(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
2020 年 4 月 2 日 (02.04.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/063003 A1

(51) 国际专利分类号:
H04L 12/801 (2013.01)

(21) 国际申请号: PCT/CN2019/094171

(22) 国际申请日: 2019 年 7 月 1 日 (01.07.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811119579.8 2018年9月25日 (25.09.2018) 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: 刘 孟 竹 (**LIU, Mengzhu**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 张 亚 丽 (**ZHANG, Yali**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 刘世兴 (**LIU, Shixing**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (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, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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,

(54) **Title:** CONGESTION CONTROL METHOD, AND NETWORK APPARATUS

(54) **发明名称**: 拥塞控制方法和网络设备

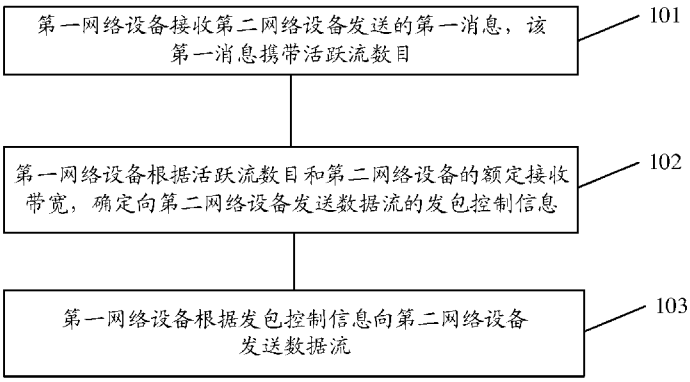


图2

101 A FIRST NETWORK APPARATUS RECEIVES A FIRST MESSAGE SENT BY A SECOND NETWORK APPARATUS, THE FIRST MESSAGE CARRYING A NUMBER OF ACTIVE FLOWS

102 THE FIRST NETWORK APPARATUS DETERMINES, ACCORDING TO THE NUMBER OF ACTIVE FLOWS AND A RATED RECEPTION BANDWIDTH OF THE SECOND NETWORK APPARATUS, PACKET SENDING CONTROL INFORMATION FOR SENDING DATA FLOWS TO THE SECOND NETWORK APPARATUS

103 THE FIRST NETWORK APPARATUS SENDS DATA FLOWS TO THE SECOND NETWORK APPARATUS ACCORDING TO THE PACKET SENDING CONTROL INFORMATION

(57) **Abstract:** The present application provides a congestion control method and a network apparatus, the method comprising: a first network apparatus receiving a first message sent by a second network apparatus, the first message carrying a number of active flows, the number of active flows being determined by the second network apparatus according to data flows containing data packets received from the first network apparatus; the first network apparatus determining, according to the number of active flows and a rated reception bandwidth of the second network apparatus, packet sending control information for sending data flows to the second network apparatus, wherein the packet sending control information is used to instruct the first network apparatus to send the data flows to the second network apparatus, such that an actual reception bandwidth of the second network apparatus reaches the rated reception bandwidth; and the first network apparatus sending the data flows to the second network apparatus according to the packet sending control information. The present application achieves better congestion control and accordingly reduces network packet loss.

WO 2020/063003 A1

[见续页]

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: 本申请提供了一种拥塞控制方法和网络设备，方法包括：第一网络设备接收第二网络设备发送的第一消息，第一消息携带活跃流数目，活跃流数目是第二网络设备根据接收到的、来自第一网络设备的数据包所归属的数据流确定的数目；第一网络设备根据活跃流数目和第二网络设备的额定接收带宽，确定向第二网络设备发送数据流的发包控制信息，其中，发包控制信息指示第一网络设备根据发包控制信息向第二网络设备发送数据流时，第二网络设备的实际接收带宽达到额定接收带宽；第一网络设备根据发包控制信息向第二网络设备发送数据流。本申请能够对拥塞进行更好地控制，进而减少网络丢包。

拥塞控制方法和网络设备

本申请要求于2018年09月25日提交中国国家知识产权局、申请号为201811119579.8、发明名称为“拥塞控制方法和网络设备”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

本申请涉及网络技术领域，并且更具体地，涉及一种拥塞控制方法和网络设备。

背景技术

随着人工智能、企业存储等新型应用的兴起，数据中心网络对低延时特性的需求也越来越迫切。而如何对数据中心网络进行拥塞控制，从而以保证数据中心网络的低延时特性是一个非常重要的问题。

传统的拥塞控制方法主要是由接收端向发送端反馈网络的拥塞信息，发送端在获知网络的拥塞信息之后，再根据拥塞信息对发送数据的速率进行调整。例如，当网络拥塞时，发送端会降低数据包的发送速率，当网络不拥塞时，发送端会逐步提高数据包的发送速率。

在上述方案中，发送端在得知网络的拥塞状态之后会逐步调整数据包的发送速率，速率的改变比较缓慢，可能会对数据包的传输过程中产生一定的延时，导致拥塞控制的效率不高。

发明内容

本申请提供一种拥塞控制方法和网络设备，能够更好地进行拥塞控制，进而减少网络丢包。

第一方面，提供了一种拥塞控制方法，所述方法应用于数据中心网络，所述数据中心网络包括第一网络设备和第二网络设备，所述第一网络设备用于向第二网络设备发送数据流，该方法包括：所述第一网络设备接收所述第二网络设备发送的第一消息，所述第一消息携带活跃流数目，所述活跃流数目是所述第二网络设备根据接收到的、来自所述第一网络设备的数据包所归属的数据流确定的数目；所述第一网络设备根据所述活跃流数目和所述第二网络设备的额定接收带宽，确定向所述第二网络设备发送所述数据流的发包控制信息，其中，所述发包控制信息指示所述第一网络设备根据所述发包控制信息向所述第二网络设备发送所述数据流时，所述第二网络设备的实际接收带宽达到所述额定接收带宽；所述第一网络设备根据所述发包控制信息向所述第二网络设备发送所述数据流。

上述活跃流数目可以是指在一定的统计周期内，正在传输数据包的数据流的数目，具体地，上述活跃流数目可以是第二网络设备根据一个统计周期内接收到的来自第一网络设备的数据包所归属的数据流确定的数据流的数目。

可选地，第一网络设备根据发包控制信息向第二网络设备发送的数据流的封装为基于融合以太的远程内存直接访问协议版本2（remote direct memory access over converged ethernet version 2，RoCEv2）报文。

可选地，上述第一消息携带在确认（Acknowledgement，ACK）报文中的option字段中。

应理解，第一网络设备根据发包控制信息向第二网络设备发送的数据流可以是一个数据流（此时，该数据流可以是第一网络设备向第二网络设备发送的所有数据流中的任意一个数据流），也可以是多个数据流（该多个数据流可以是第一网络设备向第二网络设备发送的所有的数据流）。

上述发包控制信息用于控制第一网络设备向第二网络设备发送数据流，并且，当第一

网络设备按照该发包控制信息向第二网络设备发送数据流时，第二网络设备的实际带宽能够达到第二网络设备的额定接收带宽。

第二网络设备的额定接收带宽可以是一个固定的带宽，在对第一网络设备进行配置时，可以直接将第二网络设备的额定接收带宽等信息配置给第一网络设备，使得第一网络设备能够从配置信息中获取第二网络设备的额定接收带宽。

本申请中，通过活跃流数目和第二网路设备的额定接收带宽来确定发包控制信息，并根据该发包控制信息对第一网络设备向第二网络设备发送数据流进行控制，能够更好地进行拥塞控制，从而减少网络丢包。

结合第一方面，在第一方面的某些实现方式中，所述第一网络设备根据所述活跃流数目和所述第二网络设备的额定接收带宽，确定向所述第二网络设备发送数据流的发包控制信息，包括：当所述活跃流数目小于第一阈值时，所述第一网络设备根据下列公式确定向所述第二网络设备发送所述数据流的发包控制信息；

$$V = \frac{C*T}{active_qp}$$

其中，C为所述第二网路设备的额定接收带宽，active_qp为所述活跃流数目，T是在网络空载的情况下，第一网络设备向所述第二网络设备发送的一个数据包达到所述第二网络设备所需的时间，V为T内所述第一网络设备向所述第二网络设备发送的所述数据流的数据量；所述第一网络设备根据所述发包控制信息向所述第二网络设备发送所述数据流，包括：所述第一网络设备在T内向所述第二网络设备发送的所述数据流中的数据包的数据量为V。

上述数据量（data volume）可以是数据流中的数据包的长度。例如，数据流1中包含数据包1、数据包2和数据包3，其中，每个数据包的长度为4个字节，那么，数据流1的长度为12个字节。

在获得上述T的获取方式有很多，下面对这些获取方式举例说明。

例如，可以在网络空载情况下，记录第一网络设备向第二网络设备发送的一个数据包的发送时间和到达时间，然后通过对到达时间和发送时间做差得到T。

另外，还可以在网络空载情况下，第一网络设备连续向第二网络设备发送多个数据包（例如，1000个数据包），然后通过记录的时间得到每个数据包到从第一网络设备发送到第二网络设备接收所需要的时间，接下来对这些时间进行平均，得到这些时间的平均值，最后将得到的平均值确定为T。

上述第一阈值可以根据经验来设置，例如，对于数据量较大的场景下，可以设置一个较大的数值作为第一阈值，而对于数据量较小的场景下，可以设置一个较小的数值作为第一阈值。

上述第一阈值还可以根据第二网络设备的额定接收带宽、T以及单个数据包的长度来确定。

具体地，可以将C*T/1pkt得到的数值确定为上述第一阈值。

其中，1pkt表示单个数据包的长度。1pkt的具体长度可以根据实际情况来设置，在不同的场景下，1pkt可以设置成不同的长度，例如，在某些数据传输量较小的场景下（例如，智能电表定期上报电量信息）可以为1pkt设置一个较小的长度，而在数据传输量较大的场景下，可以为1pkt设置一个较大的长度。另外，1pkt也可以直接设置为网络允许的最大传输单元（maximum transmission unit，MTU）。

结合第一方面，在第一方面的某些实现方式中，所述第一网络设备根据所述活跃流数目和所述第二网络设备的额定接收带宽，确定向所述第二网络设备发送数据流的发包控制信息，包括：当所述活跃流数目大于或者等于第一阈值时，所述第一网络设备根据下列公式确定向所述第二网络设备发送所述数据流的发包控制信息；

$$interval = \frac{1\text{pkt} * active_qp}{C}$$

其中，1pkt为单个数据包的长度，active_qp为所述活跃流数目，C为所述第二网路设备的额定接收带宽，interval为所述第一网络设备发送所述数据流时相邻数据包之间的时间间隔；所述第一网络设备根据所述发包控制信息向所述第二网络设备发送所述数据流，包括：所述第一网络设备以interval的时间间隔向所述第二网络设备发送所述数据流中的数据包。

上述单个数据包的长度具体可以是单个数据包的数量大小，例如，单个数据包的数据量大小为3字节，那么，单个数据包的长度也是3字节。

结合第一方面，在第一方面的某些实现方式中，所述第一网络设备根据所述活跃流数目和所述第二网络设备的额定接收带宽，确定向所述第二网络设备发送数据流的发包控制信息，包括：所述第一网络设备根据所述活跃流的数目、所述第二网络设备的额定接收带宽以及所述数据流的属性信息，确定向所述第二网络设备发送所述数据流的发包控制信息；其中，所述数据流的属性信息包括所述数据流的应用信息和所述数据流的数据量信息中的至少一种，所述数据流的应用信息用于指示所述数据流的数据包所属的业务类型，所述数据流的数据量信息用于指示所述第一网络设备在预设时间内发送的属于第一数据流的数据包的数据量。

本申请中，在确定数据流的发包控制信息时，能够根据数据流的属性信息为数据流合理确定发包控制信息。

结合第一方面，在第一方面的某些实现方式中，所述第一网络设备根据所述活跃流数目、所述第二网络设备的额定接收带宽以及所述数据流的属性信息，确定向所述第二网络设备发送所述数据流的发包控制信息，包括：当所述活跃流数目小于第一阈值时，所述第一网络设备根据下列公式确定向所述第二网络设备发送所述数据流的发包控制信息；

$$V = \frac{C * T * w}{active_qp}$$

其中，C为所述第二网路设备的额定接收带宽，active_qp为所述活跃流数目，T为在网络空载的情况下，所述第一网络设备向所述第二网络设备发送的一个数据包达到所述第二网络设备所需的时间，w为所述数据流的权重，w是根据所述数据流的属性信息确定的，V为T内所述第一网络设备向所述第二网络设备发送的所述数据流的数据量；所述第一网络设备根据所述发包控制信息向所述第二网络设备发送所述数据流，包括：所述第一网络设备在T内向所述第二网络设备发送的所述数据流中的数据包的数据量为V。

本申请中，在确定数据流的数据传输量时，能够根据数据流属性信息为数据流合理确定数据传输量。

结合第一方面，在第一方面的某些实现方式中，所述第一网络设备根据所述活跃流数目、所述第二网络设备的额定接收带宽以及所述数据流的属性信息，确定向所述第二网络设备发送所述数据流的发包控制信息，包括：当所述活跃流数目大于或者等于第一阈值时，所述第一网络设备根据下列公式确定向所述第二网络设备发送所述数据流的发包控制信息；

$$interval = \frac{1\text{pkt} * active_qp * w}{C}$$

其中，1pkt为单个数据包的长度，active_qp为所述活跃流数目，w为所述数据流的权重，w是根据所述数据流的属性信息确定的，C为所述第二网路设备的额定接收带宽，interval为所述第一网络设备发送所述数据流时相邻数据包之间的时间间隔；所述第一网络设备根据所述发包控制信息向所述第二网络设备发送所述数据流，包括：所述第一网络设备以interval的时间间隔向所述第二网络设备发送所述数据流中的数据包。

本申请中，在确定数据流的数据包传输时间间隔时，能够根据数据流属性信息为数据流合理确定数据包的传输时间间隔。

第二方面，提供了一种拥塞控制方法，所述方法应用于数据中心网络，所述数据中心网络包括第一网络设备和第二网络设备，所述第二网络设备用于接收所述第一网络设备发送的数据流，该方法包括：所述第二网络设备根据接收到的、来自所述第一网络设备的数据包所归属的数据流确定活跃流数目；所述第二网络设备向所述第一网络设备发送第一消息，所述第一消息携带所述活跃流数目，所述活跃流数目用于所述第一网络确定向所述第二网络设备发送数据流的发包控制信息，所述发包控制信息指示所述第一网络设备根据所述发包控制信息向所述第二网络设备发送所述数据流时，所述第二网络设备的实际接收带宽达到所述第二网络设备的额定接收带宽；所述第二网络设备按照所述额定接收带宽接收所述第一网络设备发送的所述数据流。

本申请中，第二网络设备通过向第一网络设备上报活跃流数目，使得第一网络设备能够通过活跃流数目和第二网路设备的额定接收带宽来确定发包控制信息，并根据该发包控制信息对第一网络设备向第二网络设备发送数据流进行控制，使得第二网络设备能够以额定接收带宽接收第一网络设备发送的数据流，从而达到满吞吐状态，能够提高数据流的传输效率。

结合第二方面，在第二方面的某些实现方式中，所述第二网络设备根据接收到的、来自所述第一网络设备的数据包所归属的数据流确定活跃流数目，包括：所述第二网络设备接收所述第一网络设备发送的第一组数据包的首包，并将当前活跃流数目加1，以得到第一活跃流数目；所述第二网络设备接收所述第一网络设备发送的所述第一组数据包的尾包，并根据所述第一组数据包中携带外部拥塞通知ECN标识的数据包的数量确定第一拥塞值；当所述第一拥塞值小于拥塞阈值时，所述第二网络设备将所述第一活跃流数目减1，以得到第二活跃流数目；当所述第一拥塞值大于或者等于所述拥塞阈值时，所述第二网络设备保持所述第一活跃流数目不变。

可选地，上述第一拥塞值用于指示第二网络设备接收到第一网络设备发送的数据包的第一网络的拥塞程度。

应理解，上述第一拥塞值越大，第一网络的拥塞程度越严重。

上述拥塞阈值可以是预先设置好的一个阈值，该拥塞阈值具体可以是根据网络状态估计的一个阈值。对于宽带利用率较低的场景，可以设置一下较小的拥塞阈值，而对于宽带利用率较高的场景，可以设置一个较大的拥塞阈值。

例如，在宽带利用率较低的场景下，拥塞阈值具体可以是0.3、0.4或者0.5等等，而在宽带利用率较高的场景下，拥塞阈值具体可以是0.6、0.7或者0.8等等。

应理解，在接收到第一组数据包的首包时需要将当前活跃流数目加1，以得到第一活

跃流数目。这里的当前活跃流数目是指第二网络设备在接收到第一组数据包的首包之前统计得到的活跃流的数目，在接收到第一组数据包的首包时需要更新活跃流的数目，也就是将接收到第一组数据包的首包之前统计得到的活跃流的数目加1。

而在接收到第一组数据包的尾包之后，第二网络设备需要根据第一组数据包中携带ECN标识的数据包的数量对第一活跃流数目进行再次更新，如果第一拥塞值小于拥塞阈值，那么，在接收到第一组数据包的尾包后第一网络设备统计得到的活跃流的数目为第二活跃流数目；如果第一拥塞值小于拥塞阈值，那么，在接收到第一组数据包的尾包后第一网络设备统计得到的活跃流的数目为第一活跃流数目。也就是说，在接收到第一组数据包的尾包后，统计得到的活跃流数目为第一活跃流数目（第一拥塞值大于或者等于拥塞阈值时）或者第二活跃流数目（第一拥塞值小于拥塞阈值时）。

上述ECN标识可以是第一网络设备和第二网络设备之间的其它网络设备（具体可以是交换设备）在接收到数据包时添加的一个拥塞标识。

可选地，每个数据包还可以在某个字段中携带一个标识，该标识的取值用于指示该数据包的类型（具体是首包、尾包还是首包与尾包之间的数据包）。

上述每个数据包还可以视为一个报文，每个数据包的标识可以携带在报文的某个字段中。例如，当数据包为基于融合以太的远程内存直接访问协议版本2（remote direct memory access over converged ethernet version 2，RoCEv2）报文时，数据包的标识可以携带在RoCEv2报文中的opcode字段中，或者，数据包的标识还可以携带的RoCEv2报文中的保留字段（例如，rsvd7字段）中。

本申请中，在接收到一组数据包的尾包时，可以根据该组数据包中携带ECN标识的数据包的个数来确定拥塞情况，进而可以根据拥塞情况对活跃流数目进行修正，能够对活跃流数目进行更精准的统计，进而得到更准确的活跃流数目。

具体地，传统方案在统计活跃流数目时并未考虑到网络拥塞情况，事实上，传统方案在进行活跃流数目统计是基于网络不出现拥塞的理想情况对活跃流进行统计的，当网络出现拥塞时会影响接收端在一段时间内接收到的数据包的数目，进而导致活跃流的数目统计不准。本申请中通过对每组数据包中携带ECN标识的数据包进行统计，能够实现对网络拥塞的估计，当网络拥塞程度较轻时，可以将活跃流数目减1，而当网络拥塞程度比较严重时可以保持活跃流数目不变，能够减少由于网络拥塞对活跃流数目统计的影响，能够实现更准确的统计。

结合第二方面，在第二方面的某些实现方式中，所述根据所述第一组数据包中携带ECN标识的数据包的数量确定第一拥塞值，包括：将所述第一组数据包携带有ECN标识的数据包的数量与所述第一组数据包的数据包数量的比值确定为所述第一拥塞值。

可选地，将所述第一组数据包携带有ECN标识的数据包的数量与所述第一组数据包的数据包数量的比值确定为所述第一拥塞值，包括：将所述第一组数据包携带有ECN标识的数据包的总数量与所述第一组数据包的数据包总数量的比值确定为所述第一拥塞值。

例如，第一组数据包共包含10个数据包（包括首包和尾包），其中，携带ECN标识的数据包共有5个，那么，可以计算出第一拥塞阈值为0.5。

另外，也可以先得到第一组数据包中携带有ECN标识的数据包的总数量与所述第一组数据包的数据包总数量的比值，然后再将该比值与一个修正系数的乘积作为第一拥塞阈值。该修正系数可以是根据网络的运行情况设置的一个系数。

本申请中，根据第一组数据包中携带ECN标识的数据包来确定第一拥塞值，能够实时的反映当前的拥塞情况，进而可以根据当前拥塞情况对活跃流数目作出较为准确的统计。

结合第二方面，在第二方面的某些实现方式中，在所述第二网络设备接收到所述第一网络设备发送的所述第一组数据包的尾包之后，所述方法还包括：所述第二网络设备接收所述第一网络设备发送的第二组数据包的首包，并将第三活跃流数目加1，以得到第四活跃流数目，其中，所述第二组数据包与所述第一组数据包属于同一个数据流，当所述第一拥塞值小于所述拥塞阈值时，所述第三活跃流数目等于所述第二活跃流数目，当所述第一拥塞值大于或者等于所述拥塞阈值时，所述第三活跃流数目等于所述第一活跃流数目；所述第二网络设备接收所述第一网络设备发送的所述第二组数据包的尾包，并根据所述第二组数据包中携带ECN标识的数据包的数量和所述第一拥塞值，确定所述网络的第二拥塞值；当所述第二拥塞值小于所述拥塞阈值时，所述第二网络设备将所述第四活跃流数目减1，以得到第五活跃流数目；当所述第二拥塞值大于或者等于所述拥塞阈值时，所述第二网络设备保持所述第四活跃流数目不变。

可选地，将所述第一组数据包携带有ECN标识的数据包的数量与所述第一组数据包的数据包数量的比值确定为所述第一拥塞值，包括：将所述第一组数据包携带有ECN标识的数据包的总数量与所述第一组数据包的数据包总数量的比值确定为所述第一拥塞值。

应理解，上述第一组数据包中的首包、尾包以及首包和尾包之间的数据包均可能携带ECN标识。当第一组数据包中的每个数据包均携带ECN标识时，说明网络拥塞比较严重，当第一组数据包中的携带ECN标识的数据包较小时，说明网络拥塞程度较低（或者说网络比较畅通）。

当接收到第二组数据包的首包之后，可以按照接收到第一组数据包的首包的类似的方式进行处理，而在接收到第二数据包的尾包之后，通过综合第二组数据包中携带ECN标识的数据包以及上述第一拥塞值来综合确定第二拥塞值，能够使得第二拥塞值不至于变化过大，从而得到缓慢变化的拥塞值，使得统计得到的活跃流数目不会产生突变。

可选地，当第二网络设备接收到所述第一网络设备发送的所述第二组数据包的尾包之后，还可以根据第一组数据包中携带ECN标识的数据包的数量和第二组数据包中携带ECN标识的数据包的数量来综合确定第二拥塞值。

可选地，将第一组数据包和第二组数据包中携带ECN标识的数据包的总数量与第一组数据包和第二组数据包包含的数据包的总数量的比值确定为第二拥塞值。

结合第二方面，在第二方面的某些实现方式中，所述根据所述第二组数据包中携带ECN标识的数据包的数量和所述第一拥塞值，确定所述网络的第二拥塞值，包括：根据所述第二组数据包中携带的ECN标识的数据包的数量确定所述网络的第三拥塞值；根据公式con2=x1*con3+x2* con1确定所述第二拥塞值，其中，con3为所述第三拥塞值，con1为所述第一拥塞值，con2为所述第二拥塞值，x1为预设的第一权重值，x2为预设的第二权重值。

应理解，在第二方面所示的方法中，活跃流数目的统计既可以由接收端的网络设备上执行，也可以由位于发送端和接收端之间的交换机设备上执行，当在交换机设备上执行时，交换机设备可以将统计得到的活跃流数目反馈给接收端，然后再由接收端给发送端。

第三方面，提供了一种网络设备，所述网络设备包含用于执行第一方面的各种实现方式或者第二方面的各种实现方式中的方法的模块。

第四方面，提供了一种网络设备，所述网络设备包括存储器和处理器。其中，存储器

用于存储程序，处理器用于执行存储器存储的程序，当存储器存储的程序被处理器执行时，处理器用于执行第一方面的各种实现方式或者第二方面的各种实现方式中的方法。

可选地，上述网络设备还包括收发器，当存储器存储的程序被处理器执行时，处理器和收发器用于执行第一方面的各种实现方式或者第二方面的各种实现方式中的方法。

第五方面，提供了一种计算机可读存储介质，所述计算机可读存储介质中存储有指令，当其在计算机上运行时，使得计算机执行第一方面的各种实现方式或者第二方面的各种实现方式中的方法。

第六方面，提供了一种包含指令的计算机程序产品，当其在计算机上运行时，使得计算机执行第一方面的各种实现方式或者第二方面的各种实现方式中的方法。

上述计算机具体可以是数据中心中的网络设备，例如，服务设备和传输设备等。

附图说明

图1是本申请实施例的一种可能的应用场景的示意图；

图2是本申请实施例的拥塞控制方法的示意性流程图；

图3是是本申请实施例的拥塞控制方法的示意性流程图；

图4是本申请实施例的拥塞控制方法的示意性流程图；

图5是是本申请实施例的拥塞控制方法的示意性流程图；

图6是本申请实施例的网络设备的示意性框图；

图7是本申请实施例的网络设备的示意性框图；

图8是本申请实施例的网络设备的示意性框图。

具体实施方式

图1是本申请实施例的一种可能的应用场景的示意图。

如图1所示，发送端1、发送端2和发送端3可以通过交换机1和交换机2向接收端1和接收端2发送数据流。接收端1和接收端2均可以通过交换机1和交换机2接收到发送端1至发送端3发送的数据流（交换机1和交换机2将发送端1至发送端3发送的数据流转发给接收端1和接收端2）。应理解，图1只是本申请实施例一种可能的应用场景的示意图，本申请实施例还可以应用到其他与图1类似的场景中。

以发送端1和接收端1为例，发送端1向接收端1发送的数据流可能在发送端1和接收端1之间产生拥塞，从而导致接收端1接收到的数据包的延时变大，影响网路性能。为了进行拥塞控制，传统方案中，传统方案中，接收端1会向发送端1反馈数据流的拥塞情况，当数据流出现拥塞时，发送端1会逐步降低数据流的发送速率，当数据流不拥塞时，发送端可以逐步提升数据包的发送速率。在传统方案中，发送端在得知网络的拥塞状态之后会逐步调整数据包的发送速率，速率的改变比较缓慢，并且是在获知了网络的拥塞状况之后进行的一种被动调节，拥塞控制的效果并不太好。

因此，本申请提出了一种新的拥塞控制方法，通过活跃流数目和接收端的额定接收带宽来主动调节数据流的发送速率，使得接收端的实际接收带宽能够达到额定接收带宽，使得接收端能够达到满吞吐状态，进而提高数据流的传输效率。

图2是本申请实施例的拥塞控制方法的示意性流程图。图2所示的方法可以应用于数据中心网络、城域网等网络中，第一网络设备和第二网络设备是这些网络中的设备。图2所示的方法包括步骤101至步骤103，下面对这些步骤进行详细的介绍。

101、第二网络设备向第一网络设备发送第一消息，第一网络设备接收该第一消息。

上述第一消息携带活跃流数目，该活跃流数目是第二网络设备根据接收到的、来自第一网络设备的数据包所归属的数据流确定的数目。

应理解，上述活跃流可以是第二网络设备自身对数据流的定义，对于第二网络设备来说，活跃流可以认为是传输数据包比较频繁的数据流，而非活跃流可以是传输数据包的时间间隔较长的数据流（例如，某个数据流很长时间才会传输一个数据包，这样的数据流可以认为是非活跃流）。因此，对于第二网络设备来说，第二网络设备接收到的数据流可以分成两种，即活跃流和非活跃流。

另外，上述活跃流数目可以是指在一定的统计周期内，正在传输数据包的数据流的数目，具体地，上述活跃流数目可以是第二网络设备根据一个统计周期内接收到的来自第一网络设备的数据包所归属的数据流确定的数据流的数目。

可选地，在步骤101之前，第一网络设备向第二网络设备发送数据流，第二网络设备接收第一网络设备发送的数据流。

第一网络设备在向第二网络设备发送数据流时，可按组向第二网络设备发送数据包，每组数据包包括首包核尾包，首包，尾包以及首包和尾包之间的数据包具有不同的标识，第二网络设备可以通过这些标志识别出一组数据包的首包和尾包，然后根据接收到的每组数据包的首包和尾包的情况对活跃流数目进行统计，以得到活跃流数目（具体统计活跃流的数目在下面图3所示的方法进行详细描述）。

应理解，当第一网络设备采用图1所示的拥塞控制方法对数据流的发送速率进行控制之前，可以按照初始速率向第二网络设备发送数据流。该初始速率可以按照以下两种方式来定义。

第一种方式：第一网络设备将第二网络设备的额定接收带宽确定为初始速率，并在t内以该额定接收带宽向第二网络设备发送数据流。

其中，t是从第一网络设备发送请求报文（该请求报文用于请求第二网络设备接收第一设备的发送数据流），到第一网络设备接收到第二网络设备发送的请求报文的响应报文的一段时间。

第二种方式：初始速率是基于融合以太的远程内存直接访问协议版本2（remote direct memory access over converged ethernet version 2，RoCEv2）定义的每流的初始发送速率。

具体地，作为一种可能的实现方式，初始速率可以是第一网络设备的线卡或者发送网板允许的最大的发送速率。

另外，每组数据包的首包和尾包也有两种定义方式，第一种定义方式是根据RoCEv2协议来定义，具体地，可以将数据报文按照消息（Message）来分组（该消息是应用层的逻辑分组），将消息的第一个报文视为首包，将消息的最后一个报文视为尾包，首包和尾包的具体可以定义在基于融合以太的远程内存直接访问（remote direct memory access over converged ethernet，RoCE）报文的基础传输头（Base Transport Header，BTH）的opcode字段中。

除了根据应用层的逻辑分组对数据报文的首包和尾包进行定义之外，还可以直接根据片或段（segment）对数据包的首包和尾包进行定义，具体地，可以将根据初始速率或者活跃流数目反馈的一个周期内对应的数据报文分成一组，将片或段中的第一个报文称为首包，将片或段的最后一个报文称为尾包，这两个字段可以定义在BTH头的保留字段（例如，rsvd7字段）中，每个字段的标记可以占用1bit。

可选地，上述第一消息中可以直接携带活跃流的具体数目，这样，第一网络设备可以直接根据第一消息获取活跃流数目。

102、第一网络设备根据活跃流数目和第二网络设备的额定接收带宽，确定向第二网络设备发送数据流的发包控制信息。

上述发包控制信息指示第一网络设备根据发包控制信息向所述第二网络设备发送数据流时，第二网络设备的实际接收带宽达到所述额定接收带宽。也就是说，上述发包控制信息用于控制第一网络设备向第二网络设备发送数据流，并且，当第一网络设备按照该发包控制信息向第二网络设备发送数据流时，第二网络设备的实际带宽能够达到第二网络设备的额定接收带宽。

第二网络设备的额定接收带宽可以是一个固定的带宽，在对第一网络设备进行配置时，可以直接将第二网络设备的额定接收带宽等信息配置给第一网络设备，使得第一网络设备能够从配置信息中获取第二网络设备的额定接收带宽。

103、第一网络设备根据发包控制信息向第二网络设备发送数据流，第二网络设备以额定接收带宽接收该数据流。

应理解，第一网络设备根据发包控制信息向第二网络设备发送的数据流可以是一个数据流（此时，该数据流可以是第一网络设备向第二网络设备发送的所有数据流中的任意一个数据流），也可以是多个数据流（该多个数据流可以是第一网络设备向第二网络设备发送的所有的数据流）。

本申请中，通过活跃流数目和第二网路设备的额定接收带宽来确定发包控制信息，并根据该发包控制信息对第一网络设备向第二网络设备发送数据流进行控制，使得第二网络设备能够达到满吞吐状态，可以提高数据流的传输效率。

在根据活跃流数目和第二网络设备的额定接收带宽确定发包控制信息时，可以根据活跃流数目的大小采用不同的方式来确定发包控制信息。

例如，当活跃流的数目较小时，网络对突发数据具有较强的容忍性，此时，可以仅控制数据流在一段时间内的最大数据发送量。而当活跃流的数目较多时，多流同步对网络的影响较大，为了避免突发的大量数据导致网络拥塞，这时可以控制网卡（该网卡位于接收端中，负责分配数据包包的发送周期）的分配周期，每周期单报文发送（也就是严格控制数据包之间的传输间隔，每个数据包的传输可以认为是一个周期）。

可选地，当活跃流数目小于第一阈值时，第一网络设备根据公式（1）确定向第二网络设备发送数据流的发包控制信息。

$$V = \frac{C * T}{active_qp} \qquad (1)$$

在上述公式（1）中，C为第二网路设备的额定接收带宽，active_qp为活跃流数目，T为第一网络设备发送的数据流的数据包到达第二网络设备所需的时间，V为T内第一网络设备向第二网络设备发送的数据流的数据量。

经过上述公式（1）确定了发包控制信息，该发包控制信息具体指示第一网络设备在T内向第二网络设备发送的任意一个数据流中的数据包的数据流为V。应理解，该发包控制信息此时只指示了在T内的发送的数据流中的数据包的数据量，但是对数据包之间的时间间隔并未限定，只要在T内在数据流中数据包的数据量达到V即可。

在根据上述公式（1）得到发包控制信息之后，第一网络设备根据发包控制信息向第

二网络设备发送数据流，包括：第一网络设备在T内向第二网络设备发送的数据流中的数据包的数据量为V。

上述数据量（data volume）可以是数据流中的数据包的长度。例如，数据流1中包含数据包1、数据包2和数据包3，其中，每个数据包的长度为4个字节，那么，数据流1的长度为12个字节。

上述T可以是预先设置好的，T具体可以在网络空载时按照某个速率发送数据包时，该数据到达到接收端所需要的时间。

或者，上述T具体也可以是往返时延（round-trip time，RTT），RTT表示从发送端发送数据开始，到发送端收到来自接收端的确认（接收端收到数据后便立即发送确认），总共经历的时延，RTT可以是在网络设备初始化时，在网络空载情况下通过测量静态时延得到的。

本申请中，当活跃流数目较小时，网络对突发数据有较强的容忍性，可以只控制数据流在一段时间内的最大数据发送量，能够简化控值数据流发送的操作。

可选地，当活跃流数目小于或者等于第一阈值时，第一网络设备根据公式（2）确定向第二网络设备发送数据流的发包控制信息。

$$interval = \frac{1\text{pkt} * active_qp}{C} \qquad (2)$$

在上述公式（2）中，1pkt为单个数据包的长度，active_qp为活跃流数目，C为第二网路设备的额定接收带宽，interval为第一网络设备发送数据流时相邻数据包之间的时间间隔。

经过上述公式（2）确定了发包控制信息，该发包控制信息具体指示第一网络设备发送每个数据流时要使得数据包之间的传输时间间隔达到interval。

在根据上述公式（2）得到发包控制信息之后，第一网络设备根据发包控制信息向第二网络设备发送数据流，包括：第一网络设备以interval的时间间隔向第二网络设备发送数据流中的数据包。

上述单个数据包的长度具体可以是单个数据包的数量大小，例如，单个数据包的数据量大小为3字节，那么，单个数据包的长度也是3字节。

上述公式（2）的推导过程如下：

为每个数据流评分带宽，得到每个数据流的发送速率为$\frac{C}{active_qp}$，而发送端发送的每个数据流的发送速率为$\frac{1\text{pkt}}{interval}$，为了使得接收端达到满吞吐状态，可以得到公式（3），对公式（3）进行变形可以得到上述公式（2）。

$$\frac{C}{active_qp} = \frac{1\text{pkt}}{interval} \qquad (3)$$

本申请中，当活跃流数目较多时，多流同步对网络来说是一个很大的负担，通过合理设置数据包之间的传输时间间隔，能够避免多流同步对网络造成的冲击，避免网络拥塞。

在根据活跃流数目和第二网络设备的额定接收带宽确定发包控制信息时，除了采用上述公式（1）或者公式（2）确定每个数据流的发包控制信息（这种情况下，每个数据流的发包控制信息的确定方式相同）之外，还可以针对数据流的具体情况（例如，优先级等等）来确定数据流的发包控制信息，在这种情况下，不同的数据流得到的发包控制信息可能不同。下面对这种确定数据流的发包控制信息的方式进行详细的描述。

可选地，第一网络设备根据活跃流数目和第二网络设备的额定接收带宽，确定向第二网络设备发送数据流的发包控制信息，包括：第一网络设备根据活跃流的数目、第二网络设备的额定接收带宽以及数据流的属性信息，确定向第二网络设备发送数据流的发包控制信息。

其中，数据流的属性信息包括数据流的应用信息和数据流的数据量信息中的至少一种，数据流的应用信息用于指示数据流的数据包所属的业务类型，数据流的数据量信息用于指示第一网络设备在预设时间内发送的属于第一数据流的数据包的数据量。

上述数据流的应用信息具体可以指示数据流的数据包所属的业务类型（例如，音频，视频等等），以及该业务类型对应的优先级等等。

本申请中，在确定数据流的发包控制信息时，能够根据数据流的属性信息为数据流合理确定发包控制信息。

可选地，当活跃流数目小于第一阈值时，第一网络设备根据公式（4）确定向第二网络设备发送数据流的发包控制信息。

$$V=\frac{C*T*w}{active_qp} \qquad (4)$$

在上述公式（4）中，C为第二网路设备的额定接收带宽，active_qp为活跃流数目，T为第一网络设备发送的数据流的数据包到达第二网络设备所需的时间，w为数据流的权重，w是根据数据流的属性信息确定的，V为T内第一网络设备向第二网络设备发送的数据流的数据量。

在根据上述公式（4）得到发包控制信息之后，第一网络设备根据发包控制信息向第二网络设备发送数据流，包括：第一网络设备在T内向第二网络设备发送的数据流中的数据包的数据量为V。

应理解，上述w可以根据数据流属性信息中的数据流的应用信息和/或数据流的数据量信息来确定。

可选地，在上述公式（4）中，w的取值与数据流的应用信息指示的业务类型的重要程度呈正相关的关系。

具体地，在上述公式（4）中数据流的应用信息指示的业务类型的重要程度越高，w的取值也越高。

可选地，在上述公式（4）中上述w的取值与数据流的应用信息指示的业务的优先级的高低呈正相关的关系。例如，数据流的应用信息指示的业务的优先级越高，w的取值也越高。

可选地，在上述公式（4）中，上述w的取值与数据流的数据量信息指示的数据量的大小为反相关关系。具体地，数据流的数据量信息指示的数据量越大，w的取值越小。

本申请中，在确定数据流的数据传输量时，能够根据数据流属性信息为数据流合理确定数据传输量。

可选地，当活跃流数目大于或者等于第一阈值时，第一网络设备根据公式（5）确定向第二网络设备发送数据流的发包控制信息。

$$interval=\frac{1\text{pkt}*active_qp*w}{C} \qquad (5)$$

在上述公式（5）中，1pkt为单个数据包的长度，active_qp为活跃流数目，w为数据流的权重，w是根据数据流的属性信息确定的，C为第二网路设备的额定接收带宽，interval

为第一网络设备发送数据流时相邻数据包之间的时间间隔.

在根据上述公式（5）得到发包控制信息之后，第一网络设备根据发包控制信息向第二网络设备发送数据流，包括：第一网络设备以interval的时间间隔向第二网络设备发送数据流中的数据包。

可选地，在上述公式（5）中，w的取值与数据流的应用信息指示的业务类型的重要程度呈反相关的关系。

具体地，在上述公式（5）中，数据流的应用信息指示的业务类型的重要程度越高，w的取值越小。

可选地，上述公式（5）中w的取值与数据流的应用信息指示的业务的优先级的高低呈反相关的关系。

例如，数据流的应用信息指示的业务的优先级越高，w的取值越小。

可选地，在上述公式（5）中，上述w的取值与数据流的数据量信息指示的数据量的大小为正相关关系。具体地，数据流的数据量信息指示的数据量越大，w的取值越大。

本申请中，在确定数据流的数据包传输时间间隔时，能够根据数据流属性信息为数据流合理确定数据包的传输时间间隔。

可选地，当上述w是根据数据流属性信息中的数据流的应用信息确定时，w=w1，w1的取值与数据流的应用信息指示的业务类型的重要程度呈正相关的关系。例如，数据流的应用信息指示的业务类型的重要程度越高，w1的取值也越高。此时，公式（4）可以变形为公式（6），这种情况下，可以直接根据公式（6）确定向第二网络设备发送数据流的发包控制信息。

$$V=\frac{C*T*w1}{active_qp} \qquad (6)$$

在上述公式（6）中，w1是根据数据流属性信息中的数据流的应用信息确定的，其它参数的含义与公式（4）中的含义相同。

可选地，当上述w是数据流的数据量信息来确定时，w=w2，w2的取值与数据流的数据量信息指示的数据量的大小为反相关关系。例如，数据流的数据量信息指示的数据量越大，w2的取值越小。此时，公式（4）可以变形为公式（7），这种情况下，可以直接根据公式（7）确定向第二网络设备发送数据流的发包控制信息。

$$V=\frac{C*T*w1}{active_qp} \qquad (7)$$

在上述公式（7）中，w2是根据数据流的数据量信息来确定的，其它参数的含义与公式（4）中的含义相同。

可选地，上述w可以根据数据流属性信息中的数据流的应用信息和数据流的数据量信息来确定时，w=w1*w2，其中，w1的取值与数据流的应用信息指示的业务类型的重要程度呈正相关的关系，w2的取值与数据流的数据量信息指示的数据量的大小为反相关关系。此时，公式（4）可以变形为公式（8），这种情况下，可以直接根据公式（8）确定向第二网络设备发送数据流的发包控制信息。

$$V=\frac{C*T*w1*w2}{active_qp} \qquad (8)$$

在上述公式（8）中，w1是根据数据流属性信息中的数据流的应用信息确定的，w2是根据数据流的数据量信息来确定的，其它参数的含义与公式（3）中的含义相同。

上文结合图2从发送端（第一网络设备）的角度对本申请实施例的拥塞控制方法进行了详细的描述，下面结合图3从接收端（第二网络设备）的角度对本申请实施例的拥塞控制方法进行详细的描述。

图3是是本申请实施例的拥塞控制方法的示意性流程图。图3所示的方法也可以应用于数据中心网络、城域网等网络中，第一网络设备和第二网络设备是这些网络中的设备。图3所示的方法包括步骤201至步骤203，下面对这些步骤进行详细的介绍。

201、第二网络设备根据接收到的、来自第一网络设备的数据包所归属的数据流确定活跃流数目。

202、第二网络设备向第一网络设备发送第一消息，第一消息携带活跃流数目。

上述活跃流数目用于第一网络确定向第二网络设备发送数据流的发包控制信息，发包控制信息指示第一网络设备根据发包控制信息向第二网络设备发送数据流时，第二网络设备的实际接收带宽达到第二网络设备的额定接收带宽。

203、第二网络设备按照额定接收带宽接收第一网络设备发送的数据流。

应理解，上述图2所示的方法中的步骤101至步骤103与图3所示的方法中的步骤201至203是相对应的，上文中对步骤101至步骤103的相关限定和解释也适用于步骤201至步骤203，为避免重复这里不再详细描述。

本申请中，第二网络设备通过向第一网络设备上报活跃流数目，使得第一网络设备能够通过活跃流数目和第二网路设备的额定接收带宽来确定发包控制信息，并根据该发包控制信息对第一网络设备向第二网络设备发送数据流进行控制，使得第二网络设备能够以额定接收带宽接收第一网络设备发送的数据流，从而达到满吞吐状态，能够提高数据流的传输效率。

可选地，作为一个实施例，第二网络设备根据接收到的、来自第一网络设备的数据包所归属的数据流确定活跃流数目，包括：第二网络设备接收第一网络设备发送的第一组数据包的首包，并将当前活跃流数目加1，以得到第一活跃流数目；第二网络设备接收第一网络设备发送的第一组数据包的尾包，并根据第一组数据包中携带外部拥塞通知ECN标识的数据包的数量确定第一拥塞值；当第一拥塞值小于拥塞阈值时，第二网络设备将第一活跃流数目减1，以得到第二活跃流数目；当第一拥塞值大于或者等于拥塞阈值时，第二网络设备保持第一活跃流数目不变。

下面结合图4和图5对第二网络设备根据接收到的、来自第一网络设备的数据包所归属的数据流确定活跃流数目进行详细的说明。

图4是本申请实施例的统计活跃流数的流程图。图4中的第二网络设备相当于图1中的接收端1或者接收端2，用于接收数据流，图4中的第一网络设备相当于图1中的发送端（发送端1或者发送端2或者发送端3），用于发送数据流。图4所示的过程至少可以包括步骤301至步骤304，进一步地，图4所示的过程还可以包括步骤305至步骤308，下面对这些步骤进行详细的介绍。

301、第二网络设备接收第一组数据包的首包。

本申请中，每组数据包包含一定数量的数据包，每组数据包包含首包和尾包，第二网络设备在接收到数据包之后可以通过识别该数据包是否携带首包标识和尾包标识来识别该数据包是否为首包或者尾包。

其中，首包标识和尾包标识的定义方式有两种，下面对这两种定义方式进行介绍。

第一种方式:

当本申请应用在RoCEv2场景下时，可以根据RoCEv2协议定义，将数据包按照消息（message）来分组。具体地，可以将数据包按照消息（Message）来分组（该消息是应用层的逻辑分组），将消息的第一个报文视为首包，将消息的最后一个报文视为尾包，首包和尾包的具体可以定义在RoCEv2报文的基础传输头（Base Transport Header，BTH）的opcode字段中。

第二种方式:

可以根据片或段（segment）对数据包进行分组，具体地，可以将根据初始速率或者活跃流数目反馈的一个周期内对应的数据包分成一组，将片或段中的第一个报文称为首包，将片或段的最后一个报文称为尾包，这两个字段可以定义在BTH头的保留字段（例如，rsvd7字段）中，每个字段的标记可以占用1bit。

302、第二网络设备将当前活跃流数目加1，得到第一活跃流数目。

步骤302中的当前活跃流数目是指第二网络设备在接收到第一组数据包的首包之前统计得到的活跃流的数目，在接收到第一组数据包的首包时需要更新活跃流的数目，也就是将接收到第一组数据包的首包之前统计得到的活跃流的数目加1。

303、第二网络设备接收第一组数据包的尾包。

应理解，上述第一组数据包的首包可以是第二网络设备接收到的第一组数据包中的第一个数据包，第一组数据包的尾包可以是第二网络设备接收到的第一组数据包中的最后一个数据包。每组数据包的数据包组成数量可以是预先设置好的。

另外，每个数据包可以在某个字段中携带一个标识，该标识的取值用于指示该数据包的类型（具体是首包、尾包还是首包与尾包之间的数据包）。

可选地，每个数据包还包括数据包标识，该数据包标识用于指示数据包的类型。具体地，该标识的取值可以用于指示该数据包是一组数据包的首包、尾包或者位于首包和尾包之间的数据包。

具体地，上述数据包标识可以携带在报文的报文头或者报文的有效载荷（payload）中，或者，上述数据包标识还可以携带在报文之外的字段中，携带该数据包标识的字段与报文一起传输，使得第一网络设备能够根据该字段识别数据包的类型。

例如，在基于融合以太的远程内存直接访问协议版本2（remote direct memory access over converged ethernet version 2，RoCEv2）的场景下，数据包标识可以携带在RoCEv2报文中的opcode字段中，或者，数据包标识还可以携带的RoCEv2报文中的保留字段（例如，rsvd7字段）中。

应理解，上述第一组数据包中的数据包均属于同一个数据流。

304、第二网络设备根据第一组数据包中携带外部拥塞通知ECN标识的数据包的数量确定第一拥塞值，并根据第一拥塞值域拥塞阈值的关系对活跃流数目进行更新。

可选地，在步骤304中，需要根据第一组数据包中携带ECN标识的数据包的数量对第一活跃流数目进行再次更新，如果第一拥塞值小于拥塞阈值，那么，在接收到第一组数据包的尾包后第二网络设备统计得到的活跃流的数目为第二活跃流数目；如果第一拥塞值小于拥塞阈值，那么，在接收到第一组数据包的尾包后第二网络设备统计得到的活跃流的数目为第一活跃流数目。也就是说，在接收到第一组数据包的尾包后，统计得到的活跃流数目为第一活跃流数目（第一拥塞值大于或者等于拥塞阈值时）或者第二活跃流数目（第一

拥塞值小于拥塞阈值时）。

本申请中，在接收到一组数据包的尾包时，可以根据该组数据包中携带ECN标识的数据包的个数来确定拥塞情况，进而可以根据拥塞情况对活跃流数目进行修正，能够对活跃流数目进行更精准的统计，进而得到更准确的活跃流数目。

具体而言，传统方案在统计活跃流数目时并未考虑到网络拥塞情况，事实上，传统方案在进行活跃流数目统计是基于网络不出现拥塞的理想情况对活跃流进行统计的，当网络出现拥塞时会影响接收端在一段时间内接收到的数据包的数目，进而导致活跃流的数目统计不准。本申请中通过对每组数据包中携带ECN标识的数据包进行统计，能够实现对网络拥塞的估计，当网络拥塞程度较轻时，可以将活跃流数目减1，而当网络拥塞程度比较严重时可以保持活跃流数目不变，能够减少由于网络拥塞对活跃流数目统计的影响，能够实现更准确的统计。

可选地，作为一个是实施例，根据第一组数据包中携带ECN标识的数据包的数量确定第一拥塞值，包括：将第一组数据包携带有ECN标识的数据包的数量与第一组数据包的数据包数量的比值确定为第一拥塞值。

具体地，可以将第一组数据包携带有ECN标识的数据包的总数量与所述第一组数据包的数据包数量的比值确定为第一拥塞值。

例如，第一组数据包共包含10个数据包（包括首包和尾包），其中，携带ECN标识的数据包共有5个，那么，可以计算出第一拥塞值为0.5。

另外，还可以先得到第一组数据包中携带有ECN标识的数据包的总数量与所述第一组数据包的数据包总数量的比值，然后再将该比值与一个修正系数的乘积作为第一拥塞值。该修正系数可以是根据网络的运行情况设置的一个系数。

例如，第一组数据包共包含10个数据包（包括首包和尾包），其中，携带ECN标识的数据包共有8个，那么，可以计算出携带有ECN标识的数据包的总数量与数据包的总数量的比值为0.8，假设修正系数为0.8，那么，该比值与修正系数的乘积0.64就是第一拥塞值。

可选地，在上述步骤302之前，图4所示的过程还包括：确定第一组数据包所在的数据流为第一数据流，其中，第一数据流为第一网络设备在预设时间内接收的数据包的数量大于预设数量的数据流。

上述第一数据流可以认为是大数据流，本申请中，通过仅对第一数据流进行活跃流数目的统计（大数据流适合采用本申请方案进行活跃流数目统计），使得本申请的统计活跃流数目的方案更有针对性。

具体地，由于小数据流中的数据包的数量较少，当网络发生拥塞时，小数据流受到的影响较小，而大数据流中的数据包的数量较多，当网络发生拥塞时，受到的影响较大。因此，对于小数据流，不统计活跃流（也就是说，对于小数据流来说即便是能够接收到数据包，也不将其视为活跃流），而对于大数据流，可以采用本申请的方案进行活跃流数目的统计。

上述步骤301至304仅示出了对根据一组数据包对活跃流数目进行统计的情况，事实上，本申请实施例的活跃流数目统计方法还可以根据多组数据包对活跃流数目进行统计。

可选地，在图4所示的过程中，当第二网络设备通过步骤301至步骤304实现了对第一组数据包的接收以及活跃流数目的更新后，还可以继续接收第二组数据包，并根据第二组数据包的接收情况对活跃流数目再次更新。

305、第二网络设备接收第二组数据包的首包。

应理解，第二组数据包和第一组数据包属于同一数据流，第二组数据包可以是第二网络设备在接收到第一组数据包之后接收到的一组数据包，具体地，第二组数据包和第一组数据包可以是数据流中连续的两组数据包（也就是说，第二组数据包是第二网络设备在接收到第一组数据包之后紧接着又接收到的下一组数据包）。

306、第二网络设备将当前活跃流数目加1，得到第四活跃流数目。

307、第二网络设备接收第二组数据包的尾包。

上述步骤305至步骤307与上文中的步骤301至步骤303的具体过程类似，这里不再赘述。

308、第二网络设备根据第二组数据包中携带外部拥塞通知ECN标识的数据包的数量确定第二拥塞值，并根据第二拥塞值域拥塞阈值的关系对活跃流数目进行更新。

当接收到第二组数据包的首包之后，可以按照接收到第一组数据包的首包的类似的方式进行处理，而在接收到第二数据包的尾包之后，通过综合第二组数据包中携带ECN标识的数据包以及上述第一拥塞值来综合确定第二拥塞值，能够使得第二拥塞值不至于变化过大，从而得到缓慢变化的拥塞值。

可选地，当第二网络设备接收到所述第一网络设备发送的所述第二组数据包的尾包之后，还可以根据第一组数据包中携带ECN标识的数据包的数量和第二组数据包中携带ECN标识的数据包的数量来综合确定第二拥塞值。

可选地，将第一组数据包和第二组数据包中携带ECN标识的数据包的总数量与第一组数据包和第二组数据包包含的数据包的总数量的比值确定为第二拥塞值。

例如，第一组数据包和第二组数据包均包含10个数据包，其中，第一组数据包中携带ECN标识的数据包为4个，第二组数据包中携带ECN标识的数据包为6个，那么，第一组数据包和第二组数据包携带ECN标识的数据包总数为10个，第一组数据包和第二组数据包的数据包总数量为20个，那么，得到第二拥塞值为10/20=0.5。

应理解，在确定第二拥塞值时，还可以根据第一组数据包中携带ECN标识的数据包的数量得到一个拥塞值，然后再根据第二组数据包中携带ECN标识的数据包的数量得到另一个拥塞值，接下来，将这两个拥塞值进行加权求和得到第二拥塞值。

可选地，作为一个实施例，根据第二组数据包中携带ECN标识的数据包的数量和第一拥塞值，确定网络的第二拥塞值，包括：根据第二组数据包中携带的ECN标识的数据包的数量确定网络的第三拥塞值；根据公式（9）确定第二拥塞值。

$$con2=x1*con3+x2*con1 \quad (9)$$

在上述公式（9）中，con3为第三拥塞值，con1为第一拥塞值，con2为第二拥塞值，x1为预设的第一权重值，x2为预设的第二权重值。

应理解，在本申请中，第二网络设备每接收到一组数据包（例如，前面的第一组数据包和第二组数据包）的首包时都要将当前活跃流的数目加1；当第二网络设备每接收到一组数据包的尾包时要根据该组数据中携带ECN标识的数据包的数量对活跃流数目进行调整（将活跃流数目减1或者保持不变），以实现对活跃流数目的实时更新。

为了更好地理解本申请实施例的拥塞控制方法，下面从第二网络设备的角度对本申请实施例的统计活跃数据流数目的方法进行介绍。

图5是本申请实施例的统计活跃流数的流程图。图5所示的过程至少可以包括步骤401至步骤406，进一步地，图5所示的过程还可以包括步骤407至步骤411，下面对这些步骤进

行详细的介绍。

401、开始。

步骤401表示开始对活跃流数目进行统计。步骤401可以发生在接收完一组数据包并对活跃流数目进行更新之后，接下来再接收另一组数据包之前。

402、第二网络设备接收第一网络设备发送的第一组数据包的首包，并执行counter++。

其中，counter是第二网络设备统计得到的当前活跃流数目。具体地，counter是第二网络设备在接收到第一网络设备发送的第一组数据包的首包时第二网络设备统计得到的当前活跃的数据流的数目。

步骤402中的counter++之后得到的counter是接收到第一组数据包的首包，并对当前活跃流数目进行更新后的活跃流的数目，counter++之后的counter相当于步骤102中得到的第一活跃流数目。

403、第二网络设备接收第一组数据包的尾包，并根据第一组数据包中携带外部拥塞通知ECN标识的数据包的数量确定第一拥塞值。

404、确定第一拥塞值是否大于拥塞阈值。

当第一拥塞值小于或者等于拥塞阈值时，说明网络拥塞程度较轻，此时需要将当前活跃流数目减1，也就是执行步骤405；而当第一拥塞值大于或者等于拥塞阈值时，说明网络拥塞较为严重，此时需要保持当前活跃流数目不变，也就是执行步骤406。

405、执行couner--。

在步骤405中，执行counter—得到的counter相当于上文中的第二活跃流数目。

406、保持counter不变。

步骤406中，counter相当于上文中的第一活跃流数目。

407、第二网络设备接收第一网络设备发送的第二组数据包的首包，并执行counter++。

步骤407中执行counter++之后得到的counter相当于上文中的第四活跃流数目。

408、第二网络设备接收第二组数据包的尾包，并根据第一组数据包中携带ECN标识的数据包的数量和第二组数据包中携带ECN标识的数据包的数量确定第二拥塞值。

409、确定第二拥塞值是否大于拥塞阈值。

当第二拥塞值小于或者等于拥塞阈值时，说明网络拥塞程度较轻，此时需要将当前活跃流数目减1，也就是执行步骤210；而当第二拥塞值大于或者等于拥塞阈值时，说明网络拥塞较为严重，此时需要保持当前活跃流数目不变，也就是执行步骤411。

410、执行couner--。

在步骤410中，执行counter--得到的counter相当于上文中的第五活跃流数目。

411、保持counter不变。

步骤406中，counter相当于上文中的第四活跃流数目。

应理解，在图5所示的过程中的步骤401至步骤411中，counter均表示当前的活跃流数目，经过不同的操作步骤需要对counter进行加操作或者减操作，或者保持counter不变，以实时的统计出当前活跃流数目。

上文结合图1至图5对本申请实施例的拥塞控制方法进行了详细的描述。下面结合图6至图8对本申请实施例的网络设备进行详细的描述，应理解，图6至图8所示的网络设备能够执行本申请实施例的拥塞控制方法（图6所示的网络设备500相当于上文中的第一网络设备，网络设备500可以执行上文中第一网络设备执行的各个步骤，图7所示的网络设备600

相当于上文中的第二网络设备，网络设备600可以执行上文中的第二网络设备执行的各个步骤），为了简洁，下面在介绍图6至图8所示的网络设备时，适当省略重复的描述。

图6是本申请实施例的网络设备的示意性框图。图6中的网络设备500包括:

接收模块501，用于接收第二网络设备发送的第一消息，第一消息携带活跃流数目，活跃流数目是第二网络设备根据接收到的、来自第一网络设备的数据包所归属的数据流确定的数目;

处理模块502，用于根据活跃流数目和第二网络设备的额定接收带宽，确定向第二网络设备发送数据流的发包控制信息，其中，发包控制信息指示第一网络设备根据发包控制信息向第二网络设备发送数据流时，第二网络设备的实际接收带宽达到额定接收带宽;

发送模块503，用于根据发包控制信息向第二网络设备发送数据流。

可选地，作为一个实施例，处理模块502用于：当活跃流数目小于第一阈值时，根据下列公式确定向第二网络设备发送数据流的发包控制信息;

$$V=\frac{C*T}{active_qp}$$

其中，C为第二网路设备的额定接收带宽，active_qp为活跃流数目，T为第一网络设备发送的数据流的数据包到达第二网络设备所需的时间，V为T内第一网络设备向第二网络设备发送的数据流的数据量;

发送模块503用于在T内向第二网络设备发送的数据流中的数据包的数据量为V。

可选地，作为一个实施例，处理模块502用于：当活跃流数目大于或者等于第一阈值时，根据下列公式确定向第二网络设备发送数据流的发包控制信息;

$$interval=\frac{1\text{pkt}*active_qp}{C}$$

其中，1pkt为单个数据包的长度，active_qp为活跃流数目，C为第二网路设备的额定接收带宽，interval为第一网络设备发送数据流时相邻数据包之间的时间间隔;

发送模块503用于以interval的时间间隔向第二网络设备发送数据流中的数据包。

可选地，作为一个实施例，处理模块502用于根据活跃流的数目、第二网络设备的额定接收带宽以及数据流的属性信息，确定向第二网络设备发送数据流的发包控制信息；其中，数据流的属性信息包括数据流的应用信息和数据流的数据量信息中的至少一种，数据流的应用信息用于指示数据流的数据包所属的业务类型，数据流的数据量信息用于指示第一网络设备在预设时间内发送的属于第一数据流的数据包的数据量。

可选地，作为一个实施例，处理模块502用于：当活跃流数目小于第一阈值时，根据下列公式确定向第二网络设备发送数据流的发包控制信息;

$$V=\frac{C*T*w}{active_qp}$$

其中，C为第二网路设备的额定接收带宽，active_qp为活跃流数目，T为第一网络设备发送的数据流的数据包到达第二网络设备所需的时间，w为数据流的权重，w是根据数据流的属性信息确定的，V为T内第一网络设备向第二网络设备发送的数据流的数据量;

发送模块503用于在T内向第二网络设备发送的数据流中的数据包的数据量为V。

可选地，作为一个实施例，处理模块502用于：当活跃流数目大于或者等于第一阈值时，根据下列公式确定向第二网络设备发送数据流的发包控制信息;

$$interval = \frac{1\text{pkt} * active_qp * w}{C}$$

其中，1pkt为单个数据包的长度，active_qp为活跃流数目，w为数据流的权重，w是根据数据流的属性信息确定的，C为第二网路设备的额定接收带宽，interval为第一网络设备发送数据流时相邻数据包之间的时间间隔；

发送模块503用于以interval的时间间隔向第二网络设备发送数据流中的数据包。

图7是本申请实施例的网络设备的示意性框图。图7中的网络设备600包括：

处理模块601，用于根据接收到的、来自第一网络设备的数据包所归属的数据流确定活跃流数目；

发送模块602，用于向第一网络设备发送第一消息，第一消息携带活跃流数目，活跃流数目用于第一网络确定向第二网络设备发送数据流的发包控制信息，发包控制信息指示第一网络设备根据发包控制信息向第二网络设备发送数据流时，第二网络设备的实际接收带宽达到第二网络设备的额定接收带宽；

接收模块603，用于按照额定接收带宽接收第一网络设备发送的数据流。

可选地，作为一个实施例，接收模块603用于接收第一网络设备发送的第一组数据包的首包；处理模块601用于将当前活跃流数目加1，以得到第一活跃流数目；接收模块603用于接收第一网络设备发送的第一组数据包的尾包；处理模块601用于：根据第一组数据包中携带外部拥塞通知ECN标识的数据包的数量确定第一拥塞值；当第一拥塞值小于拥塞阈值时，第二网络设备将第一活跃流数目减1，以得到第二活跃流数目；当第一拥塞值大于或者等于拥塞阈值时，第二网络设备保持第一活跃流数目不变。

可选地，作为一个实施例，处理模块601用于将第一组数据包携带有ECN标识的数据包的数量与第一组数据包的数据包数量的比值确定为第一拥塞值。

可选地，作为一个实施例，在接收模块603接收到第一网络设备发送的第一组数据包的尾包之后，接收模块603还用于接收第一网络设备发送的第二组数据包的首包；处理模块601用于将第三活跃流数目加1，以得到第四活跃流数目，其中，第二组数据包与第一组数据包属于同一个数据流，当第一拥塞值小于拥塞阈值时，第三活跃流数目等于第二活跃流数目，当第一拥塞值大于或者等于拥塞阈值时，第三活跃流数目等于第一活跃流数目；

接收模块603用于接收第一网络设备发送的第二组数据包的尾包；处理模块601还用于：根据第二组数据包中携带ECN标识的数据包的数量和第一拥塞值，确定网络的第二拥塞值；当第二拥塞值小于拥塞阈值时，第二网络设备将第四活跃流数目减1，以得到第五活跃流数目；当第二拥塞值大于或者等于拥塞阈值时，第二网络设备保持第四活跃流数目不变。

可选地，作为一个实施例，处理模块601用于：根据第二组数据包中携带的ECN标识的数据包的数量确定网络的第三拥塞值；根据公式con2=x1*con3+x2* con1确定第二拥塞值，其中，con3为第三拥塞值，con1为第一拥塞值，con2为第二拥塞值，x1为预设的第一权重值，x2为预设的第二权重值。

图8是本申请实施例的网络设备的示意性框图。

图8中的网络设备700包括：存储器701，收发器702和处理器703。其中，处理器703可以是中央处理器（central processing unit，CPU）、通用处理器，数字信号处理器（digital signal processor，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、晶体管逻

辑器件、硬件部件或者其任意组合。上述存储器701用于存储程序，处理器703可以执行存储器701中存储的程序，当存储器701中存储的程序被处理器703执行时，处理器703用于执行本申请实施例的拥塞控制方法。具体地，收发器702和处理器703可以用于执行上文中由第一网络设备或者第二网络设备执行的各个步骤。

本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统、装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统、装置和方法，可以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，所述单元的划分，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，机械或其它的形式。

所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，在本申请各个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。

所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器（read-only memory，ROM）、随机存取存储器（random access memory，RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以所述权利要求的保护范围为准。

权 利 要 求

1、一种拥塞控制方法，所述方法应用于数据中心网络，所述数据中心网络包括第一网络设备和第二网络设备，所述第一网络设备用于向第二网络设备发送数据流，其特征在于，包括:

所述第一网络设备接收所述第二网络设备发送的第一消息，所述第一消息携带活跃流数目，所述活跃流数目是所述第二网络设备根据接收到的、来自所述第一网络设备的数据包所归属的数据流确定的数目;

所述第一网络设备根据所述活跃流数目和所述第二网络设备的额定接收带宽，确定向所述第二网络设备发送所述数据流的发包控制信息，其中，所述发包控制信息指示所述第一网络设备根据所述发包控制信息向所述第二网络设备发送所述数据流时，所述第二网络设备的实际接收带宽达到所述额定接收带宽;

所述第一网络设备根据所述发包控制信息向所述第二网络设备发送所述数据流。

2、如权利要求1所述的方法，其特征在于，所述第一网络设备根据所述活跃流数目和所述第二网络设备的额定接收带宽，确定向所述第二网络设备发送数据流的发包控制信息，包括:

当所述活跃流数目小于第一阈值时，所述第一网络设备根据下列公式确定向所述第二网络设备发送所述数据流的发包控制信息;

$$V = \frac{C * T}{active_qp}$$

其中，C为所述第二网路设备的额定接收带宽，active_qp为所述活跃流数目，T为在网络空载的情况下，所述第一网络设备向所述第二网络设备发送的一个数据包达到所述第二网络设备所需的时间，V为T内所述第一网络设备向所述第二网络设备发送的所述数据流的数据量;

所述第一网络设备根据所述发包控制信息向所述第二网络设备发送所述数据流，包括:

所述第一网络设备在T内向所述第二网络设备发送的所述数据流中的数据包的数据量为V。

3、如权利要求1或2所述的方法，其特征在于，所述第一网络设备根据所述活跃流数目和所述第二网络设备的额定接收带宽，确定向所述第二网络设备发送数据流的发包控制信息，包括:

当所述活跃流数目大于或者等于第一阈值时，所述第一网络设备根据下列公式确定向所述第二网络设备发送所述数据流的发包控制信息;

$$interval = \frac{1\text{pkt} * active_qp}{C}$$

其中，1pkt为单个数据包的长度，active_qp为所述活跃流数目，C为所述第二网路设备的额定接收带宽，interval为所述第一网络设备发送所述数据流时相邻数据包之间的时间间隔;

所述第一网络设备根据所述发包控制信息向所述第二网络设备发送所述数据流，包括:

所述第一网络设备以interval的时间间隔向所述第二网络设备发送所述数据流中的数据包。

4、如权利要求1所述的方法，其特征在于，所述第一网络设备根据所述活跃流数目和

所述第二网络设备的额定接收带宽，确定向所述第二网络设备发送数据流的发包控制信息，包括:

所述第一网络设备根据所述活跃流的数目、所述第二网络设备的额定接收带宽以及所述数据流的属性信息，确定向所述第二网络设备发送所述数据流的发包控制信息;

其中，所述数据流的属性信息包括所述数据流的应用信息和所述数据流的数据量信息中的至少一种，所述数据流的应用信息用于指示所述数据流的数据包所属的业务类型，所述数据流的数据量信息用于指示所述第一网络设备在预设时间内发送的属于第一数据流的数据包的数据量。

5、如权利要求4所述的方法，其特征在于，所述第一网络设备根据所述活跃流数目、所述第二网络设备的额定接收带宽以及所述数据流的属性信息，确定向所述第二网络设备发送所述数据流的发包控制信息，包括:

当所述活跃流数目小于第一阈值时，所述第一网络设备根据下列公式确定向所述第二网络设备发送所述数据流的发包控制信息;

$$V = \frac{C * T * w}{active_qp}$$

其中，C为所述第二网路设备的额定接收带宽，active_qp为所述活跃流数目，T为在网络空载的情况下，所述第一网络设备向所述第二网络设备发送的一个数据包达到所述第二网络设备所需的时间，w为所述数据流的权重，w是根据所述数据流的属性信息确定的，V为T内所述第一网络设备向所述第二网络设备发送的所述数据流的数据量;

所述第一网络设备根据所述发包控制信息向所述第二网络设备发送所述数据流，包括:

所述第一网络设备在T内向所述第二网络设备发送的所述数据流中的数据包的数据量为V。

6、如权利要求4或5所述的方法，其特征在于，所述第一网络设备根据所述活跃流数目、所述第二网络设备的额定接收带宽以及所述数据流的属性信息，确定向所述第二网络设备发送所述数据流的发包控制信息，包括:

当所述活跃流数目大于或者等于第一阈值时，所述第一网络设备根据下列公式确定向所述第二网络设备发送所述数据流的发包控制信息;

$$interval = \frac{1\text{pkt} * active_qp * w}{C}$$

其中，1pkt为单个数据包的长度，active_qp为所述活跃流数目，w为所述数据流的权重，w是根据所述数据流的属性信息确定的，C为所述第二网路设备的额定接收带宽，interval为所述第一网络设备发送所述数据流时相邻数据包之间的时间间隔;

所述第一网络设备根据所述发包控制信息向所述第二网络设备发送所述数据流，包括:

所述第一网络设备以interval的时间间隔向所述第二网络设备发送所述数据流中的数据包。

7、一种拥塞控制方法，所述方法应用于数据中心网络，所述数据中心网络包括第一网络设备和第二网络设备，所述第二网络设备用于接收所述第一网络设备发送的数据流，其特征在于，包括:

所述第二网络设备根据接收到的、来自所述第一网络设备的数据包所归属的数据流确定活跃流数目;

所述第二网络设备向所述第一网络设备发送第一消息，所述第一消息携带所述活跃流

数目，所述活跃流数目用于所述第一网络确定向所述第二网络设备发送数据流的发包控制信息，所述发包控制信息指示所述第一网络设备根据所述发包控制信息向所述第二网络设备发送所述数据流时，所述第二网络设备的实际接收带宽达到所述第二网络设备的额定接收带宽；

所述第二网络设备按照所述额定接收带宽接收所述第一网络设备发送的所述数据流。

8、如权利要求7所述的方法，其特征在于，所述第二网络设备根据接收到的、来自所述第一网络设备的数据包所归属的数据流确定活跃流数目，包括：

所述第二网络设备接收所述第一网络设备发送的第一组数据包的首包，并将当前活跃流数目加1，以得到第一活跃流数目；

所述第二网络设备接收所述第一网络设备发送的所述第一组数据包的尾包，并根据所述第一组数据包中携带外部拥塞通知ECN标识的数据包的数量确定第一拥塞值；

当所述第一拥塞值小于拥塞阈值时，所述第二网络设备将所述第一活跃流数目减1，以得到第二活跃流数目；

当所述第一拥塞值大于或者等于所述拥塞阈值时，所述第二网络设备保持所述第一活跃流数目不变。

9、如权利要求8所述的方法，其特征在于，所述根据所述第一组数据包中携带ECN标识的数据包的数量确定第一拥塞值，包括：

将所述第一组数据包携带有ECN标识的数据包的数量与所述第一组数据包的数据包数量的比值确定为所述第一拥塞值。

10、如权利要求8或9所述的方法，其特征在于，在所述第二网络设备接收到所述第一网络设备发送的所述第一组数据包的尾包之后，所述方法还包括：

所述第二网络设备接收所述第一网络设备发送的第二组数据包的首包，并将第三活跃流数目加1，以得到第四活跃流数目，其中，所述第二组数据包与所述第一组数据包属于同一个数据流，当所述第一拥塞值小于所述拥塞阈值时，所述第三活跃流数目等于所述第二活跃流数目，当所述第一拥塞值大于或者等于所述拥塞阈值时，所述第三活跃流数目等于所述第一活跃流数目；

所述第二网络设备接收所述第一网络设备发送的所述第二组数据包的尾包，并根据所述第二组数据包中携带ECN标识的数据包的数量和所述第一拥塞值，确定所述网络的第二拥塞值；

当所述第二拥塞值小于所述拥塞阈值时，所述第二网络设备将所述第四活跃流数目减1，以得到第五活跃流数目；

当所述第二拥塞值大于或者等于所述拥塞阈值时，所述第二网络设备保持所述第四活跃流数目不变。

11、如权利要求10所述的方法，其特征在于，所述根据所述第二组数据包中携带ECN标识的数据包的数量和所述第一拥塞值，确定所述网络的第二拥塞值，包括：

根据所述第二组数据包中携带的ECN标识的数据包的数量确定所述网络的第三拥塞值；

根据公式con2=x1*con3+x2* con1确定所述第二拥塞值，其中，con3为所述第三拥塞值，con1为所述第一拥塞值，con2为所述第二拥塞值，x1为预设的第一权重值，x2为预设的第二权重值。

12、一种第一网络设备，其特征在于，所述第一网络设备包括:

接收模块，用于接收第二网络设备发送的第一消息，所述第一消息携带活跃流数目，所述活跃流数目是所述第二网络设备根据接收到的、来自所述第一网络设备的数据包所归属的数据流确定的数目;

处理模块，用于根据所述活跃流数目和所述第二网络设备的额定接收带宽，确定向所述第二网络设备发送所述数据流的发包控制信息，其中，所述发包控制信息指示所述第一网络设备根据所述发包控制信息向所述第二网络设备发送所述数据流时，所述第二网络设备的实际接收带宽达到所述额定接收带宽;

发送模块，用于根据所述发包控制信息向所述第二网络设备发送所述数据流。

13、如权利要求12所述的第一网络设备，其特征在于，所述处理模块用于:

当所述活跃流数目小于第一阈值时，根据下列公式确定向所述第二网络设备发送所述数据流的发包控制信息;

$$V = \frac{C * T}{active_qp}$$

其中，C为所述第二网路设备的额定接收带宽，active_qp为所述活跃流数目，T为在网络空载的情况下，所述第一网络设备向所述第二网络设备发送的一个数据包达到所述第二网络设备所需的时间，V为T内所述第一网络设备向所述第二网络设备发送的所述数据流的数据量;

所述发送模块用于在T内向所述第二网络设备发送的所述数据流中的数据包的数据量为V。

14、如权利要求12或13所述的第一网络设备，其特征在于，所述处理模块用于:

当所述活跃流数目大于或者等于第一阈值时，根据下列公式确定向所述第二网络设备发送所述数据流的发包控制信息;

$$interval = \frac{1\text{pkt} * active_qp}{C}$$

其中，1pkt为单个数据包的长度，active_qp为所述活跃流数目，C为所述第二网路设备的额定接收带宽，interval为所述第一网络设备发送所述数据流时相邻数据包之间的时间间隔;

所述发送模块用于以interval的时间间隔向所述第二网络设备发送所述数据流中的数据包。

15、如权利要求12所述的第一网络设备，其特征在于，所述处理模块用于根据所述活跃流的数目、所述第二网络设备的额定接收带宽以及所述数据流的属性信息，确定向所述第二网络设备发送所述数据流的发包控制信息;

其中，所述数据流的属性信息包括所述数据流的应用信息和所述数据流的数据量信息中的至少一种，所述数据流的应用信息用于指示所述数据流的数据包所属的业务类型，所述数据流的数据量信息用于指示所述第一网络设备在预设时间内发送的属于第一数据流的数据包的数据量。

16、如权利要求15所述的第一网络设备，其特征在于，所述处理模块用于:

当所述活跃流数目小于第一阈值时，根据下列公式确定向所述第二网络设备发送所述数据流的发包控制信息;

$$V = \frac{C * T * w}{active_qp}$$

其中，C为所述第二网路设备的额定接收带宽，active_qp为所述活跃流数目，T为在网络空载的情况下，所述第一网络设备向所述第二网络设备发送的一个数据包达到所述第二网络设备所需的时间，w为所述数据流的权重，w是根据所述数据流的属性信息确定的，V为T内所述第一网络设备向所述第二网络设备发送的所述数据流的数据量；

所述发送模块在T内向所述第二网络设备发送的所述数据流中的数据包的数据量为V。

17、如权利要求15或16所述的第一网络设备，其特征在于，所述处理模块用于：

当所述活跃流数目大于或者等于第一阈值时，根据下列公式确定向所述第二网络设备发送所述数据流的发包控制信息；

$$interval = \frac{1\text{pkt} * active_qp * w}{C}$$

其中，1pkt为单个数据包的长度，active_qp为所述活跃流数目，w为所述数据流的权重，w是根据所述数据流的属性信息确定的，C为所述第二网路设备的额定接收带宽，interval为所述第一网络设备发送所述数据流时相邻数据包之间的时间间隔；

所述发送模块用于以interval的时间间隔向所述第二网络设备发送所述数据流中的数据包。

18、一种第二网络设备，其特征在于，包括：

处理模块，用于根据接收到的、来自所述第一网络设备的数据包所归属的数据流确定活跃流数目；

发送模块，用于向所述第一网络设备发送第一消息，所述第一消息携带所述活跃流数目，所述活跃流数目用于所述第一网络确定向所述第二网络设备发送数据流的发包控制信息，所述发包控制信息指示所述第一网络设备根据所述发包控制信息向所述第二网络设备发送所述数据流时，所述第二网络设备的实际接收带宽达到所述第二网络设备的额定接收带宽；

接收模块，用于按照所述额定接收带宽接收所述第一网络设备发送的所述数据流。

19、如权利要求18所述的第二网络设备，其特征在于，所述接收模块用于接收所述第一网络设备发送的第一组数据包的首包；

所述处理模块用于将当前活跃流数目加1，以得到第一活跃流数目；

所述接收模块用于接收所述第一网络设备发送的所述第一组数据包的尾包；

所述处理模块用于：

根据所述第一组数据包中携带外部拥塞通知ECN标识的数据包的数量确定第一拥塞值；

当所述第一拥塞值小于拥塞阈值时，所述第二网络设备将所述第一活跃流数目减1，以得到第二活跃流数目；

当所述第一拥塞值大于或者等于所述拥塞阈值时，所述第二网络设备保持所述第一活跃流数目不变。

20、如权利要求19述的第二网络设备，其特征在于，所述处理模块用于将所述第一组数据包携带有ECN标识的数据包的数量与所述第一组数据包的数据包数量的比值确定为所述第一拥塞值。

21、如权利要求19或20所述的第二网络设备，其特征在于，在所述接收模块接收到所述第一网络设备发送的所述第一组数据包的尾包之后，所述接收模块还用于接收所述第一网络设备发送的第二组数据包的首包；

所述处理模块用于将第三活跃流数目加1，以得到第四活跃流数目，其中，所述第二组数据包与所述第一组数据包属于同一个数据流，当所述第一拥塞值小于所述拥塞阈值时，所述第三活跃流数目等于所述第二活跃流数目，当所述第一拥塞值大于或者等于所述拥塞阈值时，所述第三活跃流数目等于所述第一活跃流数目；

所述接收模块用于接收所述第一网络设备发送的所述第二组数据包的尾包；

所述处理模块还用于：

根据所述第二组数据包中携带ECN标识的数据包的数量和所述第一拥塞值，确定所述网络的第二拥塞值；

当所述第二拥塞值小于所述拥塞阈值时，所述第二网络设备将所述第四活跃流数目减1，以得到第五活跃流数目；

当所述第二拥塞值大于或者等于所述拥塞阈值时，所述第二网络设备保持所述第四活跃流数目不变。

22、如权利要求21所述的第二网络设备，其特征在于，所述处理模块用于：

根据所述第二组数据包中携带的ECN标识的数据包的数量确定所述网络的第三拥塞值；

根据公式con2=x1*con3+x2* con1确定所述第二拥塞值，其中，con3为所述第三拥塞值，con1为所述第一拥塞值，con2为所述第二拥塞值，x1为预设的第一权重值，x2为预设的第二权重值。

23、一种第一网络设备，其特征在于，所述第一网络设备包括：

存储器，用于存储程序；

处理器，用于执行所述存储器存储的程序，当所述存储器存储的程序被执行时，所述第一网络设备用于执行如权利要求1-6中任一项所述的方法。

24、一种第二网络设备，其特征在于，所述第二网络设备包括：

存储器，用于存储程序；

处理器，用于执行所述存储器存储的程序，当所述存储器存储的程序被执行时，所述第二网络设备用于执行如权利要求7-11中任一项所述的方法。

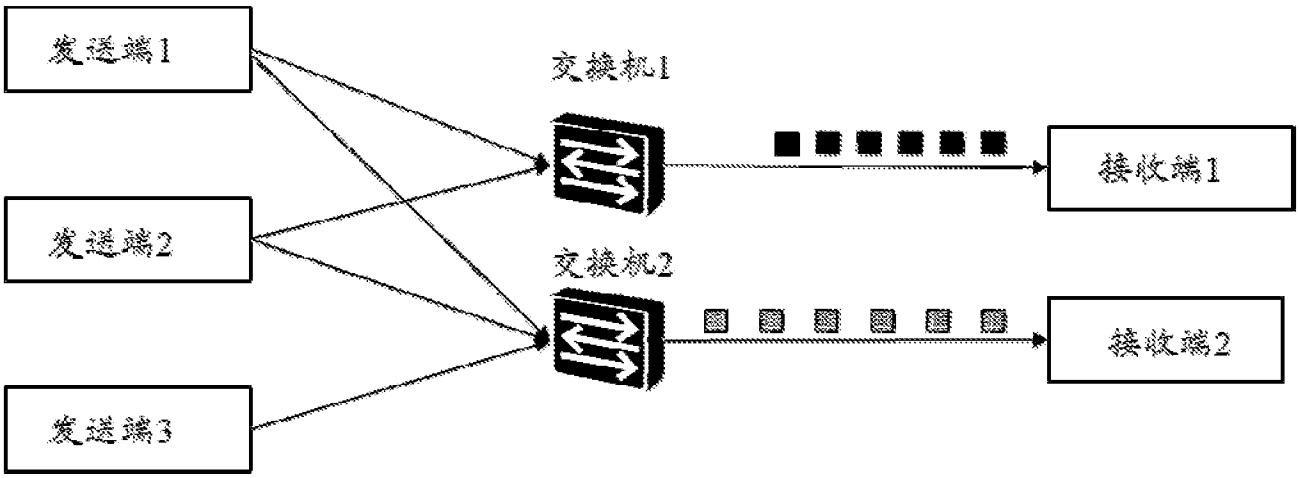

发送端1
发送端2
发送端3
交换机1
交换机2
接收端1
接收端2

图1

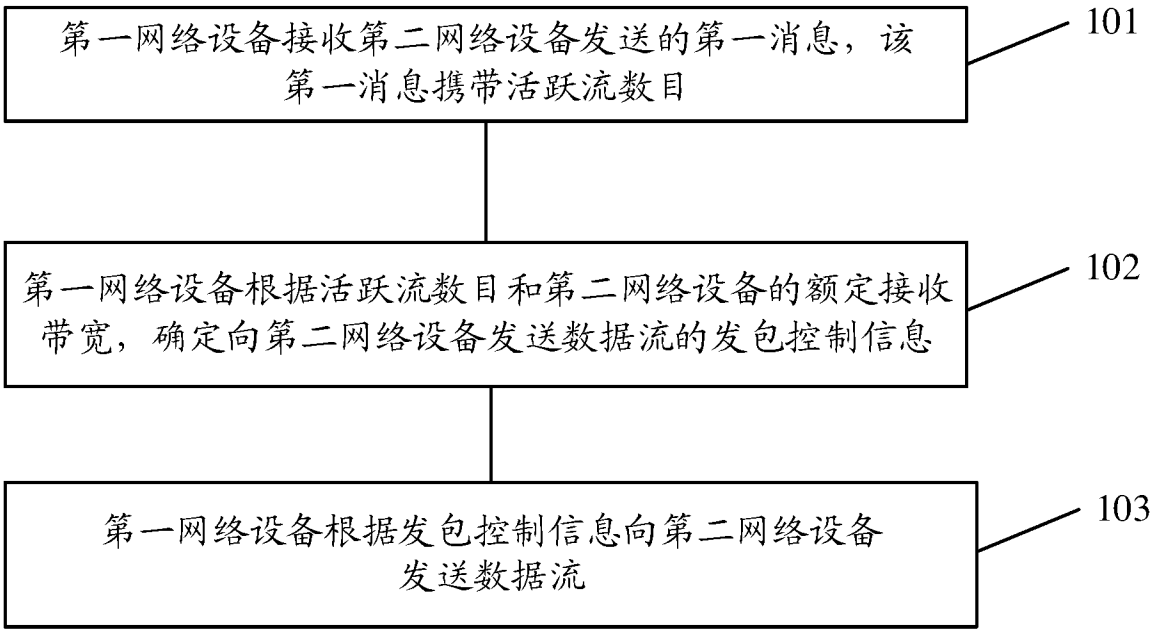

第一网络设备接收第二网络设备发送的第一消息，该第一消息携带活跃流数目
101
第一网络设备根据活跃流数目和第二网络设备的额定接收带宽，确定向第二网络设备发送数据流的发包控制信息
102
第一网络设备根据发包控制信息向第二网络设备发送数据流
103

图2

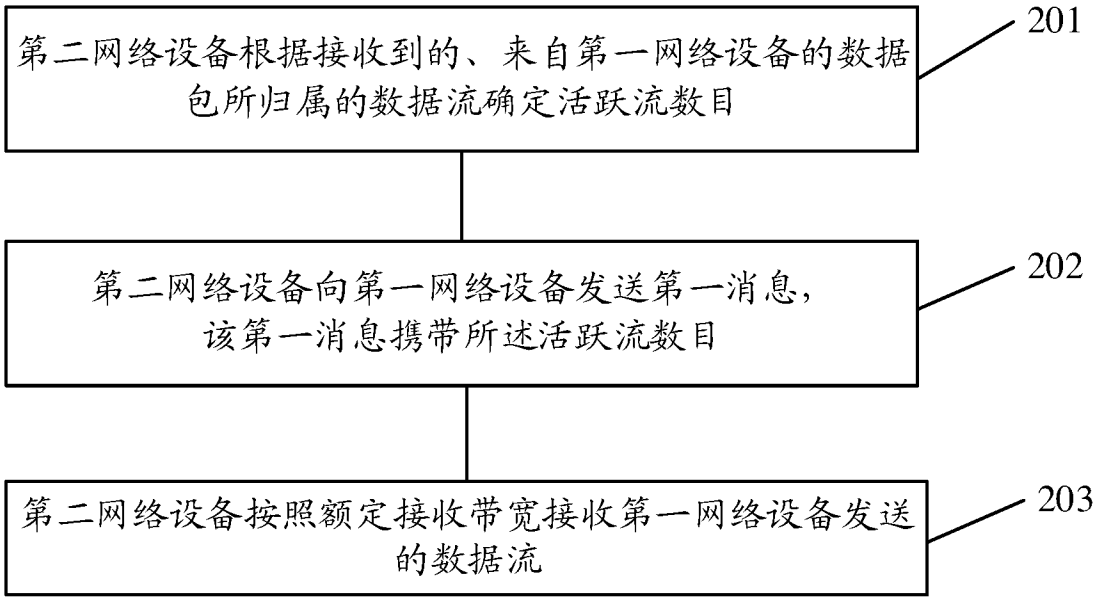

第二网络设备根据接收到的、来自第一网络设备的数据包所归属的数据流确定活跃流数目
201
第二网络设备向第一网络设备发送第一消息，该第一消息携带所述活跃流数目
202
第二网络设备按照额定接收带宽接收第一网络设备发送的数据流
203

图3

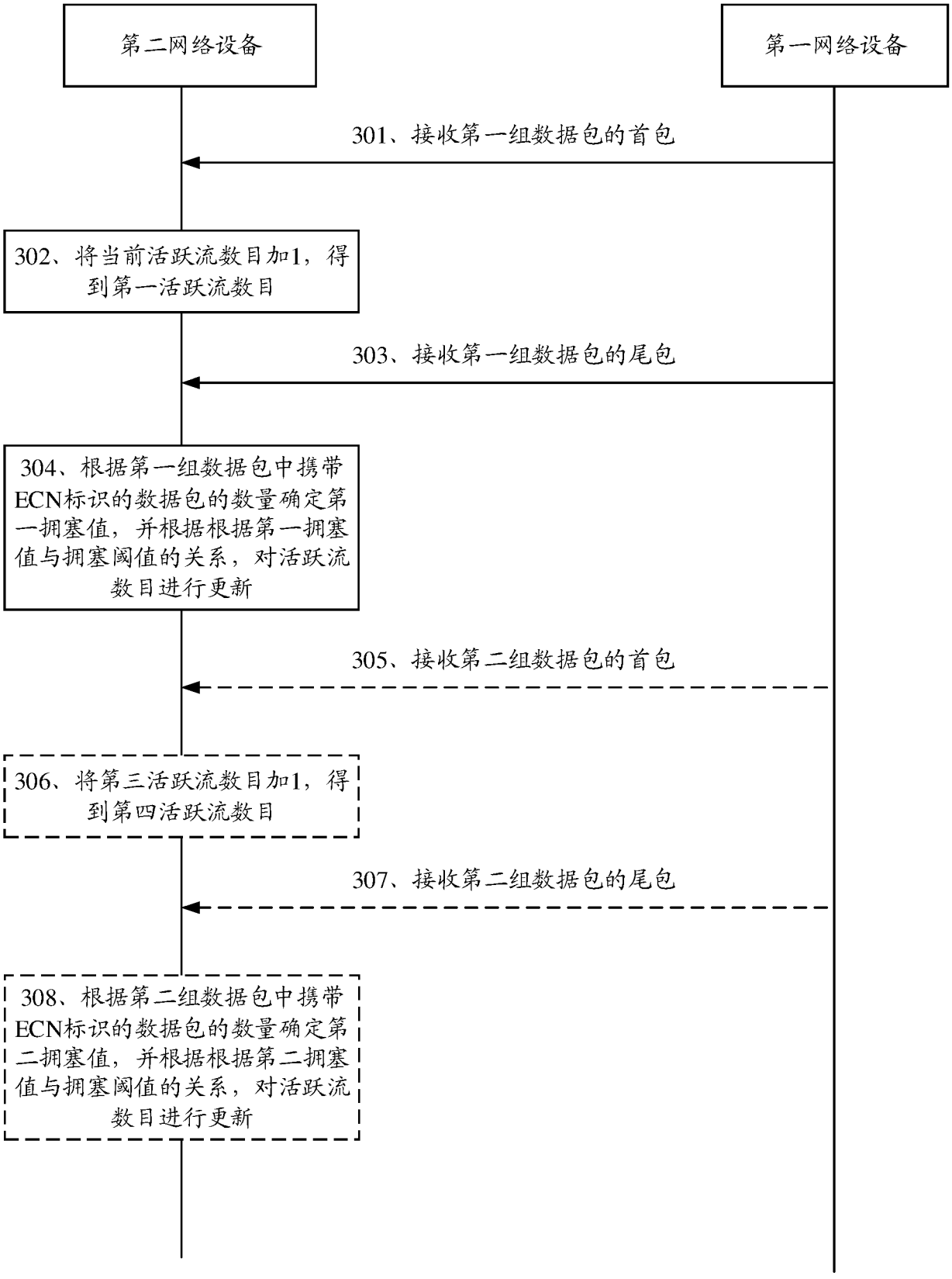

第二网络设备
第一网络设备
301、接收第一组数据包的首包
302、将当前活跃流数目加1，得到第一活跃流数目
303、接收第一组数据包的尾包
304、根据第一组数据包中携带ECN标识的数据包的数量确定第一拥塞值，并根据根据第一拥塞值与拥塞阈值的关系，对活跃流数目进行更新
305、接收第二组数据包的首包
306、将第三活跃流数目加1，得到第四活跃流数目
307、接收第二组数据包的尾包
308、根据第二组数据包中携带ECN标识的数据包的数量确定第二拥塞值，并根据根据第二拥塞值与拥塞阈值的关系，对活跃流数目进行更新

图4

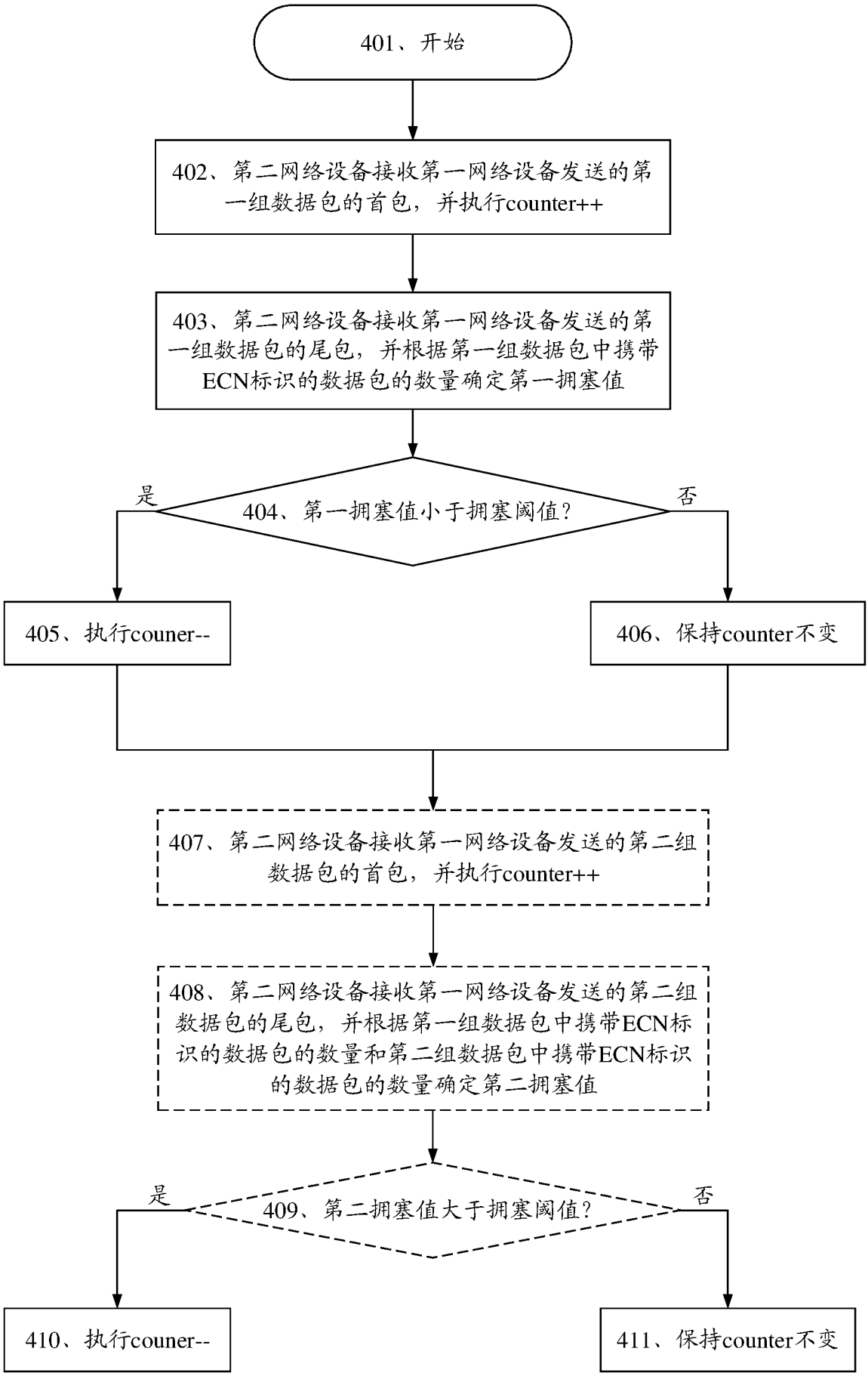

401、开始
402、第二网络设备接收第一网络设备发送的第一组数据包的首包，并执行counter++
403、第二网络设备接收第一网络设备发送的第一组数据包的尾包，并根据第一组数据包中携带ECN标识的数据包的数量确定第一拥塞值
404、第一拥塞值小于拥塞阈值？
是
否
405、执行couner--
406、保持counter不变
407、第二网络设备接收第一网络设备发送的第二组数据包的首包，并执行counter++
408、第二网络设备接收第一网络设备发送的第二组数据包的尾包，并根据第一组数据包中携带ECN标识的数据包的数量和第二组数据包中携带ECN标识的数据包的数量确定第二拥塞值
409、第二拥塞值大于拥塞阈值？
是
否
410、执行couner--
411、保持counter不变

图5

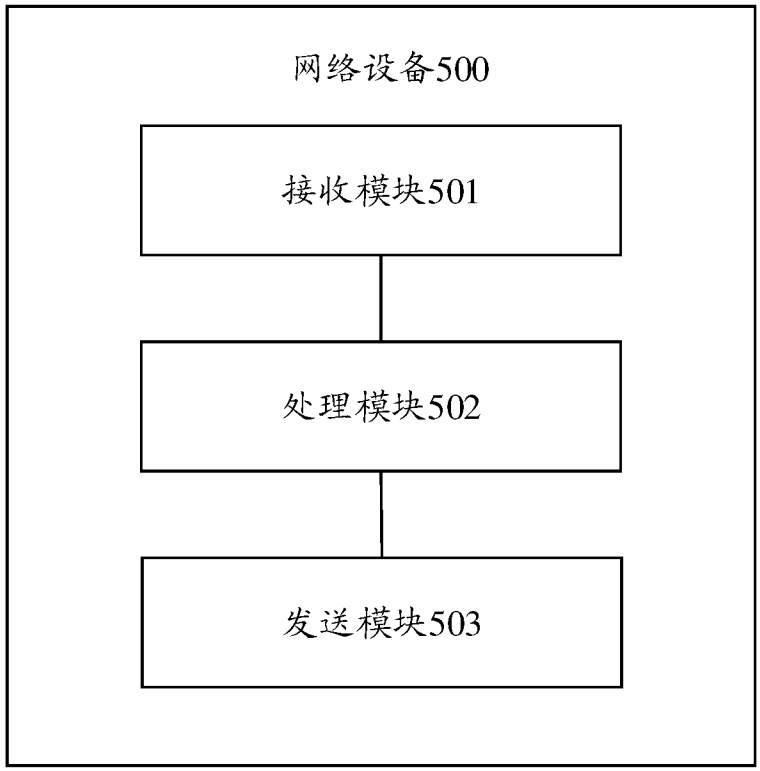

网络设备500
接收模块501
处理模块502
发送模块503

图6

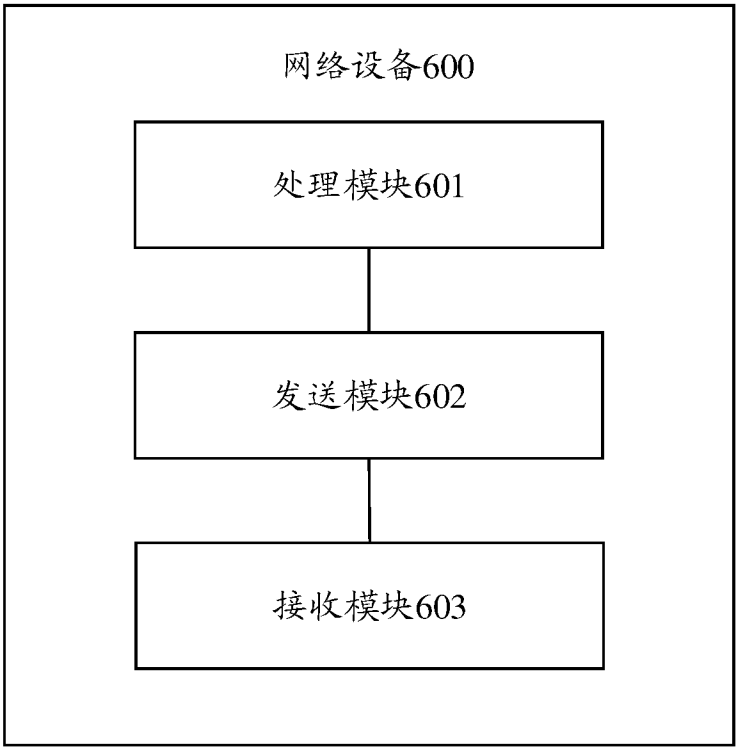

网络设备600
处理模块601
发送模块602
接收模块603

图7

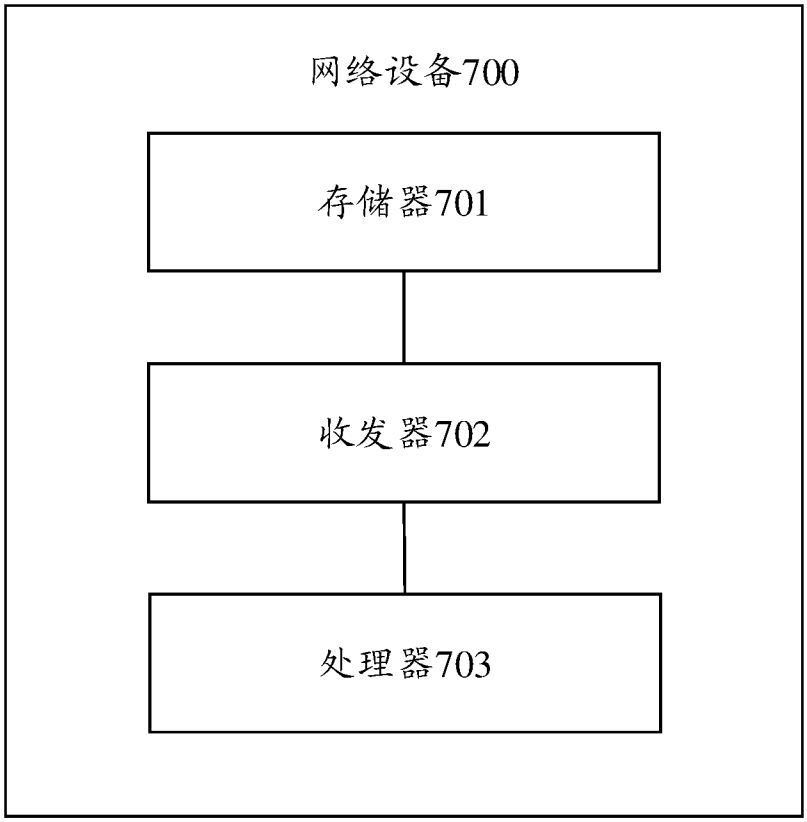


图8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/094171

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

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: 拥塞, 拥堵, 堵塞, 瓶颈, 控制, 数据流, 业务流, 活跃流, 活跃, 带宽, 额定, 目标, 速率, 调节, 调整, 吞吐, 最大, 满, congestion, jamming, bottleneck link, control, active flow, active, bandwidth, rated, rate, adjust

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109067665 A (HUAWEI TECHNOLOGIES CO., LTD.) 21 December 2018 (2018-12-21) claims 1-24	1-24
X	CN 105531968 A (CISCO TECHNOLOGY, INC.) 27 April 2016 (2016-04-27) description, paragraphs [0015]-[0038] and [0076]	1-24
A	CN 103916329 A (INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 09 July 2014 (2014-07-09) entire document	1-24
A	CN 108494698 A (TIANJIN TIANDI WEIYE ROBOT TECHNOLOGY CO., LTD.) 04 September 2018 (2018-09-04) entire document	1-24
A	CN 102404187 A (HUAWEI TECHNOLOGIES CO., LTD.) 04 April 2012 (2012-04-04) entire document	1-24
A	US 2017149665 A1 (NEC EUROPE LTD.) 25 May 2017 (2017-05-25) entire document	1-24

☐ 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
30 August 2019	**11 September 2019**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/094171

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 109067665 A	21 December 2018	None	
CN 105531968 A	27 April 2016	EP 3044918 A1	20 July 2016
		US 9521177 B2	13 December 2016
		CN 105531968 B	10 May 2019
		EP 3044918 B1	18 October 2017
		WO 2015038277 A1	19 March 2015
		US 2015074285 A1	12 March 2015
CN 103916329 A	09 July 2014	CN 103916329 B	12 April 2017
CN 108494698 A	04 September 2018	None	
CN 102404187 A	04 April 2012	WO 2011144141 A1	24 November 2011
US 2017149665 A1	25 May 2017	EP 3146682 A1	29 March 2017
		US 10135735 B2	20 November 2018
		WO 2016008519 A1	21 January 2016

国际检索报告

国际申请号
PCT/CN2019/094171

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

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: 拥塞，拥堵，堵塞，瓶颈，控制，数据流，业务流，活跃流，活跃，带宽，额定，目标，速率，调节，调整，吞吐，最大，满，congestion，jamming，bottleneck link，control，active flow，active，bandwidth，rated，rate，adjust

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
PX	CN 109067665 A (华为技术有限公司) 2018年 12月 21日 (2018-12-21) 权利要求1-24	1-24
X	CN 105531968 A (思科技术公司) 2016年 4月 27日 (2016-04-27) 说明书第[0015]-[0038]、[0076]段	1-24
A	CN 103916329 A (中国科学院计算技术研究所) 2014年 7月 9日 (2014-07-09) 全文	1-24
A	CN 108494698 A (天津天地伟业机器人技术有限公司) 2018年 9月 4日 (2018-09-04) 全文	1-24
A	CN 102404187 A (华为技术有限公司) 2012年 4月 4日 (2012-04-04) 全文	1-24
A	US 2017149665 A1 (NEC EUROPE LTD) 2017年 5月 25日 (2017-05-25) 全文	1-24

☐ 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
2019年 8月 30日	2019年 9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	李福涛
传真号 (86-10)62019451	电话号码 86-(20)-28950443

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/094171

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	109067665	A	2018年 12月 21日	无			
CN	105531968	A	2016年 4月 27日	EP	3044918	A1	2016年 7月 20日
				US	9521177	B2	2016年 12月 13日
				CN	105531968	B	2019年 5月 10日
				EP	3044918	B1	2017年 10月 18日
				WO	2015038277	A1	2015年 3月 19日
				US	2015074285	A1	2015年 3月 12日
CN	103916329	A	2014年 7月 9日	CN	103916329	B	2017年 4月 12日
CN	108494698	A	2018年 9月 4日	无			
CN	102404187	A	2012年 4月 4日	WO	2011144141	A1	2011年 11月 24日
US	2017149665	A1	2017年 5月 25日	EP	3146682	A1	2017年 3月 29日
				US	10135735	B2	2018年 11月 20日
				WO	2016008519	A1	2016年 1月 21日