(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 2 月 18 日 (18.02.2021)

WIPO | PCT

(10) 国际公布号
WO 2021/026740 A1

(51) 国际专利分类号:
H04L 29/08 (2006.01) *H04L 12/803* (2013.01)

(21) 国际申请号: PCT/CN2019/100270

(22) 国际申请日: 2019 年 8 月 12 日 (12.08.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(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: 林 云 (**LAM, Wan**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 塔尔亚利克斯 (**TAL, Alex**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 唐德智 (**TANG, Dezhi**); 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 中 博 世 达 专 利 商 标 代 理 有 限 公 司 (**BEIJING ZBSD PATENT&TRADEMARK AGENT LTD.**); 中国北京市海淀区交大东路31号11号楼8层, Beijing 100044 (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, 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))。

(54) **Title:** TRAFFIC BALANCING METHOD, NETWORK DEVICE AND ELECTRONIC DEVICE

(54) 发明名称: 流量均衡方法、网络设备及电子设备

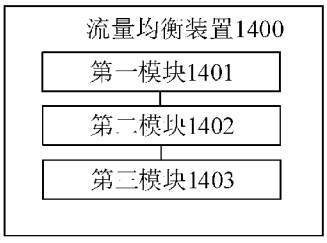


图 14

1400 Traffic balancing apparatus
1401 First module
1402 Second module
1403 Third module

(57) **Abstract:** The present application provides a traffic balancing method, a network device and an electronic device, and relates to the technical field of communications. The network device comprises: a first module, configured to obtain a data packet to be sent; a second module, configured to create or maintain stream packets and classify said data packet to a corresponding stream packet according to a destination node; and a third module, configured to send said data packet based on the stream packet to which said data packet belongs, wherein one stream packet comprises at least one data packet, the destination nodes of the data packets belonging to the same stream packet are the same, and the sending paths of the data packets belonging to the same stream packet are the same.

(57) 摘要: 本申请提供一种流量均衡方法、网络设备及电子设备，涉及通信技术领域。其中，网络设备包括：第一模块，用于获取待发送的数据包；第二模块，用于创建或维护流包，并将待发送的数据包按照目的节点划入对应的流包；第三模块，用于基于待发送的数据包所属的流包，发送所述待发送的数据包，其中，一个流包包括至少一个数据包，属于同一流包中的数据包的目的节点相同，且属于同一流包中的数据包的发送路径相同。

WO 2021/026740 A1

流量均衡方法、网络设备及电子设备

技术领域

本申请涉及通信技术领域，尤其涉及流量均衡方法、网络设备及电子设备。

背景技术

在数据中心网络（data center network，DCN）中，可以采用不同的组网模式，为DCN内的众多服务器（server）提供网络。参见图1，为多种组网模式中的一种常见的三层组网模式。其中，DCN划分为三层，接入（access）层的顶架（top of rack，TOR）)节点的下行端口连接服务器。TOR节点的上行端口连接汇聚（aggregation）层的aggregation节点的下行端口。aggregation节点的上行端口与核心（core）层的脊（spine）节点（也称为骨干节点）的下行端口连接。

从图1所示架构可以看出，从数据源端到数据目的之间可能存在多条路径可用。比如，从图1中的数据源端S0到数据目的端S4存在多条可用路径。因此，亟待提出一种流量均衡方法，以便数据均匀的分摊到多条可用路径上，以最大化DCN的带宽利用率。

发明内容

本申请实施例提供的流量均衡方法、网络设备及电子设备，能够动态的控制路径切换，实现更加灵活的流量均衡。

为达到上述目的，本申请实施例提供如下技术方案：

第一方面，本申请实施例提供一种网络设备，该网络设备可以指独立存在的设备，也可以指集成在设备中的组件，比如可以指设备中的芯片系统。该网络设备可以是具有发送节点功能的装置，比如可以是发送节点，也可以是发送节点中的芯片系统，该网络设备包括如下功能模块：第一模块，用于获取待发送的数据包；第二模块，用于创建或维护流包，并将待发送的数据包按照目的节点划入对应的流包；第三模块，用于基于待发送的数据包所属的流包，发送待发送的数据包。

其中，一个流包包括至少一个数据包，属于同一流包中的数据包的目的节点相同，且属于同一流包中的数据包的发送路径相同。

需要说明的是，流包为本申请实施例提出的一种数据组合形式。流包指的是一系列数据包的集合。不同流包可以通过流包标签等方式来区分。

相比于现有技术中难以触发足够粒度的动态负载均衡，本申请实施例提供的流量均衡方法，灵活创建或维护流包，并按照目的节点将待发送数据包划入对应流包，如此，可以将同一流包中的数据包通过同一路径发送，不同流包中的数据包通过不同路径发送，即按照本申请实施例的技术方案，能够实现根据数据包所属流包动态的控制路径切换。并且，路径切换的粒度与流包的划分方式相关。其中，当流包包括较少的数据包时，路径切换粒度较细，即每发送较少的数据包就可触发路径切换，当流包包括较多的数据包时，路径切换粒度较粗，即每发送较多的数据包才可以触发路径切换。可见，本申请实施例提供的流量均衡方法，还能够通过控制流包的划分方式灵活控制

路径切换粒度，以满足不同应用场景下的流量均衡需求。

作为一种可能的实现方式，第二模块还用于：创建或维护第一流包；将以第一节点为目标的数据包，划入第一流包；创建或维护第二流包；将之后发送的以第一节点为目标的数据包，划入第二流包。

其中，第二流包中的数据包的发送路径，与第一流包中的数据包的发送路径不同。

作为一种可能的实现方式，第二模块，还用于：判断网络均衡参数是否满足预设条件；若网络均衡参数满足预设条件，则创建或维护第二流包；网络均衡参数用于基于网络均衡原则将数据包划入对应的流包。

其中，网络均衡参数满足预设条件包括满足如下任一个条件或多个条件：

1、从发送第一流包中第一数据包的时刻开始的预设间隔内，不存在去往同一目的节点的数据包；也就是说，发送方获取某一待发送数据包，并将该数据包划入某一已创建流包，从该数据包被发送的时刻开始，若较长时间内，即预设间隔内，不存在与该数据包具有同一目的节点的其他数据包，则发送方创建新的流包，并可以将预设间隔之外的后续待发送数据包划入该新创建的流包。如此，使得时间间隔较大的两个数据包可以被划入不同流包。

2、第一流包的数据量达到预设数据量；该预设条件是说，若某一已创建流包中数据包的数据量达到预设数据量，则发送方创建新的流包。换句话说，每个流包均包括相同数据量的数据包，即每一流包包括预设数据量的数据包。如此，使得每一流包中数据包的数据量相同，以平衡不同流包中数据包占用的带宽资源。

3、第一流包的持续时长达到预设时长；已创建流包的持续时长，指的是该已创建流包中数据包的持续时长，具体的，指从该已创建流包中第一个数据包的发送时刻开始的一段时间。若从某一已创建流包中第一个数据包的发送时刻开始达到预设时长仍不存在去往同一目的节点的数据包，则发送方创建新的流包，并将预设时长之后时刻发送的数据包划入该新创建流包中。如此，使得每一流包中去往同一目的节点的数据包占用的时间资源大致相同，即能够降低某一个或几个流包中去往同一目的节点的数据包占用较多时间资源，而其他流包中数据包占用较少的时间资源的概率，平衡不同流包中数据包占用的时间资源。

4、第一流包的发送频率达到预设频率。已创建流包的发送频率，指的是。数据包的发送频率可以反映数据包发送的快慢，通常，数据包的发送频率受数据包相互之间的发送时间间隔影响。通常，数据包的发送频率与数据发送需求和/或其他因素有关。比如，当使用突发方式发送数据时，数据包的发送频率不固定，可以在某一时段内有较快的发送频率，可以在其他时段有较慢的发送频率。当采用平滑方式发送数据时，数据包的发送频率较为固定，可以是匀速发送数据包。通常，若采用突发方式发送数据，且当前已创建流包中一个或多个 flowlet 的发送频率较大，为了减轻当前已创建流包对应路径的负载，可以创建新的流包，并将之后发送的 flowlet 划入新创建的流包中。如此，以便于平衡各个流包对应路径之间的负载，降低某一个或多个流包中因 flowlet 的发送频率较快，导致该一个或多个流包对应的路径负载过重的概率。

作为一种可能的实现方式，可以通过为不同流包的数据包打不同标签来区分数据包属于哪一流包，具体的，第一流包中至少一个数据包携带第一定界标识，第二流包

中至少一个数据包携带第二定界标识，第一定界标识和第二定界标识用于区分第一流包和第二流包。

作为一种可能的实现方式，可以通过在不同流包之间插入控制包来区分数据包属于哪一流包，这种实现方式中，第三模块，还用于通过第一路径发送定界包，定界包为用于区分第一流包和第二流包的控制包，第一路径为发送第一流包中数据包的路径。

作为一种可能的实现方式，一个流包内的数据包来自相同或不同数据流。

作为一种可能的实现方式，第二模块，还用于基于网络均衡的原则为流包中的数据包设定发送路径。

第二方面，本申请实施例提供一种网络设备，该网络设备可以指独立存在的设备，也可以指集成在设备中的组件，比如可以指设备中的芯片系统。该网络设备可以是具有接收节点功能的装置，比如可以是接收节点，也可以是接收节点中的芯片系统，该网络设备包括如下功能模块：第四模块，用于判断在接收到第二流包的情况下，是否已接收到第一流包的尾包；第五模块，用于若未接收到第一流包的尾包，则通过排序通道 RC 对第一流包和第二流包中的数据包执行重排序；第六模块，用于在重排序之后，若确定已接收到第一流包的尾包，则释放 RC。

容易理解的是，接收方在接收到新流包（比如第二流包）之后，若通过某些方式确定未接收到上一流包，即第一流包的尾包，则说明第一流包、第二流包并未按照先发先至、后发后至的顺序达到接收方，导致乱序。因此，接收方通过 RC 执行重排序。本申请实施例中，接收方在接收到第一流包的尾包后，释放 RC。容易理解的是，接收方接收到第一流包的尾包，说明第一流包已接收完毕，如此，第一流包和第二流包之间不再存在乱序问题，可以释放该 RC，并将释放的 RC 用于其他重排序流程，从而提升 RC 的利用率。由于闲置 RC 还可以用于其他重排序流程，相当于拓展了可用 RC 的数目。

第三方面，本申请实施例提供一种电子设备，包括处理器和存储设备，存储设备用于存储指令，处理器用于基于指令执行下列动作：创建或维护第一流包；将以第一节点为目标的数据包，划入第一流包；在网络均衡参数满足预设条件时，创建或维护第二流包；将之后来的以第一节点为目标的数据包，划入第二流包，其中，属于同一流包中的数据包的目的节点相同，属于同一流包中的数据包的发送路径相同；第二流包中的数据包的发送路径，与第一流包中的数据包的发送路径不同。

第四方面，本申请实施例提供一种电子设备，包括处理器和存储设备，存储设备用于存储指令，处理器用于基于指令执行下列动作：判断在接收到第二流包的情况下，是否已接收到第一流包的尾包；若未接收到第一流包的尾包，则通过排序通道 RC 对第一流包和第二流包中的数据包执行重排序；在重排序之后，若确定已接收到第一流包的尾包，则释放 RC。

第五方面，本申请实施例提供一种流量均衡方法，该方法可以由上述第一方面或第三方面的装置执行。该方法包括如下步骤：获取待发送的数据包，创建或维护流包，并将待发送的数据包按照目的节点划入对应的流包；基于待发送的数据包所属的流包，发送待发送的数据包。

其中，一个流包包括至少一个数据包，属于同一流包中的数据包的目的节点相同，

且属于同一流包中的数据包的发送路径相同。

作为一种可能的实现方式，以创建或维护第一流包和第二流包为例，创建或维护流包，并划分不同数据包所属流包的方法包括如下步骤：创建或维护第一流包；将以第一节点为目标的数据包，划入第一流包；创建或维护第二流包；将之后发送的以第一节点为目标的数据包，划入第二流包。

其中，第二流包中的数据包的发送路径，与第一流包中的数据包的发送路径不同。

作为一种可能的实现方式，创建或维护第二流包，具体可以实现为如下步骤：判断网络均衡参数是否满足预设条件；若网络均衡参数满足预设条件，则创建或维护第二流包；其中，网络均衡参数用于基于网络均衡原则将数据包划入对应的流包。

网络均衡参数满足预设条件包括满足如下任一个条件或多个条件：

从发送第一流包中第一数据包的时刻开始的预设间隔内，不存在去往同一目的节点的数据包；第一流包的数据量达到预设数据量；第一流包的持续时长达到预设时长；第一流包的发送频率达到预设频率。

作为一种可能的实现方式，可以通过数据包的标签来区分数据包属于哪一流包。第一流包中至少一个数据包携带第一定界标识，第二流包中至少一个数据包携带第二定界标识，第一定界标识和第二定界标识用于区分第一流包和第二流包。

作为一种可能的实现方式，可以通过在数据包之间插入控制包来区分属于不同流包的数据包，通过第一路径发送定界包，定界包为用于区分第一流包和第二流包的控制包，第一路径为发送第一流包中数据包的路径。

作为一种可能的实现方式，一个流包内的数据包来自相同或不同数据流。

作为一种可能的实现方式，上述方法还包括：基于网络均衡的原则为流包中的数据包设定发送路径。

第六方面，本申请实施例提供一种流量均衡方法，该方法可以由上述第二方面或第四方面的装置（即具备接收节点功能的装置）执行，该方法包括如下步骤：

判断在接收到第二流包的情况下，是否已接收到第一流包的尾包；若未接收到第一流包的尾包，则通过排序通道 RC 对第一流包和第二流包中的数据包执行重排序；并在重排序之后，若确定已接收到第一流包的尾包，则释放 RC。

第七方面，本申请提供一种流量均衡装置，该流量均衡装置具有实现上述任一方面任一项的流量均衡方法的功能。实现该功能可以通过硬件实现，也可以通过硬件执行相应的软件实现。该硬件或软件包括一个或多个与上述功能相对应的模块。

第八方面，提供一种流量均衡装置，包括：处理器和存储器；该存储器用于存储计算机执行指令，当该流量均衡装置运行时，该处理器执行该存储器存储的该计算机执行指令，以使该流量均衡装置执行如上述任一方面中任一项的流量均衡方法。

第九方面，提供一种流量均衡装置，包括：处理器；处理器用于与存储器耦合，并读取存储器中的指令之后，根据指令执行如上述任一方面中任一项的流量均衡方法。

第十方面，提供一种计算机可读存储介质，该计算机可读存储介质中存储有指令，当其在计算机上运行时，使得计算机可以执行上述任一方面中任一项的流量均衡方法。

第十一方面，提供一种包含指令的计算机程序产品，当其在计算机上运行时，使得计算机可以执行上述任一方面中任一项的流量均衡方法。

第十二方面，提供一种电路系统，电路系统包括处理电路，处理电路被配置为执行如上述第五方面或者第六方面中任一项的流量均衡方法。

第十三方面，提供一种芯片，芯片包括处理器，处理器和存储器耦合，存储器存储有程序指令，当存储器存储的程序指令被处理器执行时实现上述第五方面或者第六方面任意一项的流量均衡方法。

附图说明

图 1 为本申请实施例提供的系统架构示意图；

图 2 为本申请实施例提供的流块的示意图；

图 3 为本申请实施例提供的一种负载均衡的原理示意图；

图 4 为本申请实施例提供的另一种负载均衡的原理示意图；

图 5 为本申请实施例提供的流量均衡的方法流程图；

图 6 为本申请实施例提供的流量均衡的方法流程图；

图 7 至图 10 为本申请实施例提供的流量均衡方法的原理示意图；

图 11 为本申请实施例提供的数据包排序的原理示意图；

图 12 至图 15 为本申请实施例提供的流量均衡装置的结构示意图。

具体实施方式

本申请的说明书以及附图中的术语“第一”和“第二”等是用于区别不同的对象，或者用于区别对同一对象的不同处理，而不是用于描述对象的特定顺序。

“至少一个”是指一个或者多个，

“多个”是指两个或两个以上。

“和/或”，描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 的情况，其中 A，B 可以是单数或者复数。字符“/”一般表示前后关联对象是一种“或”的关系，例如，A/B 可以表示 A 或 B。

此外，本申请的描述中所提到的术语“包括”和“具有”以及它们的任何变形，意图在于覆盖不排他的包含。例如包含了一系列步骤或单元的过程、方法、系统、产品或设备没有限定于已列出的步骤或单元，而是可选地还包括其他没有列出的步骤或单元，或可选地还包括对于这些过程、方法、产品或设备固有的其它步骤或单元。

需要说明的是，本申请实施例中，“示例性的”或者“例如”等词用于表示作例子、例证或说明。本申请实施例中被描述为“示例性的”或者“例如”的任何实施例或设计方案不应被解释为比其它实施例或设计方案更优选或更具优势。确切而言，使用“示例性的”或者“例如”等词旨在以具体方式呈现相关概念。

本申请的说明书以及附图中“的（英文：of）”，相应的“（英文 corresponding，relevant）”和“对应的（英文：corresponding）”有时可以混用，应当指出的是，在不强调其区别时，其所要表达的含义是一致的。

本申请实施例提供的流量均衡方法应用在需进行流量均衡的系统中。示例性的，应用在需进行流量均衡的 DCN 中。本申请实施例涉及的 DCN 可以具有不同组网模式也可以具有不同的层级。如下主要以应用在三层的 DCN 中为例来说明本申请实施例所适用的系统架构。参见图 1，该三层的 DCN 包括接入层、汇聚层、核心层。

其中，接入层的节点可称为接入(access)节点。在不同组网模式中，接入节点的具

体实现方式和名称可能不同。示例性的，在采用叶脊（leaf-spine）组网模式的 DCN 中，接入节点可称为叶子（leaf）节点。其中，若 leaf 节点位于机架顶部，则这种位于机架顶部的 leaf 节点还可被称为 TOR 节点。如下主要以接入节点为 TOR 节点为例说明，接入节点还可能为其他形式的节点，本申请实施例不再一一列举。TOR 节点的下行端口连接有服务器，TOR 节点可以通过服务器收发数据。

汇聚层的节点可称为汇聚节点。汇聚节点的下行端口与 TOR 节点的上行端口连接。

核心层的节点可称为核心（core）节点，在采用 leaf-spine 组网模式的 DCN 中，核心节点可称为 spine 节点。如下主要以核心层节点为 spine 节点为例进行说明，在此统一说明。核心节点还可能为 spine 节点之外的其他节点，本申请实施例不再一一列举。spine 节点的端口可以与 aggregation 节点的上行端口连接。

在本申请实施例中，上述接入节点、汇聚节点和核心节点，均可以称为交换节点。在此统一声明，下文不再赘述。

其中，一个或多个 TOR 节点以及一个或多个汇聚节点可以构成群组（point of delivery，Pod）。还可以将 spine 节点划分到不同的 spine 平面(plane)。同一 Pod 的不同 aggregation 节点可以分别连到不同的 spine 平面。示例性的，图 1 的 spine 节点划分为 3 个平面。每个平面中的 spine 节点分别和各 Pod 内的不同 aggregation 节点直接相连。

在一种可能的设计中，同一 Pod 内的的 aggregation 节点和 TOR 节点可以是全连接关系。即某一 PoD 内的每一 TOR 节点均与该 PoD 内的全部 aggregation 节点相连接，且该 PoD 内的每一 aggregation 节点均与该 PoD 内的全部 TOR 节点相连接。当然，同一 PoD 内的 TOR 节点和 aggregation 节点之间的连接关系还可能为其他形式，本申请实施例对此不进行限制。

其中，aggregation 节点，用于完成同一 Pod 内跨 TOR 节点之间的流量交换。比如图 1 中从源 d0 到目的 d1 之间的流量交换。spine 节点，用于完成跨 Pod 间流量的交换。比如图 1 中从源 S0 到目的 s4 之间的流量交换。

为了便于理解本申请实施例的内容，如下对一些技术术语进行介绍：

1、数据流：是指一组有序，有起点和终点的字节的数据序列。属于同一条数据流的数据包通常具有相同的属性，比如相同的源网络互连协议（internet protocol address，IP）地址，目的 IP 地址，目的端口（port）等。源 IP 地址，目的 IP 地址，目的端口（port）等通常可以构成五元组（5-tuple）。其中，在一种常用的数据流定义方式中，来自同一发送源且发往同一目的地的字节数据可称为同一数据流。当然，数据流还可以有其他定义。

2、流块（flowlet）：传输控制协议（transmission control protocol，TCP）通常可以采用固定速率发包，比如按照速率 5Gbps 发包，也可以不按照固定速率发包，而是可以某些时刻（或较短的时间间隔内）发送大量数据包，在另一些时刻（或较短时间间隔内）发送小量数据包，这种不按照固定速率发包的方式可称为突发（burst）发包方式。其中，每次发送的一个或多个数据包可称为一个 flowlet。通常，一条数据流可包括多个 flowlet。比如，以图 2 为例，在 t1-t2 时段内，发送节点（比如 TOR 节点）

发送数据量较小的 flowlet1，在 t3-t4 时段内，发送节点发送数据量较大的 flowlet2，在 t5-t6 时段内，发送节点发送 flowlet3。

参见图 3，a 为发送节点、b 为接收节点，flowlet1 为发送节点采用上述突发发包方式在 t3-t4 时段内发送的多个数据包，flowlet2 为发送节点采用上述突发发包方式在 t1-t2 时段内发送的多个数据包，两条链路的路径延时 delay 分别为 d1, d2，flowlet1 和 flowlet2 之间的时间间隔为 Gap。若 flowlet1 和 flowlet2 属于同一数据流，通常当 flowlet1, flowlet2 的时间间隔 Gap>|d1−d2|，则这两个 flowlet 可经不同路径发送，而无需担心乱序。这里，flowlet1, flowlet2 之间的时间间隔，指的是 flowlet1 中最后一个数据包和 flowlet2 中第二个数据包之间的时间间隔，即 t2-t3 这段时间间隔。

在本申请实施例中，flowlet 还可以称为流段，或者其他名称。

3、按流分发（也称静态负载均衡）：在一种传统的流量均衡（也称为负载均衡（load balance，LB））方式中，DCN 中的交换节点采用哈希（hash）算法对数据流（flow）进行 hash 计算，并基于 hash 结果在多条可用路径中，选择一条用于发送该数据流的路径。其中，hash 算法可以用上述提及的五元组作为输入。以图 1 为例，若仅考虑同 PoD 内的流量交换，d0 至 d2 的可用路径有 4 条，即路径 d0-s0-d2、路径 d0-s1-d2、路径 d0-s2-d2、路径 d0-s3-d2。假设定义当 hash 结果为 0 时，表示选择 PoD#1 中 s0 来转发流量，当 hash 结果为 1 时，表示选择 PoD#1 中 s1 来转发流量，hash 结果为 2 时，表示选择 PoD#1 中 s2 来转发流量，hash 结果为 3 时，表示选择 PoD#1 中 s3 来转发流量。则若当前某条数据流的 hash 结果为 2 时，通过路径 d0-s2-d2 转发该流量。

上述这种负载均衡方式，属于同一数据流的数据包（packet）通常通过同一路径发送。这种基于数据流进行的负载均衡，称为按流分发（flow-basedload balance，FLB）或者按流负载均衡，或者称为静态负载均衡（static load balance，SLB）。

上述的按流负载均衡方式，由于通常将属于同一数据流的数据包承载在同一路径上，可保证同一数据流的数据包之间不产生乱序。因此，接收端通常无需对接收的数据包进行重排序。

但是，按流负载均衡方式可能会导致数据收发端口拥塞。具体的，采用 hash 算法选路的 SLB 机制会产生哈希冲突（hash collision），比如，hash 冲突可能会发生在 TOR 节点至 aggregation 节点，或者 aggregation 节点至 spine 节点的上行端口上。以 TOR 节点和 aggregation 节点之间出现 hash 冲突为例，参见图 4，多个数据流的 hash 值可能均相同，如此，TOR 节点采用 hash 算法进行选路时，可能将 hash 值相同的多个数据流发往同一 aggregation 节点。如果存在多个同时活跃的数据流，则可能造成大量数据包同时发往同一上行端口，进而可能造成该上行端口拥塞。

4、按包（packet）分发：各交换节点均按 packet 在多条可用路径中均匀分发。如此，属于同一条数据流的多个数据包可能通过不同路径发送。这种机制可统称为按包分发（packetspray）。packetspray 是一种（dynamic load balancing，DLB）技术。与按流分发相比，按包分发技术的优点是：不存在上述哈希冲突的问题，且可以更好地利用多路径的带宽。

但是，按包分发也存在相应缺点，由于属于同一条数据流的 packet 会经不同的路径发送，不同路径的拥塞程度可能不同，造成不同路径的时延可能不同。因此，同一

数据流的不同数据包到达目的地的时间可能不同。也就是说：属于同一条数据流的packet之间可能会产生乱序（即，后发先至、先发后至的情况），这就需要接收端通过专门的设计来对乱序的数据包做重排序，比如依靠硬件或软件执行重排序。在一种可能的实现方式中，每一条需要排序的数据流可以独占或与其它数据流共用一个排序通道 (re-ordering channel，RC)，也称为重排序逻辑。RC即用于重排序的逻辑、缓存、资源、以及相关的数据结构，或者其他相关内容的统称。占用RC的时间和数据流的生命周期（比如10ms或2s）相关。如此，当数据流较多时，接收侧需有足够的RC，来对数据包进行重排序。也就是说，若每条数据流独占一个RC，那么如果有10K条数据流，则需要10K个RC，导致接收端实现的复杂性（complexity）或可扩展性（scalability）的问题。

在一个示例中，突发发包方式也可以实现DLB，即通过上述突发发包方式可以为属于同一条数据流的不同flowlet动态地切换发送路径，比如图3中将属于同一数据流的flowlet1和flowlet2通过不同路径发送，从而令流量分布更均匀。

但如下两个方面限制flowlet应用：

一方面，DCN带宽大，通常路径之间的延时差较小（比如延时差为20us），所以，过大的时间间隔Gap设置(比如在某些算法中，设置Gap=500us)难以触发足够细粒度的动态负载均衡。也就是说，当设置Gap值过大时，两个flowlet之间的时间间隔只有大于该Gap才能切换路径，即相对于高速传输的数据流来说，两个flowlet之间的时间间隔往往较小，每隔很长时间可能才触发一次路径切换，无法满足动态负载均衡的性能。

另一方面，网络中的动态行为可能会影响路径切换。例如，TCP源端采用平滑（pacing）方式发送数据，比如采用pacing发送数据，由于数据包之间的时间间隔较小，难以满足预设Gap，难以触发路径切换。另外，当交换节点资源耗尽，比如交换节点的缓冲区（buffer）耗尽时，交换节点可触发反压（backpressure），其向数据发送方发送消息，以指示数据发送方暂缓发送。如此，造成数据在发送方等待，相应的，发送数据之间的时间间隔可能无法保证大于路径间延时差。也就是说可能造成乱序。

上文已指出，按照上述按包分发方式进行流量均衡时，难以触发足够粒度的动态负载均衡，比如，可能每隔很长时间才触发一次路径切换。在按照按流分发方式进行流量均衡时，并不触发动态负载均衡。如此，导致目前的负载均衡性能较低。为了解决这一技术问题，本申请实施例提供一种流量均衡方法，参见图5，该方法包括如下步骤：

S501、发送方获取待发送的数据包。

容易理解的是，当发送方有数据需发送时，将待发送的数据按照协议定义格式封装为数据包。根据待发送数据量的大小，发送方可以将待发送数据封装成一个或多个数据包，即本申请实施例中待发送数据包可以指一个或多个数据包。

S502、发送方创建或维护流包（flowpac），并将所述待发送的数据包按照目的节点划入对应的流包。

其中，流包为本申请实施例提出的一种数据组合形式。流包指的是一系列数据包的集合，即一个流包包括一个或多个数据包。一属于同一流包中的数据包的目的节点

相同，即属于同一流包的数据包均去往同一目的节点。不同流包可以通过流包标签等方式来区分，区分流包的具体方案介绍可参见下文。需要说明的是，流包并非种协议规定的数据封装（data encapsulation）体。

在一种可能的实现方式中，若采用上述突发发包方式，流包可以包括一个或多个完整的 flowlet，比如，以图 2 为例，一个流包可以包括 flowlet1 和 flowlet2 中的全部数据包。当然，流包也可以包括 flowlet 中的部分数据包，仍以图 2 为例，一个流包可以包括 flowlet1 的全部数据包和 flowlet2 中的前 2 个数据包。flowlet 的具体介绍可参见上文，这里不再赘述。此外，流包中的数据包可以来自同一数据流，也可以来自不同数据流。即流包中的数据包可以为某一数据流中的部分或全部数据包，也可以为多个数据流中的数据包，比如，一个流包包括 flowlet1 和 flowlet2 中的全部数据包，flowlet1 和 flowlet2 属于不同数据流。又比如，一个流包包括 flowlet1 和 flowlet2 中的全部数据包，flowlet1 和 flowlet2 属于同一数据流。

本申请实施例中，发送方可以利用获取的待发送数据包创建或维护不同的流包，这里，发送方获取的多个待发送数据包可以划入同一流包，也可以将待发送数据包中的部分数据包划入某一流包，将待发送数据包中的其他部分数据包划入其他流包，即待发送数据包可被划入不同流包。以创建、维护某一流包为例，发送方获取待发送数据包后，需先创建流包，后续，发送方可以继续获取待发送数据包，来维护所创建的流包。比如，当前的待发送数据包包括 3 个数据包，发送方可根据这 3 个数据包的目的节点创建一个包括 3 个数据包的流包，即将这 3 个数据包划入该流包。之后，随着时间的推移，发送方获取后续时刻的待发送数据包，并维护之前创建的该流包，即可以将后续时刻的待发送数据包划入之前创建的该流包，更新该流包包括的数据包数目。如此，发送方获取待发送的一个或多个数据包，并创建或维护多个流包，从而将一个或多个数据包划入对应的流包，以便于后续根据每一数据包所属流包为数据包决策发送路径。

在一些实施例中，发送方将数据包具体划入哪一流包的依据可以是数据包的目的节点，即将发送至同一目的节点的数据包划入同一流包。并且，属于同一流包中的数据包的发送路径相同。

在另一些实施例中，发送方为数据包划分所属流包的依据还可以是数据包的网络均衡参数，当网络均衡参数满足预设条件时，可以创建新的流包，并将之后待发送的数据包划入该新创建的流包。网络均衡参数用于发送方基于网络均衡原则将多个数据包分别划入不同流包，且不同流包的数据包通过不同路径发送。也就是说，可以基于网络均衡原则为流包中的数据包设定发送路径。本申请实施例中，网络均衡主要指链路之间的流量均衡，流量均衡即在一段时间内，多条链路中每条链路的流量之间的差异在预设范围内。换言之，流量均衡，指的是流量通过多条链路分担，且每条链路分担的流量值差异不大。而当仅通过多条链路中的几条链路来分担大部分流量，而其他链路仅分担小部分流量时，我们可以视为链路之间的流量不均衡。本申请实施例中，发送方基于网络均衡原则为数据包确定所需划入至哪一流包，指的是，当存在较多数据包时，需要根据网络均衡参数将一部分数据包划入某一流包，将另一部分数据包划入另外流包中，如此，后续，一部分数据包可通过某一流包对应的路径发送，另一部

分数据包可通过上述另外流包对应的路径发送，即能够将多个数据包通过不同路径发送，实现多条路径之间的流量均衡。

其中，数据包的网络均衡参数可以用于表征该数据包占用的网络资源，网络资源比如可以为带宽资源、时间资源等。本申请实施例中用于流量均衡的主要网络均衡参数比如可以为如下一种或几种的组合：数据包之间的时间间隔 Gap、数据包的数据量 Size、数据包的持续时长 Time、数据包的发送频率，当然，还可以为其他各种可度量的方式。

该一个或多个网络均衡参数用于控制 DLB 的动态能力。DLB 的动态能力指的是数据包灵活选择不同路径的能力。通常，动态能力越强，不同数据包通过不同路径传输的可能性越大，或者说，较容易触发路径切换。动态能力越弱，不同数据包往往通过相同的一条或几条路径承载，换句话说，即难以触发路径切换。

当然，除了上述所列举的网络均衡参数，本申请实施例用于流量均衡的网络均衡参数还可以为其他参数，比如，交换节点从其他交换节点获取的不同路径的拥塞程度，比如 TOR 从上游 aggregation 节点获取的该 TOR 至该 aggregation 之间路径的拥塞程度，又比如，TOR 从 spine 0 获取的 spine 0 至该 TOR 之间路径的拥塞程度、某个交换节点的发送队列的深度、交换节点的滑动窗口（sliding window）的特征，比如滑动窗口大小。本段列举的这些网络均衡参数可以用于反映网络的拥塞状态。本申请实施例用于流量均衡的网络均衡参数还可以包括其他参数，本申请实施例这里不再一一列举这些参数。

上文已指出，在发送方持续获取待发送数据包，并创建或维护不同流包的过程中，若所述网络均衡参数满足预设条件，发送方可以创建新的流包，并可以将后续待发送数据包划入该新创建的流包，其中，网络均衡参数满足预设条件包括如下一项或多项的组合：

1、从发送已创建流包中第一数据包的时刻开始的预设间隔内，不存在去往同一目的节点的数据包，也就是说，发送方获取某一待发送数据包，并将该数据包划入某一已创建流包，从该数据包被发送的时刻开始，若较长时间内，即预设间隔内，不存在与该数据包具有同一目的节点的其他数据包，则发送方创建新的流包，并可以将预设间隔之外的后续待发送数据包划入该新创建的流包。如此，使得时间间隔较大的两个数据包可以被划入不同流包。

2、所述已创建流包的数据量达到预设数据量。已创建流包的数据量，指的是该已创建流包中数据包的数据量。该预设条件是说，若某一已创建流包中数据包的数据量达到预设数据量，则发送方创建新的流包。换句话说，每个流包均包括相同数据量的数据包，即每一流包包括预设数据量的数据包。如此，使得每一流包中数据包的数据量相同，以平衡不同流包中数据包占用的带宽资源。

3、所述已创建流包的持续时长达到预设时长。已创建流包的持续时长，指的是该已创建流包中数据包的持续时长，具体的，指从该已创建流包中第一个数据包的发送时刻开始的一段时间。若从某一已创建流包中第一个数据包的发送时刻开始达到预设时长仍不存在去往同一目的节点的数据包，则发送方创建新的流包，并将预设时长之后时刻发送的数据包划入该新创建流包中。如此，使得每一流包中去往同一目的节点

的数据包占用的时间资源大致相同，即能够降低某一个或几个流包中去往同一目的节点的数据包占用较多时间资源，而其他流包中数据包占用较少的时间资源的概率，平衡不同流包中数据包占用的时间资源。

4、所述已创建流包的发送频率达到预设频率。已创建流包的发送频率，指的是。数据包的发送频率可以反映数据包发送的快慢，通常，数据包的发送频率受数据包相互之间的发送时间间隔影响。通常，数据包的发送频率与数据发送需求和/或其他因素有关。比如，当使用上述突发方式发送数据时，数据包的发送频率不固定，可以在某一时段内有较快的发送频率，可以在其他时段有较慢的发送频率。当采用平滑方式发送数据时，数据包的发送频率较为固定，可以是匀速发送数据包。通常，若采用突发方式发送数据，且当前已创建流包中一个或多个 flowlet 的发送频率较大，为了减轻当前已创建流包对应路径的负载，可以创建新的流包，并将之后发送的 flowlet 划入新创建的流包中。如此，以便于平衡各个流包对应路径之间的负载，降低某一个或多个流包中因 flowlet 的发送频率较快，导致该一个或多个流包对应的路径负载过重的概率。

S503、发送方基于待发送的数据包所属的流包，发送所述待发送的数据包。

其中，属于不同流包中的数据包的发送路径不同。

相比于现有技术中难以触发足够粒度的动态负载均衡，本申请实施例提供的流量均衡方法，灵活创建或维护流包，并按照目的节点将待发送数据包划入对应流包，如此，可以将同一流包中的数据包通过同一路径发送，不同流包中的数据包通过不同路径发送，即按照本申请实施例的技术方案，能够实现根据数据包所属流包动态的控制路径切换。并且，路径切换的粒度与流包的划分方式相关。其中，当流包包括较少的数据包时，路径切换粒度较细，即每发送较少的数据包就可触发路径切换，当流包包括较多的数据包时，路径切换粒度较粗，即每发送较多的数据包才可以触发路径切换。可见，本申请实施例提供的流量均衡方法，还能够通过控制流包的划分方式灵活控制路径切换粒度，以满足不同应用场景下的流量均衡需求。

如下结合具体例子，并以创建或维护第一流包、创建或维护第二流包为例来说明本申请实施例的技术方案。参见图 6，本申请实施例提供的流量均衡方法包括如下步骤：

S601、发送方获取待发送数据包。

示例性的，待发送数据包如图 7 所示，包括 8 个数据包。

S602、发送方创建或维护第一流包，并将以第一节点为目的节点的数据包划入第一流包。

本申请实施例主要以数据包由同一源节点去往同一目的节点为例，来说明去往同一目的节点采用同一路径或者不同路径来进行流量均衡的技术方案。其中，源节点可以是如图 7~图 10 中的节点 a，目的节点可以是如图 7~图 10 中的节点 b，也就是说，第一节点可以为节点 b。

如上文阐述，除了按照数据包的目的节点为数据包划分所属流包，发送方还可以结合数据包的网络均衡参数创建或维护流包，并将待发送数据包划入对应的流包。

若网络均衡参数设置为数据包之间的时间间隔，初始时，发送方创建的第一流包可以仅包括待发送数据包中的一个数据包。比如，如图 7 所示，发送方获取的待发送

数据包包括 8 个数据包，则初始时，发送方可以将待发送的 8 个数据包中的第一个数据包（即编号为 1 的数据据包）划入第一流包，并创建仅包括该第一个数据包的第一流包。后续，发送方可以维护所创建的第一流包，并更新第一流包包括的数据包数目，比如，如图 7 所示，将与数据包 1 之间的发送时间间隔小于或等于预设间隔的数据包 2 划入第一流包。类似的，将与数据包 2 之间的发送时间间隔小于或等于预设间隔的数据包 3 也划入第一流包。也就是说，发送方检测第一流包中最后一个数据包与下一个待发送数据包之间的发送时间间隔，若下一个待发送数据包与第一流包中最后一个数据包之间的发送时间间隔小于或等于预设间隔，则发送方将该下一个待发送数据包划入第一流包。示例性的，如图 7 所示，待发送数据包中的数据包 1~数据包 5 中每两个数据包之间的发送时间间隔均小于或等于预设间隔，则发送方将数据包 1~数据包 5 均划入第一流包。

若网络均衡参数为数据包的数据量，初始时，发送方创建的第一流包可以包括预设数据量的数据包，也可以包括少于预设数据量的数据包。以每个数据包大小相同，且单个数据包大小为 16KB，设置的预设数据量为 64KB（即 4 个数据包大小）为例，在一个示例中，如图 8 中（a）所示，发送方在某一时刻，获取数据量较大的待发送数据包，假设为 128KB（8 个数据包大小），待发送数据包的数据量大于预设数据量。这种情况下，初始时，发送方可以将 8 个待发送数据包中的 4 个数据包（即图 8 中（a）所示数据包 1~数据包 4）划入第一流包，并创建包括该 4 个数据包的第一流包。该第一流包的数据量恰好为预设数据量。在另外的示例中，如图 8 中（b）所示，发送方在某一时刻，获取数据量较小的待发送数据包，假设为 32KB（2 个数据包大小），待发送数据包的数据量小于预设数据量。这种情况下，初始时，发送方可以将 2 个待发送数据包先划入第一流包，并创建包括 2 个数据包的第一流包。之后，发送发继续获取后续时刻的待发送数据包，可以理解的是，若该后续时刻的待发送数据包的数据量大于 32KB，则发送方可以仅将该后续时刻的的待发送数据包中的 32KB 数据包划入第一流包，该后续时刻的待发送数据包中的其他数据包可以划入后续创建的其他流包中。反之，若该后续时刻的待发送数据包的数据量小于或等于 32KB，则发送方可以将该后续时刻的的待发送数据包全部划入第一流包，更新第一流包中数据包的数目，并继续获取待发送数据包，直至第一流包中的数据包的数据量达到如图 8 中（b）所示的预设数据量。

若网络均衡参数设置为数据包的持续时长，初始时，发送方创建的第一流包可以仅包括待发送数据包中的一个数据包。比如，如图 9 中（a）所示，发送方获取的待发送数据包包括 8 个数据包，则初始时，发送方可以将待发送的 8 个数据包中的第一个数据包（即编号为 1 的数据据包）划入第一流包，并创建仅包括该第一个数据包的第一流包。后续，发送方可以维护所创建的第一流包，并更新第一流包包括的数据包数目。比如，如图 9 中（a）所标注，数据包 2 的发送时刻在预设时长内，则发送方将数据包 2 划入第一流包。类似的，数据包 3 的发送时刻也在预设时长内，则数据包 3 也被划入第一流包。也就是说，自第一流包中第一个数据包的发送时刻开始，在预设预设时长内发送的数据包被划入第一流包。

网络均衡参数还可以为发送周期，作为一种可能的实现方式，一个发送周期定义

为一段预设时间段，发送方周期性创建流包，即每隔一个发送周期，创建一个流包。当网络均衡参数为发送周期时，初始时，发送方创建的第一流包可以仅包括待发送数据包中的一个数据包。比如，如图 9 中（b）所示，发送方获取的待发送数据包包括 8 个数据包，则初始时，发送方可以将待发送的 8 个数据包中的第一个数据包（即编号为 1 的数据据包）划入第一流包，并创建仅包括该第一个数据包的第一流包。后续，发送方可以维护所创建的第一流包，并更新第一流包包括的数据包数目。比如，如图 9 中（b）所标注，t1~t2 为第一流包的发送周期，t2~t3 为第二流包的发送周期，数据包 2 的发送时刻在第一流包的发送周期内，则发送方将数据包 2 划入第一流包。类似的，数据包 3 的发送时刻也在第一流包的发送周期内，则数据包 3 也被划入第一流包。

若网络均衡参数为数据包的发送频率，初始时，发送方创建的第一流包可以包括待发送数据包中的一个数据包。比如，如图 10 所示，发送方获取的待发送数据包包括 8 个数据包，则初始时，发送方可以将待发送的 8 个数据包中的第一个数据包（即编号为 1 的数据据包）划入第一流包，并创建仅包括该第一个数据包的第一流包。后续，发送方可以维护所创建的第一流包，并更新第一流包包括的数据包数目。比如，如图 10 所标注，数据包 2 和数据包 3 与数据包 1 属于同一 flowlet 1，且数据包 2 与数据包 1 之间的发送时间间隔较小（并未达到预设发送间隔），数据 2 与数据包 3 之间的发送时间间隔也较小，说明 flowlet 1 的发送频率较小，则发送方将属于 flowlet 1 的数据包 2 和数据包 3 划入第一流包。类似的，之后发送的 flowlet 2 的发送频率达到预设频率，则属于 flowlet 2 的数据包 4~6 也被划入第一流包。

网络均衡参数还可以为数据包的数据量和数据包之间的时间间隔两者的组合。这种情况下，第一流包中的数据包中每两个数据包之间的发送时间间隔均小于或等于预设间隔，且第一流包的数据量为预设数据量。第一流包的具体创建和维护过程可参见上述网络均衡参数分别为数据包的数据量和时间间隔的创建和维护过程，这里不再赘述。

当然，网络均衡参数还可以是上述多个网络均衡参数中其他两个或两个以上的组合。这些情况下，第一流包的创建和维护过程也可参见上文介绍，这里不再赘述。

S603、发送方判断网络均衡参数是否满足预设条件。

其中，网络均衡参数满足预设条件，指的是如下一项或多项的组合：

所述从发送第一流包中第一数据包的时刻开始的预设间隔内，不存在去往同一目的节点的数据包，其中，第一数据包为第一流包中的某一数据包；所述第一流包的数据量达到预设数据量；所述第一流包的持续时长达到预设时长；所述第一流包的发送频率达到预设频率。当然，如上文所阐述，网络均衡参数满足预设条件，还可以是其他情况，这里不再一一列举。

S604、若网络均衡参数满足预设条件，则发送方创建或维护第二流包，并将之后发送的以第二节点为目标的数据包，划入所述第二流包。

以网络均衡参数设置为数据包之间的时间间隔为例，发送方创建或维护第一流包，若发送方检测到从发送第一流包中某一数据包的时刻开始的预设间隔内，不存在去往同一目的节点的数据包，则创建第二流包，并可以将预设间隔之后发送的数据包划入第二流包。如图 7 所示，发送方已创建和维护有第一流包，第一流包包括数据包 1~数

据包 5，发送方检测到自从数据包 5 的发送时刻开始，预设间隔内不存在去往节点 b 的数据包，则创建第二流包，并将预设间隔之后发送的数据包 6 划入第二流包，数据包 6 为第二流包中的首个数据包。其中，第二流包的具体维护过程可参见网络均衡参数为数据包之间时间间隔时上述第一流包的维护过程。

本申请实施例中，预设间隔 G 的数值可以较现有技术中 flowlet 所要求的 G 小（可以依据经验或采用其他方式确定该数值），以避免实际应用中，过大 G 往往难以触发路径切换的现象。

若网络均衡参数为数据包的数据量，参见图 8 中（a）或图 8 中（b），发送方创建或维护第一流包，当发送方检测到第一流包的数据量达到预设数据量时，则创建第二流包，并将之后发送的数据包划入第二流包。

如此，当第一流包与第二流包的时间间隔较短，即第一流包中最后一个数据包与第二流包中第一个数据包之间的发送时间间隔较短，无法满足时间间隔大于或等于预设间隔（比如 5us）的条件时，可通过每发送预设数据量（比如 256KB）的数据包构建一个流包，提升 DLB 的动态能力。其中，可以根据实际经验和统计信息确定预设间隔在满足什么条件时“较小”，比如，预设间隔小于第一阈值时，则视为预设间隔较小。第一阈值可灵活设置，本申请实施例对此不进行限制。

若网络均衡参数设置为数据包的持续时长，参见图 9 中（a），发送方创建或维护第一流包，当发送方检测到自从发送第一流包中的第一个数据包的时刻开始，预设时长到达，则发送方创建第二流包，并将该预设时长之后发送的数据包划入第二流包。

网络均衡参数还可以为发送周期，参见图 9 中（b），发送方创建或维护第一流包，当发送方检测到发送第一流包的发送周期结束，则发送方创建第二流包，并将第一流包的发送周期之后发送的数据包划入第二流包。

在另一些实施例中，发送方配置还可以配置其他网络均衡参数，并检测参数是否满足预设条件，以判断是否创建第二流包。比如，发送方配置当前路径（比如第一路径）的拥塞程度这一参数，当发送方检测到当前已创建流包中数据包通过链路发送时的拥塞程度大于或等于拥塞阈值，则创建第二流包。

需要说明的是，本申请实施例对发送方配置何种参数、以及配置参数的个数不进行限制。

当然，发送方还可以结合上述两个或多个条件来判断是否进行路径切换。比如，发送方检测到第一流包的数据量较大（即达到预设数据量），比如达到 5 个数据包，且自从发送第一流包的第五个数据包的时刻开始，预设间隔内不存在去往同一目的节点的数据包，才创建新的第二流包。在另一些实施例中，可以配置预设数据量 S 的数值较小，预设间隔 G（比如 50us）的数值较大。这样一来，当某一数据流的数据量很少时，可通过预设间隔 G 触发创建新的第二流包。即，如果两个流包之间通过不小于 50us 的空闲时间隔开。其中，第一流包与第二流包之间的空闲时间，指的是第一流包在这段空闲时间内无数据包，但这段空闲时间内允许发送其它流包的数据包。参见上文描述，关于预设数据量 S 的数值较小，以及预设间隔 G 的数值较大的具体定义，可以根据实际应用和算法确定，比如，设置阈值，当预设数据量 S 小于或等于该阈值时，就视为预设数据量 S 较小。

在一种示例性的参数配置方式中，还可以配置预设间隔 G=infinity，预设数据量 S=infinity，预设时长 T=infinity。这意味着，发送方始终保持 SLB，即，不创建新的第二流包，将属于同一数据流的数据包均属于第一流包。

S605、发送方基于待发送数据包所属流包，发送待发送数据包。

其中，所述第二流包中的数据包的发送路径，与所述第一流包中的数据包的发送路径不同。

需要说明的是，第一流包中的一个或多个数据包可以来自同一数据流，也可以来自不同数据流。即第一流包中的数据包可以为某一数据流中的部分数据包，也可以为多个数据流中的数据包（当然，要求在同一流包中的这些数据包必须要去往相同的目标设备）。类似的，第二流包中的一个或多个数据包可以来自同一数据流，也可以来自不同数据流。

第一流包和第二流包可以为同一数据流中的数据包。第一流包和第二流包也可以来自不同数据流。比如，所述第一流包为数据流 1 中的数据包，且所述第二流包为数据流 2 中的数据包；或者，所述第一流包为数据流 2 中的数据包，且所述第二流包为数据流 1 中的数据包；或者，所述第一流包包括所述数据流 1 中的数据包和所述数据流 2 中的数据包；或者，所述第二流包包括所述数据流 1 中的数据包和所述数据流 2 中的数据包。

可见，通过设置不同参数（比如设置时间间隔 Gap 或者持续时长 Time）可以实现不同性能的 DLB。或者，通过相同参数设置不同的数值，也可以实现不同性能的 DLB。并且，设置的这些参数可以分别，或者联合作用。示例性的，仅设置时间间隔这一参数时，发送周期设置为 5us 时，每隔 5us 可能就会触发一次路径切换，DLB 的性能较高。相应的，接收端可能需较多排序资源。当发送周期设置为 100us 时，每隔 100us 可能才触发一次路径切换，DLB 性能较低。相应的，接收端所需的排序资源可能较少。

在另一些实施例中，可以设置多套参数。一套参数指的是一个或多个参数，比如，第一套参数包括时间间隔、持续时长。第二套参数包括时间间隔、数据量。同一数据流使用一套参数。不同的数据流可以采用不同的参数。比如，数据流 1 使用第一套参数进行流量均衡，数据流 2 使用第二套参数。当然，不同数据流也可以使用同一套参数，本申请实施例对此不进行限制。

另外，实际应用中，不同的数据流可以基于不同参数来实现流量均衡，以满足业务差异化的要求。比如，针对数据流 1，发送方检测数据包之间的时间间隔是否满足上述预设间隔，从而为该数据流 1 创建一个或多个流包，以进行流量均衡。针对数据流 2，发送方检测数据包的数据量这一参数，每发送 256KB 数据，就切换一次路径，从而动态为数据流 2 中的数据包选择不同路径。

在一些实施例中，发送方需区分不同的流包。即，对不同流包进行定界。定界指的是，明确区分出数据包属于哪一流包。后续，接收方可以基于该定界对数据包执行重排序（re-ordering）。示例性的，以第一流包和第二流包为例，发送方可以采用如下任一种方式区分第一流包和第二流包。

1、发送方通过重新封装第一流包或第二流包的数据包，来区分第一流包和第二流包。比如，可以在数据包新增一些标签字段，并为字段设置一定数值来标识数据包属

于哪一流包。或者，直接复用现有的字段，并为字段设置新的数值来标识数据包。

相应的，接收方解封装数据包，从数据包中的相应字段获知该数据包属于哪一流包。接收方还可以根据所述定界标识判断是否接收到所述第一流包的尾包。

具体的，发送方封装数据包的具体实现方式可以为：所述第一流包中至少一个数据包携带定界标识，所述第二流包中至少一个数据包携带定界标识，所述定界标识用于区分所述第一流包和所述第二流包。

可选的，定界标识可以为流包序列号，用以显式指示数据包所归属的流包。同一流包中的数据包均携带相同的流包序列号(flowpacsequence，FSN)。比如 flowpac 0 内的所有数据包的 FSN 字段的值均为 0，flowpac 1 内的所有包的 FSN 字段的值均为 1，依此类推。其中，FSN 位宽（即所需的比特位数）需使得接收侧能区分经不同路径接收到的不同 flowpac，比如可以为 4bit。

示例性的，如果流包只携带同一 flow 的数据。接收方在接收到当前数据包之后，若解封装该当前数据包后，首次发现 FSN 字段的数值为 3，则接收方判断该当前数据包为流包 3 的首包(first packet)。

在一些实施例中，接收方一旦接收到某一流包中携带尾包标识的尾包，则将尾包信息存储在排序流表（re-ordering table，ROT）中。

后续，接收方一旦接收新的数据包，其可以查询该排序流表，若能够查询到上一流包的尾包信息，说明接收方已接收到上一流包的尾包，无需对新数据包所属 flowpac 进行重排序。比如，接收方接收到 FSN=4 的数据包，其查询排序流表发现已接收到 FSN=3 的尾包，由于流包 3、4 是按照先发先至、后发后至顺序接收的，则接收方无需执行重排序。

反之，参见图 11，接收方在接收到新流包（比如第二流包）之后，若接收方通过查询排序流表确定未接收到所述上一流包，即第一流包的尾包，则说明第一流包、第二流包并未按照先发先至、后发后至的顺序达到接收方，导致乱序。因此，接收方通过 RC 执行重排序。作为一种可能的实现方式，接收方在接收到所述第一流包的尾包后，释放所述 RC。容易理解的是，接收方接收到第一流包的尾包，说明第一流包已接收完毕，这样一来，接收方将第二流包排序在第一流包之后。如此，第一流包和第二流包之间不再存在乱序问题，可以释放该 RC，并将释放的 RC 用于其他重排序流程，从而提升 RC 的利用率。由于闲置 RC 还可以用于其他重排序流程，相当于拓展了可用 RC 的数目。

在一些实施例中，接收方在发生路径切换的情况下才需占用 RC 执行重排序。也就是说，如果两条路径的时延差最多为 20us，则发生路径切换的流包只需占用 RC 大约 20us 的时间，即保证从原路径上收到先发后至的数据后（最多只需要 20us），即可释放 RC。相比于 packetspray 机制中，在整个数据流的生命周期内（可以达到 ms 甚至 s 级）均需占用 RC，本申请实施例中的重排序流程能够降低 RC 消耗。

更进一步的，接收方的上述重排序流程中，仅在产生乱序的情况下，即在接收新流包时，还未完成上一流包的接收，才启用 RC 资源，并在确定无乱序问题后，立即释放 RC，RC 的排序周期缩短，能够降低 RC 的消耗，提升 RC 的利用率。其中，排序周期可以指接收方确定产生乱序的时刻至释放 RC 的时刻。当然，也可以指接收方

确定存在通过切换路径的流包的时刻至释放 RC 的时刻。本申请实施例对排序周期具体为哪一时间段不进行限制。

采用本申请实施例提供的流量均衡方法，发送侧可通过不同的参数，如 Gap、Size、Time，来控制 DLB 的动态能力，比如，可以控制发生路径切换的流包的数目。进一步，基于发生路径切换的流包，还能控制所需 RC 数量，和/或 RC 的排序周期。并且，参数数值不同，和/或，参数类型（比如参数类型为 Time，或者 Gap）不同时，负载均衡的性能可能不同。举例来说，当数据量的数值较大时，每发送较大的数据量，才能出发一次路径切换，DLB 的动态性能较差，但是，接收方所需的 RC 较少，RC 消耗少。

在一些实施例中，可以基于接收方可用 RC 的数目来选择不同的排序策略：

1、若可用 RC 的数目大于或等于 RC 数目阈值，则通过不同 RC 重排序不同数据流的数据包。即每一数据流可以对应一个 RC。

2、若可用 RC 的数目小于 RC 数目阈值，则通过相同 RC 重排序不同数据流的数据包。即一条数据流（比如包括上述第一流包和第二流包的数据流）或多条数据流共用一路 RC。

在一些实施例中，可使用如下可动态分配 RC(dynamic re-ordering channel，DRC)算法，用于进一步控制 RC 的数量。

可预设门限 N1，比如 N1=512。

当可用的 RC 数量大于或等于 N1，则待排序的数据流都可独占一个 RC。

否则，当可用的 RC 数量小于 N1，则一条数据流（比如包括上述第一流包和第二流包的数据流）或多条数据流共用同一 RC。

采用上述 DRC 算法，可以用有限的 RC 资源来保证大部分流或全部流都不出现乱序。另外，接收方也可以将可用 RC 资源反馈给发送方，以保证发送的 flowpac 都有对应的 RC 可用。

可选的，用于区分第一流包和第二流包的定界标识可以为首包中的某些预设标识。即发送方只在每个 flowpac 的首包携带预设标识。比如，将首包的切换完成(SwitchOver，SO)字段的值设置为 1，用以表示该流包为新流包的首包，还用于表示该流包与上一流包采用不同路径发送，即该流包相对于上一流包发生了一次路径切换。或者，首包携带其他形式的首包标识，比如首包（first packet，FP）字段值为 1。或者，首包携带包类型（packettype）字段，例如以 packettype =0 表示首包。

相应的，接收方在接收到当前数据包之后，若解封装该当前数据包后，该当前数据包的 SO 字段的数值为 1，且接收方已经接收到第一流包的一个或多个数据包，则接收方判断该当前数据包为第二流包的首包。

与上述类似，接收方可以通过查询排序流表判断是否执行重排序。比如，接收方接收到 SO=1 的当前数据包，其查询排序流表发现已接收到上一流包的尾包，由于上一流包和当前流包是按照先发先至、后发后至顺序接收的，则接收方无需执行重排序。

如此，仅通过首包就可以区分出当前数据包属于哪一 flowpac，能够降低控制信息的开销，从而提升有效的数据信息的传输效率。

可选的，定界标识可以为尾包中的某些预设标识。即发送方只在每个 flowpac 的尾包携带尾包标识，或者，携带包类型标识，包类型标识用于指示数据包为尾包。比

如，以 packettype =3 表示尾包。

可选的，packettype 的其它值可以表示 flowpac 内首、尾包之间的包。

可选的，定界标识可以为包序号（packet sequence，PSN）。其中，所述第二流包的首包携带的 PSN 为 M，M 为正整数，所述第二流包的第 i 个数据包携带的 PSN 为 i-1，i 为大于 1 的整数，M 为所述第一流包中数据包数目与 1 的差值。

比如，flowpac 3 包括 100 个数据包，则首包中 PSN=99，第二个数据包中 PSN=1，第三个数据包中 PSN=2，第四个数据包中 PSN=3，依次类推，第 100 个数据包中 PSN=99。如此，当接收方接收到具有相同 PSN 的数据包时，比如，接收方先接收到 flowpac 3 中的首包（PSN=99），之后，再次接收到 PSN 同样为 99 的数据包时，接收方可判断该数据包为 flowpac 3 的尾包。之后，接收方可准备接收下一流包。

可选的，定界标识还可以为第二流包首包携带的所述第一流包中尾包的特征值，所述特征值包括循环冗余校验（cyclic redundancy check，CRC）校验值。即在当前流包的首包中携带上一流包的尾包的一些信息。

当然，发送方可以将上述多个定界标识结合，比如，发送方在数据包中携带首包标识和尾包标识。

2、可以在两个 flowpac 之间插入专用的定界包来指示第一流包的尾包是否已收到。

具体的，当确定开始发送新的流包，并且需切换路径，发送方先通过原路径，即第一路径发送定界包（flowpac delimiter，FPD），再通过新路径，即第二路径发新流包，即第二流包。

所述定界包用于区分所述第一流包和所述第二流包。定界包为通过第一路径接收，且为第一流包的尾包的下一数据包。定界包可以为具有预设特征的控制包。比如，可以是包含预设字段的控制包，或者预设大小的控制包。本申请实施例对定界包的具体实现不进行限制。

可选的，定界包的大小较小（比如可以小于某一阈值），以使得定界包能够以较低时延到达接收方。比如，使得定界包比第二流包早些到达接收方。当然，定界包也可以比第二流包晚些到达接收方，本申请实施例对定界包到达接收方的时机并不进行限制。

相应的，接收方当前正在通过第一路径接收第一流包，后续，接收方还通过第一路径接收定界包，则接收方可确定第一流包已接收完毕。在一种示例中，接收方在接收定界包之后（第一流包已接收完毕），才接收到第二流包，第一流包和第二流包通常不存在乱序问题。进而，接收方通常无需重排序，能够降低 RC 消耗。

需要说明的是，本申请实施例中所提及的字段名称均为示例性的名称，在实际实现时还可以为其他名称，本申请实施例对此不进行限制。

上述主要描述对同一数据流进行流量均衡，以及该数据流中数据包的定界、RC、DRC 方式。

在另一些实施例中，还可以将多条数据流可以汇聚（也称聚合（merge））为一条粗粒度的流（比如通过某种 hash 算法）。示例性的，用预设算法（比如 hash 算法，或其他类似算法）对待排序的数据流做聚合。该汇聚形成的粗粒度的流在本文中可称为汇聚流。在本申请实施例中，可以针对汇聚流执行上述流量均衡方法，该汇聚流中

数据包的定界、RC、DRC 方式的具体实现方式可参考上文。

基于此，本文内的“数据流”，即可以指如上文所提及的通常意义的 TCP 流，也可以指汇聚流。

作为一种可能的实现方式，汇聚流可以共用同一 RC。其中，共用同一 RC 的流之间可能出现头阻塞（head of line blocking，HOL）。因此，接收端只有确定已接收到上一 flowpac 对应的尾包，才处理下一 flowpac。

上述主要从不同网元之间交互的角度对本申请实施例提供的方案进行了介绍。可以理解的是，发送节点和接收节点为了实现上述功能，其包含了执行各个功能相应的硬件结构和/或软件模块。结合本申请中所公开的实施例描述的各示例的单元及算法步骤，本申请实施例能够以硬件或硬件和计算机软件的结合形式来实现。某个功能究竟以硬件还是计算机软件驱动硬件的方式来执行，取决于技术方案的特定应用和设计约束条件。本领域技术人员可以对每个特定的应用来使用不同的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实施例的技术方案的范围。

本申请实施例可以根据上述方法示例对发送节点和接收节点等进行功能单元的划分，例如，可以对应各个功能划分各个功能单元，也可以将两个或两个以上的功能集成在一个处理单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。需要说明的是，本申请实施例中对单元的划分是示意性的，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。

图 12 示出了本申请实施例中所涉及的一种流量均衡装置的一种可能的示例性框图，该装置 1200 可以以软件的形式存在，也可以为网络设备，还可以为可以用于网络设备的芯片。装置 1200 包括：处理单元 1202 和通信单元 1203。处理单元 1202 用于对装置 1200 的动作进行控制管理，例如，若该装置用于实现发送节点功能，处理单元 1202 用于支持装置 1200 执行图 5 中的 S502，图 6 中的 S602、S603、S604，和/或用于本文所描述的技术的其它过程。若该装置用于实现接收节点功能，处理单元 1202 用于支持装置 1200 判断在接收到第二流包的情况下，是否已接收到第一流包的尾包，若未接收到第一流包的尾包，则通过排序通道 RC 对第一流包和第二流包中的数据包执行重排序，并在重排序之后，若确定已接收到第一流包的尾包，则释放所述 RC，和/或用于本文所描述的技术的其它过程。通信单元 1203 用于支持装置 1200 与其他网络实体的通信。装置 1200 还可以包括存储单元 1201，用于存储装置 1200 的程序代码和数据。

其中，处理单元 1202 可以是处理器或控制器，例如可以是 CPU，通用处理器，DSP，ASIC，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。其可以实现或执行结合本申请公开内容所描述的各种示例性的逻辑方框，模块和电路。所述处理器也可以是实现计算功能的组合，例如包含一个或多个微处理器组合，DSP 和微处理器的组合等等。通信单元 1203 可以是通信接口、收发器或收发电路等，其中，该通信接口是统称，在具体实现中，该通信接口可以包括多个接口，例如可以包括：发送节点和接收节点之间的接口和/或其他接口。存储单元 1201 可以是存储器，或者其他形式的存储设备。

当处理单元 1202 为处理器，通信单元 1203 为通信接口，存储单元 1201 为存储器

时，本申请实施例所涉及的装置 1200 可以为具有图 13 所示结构的装置。

参阅图 13 所示，该装置 1300 包括：处理器 1302、通信接口 1303、存储器 1301。可选的，装置 1300 还可以包括总线 1304。其中，通信接口 1303、处理器 1302 以及存储器 1301 可以通过总线 1304 相互连接；总线 1304 可以是外设部件互连标准（Peripheral Component Interconnect，简称 PCI）总线或扩展工业标准结构（Extended Industry Standard Architecture，简称 EISA）总线等。所述总线 1304 可以分为地址总线、数据总线、控制总线等。为便于表示，图 13 中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。

作为另一种可能的实现方式，在采用对应各个功能划分各个功能模块的情况下，若流量均衡装置用于实现发送节点功能，图 14 示出了上述实施例中所涉及的实现发送节点功能的装置的另一种可能的结构示意图。流量均衡装置 1400 可以包括：第一模块 1401、第二模块 1402 和第三模块 1403。第一模块 1401 用于支持流量均衡装置 1400 执行图 5 中的 S501、图 6 中的 S601，和/或用于本文所描述的方案的其它过程。第二模块 1402 用于支持流量均衡装置 1400 执行图 5 中的过程 S502，图 6 中的 S602~S604，还用于基于网络均衡的原则为流包中的数据包设定发送路径，和/或用于本文所描述的方案的其它过程。第三模块 1403 用于支持流量均衡装置 1400 执行图 5 中的过程 S503，图 6 中的 S605，还用于通过所述第一路径发送定界包，所述定界包为用于区分所述第一流包和所述第二流包的控制包，所述第一路径为发送第一流包中数据包的路径，和/或用于本文所描述的方案的其它过程。其中，上述方法实施例涉及的各步骤的所有相关内容均可以援引到对应功能模块的功能描述，在此不再赘述。当然，为了实现本申请实施例的技术方案，流量均衡装置还可能包括其他模块，这里就不再赘述。

若流量均衡装置用于实现接收节点功能，图 15 示出了上述实施例中所涉及的实现接收节点功能的装置的另一种可能的结构示意图。流量均衡装置 1500 可以包括：第四模块 1501、第五模块 1502 和第六模块 1503。第四模块 1501 用于支持流量均衡装置 1500 判断在接收到第二流包的情况下，是否已接收到第一流包的尾包，和/或用于本文所描述的方案的其它过程。第五模块 1502 用于支持流量均衡装置 1500 在未接收到第一流包的尾包的情况下，通过排序通道 RC 对第一流包和第二流包中的数据包执行重排序，和/或用于本文所描述的方案的其它过程。第六模块 1503 用于支持流量均衡装置 1500 在重排序之后，若确定已接收到第一流包的尾包，则释放所述 RC，和/或用于本文所描述的方案的其它过程。

本领域普通技术人员可以理解：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实施例所述的流程或功能。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（Digital

Subscriber Line，DSL））或无线（例如红外、无线、微波等）方式向另一个网站站点、计算机、服务器或数据中心进行传输。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括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，数字视频光盘（Digital Video Disc，DVD）)、或者半导体介质（例如固态硬盘(Solid State Disk，SSD)）等。

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设备（例如终端）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，也可以是各个功能单元独立存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能单元的形式实现。

通过以上的实施方式的描述，所属领域的技术人员可以清楚地了解到本申请可借助软件加必需的通用硬件的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。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可读取的存储介质中，如计算机的软盘，硬盘或光盘等，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的方法。

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，在本申请揭露的技术范围内的变化或替换，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 书

1、一种网络设备，其特征在于，包括：

第一模块，用于获取待发送的数据包；

第二模块，用于创建或维护流包，并将待发送的数据包按照目的节点划入对应的流包；

第三模块，用于基于待发送的数据包所属的流包，发送所述待发送的数据包，

其中，一个流包包括至少一个数据包，属于同一流包中的数据包的目的节点相同，且属于同一流包中的数据包的发送路径相同。

2、根据权利要求 1 所述的网络设备，其特征在于，所述第二模块还用于：

创建或维护第一流包；

将以第一节点为目标的数据包，划入所述第一流包；

创建或维护第二流包；

将之后发送的以第一节点为目标的数据包，划入所述第二流包，

其中，所述第二流包中的数据包的发送路径，与所述第一流包中的数据包的发送路径不同。

3、根据权利要求 2 所述的网络设备，其特征在于，所述第二模块，还用于：

判断网络均衡参数是否满足预设条件；

若网络均衡参数满足预设条件，则创建或维护所述第二流包；

所述网络均衡参数用于基于网络均衡原则将数据包划入对应的流包；

所述网络均衡参数满足预设条件包括：

从发送第一流包中第一数据包的时刻开始的预设间隔内，不存在去往同一目的节点的数据包；

所述第一流包的数据量达到预设数据量；

所述第一流包的持续时长达到预设时长；

所述第一流包的发送频率达到预设频率。

4、根据权利要求 2 或 3 所述的网络设备，其特征在于，所述第一流包中至少一个数据包携带第一定界标识，所述第二流包中至少一个数据包携带第二定界标识，所述第一定界标识和所述第二定界标识用于区分所述第一流包和所述第二流包。

5、根据权利要求 2 至 4 中任一项所述的网络设备，其特征在于，所述第三模块，还用于通过第一路径发送定界包，所述定界包为用于区分所述第一流包和所述第二流包的控制包，所述第一路径为发送第一流包中数据包的路径。

6、根据权利要求 1 至 5 中任一项所述的网络设备，其特征在于，一个流包内的数据包来自相同或不同数据流。

7、根据权利要求 1 至 6 中任一项所述的网络设备，其特征在于，所述第二模块，还用于基于网络均衡的原则为流包中的数据包设定发送路径。

8、一种网络设备，其特征在于，包括：

第四模块，用于判断在接收到第二流包的情况下，是否已接收到第一流包的尾包；

第五模块，用于若未接收到第一流包的尾包，则通过排序通道 RC 对第一流包和第二流包中的数据包执行重排序；

第六模块，用于在重排序之后，若确定已接收到第一流包的尾包，则释放所述 RC。

9、一种电子设备，其特征在于，包括处理器和存储设备，所述存储设备用于存储指令，所述处理器用于基于所述指令执行下列动作：

创建或维护第一流包；

将以第一节点为目标的数据包，划入所述第一流包；

在网络均衡参数满足预设条件时，创建或维护第二流包；

将之后来的以第一节点为目标的数据包，划入所述第二流包，

其中，属于同一流包中的数据包的目的节点相同，属于同一流包中的数据包的发送路径相同；所述第二流包中的数据包的发送路径，与所述第一流包中的数据包的发送路径不同。

10、一种电子设备，其特征在于，包括处理器和存储设备，所述存储设备用于存储指令，所述处理器用于基于所述指令执行下列动作：

判断在接收到第二流包的情况下，是否已接收到第一流包的尾包；

若未接收到第一流包的尾包，则通过排序通道 RC 对第一流包和第二流包中的数据包执行重排序；

在重排序之后，若确定已接收到第一流包的尾包，则释放所述 RC。

11、一种流量均衡方法，其特征在于，包括：

创建或维护第一流包；

将以第一节点为目标的数据包，划入所述第一流包；

在网络均衡参数满足预设条件时，创建或维护第二流包；

将之后来的以第一节点为目标的数据包，划入所述第二流包，

其中，属于同一流包中的数据包的目的节点相同，属于同一流包中的数据包的发送路径相同；所述第二流包中的数据包的发送路径，与所述第一流包中的数据包的发送路径不同。

12、一种流量均衡方法，其特征在于，包括：

判断在接收到第二流包的情况下，是否已接收到第一流包的尾包；

若未接收到第一流包的尾包，则通过排序通道 RC 对第一流包和第二流包中的数据包执行重排序；

在重排序之后，若确定已接收到第一流包的尾包，则释放所述 RC。

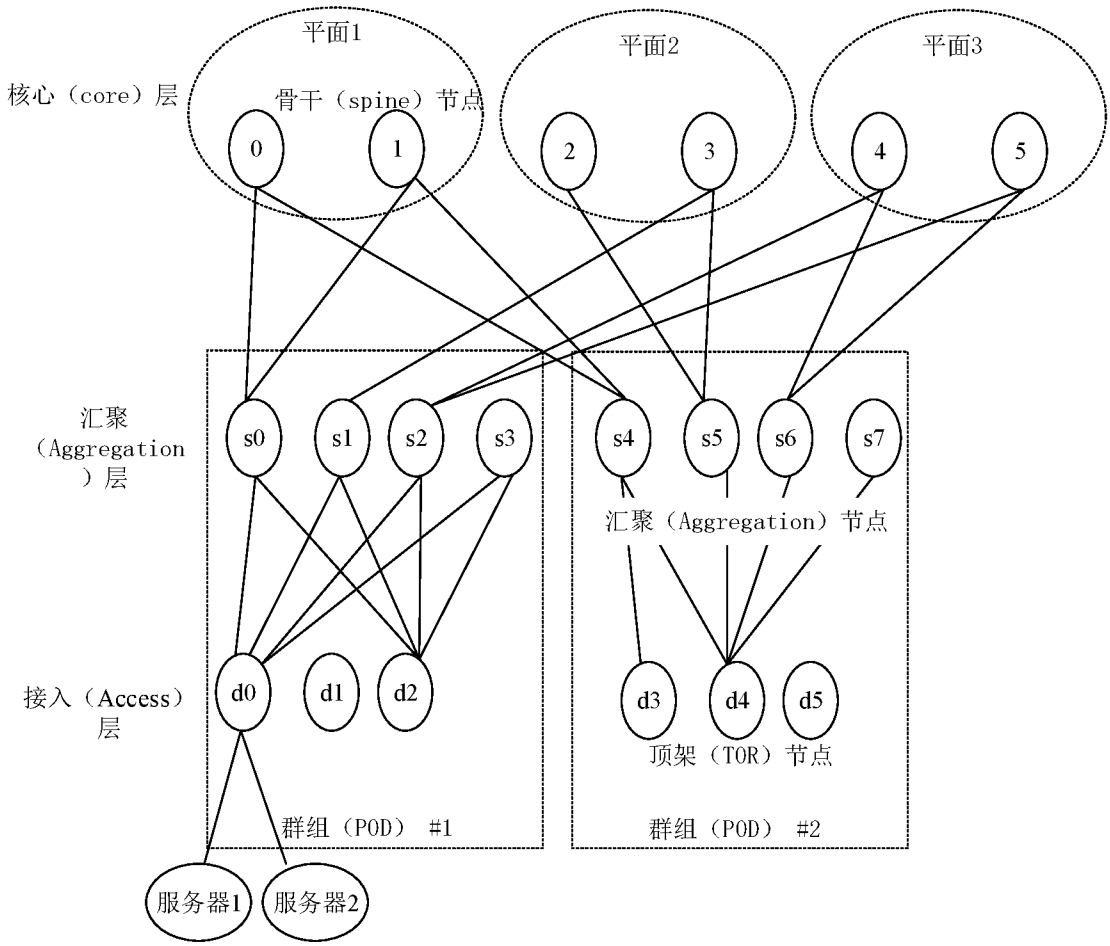

平面1
平面2
平面3
核心（core）层
骨干（spine）节点
0
1
2
3
4
5
汇聚（Aggregation）层
s0
s1
s2
s3
s4
s5
s6
s7
汇聚（Aggregation）节点
接入（Access）层
d0
d1
d2
d3
d4
d5
顶架（TOR）节点
群组（POD） #1
群组（POD） #2
服务器1
服务器2

图 1

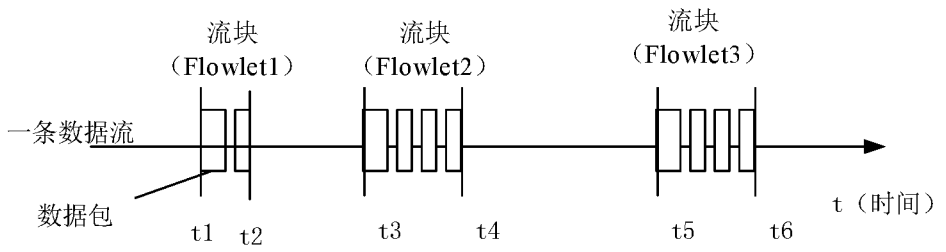

流块
（Flowlet1）
流块
（Flowlet2）
流块
（Flowlet3）
一条数据流
数据包
t1
t2
t3
t4
t5
t6
t（时间）

图 2

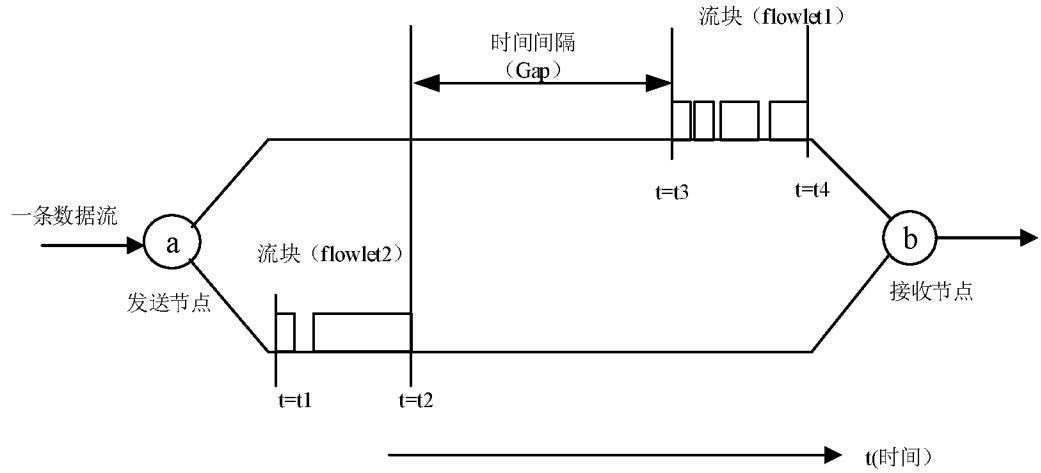

流块（flowlet1）
时间间隔
（Gap）
一条数据流
a
流块（flowlet2）
发送节点
t=t1
t=t2
t=t3
t=t4
b
接收节点
t(时间)

图 3

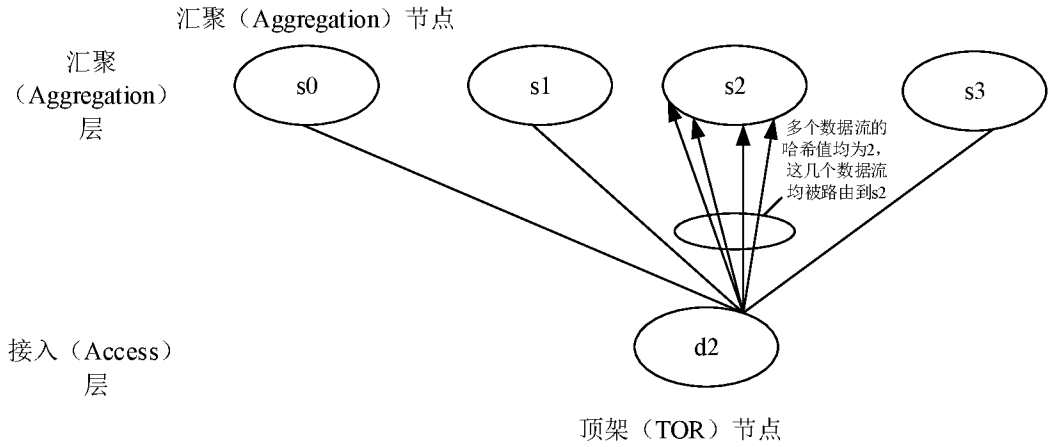

汇聚（Aggregation）节点
汇聚
（Aggregation）
层
s0
s1
s2
s3
多个数据流的
哈希值均为2，
这几个数据流
均被路由到s2
接入（Access）
层
d2
顶架（TOR）节点

图 4

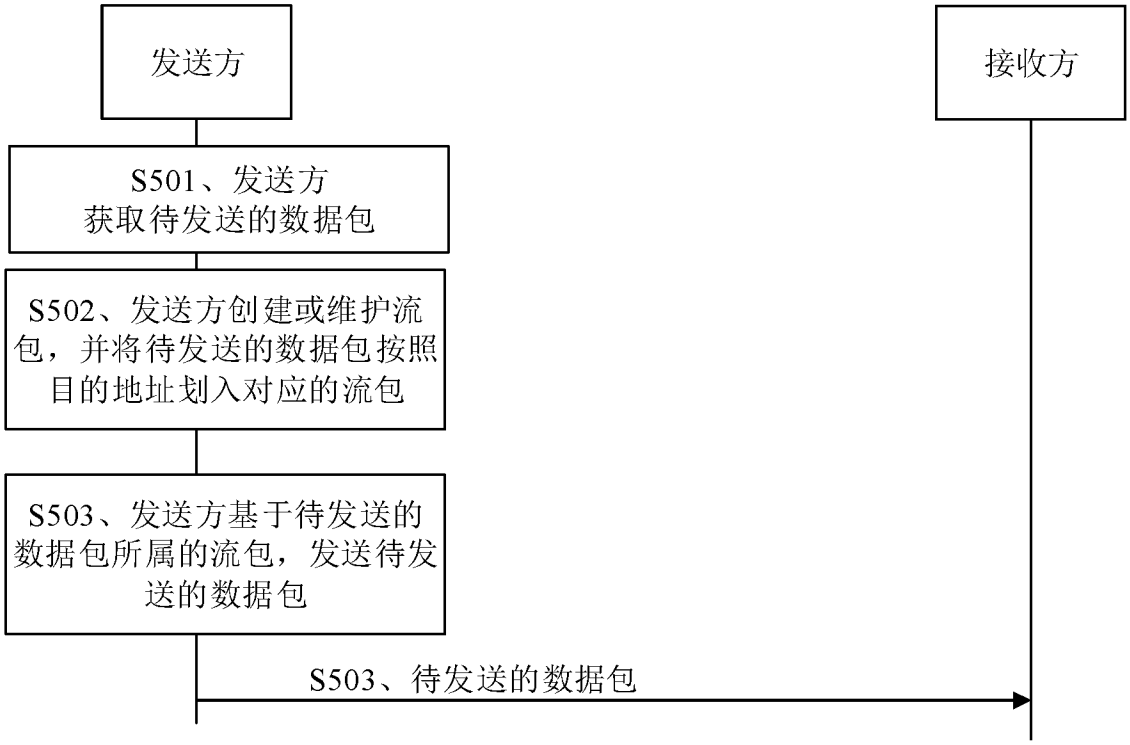

发送方
接收方
S501、发送方
获取待发送的数据包
S502、发送方创建或维护流包，并将待发送的数据包按照目的地址划入对应的流包
S503、发送方基于待发送的数据包所属的流包，发送待发送的数据包
S503、待发送的数据包

图 5

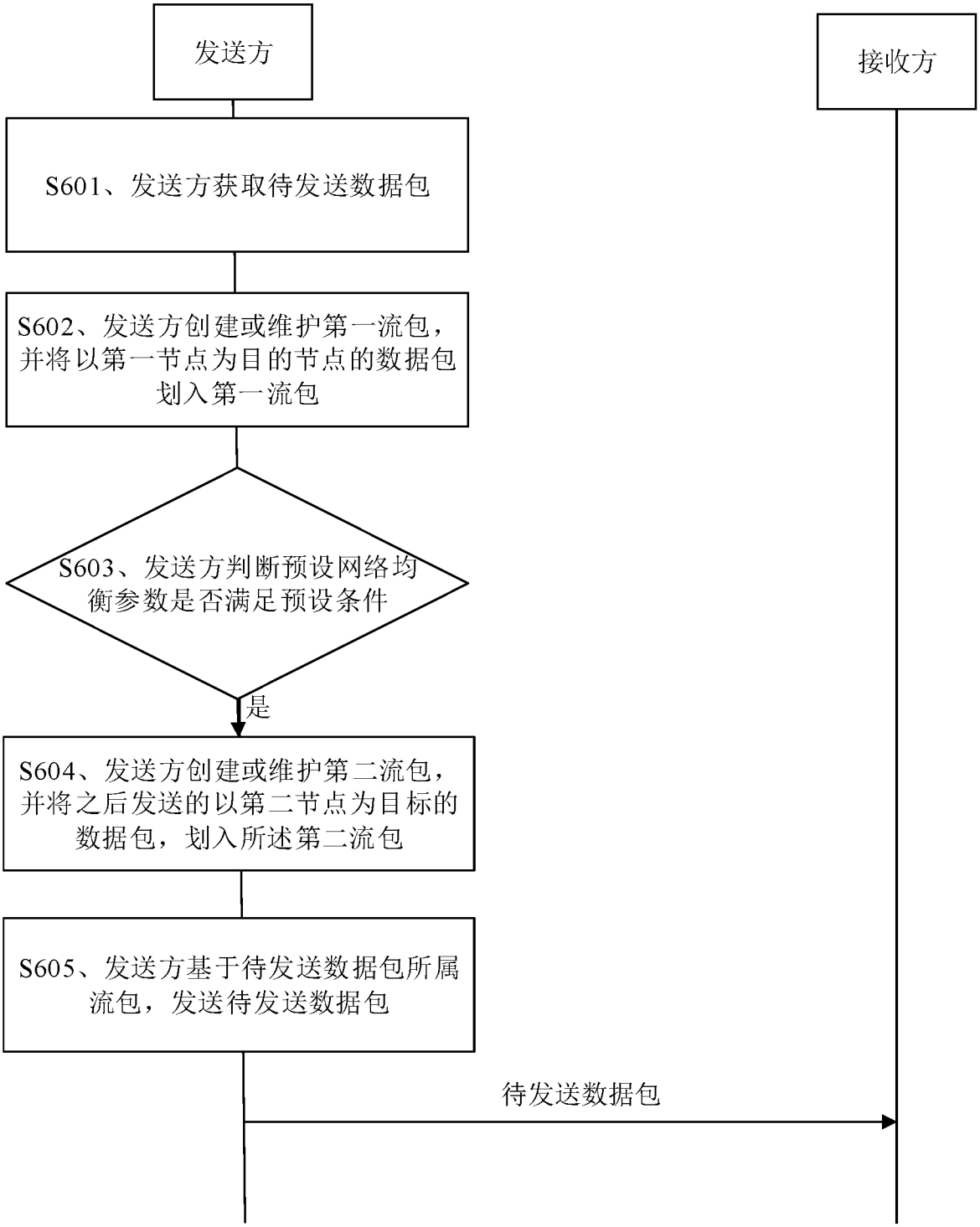

发送方
接收方
S601、发送方获取待发送数据包
S602、发送方创建或维护第一流包，并将以第一节点为目的节点的数据包划入第一流包
S603、发送方判断预设网络均衡参数是否满足预设条件
是
S604、发送方创建或维护第二流包，并将之后发送的以第二节点为目标的数据包，划入所述第二流包
S605、发送方基于待发送数据包所属流包，发送待发送数据包
待发送数据包

图 6

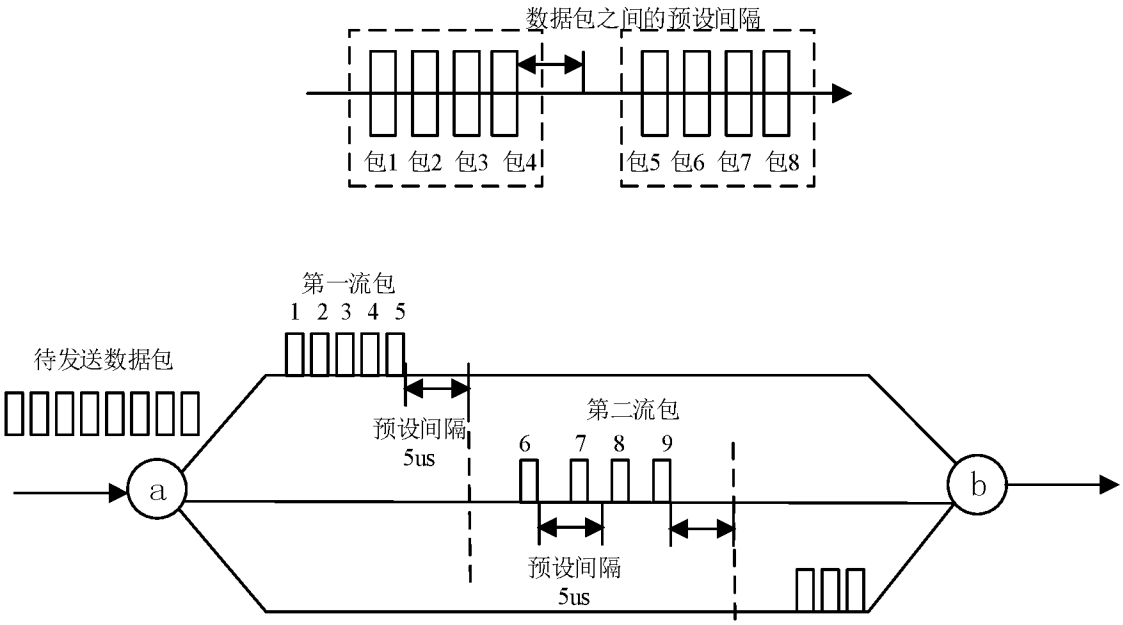
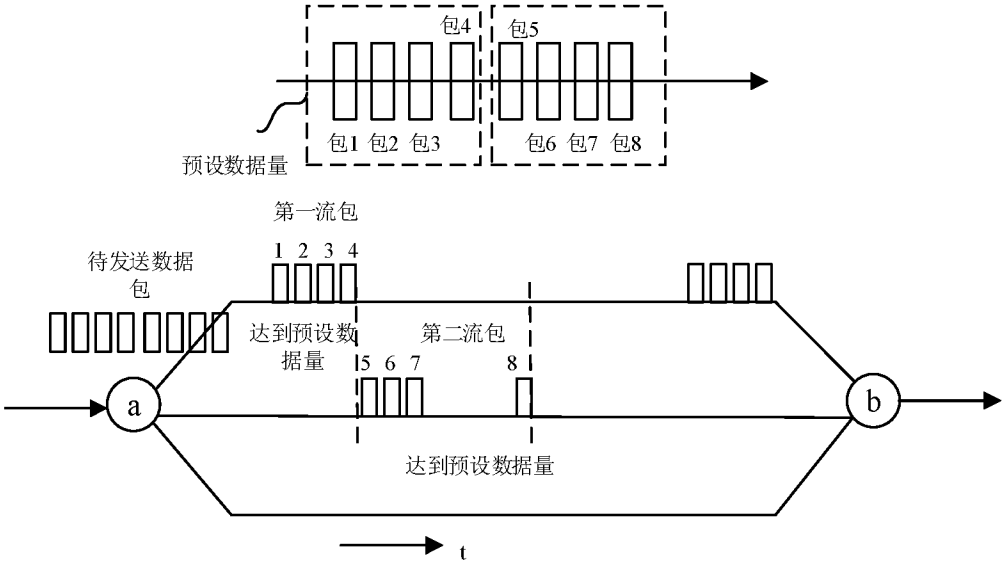
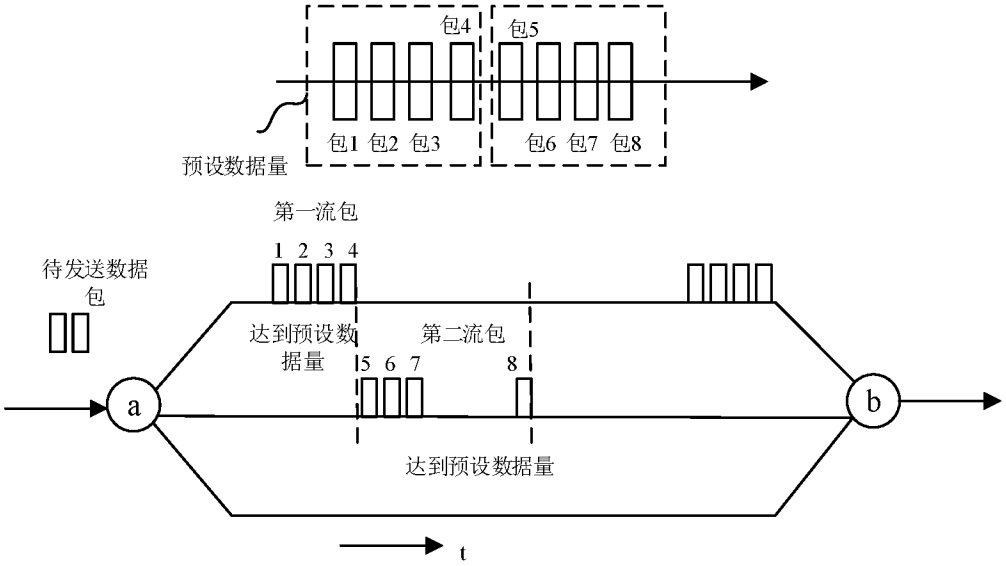

数据包之间的预设间隔
包1 包2 包3 包4
包5 包6 包7 包8
第一流包
1 2 3 4 5
待发送数据包
预设间隔
5us
第二流包
6 7 8 9
a
b
预设间隔
5us

图 7


包4
包5
预设数据量
包1 包2 包3
包6 包7 包8
第一流包
1 2 3 4
待发送数据包
达到预设数据量
第二流包
5 6 7
8
a
b
达到预设数据量
t

（a）


包4
包5
预设数据量
包1 包2 包3
包6 包7 包8
第一流包
1 2 3 4
待发送数据包
达到预设数据量
第二流包
5 6 7
8
a
b
达到预设数据量
t

（b）

图 8

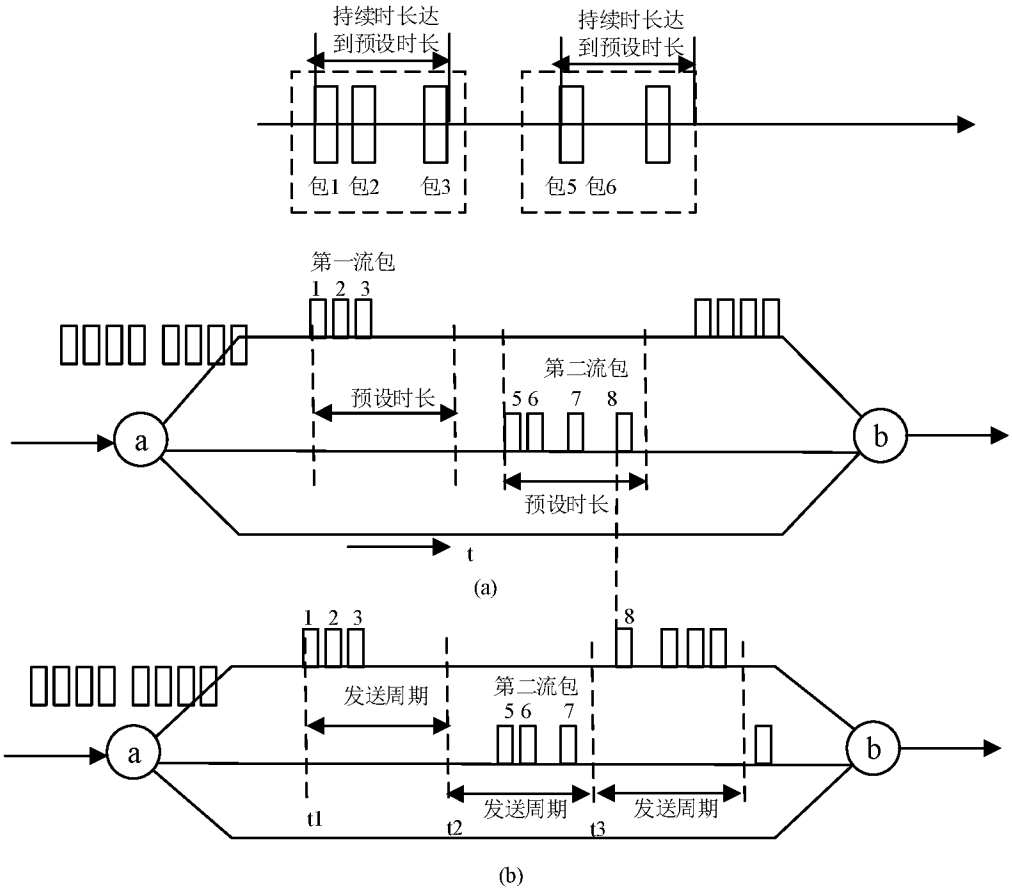

持续时长达
到预设时长
持续时长达
到预设时长
包1 包2
包3
包5 包6
第一流包
1 2 3
预设时长
第二流包
5 6 7 8
a
b
预设时长
t
(a)
1 2 3
8
发送周期
第二流包
5 6 7
a
b
t1
t2
t3
发送周期
发送周期
(b)

图 9

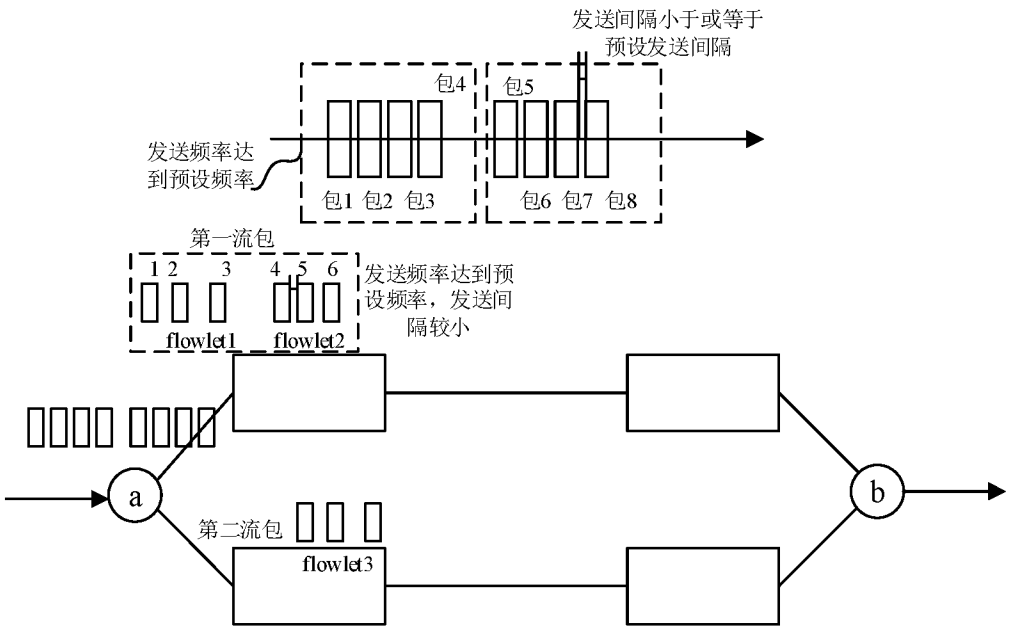

发送间隔小于或等于
预设发送间隔
包4
包5
发送频率达
到预设频率
包1 包2 包3
包6 包7 包8
第一流包
1 2 3 4 5 6
发送频率达到预
设频率，发送间
隔较小
flowlet1
flowlet2
a
b
第二流包
flowlet3

图 10

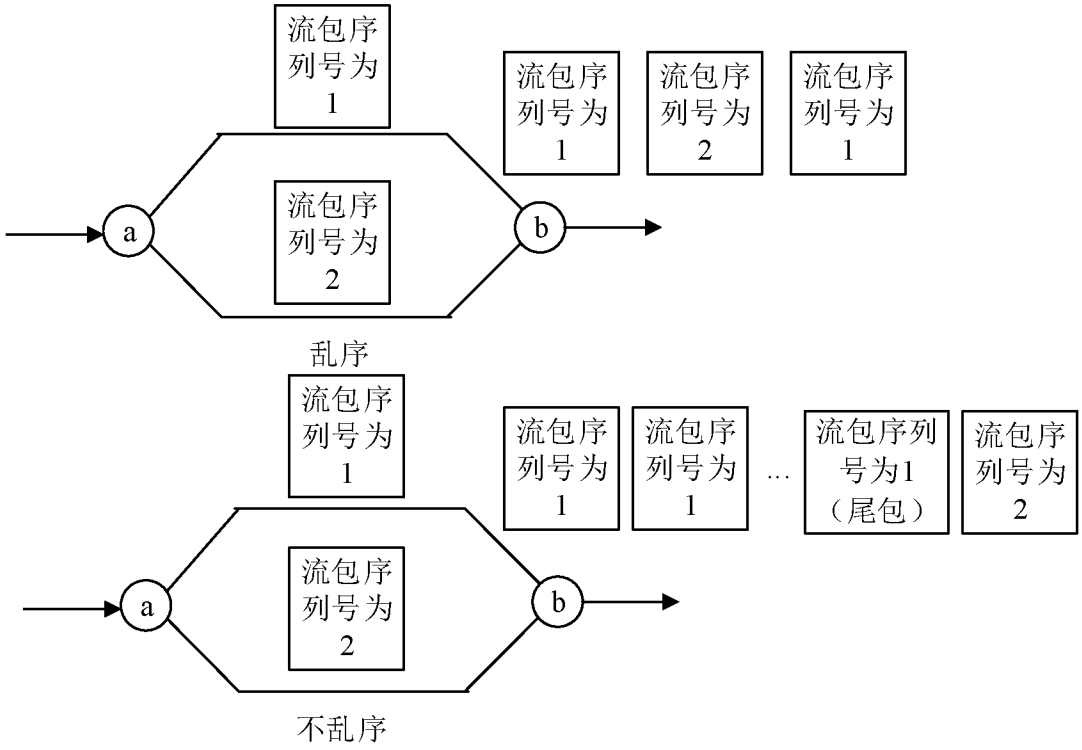

流包序列号为1
流包序列号为2
a
b
流包序列号为1
流包序列号为2
流包序列号为1
乱序
流包序列号为1
流包序列号为2
a
b
流包序列号为1
流包序列号为1
…
流包序列号为1（尾包）
流包序列号为2
不乱序

图 11

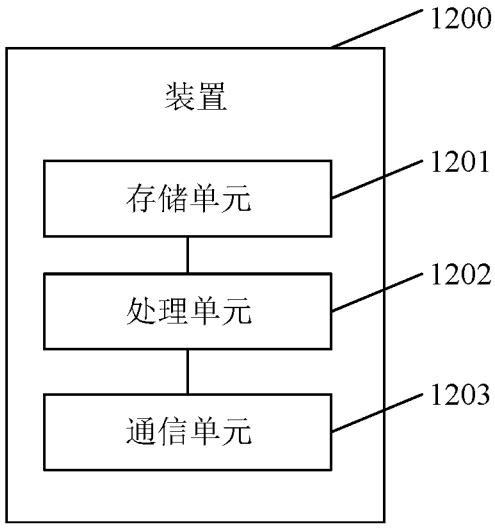

1200
装置
存储单元
1201
处理单元
1202
通信单元
1203

图 12

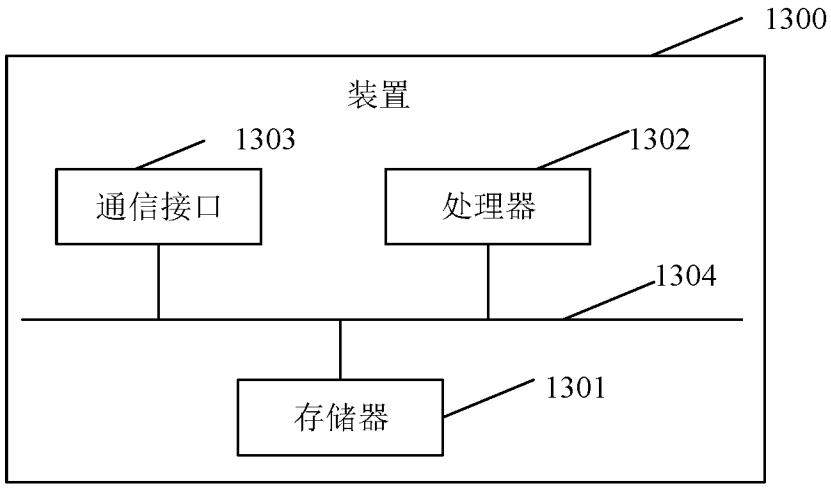

1300
装置
1303
通信接口
1302
处理器
1304
存储器
1301

图 13

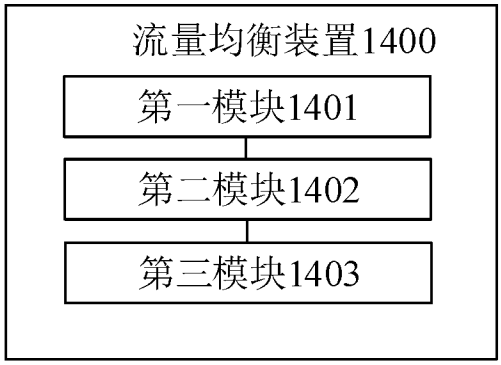

流量均衡装置1400
第一模块1401
第二模块1402
第三模块1403

图 14


流量均衡装置1500
第四模块1501
第五模块1502
第六模块1503

图 15

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/100270

A. CLASSIFICATION OF SUBJECT MATTER

H04L 29/08(2006.01)i; H04L 12/803(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)

CNKI, CNPAT, EPODOC, WPI: 数据包, 数据流, 流包, 流, 目的, 目标, 路径, 相同, 网络均衡, 排序通道, 尾包, 重排, 排序, packet, date packet, data stream, stream, forward path, path, same, balance, re-order, order, re-ordering channel, RC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 108737557 A (OPPO (CHONGQING) INTELLIGENT TECHNOLOGY CO., LTD.) 02 November 2018 (2018-11-02) description, paragraphs [0057]-[0105]	1-7, 9, 11
X	CN 103875261 A (QUALCOMM INC.) 18 June 2014 (2014-06-18) description, paragraphs [0057]-[0061]	8, 10, 12
A	CN 108390820 A (HUAWEI TECHNOLOGIES CO., LTD.) 10 August 2018 (2018-08-10) entire document	1-12
A	CN 109257282 A (BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 22 January 2019 (2019-01-22) entire document	1-12
A	US 2015124608 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 07 May 2015 (2015-05-07) entire document	1-12

☐ 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
10 April 2020	**24 April 2020**

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/CN2019/100270

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 108737557 A	02 November 2018	None	
CN 103875261 A	18 June 2014	KR 20140084001 A	04 July 2014
		JP 2014529957 A	13 November 2014
		WO 2013033106 A1	07 March 2013
		IN 201400398 P4	03 April 2015
		US 2013223336 A1	29 August 2013
		EP 2749045 A1	02 July 2014
CN 108390820 A	10 August 2018	WO 2019196630 A1	17 October 2019
CN 109257282 A	22 January 2019	None	
US 2015124608 A1	07 May 2015	CN 104618264 A	13 May 2015

国际检索报告

国际申请号
PCT/CN2019/100270

A. 主题的分类

H04L 29/08(2006.01)i; H04L 12/803(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））

CNKI, CNPAT, EPODOC, WPI：数据包，数据流，流包，流，目的，目标，路径，相同，网络均衡，排序通道，尾包，重排，排序，packet, date packet, data stream, stream, forward path, path, same, balance, re-order, order, re-ordering channel, RC

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 108737557 A (OPPO重庆智能科技有限公司) 2018年 11月 2日 (2018-11-02) 说明书第[0057]-[0105]段	1-7, 9, 11
X	CN 103875261 A (高通股份有限公司) 2014年 6月 18日 (2014-06-18) 说明书第[0057]-[0061]段	8, 10, 12
A	CN 108390820 A (华为技术有限公司) 2018年 8月 10日 (2018-08-10) 全文	1-12
A	CN 109257282 A (北京邮电大学) 2019年 1月 22日 (2019-01-22) 全文	1-12
A	US 2015124608 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2015年 5月 7日 (2015-05-07) 全文	1-12

☐ 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
2020年 4月 10日	2020年 4月 24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	魏玲
传真号 (86-10)62019451	电话号码 86-(10)-53961737

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/100270

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	108737557	A	2018年 11月 2日	无			
CN	103875261	A	2014年 6月 18日	KR	20140084001	A	2014年 7月 4日
				JP	2014529957	A	2014年 11月 13日
				WO	2013033106	A1	2013年 3月 7日
				IN	201400398	P4	2015年 4月 3日
				US	2013223336	A1	2013年 8月 29日
				EP	2749045	A1	2014年 7月 2日
CN	108390820	A	2018年 8月 10日	WO	2019196630	A1	2019年 10月 17日
CN	109257282	A	2019年 1月 22日	无			
US	2015124608	A1	2015年 5月 7日	CN	104618264	A	2015年 5月 13日