

基于远程直接数据存储的报文转发方法、装置、网卡及设备

本申请要求 2021 年 07 月 30 日递交的申请号为 202110874970.4、发明名称为“基于远程直接数据存储的报文转发方法、装置及设备”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。

5

技术领域

本发明涉及数据处理技术领域，尤其涉及一种基于远程直接数据存储的报文转发方法、装置及设备。

10 背景技术

随着网络带宽和速度的发展，大数据量数据的迁移需求越来越广泛，网络带宽增长速度远远高于处理网络流量时所必需的计算节点的能力和对内存带宽的需求，数据中心网络架构已经逐步成为计算和存储技术发展的瓶颈，迫切需要采用一种更高效的数据通讯架构。

15 在数据包处理过程中，传统的 TCP/IP 技术要经过操作系统及其他软件层，需要占用大量的服务器资源和内存总线带宽，并且，由于系统庞大的开销、数据在系统内存、处理器缓存和网络控制器缓存之间来回进行复制移动，给服务器的 CPU 和内存造成了沉重负担，尤其是面对网络带宽、处理器速度与内存带宽三者的严重“不匹配性”，更容易造成网络延迟效应的加剧。

20

发明内容

本发明实施例提供了一种基于远程直接数据存储的报文转发方法、装置及设备，实现了采用软硬件结合的方式进行报文处理操作，对网络性能和报文处理的灵活性进行了有效平衡，这样有利于减少对带宽和处理器开销的需求，同时可以显著降低时延。

25 第一方面，本发明实施例提供一种基于远程直接数据存储的报文转发方法，包括：
获取与待转发 RDMA 报文相对应的静荷数据；
利用网卡节点中的硬件模块对所述静荷数据的报文头进行封装处理，获得处理后报文；
利用网卡节点中的硬件模块或软件模块对处理后报文进行转发。

30 第二方面，本发明实施例提供一种基于远程直接数据存储的报文转发装置，包括：
物理机，用于生成与待转发 RDMA 报文相对应的静荷数据；
硬件模块，与所述物理机通信连接，用于获取与待转发 RDMA 报文相对应的静荷数据，并对所述静荷数据的报文头进行封装处理，获得处理后报文；
硬件模块/软件模块，用于对处理后报文进行转发。

第三方面，本发明实施例提供一种电子设备，包括：存储器、处理器；其中，所述存储器用于存储一条或多条计算机指令，其中，所述一条或多条计算机指令被所述处理器执行时实现上述第一方面中的基于远程直接数据存储的报文转发方法。

第四方面，本发明实施例提供了一种计算机存储介质，用于储存计算机程序，所述计算机程序使计算机执行时实现上述第一方面中的基于远程直接数据存储的报文转发方法。

第五方面，本发明实施例提供了一种计算机程序产品，包括：存储有计算机指令的计算机可读存储介质，当所述计算机指令被一个或多个处理器执行时，致使所述一个或多个处理器执行上述第一方面所示的基于远程直接数据存储的报文转发方法中的步骤。

第六方面，本发明实施例提供了一种网卡，包括：

硬件模块，用于获取与待转发 RDMA 报文相对应的静荷数据，并对所述静荷数据的报文头进行封装处理，获得处理后报文；

硬件模块/软件模块，用于对处理后报文进行转发。

本实施例提供的技术方案，通过获取与待转发 RDMA 报文相对应的静荷数据，利用网卡节点中的硬件模块对所述静荷数据的报文头进行封装处理，获得处理后报文，之后则利用所述网卡节点中的软件模块或者硬件模块对所述处理后报文进行转发，有效地实现了采用软硬件结合的方式进行报文处理操作，具体硬件部分可以负责报文的封装/解封装处理、报文的快转操作；软件部分负责报文的正常转发操作，专注于状态管理，进而实现了对网络性能和报文处理的灵活性进行了有效平衡，有利于减少对带宽和处理器开销的需求，显著降低了时延；另外，该技术方案能够基于 overlay 网络架构实现对 RDMA 技术的应用，这样不仅能够屏蔽物理网络的实现细节，并且没有物理网络范围的限制，没有同子网通信限制；此外，该报文转发方法可以复用正常的 VPC 网络流量的虚拟交换机处理逻辑，如安全组等，进一步提高了该报文转发方法使用的灵活可靠性。

附图说明

为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图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 为相关技术中实施例提供的一种报文转发方法的原理示意图；

图 2 为本发明实施例提供的一种基于远程直接数据存储的报文转发方法的原理框图；

图 3 为本发明实施例提供的一种基于远程直接数据存储的报文转发方法的流程示意图；

图 4 为本发明实施例提供的利用所述网卡节点中的硬件模块对所述静荷数据的报文头进行封装处理，获得处理后报文的流程示意图；

图 5 为本发明实施例提供的对所述处理后报文进行转发的流程示意图；

图 6 为本发明实施例提供的另一种基于远程直接数据存储的报文转发方法的流程图；

图 7 为本发明实施例提供的又一种基于远程直接数据存储的报文转发方法的流程图；

图 8 为本发明实施例提供的再一种基于远程直接数据存储的报文转发方法的流程图；

图 9 为本发明应用实施例提供的一种基于远程直接数据存储的建连过程的示意图；

图 10 为本发明应用实施例提供的一种基于远程直接数据存储的数据传输过程的示意图；

图 11 为本发明实施例提供的一种基于远程直接数据存储的数据传输装置的结构示意图；

图 12 为与图 11 所示实施例提供的数据传输装置对应的电子设备的结构示意图；

图 13 为本发明实施例提供的一种网卡的结构示意图。

具体实施方式

为使本发明实施例的目的、技术方案和优点更加清楚，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一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

在本发明实施例中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本发明。在本发明实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义，“多种”一般包含至少两种，但是不排除包含至少一种的情况。

应当理解，本文中使用的术语“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。

取决于语境，如在此所使用的词语“如果”、“若”可以被解释成为“在……时”或“当……时”或“响应于确定”或“响应于检测”。类似地，取决于语境，短语“如果确定”或“如果检测（陈述的条件或事件）”可以被解释成为“当确定时”或“响应

于确定”或“当检测（陈述的条件或事件）时”或“响应于检测（陈述的条件或事件）”。

还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的商品或者系统不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种商品或者系统所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的商品或者系统中还存在另外的相同要素。

另外，下述各方法实施例中的步骤时序仅为一种举例，而非严格限定。

为了便于理解本实施例中基于远程直接数据存储 RDMA 的报文转发方法的具体实现过程和实现效果，下面对相关技术进行简要说明：

10 随着网络带宽和速度的发展，大数据量数据的迁移需求越来越广泛，网络带宽增长速度远远高于处理网络流量时所必需的计算节点的能力和对内存带宽的需求，数据中心网络架构已经逐步成为计算和存储技术发展的瓶颈，迫切需要采用一种更高效的数据通讯架构。

在数据包处理过程中，传统的 TCP/IP 技术要经过操作系统及其他软件层，如图 1 所示的链路 1，待处理报文由应用端发送至内核，需要经过内核的套接层 Sockets、传输控制协议层 TCP、因特网互联协议 IPv4/IPv6 层、网络设备层 Network Device 以及设备驱动层 Device Driver，此时，对于待处理报文而言，不仅传输链路非常长，并且，在对待处理报文进行传输时，需要进行报文的拷贝和传输操作，而报文的拷贝和传输操作都是由软件模块 CPU 来做的，这样需要占用大量的服务器资源和内存总线带宽，进而会产生较大的延时。总的来说，由于系统庞大的开销、数据在系统内存、处理器缓存和网络控制器缓存之间来回进行复制移动，给服务器的 CPU 和内存造成了沉重负担，尤其是面对网络带宽、处理器速度与内存带宽三者的严重“不匹配性”，更容易造成网络延迟效应的加剧。

25 远程直接数据存取(Remote Direct Memory Access, 简称 RDMA)能够让计算机可以直接存取其它计算机的内存，而不需要经过处理器耗时的传输，无需双方操作系统的介入，允许高吞吐、低延迟的网络通信，尤其适合在大规模并行计算机集群中使用。如图 1 所示的链路 2，RDMA 技术具有将数据从一个系统快速移动到远程系统存储器中，而不对操作系统造成任何影响的特性，这项技术通过消除外部存储器复制和文本交换操作，因而能腾出总线空间和 CPU 周期，用于改进应用系统性能，从而减少对带宽和处理器开销的需要，显著降低时延。

30 然而，将 RDMA 技术应用在云网络中具有较大难度，具体理由如下：首先，在云网络中，用户使用的都是虚拟机，例如：云服务器(Elastic Compute Service, 简称 ECS)，其网卡通常是由软件虚拟化实现的，不具备硬件网卡的 RDMA 能力；其次，当前云网络通常都是基于 overlay 网络（将传统网络报文作为静荷进行封装，例如：封装为 Vxlan）的

私有网络(Virtual Private Cloud,简称 VPC)网络,无法使用传统的 RDMA 技术;最后,由于云网络中的网络虚拟化,用户的网络不想受到物理网络拓扑的限制,而传统 RDMA 对物理网络的感知是非常重要的。

为了能够提供一种支持在云网络 overlay 网络架构下的 RDMA 技术方案,本实施例
5 提供了一种基于远程直接数据存储 RDMA 的报文转发方法、装置及设备,上述报文转发方法的执行主体可以为报文转发系统,报文转发系统可以包括:物理机和与物理机通信连接的网卡,例如:网卡能够插在物理机上,物理机内可以根据应用场景或者使用需求启动一个或多个虚拟机 VM,如图 2 所示,网卡中可以包括软件模块(CPU)和硬件模块,其中,软件模块中可以包括负责正常的 VPC 网络流量转发操作的虚拟交换机 vswitch;
10 具体的,软件模块可以获取报文的五元组特征,通过五元组特征查询路由表、ACL 表等各种转发表,获得查询结果,而后基于查询结果和五元组特征生成第二流表。硬件模块可以采用现场可编程逻辑门阵列(Field Programmable Gate Array,简称 FPGA)或者专用集成电路(Application Specific Integrated Circuit,简称 ASIC),该硬件模块可以卸载由软件 vswitch 下发的第二流表。具体的,在进行报文转发操作时:

15 物理机/虚拟机:生成待转发报文,此时的待转发报文中可以仅包括静荷数据,将待转发报文发送至网卡;

网卡中的硬件模块:获取物理机/虚拟机所发送的待转发报文,对待转发报文的报文头进行封装处理,获得处理后报文,即将仅包括静荷数据的报文封装为完整报文。

网卡中的硬件模块/软件模块:用于对处理后报文进行转发。具体的,先检测处理后
20 报文是否命中硬件模块中预先存储的第一流表,在处理后报文命中第一流表时,则硬件模块可以直接基于第一流表对处理后报文进行转发;在处理后报文未命中第一流表时,则将处理后报文发送至网卡中的软件模块;网卡中的软件模块获取到硬件模块发送的处理后报文,并利用预先存储的与处理后报文相对应的第二流表对处理后报文进行转发处理。在对处理后报文进行转发处理后,软件模块还可以将与处理后报文相对应的第二流
25 表发送至硬件模块,以作为第一流表。

本实施例提供的技术方案,实现了采用软硬件结合的方式对报文进行处理,具体硬件部分可以负责报文的封装/解封装处理、报文的快转操作;软件部分负责报文的正常转发操作,专注于状态管理,进而实现了网络性能和灵活性的平衡,有利于减少对带宽和处理器开销的需求,显著降低了时延;另外,能够基于 overlay 网络架构实现对 RDMA
30 技术的应用,这样不仅能够屏蔽物理网络实现细节,并且没有物理网络范围的限制,没有同子网通信限制;此外,该报文转发方法可以复用正常的 VPC 网络流量的虚拟交换机处理逻辑,如安全组等,进一步提高了该报文转发方法使用的灵活可靠性。

下面结合附图,对本发明的一些实施方式作详细说明。在各实施例之间不冲突的情况下,下述的实施例及实施例中的特征可以相互组合。

图 3 为本发明实施例提供的一种基于远程直接数据存储的报文转发方法的流程示意图；参考附图 3 所示，本实施例提供了一种基于远程直接数据存储 RDMA 的报文转发方法，该方法的执行主体为基于远程直接数据存储的报文转发装置，可以理解的是，该报文转发装置可以实现为软件、或者软件和硬件的组合，具体实现时，该报文转发装置可以

5 以实现为一网卡节点。具体的，该基于远程直接数据存储的报文转发方法可以包括：

步骤 S301：获取与待转发 RDMA 报文相对应的静荷数据。

步骤 S302：利用网卡节点中的硬件模块对静荷数据的报文头进行封装处理，获得处理后报文。

步骤 S303：利用网卡节点中的硬件模块或软件模块对处理后报文进行转发。

10 下面对上述各个步骤的具体实现过程和实现效果进行详细说明：

步骤 S301：获取与待转发 RDMA 报文相对应的静荷数据。

其中，在用户存在数据访问需求时，则可以通过物理机或者虚拟机生成待转发 RDMA 报文，此时所生成的待转发 RDMA 报文中仅包括静荷数据。具体的，物理机可以提供用于供用户输入执行操作的接口或者端口，通过接口或者端口可以获取到用户输入的执行

15 操作，通过执行操作可以生成待转发 RDMA 报文，可以理解的是，在不同的应用场景中，所产生的待转发 RDMA 报文可以对应有不同的功能作用，例如：用于实现数据访问操作的待转发 RDMA 报文、用于实现控制操作的待转发 RDMA 报文、用于实现响应操作的待转发 RDMA 报文等等。

在物理机或者虚拟机生成待转发报文之后，可以将待转发报文发送至网卡节点，可

20 以理解的是，网卡节点所能够获取到的待转发报文的数量可以为一个或多个；具体实现时，物理机或者虚拟机上配置有用于 RDMA 数据传输操作的驱动模块，通过驱动模块可以将待转发报文由物理机或者虚拟机发送至网卡节点，从而使得网卡节点可以获取到与待转发 RDMA 报文相对应的静荷数据。

步骤 S302：利用网卡节点中的硬件模块对静荷数据的报文头进行封装处理，获得处

25 理后报文。

其中，网卡节点中可以包括硬件模块和软件模块，硬件模块用于实现报文的封装、解封装以及报文的快转操作；软件模块用于实现报文的正常转发操作、报文的

30 状态管理操作，上述的硬件模块可以采用 FPGA 或者 ASIC 来实现，软件模块可以由 CPU 来实现。在网卡节点获取到静荷数据之后，为了能够使得该报文转发方法可以适用于 overlay 网络架构，可以利用网卡节点中的硬件模块对静荷数据的报文头进行封装处理，静荷数据的报文头可以包括：传输控制协议 TCP 头信息、因特网互联协议 IP 头信息和介质访问控制 MAC 头信息，从而可以获得完整的处理后报文，该处理后报文可以在 overlay 网络架构中进行传输。

步骤 S303：利用网卡节点中的硬件模块或软件模块对处理后报文进行转发。

在获取到处理后报文之后，可以利用网卡节点中的硬件模块或者软件模块对处理后报文进行转发。在一些实例中，利用网卡节点中的硬件模块或软件模块对处理后报文进行转发可以包括：检测所述处理后报文是否命中第一流表，其中，所述第一流表存储在所述硬件模块中；在所述处理后报文命中第一流表时，则利用所述硬件模块和第一流表对所述处理后报文进行转发；在所述处理后报文未命中所述第一流表时，则利用所述软件模块和第二流表对所述处理后报文进行转发，其中，所述第二流表存储在所述软件模块中。

其中，对于网卡节点而言，硬件模块中预先存储有用于实现流量快转操作的第一流表，软件模块中预先存储有用于实现流量转发操作的第二流表。在硬件模块获取到处理后报文之后，可以对处理后报文进行分析检测，以检测处理后报文是否命中第一流表。具体的，检测处理后报文是否命中第一流表可以包括：确定与处理后报文相对应的五元组信息（源 IP 地址、目的 IP 地址、协议号、源端口、目的端口）以及执行操作信息，基于五元组以及执行操作检测处理后报文是否命中第一流表；在第一流表中包括与处理后报文相对应的五元组以及执行操作时，则确定处理后报文命中第一流表；在第一流表中不包括与处理后报文相对应的五元组以及执行操作时，则确定处理后报文未命中第一流表。

在处理后报文命中第一流表时，则说明此时硬件模块能够对处理后报文进行直接转发操作，进而可以直接利用硬件模块和第一流表对处理后报文进行转发操作，这样有利于提高对处理后报文进行转发操作的质量和效率。在处理后报文未命中第一流表时，则说明此时硬件模块不能对处理后报文进行直接转发操作，进而可以利用网卡节点中的软件模块和软件模块中存储的第二流表对处理后报文进行转发操作，这样有效地实现了对处理后报文进行正常的转发操作。

在又一些实例中，为了提高该方法的实用性，在利用网卡节点中的软件模块和第二流表对处理后报文进行转发之后，本实施例中的方法还可以包括：获取与处理后报文相对应的目标流表；将目标流表发送至硬件模块，以作为第一流表。

具体的，在利用网卡节点中的软件模块和第二流表对处理后报文进行转发之后，由于此时的硬件模块中并不包括与处理后报文相对应的第一流表，为了提高对之后的待处理报文进行转发操作的质量和效率，软件模块可以获取与处理后报文相对应的目标流表，而后将目标流表发送至硬件模块，以作为第一流表，从而有效地实现了第一流表的不断更新操作。

举例来说，在某一区域中的网卡节点可以获取到用户 A 的待处理报文、用户 B 的待处理报文，此时，网卡节点中的软件模块中包括与上述用户 A 和用户 B 相对应的第二流表，硬件模块中包括与上述用户 A 和用户 B 相对应的第一流表；在网卡节点第一次接收到用户 C 的待处理报文时，此时网卡节点中的硬件模块并不包括用户 C 相对应的第一流

表，因此，利用网卡节点中的软件模块对用户 C 的待处理报文进行转发处理；在处理完毕之后，则可以获取与用户 C 的待处理报文相对应的目标流表，而后将目标流表发送至硬件模块，以作为第一流表，此时，硬件模块中所包括的第一流表包括：与用户 A 相对应的第一流表、与用户 B 相对应的第一流表和与用户 C 相对应的快转流程。

- 5 这样，之后在获取到用户 C 所对应的待处理报文时，可以直接利用网卡节点中的硬件模块对待处理报文进行转发处理，有利于提高报文转发操作的质量和效率，减少报文转发延时。

本实施例提供的基于远程直接数据存储的报文转发方法，通过获取与待转发 RDMA 报文相对应的静荷数据，利用网卡节点中的硬件模块对所述静荷数据的报文头进行封装处理，获得处理后报文，之后检测所述处理后报文是否命中第一流表，在所述处理后报
10 文命中第一流表时，则利用所述硬件模块和所述第一流表对所述处理后报文进行转发；在所述处理后报文未命中所述第一流表时，则利用所述网卡节点中的软件模块和第二流表对所述处理后报文进行转发，有效地实现了采用软硬件结合的方式进行报文处理操作，具体硬件部分可以负责报文的封装/解封装处理、报文的快转操作；软件部分负责报文的正常转发操作，专注于状态管理，进而实现了对网络性能和报文处理的灵活性进行了有效平衡，有利于减少对带宽和处理器开销的需求，显著降低了时延；另外，该技术方案能够基于 overlay 网络架构实现对 RDMA 技术的应用，这样不仅能够屏蔽物理网络的实现细节，并且没有物理网络范围的限制，没有同子网通信限制；此外，该报文转发方法可以复用正常的 VPC 网络流量的虚拟交换机处理逻辑，如安全组等，进一步提高了该报
15 文转发方法使用的灵活可靠性。

图 4 为本发明实施例提供的利用网卡节点中的硬件模块对静荷数据的报文头进行封装处理，获得处理后报文的流程示意图；参考附图 4 所示，本实施例提供了一种对静荷数据的报文头进行封装处理，获得处理后报文的实现方式，具体的，本实施例中的利用网卡节点中的硬件模块对静荷数据的报文头进行封装处理，获得处理后报文可以包括：

- 25 步骤 S401：利用硬件模块获取静荷数据的数据描述符，数据描述符用于标识静荷数据的长度和 RDMA 头信息。

对于静荷数据而言，一般数据长度比较长，例如：某一静荷数据包括 1500 个字节，此时，对静荷数据进行处理操作需要占用大量的资源，为了减少数据处理所占用的资源，在硬件模块获取到静荷数据之后，可以对静荷数据进行分析处理，以获取用于标识静荷数据长度和 RDMA 头信息的数据描述符，上述的 RDMA 头信息可以包括：直接数据放置（direct data placement，简称 DDP）信息、基于标记的协议数据单元对齐头信息（marker-based, protocol data unit aligned，简称 MPA）。

步骤 S402：利用数据描述符，获取与静荷数据相对应的封装信息。

在获取到数据描述符之后，可以利用数据描述符来获取与静荷数据相对应的封装信

息，在一些实例中，封装信息存储在软件模块中，此时，利用数据描述符，获取与静荷数据相对应的封装信息可以包括：利用硬件模块将数据描述符发送至软件模块；利用软件模块和数据描述符，生成与静荷数据相对应的封装信息，并将封装信息 and 数据描述符发送至硬件模块。

- 5 由于封装信息存储在软件模块中，因此，为了能够获取到封装信息，可以利用硬件模块将数据描述符发送至软件模块，软件模块接收到数据描述符之后，可以利用数据描述符生成与静荷数据相对应的封装信息，该封装信息可以包括：数据序列号、TCP 封装信息等等。具体的，利用软件模块和数据描述符，生成与静荷数据相对应的封装信息可以包括：利用软件模块对数据描述符进行协议栈处理和拥塞控制处理，生成与静荷数据
10 相对应的封装信息，其中，封装信息中包括：用于标识拥塞控制的拥塞控制值、传输控制协议 TCP 头信息、因特网互联协议 IP 头信息和介质访问控制 MAC 头信息。

- 在软件模块获取到封装信息之后，可以将封装信息和与封装信息相对应的数据描述符发送至硬件模块，从而使得硬件模块可以准确、有效地获取到与静荷数据的封装信息。当然的，本领域技术人员也可以采用其他的方式来获取与静荷数据相对应的封装信息，
15 只要能够保证硬件模块准确地获取到封装信息即可，在此不再赘述。

 步骤 S403：利用硬件模块和封装信息对静荷数据的报文头进行封装处理，获得处理后报文。

- 在获取到封装信息之后，可以利用硬件模块和封装信息对静荷数据的报文头进行封装处理，具体的，硬件模块可以利用封装信息对静荷数据的 TCP 头信息、IP 头信息和
20 MAC 头信息进行封装操作，获得完整的处理后报文。

- 本实施例中，通过利用硬件模块获取静荷数据的数据描述符，利用数据描述符来获取与静荷数据相对应的封装信息，而后利用硬件模块和封装信息对静荷数据的报文头进行封装处理，获得处理后报文，在保证对静荷数据的报文头进行准确、有效地封装处理的基础上，减少了封装操作所需要的数据处理资源，有利于降低数据处理延时，进一步
25 提高了报文转发操作的质量和效率。

 图 5 为本发明实施例提供的对处理后报文进行转发的流程示意图；参考附图 5 所示，在上述实施例中，无论是利用硬件模块和第一流表对处理后报文进行转发，还是利用网卡节点中的软件模块和第二流表对处理后报文进行转发，本实施例提供了一种对处理后进行转发操作的实现方式，具体的，本实施例中的对处理后报文进行转发可以包括：

- 30 步骤 S501：对处理后报文进行 VxLAN 封装，获得目标报文。

 步骤 S502：获取与目标报文相对应的拥塞控制值。

 步骤 S503：将拥塞控制值拷贝到外层，并在外层生成与目标报文相对应的报文格式标识，报文格式标识用于标识目标报文为 RDMA 报文。

 步骤 S504：基于拥塞控制值、报文格式标识和流表转发规则，对目标报文进行转发。

其中，在对处理后报文进行转发处理操作时，为了保证报文转发操作的稳定可靠性，可以先对处理后报文进行 VxLAN 封装，获得目标报文。在获取到目标报文之后，可以获取与目标报文相对应的拥塞控制值，具体的，获取与目标报文相对应的拥塞控制值可以包括：对处理后报文进行分析处理，以确定与处理后报文相对应的拥塞控制值，将与处理后报文相对应的拥塞控制值确定为与目标报文相对应的拥塞控制值，该拥塞控制值用于标识对目标报文进行转发操作的网络传输速率。

由于所确定的拥塞控制值在网络内层，用户端无法感知到所获得的拥塞控制值，进而无法基于拥塞控制值对目标报文进行转发控制，因此，在获取到拥塞控制值之后，可以将拥塞控制值拷贝到外层，并在外层生成与目标报文相对应的报文格式标识，该报文格式标识用于标识目标报文为 RDMA 报文。在外层对应有拥塞控制值、报文格式标识时，则可以直接基于拥塞控制值、报文格式标识和流表转发规则对目标报文进行转发操作，进而保证了对目标报文进行转发操作的稳定可靠性。

具体的，上述实现方式可以应用在以下两种场景中：

(1) 在处理后报文命中第一流表时，硬件模块可以基于第一流表中的规则对处理后报文进行 VxLAN 封装，获得目标报文；而后获取与目标报文相对应的拥塞控制值；将拥塞控制值拷贝到外层，并在外层生成与目标报文相对应的报文格式标识，而后基于拥塞控制值、报文格式标识和第一流表中的规则对目标报文进行转发。

(2) 在处理后报文未命中第一流表时，硬件模块可以将处理后报文发送至软件模块中的虚拟交换机（vswitch）进行处理，vswitch 可以对处理后报文进行 VxLAN 封装，获得目标报文；而后获取与目标报文相对应的拥塞控制值；将拥塞控制值拷贝到外层，并在外层生成与目标报文相对应的报文格式标识，而后基于拥塞控制值、报文格式标识和第二流表中的规则对目标报文进行转发，然后转发出报文。此时，软件模块还可以将与目标报文相对应的第二流表发送至目标模块，以作为第一流表。

本实施例中，通过对处理后报文进行 VxLAN 封装，获得目标报文，获取与目标报文相对应的拥塞控制值，而后将拥塞控制值拷贝到外层，并在外层生成与目标报文相对应的报文格式标识，并基于拥塞控制值、报文格式标识和流表转发规则，对目标报文进行转发，有效地实现了对目标报文进行稳定、有效地转发操作，进一步提高了该报文转发方法使用的安全可靠。

图 6 为本发明实施例提供的另一种基于远程直接数据存储的报文转发方法的流程示意图；参考附图 6 所示，本实施例中的方法还可以包括：

步骤 S601：获取待转发报文。

在报文转发装置应用在接收端时，报文转发装置可以接收到用于转发到目的端的待转发报文，此时的待转发报文可以包括：静荷数据、拥塞控制值、传输控制协议 TCP 头信息、因特网互联协议 IP 头信息和介质访问控制 MAC 头信息等等。

步骤 S602：利用网卡节点中的硬件模块对待转发报文进行报文头的解封装操作，获得与待转发报文相对应的解封装报文。

在网卡节点获取到待转发报文之后，为了保证对待转发报文进行传输的质量和效率，可以利用网卡节点中的硬件模块对待转发报文进行报文头的解封装操作，具体可以包括对待转发报文的传输控制协议 TCP 头信息、因特网互联协议 IP 头信息和介质访问控制 MAC 头信息等进行解封装操作，从而可以获得与待转发报文相对应的解封装报文。

步骤 S603：利用所述硬件模块或软件模块对解封装报文进行转发。

在获取到解封装报文之后，可以利用网卡节点中所包括的硬件模块或软件模块对解封装报文进行转发处理，在一些实例中，利用所述硬件模块或软件模块对解封装报文进行转发可以包括：检测所述解封装报文是否命中第一流表；在所述解封装报文命中第一流表时，则利用所述硬件模块和所述第一流表对所述解封装报文进行转发；在所述解封装报文未命中所述第一流表时，则利用所述软件模块和第二流表对所述解封装报文进行转发。

具体的，在获取到解封装报文之后，可以对解封装报文进行分析处理，以检测解封装报文是否命中第一流表，其中，本实施例中步骤的实现方式和实现效果与上述实施例中步骤 S303 的具体实现方式和实现效果相类似，具体可参考上述陈述内容，在此不再赘述。

在解封装报文命中第一流表时，则说明此时硬件模块能够对解封装报文进行直接转发操作，进而可以直接利用硬件模块和第一流表对解封装报文进行转发操作，这样有利于提高对解封装报文进行转发操作的质量和效率。在一些实例中，利用硬件模块和第一流表对待转发报文进行转发可以包括：利用硬件模块和第一流表对解封装报文进行 VxLAN 解封装，获得目标报文；获取与目标报文相对应的拥塞控制值，将拥塞控制值拷贝到内层；基于拥塞控制值和第一流表，对目标报文进行转发。

具体的，在对解封装报文进行转发处理操作时，由于解封装报文是经过 VxLAN 封装操作之后的报文，此时，为了保证报文转发操作的稳定可靠性，可以先对解封装报文进行 VxLAN 解封装，获得目标报文。在获取到目标报文之后，可以获取与目标报文相对应的拥塞控制值，具体的，获取与目标报文相对应的拥塞控制值可以包括：对解封装报文进行分析处理，以确定与解封装报文相对应的拥塞控制值，将与解封装报文相对应的拥塞控制值确定为与目标报文相对应的拥塞控制值，该拥塞控制值用于标识对目标报文进行转发操作的网络传输速率。

由于所确定的拥塞控制值在外层，无法直接基于拥塞控制值对目标报文进行转发控制，因此，在获取到拥塞控制值之后，可以将拥塞控制值拷贝到内层，之后则可以直接基于拥塞控制值和第一流表中的转发规则对目标报文进行转发操作，即将目标报文发送

至目标端，进而保证了对目标报文进行转发操作的稳定可靠性。

在另一些实例中，在获得目标报文之后，本实施例中的方法还可以包括：生成并存储与目标报文相对应的数据序列值，数据序列值与用于对报文数据进行封装处理的封装信息相关。

- 5 其中，在对报文进行转发时，若所接收到的报文仅包括静荷数据时，为了能够实现对报文的正常转发操作，需要对报文进行封装操作，此时，可以利用软件模块中所包括的封装信息对报文进行封装操作，而上述的封装信息中包括数据序列值。为了能够保证不同的报文可以对应有不同的数据序列值。在获得目标报文之后，软件模块会生成并存储与目标报文相对应的数据序列值，可以理解的是，不同的目标报文可以生成不同的数据序列值。举例来说：目标报文包括：报文 1、报文 2 和报文 3，在生成报文 1 时，软件模块会生成并存储与报文 1 相对应的序列 1，例如，序列 1 可以为 202107150032；在生成报文 2 时，软件模块会生成并存储与报文 2 相对应的序列 2，例如，序列 2 可以为 202107150033；在生成报文 3 时，软件模块会生成并存储与报文 3 相对应的序列 3，例如，序列 3 可以为 202107150033。通过生成并存储与目标报文相对应的数据序列值，而所存储的数据序列值与对报文数据进行封装处理的封装信息相关，因此可以保证对报文进行封装操作的稳定可靠性。

另外，在解封装报文未命中第一流表时，则说明此时硬件模块不能对解封装报文进行直接转发操作，进而可以利用网卡节点中的软件模块和软件模块中存储的第二流表对解封装报文进行转发操作，这样有效地实现了对解封装报文进行正常的转发操作。

- 20 在一些实例中，利用软件模块和第二流表对解封装报文进行转发可以包括：利用软件模块和第二流表获取与解封装报文相对应的报文格式标识；在基于报文格式标识确定解封装报文为 RDMA 报文时，基于第二流表对解封装报文进行 VxLAN 解封装，获得目标报文；获取与目标报文相对应的拥塞控制值，并将拥塞控制值拷贝到内层；基于拥塞控制值和第二流表，对目标报文进行转发。

- 25 其中，对于软件模块而言，在接收到 RDMA 报文时，能够基于软件模块中包括的第二流表对 RDMA 报文进行正常的转发操作，而对于非 RDMA 报文则不能进行正常的转发操作。因此，在获取到解封装报文之后，可以先识别解封装报文所对应的报文类型，具体的，可以利用软件模块和第二流表获取与解封装报文相对应的报文格式标识，在报文格式标识为用于标识 RDMA 报文的标识时，则可以确定解封装报文为 RDMA 报文，
30 否则可以确定解封装报文为非 RDMA 报文。

在基于报文格式标识确定解封装报文为 RDMA 报文时，说明此时的软件模块可以对解封装报文进行正常的报文转发操作，此时，为了保证报文转发操作的稳定可靠性，可以先对解封装报文进行 VxLAN 解封装，获得目标报文。在获取到目标报文之后，可以获取与目标报文相对应的拥塞控制值，具体的，获取与目标报文相对应的拥塞控制值可

以包括：对解封装报文进行分析处理，以确定与解封装报文相对应的拥塞控制值，将与解封装报文相对应的拥塞控制值确定为与目标报文相对应的拥塞控制值，该拥塞控制值用于标识对目标报文进行转发操作的网络传输速率。

5 由于所确定的拥塞控制值在外层，无法直接基于拥塞控制值对目标报文进行转发控制，因此，在获取到拥塞控制值之后，可以将拥塞控制值拷贝到内层，之后则可以直接基于拥塞控制值和第二流表中的转发规则对目标报文进行转发操作，即将目标报文发送至目标端，进而保证了对目标报文进行转发操作的稳定可靠性。

本实施例中，通过获取待转发报文，利用网卡节点中的硬件模块对待转发报文进行报文头的解封装操作，获得与待转发报文相对应的解封装报文，检测解封装报文是否命中第一流表，在解封装报文命中第一流表时，则利用硬件模块和第一流表对解封装报文进行转发，在解封装报文未命中第一流表时，则利用软件模块和第二流表对解封装报文进行转发，有效地实现了采用软硬件结合的方式对接收端所接收到的用于发送到目的端的待转发报文进行转发操作，进一步提高了该报文转发方法的实用性和适用范围。

10 图 7 为本发明实施例提供的又一种基于远程直接数据存储的报文转发方法的流程示意图；参考附图 7 所示，在利用网卡节点中的硬件模块对待转发报文进行报文头的解封装操作之后，本实施例中的方法还可以包括：

步骤 S701：获得与待转发报文相对应的报文描述符。

步骤 S702：将解封装报文和报文描述符发送至软件模块。

20 步骤 S703：利用软件模块、解封装报文和报文描述符，对用于实现 RDMA 报文传输操作的连接管理信息进行更新。

其中，软件模块中存储有用于实现 RDMA 报文传输操作的连接管理信息，连接管理信息中包括数据序列值、序列号等等，该连接管理信息用于实现对所有的报文转发操作进行管控。为了能够实现基于连接管理信息对报文转发操作进行稳定、有效地管控，在利用网卡节点中的硬件模块对待转发报文进行报文头的解封装操作之后，即将待转发报文解封掉 ip 头、mac 头和 TCP 头之后，硬件模块可以对待转发报文进行分析处理，以获得与待转发报文相对应的报文描述符，该报文描述符用于标识待转发报文的长度信息和 RDMA 头信息。在获取到解封装报文和报文描述符之后，可以将解封装报文和报文描述符发送至软件模块，软件模块获取到解封装报文和报文描述符之后，可以基于解封装报文和报文描述符对用于实现 RDMA 报文传输操作的连接管理信息进行更新，从而可以获得更新后管理信息，而后可以利用更新后管理信息对之后的报文转发操作进行管理和控制，有效地保证了报文转发操作的稳定可靠性。

图 8 为本发明实施例提供的再一种基于远程直接数据存储的报文转发方法的流程示意图；在上述任意一个实施例的基础上，参考附图 8 所示，本实施例提供了一种用于实现 RDMA 数据传输操作的建连过程，具体的，在获取与待转发 RDMA 报文相对应的静

荷数据之前，本实施例中的方法还可以包括：

步骤 S801：获取用于实现 RDMA 建连操作的建连静荷数据。

步骤 S802：利用软件模块中的虚拟交换机对建连静荷数据的报文头进行封装处理，获得建连报文。

5 步骤 S803：利用软件模块中的软件管理器确定与建连报文相对应的建连流表，并将建连流表发送至虚拟交换机。

步骤 S804：虚拟交换机基于建连流表对建连报文进行转发处理，以实现 RDMA 建连操作。

10 其中，用于实现 RDMA 建连操作的建连静荷数据可以是主机或者虚拟机发送至基于远程直接数据存储的报文转发装置的，此时，网卡节点中的硬件模块中并没有与建连静荷数据相对应的第一流表，因此，在网卡节点获取到用于实现 RDMA 建连操作的建连静荷数据之后，可以利用软件模块中的虚拟交换机对建连静荷数据的报文头进行封装处理，具体的，可以将建连静荷报文封装 Vxlan 头部，从而可以获得建连报文。

15 在获取到建连报文之后，可以利用软件模块中的软件管理器来生成并确定与建连报文相对应的建连流表，在生成并确定建连流表之后，为了能够实现对建连报文的转发操作，则可以将建连流表发送至虚拟交换机，从而使得虚拟交换机稳定地获取到建连流表。在虚拟交换机获取到建连流表之后，可以基于建连流表对建连报文进行转发处理，以实现 RDMA 建连操作。

20 在又一些实例中，在虚拟交换机基于建连流表对建连报文进行转发处理之后，本实施例中的方法还可以包括：利用虚拟交换机将建连流表发送至硬件模块，以作为第一流表。

25 具体的，在虚拟交换机基于建连流表对建连报文进行转发处理之后，由于此时的硬件模块中并不包括与建连报文相对应的建连流表，为了能够提高对之后的建连报文进行转发操作的质量和效率，软件模块可以获取与建连报文相对应的建连流表，而后利用软件模块中的虚拟交换机将建连流表发送至硬件模块，以作为第一流表，从而有效地实现了第一流表的不断更新操作。

30 本实施例中，通过获取用于实现 RDMA 建连操作的建连静荷数据，而后利用软件模块中的虚拟交换机对建连静荷数据的报文头进行封装处理，获得建连报文，并利用软件模块中的软件管理器确定与建连报文相对应的建连流表，并将建连流表发送至虚拟交换机，而后利用虚拟交换机基于建连流表对建连报文进行转发处理，有效地实现 RDMA 建连操作，进一步提高了报文转发方法运行的稳定可靠性。

具体应用时，参考附图 9-图 10 所示，本应用实施例提供了一种基于 iwarp VPC 网络架构的 RDMA 建连过程和一种基于 iwarp VPC 网络架构的数据传输过程，上述的 iwarp (Internet Wide Area RDMA Protocol) 是一种用于远程直接内存访问、远程直接数据放置

的保护协议，它基于 TCP 协议实现。其中，为了能够实现 RDMA 建连过程，参考附图 9 所示，在物理机 host 或者虚拟机 VM 上设置一个用于实现 RDMA 传输操作的 QP 接口，通过调用上述的 QP 接口和 ib 驱动可以将 RDMA 建连信息下发至网卡节点，此时，RDMA 建连信息可以被网卡节点中的软件模块截获。具体的，基于 iwarp VPC 网络架构的 RDMA 建连过程可以包括如下步骤：

步骤 1.1：物理机/虚拟机可以生成 RDMA 建连报文，并将 RDMA 建连报文发送至网卡节点。

其中，物理机与网卡节点之间的 iwarp RDMA 连接可以通过正常的 TCP 三次握手建立连接，和正常的网络流量一样，RDMA 建连报文可以从 virtio-net 驱动（半虚拟化驱动）发出，由于此时网卡节点中的 FPGA 中没有对应的第一流表，所以可以将 RDMA 建连报文上传至网卡节点中的虚拟交换机 vswitch 进行处理。

步骤 1.2：虚拟交换机 vswitch 获取到 RDMA 建连报文之后，可以按照正常流量进行转发。

步骤 1.3：vswitch 先将 RDMA 建连报文包围封装 vxlan 头部发出。

步骤 2：网卡节点中的协议栈软件部分根据建连报文创建对应的流转发规则（即第二流表），并可以将流转发规则（即第二流表）传递给 vswitch，从而使得 vswitch 可以获取到第二流表。

步骤 2.1：vswitch 可以基于第二流表对建连报文进行转发操作，以实现 RDMA 建连操作。

步骤 2.2：vswitch 还能够将第二流表下发给网卡节点中的 FPGA，以生成第一流表。

参考附图 10 所示，基于 iwarp VPC 网络架构的数据传输过程可以包括如下步骤：

步骤 101：物理机或者虚拟机可以生成 RDMA 传输报文，此时，RDMA 传输报文中包括静荷数据。

其中，用于实现 RDMA 传输操作的 RDMA 传输报文流量从 ib 驱动设备发出，此时，RDMA 传输报文流量只有静荷数据。

步骤 102：网卡节点获取到 RDMA 传输报文，硬件模块可以将静荷数据的描述符传递给协议栈软件部分，协议栈软件部分通过协议栈处理和拥塞控制生成静荷数据所需要封装的 TCP 头和 ip 头部、mac 头部。

步骤 103：协议栈软件部分将生成静荷数据所需要封装的 TCP 头、ip 头部、mac 头部连同静荷数据的描述符一起传递回硬件 FPGA 的 RDMA 硬件部分；

步骤 104：RDMA 硬件部分根据软件传递的信息将静荷数据封装 TCP 头、ip 头和 mac 头；

步骤 105：RDMA 报文完成封装后，交由硬件模块进行处理，检测 RDMA 报文是否命中硬件模块中的第一流表，如果 RDMA 报文命中了硬件模块中的第一流表，则根据第

一流表中的流规则进行 RDMA 报文的处理操作，如封装 vxlan，将内层的拥塞控制值拷贝到外层，并在外层 vxlan 上设置对应的报文标识位来表示 RDMA 报文，然后转发出报文。

具体的，第一流表/第二流表的格式如下表 1 所示，第一流表包括两部分，一部分是查询流表的流键值（flow key），这部分由报文中提取出来；另一部分是流执行信息（flow action），这部分是根据报文提取的流键值查询命中后执行的动作信息，如 vxlan 的封装/解封装，从哪个端口发出等。

表 1

流表								
流 ID (flowid)	流键值 (flow key)						流执行信息 (flow action)	
	源端 ip/sip	目的 ip/sip	源 端 口 /sport	目的 端 口/dport	Vni/port_id	Flow dir	Action1	Action2
1	192.168.0.1	192.168.0.2	1000	2000	1024	in	Vxlan encap	Out port1
2	192.168.0.3	192.168.0.4	3000	4000	2048	out	Vxlan decap	Out port4
3	...	...	...	...	...	...	...	...

在上述的流表为第一流表时，可以检测 RDMA 报文是否命中了第一流表，即检测 RDMA 报文的五元组是否命中了第一流表，如果命中了第一流表，则会执行第一流表中所对应的流执行信息，例如：action1 或者 action2。

步骤 106：如果 RDMA 报文未命中硬件模块中的第一流表，则将 RDMA 报文上传至软件 vswitch 处理，vswitch 对 RDMA 报文进行相应处理，如封装 vxlan，将内层的拥塞控制值拷贝到外层，并在外层 vxlan 上设置对应的报文标识位来表示 RDMA 报文，然后转发出报文；并下发对应 RDMA 第二流表发送给硬件模块，以作为第一流表。

需要注意的是，接收端的处理报文与上述处理报文相类似，如果命中了硬件模块中第一流表中的 RDMA 流规则，则根据第一流表中的流规则进行 RDMA 报文的处理操作，如解封装 vxlan，并将外层的拥塞控制值拷贝到内层，设置对应的数据序列值；如果未命中，则将报文上传至软件 vswitch 进行处理，软件 vswitch 根据 vxlan 上设置的报文标识位确定是 RDMA 报文，并可以进行 RDMA 报文的流处理过程，如解封装 vxlan，并将外层的拥塞控制值拷贝到内层，设置对应的数据序列值，然后将报文发回硬件，并下发对应 RDMA 流给硬件，以作为第一流表。

此外，在硬件 RDMA 部分完成 RDMA 报文的解析之后，如解封掉 ip 头、mac 头和

TCP 头，并将描述符 mate 信息上送给 RDMA 软件部分，RDMA 软件部分更新 RDMA 的连接管理信息，然后发回给硬件送入目的端，以实现 RDMA 报文的数据处理操作。

本实施例提供的技术方案，能够基于 overlay 网络结构来实现 RDMA 数据的传输操作，此外，该技术方案能够屏蔽物理网络实现细节，并且没有物理网络范围限制，没有
5 同子网通信限制，因此不依赖中心管控节点，也没有中心节点性能瓶颈问题；另外，该技术方案采用网卡的软硬件结合方案，硬件部分负责报文的封装/解封装处理，软转部分专注状态管理，做到了报文处理性能和灵活性的平衡；此外，本实施例中的 RDMA 数据处理流程复用正常 VPC 网络流量的 vswitch 处理逻辑，例如：可以复用正常 VPC 网络的多种功能，如安全组等；并且，该技术方案采用基于 TCP 的 iwarp 方案实现，可以使用
10 TCP 的原生拥塞控制，成熟度更高，更能适应大规模网络，进一步提高了该技术方案的实用性。

图 11 为本发明应用实施例提供的一种基于远程直接数据存储的数据传输装置的结构示意图；参考附图 11 所示，本实施例提供了一种基于远程直接数据存储的数据传输装置，该数据传输装置可以执行上述图 3 所示的基于远程直接数据存储的数据传输方法。
15 具体的，该数据传输装置可以包括：相互通信连接的物理机 11、硬件模块 12 和软件模块 13；

物理机 11，用于生成与待转发 RDMA 报文相对应的静荷数据；

硬件模块 12，与所述物理机 11 通信连接，用于获取与待转发 RDMA 报文相对应的静荷数据，并对所述静荷数据的报文头进行封装处理，获得处理后报文；

20 硬件模块 12/软件模块 13，用于对处理后报文进行转发。

在一些实例中，在硬件模块 12/软件模块 13 对处理后报文进行转发时；

硬件模块 12，用于检测处理后报文是否命中第一流表，其中，第一流表存储在硬件模块 12 中；在处理后报文命中第一流表时，则基于第一流表对处理后报文进行转发；在处理后报文未命中第一流表时，将处理后报文发送至软件模块 13；

25 软件模块 13：用于接收处理后报文，并基于第二流表对处理后报文进行转发，其中，第二流表存储在软件模块 13 中。

在一些实例中，在软件模块 13 基于第二流表对处理后报文进行转发之后，本实施例中的软件模块 13 用于执行：获取与处理后报文相对应的目标流表；将目标流表发送至硬件模块 12，以作为第一流表。

30 在一些实例中，在硬件模块 12 对静荷数据的报文头进行封装处理，获得处理后报文时，该硬件模块 12 用于执行：获取静荷数据的数据描述符，数据描述符用于标识静荷数据的长度和 RDMA 头信息；利用数据描述符，获取与静荷数据相对应的封装信息；利用封装信息对静荷数据的报文头进行封装处理，获得处理后报文。

在一些实例中，在硬件模块 12 利用数据描述符，获取与静荷数据相对应的封装信息

时，该硬件模块 12 用于执行：将数据描述符发送至软件模块 13；

软件模块 13，用于利用数据描述符，生成与静荷数据相对应的封装信息，并将封装信息和数据描述符发送至硬件模块 12。

5 在一些实例中，在软件模块 13 利用数据描述符，生成与静荷数据相对应的封装信息时，该软件模块 13 用于执行：对数据描述符进行协议栈处理和拥塞控制处理，生成与静荷数据相对应的封装信息，其中，封装信息中包括：用于标识拥塞控制的拥塞控制值、传输控制协议 TCP 头信息、因特网互联协议 IP 头信息和介质访问控制 MAC 头信息。

10 在一些实例中，在硬件模块 12 或者软件模块 13 对处理后报文进行转发时，硬件模块 12 或者软件模块 13 用于执行：对处理后报文进行 VxLAN 封装，获得目标报文；获取与目标报文相对应的拥塞控制值；将拥塞控制值拷贝到外层，并在外层生成与目标报文相对应的报文格式标识，报文格式标识用于标识目标报文为 RDMA 报文；基于拥塞控制值、报文格式标识和流表转发规则，对目标报文进行转发。

在一些实例中，本实施例中的物理机 11、硬件模块 12 和软件模块 13 还用于执行以下步骤：

15 物理机 11，用于生成待转发报文，并将待转发报文发送至网卡节点；

硬件模块 12，用于获取待转发报文，并对待转发报文进行报文头的解封装操作，获得与待转发报文相对应的解封装报文；

硬件模块 12/软件模块 13，用于对解封装报文进行转发。

在一些实例中，在硬件模块 12/软件模块 13 对解封装报文进行转发时：

20 硬件模块 12，用于检测解封装报文是否命中第一流表，在解封装报文命中第一流表时，则硬件模块 12 用于基于第一流表对解封装报文进行转发；在解封装报文未命中第一流表时，则硬件模块 12 将解封装报文发送至软件模块 13；

软件模块 13，用于获取解封装报文，并基于第二流表对解封装报文进行转发。

25 在一些实例中，在硬件模块 12 基于第一流表对解封装报文进行转发时，该硬件模块 12 用于执行：利用第一流表对解封装报文进行 VxLAN 解封装，获得目标报文；获取与目标报文相对应的拥塞控制值，将拥塞控制值拷贝到内层；基于拥塞控制值和第一流表，对目标报文进行转发。

30 在一些实例中，在软件模块 13 利用第二流表对解封装报文进行转发时，该软件模块 13 用于执行：利用第二流表获取与解封装报文相对应的报文格式标识；在基于报文格式标识确定解封装报文为 RDMA 报文时，基于第二流表对解封装报文进行 VxLAN 解封装，获得目标报文；获取与目标报文相对应的拥塞控制值，并将拥塞控制值拷贝到内层；基于拥塞控制值和第二流表，对目标报文进行转发。

在一些实例中，在硬件模块 12 对待转发报文进行报文头的解封装操作之后，本实施例中的硬件模块 12 和软件模块 13 还可以用于执行以下步骤：

硬件模块 12，用于获得与待转发报文相对应的报文描述符；将解封装报文和报文描述符发送至软件模块 13；

软件模块 13，用于利用解封装报文和报文描述符，对用于实现 RDMA 报文传输操作的连接管理信息进行更新。

- 5 在一些实例中，在获得目标报文之后，本实施例中的软件模块 13 用于执行：生成并存储与目标报文相对应的数据序列值，数据序列值与用于对报文数据进行封装处理的封装信息相关。

在一些实例中，在获取与待转发 RDMA 报文相对应的静荷数据之前，本实施例中的硬件模块 12 和软件模块 13 用于执行以下步骤：

- 10 硬件模块 12，用于获取用于实现 RDMA 建连操作的建连静荷数据；

软件模块 13，用于利用软件模块 13 中的虚拟交换机对建连静荷数据的报文头进行封装处理，获得建连报文；并利用软件模块 13 中的软件管理器确定与建连报文相对应的建连流表，并将建连流表发送至虚拟交换机；虚拟交换机基于建连流表对建连报文进行转发处理，以实现 RDMA 建连操作。

- 15 在一些实例中，在虚拟交换机基于建连流表对建连报文进行转发处理之后，本实施例中的软件模块 13 用于：利用虚拟交换机将建连流表发送至硬件模块 12，以作为第一流表。

图 11 所示装置可以执行图 2-图 10 所示实施例的方法，本实施例未详细描述的部分，可参考对图 2-图 10 所示实施例的相关说明。该技术方案在执行过程和技术效果参见图 2-图 10 所示实施例中的描述，在此不再赘述。

- 20 在一个可能的设计中，图 11 所示基于远程直接数据存储的数据传输装置的结构可现为一电子设备，该电子设备可以是网卡结构、服务器等各种设备。如图 12 所示，该电子设备可以包括：第一处理器 21 和第一存储器 22。其中，第一存储器 22 用于存储相对应电子设备执行上述图 2-图 10 所示实施例中提供的基于远程直接数据存储的数据传输方法的程序，第一处理器 21 被配置为用于执行第一存储器 22 中存储的程序。

程序包括一条或多条计算机指令，其中，一条或多条计算机指令被第一处理器 21 执行时能够实现如下步骤：

获取与待转发 RDMA 报文相对应的静荷数据；

- 30 利用网卡节点中的硬件模块对所述静荷数据的报文头进行封装处理，获得处理后报文；

利用网卡节点中的硬件模块或软件模块对处理后报文进行转发。

进一步的，第一处理器 21 还用于执行前述图 2-图 10 所示实施例中的全部或部分步骤。

其中，电子设备的结构中还可以包括第一通信接口 23，用于电子设备与其他设备或

通信网络通信。

图 13 为本发明实施例提供的一种网卡的结构示意图，参考附图 13 所示，本实施例提供了一种网卡，该网卡可以执行上述图 2 所示的基于远程直接数据存储的数据传输方法。具体的，该网卡可以包括：

- 5 硬件模块 31，用于获取与待转发 RDMA 报文相对应的静荷数据，并对所述静荷数据的报文头进行封装处理，获得处理后报文；

 硬件模块 31/软件模块 32，用于对处理后报文进行转发。

- 其中，软件模块 32 中可以包括负责正常的 VPC 网络流量转发操作的虚拟交换机 vswitch；具体的，软件模块 32 可以获取报文的五元组特征，通过五元组特征查询路由表、
10 ACL 表等各种转发表，获得查询结果，而后基于查询结果和五元组特征生成第二流表。硬件模块 31 可以采用现场可编程逻辑门阵列（Field Programmable Gate Array，简称 FPGA）或者专用集成电路(Application Specific Integrated Circuit，简称 ASIC)，该硬件模块 31 可以卸载由软件 vswitch 下发的第二流表。

- 图 13 所示网卡可以执行图 2-图 10 所示实施例的方法，本实施例未详细描述的部分，
15 可参考对图 2-图 10 所示实施例的相关说明。该技术方案在执行过程和技术效果参见图 2-图 10 所示实施例中的描述，在此不再赘述。

 另外，本发明实施例提供了一种计算机存储介质，用于储存电子设备所用的计算机软件指令，其包含用于执行上述图 2-图 10 所示方法实施例中基于远程直接数据存储的数据传输方法所涉及的程序。

- 20 此外，本发明实施例提供了一种计算机程序产品，包括：存储有计算机指令的计算机可读存储介质，当所述计算机指令被一个或多个处理器执行时，致使所述一个或多个处理器执行上述图 2-图 10 所示方法实施例中基于远程直接数据存储的数据传输方法中的步骤。

- 以上所描述的装置实施例仅仅是示意性的，其中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
25 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模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下，即可以理解并实施。

- 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到各实施方式可借助加必需的通用硬件平台的方式来实现，当然也可以通过硬件和软件结合的方式来实现。
30 基于这样的理解，上述技术方案本质上或者说对现有技术做出贡献的部分可以以计算机产品的形式体现出来，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

本发明是参照根据本发明实施例的方法、设备（系统）、和计算机程序产品的流程图和 / 或方框图来描述的。应理解可由计算机程序指令实现流程图和 / 或方框图中的每一流程和 / 或方框、以及流程图和 / 或方框图中的流程和 / 或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程设备的处理器以产生一个机器，使得通过计算机或其他可编程设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制品，该指令装置实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的步骤。

在一个典型的配置中，计算设备包括一个或多个处理器（CPU）、输入/输出接口、网络接口和内存。

内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器（RAM）和/或非易失性内存等形式，如只读存储器（ROM）或闪存（flash RAM）。内存是计算机可读介质的示例。

计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存（PRAM）、静态随机存取存储器（SRAM）、动态随机存取存储器（DRAM）、其他类型的随机存取存储器（RAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、快闪记忆体或其他内存技术、只读光盘只读存储器（CD-ROM）、数字多功能光盘（DVD）或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括暂存电脑可读媒体（transitory media），如调制的数据信号和载波。

最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

权 利 要 求 书

1. 一种基于远程直接数据存储的报文转发方法，其特征在于，包括：
获取与待转发 RDMA 报文相对应的静荷数据；
利用网卡节点中的硬件模块对所述静荷数据的报文头进行封装处理，获得处理后报
5 文；
利用网卡节点中的硬件模块或软件模块对处理后报文进行转发。
2. 根据权利要求 1 所述的方法，其特征在于，利用网卡节点中的硬件模块或软件模
块对处理后报文进行转发，包括：
检测所述处理后报文是否命中第一流表，其中，所述第一流表存储在所述硬件模块
10 中；
在所述处理后报文中命中第一流表时，则利用所述硬件模块和第一流表对所述处理后
报文进行转发；
在所述处理后报文未命中所述第一流表时，则利用所述软件模块和第二流表对所述
处理后报文进行转发，其中，所述第二流表存储在所述软件模块中。
- 15 3. 根据权利要求 2 所述的方法，其特征在于，在利用所述软件模块和第二流表对所
述处理后报文进行转发之后，所述方法还包括：
获取与所述处理后报文相对应的目标流表；
将所述目标流表发送至所述硬件模块，以作为第一流表。
4. 根据权利要求 1 所述的方法，其特征在于，利用网卡节点中的硬件模块对所述静
20 荷数据的报文头进行封装处理，获得处理后报文，包括：
利用所述硬件模块获取所述静荷数据的数据描述符，所述数据描述符用于标识所述
静荷数据的长度和 RDMA 头信息；
利用所述数据描述符，获取与所述静荷数据相对应的封装信息；
利用所述硬件模块和所述封装信息对所述静荷数据的报文头进行封装处理，获得处
25 理后报文。
5. 根据权利要求 4 所述的方法，其特征在于，利用所述数据描述符，获取与所述静
荷数据相对应的封装信息，包括：
利用所述硬件模块将所述数据描述符发送至软件模块；
利用所述软件模块和所述数据描述符，生成与所述静荷数据相对应的封装信息，并
30 将所述封装信息和数据描述符发送至硬件模块。
6. 根据权利要求 5 所述的方法，其特征在于，利用所述软件模块和所述数据描述符，
生成与所述静荷数据相对应的封装信息，包括：
利用所述软件模块对所述数据描述符进行协议栈处理和拥塞控制处理，生成与所述
静荷数据相对应的封装信息，其中，所述封装信息中包括：用于标识拥塞控制的拥塞控
35 制值、传输控制协议 TCP 头信息、因特网互联协议 IP 头信息和介质访问控制 MAC 头

信息。

7. 根据权利要求 2 所述的方法，其特征在于，对所述处理后报文进行转发，包括：
对所述处理后报文进行 VxLAN 封装，获得目标报文；
获取与所述目标报文相对应的拥塞控制值；
- 5 将所述拥塞控制值拷贝到外层，并在外层生成与所述目标报文相对应的报文格式标识，所述报文格式标识用于标识目标报文为 RDMA 报文；
基于所述拥塞控制值、所述报文格式标识和流表转发规则，对所述目标报文进行转发。
8. 根据权利要求 1-7 中任意一项所述的方法，其特征在于，所述方法还包括：
- 10 获取待转发报文；
利用网卡节点中的硬件模块对所述待转发报文进行报文头的解封装操作，获得与所述待转发报文相对应的解封装报文；
利用所述硬件模块或软件模块对解封装报文进行转发。
9. 根据权利要求 8 所述的方法，其特征在于，利用所述硬件模块或软件模块对解封装报文进行转发，包括：
- 15 检测所述解封装报文是否命中第一流表；
在所述解封装报文命中第一流表时，则利用所述硬件模块和所述第一流表对所述解封装报文进行转发；
在所述解封装报文未命中所述第一流表时，则利用所述软件模块和第二流表对所述解封装报文进行转发。
- 20 10. 根据权利要求 9 所述的方法，其特征在于，利用所述硬件模块和所述第一流表对所述解封装报文进行转发，包括：
利用所述硬件模块和所述第一流表对所述解封装报文进行 VxLAN 解封装，获得目标报文；
- 25 获取与所述目标报文相对应的拥塞控制值，将所述拥塞控制值拷贝到内层；
基于所述拥塞控制值和所述第一流表，对所述目标报文进行转发。
11. 根据权利要求 9 所述的方法，其特征在于，利用所述软件模块和第二流表对所述解封装报文进行转发，包括：
利用所述软件模块和第二流表获取与所述解封装报文相对应的报文格式标识；
- 30 在基于所述报文格式标识确定所述解封装报文为 RDMA 报文时，基于所述第二流表对所述解封装报文进行 VxLAN 解封装，获得目标报文；
获取与所述目标报文相对应的拥塞控制值，并将所述拥塞控制值拷贝到内层；
基于所述拥塞控制值和第二流表，对所述目标报文进行转发。
12. 根据权利要求 9 所述的方法，其特征在于，在利用网卡节点中的硬件模块对所述待转发报文进行报文头的解封装操作之后，所述方法还包括：
- 35

获得与所述待转发报文相对应的报文描述符；

将所述解封装报文和报文描述符发送至软件模块；

利用所述软件模块、所述解封装报文和报文描述符，对用于实现 RDMA 报文传输操作的连接管理信息进行更新。

- 5 13. 根据权利要求 10 所述的方法，其特征在于，在获得目标报文之后，所述方法还包括：

生成并存储与所述目标报文相对应的数据序列值，所述数据序列值与用于对报文数据进行封装处理的封装信息相关。

- 10 14. 根据权利要求 1-7 中任意一项所述的方法，其特征在于，在获取与待转发 RDMA 报文相对应的静荷数据之前，所述方法还包括：

获取用于实现 RDMA 建连操作的建连静荷数据；

利用所述软件模块中的虚拟交换机对所述建连静荷数据的报文头进行封装处理，获得建连报文；

- 15 利用所述软件模块中的软件管理器确定与所述建连报文相对应的建连流表，并将所述建连流表发送至所述虚拟交换机；

所述虚拟交换机基于所述建连流表对所述建连报文进行转发处理，以实现 RDMA 建连操作。

15. 一种基于远程直接数据存储的报文转发装置，其特征在于，包括：

物理机，用于生成与待转发 RDMA 报文相对应的静荷数据；

- 20 硬件模块，与所述物理机通信连接，用于获取与待转发 RDMA 报文相对应的静荷数据，并对所述静荷数据的报文头进行封装处理，获得处理后报文；

硬件模块/软件模块，用于对处理后报文进行转发。

- 25 16. 一种电子设备，其特征在于，包括：存储器、处理器；其中，所述存储器用于存储一条或多条计算机指令，其中，所述一条或多条计算机指令被所述处理器执行时实现如权利要求 1-14 中任意一项所述的基于远程直接数据存储的报文转发方法。

17. 一种网卡，其特征在于，包括：

硬件模块，用于获取与待转发 RDMA 报文相对应的静荷数据，并对所述静荷数据的报文头进行封装处理，获得处理后报文；

硬件模块/软件模块，用于对处理后报文进行转发。

- 30 18. 一种计算机存储介质，其特征在于，用于储存计算机程序，所述计算机程序使计算机执行时实现如权利要求 1-14 中任意一项所述的基于远程直接数据存储的报文转发方法。