一种基于 Credit 的变长分组并行交换网络调度算法

杨君刚[①②]　刘增基[①]　赵瑞琴[①]　雒晓卓[①]
[①](西安电子科技大学综合业务网国家重点实验室　西安　710071)
[②](西安通信学院　西安　710106)

摘　要：该文提出了一种新的并行分组交换(PPS)网络调度算法。该算法通过在解复用器处采用以变长分组为业务分配单元的方式消除了信元的乱序问题；通过采用 Credit 机制进行业务分配，实现了业务到各个交换平面完全公平的分配；各个并行交换单元采用组合输入输出排队，降低了对缓存和交换平面的加速要求，同时可以充分利用现有单 Crossbar 网络调度算法的研究成果。文中证明了该算法对业务分配的公平性，对高速缓存的需求量以及整个网络的稳定性，仿真进一步证明了该算法具有良好性能。

关键词：并行分组交换；Credit 机制；业务分配；调度算法
中图分类号：TN915.07　　**文献标识码**：A　　**文章编号**：1009-5896(2008)09-2229-04

A Scheduling Algorithm Based on Credit in Variable-length Packet Parallel Switching Network

Yang Jun-gang[①②]　Liu Zeng-ji[①]　Zhao Rui-qin[①]　Luo Xiao-zhuo[①]
[①](*National Key Lab of Integrated Service Networks*, *Xidian University*, *Xi'an* 710071, *China*)
[②](*China Xi'an Communication Institute*, *Xi'an* 710106, *China*)

Abstract: A new scheduling algorithm in Parallel Packet Switch(PPS) is proposed in this paper . This algorithm eliminates out-of sequence of the cells belonging to a packet by dispatching the traffic with variable-length packet at de-multiplexer. It balances the traffic distribution to each parallel switch by a credit based mechanism. The combined input and output queued(CIOQ) mechanism is adopted with each parallel switch in this algorithm, which not only degrades the speed-up requirement of the buffer and switch, but also takes full advantage of the available research achievements of scheduling algorithm in Crossbar switches. Theoretical analysis is made on the fairness of traffic distributions, the amount of the required high-speed buffer (run in line-speed) and the stability of the network. The simulation analysis shows the good performance of this algorithm further.

Key words: Parallel Packet Switch(PPS); Credit mechanism; Traffic dispatch; Scheduling algorithm

1　引言

并行分组交换(Parallel Packet Switch, PPS)系统是为了克服随着交换网络端口速率的增加，带来高速接入缓存和高速交换网络不可实现的问题，达到通过利用多个并行的可实现的低速交换网络来构建高速交换网络的目的[1]。一个简单的 PPS 系统如图 1 所示，在图中 R 表示高速交换网络的端口速率(称为外部速率)，r 表示低速交换网络的端口速率(称为内部速率)，定义网络加速比 S 为：$S=kr/R$。可以看出一个 PPS 系统由输入端口处的解复用器(DMUX)(实现业务分配)，中间的多个并行交换单元(用多个低速交换单元完成高速信息的转发)和输出端口处的复用器(MUX)(实现业务汇聚)3 部分组成。PPS 系统调度算法的研究主要集中在其性能

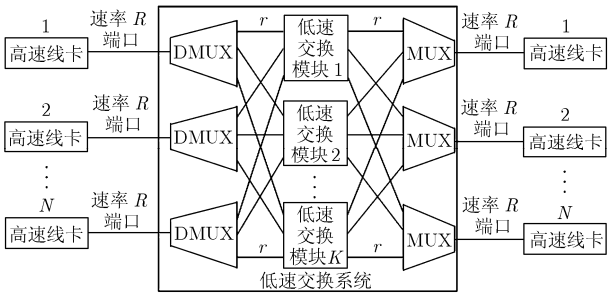

图 1 PPS 系统示意图

(稳定性)和可实现性上。最初的 PPS 系统在解复用器和复用器处不加任何缓存，所采用的调度算法是集中式调度算法，这种算法虽然可以保证网络的稳定性[1]，但是算法实现复杂，同时，需要至少 2 倍的网络加速，在实际中难以实现；在 PPS 系统解复用器和复用器处加少量缓存，可以采用分布式调度算法和取消网络加速[2]；因此，这种 PPS 系统结构成为

2007-02-13 收到，2007-09-19 改回
国家 863 计划项目(2002AA103062)，综合业务网国家重点实验室开放基金(ISN8-03)和中兴通信股份有限公司技术研究基金(ZXJS200609 120159)资助课题

了研究的重点，提出了各种不同的调度算法[2−4]，但这些算法在中间交换平面都必须采用特定的调度算法，无法很好地继承单 Crossbar 调度算法的研究成果[5−9]；同时，需要大量的开销来保持信元顺序。本文提出了一种新的并行分组交换网络调度算法，该算法具有以下特点：(1)在解复用级，采用基于变长分组的负载分配方式，使得属于同一分组的各个信元由同一个交换平面转发，从根本上消除了信元乱序问题；同时，提出了一种新型的基于 Credit 的业务分配算法，使得业务分配可以达到完全均匀；(2)中间级各个交换平面可以不加限制的对现有的单 Crossbar 调度算法进行利用，具有良好的继承性；(3)该算法是完全分布式的，网络各个模块间不需要任何的信息交换。下面对算法进行仔细描述和分析。在本文研究中系统调度是分时隙的，以速率 R 传输一个信元(一般为 64 个字节)所需要的时间为一个时隙。假设 PPS 系统是一个 $N\times N$ 的交换网络，中间并行交换单元的数目为 K 个，到达网络的业务是允许的(admissible)[8]。

2 算法描述和性能分析

由于该算法是完全分布式算法，因此按照 PPS 系统的结构，该算法可以分为 3 部分：解复用器的业务分配算法，中间级交换平面的业务调度算法和复用器的业务汇聚算法。由于中间级的调度算法可以利用现有单 Crossbar 调度算法十分丰富的研究成果[5−9]，而复用器业务汇聚算法相对比较简单，因此，解复用器处的业务分配算法是本文算法研究的重点。

2.1 解复用器处的业务分配算法

业务分配算法一方面要使属于同一分组的所有信元都从一个中间级交换平面转发；另一方面又不能因为转发分组是变长的而影响了业务到各个交换平面的均匀性，基于以上考虑，本文提出了一种基于 Credit 的业务分配算法，其基本思想是一个输入端口的每个排队利用一个均匀的 Credit 生成机制来实现业务分配的均匀性。如图 2 所示，在网络输入端口 i，按照业务到达的目的端口和转发的交换平面建立队列$(i,j,l)(1\le i,j\le N,1\le l\le K)$，因此一个解复用器处的队列总数为：$NK$ 个。解复用器的每个队列对应一个 credit 生成器，每 K 个外部时隙产生一个 Credit，逐渐积累。当一个变长分组到达输入端口 i，被分配到某个队列，那么该队列积累的 Credit 值需要减去分组的长度值(以信元为单位)。分组对队列的选择，按照 Credit 值的大小来定，每次都选 Credit 值最大的队列，这样可以通过 credit 值的调整很好地消除由于分组的变长而带来业务分配的不均匀性。下面对算法进行定量描述。

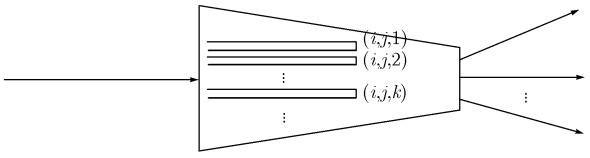

图 2 解复用器缓存组织结构示意图

用 $V_{i,j}^{(l)}(t)$ 表示在时隙 t，队列$(i,\ j,\ l)$积累的 Credit 数，$C_{i,j}^{(l)}(t)$ 表示在时隙 t 是否产生 Credit，也就是 $C_{i,j}^{(l)}(t)=\begin{cases}1, & \text{如果}t\text{是}k\text{的整数倍}\\ 0, & \text{其他}\end{cases}$，显然，$V_{i,j}^{(l)}(t)=V_{i,j}^{(l)}(t-1)+C_{i,j}^{(l)}(t)$。在时隙 t 当一个从输入端口 i 到达，目的输出端口为 j 长度为 L 个信元的分组到达，将该分组分配到队列$(i,\ j,\ m)$，其中 $m=\arg\max\limits_{1\le l\le k}V_{i,j}^{(l)}(t)$，同时，$V_{i,j}^{(m)}(t)=V_{i,j}^{(m)}(t)-L$。

用 $Q_{i,j}^{(l)}(t)$ 表示在时隙 t，队列(i,j,l)的长度。用 $A_{i,j}(t)$ 表示到时隙 t，从输入端口 i 到达，转发到输出端口 j 的分组的总信元数，令 $\lambda_{i,j}=\lim\limits_{t\to\infty}\dfrac{A_{i,j}(t)}{t}$ 表示输入端口 i 和输出端口 j 间的平均业务流量；$A_{i,j}^{(l)}(t)\ (1\le l\le K)$表示到达输入端口 i，目的输出端口为 j 的分组到时隙 t 按照以上业务分配算法从中间级交换单元 l 转发的总信元数，有以下定理 1。设到达网络的分组长度(以信元为单位)是有限的，其最大值为 $L_{\max}$。

定理 1 在任何时隙 t，$\forall k,l$，都有 $\left|A_{i,j}^{(l)}(t)-A_{i,j}^{(k)}(t)\right|\le L_{\max}$，$\left|V_{i,j}^{(l)}(t)-V_{i,j}^{(k)}(t)\right|\le L_{\max}\ (1\le l,k\le K)$。

证明 由 $V_{i,j}^{(l)}(t)=\left\lfloor\dfrac{t}{K}\right\rfloor-A_{i,j}^{(l)}(t)$，$V_{i,j}^{(k)}(t)=\left\lfloor\dfrac{t}{K}\right\rfloor-A_{i,j}^{(k)}(t)$ 可以得出：$A_{i,j}^{(l)}(t)-A_{i,j}^{(k)}(t)=-(V_{i,j}^{(l)}(t)-V_{i,j}^{(k)}(t))$，因此，只需要证明定理的前半部分，后半部分同样得证。

不失一般性，不妨假设分组长度为信元长度的整数倍，从输入端口 i 到达，转发到输出端口 j 的变长分组，在时隙 $T_1,T_2,\cdots,T_n,\cdots$, 到达。设在时隙 t，$L=\max\limits_{1\le r,p\le K}\left|A_{i,j}^{(r)}(t)-A_{i,j}^{(p)}(t)\right|$，$(l,k)=\arg\max\limits_{1\le r,p\le K}\left|A_{i,j}^{(r)}(t)-A_{i,j}^{(p)}(t)\right|$，不妨假设 $A_{i,j}^{(l)}(t)>A_{i,j}^{(k)}(t)$。由于 $A_{i,j}^{(l)}(t)(1\le l\le K)$ 只可能在 $T_n(n=1,2,\cdots)$ 发生改变，因此，当 $t\in[T_n,T_{n+1})$ 时，$A_{i,j}^{(l)}(t)=A_{i,j}^{(l)}(T_n)$ $(1\le l\le K)$，有 $L=A_{i,j}^{(l)}(T_n)-A_{i,j}^{(k)}(T_n)$。令 $f(T_n)=A_{i,j}^{(l)}(T_n)-A_{i,j}^{(k)}(T_n)$，因为 $A_{i,j}^{(l)}(0)=A_{i,j}^{(k)}(0)=0$，因此，$f(0)=0$。由于 $A_{i,j}^{(l)}(T_n)>0$，而只有在 $f(T_m)\le 0\ (0\le m<n)$时，在 T_m 时隙 $A_{i,j}^{(l)}(T_m)$ 才可能增加，所以一定可以找到一个 m_1，满足 $f(T_{m_1})\le 0$，而 $f(T_{m_2})>0(m_2=m_1+1,\cdots,n)$。由于 $f(T_{m_1})\le 0$，可得 $A_{i,j}^{(l)}(T_{m_1})\le A_{i,j}^{(k)}(T_{m_1})$，设在 T_{m1} 到达的分组长度为 L_{m1}，该分组由第 l 个交换平面转发，因此，有 $A_{i,j}^{(l)}(T_{m_1+1})=A_{i,j}^{(l)}(T_{m_1})+L_{m_1}$，可得 $f(T_{m_1+1})\le L_{m_1}$。由于 $f(T_{m_2})>0(m_2=m_1+1,\cdots,n)$，所以从 T_{m_1+2} 到 T_n，$A_{i,j}^{(l)}(T)$ 不可能再增加，而 $A_{i,j}^{(k)}(T)$ 不会减少，因此有 $L=f(T_n)\le L_{m_1}\le L_{\max}$。 证毕

定理 2 在输入端口 i 处，每个队列$(i,\ j,\ l)(1\le j\le N,1\le l\le K)$的长度都不会大于 $L_{\max}$。

证明 在时隙 t，令 $x=\arg\max\limits_{1\le l\le K}A_{i,j}^{(l)}(t)$，由定理 1 可以知道，$\forall l$ 都有 $A_{i,j}^{(l)}(t)\ge A_{i,j}^{(x)}(t)-L_{\max}$。因此有 $A_{i,j}(t)=\sum\limits_{l=1}^{K}A_{i,j}^{(l)}(t)\ge K(A_{i,j}^{(x)}(t)-L_{\max})$；又因为网络业务的到达是允

许的[8]，所以有 $A_{i,j}(t)<t$，因此有：$A_{i,j}^{(x)}(t)<\frac{t}{K}+L_{\max}$。由漏桶模型的定义可以得出定理 2 成立。　　证毕

定理 3　采用以上的输入端业务分配算法可以实现完全均匀的将输入业务分配到各个交换平面。

证明　由 $\lambda_{i,j}=\lim_{t\to\infty}\frac{A_{i,j}(t)}{t}$。因为 $A_{i,j}(t)=\sum_{l=1}^{K}A_{i,j}^{(l)}(t)$，同时结合定理 1 可以知道：$\frac{1}{K}A_{i,j}(t)-L_{\max}\le A_{i,j}^{(l)}(t)\le\frac{1}{K}A_{i,j}(t)+L_{\max}$，由极限的夹逼准则可得：$\lambda_{i,j}^{(l)}=\lim_{t\to\infty}\frac{A_{i,j}^{(l)}(t)}{t}=\frac{1}{K}\lambda_{i,j}$ $(1\le l\le K)$，很显然定理 3 成立。　　证毕

定理 4　在上述的业务分配算法中，对 $V_{i,j}^{(l)}(t)\,(1\le l\le K)$只需要用$[-L_{\max},L_{\max}]$间的整数就可以完全表示。

证明　因为在上述的业务分配算法中，只需要知道各个 $V_{i,j}^{(l)}(t)\,(1\le l\le K)$之间的大小关系，选出 $V_{i,j}^{(l)}(t)$ 最大的队列将分组放入即可。由定理 1 可以知道，在任何时隙 t，$\forall k,l$ 都有 $\left|V_{i,j}^{(l)}(t)-V_{i,j}^{(k)}(t)\right|\le L_{\max}$ $(1\le l,k\le K)$，因此，当所有 $V_{i,j}^{(l)}(t)\,(1\le l\le K)$都大于零时，只需要全部同时减去最小的 $V_{i,j}^{(l_1)}(t)$，$V_{i,j}^{(l)}(t)\,(1\le l\le K)$的范围会在$[0,L_{\max}]$；当所有 $V_{i,j}^{(l)}(t)\,(1\le l\le K)$都小于零时，只需要全部同时加上最大的 $V_{i,j}^{(l_2)}(t)$，$V_{i,j}^{(l)}(t)\,(1\le l\le K)$的范围就在[$-L_{\max}$，0]。很显然利用[$-L_{\max}$，$L_{\max}$]间的整数来表示 $V_{i,j}^{(l)}(t)$ 就已经足够了。　　证毕

从定理 2 可以知道在本算法的解复用器处，需要的缓存(以线速读写)大小为 $NK\times L_{\max}$，在现有的芯片技术下，可以在片内实现，同时，其时延也是有界的，排队稳定。从定理 4 可以知道，在本算法中 $V_{i,j}^{(l)}(t)$ 需要的开销和计算复杂度都非常的小，很利于实现。

2.2 并行交换平面调度算法

按照本算法的初衷，并行交换平面要具用良好的继承性，可以直接应用现有单 Crossbar 交换网络调度算法丰富的研究成果[5−9]，不需要因为并行交换的应用背景而进行任何的改变。在采用输入(组合输入输出)排队的单 Crossbar 交换网络中，有定长信元和变长分组两种交换机制[5]。虽然在本算法中，解复用器是基于变长分组的，但是并不意味着中间交换平面也必须采用变长分组交换机制。这是因为解复用器采用变长分组分配方式是为了让属于同一个分组的各个信元通过同一个中间级交换平面转发，以防止信元乱序问题，而在中间级交换平面不会存在信元乱序问题，因此不需要对交换机制进行任何限制。从单 Crossbar 调度的研究成果可以看出[5]，变长分组交换方式虽然可以减少网络开销，去掉分组分割和重组模块，但是变长分组交换方式的网络适应性远远比不上定长信元方式，网络的性能和到达网络业务分组长度的分布密切相关，使得网络的性能不稳定；同时，基于变长分组交换的调度算法和理论基础也远远没有定长信元方式的丰富，因此，在本文算法中中间交换平面采用定长信元交换方式。在本文算法中，变长分组的分割和重组在中间交换平面处。

由定理 2 可以得出，在解复用器处分组的排队长度是有上界的，因此，是稳定的，而在复用器处同样可以证明所需的排队缓存是有上界的，因此，整个 PPS 系统的性能(稳定性和时延)和并行交换单元的性能相关的。由于并行交换单元可以继承单 Crossbar 交换单元调度算法，从单 Crossbar 网络调度算法的研究成果可以得出，只要网络的到达业务是允许的在定长信元交换方式下，有很多种调度算法可以实现网络稳定和良好的时延特性[6−8]。因此，只需要证明在本算法中到达中间各个交换平面的业务是允许的，有以下定理 5。

定理 5　当到达 PPS 系统的业务是允许的，采用上述解复用器的业务分配算法，到达各个并行交换平面的业务也是允许的。

证明　不失一般性，不妨选中间级交换平面 l 进行讨论。从网络业务允许的条件[8]可以知道：

$$\sum_{j=1}^{N}\lambda_{i,j}\le 1;\ \sum_{i=1}^{N}\lambda_{i,j}\le 1 \tag{1}$$

因为到达交换平面 l 输入端口 i，目的端口为 j 的业务的平均流量为 $\lambda_{i,j}^{(l)}$，由定理 3 可以知道：$\lambda_{i,j}^{(l)}=\frac{1}{K}\lambda_{i,j}$，又因为在网络没有加速的情况下，一个并行交换平面的交换时隙为 PPS 系统交换时隙的 K 倍，因此，在交换平面 l 的一个交换时隙内(内部时隙)$\hat{\lambda}_{i,j}^{(l)}=K\lambda_{i,j}^{(l)}=\lambda_{i,j}$，由式(1)可得：$\sum_{j=1}^{N}\hat{\lambda}_{i,j}^{(l)}\le 1$；$\sum_{i=1}^{N}\hat{\lambda}^{(l)}{}_{i,j}\le 1$。　　证毕

2.3 复用器业务汇聚算法

由于在本算法中，信元不会乱序，而在并行交换平面已经完成分组的重组，因此，从理论上来说，在复用器处不需要缓存，直接对从各个并行交换平面来的分组进行复用就可以了，但是，这样会出现不同分组之间的交织，使得从 PPS 系统出去的是多个相互交织的分组，这样显然是不合理的。所以在本算法中，在复用器处加一定的缓存对分组进行一定的排队，使得分组可以以完整形式从输出端口输出，各个分组间不会发生交织。一个复用器对每一个并行交换平面设立一个分组排队缓存，缓存大小是$(1-1/K)L_{\max}$，一个输出端口的总缓存数为 $N(1-1/K)L_{\max}$，远小于输入端口的缓存数。

3　算法性能仿真分析

为了进一步说明算法的性能，本文在 OPNET 仿真平台下，建立了算法的性能仿真模型，在仿真中，中间级各个并行交换平面采用 islip 调度算法，叠代次数为 4 次[6]；同时，为便于对比，建立了和 PPS 系统具有相同端口数的单 Crossbar 交换单元的 iSLiP 调度算法性能仿真模型，叠代次数也为 4 次。为了提高仿真的可信度，采用文献[9]的仿真业务模型，考虑了网络中不同的业务情况，将网络业务分为高

度均匀、低度均匀和低度不均匀 3 种，按照到达业务的特性又分为非突发业务和突发业务两种，因此，网络业务分为非突发高度均匀、低度均匀、低度不均匀业务和突发高度均匀、低度均匀、低度不均匀 6 种业务类型进行分析。以 ON/OFF 源来模拟分组长度的分布情况，一个 ON 期代表一个分组的长度，各个信元的目的地址相同，不同 ON 期分组的目的地址按照以上高度均匀、低度均匀和低度不均匀业务模型分布。ON/OFF 期的长度分别服从参数为 p 和 q 的几何分布，根据业务流负载 ρ 和平均突发长度 E_{ON}，有 $p=1/E_{\mathrm{on}}$, $q=\dfrac{\rho p}{1-\rho p}$。$E_{\mathrm{ON}}$ 的概率分布如表 1 所示(在仿真中一个信元长度定义为 64 个字节)：

表 1 E_{on} 的概率分布表

E_{ON}(信元)	1	10	24
概率	0.6	0.3	0.1

在图 3 和图 4 中用 Cre 表示本文提出的 PPS 系统的调度算法，Cross 表示单 Crossbar 中的调度算法。Uni、lowban 和 lowunban 分别表示高度均匀、低度均匀和低度不均匀业务。从图 3 和图 4 可以看出无论是突发业务还是非突发业务，采用本文算法的 PPS 系统的时延和采用 islip 的单 Crossbar 系统的时延变化趋势是非常类似的，相对于单 Crossbar 系统只是增加了一个固定的值。这是由于在输入端口的业务分配算法具有一定的排队时延(其时延上界为 $L_{\max}$)，并且中间交换平面的交换速率低于单 Crossbar 的交换速率所致，这是并行交换网络为了达到简单实现所必须付出的代价。

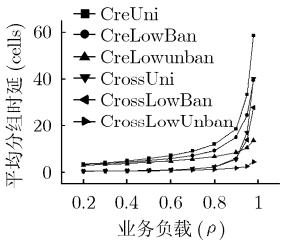

图 3 突发业务下 credit 算法和 islip 算法性能对比图

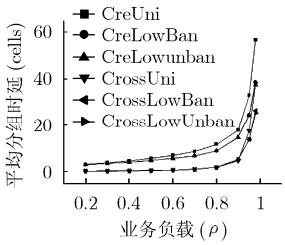

图 4 非突发业务下 credit 算法和 islip 算法性能对比图

4 结束语

本文针对现有 PPS 系统调度算法的不足，提出了一种新型的基于 Credit 的 PPS 系统调度算法。该算法通过以变长分组为负载分配单位，消除了传统 PPS 系统的信元乱序问题；通过 Credit 机制保证了业务分配的完全公平性；该算法中间交换平面可以直接继承现有单 Crossbar 调度算法的大量研究成果。文中通过严格的理论证明了算法的稳定性，通过仿真证明了算法的良好性能。 因此，该算法是一种很适合 PPS 系统应用的调度算法。

参 考 文 献

[1] Iyer S, Awadallah A, and McKeown N. Analysis of a packet switch with memories running slower than the line rate[C]. IEEE INFOCOM '00, Tel-Aviv, Israel, 2000: 529–537.

[2] Iyer S and McKeown N. Making Parallel Packet Switches Practical[C]. IEEE INFOCOM'01, Alaska, USA, 2001: 1680–1687.

[3] Zhong Hakhan, Xu Du, and Zhu Zhenyu. A parallel packet switch supporting variable-length packets[C]. International Conference on Communications, Circuits and Systems Proceedings. Hong Kong, China, 2005: 613–617.

[4] Shi Lei and Li Wenjie, *et al.*. Flow mapping in the load balancing parallel packet switches[C]. Workshop on High Performance Switching and Routing. Hong Kong, China, 2005: 254–258.

[5] Ganjali Y, Keshavarzian A, and Shah D. Input queued switches: cell switching vs. packet switching[C]. IEEE INFOCOM'03. San Francisco, USA, 2003: 1651–1658.

[6] McKeown N. iSLIP: A scheduling algorithm for input-queued switches[J]. *IEEE/ACM Trans. on Networking*, 1999, 7(2): 188–201.

[7] Prabhakar B and McKeown N. On the speedup required for combined input and output queued switching[J]. *Automatica*, 1999, 35(12): 1909–1920.

[8] Dai J G and Prabhakar B. The throughput of data switches with and without speedup[C]. IEEE INFOCOM'00, TelAviv, Israel, 2000: 556–564.

[9] Goudreau M W, *et al.*. Scheduling algorithms for input-queued switches: Randomized techniques and experimental evaluation [C]. IEEE INFOCOM'00, TelAviv, Israel, 2000: 1624–1643.

杨君刚： 男，1973 年生，博士生，讲师，研究方向为大容量路由器交换网络、下一代光网络关键技术.

刘增基： 男，1937 年生，教授，博士生导师，研究方向为宽带网络关键技术.

赵瑞琴： 女，1981 年生，博士生，研究方向为无线传感器网络、无线移动自组织网.

雒晓卓： 男，1982 年生，硕士生，研究方向为光传输设备交换网络调度和保护技术.