



(12) 发明专利申请

(10) 申请公布号 CN 103595648 A

(43) 申请公布日 2014. 02. 19

(21) 申请号 201310357302. X

H04L 29/12(2006. 01)

(22) 申请日 2013. 08. 16

H04L 12/743(2013. 01)

(30) 优先权数据

13/588, 578 2012. 08. 17 US

(71) 申请人 国际商业机器公司

地址 美国纽约

(72) 发明人 A·毕斯瓦 J·奇丹毕

S·拉贾戈帕兰 唐刚

(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 杨国权

(51) Int. Cl.

H04L 12/803(2013. 01)

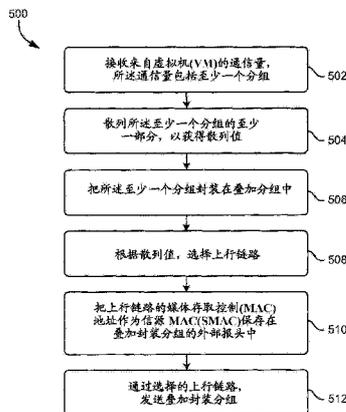
权利要求书4页 说明书13页 附图7页

(54) 发明名称

用于在服务器的接收侧进行负载均衡的方法和系统

(57) 摘要

本发明公开了一种用于在服务器的接收侧进行负载均衡的方法和系统。一种系统包括服务器,该服务器包括:适于接收来自虚拟机 (VM) 的通信量的逻辑部,该通信量包括至少一个分组,适于按照散列算法散列该至少一个分组的至少一部分以获得散列值的逻辑部,和适于根据散列值选择上行链路的逻辑部;至少一个加速网络接口卡 (NIC),每个加速 NIC 包括:包括适于与服务器和网络通信的多个快捷外设互联标准 (PCIe) 端口的网络端口,每个网络端口包括上行链路,适于把至少一个分组封装到叠加封装分组中的逻辑部,适于把与选择的上行链路对应的媒体存取控制 (MAC) 地址作为信源 MAC (SMAC) 地址保存在叠加封装分组的外部报头中的逻辑部,和适于通过选择的上行链路发送叠加封装分组的逻辑部。



1. 一种系统,包括:
 - 服务器,所述服务器包括:
 - 适合于接收来自虚拟机 VM 的通信量的逻辑部,所述通信量包括至少一个分组;
 - 适合于按照散列算法对所述至少一个分组的至少一部分进行散列以获得散列值的逻辑部;和
 - 适合于根据散列值选择上行链路的逻辑部;
 - 至少一个加速网络接口卡 NIC,每个加速 NIC 包括:
 - 多个网络端口,其包括适合于与服务器和网络通信的多个快捷外设互联标准 PCIe 端口,每个网络端口包括上行链路;
 - 适合于把所述至少一个分组封装到叠加封装分组中的逻辑部;
 - 适合于把与选择的上行链路对应的媒体存取控制 MAC 地址作为信源 MAC 地址保存在叠加封装分组的外部报头中的逻辑部,其中信源 MAC 地址即 SMAC 地址;和
 - 适合于通过选择的上行链路发送叠加封装分组的逻辑部。
2. 按照权利要求 1 所述的系统,其中选择散列算法,以对应于可用于向网络发送通信量的多个上行链路。
3. 按照权利要求 1 所述的系统,其中所述至少一个分组的报头被散列。
4. 按照权利要求 1 所述的系统,其中当所述至少一个分组被叠加封装时,所述至少一个分组的内部分组的报头被散列。
5. 按照权利要求 1 所述的系统,其中散列算法基于一个或多个参数,所述一个或多个参数包括:
 - 对应于 VM 的虚拟端口;
 - 内部分组报头 SMAC 地址;
 - 内部分组报头目的地 MAC 地址,其中目的地 MAC 地址即 DMAC 地址;
 - 内部信源网际协议 SIP 地址;和
 - 内部目的地网际协议 DIP 地址。
6. 一种用于在服务器的接收侧进行负载均衡的方法,所述方法包括:
 - 接收广播地址解析协议 ARP 请求分组;
 - 确定应当对广播 ARP 请求分组作出响应;
 - 按照散列算法对广播 ARP 请求分组的至少一部分进行散列,以获得散列值;
 - 根据所述散列值,从可用于向网络发送通信量的多个上行链路中选择上行链路;
 - 把与选择的上行链路对应的媒体存取控制 MAC 地址作为信源 MAC 地址保存在对广播 ARP 请求分组的响应中,其中信源 MAC 地址即 SMAC 地址;和
 - 在与选择的上行链路对应的 MAC 地址被保存为 SMAC 地址的情况下,把所述响应回送给广播 ARP 请求分组的信源。
7. 按照权利要求 6 所述的方法,其中选择散列算法,以对应于可用于向网络发送通信量的多个上行链路。
8. 按照权利要求 6 所述的方法,其中所述广播 ARP 请求分组的报头被散列。
9. 按照权利要求 6 所述的方法,其中当所述广播 ARP 请求分组被叠加封装时,所述广播 ARP 请求分组的内部分组的报头被散列。

10. 按照权利要求 6 所述的方法,其中散列算法基于一个或多个参数,所述一个或多个参数包括:

与为其发起广播 ARP 请求分组的虚拟机 VM 对应的虚拟端口;
内部分组报头 SMAC 地址;
内部分组报头目的地 MAC 地址,其中目的地 MAC 地址即 DMAC 地址;
内部信源网际协议 SIP 地址;和
内部目的地网际协议 DIP 地址。

11. 按照权利要求 6 所述的方法,还包括:

生成单播 ARP 请求分组,并把单播 ARP 请求分组发送给广播 ARP 请求分组的信源;
接收来自广播 ARP 请求分组的信源的响应,所述响应包含关于广播 ARP 请求分组的信源的地址信息;和

保存关于广播 ARP 请求分组的信源的地址信息,

其中当未收到来自广播 ARP 请求分组的信源的响应时,不从广播 ARP 请求分组学习地址信息。

12. 按照权利要求 6 所述的方法,其中当确定不应当对广播 ARP 请求分组作出响应时,不从广播 ARP 请求分组学习地址信息。

13. 一种用于在多个上行链路上对通信量进行负载均衡的系统,所述系统包括:

计算机可读存储介质;和

适合于执行按照权利要求 6 所述的方法的逻辑部,所述逻辑部被保存到所述计算机可读存储介质。

14. 一种用于在服务器的接收侧进行负载均衡的方法,所述方法包括:

接收单播地址解析协议 ARP 请求分组;

确定应当对单播 ARP 请求分组作出响应;

按照散列算法对单播 ARP 请求分组的至少一部分进行散列,以获得散列值;

根据所述散列值,从可用于向网络发送通信量的多个上行链路中选择上行链路;

把与选择的上行链路对应的媒体存取控制 MAC 地址作为信源 MAC 地址保存在对单播 ARP 请求分组的响应中,其中信源 MAC 地址即 SMAC 地址;和

在与选择的上行链路对应的 MAC 地址被保存为 SMAC 地址的情况下,把所述响应回送给单播 ARP 请求分组的信源。

15. 按照权利要求 14 所述的方法,其中选择散列算法,以对应于可用于向网络发送通信量的多个上行链路。

16. 按照权利要求 14 所述的方法,其中所述单播 ARP 请求分组的报头被散列。

17. 按照权利要求 14 所述的方法,其中当所述单播 ARP 请求分组被叠加封装时,所述单播 ARP 请求分组的内部分组的报头被散列。

18. 按照权利要求 14 所述的方法,其中散列算法基于一个或多个参数,所述一个或多个参数包括:

与为其发起单播 ARP 请求分组的虚拟机 VM 对应的虚拟端口;

内部分组报头 SMAC 地址;

内部分组报头目的地 MAC 地址,其中目的地 MAC 地址即 DMAC 地址;

内部信源网际协议 SIP 地址 ;和
内部目的地网际协议 DIP 地址。

19. 按照权利要求 14 所述的方法,其中当确定不应当对广播 ARP 请求分组作出响应时,不从单播 ARP 请求分组学习地址信息。

20. 一种用于在多个上行链路上对通信量进行负载均衡的系统,所述系统包括:
计算机可读存储介质 ;和

适合于执行按照权利要求 14 所述的方法的逻辑部,所述逻辑部被保存到所述计算机可读存储介质。

21. 一种用于在多个上行链路上对通信量进行负载均衡的方法,所述方法包括:

接收来自虚拟机 VM 的通信量,所述通信量包含至少一个分组;

按照散列算法,对所述至少一个分组的至少一部分进行散列,以获得散列值;

根据所述散列值,从可用于向网络发送通信量的多个上行链路中,选择上行链路;

把所述至少一个分组封装在叠加封装分组中;

把与选择的上行链路对应的媒体存取控制 MAC 地址作为信源 MAC 地址保存在叠加封装分组的外部报头中,其中信源 MAC 地址即 SMAC 地址 ;和

通过选择的上行链路,发送叠加封装分组。

22. 按照权利要求 21 所述的方法,其中选择散列算法,以对应于可用于向网络发送通信量的多个上行链路,其中所述至少一个分组的报头被散列。

23. 按照权利要求 21 所述的方法,其中散列算法基于一个或多个参数,所述一个或多个参数包括:

对应于 VM 的虚拟端口 ;

内部分组报头 SMAC 地址 ;

内部分组报头目的地 MAC 地址,其中目的地 MAC 地址即 DMAC 地址 ;

内部信源网际协议 SIP 地址 ;和

内部目的地网际协议 DIP 地址。

24. 一种用于在多个上行链路上对通信量进行负载均衡的系统,所述系统包括:

用硬件实现的处理器,所述处理器适合于执行逻辑部 ;

适合于接收地址解析协议 ARP 请求分组的逻辑部 ;

适合于按照散列算法对 ARP 请求分组的至少一部分进行散列以获得散列值的逻辑部 ;

适合于根据所述散列值从可用于向网络发送通信量的多个上行链路中选择上行链路的逻辑部 ;

适合于把与选择的上行链路对应的媒体存取控制 MAC 地址作为信源 MAC 地址保存在对 ARP 请求分组的响应中的逻辑部,其中信源 MAC 地址即 SMAC 地址 ;和

适合于在与选择的上行链路对应的 MAC 地址被保存为 SMAC 地址的情况下把所述响应回送给 ARP 请求分组的信源的逻辑部。

25. 按照权利要求 24 所述的系统,

其中当 ARP 请求分组被单播时,关于单播 ARP 请求分组的信源的地址的信息从单播 ARP 请求分组中被学习并被保存用于随后与所述信源通信,和

其中当 ARP 请求分组被广播时,所述系统还包括:

适合于生成单播 ARP 请求分组并把单播 ARP 请求分组发送给广播 ARP 请求分组的信源的逻辑部；

适合于接收来自广播 ARP 请求分组的信源的响应的逻辑部,所述响应包含关于广播 ARP 请求分组的信源的地址信息 ;和

适合于学习并保存关于广播 ARP 请求分组的信源的地址信息的逻辑部,

其中当未收到来自广播 ARP 请求分组的信源的响应时,不从广播 ARP 请求分组中学习地址信息。

用于在服务器的接收侧进行负载均衡的方法和系统

技术领域

[0001] 本发明涉及数据中心基础结构,更具体地,本发明涉及利用一组组队的网络接口卡使叠加网络通信量负载均衡。

背景技术

[0002] 网络虚拟化是新出现的数据中心和云计算趋势,其目的是以极大地简化多租户环境,以及传统环境中的网络服务开通的方式,使终端站看到的网络虚拟化。实现网络虚拟化的更常见技术之一是利用网络叠加,其中在终端站连接到的服务器、边缘网络交换机和网关之间建立隧道。隧道实际上是通过把信源终端站传送的分组封装在叠加报头(header)中实现的,所述叠加报头经基于网际协议(IP)的网络,利用用户数据报(UDP)传输,把分组从源交换机传送给目标交换机。叠加报头包括唯一地识别虚拟网络的标识符(ID)。目标交换机(隧道终点)剥离叠加报头封装、UDP传输报头和IP报头,经常规网络连接把原始分组传递给目的地终端站。除了这种隧道机制之外,边缘交换机参与地址发现协议,所述地址发现协议可以是基于学习/洪泛的,或者基于查寻的。

[0003] 叠加网络可以利用任何可用的虚拟化协议(比如虚拟可扩展局域网(VXLAN)、位置标识/身份标识分离协议(LISP)、叠加传输虚拟化(OTV)、利用通用路由封装的网络虚拟(NVGRE)等),以便利用称为隧道的结构连接地理上分离的层2(L2)网络。这些隧道是能够把分组打包到叠加分组中以跨网络传送的层3(L3)上的L2传输隧道。由一个虚拟网络中的虚拟机(VM)发起的去往在另一个物理位置的相同虚拟网络中的另一个VM或另一组VM的L2分组通过L3隧道承载。

[0004] 通过对进出每个VM的网络分组的处理进行控制的虚拟化平台,利用任意数量的VM,可实现叠加网络。一个或多个VM可以与叠加网络关联。通过把预先规定的性质和策略用于对应的叠加网络和/或VM,虚拟化平台处理与每个叠加网络和/或VM相关的网络通信量。当叠加网络的数目增大时,虚拟化平台的处理负载需求也增大。虚拟化平台的处理负载需求包括端接隧道的数目、虚拟隧道终点(VTEP)的管理、每个隧道的地址学习、每个隧道的分组封装和解封、等等。于是,对于网络的给定性能水平来说,叠加网络或VM的数目受虚拟化平台的处理能力限制。于是,需要在不增大虚拟化平台使用的可用物理资源的情况下,提高虚拟化平台的性能,以便继续扩展叠加网络的使用。

[0005] 另外,为了向服务器以及具体向服务器托管的VM提供负载均衡,可以通过静态地把VM绑定到特定上行链路(称为与交换机无关的组队,因为在上游的交换机不需要任何配置),或者通过在交换机和服务器都配置静态或动态的端口通道,并进行依赖于某个因素(比如分组报头中的各个字段的散列)的负载均衡,使服务器侧网络接口卡(NIC)组队,以使网络通信量负载均衡。配置端口通道被称为与交换机有关的组队,因为它需要在上游的交换机进行配置。对叠加通信量来说,这种负载均衡方法失效,因为来自一个或个VM的原始分组都被封装在隧道中的叠加报头中,同时外部报头包括虚拟化平台内核的媒体存取控制(MAC)地址和网际协议(IP)地址。网络中的多个组件看不见内部分组,从而使得传统的负

载均衡无效。因而,有益的是能够使叠加封装的网络通信量负载均衡。

发明内容

[0006] 在一个实施例中,一种系统包括服务器,所述服务器包括:适合于接收来自虚拟机 (VM) 的通信量的逻辑部,所述通信量包括至少一个分组,适合于按照散列算法对所述至少一个分组的至少一部分进行散列以获得散列值的逻辑部,和适合于根据散列值选择上行链路的逻辑部;至少一个加速网络接口卡 (NIC),每个加速 NIC 包括:包括适合于与服务器和网络通信的多个快捷外设互联标准 (PCIe) 端口的多个网络端口,每个网络端口包括上行链路,适合于把所述至少一个分组封装到叠加封装分组中的逻辑部,适合于把与选择的上行链路对应的媒体存取控制 (MAC) 地址作为信源 MAC (SMAC) 地址保存在叠加封装分组的外部报头中的逻辑部,和适合于通过选择的上行链路发送叠加封装分组的逻辑部。

[0007] 在另一个实施例中,一种用于在服务器的接收侧进行负载均衡的方法包括接收广播地址解析协议 (ARP) 请求分组,确定应当对广播 ARP 请求分组作出响应,按照散列算法对广播 ARP 请求分组的至少一部分进行散列,以获得散列值,根据所述散列值,从可用于向网络发送通信量的多个上行链路中选择上行链路,把与选择的上行链路对应的 MAC 地址作为 SMAC 地址保存在对广播 ARP 请求分组的响应中,和在与选择的上行链路对应的 MAC 地址被保存为 SMAC 地址的情况下,把所述响应回送给广播 ARP 请求分组的信源。

[0008] 在另一个实施例中,一种用于在服务器的接收侧进行负载均衡的方法包括接收单播 ARP 请求分组,确定应当对单播 ARP 请求分组作出响应,按照散列算法对单播 ARP 请求分组的至少一部分进行散列,以获得散列值,根据所述散列值,从可用于向网络发送通信量的多个上行链路中选择上行链路,把与选择的上行链路对应的 MAC 地址作为 SMAC 地址,保存在对单播 ARP 请求分组的响应中,和在与选择的上行链路对应的 MAC 地址被保存为 SMAC 地址的情况下,把所述响应回送给单播 ARP 请求分组的信源。

[0009] 按照另一个实施例,一种用于在多个上行链路上对具有至少一个分组的通信量进行负载均衡的方法包括接收来自 VM 的通信量,按照散列算法,对所述至少一个分组的至少一部分进行散列,以获得散列值,根据所述散列值,从可用于向网络发送通信量的多个上行链路中,选择上行链路,把所述至少一个分组封装在叠加封装分组中,把与选择的上行链路对应的 MAC 地址作为 SMAC 地址,保存在叠加封装分组的外部报头中,和通过选择的上行链路,发送叠加封装分组。

[0010] 在另一个实施例中,一种用于在多个上行链路上对通信量进行负载均衡的系统包括用硬件实现的处理器,所述处理器适合于执行逻辑部,适合于接收 ARP 请求分组的逻辑部,适合于按照散列算法,对 ARP 请求分组的至少一部分进行散列以获得散列值的逻辑部,适合于根据所述散列值从可用于向网络发送通信量的多个上行链路中选择上行链路的逻辑部,适合于把与选择的上行链路对应的 MAC 地址作为 SMAC 地址保存在对 ARP 请求分组的响应中的逻辑部,和适合于在与选择的上行链路对应的 MAC 地址被保存为 SMAC 地址的情况下把所述响应回送给 ARP 请求分组的信源的逻辑部。

[0011] 根据结合附图举例说明本发明的原理的以下详细说明,本发明的其它方面和实施例将变得明显。

附图说明

- [0012] 图 1 按照一个实施例说明网络体系结构。
- [0013] 图 2 按照一个实施例示出可与图 1 的服务器和 / 或客户端相关的代表性硬件环境。
- [0014] 图 3 是按照一个实施例的虚拟化数据中心的简化图。
- [0015] 图 4 按照一个实施例示出支持叠加的服务器。
- [0016] 图 5 按照一个实施例示出在服务器的发射侧进行负载均衡的方法的流程图。
- [0017] 图 6 按照一个实施例示出在服务器的接收侧进行负载均衡的方法的流程图。
- [0018] 图 7 按照一个实施例示出在服务器的接收侧进行负载均衡的方法的流程图。

具体实施方式

[0019] 下面说明用于举例说明本发明的一般原理,并不意图限制这里要求保护的发明概念。此外,这里说明的特定特征可以与其它说明的特征组合地用在各种可能的组合和置换的每一个中。

[0020] 除非这里明确地另有说明,否则所有术语应被给予其最宽广的可能解释,包括从说明书暗指的含义,以及本领域的技术人员理解的和 / 或在词典、论文等中定义的含义。

[0021] 另外必须注意,在说明书和附加的权利要求书中使用的单数形式的“一”、“一个”和“该”包括复数形式,除非另有说明。

[0022] 按照一个实施例,提供用于运送虚拟叠加网络通信量的网络接口卡 (NIC) 的与交换机无关的组队。在一个途径中,根据叠加封装分组的内部分组报头中的字段的散列,选择上行链路。一旦选择了上行链路,就用选择的上行链路的 MAC 地址替换分组的外部信源媒体存取控制 (MAC) 地址。这提供上游通信量的有效负载均衡,而与组队模式 (与交换机有关或者无关的组队) 无关。为了使在服务器接收的通信量负载均衡,可以利用对地址解析协议 (ARP) 分组的接收或传输采取的动作,如这里更详细所述。

[0023] 在一个实施例中,系统包括服务器,所述服务器包括:适合于接收来自虚拟机 (VM) 的通信量,所述通信量包括至少一个分组,适合于按照散列算法散列所述至少一个分组的至少一部分以获得散列值的逻辑部,和适合于根据散列值选择上行链路的逻辑部;至少一个加速 NIC,每个加速 NIC 包括:包括适合于与服务器和网络通信的多个快捷外设互联标准 (PCIe) 端口的多个网络端口,每个网络端口包括一个上行链路,适合于把所述至少一个分组封装到叠加封装分组中的逻辑部,适合于把与选择的上行链路对应的媒体存取控制 (MAC) 地址作为信源 MAC (SMAC) 地址保存在叠加封装分组的外部报头中的逻辑部,和适合于通过选择的上行链路发送叠加封装分组的逻辑部。

[0024] 在另一个实施例中,在服务器的接收侧使负载均衡的方法包括接收广播 ARP 请求分组,确定应当对广播 ARP 请求分组作出响应,按照散列算法散列广播 ARP 请求分组的至少一部分,以获得散列值,根据所述散列值,从可用于向网络发送通信量的多个上行链路中选择上行链路,把与选择的上行链路对应的 MAC 地址作为 SMAC 地址保存在对广播 ARP 请求分组的响应中,和在与选择的上行链路对应的 MAC 地址被保存为 SMAC 地址的情况下,把所述响应回送给广播 ARP 请求分组的信源。

[0025] 在另一个实施例中,在服务器的接收侧使负载均衡的方法包括接收单播 ARP 请求

分组,确定应当对单播 ARP 请求分组作出响应,按照散列算法散列单播 ARP 请求分组的至少一部分,以获得散列值,根据所述散列值,从可用于向网络发送通信量的多个上行链路中选择上行链路,把与选择的上行链路对应的 MAC 地址作为 SMAC 地址保存在对单播 ARP 请求分组的响应中,和在与选择的上行链路对应的 MAC 地址被保存为 SMAC 地址的情况下,把所述响应回送给单播 ARP 请求分组的信源。

[0026] 按照另一个实施例,在多个上行链路上使具有至少一个分组的通信量负载均衡的方法包括接收来自 VM 的通信量,按照散列算法,散列所述至少一个分组的至少一部分,以获得散列值,根据所述散列值,从可用于向网络发送通信量的多个上行链路中选择上行链路,把所述至少一个分组封装在叠加封装分组中,把与选择的上行链路对应的 MAC 地址作为 SMAC 地址保存在叠加封装分组的外部报头中,和通过选择的上行链路发送叠加封装分组。

[0027] 在另一个实施例中,在多个上行链路上使通信量负载均衡的系统包括用硬件实现的处理器,所述处理器适合于执行逻辑部,适合于接收 ARP 请求分组的逻辑部,适合于按照散列算法散列 ARP 请求分组的至少一部分以获得散列值的逻辑部,适合于根据所述散列值从可用于向网络发送通信量的多个上行链路中选择上行链路的逻辑部,适合于把与选择的上行链路对应的 MAC 地址作为 SMAC 地址保存在对 ARP 请求分组的响应中的逻辑部,和适合于在与选择的上行链路对应的 MAC 地址被保存为 SMAC 地址的情况下把所述响应回送给 ARP 请求分组的信源的逻辑部。

[0028] 所属技术领域的技术人员知道,本发明的各个方面可以实现为系统、方法或计算机程序产品。因此,本发明的各个方面可以具体实现为以下形式,即:完全的硬件实施方式、完全的软件实施方式(包括固件、驻留软件、微代码等),或硬件和软件方面结合的实施方式,这里可以统称为“逻辑部”、“电路”、“模块”或“系统”。此外,在一些实施例中,本发明的各个方面还可以实现为在一个或多个计算机可读介质中的计算机程序产品的形式,该计算机可读介质中包含计算机可读的程序代码。

[0029] 可以采用一个或多个计算机可读介质的任意组合。计算机可读介质可以是计算机可读信号介质或者非暂态计算机可读存储介质。非暂态计算机可读存储介质例如可以是一但不限于一电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者以上的任意适当组合。非暂态计算机可读存储介质的更具体的例子(非穷举列表)包括:便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM 或闪存)、便携式紧凑盘只读存储器(CD-ROM)、蓝光光盘只读存储器(BD-ROM)、光存储器件、磁存储器件、或者上述的任意适当组合。在本文的上下文中,非暂态计算机可读存储介质可以是任何能够包含或存储程序或应用软件的有形介质,该程序或应用软件可供指令执行系统、装置或者器件使用或者与其结合使用。

[0030] 计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括——但不限于——电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是非暂态计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序,比如具有一条或多条导线的电连接、光纤等等。

[0031] 计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括——但不限于——无线、有线、光缆、射频 (RF) 等等,或者上述的任意合适的组合。

[0032] 可以用一种或多种程序设计语言的任意组合编写用于执行本发明操作的计算机程序代码,所述程序设计语言包括面向对象的设计语言——诸如 Java、Smalltalk、C++ 等,还包括常规的过程式程序设计语言——诸如“C”语言或类似的设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机或服务器可以通过任意种类的网络——包括局域网 (LAN)、存储区域网 (SAN) 和 / 或广域网 (WAN)、任何虚拟网络——连接到用户计算机,或者,可以连接到外部计算机,例如利用因特网服务提供商 (ISP) 通过因特网连接。

[0033] 下面将参考按照本发明的各个实施例的方法、设备 (系统) 和计算机程序产品的流程图和 / 或框图描述本发明。应当理解,流程图和 / 或框图的每个方框以及流程图和 / 或框图中各方框的组合,都可以由计算机程序指令实现。这些计算机程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器,从而生产出一种机器,使得这些计算机程序指令在通过计算机或其它可编程数据处理装置的处理器执行时,产生了实现流程图和 / 或框图中的一个或多个方框中规定的功能 / 动作的装置。

[0034] 也可以把这些计算机程序指令存储在计算机可读介质中,这些指令使得计算机、其它可编程数据处理装置、或其他设备以特定方式工作,从而,存储在计算机可读介质中的指令就产生出包括实现流程图和 / 或框图中的一个或多个方框中规定的功能 / 动作的指令的制品 (article of manufacture)。

[0035] 计算机程序指令也可被加载到计算机、其它可编程处理设备或其它装置上,使得在所述计算机、其它可编程设备或其它装置上执行一系列的操作步骤,从而产生计算机实现的实现,以致在所述计算机或其它可编程设备上执行的指令提供实现在流程图和 / 或框图的一个或多个方框中指定的功能 / 步骤的实现。

[0036] 图 1 按照一个实施例图解说明网络体系结构 100。如图 1 中所示,提供多个远程网络 102,包括第一远程网络 104 和第二远程网络 106。网关 101 可耦接在远程网络 102 和邻近网络 108 之间。在本网络体系结构 100 的环境中,网络 104、106 都可采取任何形式,包括 (但不限于) LAN、VLAN、诸如因特网之类的 WAN、公共交换电话网 (PSTN)、内部电话网等等。

[0037] 使用中,网关 101 充当从远程网络 102 到邻近网络 108 的入口点。因而,网关 101 可起能够指引到达网络 101 的数据的特定分组的路由器作用,和为特定分组提供进出网关 101 的实际路径的交换机作用。

[0038] 另外还包括与邻近网络 108 耦接的至少一个数据服务器 114,通过网关 101,可从远程网络 102 访问所述数据服务器 114。应注意数据服务器 114 可包括任何种类的计算机设备 / 群件。与每个数据服务器 114 耦接的是多个用户设备 116。这样的用户设备 116 可包括桌上型计算机、膝上型计算机、手持式计算机、打印机和 / 或任何其它种类的包含逻辑的设备。应注意在一些实施例中,用户设备 111 也可直接耦接到网络任意之一。

[0039] 外设 120 或者一系列的外设 120 (例如,传真机、打印机、扫描仪、硬盘驱动器、联网的和 / 或本地的存储单元或系统等) 可以耦接到网络 104、106、108 中的一个或多个。应注意数据库和 / 或另外的组件可以与耦接到网络 104、106、108 的任意种类的网络元件一起使

用,或者集成到所述任意种类的网络元件中。在本说明的上下文中,网络元件指的是网络的任意组件。

[0040] 按照一些方式,这里说明的方法和系统可以用和 / 或在虚拟系统和 / 或仿真一个或多个其它系统的系统上实现,比如仿真 IBM z/OS 环境的 UNIX 系统,虚拟托管 MICROSOFT WINDOWS 环境的 UNIX 系统,仿真 IBM z/OS 环境的 MICROSOFT WINDOWS 系统,等等。在一些实施例中,通过利用 VMWARE 软件,可以增强这种虚拟化和 / 或仿真。

[0041] 在更多的途径中,一个或多个 104、106、108 可代表统称为“云”的一群系统。在云计算中,以按需的关系把诸如处理能力、外设、软件、数据、服务器之类的共享资源提供给云中的任意系统,从而允许跨越多个计算系统的服务的访问和分布。云计算一般涉及在云中工作的系统之间的因特网连接,不过本领域中已知,也可使用连接各个系统的其它技术。

[0042] 图 2 按照一个实施例示出了与图 1 的用户设备 116 和 / 或服务器 114 相关的代表性硬件环境。图 2 图解说明工作站的典型硬件结构,所述工作站具有中央处理器 (CPU) 210 (比如微处理器)和通过一个或多个总线 212 互连的多个其它单元,按照几个实施例,所述总线 212 可以是不同种类的总线,比如本地总线、并行总线、串行总线等。

[0043] 图 2 中所示的工作站包括随机存取存储器 (RAM) 214、只读存储器 (ROM) 216、用于连接诸如磁盘存储单元 220 之类的外设和一个或多个总线 212 的 I/O 适配器 218、用于把键盘 224、鼠标 226、扬声器 228、麦克风 232、和 / 或诸如触摸屏、数字照相机 (未示出) 之类的其它用户接口设备连接到一个或多个总线 212 的用户接口适配器 222、用于把工作站连接到通信网络 235 (例如,数据处理网络) 的通信适配器 234、和用于把一个或多个总线 212 连接到显示设备 238 的显示适配器 236。

[0044] 工作站上可驻留有诸如 MICROSOFT WINDOWS 操作系统 (OS)、MAC OS、UNIX OS 等之类的操作系统。应理解也可在作上述以外的平台和操作系统上实现优选实施例。可以利用 JAVA、XML、C 和 / 或 C++ 语言、或者其它程序语言、以及面向对象的编程方法,编写优选实施例。可以使用越来越多地用于开发复杂应用程序的面向对象的程序设计 (OOP)。

[0045] 现在参见图 3,图中按照一个实施例示出了叠加网络 300 的概念图。为了使网络服务虚拟化,而不是仅仅提供设备之间的架构路径 (连通性),当分组通过网关 314 时,可在分组上呈递服务,网关 314 为在非虚拟网络 312 与虚拟网络 A304 和虚拟网络 B306 之间移动的分组提供路由和转发。所述一个或多个虚拟网络 304、306 存在于物理 (真实) 网络基础结构 302 内。本领域的技术人员已知,网络基础结构 302 可包括一般与网络基础结构相关和 / 或用在网络基础结构中的任何组件、硬件、软件和 / 或功能性,包括 (但不限于) 交换机、连接器、导线、电路、电缆、服务器、主机、存储介质、操作系统、应用程序、端口、I/O, 等等。网络基础结构 302 支持至少一个非虚拟网络 312,所述非虚拟网络 312 可以是传统网络。

[0046] 每个虚拟网络 304、306 可以利用任意数目的 VM308、310。在一个实施例中,虚拟网络 A304 包括一个或多个 VM308,而虚拟网络 B306 包括一个或多个 VM310。如图 3 中所示,VM308、310 不被虚拟网络 304、306 共享,相反在任何特定时间,都只包含在仅仅一个虚拟网络 304、306 中。

[0047] 按照一个实施例,叠加网络 300 可包括与一个或多个分布式线卡 (DLC) 互连的一个或多个信元交换域可扩展架构组件 (SFC)。

[0048] 叠加网络 300 的组件一般根据称为 VNI 或 VNID 的虚拟网络标识符, 识别把分组路由到何处。虚拟网络标识符一般是排除 0x0 或 0xFFFFFFFF 之外的 24 位代码或数字。叠加网络 300 具有通过把 L2 分组封装到叠加报头中在层 3 (L3) 网络上隧道传递层 2 (L2) 分组的能力。这可利用虚拟可扩展局域网 (VXLAN) 或者某种其它支持叠加的协议 (比如位置标识 / 身份标识分离协议 (LISP)、叠加传输虚拟化 (OTV)、利用通用路由封装的网络虚拟化 (NVGRE) 等等) 来完成。

[0049] 分组也可被封装在数据报协议 (UDP) 和网际协议 (IP) UDP/IP 报头中。叠加网络 300 可包括一个或多个点对点隧道和 / 或点对多点隧道。另外, 可根据多个因素, 比如新设备被增加到叠加网络 300 中, 从叠加网络 300 中除去设备, 任意终端设备 (即, 管理隧道端点的设备, 比如虚拟叠加网络网关、管理程序、支持叠加功能的交换机等) 的启动, 创建、消除、变更和修改这些隧道任意之一。

[0050] 为了使设备可以管理隧道, 需要存在原始分组的信源地址、目的地地址和隧道标识符之间的映射。这样, 物理服务器能够把封装的原始分组转发给适当的目的地设备。

[0051] 参见图 4, 按照一个实施例示出了系统 400。系统 400 包括具有虚拟化平台 404 和与不同设备面接的一个或多个 NIC406 的服务器 402。在该例证实施例中, NIC1 被表示成耦接到交换机 408, NIC2 被表示成耦接到网络 410, 而 NIC3 被表示成耦接到支持叠加的网络网关 (GW) 412。当然, 其它设备、网络或系统可通过用来面接这样的设备或系统的 NIC406, 连接到服务器 402。

[0052] 本领域的技术人员明白, 网络 410 可包括交换机、路由器、电缆、连接器、网络设备等等。另外, 虚拟化平台 404 可提供和管理适合于向发送给 VM414 和 / 或从 VM414 接收的分组提供交换功能的虚拟交换机 (vSwitch) 416。

[0053] 每个 NIC406 能够提供服务器 402 和某些其它设备或系统之间的接口, 以及管理它们之间的通信。另外, 每个 NIC406 可包括适合于与网络 410、服务器 402 和任何其它连接的一个或多个设备通信的一个或多个连网端口。一般地, 当分组被封装在叠加报头中时, 与期望的目的地对应的 NIC406 只是把分组传给在叠加封装的分组的的外部报头中指定的目的地。

[0054] 每个 NIC406 和 / 或虚拟化平台 404 可以利用一个或多个处理器。可以使用任意种类的处理器, 比如现场可编程门阵列 (FPGA)、微处理器、微控制器、中央处理器 (CPU)、专用集成电路 (ASIC), 等等。

[0055] 虚拟化平台 404 还可为任意数目的 VM414 提供支持, 所述 VM414 可被布置在一个或多个虚拟网络中 (每个虚拟网络具有不同的虚拟网络标识符 (VNID))。图 4 中的 VM414 的数目和布置并不意图对可能的各种结构的限制, 因为附图只表示 VM414 和网络的例证布置。

[0056] 在一些实施例中, 为了向服务器 402 提供叠加功能, 虚拟化平台 404 还可与多个离散的软件引擎交互作用, 比如隧道管理器、ARP 和转发信息库 (FIB) 管理器、提供网际协议组件多播 (IPMC) 支持的引擎、分组封装和解封引擎、和本领域已知的任何其它叠加增强软件引擎。

[0057] 一般地, NIC406 看不见任何叠加封装分组的内部分组, 代替地, NIC406 把分组传送给虚拟化平台 404, 以便提供叠加功能。不过, 在另一个实施例中, NIC406 可被修改, 以提

供叠加功能,这里称为加速 NIC。

[0058] 在另一个实施例中,为了在虚拟化网络和非虚拟化网络之间进行桥接,待通过虚拟叠加网络传递的分组可被传递给支持叠加的网络网关 412,以便在服务器 402 位于的虚拟网络之外进一步转发和 / 或路由。

[0059] 现在参见图 5,图中按照一个实施例示出了在服务器的发送侧使负载均衡的方法 500。在各个实施例中,可以在图 1-4 中描述的环境等中的任意环境中,按照本发明进行方法 500。当然,当阅读本说明时,本领域的技术人员理解,方法 500 中可以包括比在图 5 中具体说明的那些操作更多或更少的操作。

[0060] 方法 500 的各个步骤可由工作环境的任何适当组件进行。例如,在各个非限制性实施例中,方法 500 可部分或完全用 NIC、支持叠加的加速 NIC、嵌入 NIC 或加速 NIC 中和 / 或在 NIC 或加速 NIC 内工作的处理器(比如 CPU、ASIC、FPGA 等)、嵌入在 NIC 或加速 NIC 内的计算机可读存储介质中的计算机程序代码等等进行。

[0061] 如图 5 中所示,方法 500 可始于操作 502,在操作 502 中,从 VM 接收通信量。通信量包括至少一个分组,按照一个实施例,通信量可由服务器或其某个组件(比如虚拟化平台、接口等)接收。一般地,分组是由 VM 产生的 IP 分组,不过在其它实施例,它可以根据各个实施例被单播、广播、叠加封装等。

[0062] 在操作 504 中,按照散列算法,散列所述至少一个分组的至少一部分,以获得散列值。可以使用本领域中已知的任何适当的散列算法。另外,散列算法可以是完全散列算法或非完全散列算法,并且可被这样选择,以致可能的散列值对应于可用于把通信量继续发送给网络的上行链路的数目,在一种途径中,例如,散列算法可完全地散列成上行链路的数目。在另一种途径中,可能的散列值的数目可对应于为其使通信量负载均衡的团队的成员,例如,如果在数目为 n 的上行链路上进行负载均衡,那么散列算法可具有 n 个可能的散列值。

[0063] 在更多的途径中,可以跨不同组的上行链路,在不同的加速 NIC 上,在不同的服务器上等等,应用不同的散列算法。

[0064] 按照一些实施例,分组可被封装在叠加分组内,在这种情况下,叠加封装的分组的内部分组被散列。

[0065] 所述至少一个分组的任意部分或者全部可被散列,比如分组的报头,分组的有效负荷,外部报头,内部报头等等。如果分组被叠加封装,那么内部报头最好被散列,不过按照各个实施例,可以散列分组的任何部分。

[0066] 在操作 506 中,本领域的技术人员会理解,所述至少一个分组被封装在叠加封装的分组中。在一种途径中,所述至少一个分组由(能够提供叠加功能的)加速 NIC 封装,不过也可由支持叠加的任何组件或设备封装。

[0067] 在操作 508 中,根据散列值,选择上行链路。以这种方式,由于上行链路是根据散列值选择的,因此可在一组上行链路、NIC 上的所有链路、服务器上的所有上行链路、在任何给定时间可用的所有上行链路等等之间,使通信量负载均衡。

[0068] 在操作 510 中,选择的上行链路的 MAC 地址作为 SMAC 地址被保存在叠加封装分组的外部报头中。这样,以期望的方式,不会向网络的任何组件增加额外处理地在上行链路之间使输出通信量负载均衡。

[0069] 在操作 512 中,通过选择的上行链路发送叠加封装的分组。这是在可用上行链路之间进行负载均衡的方式。当收到所述至少一个分组时,存在可用于继续发送所述至少一个分组的多个、一组或者一批上行链路。在操作 512 中,在选择的上行链路上发送所述至少一个分组,选择的上行链路将随由散列算法确定的多个因素而变化。

[0070] 在例证实施例中,内部分组报头字段的散列可用于选择将用于传送该分组的团队成员。散列可以基于一个或多个参数,包括:对应于 VM 的虚拟端口(比如,分配给 VM 的虚拟端口),内部分组报头 SMAC(inner_smac) 地址,内部分组报头目的地 MAC(inner_dmac) 地址,内部信源 IP(SIP) 地址,内部目的地 IP(DIP) 地址,它们的任意组合,等等。

[0071] 散列算法计算可导致选择特定团队或组中的多个可用物理 NIC 之一。随后可用选择的 NIC 的 MAC 地址重写分组报头(它最初包括虚拟 MAC(VMAC))的外部 SMAC。这导致分组在特定的 NIC 上被发送,所述特定的 NIC 总是传送相同的外部报头 SMAC(它是 NIC 的 MAC 地址)。在本例证实施例中描述的任何 NIC 可以是加速 NIC(如果叠加功能将被用于在该 NIC(一个 NIC 具有多个上行链路)或者提供多个上行链路的多个 NIC(每个 NIC 一个上行链路)的各上行链路上发送分组的话)或者在一个或多个服务器(一般,单个服务器)内的 NIC 的组合。

[0072] 按照另一个实施例中,可以利用能够执行这样的计算机可读程序代码的处理器,从计算机程序产品执行方法 500。

[0073] 在另一个实施例中,方法 500 可由系统(比如服务器、NIC、具有叠加功能的 NIC、网关、它们的某种组合等)执行。

[0074] 现在参见图 6,图中按照一个实施例示出了在服务器的接收侧使负载均衡的方法 600。在各个实施例中,可以在图 1-4 中描述的环境等中的任意环境中,按照本发明进行方法 600。当然,当阅读本说明时,本领域的技术人员理解,方法 600 中可以包括比在图 6 中具体说明的那些操作更多或更少的操作。

[0075] 方法 600 的各个步骤可由工作环境的任何适当的一个或多个组件进行。例如,在各个非限制性实施例中,方法 600 可部分或完全用支持叠加的加速 NIC、NIC、嵌入 NIC 或加速 NIC 中和/或在 NIC 或加速 NIC 内工作的处理器(比如 CPU、ASIC、FPGA 等)、嵌入在 NIC 或加速 NIC 内的计算机可读存储介质中的计算机程序代码等等进行。

[0076] 如图 6 中所示,方法 600 可始于操作 602,在操作 602 中,接收广播 ARP 请求分组。换句话说,接收不仅向方法 600 中的接收设备广播的而且可能还向其它设备广播的 ARP 请求分组。ARP 广播请求分组可由服务器、尤其是服务器的 NIC 接收。

[0077] 在操作 604 中,判定是否应当作出响应。一般根据在广播 ARP 请求分组中指定的地址是否由接收广播 ARP 请求分组的服务器托管,作出所述判定。如果是,那么应当作出响应,从而方法 600 进入操作 606。如果否,那么不应当作出响应,从而方法 600 进入操作 616。

[0078] 在操作 606 中,按照散列算法,散列广播 ARP 请求分组的至少一部分,以获得散列值。按照各个实施例,可以散列分组的任意部分或全部。

[0079] 可以使用本领域中已知的任何适当的散列算法。另外,散列算法可以是完全散列算法或非完全散列算法,并且可被这样选择,以致可能的散列值的数目对应于可用于继续发送通信量的上行链路的数目,在一种途径中,例如,散列算法可完全地散列,从而产生和上行链路的数目相同的多个可能的散列值。在另一种途径中,散列可对应于为其使通信量

负载均衡的团队的成员,例如,如果在数目为 n 的上行链路上进行负载均衡,那么散列算法可基于所述数目 n 。

[0080] 散列算法可以基于一个或多个参数,比如:与为其发起广播 ARP 请求分组的 VM 对应的虚拟端口,内部分组报头 SMAC 地址,内部分组报头 DMAC 地址,内部 SIP 地址,内部 DIP 地址,等等。

[0081] 在更多的途径中,可以跨不同组的上行链路,在不同的加速 NIC 上,在不同的服务器上等等,应用不同的散列算法。

[0082] 所述至少一个分组的任何部分或者全部可被散列,比如分组的报头、分组的有效负荷、外部报头、内部报头等等。

[0083] 在操作 608 中,根据散列值,选择上行链路。这样,由于每次收到广播 ARP 请求分组时,根据散列选择上行链路,因此能够在一组上行链路,NIC 上的所有链路,服务器上的所有上行链路,在任何给定时间可用的所有上行链路等等之间,使可从广播 ARP 请求分组的信源接收的通信量负载均衡。

[0084] 在操作 610 中,选择的上行链路的 MAC 地址作为信源 MAC (SMAC) 地址被保存到对收到的广播 ARP 请求分组的响应,该响应被发送给广播 ARP 请求分组的信源。这样,最后收到该响应分组的信源设备能够保存响应设备的选择的上行链路的 MAC 地址,以便未来在该选择的上行链路上,把通信量回送给所述响应设备。

[0085] 在可选的操作 612 中,生成单播 ARP 请求分组,并回送给接收的广播 ARP 请求分组的信源。这样,接收并响应广播 ARP 请求分组的服务器或设备可学习发端设备的地址,该地址可能不同于保存在广播 ARP 请求分组中的 SMAC。

[0086] 在可选的操作 614 中,在发送单播 ARP 请求分组之后,可收到对单播 ARP 请求分组的响应。在这种情况下,可以学习并保存关于广播 ARP 请求分组的信源的地址信息,以便随后与该特定地址通信。这样,不仅按照期望的方式,在上行链路之间使输出通信量负载均衡,而且还可按照不向网络的任何组件增加额外处理的方式,在可用链路之间使输入通信量负载均衡。

[0087] 在操作 616 中,或者因为广播 ARP 请求分组的接收者不是期望的接收者,或者因为归因于本领域的技术人员理解的多个因素,例如,认为信源不被允许与目标通信,信源是间歇性的并将会变化,目标是间歇性的等等,而认为不必学习信源,广播 ARP 请求分组不被用于学习广播 ARP 请求分组的信源。

[0088] 注意,接收的广播 ARP 请求分组中的 SMAC 不被学习,因为这会重置和 / 或废除先前已实现的任何负载均衡。

[0089] 例如,在其中远程主机利用一个或多个单播 ARP 响应分组在其可用上行链路之间使来自网络中的各个主机的输入通信负载均衡的情况下,当所述远程主机向网络中的各个主机发送广播 ARP 请求分组时,如果网络中的各个主机高速缓存包含在广播 ARP 请求分组中的 SMAC,那么这会导致重置所述远程主机早先利用一个或多个一个或多个单播 ARP 响应分组预先 (和有意地) 实现的接收负载均衡。为了避免这种情况,网络中的每个主机不把广播 ARP 请求分组看作用于 MAC 学习的信源。

[0090] 在一些实施例中,可以独立地使输出和输入通信量流负载均衡,或者可以取决于某个其它因素、编组和 / 或条件,类似地使它们负载均衡。

[0091] 按照另一个实施例,可以利用能够执行这样的计算机可读程序代码的处理器,从计算机程序产品执行方法 600。

[0092] 在另一个实施例中,方法 600 可由系统(比如服务器、NIC、具有叠加功能的 NIC、网关、它们的某种组合等)执行。

[0093] 在例证实施例中,与交换机有关的组队能够在无(需要保存在现有设备中的)新逻辑部的情况下,使接收的通信量负载均衡。在与交换机无关的组队中,通过在任意数目的可用团队成员(上行链路、NIC 等)之间使 ARP 响应分组负载均衡,可以实现接收侧负载均衡。选择的团队成员的 MAC 地址被嵌入单播 ARP 响应分组中。

[0094] 在各个实施例中,准则可用于处理 ARP 消息。在第一实施例中,广播 ARP 请求分组决不被转发给操作系统(OS)内核(相反,如果目标 IP(TIP)地址匹配 OS 内核的地址,那么组队逻辑部可构成单播 ARP 响应),可对信源 IP(SIP)地址进行散列,以选择团队成员,选择的 NIC 的 MAC 地址可用在响应中。另外,当收到广播 ARP 请求分组时,支持叠加的交换机(比如分布式叠加虚拟以太网(DOVE)交换机)向发端设备单播 ARP 请求分组,以触发发端设备答复以单播响应,所述单播响应可被 OS 内核用于学习关于发端设备的地址信息。

[0095] 在另一个实施例中,单播 ARP 响应分组可以总是被转发给 OS 内核,以便学习。这避免了组队逻辑部维持 ARP 高速缓存的需要。这还使来自不同物理主机的通信量被负载均衡到团队的不同成员。注意接收侧负载均衡非常粗略,没有能力使来自驻留在相同主机上的不同远程 VM 的通信量负载均衡——来自物理服务器上的所有 VM 的通信量将进入相同的团队成员。

[0096] 同样地,当外部网关或路由器为本地支持叠加的交换机(比如 DOVE 交换机)发送 ARP 请求时,该交换机可仅以团队成员之一进行响应。这意味从下层的网络域(比如层 2 域)外面进入支持叠加的交换机中的所有通信量只可利用团队成员之一。在不支持叠加的组队解决方案中,这种缺陷也常见。

[0097] 现在参见图 7,图中按照一个实施例示出了在服务器的接收侧使负载均衡的方法 700。在各个实施例中,可以在图 1-4 中描述的环境等中的任意环境中,按照本发明进行方法 700。当然,当阅读本说明时,本领域的技术人员理解,方法 700 中可以包括比在图 7 中具体说明的那些操作更多或更少的操作。

[0098] 方法 700 的各个步骤可以用工作环境的任何适当组件执行。例如,在各个非限制性实施例中,方法 700 可部分或完全用 NIC、支持叠加的加速 NIC、嵌入 NIC 或加速 NIC 中和/或在 NIC 或加速 NIC 内工作的处理器(比如 CPU、ASIC、FPGA 等)、嵌入在 NIC 或加速 NIC 内的计算机可读存储介质中的计算机程序代码等等进行。

[0099] 如图 7 中所示,方法 700 可始于操作 702,在操作 702 中,接收单播 ARP 请求分组。单播 ARP 请求分组可被服务器、尤其是服务器的 NIC 接收。

[0100] 在操作 704 中,判定是否应当作出响应。一般根据在单播 ARP 请求分组中指定的地址是否由接收单播 ARP 请求分组的服务器托管,作出所述判定。如果是,那么应当作出响应,从而方法 700 进入操作 706。如果否,那么不应当作出响应,从而方法 700 进入操作 716。

[0101] 在操作 706 中,按照散列算法,散列单播 ARP 请求分组的至少一部分,以获得散列值。按照各个实施例,可以散列所述分组的任意部分或者全部。

[0102] 可以使用本领域中已知的任何适当的散列算法。另外,散列算法可以是完全散列

算法或非完全散列算法,并且可被选择成对应于可用于继续发送通信量的上行链路的数目,在一种途径中,例如,散列算法可完全地散列成上行链路的数目。在另一种途径中,散列可对应于为其使通信量负载均衡的团队的成员,例如,如果在数目为 n 的上行链路上进行负载均衡,那么散列算法可基于所述数目 n 。

[0103] 散列算法可以基于一个或多个参数,比如:与为其发起单播 ARP 请求分组的 VM 对应的虚拟端口,内部分组报头 SMAC 地址,内部分组报头 DMAC 地址,内部 SIP 地址,内部 DIP 地址,等等。

[0104] 在更多的途径中,可以跨不同组的上行链路,在不同的加速 NIC 上,在不同的服务器上等等,应用不同的散列算法。

[0105] 所述至少一个分组的任意部分或者全部可被散列,比如分组的报头,分组的有效负载,外部报头,内部报头等等。

[0106] 在操作 708 中,根据散列值选择上行链路。这样,由于每次收到单播 ARP 请求分组时,根据散列选择上行链路,因此能够在一组上行链路、NIC 上的所有链路、服务器上的所有上行链路、在任何给定时间可用的所有上行链路等等之间,使可从单播 ARP 请求分组的信源接收的通信量负载均衡。

[0107] 在操作 710 中,选择的上行链路的 MAC 地址作为 SMAC 地址被保存在对收到的单播 ARP 请求分组的响应,该响应被回送给单播 ARP 请求分组的信源。这样,最后收到该响应分组的任何设备将保存该 MAC 地址,以便把通信量回送给发端设备。这样,不仅按照期望的方式,在上行链路之间使输出通信量负载均衡,而且还可按照不向网络的任何组件增加额外处理的方式,在可用链路之间使输入通信量负载均衡。

[0108] 在操作 712 中,向单播 ARP 请求分组学习信源的地址信息(例如,IP 地址,SMAC 地址等)。在这种情况下,可以学习并保存关于单播 ARP 请求分组的信源的地址信息,以便随后与该特定地址通信。

[0109] 在操作 714 中,或者因为单播 ARP 请求分组的接收者不是期望的接收者,或者因为归因于本领域的技术人员理解的多个因素,例如,认为信源不被允许与目标通信,信源是间歇性的并将会变化,目标是间歇性的等等,而认为不必学习信源,单播 ARP 请求分组不被用于学习单播 ARP 请求分组的信源。

[0110] 在一些实施例中,可以独立地使输出和输入通信流量负载均衡,或者可以取决于某个其它因素、编组和/或条件,类似地使它们负载均衡。

[0111] 按照另一个实施例,可以利用能够执行这样的计算机可读程序代码的处理器,从计算机程序产品执行方法 700。

[0112] 在另一个实施例中,方法 700 可由系统(比如服务器、NIC、具有叠加功能的 NIC、网关、它们的某种组合等)执行。

[0113] 按照例证实施例,在多个上行链路上使通信量负载均衡的系统包括用硬件实现的处理器,所述处理器适合于执行逻辑部(比如,CPU、FPGA、ASIC、微控制器等等),适合于接收 ARP 请求分组(广播 ARP 请求分组或单播 ARP 请求分组)的逻辑部,适合于按照散列算法散列 ARP 请求分组的至少一部分以获得散列值的逻辑部,适合于根据所述散列值从可用于向网络发送通信量的多个上行链路中选择上行链路的逻辑部,适合于把与选择的上行链路对应的 MAC 地址作为 SMAC 地址保存在对 ARP 请求分组的响应中的逻辑部,和适合于在与

选择的上行链路对应的 MAC 地址被保存为 SMAC 地址的情况下把所述响应回送给 ARP 请求分组的信源的逻辑部。

[0114] 在另一个实施例中,其中当 ARP 请求分组是单播的时,从 ARP 请求分组学习并保存关于 ARP 请求分组的信源的地址信息,以便随后与所述信源通信。另外,当 ARP 请求分组被广播时,系统还可包括适合于生成单播 ARP 请求分组并把单播 ARP 请求分组发送给广播 ARP 请求分组的信源的逻辑部,适合于接收来自广播 ARP 请求分组的信源的响应的逻辑部,所述响应包括关于广播 ARP 请求分组的信源的地址信息,和适合于学习并保存关于广播 ARP 请求分组的信源的地址信息的逻辑部。此外,当未收到来自广播 ARP 请求分组的信源的响应时,不从广播 ARP 请求分组学习地址信息。

[0115] 在更多的途径中,可用系统或计算机程序产品实现这里描述的任何方法。系统可包括计算机可读存储介质和适合于执行期望的方法的逻辑部,所述逻辑部被保存到计算机可读存储介质上。计算机程序产品可包括保存计算机可读程序代码的计算机可读存储介质,所述计算机可读程序代码用来执行期望的方法。

[0116] 尽管上面说明了各种实施例,不过应明白这些实施例只是作出例子给出的,而不是对本发明的限制。从而,本发明的实施例的范围不应由任意上述例证实施例限制,相反只应按照所附权利要求及其等同物限定。

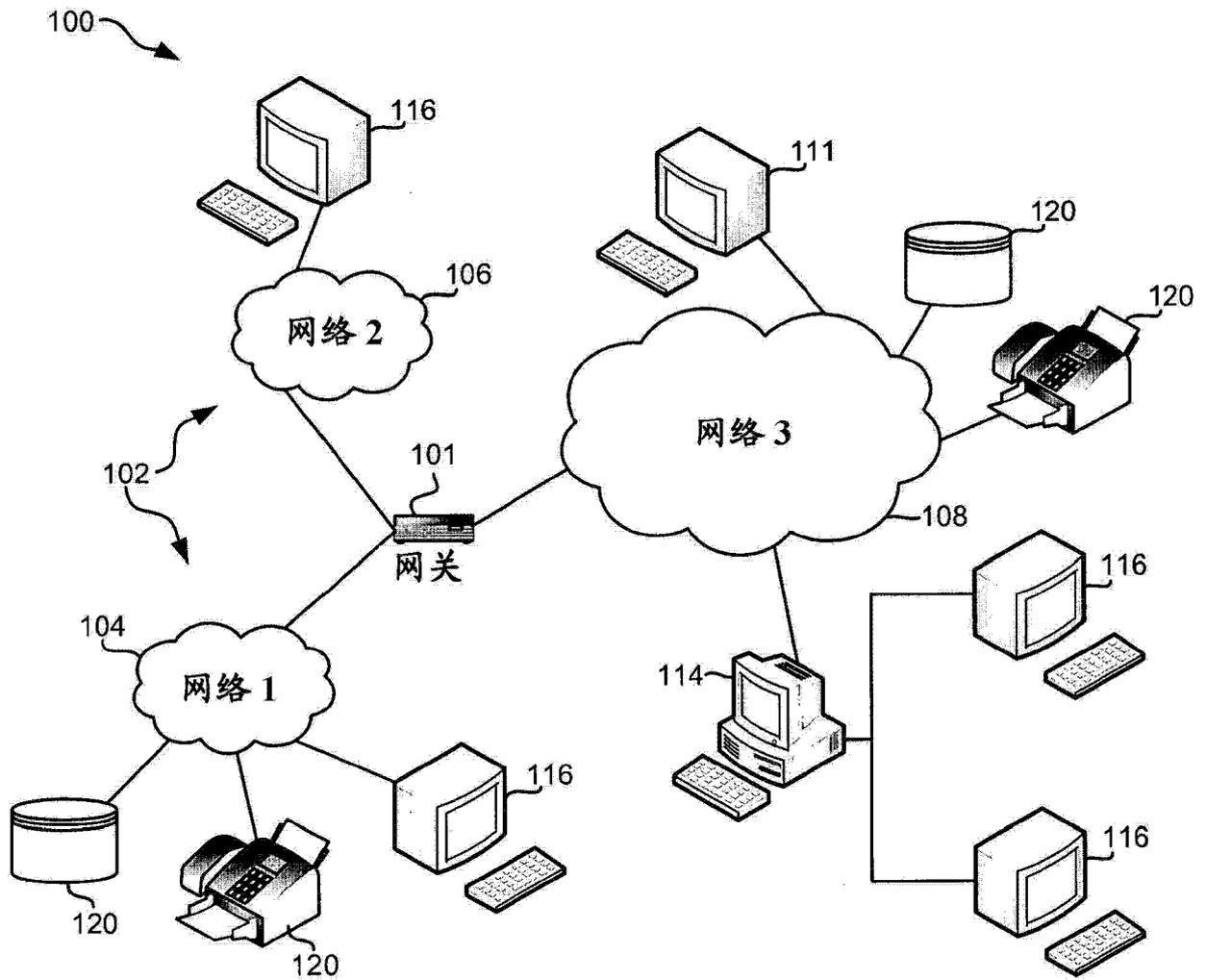


图 1

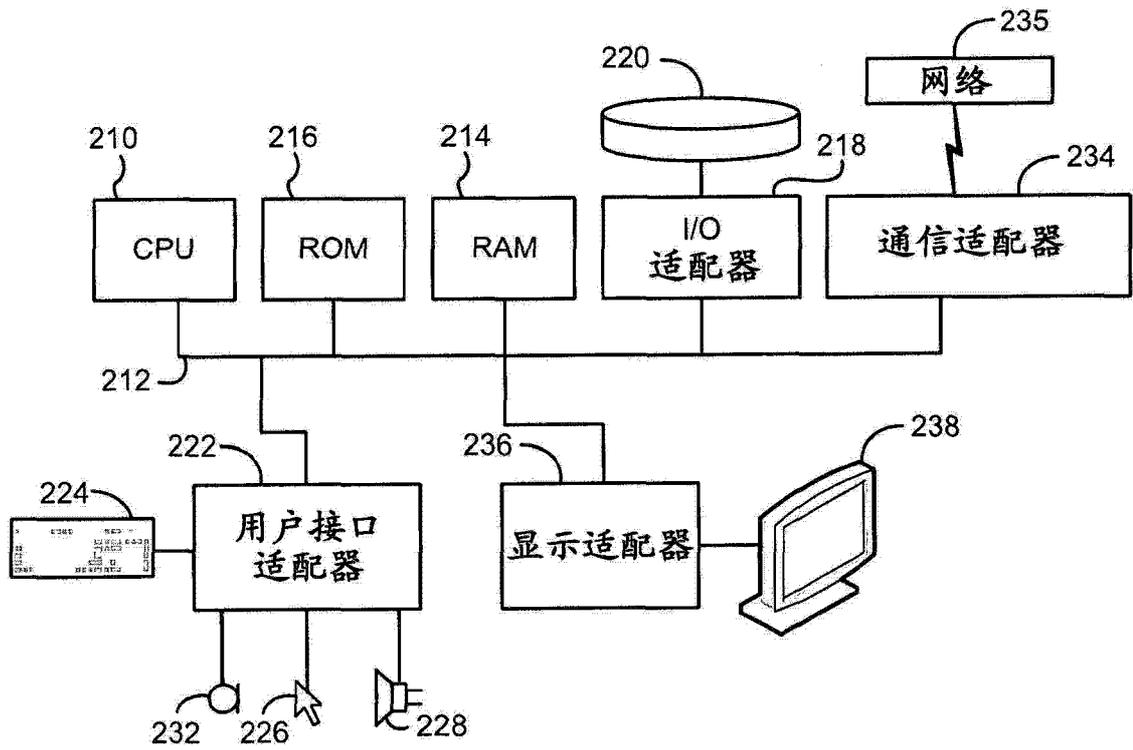


图 2

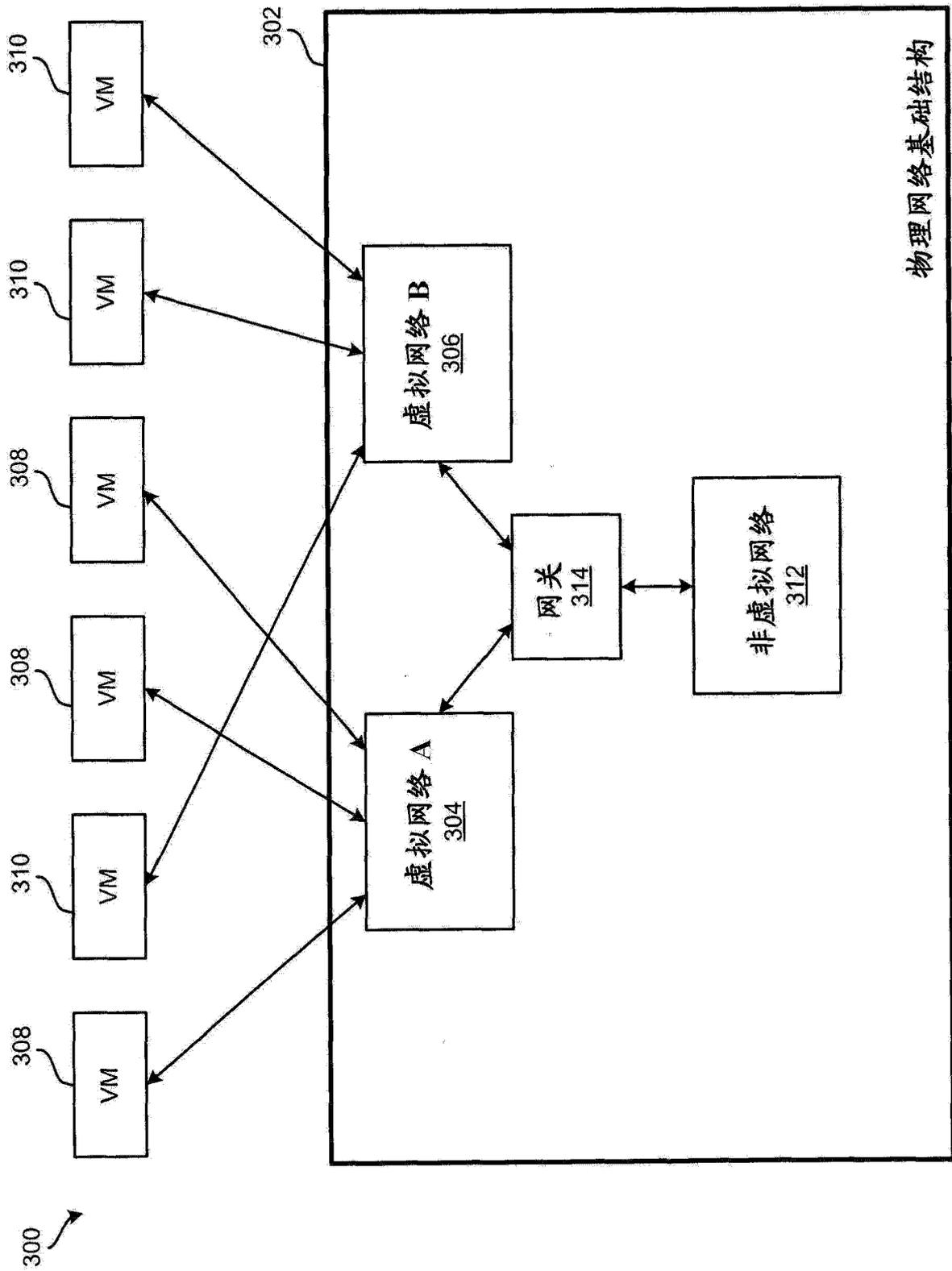


图 3

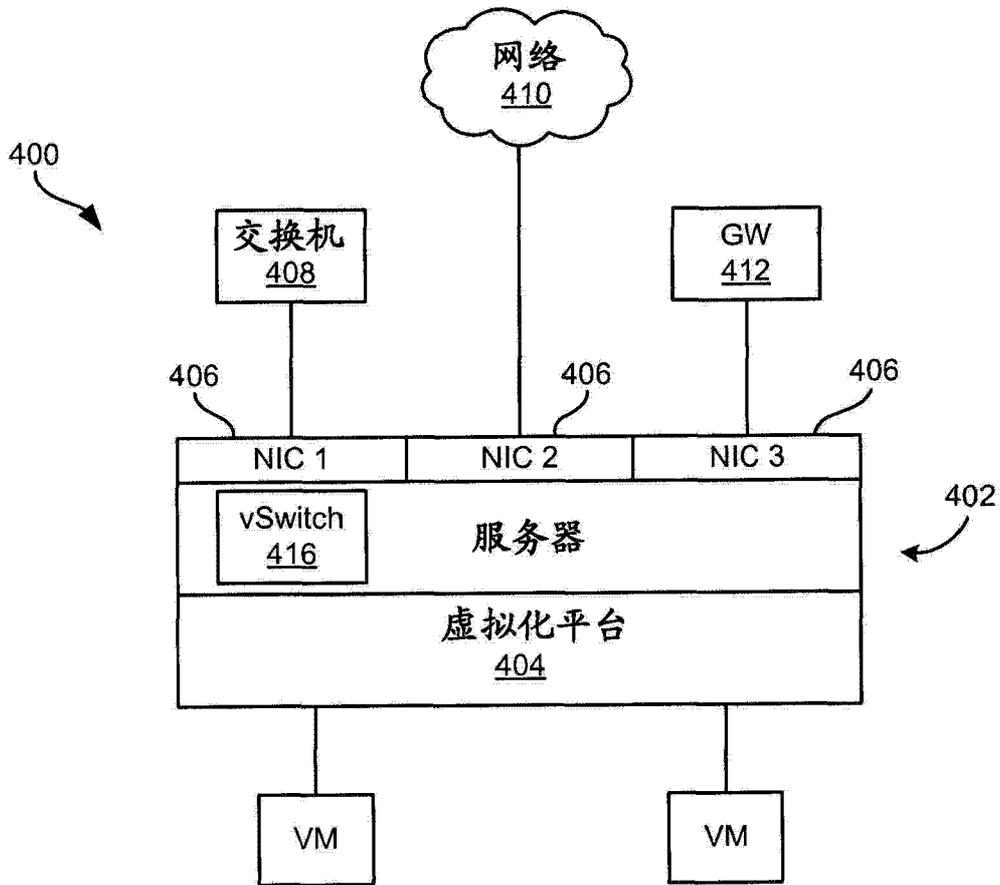


图 4

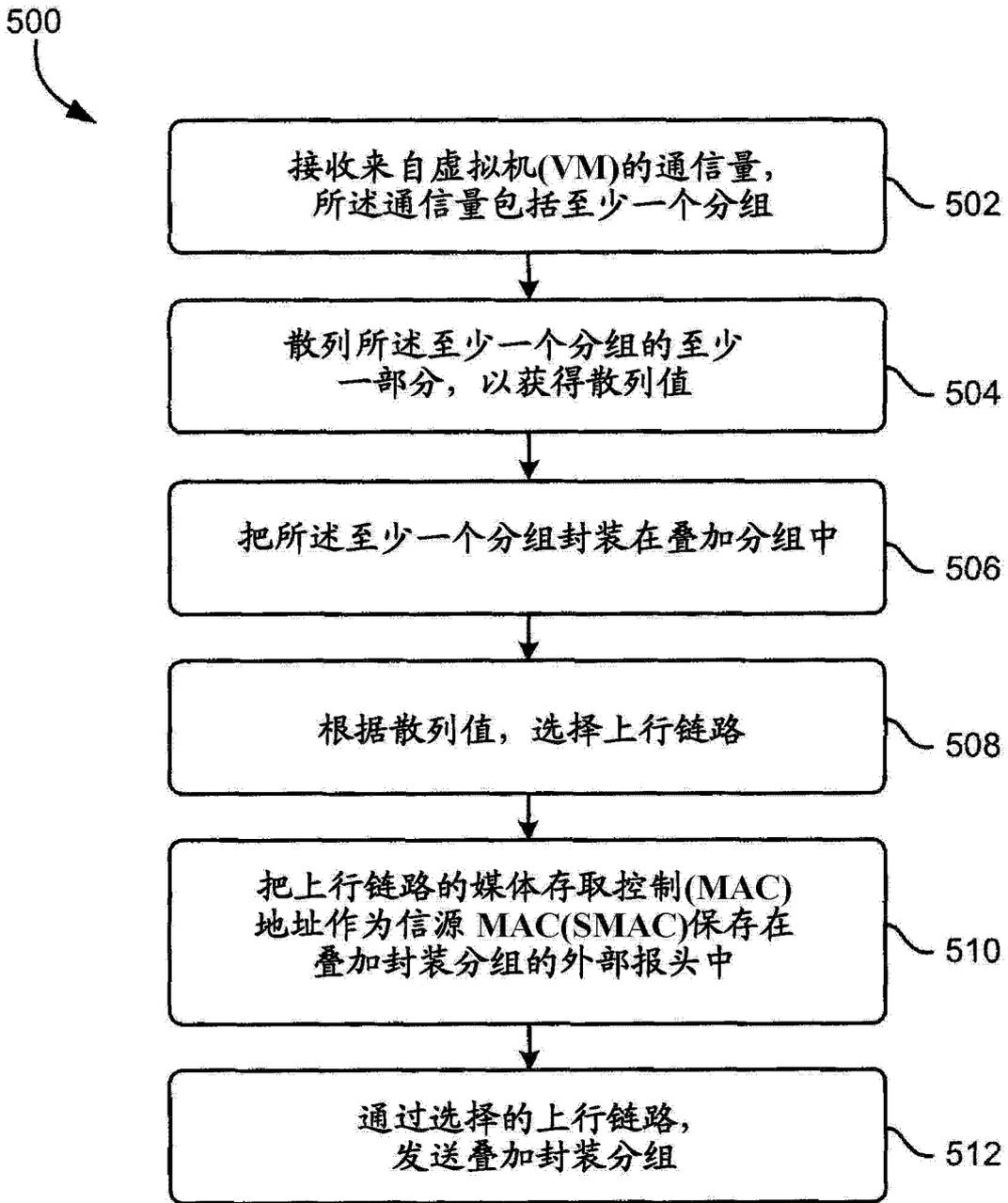


图 5

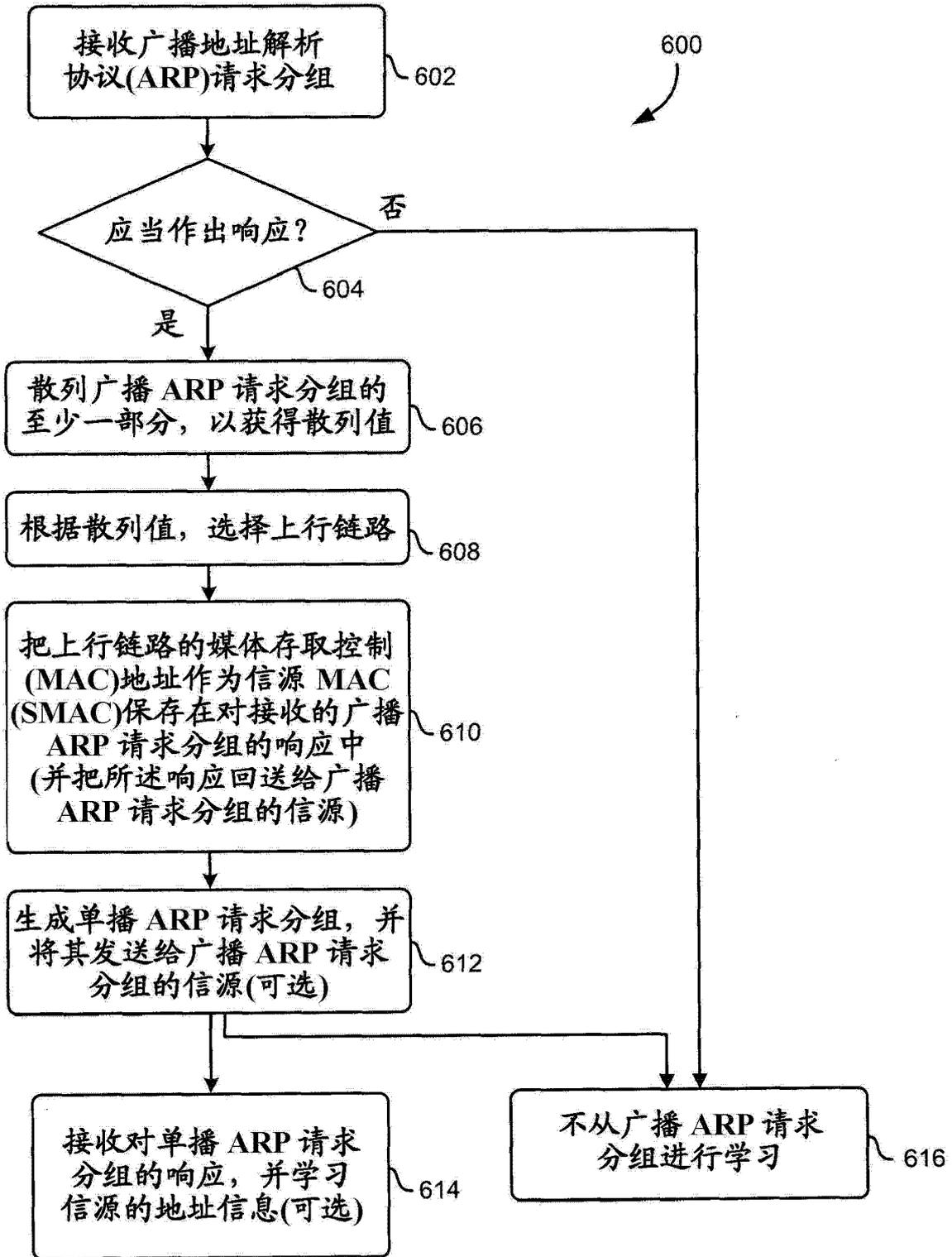


图 6

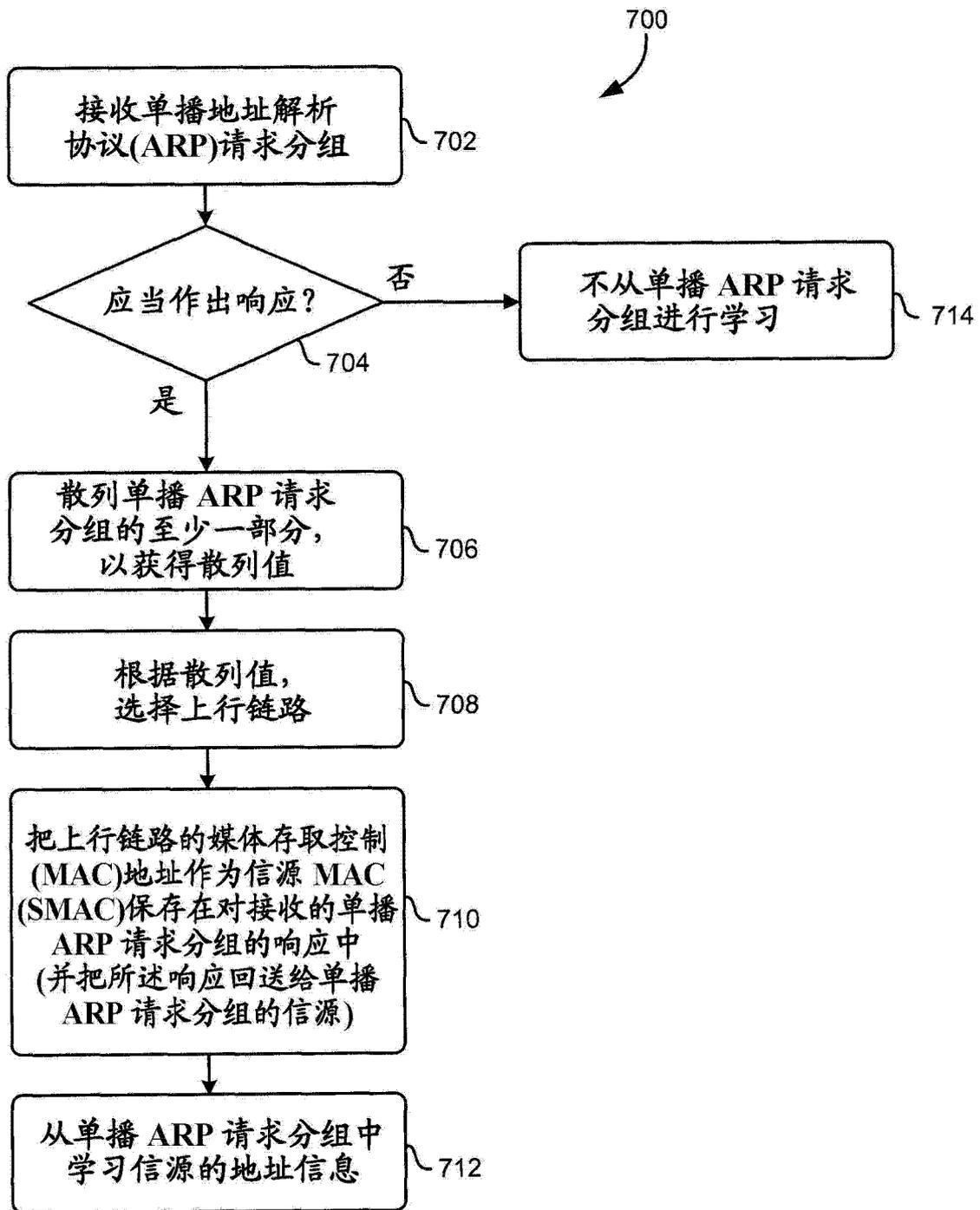


图 7