[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
H04L 12/56

[12] 发明专利申请公开说明书

[21] 申请号 02800572.4

[43] 公开日 2003 年 12 月 17 日

[11] 公开号 CN 1462528A

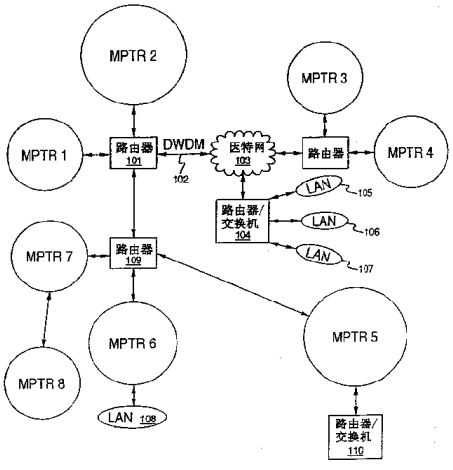
[22] 申请日 2002.3.8 [21] 申请号 02800572.4
[30] 优先权
[32] 2001. 3. 8 [33] US [31] 60/274,621
[32] 2001. 3.12 [33] US [31] 60/275,338
[32] 2002. 3. 7 [33] US [31] 10/094,035
[86] 国际申请 PCT/US02/07122 2002.3.8
[87] 国际公布 WO02/073884 英 2002.9.19
[85] 进入国家阶段日期 2002.11.8
[71] 申请人 A·迈凯提库
地址 美国加利福尼亚州
共同申请人 N·维坚 W·J·图奥依
[72] 发明人 A·迈凯提库 N·维坚
W·J·图奥依
[74] 专利代理机构 北京纪凯知识产权代理有限公司
代理人 沙 捷

权利要求书 3 页 说明书 18 页 附图 10 页

[54] 发明名称 一种用于带宽分配跟踪的方法和系统

[57] 摘要

一种在城域网中保持网络上已分配带宽的精确总量的方法和系统。 将多个进入的分组分配给城域网交换机中对应的多个队列。 使用一种公平仲裁方案，配置对应队列以特定的输出速率清空。计算每个对应队列的结束时间。 该结束时间描述了该对应队列将使用该输出速率而清空的时间。多个队列按照其各自的结束时间而分成多个组。初始组包括所具有的结束时间在第一时间增量处指示出清空状态的那些队列。 第二初始组包括所具有的结束时间在第二时间增量处指示出清空状态的那些队列，第二时间晚于第一时间增量，依此类推。 通过跟踪所有多个组中的保留速率之和来确定网络上的已分配带宽数量。 第一时间增量、第二时间增量等等都对应调度时钟而标以索引。 初始组因此指示将在调度时钟的下一时间增量处具有清空状态的那些队列。 已分配带宽数量的确定可以实时完成。



ISSN1008-4274

1．一种方法，用于在一网络中保持该网络上的已分配带宽的精确总量，所述方法包括如下步骤：

a) 把多个进入的分组分配给对应的多个队列，所述对应的队列配置为在输出速率上清空；

b) 对于每个对应的队列计算结束时间，该结束时间描述了要使用所述输出速率而将所述对应的队列清空的时刻；

c) 把所述多个队列归组到至少一个第一组和一个第二组中，其中所述第一组包括所具有的结束时间在第一时间增量处指示出清空状态的那些队列，而所述第二组包括所具有的结束时间在第二时间增量处指示出清空状态的那些队列，该第二时间增量晚于该第一时间增量；以及

d) 通过计算所述第一组和所述第二组中的队列总数来确定已分配带宽的数量。

2．一种用于在网络中分配带宽而同时保持所述网络上已分配带宽的精确总量的系统，包括：

分别的多个队列，其被配置来接收对应的多个数据流的进入的分组，所述队列包括于连接到所述网络上的 MPS（城域分组交换机）中；

一个调度时钟，其被配置为与链路上测量到的拥塞相关地递增；和

一个数据库，用于把所述多个队列归组到至少一个第一组和一个第二组中，其中所述第一组包括所具有的结束时间在第一时间增量处指示出清空状态的那些队列，而所述第二组包括所具有的结束时间在第二时间增量处指示出清空状态的那些队列，该第二时间增量晚于该第一时间增量；所述 MPS 被配置来通过计算所述第一组和所述第二组中的队列总数，从而确定已分配带宽的数量。

3．如权利要求 1 或 2 所述的网络，其中，所述网络包括一个分组数据网络。

4．如权利要求 1 所述的方法或如权利要求 2 所述的系统，其中使对应的加权与所述多个队列中的每一个相关联，并且通过计算所述第一组和所述第二组中的所述队列总数来确定已分配带宽的数量的已分配加权。

5．如权利要求 1 所述的方法，其中，所述第一时间增量和所述第二时间增量对应于一个调度时钟而标以索引，并且所述第一组表示将在该调度时钟的下一增量处具有清空状态的那些队列。

6．如权利要求 5 所述的方法或如权利要求 2 所述的系统，其中，所述调度时钟根据在一个节点处测量的相对拥塞量而递增。

7．如权利要求 1 所述的方法，其中，来自步骤 d)的总数与对应于队列数量的相应加权一起被保持在一个数据库中。

8．如权利要求 2 所述的系统，其中，队列总和的结果和那些队列的相应加权一起被保持在数据库中。

9．如权利要求 1 所述的方法或如权利要求 2 所述的系统，其中，当对应的队列接收到新的分组时，计算每一所述对应的队列的结束时间。

10．如权利要求 1 所述的方法或如权利要求 2 所述的系统，其中，实时地确定已分配带宽的数量。

11．如权利要求 1 所述的方法或如权利要求 2 所述的系统，其中，所述已分配带宽数量被用来确定未分配带宽数量，从而得以进行所述未分配带宽的分配而同时保持所述已分配带宽的服务质量。

12．如权利要求 1 所述的方法或如权利要求 2 所述的系统，其中，未分配带宽数量被描述为单个项，并且所述项被广播给许多其它节点以便所述其它节点能够形成它们的输出业务。

13．如权利要求 2 所述的系统，其中，对于每个组使用一对计数器来实现所述数据库，每一对都具有对应于该组中队列数目的一个第

一计数器和对应于该组中所述队列的各自加权的一个第二计数器。

14．如权利要求 1 所述的方法或如权利要求 2 所述的系统，其中，所述队列为虚拟队列，其中所述虚拟队列中的每一个都保持对相应数据流储备的跟踪而不必物理地缓冲所述数据流。

15．如权利要求 1～14 所述的网络，其中，所述网络是一个环布局技术城域网。

16．如权利要求 1 所述的方法，其中，所述环布局技术城域网包括一个以太网通信信道。

一种用于带宽分配跟踪的方法和系统

相关申请的交叉参考

本申请要求2001年3月8日申请的美国专利申请号No.60/274,621的在U.S.C.§119(e)下的常规优先级和优先级，其标题为“A METHOD AND SYSTEM FOR BANDWIDTH ALLOCATION TRACKING IN AN ASYNCHRONOUS METRO PACKET TRANSPORT RING NETWORK”，在此结合引述其全部内容作为参考资料。本申请还要求2001年3月12日申请的美国专利申请号No.60/275,338的在U.S.C.§119(e)下的常规优先级和优先级，其标题为“CASPIAN ARCHITECTURE”，在此结合引述其全部内容作为参考资料。本申请还要求2002年3月7日申请的美国专利申请号为____的常规优先级并且是其延续，其标题为“A METHOD AND SYSTEM FOR BANDWIDTH ALLOCATION TRACKING IN A PACKET DATA NETWORK”，在此结合引述其全部内容作为参考资料。

技术领域

本发明涉及分组数据网络领域。更具体地说，本发明属于一种数据流控制方法和系统，其根据城域分组传送环网络(metro packet transport ring network)中的可用带宽来管理数据流。本发明所公开的内容描述了一种系统和方法，其用于分组数据网中的带宽分配跟踪。

背景技术

因特网是一种通用的公共计算机网络，它使得连接到因特网上的全世界成百万的计算机和同样连接到因特网上的其他计算机进行通信并交换数字数据。随着新技术的出现，人们能够连上因特网的速度在不断增加。现在因特网上的用户们有带宽来参与聊天室中的现场讨论、玩实时游戏、观看流式视频、收听音乐、在线购物和交易等等。将来，可以想象，带宽不久将会使得视频点播、HDTV、IP 技术、视频电信会议以及其他类型的带宽强化应用成为可能。

一种正在应用的增加带宽的方法与光纤技术相关。通过比人的头发还细的玻璃光纤发送光脉冲，大量的数字数据能够以极高速度发送。并且随着密集波分多工法(dense wavelength division multiplexing）的出现，可以用同一个单根光纤来引导不同波长的光，从而使其容量增加若干倍。

可是，在将这种新的光纤网络的带宽分配给最终用户时存在问题。众所周知，一些应用对诸如等待时间和丢失分组(dropped packet)之类的带宽约束是不敏感的。例如，对等待时间和丢失分组来说，电子邮件应用和基本 Web 浏览就相对于时间是不敏感的。另一方面，诸如实时双向语音通信或视频之类的应用对于等待时间和丢失分组所导致的时间延迟就非常敏感。这些应用的可接受的性能在很大程度上取决于提供保用最小带宽。

不幸的是，由于网络通信量拥塞、网络可用性、路由条件以及其他不可控的外部因素，已经证明对于某些顾客而言，提供带宽可用性的保用水平是成问题的。一般来说，数据分组按照最佳效果传送模式来竞争可用带宽并被路由。同样，传统分组交换数据网络的可靠性有时是次优的。例如，在大多数情况下，在 IP 网上使用传统 LAN 交换机和路由器很难提供任何种类的服务质量(QoS)。QoS 是指及时提供信息传递、控制每个用户的带宽以及为选择通信量而设置优先级的保证。

不同的网络通信流(或者简称“流”)分别与不同的应用相关联。一个流是指从一个发送器到一个接收器、用以支持应用的分组传输，比如传送一个网页、实现 IP 上的语音会话、播放视频等等。有一些流被称为实时流，因其需要非常低的等待时间(例如 IP 上的语音应用)。其他流则并不在如此大程度上依赖等待时间，因其依赖恒定的数据传送速率(例如网页上的视频)。对于诸如视频点播、HDTV、语音通信等等实时应用流来说，流的丢失分组或者迟到分组会严重地干扰或者甚至破坏操作。并且对于许多因特网服务提供商（ISP）、应用服务提供商(ASP)、网站/入口（portal)以及企业而言，最重要的是：它们有能力为这些流提供某一最小门限值带宽和/或等待时间。例如，一个电子商务或者商业网站可能由于顾客不能在峰值时间访问它们的站点而使销售受损，从而丢失关键的收入。

因为有些用户对 QoS 期望如此之高，已开发了一些提供 QoS 功能的机制。一种用于实现 QoS 的原有技术方法是采用各种 TDM（时分多工）方案。一种广泛使用的 TDM 方案是实施 T 载波(T-carrier)服务（例如，以 1.544 兆位/秒来载送数据的 T1 线路，以及以 274.176 兆位/秒的快得多的速率来载送数据的 T3 线路）。这些 T1 和 T3 线路专用于由电话公司出租的点到点数据链路。电话公司通常以长话费率（例如，每个月$1,500～$20,000）来对租赁普通的老式 T1 线路收费。另一种常用的实现 QoS 的 TDM 方案涉及同步光纤网络(SONET)。与 T 载波服务相同，SONET 利用 TDM 来分配单个的信道或流，从而预先确定时隙。利用 TDM，每个信道都保证有其自身的特定时隙，在该时隙中，该信道可发送其数据。虽然 TDM 可以确保 QoS，但是它实现起来很昂贵，因为发送机和接收器二者任何时候都必须同步。用于保持这种精确同步的相关电路和系统开销是昂贵的。而且，就未使用的时隙而言，基于 TDM 的网络技术是低效率的。如果通信流是不活动的，则为其分配的带宽就被浪费了。总的来说，利用 TDM 技术，没有将不活动流中未使用的带宽分配给其他用户。

另一种原有技术方法是结合使用异步方案中各种形式的带宽保留(bandwidth reservation)。与同步 TDM 方案相比，异步数据传输方案有许多优点，并且同样，一般在语音和数字网络设备（例如，因特网中基于 IP 的网络）两方面都赶上了同步技术。在实现 QoS 时，异步方案经常通过将其一部分带宽保留用于“高优选级”的等待时间敏感的流来工作。对于大多数异步方案（例如以太网），QoS 性能随着网络中带宽使用的增加而恶化。随着网络所使用的可用带宽的百分比增加，原有技术的异步 QoS 保留方案的工作效率也就变低。这些方案或者保持大余量的未使用带宽来确保 QoS，从而实质上是保证了可用总带宽利用不足；或者是过分地分配带宽，导致对于某些用户来说突然丢失数据和/或对于任何高优选级用户来说破坏 QoS。

因此，需要这样一种解决方案：其在有效实现 QoS 的同时具备异步数据网络的优点。需要这样一种解决方案：其能够有效分配可用带宽，从而得以保证 QoS。所需要的解决方案应该能够把带宽异步地分配给各个流，而不会随规模增加（例如极大数量的流）以及网络利用

增加而导致原有技术的异步方案的恶化性能。所需要的解决方案应该能确保最少量的保留带宽，而不会招致原有技术中基于 TDM 的网络方案（其中带宽浪费在不活动的流上）的浪费带宽的问题。

所需要的解决方案应该能够在单个基础上跟踪各个流，以便确保各个流不缺乏带宽，同时同步地确保带宽没有被过分地分配给对其无需求的流。所需要的解决方案应该能够跟踪各个流何时活动以及它们何时不活动，从而使得分配给不活动流的带宽在需要时能够被重新分配给这些流。所需要的解决方案应该能够实时跟踪总的分配带宽，从而允许实时地进行未使用带宽的有效分配且同时保持 QoS。实时的总分配带宽跟踪应该允许实时地进行未使用带宽的动态分配。本发明对上述要求提供了一种新颖的解决方案。

发明概述

本发明包括一种具备异步数据网络的优点且同时有效实现 QoS 的方法和系统。本发明让可用带宽能得到有效分配，从而得以保证 QoS。本发明能够在单个基础上跟踪各个流，以便确保各个流不缺乏带宽，同时同步地确保带宽没有被过分地分配给对其无需求的流。本发明能够跟踪各个流何时活动以及它们何时不活动，从而使得分配给不活动流的带宽在需要时能够被重新分配给这些流。本发明能够实时跟踪总的分配带宽，从而允许实时地进行未使用带宽的有效分配且同时保持 QoS。实时的总分配带宽跟踪允许实时地进行未使用带宽的动态分配且同时保持 QoS。

在一个实施例中，本发明为一种系统，用于保持网络上已分配带宽的精确总量，例如实施于城域交换机(metropolitan area switch, MPS)内，该城域交换机通过分配城域网的带宽而工作。在 MPS 内，多个进入的分组分配给对应的多个 MPS 队列。计算每个对应队列的结束时间，该结束时间使用输出速率来描述各对应队列将清空的时刻。该多个队列按照其各自的结束时间而分为多个组。由于这些组包括具有相同结束时间的队列，所以将这些组称之为“桶”（bucket）。

初始组包括所具有的结束时间指示出在第一时间增量处的清空状态的那些队列的保留带宽。第二初始组包括所具有的结束时间指示出

在第二时间增量处的清空状态的那些队列的保留带宽，其中第二时间增量在第一时间增量之后，等等，依此类推。因此，例如桶 0 包含将在下一时间增量处清空的那些队列，桶 1 包含将在其后两个时间增量处清空的那些队列，等等，依此类推。在网络上的分配带宽的总量通过计算所有活动流的保留带宽来确定。

第一时间增量、第二时间增量等等相对于一个调度时钟而标以索引。调度时钟的一个增量包括 MPS 内所有活动队列的一个完整的循环仲裁（round robin arbitration）——例如输出到城域网上的每个队列。因此初始组指示在调度时钟的下一时间增量处（例如，输出循环）将具备清空状态的那些队列。当一个新的分组为对应的队列所接收时，对于每个对应的队列计算一个新的结束时间。按照这种方式，桶的系列随着调度时钟前进而顺次清空，并且新的桶随着新的队列接收要传输的新分组和新的相关清空时间而填充。在下一时间增量处为空的队列指示将在下一时间增量处不活动的那些流。分配给那些流的带宽可重新分配。按照这种方式，即可实时地完成已分配带宽总量的确定，从而能实时地进行未分配带宽的有效分配，同时保持服务质量。初始桶（例如桶 0）表示将在下一时间增量处为空的所有队列的保留速率。

因此，通过把单个的流分成如上所述的桶，本发明的各实施例即能够有效地扩大来处理一个极大数目（例如一百万或者更多）的单个流。这些流在单个的基础上如上所述地分配给各桶。它们的状态（活动与不活动）被单个地实时跟踪，从而能够实时地将不活动的流的分配带宽重新分配给活动的流。通过这种做法，因为 MPS 能够实时跟踪总的分配带宽，本发明即实现了可用带宽的有效分配。这使得能够进行未使用带宽的实时有效分配而同时保持 QoS。

附图说明

本发明通过用作示例而非用作限制的附图来说明，而且在附图中，同样的参考编号指同样的部分，其中：

图 1 显示根据本发明当前的最佳实施例的异步城域分组传送环网络的整体结构。

图 2 显示一个示例性城域分组传送环（Metro Packet Transport Ring，

MPTR）。

图 3 示出了 MPTR 组件示例的示意图。

图 4 是根据本发明一个实施例的示例性系统内实现的 MPS 单元组和环段的示意图。

图 5 示出了一个 MPS 队列及其相关结束时间的示意图。

图 6A 为一个示意图，其说明根据本发明一个实施例的多组队列进程。

图 6B 图解说明了根据本发明一个实施例的所有 r_i 和 w_i 的总和。

图 7 示出了根据本发明一个实施例的桶信息基础（bucket information base, BIB）的示意图。

图 8 显示根据本发明一个实施例的流信息基础（flow information base, FIB）。

图 9 示出了根据本发明一个实施例的带宽跟踪和分配处理的步骤流程图。

详细说明

现在将详述本发明的实施例，以附图来说明其示例。虽然结合优选实施例而描述本发明，但是应该理解，并非意在以这些实施例来将本发明局限于此。相反，本发明意在覆盖可包括在本发明的精神和范围之内的替换、改动和等价方案，而本发明的精神和范围由所附的权利要求书来规定。此外，在以下的本发明详细说明中，阐明了许多特定的细节，以便于全面理解本发明。但对于本领域普通技术人员来说很明显，没有这些具体细节也可实现本发明。在其它例子中，未详细描述公知的方法、程序、组件和电路，以免不必要地混淆本发明的各个方面。

本发明的实施例是针对一种用于保持网络上已分配带宽的精确总量的方法和系统，例如实施于城域交换机（MPS）内。本发明提供异步数据网络络的优点，同时有效实现 QoS。本发明能够进行可用带宽的有效分配，从而得以保证 QoS。本发明能够实时跟踪总的已分配带宽，从而能够实时进行未使用带宽的有效分配，同时保持 QoS。下面进一步描述本发明及其优点。

图 1 显示根据本发明目前优选实施例的异步城域分组传送环网络的整体结构。一个城域分组传送环（MPTR）由城域网（MAN）中设置用来发送数据分组的一个环构成。一个 MAN 是覆盖一个地理城域地区的主干线网络。典型地，电话公司、有线电视公司以及其他电信提供商向需要接入覆盖城域中公共可通行的网络的其他公司、企业和用户提供 MAN 服务。在目前的优选实施例中，利用已安装光纤电缆的环布局技术来提供 MPTR 的通信信道。可以部分地使用或者替代诸如混合光纤同轴电缆、同轴电缆、铜线或者甚至无线（无线电频率或者空中激光束(over-the-air laser beam)）之类的其他较为低效的传输媒体。

连接到特定 MPTR 的用户可通过该 MPTR 而在彼此间发送和接收分组数据。例如，连接到 MPTR1 的一台个人计算机能够对 / 从同样连接到 MPTR1 的一台服务器发送和接收数据分组。而且，发自一个 MPTR 的数据分组可以通过一个路由器而路由到另一 MPTR。例如，连接到 MPTR1 的一个计算机可以通过其接到路由器 101 的光纤环而将数据分组发送给 MPTR2。该数据分组然后可以通过与 MPTR2 相连的光纤环而发送到其最终目的地（例如，连接到 MPTR2 的计算机）。应当指出，MPTR 环可以具有不同的尺寸和结构。虽然目前的优选实施例构想使用了一个环，但是本发明也可利用其他类型的布局技术。各 MPTR 还也可经路由器连接到因特网主干线上。例如，MPTR1 能够通过路由器 101 连接到密集波分多工（DWDM）光纤主干线 102。因此，连接到 MPTR1 的用户可访问传统因特网 103 上可用的资源。注意：本发明可结合传统的因特网方案来使用，传统的因特网方案使用标准路由器、交换机以及其他 LAN 设备 104～107。因此可将任何数目的 MPTR 连接在一起，从而适度且成本有效地进行调节，以便符合可能提出的最严格的联网要求。并且当一个特定的环变成过载时，可增加第二、第三、第四等等 MPTR，从而容纳增加的负载。这些 MPTR 可以连接到同一个路由器（例如 MPTR5、MPTR6 和 MPTR7），或者可选择地连接到不同的路由器。

本发明的结构不但调节适度，而且它提供了巨大的灵活性。在一个实施例中，可以用一个 MPTR 来支持一个或多个 LAN。例如，MPTR6 可支持发向 / 发自 LAN108 的通信流。还可选择构想将一个 MPTR 直

接连接到另一个 MPTR。按照这种方式，流入 MPTR8 的数据能够直接与流经 MPTR7 的数据分组进行交换。可选择使单个 MPTR 具有多个入口 / 出口。例如，MPTR5 既连接到路由器 109，也连接到路由器 / 交换机 110。因此，MPTR5 上的用户即能够通过两个路由器中的任一个——109 或者 110——来发送和接收数据分组。利用本发明的 MPTR，实际上使得任何结构、协议、媒体以及布局技术都成为可能。

现在描述 MPTR 的实施和功能性。参见图 2，其中显示一个示例性城域分组传送环 200。可以看到，MPTR 200 包括两个光纤电缆环(或称环) 201 和 202；多个城域分组交换机（MPS1～MPSn）；以及一个环管理系统（Ring Management System, RMS）203。一个 MPTR 的物理层实际上由两个冗余的光纤电缆环 201 和 202 构成。数据分组通过这两个环而以相反方向流动（例如在环 201 上顺时针而在环 202 上逆时针）。沿光纤环 201 和 202 分布多个城域分组交换机（MPS）。一个 MPS 连接到两个光纤环 201 和 202。因此，如果在光纤电缆环的一个分段中出现中断，数据即可从 MPS 之一改向流经另一个可操作的 MPS。还可选择使通信流改向而将任一个环中发生的本地拥塞减至最小。

在目前的优选实施例中，每个 MPTR 可支持高达 254 个 MPS。一个 MPS 可以是设备的一个部件，其安装在特别设计的周围结构之中，或者其可以位于布线柜中，或者是置于商业地点，等等。MPS 之间的距离是可变的。通过 MPS，每一单个的最终用户得以接入光纤环 201 和 202。每一单个的最终用户把分组数据首先发送到 MPS 上。MPS 然后调度如何把分组数据放到光纤环上。类似地，在将分组数据发送给连接到 MPS 上进行接收的最终用户之前，该 MPS 首先使其脱离光纤环。在目前的优选实施例中，单个 MPS 可以支持高达 128 个最终用户。通过把一个线路接口卡插入一个 MPS，即可将一个最终用户附加到该特定的 MPS。线路接口卡提供 I/O 端口，通过该端口，数据即可在 MPS 及其最终用户之间传送。可设计不同的线路接口卡以满足对应于该特定最终用户的特定协议。所支持的一些协议包括 T1、T3、SONET、异步传输模式（ATM）、数据用户线路（DSL）以太网，等等。应当指出，可设计线路接口卡以满足未来协议的规范。依此方式，诸如主机

计算机、工作站、服务器、个人计算机、机顶盒、终端、数字设备、TV 控制台、路由器、交换机、集线器以及其他计算 / 处理设备之类的最终用户即可通过一个 MPS 接入光纤环 201 和 202。

MPS 不但对最终用户提供 I/O 端口，而且 MPS 也提供了将分组数据输入到 MPTR 以及从 MPTR 输出分组数据的装置。例如，数据分组经连接到路由器 205 上的 MPS 204 而输入到 MPTR 200。类似地，数据分组经 MPS 204 而从 MPTR 200 输出到路由器 205。

MPS 的另一功能使得发自一个上行流 MPS 的进入的数据分组传送到另一个下行流 MPS。一个 MPS 通过连接到光纤环上的输入光纤端口而从一个上行流 MPS 接收所发送的上行流数据分组。从光纤环中所接收的数据分组由该 MPS 进行检查。如果数据分组要发送到连接至该特定 MPS 上的一个最终用户，则该数据分组被路由到对应的 I/O 端口。否则，MPS 尽可能快地立即把该数据分组转发到下一个下行流 MPS。该数据分组经一个输出光纤端口而从 MPS 输出到光纤环上。应当指出，从上行流光纤环分段经 MPS 而流到下行流光纤环分段的这类通过的(pass-through)分组的优先级总是高于正等待为 MPS 插入到光纤环的分组。换言之，仅在带宽允许时，MPS 才插入由其最终用户所生成的数据分组。

现在提供一个示例，以表明数据分组如何在 MPTR 中流动。参考图 2，连接到 MPS4 的计算机 207 可以对 / 从因特网发送和接收数据，如下所述。由计算机生成的数据分组首先经由一条连接到线路接口卡——该卡设置在 MPS4 中——的线路而发送到 MPS4。这些数据分组然后被 MPS4 通过环分段 206 发送到 MPS3。MPS3 检查数据分组并通过环分段 207 把数据分组下行流传送到 MPS2；MPS2 检查数据分组并通过环分段 208 把数据分组下行流传送到 MPS1。基于包含在数据分组中的地址，MPS1 知道将这些数据分组输出到对应于路由器 205 的 I/O 端口。可以看出，MPS1 连接到路由器 205。路由器 205 对 / 从 MPTR 200、其他 MPTR 或者因特网主干线路由数据分组。在这种情形中，数据分组然后通过因特网被路由到它们的最终目的地。类似地，路由器 205 使数据分组从因特网经 MPS1 而路由到 MPTR 200。进入的数据分组然后受到检查并通过环分段 209 而从 MPS1 转发到 MPS2；经过检查并通

过环分段 210 而从 MPS2 转发到 MPS3；且经过检查并通过环分段 211 而从 MPS3 转发到 MPS4。MPS4 检查这些数据分组并确定它们要发送到计算机 207，从而 MPS4 通过其对应于计算机 207 的 I/O 端口输出该数据分组。

同样地，连接到任何 MPS 上的用户能够在同一个 MPTR 上发送并从任何其他 MPS 接收分组而无需离开该环。例如，MPS2 上的一个用户可以通过首先把分组发送到 MPS2 来把数据分组发送给 MPS4 上的一个用户；在环分段 207 上把该数据从 MPS2 发送给 MPS3；MPS3 通过环 202 而将分组发送给 MPS4；并且 MPS4 将该分组输出到对应于预定接收方的适当端口。

仍参见图 2，应当指出，本发明解决了环布局技术网络所共有的严格优先级问题。严格优先级是指上行流节点（例如一个上行流 MPS）与下行流节点相比，在通信信道中具有更多的可用带宽这一事实。例如，在环分段 210 的情况下，MPS 2 能够在 MPS3 之前将其本地输入流（例如插入通信量）插入到分段 210 上， 依此类推，MPS 3 和 MPS 4 对于环分段 211 也是如此。因此，由于在环布局技术中 MPS 4 的位置优势，与 MPS 3 和 MPS 2 相比，MPS 4 具有较少的可用带宽来插入其本地输入流。

为了避免严格优先级问题，需要关于环分段的已分配带宽的详细信息。每个 MPS 需要知道这些分段的已分配带宽，以便就任何剩余的未分配带宽的分配做出明智判断。在结合保证 QoS 而使带宽利用率达到最大时，这种信息甚至更为重要。优选地，带宽应用信息应该在“每个流”的基础上可用，应足够及时，从而允许实时作出智能分配决策。

MPTR、MPS 和 RMS 结构的附加描述可以在____日申请的、序列号为____、转让给本发明的受让人的美国专利申请“GUARANTEED QUALITY OF SERVICE IN AN ASYNCHRONOUS METRO PACKET TRANSPORT RING”中查找到，而且可以在____日申请的、序列号为____、转让给本发明的受让人的美国专利申请“PER-FLOW CONTROL FOR AN ASYNCHRONOUS METRO PACKET TRANSPORT”中查找到——在本说明中整体地结合引入该申请。

图 3 为示例性的 MPTR 组件的示意图。所示多个 MPS 301～306

连接到一个光纤环 307 上。其中两个 MPS 302 和 303 显示得尤为详细，以说明数据如何在 MPTR 中流动。所示多个计算机 308～310 连接到 MPS 302。这些计算机 308～310 中，每一台都具有一个对应的缓冲器 311～313。这些缓冲器 311～313 用来暂时存储来自其各自的计算机 308～310 的进入的数据分组。与这些缓冲器 311～313 中的每一个相关联的是对应的控制器 314～316，控制器 314～416 控制何时允许将排列在该特定缓冲器中的分组发送到环 307 上。一旦一个分组被允许从 MPS 302 中发送，则它即被插入到插入器 325 中，并附加到其他要发送的分组中，以进行该循环。一旦一个分组从 MPS 转发到环 307 上，该分组即以环 307 的最大速率而发送到其目的地，并且立即通过中介 MPS（如果有的话）转发。

在一个优选的 MPTR 实施例中，采用每个流的带宽分配概念来实现公平带宽分配（fair bandwidth allocation）。环 307 上的通信量分类成流。例如，来自一个用户的所有分组属于一个流。流的粒度可以是细密的（例如，每个对话）或者是粗糙的（例如，每个服务端口等等），并且通常可通过分组分类规则而大体上予以规定。一旦将分组分类到一个流中，则每个 MPS 能够公平地分配带宽给每个流并且监视没有流超出该分配。

因此在分组可以发送到环上之前，必须设置流。设置流包括规定若干参数。在这些参数之中，所保留的带宽 r_i 和分配加权 w_i 对于流的控制来说是必要的，其中“i”是流的唯一标识符，称为流 ID。一旦设置，即由流的唯一的流 ID 来对其进行识别。

图 4 是示意图，显示了三个 MPS 单元及其各自的环分段。如图 4 所示，三个 MPS 单元（MPS 0，MPS 1 和 MPS 2）与其各自的环分段 401～404 被示出。所示 MPS 单元具有其各自的插入通信量（I0、I1 和 I2）和其各自的出口通信量（E0、E1 和 E2）。所示各 MPS 0～2 具有用来跟踪流的多个内部队列（图中所示为每个 MPS 内有四个）。

如图 4 所示，每个 MPS 的队列跟踪每个输出环段 401～404 上的已分配带宽。如图 4 所示，输出分段上的通信量表示为：

$$\sum_{活动} r_i \text{ 和 } \sum_{活动} w_i$$

每个 MPS 的队列跟踪属于每一单个流的数据通信量（在下面将更

详细地描述）。每个分段上的通信量考虑到先前 MPS 的出口通信量、先前 MPS 的插入通信量以及环上的通过通信量。每个 MPS 的插入通信量如图 4 所示为“I”而每个 MPS 的出口通信量如图所示为“E”。插入通信量是来自连接到 MPS 上、要进到环上的用户的流（例如发往连接到另一 MPS 的用户）。每个 MPS 内的队列被使用来跟踪被一个 MPS 监视和保持的队列流（例如，具有唯一的流 ID）。跟踪输出环分段的输出流的每个队列以相当于已分配带宽的速率而排空。

队列以受其各自加权 w_i 作用的速率而清空。每个队列加权使得实现每个队列不同级别的带宽。例如，在队列具有相同加权处，各个流分组以一个相等速率从队列中被路由。一旦分组被插入到一个输出环分段上，例如，从插入通信量 I_0 的流被插入到环分段上的一个分组，则那个分组与其它输出分组相加并以有线速度会做何环发送的最大速率来沿着环分段被发送。分组通过中间 MPS 被立即转发（如果需要的话），就象通过通信量一样。一旦一个队列变成空，则其带宽分配就可用来重新分配给其它非空队列。

应当指出，在一个优选实施例中，按照本发明的一个 MPS 保持了很大的虚拟队列（virtual queue, VQ）组来监视其所有输出链路上的流活动。虚拟队列按照与上述队列相类似的方式而工作（例如，图 4 所示的 MPS 单元内示出的队列），可是，它们被实现为跟踪队列深度的计数器以便数据分组在流过它们各自的缓冲器时不被延迟。在优选实施例中实现的虚拟队列的附加描述可以在____日申请的、序列号为____、受让给本发明的受让人的美国专利申请“GUARANTEED QUALITY OF SERVICE IN AN ASYNCHRONOUS METRO PACKET TRANSPORT RING”中查找到，该申请在此被全部结合。一个 VQ 将具有结束时间，该结束时间描述所有分组以一个流分配速率 f_i 从该 VQ 中完全排出时的时间。

图 5 为队列 415 机器相关结束时间的示意图。队列 411～415 的输出速率实现了描述分别队列将被清空那个时刻的一个“结束时间”的确定。这个结束时间提供环 450 总的已分配带宽的关键测量。因此，如图 5 所述，队列 415 具有描述队列 415 将以它的输出速率被清空那个时刻的结束时间。当正如所示出的一个新的分组到达时，一个新的

结束时间被计算反映队列 415 新的深度。因此，如图 5 和 6 所示，MPS 以一个特定输出速率来路由来自各个队列中的分组，并且计算每个对应的队列的结束时间，该结束时间描述该对应的队列将使用该已分配输出速率来被清空的那个时刻。

按照这种方式，每个 MPS 保持大量队列（例如，高达一百万或者更多），每一个用于每个链路处的每个流。每个队列以属于该流的通信量速率来增长，并且以一个等于已分配带宽的速率被排出。按照所有非空（活动）队列（例如，队列 411～415）的 Σr_i 和 Σw_i 形式来测量拥塞。$\sum_{活动} r_i$ 和 $\sum_{活动} w_i$ 的最高值指示：对于 MPS 的输出链路带宽计算更多的流。每个 MPS 频繁地监视它队列的状态以便更新这两个参数。一旦检测到，则 MPS 使用 $\sum_{活动} r_i$ 和 $\sum_{活动} w_i$ 来计算每个流的带宽分配。

在一个优选实施例中，每个 MPS 计算通过每个拥塞点（例如，在输出环分段处）的所有流的带宽的公平分配。该分配是基于如下计算而被计算出来的：

$$f_i = r_i + \frac{w_i(C - \sum r_i)}{\sum_{活动} w_i},$$

在此，f_i 表示流 i 的已分配带宽，而 C 是拥塞点的链路容量。注意，项 $C-\sum_{活动} r_i$ 仅仅是 MPS 需要基于保留加权来公平地进行重新分配的链路的未保留带宽部分。这一项在下面的图 6B 中用图形来描述。

对于带宽有效性，每个 MPS 不是对于它看见的每个流都发出 f_i。取而代之的是，它发送一个容量保留比（CRR），其通常描述链路未分配带宽数量。CRR 然后可以被每个 MPS 中的每个源使用来从它的 r_i 和 w_i 状态数据库中计算出它自己的 f_i。CRR 较正式地定义如下：

$$CRR = \frac{C - \sum r_i}{\sum_{活动} w_i}$$

CRR 周期性地被广播给所有其它 MPS，以便使所有 MPS 分配未分配的链路带宽。每个 MPS 能够独立地选择更新频率。对于每个接收的 CRR，每个源使用如下等式来计算其 f_i：

$$f_i = r_i + w_i * CRR$$

因此，为了有效分布未分配的链路带宽，每个 MPS 需要跟踪已分

配带宽的总量和已分配带宽的总加权 $\sum_{活动} r_i$ 和 $\sum_{活动} w_i$。根据本发明，这些项被实时跟踪并以每个分段高达 10Gbps 的高速度跟踪流活动。本发明使用对应队列的结束时间和对应队列的已分配加权来实现 $\sum_{活动} r_i$ 和 $\sum_{活动} w_i$ 的高速跟踪方法。这些技术涉及每个流队列的使用、一个流信息基础（FIB）、一个桶信息基础（BIB）以及一个调度时钟。使用这些项，本发明的实施例能够有效地扩大来处理很大数目（例如，一百万或者更多）的单个流，同时保持在集成电路技术的性能内（例如，可以在一个 ASIC 中被实现）。单个流可以被实时跟踪，允许对于未活动流的它们的已分配带宽实时地被重新分配给活动流。

现在参见图 6A，其为描述本发明多组队列过程的示意图。图 6A 描述了被分类到多个组中的多个流，被示出为桶 0、桶 1、桶 2 等等直到桶 n。多个队列按照它们各自的结束时间被归组为多个桶或组。该结束时间根据一个调度时钟而标以索引。该调度时钟或者说全球时钟为结束时间提供时间基准。调度时钟的值表示结束时间与之进行比较多那个目前实际时间。调度时钟以与节点处的拥塞成比例的一个速率递增，如下所述。如图 6A 所示，当桶被清空时，它们从右移到左，因为每个桶渐近地到达了如图 6A 左侧上示出的“队列清空”状态。

由于这些流组包括具有与调度时间的相同结束时间的那些队列，所以这些流组被称为“桶”。例如，桶 0 包括具有与调度时钟下一增量相对应的结束时间的那些流的保留带宽和加权，而桶 n 包括具有对应于调度时钟的最长结束时间的那些流的保留带宽和加权。因此，初始桶（例如，桶 0）包括具有在第一时间增量处指示清空状态的结束时间的那些流（例如，队列），第二初始桶（例如，桶 1）包括具有在比第一时间增量更晚的一个第二时间增量处指示清空状态的结束时间的那些队列，如此类推。因此，例如桶 0 包含将在调度时钟的下一时间增量处为空的那些队列的保留带宽和加权，桶 1 包含将在调度时钟的下两个时间增量处为空的那些队列的保留带宽和加权，等等，从而来指示每个时间增量变成可用的未分配带宽数量。网络上未分配带宽数量通过对所有活动流的总分配带宽和总分配加权来确定（例如，所有桶总数）。

第一桶、第二桶等等的时间增量对应于调度时钟而标以索引。按照上面图 4 中所描述的方式，调度时钟的一个增量包括 MPS 内所有活动队列的一个完整循环仲裁（例如，输出到城域网上的每个队列）。未活动的或者空的队列未贡献于调度时钟周期。桶 0 因此指示将在调度时钟的下一时间增量（例如，输出循环）处具有清空状态的那些流。如上所述，对于每个分布的队列计算一个新的结束时间，并且因此对于每个流，当一个新的分组被对应的队列接收时。按照这种方式，桶系列随着调度时钟递增而渐进地被“清空”，并且一个新的桶随着一个新队列接收用于传输的新分组以及新的相关空时间而被填充，并且如图 6A 所示，从右到左对桶进行处理。

仍然参见图 6A，当调度时钟前进到一个桶的结束时间时，则该桶中的流以及因此的它们的队列被认为是完全被服务，并因此为空。当它们的队列为空时这些流被认为关于该链路是不活动的。

调度时钟通过每一个时间间隔 T_Sclk 来前进，给出如下：

$$T\_Sclk - \frac{\sum r_i + CRR * \sum w_i}{C}$$

该调度时钟（被表示为 SCLK）基于相应链路上的流活动性而独立前进。应当指出，SCLK 不必像传统始终那样以一个恒定速率来前进。$\sum r_i + CRR * \sum w_i$ 除以链路容量表示在目前 CRR 值处的链路利用百分比。$\sum_{活动} r_i$ 的值越高，则 SCLK 前进得越慢。队列的结束时间和调度时间之间的差值表示在清空该队列的时间量（清空时间）方面的该队列的储备（backlog）程度。

$T_{清空}=T_{结束}-T_{SCLK}$

除了确定流是活动的还是不活动的，调度时钟也可以被用来步测这些流以便确定它们中任何一个是否已经超过了它们各自的分配带宽。这可以通过确保 $T_{清空}$ 未达到很大来进行。

仍然参见图 6A，能够快速确定哪些 VQ 变空把 $\sum_{活动} r_i$ 和 $\sum_{活动} w_i$ 的计算简单化了。

这是由于可以递增进行计算的缘故。给出旧值，在如下的一个步骤中就可以计算出新的值：

$$\sum_{活动} r_i = \sum_{活动} r_i - \sum_{过期的桶} r_i$$

$$\sum_{活动} w_i = \sum_{活动} w_i - \sum_{过期的桶} w_i$$

$\sum_{过期的桶} w_i$ 也是可以递增地计算出来的一项。

当一个流从一个桶流到另一个时，则它的 r_i 和 w_i 从旧桶的总和中被减去并被加到一个新的桶上。对于从先前不活动（空）中出来的一个流，它的 r_i 和 w_I 也将被加到 $\sum_{活动} r_i$ 和 $\sum_{活动} w_i$。

图 6B 图示地描述了根据本发明一个实施例的所有 r_i 和 w_i 的总和。如图 6B 所示，横轴是带宽而纵轴是时间。链路容量如所示。轨迹示出了链路容量的使用随着时间的变化（例如，一些流变成活动的而另一些流变成不活动的）。

现在参见图 7，其为根据本发明一个实施例的一个桶信息基础（BIB）700 的示意图。如图 6A 所示的桶实施为保持在每个 MPS 内的数据库 BIB700 中的一系列计数器。如图 7 所示，每个桶实施为一个环总带宽计数器和一个相应的环总加权计数器。计数器递增，反映桶中流的数量以及其相关的加权。按照如上所述的方式，调度时钟的工作是作为依其各自的流将为清空的时间增量通过计数器循环的一个指针。因此，例如，在下一时间增量处，调度时钟指针将移动来指示与桶 1 等等相关的计数器。在一个优选实施例中，将 BIB 700 构造为一个两行 8K 长的表，如图 7 所示。BIB 700 保持每 50 纳秒（ns）16 次访问，从而例如当队列内新的分组到达时允许更新。

图 8 示出了根据本发明一个实施例的一个流信息基础（FIB）800。一个 MPS 使用 FIB 800 作为流描述器。FIB 800 包含规定被应用到属于每个流的分组上的各个动作的字段（在环上转发，从环中退出到一个特定端口等等）和保持诸如 r_i 和 w_i 之类的流参数的字段。一个流的结束时间，其跟踪它的虚拟队列深度， 被存储在 FIB 中。当分组到达时，FIB 中的结束时间被更新，并如上所述地被使用来访问 BIB。因此 FIB 只是当分组到达时被访问。

图 9 示出了根据本发明一个实施例的操作过程 900 的步骤流程图。正如图 9 中所描述的，过程 900 示出了保持网络上已分配带宽的精确总量的一个 MPS 的操作步骤，例如实施于一个 MPTR 中。

过程 900 在步骤 901 中开始，在此，用于传输的数据分组被 MPS 的队列从多个用户中接收。在该 MPS 中，来自各个用户的多个进入的分组被分配给对应的多个 MPS 队列。

在步骤 902 中，来自队列中的数据被路由到环上。使用一个公平的仲裁方案（例如循环等等），配置一个控制器，以便用特定输出速率来清空对应的队列。

在步骤 903 中，对于每个对应的队列计算结束时间。该结束时间表示将使用目前输出速率来清空对应队列的时间。

在步骤 904 中，这些队列基于其各自的结束时间而归组到对应的桶中。为了强化高速跟踪，多个队列按照其各自的结束时间而归组到多个桶或组中。由于这些组包括具有相同结束时间的那些队列，所以这些组被称为“桶”。如上所述，这些桶可以使用一个数据库中对应的计数器对而实施，该计数器对配置来跟踪具有相同结束时间的总保留 r_i 及其各自的加权。

在步骤 905 中，一个调度时钟按照控制器的循环时间而递增加一。如上所述，一个较大数目的活动流导致调度时钟的较低递增速率，或者反之亦然。结束时间根据该调度时钟而标以索引。初始桶包括具有在第一时间增量处指示清空状态的结束时间的那些队列，第二初始桶包括具有在比第一时间增量更晚的第二时间增量处指示清空状态的结束时间的那些队列，等等，依此类推。

在步骤 906 中，变为不活动的总流 r_i 和它们相关的加权使用桶而被确定。如上所述，计数器对（被配置来跟踪具有相同结束时间及其对应加权的保留的队列带宽）可以被使用来确定在下一调度时钟增量上变成未活动的数据流的已分配带宽及其相关加权。

在步骤 907 中，基于在步骤 906 中获得的信息来确定未分配带宽的总量。如上所述，网络上的未分配带宽总量通过计算 $\sum_{活动} r_i$ 和 $\sum_{活动} w_i$ 来确定。此信息允许 MPS 精确确定可用于分配给活动流的未分配带宽的总量。

在步骤 908 中，当新的数据在用于传输的队列处到达时，对于活动流计算新的结束时间。随后，在步骤 909 中，过程 900 继续重复步骤 904～909。按照这种方式，桶串随着调度时钟渐进而渐进地被“清

空”，并且新的桶随着新的队列接收用于传输的新分组和新的相关空时间而被充满。

因此，已分配带宽总量的确定可以被实时跟随，因此实现了实时的未分配带宽的有效分配同时保持服务质量。初始桶（例如桶 0）示出了在下一时间增量中将为空的所有队列。这样做时，本发明实现了可用带宽的有效分配，因为 MPS 能够实时跟踪总的已分配带宽。这使得实时地有效分配未使用的已分配带宽同时保持 QoS。

已经为了说明和描述的目的呈现了本发明特定实施例的在前描述。它们不意欲成为详尽的或者来把本发明限制为所公开的精确形式，而是很明显根据上面的指导，许多修改和变化都是可能的。这些实施例被选择和叙述以便最好地解释本发明的原理及其实际应用。本发明的范围意欲由所附权利要求书及其等价物来定义。

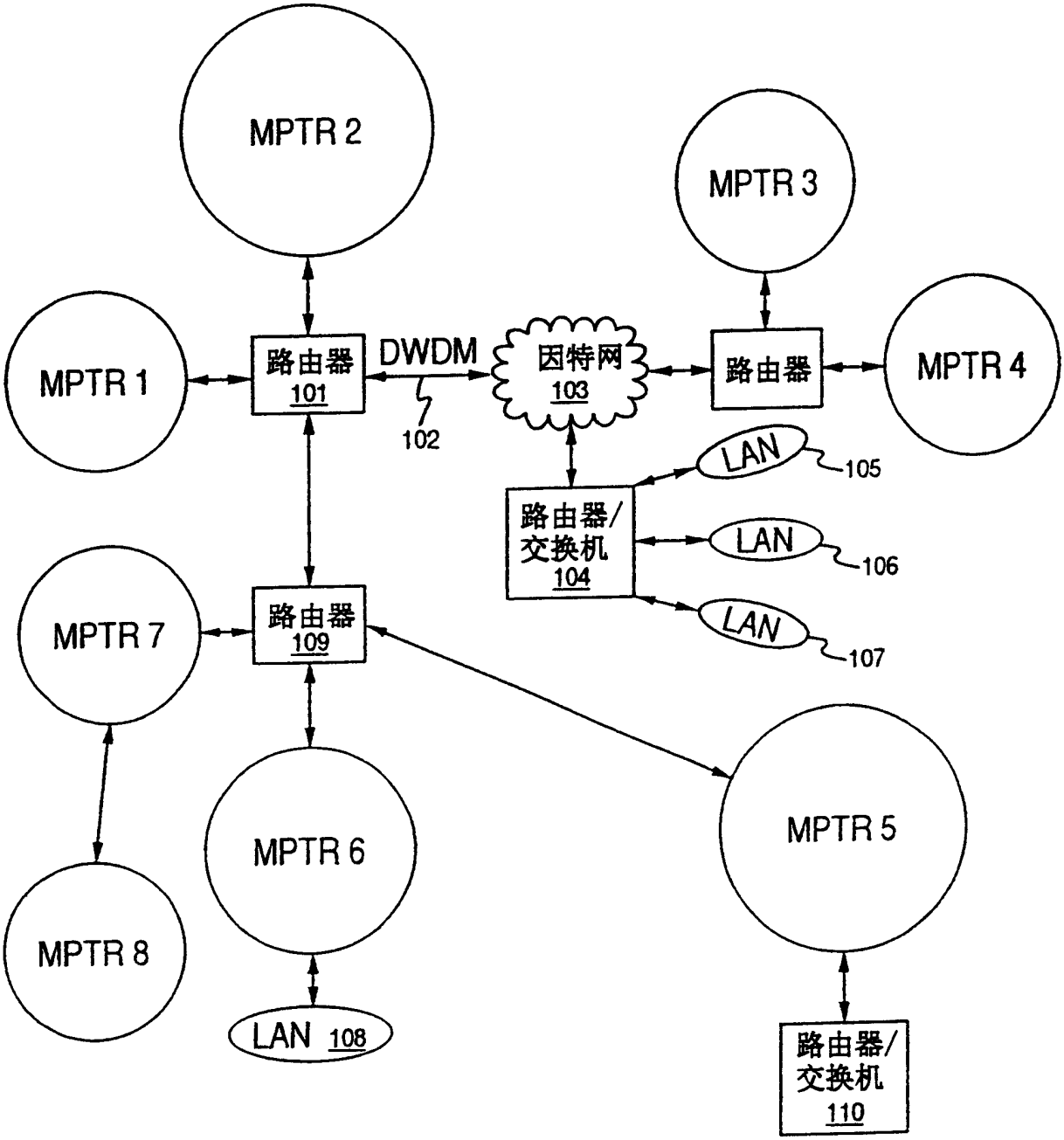

MPTR 2
MPTR 3
MPTR 1
路由器
101
DWDM
102
因特网
103
路由器
MPTR 4
LAN
105
路由器/
交换机
104
LAN
106
LAN
107
MPTR 7
路由器
109
MPTR 6
MPTR 5
MPTR 8
LAN 108
路由器/
交换机
110

图1

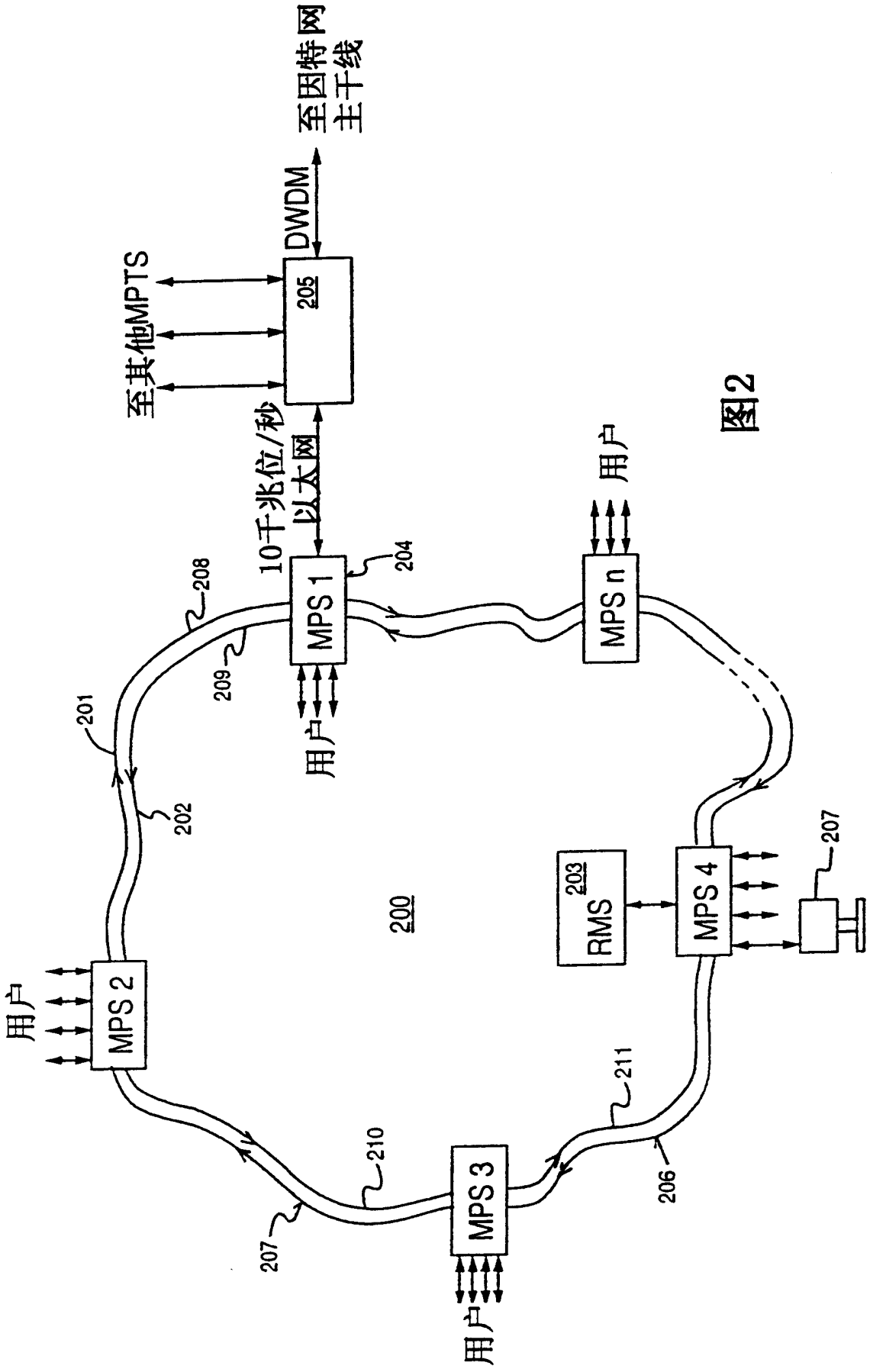

至因特网
主干线
DWDM
205
至其他MPTS
10千兆位/秒
以太网
MPS 1
204
208
209
201
202
用户
MPS n
用户
200
RMS
203
MPS 4
207
MPS 2
用户
210
211
206
207
MPS 3
用户

图2

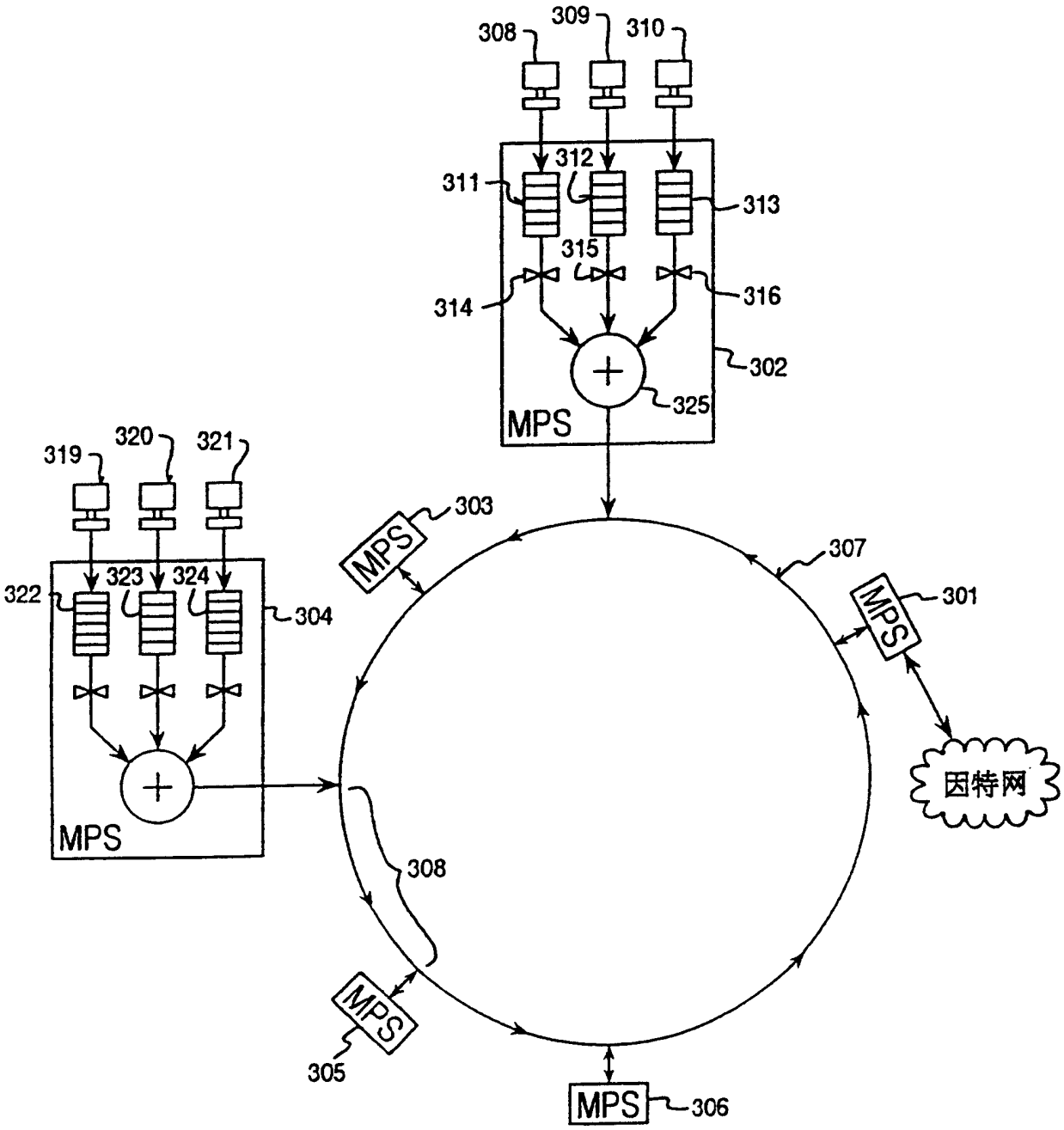

308
309
310
312
311
313
315
314
316
302
325
MPS
319
320
321
322
323
324
304
303
307
301
因特网
308
305
306

图3

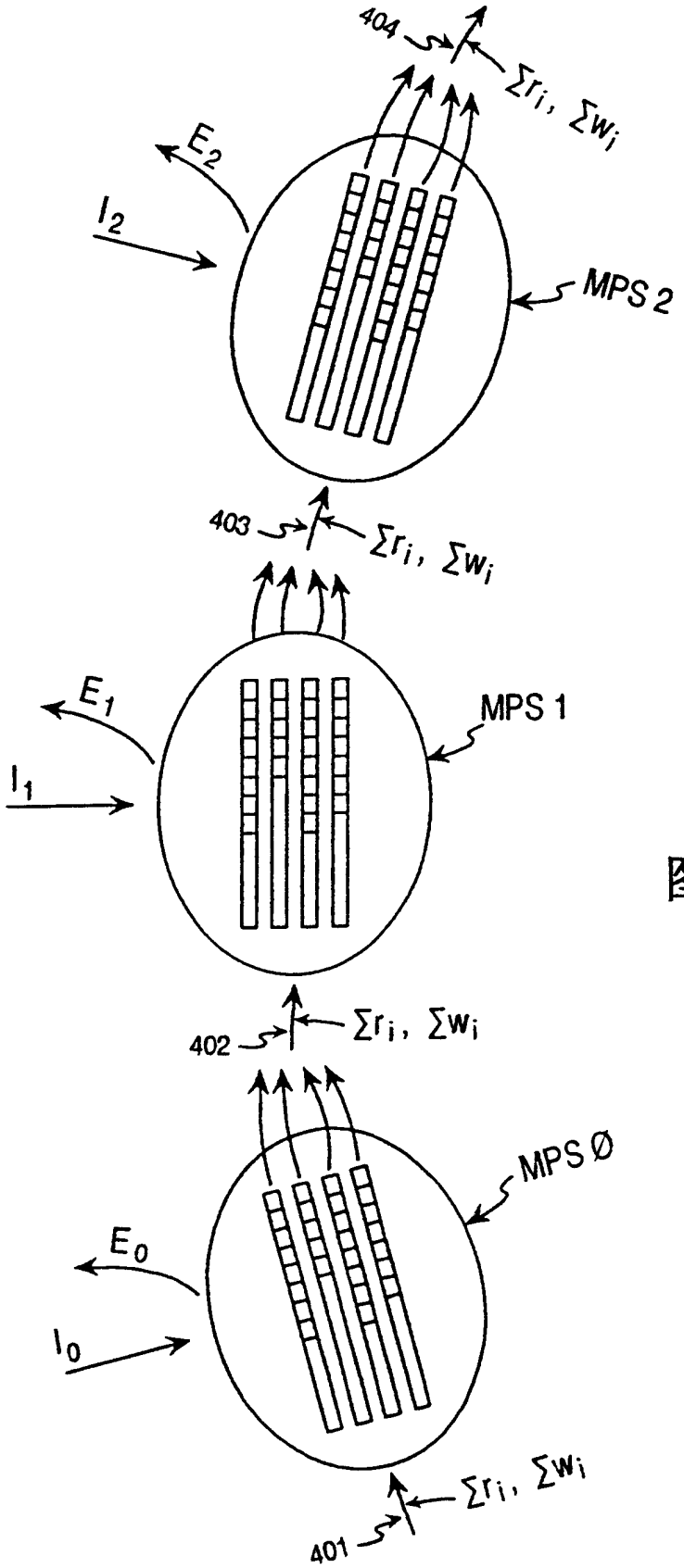

404
Σr_i, Σw_i
E_2
I_2
MPS 2
403
Σr_i, Σw_i
E_1
I_1
MPS 1
402
Σr_i, Σw_i
MPS Ø
E_0
I_0
Σr_i, Σw_i
401

图4

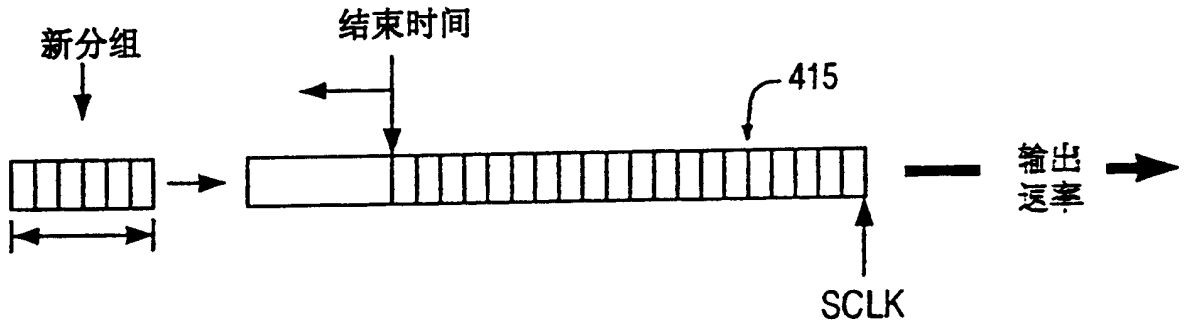


图5

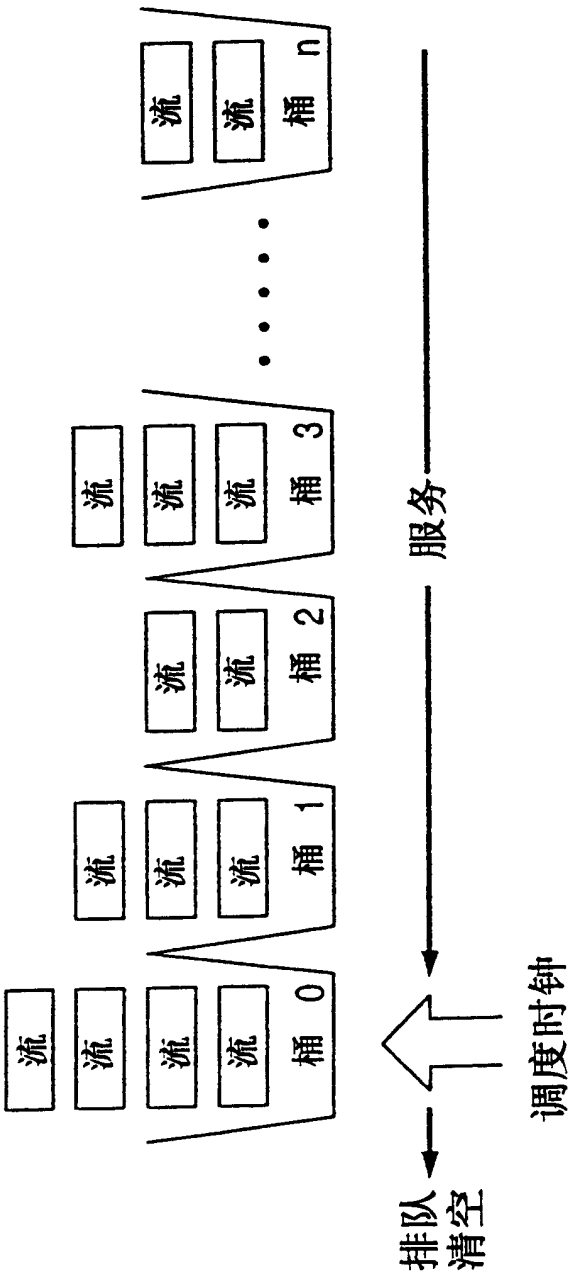


图6A

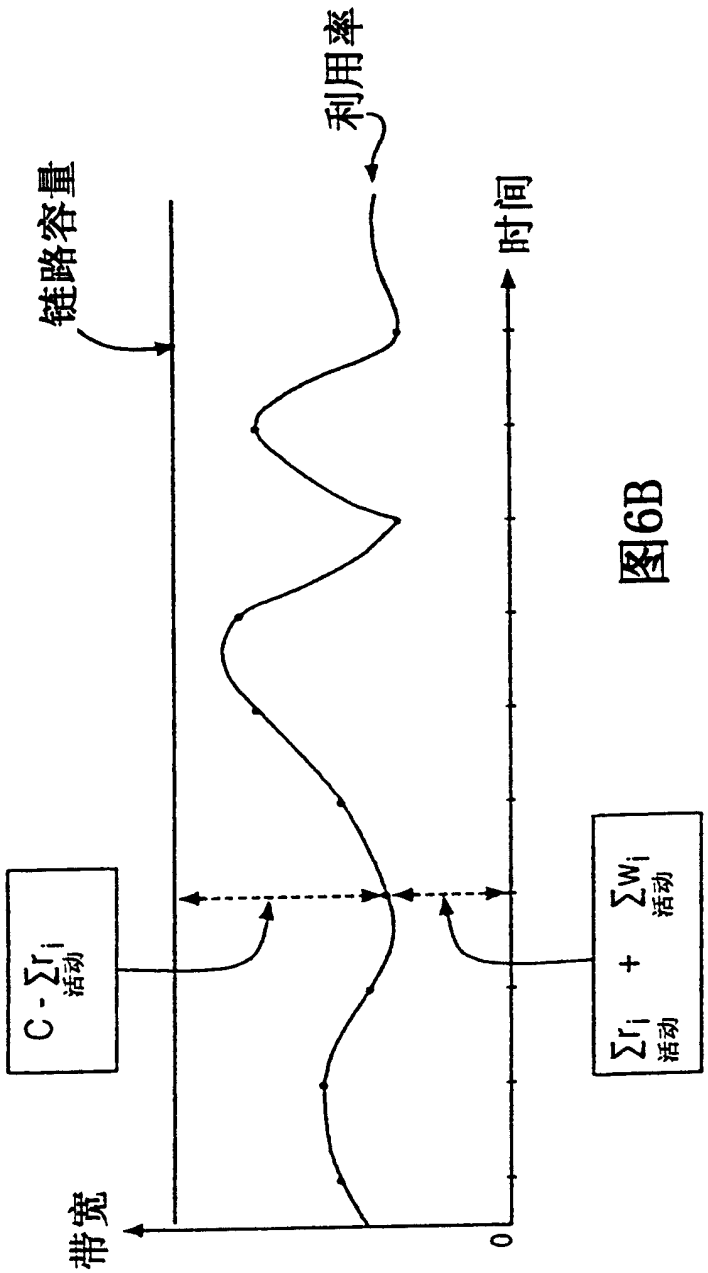


图6B

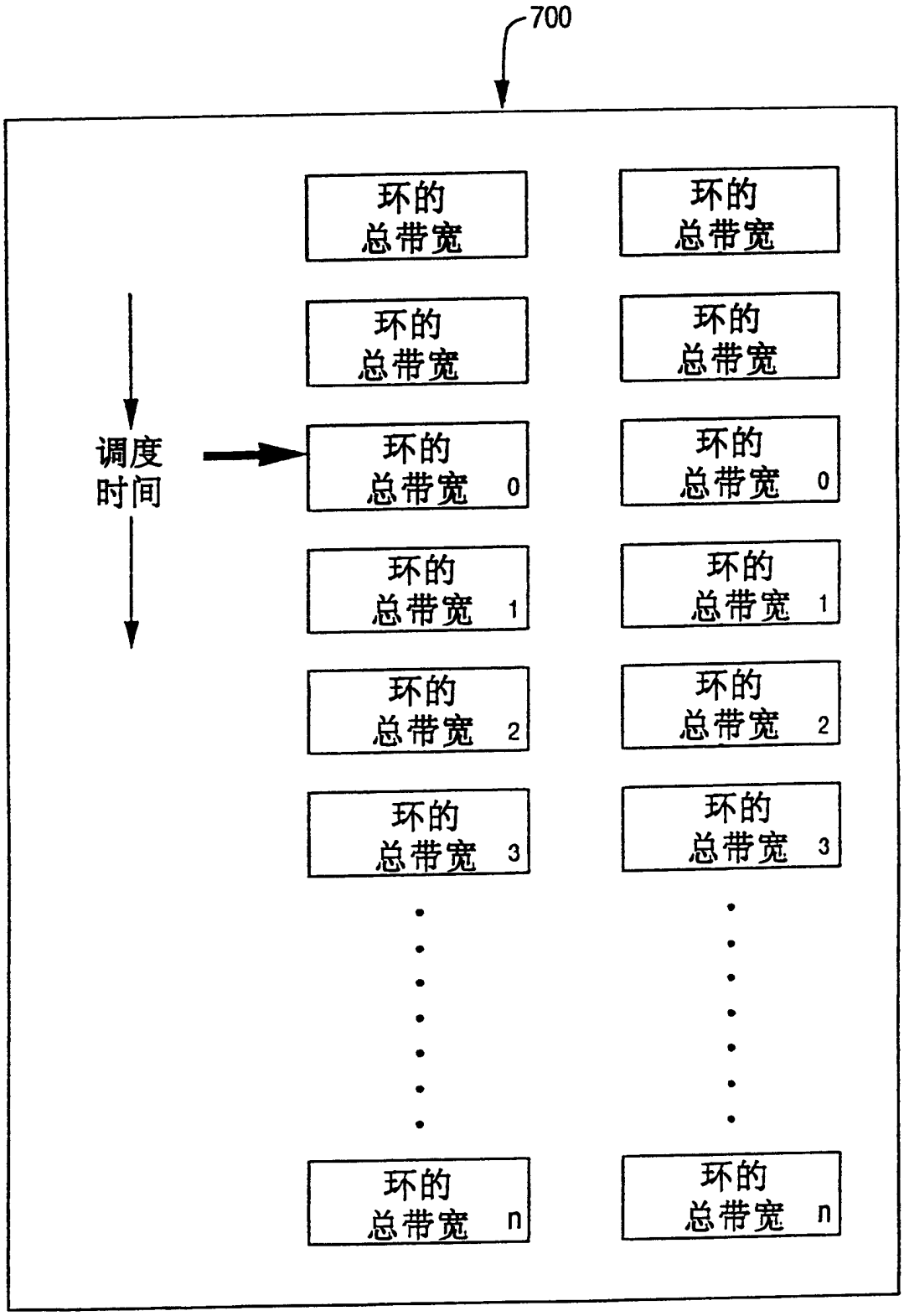

700
环的
总带宽
环的
总带宽
环的
总带宽
环的
总带宽
调度
时间
环的
总带宽 0
环的
总带宽 0
环的
总带宽 1
环的
总带宽 1
环的
总带宽 2
环的
总带宽 2
环的
总带宽 3
环的
总带宽 3
环的
总带宽 n
环的
总带宽 n

图7

800

流 ID	有效	X-OFF	活动	P_ Ring	C_ Ring	分配	SIM Fanout	MU	保护	BE 加权	保留 速率	R 结束 时间	S 结束 时间	R_ OR	S_ OR	保留	备注
0 位	1 位	1 位	1 位	1 位	1 位	2 位	9 位	1 位	1 位	5 位	14 位	32 位	32 位	1 位	1 位	4 位	8.4 秒 @5 纳秒 clk
1	1	否		R1		发送	0000000000	0	保证	0	10Mbps	435ABC	435ABC				储备流
2	1	是		R2		接收	1000000000	0	BE(STD)	1	80Mbps	空值	空值				非储备流
3	1			R2		发送	0000000000	0		0	空值	空值	空值				未用入口
4	1			R1		接收	0000000001	0		63	1Gbps						储备流
5	1			R1		复制	1000000000	1	保证	0	100Mbps						提取作多点发送
.																	
.	0																
.																	
999999	1					插入				0	8.5Gbps	空值	空值				非储备流

图 8

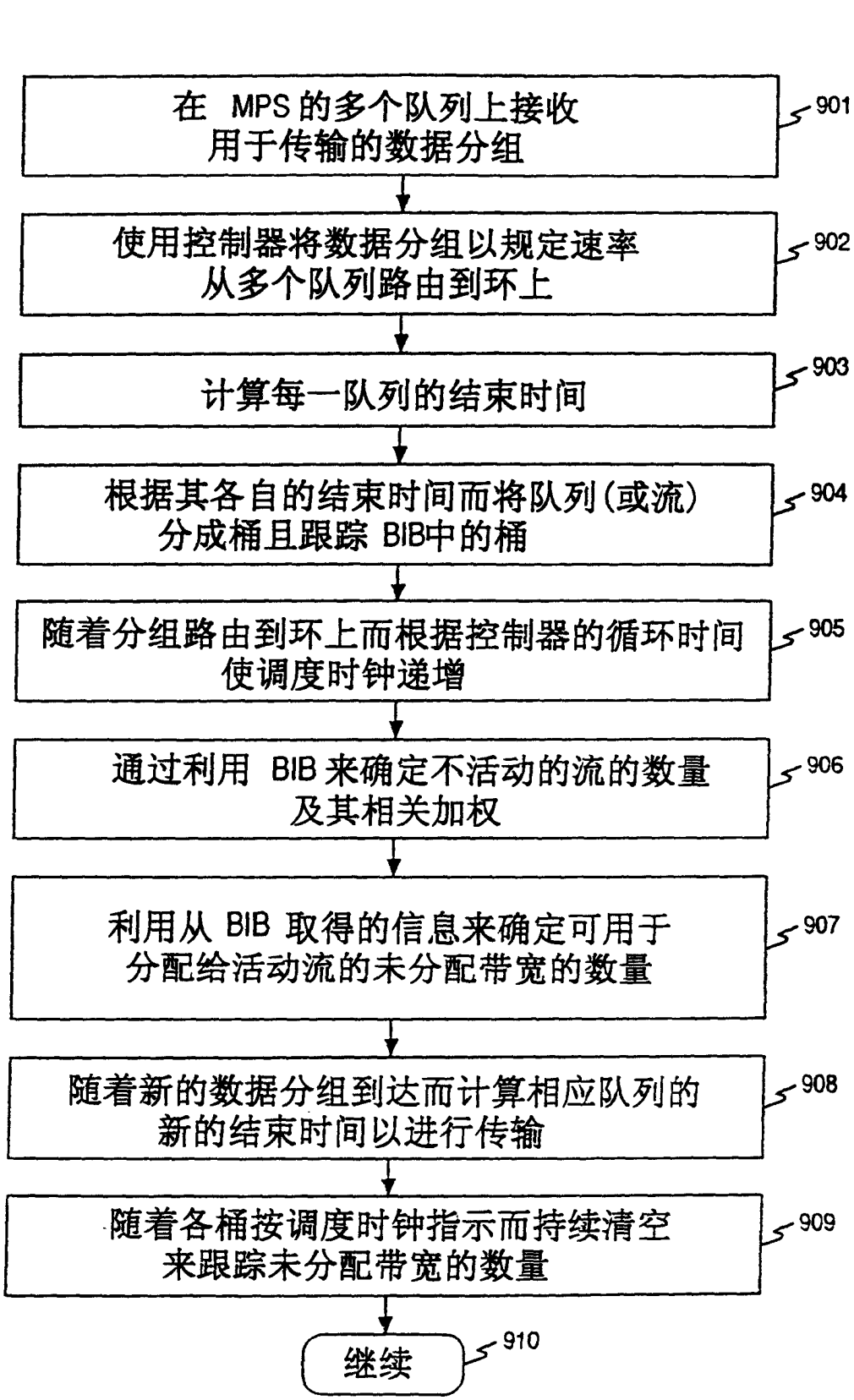

900
在 MPS 的多个队列上接收
用于传输的数据分组
901
使用控制器将数据分组以规定速率
从多个队列路由到环上
902
计算每一队列的结束时间
903
根据其各自的结束时间而将队列(或流)
分成桶且跟踪 BIB中的桶
904
随着分组路由到环上而根据控制器的循环时间
使调度时钟递增
905
通过利用 BIB 来确定不活动的流的数量
及其相关加权
906
利用从 BIB 取得的信息来确定可用于
分配给活动流的未分配带宽的数量
907
随着新的数据分组到达而计算相应队列的
新的结束时间以进行传输
908
随着各桶按调度时钟指示而持续清空
来跟踪未分配带宽的数量
909
继续
910

图9