(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104641605 B
(45)授权公告日 2018.02.23

(21)申请号 201380048770.7
(22)申请日 2013.07.09
(65)同一申请的已公布的文献号
申请公布号 CN 104641605 A
(43)申请公布日 2015.05.20
(30)优先权数据
13/554,697 2012.07.20 US
(85)PCT国际申请进入国家阶段日
2015.03.19
(86)PCT国际申请的申请数据
PCT/US2013/049762 2013.07.09
(87)PCT国际申请的公布数据
WO2014/014713 EN 2014.01.23
(73)专利权人 思科技术公司
地址 美国加利福尼亚州
(72)发明人 铃木洋 荣·潘
弗拉维欧·博诺米
拉维·普拉萨德
帕德玛·艾基瑞荣
哈里普拉萨达·R·吉因杰帕里
安德鲁·罗宾斯
(74)专利代理机构 北京东方亿思知识产权代理有限责任公司 11258
代理人 李晓冬
(51)Int.Cl.
H04L 12/801(2006.01)
(56)对比文件
US 2004/0095882 A1,2004.05.20,
US 2003/0058802 A1,2003.03.27,
CN 101133606 A,2008.02.27,
CN 1093210 A,1994.10.05,
US 2004/0004961 A1,2004.01.08,
WO 97/04543 A2,1997.02.06,
US 2009/0016222 A1,2009.01.15,
US 2007/0223373 A1,2007.09.27,
US 2010/0149978 A1,2010.06.17,
审查员 张旭

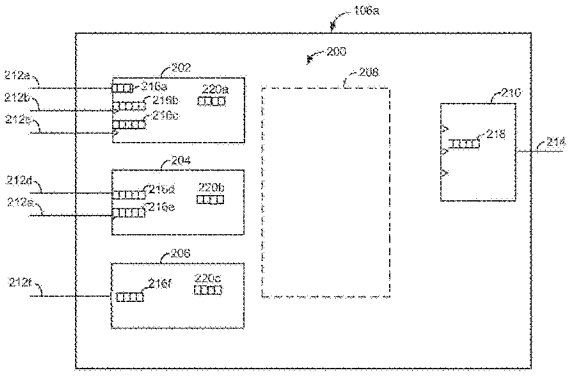
权利要求书3页 说明书11页 附图9页

(54)发明名称

用于分布式交换结构系统的智能暂停的方法、系统和装置

(57)摘要

提供了一种用于改进诸如暂停之类的流控制机制的执行的技术。在分布式系统中，该技术用来保持可用带宽的公平分配，同时还允许更少的丢弃分组、以及链路利用率最大化。例如，在一个实施例中，技术被提供以用于实现出站端口的带宽在争夺该出站端口的多个入站端口之上的公平共享分配。



CN 104641605 B

1.一种用于控制网络中的流的方法,包括:

在网络设备的入站缓冲模块处,基于从所述网络设备的第一出站端口接收到的反馈信息值Fb,计算由与所述入站缓冲模块相关联的入站端口接收流量的目标传输速率,所述反馈信息值Fb指示所述第一出站端口的阻塞程度;

在所述网络设备的入站缓冲模块处,针对所述第一出站端口将所述入站端口的计算出的目标传输速率与数据到达所述入站端口的速率进行比较;

在所述网络设备的入站缓冲模块处,基于所述比较的结果确定是否发送暂停消息,所述暂停消息包括基于在第一时间间隔Tw期间关于所述入站端口的计算出的目标传输速率与数据到达所述入站端口的速率的计算获得的暂停产生机率,

其中在多个入站缓冲模块中的每一个入站缓冲模块处针对该入站缓冲模块的关联的一个或多个入站端口执行上述的处理步骤,

其中,在周期的和/或触发的基础上计算所述一个或多个入站端口中的每一个入站端口的所述目标传输速率。

2.如权利要求1所述的方法,其中,在所述多个入站缓冲模块中的每一个入站缓冲模块上执行所述计算,以确定与所述入站缓冲模块相关联的一个或多个入站端口中的每一个入站端口的所述目标传输速率收敛于所述第一出站端口的带宽在所述入站端口之上的公平共享分配。

3.如权利要求1所述的方法,还包括:

响应于流动至所述入站端口的流量的改变,在运行时间动态地调整所述第一出站端口的带宽在与所述入站缓冲模块相关联的入站端口之上的分配。

4.如权利要求1所述的方法,其中,通过由所述入站缓冲模块执行的所述目标传输速率的计算,所述第一出站端口的带宽在与所述多个入站缓冲模块相关联的入站端口之上的改进的分配被实现。

5.如权利要求1所述的方法,其中,当预定总量的数据已经被所述第一出站端口接收到时,分组在所述第一出站端口被取样,并且所述取样触发反馈信息值Fb的重新计算。

6.如权利要求5所述的方法,其中,在所述第一出站端口的合计取样字节阈值被超过时,所述反馈信息值Fb被传输至所述一个或多个入站端口中的第一入站端口,所述第一入站端口是在所述第一出站端口处取样的分组的源。

7.如权利要求1所述的方法,其中,到达字节量AB在每个时间间隔Tw的开始处被重新设置到0和/或被重新计算,并且在时间间隔Tw期间,所述到达字节量AB的总量被所述多个入站缓冲模块中的每一个入站缓冲模块用来确定何时发送暂停消息。

8.如权利要求1所述的方法,其中,如果所述反馈信息值Fb超过了预定阈值,则包括关于反馈信息值Fb的信息的阻塞通知消息被所述第一出站端口发送至争夺所述第一出站端口的带宽的所述多个入站缓冲模块的入站端口,其中所述反馈信息值Fb与所述第一出站端口的阻塞程度相对应。

9.如权利要求1所述的方法,其中,当包括关于与所述第一出站端口处的阻塞程度相对应的反馈信息值Fb的信息的阻塞通知消息针对与所述多个入站缓冲模块中的一个入站缓冲模块相关联的所述一个或多个入站端口中的第一入站端口从所述第一出站端口被接收时,所述第一入站端口的所述目标传输速率被降低。

10.如权利要求1所述的方法，其中，在过去了时间间隔Tw而未从所述第一出站端口接收到针对与所述多个入站缓冲模块中的一个入站缓冲模块相关联的所述一个或多个入站端口中的第一入站端口的任何阻塞消息时，所述第一入站端口的所述目标传输速率被升高。

11.如权利要求1所述的方法，其中，所述暂停消息被用于命令源节点在特定的持续时间中停止至所述一个或多个入站端口中的第一入站端口的传输，所述特定的持续时间约等于时间窗口Tw减去时间t的长度，所述时间t与在时间窗口Tw期间在所述第一入站端口处接收到的与所述第一出站端口相关联的数据的到达速率超过所述第一入站端口的计算出的目标传输速率的点相对应。

12.如权利要求1所述的方法，其中，入站缓冲模块在第一时间窗口Tw期间发送针对所述一个或多个入站端口中的第一入站端口的暂停消息。

13.一种用于控制网络中的流的系统，包括：

多个网络节点；

多个网络设备，所述网络设备与所述网络节点互相连接；

其中，所述多个网络设备中的每一个网络设备被配置为：

在所述网络设备的入站缓冲模块处，基于从所述网络设备的第一出站端口接收到的反馈信息计算由与所述入站缓冲模块相关联的入站端口接收流量的目标传输速率，所述反馈信息指示所述第一出站端口的阻塞程度；

在所述网络设备的入站缓冲模块处，针对所述第一出站端口将所述入站端口的计算出的目标传输速率与数据到达所述入站端口的速率进行比较；

在所述网络设备的入站缓冲模块处，基于所述比较的结果确定是否发送暂停消息，所述暂停消息包括基于在第一时间间隔Tw期间关于所述入站端口的计算出的目标传输速率与数据到达所述入站端口的速率的计算获得的暂停产生机率，

其中，在所述网络设备的多个入站缓冲模块中的每一个入站缓冲模块上针对该入站缓冲模块的关联的一个或多个入站端口执行上述的处理步骤，

其中，在周期的和/或触发的基础上计算所述一个或多个入站端口中的每一个入站端口的所述目标传输速率。

14.一种用于控制网络中的流的装置，包括：

多个入站端口，所述多个入站端口与入站缓冲模块相关联；

多个出站端口，所述多个出站端口与出站缓冲模块相关联；

存储器；以及

处理器，所述处理器与所述多个入站缓冲模块相关联，所述处理器被配置为：

基于从所述装置的第一出站端口接收到的反馈信息值Fb计算由与所述入站缓冲模块相关联的入站端口接收流量的目标传输速率，所述反馈信息值Fb指示所述第一出站端口的阻塞程度；

针对所述第一出站端口将所述入站端口的计算出的目标传输速率与数据到达所述入站端口的速率进行比较；

基于所述比较的结果确定是否发送暂停消息，所述暂停消息包括基于在第一时间间隔Tw期间关于所述入站端口的计算出的目标传输速率与数据到达所述入站端口的速率的计

算获得的暂停产生机率，

其中，在所述装置的多个入站缓冲模块中的每一个入站缓冲模块上针对该入站缓冲模块的关联的一个或多个入站端口执行上述的处理步骤，

其中，在周期的和/或触发的基础上计算所述一个或多个入站端口中的每一个入站端口的所述目标传输速率。

15.如权利要求14所述的装置，其中，在所述多个入站缓冲模块中的每一个入站缓冲模块上执行的、用以确定与所述入站缓冲模块相关联的一个或多个入站端口中的每一个入站端口的所述目标传输速率的计算收敛于所述第一出站端口的带宽在所述入站端口之上的公平共享分配。

16.如权利要求14或15所述的装置，其中，通过由所述入站缓冲模块执行的所述目标传输速率的计算，所述第一出站端口的带宽在与所述多个入站缓冲模块相关联的入站端口之上的改进的分配被实现。

17.如权利要求14所述的装置，其中，当预定总量的数据已经被所述第一出站端口接收到时，分组在所述出站端口被取样，并且所述取样触发反馈信息值Fb的重新计算。

18.如权利要求14所述的装置，其中，在所述第一出站端口的合计取样字节阈值被超过时，所述反馈信息值Fb被传输至所述一个或多个入站端口中的第一入站端口，所述第一入站端口是在所述第一出站端口处取样的分组的源。

19.如权利要求14所述的装置，其中，到达字节量AB在每个时间间隔Tw的开始处被重新设置到0和/或被重新计算，并且在时间间隔Tw期间，所述到达字节量AB的总量被所述多个入站缓冲模块中的每一个入站缓冲模块用来确定何时发送暂停消息。

20.如权利要求14所述的装置，其中，如果所述反馈信息值Fb超过了预定阈值，则包括关于反馈信息值Fb的信息的阻塞通知消息被所述第一出站端口发送至争夺所述第一出站端口的带宽的所述多个入站缓冲模块的入站端口，其中所述反馈信息值Fb与所述第一出站端口的阻塞程度相对应。

21.如权利要求14所述的装置，其中，当包括关于与所述第一出站端口处的阻塞程度相对应的反馈信息值Fb的信息的阻塞通知消息针对与所述多个入站缓冲模块中的一个入站缓冲模块相关联的所述一个或多个入站端口中的第一入站端口从所述第一出站端口被接收时，所述第一入站端口的所述目标传输速率被降低。

22.如权利要求14所述的装置，其中，在过去了时间间隔Tw而未从所述第一出站端口接收到针对与所述多个入站缓冲模块中的一个入站缓冲模块相关联的所述一个或多个入站端口中的第一入站端口的任何阻塞消息时，所述第一入站端口的所述目标传输速率被升高。

23.如权利要求14所述的装置，其中，所述暂停消息被用于命令源节点在特定的持续时间中停止至所述一个或多个入站端口中的第一入站端口的传输，所述特定的持续时间约等于时间窗口Tw减去时间t的长度，所述时间t与在时间窗口Tw期间在所述第一入站端口处接收到的与所述第一出站端口相关联的数据的到达速率超过所述第一入站端口的计算出的目标传输速率的点相对应。

24.如权利要求14所述的装置，其中，入站缓冲模块在第一时间窗口Tw期间发送针对所述一个或多个入站端口中的第一入站端口的暂停消息。

用于分布式交换结构系统的智能暂停的方法、系统和装置

[0001] 相关申请

[0002] 本申请要求由Suziki等人于2012年7月20日提交的、标题为“用于分布式交换结构系统的智能暂停”、代理人案号为NO.CISCP625/976522的美国专利申请No.13/554,697的优先权，其全部内容被结合与此并用于所有目的。

技术领域

[0003] 本公开一般涉及在计算机网络中流控制的领域，并且更具体地涉及用于解决网络阻塞的有关暂停或类似流控制机制的技术。

背景技术

[0004] 流控制机制被用于防止并减少计算机网络中的阻塞。一些流控制机制通过临时停止至计算机网络的阻塞节点的数据传输直到在这些节点处的阻塞被减少来进行操作。例如，这种针对以太网络开发的一种流控制机制是由IEEE802.3x标准定义的暂停帧。阻塞节点可将暂停消息发送至向其发送分组的实体。该暂停消息指示发送源实体应该临时暂缓向该阻塞节点传输分组。在一个示例中，作为响应源实体可在暂停消息中指定的一段时间停止向该阻塞节点传输分组。

[0005] 随后，暂停机制被开发以提供基于优先级的流控制。这种流控制机制(例如，由IEEE 802.1Qbb标准定义)允许通过在不同服务质量(QoS)等级的流之间进行区分来更为精细地调整暂停的使用。例如，在IEEE 802.1Qbb中，使用服务类别(CoS)实现基于优先级的暂停。

[0006] 无论任何特定的实现方式，其暂停消息(和指导源实体临时停止向阻塞节点传输分组的其他流控制机制)的目标都是减少丢弃分组、链路利用率最大化、和/或防止或减缓网络节点处的阻塞。

[0007] 虽然目前用于实现暂停和类似流控制机制的现有系统提供了重要的益处，但是它们没有完全地实现这些流控制机制的潜能。本申请的各种实施例力图改进并且提供更为复杂的技术来使用这种流控制机制。

发明内容

[0008] 本申请的一方面公开了一种用于控制网络中的流的方法，包括：在网络设备的入站缓冲模块处，基于从所述网络设备的第一出站端口接收到的反馈信息值Fb，计算由与所述入站缓冲模块相关联的入站端口接收流量的目标传输速率，所述反馈信息值Fb指示所述第一出站端口的阻塞程度；在所述网络设备的入站缓冲模块处，针对所述第一出站端口将所述入站端口的计算出的目标传输速率与数据到达所述入站端口的速率进行比较；在所述网络设备的入站缓冲模块处，基于所述比较的结果确定是否发送暂停消息，所述暂停消息包括基于在第一时间间隔Tw期间关于所述入站端口的计算出的目标传输速率与数据到达所述入站端口的速率的计算获得的暂停产生机率，其中在多个入站缓冲模块中的每一个入

站缓冲模块处针对该入站缓冲模块的关联的一个或多个入站端口执行上述的处理步骤，其中，在周期的和/或触发的基础上计算所述一个或多个入站端口中的每一个入站端口的所述目标传输速率。

[0009] 本申请的另一方面公开了一种用于控制网络中的流的系统，包括：多个网络节点；多个网络设备，所述网络设备与所述网络节点互相连接；其中，所述多个网络设备中的每一个网络设备被配置为：在所述网络设备的入站缓冲模块处，基于从所述网络设备的第一出站端口接收到的反馈信息计算由与所述入站缓冲模块相关联的入站端口接收流量的目标传输速率，所述反馈信息指示所述第一出站端口的阻塞程度；在所述网络设备的入站缓冲模块处，针对所述第一出站端口将所述入站端口的计算出的目标传输速率与数据到达所述入站端口的速率进行比较；在所述网络设备的入站缓冲模块处，基于所述比较的结果确定是否发送暂停消息，所述暂停消息包括基于在第一时间间隔Tw期间关于所述入站端口的计算出的目标传输速率与数据到达所述入站端口的速率的计算获得的暂停产生机率，其中，在所述网络设备的多个入站缓冲模块中的每一个入站缓冲模块上针对该入站缓冲模块的关联的一个或多个入站端口执行上述的处理步骤，其中，在周期的和/或触发的基础上计算所述一个或多个入站端口中的每一个入站端口的所述目标传输速率。

[0010] 本申请的又一方面公开了一种用于控制网络中的流的装置，包括：多个入站端口，所述多个入站端口与入站缓冲模块相关联；多个出站端口，所述多个出站端口与所述出站缓冲模块相关联；存储器；以及处理器，所述处理器与所述多个入站缓冲模块相关联，所述处理器被配置为：基于从所述装置的第一出站端口接收到的反馈信息值Fb计算由与所述入站缓冲模块相关联的入站端口接收流量的目标传输速率，所述反馈信息值Fb指示所述第一出站端口的阻塞程度；针对所述第一出站端口将所述入站端口的计算出的目标传输速率与数据到达所述入站端口的速率进行比较；基于所述比较的结果确定是否发送暂停消息，所述暂停消息包括基于在第一时间间隔Tw期间关于所述入站端口的计算出的目标传输速率与数据到达所述入站端口的速率的计算获得的暂停产生机率，其中，在所述装置的多个入站缓冲模块中的每一个入站缓冲模块上针对该入站缓冲模块的关联的一个或多个入站端口执行上述的处理步骤，其中，在周期的和/或触发的基础上计算所述一个或多个入站端口中的每一个入站端口的所述目标传输速率。

附图说明

[0011] 通过参考以下结合示出特定示例实施例的附图的描述，本公开可以被最好地理解。

[0012] 图1是结合本公开实施例的示例计算机网络的简化框图；

[0013] 图2是结合本公开实施例的具有分布式交换结构系统的示例网络设备的简化框图；

[0014] 图3A是结合本公开实施例的具有分布式交换结构系统的示例网络设备的简化框图；

[0015] 图3B是结合本公开实施例的具有分布式交换结构系统的示例网络设备的简化框图；

[0016] 图4A根据本公开实施例描述了示出时间间隔的使用的示例时间线；

[0017] 图4B根据本公开实施例描述了示出时间间隔的使用的示例时间线；

[0018] 图5A和图5B是根据本公开实施例的描述示例方法的不同特征的简化高级流程图；以及

[0019] 图6是适用于实现本公开一些实施例的示例网络设备的简化框图。

具体实施方式

[0020] 概述

[0021] 提供了一种用于改进诸如暂停(Pause)之类的流控制机制的使用的技术。除了其他方面，提供了用于改进在具有与入站和出站端口相关联的I/O缓冲模块的分布式系统的网络设备的上下文中暂停的使用的技术。在这样的实施例中，该技术改进了出站端口的带宽在网络设备的多个入站端口上的分配，其中入站端口争夺出站端口的带宽。提供了用于智能地确定何时以及从网络设备的哪些入站端口发出暂停消息的自动化技术。

[0022] 根据本申请的实施例，在多个I/O缓冲模块中的每一个I/O缓冲模块，基于从第一出站端口接收到的阻塞通知消息中的反馈信息，计算一个或多个关联的入站端口中每一个的目标传输速率。对于每一个入站端口，计算出的入站端口的目标传输速率与第一出站端口的数据到达该入站端口的速率进行比较。基于比较的结果，做出是否从入站端口发送暂停消息的确定。根据本申请的实施例，在周期的和/或触发的基础上执行该计算。根据本申请的实施例，针对多个入站端口中的每一个执行的目标传输速率的计算随着时间收敛于第一出站端口的带宽在多个入站端口之上的公平共享分配。根据本申请的实施例，响应于流量流中的变化，在运行时间对带宽的分配自动做出调整。

[0023] 本公开的上文所述以及其他特征、实施例和优势在参考以下的说明书、权利要求和附图时会变得更为显而易见。

[0024] 示例实施例

[0025] 在现代网络中，交换机或其他网络设备可包括多个硬件和/或软件模块。例如，多个I/O缓冲模块可以被用于实现网络节点的I/O接口。线卡是这种I/O缓冲模块的示例。这些I/O缓冲模块可以位于不同的芯片或集成电路。在替代实施例中，这些缓冲模块中的一个或多个可以位于相同的芯片或集成电路。

[0026] 上述I/O缓冲模块中的每一个可以与多个入站端口或出站端口(或起到入站和出站端口作用的端口)相关联。此外，每个I/O缓冲模块可保持多个实际或虚拟的队列，例如，每个队列与关联的端口中的之一相对应。入站和出站模块可通过网络设备的交换结构或其他互连机制被互相连接。这种互连机制的示例是交换结构、环形堆栈或十字型结构。

[0027] 在具有多个互相连接的I/O缓冲模块的网络设备中，以争夺相同阻塞出站端口的所有入站端口实现出站端口的带宽的公平加权共享的方式对入站端口产生暂停帧会是非常具有挑战性的。

[0028] 在这种情况下实现暂停机制的一个挑战是想出一种方法来确定何时以及如何产生暂停，该方法考虑到在不同I/O缓冲模块端口处的流量流。

[0029] 通常，在现有系统中，用于产生暂停的触发取决于超过预定阈值的入站队列的缓冲占用率水平。例如，在一个现有机制中，当网络设备的阻塞出站端口的缓冲占用率被发现超过预定阈值时，控制信号被发送至所有的源入站端口(即，发送至阻塞端口的所有入站端

口)，或者控制信号被广播至网络设备的所有入站端口。作为响应，如果那些入站端口具有去往阻塞缓冲/端口的分组，那些入站端口会停止向该出站端口传输分组。而且，如果那些入站端口本身的入站队列被充满并超过预定缓冲占用率阈值，那么系统将从关联的入站端口发送出暂停帧直至阻塞被清除。该技术存在其他的变化，但是所有的变化通常取决于查看队列的缓冲占有率是否超过预定阈值。预定阈值不会随时间改变，并且仅与缓冲占用率进行比较。在这种系统中，实现公平共享带宽分配是很困难的。

[0030] 根据本公开的实施例，向第一出站端口传输数据的多个入站端口中的每一个入站端口的目标速率以在去往共同的出站端口的入站端口中实现公平带宽分配的方式被确定，同时在多个入站端口之上避免出站端口处的缓冲溢出/分组丢弃以及低链路利用率。

[0031] 根据本公开的实施例，针对争夺第一出站端口的多个入站端口中的每一个入站端口的目标传输速率的计算被执行，使得随着时间将这些目标传输速率的计算收敛以产生如下的针对入站端口的目标传输速率的分配，该目标传输速率的分配近似于第一出站端口的带宽的公平共享分配。

[0032] 以下描述的本发明的实施例是依据用于确定何时以及如何从网络设备发送暂停消息的技术来讨论的。然而应当理解的是除了暂停消息之外，本发明的教导也可以被用来与具有相似功能的其他流控制机制相结合。而且，应该理解的是以下描述的用于在多个I/O缓冲模块之上保持公平共享的解决方案也能够被应用于不同系统之上的公平共享(例如，流控制领域的出站和入站点之间的协作)。协作“模块”不一定是以各种实施例中描述的方式作为单个网络节点的一部分的I/O缓冲模块。以下描述的本发明的实施例并没有旨在限定本发明的范围。

[0033] 图1是结合本公开实施例的示例计算机网络的简化框图。

[0034] 计算机网络100包括经由多个通信链路被相互耦合的多个网络节点102、104和106。如图1所描述，计算机网络100包括通过多个网络设备106被耦合到彼此的用户(客户)系统102和服务器系统104。网络设备106的示例是交换机和路由器。应该理解的是比起图1所示的那些，计算机网络100还可以包括附加或不同类型的网络节点。图1所描述的计算机网络100仅仅是结合本公开的示意性实施例，并不限制如权利要求中所列举的本发明的范围。本领域的普通技术人员可以认识到其他变化、修改、以及替换。

[0035] 计算机网络100可以由很多互相连接的计算机系统和通信链路组成。例如，计算机网络110可以是LAN、广域网(WAN)、无线网络、内联网、专用网络、公共网络、交换网络、或任意其他合适的网络(比如，互联网、或任意其他计算机网络)。

[0036] 根据本发明的实施例，在图1所描述的网络环境中网络设备106被配置为发送暂停消息。根据本发明的实施例，网络设备106被配置为以有效和公平的方式对有关发送暂停消息的自动任务执行处理。这些任务可包括监控第一出站端口上的阻塞情况、对于争夺该第一出站端口的带宽的多个入站端口计算至该第一出站端口的目标传输速率、监控争夺第一出站端口的多个入站端口处的第一出站端口流量的到达速率、以及将阻塞通知消息从第一出站端口发送至多个发生争夺的入站端口。

[0037] 可以由软件模块执行处理，该软件模块可以通过网络设备106、通过耦合到网络设备106或包括在网络设备106内的硬件模块、或通过它们的组合来执行。例如，该硬件模块可由被配置为I/O线卡的专用集成电路组成。如上所述，缓冲模块可包括一个或多个入站和/

或出站端口，并且可保持与那些端口相对应的入站和出站缓冲。入站或出站缓冲模块可位于各自的ASIC上；或在替代实施例中，入站或出站缓冲模块可在相同的ASIC上实现。与入站和出站硬件模块相关联的软件模块可执行关于暂停的使用的计算。流控制公式可以被网络环境的用户(例如，最终用户、系统管理员、管理者等)配置。有关在网络设备106的出站和入站模块上执行的计算的细节会在下文中提供。

[0038] 在图1中描述的网络设备106中的每一个都具有两个在其左侧的端口108和一个在其右侧的端口110。应该理解的是在图1中所指定的端口数量仅用于示意性的目的。出于讨论的目的，假定在图1中的所有数据的流动方向是从左到右的。因此，在左侧的端口108可以被描述为入站端口，并且在右侧的端口110可以被描述为出站端口。

[0039] 在图1中，网络设备106b的入站端口108c从两个节点-用户(客户)系统102b和102c接收数据。并且入站端口108c将该数据转发至出站端口110b，出站端口110b进而将数据从网络设备106b传输至服务器系统104b或服务器系统104c。与之相比，网络设备106b的入站端口108d仅从用户(客户)系统102d接收数据。入站端口108d同样将其接收的数据转发至出站端口110b。

[0040] 根据本发明的教导，由入站端口108和出站端口110用以执行流控制的信息可以被存储在能够由执行各自的处理的入站或出站模块访问的存储器位置。例如，该信息可以被存储在入站或出站模块分别物理地位于其上的线卡或其他硬件模块。这种硬件模块中的每一个可以与一个或多端口相关联，并且还可以保持一个或多个实际的或虚拟的队列。应当理解的是关于那些端口和队列的信息可以以本领域技术人员已知的各种格式被存储。

[0041] 图2是结合本公开实施例的具有分布式交换结构系统200的示例网络设备106b的简化框图。

[0042] 在图2中，示出的网络设备具有分布式交换结构系统200，该分布式交换结构系统200包括通过互连机制108被耦合到出站模块210的多个入站模块202、204、和206。入站和出站模块202、204、206和210中的每一个可以与一个或多个入站和/或出站端口212和214相关联。

[0043] 例如，入站模块202、204、和206可以与一个或多个入站端口212相关联，并且可通过交换结构208被连接到出站模块210。出站模块210可具有出站端口214。应该理解的是入站和出站模块以及入站和出站端口的数量在不同的实施例中是可以变化的。

[0044] 入站模块202、204、和206以及出站模块210中的每一个都可以保持队列。应该理解的是这些队列可以以各种方式被实现。在图2所示的一个实现方式中，入站缓冲模块202、204、和206中的每一个都具有结构入站队列220。入站缓冲模块202、204、和204在每个入站端口基础上保持针对每个它们关联的入站端口的从出站端口214接收到的阻塞通知消息的核算(accounting)，以及其入站端口处的数据到达速率。元件216描述了这种到达各个入站端口212中的每一个并且目前被保存在入站模块202、204和206中的分组数量的每端口核算。应该理解的是在图2的实现方式中，元件216与在每个端口基础上执行的核算的概念相对应，并不代表实际硬件组件。

[0045] 根据本申请的另一个实施例(未被描述)，这些队列可以被实现为虚拟输出队列(VOQ)。单独的虚拟输出队列可针对入站端口212和出站端口214的每一个可能的组合被保持。例如，入站模块202可保持3个VOQ，每个VOQ与其入站端口212a、212b、以及212c中的一个

和出站端口214的配对相对应。

[0046] 在另一个实施例(未被描述)中,应该理解的是结构入站队列220中的每一个可以与实际的或虚拟的服务类别队列相对应。即,可针对由网络设备提供的入站端口、出站端口、和服务类别(CoS)的每一个可能的组合保持的单独入站队列或单独核算。例如,计算机网络可向用户提供对客户提供不同服务质量水平的选择;这些可通过使用服务类别被实现,并且对在管理网络阻塞以及关于整形和监管网络流做决定的情况下的不同分组之间的优先排序是有用的。本申请的发明特征在提供服务类别的系统中的应用会被本领域的技术人员所理解。

[0047] 如图2所示,入站模块202具有入站端口212a、212b、和212c,并且针对与这些入站端口中的每一个至出站端口214相关联的流保持核算216a、216b、和216c。类似地,入站模块204具有入站端口212d和212c,并且保持与从这些入站端口中的每一个至出站端口214的流相对应的核算216d和216e。

[0048] 而且,出站端口210具有出站端口队列218,该出站端口队列218保存分组直到准备好将这些分组从出站端口214发送出去。

[0049] 根据本公开的实施例,出站端口214的可用带宽在入站端口212之上的分配以实现公平并且还能有效地减少出站端口214的阻塞的方式被确定。

[0050] 为了便于说明,在图2中仅描述了一个出站端口(出站端口214)。应该理解的是,实际上网络设备106很可能会具有多个出站模块,每个出站模块与一个或多个出站端口相关联,并且每个出站端口以在图2中示出的出站端口214与入站端口212a至212f进行交互的相同方式与多个入站端口进行交互。应该理解的是关于出站端口112和其发生争夺的入站端口所描述的关于实现暂停的所有任务可以与其他出站端口和它们发生争夺的入站端口同时发生。而且,网络设备可以被配置,使得这些任务对于每个这种出站端口和其发生争夺的入站端口的组被同时执行。

[0051] 图3A是具有与图2所描述的网络设备类似配置的示例网络设备106b的简化框图。网络设备106b包括将入站和出站缓冲模块相互连接的分布式交换结构108。出于示意性的目的,图3A示出了出站端口214的带宽在网络设备106b的多个入站端口212之上的可能分配。带宽的分配示于括号中,以将带宽的分配与图3A中使用的元件标号区分开来。

[0052] 如图3A所描述的,出站端口214具有600单位的可用带宽。给出了六个入站端口(端口112a至112f),理想情况下每个入站端口112可接收到100单位的带宽作为其分配。

[0053] 然而,在执行带宽分配的一些系统中,不管与那些争夺出站端口114的带宽的入站模块相关联的入站端口112的数量如何,出站端口114的带宽在争夺出站端口114的入站模块102、104、和106之间被分配。根据这种系统,200单位的带宽分配(1/3出站端口的带宽)可以被提供给入站模块102、104和106中的每一个。根据这样的实施例,入站模块106的入站端口112f可接收出站端口114的带宽的200单位的带宽,同时入站模块104的入站端口112d和112e中的每一个接收100单位的带宽,并且入站模块102的入站端口112a、112b和112c中的每一个接收66单位的带宽。虽然该系统实现了一些公平上的改进,仍然希望有进一步的改进。

[0054] 上述的系统可能对一些入站端口进行不公平地处罚,使得这些入站端口例如因为其他入站端口引起的出站端口的过载而被惩罚。例如,考虑到入站模块104的入站端口112d

和112e具有等于或低于它们的出站端口114可用带宽的公平份额(100单位)的到达速率(即,要从出站端口114发送出去的输入分组的速率)的情况。并且入站模块102的3个入站端口具有总共高于出站端口114的总体可用容量的到达速率。在该情况中,尽管模块102是引起出站端口114的阻塞的原因,但不仅入站模块102至出站端口114的传输可能被暂缓,而且在公平共享的情况下不应被停止的入站模块104至出站端口114的传输也可能被暂缓。

[0055] 根据本公开的教导,网络设备106被配置为提供自动化系统,用于在包括多个I/O缓冲模块的分布式系统中的争夺相同阻塞出站端口的入站端口之间公平分配带宽。在理想公平共享系统中,每个争夺出站端口的带宽的入站端口可以以图3B所描述的方式获得相同份额的出站端口的带宽。

[0056] 根据本公开的实施例,基于出站缓冲的阻塞程度产生对多个入站端口的暂停,使得如图3B所示的多个入站端口得到公平共享或加权公平共享的出站端口的带宽。

[0057] 根据本公开的实施例,争夺第一出站端口的带宽的多个入站端口中每一个的目标传输率以在去往共同出站端口的入站端口中实现公平带宽分配的方式被确定,同时在出站端口上避免缓冲溢出/分组丢弃以及低链路利用率。

[0058] 根据本公开的实施例,基于在周期的和/或(在一些情况中)触发的基础上在每个发生争夺的入站端口上执行的计算,针对多个将数据发送至第一出站端口的发生争夺的入站端口的目标传输速率被确定。

[0059] 根据本公开的实施例,在出站阻塞点,出站模块将有关阻塞出站端口的阻塞程度的反馈信息(Fb)发送回争夺出站端口的入站端口。

[0060] 根据本公开的实施例,为减少进行这种计算的资源成本并且使控制流量的开销最小化,监控出站端口的阻塞仅在周期的基础上被实施。例如,每当在出站端口处接收到合计100KB的数据时,分组取样被实施。针对出站端口的取样被实施时,基于目前的出站队列占用率水平,针对出站端口的反馈信息值可以被计算。

[0061] 根据本公开的实施例,每当反馈信息(Fb)被入站缓冲模块102、104、或106接收时,对于入站端口112,部分地基于接收到的反馈信息,入站缓冲模块重新计算入站端口的目标传输速率(Tr)。根据本公开的一些实施例,目标传输速率Tr意为针对该入站端口估算公平共享分配的出站端口的带宽。

[0062] 根据本公开的实施例,针对入站端口的公平共享计算的示例在如下情况中发生:每当针对入站端口接收到包括阻塞反馈程度(Fb)信息的阻塞通知时,FairShare(公平共享)就被减少。

[0063] 例如,FairShare=FairShare-FairShare*Alpha*Fb(Alpha<1)。应该注意的是该计算是基于阻塞通知消息中从出站端口被发送回的出站队列长度/阻塞程度反馈信息(Fb)的。

[0064] 而且,根据本公开的实施例,如果在时间间隔Tw没有接收到阻塞通知消息,则入站端口的FairShare可以被增加。根据本公开的实施例,如果在间隔Tw期间没有接收到阻塞通知消息,则入站端口的目标传输速率Tr以增加该入站端口的FairShare的方式被重新计算。例如,在该环境中可通过以下公式对该入站端口计算FairShare:FairShare=FairShare+FairShare*Beta(Beta<1)。

[0065] 根据本公开的实施例,这种class_i的FairShare可通过计算Weight_class*

FairShare被进行加权。

[0066] 根据本公开的实施例，在持续时间为Tw的时间间隔期间，入站反应点(例如，入站缓冲模块)监控入站端口处的分组的到达速率，并且如果在时间间隔Tw内接收的字节数量大于目前计算的该入站端口的FairShare，则产生带有暂停量Tq(即，请求停止传输的持续时间)的暂停。在这样的实施例中，针对源节点产生的带宽可以通过如下来计算：BW＝Link Speed*(Tpause-Tq)/Tpause，其中Tpause可以是在进行暂停确定中使用的时间窗Tw的时间间隔长度(例如，秒级)。根据本公开的实施例，此概念还可被扩展至优先级暂停(PFC)的实现方式。

[0067] 根据本公开的实施例，当在时间窗口期间入站端口处的合计数据到达速率超过入站端口的目标传输速率时，产生带有反映时间间隔Tw中的剩余时间长度的暂停量Tq的暂停。

[0068] 图4A根据本公开的实施例描述了示出时间间隔的使用的示例时间线。Tw示出了本公开的各种实施例所使用的时间间隔，以触发对入站端口的目标传输速率的重新计算。当输入数据的到达速率等于或超过目前针对入站端口计算的目标传输速率时，参数“t”指示在时间间隔Tw内的时间点。

[0069] 符号Pause(Tw-t)指示暂停消息可以在入站端口处产生，并且暂停消息可指示至该入站端口的传输应该被停止Tw减t的时间长度。

[0070] 根据其他实施例，可使用不同的暂停量机制。在这样的实施例中，例如，暂停消息可命令源节点将传输停止Tpause的固定暂停长度。图4A根据本公开的实施例描述了示出暂停机率Pp以及在固定时间间隔Tpause中发生的暂停的使用的示例时间线。

[0071] 暂停长度Tpause可与在实现上述的目标速率传输的计算中使用的时间间隔Tw相对应，或在其他实施例中，Tpause可以是更短或更长的时间间隔。新变量暂停机率Pp被引入，暂停机率Pp与在给定时间间隔中暂停被要求的机率相对应。下面是可用于计算暂停机率Pp的一个公式：Arrival*(1-Pp)＝TargetRate。换句话说，根据该公式：

[0072] Pp＝1-TargetRate(目标速率)/ArrivalRate(到达速率)

[0073] 其中在暂停机率Pp超过预定阈值(例如，0)的情况下，会产生如上所述具有Ppause的固定暂停长度的暂停消息。

[0074] 根据本公开的实施例，本发明的特征是随着时间在每个反应点即每个独立的入站模块上执行的、用以重新校准其本身的目标传输速率的计算收敛于希望的稳定状态，在该稳定状态中，例如在每个向出站端口114发送流量的入站端口之上的出站端口114的带宽的公平分配可以被近似或获得。根据本公开的实施例，不需要执行总体的计算以确定公平共享分配的带宽，而是例如通过上述的分布式过程，反映了带宽的公平分配的例如在多个入站端口之上的带宽的分配可被自然而然地实现。

[0075] 根据本公开的实施例，在出站和入站端口处执行的计算收敛以提供多个入站端口之上的公平带宽分配的过程不是静态过程，而是响应于网络流中的变化的动态过程。例如，该过程可动态地响应于将分组发送至出站端口114的入站端口的数量以及由那些入站端口发送的流量水平的改变。该过程还可动态地响应于出站端口处可用带宽的量的改变。即，根据本公开的实施例，响应于网络流量流中的改变可以在运行时间进行动态地调整，以实现出站端口带宽在与多个入站缓冲模块相关联的多个入站端口之上的希望的分配。

[0076] 应该理解的是出站端口的可用带宽可以发生动态改变，例如，其中出站端口是来自下一个网络节点的暂停消息的接收端。或者，例如，如果流量根据诸如服务类别之类的服务质量分界被调控，并且服务类别之一被用在实际服务中或从实际服务中去掉，那么出站端口的可用带宽可以发生改变。

[0077] 根据本公开的实施例，相对小数量的计算被要求，并且即使是这种计算也仅需要在周期的基础上和/或响应于特定的触发被执行。例如，仅需要在周期的或触发的基础上进行出站端口处的阻塞的监控或入站端口处的目标传输速率是否需要被提高或降低的检查。此外，根据从出站端口114收到带有指示该出站端口处的阻塞情况的反馈信息Fb的阻塞通知消息，入站端口可以向下调整其目标分配。

[0078] 根据本公开的实施例，在每个时间窗口Tw结束时，执行检查以查看在该时间窗口Tw期间是否接收到针对入站端口的阻塞通知消息。如果在该时间窗口期间没有接收到阻塞通知消息，则使用在目标传输速率中计算增加的公式提高该入站端口的目标传输速率。

[0079] 到达量AB反映在特定时间窗口Tw期间在入站端口处接收到的流量的总量。根据本公开的实施例，在每个时间窗口开始时到达量可以被重新设置到0，使得到达量仅反映在单个时间窗口期间接收到的流量。在其他实施例中，在第一时间窗口期间的到达量AB(即，到达的流量的量)可在一定程度上转入到第二时间窗口。在入站端口需要不止一个时间间隔Tw来处理在第一时间窗口接收到的大量的流量的情况下，该特征是有用的，因此输入流量上需要更长的暂停。将到达速率从一个时间窗口转入到下一个时间窗口可促使提供更长的暂停。因此，根据本公开的实施例，在每个时间窗口Tw开始时，不是将到达字节AB仅重新设置到0；而是基于考虑到剩余量AR(即，在上一个时间窗口被接收但没有被完全处理的流量)的公式将AB重新计算。

[0080] 根据本公开的实施例，目标公平共享分配的计算和执行在分布式的基础上实施。入站模块确定入站端口的目标传输速率以及何时由入站模块本身发送暂停消息。不需要集中管理或控制带宽分配。而且，带宽分配以自动化的方式被执行。

[0081] 根据本公开的实施例，为了确定反馈Fb信息对出站端口处的阻塞情况的评估不仅考虑到出站队列的缓冲占用的水平，而且还考虑到自上一次评估出站队列的情况后缓冲占用水平是否被提高或降低。

[0082] 根据本公开的实施例，例如，当达到用于取样的合计字节阈值时，针对接收到的分组计算反馈信息，并且反馈信息仅被发送回是取样分组的源的入站端口，而不是发送回争夺出站端口的所有入站端口。

[0083] 图5a和5b是根据本公开实施例的描述示例方法的不同特征的简化高级流程图。图5a和5b示出的特征在一些方面与上述本公开的各种实施例的特征重叠。应该理解的是在不脱离本发明的精神的情况下，这些描述的实现方式的各种具体特征可以被改变。而且，应该理解的是在本公开中描述的不同实现方式的各种特征可以被合并在另一个实施例中。

[0084] 应该理解的是在阻塞点执行计算的公式(用于得到反馈Fb的量)以及在反应点执行计算的公式(用于得到新的目标传输速率Tr)在上述的系统中公式可从上文说明的实施例被进一步细化。例如，如下文所描述，公式可以被进一步细化以向流控制系统提供额外的优势。

[0085] 根据本公开的实施例，下面为在阻塞点(出站模块)和反应点(入站模块)执行的计

算。

[0086] 在阻塞点：

[0087] 提供1比特反馈,Fb:1(如果qlen>qref)

[0088] 对于在出站端口处接收的数据,每100KB数据进行一次取样。计算qlen是否大于qref。如果qlen>qref,则将阻塞通知消息发送至取样分组的源。

[0089] 在反应点：

[0090] 如果阻塞通知消息CN_ij(入站端口i,类别j)被接收,则使用公式TR_ij=TR_ij*Gd计算新的目标传输速率TR。应该注意的是该降低是基于当前目标传输速率TR乘以一个常数的。

[0091] 如果在时间间隔Tw期间没有接收到针对特定入站端口的阻塞通知消息CN_ij,则使用如下公式计算该入站端口的新的目标传输速率:TR_ij=TR_ij+increaseStep。应该注意的是在没有收到任何阻塞通知消息CN的情况下,由时间间隔Tw的结束触发传输速率的升高。

[0092] 如果在时间窗口中到达流量的量AB大于TR_ij*Tw,则产生暂停帧以便实现在入站端口之间的公平带宽分配。

[0093] 下面的公式进一步细化了各种发明特征。

[0094] 在阻塞点,反馈量Fb可使用如下公式计算。

[0095] ●Fb=(qlen-qref)+w*(qlen-qold)

[0096] ●对于在出站端口处接收的数据,每100KB数据进行一次取样。如果Fb>0,则将阻塞通知消息发送至取样分组的源。

[0097] 在反应点：

[0098] ●如果接收到针对入站端口的CN_ij(入站端口i,类别j),则使用如下公式计算该入站端口的新的目标传输速率：

[0099] ●TR_ij=TR_ij-TR_ij*Gd*Fb

[0100] ●应该注意的是该降低是基于当前的速率的

[0101] ●如果在时间间隔Tw未接收到CN_ij,反而使用如下公式计算新的目标传输速率：

[0102] ●TR_ij=TR_ij+TR_ij*Gi+Ri

[0103] ●应该注意的是从出站端口未收到阻塞通知的情况下,TR的升高是基于流逝的设定时间间隔的。

[0104] ●如果在时间窗口Tw的时间τ(0≤τ≤Tw)处的到达流量(以字节为单位)>TR_ij*Tw,则产生持续时间为Tw-τ的暂停帧

[0105] ●使得在入站端口之间的公平BW分配可以被实现

[0106] ●Tw是带宽测量时间窗口

[0107] 如上所述,根据本发明的实施例,网络设备106被配置为执行处理,以根据本发明的教导以智能方式发送暂停消息。图6是根据本发明实施例的网络设备106的简化框图。

[0108] 网络设备600包括主要中央处理单元(CPU)610、接口650、和总线660(例如,PCI总线)。在适当的软件或固件的控制之下进行作用时,CPU 610负责诸如交换和/或路由计算以及网络管理之类的任务。它最好在包括操作系统(例如,Cisco System,Inc(思科系统公司)的互联网络操作系统(IOS®))的软件以及任意适当的应用软件的控制之下实现所有这些

功能。CPU 610可包括一个或多个处理器630，比如，来自摩托罗拉家族的微处理器或来自MIPS家族的微处理器的处理器。在替代的实施例中，处理器630是为控制网络设备600的操作专门设计的硬件。在特定的实施例中，存储器620(比如，非易失性的RAM和/或ROM)也是CPU 610的组成部分。然而，存在很多种不同的方法将存储器耦合到系统。存储器块610可以被用于各种用途，例如，缓存和/或存储数据、程序指令等。

[0109] 接口650一般被提供为接口卡(有时被称为“线卡”)。通常，接口卡控制通过网络的分组或分组片段的发送和接收，有时支持与网络设备600一起使用的其他外围设备。可以被提供的接口当中有以太网接口、电缆接口、DSL接口等。此外，各种超高速接口可以被提供，比如，快速以太网接口、千兆以太网接口、HSSI接口、POS接口、FDDI接口等。通常，这些接口可以包括适用于与适当介质通信的端口。在一些情况下，这些接口还包括独立的处理器以及(在一些情况下)易失性RAM。独立的处理器可以控制通信密集型任务，比如，分组交换和/或路由、介质控制和管理。通过提供用于通信密集型任务的单独的处理器，这些接口允许主要微处理器610有效地执行交换/或路由计算、网络诊断、安全功能等。

[0110] 虽然在图6中示出的系统是本公开的一个具体的网络设备，但并不意味着它是能够实现本公开的唯一网络设备架构。例如，具有同样可以处理通信以及交换和/或路由计算等的单一处理器的架构被经常使用。而且，其他类型的接口和介质也可以与网络设备一起使用。

[0111] 不管网络设备的配置如何，它可以采用一个或多个被配置为存储数据、程序指令的存储器或存储器模块(例如，存储器块640)，用于一般用途的网络操作和/或本文所描述的发明技术。程序指令可控制操作系统的操作和/或一个或多个应用。一个或多个存储器还可以被配置为存储公平共享值和参数、突发阈值、最小和最大阈值、选项选择位、类别参数/规范、计时器、队列特性、分组到达历史参数等。

[0112] 因为可以采用这种信息和程序指令以实现本文所描述的系统/方法，所以本公开涉及包括程序指令、状态信息等的机器可读介质以用于执行本文所描述的各种操作。机器可读介质的示例包括(但并不限于)诸如硬盘、软盘、和磁带之类的磁性介质；诸如CD-ROM盘和DVD之类的光学介质；诸如光磁软盘之类的磁光介质；以及专门被配置为存储程序指令的示例的硬件设备，程序指令的示例包括比如由编译器产生的机器代码，以及包括可以由计算机使用解释器执行的高级代码的文件。

[0113] 虽然通过参考其中具体的实施例明确地示出并描述了本公开，但是应该理解的是在不脱离本公开的精神或范围的情况下，本领域的技术人员可以对公开实施例的形式和细节进行改变。例如，本公开的实施例可采用各种网络协议和架构。因此，本公开旨在被解释为包括所有落入本公开的真实精神和范围内的变化和等同物。

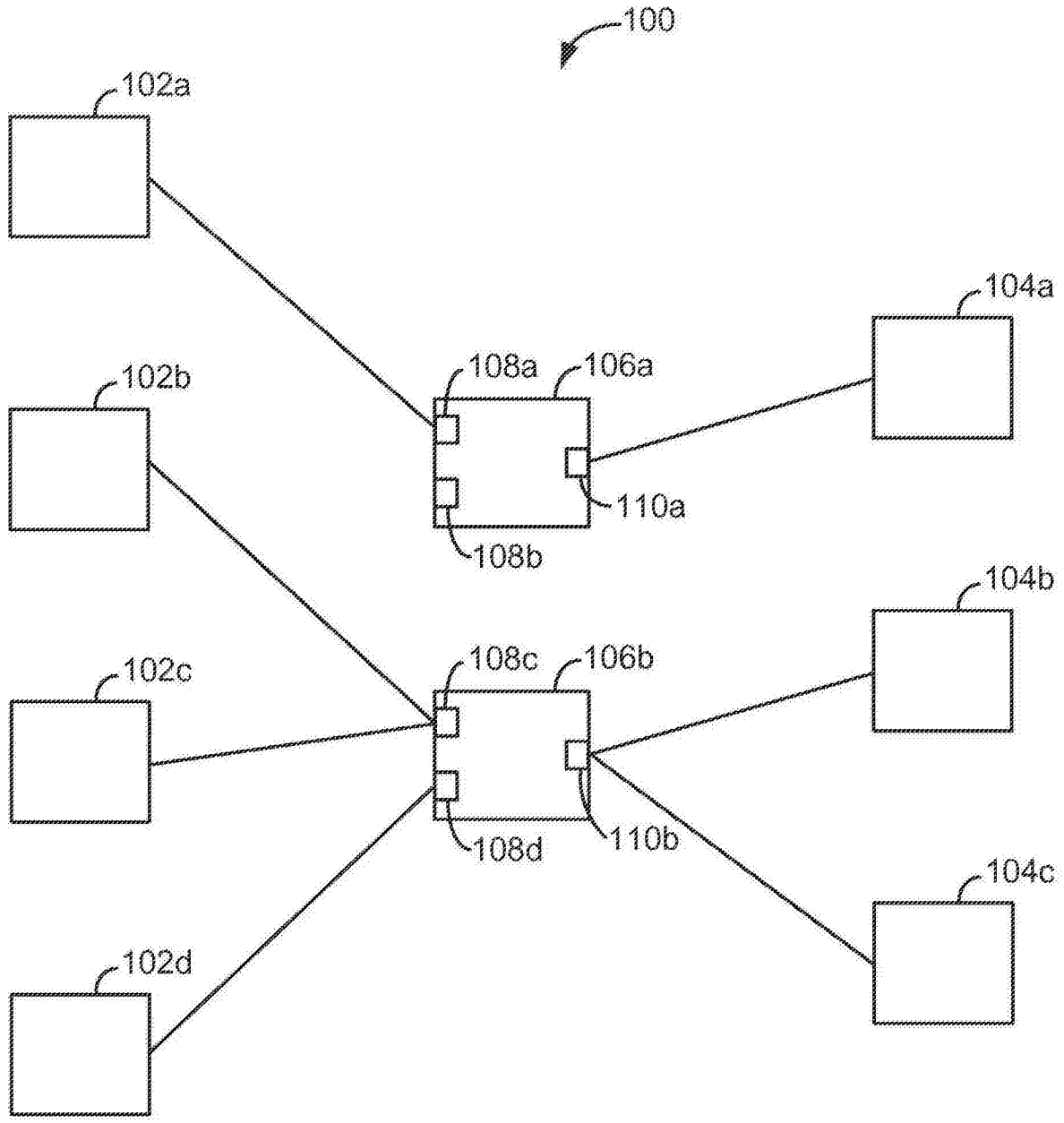

100
102a
102b
102c
102d
104a
104b
104c
108a
106a
110a
108b
108c
106b
108d
110b

图1

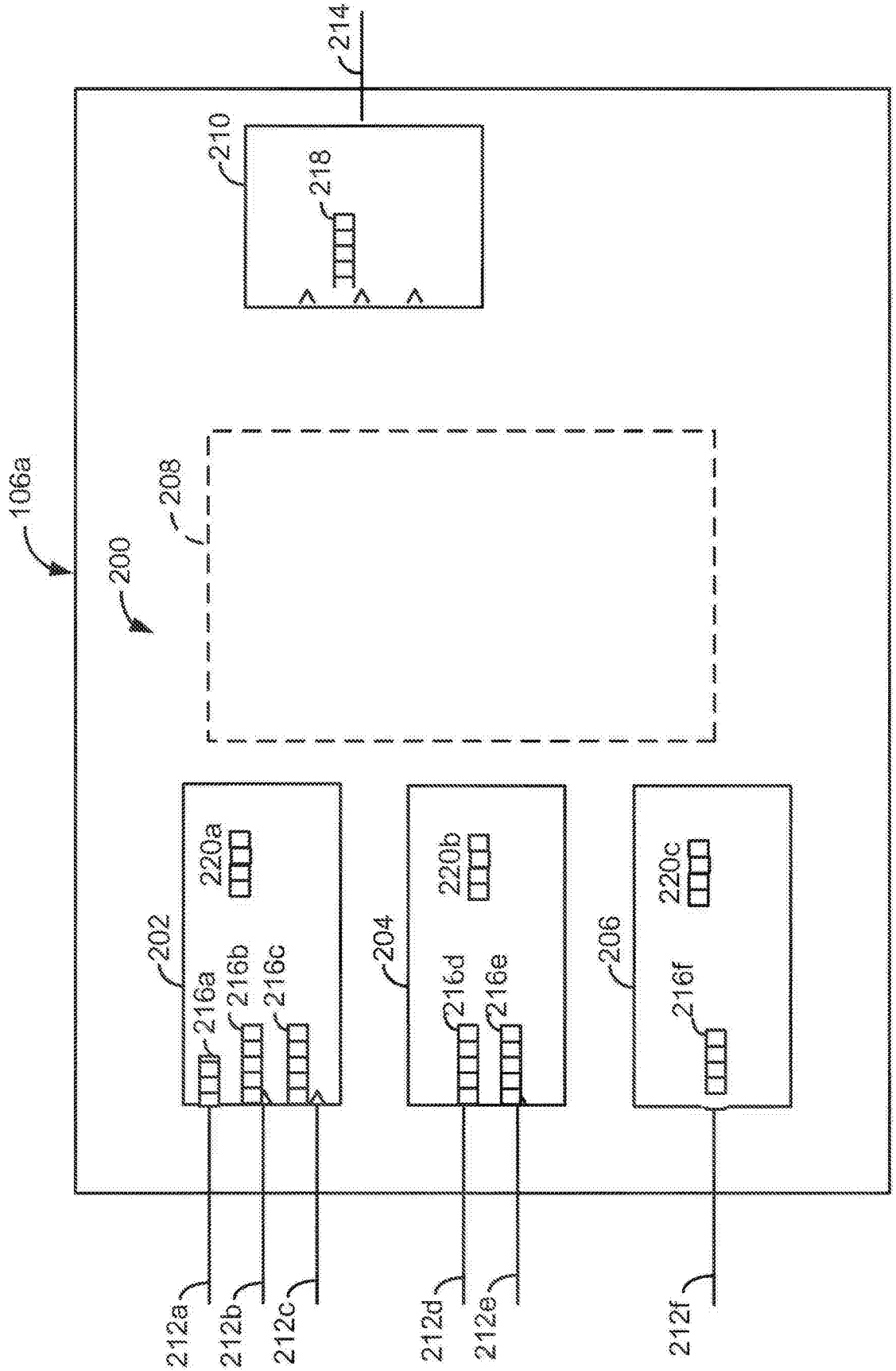

214
210
218
106a
200
208
202
220a
216a
216b
216c
204
220b
216d
216e
206
220c
216f
212a
212b
212c
212d
212e
212f

图2

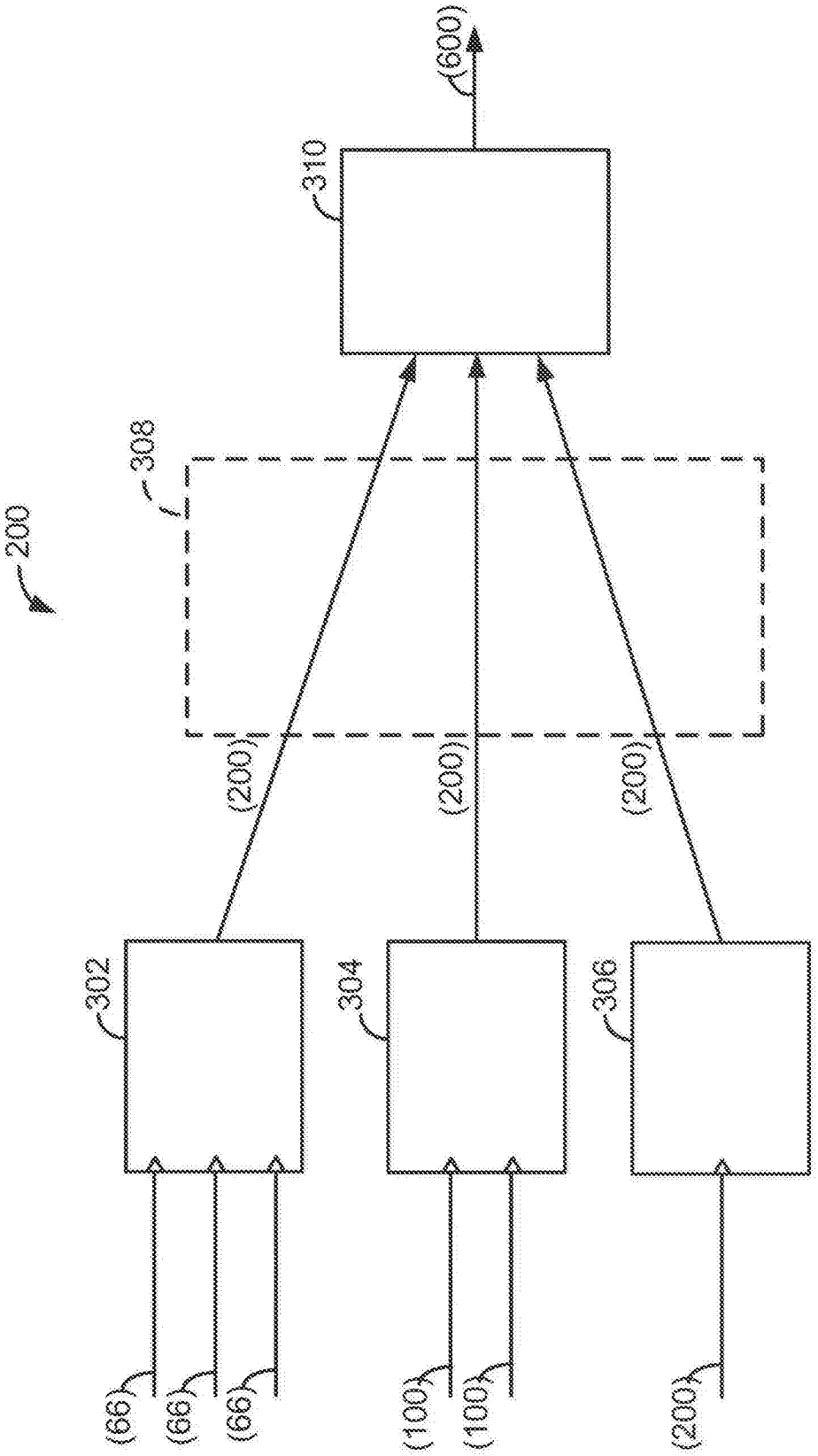


图3A

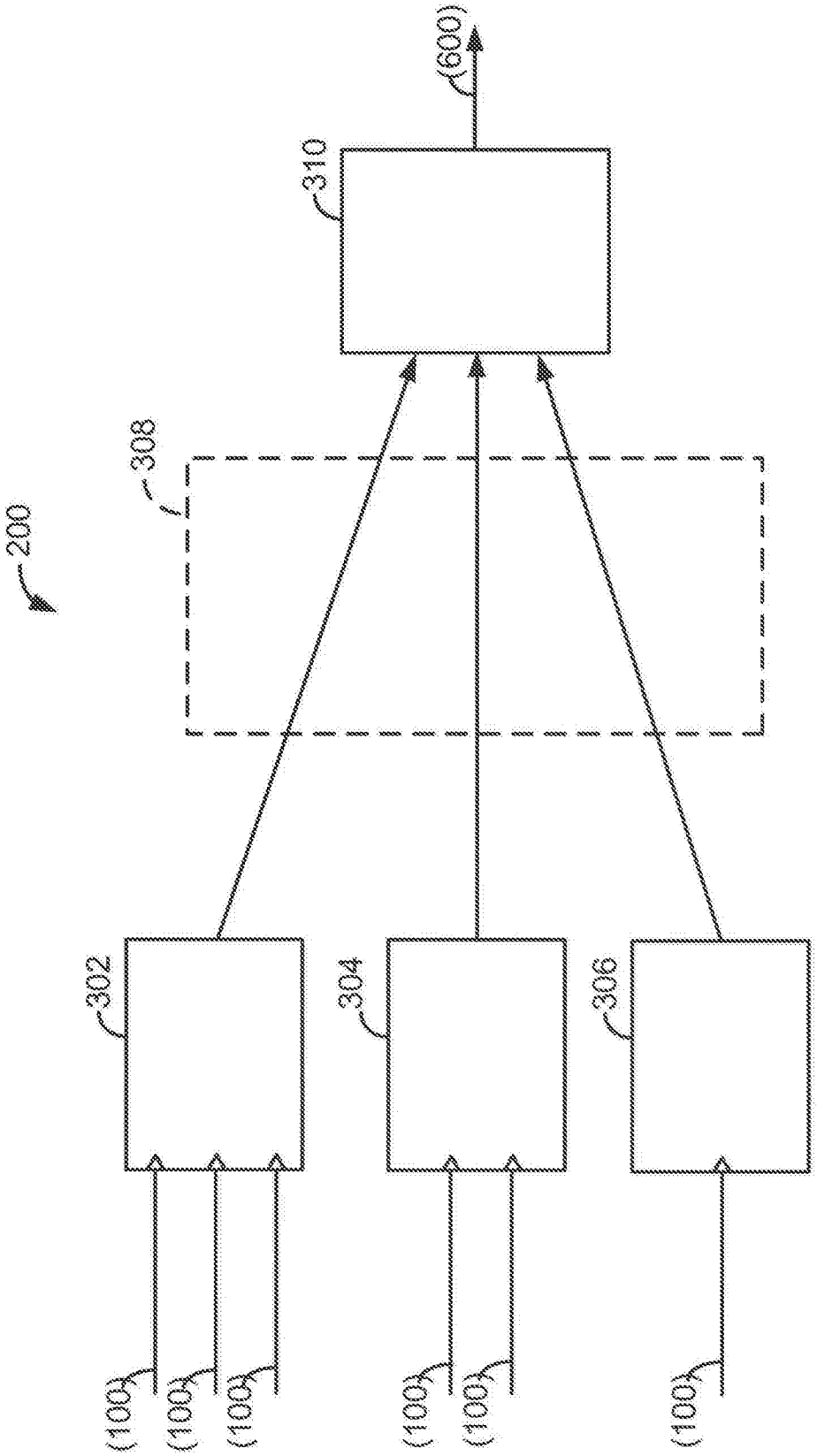


图3B

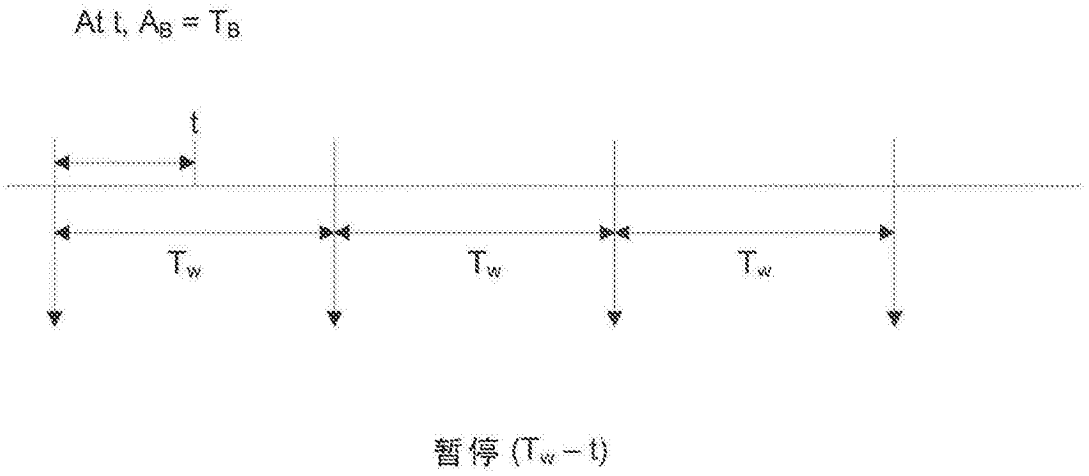


图4A

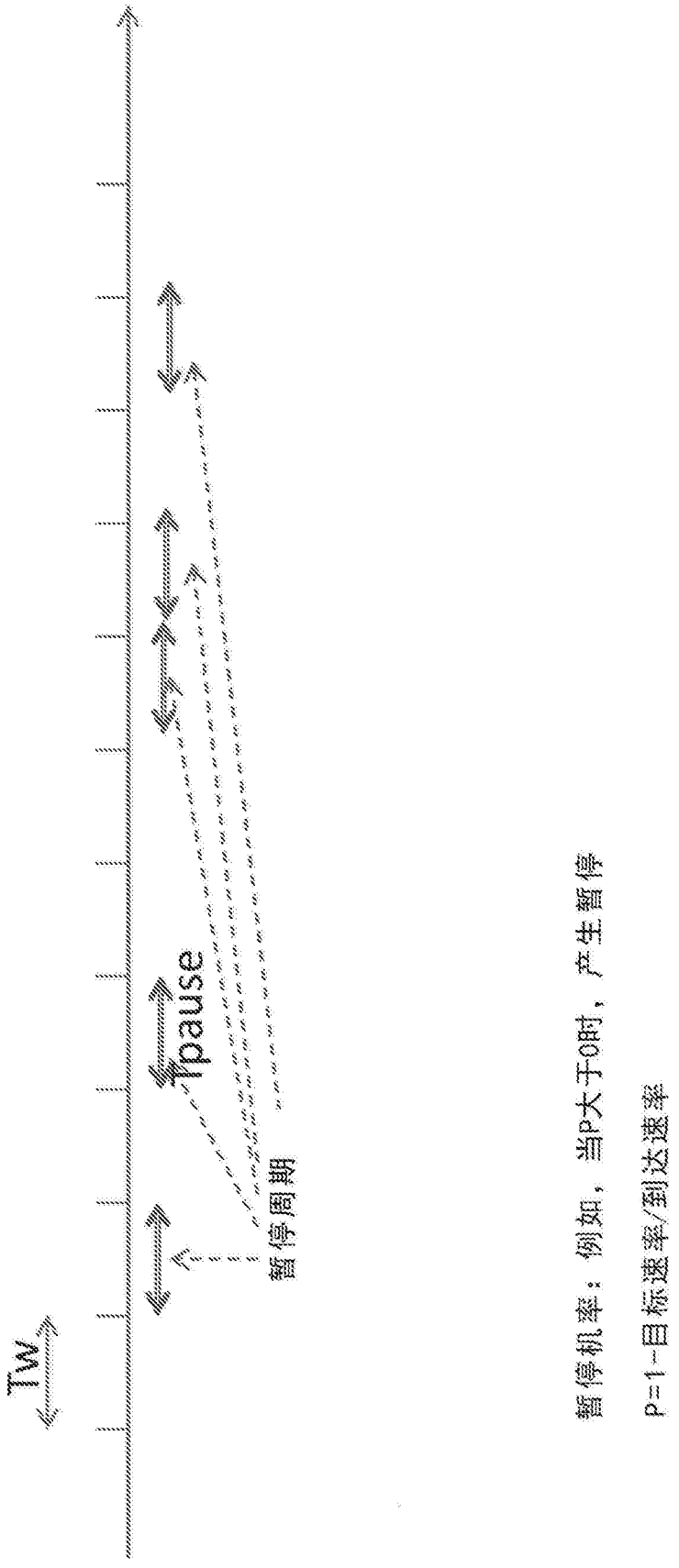


图4B

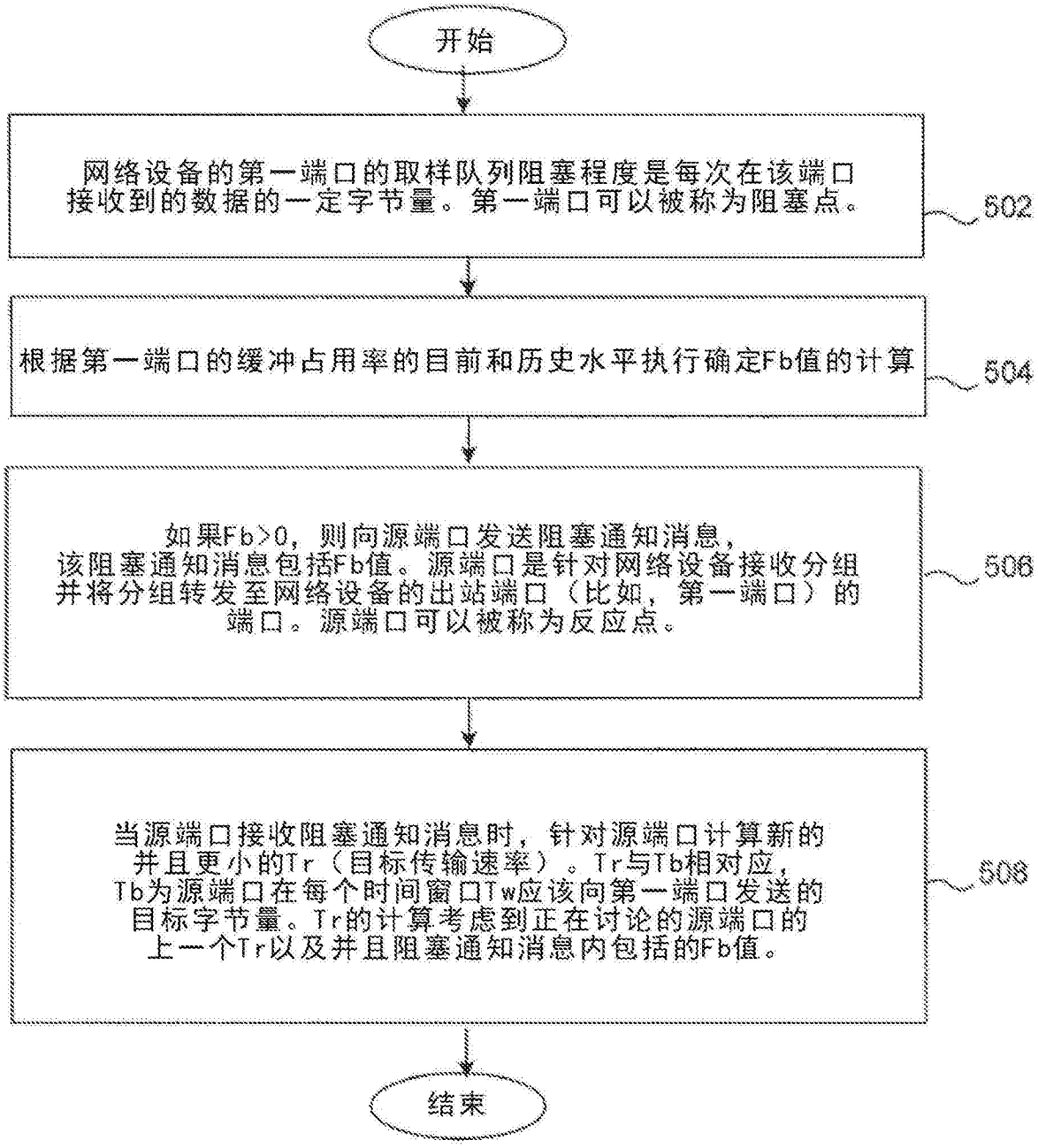

开始
网络设备的第一端口的取样队列阻塞程度是每次在该端口接收到的数据的一定字节量。第一端口可以被称为阻塞点。
502
根据第一端口的缓冲占用率的目前和历史水平执行确定Fb值的计算
504
如果Fb>0，则向源端口发送阻塞通知消息，该阻塞通知消息包括Fb值。源端口是针对网络设备接收分组并将分组转发至网络设备的出站端口（比如，第一端口）的端口。源端口可以被称为反应点。
506
当源端口接收阻塞通知消息时，针对源端口计算新的并且更小的Tr（目标传输速率）。Tr与Tb相对应，Tb为源端口在每个时间窗口Tw应该向第一端口发送的目标字节量。Tr的计算考虑到正在讨论的源端口的上一个Tr以及并且阻塞通知消息内包括的Fb值。
508
结束

图5A

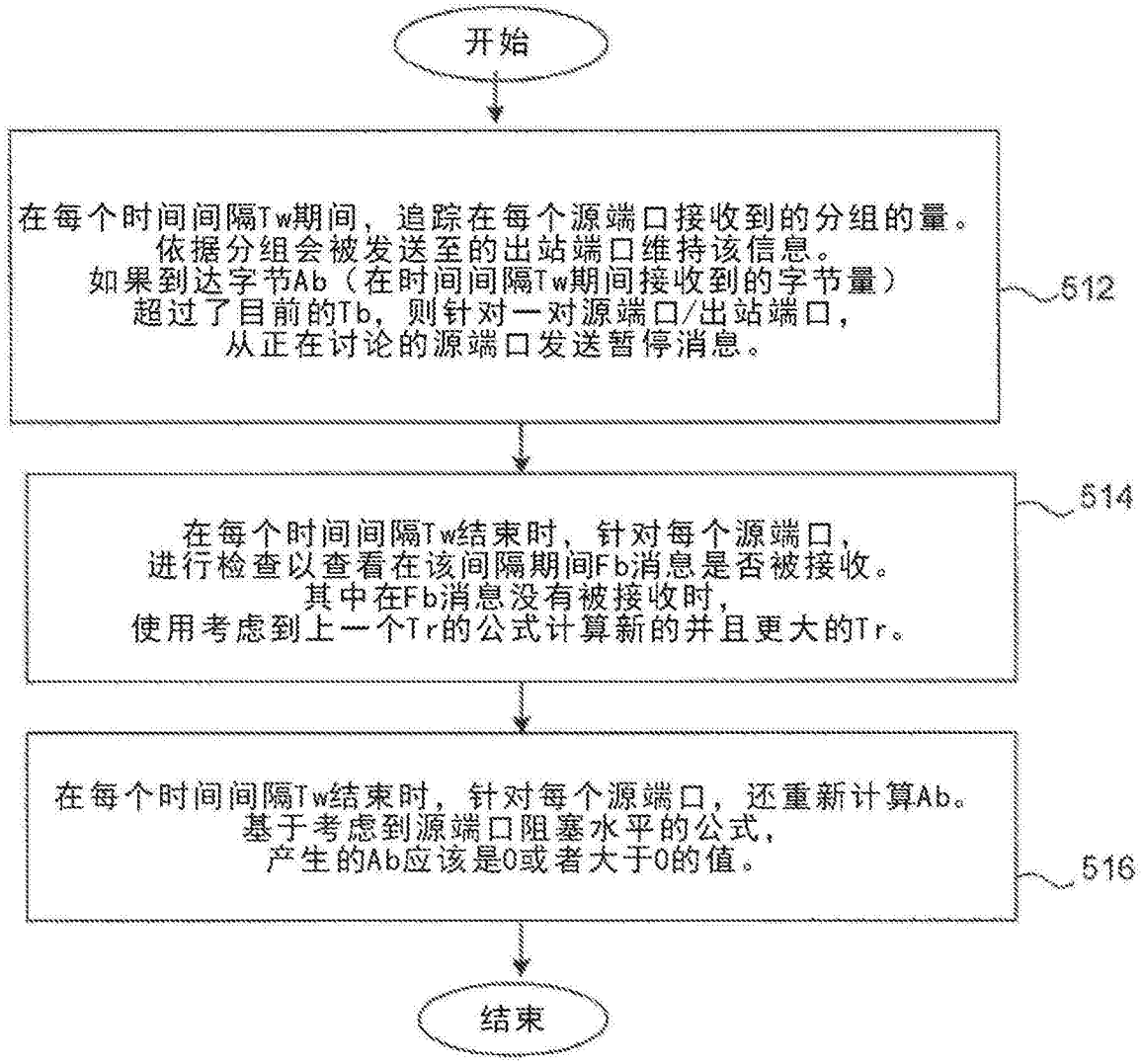

开始
在每个时间间隔Tw期间，追踪在每个源端口接收到的分组的量。
依据分组会被发送至的出站端口维持该信息。
如果到达字节Ab（在时间间隔Tw期间接收到的字节量）
超过了目前的Tb，则针对一对源端口/出站端口，
从正在讨论的源端口发送暂停消息。
512
在每个时间间隔Tw结束时，针对每个源端口，
进行检查以查看在该间隔期间Fb消息是否被接收。
其中在Fb消息没有被接收时，
使用考虑到上一个Tr的公式计算新的并且更大的Tr。
514
在每个时间间隔Tw结束时，针对每个源端口，还重新计算Ab。
基于考虑到源端口阻塞水平的公式，
产生的Ab应该是0或者大于0的值。
516
结束

图5B

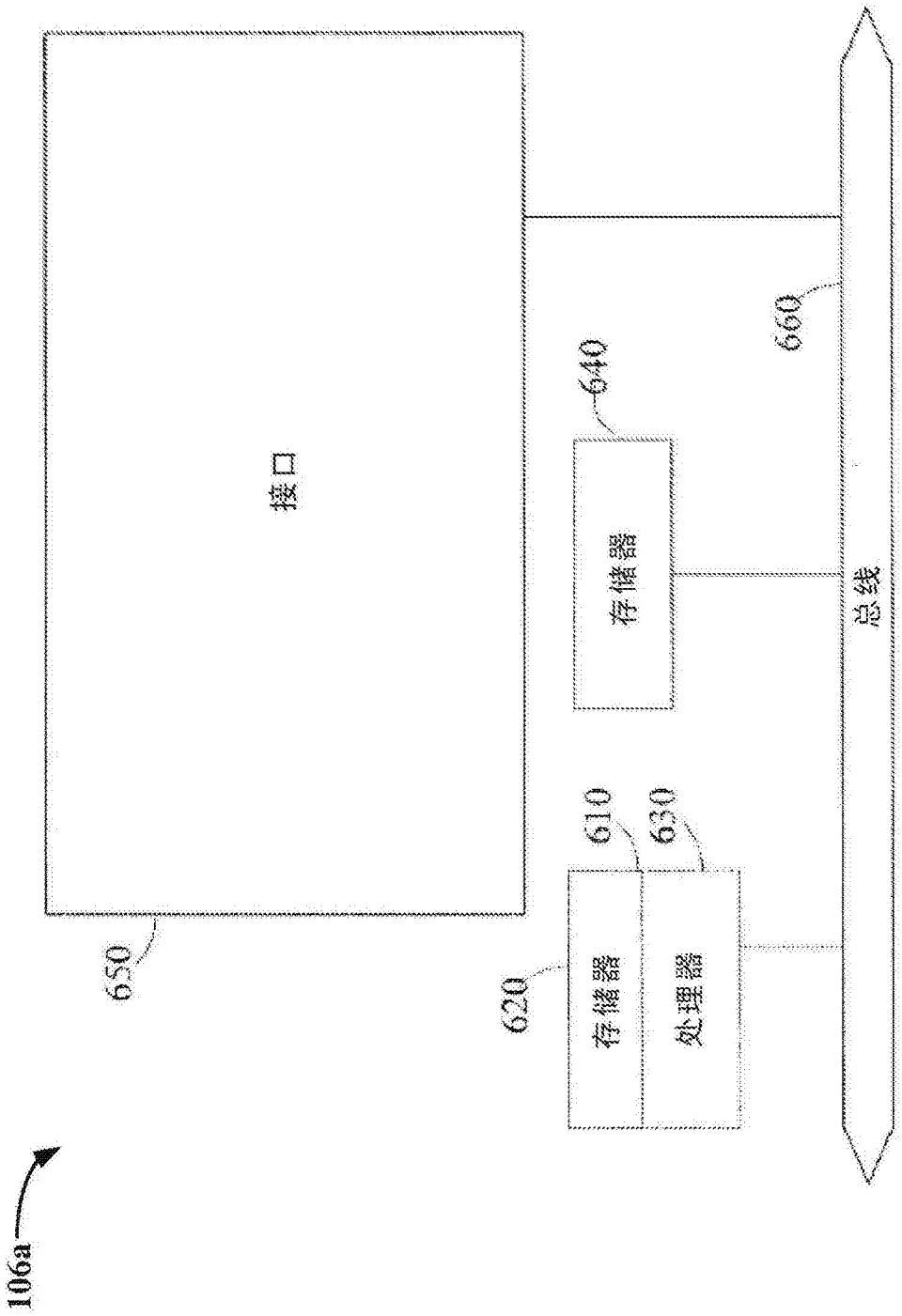


图6