

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2023 年 3 月 16 日 (16.03.2023)



(10) 国际公布号
WO 2023/035722 A1

- (51) 国际专利分类号:
H04L 67/025 (2022.01) **G06F 9/54** (2006.01)
- (21) 国际申请号: PCT/CN2022/099788
- (22) 国际申请日: 2022 年 6 月 20 日 (20.06.2022)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202111045146.4 2021 年 9 月 7 日 (07.09.2021) CN
- (71) 申请人: 华为技术有限公司 (HUAWEI TECHNOLOGIES CO., LTD.) [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。
- (72) 发明人: 曲会春 (QU, Huichun); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129

(CN)。邱军 (QIU, Jun); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。武雪平 (WU, Xueping); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。张锦彬 (ZHANG, Jinbin); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。吴沛 (WU, Pei); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(74) 代理人: 深圳市深佳知识产权代理事务所 (普通合伙) (SHENPAT INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市罗湖区南湖街道春风路庐山大厦 B 座 18C2、18D、18E、18E2, Guangdong 518001 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU,

(54) **Title:** RDMA DATA TRANSMISSION SYSTEM, RDMA DATA TRANSMISSION METHOD, AND NETWORK DEVICE

(54) 发明名称: RDMA 数据传输系统、RDMA 数据传输方法及网络设备

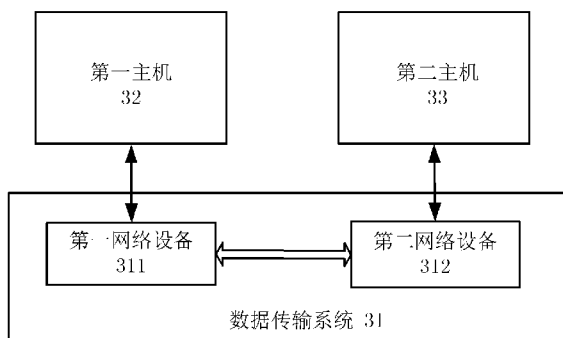


图 3

- 31 Data transmission system
32 First host
33 Second host
311 First network device
312 Second network device

(57) **Abstract:** An RDMA data transmission system, an RDMA data transmission method, and a network device, usable in, for example, a network device such as a smart network interface card (Smart NIC) or a data processing unit (DPU), and applicable to, for example, RoCEv1, RoCEv2, and other new RDMA scenarios. The RDMA data transmission system may comprise a first network device in a first host and a second network device in a second host. The first network device can create an SSQ used by a plurality of processes running by the first host, obtain an RDMA data transmission message of a first process from the SSQ, and encapsulate a first identifier corresponding to the first process into a first packet encapsulating the RDMA data transmission message. The second network device encapsulates the first identifier into a second packet encapsulating a feedback message. In this way, upon receiving the second packet, the first network device can efficiently instruct, according to the first identifier in the second packet, the first process that the RDMA data transmission message has been processed, thereby facilitating improving the running efficiency of the first process.

CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

- (84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

- 包括国际检索报告(条约第21条(3))。

(57) 摘要: 一种RDMA数据传输系统、RDMA数据传输方法及网络设备, 例如可以用于智能网卡(Smart NIC)或数据处理单元(DPU)等网络设备, 例如可以应用于RoCEv1、RoCEv2以及其他新RDMA场景。该RDMA数据传输系统可以包括第一主机中的第一网络设备和第二主机中的第二网络设备。第一网络设备可以创建供第一主机运行的多个进程使用的SSQ, 从SSQ获取第一进程的RDMA数据传输消息, 并且在封装RDMA数据传输消息的第一报文中同时封装第一进程对应的第一标识。第二网络设备在封装反馈消息的第二报文中同时封装第一标识, 这样, 第一网络设备在接收到第二报文后, 可以根据第二报文中的第一标识高效的通知第一进程已处理该RDMA数据传输消息, 进而有利于提高第一进程的运行效率。

RDMA 数据传输系统、RDMA 数据传输方法及网络设备

本申请要求于 2021 年 9 月 7 日提交中国国家知识产权局、申请号为 202111045146.4、发明名称为“RDMA 数据传输系统、RDMA 数据传输方法及网络设备”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。

技术领域

本申请涉及通信技术领域，尤其涉及一种 RDMA 数据传输系统、RDMA 数据传输方法及网络设备。

背景技术

远程直接内存访问 (remote direct memory access, RDMA) 是一种绕过远程主机 CPU 的核心访问 CPU 的内存中数据的技术，由于不经过 CPU 核心，不仅节省了大量 CPU 核心资源，同样也提高了系统吞吐量、降低了系统的网络通信延迟，尤其适合在大规模并行计算机集群中有广泛应用。

本端主机中的应用进程执行 RDMA 数据传输消息时，本端主机的网卡从主机或网卡的内存中的发送队列中将该 RDMA 数据传输消息读取到网卡的缓冲区中，之后通过网络将该 RDMA 数据传输消息发送到对端主机的网卡。之后，对端主机的网卡发送反馈信息给本端主机的网卡，本端主机的网卡根据反馈信息通知应用进程该 RDMA 数据传输消息已被处理完。

但是，当本端主机的网卡的发送队列被多个应用进程共享的时候，反馈信息无法快速将该 RDMA 数据传输消息的完成情况通知对应的应用进程。

发明内容

本申请实施例提供一种数据传输系统、数据传输方法及网络设备，用于快速将该 RDMA 数据传输消息的完成情况通知对应的应用进程。

为解决上述技术问题，本申请实施例提供以下技术方案：

第一网络设备可以创建供第一主机运行的多个进程使用的共享发送队列，第一网络设备可以从共享发送队列获取第一进程的 RDMA 数据传输消息，并且在封装 RDMA 数据传输消息的报文（称作第一报文）中同时封装第一进程对应的第一标识，通过 RDMA 方式向第二网络设备发送第一报文。第二网络设备接收到第一报文后，可以在封装反馈消息的报文（称作第二报文）中同时封装第一标识，之后通过 RDMA 方式向第一网络设备发送第二报文。这样，第一网络设备在接收到第二报文后，无需根据共享发送队列的上下文来确定第二报文中的反馈消息对应于第一进程，而是可以根据第二报文中的第一标识确定该反馈消息对应于第一进程，有利于高效的通知第一进程已处理该 RDMA 数据传输消息，进而有利于提高第一进程的运行效率。

可选的，第一报文为第一网络设备根据 RDMA 协议封装得到的，第二报文为第二网络设备根据 RDMA 协议封装得到的，该 RDMA 协议可以例如为无线带宽 (infiniband) 协议、或采用聚合以太网版本 1 的 RDMA (RDMA over Converged Ethernet version 1, RoCE1) 协议、或采用聚合以太网版本 2 的 RDMA (RDMA over Converged Ethernet version 2, RoCE2) 协议、

或互联网广域 RDMA 协议 (internet wide area RDMA protocol, iWARP) 等。

可选的, 第一网络设备可以通过该共享发送队列所绑定的传输通道向第二网络设备发送第一报文。该传输通道用于实现第一网络设备和第二网络设备之间的网络连接, 该传输通道可以例如为输入输出通道 (channel-I/O)。

5 可选的, 第二网络设备通过该传输通道接收第一报文后, 还可以通过该传输通道向第二网络设备发送第二报文。

可选的, 第一网络设备可以通过该传输通道接收来自第二网络设备的第二报文。

可选的, 所述第一报文和所述第二报文还包括第二标识, 所述第二标识用于从所述第一进程的多个 RDMA 数据传输消息中确定所述 RDMA 数据传输消息; 所述第一网络设备具体用于根据所述第一标识、所述反馈消息和所述第二标识通知所述第一进程所述 RDMA 数据传输消息的完成情况。通过在第一报文和第二报文中携带第二标识, 有利于第一网络设备高效、准确的通知第一进程 RDMA 数据传输消息的完成情况。

可选的, 所述共享发送队列具体用于保存来自所述多个进程的工作请求, 所述第一网络设备具体用于从所述共享发送队列获取来自所述第一进程的第一工作请求, 根据所述第一工作请求获取所述 RDMA 数据传输消息, 所述第一工作请求描述所述 RDMA 数据传输消息。

可选的, 共享发送队列可以设置在第一网络设备的内存中。或者, 可选的, 共享发送队列可以设置在第一主机的内存中。

和根据共享发送队列的上下文来从共享发送队列读取第一工作请求来确定第一标识相比, 本申请实施例通过第二报文获取第一标识, 有利于尽量避免从共享发送队列中读取工作请求, 有利于节约第一网络设备的缓存资源, 减少工作请求的完成过程中的延时。

可选的, 所述第一网络设备具体用于根据所述第一标识从多个完成队列中确定所述第一进程对应的第一完成队列, 根据所述反馈消息在所述第一完成队列中写入工作完成元素, 所述工作完成元素描述所述 RDMA 数据传输消息的完成情况。

第一标识可以用于确定第一完成队列, 本申请实施例可以通过第一报文和第二报文携带的第一标识来高效准确的确定第一完成队列, 有利于高效准确的通知第一进程已处理该 RDMA 数据传输消息, 进而有利于提高第一进程的运行效率。

第二方面, 本申请实施例提供一种 RDMA 数据传输方法, 包括: 第一网络设备从共享发送队列获取第一进程的 RDMA 数据传输消息, 所述第一网络设备设置在第一主机上, 所述第一进程是所述第一主机运行的使用所述共享发送队列的多个进程中的任意一个进程; 所述第一网络设备向第二网络设备发送第一报文, 所述第二网络设备设置在第二主机上, 所述第一报文包括所述 RDMA 数据传输消息和所述第一进程对应的第一标识; 所述第一网络设备接收来自所述第二网络设备的第二报文, 所述第二报文包括所述第一标识和反馈消息, 所述反馈消息指示所述 RDMA 数据传输消息的完成情况; 所述第一网络设备根据所述第二报文中的所述第一标识和所述反馈消息通知所述第一进程所述 RDMA 数据传输消息的完成情况。

35 可选的, 第一报文为第一网络设备根据 RDMA 协议封装得到的, 第二报文为第二网络设备根据 RDMA 协议封装得到的, 该 RDMA 协议可以例如为无线带宽 (infiniband) 协议、或采用聚合以太网版本 1 的 RDMA (RDMA over Converged Ethernet version 1, RoCE1) 协议、或采用聚合以太网版本 2 的 RDMA (RDMA over Converged Ethernet version 2, RoCE2) 协议、

或互联网广域 RDMA 协议 (internet wide area RDMA protocol, iWARP) 等。

可选的, 第一网络设备可以通过该共享发送队列所绑定的传输通道向第二网络设备发送第一报文。该传输通道用于实现第一网络设备和第二网络设备之间的网络连接, 该传输通道可以例如为输入输出通道 (channel-I/O)。

5 可选的, 第一网络设备可以通过该传输通道接收来自第二网络设备的第二报文。

第一网络设备通过在第一报文中携带第一标识, 有利于指示第二网络设备在第二报文中携带第一标识, 这样, 第一网络设备在接收到第二报文后, 无需根据共享发送队列的上下文来确定第二报文中的反馈消息对应于第一进程, 而是可以根据第二报文中的第一标识确定该反馈消息对应于第一进程, 有利于高效的通知第一进程已处理该 RDMA 数据传输消息, 进而有利于提高第一进程的运行效率。

10 可选的, 所述第一报文和所述第二报文还包括第二标识, 所述第二标识用于从所述第一进程的多个 RDMA 数据传输消息中确定所述 RDMA 数据传输消息; 所述第一网络设备根据所述第二报文中的所述第一标识和所述反馈消息通知所述第一进程所述 RDMA 数据传输消息的完成情况包括: 所述第一网络设备具体用于根据所述第一标识、所述反馈消息和所述第二标识通知所述第一进程所述 RDMA 数据传输消息的完成情况。

15 第一网络设备通过在第一报文中携带第二标识, 有利于指示第二网络设备在第二报文中携带第二标识, 进而有利于第一网络设备根据第二报文中的第二标识, 高效、准确的通知第一进程 RDMA 数据传输消息的完成情况。

20 可选的, 所述共享发送队列具体用于保存来自所述多个进程的工作请求; 所述第一网络设备从共享发送队列获取第一进程的 RDMA 数据传输消息包括: 所述第一网络设备从所述共享发送队列获取来自所述第一进程的第一工作请求, 所述第一工作请求描述所述 RDMA 数据传输消息; 所述第一网络设备根据所述第一工作请求获取所述 RDMA 数据传输消息。

可选的, 共享发送队列可以设置在第一网络设备的内存中。或者, 可选的, 共享发送队列可以设置在第一主机的内存中。

25 和根据共享发送队列的上下文来从共享发送队列读取第一工作请求来确定第一标识相比, 本申请实施例中, 第一网络设备通过在第一报文中携带第一标识, 有利于指示第二网络设备在第二报文中携带第一标识, 这样, 第一网络设备在接收到第二报文后, 通过第二报文获取第一标识, 有利于尽量避免从共享发送队列中读取工作请求, 有利于节约第一网络设备的缓存资源, 减少工作请求的完成过程中的延时。

30 可选的, 所述第一网络设备根据所述第二报文中的所述第一标识和所述反馈消息通知所述第一进程所述 RDMA 数据传输消息的完成情况, 包括: 所述第一网络设备根据所述第一标识从多个完成队列中确定所述第一进程对应的第一完成队列; 所述第一网络设备根据所述反馈消息在所述第一完成队列中写入工作完成元素, 所述工作完成元素用于通知所述第一进程所述 RDMA 数据传输消息的完成情况。

35 第一标识可以用于确定第一完成队列, 本申请实施例中, 第一网络设备通过在第一报文中携带第一标识, 有利于指示第二网络设备在第二报文中携带第一标识, 这样, 第一网络设备在接收到第二报文后, 可以通过第二报文携带的第一标识来高效准确的确定第一完成队列, 有利于高效准确的通知第一进程已处理该 RDMA 数据传输消息, 进而有利于提高第一进程的运

行效率。

第三方面，本申请实施例提供一种 RDMA 数据传输方法，包括：第二网络设备接收来自第一网络设备的第一报文，所述第一网络设备设置于第一主机上，所述第二网络设备设置于第二主机上，所述第一报文包括第一进程的 RDMA 数据传输消息和所述第一进程对应的第一标识，所述 RDMA 数据传输消息为所述第一网络设备从共享发送队列获取的，所述第一进程是所述第一主机运行的使用所述共享发送队列的多个进程中的任意一个进程；所述第二网络设备根据所述第一报文向所述第一网络设备发送第二报文，所述第二报文包括所述第一标识和反馈消息，所述第一标识和所述反馈消息指示所述第一网络设备通知所述第一进程所述 RDMA 数据传输消息的完成情况。

第二网络设备解析到第一报文中的第一标识后，可以在第二报文中携带第一标识，这样，第一网络设备在接收到第二报文后，无需根据共享发送队列的上下文来确定第二报文中的反馈消息对应于第一进程，而是可以根据第二报文中的第一标识确定该反馈消息对应于第一进程，有利于高效的通知第一进程已处理该 RDMA 数据传输消息，进而有利于提高第一进程的运行效率。

可选的，第一报文为第一网络设备根据 RDMA 协议封装得到的，第二报文为第二网络设备根据 RDMA 协议封装得到的，该 RDMA 协议可以例如为无线带宽（infiniband）协议、或采用聚合以太网版本 1 的 RDMA（RDMA over Converged Ethernet version 1, RoCE1）协议、或采用聚合以太网版本 2 的 RDMA（RDMA over Converged Ethernet version 2, RoCE2）协议、或互联网广域 RDMA 协议（internet wide area RDMA protocol, iWARP）等。

可选的，第一报文为第一网络设备通过该共享发送队列所绑定的传输通道向第二网络设备发送的。该传输通道用于实现第一网络设备和第二网络设备之间的网络连接，该传输通道可以例如为输入输出通道（channel-I/O）。

可选的，第二网络设备通过该传输通道接收第一报文后，还可以通过该传输通道向第二网络设备发送第二报文。

可选的，所述第一报文和所述第二报文还包括第二标识，所述第二标识用于从所述第一进程的多个 RDMA 数据传输消息中确定所述 RDMA 数据传输消息，所述第一标识、所述反馈消息和所述第二标识指示所述第一网络设备通知所述第一进程所述 RDMA 数据传输消息的完成情况。

第二网络设备解析到第一报文中的第二标识后，可以在第二报文中携带第二标识，这样，第一网络设备在接收到第二报文后，通过在第二报文中携带第二标识，有利于第一网络设备高效、准确的通知第一进程 RDMA 数据传输消息的完成情况。

第四方面，本申请实施例提供一种网络设备，包括：获取单元，用于从共享发送队列获取第一进程的 RDMA 数据传输消息，所述网络设备设置于第一主机上，所述第一进程是所述第一主机运行的使用所述共享发送队列的多个进程中的任意一个进程；发送单元，用于向第二网络设备发送第一报文，所述第二网络设备设置于第二主机上，所述第一报文包括所述 RDMA 数据传输消息和所述第一进程对应的第一标识；接收单元，用于接收来自所述第二网络设备的第二报文，所述第二报文包括所述第一标识和反馈消息，所述反馈消息指示所述 RDMA 数据传输消息的完成情况；完成单元，用于根据所述第二报文中的所述第一标识和所述反馈消息

通知所述第一进程所述 RDMA 数据传输消息的完成情况。

可选的，第一报文为第一网络设备根据 RDMA 协议封装得到的，第二报文为第二网络设备根据 RDMA 协议封装得到的，该 RDMA 协议可以例如为无线带宽 (infiniband) 协议、或采用聚合以太网版本 1 的 RDMA (RDMA over Converged Ethernet version 1, RoCE1) 协议、或采用聚合以太网版本 2 的 RDMA (RDMA over Converged Ethernet version 2, RoCE2) 协议、或互联网广域 RDMA 协议 (internet wide area RDMA protocol, iWARP) 等。

可选的，发送单元可以通过该共享发送队列所绑定的传输通道向第二网络设备发送第一报文。该传输通道用于实现第一网络设备和第二网络设备之间的网络连接，该传输通道可以例如为输入输出通道 (channel-I/O)。

可选的，接收单元可以通过该传输通道接收第二报文。

可选的，所述第一报文和所述第二报文还包括第二标识，所述第二标识用于从所述第一进程的多个 RDMA 数据传输消息中确定所述 RDMA 数据传输消息；所述完成单元具体用于根据所述第一标识、所述反馈消息和所述第二标识通知所述第一进程所述 RDMA 数据传输消息的完成情况。

可选的，所述共享发送队列具体用于保存来自所述多个进程的工作请求；所述获取单元具体用于：从所述共享发送队列获取来自所述第一进程的第一工作请求，所述第一工作请求描述所述 RDMA 数据传输消息；根据所述第一工作请求获取所述 RDMA 数据传输消息。

可选的，所述完成单元具体用于：根据所述第一标识从多个完成队列中确定所述第一进程对应的第一完成队列；根据所述反馈消息在所述第一完成队列中写入工作完成元素，所述工作完成元素用于通知所述第一进程所述 RDMA 数据传输消息的完成情况。

第五方面，本申请实施例提供一种网络设备，包括：接收单元，用于接收来自第一网络设备的第一报文，所述第一网络设备设置于第一主机上，所述网络设备设置于第二主机上，所述第一报文包括第一进程的 RDMA 数据传输消息和所述第一进程对应的第一标识，所述 RDMA 数据传输消息为所述第一网络设备从共享发送队列获取的，所述第一进程是所述第一主机运行的使用所述共享发送队列的多个进程中的任意一个进程；发送单元，用于根据所述第一报文向所述第一网络设备发送第二报文，所述第二报文包括所述第一标识和反馈消息，所述第一标识和所述反馈消息指示所述第一网络设备通知所述第一进程所述 RDMA 数据传输消息的完成情况。

可选的，第一报文为第一网络设备根据 RDMA 协议封装得到的，第二报文为第二网络设备根据 RDMA 协议封装得到的，该 RDMA 协议可以例如为无线带宽 (infiniband) 协议、或采用聚合以太网版本 1 的 RDMA (RDMA over Converged Ethernet version 1, RoCE1) 协议、或采用聚合以太网版本 2 的 RDMA (RDMA over Converged Ethernet version 2, RoCE2) 协议、或互联网广域 RDMA 协议 (internet wide area RDMA protocol, iWARP) 等。

可选的，第一报文可以是第一网络设备通过该共享发送队列所绑定的传输通道向第二网络设备发送的。该传输通道用于实现第一网络设备和第二网络设备之间的网络连接，该传输通道可以例如为输入输出通道 (channel-I/O)。

可选的，接收单元可以通过该传输通道接收第一报文。

可选的，发送单元可以通过该传输通道向第二网络设备发送第二报文。

可选的, 所述第一报文和所述第二报文还包括第二标识, 所述第二标识用于从所述第一进程的多个 RDMA 数据传输消息中确定所述 RDMA 数据传输消息, 所述第一标识、所述反馈消息和所述第二标识指示所述第一网络设备通知所述第一进程所述 RDMA 数据传输消息的完成情况。

5 第六方面, 本申请提供一种计算设备, 该计算设备包括处理器和存储器, 该处理器和该存储器耦合, 该存储器用于存储程序代码, 该处理器执行该存储器中存储的程序代码时, 能够执行第二方面或第二方面的任意一种可能的实现方式或第三方面或第三方面的任意一种可能的实现方式所描述的方法。

10 在一种可能的实现方式中, 该计算设备还可以包括通信接口, 该处理器通过该通信接口例如能够接收或发送报文等。

本申请第七方面提供了一种芯片系统, 该芯片系统包括处理器和接口电路, 该处理器通过该接口电路与存储器耦合, 该处理器用于执行该存储器中的程序代码, 以执行第二方面或第二方面的任意一种可能的实现方式或第三方面或第三方面的任意一种可能的实现方式所描述的方法。该芯片系统可以由芯片构成, 也可以包含芯片和其他分立器件。

15 本申请第八方面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了程序代码, 这些程序代码在计算机设备上运行时, 使得计算机设备执行如本申请能够执行第二方面或第二方面的任意一种可能的实现方式或第三方面或第三方面的任意一种可能的实现方式所描述的方法。

20 本申请第九方面提供一种计算机程序产品, 该计算机程序产品包含的程序代码被计算机设备执行时, 以执行第二方面或第二方面的任意一种可能的实现方式或第三方面或第三方面的任意一种可能的实现方式所描述的方法。

由于本申请提供的各装置可用于执行前述对应的方法, 因此本申请各装置所能获得到的技术效果可参考前述对应的方法所获得的技术效果, 此处不再赘述。

25 附图说明

图 1A、图 1B 和图 1C 分别示出了本申请实施例一种可能的应用场景;

图 2A 和图 2B 分别示出了计算机系统一种可能的结构;

图 3 示出了本申请实施例计算机系统另一种可能的结构;

图 4 示出了第一工作请求一种可能的结构;

30 图 5A 和图 5B 分别示出了第一网络设备对 RDMA 数据传输消息的处理过程;

图 6A 示出了发送操作对应的第一报文和第二报文的传输过程;

图 6B 示出了写操作对应的第一报文和第二报文的传输过程;

图 6C 示出了原子操作对应的第一报文和第二报文的传输过程;

图 6D 示出了读操作对应的第一报文和第二报文的传输过程;

35 图 7 示出了本申请实施例 RDMA 数据传输方法一种可能的流程;

图 8 示出了本申请实施例网络设备一种可能的结构;

图 9 示出了本申请实施例网络设备一种可能的结构。

具体实施方式

下面，首先对本申请实施例的应用场景进行举例介绍。

图 1A 示出了本申请实施例一种可能的应用场景。图 1A 所示的计算机系统至少包括节点 #1 和节点#2。节点#1 或节点#2 可以为连接到通信网络的一个或多个链路并且形成通信网络内数据传输消息（简称消息）的起源和/或目的地的任意设备。节点#1 或节点#2 中可以运行一个或多个应用进程（简称进程）。图 1A 以节点#1 运行进程#1 和进程#2 为例，以节点#2 运行进程#3 为例。可以理解的是，节点#1 或节点#2 可以运行更多或更少的进程。以节点#1 为例，节点#1 发出的消息可以来自进程#1 或进程#2，节点#1 接收的消息可以去往进程#1 或进程#2。

节点#1 和节点#2 可以通过数据传输系统进行通信连接。图 1A 所示的数据传输系统可以至少包括节点#1 中的网络设备#1 和节点#2 中的网络设备#2。可选的，网络设备#1 或网络设备#2 可以为网络适配器（network interface card, NIC）或称网卡。可选的，图 1A 所示的数据传输系统可以为远程直接内存访问（remote direct memory access, RDMA）系统，该数据传输系统中传输的消息可以为 RDMA 数据传输消息。

图 1A 所示的计算机系统中，数据传输系统可以基于节点运行的进程来创建数据传输通道（channel）。假设进程#1 和进程#2 需要分别向进程#3 发送消息，网络设备#1 和网络设备#2 需要分别为进程#1 和进程#2 创建数据传输通道，分别为如图 1A 所示的通道 p1-p3 和通道 p2-p3。进程#1 的消息通过通道 p1-p3 发送给进程#3，进程#2 的消息通过通道 p2-p3 发送给进程#3。

随着计算机系统中节点数目的增加以及节点中进程数目的增加，图 1A 所示的数据传输系统需要为大量进程传输消息，相应的需要创建大量数据传输通道。假设图 1A 所示的计算机系统包括 n 个节点，每个节点运行 m 个进程，那么对于任意一个节点，数据传输系统需要创建 $m*n$ 个数据传输通道。由于每个数据传输通道需要在相应节点中创建至少一个队列对（queue pair, QP），而大量 QP 将占用节点过多的存储资源，因此，数据传输系统基于节点运行的进程来创建数据传输通道将影响数据传输系统的可伸缩性。

图 1B 示出了本申请实施例另一种可能的应用场景。图 1B 所示的计算机系统可以参考图 1A 所示的计算机系统理解，但是和图 1A 不同的是，在图 1B 所示的计算机系统中，数据传输系统是基于计算机系统节点来创建数据传输通道。仍然假设进程#1 和进程#2 需要分别向进程#3 发送消息，在图 1B 所示的数据传输系统中，网络设备#1 和网络设备#2 为节点#1 和节点#2 创建一个数据传输通道，例如为图 1B 所示的通道 n1-n2。进程#1 的消息和进程#2 的消息均通过通道 n1-n2 发送给进程#3。

假设图 1B 所示的计算机系统包括 n 个节点，每个节点运行 m 个进程，那么对于任意一个节点，数据传输系统最多需要创建 $n-1$ 个数据传输通道。和图 1A 所示的数据传输系统相比，通过减少数据传输通道的数目，有利于减少 QP 占用的存储资源，可见，数据传输系统基于节点来创建数据传输通道将减少对数据传输系统的可伸缩性的影响。

以 n 为 4 为例，图 1C 示出了本申请实施例另一种可能的应用场景。和图 1B 相比，图 1C 所示的计算机系统除了包括节点#1 和节点#2，还包括节点#3 和节点#4。图 1C 所示的数据传输系统基于节点创建数据传输通道，分别例如为图 1C 所示的通道 n1-n2、通道 n1-n3、通道

n1-n4、通道 n2-n3、通道 n2-n4 和通道 n3-n4。对于图 1C 所示 4 个节点中的任意一个节点，数据传输系统为其创建 3 条数据传输通道。

下面，以节点#1 和节点#2 为例，进一步介绍图 1B 或图 1C 所示的计算机系统。

5 参考图 2A，计算机系统可以至少包括节点#1 和节点#2，节点#1 包括主机 11、支持 RMDA 的网络适配器 (network interface card that supports RDMA, RNIC) 12 和内存 13，节点 #2 包括主机 21 和 RNIC 22。RNIC 22 可以与主机 21 通信相连，RNIC 12 分别与主机 11 和内存 13 通信相连。

图 2A 中的 RNIC 12 可以视为图 1B 或图 1C 中的网络设备#1，或者，图 2A 中的 RNIC 12 和内存 13 可以视为图 1B 或图 1C 中的网络设备#1。图 2A 中的 RNIC 22 可以视为图 1B 或图 1C 中的网络设备#2。图 2A 以图 1B 或图 1C 所示的网络设备#1 包括 RNIC 12、网络设备包括 RNIC 22 为例，本申请实施例不限定图 1B 或图 1C 所示的网络设备#1 或网络设备#2 的具体类型。

其中，主机 11 或主机 21 可以包括：处理器、通信接口以及存储器，处理器、通信接口 15 以及存储器通过内部总线相互连接。其中，处理器可以由一个或者多个通用处理器构成，例如中央处理器 (central processing unit, CPU)，或者 CPU 和硬件芯片的组合。主机 11 或主机 21 的存储器可以存储系统应用和/或应用进程的代码，处理器执行这些代码可以实现 CPU 核心 113 和/或进程的功能的功能。

其中，RNIC 12 可以包括处理器 122 和缓存 121，RNIC 22 可以包括处理器 222 和缓存 20 221。其中，处理器 122 或处理器 222 可以由一个或者多个通用处理器构成，例如 CPU，或者 CPU 和硬件芯片的组合。处理器 122 和缓存 121 之间可以通过总线相连 (图 2A 未示出)，也可以通过其他方式相连。处理器 222 和缓存 221 之间可以通过总线相连 (图 2A 未示出)，也可以通过其他方式相连。

上述硬件芯片可以是专用集成电路 (application-specific integrated circuit, 25 ASIC)、可编程逻辑器件 (programmable logic device, PLD) 或其组合。上述 PLD 可以是复杂可编程逻辑器件 (complex programmable logic device, CPLD)、现场可编程逻辑门阵列 (field-programmable gate array, FPGA)、通用阵列逻辑 (generic array logic, GAL) 或其任意组合。

其中，存储器或内存 13 可以包括易失性存储器 (volatile memory)，例如随机存取存 30 储器 (random access memory, RAM)；存储器或内存 13 也可以包括非易失性存储器 (non-volatile memory)，例如只读存储器 (read-only memory, ROM)、快闪存储器 (flash memory)、硬盘 (hard disk drive, HDD) 或固态硬盘 (solid-state drive, SSD) 或为双倍速率同步动态随机存储器 (Double Data Rate Synchronous Dynamic Random Access Memory, DDR SDRAM)；存储器或内存 13 还可以包括上述种类的组合。其中，DDR SDRAM 可以简称为 DDR。缓存 121 或缓存 221 可以为一级或多级缓存，例如，可以采用寄存器和/或 35 静态随机存取存储器 (static random access memory, SRAM) 实现。

在一种可能的实现方式中，RNIC 12 和内存 13 可以集成在同一芯片上，或者说，该内存 13 可以为 RNIC 12 对应的网络设备#1 中的内存。以内存 13 为 DDR 为例，可选的，该 DDR 可

以支持多通道技术, RNIC 12 可以通过多个通道访存内存 13。

或者, 在一种可能的实现方式中, 该内存 13 可以为主机 11 中的内存。RNIC 12 与主机 11 之间可以通过输入输出 (input output, IO) 接口相连, 该 IO 接口可以包括但不限于周边装置互连高速 (peripheral component interconnect express, PCIe) 等 IO 结构 (fabric) 接口。若内存 13 为主机 11 中的内存, 那么 RNIC 12 可以通过与主机 11 之间的 IO 接口访存内存 13。

RNIC 12 和 RNIC 22 之间创建有数据传输通道, 该数据传输通道可以理解为图 1B 或图 1C 或图 2A 所示的通道 n1-n2, RNIC 12 和 RNIC 22 通过通道 n1-n2 传输节点 1 和节点 2 之间的消息。RNIC 12 可以在内存 13 中创建绑定通道 n1-n2 的至少一个共享发送队列 (Shard Send Queue, SSQ), SSQ 可以用于处理来自节点#1 中多个进程的、待传输至节点#2 的消息。图 2A 示例性示出了该至少一个 SSQ 中的 SSQ#1。可选的, 若该至少一个 SSQ 还包括除 SSQ#1 以外的其他 SSQ (例如 SSQ#2), SSQ#1 和 SSQ#2 可以对应于不同的优先级, RNIC 12 可以根据待传输至节点 2 的消息的优先级使用相应的 SSQ。

主机 11 和/或主机 21 可以运行一个或多个进程。图 2A 示意性示出主机 11 运行的两个进程 (进程#1 和进程#2) 和主机 21 运行的两个进程 (进程#3 和进程#4)。RNIC 12 待传输至节点 2 的消息可以来自进程#1 和/或进程#2。

假设进程#1 和进程#2 需要向主机 21 发送消息, 进程#1 和进程#2 可以分别向 RNIC 12 提交相应的工作请求 (work request, WR)。为了便于区分, 本申请实施例将进程#1 待发送的一个消息称作 M1, 将进程#2 待发送的一个消息称作 M2, 将 M1 对应的 WR 称作 WR1, 将 M2 对应的 WR 称作 WR2。作为举例, 进程#1 和进程#2 可以分别调用主机 11 中的程序接口和 RNIC 12 的驱动程序向 RNIC 12 提交 WR1 和 WR2。在一种可能的实现方式中, M1 和 M2 可以为 RDMA 消息。

由于 M1 和 M2 均用于发送至节点 2, 因此, RNIC 12 接收到 WR1 和 WR2 后, 可以分别将 WR1 和 WR2 写入 SSQ#1 中。本申请实施例将写入 SSQ#1 中的 WR1 和 WR2 分别称作工作队列元素 (work queue element, WQE) 1 和 WQE2。WQE1 和 WQE2 分别描述了 M1 和 M2, 例如, WQE1 包括 M1 在内存 111 中的存储地址 1, WQE2 包括 M2 在内存 13 中的存储地址 2。

在一种可能的实现方式中, 内存 13 中还包括 SSQ#1 的上下文 (简称上下文 S1), 上下文 S1 记录了 SSQ#1 的使用信息。示例性的, 上下文 S1 可以包括 SSQ#1 的写入索引 (简称 SSQ#1 的 PI) 和读取索引 (简称 SSQ#1 的 CI)。RNIC 12 可以根据 SSQ#1 的 PI 的取值确定 WQE 在 SSQ#1 中的写入位置, 根据 SSQ#1 的 CI 的取值确定 WQE 在 SSQ#1 中的读取位置。

图 2A 示例性示出了 SSQ#1 中的三个放置 WQE 的存储位置, SSQ#1 还可以包括更多或更少的 WQE, 并且, 为了便于理解, 图 2A 将 SSQ#1 中的用于保存 WQE 的三个存储位置绘制为从上至下依次排列的三个矩形。本申请实施例将图 2A 所示的 SSQ#1 中用于保存 WQE 的三个存储位置从上至下分别称作存储位置 T、存储位置 M 和存储位置 B。本申请实施例不限定 SSQ#1 中各 WQE 的存储方式, 例如, SSQ#1 可以对应于内存 13 中的一块连续存储空间, 或者, 可以对应于内存 13 中离散的存储空间。

作为举例, RNIC 12 接收到 WR1 后, 可以根据上下文 S1 中 SSQ#1 的 PI 指向 SSQ#1 的存

储位置 T，之后可以将 WQE1 写入存储位置 T，并更新 SSQ#1 的 PI 的取值，使其指向 SSQ#1 的存储位置 M。RNIC 12 接收到 WR2 后，可以根据上下文 S1 中 SSQ#1 的 PI 的取值将 WQE2 写入存储位置 M，并更新 SSQ#1 的 PI 的取值，使其指向 SSQ#1 的存储位置 B。

图 2A 示出了 RNIC 12 在 SSQ#1 中写入 WQE1 和 WQE2 后 CI 和 PI 在内存 13 中指向的存储位置。在一种可能的实现方式中，SSQ#1 用于顺序处理写入 SSQ#1 中的 WQE。假设 WQE1 先于 WQE2 写入 SSQ#1，那么 SSQ#1 的 CI 的取值可以先指向 WQE1，待 RNIC 12 处理 WQE1 后，将 SSQ#1 的 CI 的取值更新为指向 WQE2（参考图 2B），以使得 RNIC 12 继续处理 WQE2。

下面结合图 2A 中代表数据传输过程的虚线及其序号，介绍本申请实施例提供的 M1 发送至主机 21 的过程。

步骤 1. RNIC 12 将 SSQ#1 中的 WQE1 读取到缓存 121 中；

示例性的，RNIC 12 可以根据上下文 S1 确定 SSQ#1 的 CI 的取值，根据 CI 的取值确定 CI 指向 SSQ#1 中的存储位置 T，进而从存储位置 T 取出 WQE（即 WQE1）。

步骤 2. RNIC 12 访问内存 111 中 WQE1 所指示的存储位置 1；

WQE1 可以包括 M1 在内存 111 中的存储地址 1，RNIC 12 将 WQE1 读取到缓存 121 后，可以从 WQE1 中解析到存储地址 1，之后可以访问内存 111 中的存储位置 1。

步骤 3. RNIC 12 将存储位置 1 中的 M1 读取到缓存 121；

RNIC 12 可以将存储位置 1 中的数据（即 M1）读取到缓存 121。

步骤 4. RNIC 12 将 M1 封装为报文 1，并将报文 1 通过通道 n1-n2 发送到 RNIC 22；

可选的，RNIC 12 可以根据 RDMA 协议将 M1 封装为报文 1，并将报文 1 通过 SSQ#1 所绑定的通道 n1-n2 发送到 RNIC 22。示例性的，该 RDMA 协议可以例如为无线带宽（infiniband）协议、或采用聚合以太网版本 1 的 RDMA（RDMA over Converged Ethernet version 1，RoCE1）协议、或采用聚合以太网版本 2 的 RDMA（RDMA over Converged Ethernet version 2，RoCE2）协议、或互联网广域 RDMA 协议（internet wide area RDMA protocol，iWARP）等。

RNIC 22 接收到报文 1 后，可以对报文 1 进行解封装，得到 M1。

步骤 5. RNIC 22 将 M1 存储在主机 21 中；

假设 M1 为待存储在主机 21 中的数据，那么 RNIC 22 可以将 M1 存储在主机 21 中。进一步的，假设 M1 用于写入进程#3 的内存 211，作为举例，RNIC 22 对报文 1 进行解封装后还可以得到 M1 在内存 211 中的写入位置，RNIC 22 可以根据该写入位置将 M1 写入内存 211 中。

后文将结合 RDMA 的发送（send）模型、写（write）模型、读（read）模型和原子（atomic）模型具体介绍步骤 4 和步骤 5，此处暂不展开介绍。

图 2B 所示的计算机系统通过代表数据传输过程的虚线及其序号，描述了本申请实施例提供的数据传输系统向进程#1 通知 M1 的完成情况的过程。图 2B 中的结构可以解释为图 2A 所示的计算机系统结构中的相应结构，此处不再赘述。

参考图 2B，RNIC 12 可以在内存 13 中分别为进程#1 和进程#2 创建完成队列（complete queue，CQ）。本申请实施例将进程#1 对应的 CQ 称作 CQ#1，将进程#2 对应的 CQ 称作 CQ#2。CQ#1 用于向进程#1 反馈进程#1 向 RNIC 12 提交的消息（例如 M1）的完成情况，CQ#2 用于向

进程#2 反馈进程#2 向 RNIC 12 提交的消息（例如 M2）的完成情况。

下面结合图 2B 介绍本申请实施例提供的 RNIC 12 和 RNIC 22 向进程#1 通知 M1 的完成过程。

步骤 6. RNIC 22 通过通道 n1-n2 向 RNIC 12 发送携带 R1 的报文 2；

5 RNIC 22 接收到报文 1 后，可以通过通道 n1-n2 向 RNIC 12 发送报文 2，报文 2 携带反馈消息（简称 R1），R1 描述消息的完成情况。例如，消息的完成情况可以为 RNIC 22 成功接收消息，或 RNIC 22 成功将消息写入主机 21，或 RNIC 22 未接收到消息，或消息写入失败等。RNIC 12 接收到报文 2 后，可以对报文 2 进行解封装，得到 R1。RNIC 12 可以确定 R1 所描述的消息的完成情况。

10 可选的，R1 为报文 2 中的某个字段（称作字段 R），字段 R 的不同取值对应于消息的不同完成情况。例如，字段 R 的取值为 0 时，RNIC 12 可以确定消息的完成情况为成功，具体例如为 RNIC 22 成功接收消息，字段 R 的取值为 1 时，RNIC 12 可以确定消息的完成情况为失败，具体例如为 RNIC 22 未接收到消息。

步骤 7. RNIC 12 将 SSQ#1 中的 WQE1 读取到缓存 121 中，以获得 CQ#1 的标识；

15 RNIC 12 从通道 n1-n2 获取报文 2 后，虽然可以确定报文 2 中的 R1 对应于 SSQ#1 中的 WQE，但是由于 SSQ#1 对应于主机 11 运行的多个进程，因此 RNIC 12 无法确定 R1 对应的完成情况对应于哪个进程，进而无法确认该 R1 应该交由哪个 CQ 来处理。为此，RNIC 12 在确认绑定通道 n1-n2 的 SSQ 为 SSQ#1 之后，可以根据 SSQ#1 的上下文 S1 确定 R1 对应于 SSQ#1 中的 WQE1。本申请实施例中，可以在 WQE1 中携带 CQ#1 的标识。相应的，RNIC 12 将 WQE1 从
20 SSQ#1 中读取到缓存 121 中后，可以解析 WQE1 以获取 WQE1CQ#1 的标识，从而确定需要将 R1 对应的完成情况交由 CQ#1 来处理。

步骤 8. RNIC 12 根据 CQ#1 的标识和 R1 向 CQ#1 中写入 CQE1；

步骤 9. RNIC 12 处理 CQ#1 中的 CQE，当处理到 CQE1 时，向进程#1 通知 M1 的完成情况。

下面对步骤 8 和步骤 9 进行介绍。

25 RNIC 12 获取 CQ#1 的标识后，可以根据 R1 向 CQ#1 中写入完成队列元素（CQ element，CQE）1，CQE1 用于确定 M1 的完成情况。RNIC 12 在处理 CQ#1 中的 CQE 的过程中，当处理到 CQE1 时，可以向进程#1 通知 M1 的完成情况。

可选的，RNIC 12 向进程#1 通知 M1 的完成情况可以指，进程#1 调用程序接口和 RNIC 12 的驱动程序对 CQ#1 中的 CQE 进行检索，当检索到 CQE1 时，进程#1 能够获取 M1 的完成情况。

30 可选的，CQE 通过携带的部分字段（例如称作错误码）来指示相应消息的完成情况。例如，进程#1 解析 CQE1，确定 CQE1 中错误码的取值为 0 时，进程#1 可以确定 M1 的完成情况为成功，具体例如为 M1 对应的数据传输任务已完成。例如，进程#1 解析 CQE1，确定 CQE1 中错误码的取值为 1 时，进程#1 可以确定 M1 的完成情况为失败，具体例如为 M1 对应的数据传输任务未完成。

35 可选的，RNIC 12 可以根据 R1 确定 CQE1 中错误码的取值。可选的，CQE1 中错误码的取值所描述的完成情况与 R1 所描述的完成情况一致。例如，若 R1 描述消息的完成情况为成功，那么 CQE1 中错误码的取值可以为 0，若 R1 描述消息的完成情况为失败，那么 CQE1 中错误码的取值可以为 1。或者，可选的，CQE1 中错误码的取值所描述的完成情况与 R1 描述的完成情

况不一致。例如，若 R1 描述消息的完成情况为成功，但是 RNIC 12 因故障可能未正确的封装报文 1（M1 错误或目的地址错误等），那么 CQE1 中错误码的取值可以为 1。

下面举例介绍 RNIC 12 在 CQ#1 中写入 CQE 和读取 CQE 的过程。

在一种可能的实现方式中，内存 13 中还包括 CQ#1 的上下文（简称上下文 C1），上下文 C1 记录了 CQ#1 的使用信息，示例性的，上下文 C1 可以包括 CQ#1 的写入索引（简称 CQ#1 的 PI）和读取索引（简称 CQ#1 的 CI）。RNIC 12 可以根据 PI 的取值确定 CQE 在 CQ#1 中的写入位置，根据 CI 的取值确定 CQE 在 CQ#1 中的读取位置。

图 2B 仅示例性示出了 CQ#1 中的三个放置 CQE 的存储位置，CQ#1 还可以包括更多或更少的 CQE，并且，为了便于理解，图 2B 将 CQ#1 中的用于保存 CQE 的三个存储位置绘制为从上至下依次排列的三个矩形。本申请实施例将图 2B 所示的 CQ#1 中用于保存 CQE 的三个存储位置从上至下分别称作存储位置 T、存储位置 M 和存储位置 B。本申请实施例不限定 CQ#1 中各 CQE 的存储方式，例如，CQ#1 可以对应于内存 13 中的一块连续存储空间，或者，可以对应于内存 13 中离散的存储空间。

参考图 2B，作为举例，RNIC 12 生成 CQE4 后，可以根据上下文 C1 确定 CQ#1 的 PI 指向 CQ#1 的存储位置 T，之后在存储位置 T 写入 CQE4，并更新 CQ#1 的 PI 的取值，以使其指向 CQ#1 的存储位置 M。之后，RNIC 12 生成 CQE3 后，可以根据上下文 C1 中 CQ#1 的 PI 的取值在存储位置 M 写入 CQE3，并更新 CQ#1 的 PI 的取值，使其指向存储位置 B。

图 2B 示出了 RNIC 12 在 CQ#1 中写入 CQE4 和 CQE2 后 CI 和 PI 在内存 13 中指向的存储位置。在一种可能的实现方式中，CQ#1 用于顺序处理写入 CQ#1 中的 CQE。假设 CQE4 先于 CQE3 写入 CQ#1，那么 CQ#1 的 CI 的取值可以先指向 CQE4，RNIC 12 处理 CQE4，向进程#1 通知 CQE4 对应的消息（称作 M4）的完成情况，之后，将 CQ#1 的 CI 的取值更新为指向 CQE3，以使得 RNIC 12 继续处理 CQE3，向进程#1 通知 CQE3 对应的消息（称作 M3）的完成情况。

参考图 2B，假设 RNIC 12 生成 CQE1 后，根据 CQ#1 的标识从上下文 C1 的取值向 CQ#1 的存储位置 B 写入 CQE1。待 RNIC 12 依次处理 CQE4 和 CQE3 后，将 CQ#1 的 CI 的取值更新为指向 CQ#1 的存储位置 B，以使得 RNIC 12 继续处理 CQ#1 的存储位置 B 中的 CQE3，向进程#1 通知 M1 的完成情况。

以上结合图 2A 和图 2B 介绍了步骤 1 至步骤 9 对应的数据传输流程。在上述步骤 6 后，由于通道 n1-n2 所绑定的 SSQ 保存主机 11 中多个进程的 WQE，因此，RNIC 12 无法确定通道 n1-n2 返回的报文 2 对应于主机 11 中的哪个进程，也无法确定需要通过内存 13 中的哪个 CQ 来处理报文 2 中的 R1。为此，RNIC 12 需要通过步骤 7 来获取相应 CQ 的标识（即 CQ#1 的标识），具体的，先确认绑定通道 n1-n2 的 SSQ 为 SSQ#1，之后根据 SSQ#1 的上下文 S1 确定 R1 对应于 SSQ#1 中的 WQE1，将 WQE1 从 SSQ#1 中读取到缓存 121 中，解析 WQE1 以获取 CQ#1 的标识。由于步骤 7 包括 RNIC 12 从内存 13 中读取 WQE1 的过程，因此步骤 7 增加了 RNIC 12 对内存 13 的读时延，降低了 RNIC 12 向进程#1 通知 M1 的完成情况的效率。并且，由于缓存 121 的空间有限，将 WQE1 从内存 13 中读取到缓存 121 中，可能导致预取到缓存 121 中的待发送的 WQE 被老化删除，导致发送业务的延时。

在分析发现了上述延时以及导致延时的原因后，本申请实施例对步骤 1 至步骤 9 中的部

分内容进行优化。下面具体介绍优化方案。

1、对步骤 4 进行优化。对步骤 4 优化前，报文 1 包括 M1，对步骤 4 优化后，参考图 2A 步骤 4 的括号中的内容，在 M1 的基础上，在报文 1 中增加标识信息。可选的，该标识信息可以为 CQ#1 的标识和/或进程#1 的标识。

5 2、对步骤 6 进行优化。对步骤 6 优化前，报文 2 包括 R1，对步骤 6 优化后，参考图 2B 中步骤 6 的括号中的内容，在报文 2 中增加报文 1 所携带的全部或部分标识信息。

3、不执行步骤 7。由于报文 2 携带 CQ#1 的标识，因此，参考 2B 中步骤 7 对应的虚线上的“×”，RNIC 12 可以无需执行步骤 7 来获取 CQ#1 的标识。

10 基于上述步骤 1 至步骤 9 以及优化内容的构思，图 3 示出了本申请计算机系统另一种可能的实施例。

参考 3，计算机系统 3 可以包括数据传输系统 31、第一主机 32 和第二主机 33。数据传输系统 31 包括第一网络设备 311 和第二网络设备 312，第一网络设备 311 设置于第一主机 32 上，第二网络设备 312 设置于第二主机上。该数据传输系统 31 可以为 RDMA 数据传输系统，第一网络设备 311 和第二网络设备 312 之间可以传输 RDMA 数据传输消息。

15 可选的，计算机系统 3 可以解释为图 1B 或图 1C 或图 2A 或图 2B 所示的计算机系统。可选的，数据传输系统 31 可以解释为图 1B 或图 1C 或图 2A 或图 2B 所示的数据传输系统。可选的，第一主机 32 可以解释为包括图 2A 或图 2B 所示的主机 11。可选的，第二主机 33 可以解释为包括图 2A 或图 2B 所示的主机 21。可选的，第一网络设备 311 可以解释为包括图 20 1B 或图 1C 所示的网络设备#1，或者解释为包括图 2A 或图 2B 所示的 RNIC 12，或者解释为包括图 2A 或图 2B 所示的 RNIC 12 和内存 13。可选的，第二网络设备 312 可以解释为包括图 1B 或图 1C 所示的网络设备#2，或者解释为包括图 2A 或图 2B 所示的 RNIC 22。

其中，第一主机 32 可以运行一个或多个进程，第一网络设备 311 可以创建供这一个或多个进程中的多个进程使用的 SSQ。参考图 2A 或图 2B 所示的计算机系统，可选的，第一主机 25 32 运行的一个或多个进程可以例如包括进程#1 和进程#2，第一网络设备 311 创建的 SSQ 可以例如为内存 13 中的 SSQ#1。

可选的，SSQ 可以设置在第一网络设备的内存中，或者可以设置在第一主机的内存中。示例性的，SSQ 可以设置在图 2A 或图 2B 所示的内存 13 中。

30 为了便于描述，本申请实施例将使用该 SSQ 的多个进程中的一个进程称作第一进程，例如，第一进程可以解释为图 2A 或图 2B 对应的实施例中的进程#1。下面以数据传输系统 31 处理第一进程的数据传输消息的过程为例，介绍数据传输系统 31 中第一网络设备 311 和第二网络设备 312 的功能。

第一网络设备 311 可以用于从 SSQ 获取第一进程的数据传输消息。可选的，该数据传输 35 消息可以解释为图 2A 或图 2B 对应的实施例中的 M1。可选的，该 SSQ 具体用于保存来自多个进程的工作请求，第一网络设备 311 从 SSQ 获取数据传输消息具体可以指，第一网络设备 311 用于从 SSQ 获取来自第一进程的第一工作请求，之后根据第一工作请求获取数据传输消息。其中，第一工作请求描述数据传输消息。

可选的，SSQ 所保存的来自多个进程的工作请求可以例如分别参考图 2A 或图 2B 所示的 WQE1 和 WQE2 进行理解，第一工作请求可以参考图 2A 或图 2B 所示的 WQE1 进行理解。

图 4 示出了第一工作请求一种可能的结构。参考图 4，第一工作请求可以包括控制字段、任务字段和内存描述字段。

5 其中，任务字段可以用于描述第一工作请求的格式信息。其中，任务字段可以包括第一网络设备 311 处理数据传输消息的指示信息，例如，任务字段可以包括标识信息，可选的，该标识信息可以包括第一进程对应的第一标识。可选的，该标识信息中的第一标识可以解释为图 2A 或图 2B 对应的实施例中的 CQ#1 的标识和/或进程#1 的标识。该标识信息还可以包括其他信息，后文将介绍该其他信息的可能内容，此处暂不展开介绍。

10 可选的，第一工作请求还可以包括内存描述字段。内存描述字段可以用于描述第一网络设备 311 和/或第二网络设备 312 注册的内存空间。第一网络设备 311 可以根据内存描述字段从第一主机 32 中获取数据传输消息。可选的，内存描述字段可以包括地址字段，地址字段可以用于确定该内存空间的起始位置。可选的，内存描述字段还可以包括长度（length）字段，用于确定该内存空间的长度。可选的，内存描述字段还可以包括钥匙（key）字段，用于
15 唯一标识该内存空间。

第一网络设备 311 还可以根据该数据传输消息和标识信息封装第一报文，之后向第二网络设备发送第一报文。可选的，第一网络设备 311 可以根据 RDMA 协议封装第一报文。

20 可选的，该第一报文可以解释为图 2A 或图 2B 对应的优化后的实施例中的报文 1，即第一报文可以包括数据传输消息和标识信息。

第二网络设备 312 可以用于接收第一报文，对第一报文进行解封装，以获得数据传输消息和标识信息。之后，第二网络设备 312 可以根据数据传输消息生成反馈消息，用以指示数据传输消息的完成情况。可选的，第二网络设备 312 在获得数据传输消息后生成反馈消息，用以通知第一网络设备 311 已经成功接收数据传输消息。或者，第二网络设备 312 可以根据
25 数据传输消息是否成功写入第二主机 33 来生成相应的反馈消息。例如，若数据传输消息成功写入第二主机 33，该反馈消息指示的完成情况可以为成功，若数据传输消息写入第二主机 33 失败，该反馈消息指示的完成情况可以为失败。可选的，该反馈消息可以解释为图 2B 对应的实施例中的 R1。

30 第二网络设备 312 还可以用于根据反馈消息和标识信息封装第二报文，向第一网络设备 311 发送第二报文。可选的，第二报文可以解释为图 2B 对应的优化后的实施例中的报文 2，即第二报文可以包括第一标识和反馈消息。

35 第一网络设备 311 还可以用于接收第二报文，对第二报文进行解封装，以获得反馈消息和标识信息。之后，第一网络设备 311 可以根据第二报文中的标识信息和反馈消息通知第一进程数据传输消息的完成情况。

可选的，第一网络设备 311 可以为第一主机 32 运行的多个进程创建多个 CQ，每个进程对应多个 CQ 中的部分 CQ（例如一个 CQ），每个 CQ 用于向相应进程通知消息的完成情况。本

申请实施例将该多个 CQ 中对应于第一进程的 CQ 称作第一 CQ。

可选的，第一网络设备 311 所创建的 CQ 可以解释为图 2B 对应的实施例中的 CQ#1 或 CQ#2。示例性的，参考图 2A 或图 2B 对应的实施例，第一网络设备 311 可以为进程#1 创建 CQ#1，为进程#2 创建 CQ#2。其中，CQ#1 用于向进程#1 通知进程#1 提交的消息的完成情况，CQ#2 用于向进程#2 通知进程#2 提交的消息的完成情况。

第一网络设备 311 根据第二报文中的第一标识和反馈消息通知第一进程数据传输消息的完成情况可以指，第一网络设备 311 从多个 CQ 中确定第一标识对应的 CQ（称作第一 CQ），根据反馈消息在第一 CQ 中写入 CQE（称作第一 CQE），第一 CQE 描述数据传输消息的完成情况。可选的，第一 CQE 可以解释为图 2B 对应的实施例中的 CQE1。

可选的，第一 CQE 通过携带的部分字段（例如称作错误码）来指示数据传输消息的完成情况。例如，第一进程解析第一 CQE，若第一 CQE 中错误码的取值为 0，第一进程可以确定数据传输消息的完成情况为成功，具体例如为数据传输消息已完成传输等。例如，第一进程解析第一 CQE，若第一 CQE1 中错误码的取值为 1，第一进程可以确定该数据传输消息的完成情况为失败，具体例如为数据传输消息未完成传输等。

可选的，第一网络设备 311 通知的数据传输消息的完成情况可以与反馈消息所描述的完成情况一致。例如，若反馈消息描述数据传输消息的完成情况为成功，那么第一网络设备 311 可以通知第一进程数据传输消息已完成或传输成功。若反馈消息描述数据传输消息的完成情况为失败，那么第一网络设备 311 可以通知第一进程数据传输消息未完成或传输失败。

或者，可选的，第一网络设备 311 通知的数据传输消息的完成情况可以与反馈消息所描述的完成情况不一致。例如，若反馈消息描述数据传输消息的完成情况为成功，但是第一网络设备 311 因故障可能未正确的封装第一报文（例如封装的数据传输消息错误或目的地址错误等），那么第一网络设备 311 可以通知第一进程数据传输消息未完成或传输失败。

图 3 对应的实施例中，第一网络设备 311 可以创建供第一主机 32 运行的多个进程使用的 SSQ，第一网络设备 311 从 SSQ 获取第一进程的数据传输消息后，可以在封装数据传输消息的报文（即第一报文）中同时封装第一进程对应的第一标识。第二网络设备 312 接收到第一报文后，可以在封装反馈消息的报文（即第二报文）中同时封装第一标识，这样，第一网络设备 312 在接收到第二报文后，无需根据 SSQ 的上下文来确定第二报文中的反馈消息对应于第一进程，而是可以根据第二报文中的第一标识确定该反馈消息对应于第一进程，有利于高效的通知第一进程已处理该数据传输消息，进而有利于提高第一进程的运行效率。

下面对图 3 对应的实施例进行补充介绍。

一、上述图 3 对应的实施例提到，第一标识可以解释为图 2A 或图 2B 对应的实施例中的 CQ#1 的标识和/或进程#1。可选的，本申请实施例不限定第一标识的具体含义，只要第一标识能够用于确定第一 CQ 即可。

以第一标识为第一进程的标识为例，介绍第一网络设备 311 根据第一进程的标识确定第一 CQ 的方法。第一网络设备 311 可以保存映射表，该映射表记录进程与 CQ 之间的对应关系。第一网络设备 311 解封装第二报文得到第一进程的标识后，可以在映射表中查找第一进程与第一 CQ 对应关系，进而确定第一 CQ 的标识。继续参考图 2B 对应的实施例，RNIC 12 在获取

报文 2 携带的进程#1 的标识后, 可以从映射表中获取 CQ#1 的标识, 进而将 CQE1 写入 CQ#1。

可选的, 第一标识可以包括第一进程的标识和第一 CQ 的标识, 第一 CQE 可以包括第一进程的标识。可选的, 第一进程和其他进程可以共享同一 CQ, 第一进程通过第一 CQE 携带的第一进程的标识可以确定第一 CQE 为自己的 CQE。这样, 虽然第一网络设备 311 创建的 CQ 仍然不被所有进程共享, 但是通过让部分进程共享一个 CQ, 有利于减少创建的 CQ 的数目, 从而有利于节约内存空间。

二、图 5A 通过带有箭头的实线线段示例性示出了第一网络设备 311 获取进程的数据传输消息的过程, 通过带有箭头的虚线线段示例性示出了第一网络设备 311 通知进程数据传输消息的完成情况的过程。结合图 2A 或图 2B 对应的实施例, 下面介绍图 5A 所描述的过程。

参考图 1C 对应的实施例, 假设图 2A 或图 2B 所示的计算机系统还包括节点#3, RNIC 12 创建的 SSQ#2 可以绑定节点#1 和节点#3 之间的通道 n1-n3, SSQ#2 用于获取发往节点#3 的消息。假设进程#1 待传输的消息依次包括 M4、M3 和 M1, 其中, M4 和 M3 用于发往节点#3。假设进程#2 待传输的消息依次包括 M5 和 M2, 其中, M5 用于发往节点#3。

RNIC 12 可以根据 SSQ#2 的 PI 的取值依次将 M5 对应的 WQE5、M4 对应的 WQE4 和 M3 对应的 WQE3 写入 SSQ#2 中, 并且根据 SSQ#1 的 PI 的取值依次将 M1 对应的 WQE1 和 M2 对应的 WQE2 写入 SSQ#1。之后, RNIC 12 可以根据 SSQ#2 的 CI 的取值依次处理 SSQ#2 中的 WQE, 发送相应的消息, 之后 RNIC 12 可以根据 SSQ#1 的 CI 的取值依次处理 SSQ#1 中的 WQE, 发送相应的消息。假设 RNIC 12 从先至后依次发送 M5、M4、M3、M1 和 M2。

RNIC 12 在接收到携带反馈消息的报文后, 可以根据报文携带的第一标识确定相应的 CQ。假设 RNIC 12 从先至后依次接收 M5 对应的报文 a、M4 对应的报文 b、M3 对应的报文 c、M1 对应的报文 d 和 M2 对应的报文 e。RNIC 12 根据报文 a 携带的 CQ#2 的标识在 CQ#2 中新增 CQE-a, 根据报文 b 携带的 CQ#1 的标识在 CQ#1 中新增 CQE-b, 类似的在 CQ#1 中新增 WQE-c 和 CQE-d, 在 CQ#2 中新增 CQE-e。

进程#1 根据 CQ#1 中各 CQE 的写入顺序, 可以分别确定 CQE-b 对应于 M4、CQE-c 对应于 M3、CQE-d 对应于 M1。类似的, 进程#2 根据 CQ#2 中各 CQE 的写入顺序, 可以分别确定 CQE-a 对应于 M5、CQE-e 对应于 M2。

为了更加高效准确的知道进程消息的完成情况, 可选的, 第一报文和第二报文中的标识信息还可以包括第二标识, 第二标识用于从第一进程的多个数据传输消息中确定数据传输消息。相应的, 第一网络设备具体用于根据第一标识、反馈消息和第二标识通知第一进程数据传输消息的完成情况。可选的, 第一 CQE 可以包括第二标识。

继续参考图 5A, 可选的, M4、M3 和 M1 对应的第二标识可以分别为相应消息的序号。例如, M4、M3 和 M1 的第二标识可以分别为 1、2 和 3。CQE-b、CQE-c 和 CQE-d 携带的第二标识分别为 1、2 和 3。进程#1 可以根据 CQ#1 中各 CQE 的第二标识确定 CQE-b 对应于 M4、CQE-c 对应于 M3、CQE-d 对应于 M1。

参考图 5B, 假设进程#1 通过调用程序接口和 RNIC 12 的驱动程序依次向 RNIC 12 提交描述 M4 的 WR4、描述 M3 的 WR3 和描述 M1 的 WR1, 进程#2 通过调用程序接口和 RNIC 12 的驱动程序依次向 RNIC 12 提交描述 M5 的 WR5 和描述 M2 的 WR2。RNIC 12 依次根据 WR5、WR4、WR3、

WR1 和 WR2 在 SSQ#2 和 SSQ#1 中新增相应的 WQE。那么，可选的，消息的第二标识可以为相应 WR 的标识（例如 WR 的序号）。RNIC 12 在 CQ#1 和 CQ#2 中分别新增相应消息的 CQE 后，进程 #1 可以根据 CQ#1 中各 CQE 的第二标识确定 CQE 对应的消息，例如，通过确定 CQE-b 对应于 WR4，进而确定 CQE-b 对应于 M4。

5

三、下面对图 3 对应的实施例所涉及的数据传输消息、第一报文和第二报文进行补充介绍。

1) 可选的，第一进程的数据传输消息可以包括待写入第二主机 33 中的第一数据。第二网络 312 设备用于在获得第一数据后，在第二主机上存储第一数据。

10 可选的，第一网络设备 311 可以将第一数据分为多个数据片（segment），并且可以将多个片段依次封装至多个报文（称作第一报文序列）中。相应的，图 3 对应的实施例中涉及的第一报文可以指第一报文序列。

可选的，第一网络设备 311 可以在第一报文序列中的每个报文中均封装标识信息，或者，仅在第一报文序列中的最后一个报文中封装标识信息。可选的，标识信息可以封装在报文的扩展头中，第一数据或数据片可以作为报文的净荷数据。下面，以第一网络设备 311 在第一报文序列中的每个报文中均封装标识信息为例，介绍第一报文和第二报文的发送过程。

15 示例性的，图 3 对应的实施例中，数据传输系统 31 可以采用发送（send）操作或写（write）操作或原子（atomic）操作来执行数据传输消息对应的工作请求。

图 6A 示出了发送操作对应的第一报文和第二报文的传输过程。参考图 6A，第一报文序列包括报文#1、报文#2、……、报文#k，报文#1、报文#2、……、报文#k 中分别携带第一数据中的数据片#1、数据片#2、……、数据片#k，k 为任意一个正整数。并且，第一报文序列中的每个报文均携带标识信息。可选的，该标识信息可以包括前文介绍的第一标识，还可以包括第二标识。此外，该标识信息还可以第三标识。该第三标识可以包括第二进程的标识和/或第二进程对应的 RQ 的标识。图 6A 示例性示出了标识信息中的第一标识和 RQ 的标识。

25 以第三标识包括第二进程对应的 RQ 为例，第二网络设备 312 接收到第一网络设备 311 发送的第一报文序列后，可以从第三标识对应的 RQ 中读取 WQE，该 WQE 描述第二进程对应的存储空间。之后第二网络设备 312 可以将第一数据保存在 WQE 描述的存储空间。第二网络设备 312 接收第一报文序列后，可以向第一网络设备 311 发送确认报文，该确认报文携带确认字符（acknowledge character，ACK），因此，本申请实施例将该确认报文称作 ACK 报文。该 ACK 报文可以携带前文介绍的反馈消息（例如为确认字符），还可以携带第一报文序列中标识信息中的全部或部分内容。例如，ACK 报文还可以携带第一标识。

图 6B 示出了写操作对应的第一报文和第二报文的传输过程。参考图 6B，第一报文序列包括报文#1、报文#2、……、报文#k，报文#1、报文#2、……、报文#k 中分别携带第一数据中的数据片#1、数据片#2、……、数据片#k，k 为大于或等于 3 的正整数。并且，第一报文序列中的每个报文均携带标识信息。可选的，该标识信息可以包括前文介绍的第一标识，可选的，还可以包括第二标识。图 6B 示例性示出了标识信息中的第一标识。

35 此外，报文#1 中还可以封装有第一数据在第二主机 33 中的写入位置。第二网络设备 312 接收到第一报文序列后，可以根据该写入位置将第一数据存储在第二主机 33。第二网络

设备 312 接收第一报文序列后，可以向第一网络设备 311 发送 ACK 报文。该 ACK 报文可以携带前文介绍的反馈消息（例如为确认字符），还可以携带第一报文序列中标识信息中的全部或部分内容。例如，ACK 报文还可以携带第一标识。

图 6C 示出了原子操作对应的第一报文和第二报文的传输过程。参考图 6C，第一报文包括第一数据和标识信息。可选的，该标识信息可以包括前文介绍的第一标识，可选的，还可以包括第二标识。图 6C 示例性示出了标识信息中的第一标识。第二网络设备 312 接收到第一报文后，可以将第一数据存储在第二主机 33。第二网络设备 312 接收第一报文后，可以向第一网络设备 311 发送第二报文，该第二报文可以例如为原子反馈（atomic response）报文或原子确认（atomic acknowledge）报文。第二报文可以携带前文介绍的反馈消息（例如为确认字符），还可以携带第一报文序列中标识信息中的全部或部分内容。例如，第二报文还可以携带第一标识。

2) 可选的，数据传输消息可以包括第一数据的源地址和目的地址，并且，第一数据的源地址指向第二主机 33，第一数据的地址指向第一主机 32，因此，数据传输消息可以不携带第一数据。

第二网络设备 312 可以用于在获得第一报文中封装的源地址、目的地址和标识信息后，根据源地址在第二主机 33 上读取第一数据，根据第一数据、标识信息、反馈消息和目的地址封装第二报文，并向第一主机 32 发送第二报文。图 6B 示例性示出了标识信息中的第一标识。

可选的，标识信息可以封装在报文的扩展头中。第一网络设备 311 用于在获得第二报文中的第一数据和目的地址后，根据目的地址在第一主机上存储第一数据。第一网络设备 311 还用于根据第二报文中的第一标识和反馈消息通知第一进程数据传输消息的完成情况。

可选的，第二网络设备 312 可以将第一数据分为多个片段，并且可以将多个片段依次封装至多个报文（称作第二报文序列）中。相应的，图 3 对应的实施例中涉及的第二报文可以指第二报文序列。

可选的，第二网络设备 312 可以在第二报文序列中的每个报文中均封装标识信息中的部分或全部内容，或者，仅在第二报文序列中的最后一个报文中封装标识信息中的部分或全部内容。可选的，标识信息可以封装在报文的扩展头中，第一数据或数据片可以作为报文的净荷数据。下面，以第一网络设备 311 在第一报文序列中的每个报文中均封装第一标识为例，介绍第一报文和第二报文的发送过程。

图 6D 示出了读（read）操作对应的第一报文和第二报文的传输过程。参考图 6D，第一报文包括源地址、目的地址和第一标识，第二网络设备 312 接收到第一报文后，可以根据源地址从第二主机 33 中读取第一数据，之后对第一数据进行分片，依次得到数据片#1、数据片#2、……、数据片#k，k 为大于或等于 3 的正整数。之后第二网络设备 312 可以封装并发送第二报文序列，第二报文序列包括报文#1、报文#2、……、报文#k，报文#1、报文#2、……、报文#k 中分别携带数据片#1、数据片#2、……、数据片#k。此外，第二报文序列中的每个报文还可以包括目的地址、第一标识和反馈消息。第一网络设备 311 接收到第二报文后，可以根据目的地址将第一数据存储在第二主机 32 中，根据第一标识和反馈消息通知第一进程该数据传输消息的完成情况。

以上介绍了本申请实施例提供的计算机系统和数据传输系统，基于同一发明构思，本申请实施例还提供一种数据传输方法，该方法可以为 RDMA 数据传输方法。参考图 7，该数据传输方法可以包括步骤 S701 至步骤 S704。

5 S701、第一网络设备从 SSQ 获取第一进程的 RDMA 数据传输消息；

第一网络设备可以设置在第一主机上，第一进程是第一主机运行的使用共享发送队列的多个进程中的任意一个进程，该数据传输消息可以为 RDMA 数据传输消息。

S702、第一网络设备向第二网络设备发送第一报文；

10 第二网络设备设置在第二主机上，第一报文包括数据传输消息和第一进程对应的第一标识。

第二网络设备接收来自第一网络设备的第一报文，第一网络设备设置于第一主机上，第二网络设备设置于第二主机上，第一报文包括第一进程的数据传输消息和第一进程对应的第一标识，数据传输消息为第一网络设备从共享发送队列获取的，第一进程是第一主机运行的使用共享发送队列的多个进程中的任意一个进程。

15 S703、第二网络设备根据第一报文向第一网络设备发送第二报文，第二报文包括第一标识和反馈消息；

S704、第一网络设备根据第二报文中的第一标识和反馈消息通知第一进程数据传输消息的完成情况。

20 第一网络设备接收来自第二网络设备的第二报文，第二报文包括第一标识和反馈消息，反馈消息指示数据传输消息的完成情况。第一网络设备根据第二报文中的第一标识和反馈消息通知第一进程数据传输消息的完成情况。

需要说明的是，步骤 S701、步骤 S702 和步骤 S704 对应的方法为第一网络设备所执行的方法，该方法可以视为图 3 对应实施例中第一网络设备所执行的方法。图 7 所示的第一网络设备所执行的步骤 S701、步骤 S702 和步骤 S704 中未详尽描述的实现方式和技术效果可以参见图 3 所示的实施例中第一网络设备的相关描述。对于步骤 S701、步骤 S702 和步骤 S704 对应的方法，本申请实施例不限定步骤 S703 中第二网络设备具体执行的方法，只要第一网络设备向第二网络设备发送第一报文后，第二网络设备可以向第一网络设备发送包括第一标识的第二报文即可。

30 需要说明的是，步骤 S703 对应的方法为第二网络设备所执行的方法，该方法可以视为图 3 对应实施例中第二网络设备所执行的方法。图 7 所示的第二网络设备所执行的步骤 S703 中未详尽描述的实现方式和技术效果可以参见图 3 所示的实施例中第二网络设备的相关描述。对于步骤 S702 对应的方法，本申请实施例不限定第一网络设备所执行的步骤，只要第一网络设备能够向第二网络设备发送包括第一标识的第一报文即可。

35 上述详细阐述了本申请实施例的方法，为了便于更好的实施本申请实施例的上述方案，相应地，下面还提供用于配合实施上述方案的相关设备。

参见图 8，图 8 是本申请实施例提供的一种网络设备的结构示意图，该网络设备可以是上述图 3 或图 7 对应的实施例中的第一网络设备，可以执行相应实施例中第一网络设备所执

行的方法和步骤。

如图 8 所示, 该网络设备 800 可以包括获取单元 801, 用于从共享发送队列获取第一进程的 RDMA 数据传输消息, 网络设备设置在第一主机上, 第一进程是第一主机运行的使用共享发送队列的多个进程中的任意一个进程。网络设备 800 还可以包括发送单元 802, 用于向第二网络设备发送第一报文, 第二网络设备设置在第二主机上, 第一报文包括 RDMA 数据传输消息和第一进程对应的第一标识。还可以包括接收单元 803, 用于接收来自第二网络设备的第二报文, 第二报文包括第一标识和反馈消息, 反馈消息指示 RDMA 数据传输消息的完成情况。还可以包括完成单元 804, 用于根据第二报文中的第一标识和反馈消息通知第一进程 RDMA 数据传输消息的完成情况。

在一种可能的实现方式中, 第一报文和第二报文还包括第二标识, 第二标识用于从第一进程的多个 RDMA 数据传输消息中确定 RDMA 数据传输消息。完成单元 804 具体用于根据第一标识、反馈消息和第二标识通知第一进程 RDMA 数据传输消息的完成情况。

在一种可能的实现方式中, 共享发送队列具体用于保存来自多个进程的工作请求。获取单元 801 具体用于从共享发送队列获取来自第一进程的第一工作请求, 第一工作请求描述 RDMA 数据传输消息, 根据第一工作请求获取 RDMA 数据传输消息。

在一种可能的实现方式中, 完成单元 804 具体用于, 根据第一标识从多个完成队列中确定第一进程对应的第一完成队列, 根据反馈消息在第一完成队列中写入工作完成元素, 工作完成元素用于通知第一进程 RDMA 数据传输消息的完成情况。

应理解, 网络设备 800 包括的各单元可以软件模块、也可以为硬件模块、或部分为软件模块部分为硬件模块。

关于网络设备 800 的可能实现方式及有益效果可以参考图 3 或图 7 对应的实施例的相关内容, 此处不再赘述。

需要说明的是, 上述网络设备 800 的结构仅仅作为一种示例, 不应构成具体限定, 可以根据需要对网络设备中的各个单元进行增加、减少或合并。此外, 该网络设备 800 中的各个单元的操作和/或功能是为了实现上述图 3 或图 7 所描述的第一网络设备的功能或方法, 为了简洁, 在此不再赘述。

参见图 9, 图 9 是本申请实施例提供的另一种网络设备的结构示意图, 该网络设备可以是上述图 3 或图 7 对应的实施例中的第二网络设备, 可以执行相应实施例中第二网络设备所执行的方法和步骤。

如图 9 所示, 该网络设备 900 可以包括接收单元 901, 用于接收来自第一网络设备的第一报文, 第一网络设备设置于第一主机上, 网络设备设置于第二主机上, 第一报文包括第一进程的 RDMA 数据传输消息和第一进程对应的第一标识, RDMA 数据传输消息为第一网络设备从共享发送队列获取的, 第一进程是第一主机运行的使用共享发送队列的多个进程中的任意一个进程。还可以包括发送单元 902, 用于根据第一报文向第一网络设备发送第二报文, 第二报文包括第一标识和反馈消息, 第一标识和反馈消息指示第一网络设备通知第一进程 RDMA 数据传输消息的完成情况。

在一种可能的实现方式中, 第一报文和第二报文还包括第二标识, 第二标识用于从第一进程的多个 RDMA 数据传输消息中确定 RDMA 数据传输消息, 第一标识、反馈消息和第二标识

指示第一网络设备通知第一进程 RDMA 数据传输消息的完成情况。

应理解，网络设备 900 包括的各单元可以软件模块、也可以为硬件模块、或部分为软件模块部分为硬件模块。

5 关于网络设备 900 的可能实现方式及有益效果可以参考图 3 或图 7 对应的实施例的相关内容，此处不再赘述。

需要说明的是，上述网络设备 900 的结构仅仅作作为一种示例，不应构成具体限定，可以根据需要对网络设备中的各个单元进行增加、减少或合并。此外，该网络设备 900 中的各个单元的操作和/或功能是为了实现上述图 3 或图 7 所描述第二网络设备的功能或方法，为了简洁，在此不再赘述。

10 本申请还提供一种计算机可读存储介质，其中，计算机可读存储介质存储有计算机程序，当该计算机程序被处理器执行时，可以实现上述方法实施例中记载的任意一种的部分或全部步骤。

本发明实施例还提供一种计算机程序，该计算机程序包括指令，当该计算机程序被计算机执行时，使得计算机可以执行任意一种方法的部分或全部步骤。

15 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中未详述的部分，可以参见其它实施例的相关描述。

需要说明的是，对于前述的各方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请并不受所描述的动作顺序的限制，因为依据本申请，某些步骤可能可以采用其它顺序或者同时进行。其次，本领域技术人员也应该知悉，20 说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本申请所必须的。

在本申请所提供的几个实施例中，应该理解到，所揭露的装置，可通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如上述单元的划分，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，例如多个单元或组件可以结合或者可以集成到25 另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性或其它的形式。

上述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，即可以位于一个地方，或者也可以分布到多个网络单元30 上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，在本申请各实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

本申请实施例中，多个是指两个或两个以上，本申请不做限制。在本申请实施例中，“/”35 可以表示前后关联的对象是一种“或”的关系，例如，A/B 可以表示 A 或 B；“和/或”可以用于描述关联对象存在三种关系，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，其中 A，B 可以是单数或者复数。为了便于描述本申请实施例的技术方案，在本申请实施例中，可以采用“第一”、“第二”等字样对功能相同或相似的技术特

征进行区分。该“第一”、“第二”等字样并不对数量和执行次序进行限定，并且“第一”、“第二”等字样也并无限定一定不同。在本申请实施例中，“示例性的”或者“例如”等词用于表示例子、例证或说明，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它实施例或设计方案更优选或更具优势。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，便于理解。

本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于系统实施例而言，由于其基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

显然，本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的范围。这样，倘若本申请的这些修改和变型属于本发明权利要求的范围之内，则本发明也意图包括这些改动和变型在内。

权 利 要 求

1. 一种远程直接内存访问 RDMA 数据传输系统，其特征在于，所述 RDMA 数据传输系统包括第一网络设备和第二网络设备，所述第一网络设备设置于第一主机上，所述第二网络设备设置于第二主机上；

5 所述第一网络设备，用于从共享发送队列获取第一进程的 RDMA 数据传输消息，向所述第二网络设备发送第一报文，其中，所述第一进程是所述第一主机运行的使用所述共享发送队列的多个进程中的任意一个进程，所述第一报文包括所述 RDMA 数据传输消息和所述第一进程对应的第一标识；

10 所述第二网络设备，用于接收所述第一报文，根据所述第一报文向所述第一网络设备发送第二报文，其中，所述第二报文包括所述第一标识和反馈消息，所述反馈消息指示所述 RDMA 数据传输消息的完成情况；

 所述第一网络设备，还用于接收所述第二报文，根据所述第二报文中的所述第一标识和所述反馈消息通知所述第一进程所述 RDMA 数据传输消息的完成情况。

15 2. 根据权利要求 1 所述的系统，其特征在于，所述第一报文和所述第二报文还包括第二标识，所述第二标识用于从所述第一进程的多个 RDMA 数据传输消息中确定所述 RDMA 数据传输消息；

 所述第一网络设备具体用于根据所述第一标识、所述反馈消息和所述第二标识通知所述第一进程所述 RDMA 数据传输消息的完成情况。

20 3. 根据权利要求 1 或 2 所述的系统，其特征在于，所述共享发送队列具体用于保存来自所述多个进程的工作请求；

 所述第一网络设备具体用于从所述共享发送队列获取来自所述第一进程的第一工作请求，根据所述第一工作请求获取所述 RDMA 数据传输消息，其中，所述第一工作请求描述所述 RDMA 数据传输消息。

25 4. 根据权利要求 1 至 3 中任一项所述的系统，其特征在于，所述第一网络设备具体用于根据所述第一标识从多个完成队列中确定所述第一进程对应的第一完成队列，根据所述反馈消息在所述第一完成队列中写入工作完成元素，其中，所述工作完成元素描述所述 RDMA 数据传输消息的完成情况。

5. 一种远程直接内存访问 RDMA 数据传输方法，其特征在于，包括：

30 第一网络设备从共享发送队列获取第一进程的 RDMA 数据传输消息，所述第一网络设备设置于第一主机上，所述第一进程是所述第一主机运行的使用所述共享发送队列的多个进程中的任意一个进程；

 所述第一网络设备向第二网络设备发送第一报文，所述第二网络设备设置在第二主机上，所述第一报文包括所述 RDMA 数据传输消息和所述第一进程对应的第一标识；

35 所述第一网络设备接收来自所述第二网络设备的第二报文，所述第二报文包括所述第一标识和反馈消息，所述反馈消息指示所述 RDMA 数据传输消息的完成情况；

 所述第一网络设备根据所述第二报文中的所述第一标识和所述反馈消息通知所述第一进程所述 RDMA 数据传输消息的完成情况。

6. 根据权利要求 5 所述的方法，其特征在于，所述第一报文和所述第二报文还包括第二

标识, 所述第二标识用于从所述第一进程的多个 RDMA 数据传输消息中确定所述 RDMA 数据传输消息;

所述第一网络设备根据所述第二报文中的所述第一标识和所述反馈消息通知所述第一进程所述 RDMA 数据传输消息的完成情况包括:

5 所述第一网络设备根据所述第一标识、所述反馈消息和所述第二标识通知所述第一进程所述 RDMA 数据传输消息的完成情况。

7. 根据权利要求 5 或 6 所述的方法, 其特征在于, 所述共享发送队列具体用于保存来自所述多个进程的工作请求;

所述第一网络设备从共享发送队列获取第一进程的 RDMA 数据传输消息包括:

10 所述第一网络设备从所述共享发送队列获取来自所述第一进程的第一工作请求, 所述第一工作请求描述所述 RDMA 数据传输消息;

所述第一网络设备根据所述第一工作请求获取所述 RDMA 数据传输消息。

8. 根据权利要求 5 至 7 中任一项所述的方法, 其特征在于, 所述第一网络设备根据所述第二报文中的所述第一标识和所述反馈消息通知所述第一进程所述 RDMA 数据传输消息的完成
15 情况, 包括:

所述第一网络设备根据所述第一标识从多个完成队列中确定所述第一进程对应的第一完成队列;

所述第一网络设备根据所述反馈消息在所述第一完成队列中写入工作完成元素, 所述工作完成元素用于通知所述第一进程所述 RDMA 数据传输消息的完成情况。

20 9. 一种远程直接内存访问 RDMA 数据传输方法, 其特征在于, 包括:

第二网络设备接收来自第一网络设备的第一报文, 所述第一网络设备设置于第一主机上, 所述第二网络设备设置于第二主机上, 所述第一报文包括第一进程的 RDMA 数据传输消息和所述第一进程对应的第一标识, 所述 RDMA 数据传输消息为所述第一网络设备从共享发送队列获取的, 所述第一进程是所述第一主机运行的使用所述共享发送队列的多个进程中的任意一个
25 进程;

所述第二网络设备根据所述第一报文向所述第一网络设备发送第二报文, 所述第二报文包括所述第一标识和反馈消息, 所述第一标识和所述反馈消息指示所述第一网络设备通知所述第一进程所述 RDMA 数据传输消息的完成情况。

10. 根据权利要求 9 所述的方法, 其特征在于, 所述第一报文和所述第二报文还包括第二
30 标识, 所述第二标识用于从所述第一进程的多个 RDMA 数据传输消息中确定所述 RDMA 数据传输消息, 所述第一标识、所述反馈消息和所述第二标识指示所述第一网络设备通知所述第一进程所述 RDMA 数据传输消息的完成情况。

11. 一种网络设备, 其特征在于, 包括:

获取单元, 用于从共享发送队列获取第一进程的 RDMA 数据传输消息, 所述网络设备设置
35 在第一主机上, 所述第一进程是所述第一主机运行的使用所述共享发送队列的多个进程中的任意一个进程;

发送单元, 用于向第二网络设备发送第一报文, 所述第二网络设备设置在第二主机上, 所述第一报文包括所述 RDMA 数据传输消息和所述第一进程对应的第一标识;

接收单元，用于接收来自所述第二网络设备的第二报文，所述第二报文包括所述第一标识和反馈消息，所述反馈消息指示所述 RDMA 数据传输消息的完成情况；

完成单元，用于根据所述第二报文中的所述第一标识和所述反馈消息通知所述第一进程所述 RDMA 数据传输消息的完成情况。

- 5 12. 根据权利要求 11 所述的网络设备，其特征在于，所述第一报文和所述第二报文还包括第二标识，所述第二标识用于从所述第一进程的多个 RDMA 数据传输消息中确定所述 RDMA 数据传输消息；

所述完成单元具体用于根据所述第一标识、所述反馈消息和所述第二标识通知所述第一进程所述 RDMA 数据传输消息的完成情况。

- 10 13. 根据权利要求 11 或 12 所述的网络设备，其特征在于，所述共享发送队列具体用于保存来自所述多个进程的工作请求；

所述获取单元具体用于：

从所述共享发送队列获取来自所述第一进程的第一工作请求，所述第一工作请求描述所述 RDMA 数据传输消息；

- 15 根据所述第一工作请求获取所述 RDMA 数据传输消息。

14. 根据权利要求 11 至 13 中任一项所述的网络设备，其特征在于，所述完成单元具体用于：

根据所述第一标识从多个完成队列中确定所述第一进程对应的第一完成队列；

根据所述反馈消息在所述第一完成队列中写入工作完成元素，所述工作完成元素用于通知所述

- 20 所述第一进程所述 RDMA 数据传输消息的完成情况。

15. 一种网络设备，其特征在于，包括：

接收单元，用于接收来自第一网络设备的第一报文，所述第一网络设备设置于第一主机上，所述网络设备设置于第二主机上，所述第一报文包括第一进程的 RDMA 数据传输消息和所述

- 25 所述第一进程对应的第一标识，所述 RDMA 数据传输消息为所述第一网络设备从共享发送队列获取的，所述第一进程是所述第一主机运行的使用所述共享发送队列的多个进程中的任意一个进程；

发送单元，用于根据所述第一报文向所述第一网络设备发送第二报文，所述第二报文包括所述第一标识和反馈消息，所述第一标识和所述反馈消息指示所述第一网络设备通知所述第一进程所述 RDMA 数据传输消息的完成情况。

- 30 16. 根据权利要求 15 所述的网络设备，其特征在于，所述第一报文和所述第二报文还包括第二标识，所述第二标识用于从所述第一进程的多个 RDMA 数据传输消息中确定所述 RDMA 数据传输消息，所述第一标识、所述反馈消息和所述第二标识指示所述第一网络设备通知所述第一进程所述 RDMA 数据传输消息的完成情况。

17. 一种网络设备，其特征在于，包括：处理器以及存储器，

- 35 所述处理器用于执行所述存储器中存储的指令，使得所述网络设备执行权利要求 5 至 8 中任一项所述的方法。

18. 一种网络设备，其特征在于，包括：处理器以及存储器，

所述处理器用于执行所述存储器中存储的指令，使得所述网络设备执行权利要求 9 或 10

所述的方法。

计算机系统

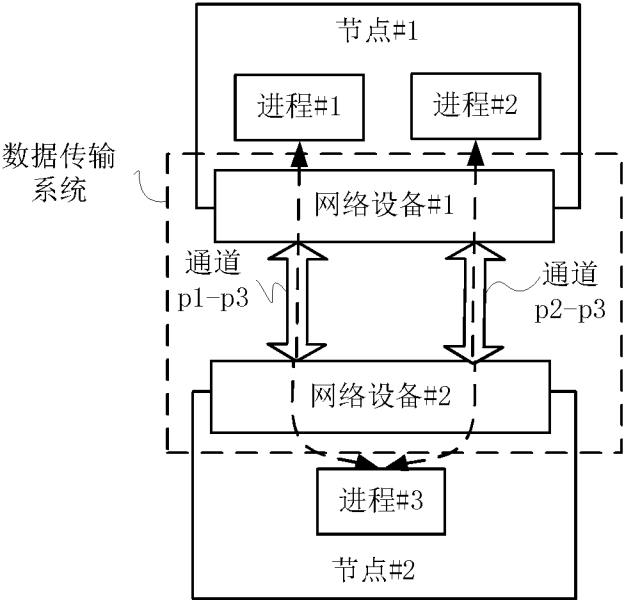


图 1A

计算机系统

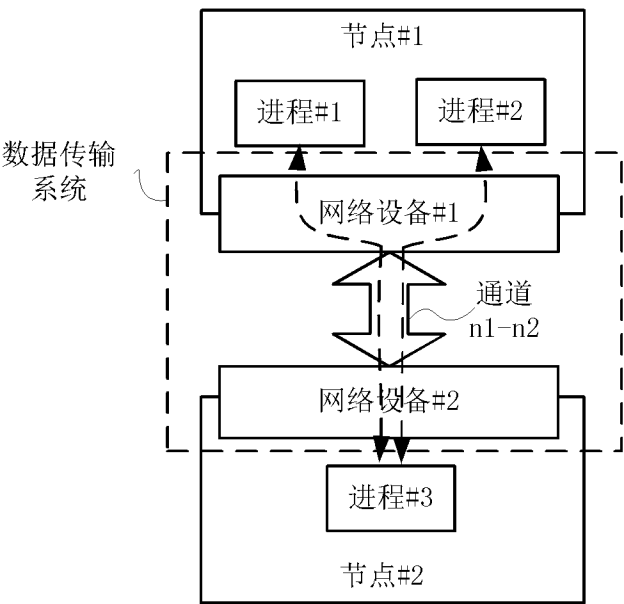


图 1B

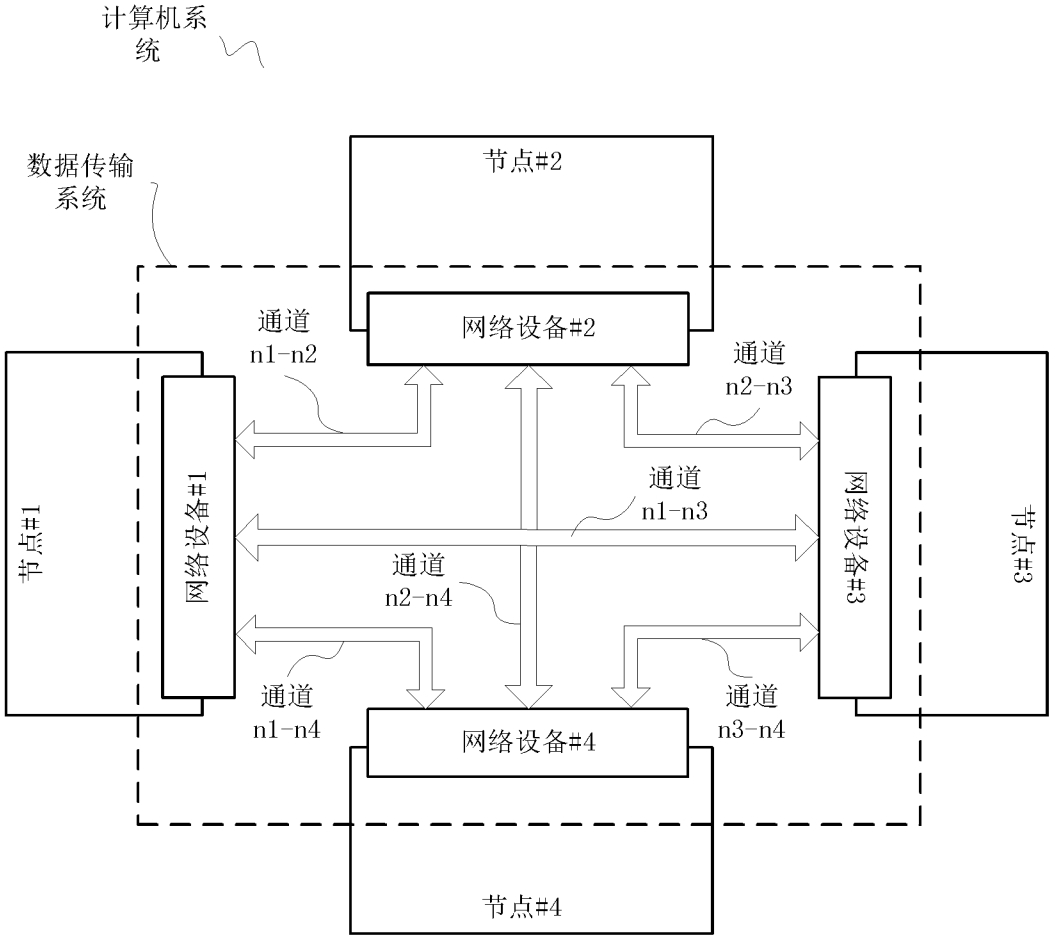


图 1C

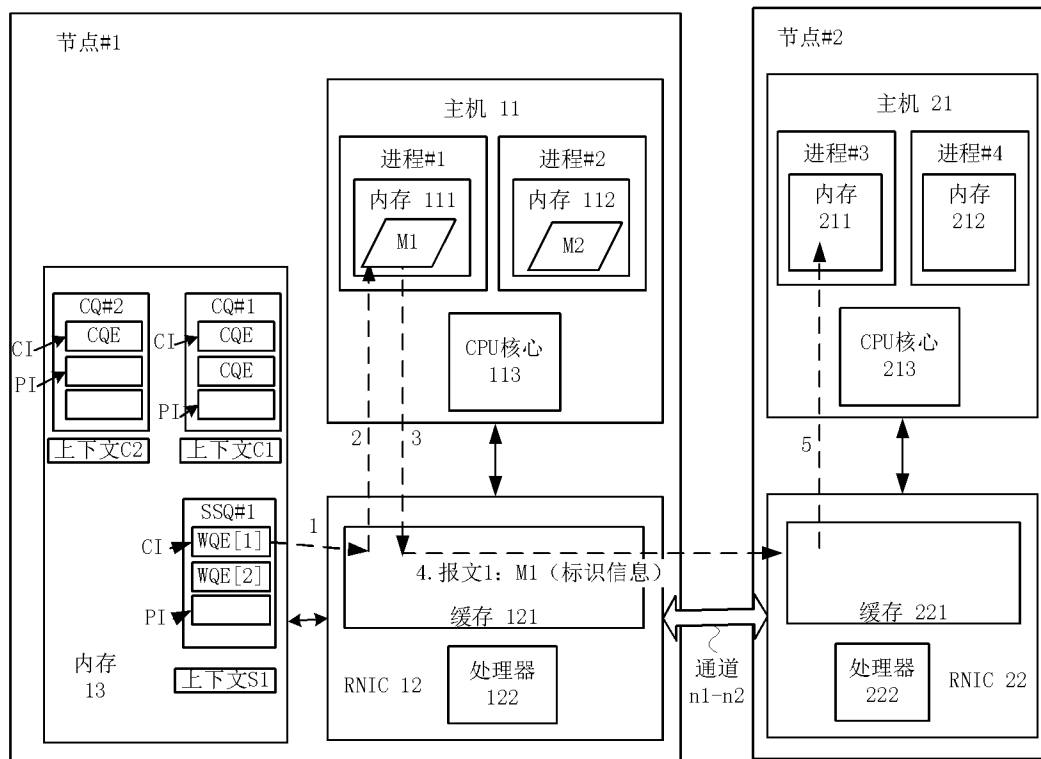


图 2A

计算机系统 ✓

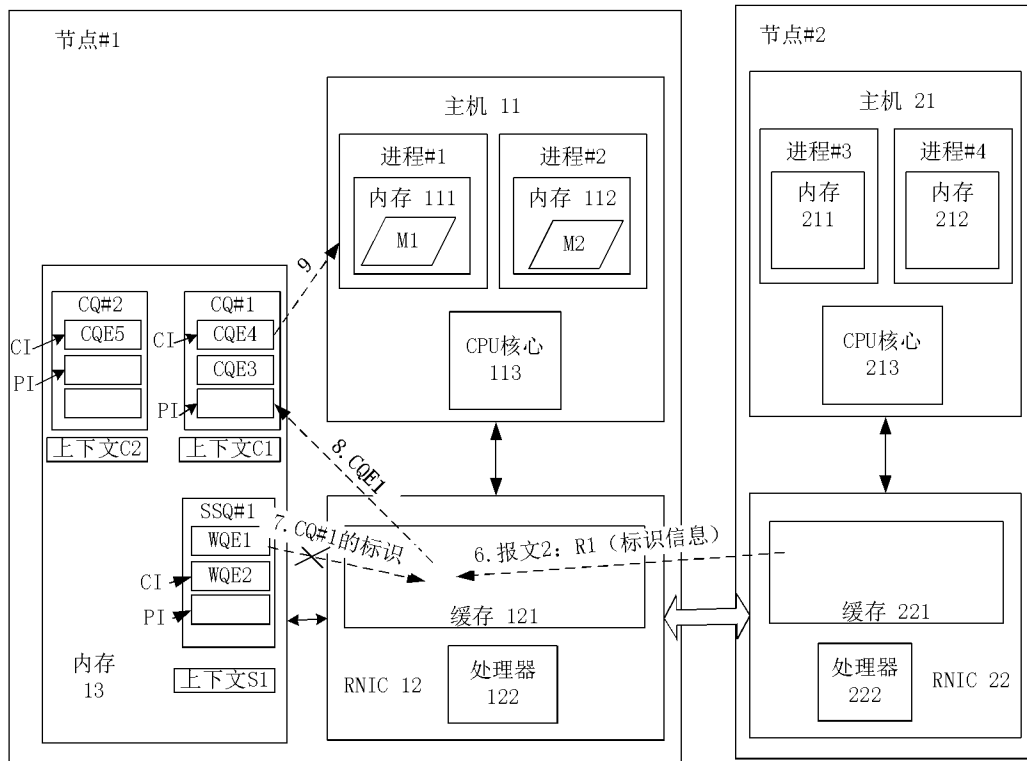


图 2B

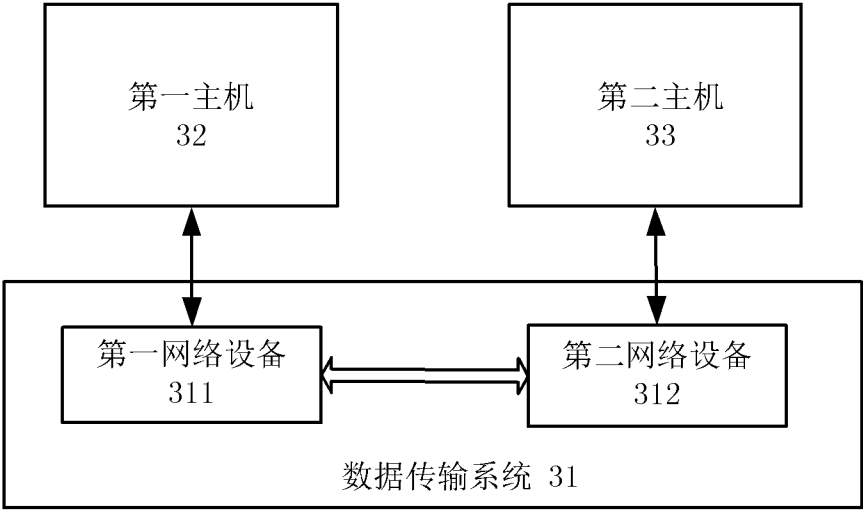


图 3

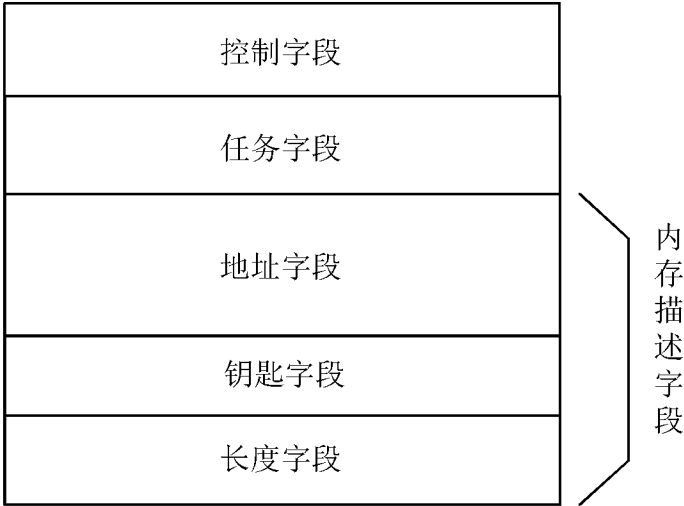


图 4

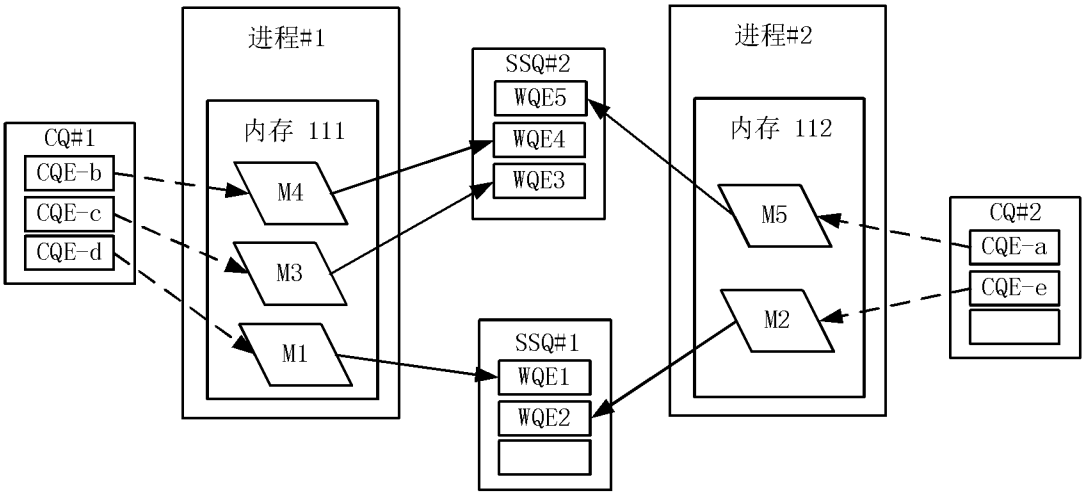


图 5A

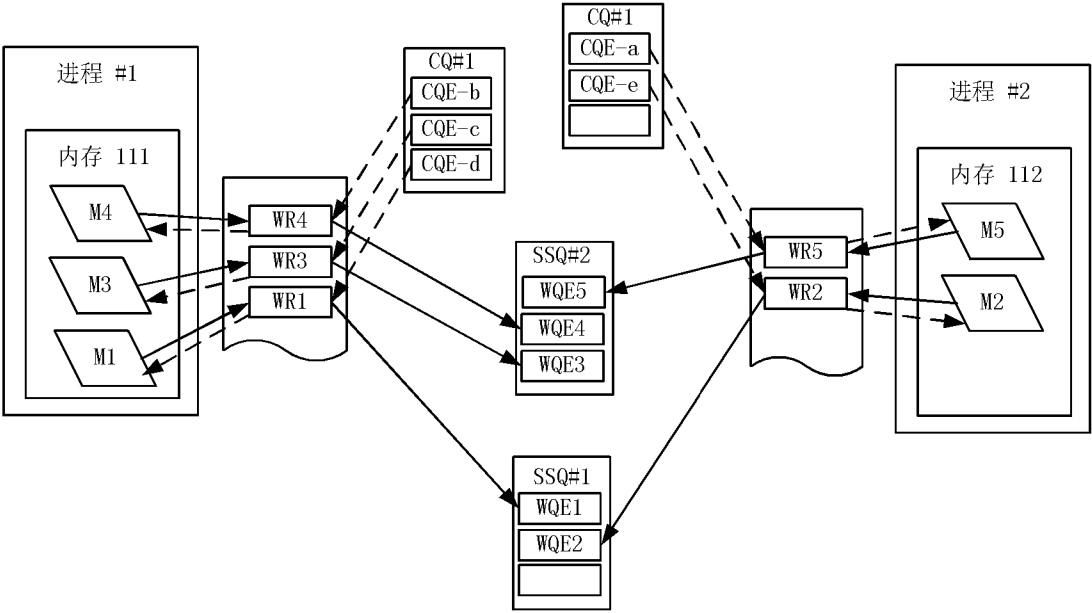


图 5B

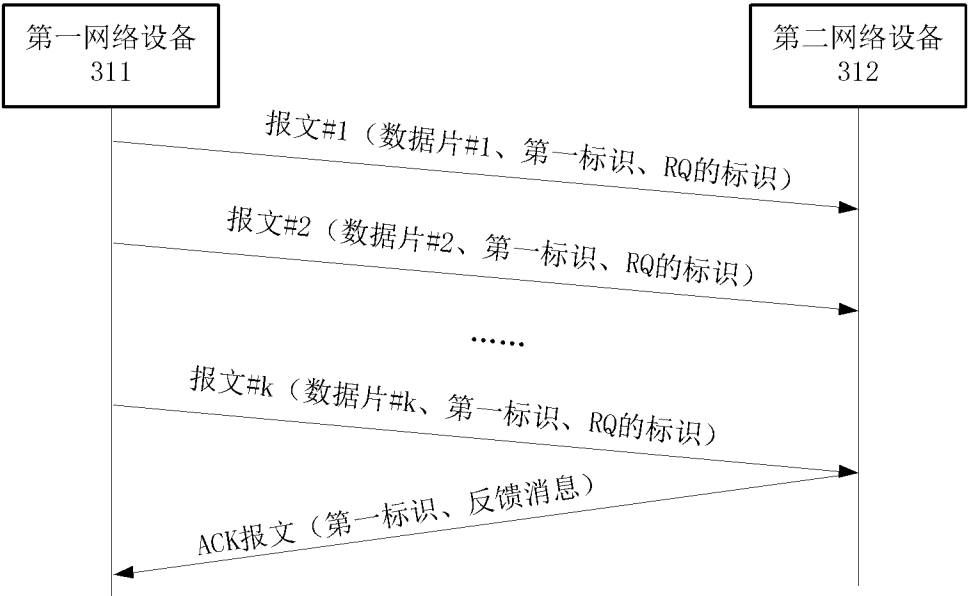


图 6A

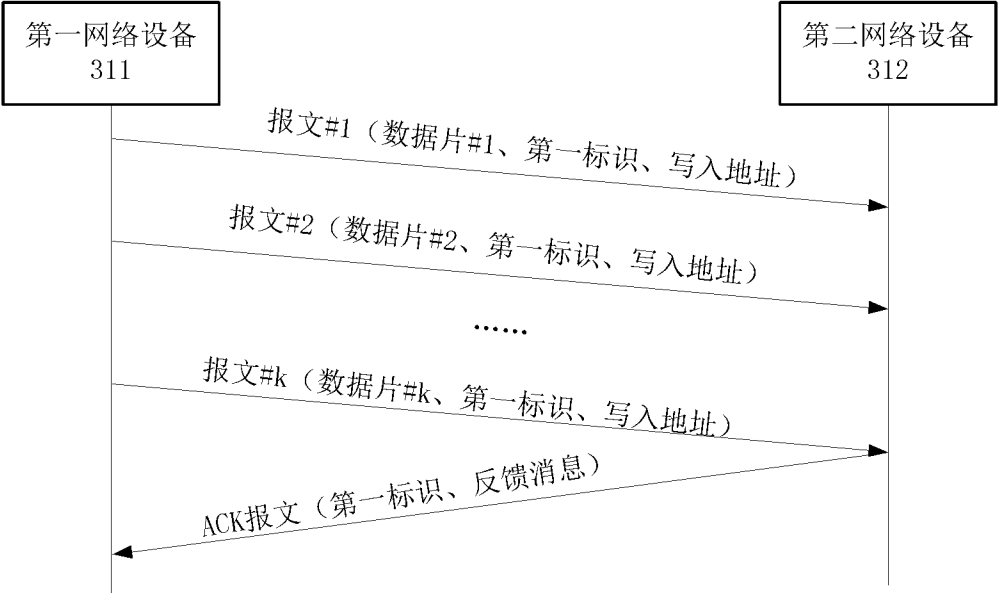


图 6B

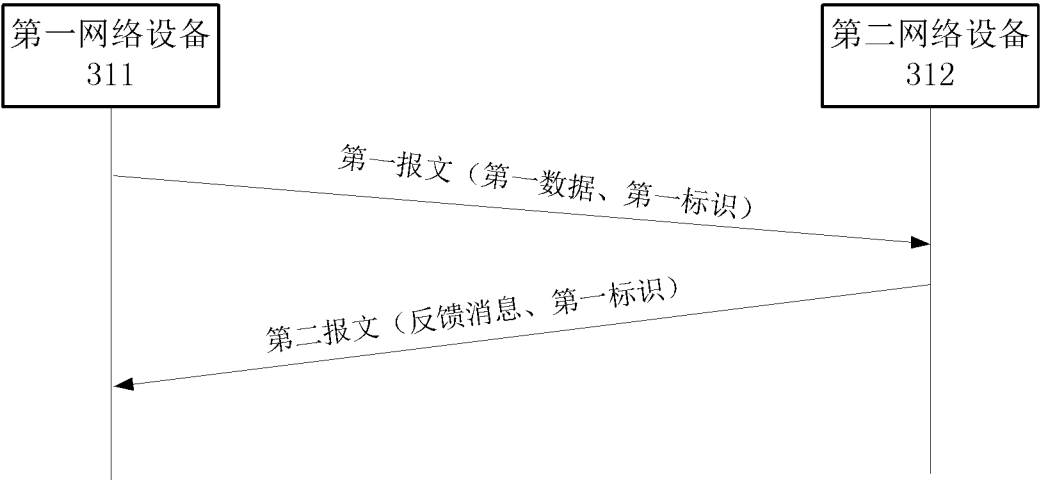


图 6C

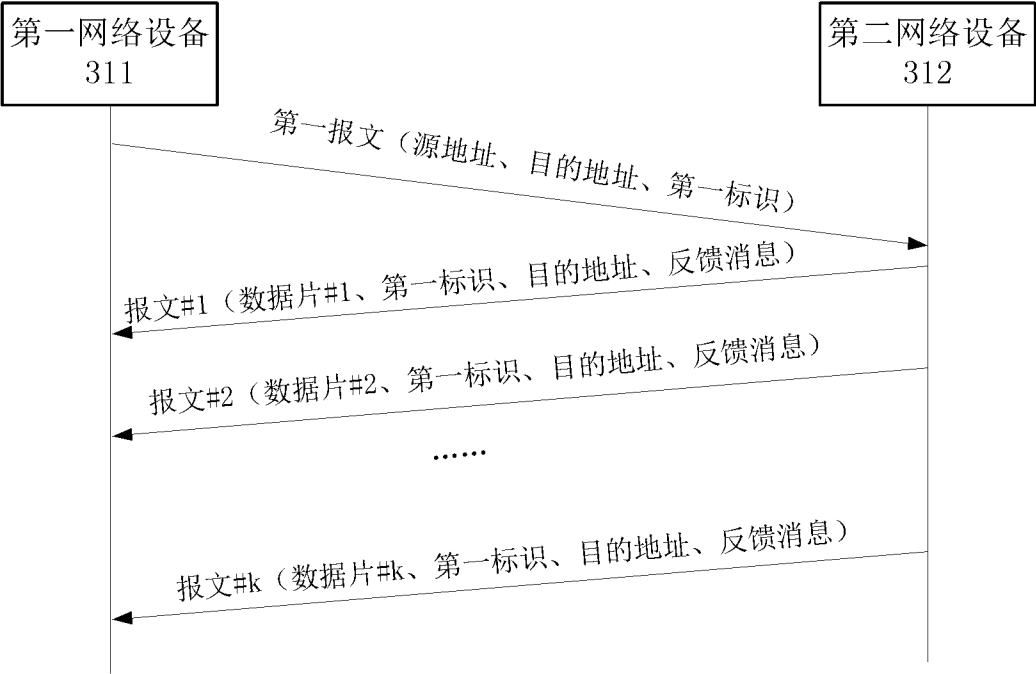


图 6D

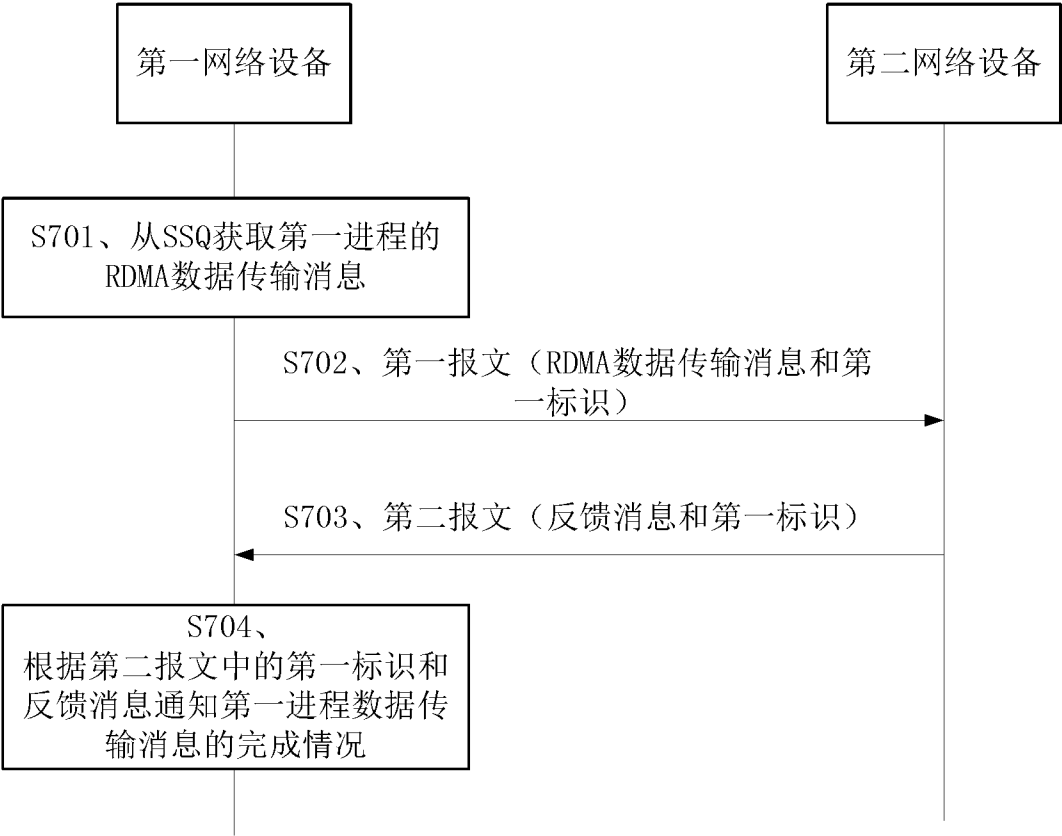


图 7

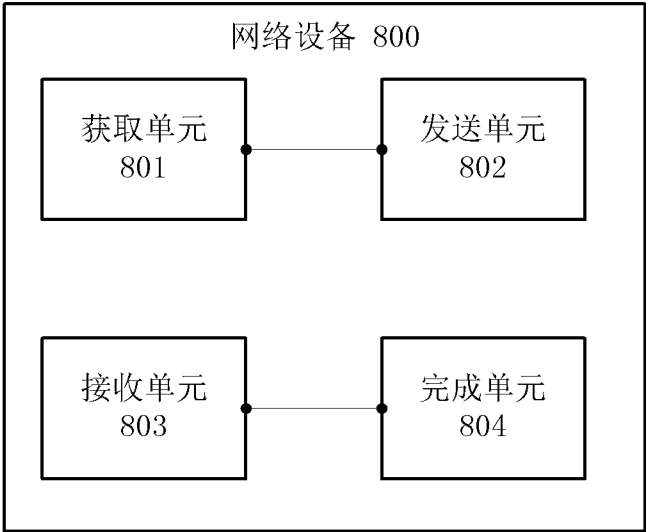


图 8

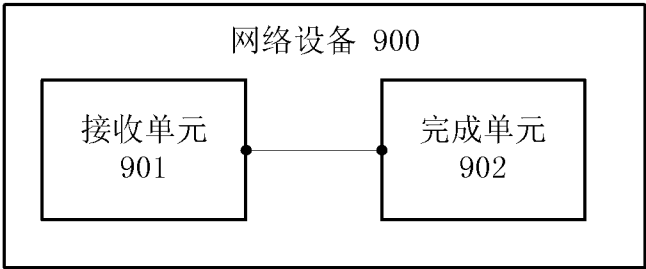


图 9

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/099788

A. CLASSIFICATION OF SUBJECT MATTER

H04L 67/025(2022.01)i; G06F 9/54(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

H04L; G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT; VEN; CNKI; ENTXTC; IEEE Xplore; IETF: 进程, 应用, 标识, 共享发送队列, 上下文, 封装, 完成通知, 反馈, 响应, RDMA, IB, RoCE, iWARPF, shared send queue, SSQ, context, complete, ID, app+, process, consumer

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 103999068 A (INTEL CORP.) 20 August 2014 (2014-08-20) description, paragraphs [0008]-[0066]	1-18
Y	CN 111782417 A (BEIJING INSTITUTE OF COMPUTER TECHNOLOGY AND APPLICATION) 16 October 2020 (2020-10-16) description, paragraphs [0004]-[0043]	1-18
A	CN 108009032 A (ZHENGZHOU YUNHAI INFORMATION TECHNOLOGY CO., LTD.) 08 May 2018 (2018-05-08) entire document	1-18
A	US 2016072906 A1 (MELLANOX TECHNOLOGIES LTD.) 10 March 2016 (2016-03-10) entire document	1-18



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

31 August 2022

Date of mailing of the international search report

08 September 2022

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2022/099788

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	103999068	A	20 August 2014	US	2014310369	A1	16 October 2014
				WO	2013095654	A1	27 June 2013
				DE	112011106016	T5	11 September 2014
CN	111782417	A	16 October 2020	None			
CN	108009032	A	08 May 2018	None			
US	2016072906	A1	10 March 2016	US	9742855	B2	22 August 2017

国际检索报告

国际申请号

PCT/CN2022/099788

A. 主题的分类 H04L 67/025 (2022.01) i; G06F 9/54 (2006.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类		
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) H04L; G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNTXT; VEN; CNKI; ENTXTC; IEEE Xplore; IETF: 进程, 应用, 标识, 共享发送队列, 上下文, 封装, 完成通知, 反馈, 响应, RDMA, IB, RoCE, iWARP, shared send queue, SSQ, context, complete, ID, app+, process, consumer		
C. 相关文件		
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
Y	CN 103999068 A (英特尔公司) 2014年8月20日 (2014 - 08 - 20) 说明书第[0008]-[0066]段	1-18
Y	CN 111782417 A (北京计算机技术及应用研究所) 2020年10月16日 (2020 - 10 - 16) 说明书第[0004]-[0043]段	1-18
A	CN 108009032 A (郑州云海信息技术有限公司) 2018年5月8日 (2018 - 05 - 08) 全文	1-18
A	US 2016072906 A1 (MELLANOX TECHNOLOGIES LTD) 2016年3月10日 (2016 - 03 - 10) 全文	1-18
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。 * 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件		
国际检索实际完成的日期 2022年8月31日		国际检索报告邮寄日期 2022年9月8日
ISA/CN的名称和邮寄地址 中国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		受权官员 徐苏宁 电话号码 (86-512)88996068

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/099788

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	103999068	A	2014年8月20日	US	2014310369	A1	2014年10月16日
				WO	2013095654	A1	2013年6月27日
				DE	112011106016	T5	2014年9月11日
CN	111782417	A	2020年10月16日	无			
CN	108009032	A	2018年5月8日	无			
US	2016072906	A1	2016年3月10日	US	9742855	B2	2017年8月22日