(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
2021 年 12 月 9 日 (09.12.2021)

WIPO | PCT

(10) 国际公布号
WO 2021/244240 A1

(51) 国际专利分类号:
H04L 12/801 (2013.01)

(21) 国际申请号: PCT/CN2021/093165

(22) 国际申请日: 2021 年 5 月 11 日 (11.05.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010480552.2 2020年5月30日 (30.05.2020) 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: 严金丰 (YAN, Jinfeng); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 郑合文 (ZHENG, Hewen); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 韩磊 (HAN, Lei); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 刘和洋 (LIU, Heyang); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 陶佩莹 (TAO, Peiying); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 王煜 (WANG, Yu); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 姚学军 (YAO, Xuejun); 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: 北京三高永信知识产权代理有限责任公司 (BEIJING SAN GAO YONG XIN INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国北京市

(54) **Title:** NETWORK CONGESTION CONTROL METHOD AND APPARATUS, DEVICE, SYSTEM, AND STORAGE MEDIUM

(54) **发明名称:** 网络拥塞的控制方法、装置、设备、系统及存储介质

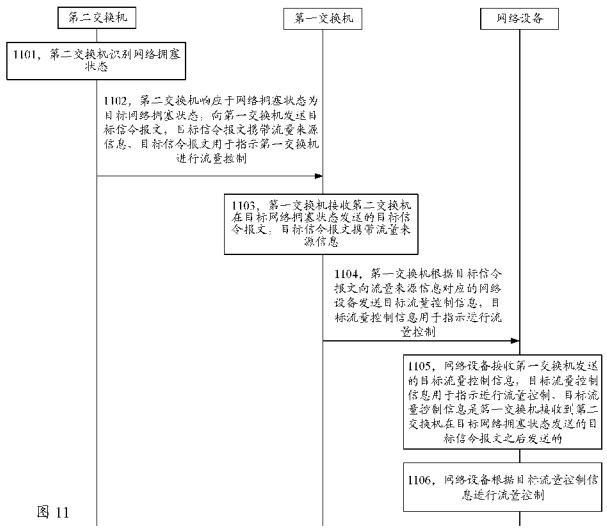

图 11

1101 A second switch recognizes a network congestion state
1102 In response to the network congestion state being a target network congestion state, the second switch sends a target signaling packet to a first switch, the target signaling packet carrying traffic source information, the target signaling packet being used for instructing the first switch to perform traffic control
1103 The first switch receives the target signaling packet sent by the second switch in the target network congestion state, the target signaling packet carrying traffic source information
1104 The first switch sends, according to the target signaling packet, target traffic control information to a network device corresponding to the traffic source information, the target traffic control information being used for instructing to perform traffic control
1105 The network device receives the target traffic control information sent by the first switch, the target traffic control information being used for instructing to perform traffic control, the target traffic control information being sent after the first switch receives the target signaling packet sent by the second switch in the target network congestion state
1106 The network device performs traffic control according to the target traffic control information

(57) **Abstract:** Disclosed in the present application are a network congestion control method and apparatus, a device, a system, and a storage medium. The method comprises: a first switch receives a target signaling packet sent by a second switch in a target network congestion state, the target signaling package carrying traffic source information; the first switch sends, according to the target signaling packet, target traffic control information to a network device corresponding to the traffic source information, the target traffic control information being used for instructing to perform traffic control. Upon receiving a target signaling packet sent by a second switch in a target network congestion state, a first switch sends target traffic control information to a network device corresponding to traffic source information carried in the target signaling packet to instruct to perform traffic control, thereby inhibiting queue backlog on a congestion side and ensuring low service delay without affecting service throughput. The present application can support large-scale RoCE networking and solve the problem of a DCQCN speed control failure in a large-scale high concurrency scene.

WO 2021/244240 A1

[见续页]

海淀区学院路蓟门里和景园A座1单元102室, Beijing 100088 (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, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, 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：本申请公开了网络拥塞的控制方法、装置、设备、系统及存储介质，该方法包括：第一交换机接收第二交换机在目标网络拥塞状态发送的目标信令报文，该目标信令报文携带流量来源信息。第一交换机根据该目标信令报文向流量来源信息对应的网络设备发送目标流量控制信息，该目标流量控制信息用于指示进行流量控制。第一交换机接收到第二交换机在目标网络拥塞状态发送的目标信令报文后，通过向目标信令报文中携带的流量来源信息所对应的网络设备发送目标流量控制信息，以指示进行流量控制，从而抑制拥塞侧的队列积压，保证业务低时延，且不影响业务的吞吐量，能够支持大规模RoCE组网，解决了大规模高并发场景下DCQCN控速失效的问题。

网络拥塞的控制方法、装置、设备、系统及存储介质

本申请要求于2020年05月30日提交的申请号为202010480552.2、发明名称为“网络拥塞的控制方法、装置、设备、系统及存储介质”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，特别涉及一种网络拥塞的控制方法、装置、设备、系统及存储介质。

背景技术

随着高性能计算、分布式存储等应用的出现和广泛使用，对数据中心网络和协议提出高吞吐、低时延、低中央处理单元（central processing unit，CPU）开销的需求。由于传统的传输控制协议/网际协议（transmission control protocol/internet protocol，TCP/IP）协议CPU开销极大，不能很好的满足这些应用的需求。因此，允许用户态的应用程序直接读取和写入远程内存，而无内核干预和内存拷贝发生的远程直接内存访问（remote direct memory access，RDMA）协议应运而生。

目前运用比较广泛的RDMA协议是聚合以太网上的RDMA（RDMA over converged ethernet，RoCE）协议，在RoCE网络中，对网络拥塞进行有效的控制，是降低业务时延，支持大规模RoCE组网的关键。

发明内容

本申请实施例提供了一种网络拥塞的控制方法、装置、设备、系统及存储介质，以解决相关技术提供的问题，技术方案如下：

第一方面，提供了一种网络拥塞的控制方法，以该方法应用于第一交换机为例，该方法包括：第一交换机接收第二交换机在目标网络拥塞状态发送的目标信令报文，该目标信令报文携带流量来源信息。第一交换机根据该目标信令报文向流量来源信息对应的网络设备发送目标流量控制信息，该目标流量控制信息用于指示进行流量控制。

本申请实施例提供的方法，接收到第二交换机在目标网络拥塞状态发送的目标信令报文后，通过向目标信令报文中携带的流量来源信息所对应的网络设备发送目标流量控制信息，以指示进行流量控制，从而抑制拥塞侧的队列积压，保证业务低时延，且不影响业务的吞吐量，能够支持大规模RoCE组网，解决了大规模高并发场景下DCQCN控速失效的问题。

在第一方面的一种可能的实现方式中，根据目标信令报文向流量来源信息对应的网络设备发送目标流量控制信息，包括：根据目标信令报文向流量来源信息对应的网络设备发送第一流量控制信息，第一流量控制信息用于指示网络设备暂停发送目标队列的数据包，目标队列为网络设备的一个或多个队列。

通过第一流量控制信息来指示流量来源信息对应的网络设备暂停发送目标队列的数据包，

能够有效抑制拥塞侧的队列积压，进一步保证业务低时延。

在第一方面的一种可能的实现方式中，根据目标信令报文向流量来源信息对应的网络设备发送第一流量控制信息，包括：根据目标信令报文构造第一基于优先级的流量控制PFC报文，第一PFC报文的时间字段的值为第一值，第一值用于指示第一流量控制信息；向流量来源信息对应的网络设备发送第一PFC报文。

在第一方面的一种可能的实现方式中，接收第二交换机在目标网络拥塞状态发送的目标信令报文，包括：接收第二交换机在目标网络拥塞状态发送的第一信令报文，第一信令报文用于指示发送第一流量控制信息。

在第一方面的一种可能的实现方式中，接收第二交换机在目标网络拥塞状态发送的第一信令报文，包括：接收第二交换机在目标网络拥塞状态发送的第一拥塞通知包CNP报文，第一CNP报文的帧头中的指定字段的值为第一特征值，第一特征值用于指示发送第一流量控制信息。

在第一方面的一种可能的实现方式中，根据目标信令报文向流量来源信息对应的网络设备发送目标流量控制信息，包括：根据目标信令报文向流量来源信息对应的网络设备发送第二流量控制信息，第二流量控制信息用于指示网络设备继续发送目标队列的数据包，目标队列为网络设备的一个或多个队列。

通过第二流量控制信息来指示流量来源信息对应的网络设备继续发送目标队列的数据包，从而不影响业务的吞吐量。

在第一方面的一种可能的实现方式中，根据目标信令报文向流量来源信息对应的网络设备发送第二流量控制信息，包括：根据目标信令报文构造第二基于优先级的流量控制PFC报文，第二PFC报文的时间字段的值为第二值，第二值用于指示第二流量控制信息；向流量来源信息对应的网络设备发送第二PFC报文。

在第一方面的一种可能的实现方式中，接收第二交换机在目标网络拥塞状态发送的目标信令报文，包括：接收第二交换机在目标网络拥塞状态发送的第二信令报文，第二信令报文用于指示发送第二流量控制信息。

在第一方面的一种可能的实现方式中，接收第二交换机在目标网络拥塞状态发送的第二信令报文，包括：接收第二交换机在目标网络拥塞状态发送的第二拥塞通知包CNP报文，第二CNP报文的帧头中的指定字段的值为第二特征值，第二特征值用于指示发送第二流量控制信息。

在第一方面的一种可能的实现方式中，根据目标信令报文向流量来源信息对应的网络设备发送目标流量控制信息，包括：根据目标信令报文携带的流量来源信息确定流量来源端口；通过流量来源端口向流量来源信息对应的网络设备发送目标流量控制信息。

在第一方面的一种可能的实现方式中，通过流量来源端口向流量来源信息对应的网络设备发送目标流量控制信息，包括：通过流量来源端口向流量来源信息对应的网络设备发送第三流量控制信息，所述第三流量控制信息用于指示暂停发送所述流量来源端口所对应的队列的数据包。

在第一方面的一种可能的实现方式中，通过流量来源端口向流量来源信息对应的网络设备发送目标流量控制信息，包括：通过流量来源端口向流量来源信息对应的网络设备发送第四流量控制信息，所述第四流量控制信息用于指示继续发送所述流量来源端口所对应的队列

的数据包。

第二方面，提供了一种方法应用于第二交换机，方法包括：第二交换机识别网络拥塞状态；响应于网络拥塞状态为目标网络拥塞状态，向第一交换机发送目标信令报文，目标信令报文携带流量来源信息，目标信令报文用于指示第一交换机进行流量控制。

本申请实施例提供的方法，通过识别网络拥塞状态，在目标网络拥塞状态，通过目标信令报文指示第一交换机进行流量控制，从而抑制拥塞侧的队列积压，保证业务低时延，且不影响业务的吞吐量，能够支持大规模RoCE组网，解决大规模高并发场景下DCQCN控速失效的问题。

在第二方面的一种可能的实现方式中，目标信令报文包括第一信令报文或第二信令报文，响应于网络拥塞状态为目标网络拥塞状态，向第一交换机发送目标信令报文，包括：响应于网络拥塞状态为目标网络拥塞状态，且当前的队列长度大于第一阈值，向第一交换机发送第一信令报文，第一信令报文用于指示第一交换机发送第一流量控制信息，第一流量控制信息用于指示流量来源信息对应的网络设备暂停发送目标队列的数据包，目标队列为网络设备的一个或多个队列；或者，

响应于网络拥塞状态为目标网络拥塞状态，且当前的队列长度小于第二阈值，向第一交换机发送第二信令报文，第二信令报文用于指示第一交换机发送第二流量控制信息，第二流量控制信息用于指示网络设备继续发送目标队列的数据包，第二阈值小于第一阈值。

在第二方面的一种可能的实现方式中，向第一交换机发送第一信令报文之前，还包括：获取第一CNP报文，将第一CNP报文的帧头中的指定字段的值设为第一特征值，将第一CNP报文作为第一信令报文；

向第一交换机发送第二信令报文之前，还包括：获取第二CNP报文，将第二CNP报文的帧头中的指定字段的值设为第二特征值，将第二CNP报文作为第二信令报文。

通过CNP报文来构造第一信令报文或第二信令报文只是一种示例，还可以通过其他类型的报文格式来构造第一信令报文或第二信令报文，本申请实施例对此不进行限定。

在第二方面的一种可能的实现方式中，识别网络拥塞状态，包括：读取当前的队列长度及显式拥塞通知ECN阈值范围，ECN阈值范围用于指示添加ECN标识的概率，ECN标识用于指示网络发生拥塞；根据当前的队列长度及ECN阈值范围识别网络拥塞状态。

通过识别不同网络拥塞状态，便于后续基于不同的网络拥塞状态进行相应的网络拥塞的控制。

第三方面，提供了一种网络拥塞的控制方法，该方法应用于网络设备，方法包括：网络设备接收第一交换机发送的目标流量控制信息，目标流量控制信息用于指示进行流量控制，所示目标流量控制信息是第一交换机接收到第二交换机在目标网络拥塞状态发送的目标信令报文之后发送的；根据目标流量控制信息进行流量控制。

本申请实施例提供的方法，接收到第一交换机发送的目标流量控制信息后，基于该目标流量控制信息进行流量控制，从而抑制拥塞侧的队列积压，保证业务低时延，且不影响业务的吞吐量，能够支持大规模RoCE组网，解决大规模高并发场景下DCQCN控速失效的问题。

在第三方面的一种可能的实现方式中，接收第一交换机发送的目标流量控制信息，包括：

接收第一交换机发送的第一流量控制信息，第一流量控制信息用于指示暂停发送目标队列的数据包，目标队列为网络设备的一个或多个队列；

根据目标流量控制信息进行流量控制，包括：根据第一流量控制信息暂停发送目标队列的数据包。

在第三方面的一种可能的实现方式中，接收第一交换机发送的第一流量控制信息，包括：接收第一交换机发送的第一PFC报文，第一PFC报文的时间字段的值为第一值，第一值用于指示第一流量控制信息；

根据第一流量控制信息暂停发送目标队列的数据包，包括：根据第一PFC报文的时间字段的值确定暂停发送数据包的时间长度，在时间长度内暂停发送目标队列的数据包。

在第三方面的一种可能的实现方式中，接收第一交换机发送的目标流量控制信息，包括：接收第一交换机发送的第二流量控制信息，第二流量控制信息用于指示继续发送目标队列的数据包，目标队列为网络设备的一个或多个队列；

根据目标流量控制信息进行流量控制，包括：根据第二流量控制信息继续发送目标队列的数据包。

在第三方面的一种可能的实现方式中，接收第一交换机发送的第二流量控制信息，包括：接收第一交换机发送的第二PFC报文，第二PFC报文的时间字段的值为第二值，第二值用于指示第二流量控制信息；

根据第二流量控制信息继续发送目标队列的数据包，包括：根据第二PFC报文的时间字段的值继续发送目标队列的数据包。

第四方面，提供了一种网络拥塞的控制装置，装置包括：

接收模块，用于接收第二交换机在目标网络拥塞状态发送的目标信令报文，目标信令报文携带流量来源信息；

发送模块，用于根据目标信令报文向流量来源信息对应的网络设备发送目标流量控制信息，目标流量控制信息用于指示进行流量控制。

在第四方面的一种可能的实现方式中，发送模块，用于根据目标信令报文向流量来源信息对应的网络设备发送第一流量控制信息，第一流量控制信息用于指示网络设备暂停发送目标队列的数据包，目标队列为网络设备的一个或多个队列。

在第四方面的一种可能的实现方式中，发送模块，用于根据目标信令报文构造第一基于优先级的流量控制PFC报文，第一PFC报文的时间字段的值为第一值，第一值用于指示第一流量控制信息；向流量来源信息对应的网络设备发送第一PFC报文。

在第四方面的一种可能的实现方式中，接收模块，用于接收第二交换机在目标网络拥塞状态发送的第一信令报文，第一信令报文用于指示发送第一流量控制信息。

在第四方面的一种可能的实现方式中，接收模块，用于接收第二交换机在目标网络拥塞状态发送的第一拥塞通知包CNP报文，第一CNP报文的帧头中的指定字段的值为第一特征值，第一特征值用于指示发送第一流量控制信息。

在第四方面的一种可能的实现方式中，发送模块，用于根据目标信令报文向流量来源信息对应的网络设备发送第二流量控制信息，第二流量控制信息用于指示网络设备继续发送目标队列的数据包，目标队列为网络设备的一个或多个队列。

在第四方面的一种可能的实现方式中，发送模块，用于根据目标信令报文构造第二基于优先级的流量控制PFC报文，第二PFC报文的时间字段的值为第二值，第二值用于指示第二流量控制信息；向流量来源信息对应的网络设备发送第二PFC报文。

在第四方面的一种可能的实现方式中，接收模块，用于接收第二交换机在目标网络拥塞状态发送的第二信令报文，第二信令报文用于指示发送第二流量控制信息。

在第四方面的一种可能的实现方式中，接收模块，用于接收第二交换机在目标网络拥塞状态发送的第二拥塞通知包CNP报文，第二CNP报文的帧头中的指定字段的值为第二特征值，第二特征值用于指示发送第二流量控制信息。

在第四方面的一种可能的实现方式中，发送模块，用于根据目标信令报文携带的流量来源信息确定流量来源端口；通过流量来源端口向流量来源信息对应的网络设备发送目标流量控制信息。

在第四方面的一种可能的实现方式中，发送模块，用于通过流量来源端口向流量来源信息对应的网络设备发送第三流量控制信息，所述第三流量控制信息用于指示暂停发送所述流量来源端口所对应的队列的数据包。

在第四方面的一种可能的实现方式中，发送模块，用于通过流量来源端口向流量来源信息对应的网络设备发送第四流量控制信息，所述第四流量控制信息用于指示继续发送所述流量来源端口所对应的队列的数据包。

第五方面，提供了一种网络拥塞的控制装置，装置包括:

识别模块，用于识别网络拥塞状态;

发送模块，用于响应于网络拥塞状态为目标网络拥塞状态，向第一交换机发送目标信令报文，目标信令报文携带流量来源信息，目标信令报文用于指示第一交换机进行流量控制。

在第五方面的一种可能的实现方式中，目标信令报文包括第一信令报文或第二信令报文，发送模块，用于响应于网络拥塞状态为目标网络拥塞状态，且当前的队列长度大于第一阈值，向第一交换机发送第一信令报文，第一信令报文用于指示第一交换机发送第一流量控制信息，第一流量控制信息用于指示流量来源信息对应的网络设备暂停发送目标队列的数据包，目标队列为网络设备的一个或多个队列；或者，

响应于网络拥塞状态为目标网络拥塞状态，且当前的队列长度小于第二阈值，向第一交换机发送第二信令报文，第二信令报文用于指示第一交换机发送第二流量控制信息，第二流量控制信息用于指示网络设备继续发送目标队列的数据包，第二阈值小于第一阈值。

在第五方面的一种可能的实现方式中，装置还包括:

获取模块，用于获取第一CNP报文，将第一CNP报文的帧头中的指定字段的值设为第一特征值，将第一CNP报文作为第一信令报文;

或者，获取模块，用于获取第二CNP报文，将第二CNP报文的帧头中的指定字段的值设为第二特征值，将第二CNP报文作为第二信令报文。

在第五方面的一种可能的实现方式中，识别模块，用于读取当前的队列长度及显式拥塞通知ECN阈值范围，ECN阈值范围用于指示添加ECN标识的概率，ECN标识用于指示网络发生拥塞；根据当前的队列长度及ECN阈值范围识别网络拥塞状态。

在第五方面的一种可能的实现方式中，目标网络拥塞状态包括ECN失效状态或拥塞通知

包CNP失效状态，识别模块，用于响应于当前的队列长度大于参考范围的最大值，且未补充CNP报文，则网络拥塞状态为ECN失效状态，参考范围基于ECN阈值范围确定；响应于当前的队列长度大于参考范围的最大值，且已补充CNP报文，则网络拥塞状态为CNP失效状态。

第六方面，提供了一种网络拥塞的控制装置，装置包括：

接收模块，用于接收第一交换机发送的目标流量控制信息，目标流量控制信息用于指示进行流量控制，所示目标流量控制信息是第一交换机接收到第二交换机在目标网络拥塞状态发送的目标信令报文之后发送的；

控制模块，用于根据目标流量控制信息进行流量控制。

在第六方面的一种可能的实现方式中，接收模块，用于接收第一交换机发送的第一流量控制信息，第一流量控制信息用于指示暂停发送目标队列的数据包，目标队列为网络设备的一个或多个队列；

控制模块，用于根据第一流量控制信息暂停发送目标队列的数据包。

在第六方面的一种可能的实现方式中，接收模块，用于接收第一交换机发送的第一PFC报文，第一PFC报文的时间字段的值为第一值，第一值用于指示第一流量控制信息；

控制模块，用于根据第一PFC报文的时间字段的值确定暂停发送数据包的时间长度，在时间长度内暂停发送目标队列的数据包。

在第六方面的一种可能的实现方式中，接收模块，用于接收第一交换机发送的第二流量控制信息，第二流量控制信息用于指示继续发送目标队列的数据包，目标队列为网络设备的一个或多个队列；

控制模块，用于根据第二流量控制信息继续发送目标队列的数据包。

在第六方面的一种可能的实现方式中，接收模块，用于接收第一交换机发送的第二PFC报文，第二PFC报文的时间字段的值为第二值，第二值用于指示第二流量控制信息；

控制模块，用于根据第二PFC报文的时间字段的值继续发送目标队列的数据包。

在第一方面至第六方面的一种可能的实现方式中，目标网络拥塞状态包括显式拥塞通知（explicit congestion notification，ECN）失效状态或拥塞通知包（congestion notification packet，CNP）失效状态；

ECN失效状态是指第二交换机当前的队列长度大于参考范围的最大值，且未补充CNP报文的状态；CNP失效状态是指第二交换机当前的队列长度大于参考区域的最大值，且已补充CNP报文的状态；参考范围基于ECN阈值范围确定，ECN阈值范围用于指示添加ECN标识的概率，ECN标识用于指示网络发生拥塞。

还提供一种网络拥塞的控制设备，该设备包括：存储器及处理器，所述存储器中存储有至少一条指令，所述至少一条指令由所述处理器加载并执行，以实现上述第一方面任一所述的网络拥塞的控制方法。

还提供一种网络拥塞的控制设备，该设备包括：存储器及处理器，所述存储器中存储有至少一条指令，所述至少一条指令由所述处理器加载并执行，以实现上述第二方面任一所述的网络拥塞的控制方法。

还提供一种网络拥塞的控制设备，该设备包括：存储器及处理器，所述存储器中存储有

至少一条指令，所述至少一条指令由所述处理器加载并执行，以实现上述第三方面任一所述的网络拥塞的控制方法。

还提供了一种网络拥塞的控制系统，该系统包括上述三种设备。

还提供了一种计算机可读存储介质，所述存储介质中存储有至少一条指令，所述指令由处理器加载并执行以实现如上第一方面至第三方面中任一所述的网络拥塞的控制方法。

提供了另一种通信装置，该装置包括：收发器、存储器和处理器。其中，该收发器、该存储器和该处理器通过内部连接通路互相通信，该存储器用于存储指令，该处理器用于执行该存储器存储的指令，以控制收发器接收信号，并控制收发器发送信号，并且当该处理器执行该存储器存储的指令时，使得该处理器执行第一方面或第一方面的任一种可能的实施方式中的方法。或者，当该处理器执行该存储器存储的指令时，使得该处理器执行第二方面或第二方面的任一种可能的实施方式中的方法。或者，当该处理器执行该存储器存储的指令时，使得该处理器执行第三方面或第三方面的任一种可能的实施方式中的方法。

作为一种示例性实施例，所述处理器为一个或多个，所述存储器为一个或多个。

作为一种示例性实施例，所述存储器可以与所述处理器集成在一起，或者所述存储器与处理器分离设置。

在具体实现过程中，存储器可以为非瞬时性（non-transitory）存储器，例如只读存储器（read only memory，ROM），其可以与处理器集成在同一块芯片上，也可以分别设置在不同的芯片上，本申请实施例对存储器的类型以及存储器与处理器的设置方式不做限定。

提供了一种计算机程序（产品），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被计算机运行时，使得所述计算机执行上述各方面中的方法。

提供了一种芯片，包括处理器，用于从存储器中调用并运行所述存储器中存储的指令，使得安装有所述芯片的通信设备执行上述各方面中的方法。

提供另一种芯片，包括：输入接口、输出接口、处理器和存储器，所述输入接口、输出接口、所述处理器以及所述存储器之间通过内部连接通路相连，所述处理器用于执行所述存储器中的代码，当所述代码被执行时，所述处理器用于执行上述各方面中的方法。

附图说明

图 1 为相关技术中提供的一种网络架构示意图；

图 2 为本申请实施例提供的 ECN 标识模型示意图；

图 3 为本申请实施例提供的流量模型示意图；

图 4 为本申请实施例提供的 ECN 打标率与队列积压关系示意图；

图 5 为本申请实施例提供的流传输过程示意图；

图 6 为本申请实施例提供的流量模型示意图；

图 7 为本申请实施例提供的服务器 qp 数与时延的关系示意图；

图 8 为本申请实施例提供的流量模型示意图；

图 9 为相关技术提供的流速率与 CNP 数目变化示意图；

图 10 为本申请实施例提供的数据中心网络的结构示意图；

图 11 为本申请实施例提供的网络拥塞的控制方法流程图；

图 12 为本申请实施例提供的应用场景示意图；

图 13 为本申请实施例提供的网络拥塞的控制装置的结构示意图；
图 14 为本申请实施例提供的网络拥塞的控制装置的结构示意图；
图 15 为本申请实施例提供的网络拥塞的控制装置的结构示意图；
图 16 为本申请实施例提供的网络拥塞的控制设备的结构示意图。

具体实施方式

本申请的实施方式部分使用的术语仅用于对本申请的实施例进行解释，而非旨在限定本申请。

由于传统的TCP/IP协议CPU开销极大，不能很好的满足数据中心网络和协议提出的高吞吐、低时延、低CPU开销的需求，因此，RDMA协议应运而生。

目前运用比较广泛的RDMA协议是RoCE协议，RoCE协议有RoCE1和RoCE2两个版本。其中，RoCE1是基于以太网链路层实现的RDMA协议，RoCEv2则是基于以太网TCP/IP协议中用户数据报协议（user datagram protocol，UDP）层实现的RDMA协议。无论是哪个版本的RoCE协议，在RoCE网络中，对网络拥塞进行有效的控制，是降低业务时延，支持大规模RoCE组网的关键。

对此，相关技术提供了一种数据中心量化拥塞通知（data center quantized congestion notification，DCQCN）算法来对RoCE网络进行拥塞控制。以图1所示的网络架构为例，该网络架构包括发送端网络接口控制器（network interface controller，NIC）、交换机及接收端NIC。其中，发送端NIC为响应点（Reaction Point，RP），交换机为拥塞点(congestion point，CP)，接收端NIC为通知点（Notification Point，NP）。

在交换机侧，当一个数据包到达交换机的出端口时，交换机查看交换机的出端口的队列缓存长度。如果交换机的出口端的队列缓存长度超过给定阈值，则对这个数据包按标记概率标记显式拥塞通知（explicit congestion notification，ECN）标识。ECN最初在RFC 3168中定义，交换机会在检测到拥塞时，通过在IP头部嵌入一个拥塞指示器和在TCP头部嵌入一个拥塞确认实现。RoCEv2标准定义了RoCEv2拥塞管理(RCM)。启用了ECN之后，交换机一旦检测到RoCEv2流量出现了拥塞，会在数据包的IP头部ECN域进行标记。

其中，拥塞指示器被目的终端节点按照存在于IB数据段中的基本传输头(base transport header，BTH)中的前向显式拥塞通告（forward explicit congestion notification，FECN）拥塞指示标识来解释意义。换句话说，当被ECN标记过的数据包到达原本要到达的目的地时，拥塞通知就会被反馈给源节点，源节点再通过对有问题的队列对（queue pairs，QP）进行网络数据包的速率限制来回应拥塞通知。

以基于图2所示的ECN标识模型对数据包按标记概率标记ECN为例，当交换机的出口端的队列缓存长度小于给定下限阈值Kmin时，则标记概率为0%；当交换机的出口端的队列缓存长度大于给定上限阈值Kmax时，则标记概率为100%；当交换机的出口端的队列缓存长度介于Kmin和Kmax之间时，标记概率随队列缓存长度线性增加。

在接收端侧，当携带ECN标识的数据包到达接收端NIC时，表明网络中发生了拥塞，则接收端NIC针对该携带ECN标识的报文返回拥塞通知包（congestion notification packet，CNP）报文，以通过该CNP报文将拥塞信息传递给发送端NIC。示例性地，如果某条数据流中携带ECN标识的数据包到达接收端NIC，且在之前的参考时间段内没有相应CNP报文被发送，则

接收端NIC立刻发送一个CNP报文。其中，参考时间段的长度为N微妙，N的值可配置为0，即接收端NIC每收到一个携带ECN标识的报文就返回一个CNP报文。

CNP报文作为拥塞控制报文，也会存在延迟和丢包，从发送端到接收端经过的每一跳设备、每一条链路都会有一定的延迟，会最终加大发送端接收到CNP报文的时间。与此同时，交换机的出端口下的拥塞也会逐步增多，若发送端不能及时降速，仍然可能造成丢包。因此，在发送端侧，发送端NIC控制每条数据流的发送速率。例如，通过收到的CNP报文来触发对数据流的降速控制，通过时间定时器和字节计数器来触发对数据流的升速控制，升速控制与降速控制相互独立。

通过上述过程不难看出，DCQCN 算法的升速控制基于发送端 NIC 的定时器触发，降速控制基于接收端 NIC 发送的 CNP 报文触发。当升速控制和降速控制不能协调工作时，将会导致 DCQCN 控速失效。尤其是网络数据流规模增大时，更容易发生 DCQCN 控速失效的问题。

例如，随着数据流数目的增加，每条数据流分到的带宽随之减小，单位时间内发出的 CNP 报文数量也随之变少，而 CNP 报文间隔随之变大。比如发送端带宽限制每秒发出两千个数据包，每个数据包的间隔是 500us，这样即使每个数据包都标记 ECN 标识，那么接收端返回的 CNP 报文间隔最小也是 500us。但是 DCQCN 源端的升速周期是 300us，这时 CNP 的间隔大于升速周期。当拥塞发生时，会使得发送端的数据流在拥塞状态下仍错误的升速，导致DCQCN控速失败。

以图 3 所示的数据流量模型，由单个柜顶（top of rack，ToR）交换机实现多个发送端到一个接收端的数据流传输，且以 TCP 协议的报文与 RoCE 协议的报文按照 9:1 的带宽比例进行传输为例。数据流数、ECN 和队列积压的关系如图 4 所示。由图 4 可以看出，当数据流数目较少，例如数据流数目低于 4 时，DCQCN 工作的很好，ECN 打标率随着数据流数目的增加而增加，队列缓存长度控制的很低，业务时延也很小。但是，当数据流数目达到一定程度后，例如图 4 所示的数据流数目超过 4 以后，即使 ECN 达标率已经达到 100%了，但是队列缓存长度突变为 MB 级，业务时延劣化为 ms 级，这时 ECN 失效。

为了解决ECN失效的问题，相关技术还提出了一种在交换机侧记录交换机的出端口的拥塞程度和收到的CNP报文时间，根据交换机的出端口的拥塞程度来确定是否补充CNP报文的方法。以图5所示的数据流传输示意图为例，当交换机的出端口进入拥塞状态时，则交换机检查经过该出端口的数据流的CNP报文的时间间隔。如果在DCQCN的升速周期内没有CNP报文经过，则交换机侧补充一个CNP报文发往发送端并更新CNP报文时间。如果在DCQCN的升速周期内有CNP报文经过，则交换机只更新CNP报文经过的时间，不再补充CNP报文。当交换机的出端口不拥塞时，则只更新CNP报文经过的时间，不再补充CNP报文。

然而，上述补充CNP报文的方式，在数据流进一步增加时仍会出现队列积压，时延劣化的问题。例如，以图6所示的流量模型，由单个ToR交换机实现多个发送端到一个接收端的数据流传输，且以TCP协议的报文与RoCE协议的报文按照9:1的带宽比例进行传输为例，不断增加每个服务器的数据流数（qp数）。服务器的qp数与时延的关系可如图7的测试结果所示，可以看出当每个服务器输出的数据流数小于9时，时延较小。但是当服务器输出的数据流数大于9时，时延突变劣化为ms级，导致该方案仍然存在CNP失效。

为便于理解，以图8所示的流量模型做进一步分析。如图8所示，流量模型中的发送端1

（sender1）和发送端2（sender2）同时向接收端（receiver）发送数据流，receiver可根据需要构造任意多的CNP报文给sender1和sender2。服务器每秒处理的CNP报文数与流速率的关系可如图9所示，服务器的速率随着单位时间内处理的CNP报文数增加而减少，但是当速率下降到一定值后，不再下降，称这种现象为CNP失效。因此当大规模高并发场景下，每个数据流分配的带宽小于网卡能控速的极限值时，CNP失效必然出现，这时队列会压不住，业务时延会很大。

例如，以在图6所示的流量模型下进一步测试为例，得到如表1所示的测试结果。从下面表1所示的测试结果可以看出，在任何比例下，CNP失效都会出现，这时队列积压很深，业务时延很大。

表1

TCP: RoCE带宽配比	数据流数目	端口堆积	时延
90:10	7*11	19446KB	31637.50us
70:30	7*40	19757KB	10360.06us
50:50	7*65	18845KB	5987.04us
30:70	7*85	16829KB	3824.71us
10:90	7*115	18160KB	3206.68us

对此，为了解决CNP失效的问题，也即解决大规模高并发场景下DCQCN控速失效的问题，本申请实施例提供了一种网络拥塞的控制方法。该方法将网络的拥塞状态分为ECN正常状态、ECN失效状态和CNP失效状态三种，通过识别当前的网络拥塞的状态，根据网络拥塞的状态按照路径级触发式产生优先级的流量控制（priority-based flow control，PFC）报文，通过PFC控制流量，从而抑制拥塞侧的队列积压，保证业务低时延的同时不影响业务的吞吐，进而支持大规模RoCE组网。

示例性地，本申请实施例提供的方法可应用于图10所示的数据中心网络中。该数据中心网络是一种Clos网络架构，Clos网络架构是一种交换架构，能够做到可重排无阻塞、可扩展，用于高性能计算、高性能分布式存储、大数据、人工智能等。该Clos网络架构中包括：第一交换机、第二交换机和源服务器。第一交换机如图10中T所示的ToR交换机，第二交换机如图10中与第一交换机不同的T所示的ToR交换机，源服务器如图10中H所示的服务器（Host）。第一交换机可作为源端交换机，第二交换机作为拥塞侧交换机，源服务器作为流量来源对应的网络设备。其中，各个交换机的功能如下。

第二交换机（T）：用于读取队列长度、当前ECN的阈值范围（如水线kmin、kmax值）。如果队列长度在kmin-kmax附近，则识别网络拥塞状态为ECN正常状态；如果队列长度远大于kmax，则识别网络拥塞状态为ECN失效状态，并智能添加CNP报文；如果添加CNP报文后，队列长度仍远大于kmax，则识别网络拥塞状态为CNP失效状态。ECN失效状态和CNP失效状态可以作为目标网络拥塞状态，识别当前网络拥塞状态为目标网络拥塞状态时，该第二交换机构造目标信令报文，通知源端交换机即第一交换机当前进入目标网络拥塞状态。例如，进入CNP失效状态，进行补PFC操作。例如，如果队列长度大于给定上限值，则第二交换机构造第一信令报文发往第一交换机，以第一信令报文为CNP报文为例，第二交换机在CNP报文的帧头的指定字段例如保留（reserved）字段填入第一特征值，例如1。如果队列长度小于给

定下限值，则第二交换机构造第二信令报文发往第一交换机，仍以第一信令报文为CNP报文为例，第二交换机在CNP报文的帧头的reserved字段填入第二特征值，例如2。

第一交换机（T）：第一交换机接收并解析第二交换机发送的目标信令报文，发现通告对象是第一交换机自身，如果reserved字段为第一特征值1，标识该目标信令报文是第一信令报文，则第一交换机构造第一PFC报文发送给对应的源服务器（H），以触发对应的源服务器暂停发送目标队列的数据包。例如，第一交换机构造的第一PFC报文为xoff PFC报文。如果reserved字段为第二特征值2，标识该目标信令报文是第二信令报文，则第一交换机构造第二PFC报文发送给对应的源服务器，以触发对应的源服务器继续发送目标队列的数据包。例如，第一交换机构造的第二PFC报文为xon PFC报文。其中，目标队列是源服务器的一个或多个队列，可通过PFC报文来指示。此外，第一交换机也可通过流量来源端口来进行流量控制。例如，通过该流量来源端口向流量来源信息对应的源服务器发送对应的目标流量控制信息，以指示暂停或继续发送该流量来源端口所对应的队列的数据包。

该Clos网络架构中除了包括第一交换机和第二交换机外，在示例性实施例中，第一交换机与第二交换机之间还具有中间交换机，该中间交换机如图10中L所示的叶（Leaf）交换机以及图10中S所示的汇聚（Spine）交换机。

中间交换机（L、S）：中间交换机接收并解析第二交换机发送的目标信令报文，如果通告对象不是中间交换机自身，则中间交换机不作特殊操作，将该目标信令报文继续转发。

需要说明的是，图10所示的各个交换机的数量仅是示例性说明，本申请实施例不对各个交换机的数量进行限定。此外，源服务器仅是一种作为流量来源对应的网络设备的举例，除了源服务器，还可以是其他交换机，本申请实施例不对流量来源对应的网络设备的类型进行限定。再有，本申请实施例提供的方法不仅仅局限于图10所示数据中心网络场景，也可应用于其他使用DCQCN技术的场景，本申请实施例对应用场景不进行限定。

接下来，对本申请实施例提供的网络拥塞的控制方法进行说明。该方法可通过第一交换机、第二交换机及流量来源对应的网络设备之间的交互实现，示例性地，网络设备为源服务器。参见图11，本申请实施例提供的方法包括如下几个步骤。

1101，第二交换机识别网络拥塞状态。

在示例性实施例中，第二交换机识别网络拥塞状态，包括但不限于如下1101A和1101B两个过程。

1101A：读取当前的队列长度及ECN的阈值范围，ECN的阈值范围用于指示添加ECN标识的概率，ECN标识用于指示网络发生拥塞。

在示例性实施例中，第二交换机为检测拥塞的交换机，当前的队列长度为第二交换机的出端口当前缓存数据的队列长度。当一个数据包到达第二交换机的出端口时，该第二交换机查看第二交换机的出端口缓存数据的队列长度。如果第二交换机的出口端缓存数据的队列长度超过给定阈值，则对这个数据包按标记概率标记ECN标识。因此，ECN标识能够用于指示网络拥塞。本申请实施例不对给定阈值的大小进行限定，例如可基于应用场景设置，也可以基于经验设置。

对数据包标记ECN标识时，标记概率基于ECN的阈值范围和队列长度来确定，例如图2所示的ECN标记模型所示，该ECN的阈值范围是指给定下限阈值Kmin与给定上限阈值Kmax

构成的范围。该ECN的阈值范围用于指示添加ECN标识的概率，网络拥塞越严重，具有ECN标识的报文数量越多。

本申请实施例提供的方法通过当前的队列长度和ECN的阈值范围来识别网络拥塞状态。为便于理解，以该方法应用于图12所示的应用场景为例。如图12所示，该网络架构包括4个L交换机，分别为交换机L1、交换机L2、交换机L3和交换机L4，包括2个S交换机，分别为交换机S1和交换机S2。以交换机L4侧发生拥塞为例，该交换机L4即为第二交换机，读取当前的队列长度、当前ECN的kmin和kmax值。

1101B: 第二交换机根据当前的队列长度及ECN的阈值范围识别网络拥塞状态。

示例性地，网络拥塞状态包括但不限于ECN正常状态、ECN失效状态和拥塞通知包CNP失效状态。

在示例性实施例中，根据当前的队列长度及ECN的阈值范围识别网络拥塞状态，包括但不限于如下三种识别结果：

识别结果一： 响应于当前的队列长度在参考范围内，则网络拥塞状态为ECN正常状态。

其中，本申请实施例不对参考范围进行限定，可基于经验设置，也可基于应用场景进行调整。例如，参考范围基于ECN阈值范围确定。以将0与1.5倍的最大值Kmax之间的范围作为参考范围为例，如果当前的队列长度在0-1.5Kmax的参考范围内，则网络拥塞状态为ECN正常状态。

识别结果二： 响应于当前的队列长度大于参考范围的最大值，且未补充CNP报文，则网络拥塞状态为ECN失效状态。

例如，仍以0与1.5倍的最大值Kmax之间的区域作为参考范围为例，参考范围的最大值为1.5Kmax，如果当前的队列长度大于1.5Kmax，且未补充CNP报文，则网络拥塞状态为ECN失效状态。针对识别结果是网络拥塞状态为ECN失效状态的情况，第二交换机可补充CNP报文。例如，在第二交换机侧记录第二交换机的出端口的拥塞程度和收到的CNP报文时间，根据第二交换机的出端口的拥塞程度来确定是否补充CNP报文。补充CNP报文的方式参见上述图5的相关描述，此处不再赘述。

识别结果三： 响应于当前的队列长度大于参考范围的最大值，且已补充CNP报文，则网络拥塞状态为CNP失效状态。

例如，仍以0与1.5倍的最大值Kmax之间的范围作为参考范围为例，参考范围的最大值为1.5Kmax，如果当前的队列长度大于1.5Kmax，且已补充CNP报文，则网络拥塞状态为CNP失效状态。

综上，以如图12所示的应用场景为例，交换机L4发生拥塞，读取当前的队列长度、当前ECN的kmin和kmax值后，如果当前的队列长度在kmin-kmax的参考范围附近，例如参考区域为[0,1.5*kmax]，则交换机L4识别当前网络拥塞状态为ECN正常状态。如果队列长度远大于1.5kmax，例如大于3*kmax，且未补充CNP报文，则交换机L4识别当前网络拥塞状态为ECN失效状态，打开智能补充CNP报文功能。如果队列长度还是远大于1.5kmax，例如大于3*kmax，则交换机L4识别当前网络拥塞状态为CNP失效状态。

1102，第二交换机响应于网络拥塞状态为目标网络拥塞状态，向第一交换机发送目标信令报文，目标信令报文携带流量来源信息，目标信令报文用于指示第一交换机进行流量控制。

在示例性实施例中，目标信令报文包括第一信令报文或第二信令报文，响应于网络拥塞

状态为目标网络拥塞状态，向第一交换机发送目标信令报文，包括但不限于如下两种发送情况。

发送情况一：响应于网络拥塞状态为目标网络拥塞状态，且当前的队列长度大于第一阈值，向第一交换机发送第一信令报文，第一信令报文用于指示第一交换机发送第一流量控制信息，该第一流量控制信息用于指示流量来源信息对应的网络设备暂停发送目标队列的数据包，目标队列为网络设备的一个或多个队列。

针对发送情况一，第一阈值的大小可基于经验设置，也可基于应用场景设置，还可在实施方法过程中进行调整。当前的队列长度大于第一阈值，说明拥塞情况较为严重，需要启动流量控制。例如，通过第一信令报文来指示第一交换机进行第一类流量控制，即指示第一交换机发送第一流量控制信息。其中，第一信令报文可是任意格式的报文，能够指示第一交换机进行第一类流量控制即可。

示例性地，向第一交换机发送第一信令报文之前，还包括：获取第一CNP报文，将第一CNP报文的帧头中的指定字段的值设为第一特征值，将第一CNP报文作为第一信令报文。

以CNP报文的格式如表2所示为例，表2中的第一行为各个字段的比特数，第二行为各个字段的名称，第三行为各个字段的值。该CNP报文的帧头包括8比特的操作码（pocode）字段、1比特的请求事件（solicited event，SE）字段、1比特的迁移请求（migreq，M）字段、2比特的填充计数（Pad Count）字段、4比特的头版本（Head version）字段、16比特的分区键（Partition Key，P_KEY）字段、8比特的保留（Reserved）字段、24比特的目的队列对（DestQP）字段、1比特的确认请求（Ack request）字段、7比特的保留字段以及24比特的数据包序列号(packet sequence number，PSN)字段。

表2

8	1	1	2	4	16	8	24	1	7	24
pocode	SE	M	Pad Count	Head version	P_KEY	Reserved	DestQP	Ack request	Reserved	PSN
0x81	0	0	00	0000	0xffff	01000000	DQP	0		全0

在本申请实施例中，将CNP报文的第2个保留字段即7比特的保留字段的值设为第一特征值。例如，在表2所示的CNP报文的帧头基础上，将CNP报文的帧头中的7比特的保留字段的值设为第一特征值，以该第一特征值为1，用0000001表示为例，该CNP报文的帧头如表3所示。

表3

8	1	1	2	4	16	8	24	1	**7**	24
pocode	SE	M	Pad Count	Head version	P_KEY	Reserved	DestQP	Ack request	**Reserved**	PSN
0x81	0	0	00	0000	0xffff	01000000	DQP	0	**0000001**	全0

发送情况二：响应于网络拥塞状态为目标网络拥塞状态，且当前的队列长度小于第二阈值，向第一交换机发送第二信令报文，第二信令报文用于指示第一交换机发送第二流量控制信息，第二流量控制信息用于指示网络设备继续发送目标队列的数据包，第二阈值小于第一阈值。

针对发送情况二，第二阈值的大小可基于经验设置，也可基于应用场景设置，还可在实

施方法过程中进行调整。第二阈值小于第一阈值，当前的队列长度小于第二阈值，说明拥塞情况得到缓解，需要启动另一类流量控制。例如，通过第二信令报文来指示第一交换机进行第二类流量控制。其中，第二信令报文可是任意格式的报文，能够指示第一交换机进行第二类流量控制即可。

示例性地，向第一交换机发送第二信令报文之前，还包括：获取第二CNP报文，将第二CNP报文的帧头中的指定字段的值设为第二特征值，将第二CNP报文作为第二信令报文。

仍以上面表2所示的CNP报文的帧头的格式为例，将CNP报文的帧头中的7比特的保留字段的值设为第二特征值，以该第二特征值为2，用0000010表示为例，该CNP报文的帧头如表4所示。

表4

8	1	1	2	4	16	8	24	1	7	24
pocode	SE	M	Pad Count	Head version	P_KEY	Reserved	DestQP	Ack request	**Reserved**	PSN
0x81	0	0	00	0000	0xffff	01000000	DQP	0	**0000010**	全0

仍以图12所示的应用场景为例，当交换机L4识别当前的网络拥塞状态为目标网络拥塞状态时，则比较当前的队列长度与给定门限第一阈值thh和第二阈值thl。如果当前的队列长度大于thh，则交换机L4构造第一信令报文发往第一交换机。例如，使用CNP报文作为信令报文，在CNP报文的帧头的reserved字段填入特征值1。如果当前的队列长度小于thl，则交换机L4构造第二信令报文发往第一交换机。例如，使用CNP报文作为信令报文，在CNP报文的帧头的reserved字段填入特征值2。CNP报文的帧头中reserved字段为保留字段，默认为全0。本申请实施例利用该保留字段构造第一、第二信令报文，通告第一交换机做不同动作，即采用不同的流量控制方式。

在示例性实施例中，目标网络拥塞状态包括ECN失效状态或CNP失效状态；其中，ECN失效状态是指第二交换机当前的队列长度大于参考范围的最大值，且未补充CNP报文的状态；CNP失效状态是指第二交换机当前的队列长度大于参考区域的最大值，且已补充CNP报文的状态。

1103，第一交换机接收第二交换机在目标网络拥塞状态发送的目标信令报文，目标信令报文携带流量来源信息。

在示例性实施例中，针对1102中第二交换机发送目标信令报文的两种情况，第一交换机接收第二交换机在目标网络拥塞状态发送的目标信令报文，包括但不限于如下两种接收情况。

接收情况一：接收第二交换机在目标网络拥塞状态发送的第一信令报文，第一信令报文用于指示第一交换机发送第一流量控制信息，进行第一类流量控制。

示例性地，接收第二交换机在目标网络拥塞状态发送的第一信令报文，包括：接收第二交换机在目标网络拥塞状态发送的第一CNP报文，第一CNP报文的帧头中的指定字段的值为第一特征值，第一特征值用于指示发送第一流量控制信息。

接收情况二：接收第二交换机在目标网络拥塞状态发送的第二信令报文，第二信令报文用于指示第一交换机发送第二流量控制信息，进行第二类流量控制。

示例性地，接收第二交换机在目标网络拥塞状态发送的第二信令报文，包括：接收第二

交换机在目标网络拥塞状态发送的第二拥塞通知包CNP报文，第二CNP报文的帧头中的指定字段的值为第二特征值，第二特征值用于指示发送第二流量控制信息。

1104，第一交换机根据目标信令报文向流量来源信息对应的网络设备发送目标流量控制信息，目标流量控制信息用于指示进行流量控制。

在示例性实施例中，第一交换机根据目标信令报文向流量来源信息对应的网络设备发送目标流量控制信息，包括但不限于如下两种情况。

情况一：根据目标信令报文向流量来源信息对应的网络设备发送第一流量控制信息，第一流量控制信息用于指示网络设备暂停发送目标队列的数据包，目标队列为网络设备的一个或多个队列。

示例性地，根据目标信令报文向流量来源信息对应的网络设备发送第一流量控制信息，包括：根据目标信令报文构造第一PFC报文，第一PFC报文的时间字段的值为第一值，第一值用于指示第一流量控制信息；向流量来源信息对应的网络设备发送第一PFC报文。

情况二：根据目标信令报文向流量来源信息对应的网络设备发送第二流量控制信息，第二流量控制信息用于指示网络设备继续发送目标队列的数据包，目标队列为网络设备的一个或多个队列。

示例性地，根据目标信令报文向流量来源信息对应的网络设备发送第二流量控制信息，包括：根据目标信令报文构造第二基于优先级的流量控制PFC报文，第二PFC报文的时间字段的值为第二值，第二值用于指示第二流量控制信息；向流量来源信息对应的网络设备发送第二PFC报文。

PFC是IEEE数据中心桥接(Data Center Bridge，DCB)协议族中的技术，是流量控制的增强版。本申请实施例提供的方法在识别网络拥塞状态为CNP失效状态后，触发第一交换机向流量来源信息对应的网络设备发送对应的PFC报文，以进行流量控制。

例如，如表3所示，如果第一交换机接收到的目标信令报文的reserved字段为00000001，则表示第一交换机接收到的目标信令报文为第一信令报文，第一交换机需要给对应的网络设备发送第一PFC报文。示例性地，该第一PFC报文包括但不限于为XOFF的PFC报文。如果第一交换机接收到的目标信令报文的reserved字段为00000010，则表示第一交换机接收到的目标信令报文为第二信令报文，第一交换机需要给对应的网络设备发送第二PFC报文。示例性地，该第二PFC报文包括但不限于为XON的PFC报文。

XON/XOFF是一种在计算机和其它设备之间控制数据流的软件数据流通信协议。其中，X代表发射器。XON/XOFF常指为“软件流控制”。典型地，接收器将发送一个XOFF字符，当它不能接收任何更多的数据时（例如，它可能需要时间来处理一些事情），当它能够再次接收更多的数据时，将发送一个XON字符给发射器。在本申请实施例中，将XON/XOFF的PFC报文作为流量控制报文，实现基于优先级的流量控制。示例性地，PFC报文的格式如表5所示。

表5

字段名	含义
目的MAC地址（Destination Mac Address）	目的MAC地址域，6字节，要求为01-80-C2-00-00-01
源MAC地址（Source Mac Address）	源MAC地址域，6字节，为本设备MAC地址
类型（type）/长度（len）	以太网帧长度或类型域，要求为88-08，用于标

	明本帧的类型为MAC控制帧
控制操作码（Control opcode）	MAC控制操作码，2字节。PFC帧（报文）仅是MAC控制帧的一种，对于PFC帧，其在MAC控制帧中的操作码为01-01
优先级_启用_向量（Priority_enable_vector）	2字节，高字节置0，低字节的每个位代表相应的time[n]有效；当e[n]为0，表示time[n]无效
时间（time）	包含time[0]至time[7]的8个数组单元，每个数组单元为2字节。当e[n]为0时, time[n]没有意义。当e[n]为1时, time[n]代表接收站点抑制优先级为n的报文发送的时间，时间的单位为物理层芯片发送512位数据所需要的时间。所以发送一次PFC PAUSEI帧，要求对端设备暂停发送的时间长度最长为：65535×物理层芯片发送512位数据所需要的时间

仍以图12所示的应用场景为例，中间交换机接收目标信令报文并解析，如果通告对象不是中间交换机自身，则不作特殊操作，将目标信令报文继续转发。

例如，交换机L1收到目标信令报文，从帧头中提取reserved字段并解析，发现是特征值1，则构造如表5所示格式的PFC报文发送给对应的网络设备，例如源服务器。之后，可将该目标信令报文丢弃。其中，PFC报文的time[n]字段为65535，用于指示在65535所示时间内暂停发送目标队列n的数据包。

又例如，第一交换机L2收到目标信令报文，从帧头中提取reserved字段并解析，发现是特征值2，则构造表5所示格式的PFC报文给对应的网络设备，例如源服务器。之后，将该目标信令报文丢弃。其中，PFC报文的time[n]字段为0，用于指示继续发送目标队列n的数据包。

需要说明的是，由于PFC是基于优先级的流量控制报文，Priority_enable_vector字段e[n]指示优先级为n的队列的time值是否有效。以网络设备具有n个优先级的队列为例，如果需要网络设备暂停发送所有队列的数据包，则目标队列包括n个队列，则e[1]至e[n]的值均为非0，将time的值按照暂停时间进行设置。如果需要网络设备暂停发送部分队列的数据包，以目标队列包括1个队列为例，则目标队列是哪个优先级的队列，哪个优先级队列所对应的e[n]的值为非0，将time的值按照暂停时间进行设置。

除了通过上述方式由第一交换机向网络设备发送目标流量控制信息来指示网络设备暂停或继续发送目标队列的数据包的方式外，第一交换机也可以不指定目标队列，只是通过目标流量控制信息指示网络设备暂停或继续发送数据包，由网络设备来确定暂停或发送哪个队列的数据包。

另外，本申请实施例上述过程仅以构造PFC报文来携带目标流量控制信息的方式为例进行说明。除了PFC报文之外，还可以采用PAUSE报文实现，本申请实施例不对携带目标流量控制信息的报文的类型进行限定。

PAUSE报文是一种用于控制MAC数据流量的报文。当对端数据量过大，将无法及时处理数据时，会向数据上游MAC（在本申请实施例中是流量来源对应的网络设备）发送PAUSE报

文，通知上游MAC在一段时间内停止发送数据，停止时间记录在PAUSE报文的PAUSE_TIMING字段。也就是说，该记录有停止时间的PAUSE_TIMING字段用于携带目标流量控制信息。当上游MAC接收到对端的有效PAUSE报文时，会开始计时，并会停止发送数据，防止对端无法及时处理数据，导致对端FIFO溢出或者数据丢失。若计时结束，并且没有收到新的PAUSE报文，将重新发送数据。若计时没有结束，且新收到的PAUSE报文PAUSE_TIMING字段为全0，则表示可以重新发送数据，此时停止计时，重新开始发送数据。

在示例性实施例中，第一交换机根据目标信令报文向流量来源信息对应的网络设备发送目标流量控制信息，包括：根据目标信令报文携带的流量来源信息确定流量来源端口；通过流量来源端口向流量来源信息对应的网络设备发送目标流量控制信息。

示例性地，通过流量来源端口向流量来源信息对应的网络设备发送目标流量控制信息，包括：通过流量来源端口向流量来源信息对应的网络设备发送第三流量控制信息，该第三流量控制信息用于指示暂停发送流量来源端口所对应的队列的数据包。该种情况下，通过第三流量控制信息来控制暂停发送该流量来源端口所对应的所有队列的数据包，实现了端口级控制。本申请实施例不对第三流量控制信息的发送方式进行限定，例如，向流量来源信息对应的网络设备发送PAUSE报文，通过PAUSE报文来携带第三流量控制信息，以指示该流量来源端口所对应的队列的数据包均暂停发送。例如，将PAUSE报文的PAUSE_TIMING字段的值设为非0，以用来携带第三流量控制信息，用于指示在该字段所表示的时间内暂停发送数据包。

示例性地，通过流量来源端口向流量来源信息对应的网络设备发送目标流量控制信息，包括：通过流量来源端口向流量来源信息对应的网络设备发送第四流量控制信息，该第四流量控制信息用于指示继续发送所述流量来源端口所对应的队列的数据包。该种情况下，通过第四流量控制信息来控制继续发送该流量来源端口所对应的所有队列的数据包，实现了端口级控制。本申请实施例不对第四流量控制信息的发送方式进行限定，例如，向流量来源信息对应的网络设备发送PAUSE报文，通过PAUSE报文来携带第四流量控制信息，以指示该流量来源端口所对应的队列的数据包均继续发送。例如，将PAUSE报文的PAUSE_TIMING字段的值设为0，以用来携带第四流量控制信息，用于指示继续发送数据包。

以第一交换机与流量来源信息对应的网络设备之间的交互过程为例，如图11所示，该网络拥塞的控制方法包括如下几个过程。

1105，网络设备接收第一交换机发送的目标流量控制信息，目标流量控制信息用于指示进行流量控制，目标流量控制信息是第一交换机接收到第二交换机在目标网络拥塞状态发送的目标信令报文之后发送的。

在示例性实施例中，接收第一交换机发送的目标流量控制信息，包括但不限于如下两种情况。

情况一：接收第一交换机发送的第一流量控制信息，第一流量控制信息用于指示暂停发送目标队列的数据包，目标队列为网络设备的一个或多个队列。

示例性地，接收第一交换机发送的第一流量控制信息，包括：接收第一交换机发送的第一PFC报文，第一PFC报文的时间字段的值为第一值，第一值用于指示第一流量控制信息。

情况二：接收第一交换机发送的第二流量控制信息，第二流量控制信息用于指示继续发送目标队列的数据包，目标队列为网络设备的一个或多个队列。

示例性地，接收第一交换机发送的第二流量控制信息，包括：接收第一交换机发送的第二PFC报文，第二PFC报文的时间字段的值为第二值，第二值用于指示第二流量控制信息。

1106，网络设备根据目标流量控制信息进行流量控制。

在示例性实施例中，根据目标流量控制信息进行流量控制，包括但不限于如下两种控制方式。

控制方式一：根据第一流量控制信息暂停发送目标队列的数据包。

示例性地，根据第一流量控制信息暂停发送目标队列的数据包，包括：根据第一PFC报文的时间字段的值确定暂停发送数据包的时间长度，在时间长度内暂停发送目标队列的数据包。

控制方式二：根据第二流量控制信息继续发送目标队列的数据包。

示例性地，根据第二流量控制信息继续发送目标队列的数据包，包括：根据第二PFC报文的时间字段的值继续发送目标队列的数据包。

采用本申请实施例提供的方法进行网络拥塞的控制，针对CNP失效状态，将本申请实施例提供的方法与补充CNP报文的相关技术进行对比，得到的测试结果如表6所示。

表6

TCP：RoCE带宽配比	数据流数目	相关技术的端口堆积	相关技术的时延	本申请实施例的端口堆积	本申请实施例的时延
90:10	7*11	19446KB	31637.50us	62KB	687us
70:30	7*40	19757KB	10360.06us	68KB	727us
50:50	7*65	18845KB	5987.04us	53KB	713us
30:70	7*85	16829KB	3824.71us	70KB	661.57us
10:90	7*115	18160KB	3206.68us	81KB	655.78us

从表6所示的CNP失效场景下的测试结果可以看出，本申请实施例提供的方法在端口堆积和业务时延均得到了数量级的优化，且对吞吐几乎无影响。

本申请实施例提供的方法，通过识别网络拥塞状态，当处于CNP失效状态时，通过信令报文指示第一交换机进行流量控制，从而抑制拥塞侧的队列积压，保证业务低时延，且不影响业务的吞吐量，能够支持大规模RoCE组网。不仅解决了CNP失效的问题，还解决了大规模高并发场景下DCQCN控速失效的问题。

本申请实施例提供了一种网络拥塞的控制装置，该装置用于执行图11所示的网络拥塞的控制方法中第一交换机所执行的功能。参见图13，该装置包括：

接收模块1301，用于接收第二交换机在目标网络拥塞状态发送的目标信令报文，目标信令报文携带流量来源信息；

发送模块1302，用于根据目标信令报文向流量来源信息对应的网络设备发送目标流量控制信息，目标流量控制信息用于指示进行流量控制。

在示例性实施例中，发送模块1302，用于根据目标信令报文向流量来源信息对应的网络设备发送第一流量控制信息，第一流量控制信息用于指示网络设备暂停发送目标队列的数据包，目标队列为网络设备的一个或多个队列。

在示例性实施例中，发送模块1302，用于根据目标信令报文构造第一基于优先级的流量

控制PFC报文，第一PFC报文的时间字段的值为第一值，第一值用于指示第一流量控制信息；向流量来源信息对应的网络设备发送第一PFC报文。

在示例性实施例中，接收模块1301，用于接收第二交换机在目标网络拥塞状态发送的第一信令报文，第一信令报文用于指示发送第一流量控制信息。

在示例性实施例中，接收模块1301，用于接收第二交换机在目标网络拥塞状态发送的第一拥塞通知包CNP报文，第一CNP报文的帧头中的指定字段的值为第一特征值，第一特征值用于指示发送第一流量控制信息。

在示例性实施例中，发送模块1302，用于根据目标信令报文向流量来源信息对应的网络设备发送第二流量控制信息，第二流量控制信息用于指示网络设备继续发送目标队列的数据包，目标队列为网络设备的一个或多个队列。

在示例性实施例中，发送模块1302，用于根据目标信令报文构造第二基于优先级的流量控制PFC报文，第二PFC报文的时间字段的值为第二值，第二值用于指示第二流量控制信息；向流量来源信息对应的网络设备发送第二PFC报文。

在示例性实施例中，接收模块1301，用于接收第二交换机在目标网络拥塞状态发送的第二信令报文，第二信令报文用于指示发送第二流量控制信息。

在示例性实施例中，接收模块1301，用于接收第二交换机在目标网络拥塞状态发送的第二拥塞通知包CNP报文，第二CNP报文的帧头中的指定字段的值为第二特征值，第二特征值用于指示发送第二流量控制信息。

在示例性实施例中，发送模块1302，用于根据目标信令报文携带的流量来源信息确定流量来源端口；通过流量来源端口向流量来源信息对应的网络设备发送目标流量控制信息。

在示例性实施例中，发送模块1302，用于通过流量来源端口向流量来源信息对应的网络设备发送第三流量控制信息，第三流量控制信息用于指示暂停发送流量来源端口所对应的队列的数据包。

在示例性实施例中，发送模块1302，用于通过流量来源端口向流量来源信息对应的网络设备发送第四流量控制信息，第四流量控制信息用于指示继续发送流量来源端口所对应的队列的数据包。

本申请实施例提供的装置，接收到第二交换机在目标网络拥塞状态发送的目标信令报文后，通过向目标信令报文中携带的流量来源信息所对应的网络设备发送目标流量控制信息，以指示进行流量控制，从而抑制拥塞侧的队列积压，保证业务低时延，且不影响业务的吞吐量，能够支持大规模RoCE组网，解决了大规模高并发场景下DCQCN控速失效的问题。

本申请实施例提供了一种网络拥塞的控制装置，该装置用于执行图11所示的网络拥塞的控制方法中第二交换机所执行的功能。参见图14，该装置包括：

识别模块1401，用于识别网络拥塞状态；

发送模块1402，用于响应于网络拥塞状态为目标网络拥塞状态，向第一交换机发送目标信令报文，目标信令报文携带流量来源信息，目标信令报文用于指示第一交换机进行流量控制。

在示例性实施例中，目标信令报文包括第一信令报文或第二信令报文，发送模块1402，用于响应于网络拥塞状态为目标网络拥塞状态，且当前的队列长度大于第一阈值，向第一交

换机发送第一信令报文，第一信令报文用于指示第一交换机发送第一流量控制信息，第一流量控制信息用于指示流量来源信息对应的网络设备暂停发送目标队列的数据包，目标队列为网络设备的一个或多个队列；或者，

响应于网络拥塞状态为目标网络拥塞状态，且当前的队列长度小于第二阈值，向第一交换机发送第二信令报文，第二信令报文用于指示第一交换机发送第二流量控制信息，第二流量控制信息用于指示网络设备继续发送目标队列的数据包，第二阈值小于第一阈值。

在示例性实施例中，该装置还包括:

获取模块，用于获取第一CNP报文，将第一CNP报文的帧头中的指定字段的值设为第一特征值，将第一CNP报文作为第一信令报文;

或者，获取模块，用于获取第二CNP报文，将第二CNP报文的帧头中的指定字段的值设为第二特征值，将第二CNP报文作为第二信令报文。

在示例性实施例中，识别模块1401，用于读取当前的队列长度及显式拥塞通知ECN阈值范围，ECN阈值范围用于指示添加ECN标识的概率，ECN标识用于指示网络发生拥塞；根据当前的队列长度及ECN阈值范围识别网络拥塞状态。

本申请实施例提供的装置，通过识别网络拥塞状态，在目标网络拥塞状态，通过目标信令报文指示第一交换机进行流量控制，从而抑制拥塞侧的队列积压，保证业务低时延，且不影响业务的吞吐量，能够支持大规模RoCE组网，解决大规模高并发场景下DCQCN控速失效的问题。

本申请实施例提供了一种网络拥塞的控制装置，该装置用于执行图11所示的网络拥塞的控制方法中网络设备所执行的功能。参见图15，该装置包括:

接收模块1501，用于接收第一交换机发送的目标流量控制信息，目标流量控制信息用于指示进行流量控制，所示目标流量控制信息是第一交换机接收到第二交换机在目标网络拥塞状态发送的目标信令报文之后发送的;

控制模块1502，用于根据目标流量控制信息进行流量控制。

在示例性实施例中，接收模块1501，用于接收第一交换机发送的第一流量控制信息，第一流量控制信息用于指示暂停发送目标队列的数据包，目标队列为网络设备的一个或多个队列;

控制模块1502，用于根据第一流量控制信息暂停发送目标队列的数据包。

在示例性实施例中，接收模块1501，用于接收第一交换机发送的第一PFC报文，第一PFC报文的时间字段的值为第一值，第一值用于指示第一流量控制信息;

控制模块1502，用于根据第一PFC报文的时间字段的值确定暂停发送数据包的时间长度，在时间长度内暂停发送目标队列的数据包。

在示例性实施例中，接收模块1501，用于接收第一交换机发送的第二流量控制信息，第二流量控制信息用于指示继续发送目标队列的数据包，目标队列为网络设备的一个或多个队列;

控制模块1502，用于根据第二流量控制信息继续发送目标队列的数据包。

在示例性实施例中，接收模块1501，用于接收第一交换机发送的第二PFC报文，第二PFC报文的时间字段的值为第二值，第二值用于指示第二流量控制信息;

控制模块1502，用于根据第二PFC报文的时间字段的值继续发送目标队列的数据包。

本申请实施例提供的装置，接收到第一交换机发送的目标流量控制信息后，基于该目标流量控制信息进行流量控制，从而抑制拥塞侧的队列积压，保证业务低时延，且不影响业务的吞吐量，能够支持大规模RoCE组网，解决大规模高并发场景下DCQCN控速失效的问题。

应理解的是，上述图13-图15提供的装置在实现其功能时，仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将设备的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的装置与方法实施例属于同一构思，其具体实现过程详见方法实施例，这里不再赘述。此外，在示例性实施例中，上述图13-图15提供的装置在实现其功能时，涉及的目标网络拥塞状态包括但不限于ECN失效状态或CNP失效状态；其中，ECN失效状态是指第二交换机当前的队列长度大于参考范围的最大值，且未补充CNP报文的状态；CNP失效状态是指第二交换机当前的队列长度大于参考区域的最大值，且已补充CNP报文的状态。

图 16 为本申请实施例的网络拥塞的控制设备 1600 的硬件结构示意图。图 16 所示的网络拥塞的控制设备 1600 可以执行上述图 11 所示实施例提供的网络拥塞的控制方法中的相应步骤。

如图 16 所示，网络拥塞的控制设备 1600 包括处理器 1601、存储器 1602、接口 1603 和总线 1604。其中接口 1603 可以通过无线或有线的方式实现，示例性地，该接口 1603 可以是网卡。上述处理器 1601、存储器 1602 和接口 1603 通过总线 1604 连接。

接口1603可以包括发送器和接收器，用于与其他通信设备通信。处理器1601用于执行上述图3所示实施例中301-304的处理相关步骤。处理器1601和/或用于本文所描述的技术的其他过程。

例如，图 16 所示的网络拥塞的控制设备 1600 为图 11 中的第一交换机，处理器 1602 读取存储器 1601 中的指令，使图 16 所示的网络拥塞的控制设备 1600 能够执行第一交换机所执行的全部或部分操作。

又例如，图 16 所示的网络拥塞的控制设备 1600 为图 11 中的第二交换机，处理器 1602 读取存储器 1601 中的指令，使图 16 所示的网络拥塞的控制设备 1600 能够执行第二交换机所执行的全部或部分操作。

又例如，图 16 所示的网络拥塞的控制设备 1600 为图 11 中的网络设备，处理器 1602 读取存储器 1601 中的指令，使图 16 所示的网络拥塞的控制设备 1600 能够执行网络设备所执行的全部或部分操作。

存储器1602包括操作系统16021和应用程序16022，用于存储程序、代码或指令，当处理器或硬件设备执行这些程序、代码或指令时可以完成方法实施例中涉及网络拥塞的控制设备1600的处理过程。可选的，存储器1602可以包括只读存储器（英文：Read-only Memory，缩写：ROM）和随机存取存储器（英文：Random Access Memory，缩写：RAM）。其中，ROM包括基本输入/输出系统（英文：Basic Input/Output System，缩写：BIOS）或嵌入式系统；RAM包括应用程序和操作系统。当需要运行网络拥塞的控制设备1600时，通过固化在ROM中的BIOS或者嵌入式系统中的bootloader引导系统进行启动，引导网络拥塞的控制设备1600进入正

常运行状态。在网络拥塞的控制设备1600进入正常运行状态后，运行在RAM中的应用程序和操作系统，从而，完成方法实施例中涉及网络拥塞的控制设备1600的处理过程。

可以理解的是，图 16 仅仅示出了网络拥塞的控制设备 1600 的简化设计。在实际应用中，网络拥塞的控制设备 1600 可以包含任意数量的接口，处理器或者存储器。

应理解的是，上述处理器可以是中央处理器（Central Processing Unit，CPU），还可以是其他通用处理器、数字信号处理器（digital signal processing，DSP）、专用集成电路（application specific integrated circuit ，ASIC）、现场可编程门阵列（field－programmable gate array，FPGA）或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者是任何常规的处理器等。值得说明的是，处理器可以是支持进阶精简指令集机器（advanced RISC machines，ARM）架构的处理器。

进一步地，在一种可选的实施例中，上述存储器可以包括只读存储器和随机存取存储器，并向处理器提供指令和数据。存储器还可以包括非易失性随机存取存储器。例如，存储器还可以存储设备类型的信息。

该存储器可以是易失性存储器或非易失性存储器，或可包括易失性和非易失性存储器两者。其中，非易失性存储器可以是只读存储器（read-only memory，ROM）、可编程只读存储器（programmable ROM，PROM）、可擦除可编程只读存储器（erasable PROM，EPROM）、电可擦除可编程只读存储器（electrically EPROM，EEPROM）或闪存。易失性存储器可以是随机存取存储器（random access memory，RAM），其用作外部高速缓存。通过示例性但不是限制性说明，许多形式的RAM可用。例如，静态随机存取存储器（static RAM，SRAM）、动态随机存取存储器（dynamic random access memory，DRAM）、同步动态随机存取存储器（synchronous DRAM，SDRAM）、双倍数据速率同步动态随机存取存储器（double data date SDRAM，DDR SDRAM）、增强型同步动态随机存取存储器（enhanced SDRAM，ESDRAM）、同步连接动态随机存取存储器（synchlink DRAM，SLDRAM）和直接内存总线随机存取存储器（direct rambus RAM，DR RAM）。

还提供了一种计算机可读存储介质，存储介质中存储有至少一条指令，指令由处理器加载并执行以实现如上任一所述的网络拥塞的控制方法。

本申请提供了一种计算机程序，当计算机程序被计算机执行时，可以使得处理器或计算机执行上述方法实施例中对应的各个步骤和/或流程。

提供了一种芯片，包括处理器，用于从存储器中调用并运行所述存储器中存储的指令，使得安装有所述芯片的通信设备执行上述各方面中的方法。

提供另一种芯片，包括：输入接口、输出接口、处理器和存储器，所述输入接口、输出接口、所述处理器以及所述存储器之间通过内部连接通路相连，所述处理器用于执行所述存储器中的代码，当所述代码被执行时，所述处理器用于执行上述各方面中的方法。

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。当使用软件实现时，可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时，全部或部分地产生按照本申请所述的流程或功能。所述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。所述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，所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线（例如同轴电缆、光纤、数字用户线）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。所述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用介质集成的服务器、数据中心等数据存储设备。所述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 State Disk）等。

本申请中术语“第一”、“第二”等字样用于对作用和功能基本相同的相同项或相似项进行区分， 应理解，“第一”、“第二”、“第 n”之间不具有逻辑或时序上的依赖关系，也不对执行顺序进行限定。还应理解，尽管描述中使用术语第一、第二等来描述各种元素，但这些元素不应受术语的限制。这些术语只是用于将一元素与另一元素区别分开。

以上所述的具体实施方式，对本申请的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本申请的具体实施方式而已，并不用于限定本申请的保护范围，凡在本申请的技术方案的基础之上，所做的任何修改、等同替换、改进等，均应包括在本申请的保护范围之内。

权利要求书

1.一种网络拥塞的控制方法，其特征在于，所述方法应用于第一交换机，所述方法包括:

所述第一交换机接收第二交换机在目标网络拥塞状态发送的目标信令报文，所述目标信令报文携带流量来源信息;

根据所述目标信令报文向所述流量来源信息对应的网络设备发送目标流量控制信息，所述目标流量控制信息用于指示进行流量控制。

2.根据权利要求1所述的方法，其特征在于，所述根据所述目标信令报文向所述流量来源信息对应的网络设备发送目标流量控制信息，包括:

根据所述目标信令报文向所述流量来源信息对应的网络设备发送第一流量控制信息，所述第一流量控制信息用于指示所述网络设备暂停发送目标队列的数据包，所述目标队列为所述网络设备的一个或多个队列。

3.根据权利要求2所述的方法，其特征在于，所述根据所述目标信令报文向所述流量来源信息对应的网络设备发送第一流量控制信息，包括:

根据所述目标信令报文构造第一基于优先级的流量控制PFC报文，所述第一PFC报文的时间字段的值为第一值，所述第一值用于指示所述第一流量控制信息;

向所述流量来源信息对应的网络设备发送所述第一PFC报文。

4.根据权利要求2或3所述的方法，其特征在于，所述接收第二交换机在目标网络拥塞状态发送的目标信令报文，包括:

接收所述第二交换机在目标网络拥塞状态发送的第一信令报文，所述第一信令报文用于指示发送所述第一流量控制信息。

5.根据权利要求4所述的方法，其特征在于，所述接收所述第二交换机在目标网络拥塞状态发送的第一信令报文，包括:

接收所述第二交换机在目标网络拥塞状态发送的第一拥塞通知包CNP报文，所述第一CNP报文的帧头中的指定字段的值为第一特征值，所述第一特征值用于指示发送所述第一流量控制信息。

6.根据权利要求1所述的方法，其特征在于，所述根据所述目标信令报文向所述流量来源信息对应的网络设备发送目标流量控制信息，包括:

根据所述目标信令报文向所述流量来源信息对应的网络设备发送第二流量控制信息，所述第二流量控制信息用于指示所述网络设备继续发送目标队列的数据包，所述目标队列为所述网络设备的一个或多个队列。

7.根据权利要求6所述的方法，其特征在于，所述根据所述目标信令报文向所述流量来源

信息对应的网络设备发送第二流量控制信息，包括:

根据所述目标信令报文构造第二基于优先级的流量控制PFC报文，所述第二PFC报文的时间字段的值为第二值，所述第二值用于指示所述第二流量控制信息;

向所述流量来源信息对应的网络设备发送所述第二PFC报文。

8.根据权利要求6或7所述的方法，其特征在于，所述接收第二交换机在目标网络拥塞状态发送的目标信令报文，包括:

接收所述第二交换机在目标网络拥塞状态发送的第二信令报文，所述第二信令报文用于指示发送所述第二流量控制信息。

9.根据权利要求8所述的方法，其特征在于，所述接收所述第二交换机在目标网络拥塞状态发送的第二信令报文，包括:

接收所述第二交换机在目标网络拥塞状态发送的第二拥塞通知包CNP报文，所述第二CNP报文的帧头中的指定字段的值为第二特征值，所述第二特征值用于指示发送所述第二流量控制信息。

10.根据权利要求1-9任一所述的方法，其特征在于，所述根据所述目标信令报文向所述流量来源信息对应的网络设备发送目标流量控制信息，包括:

根据所述目标信令报文携带的流量来源信息确定流量来源端口;

通过所述流量来源端口向所述流量来源信息对应的网络设备发送目标流量控制信息。

11.根据权利要求1-10任一所述的方法，其特征在于，所述目标网络拥塞状态包括显式拥塞通知ECN失效状态或拥塞通知包CNP失效状态;

所述ECN失效状态是指所述第二交换机当前的队列长度大于参考范围的最大值，且未补充CNP报文的状态;所述CNP失效状态是指所述第二交换机当前的队列长度大于所述参考区域的最大值，且已补充CNP报文的状态;所述参考范围基于ECN阈值范围确定，所述ECN阈值范围用于指示添加ECN标识的概率，所述ECN标识用于指示网络发生拥塞。

12.一种网络拥塞的控制方法，其特征在于，所述方法应用于第二交换机，所述方法包括:

所述第二交换机识别网络拥塞状态;

响应于网络拥塞状态为目标网络拥塞状态，向第一交换机发送目标信令报文，所述目标信令报文携带流量来源信息，所述目标信令报文用于指示所述第一交换机进行流量控制。

13.根据权利要求12所述的方法，其特征在于，所述目标信令报文包括第一信令报文或第二信令报文，所述响应于网络拥塞状态为目标网络拥塞状态，向第一交换机发送目标信令报文，包括:

响应于网络拥塞状态为目标网络拥塞状态，且当前的队列长度大于第一阈值，向所述第一交换机发送第一信令报文，所述第一信令报文用于指示所述第一交换机发送第一流量控制信息，所述第一流量控制信息用于指示所述流量来源信息对应的网络设备暂停发送目标队列

的数据包，所述目标队列为所述网络设备的一个或多个队列；或者，

响应于网络拥塞状态为目标网络拥塞状态，且所述当前的队列长度小于第二阈值，向所述第一交换机发送第二信令报文，所述第二信令报文用于指示所述第一交换机发送第二流量控制信息，所述第二流量控制信息用于指示所述网络设备继续发送目标队列的数据包，所述第二阈值小于所述第一阈值。

14.根据权利要求13所述的方法，其特征在于，所述向所述第一交换机发送第一信令报文之前，还包括：获取第一拥塞通知包CNP报文，将所述第一CNP报文的帧头中的指定字段的值设为第一特征值，将所述第一CNP报文作为所述第一信令报文；

所述向所述第一交换机发送第二信令报文之前，还包括：获取第二CNP报文，将所述第二CNP报文的帧头中的指定字段的值设为第二特征值，将所述第二CNP报文作为所述第二信令报文。

15.根据权利要求12-14任一所述的方法，其特征在于，所述识别网络拥塞状态，包括：

读取当前的队列长度及显式拥塞通知ECN阈值范围，所述ECN阈值范围用于指示添加ECN标识的概率，所述ECN标识用于指示网络发生拥塞；

根据所述当前的队列长度及所述ECN阈值范围识别网络拥塞状态。

16.根据权利要求15所述的方法，其特征在于，所述目标网络拥塞状态包括ECN失效状态或拥塞通知包CNP失效状态，所述根据所述当前的队列长度及所述ECN阈值范围识别网络拥塞状态，包括：

响应于所述当前的队列长度大于参考范围的最大值，且未补充CNP报文，则所述网络拥塞状态为ECN失效状态，所述参考范围基于所述ECN阈值范围确定；

响应于所述当前的队列长度大于所述参考范围的最大值，且已补充CNP报文，则所述网络拥塞状态为CNP失效状态。

17.一种网络拥塞的控制方法，其特征在于，所述方法应用于网络设备，所述方法包括：

所述网络设备接收第一交换机发送的目标流量控制信息，所述目标流量控制信息用于指示进行流量控制，所示目标流量控制信息是所述第一交换机接收到第二交换机在目标网络拥塞状态发送的目标信令报文之后发送的；

根据所述目标流量控制信息进行流量控制。

18.根据权利要求17所述的方法，其特征在于，所述接收第一交换机发送的目标流量控制信息，包括：

接收所述第一交换机发送的第一流量控制信息，所述第一流量控制信息用于指示暂停发送目标队列的数据包，所述目标队列为所述网络设备的一个或多个队列；

所述根据所述目标流量控制信息进行流量控制，包括：

根据所述第一流量控制信息暂停发送所述目标队列的数据包。

19.根据权利要求18所述的方法，其特征在于，所述接收所述第一交换机发送的第一流量控制信息，包括:

接收所述第一交换机发送的第一基于优先级的流量控制PFC报文，所述第一PFC报文的时间字段的值为第一值，所述第一值用于指示所述第一流量控制信息;

所述根据所述第一流量控制信息暂停发送所述目标队列的数据包，包括:

根据所述第一PFC报文的时间字段的值确定暂停发送数据包的时间长度，在所述时间长度内暂停发送所述目标队列的数据包。

20.根据权利要求17所述的方法，其特征在于，所述接收第一交换机发送的目标流量控制信息，包括:

接收所述第一交换机发送的第二流量控制信息，所述第二流量控制信息用于指示继续发送目标队列的数据包，所述目标队列为所述网络设备的一个或多个队列;

所述根据所述目标流量控制信息进行流量控制，包括:

根据所述第二流量控制信息继续发送所述目标队列的数据包。

21.根据权利要求20所述的方法，其特征在于，所述接收所述第一交换机发送的第二流量控制信息，包括:

接收所述第一交换机发送的第二基于优先级的流量控制PFC报文，所述第二PFC报文的时间字段的值为第二值，所述第二值用于指示所述第二流量控制信息;

所述根据所述第二流量控制信息继续发送所述目标队列的数据包，包括:

根据所述第二PFC报文的时间字段的值继续发送所述目标队列的数据包。

22.根据权利要求17-21任一所述的方法，其特征在于，所述目标网络拥塞状态包括显式拥塞通知ECN失效状态或拥塞通知包CNP失效状态;

所述ECN失效状态是指所述第二交换机当前的队列长度大于参考范围的最大值，且未补充CNP报文的状态；所述CNP失效状态是指所述第二交换机当前的队列长度大于所述参考区域的最大值，且已补充CNP报文的状态；所述参考范围基于ECN阈值范围确定，所述ECN阈值范围用于指示添加ECN标识的概率，所述ECN标识用于指示网络发生拥塞。

23.一种网络拥塞的控制装置，其特征在于，所述装置包括:

接收模块，用于接收第二交换机在目标网络拥塞状态发送的目标信令报文，所述目标信令报文携带流量来源信息;

发送模块，用于根据所述目标信令报文向所述流量来源信息对应的网络设备发送目标流量控制信息，所述目标流量控制信息用于指示进行流量控制。

24.根据权利要求23所述的装置，其特征在于，所述发送模块，用于根据所述目标信令报文向所述流量来源信息对应的网络设备发送第一流量控制信息，所述第一流量控制信息用于指示所述网络设备暂停发送目标队列的数据包，所述目标队列为所述网络设备的一个或多个队列。

25.根据权利要求24所述的装置，其特征在于，所述发送模块，用于根据所述目标信令报文构造第一基于优先级的流量控制PFC报文，所述第一PFC报文的时间字段的值为第一值，所述第一值用于指示所述第一流量控制信息；向所述流量来源信息对应的网络设备发送所述第一PFC报文。

26.根据权利要求24或25所述的装置，其特征在于，所述接收模块，用于接收所述第二交换机在目标网络拥塞状态发送的第一信令报文，所述第一信令报文用于指示发送所述第一流量控制信息。

27.根据权利要求26所述的装置，其特征在于，所述接收模块，用于接收所述第二交换机在目标网络拥塞状态发送的第一拥塞通知包CNP报文，所述第一CNP报文的帧头中的指定字段的值为第一特征值，所述第一特征值用于指示发送所述第一流量控制信息。

28.根据权利要求23所述的装置，其特征在于，所述发送模块，用于根据所述目标信令报文向所述流量来源信息对应的网络设备发送第二流量控制信息，所述第二流量控制信息用于指示所述网络设备继续发送目标队列的数据包，所述目标队列为所述网络设备的一个或多个队列。

29.根据权利要求28所述的装置，其特征在于，所述发送模块，用于根据所述目标信令报文构造第二基于优先级的流量控制PFC报文，所述第二PFC报文的时间字段的值为第二值，所述第二值用于指示所述第二流量控制信息；向所述流量来源信息对应的网络设备发送所述第二PFC报文。

30.根据权利要求28或29所述的装置，其特征在于，所述接收模块，用于接收所述第二交换机在目标网络拥塞状态发送的第二信令报文，所述第二信令报文用于指示发送所述第二流量控制信息。

31.根据权利要求30所述的装置，其特征在于，所述接收模块，用于接收所述第二交换机在目标网络拥塞状态发送的第二拥塞通知包CNP报文，所述第二CNP报文的帧头中的指定字段的值为第二特征值，所述第二特征值用于指示发送所述第二流量控制信息。

32.根据权利要求23-31任一所述的装置，其特征在于，所述发送模块，用于根据所述目标信令报文携带的流量来源信息确定流量来源端口；通过所述流量来源端口向所述流量来源信息对应的网络设备发送目标流量控制信息。

33.根据权利要求23-32任一所述的装置，其特征在于，所述目标网络拥塞状态包括显式拥塞通知ECN失效状态或拥塞通知包CNP失效状态；

所述ECN失效状态是指所述第二交换机当前的队列长度大于参考范围的最大值，且未补

充CNP报文的状态；所述CNP失效状态是指所述第二交换机当前的队列长度大于所述参考区域的最大值，且已补充CNP报文的状态；所述参考范围基于ECN阈值范围确定，所述ECN阈值范围用于指示添加ECN标识的概率，所述ECN标识用于指示网络发生拥塞。

34.一种网络拥塞的控制装置，其特征在于，所述装置包括:

识别模块，用于识别网络拥塞状态;

发送模块，用于响应于网络拥塞状态为目标网络拥塞状态，向第一交换机发送目标信令报文，所述目标信令报文携带流量来源信息，所述目标信令报文用于指示所述第一交换机进行流量控制。

35.根据权利要求34所述的装置，其特征在于，所述目标信令报文包括第一信令报文或第二信令报文，所述发送模块，用于响应于网络拥塞状态为目标网络拥塞状态，且当前的队列长度大于第一阈值，向所述第一交换机发送第一信令报文，所述第一信令报文用于指示所述第一交换机发送第一流量控制信息，所述第一流量控制信息用于指示所述流量来源信息对应的网络设备暂停发送目标队列的数据包，所述目标队列为所述网络设备的一个或多个队列;或者，

响应于网络拥塞状态为目标网络拥塞状态，且所述当前的队列长度小于第二阈值，向所述第一交换机发送第二信令报文，所述第二信令报文用于指示所述第一交换机发送第二流量控制信息，所述第二流量控制信息用于指示所述网络设备继续发送目标队列的数据包，所述第二阈值小于所述第一阈值。

36.根据权利要求35所述的装置，其特征在于，所述装置还包括:

获取模块，用于获取第一拥塞通知包CNP报文，将所述第一CNP报文的帧头中的指定字段的值设为第一特征值，将所述第一CNP报文作为所述第一信令报文;

或者，所述获取模块，用于获取第二CNP报文，将所述第二CNP报文的帧头中的指定字段的值设为第二特征值，将所述第二CNP报文作为所述第二信令报文。

37.根据权利要求34-36任一所述的装置，其特征在于，所述识别模块，用于读取当前的队列长度及显式拥塞通知ECN阈值范围，所述ECN阈值范围用于指示添加ECN标识的概率，所述ECN标识用于指示网络发生拥塞；根据所述当前的队列长度及所述ECN阈值范围识别网络拥塞状态。

38.根据权利要求37所述的装置，其特征在于，所述目标网络拥塞状态包括ECN失效状态或拥塞通知包CNP失效状态，所述识别模块，用于响应于所述当前的队列长度大于参考范围的最大值，且未补充CNP报文，则所述网络拥塞状态为ECN失效状态，所述参考范围基于所述ECN阈值范围确定;响应于所述当前的队列长度大于所述参考范围的最大值，且已补充CNP报文，则所述网络拥塞状态为CNP失效状态。

39.一种网络拥塞的控制装置，其特征在于，所述装置包括:

接收模块，用于接收第一交换机发送的目标流量控制信息，所述目标流量控制信息用于指示进行流量控制，所示目标流量控制信息是所述第一交换机接收到第二交换机在目标网络拥塞状态发送的目标信令报文之后发送的；

控制模块，用于根据所述目标流量控制信息进行流量控制。

40.根据权利要求39所述的装置，其特征在于，所述接收模块，用于接收所述第一交换机发送的第一流量控制信息，所述第一流量控制信息用于指示暂停发送目标队列的数据包，所述目标队列为所述网络设备的一个或多个队列；

所述控制模块，用于根据所述第一流量控制信息暂停发送所述目标队列的数据包。

41.根据权利要求40所述的装置，其特征在于，所述接收模块，用于接收所述第一交换机发送的第一基于优先级的流量控制PFC报文，所述第一PFC报文的时间字段的值为第一值，所述第一值用于指示所述第一流量控制信息；

所述控制模块，用于根据所述第一PFC报文的时间字段的值确定暂停发送数据包的时间长度，在所述时间长度内暂停发送所述目标队列的数据包。

42.根据权利要求39所述的装置，其特征在于，所述接收模块，用于接收所述第一交换机发送的第二流量控制信息，所述第二流量控制信息用于指示继续发送目标队列的数据包，所述目标队列为所述网络设备的一个或多个队列；

所述控制模块，用于根据所述第二流量控制信息继续发送所述目标队列的数据包。

43.根据权利要求42所述的装置，其特征在于，所述接收模块，用于接收所述第一交换机发送的第二基于优先级的流量控制PFC报文，所述第二PFC报文的时间字段的值为第二值，所述第二值用于指示所述第二流量控制信息；

所述控制模块，用于根据所述第二PFC报文的时间字段的值继续发送所述目标队列的数据包。

44.根据权利要求39-43任一所述的装置，其特征在于，所述目标网络拥塞状态包括显式拥塞通知ECN失效状态或拥塞通知包CNP失效状态；

所述ECN失效状态是指所述第二交换机当前的队列长度大于参考范围的最大值，且未补充CNP报文的状态；所述CNP失效状态是指所述第二交换机当前的队列长度大于所述参考区域的最大值，且已补充CNP报文的状态；所述参考范围基于ECN阈值范围确定，所述ECN阈值范围用于指示添加ECN标识的概率，所述ECN标识用于指示网络发生拥塞。

45.一种网络拥塞的控制设备，其特征在于，所述设备包括：

存储器及处理器，所述存储器中存储有至少一条指令，所述至少一条指令由所述处理器加载并执行，以实现权利要求1-11中任一所述的网络拥塞的控制方法。

46.一种网络拥塞的控制设备，其特征在于，所述设备包括：

存储器及处理器，所述存储器中存储有至少一条指令，所述至少一条指令由所述处理器加载并执行，以实现权利要求12-16中任一所述的网络拥塞的控制方法。

47.一种网络拥塞的控制设备，其特征在于，所述设备包括:

存储器及处理器，所述存储器中存储有至少一条指令，所述至少一条指令由所述处理器加载并执行，以实现权利要求17-22中任一所述的网络拥塞的控制方法。

48.一种网络拥塞的控制系统，其特征在于，所述系统包括: 所述权利要求45所述的设备、所述权利要求46所述的设备及所述权利要求47所述的设备。

49.一种计算机可读存储介质，其特征在于，所述存储介质中存储有至少一条指令，所述指令由处理器加载并执行以实现如权利要求 1-22 中任一所述的网络拥塞的控制方法。

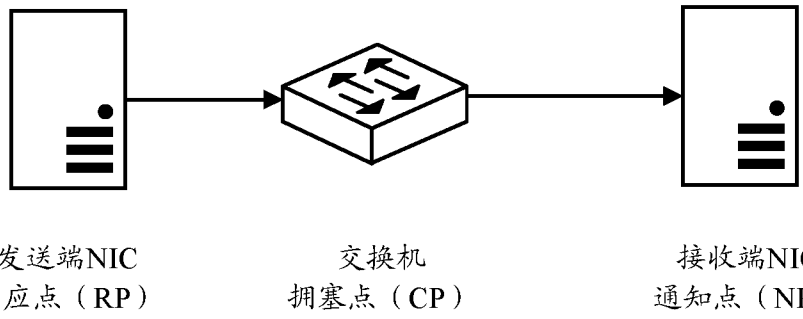
发送端NIC
响应点（RP）
交换机
拥塞点（CP）
接收端NIC
通知点（NP）

图 1

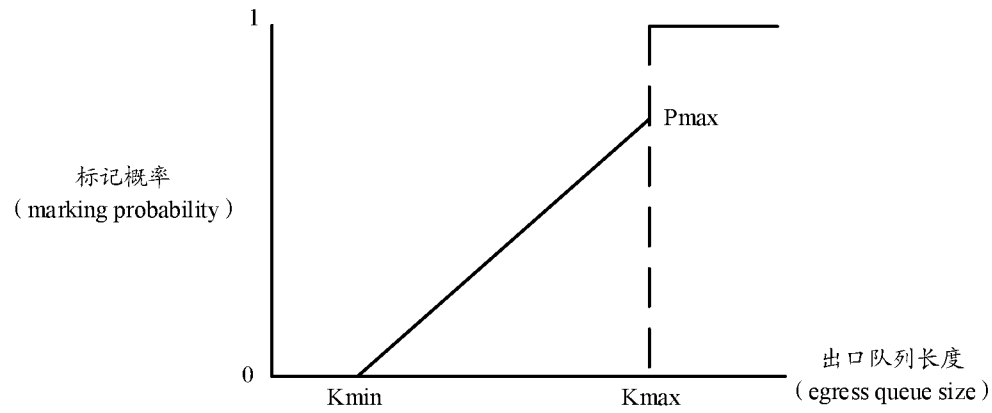
1
Pmax
标记概率
（marking probability）
0
Kmin
Kmax
出口队列长度
（egress queue size）

图 2

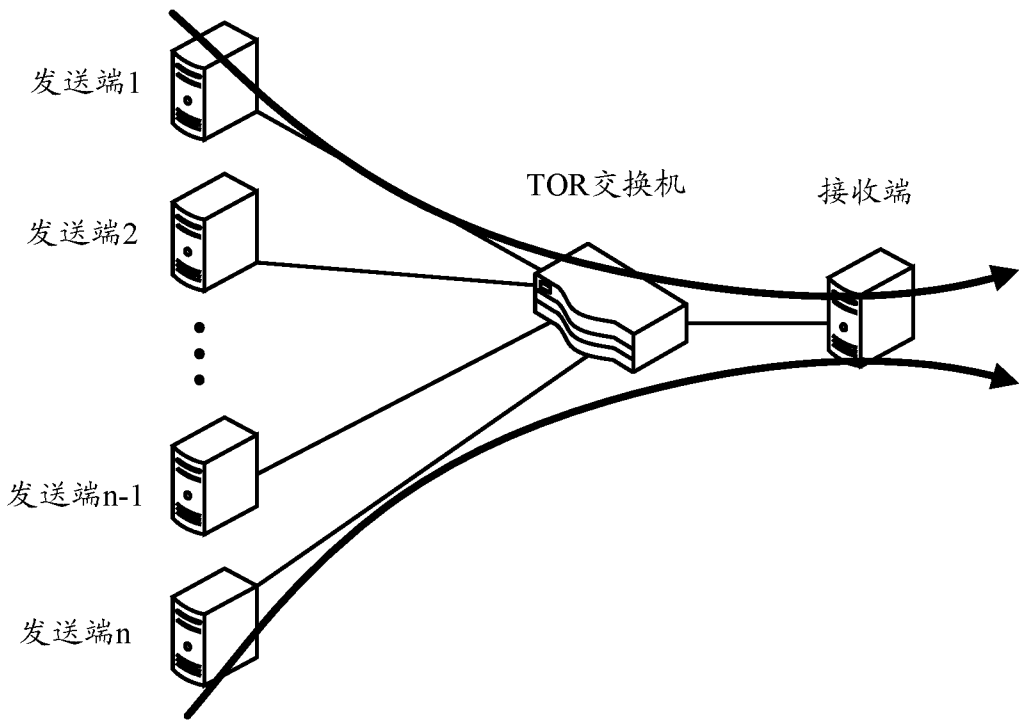
发送端1
发送端2
发送端n-1
发送端n
TOR交换机
接收端

图 3

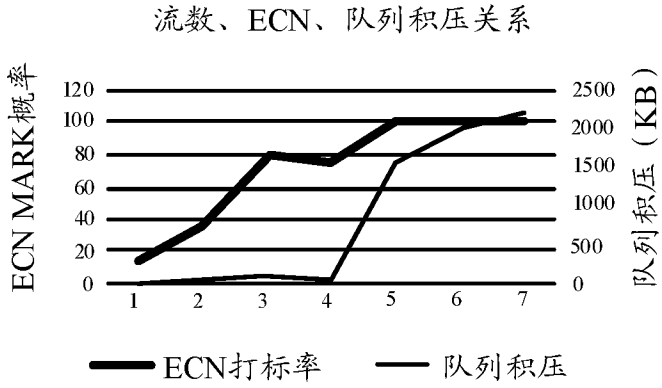
流数、ECN、队列积压关系
ECN MARK概率
120
100
80
60
40
20
0
2500
2000
1500
1000
500
0
队列积压（KB）
1
2
3
4
5
6
7
ECN打标率
队列积压

图 4

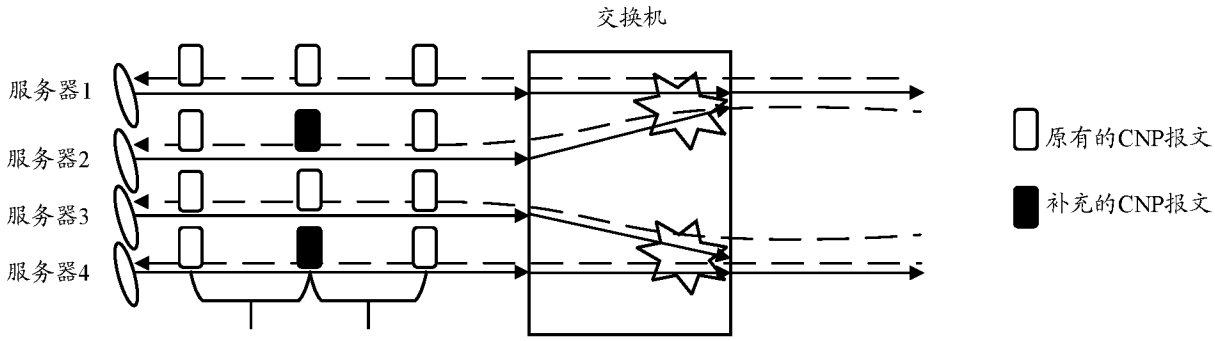
交换机
服务器1
服务器2
服务器3
服务器4
原有的CNP报文
补充的CNP报文

图 5

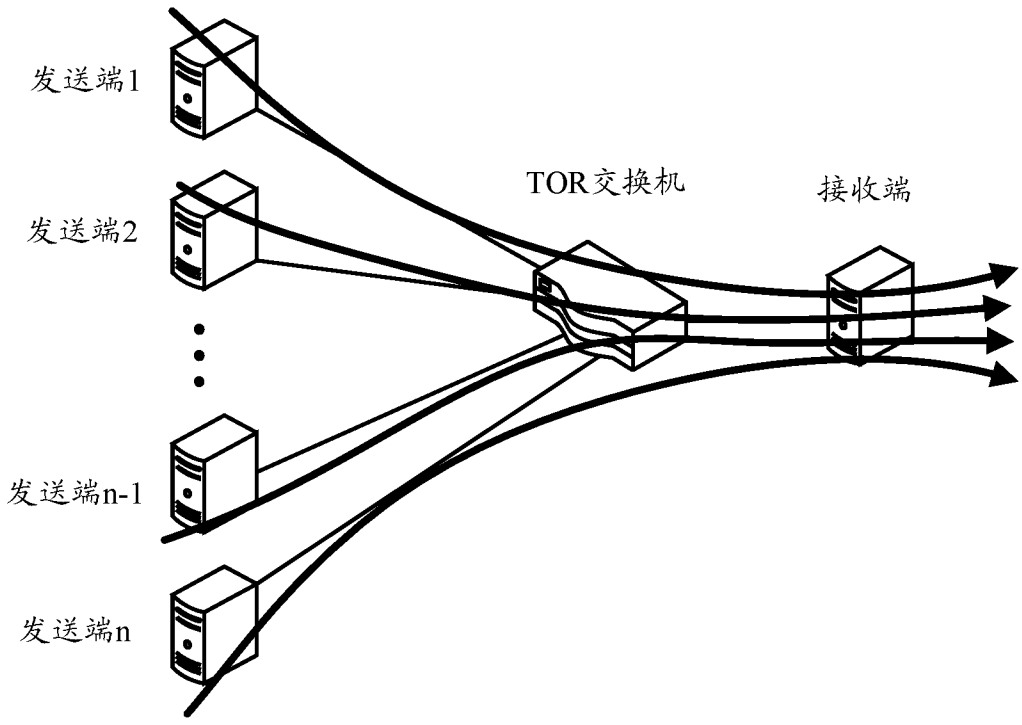
发送端1
发送端2
发送端n-1
发送端n
TOR交换机
接收端

图 6

时延随qp数变化图

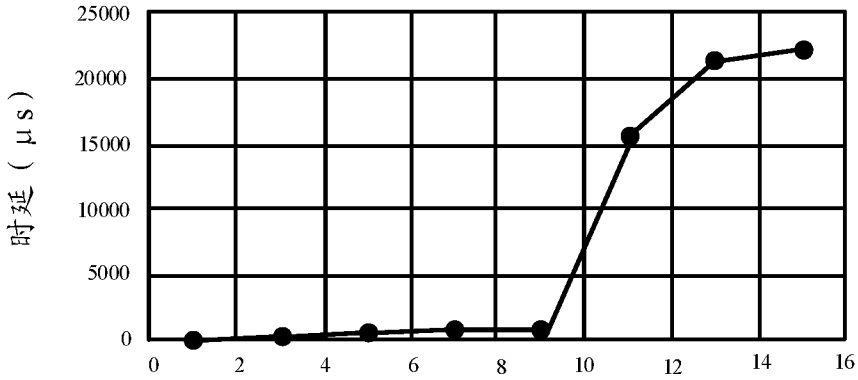

图 7

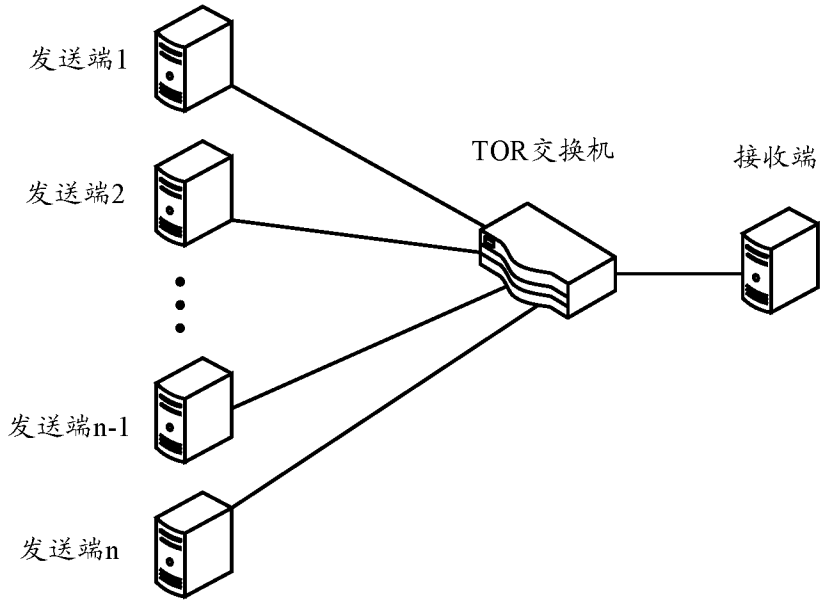

图 8

流速率与CNP数目变化图

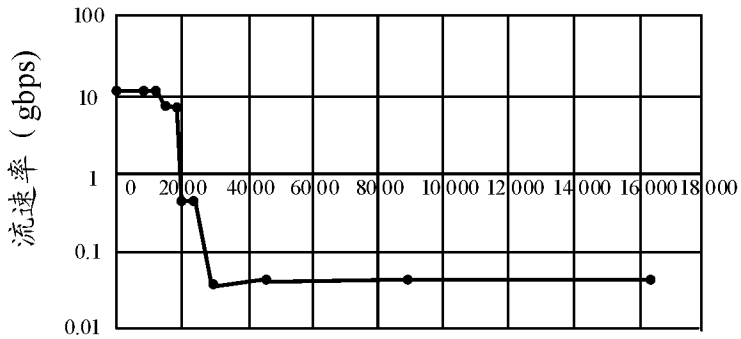

图 9

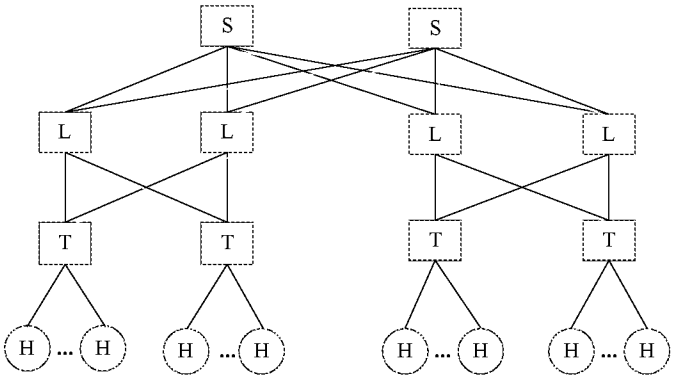
S
S
L
L
L
L
T
T
T
T
H ... H
H ... H
H ... H
H ... H

图 10

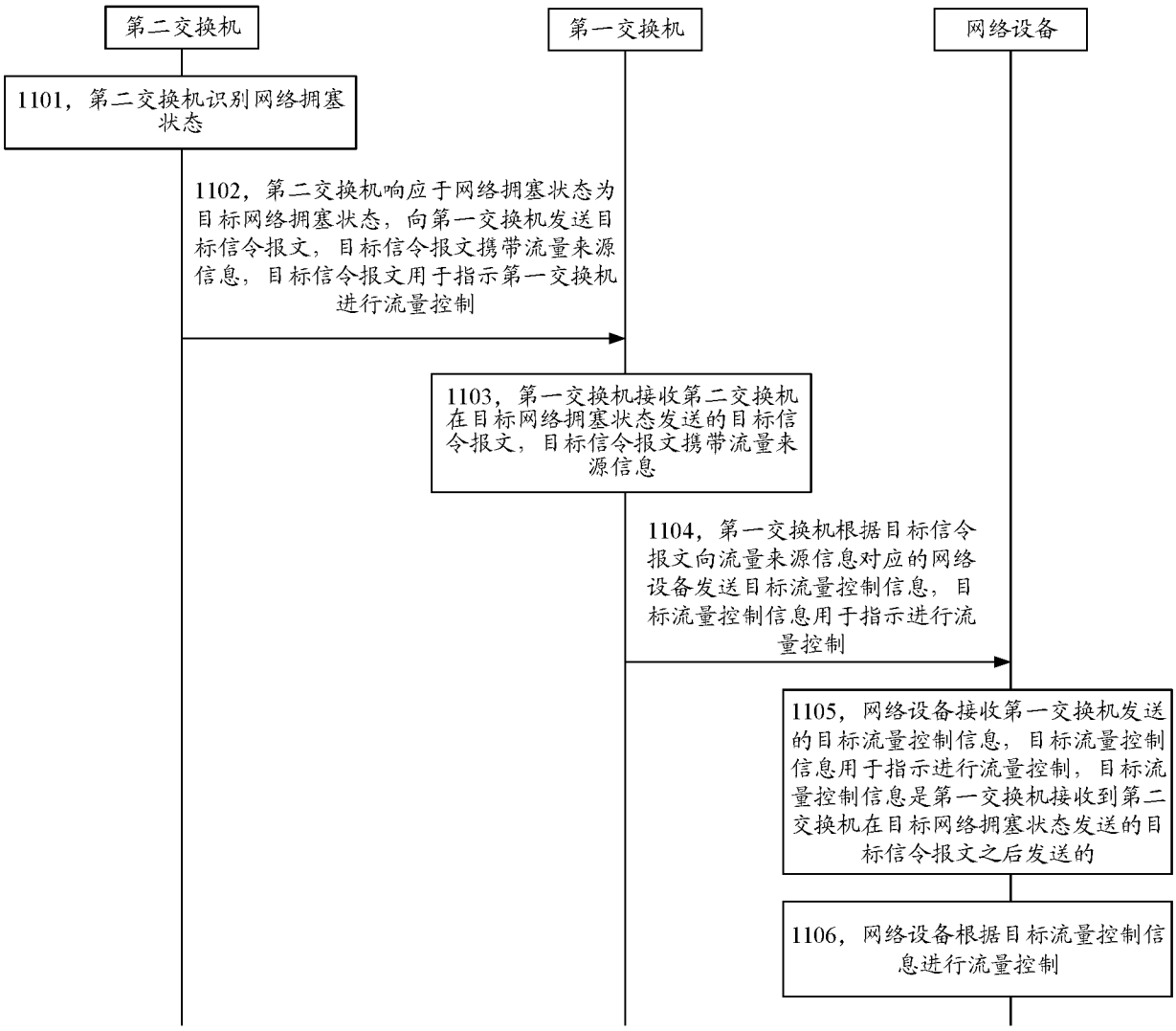
第二交换机
第一交换机
网络设备
1101，第二交换机识别网络拥塞状态
1102，第二交换机响应于网络拥塞状态为目标网络拥塞状态，向第一交换机发送目标信令报文，目标信令报文携带流量来源信息，目标信令报文用于指示第一交换机进行流量控制
1103，第一交换机接收第二交换机在目标网络拥塞状态发送的目标信令报文，目标信令报文携带流量来源信息
1104，第一交换机根据目标信令报文向流量来源信息对应的网络设备发送目标流量控制信息，目标流量控制信息用于指示进行流量控制
1105，网络设备接收第一交换机发送的目标流量控制信息，目标流量控制信息用于指示进行流量控制，目标流量控制信息是第一交换机接收到第二交换机在目标网络拥塞状态发送的目标信令报文之后发送的
1106，网络设备根据目标流量控制信息进行流量控制

图 11

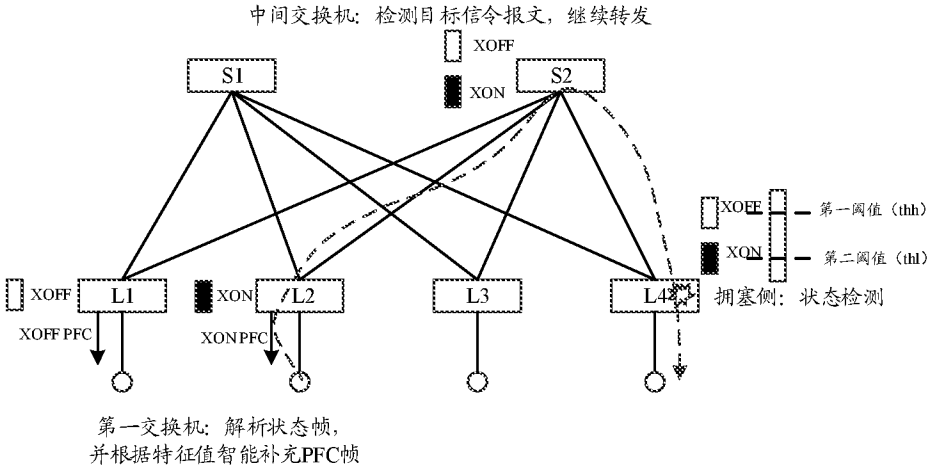
中间交换机：检测目标信令报文，继续转发
XOFF
XON
S1
S2
XOFF
第一阈值（thh）
XON
第二阈值（thl）
XOFF
L1
XON
L2
L3
L4
拥塞侧：状态检测
XOFF PFC
XON PFC
第一交换机：解析状态帧，
并根据特征值智能补充PFC帧

图 12

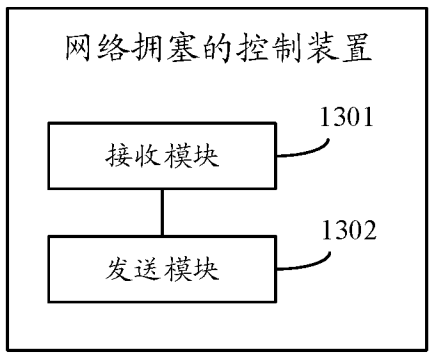
网络拥塞的控制装置
1301
接收模块
1302
发送模块

图 13

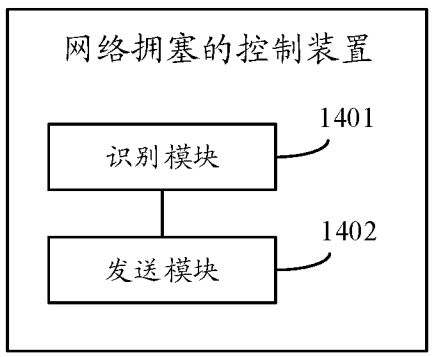
网络拥塞的控制装置
1401
识别模块
1402
发送模块

图 14

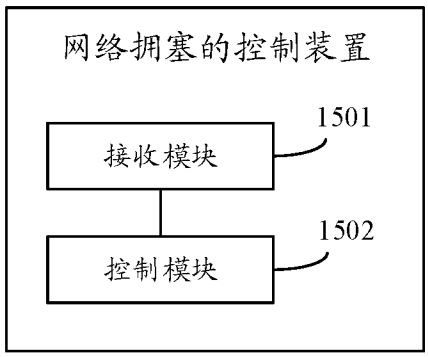
网络拥塞的控制装置
1501
接收模块
1502
控制模块

图 15

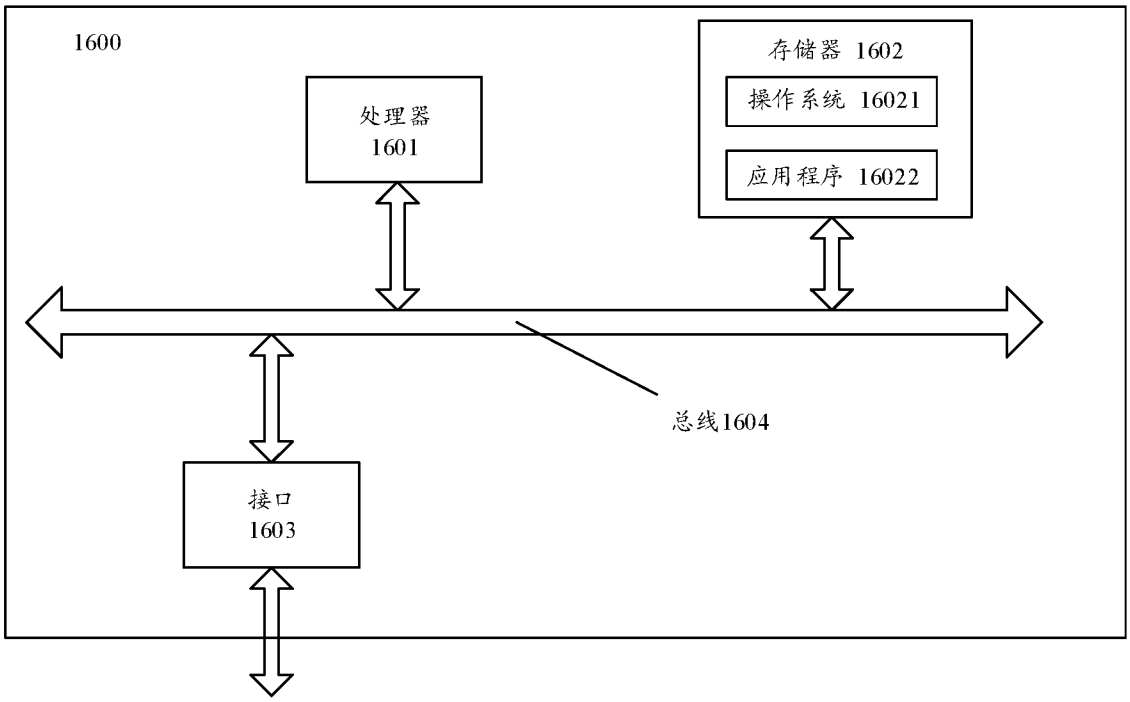
1600
处理器
1601
存储器 1602
操作系统 16021
应用程序 16022
总线1604
接口
1603

图 16

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/093165

A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/801(2013.01)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;H04W

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)

CNABS, CNTXT, CNKI, VEN, USTXT, WOTXT, EPTXT, 3GPP, IETF, IEEE: 网络拥塞, 流量, 控制, 显式拥塞通知, 拥塞通知包, 拥塞通知报文, 优先级的流量控制, 失效, 暂停, network, congest+, flow, control+, ECN, CNP, PFC, invalid+, paus???, break

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 109981471 A (HUAWEI TECHNOLOGIES CO., LTD.) 05 July 2019 (2019-07-05) description paragraphs 0032-0105	1-49
A	CN 106330742 A (HUAWEI TECHNOLOGIES CO., LTD.) 11 January 2017 (2017-01-11) entire document	1-49
A	CN 109802894 A (CHINA UNITED NETWORK COMMUNICATIONS GROUP CO., LTD.) 24 May 2019 (2019-05-24) entire document	1-49
A	US 2020021532 A1 (CISCO TECH. INC.) 16 January 2020 (2020-01-16) the whole document	1-49
A	KR 101992750 B1 (UNIST (ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY)) 25 June 2019 (2019-06-25) the whole document	1-49

☐ 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	Date of mailing of the international search report
23 July 2021	**11 August 2021**
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/CN2021/093165

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 109981471 A	05 July 2019	None	
CN 106330742 A	11 January 2017	CN 106330742 B	06 December 2019
CN 109802894 A	24 May 2019	None	
US 2020021532 A1	16 January 2020	US 2020296049 A1	17 September 2020
		US 10785161 B2	22 September 2020
KR 101992750 B1	25 June 2019	None	

国际检索报告

国际申请号
PCT/CN2021/093165

A. 主题的分类

H04L 12/801(2013.01)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;H04W

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））

CNABS,CNTXT,CNKI,VEN,USTXT,WOTXT,EPTXT,3GPP,IETF,IEEE:网络拥塞，流量，控制，显式拥塞通知，拥塞通知包，拥塞通知报文，优先级的流量控制，失效，暂停，network, congest+, flow, control+, ECN, CNP, PFC,invalid+, paus???, break

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
X	CN 109981471 A (华为技术有限公司) 2019年 7月 5日 (2019-07-05) 说明书第0032-0105段	1-49
A	CN 106330742 A (华为技术有限公司) 2017年 1月 11日 (2017-01-11) 全文	1-49
A	CN 109802894 A (中国联合网络通信集团有限公司) 2019年 5月 24日 (2019-05-24) 全文	1-49
A	US 2020021532 A1 (CISCO TECH INC) 2020年 1月 16日 (2020-01-16) the whole document	1-49
A	KR 101992750 B1 (ULSAN NAT INST SCIENCE & TECH UNIST) 2019年 6月 25日 (2019-06-25) the whole document	1-49

☐ 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	国际检索报告邮寄日期
2021年 7月 23日	2021年 8月 11日

ISA/CN的名称和邮寄地址	受权官员
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	李凡
传真号 (86-10)62019451	电话号码 86-(010)-62089572

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2021/093165

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	109981471	A	2019年 7月 5日		无		
CN	106330742	A	2017年 1月 11日	CN	106330742	B	2019年 12月 6日
CN	109802894	A	2019年 5月 24日		无		
US	2020021532	A1	2020年 1月 16日	US	2020296049	A1	2020年 9月 17日
				US	10785161	B2	2020年 9月 22日
KR	101992750	B1	2019年 6月 25日		无		