(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114465899 A
(43)申请公布日 2022.05.10

(21)申请号 202210120548.4

(22)申请日 2022.02.09

(71)申请人 浪潮云信息技术股份公司
地址 250100 山东省济南市高新区浪潮路1036号浪潮科技园S01号楼

(72)发明人 李彦君 孙思清 高传集 胡章丰

(74)专利代理机构 济南信达专利事务所有限公司 37100
专利代理师 陈婷婷

(51)Int.Cl.
H04L 41/083(2022.01)
H04L 41/40(2022.01)
H04L 41/0895(2022.01)
H04L 45/645(2022.01)
H04L 49/101(2022.01)
H04L 49/111(2022.01)
H04L 49/35(2022.01)
H04L 67/1001(2022.01)

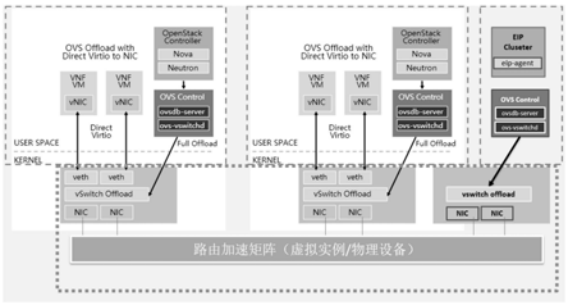
权利要求书1页 说明书6页 附图1页

(54)发明名称

复杂云计算环境下的网络加速方法、系统及装置

(57)摘要

本发明公开了复杂云计算环境下的网络加速方法、系统及装置，属于云计算及计算机网络技术领域，该方法将计算与网络处理解耦，卸载网络任务给相应的设备处理：将虚拟网络的转发逻辑通过流表化处理的方式融合在虚拟交换设备中一体化完成，虚拟交换设备的流表化处理卸载至专用的硬件加速处理单元阵列；将专用的硬件加速处理单元通过PCI-E接口进行模块集成化，通过设计包括接入子模块和交换芯片背板的专用硬件网络加速盒子，以内部数字电路与高速交换芯片相连，构成完整的交换矩阵。本发明能够解决虚拟网络性能低、受计算影响大、网络部署复杂的问题，大大提升虚拟网性能和东西向扩展能力，并降低数据中心网络的部署复杂度。



CN 114465899 A

1.复杂云计算环境下的网络加速方法,其特征在于,将计算与网络处理解耦,卸载网络任务给相应的设备处理:

将虚拟网络的转发逻辑通过流表化处理的方式融合在虚拟交换设备中一体化完成,

虚拟交换设备的流表化处理卸载至专用的硬件加速处理单元阵列;

将专用的硬件加速处理单元通过PCI-E接口进行模块集成化,通过设计包括接入子模块和交换芯片背板的专用硬件网络加速盒子,以内部数字电路与高速交换芯片相连,构成完整的交换矩阵。

2.根据权利要求1所述的复杂云计算环境下的网络加速方法,其特征在于,所述硬件网络加速盒子外置多个高速上联口,用于连接核心交换机,以组成更大规模的交换矩阵。

3.根据权利要求2所述的复杂云计算环境下的网络加速方法,其特征在于,所述高速上联口为4-8个。

4.根据权利要求2所述的复杂云计算环境下的网络加速方法,其特征在于,通过所述硬件网络加速盒,同一模块内部的虚拟机之间可以通过硬件卸载的虚拟交换机处理,跨节点之间的流量,则通过加速盒内置的高速交换矩阵进行交换。

5.根据权利要求1或2或4所述的复杂云计算环境下的网络加速方法,其特征在于,每一条物理PCI-E对应一个子模块,每个子模块对应一台物理机;

每一条物理PCI-E又通过SR-IOV标准,虚拟成多个逻辑PCI-E设备,对应多块虚拟网卡,供同一个物理机上的多个虚机使用。

6.根据权利要求1或2所述的复杂云计算环境下的网络加速方法,其特征在于,虚拟网的路由、SNAT和Floating IP三层功能全部使用OpenFlow流表的方式实现;

报文的处理在OpenVSwitch交换机中一次性处理完成。

7.根据权利要求1所述的复杂云计算环境下的网络加速方法,其特征在于,所述交换矩阵,通过线缆及交换设备与其它网络加速盒相连,构成一个普通的Fabric网络。

8.根据权利要求1或7所述的复杂云计算环境下的网络加速方法,其特征在于,交换矩阵本身支持Trunk功能,它和外部的交换机相连,可以做Trunk,从而支持Frame级的负载均衡。

9.复杂云计算环境下的网络加速系统,其特征在于,将传统网卡从宿主机上迁移到模块化形态的加速单元设备上,这些单元设备在硬件化后可通过PCI-E连接至宿主机;

该系统实现权利要求1-8任一项所述的复杂云计算环境下的网络加速方法。

10.复杂云计算环境下的网络加速装置,其特征在于,包括:至少一个存储器和至少一个处理器、以及权利要求1至8任一所述的硬件网络加速盒子;

所述至少一个存储器,用于存储机器可读程序;

所述至少一个处理器,用于调用所述机器可读程序,执行权利要求1至8任一所述的复杂云计算环境下的网络加速方法。

复杂云计算环境下的网络加速方法、系统及装置

技术领域

[0001] 本发明涉及云计算及计算机网络技术领域，具体地说是复杂云计算环境下的网络加速方法、系统及装置。

背景技术

[0002] 人工智能、机器学习、网络安全、超大规模架构和云服务等趋势的兴起，对网络提出了前所未有的要求，特别是在性能和高可用方面。这些因素加上无线网络和远程办公带来的网络使用激增，正推动着网络带宽、用户数量和活跃网络流量数量的增加，而网络流量的增长和流量的复杂性给服务器基础设施计算节点的CPU带来了巨大的压力。

[0003] 云计算环境下，网络通常具有较高的复杂性，虚拟网络和物理网络混合在一起，随着边缘计算的兴起，很多云服务都要求更低的延迟来支持部署在端系统的实时应用程序和服务，如视频会议、5G和自动驾驶汽车等，其他因素还包括需要支持传统网络服务，这些都对云计算网络提出了很高的性能要求。

发明内容

[0004] 本发明的技术任务是针对以上不足之处，提供复杂云计算环境下的网络加速方法、系统及装置，能够解决虚拟网络性能低、受计算影响大、网络部署复杂的问题，大大提升虚拟网性能和东西向扩展能力，并降低数据中心网络的部署复杂度。

[0005] 本发明解决其技术问题所采用的技术方案是：

[0006] 复杂云计算环境下的网络加速方法，将计算与网络处理解耦，卸载网络任务给相应的设备处理：

[0007] 将虚拟网络的转发逻辑通过流表化处理的方式融合在虚拟交换设备中一体化完成，

[0008] 虚拟交换设备的流表化处理卸载至专用的硬件加速处理单元阵列，使得网络具有更低时延、更高带宽、更少的CPU消耗；

[0009] 将专用的硬件加速处理单元通过PCI-E接口进行模块集成化，通过设计包括接入子模块和交换芯片背板的专用硬件网络加速盒子，以内部数字电路与高速交换芯片相连，构成完整的交换矩阵。

[0010] 在任何虚拟化的网络基础设施中，服务器内部都有大量的数据平面网络需求。网络工作负载在计算方面特别昂贵。单是虚拟交换一项就可以占用服务器90％以上的可用CPU资源。将计算与网络处理解耦，卸载网络任务给相应的设备处理，可以将这些重要的资源返回给应用层，同时提升虚拟网络的性能效率。

[0011] 进一步的，所述硬件网络加速盒子外置多个高速上联口，用于连接核心交换机，以组成更大规模的交换矩阵。

[0012] 优选的，所述高速上联口为4-8个。

[0013] 优选的，通过所述硬件网络加速盒，同一模块内部的虚拟机之间可以通过硬件卸

载的虚拟交换机处理，跨节点之间的流量，则通过加速盒内置的高速交换矩阵进行交换。

[0014] 优选的，每一条物理PCI-E对应一个子模块，每个子模块对应一台物理机；

[0015] 每一条物理PCI-E又通过SR-IOV标准，虚拟成多个逻辑PCI-E设备，对应多块虚拟网卡，供同一个物理机上的多个虚机使用。

[0016] 每个接入子模块包括支持虚拟通道化的PCI-E网卡、内部交换结构等，接入子模块的功能与普通智能网卡类似，只不过接入子模块无需光模块进行光电信号转化，接入子模块通过内部数字电路与高速交换芯片相连，可构成完整的交换矩阵。

[0017] 优选的，虚拟网的路由、SNAT和Floating IP三层功能全部使用OpenFlow流表的方式实现；

[0018] 报文的处理在OpenVSwitch交换机中一次性处理完成。

[0019] 虚拟网络的三层功能主要包括路由、SNAT和Floating IP，一般是通过Linux kernel的Namespace来实现的，每个路由器对应一个Namespace，利用Linux TCP/IP协议栈来做路由转发。

[0020] 传统的虚拟网络架构中，虚机端口通过tap口连接到Linux网桥上，进行防火墙网络安全相关功能的处理，处理完成后通过虚拟路由器连接到OpenVSwitch的集成网桥上，进行二层报文处理，如果需要三层路由，再从集成网桥的端口送入虚拟路由器的Linux namespace，查找namespace中配置的网段路由，如果是FIP流量需要再次送入FIP的Linux namespace再次查找，通过namespace中配置的iptables规则对报文进行NAT，再通过浮动IP网关将报文送入Openvswitch的集成网桥中，再通过patch端口将报文送入外网口中发出。总之，虚机流量所经过的冗长的链路跳数，网络性能较低，网络整体性能受限于其中的最短板，拓扑太长，定位问题开销较大，性能受限于iptables条数和namespace个数等。冗长链路上的任何一个节点都可能成为总体网络带宽的瓶颈，网络出现问题后定位的路径较长，网络面临低带宽高延时的问题迫切需要改善。

[0021] 针对虚拟网络现状中存在的上述问题，将报文的处理在OpenVSwitch交换机中一次性处理完成，将上述描述的一个三层报文需要多次进出多个bridge和多个linux namespace才能完成的功能一次性在流表中处理完成，大大缩短包延时，大大提高带宽。除了上述举例中典型的三层转发外，加速功能还包含负载均衡能力、EIP包处理能力等所有虚拟网络功能。

[0022] 为了提升网络性能，将虚拟网的三层功能全部使用OpenFlow流表的方式实现。以安全组为例，传统虚拟网每创建一个端口，就需要创建相应的网络连接设备，而通过流表化的方式，每创建一个端口，直接将其挂载在虚拟交换机(OVS)上即可，从而大大减少了跳数。

[0023] 优选的，所述交换矩阵可视为一台交换机，其通过线缆及交换设备与其它网络加速盒相连，构成一个普通的Fabric网络。

[0024] 进一步的，交换矩阵本身支持Trunk功能，它和外部的交换机相连，可以做Trunk，从而支持Frame级的负载均衡，使得在数据传输时保证没有拥塞出现。

[0025] 云计算环境下，网络通常具有较高的复杂性，虚拟网络和物理网络混合在一起，很多云服务都要求更低的延迟来支持部署在端系统的实时应用程序和服务。当前云计算中主流虚拟网络模型实现中，大量使用了Linux网络的命名空间技术。在使用分布式路由的场景下，每个计算终端均需要创建虚拟路由器和相应的网络命名空间，来实现流量的路由转发

功能。此类被大量使用的虚拟网络路由处理方式其实存在明显的性能问题，从一个虚拟主机上发出的流量，需要经过多个虚拟网络设备的处理，其网络性能损耗巨大。尤其在使用DPDK软加速技术的场景下，由于流量需要在用户态和内核态之间反复切换处理，导致加速性能远低于预期性能。

[0026] 本方法针对复杂云计算环境中虚拟网性能差，扩展性能力受限的问题，提出了基于流表卸载与直通处理的方法，通过软件与硬件的设计，可大大提升虚拟网性能和东西向扩展能力，并降低了数据中心网络的部署复杂度。

[0027] 本发明还要求保护复杂云计算环境下的网络加速系统，该系统将传统网卡从宿主机上迁移到模块化形态的加速单元设备上，这些单元设备在硬件化后可通过PCI-E连接至宿主机；

[0028] 该系统实现上述的复杂云计算环境下的网络加速方法。

[0029] 本发明还要求保护复杂云计算环境下的网络加速装置，包括：至少一个存储器和至少一个处理器、以及上述的硬件网络加速盒子；

[0030] 所述至少一个存储器，用于存储机器可读程序；

[0031] 所述至少一个处理器，用于调用所述机器可读程序，执行上述的复杂云计算环境下的网络加速方法。

[0032] 本发明的复杂云计算环境下的网络加速方法、系统及装置与现有技术相比，具有以下有益效果：

[0033] 将虚拟网络所有的转发逻辑通过流表化的方式融合在虚拟交换设备中一体化完成，大大减少了跳数，同时，将虚拟交换机中的处理进一步offload至专用的软硬件处理单元，从而进一步提升效率，使得网络具有更低延时，更高带宽，更少的CPU消耗策略路由实现全路径正反向对称引流，可支持各种非业务类型的虚拟设备，包括各种安全设备、监控设备、日志审计设备等等；

[0034] 可提供更好的网络服务能力，装置的集成化，大大节省了网络布线成本和运维成本，真正实现了低部署门槛的高效网络优化服务；

[0035] 将虚拟网的传统网络功能全部通过流表化集中处理，既可大大优化传统虚拟网络性能，也可提升相应的网络功能的处理性能；同时，涉及路由与转发性能需求的相关网络产品如负载均衡、集群网关等均可以通过本发明的方式获得并发能力和处理能力的大幅提升。

附图说明

[0036] 图1是本发明实施例提供的复杂云计算环境下的网络加速方法原理图；

[0037] 图2是本发明实施例提供的网络加速方法的实现过程图。

具体实施方式

[0038] 下面结合附图和具体实施例对本发明作进一步说明。

[0039] 在当前云计算中主流虚拟网络模型实现中，大量使用了Linux网络的命名空间：如Linux bridge，IPtables，tap/tun/veth接口等技术。而在使用分布式路由的场景下，每个计算终端均需要创建虚拟路由器和相应的网络命名空间，来实现流量的路由转发功能。此

类被大量使用的虚拟网络路由处理方式其实存在明显的性能问题，从一个虚拟主机上发出的流量，需要经过多个虚拟网络设备的处理，其网络性能损耗巨大。尤其在使用DPDK软加速技术的场景下，使用tap和veth接口后，由于流量需要在用户态和内核态之间反复切换处理，导致加速性能远低于预期性能。

[0040] 目前业界也有一些相应的解决方案，如比较热门的OVN技术，通过将控制命令流表化，从而减少虚拟转发跳数，提升系统性能。然而从实际使用的角度看，OVS还存在很多问题，如有些较低版本的云操作系统不支持OVN功能、数据库节点存在瓶颈问题、已经部署OVS环境迁移到OVN环境需要很多额外的工作，对于网络问题追溯更为复杂，环境的版本更新升级引入更多步骤等。且OVS并不是一种完全的卸载，虚拟网口的功能仍然留存在虚机里需要占用相应的计算资源，成为系统瓶颈。

[0041] 此外，由于高可用与网络隔离的需求，现有云数据中心的Tor交换机处网络连线非常复杂，故障排查与流量模型分析也较为困难。

[0042] 针对上述虚拟网络性能低、受计算影响大、网络部署复杂的问题，

[0043] 本发明实施例提供了一种复杂云计算环境下的网络加速方法，将计算与网络处理解耦，卸载网络任务给相应的设备处理：

[0044] 将虚拟网络所有的转发逻辑通过流表化处理的方式融合在虚拟交换设备中一体化完成，

[0045] 将虚拟网的2、3层功能全部通过流表化集中处理，这种处理方式不限于虚拟交换机，涉及网络路由与转发性能需求的相关产品均可通过本装置获得并发能力和处理能力的大幅提升；

[0046] 虚拟交换设备的流表化处理卸载至专用的硬件加速处理单元阵列，使得网络具有更低时延、更高带宽、更少的CPU消耗；

[0047] 将专用的硬件加速处理单元通过PCI-E接口进行模块集成化，通过设计包括接入子模块和交换芯片背板的专用硬件网络加速盒子，以内部数字电路与高速交换芯片相连，构成完整的交换矩阵。

[0048] 在任何虚拟化的网络基础设施中，服务器内部都有大量的数据平面网络需求。网络工作负载在计算方面特别昂贵。单是虚拟交换一项就可以占用服务器90％以上的可用CPU资源。将计算与网络处理解耦，卸载网络任务给相应的设备处理，可以将这些重要的资源返回给应用层，同时提升虚拟网络的性能效率。

[0049] 将传统网卡从宿主机上迁移到模块化形态的加速单元设备上，这些单元设备在硬件化后可通过PCI-E连接至宿主机，在该网络加速方法中，每一条物理PCI-E对应一个子模块，每个子模块对应一台物理机，而每一条物理PCI-E又通过SR-IOV标准，虚拟成多个逻辑PCI-E设备，对应多块虚拟网卡，供同一个物理机上的多个虚机使用。由于其可在硬件中实现，因此可以获得能够和物理网卡媲美的I/O性能，通过该标准，虚拟服务器可直接连接到I/O设备，越过了Hypervisor和虚拟交换层，带来极低的处理延迟和接近线缆的速度。

[0050] 同时，通过这些加速单元，虚拟机可将数据包处理工作负载从CPU转移到可编程的物理加速卡上。通过卸载服务器CPU的网络处理工作负载和任务，网络加速装置可大大提高了云和私有数据中心的服务器网络性能。在数据中心网络流量和计算复杂性不断增长的推动下，采用网络加速装置提供了一种处理架构，通过加速盒单元为某些工作负载提供计算，

并从通用计算内核中卸载这些工作负载,从而提高整体虚拟网络解决方案的效率。

[0051] 而对于虚拟网卡上行流量的处理,通过本发明中的加速盒,同一模块内部的虚拟机之间可以通过硬件卸载的虚拟交换机处理,跨节点之间的流量,则通过加速盒内置的高速交换矩阵进行交换。内置高速交换矩阵类似刀片交换机,具有若干外部接口,更多的端口分布在内部,直接连接网络加速单元。硬件加速盒集成了以太网管理接口和web服务器,可以通过管理架构进行管理或web方式进行配置。通过内置的交换矩阵,大大降低了云数据中心网络的线缆数量和排障难度。同时小型化了云计算网络,降低了能耗。不同网络加速盒之间的流量可以通过上联的高速交换矩阵进行交互。

[0052] 高速交换矩阵本身可视为一台交换机,可以通过线缆及交换设备与其它网络加速盒相连,构成一个普通的Fabric网络。高速交换矩阵本身也支持Trunk功能,它和外部的交换机相连,可以做Trunk,从而支持Frame级的负载均衡,使得在数据传输时保证没有拥塞出现。

[0053] 本方法的具体技术实现包括两个方面:

[0054] 1、软件实现

[0055] 虚拟网络的三层功能主要包括路由、SNAT和Floating IP,一般是通过Linux kernel的Namespace来实现的,每个路由器对应一个Namespace,利用Linux TCP/IP协议栈来做路由转发。

[0056] 传统的虚拟网络架构中,虚机端口通过tap口连接到Linux网桥上,进行防火墙网络安全相关功能的处理,处理完成后通过虚拟路由器连接到OpenVSwitch的集成网桥上,进行二层报文处理,如果需要三层路由,再从集成网桥的端口送入虚拟路由器的Linux namespace,查找namespace中配置的网段路由,如果是FIP流量需要再次送入FIP的Linux namespace再次查找,通过namespace中配置的iptables规则对报文进行NAT,再通过浮动IP网关将报文送入Openvswitch的集成网桥中,再通过patch端口将报文送入外网口中发出。总之,虚机流量所经过的冗长的链路跳数,网络性能较低,网络整体性能受限于其中的最短板,拓扑太长,定位问题开销较大,性能受限于iptables条数和namespace个数等。冗长链路上的任何一个节点都可能成为总体网络带宽的瓶颈,网络出现问题后定位的路径较长,网络面临低带宽高延时的问题迫切需要改善。

[0057] 本方法针对虚拟网络现状中存在的上述问题,将报文的处理在OpenVSwitch交换机中一次性处理完成,将上述描述的一个三层报文需要多次进出多个bridge和多个linux namespace才能完成的功能一次性在流表中处理完成,大大缩短包延时,大大提高带宽。除了上述举例中典型的三层转发外,加速功能还包含负载均衡能力、EIP包处理能力等所有虚拟网络功能。

[0058] 为了提升网络性能,本方法将虚拟网的三层功能全部使用OpenFlow流表的方式实现。以安全组为例,传统虚拟网每创建一个端口,就需要创建相应的网络连接设备,而通过流表化的方式,每创建一个端口,直接将其挂载在虚拟交换机(OVS)上即可,从而大大减少了跳数。

[0059] 2、硬件实现

[0060] 基于流表处理的软件处理方式,大幅缩减了路由节点跳数,在此基础上可进一步将处理单元硬件化,迁移到专用的网络加速盒上,从而进一步提升系统的处理性能和稳定

性。如图2所示，其基本的设计思想包括以下几个方面：

[0061] 首先，将网卡从宿主机上迁移到硬件化的网络加速盒上，并通过PCI-E直连宿主机，硬件加速盒内，每一条物理PCI-E对应一个子模块，每个子模块对应一台物理机，每一条物理PCI-E可以虚拟成多个逻辑PCI-E设备，对应多块虚拟网卡，供同一物理机上的多个虚机使用。

[0062] 而每个接入子模块包括支持虚拟通道化的PCI-E网卡、内部交换结构等，接入子模块的功能与普通智能网卡类似，只不过接入子模块无需光模块进行光电信号转化，接入子模块通过内部数字电路与高速交换芯片相连，可构成完整的交换矩阵。

[0063] 最后，在交换矩阵侧提供4-8个高速上联口，用于连接核心交换机，组成更大规模的网络矩阵。

[0064] 通过这个硬件网络加速盒，虚拟机的虚拟网络能力可全部offload到网络侧硬件处理，无需再经过物理机协议栈和内部复杂的虚拟链路，进而提供更好的网络服务能力。

[0065] 本发明实施例还提供一种复杂云计算环境下的网络加速系统，该系统将传统网卡从宿主机上迁移到模块化形态的加速单元设备上，这些单元设备在硬件化后可通过PCI-E连接至宿主机；

[0066] 该系统实现本发明上述实施例所述的复杂云计算环境下的网络加速方法。

[0067] 本发明实施例还提供一种复杂云计算环境下的网络加速装置，包括：至少一个存储器和至少一个处理器、以及上述的硬件网络加速盒子；

[0068] 所述至少一个存储器，用于存储机器可读程序；

[0069] 所述至少一个处理器，用于调用所述机器可读程序，执行本发明上述实施例所述的复杂云计算环境下的网络加速方法。

[0070] 上文通过附图和优选实施例对本发明进行了详细展示和说明，然而本发明不限于这些已揭示的实施例，基与上述多个实施例本领域技术人员可以知晓，可以组合上述不同实施例中的代码审核手段得到本发明更多的实施例，这些实施例也在本发明的保护范围之内。

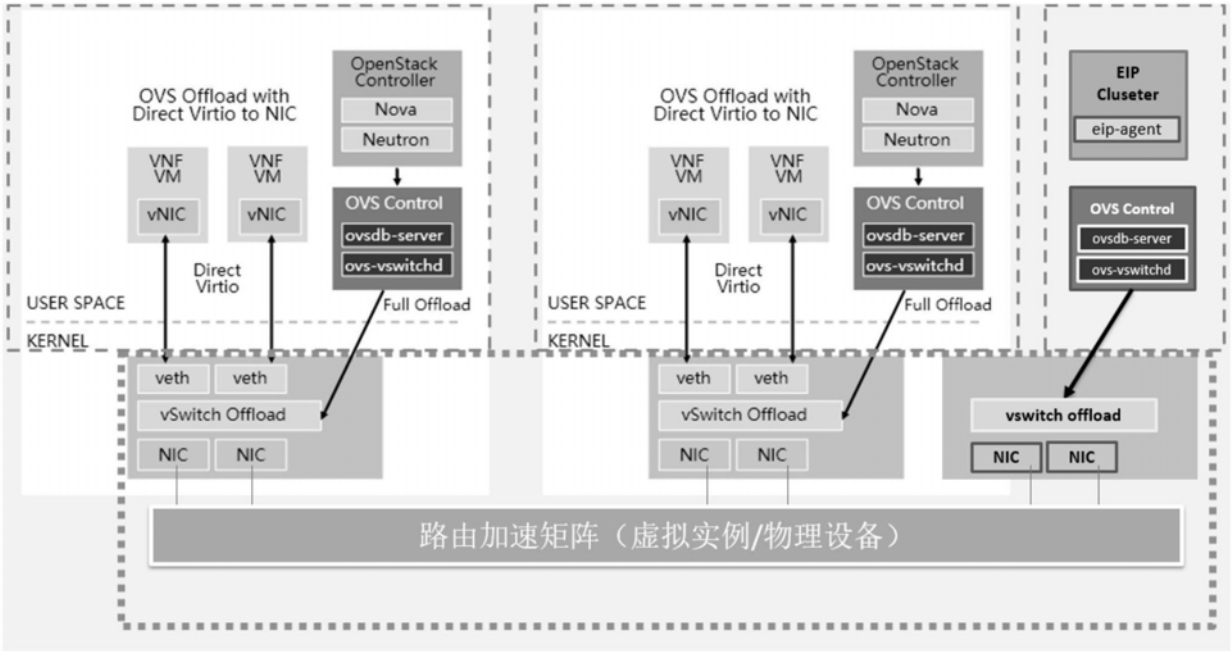

OVS Offload with Direct Virtio to NIC
OpenStack Controller
Nova
Neutron
VNF VM
vNIC
VNF VM
vNIC
OVS Control
ovsdb-server
ovs-vswitchd
Direct Virtio
USER SPACE
KERNEL
Full Offload
veth
veth
vSwitch Offload
NIC
NIC
OVS Offload with Direct Virtio to NIC
OpenStack Controller
Nova
Neutron
VNF VM
vNIC
VNF VM
vNIC
OVS Control
ovsdb-server
ovs-vswitchd
Direct Virtio
USER SPACE
KERNEL
Full Offload
veth
veth
vSwitch Offload
NIC
NIC
EIP Cluseter
eip-agent
OVS Control
ovsdb-server
ovs-vswitchd
vswitch offload
NIC
NIC
路由加速矩阵（虚拟实例/物理设备）

图1

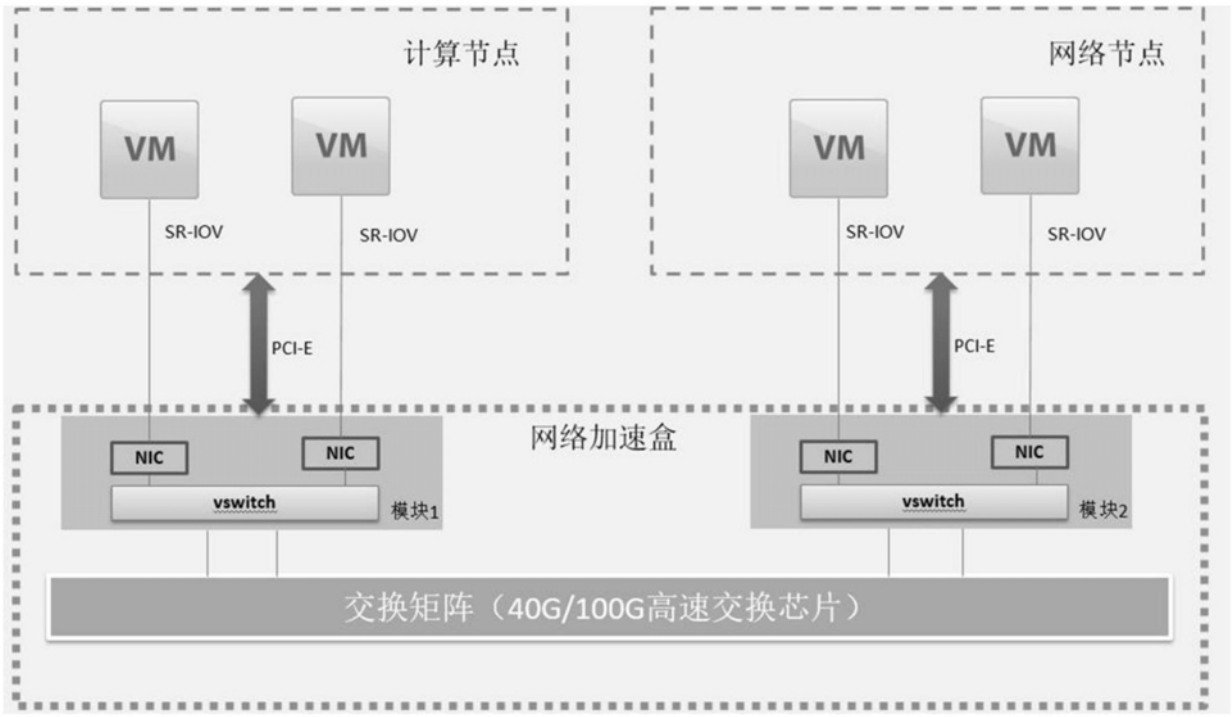

计算节点
VM
VM
SR-IOV
SR-IOV
网络节点
VM
VM
SR-IOV
SR-IOV
PCI-E
PCI-E
网络加速盒
NIC
NIC
vswitch
模块1
NIC
NIC
vswitch
模块2
交换矩阵（40G/100G高速交换芯片）

图2