Performance Evaluation of the MSMPS Algorithm under Different Distribution Traffic

Grzegorz Danilewicz and Marcin Dziuba

Faculty of Electronics and Telecommunications, Poznan University of Technology, Poznan, Poland

Abstract—In this paper, the Maximal Size Matching with Permanent Selection (MSMPS) scheduling algorithm and its performance evaluation, under different traffic models, are described. In this article, computer simulation results under nonuniformly, diagonally and lin-diagonally distributed traffic models are presented. The simulations was performed for different switch sizes: 4×4, 8×8 and 16×16. Results for MSMPS algorithm and for other algorithms well known in the literature are discussed. All results are presented for 16×16 switch size but simulation results are representative for other switch sizes. Mean Time Delay and efficiency were compared and considered. It is shown that our algorithm achieve similar performance results like another algorithms, but it does not need any additional calculations. This information causes that MSMPS algorithm can be easily implemented in hardware.

Keywords—connection pattern, diagonally distributed traffic, lin-diagonally distributed traffic, MQL matrix, non-uniformly distributed traffic, switching fabric.

1. Introduction

Several well known scheduling algorithms have been proposed in the literature [1]–[6]. All these algorithms, which are responsible for configuration of a switching fabric, are very sophisticated and they achieve a good efficiency and short time delay. During designing of a new algorithm, a theoretical approach is applied. It means that designers do not pay attention to algorithm implementation constraints. Most of well known algorithms, which achieve the good performance results, are very difficult for implementation in the real switching fabric hardware. This is due to very complicated calculations which must be performed during algorithms work. The high calculations complexity makes this algorithms impractical. Instead, most of the new generation switches and routers use much simpler scheduling algorithms to control and configure switching fabric. One of this kind of algorithms is MSMPS [7], which achieve the similar performance results like the rest of algorithms but does not need to perform a lot of complicated calculation.

Other important fact, which influence on switches and routers performance, is switching fabric buffers architecture. In our research we study a switching fabric with VOQ (Virtual Output Queue) system [6], [8]. This buffering system has been proposed to solve a HOL (Head of Line) effect. In VOQ system each switching fabric input has a separate queue for a packet directed to particular output of a switching fabric. Using this kind of architecture, its performance depends only on a good scheduling algorithm. Algorithm should be very fast, achieve the good results (high efficiency and short time delay) and be easy to implement in hardware.

Before each packet will be send through the switch, it should be decided which packet, from which VOQ will be chosen. This decision is taken in each time slot – the basic unit of time in simulation environment. To solve this problem in hardware, a few scheduling mechanisms are used. There are three basic methods: random selection, first in first out (FIFO) and round-robin. In the presented architecture centralized scheduling mechanism is used. In this mechanism all decisions considering setting up connections between switching fabric inputs and outputs (connection patterns) are made by algorithm or driver implemented in a separated control module. Driver can control some connected switching fabrics located in different equipments (i.e., routers). Such solution can be used in the new generation networks for example in Software Defined Networks (SDN) [9]. Routers are responsible for direct packets in data paths but high level decisions (routing) are moved to separate module or device which is located out of routers. Routing decisions are sent to routers to execute suitable connection patterns in each switching fabric of each router. Centralized scheduling mechanism has a huge advantage over traditional scheduling mechanism. In todays network nodes, where 10 Gbit/s ports are used, each time slot is equivalent to the 50 ns. This time in not enough to realize traditional scheduling mechanism, which based on sending control signal. This signal consists of three parts: demand, confirmation and acceptance. Nowadays, all algorithms are designed in such a way, that the number of control signals is minimized. The best solution is sending only one signal between control module and switching fabric. All this things are fulfilled by MSMPS algorithm.

This paper is organized as follows. In Section 2, the switch architecture is discussed. In Section 3 of this article, all simulation parameters are explained. In Section 4 traffic distribution models which are used in our research, are described. Then in Section 5 computer simulation results under different traffic patterns are shown. Results achieved for MSMPS algorithm, are compared with another algorithms well known in the literature. In Section 6, same conclusion are given.

2. Switch Architecture

The general VOQ switch architecture is presented in Fig. 1 [10].

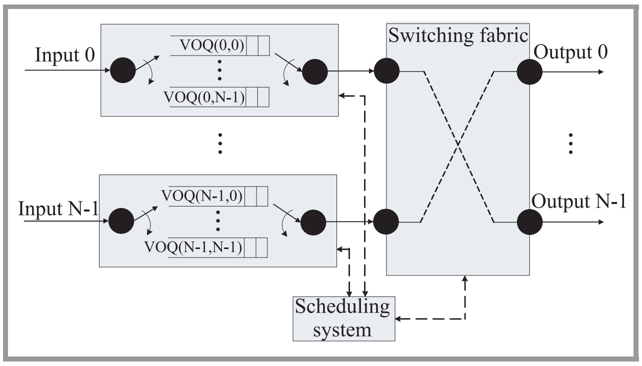

Fig. 1. General VOQ switch architecture.

In our research we use switching fabric with input queuing system (Input Queued switches), where buffers are placed at the inputs. Each input has separated queue which is divided into N independent VOQs. The total number of virtual queues depends on the number of inputs and outputs. It was assumed that in presented switch, the number of inputs and outputs is equal and in general case is N. Based on this assumption, total number of VOQs in switching fabric, with N number of inputs/outputs, is equal to N^2. Additionally, each virtual queue is denoted by VOQ (i,j), where i is the input port number and j is the output port number. It can be assumed that: $0 \le i \le N-1$ and $0 \le j \le N-1$.
Between inputs and outputs modules, the switching fabric is placed. In the switching fabric, there are electrical or optical equipments which have to be properly configured when all connections between inputs and outputs are established. Implemented algorithm is responsible for a proper configuration of mentioned equipments.
The most important module, in presented symmetrical switch, is scheduling system module. This is a module, where algorithms are implemented. In the scheduling module all information about queues conditions are stored. It means that scheduling system has knowledge about numbers of packets waiting in all queues, to be send through the switch. This information is necessary to make a right decision by MSMPS algorithm about connection pattern in the switching fabric.

3. Algorithm Description

MSMPS algorithm is based on permanent connections pattern between inputs and outputs. For example, from Fig. 2, connection pattern for 4×4 switch can be observed.

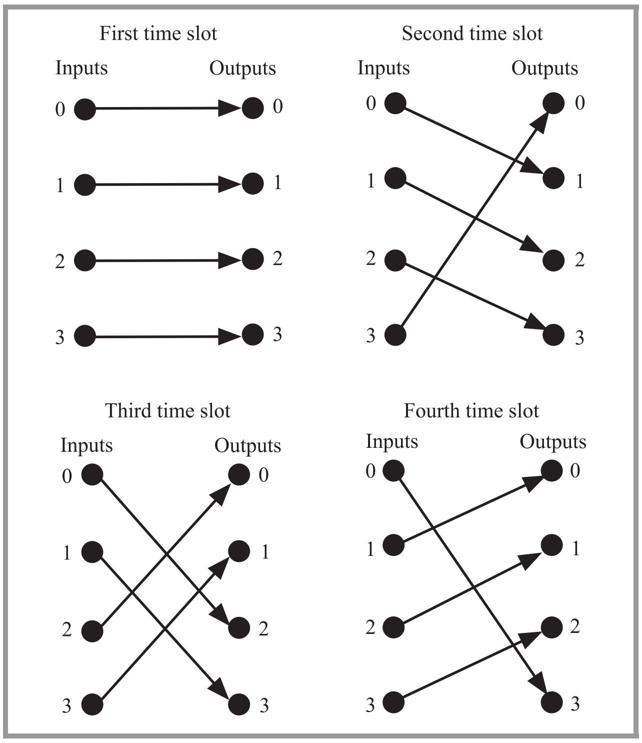

Fig. 2. Connection pattern for 4×4 switch.

Permanent connections pattern provides fair access to the each output. It means that all outputs in switch are treated equally. As mentioned before, scheduling module has information about VOQ conditions. This information is stored in MQL matrix (Matrix of Queue Lengths). This kind of matrix was the easiest way to store this information. Figure 3 shows MQL matrix for 4×4 switch.

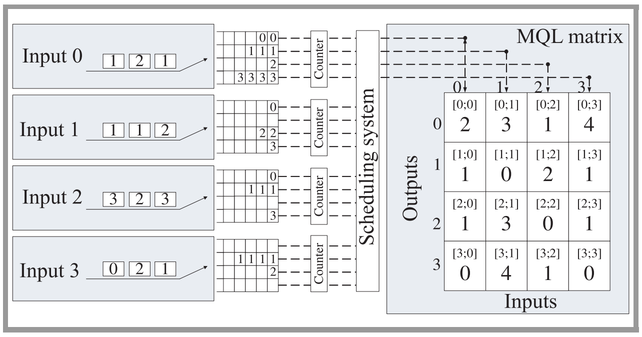

Fig. 3. MQL matrix for 4×4 switch.

Information is updated in each time slot. Each cell (one position in matrix) in matrix MQL and each VOQ has unique address. This correlation allows attribute one cell to one VOQ. For example cell [0;0] corresponds to the VOQ (0,0). In cell [0;0] information about number of packets waiting in VOQ (0,0) are stored. If there is no packets in VOQ, suitable position in matrix is filled by 0. It can be seen from Fig. 3 can be observed that matrix has N rows and N columns. It corresponds to the 4×4 switch, which is presented in our example. Based on permanent connections and information, stored in MQL matrix, MSMPS algorithm makes decisions about connections to be set up in switching fabric. The main purpose is to avoid empty connections. Empty connection means that there is no packets to be send from an input to an output. Algorithm gives priority to the most filled VOQs. More details about MSMPS algorithm can be found in [7].

4. Simulation Conditions

In this paper, performance results for some scheduling algorithms, well known in the literature, and for MSMPS algorithm are presented. All graphs are plotted as the results of computer simulations. Packets are incoming at the inputs according to Bernoulli arrival model [11], [12]. Under this model, only one packet can arrive at the input in each time slot (basic of time unit). It was assumed that one packet may occupy only one time slot. In Bernoulli model, probability that packet will arrive at the input is equal to p, where:

$$p\,\varepsilon(0 < p \leq 1)\,. \tag{1}$$

Simulation results are presented as a mean value of ten independent simulation runs. Number of iteration in one simulation run is equal to 500,000, where the first 30,000 steps are reserved for obtaining convergence in the simulation environment. It was assumed also that our switching fabric is strict sense nonblocking. It means that there is always possible to establish connection between each suitable and idle input and suitable and idle output of the switching fabric. Performance results consider the efficiency and Mean Time Delay parameters.

Efficiency is parameter which was calculated according to Eq. (2). Numerator is the number of packets passed in n-th time slots through the switching fabric. Denominator is the number of packets which have arrived to the switch buffers in n-th time slot [7].

$$q = \frac{\sum\limits_n a_n}{\sum\limits_n b_n}\,, \tag{2}$$

where:

n – time slot number,

a_n – number of packets passed in n time slot through the switching fabric,

b_n – number of packets which can be send through the switching fabric in n time slot.

Mean Time Delay (MTD) is calculated according to Eq. (3). Numerator is a sum of difference between time when a packet is transferred by the switch and the time when the packet has arrived to the buffer system. Denominator is a total number of packets served by the switching fabric.

$$MTD = \frac{\sum\limits_n t_{out} - t_{in}}{\sum\limits_n k_n}\,, \tag{3}$$

where:

MTD – Mean Time Delay,

n – time slots number,

t_{in} – time when a packet arrived to the VOQ,

t_{out} – time when the same packet is transferred by the switching fabric,

k – number of packets.

Three distributed traffic models were taken into account in this paper. Each of this model determines the probability that packet which appear at the input, will be directed to the certain output. These considered traffic models are described in following subsections.

4.1. Non-uniformly Distributed Traffic

The probability of the packet arriving at the input i, directed to the output j is presented in Table 1. For readability, table shows traffic distribution in 4 × 4 switch. Analogous traffic distribution is used for other switch sizes: 8 × 8 and 16 × 16. It can be observed from Table 1 that in this type of traffic model, some outputs have higher probability of being selected [13]. This probability can be defined as: p_{ij} and it can be calculated according to the Eq. 4:

$$p_{ij} \begin{cases} \dfrac{1}{2} & \text{for } i = j, \\ \dfrac{1}{2(N-1)} & \text{for } i \neq j. \end{cases}, \tag{4}$$

where:

N – number of switch inputs/outputs.

Table 1
Non-uniformly distributed traffic in 4 × 4 switch with VOQ

	Output 0	Output 1	Output 2	Output 2
Input 0	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
Input 1	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{6}$
Input 2	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{6}$
Input 3	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$

4.2. Diagonally Distributed Traffic

In this type of distribution model, the traffic is concentrated in two diagonals of the table (traffic matrix). The probability that packet is appeared at the suitable input i and it will be directed to the output j is equal to $p_{ij} = \frac{1}{2}$. Probability for the rest of inputs (not placed in two diagonals) is p_{ij} = 0 [12], [14]–[16]. From Table 2 it can be observed that input i has packets only for output i and for output ((i + (N-1)) mod N).

Table 2
Diagonally distributed traffic in 4 × 4 switch with VOQ

	Output 0	Output 1	Output 2	Output 2
Input 0	$\frac{1}{2}$	0	0	$\frac{1}{2}$
Input 1	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Input 2	0	$\frac{1}{2}$	$\frac{1}{2}$	0
Input 3	0	0	$\frac{1}{2}$	$\frac{1}{2}$

4.3. Lin-diagonally Distributed Traffic

Lin-diagonally distributed model is a modification of diagonally distributed model. Considered lin-diagonally model and its probabilities are presented in Table 3. It can be seen from this table that a load decrease linearly from one diagonal to the other. In general case, probability can be calculated according to the following formula [17]:

$$p_d = p\frac{N-d}{N(N+1)/2} \tag{5}$$

with $d = 0, \ldots, N-1$, then $p_{ij} = p_d$ if $j = (i+d) \mod N$, and where:

p_d – probability of packet arriving in lin-diagonally distributed traffic,

p – probability of packet arriving in Bernoulli, process,

N – number of switch inputs/outputs,

d – output number.

Table 3
Lin-diagonally distributed traffic in 4 × 4 switch with VOQ

	Output 0	Output 1	Output 2	Output 2
Input 0	$\frac{4}{10}$p	$\frac{1}{10}$p	$\frac{2}{10}$p	$\frac{1}{10}$p
Input 1	$\frac{3}{10}$p	$\frac{4}{10}$p	$\frac{1}{10}$p	$\frac{2}{10}$p
Input 2	$\frac{2}{10}$p	$\frac{3}{10}$p	$\frac{4}{10}$p	$\frac{1}{10}$p
Input 3	$\frac{3}{10}$p	$\frac{2}{10}$p	$\frac{3}{10}$p	$\frac{4}{10}$p

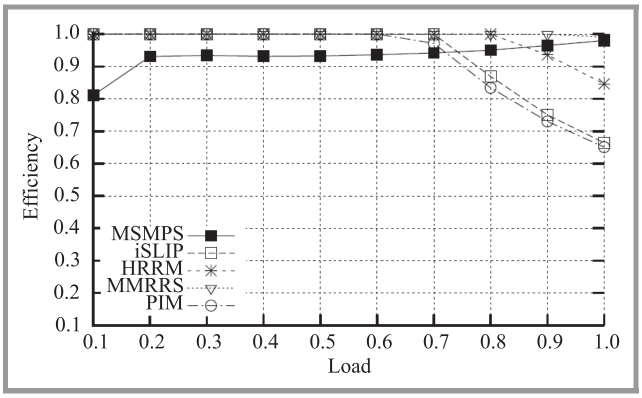

Fig. 4. The efficiency for Bernoulli arrivals with nonuniformly distributed traffic in 16 × 16 switches.

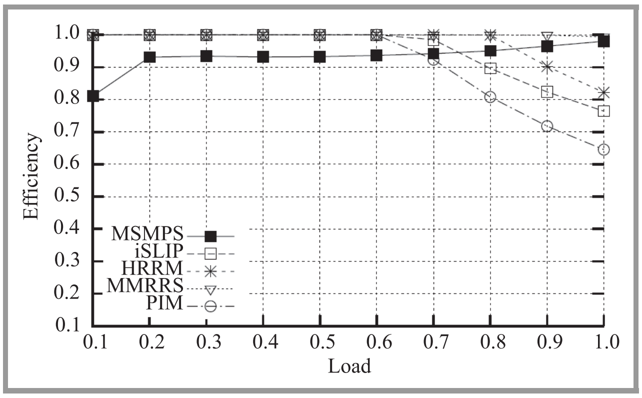

Fig. 5. The efficiency for Bernoulli arrivals with lin-diagonally distributed traffic in 16 × 16 switches.

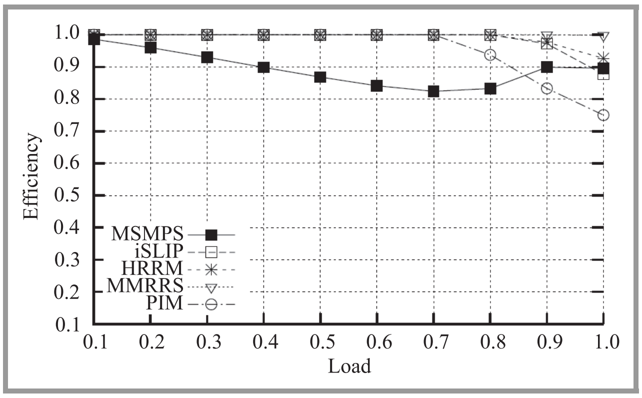

Fig. 6. The efficiency for Bernoulli arrivals with diagonally distributed traffic in 16 × 16 switches.

5. Simulation Results Analysis

In this section performance of the MSMPS algorithm will be compared with another algorithms for VOQ switches. Up today, several scheduling algorithms are presented in the literature [1]–[6]. It was compared and analyzed results for: iSLIP which was presented in [1], Maximal Matching with Round-Robin Selection (MMRRS) [2], [3], [4], Hierarchical Round-Robin Matching (HRRM) [5] and Parallel Iterative Matching (PIM) [6].

The efficiency is plotted in Figs. 4, 5 and 6. This parameter was calculated according to Eq. 2. Similarly as for MTD, results only for 16 × 16 switch size are presented. From Figs. 4 and 5 it can be observed that for low traffic load (between 10 – 20%) our algorithm achieve the worst results compared to other algorithms. Conducted simulations confirm, that MSMPS algorithm can not cope with low traffic load for different traffic models. The reason is that our algorithm focused very much on access alignment for all outputs, instead of avoiding of empty connections. Connections where no packets are to be send through the switch [7]. Above 20% load, efficiency of MSMPS algorithm increases and reaches mean value about 0.95 with growing tendency. Different phenomena can be observed for other algorithms. All of them maintain efficiency on a high level about 1. But above 60% load, PIM and iSLIP rapidly decreases with nonuniformly and lin-diagonally distributed traffic. Only MMRRS maintain efficiency about 1 for both mentioned traffic distributions. It looks different with diagonally distributed traffic. Efficiency for MSMPS algorithm systematically decreases for over 40% load, efficiency is under 0.9. This type of distribution caused that packets are concentrated in two diagonals of the traffic matrix (Table 2). For this traffic model our algorithm achieve the worst results.

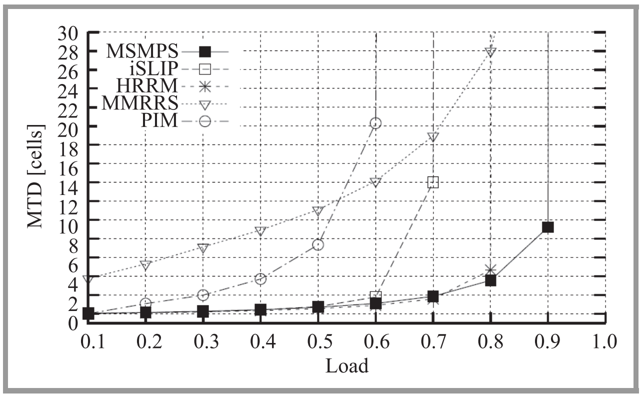

Fig. 7. The MTD for Bernoulli arrivals with nonuniformly distributed traffic in 16 × 16 switches.

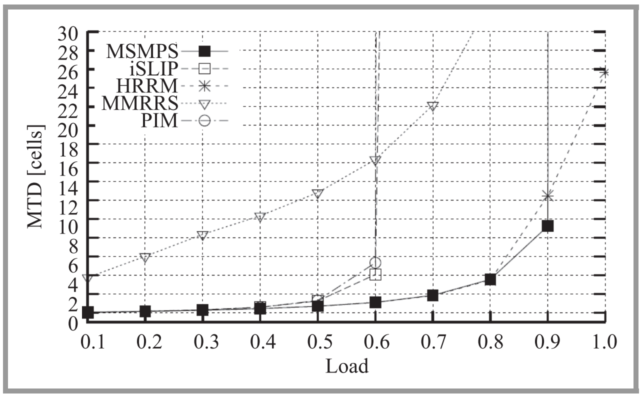

Fig. 8. The MTD for Bernoulli arrivals with lin-diagonally distributed traffic in 16 × 16 switches.

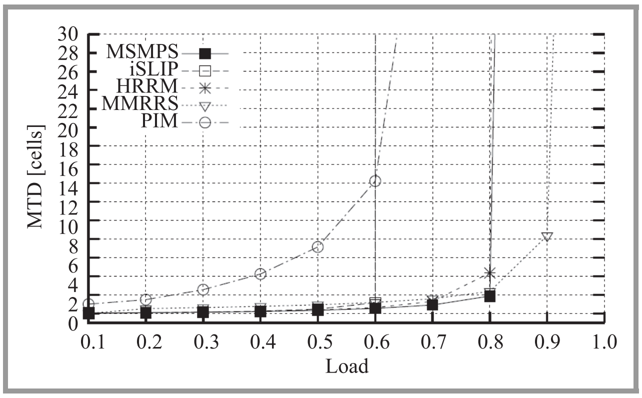

Fig. 9. The MTD for Bernoulli arrivals with diagonally distributed traffic in 16 × 16 switches

The MTD is a function of traffic load and is plotted in Figs. 7, 8 and 9. MTD is measured in time slots, where one slot is the basic of time unit in presented system. Computer simulations were performed for different switch sizes. Only the results for 16×16 switch size are shown. The authors assume that the input buffers are infinitely long, and have presented results for Bernoulli arrivals with different distribution traffic. From Fig. 7 it can be seen that for nonuniformly distributed traffic MSMPS algorithm achieve the best results (the lowest MTD) compared to other algorithms. Up to 75% load, only HRRM algorithm achieve similar results. The highest MTD, for this type of distribution, has reached MMRRS algorithm. For 10% load, MMRRS algorithm has already achieved 4 cells delay, when the rest of algorithms reached results close to 0. Very similar results are achieved by all algorithms with lin-diagonally distribution traffic – Fig. 8. MSMPS algorithm achieve almost the same results like for nonuniformly distribution. The same situation can be observed with MMRRS algorithm. Interesting situation occurred above 60% load, when MTD for PIM and iSLIP algorithm rapidly increase. It can be caused by arbiters synchronization problem. From the Fig. 9, with results for diagonal distribution traffic, it can be seen that MTD for our algorithm rapidly increased. This is due to our algorithm based on permanent connection patterns and for high load some outputs are blocked. According to this fact, to much empty connections are established. This effect can be eliminated by set up connections (between inputs and outputs) for more than one time slot. Acceptable results are reached by MMRRS algorithm which behave extremely well for diagonal distribution traffic.

6. Conclusions and Future Work

In this paper, performance results for MSMPS scheduling algorithm for VOQ switches under different traffic patterns were shown and described. Its performance confirms that MSMPS algorithm can be used in practice. This algorithm achieved high efficiency and in the same time low latency is provided. In the next studies, implementation of MSMPS algorithm in separate chips or in the switching fabric equipment will be discussed. Our algorithm works in simply way and there is no additional calculation needed. MSMPS algorithm can be also modified to support different traffic priorities and switch architectures.

Acknowledgements

The work described in this paper was financed from the research funding as research grant UMO-2011/01/B/ST7/03959.

References

[1] N. McKeown, "The iSLIP scheduling algorithm for input-queued switches", *IEEE/ACM Trans. Netw.*, vol. 7, pp. 188–200, 1999.

[2] A. Baranowska and W. Kabaciński, "The new packet scheduling algorithms for VOQ switches", in *Telecommunications and Networking – ICT 2004*, J. Neuman de Souza, P. Dini, and P. Lorenz, Eds. LNCS 3124, pp. 711–716. Springer, 2004.

[3] A. Baranowska and W. Kabaciński, "MMRS and MMRRS packet scheduling algorithms for VOQ switches", in *Proc. MMB & PGTS 2004 – 12th GI/ITG Conf. Measur. Eval. Comp. Commun. Sys. (MMB) & 3rd Polish-German Teletr. Symp. (PGTS)*, Dresden, Germany, 2004.

[4] A. Baranowska and W. Kabaciński, "Evaluation of MMRS and MMRRS packet scheduling algorithms for VOQ switches under bursty packet arrivals", in *Proc. Worksh. High Perfor. Switch. Rout. HPSR 2005*, Hong Kong, China, 2005, pp. 327–331.

[5] A. Baranowska and W. Kabaciński, "Hierarchical round-robin matching for virtual output queuing switches", in *Proc. Adv. Industr. Conf. Telecommun. AICT 2005*, Lisbon, Portugal, 2005, pp. 196–201.

[6] T. Anderson *et al.*, "High-speed switch scheduling for local-area networks", *ACM Trans. Comp. Sys.*, vol. 11, no. 4, pp. 319–352, 1993.

[7] G. Danilewicz and M. Dziuba, "The new MSMPS packet scheduling algorithm for VOQ switches", in *Proc. 8th IEEE, IET Int. Symp. Commun. Sys. Netw. Digit. Sig. Process. CSNDSP 2012*, Poznań, Poland, 2012.

[8] Y. Tamir and G. Frazier, "High performance multiqueue buffers for VLSI communication switches", in *Proc. 15th Ann. Int. Symp. Comp. architec. ISCA 1988*, Honolulu, Hawaii, USA, 1988, pp. 343–354.

[9] Myung-Ki Shin, Ki-Hyuk Nam, and Hyoung-Jun Kim, *Software-defined networking (SDN): A reference architecture and open APIs*, in *Proc. Int. Conf. ICT Converg. ICTC 2012*, Jeju, Korea, 2012.

[10] A. Baranowska and W. Kabaciński, "Hierarchiczny algorytm planowania przesyłania pakietów dla przełącznika z VOQ", in *Poznańskie Warsztaty Telekomunikacyjne PWT 2004*, Poznań, Poland, 2004 (in Polish).

[11] H. Jonathan Chao and B. Liu, *High Performance Switches and Routers*. New Jersey: Wiley, 2007, pp. 195–197.

[12] P. Giaccone, D. Shah, and S. Prabhakar, "An implementable parallel scheduler for input-queued switches", *IEEE Micro*, vol. 22, no. 1, pp. 19–25, 2002.

[13] K. Yoshigoe and K. J. Christensen, "An evolution to crossbar switches with virtual ouptut queuing and buffered cross points", *IEEE Network*, vol. 17, no. 5, pp. 48–56, 2003.

[14] D. Shah, P. Giaccone, and B. Prabhakar, "Efficent randomized algorithms for input-queued switch scheduling", *IEEE Micro*, vol. 22, no. 1, pp. 10–18, 2002.

[15] P. Giaccone, B. Prabhakar, and D. Shah, "Randomized scheduling algorithms for high-aggregate bandwidth switches", *IEEE J. Selec. Areas Commun.*, vol. 21, no. 4, pp. 546–559, 2003.

[16] Y. Jiang and M. Hamdi, "A fully desynchronized round-robin matching scheduler for a VOQ packet switch architecture", in Proc. IEEE Worksh. High Perform. Switch. Routing HPSR 2001, Dallas, TX, USA, 2001, pp. 407–411.

[17] A. Bianco, P. Giaccone, E. Leonardi, and F. Neri, "A framework for differential frame-based matching algorithms in input-queued switches", in *Proc. 23rd Ann. Joint Conf. IEEE Comp. Commun. Soc. IEEE INFOCOM 2004*, Hong Kong, China, 2004.

Grzegorz Danilewicz received the M.Sc. and Ph.D. degrees in Telecommunications from the Poznan University of Technology, Poland, in 1993 and 2001, respectively. Since 1993, he has been working in the Institute of Electronics, Poznan University of Technology. Currently he is a Professor of Poznan University of Technology and working as a Vice Dean of the Faculty of Electronics and Telecommunications. His scientific interests cover photonic broadband switching systems with special regard to the realization of multicast connections in such systems. He has published about 40 papers.

E-mail: Grzegorz.Danilewicz@et.put.poznan.pl
Chair of Communication and Computer Networks
Faculty of Electronics and Telecommunications
Poznan University of Technology
Polanka st 3
60-965 Poznan, Poland

Marcin Dziuba received the M.Sc. degree in Computer Science and Robotics from the Poznan University of Technology, Poland, in 2010. Since 2010, he is a Ph.D. student at Poznan University of Technology, Chair of Communication and Computer Networks Faculty of Electronics and Telecommunications.

E-mail: Marcin.Dziuba@put.poznan.pl
Chair of Communication and Computer Networks
Faculty of Electronics and Telecommunications
Poznan University of Technology
Polanka st 3
60-965 Poznan, Poland