

(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 12 月 10 日 (10.12.2020)



(10) 国际公布号
WO 2020/244448 A1

- (51) 国际专利分类号:
H04L 12/825 (2013.01) **H04L 1/16** (2006.01)
- (21) 国际申请号: PCT/CN2020/093138
- (22) 国际申请日: 2020 年 5 月 29 日 (29.05.2020)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201910487772.5 2019年6月5日 (05.06.2019) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (**ALIBABA GROUP HOLDING LIMITED**) [—/CN]; 开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。
- (72) 发明人: 缪睿 (**MIAO, Rui**); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。刘洪强 (**LIU, Hongqiang**); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

张铭 (**ZHANG, Ming**); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。李宇亮 (**LI, Yuliang**); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京三友知识产权代理有限公司 (**BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.**); 中国北京市金融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (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,

(54) **Title:** CONGESTION CONTROL METHOD AND DEVICE, AND COMPUTER-READABLE MEDIUM

(54) 发明名称: 一种拥塞控制方法、设备及计算机可读介质

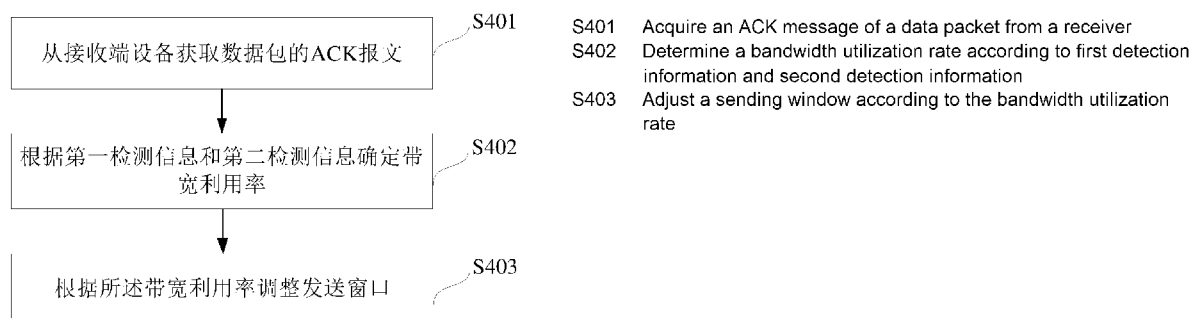


图 4

(57) **Abstract:** Provided are a congestion control method and device, and a computer-readable medium. In an embodiment of the method, a switch may insert detection information into a data packet when transmitting data, and a sender can acquire, from a receiver, an ACK message of the data packet, wherein the ACK message contains the detection information; and the sender can determine a bandwidth utilization rate by means of performing calculation according to first detection information in an ACK message corresponding to the next RTT interval and second detection information in an ACK message corresponding to the previous RTT interval, thereby adjusting a sending window according to the bandwidth utilization rate. Since, according to the solution, the bandwidth utilization rate can be calculated, when each RTT starts, by using the detection information that may be inserted into the data packet by the switch when transmitting data, performing closed-loop adjustment on the sending window according to the bandwidth utilization rate can adapt to different degrees of network congestion, and enables control to be more accurate, and throughout the whole process, no queue is relied on, and no increase in delay may be caused.

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: 本申请提供了一种拥塞控制方法、设备及计算机可读介质。所述方法的一个实施例中, 交换机在传输数据时会向数据包插入的检测信息, 而发送端设备可以从接收端设备获取数据包的ACK报文, 该ACK报文中所述检测信息, 发送端设备根据下一个RTT区间对应的ACK报文中的第一检测信息以及上一个RTT区间对应的ACK报文中的第二检测信息进行计算, 可以确定带宽利用率, 进而根据所述带宽利用率调整发送窗口。由于本方案是在每个RTT开始时都可以利用交换机在传输数据时会向数据包插入的检测信息计算带宽利用率, 据此对发送窗口进行闭环调整, 能够适应不同的网络拥塞程度, 使得控制更加准确, 并且在整个过程中不依赖队列, 不会造成延迟的增加。

一种拥塞控制方法、设备及计算机可读介质

本申请要求 2019 年 06 月 05 日递交的申请号为 201910487772.5、发明名称为“一种拥塞控制方法、设备及计算机可读介质”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。

5

技术领域

本申请涉及信息技术领域，尤其涉及一种拥塞控制方法、设备及计算机可读介质。

背景技术

10 RDMA (Remote Direct Memory Access, 远程直接内存访问) 技术能够提供高带宽、低时延和低 CPU 负载的特性, 在数据中心中的应用越来越广泛。RDMA 在以太网上的实现标准叫 RoCE (RDMA over Converged Ethernet), 由于以太网不能保证不丢包, 为了到达不丢包的目的, 引入了 PFC (Priority-Based Flow Control, 基于优先级流量控制) 技术。

15 然而, 虽然 PFC 可以保证不丢包, 但这个技术也存在相应的缺点, 比如不公平、拥塞扩散和网络死锁等问题, 所以在网络中需要尽量少的使用 PFC。在这种需求下, 出现了 DCQCN (Data Center Quantized Congestion Notification) 方案实现网络的拥塞控制。

DCQCN 算法中涉及的参与方如图 1 所示, 包括发送端设备 (Sender), 接收端设备 (Receiver) 和交换机 (Switch) 三个参与方。交换机负责转发从发送端设备到接收端设备的数据包。当交换机中的队列出现拥塞时, 交换机在数据包上打上 ECN (Explicit Congestion Notification) 标识。当接收端设备收到带有 ECN 标识的数据包后, 会相应地产生 CNP (Congestion Notification Packets) 报文告知发送端设备, 从而引导发送端设备调整发送速度。DCQCN 方案可以有效的在网络拥塞的早期控制发送端设备的吞吐量, 从而避免大量 PFC 的出现。

25 但是, DCQCN 算法具有深层次的设计缺陷, 这种方案是一种开环的、模糊的、依靠队列的控制机制。

1) 开环的控制是指, DCQCN 的升速和减速的调整的时间周期、幅度, 均是依据一组预设的固定参数。而这一组固定的参数无法适应于不同的网络拥塞场景以及网络流量的动态变化。

2) 模糊的控制是指, DCQCN 响应网络的拥塞状态, 依据的是网络反馈的单一比特的 ECN 信号, 因此算法对于网络拥塞的程度、变化, 都不能做及时且精确的响应。

3) 依靠队列的控制是指, DCQCN 算法需要在网络中维护一个较大的队列, 并且依靠队列变化来反馈一个模糊的网络拥塞的程度, 但是这就造成了网络端到端延迟的增加,

5 同 RDMA 提供低延迟网络的初衷是违背的。

此外, 还有 DCTCP (Data Center TCP)、Timely 以及如 BBR、CUBIC 等通用的拥塞控制算法。其中, DCTCP 算法同样需要利用 ECN 标识, 并且依靠队列变化来反馈一个模糊的网络拥塞的程度, 因此导致该算法对于网络拥塞的程度、变化, 都不能做及时且精确的响应, 并且会造成了网络端到端延迟的增加。

10 Timely 算法是利用网卡硬件测量网络的 RTT (Round-Trip Time) 值, 从而检测并控制拥塞的算法。该算法的核心思想是基于硬件的 RTT 测量从而排除掉软件层面影响, 因此条件下测量的 RTT 值仅仅包含网络传输的延迟。在获得 RTT 值之后, 可以根据 RTT 值的大小执行增速或者减速, 此外还可以根据 RTT 值的变化方向进行动态调整。该算法的不足在于: 首先, 需要依赖于网络拥塞造成网络排队的增加, 因此其同样需要在网络
15 中维护一个较大的队列, 并且依靠队列变化来反馈网络拥塞的程度, 但是这就造成了网络端到端延迟的增加; 其次, Timely 算法的拥塞调整手段拥塞收敛较慢; 最后, 依靠 RTT 调整拥塞会受反向拥塞的影响, 且有网络公平性的问题。

而 BBR、CUBIC 等通用的拥塞控制算法的主要问题在于拥塞的检测仅仅是基于丢包, 而往往这种检测存在滞后性, 若网络有大量的排队延迟, 在网络发生丢包减速前,
20 拥塞的影响已经造成。并且丢包重传也会造成延迟, 甚至超时 (timeout), 因此拥塞控制的整体效果不佳。

发明内容

本申请的一个目的是提供一种拥塞控制方案, 用以解决现有方案在实现拥塞控制时
25 准确性不佳, 且会造成延迟增加的问题。

本申请实施例提供了一种拥塞控制方法, 该方法包括:

发送端设备从接收端设备获取数据包的 ACK 报文, 其中, 所述 ACK 报文中包含交换机向数据包插入的检测信息;

所述发送端设备根据第一检测信息和第二检测信息确定带宽利用率, 其中, 所述第
30 一检测信息为下一个 RTT 区间对应的 ACK 报文中的检测信息, 所述第二检测信息为上

一个 RTT 区间对应的 ACK 报文中的检测信息；

所述发送端设备根据所述带宽利用率调整发送窗口。

本申请实施例还提供了一种拥塞控制方法，该方法包括：

交换机接收由发送端设备发送的数据包，其中，所述发送端设备发送数据包的发送
5 窗口由所述发送端设备根据带宽利用率调整获得，所述带宽利用率由所述发送端设备第一检测信息和第二检测信息确定；

所述交换机向所述数据包中插入检测信息，并将插入检测信息后的数据包发送至接收端设备，以使所述接收端设备在接收到所述数据包后向所述发送端设备返回所述数据包的 ACK 报文，其中，所述 ACK 报文中包含交换机向数据包插入的检测信息，所述第
10 一检测信息为下一个 RTT 区间对应的 ACK 报文中的检测信息，所述第二检测信息为上一个 RTT 区间对应的 ACK 报文中的检测信息。

本申请实施例还提供了一种实现拥塞控制的发送端设备，该发送端设备包括：

接收模块，用于从接收端设备获取数据包的 ACK 报文，其中，所述 ACK 报文中包含交换机向数据包插入的检测信息；

15 控制模块，用于根据第一检测信息和第二检测信息确定带宽利用率，其中，所述第一检测信息为下一个 RTT 区间对应的 ACK 报文中的检测信息，所述第二检测信息为上一个 RTT 区间对应的 ACK 报文中的检测信息，以及根据所述带宽利用率调整发送窗口；

发送模块，基于发送窗口发送向接收端设备发送数据包。

此外，本申请的一些实施例还提供了一种计算设备，该设备包括用于存储计算机程序指令的存储器和用于执行计算机程序指令的处理器，其中，当该计算机程序指令被该
20 处理器执行时，触发所述设备执行所述的拥塞控制方法。

本申请实施例还提供了一种实现拥塞控制的交换机，该交换机包括：

接收模块，用于接收由发送端设备发送的数据包，其中，所述发送端设备发送数据包的发送窗口由所述发送端设备根据带宽利用率调整获得，所述带宽利用率由所述发送
25 端设备第一检测信息和第二检测信息确定；

信息插入模块，用于向所述数据包中插入检测信息，

发送模块，用于将插入检测信息后的数据包发送至接收端设备，以使所述接收端设备在接收到所述数据包后向所述发送端设备返回所述数据包的 ACK 报文，其中，所述 ACK 报文中包含交换机向数据包插入的检测信息，所述第一检测信息为下一个 RTT 区
30 间对应的 ACK 报文中的检测信息，所述第二检测信息为上一个 RTT 区间对应的 ACK 报

文中的检测信息。

本申请的另一些实施例还提供了一种计算机可读介质，其上存储有计算机程序指令，所述计算机可读指令可被处理器执行以实现所述的拥塞控制方法。

本申请实施例提供方案中，交换机在传输数据时会向数据包插入的检测信息，而发送端设备可以从接收端设备获取数据包的 ACK 报文，该 ACK 报文中所述检测信息，发送端设备根据下一个 RTT 区间对应的 ACK 报文中的第一检测信息以及上一个 RTT 区间对应的 ACK 报文中的第二检测信息进行计算，可以确定带宽利用率，进而根据所述带宽利用率调整发送窗口。由于本方案是在每个 RTT 开始时都可以利用交换机在传输数据时会向数据包插入的检测信息计算带宽利用率，据此对发送窗口进行闭环调整，能够适应不同的网络拥塞程度，使得控制更加准确，并且在整个过程中不依赖队列，不会造成延迟的增加。

附图说明

通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本申请的其它特征、目的和优点将会变得更明显：

图 1 为拥塞控制方案中涉及的参与方的示意图；

图 2 为本申请实施例中一种拥塞控制方案所涉及的场景示意图；

图 3 为本申请实施例中交换机对数据包插入的 INT 信息的一种格式示意图；

图 4 为发送端设备在实现本申请实施例提供的拥塞控制方法时的处理流程图；

图 5 为本申请实施例中一种调整发送窗口的方式的处理逻辑流程图；

图 6 为本申请实施例提供的一种发送端设备的结构示意图；

图 7 为本申请实施例提供的一种计算设备的结构示意图；

附图中相同或相似的附图标记代表相同或相似的部件。

具体实施方式

下面结合附图对本申请作进一步详细描述。

在本申请一个典型的配置中，终端、服务网络的设备均包括一个或多个处理器 (CPU)、输入/输出接口、网络接口和内存。

内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器 (RAM) 和

/或非易失性内存等形式，如只读存储器（ROM）或闪存（flash RAM）。内存是计算机可读介质的示例。

计算机可读介质包括永久性和非永久性、可移动和非可移动媒体，可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的装置或其他数据。计算机的存储介质的例子包括，但不限于相变内存（PRAM）、静态随机存取存储器（SRAM）、动态随机存取存储器（DRAM）、其他类型的随机存取存储器（RAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、快闪记忆体或其他内存技术、只读光盘（CD-ROM）、数字多功能光盘（DVD）或其他光学存储、磁盒式磁带，磁带磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。

本申请实施例提供了一种拥塞控制方法，该方法在每个 RTT 开始时都可以利用交换机在传输数据时会向数据包插入的检测信息计算带宽利用率，据此对发送窗口进行闭环调整，能够适应不同的网络拥塞程度，使得控制更加准确，并且在整个过程中不依赖队列，不会造成延迟的增加。

在实际场景中，拥塞控制的过程中涉及如图 1 所示的三个参与方，包括发送端设备（Sender），接收端设备（Receiver）和交换机（Switch）。其中，发送端设备用于根据本申请实施例中提供的拥塞控制方法来调整发送窗口，以对链路上的拥塞情况进行控制。

交换机用于转发发送端设备发出的数据包，并在转发过程中逐个或者抽样在数据包中插入检测信息，即可以对每个转发的数据包都插入检测信息，或者也可以在每 N 个数据包选择一个数据包插入检测信息。此外，当链路中存在多个交换机时，每个交换机都可以在数据包中插入检测信息。例如在图 2 所示的场景中，包含两个交换机 Link-1 和 Link-2，其插入的检测信息为 INT（In-band Network Telemetry）信息，当发送端设备 Sender 发送数据包 pkt 后，首先由交换机 Link-1 转发该数据包 pkt，此时交换机 Link-1 会在其中插入 INT 信息 INT1，然后再由交换机 Link-2 转发该数据包 pkt，此时交换机 Link-2 会在其中插入 INT 信息 INT2，再将数据包 pkt 发送给接收端设备 Receiver。

图 3 示出了本申请实施例中交换机对数据包插入的 INT 信息的一种格式，其中，INT 信息包括以下几部分：nHop 用于表示交换机序号，每个交换机加 1，pathID 是

逐跳交换机 ID 进行 XOR（异或）运算后的结果，1st Hop、2nd Hop 分别表示各个交换机插入的 B、ts、txBytes 和 qlen，其中，B 是端口速率，ts 是插入 INT 信息时的 timestamp（时间戳），即插入时钟，txBytes 是发送队列发送数据计数，即累积的端口发送字节数，qlen 是发送队列长度。在实际场景中，根据设备的具体情况，还可以增加其它信息，例如 INT 包识别头、INT 信息类别等等。

接收端设备 Receiver 接收带有检测信息的数据包，会产生关于该数据包的 ACK 报文，并将该 ACK 反馈至发送端设备。该 ACK 报文中会携带交换机向数据包中插入的检测信息，若有多个交换机向数据包中插入了多个检测信息，ACK 报文中会携带所有的检测信息。以图 2 所示的场景为例，接收端设备返回的 ACK 报文中，会同时携带 Link-1 和 Link-2 插入的 INT 信息 INT1 和 INT2。

图 4 示出了发送端设备在实现本申请实施例提供的拥塞控制方法时的处理流程，包括以下处理步骤：

步骤 S401，从接收端设备获取数据包的 ACK 报文。其中，所述 ACK 报文中包含交换机向数据包插入的检测信息，所述检测信息是能够用于计算确定链路的带宽利用率的信息，例如在本申请的一些实施例中，所述检测信息可以是至少包括了发送队列长度（qlen）、发送队列发送数据计数（txBytes）、插入时钟（ts）和端口速率（B）的 INT 信息。

发送端设备发出的数据包经由交换机转发至接收端设备，交换机在转发时可以向数据包中插入相应的检测信息，即在数据包的交互过程中，交换机执行以下的处理以实现拥塞控制：交换机接收由发送端设备发送的数据包，向所述数据包中插入检测信息，并将插入检测信息后的数据包发送至接收端设备。接收端设备在获取到包含检测信息的数据包之后，会产生并向发送端设备返回关于该数据包的 ACK 报文，该 ACK 报文中会携带数据包中的检测信息，以便于接收端设备能够从收到的 ACK 报文获得检测信息，从而以此为依据实现拥塞控制。为了确保拥塞控制的实时性，接收端设备对于携带检测信息的数据包，可以优先产生并返回 ACK 报文，以使发送端设备能够尽快获得检测信息，以达到迅速响应的目的。

步骤 S402，根据第一检测信息和第二检测信息确定带宽利用率。其中，所述第一检测信息为下一个 RTT 区间对应的 ACK 报文中的检测信息，所述第二检测信息为上一个 RTT 区间对应的 ACK 报文中的检测信息。所述 RTT 是指从发送端设备发送一

个数据包到达接收端设备，并收到接收端设备返回 ACK 报文的时间，RTT 区间则表示该一个 RTT 所对应的时间区间。

RTT 区间的区分可以采用如下的方式：首先，在发送数据包时记录下一个 RTT 区间对应数据包的序列号 `snd_nxt`，而每个数据包对应的 ACK 报文返回时携带的序列号是接收端设备收到的数据包序列号加 1。当发送端设备收到 ACK 报文时，可以根据 ACK 报文的序列号判断是否经历了一个 RTT 区间。

例如，发送端设备在准备发送序列号为 N 的数据包时，可以记录下一个 RTT 区间（即该序列号为 N 的数据包的发送过程所对应的 RTT 区间）对应的数据包的序列号 N，接收端设备收到序列号为 N 的数据包，会产生序列号为 N+1 的 ACK 报文，表示接收端设备下一次期待收到序列号为 N+1 的数据包。由此，发送端设备会收到序列号为 N+1 的 ACK 报文，通过比较收到 ACK 报文的序列号 N+1 以及发出的数据包的序列号 N，可以确定是否经历了一个 RTT 区间。以本实施例中的情况为例，由于 $N+1 > N$ ，则表示收到的 ACK 报文时，已经经历了一个 RTT 区间。

在本申请的一些实施例中，可以在从接收端设备获取数据包的 ACK 报文之后，根据 ACK 报文的序列号判断接收到的 ACK 报文是否为下一个 RTT 区间对应的 ACK 报文。若判断结果为下一个 RTT 区间对应的 ACK 报文，则执行后续步骤更新发送窗口，否则可以不做调整或者仅在原发送窗口的基础上进行小幅度的调整。由此，可以在每个 RTT 区间更新一次发送窗口。

在计算带宽利用率时，可以先基于根据第一检测信息和第二检测信息，计算链路在下一个 RTT 区间中需要处理的数据量和链路的带宽延迟积（BDP，Bandwidth-Delay Product），然后根据链路在下一个 RTT 区间中需要处理的数据量以及链路的带宽延迟，可以计算带宽利用率。其中，下一个 RTT 区间中需要处理的数据量表示了链路在当前的 RTT 区间中还需要处理的数据总量，即 in-flight 包的数据量，而链路的带宽延迟积则表示了链路在一个 RTT 区间中实际可以处理的数据量。由此，利用 in-flight 包的数据量除以 BDP，就可以获得整个链路的带宽利用率。

例如，在本申请实施例中可以将第一检测信息记作 `ack.L`，将第二检测信息记作 `L`。当再次经历一个 RTT 区间之后，会收到新的 `ack.L`，而原来的 `ack.L` 将作为新的 `L`。`ack.L` 和 `L` 中都可以包含 `qlen`、`txBytes`、`ts`、`B` 四种信息，以用于计算带宽利用率。对于包含多个交换机的场景，对于各个交换机插入的检测信息，可

以记为 $\text{ack.L}[i]$ 、 $L[i]$ ， i 表示不同的交换机，例如前述 Link-1 插入的 INT 信息即为 $\text{ack.L}[1]$ 和 $L[1]$ ，Link-2 插入的 INT 信息即为 $\text{ack.L}[2]$ 和 $L[2]$ 。

对于各个交换机所插入的检测信息，可以采用如下的方式计算下一个 RTT 区间中需要处理的数据量以及链路的带宽延迟积：可以根据第一检测信息中的发送队列发送数据计数和第二检测信息中的发送队列发送数据计数之差，以及第一检测信息中的插入时钟和第二检测信息中的插入时钟之差，计算发送速率 txRate ，即：

$$\text{txRate} = (\text{ack.L}[i].\text{txBytes} - L[i].\text{txBytes}) / (\text{ack.L}[i].\text{ts} - L[i].\text{ts})$$

而后，可以根据所述发送速率、基线 RTT 值和第一检测信息中的发送队列长度，计算链路在下一个 RTT 区间中需要处理的数据量，即为 $\text{ack.L}[i].\text{qlen} + \text{txRate} \times T$ ，其中， T 为基线 RTT 值。此外，可以根据第一检测信息中的端口速率和基线 RTT 值，计算链路的带宽延迟积 $\text{ack.L}[i].B \times T$ 。由此可以计算出带宽利用率为：

$$\begin{aligned} u &= (\text{ack.L}[i].\text{qlen} + \text{txRate} \times T) / \text{ack.L}[i].B \times T \\ &= \text{ack.L}[i].\text{qlen} / (\text{ack.L}[i].B \times T) + \text{txRate} / \text{ack.L}[i].B \end{aligned}$$

由于此时计算获得带宽利用率 u 是基于两个 RTT 区间（即前一个 RTT 区间和下一个 RTT 区间）对应的检测信息计算获得带宽利用率，该带宽利用率表示了短时间内的带宽利用情况，其结果有可能会收到短时间内的网络环境变化影响。因此，可以基于计算获得的该带宽利用率进行滑动平均处理，以获得能够更加有效地表示实际带宽利用情况的平均带宽利用率。

由此，本申请的一些实施例中，在根据链路在下一个 RTT 区间中需要处理的数据量以及链路的带宽延迟积计算带宽利用率时，可以先根据链路在下一个 RTT 区间中需要处理的数据量以及链路的带宽延迟积，计算第一瞬时带宽利用率，该第一瞬时带宽利用率为当前计算获得的瞬时带宽利用率 u 。然后可以根据第一瞬时带宽利用率和第二平均带宽利用率进行滑动平均处理，计算第一平均带宽利用率 U 。其中，所述第一平均带宽利用率为当前计算获得的平均带宽利用率，所述第二平均带宽利用率为上一次计算获得的平均带宽利用率。在本申请的一些实施例中，可以采用如下的公式计算第一平均带宽利用率

$$U = (1 - \tau/T) \times U' + \tau/T \times u$$

其中， U' 为第二平均带宽利用率， T 为基线 RTT 值， τ 为 $(\text{ack.L}[i].\text{ts} - L[i].\text{ts})$ 与 T 中的较小值，即 $\tau = \min(\text{ack.L}[i].\text{ts} - L[i].\text{ts}, T)$ ， u 为第一瞬时带宽利用率。

在此，本领域技术人员应当理解，上述计算第一平均带宽利用率的方式仅为举例，现有或今后出现的基于类似原理的其它方式如果能够适用于本申请，也应该包含在本申请的保护范围内，并以引用的形式包含于此。

在实际场景中，对于第一个 RTT 区间，由于不存在上一次计算获得的平均带宽
5 利用率，因此，该次计算时的第二平均带宽利用率可以根据实际发送窗口的情况进行确定，例如，若发送端设备对于一个新的数据流都采用满线速发送整个数据，此时在第一个 RTT 区间计算第一平均带宽利用率时的第二平均带宽利用率可以设定为 1。

在本申请的一些实施例中，可以在根据第一瞬时带宽利用率和第二平均带宽
10 利用率进行滑动平均处理，计算第一平均带宽利用率之前，判断所述第一瞬时带宽利用率是否大于第二瞬时带宽利用率，并仅在判断结果为是时执行后续处理。其中，所述第二瞬时带宽利用率为上一次计算获得的瞬时带宽利用率。由此，可以仅在比上一个 RTT 区间的带宽利用率更高时，更新第一平均带宽利用率。

步骤 S403，根据所述带宽利用率调整发送窗口。其中，调整发送窗口的原则如
15 下：若带宽利用率较高，则可以适当减速，即减小发送窗口；若带宽利用率较低，则可以适当加速，即增大发送窗口。

在本申请的一些实施例中，可以采用如下的方式调整发送窗口：当所述带宽
20 利用率大于等于预设的利用率阈值时，乘性地减小发送窗口。其中，乘性地调整即为将基线值与某一系数相乘，以实现快速的增大或减小。例如，本实施例中可以设置预设的利用率阈值为 U_{target} ，计算获得的带宽利用率为前述的第一平均带宽利用率 U ，当 $U \geq U_{\text{target}}$ 时，可以采用如下的公式计算发送窗口 W ：

$$R = R^c / (U / U_{\text{target}})$$

$$W = R \times T$$

其中， R^c 为每次调整的速率基线值，即上一次调整后的 R 或者第一次调整时的
25 初始值，乘性的调整系数为带宽利用率与利用率阈值之比。在本申请的另一些实施中，还可以采用如下方式调整发送窗口：

$$R = R^c / (U / U_{\text{target}}) + \text{aiRate}$$

$$W = R \times T$$

其中，aiRate 为预先设置的数量级小于 R 的预设调整值，例如在万兆以太网场

景下，R 可以是 100Gbps，而 aiRate 可以是 100Mbps 以内。

而当所述带宽利用率小于预设的利用率阈值时，可以加性地增大发送窗口。其中，加性地调整即为将基线值与某一系数相加，以实现慢速的增大或减小。例如，本申请实施例有加性的调整系数可以是预设调整值 aiRate，当 $U < U_{\text{target}}$ 时，可以采

5 用如下的公式计算发送窗口 W：

$$R = R^c + \text{aiRate}$$

$$W = R \times T$$

由此，本申请实施例提供的调整发送窗口的方式服从 AIMD (Additive-Increase Multiplicative-Decrease) 准则，从而能够体现公平性。

10 此外，当连续加性地增大发送窗口的次数大于等于预设次数时，此时一般表示链路还有大量带宽未被利用，因此可以快速增大发送窗口，即乘性地增大发送窗口。

在实际场景中，可以设置一个参数 incStage，在每次加性地增大发送窗口后，该参数 incStage 自增 1，在每次乘性地减小发送窗口后，该参数 incStage 置 0，由此可以在通过比较该参数 incStage 与预设次数 maxStage，来确定是否需要乘性地增
15 大发送窗口。图 5 示出了本申请实施例中一种调整发送窗口的方式的处理逻辑。

由于本申请实施例的方案中，带宽利用率会在经过每个 RTT 区间时进行计算，因此调整发送窗口也会在每个 RTT 区间开始时进行，由此使得拥塞控制能够适应网络拥塞的程度和不同流量的影响。并且，由于发送窗口的调整是依赖于网络的拥塞情况，即发送队列长度、发送队列发送数据计数、时钟等，相对于单依靠单一的 ECN
20 信息，本方案能够更加精准的确定网络的拥塞程度，而且使得收敛更加快速，一般可以在一个 RTT 实现收敛。此外，由于本方案并不依赖队列的反馈，而是基于交换机插入的检测信息直接计算带宽利用率进而调整发送窗口，因此不会增加网络延迟。

由于实际场景中可能存在一些特殊情况，例如部分传输协议无法支持每个数据包都产生 ACK 报文，或者部分传输协议会在收到多个数据包才产生一个 ACK 报文，
25 甚至不需要产生 ACK 报文（如 RDMA READ），或者在双向数据传输时 ACK 报文会与数据包共享一个消息队列造成 INT 信息的延迟等。在这些场景下，可以使用专用于携带检测信息的探测数据包（probing 包）来获取检测信息，probing 包可以用于数据包相同的五元组，或者特殊的五元组，但都需要使其能够测量链路的拥塞情况。

基于同一发明构思，本申请实施例中还提供了一种实现拥塞控制的发送端设备，

所述发送端设备对应的方法是前述实施例中拥塞控制方法，并且其解决问题的原理与该方法相似。

本申请实施例提供的一种发送端设备在实现拥塞控制时，在每个 RTT 开始时都可以利用交换机在传输数据时会向数据包插入的检测信息计算带宽利用率，据此对发送窗口进行闭环调整，能够适应不同的网络拥塞程度，使得控制更加准确，并且在整个过程中不依赖队列，不会造成延迟的增加。

在实际场景中，拥塞控制的过程中涉及如图 1 所示的三个参与方，包括发送端设备（Sender），接收端设备（Receiver）和交换机（Switch）。其中，发送端设备用于根据本申请实施例中提供的拥塞控制方法来调整发送窗口，以对链路上的拥塞情况进行控制。

交换机用于转发发送端设备发出的数据包，并在转发过程中逐个或者抽样在数据包中插入检测信息，即可以对每个转发的数据包都插入检测信息，或者也可以在每 N 个数据包选择一个数据包插入检测信息。此外，当链路中存在多个交换机时，每个交换机都可以在数据包中插入检测信息。例如在图 2 所示的场景中，包含两个交换机 Link-1 和 Link-2，其插入的检测信息为 INT（In-band Network Telemetry）信息，当发送端设备 Sender 发送数据包 pkt 后，首先由交换机 Link-1 转发该数据包 pkt，此时交换机 Link-1 会在其中插入 INT 信息 INT1，然后再由交换机 Link-2 转发该数据包 pkt，此时交换机 Link-2 会在其中插入 INT 信息 INT2，再将数据包 pkt 发送给接收端设备 Receiver。

图 3 示出了本申请实施例中交换机对数据包插入的 INT 信息的一种格式，其中，INT 信息包括以下几部分：nHop 用于表示交换机序号，每个交换机加 1，pathID 是逐跳交换机 ID 进行 XOR（异或）运算后的结果，1st Hop、2nd Hop 分别表示各个交换机插入的 B、ts、txBytes 和 qlen，其中，B 是端口速率，ts 是插入 INT 信息时的 timestamp（时间戳），即插入时钟，txBytes 是发送队列发送数据计数，即累积的端口发送字节数，qLen 是发送队列长度。在实际场景中，根据设备的具体情况，还可以增加其它信息，例如 INT 包识别头、INT 信息类别等等。

接收端设备 Receiver 接收带有检测信息的数据包，会产生关于该数据包的 ACK 报文，并将该 ACK 反馈至发送端设备。该 ACK 报文中会携带交换机向数据包中插入的检测信息，若有多个交换机向数据包中插入了多个检测信息，ACK 报文中会携带所

有的检测信息。以图 2 所示的场景为例，接收端设备返回的 ACK 报文中，会同时携带 Link-1 和 Link-2 插入的 INT 信息 INT1 和 INT2。

图 6 示出了本申请实施例提供的一种实现拥塞控制的发送端设备的结构，包括接收模块 610、控制模块 620 和发送模块 630。接收模块 610 用于从接收端设备获取数据包的 ACK 报文，控制模块 620 用于根据第一检测信息和第二检测信息确定带宽利用率，以及根据所述带宽利用率调整发送窗口。发送模块 630 基于发送窗口发送向接收端设备发送数据包。

所述 ACK 报文中包含交换机向数据包插入的检测信息，所述检测信息是能够用于计算确定链路的带宽利用率的信息，例如在本申请的一些实施例中，所述检测信息可以是至少包括了发送队列长度 (qlen)、发送队列发送数据计数 (txBytes)、插入时钟 (ts) 和端口速率 (B) 的 INT 信息。

发送端设备发出的数据包经由交换机转发至接收端设备，交换机在转发时可以向数据包中插入相应的检测信息。即在数据包的交互过程中，交换机执行以下的处理以实现拥塞控制：交换机的接收模块接收由发送端设备发送的数据包，交换机的信息插入模块向所述数据包中插入检测信息，交换机的发送模块将插入检测信息后的数据包发送至接收端设备。接收端设备在获取到包含检测信息的数据包之后，会产生并向发送端设备返回关于该数据包的 ACK 报文，该 ACK 报文中会携带数据包中的检测信息，以便于接收端设备能够从收到的 ACK 报文获得检测信息，从而以此为依据实现拥塞控制。为了确保拥塞控制的实时性，接收端设备对于携带检测信息的数据包，可以优先产生并返回 ACK 报文，以使发送端设备能够尽快获得检测信息，以达到迅速响应的目的。

所述第一检测信息为下一个 RTT 区间对应的 ACK 报文中的检测信息，所述第二检测信息为上一个 RTT 区间对应的 ACK 报文中的检测信息。所述 RTT 是指从发送端设备发送一个数据包到达接收端设备，并收到接收端设备返回 ACK 报文的时间，RTT 区间则表示该一个 RTT 所对应的时间区间。

RTT 区间的区分可以采用如下的方式：首先，在发送数据包时记录下一个 RTT 区间对应数据包的序列号 `snd_nxt`，而每个数据包对应的 ACK 报文返回时携带的序列号是接收端设备收到的数据包序列号加 1。当发送端设备收到 ACK 报文时，可以根据 ACK 报文的序列号判断是否经历了一个 RTT 区间。

例如，发送端设备在准备发送序列号为 N 的数据包时，可以记录下一个 RTT 区间（即该序列号为 N 的数据包的发送过程所对应的 RTT 区间）对应的数据包的序列号 N ，接收端设备收到序列号为 N 的数据包，会产生序列号为 $N+1$ 的 ACK 报文，表示接收端设备下一次期待收到序列号为 $N+1$ 的数据包。由此，发送端设备会收到序列号为 $N+1$ 的 ACK 报文，通过比较收到 ACK 报文的序列号 $N+1$ 以及发出的数据包的序列号 N ，可以确定是否经历了一个 RTT 区间。以本实施例中的情况为例，由于 $N+1 > N$ ，则表示收到的 ACK 报文时，已经经历了一个 RTT 区间。

在本申请的一些实施例中，控制模块可以在从接收端设备获取数据包的 ACK 报文之后，根据 ACK 报文的序列号判断接收到的 ACK 报文是否为下一个 RTT 区间对应的 ACK 报文。若判断结果为下一个 RTT 区间对应的 ACK 报文，则执行后续步骤更新发送窗口，否则可以不做调整或者仅在原发送窗口的基础上进行小幅度的调整。由此，可以在每个 RTT 区间更新一次发送窗口。

在计算带宽利用率时，控制模块可以先基于根据第一检测信息和第二检测信息，计算链路在下一个 RTT 区间中需要处理的数据量和链路的带宽延迟积（BDP，Bandwidth-Delay Product），然后根据链路在下一个 RTT 区间中需要处理的数据量以及链路的带宽延迟，可以计算带宽利用率。其中，下一个 RTT 区间中需要处理的数据量表示了链路在当前的 RTT 区间中还需要处理的数据总量，即 in-flight 包的数据量，而链路的带宽延迟积则表示了链路在一个 RTT 区间中实际可以处理的数据量。由此，利用 in-flight 包的数据量除以 BDP，就可以获得整个链路的带宽利用率。

例如，在本申请实施例中可以将第一检测信息可以记作 $ack.L$ ，将第二检测信息可以记作 L 。当再次经历一个 RTT 区间之后，会收到新的 $ack.L$ ，而原来的 $ack.L$ 将作为新的 L 。 $ack.L$ 和 L 中都可以包含 $qlen$ 、 $txBytes$ 、 ts 、 B 四种信息，以用于计算带宽利用率。对于包含多个交换机的场景，对于各个交换机插入的检测信息，可以记为 $ack.L[i]$ 、 $L[i]$ ， i 表示不同的交换机，例如前述 Link-1 插入的 INT 信息即为 $ack.L[1]$ 和 $L[1]$ ，Link-2 插入的 INT 信息即为 $ack.L[2]$ 和 $L[2]$ 。

对于各个交换机所插入的检测信息，控制模块可以采用如下的方式计算下一个 RTT 区间中需要处理的数据量以及链路的带宽延迟积：可以根据第一检测信息中的发送队列发送数据计数和第二检测信息中的发送队列发送数据计数之差，以及第一检测信息中的插入时钟和第二检测信息中的插入时钟之差，计算发送速率 $txRate$ ，即：

$$txRate = (ack.L[i].txBytes - L[i].txBytes) / (ack.L[i].ts - L[i].ts)$$

而后，控制模块可以根据所述发送速率、基线 RTT 值和第一检测信息中的发送队列长度，计算链路在下一个 RTT 区间中需要处理的数据量，即为 $ack.L[i].qlen + txRate \times T$ ，其中，T 为基线 RTT 值。此外，可以根据第一检测信息中的端口速率和

5 基线 RTT 值，计算链路的带宽延迟积 $ack.L[i].B \times T$ 。由此可以计算出带宽利用率为：

$$u = (ack.L[i].qlen + txRate \times T) / ack.L[i].B \times T$$

$$= ack.L[i].qlen / (ack.L[i].B \times T) + txRate / ack.L[i].B$$

由于此时计算获得带宽利用率 u 是基于两个 RTT 区间（即前一个 RTT 区间和下一个 RTT 区间）对应的检测信息计算获得带宽利用率，该带宽利用率表示了短时间

10 内的带宽利用情况，其结果有可能会收到短时间内的网络环境变化影响。因此，可以基于计算获得的该带宽利用率进行滑动平均处理，以获得能够更加有效地表示实际带宽利用情况的平均带宽利用率。

由此，本申请的一些实施例中，在根据链路在下一个 RTT 区间中需要处理的数据量以及链路的带宽延迟积计算带宽利用率时，控制模块可以先根据链路在下一个

15 RTT 区间中需要处理的数据量以及链路的带宽延迟积，计算第一瞬时带宽利用率，该第一瞬时带宽利用率为当前计算获得的瞬时带宽利用率 u。然后可以根据第一瞬时带宽利用率和第二平均带宽利用率进行滑动平均处理，计算第一平均带宽利用率 U。其中，所述第一平均带宽利用率为当前计算获得的平均带宽利用率，所述第二平均带宽利用率为上一次计算获得的平均带宽利用率。在本申请的一些实施例中，可以采用

20 用如下的公式计算第一平均带宽利用率

$$U = (1 - \tau/T) \times U' + \tau/T \times u$$

其中，U' 为第二平均带宽利用率，T 为基线 RTT 值， τ 为 $(ack.L[i].ts - L[i].ts)$ 与 T 中的较小值，即 $\tau = \min(ack.L[i].ts - L[i].ts, T)$ ，u 为第一瞬时带宽利用率。在此，本领域技术人员应当理解，上述计算第一平均带宽利用率的方式仅为举例，

25 现有或今后出现的基于类似原理的其它方式如果能够适用于本申请，也应该包含在本申请的保护范围内，并以引用的形式包含于此。

在实际场景中，对于第一个 RTT 区间，由于不存在上一次计算获得的平均带宽利用率，因此，该次计算时的第二平均带宽利用率可以根据实际发送窗口的情况进行确定，例如，若发送端设备对于一个新的数据流都采用满线速发送整个数据，此

时在第一个 RTT 区间计算第一平均带宽利用率时的第二平均带宽利用率可以设定为 1。

在本申请的一些实施例中，控制模块可以在根据第一瞬时带宽利用率和第二平均带宽利用率进行滑动平均处理，计算第一平均带宽利用率之前，判断所述第一瞬时带宽利用率是否大于第二瞬时带宽利用率，并仅在判断结果为是时执行后续处理。其中，所述第二瞬时带宽利用率为上一次计算获得的瞬时带宽利用率。由此，可以仅在比上一个 RTT 区间的带宽利用率更高时，更新第一平均带宽利用率。

调整发送窗口的原则如下：若带宽利用率较高，则可以适当减速，即减小发送窗口；若带宽利用率较低，则可以适当加速，即增大发送窗口。

在本申请的一些实施例中，控制模块可以采用如下的方式调整发送窗口：当所述带宽利用率大于等于预设的利用率阈值时，乘性地减小发送窗口。其中，乘性地调整即为将基线值与某一系数相乘，以实现快速的增大或减小。例如，本实施例中可以设置预设的利用率阈值为 U_{target} ，计算获得的带宽利用率为前述的第一平均带宽利用率 U ，当 $U \geq U_{target}$ 时，可以采用如下的公式计算发送窗口 W ：

$$R = R^c / (U / U_{target})$$

$$W = R \times T$$

其中， R^c 为每次调整的速率基线值，即上一次调整后的 R 或者第一次调整时的初始值，乘性的调整系数为带宽利用率与利用率阈值之比。在本申请的另一些实施中，还可以采用如下方式调整发送窗口：

$$R = R^c / (U / U_{target}) + aiRate$$

$$W = R \times T$$

其中， $aiRate$ 为预先设置的数量级小于 R 的预设调整值，例如在万兆以太网场景下， R 可以是 100Gbps，而 $aiRate$ 可以是 100Mbps 以内。

而当所述带宽利用率小于预设的利用率阈值时，可以加性地增大发送窗口。其中，加性地调整即为将基线值与某一系数相加，以实现慢速的增大或减小。例如，本申请实施例中加性的调整系数可以是预设调整值 $aiRate$ ，当 $U < U_{target}$ 时，可以采用如下的公式计算发送窗口 W ：

$$R = R^c + aiRate$$

$$W = R \times T$$

由此，本申请实施例提供的调整发送窗口的方式服从 AIMD (Additive-Increase Multiplicative-Decrease) 准则，从而能够体现公平性。

此外，当连续加性地增大发送窗口的次数大于等于预设次数时，此时一般表示链路还有大量带宽未被利用，因此控制模块可以快速增大发送窗口，即乘性地增大发送窗口。

在实际场景中，可以设置一个参数 incStage，在每次加性地增大发送窗口后，该参数 incStage 自增 1，在每次乘性地减小发送窗口后，该参数 incStage 置 0，由此可以在通过比较该参数 incStage 与预设次数 maxStage，来确定是否需要乘性地增大发送窗口。图 5 示出了本申请实施例中一种调整发送窗口的方式的处理逻辑。

由于本申请实施例的方案中，带宽利用率会在经过每个 RTT 区间时进行计算，因此调整发送窗口也会在每个 RTT 区间开始时进行，由此使得拥塞控制能够适应网络拥塞的程度和不同流量的影响。并且，由于发送窗口的调整是依赖于网络的拥塞情况，即发送队列长度、发送队列发送数据计数、时钟等，相对于单依靠单一的 ECN 信息，本方案能够更加精准的确定网络的拥塞程度，而且使得收敛更加快速，一般可以在一个 RTT 实现收敛。此外，由于本方案并不依赖队列的反馈，而是基于交换机插入的检测信息直接计算带宽利用率进而调整发送窗口，因此不会增加网络延迟。

由于实际场景中可能存在一些特殊情况，例如部分传输协议无法支持每个数据包都产生 ACK 报文，或者部分传输协议会在收到多个数据包才产生一个 ACK 报文，甚至不需要产生 ACK 报文（如 RDMA READ），或者在双向数据传输时 ACK 报文会与数据包共享一个消息队列造成 INT 信息的延迟等。在这些场景下，可以使用专用于携带检测信息的探测数据包（probing 包）来获取检测信息，probing 包可以使用与数据包相同的五元组，或者特殊的五元组，但都需要使其能够测量链路的拥塞情况。

另外，本申请的一部分可被应用为计算机程序产品，例如计算机程序指令，当其被计算机执行时，通过该计算机的操作，可以调用或提供根据本申请的方法和/或技术方案。而调用本申请的方法的程序指令，可能被存储在固定的或可移动的记录介质中，和/或通过广播或其他信号承载媒体中的数据流而被传输，和/或被存储在根据程序指令运行的计算机设备的工作存储器中。在此，根据本申请的一些实施例包括一个如图 7 所示的计算设备，该设备包括存储有计算机可读指令的一个或多个存储器 710 和用于执行计算机可读指令的处理器 720，其中，当该计算机可读指令被

该处理器执行时，使得所述设备执行基于前述本申请的多个实施例的方法和/或技术方案。

此外，本申请的一些实施例还提供了一种计算机可读介质，其上存储有计算机程序指令，所述计算机可读指令可被处理器执行以实现前述本申请的多个实施例的方法和/或技术方案。

需要注意的是，本申请可在软件和/或软件与硬件的组合体中被实施，例如，可采用专用集成电路（ASIC）、通用目的计算机或任何其他类似硬件设备来实现。在一些实施例中，本申请的软件程序可以通过处理器执行以实现上文步骤或功能。同样地，本申请的软件程序（包括相关的数据结构）可以被存储到计算机可读记录介质中，例如，RAM 存储器，磁或光驱动器或软磁盘及类似设备。另外，本申请的一些步骤或功能可采用硬件来实现，例如，作为与处理器配合从而执行各个步骤或功能的电路。

对于本领域技术人员而言，显然本申请不限于上述示范性实施例的细节，而且在不背离本申请的精神或基本特征的情况下，能够以其他的具体形式实现本申请。

因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本申请的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本申请内。不应将权利要求中的任何附图标记视为限制所涉及的权利要求。此外，显然“包括”一词不排除其他单元或步骤，单数不排除复数。装置权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第一，第二等词语用来表示名称，而并不表示任何特定的顺序。

权 利 要 求 书

1. 一种拥塞控制方法，其中，该方法包括：

发送端设备从接收端设备获取数据包的 ACK 报文，其中，所述 ACK 报文中包含交换机向数据包插入的检测信息；

5 所述发送端设备根据第一检测信息和第二检测信息确定带宽利用率，其中，所述第一检测信息为下一个 RTT 区间对应的 ACK 报文中的检测信息，所述第二检测信息为上一个 RTT 区间对应的 ACK 报文中的检测信息；

所述发送端设备根据所述带宽利用率调整发送窗口。

2. 根据权利要求 1 所述的方法，其中，所述发送端设备根据第一检测信息和第二检测信息确定带宽利用率，包括：

所述发送端设备根据第一检测信息和第二检测信息，计算链路在下一个 RTT 区间中需要处理的数据量和链路的带宽延迟积；

所述发送端设备根据链路在下一个 RTT 区间中需要处理的数据量以及链路的带宽延迟积计算带宽利用率。

15 3. 根据权利要求 2 所述的方法，其中，所述检测信息包括发送队列长度、发送队列发送数据计数、插入时钟和端口速率；

所述发送端设备根据第一检测信息和第二检测信息，计算链路在下一个 RTT 区间中需要处理的数据量和链路的带宽延迟积，包括：

20 所述发送端设备根据第一检测信息中的发送队列发送数据计数和第二检测信息中的发送队列发送数据计数之差，以及第一检测信息中的插入时钟和第二检测信息中的插入时钟之差，计算发送速率；

所述发送端设备根据所述发送速率、基线 RTT 值和第一检测信息中的发送队列长度，计算链路在下一个 RTT 区间中需要处理的数据量；

25 所述发送端设备根据第一检测信息中的端口速率和基线 RTT 值，计算链路的带宽延迟积。

4. 根据权利要求 2 所述的方法，其中，所述发送端设备根据链路在下一个 RTT 区间中需要处理的数据量以及链路的带宽延迟积，计算带宽利用率，包括：

30 所述发送端设备根据链路在下一个 RTT 区间中需要处理的数据量以及链路的带宽延迟积，计算第一瞬时带宽利用率，其中，所述第一瞬时带宽利用率为当前计算获得的瞬时带宽利用率；

所述发送端设备根据第一瞬时带宽利用率和第二平均带宽利用率进行滑动平均处理，计算第一平均带宽利用率，其中，所述第一平均带宽利用率为当前计算获得的平均带宽利用率，所述第二平均带宽利用率为上一次计算获得的平均带宽利用率。

- 5 根据权利要求 4 所述的方法，其中，所述发送端设备在根据第一瞬时带宽利用率和第二平均带宽利用率进行滑动平均处理，计算第一平均带宽利用率之前，还包括：

所述发送端设备判断所述第一瞬时带宽利用率是否大于第二瞬时带宽利用率，并确定判断结果为是，其中，所述第二瞬时带宽利用率为上一次计算获得的瞬时带宽利用率。

6. 根据权利要求 1 所述的方法，其中，所述发送端设备在从接收端设备获取数据包的 ACK 报文之后，还包括：

- 10 所述发送端设备根据 ACK 报文的序列号判断接收到的 ACK 报文是否为下一个 RTT 区间对应的 ACK 报文。

7. 根据权利要求 1 所述的方法，其中，所述发送端设备根据所述带宽利用率调整发送窗口，包括：

- 15 当所述带宽利用率大于等于预设的利用率阈值时，所述发送端设备乘性地减小发送窗口；

当所述带宽利用率小于预设的利用率阈值时，所述发送端设备加性地增大发送窗口。

8. 根据权利要求 7 所述的方法，其中，该方法还包括：

- 20 当连续加性地增大发送窗口的次数大于等于预设次数时，所述发送端设备乘性地增大发送窗口。

9. 根据权利要求 7 或 8 所述的方法，其中，乘性的调整系数为带宽利用率与利用率阈值之比，加性的调整系数为预设调整值。

10. 根据权利要求 1 所述的方法，其中，所述数据包包括专用于携带检测信息的探测数据包。

- 25 11. 一种拥塞控制方法，其中，该方法包括：

交换机接收由发送端设备发送的数据包，其中，所述发送端设备发送数据包的发送窗口由所述发送端设备根据带宽利用率调整获得，所述带宽利用率由所述发送端设备第一检测信息和第二检测信息确定；

- 30 所述交换机向所述数据包中插入检测信息，并将插入检测信息后的数据包发送至接收端设备，以使所述接收端设备在接收到所述数据包后向所述发送端设备返回所述数据

包的 ACK 报文, 其中, 所述 ACK 报文中包含交换机向数据包插入的检测信息, 所述第一检测信息为下一个 RTT 区间对应的 ACK 报文中的检测信息, 所述第二检测信息为上一个 RTT 区间对应的 ACK 报文中的检测信息。

12. 一种实现拥塞控制的发送端设备, 其中, 该发送端设备包括:

5 接收模块, 用于从接收端设备获取数据包的 ACK 报文, 其中, 所述 ACK 报文中包含交换机向数据包插入的检测信息;

 控制模块, 用于根据第一检测信息和第二检测信息确定带宽利用率, 其中, 所述第一检测信息为下一个 RTT 区间对应的 ACK 报文中的检测信息, 所述第二检测信息为上一个 RTT 区间对应的 ACK 报文中的检测信息, 以及根据所述带宽利用率调整发送窗口;

10 发送模块, 基于发送窗口发送向接收端设备发送数据包。

 13. 根据权利要求 12 所述的发送端设备, 其中, 所述控制模块, 用于根据第一检测信息和第二检测信息, 计算链路在下一个 RTT 区间中需要处理的数据量和链路的带宽延迟积; 以及根据链路在下一个 RTT 区间中需要处理的数据量以及链路的带宽延迟积计算带宽利用率。

15 14. 根据权利要求 13 所述的发送端设备, 其中, 所述检测信息包括发送队列长度、发送队列发送数据计数、插入时钟和端口速率;

 所述控制模块, 用于根据第一检测信息中的发送队列发送数据计数和第二检测信息中的发送队列发送数据计数之差, 以及第一检测信息中的插入时钟和第二检测信息中的插入时钟之差, 计算发送速率; 根据所述发送速率、基线 RTT 值和第一检测信息中的
20 发送队列长度, 计算链路在下一个 RTT 区间中需要处理的数据量; 以及根据第一检测信息中的端口速率和基线 RTT 值, 计算链路的带宽延迟积。

 15. 根据权利要求 13 所述的发送端设备, 其中, 所述控制模块, 用于根据链路在下一个 RTT 区间中需要处理的数据量以及链路的带宽延迟积, 计算第一瞬时带宽利用率, 其中, 所述第一瞬时带宽利用率为当前计算获得的瞬时带宽利用率; 以及根据第一瞬时
25 带宽利用率和第二平均带宽利用率进行滑动平均处理, 计算第一平均带宽利用率, 其中, 所述第一平均带宽利用率为当前计算获得的平均带宽利用率, 所述第二平均带宽利用率为上一次计算获得的平均带宽利用率。

 16. 根据权利要求 15 所述的发送端设备, 其中, 所述控制模块, 还用于在根据第一瞬时带宽利用率和第二平均带宽利用率进行滑动平均处理, 计算第一平均带宽利用率之前, 判断所述第一瞬时带宽利用率是否大于第二瞬时带宽利用率, 并确定判断结果是为是,
30

其中，所述第二瞬时带宽利用率为上一次计算获得的瞬时带宽利用率。

17. 根据权利要求 12 所述的发送端设备，其中，所述控制模块，用于在所述带宽利用率大于等于预设的利用率阈值时，乘性地减小发送窗口；以及在所述带宽利用率小于预设的利用率阈值时，加性地增大发送窗口。

5 18. 根据权利要求 17 所述的发送端设备，其中，所述控制模块，还用于在连续加性地增大发送窗口的次数大于等于预设次数时，乘性地增大发送窗口。

19. 根据权利要求 17 或 18 所述的发送端设备，其中，乘性的调整系数为带宽利用率与利用率阈值之比，加性的调整系数为预设调整值。

20. 一种实现拥塞控制的交换机，其中，该交换机包括：

10 接收模块，用于接收由发送端设备发送的数据包，其中，所述发送端设备发送数据包的发送窗口由所述发送端设备根据带宽利用率调整获得，所述带宽利用率由所述发送端设备第一检测信息和第二检测信息确定；

信息插入模块，用于向所述数据包中插入检测信息，

15 发送模块，用于将插入检测信息后的数据包发送至接收端设备，以使所述接收端设备在接收到所述数据包后向所述发送端设备返回所述数据包的 ACK 报文，其中，所述 ACK 报文中包含交换机向数据包插入的检测信息，所述第一检测信息为下一个 RTT 区间对应的 ACK 报文中的检测信息，所述第二检测信息为上一个 RTT 区间对应的 ACK 报文中的检测信息。

20 21. 一种计算设备，其中，该设备包括用于存储计算机程序指令的存储器和用于执行计算机程序指令的处理器，其中，当该计算机程序指令被该处理器执行时，触发所述设备执行权利要求 1 至 11 中任一项所述的方法。

22. 一种计算机可读介质，其上存储有计算机程序指令，所述计算机可读指令可被处理器执行以实现如权利要求 1 至 11 中任一项所述的方法。

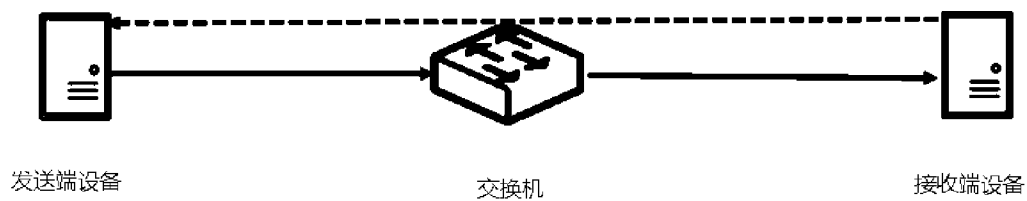


图 1

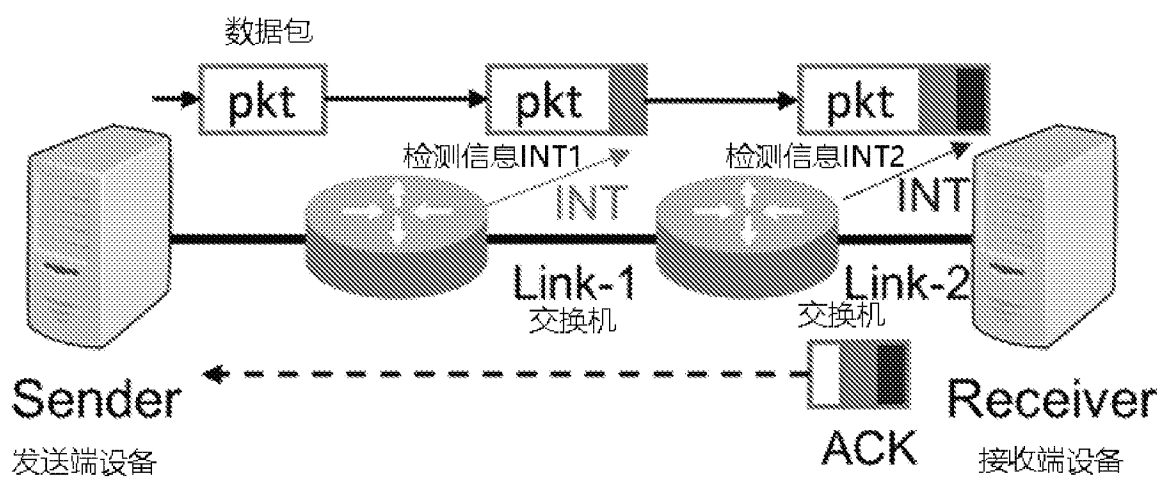


图 2

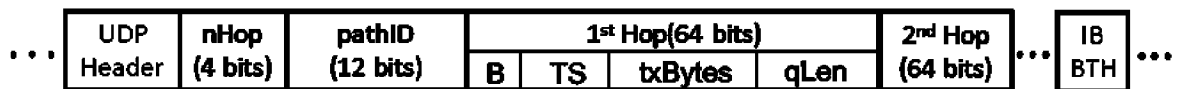


图 3

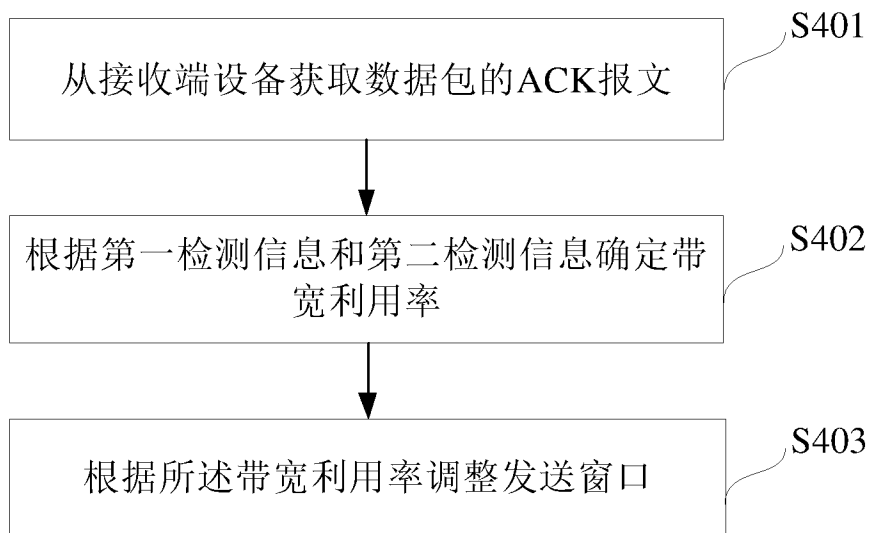


图 4

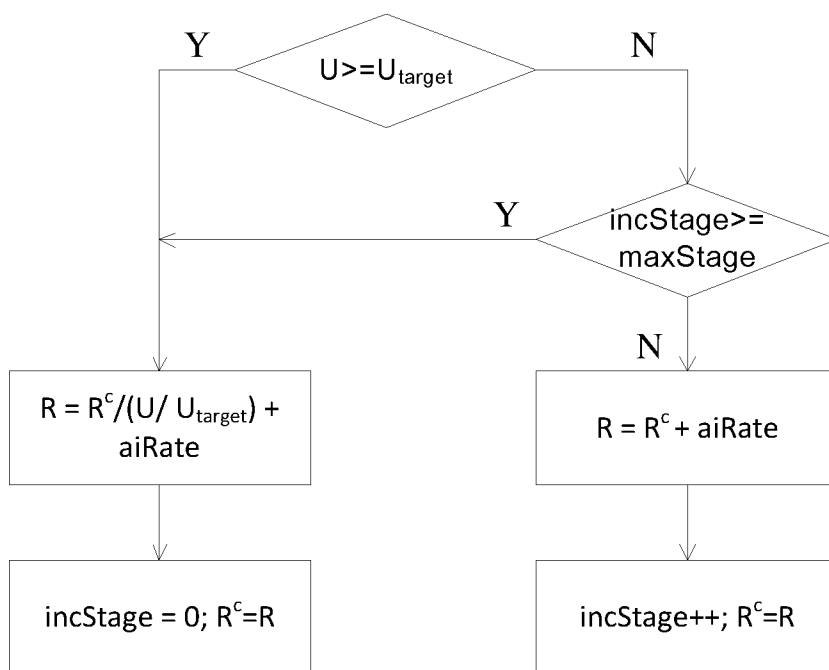


图 5

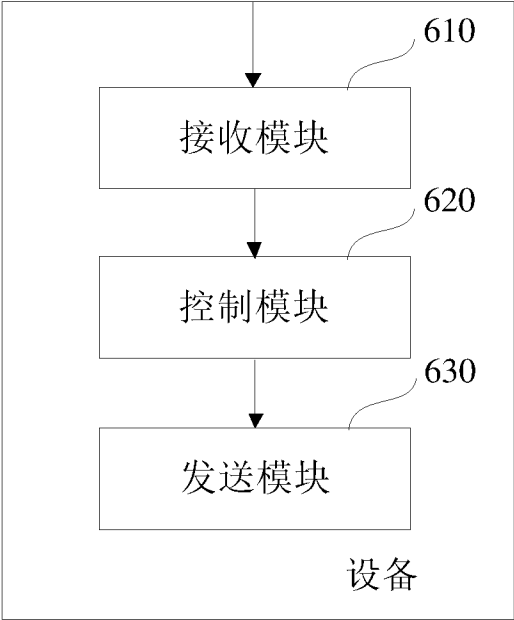


图 6

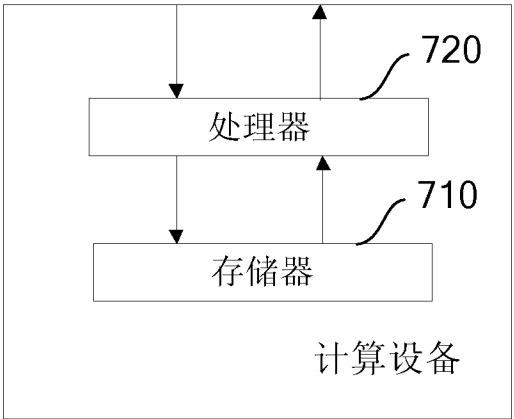


图 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/093138

A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/825(2013.01)i; H04L 1/16(2006.01)n

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)

WPI, EPODOC, CNPAT, CNKI: 拥塞, 控制, 发送, 接收, 插入, 加入, 增加, 添加, 带宽利用率, 反馈, 返回, 确认, 调整, 改变, 增大, 减小, 窗口, congest+, control, transmit, receive, insert, add, bandwidth, usage, utilization, return, feedback, ACK, adjust, change, window

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 101056259 A (CENTRAL SOUTH UNIVERSITY) 17 October 2007 (2007-10-17) description page 7 line 3 from the bottom to page 10, line 7, claim 1, figures 1-2	1-22
Y	US 2013163428 A1 (THE CHINESE UNIVERSITY OF HONG KONG) 27 June 2013 (2013-06-27) description, paragraphs [0047]-[0052]	1-22
Y	CN 106911583 A (ECONOMICS & TECHNOLOGY RESEARCH INSTITUTE, STATE GRID SICHUAN ELECTRIC POWER COMPANY) 30 June 2017 (2017-06-30) description paragraphs [0045]-[0102], figure 1	1-22
A	CN 106059951 A (CENTRAL SOUTH UNIVERSITY et al.) 26 October 2016 (2016-10-26) entire document	1-22
A	CN 102388584 A (HUAWEI TECHNOLOGIES CO., LTD.) 21 March 2012 (2012-03-21) entire document	1-22
A	CN 101645765 A (SICHUAN UNIVERSITY) 10 February 2010 (2010-02-10) entire document	1-22



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

13 August 2020

Date of mailing of the international search report

27 August 2020

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

International application No.

PCT/CN2020/093138

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2017192414 A1 (MICROSOFT TECHNOLOGY LICENSING, LLC) 09 November 2017 (2017-11-09) entire document	1-22

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/093138

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	101056259	A	17 October 2007	None			
US	2013163428	A1	27 June 2013	WO	2013000432	A1	03 January 2013
CN	106911583	A	30 June 2017	None			
CN	106059951	A	26 October 2016	None			
CN	102388584	A	21 March 2012	WO	2012149762	A1	08 November 2012
CN	101645765	A	10 February 2010	None			
WO	2017192414	A1	09 November 2017	US	2017324668	A1	09 November 2017

国际检索报告

国际申请号

PCT/CN2020/093138

A. 主题的分类 H04L 12/825(2013.01)i; H04L 1/16(2006.01)n 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)) WPI, EPDOC, CNPAT, CNKI: 拥塞, 控制, 发送, 接收, 插入, 加入, 增加, 添加, 带宽利用率, 反馈, 返回, 确认, 调整, 改变, 增大, 减小, 窗口, congest+, cnotrol, transmit, receive, insert, add, bandwidth, usage, utilization, retur, feedback, ACK, adjust, change, window																							
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>Y</td> <td>CN 101056259 A (中南大学) 2007年 10月 17日 (2007 - 10 - 17) 说明书第7页倒数第3行-第10页7行, 权利要求1, 附图1-2</td> <td>1-22</td> </tr> <tr> <td>Y</td> <td>US 2013163428 A1 (THE CHINESE UNIVERSITY OF HONG KONG) 2013年 6月 27日 (2013 - 06 - 27) 说明书第[0047]-[0052]段</td> <td>1-22</td> </tr> <tr> <td>Y</td> <td>CN 106911583 A (国网四川省电力公司经济技术研究院) 2017年 6月 30日 (2017 - 06 - 30) 说明书第[0045]-[0102]段, 附图1</td> <td>1-22</td> </tr> <tr> <td>A</td> <td>CN 106059951 A (中南大学等) 2016年 10月 26日 (2016 - 10 - 26) 全文</td> <td>1-22</td> </tr> <tr> <td>A</td> <td>CN 102388584 A (华为技术有限公司) 2012年 3月 21日 (2012 - 03 - 21) 全文</td> <td>1-22</td> </tr> <tr> <td>A</td> <td>CN 101645765 A (四川大学) 2010年 2月 10日 (2010 - 02 - 10) 全文</td> <td>1-22</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	Y	CN 101056259 A (中南大学) 2007年 10月 17日 (2007 - 10 - 17) 说明书第7页倒数第3行-第10页7行, 权利要求1, 附图1-2	1-22	Y	US 2013163428 A1 (THE CHINESE UNIVERSITY OF HONG KONG) 2013年 6月 27日 (2013 - 06 - 27) 说明书第[0047]-[0052]段	1-22	Y	CN 106911583 A (国网四川省电力公司经济技术研究院) 2017年 6月 30日 (2017 - 06 - 30) 说明书第[0045]-[0102]段, 附图1	1-22	A	CN 106059951 A (中南大学等) 2016年 10月 26日 (2016 - 10 - 26) 全文	1-22	A	CN 102388584 A (华为技术有限公司) 2012年 3月 21日 (2012 - 03 - 21) 全文	1-22	A	CN 101645765 A (四川大学) 2010年 2月 10日 (2010 - 02 - 10) 全文	1-22
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																					
Y	CN 101056259 A (中南大学) 2007年 10月 17日 (2007 - 10 - 17) 说明书第7页倒数第3行-第10页7行, 权利要求1, 附图1-2	1-22																					
Y	US 2013163428 A1 (THE CHINESE UNIVERSITY OF HONG KONG) 2013年 6月 27日 (2013 - 06 - 27) 说明书第[0047]-[0052]段	1-22																					
Y	CN 106911583 A (国网四川省电力公司经济技术研究院) 2017年 6月 30日 (2017 - 06 - 30) 说明书第[0045]-[0102]段, 附图1	1-22																					
A	CN 106059951 A (中南大学等) 2016年 10月 26日 (2016 - 10 - 26) 全文	1-22																					
A	CN 102388584 A (华为技术有限公司) 2012年 3月 21日 (2012 - 03 - 21) 全文	1-22																					
A	CN 101645765 A (四川大学) 2010年 2月 10日 (2010 - 02 - 10) 全文	1-22																					
☒ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																							
<table border="0"> <tr> <td> * 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 </td> <td> “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 </td> </tr> </table>			* 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																			
* 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																						
2020年 8月 13日	2020年 8月 27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	郭风顺																						
传真号 (86-10)62019451	电话号码 86-10-53961601																						

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
A	WO 2017192414 A1 (MICROSOFT TECHNOLOGY LICENSING, LLC) 2017年 11月 9日 (2017 - 11 - 09) 全文	1-22

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/093138

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	101056259	A	2007年 10月 17日	无			
US	2013163428	A1	2013年 6月 27日	W0	2013000432	A1	2013年 1月 3日
CN	106911583	A	2017年 6月 30日	无			
CN	106059951	A	2016年 10月 26日	无			
CN	102388584	A	2012年 3月 21日	W0	2012149762	A1	2012年 11月 8日
CN	101645765	A	2010年 2月 10日	无			
W0	2017192414	A1	2017年 11月 9日	US	2017324668	A1	2017年 11月 9日