

(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
2018 年 7 月 5 日 (05.07.2018)



(10) 国际公布号
WO 2018/120986 A1

- (51) 国际专利分类号:
H04L 12/861 (2013.01)
- (21) 国际申请号: PCT/CN2017/105823
- (22) 国际申请日: 2017 年 10 月 12 日 (12.10.2017)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201611228028.6 2016 年 12 月 27 日 (27.12.2016) 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: 郭峰 (GUO, Feng); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 林海锋 (LIN, Haifeng); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 张海波 (ZHANG, Haibo); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。
- (74) 代理人: 北京龙双利达知识产权代理有限公司 (LONGSUN LEAD IP LTD.); 中国北京市海淀区北清路 68 号院 3 号楼 101, Beijing 100094 (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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,

(54) Title: METHOD FOR FORWARDING PACKET AND PHYSICAL HOST

(54) 发明名称: 转发报文的方法和物理主机

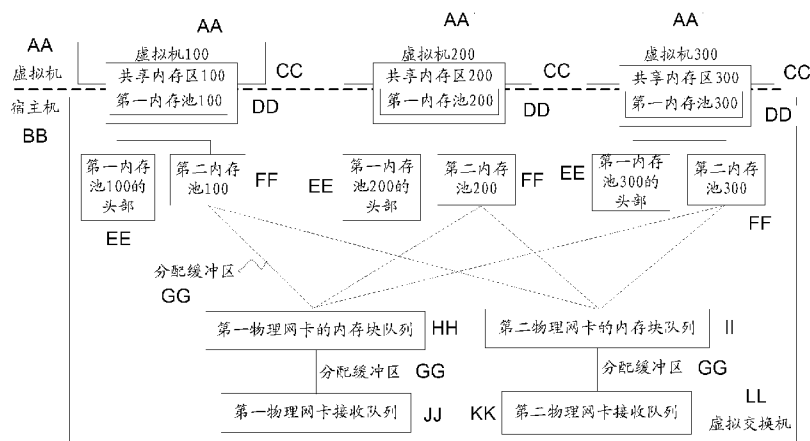


图 4

- AA Virtual machine
BB Host machine
CC Shared memory region
DD First memory pool
EE Head of the first memory pool
FF Second memory pool
GG Allocation buffer region
HH Memory block queue of a first physical network interface card
II Memory block queue of a second physical network interface card
JJ Receiving queue of the first physical network interface card
KK Receiving queue of the second physical network interface card
LL Virtual switch

(57) Abstract: Embodiments of the present invention provide a method for forwarding a packet and a physical host. The physical host comprises a first virtual switch and at least two virtual machines; each of the at least two virtual machines is provided with a shared memory region that can be accessed by both the virtual machine and the physical host; each shared memory region comprises a first memory pool; each first memory pool comprises at least one memory block; an index field is configured for the memory block in each first memory pool for identifying the virtual machine to which the memory block belongs; a first shared memory region corresponding to a first virtual machine in the at least two virtual machines is inaccessible for other virtual machines except for the first virtual machine in the at least two virtual machines. The physical host separately establishes a shared memory region with the at least two virtual

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)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

machines, and the shared memory regions are isolated from each other, so that the possibility of a network packet stored in the shared memory region being modified by other virtual machines is avoided, thereby improving the reliability of the system.

(57) 摘要: 本发明实施例提供了一种转发报文的方法和物理主机。该物理主机上包括第一虚拟交换机和至少两个虚拟机, 该至少两个虚拟机中的每个虚拟机上配置有能够与该物理主机共同访问的共享内存区, 每个共享内存区配置有第一内存池, 每个第一内存池配置有至少一个内存块, 该每个第一内存池中的内存块配置有用于标识所属虚拟机的索引字段, 该至少两个虚拟机中的第一虚拟机对应的第一共享内存区禁止被该至少两个虚拟机中除该第一虚拟机之外的其它虚拟机访问。该物理主机通过与该至少两个虚拟机分别建立共享内存区, 并且每个共享内存区之间相互隔离, 解决共享内存区存储的网络报文被其它虚拟机修改的可能, 能够提升系统的可靠性。

转发报文的方法和物理主机

5 本申请要求于2016年12月27日提交中国专利局、申请号为201611228028.6、发明名称为“转发报文的方法和物理主机”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

10 本发明实施例涉及通信领域，并且更具体地，涉及一种转发报文的方法和物理主机。

背景技术

网络功能虚拟化（Network Functions Virtualization, NFV）是由电信网络运营商提出的，是指用借助信息技术（Information Technology, IT）的虚拟化技术，采用业界标准的大容量服务器、存储器和交换机承载各种各样的网络软件功能的技术标准。

15 虚拟化技术在一个物理服务器上可以同时运行多个虚拟机（Virtual machine, VM）。这些VM一方面需要共享物理网卡与外界资源通信，另一方面，VM之间也需要通信。为此，在虚拟机监控器（Virtual Machine Monitor, VMM）中增加虚拟交换机（Virtual switch, Vswtich）来解决上述问题。为了提升Vswtich的性能，业界推出了用户态Vswtich，使用数据面开发包（Data Plane Development Kit, DPDK）从物理网卡收发网络报文。

20 现有技术中，Ivshmem方案通过配置虚拟操作系统模拟器（QEMU）命令可以使得宿主主机（Host）和所有VM共享同一块内存存储区。由于Host和Host上的VM共享一块内存，可以在该共享内存区建立内存池（mbufpool），该内存池可以配置有多个内存块（mbuf），VM收发报文和Host的物理网卡收发报文都可以使用该共享内存区的内存块。由于mbuf存在于共享内存区，Host可以直接把物理网卡收到的网络报文传给VM，而不用拷贝网络报文的内容。

25 但是，当Host共享内存给所有的VM时，Host和所有的VM都能读写该共享内存区，在共享内存中建立的mbuf信息能被所有的虚拟机改写，这将可能会给系统带来灾难性的后果，降低了系统的可靠性。

30

发明内容

本发明实施例提供了一种转发报文的方法和物理主机，能够提升系统可靠性。

第一方面，提供了一种转发报文的方法，所述方法应用于物理主机，所述物理主机上包括：

35 第一虚拟交换机和至少两个虚拟机，所述至少两个虚拟机中的每个虚拟机上配置有能够与所述物理主机共同访问的共享内存区，每个共享内存区配置有第一内存池，每个第一内存池配置有至少一个内存块，所述每个第一内存池中的内存块配置有用于标识所属虚拟机的索引字段，所述至少两个虚拟机中的第一虚拟机对应的第一共享内存区禁止被所述至

少两个虚拟机中除所述第一虚拟机之外的其它虚拟机访问;

所述方法包括:

所述第一虚拟交换机获取存有第一报文的第一内存块,所述第一内存块携带有助于标识所述第一内存块的所属虚拟机的索引字段;

5 所述第一虚拟交换机根据所述第一内存块的索引字段,转发所述第一报文。

由于该物理主机与该至少两个虚拟机分别建立共享内存区,并且每个共享内存区之间相互隔离,解决共享内存区存储的网络报文被其它虚拟机修改的可能,能够提升系统的可靠性。

10 另外,每个共享内存区存储的网络报文,能够被对应的虚拟机和物理主机共同访问,解决了网络报文在物理主机和虚拟机间传输时,需要拷贝报文内容的问题,达到节省 CPU 消耗,提升系统转发性能的效果。

15 在一个可能的设计中,所述每个第一内存池中的内存块包括第一内存分段和第二内存分段,所述第一内存分段在所述第二内存分段之前,所述第一内存分段用于存储所述第二内存分段的地址信息,所述第二内存分段用于存储报文,所述物理主机上配置有至少两个第二内存池,所述至少两个第二内存池中的每个第二内存池和所述每个第一内存池一一对应,每个第二内存池配置有对应的第一内存池中的第一内存分段;

其中,所述第一虚拟交换机根据所述第一内存块的索引字段,转发所述第一报文,包括:

20 所述第一虚拟交换机根据所述至少两个第二内存池和所述第一内存块的索引字段,转发所述第一报文。

由于该第二内存池中的内存块只包括内存块的 head,虚拟机操作内存块时可以使用共享内存区中的 head,物理主机操作内存块时可以使用第二内存池中的 head,解决了在物理主机操作内存块时,虚拟机修改该内存块的 head 导致物理主机系统异常的问题,进一步提升了系统的可靠性。

25 在一个可能的设计中,所述物理主机根据所述至少两个第二内存池配置有物理网卡接收队列,所述第一虚拟机配置有发送队列和/或接收队列,所述第一虚拟机的接收队列或者发送队列包括所述第一虚拟机的第一内存池中的部分或者全部内存块;

其中,所述第一虚拟交换机获取存有第一报文的第一内存块,包括:

30 第一虚拟交换机获取所述物理网卡接收队列中的所述第一内存块,或者,第一虚拟交换机获取所述第一虚拟机的发送队列中的所述第一内存块。

在一个可能的设计中,所述第一虚拟交换机用于向所述第一虚拟机转发报文,其中,所述第一虚拟交换机根据所述至少两个第二内存池和所述第一内存块的索引字段,转发所述第一报文,包括:

35 若所述第一内存块的索引字段标识的虚拟机为所述第一虚拟机,所述第一虚拟交换机从所述第一虚拟机的第一内存池中调度一个内存块作为第二内存块;所述第一虚拟交换机将所述第一内存块的地址信息赋予所述第二内存块;所述第一虚拟交换机将所述第二内存块填充到所述第一虚拟机的接收队列。

在一个可能的设计中,所述第一虚拟交换机用于向所述第一虚拟机转发报文,其中,所述第一虚拟交换机根据所述至少两个第二内存池和所述第一内存块的索引字段,转发所

述第一报文, 包括:

若所述第一内存块的索引字段标识的虚拟机不是所述第一虚拟机, 所述第一虚拟交换机从所述至少两个第二内存池中调度一个内存块作为第三内存块, 所述第三内存块的索引
5 字段标识的虚拟机为所述第一虚拟机; 所述第一虚拟交换机将所述第一内存块的内容拷贝
到所述第三内存块; 所述第一虚拟交换机从所述第一虚拟机的第一内存池中调度一个内存
块作为第二内存块; 所述第一虚拟交换机将所述第三内存块的地址信息, 赋予所述第二内
存块; 所述第一虚拟交换机将所述第二内存块填充到所述第一虚拟机的接收队列。

10 由于转发报文的内存块从第二内存池中调取, 而第二内存池的数据区是不共享的, 并
且包括 head 信息, 因此, 能够有效防止报文转发过程中内存块的 head 信息被虚拟机破坏,
提高了系统的可靠性。

另外, 每个共享内存区存储的网络报文, 能够被对应的虚拟机和物理主机共同访问,
解决了网络报文在物理主机和虚拟机间传输时, 需要拷贝报文内容的问题, 达到节省 CPU
消耗, 提升系统转发性能的效果。

15 在一个可能的设计中, 所述第一虚拟交换机用于向所述物理网卡接收队列转发所述第
一报文, 其中, 所述第一虚拟交换机根据所述至少两个第二内存池和所述第一内存块的索
引字段, 转发所述第一报文, 包括:

20 所述第一虚拟交换机从所述至少两个第二内存池中获取第四内存块, 所述第四内存块
的索引字段标识的虚拟机为所述第一虚拟机; 所述第一虚拟交换机将所述第一内存块的地
址信息, 赋予所述第四内存块; 所述第一虚拟交换机将所述第四内存块填充到所述物理网
卡接收队列。

由于转发报文的内存块从第二内存池中调取, 而第二内存池的数据区是不共享的且包
括 head 信息, 能够有效防止报文转发过程中内存块的 head 信息被虚拟机破坏, 提高了系
统的可靠性。

25 另外, 每个共享内存区存储的网络报文, 能够被对应的虚拟机和物理主机共同访问,
解决了网络报文在物理主机和虚拟机间传输时, 需要拷贝报文内容的问题, 达到节省 CPU
消耗, 提升系统转发性能的效果。

第二方面, 本发明实施例提供了一种物理主机, 所述物理主机上包括:

30 第一虚拟交换机和至少两个虚拟机, 所述至少两个虚拟机中的每个虚拟机上配置有能
够与所述物理主机共同访问的共享内存区, 每个共享内存区配置有第一内存池, 每个第
一内存池配置有至少一个内存块, 所述每个第一内存池中的内存块配置有用于标识所属虚
拟机的索引字段, 所述至少两个虚拟机中的第一虚拟机对应的第一共享内存区禁止被所述至
少两个虚拟机中除所述第一虚拟机之外的其它虚拟机访问;

35 其中, 所述第一虚拟交换机用于获取存有第一报文的第一内存块, 所述第一内存块携
带有用于标识所述第一内存块的所属虚拟机的索引字段; 所述第一虚拟交换机还用于根据
所述第一内存块的索引字段, 转发所述第一报文。

在一个可能的设计中, 所述每个第一内存池中的内存块包括第一内存分段和第二内存
分段, 所述第一内存分段在所述第二内存分段之前, 所述第一内存分段用于存储所述第二
内存分段的地址信息, 所述第二内存分段用于存储报文, 所述物理主机上配置有至少两个
第二内存池, 所述至少两个第二内存池中的每个第二内存池和所述每个第一内存池一一

应, 每个第二内存池配置有对应的第一内存池中的第一内存分段,

其中, 所述第一虚拟交换机具体用于根据所述至少两个第二内存池和所述第一内存块的索引字段, 转发所述第一报文。

5 在一个可能的设计中, 所述物理主机根据所述至少两个第二内存池配置有物理网卡接收队列, 所述第一虚拟机配置有发送队列和/或接收队列, 所述第一虚拟机的接收队列或者发送队列包括所述第一虚拟机的第一内存池中的部分或者全部内存块;

其中, 第一虚拟交换机具体用于获取所述物理网卡接收队列中的所述第一内存块, 或者, 第一虚拟交换机具体用于获取所述第一虚拟机的发送队列中的所述第一内存块。

10 在一个可能的设计中, 所述第一虚拟交换机用于向所述第一虚拟机转发报文, 其中, 若所述第一内存块的索引字段标识的虚拟机为所述第一虚拟机, 所述第一虚拟交换机从所述第一虚拟机的第一内存池中调度一个内存块作为第二内存块; 所述第一虚拟交换机用于将所述第一内存块的地址信息赋予所述第二内存块; 所述第一虚拟交换机用于将所述第二内存块充填到所述第一虚拟机的接收队列。

15 在一个可能的设计中, 所述第一虚拟交换机用于向所述第一虚拟机转发报文, 其中, 若所述第一内存块的索引字段标识的虚拟机不是所述第一虚拟机, 所述第一虚拟交换机从所述至少两个第二内存池中调度一个内存块作为第三内存块, 所述第三内存块的索引字段标识的虚拟机为所述第一虚拟机; 所述第一虚拟交换机用于将所述第一内存块的内容拷贝到所述第三内存块; 所述第一虚拟交换机用于从所述第一虚拟机的第一内存池中调度一个内存块作为第二内存块; 所述第一虚拟交换机用于将所述第三内存块的地址信息, 赋予所述
20 第二内存块; 所述第一虚拟交换机用于将所述第二内存块充填到所述第一虚拟机的接收队列。

在一个可能的设计中, 所述第一虚拟交换机用于向所述物理网卡接收队列转发所述第一报文, 其中, 所述第一虚拟交换机用于从所述至少两个第二内存池中获取第四内存块, 所述第四内存块的索引字段标识的虚拟机为所述第一虚拟机; 所述第一虚拟交换机用于将
25 所述第一内存块的地址信息, 赋予所述第四内存块; 所述第一虚拟交换机用于将所述第四内存块填充到所述物理网卡接收队列。

第三方面, 提供了一种物理主机, 本发明实施例提供的物理主机包括输入设备、输出设备、处理器和存储器, 其中, 所述处理器、所述存储器和接口之间通过内部连接通路互相通信, 传递控制和/或数据信号。

30 所述处理器通过调用存储器存储的虚拟机软件在物理主机上模拟出第一虚拟交换机和至少两个虚拟机, 该至少两个虚拟机中的每个虚拟机上配置有能够与该物理主机共同访问的共享内存区, 每个共享内存区配置有第一内存池, 每个第一内存池配置有至少一个内存块, 该每个第一内存池中的内存块配置有用于标识所属虚拟机的索引字段, 该至少两个虚拟机中的第一虚拟机对应的第一共享内存区禁止被该至少两个虚拟机中除该第一虚拟机
35 之外的其它虚拟机访问, 该第一虚拟交换机用于:

获取存有第一报文的第一内存块, 该第一内存块携带有用于标识该第一内存块的所属虚拟机的索引字段; 根据该第一内存块的索引字段, 转发该第一报文。

结合上述各个方面, 在一个可能的设计中, 所述地址信息包括数据长度字段的信息和数据起始位置偏移字段的信息。

结合上述各个方面,在一个可能的设计中,所述物理主机配置有物理网卡内存块队列,所述方法还包括:所述物理主机根据所述至少两个第二内存池中的内存块,填充所述物理网卡内存块队列。

5 结合上述各个方面,在一个可能的设计中,在所述物理主机根据所述至少两个第二内存池中的内存块,填充所述物理网卡内存块队列之前,所述方法还包括:

结合上述各个方面,所述物理主机获取所述至少两个虚拟机中每个虚拟机的流量统计信息;所述物理主机根据所述流量统计信息,确定所述至少两个虚拟机中每个虚拟机的流量与所述至少两个虚拟机的总流量之间的流量比值;其中,所述物理主机根据所述至少两个第二内存池中的内存块,填充所述物理网卡内存块队列,包括:所述物理主机将所述第
10 一虚拟机的流量比值确定为,填充到所述物理网卡内存块队列中的所述第一虚拟机对应的内存块数量与填充到所述物理网卡内存块队列中的内存块总数量之间的比值。

本发明实施例的物理网卡内存块队列,通过对进入每个虚拟机的网络流量监控,动态调整物理网卡内存块队列中来自各第二内存池的内存块比例,能够解决存储报文的内存块与目的虚拟机的内存块匹配概率低的问题。

15 结合上述各个方面,在一个可能的设计中,所述物理主机通过为所述第一虚拟机配置元数据文件生成所述第一共享内存区,所述元数据文件包括所述第一虚拟机的端口信息和所述第一虚拟机的第一内存池中的内存块。

由于该元数据文件包括第一虚拟机的端口信息,而不包括其他虚拟机的端口信息,能够有效防止本虚拟端口破坏其它虚拟端口的重要信息。

20 另外,将该元数据文件包括第一虚拟机的第一内存池的内存块,而不包括第一虚拟机的第一内存池的头部,能够防止第一内存池的头部管理信息被虚拟机破坏,进一步提高系统的可靠性。

结合上述各个方面,在一些可能的设计中,所述物理主机还可以根据所述至少两个虚拟机之间的流量比值;从各第二内存池中申请相应数量的内存块放入物理网卡内存块队列
25 中。

附图说明

图1是本发明实施例的虚拟机架构的示意性框图。

图2是本发明实施例的内存块的示意性结构图。

30 图3是现有技术中的转发报文的装置的原理的示意性框图。

图4是根据本发明实施例的转发报文的物理主机的内部结构的示意性框图。

图5是根据本发明实施例的第一共享内存区对应的第二内存池的示意性结构框图。

图6是根据本发明实施例的物理网卡接收队列的示意性结构图。

图7是根据本发明实施例的共享内存区和第二内存池的示意性结构框图。

35 图8是根据本发明实施例的第一虚拟交换机向第一虚拟机转发报文的方法的示意性流程图。

图9是根据本发明实施例的第一虚拟交换机向物理网卡接收队列转发报文的方法的示意性流程图。

图10是根据本发明实施例的物理主机的示意性框图。

具体实施方式

本发明实施例提供的方法和物理主机可以应用于包括多种操作系统的虚拟机架构中。例如，可以应用于包括 Linux、Windows 和 Unix 等一种或多种操作系统的虚拟机架构中，虚拟机架构中还可以包括其它一些操作系统，本发明实施例对此不做限定。

为了便于理解，首先结合图 1 至图 3，对适用于本发明实施例的转发报文的方法的通信系统的架构以及通信系统中各设备的功能进行说明。

图 1 是本发明实施例的虚拟机架构的示意性框图。

如图 1 所示，虚拟机包括用户态和内核态，是操作系统的两种运行级别。用户在用户态上发出各种指令，用户发出的指令从虚拟机的用户态生成下发到虚拟机的内核态。

虚拟机监控器（Virtual Machine Monitor, VMM）与虚拟机通信连接，VMM 也可以称为 Hypervisor，用于管理上述一个或多个虚拟机。本发明实施例中的 VMM 是一种运行在物理服务器和操作系统之间的中间软件层，VMM 可以允许多个操作系统和应用共享一套基础物理硬件。因此，VMM 也可以看作是虚拟环境中的“元”操作系统，它可以协调访问服务器上的所有硬件层设备和虚拟机。VMM 的基本功能是非中断地支持多工作负载迁移的能力，当服务器启动并执行 VMM 时，它会给每一台虚拟机分配适量的内存、CPU、网络和磁盘，并加载所有虚拟机的客户操作系统。

其中，C 语言中分配内存的相关函数主要有 alloc、calloc、malloc、free、realloc、sbrk 等。其中，alloc 是申请内存无需释放。

应理解，虚拟机架构（或系统）中可以包括一个或多个虚拟机。这些虚拟机一方面需要共享物理网卡与外界资源通信，另一方面，虚拟机之间也需要通信，为此可以在 Hypervisor 中增加虚拟交换机来解决上述问题。为了进一步提升虚拟交换机的性能，业界推出了用户态虚拟交换机（Vswtich），基于数据面开发包（Data Plane Development Kit, DPDK）的应用软件根据 mbuf 的数据结构从物理网卡的 mbuf 收发网络报文。

需要注意的是，在虚拟交换环境中，物理网卡为宿主机（Host）持有，虚拟机必须通过 Host 才能利用物理网卡和外部进行网络交互，而在虚拟化环境下，虚拟机和虚拟机之间、虚拟机和主机之间的内存空间彼此隔离，必须通过特殊手段才能在虚拟机和 Host 间传递网络报文。

其中，物理网卡也叫网络适配器（Network Interface Card, NIC）。

图 2 示出了本发明实施例的内存块的示意性结构图。

如图 2 所示，mbuf 的数据结构，分为第一内存分段，用于存储该内存块中用于存储报文的第二内存块分段的地址信息，该第一内存分段也可以称为该内存块的头部（head），该第二内存分段也可以称为报文存储部分（data）。其中，head 中的地址信息可以通过 m->buf_addr 指令指示报文的 mbuf 地址，通过 m->buf_len 指示报文的 mbuf 长度，通过指针 m->pkt.data 指向报文的存储区，通过 m->pkt.data_len==m->pkt.pkt_len 指示报文的实际长度，在第二内存分段前后分别保留头部空间（headroom）和尾部空间（tailroom），以方便应用解封报文。

应理解，mbuf 由内存池（rte_mempool）管理，rte_mempool 在初始化时一次申请多个 mbuf，申请的 mbuf 个数和长度可以都由用户指定，其中，rte_mempool 可以通过函数

rte_mempool_create()创建。具体地包括:

1) 计算需要为 rte_mempool 申请的内存空间。

2) 通过函数 rte_config.mem_config->free_memseg[]从系统内存中取出合适大小的内存,记录到内存池(rte_config.mem_config->memzone[])中。

5 3) 初始化新创建的 rte_mempool,并调用函数 rte_pktmbuf_pool_init()初始化 rte_mempool 的私有数据结构。

4) 调用函数 mempool_populate(),以及函数 rte_pktmbuf_init()初始化 rte_mempool 的每个 mbuf。

10 应理解,申请的 rte_mempool 可以包括:最前面的内存池数据结构(struct rte_mempool)、私有数据结构(rte_pktmbuf_pool_private),以及分配的内存块(mbuf)。其中,rte_pktmbuf_pool_private 也可以称为内存池的头部。

用户态虚拟交换机或者基于 DPDK 的应用软件通过持有 mbuf 的 head 来访问 data 存储的网络报文内容。

图 3 示出了现有技术中的转发报文的装置的示意性框图。

15 如图 3 所示,用户态虚拟主机管理系统(vHost-user)方案主要由三部分组成:用户态 vHost-user 进程、Qemu 进程的 vHost-user 模块、客户机操作系统(Guest OS)的前端驱动(virtio-net)。其中,用户态 vHost-user 进程的作用是从 Guest OS 的网卡接收报文或往 Guest OS 的网卡发送报文;Qemu 进程的 vHost-user 模块的作用是建立 Guest OS 和用户态 vHost-user 进程之间消息通道;Guest OS 的 virtio-net 网卡的作用是提供统一的用户态

20 的网络通信接口,用户可以通过这个接口发送接收报文。

如果 Guest OS 使用的是内核态驱动还需要算上内核的多计算机切换器(KVM)模块的通知部分,如果是用户态轮询(Poll Mode Drivers, PMD)驱动不需要通知部分。

该装置的初始化基本流程包括:

25 1) 启动用户态 vHost-user 进程,初始化服务端的套接口(socket)并监听客户端的 Qemu 的 socket 事件。

2) 启动 Qemu 进程,连接用户态 vHost-user 进程,通过内存映射(mmap)下发 Guest OS 内存信息给用户态 vHost-user 进程;用户态 vHost-user 进程将接收到的内存信息内存映射到自己的地址空间,获取 Guest OS 内存的读写权限,以使用户态 vHost-user 进程能够合法的访问 Guest OS 的内存信息。

30 其中,mmap 系统调用可以将一个磁盘文件的全部或部分内容映射到用户空间中,进程读写文件的操作变成了读写内存的操作,实现了更有用的动态内存分配功能。

35 3) 启动 Guest OS,加载 virtio-net 网卡驱动,下发 virtio-net 网卡收发队列信息给用户态 vHost-user 进程,并和用户态 vHost-user 进程协商网卡特性;用户态 vHost-user 进程接收 virtio-net 网卡收发队列信息后,转换为自己进程的地址,即可访问和操作 virtio-net 网卡的收发队列。

在 VM 收包流程中,virtio-net 网卡初始化会往 virtio-net 网卡接收队列(例如,vring 队列)填充内存;用户态 vHost-user 进程从物理网卡或虚拟网卡接收报文并存储在分配的内存上,然后获取 virtio-net 网卡的接收队列可用的内存地址,将报文拷贝到 virtio-net 网卡的接收队列。

在 VM 发包流程中, Guest OS 的用户态的进程在生成一个报文后, 通过 virtio-net 网卡的发送接口将报文填入发送队列 (例如, vring 队列); 用户态 vHost-user 进程轮询到发送队列有新的报文时, 分配一块内存, 将发送队列中的新报文拷贝到自己的内存, 然后处理报文, 将报文发往物理网卡或虚拟端口。

- 5 由于在虚拟化环境中, VM 没有访问 Host 内存的权限, Host 可以访问 VM 的 virtio 网卡的 vring 队列中的内存, 而物理网卡从 Host 外接收网络报文的缓冲区使用的是 Host 的内存。

10 所以, 网络报文经过 Vswtich 交换后, 必须由 Host 拷贝网络报文内容到 VM 的 vring 的内存中, 这样 VM 才能处理网络报文。由于网络报文在进虚拟机必须进行报文内容的拷贝, 大量消耗 Host 的中央处理器 (Central Processing Unit, CPU), 从而显著影响到 Vswtich 的转发性能。

15 为此, DPDK 程序通过 QEMU 的 Ivshmem 机制, 为虚拟机提供了快速 (Host to guest 和 guest to Host) 的零拷贝数据共享方案。具体而言, Ivshmem 方案通过配置 QEMU 命令, 把一到多个物理大页映射到一个 Ivshmem 设备中。Ivshmem 设备被 QEMU 当作 PCI 内存设备映射到虚拟机中。通过 Ivshmem 可以使得 Host 和所有 VM 共享同一块内存存储区。由于 Host 和 Host 上的 VM 共享一块内存, 可以在该共享内存区建立 mbufpool, VM 收发报文和 Host 的物理网卡收发报文都可以使用该共享内存区的 mbuf。由于 mbuf 存在于共享内存区, Host 可以直接把物理网卡收到的网络报文传给 VM, 而不用拷贝网络报文的内容。同样 Vswtich 也可以把一个 VM 发出的网络报文直接转发给另一个 VM, 不用进行报文内容的拷贝。

20 由于 Ivshmem 方案从物理网卡接收网络报文到进出 VM 都不用进行报文内存的拷贝, 因此, 能节省大量的 CPU 消耗, 提升 vswitch 的转发性能。

25 但是, 由于 Host 可以共享内存给所有的 VM, Host 和所有的 VM 都能读写该共享内存区, 在共享内存中建立的 mbuf 信息能被所有的虚拟机改写, 这将可能会给系统带来灾难性的后果, 降低系统的可靠性。

 例如, 其中某个 VM 程序异常导致改写共享内存区中 head 的指针信息, 例如, m->pkt.data 改写为 NULL, Host 或其它 VM 访问该指针时, 将导致 Host 或 VM 出现程序异常。

30 再例如, 当某个 VM 在处理 data 区存储的网络报文时, 其它 VM 程序异常改写 IP 首部的长度信息, 该 VM 使用异常的 IP 长度信息将导致不可预知的异常, 使 VM 不可用。

 这些由于单个 VM 程序异常导致 Host 和其它 VM 异常的问题, 在商用领域是不可接受的, 降低了系统的可靠性。

 本发明实施例提出了一种高性能、高可靠的虚拟机和 Host 间传递网络报文的方法和物理主机, 能够提高系统的可靠性。

35 可选地, 该物理主机包括至少两个虚拟机, 该至少两个虚拟机中的每个虚拟机上配置有能够与该第一物理主机共同访问的共享内存区, 每个共享内存区配置有第一内存池, 每个第一内存池配置有至少一个内存块, 每个第一内存池中的内存块配置有用于标识所属虚拟机的索引字段, 该至少两个虚拟机中的第一虚拟机对应的第一共享内存区禁止被该至少两个虚拟机中除该第一虚拟机之外的其它虚拟机访问。

由于该物理主机与该至少两个虚拟机分别建立共享内存区，并且禁止其它虚拟机访问该共享内存区，解决共享内存区存储的网络报文被其它虚拟机修改的可能，能够提升系统的可靠性。

5 另外，每个共享内存区存储的网络报文，能够被对应的虚拟机和物理主机共同访问，解决了网络报文在物理主机和虚拟机间传输时，需要拷贝报文内容的问题，达到节省 CPU 消耗，提升系统转发性能的效果。

10 可选地，第一内存池中的内存块包括第一内存分段和第二内存分段，该第一内存分段在该第二内存分段之前，该第一内存分段用于存储该第二内存分段的地址信息，该第二内存分段用于存储报文，该物理主机上配置有至少两个第二内存池，该至少两个第二内存池中的每个第二内存池和每个第一内存池一一对应，每个第二内存池配置有对应的第一内存池中的第一内存分段，可选地，该地址信息包括数据长度字段（data_len）的信息和数据起始位置偏移字段（data_off）的信息。

15 具体而言，可以通过在内存块的数据结构中增加索引字段（index），用于标识内存块所属的虚拟机。可选地，在第二内存池中内存块初始化阶段，对其进行重新赋值，使第二内存池与第一内存池中 index 相同的内存块具有相同的数据结构，即每个第二内存池配置有对应的第一内存池中的第一内存分段。

20 由于该第二内存池中的内存块只包括内存块的 head。虚拟机操作内存块时使用共享内存区中的 head，物理主机操作内存块时使用第二内存池中的 head。解决了在物理主机操作内存块时，由于虚拟机修改该内存块的 head 导致物理主机系统异常的问题，进一步提升了系统的可靠性。

图 4 是根据本发明实施例的转发报文的物理主机的内部结构的示意性框图。

25 如图 4 所示，物理主机和虚拟机（VM）100、虚拟机 200 和虚拟机 300 分别建立可共同访问的共享内存区 100、共享内存区 200 和共享内存区 300，共享内存区 100 禁止被该物理主机内除虚拟机 100 之外的其它虚拟机访问，共享内存区 200 禁止被该物理主机内除虚拟机 200 之外的其它虚拟机访问，共享内存区 300 禁止被该物理主机内除虚拟机 300 之外的其它虚拟机访问。在这各共享内存区中构建物理网卡接收报文用 mbuf 的第一内存池（mempool），每个第一内存池包含至少一个内存块。物理主机在物理内存中建立第二内存池 100、第二内存池 200 和第二内存池 300。物理主机根据第二内存池 100、第二内存池 200 和第二内存池 300 建立物理主机中的第一物理网卡的内存块队列和第二物理网卡的内存块队列，同时每个物理网卡内存块队列对应有物理网卡的接收队列（rx queue），物理网卡接收队列从对应的物理网卡内存块队列中调用内存块。

35 在本发明实施例中，第一内存池也可以称为 vm mempool，相应的，第一内存池的头部可以称为 vm mempool head。第二内存池也可以称为 Host-vm mempool，第二内存池的头部也可以称为 Host-vm mempool head，内存块也可以称为 mbuf，本发明实施例不作具体限定。

应理解，图 4 示例性的描述了本发明实施例的物理主机的内部结构，本发明实施例不作限定。

例如，该物理主机上可以配置有一个或者多个物理网卡内存块队列，以及一个或者多个物理网卡接收队列。

又例如，该物理主机上包括一个或者多个虚拟机等。

图 5 是根据本发明实施例的第一共享内存区对应的第二内存池的示意性结构框图。

如图 5 所示，物理主机上运行的第一虚拟机配置有第一共享内存区，第一共享内存区中的内存池配置有第一内存块 110、第一内存块 120、第一内存块 130，每个内存块都包
5 括第一内存分段和第二内存分段。物理主机配置有第二内存池，第二内存池上配置有每个内存块的第一内存分段。其中，第二内存池与第一内存池中 index 相同的第一内存分段对应同一个第二内存分段，例如，第一内存池中的第一内存分段 110 和第二内存池中的第一内存分段 110 对应第二内存分段 110。在本发明实施例中，第二内存池中的内存块禁止被该物理主机上的虚拟机访问，第一内存池中的内存块可以被物理主机和第一虚拟机访问。

10 应理解，图 5 示例性的描述了第一虚拟机的第一内存池和对应的物理主机上的第二内存池，本发明实施例不作限定。

例如，该物理主机上其它虚拟机也可以配置有第二内存池。

可选地，该物理主机可以根据该至少两个第二内存池中的内存块，填充物理网卡内存块队列，该物理网卡内存块队列作为对应的物理网卡接收队列调用内存块的队列。

15 具体而言，物理网卡接收队列接收报文时，将报文存储于该物理网卡接收队列中的内存块，同时从对应的物理网卡内存块队列直接调用内存块填充该物理网卡接收队列空出的位置。

20 可选地，该物理主机获取该至少两个虚拟机中每个虚拟机的流量统计信息；根据该流量统计信息，确定该至少两个虚拟机中每个虚拟机的流量与该至少两个虚拟机的总流量之间的流量比值；将该第一虚拟机的流量比值确定为，填充到该物理网卡内存块队列中的该第一虚拟机对应的内存块数量与填充到该物理网卡内存块队列中的内存块总数量之间的比值。

图 6 是根据本发明实施例的物理网卡内存块队列的示意性结构图。

25 如图 6 所示，物理主机的内存上配置有虚拟机 100 对应的第二内存池 100、虚拟机 200 对应的第二内存池 200、以及虚拟机 300 对应的第二内存池 300。其中，第二内存池 100 配置有至少一个内存块 100，第二内存池 200 配置有至少一个内存块 200，第二内存池 300 配置有至少一个内存块 300。该物理主机上根据该至少一个内存块 100、该至少一个内存块 200、以及该至少一个内存块 300 确定物理网卡内存块队列中的内存块。

30 具体地，如图 6 所示，物理主机通过统计确定虚拟机 100 流量：虚拟机 200 流量：虚拟机 300 流量=0.25: 0.5: 0.25，则确定填充该物理网卡内存块队列时内存块 100: 内存块 200: 内存块 300=0.25: 0.5: 0.25，根据上述比值和物理网卡内存块队列的容量从第二内存池 100、第二内存池 200 和第三内存池 300 中抽取内存块。

应理解，图 6 示例性的描述了物理网卡内存块队列从第二内存池 100、第二内存池 200 和第三内存池 300 中抽取内存块，本发明实施例不作限定。

35 可选地，该物理主机还可以根据该至少两个虚拟机之间的流量比值；从各第二内存池中申请相应数量的内存块放入物理网卡内存块队列中。

具体而言，该物理主机获取该周期性统计各个虚拟机收到的来自物理网卡的数据包流量，并由统计结果计算出各虚拟机的流量比例；或者每个虚拟机的流量与该至少两个虚拟机的总流量之间的流量比值。

因此,本发明实施例的物理网卡内存块队列,通过对进入每个虚拟机的网络流量监控,动态调整物理网卡内存块队列中来自各第二内存池的内存块比例,能够解决存储报文的内存块与目的虚拟机的内存块匹配概率低的问题。

上文结合图 1 至图 6 介绍了本发明实施例的物理主机上运行的虚拟机框架及虚拟机配置。下面将结合附图 7 详细描述物理主机为每个虚拟机建立共享内存区的具体实现方式。

可选地,该物理主机通过为该第一虚拟机配置元数据文件生成该第一共享内存区,该元数据文件包括该第一虚拟机的端口信息和该第一虚拟机的第一内存池中的内存块。

具体而言,物理主机为每一个虚拟机申请四段内存单元(memzone),其中一段用于存放虚拟端口信息(vport info),其中,物理主机上的每个虚拟机可以分别对应一个虚拟端口,另外三段 memzone 分别用于存放第一内存池的头部,第一内存池的内存块和第二内存池。其中,第一内存池的头部用于存储该第一内存池的控制管理信息。

可选地,将需要共享的 memzone 添加到元数据(metadata)文件,通过使用 QEMU 程序共享给该指定的虚拟机。

例如,物理主机为第一虚拟机创建虚拟端口时,将用于存放该第一虚拟机的第一内存池的内存块的 memzone 和用于存放该第一虚拟机的 vport info 的 memzone 添加到 metadata 文件。

由于该 metadata 文件只添加了用于存放第一虚拟机的 vport info 的 memzone,而没有添加用于存放其他 vport info 的 memzone,能够有效防止本虚拟端口破坏其它虚拟端口的重要信息。

另外,将该 metadata 文件只添加用于存放第一虚拟机的第一内存池的内存块的 memzone,而没有添加用于存放第一虚拟机的第一内存池的头部的 memzone。能够防止第一内存池的头部管理信息被虚拟机破坏,进一步提高系统的可靠性。

图 7 是根据本发明实施例的共享内存区和第二内存池的示意性结构框图。

具体地,如图 7 所示,虚拟机 100 的第一内存池 100 中包括一个或者多个内存块,每个内存块包括第一内存分段和第二内存分段。其中,第一内存池 100 头部和内存块分开存储。物理主机配置有该虚拟机 100 对应的第二内存池 100,该第二内存池 100 只包括第一内存池 100 中内存块的第一内存分段。物理主机为该虚拟机 100 创建共享内存区时,只将用于存放第一内存池 100 的内存块的 memzone 和用于存放第一内存池 100 头部的 memzone 添加到 metadata 文件。

应理解,图 7 示例性的描述了物理物理主机为第一虚拟机建立共享内存区的方法,本发明实施例不作限定。

例如,该物理主机也可以通过上述方法为其它虚拟机建立共享内存区。

因此,本发明实施例的物理主机,通过为该物理主机上运行的虚拟机分别建立共享内存区,并且禁止其它虚拟机访问该共享内存区,解决共享内存区存储的网络报文被其它虚拟机修改的可能,能够提升系统的可靠性。

另外,每个共享内存区存储的网络报文,能够被对应的虚拟机和物理主机共同访问,解决了网络报文在物理主机和虚拟机间传输时,需要拷贝报文内容的问题,达到节省 CPU 消耗,提升系统转发性能的效果。

下面将结合图 8 和图 9 详细介绍基于上文配置的物理主机和虚拟机,转发报文的方法。

物理主机配置的至少一个虚拟交换机中的第一虚拟交换机获取存有第一报文的第一内存块，该第一内存块携带有用于标识该第一内存块的所属虚拟机的索引字段；该第一虚拟交换机根据该第一内存块的索引字段，转发该第一报文。

5 在本发明实施例中，第一虚拟交换机转发第一报文分为两种情况：第一虚拟交换机向第一虚拟机转发该第一报文，以及第一虚拟交换机向物理网卡转发该第一报文。具体地，第一虚拟交换机获取该物理网卡接收队列中的该第一内存块，向第一虚拟机转发该第一报文；或者，第一虚拟交换机获取该第一虚拟机的发送队列中的该第一内存块，向物理网卡接收队列转发该第一报文。

可选地，第一虚拟交换机用于向第一虚拟机转发该第一报文。

10 具体而言，若该第一内存块的索引字段标识的虚拟机为该第一虚拟机，该第一虚拟交换机从该第一虚拟机的第一内存池中调度一个内存块作为第二内存块；该第一虚拟交换机将该第一内存块的地址信息赋予该第二内存块；该第一虚拟交换机将该第二内存块填充到该第一虚拟机的接收队列。

15 若该第一内存块的索引字段标识的虚拟机不是该第一虚拟机，该第一虚拟交换机从该至少两个第二内存池中调度一个内存块作为第三内存块，该第三内存块的索引字段标识的虚拟机为该第一虚拟机；该第一虚拟交换机将该第一内存块的内容拷贝到该第三内存块；该第一虚拟交换机从该第一虚拟机的第一内存池中调度一个内存块作为第二内存块；该第一虚拟交换机将该第三内存块的地址信息，赋予该第二内存块；该第一虚拟交换机将该第二内存块填充到该第一虚拟机的接收队列。

20 图 8 是根据本发明实施例的第一虚拟交换机向第一虚拟机转发报文的方法的示意性流程图。

如图 8 所示，第一虚拟交换机向第一虚拟机转发报文的流程具体包括：

S110，获取物理网卡接收队列中的第一内存块上的第一报文。

S120，该第一内存块的索引字段标识的虚拟机是否为第一虚拟机？

25 S130，若该第一内存块的索引字段标识的虚拟机不是第一虚拟机，从该第一虚拟机对应的第二内存池中调度一个内存块作为第三内存块，并将第一内存块的内容拷贝到该第三内存块。

S140，若该第一内存块的索引字段标识的虚拟机是第一虚拟机，从该第一虚拟机的第一内存池中取出一个内存块作为第二内存块。

30 S150，将该第一内存块、该第三内存块中的 `date_len` 的值和 `date_off` 字段的值赋予该第二内存块。

S160，将该第二内存块放入第一虚拟机的接收队列。

35 在本发明实施例中，由于转发报文的内存块从第二内存池中调取，而第二内存池的数据区是不共享的且包括 `head` 信息，因此，能够有效防止报文转发过程中内存块的 `head` 信息被虚拟机破坏，提高了系统的可靠性。

另外，每个共享内存区存储的网络报文，能够被对应的虚拟机和物理主机共同访问，解决了网络报文在物理主机和虚拟机间传输时，需要拷贝报文内容的问题，达到节省 CPU 消耗，提升系统转发性能的效果。

可选地，第一虚拟交换机用于向物理网卡接收队列转发该第一报文。

具体而言,该第一虚拟交换机从该至少两个第二内存池中获取第四内存块,该第四内存块的索引字段标识的虚拟机为该第一虚拟机;该第一虚拟交换机将该第一内存块的地址信息,赋予该第四内存块;该第一虚拟交换机将该第四内存块填充到该物理网卡接收队列。该地址信息包括数据长度字段的信息和数据起始位置偏移字段的信息。

5 图 9 是根据本发明实施例的第一虚拟交换机向物理网卡接收队列转发报文的方法的示意性流程图。

如图 9 所示,第一虚拟交换机向物理网卡接收队列转发报文的流程具体包括:

S210,获取第一虚拟机的发送队列中的第一内存块。

S220,第一虚拟机的发送队列中有内存块?

10 S230,若第一虚拟机的发送队列中有内存块,从第一虚拟机对应的第二内存池中取出与第四内存块。

S240,将该第一内存块中的 date_len 的值和 date_off 字段的值赋予该第四内存块。

S250,将该第四内存块填充到所述物理网卡接收队列。

15 在本发明实施例中,由于转发报文的内存块从第二内存池中调取,而第二内存池的数据区是不共享的且包括 head 信息,能够有效防止报文转发过程中内存块的 head 信息被虚拟机破坏,提高了系统的可靠性。

另外,每个共享内存区存储的网络报文,能够被对应的虚拟机和物理主机共同访问,解决了网络报文在物理主机和虚拟机间传输时,需要拷贝报文内容的问题,达到节省 CPU 消耗,提升系统转发性能的效果。

20 上述本发明实施例揭示的方法可以应用于处理器中,或者由处理器实现。处理器可能是一种集成电路芯片,具有信号的处理能力。在实现过程中,上述方法的各步骤可以通过处理器中的硬件的集成逻辑电路或者软件形式的指令完成。

图 10 是根据本发明实施例的物理主机 400 的示意性框图。

25 如图 10 所示,本发明实施例提供的物理主机包括输入设备 403(可选的)、输出设备 404(可选的)、处理器 401 和存储器 405。其中,图 10 中的处理器 401、存储器 405 和接口之间通过内部连接通路互相通信,传递控制和/或数据信号。

其中,存储器 405 可以包括只读存储器和随机存取存储器,并向处理器 401 提供指令和数据。存储器 405 的一部分还可以包括非易失性随机存取存储器(NVRAM)。存储器 405 存储了可执行模块或者数据结构,或者它们的子集,或者它们的扩展集。例如,操作指令:包括各种操作指令,用于实现各种操作。又例如,操作系统:包括各种系统程序,用于实现各种基础业务以及处理基于硬件的任务。

处理器 401 通过调用存储器 405 存储的操作指令(该操作指令可存储在操作系统中)。

35 在本发明实施例中,处理器 401 通过调用存储器 405 存储虚拟机软件在物理主机 400 上模拟出第一虚拟交换机 4052 和至少两个虚拟机,该至少两个虚拟机中的每个虚拟机上配置有能够与该物理主机共同访问的共享内存区,每个共享内存区配置有第一内存池,每个第一内存池配置有至少一个内存块,该每个第一内存池中的内存块配置有用于标识所属虚拟机的索引字段,该至少两个虚拟机中的第一虚拟机 4051 对应的第一共享内存区禁止被该至少两个虚拟机中除该第一虚拟机 4501 之外的其它虚拟机访问,该第一虚拟交换机 4502 用于:

获取存有第一报文的第一内存块，该第一内存块携带有用于标识该第一内存块的所属虚拟机的索引字段；根据该第一内存块的索引字段，转发该第一报文。

5 可选地，该每个第一内存池中的内存块包括第一内存分段和第二内存分段，该第一内存分段在该第二内存分段之前，该第一内存分段用于存储该第二内存分段的地址信息，该第二内存分段用于存储报文，该物理主机 400 上配置有至少两个第二内存池，该至少两个第二内存池中的每个第二内存池和该每个第一内存池一一对应，每个第二内存池配置有对应的第一内存池中的第一内存分段。

其中，该第一虚拟交换机 4502 具体用于：根据该至少两个第二内存池和该第一内存块的索引字段，转发该第一报文。

10 可选地，该物理主机 400 根据该至少两个第二内存池配置有物理网卡接收队列，该第一虚拟机 4501 配置有发送队列和/或接收队列，该接收队列或者该发送队列包括该第一虚拟机 4501 的第一内存池中的部分或者全部内存块。

其中，第一虚拟交换机 4502 具体用于：获取该物理网卡接收队列中的该第一内存块，或者，获取该第一虚拟机 4501 的发送队列中的该第一内存块。

15 可选地，该第一虚拟交换机 4502 用于向该第一虚拟机 4501 转发报文。

具体而言，若该第一内存块的索引字段标识的虚拟机为该第一虚拟机 4501，该第一虚拟交换机 4502 具体用于：

20 从该第一虚拟机 4501 的第一内存池中调度一个内存块作为第二内存块；将该第一内存块的地址信息赋予该第二内存块；将该第二内存块填充到该第一虚拟机 4501 的接收队列。

若该第一内存块的索引字段标识的虚拟机不是该第一虚拟机 4501，该第一虚拟交换机 4502 具体用于：

25 从该至少两个第二内存池中调度一个内存块作为第三内存块，该第三内存块的索引字段标识的虚拟机为该第一虚拟机 4501；将该第一内存块的内容拷贝到该第三内存块；从该第一虚拟机 4501 的第一内存池中调度一个内存块作为第二内存块；将该第三内存块的地址信息，赋予该第二内存块；将该第二内存块填充到该第一虚拟机 4501 的接收队列。

30 可选地，该第一虚拟交换机 4502 用于向该物理网卡接收队列转发该第一报文，其中，该第一虚拟交换机 4502 具体用于：从该至少两个第二内存池中获取第四内存块，该第四内存块的索引字段标识的虚拟机为该第一虚拟机 4501；将该第一内存块的地址信息，赋予该第四内存块；将该第四内存块填充到该物理网卡接收队列。

可选地，该地址信息包括数据长度字段的信息和数据起始位置偏移字段的信息。

可选地，该物理主机配置有物理网卡内存块队列，该物理主机还用于根据该至少两个第二内存池中的内存块，填充该物理网卡内存块队列。

35 可选地，在该物理主机根据该至少两个第二内存池中的内存块，填充该物理网卡内存块队列之前，该物理主机用于获取该至少两个虚拟机中每个虚拟机的流量统计信息；该物理主机用于根据该流量统计信息，确定该至少两个虚拟机中每个虚拟机的流量与该至少两个虚拟机的总流量之间的流量比值；该物理主机用于将该第一虚拟机 4501 的流量比值确定为，填充到该物理网卡内存块队列中的该第一虚拟机 4501 对应的内存块数量与填充到该物理网卡内存块队列中的内存块总数量之间的比值。

可选地，该物理主机还用于通过为该第一虚拟机 4501 配置元数据文件生成该第一共享内存区，该元数据文件包括该第一虚拟机 4501 的端口信息和该第一虚拟机 4501 的第一内存池。

应理解，图 10 只是示例性示出了本发明实施例的物理主机的框图，本发明实施例不作限定。例如，该物理主机上运行有多个虚拟机和多个虚拟交换机等等。

在实现过程中，上述软件模块可以位于随机存储器，闪存、只读存储器，可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器，处理器执行存储器中的指令，结合其硬件完成上述方法的步骤。为避免重复，这里不再详细描述。

本文中术语“和/或”，仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况。另外，本文中字符“/”，一般表示前后关联对象是一种“或”的关系。

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实施例的范围。

所属领域的技术人员可以清楚地了解到，为了描述的方便和简洁，上述描述的系统、物理主机和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

在本发明实施例所提供的几个实施例中，应该理解到，所揭露的系统、物理主机和方法，可以通过其它的方式实现。例如，以上所描述的物理主机的实施例仅是示意性的，例如，该单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另外，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口、物理主机或单元的间接耦合或通信连接，也可以是电的，机械的或其它的形式连接。

作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本发明实施例的方案的目的。

另外，在本发明实施例各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以是两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

该集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个计算机可读存储介质中。基于这样的理解，本发明实施例的技术方案本质上或者说对现有技术做出贡献的部分，或者该技术方案的全部或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计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本发明实施例各个实施例的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（Read-Only Memory，

ROM)、随机存取存储器(Random Access Memory, RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

5 以上所述,仅为本发明实施例的具体实施方式,但本发明实施例的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本发明实施例揭露的技术范围内,可轻易想到各种等效的修改或替换,这些修改或替换都应涵盖在本发明实施例的保护范围之内。因此,本发明实施例的保护范围应以权利要求的保护范围为准。

权 利 要 求 书

1、一种转发报文的方法，其特征在于，应用于物理主机，所述物理主机上包括第一虚拟交换机和至少两个虚拟机；

5 所述至少两个虚拟机中的每个虚拟机上配置有能够与所述物理主机共同访问的共享内存区，每个共享内存区配置有第一内存池，每个第一内存池配置有至少一个内存块，所述每个第一内存池中的内存块配置有用于标识所属虚拟机的索引字段，所述至少两个虚拟机中的第一虚拟机对应的第一共享内存区禁止被所述至少两个虚拟机中除所述第一虚拟机之外的其它虚拟机访问；

10 所述方法包括：

所述第一虚拟交换机获取存有第一报文的第一内存块，所述第一内存块携带有用于标识所述第一内存块的所属虚拟机的索引字段；

所述第一虚拟交换机根据所述第一内存块的索引字段，转发所述第一报文。

2、根据权利要求1所述的方法，其特征在于，

15 所述每个第一内存池中的内存块包括第一内存分段和第二内存分段，所述第一内存分段在所述第二内存分段之前，所述第一内存分段用于存储所述第二内存分段的地址信息，所述第二内存分段用于存储报文；

20 所述物理主机上配置有至少两个第二内存池，所述至少两个第二内存池中的每个第二内存池和所述每个第一内存池一一对应，每个第二内存池配置有对应的第一内存池中的第一内存分段；

其中，所述第一虚拟交换机根据所述第一内存块的索引字段，转发所述第一报文，包括：

所述第一虚拟交换机根据所述至少两个第二内存池和所述第一内存块的索引字段，转发所述第一报文。

25 3、根据权利要求2所述的方法，其特征在于，所述物理主机根据所述至少两个第二内存池配置有物理网卡接收队列，所述第一虚拟机配置有发送队列和/或接收队列，所述第一虚拟机的接收队列或者发送队列包括所述第一虚拟机的第一内存池中的部分或者全部内存块；

其中，所述第一虚拟交换机获取存有第一报文的第一内存块，包括：

30 第一虚拟交换机获取所述物理网卡接收队列中的所述第一内存块，或者，
第一虚拟交换机获取所述第一虚拟机的发送队列中的所述第一内存块。

4、根据权利要求3所述的方法，其特征在于，所述第一虚拟交换机用于向所述第一虚拟机转发报文；

35 其中，所述第一虚拟交换机根据所述至少两个第二内存池和所述第一内存块的索引字段，转发所述第一报文，包括：

若所述第一内存块的索引字段标识的虚拟机为所述第一虚拟机，所述第一虚拟交换机从所述第一虚拟机的第一内存池中调度一个内存块作为第二内存块；

所述第一虚拟交换机将所述第一内存块的地址信息赋予所述第二内存块；

所述第一虚拟交换机将所述第二内存块填充到所述第一虚拟机的接收队列。

5、根据权利要求 3 或 4 所述的方法，其特征在于，所述第一虚拟交换机用于向所述第一虚拟机转发报文；

5 其中，所述第一虚拟交换机根据所述至少两个第二内存池和所述第一内存块的索引字段，转发所述第一报文，包括：

若所述第一内存块的索引字段标识的虚拟机不是所述第一虚拟机，所述第一虚拟交换机从所述至少两个第二内存池中调度一个内存块作为第三内存块，所述第三内存块的索引字段标识的虚拟机为所述第一虚拟机；

所述第一虚拟交换机将所述第一内存块的内容拷贝到所述第三内存块；

10 所述第一虚拟交换机从所述第一虚拟机的第一内存池中调度一个内存块作为第二内存块；

所述第一虚拟交换机将所述第三内存块的地址信息，赋予所述第二内存块；

所述第一虚拟交换机将所述第二内存块填充到所述第一虚拟机的接收队列。

15 6、根据权利要求 3 所述的方法，其特征在于，所述第一虚拟交换机用于向所述物理网卡接收队列转发所述第一报文；

其中，所述第一虚拟交换机根据所述至少两个第二内存池和所述第一内存块的索引字段，转发所述第一报文，包括：

所述第一虚拟交换机从所述至少两个第二内存池中获取第四内存块，所述第四内存块的索引字段标识的虚拟机为所述第一虚拟机；

20 所述第一虚拟交换机将所述第一内存块的地址信息，赋予所述第四内存块；

所述第一虚拟交换机将所述第四内存块填充到所述物理网卡接收队列。

7、根据权利要求 4 至 6 中任一项所述的方法，其特征在于，所述地址信息包括数据长度字段的信息和数据起始位置偏移字段的信息。

25 8、根据权利要求 2 至 7 中任一项所述的方法，其特征在于，所述物理主机配置有物理网卡内存块队列，所述方法还包括：

所述物理主机根据所述至少两个第二内存池中的内存块，填充所述物理网卡内存块队列。

9、根据权利要求 8 所述的方法，其特征在于，在所述物理主机根据所述至少两个第二内存池中的内存块，填充所述物理网卡内存块队列之前，所述方法还包括：

30 所述物理主机获取所述至少两个虚拟机中每个虚拟机的流量统计信息；

所述物理主机根据所述流量统计信息，确定所述至少两个虚拟机中每个虚拟机的流量与所述至少两个虚拟机的总流量之间的流量比值；

其中，所述物理主机根据所述至少两个第二内存池中的内存块，填充所述物理网卡内存块队列，包括：

35 所述物理主机将所述第一虚拟机的流量比值确定为，填充到所述物理网卡内存块队列中的所述第一虚拟机对应的内存块数量与填充到所述物理网卡内存块队列中的内存块总数量之间的比值。

10、根据权利要求 1 至 9 中任一项所述的方法，其特征在于，所述物理主机通过为所述第一虚拟机配置元数据文件生成所述第一共享内存区，所述元数据文件包括所述第一虚

拟机的端口信息和所述第一虚拟机的第一内存池中的内存块。

11、一种物理主机，其特征在于，所述物理主机上包括第一虚拟交换机和至少两个虚拟机；

5 所述至少两个虚拟机中的每个虚拟机上配置有能够与所述物理主机共同访问的共享内存区，每个共享内存区配置有第一内存池，每个第一内存池配置有至少一个内存块，所述每个第一内存池中的内存块配置有用于标识所属虚拟机的索引字段，所述至少两个虚拟机中的第一虚拟机对应的第一共享内存区禁止被所述至少两个虚拟机中除所述第一虚拟机之外的其它虚拟机访问；

所述第一虚拟交换机用于：

10 获取存有第一报文的第一内存块，所述第一内存块携带有用于标识所述第一内存块的所属虚拟机的索引字段；

根据所述第一内存块的索引字段，转发所述第一报文。

12、根据权利要求 11 所述的物理主机，其特征在于，

15 所述每个第一内存池中的内存块包括第一内存分段和第二内存分段，所述第一内存分段在所述第二内存分段之前，所述第一内存分段用于存储所述第二内存分段的地址信息，所述第二内存分段用于存储报文；

所述物理主机上配置有至少两个第二内存池，所述至少两个第二内存池中的每个第二内存池和所述每个第一内存池一一对应，每个第二内存池配置有对应的第一内存池中的第一内存分段；

20 其中，所述第一虚拟交换机具体用于：根据所述至少两个第二内存池和所述第一内存块的索引字段，转发所述第一报文。

13、根据权利要求 12 所述的物理主机，其特征在于，所述物理主机根据所述至少两个第二内存池配置有物理网卡接收队列，所述第一虚拟机配置有发送队列和/或接收队列，所述所述第一虚拟机的接收队列或者发送队列包括所述第一虚拟机的第一内存池中的部分或

25 者全部内存块；

其中，第一虚拟交换机具体用于：获取所述物理网卡接收队列中的所述第一内存块，或者，

获取所述第一虚拟机的发送队列中的所述第一内存块。

14、根据权利要求 13 所述的物理主机，其特征在于，所述第一虚拟交换机用于向所

30 述第一虚拟机转发报文；

其中，若所述第一内存块的索引字段标识的虚拟机为所述第一虚拟机，所述第一虚拟交换机从所述第一虚拟机的第一内存池中调度一个内存块作为第二内存块；

所述第一虚拟交换机用于：

将所述第一内存块的地址信息赋予所述第二内存块；

35 将所述第二内存块充填到所述第一虚拟机的接收队列。

15、根据权利要求 13 或 14 所述的物理主机，其特征在于，所述第一虚拟交换机用于向所述第一虚拟机转发报文；

其中，若所述第一内存块的索引字段标识的虚拟机不是所述第一虚拟机，所述第一虚拟交换机从所述至少两个第二内存池中调度一个内存块作为第三内存块，所述第三内存块

的索引字段标识的虚拟机为所述第一虚拟机;

所述第一虚拟交换机用于:

将所述第一内存块的内容拷贝到所述第三内存块;

从所述第一虚拟机的第一内存池中调度一个内存块作为第二内存块;

5 将所述第三内存块的地址信息, 赋予所述第二内存块;

将所述第二内存块填充到所述第一虚拟机的接收队列。

16、根据权利要求 13 所述的物理主机, 其特征在于, 所述第一虚拟交换机用于向所述物理网卡接收队列转发所述第一报文;

10 其中, 所述第一虚拟交换机用于从所述至少两个第二内存池中获取第四内存块, 所述第四内存块的索引字段标识的虚拟机为所述第一虚拟机;

所述第一虚拟交换机用于:

将所述第一内存块的地址信息, 赋予所述第四内存块;

将所述第四内存块填充到所述物理网卡接收队列。

15 17、根据权利要求 14 至 16 中任一项所述的物理主机, 其特征在于, 所述地址信息包括数据长度字段的信息和数据起始位置偏移字段的信息。

18、根据权利要求 12 至 17 中任一项所述的物理主机, 其特征在于, 所述物理主机配置有物理网卡内存块队列, 所述物理主机还用于根据所述至少两个第二内存池中的内存块, 填充所述物理网卡内存块队列。

20 19、根据权利要求 18 所述的物理主机, 其特征在于, 在所述物理主机根据所述至少两个第二内存池中的内存块, 填充所述物理网卡内存块队列之前;

所述物理主机用于:

获取所述至少两个虚拟机中每个虚拟机的流量统计信息;

根据所述流量统计信息, 确定所述至少两个虚拟机中每个虚拟机的流量与所述至少两个虚拟机的总流量之间的流量比值;

25 将所述第一虚拟机的流量比值确定为, 填充到所述物理网卡内存块队列中的所述第一虚拟机对应的内存块数量与填充到所述物理网卡内存块队列中的内存块总数量之间的比值。

30 20、根据权利要求 11 至 19 中任一项所述的物理主机, 其特征在于, 所述物理主机还用于通过为所述第一虚拟机配置元数据文件生成所述第一共享内存区, 所述元数据文件包括所述第一虚拟机的端口信息和所述第一虚拟机的第一内存池中的内存块。

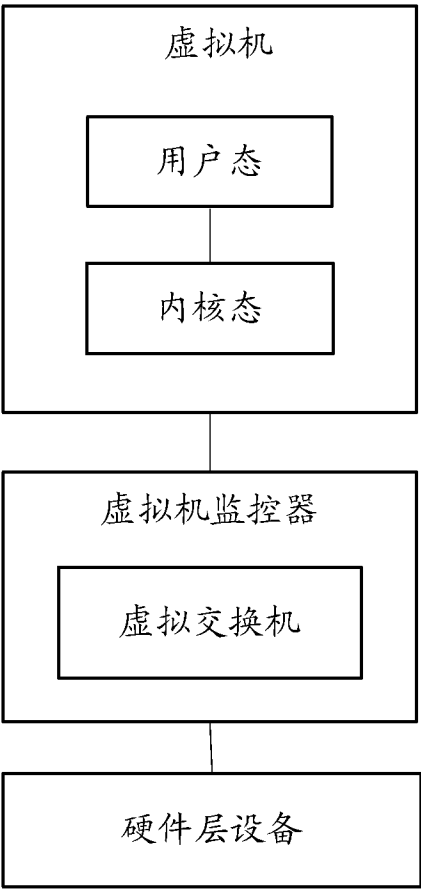


图 1

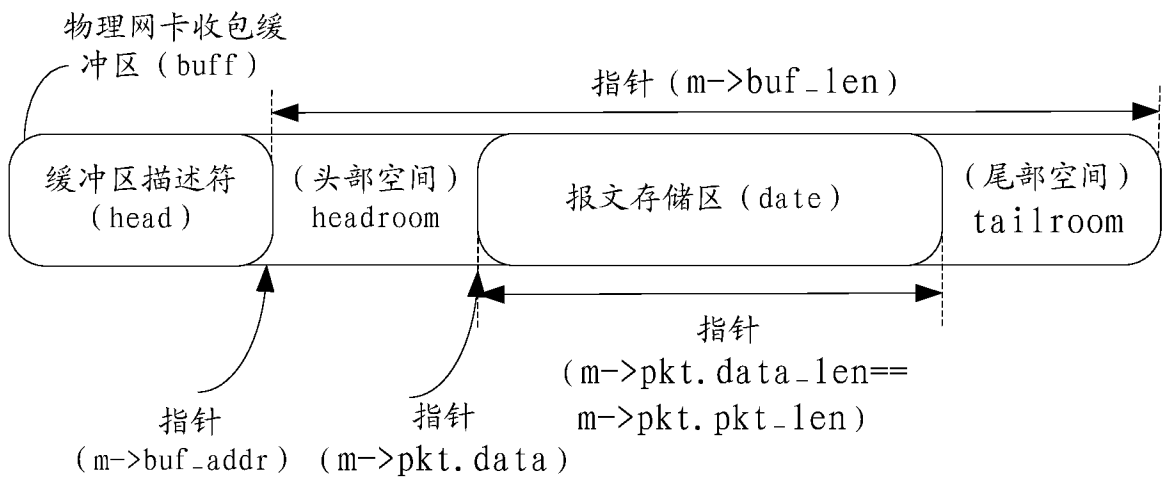


图 2

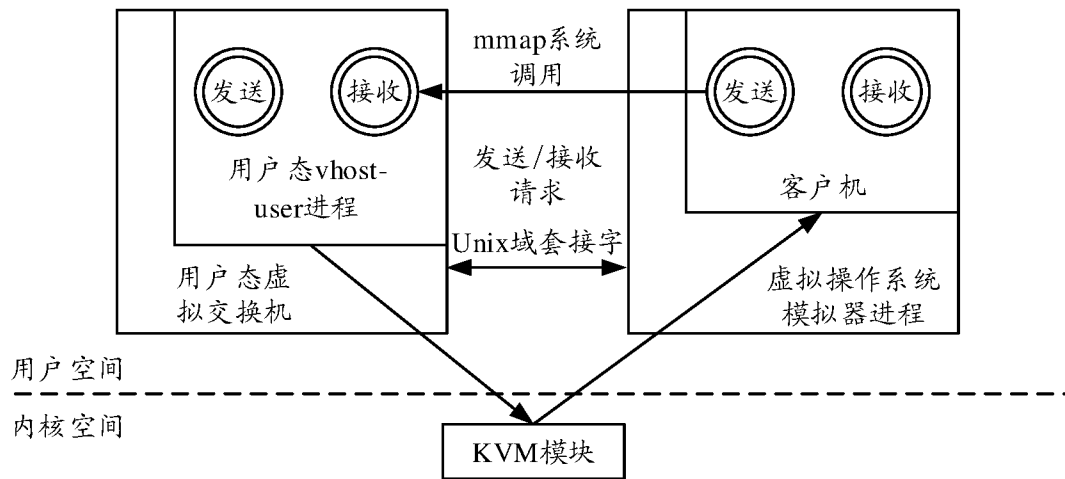


图 3

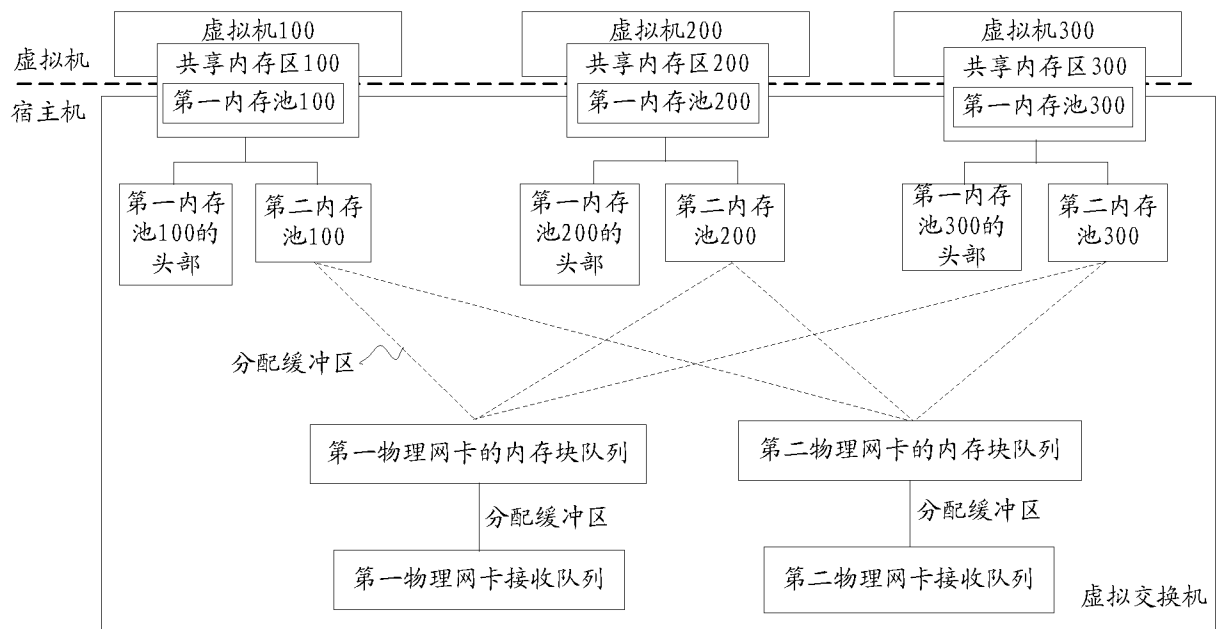


图 4

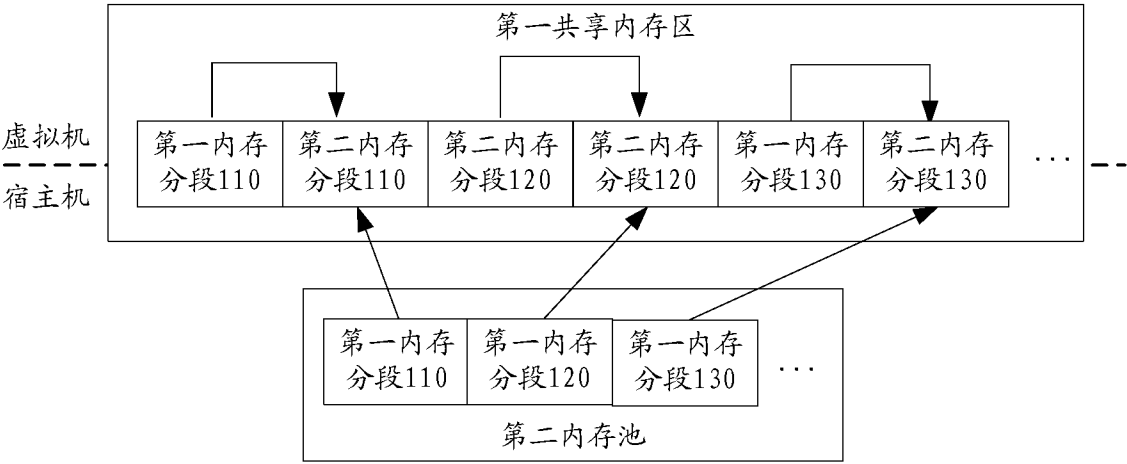


图 5

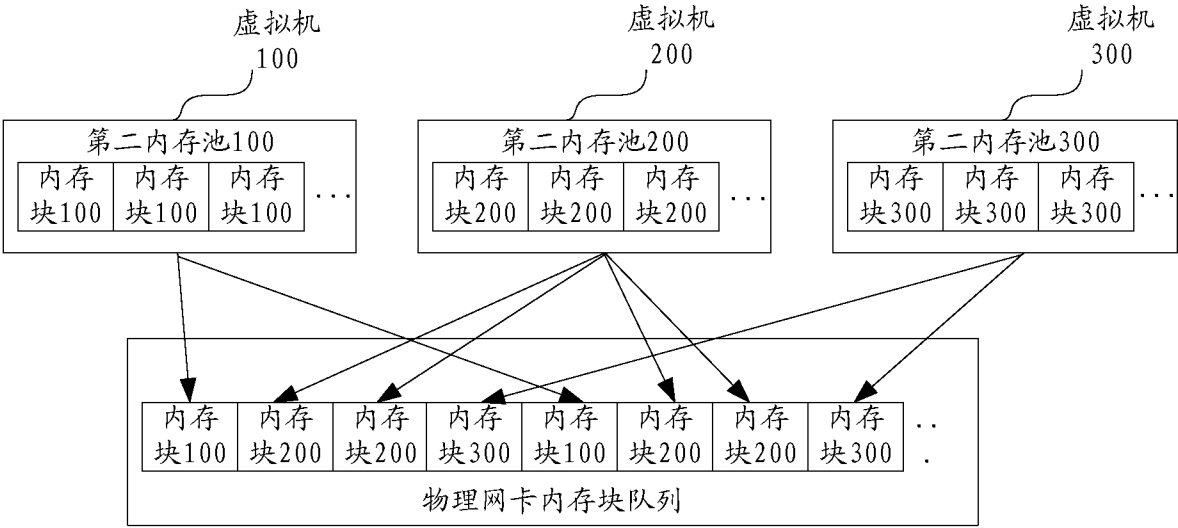


图 6

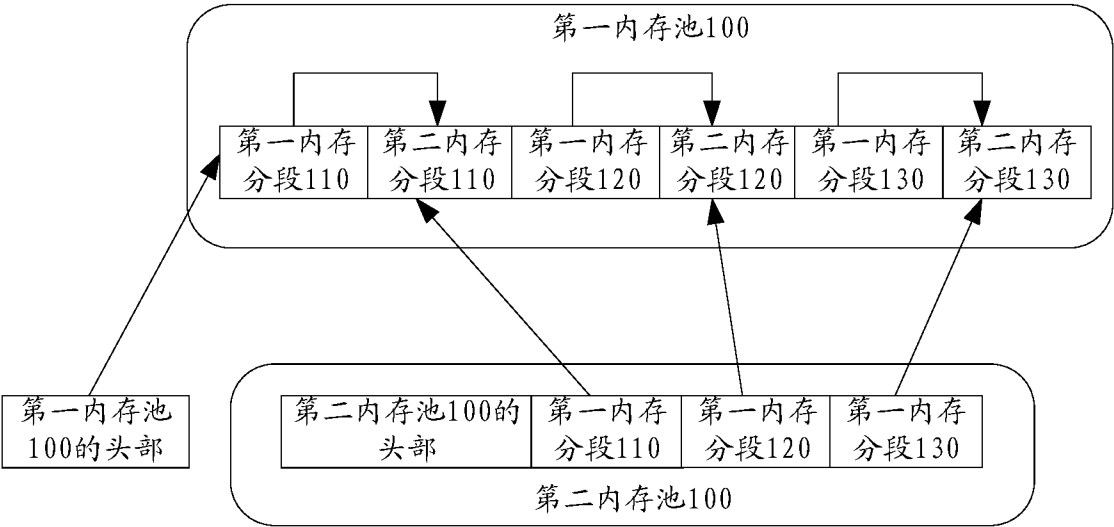


图 7

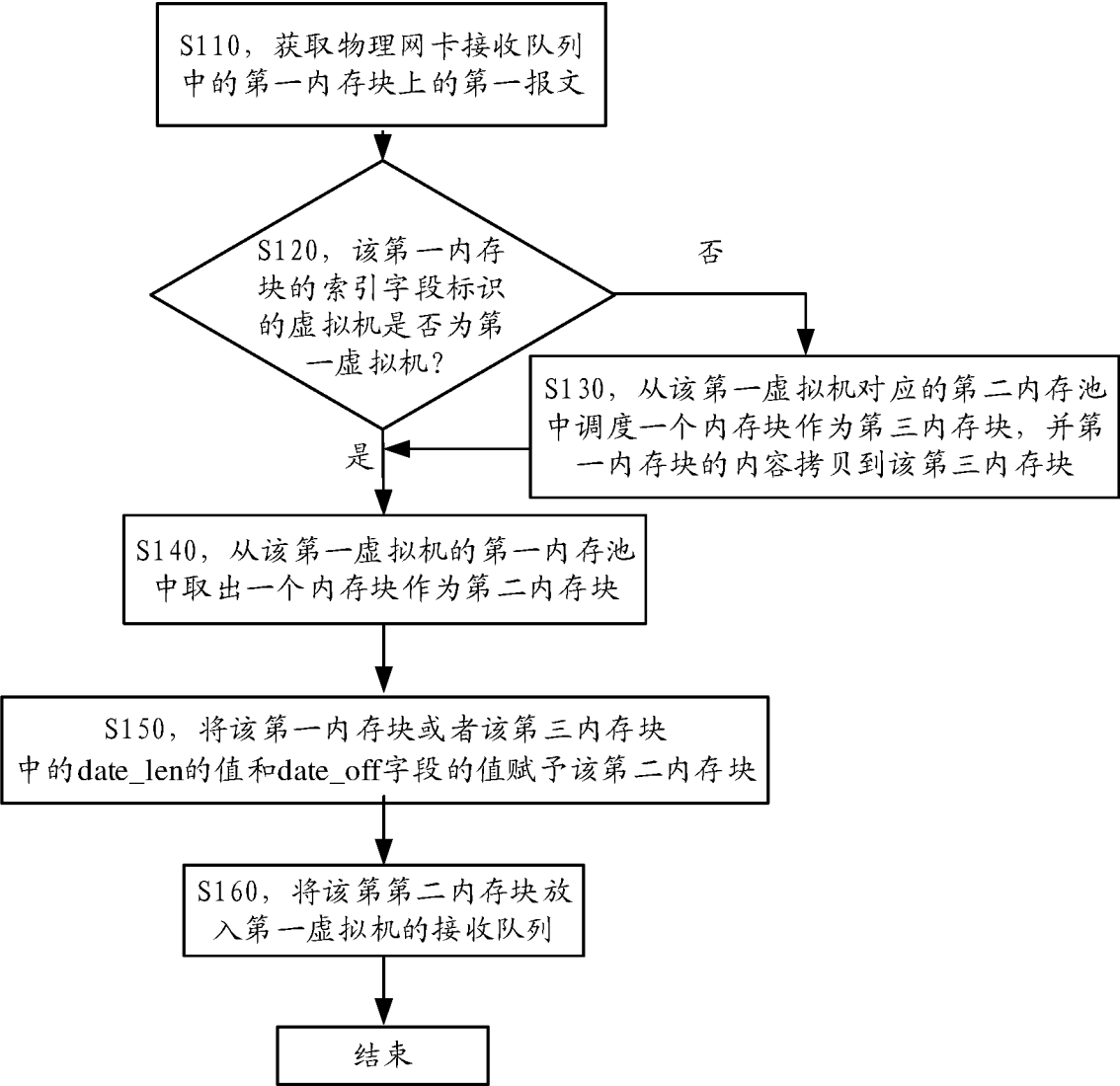


图 8

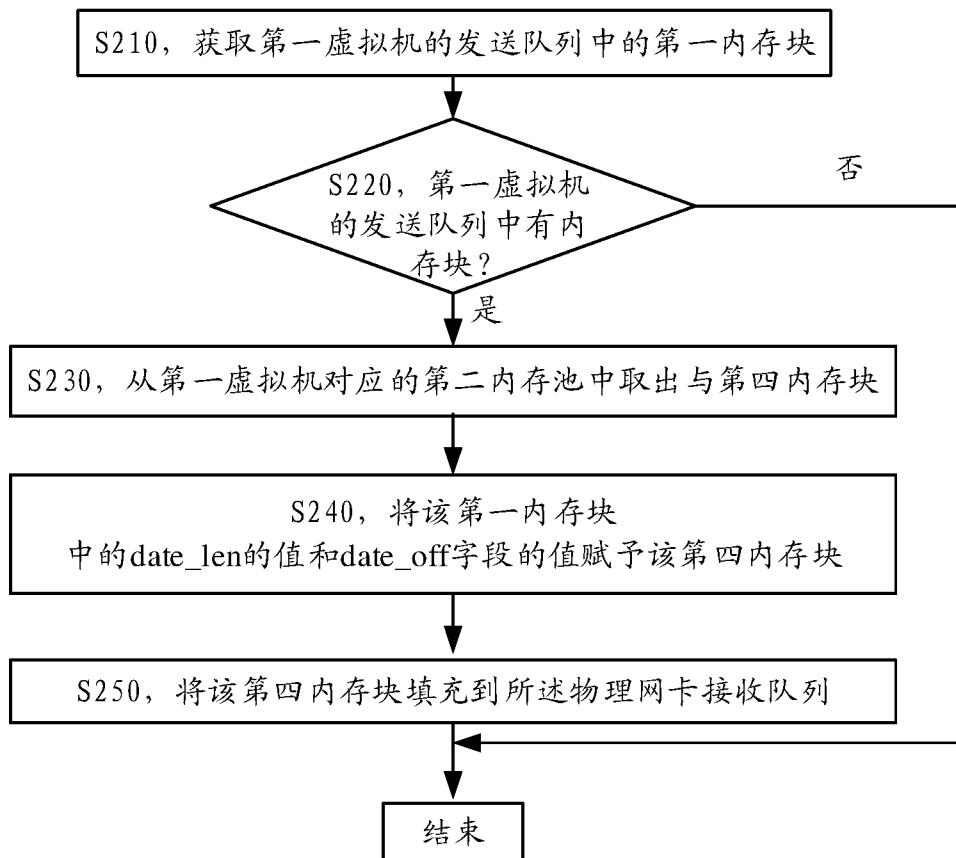


图 9

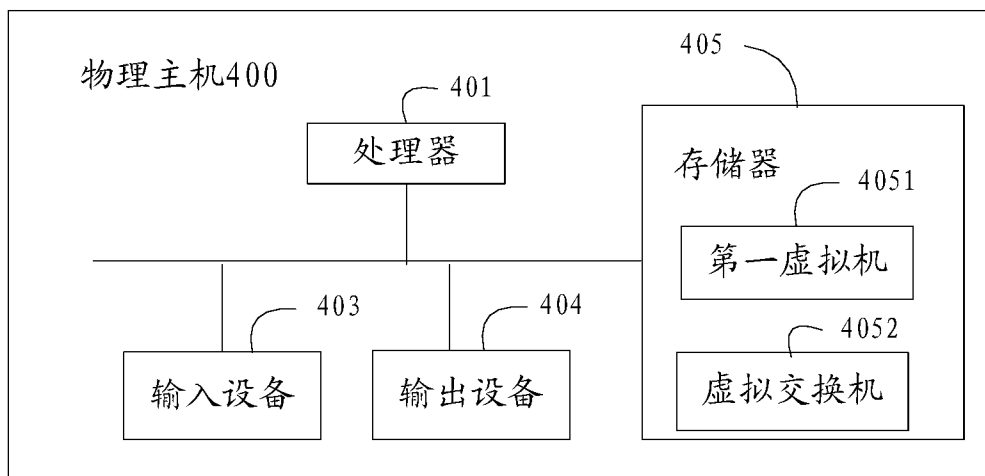


图 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/105823

A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/861 (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; G06F

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)

CPRSABS; CNTXT; SIPOABS; VEN: 虚拟交换机, 虚拟机, 共享内存区, 内存池, 内存块, 索引字段, 转发, 报文, 物理主机,
virtual switch, virtual machine, share section, memory pool, memory block, index field, forward, message, physical host

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 102004675 A (FUJIAN STAR-NET RUIJIE NETWORKS CO., LTD.) 06 April 2011 (06.04.2011), claims 1-20	1-20
A	CN 106161110 A (NEUSOFT GROUP CORP.) 23 November 2016 (23.11.2016), entire document	1-20
A	WO 2007075258 A2 (CISCO TECHNOLOGY, INC.) 05 July 2007 (05.07.2007), entire document	1-20

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search 23 November 2017	Date of mailing of the international search report 07 December 2017
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 XU, Gang Telephone No. (86-10) 62411238

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/105823

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102004675 A	06 April 2011	None	
CN 106161110 A	23 November 2016	None	
WO 2007075258 A2	05 July 2007	US 2007147279 A1	28 June 2007
		EP 1967035 A4	10 April 2013
		CN 101347038 B	01 February 2012
		US 7821981 B2	26 October 2010
		EP 1967035 A2	10 September 2008
		CN 101347038 A	14 January 2009
		WO 2007075258 A3	13 December 2007

国际检索报告

国际申请号

PCT/CN2017/105823

A. 主题的分类

H04L 12/861(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; G06F

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))

CPRSABS;CNTXT;SIPOABS;VEN:虚拟交换机, 虚拟机, 共享内存区, 内存池, 内存块, 索引字段, 转发, 报文, 物理主机, virtual switch, virtual machine, share section, memory pool, memory block, index field, forward, message, physical host

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	CN 102004675 A (福建星网锐捷网络有限公司) 2011年 4月 6日 (2011 - 04 - 06) 权利要求1-20	1-20
A	CN 106161110 A (东软集团股份有限公司) 2016年 11月 23日 (2016 - 11 - 23) 全文	1-20
A	WO 2007075258 A2 (思科技术公司) 2007年 7月 5日 (2007 - 07 - 05) 全文	1-20

☐ 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

2017年 11月 23日

国际检索报告邮寄日期

2017年 12月 7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)
中国北京市海淀区蓟门桥西土城路6号 100088

受权官员

徐刚

传真号 (86-10)62019451

电话号码 (86-10)62411238

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2017/105823

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	102004675	A	2011年 4月 6日	无			
CN	106161110	A	2016年 11月 23日	无			
WO	2007075258	A2	2007年 7月 5日	US	2007147279	A1	2007年 6月 28日
				EP	1967035	A4	2013年 4月 10日
				CN	101347038	B	2012年 2月 1日
				US	7821981	B2	2010年 10月 26日
				EP	1967035	A2	2008年 9月 10日
				CN	101347038	A	2009年 1月 14日
				WO	2007075258	A3	2007年 12月 13日

表 PCT/ISA/210 (同族专利附件) (2009年7月)