUX 212(1)-212(n) 注册到路由器。基本上，IIP 模块可以在路由器上建立 IIP 地址和 MUX MAC 地址的 IP-MAC 映射。

[0039] 考虑示例函数 'bool AddIP(IP Address iip)'：在这一示例函数中，IIP 地址可作为次级 IP 地址合计在 MUX 接口上。注意，MUX 可能具有多个次级 IP 地址。'AddIP()' 可以使得 MUX 网络栈发送 3 个无偿 ARP(G-ARP) 请求，这些无偿 ARP 请求可以更新路由器的 ARP 表（或触发其自身上的 IP 地址冲突检测）。

[0040] 出于解释的目的，考虑示例函数 "RemoveIP(IPAddress iip)"：此示例函数可以从 MUX 接口移除 IIP 地址。还考虑示例函数 "SendARP()"。此示例函数可以强制发送 G-ARP 请求。这一 G-ARP 请求可以作为 IIP-MAC 映射的正确性的预防措施而发送。

[0041] G-ARP 和地址冲突检测

[0042] 在将 IP 地址添加到该接口时，操作系统 (OS) 可以广播 G-ARP（在同一 L2 域内）。此 G-ARP 请求可以请求它正在主张的 IP 地址。如果没有其他机器以此 IP 地址应答，则可以成功添加该 IP 地址。否则，可以检测到 IP 地址冲突，且 MUX 栈可以阻止该机器主张此 IP 地址。如果另一 MUX 已经主张该 IIP（例如故障切换）且未能移除它，则这种情况可以发生。可以通过外部措施（例如切断防护机器）来应对这种场景。

[0043] 出于示例的目的，在新 MUX MUX "B" 需要代替 MUX "A"（例如，因为 MUX A 的计划停机时间和 / 或在 MUX A 处的系统故障）时，该新 MUX B 可以将 MUX A 的（各）IIP 添加到其自己的接口。

[0044] 在至少一种实施方式中，例如上面所述的模块可以将分组流定向到服务器池中的一个或多个有状态模块中，其中有状态模块可以保持每个流状态。在这种情况中，入站分组

可以流过从客户机到模块到有状态模块到处理相关联的请求的目标服务器的路线。出站流可以流过从目标服务器到有状态模块到客户机的路线。在有状态模块处的每个流状态可以允许各单独的有状态模块应用流级策略以便支持附加的负载平衡特征。具体而言，有状态模块可以，例如，检查 cookies 或 URL 以便将到目标服务器的负载平衡自定义为取决于应用、客户机请求和 / 或服务器和网络元件的角色和 / 或负载和 / 或状况。此实施方式是有益的，这是因为它可以将 CPU 和状态密集的工作量分散到必要多的服务器。

[0045] 在至少一种实施方式中，该模块可以使其到有状态模块的路由适应为取决于比 TCP/IP 首部和应用首部中携带的首部信息更深的信息。具体而言，为了支持直接访问函数，例如在 Windows 7®中，该模块可以学习或参与密码协议，从而允许对分组的各部分进行解密。然后，有状态模块对目标服务器的选择可以取决于这些经解密的部分。该机制可以被构建为使得目标服务器将出站流返回给能够（且可能最适合）处理它的有状态模块。这可以受益于使用可编程 CPU 来实现该模块。

[0046] 在至少一种实施方式中，该模块可以在诸如网际协议 (IP) 选项等分组首部的某一部分中包括原始目的地地址，并将该分组发送给目标设备。目标设备可以从分组首部提取这一信息并使用它来将传出分组直接发送给源（例如，外部客户机），其中分组中的一些不穿过该模块。

[0047] 图 6 示出可以实现上文和下文所描述的概念的示例可扩展负载平衡系统体系结构 600。在这种情况中，可扩展负载平衡系统体系结构 600 可以包括可扩展负载平衡管理器 602、在 604 处表示的 MUX 角色和在 606 处表示的 DIP 角色。负载平衡系统体系结构 600 还可以包括健康监视器 608、健康探头 (probe)610 和路由管理器 612。MUX 角色 604 可以涉及在用户模式 616 中操作的 MUX 控制器 614 和在内核模式 620 中操作的 MUX 驱动程序 618。DIP 角色 606 可以涉及在用户模式 624 中操作的 DIP 控制器 622 和在内核模式 628 中操作的解封驱动程序 626。

[0048] 可扩展负载平衡管理器 602 可以被认为是与可扩展负载平衡系统体系结构 600 交互的入口点。可扩展负载平衡管理器 602 可以提供可以被用来管理可扩展负载平衡概念的实例的 API。可以使用 XML 配置或 API 来指定可扩展负载平衡实例。

[0049] 可扩展负载平衡管理器 602 可以负责在 MUX 机器上配置 VIP:DIP 映射并确保 MUX 机器保持同步。此外，在向池添加 DIP 或从中优雅地移除 DIP 时，可扩展负载平衡管理器 602 还可以便于保留长期运行的连接。下面参考图 9 更详细地描述这种特征。

[0050] 为了增加可用性，可以复制可扩展负载平衡管理器 602 且可以使用主可扩展负载平衡管理器选择算法来确保状态的一致性。

[0051] MUX 角色 604 可以被配置为带有一个或多个中间 IP 地址 (IIP)。如上面参考图 4 所描述的，诸如路由器 404(1) 等路由器可以被配置为向一组 IIP 转发去往 VIP 的分组。被配置为带有给定 IIP 的 MUX 将执行向该 IIP 转发的分组的 MUX 处理。

[0052] MUX 控制器 614 可以控制 MUX 驱动程序 618。MUX 控制器可以导出由可扩展负载平衡管理器 602 用来控制该 MUX 的 web 服务 API。在一些实现中，MUX 控制器可以执行下列功能：

[0053] 1. 将 VIP:DIP 下载到驱动程序；

[0054] 2. 向该驱动程序告知长期运行的连接；

[0055] 3. 从该驱动程序收集统计数据；

[0056] 4. 在网络接口上配置 IIP；

[0057] 5. 在网络上发出针对所指定的 IIP 的 G-ARP 分组，以便由路由器或网络上的其他主机将向该 IIP 转发的任何分组吸引到该 MUX。

[0058] MUX 驱动程序 618 可以实现基础分组修改功能。MUX 驱动程序可以对传入分组的首部字段进行散列，基于散列值和当前 VIP 映射来拾取针对它的 DIP，以及封装该分组以供传输。除了映射之外，MUX 驱动程序 618 还可以维持散列的高速缓存：每一 VIP 的所有长期运行连接的 DIP 映射。

[0059] DIP 控制器 622 可以控制 DIP 机器上的解封驱动程序 626。类似于 Mux 控制器 614，DIP 控制器 622 可以导出由可扩展负载平衡管理器 602 用来控制和查询 DIP 机器的 web 服务 API。在一些实现中，DIP 控制器 622 可以执行下列功能：

[0060] 1. 在回送接口上配置各 VIP；

[0061] 2. 配置用于所指定的 VIP 的解封；

[0062] 3. 查询 DIP 机器以寻找当前活动的连接；

[0063] 4. 查询 DIP 机器的健康（取决于健康监视器实现，这是可选的）。

[0064] 解封驱动程序 626 可以解封去往所指定的 VIP 的 IP-in-IP 分组。这一特征帮助避免断开与具体应用的正在进行的通信。例如，如果存在正在使用原始套接字来发送 IP-in-IP 的应用（例如，虚拟专用网 VPN 应用），则解封驱动程序 626 不解封那些。

[0065] 路由管理器 612 可以负责在向池添加或从中移除 MUX 机器时配置路由器。路由管理器可以使用诸如 OSPF 或 BGP 等路由协议或接口来在各路由器上配置静态路线。

[0066] 健康监视器 608 可以负责维持 MUX 和 DIP 机器的健康状态，且可能负责在请求处理中所涉及的路线。为此，健康监视器可以监视一个或多个网络参数，这些参数在确定网络和 / 或网络组件的健康时是有价值的。可扩展负载平衡管理器 602 可以将健康监视器 608 用作关于 MUX 和 DIP 的健康信息的权威源。如果健康监视器 608 向可扩展负载平衡管理器 602 通知健康改变事件，则可扩展负载平衡管理器可以采取向相应的池添加或从中移除该节点的适当动作。

[0067] 从一个角度来看，健康监视器 608 可以被用来监视 MUX、DLB 的健康和 / 或到这些机器的路线。

[0068] 在至少一些实现中，健康监视器 608 可以包括三个模块：VPN 拨号器、MUX 监视器和 DLB 监视器。DLB 可以提供 HTTP 接口。健康监视器 608 可以采用各种健康探头 610 来确立目标组件的健康。例如，健康监视器可以发送“http get”以便从 DLB 获取小的文本 / xml 文件。如果该文件包含健康监视器和 DLB 所达成一致的‘魔术词 (magic word)’，则健康监视器可以认为该 DLB 已启动且正在运行，并确定 DLB 或 MUX 是否正在如所预期的那样运行。此外，在至少一些实施方式中，健康监视器组件可以存在于分开的设备上而不是 MUX 设备上。

[0069] 健康探头 610 可以由健康监视器 608 使用。举例来说，健康监视器可以使用各种健康探头来完成其作业。健康探头 610 可以主动地监视目标机器的健康的一个方面，例如，ping 探头监视机器的连接性和活性。其他健康探头可以简单地查询机器 / 角色以得到其健康——该机器 / 角色可以负责维护其健康的记录，探头只是周期性地查询它。

[0070] 如果 HTTP 探头是成功的，则这可以指示所有事物都开启且正在运行。但是由于它在 TCP 上运行，所以 DLB 临时地用光了套接字或其他资源是可能的。在拒绝服务 (DoS) 攻击期间，DLB 在持续时间周期内用光了资源（例如，套接字）也是可能的。对此的一种解决方案可以是维持持久的 HTTP 连接。然而，大多数服务器 / 浏览器实现将使得持久的 TCP 连接超时。例如，一些浏览器可以在 60 秒之后使得持久连接超时。因此，健康监视器准备在持久连接被关闭的情况下重建该持久连接，且不必将持久连接的关闭看作是指示 DIP 故障。

[0071] 如果另一 MUX 可以接管故障的 MUX，则由于所有 MUX 运行相同的一致散列函数，所以各分组将被转发给相同的 DLB。因此，该流（例如，TCP 连接）不会被中断。

[0072] 一分开的 MUX 池可以被用作活动 MUX 的热备份。一旦检测到 MUX 故障，健康监视器 608 就可以开始一个或多个 MUX 以便接管故障 MUX 的各 IIP。在同一时刻，健康监视器可以切断该故障 MUX。为了处理 MUX 的计划停机时间，可以使用与用于热备份的技术相似的技术。由于 MUX 以无状态模式操作，一些实现可以在从该 MUX 排出了所有分组之后安全地切断该 MUX。

[0073] 在至少一种实施方式中，可以通过有状态 MUX 映射过渡来处理 DLB 计划停机时间。

[0074] 1. MUX 正在使用使用 DLB(D) 的 VipMap(V)；

[0075] 2. 向 MUX 通知 DLB(D) 在 T 时刻停机；

[0076] 3. MUX 计算不使用 DLB(D) 的新 VipMap(V')；

[0077] 4. MUX 将驱动程序置于 (V- ＞ V' 过渡模式)；

[0078] 5. 在过渡中，保持状态表，且每一 TCP SYN 将在表中造成新条目；

[0079] a. 如果分组与状态表中的一条目匹配，则它是新的流且因此使用 V'；

[0080] b. 否则，使用旧的 V；

[0081] 注意：在这一过渡时间段期间，任何新的流将切换到新 VipMap(V')，从而避开了 DLB(D)。

[0082] 6. DLB(D) 保持对（到 VIP 的）活动 TCP 连接的数量进行计数。在计数器达到 0 时，它向 MUX 通知该过渡已完成。

[0083] 7. 替代地，MUX 可以将长期运行连接标识为不与状态表中的任何条目相匹配的连接。

[0084] 8. 在达到时间 T 时，强加过渡 V- ＞ V'。MUX 将基于 V' 来转发所有流量。

[0085] 在一种实施方式中，经由下列步骤来处理 MUX 计划停机时间：

[0086] 1. 在新 MUX(M') 上设置 VipMap；

[0087] 2. 设置旧 MUX(M) 以便将所有 VIP 流量转发给 M'，M' 照常将流量转发给 DLB；

[0088] 3. 从旧的 MUX(M) 移除 IIP；

[0089] 4. 将 IIP 添加到新 MUX(M')；以及，

[0090] 5. 路由器应开始向新 MUX 转发。

[0091] 在至少一种实施方式中，健康监视器 608 可以将周期性的探头发送给 MUX 和 DLB 以便监视非预期故障。在观察到 DLB 故障时，健康监视器可以指示 MUX 更新其 VipMap 以便避免使用故障的 DLB。在观察到 MUX 故障时，健康监视器可以指示在同一 VLAN 中的另一 MUX 安装该 IIP（且使用 G-ARP 来向路由器通告）。在至少一种实施方式中，健康监视器可以每两秒发送 KeepAlive（保持活动）探头，且在 3 次连续失败之后通告 MUX/DLB 死亡。

[0092] 为了实现用于关键任务 VIP 的快速 MUX 故障切换（＜＜ 1 秒），可以利用每一 IIP 的 MUX 虚拟组。这种快速故障切换的成本可以是在正常操作期间更多使用网络。可以使用下列步骤用来为 VIP 管理 MUX 和 IIP：

[0093] A. 每一 IIP 可以是多播地址。每一 VIP 具有被分配给它的一组 MUX。

[0094] B. 该组的主 MUX 是该 IIP 的实际持有者。

[0095] C. 主 MUX 向该组的所有成员发送它是此 VIP 的活动 MUX 的多播通告。以高速率（＜＜ 1 秒）发送这一通告。这一通告也防止其他 MUX 开始新的主 MUX 选举过程。

[0096] D. 由于 IIP 可以是多播地址，所以上游路由器将它所接收到的每一分组复制到该 VIP 组中的各 MUX 成员（主 MUX 和所有备份 MUX）。

[0097] E. 所指定的备份 MUX 将这些分组存储所指定的时间 T。

[0098] F. 主 MUX 对这些分组执行负载平衡功能并将它们转发给各 DLB。

[0099] G. 如果在给定的时间 T 内没有接收到主 MUX 活动的通告，则所指定的备份 MUX 将开始负载平衡并转发其缓冲区中的所有分组。

[0100] H. 这一组中的各备份 MUX 将开始新的主 MUX 选举过程。在一些配置中，所指定的备份 MUX 可以变成新的主 MUX。

[0101] I. 步骤 G 可以使得 DLB 两次接收一些分组，但是 TCP 足够好地容忍重复分组和暂时的分组丢失。注意，只要上游路由器活动且执行良好，就可以不发生分组丢失。

[0102] 图 7 示出根据一个或多个实施方式的 MUX 212(1)（在图 2 中已介绍）的示例配置。图 7 和图 8 一起示出如何沿着路径封装和解封各分组。

[0103] 图 7 涉及用户模式 702 和内核模式 704，但是集中于由 MUX 的在内核模式中的 MUX 驱动程序 618 提供的功能。在这种情况中，MUX 驱动程序被实现为网络栈的 IP 层的扩展。

[0104] 在这一示例中，由 MUX 驱动程序 618 例如从应用服务器接收分组 706。该分组包括在 708 处的源客户机地址和在 710 处的目的地 VIP 地址。分组迁移通过物理网络接口卡 (NIC) 层 712 和网络驱动程序接口规范 (NDIS) 层 714。该分组由 IP 层 718 中的 MUX 驱动程序的转发器 716 来处理。该转发器封装分组 706 以便生成分组 720。此分组包括由源 MUX 地址 722 和目的地 DIP 地址 724 封装的在 708 处的源客户机地址和在 710 处的目的地 VIP 地址。因而，以给出它是来自 MUX212(1) 而不是来自客户机 708 的印象的方式将原始分组 706 封装在分组 720 中。

[0105] MUX 212(1) 可以实现也称为 VIP:DIP 映射的层 4 负载平衡。可以由层 1 将来自客户机的流量发送到各 MUX 节点中的一个（通常经由等价多路径 (ECMP) 路由）。在 MUX 212(1) 接收到分组 706 时，它可以将各分组首部字段进行散列（在散列哪些字段方面是灵活的）且可以基于此散列来拾取 DIP。（下面参考图 9 描述此过程的示例）。然后，该 MUX 可以将原始分组 706 封装在新的 IP 首部中，该新的 IP 首部将所选择的 DIP 指示为目的地（即，目的地 DIP 地址 724）并将该 MUX 指示为源 IP 地址。（替代地，MUX 可以将原始发送者用作源 IP。）

[0106] 负载平衡集群中的各 MUX 节点可以使用相同的散列函数。此外，MUX 节点可以在 DIP 的添加和优雅删除期间维持状态。这可以允许将给定流的各分组转发给下一层中的相同服务器而不考虑哪个 MUX 接收该分组。

[0107] 图 8 示出上面参考图 6 介绍的 DIP 角色 606 的示例。简言之，在这种情况中，DIP

解封驱动程序 626 可以对图 7 中所介绍的经封装分组 720 执行解封。在这种配置中，将 DIP 解封驱动程序被实现成网络栈的 IP 层的扩展。如上所述，图 7 提供用于在传输路径的前端处实现封装的示例，图 8 提供在后端处解封上面介绍的原始分组 706 的示例。

[0108] 在这一示例中，解封驱动程序 626 可以接收经封装的分组 720。一旦经封装的分组在路径上行进且准备好传输给目的地 VIP 地址 710，解封驱动程序就可以移除封装（即，源 MUX 地址 722 和目的地 DIP 地址 724）以便产生分组 706。

[0109] 上面所描述的 MUX 212(1) 和 DIP 角色 606 可以与本发明的各概念一起使用以便于诸如分组 706 等分组的封装，该分组 706 与带有位置地址（即，目的地 DIP 724）的应用地址（即，目的地 VIP 地址 710）相关联，以使得分组 706 可以在层 3 基础设施上传输且最终被递送到层 2 目的地 VIP 地址 710。此外，经封装的分组可以在由该封装所定义的路径上行进，且可以容易地为随后的分组重新选择所选择的路径以便避免拥塞。

[0110] 此外，这种配置可以便于网络节点池（即，可扩展负载平衡系统 104、204 和 / 或 304 的各组件）的免中断（或减少的中断）的生长和缩小。简言之，可扩展负载平衡系统状态不倾向于是静态的。举例来说，更多的应用服务器可以上线和 / 或各应用服务器可以下线，交换机可以来来往往，通信可被发起和结束等等。本发明的各概念可以允许从现有的可扩展负载平衡系统映射到新的可扩展负载平衡系统映射的优雅过渡。举例来说，本发明的各概念可以跟踪现有映射的现有或正在进行的通信。在利用反映可扩展负载平衡系统针对新通信的改变的新映射的同时，一些实现可以尝试利用现有映射来维持那些正在进行的通信的连续性。然后，这些实现可以以相对无缝的方式从旧映射‘优雅地’过渡到新映射。

[0111] 图 9 示出将散列空间映射到 DIP 池的示例方法 900。举例来说，该映射可以允许从 VIP 池移除 DIP 而不中断对不去往受影响的 DIP 的流量。举例来说，在 902 处示出在散列空间（即，可能的散列值）和可用的 DIP 的池之间的第一映射。在 904 处示出在该散列空间和可用的 DIP 的不同的池的第二映射。在这种情况中，第二映射 904 作为如在 906 处所示的 DIP 1 停机（即，变得不可用）的结果而发生。最初参见第一映射 902，各散列值在 908(1)、908(2)、和 908(3) 处被映射到 DIP 1，在 910(1)、910(2) 和 910(3) 处被映射到 DIP 2，在 912(1)、912(2)、和 912(3) 处被映射到 DIP3，且在 914(1)、914(2) 和 914(3) 处被映射到 DIP 4。因而，以可以减少或避免瓶颈的方式在可用的 DIP 当中分配各散列值。

[0112] 在 906 处在 DIP 1 消失的情况下，这一实现以避免突然使得任何单独的可用 DIP 过载的方式在剩余的可用 DIP 之间重新分配 DIP 1 的负载。举例来说，在第二映射 904 中，在第一映射 902 中的散列的在 908(1) 处被映射到 DIP 1 的第一部分被重新分配给 DIP 2，如在 916 处所指示。DIP 1 的第二部分 908(2) 被重新分配给 DIP3，如在 918 处所指示。DIP 1 的第三部分 908(3) 被重新分配给 DIP 4，如在 920 处所指示。因而，这一实现以可以避免使得任何其余 DIP 过载且由此避免了潜在地造成与过载的 DIP 相关联的瓶颈的平衡方式无缝地将分组流从如第一映射 902 中所示的四路分布重新分配成第二映射 904 中所示的三路分布。

[0113] 出于更详尽的解释的目的，考虑具有确定 VIP 到一个或多个应用服务器 (DLB) 的映射的 VIP-DIP 映射 M 的 MUX（例如 MUX 212(1)）。现在考虑其中要将 M 改变成 M’的场景。利用所描述的技术，可以将 M 优雅地改变成 M’。由于可以存在长寿命的连接，所以可选地，可以定义最后期限 T。然后，一旦达到 T 或者一旦完成了优雅改变，该 MUX 就可以将 M 改变

成 M'。

[0114] 下面描述的仅是将 M 优雅地改变成 M' 的方式一个示例 :

[0115] 对于分组 P, MUX 可以计算 H(P) 和 H'(P) 两者,其中可以使用映射 M 来计算 H(P) 且可以使用映射 M' 来计算 H'(P)。

[0116] - 如果 H(P) = H'(P),则转发给 H(P) 等效于转发给 H'(P) ;

[0117] - 如果 H(P) ! = H'(P) 且 P 是 SYN(TCP SYN 分组,其可以发起 TCP 连接),则 P 可以被用来建立新的连接,该新连接应去往 H'(P),且插入散列 (P)(hash(P)) → H'(P) 也可以被插入到状态表 S,以使得此流可以被认为是已经被移到 M' ;

[0118] - 如果 H(P) ! = H'(P) 且 P 不是 SYN,并且散列 (P) 不在 S 中,则这可以是正在进行的到 H(P) 的连接的一部分,因此继续到 H(P) ;

[0119] - 如果 H(P) ! = H'(P) 且 P 不是 SYN,并且散列 (P) 在 S 中,则这可以是正在进行的已经被移到 M' 的连接的一部分,因此继续到 H'(P) ;

[0120] - 在达到 T 或所有 DLB 告知过渡已完成时,可以将映射从 M 改变为 M',且可以对状态表 S 进行转储清除。

[0121] 相应地,DLB 可被告知相同的 M- > M' 过渡,且然后,则它可以计算它 (即,该 DLB) 是否受到此过渡的影响。

[0122] 如果 DLB 确定它正在被过渡出去,则它可以优雅地耗尽它拥有的连接。

[0123] 对于持久的 HTTP 连接,DLB HTTP 服务器可以禁用 'HTTP Keepalive'(HTTP 保活)。如此,DLB HTTP 服务器可以使用 FIN 来终止 (TCP FIN 分组,其结束 TCP 连接) 底层 TCP 连接。FIN 可以被认为是 TCP 首部中指示此分组的发送者想要终止连接的标志。外部客户机可以重启连接。然而,这将可能开始新的握手,对于该新的握手, MUX 可以将该新的 TCP 连接路由到新的 DLB。

[0124] 替代地,可以类似于下面描述的所建立的 TCP 连接来处理持久的 HTTP 连接。

[0125] - 在过渡时间段期间,所建立的 TCP 连接可以是寂静的或繁忙的,且可以预期 HTTP 关闭它。一些可能的动作是 :

[0126] ● 1. 令 TCP 连接在客户机侧超时。基本上,此技术仅仅忽略这些 TCP 连接。

[0127] ● 2. 在达到时间 T 时向客户机强制发送 TCP RST 以便告知该客户机。发送 RST 不要求具有正确的序列号。因而,此技术可以只是枚举“已建立的”连接且结束所有已建立的连接。

[0128] ● 3.MUX 可以维持持久连接的状态,直到 DLB 确定已经终止受过渡影响的各连接。

[0129] - 在打开 TCP 套接字的数量是 0 时,可以告知 MUX 可以从该池安全移除该节点。

[0130] 总之,本发明的各实现可以使用 IP-in-IP 封装以便可以跨越可能全部目标设备而不仅是子网来使用 DSR。此外,可以按需将负载平衡器实现为可扩展逻辑层。各概念也可以在系统过渡期间保留连接。举例来说,在优雅地过渡各连接的同时,可以添加或移除 DIP、可以重新平衡负载并且 / 或者可以调整系统容量。可以在 MUX 层实现一致散列以便允许可扩展性并允许移除故障的 DIP 而不保持状态。此外,系统监视、控制和 / 或管理功能可以与负载平衡功能位于一处。这可以允许主 MUX 确保在各 MUX 之间的地址连续性以及其他潜在的优点。

[0131] 第一方法示例

[0132] 图 10 示出根据一个或多个实施方式的参考 VIP 的 DIP 池的扩展来描述与保留长期运行连接相关联的示例的各步骤或动作的方法 1000 的流程图。

[0133] 可以结合任何合适的硬件、软件（例如，包括固件）或其任何组合实现该方法。在一些情况中，该方法可以被存储在可由计算设备的处理器执行以便执行该方法的计算机可读存储介质上。此外，可以重复该方法的各步骤中的一个或多个任何次数。另外或替代地，在至少一些实施方式中可以省略各步骤中的一个或多个。

[0134] 在步骤 1002，标识网络或可扩展负载平衡系统的新连接。在至少一些实施方式中，这可以通过查找 TCP SYN 来完成。

[0135] 在步骤 1004，保持新连接的状态。

[0136] 在步骤 1006，对现有或旧连接使用现有或旧散列，且可对新连接使用新散列。

[0137] 在步骤 1008，查询各 DIP。在至少一些实施方式中，这可以包括查询 DIP 以便得到要保持的长期运行连接。替代地，负载平衡系统可以通过解释分组首部来确定活动连接。

[0138] 在步骤 1010，使新连接的状态过期。

[0139] 在步骤 1012，使所保持的连接的状态过期。在至少一些实施方式中，这可以包括在所保持的连接在 DIP 处终止时使得所保持的连接的状态过期。

[0140] 出于解释性目的提供方法 1000，且不应以限制方式来理解方法 1000。举例来说，在过渡期间可以采用的替代方法可以利用下列算法：

[0141] 1. 通过解释分组首部（例如查找 TCP SYN）来标识新连接发起分组；

[0142] 2. 如果它是新连接发起分组，则仅根据新映射来发送它；

[0143] 3. 否则根据旧映射和新映射两者发送该分组；

[0144] 4. 通过询问 DIP 或跟踪在某一时间段期间在负载平衡器处的状态来标识旧连接；

[0145] 5. 根据旧映射发送旧连接且根据新映射发送新连接；以及

[0146] 6. 在超时之后或在旧连接在 DIP 上终止时，使得关于旧连接的状态过期。

[0147] 第二方法示例

[0148] 图 11 示出描述示例方法 1100 的各步骤或动作的流程图。可以结合任何合适的硬件、软件（例如，包括固件）或其任何组合实现该方法。在一些情况中，该方法可以被存储在可由计算设备的处理器执行以便执行该方法的计算机可读存储介质上。此外，可以重复该方法的各步骤中的一个或多个任何次数。另外或替代地，在至少一些实施方式中可以省略各步骤中的一个或多个。

[0149] 在步骤 1102，可以将各网络分组分散在一系列模块之间。在至少一种实施方式中，各模块是被配置为在服务器上和 / 或在路由器中实现的 MUX 模块。除了可以将到一目的地的各分组递送到包含处理该目的地的各分组所需要的状态的 MUX 模块之外，分散可以无视各分组的各单独的特性。在至少一些实施方式中，使用 ECMP 路由器将各单独的网络分组分散在各模块之间。

[0150] 在步骤 1104，网络分组可以被封装在各单独的模块。在至少一些实施方式中，该分组的封装包括 IP-in-IP 封装和 / 或保留该分组被发送到的一个或多个 VIP 地址。在这一点上，应注意，在此描述的各技术的可能有价值的特征与以下相关联：基于各分组的特性（例如，5 元组，IP 源地址、IP 目的地地址、IP 协议号、TCP 源端口和 / 或 TCP 目的地端口）封装各网络分组，以使得作为同一请求的一部分的各分组在一些实施方式中可以全部由同

一目标设备处理，而不管哪个 MUX 模块封装了该分组。

[0151] 在步骤 1106，可以使用在各模块之间共享的状态来选择将各网络分组封装到的目标设备。在至少一些实施方式中，在各模块之间共享的状态是一致散列函数的键空间。另外或替代地，在至少一些实施方式中，可以响应于目标设备的故障而改变在各模块之间共享的状态。

[0152] 在步骤 1108，可以从各模块转发各网络分组。

[0153] 在步骤 1110，可以监视目标设备、MUX 模块、路由器的健康和在各组件之间的路线。

[0154] 结论

[0155] 尽管已经用对结构特征和 / 或方法论动作来说专用的语言描述了关于负载平衡场景的各技术、方法、设备、系统等等，但应理解，在所附权利要求中限定的主题并不必定限于所描述的具体特征或动作。相反，这些具体特征和动作是作为实现所要求保护的方法、设备、系统等等的示例性形式而公开的。

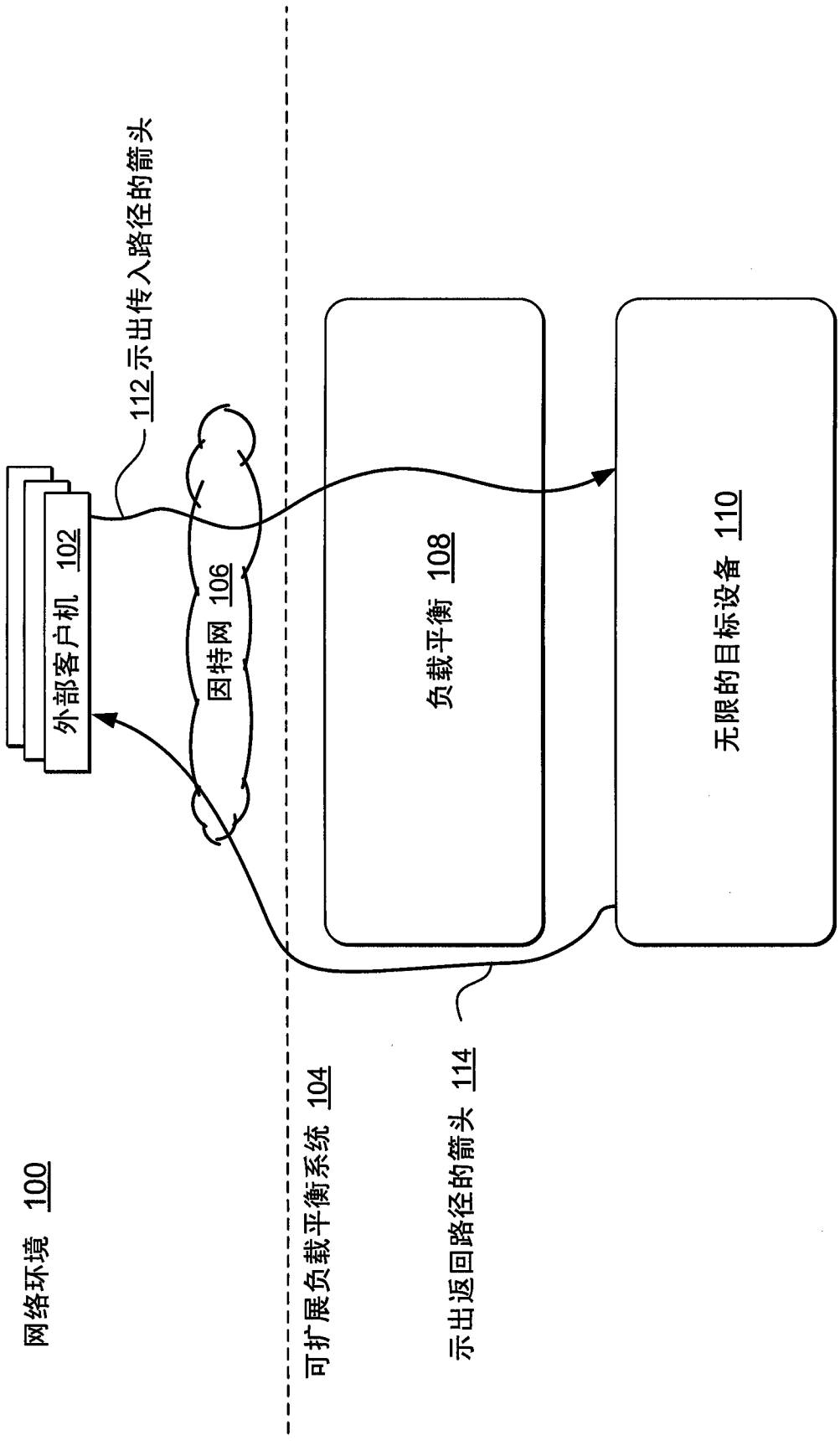

网络环境 100
外部客户机 102
112 示出传入路径的箭头
因特网 106
可扩展负载平衡系统 104
负载平衡 108
无限的目标设备 110
示出返回路径的箭头 114

图 1

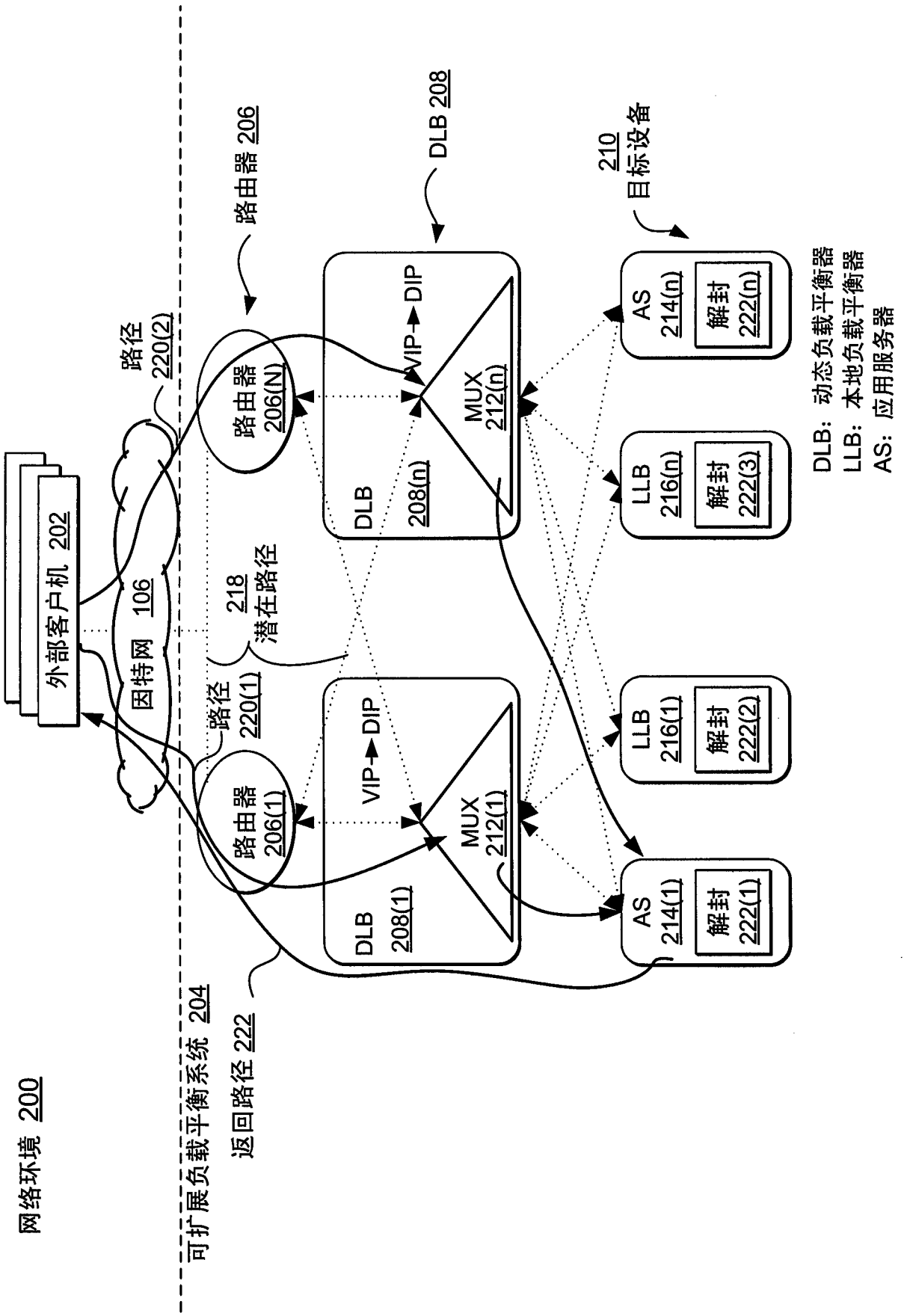

网络环境 200
可扩展负载平衡系统 204
外部客户机 202
因特网 106
路径 220(2)
路径 220(1)
返回路径 222
218 潜在路径
路由器 206
路由器 206(N)
路由器 206(1)
DLB 208
DLB 208(n)
DLB 208(1)
VIP→DIP
MUX 212(n)
MUX 212(1)
210 目标设备
AS 214(n)
AS 214(1)
LLB 216(n)
LLB 216(1)
解封 222(n)
解封 222(3)
解封 222(2)
解封 222(1)
DLB：动态负载平衡器
LLB：本地负载平衡器
AS：应用服务器

图 2

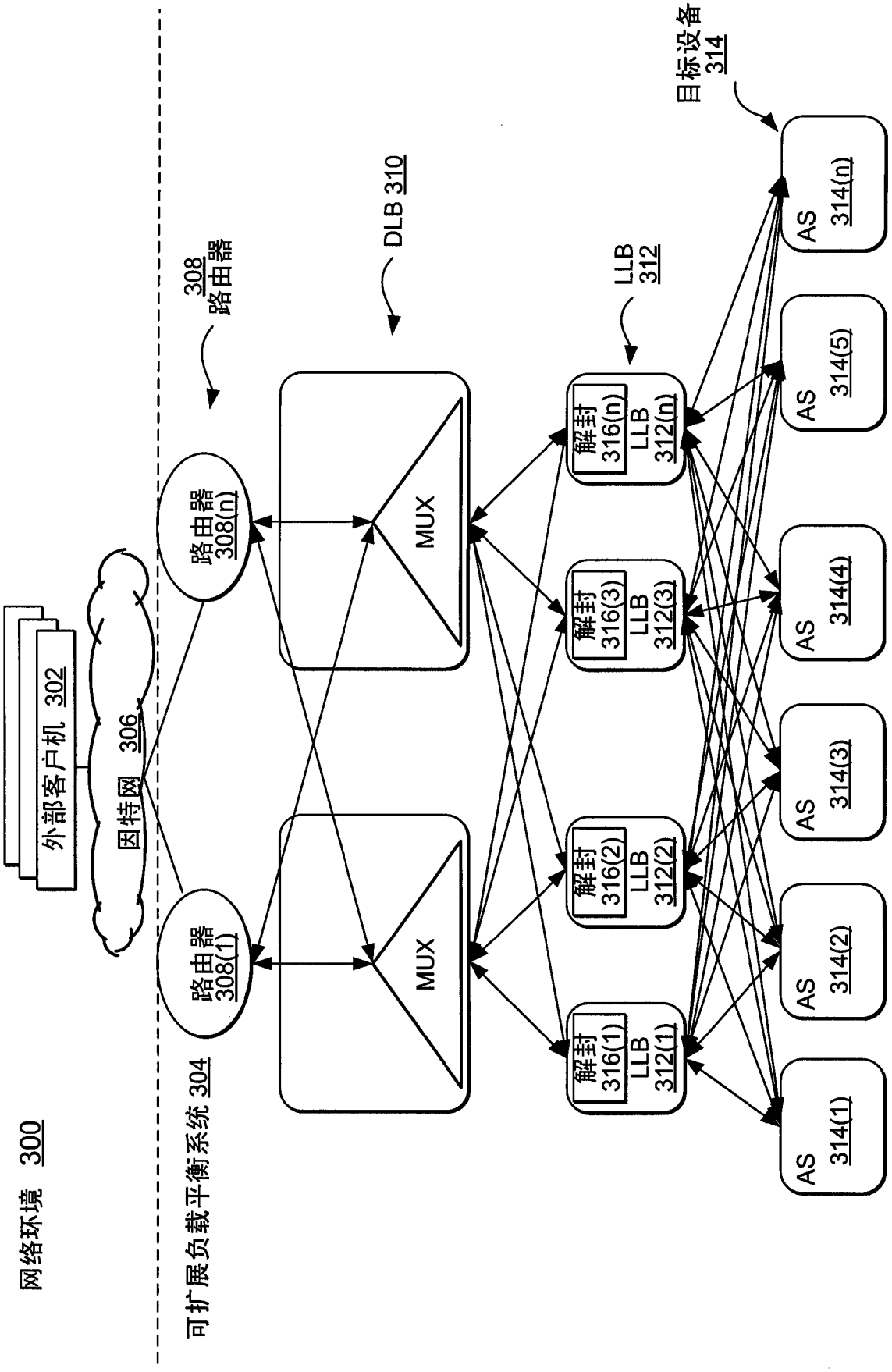

网络环境 300
外部客户机 302
因特网 306
可扩展负载平衡系统 304
路由器 308(1)
路由器 308(n)
308 路由器
DLB 310
MUX
MUX
LLB 312
解封 316(1) LLB 312(1)
解封 316(2) LLB 312(2)
解封 316(3) LLB 312(3)
解封 316(n) LLB 312(n)
目标设备 314
AS 314(1)
AS 314(2)
AS 314(3)
AS 314(4)
AS 314(5)
AS 314(n)

图 3

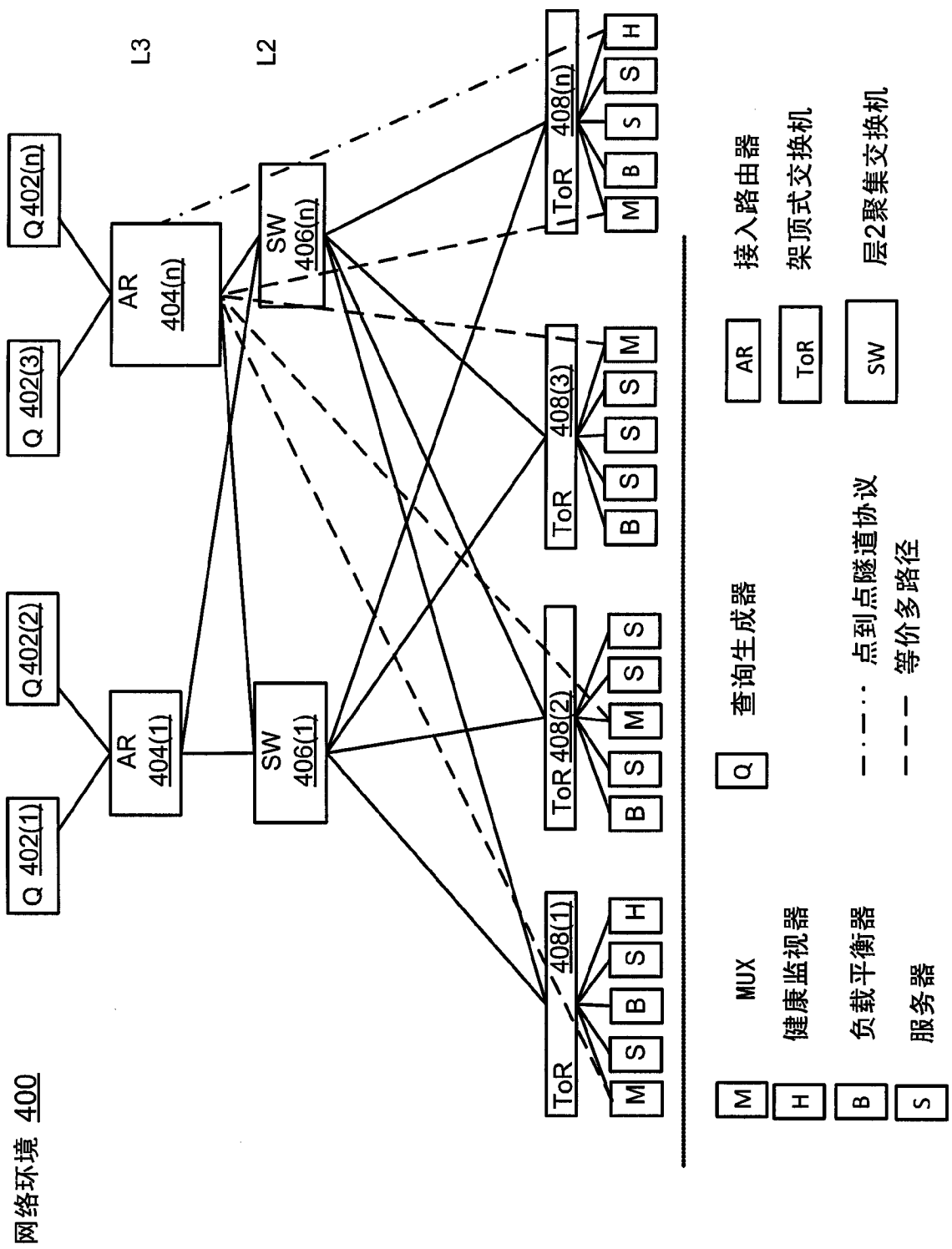

网络环境 400
L3
L2
Q 402(1)
Q 402(2)
Q 402(3)
Q 402(n)
AR 404(1)
AR 404(n)
SW 406(1)
SW 406(n)
ToR 408(1)
ToR 408(2)
ToR 408(3)
ToR 408(n)
M
B
S
H
Q 查询生成器
AR 接入路由器
ToR 架顶式交换机
SW 层2聚集交换机
MUX
健康监视器
负载平衡器
服务器
点到点隧道协议
等价多路径

图 4

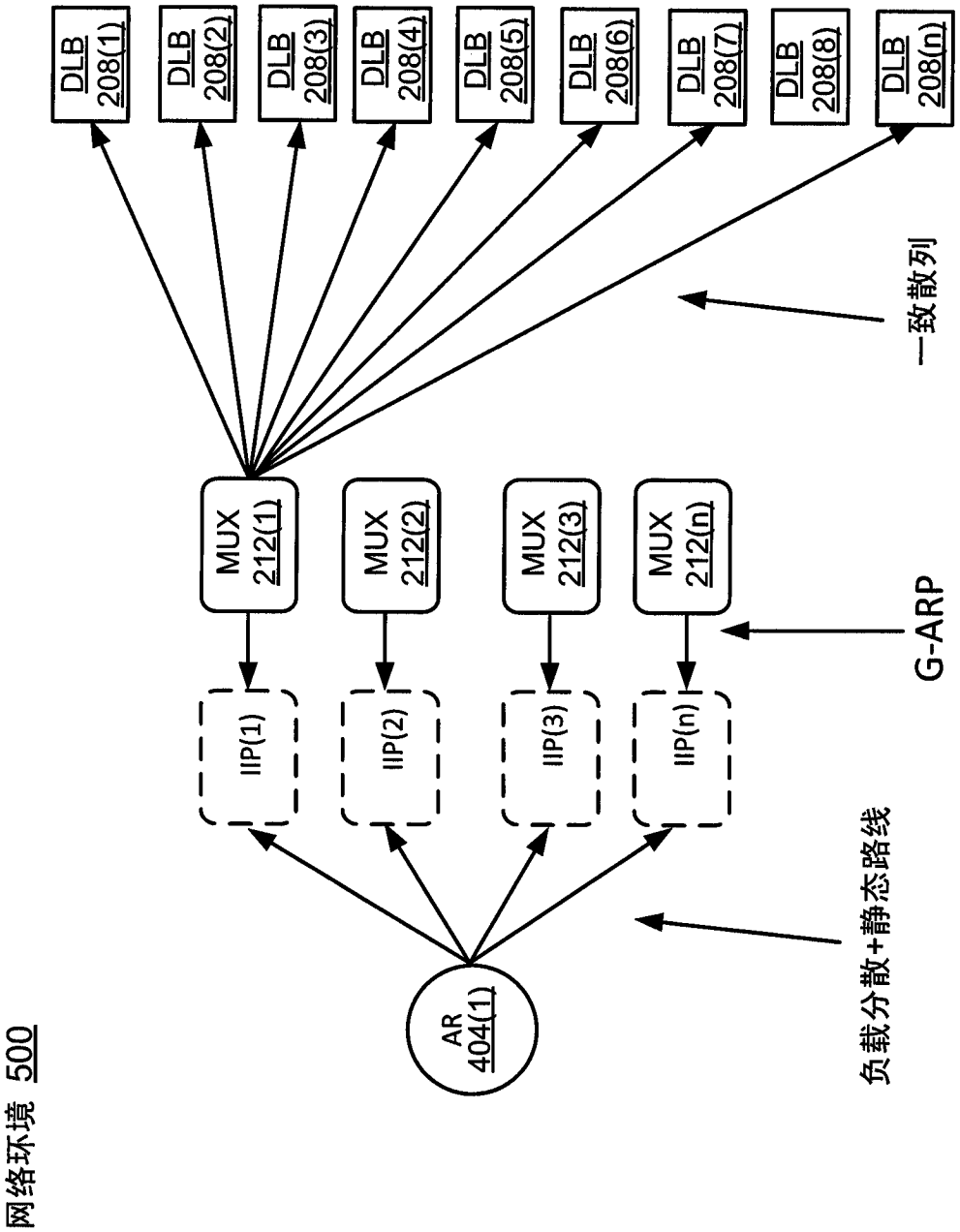

网络环境 500
DLB 208(1)
DLB 208(2)
DLB 208(3)
DLB 208(4)
DLB 208(5)
DLB 208(6)
DLB 208(7)
DLB 208(8)
DLB 208(n)
一致散列
MUX 212(1)
MUX 212(2)
MUX 212(3)
MUX 212(n)
G-ARP
IIP(1)
IIP(2)
IIP(3)
IIP(n)
AR 404(1)
负载分散+静态路线

图 5

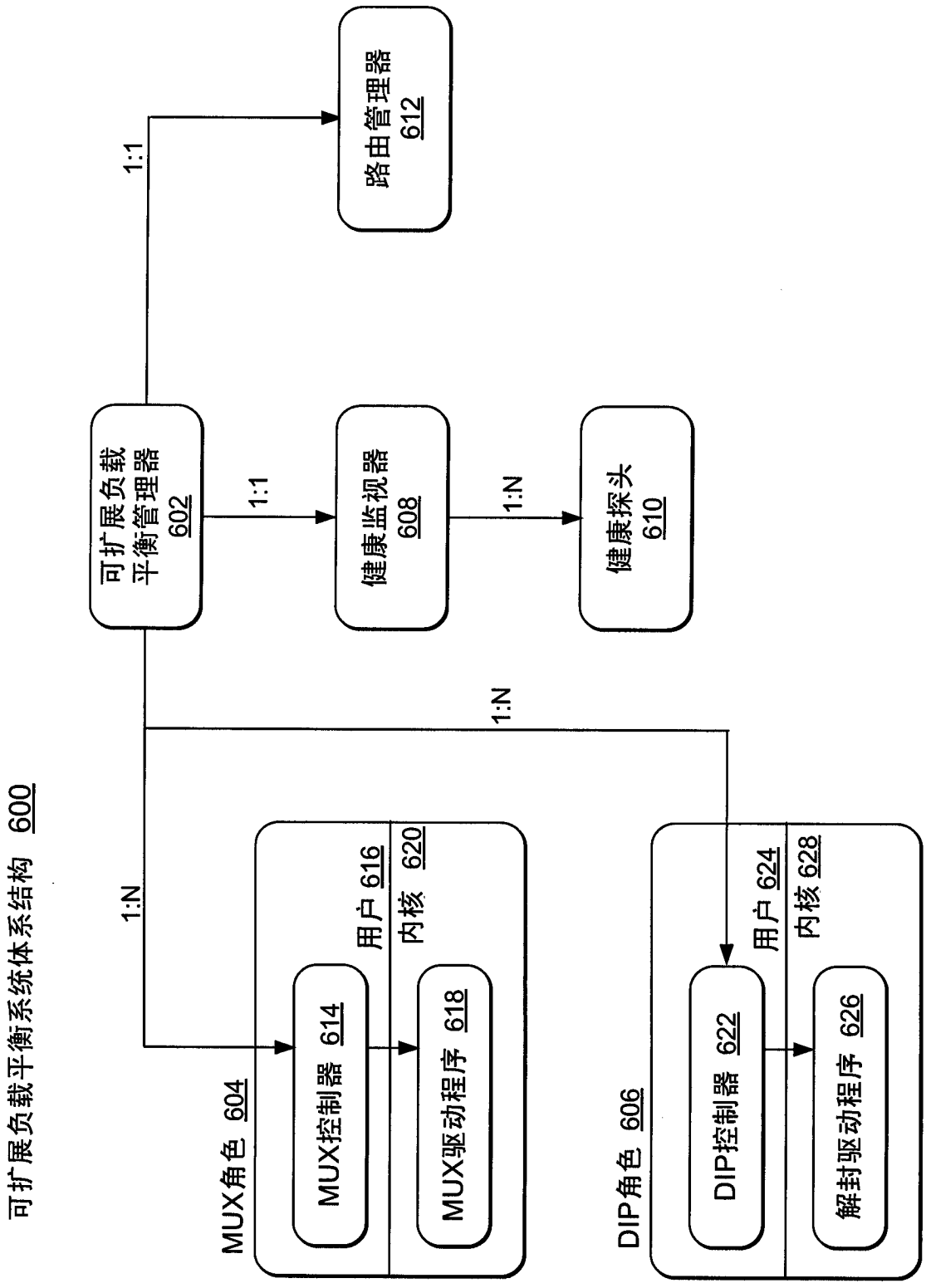

可扩展负载平衡系统体系结构 600
可扩展负载
平衡管理器
602
1:1
路由管理器
612
1:1
健康监视器
608
1:N
健康探头
610
1:N
1:N
MUX角色 604
用户 616
内核 620
MUX控制器 614
MUX驱动程序 618
DIP角色 606
用户 624
内核 628
DIP控制器 622
解封驱动程序 626

图 6

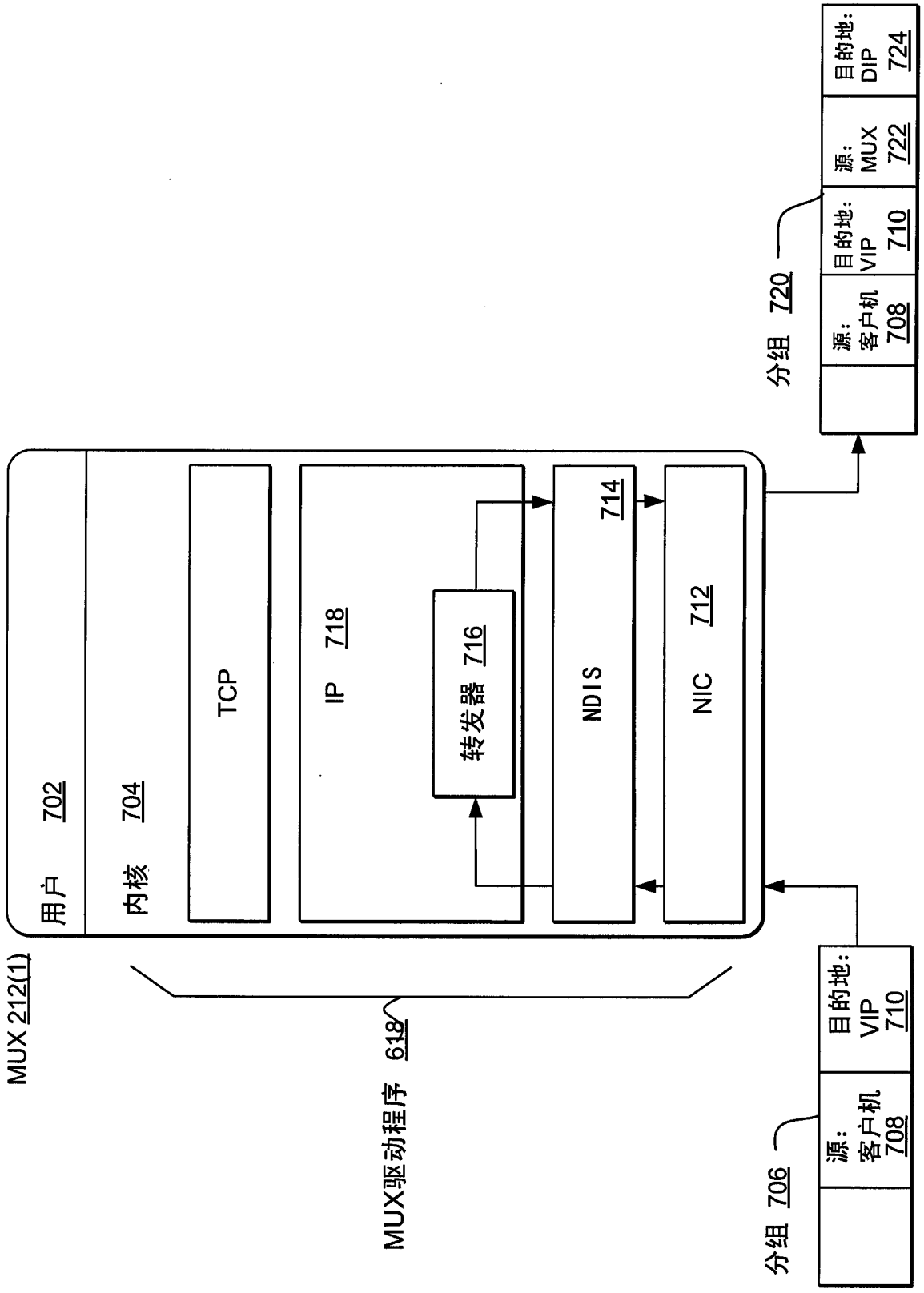

MUX 212(1)
用户 702
内核 704
TCP
IP 718
转发器 716
NDIS 714
NIC 712
MUX驱动程序 618
分组 706
源：客户机 708
目的地：VIP 710
分组 720
源：客户机 708
目的地：VIP 710
源：MUX 722
目的地：DIP 724

图 7

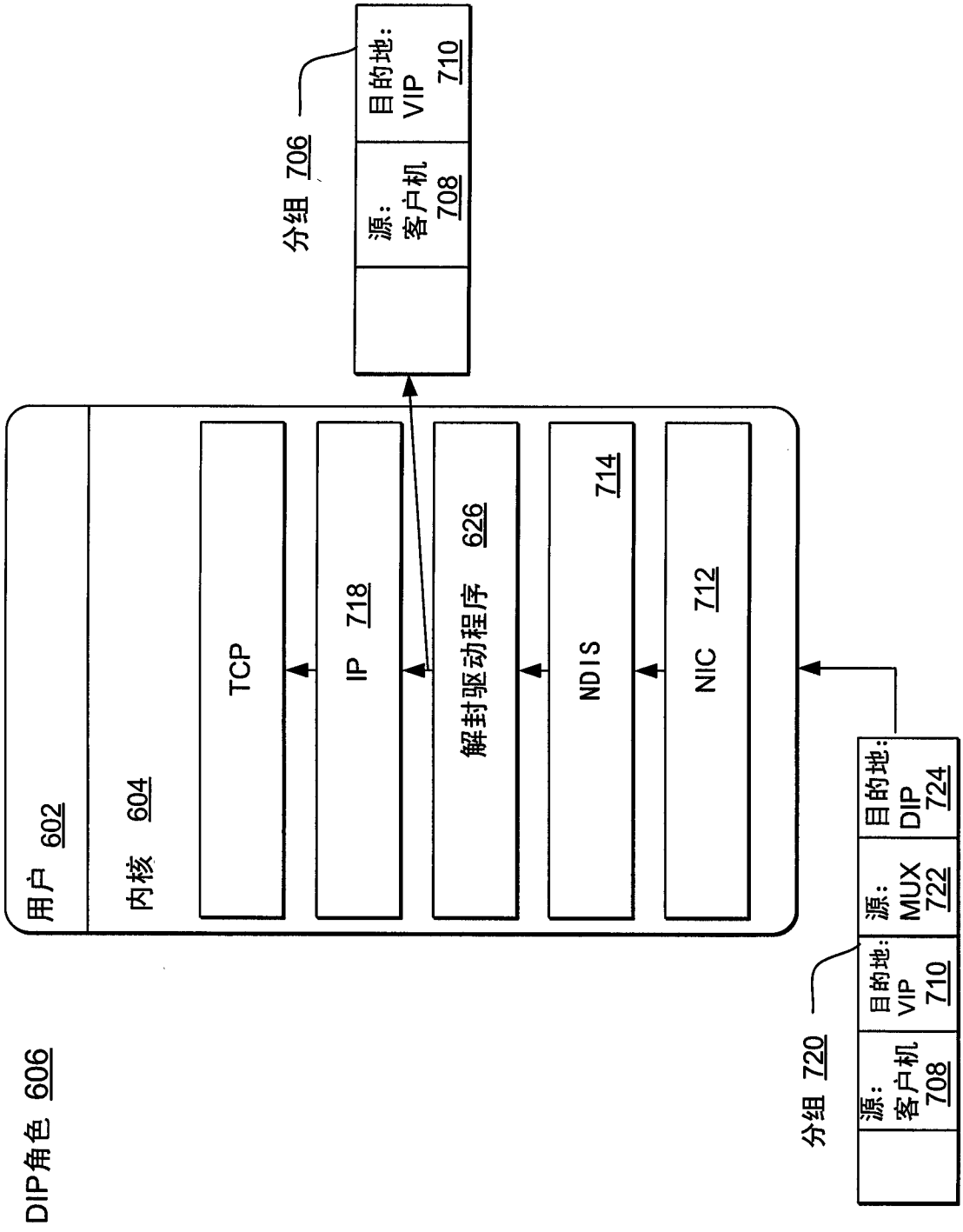

DIP角色 606
用户 602
内核 604
TCP
IP 718
解封驱动程序 626
NDIS 714
NIC 712
分组 706
源：客户机 708
目的地：VIP 710
分组 720
源：客户机 708
目的地：VIP 710
源：MUX 722
目的地：DIP 724

图 8

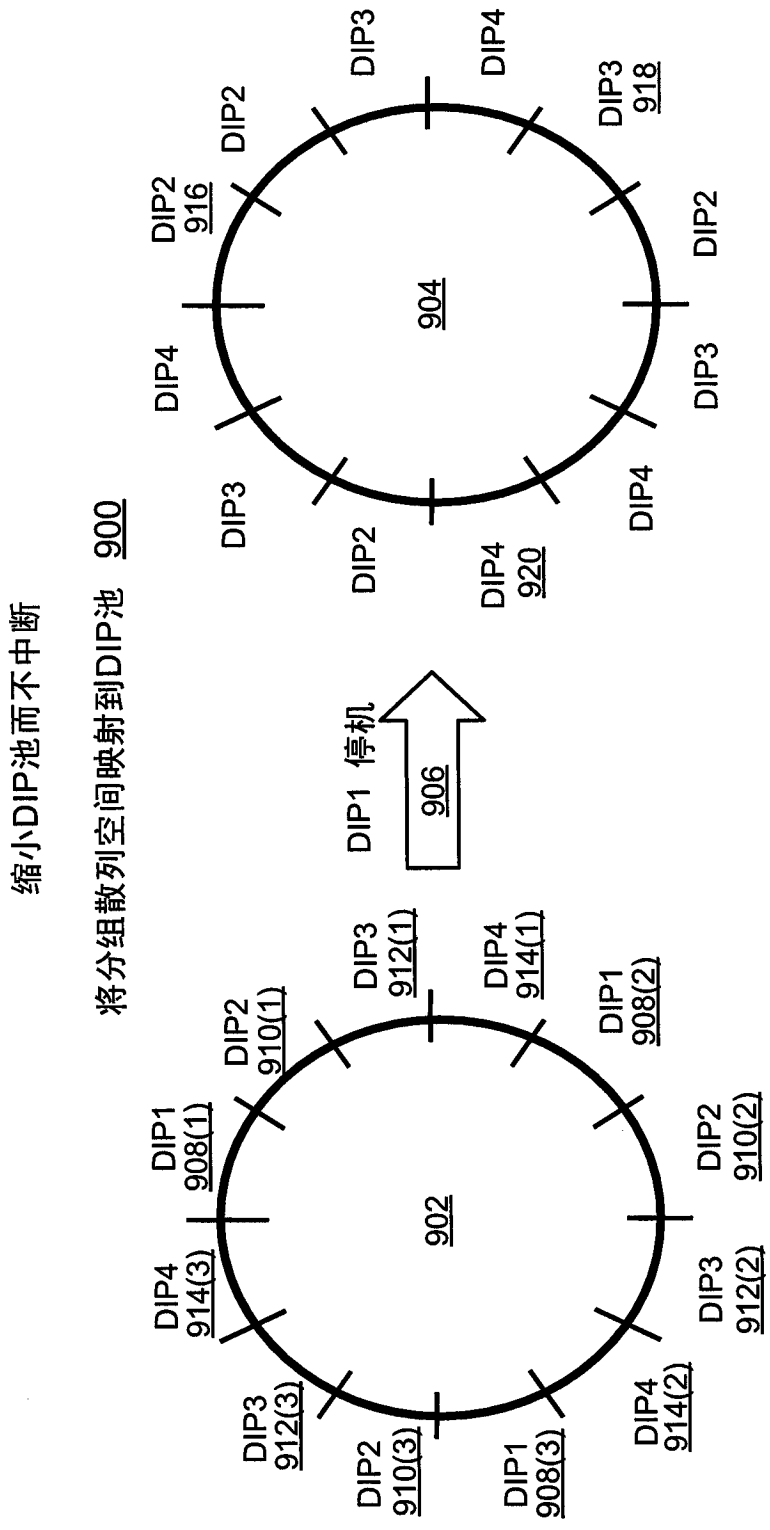

缩小DIP池而不中断
将分组散列空间映射到DIP池 900
DIP1 停机
906
902
904
DIP1 908(1)
DIP2 910(1)
DIP3 912(1)
DIP4 914(1)
DIP1 908(2)
DIP2 910(2)
DIP3 912(2)
DIP4 914(2)
DIP1 908(3)
DIP2 910(3)
DIP3 912(3)
DIP4 914(3)
DIP2 916
DIP3 918
DIP4 920

图 9

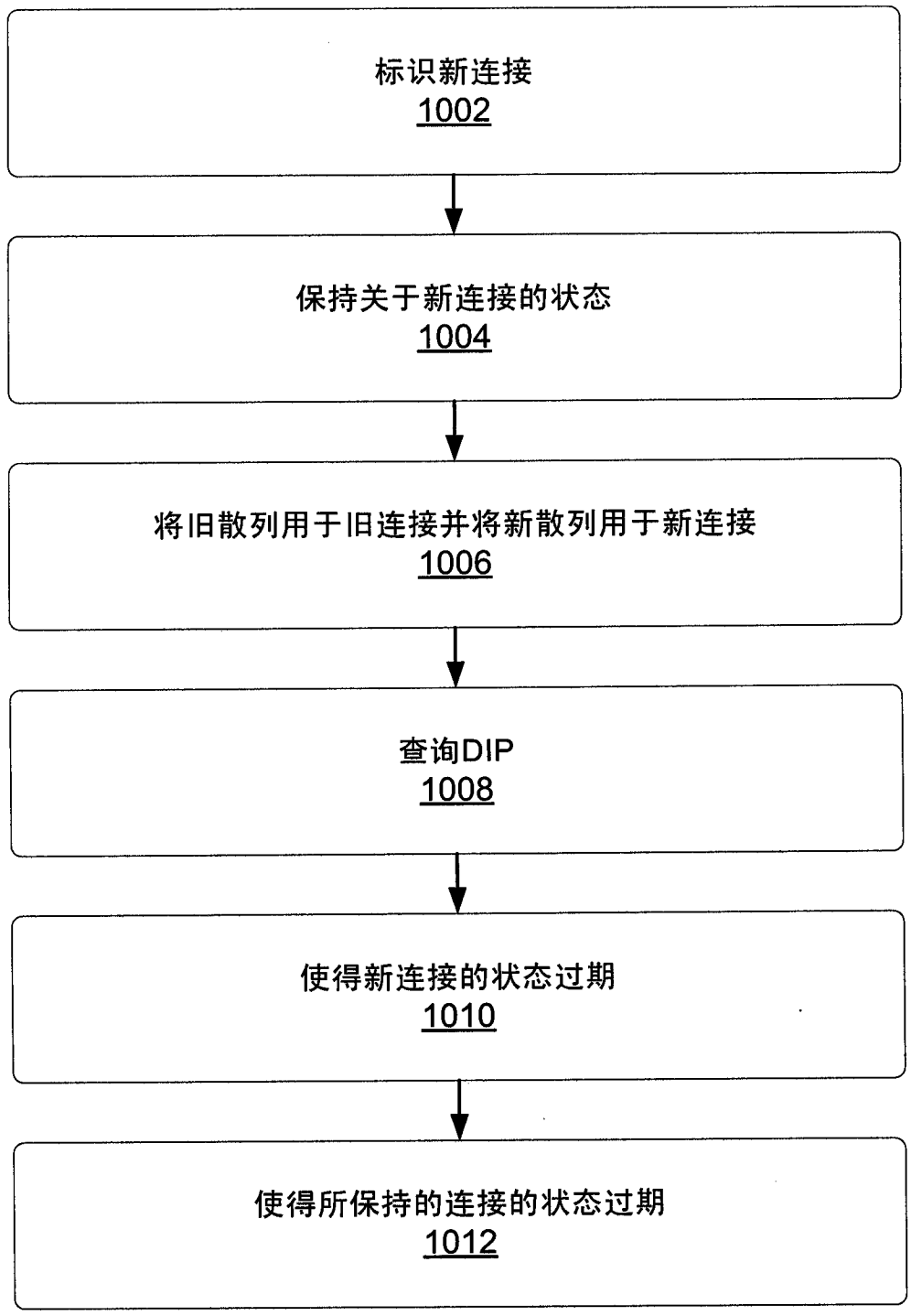

方法 1000
标识新连接
1002
保持关于新连接的状态
1004
将旧散列用于旧连接并将新散列用于新连接
1006
查询DIP
1008
使得新连接的状态过期
1010
使得所保持的连接的状态过期
1012

图 10

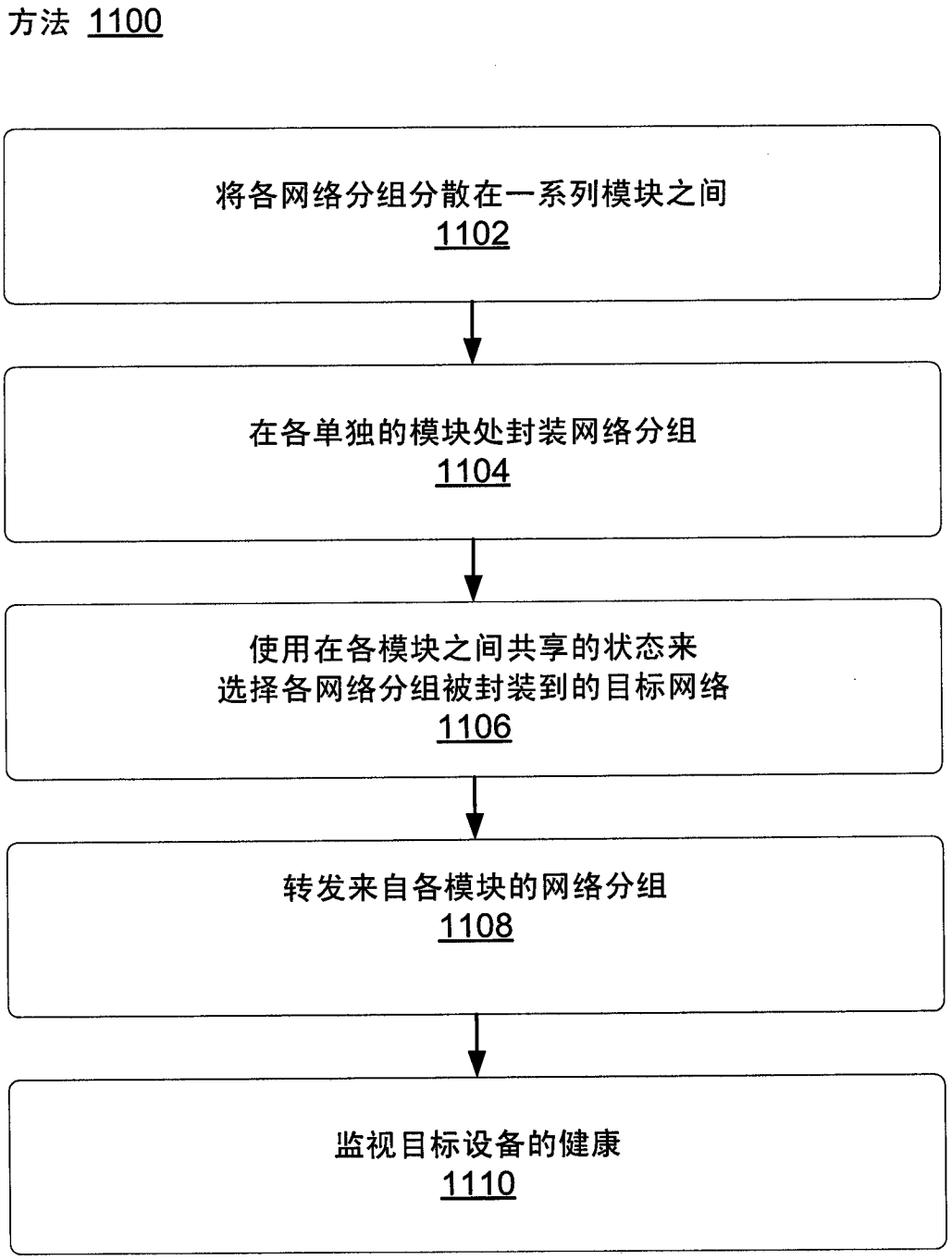

方法 1100
将各网络分组分散在一系列模块之间
1102
在各单独的模块处封装网络分组
1104
使用在各模块之间共享的状态来
选择各网络分组被封装到的目标网络
1106
转发来自各模块的网络分组
1108
监视目标设备的健康
1110

图 11