



(12) 发明专利申请

(10) 申请公布号 CN 103227843 A

(43) 申请公布日 2013. 07. 31

(21) 申请号 201310121091. X

(22) 申请日 2013. 04. 09

(66) 本国优先权数据

201210318996. 1 2012. 08. 31 CN

(71) 申请人 杭州华三通信技术有限公司

地址 310053 浙江省杭州市高新技术产业开发区之江科技工业园六和路 310 号华为杭州生产基地

(72) 发明人 宋玉兵

(74) 专利代理机构 北京鑫媛睿博知识产权代理有限公司 11297

代理人 龚家骅

(51) Int. Cl.

H04L 29/12(2006. 01)

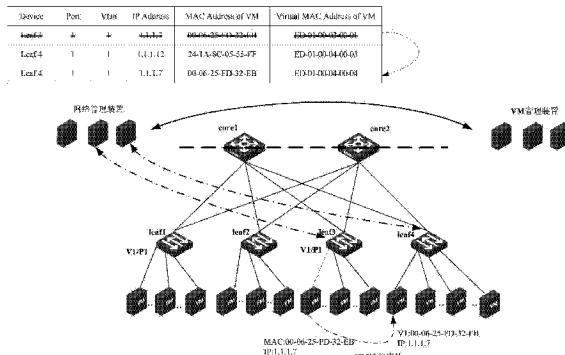
权利要求书6页 说明书30页 附图7页

(54) 发明名称

一种物理链路地址管理方法及装置

(57) 摘要

本发明公开了一种物理链路地址管理方法及装置，其中，网络管理装置为大二层网络中各数据中心的虚拟机分配虚拟 MAC 地址，且保证同一接入层设备连接的所有虚拟机的虚拟 MAC 地址构成连续地址空间，每个数据中心内的所有虚拟机的虚拟 MAC 地址构成连续地址空间；网络管理装置根据分配的虚拟 MAC 地址，为大二层网络中的各数据中心中的接入层设备和核心层设备配置基于 MAC 地址掩码的二层转发表，通过 MAC 地址掩码对虚拟 MAC 地址进行聚合，有效地减少了二层转发表中表项的数量。



1. 一种物理链路地址管理方法,应用于包括多个数据中心的大二层网络,其特征在于,该方法包括:

 网络管理装置根据预设的虚拟 MAC 地址配置规则为每个虚拟机设置一个虚拟媒体接入控制 MAC 地址;

 所述虚拟 MAC 地址配置规则包括:每个所述虚拟 MAC 地址由唯一性标识、网络标识、接入层设备标识以及主机标识组成且字节数等于每个所述虚拟机自身的真实 MAC 地址的字节数;

 位于同一个数据中心且接入到相同接入层设备的所有虚拟机的虚拟 MAC 地址具有相同的唯一性标识、相同的网络标识、相同的接入层设备标识以及不同的主机标识;位于同一个数据中心且接入到不同接入层设备的所有虚拟机的虚拟 MAC 地址具有相同的唯一性标识、相同的网络标识、不同的接入层设备标识。

2. 如权利要求 1 所述的方法,其特征在于,所述方法还包括:

 所述网络管理装置在一个接入层设备的二层转发表中配置至少一对虚拟机表项;每对虚拟机表项关联于所述接入层设备接入的一个虚拟机;每对虚拟机表项包括第一虚拟机表项和第二虚拟机表项;

 其中,第一虚拟机表项包含虚拟机所属虚拟局域网 VLAN 的标识、所述虚拟机的虚拟 MAC 地址、主机掩码、映射于所述虚拟 MAC 地址的所述虚拟机的真实 MAC 地址以及指向所述虚拟机的出端口;

 所述第二虚拟机表项包括虚拟机所属 VLAN 的标识,所述虚拟机的真实 MAC 地址,主机掩码,映射于所述真实 MAC 地址的所述虚拟机的虚拟 MAC 地址,以及指向所述虚拟机的出端口。

3. 如权利要求 2 所述的方法,其特征在于,该方法还包括:

 所述网络管理装置接收到虚拟机迁移的通知信息,删除源接入层设备的二层转发表中关联于迁移虚拟机的一对虚拟机表项,为迁移虚拟机重新配置虚拟 MAC 地址,在目标接入层设备的二层转发表中添加关联于迁移虚拟机的另一对虚拟机表项;所述源接入层设备连接于所述迁移虚拟机所在的源物理服务器,所述目标接入层设备连接于所述迁移虚拟机所在的目标物理服务器。

4. 如权利要求 2 所述的方法,其特征在于,该方法还包括:

 所述网络管理装置接收删除虚拟机的通知信息,删除源接入层设备的二层转发表中关联于被删除的虚拟机的一对虚拟机表项;所述源接入层设备连接于连接于所述被删除虚拟机所在的物理服务器。

5. 如权利要求 2 所述的方法,其特征在于,该方法还包括:

 所述网络管理装置接收增加虚拟机的通知信息,为新增的虚拟机配置虚拟 MAC 地址,在目标接入层设备的二层转发表中添加关联于新增的虚拟机的一对虚拟机表项;所述目标接入层设备连接于所述新增虚拟机所在的物理服务器。

6. 如权利要求 1 所述的方法,其特征在于,所述方法还包括:

 所述网络管理装置在一个接入层设备的二层转发表中配置至少一个接入设备表项;每个所述接入设备表项关联于同一数据中心内的一个其他接入层设备;每个接入设备表项包括:基于接入设备掩码的聚合虚拟 MAC 地址、接入设备掩码、以及到达所述接入设备表项所

关联的接入层设备的出接口；

其中，基于接入设备掩码的聚合虚拟 MAC 是通过接入设备掩码将接入到所述接入设备表项所关联的接入层设备的所有虚拟机的虚拟 MAC 聚合成的一个虚拟 MAC 地址，接入设备掩码的长度等于所述唯一性标识的字节数、所述网络标识的字节数以及所述接入层设备标识的字节数的和。

7. 如权利要求 1 所述的方法，其特征在于，所述方法还包括：

所述网络管理装置在一个接入层设备的二层转发表中配置至少一个数据中心表项，每个所述数据中心表项关联于一个其他数据中心；每个数据中心表项包括：基于数据中心掩码的聚合虚拟 MAC 地址、数据中心掩码以及到达一个核心层设备出接口；

其中，所述核心层设备负责与关联于所述数据中心表项的数据中心通信，所述基于数据中心掩码的聚合虚拟 MAC 地址是通过数据中心掩码将关联于所述数据中心表项的数据中心内所有虚拟机的虚拟 MAC 地址的聚合成的一个虚拟 MAC 地址；所述数据中心掩码等于所述唯一性标识的字节数与所述网络标识的字节数的和。

8. 如权利要求 1 所述的方法，其特征在于，所述方法还包括：

所述网络管理装置在一个接入层设备的二层转发表中配置至少一个网关转发表项，每个所述网关转发表项包括网关的所属 VLAN 的标识，所述网关的真实 MAC 地址，主机掩码，以及指向所述网关的出接口；

其中，所述主机掩码的长度等于虚拟机的真实 MAC 地址的长度。

9. 如权利要求 1 所述的方法，其特征在于，所述方法还包括：

所述网络管理装置在接入组播源和组播接收端的一个接入层设备的二层转发表中配置组播转发表项；所述组播转发表项包括组播组所属 VLAN 的标识，组播地址，主机掩码，以及指向所述组播组的组播树的树根的出接口和指向组播接收端的出端口；

所述网络管理装置在接入组播源的一个接入层设备的二层转发表中配置组播转发表项；所述组播转发表项包括组播组所属 VLAN 的标识，组播地址，主机掩码，以及指向所述组播组的组播树的树根的出接口；

所述网络管理装置在接入组播接收端的一个接入层设备的二层转发表中配置组播转发表项；所述组播转发表项包括组播组所属 VLAN 的标识，组播地址，主机掩码，以及指向所述组播接收端的出端口；

其中，所述主机掩码长度等于虚拟机的真实 MAC 地址的长度。

10. 如权利要求 1 所述的方法，其特征在于，所述方法还包括：

所述网络管理装置在一个核心层设备的二层转发表配置至少一个接入设备表项，每个所述接入设备表项关联于相同数据中心内的一个接入层设备；所述接入设备表项包括：基于接入设备掩码的聚合虚拟 MAC 地址、接入设备掩码以及到达所述接入设备表项所关联的接入层设备的出接口；其中，所述基于接入设备掩码的聚合虚拟 MAC 地址是通过接入设备掩码将接入到所述接入设备表项所关联的接入层设备的所有虚拟机的虚拟 MAC 地址的聚合成的一个虚拟 MAC 地址；所述接入设备掩码的长度等于所述唯一性标识的字节数、所述网络标识的字节数以及所述接入层设备标识的字节数的和。

11. 如权利要求 1 所述的方法，其特征在于，所述方法还包括：

所述网络管理装置在一个核心层设备的二层转发表配置至少一个数据中心表项，每个

所述数据中心表项关联于所述核心层设备所在数据中心之外的其他数据中心；所述数据中心表项包括：基于数据中心掩码的聚合虚拟 MAC 地址、数据中心掩码以及到达所述数据中心表项所关联的数据中心的出接口；其中，所述基于数据中心掩码的聚合虚拟 MAC 地址是通过数据中心掩码将所述数据中心表项所关联的数据中心的所有虚拟机的虚拟 MAC 地址的聚合成的一个虚拟 MAC 地址；所述数据中心掩码的长度等于所述唯一性标识的字节数与所述网络标识的字节数的和。

12. 如权利要求 1 所述的方法，其特征在于，所述方法还包括：

所述网络管理装置在一个核心层设备的二层转发表配置至少一个网关转发表项，每个所述网关转发表项包括网关所属 VLAN 的标识，所述网关的真实 MAC 地址，主机掩码，以及指向所述网关的出接口；其中，所述主机掩码的长度等于虚拟机的真实 MAC 地址的长度。

13. 如权利要求 1 所述的方法，其特征在于，该方法还包括：

所述网络管理装置接收私有地址解析协议 ARP 请求报文，根据所述 ARP 请求报文的目标端 IP 地址查询对应的虚拟 MAC 地址，将所述 ARP 请求报文的目标端 IP 地址设置为 ARP 响应报文的发送端 IP 地址，将查询到的虚拟 MAC 地址设置为所述 ARP 响应报文的发送端 MAC 地址，将所述 ARP 请求报文的发送端 IP 地址和发送端 MAC 地址设置为所述 ARP 响应报文的目标端 IP 地址和目标端 MAC 地址，将所述 ARP 响应报文封装为单播的私有 ARP 响应报文并发送；

或者，所述网络管理装置接收私有地址解析协议 ARP 请求报文，根据所述 ARP 请求报文的目标端 IP 地址查询对应的虚拟 MAC 地址，将所述 ARP 请求报文的目标端 IP 地址设置为 ARP 响应报文的发送端 IP 地址，将查询到的虚拟 MAC 地址设置为所述 ARP 响应报文的发送端 MAC 地址，将所述 ARP 请求报文的发送端 IP 地址设置为所述 ARP 响应报文的目标端 IP 地址，将所述 ARP 请求报文的发送端 IP 地址对应的虚拟 MAC 地址设置为所述 ARP 响应报文的目标 MAC 地址，将所述 ARP 响应报文封装为单播的私有 ARP 响应报文并发送。

14. 一种网络管理装置，应用于包括多个数据中心的大二层网络，其特征在于，所述网络管理装置包括：

虚拟 MAC 地址分配模块，用于根据预设的虚拟 MAC 地址配置规则为每个虚拟机设置一个虚拟媒体接入控制 MAC 地址；

所述虚拟 MAC 地址配置规则包括：每个所述虚拟 MAC 地址由唯一性标识、网络标识、接入层设备标识以及主机标识组成且字节数等于每个所述虚拟机自身的真实 MAC 地址的字节数；

位于同一个数据中心且接入到相同接入层设备的所有虚拟机的虚拟 MAC 地址具有相同的唯一性标识、相同的网络标识、相同的接入层设备标识以及不同的主机标识；位于同一个数据中心且接入到不同接入层设备的所有虚拟机的虚拟 MAC 地址具有相同的唯一性标识、相同的网络标识、不同的接入层设备标识。

15. 如权利要求 14 的网络管理装置，其特征在于，还包括：

二层转发表配置模块，用于在一个接入层设备的二层转发表中配置至少一对虚拟机表项；每对虚拟机表项关联于所述接入层设备接入的一个虚拟机；每对虚拟机表项包括第一虚拟机表项和第二虚拟机表项；

其中，第一虚拟机表项包含虚拟机所属虚拟局域网 VLAN 的标识、所述虚拟机的虚拟

MAC 地址、主机掩码、映射于所述虚拟 MAC 地址的所述虚拟机的真实 MAC 地址以及指向所述虚拟机的出端口；

所述第二虚拟机表项包括虚拟机所属 VLAN 的标识，所述虚拟机的真实 MAC 地址，主机掩码，映射于所述真实 MAC 地址的所述虚拟机的虚拟 MAC 地址，以及指向所述虚拟机的出端口。

16. 如权利要求 15 所述的网络管理装置，其特征在于，还包括：

虚拟机迁移管理模块，用于在接收到虚拟机迁移的通知信息后，指示所述二层转发表配置模块删除源接入层设备的二层转发表中关联于迁移虚拟机的一对虚拟机表项，指示所述虚拟 MAC 地址分配模块为迁移虚拟机重新配置虚拟 MAC 地址，所述二层转发表配置模块在目标接入层设备的二层转发表中添加关联于迁移虚拟机的另一对虚拟机表项；所述源接入层设备连接于所述迁移虚拟机所在的源物理服务器，所述目标接入层设备连接于所述迁移虚拟机所在的目标物理服务器。

17. 如权利要求 15 所述的网络管理装置，其特征在于，还包括：

虚拟机迁移管理模块，用于在接收到删除虚拟机的通知信息后，指示所述二层转发表配置模块删除源接入层设备的二层转发表中关联于被删除的虚拟机的一对虚拟机表项；所述源接入层设备连接于所述被删除虚拟机所在的物理服务器。

18. 如权利要求 15 所述的网络管理装置，其特征在于，还包括：

虚拟机迁移模块，用于在接收到增加虚拟机的通知信息后，指示所述虚拟 MAC 地址分配模块为新增的虚拟机配置虚拟 MAC 地址，指示所述二层转发表配置模块在目标接入层设备的二层转发表中添加关联于新增的虚拟机的一对虚拟机表项；所述目标接入层设备连接于所述新增虚拟机所在的物理服务器。

19. 如权利要求 14 所述的网络管理装置，其特征在于，还包括：

二层转发表配置模块，用于在一个接入层设备的二层转发表中配置至少一个接入设备表项；每个所述接入设备表项关联于同一数据中心内的一个其他接入层设备；每个接入设备表项包括：基于接入设备掩码的聚合虚拟 MAC 地址、接入设备掩码、以及到达所述接入设备表项所关联的接入层设备的出接口；

其中，基于接入设备掩码的聚合虚拟 MAC 是通过接入设备掩码将接入到所述接入设备表项所关联的接入层设备的所有虚拟机的虚拟 MAC 聚合成的一个虚拟 MAC 地址，接入设备掩码的长度等于所述唯一性标识的字节数、所述网络标识的字节数以及所述接入层设备标识的字节数的和。

20. 如权利要求 14 所述的网络管理装置，其特征在于，还包括：

二层转发表配置模块，用于在一个接入层设备的二层转发表中配置至少一个数据中心表项，每个所述数据中心表项关联于一个其他数据中心；每个数据中心表项包括：基于数据中心掩码的聚合虚拟 MAC 地址、数据中心掩码以及到达一个核心层设备的出接口；

其中，所述核心层设备负责与关联于所述数据中心表项的数据中心通信；所述基于数据中心掩码的聚合虚拟 MAC 地址是通过数据中心掩码将关联于所述数据中心表项的数据中心内所有虚拟机的虚拟 MAC 地址的聚合成的一个虚拟 MAC 地址；所述数据中心掩码等于所述唯一性标识的字节数与所述网络标识的字节数的和。

21. 如权利要求 14 所述的网络管理装置，其特征在于，还包括：

二层转发表配置模块,用于在一个接入层设备的二层转发表中配置至少一个网关转发表项,每个所述网关转发表项包括网关的所属 VLAN 的标识,所述网关的真实 MAC 地址,主机掩码,以及指向所述网关的出接口;

其中,所述主机掩码的长度等于虚拟机的真实 MAC 地址的长度。

22. 如权利要求 14 所述的网络管理装置,其特征在于,还包括:

二层转发表配置模块,用于在接入组播源和组播接收端的一个接入层设备的二层转发表中配置组播转发表项;所述组播转发表项包括组播组所属 VLAN 的标识,组播地址,主机掩码,以及指向所述组播组的组播树的树根的出接口和指向组播接收端的出端口;

所述网络管理装置在接入组播源的一个接入层设备的二层转发表中配置组播转发表项;所述组播转发表项包括组播组所属 VLAN 的标识,组播地址,主机掩码,以及指向所述组播组的组播树的树根的出接口;

所述网络管理装置在接入组播接收端的一个接入层设备的二层转发表中配置组播转发表项;所述组播转发表项包括组播组所属 VLAN 的标识,组播地址,主机掩码,以及指向所述组播接收端的出端口;

其中,所述主机掩码长度等于虚拟机的真实 MAC 地址的长度。

23. 如权利要求 14 所述的网络管理装置,其特征在于,还包括:

二层转发表配置模块,用于在一个核心层设备的二层转发表配置至少一个接入设备表项,每个所述接入设备表项关联于相同数据中心内的一个接入层设备;所述接入设备表项包括:基于接入设备掩码的聚合虚拟 MAC 地址、接入设备掩码以及到达所述接入设备表项所关联的接入层设备的出接口;其中,所述基于接入设备掩码的聚合虚拟 MAC 地址是通过接入设备掩码将接入到所述接入设备表项所关联的接入层设备的所有虚拟机的虚拟 MAC 地址的聚合成的一个虚拟 MAC 地址;所述接入设备掩码的长度等于所述唯一性标识的字节数、所述网络标识的字节数以及所述接入层设备标识的字节数的和。

24. 如权利要求 14 所述的网络管理装置,其特征在于,还包括:

二层转发表配置模块,用于在一个核心层设备的二层转发表配置至少一个数据中心表项,每个所述数据中心表项关联于所述核心层设备所在数据中心之外的其他数据中心;所述数据中心表项包括:基于数据中心掩码的聚合虚拟 MAC 地址、数据中心掩码以及到达所述数据中心表项所关联的数据中心的出接口;其中,所述基于数据中心掩码的聚合虚拟 MAC 地址是通过数据中心掩码将所述数据中心表项所关联的数据中心的所有虚拟机的虚拟 MAC 地址的聚合成的一个虚拟 MAC 地址;所述数据中心掩码的长度等于所述唯一性标识的字节数与所述网络标识的字节数的和。

25. 如权利要求 14 所述的网络管理装置,其特征在于,还包括:

二层转发表配置模块,用于在一个核心层设备的二层转发表配置至少一个网关转发表项,每个所述网关转发表项包括网关所属 VLAN 的标识,所述网关的真实 MAC 地址,主机掩码,以及指向所述网关的出接口;其中,所述主机掩码的长度等于虚拟机的真实 MAC 地址的长度。

26. 如权利要求 14 所述的网络管理装置,其特征在于,还包括:

ARP 处理模块,用于接收私有地址解析协议 ARP 请求报文,根据所述 ARP 请求报文的目标端 IP 地址查询对应的虚拟 MAC 地址,将所述 ARP 请求报文的目标端 IP 地址设置为 ARP 响

应报文的发送端 IP 地址,将查询到的虚拟 MAC 地址设置为所述 ARP 响应报文的发送端 MAC 地址,将所述 ARP 请求报文的发送端 IP 地址设置为所述 ARP 响应报文的目标端 IP 地址,将所述 ARP 请求报文的发送端 MAC 地址设置为所述 ARP 响应报文的目标 MAC 地址,将所述 ARP 响应报文封装为单播的私有 ARP 响应报文并发送;

或者,用于接收私有 ARP 请求报文,根据所述 ARP 请求报文的目标端 IP 地址查询对应的虚拟 MAC 地址,将所述 ARP 请求报文的目标端 IP 地址设置为 ARP 响应报文的发送端 IP 地址,将查询到的虚拟 MAC 地址设置为所述 ARP 响应报文的发送端 MAC 地址,将所述 ARP 请求报文的发送端 IP 地址设置为所述 ARP 响应报文的目标端 IP 地址,将所述 ARP 请求报文的发送端 IP 地址对应的虚拟 MAC 地址设置为所述 ARP 响应报文的目标 MAC 地址,将所述 ARP 响应报文封装为单播的私有 ARP 响应报文并发送。

一种物理链路地址管理方法及装置

技术领域

[0001] 本发明涉及通信技术领域，尤其涉及一种物理链路地址管理方法及装置。

背景技术

[0002] 服务器虚拟化技术可以在一台物理的服务器上虚拟出几十个甚至上百个虚拟机(Virtual Machine, VM)，以提升服务器的利用率。为了提升服务器的高可用性(High Availability, HA)，需要 VM 能够在同一接入设备的不同端口之间以及不同接入设备之间迁移。不同的标准组织制定了不同标准化协议，如多链路透明互联(Transparent Interconnection of Lots of Links, Trill)协议、最短路径桥(Shortest Path Bridging, SPB)协议等等，用以构建大二层网络(Large scale layer-2network)，实现 VM 迁移。

[0003] 以包含多个数据中心的大二层网络(Very Large Layer-2Network)为例，大二层组网技术分为数据中心内部的大二层网络技术和数据中心之间互联的大二层网络技术。前者可以实现单个数据中心内单台接入设备不同端口之间的 VM 迁移以及不同接入设备之间的 VM 迁移，后者可以实现不同数据中心的接入设备之间的 VM 迁移。

[0004] 目前，数据中心内的大二层网络的接入层设备能够提供超过 12000 个以上的万兆口，用于连接 12000 台万兆物理服务器。单台万兆物理服务器又能够虚拟 200 个以上的 VM。数据中心内的大二层网络的 12000 台万兆物理服务器能够虚拟 2.4 兆(M)个以上的 VM，导致单个数据中心内大二层网络的 VM 所需的 MAC(Media Access Control, 介质访问控制)地址数量高达 2.4M 个以上。基于多租户(multi tenant)模型的数据中心内的 VM 数量更大，需要的 MAC 地址将更多。

[0005] 相应地，在数据中心内部的大二层网络内，接入设备以及网关设备需要在二层转发表中学习大量的 MAC 地址，从而对自身的二层转发表的表项数目产生巨大的压力。

[0006] 交换机在二层转发表中记录 MAC 地址表项的方式类似于网络设备在三层转发表中记录主机路由表项的方式，导致交换机设备的二层转发表的表项数目庞大，难以适应未来大规模数据中心的需求。通常，交换机学习 MAC 地址采用 HASH(哈希)方式，而在 MAC 地址学习数量大的情况下，交换机学习 MAC 地址会发生 HASH 冲突，导致交换机记录的 MAC 地址转发表项数量达不到规格定义的 MAC 地址转发表项数量的最大值。

发明内容

[0007] 本发明的目的在于提供一种物理链路地址表管理方法及装置。为实现该目的，本发明提供了以下技术方案：

[0008] 一种物理链路地址管理方法，应用于包括多个数据中心的大二层网络，该方法包括：

[0009] 网络管理装置根据预设的虚拟 MAC 地址配置规则为每个虚拟机设置一个虚拟媒体接入控制 MAC 地址；

[0010] 所述虚拟 MAC 地址配置规则包括：每个所述虚拟 MAC 地址由唯一性标识、网络标

识、接入层设备标识以及主机标识组成且字节数等于每个所述虚拟机自身的真实 MAC 地址的字节数；

[0011] 位于同一个数据中心且接入到相同接入层设备的所有虚拟机的虚拟 MAC 地址具有相同的唯一性标识、相同的网络标识、相同的接入层设备标识以及不同的主机标识；位于同一个数据中心且接入到不同接入层设备的所有虚拟机的虚拟 MAC 地址具有相同的唯一性标识、相同的网络标识、不同的接入层设备标识。

[0012] 一种网络管理装置，应用于包括多个数据中心的大二层网络，所述网络管理装置包括：

[0013] 虚拟 MAC 地址分配模块，用于根据预设的虚拟 MAC 地址配置规则为每个虚拟机设置一个虚拟媒体接入控制 MAC 地址；

[0014] 所述虚拟 MAC 地址配置规则包括：每个所述虚拟 MAC 地址由唯一性标识、网络标识、接入层设备标识以及主机标识组成且字节数等于每个所述虚拟机自身的真实 MAC 地址的字节数；

[0015] 位于同一个数据中心且接入到相同接入层设备的所有虚拟机的虚拟 MAC 地址具有相同的唯一性标识、相同的网络标识、相同的接入层设备标识以及不同的主机标识；位于同一个数据中心且接入到不同接入层设备的所有虚拟机的虚拟 MAC 地址具有相同的唯一性标识、相同的网络标识、不同的接入层设备标识。

[0016] 本发明的上述实施例中，网络管理装置为大二层网络中各数据中心的虚拟机分配虚拟 MAC 地址，且同一接入层设备连接的物理服务器上的所有虚拟机的虚拟 MAC 地址能够被聚合一个虚拟 MAC 地址，以及同一个数据中心内各接入层设备连接的物理服务器的所有虚拟机的虚拟 MAC 地址能够被聚合为一个虚拟 MAC 地址，以及基于各聚合的虚拟 MAC 地址进行二层转发表项的配置，减少了二层转发表中表项的数量。

附图说明

[0017] 图 1 为 Trill 组网示意图；

[0018] 图 2 为本发明实施例中的 VM 迁移的示意图；

[0019] 图 3A 为本发明实施例中数据中心内相同 VLAN 内报文转发的示意图；

[0020] 图 3B 为本发明实施例提供的私有 ARP 报文的示意图；

[0021] 图 4 为本发明实施例中组播报文转发的示意图；

[0022] 图 5 为本发明实施例中数据中心内的报文转发至外网的示意图；

[0023] 图 6 为本发明实施例中数据中心内不同 VLAN 报文转发的示意图；

[0024] 图 7 为本发明实施例中的不同数据中心之间的二层转发的示意图；

[0025] 图 8 为本发明实施例提供的网络管理装置的结构示意图。

具体实施方式

[0026] 本发明实施例不限定数据中心内和数据中心之间使用何种大二层技术。单个数据中心内可使用 Trill、SPB 等大二层网络技术，各数据中心互联可使用 MAC over IP (如 OTV (Overlay Transport Virtualization, 覆盖传输虚拟化) 协议、EVI (Ethernet Virtualization Interconnection, 以太网虚拟互联) 协议、VPLS (Virtual Private LAN

Service, 虚拟专用局域网业务) 协议等大二层网络技术。

[0027] 本发明实施例以四个数据中心(Data Centre, DC) 互联构成的大二层网络的架构为例进行描述。该大二层网络中还包括与四个数据中心连接的网络管理装置(也即 network management plane, 网络管理平面)以及服务器管理装置(也即 VM management plane, VM 管理装置或 VM 管理平面)。该架构中, 各数据中心内的大二层网络采用 Trill 技术, 各数据中心间的大二层网络采用 MACover IP 技术。

[0028] 图 1 示出了 Trill 技术构建的数据中心的大二层网络的架构。其它数据中心的大二层网络架构类似于图 1 数据中心大二层网络架构。

[0029] 如图 1 所示, 数据中心 1 的大二层网络包括核心层、接入层。leaf1、leaf2、leaf3、leaf4 是位于接入层设备, core1 和 core2 是核心层设备。

[0030] 数据中心 1 的接入层的 leaf1、leaf2、leaf3、leaf4 以及核心层的 core1、core2 运行 Trill 协议, 这些运行 Trill 协议的设备称为路由桥(Routing Bridge, RBridge), 构成了 Trill 网络。各路由桥之间通过链路状态协议获取 Trill 网络拓扑。每个路由桥使用最短路径树算法生成从本路由桥到达 Trill 网络里的其它各个路由桥的路由转发表(称之为 Trill 路由表)。

[0031] 本发明实施例中, “网络管理装置”按照虚拟 MAC 编码规则为各数据中心内的每个 VM 配置虚拟 MAC 地址。每个虚拟 MAC 地址为 6 字节的 2 进制数, 并包含如下标识:

[0032] 唯一性标识(1 字节): 与现有已分配的 OUI (Organizationally unique identifier, 组织唯一标识符) 不冲突, 可以使用 OUI 尚未被分配的标识符, 例如: ED, 22 等。

[0033] Data Centre ID (1 字节): 数据中心标识或称网络标识。

[0034] Device ID (2 字节): 接入层设备的标识。

[0035] Host ID (2 字节): VM 的标识, 也即主机标识。同一个物理接入层路由桥连接的物理服务器上承载的 VM 的 host ID 不能相同。不同物理接入层设备连接物理服务器上承载的 VM 的 host ID 可以相同。

[0036] 基于以上 VM 的虚拟 MAC 地址编码规则, 本发明实施例定义如下掩码:

[0037] 主机掩码(Host mask): ff-ff-ff-ff-ff-ff。

[0038] 接入设备掩码(access device mask): ff-ff-ff-ff-00-00。

[0039] 数据中心掩码(data centre mask): ff-ff-00-00-00-00。

[0040] 网络管理装置可通过运行批量配置工具, 为全网的 VM 配置虚拟 MAC 地址。在配置虚拟 MAC 地址过程中, 网络管理装置从 VM 管理装置获取包括整网 VM 的信息表以及物理设备的连接关系。在该表基础上, 根据上述虚拟 MAC 地址编码规则, 在该表中添加虚拟 MAC 地址, 并维护该表。如表 1 所示, 网络管理装置维护的整网设备及 VM 的信息表至少包含以下信息(其中仅示出了数据中心 1 的相关配置信息):

[0041] 表 1:

[0042]

Device	Nickname	IP Address (MAC over IP)	Port	VLAN ID	IP address of VM	MAC address of VM	Virtual MAC address of VM	access device mask based virtual MAC Address	data centre mask based virtual MAC Address

[0043]

leaf1	DC1_leaf1	IP1	Port1	1	1.1.1.1	00-11-11-11-11-11	ED-01-00-01-00-01		
	DC1_leaf1	IP1	Port1	1	1.1.1.2	00-E0-FC-03-42-24	ED-01-00-01-00-02		
	DC1_leaf1	IP1	Port1	1	1.1.1.3	00-14-2A-EB-74-2F	ED-01-00-01-00-03		
	DC1_leaf1	IP1	Port2	2	2.2.2.1	00-05-5B-A4-6B-28	ED-01-00-01-00-04		
	DC1_leaf1	IP1	Port2	2	2.2.2.2	00-0F-e2-0F-9a-86	ED-01-00-01-00-05	ED-01-00-01-00-00	
	DC1_leaf1	IP1	Port2	2	2.2.2.3	00-0C-76-0A-17-2D	ED-01-00-01-00-06		
	DC1_leaf1	IP1	Port3	3	3.3.3.1	00-0D-88-F6-44-C1	ED-01-00-01-00-07		
	DC1_leaf1	IP1	Port3	3	3.3.3.2	00-0D-88-F7-9F-7D	ED-01-00-01-00-08		
	DC1_leaf1	IP1	Port3	3	3.3.3.3	00-0D-88-F7-B0-90	ED-01-00-01-00-09		
leaf2	DC1_leaf2	IP1	Port1	1	1.1.1.4	00-22-22-22-22-22	ED-01-00-02-00-01		
	DC1_leaf2	IP1	Port1	1	1.1.1.5	00-6B-28-07-44-3F	ED-01-00-02-00-02	ED-01-00-00-00	
	DC1_leaf2	IP1	Port1	1	1.1.1.6	00-14-3A-EB-84-2F	ED-01-00-02-00-03		
	DC1_leaf2	IP1	Port2	2	2.2.2.4	00-05-6B-A4-6B-38	ED-01-00-02-00-04		
	DC1_leaf2	IP1	Port2	2	2.2.2.5	00-0D-88-F7-B0-94	ED-01-00-02-00-05	ED-01-00-02-00-00	
	DC1_leaf2	IP1	Port2	2	2.2.2.6	00-0D-98-F8-4E-88	ED-01-00-02-00-06		
	DC1_leaf2	IP1	Port3	3	3.3.3.4	04-37-1A-44-55-66	ED-01-00-02-00-07		
	DC1_leaf2	IP1	Port3	3	3.3.3.5	06-22-23-AA-BB-CC	ED-01-00-02-00-08		
	DC1_leaf2	IP1	Port3	3	3.3.3.6	08-53-26-3B-7C-FD	ED-01-00-02-00-09		
leaf3	DC1_leaf3	IP1	Port1	1	1.1.1.7	00-06-25-FD-32-EB	ED-01-00-03-00-01		
	DC1_leaf3	IP1	Port1	1	1.1.1.8	00-1D-A1-75-28-70	ED-01-00-03-00-02	ED-01-00-03-00-00	
	DC1_leaf3	IP1	Port1	1	1.1.1.9	00-09-92-01-CA-D7	ED-01-00-03-00-03		
	DC1_leaf3	IP1	Port2	2	2.2.2.7	00-25-9C-2F-63-FE	ED-01-00-03-00-04		

[0044]

	DC1_leaf3	IP1	Port2	2	2.2.2.8	FC-FB-FB-11-22-33	ED-01-00-03-00-05	
	DC1_leaf3	IP1	Port2	2	2.2.2.9	F8-83-88-47-77-98	ED-01-00-03-00-06	
	DC1_leaf3	IP1	Port3	3	3.3.3.7	10-11-23-5A-8B-CF	ED-01-00-03-00-07	
	DC1_leaf3	IP1	Port3	3	3.3.3.8	28-47-6c-66-77-88	ED-01-00-03-00-08	
	DC1_leaf3	IP1	Port3	3	3.3.3.9	3C-4B-5A-99-3D-57	ED-01-00-03-00-09	
leaf4	DC1_leaf4	IP1	Port1	1	1.1.1.10	20-47-FC-13-34-57	ED-01-00-04-00-01	ED-01-00-04-00-00
	DC1_leaf4	IP1	Port1	1	1.1.1.11	FC-FB-FB-01-33-45	ED-01-00-04-00-02	
	DC1_leaf4	IP1	Port1	1	1.1.1.12	24-1A-8C-05-55-FF	ED-01-00-04-00-03	
	DC1_leaf4	IP1	Port2	2	2.2.2.10	24-37-EF-AA-97-A8	ED-01-00-04-00-04	
	DC1_leaf4	IP1	Port2	2	2.2.2.11	00-00-01-17-4d-F9	ED-01-00-04-00-05	
	DC1_leaf4	IP1	Port2	2	2.2.2.12	00-E0-FC-37-45-98	ED-01-00-04-00-06	
	DC1_leaf4	IP1	Port3	3	3.3.3.10	58-66-BA-03-27-99	ED-01-00-04-00-07	
	DC1_leaf4	IP1	Port3	3	3.3.3.11	C4-CA-D9-70-90-58	ED-01-00-04-00-08	
	DC1_leaf4	IP1	Port3	3	3.3.3.12	00-0D-EF-33-44-55	ED-01-00-04-00-09	
core1	DC1_core1		L3 interface	1	1.1.1.100	00-E0-FC-11-11-11		
			L3 interface	2	2.2.2.100	00-E0-FC-22-22-22		
core2	DC1_core2		L3 interface	3	3.3.3.100	00-E0-FC-33-33-33		

[0045] “device”(设备)分别是 VM 所在的物理服务器连接的接入层设备以及所在数据中心的核心层设备,Nickname(昵称)分别是 VM 所在的物理服务器连接的接入层设备在 Trill 网络内的转发标识以及 VM 所在的物理服务器所属数据中心的核心层设备在 Trill 网络内的转发标识;“Port”是接入层设备连接 VM 所在物理服务器的端口,“VLAN ID”(虚拟局域网标识)描述了 VM 所在虚拟机局域网的标识。

[0046] “IP address of VM”(虚拟机 IP 地址)是各虚拟机的 IP 地址;“MAC address of VM”为 VM 的真实 MAC 地址,“Virtual MAC address of VM”为 VM 的虚拟 MAC 地址。

00-E0-FC-11-11-11、00-E0-FC-22-22-22、00-E0-FC-33-33-33 分别为配置在核心层路由桥 core1 和 core2 三层接口上的 VLAN 网关地址, 其中 00-E0-FC-11-11-11 为 VLAN1 的网关 MAC 地址, 00-E0-FC-22-22-22 为 VLAN2 的网关地址, 00-E0-FC-33-33-33 为 VLAN3 的网关地址。

[0047] 将每个 VM 的虚拟 MAC 地址与接入设备掩码 ff-ff-ff-ff-00-00 进行逻辑与 (and) 运算, 运算结果同为 48 位的聚合 MAC 地址 ED-01-00-01-00-00, 因而接入同一接入层设备的 VM 的虚拟 MAC 地址聚合为基于接入设备掩码的虚拟 MAC 地址 (access device mask based virtual MAC Address)。即, 接入 leaf1 的所有 VM 的虚拟 MAC 地址可以聚合为 ED-01-00-01-00-00。接入到 leaf2、leaf3 以及 leaf4 的 VM 的虚拟 MAC 地址分别聚合为 ED-01-00-02-00-00、ED-01-00-03-00-00 以及 ED-01-00-04-00-00。按照相同的虚拟 MAC 地址编码规则, 其他数据中心的 VM 的虚拟 MAC 地址可以基于接入设备掩码聚合, 在此不再举例。

[0048] 将每个 VM 的虚拟 MAC 地址与数据中心掩码 ff-ff-00-00-00-00 进行逻辑与 (and) 运算, 运算结果同为 48 位的聚合 MAC 地址 ED-01-00-00-00-00, 从而同一数据中心内的 VM 的虚拟 MAC 地址能够聚合为基于数据中心掩码的虚拟 MAC 地址 (Data Centre Mask based Virtual MAC address)。即, 数据中心 1 内的 VM 的虚拟 MAC 地址能够聚合为 ED-01-00-00-00-00。同样地, 按照相同的虚拟 MAC 地址编码, 数据中心 2、数据中心 3、数据中心 4 的 VM 的虚拟 MAC 地址基于数据中心掩码分别聚合为 ED-02-00-00-00-00、ED-03-00-00-00-00 以及 ED-04-00-00-00-00。

[0049] 需要说明的是, 不同数据中心的 VM 的唯一性标识和数据中心标识可以相同, 也可以不同。只需要保证同一个数据中心内的 VM 的唯一性标识和数据中心标识相同, 确保同一数据中心的 VM 的聚合关系不被破坏, 确保接入到相同接入层设备的 VM 的聚合关系不被破坏。

[0050] 在报文转发前, 网络管理装置根据其所维护的整网设备以及 VM 信息, 在各数据中心的接入层路由桥以及核心层路由桥上配置二层转发表。

[0051] 表 2.1 示出了配置于数据中心 1 的接入层路由桥 leaf1 上的二层转发表。

[0052] 表 2.1 :

[0053]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
1	00-11-11-11-11-11	ff-ff-ff-ff-ff-ff	ED-01-00-01-00-01	Port1
1	ED-01-00-01-00-01	ff-ff-ff-ff-ff-ff	00-11-11-11-11-11	Port1
1	00-E0-FC-03-42-24	ff-ff-ff-ff-ff-ff	ED-01-00-01-00-02	Port1
1	ED-01-00-01-00-02	ff-ff-ff-ff-ff-ff	00-E0-FC-03-42-24	Port1
1	00-14-2A-EB-74-2F	ff-ff-ff-ff-ff-ff	ED-01-00-01-00-03	Port1
1	ED-01-00-01-00-03	ff-ff-ff-ff-ff-ff	00-14-2A-EB-74-2F	Port1
2	00-05-5B-A4-6B-28	ff-ff-ff-ff-ff-ff	ED-01-00-01-00-04	Port2
2	ED-01-00-01-00-04	ff-ff-ff-ff-ff-ff	00-05-5B-A4-6B-28	Port2
2	00-0F-E2-0F-9A-86	ff-ff-ff-ff-ff-ff	ED-01-00-01-00-05	Port2
2	ED-01-00-01-00-05	ff-ff-ff-ff-ff-ff	00-0F-E2-0F-9A-86	Port2
2	00-0C-76-0A-17-2D	ff-ff-ff-ff-ff-ff	ED-01-00-01-00-06	Port2
2	ED-01-00-01-00-06	ff-ff-ff-ff-ff-ff	00-0C-76-0A-17-2D	Port2
3	00-0D-88-F6-44-C1	ff-ff-ff-ff-ff-ff	ED-01-00-01-00-07	Port3
3	ED-01-00-01-00-07	ff-ff-ff-ff-ff-ff	00-0D-88-F6-44-C1	Port3
3	00-0D-88-F7-9F-7D	ff-ff-ff-ff-ff-ff	ED-01-00-01-00-08	Port3
3	ED-01-00-01-00-08	ff-ff-ff-ff-ff-ff	00-0D-88-F7-9F-7D	Port3
3	00-0D-88-F7-B0-90	ff-ff-ff-ff-ff-ff	ED-01-00-01-00-09	Port3
3	ED-01-00-01-00-09	ff-ff-ff-ff-ff-ff	00-0D-88-F7-B0-90	Port3
VLAN unaware	ED-01-00-02-00-00	ff-ff-ff-ff-00-00		DC1_leaf2
VLAN unaware	ED-01-00-03-00-00	ff-ff-ff-ff-00-00		DC1_leaf3
VLAN	ED-01-00-04-00-00	ff-ff-ff-ff-00-00		DC1_leaf4

[0054]

unaware				
VLAN unaware	ED-02-00-00-00-00	ff-ff-00-00-00-00		DC1_core1
VLAN unaware	ED-03-00-00-00-00	ff-ff-00-00-00-00		DC1_core1
VLAN unaware	ED-04-00-00-00-00	ff-ff-00-00-00-00		DC1_core1
1	00-E0-FC-11-11-11	ff-ff-ff-ff-ff-ff		DC1_core1
2	00-E0-FC-22-22-22	ff-ff-ff-ff-ff-ff		DC1_core1
3	00-E0-FC-33-33-33	ff-ff-ff-ff-ff-ff		DC1_core2

[0055] Port1、Port2、Port3 分别为 leaf1 连接 VM 所在服务器的端口 ;DC1_leaf2、DC1_leaf3、DC1_leaf4 分别为数据中心 1 中的相应接入层路由桥的 nickname ;DC1_core1 为数据中心 1 的 core1 的 nickname, DC1_core2 为数据中心 1 的 core2 的 nickname。00-E0-FC-11-11-11、00-E0-FC-22-22-22、00-E0-FC-33-33-33 分别为 VLAN1、VLAN2、VLAN3 的网关 MAC 地址。

[0056] 表 2.1 中, core1 被配置为负责数据中心 1 与其它数据中心(数据中心 2、数据中心 3、数据中心 4)之间的流量转发, 基于数据中心掩码的虚拟 MAC 地址的 3 个表项(即初始 MAC 地址是 ED-02-00-00-00-00、ED-03-00-00-00-00、ED-04-00-00-00-00 的三个表项)中的出接口配置为 core1 的 nickname (DC1_core1)。

[0057] 当 core1 被配置为负责数据中心 1 与数据中心 2、数据中心 3 之间的流量转发时, 则将初始 MAC 地址是 ED-02-00-00-00-00、ED-03-00-00-00-00 的两个表项的出接口将被配置为 core1 的 nickname (DC1_core1), core2 被配置为负责数据中心 1 与数据中心 4 之间的流量转发时, 则初始 MAC 地址为 ED-04-00-00-00-00 的表项中的出接口配置为 core2 的 nickname (DC1_core2)。

[0058] 当 core1 和 core2 堆叠构成一台虚拟设备, 该虚拟设备被配置为负责数据中心 1 与其他三个数据中心之间的流量转发时, 则将初始 MAC 地址为 ED-02-00-00-00-00、ED-03-00-00-00-00 和 ED-04-00-00-00-00 的三个表项的出接口均配置为虚拟设备的 nickname。

[0059] 进一步的, 针对组播业务, 还需要在接入层路由桥设备配置相应组播转发表项。比如, 在数据中心 1 内, 组播组的组播树根为 core1(nickname 为 DC1_core1), 对于 VLAN1 内的组播组 1 (组播地址为 01-00-5E-XX-XX-XX), 作为该组播组的组播源的 VM 所在的物理服务器连接于 leaf1, 作为组播组的组播接收端 VM 所在的物理服务器连接于 leaf3 的 Port1, 另一个作为组播组的组播接收端 VM 所在物理服务器连接于 leaf4 的 Port1. 则 leaf1、leaf3、leaf4 上的二层转发表中的相应组播转发表项包括 :

[0060] 表 2.2 :leaf1 上二层转发表中的组播转发表项

[0061]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
1	01-00-5E-XX-XX-XX	ff-ff-ff-ff-ff-ff		DC1_core1

[0062] 表 2.3 :leaf3 上二层转发表中的组播转发表项

[0063]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
1	01-00-5E-XX-XX-XX	ff-ff-ff-ff-ff-ff		Port1

[0064] 表 2.4 :leaf4 上二层转发表中的组播转发表项

[0065]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
1	01-00-5E-XX-XX-XX	ff-ff-ff-ff-ff-ff		Port1

[0066] 作为组播组的接收端的 VM 所在的服务器通过 Port2 连接到 leaf1，则在表 2.2 所示的组播转发表项的出接口中增加 Port2。

[0067] 数据中心 2、数据中心 3、数据中心 4 的接入层路由桥的二层转发表的组播转发表项的配置方式与表 2.1- 表 2.4 中数据中心的接入层设备的组播转发表项的配置方式相同。

[0068] core1 的二层转发表至少包括表 2.5 所示的表项：

[0069] 表 2.5

[0070]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
VLAN unaware	ED-01-00-01-00-00	ff-ff-ff-ff-00-00		DC1_leaf1
VLAN unaware	ED-01-00-02-00-00	ff-ff-ff-ff-00-00		DC1_leaf2
VLAN unaware	ED-01-00-03-00-00	ff-ff-ff-ff-00-00		DC1_leaf3
VLAN unaware	ED-01-00-04-00-00	ff-ff-ff-ff-00-00		DC1_leaf4
VLAN unaware	ED-02-00-00-00-00	ff-ff-00-00-00-00		IP2
VLAN unaware	ED-03-00-00-00-00	ff-ff-00-00-00-00		IP3
VLAN unaware	ED-04-00-00-00-00	ff-ff-00-00-00-00		IP4
1	00-E0-FC-11-11-11	ff-ff-ff-ff-ff-ff		L3
2	00-E0-FC-22-22-22	ff-ff-ff-ff-ff-ff		L3
m	Next-hop MAC	ff-ff-ff-ff-ff-ff		Port m

[0071] core2 的二层转发表至少包含表 2.6 所示的表项：

[0072] 表 2.6

[0073]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
VLAN unaware	ED-01-00-01-00-00	ff-ff-ff-ff-00-00		DC1_leaf1
VLAN unaware	ED-01-00-02-00-00	ff-ff-ff-ff-00-00		DC1_leaf2
VLAN unaware	ED-01-00-03-00-00	ff-ff-ff-ff-00-00		DC1_leaf3
VLAN unaware	ED-01-00-04-00-00	ff-ff-ff-ff-00-00		DC1_leaf4
3	00-E0-FC-33-33-33	ff-ff-ff-ff-ff-ff		L3
n	Next-hop MAC'	ff-ff-ff-ff-ff-ff		Port n

[0074] 其中,DC1_leaf1、DC1_leaf2、DC1_leaf3、DC1_leaf4 分别为数据中心 1 中 leaf1—leaf4 的 nickname ;IP2、IP3、IP4 是根据协议定义的数据中心 2、数据中心 3 和数据中心 4 的 IP 地址,数据中心 1 的 core1 向这三个数据中心发送数据时,可将这些 IP 地址作为 MAC over IP 的隧道(tunnel)的目的 IP 地址;相应地,IP1 是数据中心 1 的 IP 地址,其他三个数据中心的核心层设备向数据中心 1 发送数据时,可将 IP1 作为 MAC over IP 隧道的目的 IP 地址。00-E0-FC-11-11-11、00-E0-FC-22-22-22、00-E0-FC-33-33-33 分别为 VLAN1、VLAN2、VLAN3 的网关 MAC 地址;L3 为三层转发标识,用于表示对目的 MAC 地址为这三个 MAC 地址的以太网报文执行三层转发。

[0075] 数据中心 2、数据中心 3、数据中心 4 中的核心层路由桥的二层转发表,与表 2.5 或 2.6 所示的二层转发表的配置方式相同,本实施例不再详细描述。

[0076] 为支持到外网的 VLAN(如 VLAN m)的报文转发,表 2.5 中 core1 的二层转发表中包含对应的表项(表 2.5 中的最后一行),该表项包括:VLAN 标识 m、本路由桥到 VLAN m 的下一跳设备的 MAC 地址 Next-hop MAC、出端口 Port m。core1 根据 ARP(Address Resolution Protocol,地址解析协议)报文学习下一跳设备的 MAC 地址。core1 和 core2 的下一跳设备不同,因此 core2 学习到的下一跳设备表项(表 2.6 的最后一行所示的表项)包括:VLAN 标识 n,本路由桥到 VLAN n 的下一跳设备的 MAC 地址 Next-hop MAC',出端口 Port n。

[0077] 基于图 1 所示的组网架构,图 2 所示为本发明实施例中 VM 迁移的示意图。

[0078] VM 从连接到 leaf3 的物理服务器迁移到连接于 leaf4 的物理服务器。虚拟机的真实 MAC 地址和 IP 地址不变。

[0079] 迁移前,虚拟机所在的物理服务器视为源物理服务器,连接于源物理服务器 leaf3 可视为迁移主机的源接入设备。迁移后,VM 所在的物理服务器视为目标物理服务器,Leaf4 连接于目标物理服务器可视为迁移主机的目标接入设备。

[0080] leaf3 将 VM 迁移事件通知给网络管理装置(network management plane)。根据 802.1Qbg 定义的 VDP(VSI Discovery and Configuration Protocol, VSI 发现和配置协议),物理交换机可感知 VM 的迁移过程,并将变化信息通知给网络管理装置。

[0081] 网络管理装置根据 leaf3 通知的信息,在 leaf3 接入的 VM 信息中,删除迁移的 VM。网络管理装置按照虚拟 MAC 地址编码规则为迁移到 leaf4 的 VM 设置新的虚拟 MAC 地址,在 leaf4 接入的 VM 信息中增加迁移的 VM 信息,以保证对外的聚合关系不受破坏。

[0082] 网络管理装置在 leaf4 的二层转发表中添加关联于迁移的 VM 的真实 MAC 地址和虚拟 MAC 地址的二层转发表项,并在 leaf3 的二层转发表中删除关联于迁移的 VM 的真实 MAC 地址和虚拟 MAC 地址的二层转发表项。

[0083] 本步骤的目的在于减少无效表项的占用。本步骤还可采用其它实施方式,如将 leaf3 上迁出的 VM 的二层转发表项标记为无效。

[0084] 迁移的 VM 广播免费 ARP 报文。leaf4 收到免费 ARP 报文后,根据二层转发表将该免费 ARP 报文以太网头的“源 MAC 地址(Source MAC address)”和“发送端 MAC 地址(Sender MAC Address)”信息都替换为 VM 的虚拟 MAC 地址,广播收到的 ARP 报文。leaf4 在相同 VLAN 的其他端口广播免费 ARP 报文,将免费 ARP 报文进行 Trill 封装,在 Trill 网络内广播。网关和相同 VLAN 内的其他虚拟机根据免费 ARP 报文学习 ARP 表项,将迁移的 VM 的 IP 地址对应的原虚拟 MAC 地址刷新为新分配的虚拟 MAC 地址。

[0085] 按照 ARP 协议,设备发送免费 ARP 报文时,将设备的真实 MAC 地址写入发送端 MAC 地址;其他设备收到免费 ARP 报文时,根据“发送端 IP 地址”和“发送端 MAC 地址”学习 ARP 表项。

[0086] 如果 leaf4 不修改免费 ARP 报文的发送端 MAC 地址,网关以及相同 VLAN 内的其他 VM 学习的 ARP 表项中的 MAC 地址是迁移的 VM 的真实 MAC 地址。相同 VLAN 的其他 VM 或网关向迁移的 VM 发送以太网报文,将迁移的 VM 的真实 MAC 地址作为目的 MAC 地址。Leaf4 收到的以太网报文的目的 MAC 地址是迁移的 VM 的真实 MAC 地址时,根据二层转发表将报文的目的 MAC 映射为迁移 VM 的虚拟 MAC 地址并发送给迁移的 VM。迁移的 VM 收到目的 MAC 地址为自身虚拟 MAC 地址的以太网报文时,执行丢弃,导致了报文丢失。

[0087] 如图 2 所示,IP 地址为 1.1.1.7 的 VM 从源接入设备连接的物理服务器迁移到目标接入设备连接的物理服务器后,网络管理装置为迁移的虚拟机重新分配的虚拟 MAC 地址为 ED-01-01-04-00-04。网络管理装置在 leaf4 配置迁移的虚拟机关联的转发表项。当迁移的虚拟机发送免费 ARP 报文时,leaf4 根据配置的转发表项,替换免费 ARP 报文的以太网头的源 MAC 地址以及免费 ARP 报文的发送端 MAC 地址。网关以及同 VLAN 的其他 VM 学习的 ARP 表项中,IP 地址 1.1.1.7 对应的 MAC 地址为 ED-01-01-04-00-04。

[0088] 在另一场景下,物理服务器的任一 VM 被删除时,接入层设备将该 VM 事件通知给网络管理装置。网络管理装置根据接入层设备的通知,删除被删除的 VM 的信息,删除关联于被删除的 VM 的一对二层转发表项。(图 2 未示)

[0089] 在另一场景下,物理服务器上新增加 VM 时,接入层设备将该 VM 事件通知给网络管理装置。网络管理装置按照虚拟 MAC 地址编码规则为新增 VM 设置虚拟 MAC 地址,在该接入层设备的 VM 信息中增加新增 VM 的信息,以保证对外的聚合关系不受破坏。

[0090] 网络管理装置在接入层设备的二层转发表中配置关联于新增 VM 的一对二层转发表项。新增的 VM 广播免费 ARP 报文,接入层设备根据配置的二层转发表项替换免费 ARP 报文的源 MAC 地址和发送端 MAC 地址,在相同 VLAN 内和 Trill 网络内广播 ARP 报文,从而使网关以及同 VLAN 的其他 VM 学习到 ARP 表项中记录新增 VM 的 IP 地址和虚拟 MAC 地址(图 2 未示)。

[0091] 本发明实施例对大二层网络的报文转发机制进行了相应改进,这些改进主要包括以下几个方面:

[0092] (1) 路由桥根据以太网报文的源 MAC 地址(相较于 Trill 封装的报文,该源 MAC 地址为内层源 MAC 地址),在二层转发表查找到匹配表项且该表项中包含映射 MAC 地址(Mapped MAC address),则将以太网报文的源 MAC 地址替换为该映射 MAC 地址。同理,路由桥根据以太网报文的目的 MAC 地址(相较于 Trill 封装的报文,该目的 MAC 地址为内层的目的 MAC 地址),在二层转发表查找到匹配表项且该表项中包含映射 MAC 地址(Mapped MAC address),则将以太网报文的目的 MAC 地址替换为该映射 MAC 地址。

[0093] (2) 路由桥支持基于掩码的 MAC 地址查找方式。路由桥用以太网头(Trill 报文的内层以太网头)的源 MAC 地址或目的 MAC 地址与各表项的“mask”进行“逻辑与”运算,再将运算结果与相应表项中的初始 MAC 地址“Initial MAC address”比较,如果一致,则确定查找到的匹配表项。

[0094] (3) 路由桥支持已知 VLAN(VLAN aware)转发以及未知 VLAN(VLAN unaware)转

发。VLAN aware 方式利用以太网头的 VLAN ID 以及 MAC 地址方式查找记录了 VLAN ID 的表项,VLAN unaware 方式利用内层以太网头的源 MAC 地址和目的 MAC 地址查找未记录 VLAN ID 的表项。

[0095] 通过 VLAN unaware 表项,连接至同一个接入设备的不同 VLAN 的 VM 的虚拟 MAC 地址被聚合成一个表项,同一个数据中心的不同 VLAN 的 VM 的虚拟 MAC 地址被聚合成一个表项,显著减少了接入层设备和核心层设备的二层转发表项的数目。

[0096] (4) 接入层路由桥的二层转发表由网络管理装置配置,核心层路由桥的二层转发表包括网络管理装置配置的表项以及根据已有 MAC 地址学习机制动态记录生成的表项。譬如,核心设备对外的接口使能 MAC 地址学习。其中,核心层设备学习到的 MAC 地址在二层转发表中同样设置 48 位的主机掩码。核心层设备和接入层设备的 MAC 地址学习可根据具体组网按全局和按端口灵活使能和去使能。

[0097] (5) VM 在发送多目的 MAC 地址的报文(如广播报文、已知组播报文、未知组播报文或未知单播报文)时,报文中的目的 MAC 地址不进行聚合处理。

[0098] 为了更清楚的说明本发明实施例的报文转发机制,下面以图 1 所示的组网架构和前述虚拟 MAC 地址编码规则为例,对几种典型场景下的报文转发流程进行描述。

[0099] 参见图 3A,为本发明实施例提供的一个数据中心内的相同 VLAN 内的报文转发示意图。其中,源 VM 位于 leaf1 的 port1 连接的物理服务器上,目的 VM 位于 leaf3 的 port1 连接的物理服务器上;源 VM 和目的 VM 的 IP 地址分别是 1.1.1.1 和 1.1.1.8。

[0100] 图 3A 中,leaf1 的二层转发表中至少包括表 3.1 所示的表项:

[0101] 表 3.1

[0102]

VLAN ID	Initial MAC address	Mask	Mapped MAC address	Egress Port
1	00-11-11-11-11-11	ff-ff-ff-ff-ff-ff	ED-01-00-01-00-01	Port1
1	ED-01-00-01-00-01	ff-ff-ff-ff-ff-ff	00-11-11-11-11-11	Port1
VLAN unaware	ED-01-00-03-00-00	ff-ff-ff-ff-00-00		DC1_leaf3
1	00-E0-FC-11-11-11	ff-ff-ff-ff-ff-ff		DC1_core1
2	00-E0-FC-22-22-22	ff-ff-ff-ff-ff-ff		DC1_core1
3	00-E0-FC-33-33-33	ff-ff-ff-ff-ff-ff		DC1_core2

[0103] leaf3 的二层转发表中至少包括表 3.2 所示的表项:

[0104] 表 3.2

[0105]

VLAN ID	Initial MAC address	Mask	Mapped MAC address	Egress Port
1	00-1D-A1-75-28-70	ff-ff-ff-ff-ff-ff	ED-01-00-03-00-02	Port1
1	ED-01-00-03-00-02	ff-ff-ff-ff-ff-ff	00-1D-A1-75-28-70	Port1
VLAN unaware	ED-01-00-01-00-00	ff-ff-ff-ff-00-00		DC1_leaf1
1	00-E0-FC-11-11-11	ff-ff-ff-ff-ff-ff		DC1_core1
2	00-E0-FC-22-22-22	ff-ff-ff-ff-ff-ff		DC1_core1
3	00-E0-FC-33-33-33	ff-ff-ff-ff-ff-ff		DC1_core2

[0106]

[0107] core1 的二层转发表中至少包括表 3.3 所示的表项：

[0108] 表 3.3

[0109]

VLAN ID	Initial MAC address	Mask	Mapped MAC address	Egress Port
VLAN unaware	00-5F-AA-95-82-07	ff-ff-ff-ff-ff-ff		DC1_leaf2
VLAN unaware	ED-01-00-02-00-00	ff-ff-ff-ff-00-00		DC1_leaf2
VLAN unaware	ED-01-00-03-00-00	ff-ff-ff-ff-00-00		DC1_leaf3
VLAN unaware	ED-01-00-04-00-00	ff-ff-ff-ff-00-00		DC1_leaf4
1	00-E0-FC-11-11-11	ff-ff-ff-ff-ff-ff		DC1_core1
2	00-E0-FC-22-22-22	ff-ff-ff-ff-ff-ff		DC1_core1

[0110] 源 VM 发送以太网报文给 leaf1，该以太网报文的源 MAC 地址为 00-11-11-11-11-11，目的 MAC 地址为 ED-01-00-03-00-02。

[0111] 源 VM 确定 IP 报文的目的 IP 地址 1.1.1.8 与自身的 IP 地址 1.1.1.1 处于同一 IP 网段，源 VM 根据目的 IP 地址 1.1.1.8 查找 ARP 表，查找到 ARP 表项记录的 MAC 地址是虚拟 MAC 地址 ED-01-00-03-00-02。源 VM 将 IP 报文封装为以太网报文，其中，源 MAC 地址 =00-11-11-11-11-11，目的 MAC 地址 =ED-01-00-03-00-02。

[0112] leaf1 将收到的以太网报文的源 MAC 地址 00-11-11-11-11-11 与二层转发表项（表 3.1 的第二行所示的表项）的 48 位掩码 ff-ff-ff-ff-ff-ff 进行“逻辑与”运算，运算结果 00-11-11-11-11-11 与该二层转发表项的初始 MAC 地址 00-11-11-11-11-11 一致，确定查找

到表项, leaf1 将找到的表项的映射 MAC 地址(Mapped MAC address)替换以太网报文的源 MAC 地址。

[0113] leaf1 将目的 MAC 地址 ED-01-00-03-00-02 与一个表项(表 3.1 的第三行所示的表项)的 32 位接入设备掩码 ff-ff-ff-ff-00-00 进行“逻辑与”运算, leaf1 确定运算结果 ED-01-00-03-00-00 与该表项的初始 MAC 地址 ED-01-00-03-00-00 一致。leaf1 根据找到的表项的 Egress Port 转发以太网报文到 leaf3。leaf1 根据 DC1_leaf3 为收到的以太网头封装 Trill 头;其中, Ingress nickname = leaf1 的 nickname, Egress Nickname = DC1_leaf3(即 leaf3 的 nickname)。然后, leaf1 为以太网报文封装下一跳头;其中, 源 MAC 地址 = leaf1 的 MAC 地址;目的 MAC 地址 = Core1 的 MAC 地址;VLAN ID = 指定 VLAN(Designated VLAN)ID。该下一跳头是一个逐跳头。本实施例中, RBridge 之间是以太网链路, 因此 leaf1 在 Trill 头外封装的下一跳头视为外层以太网头;Leaf1 可以根据 Trill 路由表确定到达 leaf3 的下一跳是 core1, 然后根据 Trill 邻接表查找到 core1 的 MAC 地址。下一跳头的源 MAC 地址以及目的 MAC 地址专用于识别发送 RBridge (transmitting RBridge) 以及下一跳 RBridge (Next hop RBridge)。leaf1 转发 Trill 封装的报文到下一跳 core1。

[0114] core1 收到 Trill 封装的报文, 解封装外层以太网头, 根据 Trill 头的 Egress Nickname 重新封装下一跳头, 其中, 源 MAC 地址 = core1 的 MAC 地址, 目的 MAC 地址 = leaf3 的 MAC 地址, VLAN ID = 指定 VLAN ID。core1 转发重新封装的 Trill 报文到下一跳 leaf3。

[0115] leaf3 收到 Trill 报文后, 发现下一跳头的目的 MAC 地址为本设备 MAC 地址则移除下一跳头, leaf3 确定 Trill 头的 Egress nickname 是本设备的 DC1_leaf3 则移除 Trill 头, leaf3 获得内层以太网报文。

[0116] leaf3 根据内层以太网报文的源 MAC 地址 ED-01-00-01-00-01 在表 3.2 所示二层转发表中查找到表项(表 3.2 第三行所示的表项), 该表项未包含映射 MAC 地址, 则不替换以太网报文的源 MAC 地址;leaf3 根据以太网报文的目的 MAC 地址 ED-01-00-03-00-02 查找到包含映射 MAC 地址的表项(表 3.2 第三行所示的表项)。leaf3 将以太网报文中的目的 MAC 替换为 00-1D-A1-75-28-70, 将该以太网报文通过表项的 port1, 转发以太网报文。

[0117] leaf3 将源 MAC 地址 ED-01-00-01-00-01 与表项 32 位掩码 ff-ff-ff-ff-00-00 进行“逻辑与”运算, 运算结果 ED-01-00-01-00-00 与该初始 MAC 地址 ED-01-00-01-00-00 一致, 则确定查找到表项。leaf3 将目的 MAC 地址 ED-01-00-03-00-02 与表项的 48 位的掩码 ff-ff-ff-ff-ff-ff 进行“逻辑与”运算, 运算结果表项中初始 MAC 地址 ED-01-00-03-00-02 一致, 则确定查找到表项。

[0118] 图 3A 中, 若源 VM 在 ARP 表中未查找到目的 IP 地址 1.1.1.8 对应的 ARP 表项, 则发送 ARP 请求报文, 以请求该目的 IP 地址 1.1.1.8 对应的 MAC 地址。其中, ARP 请求报文的 Sender IP 地址是 1.1.1.1, Sender MAC 地址是 00-11-11-11-11-11, Target IP 地址是 1.1.1.8, Target MAC 地址是全 0 的 MAC 地址。ARP 请求报文的以太网头的源 MAC 地址和目的 MAC 地址分别是 00-11-11-11-11-11 和全 F 的广播地址。

[0119] leaf1 收到 ARP 请求报文后, 不在 Trill 网络广播 ARP 请求报文, 而是将收到的 ARP 请求报文转化为私有 ARP 请求报文(如图 3B 所示), 将其单播发送给网络管理装置。

[0120] leaf1 移除收到的 ARP 请求报文的以太网头, 封装 IP 头, 其中, 源 IP 地址为 leaf1 的 IP 地址 1.1.1.30, 目的 IP 地址是网络管理装置的 IP 地址 122.1.2.1。然后, leaf1 在

IP 报文头外封装一个逐跳头。本实施例中,该逐跳头是以太网头,其中,源 MAC 地址是对应 leaf1 的 MAC 地址 00-5F-AA-95-82-07 (对应于 IP 头的源 IP 地址 1.1.1.30),目的 MAC 地址是 VLAN1 网关的 MAC 地址 00-E0-FC-11-11-11。

[0121] Leaf1 根据私有 ARP 请求报文的源 MAC 地址未查找到二层转发表项。leaf1 根据目的 MAC 地址 00-E0-FC-11-11-11 查找到未包含映射 MAC 地址的二层转发表项(表 3.1 第五行所示的表项),根据查找到的表项的出接口将私有 ARP 报文封装为 Trill 封装的私有 ARP 请求报文发往 core1。leaf1 在私有 ARP 请求报文外封装 Trill 头和以太网头(外层的以太网头),私有 ARP 请求报文的以太网头位于 Trill 头与 IP 头之间,仍可以视为内层的以太网头,Trill 头外的以太网头仍视为外层的以太网头。

[0122] core1 收到 Trill 封装的报文,移除 Trill 封装(外层的以太网头和 Trill 头),移除内层的以太网头,根据私有 ARP 请求报文的 IP 头的目的 IP 地址重新封装以太网头(逐跳头),其中,目的 MAC 地址是到达目的 IP 地址的下一跳的 MAC 地址,VLAN ID 是下一跳设备所在的 VLAN 的标识;源 MAC 地址是 core1 的一个三层接口 MAC 地址,该三层接口所在的 VLAN 与下一跳设备位于相同的 VLAN。下一跳设备收到 core1 重新封装的私有 ARP 请求报文后,根据私有 ARP 请求报文的 IP 头的目的 IP 地址执行三层转发,方式与 core1 的转发过程类似。

[0123] 网络管理装置收到该私有 ARP 请求报文后,在表 1 所示的全网设备以及 VM 信息表查找 IP 地址 1.1.1.8 对应的虚拟 MAC 地址 ED-01-00-03-00-02,将 IP 地址 1.1.1.1 和 MAC 地址 00-11-11-11-11-11 设置为 ARP 响应报文 Target IP 地址和 Target MAC 地址(IP 地址 1.1.1.1 和 MAC 地址 00-11-11-11-11-11 分别是网络管理装置收到的 ARP 请求报文中的 Sender IP 地址和 Sender MAC 地址),将目的 VM 的 IP 地址 1.1.1.8 和虚拟 MAC 地址 ED-01-00-03-00-02 设置为 ARP 响应报文的 Sender IP 地址和 Sender Target MAC 地址,将 ARP 响应报文封装单播的私有 ARP 响应报文(如图 3B 所示)。即,网络管理装置在 ARP 响应报文封装 IP 头和以太网头(逐跳头)。私有 ARP 响应报文的 IP 头的源 IP 地址是其自身的 IP 地址 122.1.2.1,目的 IP 地址是 Leaf1 的 IP 地址 1.1.1.30。私有 ARP 响应报文的以太网头的源 MAC 地址是网络管理装置的 MAC 地址,目的 MAC 地址是到达目的 IP 地址的下一跳设备的 MAC 地址。这样,私有 ARP 响应报文的以太网头的源 MAC 地址和目的 MAC 地址会逐跳改变,但是私有 ARP 响应报文的目的 IP 地址不变,这样私有 ARP 报文被逐跳发送到作为 VLAN1 的网关 core1。

[0124] core1 收到私有 ARP 响应报文,根据私有 ARP 响应报文的 IP 头的目的 IP 地址执行三层转发,将私有 ARP 响应报文的以太网头的源 MAC 地址和目的 MAC 地址分别修改为 VLAN1 网关的 MAC 地址 00-E0-FC-11-11-11 和 leaf1 的 MAC 地址 00-5F-AA-95-82-07。core1 根据私有 ARP 响应报文的以太网头的目的 MAC 地址查找到表项(表 3.3 第二行所示的表项),对私有 ARP 响应报文进行 Trill 封装,在 Trill 域内将 Trill 封装的私有 ARP 响应报文发送到 leaf1。

[0125] leaf1 收到 Trill 封装的私有 ARP 响应报文,移除外层以太网头和 Trill 头,将私有 ARP 响应报文的以太网头和 IP 头移除,为 ARP 响应报文设置以太网头,其中,源 MAC 地址 = ED-01-00-03-00-02;目的 MAC 地址 = ED-01-00-01-00-01。leaf1 根据 Sender MAC 地址 ED-01-00-03-00-02 地址查找到的表项不包含映射的虚拟 MAC 地址,则将 Sender MAC 地

址设置为 ARP 响应报文的源 MAC 地址。leaf1 根据 Target MAC 地址 00-11-11-11-11-11 在二层转发表中查找到映射的虚拟 MAC 地址 ED-01-00-01-00-01，将其作为 ARP 响应报文的目的 MAC 地址。

[0126] leaf1 根据转发 ARP 响应报文到源 VM。leaf1 根据源 MAC 地址查找到未包含映射 MAC 地址的表项，根据目的 MAC 地址 ED-01-00-01-00-01 查找到包含映射 MAC 地址的表项，将 ARP 响应报文的以太网头的目的 MAC 地址 ED-01-00-01-00-01 替换为 00-11-11-11-11-11，通过端口 Port1 发送 ARP 响应报文至源 VM。源 VM 根据收到的 ARP 响应报文学习 ARP 表项，该 ARP 表项记录了 IP 地址 1.1.1.8 与虚拟 MAC 地址 ED-01-00-03-00-02 的映射关系。

[0127] 本发明实施例中，网络管理装置可采用其他方式设置私有 ARP 响应报文的一对发送端地址和一对目标端地址。网络管理装置将 IP 地址 1.1.1.1 和 MAC 地址 ED-01-00-01-00-01 设置为 ARP 响应报文 Target IP 地址和 Target MAC 地址 (Target IP 地址 1.1.1.1 是收到的 ARP 请求报文中的 Sender IP 地址，Target MAC 地址 ED-01-00-01-00-01 映射于收到 ARP 请求报文的 Sender MAC 地址的虚拟 MAC 地址)，将目的 VM 的 IP 地址 1.1.1.8 和虚拟 MAC 地址 ED-01-00-03-00-02 设置为 ARP 响应报文的 Sender IP 地址和 Sender MAC 地址。

[0128] 网络管理装置将 ARP 响应报文封装为单播的私有 ARP 响应报文。私有 ARP 报文逐跳地发送 core1。core1 收到私有 ARP 响应报文，根据私有 ARP 响应报文的 IP 头的目的 IP 地址执行路由转发，修改私有 ARP 响应报文的以太网头的源 MAC 地址和目的 MAC 地址，其中，源 MAC 地址和目的 MAC 地址分别是 VLAN1 网关的 MAC 地址 00-E0-FC-11-11-11 和 leaf1 的 MAC 地址 00-5F-AA-95-82-07。core1 根据私有 ARP 响应报文的以太网头的目的 MAC 地址查找到表项(表 3.3 第二行所示的表项)，对私有 ARP 响应报文进行 Trill 封装，在 Trill 域内将 Trill 封装的私有 ARP 响应报文发送到 leaf1。

[0129] leaf1 收到 Trill 封装的私有 ARP 响应报文，移除外层以太网头和 trill 头，将私有 ARP 响应报文的以太网头和 IP 头移除，将 ARP 响应报文中的 Sender MAC 地址 ED-01-00-03-00-02 和 Target MAC 地址 ED-01-00-01-00-01 分别设置为 ARP 响应报文的源 MAC 地址和目的 MAC 地址。

[0130] leaf1 转发 ARP 响应报文到源 VM。leaf1 根据源 MAC 地址查找到未包含映射 MAC 地址的表项。leaf1 根据配置的二层转发表项将 ARP 响应报文的以太网头的目的 MAC 地址 ED-01-00-01-00-01 替换为 00-11-11-11-11-11，通过端口 Port1 发送 ARP 响应报文至源 VM。源 VM 根据收到的 ARP 响应报文学习 ARP 表项，该 ARP 表项记录了 IP 地址 1.1.1.8 与虚拟 MAC 地址 ED-01-00-03-00-02 的映射关系。

[0131] 需要说明的是，接入层路由桥对于从普通接口收到的 ARP 请求报文进行截获，而对使能 Trill 协议的接口收到的 ARP 请求报文的不进行截获。大二层网络中，如果核心层路由桥的三层 L3 接口以广播方式发送 ARP 请求报文以学习 VM 的 ARP 表项，为了控制 ARP 请求报文的洪泛(flooding)，可同样使用上述的 ARP 截获机制。

[0132] 例如，图 1 中的三层设备 core1 发送单播的私有 ARP 请求报文至网络管理装置，请求数数据中心内 VLAN1 内所有 VM 的 ARP 信息。

[0133] 或者，core1 发送单播的私有 ARP 请求报文至网络管理装置，请求某一个 VM 的 ARP

信息。仍以目的 VM 为例, core1 发送私有 ARP 请求报文。私有 ARP 请求报文的 Sender IP 地址是 VLAN1 网关的 IP 地址, Sender MAC 地址 VLAN1 网关的 MAC 地址 00-E0-FC-11-11-11, Target IP 地址是 1.1.1.8, Target MAC 地址是全 0 的 MAC 地址。私有 ARP 请求报文的源 IP 地址 VLAN1 网关的 IP 地址 1.1.1.30, 目的 IP 地址是网络管理装置的 IP 地址 122.1.2.1。然后, core1 在 IP 报文头外封装一个逐跳的以太网头。最终, 私有 ARP 请求报文被逐跳的发送到网络管理装置。

[0134] 网络管理装置根据私有 ARP 请求报文的 Target IP 地址 1.1.1.8 查找到对应的虚拟 MAC 地址 ED-01-00-03-00-02, 将 IP 地址 1.1.1.8 和虚拟 MAC 地址 D-01-00-03-00-02 设置为私有 ARP 响应报文 Sender IP 地址和 Sender MAC 地址, 将收到的私有 ARP 请求报文中的 Sender IP 地址和 Sender MAC 地址设置为私有 ARP 报文的 Target IP 地址和 Target MAC 地址。网络管理装置将私有 ARP 响应报文的源 IP 地址设置自身的 IP 地址 122.1.2.1, 将私有 ARP 响应报文的目的 IP 地址设置为 VLAN1 网关的 IP 地址 1.1.1.30。私有 ARP 响应报文的以太网头的源 MAC 地址是网络管理装置的 MAC 地址, 目的 MAC 地址是到达 core1 的下一跳设备的 MAC 地址。这样, 私有 ARP 响应报文的被逐跳发送到作为 VLAN1 的网关 core1。

[0135] core1 收到私有 ARP 响应报文, 根据 Sender IP 地址和 Sender MAC 地址学习 ARP 表项。

[0136] 而外网 VLAN (如 VLAN m) 与数据中心 Trill 网络无关, 因此 core1 仍可按照 ARP 协议机制学习外网 VLAN m 三层接口的 ARP 表项。

[0137] 因此, ARP 请求报文是否采用截获处理方式可以在设备的 VLAN 和 port 模式下进行配置以示区分。

[0138] 参见图 4, 为本发明实施例提供的组播报文转发的示意图。其中, 源 VM 为 leaf1 的 port1 所连接的 MAC 地址为 00-11-11-11-11-11 的 VM。

[0139] leaf1 的二层转发表中至少包括表 4.1 所示的表项:

[0140] 表 4.1

[0141]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
1	00-11-11-11-11-11	ff-ff-ff-ff-ff-ff	ED-01-00-01-00-01	Port1
1	ED-01-00-01-00-01	ff-ff-ff-ff-ff-ff	00-11-11-11-11-11	Port1
1	01-00-5E-XX-XX-XX	ff-ff-ff-ff-ff-ff		DC1_core1

[0142] leaf3 的二层转发表中至少包括表 4.2 所示的表项:

[0143] 表 4.2

[0144]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
VLAN unaware	ED-01-00-01-00-00	ff-ff-ff-ff-00-00		DC1_leaf1
1	01-00-5E-XX-XX-XX	ff-ff-ff-ff-ff-ff		Port1

[0145] leaf4 的二层转发表中至少包括表 4.2 所示的以下表项：

[0146] 表 4.3

[0147]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
1	01-00-5E-XX-XX-XX	ff-ff-ff-ff-ff-ff		Port1
VLAN unaware	ED-01-00-01-00-00	ff-ff-ff-ff-00-00		DC1_leaf1

[0148] 如图 4 所示,作为组播源的 VM 发送组播报文给 leaf1,该报文的源 MAC 地址为源 VM 的真实 MAC 地址 00-11-11-11-11-11,目的 MAC 地址为组播组 1 的 MAC 地址 01-00-5E-XX-XX-XX。

[0149] leaf1 接收到组播报文后,根据该报文的源 MAC 地址 00-11-11-11-11-11 查找到表项(表 4.1 第二行所示的表项);根据该表项的映射 MAC 地址 ED-01-00-01-00-01 替换源 MAC 地址。

[0150] leaf1 根据该以太网报文的目的 MAC 地址 01-00-5E-XX-XX-XX,查找到表项(表 4.1 第四行所示的表项),根据查找到表项的出接口 DC1_core1 将收到的组播报文封装为 Trill 报文并将 Trill 封装的报文分发到 Trill 网络。即,core1 的 nickname 是目的组播组所在组播树的树根的 nickname),

[0151] 本步骤中,leaf1 将该出接口的 Nickname 作为 Egress nickname,将自身的 DC1_leaf1 作为 Ingress Nickname,为组播该报文封装 Trill 报文头。leaf1 将本设备的 MAC 地址和 Trill 定义的特定组播 MAC 地址作为下一跳头的源 MAC 地址和目的 MAC 地址。leaf1 通过 Trill 封装的报文至该组播组的组播树的树根 core1。

[0152] core1 接收到 Trill 封装的组播报文后,移除外层的以太网头和 Trill 头,根据 Trill 头的 Egress Nickname 在 Trill 组播表确定该组播组在 VLAN1 的组播转发树有两个下游节点 leaf3 和 leaf4,因此复制两份报文并分别封装为 Trill 封装的组播报文,然后发送给两个下游路由器 leaf3 和 leaf4。

[0153] leaf3 和 leaf4 各自收到 Trill 封装的组播报文,解封装获得内层组播报文,然后根据该内层以太网头的源 MAC 地址分别在表 4.2、表 4.3 所示的二层转发表进行查找。leaf3 和 leaf4 分别查找到表项(表 4.2 第三行所示的表项以及表 4.3 所示的第三行所示的表项),这些表项未包含映射 MAC 地址,leaf3 和 leaf4 不替换组播报文的源 MAC 地址,leaf3 和 leaf4 各自根据组播 MAC 地址查到对应的表项(表 4.2 第二行所示的表项以及表 4.3 所示的第二行所示的表项)二层转发表,然后根据表项中的出接口 port1 发送组播报文。

[0154] 参见图 5,为本发明实施例中数据中心内的报文转发至外网的示意图。其中,源 VM 为数据中心 1 中的 leaf1 的 port1 所接入的 VM,源 VM 的 IP 地址为 1.1.1.1。目的端的 IP 地址为 172.1.1.1,该 IP 地址是用户的业务 IP 地址而非跨数据中心二层互联用的隧道 IP 地址,VLAN 为 VLAN m。

[0155] 数据中心 1 核心层的 core1 和 core2 采用堆叠协议构成的一个虚拟设备作为网关,以实现负载均衡与备份。该虚拟设备是一个虚拟的核心层设备。leaf1 将自身连接 core1 和 core2 的链路绑定为一个链路聚合组。leaf2—leaf4 各自将自身连接 core1 和

core2 的链路绑定为一个链路聚合组。

[0156] 表 1 中数据中心 1 的 core1 和 core2 的配置信息按表 5.1 做如下修改：

[0157] 表 5.1

[0158]

Device	Nickname	IP Address (MAC over IP)	Port	VLA N	IP address of VM	MAC address of VM	Virtual MAC address of VM	Access Device Mask Based Virtual MAC Address	Data Centre Mask Based virtual MAC Address
core1	DC1_core		L3 interface	1	1.1.1.100	00-E0-FC-11-11-11		ED-01-00-00-00-00	
			L3 interface	2	2.2.2.100	00-E0-FC-22-22-22			
			L3 interface	3	3.3.3.100	00-E0-FC-33-33-33			
core2	DC1_core		L3 interface	1	1.1.1.100	00-E0-FC-11-11-11		ED-01-00-00-00-00	
			L3 interface	2	2.2.2.100	00-E0-FC-22-22-22			
			L3 interface	3	3.3.3.100	00-E0-FC-33-33-33			

[0159] 本实施例中，DC1_core 是虚拟设备的 Nickname。core2 是虚拟设备的 master 设备。

[0160] 图 5 中 leaf1 的二层转发表中至少包括表 5.2 所示的以下表项：

[0161] 表 5.2

[0162]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
1	00-11-11-11-11-11	f-ff-ff-ff-ff-ff	ED-01-00-01-00-01	Port1
1	ED-01-00-01-00-01	ff-ff-ff-ff-ff-ff	00-11-11-11-11-11	Port1
1	00-E0-FC-11-11-11	ff-ff-ff-ff-ff-ff		DC1_core
2	00-E0-FC-22-22-22	ff-ff-ff-ff-ff-ff		DC1_core
3	00-E0-FC-33-33-33	ff-ff-ff-ff-ff-ff		DC1_core

[0163] core1 和 core2 的二层转发表中至少包括表 5.3 所示的以下表项：

[0164] 表 5.3

[0165]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
1	00-E0-FC-11-11-11	ff-ff-ff-ff-ff-ff		L3
2	00-E0-FC-22-22-22	ff-ff-ff-ff-ff-ff		L3
3	00-E0-FC-33-33-33	ff-ff-ff-ff-ff-ff		L3
VLAN unaware	ED-01-00-01-00-00	ff-ff-ff-ff-00-00		DC1_leaf1
m	Next-hop MAC	ff-ff-ff-ff-ff-ff		Port m

[0166] 如图 5 所示,源 VM 发送以太网报文给 leaf1,该报文的源 MAC 地址为该 VM 的真实 MAC 地址 00-11-11-11-11-11,目的 MAC 地址为 VLAN1 网关的 MAC 地址 00-E0-FC-11-11-11。

[0167] 此步骤中,源 VM 确定 IP 地址 1.1.1.1 与目的 IP 地址 172.1.1.1 不在同一网段,因此通过查询 VM 本地路由表获得 VLAN1 的网关的 IP 地址(网关 IP 地址可以是通过静态配置方式或动态主机设置协议方式配置)。

[0168] 若源 VM 未查找到 VLAN1 网关 IP 地址匹配的 ARP 表项,则广播 ARP 请求报文,以请求 VLAN1 网关的 IP 地址 1.1.1.100 对应的 MAC 地址。具体过程本实施例不再赘述。

[0169] 源 VM 根据网关的 IP 地址为 1.1.1.100 在 ARP 表项中查找到对应的 MAC 地址为 00-E0-FC-11-11-11,将该 MAC 地址作为以太网头的目的 MAC 地址。

[0170] leaf1 收到该报文后,根据该报文的源 MAC 地址 00-11-11-11-11-11 在二层转发表查找到表项(表 5.2 第二行所示的表项),将收到的以太网报文的源 MAC 地址 00-11-11-11-11-11 替换为该表项的映射 MAC 地址 ED-01-00-01-00-01。

[0171] leaf1 根据该报文的目的 MAC 地址 00-E0-FC-11-11-11 在二层转发表中查找到表项(表 5.2 第四行所示的表项),且该表项未包含对应的 mapped MAC 地址,leaf1 根据表项的出接口 Egress nickname(即 core1 和 core2 堆叠构成的逻辑节点的 nickname :DC1_core),将收到的以太网报文外封装为 Trill 封装的报文。

[0172] leaf1 将 DC1_core(core1 和 core2 堆叠构成的逻辑节点的 nickname)作为 Egress nickname,将自身的 DC1_leaf1 作为 Ingress nickname,封装 Trill 头。leaf1 根据 Trill 路由表,确定到达 Egress Nickname 的下一跳是 DC1_core,在 Trill 邻接表中查找到 DC1_core 的 MAC 地址,在 Trill 报文头外封装下一跳头(Next-Hop header);其中,源 MAC 地址为 leaf1 的 MAC 地址,目的 MAC 地址为 DC1_core1 的 MAC 地址,VLAN ID 是指定 VLAN(Designated VLAN) 标识。leaf1 转发 Trill 封装的报文至 DC1_core。

[0173] DC1_core 的主设备 core2 收到 Trill 封装的以太网报文,移除下一跳头和 Trill 头,根据内层以太网报文的源 MAC 地址 ED-01-00-01-00-01 在二层转发表中查找到表项(表 5.3 第五行所示的表项),该表项不包括 mapped MAC 地址,core1 不替换源 MAC 地址。然后 core1 根据内层以太网头的目的 MAC 地址 00-E0-FC-11-11-11 在二层转发表中查找到表项(表 5.3 第二行所示的表项)且该表项的出端口信息为 L3 标记(表示使能 L3 转发处理),因

此 core2 转入三层转发处理流程：在路由表中查目的 IP 地址 172.1.1.1 的路由表项，确定到达目的 IP 地址的下一跳，查询下一跳的 MAC 地址，将解 Trill 封装后的以太网报文的源 MAC 地址设置虚拟设备 DC1_core 的 VLAN m 的接口的 MAC 地址，将以太网报文的目的 MAC 地址设置为下一跳的 MAC 地址（next-hop MAC address），报文在 IP 网络内基于路由被逐跳转发到至 IP 地址为 172.1.1.1 的目的端设备。

[0174] IP 地址为 172.1.1.1 的设备发给源 VM 的 IP 报文在 IP 网络内被逐跳地发到虚拟设备。

[0175] DC1_core 的成员设备 core1 收到来自数据中心外的以太网报文时，其中，源 MAC 是 DC1_core 学习到的下一跳的 MAC 地址；目的 MAC 地址是 DC1_core 的 VLAN m 的三层接口 MAC 地址。core1 根据目的 MAC 地址在二层转发表进行查找，确定执行三层转发。core1 根据目的 IP 地址 1.1.1.1 查询 ARP 表，确定对应的 MAC 地址是 ED-01-00-01-00-01，将 1.1.1.1 所在 VLAN 的标识 VLAN1 设置为收到的以太网报文的 VLAN ID，将根据 ARP 表查询到的 MAC 地址设置为收到以太网报文的目的 MAC 地址，将 VLAN1 网关的 MAC 地址 00-E0-FC-11-11-11 设置为收到的以太网报文的源 MAC 地址。

[0176] core1 根据源 MAC 地址 00-E0-FC-11-11-11 和目的 MAC 地址 ED-01-00-01-00-01 分别在二层转发表执行查找，core1 查找到的表项（表 5.3 第二行以及第五行所示的表项）不包含映射的虚拟 MAC 地址。core1 根据目的 MAC 地址匹配的二层转发表项中的出接口，封装收到的以太网报文封装为 Trill 报文，其中，Egress nickname 是 DC1_leaf1，Ingress nickname 是 DC1_core。core1 根据 Trill 路由表，确定到达 Egress Nickname 的下一跳是 Leaf1，在 Trill 邻接表查找到 leaf1 的 MAC 地址，在 Trill 报文头外封装下一跳头（Next-Hop header）；其中，源 MAC 地址为 DC1_core1 的 MAC 地址，目的 MAC 地址为 leaf1 的 MAC 地址，VLAN ID 是指定 VLAN（Designated VLAN）标识。core1 转发 Trill 封装的报文至 leaf1。

[0177] 参见图 6，为本发明实施例中数据中心内不同 VLAN 报文转发的示意图。其中，源 VM 所在的物理服务器连接到 leaf1 的 port1，源 VM 的 IP 地址为 1.1.1.1，所属 VLAN 为 VLAN1。目的 VM 所在的物理服务器连接到 leaf3 的 port2，目的 VM 的 IP 地址为 2.2.2.7，所属 VLAN 为 VLAN2。

[0178] 图 6 中，数据中心 1 核心层的 core1 和 core2 堆叠构成的一个虚拟设备，以实现负载均衡与备份。虚拟设备的 Nickname 为 DC1_core。core2 是虚拟设备的 master 设备。leaf1 – leaf4 各自将自身连接 core1 和 core2 的链路绑定为一个链路聚合组。

[0179] leaf1 的二层转发表中至少包括表 6.1 所示的以下表项：

[0180] 表 6.1

[0181]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
1	00-11-11-11-11-11	ff-ff-ff-ff-ff-ff	ED-01-00-01-00-01	Port1

[0182]

1	ED-01-00-01-00-01	ff-ff-ff-ff-ff-ff	00-11-11-11-11-11	Port1
1	00-E0-FC-11-11-11	ff-ff-ff-ff-ff-ff		DC1_core
2	00-E0-FC-22-22-22	ff-ff-ff-ff-ff-ff		DC1_core

[0183] core1 和 core2 的二层转发表中至少包括表 6.2 所示的以下表项：

[0184] 表 6.2

[0185]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
1	00-E0-FC-11-11-11	ff-ff-ff-ff-ff-ff		L3
2	00-E0-FC-22-22-22	ff-ff-ff-ff-ff-ff		L3
VLAN unaware	ED-01-00-01-00-00	ff-ff-ff-ff-00-00		DC1_leaf1
VLAN unaware	ED-01-00-02-00-00	ff-ff-ff-ff-00-00		DC1_leaf2
VLAN unaware	ED-01-00-03-00-00	ff-ff-ff-ff-00-00		DC1_leaf3
VLAN unaware	ED-01-00-04-00-00	ff-ff-ff-ff-00-00		DC1_leaf4

[0186] leaf3 的二层转发表中至少包括表 6.3 所示的以下表项：

[0187] 表 6.3

[0188]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
1	00-E0-FC-11-11-11	ff-ff-ff-ff-ff-ff		DC1_core
2	00-E0-FC-22-22-22	ff-ff-ff-ff-ff-ff		DC1_core
1	00-25-9C-2F-63-FE	ff-ff-ff-ff-ff-ff	ED-01-00-03-00-04	Port1
1	ED-01-00-03-00-04	ff-ff-ff-ff-ff-ff	00-25-9C-2F-63-FE	Port1

[0189] 如图 6 所示, 源 VM 发送以太网报文给 leaf1, 该报文的源 MAC 地址为该 VM 的真实 MAC 地址 00-11-11-11-11-11, 目的 MAC 地址为 VLAN1 网关的 MAC 地址 00-E0-FC-11-11-11。

[0190] leaf1 接收到以太网报文后, 根据源 MAC 地址 00-11-11-11-11-11 查二层转发表, 查找到包含映射 MAC 地址的表项(表 6.1 第二行所示表项), 将报文的源 MAC 地址 00-11-11-11-11-11 替换为映射 MAC 地址 ED-01-00-01-00-01。

[0191] leaf1 根据该报文的目的 MAC 地址 00-E0-FC-11-11-11 查二层转发表, leaf1 查找到未包含映射 MAC 地址的表项(表 6.1 第四行表项所示的表项), leaf1 根据查找到的表

项出接口 DC1_core 将收到的以太网报文封装为 Trill 报文,然后根据 Trill 路由表发送到 Trill 网络进行转发。

[0192] DC1_core 的主设备 core2 收到 Trill 封装的报文,移除下一跳头和 Trill 头,根据内层以太网头报文的源 MAC 地址 ED-01-00-01-00-01 查二层转发表,core2 查找到未包含映射 MAC 地址的表项(表 6.2 第四行所示的表项),core2 不替换内层以太网头的源 MAC 地址。

[0193] core2 根据内层以太网头的目的 MAC 地址 00-E0-FC-11-11-11 查二层转发表,core2 查找到未包含映射 MAC 地址且出接口信息为 L3 属性的表项(表 6.2 第二行所示的表项)。core2 路由表中查目的 IP 地址 2.2.2.7 的路由表项,然后根据命中的路由表项中的 IP 地址在 ARP 表查找到对应的虚拟 MAC 地址 ED-01-00-03-00-04,将内层以太网报文的 VLAN ID 由 VLAN1 更换为 VLAN2,将源 MAC 地址设置为 VLAN2 网关的三层接口 MAC 地址 00-E0-FC-22-22-22,将内层以太网头的目的 MAC 地址设置为 ED-01-00-03-00-04。

[0194] DC1_core 的主设备 core2 根据目的 MAC 地址 ED-01-00-03-00-04 查找到未包含映射 MAC 地址的表项(表 6.2 第六行所示的表项),然后根据查找到的表项的出接口将以太网报文封装为 Trill 封装的以太网报文,发送 Trill 封装的以太网到 leaf3。

[0195] leaf3 收到 Trill 封装的以太网报文后,移除 Trill 头和下一跳头。Leaf3 根据以太网报文的源 MAC 地址 00-E0-FC-22-22-22 查二层转发表,leaf3 查找到未包含映射 MAC 地址的表项(表 6.3 第三行所示表项),leaf3 不替换源 MAC 地址;leaf3 根据以太网报文的目的 MAC 地址 ED-01-00-03-00-04 查找到包含映射 MAC 地址的表项(表 6.3 最后一行所示的表项),leaf3 将以太网报文的目的 MAC 地址 ED-01-00-03-00-04 替换为表项中的映射 MAC 地址 00-25-9C-2F-63-FE,leaf3 根据查找到的表项的物理端口 port1 转发替换了目的 MAC 地址的以太网报文给目的 VM。

[0196] 图 7 为本发明实施例中的不同数据中心之间的二层转发的示意图。该组网中,数据中心互联网络采用 MAC over IP 技术,每个数据中心设备同时支持 Trill 和 MAC over IP,并支持 Trill 和 MAC over IP 的双向转换,即 Trill 终结后继续封装 MAC over IP 报文,MAC over IP 终结后继续封装 Trill 报文。

[0197] 在数据中心 1 中,核心层的 core1 和 core2 堆叠构成的一个虚拟设备,以实现负载均衡与备份。core1 和 core2 堆叠构成的虚拟设备的 Nickname 为 DC1_core,core2 是逻辑节点的 master 设备。

[0198] 在数据中心 2 中,核心层的 core1' 和 core2' 堆叠构成的一个虚拟设备,以实现负载均衡与备份。core1' 和 core2' 堆叠构成的虚拟设备的 Nickname 为 DC2_core,core1' 是逻辑节点的 master 设备。

[0199] 该流程中,源 VM 所在的物理服务器连接于数据中心 1 的 leaf1,源 VM 的 IP 地址为 1.1.1.1,所属 VLAN 为 VLAN1。目的 VM 所在的物理服务器连接于数据中心 2 的 leaf1',目的 VM 的 IP 地址为 1.1.1.100,所属 VLAN 为 VLAN1。

[0200] leaf1 的二层转发表中至少包括表 7.1 所示的以下表项:

[0201] 表 7.1

[0202]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
1	00-11-11-11-11-11	ff-ff-ff-ff-ff-ff	ED-01-00-01-00-01	Port1
1	ED-01-00-01-00-01	ff-ff-ff-ff-ff-ff	00-11-11-11-11-11	Port1
VLAN unaware	ED-02-00-00-00-00	ff-ff-00-00-00-00		DC1_core

[0203] core1 和 core2 各自的二层转发表至少包括表 7.2 所示的以下表项：

[0204] 表 7.2

[0205]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
VLAN unaware	ED-01-00-01-00-00	ff-ff-ff-ff-00-00		DC1_leaf1
VLAN unaware	ED-02-00-00-00-00	ff-ff-00-00-00-00		IP2

[0206] core1' 和 core2' 各自转发二层转发表中至少包括表 7.3 所示的以下表项：

[0207] 表 7.3

[0208]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
VLAN unaware	ED-01-00-01-00-00	ff-ff-00-00-00-00		IP1
VLAN unaware	ED-02-00-01-00-00	ff-ff-ff-ff-00-00		DC2_leaf1

[0209] leaf1' 的二层转发表中至少包括表 7.4 所示的以下表项：

[0210] 表 7.4

[0211]

VLAN	Initial MAC address	Mask	Mapped MAC address	Egress Port
1	ED-02-00-01-00-01	ff-ff-ff-ff-ff-ff	00-20-00-20-20-20	Port1
1	00-20-00-20-20-20	ff-ff-ff-ff-ff-ff	ED-02-00-01-00-01	Port1
VLAN unaware	ED-01-00-00-00-00	ff-ff-00-00-00-00		DC2_core

[0212] 如图 7 所示, 源 VM 发送以太网报文到 leaf1, 该以太网报文的源 MAC 地址为真实 MAC 地址 00-11-11-11-11-11, 目的 MAC 地址为目的 IP 地址对应的虚拟 MAC 地址 ED-02-00-01-00-01。

[0213] leaf1 收到该以太网报文, 根据源 MAC 地址 00-11-11-11-11-11 在二层转发表中查找到具有映射 MAC 地址的表项(表 7.1 第二行所示的表项), 将报文的源 MAC 地址替换为

映射 MAC 地址 ED-01-00-01-00-01 ;leaf1 根据目的 MAC 地址 ED-02-00-01-00-01 在二层转发表查找到未包含映射 MAC 地址表项(表 7.1 第四行所示的表项),根据表项的出接口 DC1_core 将收到的以太网报文封装为 Trill 报文,在 Trill 网络内发送到出口设备 DC1_core。

[0214] DC1_core 的主设备 core2 接收到 Trill 报文,解封装得到以太网报文,根据解封装后的以太网报文的源 MAC 地址 ED-01-00-01-00-01 查找到未包含映射 MAC 地址的表项(表 7.2 第二行所示的表项),core1 不替换源 MAC 地址。core2 根据解封装后的以太网报文的目的 MAC 地址 ED-02-00-01-00-01 查找到未包含映射 MAC 地址的表项(表 7.2 第三行所示的表项),core2 根据查找到表项中的出接口 IP2 将解封装后的以太网报文进行 MAC over IP 封装,core2 为报文封装 MAC over IP 报文头,其中,外层的源 IP 地址 = IP1、外层的目的 IP 地址 = IP2,core2 根据 IP2 进行路由转发,在外层 IP 头外封装逐跳的以太网头,从而将 MAC over IP 封装后的报文通过数据中心互联(DCI, data centre interconnecting) 网络被逐跳的发送到数据中心 2。

[0215] DC2_core 的主设备 core1' 收到 MAC over IP 封装的报文,解封装外层的以太网头和 IP 头,得到内层的以太网报文,根据内层以太网头的源 MAC 地址 ED-01-00-01-00-01 在二层转发表查找到未包含映射 MAC 地址的表项(表 7.3 第二行所示的表项),core1' 不替换源 MAC 地址。core1' 根据解封装后的以太网报文目的 MAC 地址 ED-02-00-01-00-01 查找到未包含映射 MAC 地址的表项(表 7.3 第三行所示的表项),core1' 根据该表项中的出端口 DC2_leaf1 将内层的以太网报文封装为 Trill 封装的以太网报文,在数据中心 2 的 Trill 网络内将 Trill 封装的以太网报文发到 Leaf1'。

[0216] leaf1' 收到 Trill 封装的报文,移除下一跳头和 Trill 头,根据源 MAC 地址 ED-01-00-01-00-01 在二层转发表查找到未包含映射 MAC 地址的表项,(表 7.4 最后一行所示的表项),leaf1' 不替换源 MAC 地址学习和替换;leaf1' 根据目的 MAC 地址 ED-02-00-01-00-01 在二层转发表查找到包含映射 MAC 地址的表项(表 7.4 第二行所示的表项),则将目的 MAC 地址替换为映射 MAC 地址 00-20-00-20-20-20,通过出端口 Port1 将替换了目的 MAC 地址的以太网报文发送给目的 VM。

[0217] 上述实施例通过将地址层次化和掩码的机制引入到二层转发表的管理,实现基于 mask 的二层转发表,极大的缩减二层转发表的表项的数量。通过缩减二层转发表的表项数量,可以有效解决大数据中心内二层转发的表项数目过大问题。同时,可以避免 MAC 地址学习时,因 HASH 冲突而使二层转发表的表项的实际数量不能达到设备支持的表项数量最大值的问题。

[0218] 需要说明的是,本发明实施例中仅以虚拟 MAC 地址格式为 6 字节的 OUI-DC ID-Device ID-host ID,掩码为 32 位的接入设备掩码和 16 位的数据中心掩码,进行了具体描述,并以此为基础描述了基于掩码的二层转发表的配置方式,以及各种场景下的基于二层转发表的报文转发流程。本领域技术人员应该能够理解,依据本发明实施例提供的原理,还可以设计出其它的虚拟 MAC 地址格式和相应的不同层次的 MAC 地址掩码,以及以此为基础的基于掩码的二层转发表及各种场景下的基于二层转发表实现的报文转发流程,只要能够通过不同层次的掩码将 VM 的虚拟 MAC 地址进行层次化的聚合,均应在本发明的保护范围之内。

[0219] 基于相同的技术构思,本发明实施例还提供了一种应用于大二层网络管理装置。

[0220] 参见图 8,为本发明实施例提供的网络管理装置 100 的结构示意图。该装置包括：虚拟 MAC 地址分配模块 81,进一步的,还可包括二层转发表配置模块 82、虚拟机迁移处理模块 83 和 ARP 处理模块 84 中的一个或组合,其中：

[0221] 虚拟 MAC 地址分配模块 81,用于根据预设的虚拟 MAC 地址配置规则为每个虚拟机设置一个虚拟媒体接入控制 MAC 地址。所述虚拟 MAC 地址配置规则包括：每个所述虚拟 MAC 地址由唯一性标识、网络标识、接入层设备标识以及主机标识组成且字节数等于每个所述虚拟机自身的真实 MAC 地址的字节数；其中，位于同一个数据中心且接入到相同接入层设备的所有虚拟机的虚拟 MAC 地址具有相同的唯一性标识、相同的网络标识、相同的接入层设备标识以及不同的主机标识；位于同一个数据中心且接入到不同接入层设备的所有虚拟机的虚拟 MAC 地址具有相同的唯一性标识、相同的网络标识、不同的接入层设备标识。

[0222] 进一步的,二层转发表配置模块 82,用于在一个接入层设备的二层转发表中配置至少一对虚拟机表项；每对虚拟机表项关联于所述接入层设备接入的一个虚拟机；每对虚拟机表项包括第一虚拟机表项和第二虚拟机表项。其中,第一虚拟机表项包含虚拟机所属虚拟局域网 VLAN 的标识、所述虚拟机的虚拟 MAC 地址、主机掩码、映射于所述虚拟 MAC 地址的所述虚拟机的真实 MAC 地址以及指向所述虚拟机的出端口；所述第二虚拟机表项包括虚拟机所属 VLAN 的标识,所述虚拟机的真实 MAC 地址,主机掩码,映射于所述真实 MAC 地址的所述虚拟机的虚拟 MAC 地址,以及指向所述虚拟机的出端口。

[0223] 进一步的,二层转发表配置模块 82,用于在一个接入层设备的二层转发表中配置至少一个接入设备表项；每个所述接入设备表项关联于同一数据中心内的一个其他接入层设备；每个接入设备表项包括：基于接入设备掩码的聚合虚拟 MAC 地址、接入设备掩码、以及到达所述接入设备表项所关联的接入层设备的出接口。其中,基于接入设备掩码的聚合虚拟 MAC 是通过接入设备掩码将接入到所述接入设备表项所关联的接入层设备的所有虚拟机的虚拟 MAC 聚合成的一个虚拟 MAC 地址,接入设备掩码的长度等于所述唯一性标识的字节数、所述网络标识的字节数以及所述接入层设备标识的字节数的和。

[0224] 进一步的,二层转发表配置模块 82,用于在一个接入层设备的二层转发表中配置至少一个数据中心表项,每个所述数据中心表项关联于一个其他数据中心；每个数据中心表项包括：基于数据中心掩码的聚合虚拟 MAC 地址、数据中心掩码以及到达一个核心层设备的出接口。其中,所述核心层设备负责与关联于所述数据中心表项的数据中心通信,所述基于数据中心掩码的聚合虚拟 MAC 地址是通过数据中心掩码将所述数据中心表项所关联的数据中心的所有虚拟机的虚拟 MAC 地址的聚合成的一个虚拟 MAC 地址；所述数据中心掩码等于所述唯一性标识的字节数与所述网络标识的字节数的和。

[0225] 进一步的,二层转发表配置模块 82,用于在一个接入层设备的二层转发表中配置至少一个网关转发表项,每个所述网关转发表项包括网关的所属 VLAN 的标识,所述网关的真实 MAC 地址,主机掩码,以及指向所述网关的出接口。其中,所述主机掩码的长度等于虚拟机的真实 MAC 地址的长度。

[0226] 进一步的,二层转发表配置模块 82,用于在接入组播源和组播接收端的一个接入层设备的二层转发表中配置组播转发表项；所述组播转发表项包括组播组所属 VLAN 的标识,组播地址,主机掩码,以及指向所述组播组的组播树的树根的出接口和指向组播接收端的出端口。所述网络管理装置在接入组播源的一个接入层设备的二层转发表中配置组播转

发表项 ;所述组播转发表项包括组播组所属 VLAN 的标识,组播地址,主机掩码,以及指向所述组播组的组播树的树根的出接口 ;所述网络管理装置在接入组播接收端的一个接入层设备的二层转发表中配置组播转发表项 ;所述组播转发表项包括组播组所属 VLAN 的标识,组播地址,主机掩码,以及指向所述组播接收端的出端口 ;其中,所述主机掩码长度等于虚拟机的真实 MAC 地址的长度。

[0227] 进一步的,二层转发表配置模块 82,用于在一个核心层设备的二层转发表配置至少一个接入设备表项,每个所述接入设备表项关联于相同数据中心内的一个接入层设备 ;所述接入设备表项包括 :基于接入设备掩码的聚合虚拟 MAC 地址、接入设备掩码以及到达所述接入设备表项所关联的接入层设备的出接口 ;其中,所述基于接入设备掩码的聚合虚拟 MAC 地址是通过接入设备掩码将接入到所述接入设备表项所关联的接入层设备的所有虚拟机的虚拟 MAC 地址的聚合成的一个虚拟 MAC 地址 ;所述接入设备掩码的长度等于所述唯一性标识的字节数、所述网络标识的字节数以及所述接入层设备标识的字节数的和。

[0228] 进一步的,二层转发表配置模块 82,用于在一个核心层设备的二层转发表配置至少一个数据中心表项,每个所述数据中心表项关联于所述核心层设备所在数据中心之外的其他数据中心 ;所述数据中心表项包括 :基于数据中心掩码的聚合虚拟 MAC 地址、数据中心掩码以及到达所述数据中心表项所关联的数据中心的出接口 ;其中,所述基于数据中心掩码的聚合虚拟 MAC 地址是通过数据中心掩码将所述数据中心表项所关联的数据中心的所有虚拟机的虚拟 MAC 地址的聚合成的一个虚拟 MAC 地址 ;所述数据中心掩码的长度等于所述唯一性标识的字节数与所述网络标识的字节数的和。

[0229] 进一步的,二层转发表配置模块 82,用于在一个核心层设备的二层转发表配置至少一个网关转发表项,每个所述网关转发表项包括网关所属 VLAN 的标识,所述网关的真实 MAC 地址,主机掩码,以及指向所述网关的出接口 ;其中,所述主机掩码的长度等于虚拟机的真实 MAC 地址的长度。

[0230] 进一步的,虚拟机迁移管理模块 83,用于在接收到虚拟机迁移的通知信息后,指示所述二层转发表配置模块删除源接入层设备的二层转发表中关联于迁移虚拟机的一对虚拟机表项,指示所述虚拟 MAC 地址分配模块为迁移虚拟机重新配置虚拟 MAC 地址,所述二层转发表配置模块在目标接入层设备的二层转发表中添加关联于迁移虚拟机的另一对虚拟机表项 ;所述源接入层设备连接于所述迁移虚拟机所在的源物理服务器,所述目标接入层设备连接于所述迁移虚拟机所在的目标物理服务器。

[0231] 进一步的,虚拟机迁移管理模块 83,用于在接收到删除虚拟机的通知信息后,指示所述二层转发表配置模块删除源接入层设备的二层转发表中关联于被删除的虚拟机的一对虚拟机表项 ;所述源接入层设备连接于所述被删除虚拟机所在的物理服务器。

[0232] 进一步的,虚拟机迁移模块 83,用于在接收到增加虚拟机的通知信息后,指示所述虚拟 MAC 地址分配模块为新增的虚拟机配置虚拟 MAC 地址,指示所述二层转发表配置模块在目标接入层设备的二层转发表中添加关联于新增的虚拟机的一对虚拟机表项 ;所述目标接入层设备连接于所述新增虚拟机所在的物理服务器。

[0233] 进一步的,ARP 处理模块 84,用于接收私有 ARP 请求报文,根据所述 ARP 请求报文的目标端 IP 地址查询对应的虚拟 MAC 地址,将所述 ARP 请求报文的目标端 IP 地址设置为 ARP 响应报文的发送端 IP 地址,将查询到的虚拟 MAC 地址设置为所述 ARP 响应报文的发送

端 MAC 地址,将所述 ARP 请求报文的发送端 IP 地址设置为所述 ARP 响应报文的目标端 IP 地址,将所述 ARP 请求报文的发送端 MAC 地址设置为所述 ARP 响应报文的目标 MAC 地址,将所述 ARP 响应报文封装为单播的私有 ARP 响应报文并发送。或者,ARP 处理模块 84 也可用于接收私有地址解析协议 ARP 请求报文,根据所述 ARP 请求报文的目标端 IP 地址查询对应的虚拟 MAC 地址,将所述 ARP 请求报文的目标端 IP 地址设置为 ARP 响应报文的发送端 IP 地址,将查询到的虚拟 MAC 地址设置为所述 ARP 响应报文的发送端 MAC 地址,将所述 ARP 请求报文的发送端 IP 地址设置为所述 ARP 响应报文的目标端 IP 地址,将所述 ARP 请求报文的发送端 IP 地址对应的虚拟 MAC 地址设置为所述 ARP 响应报文的目标 MAC 地址,将所述 ARP 响应报文封装为单播的私有 ARP 响应报文并发送。

[0234] 本发明实施例在实现方式上,如果对于处理性能要求很高的以太网交换机设备来说,需要使用硬件 ASIC 的方式来实现,如果性能要求不高,如路由器,vswitch 等,可以使用纯软件的方式来实现。

[0235] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到本发明可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台设备执行本发明各个实施例所述的方法。

[0236] 以上所述仅是本发明的优选实施方式,应当指出,对于本技术领域的普通技术人员来说,在不脱离本发明原理的前提下,还可以做出若干改进和润饰,这些改进和润饰也应视本发明的保护范围。

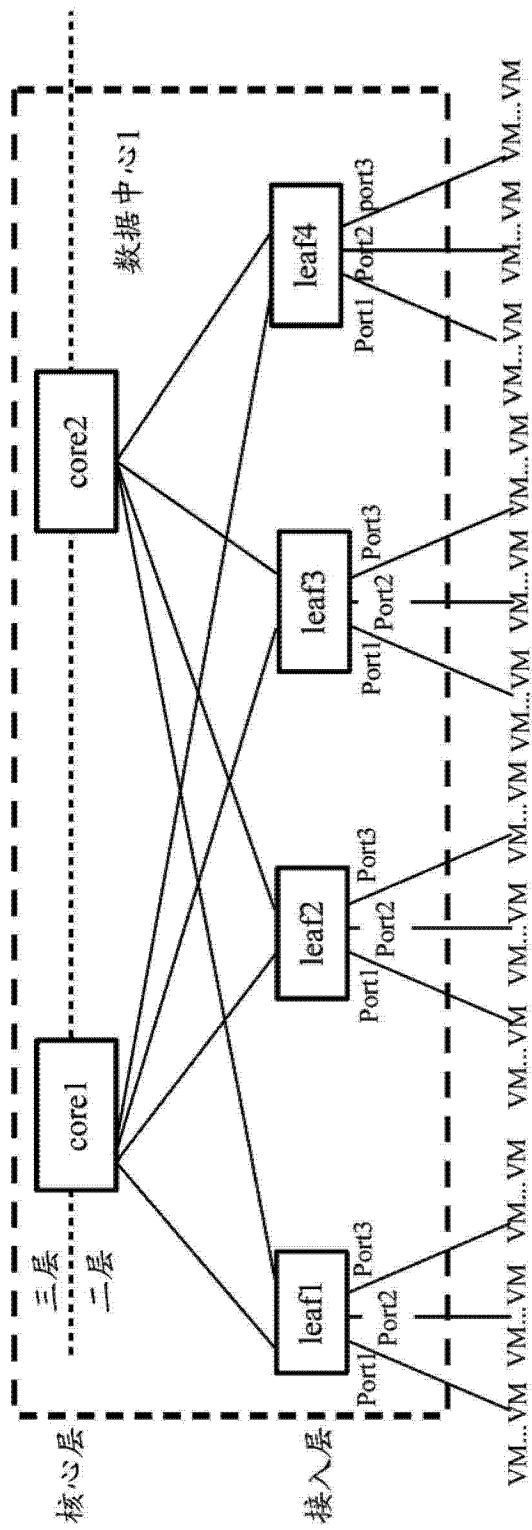


图 1

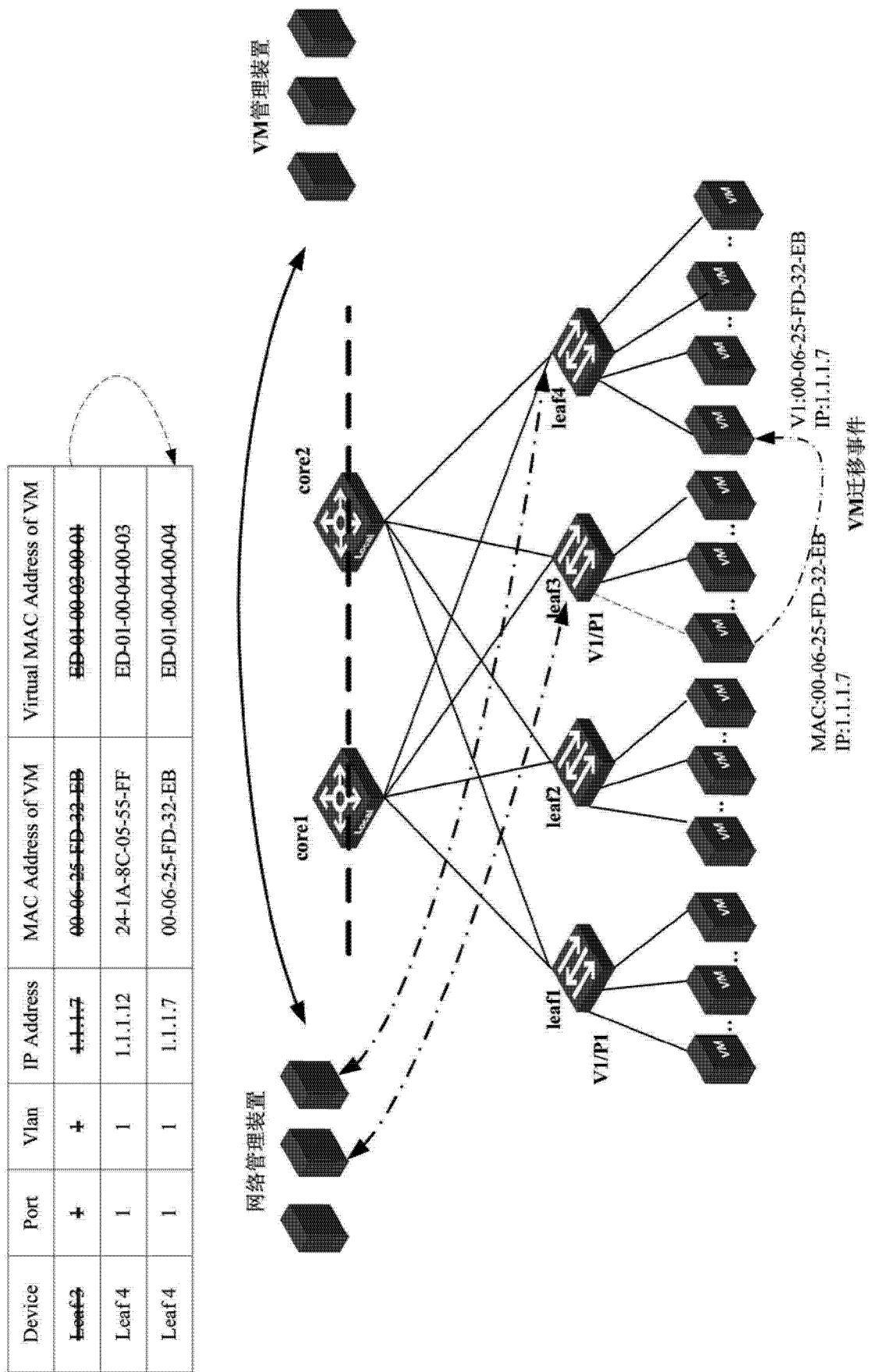


图 2

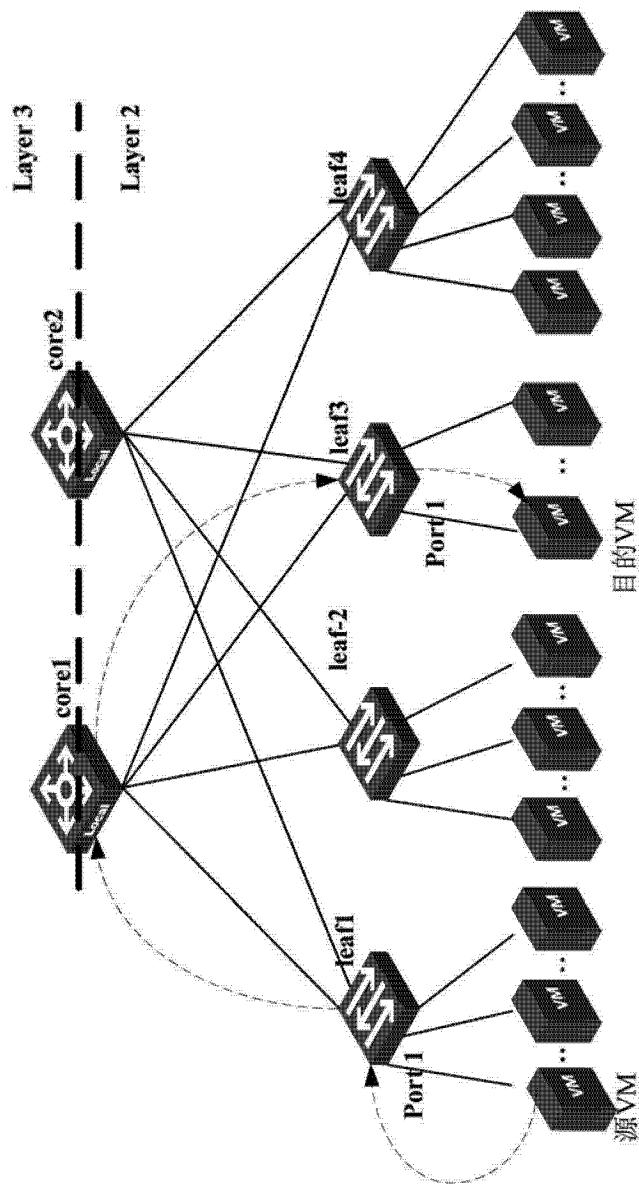
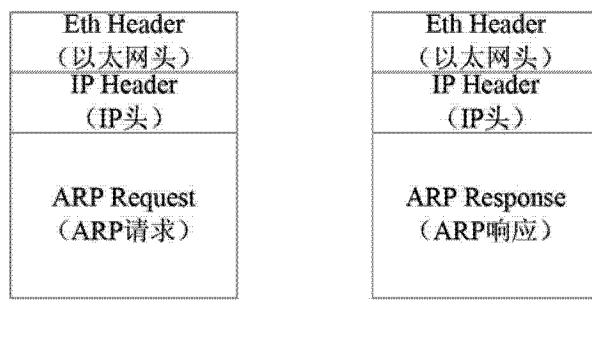


图 3A



私有ARP请求报文

私有ARP响应报文

图 3B

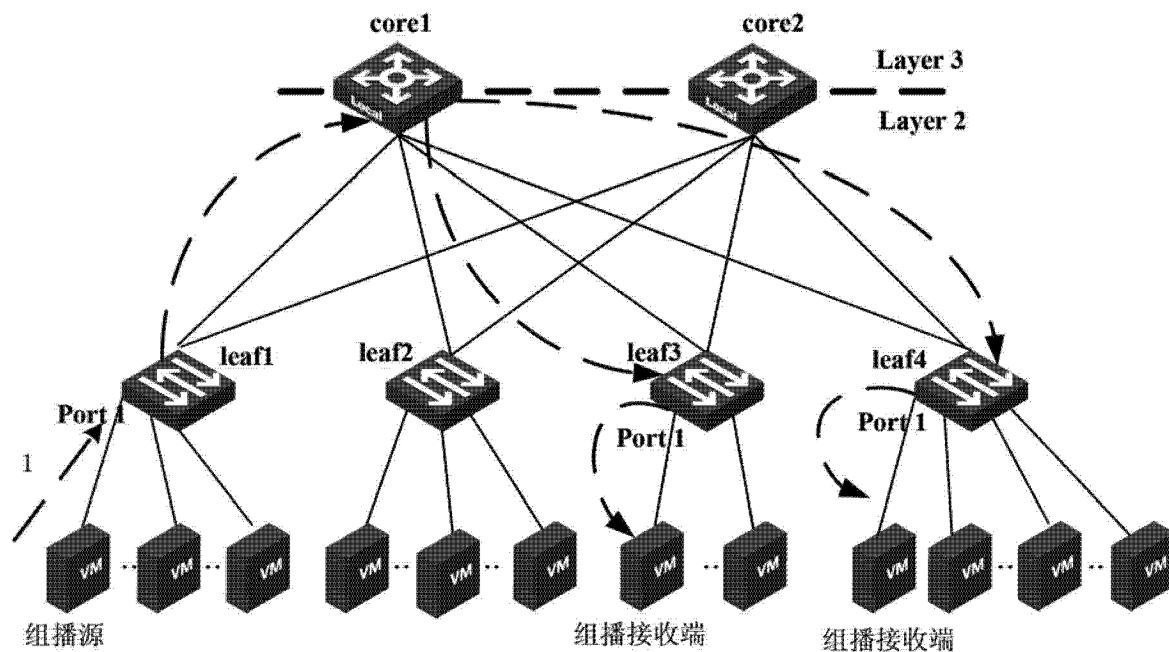


图 4

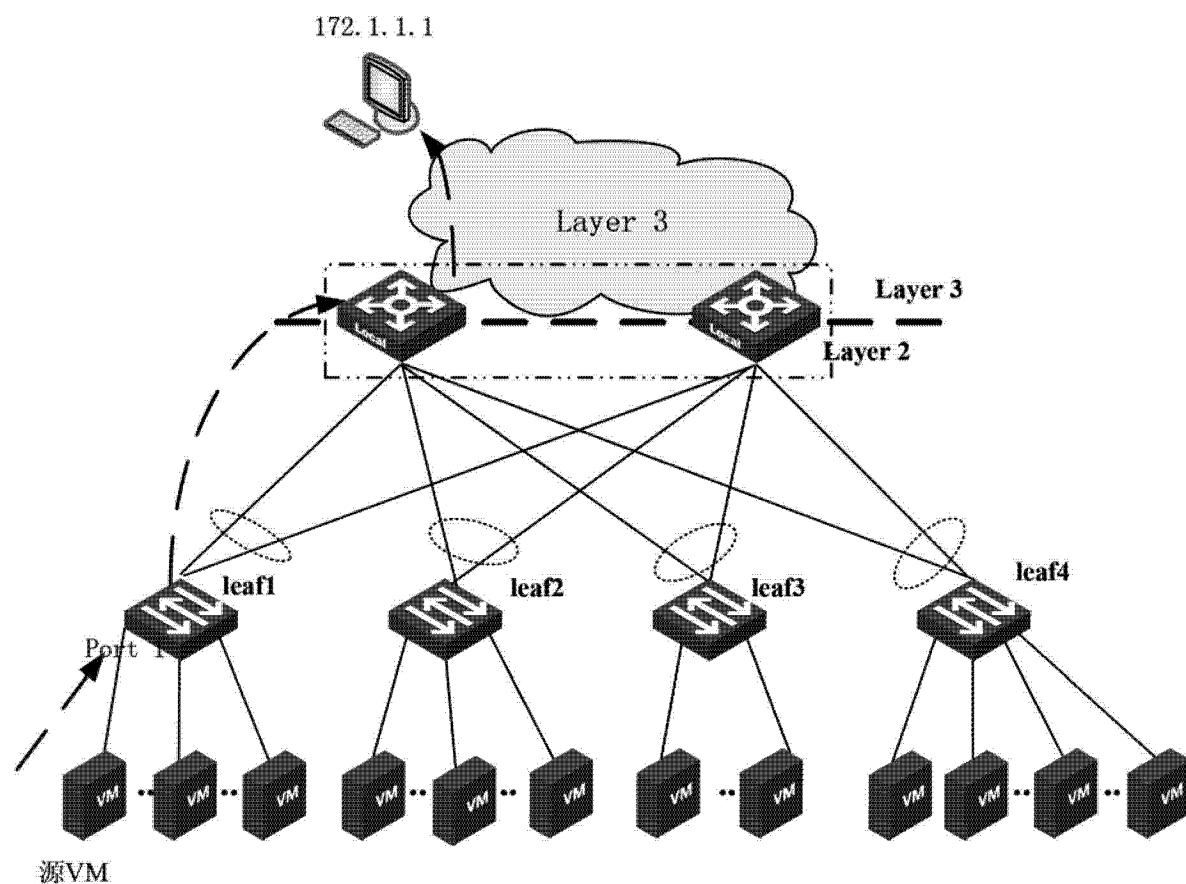


图 5

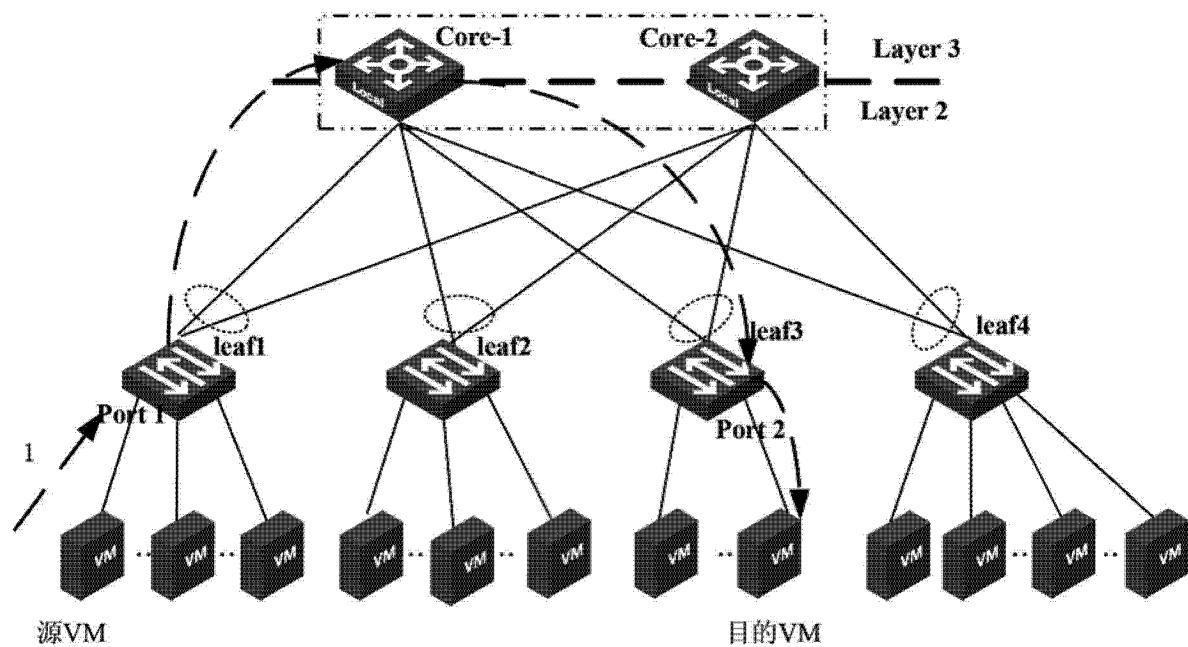


图 6

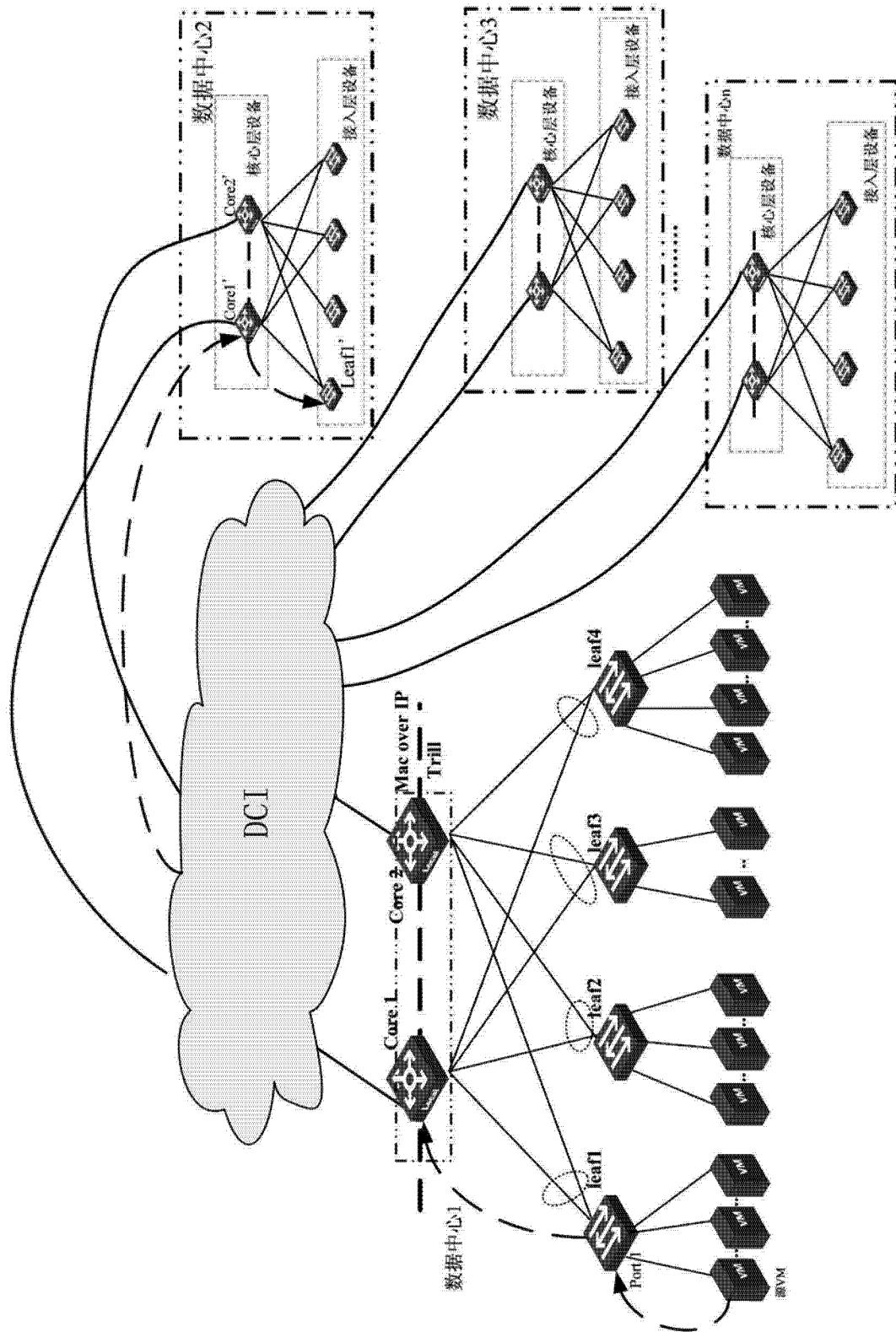


图 7

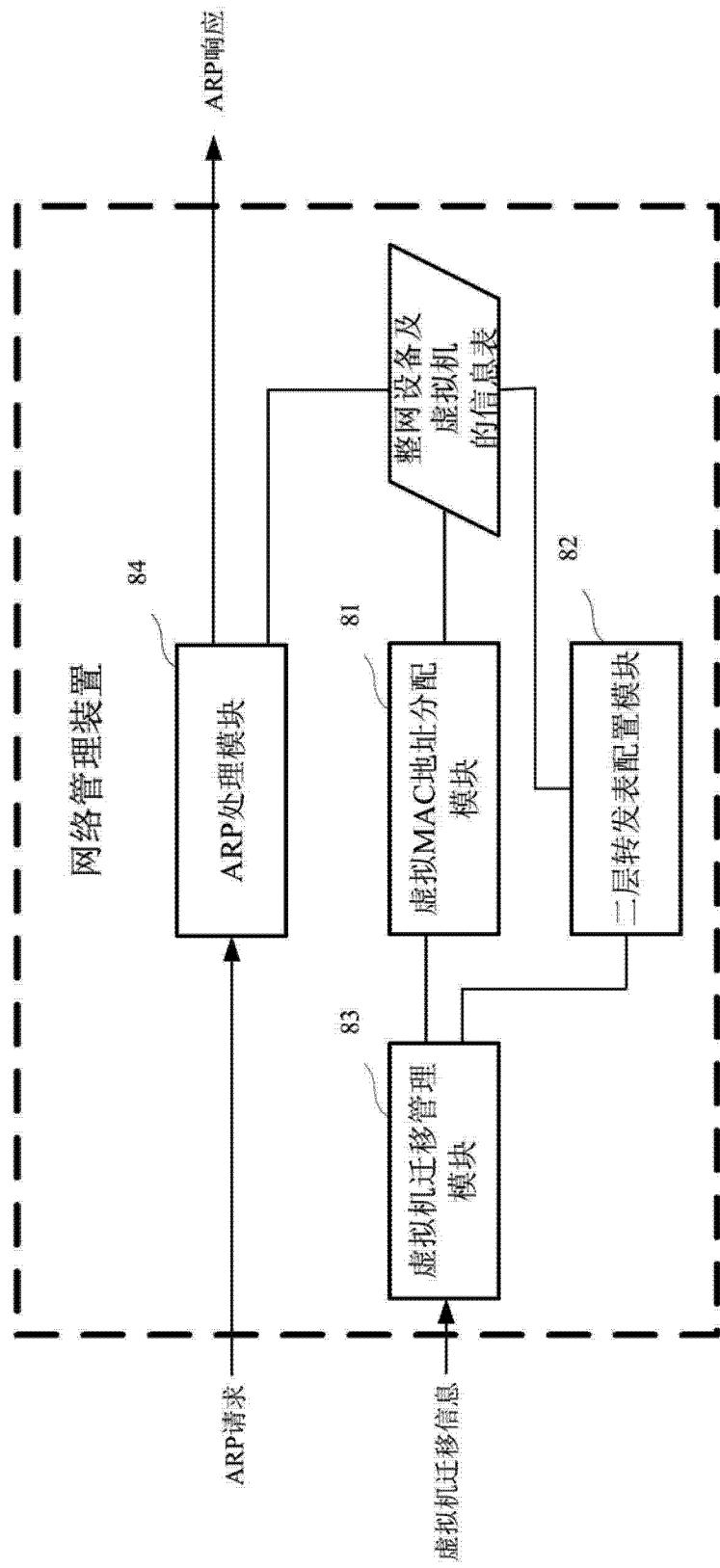


图 8