

(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
2010 年 1 月 7 日 (07.01.2010)

PCT

(10) 国际公布号
WO 2010/000191 A1

- (51) 国际专利分类号:
H04L 12/56 (2006.01)
- (21) 国际申请号: PCT/CN2009/072494
- (22) 国际申请日: 2009 年 6 月 26 日 (26.06.2009)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
200810129144.1 2008 年 6 月 30 日 (30.06.2008) CN
- (71) 申请人 (对除美国外的所有指定国): 中兴通讯股份有限公司 (ZTE CORPORATION) [CN/CN]; 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦, Guangdong 518057 (CN)。
- (72) 发明人: 及
- (75) 发明人/申请人 (仅对美国): 孙明施 (SUN, Ming-shi) [CN/CN]; 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦, Guangdong 518057 (CN)。 杨宜 (YANG, Yi) [CN/CN]; 中国广东省深

圳市南山区高新技术产业园科技南路中兴通讯大厦, Guangdong 518057 (CN)。 张志伟 (ZHANG, Zhiwei) [CN/CN]; 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦, Guangdong 518057 (CN)。 李国栋 (LI, Guodong) [CN/CN]; 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦, Guangdong 518057 (CN)。 黄炜 (HUANG, Wei) [CN/CN]; 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦, Guangdong 518057 (CN)。 赖伟 (LAI, Wei) [CN/CN]; 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦, Guangdong 518057 (CN)。

- (74) 代理人: 北京康信知识产权代理有限公司 (KANGXIN PARTNERS, P.C.); 中国北京市海淀区知春路甲 48 号盈都大厦 A 座 16 层, Beijing 100098 (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, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,

[见续页]

(54) Title: METHOD AND DEVICE FOR PACKET SCHEDULING

(54) 发明名称: 一种分组调度方法和装置

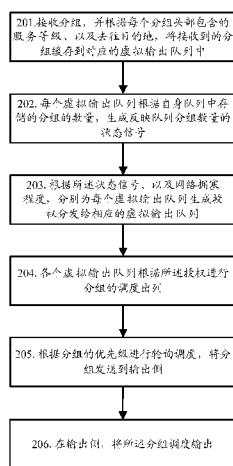


图 2 /Fig.2

(57) Abstract: A method for packet scheduling is disclosed by the present invention. The method includes: A, the received packet is buffered to the corresponding virtual output queue (VOQ) according to the class of service contained in the packet header and destination that the packet will be transmitted to; B, a grant is generated respectively for each VOQ according to the amount of packets in each VOQ and network congestion degree, and the grants are distributed to each VOQ; C, each VOQ performs packet scheduling and dequeuing according to the grant; D, a round-robin scheduling is performed according to the priority of the dequeued packet, and then the packet is scheduled and outputted. A device for implementing packet scheduling is disclosed by the present invention as well. The method and device enable the distributed scheduling of VOQ, and flexible scheduling, and thereby the hardware logic resources are saved.

[见续页]

201 A PACKET IS RECEIVED, AND THE RECEIVED PACKET IS BUFFERED TO THE CORRESPONDING VIRTUAL OUTPUT QUEUE ACCORDING TO THE CLASS OF SERVICE CONTAINED IN THE PACKET HEADER AND DESTINATION THAT THE PACKET WILL BE TRANSMITTED TO
202 EACH VIRTUAL OUTPUT QUEUE GENERATES THE STATE SIGNAL THAT REFLECTS THE PACKET AMOUNT OF THE QUEUE ACCORDING TO THE STORED PACKET AMOUNT IN ITS OWN QUEUE
203 A GRANT IS GENERATED RESPECTIVELY FOR EACH VIRTUAL OUTPUT QUEUE ACCORDING TO THE STATE SIGNAL AND NETWORK CONGESTION DEGREE, AND THE GRANTS ARE DISTRIBUTED TO THE CORRESPONDING VIRTUAL OUTPUT QUEUE
204 EACH VIRTUAL OUTPUT QUEUE PERFORMS PACKET SCHEDULING AND DEQUEUING ACCORDING TO THE GRANT
205 A ROUND-ROBIN SCHEDULING IS PERFORMED ACCORDING TO THE PACKET PRIORITY, AND THE PACKET IS TRANSMITTED TO THE OUTPUT SIDE
206 THE PACKET IS SCHEDULED AND OUTPUTTED IN THE OUTPUT SIDE

WO 2010/000191 A1



GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

KG, KZ, MD, RU, TJ, TM), 欧洲 (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, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

本国际公布:

(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, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY,

— 包括国际检索报告(条约第 21 条(3))。

(57) 摘要:

本发明公开了一种分组调度方法, 该方法包括: A、将接收到的分组根据分组头部包含的服务等级、以及去往目的地, 缓存到相应的 VOQ 中; B、根据各个 VOQ 中的分组的数量、以及网络拥塞程度, 分别为各个 VOQ 生成授权, 将授权分发到各个 VOQ; C、各个 VOQ 根据授权进行分组的调度出列; D、根据出列分组的优先级进行轮询调度, 之后, 将分组调度输出。本发明同时公开了一种实现分组调度的装置, 该方法和装置可以实现对 VOQ 的分布式调度, 调度灵活, 节约了硬件逻辑资源。

一种分组调度方法和装置

技术领域

本发明涉及网络的分组传送技术，尤其涉及一种分组调度方法和装置。

背景技术

5 随着通信技术的不断发展，运营商需要满足用户全方位的业务需求。从电信网络、广电网络和移动网络的融合，以及目前国内的电信重组可以看出，通过各种业务的融合以使运营商具备全业务运营能力为大势所趋。这就要求一个通信网络必须支持多种业务，但是，每种业务有不同的服务质量（QoS）要求。QoS 是指网络在传输数据流时要求满足的一系列服务请求及实现这些
10 服务请求的机制，这些服务请求可用一系列指标来衡量，例如：带宽要求、传输延迟、抖动、丢包率、吞吐量等。

 通信网络的 QoS 能力与所述通信网络采用的分组调度机制有着密切的关系。一个通信网络可以看成是一个非常复杂的排队系统，调度算法要在延迟、实现的复杂性和公平性之间进行综合考虑，以达到在保证服务质量的同时
15 提高网络资源利用率的目的。

 为很好应对上述挑战，一方面，分组调度机制应该使运营商可以提供基于网络间互联协议（IP）的各种差异化服务（基于服务层协议），如网络电话（VoIP）、交互式网络电视（IPTV）、视频等服务，以便给运营商提供新的利润增长点；另一方面，分组调度机制应该使运营商能够充分利用现有的网络
20 资源，实现基于业务的精细化带宽管理和分配。

 目前的分组调度机制中常采用数据推动型的调度结构，即：只有分组到达调度节点后，才进行相应的调度。这种调度结构一般分两次调度，输入侧的调度和输出侧的调度。输入侧将用户的流量进行调度后传输给输出侧，由输出侧进行调度输出。下面详细介绍输入侧的调度，输入侧调度一般基于队列进行，按层次进行调度，目前常见的设计一般支持 3~5 级层次。不同层次
25 的调度原理基本相同，但是，层次越多，调度越灵活。

下面，通过图 1 所示的抽象模型对输入侧的调度进行详细说明。上述的按层次进行调度抽象模型中各部分的主要功能以及处理过程如下：

A、接入网设备感知业务类型：根据不同的业务类型如 VoIP、IPTV、数据业务等，以及不同的用户，把分组放入相应的队列中。

5 并且，接入网设备可以根据要求对不同的分组进行流量的监管和控制。

B、第一层调度节点一般用来表示不同的用户，用于直接获得用户的流量。第一层调度节点根据调度节点签订的服务层协议的不同，进行相应的流量的监管和控制，根据流量限制的结果来调度下一层的分组。并且，由于用户优先级的不同，可以配合相应的调度算法，例如严格优先级（SP）、加权公平队列（WFQ）、轮询（RR）等对用户进行调度，首先保证高优先级用户的流量，剩余流量在优先级相同的用户之间分配。

10

C、第二、第三层次的调度节点连接下一层次的节点，一般可以表示接入层的设备，具有一定的带宽限制。各个调度节点根据自身上一层次的需求来调度自身下一个层次的分组，可以根据对应带宽的不同，进行相应的流量控制和调度。第二、第三层次的调度节点也应该支持基本的调度算法。

15

D、第四层次的调度节点位于调度模型的最高层次，可以用来表示该设备的实际输出接口（Egress Port），用于根据输出接口的配置情况进行流量控制，它同样也支持多种调度算法，以从不同的下层节点中选择合适的报文。

这种数据推动型调度方式的配置较为简单，节点数目较多。但由于调度层次已经固定，存在灵活性较低的问题。同时，这种调度方式只有在流量到达调度节点后才能进行调度，并且，输入侧的调度未能考虑网络拥塞程度，从而，输入侧的流量将全部被调度传输到输出侧进行缓存，等待输出侧进行调度输出，这不但可能导致输出侧的调度出现拥塞，使得输入侧和输出侧的配合出现问题，而且，大大耗费了输出侧的硬件逻辑资源。另外，由于调度层次固定，且不可配置，所以对于各种标准的支持有限，只能保证有限的服务质量。

20

25

发明内容

有鉴于此，本发明的主要目的在于提供一种分组调度方法和装置，可以实现对虚拟输出队列（VOQ）的分布式调度，调度灵活，且节约了硬件逻辑

资源。

为实现上述目的，根据本发明的一个方面，提供了一种分组调度方法。

根据本发明的分组调度方法包括：

5 A、将接收到的分组根据分组头部包含的服务等级（CoS）及去往目的地，缓存到相应的 VOQ 中；

 B、根据各个 VOQ 中分组的数量、以及网络拥塞程度，分别为各个 VOQ 生成授权，将各个授权分发到各个 VOQ；

 C、各个 VOQ 根据接收到的授权进行分组的调度出列；

 D、根据出列分组的优先级进行轮询调度，之后，将分组调度输出。

10 其中，根据各个 VOQ 中分组的数量、以及网络拥塞程度，为各个 VOQ 相应分配授权具体为：

 首先各个 VOQ 根据自身的分组数量生成状态信号；

 之后，根据上述状态信号、以及网络拥塞程度，分别为各个 VOQ 生成授权。

15 优选地，将授权分发的方法具体为：

 A1、根据应用场景和服务层协议，建立满足业务应用的网络模型，并根据该网络模型，配置调度单元和调度通道的连接关系，建立与网络模型层次相匹配的调度层次；

20 B1、将生成的授权在调度层次中通过调度单元与调度通道逐层下发，直到调度层次的最底层调度层；

 C1、最底层调度层中的调度单元把授权相应分发到各个 VOQ。

为实现上述目的，根据本发明的另一方面，提供了一种授权分发方法。

根据本发明的授权分发方法包括：

25 a、根据应用场景和服务层协议，建立满足业务应用的网络模型，并根据该网络模型，配置调度单元和调度通道的连接关系，建立与网络模型层次

相匹配的调度层次;

b、将生成的授权在调度层次中通过调度单元与调度通道逐层下发,直到调度层次的最底层调度层;

c、最底层调度层中的调度单元把授权相应分发到各个 VOQ。

5 为实现上述目的,根据本发明的再一个方面,提供了一种实现分组调度的装置。

根据本发明的实现分组调度的装置包括:入队管理模块、VOQ、队列状态管理模块、调度模块、授权调度模块、以及输出模块,其中,

10 入队管理模块,用于接收分组,并根据每个分组头部包含的 CoS 及去往目的地,将接收到的分组缓存到对应的 VOQ;

VOQ,用于缓存分组,并在调度模块进行分组调度时,输出相应的分组;

15 队列状态管理模块,用于根据各个 VOQ 中存储的分组的数量,生成反映队列分组数量的状态信号,并将该状态信号发送给授权调度模块;还用于接收授权调度模块分发来的授权,并比较各个 VOQ 的授权与 VOQ 的分组数量,当确定允许分组出列时,发送相应的出列允许信号到调度模块;

调度模块,用于接收到上述出列允许信号后,根据 VOQ 的优先级对相应 VOQ 中的分组进行轮询调度,之后,将调度得到的分组向外发送;

20 授权调度模块,用于根据上述状态信号、以及网络拥塞程度,生成各个 VOQ 的授权,并将授权分发到相应的队列状态管理模块。

优选地,该装置进一步包括:

输出模块,用于接收调度模块发来的分组,并将接收到的分组缓存,对分组进行调度输出。

25 优选地,授权调度模块进一步包括:授权生成模块、调度单元、以及调度通道,其中,

授权生成模块,用于生成授权,并将授权发送到相应的调度单元;

调度单元，用于将接收到的授权发送给相应的位于自身下层的调度单元；还用于在调度单元为最底层调度单元时，将授权分发给相应的 VOQ；

调度通道，用于在调度单元之间传输授权；还用于将最底层调度单元发送的授权传输给队列状态管理模块。

- 5 调度模块进一步还可以用于：将需要发送给输出模块的分组进行切片处理，将获得的数据信元发送给输出模块；

相应的，输出模块进一步还可以用于：将接收到的上述数据信元重组，将重组得到的分组进行缓存。

优选地，输出模块进一步包括：

- 10 分组重组模块，用于将接收到的数据信元进行重组，将重组得到的分组发送给调度输出模块；

调度输出模块，用于缓存分组，并对分组进行调度输出。

为实现上述目的，根据本发明的再一个方面，提供了一种实现授权分发的装置。

- 15 根据本发明的实现授权分发的装置包括：授权生成模块、调度单元、以及调度通道，其中，

授权生成模块，用于生成授权，并将授权发送到相应的调度单元；

调度单元，用于将接收到的授权发送给相应的位于自身下层的调度单元；还用于在调度单元为最底层调度单元时，将授权分发给相应的 VOQ；

- 20 调度通道，用于在调度单元之间传输授权；还用于将最底层调度单元发送的授权传输给相应的 VOQ。

- 25 本发明所提供的分组调度方法和装置，分组存储于 VOQ 中，对分组进行调度时，通过为 VOQ 分配授权的方式对 VOQ 中的分组进行调度，无需各个分组到达各个调度节点完成调度，而是以授权逐层分发到 VOQ 进行分组的调度，实现了对于 VOQ 的分布式调度。

而且，通过本发明可以任意划分授权分发层次，使用任意的调度算法，

使得调度过程更为灵活。另外，通过本发明可以在输入侧进行分组的缓存、以及简单的出列调度，在输出侧根据 VOQ 的状态信息、以及网络拥塞程度完成授权的生成与分发，从而使得输出侧只需缓存被调度出列的部分分组，而无需对所有分组进行缓存，节省了输出侧的硬件逻辑资源。

5 附图说明

图 1 是根据现有技术输入侧的调度方式的抽象模型结构示意图；

图 2 是根据本发明实施例的分组调度方法流程示意图；

图 3 是根据本发明实施例的授权分发方法流程示意图；

图 3a 是根据本发明实施例的调度单元中调度结构示例图；

10 图 4 是根据本发明实施例的实现分组调度的装置结构示意图；

图 5 是根据本发明实施例的实现授权分发的装置结构示意图。

具体实施方式

本发明的基本思想是：将接收到的分组根据分组头部包含的服务等级（CoS）及去往目的地，缓存到相应的 VOQ 中；根据各个 VOQ 中分组的数量、以及网络拥塞程度，为各个 VOQ 相应分配授权；各个 VOQ 根据所述授权进行分组的调度出列。

15

以下，通过具体实施例结合附图详细说明本发明分组调度方法和装置的实现。需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。

20 根据本发明的实施例，提供了一种分组调度方法。

图 2 为本发明分组调度方法流程示意图，该分组调度方法中分为输入侧、以及输出侧的调度，输入侧与输出侧之间通过交换网实现通信，如图 2 所示，该方法包括：

步骤 201：接收分组，并根据每个分组头部包含的 CoS、以及去往目的地，将接收到的分组缓存到对应的 VOQ 中。

25

CoS 是指把流量分配为不同的等级，每个等级可以提供不同的时延、抖动和丢包特点。

可以在本步骤中执行之前根据 CoS、以及流量的去往目的地建立不同的 VOQ，从而当接收到分组后，可以根据 CoS、以及去往目的地将分组区分开

5

步骤 202：每个 VOQ 根据自身队列中存储的分组的数量，生成反映队列分组数量的状态信号。

其中，状态信号可以分为：常态、缓态、以及空态，分别表示 VOQ 中具有较多分组、较少分组、或没有分组。其中，在实际应用中，可以为每个 VOQ 设定两个门限值 A、B ($A > B$)，当 VOQ 需要生成状态信号时，将自身存储的分组数量与上述两个门限值 A、B 进行比较，当分组数量大于等于 A 时，确定状态信号为常态，当分组数量处于 A 与 B 之间时，确定状态信号为缓态，而当分组数量小于 B 时，确定状态信号为空态，从而完成上述三种状态的区分。

10

各个 VOQ 的门限值可以分别设定，也可以为所有的 VOQ 设定统一的门限值，这里并不限定。但是，所述门限值最好根据各个 VOQ 的类型来配置。

15

步骤 203：根据所述状态信号、以及网络拥塞程度，分别为每个 VOQ 生成授权，并把生成的授权分发给相应的 VOQ。

将授权分发给各个 VOQ 的处理方法可以使用现有技术中的相关方法完成，或者，也可以使用图 3 所示的授权分发方法完成。

20

步骤 204：各个 VOQ 根据所述授权进行分组的调度出列。

其中，上述对分组进行调度，并输出分组的过程也称为分组的调度出列。各个 VOQ 对于分组的调度出列是通过将自身所包括的分组的数量与 VOQ 自身缓存的授权所代表的总字节数比较后确定的。每个授权代表一定的字节数，如果某一个 VOQ 自身累积的授权所代表的字节总数大于该 VOQ 头部的分组大小，就允许该分组出列，VOQ 缓存的授权不足就不能进行分组的调度出列。通过控制分发给各 VOQ 的授权数量，决定了各队列的输出带宽。

25

其中，本步骤中的各个 VOQ 的授权可以通过为每个 VOQ 维护一个令

牌桶整形器完成。将分发给各个 VOQ 的授权存储于相应的令牌桶整形器中，当 VOQ 中缓存有分组，且令牌桶整形器中缓存的令牌数足够时，允许对该 VOQ 进行调度。

步骤 205：根据分组的优先级进行轮询调度，将分组发送到输出侧。

5 步骤 206：在输出侧，将接收到的分组调度输出。

其中，输出侧对于分组的调度输出可以使用现有技术中的相关技术完成，这里不再赘述。

在步骤 205 和步骤 206 中，分组发送到输出侧并最终调度输出的方法可以认为：首先将分组进行切片处理，即：分组被切片成一个个固定大小的数据信元；数据信元通过交换网路由到输出侧；相应的，在输出侧，根据所述数据信元进行分组的重组，重组后的分组在输出侧进行缓存，等待被调度输出。

对于图 2 所示的方法中，一般将分组的入队、VOQ 的储存与出列调度放置于输入侧，而将授权的生成与分发、以及分组的调度输出放置于输出侧，输入侧与输出侧之间通过交换网相互通信，即：步骤 201、步骤 202、步骤 15 204、以及步骤 205 在输入侧执行，而步骤 203、步骤 206 在输出侧执行，执行过程中的信息交互，例如 VOQ 的状态信息、为 VOQ 生成的授权、以及发往输出侧的分组均通过交换网进行传输。这样，无需各个分组到达输入侧的各个调度节点完成调度，而是以授权逐层分发到 VOQ 进行分组的调度，实现了对于 VOQ 的分布式调度；且，节省了输出侧的硬件逻辑资源。

20 另外，所述状态信息、授权在通过交换网进行传输时，均可以为状态信息、或授权添加信元头，以状态信元、或授权信元的方式进行所述传输。

根据本发明的实施例，还提供了一种授权分发方法。

图 3 为本发明授权分发方法流程示意图，如图 3 所示，该方法包括：

步骤 301：根据应用场景和服务层协议，建立满足业务应用的网络模型，并根据所述网络模型，配置调度单元和调度通道的连接关系，建立与网络模型层次相匹配的调度层次。

例如：对于一个城域接入网，可以配置每个家庭对应两个队列，每栋楼对应第一层调度层，每个小区对应第二层调度层，每个街道对应第三层调度

层，每个区对应第四层调度层。如果把第四层设置为端口层，就相当于建立了一个四层调度结构。

步骤 302：根据 VOQ 的状态信号、以及网络拥塞程度，为每个 VOQ 生成相应的授权，并将授权逐层下发，直到调度层次的最底层调度层。

5 其中，授权在层次内的分发是通过调度单元或调度单元的集合完成的，不同层次间授权的传递是通过调度通道完成的。一个或多个调度单元完成一个调度层次的调度算法，并负责该层次内授权的分发。调度单元对授权进行调度时，可以首先对进入调度单元的授权使用严格优先级调度方法进行调度，之后，再根据不同的优先级对授权使用不同的调度方法，例如公平调度、加
10 权公平调度等进行调度，在上述调度方式下，调度单元内部的调度结构如图 3a 所示。调度通道对由调度单元构成的调度层次进行互联，并在层次间传送授权。

步骤 303：最底层调度层把授权相应分发到各个 VOQ。

15 最底层调度层接收来自 VOQ 的请求信元，并根据该层的调度算法，完成授权分发。完成分发的每个授权由调度通道输出，并加上信元头组成授权信元，通过交换网传送到 VOQ。

根据本发明的实施例，提供了一种实现分组调度的装置。

20 图 4 为本发明实现分组调度的装置结构示意图，如图 4 所示，该装置包括：入队管理模块 410、队列状态管理模块 420、调度模块 430、授权调度模块 440、输出模块 450、以及 VOQ 460，其中，

入队管理模块 410，用于接收分组，并根据每个分组头部包含的 CoS、以及去往目的地，将接收到的分组缓存到对应的 VOQ 460 中。

25 其中，在实际应用中，入队管理模块 410 还可以进一步增加功能，例如：对入队列的分组执行拥塞管理，具体地，拥塞管理可以采用加权随机早期丢弃、从头丢弃和尾部丢弃等算法来实现。

队列状态管理模块 420，用于根据各个 VOQ 460 中存储的分组的数量，生成反映队列分组数量的状态信号，并将该状态信号发送给授权调度模块 440。并且，队列状态管理模块 420 还用于接收授权调度模块 440 分发来的授权，并比较各个 VOQ 460 的授权与 VOQ 460 的分组数量，当确定允许分

组出列时，发送出列允许信号到调度模块 430。

其中，队列状态管理模块 420 的实现方式可以为：通过建立数据库维护并输出各个 VOQ 460 的状态；并对每个 VOQ 460 维护一个令牌桶整形器，接收到授权调度模块 440 发来的授权就存储在相应的令牌桶整形器中；只有
5 VOQ 460 中缓存有分组，并且对应的令牌桶整形器中缓存的令牌数足够时，输出相应的出列允许信号给调度模块 430。

在本发明所述的装置中，队列状态管理模块 420 的数量不少于一个。

其中，上述状态信号发送给授权调度模块 440 时，可以为状态信号加上信元头，以请求信元的方式完成发送；同样的，也可以为各个队列状态管理
10 模块 420 的授权加上信元头，以授权信元的方式完成发送。

调度模块 430，用于接收到上述出列允许信号后，根据 VOQ 460 的优先级对相应 VOQ 460 中的分组进行轮询调度，之后，将调度得到的分组发送给输出模块 450。

调度模块 430 的实现可以为：调度模块 430 根据调度出列允许信号建立一个 VOQ 的链表，该链表是一个所有满足分组出列条件的 VOQ 集合，调度
15 模块 430 通过该链表实现对分组的调度。

授权调度模块 440，用于根据上述状态信号、以及网络拥塞程度，生成各个 VOQ 460 的授权，将授权分发给队列状态管理模块 420。

输出模块 450，用于将得到的分组缓存，并进行调度输出。

20 VOQ 460，用于缓存分组，并在调度模块 430 进行分组调度时，输出相应的分组。

在图 4 所示的装置中，分组发送给输出模块 450 的方式可以信元的方式，即：调度模块 430 将需要发送的分组进行切片处理，获得数据信元发送给输出模块 450。相应的，输出模块 450 进一步用于：将接收到的数据信元进行
25 重组，将重组得到的分组进行缓存。此时，输出模块 450 还可以进一步分为两个模块：分组重组模块、以及调度输出模块，其中，

分组重组模块，用于将接收到的数据信元进行重组，将重组得到的分组发送给调度输出模块；

调度输出模块，用于缓存分组，并对分组进行调度输出。

对于图 4 所示的实现分组调度的装置，一般将入队管理模块 410、队列状态管理模块 420、以及调度模块 430 置于输入侧，而将授权调度模块 440、输出模块 450 置于输出侧，输入侧和输出侧之间通过交换网进行通信。

5 根据本发明的实施例，提供了一种实现授权分发的装置。

图 5 为实现授权分发的装置结构示意图，该装置可以应用于图 4 所示装置中，作为其中的授权调度模块 440。如图 5 所示，该装置包括：授权生成模块 510、调度单元 520、以及调度通道 530，其中，

10 授权生成模块 510，用于生成授权，并将授权发送到相应的调度单元 520。

调度单元 520，用于将接收到的授权发送给相应的位于自身下层的调度单元 520。还用于当调度单元 520 为最底层调度单元时，将授权分发给相应的 VOQ。

15 调度通道 530，用于在调度单元 520 之间传输授权；还用于将最底层调度单元 520 发送的授权传输给不同的 VOQ。

其中，上述将授权传输给不同的 VOQ 仅用于表示最底层调度单元分发的授权与 VOQ 之间的映射关系，在实际应用中，根据管理 VOQ 授权的装置、或模块不同，将相应将各个 VOQ 的授权发送给管理该 VOQ 授权的模块或装置。例如：当图 5 所示的装置作为图 4 所示装置中的授权调度模块 440 时，
20 由于图 4 中由队列状态管理模块 420 进行 VOQ 授权的管理，因此，调度通道 530 会将最底层调度单元 520 发送的授权传输给队列状态管理模块 420。

其中，由一定数量的调度单元构成一个调度层次。调度单元的数量、以及调度层次的数量根据具体情况自主设定，这里并不限制。例如，在步骤 301 的举例中，将调度层次分为了 4 层。

25 如上所述，通过本发明实施例提供的技术方案，将分组存储于 VOQ 中，对分组进行调度时，通过为 VOQ 分配授权的方式对 VOQ 中的分组进行调度，无需各个分组到达各个调度节点完成调度，而是以授权逐层分发到 VOQ 进行分组的调度，实现了对于 VOQ 的分布式调度。而且，通过本发明可以任意划分授权分发层次，使用任意的调度算法，使得调度过程更为灵活。另外，

通过本发明可以在输入侧进行分组的缓存、以及简单的出列调度，在输出侧根据 VOQ 的状态信息、以及网络拥塞程度完成授权的生成与分发，从而使得输出侧只需缓存被调度出列的部分分组，而无需对所有分组进行缓存，节省了输出侧的硬件逻辑资源。

- 5 以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

权利要求书

1. 一种分组调度方法，其特征在于，该方法包括：
 - A、将接收到的分组根据分组头部包含的服务等级 CoS 及去往目的地，缓存到相应的虚拟输出队列 VOQ 中；
 - B、根据各个 VOQ 中分组的数量、以及网络拥塞程度，分别为各个 VOQ 生成授权，将所述授权分发到各个 VOQ；
 - C、各个 VOQ 根据所述授权进行分组的调度出列；
 - D、根据出列分组的优先级进行轮询调度，之后，将分组调度输出。
2. 根据权利要求 1 所述的方法，其特征在于，所述根据各个 VOQ 中分组的数量、以及网络拥塞程度，为各个 VOQ 相应分配授权具体为：
 - 首先各个 VOQ 根据自身的分组数量生成状态信号；
 - 之后，根据所述状态信号、以及网络拥塞程度，分别为各个 VOQ 生成授权。
3. 根据权利要求 1 或 2 所述的方法，其特征在于，所述分发的方法具体为：
 - A1、根据应用场景和服务层协议，建立满足业务应用的网络模型，并根据所述网络模型，配置调度单元和调度通道的连接关系，建立与网络模型层次相匹配的调度层次；
 - B1、将生成的授权在调度层次中通过调度单元与调度通道逐层下发，直到调度层次的最底层调度层；
 - C1、最底层调度层中的调度单元把授权相应分发到各个 VOQ。
4. 一种授权分发方法，其特征在于，该方法包括：
 - a、根据应用场景和服务层协议，建立满足业务应用的网络模型，并根据所述网络模型，配置调度单元和调度通道的连接关系，建立与网络模型层次相匹配的调度层次；

b、将生成的授权在调度层次中通过调度单元与调度通道逐层下发，直到调度层次的最底层调度层；

c、最底层调度层中的调度单元把授权相应分发到各个 VOQ。

5. 一种实现分组调度的装置，其特征在于，该装置包括：入队管理模块、VOQ、队列状态管理模块、调度模块、授权调度模块、以及输出模块，其中，

入队管理模块，用于接收分组，并根据每个分组头部包含的 CoS 及去往目的地，将接收到的分组缓存到对应的 VOQ；

VOQ，用于缓存分组，并在调度模块进行分组调度时，输出相应的分组；

队列状态管理模块，用于根据各个 VOQ 中存储的分组的数量，生成反映队列分组数量的状态信号，并将所述状态信号发送给授权调度模块；还用于接收授权调度模块分发来的授权，并比较各个 VOQ 的授权与 VOQ 的分组数量，当确定允许分组出列时，发送相应的出列允许信号到调度模块；

调度模块，用于接收到所述出列允许信号后，根据 VOQ 的优先级对相应 VOQ 中的分组进行轮询调度，之后，将调度得到的分组向外发送；

授权调度模块，用于根据所述状态信号、以及网络拥塞程度，生成各个 VOQ 的授权，并将所述授权分发到相应的队列状态管理模块。

6. 根据权利要求 5 所述的装置，其特征在于，该装置进一步包括：

输出模块，用于接收调度模块发来的分组，并将接收到的分组缓存，对分组进行调度输出。

7. 根据权利要求 5 或 6 所述的装置，其特征在于，所述授权调度模块进一步包括：授权生成模块、调度单元、以及调度通道，其中，

授权生成模块，用于生成授权，并将授权发送到相应的调度单元；

调度单元，用于将接收到的授权发送给相应的位于自身下层的调度单元；还用于在调度单元为最底层调度单元时，将所述授权分发给相应的 VOQ；

调度通道，用于在调度单元之间传输授权；还用于将最底层调度单元发送的授权传输给队列状态管理模块。

8. 根据权利要求 5 或 6 所述的装置，其特征在于，调度模块进一步用于：将需要发送给输出模块的分组进行切片处理，将获得的数据信元发送给输出模块；

相应的，输出模块进一步用于：将接收到的所述数据信元重组，将重组得到的分组进行缓存。

9. 根据权利要求 8 所述的装置，其特征在于，所述输出模块进一步包括：

分组重组模块，用于将接收到的数据信元进行重组，将重组得到的分组发送给调度输出模块；

调度输出模块，用于缓存所述分组，并对分组进行调度输出。

10. 一种实现授权分发的装置，其特征在于，该装置包括：授权生成模块、调度单元、以及调度通道，其中，

授权生成模块，用于生成授权，并将授权发送到相应的调度单元；

调度单元，用于将接收到的授权发送给相应的位于自身下层的调度单元；还用于在调度单元为最底层调度单元时，将所述授权分发给相应的 VOQ；

调度通道，用于在调度单元之间传输授权；还用于将最底层调度单元发送的授权传输给相应的 VOQ。

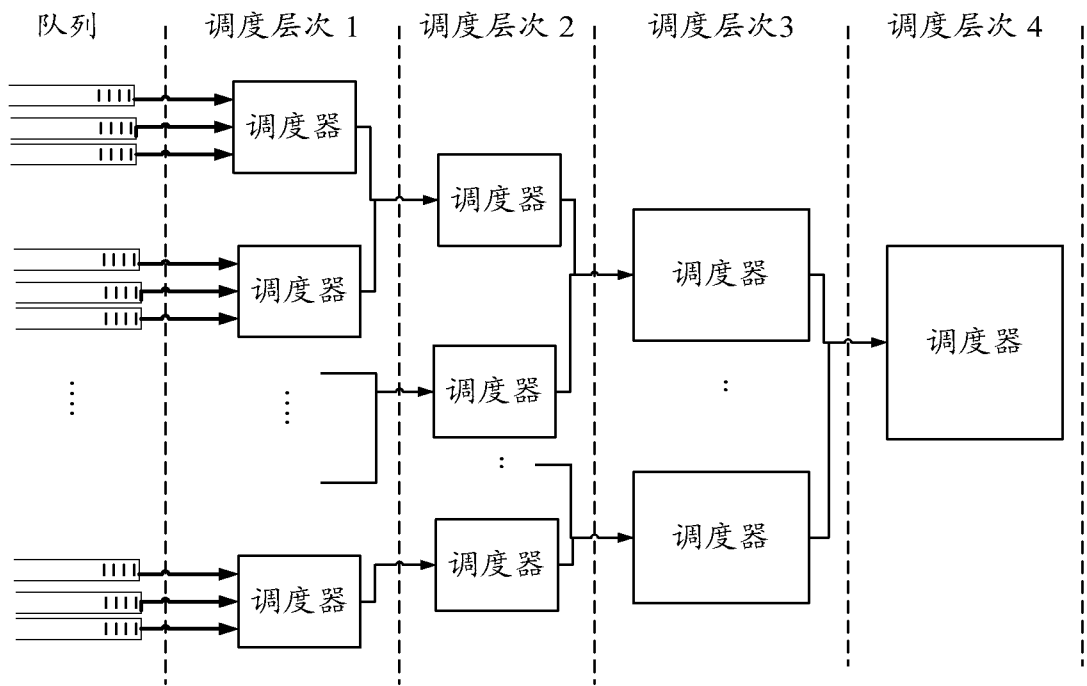


图 1

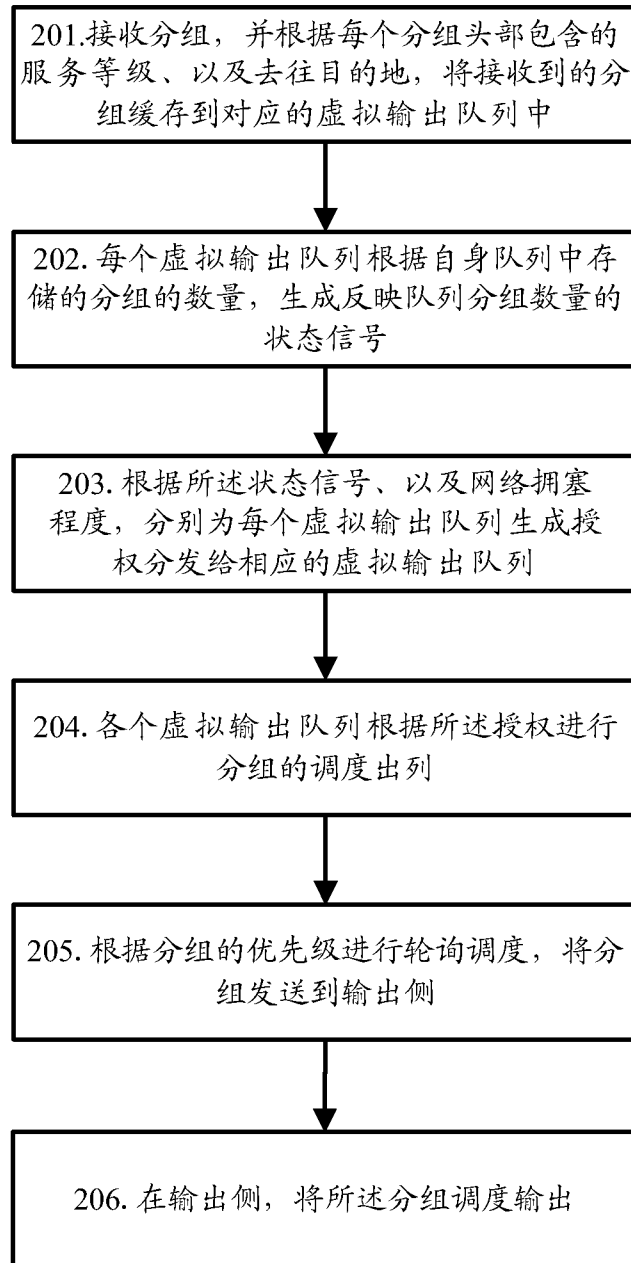


图 2

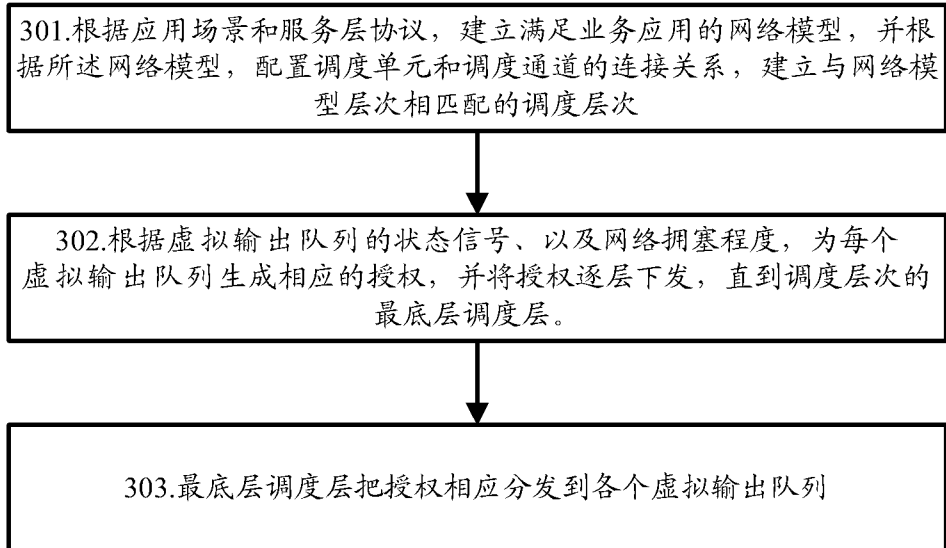


图 3

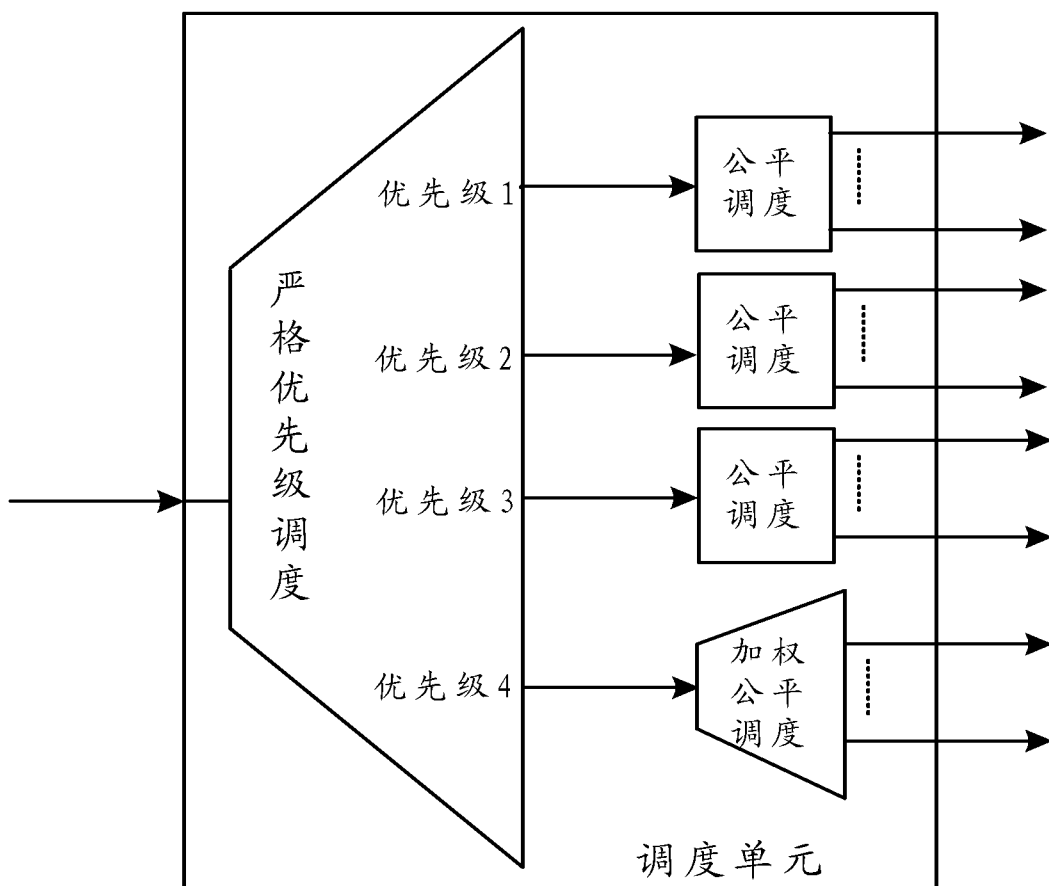


图 3a

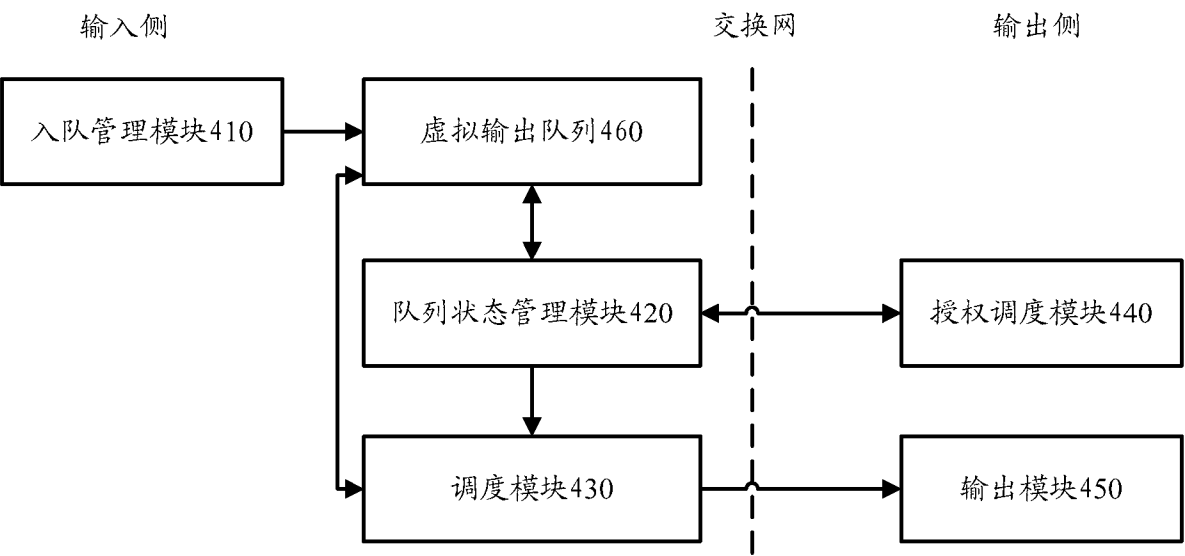


图 4

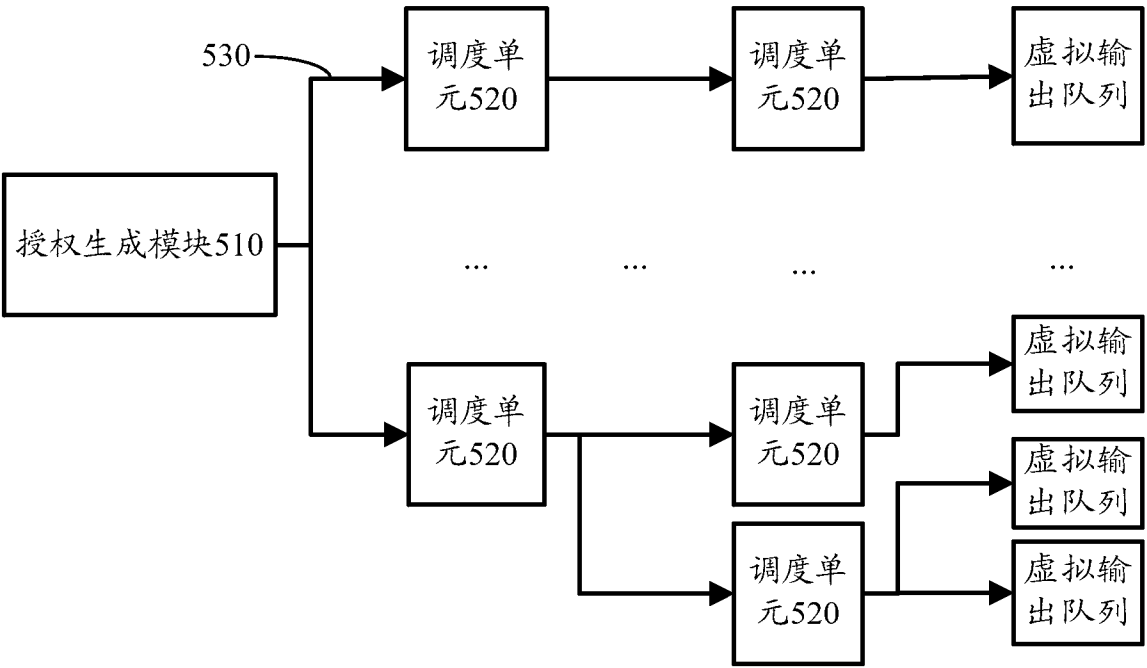


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2009/072494

A. CLASSIFICATION OF SUBJECT MATTER

H04L12/56(2006.01)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)

IPC: H04L12

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)

CNPAT, WPI, PAJ, EPODOC, IEEE, CNKI: CoS, VOQ, virtual, queue+, prior, congestion, destination, class, grant, load, address, schedule+, round, robin

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US2003/0227926A1(RAMAMURTHY, Gopalakrishnan et al.) 11 Dec.2003(11.12.2003) The whole document	1-10
A	US7120160B2 (TAKASE, Masayuki et al.) 10 Aug. 2006 (10.08.2006) The whole document	1-10
A	CN1826768A (CISCO TECHNOLOGY INC.) 30 Aug. 2006 (30.08.2006) The whole document	1-10

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&”document member of the same patent family
“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	

Date of the actual completion of the international search
02 Sep. 2009(02.09.2009)Date of mailing of the international search report
01 Oct. 2009 (01.10.2009)Name and mailing address of the ISA/CN
The State Intellectual Property Office, the P.R.China
6 Xitucheng Rd., Jimen Bridge, Haidian District, Beijing, China
100088
Facsimile No. 86-10-62019451Authorized officer
YANGZhen
Telephone No. (86-10)62413538

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2009/072494

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
US2003/0227926A1	11.12.2003	US2003227932A1	11.12.2003
		US2003227945A1	11.12.2003
US7120160B2	10.08.2006	JP2002252629A	06.09.2002
CN1826768A	30.08.2006	WO2005025153A1	17.03.2005
		US2005190779A1	01.09.2005
		EP1665663A1	07.06.2006

国际检索报告

国际申请号

PCT/CN2009/072494

A. 主题的分类

H04L12/56 (2006.01)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)

IPC: H04L12

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))

CNPAT, WPI, PAJ, EPODOC, IEEE, CNKI:调度, 虚拟队列, 服务等级, 优先, 拥塞, 授权, 轮询, 排队, 目的, 地址, CoS, VOQ, virtual, queue+, prior, congestion, destination, class, grant, load, address, schedule+, round, robin

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	US2003/0227926A1(RAMAMURTHY, Gopalakrishnan 等) 11.12 月 2003 (11.12.2003) 全文	1-10
A	US7120160B2 (TAKASE, Masayuki 等) 10.8 月 2006(10.08.2006) 全文	1-10
A	CN1826768A (思科技公司) 30.8 月 2006 (30.08.2006) 全文	1-10

☐ 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

02.9 月 2009 (02.09.2009)

国际检索报告邮寄日期

01.10 月 2009 (01.10.2009)

ISA/CN 的名称和邮寄地址:

中华人民共和国国家知识产权局

中国北京市海淀区蓟门桥西土城路 6 号 100088

传真号: (86-10)62019451

授权官员

杨震

电话号码: (86-10) 62413538

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2009/072494

检索报告中引用的 专利文件	公布日期	同族专利	公布日期
US2003/0227926A1	11.12.2003	US2003227932A1	11.12.2003
		US2003227945A1	11.12.2003
US7120160B2	10.08.2006	JP2002252629A	06.09.2002
CN1826768A	30.08.2006	WO2005025153A1	17.03.2005
		US2005190779A1	01.09.2005
		EP1665663A1	07.06.2006