

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



(10) 国际公布号
WO 2016/198022 A1

(43) 国际公布日
2016年12月15日 (15.12.2016)

- (51) 国际专利分类号:
H04L 12/761 (2013.01)
- (21) 国际申请号: PCT/CN2016/087112
- (22) 国际申请日: 2016年6月24日 (24.06.2016)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201510647010.9 2015年10月9日 (09.10.2015) CN
- (71) 申请人: 中兴通讯股份有限公司 (ZTE CORPORATION) [CN/CN]; 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦, Guangdong 518057 (CN)。
- (72) 发明人: 王翠 (WANG, Cui); 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦中兴通讯股份有限公司转交, Guangdong 518057 (CN)。张征 (ZHANG, Zheng); 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦中兴通讯股份有限公司转交, Guangdong 518057 (CN)。胡方伟 (HU, Fangwei); 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦中兴通讯股份有限公司转交, Guangdong 518057 (CN)。黄孙亮 (HUANG, Sunliang); 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦中兴通讯股份有限公司转交, Guangdong 518057 (CN)。
- (74) 代理人: 北京安信方达知识产权代理有限公司 (AFD CHINA INTELLECTUAL PROPERTY LAW OFFICE); 中国北京市海淀区学清路8号B座1601A, Beijing 100192 (CN)。

- (81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。
- (84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

根据细则 4.17 的声明:

- 关于申请人有权申请并被授予专利(细则 4.17(ii))
- 发明人资格(细则 4.17(iv))

本国际公布:

- 包括国际检索报告(条约第 21 条(3))。
- 在修改权利要求的期限届满之前进行, 在收到该修改后将重新公布(细则 48.2(h))。
- 根据申请人的请求, 在条约第 21 条(2)(a)所规定的期限届满之前进行。

(54) Title: METHOD FOR IMPLEMENTING VIRTUALIZATION NETWORK OVERLAY AND NETWORK VIRTUALIZATION EDGE NODE

(54) 发明名称: 一种实现虚拟化网络叠加的方法与网络虚拟化边缘节点



图 8
11 ACQUIRING A VNI OF A CONNECTED VIRTUAL NETWORK
12 NOTIFYING THE VNI THROUGH A ROUTING PROTOCOL

(57) Abstract: A method for implementing virtualization network overlay, comprising: acquiring a virtual network identifier of a connected virtual network; and notifying the virtual network identifier through a routing protocol. By means of the solution, the burden of a current data centre on a data plane and a control plane for a BUM traffic forwarding method can be reduced.

(57) 摘要: 一种实现虚拟化网络叠加的方法, 包括: 获取所连接的虚拟网络的虚拟网络标识; 通过路由协议通告所述虚拟网络标识。通过上述方案可以减轻当前数据中心对于 BUM 流量转发方法的数据面和控制面上的负担。



WO 2016/198022 A1

一种实现虚拟化网络叠加的方法与网络虚拟化边缘节点

技术领域

本申请涉及但不限于网络虚拟化技术领域，特别是一种实现虚拟化网络
5 叠加的方法及网络虚拟化边缘节点。

背景技术

比特位索引显示复制（Bit Index Explicit Replication，简称 BIER）技术
是近两年在 IETF（Internet Engineering Task Force，互联网工程任务组）开始
10 研究的组播技术，如图 1 所示，其基本原理是为每一个 BIER 域内的节点分
配一个唯一的 BFR-id（Bit-Forwarding Router Identifier，比特位转发路由器标
识）。一般情况下，BFR-id 通过<SI:XYZW>的格式标识，其中，SI 是 Set
Identifier（集标识），当比特位串长度（BSL，BitStringLength）不足以标识
域内所有 BIER 节点时，会引入 SI。XYZW 标识 BitPosition（比特位位置），
15 比特位位置中的每一位 bit 都对应于一个 BFR（比特位转发路由器），其长
度标识比特位串长度（BitStringLength）。例如，当 BIER 域内有 5 个节点，
BSL 为 5 时，可以将这 5 个节点放在同一个 SI 中，即 BFR-id 为 1 的 BFR-1
对应的 SI 为 0，BitPosition 是 00001，BFR-id 为 2 的 BFR-2 对应的 SI 为 0，
BitPosition 是 00010，以此类推。当 BIER 域内有 10 个节点，BSL 为 5 时，
20 需要将这 10 个节点分放在 2 个 SI 中，一个 SI 中 5 个节点。比特位串（BitString）
中的每一位 Bit 都对应于一个 BFR-id。例如，BFR-id 为 1 的 BFR-1 对应的
BitString 是 00001，BFR-id 为 2 的 BFR-2 对应的 BitString 是 00010，以此类推。
当组播报文到达 BFR-1 时，此时 BFR-1 作为 BFIR（Bit-Forwarding Ingress
Router，比特位转发入口路由器），BFR-1 通过某种方式决定哪些 BFERs
25 （Bit-Forwarding Egress Router，比特位转发出口路由器）需要这个组播流量，
例如，获取到 BFR-2 和 BFR-3 属于同一个子集 Set Identifier，且均需要该组
播流量，则将这些需要此组播流量的 BFERs 对应的 BFR-id 解析成集标识 SI，
并将 BFR-2 和 BFR-3 对应的 BitPosition 组合成 BitString 00110 封装在 BIER
报文头中，然后通过扩展 IGP（Interior Gateway Protocol，内部网关协议）生

成的比特位索引转发表 (Bit Index Forwarding Table, 简称 BIFT) 转发此封装有 BIER 头的组播数据报文。

上面提到, BIFT 是基于 IGP 协议进行扩展的, 当前支持扩展的 IGP 协议主要包括 IS-IS (Intermediate System-to-Intermediate System, 中间系统到中间系统) 协议和 OSPF (Open Shortest Path First, 开放式最短路径优先) 协议。
5 图 2 所示是 IS-IS 协议为了支持 BIER 技术的协议扩展 IS-IS LSA (Link-State Advertisement, 链路状态通告); 图 3 所示是 OSPF 协议为了支持 BIER 技术的协议扩展 OSPF-LSA。

基于 BIER 技术, 网路节点不再需要支持组播协议以及维护组播每流状态, 极大地简化了组播控制面的实现以及网络节点的性能。而且, BIER 技术有效地利用了当前的 IGP 协议, 只需要对当前 IGP 协议做个小小的扩展便能实现和提高组播的部署。进一步地, BIER 技术也可以和当前软件定义网络技术相结合, 为未来的软件定义网络的发展减轻阻碍。
10

另一方面, 虚拟化数据中心技术被越来越广泛的应用到私有云/公有云/混合云的数据中心部署中。IETF 国际标准组织提出的一种针对虚拟化数据中心的解决方案 NVO3 (Network Virtualization using Overlays over Layer 3, 基于层三的网络虚拟化叠加) 技术也逐步进行着其标准化工作。其中, 最广泛使用的就是虚拟可扩展局域网技术 (Virtual eXtensible Local Area Network, 简称 VXLAN)。
15

虚拟化技术使得每一台物理的服务器可以虚拟化为多台虚拟机 (Virtual Machines, 简称 VMs), 属于同一 VLAN (Virtual Local Area Network, 虚拟局域网) 域的虚拟机可以互通。但是由于 VLAN 只支持 4096 个, 故极大地限制了当前租户的数量。而且, 由于当前数据中心的大二层网络结构, 为了防止环路, 使用了分发树协议 (Spanning Tree Protocol), 这也导致了大量的端口或者链路被失效和浪费掉。进而, 三层 IP 技术逐渐被引入到数据中
20 25 这就意味着, 对属于不同网络的虚拟机需要互通时不仅需要跨越二层网络, 还需要跨越三层网络。各种原因的催生下, 虚拟可扩展局域网技术 (VXLAN) 应运而生。首先, VXLAN 技术使用了 24-bit 的虚拟网络标识 (Virtual Network Identifier, 简称 VNI) 来标示 VXLAN 域, 即支持 16M 的 VXLAN 用户。另

外，VXLAN 是一种叠加（overlay）技术，无论传输网络是二层还是三层，VXLAN 技术可以将原始报文打上 VXLAN 标识，然后封装在隧道（Tunnel）中转发到远端，为虚拟化后的属于同一租户的虚拟机实现互通。

数据中心中，存在两种类型的流量，一种是单播流量，另一种是 BUM 流量（Broadcast\Unknown\Multicast，广播\未知\组播），如 ARP（Address Resolution Protocol，地址解析协议）/ND（Neighbor Discovery，邻居发现协议）、DHCP（Dynamic Host Configuration Protocol，动态主机设置协议）和 mDNS（multicast DNS，组播 DNS（Domain Name System，域名系统））等。对于单播流量，当前 NVO3 的技术架构如图 4 所示，其中 Server（服务器）1/Server2/Server3 分别虚拟化成不同的虚拟机 VM1 至 VM6，隶属于不同的租户。属于同一租户的虚拟机之间构成一个虚拟网络（Virtual Network）；例如，VM1 和 VM3 属于同一个租户 A，VM2 和 VM5 属于同一个租户 B，VM4 和 VM6 属于同一个租户 C。NVE（Network Virtualization Edge，网络虚拟化边缘）是执行隧道封装/解封装的节点。NVE 之间的叠加隧道可以选择 VXLAN 隧道。图 5 所示是 VXLAN 数据面的报文头结构。图 6 所示是 NVE 封装后的隧道报文数据结构。数据面转发时，NVE 上对原始报文进行 VXLAN 报文头的封装后，根据该 VXLAN 对应的隧道目的 IP 地址进行外层隧道的封装，然后单播转发报文到远端 NVE，例如，VM1 发起的租户 A 的数据流量到达 NVE1 后，NVE1 封装上携带有 VNI（Virtual Network Identifier）为 1 的 VXLAN 头，进而根据该 VXLAN 对应的隧道目的 IP 地址 NVE2 进行外层隧道的封装，然后转发到远端 NVE2。远端 NVE2 接收到报文后，解封装外层隧道，根据 VXLAN 报文头中的 VNI 将报文转发到属于该 VXLAN 的租户网络 A 的 VM3 中。租户 B 和租户 C 的转发类似。VXLAN 隧道中，VNI 特定指的是 VXLAN 网络标识（VXLAN Network Identifier）。

对于 BUM 流量，当前 NVO3 的部署如图 7 所示，比如，VM1 和 VM3 和 VM5 属于同一租户，NVE1 收到 BUM 流量后，方法一是在 NVE1 端点上进行入口复制，将组播流量复制一份，分别封装上 VXLAN 报文头，进一步封装上该 VXLAN 对应的多个隧道目的 IP 地址（NVE2 和 NVE3），分别发向不同的远端 NVE2 和 NVE3。但是这个方法仅适用于小型网络，当租户网

络较大时，入口端点上会存在大量负担去复制组播报文，同时也大量浪费了 NVEs 之间的带宽，而且，NVE 上 VXLAN 对应隧道目的地址的映射需要额外的控制面技术协助下发。方法二是在 NVEs 之间的网络上运行组播协议 PIM (Protocol Independent Multicast, 协议无关组播) 建立组播分发树，然后

5 当 NVE1 收到 BUM 报文后，查收该 BUM 报文属于哪个 VXLAN，然后查找该 VXLAN 对应的组播组映射，然后将 BUM 报文封装上 VXLAN 报文头，然后沿着 NVEs 之间建立的属于该对应组播组的组播分发树进行转发。这个方法在一定程度上可以解决入口端点的负担以及 NVEs 之间的带宽，但是需要 NVEs 之间运行三层的 PIM 协议，并且还需要全网维护组播树，在另一种

10 程度上又增加了网络复杂性和可部署性，而且，NVE 上 VXLAN 对应组播组的映射也需要额外的控制面技术协助下发。

发明内容

以下是对本文详细描述的主题的概述。本概述并非是为了限制权利要求

15 的保护范围。

本发明实施例提供一种实现虚拟化网络叠加的方法及 NVE 节点，以减轻当前数据中心对于 BUM 流量转发方法的数据面和控制面上的各种弊端。

本发明实施例提供了一种实现虚拟化网络叠加的方法，应用于虚拟化数据中心的网络虚拟化边缘节点，包括：

20 获取所连接的虚拟网络的虚拟网络标识；

通过路由协议通告所述虚拟网络标识。

可选地，上述方法还具有下面特点：所述通告所述虚拟网络标识包括：

通告有效的虚拟网络标识；和/或，

通告撤销的虚拟网络标识。

25 可选地，上述方法还具有下面特点：

所述路由协议包括以下的任一种：中间系统到中间系统协议、开放式最短路径优先协议和边界网关协议。

可选地，上述方法还具有下面特点：

所述路由协议支持 IPv4 网络协议和 IPv6 网络协议。

可选地，上述方法还具有下面特点：

所述虚拟网络标识包括虚拟可扩展局域网的网络标识。

5 本发明实施例还提供了一种网络虚拟化边缘节点，其中，包括：

获取模块，设置为获取所连接的虚拟网络的虚拟网络标识；

通告模块，设置为通过路由协议通告所述虚拟网络标识。

可选地，上述网络虚拟化边缘节点还具有下面特点：

10 所述通告模块设置为：通告有效的虚拟网络标识；和/或，通告撤销的虚拟网络标识，其中，所述路由协议包括以下的任一种：中间系统到中间系统协议、开放式最短路径优先协议、边界网关协议，所述路由协议支持 IPv4 网络协议和 IPv6 网络协议，所述虚拟网络标识包括虚拟可扩展局域网的网络标识。

15 本发明实施例还提供了一种实现虚拟化网络叠加的方法，应用于虚拟化数据中心的网络虚拟化边缘节点，包括：

接收携带有虚拟网络标识的通告报文；

解析所述虚拟网络标识，根据所述虚拟网络标识建立或更新对应的虚拟网络标识与发送相同虚拟网络标识的节点的比特位串的映射关系。

可选地，上述方法还具有下面特点：

20 所述通告报文包括：携带有有效的虚拟网络标识和/或携带有撤销的虚拟网络标识的通告报文。

可选地，上述方法还包括：

25 接收到租户的广播\未知\组播 BUM 流量时，查找所述租户隶属的虚拟网络标识，封装上相应的虚拟网络报文头，查找对应该虚拟网络标识的比特位串，封装上所述比特位串对应的比特位索引显示复制（BIER）头，按照比特位索引转发表进行转发。

本发明实施例还提供了了一种网络虚拟化边缘节点，其中，包括：

接收模块，设置为接收携带有虚拟网络标识的通告报文；

5 处理模块，设置为解析所述虚拟网络标识，根据所述虚拟网络标识建立或更新对应的虚拟网络标识与发送相同虚拟网络标识的节点的比特位串的映射关系。

可选地，上述网络虚拟化边缘节点还具有下面特点：

所述接收模块接收到的所述通告报文包括：携带有有效的虚拟网络标识的通告报文和/或携带有撤销的虚拟网络标识的通告报文。

可选地，上述网络虚拟化边缘节点还具有下面特点：

10 所述接收模块，还设置为接收到租户的广播\未知\组播 BUM 流量；

所述处理模块，还设置为查找所述租户隶属的虚拟网络标识，封装上相应的虚拟网络报文头，查找对应该虚拟网络标识的比特位串，封装上所述比特位串对应的比特位索引显示复制头，按照比特位索引转发表进行转发。

15 本发明实施例还提供一种计算机可读存储介质，存储有计算机可执行指令，所述计算机可执行指令被执行时实现发送侧的上述实现虚拟化网络叠加的方法。

本发明实施例还提供一种计算机可读存储介质，存储有计算机可执行指令，所述计算机可执行指令被执行时实现接收侧的上述实现虚拟化网络叠加的方法。

20 综上，本发明实施例提供一种实现虚拟化网络叠加的方法及 NVE 节点，以减轻当前数据中心对于 BUM 流量转发方法的数据面和控制面上的负担。

在阅读并理解了附图和详细描述后，可以明白其他方面。

附图概述

25 附图用来提供对本申请的进一步理解，并且构成说明书的一部分，与本申请的实施例一起用于解释本申请，并不构成对本申请的限制。在附图中：

图 1 是相关的 BIER 技术架构的示意图；

- 图 2 是相关技术的 IS-IS 协议扩展实现 BIER 控制面的示意图；
- 图 3 是相关技术的 OSPF 协议扩展实现 BIER 控制面的示意图；
- 图 4 是相关技术的 NVO3 技术架构（单播场景）的示意图；
- 图 5 是相关技术的 VXLAN 报文头结构的示意图；
- 5 图 6 是相关技术的 NVE 封装后隧道上转发的报文结构的示意图；
- 图 7 是相关技术的 NVO3 技术架构（BUM 场景）的示意图；
- 图 8 为本发明实施例的发送侧的实现虚拟化网络叠加的方法的流程图；
- 图 9 为本发明实施例的发送侧的 NVE 节点的示意图；
- 图 10 为本发明实施例的接收侧的实现虚拟化网络叠加的方法的流程图；
- 10 图 11 为本发明实施例的接收侧的 NVE 节点的示意图；
- 图 12 为本发明实施例的基于 IS-IS 协议扩展携带虚拟网络标识的示意图；
- 图 13 为本发明实施例的应用场景的示意图。

本发明的实施方式

- 15 下文中将结合附图对本发明实施例进行详细说明。需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互任意组合。

20 鉴于相关技术存在的问题，如果能将 BIER 技术引入到虚拟化网络叠加 NVO3 中，用于实现数据中心 BUM 流量的数据面的转发技术，同时，在控制面，引入 IGP-BIER 和 BGP-BIER 的扩展，用于源端 NVE 发现远端属于同一 VXLAN 的 NVEs 的控制面技术。这样，将极大地减轻当前数据中心对于 BUM 流量转发方法的数据面和控制面上的各种弊端，进一步加快虚拟化数据中心的部署和 BIER 的部署。本发明实施例试图在上述虚拟化数据中心网络中，引入 BIER 技术，实现虚拟化网络中 BUM 流量转发的最优实现。

- 25 图 8 为本发明实施例的发送侧的实现虚拟化网络叠加的方法的流程图，如图 8 所示，本实施例的方法应用于虚拟化数据中心的 NVE 节点，包括以下步骤：

步骤 11、获取所连接的虚拟网络的虚拟网络标识 (VNI, Virtual Network Identifier) ;

步骤 12、通过路由协议通告所述 VNI。

其中, 通知所述 VNI 包括: 通告有效的 VNI; 和/或, 通告撤销的 VNI。

5 其中, 所述路由协议包括以下的任一种: IS-IS 协议、OSPF 协议和 BGP (Border Gateway Protocol, 边界网关协议); 所述路由协议支持互联网协议第四版 (IPv4) 网络协议和互联网协议第六版 (IPv6) 网络协议。

其中, 所述 VNI 包括 VXLAN 的网络标识。

10 图 9 为本发明实施例的发送侧的 NVE 节点的示意图, 如图 9 所示, 本实施例的 NVE 节点可以包括:

获取模块, 设置为获取所连接的虚拟网络的 VNI;

通告模块, 设置为通过路由协议通告所述 VNI。

在一可选实施例中, 所述通告模块设置为: 通告有效的 VNI; 和/或, 通告撤销的 VNI。

15 其中, 所述路由协议包括以下的任一种: IS-IS 协议、OSPF 协议和 BGP 协议; 所述路由协议支持 IPv4 网络协议和 IPv6 网络协议; 所述虚拟网络标识包括 VXLAN 的网络标识。

20 图 10 为本发明实施例的接收侧的实现虚拟化网络叠加的方法的流程图, 如图 10 所示, 本实施例的方法应用于虚拟化数据中心的 NVE 节点, 包括以下步骤:

步骤 21、接收携带有 VNI 的通告报文;

步骤 22、解析所述 VNI, 根据所述 VNI 建立或更新对应的 VNI 与发送相同 VNI 的节点的比特位串的映射关系。

25 其中, 所述通告报文包括: 携带有有效的虚拟网络标识和/或携带有撤销的虚拟网络标识的通告报文。

本实施例的方法还可以包括:

接收到租户的 BUM 流量时，查找所述租户隶属的 VNI，封装上相应的虚拟网络报文头，查找所述 VNI 对应的比特位串，封装上所述比特位串对应的比特位索引显示复制（BIER）头，按照比特位索引转发表（BIFT）进行转发。

5 图 11 为本发明实施例的接收侧的 NVE 节点的示意图，如图 11 所示，本实施例的 NVE 节点可以包括：

接收模块，设置为接收携带有 VNI 的通告报文；

处理模块，设置为解析所述 VNI，根据所述 VNI 建立或更新对应的 VNI 与发送相同 VNI 的节点的比特位串的映射关系。

10 可选地，所述接收模块接收到的所述通告报文包括：携带有有效的虚拟网络标识的通告报文和/或携带有撤销的虚拟网络标识的通告报文。

在一可选实施例中，所述接收模块，还可以设置为接收到租户的 BUM 流量；

15 所述处理模块，还可以设置为查找所述租户隶属的 VNI，封装上相应的虚拟网络报文头，查找所述 VNI 对应的比特位串，封装上所述比特位串对应的比特位索引显示复制头，按照比特位索引转发表进行转发。

下面结合实施例阐述本申请。

实施例一

20 当前，IS-IS 协议扩展实现 BIER 控制面时，对于 IPv4（互联网协议的第四版）网络，在 IS-IS 协议的 Extended IP reachability TLV（扩展 IP 可达性 TLV（Type\Length\Value，类型\长度\值））（TLV 类型为 135）和 Multi-Topology Reachable IPv4 Prefixes TLV（多拓扑可达 IPv4 前缀 TLV）（TLV 类型为 235）下进行了扩展；以及对于 IPv6（互联网协议的第六版）网络，在 IS-IS 的 IPv6 Reachability TLV（TLV 类型为 236）和 Multi-Topology Reachable IPv6 Prefixes
25 TLV（TLV 类型为 237）下进行了扩展，具体扩展格式见图 2。

本发明实施例试图将 BIER 技术应用在虚拟化数据中心的控制面，于是在上述图 2 定义的 IS-IS 扩展中进一步定义了一个新的 sub-sub-TLV（子子

TLV)，用于通告虚拟网络的虚拟网络标识。基于 IS-IS 协议扩展携带虚拟网络标识的报文参考格式如图 12 所示。

其中，Type 标识该 sub-sub-TLV 的类型，本发明实施例用于标识虚拟网络 sub-sub-TLV；Length 标识该 sub-sub-TLV 中 Value 部分的长度；Virtual Network Identifier 是虚拟网络标识，24-bit，唯一标识虚拟网络。

实施例二

当前，OSPF 协议扩展实现 BIER 控制面，对 IPv4 网络，在 OSPFv2 协议的 Extended Prefix TLV（扩展前缀 TLV）下进行了扩展；以及对于 IPv6 网络，在 OSPFv3 的 Extended LSA TLV（扩展链路状态通告 TLV）下进行了扩展，具体扩展格式见图 3。

本发明实施例试图将 BIER 技术应用在虚拟化数据中心的控制面，于是在上述图 3 定义的 OSPF 和 OSPFv3 扩展中进一步定义了一个新的 sub-sub-TLV，用于通告虚拟网络的虚拟网络标识。基于 OSPF 和 OSPFv3 协议扩展携带虚拟网络标识的报文参考格式也如图 12 所示。

其中，Type 标识该 sub-sub-TLV 的类型，本发明实施例用于标识虚拟网络 sub-sub-TLV；Length 标识该 sub-sub-TLV 中 Value 部分的长度；Virtual Network Identifier 是虚拟网络标识，24-bit，唯一标识虚拟网络。

实施例三

本发明实施例试图将 BIER 技术应用在虚拟化数据中心的控制面，于是参考在 BGP 协议的 BGP BIER 属性下进一步扩展了一个新的 sub-TLV，用于通告虚拟网络标识；或者在 BGP 协议的网络层可达信息（NLRI，Network Layer Reachable Information）下进行扩展，用于通告虚拟网络标识。

实施例四

OSPF 协议可以支持在 BIER 域的 BFIR 和 BFER 设备上建立 OSPF 虚链，通过虚链，将本发明实施例所提到的扩展 TLV 信息直接发送到 BIER 域的边缘设备，BFIR 和 BFER 设备直接互相交互所连接的 VNI 信息，减少 BIER 域中间节点的信息存储。所通告的格式仍然基于 OSPF 协议扩展携带虚拟网络标识的报文参考格式也如图 12 所示。

其中，Type 标识该 sub-sub-TLV 的类型，本发明实施例用于标识虚拟网络 sub-sub-TLV；Length 标识该 sub-sub-TLV 中 Value 部分的长度；Virtual Network Identifier 是虚拟网络标识，24-bit，唯一标识虚拟网络。

实施例五

5 如图 13 所示，NVEs (NVE1/NVE2/NVE3) 之间运行 IGP 或 BGP 协议，隶属于租户 A (VXLAN 标识 10) 的 VM1 连接至 NVE1，NVE1 的 BFRID 为 1，对应的 BitString 为 001；同时，隶属于租户 A (VXLAN 标识 10) 的 VM3 连接至 NVE2，NVE2 的 BFRID 为 2，对应的 BitString 为 010；隶属于租户 A (VXLAN 标识 10) 的 VM5 连接至 NVE3，NVE3 的 BFRID 为 3，
10 对应的 BitString 为 100。NVE1/NVE2/NVE3 通过 IGP 协议扩展或者 BGP 协议扩展携带 BIER 信息和 VXLAN 信息。

当 NVEs 之间直接连接时，可以直接通过实施例一或者实施例二中扩展的 IGP 格式通告 VXLAN 信息。

15 例如，NVE2 通告 BIER 信息和 VXLAN 信息，NVE1 收到后，本地建立 VXLAN 信息和 BitString 的映射关系[VXLAN 10: 010]，同样，NVE3 通告 BIER 信息和 VXLAN 信息，NVE1 也收到后，更新本地映射为[VXLAN 10: 110]。当 NVE1 接收到来自于 VM1 的租户组播流量时，会查找该组播流量属于 VXLAN 10，封装上 VXLAN 报文头，进一步查找，属于该 VXLAN 的远端 NVEs 对应的 BitString 为 110，于是进一步封装上 BIER 报文头，转发组
20 播报文。

当 NVEs 之间非直接连接，而是需要经过多个节点才能达到互通时，有以下两种方法可以实现：

25 方法一：NVEs 通过实施例一和实施例二中扩展的 IGP 格式通告 VXLAN 信息；中间节点接收到 VXLAN 信息发现不识别，则按照 IGP 规则转发该 IGP 通告消息即可。

方法二：NVEs 通过实施例三或者实施例四的实现，在 NVEs 之间建立 BGP 邻居或者 OSPF 虚链，直接在 NVEs 之间通告 VXLAN 信息，中间节点无需处理。

同样的，NVE2 通告 BIER 信息和 VXLAN 信息，无论通过方法一还是方法二，NVE1 收到后，本地建立 VXLAN 信息和 BitString 的映射关系 [VXLAN 10: 010]，同样，NVE3 通告 BIER 信息和 VXLAN 信息，无论通过方法一还是方法二，NVE1 也收到后，更新本地映射为 [VXLAN 10: 110]。当 NVE1 接收到来自于 VM1 的租户组播流量时，会查找该组播流量属于 VXLAN 10，封装上 VXLAN 报文头，进一步查找，属于该 VXLAN 的远端 NVEs 对应的 BitString 为 110，于是进一步封装上 BIER 报文头，转发组播报文。

实施例六

10 本实施例基于虚拟机迁移导致转发面更新，仍然如图 13 所示，当虚拟机 VM5 发生迁移，从隶属于的 VNI A 迁移到 VNI B 时，连接 VM5 的节点 NVE3 发现隶属于 VXLAN 10 的用户迁移了，于是通过 IGP 协议或者 BGP 协议通告撤销 VXLAN 信息。

当 NVEs 之间直接连接时，可以直接通过实施例一和实施例二中扩展的 IGP 格式通告撤销 VXLAN 信息；例如，NVE3 通告撤销 VXLAN 信息，NVE1 收到后，本地更新原来保存的 VXLAN 信息和 BitString 的映射关系，从 [VXLAN 10: 110]更新到[VXLAN 10: 010]。这样，当 NVE1 接收到来自于 VM1 的后续租户组播流量时，仍然会先查找该组播流量属于 VXLAN 10，封装上 VXLAN 报文头，进一步查找，发现属于该 VXLAN 的远端 NVEs 对应的 BitString 更新为 010，于是进一步封装上更新后的 BIER 报文头，转发组播报文。

当 NVEs 之间非直接连接，而是需要经过多个节点才能达到互通时，有两种方法可以实现本发明：

方法一：NVE3 通过实施例一和实施例二中扩展的 IGP 格式通告撤销 VXLAN 信息；中间节点接收到撤销 VXLAN 信息发现不识别，则按照 IGP 规则转发该 IGP 通告消息即可。

方法二：NVE3 通过实施例三或者实施例四中的实现，在 NVE1 和 NVE3 之间建立 BGP 邻居或者 OSPF 虚链，直接在 NVE1 和 NVE3 之间通告撤销

VXLAN 信息。中间节点无需处理。

5 这样，无论通过方法一还是方法二，NVE1 收到 VXLAN 撤销消息后，更新本地映射为[VXLAN 10: 010]。当 NVE1 接收到来自于 VM1 的后续租户组播流量时，仍然会查找该组播流量属于 VXLAN 10，封装上 VXLAN 报文头，进一步查找，发现属于该 VXLAN 的远端 NVEs 对应的 BitString 更新为 010，于是进一步封装上更新后的 BIER 报文头，转发组播报文。

本发明实施例还提供一种计算机可读存储介质，存储有计算机可执行指令，所述计算机可执行指令被执行时实现发送侧的上述实现虚拟化网络叠加的方法。

10 本发明实施例还提供一种计算机可读存储介质，存储有计算机可执行指令，所述计算机可执行指令被执行时实现接收侧的上述实现虚拟化网络叠加的方法。

本领域普通技术人员可以理解上述方法中的全部或部分步骤可通过程序来指令相关硬件（例如处理器）完成，所述程序可以存储于计算机可读存储介质中，如只读存储器、磁盘或光盘等。可选地，上述实施例的全部或部分步骤也可以使用一个或多个集成电路来实现。相应地，上述实施例中的各模块/单元可以采用硬件的形式实现，例如通过集成电路来实现其相应功能，也可以采用软件功能模块的形式实现，例如通过处理器执行存储于存储器中的程序/指令来实现其相应功能。本申请不限制于任何特定形式的硬件和软件的结合。

20 以上仅为本申请的可选实施例，本申请还可有其他多种实施例，在不背离本申请精神及其实质的情况下，熟悉本领域的技术人员当可根据本申请作出各种相应的改变和变形，但这些相应的改变和变形都应属于本申请所附的权利要求的保护范围。

25

工业实用性

本申请实施例提供一种实现虚拟化网络叠加的方法及 NVE 节点，能够减轻当前数据中心对于 BUM 流量转发方法的数据面和控制面上的负担。

权 利 要 求 书

1、一种实现虚拟化网络叠加的方法，应用于虚拟化数据中心的网络虚拟化边缘节点，包括：

获取所连接的虚拟网络的虚拟网络标识；

5 通过路由协议通告所述虚拟网络标识。

2、如权利要求 1 所述的方法，其中，所述通告所述虚拟网络标识包括：

通告有效的虚拟网络标识；和/或，

通告撤销的虚拟网络标识。

3、如权利要求 1 所述的方法，其中，所述路由协议包括以下的任一种：

10 中间系统到中间系统协议、开放式最短路径优先协议和边界网关协议。

4、如权利要求 1 所述的方法，其中，所述路由协议支持互联网协议第四版 IPv4 网络协议和互联网协议第六版 IPv6 网络协议。

5、如权利要求 1 至 4 任一项所述的方法，其中，所述虚拟网络标识包括虚拟可扩展局域网的网络标识。

15 6、一种网络虚拟化边缘节点，包括：

获取模块，设置为获取所连接的虚拟网络的虚拟网络标识；

通告模块，设置为通过路由协议通告所述虚拟网络标识。

7、如权利要求 6 所述的网络虚拟化边缘节点，其中，所述通告模块设置为：通告有效的虚拟网络标识；和/或，通告撤销的虚拟网络标识；所述路由
20 协议包括以下的任一种：中间系统到中间系统协议、开放式最短路径优先协议、边界网关协议；所述路由协议支持互联网协议第四版 IPv4 网络协议和互联网协议第六版 IPv6 网络协议；所述虚拟网络标识包括虚拟可扩展局域网的网络标识。

8、一种实现虚拟化网络叠加的方法，应用于虚拟化数据中心的网络虚拟化边缘节点，包括：

接收携带有虚拟网络标识的通告报文；

解析所述虚拟网络标识，根据所述虚拟网络标识建立或更新对应的虚拟网络标识与发送相同虚拟网络标识的节点的比特位串的映射关系。

9、如权利要求 8 所述的方法，其中，所述通告报文包括：携带有有效的虚拟网络标识和/或携带有撤销的虚拟网络标识的通告报文。

5 10、如权利要求 8 或 9 所述的方法，所述方法还包括：

接收到租户的广播\未知\组播 BUM 流量时，查找所述租户隶属的虚拟网络标识，封装上相应的虚拟网络报文头，查找对应该虚拟网络标识的比特位串，封装上所述比特位串对应的比特位索引显示复制 BIER 头，按照比特位索引转发表进行转发。

10 11、一种网络虚拟化边缘节点，包括：

接收模块，设置为接收携带有虚拟网络标识的通告报文；

处理模块，设置为解析所述虚拟网络标识，根据所述虚拟网络标识建立或更新对应的虚拟网络标识与发送相同虚拟网络标识的节点的比特位串的映射关系。

15 12、如权利要求 11 所述的网络虚拟化边缘节点，其中，所述接收模块接收到的所述通告报文包括：携带有有效的虚拟网络标识的通告报文和/或携带有撤销的虚拟网络标识的通告报文。

13、如权利要求 11 或 12 所述的网络虚拟化边缘节点，其中，

所述接收模块，还设置为接收到租户的广播\未知\组播 BUM 流量；

20 所述处理模块，还设置为查找所述租户隶属的虚拟网络标识，封装上相应的虚拟网络报文头，查找对应该虚拟网络标识的比特位串，封装上所述比特位串对应的比特位索引显示复制头，按照比特位索引转发表进行转发。

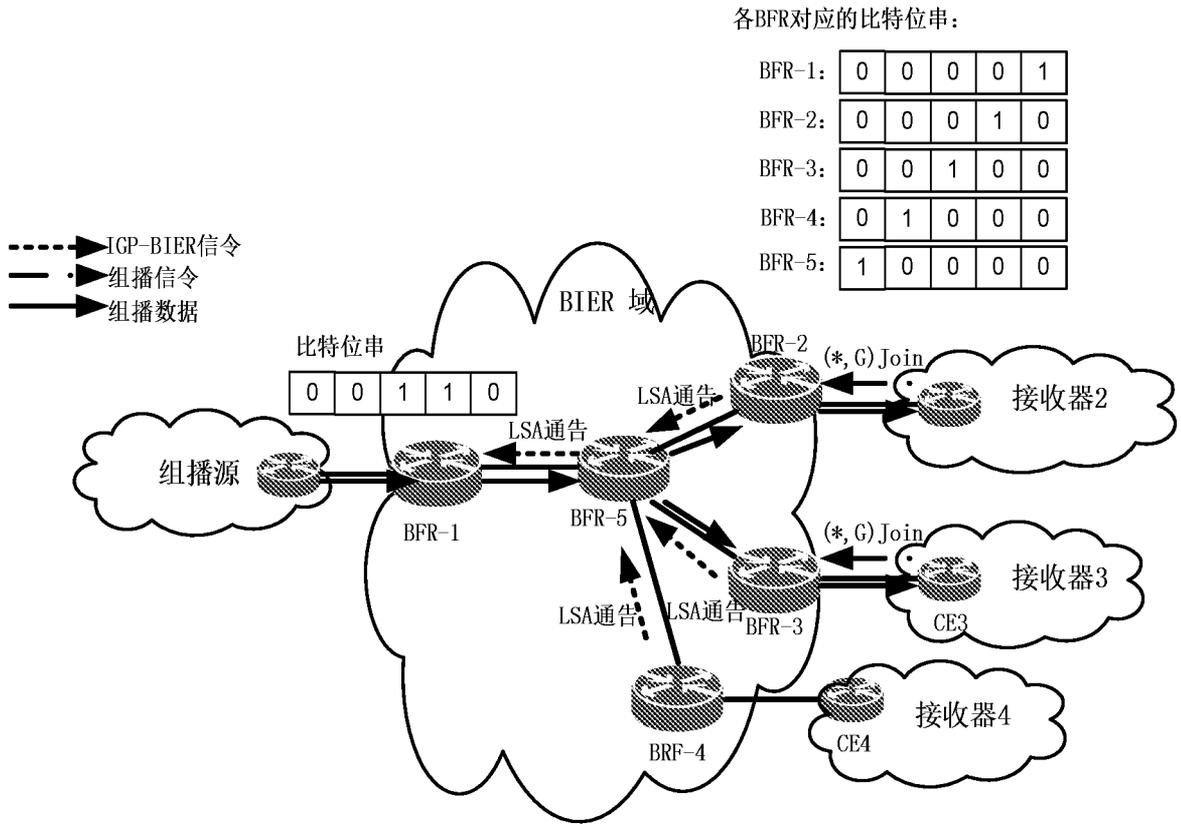


图 1

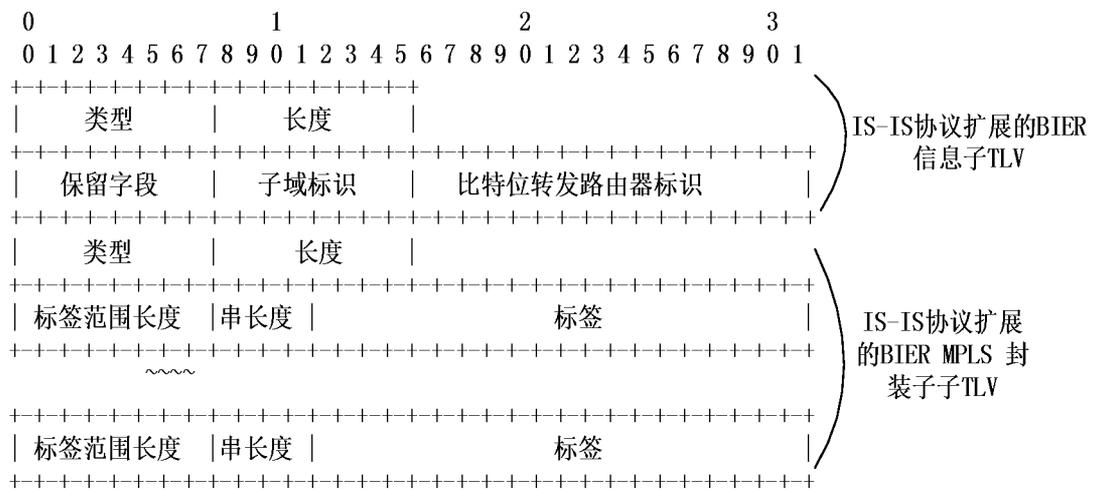


图 2

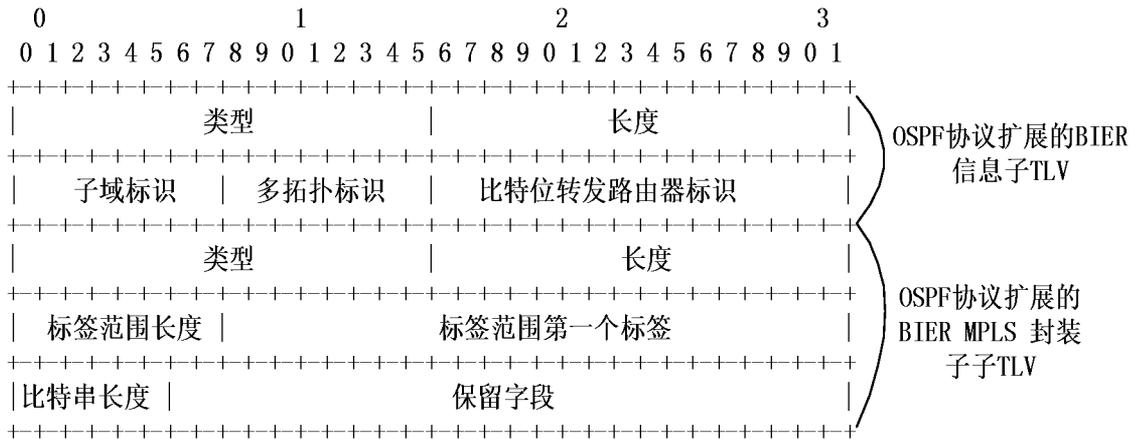


图 3

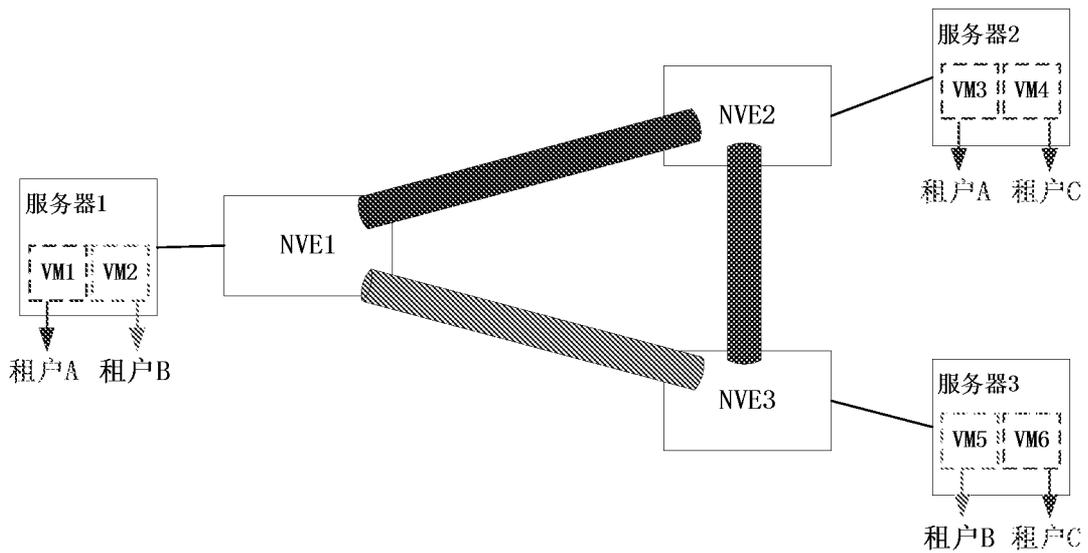


图 4

VXLAN头:

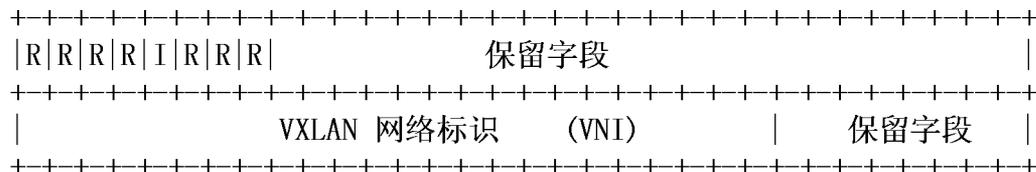


图 5

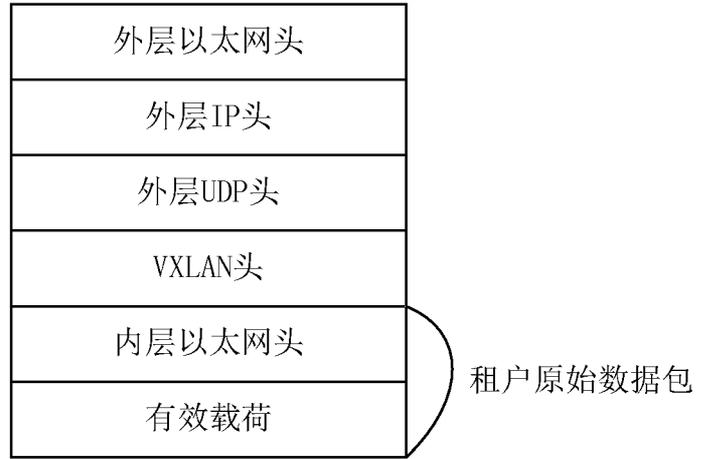


图 6

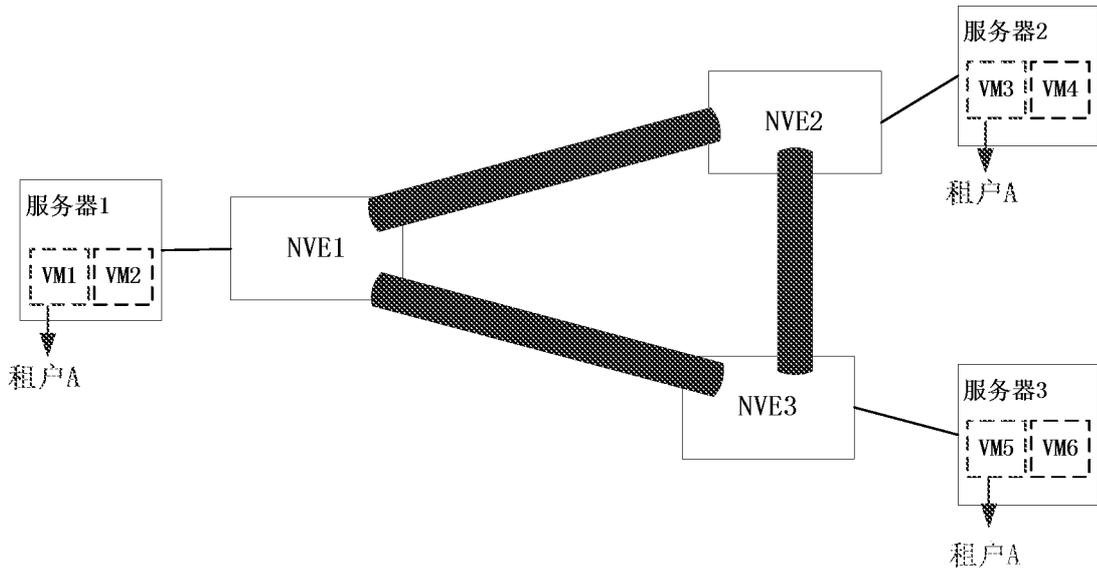


图 7

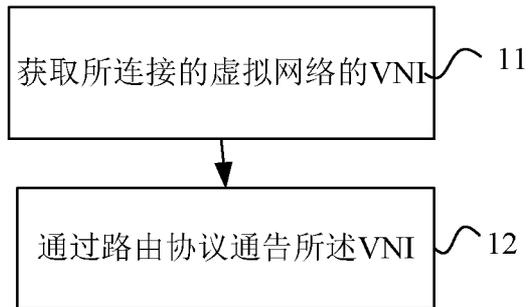


图 8

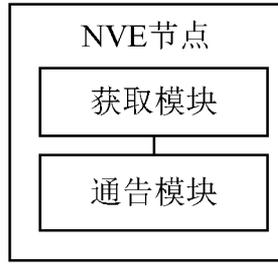


图 9

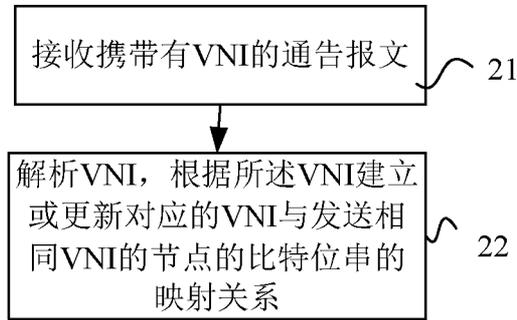


图 10

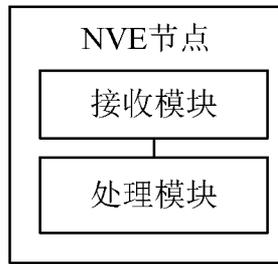


图 11



虚拟网络子TLV

图 12

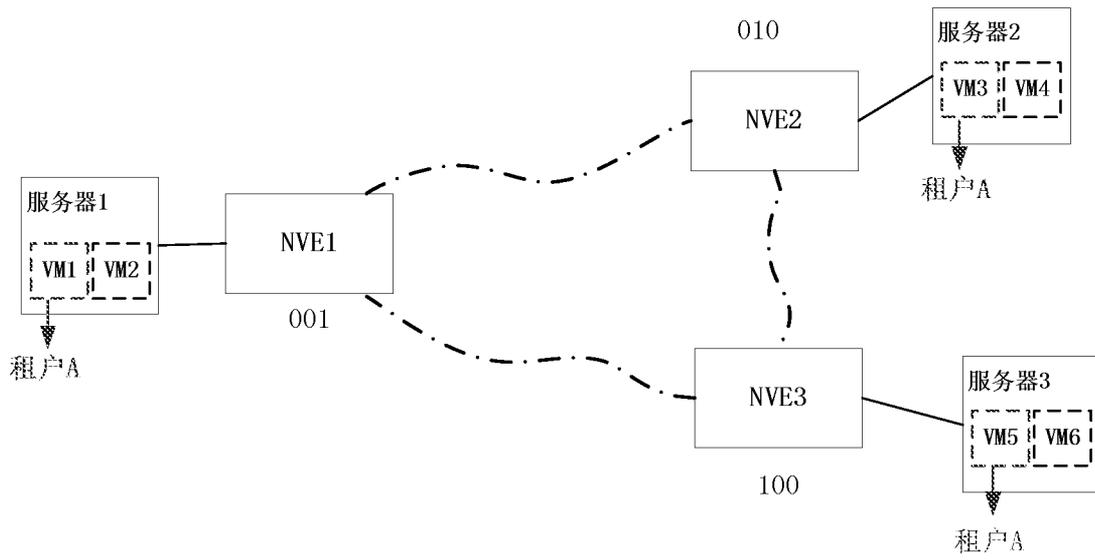


图 13

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2016/087112

A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/761 (2013.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

H04L; H04W; H04Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI: network overlay, edge node, identification, BIER, VXLAN, NVE, VM, edge, network, virtualization, ID, notify, routing, address, extension, expansion, protocol, map, bit, index, forward

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 104871495 A (HUAWEI TECHNOLOGIES CO., LTD.), 26 August 2015 (26.08.2015), description, paragraphs [0033]-[0043], [0051]-[0058] and [0068]-[0072]	
X	KAWASHIMA, R. et al., "Performance Evaluation of Non-Tunneling Edge-Overlay Model on 40GbE Environment", 2014 IEEE 3RD SYMPOSIUM ON NETWORK CLOUD COMPUTING AND APPLICATIONS, 31 December 2014 (31.12.2014), page 69, paragraph 3 to page 71, paragraph 1	1-13
X	KAWASHIMA, R. et al., "Non-Tunneling Edge-Overlay Model using OpenFlow for Cloud Datacenter Networks", 2013 IEEE INTERNATIONAL CONFERENCE ON CLOUD COMPUTING TECHNOLOGY AND SCIENCE, 31 December 2013 (31.12.2013), page 177, paragraph 3 to page 179, last paragraph	1-13
A	CN 103581277 A (ZTE CORP.), 12 February 2014 (12.02.2014), the whole document	1-13
A	CN 104348724 A (HUAWEI TECHNOLOGIES CO., LTD.), 11 February 2015 (11.02.2015), the whole document	1-13

Further documents are listed in the continuation of Box C.

See patent family annex.

<p>* Special categories of cited documents:</p> <p>"A" document defining the general state of the art which is not considered to be of particular relevance</p> <p>"E" earlier application or patent but published on or after the international filing date</p> <p>"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)</p> <p>"O" document referring to an oral disclosure, use, exhibition or other means</p> <p>"P" document published prior to the international filing date but later than the priority date claimed</p>	<p>"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention</p> <p>"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone</p> <p>"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art</p> <p>"&" document member of the same patent family</p>
---	---

Date of the actual completion of the international search
03 August 2016 (03.08.2016)

Date of mailing of the international search report
24 August 2016 (24.08.2016)

Name and mailing address of the ISA/CN:
State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao
Haidian District, Beijing 100088, China
Facsimile No.: (86-10) 62019451

Authorized officer
YANG, Yingxiao
Telephone No.: (86-10) **62413196**

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2016/087112

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2015089583 A1 (WANSER, K. et al.), 26 March 2015 (26.03.2015), the whole document	1-13

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2016/087112

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104871495 A	26 August 2015	WO 2014052485 A1	03 April 2014
		EP 2891277 A1	08 July 2015
		US 2014086253 A1	27 March 2014
CN 103581277 A	12 February 2014	WO 2014023255 A1	13 February 2014
CN 104348724 A	11 February 2015	WO 2015014187 A1	05 February 2015
US 2015089583 A1	26 March 2015	US 2014123212 A1	01 May 2014
		EP 2915090 A1	09 September 2015
		WO 2014070773 A1	08 May 2014
		US 2014123211 A1	01 May 2014

<p>A. 主题的分类</p> <p>H04L 12/761(2013.01)i</p> <p>按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类</p>																				
<p>B. 检索领域</p> <p>检索的最低限度文献(标明分类系统和分类号)</p> <p>H04L; H04W; H04Q</p> <p>包含在检索领域中的除最低限度文献以外的检索文献</p> <p>在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))</p> <p>CNPAT, WPI, EPODOC, CNKI: 网络叠加, 虚拟化, 边缘节点, 网络, 标识, 路由, 通告, 地址, 扩展, 协议, 映射, 比特, 索引, 转发, BIER, VXLAN, NVE, VM, edge, network, virtualization, ID, notify, routing, address, extension, expansion, protocol, map, bit, index, forward</p>																				
<p>C. 相关文件</p> <table border="1"> <thead> <tr> <th>类型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 104871495 A (华为技术有限公司) 2015年 8月 26日 (2015 - 08 - 26) 说明书第[0033]-[0043], [0051]-[0058], [0068]-[0072]段</td> <td>1-13</td> </tr> <tr> <td>X</td> <td>RYOTA Kawashima 等. "Performance Evaluation of Non-Tunneling Edge-Overlay Model on 40GbE Environment" 2014 IEEE 3rd Symposium on Network Cloud Computing and Applications, 2014年 12月 31日 (2014 - 12 - 31), 第69页第3段-71页第1段</td> <td>1-13</td> </tr> <tr> <td>X</td> <td>RYOTA Kawashima 等. "Non-Tunneling Edge-Overlay Model using OpenFlow for Cloud Datacenter Networks" 2013 IEEE International Conference on Cloud Computing Technology and Science, 2013年 12月 31日 (2013 - 12 - 31), 第177页第3段-179页最后1段</td> <td>1-13</td> </tr> <tr> <td>A</td> <td>CN 103581277 A (中兴通讯股份有限公司) 2014年 2月 12日 (2014 - 02 - 12) 全文</td> <td>1-13</td> </tr> <tr> <td>A</td> <td>CN 104348724 A (华为技术有限公司) 2015年 2月 11日 (2015 - 02 - 11) 全文</td> <td>1-13</td> </tr> </tbody> </table>			类型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 104871495 A (华为技术有限公司) 2015年 8月 26日 (2015 - 08 - 26) 说明书第[0033]-[0043], [0051]-[0058], [0068]-[0072]段	1-13	X	RYOTA Kawashima 等. "Performance Evaluation of Non-Tunneling Edge-Overlay Model on 40GbE Environment" 2014 IEEE 3rd Symposium on Network Cloud Computing and Applications, 2014年 12月 31日 (2014 - 12 - 31), 第69页第3段-71页第1段	1-13	X	RYOTA Kawashima 等. "Non-Tunneling Edge-Overlay Model using OpenFlow for Cloud Datacenter Networks" 2013 IEEE International Conference on Cloud Computing Technology and Science, 2013年 12月 31日 (2013 - 12 - 31), 第177页第3段-179页最后1段	1-13	A	CN 103581277 A (中兴通讯股份有限公司) 2014年 2月 12日 (2014 - 02 - 12) 全文	1-13	A	CN 104348724 A (华为技术有限公司) 2015年 2月 11日 (2015 - 02 - 11) 全文	1-13
类型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
X	CN 104871495 A (华为技术有限公司) 2015年 8月 26日 (2015 - 08 - 26) 说明书第[0033]-[0043], [0051]-[0058], [0068]-[0072]段	1-13																		
X	RYOTA Kawashima 等. "Performance Evaluation of Non-Tunneling Edge-Overlay Model on 40GbE Environment" 2014 IEEE 3rd Symposium on Network Cloud Computing and Applications, 2014年 12月 31日 (2014 - 12 - 31), 第69页第3段-71页第1段	1-13																		
X	RYOTA Kawashima 等. "Non-Tunneling Edge-Overlay Model using OpenFlow for Cloud Datacenter Networks" 2013 IEEE International Conference on Cloud Computing Technology and Science, 2013年 12月 31日 (2013 - 12 - 31), 第177页第3段-179页最后1段	1-13																		
A	CN 103581277 A (中兴通讯股份有限公司) 2014年 2月 12日 (2014 - 02 - 12) 全文	1-13																		
A	CN 104348724 A (华为技术有限公司) 2015年 2月 11日 (2015 - 02 - 11) 全文	1-13																		
<p><input checked="" type="checkbox"/> 其余文件在C栏的续页中列出。 <input checked="" type="checkbox"/> 见同族专利附件。</p>																				
<p>* 引用文件的具体类型:</p> <p>"A" 认为不特别相关的表示了现有技术一般状态的文件</p> <p>"E" 在国际申请日的当天或之后公布的在先申请或专利</p> <p>"L" 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)</p> <p>"O" 涉及口头公开、使用、展览或其他方式公开的文件</p> <p>"P" 公布日先于国际申请日但迟于所要求的优先权日的文件</p> <p>"T" 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件</p> <p>"X" 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性</p> <p>"Y" 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性</p> <p>"&" 同族专利的文件</p>																				
<p>国际检索实际完成的日期</p> <p>2016年 8月 3日</p>		<p>国际检索报告邮寄日期</p> <p>2016年 8月 24日</p>																		
<p>ISA/CN的名称和邮寄地址</p> <p>中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088</p> <p>传真号 (86-10)62019451</p>		<p>授权官员</p> <p>杨盈霄</p> <p>电话号码 (86-10)62413196</p>																		

C. 相关文件		
类型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	US 2015089583 A1 (WANSER, KELLY 等) 2015年 3月 26日 (2015 - 03 - 26) 全文	1-13

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/087112

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	104871495	A	2015年 8月 26日	WO	2014052485	A1	2014年 4月 3日
				EP	2891277	A1	2015年 7月 8日
				US	2014086253	A1	2014年 3月 27日
CN	103581277	A	2014年 2月 12日	WO	2014023255	A1	2014年 2月 13日
CN	104348724	A	2015年 2月 11日	WO	2015014187	A1	2015年 2月 5日
US	2015089583	A1	2015年 3月 26日	US	2014123212	A1	2014年 5月 1日
				EP	2915090	A1	2015年 9月 9日
				WO	2014070773	A1	2014年 5月 8日
				US	2014123211	A1	2014年 5月 1日