(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104580011 B
(45)授权公告日 2017.12.15

(21)申请号 201310505563.1

(22)申请日 2013.10.23

(65)同一申请的已公布的文献号
申请公布号 CN 104580011 A

(43)申请公布日 2015.04.29

(73)专利权人 新华三技术有限公司
地址 310052 浙江省杭州市滨江区长河路466号

(72)发明人 魏初舜

(74)专利代理机构 北京德琦知识产权代理有限公司 11018
代理人 谢安昆 宋志强

(51)Int.Cl.
H04L 12/867(2013.01)
H04L 29/08(2006.01)

(56)对比文件
US 2004160903 A1,2004.08.19,
US 2003200315 A1,2003.10.23,
US 2010071025 A1,2010.03.18,
US 2013145072 A1,2013.06.06,
US 6880089 B1,2005.04.12,
US 2004125746 A1,2004.07.01,
CN 102821082 A,2012.12.12,

审查员 李珍珍

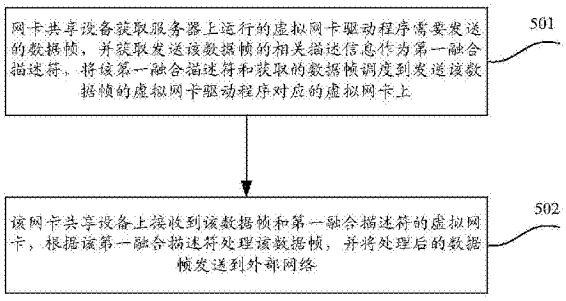
权利要求书5页 说明书20页 附图3页

(54)发明名称

一种数据转发装置和方法

(57)摘要

本发明公开了一种数据转发装置，应用于包括多个服务器、一个网卡共享设备的网络系统中的网卡共享设备上，在网卡共享设备上虚拟多个虚拟网卡单元，并在各服务器上运行一个或多个与虚拟网卡单元一一对应的虚拟网卡驱动程序，将各服务器发送的数据帧通过该共享网卡设备对应的虚拟网卡单元转发到外部网络。基于同样的发明构思，本申请还提出一种方法，能够使多台服务器之间共享网卡资源。



CN 104580011 B

1.一种数据转发装置，其特征在于，应用于包括多个服务器、一个网卡共享设备的网络系统中的网卡共享设备上，该装置包括：多个服务器接口单元、队列池及调度单元、多个虚拟网卡单元和网络接口单元；

所述服务器接口单元，用于获取对应的服务器上运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符，将该第一融合描述符和获取的数据帧发送给队列池及调度单元；其中，第一融合描述符包含描述符类型和数据帧长度；每个服务器与一个服务器接口单元对应，每个服务器上运行一个或多个虚拟网卡驱动程序，且每个虚拟网卡驱动程序与虚拟网卡单元一一对应；

所述队列池及调度单元，用于将接收到的第一融合描述符和数据帧，调度到发送该数据帧的虚拟网卡驱动程序对应的虚拟网卡单元；

所述虚拟网卡单元，用于接收到第一融合描述符合数据帧时，根据该第一融合描述符处理该数据帧，并将处理后的数据帧发送给所述网络接口单元；

所述网络接口单元，用于将从所述虚拟网卡单元接收到的数据帧转发到外部网络。

2.根据权利要求1所述的装置，其特征在于，该装置进一步包括：管理单元；

所述管理单元，用于配置数据帧的信息字段与虚拟网卡单元标识的对应关系；

所述网络接口单元，进一步用于接收外部网络发送的数据帧，根据该数据帧的信息字段与虚拟网卡单元标识的对应关系，匹配到对应的虚拟网卡单元标识，并将该数据帧发送给匹配到的虚拟网卡单元标识对应的虚拟网卡单元；

所述虚拟网卡单元，进一步用于接收到所述网络接口单元发送的数据帧时，对该数据帧进行处理，并根据处理结果为该数据帧构造第二融合描述符，并将该数据帧以及构造的第二融合描述符发送给所述队列池及调度单元；其中，该第二融合描述符包含描述符类型和数据帧长度；

所述队列池及调度单元，进一步用于将所述第二融合描述符和数据帧，调度到对应的服务器接口单元，该服务器接口单元与运行发送该第二融合描述符的虚拟网卡单元对应的虚拟网卡驱动程序的服务器对应；

所述服务器接口单元，进一步用于将该数据帧发送给与本服务器接口单元对应的服务器，并将第二融合描述符的内容发送给该服务器，使所述服务器上运行的对应虚拟网卡驱动程序根据接收到的第二融合描述符的内容对接收到的数据帧进行处理。

3.根据权利要求2所述的装置，其特征在于，

所述第一融合描述符的内容还包含下述之一或任意组合：

数据帧的帧格式、是否添加或修改网络节点标识ID信息、是否重新计算校验和、是否进行加密、指导虚拟网卡单元如何处理数据帧的其它信息；

所述第二融合描述符的内容还包含下述之一或任意组合：

数据帧的帧格式、是否出现错误、虚拟网卡单元对数据帧的字段的判断结果，虚拟网卡单元提取或丢弃数据帧中的信息、虚拟网卡单元对数据帧进行的修改、虚拟网卡单元是否完成解密、虚拟网卡单元发现或已经处理的其它信息。

4.根据权利要求2所述的装置，其特征在于，

所述管理单元，进一步用于对所述队列池及调度单元中的各队列配置传输速率、优先级、调度策略和各队列的当前状态；

所述队列池及调度单元，进一步用于配置多个队列；根据所述管理单元对各个队列的配置丢弃部分数据帧，或将数据帧调度到对应的服务器接口单元或虚拟网卡单元。

5.根据权利要求2-4任意一项所述的装置，其特征在于，

所述服务器接口单元，用于与对应的服务器通过外设组件互连标准的总线接口PCI Express点到点连接时，作为PCI Express链路的下游端点，配置多个发送引擎和接收引擎，并与所述队列池及调度单元中配置的队列一一对应；具体用于在对应的服务器上运行的虚拟网卡驱动程序需要发送数据帧时，根据所述发送引擎指向的当前有效的发送缓存buffer描述符从所述服务器内存中读取需要发送的数据帧，将该发送buffer描述符中除buffer空间起始地址之外的内容构造为第一融合描述符，并将该第一融合描述符和该数据帧写入所述队列池及调度单元中对应的队列；通过接收引擎读取服务器当前有效的接收buffer描述符，在该接收引擎对应的队列中有数据帧和第二融合描述符时，读取第二融合描述符，根据该第二融合描述符读取其后的数据帧，并将该数据帧写入所述接收buffer描述符所指的服务器的buffer中，并在回写所述接收buffer描述符时进一步携带该第二融合描述符的内容。

6.根据权利要求2-4任意一项所述的装置，其特征在于，

所述服务器接口单元，用于与对应的服务器通过以太网点到点连接时，配置多个发送引擎和接收引擎，并与所述队列池及调度单元中的配置的队列一一对应；具体用于发送引擎接收到服务器发送的描述符和数据帧时，将该描述符的格式变换为第一融合描述符的格式作为第一融合描述符，并将该第一融合描述符和该数据帧发送给所述队列池及调度单元中对应的队列；在所述接收引擎对应的队列中有数据帧和第二融合描述符时，读取第二融合描述符，根据该第二融合描述符读取其后的数据帧，将第二融合描述符以及读取的数据帧发送给对应的服务器，使该服务器上运行的对应虚拟网卡驱动程序进一步处理所述第二融合描述符和数据帧。

7.根据权利要求2-4任意一项所述的装置，其特征在于，

所述服务器接口单元，用于获取对应的服务器上的虚拟机VM运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符，将该第一融合描述符和获取的数据帧发送给所述队列池及调度单元；将接收到的数据帧，以及第二融合描述符的内容，发送给发送该数据帧的虚拟网卡单元对应虚拟网卡驱动程序运行的VM，使该VM上运行的对应虚拟网卡驱动程序根据接收到的第二融合描述符的内容对接收到的数据帧进行处理；其中，各服务器上通过虚拟机管理程序VMM实现多个虚拟机VM的虚拟环境，每个VM上运行一个或多个虚拟网卡驱动程序，且每个虚拟网卡驱动程序与虚拟网卡单元一一对应。

8.根据权利要求7所述的装置，其特征在于，

所述管理单元，进一步用于获知任一VM迁移时，停止所述VM上运行的虚拟网卡驱动程序对应的虚拟网卡单元接收外部网络发送数据帧的功能；使队列池及调度单元完成数据帧的收发，并停止接收功能的虚拟网卡单元上的相关内容复制到目的虚拟网卡单元，所述目的虚拟网卡单元为迁移后的VM上运行的虚拟网卡驱动程序对应的虚拟网卡单元；当所述迁移后的VM上的虚拟网卡驱动程序启动时，启动所述目的虚拟网卡单元的收发功能。

9.根据权利要求1-4任意一项所述的装置，其特征在于，该装置还包括：一个或多个共

享加速单元；

所述共享加速单元，用于接收到运行该共享加速单元对应的共享设备驱动程序的服务器发送的数据帧时，对该数据帧根据配置进行加速处理，并将处理结果返回发送该数据帧的服务器；若具有网络通信功能时，将处理结果发送给网络接口单元或返回发送该数据帧的服务器。

10.一种数据转发方法，其特征在于，应用于包括多个服务器、一个网卡共享设备的网络系统中的网卡共享设备上，所述网卡共享设备上虚拟多个虚拟网卡；每个服务器上运行一个或多个虚拟网卡驱动程序，且每个虚拟网卡驱动程序与共享网卡设备上的虚拟网卡一一对应；所述方法包括：

获取所述服务器上运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符，将该第一融合描述符和获取的数据帧调度到发送该数据帧的虚拟网卡驱动程序对应的虚拟网卡上；其中，第一融合描述符包含描述符类型和数据帧长度；

接收到该数据帧和第一融合描述符的虚拟网卡，根据该第一融合描述符处理该数据帧，并将处理后的数据帧发送到外部网络。

11.根据权利要求10所述的方法，其特征在于，所述方法进一步包括：配置数据帧的信息字段与虚拟网卡标识的对应关系；

接收到外部网络发送的数据帧时，根据该数据帧的信息字段匹配到对应的虚拟网卡标识，并将该数据帧发送给匹配到的虚拟网卡标识对应的虚拟网卡；

接收到该数据帧的虚拟网卡对该数据帧进行处理，根据处理结果为该数据帧构造第二融合描述符，并将该数据帧以及构造的第二融合描述符的内容发送给对应的服务器，使所述服务器上运行的对应虚拟网卡驱动程序根据接收到的第二融合描述符的内容对接收到的数据帧进行处理。

12.根据权利要求11所述的方法，其特征在于，

所述第一融合描述符的内容还包含下述之一或任意组合：

数据帧的帧格式、是否添加或修改网络节点ID信息、是否重新计算校验和、是否进行加密、指导虚拟网卡单元如何处理数据帧的其它信息；

所述第二融合描述符的内容还包含下述之一或任意组合：

数据帧的帧格式、是否出现错误、虚拟网卡单元对数据帧的字段的判断结果，虚拟网卡单元提取或丢弃数据帧中的信息、虚拟网卡单元对数据帧进行的修改、虚拟网卡单元是否完成解密、虚拟网卡单元发现或已经处理的其它信息。

13.根据权利要求11所述的方法，其特征在于，所述方法进一步包括：

根据预先配置的传输速率、优先级、调度策略，对接收到的数据帧进行部分丢弃处理，或将接收到的数据帧调度给服务器或虚拟网卡。

14.根据权利要求11-13任意一项所述的方法，其特征在于，

本网卡共享设备与所述服务器通过外设组件互连标准的总线接口PCI Express点到点连接时，作为PCI Express链路的下游端点，所述方法进一步包括：配置多个发送引擎和接收引擎，并分别对应一个队列；

所述获取服务器上运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧

的相关描述信息作为第一融合描述符，将该第一融合描述符和获取的数据帧调度到发送该数据帧的虚拟网卡驱动程序对应的虚拟网卡上，包括：根据所述发送引擎指向的当前有效的发送缓存buffer描述符从所述服务器内存中读取需要发送的数据帧，将该发送buffer描述符中除buffer空间起始地址之外的内容构造为第一融合描述符，并将该第一融合描述符和获取的数据帧通过对应的队列调度到对应的虚拟网卡上；

所述将该数据帧以及构造的第二融合描述符的内容发送给对应的服务器，包括：通过接收引擎读取服务器当前有效的接收buffer描述符，在该接收引擎对应的队列中有数据帧和第二融合描述符时，读取第二融合描述符，根据该第二融合描述符读取其后的数据帧，并将该数据帧写入所述接收buffer描述符所指的服务器的buffer中，并在回写所述接收buffer描述符时进一步携带该第二融合描述符的内容。

15.根据权利要求11-13任意一项所述的方法，其特征在于，

本网卡共享设备与所述服务器通过以太网点到点连接时，所述方法进一步包括：配置多个发送引擎和接收引擎，并分别对应一个队列；

所述获取服务器上运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符，包括：发送引擎接收到服务器发送的描述符和数据帧时，将该描述符的格式变换为第一融合描述符的格式作为第一融合描述符，并将该第一融合描述符和该数据帧发送给对应的队列；

所述将该数据帧以及构造的第二融合描述符的内容发送给对应的服务器，包括：通过所述接收引擎在该接收引擎对应的队列中有数据帧和第二融合描述符时，读取第二融合描述符，根据该第二融合描述符读取其后的数据帧，将第二融合描述符以及该数据帧发送给对应的服务器。

16.根据权利要求11-13任意一项所述的方法，其特征在于，

所述获取服务器上运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符，包括：获取服务器上的虚拟机VM运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符；

所述将该数据帧以及构造的第二融合描述符的内容发送给对应的服务器，使所述服务器上运行的对应虚拟网卡驱动程序根据接收到的第二融合描述符的内容对接收到的数据帧进行处理，包括：将该数据帧，以及第二融合描述符的内容发送给，发送该数据帧的虚拟网卡对应的虚拟网卡驱动程序运行的VM，使所述VM上运行的对应虚拟网卡驱动程序根据接收到的第二融合描述符的内容对接收到的数据帧进行处理；

其中，各服务器上通过虚拟机管理软件VMM实现多个虚拟机VM的虚拟环境，各VM上运行一个或多个虚拟网卡驱动程序，且每个虚拟网卡驱动程序与虚拟网卡单元一一对应。

17.根据权利要求16所述的方法，其特征在于，所述方法进一步包括：

获知任一VM迁移时，停止所述VM上运行的虚拟网卡驱动程序对应的虚拟网卡接收外部网络发送数据帧的功能；完成对已接收到的数据帧的转发，并将停止接收功能的虚拟网卡上的相关内容复制到目的虚拟网卡，所述目的虚拟网卡为迁移后的VM上运行的虚拟网卡驱动程序对应的虚拟网卡；

当所述迁移后的VM上的虚拟网卡驱动程序启动后，启动所述目的虚拟网卡的收发功能。

18.根据权利要求10-13任意一项所述的方法，其特征在于，所述方法进一步包括：

接收到运行共享设备驱动程序的服务器发送的数据帧时，对该数据帧根据配置进行加速处理，并将处理结果返回发送该数据帧的服务器；若具有网络通信功能时，将处理结果发送给外部网络或返回发送该数据帧的服务器。

一种数据转发装置和方法

技术领域

[0001] 本发明涉及通信技术领域，特别涉及一种数据转发装置和方法。

背景技术

[0002] 10G以太网的出现，加之行业为满足存储和集群互联需求而对以太网协议进行的关键扩展，数据中心桥接协议(DCB)得到了发展，包括流量优先级控制，带宽管理和拥塞管理，可将使现有结构融合为基于以太网的统一整合型网络结构。这种结构将提供对其所支持的存储和计算处理资源的无缝访问。

[0003] 在每台服务器上安装一块融合局域网(LAN)/存储网络(SAN)/进程间通信(IPC)数据流的融合网络适配器(CNA)。与普通网络适配器一样，融合网络适配器直接焊接在服务器主板上，或独立设计为一块插卡并通过服务器上的主板插槽，如PCI Express Slot，与CPU/BMC紧密耦合，各服务器独立、直接管理和使用此融合网络适配器。

[0004] 由于各服务器独立配置一块融合网络适配器，无法多台服务器共享一块融合网络适配器。

[0005] 刀片服务器是指在标准高度的机架式机箱内可插装多个卡式的服务器单元，是一种实现HAHD(High Availability High Density，高可用高密度)的服务器平台，为特殊应用行业和高密度计算环境专门设计。每一块“刀片”实际上就是一块系统主板。它们可以通过"板载"硬盘启动自己的操作系统，如Windows、Linux等，类似于一个个独立的服务器，在这种模式下，每一块主板运行自己的系统，服务于指定的不同用户群，相互之间没有关联。不过，管理员可以使用系统软件将这些母板集合成一个服务器集群。在集群模式下，所有的母板可以连接起来提供高速的网络环境，并同时共享资源，为相同的用户群服务。在集群中插入新的"刀片"，就可以提高整体性能。而由于每块"刀片"都是热插拔的，所以，系统可以轻松地进行替换，并且将维护时间减少到最小。

[0006] 这种结构可以大大减少互联电缆和光纤收发器，可以大大降低由于线缆连接故障带来的隐患，提高系统可靠性。最大限度地节约服务器的使用空间和费用。每个刀片服务器独立拥有其一块或多块网卡的资源，无法在不同刀片服务器间共享。

[0007] 综上所述，现有实现中无论是使用刀片服务器还是为每台服务器配置一块网络适配器，都不能在不同服务器间共享网卡资源。

发明内容

[0008] 有鉴于此，本发明提供一种数据转发装置和方法，能够使多台服务器之间共享网卡资源。

[0009] 为解决上述技术问题，本发明的技术方案是这样实现的：

[0010] 一种数据转发装置，应用于包括多个服务器、一个网卡共享设备的网络系统中的网卡共享设备上，该装置包括：多个服务器接口单元、队列池及调度单元、多个虚拟网卡单元和网络接口单元；

[0011] 所述服务器接口单元，用于获取对应的服务器上运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符，将该第一融合描述符和获取的数据帧发送给队列池及调度单元；其中，第一融合描述符包含描述符类型和数据帧长度；每个服务器与一个服务器接口单元对应，每个服务器上运行一个或多个虚拟网卡驱动程序，且每个虚拟网卡驱动程序与虚拟网卡单元一一对应；

[0012] 所述队列池及调度单元，用于将接收到的第一融合描述符和数据帧，调度到发送该数据帧的虚拟网卡驱动程序对应的虚拟网卡单元；

[0013] 所述虚拟网卡单元，用于接收到第一融合描述符合数据帧时，根据该第一融合描述符处理该数据帧，并将处理后的数据帧发送给所述网络接口单元；

[0014] 所述网络接口单元，用于将从所述虚拟网卡单元接收到的数据帧转发到外部网络。

[0015] 一种数据转发方法，应用于包括多个服务器、一个网卡共享设备的网络系统中的网卡共享设备上，所述网卡共享设备上虚拟多个虚拟网卡；每个服务器上运行一个或多个虚拟网卡驱动程序，且每个虚拟网卡驱动程序与共享网卡设备上的虚拟网卡一一对应；所述方法包括：

[0016] 获取所述服务器上运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符，将该第一融合描述符和获取的数据帧调度到发送该数据帧的虚拟网卡驱动程序对应的虚拟网卡上；其中，第一融合描述符包含描述符类型和数据帧长度；

[0017] 接收到该数据帧和第一融合描述符的虚拟网卡，根据该第一融合描述符处理该数据帧，并将处理后的数据帧发送到外部网络。

[0018] 综上所述，本发明通过在网卡共享设备上虚拟多个虚拟网卡单元，并在各服务器上运行一个或多个与虚拟网卡单元一一对应的虚拟网卡驱动程序，将各服务器发送的数据帧通过该共享网卡设备对应的虚拟网卡单元转发到外部网络，能够使多台服务器之间共享网卡资源。

附图说明

[0019] 图1为本发明具体实施例中资源共享系统示意图；

[0020] 图2为队列池及调度单元结构示意图；

[0021] 图3为本发明实施例中服务器虚拟VM时共享资源系统示意图；

[0022] 图4为本发明实施例一中装置的硬件架构组成示意图；

[0023] 图5为本发明实施例中接收服务器发送的数据帧的处理方式流程示意图；

[0024] 图6为本发明实施例中接收到外部网络发送的数据帧的处理方式流程示意图。

具体实施方式

[0025] 为使本发明的目的、技术方案及优点更加清楚明白，以下参照附图并举实施例，对本发明所述方案作进一步地详细说明。

[0026] 本发明具体实施例中提出一种数据转发装置，应用于包括多个服务器、和一个网卡共享设备的网络系统中的网卡共享设备上。通过在网卡共享设备上虚拟多个虚拟网卡单

元，并在各服务器上运行一个或多个与虚拟网卡单元一一对应的虚拟网卡驱动程序，将各服务器发送的数据帧通过该共享网卡设备转发到外部网络，能够使多台服务器之间共享网卡资源。

[0027] 该网卡共享设备可以为在网络系统中新增的设备，也可以为网络系统中与各服务器相连的交换设备，并在该交换设备中配置多个虚拟网卡来实现即可。

[0028] 参见图1，图1为本发明具体实施例中资源共享系统示意图。该资源共享系统中包括n个服务器，一个网卡共享设备。数据转发装置应用于该网卡共享设备上。该装置包括n个服务器接口单元，与n个服务器一一对应连接、一个队列池及调度单元、m个虚拟网卡单元和1个网络接口单元，其中，n和m可以相同也可以不相同，且n、m为大于1的自然数。

[0029] 每个服务器上运行一个或多个虚拟网卡驱动程序，且每个虚拟网卡驱动程序与虚拟网卡单元一一对应；其中，任一服务器上运行的虚拟网卡驱动程序与其他服务器上运行的虚拟网卡驱动程序对应的虚拟网卡单元的标识不相同，即各服务器上运行的虚拟网卡驱动程序对应的虚拟网卡单元标识均不相同。

[0030] 以该装置接收服务器的虚拟网卡驱动程序发送数据帧，并转发给外部网络为例：

[0031] 第一步，服务器接口单元(如服务器接口单元1)，获取对应服务器(服务器1)上运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符，将该第一融合描述符和获取的数据帧发送给队列池及调度单元。

[0032] 其中，第一融合描述符至少包含描述符类型和数据帧长度。第一融合描述符还可以包含的内容为下述之一或任意组合：

[0033] 数据帧的帧格式、是否添加或修改网络节点ID信息、是否重新计算校验和、是否进行加密、指导虚拟网卡单元如何处理数据帧的其它信息。其中，数据帧的帧格式用于指示，虚拟网卡单元采用第一融合描述符支持的多种帧格式中的哪种帧格式进行发送。

[0034] 服务器目前广泛采用的高速串行IO链路有外设组件互连标准的总线接口(PCI Express)、串行快速输入输出互连总线(RapidIO)、以太网等多种点到点形式。

[0035] 从技术上讲，这些类型的服务器接口单元在服务器的配合下均可以实现一个服务器接口单元关联多个队列(具体实现在下文描述)，虚拟网卡驱动程序发送一帧数据时通过一个下行队列往虚拟网卡单元发送第一融合描述符和数据帧。虚拟网卡驱动程序接收一帧数据时通过一个上行队列从虚拟网卡单元得到一个第二融合描述符和一个数据帧。

[0036] 下面给出在不同链接方式下，服务器接口单元的具体处理过程：

[0037] 1)、当服务器接口单元与对应服务器通过PCI Express点到点连接时，作为PCI Express链路的下游端点，配置多个发送引擎和接收引擎，与队列池及调度单元中配置的队列一一对应。

[0038] 图1中的服务器包括内存、硬盘、CPU和IO接口等。服务器的IO接口作为PCI Express链路的上游端点。服务器上运行的任一虚拟网卡驱动程序需要发送数据帧时，将该数据帧放置在服务器内存中一个buffer空间，并在一个发送buffer描述符循环队列中设置一个buffer描述符，现有网卡广泛采用，不再描述。

[0039] buffer描述符的内容中除了包含buffer空间起始地址、数据帧长度等信息，还有指示网卡如何发送的信息，如本buffer的数据属于哪种帧格式、是否添加或修改网络节点ID信息、是否重新计算某个校验和、是否需要进一步加密、是否需要以TCP负荷形式发送。可

选的，进一步还可包含通过多个队列中的哪个队列进行发送。

[0040] 当对应的服务器(如服务器1)上运行的虚拟网卡驱动程序需要发送数据帧时，服务器接口单元(服务器接口单元1)根据所述发送引擎指向的当前有效的发送buffer描述符从所述服务器内存中读取需要发送的数据帧，将该发送buffer描述符中除buffer空间起始地址之外的内容构造为第一融合描述符，并将该第一融合描述符和该数据帧写入所述队列池及调度单元中对应的队列。

[0041] 具体实现时，各发送引擎有地址寄存器，其指向一个buffer描述符循环队列中的一个buffer描述符。地址寄存器初始值由驱动软件设置。发送引擎根据地址寄存器的指示通过PCI Express存储器读操作来读取的当前buffer描述符。如果读取的buffer描述符无效，即没有待发送数据，继续读取当前指向的buffer描述表项。如果读取的buffer描述符有效，准备发送。

[0042] 每读取一个有效buffer描述符，判断对应的下行队列是否有足够空间。有足够空间时，发送引擎将buffer描述符中全部或部分信息以第一融合描述符的格式写入一个下行队列；接着PCI Express存储器读操作读取buffer空间数据附加在后面。其中，无论是buffer描述符中的全部或部分信息，都不包括buffer描述符中的buffer空间起始地址。

[0043] 完成数据帧发送后，一般地触发中断并通过PCI Express存储器写操作回写buffer描述符为无效，以便指示此buffer描述符已经处理完成。接着主动更新其地址寄存器以指向buffer描述符循环队列中下一个buffer描述符。

[0044] 2)、当服务器接口单元与对应服务器通过以太网点到点连接时，配置多个发送引擎，并与所述队列池及调度单元中配置的队列一一对应。

[0045] 服务器上运行的虚拟网卡驱动程序需要发送数据帧时，通过IO接口将该数据帧，以及发送该数据帧的描述符发送给对应的服务器接口单元。指明哪个虚拟网卡驱动程序发送了该数据帧，如在描述符中携带，或在以太网帧增加一个VLAN标签，以便服务器接口单元关联到具体队列。

[0046] 服务器接口单元通过所述发送引擎接收到对应服务器发送的描述符和数据帧时，将该描述符的格式变换为第一融合描述符的格式作为第一融合描述符，并将该第一融合描述符和该数据帧发送给所述队列池及调度单元中对应的队列。

[0047] 服务器接口单元接收到的描述符中如果有buffer空间起始地址，还在变化为第一融合描述符格式时，删除描述符中包含的buffer空间起始地址。

[0048] 3)当服务器接口单元与对应服务器通过串行RapidIO点到点连接时，因为串行RapidIO既能工作在类似PCI Express的存储器读写模式，也能工作在类似以太网的消息传递模式，因此可以参考PCI Express或以太网来实现获取对应服务器上运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符，将该第一融合描述符和获取的数据帧发送给队列池及调度单元。

[0049] 第二步，队列池及调度单元将接收到的第一融合描述符和数据帧，调度到发送该数据帧的虚拟网卡驱动程序对应的虚拟网卡单元。

[0050] 该装置还可以包括：管理单元，对所述队列池及调度单元中的各队列配置传输速率、优先级、调度策略和各队列的当前状态。

[0051] 队列池及调度单元，还可以用于配置多个队列，并根据管理单元对各个队列的配

置丢弃部分数据帧，或将数据帧调度到对应的虚拟网卡单元。

[0052] 参见图2，图2为队列池及调度单元结构示意图。图2中，队列池及调度单元中配置多个队列，多个队列分为多组上下行队列，服务器上的虚拟网卡驱动程序与虚拟网卡单元之间通过上行队列和下行队列组成的双向队列池进行通信。

[0053] 一个服务器接口单元可能关联一组上下行队列，也可能关联多组上下行队列，图2中显示一个服务器接口单元关联了n组上下行队列。

[0054] 每个服务器接口单元能感知每个关联队列的状态，如每个下行和上行队列中的使用量，并根据预先设置针对每个关联队列的状态给服务器发出信息，如中断消息或流控消息。

[0055] 服务器上的虚拟网卡驱动程序发送数据帧时，对应的服务器接口单元将该数据帧和第一融合描述符通过关联的一个下行队列发送到队列调度单元，再由队列调度单元调度到对应的虚拟网卡单元。

[0056] 服务器接口单元将数据帧和第一融合描述符通过其关联的哪个下行队列发送，根据具体配置实现，如，每个虚拟网卡驱动程序对应一个双向队列，和一个虚拟网卡单元。或每个发送引擎对应一个虚拟网卡驱动程序，一个双向队列，和一个虚拟网卡单元。

[0057] 第三步，虚拟网卡单元在接收到第一融合描述符和数据帧时，根据该第一融合描述符处理该数据帧，并将处理后的数据帧发送给所述网络接口单元。

[0058] 如果发送该数据帧的虚拟网卡驱动程序在服务器1上，且该虚拟网卡驱动程序与虚拟网卡单元1对应，则第三步中的虚拟网卡单元为虚拟网卡单元1。

[0059] 虚拟网卡单元根据第一融合描述符处理数据帧，同现有实现中，根据普通描述符处理数据帧的方式一致，只是本发明实施例中的第一融合描述符不包含原普通描述符中的buffer起始地址。

[0060] 第四步，网络接口单元将从虚拟网卡单元接收到的数据帧转发到外部网络。

[0061] 该装置还可以包括管理单元，管理单元可以为不同虚拟网卡单元发送的数据帧分配不同的VLAN标识；分配结束后，可以在本地保存，也可以下发给网络接口单元进行保存。

[0062] 网络接口单元接收到虚拟网卡单元发送来的数据帧时，将数据帧发送到外部网络。可选的，根据发送该虚拟网卡单元的标识匹配对应的VLAN标识，并使用匹配到的VLAN标识为该数据帧添加对应的VLAN标签，再将该添加VLAN标签的数据帧发送到外部网络。

[0063] 可选的，管理单元也可以通过网络接口单元给外部网络发送数据帧。

[0064] 以该装置接收到外部网络发送的数据帧向服务器转发为例：

[0065] 在该实例中，该装置还包括管理单元，用于配置数据帧的信息字段与虚拟网卡单元标识的对应关系。该数据帧的信息字段可以为VLAN ID，即不同的VLAN ID对应不同的虚拟网卡单元。根据数据帧的信息字段可以获知由哪个虚拟网卡单元来处理该接收的数据帧。

[0066] 具体处理流程如下：

[0067] 第一步，网络接口单元接收到外部网络发送的数据帧，根据该数据帧的信息字段与虚拟网卡单元标识的对应关系，匹配到对应的虚拟网卡单元标识，并将该数据帧发送匹配到的虚拟网卡单元标识对应的虚拟网卡单元。

[0068] 第二步，虚拟网卡单元接收到所述网络接口单元发送的数据帧时，对该数据帧进

行处理，并根据处理结果为该数据帧构造第二融合描述符，并将该数据帧以及构造的第二融合描述符发送给所述队列池及调度单元；

[0069] 虚拟网卡单元对数据帧的处理，同现有实现中网卡对数据帧的处理。为该数据帧构造的第二融合描述符至少包含描述符类型和数据帧长度。

[0070] 该第二融合描述符的内容还包含下述之一或任意组合：

[0071] 数据帧的帧格式、是否出现错误、虚拟网卡单元对数据帧的字段的判断结果，虚拟网卡单元提取或丢弃数据帧中的信息、虚拟网卡单元对数据帧进行的修改、虚拟网卡单元是否完成解密、虚拟网卡单元发现或已经处理的其它信息。

[0072] 第三步，队列池及调度单元，将接收到的第二融合描述符和数据帧，调度到对应的服务器接口单元。

[0073] 该服务器接口单元与运行发送该第二融合描述符的虚拟网卡单元对应的虚拟网卡驱动程序的服务器对应。

[0074] 若管理单元对所述队列池及调度单元中的各队列配置传输速率、优先级、调度策略和各队列的当前状态，队列池及调度单元，该队列池及调度单元还可以用于根据管理单元对各个队列的配置丢弃部分数据帧，或将数据帧调度到对应的服务器接口单元。

[0075] 如图2中，队列池及调度单元可以通过对应的上行队列将该数据帧和第二融合描述符调度到对应的服务器接口单元。

[0076] 第四步，服务器接口单元，将该数据帧发送给与本服务器接口单元对应的服务器中，并将第二融合描述符的内容发送给该服务器，使所述服务器上运行的对应虚拟网卡驱动程序根据接收到的第二融合描述符的内容对接收到的数据帧进行处理。

[0077] 服务器目前广泛采用的高速串行IO链路有PCI Express、串行RapidIO、以太网等多种点到点形式。下面详细描述通过不同方式连接时，服务器接口单元的具体处理过程：

[0078] 1)、当服务器接口单元与所述服务器通过PCI Express点到点连接时，作为PCI Express链路的下游端点，配置多个接收引擎，与所述队列池及调度单元中的队列对应。

[0079] 接收引擎主动将上行队列中数据帧搬移到服务器的内存中一个个buffer空间，与一般的网络数据接收机制类似：

[0080] 服务器上运行的虚拟网卡驱动程序需要在服务器内存中预留一组或多组buffer空间，对应地在一个或多个接收buffer描述符循环队列中的设置一组buffer描述符，现有网卡广泛采用，不再描述，每个buffer描述符其内包含buffer空闲标识、buffer空间起始地址、buffer长度等信息。可选的，进一步通过多个队列中的哪个队列进行接收。

[0081] 服务器接口单元通过接收引擎读取对应的服务器当前有效的接收buffer描述符，在该接收引擎对应的队列中有数据帧和第二融合描述符时，读取第二融合描述符，根据该第二融合描述符读取其后的数据帧，并将该数据帧写入所述接收buffer描述符所指的服务器的buffer中，并在回写所述接收buffer描述符时进一步携带该第二融合描述符的内容。

[0082] 具体实现时，接收引擎有地址寄存器，其指向一个buffer描述符循环队列中的一个buffer描述符。地址寄存器初始值由驱动软件设置。接收引擎根据地址寄存器的指示通过PCI Express存储器读操作来读取的当前buffer描述符。如果读取的buffer描述符无效，即非空闲buffer，继续读取当前指向的buffer描述符。如果读取的buffer描述符有效，即空闲buffer，准备接收。

[0083] 每读取一个有效buffer描述符，判断对应的上行行队列是否有数据可以读取。有数据时，接收引擎先读取一个第二融合描述符；根据第二融合描述符接着读取附加在后面数据帧，通过PCI Express存储器写操作写入buffer描述符指示的服务器内存中的一个buffer空间。

[0084] 完成数据帧接收后，一般地触发中断并PCI Express存储器写操作回写buffer描述符以便指示此buffer描述符已经处理完成，指示对应buffer为非空闲状态。回写的buffer描述符进一步携带第二融合接收描述符中部分或全部信息，接着主动更新其地址寄存器指向下一个buffer描述符。

[0085] 2)、当服务器接口单元与对应服务器通过以太网点到点连接时，配置多个接收引擎，与所述队列池及调度单元中的队列对应。

[0086] 服务器接口单元，用于在所述接收引擎对应的队列中有数据帧和第二融合描述符时，读取第二融合描述符，根据该第二融合描述符读取其后的数据帧，将第二融合描述符以及该数据帧发送给对应的服务器，使该服务器上运行的对应虚拟网卡驱动程序进一步处理所述第二融合描述符和数据帧。

[0087] 服务器上运行的对应虚拟网卡驱动程序根据接收到的第二融合描述符的内容处理接收到的数据帧。其中，所述对应的虚拟网卡驱动程序，为发送该数据帧的虚拟网卡单元对应的虚拟网卡驱动程序。

[0088] 可选的，所述第二融合描述符格式与服务器IO接口接收数据帧时的描述符格式可能不同，服务器上运行的对应虚拟网卡驱动程序同时结合这两个描述符来处理接收到的数据帧。

[0089] 可选的，指明哪个虚拟网卡驱动程序需要接收该数据帧，如在第二融合描述符中携带，或在以太网帧中增加一个VLAN标签，以便服务器关联到多个虚拟网卡驱动程序的一个。

[0090] 3)、当服务器接口单元与对应服务器通过串行RapidIO点到点连接时，因为串行RapidIO既能工作在类似PCI Express的存储器读写模式，也能工作在类似以太网的消息传递模式，因此可以参考PCI Express或以太网来实现将数据帧发送给与本服务器接口单元对应的服务器中，并将第二融合描述符的内容发送给该服务器，使所述服务器上运行的对应虚拟网卡驱动程序根据接收到的第二融合描述符的内容对接收到的数据帧进行处理。

[0091] 可选的，管理单元也通过网络接口单元接收到外部网络发送的数据帧。如通过网络接口单元接收到管理设备发送的控制报文，进行处理后，再通过网络接口单元响应管理设备。

[0092] 下面以具体实施例详细描述第一融合描述符和第二融合描述符的实现。

[0093] 第一融合描述符指示虚拟网卡单元如何发送一个数据帧。一般至少包含描述符的类型和数据帧的长度信息。

[0094] 还可以包括如下具体信息：

[0095] 数据帧的帧格式，也可称为网络格式，如以太网、光纤通道(FC)、因特网小型计算机系统接口(iSCSI)、快速输入输出互连总线(RapidIO)、多并发互连总线(Infiniband)、远程直接存储器访问(RDMA)等；

[0096] 是否添加或修改网络节点ID信息、如以太网的MAC地址和VLAN标签等；

[0097] 是否重新计算校验和，如以太网帧的FCS、IP头校验和、TCP校验和、UDP校验和等；

[0098] 是否进行加密，如IPSec安全联盟信息等；

[0099] 指导虚拟网卡单元如何处理数据帧的其它信息等其他可选信息。

[0100] 为了更清楚、简洁表示各种网络发送和接收特征，第一融合描述符的类型不止一种。

[0101] 第一融合描述符可以统一定义多种可以区分的格式，如，针对以太网和FCoE定义2种格式、针对RapidIO定义1种格式、针对Infiniband定义1种格式、针对RDMA定义1种格式等。

[0102] 举例如下，其在64比特普通描述符基础上进一步定义扩展类型，普通发送描述符适用于普通以太网帧发送；各种扩展发送描述符适用于其它帧格式或指示虚拟网卡分担更多功能。

[0103] (1)普通发送描述符，适用于普通以太网帧发送。适用于普通以太网帧发送的第一融合描述符包含的内容见表1。

[0104]

63－48	47－40	39－32	31－24	23－16	15－0
VLAN	HEADLEN	MACLEN	CMD	RSV	Length

[0105] 表1

[0106] 其中，Length：待发送的数据长度。

[0107] MACLEN：表示MAC头+VLAN长度，以便虚拟网卡进行IP头checksum计算。

[0108] HEADLEN：表示IP头长度，以便虚拟网卡计算IP头checksum。

[0109] VLAN：提供802.1q/802.1ac标签信息。

[0110] CMD：Command Byte，进一步展开后包含的内容见表2。

[0111]

7	6	5	4	3	2	1	0
DEXT	VLE	RSV	RSV	IXSM	TXSM	IFCS	EOP

[0112] 表2

[0113] 其中，DEXT：Descriptor extension描述符扩展标识。1’b0表示非扩展(即普通描述符)；1’b1表示扩展。

[0114] VLE：VLAN Packet Enable，VLAN使能，表示本帧发送时需要添加VLAN标签。

[0115] TXSM：指示虚拟网卡添加TCP/UDP checksum。

[0116] IXSM：指示虚拟网卡添加IP checksum。

[0117] IFCS：Insert FCS，插入FCS，表示需要添加以太网帧的FCS字段。

[0118] EOP：End of Packet，对应一个帧的最后一个描述符。

[0119] RSV：表示保留未用。

[0120] (2)扩展发送描述符之一，适用于以太网帧和FCoE增强发送。适用于以太网帧和FCoE增强发送的第一融合描述符包含的内容见表3。

[0121]

63－48	47－40	39－32	31－24	23－16	15－0
VLAN	HEADLEN	MACLEN	CMD	ETYPE	Length

Ipsec SA IDX	L4LEN	RSV	ECMD	FCoEF	MSS

[0122] 表3

[0123] 其中,Length:待发送的数据长度。

[0124] MACLEN:对于非FCoE帧,表示MAC头+VLAN长度,以便虚拟网卡进行IP头checksum计算;对于FCoE帧,表示MAC头+VLAN+FCoE头长度,以便虚拟网卡进行FC-CRC计算。

[0125] HEADLEN:对于IP帧,表示IP头长度,以便虚拟网卡计算IP头checksum;对于FCoE帧,表示FCoE帧头长度,包括MAC头+VLAN+FCoE头+FC头长度。

[0126] VLAN:提供802.1q/802.1ac标签信息。

[0127] L4LEN:L4头长度。

[0128] Ipsec SA IDX:IPsec SA Index,指示安全联盟表的一个表项,以便虚拟网卡采用对应密钥进行加密。

[0129] MSS:Maximum Segment Size。TCP 和FCoE帧的最大分片字节数。

[0130] ETYPE:扩展描述符类型编码,8'h02为扩展描述符之一。

[0131] FCoEF:指示虚拟网卡如何给FCoE帧添加E-SOF和E-EOF。

[0132] CMD:Command Byte,进一步展开后包含的具体内容参见表4。

[0133]

7	6	5	4	3	2	1	0
DEXT	VLE	FCoE	RSV	IXSM	TXSM	IFCS	EOP

[0134] 表4

[0135] 其中,DEXT:Descriptor extension描述符扩展标识。1'b0表示非扩展;1'b1表示扩展(本描述符设置为1'b1)。

[0136] VLE:VLAN Packet Enable,VLAN使能,表示本帧发送时需要添加VLAN标签。

[0137] FCoE:指示虚拟网卡是按FCoE帧处理,还是非FCoE帧。

[0138] TXSM:指示虚拟网卡添加TCP/UDP checksum。

[0139] IXSM:指示虚拟网卡添加IP checksum。

[0140] IFCS:Insert FCS,插入FCS,表示需要添加以太网帧的FCS字段。

[0141] EOP:End of Packet,对应一个帧的最后一个描述符。

[0142] 表3中ECMD :Extension Command Byte,进一步展开后包含的具体内容参见表5。

[0143]

7	6-5	4	3	2	1	0
TSE	L4T	IPV4	Encrypt	IPSEC_TYPE	RSV	RSV

[0144] 表5

[0145] 其中,TSE:指示虚拟网卡必要时启动TCP 和FCoE帧分片。

[0146] L4T:L4负载类型(00:UDP;01:TCP;10:SCTP;11:RSV)。

[0147] IPV4:IP包类型(1:IPv4;0:IPv6)。

[0148] Encrypt:指示虚拟网卡是否启动IPSec加密。

[0149] IPSEC_TYPE:是ESP还是HA。

[0150] (3)扩展发送描述符之二,适用于RDMA操作,适用于RDMA操作的第一描述包含的内容参见表6。

[0151]

63－40	39－32	31－24	23－16	15－0
RSV	SEQ	CMD	ETYPE	Length
Source Node ID				
Source Memory Address				
Remote Node ID				
Remote Memory Address				

[0152] 表6

[0153] 其中，Length：待发送的数据长度，DMA的数据长度。

[0154] SEQ：系列号，记录本连接操作序号。

[0155] Source Node ID：本地Node ID，IP+TCP端口号。

[0156] Source Memory Address：本地服务器64比特内存物理地址，DMA的起始地址。

[0157] Remote Node ID：远端Node ID，IP+TCP端口号。

[0158] Remote Memory Address：远端服务器64比特内存物理地址，DMA的起始地址。

[0159] ETYPE：扩展描述符类型编码，8'h03为扩展描述符之二。

[0160] CMD：Command Byte，进一步展开后的内容参见表7。

[0161]

7	6	5	4	3	2	1	0
DEXT	VLE	TCP/ETH	WE/RD	TOE	RSV	IFCS	RSV

[0162] 表7

[0163] DEXT：Descriptor extension描述符扩展标识。1'b0表示非扩展；1'b1表示扩展（本描述符设置为1'b1）。

[0164] VLE：VLAN Packet Enable，VLAN使能，表示本帧发送时需要添加VLAN标签。

[0165] TCP/ETH：指示虚拟网卡是按RDMA over TCP还是RDMA over Ethernet。

[0166] WE/RD：指示是RDMA读操作，还是RDMA写操作。RDMA读操作

[0167] TOE：指示虚拟网卡执行TCP协议栈。

[0168] IFCS：Insert FCS，插入FCS，表示需要添加以太网帧的FCS字段。

[0169] 第二融合描述符表示虚拟化网卡单元接收一个数据帧时发现的一些信息。一般至少包含数据帧长度和第二融合描述符的类型，还可以包括如下具体信息：

[0170] 数据帧的帧格式，如以太网、FC、iSCSI、RapidIO、Infiniband、RDMA等；

[0171] 是否出现某些错误，如某校验和出错、数据帧长度异常等；

[0172] 虚拟网卡是否已经剥离数据帧的某些字段，如以太网帧的FCS等；

[0173] 虚拟网卡是否完成解密，如IPSec等；

[0174] 虚拟网卡从数据帧提取了某些字段，如以太网帧的VLAN标签、IP报文五元组信息等。

[0175] 虚拟网卡单元对数据帧的字段的判断结果、虚拟网卡单元对数据帧进行的修改等其它可选信息。

[0176] 为了更清楚、简洁表示各种网络发送和接收特征，第二融合描述符的类型不止一

种。

[0177] 第二融合描述符可以统一定义多种可以区分的格式，如针对以太网和FCoE定义2种格式、针对RapidIO定义1种格式、针对Infiniband定义1种格式、针对RDMA定义1种格式等。举例如下，其在64比特普通描述符基础上进一步定义扩展类型，普通接收描述符适用于普通以太网帧接收；各种扩展接收描述符适用于其它帧格式或指示虚拟网卡分担更多功能。

[0178] (1)普通接收描述符，适用于普通以太网帧接收，适用于普通以太网帧接收的第二融合描述符包含的内容参见表8。

[0179]

63－48	47－40	39－32	31－24	23－16	15－0
VLAN Tag	Errors	RSV	Status	RSV	Length

[0180] 表8

[0181] 其中，Length：接收的数据长度。

[0182] VLAN：提取的802.1q/802.1ac标签信息。

[0183] Status：状态信息字节，进一步展开后的内容参见表9。

[0184]

7	6	5	4	3	2	1	0
PIF	IPCS	L4CS	UDPCS	VP	EOP	SOP	DEXT

[0185] 表9

[0186] 其中，DEXT：Descriptor extension描述符扩展标识。1'b0表示非扩展(本描述符设置为1'b0)；1'b1表示扩展。

[0187] VP：VLAN Packet，指示输入帧是否携带VLAN标签。

[0188] IPCS：Ipv4 Checksum，指示完成了IP头校验，结果在IPE。

[0189] L4CS：L4 Checksum，指示完成了L4校验，结果在L4E。

[0190] UDPCS：UDP Checksum，指示完成了L4校验是UDP还是TCP。

[0191] PIF：Non Unicast Address，指示输入帧的MAC是否为单播。

[0192] EOP：End of Packet，对应一个帧的最后一个描述符。

[0193] SOP：Start of Packet，对应一个帧的第一个描述符。

[0194] 表8中Errors：错误信息字节，进一步展开后包含的内容参见表10。

[0195]

7	6	5	4	3	2	1	0
IPE	L4E	RSV	RSV	RSV	RSV	RSV	RXE

[0196] 表10

[0197] 其中，IPE：Ipv4 Checksum Error，IP头校验结果。

[0198] L4E：L4校验结果，如TCP/UDP Checksum Error。

[0199] RXE：其它以太网帧错误，如CRC错误、链路错误、各种长度错误。

[0200] 表8中RSV：表示保留未用。

[0201] (2)扩展接收描述符之一，适用于以太网帧和FCoE增强接收，适用于太帧和FCoE增强接收的第二融合描述符包含的内容参见表11。

[0202]

63 - 48	47 - 40	39 - 32	31 - 24	23 - 16	15 - 0
VLAN Tag	Errors	Ext. Status	Status	ETYPE	Length
Packet Type	HDR_LEN	Ext. Errors	FCoE_PARAM/Fragment Checksum/ RSS Hash/RSS TYPE		

[0203] 表11

[0204] 其中,Length:接收的数据长度。

[0205] ETYPE:扩展描述符类型编码,8'h01为扩展描述符之一。

[0206] VLAN:提取的802.1q/802.1ac标签信息。

[0207] Status:状态信息字节,进一步展开包含的内容参见表12。

[0208]

7	6	5	4	3	2	1	0
PIF	IPCS/ FCEOFs	L4CS/ FCSTAT[1]	UDPCS/ FCSTAT[0]	VP	EOP	SOP	DEXT

[0209] 表12

[0210] 其中,DEXT:Descriptor extension描述符扩展标识。1'b0表示非扩展;1'b1表示扩展(本描述符设置为1'b1)。

[0211] VP:VLAN Packet,指示输入帧是否携带VLAN标签。

[0212] IPCS:Ipv4 Checksum,指示完成了IP头校验,结果在IPE。

[0213] L4CS:L4 Checksum,指示完成了L4校验,结果在L4E。

[0214] UDPCS:UDP Checksum,指示完成了L4校验是UDP还是TCP。

[0215] PIF:Non Unicast Address,指示输入帧的MAC是否为单播。

[0216] EOP:End of Packet,对应一个帧的最后一个描述符。

[0217] SOP:Start of Packet,对应一个帧的第一个描述符。

[0218] FCSTAT:FCoE Status,FCoE帧的FC状态。

[0219] FCEOFs:与错误信息的FCEOFe共同表示EOF/SOF系列状态。

[0220] 表11中Ext. Status:状态信息扩展字节,进一步展开包含的内容参见表12。

[0221]

7	6	5	4	3	2	1	0
SECP	UDPV	VEXT	RSV	RSV	RSV	RSV	RSV

[0222] 表13

[0223] 其中,SECP:IPSec命中SA,并进行了处理。

[0224] UDPV:UDP Checksum Valid,表示接收帧为UDP,且含有非零Checksum,Fragment Checksum字段有效。

[0225] VEXT:双VLAN帧。

[0226] 表11中Errors:错误信息字节,进一步展开包含的内容见表14。

[0227]

7	6	5	4	3	2	1	0
IPE/ FCEOFe	L4E	RSV	RSV	FCERR			RXE

[0228] 表14

[0229] 其中IPE:Ipv4 Checksum Error,IP头校验结果。

[0230] FCEOFe:与状态信息的FCEOFs共同表示EOF/SOF系列状态。

[0231] L4E:L4校验结果,如TCP/UDP Checksum Error。

[0232] RXE:其它以太网帧错误,如CRC错误、链路错误、各种长度错误。

[0233] FCERR:FCoE错误码,3'b000表示无错误;3'b001表示错误的FC CRC;依此定义。

[0234] 表11中Ext. Errors:错误信息扩展字节,进一步展开包含的内容见表15。

[0235]

7	6	5	4	3	2	1	0
RSV		RSV	RSV	SECERR			RSV

[0236] 表15

[0237] 其中,SECERR:IPSec错误码,3'b000表示无错误;2'b001表示SA没命中;2'b010表示摘要错误;依此定义。

[0238] 表11中HDR_LEN:头部长度,不同的帧类型长度不同。

[0239] Packet Type:识别的帧类型,分为L2还是非L2,进一步展开包含的内容见表16。

[0240]

15	14-10	9	8	7	6	5	4	3	2	1	0
0=L3	RSV	IPV4	IPV4E	IPV6	IPV6E	TCP	UDP	SCTP	NFS	IPSec ESP	IPSec AH
1=L2	RSV	802.1x	RSV	FCoE	RSV	RSV	RSV	RSV	RSV	RSV	RSV

[0241] 表16

[0242] 其中,表11中FCoE_PARAM:针对FCoE帧,提取的FCoE一些参数。

[0243] RSS Hash/RSS TYPE:HASH某些字段,以便实现分配到多核CPU之一。

[0244] Fragment Checksum:针对UDP帧,状态信息中UDPV有效时,此字段有效。

[0245] (3)扩展接收描述符之二,适用于RDMA读操作,适用于RDMA操作的第二描述符的内容参见表17。

[0246]

63－48	47－40	39－32	31－24	23－16	15－0
VLAN	SEQ	Errors	Status	ETYPE	Length
Source Node ID					
Source Memory Address					
Remote Node ID					
Remote Memory Address					

[0247] 表17

[0248] 其中,Length:接收的数据长度。

[0249] ETYPE:扩展描述符类型编码,8'h03为扩展描述符之二。

[0250] VLAN:提取的802.1q/802.1ac标签信息。

[0251] SEQ:系列号,记录本连接操作序号。

[0252] Source Node ID:本地Node ID,IP+TCP端口号。

[0253] Source Memory Address:本地服务器64比特内存物理地址,DMA的起始地址。

[0254] Remote Node ID:远端Node ID,IP+TCP端口号。

[0255] Remote Memory Address:远端服务器64比特内存物理地址,DMA的起始地址。

[0256] Status:状态信息字节,进一步展开包含的内容见表18。

[0257]

7	6	5	4	3	2	1	0
RSV	IPCS	L4CS	RSV	VP	EOP	SOP	DEXT

[0258] 表18

[0259] 其中,DEXT:Descriptor extension描述符扩展标识。1'b0表示非扩展(本描述符设置为1'b0);1'b1表示扩展。

[0260] VP:VLAN Packet,指示输入帧是否携带VLAN标签。

[0261] IPCS:Ipv4 Checksum,指示完成了IP头校验,结果在IPE。

[0262] L4CS:L4 Checksum,指示完成了L4校验,结果在L4E。

[0263] EOP:End of Packet,对应一个帧的最后一个描述符。

[0264] SOP:Start of Packet,对应一个帧的第一个描述符。

[0265] 表17中Errors:错误信息字节,进一步展开包含的内容参见表19。

[0266]

7	6	5	4	3	2	1	0
IPE	L4E	RSV	RSV	RSV	RSV	RSV	RXE

[0267] 表19

[0268] 其中,IPE:Ipv4 Checksum Error,IP头校验结果。

[0269] L4E:L4校验结果,如TCP Checksum Error。

[0270] RXE:其它以太网帧错误,如CRC错误、链路错误、各种长度错误。

[0271] 上文列举了不同格式下第一融合描述符合第二融合描述符包含的内容,在具体实现时,可以减少融合交换格式描述符中的内容,也可以在预留字段中再增加内容,但是,与

现有各种网络适配器实现区别明显的地方是，第一融合描述符和第二融合描述符中均不包含buffer空间起始地址，即与服务器CPU的IO地址没有关联。

[0272] 进一步，各物理服务器上借助虚拟化VMM实现多个VM的虚拟化运行环境，为每个VM提供虚拟化的CPU、内存、存储、网卡等。

[0273] 这样在各VM上运行与一个或多个虚拟网卡驱动程序，且每个虚拟网卡驱动程序与虚拟网卡单元一一对应。不同VM上运行的虚拟网卡驱动程序对应的虚拟网卡单元的标识均不相同。

[0274] VM上运行的某个虚拟网卡驱动程序需要发送数据帧时，服务器接口单元获取该VM上运行的该虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符，将该第一融合描述符和获取的数据帧发送给队列池及调度单元。

[0275] 服务器接口单元需向VM发送数据帧和第二融合描述符时，将该数据帧发送给与本服务器接口单元对应的服务器上的对应VM，并将第二融合描述符的内容发送给，发送该数据帧的虚拟网卡单元对应的虚拟网卡驱动程序运行的VM，使该VM上运行的对应虚拟网卡驱动程序根据接收到的第二融合描述符的内容对接收到的数据帧进行处理。

[0276] 同获取服务器的上需要发送的数据帧处理过程一致，只是在虚拟多个VM时，在对应的VM上获取数据帧和描述符。在向服务器发送数据帧和描述符时，将数据帧和描述符发送给服务器中对应的VM中即可。

[0277] 当一个VM从一个源服务器迁移到另外一个目的服务器上时，需在目的服务器上运行一个或多个虚拟网卡驱动程序，并建立与目的网卡共享设备中一个或多个虚拟网卡单元的一一对应关系。因为虚拟网卡驱动程序采用队列方式与虚拟网卡单元进行消息传递，第一融合描述符合第二融合描述符与CPU的IO地址没有关联，极大地降低了VM与网卡的关联，因此，容易实现VM的迁移。

[0278] 实现VM迁移的过程具体如下：

[0279] (1)、服务器上停止即将迁移的VM上的虚拟网卡驱动程序的发送功能。

[0280] (2)、管理单元获知任一VM迁移时，即VM上的虚拟网卡驱动程序的发送功能停止时，停止该VM上运行的虚拟网卡驱动程序对应的虚拟网卡单元接收外部网络发送数据帧的功能；使队列池及调度单元完成数据帧的收发。

[0281] (3)、复制源VM的软件现场到目的VM的相同操作系统上。管理单元将所述停止接收功能的虚拟网卡单元上的相关内容复制到目的虚拟网卡单元和目的融合交换单元；所述目的虚拟网卡单元为迁移后的VM上运行的虚拟网卡驱动程序对应的虚拟网卡单元。

[0282] (4)、当所述迁移后的VM上的虚拟网卡驱动程序启动时，管理单元启动目的虚拟网卡单元的收发功能。

[0283] 为了增强该装置的功能，该装置还可增加一个或多个共享加速单元。

[0284] 共享加速单元接收到运行该共享加速单元对应的共享设备驱动程序的服务器发送的数据帧时，对该数据帧根据配置进行加速处理，并将处理结果返回发送该数据帧的服务器。

[0285] 其中，共享加速单元对数据帧的加速处理包括：浮点计算、加解密、压缩解压缩、图形图像处理等。

[0286] 该共享加速单元若具有网络通信功能，则将处理结果发送给网络接口单元或返回

发送该数据帧的服务器。

[0287] 参见图3，图3为本发明实施例中服务器虚拟VM时共享资源系统示意图。图3中的VM1上运行虚拟网卡单元1和虚拟网卡单元2对应的虚拟网卡驱动程序；VM5上运行共享加速单元1的共享设备驱动程序。

[0288] 当虚拟网卡单元1对应的虚拟网卡驱动程序需要发送数据帧时，服务器接口单元从VM1上获取需要发送的数据帧，以及发送该数据帧的相关描述信息，根据该相关描述信息构造第一融合描述符，并将构造的第一融合描述符以及数据帧通过队列池及调度单元调度到虚拟网卡单元1。

[0289] 虚拟网卡单元1根据第一融合描述符处理该数据帧，并发送给网络接口单元。

[0290] 网络接口单元，将虚拟网卡单元1发送的数据帧转发到外部网络。在向外部网络转发时，可以根据管理单元的配置确定是否给该数据帧添加标签等操作。如果接收到外部网络发送的数据帧时，根据该数据帧的信息字段与虚拟网卡单元标识的对应关系，将该数据帧发送给对应的虚拟网卡单元，如虚拟网卡单元2。

[0291] 虚拟网卡单元2接收到网络接口单元转发的数据帧时，对该数据帧处理，并构造第二融合描述符发送给队列池及调度单元。

[0292] 队列池及调度单元将数据帧和第二融合描述符发送给服务器接口单元1，因为虚拟网卡单元2的虚拟网卡驱动程序运行在服务器1上的VM1上，且服务器1对应的服务器接口单元为服务器接口单元1，因此，队列池及调度单元将数据帧和第二融合描述符调度到服务器接口单元1上。

[0293] 服务器接口单元1将第二融合描述符的内容以及数据帧写入对应服务器1中的对应VM1，使该VM1使用第二融合描述符的内容处理对应的数据帧。

[0294] 当VM1要从服务器1迁移到服务器n时，具体过程如下：

[0295] 第一步、VM1停止其上的虚拟网卡驱动程序的发送功能，网卡共享设备停止虚拟网卡单元1和虚拟网卡单元2的接收外部网络发送的数据帧。

[0296] 第二步、队列池及调度单元完成数据帧的收发。

[0297] 第三步、服务器1上的VM1的软件现场到目的VM的相同操作系统上；网卡共享设备复制虚拟网卡单元1和虚拟网卡单元2的现场到目的网卡共享设备上的虚拟网卡单元上和融合交换单元上。

[0298] 第四步、启动目的VM上的虚拟网卡驱动程序，以及目的虚拟网卡单元的收发功能。至此，VM迁移结束。

[0299] 在具体实现时，有些VM的从1个服务器迁移到另外一个服务器上，且迁移前的服务器和迁移后的服务器连接的不是同一网卡共享设备，因此，需要将源网卡共享设备上的现场相关内容，全部复制到目的网卡共享设备上。

[0300] 由于VM5上运行共享加速单元1的共享设备驱动程序，当共享加速单元1接收到VM5发送的数据帧时，对该数据帧进行加解密、浮点技术、压缩/解压缩、图形图像处理等处理后，发送回服务器n的VM5上，如果共享加速单元1具有通信功能，将处理后的数据帧发送给网络接口单元。

[0301] 上述实施例的单元可以集成于一体，也可以分离部署；可以合并为一个单元，也可以进一步拆分成多个子单元。

[0302] 以上实施例对本申请具体实施例中的数据转发装置进行了说明，本实施例给出本申请实施例一中装置的硬件架构组成。

[0303] 该装置是可以软硬件结合的可编程设备，具体参见图4，图4为本发明实施例一中装置的硬件架构组成示意图，该装置包括：FPGA/ASIC和CPU(中央处理器)小系统；其中，

[0304] FPGA/ASIC，用于完成装置中的n个服务器接口单元、1个队列池及调度单元、m虚拟网卡单元、1个网络接口单元和1个或多个共享加速单元等单元完成的功能，这里不再详述，在该实施例中以2个加速单元为例。

[0305] CPU小系统，包含CPU、以及正常工作必备的存储器和其它硬件，用于完成装置中的管理单元功能，与FPGA/ASIC互连。

[0306] 其中，任一服务器接口单元用于获取服务器上运行的VM发送的数据帧和发送该数据帧的描述符，构造第一融合描述符，并发送给，队列池及调度单元；获取队列池调度单元中的数据帧和第二融合描述符，并将获取的数据帧以及第二融合描述符的内容写入对应的服务器；

[0307] 队列池及调度单元，用于通过队列将服务器接口单元发送的数据帧和第一融合描述符调度给虚拟网卡单元；接收到虚拟网卡单元发送的第二融合描述符时，通过队列将该数据帧和第二融合描述符调度到对应的服务器接口单元；

[0308] 任一虚拟网卡单元，接收到队列池及调度单元调度来的数据帧和第一融合描述符时，根据第一融合描述符处理该数据帧，并将处理后的数据帧发送给网络接口单元；接收到网络接口单元发送的数据帧时，为该数据帧构造第二融合描述符，并将该数据帧以及构造的第二融合描述符发送给服务器接口单元；

[0309] 网络接口单元，接收虚拟网卡单元发送的数据帧时，转发到外部网络，接收到外部网络发送的数据帧时，根据管理单元配置的对应关系，将该数据帧发送给对应的虚拟网卡单元；

[0310] 管理单元，配置数据帧的信息字段与虚拟网卡单元标识的对应关系，对所述队列池及调度单元中的各队列配置的传输速率、优先级、调度策略和各队列的当前状态，并将配置的信息存储到本单元中，和/或，将数据帧的信息字段与虚拟网卡单元标识的对应关系存储到网络接口单元；将对所述队列池及调度单元中的各队列配置的传输速率、优先级、调度策略和各队列的当前状态存储到队列池及调度单元对应单元中。管理单元也可以通过网络接口单元与外部网络进行数据帧收发。

[0311] 需要说明的是，图4所示的装置只是一个具体的例子，也可以通过其他的与本实施例描述不同结构实现，如可将FPGA/ASIC中的部分功能采用CPU上运行的程序来实现，或将网络接口单元采用与CPU直接互连的普通以太网网卡来实现，因此，本申请对装置的具体结构不作具体限定。

[0312] 本实施中通过在网卡共享设备中虚拟多个虚拟网卡单元，来一一对应完成各服务器上的数据帧的收发，实现了多个服务器资源共享。并且由于本发明具体实现数据帧转发时的第一融合描述符合第二融合描述符中不包含buffer空间起始地址，即与CPU的IO地址没有关联，降低了VM与虚拟网卡单元的关联，因此更容易实现VM的迁移。

[0313] 本发明具体实施例中基于与上述技术同样的发明构思，还提出一种数据转发方法。应用于包括多个服务器、一个网卡共享设备的网络系统中的网卡共享设备上，所述网卡

共享设备虚拟多个虚拟网卡；每个服务器上运行一个或多个虚拟网卡驱动程序，且每个虚拟网卡驱动程序与共享网卡设备上的虚拟网卡一一对应。

[0314] 参见图5，图5为本发明实施例中接收服务器发送的数据帧的处理方式流程示意图。具体步骤为：

[0315] 步骤501，网卡共享设备获取服务器上运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符，将该第一融合描述符和获取的数据帧调度到发送该数据帧的虚拟网卡驱动程序对应的虚拟网卡上。

[0316] 其中，第一融合描述符包含描述符类型和数据帧长度。

[0317] 第一融合描述符的内容还包含下述之一或任意组合：

[0318] 数据帧的帧格式、是否添加或修改网络节点ID信息、是否重新计算校验和、是否进行加密、指导虚拟网卡单元如何处理数据帧的其它信息；

[0319] 该网卡共享设备与所述服务器通过PCI Express点到点连接时，作为PCIExpress链路的下游端点，所述方法进一步包括：配置多个发送引擎和接收引擎，并分别对应一个队列。

[0320] 步骤501中获取服务器上运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符，包括：根据所述发送引擎指向的当前有效的发送buffer描述符从所述服务器内存中读取需要发送的数据帧，将该发送buffer描述符中除buffer空间起始地址之外的内容构造为第一融合描述符，并将该第一融合描述符和该数据帧通过对应的队列调度到对应的虚拟网卡上。

[0321] 该网卡共享设备与所述服务器通过以太网点到点连接时，所述方法进一步包括：配置多个发送引擎和接收引擎，并分别对应一个队列；

[0322] 步骤501中获取服务器上运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符，包括：发送引擎接收到服务器发送的描述符和数据帧时，将该描述符的格式变换为第一融合描述符的格式作为第一融合描述符，并将该第一融合描述符和该数据帧发送给对应的队列；

[0323] 步骤502，该网卡共享设备上接收到该数据帧和第一融合描述符的虚拟网卡，根据该第一融合描述符处理该数据帧，并将处理后的数据帧发送到外部网络。

[0324] 参见图6，图6为本发明实施例中接收到外部网络发送的数据帧的处理方式流程示意图。具体步骤为：

[0325] 配置数据帧的信息字段与虚拟网卡标识的对应关系；

[0326] 步骤601，网卡共享设备接收到外部网络发送的数据帧时，根据该数据帧的信息字段匹配到对应的虚拟网卡标识，并将该数据帧发送给匹配到的虚拟网卡标识对应的虚拟网卡。

[0327] 步骤602，该网卡共享设备接收到该数据帧的虚拟网卡对该数据帧进行处理，根据处理结果为该数据帧构造第二融合描述符，并将该数据帧以及构造的第二融合描述符的内容发送给对应的服务器，使所述服务器上运行的对应虚拟网卡驱动程序根据接收到的第二融合描述符的内容对接收到的数据帧进行处理。

[0328] 其中，该第二融合描述符至少包含描述符类型和数据帧长度。

[0329] 第二融合描述符的内容还包含下述之一或任意组合：

[0330] 数据帧的帧格式、是否出现错误、虚拟网卡单元对数据帧的字段的判断结果，虚拟网卡单元提取或丢弃数据帧中的信息、虚拟网卡单元对数据帧进行的修改、虚拟网卡单元是否完成解密、虚拟网卡单元发现或已经处理的其它信息。

[0331] 该网卡共享设备与所述服务器通过PCI Express点到点连接时，作为PCI Express链路的下游端点，所述方法进一步包括：配置多个发送引擎和接收引擎，并分别对应一个队列。

[0332] 步骤602中将该数据帧以及构造的第二融合描述符的内容写入对应的服务器，使所述服务器根据写入的第二融合描述符的内容对所述写入的数据帧进行处理，包括：在该接收引擎对应的队列中有数据帧和第二融合描述符时，读取第二融合描述符，根据该第二融合描述符读取其后的数据帧，并将该数据帧写入所述接收buffer描述符所指的服务器的buffer中，并在回写所述接收buffer描述符时进一步携带该第二融合描述符的内容。

[0333] 该网卡共享设备还可根据预先配置的传输速率、优先级、调度策略对接收到的数据帧进行部分丢弃处理，或将接收到的数据帧调度给服务器或虚拟网卡。

[0334] 该网卡共享设备与所述服务器通过以太网点到点连接时，所述方法进一步包括：配置多个发送引擎和接收引擎，并分别对应一个队列。

[0335] 步骤602中，将该数据帧以及构造的第二融合描述符的内容写入对应的服务器，使所述服务器根据写入的第二融合描述符的内容对所述写入的数据帧进行处理，包括：在该接收引擎对应的队列中有数据帧和第二融合描述符时，读取第二融合描述符，根据该第二融合描述符读取其后的数据帧，将第二融合描述符以及该数据帧发送给对应的服务器，使该服务器上运行的对应虚拟网卡驱动程序进一步处理所述第二融合描述符和数据帧。

[0336] 本实施例中进一步包括：各服务器上通过VMM实现多个虚拟机VM的虚拟环境，各VM上运行一个或多个虚拟网卡驱动程序，且每个虚拟网卡驱动程序与虚拟网卡单元一一对应。

[0337] 步骤601中获取服务器上运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符，包括：

[0338] 获取服务器上的VM运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符，将该第一融合描述符和获取的数据帧发送给该数据帧的虚拟网卡驱动程序对应的虚拟网卡；

[0339] 步骤602中将该数据帧以及构造的第二融合描述符的内容发送给对应的服务器，使所述服务器根据接收到的第二融合描述符的内容对接收到的数据帧进行处理，包括：

[0340] 将该数据帧，以及第二融合描述符的内容发送给，发送该数据帧的虚拟网卡对应的虚拟网卡驱动程序运行的VM，使所述VM上运行的对应虚拟网卡驱动程序根据接收到的第二融合描述符的内容对接收到的数据帧进行处理。

[0341] 该网卡共享设备获知任一VM迁移时，停止该VM上运行的虚拟网卡驱动程序对应的虚拟网卡接收外部网络发送数据帧的功能；完成对已接收到的数据帧的转发，并将该虚拟网卡上的相关内容复制到目的虚拟网卡，该目的虚拟网卡为迁移后的VM上运行的虚拟网卡驱动程序对应的虚拟网卡。

[0342] 当所述迁移后的VM上的虚拟网卡驱动程序启动时，启动所述目的虚拟网卡的收发功能。

[0343] 本实施例中的网卡共享设备还进一步配置共享加速功能时，所述方法进一步包括：

[0344] 该网卡共享设备接收到运行共享设备驱动程序的服务器发送的数据帧时，对该数据帧根据配置进行加速处理，并将处理结果返回发送该数据帧的服务器；若具有网络通信功能时，将处理结果发送给外部网络或返回发送该数据帧的服务器。

[0345] 综上所述，本发明具体实施例中通过在网卡共享设备中虚拟多个虚拟网卡单元，在各服务器上运行一个或多个虚拟网卡单元对应的虚拟网卡驱动程序，来一一对应完成各服务器上的数据帧的收发，实现了多个服务器资源共享。

[0346] 并且由于本发明具体实现数据帧转发时的第一融合描述符合第二融合描述符中不包含buffer空间起始地址，即与CPU的IO地址没有关联，降低了VM与虚拟网卡单元的关联，因此更容易实现虚拟了VM的多个服务器之间共享网卡单元，以及服务器上VM的迁移。

[0347] 在网卡共享设备中还增加了共享加速单元，对应在服务器的操作系统或VM的操作系统上运行共享加速单元的共享驱动程序，来实现对服务器或服务器上的VM实现加速处理的功能，以提高服务器的处理速度。

[0348] 以上所述，仅为本发明的较佳实施例而已，并非用于限定本发明的保护范围。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

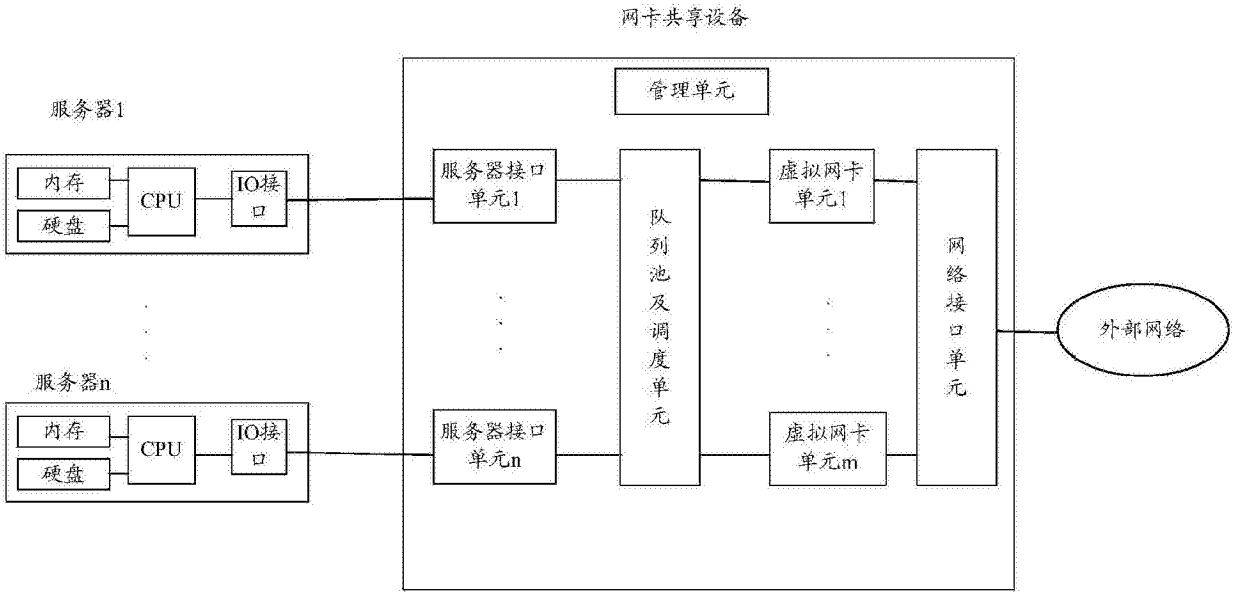

网卡共享设备
管理单元
服务器1
内存
硬盘
CPU
IO接口
服务器接口单元1
队列池及调度单元
虚拟网卡单元1
网络接口单元
外部网络
服务器n
内存
硬盘
CPU
IO接口
服务器接口单元n
虚拟网卡单元m

图1

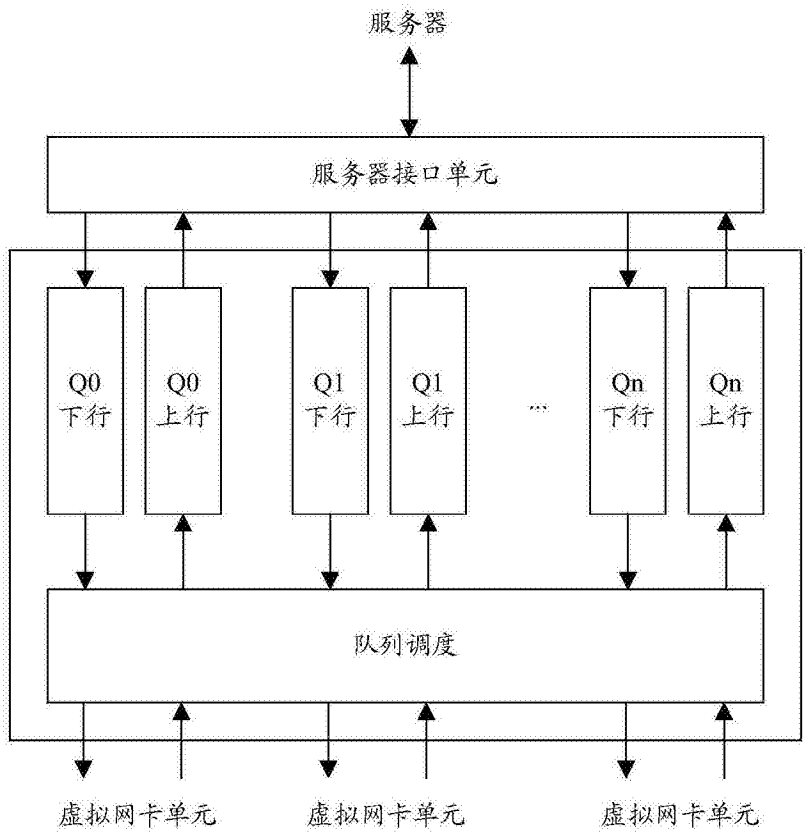

服务器
服务器接口单元
Q0 下行
Q0 上行
Q1 下行
Q1 上行
...
Qn 下行
Qn 上行
队列调度
虚拟网卡单元
虚拟网卡单元
虚拟网卡单元

图2

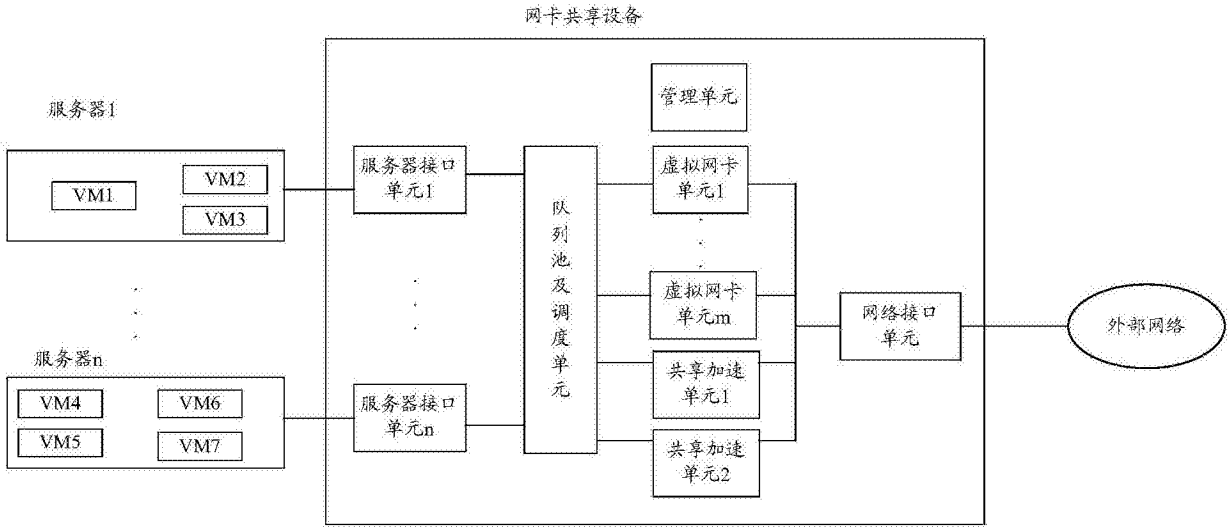

网卡共享设备
服务器1
VM1
VM2
VM3
服务器n
VM4
VM5
VM6
VM7
服务器接口单元1
服务器接口单元n
队列池及调度单元
管理单元
虚拟网卡单元1
虚拟网卡单元m
共享加速单元1
共享加速单元2
网络接口单元
外部网络

图3

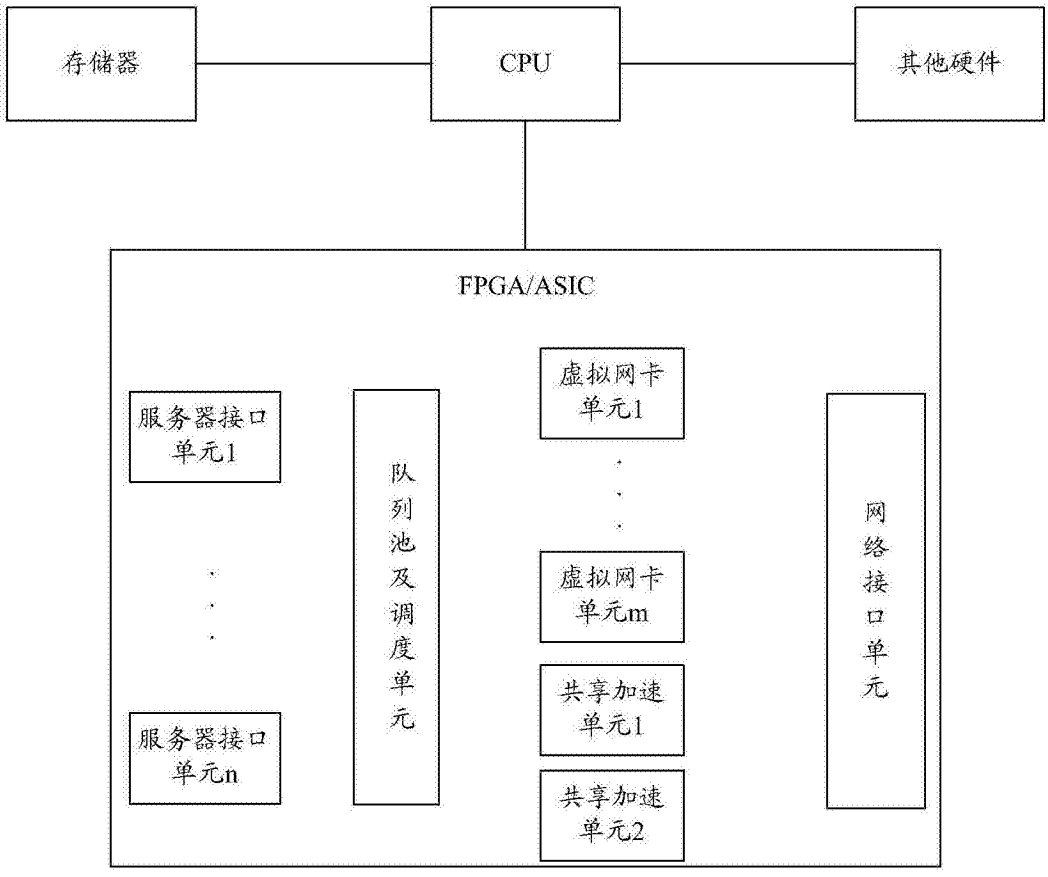

存储器
CPU
其他硬件
FPGA/ASIC
服务器接口单元1
服务器接口单元n
队列池及调度单元
虚拟网卡单元1
虚拟网卡单元m
共享加速单元1
共享加速单元2
网络接口单元

图4

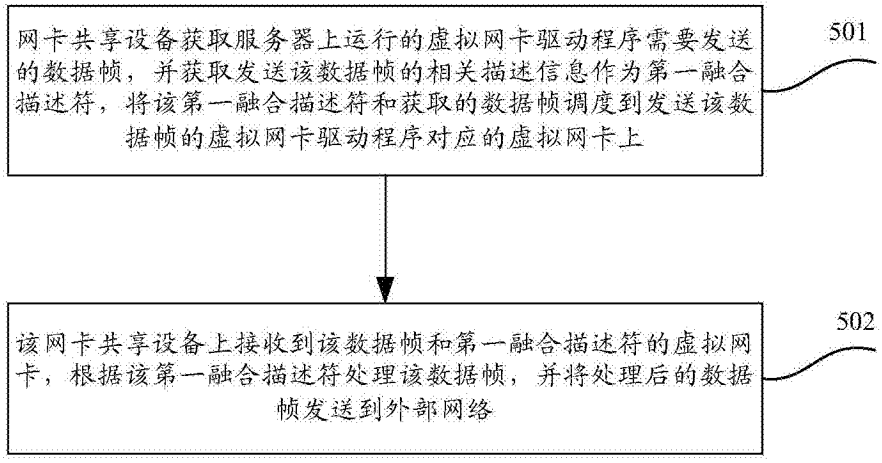

网卡共享设备获取服务器上运行的虚拟网卡驱动程序需要发送的数据帧，并获取发送该数据帧的相关描述信息作为第一融合描述符，将该第一融合描述符和获取的数据帧调度到发送该数据帧的虚拟网卡驱动程序对应的虚拟网卡上
501
该网卡共享设备上接收到该数据帧和第一融合描述符的虚拟网卡，根据该第一融合描述符处理该数据帧，并将处理后的数据帧发送到外部网络
502

图5

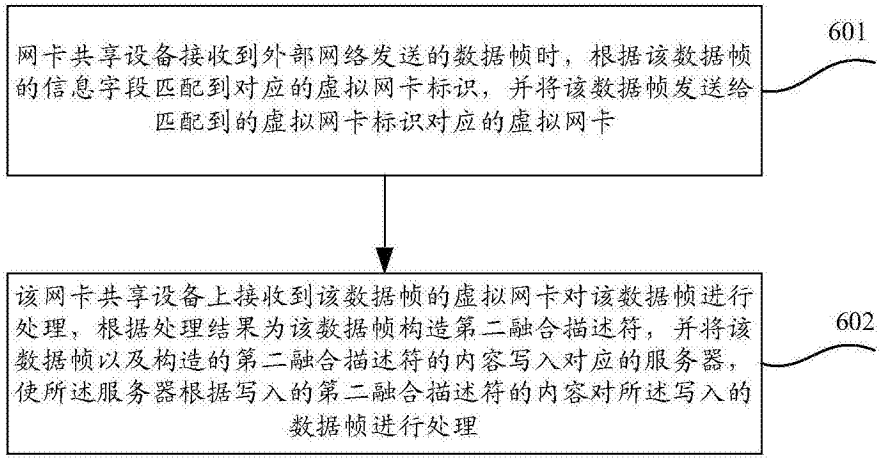

网卡共享设备接收到外部网络发送的数据帧时，根据该数据帧的信息字段匹配到对应的虚拟网卡标识，并将该数据帧发送给匹配到的虚拟网卡标识对应的虚拟网卡
601
该网卡共享设备上接收到该数据帧的虚拟网卡对该数据帧进行处理，根据处理结果为该数据帧构造第二融合描述符，并将该数据帧以及构造的第二融合描述符的内容写入对应的服务器，使所述服务器根据写入的第二融合描述符的内容对所述写入的数据帧进行处理
602

图6