(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118714183 A
(43)申请公布日 2024.09.27

(21)申请号 202410929120.3
(22)申请日 2023.04.28
(62)分案原申请数据
202310483155.4 2023.04.28
(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼
(72)发明人 屈向峰 程中武 李冰 程传宁 谭焜
(74)专利代理机构 广州三环专利商标代理有限公司 44202
专利代理师 刘丽萍
(51)Int.Cl.
H04L 67/141(2022.01)
H04L 67/14(2022.01)
H04L 69/163(2022.01)

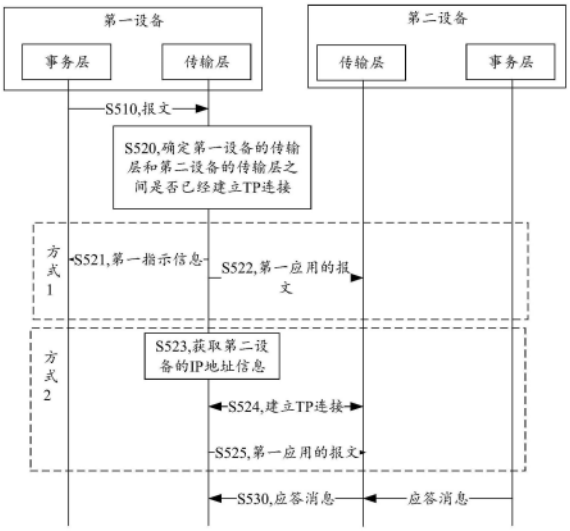
权利要求书3页 说明书17页 附图6页

(54)发明名称

报文传输的方法和装置

(57)摘要

本申请提供了一种报文传输的方法,应用于第一设备,第一设备包括传输层和事务层,该方法包括:第一设备接收待发送给第二设备的第一应用的报文;第一设备确定第一设备的传输层和第二设备的传输层之间是否建立传输层TP连接;当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TP连接时,第一设备确定该至少一个TP连接用于传输第一应用的报文,其中,至少一个TP连接是第一设备为了传输第二应用的报文而建立的TP连接。通过该报文传输的方法,不同的应用的报文可以通过相同的TP连接传输,无需针对不同的应用分别建立可靠连接,从而可以降低建立连接所需的资源开销。



CN 118714183 A

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获取待发送给第二设备的第一应用的报文,所述报文来自所述第一设备的事务层,所述第一设备和所述第二设备通过网络通信;

当所述第一设备的传输层和所述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连接用于传输所述第一设备的多个应用的报文,所述多个应用包括所述第一应用和第二应用。

2.根据权利要求1所述的方法,其特征在于,在所述当所述第一设备的传输层和所述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之前,所述方法还包括:

当所述第一设备的传输层和所述第二设备的传输层之间未建立TP连接时,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按照负载均衡方式通过所述至少一个第一TP连接中的一个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3所述的方法,其特征在于,所述第一设备的传输层按照负载均衡方式通过所述至少一个第一TP连接中的一个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:

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。

5.根据权利要求1至4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获取应答消息,所述应答消息来自所述第二设备的传输层,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。

6.根据权利要求5所述的方法,其特征在于,若所述应答消息为事物应答TAACK,则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。

7.根据权利要求5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,则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获取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。

8.根据权利要求1至7中任一项所述的方法,其特征在于,所述第一设备的传输层获取待发送给第二设备的第一应用的报文,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。

9.根据权利要求8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发送的所述SQE。

10.根据权利要求8或9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。

11.一种报文传输的装置,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,用于获取待发送给第二设备的第一应用的报文,所述报文来自所述事务层,所述第一设备和所述第二设备通过网络通信;

当所述装置的传输层和所述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连接用于传输所述第一设备的多个应用的报文,所述多个应用包括所述第一应用和第二应用。

12.根据权利要求11所述的装置,其特征在于,所述装置还包括:

传输层处理模块,用于当所述第一设备的传输层和所述第二设备的传输层之间未建立TP连接时,获取所述第二设备的互联网协议IP地址信息,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。

13.根据权利要求11所述的装置,其特征在于,包括:

所述传输层发送模块,具体用于按照负载均衡方式通过所述至少一个TP连接中的一个的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。

14.根据权利要求11所述的装置,其特征在于,包括:

所述传输层发送模块,具体用于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。

15.根据权利要求11至14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。

16.根据权利要求11至15中任一所述的装置,其特征在于,若所述应答消息为事务应答TAACK,则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。

17.根据权利要求11至15中任一所述的装置,其特征在于,若所述第一响应消息为事物否定应答TANAK,则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。

18.根据权利要求11至17中任一项所述的装置,其特征在于,包括:

所述传输层接收模块,具体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。

19.根据权利要求18所述的装置,其特征在于,包括:

所述装置的传输层,具体用于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。

20.根据权利要求18或19所述的装置,其特征在于,

所述传输层处理模块,还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。

21.一种报文传输的装置,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。

22.一种报文传输的装置,其特征在于,包括用于执行如权利要求1至10中任一项所述的方法的单元。

23.一种芯片,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10中任一项所述的方法。

24.一种计算机设备,其特征在于,包括:如权利要求23所示的芯片。

25.一种计算机程序产品,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10中任一项所述的方法被执行。

26.一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10中任意一项所述的方法。

27.一种报文传输的系统,其特征在于,包括用于执行如权利要求1至10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。

报文传输的方法和装置

[0001] 本申请是分案申请,原申请的申请号是202310483155.4,原申请日是2023年4月28日,原申请的全部内容通过引用结合在本申请中。

技术领域

[0002] 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

[0003] 通信设备跨网络通信的应用场景中对网络的述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。

[0004] 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

[0005] 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。

[0006] 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:

[0007] 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。

[0008] 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。

[0009] 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。

[0010] 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。

[0011] 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。

[0012] 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。

[0013] 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。

[0014] 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。

[0015] 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。

[0016] 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。

[0017] 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。

[0018] 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。

[0019] 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。

[0020] 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。

[0021] 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。

[0022] 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。

[0023] 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。

[0024] 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。

[0025] 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。

[0026] 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。

[0027] 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。

[0028] 第八方面,提供一种终端设备,该终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。

[0029] 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

[0030] 图1为本申请实施例提供的计算机设备的结构示意图。

[0031] 图2为本申请实施例提供的数据中心示意图。

[0032] 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。

[0033] 图4是本申请实施例提供的一种主机的结构示意图。

[0034] 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。

[0035] 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。

[0036] 图7是本申请提供的多个SQE写入一个JFS的示意图。

[0037] 图8是本申请提供的负载均衡的示意图。

[0038] 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。

[0039] 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。

[0040] 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。

[0041] 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

[0042] 下面将结合附图,对本申请实施例中的技术方案进行描述。

[0043] 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。

[0044] 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。

[0045] 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。

[0046] 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)。

[0047] 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。

[0048] 内存103也称作内存储器或主存储器,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。

[0049] 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。

[0050] 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。

[0051] 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102可见的虚拟地址(也可以称作设备地址或IO地址)映射为物理地址。

[0052] 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,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)1071和固态硬盘(solid state drive,SSD)。

[0053] 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送信息的公共通信干线。总线108可以是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。

[0054] 其中,内部总线采用内部总线协议传送信息。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送信息。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。

[0055] 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(IntelTM 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 systeminterface,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。

[0056] 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。

[0057] 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。

[0058] 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。

[0059] 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。

[0060] 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。

[0061] 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:

[0062] 表1

[0063]

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
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。

[0064] 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)。

[0065] 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。

[0066] 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。

[0067] 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。

[0068] 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。

[0069] 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。

[0070] 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。

[0071] 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。

[0072] 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。

[0073] 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。

[0074] 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。

[0075] 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。

[0076] 7、页错误(Page fault):用虚拟地址读写内存时,发生缺页。

[0077] 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)。

[0078] 另外,为了便于理解本申请实施例,做出以下几点说明。

[0079] 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。

[0080] 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。

[0081] 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。

[0082] 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时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。

[0083] 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]”。

[0084] 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。

[0085] 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。

[0086] 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。

[0087] 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。

[0088] 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。

[0089] 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。

[0090] 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。

[0091] 具体地,图5所示的方法包括以下步骤:

[0092] S510,第一设备的传输层接收来自第一设备的事务层的报文。

[0093] 该报文为待发送给第二设备的第一应用的报文。

[0094] 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。

[0095] 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物层和传输层,第二设备也包括事物层和传输层,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。

[0096] 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)。

[0097] 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:

[0098] 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。

[0099] 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。

[0100] 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。

[0101] 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。

[0102] 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:

[0103] 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。

[0104] 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。

[0105] 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。

[0106] 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。

[0107] 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*P*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。

[0108] 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。

[0109] 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。

[0110] 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。

[0111] 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:

[0112] S523,第一设备的传输层获取第二设备的IP地址信息。

[0113] 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。

[0114] 该至少一个第二TP连接用于传输第一应用的报文。

[0115] 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。

[0116] 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。

[0117] 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。

[0118] 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:

[0119] 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;

[0120] VM粒度共享:每个VM内的不同应用,共享传输通道;

[0121] 应用内共享:在应用内,不同的Jetty共享传输通道;

[0122] Jetty独占:在应用内,每个Jetty使用单独的共享传输通道。

[0123] 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:

[0124] 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。

[0125] 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。

[0126] 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:

[0127] 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:

[0128] 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。

[0129] 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。

[0130] 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。

[0131] 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。

[0132] 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。

[0133] 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。

[0134] 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:

[0135] 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。

[0136] 可选地,建链响应报文内容包括:第二设备的TPN,IP地址等。

[0137] 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。

[0138] 步骤九:建链完成。

[0139] 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。

[0140] 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。

[0141] 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。

[0142] 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。

[0143] 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。

[0144] 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。

[0145] 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。

[0146] 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#3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。

[0147] 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。

[0148] 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:

[0149] 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。

[0150] 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。

[0151] 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。

[0152] 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:

[0153] S530,第二设备的传输层向第一设备的传输层发送应答消息。

[0154] 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。

[0155] 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。

[0156] 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。

[0157] 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。

[0158] 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。

[0159] 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:

[0160] S910,第一设备的应用进程下发‘Send’事务到JFS。

[0161] 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。

[0162] 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。

[0163] 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’。

[0164] S950,第二设备的传输层解析包,得知该包的目的地是某个VM。

[0165] 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。

[0166] 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。

[0167] 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。

[0168] 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。

[0169] 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。

[0170] 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。

[0171] 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。

[0172] 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。

[0173] 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。

[0174] 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可以包括获取单元1010和处理单元1020。

[0175] 获取单元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,所述获取单元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;所述获取单元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。

[0176] 作为示例,结合图5,获取单元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。

[0177] 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。

[0178] 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。

[0179] 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。

[0180] 作为示例,结合图5,收发单元1030可以用于执行S522、S524、S525和S530。

[0181] 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。

[0182] 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)。

[0183] 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。

[0184] 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。

[0185] 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。

[0186] 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。

[0187] 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相关联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。

[0188] 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、紧密盘(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。

[0189] 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。

[0190] 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。

[0191] 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。

[0192] 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。

[0193] 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。在计算机上和执行计算机执行指令时,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。

[0194] 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(例如同轴电缆、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。

[0195] 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。

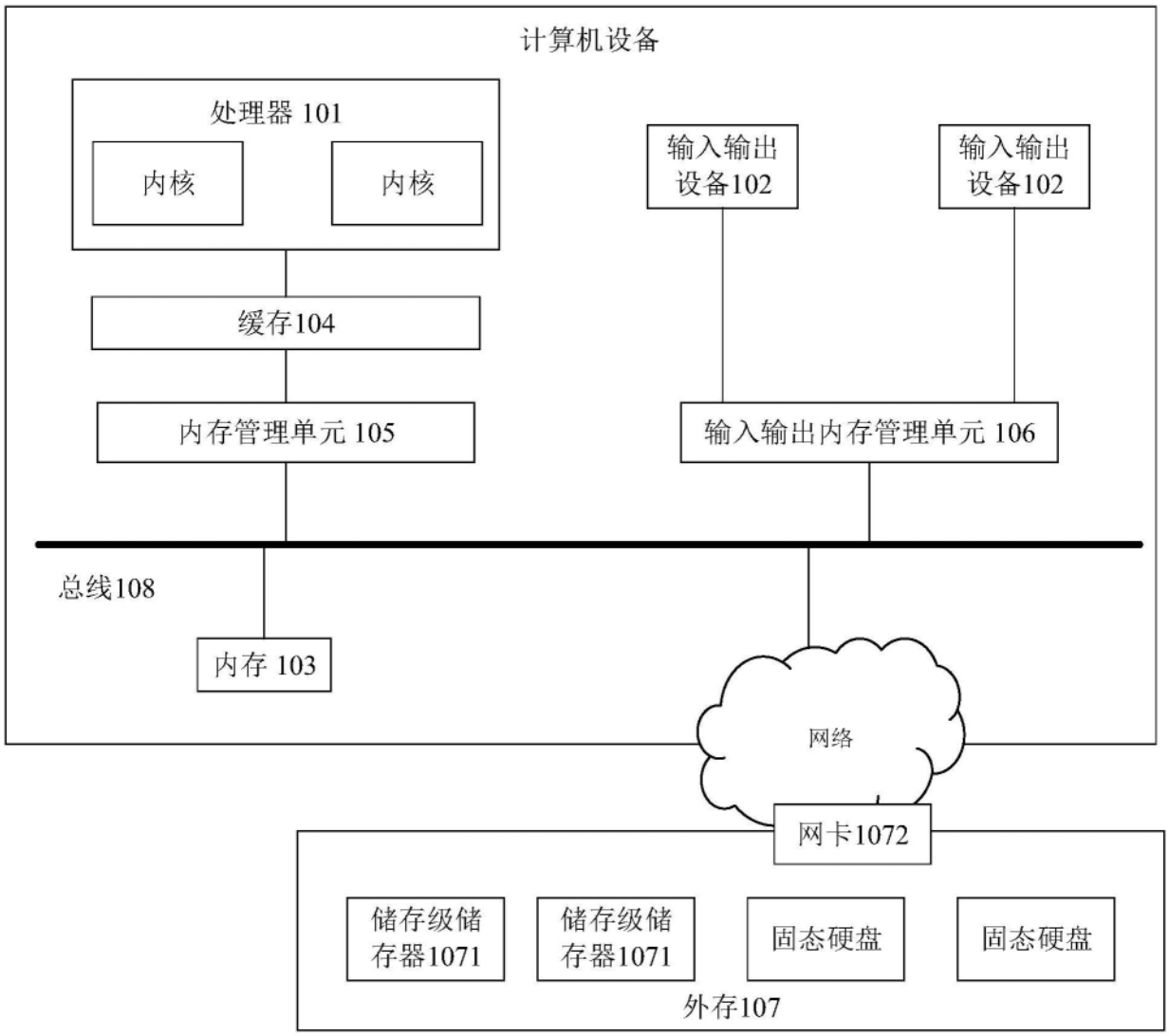

计算机设备
处理器 101
内核
内核
输入输出设备102
输入输出设备102
缓存104
内存管理单元 105
输入输出内存管理单元 106
总线108
内存 103
网络
网卡1072
储存级储存器1071
储存级储存器1071
固态硬盘
固态硬盘
外存107

图1

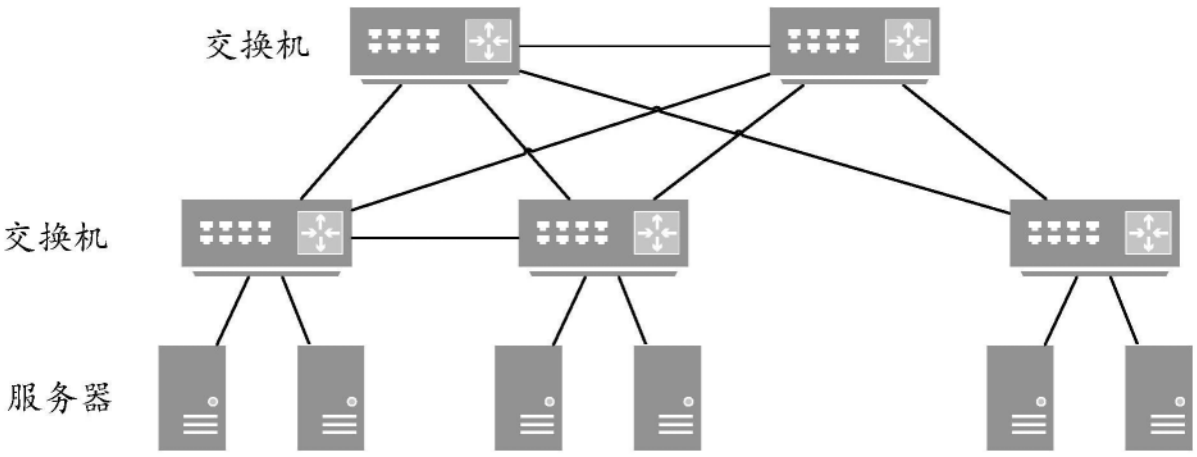

交换机
交换机
服务器

图2

UBLINK	NPI	IP	UDP	TPH	UPI	UEID	TAH	Payload	ICRC

图3

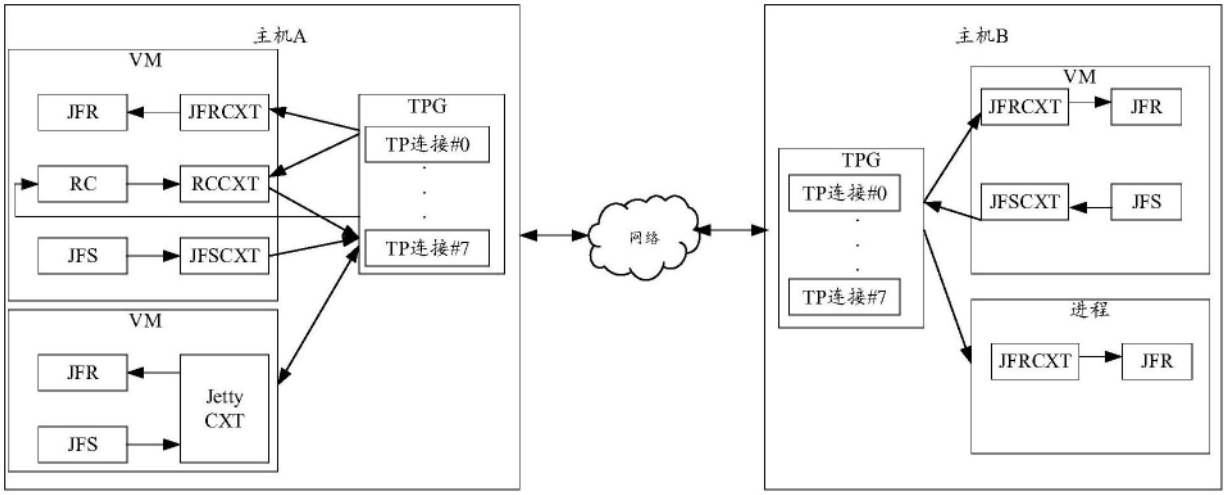


图4

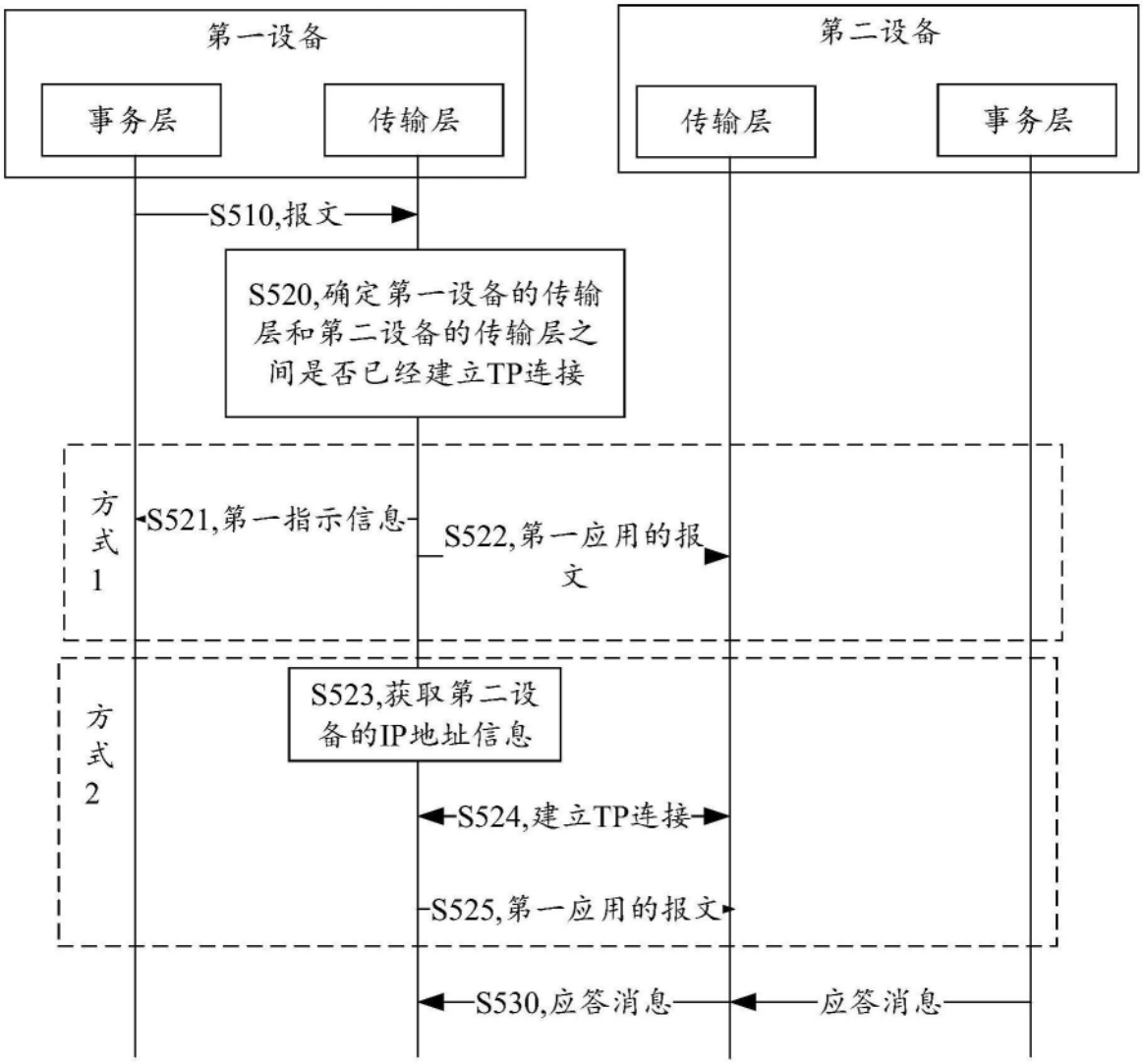

第一设备
事务层
传输层
第二设备
传输层
事务层
S510,报文
S520,确定第一设备的传输层和第二设备的传输层之间是否已经建立TP连接
方式1
S521,第一指示信息
S522,第一应用的报文
方式2
S523,获取第二设备的IP地址信息
S524,建立TP连接
S525,第一应用的报文
S530,应答消息
应答消息

图5

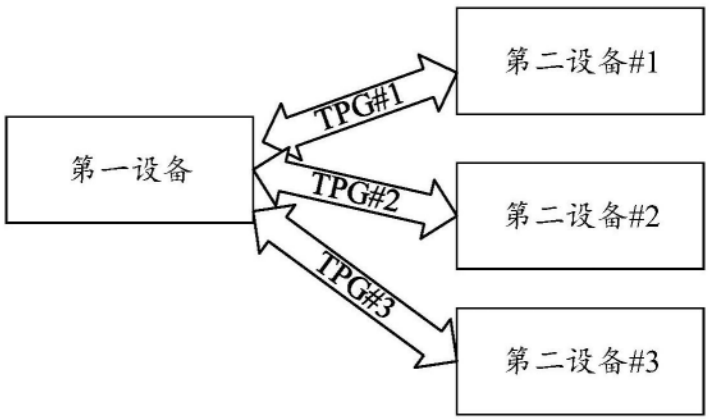

第一设备
TPG#1
TPG#2
TPG#3
第二设备#1
第二设备#2
第二设备#3

图6

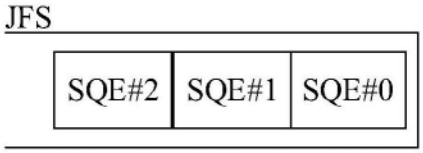

JFS
SQE#2
SQE#1
SQE#0

图7

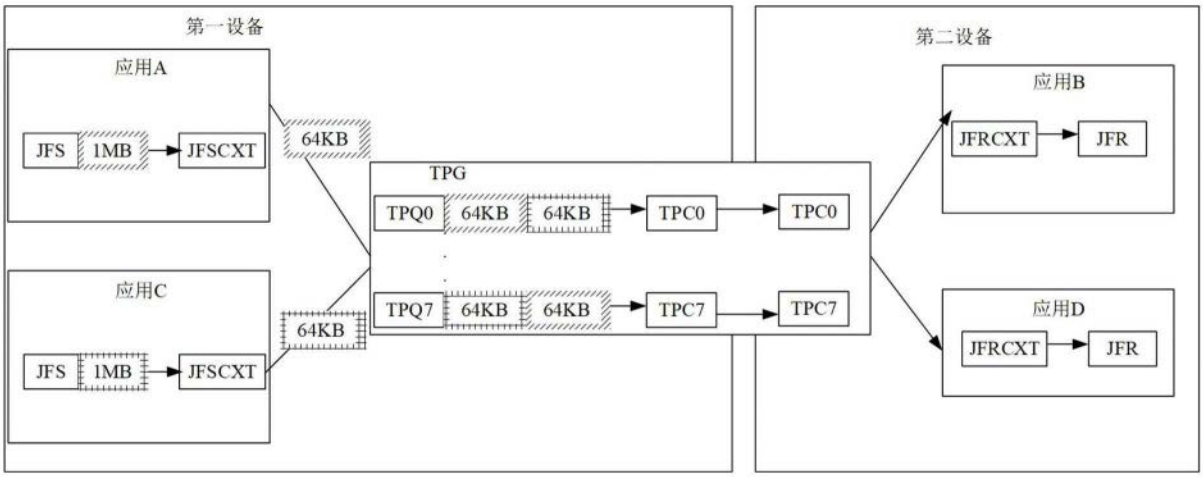

第一设备
应用A
JFS
1MB
JFSCXT
64KB
应用C
JFS
1MB
JFSCXT
64KB
TPG
TPQ0
64KB
64KB
TPC0
TPC0
TPQ7
64KB
64KB
TPC7
TPC7
第二设备
应用B
JFRCXT
JFR
应用D
JFRCXT
JFR

图8

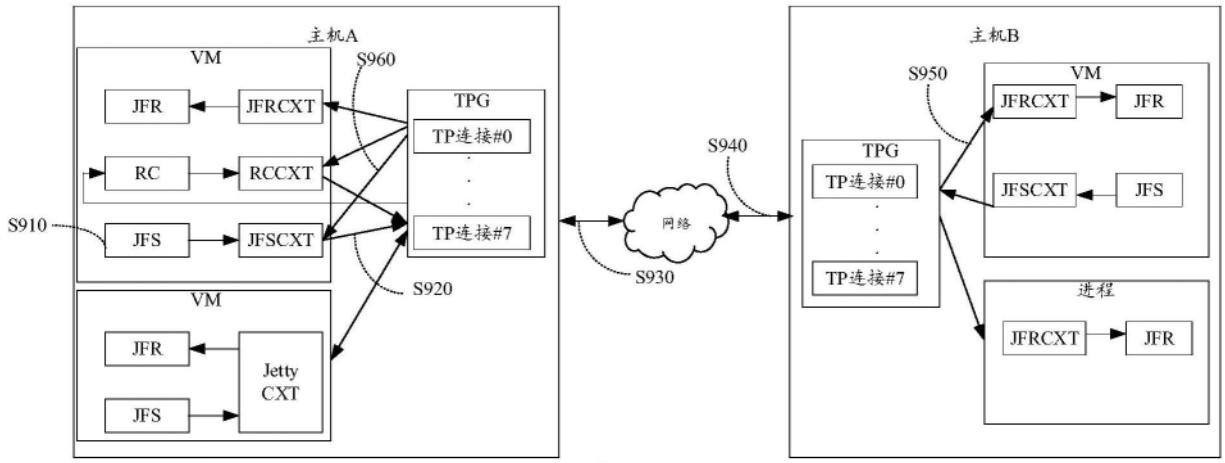

主机A
VM
JFR
JFRCXT
RC
RCCXT
JFS
JFSCXT
S910
S960
TPG
TP连接#0
TP连接#7
S920
VM
JFR
Jetty CXT
JFS
S940
网络
S930
主机B
S950
TPG
TP连接#0
TP连接#7
VM
JFRCXT
JFR
JFSCXT
JFS
进程
JFRCXT
JFR

图9

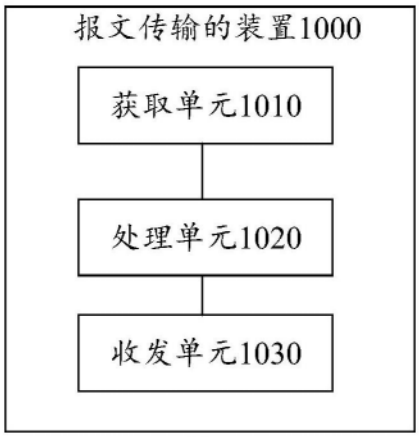

报文传输的装置1000
获取单元1010
处理单元1020
收发单元1030

图10

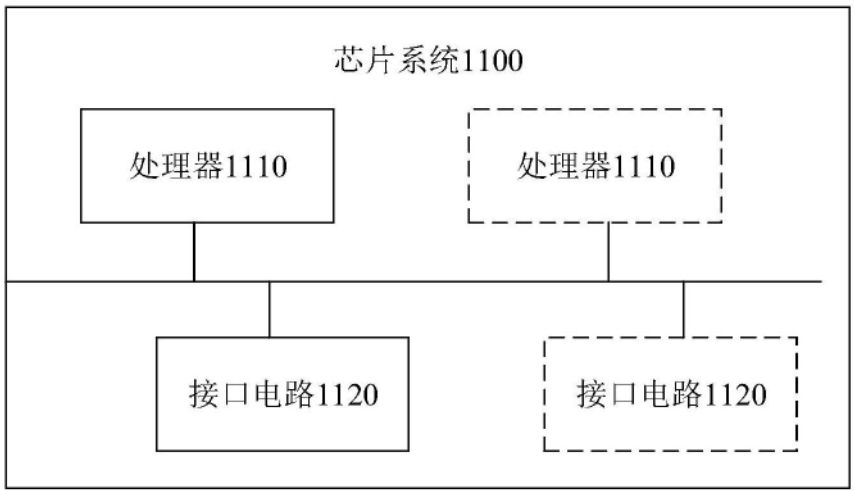

芯片系统1100
处理器1110
处理器1110
接口电路1120
接口电路1120

图11

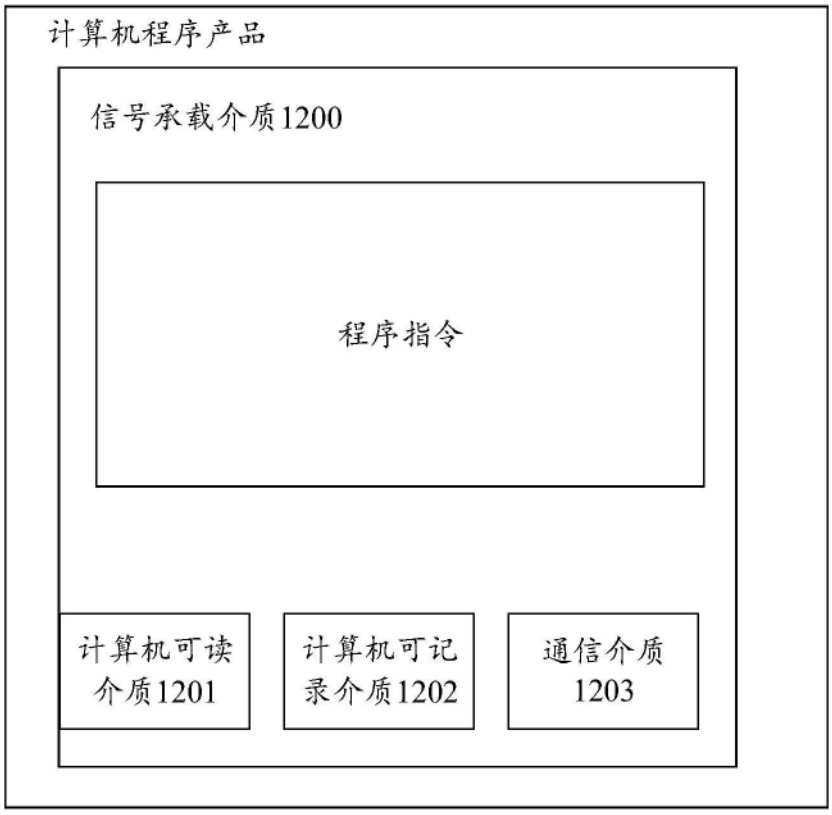

计算机程序产品
信号承载介质1200
程序指令
计算机可读介质1201
计算机可记录介质1202
通信介质1203

图12