

报文传输的方法和装置

本申请要求于 2023 年 4 月 28 日提交中国专利局、申请号为 202310483155.4、申请名称为“报文传输的方法和装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请实施例涉及计算机技术领域，特别涉及一种报文传输的方法和装置。

背景技术

通信设备跨网络通信的应用场景中对网络的诉求体现在高吞吐和低时延两个关键指标上，为实现高吞吐和低时延，业界一般都采用远程直接内存访问（Remote Direct Memory Access, RDMA）技术替代传统的传输控制协议（Transmission Control Protocol, TCP）技术，实现时延的下降和降低对数据中心中央处理机（Central Processing Unit, CPU）的占用率。

目前无限带宽技术 RDMA 协议（InfiniBand, IB）和基于以太网物理层传输的 RDMA 技术（RDMA over Converged Ethernet, RoCE）作为业界主流的 RDMA 技术，在高性能数据中心互联领域得到了广泛的应用。IB 和 RoCE 是为 RDMA 设计的网络协议，IB 和 RoCE 支持可靠连接（Reliable Connect, RC）模式，在 RC 模式下，消息按顺序到达并可靠交付。但是 RC 模式下组网规模越大，所需连 RC 连接越多，资源消耗越大。因此如何降低 RC 连接数量成为亟待解决的问题。

发明内容

本申请实施例提供一种报文传输的方法，应用于第一设备，该第一设备中不同的应用（如，第一应用和第二应用）的报文可以共享连接进行传输，以期实现降低连接数量的目的。

第一方面，提供了一种报文传输的方法，应用于第一设备。该方法可以由第一设备执行，或者，也可以由配置于第一设备中电路执行，本申请对此不作限定。为了方便，以下以第一设备执行为例进行说明。该第一设备包括事务层和传输层，该报文传输的方法，包括：

所述第一设备的传输层接收来自所述第一设备的事务层的报文，所述报文为待发送给第二设备的第一应用的报文；当所述第一设备的传输层和第二设备的传输层之间已建立至少一个第一传输层 TP 连接时，所述第一设备的传输层向所述第一设备的事务层发送第一指示信息，所述第一指示信息用于指示所述至少一个第一 TP 连接用于传输所述第一应用的报文；所述第一设备的传输层通过所述至少一个第一 TP 连接向所述第二设备的传输层发送所述第一应用的报文，其中，所述至少一个第一 TP 连接是所述第一设备的传输层用于传输第二应用的报文的 TP 连接。

基于上述技术方案，当第一设备的传输层接收到第一应用的待发送给第二设备的报文之后，确定第一设备的传输层和第二设备的传输层之间是否已经有建立的传输层连接，若已经有建立的传输层连接，第一应用的报文可以通过已经建立的传输层连接发送至第二设备的传输层，无需针对该第一应用的报文重新建立连接。该已经建立的传输层连接是第一设备的传输层针对第二应用的报文建立的连接，该报文传输的方法中，不同的应用可以共享传输层的连接，无需针对每个应用建立不同的连接，从而可以降低连接数量，节约建立连接所需的资源。

结合第一方面，在第一方面的某些实现方式中，当所述第一设备的传输层和第二设备的传输层之间未建立 TP 连接时，所述方法还包括：所述第一设备的传输层获取所述第二设备的互联网协议 IP 地址信息；所述第一设备的传输层根据所述第二设备的 IP 地址信息，建立所述第一设备的传输层和第二设备的传输层之间至少一个第二 TP 连接，所述至少一个第二 TP 连接用于传输所述第一应用的报文。

基于上述技术方案，当第一设备的传输层接收到第一应用的待发送给第二设备的报文之后，确定第一设备的传输层和第二设备的传输层之间还未建立传输层连接时，第一设备的传输层可以获取第二设备的 IP 地址信息，并且基于获取的 IP 地址信息和自身的及时建立传输报文所需的传输层连接，保证报文传输的及时性。需要说明的是，当第一设备的传输层和第二设备的传输层之间建立了至少一个第二 TP 连接

之后，若第一设备的传输层接收到待发送给第二设备的第三应用的报文，可以通过该至少一个第二 TP 连接传输第三应用的报文，无需针对第三应用建立不同的连接，从而可以降低连接数量，节约建立连接所需的资源。

结合第一方面，在第一方面的某些实现方式中，所述第一设备的传输层通过所述至少一个第一 TP 连接向所述第二设备的传输层发送所述第一应用的报文，包括：所述第一设备的传输层通过所述至少一个第一 TP 连接中的负载最轻的 TP 连接向所述第二设备的传输层发送所述第一应用的报文。

基于上述技术方案，第一设备的传输层可以选择负载最轻的 TP 连接传输报文，实现负载均衡。

结合第一方面，在第一方面的某些实现方式中，所述方法还包括：所述第一设备的传输层接收来自所述第二设备的传输层的应答消息，所述应答消息用于指示所述第一应用的报文是否成功被所述第二设备的事务层执行。

基于上述技术方案，第二设备可以通过第二设备的传输层向第一设备的传输层发送指示报文是否成功被所述第二设备的事务层执行的应答消息，可以及时上报多个应用共享传输层连接传输报文时带来的阻塞问题。

结合第一方面，在第一方面的某些实现方式中，若所述应答消息为事物应答 TAACK，指示所述第二设备的事务层成功执行所述第一应用的报文，所述方法还包括：所述第一设备的传输层向所述第一设备的事务层发送完成队列条目 CQE，所述 CQE 用于指示所述第一应用的报文成功传输。

结合第一方面，在第一方面的某些实现方式中，若所述第一响应消息为事物否定应答 TANAK，指示所述第二设备的事务层未成功执行所述第一应用的报文，所述方法还包括：所述第一设备的传输层向所述第一设备的事务层发送所述 TANAK；所述第一设备的传输层接收来自所述第一设备的事务层的所述第一应用的报文，并重新向所述第二设备的传输层发送所述第一应用的报文。

基于上述技术方案，若第二设备的事务层未成功接收第一应用的报文，第一设备的传输层可以重传该第一应用的报文，实现可靠传输。

结合第一方面，在第一方面的某些实现方式中，所述第一设备的传输层接收来自所述第一设备的事务层的报文，包括：所述第一设备的传输层接收来自所述第一设备的事务层的发送队列条目 SQE，所述 SQE 的目的地址为所述第二设备的地址，所述 SQE 承载发送给所述第二设备所述第一应用的报文。

结合第一方面，在第一方面的某些实现方式中，所述第一设备的传输层接收来自所述第一设备的事务层的 SQE，包括：所述第一设备的传输层接收来自所述第一设备的至少一个发送接口的所述 SQE。

结合第一方面，在第一方面的某些实现方式中，在所述第一设备的传输层通过所述至少一个第一 TP 连接向所述第二设备的传输层发送所述第一应用的报文之前，所述方法还包括：所述第一设备的传输层根据所述 SQE 中携带的目的实体标识 DEID 确定所述至少一个第一 TP 连接。

第二方面，提供了一种报文传输的装置，该装置包括：存储模块，用于存储程序；处理模块，用于执行存储模块存储的程序，当存储模块存储的程序被执行时，处理模块用于执行上述第一方面提供的方法。

第三方面，提供一种计算机可读存储介质，该计算机可读介质存储用于设备执行的程序代码，该程序代码包括用于执行上述第一方面提供的方法。

第四方面，提供一种包含指令的计算机程序产品，当该计算机程序产品在计算机上运行时，使得计算机用于执行上述第一方面提供的方法。

第五方面，提供一种芯片，该芯片包括处理模块与通信接口，该处理模块通过该通信接口读取存储器上存储的指令，用于执行上述第一方面提供的方法。

可选地，作为一种实现方式，该芯片还可以包括存储模块，该存储模块中存储有指令，该处理模块用于执行该存储模块上存储的指令，当该指令被执行时，该处理模块用于执行上述第一方面提供的方法。

第六方面，提供一种芯片，该芯片包括用于执行第一方面提供的方法的第一设备和用于接收报文的第二设备。

第七方面，提供一种计算机设备，该计算机设备包括第六方面所示的芯片。例如，计算机设备包括但不限于数据中心中的交换机或服务器。

第八方面，提供一种终端设备，该终端设备包括第六方面所示的芯片。例如，终端设备包括但不限于手机、车辆等终端。

第九方面，提供一种报文传输的系统，该系统包括用于执行第一方面提供的方法的第一设备和用于

接收报文的第二设备。

附图说明

- 图 1 为本申请实施例提供的计算机设备的结构示意图。
图 2 为本申请实施例提供的数据中心示意图。
图 3 是本申请实施例提供的一种通用总线协议报文格式示意图。
图 4 是本申请实施例提供的一种主机的结构示意图。
图 5 是本申请提供的一种报文传输的方法的示意性流程图。
图 6 是本申请提供的两个设备的传输层之间建立 TPG 的示意图。
图 7 是本申请提供的多个 SQE 写入一个 JFS 的示意图。
图 8 是本申请提供的负载均衡的示意图。
图 9 是本申请提供的另一种报文传输的方法的示意性流程图。
图 10 示出了本申请实施例提供的一种报文传输的装置 1000 的结构示意图。
图 11 示出了本申请实施例提供的一种芯片系统 1100 的结构示意图。
图 12 示意性地示出本申请实施例提供的计算机程序产品的概念性局部视图。

具体实施方式

下面将结合附图，对本申请实施例中的技术方案进行描述。

在很多应用中，部署应用的计算机设备需要访问数据，以实现应用的功能。例如部署数据库应用的计算机设备需要进行大量的数据访问，以更新数据库中的数据，或者是响应数据查询请求，向用户返回查询结果。又例如，部署 web 应用的计算机设备需要进行大量的数据访问，以向用户返回请求的内容。

计算机设备可以是服务器，或者是终端。终端包括但不限于台式机、笔记本电脑、智能手机等用户设备。为了便于理解，下面对计算机设备的结构进行介绍。

参见图 1 所示的计算机设备的结构示意图，计算机设备包括处理器 101、输入输出设备(input output device, IO device)102、内存 103、缓存 104、内存管理单元(memory management unit, MMU)105、输入输出内存管理单元(input output management unit, IOMMU)106、外存 107 和总线 108。

处理器 101 包括至少一个内核(core)。该内核也称作计算引擎。其中，每个内核可以独立地执行任务。当处理器 101 包括多个内核时，可以对来自应用的任务进行划分，使得应用能够充分利用多个内核，在特定的时间内执行更多任务。在本实施例中，处理器 101 可以是主处理器，例如为中央处理机 (Central Processing Unit, CPU)。

输入输出设备 102 是指具有输入数据和/或输出数据能力的硬件设备。输入输出设备 102 可以分为输入设备和输出设备。其中，输入设备可以包括鼠标、键盘、操作杆、触控笔、麦克风等设备，输出设备可以包括显示器、扬声器等设备。

内存 103 也称作内存存储器或主存储器，用于暂时存放处理器 101 中的运算数据。进一步地，内存 103 还用于暂时存放与外存 107 交换的数据。内存 103 通常可以采用动态随机存取存储器 DRAM 或者静态随机存取存储器(static random access memory, SRAM)等存储介质实现。

缓存 104(本实施例中是指处理器缓存，如 CPU 缓存)是用于减少处理器 101 访问内存 103 所需平均时间的部件。参见图 1，在金字塔式存储体系中，缓存 104 位于自顶向下的第二层，仅次于处理器 101 的寄存器(图 1 中未示出)，高于内存 103(内存 103 位于自顶向下的第三层)。通常情况下，缓存 104 的容量远小于内存 103，但访问速度可以接近处理器 101 的频率。

内存管理单元 105 是一种用于处理数据访问请求的计算机硬件。内存管理单元 105 具体用于对数据访问请求中的虚拟地址(virtual address, VA)进行映射。其中，内存管理单元 105 可以截获处理器 101 的内核发出的数据访问请求，将数据访问请求中的虚拟地址映射(或者翻译)为物理地址(physical address, PA)，以便于根据该物理地址访问内存 103。

输入输出内存管理单元 106 实质是一种内存管理单元。类似于内存管理单元 105 将处理器 101 可见的虚拟地址映射为物理地址，输入输出内存管理单元 106 用于将输入输出设备可见 102 的虚拟地址(也可以称作设备地址或 IO 地址)映射为物理地址。

外存 107 也称作外部存储器、辅存，通常用于持久化保存数据。例如，外存 107 可以持久化存储处

理器 101 中的运算数据。即使供电异常，已经写入该外存 107 的数据仍然能够保存，避免了数据丢失。具体实现时，外存 107 包括至少一个非易失性存储器 1071，当外存包括多个非易失性存储器时，这多个非易失性存储器可以是相同类型，也可以是不同类型。例如，在图 1 的示例中，外存 107 可以包括两种类型的非易失性存储器，例如为存储级存储器(storage class memory, SCM)和固态硬盘(solid state drive, SSD)。

总线 108 用于将计算机设备的各个功能部件连接。总线 108 是计算机设备各种功能部件之间传递信息的公共通信干线。总线 108 可以由导线形成的传输线束。根据连接对象不同，总线 108 还可以分为内部总线和外部总线。

其中，内部总线采用内部总线协议传递信息。内部总线协议包括用于访问所述计算机设备的内存空间的总线协议。外部总线采用外部总线协议传递信息。外部总线协议包括用于访问所述计算机设备的外存空间的总线协议。其中，内存空间是指内存的地址空间，外存空间是指外存的地址空间。

在一些实施例中，内部总线协议包括但不限于外设部件互连标准(peripheral component interconnect, PCI)总线、外设部件互连标准高速(PCI Express, PCI-E)协议、快速通道互联(Intel™ Quick Path Interconnect, QPI)协议、通用总线(Unified Bus, UB)协议。外部总线协议包括但不限于小型计算机系统专用接口(small computer system interface, SCSI)协议或者串行连接小型计算机系统专用接口(Serial Attached SCSI, SAS)协议。

需要说明的是，图 1 所示的计算机设备是以外存 107 为远端外存进行示例说明。如图 1 所示，外存 107 包括网卡 1072。该网卡 1072 例如可以是 smart NIC 网络接口卡(也即网络适配卡)。外存 107 通过该网卡 1072 接入网络，进而通过网络与计算机设备 101 的其他部件连接。网络可以是有线通信网络，如光纤通信网络，也可以是无线通信网络，例如是无线局域网(wireless local area network, WLAN)或者是第五代(the fifth generation, 5G)移动通信网络。

在一些可能的实现方式中，计算机设备的外存 107 也可以是本地外存，计算机设备的其他部件如处理器 101 可以通过总线 108 连接上述本地外存。在另一些可能的实现方式中，计算机设备可以既包括远端外存，又包括本地外存。此外，本申请实施例可以适用于集中式存储，或者分布式存储场景，本实施例对此不作限定。

本申请主要涉及跨网络的数据传输，示例性地，应用于需要跨网络通信的服务器集群，如图 2 所示的数据中心。其中，图 2 中所示的交换机或服务器内部结构如上文图 1 中所示。

另外，本申请中涉及的传输报文的设备是支持通用总线(Unified Bus, UB)协议的，设备之间可以建立传输层的连接。其中，通用总线协议还可以称为灵衢总线或统一总线，一种总线协议标准，本申请对于该通用总线的名称不做限定。

通用总线协议打破现有各种协议壁垒，去除中间不必要的转换开销，从而实现极致低时延。通用总线协议定义了独立的事务层和传输层。传输层之间有连接，事务层之间无连接。一个 Host 内的所有事务都承载在一个传输层上。通用总线协议包含传输层和事务层，传输层负责网络丢包重传，保证可靠传输，事务层处理各自不同的事务。传输层从网络收到包，剥离传输层包头，转发给事务层。

通用总线协议报文格式如图 3 所示。具体地，通用总线协议报文格式中的字段定义如下表 1 所示：

表 1

名称(name)	描述(description)
UB 链接层(UBLINK)	UB 协议定义的链接(LINK)层，UBLINK 也可替换为以太 MAC，UB 协议定义此形态为 UB over Ethernet。
网络分区 ID (Network Partition ID, NPI)	用于物理网络隔离。
IP	IP 协议头
用户数据报协议(User Datagram Protocol, UDP)	UDP 协议头，UDP 目的端口等于 4792 表示 UB 报文，后面是传输层头(Trans Port Header, TPH)。
TPH	传输层头，包含 TPopcode、源 TPN、目的 TPN、包序列号(Packet Sequence Number)等内容。
UB 分区 ID (UB partition ID, UPI)	用于租户隔离。

UB 实体 ID (UB Entity ID , UEID)	包括源实体标识 (Source Entity ID) 和目的实体标识 (Destination Entity ID)。 Entity ID 位宽 128bit, 全网唯一。一个实体标识 (EID) 可以表示一个虚拟机, 也可以表示一个 SSD 控制器。
事务层头 (Transaction Header, TAH)	事务层头。包含 TAOpcode (表示事务操作类型, 如 send、read、write、atomic 或事务层应答)、目的 JFR 号、事务层顺序段号 (Segment Sequence Number, TASSN)、读写地址和长度等。

具体地, 通用总线协议的事务层和应用的交互接口称为 Jetty, 应用的消息可以通过一个 Jetty 发到任何目的地, 也可以通过一个 Jetty 接收来自任何源的消息。只能发送的 Jetty, 定义为 (Jetty For Send, JFS); 只能接收的 Jetty, 定义为 (Jetty For Receive, JFR)。

图 4 是本申请实施例提供的一种主机的结构示意图。该主机 (如图 4 中所示的主机 A 和主机 B) 可以应用于图 2 所示的跨网络通信的应用场景中。如图 4 所示, 主机 A 包含若干个虚拟机 (Virtual Machine, VM), 一个 VM 包含若干个进程和若干个远端命令 (Remote Command, RC) 表, 用于接收远端 Read 命令。一个进程包含若干个通信接口 (如, 图 4 中所示的 Jetty)、只能发送的通信接口 (如, 图 4 中所示的 JFS)、只能发送的通信接口 (如, 图 4 中所示的 JFR)。其中, Jetty 是双向的, 既可以收也可以发; JFS 是单向的, 只能发送; JFR 是单向的, 只能接收。

Jetty、JFS、JFR、RC 有各自的上下文 (context, CXT)。Jetty、JFS、JFR、RC 属于事务层。两个主机间建立若干个传输层 (Transport, TP) 连接 (如, 图 4 中所示的 TP 连接#0 至 TP 连接#7, 共 8 个 TP 连接), 这 8 个 TP 连接可以组成一个传输层组 (Transport Group, TPG), 两个主机之间所有的流量都通过这个 TPG。这 8 个 TP 连接可以分布在不同物理端口, 流量在 8 个 TP 连接之间均衡发送, 实现多端口多路径。TP 连接、TPG 属于传输层。两个主机之间经过网络, 网络可能丢包, TP 连接负责网络丢包重传, 保证端到端的可靠性。TP 连接负责端到端拥塞控制。

图 4 中主机 B 包含若干个 VM 和若干个进程 (process)。VM 和 SSD 控制器属于事务层。

上文中结合图 1 至图 4 简单介绍本申请能够应用的场景以及涉及的主机内部逻辑单元, 为了便于理解本申请实施例, 对本申请涉及的一些基本概念做简要说明。

1、包序列号 (packet sequence number, PSN): 发送侧传输层发包时, 给每个包打上一个 PSN, PSN 逐包递增。接收侧收到包, 返回 TPACK (携带已收包的 PSN), 告知发送侧传输层包已正确接收。如果接收侧收到包, 发现比此包 PSN 小的包未收到, 则判定‘比此包 PSN 小的包’在网络丢失, 返回 TPSACK (携带已收包的 PSN, 丢失包的 PSN), 发送侧传输层收到 TPSACK, 重传丢失的包。

2、切片序列号 (Segment sequence number, SSN): 事务层的消息可能比较大, 例如 16MB。UB 协议中, 多个事务层共用一个传输层, 为了防止一个事务层的消息长时间占用传输层连接, 事务层把消息发给传输层时, 把消息切成多个切片, 例如一个切片 64KB, 一个事务层每次只发一个切片给传输层。

3、事物应答 (Transaction ACK, TAACK): 接收侧收齐一个 Segment (一个 Segment 在传输层被拆成多个包) 且正确执行 (例如正确读写内存) 后, 返回 TAACK, 告知发送侧事务层, 该切片已正确执行, 或者接收侧成功接收发送侧的消息, 返回 TAACK, 告知发送侧事务层, 消息已成功接收。

4、事物否定应答 (Transaction No OK ACK, TANAK): 接收侧收到一个 Segment, 执行出错 (例如读写内存出现 pagefault), 则返回 TANAK, 告知发送侧事务层, 重传该切片, 或者接收侧未成功接收发送侧的消息, 返回 TANAK, 告知发送侧事务层, 重传该消息。

5、远程直接数据存取 (Remote Direct Memory Access, RDMA): 通过网络把数据直接传入计算机的存储区, 将数据从一个系统快速移动到远程系统存储器中, 而无需两台计算机设备的操作系统或内核介入。RDMA 消除了外部存储器复制和上下文切换的开销, 因此能解放内存带宽和 CPU 周期用于改进应用系统性能。

6、可靠连接 (Reliable Connect, RC): 一个队列对 (Queue Pair, QP) 只和一个另外的 QP 相连。消息通过一个 QP 的发送队列可靠地传输到另一个 QP 的接收队列。数据包按序交付。RC 连接很类似于 TCP 连接。

具体地, RDMA 在双方建立 RC 连接之前, 首先创建上下文, 并且创建保护域 (Protection Domain, PD) 以将队列对 QP 与内存区域 (Memory Region, MR) 关联, 接着双方创建 QP, 每个 QP 包含两个先进先

出(first in first out, FIFO)的工作队列:即发送队列(send Queue, SQ)用于发送请求,和接收队列(receive Queue, RQ)用于接收请求,它们每个都与完成队列(Completion Queue, CQ)关联。为了允许 RDMA 网卡有权访问内存,需要对内存进行注册,双方都可以任意访问的此类内存称为支持 RDMA 访问的内存,而远程地址是此内存区域的起始虚拟地址。创建 MR 后,将生成一个 8 字节类型的密钥,双方应交换其密钥和已注册内存的虚拟地址。每个 MR 都包含自己的密钥(lkey)和远端内存的密钥(rkey),必须使用 rkey 来对远端已经注册的内存进行访问。在一般情况下,通过 RDMA 的原生通信库 RDMA_CM,创建 4096 个连接需要花费 10s 时间左右,平均创建每个连接需要花费 1ms~5ms 之间不等,而 RDMA 在负载为 32 字节的情况下,每次延迟大概为 4us,因此为了创建连接,需要延后数百个网络数据报文的发送。

7、页错误(Page fault):用虚拟地址读写内存时,发生缺页。

8、进程:是计算机中的程序关于某数据集合上的一次运行活动,是系统进行资源分配的基本单位,是操作系统结构的基础。在早期面向进程设计的计算机结构中,进程是程序的基本执行实体;在当代面向线程设计的计算机结构中,进程是线程的容器。程序是指令、数据及其组织形式的描述,进程是程序的实体。下文中进程也可以称为应用(application, APP)。

另外,为了便于理解本申请实施例,做出以下几点说明。

第一,在本申请中示出的“至少一个”是指一个或者多个,“多个”是指两个或两个以上。另外,在本申请的实施例中,“第一”、“第二”以及各种数字编号(例如,“#1”、“#2”等)只是为了描述方便进行的区分,并不用来限制本申请实施例的范围。下文各过程的序号的大小并不意味着执行顺序的先后,各过程的执行顺序应以其功能和内在逻辑确定,而不应对本申请实施例的实施过程构成任何限定,应该理解这样描述的对象在适当情况下可以互换,以便能够描述本申请的实施例以外的方案。此外,在本申请实施例中,“S310”等字样仅为了描述方便作的标识,并不是对执行步骤的次序进行限定。

第二,本申请实施例中,“示例性的”或者“例如”等词用于表示作例子、例证或说明。本申请中被描述为“示例性的”或者“例如”的任何实施例或设计方案不应被解释为比其他实施例或设计方案更优选或更具优势。确切而言,使用“示例性的”或者“例如”等词旨在以具体方式呈现相关概念。

第三,本申请实施例中涉及的“保存”,可以是指的保存在一个或者多个存储器中。该一个或者多个存储器,可以是单独的设置,也可以是集成在编码器或者译码器,处理器、或通信装置中。该一个或者多个存储器,也可以是一部分单独设置,一部分集成在译码器、处理器、或通信装置中。存储器的类型可以是任意形式的存储介质,本申请并不对此限定。

第四,本申请实施例中涉及的“包括”(也称“includes”、“including”、“comprises”和/或“comprising”)当在本说明书中使用指定存在所陈述的特征、整数、步骤、操作、元素、和/或部件,但是并不排除存在或添加一个或多个其他特征、整数、步骤、操作、元素、部件、和/或其分组。

第五,本申请实施例中涉及的“如果”可被解释为意指“当...时”(“when”或“upon”)或“响应于确定”或“响应于检测到”。类似地,根据上下文,短语“如果确定...”或“如果检测到[所陈述的条件或事件]”可被解释为意指“在确定...时”或“响应于确定...”或“在检测到[所陈述的条件或事件]时”或“响应于检测到[所陈述的条件或事件]”。

第六,本申请实施例中对各种所述示例的描述中所使用的术语只是为了描述特定示例,而并非旨在进行限制。如在对各种所述示例的描述和所附权利要求书中所使用的那样,数形式“一个(a”, “an”)和“该”旨在也包括复数形式,除非上下文另外明确地指示。

第七,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。

上文结合图 2 简单介绍了本申请提供的报文传输的方法适用的场景,以及介绍了本申请涉及的基本概念。并在基本概念中介绍了 RC,组网规模越大,所需连接数量越多,资源消耗越大。例如,组网中有 N 个结点(Host),每个结点有 P 个进程,全互联通信,则一个结点需要建立 N*P*P 个 RC 连接。

为了解决目前的基于 RDMA 组网中 RC 连接数量越多,资源消耗越大的问题,本申请提出一种基于 RDMA 的信息传输的方法,通过采用 UB 协议,降低组网中 RC 连接数量。下面将结合附图详细介绍本申请提供的基于 RDMA 的信息传输的方法。

应理解,本申请实施例提供的基于 RDMA 的信息传输的方法可以应用于计算机系统中,例如,图 2 所示的跨网络通信系统中。

还应理解，下文示出的实施例并未对本申请实施例提供的方法的执行主体的具体结构特别限定，只要能够通过运行记录有本申请实施例的提供的方法的代码的程序，以实现本申请实施例提供的方法即可。例如，本申请实施例提供的方法的执行主体可以是设备，或者，是设备中能够调用程序并执行程序的功能模块。

图 5 是本申请提供的一种报文传输的方法的示意性流程图。应用于第一设备和第二设备之间传输报文的场景下，如图 2 所示的场景中。该实施例中第一设备和第二设备均支持通用总线协议，即第一设备包括事务层和传输层，第二设备包括事务层和传输层。

具体地，图 5 所示的方法包括以下步骤：

S510，第一设备的传输层接收来自第一设备的事务层的报文。

该报文为待发送给第二设备的第一应用的报文。

该实施例中第一设备和第二设备可以为组网中任意的两个计算机设备，第一设备和第二设备之间可以跨网络通信，例如，第一设备和第二设备为图 2 中所示的两个服务器，或两个交换机，或一个为服务器一个为交换机。该实施例中对于第一设备和第二设备的具体形式不做限定，可以为任意的跨网络通信的计算机设备。

具体地，该实施例中第一设备和第二设备均为支持通用总线协议的计算机设备，从而第一设备包括事务层和传输层，第二设备也包括事务层和传输层，第一设备的传输层和第二设备的传输层之间可以建立至少一个 TP 连接，该至少一个 TP 连接可以组成一个 TPG，如图 4 中所示，第一设备可以为图 4 中所示的主机 A，第二设备可以为图 4 中所示的主机 B，第一设备的传输层和第二设备的传输层之间建立的 TPG 可以为图 4 中所示的 TPG，可以包括 8 个 TP 连接（如，图 4 中所示的 TP 连接#0 至 TP 连接#7，共 8 个 TP 连接），第一设备和第二设备之间所有的信息（包括数据、信令、报文等）都通过这个 TPG 传输。

应理解，上述第一设备的传输层和第二设备的传输层之间建立 TPG 仅为示例，对本申请的保护范围不构成任何的限定，组网中任意的两个计算机设备的传输层之间均可以建立 TPG。例如，如图 6 所示，组网中包括第一设备、第二设备#1、第二设备#2 和第二设备#3，第一设备的传输层可以分别和第二设备#1 的传输层、第二设备#2 的传输层和第二设备#3 的传输层建立 TPG（如，图 6 中所示的 TPG#1、TPG#2 和 TPG#3）。

进一步地，该实施例中第一设备的传输层接收到来自第一设备的事务层的报文之后，判断第一设备的传输层和第二设备的传输层之间是否已经建立至少一个第一 TP 连接，则图 5 所示的方法流程还包括：

S520，第一设备的传输层确定第一设备的传输层和第二设备的传输层之间是否已经建立 TP 连接。

方式 1：第一设备的传输层确定第一设备的传输层和第二设备的传输层之间已经建立至少一个第一 TP 连接。

例如，在第一设备的传输层接收来自第一设备的事务层的第一应用的报文之前，该第一设备的传输层接收来自第一设备的事务层的第二应用的报文，并且为了传输该第二应用的报文和第二设备的传输层之间建立至少一个第一 TP 连接，通过该至少一个第一 TP 连接传输第二应用的报文。

在该方式 1 所示的情况下，第一设备的传输层可以通过第一指示信息通知第一设备的事务层可以通过至少一个第一 TP 连接传输第一应用的报文。

具体地，在该方式 1 所示的情况下，图 5 所示的方法流程还包括：

S521，第一设备的传输层向第一设备的事务层发送第一指示信息。

该第一指示信息用于指示至少一个第一 TP 连接用于传输第一应用的报文。

S522，第一设备的传输层通过至少一个第一 TP 连接向第二设备的传输层发送第一应用的报文。

由前文基本概念中关于通用总线协议的介绍可知，通用总线协议定义了独立的事务层和传输层。两个设备的传输层之间有连接，事务层之间无连接。一个设备内的所有事务都承载在一个传输层上。因此，该实施例中不同应用的报文可以共享传输层的 TP 连接，从而设备之间建立的 TP 连接远少于需要建立的连接（如，基于 IB 协议所建立的 RC 连接）数量，可以降低设备之间的连接数量。

例如，组网中存在 N 个结点，每个结点 P 个应用，若基于 RC 连接实现互联通信，则一个结点需要建立 $N \times P \times P$ 个 RC 连接；若基于 TP 连接实现互联通信，则一个结点建立 N 个 TP 连接即可。

另外，TP 连接是常连接，不需要动态拆连接、建连接，从而避免了动态拆连接、建连接导致的资源消耗。

方式 2：第一设备的传输层确定第一设备的传输层和第二设备的传输层之间未建立 TP 连接。

例如，第一设备的传输层接收到的第一应用的报文为首个需要传输至第二设备的报文，第一设备的传输层为了传输该第一应用的报文，第一设备的传输层和第二设备的传输层之间需要建立至少一个第二 TP 连接，用于传输该第一应用的报文。

在方式 2 所示的情况下，第一设备的传输层和第二设备的传输层之间需要建立至少一个第二 TP 连接，则图 5 所示的方法流程还包括：

S523，第一设备的传输层获取第二设备的 IP 地址信息。

S524，第一设备的传输层根据第二设备的 IP 地址信息，建立第一设备的传输层和第二设备的传输层之间至少一个第二 TP 连接。

该至少一个第二 TP 连接用于传输第一应用的报文。

S525，第一设备的传输层通过至少一个第二 TP 连接向第二设备的传输层发送第一应用的报文。

在方式 2 所示的情况下，第一设备的传输层和第二设备的传输层之间建立至少一个第二 TP 连接的过程，可以理解为共享连接的建连过程，当第一设备的传输层和第二设备的传输层之间建立了至少一个第二 TP 连接之后，若第一设备的传输层接收到待发送给第二设备的第三应用的报文，可以通过该至少一个第二 TP 连接传输第三应用的报文，无需针对第三应用建立不同的连接，从而可以降低连接数量，节约建立连接所需的资源。

具体地，共享连接的建连过程可以理解为：当有应用的报文需要传输时，URMA 软件栈会根据目标 EID (Target EID) 确定对端设备的 IP 地址，然后由 UMDK 内的传输通道建连服务与外部节点的 URMA 软件栈建连服务相互协商，交换 PSN 和路径等信息，完成两端 IP 地址之间传输通道的建立。

需要说明的是，为了降低资源开销，URMA 软件栈提供了多种级别的传输通道共享策略，包括但不限于：

设备粒度共享：所有的应用、VM 均共享两个设备之间的传输通道；

VM 粒度共享：每个 VM 内的不同应用，共享传输通道；

应用内共享：在应用内，不同的 Jetty 共享传输通道；

Jetty 独占：在应用内，每个 Jetty 使用单独的共享传输通道。

本申请主要涉及设备粒度共享，即所有的应用、VM 共享两个设备之间的传输通道，示例性地，在方式 2 所示的情况下，第一设备的传输层和第二设备的传输层之间建立 TP 连接通信信息交换包括以下几种可能的实现方式：

作为一种可能的实现方式，第一设备使用业务通道或已有的 IP/TCP 连接来交换通信关系，例如，获取第二设备侧的 Segment 信息或 Jetty 信息，称为带外交换。

作为另一种可能的实现方式，第一设备使用 URMA 的公知 Jetty 的消息语义与第二设备交互，获取第二设备侧的 Segment 信息或 Jetty 信息，称为带内交换。其中，URMA 公知 Jetty 是指第二设备在创建 Jetty 时，可指定 Jetty 号，第一设备可通过约定或配置的方式，获取到公知 Jetty 号，然后使用本地 Jetty 与第二设备侧的公知 Jetty 交互。

为了便于理解，下面结合具体的示例，说明第一设备的传输层和第二设备的传输层之间建立 TP 连接的流程：

示例一：第一设备的传输层和第二设备的传输层之间建立 TP 连接，包括以下步骤：

步骤一：第一设备侧的应用#1 获取第二设备侧的应用#2 的 home_seg 信息（如，第二设备的 EID）。获取方式使用前面介绍的带外交换或带内交换。

步骤二：第一设备侧的应用#1 调用函数（如，调用 `urma_import_seg()`）时，陷入内核。

步骤三：URMA 软件栈检查第二设备的 EID 确定第一设备和第二设备之间有没有创建 TP，如果有则直接把 TP 连接号（TP Number, TPN）放在 target seg 数据结构中。

步骤四：如果第一设备和第二设备之间没有创建 TP，继续下述步骤五至步骤九。

步骤五：第一设备根据第二设备的 EID 获取第二设备的 IP 地址。当直接用物理机（Physical Function, PF）时，EID 和 IP 地址相同可直接转换，当使用多通道（Multipath）设备时，EID 相当于主机名称（host name），可通过查表方式或者 DNS 方式得到多个 IP 地址。

步骤六：创建建链通道：使用第一设备和第二设备的公知 Jetty，或者使用获取到 IP 地址建立 TCP socket

步骤七：第一设备通过建链通道向第二设备发送建链请求，协商传输层（Transport）的初始信息：

可选地，建链请求报文内容包括：第一设备的 TPN，初始 PSN，IP 地址，第一设备的身份凭据、建链策略和参数（例如，src/dst EID）、拥塞控制策略和参数等。

可选地，建链响应报文内容包括：第二设备的 TPN，IP 地址等。

步骤八：第一设备记录第二设备的 EID 到 TPN 的映射关系，第二设备记录第一设备的 EID 到 TPN 的关系。第一设备将 TPN 返回，放在 target seg 数据结构中。

步骤九：建链完成

可选地，如果是创建 TPG，则在建链请求报文中一次传递多组建链信息，批量完成多个 TP 建链。

另外，该实施例中为了实现不同 VM 共享物理传输通道节省资源的目的，物理传输通道在主机（Host）中维护，VM 内无需建立物理传输通道，VM 通知 Host 建链服务（TP Service）完成物理传输通道的建立即可。具体地，VM 内的 URMA 软件栈建立传输通道时，UBN 设备定义为支持分流（Offloading）能力，则 VM 内的 URMA 软件栈不再执行建链，URMA 软件栈把建立传输通道请求发给 UBN 设备驱动，硬件将此建立传输通道的请求通过邮箱（Mailbox）发给 Host 上 URMA 软件栈建链服务，为此 VM 分配虚拟传输通道（virtual TP，vTP），以及创建 vTP 和物理传输通道 TP 或 TPG 的映射关系，然后再通过 Mailbox 返回 vTP 给 VM 的 URMA 软件栈，后续 VM 内的 URMA 软件栈则可使用 vTP 进行通信。

由上述的方式 1 和方式 2 可知，该实施例中第一设备的传输层接收到待发送给第二设备的第一应用的报文之后，若确定第一设备和第二设备的传输层之间已经建立至少一个第一 TP 连接，可以通过该至少一个第一 TP 连接直接向第二设备的传输层发送报文，若确定第一设备和第二设备的传输层之间未建立连接，需要建立用于传输第一应用的报文的至少一个第二 TP 连接，为了便于描述下文中以通过至少一个第一 TP 连接传输第一应用的报文为例进行说明。

具体地，第一设备的事务层需要向第二设备发送第一应用的报文时，该第一应用的报文由发送队列条目（Send Queue Entry，SQE）承载，该 SQE 的目的地址为第二设备的地址。例如，该 SQE 的目的地址为第二设备的 IP 地址。其中，第二设备的 IP 地址可以是第二设备中的一个虚拟机的 Entity ID。

应理解，当第一设备的传输层和多个不同的设备的传输层之间建立了 TP 连接的情况下，第一设备的某个应用可以向不同的设备发送报文。例如，第一设备一个应用需要分别发 3 个报文给 3 个不同的设备（如，第一设备的应用需要向图 6 中所示的第二设备#1、第二设备#2 和第二设备#3 分别发送报文），3 个报文分别由 SQE#0、SQE#1、SQE#2 承载。其中 SQE#0 目的地是第二设备#1，SQE#1 目的地是第二设备#1，SQE#2 目的地是第二设备#1。具体地，第一设备的应用并不感知传输层的 TP 连接，只需在相应的 SQE 中指定目的地即可。

该实施例中，第一设备的应用可以将 SQE 写入 JFS。可选地，若第一设备的应用需要向不同的设备分别发送不同的报文，则可以将多个 SQE 写入同一个 JFS。例如，如图 7 所示第一设备的应用将目的地分别为第二设备#1、第二设备#2 和第二设备#3 的 SQE#0、SQE#1 和 SQE#2 写入同一个 JFS。

具体地，第一设备的应用将 SQE 写入 JFS 之后，SQE 可以从 JFS 调出，发送到第一设备的传输层。第一设备的传输层根据 SQE 中的目的实体标识（Destination Entity ID，DEID），找到对应的 TPG，把 SQE 发到对应的 TPG 上。

例如，JFS 中写入了 SQE#0、SQE#1 和 SQE#2，其中，SQE#0 中的 DEID 与 TPG#1 对应；SQE#1 中的 DEID 与 TPG#2 对应；SQE#2 中的 DEID 与 TPG#3 对应，则第一设备的传输层接收到 SQE#0 之后可以通过 SQE#0 中的 DEID 确定通过 TPG#1 中的 TP 连接传输该 SQE#0，可以将该 SQE#0 传输至正确的目的地。

可选地，当一个 JFS 中写入了多个 SQE 的情况下，多个 SQE 可以并行调出 JFS（如，多个 TPG 并行调出 SQE），发到不同设备。

示例性地，第一设备的传输层通过至少一个第一 TP 连接向第二设备的传输层发送所述第一应用的报文，包括：

第一设备的传输层通过至少一个第一 TP 连接中的负载最轻的 TP 连接向第二设备的传输层发送第一应用的报文。

例如，如图 8 所示，第一设备的传输层和第二设备的传输层之间建立了 8 个 TP 连接，每个 TP 连接有各自的 TP 队列（TP Queue）和 TP 上下文（TP context）。第一设备的应用 A 和应用 C 各需要向第二设备发送一个 1MB 的消息，应用 A 和应用 C 并不感知 TPG 内包含多少 TP 连接。第一设备的 JFS 可以把

1MB 的消息拆成 16 个 64KB 的 segment，每次发一个 segment 给 TPG。TPG 选择负载最轻的 TP 连接发送这个 segment。即流量在 TPG 内的多个 TP 连接间负载均衡，充分利用多端口多路径，减小消息传输时间。

可选地，发生网络丢包，第一设备的 TP 连接负责重传，从而可以保证对第一设备的应用进程提供可靠的服务。

示例性地，第一设备的传输层向第二设备的传输层发送 SQE 之后，第二设备的传输层可以向第一设备的传输层反馈应答消息，以告知第一设备是否成功传输 SQE，则图 5 所示的方法流程还包括：

S530，第二设备的传输层向第一设备的传输层发送应答消息。

作为一种可能的实现方式，应答消息为 TAACK，指示报文在第二设备的事务层执行成功。在该实现方式下，第一设备的传输层向第一设备的 JFS 发送 TAACK，第一设备的 JFS 收到 TAACK 后，向应用进程上报完成队列条目（Complete Queue Entry，CQE），告知应用进程 SQE 执行完成。

作为另一种可能的实现方式，应答消息为 TANAK，指示报文在第二设备的事务层执行失败。在该实现方式下，第一设备的传输层向第一设备的 JFS 发送 TANAK，第一设备的 JFS 收到 TANAK 后，重新向第一设备的传输层发送 SQE，第一设备的传输层重新向所述第二设备的传输层发送第一应用的报文。

图 5 所示的报文传输的方法，发送端和接收端之间建立 TP 连接，报文通过 TP 连接传输，相比于建立 RC 连接可以降低建立连接的数量，从而降低资源消耗。并且，一个 TPG 内包含多个 TP 连接，应用不感知 TPG 内包含多少 TP 连接，事务层将事务（如，上述的 SQE）发给 TPG，事务在 TPG 内多条 TP 连接均衡传输，实现多端口多路径，提升网络利用率，缩短事务传输时间。

进一步地，基于 TP 连接实现网络拥塞控制。不同应用共用 TP 连接，共用丢包重传和拥塞控制，相比 IB 中每个 QP 各自为政的拥塞控制，基于 TP 连接的拥塞控制更能统揽全局，有助于减少网络上的突发，减少网络排队。而且事务层和传输层分开，事务层异常，在事务层重传，不阻塞共享的传输层。

为了便于理解，下面结合具体的示例说明基于第一设备和第二设备之间的报文传输流程。

示例二：如图 9 所示，以发送（Send）消息的传输流程为例说明第一设备和第二设备之间的信息传输流程，Send 消息的传输包括以下步骤：

S910，第一设备的应用进程下发‘Send’事务到 JFS。

S920，第一设备的 JFS 把‘Send’事务发给 TPG，TPG 中选负载最轻的 TP 连接，为‘Send’事务封装 TPH。

S930，Link 层再为‘Send’事务封装 Link header 得到基于 UB 协议封装的‘Send’事务得到包，通过网络将包发到第二设备的传输层。

S940，第二设备的传输层收到包，如果包校验通过，返回 TPACK（传输层应答），告知第一设备的传输层‘包已经正确接收’，如果检查 TPH 中的 PSN，发现网络有丢包，则返回传输层选择性应答（Transport Selective ACK，TPSACK），告知第一设备的传输层‘该包已正确接收，但有丢包，要重传丢失的 PSN 对应的包’。

S950，第二设备的传输层解析包，得知该包的目的地是某个 VM。

例如，第二设备的传输层通过 DEID 查表确定包的目的地。第二设备的传输层把包中 TPH 及之前的部分剥离，剩余部分转给 VM。

S960，第二设备的 VM 的某个 JFR 接收此包，通过 UBMMU 校验，如果正确写入内存，则产生 TAACK 发给第一设备，如果写内存时发生 pagefault 或 JFR 无资源可接收，则产生 TANAK 发给第一设备。从第二设备的传输层看，TAACK、TANAK 也是一个事务，第二设备的传输层也会保证可靠到达第一设备。

第一设备的 JFS 收到 TAACK，知道该事务已经被正确执行，则产生 CQE 告知应用进程。如果收到 TANAK，指示第二设备侧出现 pagefault 或资源不足，则第一设备的 JFS 重传该事务。

应理解，本申请实施例中的图 5 至图 9 所示的具体的例子只是为了帮助本领域技术人员更好地理解本申请实施例，而非限制本申请实施例的范围。还应理解，上述各过程的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不应对本申请实施例的实施过程构成任何限定。

还应理解，在本申请的各个实施例中，如果没有特殊说明以及逻辑冲突，不同的实施例之间的术语和/或描述具有一致性、且可以相互引用，不同的实施例中的技术特征根据其内在的逻辑关系可以组合形成新的实施例。

上述主要从方法的角度对本申请实施例提供的方案进行了介绍。为了实现上述功能，其包含了执行各个功能相应的硬件结构和/或软件模块。本领域技术人员应该很容易意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，本申请能够以硬件或硬件和计算机软件的结合形式来实现。某个功能究竟以硬件还是计算机软件驱动硬件的方式来执行，取决于技术方案的特定应用和设计约束条件。专业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申请的范围。

以下，结合图 10 至图 12 详细说明本申请实施例提供的报文传输的装置。应理解，装置实施例的描述与方法实施例的描述相互对应，因此，未详细描述的内容可以参见上文方法实施例，为了简洁，部分内容不再赘述。

本申请实施例可以根据上述方法示例对发送端设备或者接收端设备进行功能模块的划分，例如，可以对应各个功能划分各个功能模块，也可以将两个或两个以上的功能集成在一个处理模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。需要说明的是，本申请实施例中对模块的划分是示意性的，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。下面以采用对应各个功能划分各个功能模块为例进行说明。

图 10 示出了本申请实施例提供的一种报文传输的装置 1000 的结构示意图。

一种示例，报文传输的装置 1000 可以应用于第一设备，报文传输的装置 1000 可以用于执行上述的报文传输的方法，例如用于执行图 5 所示的方法。其中，报文传输的装置 1000 包括事务层和传输层。具体地，报文传输的装置 1000 可以包括收发单元 1010 和处理单元 1020。

收发单元 1010，用于接收来自所述第一设备的事务层的报文，所述报文为待发送给第二设备的第一应用的报文；当所述第一设备的传输层和第二设备的传输层之间已建立至少一个第一传输层 TP 连接时，所述收发单元 1010，用于向所述第一设备的事务层发送第一指示信息，所述第一指示信息用于指示所述至少一个第一 TP 连接用于传输所述第一应用的报文；所述收发单元 1010，用于通过所述至少一个第一 TP 连接向所述第二设备的传输层发送所述第一应用的报文，其中，所述至少一个第一 TP 连接是所述第一设备的传输层为了传输第二应用的报文而建立的 TP 连接。

作为示例，结合图 5，收发单元 1010 可以用于执行 S510、S521、S522、S524、S525 和 S530，处理单元 1020 可以用于执行 S520 和 S523。

需要说明的是，图 10 所述的装置也可以用于执行前述提及的附图所示的实施例变形所涉及的方法步骤，在此不再赘述。

另一种示例，报文传输的装置 1000 可以应用于第二设备，报文传输的装置 1000 可以用于执行上述的报文传输的方法，例如用于执行图 5 所示的方法。其中，报文传输的装置 1000 包括事务层和传输层。

收发单元 1030，用于接收来自第一设备的传输层的第一应用的报文。

作为示例，结合图 5，收发单元 1010 可以用于执行 S522、S524、S525 和 S530。

需要说明的是，图 10 所述的装置也可以用于执行前述提及的附图所示的实施例变形所涉及的方法步骤，在此不再赘述。

本申请实施例还提供一种芯片系统 1100，如图 11 所示，该芯片系统 1100 包括至少一个处理器和至少一个接口电路。作为示例，当该芯片系统 1100 包括一个处理器和一个接口电路时，则该一个处理器可以是图 11 中实线框所示的处理器 1110(或者是虚线框所示的处理器 1110)，该一个接口电路可以是图 11 中实线框所示的接口电路 1120(或者是虚线框所示的接口电路 1120)。

当该芯片系统 1100 包括两个处理器和两个接口电路时，则该两个处理器包括图 11 中实线框所示的处理器 1110 和虚线框所示的处理器 1110，该两个接口电路包括图 11 中实线框所示的接口电路 1120 和虚线框所示的接口电路 1120。对此不作限定。处理器 1110 和接口电路 1120 可通过线路互联。例如，接口电路 1120 可用于接收信号(例如存储器中存储的指令等)。又例如，接口电路 1120 可用于向其它装置(例如处理器 1110)发送信号。

示例性的，接口电路 1120 可读取存储器中存储的指令，并将该指令发送给处理器 1110。当该指令被处理器 1110 执行时，可使得报文传输的装置执行上述实施例中的各个步骤。当然，该芯片系统 1100 还可以包含其他分立器件，本申请实施例对此不作具体限定。

本申请另一实施例还提供一种计算机可读存储介质，该计算机可读存储介质中存储有指令，当指令在报文传输的装置上运行时，该报文传输的装置执行上述方法实施例所示的方法流程中该报文传输的装

置执行的各个步骤。在一些实施例中，所公开的方法可以实施为以机器可读格式被编码在计算机可读存储介质上的或者被编码在其它非瞬时性介质或者制品上的计算机程序指令。

图 12 示意性地示出本申请实施例提供的计算机程序产品的概念性局部视图，该计算机程序产品包括用于在计算机设备上执行计算机进程的计算机程序。

在一个实施例中，计算机程序产品是使用信号承载介质 1200 来提供的。该信号承载介质 1200 可以包括一个或多个程序指令，其当被一个或多个处理器运行时可以提供以上针对图 5 和图 9 描述的功能或者部分功能。因此，例如，参考图 5 中 S510~S530 的一个或多个特征可以由与信号承载介质 1200 相关的一个或多个指令来承担。此外，图 12 中的程序指令也描述示例指令。

在一些示例中，信号承载介质 1200 可以包含计算机可读介质 1201，诸如但不限于，硬盘驱动器、紧凑盘(CD)、数字视频光盘(DVD)、数字磁带、存储器、只读存储记忆体 readonly memory, ROM)或随机存储记忆体(random access memory, RAM)等等。

在一些实施方式中，信号承载介质 1200 可以包含计算机可记录介质 1202，诸如但不限于，存储器、读/写(R/W)CD、R/W DVD、等等。

在一些实施方式中，信号承载介质 1200 可以包含通信介质 1203，诸如但不限于，数字和/或模拟通信介质(例如，光纤电缆、波导、有线通信链路、无线通信链路、等等)。信号承载介质 1200 可以由无线形式的通信介质 1203(例如，遵守 IEEE 1502 .11 标准或者其它传输协议的无线通信介质)来传达。一个或多个程序指令可以是，例如，计算机可执行指令或者逻辑实施指令。

在一些示例中，诸如针对图 5 报文传输的装置可以被配置为，响应于通过计算机可读介质 1201、计算机可记录介质 1202、和/或通信介质 1203 中的一个或多个程序指令，提供各种操作、功能、或者动作。

应该理解，这里描述的布置仅仅是用于示例的目的。因而，本领域技术人员将理解，其它布置和其它元素(例如，机器、接口、功能、顺序、和功能组等等)能够被取而代之地使用，并且一些元素可以根据所期望的结果而一并省略。另外，所描述的元素中的许多是可以被实现为离散的或者分布式的组件的、或者以任何适当的组合和位置来结合其它组件实施的功能实体。

在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件程序实现时，可以全部或部分地以计算机程序产品的形式来实现。该计算机程序产品包括一个或多个计算机指令。在计算机上和/或执行计算机指令时，全部或部分地产生按照本申请实施例的流程或功能。计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。

计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一个计算机可读存储介质传输，例如，计算机指令可以从一个网站站点、计算机、服务器或者数据中心通过有线(如同轴电缆、光纤、数字用户线(digital subscriber line, DSL))或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可以用介质集成的服务器、数据中心等数据存储设备。可用介质可以是磁性介质(例如，软盘、硬盘、磁带)，光介质(例如，DVD)、或者半导体介质(例如固态硬盘(solid statedisk, SSD))等。

以上所述，仅为本申请的具体实施方式，但本申请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申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以所述权利要求的保护范围为准。

权利要求书

1. 一种报文传输的方法，其特征在于，应用于第一设备，所述第一设备包括事务层和传输层，所述方法包括：

所述第一设备的传输层接收来自所述第一设备的事务层的报文，所述报文为待发送给第二设备的第一应用的报文；

当所述第一设备的传输层和第二设备的传输层之间已建立至少一个第一传输层 TP 连接时，所述第一设备的传输层向所述第一设备的事务层发送第一指示信息，所述第一指示信息用于指示所述至少一个第一 TP 连接用于传输所述第一应用的报文；

所述第一设备的传输层通过所述至少一个第一 TP 连接向所述第二设备的传输层发送所述第一应用的报文，其中，所述至少一个第一 TP 连接是所述第一设备的传输层用于传输第二应用的报文的 TP 连接。

2. 根据权利要求 1 所述的方法，其特征在于，当所述第一设备的传输层和第二设备的传输层之间未建立 TP 连接时，所述方法还包括：

所述第一设备的传输层获取所述第二设备的互联网协议 IP 地址信息；

所述第一设备的传输层根据所述第二设备的 IP 地址信息，建立所述第一设备的传输层和第二设备的传输层之间至少一个第二 TP 连接，所述至少一个第二 TP 连接用于传输所述第一应用的报文。

3. 根据权利要求 1 或 2 所述的方法，其特征在于，所述第一设备的传输层通过所述至少一个第一 TP 连接向所述第二设备的传输层发送所述第一应用的报文，包括：

所述第一设备的传输层通过所述至少一个第一 TP 连接中的负载最轻的 TP 连接向所述第二设备的传输层发送所述第一应用的报文。

4. 根据权利要求 1 至 3 中任一项所述的方法，其特征在于，所述方法还包括：

所述第一设备的传输层接收来自所述第二设备的传输层的应答消息，所述应答消息用于指示所述第一应用的报文是否成功被所述第二设备的事务层执行。

5. 根据权利要求 4 所述的方法，其特征在于，若所述应答消息为事物应答 TAACK，指示所述第二设备的事务层成功执行所述第一应用的报文，所述方法还包括：

所述第一设备的传输层向所述第一设备的事务层发送完成队列条目 CQE，所述 CQE 用于指示所述第一应用的报文成功传输。

6. 根据权利要求 4 所述的方法，其特征在于，若所述第一响应消息为事物否定应答 TANAK，指示所述第二设备的事务层未成功执行所述第一应用的报文，所述方法还包括：

所述第一设备的传输层向所述第一设备的事务层发送所述 TANAK；

所述第一设备的传输层接收来自所述第一设备的事务层的所述第一应用的报文，并重新向所述第二设备的传输层发送所述第一应用的报文。

7. 根据权利要求 1 至 6 中任一项所述的方法，其特征在于，所述第一设备的传输层接收来自所述第一设备的事务层的报文，包括：

所述第一设备的传输层接收来自所述第一设备的事务层的发送队列条目 SQE，所述 SQE 的目的地址为所述第二设备的地址，所述 SQE 承载发送给所述第二设备所述第一应用的报文。

8. 根据权利要求 7 所述的方法，其特征在于，所述第一设备的传输层接收来自所述第一设备的事务层的 SQE，包括：

所述第一设备的传输层接收来自所述第一设备的至少一个发送接口的所述 SQE。

9. 根据权利要求 7 或 8 所述的方法，其特征在于，在所述第一设备的传输层通过所述至少一个第一 TP 连接向所述第二设备的传输层发送所述第一应用的报文之前，所述方法还包括：

所述第一设备的传输层根据所述 SQE 中携带的目的实体标识 DEID 确定所述至少一个第一 TP 连接。

10. 一种报文传输的装置，其特征在于，所述装置包括事务层和传输层，所述装置包括：

传输层接收模块，用于接收来自所述事务层的报文，所述报文为待发送给第二设备的第一应用的报文；

当所述装置的传输层和第二设备的传输层之间已建立至少一个第一传输层 TP 连接时，传输层发送模块，用于向所述事务层发送第一指示信息，所述第一指示信息用于指示所述至少一个第一 TP 连接用于传

输所述第一应用的报文；

所述传输层发送模块，还用于通过所述至少一个第一 TP 连接向所述第二设备的传输层发送所述第一应用的报文；

其中，所述至少一个第一 TP 连接是所述装置的传输层用于传输第二应用的报文的 TP 连接。

11. 根据权利要求 10 所述的装置，其特征在于，当所述装置的传输层和第二设备的传输层之间未建立 TP 连接时，所述传输层接收模块，还用于获取所述第二设备的互联网协议 IP 地址信息；

所述装置还包括：

传输层处理模块，用于根据所述第二设备的 IP 地址信息，建立所述装置的传输层和第二设备的传输层之间的至少一个第二 TP 连接，所述至少一个第二 TP 连接用于传输所述第一应用的报文。

12. 根据权利要求 10 或 11 所述的装置，其特征在于，所述传输层发送模块通过所述至少一个第一 TP 连接向所述第二设备的传输层发送所述第一应用的报文，包括：

所述传输层发送模块通过所述至少一个 TP 连接中的负载最轻的 TP 连接向所述第二设备的传输层发送所述第一应用的报文。

13. 根据权利要求 10 至 12 中任一项所述的装置，其特征在于，所述传输层接收模块，还用于接收来自所述第二设备的传输层的应答消息，所述应答消息用于指示所述第一应用的报文是否成功被所述第二设备的事务层执行。

14. 根据权利要求 13 所述的装置，其特征在于，若所述应答消息为事务应答 TAACK，指示所述第二设备的事务层成功执行所述第一应用的报文，所述传输层发送模块，还用于向所述第一设备的事务层发送完成队列条目 CQE，所述 CQE 用于指示所述第一应用的报文成功传输。

15. 根据权利要求 13 所述的装置，其特征在于，若所述第一响应消息为事物否定应答 TANAK，指示所述第二设备的事务层未成功执行所述第一应用的报文，所述的传输层发送模块，还用于向所述第一设备的事务层发送所述 TANAK；

所述传输层接收模块，还用于接收来自所述第一设备的事务层的所述第一应用的报文；

所述传输层发送模块，还用于重新向所述第二设备的传输层发送所述第一应用的报文。

16. 根据权利要求 10 至 15 中任一项所述的装置，其特征在于，所述传输层接收模块接收来自所述事务层的报文，包括：

所述传输层接收模块，用于接收来自所述事务层的发送队列条目 SQE，所述 SQE 的目的地址为所述第二设备的地址，所述 SQE 承载发送给所述第二设备所述第一应用的报文。

17. 根据权利要求 16 所述的装置，其特征在于，所述传输层接收模块接收来自所述事务层的 SQE，包括：

所述装置的传输层接收模块接收来自所述装置的至少一个发送接口的所述 SQE。

18. 根据权利要求 16 或 17 所述的装置，其特征在于，所述装置包括：

传输层处理模块，用于在所述传输层发送模块通过所述至少一个第一 TP 连接向所述第二设备的传输层发送所述第一应用的报文之前，根据所述 SQE 中携带的目的实体标识 DEID 确定所述至少一个第一 TP 连接。

19. 一种报文传输的装置，其特征在于，包括：处理器，用于读取存储器中存储的指令，当所述处理器执行所述指令时，使得所述用于访问内存的装置实现权利要求 1 至 9 中任一项所述的方法。

20. 一种报文传输的装置，其特征在于，包括用于执行如权利要求 1 至 9 中任一项所述的方法的单元。

21. 一种芯片，其特征在于，包括：

至少一个处理核，用于执行如权利要求 1 至 9 中任一项所述的方法。

22. 一种计算机设备，其特征在于，包括：如权利要求 21 所示的芯片。

23. 一种计算机程序产品，其特征在于，所述计算机程序产品包括计算机程序代码，当所述计算机程序代码在计算机上运行时，权利要求 1 至 9 中任一项所述的方法被执行。

24. 一种计算机可读存储介质，其特征在于，包括计算机程序，当其在计算机设备上运行时，使得所述计算机设备中的处理模块执行如权利要求 1 至 9 中任意一项所述的方法。

25. 一种报文传输的系统，其特征在于，包括用于执行如权利要求 1 至 9 中任一项所述的方法的第一设备和用于接收报文的第二设备。