



(12)发明专利

(10)授权公告号 CN 103259727 B

(45)授权公告日 2016.12.28

(21)申请号 201310121235.1

(22)申请日 2013.04.09

(65)同一申请的已公布的文献号

申请公布号 CN 103259727 A

(43)申请公布日 2013.08.21

(66)本国优先权数据

201210318996.1 2012.08.31 CN

(73)专利权人 杭州华三通信技术有限公司

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

(72)发明人 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 12/741(2013.01)

H04L 12/947(2013.01)

(56)对比文件

CN 101557341 A, 2009.10.14,

CN 1852253 A, 2006.10.25,

US 2010290391 A1, 2010.11.18,

US 2007162594 A1, 2007.07.12,

审查员 徐佳

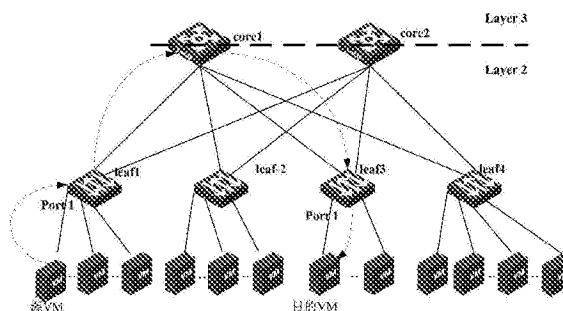
权利要求书4页 说明书29页 附图7页

(54)发明名称

一种报文转发方法及设备

(57)摘要

本发明公开了一种报文转发方法及设备,用以根据基于MAC地址掩码的二层转发表实现报文转发。本发明中,大二层网络中的接入层设备在进行报文转发时,基于MAC地址掩码的二层转发表,分别根据以太网报文的源MAC地址和目的MAC地址匹配二层转发表项,并当匹配到的二层转发表项中包含映射的MAC地址时,将所述以太网报文中的相应MAC地址替换为所述映射的MAC地址,接入层设备根据与目的MAC地址匹配的二层转发表项中的出接口转发以太网报文,从而实现基于MAC地址掩码的二层转发表项进行报文转发。



1. 一种报文转发方法,其特征在于,所述方法包括:

接收以太网报文;

根据接收到的以太网报文的源媒体接入控制MAC地址在二层转发表中执行基于掩码的MAC地址查找,确定查找到的表项中包含映射于所述源MAC地址的MAC地址,则根据映射于所述源MAC地址的MAC地址替换接收到的以太网报文中的源MAC地址;

根据接收到的以太网报文的的目的MAC地址在二层转发表中执行基于掩码的MAC地址查找,确定查找到的表项中未包含映射于所述目的MAC地址的MAC地址,则根据查找到的表项中的出接口对替换了源MAC地址的以太网报文进行大二层网络协议封装,并发送大二层网络协议封装后的以太网报文。

2. 如权利要求1所述的方法,其特征在于,根据接收到的以太网报文的的目的MAC地址查找到的表项为关联于同一数据中心内的一个其它接入层设备的接入设备表项;

所述根据查找到的表项中的出接口对替换了源MAC地址的以太网报文进行大二层网络协议封装,并发送大二层网络协议封装后的以太网报文,具体包括:

将查找到的接入设备表项中到达所述接入设备表项所关联的接入层设备的出接口,作为大二层网络协议封装中的出口信息,对替换了源MAC地址的以太网报文进行大二层网络协议封装,并从查找到的接入设备表项中的出接口向到达所述接入设备表项所关联的接入层设备发送大二层网络协议封装后的以太网报文。

3. 如权利要求1所述的方法,其特征在于,根据接收到的以太网报文的的目的MAC地址查找到的表项为关联于其它数据中心的数据中心表项;

所述根据查找到的表项中的出接口对替换了源MAC地址的以太网报文进行大二层网络协议封装,并发送大二层网络协议封装后的以太网报文,具体包括:

将查找到的数据中心表项中到达所述数据中心表项所关联的数据中心的出接口,作为大二层网络协议封装中的出口信息,对替换了源MAC地址的以太网报文进行大二层网络协议封装,并从查找到的数据中心表项中的出接口向到达所述数据中心表项所关联的数据中心发送大二层网络协议封装后的以太网报文。

4. 如权利要求1所述的方法,其特征在于,根据接收到的以太网报文的的目的MAC地址查找到的表项为网关转发表项;

所述根据查找到的表项中的出接口对替换了源MAC地址的以太网报文进行大二层网络协议封装,并发送大二层网络协议封装后的以太网报文,具体包括:

将查找到的网关转发表项中指向所述网关的出接口,作为大二层网络协议封装中的出口信息,对替换了源MAC地址的以太网报文进行大二层网络协议封装,并从查找到的网关转发表项中的出接口向所指向的网关发送大二层网络协议封装后的以太网报文。

5. 如权利要求1所述的方法,其特征在于,所述以太网报文为组播报文,根据接收到的以太网报文的的目的MAC地址查找到的表项为组播转发表项,所述组播转发表项的出接口包含指向所述组播组的组播树的树根的出接口;

所述根据查找到的表项中的出接口对替换了源MAC地址的以太网报文进行大二层网络协议封装,并发送大二层网络协议封装后的以太网报文,具体包括:

将查找到的组播转发表项中指向所述组播组的组播树的树根的出接口,作为大二层网络协议封装中的出口信息,对替换了源MAC地址的以太网报文进行大二层网络协议封装,并

从查找到的组播转发表项中的出接口向所述组播组的组播树的树根发送大二层网络协议封装后的组播报文。

6. 如权利要求1所述的方法, 其特征在于, 所述以太网报文为组播报文, 根据接收到的以太网报文的目MAC地址查找到的表项为组播转发表项, 所述组播转发表项的出接口包含指向组播接收端的出端口;

所述方法还包括:

根据查找到的组播转发表项中的出接口向所指向的组播接收客户端发送替换了源MAC地址的组播报文。

7. 如权利要求1所述的方法, 其特征在于, 所述方法还包括:

根据接收到的以太网报文的目MAC地址在二层转发表中执行基于掩码的MAC地址查找时, 确定查找到的表项中包含映射于所述目MAC地址的MAC地址, 则根据所述映射于所述目MAC地址的MAC地址替换接收到的以太网报文中的目MAC地址, 并根据查找到的表项中的出端口发送替换了源MAC地址和目MAC地址的以太网报文。

8. 如权利要求1所述的方法, 其特征在于, 所述方法还包括:

接收大二层网络协议封装的以太网报文;

根据所述以太网报文的大二层网络协议封装确定本设备为所述大二层网络协议封装的以太网报文的目设备, 则解除大二层网络协议封装;

根据解封装后的以太网报文的目MAC地址在二层转发表中执行基于掩码的MAC地址查找, 根据与所述目MAC地址匹配的二层转发表项中映射于所述目MAC地址的MAC地址, 替换解封装后的以太网报文的目MAC地址, 通过与所述目MAC地址匹配的二层转发表项的出端口发送替换了目MAC地址的以太网报文。

9. 如权利要求1所述的方法, 其特征在于, 所述方法还包括:

接收地址解析协议ARP请求报文;

将收到的ARP请求报文转换为私有ARP请求报文;

根据所述私有ARP报文的源MAC地址和目MAC地址在所述二层转发表中执行基于掩码的MAC地址查找;

根据所述私有ARP报文的目MAC地址的匹配表项中到达大二层网络内网关的出接口对替换了私有ARP报文进行大二层网络协议封装, 发送大二层网络协议封装后的ARP请求报文。

10. 如权利要求1所述的方法, 其特征在于, 所述方法还包括:

接收大二层网络协议封装的私有ARP响应报文;

解除大二层网络协议封装, 将私有ARP响应报文转换为带有以太网头的ARP响应报文;

根据所述ARP响应报文的目MAC地址的匹配表项的映射MAC地址替换所述ARP响应报文的目MAC地址;

根据所述ARP响应报文的目MAC地址的匹配表项的出端口发送替换了目MAC地址的ARP响应报文。

11. 如权利要求1-10任意项所述的方法, 其特征在于, 所述大二层网络协议是多链路透明互联Trill协议或最短路径桥SPB协议。

12. 一种路由交换设备, 其特征在于, 所述路由交换设备包括:

接收模块,用于接收以太网报文;

转发模块,用于根据接收到的以太网报文的源媒体接入控制MAC地址在二层转发表中执行基于掩码的MAC地址查找,确定查找到的表项中包含映射于所述源MAC地址的MAC地址,则根据映射于所述源MAC地址的MAC地址替换接收到的以太网报文中的源MAC地址;根据接收到的以太网报文的目的地MAC地址在二层转发表中执行基于掩码的MAC地址查找,确定查找到的表项中未包含映射于所述目的地MAC地址的MAC地址,则根据查找到的表项中的出接口对替换了源MAC地址的以太网报文进行大二层网络协议封装;

发送模块,用于发送大二层网络协议封装后的以太网报文。

13.如权利要求12所述的路由交换设备,其特征在于,所述转发模块根据接收到的以太网报文的目的地MAC地址查找到的表项为关联于同一数据中心内的一个其它接入层设备的接入设备表项;

所述转发模块具体用于,将查找到的接入设备表项中到达所述接入设备表项所关联的接入层设备的出接口,作为大二层网络协议封装中的出口信息,对替换了源MAC地址的以太网报文进行大二层网络协议封装;

所述发送模块具体用于,从查找到的接入设备表项中的出接口向到达所述接入设备表项所关联的接入层设备发送大二层网络协议封装后的以太网报文。

14.如权利要求12所述的路由交换设备,其特征在于,所述转发模块根据接收到的以太网报文的目的地MAC地址查找到的表项为关联于其它数据中心的数据中心表项;

所述转发模块具体用于,将查找到的数据中心表项中到达所述数据中心表项所关联的数据中心的出接口,作为大二层网络协议封装中的出口信息,对替换了源MAC地址的以太网报文进行大二层网络协议封装;

所述发送模块具体用于,从查找到的数据中心表项中的出接口向到达所述数据中心表项所关联的数据中心发送大二层网络协议封装后的以太网报文。

15.如权利要求12所述的路由交换设备,其特征在于,所述转发模块根据接收到的以太网报文的目的地MAC地址查找到的表项为网关转发表项;

所述转发模块具体用于,将查找到的网关转发表项中指向所述网关的出接口,作为大二层网络协议封装中的出口信息,对替换了源MAC地址的以太网报文进行大二层网络协议封装;

所述发送模块具体用于,从查找到的网关转发表项中的出接口向所指向的网关发送大二层网络协议封装后的以太网报文。

16.如权利要求12所述的路由交换设备,其特征在于,所述接收模块接收到的所述以太网报文为组播报文;

所述转发模块根据接收到的以太网报文的目的地MAC地址查找到的表项为组播转发表项,所述组播转发表项的出接口包含指向所述组播组的组播树的树根的出接口;

所述转发模块具体用于,将查找到的组播转发表项中指向所述组播组的组播树的树根的出接口,作为大二层网络协议封装中的出口信息,对替换了源MAC地址的以太网报文进行大二层网络协议封装;

所述发送模块具体用于,从查找到的组播转发表项中的出接口向所述组播组的组播树的树根发送大二层网络协议封装后的组播报文。

17. 如权利要求12所述的路由交换设备,其特征在于,所述接收模块接收到的以太网报文为组播报文;

所述转发模块根据接收到的以太网报文的目的地MAC地址查找到的表项为组播转发表项,所述组播转发表项的出接口包含指向组播接收端的出端口;

所述发送模块还用于,根据查找到的组播转发表项中的出接口向所指向的组播接收客户端发送替换了源MAC地址的组播报文。

18. 如权利要求12所述的路由交换设备,其特征在于,所述转发模块还用于,根据接收到的以太网报文的目的地MAC地址在二层转发表中执行基于掩码的MAC地址查找时,确定查找到的表项中包含映射于所述目的地MAC地址的MAC地址,则根据所述映射于所述目的地MAC地址的MAC地址替换接收到的以太网报文中的目的地MAC地址;

所述发送模块还用于,根据查找到的表项中的出端口发送替换了源MAC地址和目的地MAC地址的以太网报文。

19. 如权利要求12所述的路由交换设备,其特征在于,所述接收模块还用于,接收大二层网络协议封装的以太网报文;

所述转发模块还用于,根据所述以太网报文的大二层网络协议封装确定本设备为所述大二层网络协议封装的以太网报文的目的地设备,则解除大二层网络协议封装;根据解封装后的以太网报文的目的地MAC地址在二层转发表中执行基于掩码的MAC地址查找,根据与所述目的地MAC地址匹配的二层转发表项中映射于所述目的地MAC地址的MAC地址,替换解封装后的以太网报文的目的地MAC地址;

所述发送模块还用于,通过与所述目的地MAC地址匹配的二层转发表项的出端口发送替换了目的地MAC地址的以太网报文。

20. 如权利要求12所述的路由交换设备,其特征在于,还包括:地址解析协议ARP处理模块;

所述接收模块还用于,接收ARP请求报文;

所述ARP处理模块用于,将收到的ARP请求报文转换为私有ARP请求报文;根据所述私有ARP报文的源MAC地址和目的地MAC地址在所述二层转发表中执行基于掩码的MAC地址查找;根据所述私有ARP报文的目的地MAC地址的匹配表项中到达大二层网络内网关的出接口对替换了私有ARP报文进行大二层网络协议封装;

所述发送模块还用于,发送大二层网络协议封装后的ARP请求报文。

21. 如权利要求12所述的路由交换设备,其特征在于,还包括:ARP处理模块;

所述接收模块还用于,接收大二层网络协议封装的私有ARP响应报文;

所述ARP处理模块还用于,解除大二层网络协议封装,将私有ARP响应报文转换为带有以太网头的ARP响应报文;根据所述ARP响应报文的目的地MAC地址的匹配表项的映射MAC地址替换所述ARP响应报文的目的地MAC地址;

所述发送模块还用于,根据所述ARP响应报文的目的地MAC地址的匹配表项的出端口发送替换了目的地MAC地址的ARP响应报文。

22. 如权利要求12-21任意项所述的路由交换设备,其特征在于,所述大二层网络协议是多链路透明互联Trill协议或最短路径桥SPB协议。

一种报文转发方法及设备

技术领域

[0001] 本发明涉及通信技术领域,尤其涉及一种报文转发方法及设备。

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-2 network),实现VM迁移。

[0003] 以包含多个数据中心的大二层网络(Very Large Layer-2 Network)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,接入设备需要在二层转发表中学习大量的MAC地址,执行转发。

发明内容

[0006] 本发明的目的在于提供一种报文转发方法及设备。为实现该目的,本发明实施例提供了以下技术方案:

[0007] 一种报文转发方法,所述方法包括:

[0008] 接收以太网报文;

[0009] 根据接收到的以太网报文的源媒体接入控制MAC地址在二层转发表中执行查找,确定查找到的表项中包含映射于所述源MAC地址的MAC地址,则根据映射于所述源MAC地址的MAC地址替换接收到的以太网报文中的源MAC地址;

[0010] 根据接收到的以太网报文的目的MAC地址在二层转发表中执行查找,确定查找到的表项中未包含映射于所述目的MAC地址的MAC地址,则根据查找到的表项中的出接口对替换了源MAC地址的以太网报文进行大二层网络协议封装,并发送大二层网络协议封装后的以太网报文。

[0011] 一种路由交换设备,包括:

[0012] 接收模块,用于接收以太网报文;

[0013] 转发模块,用于根据接收到的以太网报文的源媒体接入控制MAC地址在二层转发表中执行查找,确定查找到的表项中包含映射于所述源MAC地址的MAC地址,则根据映射于所述源MAC地址的MAC地址替换接收到的以太网报文中的源MAC地址;根据接收到的以太网报文的目的MAC地址在二层转发表中执行查找,确定查找到的表项中未包含映射于所述目的MAC地址的MAC地址,则根据查找到的表项中的出接口对替换了源MAC地址的以太网报文进行大二层网络协议封装;

[0014] 发送模块,用于发送大二层网络协议封装后的以太网报文。

[0015] 本发明的上述实施例中,大二层网络中的接入层设备在进行报文转发时,基于MAC地址掩码的二层转发表,根据以太网报文的目的MAC地址匹配二层转发表项,根据与目的MAC地址匹配的二层转发表项中的出接口转发以太网报文,从而实现基于MAC地址掩码的二层转发表项进行报文转发。

附图说明

[0016] 图1为Trill组网示意图;

[0017] 图2为本发明实施例中的VM迁移的示意图;

[0018] 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;

[0019] 图3B为本发明实施例提供的私有ARP报文的示意图;

[0020] 图4为本发明实施例中组播报文转发的示意图;

[0021] 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;

[0022] 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;

[0023] 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;

[0024] 图8为本发明实施例提供的路由交换设备的结构示意图。

具体实施方式

[0025] 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。

[0026] 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MAC over IP技术。

[0027] 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。

[0028] 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。

[0029] 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)。

[0030] 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:

[0031] 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。

[0032] Data Centre ID(1字节):数据中心标识或称网络标识。

[0033] Device ID(2字节):接入层设备的标识。

[0034] 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。

[0035] 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:

[0036] 主机掩码(Host mask):ff-ff-ff-ff-ff-ff。

[0037] 接入设备掩码(access device mask):ff-ff-ff-ff-00-00。

[0038] 数据中心掩码(data centre mask):ff-ff-00-00-00-00。

[0039] 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):

[0040] 表1:

[0041]

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
leaf1	DC1_leaf1	IP1	Port1	1	1.1.1.1	00-11-11-11-11-11	ED-01-00-01-00-01	ED-01-00-01-00-00	ED-01-00-00-00-00
	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-c2-0f-9a-86	ED-01-00-01-00-05		
	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	ED-01-00-02-00-00	
	DC1_leaf2	IP1	Port1	1	1.1.1.5	00-6B-28-07-44-3F	ED-01-00-02-00-02		
	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-	ED-01-00-		

[0042]

						F7-B0-94	02-00-05		
	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		
	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		
	DC1_leaf3	IP1	Port2	2	2.2.2.8	FC-FB-FB-11-22-33	ED-01-00-03-00-05	ED-01-00-03-00-00	
	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		
	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	ED-01-00-04-00-00	
	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		

[0043]

	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			

[0044] “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。

[0045] “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。

[0046] 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。

[0047] 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。

[0048] 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。

[0049] 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。

[0050] 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。

[0051] 表2.1：

[0052]

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

[0053]

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 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		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

[0054] 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。

[0055] 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）。

[0056] 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)。

[0057] 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。

[0058] 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:

[0059] 表2.2:leaf1上二层转发表中的组播转发表项

[0060]

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

[0061] 表2.3:leaf3上二层转发表中的组播转发表项

[0062]

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

[0063] 表2.4:leaf4上二层转发表中的组播转发表项

[0064]

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

[0065] 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。

[0066] 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。

[0067] core1的二层转发表至少包括表2.5所示的表项:

[0068] 表2.5

[0069]

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

[0070] core2的二层转发表至少包含表2.6所示的表项：

[0071] 表2.6

[0072]

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

[0073] 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。

[0074] 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。

[0075] 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。

[0076] 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。

[0077] 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。

[0078] 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。

[0079] 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。

[0080] 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。

[0081] 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。

[0082] 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。

[0083] 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。

[0084] 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。

[0085] 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。

[0086] 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。

[0087] 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)

[0088] 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。

[0089] 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)。

[0090] 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:

[0091] (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。

[0092] (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。

[0093] (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。

[0094] 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。

[0095] (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。

[0096] (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。

[0097] 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。

[0098] 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。

[0099] 图3A中,leaf1的二层转发表中至少包括表3.1所示的表项:

[0100] 表3.1

[0101]

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

[0102]

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] core1的二层转发表中至少包括表3.3所示的表项:

[0107] 表3.3

[0108]

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

[0109] 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。

[0110] 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。

[0111] 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。

[0112] 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。

[0113] 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。

[0114] 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。

[0115] 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。

[0116] 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-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。

[0117] 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。

[0118] 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。

[0119] 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。

[0120] 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。

[0121] 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。

[0122] 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VLAN 1的网关core1。

[0123] 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VLAN 1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。

[0124] 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。

[0125] 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。

[0126] 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。

[0127] 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VLAN 1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。

[0128] 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。

[0129] 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。

[0130] 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。

[0131] 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,请求数据中心内VLAN1内所有VM的ARP信息。

[0132] 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。

[0133] 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地址是设置为VLAN 1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VLAN 1的网关core1。

[0134] 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。

[0135] 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。

[0136] 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。

[0137] 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。

[0138] leaf1的二层转发表中至少包括表4.1所示的表项：

[0139] 表4.1

[0140]

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

[0141] leaf3的二层转发表中至少包括表4.2所示的表项：

[0142] 表4.2

[0143]

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

[0144] leaf4的二层转发表中至少包括表4.2所示的以下表项：

[0145] 表4.3

[0146]

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

[0147]

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

[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桥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	VLAN	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	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-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VLAN 1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
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

[0182] core1和core2的二层转发表中至少包括表6.2所示的以下表项:

[0183] 表6.2

[0184]

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

[0185] leaf3的二层转发表中至少包括表6.3所示的以下表项:

[0186] 表6.3

[0187]

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

[0188] 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VLAN 1网关的MAC地址00-E0-FC-11-11-11。

[0189] 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。

[0190] 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。

[0191] 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。

[0192] 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VLAN 2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。

[0193] 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。

[0194] 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。

[0195] 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。

[0196] 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。

[0197] 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。

[0198] 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。

[0199] leaf1的二层转发表中至少包括表7.1所示的以下表项:

[0200] 表7.1

[0201]

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

[0202] 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:

[0203] 表7.2

[0204]

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

[0205] 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:

[0206] 表7.3

[0207]

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

[0208] leaf1'的二层转发表中至少包括表7.4所示的以下表项:

[0209] 表7.4

[0210]

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

[0211]

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,本发明实施例还通过了一种路由交换设备。该路由交换设备可作为接入层设备应用于上述实施例中。

[0220] 参见图8,为本发明实施例提供的路由交换设备的结构示意图。如图所示,该路由交换设备可包括接收模块81、发送模块82、转发模块83,以及至少一个端口85,还可以进一步包括ARP处理模块84。接收模块81、发送模块82、转发模块83和ARP处理模块84之间通过总线80进行信息交互。接收模块81可通过本设备的端口85接收报文,发送模块82可通过本设备的端口85发送报文。

[0221] 下面详细描述该路由交换设备中的各功能模块。

[0222] 接收模块81用于接收以太网报文。转发模块82用于根据接收到的以太网报文的源MAC地址在二层转发表中执行查找,确定查找到的表项中包含映射于所述源MAC地址的MAC

地址,则根据映射于所述源MAC地址的MAC地址替换接收到的以太网报文中的源MAC地址;根据接收到的以太网报文的的目的MAC地址在二层转发表中执行查找,确定查找到的表项中未包含映射于所述目的MAC地址的MAC地址,则根据查找到的表项中的出接口对替换了源MAC地址的以太网报文进行大二层网络协议封装。发送模块83用于发送大二层网络协议封装后的以太网报文。其中,所述大二层网络协议可以是Trill协议或SPB协议。

[0223] 具体的,转发模块83根据接收到的以太网报文的的目的MAC地址查找到的表项为关联于同一数据中心内的一个其它接入层设备的接入设备表项。相应的,转发模块83将查找到的接入设备表项中到达所述接入设备表项所关联的接入层设备的出接口,作为大二层网络协议封装中的出口信息,对替换了源MAC地址的以太网报文进行大二层网络协议封装;发送模块82从查找到的接入设备表项中的出接口向到达所述接入设备表项所关联的接入层设备发送大二层网络协议封装后的以太网报文。

[0224] 具体的,转发模块84根据接收到的以太网报文的的目的MAC地址查找到的表项为关联于其它数据中心的数据中心表项。相应的,转发模块82将查找到的数据中心表项中到达所述数据中心表项所关联的数据中心的出接口,作为大二层网络协议封装中的出口信息,对替换了源MAC地址的以太网报文进行大二层网络协议封装;发送模块82从查找到的数据中心表项中的出接口向到达所述数据中心表项所关联的数据中心发送大二层网络协议封装后的以太网报文。

[0225] 具体的,转发模块83根据接收到的以太网报文的的目的MAC地址查找到的表项为网关转发表项。相应的,转发模块83将查找到的网关转发表项中指向所述网关的出接口,作为大二层网络协议封装中的出口信息,对替换了源MAC地址的以太网报文进行大二层网络协议封装;发送模块82从查找到的网关转发表项中的出接口向所指向的网关发送大二层网络协议封装后的以太网报文。

[0226] 具体的,接收模块81接收到的所述以太网报文为组播报文。相应的,转发模块83根据接收到的以太网报文的的目的MAC地址查找到的表项为组播转发表项,所述组播转发表项的出接口包含指向所述组播组的组播树的树根的出接口;转发模块83将查找到的组播转发表项中指向所述组播组的组播树的树根的出接口,作为大二层网络协议封装中的出口信息,对替换了源MAC地址的以太网报文进行大二层网络协议封装;发送模块82从查找到的组播转发表项中的出接口向所述组播组的组播树的树根发送大二层网络协议封装后的组播报文。

[0227] 具体的,接收模块81接收到的以太网报文为组播报文。相应的,转发模块83根据接收到的以太网报文的的目的MAC地址查找到的表项为组播转发表项,所述组播转发表项的出接口包含指向组播接收端的出端口;发送模块82根据查找到的组播转发表项中的出接口向所指向的组播接收客户端发送替换了源MAC地址的组播报文。

[0228] 进一步的,转发模块83根据接收到的以太网报文的的目的MAC地址在二层转发表中执行查找时,确定查找到的表项中包含映射于所述目的MAC地址的MAC地址,则根据所述映射于所述目的MAC地址的MAC地址替换接收到的以太网报文中的目的MAC地址。相应的,发送模块82根据查找到的表项中的出端口发送替换了源MAC地址和目的MAC地址的以太网报文。

[0229] 进一步的,接收模块81还用于接收大二层网络协议封装的以太网报文。相应的,转发模块83还用于根据所述以太网报文的大二层网络协议封装确定本设备为所述大二层网

络协议封装的以太网报文的目的设备,则解除大二层网络协议封装;根据解封装后的以太网报文的目的MAC地址在二层转发表中执行查找,根据与所述目的MAC地址匹配的二层转发表项中映射于所述目的MAC地址的MAC地址,替换解封装后的以太网报文的目的MAC地址;发送模块82还用于通过与所述目的MAC地址匹配的二层转发表项的出端口发送替换了目的MAC地址的以太网报文。

[0230] 进一步的,接收模块还用于接收ARP请求报文。相应的,ARP处理模块84将收到的ARP请求报文转换为私有ARP请求报文;根据所述私有ARP报文的源MAC地址和目的MAC地址在所述二层转发表中执行查找;根据所述私有ARP报文的目的MAC地址的匹配表项中到达大二层网络内网关的出接口对替换了私有ARP报文进行大二层网络协议封装;发送模块82还用于发送大二层网络协议封装后的ARP请求报文。

[0231] 进一步的,接收模块81还用于接收大二层网络协议封装的私有ARP响应报文。相应的,ARP处理模块84还用于解除大二层网络协议封装,将私有ARP响应报文转换为带有以太网头的ARP响应报文,根据所述ARP响应报文的目的MAC地址的匹配表项的映射MAC地址替换所述ARP响应报文的目的MAC地址;发送模块82还用于根据所述ARP响应报文的目的MAC地址的匹配表项的出端口发送替换了目的MAC地址的ARP响应报文。

[0232] 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。

[0233] 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。

[0234] 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。

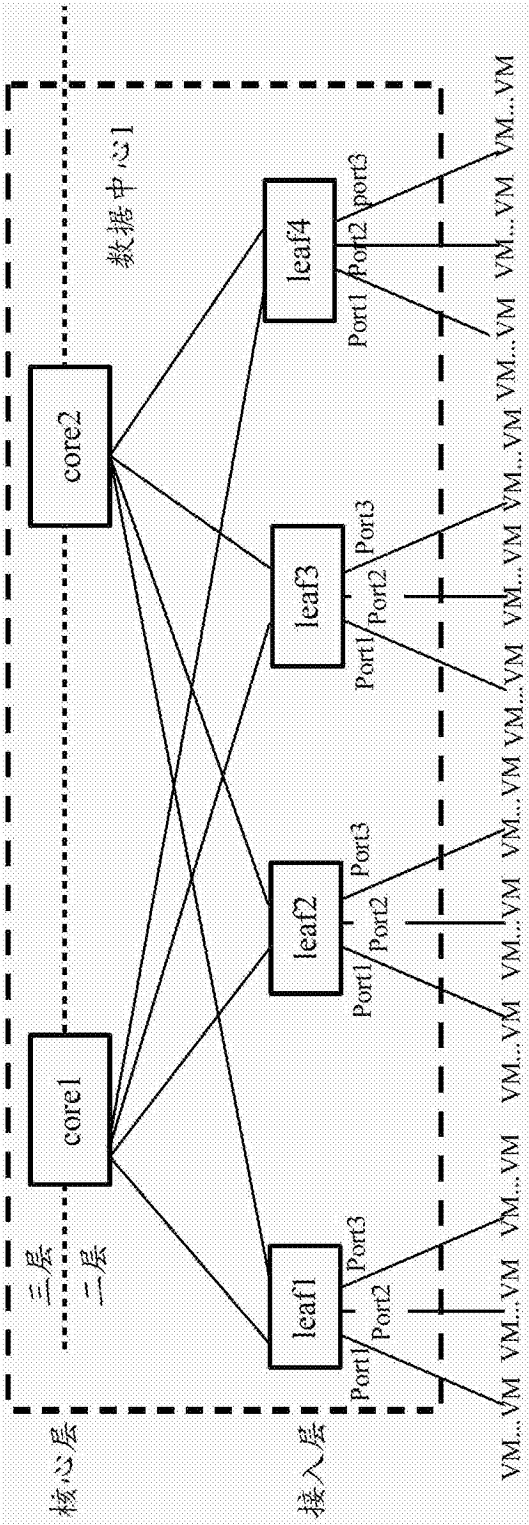


图1

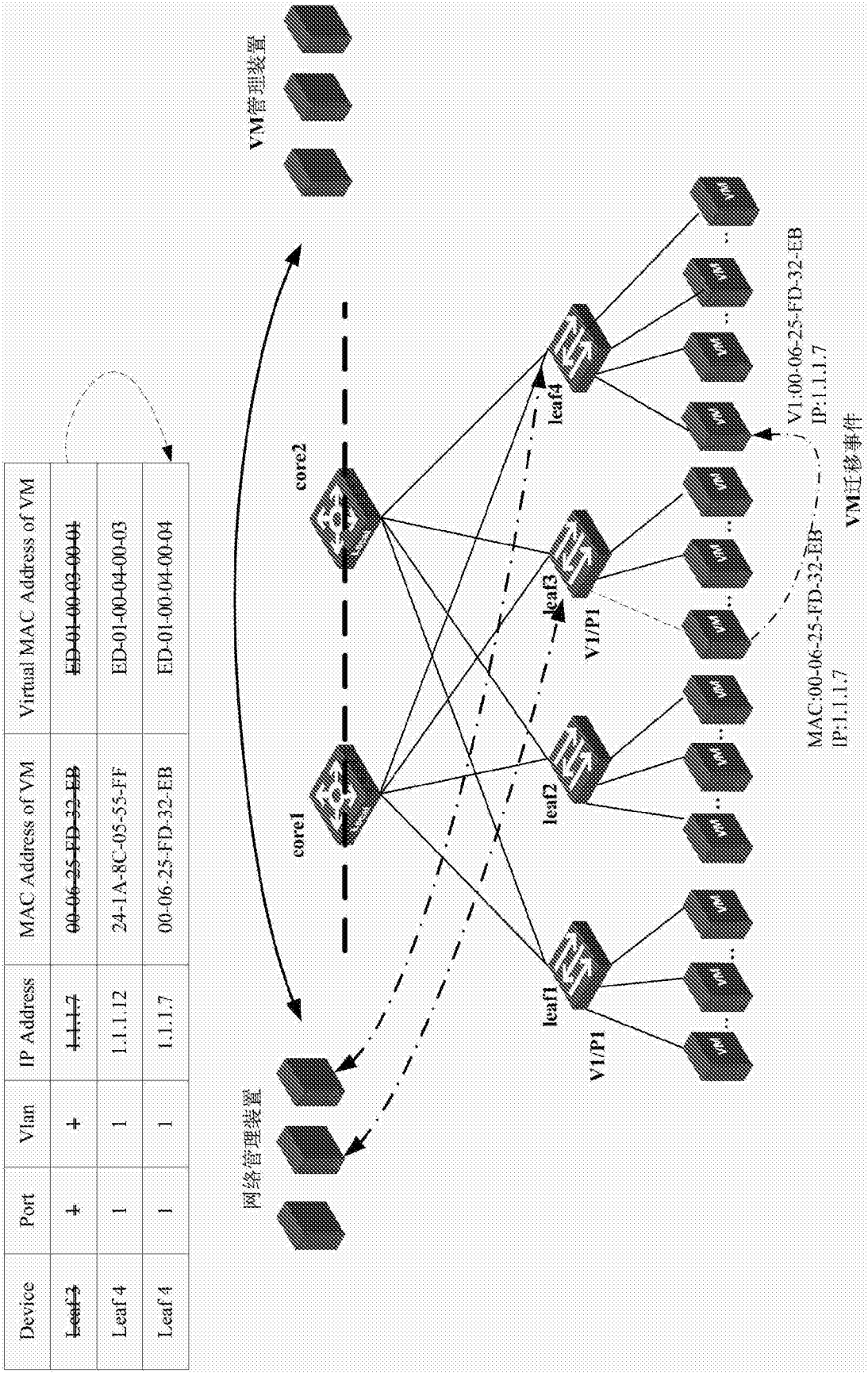


图2

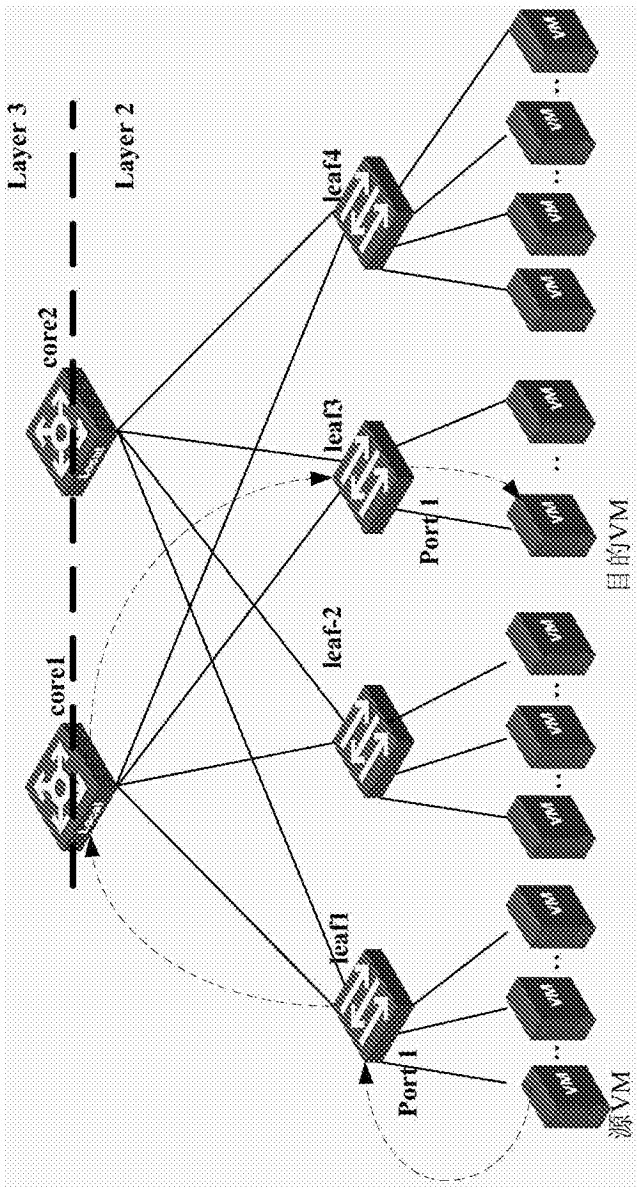
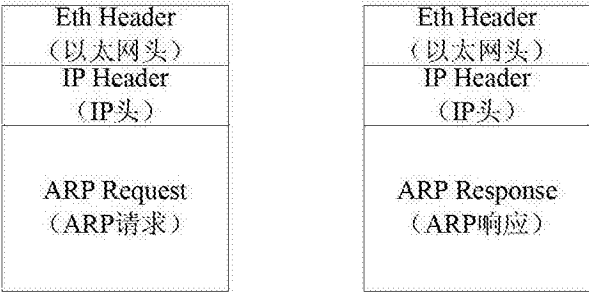


图3A



私有ARP请求报文

私有ARP响应报文

图3B

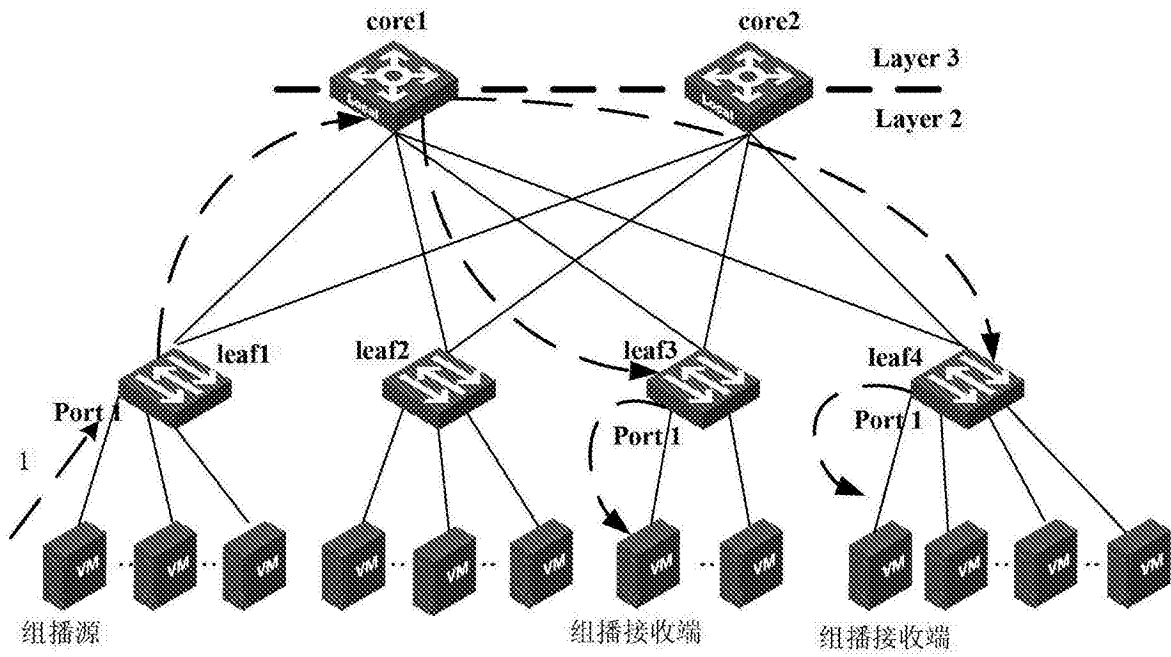


图4

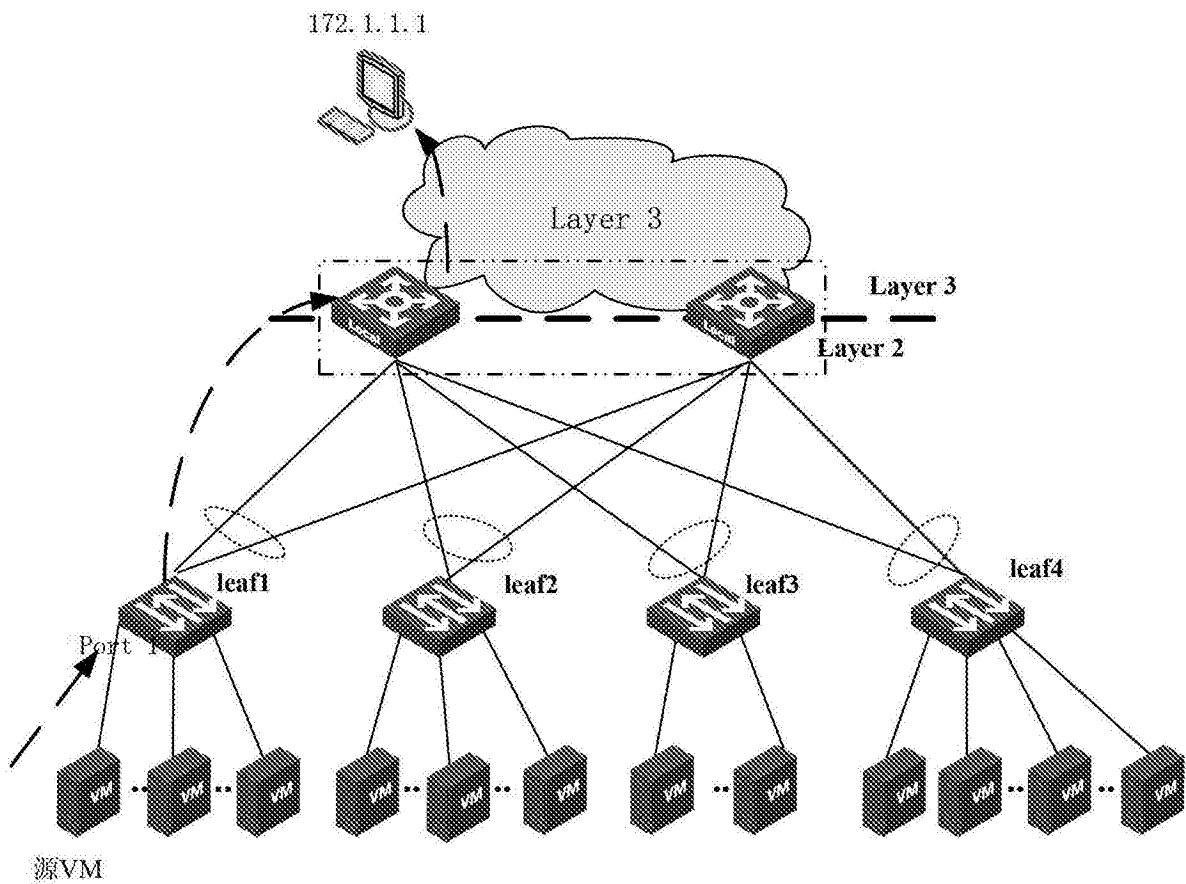


图5

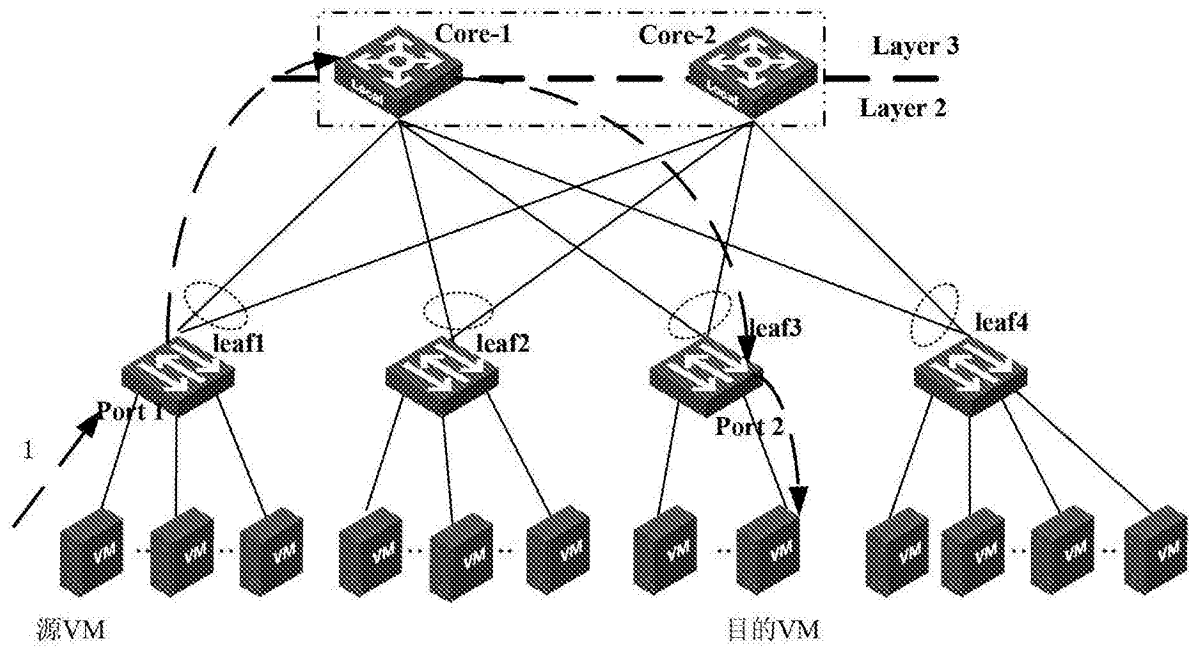


图6

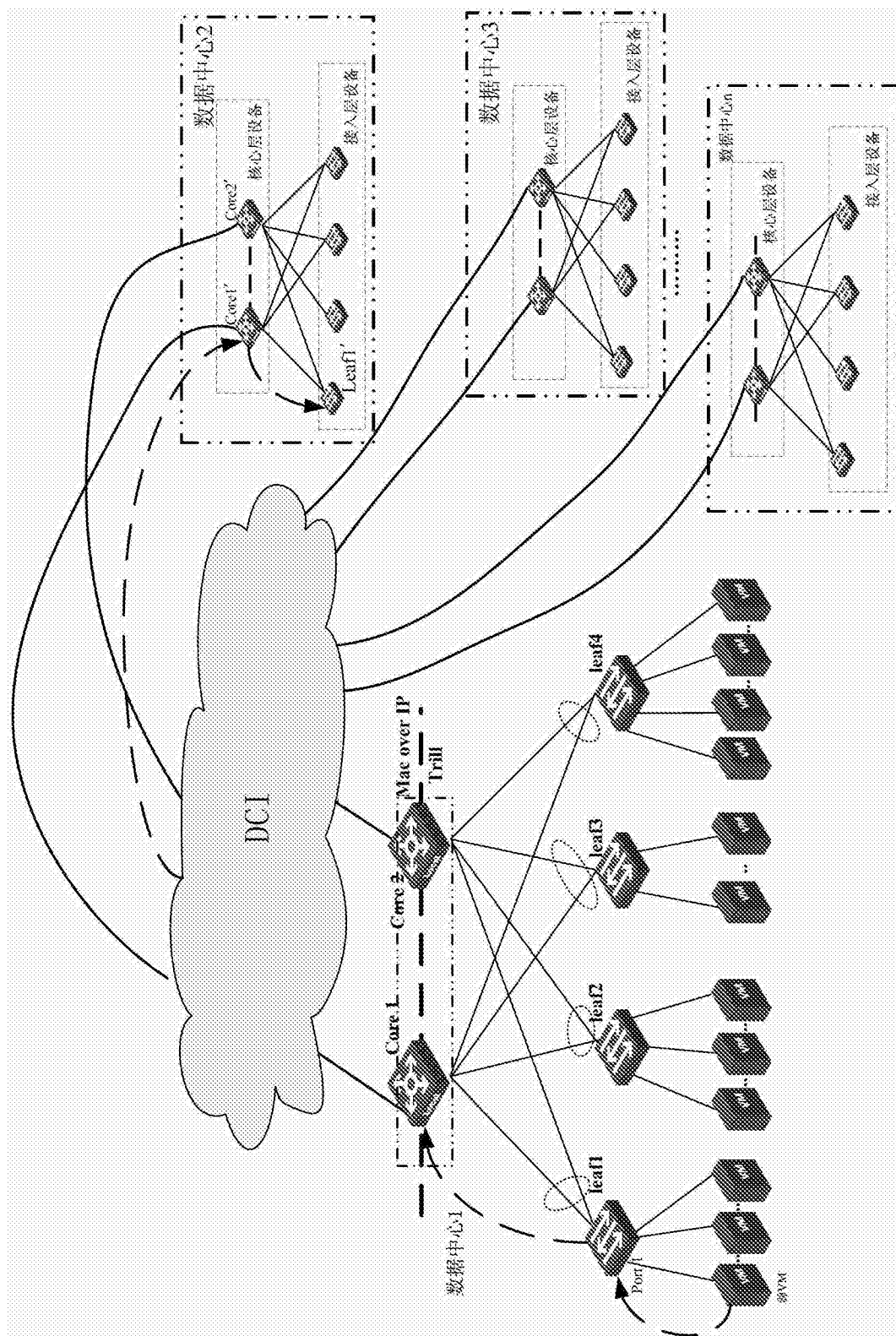


图7

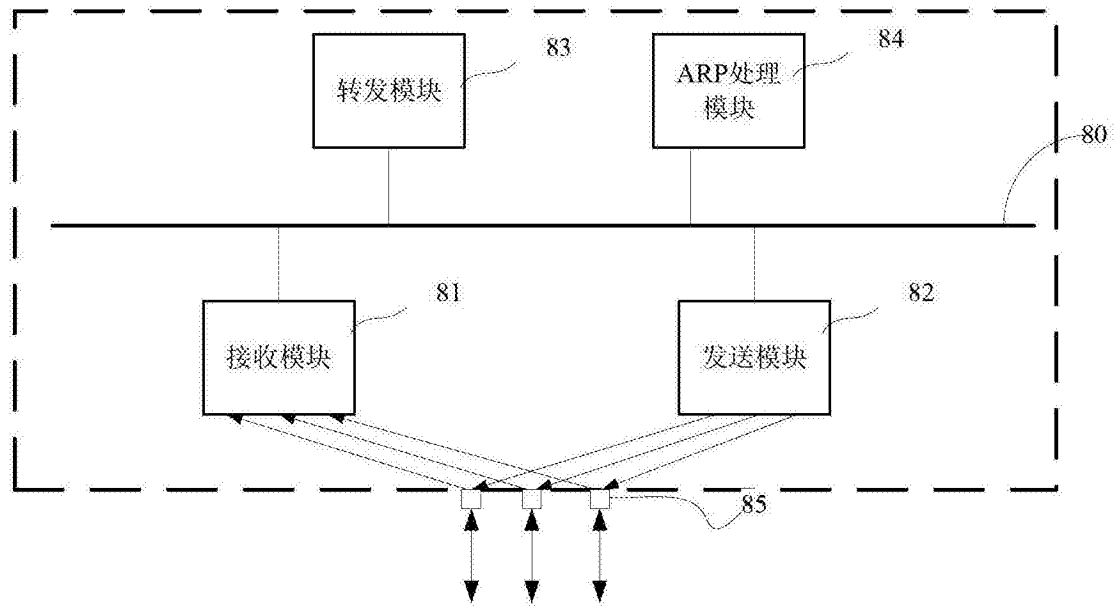


图8