



(12)发明专利

(10)授权公告号 CN 104871483 B

(45)授权公告日 2018.09.04

(21)申请号 201380064577.2

(22)申请日 2013.10.03

(65)同一申请的已公布的文献号
申请公布号 CN 104871483 A

(43)申请公布日 2015.08.26

(30)优先权数据
13/649026 2012.10.10 US

(85)PCT国际申请进入国家阶段日
2015.06.10

(86)PCT国际申请的申请数据
PCT/IB2013/059100 2013.10.03

(87)PCT国际申请的公布数据
W02014/057402 EN 2014.04.17

(73)专利权人 瑞典爱立信有限公司

地址 瑞典斯德哥尔摩

(72)发明人 R.米什拉 H.格林 S.鲍克
J.肯普夫 M.塔蒂帕穆拉

(74)专利代理机构 中国专利代理(香港)有限公
司 72001

代理人 朱君 姜甜

(51)Int.Cl.
H04L 12/18(2006.01)

(56)对比文件
US 5764875 A, 1998.06.09,
US 6735200 B1, 2004.05.11,
CN 102449963 A, 2012.05.09,

审查员 王晓玲

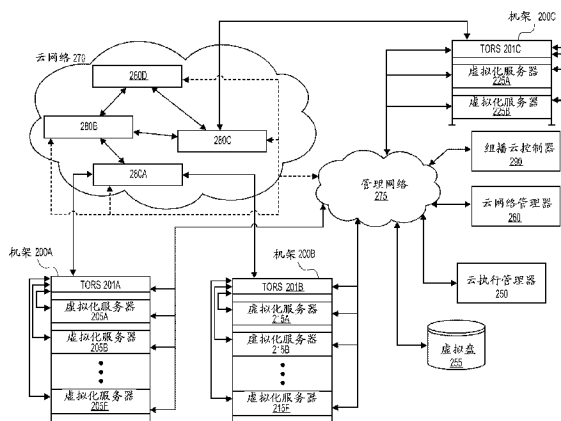
权利要求书5页 说明书20页 附图21页

(54)发明名称

用于管理云网络中的组播业务的方法和装置

(57)摘要

云系统中的组播云控制器(“MCC”)实现处理器来管理云网络中的组播业务。MCC耦合于至少一个虚拟化服务器用于托管一个或多个虚拟机(“VM”),其中虚拟化服务器包括至少一个虚拟交换机(“VS”),其支持多协议标记交换(MPLS),并且虚拟交换机耦合于支持MPLS的架顶式交换机(“TORS”)。MPLS用于在云系统中支持组播数据业务使得系统和方法减少状态并且是可标度的。



1. 一种由云系统中的组播云控制器“MCC”执行来管理云网络中的组播业务的方法, 所述MCC耦合于用于托管一个或多个虚拟机“VM”的至少一个虚拟化服务器, 其中所述虚拟化服务器包括至少一个虚拟交换机“VS”, 所述虚拟交换机支持多协议标记交换MPLS, 并且所述虚拟交换机耦合于支持MPLS的架顶式交换机“TORS”, 所述方法包括以下步骤:

在所述MCC接收(1201)加入消息, 其中所述加入消息指示VM请求了加入组播组;

将所述VM添加(1203)到租户的VM集, 其中所述VM集预订所述组播组;

通过所述MCC使用所述MCC维持的VM记录查找(1205)对于所述VM的VS和TORS;

对于所述VM的VS处安装(1207)至少一个流表条目, 所述至少一个流表条目配置所述组播组的入站组播业务转发到所述VM并且配置所述组播组的出站组播业务发送到所述VM的所述TORS; 以及

在所述VM的所述TORS处安装(1211)至少一个流表条目, 所述TORS的所述至少一个流表条目配置所述TORS为通过MPLS转发所述组播组的出站组播业务到具有预订所述组播组的VM的所述云系统的其他TORS并且配置所述TORS转发所述组播组的进入组播业务到所述VM的所述VS。

2. 如权利要求1所述的方法, 其进一步包括以下步骤:

在所述VM的VS处安装流表条目以放弃从所述VM的TORS环回的组播组业务。

3. 如权利要求1所述的方法, 其中接收所述加入消息进一步包括以下步骤:

接收来自具有组播组地址的VM的互联网组组播协议(IGMP)报告消息或组播发现侦听器(MLD)报告消息中的任何一个。

4. 如权利要求1所述的方法, 其进一步包括以下步骤:

通过所述MCC使用在所述加入消息中的所述VM的MAC地址查找租户ID。

5. 如权利要求1所述的方法, 其进一步包括以下步骤:

通过所述MCC使用从所述加入消息导出的租户ID查找租户的组播组管理表。

6. 如权利要求1所述的方法, 其进一步包括以下步骤:

构造具有到VS记录的指针和到TORS记录的指针的VM记录, 所述VM记录与所述VM关联, 所述VS记录与所述VM的所述VS关联并且所述TORS记录与所述VM的TORS关联。

7. 如权利要求1所述的方法, 其进一步包括以下步骤:

插入所述VM的VM记录到在加入消息中的组播组地址所识别的所述组播组的组播组管理表条目。

8. 如权利要求1所述的方法, 其进一步包括以下步骤:

从在加入消息中的组播组地址所识别的所述组播组的组播组管理表条目检索所述组播组的MPLS标记用于配置所述VM的VS和所述VM的TORS以处理组播业务。

9. 如权利要求1所述的方法, 其进一步包括以下步骤:

在VM的VS处安装流表条目以匹配包的目的互联网协议地址与组播组地址, 以匹配包源地址与VM的互联网协议地址并且匹配源MAC地址与所述VM的MAC地址, 并且将所述组播组的MPLS标记压入所述包并且将包复制到输出组。

10. 如权利要求1所述的方法, 其进一步包括以下步骤:

在VM的VS处安装流表条目以匹配所述包的源IP地址和所述VM的IP地址并且匹配TORS的端口, 并且放弃所述包以防止循环。

11. 如权利要求1所述的方法,其进一步包括以下步骤:

安装所述VS的输入组中的转发动作以转发组播组业务到所述VM的互联网协议地址。

12. 如权利要求1所述的方法,其进一步包括以下步骤:

检查所述VM是否是在TORS预订组播组的第一VM。

13. 如权利要求12所述的方法,在所述VM是所述第一VM的情况下,还包括下列步骤:

安装在所述TORS的流表条目以将骨干路由标记压入到包中。

14. 如权利要求12所述的方法,在所述VM是所述第一VM的情况下,还包括下列步骤:

安装在所述TORS的流表条目以匹配组播标记和转发到目的TORS的输出端口。

15. 如权利要求12所述的方法,在所述VM是所述第一VM的情况下,还包括下列步骤:

安装在其他TORS的流表条目以转发组播组包到具有预订VM的其他TORS。

16. 如权利要求12所述的方法,在所述VM是所述第一VM的情况下,还包括下列步骤:

安装在所述TORS的流表条目以转发组播组包到具有预订所述组播组的VM的机架中的所有VS。

17. 如权利要求12所述的方法,在所述VM是所述第一VM的情况下,还包括下列步骤:

安装在所述TORS的流表条目以匹配包中的骨干路由标记,以解封所述骨干路由标记并且以转发所述包到所述TORS的下个表。

18. 如权利要求17所述的方法,还包括下列步骤:

安装流表条目以匹配所述包中组播组标记,以解封所述组播组标记,并且以经由关联VS使用输入组转发所述包到预订VM。

19. 如权利要求1所述的方法,还包括下列步骤:

推断从组播地址的组播源的加入,以及

安装在所述组播源的TORS处的至少一个流表条目,所述组播源的所述TORS处的至少一个流表条目配置所述TORS通过MPLS转发所述组播组的所述出站组播业务到具有预订所述组播组的VM的所述云系统的其他TORS。

20. 一种由云系统中的组播云控制器“MCC”执行来管理云网络中的组播业务的方法,所述MCC耦合于用于托管一个或多个虚拟机“VM”的至少一个虚拟化服务器,其中所述虚拟化服务器包括至少一个虚拟交换机“VS”,其支持多协议标记交换MPLS,并且所述虚拟交换机耦合于支持MPLS的架顶式交换机“TORS”,所述方法包括以下步骤:

在所述MCC接收(1201)以互联网组管理协议报告消息或组播侦听器发现报告消息的形式的加入消息,其中所述加入消息指示VM请求了加入组播组;

将所述VM添加(1203)到租户的VM集,其中所述VM集预订所述组播组,所述VM集在组播组管理表中连同组播组标记跟踪;

通过所述MCC使用所述MCC维持的VM记录查找(1205)对于所述VM的VS和TORS;

在所述VS的输入组中安装(1321)转发动作以将组播组业务转发到所述VM的互联网协议地址;以及

检查(1501)所述VM是否是在TORS预订组播组的第一VM,在所述VM是所述第一VM的情况下,

在TORS处安装(1503)流表条目以将骨干路由标记压入包上,

在TORS处安装(1505)流表条目以匹配组播标记并且转发所述包到目的TORS的输出端

口,

在其他TORS处安装(1507)流表条目以将组播组包转发到具有预订VM的其他TORS,

在所述TORS处安装(1509)流表条目以将组播组包转发到具有预订组播组的VM的机架中的所有VS,

在所述TORS处安装(1511)流表条目以匹配包中的骨干路由标记,解封所述骨干路由标记并转发所述包到所述TORS的下个表,以及

安装(1513)流表条目以与所述包中的组播组标记匹配,对所述组播组标记解封并且经由关联VS使用输入组将所述包转发到预订VM。

21. 一种在云系统中实现的组播云控制器“MCC”(290),所述MCC耦合于用于托管一个或多个虚拟机“VM”的至少一个虚拟化服务器,其中所述虚拟化服务器包括至少一个虚拟交换机“VS”,其支持多协议标记交换MPLS,并且所述虚拟交换机耦合于支持MPLS的架顶式交换机“TORS”,所述MCC包括:

处理器,用于执行加入组模块,

所述加入组模块管理组播组加入操作,其中所述加入组模块配置成在所述MCC接收以互联网组管理协议报告消息或组播侦听器发现报告消息的形式的加入消息,其中所述加入消息指示VM请求了加入组播组,所述加入组模块还配置成将所述VM添加到租户的VM集,其中所述VM集预订所述组播组,配置成通过MCC使用所述MCC维持的VM记录查找所述VM的VS和TORS,配置成安装在所述VM的VS的至少一个流表条目,所述至少一个流表条目配置所述组播组的入站组播业务为转发到所述VM并且配置所述组播组的出站组播业务被发送到所述VM的TORS,并且配置成在所述VM的TORS处安装至少一个流表条目,所述TORS的所述至少一个流表条目配置所述TORS为通过MPLS转发所述组播组的出站组播业务到具有预订所述组播组的VM的云系统的其他TORS并且配置所述TORS为转发进入所述组播组的组播业务到所述VM的所述VS。

22. 如权利要求21所述的MCC,其中所述加入组模块还配置成在所述VM的VS安装流表条目以放弃从所述VM的TORS环回的组播组业务。

23. 如权利要求21所述的MCC,其中接收所述加入消息和其中所述加入组模块还配置成接收来自具有组播组地址的VM的互联网组组播协议(IGMP)报告消息或组播发现侦听器(MLD)报告消息中的任一个。

24. 如权利要求21所述的MCC,其中所述加入组模块还配置成通过MCC使用在所述加入消息中的VM的MAC地址查找租户ID。

25. 如权利要求21所述的MCC,其中所述加入组模块还配置成通过MCC使用从所述加入消息导出的租户ID查找租户的组播组管理表。

26. 如权利要求21所述的MCC,其中所述加入组模块配置成构造具有到VS记录的指针和到TORS记录的指针的VM记录,所述VM记录与所述VM关联,所述VS记录与所述VM的所述VS关联并且所述TORS记录与所述VM的TORS关联。

27. 如权利要求21所述的MCC,其中所述加入组模块插入所述VM的VM记录到在所述加入消息中的组播组地址所识别的所述组播组的组播组管理表条目。

28. 如权利要求21所述的MCC,其中所述加入组模块配置成从在所述加入消息中的组播组地址所识别的所述组播组的组播组管理表条目检索所述组播组的MPLS标记用于配置所

述VM的VS和所述VM的TORS以处理组播业务。

29. 如权利要求21所述的MCC,其中所述加入组模块配置成在VM的VS处安装流表条目以匹配包的目的互联网协议地址与所述组播组地址,以匹配包源地址与VM的互联网协议地址并且匹配源MAC地址与所述VM的MAC地址,并且将所述组播组的MPLS标记压入所述包并且将包复制到输出组。

30. 如权利要求21所述的MCC,其中所述加入组模块配置成在VM的VS安装流表条目以将所述包的源IP地址与所述VM的IP地址匹配以及匹配所述TORS的端口,然后放弃所述包以防止循环。

31. 如权利要求21所述的MCC,其中所述加入组模块配置成在所述VS的输入组中安装转发动作以将组播组业务转发到所述VM的互联网协议地址。

32. 如权利要求21所述的MCC,其中所述加入组模块检查所述VM是否是在所述TORS预订组播组的第一VM。

33. 如权利要求32所述的MCC,其中所述加入组模块配置成在所述TORS安装流表条目以将骨干路由标记压入到包中。

34. 如权利要求32所述的MCC,其中所述VM是所述第一VM,并且其中所述加入组模块配置成安装在所述TORS的流表条目以匹配组播标记和转发到目的TORS的输出端口。

35. 如权利要求32所述的MCC,其中所述VM是所述第一VM,并且其中所述加入组模块配置成安装在所述TORS的流表条目以转发组播组包到具有预订VM的其他TORS。

36. 如权利要求32所述的MCC,其中所述VM是所述第一VM,并且所述加入组模块配置成安装在所述TORS的流表条目以转发组播组包到具有预订所述组播组的VM的机架中的所有VS。

37. 如权利要求32所述的MCC,其中所述VM是所述第一VM,并且所述加入组模块配置成安装在所述TORS的流表条目以匹配包中的骨干路由标记,以解封所述骨干路由标记并且以转发所述包到所述TORS的下个表。

38. 如权利要求37所述的MCC,其中所述加入组模块配置成安装流表条目以匹配所述包中的组播组标记,以解封所述组播组标记,并且以经由关联VS使用输入组转发所述包到预订VM。

39. 如权利要求21所述的MCC,其中所述加入组模块配置成推断从组播地址的组播源的加入,以及安装在所述组播源的TORS处的至少一个流表条目,所述组播源的所述TORS处的至少一个流表条目配置所述TORS通过MPLS转发所述组播组的所述出站组播业务到具有预订所述组播组的VM的所述云系统的其他TORS。

40. 一种在云系统中实现来管理云网络中的组播业务的组播云控制器“MCC”(290),所述MCC耦合于用于托管一个或多个虚拟机“VM”的至少一个虚拟化服务器,其中所述虚拟化服务器包括至少一个虚拟交换机“VS”,所述虚拟交换机支持多协议标记交换MPLS,并且所述虚拟交换机耦合于支持MPLS的架顶式交换机“TORS”,所述MCC包括:

处理器,用于执行加入组模块,

所述加入组模块(425)用于管理组播组加入操作,其中所述加入组模块配置成将所述VM安装到租户的VM集,其中所述VM集预订所述组播组,所述VM集在组播组管理表中连同组播组标记跟踪,

配置成使用所述MCC维持的VM记录查找所述VM的VS和TORS,配置成在所述VS的输入组中安装转发动作以将组播组业务转发到所述VM的互联网协议地址,以及检查所述VM是否是在所述TORS预订组播组的第一VM,在所述VM是所述第一VM的情况下,配置成在所述TORS处安装流表条目以将骨干路由标记压入包上,配置成在所述TORS处安装流表条目以匹配组播标记并且转发所述包到目的TORS的输出端口,配置成在其他TORS处安装流表条目以将组播组包转发到具有预订VM的其他TORS,配置成在所述TORS处安装流表条目以将组播组包转发到具有预订组播组的VM的机架中的所有VS,配置成在所述TORS处安装流表条目以匹配包中的骨干路由标记,以解封所述骨干路由标记并转发所述包到所述TORS的下个表,以及安装流表条目以与所述包中的组播组标记匹配,以对所述组播组标记解封并且经由关联VS使用输入组将所述包转发到预订VM。

用于管理云网络中的组播业务的方法和装置

技术领域

[0001] 本发明的实施例涉及云计算领域;并且更具体地,涉及云计算中的组播支持。

背景技术

[0002] 大型公司很多年来一直专注于它们在数据中心中的计算资源。该趋势随着服务器虚拟化技术变得越来越普遍而在过去几年加速。因为数据中心变得更大,一些数据中心运营商开始向外部客户提供计算、存储和网络通信资源,这些客户然后变成数据中心的租户。提供的服务典型地由弹性、按需处理存储组成,其对于多数实用目的而言仅受到客户支付能力和进入互联网的网络带宽的限制。该开发叫作云计算。

[0003] 服务器虚拟化技术允许将服务器池作为基本上一个大的计算机资源来管理。叫作管理程序的软件层位于操作系统与硬件之间。该管理程序调度虚拟机(VM)在虚拟化服务器上的执行。VM是封装有一些应用的操作系统图像。管理程序允许暂停VM并且使其在服务器之间移动到负载平衡。用于捕捉崩溃的VM执行的负载平衡和监视对用高得多的专业解决方案成本实现的企业应用提供相同种类的容错和可标度性服务。这些系统在本文称为“云系统”。云管理器系统监管云系统中VM的执行;调度执行来满足需求、优化服务器利用并且使功耗最小化。云执行管理器可以调度执行以允许硬件和服务中升级而不影响正在进行中的服务预备。

[0004] 为了支持VM在机器之间的任意移动,云系统内的联网也必须虚拟化。大部分云系统现今通过使虚拟交换机并入管理程序而使网络虚拟化。虚拟交换机向在管理程序的控制下执行的VM提供虚拟网络端口。虚拟交换机软件还允许网络资源采用与服务器资源如何被管理程序虚拟化相似的方式虚拟化。管理程序和虚拟交换机由此可以合作以允许VM在服务器之间移动。在管理程序使VM移动时,它关于新的位点来与虚拟交换机通信,并且虚拟交换机确保对于VM地址(层2媒体访问控制(MAC)地址,和层3互联网协议(IP)地址)的网络路由表被升级,因此包被路由到新的位点。

[0005] 组播是由源节点将数据交付给一组预订节点(例如,VM)的系统。由云系统内的VM提供的一些服务可以从数据中心内组播协议的实现获益。然而,这样的组播协议支持不被云系统所支持,因为这些组播协议需要维持超出主机容量的大量控制状态并且将使数据中心和所有云服务的性能下降。

发明内容

[0006] 由云系统中的组播云控制器(“MCC”)执行来管理云网络中的组播业务的过程。该MCC耦合于用于托管一个或多个虚拟机(“VM”)的至少一个虚拟化服务器,其中该虚拟化服务器包括至少一个虚拟交换机(“VS”),该虚拟交换机支持多协议标记交换MPLS,并且该虚拟交换机耦合于支持MPLS的架顶式交换机(“TORS”),管理组播业务的过程包括以下步骤:在该MCC接收加入消息,其中该加入消息指示VM请求了加入组播组;将该VM加入到租户的VM集,其中该VM集预定该组播组;通过该MCC使用该MCC维持的VM记录查找对于该VM的VS和

TORS;在对于该VM的VS处安装至少一个流表条目,该至少一个流表条目配置该组播组的进站组播业务转发到该VM并且配置该组播组的出站组播业务发送到该VM的TORS;以及在该VM的TORS处安装至少一个流表条目,该TORS的至少一个流表条目配置该TORS为通过MPLS转发该组播组的出站组播业务到具有预定该组播组的VM的该云系统的其他TORS并且配置该TORS转发该组播组的进入组播业务到该VM的VS。

[0007] 由云系统中的组播云控制器(“MCC”)执行该方法来管理云网络中的组播业务。该MCC耦合于用于托管一个或多个虚拟机(“VM”)的至少一个虚拟化服务器,其中该虚拟化服务器包括至少一个虚拟交换机(“VS”),其支持多协议标记交换MPLS,并且该虚拟交换机耦合于支持MPLS的架顶式交换机(“TORS”)。在该MCC接收以互联网组管理协议报告消息或组播侦听器发现报告消息的形式的加入消息,其中该加入消息指示VM请求了加入组播组;将该VM加入到租户的VM集,其中该VM集被预定到该组播组,该VM集在组播组管理表中连同组播组标记跟踪;通过该MCC使用该MCC维持的VM记录查找对于该VM的VS和TORS;在该VS的输入组中安装转发动作以将组播组业务转发到该VM的互联网协议地址;以及检查该VM是否是第一VM以在TORS预定组播组,其中该VM是该第一VM,在TORS处安装流表条目以将骨干路由标记压入包上,在TORS处安装流表条目以匹配组播标记并且转发该包到目的TORS的输出端口,在其他TORS处安装流表条目以将组播组包转发到具有预定VM的其他TORS,在该TORS处安装流表条目以将组播组包转发到具有预订组播组的VM的机架中的所有VS,在该TORS处安装流表条目以匹配包中的骨干路由标记,解封该骨干路由标记并转发该包到该TORS的下个表,以及安装流表条目以与该包中的组播组标记匹配,对该组播组标记解封并且经由关联VS使用输入组将该包转发到预订VM。

[0008] 在云系统中实现的组播云控制器(“MCC”),该MCC耦合于用于托管一个或多个虚拟机(“VM”)的至少一个虚拟化服务器,其中该虚拟化服务器包括至少一个虚拟交换机(“VS”),其支持多协议标记交换MPLS,并且该虚拟交换机耦合于支持MPLS的架顶式交换机(“TORS”),该MCC包括:处理器,用于执行加入组模块,该加入组模块管理组播组加入操作,其中该加入组模块配置成在该MCC接收以互联网组管理协议报告消息或组播侦听器发现报告消息的形式的加入消息,其中该加入消息指示VM请求了加入组播组,该加入组模块还配置成将该VM添加到租户的VM集,其中该VM集被预定到该组播组,配置成通过MCC使用该MCC维持的VM记录查找该VM的VS和TORS,配置成安装在该VM的VS的至少一个流表条目,该至少一个流表条目配置该组播组的进站组播业务为转发到该VM并且配置该组播组的出站组播业务被发送到该VM的TORS,并且配置成在该VM的TORS处安装至少一个流表条目,该TORS的该至少一个流表条目配置该TORS为通过MPLS转发该组播组的出站组播业务到具有预定该组播组的VM的云系统的其他TORS并且配置该TORS为转发进入该组播组的组播业务到该VM的该VS。

[0009] 在云系统中实现来管理云网络中的组播业务的组播云控制器(“MCC”),该MCC耦合于用于托管一个或多个虚拟机(“VM”)的至少一个虚拟化服务器,其中该虚拟化服务器包括至少一个虚拟交换机“VS”,该虚拟交换机支持多协议标记交换MPLS,并且该虚拟交换机耦合于支持MPLS的架顶式交换机(“TORS”),该MCC包括:处理器,用于执行加入组模块,该加入组模块用于管理组播组加入操作,其中该加入组模块配置成将该VM安装到租户的VM集,其中该VM集被预定到该组播组,该VM集在组播组管理表中连同组播组标记跟踪,配置成使用

该MCC维持的VM记录查找该VM的VS和TORS,配置成在该VS的输入组中安装转发动作以将组播组业务转发到该VM的互联网协议地址,以及检查该VM是否是第一VM以在该TORS预定组播组,其中该VM是该第一VM,配置成在该TORS处安装流表条目以将骨干路由标记压入包上,配置成在该TORS处安装流表条目以匹配组播标记并且转发该包到目的TORS的输出端口,配置成在其他TORS处安装流表条目以将组播组包转发到具有预定VM的其他TORS,配置成在该TORS处安装流表条目以将组播组包转发到具有预订组播组的VM的机架中的所有VS,配置成在该TORS处安装流表条目以匹配包中的骨干路由标记,以解封该骨干路由标记并转发该包到该TORS的下个表,以及安装流表条目以与该包中的组播组标记匹配,以对该组播组标记解封并且经由关联VS使用输入组将该包转发到预订VM。

附图说明

[0010] 本发明通过参考下列描述和附图而最好地理解,这些附图用于图示本发明的实施例。在图中:

[0011] 图1A和1B是标记栈的图。

[0012] 图2是云系统中的部件的一个实施例的图;

[0013] 图3A-C是由组播控制器(MCC)维持的数据结构的一个实施例的图。

[0014] 图4是云系统中的部件的另一个实施例的图。

[0015] 图5是用于激活云系统中的虚拟机的过程的一个实施例的流程图。

[0016] 图6是用于激活虚拟机的过程的一个实施例的图。

[0017] 图7是用于拦截来自VM的DHCP请求的过程的一个实施例的流程图。

[0018] 图8是用于拦截来自虚拟机的DHCP请求的过程的一个实施例的图。

[0019] 图9是用于拦截来自VM的ARP请求的过程的一个实施例的流程图。

[0020] 图10是用于拦截来自虚拟机的ARP请求的过程的一个实施例的图。

[0021] 图11是在云计算网络中使VM从第一虚拟化服务器移到第二虚拟化服务器的过程的一个实施例的图。

[0022] 图12是在云系统中对于VM预订组播的过程的一个实施例的流程图。

[0023] 图13是在云系统(其中配置虚拟交换机)中对于VM预订组播组的过程的另一个实施例的流程图。

[0024] 图14是配置成在云系统中支持预订组播组的VM的流表的一个实施例的图。

[0025] 图15是在云系统(其中配置架顶式交换机)中对于VM预订组播组的过程的另一个实施例的流程图。

[0026] 图16是配置成在云系统中支持预订组播组的VM的架顶式交换机的流表的一个实施例的图。

[0027] 图17是用于在云系统中支持离开组播组的VM的过程的一个实施例的流程图。

[0028] 图18是用于在云系统中支持VM在虚拟服务器之间的移动的过程的一个实施例的流程图。

[0029] 图19图示根据本发明的实施例的示范性计算机系统的框图。

具体实施方式

[0030] 下列说明描述用于管理多协议标记交换 (MPLS) 流表以在云网络中支持组播的方法和装置。在下列说明中, 阐述许多具体细节, 例如逻辑实现、操作码、规定操作码的工具、资源划分/共享/复制实现、系统部件的类型和相互关系和逻辑划分/集成选择, 以便提供对本发明的更全面理解。然而, 本领域内技术人员将意识到可在没有这样的具体细节的情况下实践本发明。在其他实例中, 为了不使本描述难以理解, 未详细地示出众所周知的控制结构、门级电路和全软件指令序列。本领域内普通技术人员利用包括的说明将能够实现适当的功能性而没有过度实验。

[0031] 在说明书中对“一个实施例”、“实施例”、“示例实施例”等的引用指示描述的实施例可包括特定特征、结构或特性, 但每个实施例可不一定包括该特定特征、结构或特性。此外, 这样的短语不一定都指相同的实施例。此外, 当特定特征、结构或特性连同实施例描述时, 认为连同其他无论是否明确描述的实施例实现这样的特征、结构或特性, 这在本领域内技术人员的知识内。

[0032] 在下列说明和权利要求中, 可使用术语“耦合”和“连接”连同它们的派生词。应理解这些术语不规定为是彼此的同义词。“耦合”用于指示彼此可或可不直接物理或电接触的两个或以上的元件彼此共同操作或相互作用。“连接”用于指示彼此耦合的两个或以上的元件之间的通信的建立。

[0033] 为了便于理解, 在图中使用虚线来表示某些项的可选性质 (例如, 本发明的指定实现不支持的特征; 指定实现支持但在一些情况下而不是其他情况下使用的特征)。

[0034] 在图中示出的技术可以使用在一个或多个电子设备上存储并且执行的代码和数据来实现。这样的电子设备使用非暂时性有形计算机可读介质 (例如, 磁盘; 光盘; 只读存储器; 闪存存储器设备; 相变存储器) 和暂时性计算机可读通信介质 (例如, 电、光、声或其他形式的传播信号-例如载波、红外信号、数字信号, 等) 来存储和传达 (在内部和/或通过网络利用其他电子设备) 代码和数据。另外, 这样的电子设备典型地包括一个或多个处理器集, 其耦合于一个或多个其他部件, 例如存储设备、一个或多个输入/输出设备 (例如, 键盘、触屏和/或显示器) 和网络连接。处理器集和其他部件的耦合典型地通过一个或多个总线和网桥 (也称作总线控制器)。承载网络业务的存储设备和信号分别代表一个或多个非暂时性有形计算机可读介质和暂时性计算机可读通信介质。从而, 指定电子设备的存储设备典型地存储代码和/或数据以供在该电子设备的一个或多个处理器集上执行。当然, 本发明的实施例的一个或多个部分可使用软件、固件和/或硬件的不同组合来实现。

[0035] 概览

[0036] 如在下文描述实施例提供使用多协议标记切换 (MPLS) 服务来支持互联网协议 (IP) 组播的方法和装置。特别地, 实施例使用 MPLS 服务以在数据中心中支持 IP 组播来支持云计算服务。

[0037] 数据中心通常组织为执行虚拟化服务的物理计算设备集。这些物理计算设备可以组织成可以独立或组合地执行虚拟化服务的设备机架或组。物理计算设备可以通过架顶式交换机 (TORS) 集而与机架中的其他物理设备并且与其他机架中的物理设备联网。这些 TORS 连接到它们机架中的每个物理设备并且然后连接到彼此或数据中心中的 TORS 子集。TORS 可以在网状拓扑中与数据中心的 TORS 联网。该网状或网络结构可以包括任何数量的中间联网设备, 用于连接 TORS 和其他设备。TORS 还可以直接或间接与数据中心外部的设备连接。

该数据中心架构进一步在下文更详细描述。

[0038] 充当云计算设施的许多数据中心仅支持Web服务应用。Web服务应用由负载平衡前端组成,其向Web服务器池派发请求。这些请求在概念上源自互联网上的应用并且因此安全性和隐私要求比私人企业网中的应用宽松得多。另一个云计算趋势是安全多租户,其中云服务提供商提供虚拟私有网 (VPN),像在云外部的租户的分布式办公网络与云内的VPN之间的连接。这允许在云内的租户应用在与企业广域网 (WAN) 类似的网络环境中操作。对于其中这些服务仅提供给拥有数据中心的企业内的客户的私有数据中心,对于多租户的安全性和隐私要求放宽。对于公共数据中心,云服务运营商必须确保来自多个租户的业务与其他租户隔离并且没有业务从一个租户到达另一个的可能性。在任一情况下,云计算设施趋于使用媒体访问控制 (MAC) 层虚拟局域网 (VLAN) 来实现云计算网络。

[0039] 例如,可以对两个不同外部企业客户设置两个虚拟私有云 (VPC)。VPC由VM的集合、存储和向租用云中的空间的企业提供安全多租户的联网资源组成。企业客户经由VPN通过在公共运营商网络上运行的互联网而连接到VPC内。

[0040] 为了向VPC添加新的服务实例(新的虚拟机“VM”),云执行管理器使VM初始化以在虚拟化服务器上的管理程序上运行。虚拟化服务器上的虚拟交换机配置成在VLAN中包括VM,该VLAN是对于添加新VM的租户的VPN的部分。在一些情况下,对于新的服务升级虚拟客户边缘路由器并且用新的服务升级云计算设施中的供应商边缘路由器。

[0041] 为了提供VPN,云计算设施实现三个解决方案中的一个。在第一可能解决方案中,每个租户接收独立VLAN片 (VLAN slice)。在第二可能解决方案中,租户VPN使用IP封装来实现。在第三可能解决方案中,租户VPN使用MAC地址封装来实现。这些解决方案中的每个面临不同的缺点。如果云系统实现VLAN隔离,则每个租户被分配独立VLAN标签并且云网络作为平面层2网络运行。VLAN标签具有12个位,因此如果云运营商使用VLAN隔离,租户的数量限制在4096。对可能租户数量的该限制是对云系统可标度性的主要限制。

[0042] VLAN隔离的另一个问题是标准区域联网(例如,LAN、WAN、MAN;电气和电子工程师协会(“IEEE”)802.1)交换使用生成树协议(STP)来设置路由。为了去除路由环的可能性,STP在源地址与目的地址之间指定一个并且仅仅一个路径,而不管是否存在多个路由。这在生成树协议遭受业务压力并且忽略备选路由时可以导致交换结构的拥挤和利用不足。

[0043] 如果云系统实现IP封装,则云系统作为路由IP网络运行并且IP隧道用于隔离每个租户的业务。来自租户的业务封装在IP包中(典型地使用通用路由封装(GRE),其中隧道端点是VM运行所在的源和目的虚拟化服务器上的源与目的虚拟交换机。

[0044] IP封装允许客户端通过将以太网帧封装在层2隧道协议(L2TP)中而定义在顶部的任意层2服务。它还允许有大量租户,其仅受云范围内IP地址空间的约束。租户还可以在IP隧道顶部部署它们自己的IP地址空间。然而,在没有其他度量的情况下,IP路由还选择单个路由并且因此忽略多径路由,从而导致不必要的拥挤。云路由可以利用等价多径以在多个链路上散布包但以额外配置复杂性为代价。一般,配置IP路由网络是耗时的并且路由器趋于成为比简单交换机更昂贵的设备。另外,IP网络具有有限手段来预备专用带宽,这对于大的数据流可是必需的。

[0045] 如果云系统实现MAC封装,则独立租户VPN通过封装在MAC隧道内部而实现,这与IP封装相似。MAC隧道的端点典型地是VM运行所在的源和目的虚拟化服务器上的虚拟交换机。

MAC封装提供与通过VLAN隔离的IP封装相似的益处,只是增加了云可以作为平面层2网络运行这一益处(如期望的话)。缺点是对于MAC封装存在很少对于信令协议的标准(与IP不同),并且尽管存在对于数据平面封装的标准,不在一些现有云管理软件产品中使用它们。这要冒着不与某些类型的硬件一起工作的风险。配置并且维持MAC封装网络也比维持VLAN隔离网络更复杂。

[0046] 在本发明的实施例中,云系统使用MPLS来形成对于VPC租户的VPN,并且对在虚拟交换机和TORS中的流路由编程。例如,MPLS交换机可以经由如在2011年2月的OpenFlow交换机规范版本1.1.0中描述的OpenFlow对流路由编程。虚拟交换机可以使用以太网MAC地址来将包路由到虚拟化服务器上的VM,并且将包分类成特定MPLS标记交换路径(LSP),但在虚拟交换后,以太网和IP地址都未用于路由。相反,路由使用基于IP和MAC目的地址而分配的MPLS标记严格进行。过程与广域网中的虚拟专线服务(VPWS)或层3 VPN(L3VPN)相似但不相同。

[0047] 在一个实施例中,由OpenFlow控制器执行的云网络管理器(CNM)使用OpenFlow将流路由安装到VM运行所在的虚拟化服务器上的虚拟交换机内。流路由将来自虚拟化服务器上的租户VM的以太网帧与指示在TORS与目的方上的虚拟交换机之间的路由并且对其之间的租户特定的MPLS标记一起封装。包然后转发到源VM的虚拟化服务器的TORS。源TORS然后将包路由到数据中心网络结构内。如果云网络结构是MPLS网络,则源TORS压入另一个MPLS标记,其规定源TORS于目的TORS之间的路由。目的TORS弹出MPLS路由标记并且使用由源虚拟交换机推上的标记来将包路由到合适的目的服务器和虚拟交换机。目的虚拟交换机弹出TORS到虚拟交换机标记并且基于目的MAC地址将包转发到正确的VM。如果云交换机结构支持某一其他技术,例如光,TORS进行适合的适配。

[0048] MPLS在云系统中的该使用提供许多优势。例如,在数据平面上不需要IP路由,由此去除具有IP路由云网络的成本、电力使用和复杂性。云架构仅需要廉价的MPLS交换机。云系统中的交换机结构可以是标准以太网交换机、有MPLS能力的交换机、光电路交换机或甚至混合物,例如对高容量流的光和对于其他业务的电交换机。交换机结构甚至可以是分布式计算资源之间的广域网。TORS适当地处理封装。可以支持的租户的数量不限于MPLS标记空间的大小。租户可以管理它自己的IP地址空间(如期望的话)和DNS服务器(如期望的话)。到云交换机结构内的路由细节由TORS处理,其中可以实现可标度性和聚合的正确混合。

[0049] 对在数据平面上使用MPLS并且在控制平面上使用OpenFlow的VPC中的VM之间的IP单播使用VPN的云系统中该基于MPLS的路由顶部,实施例提供在基于MPLS的VPC网络中处理组播的方法。因为云系统在数据平面中使用MPLS,云系统不需要MPLS组播的控制平面复杂性,其用别的方式在分布式控制平面网络中是必需的。避免该开销从可标度性和带宽利用方面提供优势使得组播可以实际上且有效地在云系统中实现。

[0050] 在高级别处,实施例如下工作。虚拟交换机(VS)将对于IPv4的互联网组管理协议(IGMP)包或对于IPv6的组播发现侦听器(MLD)包从VM转发到组播云控制器(MCC)。MCC向组播组添加VM并且在源侧VS和将组播包从VM路由到组播组的TORS上安装路由。对每个组播组分配MPLS标记。如果这是机架上要加入组的第一VM,MCC还在服务于机架(具有组中的VM)的其他TORS上安装路由,从而将对于组的包路由到新的TORS。

[0051] 该组播实现的优势包括在数据平面上不需要有IP路由,从而去除具有IP路由云网

络的成本、电力使用和复杂性。实施例可以使用廉价的MPLS交换机来实现。云系统中TORS之间的网络结构可以是标准以太网交换机、有MPLS能力的交换机、光电路交换机或甚至混合物,例如对高容量流的光和对于其他业务的电交换机。网络结构甚至可以是分布式计算资源之间的广域网。TORS适当地处理封装。到云网络结构内的路由细节由TORS处理,其中可以实现可标度性和聚合的正确混合。

[0052] 如在上文在一定程度上论述的,现有解决方案存在多种问题。在现今的多租户云系统中,云系统运营商因为它在路由器中需要太多控制状态并且干扰VM移动而未实现组播。相比之下,实施例从路由器去除控制状态并且将它放置在OpenFlow控制器内,这可以在具有大量存储器的服务器中运行。另外,因为数据中心网络是这么的规则(由架顶式交换机、聚合交换机和路由器以及虚拟交换机组成),实施例从根本上减少需要的控制状态量以便处理组播。实施例还正确处理VM移动。

[0053] 实施例以集中控制平面为特征,其中集中OpenFlow控制器直接对架顶式交换机和虚拟交换机编程。可标度性通过保持服务器上的所有控制平面状态来处理,在该服务器中存储器的数量可以是非常大的,但实现转发所必需的流表状态除外。OpenFlow控制器直接对交换机编程。因为实施例基于集中控制平面,实施例不需要在云系统内使用链路发现协议(LDP)或边界网关协议(BGP)。业务在入口侧上每TOR交换机复制,但仅发送到服务于机架(具有预订组播组的VM的TOR交换机,实施例是非常高效的。

[0054] 实施例具有唯一标记栈。建立组播所在的标准(RFC 4762) VPLS标记栈与图1A看起来相似。实施例的标记栈在图1B中示出。实施例与VPLS标记栈之间的重要区别是,在VPLS中,来自在不同VLAN上运行的不同主机但对于相同租户的组播包将具有不同的VLAN标记,而实施例对于相同租户的组播组利用相同标记。在实施例中,组播树标记邻近租户包的IP报头并且骨干网状标记在栈顶部。在一个实施例中,本文描述的MPLS组播因为标记栈与标准MPLS VPN标记栈不兼容而在站点边界路由器(端设备(premise equipment, PE)或客户设备(CE)路由器)处终止。

[0055] 云系统架构

[0056] 图2图示根据本发明的一个实施例的云系统中的部件。该云系统包括通过云网络270而耦合的多个机架200A-D中的多个虚拟化服务器。在这些虚拟化服务器和机架中,图示三个机架200A、200B和200C。每个图示的机架包括架顶式交换机(TORS)和多个虚拟化服务器。机架200A包括TORS 201A和多个虚拟化服务器205A-205N。机架200B包括TORS 201B和多个虚拟化服务器215A-215N。机架200C包括TORS 201C、虚拟化服务器225A和虚拟化服务器225B。每个TORS 200A-200C使该机架内的多个虚拟化服务器耦合并且进一步与云网络270耦合。云网络270的架构可以支持任何类型的交换或路由协议,例如光电路交换网络或OpenFlow MPLS交换机网络。云网络270甚至可以在分布式服务器机架之间运行的广域网(WAN)。图中的连接性通过示例而非限制的方式提供,存在许多方式来使云网络270中的TORS耦合,在本文也称为网络结构。

[0057] 在图2的示例中,云网络270包括多个交换元件(例如,交换机或路由器)280A-280D。每个TORS 200A-200C和每个虚拟化服务器205A-205N、215A-215N、225A和225B可以进一步与管理网络275耦合。该管理网络275使TORS和虚拟化服务器与管理元件耦合。管理元件包括云执行管理器250、虚拟盘255、组播云控制器(MCC) 290和云网络管理器260。

[0058] 云执行管理器 (CEM) 250 调度 VM 以供执行, 并且包括对象管理器, 用于将 VM 映射到虚拟盘 255 内和外, 这提供对于每个 VM 图像的永久存储。云网络管理器 (CNM) 260 对于租户的 VPC 来管理 VPN。

[0059] MCC 在 TORS 中配置流表条目用于实现组播并且跟踪实现组播所需要的状态数据。MCC 可以维持对于每个租户的组播组管理表 (MGMT)。MCC 还可以在每租户基础上跟踪对所有组播组的每一个租户的 MAC 地址到租户 ID 映射 (1:1)、租户 ID 到 MAC 地址映射 (1:m) 和 OpenFlow 组 ID。MCC 每租户包含一个 MGMT。租户通过租户 ID (Tid_n) 来识别, 其对于每一个组合是唯一的。在新的租户发起创建 VPC 的请求时, 也在 MCC 中创建对应条目。MGMT 包含对于组播组的记录。每个记录包含下列信息: 组的组播 IP 地址、对于组播组中包含的 VM 的 VM 记录集和对于组的 MPLS 路由标记。

[0060] 组播数据业务可以每租户分离并且 MCC 根据需要对每租户的组播地址创建 OpenFlow 组播组标识符 (ID) $IG_1 \cdots IG_n$ 。在一个实施例中, OpenFlow 表中的组播组 ID 是组播地址和租户 ID 的组合。组播组可以在租户请求虚拟私有云 (VPC) 时在 OpenFlow 表中预先创建或可以在对指定组播地址接收第一加入操作的运行时间创建。在租户加入组播组时, 由 MCC 维持的数据结构集用条目填充。这些数据结构在下文进一步论述。

[0061] 图 3A-C 是由 MCC 维持的数据结构中的一些的一个实施例的图。图 3A 图示租户跟踪结构 329, 其存储租户 ID 与关联 MAC 地址之间的相关性, 而每个 MAC 地址相似地与租户 ID 相关。图 3B 图示 MGMT 跟踪表 301, 其包含指向每个租户的每个 MGMT 303 的指针, 其中租户通过租户 ID (Tid_n) 来识别。MGMT 303 包括使租户 ID 305 与组播地址 (MG_n) 307、对于组播组的 MPLS 路由标记 (RL_{MG_n}) 309 和对于预订组播组的租户的每个 VM 的 VM 记录 311 的集相关的条目。

[0062] 图 3C 图示 VM 记录 325 (例如, 在 VM 记录 311 的集中) 另外包含指向 VS 的指针 (VS_{VM_n}), 其支持对于 VM 执行所在的机架的 VM (其包括 VS 的 MAC 和服务器标识) 和 TORS (其包括 TORS 的 MAC)。图 3A-C 的数据结构中的每个由 MCC 在 OpenFlow 控制器处维持, 其具有足够的工作存储器或相似的存储设备来将对于所有组播的数据结构存储在云系统内。

[0063] 图 4 是根据本发明的一个实施例在使用 MPLS 支持 VM 迁移和 VM 通信的云系统中的部件。图 4 示出 TORS 201A, 其包括一个或多个流条目表 212A-Z。这些流条目表 212A-Z 用于使 MPLS 标记与 MPLS 动作匹配, 如参考 OpenFlow 规范描述的。TORS 201A 与一个或多个虚拟化服务器 215A-N 耦合。

[0064] 每个虚拟化服务器 215A-N 配置成执行一个或多个 VM。在图 4 中, 虚拟化服务器 215A 包括管理程序 216A, 其调度 VM (例如, VM T1 和 VM T2) 的执行。虚拟化服务器 215A 进一步包括虚拟交换机 217A。该虚拟交换机 217A 向在管理程序 216A 的控制下执行的 VM 提供虚拟网络端口。虚拟交换机 217A 是启用 MPLS 的交换机, 其包括一个或多个流条目表 218A-Z, 用于使 MPLS 标记与 MPLS 动作匹配, 如参考 OpenFlow 规范描述的。

[0065] TORS 201A 与云网络 270 耦合, 例如在图 2 中描述的云网络。TORS 201A 进一步与管理网络 275 耦合。如在图 2 中的, 管理网络 275 使 TORS 和虚拟化服务器与例如 CEM 250、MCC 290、虚拟盘 255 和 CNM 260 等管理元件耦合。CNM 260 包括描述 VPN 的表, 其包括地址映射表 262 和标记映射表 264。地址映射表 262 在 VM 的 MAC 地址、运行 VM 的虚拟化服务器的 MAC 地址 (其是虚拟交换机 217A 的 MAC 地址) 和 VM 的 IP 地址之间映射。标记映射表 264 将虚拟化服务器 215A 的 MAC 地址 (其是虚拟交换机 217A 的 MAC 地址) 映射到对于虚拟交换机 217A 与 TORS 201A

之间的链接的TORS链接标记。在其他实施例中,CNM 260了解哪些动态主机配置协议(“DHCP”)和域名服务(“DNS”)(如有的话)服务器属于哪个租户。

[0066] 在一个实施例中,CNM假设租户VM关于它们如何配置它们的IP栈以及它们未任意改变它们的IP/MAC地址而表现良好。如果租户VM表现不好并且做出这样的改变,它的包将路由到CNM并且CNM将告知CEM。CEM然后可以关闭出错(offending)VM。

[0067] MCC 290包括MGMT 303、VM记录325和用于管理组播业务的其他数据结构,如上文描述的。MCC 290或CNM 260还可以在VS 217A中配置流表条目218A来将所有组播报告消息(例如互联网组播协议(IGMP)或组播发现侦听器(MLD)消息)转发到MCC 290。MCC 290或CNM 260还可以在流条目表212A或VS 217A流条目表218A中创建或删除流条目来管理组播在云系统中的实现,例如对组播组实行加入操作、离开组播组或VM迁移,如在下文进一步描述的。

[0068] 在一个实施例中,CNM 260、MCC 290、CEM 250和相似的部件可以在OpenFlow控制器450中实现。OpenFlow控制器450由此可以实现对于云系统网络结构的控制平面。TORS 201A和VS 215A的管理因此受到流表212A和218中流表条目的配置的影响。流表由如在OpenFlow规范中定义的OpenFlow流表更新消息的应用来维持。

[0069] VM激活

[0070] 图5图示根据本发明的一个实施例用于激活虚拟机的框图。在框500中,CNM接收为了在虚拟化服务器上激活而调度VM的通知。例如,CEM将CEM已经为了激活而调度VM通知CNM。在框510中,CNM确定VM的MAC地址。在一个实施例中,CEM对CNM提供VM的MAC地址连同通知,而在另一个实施例中,CEM在独立消息中对CNM提供VM的MAC地址。在框515中,CNM确定虚拟交换机的MAC地址,其与驻存在虚拟服务器(VM将在其他激活)上的虚拟交换机相对应。在框520中,CNM在CNM的地址映射表中记录VM的MAC地址和虚拟交换机的MAC地址之间的关联。在框525中,CNM进一步确定MPLS标记,其使虚拟交换机与TORS(其与虚拟交换机耦合)关联。在框530中,CNM向虚拟交换机发送流条目修改消息来指示与MPLS标记和VM的MAC地址匹配的数据包应转发到VM。在一个实施例中,在框532中,CNM向虚拟交换机发送流条目修改消息来指示源自VM的DHCP请求应转发到CNM。在另一个实施例中,在框533中,CNM向虚拟交换机发送流条目修改消息来指示源自VM的地址分辨协议(“ARP”)请求应转发到CNM。

[0071] 图6图示根据本发明的一个实施例用于激活虚拟机的方法。在图6中,TORS 201A与虚拟化服务器215A耦合。该虚拟化服务器215A包括管理程序216A和虚拟交换机217A。在图6中,虚拟化服务器215A具有两个活动VM T1和T2 490A和490B。此外,图6示出CEM 250、虚拟盘255和CNM 260。

[0072] 在图6中,CEM 250调度第三VM T3 490C以供在虚拟化服务器215A上执行。该图图示在虚拟机激活期间的各种点。在点A处,CEM 250调度VM T3 490C(例如属于租户1的VM)以供执行。对象文件通过CEM的对象管理器从虚拟盘255映射。CEM编辑VM 490C栈来将VM 490C的MAC地址设置成分配给租户1的MAC地址中的一个。在点B处,CEM 250将激活告知CNM 260。它对CNM 260提供VM 490C的MAC地址和VM执行所在的虚拟化服务器215A的MAC地址(其是虚拟化服务器215A的虚拟交换机217A的MAC地址)。

[0073] 在点C处,CNM 260在CNM的地址映射表(未示出)中记录VM 490C的MAC地址连同调度它所在的虚拟化服务器215A的MAC地址。在点D处,CNM 260对于VM 490C运行所在的虚拟

化服务器215A与TORS 201A之间的链接的TORS 201A MPLS标记(这也称为源TORS“STORS”链接标记)从CNM 260标记映射表来查找虚拟交换机。在点E处,CNM 260向启用MPLS的虚拟交换机217A发送流条目修改消息来安装使VM 490C的MAC和STORS链接标记与弹出STORS链接标记的动作匹配的规则,并且将包转发到VM 490C。

[0074] 在一个实施例中,在点F处,CNM 260向虚拟交换机217A发送第二流条目修改消息来安装下列规则。如果源MAC与对于VM 490C的MAC地址匹配并且协议是ARP,将包路由到CNM 260以允许CNM 260服务于ARP请求。该规则使ARP业务减少并且允许CNM 260从CNM地址映射表答复。

[0075] 在另一个实施例中,在点G处,CNM 260向虚拟交换机217A发送第三流条目修改消息来安装下列规则。如果源MAC与对于VM 490C的MAC地址匹配并且协议是DHCP,将包路由到CNM 260以允许CNM 260服务于DHCP请求。该规则允许CNM 260通过拦截来自DHCP服务器的应答包而在CNM地址映射表中记录MAC地址到IP地址映射。

[0076] 图6图示假设之前描述的全部三个流条目修改消息被传达的情况下虚拟交换机217A的流条目表218A中的三个条目。流条目表218A具有图示的五个列:源MAC、目的MAC、MPLS标记、IP协议和动作。流条目对应于490C的激活。

[0077] 图示的流条目表218A中的第一条目与作为源MAC地址的VM T3 490C的MAC地址匹配。目的MAC地址和MPLS标记设置成匹配任何条目,而IP协议设置成关于ARP协议匹配。动作则是转发到CNM 260。该规则与上文描述的步骤F相对应。

[0078] 图示的流条目表218A中的第二条目与作为源MAC地址的VM T3 490C的MAC地址匹配。目的MAC地址和MPLS标记设置成匹配任何条目,而IP协议设置成关于DHCP协议匹配。动作则是转发到CNM 260。该规则与上文描述的步骤G相对应。

[0079] 图示的流条目表218A中的第三条目与作为源MAC地址的所有MAC地址匹配。目的MAC地址设置成匹配VM T3 490C的MAC地址,而MPLS标记设置成TORS 201标记。IP协议设置成关于所有条目匹配。动作则是弹出MPLS标记并且将包转发到VM T3 490C。该规则与上文描述的步骤E相对应。

[0080] 如上文描述的,CNM维持VM MAC地址、虚拟化服务器MAC和VM IP地址之间的映射。如果租户选择不使用DHCP地址,则CEM在执行调度之前修改VM的IP栈并且安装静态IP地址。然而,在其他实施例中,租户将选择使用DHCP地址。在这样的情况下,CNM将拦截DHCP请求并且根据2001年1月的“DHCP中继代理信息选项”IETF、RFC 3046充当中继器。

[0081] DHCP请求

[0082] 图7图示根据本发明的一个实施例用于拦截来自VM的DHCP请求的方法的框图。在框700处,CNM接收源自VM的DHCP请求。在框705处,CNM根据RFC 3046将DHCP请求中继到DHCP服务器。在框710处,CNM接收来自DHCP服务器的DHCP应答。在框715处,CNM在地址映射表中记录VM的MAC地址与在DHCP答复中接收的IP地址之间的关联。根据本发明的实施例,该关联进一步包括执行VM的虚拟服务器的MAC地址。在框720处,CNM将DHCP答复转发到VM。

[0083] 图8图示根据本发明的一个实施例用于拦截来自虚拟机的DHCP请求的方法。在图8中,虚拟化服务器215A包括管理程序216A和虚拟交换机217A。在图8中,虚拟化服务器215A具有三个活动VM T1、T2和T3 490A-490C。虚拟交换机217A包括流条目表218A,其包括将DHCP请求从VM T3 490C转发到CNM 660(其与DHCP服务器630耦合)的规则。

[0084] 该图图示在拦截来自虚拟机的DHCP请求期间的各种点。在点A处,VM T3 490C发出DHCP请求来获得IP地址和其他配置信息,例如DNS服务器。在点B处,虚拟交换机217A使请求与流规则匹配以将DHCP请求路由到CNM 660。在点C处,CNM 260充当DHCP中继器、插入标准DHCP中继选项并且将请求转发到租户的DHCP服务器630。CNM 260可以通过检查DHCP请求的源MAC地址来确定VM T3 490C属于哪个租户。在点D处,租户的DHCP服务器230用包含对于VM T3 490C的IP地址和其他配置信息的DHCP应答作出响应。在点E处,CNM 260在CNM的地址映射表262中记录VM T3 490C MAC地址和IP地址之间的映射。在点F处,CNM 260去掉DHCP中继选项并且将DHCP应答转发到虚拟交换机217A,其在点G处将请求转发回到VM T3 490C。

[0085] 在典型的IP LAN场景中,在服务器中运行的希望与目的应用通信的源应用使用DNS来发现目的应用的IP地址。在找到IP地址后,源服务器中的IP栈广播ARP消息来发现IP地址与目的MAC地址之间的映射。如果目的地不在本地LAN上,它对ARP做出应答。如果目的地不在本地LAN上运行,源服务器IP栈使用本地路由器的MAC地址作为到其他子网的网关。

[0086] 在云中,VM在虚拟化LAN顶部运行。租户运行它自己的DNS服务器,或从云运营商购买DNS服务。VM用它要使用的DNS服务器的地址来配置。在租户希望联系服务时,它向DNS服务器发送DNS请求来发现IP地址,正如在正常企业LAN中的一样。DNS服务器可以在云运营商的控制下或它可以由个体租户控制。在DNS应答返回时,VM广播ARP请求来确定IP到MAC地址映射。

[0087] 在本发明的实施例中,虚拟交换机拦截ARP消息并且对于CNM来应答它而不是广播它。CNM提供地址映射服务使得ARP消息不必在整个云网络中广播。

[0088] ARP请求

[0089] 图9图示根据本发明的一个实施例用于拦截来自VM的ARP请求的方法的框图。在框900处,CNM接收源自VM的ARP请求。在框905处,CNM从ARP请求确定目的IP地址。在框910处,CNM确定对应于IP地址的目的VM MAC地址和目的虚拟交换机MAC地址。例如,CNM可以基于ARP请求内的IP地址从CNM的地址映射表查找映射。在框915处,CNM确定目的MPLS标记,其与目的虚拟交换机MAC地址相对应。在框920处,CNM向执行源VM的虚拟服务器上的虚拟交换机发送流条目修改消息。该条目修改消息指示与源VM MAC地址和目的VM IP地址匹配的包应推上租户id标记并且将其转发到对于源虚拟交换机的TORS。在框925处,CNM生成ARP应答,其指示目的VM MAC地址对应于目的IP地址。

[0090] 在框930中,CNM将ARP应答转发到源VM。

[0091] 图10图示根据本发明的一个实施例用于拦截来自虚拟机的ARP请求的方法。在图10中,虚拟化服务器215A包括管理程序216A和虚拟交换机217A。在图10中,虚拟化服务器215A具有活动的VM T3 490C。虚拟交换机217A包括流条目表218A,其包括将ARP请求从VM T3 490C转发到CNM 260的规则。虚拟化服务器216A与TORS 201(其通过云网络870耦合于另一个TORS 201C)耦合。TORS 201C与虚拟化服务器225B耦合,该虚拟化服务器225B包括管理程序226B和虚拟交换机227B。虚拟化服务器225B执行VM T4 295B。

[0092] 该图图示在拦截来自VM T3 490C的ARP请求期间的各种点。在点A处,VM T3 490C广播ARP请求。其余的拦截假设ARP请求指示驻存在虚拟服务器225B上的目的VM T4 495B的IP地址。在点B处,虚拟交换机217A拦截ARP请求并且使请求与流条目表218A中的将ARP请求路由到CNM 260的规则匹配。

[0093] 在点C处,CNM 260使用请求的目的IP地址(“DIP”)作为在CNM的地址映射表262中查找目的MAC地址(“DMAC”)的键。如果DIP不存在,目的地不在云内并且CNM未对ARP请求做出应答。VM然后将使用提供的本地路由器的MAC地址。如果在CNM的地址映射表262中存在DIP地址,则CNM 260检索DMAC VM地址和DMAC虚拟化服务器地址。在该情况下,CNM将检索VM T4 295B的MAC作为DMAC VM地址并且检索虚拟交换机227B MAC作为DMAC虚拟化服务器地址。

[0094] 在点D处,CNM对于目的TORS与DMAC虚拟化服务器地址之间的链接从CNM的标记映射表264查找MPLS流标记,也称为目的TORS(“DTORS”)链接标记。该流标记描述在DMAC虚拟化服务器处TORS 201A与虚拟化服务器之间的路由。在该情况下,将检索标记TORS:VS 201C:227B LL来指示从TORS 201A到TORS 201C的路由。

[0095] 在点E处,CNM 260在虚拟交换机217A的流条目表218A中安装流路由。规则匹配对于VM T4 495B的DMAC VM地址。动作将DTORS链接标记TORS:VS 201C:227B LL推到由VM T3 290C发送到VM T4 895B的业务包上,然后将包转发到TORS 201A。在点F处,CNM 260将ARP应答返回虚拟交换机217A,其将应答转发到VM T3 490C。

[0096] 在一个实施例中,CNM 260在将ARP应答返回VM之前将TORS-VS规则和动作安装到TORS 201的流条目表(未示出)。如果云网络270是MPLS网络,对于云网络270交换机结构内的TORS之间的LSP的规则不必在返回ARP应答时安装。如果云网络270交换机另外支持OpenFlow,CNM 260在安装新的TORS或现有TORS退役的时候使用OpenFlow仅需要将MPLS转发规则安装到较高级云网络270交换机内。然而,如果较高级数据中心交换机不支持OpenFlow,TORS相反使用标记分配协议(LDP)来设置它自己与另一TORS之间的LSP,并且应配置这些LSP。云网络270交换机不必支持支持IP路由协议,它们也不必进行路由。

[0097] 表1图示根据本发明的实施例对于应对路由到目的TORS的业务(即,传出业务和当前机架中目的服务器的业务)的流表的部分的TORS的OpenFlow流表的规则结构。稍后进一步解释对于进入业务的规则结构。在表中存在对于当前在使用中的作为机架中的源VM的目的地的每个DTORS(目的TORS)-DVS(目的虚拟化服务器)链接标记的流规则。TORS可另外高速缓存一些流规则以在VM移动时捕捉入站业务。来自机架中的源VM的业务通过匹配MPLS DTORS链接标记(其被虚拟交换机压入)而路由,并且与规则关联的动作被激活并且应用于包。动作根据云网络的性质而不同。

链接标记	其他字段	动作
TORS:VS 201C:227B LL	...	云网络特定动作
TORS:VS 201A:217A LL	...	云网络特定动作

[0098] 表1

[0099] 如果云网络不是MPLS网络,TORS通过OpenFlow虚拟端口使业务适应于云网络。与关于流匹配规则的动作关联的虚拟端口处理包并且添加必需的任何报头来将包路由到云网络内。例如,如果网络是光电路交换网络,虚拟端口可通过光交叉连接来路由它。本领域内技术人员将理解其他类型的云网络将利用其他适配虚拟端口。

[0100] 然而,如果云网络是MPLS网络,则示范性动作在表2示出。每个动作压入STORS到DTORS路由标记,然后向OpenFlow SELECT组发送包。SELECT组使用等价多径(ECMP)或某一其他负载平衡算法以从输出端口桶列表选择一个输出端口桶并且将包转发到输出端口。负

载平衡有助于避免拥挤并且促进两个TORS之间的所有路由的多径利用。

链接标记	其他字段	动作
TORS:VS 201C:227B LL	...	压入S->D标记;路由标记
TORS:VS 201A:217A LL	...	压入S->D标记;路由标记

[0101] 表2。

[0102] 一旦包到达目的TORS,它通过例如在表3和4中的示出等OpenFlow流表而路由。假设包在目的TORS处具有“LABEL-N”标记,规则与指示包去往该TORS的DTORS路由标记“LABEL-N”匹配。动作弹出DTORS路由标记并且将包发送到表4用于处理。表4规则与对于机架中的服务器的TORS-VS标记匹配,从而指示包去往哪个虚拟化服务器。动作将包转发出合适的端口用于到达该虚拟化服务器。

链接标记	其他字段	动作
LABEL-N	...	压入S->D标记;路由标记
LABEL-X	...	压入S->D标记;路由标记

[0103] 表3。

链接标记	其他字段	动作
TORS:VS 201C:227B LL	...	将包路由到虚拟交换机227B
TORS:VS 201A:217A LL	...	压入S->D标记;路由标记

[0104] 表4。

[0105] 在虚拟交换机227B处,触发使VM T4 227B的MAC地址和VS-TORS路由标记匹配的规则。VS-TORS路由标记被弹出并且包转发到VM T4 427B。

[0106] 云联网的最困难方面之一是应对VM从第一虚拟化服务器到第二虚拟化服务器的移动。CEM可以在任何时候使VM从一个服务器移到另一个。可进行移动来优化服务器利用、减少功耗或热生成,或将VM定位成更接近外部资源,例如数据库。在出现这样的移动时,到旧服务器处的VM的路由不再有效。在VM移动时可以放弃“飞行中”的包并且新的包可未到达VM的新位点,除非采取措施来修正路由。

[0107] 基本VM迁移

[0108] 图11图示在云计算网络中使VM从第一虚拟化服务器移到第二虚拟化服务器的方法。在图11中,第一虚拟化服务器215A(也称为旧虚拟化服务器)与TORS 201A(其耦合于云网络270)耦合。云网络270使TORS 201A与另一个TORS 201C耦合,该TORS 201C耦合于第二虚拟化服务器225A(也称为新虚拟化服务器)。

[0109] 每个虚拟化服务器215A和225A配置成运行多个VM并且包括对应的管理程序216A和226A和启用MPLS的虚拟交换机217A和227A。每个虚拟交换机217A和227A包括流条目表218A和228A,其包括MPLS规则来指示对应的虚拟交换机应如何处理它接收的包。用于VM移动的特定条目将在下文详细论述。在图示图9之前,旧虚拟化服务器915A具有三个活动的VM:VM T1 490A、VM T2 490B和VM T3 490C。图9进一步图示通过管理网络975而与CNM 260耦合的CEM 250。管理网络进一步使CEM 250和CEM 260与TORS 215A和225A耦合,由此提供到虚拟化服务器215A和225A的通信路径。

[0110] 该图图示在VM T3 490C从旧虚拟化服务器215A移到新虚拟化服务器225A期间的各种点。术语旧虚拟化服务器和新虚拟化服务器反映在移动之前VM T3 490C在旧虚拟化服

务器处是活动的并且在移动后VM T3 990C将在新虚拟化服务器225A处是活动的这一事实。

[0111] 在点A处,CEM 250开始使VM T3 490C从旧虚拟化服务器215A移到新虚拟化服务器225A。点A指示消息从CEM 250传送到旧虚拟化服务器215A,从而指示发起VM T3 490C的移动。只要VM T3 490C在旧虚拟化服务器215A上是活动的,则VM T3 490C可以继续服务于业务并且旧流规则在流条目表218A中维持。管理程序216A通过保存VM T3 290C的状态信息并且指示VM T3 290C进入待命或非活动状态而使VM T3 490C准备传输到新虚拟化服务器225A。在点B处,VM T3 490C的状态信息从旧虚拟化服务器215A传输到新虚拟化服务器225A。

[0112] 在这一点(点C)处,在停用VM T3 490C期间,CEM 450决定VM T3 490C可以不再服务于业务。此时,CEM 250将VM T3 490C因为VM T3 490C被移动而在此时无法服务于业务通知CNM 260。对该消息作出反应,CNM 260修改虚拟交换机217A的流条目表218A中的规则。从而,在点D处,CNM 260在流条目表218A中安装规则,其指示使VM T3 490C的MAC地址(作为目的MAC)与链接虚拟化服务器215A和TORS 201A的MPLS标记匹配的业务应转发到CNM 260(或到某一其他实体)。CNM 260或可选地某一其他实体将对智能提供如何应对被指引到针对VM T3 490C的虚拟化服务器215A的业务。

[0113] 在一个实施例中,CNM 260或为该目的设计的某一其他实体包括包缓冲器1166,用于在过渡期间存储针对VM T3 490C的包。这样,包缓冲器1166中针对VM T3 490C的包一旦在新虚拟化服务器225A处激活VM T3 490C则可以发送到VM T3 490C。

[0114] 在一个实施例中,在VM T3 490准备服务于新虚拟化服务器225A处的数据时,CEM 250将VM T3 490在点E处再次是活动的通知CNM 260。在另一个实施例中,CNM 260可以检测到VM T3 490因为CNM 260在VM T3 490被停用后接收源自它的包(例如ARP请求)而是活动的。CNM 260通过修改流条目表228A而对准备服务于数据的VM T3 490做出反应。

[0115] 在点F处,CNM 260将一个或多个条目添加到流条目表228A。CNM 960将匹配VM T3 490C的MAC地址(作为目的MAC)连同匹配链接虚拟化服务器225A和TORS 201C的MPLS标记的流规则添加到流条目表228A。该规则告知虚拟交换机227A匹配包应弹出两个MPLS标记并且将其转发到VM T3 490C。本发明的其他实施例支持额外规则。例如,一个实施例添加指示源自VM T3 490C的DHCP包应转发到CNM 260的规则。另一个实施例添加指示源自VM T3 490C的ARP请求应转发到CNM 260的规则。

[0116] 在点G处,CNM 260(或设计成在VM从旧循环服务器到新虚拟化服务器的过渡期间缓冲包的某一其他实体)将去往VM T3 490C的累积包转发到虚拟化服务器225A,因此那些包可交付给VM T3 490C。

[0117] 在点G处,CNM 260修改CNM 260的地址映射表262表来使对于VM T3 990C的虚拟化服务器MAC地址条目变成目的新虚拟化服务器225A的MAC地址。在这些改变后,根据图9的对于VM T3 490C的任何新的ARP请求将导致在源虚拟交换机处指示业务应转发到225A的流条目。

[0118] 在一个实施例中,CNM 260使流规则保持在对于任何对等VM(其可仍然与VM T3 490C交换业务)的旧虚拟化服务器215A的虚拟交换机217A上的位置,因为那些对等VM可仍然期望VM T3 490C在旧虚拟化服务器215A上是活动的。如果这样的对等VM应向VM T3 490C发送对等数据包,它将根据在点D处安装的规则而被转发到CNM 260。CNM 260然后修改对等

VM的虚拟化服务器上的流规则以使用包上的MAC和IP地址将业务转发到新虚拟化服务器225A上的VM T3 490C来定位VM T3 490C。CNM 260将对等数据包一直转发到新虚拟化服务器225A上的VM T3 490C。这样,所有流规则最终超时或被修改以将业务发送到新虚拟化服务器225A上的VM T3 490C。CNM可以允许源虚拟交换机流规则在预定宽限期后超时。

[0119] 管理云系统中的组播组成员-概览

[0120] 图12是对于云系统中要加入组播组的虚拟机的过程的一个实施例的流程图。在一个实施例中,过程由MCC在OpenFlow控制器处实现。在框1201处,过程可以响应于在MCC处接收的组播加入消息(例如IGMP或MLD报告消息)而发起。例如,VM可以通过经由IPv4发出IGMP报告消息或通过经由IPv6发出MLD报告消息而加入组播组。这些加入消息包括组播组地址以及VM的源地址。在一个实施例中,在云系统的CNM或相似部件的启动期间,每个VS的流表配置成包括这样的条目,其检测所有IGMP或MLD消息并且将它们转发到MCC以供处理。

[0121] MCC维持组播组管理表(MGMT),其跟踪具有在每租户基础上预订的至少一个VM的每个组播组。MGMT在下文进一步详细描述。在接收VM加入消息时,在框1203处,将VM添加到对于预订加入消息中规定的组播组的租户的VM集。

[0122] 如在上文描述的,MCC还可以跟踪与每个租户关联的VM集以及对于每个VM的关联VS和TORS。在接收加入消息并且将VM添加到MGMT时,则在框1205处,可以使用这些数据结构进行查找来确定发送加入消息的VM的VS和TORS。VS和TORS信息然后可以用于配置每一个的流表条目。在框1207处,VS配置成将出站(相对于VM)组播业务转发到具有经由VM的TORS而预订组播组的VM的TORS集。在框1209处,VS还可以配置成放弃源自VM、从TORS接收的组播组业务,因为该组播数据环回并且因此可以被丢弃。TORS的流表条目可以用VS的流表条目相继或并行更新。在框1211处,TORS流表条目被更新以使用MPLS将出站组播业务转发到预订VM的TORS并且将进入组播业务转发到机架中预订组播组的VM的VS。过程的一个实施例在下文进一步详细描述。

[0123] 在VM离开组播组时,离开消息相似地由VM发送并且在VS处被拦截。拦截的离开消息被转发到MCC。MCC更新MGMT以在离开消息中从VM集去除预订识别的组播组的VM。VS流表然后被更新以从VM集去除接收组播业务的VM并且相似地可更新TORS来如果它是与预订组播组的VS关联的最后的VM则去除VM的VS。也可以更新云系统中的每个TORS以如果VM是与预订组播组的TORS关联的最后的VM则从对于组播组的转发去除VM的TORS。

[0124] 加入组播组-虚拟交换机配置

[0125] 图13是用于由VM加入组播组并且配置VM的虚拟交换机的过程的一个实施例的流程图。在一个实施例中,在框1301处,加入过程在MCC从VM接收报告消息时发起。如上文提到的,CNM或相似的部件配置每个VS来拦截IGMP报告消息和MLD报告消息并且将其转发到MCC以供处理。IGMP报告消息或MLD报告消息包括VM MAC地址作为源地址。在框1303处,MCC使用在上文描述的数据结构(例如租户表)来对VM的MAC地址进行查找,来确定对于VM的租户ID。在框1305处,MCC然后在具有租户id的MCC表中进行第二查找来获得指针以提取对应的MGMT条目。

[0126] 如果VM记录不存在,则在框1307处,MCC构造VM记录,其包括指向与VM关联的VS和TORS的VS和TORS记录的指针。在框1309处,VM记录或指向VM记录的指针插入由IGMP或MLD报告消息识别的组播组的MGMT条目内。对于每个组播组的MGMT条目可以响应于加入组的第一

请求在组播源启动时、在MCC启动时或在相似时间创建。MGMT条目包括要用于每个组播组的MPLS标记。在框1311处，MCC检索对于组播组的MPLS标记以供在对与VM关联的VS和TORS的流表编程中使用。

[0127] 如果新近预订的VM是与要预订对于该租户的组播组的VS关联的第一VM，则MCC在VS的流表中设置下列流表条目：

[0128] 如果接收的包的MPLS标记与MPLS组播路由标记匹配，在框1315处，弹出标记并且将包复制到处理输入的OpenFlow组，例如IG_n。

[0129] OpenFlow组表使OpenFlow能够代表转发的额外方法（例如，选择及全部）。每个组条目由它的组标识符识别并且具有有序的动作桶列表。每个桶包含要执行的动作和关联参数的列表。IG_n是在对租户添加新的组播组时创建的输入组。在它来自具有MPLS标记的TORS时，它包含需要转发包的所有VM的列表。

[0130] IG具有类型ALL，其导致包被复制并且发送到组中的每个桶。在将VM添加到组播组来包含到新近预订的VM的路由时，递增地修改IG_n。在新的租户请求到达控制器时，在虚拟交换机中对每一个组播组创建IG_n。

[0131] 除上文的配置外，对于特定租户的任何新近预订VM，MCC通过将下列流条目和修改添加到IG来修改VS的流表：

[0132] 如果接收的包的目的IP地址与组播组地址DIP_{MG1} 匹配，租户VM位于的输入端口压入MPLS组播路由标记ML_{MG1}并且将包复制到处理输出的OpenFlow组OG₁。OG₁是在服务器上对每一个租户创建的组并且在租户加入组播组时用条目填充。

[0133] 如果源地址与新近预订VM的源地址SIP_{VM1}匹配，并且包在TORS端口PTORS₁上进入，放弃包。该规则的目的是防止路由环并且将在下一章节中更详细解释。

[0134] 在IG₁中添加将包转发到DIP_{VM1}的转发动作。

[0135] 关于对VS的流表定义的输入组IG和输出组OG的构造。OG具有类型ALL。从而，包被复制并且发送到组中的每个桶。OG的桶包含下列路由：一个桶将包转发到输出端口外，从而通向TORS；并且余下的桶将包转发到服务器上的其他VM，其是组播组的成员，但VM₁除外。IG组也具有类型ALL。它在对于来自具有MPLS标记的TORS的进入包的时候使用。MPLS标记被弹出并且包发送到IG的所有成员VM。

[0136] MCC利用组播组中的VM列表以通过搜遍VM列表并且确定哪些VM具有与VM₁相同的VS而确定哪些VM在与VM₁相同的VS上，以便构造OG₁。递增地构造IG₁。在VS上的VM预订组播组时，将转发动作添加到IG₁。

[0137] 图14是虚拟交换机的示例流表的一个实施例的图。VS 1401可以处理对于云系统中的VM 1405的集的单播和组播数据业务。VM 1405和VS可以由执行管理程序或相似系统的服务器托管或执行，如在上文进一步描述的。到和来自VM 1405的所有数据业务经过VS 1401，因为它充当VM 1405的网络通信端口的抽象/虚拟化，从而实现它们跨云系统的移动。

[0138] VS 1401的流表1403包括流表条目集，每个分成规则和动作。规则定义这样的条件，在该条件下执行动作。在需要采取多个动作的情景中，可以定义动作组，例如OG 1407和IG 1409。可以配置这些组使得组中动作的任何组合可以根据规则和组的配置来采取，其中特定标准可以进一步影响组中动作的执行。

[0139] 在图14的示例实施例中，流表已经用示例规则和动作集配置。在一流表条目中

管理源自于VS关联的VM的出站组播业务。流表条目寻找具有来自本地VM的源IP地址和与组播组关联的目的IP地址的包。在找到匹配时,动作包括将组播组的MPLS标记封装或‘压入’包上,然后将封装包转发到出站组OG,其定义要采取的额外动作,这些动作包括将包转发到TORS和与VS关联的其他预订VM。

[0140] 下一个流表条目通过寻找对于组播组的匹配MPLS标记、然后作为动作去除标记来管理进入包。解封的包然后转发到与VS关联的本地预订VM。最后的流表条目寻找源IP和TORS端口匹配并且定义放弃动作来丢弃这样的包,因为它们是已经从TORS环回到VS的包。IGMP报告仅在节点将希望接收组播时发出。在地址处进行组播的节点未发出IGMP报告消息。因此,“加入”通过将组播地址转发到控制器而在组播地址上推断。控制器然后必须安装流,正如在接收IGMP报告消息时如何安装流一样。

[0141] 加入组播组-架顶式交换机路由

[0142] 图15是用于对于加入过程配置TORS的过程的一个实施例的图。在安装VS路由后,或并行地,如果还未安装组播组的流路由,MCC在TORS中安装那些路由,即与VM关联的TORS和云系统中的其他TORS。如果相同服务器或VS上的另一个VM已经预订组播组,则额外TORS路由对于该TORS不是必需的,在框1501处,对云系统中的每个TORS或只对于与发出加入消息的VM关联的TORS进行该检查。

[0143] 对于出站包,MCC在TORS上安装第一流中的流路由,其的规则与具有对于组播组的MPLS标记(即,MLMG1)的包匹配。动作复制包并且将副本发送到TORS输出组TOG1。TOG1中的每个动作桶包含下列动作:

[0144] 在框1503处,将骨干路由标记BRL_n(其对应于目的TORS DTORS_n)压入包上;以及

[0145] 在框1505处,将包转发到输出端口外,从而通向DTORS_n。

[0146] 如果已经存在对于DTORS_n的动作桶,则MCC不必写入任何新的流规则。MCC通过遍历VM列表并且提取出将包路由到组播组中的VM的TORS集而确定哪些DTORS接收包。注意该规程仅需要在机架中的第一VM初始加入组播组时进行。在那之后,在其他机架上的其他首次加入VM新近加入组播组时,递增地更新TOG,如在下一个段落中描述的。

[0147] 如果VM是机架上加入组播组的第一VM,在框1507处,MCC也必须在服务于机架(其具有是组播组成员的VM)的所有其他TORS中安装到新近预订VM的TORS的路由。在由源TORS服务的机架上的第一VM初始加入组播组时,该规程应只有一次是必需的。MCC通过搜遍组播组中的VM列表并且提取服务于组中的VM的TORS列表来确定哪些TORS需要路由。然后MCC在列表中的每个TORS的TOG中安装将包路由到新的源TORS的动作。

[0148] 除其他机架上的服务器外,在框1509处,TORS需要将组播包转发到相同机架上预订组播组的其他服务器。对于具有预订组的VS的机架上的每个服务器,存在动作桶,其具有朝VS转发包的動作。注意这意指包源自的VS也将接收包。正是为了这个目的,VS包含放弃从具有与发送VM相同的源地址的TORS进入的包的规则,来防止路由环。

[0149] 对于进入包,使用两个流表。第一流表包含与每个骨干路由标记(其对应于服务于组播组中的VM的TORS)匹配的一个规则。在框1511处,动作弹出骨干路由标记并且将包发送到下一个表。第二表包含与组播路由标记MLMG1匹配的規則。动作复制包并且将它发送到TORS输入组TIG1。在框1513处,输入组中的每个桶包含将包转发到至少一个VM(其是组播组的成员)运行所在的机架上的VS或服务器中的一个的动作。正如对于输出一样,在来自机架

的VM是第一个加入时,MCC必须在服务于具有作为组播组成员的VM的所有TORS中安装规则。

[0150] 图16是具有示例配置的TORS的流表的图。TORS对于多级包处理利用两个流表配置。在第一流表1601中,示出示例条目集。第一条目管理对于特定组播组使用MPLS标记来寻找匹配的出站组播业务。这些匹配包与骨干MPLS标记一起封装并且使用TORS输出组1603发送到云系统中具有对于组播组的预订VM的每个TORS。用第一流表中的第二流表条目处理入站业务,其与骨干标记上的包匹配并且对这些包解封。解封的包然后发送到第二流表1607。

[0151] 在第二流表1607中,包对于MPLS标记匹配来确定它是否与组播组MPLS标记匹配。如果找到匹配,将包复制到TORS输入组,其将包转发到在机架中具有预订VM的每个VS。

[0152] 注意如在上文描述的单播过程中的,骨干路由将不改变并且因此必须接触云系统骨干中的交换机。另外,骨干标记可以与对于上文描述的单播情况完全相同,或云系统运营商可选择为组播业务预留某些路由并且用特定标记指出它们。

[0153] 离开组播组

[0154] 图17是用于在云系统中处理离开组消息的过程的一个实施例的流程图。在VM离开组播组时,它采用IPv4发送IGMP离开组消息并且采用IPv6发送MLD离开组消息。在服务器启动时,CNM在服务器的VS中安装将所有离开组消息转发到MCC的OpenFlow规则。根据什么VM预订组,存在影响需要修改流路由的地方的三个情况:

[0155] 首先,在VM离开组播组所在的VS的服务器上,存在预订组播组的其他VM;

[0156] 其次,VM是服务器或VS上最后离开组的,但存在预订相同组播组的其他机架上的其他VM或机架中的VS的其他服务器;以及

[0157] 第三,VM是机架上最后离开组播组的。

[0158] MCC通过搜遍组播组中的VM列表、使用合适的MGMT并且根据哪些VM在与离开VM VS的相同服务器上以及哪些在机架中的VS的其他服务器上对它们分类而确定要考虑这些情况中的哪个。本文如下解决可能的情况中的每一个。

[0159] 相同VS上仍然预订的其他VM

[0160] 如果在VS的相同服务器上存在其他预订VM,则MCC仅需要在VS的流表中删除下列规则和动作:

[0161] 与VM离开的源IP地址(SIP_{VM1})和组播组的目的IP地址(DIP_{MG1})匹配并且将传出包从VM₁转发的规则。该规则和关联的动作和OG管理来自离开VM的出站组播业务并且因此可以在框1703处删除。

[0162] IG₁中将进入包转发到VM₁的动作,而不是整个IG₁。该规则将进入组播组业务转发到离开VM并且因此可以在框1705处删除。

[0163] 与放弃包以防止路由环的 SIP_{VM1} 和 P_{TORS1} 匹配的规则。该规则防止离开VM的出站组播业务的环回并且因此可以在框1707处删除。

[0164] 然后可以在框1709处和1715处进行是否应用其他情况中的任一个的检查。如果否的话,则不需要修改其他规则并且过程结束。

[0165] VS上离开组的最后的VM

[0166] 如果离开VM是VS上离开组播组的最后的VM,则另外在框1711处删除将进入包转发到VM和关联的IG₁的规则。另外,TORS输出组(TOG_1)和TORS输入组(TIG_1)中将组播组的包转发到与VM关联的VS₁的动作在框1713处删除。

[0167] 然后可以在框1715处进行离开VM是否是机架上(即,与TORS关联)预订组播组的最后的VM的检查。如果它不是TORS上预订组播组的最后的VM,则过程结束。

[0168] 机架上离开组的最后的VM

[0169] 如果VM是机架上离开组播组的最后的VM,则TORS₁的第一流表中的两个规则和第二流表中的一个规则连同它们相应的输入和输出组另外在框1717处去除。

[0170] 处理VM移动

[0171] 图18是用于处理VM移动以在云系统中使用MPLS支持组播的过程的一个实施例的流程图。服务器与VS之间的VM移动可以通过上文描述的技术和在上文关于单播描述的一般VM移动技术的混合来处理。该过程为了负载平衡或相似的原因而通过VM的移动在框1801处发起。

[0172] 在移动期间,将组播业务指引到与移动VM关联的VS和TORS中的移动VM的路由被维持直到VM在新的位点中执行并且准备接收业务。在VM移动期间,根据在上文关于单播描述的规程,在框1803中,新的路由(即,流表条目)被放置在旧的VS中来捕捉去往移动VM的多播以及单播业务,并且将它发送到CNM用于缓冲。

[0173] 在框1805处,缓冲继续直到完成VM移动。在框1807处,然后可以通知CNM将缓冲包转发到新的位点处的VM。在新的VM变成活动之前,如果VM是组播组的成员,在框1809处,需要修改组播组VM记录中的VS和TORS记录来反映新的位点。另外,在框1811处,对于VM预订的每个组播组,上文描述的对于加入组播组的规程必须新的VS和TORS上执行。在框1813中,离开操作也在每个预订组播组中对于旧的VS和TORS进行。

[0174] 图19图示根据本发明的实施例的示范性计算机系统的框图。该示范性计算机系统1000在图10中的可以用于实现TORS 201A、虚拟化服务器205A、CEM 250、MCC 290和/或CNM 260。如果CEM 250、MCC 290和/或CNM 260在OpenFlow控制器处实现,该系统可以执行OpenFlow控制器。本领域内技术人员将认识到用于实现这些设备的其他计算机系统可具有更多或更少的部件并且可在本发明的实施例中使用。

[0175] 计算机系统1900包括总线1050,其与处理系统1915、电力供应1920、易失性存储器1925(例如,双数据速率随机存取存储器(DDR-RAM)、单数据速率(SDR)RAM)、非易失性存储器1930(例如,硬驱动器、闪速存储器、相变存储器(PCM))耦合。处理系统1915可进一步耦合于处理系统高速缓存1910。处理系统1915可从易失性存储器1925和/或非易失性存储器1930检索指令,并且执行指令来进行上文描述的操作。总线1950使上文的部件耦合在一起并且进一步使显示控制器1070、一个或多个输入/输出设备1080(例如,网络接口卡、光标控制(例如,鼠标、轨迹球、触屏、触摸板,等)键盘、等)和可选地一个或多个无线收发器(例如,蓝牙、WiFi、红外,等)耦合。在一个实施例中,显示控制器1970进一步耦合于显示设备1975。

[0176] 如本文描述的,指令可指硬件的特定配置,例如专用集成电路(ASIC),其配置成进行某些操作或具有存储在存储器(包含在非易失性计算机可读介质中)中的预定功能性或软件指令。从而,在图中示出的技术可以使用在一个或多个电子设备(例如,终端站、网络元件)上存储并且执行的代码和数据来实现。这样的电子设备使用例如非暂时性计算机可读存储介质(例如磁盘;光盘;只读存储器;随机存取存储器;只读存储器;闪速存储器设备;相变存储器)和暂时性机器可读通信介质(例如,电、光、声或其他形式的传播信号-例如载波、红外信号、数字信号)等计算机可读介质来存储和传达(在内部和/或通过网络利用其他电

子设备)代码和数据。另外,这样的电子设备典型地包括一个或多个处理器集,其耦合于一个或多个其他部件,例如一个或多个存储设备(非暂时性机器可读存储介质)、用户输入/输出设备(例如,键盘、触屏和/或显示器)和网络连接。处理器集和其他部件的耦合典型地通过一个或多个总线和网桥(也称作总线控制器)。从而,指定电子设备的存储设备典型地存储代码和/或数据以供在该电子设备的一个或多个处理器集上执行。当然,本发明的实施例的一个或多个部分可使用软件、固件和/或硬件的不同组合来实现。

[0177] 尽管图中的流程图示出由本发明的某些实施例执行的操作的特定顺序,应理解这样的顺序是示范性的(例如,备选实施例可按不同顺序进行操作、使某些操作组合、与某些操作重叠,等)。

[0178] 尽管本发明已经从若干实施例方面描述,本领域内技术人员将认识到本发明不限于描述的实施例,可以在附上的权利要求的精神和范围内带有修改和更改地实践。描述从而视为说明性而非限制性的。



图 1A



图 1B

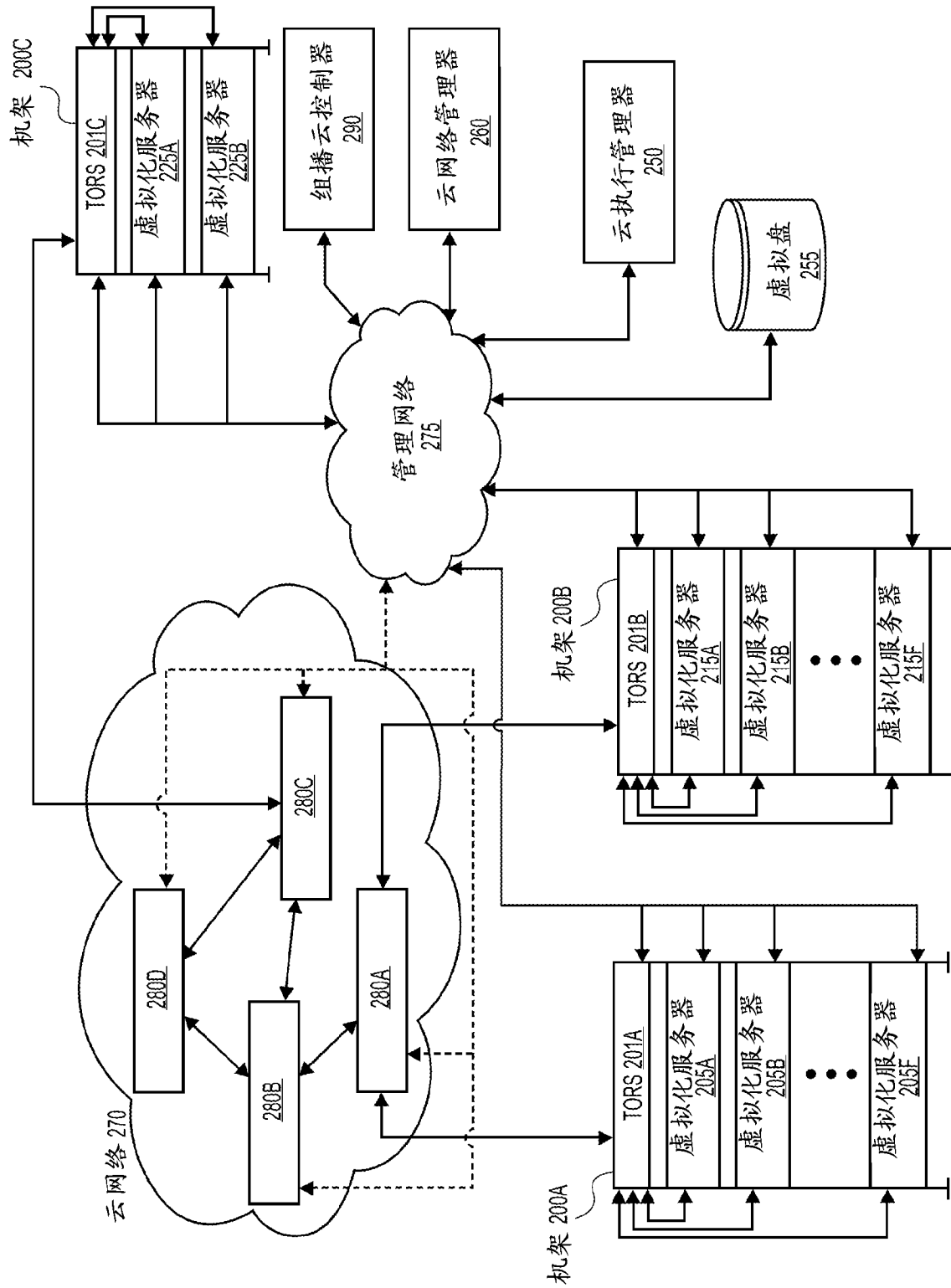


图 2

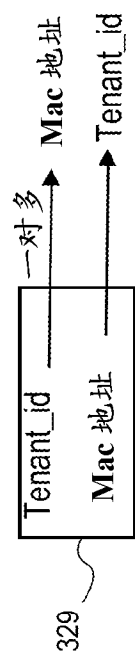


图 3A

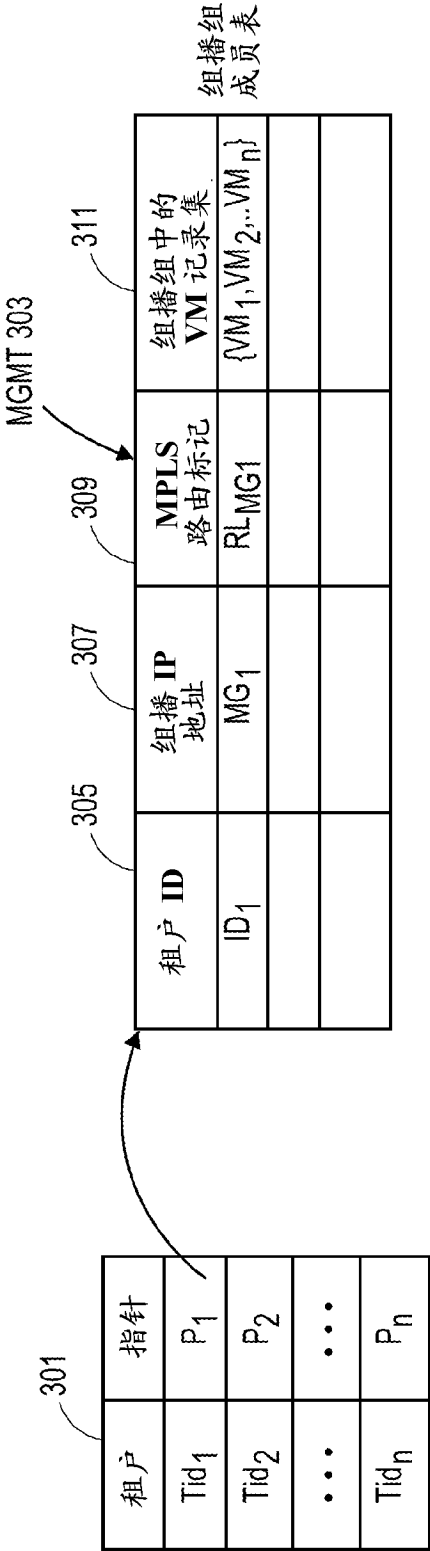


图 3B

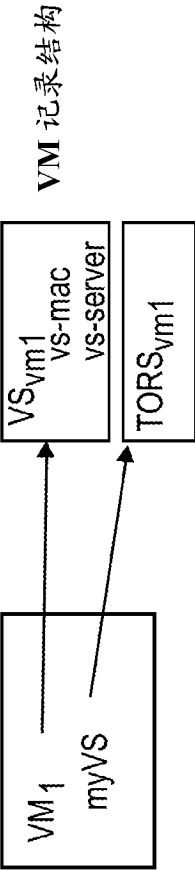


图 3C

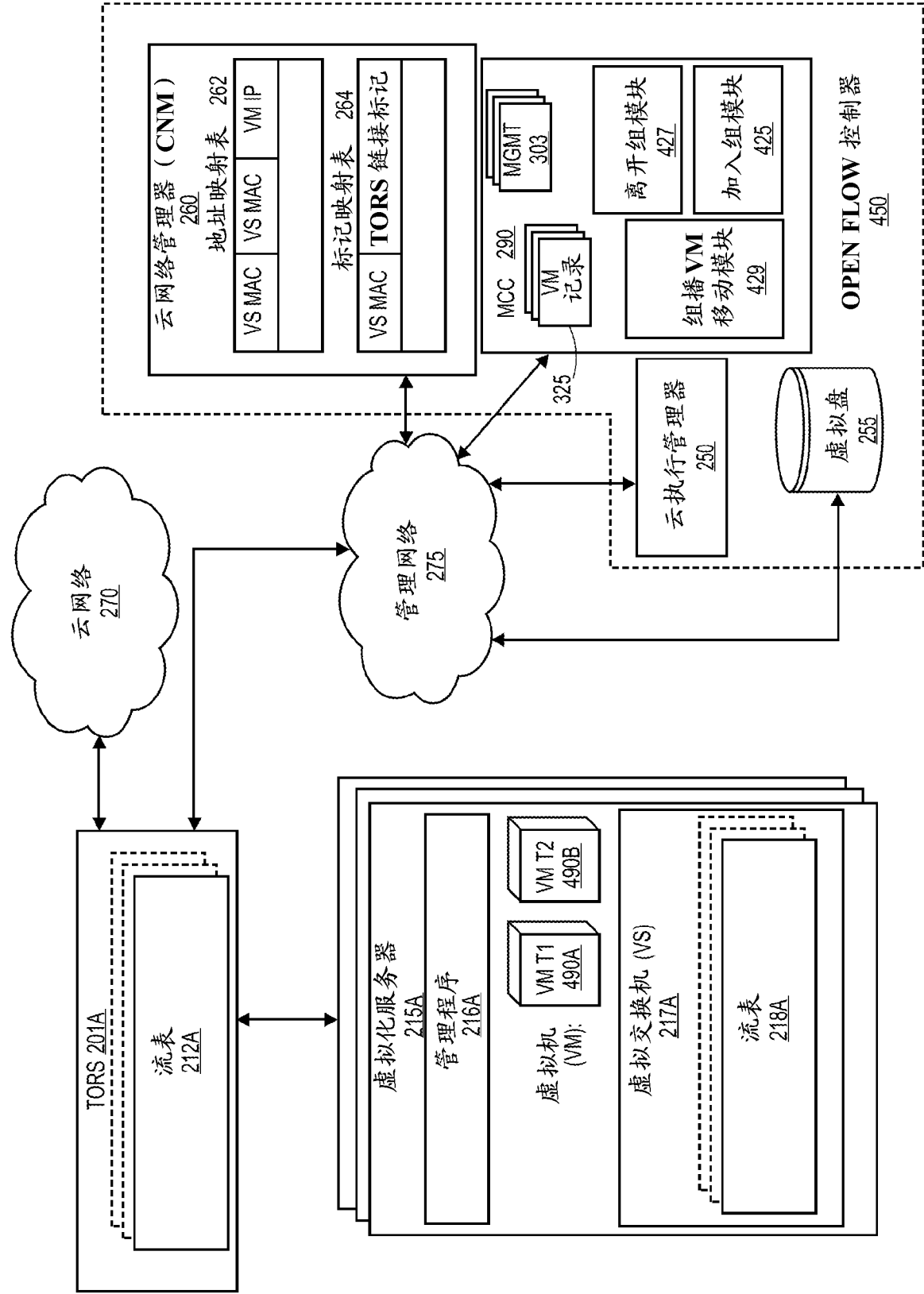


图 4

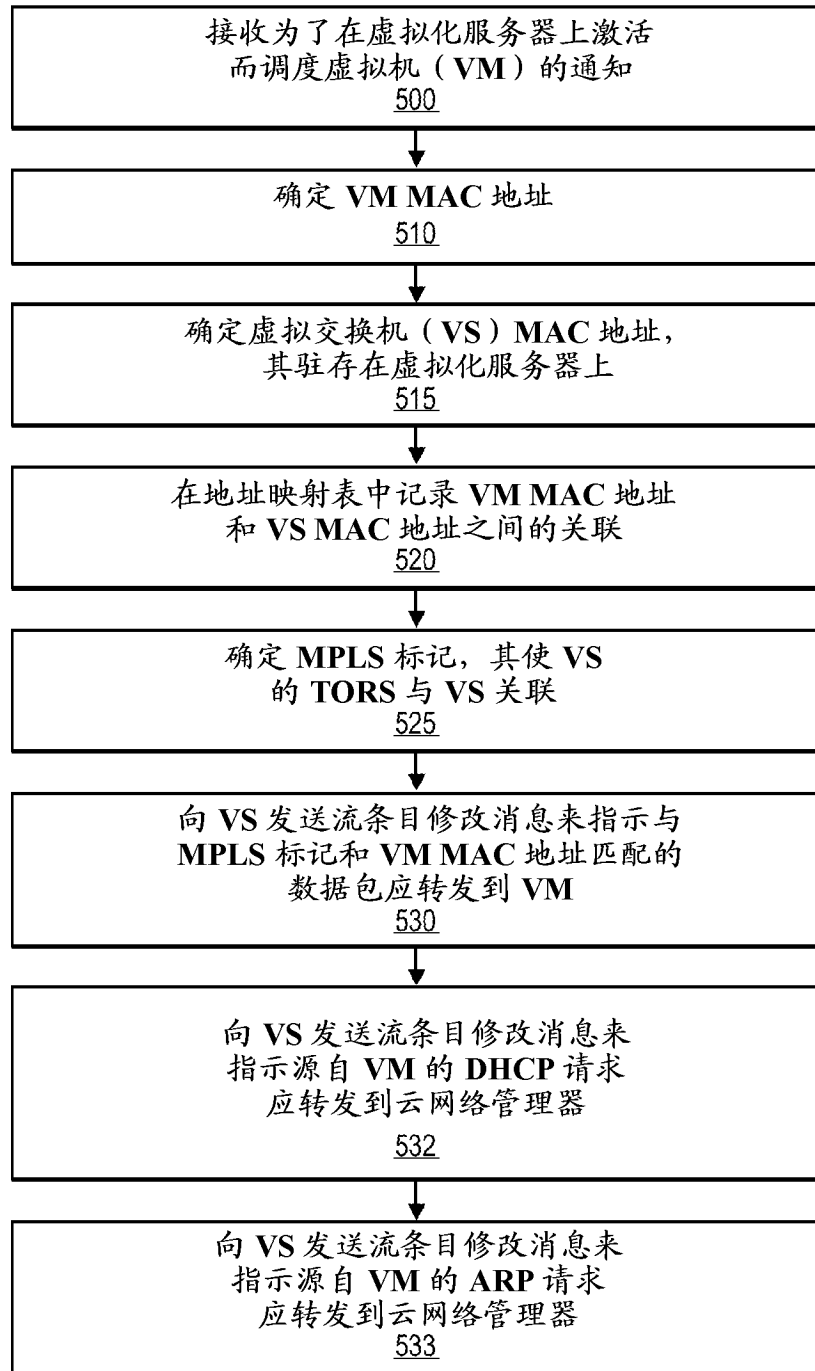


图 5

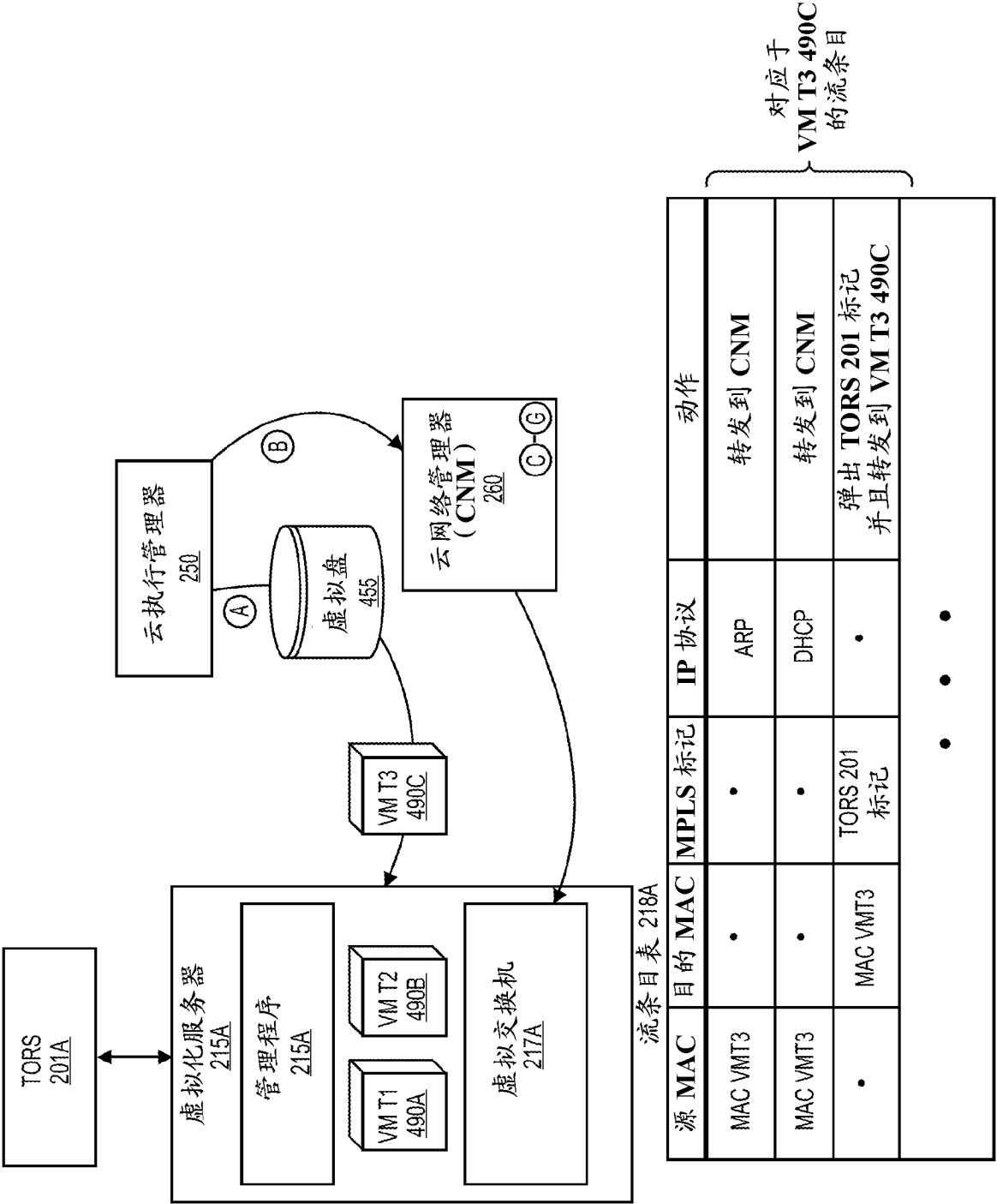


图 6



图 7

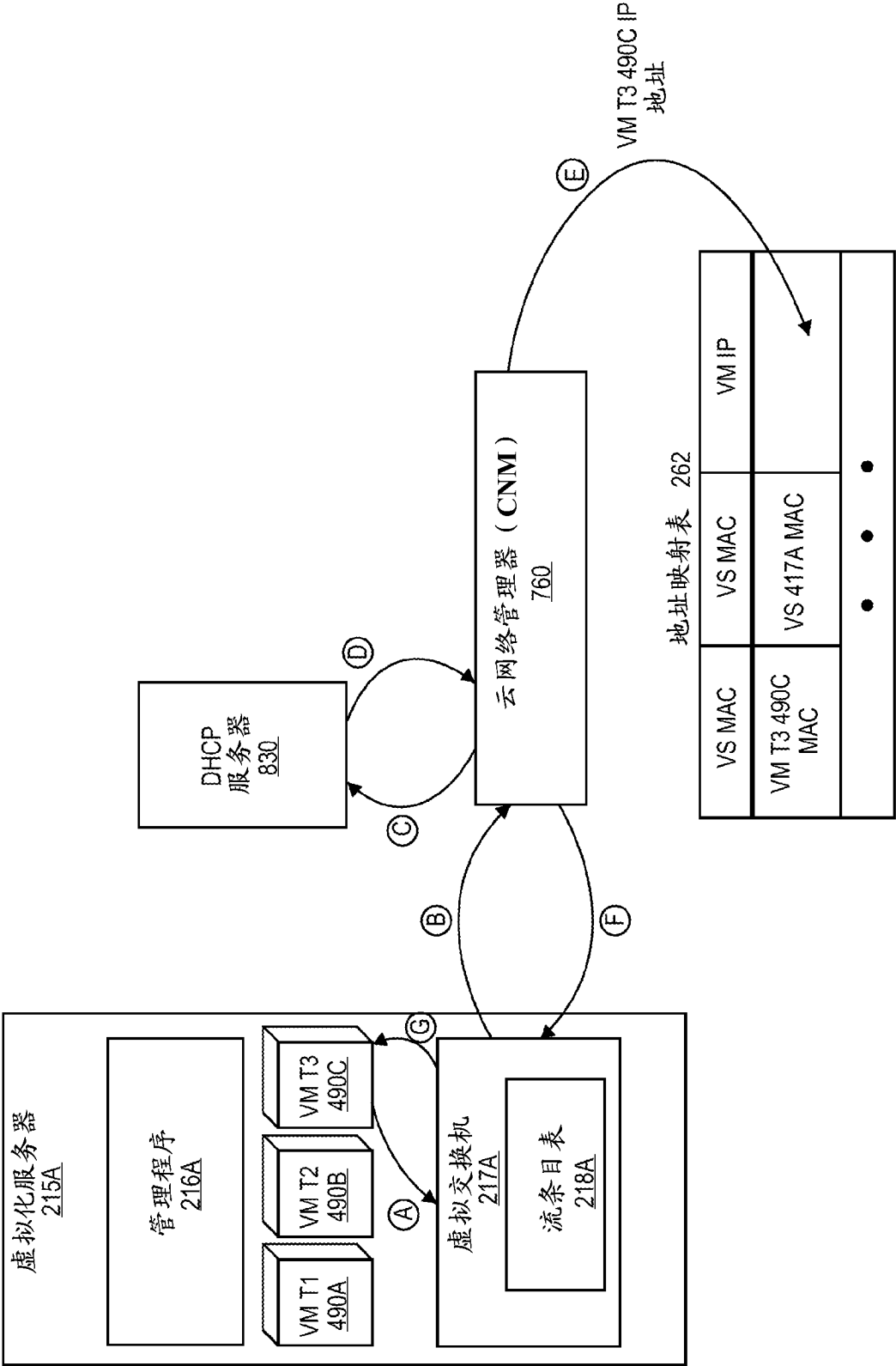


图 8

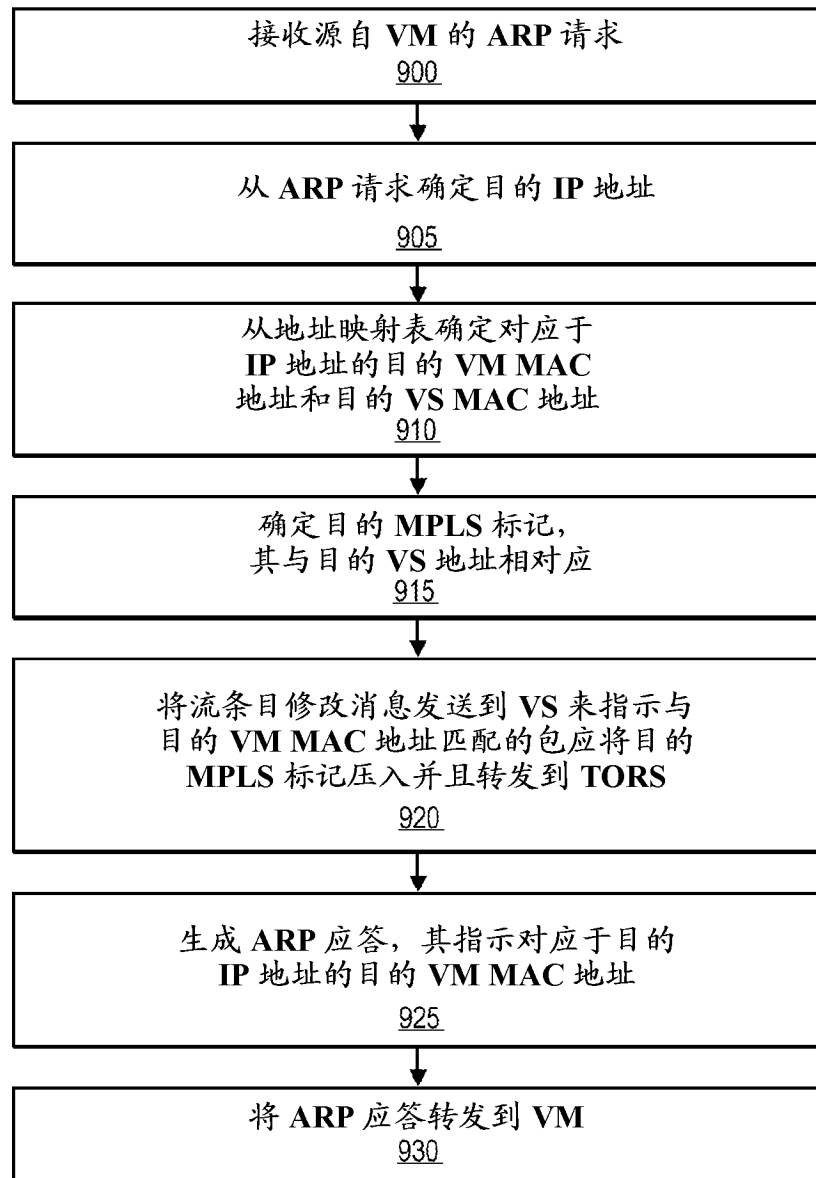


图 9

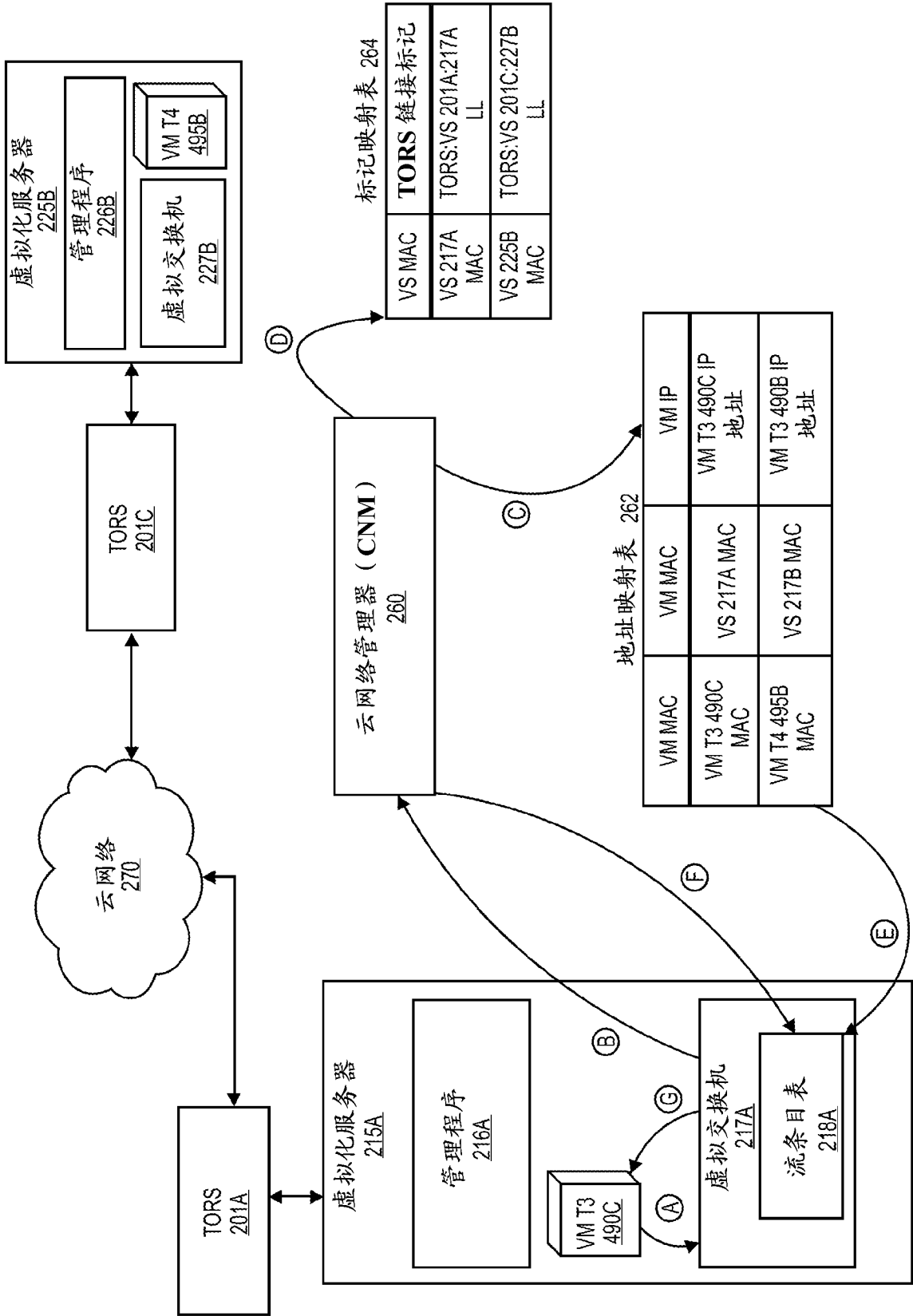


图 10

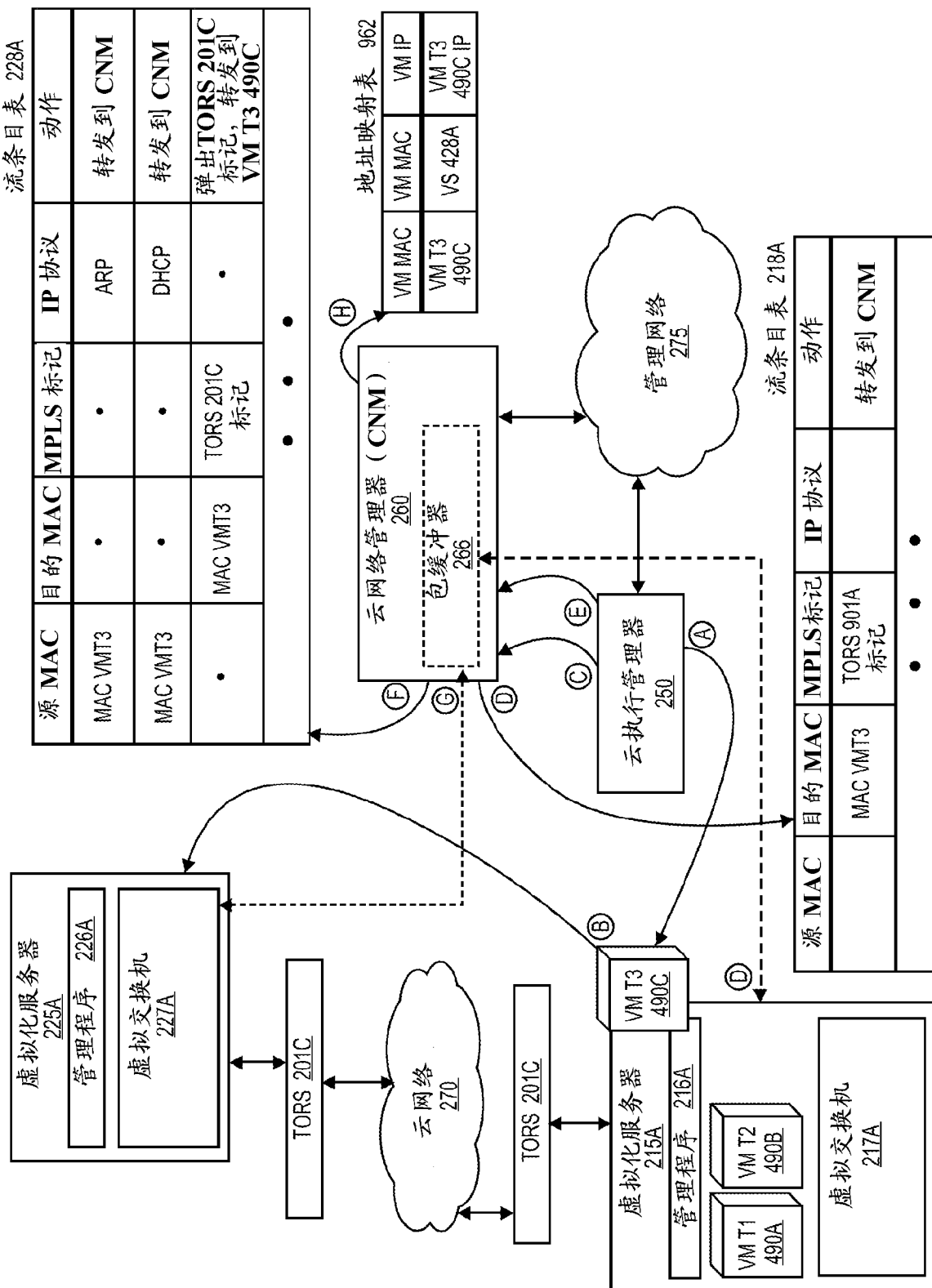


图 11

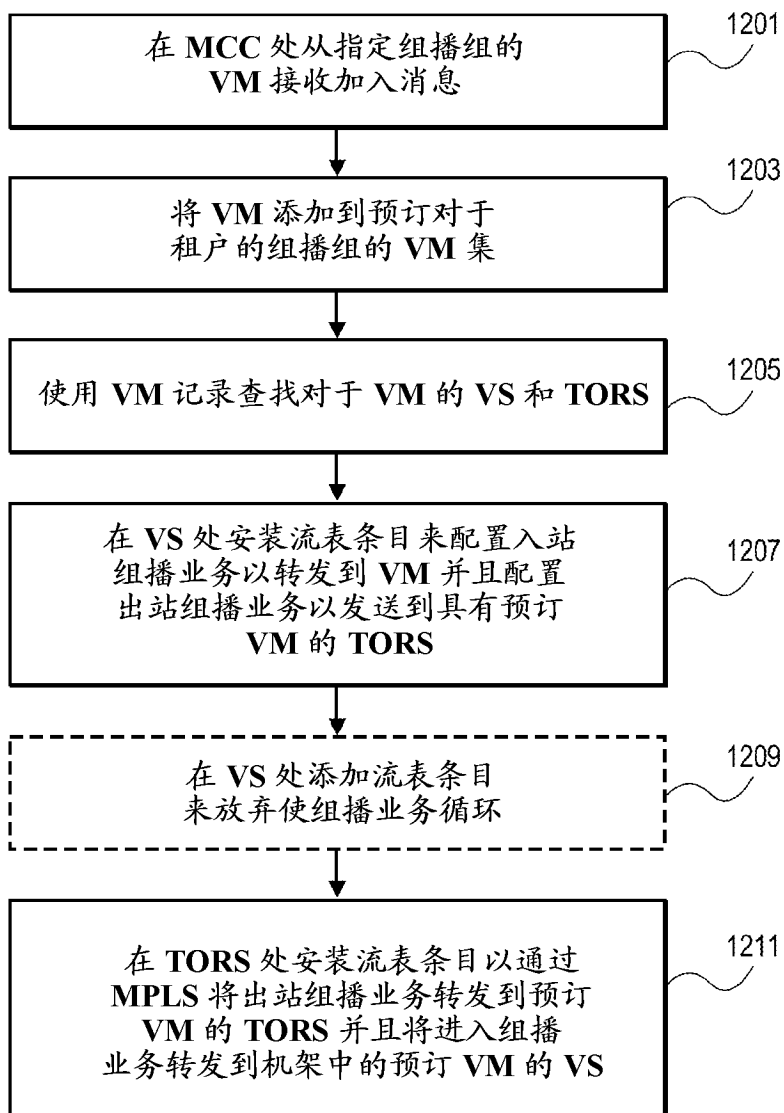


图 12

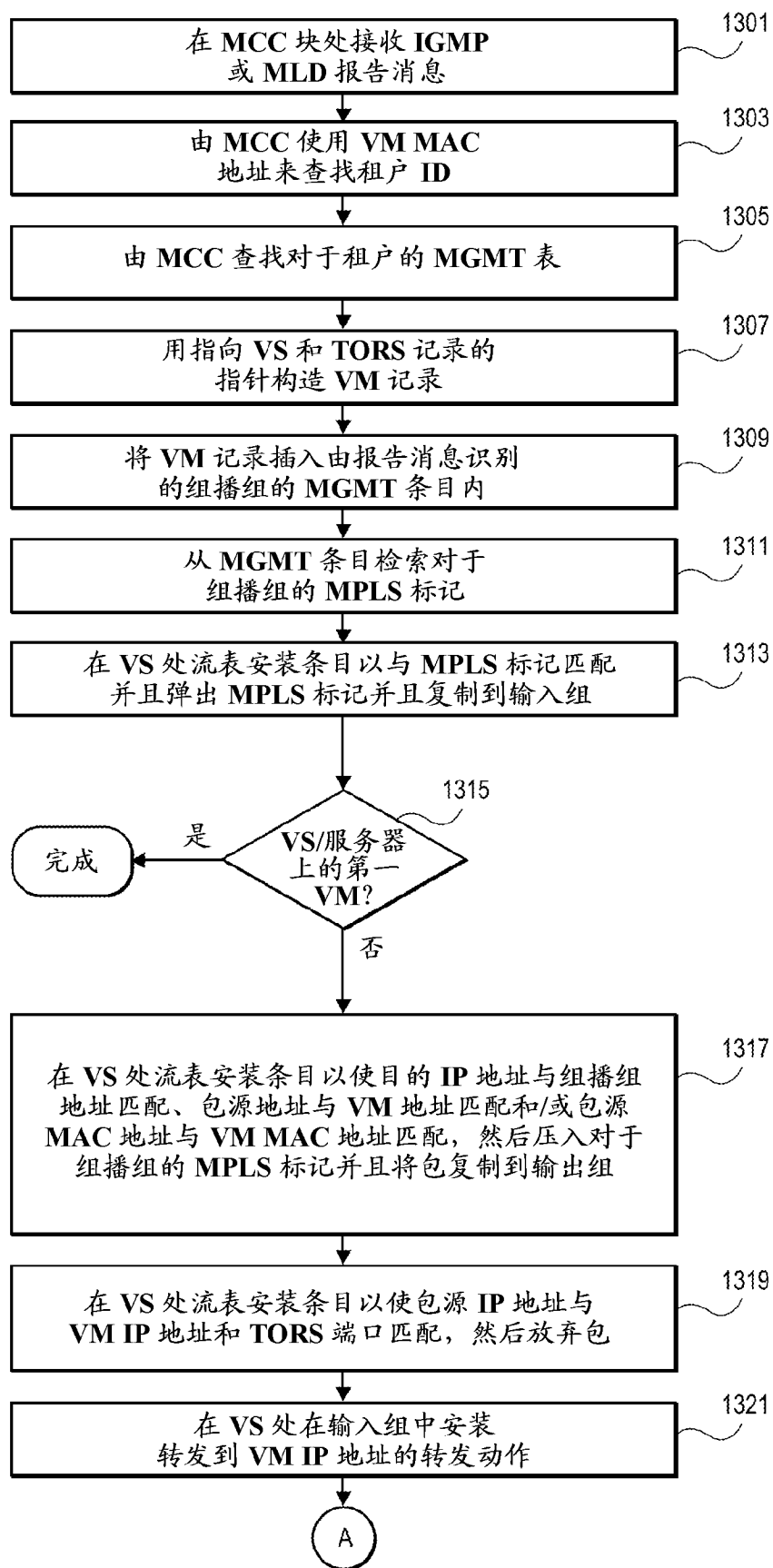


图 13

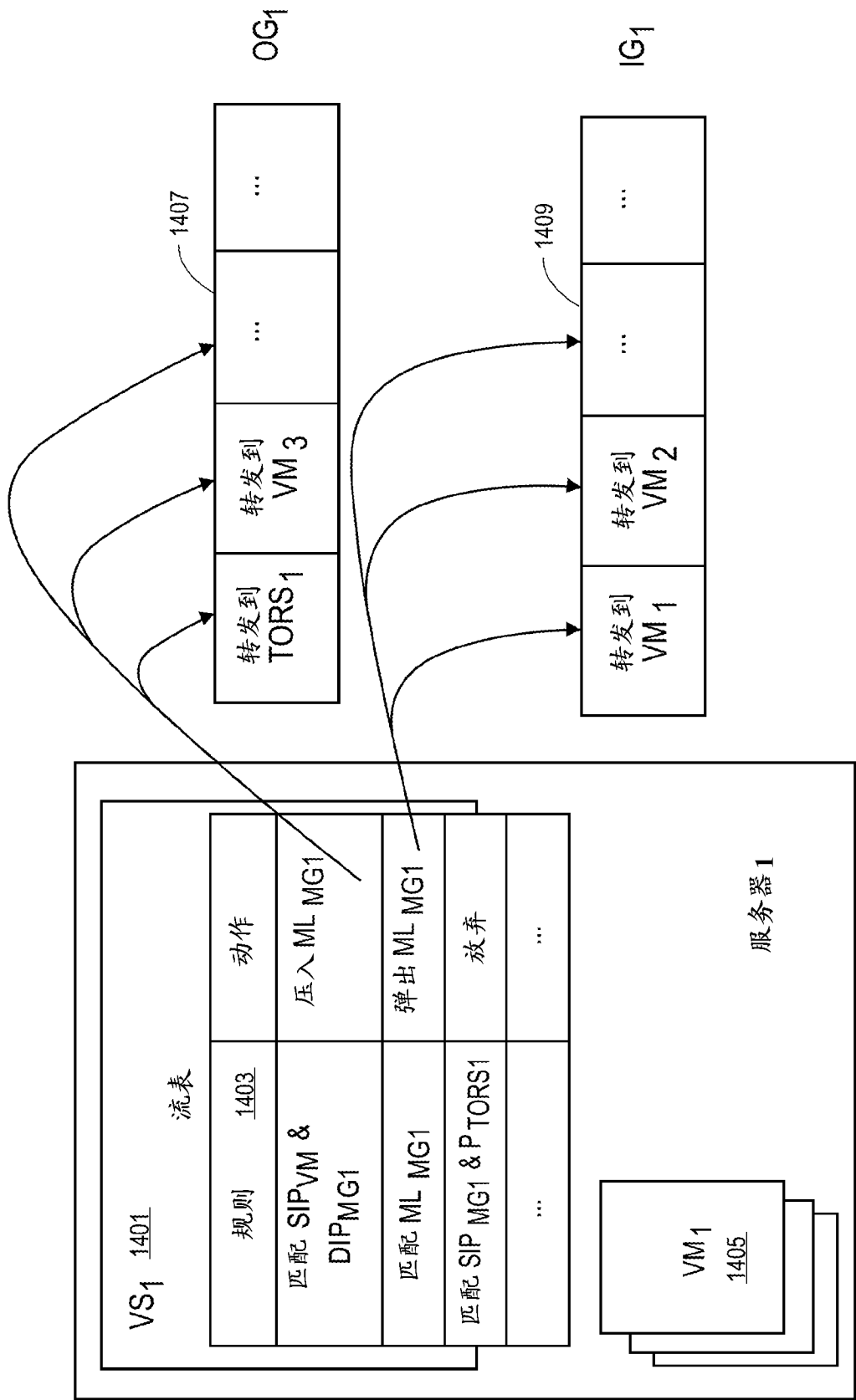


图 14

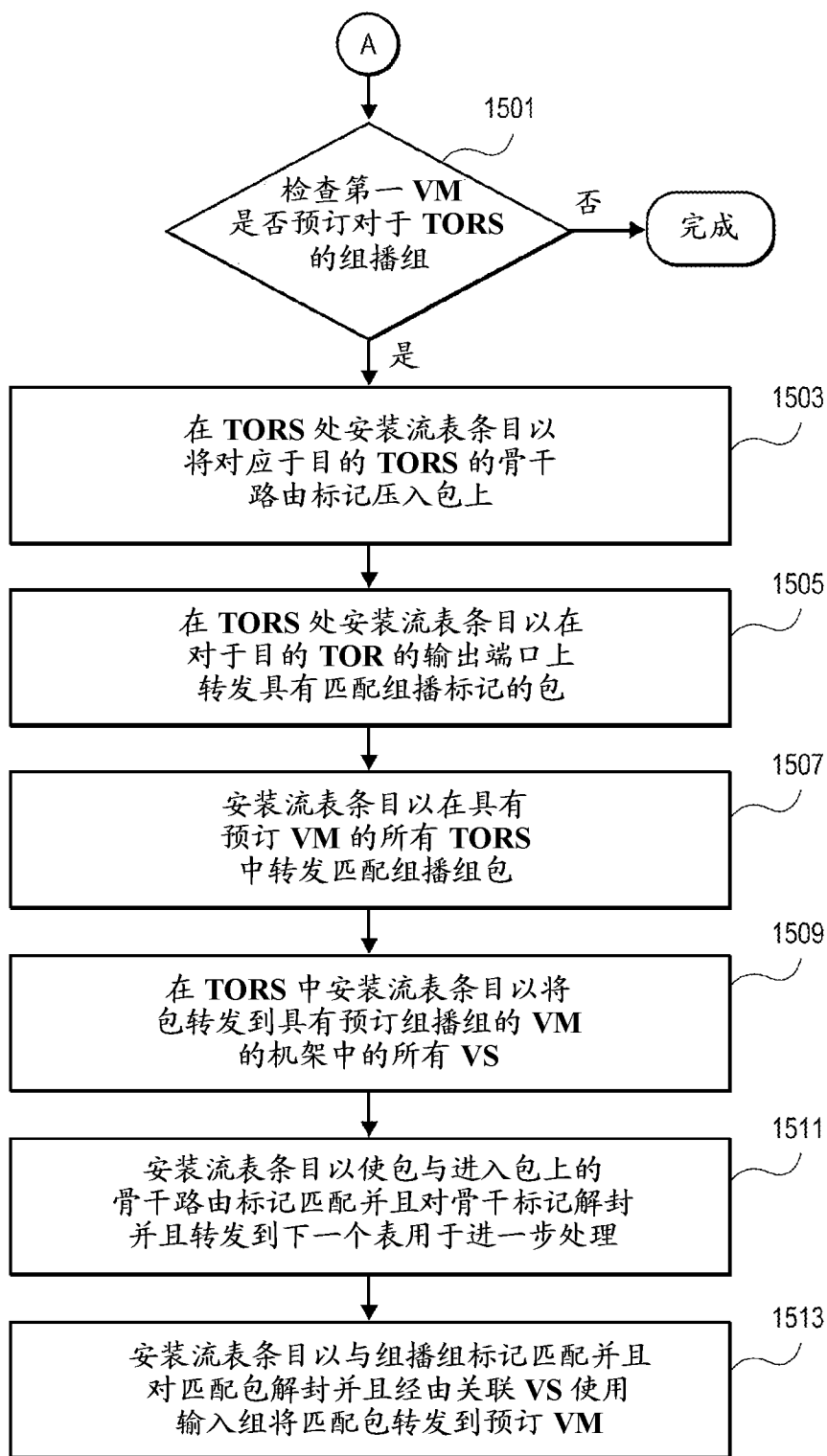


图 15

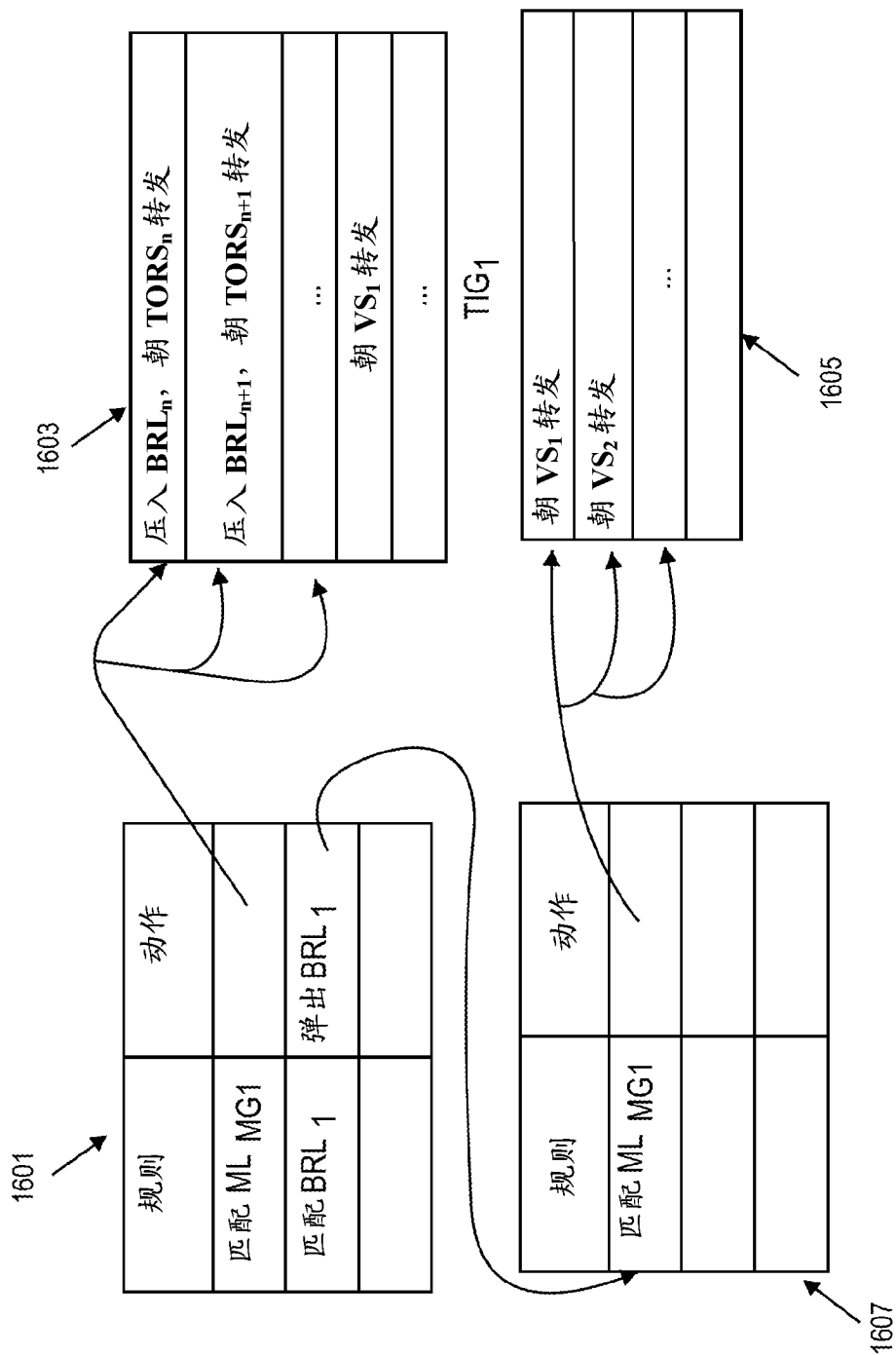


图 16

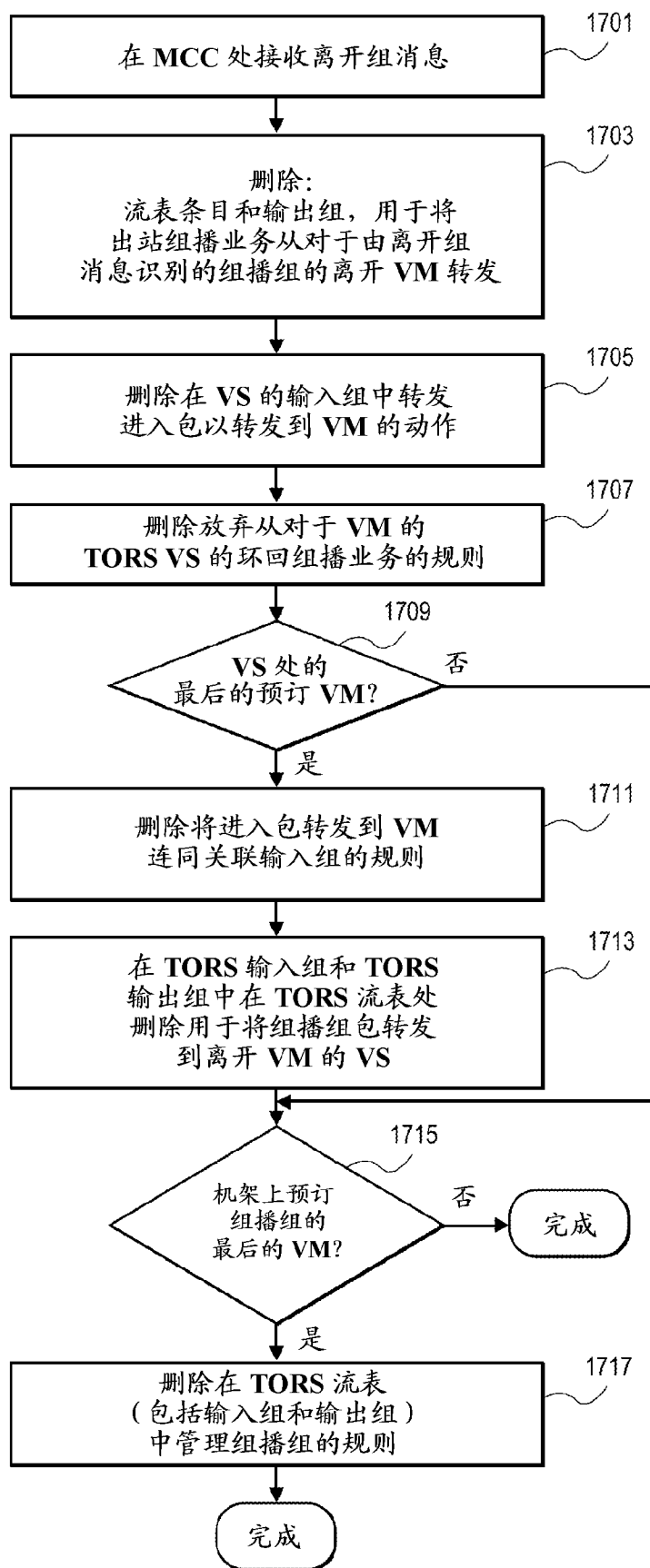


图 17

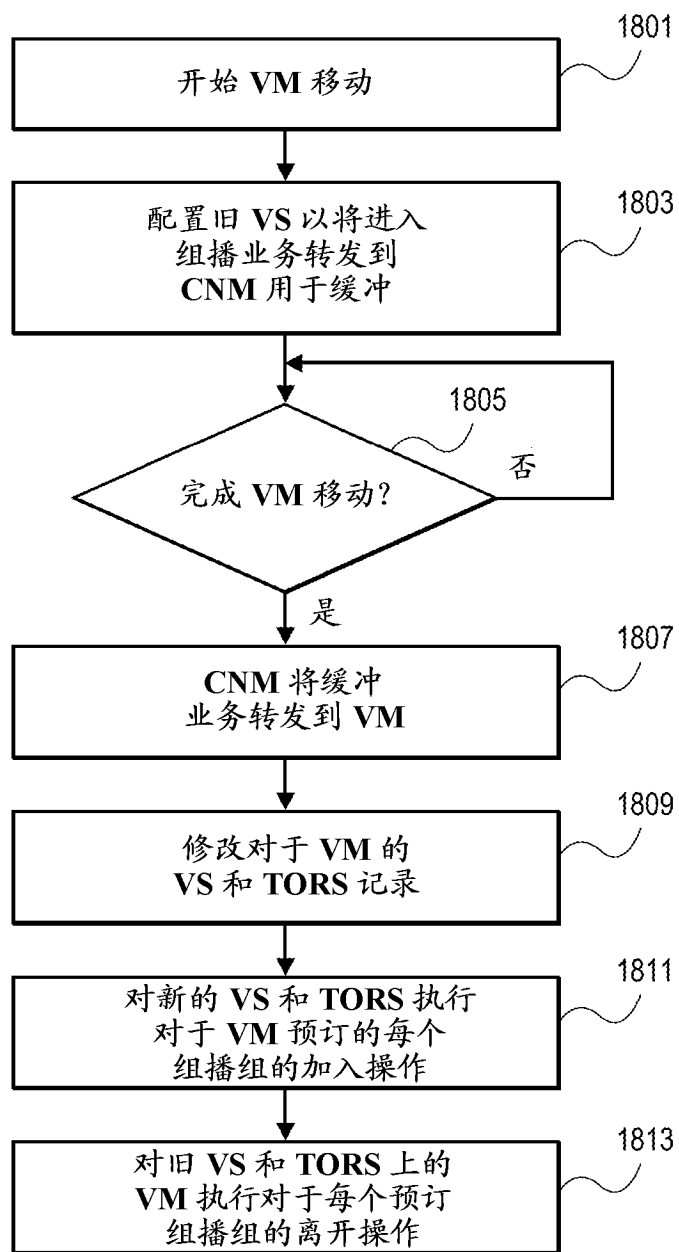


图 18

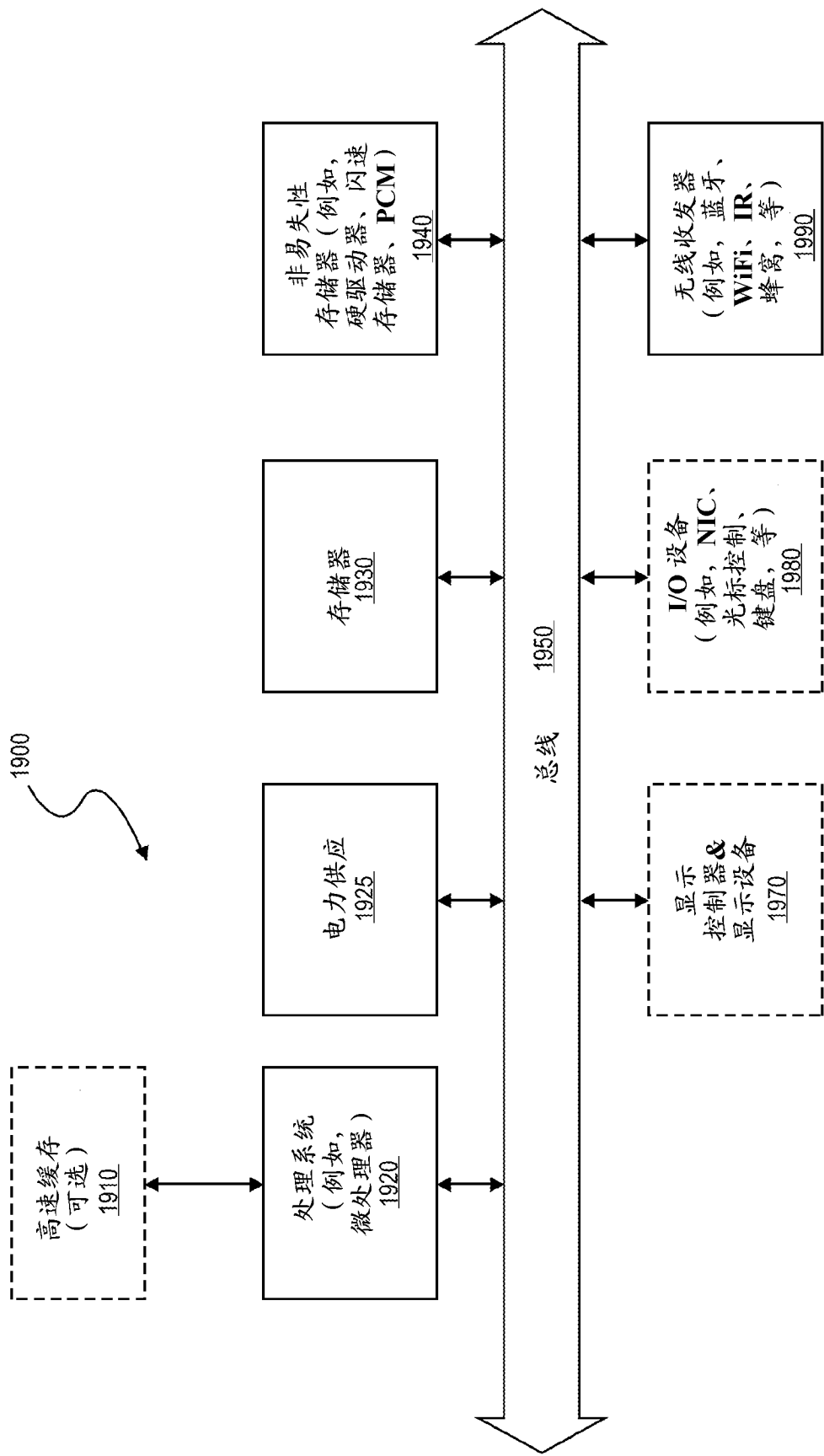


图 19