[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
H04L 12/56 (2006.01)
H04L 12/24 (2006.01)

[12] 发明专利说明书

专利号 ZL 200310124076.7

[45] 授权公告日 2006 年 10 月 11 日

[11] 授权公告号 CN 1279728C

[22] 申请日 2003.10.29
[21] 申请号 200310124076.7
[30] 优先权
[32] 2002.10.29 [33] FI [31] 20021921
[71] 专利权人 泰勒比斯股份公司
地址 芬兰埃斯波
[72] 发明人 J·韦内宁
审查员 陈 军

[74] 专利代理机构 中国专利代理(香港)有限公司
代理人 刘 杰 张志醒

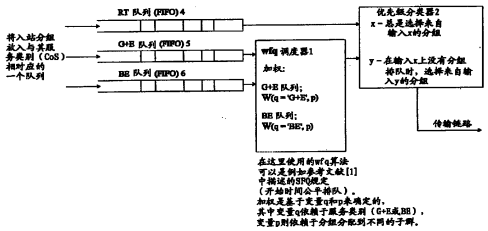
权利要求书 2 页 说明书 7 页 附图 4 页

[54] 发明名称

在分组交换数据流之间调度可用链路带宽的方法和设备

[57] 摘要

本发明涉及一种方法和设备，用于在分组交换数据流之间调度传输链路，以使预期 CoS(服务类别)具有使用数据传输网络的瞬时可用带宽的能力，同时提供了一个保证的最小数据速率(保证数据速率和尽最大努力)，但却没有损害这种没有数据速率保证下限而是具有通过使用瞬时可用带宽来加以实施的服务的类别(尽最大努力)的操作。 本发明则是基于在调度器控制中使用指示服务类别的信息以及指示服务类别内部子群的信息(例如丢弃优先权)。 按照惯例，指示子群的信息只被用于拥塞控制目的。



1．一种用于在不同的分组交换数据流之间调度链路带宽的方法，该方法包括以下步骤：

以固定或可变长度分组的形式来传送数字数据，

使用允许将分组分类为至少两个服务类别的标识符信息来标记分组，

基于指示服务类别的信息，将每一个入站分组单独分类到服务类别特定的并行FIFO队列（4，5，6）之一中，其中队列的数目是每个服务类别一个，

至少一个服务类别利用允许将分组分类到所述服务类别内的至少两个内部子群的标识符信息来为其分组加上标签，

给定服务类别的分组形成数据流，其中不管分组中承载的子群定义标识符信息，都保持分组的转发顺序，以及

系统的一个或多个出站链路的可用带宽是使用基于加权的调度规定、基于优先级的调度规定或其组合在所述服务类别特定的FIFO队列之间进行调度的，

其特征在于：基于优先级的调度规定中的分组特定的优先级值和/或基于加权的调度规定中的加权是使用变量q和p的组合结果来确定的，由此变量q的值依赖于给所述分组所传送的数据流指派的服务类别，而变量p的值则依赖于所述分组归属的子群和/或依赖于在所述分组之前或之后在调度器输入端口上接收的同一服务类别的入站分组的子群中的分类。

2．权利要求1的方法，其特征在于：基于所述分组归属的子群和/或如何在子群之间分配在所述分组之前或之后在调度器输入端口上接收的同一服务类别的入站分组，在基于加权的或基于优先级的调度规定的使用之间进行选择。

3．权利要求1的方法，其特征在于：所述基于加权的调度规定是开始时间公平排队SFQ规定。

4．权利要求1的方法，其特征在于：所述基于加权的调度规定是加权公平排队WFQ规定。

5．一种用于在不同的分组交换数据流之间调度链路带宽的设备，该

设备包括：

用于接收固定或可变长度分组形式的数字数据的装置，

用于读取所述入站分组中承载的并且允许将所述分组分类为至少两个不同服务类别的标识符信息的装置；

用于将所述入站分组单独分类为至少两个不同服务类别的装置；

用于所述服务类别中的每一个服务类别的FIFO分组队列（4，5，6），

用于基于各个服务类别特定FIFO队列中的给定分组的服务类别特定的标识符信息来引导所述给定分组的装置，

用于从给定分组中读取其标识符信息的装置，其中所述标识符信息允许将所述分组分类到被指派给该分组的服务类别的内部子群，

调度器，用于使用基于加权的调度规定、基于优先级的调度规定或其组合来为服务类别特定的FIFO队列（4，5，6）调度该系统的出站链路的可用带宽，以及

用于按照所述调度器确定的分组转发顺序将分组发送至出站链路的装置，

其特征在于：所述设备包括用于在每个分组基础上使用变量q和p的组合结果来确定基于优先级的调度规定中的优先级值和/或基于加权的调度规定中的加权的装置，由此变量q的值依赖于给所述分组所传送的数据流指派的服务类别，而变量p的值则依赖于所述分组归属的子群和/或依赖于在所述分组之前或之后在调度器输入端口上接收的同一服务类别的入站分组的子群中的分类。

6．权利要求5的设备，其特征在于：所述设备包括用于在基于加权的调度规定或基于优先级的调度规定的使用之间进行判定的装置，其中基于所述分组归属的子群和/或如何在子群之间分配在所述分组之前或之后在调度器输入端口上接收的同一服务类别的入站分组进行判定。

7．权利要求5的设备，其特征在于：所述设备包括用于使用开始时间公平排队SFQ规定来执行基于加权的调度规定的装置。

8．权利要求5的设备，其特征在于：所述设备包括用于使用加权公平排队WFQ规定来执行基于加权的调度规定的装置。

在分组交换数据流之间调度可用链路带宽的方法和设备

技术领域

本发明涉及一种根据权利要求1的方法，用于在分组交换数据流之间调度可用链路带宽。

本发明还涉及一种根据权利要求5的设备，用于在分组交换数据流之间调度可用链路带宽。

背景技术

在以下详细描述现有技术和本发明的文本中将会用到下列缩略语：

BE 用于如下应用的服务类别（尽最大努力），所述应用被允许使用网络瞬时可用带宽，但不保证最小数据速率，也不保证分组传输延迟与延迟抖动的上限，

CoS 服务类别，

DSCP 指示分组服务类别的分组报头信息（区分服务编码点），

FIFO 先入先出的排队规定，

G+E 用于如下应用的服务类别（保证速率和尽最大努力），所述应用被允许使用网络瞬时可用带宽并且保证一个最小数据速率，但不保证分组传送延迟与延迟抖动的上限，

QoS 服务质量，

RT 用于如下应用的服务类别（实时），对这些应用而言，分组传送延迟与延迟抖动将会减至最小并且保证了最小数据速率，但是这些应用无法使用网络的瞬时可用带宽，

SFQ 开始时间公平排队，一种加权排队规定[1]，

wfq 加权排队规定，用作一个广义概念的缩略语（加权公平排队），

WFQ 加权公平排队，一种特定的加权排队规定[1]，

WRED加权拥塞避免算法[3,4]（加权随机早期检测）。

在分组交换网络中，一方面由于使用数据网络服务的各种应用的需要，另

一方面则由于电信服务供应商与其客户的QoS等级协定，因此通常有利的是将所要传送的数据分组分为不同的服务类别（CoS）。例如在与常规电话连接相结合的时候，有必要使得应用所需要的带宽在预定时间可用并且伴随足够低的数据传送延迟和延迟抖动。在一个电话应用中，在网络负载低的状态下，用户可以使用暂时较高的链路带宽，但用户并不会从中获益。与此相反，举例来说，在下载万维网页面的过程中，如果可以使用网络的所有临时可用带宽，那将是极为有利的。

接下来对一种情况进行检查，其中电信服务供应商提供了以下服务类别：

- RT（实时）：用于那些保证最小数据速率的应用的服务类别，并且分组传输延迟和延迟抖动都减至最小，即使施加到通信网络上的业务量负载暂时处于很低等级，也不会进行任何尝试来增加提供给指定应用的瞬时数据速率。

- G+E（保证速率和尽最大努力）：用于那些确保一个给定最小数据速率的应用的服务类别，并且附加提供数据传输系统的所有瞬时可用带宽，以供所述应用使用。然而并没有为分组传输延迟与延迟抖动的保证上限给出任何约定。

- BE（尽最大努力）：用于那些被分配使用网络瞬时可用带宽但没有任何保证的最小数据传输速率的应用的服务类别。此外也没有为指定给分组传送延迟和延迟抖动的上限做出任何约定。

图1显示了用于在代表以上列举的服务类别的数据流之间调度公共数据传输链路带宽的常规方案。而图1所示系统的功能则如下所示：

- 指派给指定分组的服务类别可以通过分组中传送的报头信息（例如DSCP，也就是区分服务编码点[2]）来识别。

- 将接收到的分组调度到相应的服务类别特定的FIFO队列（RT、G+E和BE队列）。

- 把归入服务类别G+E的各个分组进一步指派到CoS的一个内部子群，至少允许判定所述分组属于被约定为保证最小速率的业务量部分（下文称为G部分），还是属于超出了保证最小速率的业务量部分（下文称为E部分）。举例来说，可以借助DSCP[2]中传送的优先级信息（例如丢弃优先权）来表示将一个分组指派到给定子群。在队列拥塞而需要判定应该将拥塞控制策略操作应用于哪些分组的时候将会用到所述子群信息。这种方法的一个实例是WRED（加

权随机早期检测）拥塞控制方法[3,4]。

- 使用加权调度规定（例如SFQ[1]）来为RT队列1、G+E队列5和BE队列6的数据流调度链路带宽，因此，相对于G+E和BE队列的加权（W_{G+E}和W_{BE}）而把RT队列4的加权（W_{RT}）选得很大，以使类别RT的业务量在所有条件下都可以使用为其分配的最小带宽，同时相对于BE队列6的加权而把G+E队列5的加权选得很大，以使类别G+E的业务量在所有状态下都被准许使用所述的保证最小数据速率。

- 假设在调度器之前，类别RT与类别G+E的G部分的业务量都是带宽受限的。

图2显示了在代表以上列出的服务类别的数据流之间调度公共链路带宽的另一种常规方案。图2所示系统的功能不同于图1所示系统的功能，其差别在于：图2所示系统以先于G+E 5和BE队列的优先级来为RT队列4调度链路带宽。由于在把RT队列4的业务量输入调度器的输入端口之前，将其假设为带宽受限，因此可能将优先级调度规定运用于RT队列4。

图1和2所示调度方案的问题在于：在调度器中，类别BE的业务量受到加权为W_{G+E}的类别G+E中E部分业务流的竞争，其中，相对于类别BE的加权W_{BE}而言，加权W_{G+E}的值是基于类别G+E的保证最小数据速率（为G部分所保证的）来选择的。因此，当类别G+E中E部分业务流也在同一时间尝试使用同一瞬时空闲链路带宽时，类别BE使用瞬时可用带宽的能力将会是非常差的。而这与类别BE的业务量的基本思想恰恰相反，类别BE的业务量的基本思想是确保没有数据传输速率下限，取而代之的是提供用户充分使用瞬时可用带宽的服务。

图3描述的典型事例（a）和（b）说明了这种情况。图中，典型事例（a）对应的是：当以最大可能数量发送来自各个服务类别的业务量时，在不同服务类别的业务流之间进行带宽共享。在这里，类别G+E的业务量使用的带宽数量（B_{G+E}）与类别BE所用带宽数量（B_{BE}）的比值是W_{G+E}/W_{BE}。而典型事例（b）对应的则是：除了类别RT的业务量所使用的带宽部分小于为所述类别保留的带宽，同时尽可能多地传送类别G+E和BE的业务量之外，当为类别RT以及类别G+E中G部分的业务量所保留的带宽与典型事例（a）中相同时，在不同服务类别的数据流之间进行带宽共享。在这种情况下，带宽使用率为$B_{G+E}/B_{BE}=W_{G+E}/W_{BE}$。从典型事例（b）中可以明显看出，类别RT的业务量仍未

使用的带宽部分几乎全部提供给了类别G+E中的E部分。

必须注意的是，由于不允许调度器改变类别G+E的业务流中的分组转发顺序，因此无法将类别G+E中G部分与E部分划分为可以给予相互独立的调度加权的不同队列。

发明内容

本发明的一个目的是克服上述现有技术的缺陷并提供一种全新类型的方法和设备，用于在不同分组交换数据流之间调度瞬时可用带宽。更具体的说，本发明涉及一种方法，该方法能够实现一个调度器，以便在类别G+E中E部分业务流与类别BE的业务流之间以预期比率（例如1：1）来分配瞬时可用带宽。

本发明的目的是通过使用调度器操作控制中的子群信息（例如丢弃优先权）来实现的。在现有技术中，仅仅将子群信息用在拥塞控制系统中（例如WRED）。然而，根据本发明的调度方法并不排斥将子群信息（例如丢弃优先权）用在拥塞控制系统中。

更具体的说，根据本发明的方法，其特征即为权利要求1的特征部分所描述的内容。

此外，根据本发明的设备，其特征即为权利要求5的特征部分所描述的内容。

本发明允许使用一种在类别G+E中E部分数据流与类别BE数据流之间以预期比率（例如1:1）分配可用剩余带宽的方式来实现调度引擎，从而提供了超越现有技术的显著优点。因此，可以提供一种服务类别（G+E），以便能够使用数据传输网络的瞬时可用带宽，同时确保一个保证最大数据速率，而不损害这种不具有数据传输速率保证下限而是具有通过使用瞬时可用带宽所实现的服务的类别（例如BE）中的服务质量。

附图说明

下文中通过参考附图并且根据例示实施例而对本发明进行更为详细的描述，其中

图1显示了用于为上述服务类别（RT，G+E，BE）的数据流调度公共数据传输链路带宽的现有技术系统的框图；

图2显示了用于为上述服务类别的数据流调度公共数据传输链路带宽的另一种现有技术系统的框图；以及

图3显示了在不同服务类别的数据流之间划分瞬时可用带宽的两个典型事

例（a）和（b）。典型事例（a）中传送的是每个服务类别上的最大量的业务量。而在典型事例（b）中，分别为类别RT和类别G+E中G部分保留的带宽部分与典型事例（a）中相同，但是类别RT的业务量使用的带宽小于为所述类别保留的带宽上限，同时类别G+E与BE的业务量是以最大带宽来传送的；以及

图4显示了一个根据本发明而在上述服务类别的数据流之间调度公共数据传送链路带宽的系统的框图。

具体实施方式

后续描述中将对根据本发明的方法的理论基础进行说明。

在基于加权的调度方法中，在调度器输入端口接收的分组是用一个转发顺序标识符（例如SFQ方法[1]中的Start_tag）来标记的，其中所述标识符声明了调度转发所述分组的时刻。因此，所要转发的第一个分组是具有这样一个顺序标识符的分组，所述顺序标识符具有一个指示最早转发时刻的数值。传输顺序指示无需与实际时间同步，而是只要各个分组的转发标识符相对彼此都处于正确的传输顺序即可。

在为一个从给定服务类别队列接收的分组产生转发顺序标识符的过程中，分组加权是根据各个服务类别来指派的。如果队列J1具有高于队列J2的加权，那么相对于队列J2的相应转发标识符序列而言，队列J1的连续分组转发标识符序列具有这样一种特性，那就是队列J1得到了较大部分的调度器输出容量。

在基于优先级的排序方法中，将一个优先级数值指派给在调度器1的输入端口接收的各个分组。分组优先级数值则确定了接下来将要转发哪个分组。

然而在根据本发明的方法中，指派给一个分组的优先级数值或者用于产生分组转发顺序标识符的加权不但分别依赖于分组服务类别（下文用符号q表示），而且还依赖于所述分组和/或处于同一服务类别中的所述分组之前或之后的分组的子群信息（下文用符号p表示，例如，这种优先级信息可以是分组丢弃优先权[2]），如图4所示。由于这些分组包含在序列发生器1前面的服务类别特定队列中，因此可以了解的是，在由所述队列包含的分组数量所确定的限度内，预定在不久的将来将会输入到调度器中的分组所携带的究竟是什么类型的子群信息。

在根据本发明的方法中，子群信息中的一条或多条还可以确定：对于一个给定分组的调度判定是使用基于加权还是基于优先级的调度机制来做出的。

与此相反，现有技术的系统则是把子群信息（p）用于拥塞控制操作而不

是进行调度。

接下来描述的是根据本发明的调度器的实施例，所述描述针对的是调度器使用SFQ算法[1]来调度类别G+E和BE的业务流的功能。在这里所论述的本发明的示范性实施例中，分组特定的加权是基于所研究的分组所归属的子群来进行选择的。对类别G+E中的分组i以及类别BE中的分组j而言，分别如下计算其转发顺序标识符（S_{G+E}（i）和S_{BE}（j））：

$$S_{G+E}(i)=\max\{v,S_{G+E}(i-1)+L(i-1)/W(q,p)\} \quad (1)$$

$$S_{BE}(j)=\max\{v,S_{BE}(j-1)+L(j-1)/W(q,p)\} \quad (2)$$

其中L（i-1），L（j-1）是以字节为单位的分组大小，例如，变量p和q确定了加权W的值，因此变量q依赖于指派给所检查的分组（i或j）的服务类别（G+E或BE），而变量p则依赖于指派给所检查的分组（i或j）的子群，v则是所传送分组的转发顺序标识符（虚拟时间戳）。

转发顺序标识符的值是在调度器的服务类别特定输入端口接收到分组的时候计算得到的，即使应该改变v的值，也不会后来对转发顺序标识符的值进行更新。在入站分组中，所要转发的第一个分组是一个具有较低的转发标识符的值的分组（i或j）。

在这里论述的典型事例中，假定如下所述来选择特定子群的加权：

如果类别G+E的分组属于G部分，则W（q＝‘G+E’，p＝‘G’）＝W_G，

如果类别G+E的分组属于E部分，则W（q＝‘G+E’，p＝‘G’）＝W_E，

不管子群信息怎样，类别BE的分组都具有相同加权，即W（q＝‘BE’，p：不相关）＝W_{BE}，

在下文中，一个简单的测试或模拟足以验证以下事实：如果系统在给定时段传送平均数量W_G的G部分分组字节（或比特），那么在这个时段，系统会传送平均数量W_{BE}的类别BE的分组字节（或比特），如果系统在给定时段传送平均数量W_E的E部分分组字节（或比特），那么系统也会传送平均数量W_{BE}的类别BE的分组字节（或比特）。为了简化情况，可以假设所有分组都具有同等大小，由此以上描述是不仅对分组字节来说是成立的，对完整的分组而言，以上描述同样是成立的。

通过适当选择用于加权W_E和W_{BE}的值，可以实现一种调度设备，以便在类别G+E的E部分数据流与类别BE的数据流之间以预期比率来分配可用带宽。

上述情况的一个替换实施例可以通过为加权W_G指派一个无限值来实现。在实践中，这意味着G部分分组是基于优先权而不是使用SFQ规定来进行调度的。然后，无论在为类别BE的数据流提供了服务的输出端口上的分组队列的转发顺序指示是什么，都会以区分优先级的方式来转发一个接收于调度器输入端口的分组，其中所述端口是为那些处于类别G+E的队列中的子群G的分组而指派的。由于将类别G+E中G部分业务量假设为带宽受限，因此上述操作是切实可行的。

参考文献：

[1] Pawan Goyal, Harric M. Vin, Haichen Cheng. 的“开始时间公平排队：用于综合业务分组交换网的调度算法（Start-time Fair Queuing: A scheduling Algorithm for Integrated Services Packet Switching Networks），美国奥斯汀德州大学计算机科学系的技术报告 TR-96-02 。

[2] Brace Davie, Yakov Rekhter 的“MPLS 技术与应用（MPLS Technology and Applications），美国加州科学出版社（ Academic Press）, 2000, (www.academicpress.com)。

[3] Sally Floyd, Van Jacobson的“用于拥塞避免的随机早期检测网关（Random Early Detection Gateways for Congestion Avoidance）”，美国加州的加州大学劳伦斯伯克力实验室 ，1993。

[4] 从 http://www.junipef.net/techcenter/techpapers/200021-01 .html可获得的关于WRED规定的白皮书。

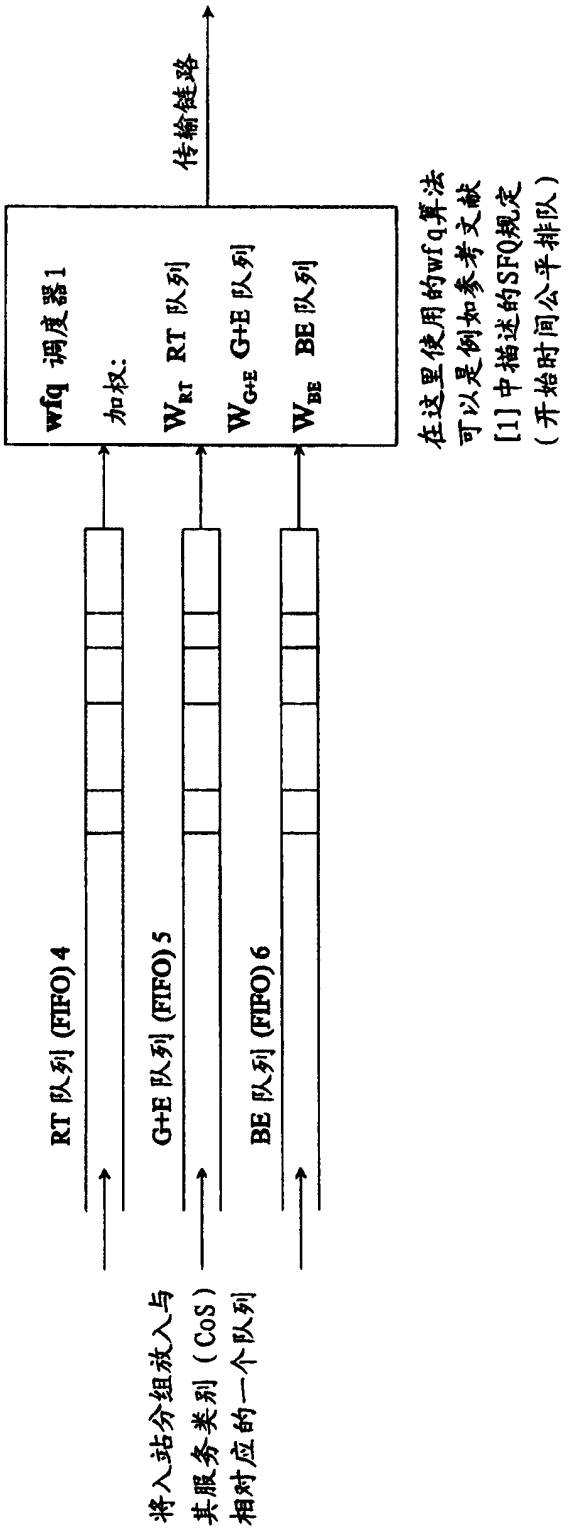

wfq 调度器 1
加权:
W_{RT} RT 队列
W_{G+E} G+E 队列
W_{BE} BE 队列
传输链路
RT 队列 (FIFO) 4
G+E 队列 (FIFO) 5
BE 队列 (FIFO) 6
将入站分组放入与
其服务类别（CoS）
相对应的一个队列
在这里使用的wfq算法
可以是例如参考文献
[1]中描述的SFQ规定
（开始时间公平排队）

图 1

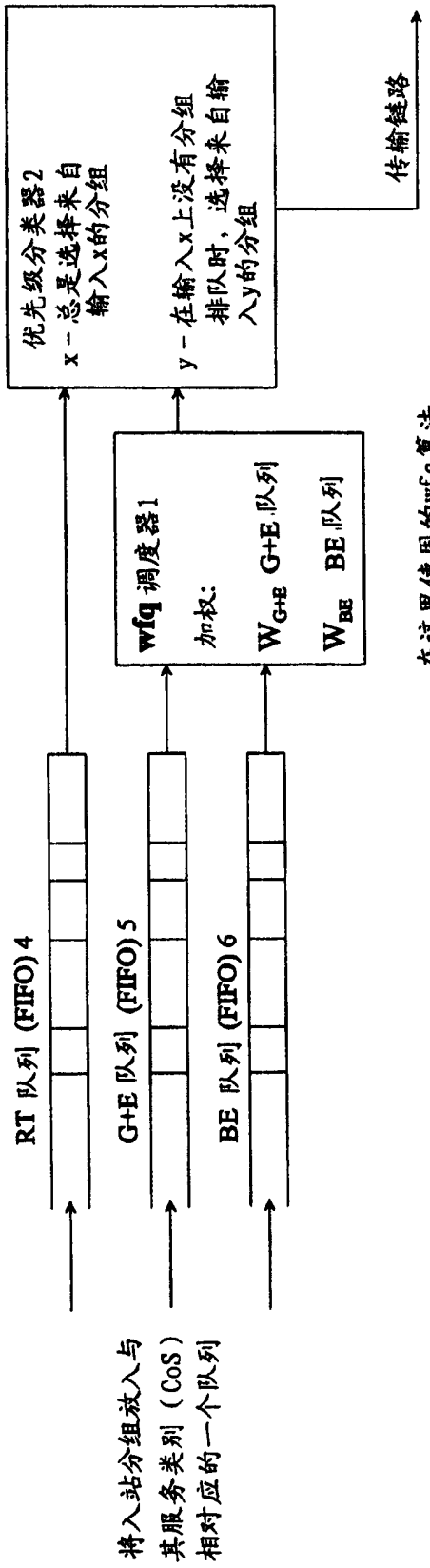

优先级分类器2
x－总是选择来自输入x的分组
y－在输入x上没有分组排队时，选择来自输入y的分组
传输链路
wfq 调度器1
加权:
W_{G+E} G+E 队列
W_{BE} BE 队列
RT 队列 (FIFO) 4
G+E 队列 (FIFO) 5
BE 队列 (FIFO) 6
将入站分组放入与其服务类别（CoS）相对应的一个队列
在这里使用的wfq算法可以是例如参考文献[1]中描述的SFQ规定（开始时间公平排队）

图 2

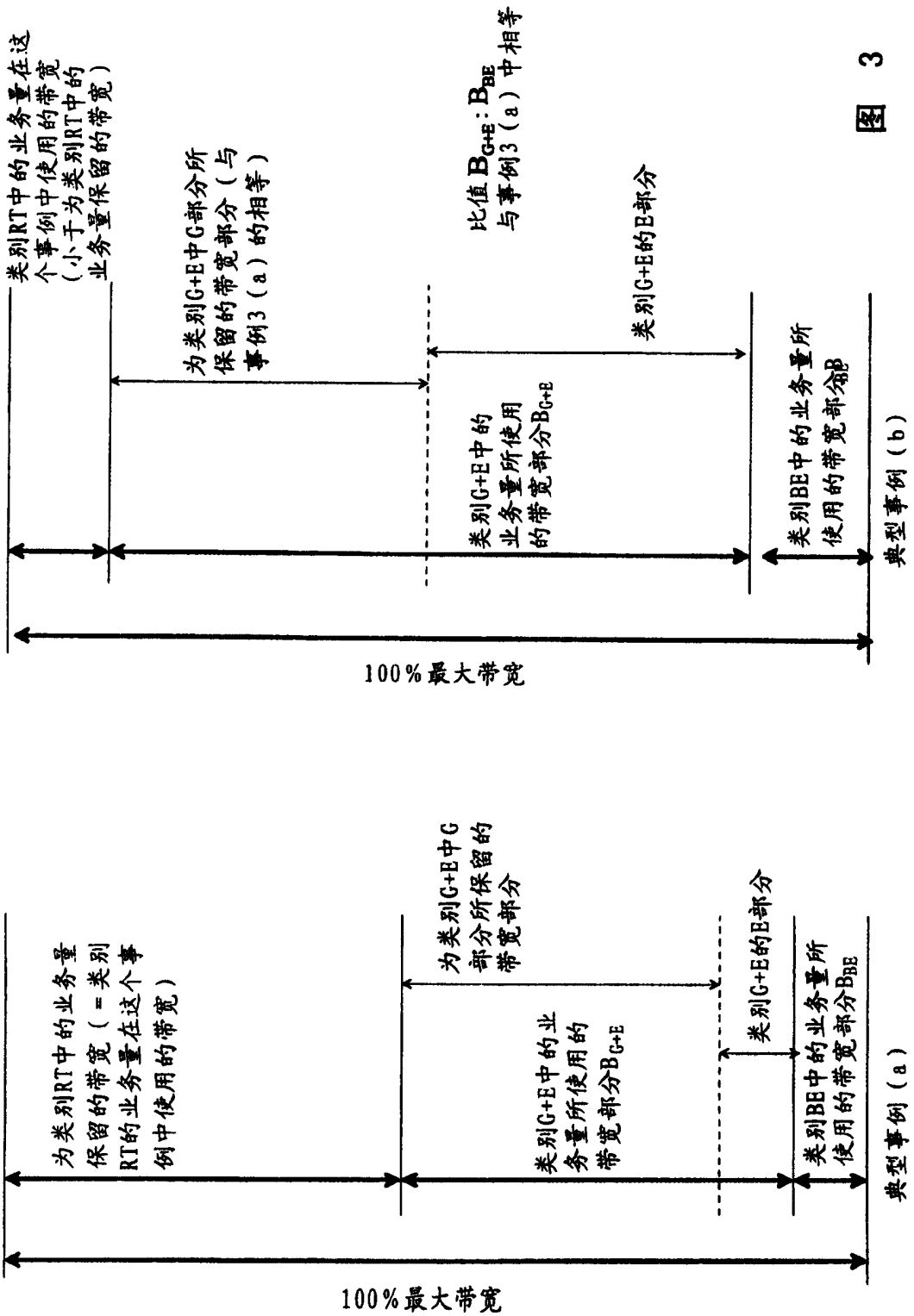

类别RT中的业务量在这个事例中使用的带宽（小于为类别RT中的业务量保留的带宽）
为类别G+E中G部分所保留的带宽部分（与事例3（a）的相等）
比值 $B_{G+E}:B_{BE}$ 与事例3（a）中相等
类别G+E中的业务量所使用的带宽部分B_{G+E}
类别G+E的E部分
类别BE中的业务量所使用的带宽部分B_{BE}
典型事例（b）
100%最大带宽
为类别RT中的业务量保留的带宽（=类别RT的业务量在这个事例中使用的带宽）
为类别G+E中G部分所保留的带宽部分
类别G+E中的业务量所使用的带宽部分B_{G+E}
类别G+E的E部分
类别BE中的业务量所使用的带宽部分B_{BE}
典型事例（a）
100%最大带宽

图 3

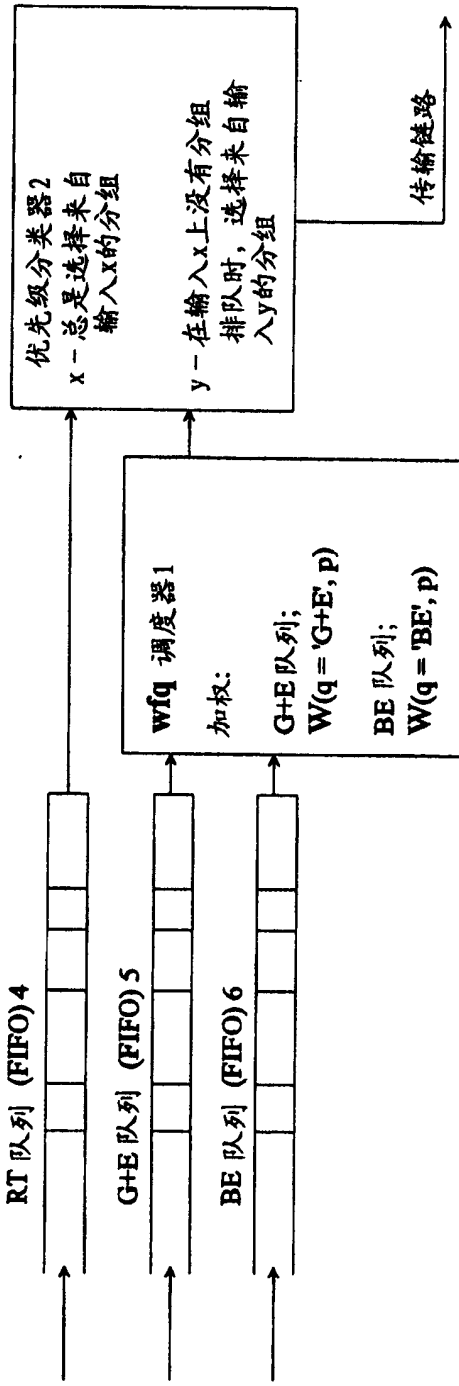

将入站分组放入与
其服务类别（CoS）
相对应的一个队列
RT 队列 (FIFO) 4
G+E 队列 (FIFO) 5
BE 队列 (FIFO) 6
wfq 调度器1
加权:
G+E 队列;
W(q = 'G+E', p)
BE 队列;
W(q = 'BE', p)
优先级分类器2
x－总是选择来自
输入x的分组
y－在输入x上没有分组
排队时，选择来自输
入y的分组
传输链路
在这里使用的wfq算法
可以是例如参考文献[1]
中描述的SFQ规定
（开始时间公平排队）。
加权是基于变量q和p来确定的，
其中变量q依赖于服务类别（G+E或BE），
变量p则依赖于分组分配到不同的子群。

图 4