(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

(43) 国际公布日
2014 年 12 月 31 日 (31.12.2014)

WIPO | PCT

(10) 国际公布号
WO 2014/206105 A1

(51) 国际专利分类号:
H04L 12/931 (2013.01)

(21) 国际申请号: PCT/CN2014/072502

(22) 国际申请日: 2014 年 2 月 25 日 (25.02.2014)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201310270272.9 2013 年 6 月 28 日 (28.06.2013) CN

(71) 申请人: **华为技术有限公司 (HUAWEI TECHNOLOGIES CO., LTD.)** [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(72) 发明人: **林洋 (LIN, Yang)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。
郑坤 (ZHENG, Kun); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (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, LT, 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, 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)。

本国际公布:
— 包括国际检索报告(条约第 21 条(3))。

(54) **Title**: VIRTUAL SWITCH METHOD, RELEVANT APPARATUS, AND COMPUTER SYSTEM

(54) **发明名称** : 虚拟交换方法、相关装置和计算机系统

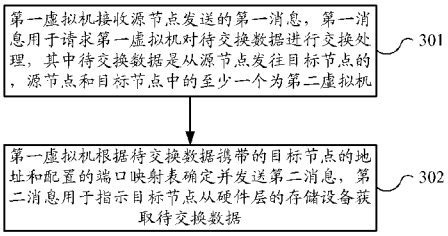


图3 /Fig.3

301 A first virtual machine receives a first message sent by a source node, where the first message is used for requesting the first virtual machine to perform switch processing on data to be switched, the data to be switched is sent from the source node to a target node, and at least one of the source node and the target node is a second virtual machine

302 The first virtual machine determines a second message according to an address of the target node and a configured port mapping table carried in the data to be switched and sends the second message, where the second message is used for instructing the target node to acquire the data to be switched from a storage device in a hardware layer

(57) Abstract: Embodiments of the present invention provide a virtual switch method, a relevant apparatus, and a computer system. The method comprises: receiving a first message sent by a source node, where the first message is used for requesting a first virtual machine to perform switch processing on data to be switched, the data to be switched is sent from the source node to a target node, and at least one of the source node and the target node is a second virtual machine; determining a second message according to an address of the target node and a configured port mapping table carried in the data to be switched and sending the second message, where the second message is used for instructing the target node to acquire the data to be switched from a storage device in a hardware layer. In the embodiment of the present invention, a virtual switch function is deployed to a virtual machine, so that the virtual machine having the virtual switch function and another ordinary virtual machine are equal in rank, so as to facilitate management on a virtual network by a Host and effective and proper allocation of network resources. Meanwhile, because a virtual switch function is stripped from the core of a Host, the expandability is enhanced.

(57) 摘要: 本发明实施例提供一种虚拟交换方法、相关装置和计算机系统，该方法包括：接收源节点发送的第一消息，第一消息用于请求第一虚拟机对待交换数据进行交换处理，其中待交换数据是从源节点发往目标节点的，源节点和目标节点中的至少一个为第二虚拟机；根据待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，第二消息用于指示目标节点从硬件层的存储设备获取待交换数据。本发明实施例通过将虚拟交换功能部署到虚拟机中，使得具有虚拟交换功能的虚拟机与其他普通虚拟机处于同等地位，从而有利于 Host 对虚拟网络进行管理并进行高效、合理的网络资源分配。并且由于虚拟交换功能从 Host 核心中剥离，从而增强了扩展性。

WO 2014/206105 A1

虚拟交换方法、相关装置和计算机系统

本申请要求于 2013 年 06 月 28 日提交中国专利局、申请号为 201310270272.9、发明名称为“**虚拟交换方法、相关装置和计算机系统**”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本发明涉及计算机技术领域，并且更具体地，涉及虚拟交换方法、相关装置和计算机系统。

背景技术

网络虚拟化是使用基于软件的抽象从物理网络元素中分离网络流量的一种方式。网络虚拟化与其他形式的虚拟化有很多共同之处。

对网络虚拟化来说，抽象隔离了网络中的交换机、网络端口、路由器以及其他物理元素的网络流量。每个物理元素被网络元素的虚拟表示形式所取代。管理员能够对虚拟网络元素进行配置以满足其独特的需求。网络虚拟化在此处的主要优势是将多个物理网络整合进更大的逻辑网络中。

现有的网络虚拟化主要方案为 VMware 的开放式虚拟交换（Open Virtual Switch，OVS）和分布式虚拟交换（Distributed Virtual Switch，DVS）。针对主流的 OVS 架构，虚拟交换机（Virtual Switch，vSwitch）在主机 Host 内核中实现，即在（Virtual Machine Monitor，VMM）内核中实现，处于虚拟网络的核心位置，其架构如图 1 所示。其中 vSwich 使用虚拟端口 port，通过 FE/BE 与连接虚拟机（Virtual Machine，VM）以及底层网卡（Network Interface Card，NIC）。Host 为其上运行的虚拟机以及各种虚拟硬件分配诸如 CPU、内存等物理资源，这些物理资源划分为内核空间物理资源和用户空间物理资源，vSwitch 在交换处理过程中需要申请占用较多的 Host 内核空间物理资源，因而非常不利于 Host 对虚拟网络进行管理和资源分配。vSwitch 负担了诸多任务和功能，例如图 1 中示出的虚拟局域网（Virtual Local Area Network，VLAN）、负载均衡 Load-balance、隧道 Tunneling、安全 Security、链路汇聚控制协议（Link

Aggregation Control Protocol，LACP）、服务质量（Quality of Service，QoS）等等，其设计非常庞大和复杂，vSwich 与 Host 内核的紧密耦合使得 vSwitch 以及整个虚拟网络的扩展性和灵活性都很差。

发明内容

本发明实施例提供一种虚拟交换方法、相关装置和计算机系统，将虚拟交换功能从内核中剥离，提高了虚拟交换设备的扩展性和灵活性，并将虚拟交换功能部署在虚拟机上，与普通虚拟机形成对等节点，从而有利于 Host 对虚拟网络进行管理并进行高效、合理的资源分配。

第一方面，提供了一种虚拟交换的方法，应用于计算节点上，所述计算节点包括：硬件层、运行在所述硬件层之上的宿主机 Host、以及运行在所述 Host 之上的至少一个虚拟机 VM，其中，所述硬件层包括输入/输出 I/O 设备和存储设备，所述至少一个虚拟机 VM 包括具有虚拟交换功能的第一虚拟机，所述至少一个 VM 还包括第二虚拟机，所述方法包括：所述第一虚拟机接收源节点发送的第一消息，所述第一消息用于请求所述第一虚拟机对待交换数据进行交换处理，其中所述待交换数据是从所述源节点发往目标节点的，所述源节点和所述目标节点中的至少一个为所述第二虚拟机；所述第一虚拟机根据所述待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，所述第二消息用于指示所述目标节点从所述硬件层的存储设备获取所述待交换数据。

结合第一方面，在其第一种实现方式中，所述第一虚拟机接收源节点发送的第一消息之前，还包括：所述第一虚拟机接收所述 Host 发送的配置命令；所述第一虚拟机根据所述配置命令配置用于与所述第二虚拟机进行通信的所述第一虚拟机的第一虚拟端口，并配置用于与所述 I/O 设备进行通信的所述第一虚拟机的第二虚拟端口；所述第一虚拟机建立所述第一虚拟端口与所述第二虚拟端口之间的映射关系，以生成所述端口映射表。

结合第一方面及其上述实现方式，在其第二种实现方式中，所述接收所述 Host 发送的配置命令之后，还包括：所述第一虚拟机根据所述配置命令配置

所述第二虚拟机对应的第一共享内存，其中所述第一共享内存为所述硬件层的存储设备上的指定存储区域。

结合第一方面及其上述实现方式，在其第三种实现方式中，当所述源节点为所述第二虚拟机，所述目标节点为所述 I/O 设备时，所述第一虚拟机接收源节点发送的第一消息，包括：所述第一虚拟机通过所述第一虚拟端口接收所述第二虚拟机发送的所述第一消息，所述第一消息包括用于向所述第一虚拟机指示所述第二虚拟机已完成将所述待交换数据写入所述第一共享内存的写完中断；所述第一虚拟机根据所述待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，包括：所述第一虚拟机根据用于接收所述第一消息的所述第一虚拟端口确定对应的所述第一共享内存的地址；从所述第一共享内存获取所述待交换数据，根据所述待交换数据携带的所述 I/O 设备的地址从所述端口映射表中确定与所述 I/O 设备对应的所述第二虚拟端口；确定携带有所述第一共享内存的地址和读取指令的所述第二消息，并通过所述第二虚拟端口向所述 I/O 设备发送所述第二消息，以便于所述 I/O 设备从所述第一共享内存读取所述待交换数据。

结合第一方面及其上述实现方式，在其第四种实现方式中，当所述源节点为所述 I/O 设备，所述目标节点为所述第二虚拟机时，所述第一虚拟机接收源节点发送的第一消息之后还包括：所述第一虚拟机从所述 I/O 设备获取所述待交换数据携带的目标节点的地址，所述目标节点的地址为所述第二虚拟机的地址；所述第一虚拟机根据所述待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，包括：所述第一虚拟机根据所述第二虚拟机的地址查询所述端口映射表以确定与所述第二虚拟机对应的第一虚拟端口并确定与所述第二虚拟机对应的第一共享内存的地址；通过所述 I/O 设备所对应的所述第二虚拟端口向所述I/O设备发送携带有所述第一共享内存的地址的回复消息，以便于所述 I/O 设备根据所述回复消息将所述待交换数据写入所述第一共享内存；在所述第一虚拟机接收到所述 I/O 设备发送的用于向所述第一虚拟机指示所述I/O设备已完成将所述待交换数据写入所述第一共享内存的写完中断时，确定携带有读取指令的所述第二消息，通过所述第一虚拟端口向所述第二虚拟机发送所述第二消息，以便于所述第二虚拟机从所述第一共

享内存读取所述待交换数据。

结合第一方面及其上述实现方式，在其第五种实现方式中，所述至少一个VM还包括第三虚拟机，当所述源节点为所述第二虚拟机，所述目标节点为所述第三虚拟机时，所述第一虚拟机接收源节点发送的第一消息，包括：所述第一虚拟机通过所述第一虚拟端口接收所述第二虚拟机发送的所述第一消息，所述第一消息包括用于向所述第一虚拟机指示所述第二虚拟机已完成将所述待交换数据写入所述第二虚拟机与所述第三虚拟机通过所述第一虚拟机预先协商的第二共享内存的写完中断，其中所述第二共享内存为所述硬件层的存储设备上的指定存储区域；所述第一虚拟机根据所述待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，包括：所述第一虚拟机根据用于接收所述第一消息的所述第一虚拟端口确定与所述第一虚拟端口对应的所述第二虚拟机的地址；根据所述第二虚拟机的地址和所述待交换数据携带的第三虚拟机的地址确定所述第二共享内存的地址；确定携带有所述第二共享内存的地址和读取指令的所述第二消息，并向所述第三虚拟机发送所述第二消息，以便于所述第三虚拟机从所述第二共享内存读取所述待交换数据。

结合第一方面及其上述实现方式，在其第六种实现方式中，所述方法还包括：接收所述目标节点发送的读完指示信息，以便于所述第一共享内存或所述第二共享内存被释放。

结合第一方面及其上述实现方式，在其第七种实现方式中，所述第一虚拟机接收源节点发送的第一消息之后，还包括：所述第一虚拟机根据所述待交换数据携带的目标节点的地址，在配置的开放流 Openflow 流表中确定与所述目标节点的地址所匹配的表项，其中，所述 Openflow 流表中包括至少一个表项，所述表项包括地址、虚拟端口和执行动作参数；如果所述匹配的表项存在，所述第一虚拟机根据所述匹配的表项中与所述目标节点的地址所对应的执行动作参数处理所述待交换数据；如果所述匹配的表项不存在，所述第一虚拟机建立能够与所述待交换数据匹配的新表项，并在所述 Openflow 流表中插入所述新表项。

第二方面，提供了一种宿主机，其特征在于，包括：创建模块，用于在输

入/输出 I/O 设备的 I/O 虚拟功能启动后，在宿主机 Host 之上产生至少一个虚拟机 VM，其中所述至少一个 VM 包括具有虚拟交换功能的第一虚拟机，所述至少一个 VM 还包括第二虚拟机；配置模块，用于向所述第一虚拟机发送配置命令，以便于所述第一虚拟机根据所述配置命令配置用于与所述第二虚拟机进行通信的所述第一虚拟机的第一虚拟端口，并配置用于与所述 I/O 设备进行通信的所述第一虚拟机的第二虚拟端口。

第三方面，提供了一种计算节点，其特征在于，运行在宿主机 Host 之上，所述 Host 运行在硬件层之上，所述硬件层包括输入/输出 I/O 设备和存储设备，所述虚拟机包括：接收模块，用于接收源节点发送的第一消息，所述第一消息用于请求所述虚拟机对待交换数据进行交换处理，其中所述待交换数据是从所述源节点发往目标节点的，所述源节点和所述目标节点中的至少一个为第二虚拟机，所述第二虚拟机运行在所述 Host 之上；交换处理模块，用于根据所述待交换数据携带的目标节点的地址和所述虚拟机配置的端口映射表确定第二消息，所述第二消息用于指示所述目标节点从所述硬件层的存储设备获取所述待交换数据；发送模块，用于向所述目标节点发送所述第二消息。

结合第三方面，在其第一种实现方式中，其特征在于，包括：代理 Agent 模块，用于根据所述 Host 发送的配置命令，配置用于与所述第二虚拟机进行通信的所述虚拟机的第一虚拟端口，并配置用于与所述 I/O 设备进行通信的所述虚拟机的第二虚拟端口；生成模块，用于建立所述第一虚拟端口与所述第二虚拟端口之间的映射关系，以生成所述端口映射表。

结合第三方面及其上述实现方式，在其第二种实现方式中，其特征在于，所述 Agent 模块，还用于根据所述配置命令配置所述第二虚拟机对应的第一共享内存，其中所述第一共享内存为所述硬件层的存储设备上的指定存储区域。

结合第三方面及其上述实现方式，在其第三种实现方式中，所述接收模块，具体用于通过所述第一虚拟端口接收所述第一消息，所述第一消息包括用于向所述虚拟机指示所述源节点已完成将所述待交换数据写入所述第一共享内存的写完中断；所述交换处理模块，具体用于根据用于接收所述第一消息的所述第一虚拟端口确定对应的所述第一共享内存的地址；从所述第一共享内存获取所述待交换数据，根据所述待交换数据携带的所述目标节点的地址从所述端口

映射表中确定与所述目标节点对应的所述第二虚拟端口；确定携带有所述第一共享内存的地址和读取指令的所述第二消息；所述发送模块，具体用于通过所述第二虚拟端口向所述目标节点发送所述第二消息；其中，所述源节点为所述第二虚拟机，所述目标节点为所述 I/O 设备。

结合第三方面及其上述实现方式，在其第四种实现方式中，所述接收模块，具体用于接收源节点发送的所述第一消息；所述交换处理模块，具体用于获取所述待交换数据携带的目标节点的地址；根据所述目标节点的地址查询所述端口映射表以确定与所述目标节点对应的第一虚拟端口并确定与所述目标节点对应的第一共享内存的地址；所述发送模块，具体用于通过所述源节点所对应的所述第二虚拟端口向所述源节点发送携带有所述第一共享内存的地址的回复消息；所述交换处理模块，还用于在接收到所述源节点发送的用于向所述虚拟机指示所述源节点已完成将所述待交换数据写入所述第一共享内存的写完中断时，确定携带有读取指令的所述第二消息；所述发送模块，还用于通过所述第一虚拟端口向所述目标节点发送所述第二消息；所述接收模块，还用于接收所述源节点发送的指示所述源节点已完成将所述待交换数据写入所述第一共享内存的写完中断；其中，所述源节点为所述 I/O 设备，所述目标节点为所述第二虚拟机。

结合第三方面及其上述实现方式，在其第五种实现方式中，所述接收模块，具体用于通过所述第一虚拟端口接收所述源节点发送的所述第一消息，所述第一消息包括写完中断；所述交换处理模块，具体用于根据用于接收所述第一消息的所述第一虚拟端口确定所述第一虚拟端口对应的所述源节点的地址；根据所述源节点的地址和所述待交换数据携带的目标节点的地址确定所述第二共享内存的地址；确定携带有所述第二共享内存的地址和读取指令的所述第二消息；所述发送模块，具体用于向所述目标节点发送所述第二消息；其中，所述至少一个 VM 还包括第三虚拟机，所述源节点为所述第二虚拟机，所述目标节点为所述第三虚拟机。

第四方面，提供了一种计算节点，包括：硬件层、运行在所述硬件层之上的宿主机 Host、以及运行在所述 Host 之上的至少一个虚拟机 VM，其中，所述硬件层包括输入/输出 I/O 设备和存储设备，所述至少一个虚拟机 VM 包括

具有虚拟交换功能的第一虚拟机，所述至少一个 VM 还包括第二虚拟机，其中：所述第一虚拟机，用于接收源节点发送的第一消息，所述第一消息用于请求所述第一虚拟机对待交换数据进行交换处理，其中所述待交换数据是从所述源节点发往目标节点的，所述源节点和所述目标节点中的至少一个为所述第二虚拟机；所述第一虚拟机，还用于根据所述待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，所述第二消息用于指示所述目标节点从所述硬件层的存储设备获取所述待交换数据。

结合第四方面，在其第一种实现方式中，所述 Host，用于向所述第一虚拟机发送配置命令；所述第一虚拟机，还用于根据所述配置命令配置用于与所述第二虚拟机进行通信的所述第一虚拟机的第一虚拟端口，并配置用于与所述 I/O 设备进行通信的所述第一虚拟机的第二虚拟端口；所述第一虚拟机，还用于建立所述第一虚拟端口与所述第二虚拟端口之间的映射关系，以生成所述端口映射表。

结合第四方面及其上述实现方式，在其第二种实现方式中，所述第一虚拟机，还用于根据所述配置命令配置所述第二虚拟机对应的第一共享内存，其中所述第一共享内存为所述硬件层的存储设备上的指定存储区域。

结合第四方面及其上述实现方式，在其第三种实现方式中，所述源节点，用于将所述待交换数据写入所述第一共享内存；所述源节点，还用于向所述第一虚拟机发送所述第一消息；所述第一虚拟机，具体用于通过所述第一虚拟端口接收所述第一消息，所述第一消息包括用于向所述第一虚拟机指示所述源节点已完成将所述待交换数据写入所述第一共享内存的写完中断；以及根据用于接收所述第一消息的所述第一虚拟端口确定对应的所述第一共享内存的地址；从所述第一共享内存获取所述待交换数据，根据所述待交换数据携带的所述 I/O 设备的地址从所述端口映射表中确定与所述 I/O 设备对应的所述第二虚拟端口；确定携带有所述第一共享内存的地址和读取指令的所述第二消息，并通过所述第二虚拟端口向所述目标节点发送所述第二消息；所述目标节点，用于根据所述第二消息从所述第一共享内存读取所述待交换数据；其中，所述源节点为所述第二虚拟机，所述目标节点为所述 I/O 设备。

结合第四方面及其上述实现方式，在其第四种实现方式中，所述第一虚拟

机，具体用于接收源节点发送的所述第一消息，获取所述待交换数据携带的目标节点的地址；根据所述目标节点的地址查询所述端口映射表以确定与所述目标节点对应的第一虚拟端口并确定与所述目标节点对应的第一共享内存的地址；通过所述源节点所对应的所述第二虚拟端口向所述源节点发送携带有所述第一共享内存的地址的回复消息；以及，在接收到所述源节点发送的用于向所述第一虚拟机指示所述源节点已完成将所述待交换数据写入所述第一共享内存的写完中断时，确定携带有读取指令的所述第二消息，通过所述第一虚拟端口向所述目标节点发送所述第二消息；所述源节点，还用于根据所述回复消息中的所述第一共享内存的地址将所述待交换数据写入所述第一共享内存；所述源节点，还用于向所述第一虚拟机发送指示所述源节点已完成将所述待交换数据写入所述第一共享内存的写完中断；所述目标节点，用于根据所述第二消息从所述第一共享内存读取所述待交换数据；其中，所述源节点为所述 I/O 设备，所述目标节点为所述第二虚拟机。

结合第四方面及其上述实现方式，在其第五种实现方式中，所述源节点，还用于将所述待交换数据写入所述源节点与所述目标节点通过所述第一虚拟机预先协商的第二共享内存，其中所述第二共享内存为所述硬件层的存储设备上的指定存储区域；所述源节点，还用于通过所述第一虚拟端口向所述第一虚拟机发送所述第一消息，所述第一消息包括写完中断；所述第一虚拟机，具体用于根据用于接收所述第一消息的所述第一虚拟端口确定所述第一虚拟端口对应的所述源节点的地址；根据所述源节点的地址和所述待交换数据携带的目标节点的地址确定所述第二共享内存的地址；确定携带有所述第二共享内存的地址和读取指令的所述第二消息，并向所述目标节点发送所述第二消息；所述目标节点，用于根据所述第二消息从所述第二共享内存读取所述待交换数据；其中，所述至少一个 VM 还包括第三虚拟机，所述源节点为所述第二虚拟机，所述目标节点为所述第三虚拟机。

结合第四方面及其上述实现方式，在其第六种实现方式中，所述目标节点根据所述第二消息从所述共享内存读取所述待交换数据之后，所述目标节点，还用于向所述第一虚拟机发送读完指示信息，以便于所述第一共享内存或所述第二共享内存被释放；所述第一虚拟机，还用于释放所述第一共享内存或所述

第二共享内存。

结合第四方面及其上述实现方式，在其第七种实现方式中，在接收源节点发送的第一消息之后，所述第一虚拟机，还用于根据所述待交换数据携带的目标节点的地址，在配置的开放流 Openflow 流表中确定与所述目标节点的地址所匹配的表项，其中，所述 Openflow 流表中包括至少一个表项，所述表项包括地址、虚拟端口和执行动作参数；如果所述匹配的表项存在，根据所述匹配的表项中与所述目标节点的地址所对应的执行动作参数处理所述待交换数据；如果所述匹配的表项不存在，建立能够与所述待交换数据匹配的新表项，并在所述 Openflow 流表中插入所述新表项。

第五方面，提供了一种计算机系统，包括：至少一个如第四方面所述的计算节点。

由上可见，本发明实施例中的计算节点包括：硬件层、运行在所述硬件层之上的宿主机 Host、以及运行在所述 Host 之上的至少一个虚拟机 VM，其中，所述硬件层包括输入/输出 I/O 设备和存储设备，所述至少一个虚拟机 VM 包括具有虚拟交换功能的第一虚拟机，所述至少一个 VM 还包括第二虚拟机；如此，将虚拟交换功能实现在虚拟机上，使得虚拟交换机与普通 VM 处于同等优先级，形成对等的网络虚拟化架构，在进行资源分配时虚拟交换机和普通 VM 一样使用用户空间的物理资源，这样便于 Host 进行管理和高效合理地进行带宽、CPU、存储等资源的分配。应用于该计算节点上的虚拟交换方法包括：所述第一虚拟机接收源节点发送的第一消息，所述第一消息用于请求所述第一虚拟机对待交换数据进行交换处理，其中所述待交换数据从所述源节点发往目标节点，所述源节点和所述目标节点中的至少一个为所述第二虚拟机；所述第一虚拟机根据所述待交换数据携带的目标节点的地址和所述配置的端口映射表确定第二消息并发送所述第二消息，所述第二消息用于指示所述目标节点从所述硬件层的存储设备获取所述待交换数据。该方法将虚拟交换功能从 Host 内核中剥离解耦，降低与 Host 的耦合性，可以在同一 Host 内部署多个 vSwitch，不受 Host 约束，因此具有更强的扩展性，并且解耦后 vSwtich 不再依赖 Host 内核中的操作系统，变得更加易于部署，所以获得了更好的移植性，并且由于配置模块（agent）与待交换数据交换转发模块（端口映射表）相分离，更符

合软件定义网络的要求。

附图说明

为了更清楚地说明本发明实施例的技术方案，下面将对本发明实施例中所需要使用的附图作简单地介绍，显而易见地，下面所描述的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1 是现有技术中 OVS 的架构图。

图 2 是本发明一个实施例的虚拟化软硬件体系架构示意图。

图 3 是本发明一个实施例的虚拟交换方法的流程图。

图 4 是本发明一个实施例的虚拟交换数据流的示意图。

图 5 是本发明另一实施例的虚拟交换数据流的示意图。

图 6 是本发明另一实施例的虚拟交换数据流的示意图。

图 7 是本发明另一实施例的用于软件定义网络 SDN 的虚拟交换设备的示意图。

图 8 是本发明另一实施例的分布式实施的示意图。

图 9 是本发明另一实施例的分布式实施的流程图。

图 10 是本发明一个实施例的宿主机的模块架构示意图。

图 11 是本发明一个实施例的虚拟机的模块架构示意图。

图 12 是本发明一个实施例的计算机节点的示意图。

图 13 是本发明一个实施例的计算机系统的示意图。

具体实施方式

下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明的一部分实施例，而不是全

部实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例，都应属于本发明保护的范围。

为了方便理解本发明实施例，首先在此介绍本发明实施例描述中会引入的几个术语；

虚拟机 VM:

通过虚拟机软件可以在一台物理计算机上模拟出一台或者多台虚拟的计算机，而这些虚拟机就像真正的计算机那样进行工作，虚拟机上可以安装操作系统和应用程序，虚拟机还可访问网络资源。对于在虚拟机中运行的应用程序而言，虚拟机就像是在真正的计算机中进行工作。

硬件层:

虚拟化环境运行的硬件平台。其中，硬件层可包括多种硬件，例如某计算节点的硬件层可包括 CPU 和内存，还可以包括网卡（Network Interface Card，NIC）、存储器等等高速/低速输入/输出（I/O，Input/Output）设备，其中 NIC 为底层物理网卡，以下简称 Host NIC 来区别于虚拟机的虚拟网卡 VM NIC。

宿主机（Host）:

作为管理层，用以完成硬件资源的管理、分配；为虚拟机呈现虚拟硬件平台；实现虚拟机的调度和隔离。其中，Host 可能是虚拟机监控器（VMM）；或者，有时 VMM 和 1 个特权虚拟机配合，两者结合组成 Host。其中，虚拟硬件平台对其上运行的各个虚拟机提供各种硬件资源，如提供虚拟 CPU、内存、虚拟磁盘、虚拟网卡等等。其中，该虚拟磁盘可对应 Host 的一个文件或者一个逻辑块设备。虚拟机则运行在 Host 为其准备的虚拟硬件平台上，Host 上运行一个或多个虚拟机。

虚拟交换机（Virtual Switch，vSwitch）:

虚拟交换机在 Host 的控制下将虚拟机互相连接起来，并且接入到物理网络当中，虚拟交换机就像真正的虚拟机那样工作，现有的虚拟交换机在 Host 内核中实现，处于虚拟网络的核心位置，负担虚拟局域网（Virtual Local Area Network，VLAN）、负载均衡 Load-balance、隧道 Tunneling、安全 Security、

链路汇聚控制协议（Link Aggregation Control Protocol，LACP）、服务质量（Quality of Service，QoS）等等诸多功能。

共享内存:

操作系统进程间通信（Inter-Process Communication，IPC）的一种机制，共享内存是进程间通信中最简单的方式之一，共享内存允许两个或更多进程访问同一块内存，在网络虚拟化中，共享内存允许两个或者更多的虚拟机、虚拟硬件访问同一块内存。共享内存在各种进程间通信方式中具有最高效率。

零拷贝:

避免 CPU 将数据从一块存储拷贝到另外一块存储的技术，通过减少或消除关键通信路径影响速率的操作，降低数据传输的开销，从而有效的提高通信性能，实现高速数据传输，实现有 IO 直通，MMAP 等方式

软件定义网络（Software Defined Network，SDN）:

SDN 是新一代网络架构，其核心技术开放流 Openflow 通过将网络设备控制面与数据面分离开来，从而实现了网络流量的灵活控制，为核心网络以及应用的创新提供了良好的平台。

图 2 示出了本发明实施例中将 vSwitch 部署到 VM 中的虚拟化方案的软硬件体系架构示意图，该体系架构主要包括三个层次：硬件层、Host 和虚拟机（VM）。其中硬件层包括 I/O 设备，即物理网卡 NIC，通过该 NIC 可以与外界其他 Host 或网络进行通信，硬件层还可以包括存储设备，例如内存、硬盘等等。Host 运行在硬件层之上，其中 Host 可能是虚拟机监控器（VMM），或者，有时 VMM 和 1 个特权虚拟机配合，两者结合组成 Host，图 2 中示出的为第二种情况，然而这仅仅为一个示例，本发明对此不作限定。在 Host 之上运行的至少一个虚拟机 VM，其中一个 VM 为本发明中的具有虚拟交换功能的虚拟机（第一虚拟机），同时还可以有若干个普通的虚拟机（第二虚拟机、第三虚拟机等等）。

以该体系架构建立虚拟化网络环境的过程中，Host 中的配置管理模块（Config and Manage Module，CMM）可以向具有虚拟交换功能的第一虚拟机（以下用 vSwitch 代称）发送配置命令来进行虚拟网络环境的配置以及 vSwitch

的配置。具体地，CMM 可以通过 vSwitch 中的配置代理模块（agent）来进行配置，包括端口映射表，VLAN 表，访问控制列表（Access Control List，ACL）等的管理和配置。其中该 Host 中的配置管理模块可以通过 IPC（例如 IOCTL，NETLINK，SOCKET 等）与 vSwitch 的 agent 模块相连接，从而可以将 Host 虚拟环境的配置传入 vSwitch，具体可以包括 Host NIC、VM 的后端 BE、共享内存、DMA 中断等配置信息，使得 vSwitch 获得虚拟环境信息，从而建立相应的虚拟网络环境。

具体地，可以在 VM 创建好后，由配置管理模块为 VM 创建虚拟 NIC 接口，而后配置管理模块可以通过 agent 模块协商 vSwitch 与 Host NIC 的通信机制（通信方式）和端口映射，且协商 vSwitch 与 VMM NIC 之间的通信机制（通信方式）和端口映射，还可以进一步地协商 vSwitch 与 VMM NIC 之间的共享内存等。其中，vSwitch 与 Host NIC 之间可以使用 IO 直通、零拷贝等方式通信，vSwitch 与 VM 之间可以使用共享内存、前后端 FE/BE 事件通道等技术进行通信。根据协商好的各项配置的对应关系建立表项，以生成映射表，例如将 VM 的地址、该 VM 所对应的 vSwitch 的虚拟端口的端口号、该 VM 与 vSwitch 之间协商的共享内存地址建立对应关系，以形成表项，其中该 VM 为普通虚拟机，例如第二虚拟机。

该虚拟化网络环境搭建好后，进行数据交换时：该第一虚拟机（vSwitch），用于接收源节点发送的第一消息，该第一消息用于请求该第一虚拟机对待交换数据进行交换处理，其中该待交换数据从该源节点发往目标节点，该源节点和该目标节点中的至少一个为该第二虚拟机；该第一虚拟机，还用于根据该待交换数据携带的目标节点的地址和该配置的端口映射表确定第二消息并发送所述第二消息，该第二消息用于指示该目标节点从该硬件层的存储设备获取该待交换数据。从而通过 vSwitch 的信令控制，和交换处理，实现了待交换数据的转发。如此，虚拟交换功能从 Host 内核中剥离解耦，转而在虚拟机上实现虚拟交换的功能，简化了 Host 内核的设计和负担，并且由于 VM 具有灵活性和很好的扩展性，从而使得 vSwitch 以及整个虚拟网络的扩展性和灵活性都得到了提高。进一步地，因为将虚拟交换功能实现在虚拟机上，使得虚拟交换机与普通 VM 处于同等优先级，形成对等的网络虚拟化架构，在进行资源分配时虚

拟交换机和普通 VM 一样使用用户空间的物理资源，这样便于 Host 进行管理和高效合理地进行资源分配。

图 3 是本发明一个实施例的虚拟交换方法的流程图。图 2 的方法由具有虚拟交换功能的虚拟机（下文简称为第一虚拟机）执行。

301，第一虚拟机接收源节点发送的第一消息，第一消息用于请求第一虚拟机对待交换数据进行交换处理，其中待交换数据从源节点发往目标节点，源节点和目标节点中的至少一个为第二虚拟机。

第一虚拟机为具有虚拟交换功能的虚拟机，与其他普通虚拟机处于同等地位并运行在 Host 之上。其中源节点可以是该 Host 上的普通虚拟机 VM，应当理解的是：这里的普通虚拟机是相对于具有虚拟交换功能的虚拟机而言，也可以是该 Host 外部的虚拟机或物理机，然而由于该 Host 是通过 Host NIC 与外界进行通信的，所以与该 Host 外部的虚拟机或物理机的通信都简化地描述成与 Host NIC 进行通信，即源节点也可以是 Host NIC。同样地，目标节点也可以是该 Host 上的普通虚拟机 VM，也可以是 Host NIC。

302，第一虚拟机根据待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，第二消息用于指示目标节点从硬件层的存储设备获取待交换数据。

应理解，上述步骤 302 中，配置的端口映射表可以由第一虚拟机来进行配置，包括虚拟化网络建立初期端口映射表的初始化配置以及虚拟化网络后期运行时端口映射表的动态维护。而第一虚拟机可以仅仅是配置命令的执行者，而配置命令可以由 Host 或者网络维护人员配置。

本发明实施例通过将虚拟交换功能部署到虚拟机中，简化了 VMM，有利于 Host 对虚拟网络进行管理并进行高效、合理的网络资源分配。

可选地，作为一个实施例，步骤 301 之前，还包括：接收 Host 发送的配置命令；根据配置命令配置用于与第二虚拟机进行通信的第一虚拟机的第一虚拟端口，并配置用于与 I/O 设备进行通信的第一虚拟机的第二虚拟端口；建立第一虚拟端口与第二虚拟端口之间的映射关系，以生成端口映射表。

可选地，作为另一个实施例，第一虚拟机根据配置命令配置第二虚拟机对

应的第一共享内存，其中第一共享内存为硬件层的存储设备上的指定存储区域。

具体地，Host 中的配置管理模块可以通过 vSwitch 中的 agent 模块协商 vSwitch 与 Host NIC 的通信机制（通信方式）和端口映射，且协商 vSwitch 与 VMM NIC 之间的通信机制（通信方式）和端口映射，可选地，还可以进一步地协商 vSwitch 与 VMM NIC 之间的共享内存等，其中共享内存为硬件层的存储设备上的指定存储区域。而后可以将协商好的各项配置的对应关系建立表项，生成端口映射表，例如，将 VM 的地址、该 VM 所对应的 vSwitch 的端口号、该 VM 与 vSwitch 之间协商的共享内存地址建立对应的关系，生成端口映射表的表项。在进行虚拟交换时，第一虚拟机从第一虚拟机的第一虚拟端口接收待交换数据，其中第一虚拟端口对应于源节点；通过第一虚拟机的第二虚拟端口向目标节点发送待交换数据，其中第二虚拟端口是第一虚拟机根据第一虚拟端口和预先配置的端口映射表确定的。上述从第一虚拟端口接收待交换数据，并通过第二虚拟端口向目标节点发送待交换数据的过程为第一虚拟机的逻辑交换过程。其中，第一虚拟机与源节点通信的第一虚拟端口，第一虚拟机与目标节点通信的第二虚拟端口都是预先协商并配置好的。

可选地，作为另一个实施例，当源节点为第二虚拟机，目标节点为 I/O 设备时，第一虚拟机接收源节点发送的第一消息，包括：第一虚拟机通过第一虚拟端口接收第二虚拟机发送的第一消息，第一消息包括用于向第一虚拟机指示第二虚拟机已完成将待交换数据写入共享内存的写完中断；第一虚拟机根据用于接收第一消息的第一虚拟端口确定对应的第一共享内存的地址；从第一共享内存获取待交换数据，根据待交换数据携带的 I/O 设备的地址从端口映射表中确定与 I/O 设备对应的第二虚拟端口；确定携带有第一共享内存的地址和读取指令的第二消息，并通过第二虚拟端口向 I/O 设备发送第二消息，以便于 I/O 设备从第一共享内存读取待交换数据。

具体地，作为源节点的 Host 中的第二虚拟机与第一虚拟端口建立虚连接，其中第一虚拟端口是第一虚拟机预先配置的与该第二虚拟机对应的虚拟端口。第二虚拟机向第一虚拟端口发送待交换数据，该待交换数据实际写入该第二虚拟机与第一虚拟机预先协商的共享内存中。写入完毕后，第二虚拟机向第一虚

拟机发送写完指示信息，第一虚拟机查询内部与配置的端口映射表，以确定第二虚拟端口以及与第二虚拟端口对应的主机网卡 Host NIC，通过第二虚拟端口向 Host NIC 发送读取指示信息，令 Host NIC 从共享内存中读取该待交换数据，以便于 Host NIC 进一步向 Host 外部的目标节点发送该待交换数据。应理解，在第二虚拟机向 Host 外部发送待交换数据的过程中，目标节点也可以理解为 Host NIC。

可选地，作为另一个实施例，当源节点为 I/O 设备，目标节点为第二虚拟机时，第一虚拟机接收源节点发送的第一消息之后还包括：第一虚拟机接收源节点发送的第一消息之后还包括：第一虚拟机从 I/O 设备获取待交换数据携带的目标节点的地址，目标节点的地址为第二虚拟机的地址；第一虚拟机根据待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，包括：第一虚拟机根据第二虚拟机的地址查询端口映射表以确定与第二虚拟机对应的第一虚拟端口并确定与第二虚拟机对应的第一共享内存的地址；通过 I/O 设备所对应的第二虚拟端口向 I/O 设备发送携带有第一共享内存的地址的回复消息，以便于 I/O 设备根据回复消息将待交换数据写入第一共享内存；在第一虚拟机接收到 I/O 设备发送的用于向第一虚拟机指示 I/O 设备已完成将待交换数据写入第一共享内存的写完中断时，确定携带有读取指令的第二消息，通过第一虚拟端口向第二虚拟机发送第二消息，以便于第二虚拟机从第一共享内存读取待交换数据。

具体地，第一虚拟机从 I/O 设备获取待交换数据携带的目标节点的地址，是由第一虚拟机在接收到第一消息的通知后，得知 I/O 设备（即底层物理网卡）接收到了待交换数据，之后第一虚拟机则可以通过驱动层直接访问该待交换数据以获取其携带的目标节点的地址。

可选地，作为另一个实施例，至少一个 VM 还包括第三虚拟机，当源节点为第二虚拟机，目标节点为第三虚拟机时，即源节点和目标节点均为 Host 上的普通 VM 时，第一虚拟机接收源节点发送的第一消息，包括：第一虚拟机通过第一虚拟端口接收第二虚拟机发送的第一消息，第一消息包括用于向第一虚拟机指示第二虚拟机已完成将待交换数据写入第二虚拟机与第三虚拟机通过第一虚拟机预先协商的第二共享内存的写完中断，其中第二共享内存为硬件层

的存储设备上的指定存储区域; 第一虚拟机根据待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，包括: 第一虚拟机根据用于接收第一消息的第一虚拟端口确定与第一虚拟端口对应的第二虚拟机的地址; 根据第二虚拟机的地址和待交换数据携带的第三虚拟机的地址确定第二共享内存的地址; 确定携带有第二共享内存的地址和读取指令的第二消息，并向第三虚拟机发送第二消息，以便于第三虚拟机从第二共享内存读取待交换数据。

其中，第二共享内存是第二虚拟机与第三虚拟机通过第一虚拟机进行协商的，具体可以通过 Xen 的事件通道（Event Channel）进行协商。

可选地，作为另一个实施例，上述方法还包括: 接收目标节点发送的读完指示信息，以便于释放第一共享内存或者第二共享内存。具体地，目标节点在读取完待交换数据后向第一虚拟机发送读完指示信息，第一虚拟机接收到读完指示信息后，恢复共享内存的可写权限，即释放该共享内存。

应理解，以上所述第一共享内存和第二共享内存仅仅为了区分，对本发明不构成限定。第一共享内存和第二共享内存都是硬件层存储设备上指定的一部分内存空间，具有随机性和不确定性。例如，第一共享内存被释放后，也可能转而被分配作为第二共享内存，在这种情况下，第一共享内存和第二共享内存对应相同的内存空间。

可选地，作为另一个实施例，在端口映射表为开放流 Openflow 流表时，第一虚拟机根据待交换数据携带的目标节点的地址，在该 Openflow 流表中确定与目标节点的地址所匹配的表项，其中，Openflow 流表中包括至少一个表项，表项包括地址、虚拟端口和执行动作参数; 如果匹配的表项存在，第一虚拟机根据匹配的表项中与目标节点的地址所对应的执行动作参数处理待交换数据; 如果匹配的表项不存在，第一虚拟机建立能够与待交换数据匹配的新表项，并在 Openflow 流表中插入新表项。

由上可见，本发明实施例中的计算节点包括: 硬件层、运行在硬件层之上的宿主机 Host、以及运行在 Host 之上的至少一个虚拟机 VM，其中，硬件层包括输入/输出 I/O 设备和存储设备，至少一个虚拟机 VM 包括具有虚拟交换功能的第一虚拟机，至少一个 VM 还包括第二虚拟机; 如此，将虚拟交换功能

实现在虚拟机上，使得虚拟交换机与普通 VM 处于同等优先级，形成对等的网络虚拟化架构，在进行资源分配时虚拟交换机和普通 VM 一样使用用户空间的物理资源，这样便于 Host 进行管理和高效合理地进行资源分配。应用于该计算节点上的虚拟交换方法包括：第一虚拟机接收源节点发送的第一消息，第一消息用于请求第一虚拟机对待交换数据进行交换处理，其中待交换数据从源节点发往目标节点，源节点和目标节点中的至少一个为第二虚拟机；第一虚拟机根据待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，第二消息用于指示目标节点从硬件层的存储设备获取待交换数据。该方法将虚拟交换功能从 Host 内核中剥离解耦，转而在虚拟机上实现虚拟交换的功能，简化了 Host 内核的设计和负担，并且由于 VM 具有灵活性和很好的扩展性，从而使得 vSwitch 以及整个虚拟网络的扩展性和灵活性都得到了提高，便于控制面和数据面的分离，使其满足 SDN 的需求，支持 Openflow。

图 4 是本发明一个实施例的虚拟交换数据流的示意图。如图 4 所示，虚拟交换机 vSwitch（虚拟交换功能）部署于第一虚拟机上，使该第一虚拟机成为虚拟交换设备，并与普通的虚拟机 VM1、VM2 处于同等地位。其中第一虚拟机中的代理 Agent 模块与主机 Host 中的配置管理模块（Config and Manage Module）连接，以便于系统管理员对第一虚拟机进行配置。第一虚拟机的虚拟端口 port 可以与 VM1、VM2 或者 VMM 的底层物理网卡 HOST NIC 进行连接。以下通过数据流详细说明 Host 中的普通 VM（例如 VM1）向外界（HOST NIC）发送待交换数据的过程。应理解，图 4 所示的系统架构仅仅为一个示例，其中 VM、port、Host NIC 等模块的数量可以进行扩展。

401，预先配置。

在进行虚拟交换之前，需要构建虚拟网络，并对虚拟交换机 vSwitch（第一虚拟机）进行预先配置。具体可以通过 Host 上的 Config and Manage Module 向第一虚拟机中的 Agent 模块发送配置命令，使得 Agent 模块对 vSwitch 的端口映射、VLAN 管理等进行配置。

具体地，可以协商普通 VM 与 vSwitch 的通信方式、共享内存 Share Memory 和端口，协商 vSwitch 与 HOST NIC 的通信方式和端口，配置 vSwich 的端口

映射，以生成端口映射表。其中，通信方式可以包括共享内存、IO 直通、零拷贝或直接内存存取（Direct Memory Access，DMA）等。共享内存是操作系统进程间通信（IPC）的一种机制，零拷贝为避免中央处理器 CPU 将数据从一块存储拷贝到另外一块存储的技术，其实现由 IO 直通、MMAP 等方式。其中，作为更优选的实施例，普通 VM 与 vSwitch 通过共享内存的方式进行通信，vSwitch 与 Host NIC 通过 IO 直通或 DMA 方式进行通信，可以使得本发明所涉及的交换设备实现零拷贝，从而降低了资源开销，提高了交换效率。

402，建立虚连接。

当 VM1 需要向 Host 外部（Host NIC）发送数据时，VM1 首先与 vSwitch 的第一虚拟端口 port1 建立虚连接，其中 port1 是步骤 401 中 Agent 模块预先配置与 VM1 相对应的虚拟端口。相应的物理过程为，VM1 通过其虚拟网卡 VM NIC 映射到 VM1 对应的共享内存。

403，写入待交换数据。

之后，VM1 通过其 NIC 向 port1 发送待交换数据。相应的实际物理过程为，将待交换数据写入 VM1 对应的共享内存。VM1 写入完毕后，通过 port1 向 vSwitch 发送写完指示信息，以通知 vSwitch 进行下一步操作。具体地，该写完指示信息可以是写完中断。

404，交换处理过程。

vSwitch 接收到 VM1 发送的写完指示信息后，转入交换处理过程，查询 vSwitch 内部的由 Agent 模块配置的端口映射表，以确定待交换数据的流出端口（第二虚拟端口 port2）以及相对应的 Host NIC。具体地，端口映射表中存有配置输入端口、输出端口、源地址、目标地址等信息的对应关系。从而 vSwitch 可以根据待交换数据中携带的目标地址和端口等信息可以确定输出端口，从而完成交换处理过程。这里的输入/输出端口信息可以是 vSwitch 的虚拟端口的端口号，源地址/目标地址可以是源节点/目标节点的互联网协议 IP 地址或多媒体访问控制 MAC 地址

405，读取待交换数据。

确定 port2 后，vSwitch 通过 port2 向 Host NIC 发送读取指示信息，该读取

指示信息中可携带待交换数据存入的共享内存的地址，令其读取共享内存中的待交换数据。Host NIC 读取数据完毕后，可以向 Host 外部连接的设备或节点发送待交换数据，并通过 port2 向 vSwitch 发送读完指示信息，以便于 vSwitch 恢复共享内存的可写权限，即释放该共享内存，其中读完指示信息可以为读完中断。

应理解，为了方便描述，本发明实施例中以待交换数据为例来说明虚拟交换的具体过程，事实上，实际的虚拟交换还可以是数据流、信令、消息等，本发明对此不做限定。

由上可见，本发明实施例将虚拟交换功能实现在虚拟机上，使得虚拟交换机与普通 VM 处于同等优先级，形成对等的网络虚拟化架构，在进行资源分配时虚拟交换机和普通 VM 一样使用用户空间的物理资源，这样便于 Host 进行管理和高效合理地进行资源分配。应用于该计算节点上的虚拟交换方法包括：第一虚拟机接收源节点发送的第一消息，第一消息用于请求第一虚拟机对待交换数据进行交换处理，其中待交换数据从源节点发往目标节点，源节点和目标节点中的至少一个为第二虚拟机；第一虚拟机根据待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，第二消息用于指示目标节点从硬件层的存储设备获取待交换数据。该方法将虚拟交换功能从 Host 内核中剥离解耦，转而在虚拟机上实现虚拟交换的功能，简化了 Host 内核的设计和负担，并且由于 VM 具有灵活性和很好的扩展性，从而使得 vSwitch 以及整个虚拟网络的扩展性和灵活性都得到了提高。

图 5 是本发明另一实施例的虚拟交换数据流的示意图。如图 5 所示，虚拟交换机 vSwitch（虚拟交换功能）部署于第一虚拟机上，使该第一虚拟机成为虚拟交换设备，并与普通的虚拟机 VM1、VM2 处于同等地位。其中第一虚拟机中的代理 Agent 模块与主机 Host 中的配置管理模块（Config and Manage Module）连接，以便于系统管理员对第一虚拟机进行配置。第一虚拟机的虚拟端口 port 可以与 VM1、VM2 或者 VMM 的底层物理网卡 HOST NIC 进行连接。以下通过数据流详细说明由 Host 外界（Host NIC）向 Host 中的普通 VM（例如 VM1）发送待交换数据的过程。应理解，图 5 所示的系统架构仅仅为一个示例，其中 VM、port、Host NIC 等模块的数量可以进行扩展。

501，预先配置。

在进行虚拟交换之前，需要构建虚拟网络，并对虚拟交换机 vSwitch（第一虚拟机）进行预先配置。具体可以通过 Host 上的 Config and Manage Module 向第一虚拟机中的 Agent 模块发送配置命令，使得 Agent 模块对 vSwitch 的端口映射、VLAN 管理等进行配置。具体的配置过程和配置项目与上述图 3 中步骤 301 相类似，此处不再赘述。

502，确定共享内存。

Host NIC 接收到从外界（源节点）传入的待交换数据后，查询目标节点（VM1）的地址，并通过 port1 向 vSwitch 发送携带有 VM1 的地址的请求信息，其中 port1 是步骤 501 中 Agent 模块预先配置与 Host NIC 相对应的虚拟端口，之后 vSwitch 驱动层直接访问该待交换数据，查询 vSwitch 内部的由 Agent 模块预先配置的端口映射表，以确定待交换数据的流出端口（第二虚拟端口 port2）以及相对应的共享内存。而后通过 port1 向 Host NIC 发送接待有共享内存地址的回复消息。

503，写入待交换数据。

Host NIC 接收到共享内存地址后，将待交换数据写入共享内存中。写入方式由步骤 501 中 Agent 模块预先配置，例如，通过 DMA 方式写入。Host NIC 写入完毕后，通过 port1 向 vSwitch 发送写完指示信息，以通知 vSwitch 进行下一步操作，其中写完指示信息可以为写完中断。

504，读取待交换数据。

vSwitch 收到写完指示信息后，通过 port2 向 VM1 发送读取指示信息，以通知其新数据到来。VM1 从共享内存中读取待交换数据完毕后，通过 port2 向 vSwitch 发送读完指示信息，以便于 vSwitch 恢复共享内存的可写权限，即释放该共享内存。

应理解，为了方便描述，本发明实施例中以待交换数据为例来说明虚拟交换的具体过程，事实上，实际的虚拟交换还可以是数据流、信令、消息等，本发明对此不做限定。

由上可见，本发明实施例将虚拟交换功能实现在虚拟机上，使得虚拟交换机与普通 VM 处于同等优先级，形成对等的网络虚拟化架构，在进行资源分配时虚拟交换机和普通 VM 一样使用用户空间的物理资源，这样便于 Host 进行管理和高效合理地进行资源分配。应用于该计算节点上的虚拟交换方法包括：第一虚拟机接收源节点发送的第一消息，第一消息用于请求第一虚拟机对待交换数据进行交换处理，其中待交换数据从源节点发往目标节点，源节点和目标节点中的至少一个为第二虚拟机；第一虚拟机根据待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，第二消息用于指示目标节点从硬件层的存储设备获取待交换数据。该方法将虚拟交换功能从 Host 内核中剥离解耦，转而在虚拟机上实现虚拟交换的功能，简化了 Host 内核的设计和负担，并且由于 VM 具有灵活性和很好的扩展性，从而使得 vSwitch 以及整个虚拟网络的扩展性和灵活性都得到了提高。

图 6 是本发明另一实施例的虚拟交换数据流的示意图。如图 6 所示，虚拟交换机 vSwitch（虚拟交换功能）部署于第一虚拟机上，使该第一虚拟机成为虚拟交换设备，并与普通的虚拟机 VM1、VM2 处于同等地位。其中第一虚拟机中的代理 Agent 模块与主机 Host 中的配置管理模块（Config and Manage Module）连接，以便于系统管理员对第一虚拟机进行配置。第一虚拟机的虚拟端口 port 可以与 VM1、VM2 或者 VMM 的底层物理网卡 HOST NIC 进行连接。以下通过数据流详细说明 Host 中普通 VM 之间（VM1 与 VM2）待交换数据发送的过程。应理解，图 6 所示的系统架构仅仅为一个示例，其中 VM、port、Host NIC 等模块的数量可以进行扩展。

601，预先配置。

在进行虚拟交换之前，需要构建虚拟网络，并对虚拟交换机 vSwitch（第一虚拟机）进行预先配置。具体可以通过 Host 上的 Config and Manage Module 向第一虚拟机中的 Agent 模块发送配置命令，使得 Agent 模块对 vSwitch 的端口映射、VLAN 管理等进行配置。具体的配置过程和配置项目与上述图 3 中步骤 301 相类似，此处不再赘述。

602，共享内存协商。

Host 中普通 VM 之间需要通过 vSwitch 协商共享内存以供通信。具体地，

VM1 可以通过 vSwitch 与 VM2 协商，由 vSwitch 创建一个共享内存以供 VM1 和 VM2 共享。具体协商过程可以利用 Xen 事件通道（Event Channel）的机制进行。VM1 与 vSwitch 的第一虚拟端口 port1 建立虚连接，其中 port1 是步骤 601 中 Agent 模块预先配置与 VM1 相对应的虚拟端口。相应的物理过程为，VM1 通过其虚拟网卡 VM NIC 映射到 VM1 与 VM2 协商的共享内存。

603，写入待交换数据。

之后，VM1 通过其 NIC 向 port1 发送待交换数据。相应的实际物理过程为，将待交换数据写入 VM1 对应的共享内存。VM1 写入完毕后，通过 port1 向 vSwitch 发送写完指示信息，以通知 vSwitch 进行下一步操作。

604，读取待交换数据。

vSwitch 向 VM2 发送读取指示信息，令其读取共享内存中的待交换数据。Host NIC 读取数据完毕后，向 Host 外部的目标节点发送待交换数据，并向 vSwitch 发送读完指示信息，以便于 vSwitch 恢复共享内存的可写权限，即释放该共享内存。

应理解，为了方便描述，本发明实施例中以待交换数据为例来说明虚拟交换的具体过程，事实上，实际的虚拟交换还可以是数据流、信令、消息等，本发明对此不做限定。

由上可见，本发明实施例将虚拟交换功能实现在虚拟机上，使得虚拟交换机与普通 VM 处于同等优先级，形成对等的网络虚拟化架构，在进行资源分配时虚拟交换机和普通 VM 一样使用用户空间的物理资源，这样便于 Host 进行管理和高效合理地进行资源分配。应用于该计算节点上的虚拟交换方法包括：第一虚拟机接收源节点发送的第一消息，第一消息用于请求第一虚拟机对待交换数据进行交换处理，其中待交换数据从源节点发往目标节点，源节点和目标节点中的至少一个为第二虚拟机；第一虚拟机根据待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，第二消息用于指示目标节点从硬件层的存储设备获取待交换数据。该方法将虚拟交换功能从 Host 内核中剥离解耦，转而在虚拟机上实现虚拟交换的功能，简化了 Host 内核的设计和负担，并且由于 VM 具有灵活性和很好的扩展性，从而使得 vSwitch

以及整个虚拟网络的扩展性和灵活性都得到了提高。

图 7 是本发明另一实施例的用于软件定义网络 SDN 的虚拟交换设备的示意图。

本发明通过将虚拟交换机 vSwitch 与 Host 内核解耦，并将 vSwitch 部署到第一虚拟机中，简化了 Host 内核的设计和复杂程度。并且由于虚拟机的可配置性、扩展性和灵活性较高，从而也提高了 vSwitch 乃至整个虚拟化网络的扩展性和灵活性，使得本发明实施例的虚拟交换设备可以实现控制面 control plane 和数据面 data plane 的分离，也就是说，满足 SDN 的需求。

SDN 是新一代网络架构，与传统网络架构将协议分层，控制面和数据面相融合的做法不同，SDN 在操作和控制层面将协议融合处理，并将控制面和数据面分开。典型的 SDN 方案为开放流 Openflow，具体到在本发明实施例的具有虚拟交换功能的第一虚拟机上实现 Openflow，可以将交换设备的逻辑实现分为两个部分: 开放流控制器(Openflow Controller)和开放流流表(Openflow Flowtable)，其中开放流控制器负责控制面，用于网络拓扑配置，数据转发策略调整，配置和维护 Openflow 流表，Openflow 流表则负责数据面，是数据流转发的查询映射表。为了满足 SDN 架构对交换设备的需求，本发明可采用如下两种部署方式:

第一种，Openflow Controller 和 Openflow Flowtable 实现在同一个 VM 中，也就是本发明中的具有虚拟交换功能的第一虚拟机，其中 Openflow Controller 实现在用户空间，而 Flowtable 可实现在用户空间，亦可实现在内核空间;

第二种，Openflow Controller 和 Openflow Flowtable 分别实现在两个具有虚拟交换功能的虚拟机中，例如，可以将 Openflow Controller 部署在第一虚拟机中，运行在 Host 之上的至少一个 VM 之中还包括具有虚拟交换功能的第四虚拟机，第四虚拟机与第一虚拟机相类似，两者使用 VM 间的通信技术交互信息，例如 Xen 的事件 Event Channel。

具体地，如图 7 所示，虚拟交换机 vSwitch 的 Controller 和 FlowTable 部署于第一虚拟机上，或者部署于不同的两个虚拟机上，使得该 vSwitch 与普通的虚拟机 VM1、VM2 处于同等地位。其中 Controller 中的代理 Agent 模块与

主机 Host 中的配置管理模块（Config and Manage Module）连接，以便于系统管理员对 vSwitch 进行配置。Flowtable 部分的虚拟端口 port 可以与 VM1、VM2 或者 VMM 的底层物理网卡 HOST NIC 进行连接。应理解，图 7 所示的系统架构仅仅为一个示例，其中 VM、port、Host NIC 等模块的数量可以进行扩展。

Openflow Controller 和 Flowtable 相互配合实现业务流的转发，其中 Controller 包含用户配置数据库和一个规则库，Flowtable 是一个以业务流为单位的表结构，包含匹配和执行和部分。Flowtable 每一个表项 entry 代表一个业务流，匹配部分是待交换数据 IP、MAC 和 Port 等字段，执行部分表示对匹配待交换数据的处理，包括转发、丢包和向 Controller 申请新的表项。例如，每当有待交换数据到达 vSwitch 时，vSwitch 检查待交换数据的 IP，Mac 和 Port 等字段，并搜索 Flowtable，寻找匹配 entry；若找到匹配表项，按照 Action 执行操作；若未找到匹配表项，Flowtable 向 Controller 发送表项建立请求，Controller 收到请求后，查询规则库，建立新的表项，并发给 Flowtable; Flowtable 插入新的表项，并将后续符合此表项的待交换数据按规则转发。

由上可见，本发明实施例将虚拟交换功能实现在虚拟机上，使得虚拟交换机与普通 VM 处于同等优先级，形成对等的网络虚拟化架构，在进行资源分配时虚拟交换机和普通 VM 一样使用用户空间的物理资源，这样便于 Host 进行管理和高效合理地进行资源分配。该方法将虚拟交换功能从 Host 内核中剥离解耦，降低 Host 与 vSwitch 的耦合性，可以在同一 Host 内部署多个 vSwitch 不受 Host 的约束，并且由于 VM 具有灵活性和很好的扩展性，从而使得 vSwitch 以及整个虚拟网络的扩展性和灵活性都得到了提高。本发明还将配置模块与待交换数据交换转发模块相分离，更加地符合可编程网络设计，从而能够在本发明实施例的虚拟化网络架构上实现 SDN。

图 8 是本发明另一实施例的分布式实施的示意图。

如图 8 所示。本发明实施例的配置架构包括一个主虚拟交换机 Master vSwitch 和两个个从属虚拟交换机 Slave vSwitch，应理解的是，图 8 为了方便描述，仅仅示出两个从属 vSwitch，这对本发明并不造成限定，事实上可以有若干个从属 vSwitch。图 8 中的每一个主机 Host 都与上述实施例中描述的运行在硬件层之上的 Host 相同，且这些 Host 可以是运行在同一个物理机的硬件层

之上的 Host，也可以是运行在不同物理机的硬件层之上的 Host，本发明对此不做限定。其中，每一个 vSwitch 均为本发明所涉及的具有虚拟交换功能的虚拟机，也就是说每一个 vSwitch 都与上述实施例中的具有虚拟交换功能的第一虚拟机相类似。各个 Host 中的主管理模块和从属管理模块都可以对应于上述实施例 Host 中的配置管理模块 Config and Manage Module，相应地，Master vSwitch 的控制管理模块被设定为主管理模块 Master Manager，Slave vSwitch 的控制管理模块被设定为从属管理模块 Slave Manager。Master Manager 和 Slave Manager 对其 Host 的 vSwitch 管理方式与上述各实施例的方式相同，可以通过 vSwitch 中的 agent 模块来配置和管理 vSwitch(agent 未在图 8 中示出)。其中 Master Manager 是用户配置的接口，可以由用户通过客户端程序直接进行配置，Master Manager 通过协议与 Slave Manager 通信，协商各个 vSwitch 之间的端口映射，而 Master Manager 与 Slave Manager 之间的通信则为控制流，主从 vSwitch 之间的通信则为数据流。

具体地，本发明实施例的分布式 vSwitch 的配置过程为：首先在一个 Host 上创建 Master vSwitch，之后创建 vSwitch 级联配置，包括各个 Slave vSwitch，以及所有 vSwitch 上的 IP 地址和端口映射；之后通过配置协议将上述配置信息发送到其他 Host，至此，承载 Master vSwitch 的 Host 为主 Host，接收配置信息的其他 Host 为从属 Host；之后，接收到配置信息的各个从属 Host 创建控制管理模块，即从属管理模块；最后，各个从属管理模块按照接收到的配置信息配置器对应的 Slave vSwitch 上的 IP 地址和端口。应理解，本发明实施例所涉及的配置协议，包括但不限于可扩展标记语言 XML、超文本传输协议 HTTP 等应用协议。

作为一个具体的例子，本发明实施例的分布式交换架构的配置过程如图 9 所示：

901，用户登陆 Host0 中的 Manage Module 创建一个 vSwitch 实例，并将其定义为 Master。

902，通过通信协议，将配置消息传输至 Host1 和 Host2 的 Manage Moduel。

903，Host1 和 Host2 的 Manage Module 接收到配置消息，按照配置要求创建 vSwitch 实例，并定义为 Slave，然后将其 Master 指针指向 Host0 的 vSwitch；

根据配置中的端口映射，配置其 vSwitch 的端口映射。

本发明实施例将虚拟交换功能从 Host 内核中剥离解耦，降低 Host 与 vSwitch 的耦合性，可以在同一 Host 内部署多个 vSwitch 不受 Host 的约束，并且由于 vSwitch 在用户操作系统 Guest OS 中实现，无需再依赖内核操作系统 Host OS/VMM OS，使得 vSwitch 非常容易部署，具有良好的移植性，从而使得 vSwitch 以及整个虚拟网络的扩展性和灵活性都得到了提高，本发明实施例中的分布式架构将多个 vSwitch 级联，使得虚拟网络得到大幅扩展以及虚拟交换能力得到大幅提升。

图 10 是本发明一个实施例的宿主机的模块架构示意图。图 10 的宿主机 1000 包括创建模块 1001 和配置模块 1002。

创建模块 1001，用于在输入/输出 I/O 设备的 I/O 虚拟功能启动后，在宿主机 Host 中产生至少一个虚拟机 VM，其中至少一个 VM 包括具有虚拟交换功能的第一虚拟机，至少一个 VM 还包括第二虚拟机；

配置模块 1002，用于向第一虚拟机发送配置命令，以便于第一虚拟机根据配置命令配置用于与第二虚拟机进行通信的第一虚拟机的第一虚拟端口，并配置用于与 I/O 设备进行通信的第一虚拟机的第二虚拟端口。

可以理解的是，本实施例宿主机 1000 可如上述方法实施例中的 Host，其各个功能模块的功能可以根据上述方法实施例中的方法具体实现，其具体实现过程可以参照上述方法实施例的相关描述，此处不再赘述。

由上可见，本实施例中在 I/O 设备的 I/O 虚拟功能启动后，Host1000 通过创建模块 1001 产生至少一个运行在 Host1000 之上的虚拟机。具体地，该创建模块 1001 可以是配置管理模块（Config and Manage Module），创建模块 1001 还可以通过使用 Qemu 等工具创建虚拟机的虚拟网卡接口（VM NIC），由创建模块 1001 产生的虚拟机中有至少一个具有虚拟交换功能的第一虚拟机 vSwitch 以及若干普通虚拟机 VM。

之后配置模块 1002，即 Config and Manage Module，向 Agent 模块发送配置命令，其中配置模块 1002 通过进程间通信技术 IPC（如 IOCTL，NETLINK，SOCKET 等）与 Agent 相连接，配置模块 1002 将 Host 虚拟环境的配置传入第

一虚拟机的 Agent，具体可以包括 Host1000 下层物理网卡、虚拟机的前后端 FE/BE、共享内存、DMA 中断等配置信息，使得第一虚拟机获得虚拟环境信息，从而建立相应的虚拟网络环境。

由上可见，通过 Host1000 搭建的虚拟网络环境，虚拟交换功能得以从 Host1000 内核中剥离解耦，转而在第一虚拟机上实现虚拟交换的功能，简化了 Host 内核的设计和负担，并且由于 VM 具有灵活性和很好的扩展性，从而使得 vSwitch 以及整个虚拟网络的扩展性和灵活性都得到了提高。进一步地，因为将虚拟交换功能实现在虚拟机上，使得虚拟交换机与普通 VM 处于同等地位，具有相同的优先级，形成对等的网络虚拟化架构，在进行资源分配时虚拟交换机和普通 VM 一样使用用户空间的物理资源，这样便于 Host 进行管理和高效合理地进行资源分配。

图 11 是本发明一个实施例的虚拟机的模块架构示意图。图 11 的虚拟机 1100 包括接收模块 1101、交换处理模块 1102 和发送模块 1103。

接收模块 1101，用于接收源节点发送的第一消息，所述第一消息用于请求所述虚拟机 1100 对待交换数据进行交换处理，其中所述待交换数据是从所述源节点发往目标节点的，所述源节点和所述目标节点中的至少一个为第二虚拟机，所述第二虚拟机运行在所述 Host 之上；

交换处理模块 1102，用于根据所述待交换数据携带的目标节点的地址和所述虚拟机 1100 配置的端口映射表确定第二消息，所述第二消息用于指示所述目标节点从所述硬件层的存储设备获取所述待交换数据；

发送模块 1103，用于向所述目标节点发送所述第二消息。

本发明实施例的虚拟机 1100 为具有虚拟交换功能的虚拟机，与其他普通虚拟机具有同等地位，部署于 Host 上。其中源节点可以是 Host 上的普通虚拟机，也可以是 Host 外部的虚拟机或物理机。同样地，目标节点也可以是 Host 上的普通虚拟机，也可以是 Host 外部的虚拟机或物理机。

可以理解的是，本发明实施例的虚拟机 1100 可如上述方法实施例中的具有虚拟交换功能的第一虚拟机，其各个功能模块的功能可以根据上述方法实施例中的方法具体实现，其具体实现过程可以参照上述方法实施例的相关描述，

此处不再赘述。

本发明实施例通过将虚拟交换功能部署到虚拟机中，简化了 VMM，有利于 Host 对虚拟网络进行管理并进行高效、合理的网络资源分配。

可选地，作为一个实施例，虚拟机 1100 还包括代理 Agent 模块 1104 和生成模块 1105。具体地，代理 Agent 模块 1104，用于根据 Host 发送的配置命令，配置用于与第二虚拟机进行通信的虚拟机的第一虚拟端口 1106，并配置用于与 I/O 设备进行通信的虚拟机的第二虚拟端口 1107。生成模块 1105，用于建立第一虚拟端口 1106 与第二虚拟端口 1107 之间的映射关系，以生成端口映射表。

可选地，作为一个实施例，Agent 模块 1104，还用于根据配置命令配置第二虚拟机对应的第一共享内存，其中第一共享内存为硬件层的存储设备上的指定存储区域。具体可以通过第二虚拟机与虚拟机 1100 之间的事件通道协商第一共享内存。接收模块 1101，具体用于通过第一虚拟端口 1106 接收第一消息，第一消息包括用于向虚拟机 1100 指示源节点已完成将待交换数据写入第一共享内存的写完中断；交换处理模块 1102，具体用于根据用于接收第一消息的第一虚拟端口 1106 确定对应的第一共享内存的地址；从第一共享内存获取待交换数据携带的目标节点的地址，以便于确定目标节点所对应的第二虚拟端口 1107；确定携带有第一共享内存的地址和读取指令的第二消息。发送模块 1103，具体用于通过端口映射表中与第一虚拟端口 1106 对应的第二虚拟端口 1107 向目标节点发送第二消息；其中，源节点为第二虚拟机，目标节点为 I/O 设备。

可选地，作为一个实施例，接收模块 1101，具体用于接收源节点发送的第一消息；交换处理模块 1102，具体用于获取待交换数据携带的目标节点的地址；根据目标节点的地址查询端口映射表以确定与目标节点对应的第一虚拟端口 1106 并确定与第二虚拟机对应的第一共享内存的地址；发送模块 1103，具体用于通过 I/O 设备所对应的第二虚拟端口 1107 向目标节点发送携带有第一共享内存的地址的回复消息；交换处理模块 1102，还用于在接收到源节点发送的用于向虚拟机 1100 指示源节点已完成将待交换数据写入第一共享内存的写完中断时，确定携带有读取指令的第二消息；发送模块 1103，还用于通过第一虚拟端口 1106 向目标节点发送第二消息；接收模块 1101，还用于接收

源节点发送的指示源节点已完成将待交换数据写入第一共享内存的写完中断；其中，源节点为 I/O 设备，目标节点为第二虚拟机。

可选地，作为一种实现方式，接收模块 1101，具体用于通过第一虚拟端口 1106 接收源节点发送的第一消息，第一消息包括写完中断；交换处理模块 1102，具体用于根据用于接收第一消息的第一虚拟端口 1106 确定对应的源节点的地址；根据源节点的地址和待交换数据携带的目标节点的地址确定第二共享内存的地址；确定携带有第二共享内存的地址和读取指令的第二消息；发送模块 1103，具体用于向目标节点发送第二消息。

可选地，作为另一个实施例，接收模块 1101 还用于：接收目标节点发送的读完指示信息，以便于虚拟机 1100 释放第一共享内存或第二共享内存。

具体地，第一虚拟机从 I/O 设备获取待交换数据携带的目标节点的地址，是由第一虚拟机在接收到第一消息的通知后，得知 I/O 设备（即底层物理网卡）接收到了待交换数据，之后第一虚拟机则可以通过驱动层直接访问该待交换数据以获取其携带的目标节点的地址。

可选地，在一种实现方式下，在端口映射表为开放流 Openflow 流表时，第一虚拟机 1231 还包括包含 Agent 模块 1104 的 Openflow 控制器，其中：在接收模块 1101 接收源节点发送的第一消息之后，，交换处理模块 1102 还用于，根据待交换数据携带的目标节点的地址，在 Openflow 流表中确定与目标节点的地址所匹配的表项，其中，Openflow 流表中包括至少一个表项，表项包括地址、虚拟端口和执行动作参数；如果匹配的表项存在，根据匹配的表项中与目标节点的地址所对应的执行动作参数处理待交换数据；如果匹配的表项不存在，向 Openflow 控制器发送表项建立请求，以便于 Openflow 控制器根据表项建立请求建立能够与待交换数据匹配的新表项，并在 Openflow 流表中插入新表项。

本发明实施例通过将虚拟交换功能部署到虚拟机 1100 中，使得具有虚拟交换功能的虚拟机 1100 与其他普通虚拟机处于同等地位，从而有利于 Host 对虚拟网络进行管理并进行高效、合理的网络资源分配。并且由于虚拟交换功能从 Host 核心中剥离，从而增强了扩展性，使其虚拟机 1100 满足 SDN 的需求，支持 Openflow。

图 12 是本发明一个实施例的计算机节点的示意图。图 12 所示的计算节点 1200 可包括:

硬件层 1210、运行在硬件层 1210 之上的宿主机 Host 1220、以及运行在 Host 1220 之上的至少一个虚拟机 VM1230;

其中，硬件层 1210 包括输入/输出 I/O 设备 1211 和存储设备 1212，至少一个虚拟机 VM1230 包括具有虚拟交换功能的第一虚拟机 1231，至少一个 VM1230 还包括第二虚拟机 1232。

第一虚拟机 1231，用于接收源节点发送的第一消息，第一消息用于请求第一虚拟机 1231 对待交换数据进行交换处理，其中待交换数据是从源节点发往目标节点的，源节点和目标节点中的至少一个为第二虚拟机 1232;

第一虚拟机 1231，还用于根据待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，第二消息用于指示目标节点从硬件层的存储设备 1212 获取待交换数据。

此外，Host1220，用于向第一虚拟机 1231 发送配置命令;

第一虚拟机 1231，还用于根据配置命令，通过第一虚拟机的代理 Agent 模块配置用于与第二虚拟机进行通信的第一虚拟机的第一虚拟端口，并配置用于与 I/O 设备 1211 进行通信的第一虚拟机的第二虚拟端口;

第一虚拟机 1231，还用于建立第一虚拟端口与第二虚拟端口之间的映射关系，以生成端口映射表。

可选地，第一虚拟机 1231，还用于根据配置命令配置第二虚拟机 1232 对应的第一共享内存，其中第一共享内存为硬件层 1210 的存储设备 1212 上的指定存储区域。

具体地，作为一个数据流和信令流交互的例子，当源节点为第二虚拟机 1232，目标节点为 I/O 设备 1211 时:

源节点 1232，用于将待交换数据写入第一共享内存;

源节点 1232，还用于向第一虚拟机 1231 发送第一消息;

第一虚拟机 1231，具体用于通过第一虚拟端口接收第一消息，第一消息

包括用于向第一虚拟机1231指示源节点1232已完成将待交换数据写入第一共享内存的写完中断；以及根据用于接收第一消息的第一虚拟端口确定对应的第一共享内存的地址；从第一共享内存获取待交换数据携带的目标节点1211的地址，以便于确定目标节点1211所对应的第二虚拟端口；确定携带有第一共享内存的地址和读取指令的第二消息，并通过端口映射表中与第一虚拟端口对应的第二虚拟端口向目标节点1211发送第二消息；

目标节点1211，用于根据第二消息从第一共享内存读取待交换数据；

具体地，作为一个数据流和信令流交互的例子，当源节点为I/O设备1211，目标节点为第二虚拟机1232时：

第一虚拟机1231，具体用于接收源节点1211发送的第一消息，获取待交换数据携带的目标节点1232的地址；根据目标节点1232的地址查询端口映射表以确定与目标节点1232对应的第一虚拟端口并确定与第二虚拟机1232对应的第一共享内存的地址；通过I/O设备1211所对应的第二虚拟端口向目标节点1232发送携带有第一共享内存的地址的回复消息；以及，在接收到源节点1211发送的用于向第一虚拟机指示源节点1211已完成将待交换数据写入第一共享内存的写完中断时，确定携带有读取指令的第二消息，通过第一虚拟端口向目标节点1232发送第二消息；

源节点1211，还用于根据回复消息中的第一共享内存的地址将待交换数据写入第一共享内存；

源节点1211，还用于向第一虚拟机发送指示源节点1211已完成将待交换数据写入第一共享内存的写完中断；

目标节点1232，用于根据第二消息从第一共享内存读取待交换数据；

具体地，作为一个数据流和信令流交互的例子，当源节点和目标节点同为至少一个VM1230中的普通虚拟机时，假设源节点为第二虚拟机1232，目标节点为第三虚拟机1233：

源节点1232，还用于将待交换数据写入源节点1232与目标节点1233通过第一虚拟机1231预先协商的第二共享内存，其中第二共享内存为硬件层1210的存储设备1212上的指定存储区域；

源节点 1232，还用于通过第一虚拟端口向第一虚拟机发送第一消息，第一消息包括写完中断；

第一虚拟机 1231，具体用于根据用于接收第一消息的第一虚拟端口确定对应的源节点 1232 的地址；根据源节点 1232 的地址和待交换数据携带的目标节点 1233 的地址确定第二共享内存的地址；确定携带有第二共享内存的地址和读取指令的第二消息，并向目标节点 1233 发送第二消息；

目标节点 1233，用于根据第二消息从第二共享内存读取待交换数据。

可选地，作为一个实施例，在目标节点根据第二消息从共享内存读取待交换数据之后，目标节点可以向第一虚拟机 1231 发送读完指示信息，以便于第一共享内存或第二共享内存被释放；第一虚拟机 1231，在接收到该读完指示信息后，释放第一共享内存或第二共享内存。

具体地，第一虚拟机从 I/O 设备获取待交换数据携带的目标节点的地址，是由第一虚拟机在接收到第一消息的通知后，得知 I/O 设备（即底层物理网卡）接收到了待交换数据，之后第一虚拟机则可以通过驱动层直接访问该待交换数据以获取其携带的目标节点的地址。

可选地，在端口映射表为开放流 Openflow 流表时，在接收源节点发送的第一消息之后，第一虚拟机 1231 还用于：根据待交换数据携带的目标节点的地址，配置的 Openflow 流表中确定与目标节点的地址所匹配的表项，其中，Openflow 流表中包括至少一个表项，表项包括地址、虚拟端口和执行动作参数；如果匹配的表项存在，根据匹配的表项中与目标节点的地址所对应的执行动作参数处理待交换数据；如果匹配的表项不存在，建立能够与待交换数据匹配的新表项，并在该 Openflow 流表中插入新表项。

综上，本发明实施例中计算节点 1200 可包括：硬件层 1210、运行在所述硬件层 1210 之上的宿主机 Host1220、以及运行在所述 Host1220 之上的至少一个虚拟机 VM1230，其中，所述硬件层包括输入/输出 I/O 设备 1211 和存储设备 1212，所述至少一个虚拟机 VM 包括具有虚拟交换功能的第一虚拟机 1231，所述至少一个 VM 还包括第二虚拟机 1232；如此，将虚拟交换功能实现在虚拟机上，使得虚拟交换机与普通 VM 处于同等优先级，形成对等的网络虚拟化

架构，在进行资源分配时虚拟交换机和普通 VM 一样使用用户空间的物理资源，这样便于 Host 进行管理和高效合理地进行带宽、CPU、存储等资源的分配。应用于该计算节点上的虚拟交换方法包括：所述第一虚拟机接收源节点发送的第一消息，所述第一消息用于请求所述第一虚拟机对待交换数据进行交换处理，其中所述待交换数据从所述源节点发往目标节点，所述源节点和所述目标节点中的至少一个为所述第二虚拟机；所述第一虚拟机根据所述待交换数据携带的目标节点的地址和所述配置的端口映射表确定第二消息并发送所述第二消息，所述第二消息用于指示所述目标节点从所述硬件层的存储设备获取所述待交换数据。该方法将虚拟交换功能从 Host 内核中剥离解耦，降低与 Host 的耦合性，可以在同一 Host 内部署多个 vSwitch，不受 Host 约束，因此具有更强的扩展性，并且解耦后 vSwtich 不再依赖 Host 内核中的操作系统，变得更加易于部署，所以获得了更好的移植性，并且由于配置模块（Agent）与待交换数据交换转发模块（端口映射表）相分离，更符合软件定义网络的要求。

图 13 是本发明一个实施例的计算机系统的示意图。参见图 13，本发明实施例还提供一种计算机系统 1300，可包括：

至少一个计算节点 1200。

需要说明的是，对于前述的各方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本发明并不受所描述的动作顺序的限制，因为依据本发明，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本发明所必须的。

在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其他实施例的相关描述。

综上，本发明实施例的计算机系统 1300 中的计算节点 1200 可包括：硬件层、运行在所述硬件层之上的宿主机 Host、以及运行在所述 Host 之上的至少一个虚拟机 VM，其中，所述硬件层包括输入/输出 I/O 设备和存储设备，所述至少一个虚拟机 VM 包括具有虚拟交换功能的第一虚拟机，所述至少一个 VM 还包括第二虚拟机；如此，将虚拟交换功能实现在虚拟机上，使得虚拟交换机与普通 VM 处于同等优先级，形成对等的网络虚拟化架构，在进行资源分配时

虚拟交换机和普通 VM 一样使用用户空间的物理资源，这样便于 Host 进行管理和高效合理地进行带宽、CPU、存储等资源的分配。应用于该计算节点上的虚拟交换方法包括：所述第一虚拟机接收源节点发送的第一消息，所述第一消息用于请求所述第一虚拟机对待交换数据进行交换处理，其中所述待交换数据从所述源节点发往目标节点，所述源节点和所述目标节点中的至少一个为所述第二虚拟机；所述第一虚拟机根据所述待交换数据携带的目标节点的地址和所述配置的端口映射表确定第二消息并发送所述第二消息，所述第二消息用于指示所述目标节点从所述硬件层的存储设备获取所述待交换数据。该方法将虚拟交换功能从 Host 内核中剥离解耦，降低与 Host 的耦合性，可以在同一 Host 内部署多个 vSwitch，不受 Host 约束，因此具有更强的扩展性，并且解耦后 vSwtich 不再依赖 Host 内核中的操作系统，变得更加易于部署，所以获得了更好的移植性，并且由于配置模块（Agent）与待交换数据交换转发模块（端口映射表）相分离，更符合软件定义网络的要求。

本领域普通技术人员可以意识到，结合本文中所公开的实施例中描述的各方法步骤和单元，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各实施例的步骤及组成。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。本领域普通技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本发明的范围。

结合本文中所公开的实施例描述的方法或步骤可以用硬件、处理器执行的软件程序，或者二者的结合来实施。软件程序可以置于随机存储器（RAM）、内存、只读存储器（ROM）、电可编程 ROM、电可擦除可编程 ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的存储介质中。

尽管通过参考附图并结合优选实施例的方式对本发明进行了详细描述，但本发明并不限于此。在不脱离本发明的精神和实质的前提下，本领域普通技术人员可以对本发明的实施例进行各种等效的修改或替换，而这些修改或替换都应在本发明的涵盖范围内。

权 利 要 求

1. 一种虚拟交换方法，其特征在于，应用于计算节点上，所述计算节点包括：硬件层、运行在所述硬件层之上的宿主机 Host、以及运行在所述 Host 之上的至少一个虚拟机 VM，其中，所述硬件层包括输入/输出 I/O 设备和存储设备，所述至少一个虚拟机 VM 包括具有虚拟交换功能的第一虚拟机，所述至少一个 VM 还包括第二虚拟机：

所述方法包括：

所述第一虚拟机接收源节点发送的第一消息，所述第一消息用于请求所述第一虚拟机对待交换数据进行交换处理，其中所述待交换数据是从所述源节点发往目标节点的，所述源节点和所述目标节点中的至少一个为所述第二虚拟机；

所述第一虚拟机根据所述待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，所述第二消息用于指示所述目标节点从所述硬件层的存储设备获取所述待交换数据。

2. 根据权利要求 1 所述的方法，其特征在于，所述第一虚拟机接收源节点发送的第一消息之前，还包括：

所述第一虚拟机接收所述 Host 发送的配置命令；

所述第一虚拟机根据所述配置命令配置用于与所述第二虚拟机进行通信的所述第一虚拟机的第一虚拟端口，并配置用于与所述 I/O 设备进行通信的所述第一虚拟机的第二虚拟端口；

所述第一虚拟机建立所述第一虚拟端口与所述第二虚拟端口之间的映射关系，以生成所述端口映射表。

3. 根据权利要求 2 所述的方法，其特征在于，所述接收所述 Host 发送的配置命令之后，还包括：所述第一虚拟机根据所述配置命令配置所述第二虚拟机对应的第一共享内存，其中所述第一共享内存为所述硬件层的存储设备上的指定存储区域。

4. 根据权利要求 3 所述的方法，其特征在于，当所述源节点为所述第二虚拟机，所述目标节点为所述 I/O 设备时，

所述第一虚拟机接收源节点发送的第一消息，包括：所述第一虚拟机通过所述第一虚拟端口接收所述第二虚拟机发送的所述第一消息，所述第一消息包

括用于向所述第一虚拟机指示所述第二虚拟机已完成将所述待交换数据写入所述第一共享内存的写完中断;

所述第一虚拟机根据所述待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，包括: 所述第一虚拟机根据用于接收所述第一消息的所述第一虚拟端口确定对应的所述第一共享内存的地址; 从所述第一共享内存获取所述待交换数据，根据所述待交换数据携带的所述I/O 设备的地址从所述端口映射表中确定与所述 I/O 设备对应的所述第二虚拟端口; 确定携带有所述第一共享内存的地址和读取指令的所述第二消息，并通过所述第二虚拟端口向所述 I/O 设备发送所述第二消息，以便于所述 I/O 设备从所述第一共享内存读取所述待交换数据。

5. 根据权利要求 3 所述的方法，其特征在于，当所述源节点为所述 I/O 设备，所述目标节点为所述第二虚拟机时，

所述第一虚拟机接收源节点发送的第一消息之后还包括: 所述第一虚拟机从所述 I/O 设备获取所述待交换数据携带的目标节点的地址，所述目标节点的地址为所述第二虚拟机的地址;

所述第一虚拟机根据所述待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，包括: 所述第一虚拟机根据所述第二虚拟机的地址查询所述端口映射表以确定与所述第二虚拟机对应的第一虚拟端口并确定与所述第二虚拟机对应的第一共享内存的地址; 通过所述 I/O 设备所对应的所述第二虚拟端口向所述I/O设备发送携带有所述第一共享内存的地址的回复消息，以便于所述 I/O 设备根据所述回复消息将所述待交换数据写入所述第一共享内存; 在所述第一虚拟机接收到所述 I/O 设备发送的用于向所述第一虚拟机指示所述I/O设备已完成将所述待交换数据写入所述第一共享内存的写完中断时，确定携带有读取指令的所述第二消息，通过所述第一虚拟端口向所述第二虚拟机发送所述第二消息，以便于所述第二虚拟机从所述第一共享内存读取所述待交换数据。

6. 根据权利要求 2 所述的方法，其特征在于，所述至少一个 VM 还包括第三虚拟机，当所述源节点为所述第二虚拟机，所述目标节点为所述第三虚拟机时，

所述第一虚拟机接收源节点发送的第一消息，包括: 所述第一虚拟机通过

所述第一虚拟端口接收所述第二虚拟机发送的所述第一消息，所述第一消息包括用于向所述第一虚拟机指示所述第二虚拟机已完成将所述待交换数据写入所述第二虚拟机与所述第三虚拟机通过所述第一虚拟机预先协商的第二共享内存的写完中断，其中所述第二共享内存为所述硬件层的存储设备上的指定存储区域；

所述第一虚拟机根据所述待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，包括：所述第一虚拟机根据用于接收所述第一消息的所述第一虚拟端口确定与所述第一虚拟端口对应的所述第二虚拟机的地址；根据所述第二虚拟机的地址和所述待交换数据携带的第三虚拟机的地址确定所述第二共享内存的地址；确定携带有所述第二共享内存的地址和读取指令的所述第二消息，并向所述第三虚拟机发送所述第二消息，以便于所述第三虚拟机从所述第二共享内存读取所述待交换数据。

7. 根据权利要求 4 至 6 中任意一项所述的方法，其特征在于，所述方法还包括：接收所述目标节点发送的读完指示信息，以便于所述第一共享内存或所述第二共享内存被释放。

8. 根据权利要求 1 至 7 中任意一项所述的方法，其特征在于，在所述端口映射表为开放流 Openflow 流表时，所述第一虚拟机接收源节点发送的第一消息之后，还包括：

所述第一虚拟机根据所述待交换数据携带的目标节点的地址，在所述 Openflow 流表中确定与所述目标节点的地址所匹配的表项，其中，所述 Openflow 流表中包括至少一个表项，所述表项包括地址、虚拟端口和执行动作参数；

如果所述匹配的表项存在，所述第一虚拟机根据所述匹配的表项中与所述目标节点的地址所对应的执行动作参数处理所述待交换数据；

如果所述匹配的表项不存在，所述第一虚拟机建立能够与所述待交换数据匹配的新表项，并在所述 Openflow 流表中插入所述新表项。

9. 一种宿主机，其特征在于，包括：

创建模块，用于在输入/输出 I/O 设备的 I/O 虚拟功能启动后，在宿主机 Host 之上产生至少一个虚拟机 VM，其中所述至少一个 VM 包括具有虚拟交换功能的第一虚拟机，所述至少一个 VM 还包括第二虚拟机；

配置模块，用于向所述第一虚拟机发送配置命令，以便于所述第一虚拟机根据所述配置命令配置用于与所述第二虚拟机进行通信的所述第一虚拟机的第一虚拟端口，并配置用于与所述 I/O 设备进行通信的所述第一虚拟机的第二虚拟端口。

10. 一种虚拟机，其特征在于，运行在宿主机 Host 之上，所述 Host 运行在硬件层之上，所述硬件层包括输入/输出 I/O 设备和存储设备，所述虚拟机包括:

接收模块，用于接收源节点发送的第一消息，所述第一消息用于请求所述虚拟机对待交换数据进行交换处理，其中所述待交换数据是从所述源节点发往目标节点的，所述源节点和所述目标节点中的至少一个为第二虚拟机，所述第二虚拟机运行在所述 Host 之上;

交换处理模块，用于根据所述待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息，所述第二消息用于指示所述目标节点从所述硬件层的存储设备获取所述待交换数据;

发送模块，用于向所述目标节点发送所述第二消息。

11. 根据权利要求 10 所述的虚拟机，其特征在于，包括:

代理 Agent 模块，用于根据所述 Host 发送的配置命令，配置用于与所述第二虚拟机进行通信的所述虚拟机的第一虚拟端口，并配置用于与所述 I/O 设备进行通信的所述虚拟机的第二虚拟端口;

生成模块，用于建立所述第一虚拟端口与所述第二虚拟端口之间的映射关系，以生成所述端口映射表。

12. 根据权利要求 11 所述的虚拟机，其特征在于，所述 Agent 模块，还用于根据所述配置命令配置所述第二虚拟机对应的第一共享内存，其中所述第一共享内存为所述硬件层的存储设备上的指定存储区域。

13. 根据权利要求 12 所述的虚拟机，其特征在于，

所述接收模块，具体用于通过所述第一虚拟端口接收所述第一消息，所述第一消息包括用于向所述虚拟机指示所述源节点已完成将所述待交换数据写入所述第一共享内存的写完中断;

所述交换处理模块，具体用于根据用于接收所述第一消息的所述第一虚拟端口确定对应的所述第一共享内存的地址; 从所述第一共享内存获取所述待交

换数据，根据所述待交换数据携带的所述目标节点的地址从所述端口映射表中确定与所述目标节点对应的所述第二虚拟端口；确定携带有所述第一共享内存的地址和读取指令的所述第二消息；

所述发送模块，具体用于通过所述第二虚拟端口向所述目标节点发送所述第二消息；

其中，所述源节点为所述第二虚拟机，所述目标节点为所述 I/O 设备。

14. 根据权利要求 12 所述的虚拟机，其特征在于，

所述接收模块，具体用于接收源节点发送的所述第一消息；

所述交换处理模块，具体用于获取所述待交换数据携带的目标节点的地址；根据所述目标节点的地址查询所述端口映射表以确定与所述目标节点对应的第一虚拟端口并确定与所述目标节点对应的第一共享内存的地址；

所述发送模块，具体用于通过所述源节点所对应的所述第二虚拟端口向所述源节点发送携带有所述第一共享内存的地址的回复消息；

所述交换处理模块，还用于在接收到所述源节点发送的用于向所述虚拟机指示所述源节点已完成将所述待交换数据写入所述第一共享内存的写完中断时，确定携带有读取指令的所述第二消息；

所述发送模块，还用于通过所述第一虚拟端口向所述目标节点发送所述第二消息；

所述接收模块，还用于接收所述源节点发送的指示所述源节点已完成将所述待交换数据写入所述第一共享内存的写完中断；

其中，所述源节点为所述 I/O 设备，所述目标节点为所述第二虚拟机。

15. 根据权利要求 11 所述的虚拟机，其特征在于，

所述接收模块，具体用于通过所述第一虚拟端口接收所述源节点发送的所述第一消息，所述第一消息包括写完中断；

所述交换处理模块，具体用于根据用于接收所述第一消息的所述第一虚拟端口确定所述第一虚拟端口对应的所述源节点的地址；根据所述源节点的地址和所述待交换数据携带的目标节点的地址确定所述第二共享内存的地址；确定携带有所述第二共享内存的地址和读取指令的所述第二消息；

所述发送模块，具体用于向所述目标节点发送所述第二消息；

其中，所述至少一个 VM 还包括第三虚拟机，所述源节点为所述第二虚拟

机，所述目标节点为所述第三虚拟机。

16. 一种计算节点，其特征在于，包括：硬件层、运行在所述硬件层之上的宿主机 Host、以及运行在所述 Host 之上的至少一个虚拟机 VM，其中，所述硬件层包括输入/输出 I/O 设备和存储设备，所述至少一个虚拟机 VM 包括具有虚拟交换功能的第一虚拟机，所述至少一个 VM 还包括第二虚拟机，其中：

所述第一虚拟机，用于接收源节点发送的第一消息，所述第一消息用于请求所述第一虚拟机对待交换数据进行交换处理，其中所述待交换数据是从所述源节点发往目标节点的，所述源节点和所述目标节点中的至少一个为所述第二虚拟机；

所述第一虚拟机，还用于根据所述待交换数据携带的目标节点的地址和配置的端口映射表确定第二消息并发送所述第二消息，所述第二消息用于指示所述目标节点从所述硬件层的存储设备获取所述待交换数据。

17. 根据权利要求 16 所述的计算节点，其特征在于，

所述 Host，用于向所述第一虚拟机发送配置命令；

所述第一虚拟机，还用于根据所述配置命令配置用于与所述第二虚拟机进行通信的所述第一虚拟机的第一虚拟端口，并配置用于与所述 I/O 设备进行通信的所述第一虚拟机的第二虚拟端口；

所述第一虚拟机，还用于建立所述第一虚拟端口与所述第二虚拟端口之间的映射关系，以生成所述端口映射表。

18. 根据权利要求 17 所述的计算节点，其特征在于，

所述第一虚拟机，还用于根据所述配置命令配置所述第二虚拟机对应的第一共享内存，其中所述第一共享内存为所述硬件层的存储设备上的指定存储区域。

19. 根据权利要求 18 所述的计算节点，其特征在于，

所述源节点，用于将所述待交换数据写入所述第一共享内存；

所述源节点，还用于向所述第一虚拟机发送所述第一消息；

所述第一虚拟机，具体用于通过所述第一虚拟端口接收所述第一消息，所述第一消息包括用于向所述第一虚拟机指示所述源节点已完成将所述待交换数据写入所述第一共享内存的写完中断；以及根据用于接收所述第一消息的所述第一虚拟端口确定对应的所述第一共享内存的地址；从所述第一共享内存获

取所述待交换数据，根据所述待交换数据携带的所述 I/O 设备的地址从所述端口映射表中确定与所述 I/O 设备对应的所述第二虚拟端口；确定携带有所述第一共享内存的地址和读取指令的所述第二消息，并通过所述第二虚拟端口向所述目标节点发送所述第二消息；

所述目标节点，用于根据所述第二消息从所述第一共享内存读取所述待交换数据；

其中，所述源节点为所述第二虚拟机，所述目标节点为所述 I/O 设备。

20. 根据权利要求 18 所述的计算节点，其特征在于，

所述第一虚拟机，具体用于接收源节点发送的所述第一消息，获取所述待交换数据携带的目标节点的地址；根据所述目标节点的地址查询所述端口映射表以确定与所述目标节点对应的第一虚拟端口并确定与所述目标节点对应的第一共享内存的地址；通过所述源节点所对应的所述第二虚拟端口向所述源节点发送携带有所述第一共享内存的地址的回复消息；以及，在接收到所述源节点发送的用于向所述第一虚拟机指示所述源节点已完成将所述待交换数据写入所述第一共享内存的写完中断时，确定携带有读取指令的所述第二消息，通过所述第一虚拟端口向所述目标节点发送所述第二消息；

所述源节点，还用于根据所述回复消息中的所述第一共享内存的地址将所述待交换数据写入所述第一共享内存；

所述源节点，还用于向所述第一虚拟机发送指示所述源节点已完成将所述待交换数据写入所述第一共享内存的写完中断；

所述目标节点，用于根据所述第二消息从所述第一共享内存读取所述待交换数据；

其中，所述源节点为所述 I/O 设备，所述目标节点为所述第二虚拟机。

21. 根据权利要求 17 所述的计算节点，其特征在于，

所述源节点，还用于将所述待交换数据写入所述源节点与所述目标节点通过所述第一虚拟机预先协商的第二共享内存，其中所述第二共享内存为所述硬件层的存储设备上的指定存储区域；

所述源节点，还用于通过所述第一虚拟端口向所述第一虚拟机发送所述第一消息，所述第一消息包括写完中断；

所述第一虚拟机，具体用于根据用于接收所述第一消息的所述第一虚拟端

口确定所述第一虚拟端口对应的所述源节点的地址；根据所述源节点的地址和所述待交换数据携带的目标节点的地址确定所述第二共享内存的地址；确定携带有所述第二共享内存的地址和读取指令的所述第二消息，并向所述目标节点发送所述第二消息；

所述目标节点，用于根据所述第二消息从所述第二共享内存读取所述待交换数据；

其中，所述至少一个 VM 还包括第三虚拟机，所述源节点为所述第二虚拟机，所述目标节点为所述第三虚拟机。

22. 根据权利要求 19 至 21 中任意一项所述的计算节点，其特征在于，所述目标节点根据所述第二消息从所述共享内存读取所述待交换数据之后，

所述目标节点，还用于向所述第一虚拟机发送读完指示信息，以便于所述第一共享内存或所述第二共享内存被释放；

所述第一虚拟机，还用于释放所述第一共享内存或所述第二共享内存。

23. 根据权利要求 16 至 22 中任意一项所述的计算节点，其特征在于，在所述端口映射表为开放流 Openflow 流表时，在接收源节点发送的第一消息之后，

所述第一虚拟机，还用于根据所述待交换数据携带的目标节点的地址，在所述 Openflow 流表中确定与所述目标节点的地址所匹配的表项，其中，所述 Openflow 流表中包括至少一个表项，所述表项包括地址、虚拟端口和执行动作参数；

如果所述匹配的表项存在，根据所述匹配的表项中与所述目标节点的地址所对应的执行动作参数处理所述待交换数据；

如果所述匹配的表项不存在，建立能够与所述待交换数据匹配的新表项，并在所述 Openflow 流表中插入所述新表项。

24. 一种计算机系统，其特征在于，包括：至少一个如权利要求 16 至 23 任意一项所述的计算节点。

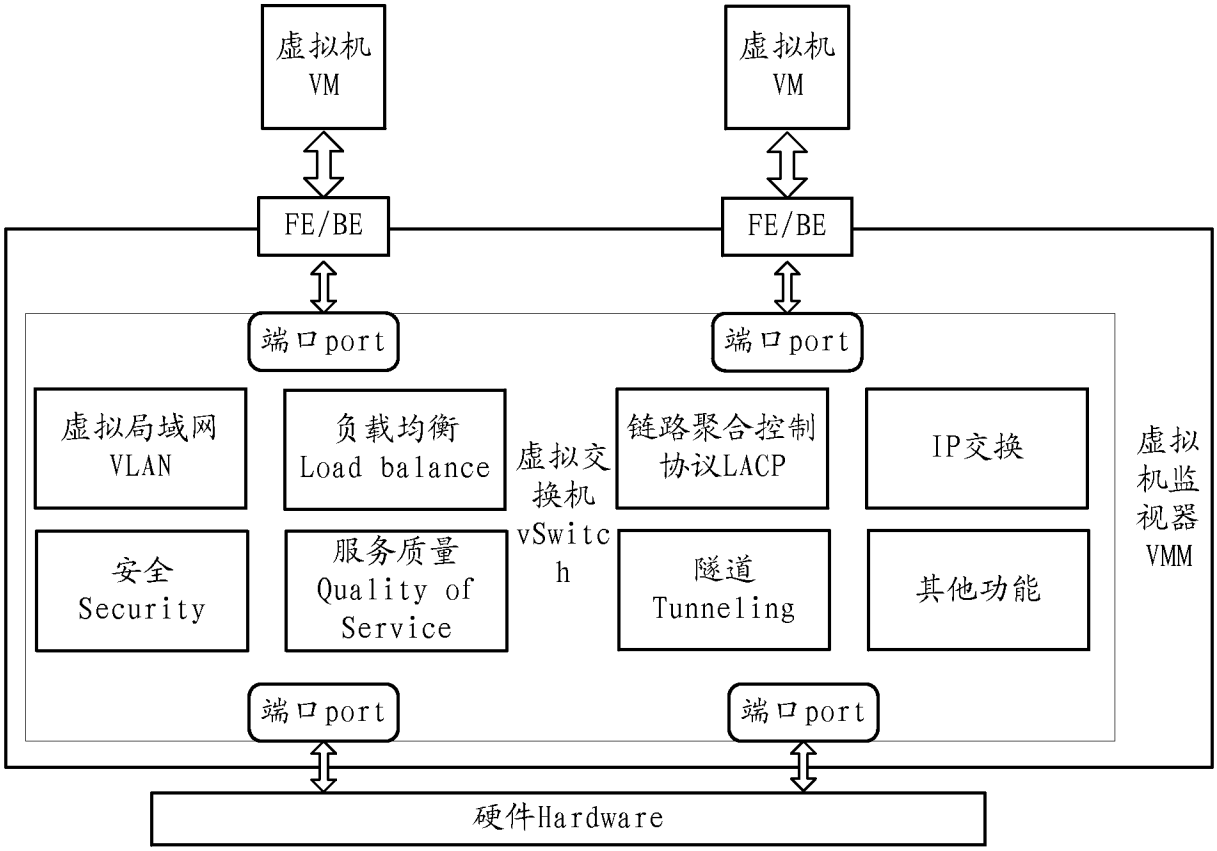

虚拟机
VM
虚拟机
VM
FE/BE
FE/BE
端口port
端口port
虚拟局域网
VLAN
负载均衡
Load balance
虚拟交换机
vSwitch
链路聚合控制
协议LACP
IP交换
虚拟机监视器
VMM
安全
Security
服务质量
Quality of Service
隧道
Tunneling
其他功能
端口port
端口port
硬件Hardware

图1

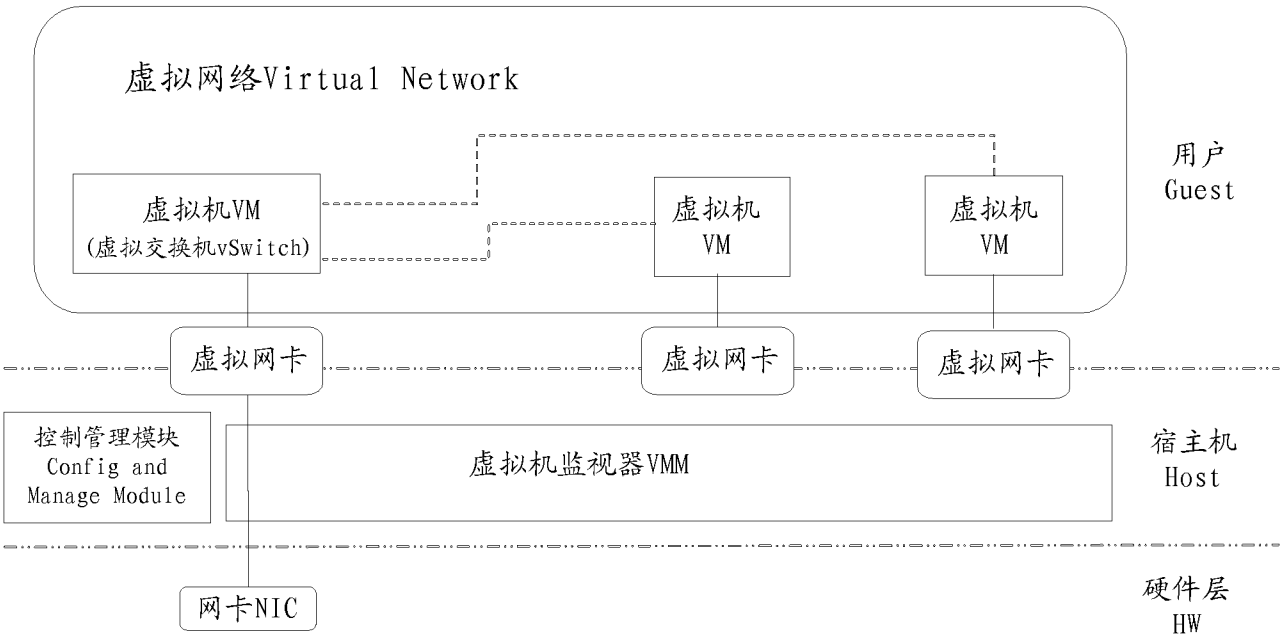

虚拟网络Virtual Network
用户
Guest
虚拟机VM
(虚拟交换机vSwitch)
虚拟机
VM
虚拟机
VM
虚拟网卡
虚拟网卡
虚拟网卡
控制管理模块
Config and
Manage Module
虚拟机监视器VMM
宿主机
Host
网卡NIC
硬件层
HW

图2

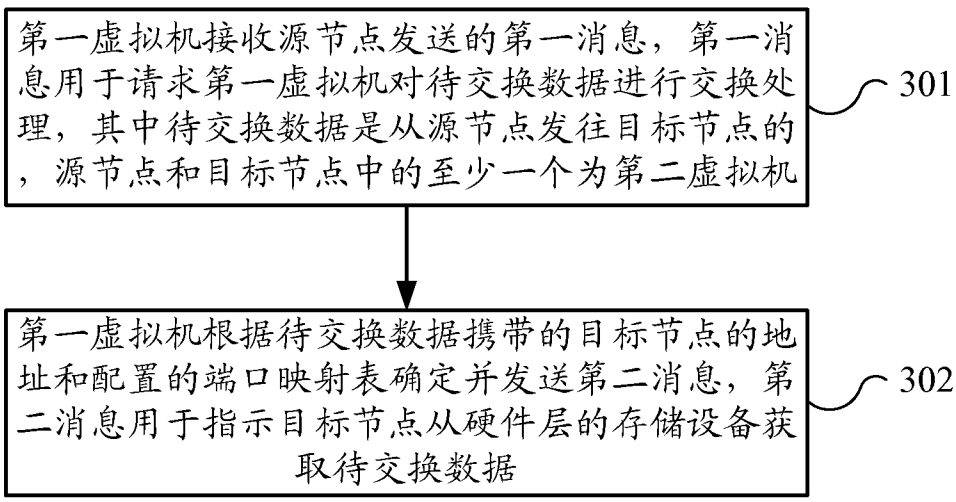

第一虚拟机接收源节点发送的第一消息，第一消息用于请求第一虚拟机对待交换数据进行交换处理，其中待交换数据是从源节点发往目标节点的，源节点和目标节点中的至少一个为第二虚拟机
301
第一虚拟机根据待交换数据携带的目标节点的地址和配置的端口映射表确定并发送第二消息，第二消息用于指示目标节点从硬件层的存储设备获取待交换数据
302

图3

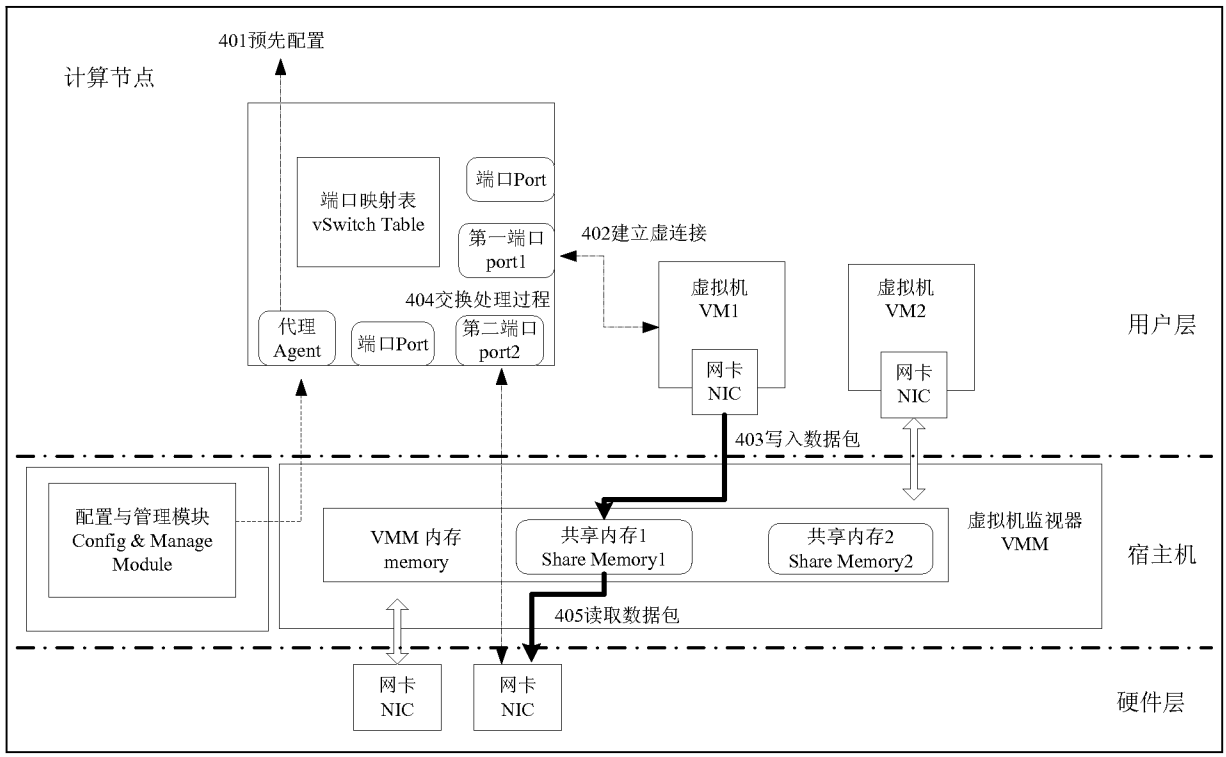

401预先配置
计算节点
端口映射表
vSwitch Table
端口Port
第一端口
port1
402建立虚连接
404交换处理过程
代理
Agent
端口Port
第二端口
port2
虚拟机
VM1
网卡
NIC
虚拟机
VM2
网卡
NIC
用户层
403写入数据包
配置与管理模块
Config & Manage
Module
VMM 内存
memory
共享内存1
Share Memory1
共享内存2
Share Memory2
虚拟机监视器
VMM
宿主机
405读取数据包
网卡
NIC
网卡
NIC
硬件层

图4

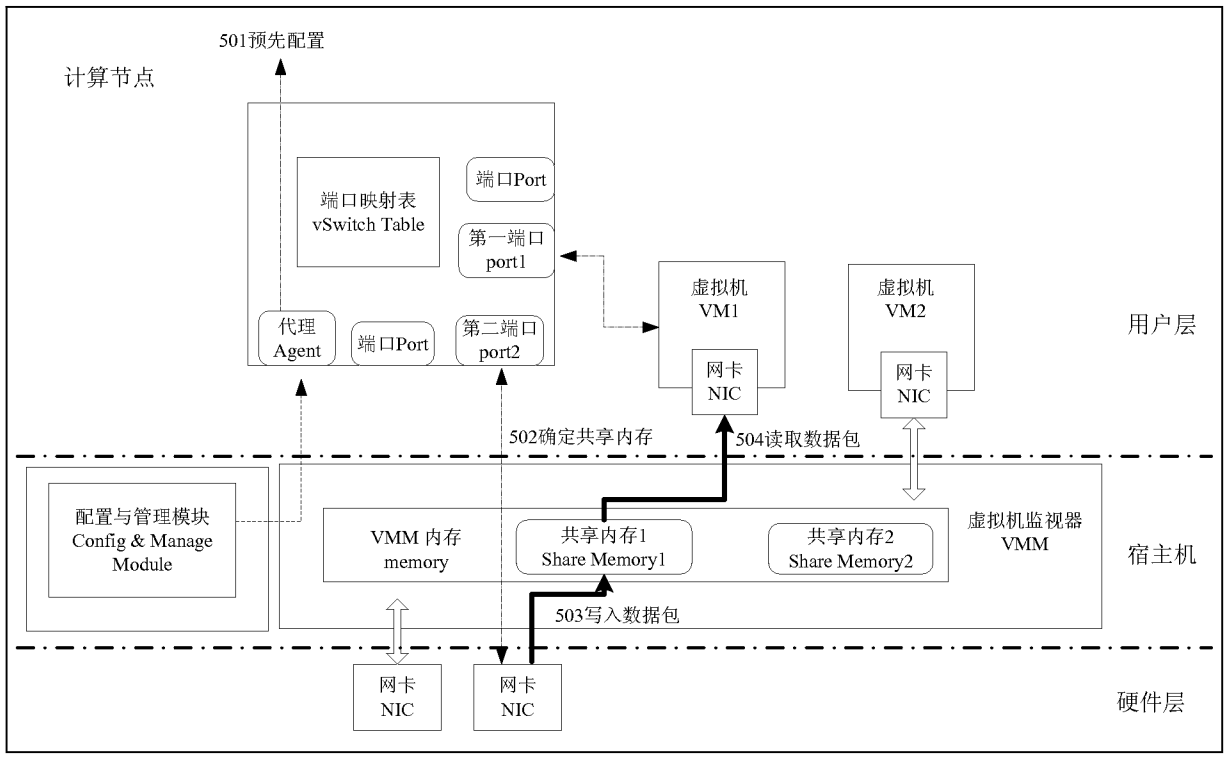

501预先配置
计算节点
端口映射表
vSwitch Table
端口Port
第一端口
port1
代理
Agent
端口Port
第二端口
port2
虚拟机
VM1
网卡
NIC
虚拟机
VM2
网卡
NIC
用户层
502确定共享内存
504读取数据包
配置与管理模块
Config & Manage
Module
VMM 内存
memory
共享内存1
Share Memory1
共享内存2
Share Memory2
虚拟机监视器
VMM
宿主机
503写入数据包
网卡
NIC
网卡
NIC
硬件层

图5

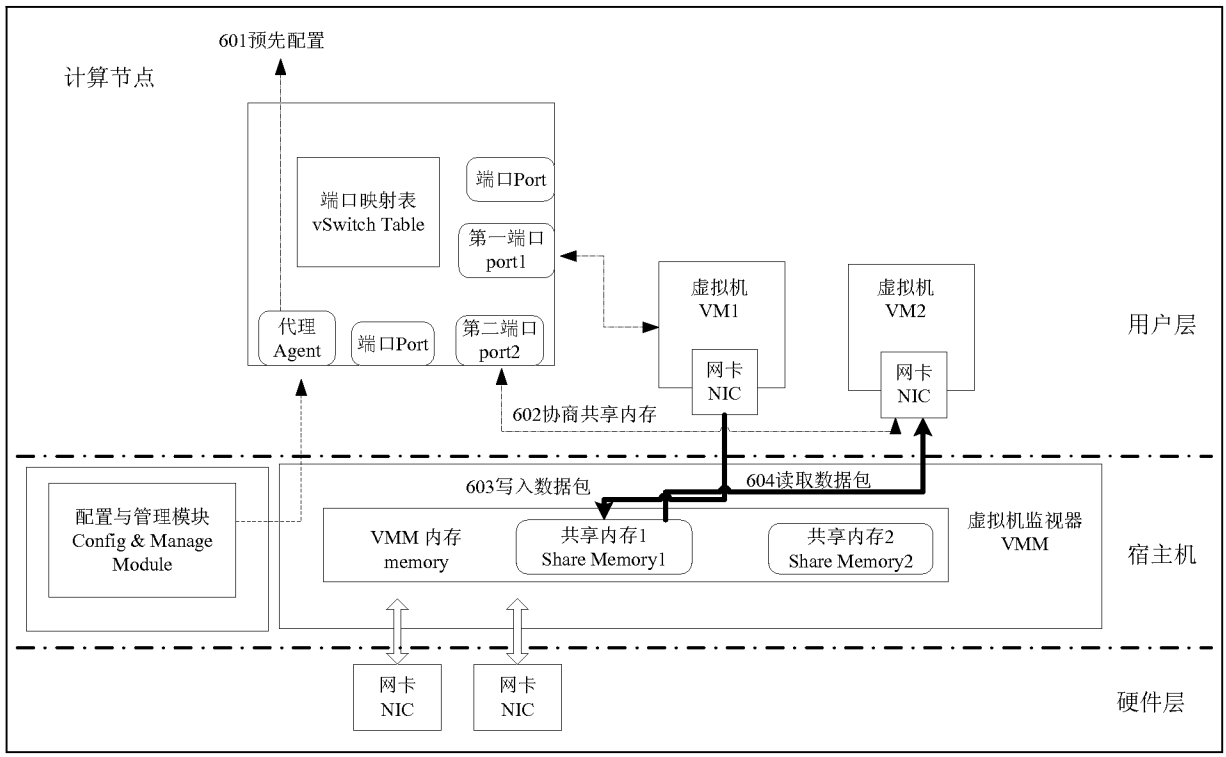

601预先配置
计算节点
端口映射表
vSwitch Table
端口Port
第一端口
port1
代理
Agent
端口Port
第二端口
port2
虚拟机
VM1
网卡
NIC
虚拟机
VM2
网卡
NIC
用户层
602协商共享内存
603写入数据包
604读取数据包
配置与管理模块
Config & Manage
Module
VMM 内存
memory
共享内存1
Share Memory1
共享内存2
Share Memory2
虚拟机监视器
VMM
宿主机
网卡
NIC
网卡
NIC
硬件层

图6

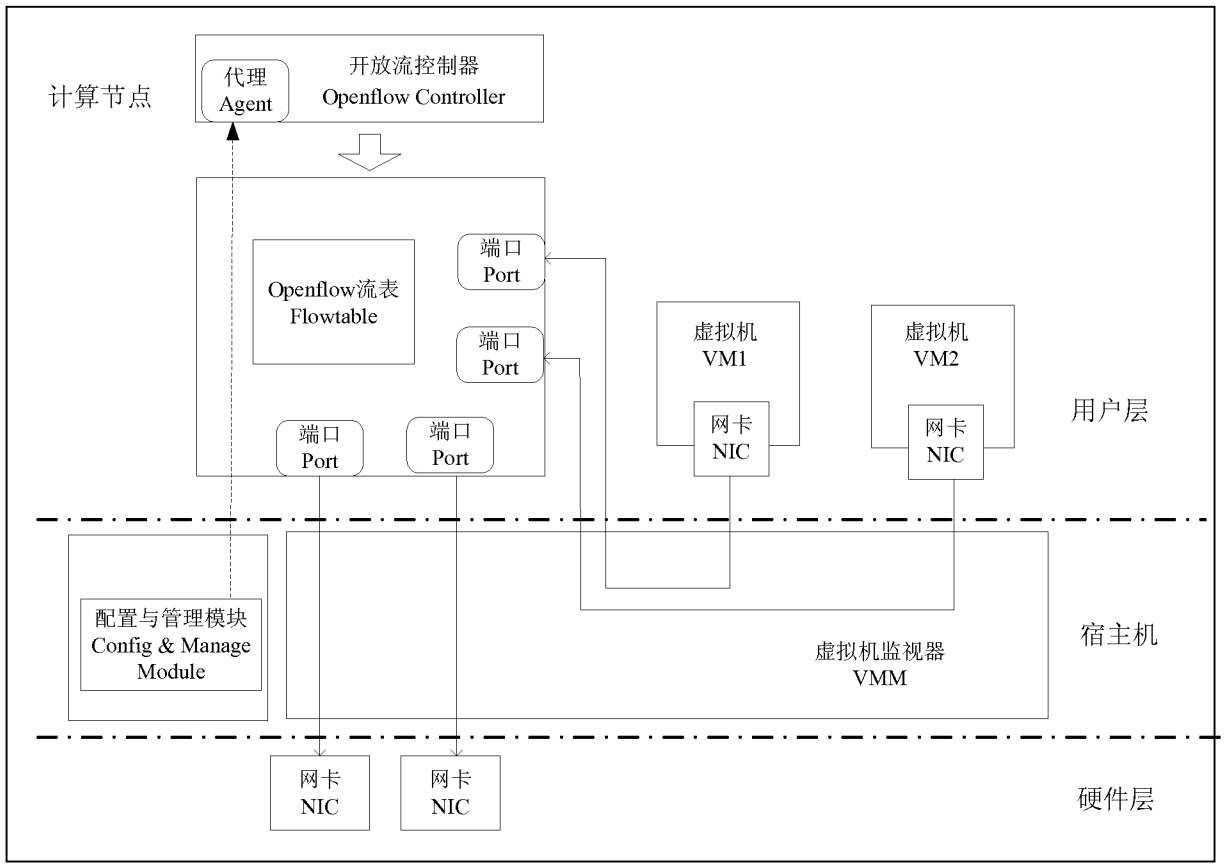

计算节点
代理
Agent
开放流控制器
Openflow Controller
Openflow流表
Flowtable
端口
Port
端口
Port
端口
Port
端口
Port
虚拟机
VM1
网卡
NIC
虚拟机
VM2
网卡
NIC
用户层
配置与管理模块
Config & Manage
Module
虚拟机监视器
VMM
宿主机
网卡
NIC
网卡
NIC
硬件层

图7

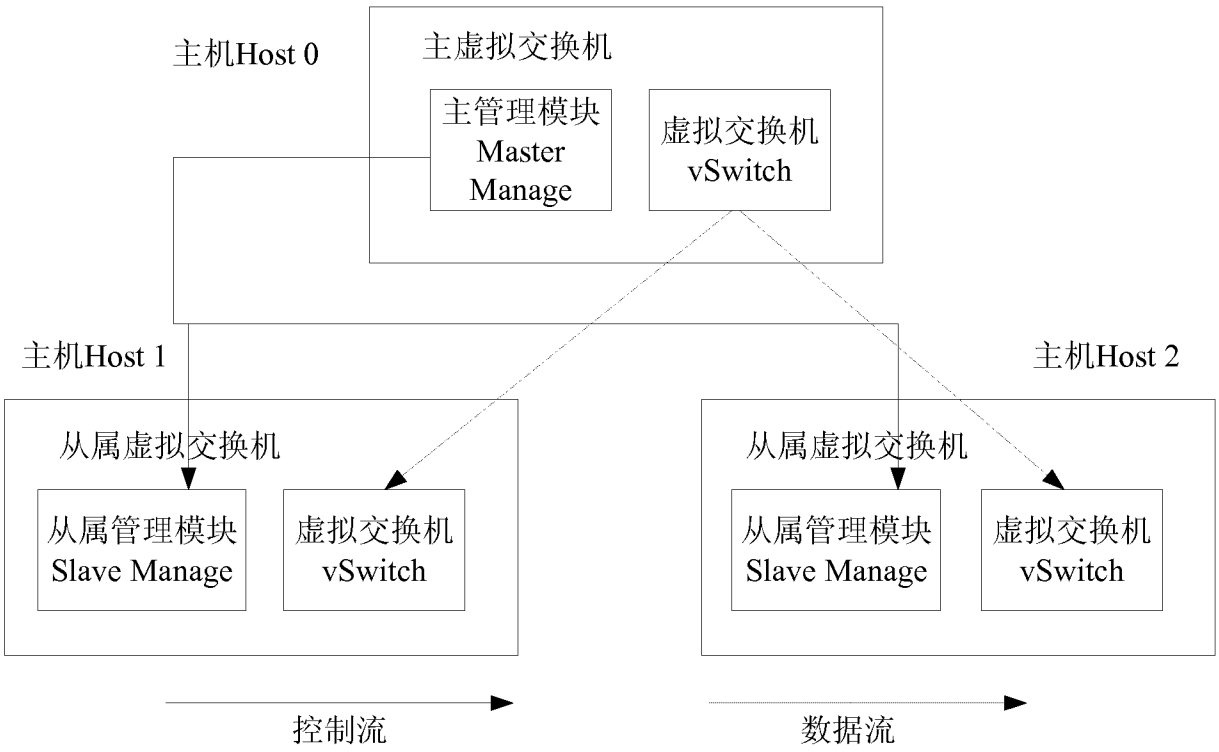

主机Host 0
主虚拟交换机
主管理模块
Master
Manage
虚拟交换机
vSwitch
主机Host 1
主机Host 2
从属虚拟交换机
从属管理模块
Slave Manage
虚拟交换机
vSwitch
从属虚拟交换机
从属管理模块
Slave Manage
虚拟交换机
vSwitch
控制流
数据流

图8

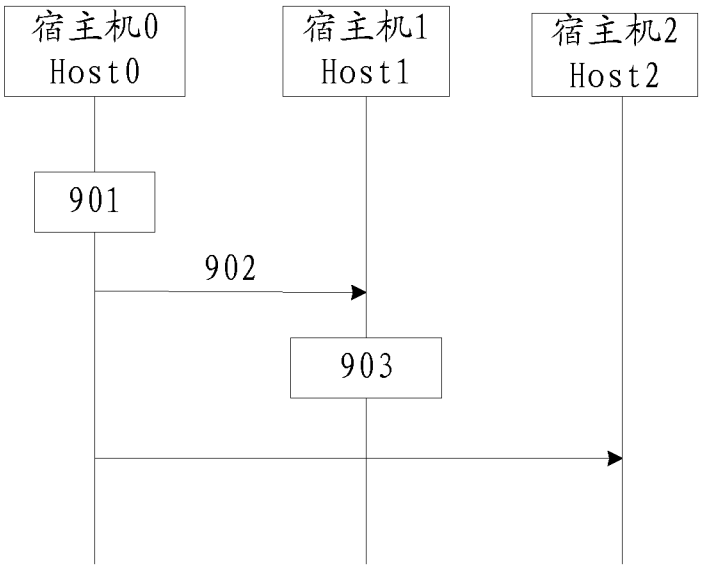

宿主机0
Host0
宿主机1
Host1
宿主机2
Host2
901
902
903

图9

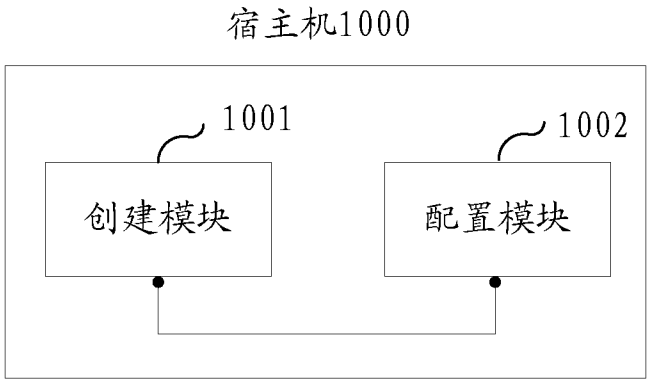

宿主机1000
1001
1002
创建模块
配置模块

图10

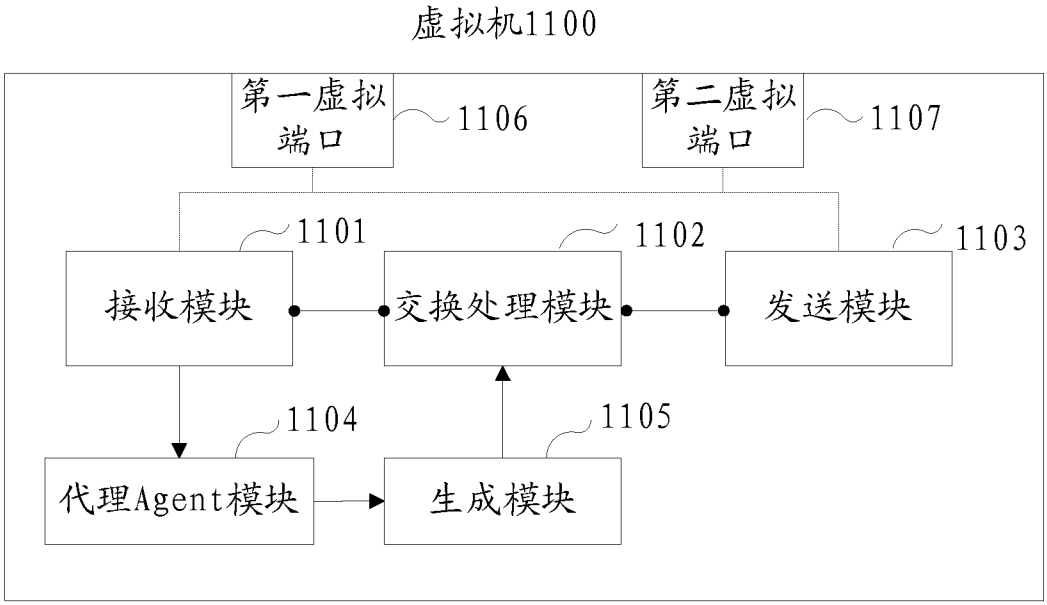

虚拟机1100
第一虚拟端口
1106
第二虚拟端口
1107
1101
1102
1103
接收模块
交换处理模块
发送模块
1104
1105
代理Agent模块
生成模块

图11

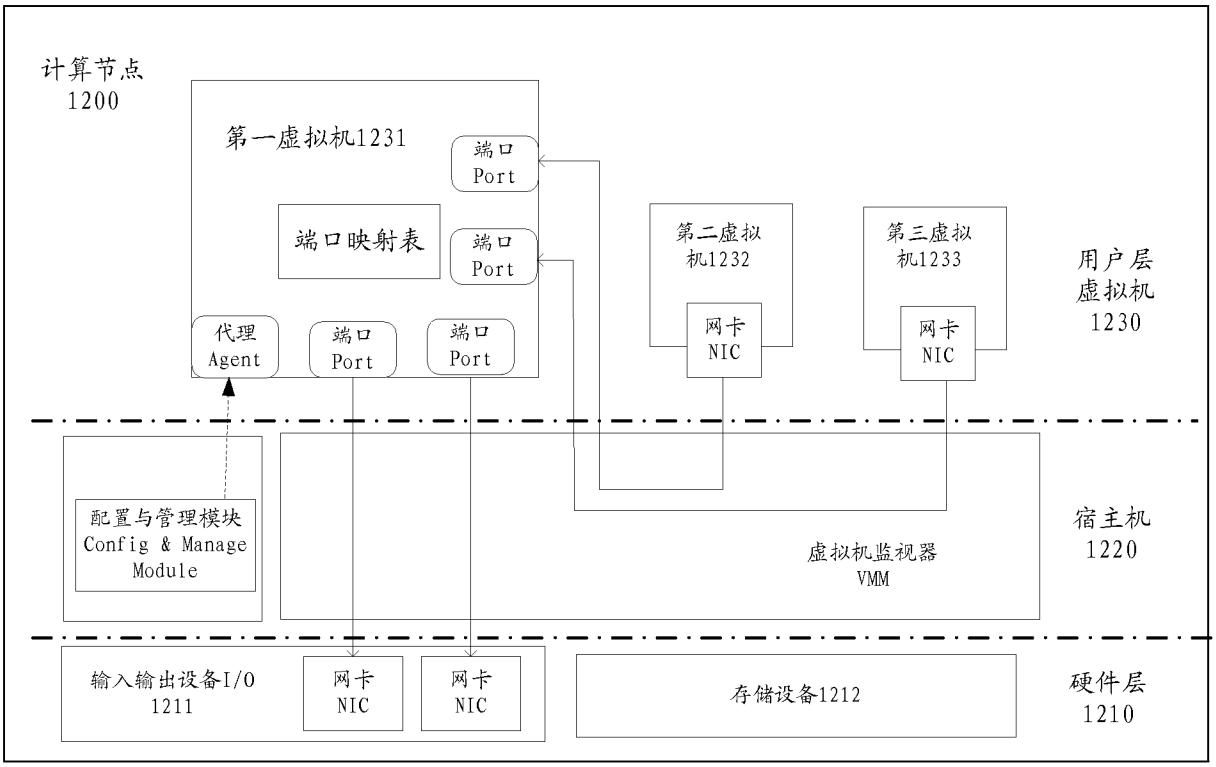

计算节点
1200
第一虚拟机1231
端口
Port
端口映射表
端口
Port
代理
Agent
端口
Port
端口
Port
第二虚拟机1232
网卡
NIC
第三虚拟机1233
网卡
NIC
用户层
虚拟机
1230
配置与管理模块
Config & Manage
Module
虚拟机监视器
VMM
宿主机
1220
输入输出设备I/O
1211
网卡
NIC
网卡
NIC
存储设备1212
硬件层
1210

图12

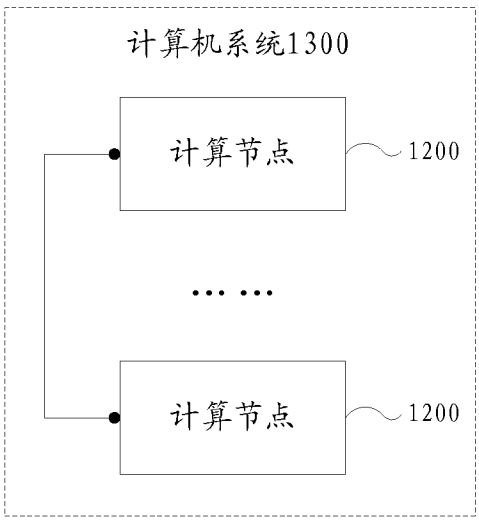

计算机系统1300
计算节点
1200
计算节点
1200

图13

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2014/072502

A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/931 (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

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)

CNABS, CNTXT, CNKI, VEN: host, manage, virtual, machine, VM, VS, vswitch, switch, config+, set, first, second, exchange, main, receiv+, send, address, map+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 103346981 A (HUAWEI TECHNOLOGIES CO., LTD.) 09 October 2013 (09.10.2013) claims 1 to 24	1-24
A	CN 101630270 A (CHENGDU HUAWEI SYMANTEC TECH) 20 January 2010 (20.01.2010) the whole document	1-24
A	CN 103095546 A (HUAWEI TECHNOLOGIES CO., LTD.) 08 May 2013 (08.05.2013) the whole document	1-24
A	CN 102132511 A (CISCO TECH INC.) 20 July 2011 (20.07.2011) the whole document	1-24
A	US 2009268608 A1 (NOKIA SIEMENS NETWORKS GMBH) 29 October 2009 (29.10.2009) the whole document	1-24

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"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)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"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

"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

"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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
11 May 2014	11 June 2014
Name and mailing address of the ISA 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 CAO, Yuanyuan Telephone No. (86-10) 62089367

Form PCT/ISA /210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2014/072502

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 103346981 A	09 October 2013	None	
CN 101630271 A	20 January 2010	CN 101630270 B	26 June 2013
		WO 2011009406 A1	27 January 2011
		US 2012110578 A1	03 May 2012
CN 103095546 A	08 May 2013	None	
CN 102132511 A	20 July 2011	EP 2319204 A1	11 May 2011
		CA 2735154 A1	04 March 2010
		AU 2009285861 A1	04 March 2010
		AU 2009285861 A2	14 April 2011
		CN 102132511 B	09 April 2014
		US 2010054129 A1	04 March 2010
		WO 2010025127 A1	04 March 2010
		US 8385202 B2	26 February 2013
		IN 201101897 P4	03 February 2012
US 2009268608 A1	20 October 2009	EP 1958461 B1	16 March 2011
		WO 2007062884 A1	07 June 2007
		AT 502487 T	15 April 2011
		DE 502006009124 D1	28 April 2011
		US 7961597 B2	14 June 2011
		EP 1958461 A1	20 August 2008

国际检索报告

国际申请号 PCT/CN2014/072502

A. 主题的分类

H04L 12/931 (2013.01) i

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

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

H04L; H04W

包含在检索领域中的除最低限度文献以外的检索文献

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））

CNABS，CNTXT， CNKI： host，管理，虚拟机，虚机，VM，VS，vswitch，虚拟交换，配置，作为，第一，第二，交换，主机，接收，发送，地址，映射 VEN：VS, vswitch, virtual, VM, config+, set, map, exchange

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
PX	CN 103346981A (华为技术有限公司) 2013年 10月 09日 (2013-10-09) 权利要求1-24	1-24
A	CN 101630270A (成都市华为赛门铁克科技有限公司) 2010年 1月 20日 (2010-01-20) 全文	1-24
A	CN 103095546A (华为技术有限公司) 2013年 5月 08日 (2013-05-08) 全文	1-24
A	CN 102132511A (思科技术公司) 2011年 7月 20日 (2011-07-20) 全文	1-24
A	US 2009268608A1 (NOKIA SIEMENS NETWORKS GMBH) 2009年 10月 29日 (2009-10-29) 全文	1-24

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“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）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“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

“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

“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

“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

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2014年 5月 11日	2014年 6月 11日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(ISA/CN) 北京市海淀区蓟门桥西土城路6号 100088 中国 传真号 (86-10)62019451	受权官员 曹元媛 电话号码 (86-10)62089367

表 PCT/ISA/210 (第2页) (2009年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2014/072502

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	103346981A	2013年 10月 09日	无		
CN	101630270A	2010年 1月 20日	CN	101630270B	2013年 6月 26日
			WO	2011009406A1	2011年 1月 27日
			US	2012110578A1	2012年 5月 03日
CN	103095546A	2013年 5月 08日	无		
CN	102132511A	2011年 7月 20日	EP	2319204A1	2011年 5月 11日
			CA	2735154A1	2010年 3月 04日
			AU	2009285861A1	2010年 3月 04日
			AU	2009285861A2	2011年 4月 14日
			CN	102132511B	2014年 4月 09日
			US	2010054129A1	2010年 3月 04日
			WO	2010025127A1	2010年 3月 04日
			US	8385202B2	2013年 2月 26日
			IN	201101897P4	2012年 2月 03日
US	2009268608A1	2009年 10月 29日	EP	1958461B1	2011年 3月 16日
			WO	2007062884A1	2007年 6月 07日
			AT	502487T	2011年 4月 15日
			DE	502006009124D1	2011年 4月 28日
			US	7961597B2	2011年 6月 14日
			EP	1958461A1	2008年 8月 20日