(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
2008 年8 月7 日 (07.08.2008)

PCT

(10) 国际公布号
WO 2008/092404 A1

(51) 国际专利分类号:
H04L 12/56 (2006.01)

(21) 国际申请号: PCT/CN2008/070188

(22) 国际申请日: 2008 年1 月25 日 (25.01.2008)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
200710002741.3
2007 年1 月25 日 (25.01.2007) CN

(71) 申请人 (对除美国外的所有指定国): **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)。

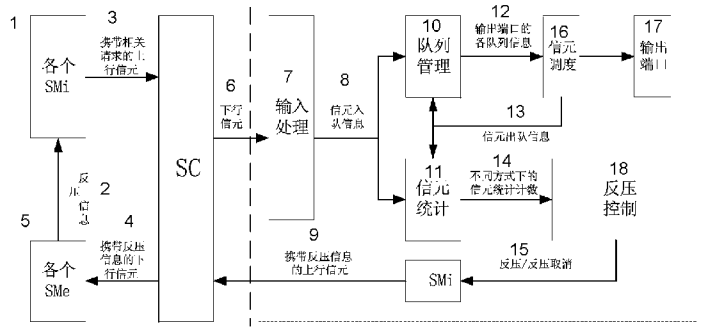
(72) 发明人; 及

(75) 发明人/申请人 (仅对美国): **杜文华(DU, Wenhua)** [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)。 **吴振耀(WU, Zhenyao)** [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)。 **唐德智(TANG, Dezhi)** [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)。 **罗焰斌(LUO, Yanbin)** [CN/CN]; 中国广东省深圳

[见续页]

(54) **Title:** AN OUTPUT QUEUE-BASED FLOW CONTROL REALIZATION METHOD AND APPARATUS

(54) **发明名称:** 一种基于输出队列的流控实现方法及装置



1 EACH SMI
2 BACK PRESSURE INFORMATION
3 UP-LINE CELLS CARRYING THE ASSOCIATED REQUEST
4 DOWN-LINE CELLS CARRYING THE BACK PRESSURE INFORMATION
5 EACH SME
6 DOWN-LINE CELLS
7 INPUTING PROCESS
8 CELLS INPUT QUEUE INFORMATION
9 UP-LINE CELLS CARRYING THE BACK PRESSURE INFORMATION
10 QUEUE MANAGEMENT
11 CELL STATISTIC
12 EACH QUEUE INFORMATION OF THE OUTPUT PORT
13 OUTPUT QUEUE INFORMATION OF THE CELLS
14 THE CELL STATISTICAL COUNTS BY THE DIFFERENT WAYS
15 BACK PRESSURE/BACK PRESSURE CANCELLATION
16 CELL SCHEDULING
17 OUTPUT PORT
18 BACK PRESSURE CONTROL

(57) Abstract: An output queue-based flow control realization method and apparatus mainly include: scheduling the queue and controlling the flow are carried out by calculating a number of cells using output port-based cell rank from different angles. In this system, the flow control management and the queue management are carried out dividually; the queue management is directly applied to schedule the cells, the flow control does not directly depend on the cell statistical result in the queue management, but calculates a number of cells using the different combination, from the angles of the priority of cells, the output port, the source chip number of cells etc. and the flow control is realized based on this. So the invention can reduce and simplify the number of the queues joining the scheduling and can make the back pressure control force exact and flexible.

[见续页]

WO 2008/092404 A1

市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(74) 代理人：北京凯特来知识产权代理有限公司(BEIJING CATALY IP ATTORNEY AT LAW); 中国北京市海淀区大柳树甲2号中铁科大厦8层南区, Beijing 100081 (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, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, 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, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, 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, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, 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, 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：
— 包括国际检索报告。

(57) 摘要：

公开了一种基于输出队列的流控实现方法及装置。主要包括：采用基于输出端口的信元排队和对信元从不同角度进行统计计数来实现队列的调度和流控。在这种体系下，流控和队列的管理是分开进行；所述的队列管理直接应用于信元的调度，所述的流控不是直接依赖于队列管理中的信元统计结果，而是通过从信元的优先级、输出端口、信元的源芯片号等角度出发，采用不同的组合对信元进行统计计数，并在此基础上实现流控。因此，本发明的实现可以使参加调度的队列数目少而简单，而且可以使反压控制力度做的很细，并且十分灵活。

说明书

一种基于输出队列的流控实现方法及装置

[1] 技术领域

[2] 本发明涉及网络通信领域，尤其涉及一种流控实现技术。

[3] 发明背景

[4] 目前，三级交换系统已经广泛应用于通信设备中。如图1所示，组成三级交换系统的芯片主要包括：

[5] （1）共享缓存交换芯片（SM，Switching Memory Chip），主要实现信元缓存、队列管理和调度等功能。

[6] （2）纵横交叉交换芯片（SC，Switching Crossbar Chip），主要完成对SM的请求进行仲裁、实现无阻塞空分交换以及对收集到的反压信息处理后转发给相应的SM。

[7] 其中，如图1所示，所述的SM进一步又可以分成以下两部分：

[8] （1）SMi，即SM的上行部分，有输入端口和输出端口，主要实现对需要交换的信元进行缓存、进行队列管理、发送请求给SC、根据SC送来的仲裁信号完成调度的功能。

[9] （2）SMe，即SM的下行部分，也是既有输入端口也有输出端口，主要实现把SC交换过来的信元进行缓存、实现队列管理、向SMi反映SMe的反压状况、完成信元的调度等功能。

[10] 在三级交换系统中，任何一个SMe可以接收来自任何一个SMi的信元；任何一个SMi发出的信元都可以通过SC交换到任何一个SMe。

[11] 这样，三级交换系统中，当各个SMi在一段时间内同时向某一个SMe发送信元时，就需要对来自于各个SMi的信元进行合理的管理，从而实现信元的合理调度。

[12] 另外，三级交换系统中，还需要提供相应的流控措施，以保证整个信元传输过程可以正常进行。

[13] 目前，在三级交换系统中，如图2所示，采用了相应的反压控制措施对相应的

信元传输过程进行控制，具体方案为：

[14] 当SMe检测到空闲缓存小于设定的阈值时，该SMe可以将相关信息（即反压信息）通知SMi，SMi将反压信息放在信元头发送给SC；

[15] SC通过信元头获取反压信息后，将反压信息发往各个SMe；

[16] 各个SMe收到反压信息后，将反压信息通知各自的SMi，各个SMi收到反压信息后，相应的有发送请求的SMi将停止发送请求。

[17] 当SMe_BP（指发生反压的SMe，BP，即back pressure，反压）发现空闲缓存大于相应的阈值，则采用和发送反压信息相同的手法，将反压取消的信息通过SC告诉各个SM，具体如图3所示，各个SM检测到反压信息已经被取消后，需要发送信息的SMi重新向SC发送请求，并可以通过SC继续将信元发给SMe。

[18] 发明内容

[19] 本发明的实施例提供一种基于输出队列的流控实现方法及装置，具有易于实现及流控机制可以灵活设置以满足实际应用的特点。

[20] 本发明的实施例提供了一种基于输出队列的流控实现方法，该方法包括：

[21] 统计获取信息接收端输入处理过程和信元调度过程中的信元处理信息；

[22] 根据统计获取的信元处理信息产生反压信息；所述反压信息用于对需要进行反压控制的信息发送端进行反压控制。

[23] 本发明的实施例提供了一种基于输出队列的流控实现装置，该装置包括：

[24] 信元统计单元，用于分别获取并统计信息接收端中输入处理过程和信元调度过程中的信元处理信息；

[25] 反压控制单元，用于根据统计获取的信元处理信息产生反压信息；所述反压控制信息用于对需要进行反压控制的信息发送端进行反压控制。

[26] 本发明的实施例提供了一种交换系统，所述交换系统包括：

[27] 共享缓存交换单元SM，所述SM包括：

[28] 信元统计单元，用于分别获取并统计信息接收端中输入处理过程和信元调度过程中的信元处理信息；

[29] 反压控制单元，用于根据统计获取的信元处理信息产生反压信息，所述反压信

息用于对需要进行反压控制的信息发送端进行反压控制；

[30] 所述系统还包括纵横交叉交换单元SC，用于接收所述反压信息，根据所述反压信息确定需要进行反压控制的信息发送端，并对所述确定的信息发送端进行反压控制。

[31] 附图简要说明

[32] 图1为现有技术提供的由SM和SC组成的三级交换系统的示意图；

[33] 图2为现有技提供的实现反压控制的流程图；

[34] 图3为现有技提供的实现反压取消控制的流程图；

[35] 图4为本发明实施例提供的实现信元流控机制的装置结构图；

[36] 图5为本发明实施例提供的实现信元调度的流程图；

[37] 图6为本发明实施例提供的实现信元流控的流程图。

[38] 实施本发明的方式

[39] 本发明实施例是从不同角度对信元进行统计计数以实现针对队列的流控。本发明实施例中，相应的流控功能的实现不再是直接依赖于队列管理中的信元统计结果，而是通过从信元的优先级、输出端口、信元的源芯片号等多角度出发，对信元进行统计计数，并在此基础上实现流控。

[40] 本发明实施例的信元队列流控技术方案应用的系统包括信息发送端和信息接收端，且在信息接收端中包括：用于接收输入的信元的输入处理单元及用于对队列管理单元管理的信元队列进行调度发送处理的信元调度单元；其中，所述的队列管理单元采用基于输出端口的队列管理方式。

[41] 基于上述系统，本发明实施例提供的对信元传输的流控机制则是根据对信元的各种统计信息，产生各种反压信息和反压取消信息，以达到对信元流控的目的，相应的具体处理过程包括：

[42] 1、统计信息接收端中所述输入处理单元和信元调度单元中的信元处理信息，具体包括：

[43] 对信元调度单元送进来的出队信息，根据信元的优先级和输出端口，确定信元的源芯片号，另一方面，结合输入处理单元送进来的入队信息，从信元的优先级、输出端口、信元的源芯片号等角度考虑，统计出相关信元的数目，并将统

计结果送给反压控制单元；

[44] 2、根据统计获取的针对各个单元的信元处理信息确定需要进行反压控制的信息发送端，并对相应的信息发送端进行反压控制；

[45] 在该过程中，所述的对相应的信息发送端进行的反压控制包括发起反压控制和反压取消控制两个过程，其中：

[46] （1）所述的发起反压控制的具体处理过程包括：

[47] 根据信息接收端的承载能力预先设定一组信元数量阈值，即设置相应的反压控制信元数量阈值，当统计输出端口的相关信元数量大于预先设定的该阈值，则针对该相关信元产生反压信息，并根据该相关信元处理信息的统计结果，通知该相关信元的信息发送端暂停发送；下面将举例对反压控制过程进行说明：

[48] 例一：设定针对某一优先级的反压控制信元数量阈值，若统计的各单元收到或处理的具有某一优先级的信元的数量超过该阈值，则产生反压信息，对发送相应优先级信元的信息发送端进行反压控制；另一方面，设定全局反压的阈值，当缓存中所有信元的数目超过该阈值，则可以发送全局反压信息，控制所有的SM停止往该SM发送信元；

[49] 相应的反压控制的应用实例可以为：

[50] 假设

SMe（A）的总的缓存数目为10000，对SMe（A）进行如下的设置：优先级为0的信元反压阈值为5000，优先级为1、2、3的信元的反压控制信元反压阈值为2000，全局反压控制阈值设置为7500，假设在某一时刻，统计单元的统计结果为：优先级为0的信元为4998个，优先级为1、2、3的信元为800个。假设在接下来的时间内，输入处理单元通知信元统计单元：在这一段时间内，优先级为0的信元新增加了20个，优先级为1、2、3的信元新增加3个。另一方面，信元调度单元告诉信元统计单元：在这一段时间内，优先级为0的信元被送出了2个，优先级为1、2、3的信元被送出了2个。信元统计单元根据输入处理单元和信元调度单元送来的信息，得出以下的统计信息：目前优先级为0的信元为5016个，优先级为1、2、3的信元为801个，缓存中信元的总数为7419个。信元统计单元把统计信息发送给反压控制单元，由于优先级为0的信元数目为5016个，超过了设定

的阈值5000，因此反压控制单元就会把优先级为0的反压信息通过上行的信元头携带给SC，SC收到该反压信息，就会告诉所有的SM停止往该SM发送优先级为0的信元，但是可以发送其他优先级的信元。这样经过一段时间，由于其他SM停止向该SMe发送优先级为0的信元，而该SMe（A）又不断的把优先级为0的信元调度出去，因此在缓存中的信元数目必然会不断减少，从而确保了缓存中优先级为0的信元不会严重偏离5000这一阈值。另一方面，假设又经过一段时间后，统计单元的统计结果为：优先级为0的信元为3000个，优先级为1、2、3的信元为1502个。这样整个缓存中的信元数目为7506个，超过了全局反压阈值7500。反压控制单元就会把该信息通过SC告诉所有的SM，所有SM就会停止往该SMe（A）发送信元，这样就可以避免SMe（A）中缓存被耗尽而不得不丢弃信元的现象。

[51] 上面所说的仅仅是实际应用中的一个例子。事实上，采用什么样的流控手段完全是根据实际需要来确定，比如说：如果实际应用中需要对各个SM进行不同的流控，可以针对不同的SM设定不同的阈值。整个反压的产生和取消的过程同上述的例子都是一样的。

[52] 上面提到的反压控制信元数量阈值是根据信息接收端的承载能力，从信元的优先级信息、信元的输出端口和/或信元的源端信息等角度考虑，并结合实际需要进行相关的设置，具体可以包括针对整个缓存全局反压的信元数量阈值，针对某个输出端口反压的信元数量阈值和/或针对某一个信元队列反压的信元数量阈值等；当然，还可以包含其他阈值，例如可以设置来自某个芯片或者某个优先级的信元数量阈值等。

[53] 反压控制信元数量阈值体现了不同的反压控制力度。这里举一些例子来说明。比如，统计输入处理单元缓存的所有信元的数目，可以用来发送当缓存不足的反压信号，这种反压称为全局反压；统计输出端口的信元数目或者统计某个输出端口中具有同一优先级的信元的数目，可以对需要从该输出端口出去的队列进行反压控制；对具有同源芯片号、同输出端口、同优先级的信元进行统计，可以具体控制某个队列的反压；总的来说，反压的控制力度都可以针对信元的优先级、输出端口、信元的源芯片号这几方面根据需要进行相关配置，通过不

同的组合方式来进行使用。

[54] （2）所述的反压取消控制的具体处理过程包括：

[55] 根据信息接收端的承载能力以及实际应用的需要预先设定一组信元数量阈值，即设置相应的反压取消控制信元数量阈值，当统计输出端口的相关信元数量小于或等于预先设定的该阈值，则针对该相关信元产生反压取消信息，并根据该相关信元处理信息的统计结果，通知该相关信元的信息发送端恢复发送；该反压取消控制信元数量阈值可以与之前的反压控制信元数量阈值相同，也可以各自独立设置；

[56] 其中，所述的反压取消控制信元数量阈值是根据信息接收端的承载能力，从信元的优先级信息、信元的输出端口和/或信元的源端信息等角度考虑，并结合实际需要进行相关的设置。反压取消控制信元数量阈值和前面提到的反压控制信元数量阈值是配套的，它的值可以与反压控制信元数量阈值相同，也可以各自独立设置。

[57] 由上述本发明实施例提供的技术方案可以看出，本发明的实施例可以使得对反压的控制基于对信元的各种统计来实现，而队列的管理直接作用于信元的调度。本发明实施例由于对反压控制和队列的管理分开，因此，可以使反压控制粒度根据需要进行相关调整，选择不同的组合方式来进行使用，从而在需要时可以使控制粒度做到很细。而且本发明实施例还具有管理简单，易于实现，占用资源较少，并具有灵活的流控机制以满足实际的应用的特点。

[58] 本发明实施例还提供有其它两种可供选择的流控和调度实现方案，下面将分别对这两种实现方案进行说明。

[59] （一）实现方案一

[60] 该方案为基于源芯片排队的流控和调度实现方案，该方案中具体是采用一种特定的队列管理机制，利用该机制同时实现流控和信元的调度。

[61] 基于源芯片排队的队列管理机制的原理为：假设在交换系统中有N个SM，每个SMe有M个输出端口，信元的优先级有W种。对于每一个SMe，其对收到的信元按照具有相同源SM号、相同输出端口、相同优先级的信元构成一个队列的原则进行分类，则将需要维护N×M×W个队列的信息，队列的信息包括队列的长度以

及队列中每一个信元在缓存中的位置等信息。

[62] 基于上述维护的N×M×W个队列的信息，在基于源芯片排队的队列管理机制中，相应的输出端口调度方案为：首先，选择输出端口，之后，选择源芯片号及优先级，最后，根据选择的结果在对应的队列中选择一个信元，并输出。当一个信元可以从SMe的某个输出端口出去时，则相应的队列的信元数目减一，并更新队列信息。

[63] 在基于源芯片排队的队列管理机制下，相应的队列反压的流控实现方案为：某个SMe（比如编号A），检测到来自一个SMi（比如编号B）的信元或者数据包的队列长度超过预定的阈值后，就单独通知SMi（B），让其停止或者降低速率向SMe（A）发送数据的流量。也就是说，SMe（A）对所有SMi的流控是可以做到各自独立的。相应的全局反压的流控实现方案为：当检测到某一SMe空闲缓存低于一定阈值时，则生成相应的全局反压信号，并通过本芯片的SMi传递给SC，由SC通知所有的SMi停止向该SMe发送信元。

[64] （二）实现方案二

[65] 该方案是基于输出端口排队实现流控和调度，即采用一种特定的队列管理机制，利用该机制可以同时实现基于输出端口的流控和信元的调度。

[66] 基于输出端口排队的队列管理机制的原理为：假设在交换系统中，每个SMe有M个输出端口，信元的优先级有W种。对于每一个SMe，其对收到的信元按照所有相同输出端口、相同优先级的信元构成一个队列的规则进行分类，共有M×W个队列，该队列管理机制需要维护M×W个队列的信息，队列的信息包括队列的长度以及队列中每一个信元在缓存中的位置等信息。

[67] 在基于上述的输出端口排队的队列管理机制下，对信元的调度方案包括：首先确定输出端口，接着选择优先级，然后根据选择的结果在对应的队列中选择一个信元并输出，当一个信元从SMe的某个输出端口出去的时候，相应的该队列的信元数目减一，同时需要对信元占用的缓存进行回收，并更新队列信息。

[68] 在基于上述的输出端口排队的队列管理机制下，对信元反压的流控实现过程包括：当某个SMe检测到需要从该端口出去的信元数目超出阈值的时候，该SMe就会生成相应的反压信号，并通过本芯片的SMi传递给SC，再由SC通知所有SMi都

停止向该端口发送数据流。相应的全局反压的流控实现方案包括：当检测到空闲缓存低于一定阈值的时候，生成相应的反压信号，该反压信号就会通过本芯片的SMi传递给SC，再由SC告诉所有的SMi停止往该SMe发送信元。

[69] 本发明实施例还提供了一种基于输出队列的流控实现装置，该装置应用的系统中包括信息发送端和信息接收端，且在信息接收端中依次包括输入处理单元、队列管理单元和信元调度单元，其中，所述的队列管理单元采用基于输出端口排队的队列管理方式直接应用于信元的调度；为实现本发明，在该装置中还设置了以下两个处理单元，具体为：

[70] 信元统计单元，一方面用于对信元调度单元送进来的出队信息，根据信元的优先级和输出端口，确定信元的源芯片号，另一方面，用于结合输入处理单元送进来的入队信息，从信元的优先级、输出端口、信元的源芯片号等角度出发，统计出相关信元的数目，相应的统计结果将被发送给反压控制单元；

[71] 反压控制单元，用于根据统计获取的信元处理信息确定需要进行反压控制的信息发送端，并对相应的信息发送端进行反压控制。

[72] 为了便于理解本发明，下面将结合如图4所示的具体实现结构图对各个单元的作用作具体的说明。

[73] （一）输入处理单元

[74] 所述的输入处理单元负责接收下行信元，同时把信元的入队信息作为信元处理信息发给队列管理单元和信元统计单元；其中，所述的入队信息主要包括信元的源端信息（即发送信元的源芯片）、信元的输出端口信息、信元的优先级信息等。

[75] （二）队列管理单元

[76] 所述的队列管理单元主要有两方面的功能，即入队管理和出队管理。

[77] 所述的入队管理是指队列管理单元会根据输入处理单元送来的入队信息，采用基于输出端口排队的方式，令所有具有相同输出端口、相同优先级的信元构成一个队列，例如，输出端口数为M，有W种优先级，则同时需要管理M×W个队列的信息；

[78] 所述的出队管理是指队列管理单元会根据信元调度单元送来的出队信息，将调

度出去的信元信息从相关的队列中删除。

[79] （三）信元统计单元

[80] 所述的信元统计单元根据当前的入队信息和出队信息（其中，在处理出队信息时，还会根据出队信元的输出端口和优先级找出信元的源芯片号），从信元的优先级、输出端口、信元的源芯片号等角度出发，统计出相关信元的数目，并且把统计结果送给反压控制单元，作为反压产生或取消的依据；比如说，统计缓存中所有信元的数目，可以用来发送当缓存不足的反压信号；统计输出端口的信元数目或者统计某个输出端口中具有同一优先级的信元的数目，可以对需要从该输出端口出去的队列进行反压控制，对具有同源芯片号、同输出端口、同优先级的信元进行统计，可以具体控制某个队列的反压；

[81] 所述的信元统计单元具体为针对信元的优先级、输出端口、信元的源芯片号等任意一种或多种信息的统计，以作为反压或反压取消控制的依据。

[82] （四）信元调度单元

[83] 所述的信元调度单元会根据队列管理单元送来的各个端口的信元数目来决定当前是否有信元可调度，如果当前端口有信元可调度，则进一步进行优先级的选择，从而完成从M×W个队列中选出一个队列的操作，然后把排在该队列最前面的信元调度出去，这样就确保了最早到达的信元可以最早得到服务，实现了信元的公平调度；同时还会把出队信息通知队列管理单元和信元统计单元。

[84] （五）反压控制单元

[85] 所述的反压控制单元主要是根据信元统计单元送过来的各种统计信息，发送各种反压信息以及各种反压取消的信息；

[86] 所述的反压控制单元进行反压的处理过程包括：根据预先设定的信元数量阈值，当统计输出端口的相关信元数量大于预先设定的阈值，则针对该相关信元产生反压信息；

[87] 同理，所述的反压控制单元进行反压取消的处理过程包括：根据预先设定的信元数量阈值，当统计输出端口的相关信元数量小于或等于预先设定的阈值，则针对该相关信元产生反压取消信息。

[88] 本发明实施方式中的反压控制单元包括发起反压控制模块和执行反压控制模块

。

[89] 发起反压控制模块根据信元统计单元发送来的各种统计信息确定出输出端口的相关信元数量大于预先设定的反压控制信元数量阈值时，通知执行反压控制模块。

[90] 执行反压控制模块根据发起反压控制模块的通知针对该相关信元产生反压信息，并发送。

[91] 本发明实施方式中的反压控制单元还包括反压取消控制模块。

[92] 在发起反压控制模块通知执行反压控制模块后，反压取消控制模块根据信元统计单元发送来的各种统计信息确定出输出端口的相关信元数量小于或等于预先设定的反压取消控制信元数量阈值时，通知执行反压控制模块。

[93] 执行反压控制模块根据反压取消控制模块的通知针对该相关信元产生反压取消信息，并发送。

[94] 所述执行反压控制模块可以包括：全局反压模块、优先级反压模块、输出端口反压模块和队列反压模块中的一个或多个。

[95] 全局反压模块根据发起反压控制模块的通知产生针对整个缓存的信元进行反压控制的反压信息，并发送。

[96] 优先级反压模块根据发起反压控制模块的通知产生针对某个优先级的信元进行反压控制的反压信息，并发送。

[97] 输出端口反压模块根据发起反压控制模块的通知产生针对某个输出端口信元进行反压控制的反压信息，并发送。

[98] 队列反压模块根据发起反压控制模块的通知产生针对某一个队列的信元进行反压控制的反压信息，并发送。

[99] 本发明实施方式还提供一种交换系统，该交换系统包括：SC和至少一个SM，SM包括：信元统计单元和反压控制单元。

[100] 信元统计单元和反压控制单元如上述实施方式中的描述，在此不再重复说明。

[101] 下面将结合附图对本发明在某一信元队列进来后，各个单元是如何相互合作来实现信元的调度和流控为实施例对本发明的实现方案进行详细的说明。

[102] （一）图5所示的是本发明信元由输入到调度出去的过程，其具体步骤包括：

[103] 步骤51：信元从输入端口进来，输入处理单元把信元缓存起来，然后把相关的入队信息告诉队列管理单元和信元统计单元。所述的入队信息主要包括信元来自哪个芯片、要从哪个端口输出、信元的优先级等信息。

[104] 步骤52：队列管理单元和信元统计单元对入队信息进行处理；

[105] 所述的队列管理模块对入队信息的入队管理具体包括：根据输入处理单元送来的入队信息，采用基于输出端口排队的方式，让所有相同输出端口，相同优先级的信元构成一个队列。所述的信元统计单元对入队信息的处理具体包括：对输入处理单元送来的入队信息，根据信元的优先级、输出端口、信元的源芯片号等这几方面，通过不同的组合方式进行统计，统计信元的数目，并把统计结果发送给反压控制单元，作为反压产生的依据。

[106] 步骤53：信元调度单元检查到输出队列中信元个数不为0，把信元调度出去，同时返回出队信元的信息给队列管理单元和信元统计单元。

[107] 步骤54、步骤55：队列管理单元和信元统计单元对出队信息进行处理。所述的出队信息主要包括出队信元输出端口、信元的优先级等信息。所述的队列管理单元对出队信息的管理包括把已经调度出去的信元信息从队列中删除。所述的信元统计单元根据信元的出队信息，基于信元的优先级、输出端口确定信元的源芯片号，并根据信元的优先级、输出端口、信元的源芯片号等进行相关信元的数目的统计，之后，再将统计结果发给反压控制单元，以作为反压取消的依据。

[108] 通过上述过程，共享缓存交换芯片的下行部分SMe可以在接收到信元后，基于信元的优先级、输出端口等信息，对输出端口的信元队列进行排队来管理每一个端口的信元，从而实现信元的调度。

[109] （二）图6所示的是本发明实现信元流控的具体过程，为了描述方便，这里以如何对来自芯片A优先级为0的队列实现流控为例进行说明，实现其他类型的流控过程与该过程类似，其具体步骤如下：

[110] 反压控制过程包括：

[111] 步骤61：输入信元被正确接收并缓存后，输入处理单元将相关的队列信息发送给信元统计单元；

[112] 步骤66：当芯片A优先级0的信元被信元调度单元调度出去以后，相关的出队信息发给信元统计单元；

[113] 步骤62：信元统计单元根据入队信息和出队信息，对信元进行统计，同时把各种统计结果发送给反压控制单元；

[114] 步骤63：反压控制单元根据统计结果产生反压信号；

[115] 具体为：反压控制单元处理各种统计结果，并检测来自芯片A优先级为0的信元个数是否超出预先设置的阈值，如果超出阈值，则产生反压信号，并继续执行步骤64；

[116] 步骤64：反压信号通过下行信元发送给纵横交叉交换芯片SC，并通过SC转发该反压信号；

[117] 步骤65：纵横交叉交换芯片SC将反压信号发送给芯片A后，芯片A接收该反压信号则暂时停发0优先级的信元。

[118] 反压取消控制过程包括：

[119] 步骤61：输入信元被正确接收并缓存后，输入处理单元把相关的队列信息发送给信元统计单元；

[120] 步骤66：当芯片A优先级0的信元被信元调度单元调度出去以后，相关的出队信息发给信元统计单元；

[121] 步骤62：信元统计单元根据入队信息和出队信息进行统计，同时把各种统计结果发送给反压控制单元；

[122] 步骤67：反压控制单元产生反压取消信号；

[123] 具体为：反压控制单元处理各种统计结果，并检测来自芯片A优先级为0的信元个数是否低于阈值，如果是，则产生反压取消信号；

[124] 步骤68：反压取消信号通过下行信元发送给纵横交叉交换芯片SC，并由SC转发该反压取消信号；

[125] 步骤69：SC将反压取消信息发送给芯片A后，芯片A接收该反压取消信号则可以重新发送0优先级的信元。

[126] 通过上述过程，SMe可以在接收到信元后，基于对信元的优先级、输出端口、信元的源芯片号等不同角度进行设置，统计信元的数目，从而实现从粗到细对

反压进行控制，最粗可以实现针对整个缓存的全局反压，最细可以对来自某个芯片的某个优先级队列进行反压。

[127] 综上所述，本发明不仅可以实现对信元的调度，而且使参加调度的信元队列少，易于管理；同时可以实现多种反压控制粒度。

[128] 以上所述，仅为本发明较佳的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到的变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应该以权利要求的保护范围为准。

权利要求书

[1] 1、一种基于输出队列的流控实现方法，其特征在于，该方法包括：

统计获取信息接收端输入处理过程和信元调度过程中的信元处理信息；

根据统计获取的信元处理信息产生反压信息；所述反压信息用于对需要进行反压控制的信息发送端进行反压控制。

[2] 2、根据权利要求1所述的方法，其特征在于，所述的信元处理信息包括：

输入处理单元的入队信息和信元调度单元中的信元出队信息。

[3] 3、根据权利要求1所述的方法，其特征在于，所述的对输入处理单元以及信元调度单元的信元处理信息的统计处理包括：

对信元调度单元送进来的出队信息，根据信元的优先级和输出端口确定信元的源芯片号，并结合输入处理单元送进来的入队信息，基于信元的优先级、输出端口和源芯片号中的至少一项对相关信元的处理数量进行统计。

[4] 4、根据权利要求3所述的方法，其特征在于，所述根据统计获取的信元处理信息产生反压控制信息具体包括：当统计输出端口的相关信元数量大于预先设定的阈值，则针对该相关信元产生反压信息。

[5] 5、根据权利要求4所述的方法，其特征在于，所述的方法还包括对信息发送端进行反压取消控制，具体包括：当统计输出端口的相关信元数量小于或等于预先设定的阈值，则针对该相关信元产生反压取消信息；该反压取消信息用于通知该相关信元的信息发送端恢复信元的发送操作。

[6] 6、根据权利要求4所述的方法，其特征在于，所述预先设定的阈值为：根据信息接收端的承载能力及应用需求预先设定的一组反压控制信元数量阈值。

[7] 7、根据权利要求1、2或3所述的方法，其特征在于，所述的反压控制包括下述一项或任意多项：

针对整个缓存信元的全局反压控制；

针对某个优先级的信元的反压控制；

针对某个输出端口信元的反压控制；

针对某一个队列的信元的反压控制。

[8] 8、一种基于输出队列的流控实现装置，其特征在于，该装置包括：

信元统计单元，用于分别获取并统计信息接收端中输入处理过程和信元调度过程中的信元处理信息；

反压控制单元，用于根据统计获取的信元处理信息产生反压信息；所述反压控制信息用于对需要进行反压控制的信息发送端进行反压控制。

[9] 9、根据权利要求8所述的装置，其特征在于，所述的反压控制单元进行的根据所述统计获取的信元处理信息产生反压信息的处理过程包括：当统计输出端口的相关信元数量大于预先设定的反压控制信元数量阈值，则针对该相关信元产生反压信息。

[10] 10、根据权利要求8所述的装置，其特征在于，所述的反压控制单元还用于反压取消控制：根据预先设定的反压取消控制信元数量阈值，当统计输出端口的相关信元数量小于或等于预先设定的反压取消控制信元数量阈值，则针对该相关信元产生反压取消信息。

[11] 11、根据权利要求8、9或10所述的装置，其特征在于，所述的反压控制单元产生的反压信息包括针对整个缓存信元的全局反压控制的反压信息、针对某个优先级的信元的反压控制的反压信息、针对某个输出端口信元的反压控制的反压信息和针对某一个队列的信元的反压控制的反压信息中的至少一项。

[12] 12、一种交换系统，其特征在于，所述交换系统包括：

共享缓存交换单元SM，所述SM包括：

信元统计单元，用于分别获取并统计信息接收端中输入处理过程和信元调度过程中的信元处理信息；

反压控制单元，用于根据统计获取的信元处理信息产生反压信息，所述反压信息用于对需要进行反压控制的信息发送端进行反压控制；

所述系统还包括纵横交叉交换单元SC，用于接收所述反压信息，根据所述反压信息确定需要进行反压控制的信息发送端，并对所述确定的信息发送端进行反压控制。

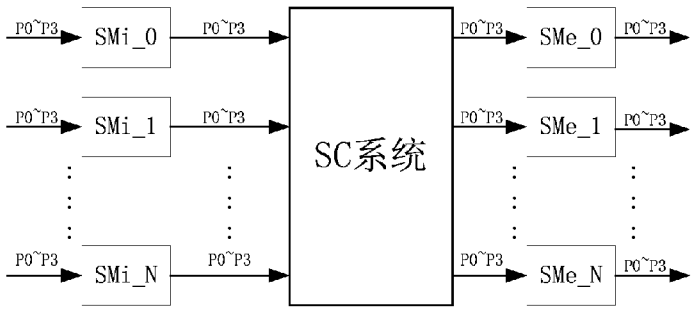

P0~P3
SMi_0
SMi_1
SMi_N
SC系统
SMe_0
SMe_1
SMe_N

图1

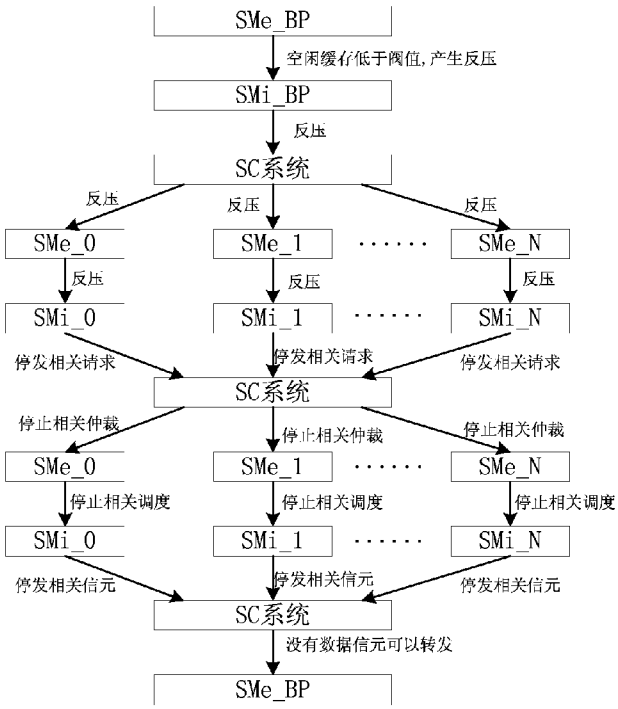

SMe_BP
空闲缓存低于阀值，产生反压
SMi_BP
反压
SC系统
SMe_0
SMe_1
SMe_N
SMi_0
SMi_1
SMi_N
停发相关请求
停止相关仲裁
停止相关调度
停发相关信元
没有数据信元可以转发

图2

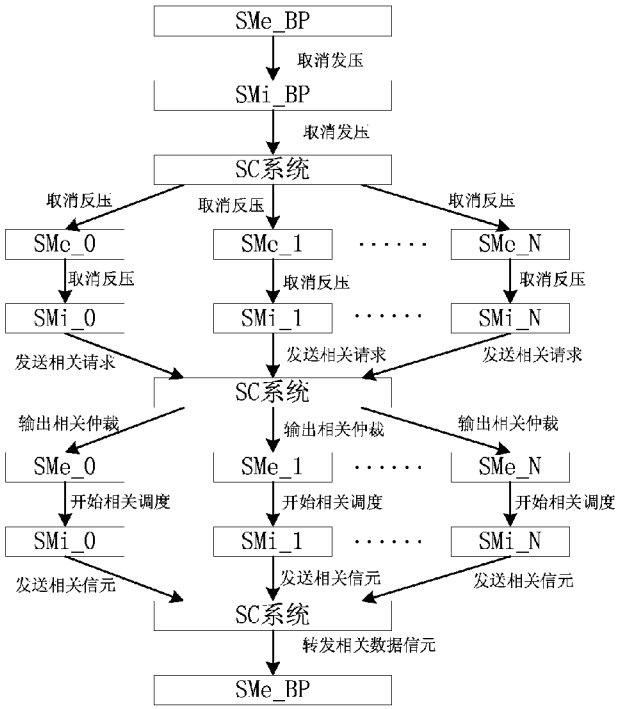

SMe_BP
取消发压
SMi_BP
取消发压
SC系统
取消反压
取消反压
取消反压
SMc_0
SMc_1
SMc_N
取消反压
取消反压
取消反压
SMi_0
SMi_1
SMi_N
发送相关请求
发送相关请求
发送相关请求
SC系统
输出相关仲裁
输出相关仲裁
输出相关仲裁
SMe_0
SMe_1
SMe_N
开始相关调度
开始相关调度
开始相关调度
SMi_0
SMi_1
SMi_N
发送相关信元
发送相关信元
发送相关信元
SC系统
转发相关数据信元
SMe_BP

图3

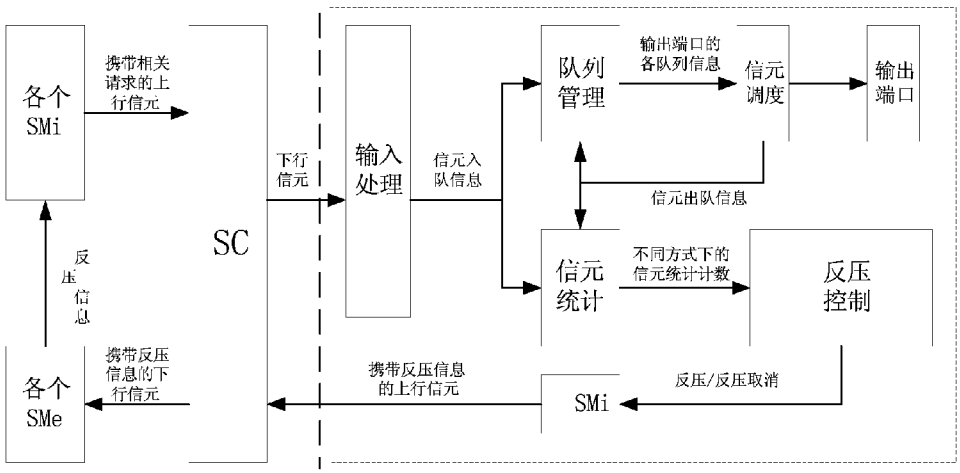

各个
SMi
携带相关
请求的上
行信元
SC
下行
信元
输入
处理
信元入
队信息
队列
管理
输出端口的
各队列信息
信元
调度
输出
端口
信元出队信息
信元
统计
不同方式下的
信元统计计数
反压
控制
反
压
信
息
各个
SMe
携带反压
信息的下
行信元
携带反压信息
的上行信元
SMi
反压/反压取消

图4

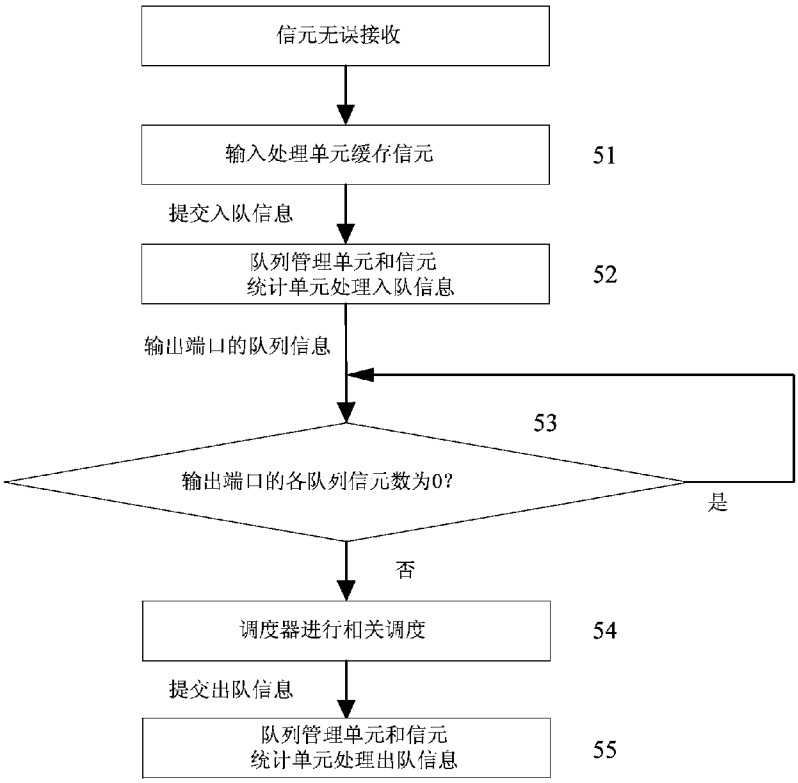

信元无误接收
输入处理单元缓存信元
51
提交入队信息
队列管理单元和信元
统计单元处理入队信息
52
输出端口的队列信息
53
输出端口的各队列信元数为0?
是
否
调度器进行相关调度
54
提交出队信息
队列管理单元和信元
统计单元处理出队信息
55

图5

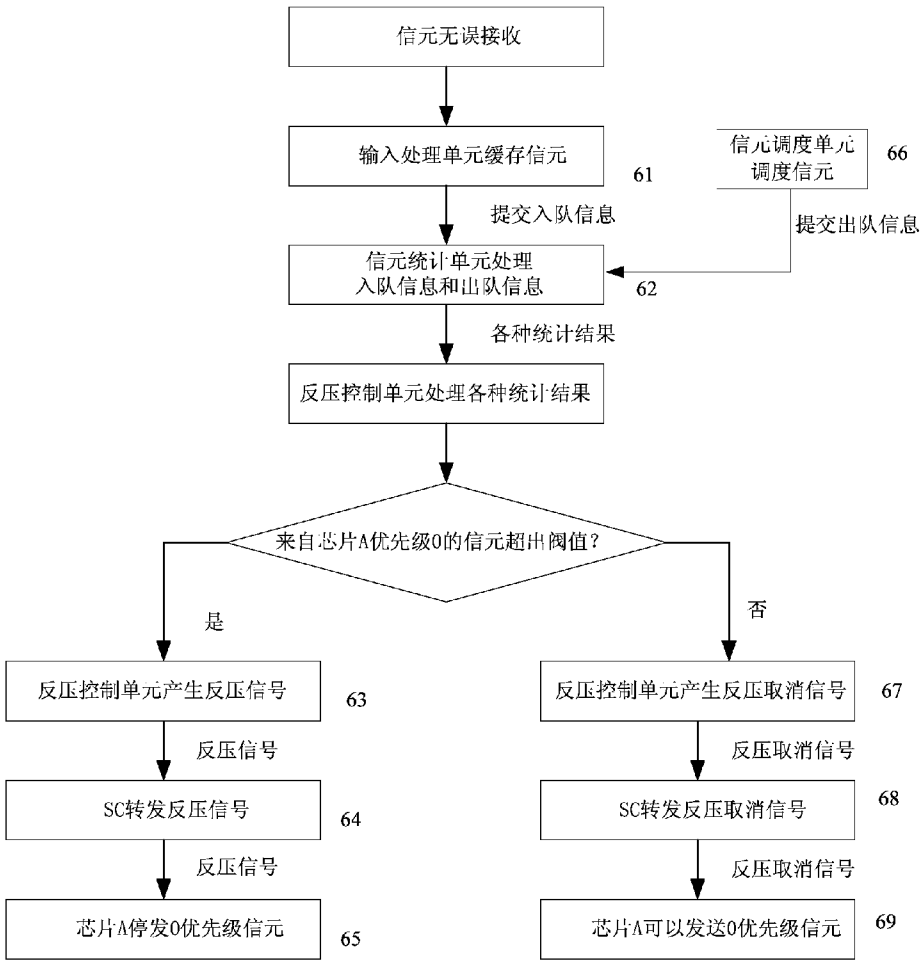

信元无误接收
输入处理单元缓存信元
61
信元调度单元调度信元
66
提交入队信息
提交出队信息
信元统计单元处理入队信息和出队信息
62
各种统计结果
反压控制单元处理各种统计结果
来自芯片A优先级0的信元超出阀值？
是
否
反压控制单元产生反压信号
63
反压信号
SC转发反压信号
64
反压信号
芯片A停发0优先级信元
65
反压控制单元产生反压取消信号
67
反压取消信号
SC转发反压取消信号
68
反压取消信号
芯片A可以发送0优先级信元
69

图6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2008/070188

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:H04L,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)

WPI, EPODOC, PAJ, CNPAT,CNKI: schedul+, queue, orde+, rank, priority, stat+, comput+, bk, backpress+, back press+, input+, output+, ingress, egress

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Journal of electronics and information technology, Vol.25 No.4, Apr. 2003(04.2003), u Xiaodong, Li Lemin, "the research on distributed packet scheduling method for an input and output queuing switch", pages 515-516	1-12
PX	CN101035067A (HUAWEI TECHNOLOGIES CO LTD) PD:12 Sep. 2007 (12.09.2007) AD:25 Jan. 2007 (25.01.2007) the whole document	1-12
A	CN1848803A (HUAWEI TECHNOLOGIES CO LTD) 18 Oct. 2006(18.10.2006) the whole document	1-12
A	US6195335B1 (INT BUSINESS MACHINES CORP) 27 Feb. 2001(27.02.2001) the whole document	1-12
A	US6519225B1 (NORTEL NETWORKS CORP et al.) 11 Feb. 2003(11.02.2003) the whole 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
16 Apr. 2008(16.04.2008)	**08 May 2008 (08.05.2008)**
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. 86-10-62019451	Authorized officer YANG Hongli Telephone No. (86-10)62411277

Form PCT/ISA/210 (second sheet) (April 2007)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2008/070188

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN101035067A	12.09.2007	NONE	
CN1848803A	18.10.2006	NONE	
US6195335B1	27.02.2001	NONE	
US6519225B1	11.02.2003	EP1052816A2	15.11.2000
		CA2308353A1	14.11.2000
		US6771596B1	03.08.2004

国际检索报告

国际申请号
PCT/CN2008/070188

A. 主题的分类

H04L12/56(2006.01)i

按照国际专利分类表(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:H04L,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）)

WPI, EPODOC, PAJ, CNPAT,CNKI:调度，排队，队列，统计，计算，反压，流控，输入，输出，入队，出队，schedul+, queue, orde+, rank, priority, stat+, comput+, bk, backpress+, back press+, input+, output+, ingress, egress

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	电子与信息学报，第 25 卷第 4 期，4 月 2003（04.2003），涂晓东，李乐民，“一种输入输出排队交换机中分布式分组调度方法的研究”，第 515-516 页	1-12
PX	CN101035067A （华为技术有限公司） PD:12.9 月 2007（12.09.2007）AD:25.1 月 2007（25.01.2007）全文	1-12
A	CN1848803A （华为技术有限公司） 18.10 月 2006（18.10.2006） 全文	1-12
A	US6195335B1 （INT BUSINESS MACHINES CORP） 27.2 月 2001（27.02.2001） 全文	1-12
A	US6519225B1 （NORTEL NETWORKS CORP et al.） 11.2 月 2003（11.02.2003） 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

“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
16.4 月 2008（16.04.2008）	**08.5 月 2008 (08.05.2008)**
中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路 6 号 100088 传真号： (86-10)62019451	受权官员 杨红丽 电话号码：(86-10) **62411277**

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2008/070188

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN101035067A	12.09.2007	无	
CN1848803A	18.10.2006	无	
US6195335B1	27.02.2001	无	
US6519225B1	11.02.2003	EP1052816A2	15.11.2000
		CA2308353A1	14.11.2000
		US6771596B1	03.08.2004