Delay Performance Analysis for an Agile All-Photonic Star Network

Cheng Peng, Peng He, Gregor v. Bochmann and Trevor J. Hall

Centre for Research in Photonics
School of Information Technology and Engineering
University of Ottawa, Ottawa, ON, K1N 6N5, Canada
{cpeng, penghe, bochmann, thall}@site.uottawa.ca

Abstract. In this paper, we study the delay performance of a centrally-controlled agile all-photonic star WDM network that provides multiplexing in the time domain over each wavelength. We consider two timeslot allocation strategies, First-Fit (FF) and First-Fit+Random (FFR), as well as network scenarios with different propagation delays. Both theoretical analyses and simulation experiments are conducted to evaluate the delay performance of the network. Through analytical and simulation results, we show that allocating residual free bandwidth can significantly improve queuing delay performance under light traffic load while maintaining good delay performance under heavy traffic load, especially for a network scenario with large propagation delays. The results obtained can be used to guide the design of scheduling algorithms especially for large-scale networks.

1 Introduction

The term “agility” in optical networks describes the ability to deploy bandwidth on demand at fine granularity, which radically increases network efficiency and brings to the user much higher performance at reduced cost. One possible scheme to provide such agility in WDM networks is multiplexing in the time domain, which is based on the principle of Time Division Multiplexing (TDM)[1]. In such a context, optical switches along lightpaths must be scheduled to reconfigure every timeslot, or every few timeslots, for bandwidth sharing. The centrally-controlled Agile All-Photonic Networks (AAPN)[1][2] can provide such agility.

As shown in Figure 1, an AAPN consists of a number of hybrid photonic/electronic edge nodes connected together via a core node that contains a stack of bufferless transparent photonic space switches one for each wavelength. A scheduler at a core node is used to dynamically allocate timeslots over the various wavelengths to each edge node. An edge node contains a separate buffer for the traffic destined to each of the other edge nodes. These buffers are called Virtual Output Queues (VOQs) [3] and are used to eliminate the Head-Of-Line blocking problem associated with First-In-First-Out (FIFO) queuing [4]. Traffic aggregation is performed in these buffers, where packets are collected together in fixed-size slots (or,

alternatively, bursts) that are then transmitted as single units across the network via optical links. At the destination edge node the slots are partitioned, with reassembly as necessary, into the original packets. The optical core network does not provide wavelength conversion or buffering, which provides major network architecture simplifications and hardware reductions.

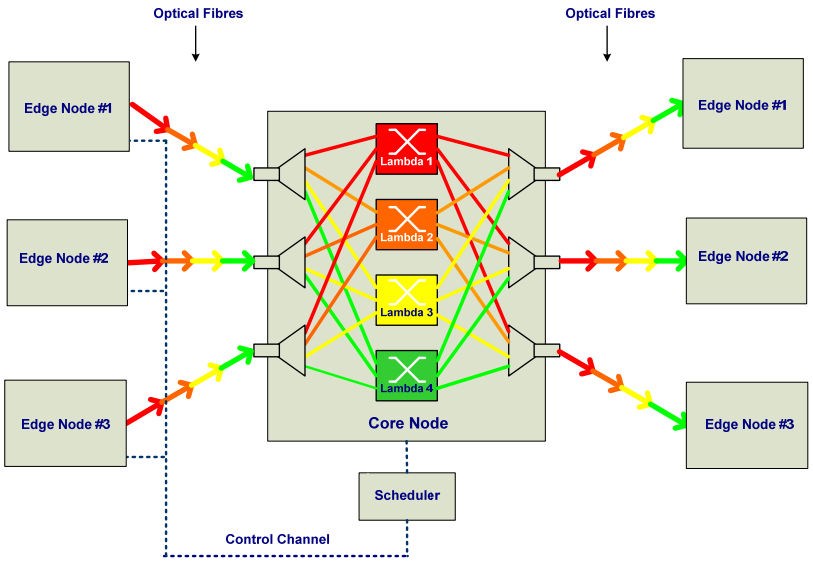

Fig. 1. AAPN star topology

The focus of this paper is on the delay performance of the AAPN with two time-slot allocation strategies as well as network scenarios with different propagation delays. We first review related work in the literature and introduce the signaling protocols between core and edge nodes. Then we provide both theoretical analyses and simulation experiments to evaluate the delay performance of the network by applying these two strategies to AAPN. In the last part of the paper, we conclude that allocating residual free bandwidth can significantly improve queuing delay performance under light traffic load while maintaining good delay performance under heavy traffic load.

2 Related work

In the literature, the delay performance of different scheduling algorithms for an Input Queued (IQ) switch (a switching fabric equipped with buffers at its input ports) have been studied extensively (e.g. [5][6]). The work is relevant to AAPN since the star network formed by each wavelength space switch with its attached edge nodes

can be viewed as a distributed IQ switch. In fact, the AAPN architecture may be viewed as a distributed three-stage Clos packet switch: the edge nodes can be logically split in two parts (the source and the destination modules) and the core node may be explicitly drawn with its de-multiplexers/multiplexers and its wavelength space switches in parallel. The connections between the respective source/destination edge nodes and the core node are seen now as unidirectional (shown in Figure 1).

The delay performance analyses for an IQ switch usually shown in the literature cannot be compared to AAPN because it implies zero propagation delay between the edge nodes (input buffers) and the core node (switch fabric). The propagation delay, however, cannot be ignored since AAPN can be deployed in the backbone of national or large metropolitan networks.

The performance of passive star optical networks was studied extensively in [7][8][9]. By "passive" it is meant that the core node uses couplers or Array Waveguide Grating (AWG) optical devices. Though the network topology is the same as AAPN the switching mechanisms applied to the core node are quite different since passive AWG optical devices are used, as opposed to the dynamic photonic switch fabrics used in the AAPN core node. A scheduling algorithm based on time-slot allocation was proposed in [9] but the delay performance was not studied. The scheduling algorithms discussed in [7] and [8] are quite different from our work because they do not multiplex slots in the time domain over each wavelength.

3 Signaling and scheduling strategies in a TDM-AAPN

3.1 Signaling protocols

When the AAPN operates in TDM mode (TDM-AAPN), each edge node signals a bandwidth request (estimated mainly from queue state information) to the core along control channels (shown as dotted lines in Figure 1) before sending the slots. The scheduler at the core allocates timeslots to each edge node over an appropriate wavelength based on the request. The schedule is signaled back to inform each edge node of the timeslots that it may use to transmit its traffic for each destination, and the core wavelength switches are re-configured in coordination with the edge nodes according to the bandwidth allocated.

3.2 Scheduling strategies

The main task of a scheduler at a core node is to allocate timeslots over an appropriate wavelength to edge nodes.

An ideal scheduler may allocate timeslots in such a way that each edge node may be granted the earliest possible available timeslots based on its bandwidth request. By "earliest possible" we mean the earliest timeslot that an edge node can send a slot without contention. In order to simulate the ideal bandwidth allocation mechanism, a

scheduling strategy, called *First-Fit* (FF), is studied, where the earliest possible timeslots can always be allocated without blocking. The timeslots granted by the scheduler in response to the request are called *reserved* timeslots for the slot for which the request was issued.

Another scheduling strategy is called *First-Fit+Random* (FFR) where the scheduler allocates the earliest possible available timeslot for each edge node based on its bandwidth request and randomly allocates residual free timeslots over all the output links of the AAPN core node. Such randomly allocated timeslots are referred to as *unreserved*. The benefits of this strategy are that slots may be transferred without waiting for their reserved timeslots if an unreserved timeslot assigned to the correct destination is available earlier, which reduces the queuing delay of the slots. Evidently, a slot cannot be sent later than its reserved timeslot.

4 Analytical analyses of delay performance

In this section, analytical analyses of the delay performance for the two timeslot allocation strategies, FF and FFR, over a single wavelength are discussed in the context of AAPN.

4.1 Assumptions

We assume that an AAPN core contains a $N \times N$ photonic space switch for a single wavelength. Each source edge node maintains a VOQ with a FIFO queuing policy for each destination edge node. The capacity of these VOQs is assumed to be infinite. It is also assumed that, in every timeslot, there is an independent and identical probability ρ (load) that traffic arrives at an edge node. The arrivals therefore follow a Poisson distribution. We assume that traffic is equally likely destined for each edge node. The longest propagation delay between core and edge nodes is assumed to be d.

4.2 Analytical analyses of delay performance

Two factors affect the queuing delay performance. One is the scheduling delay Δ; the other is the signaling delay $d_{signaling}$. The scheduling delay describes the waiting time introduced by the scheduler to resolve contention. A bound for Δ follows Shah's delay model in [6] which is slightly tighter than Leonardi's [5] for uniform traffic. According to Section 3, the signaling delay is defined as the round-trip time (measured in timeslots) between the core and the edge nodes.

$$d_{signaling} = 2d \quad (1)$$

4.2.1 FF strategy

The waiting time of a slot that is transmitted through its reserved timeslot is denoted by D_{rsv}. The D_{rsv} is determined by the sum of the signaling and scheduling delay, that is:

$$D_{rsv} = d_{signaling} + \Delta \tag{2}$$

The signaling delay $d_{signaling}$ can be calculated according to (1). The scheduling delay Δ can be calculated according to following arguments.

Given a $N \times N$ photonic space switch, the probability P_{dest} that a slot from a given source edge node is destined to a given destination edge node is:

$$P_{dest} = \frac{1}{N} \tag{3}$$

The probability ρ_{dest} that a slot arrives at a given source edge node and is destined to a given destination edge node is

$$\rho_{dest} = \rho \times P_{dest} = \frac{\rho}{N} \tag{4}$$

In the context of AAPN, there are a total of N VOQs (one for each edge node) with slots destined to the same edge node that contend for the output link of the core node, and there are a total of N VOQs (in one edge node) that contend for the input link of the core node. Therefore, the link load ρ_{link} is:

$$\rho_{link} = N \cdot \rho_{dest} = \rho \tag{5}$$

The scheduling delay that a slot waits in a VOQ can be approximated [6] as,

$$\Delta = \frac{N-1}{N} \cdot \frac{\rho_{link}}{1-\rho_{link}} \cdot \frac{1}{\rho_{link}/N} = \frac{N-1}{1-\rho} \tag{6}$$

Introducing (1) and (6) to (2), we have

$$D_{rsv} = 2d + \frac{N-1}{1-\rho} \tag{7}$$

For FF strategy, a slot at an edge node can only be sent out through its reserved timeslot. Hence the queuing delay of FF strategy D_{FF} is:

$$D_{FF} = D_{rsv} = 2d + \frac{N-1}{1-\rho} \tag{8}$$

4.2.2 FFR strategy

For FFR strategy, it is not necessary for a slot to wait D_{rsv} timeslots for transmission because the slot may be sent out earlier by using an unreserved timeslot. We denote by D_{ursv} the waiting time of a slot that is transmitted through an unreserved timeslot. The D_{ursv} is determined by the scheduling delay only, i.e., when a slot reaches the head of its VOQ, it will depart on the first unreserved timeslot if one is available before its reserved timeslot; otherwise it waits for its reserved timeslot. Hence,

$$D_{ursv} = \Delta \tag{9}$$

For the AAPN deployed in the backbone of national or large metropolitan networks (optical links are more than 100km), nearly all slots are transmitted through unreserved timeslots if the load is less than 0.5 due to the facts that (1) the number of the waiting slots is on average less than that of the unreserved timeslots and (2) the time that a slot waiting for its reserved timeslot is rather long. In case the load exceeds 0.5, it means that some slots are forced to be transmitted through their reserved timeslots because of the shortage of the unreserved timeslots. Based on the thoughts, we discuss the expected queuing delay of the FFR in two scenarios.

1. The expected queuing delay of the FFR when $\rho \geq 0.5$

The probability P_{su} that a slot is sent out through an unreserved timeslot is determined by three conditions: (1) the timeslot is unreserved; (2) the timeslot is destined to the same edge node as the slot; (3) the corresponding VOQ is not empty.

Let us consider a timeslot allocation $A(i)$ by a FFR scheduler for a given edge i, where $i \in \{0,1,...N-1\}$.

Denoted by P_{ursv}, the probability that $A(i)$ is an unreserved timeslot departing from edge node i is:

$$P_{ursv} = 1 - \rho_{link} \tag{10}$$

The probability that $A(i)$ is destined to a given edge j is denoted by P_j where $j \in \{0,1,...N-1\}$. According to the uniform traffic assumption, the reserved timeslots are equally likely destined for each edge node. Due to the random allocation for the residual bandwidth, the unreserved timeslots are equally likely destined for each edge node as well. Thus, for any $j \in \{0,1,...N-1\}$, we have

$$P_j = \frac{1}{N} \tag{11}$$

The probability that the j th VOQ is not empty is denoted by P_{VOQ_j} for any $j \in \{0,1,...N-1\}$. In the context of AAPN, there are a total of N VOQs (one for each edge node) with slots destined to the same edge node that contend for the output

link of the core node, and there are a total of N VOQs (in one edge node) that contend for the input link of the core node. Hence,

$$P_{VOQ_j} = \rho_{link} \tag{12}$$

By (5), (10), (11) and (12), we have

$$P_{su} = P_{ursv} \cdot P_j \cdot P_{VOQ_j} = \frac{\rho(1-\rho)}{N} \tag{13}$$

Accordingly, the probability P_{sr} that a slot is sent out through its reserved timeslot is

$$P_{sr} = (1 - P_{su}) = 1 - \frac{\rho(1-\rho)}{N} \tag{14}$$

Consequently, the expected queuing delay $D_{\geq 0.5}$ for $\rho \geq 0.5$ can be calculated by the following equation.

$$D_{\geq 0.5} = P_{sr} D_{rsv} + P_{su} D_{ursv} \tag{15}$$

Introducing (1), (2), (7), (9), (13) and (14) to (15), we have

$$D_{\geq 0.5} = 2d\left(1 - \frac{\rho(1-\rho)}{N}\right) + \frac{N-1}{1-\rho} \tag{16}$$

2. The expected queuing delay of the FFR when $\rho < 0.5$

For $\rho < 0.5$, more unreserved than reserved timeslots are allocated by the FFR. Almost all slots can be sent out through unreserved timeslots. Based on these thoughts, we assume for now that all slots are sent out through unreserved timeslots, which is equivalent to the case that the core node randomly allocates unreserved timeslots without knowing the bandwidth requests and signals the allocations back to the edge nodes.

According to the assumption, the equivalent effective bandwidth for each link degrades to $1-\rho$ of the link bandwidth since we assume no slots are transmitted through reserved timeslots. Hence, the equivalent link load $\rho_{eqv-link}$ increases up to

$$\rho_{eqv-link} = \frac{\rho}{1-\rho} \tag{17}$$

The scheduling delay Δ that a slot waits in a VOQ can be approximated [6] as

$$\Delta = \frac{N-1}{N} \cdot \frac{\rho_{eqv-link}}{1-\rho_{eqv-link}} \cdot \frac{1}{\rho_{eqv-link}/N} = \frac{N-1}{1-\rho_{eqv-link}} = \frac{N-1}{1-2\rho}(1-\rho) \tag{18}$$

Consequently, the expected queuing delay $D_{<0.5}$ for $\rho < 0.5$ is

$$D_{<0.5} = \Delta = \frac{N-1}{1-2\rho}(1-\rho) \tag{19}$$

3. The expected queuing delay of the FFR

For (19), given a N, we have

$$\lim_{\rho \to 0.5} D_{<0.5} = \infty \tag{20}$$

Equation (20) indicates that the delay will become infinite when the load approaches to 0.5, which is unrealistic. The reason of this result is the assumption that all slots are sent out through unreserved timeslots when $\rho < 0.5$. Actually, when the load increases, some slots have to use their reserved timeslots so the queuing delay would be bounded by Equation (16). Consequently, the expected queuing delay of the FFR D_{FFR} can be expressed by the following equation.

$$D_{FFR} = \begin{cases} D_{\geq 0.5} & \rho \geq 0.5 \\ \min(D_{<0.5}, D_{\geq 0.5}) & otherwise \end{cases} \tag{21}$$

Note that Equation (21) can be solved by combining (16) and (19).

5 Simulation Results and Discussions

In this section, we present both the analytical and simulation results of the delay performance of AAPN using the two scheduling strategies, FF and FFR, with different propagation delays. We demonstrate the accuracy of our analytical techniques by comparing analytical results to simulation. In the simulations, the traffic requests arrive at the network following a Poisson process, and the holding time is exponentially distributed. We assume that all the source-destination node pairs have the same traffic load. The duration of a timeslot is 10 microseconds. The simulation lasts 1,000,000 timeslots.

There are three factors that affect the queuing delay under a given load, i.e., the scheduling strategies (whether using unreserved timeslots or not), the scalability of the AAPN core node (measured in the port dimension N), and the longest propagation delay between the core and the edge nodes, d (measured in timeslot). We study how these factors affect the delay performance in this section.

In Section 4, we assume that all slots are sent out through unreserved timeslots when $\rho < 0.5$. With this assumption, we analyze the delay performance of the FFR. Figure 3 shows the simulation result of the probability (P_{sr} in Equation (14)) that a slot is transmitted through its reserved timeslot under the FFR scheduling strategy. As we can see, when the load is less than 0.5 (light load), the probability is almost zero and thus can be ignored. When the load becomes larger than 0.5, the probability increases very fast for all three propagation delays and hence cannot be ignored. This simulation shows that the assumption for $\rho < 0.5$ is correct.

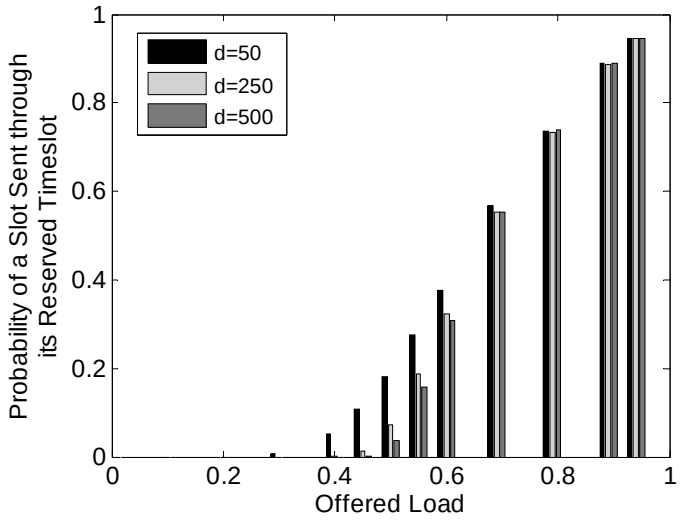

Fig. 3. The probability of a slot transmitted through its reserved timeslot. (*N*=8)

In Figure 4 (Equation (8) and (21)), three curves are presented. The solid ones show the expected queuing delay of the FFR, where some slots are sent out earlier by using unreserved timeslots. The simulation results of the FFR are shown as point marks (no lines). The dotted lines, as a benchmark, give the expected queuing delay by applying the FF strategy, where all slots are sent out through their reserved timeslots.

It is shown that the delay curves of the FFR (Equation (21)) are matched well with the simulation curves, both in amplitude and in the trend of curves in the low-load and high-load regions, which confirms the correctness and exactness of our analytical analysis proposed in Section 4 in general. However, as mentioned earlier, in the region of medium load, the analytical model is not precise because of the assumptions made.

As shown in Figure 4, the delay curves can be divided into three regions according to load, i.e., $\rho < 0.5$, $\rho \approx 0.5$ and $\rho > 0.5$.

In the first region ($0 \le \rho < 0.5$), where the load is light, both FF and FFR curves are fairly "flat". The FF curves are just above $2d$ because of the signaling delay that is equal to $2d$. Compared to the scheduling delay, the signaling delay of the FF dominates the queuing delay in the low load region. As shown in Figure 4, the FFR curves are just above zero. Apparently, FFR outperforms FF greatly because nearly all the slots can be transported by unreserved timeslots.

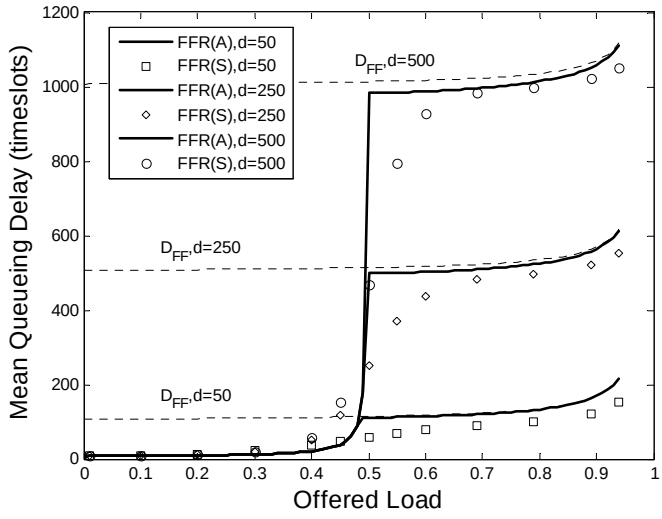

Fig. 4. Queuing delay as a function of offered load for AAPN under various scheduling strategies and different propagation delay (N=8).

The scale of such an improvement is different for different propagation delays. It is shown that FFR can contribute a significant improvement to the delay performance compared to FF in an AAPN with large propagation delays. The reason is that the propagation delay has nearly no influence on the delay performance for FFR (Equation (19)) when $\rho < 0.5$. It hence can be concluded that allocating residual free bandwidth can significantly improve queuing delay performance in the light load region, especially when the propagation delay is large, e.g., for a WAN.

In the second region ($\rho \approx 0.5$), the FFR curves dramatically rise and approach the FF curves. This can be explained by Equation (20) that is the slots are largely accumulated in the VOQs so that some of them are forced to use their reserved timeslots for transmission. It is observed that the simulated FFR curves are not as sharp as the analytical ones. This is because the analytical model assumes that all slots are sent out through unreserved timeslots when $\rho < 0.5$, which ignores the transition by which slots gradually use more reserved timeslots as ρ increases in the first region.

In the third region ($0.5 < \rho < 1$), where the load is heavy, slots are more likely to be sent through their reserved timeslots and hence experience both the signaling delay and the scheduling delay. In this region, signaling delay dominates the delay performance. Thus the FFR has only a limited improvement to the delay performance of the FF. It can be concluded that allocating residual free bandwidth gives a small improvement to the queuing delay performance in the heavy load region.

6 Conclusions and Remarks

In this paper, we study the delay performance of an agile all-photonic star network with centralized schedulers working in TDM mode. A scheduling strategy, called *First-Fit* (FF), which emulates an ideal bandwidth allocation that allocates the earliest available timeslot for each edge node based on its bandwidth request, is studied. The FF shows long queuing delay especially for large-scale networks even if the traffic load is light. Another scheduling strategy, called *First-Fit+Random* (FFR), is studied and shows a significant improvement of the queuing delay performance in the light load region. Through analytical and simulation results, we show that allocating residual free bandwidth can significantly improve the queuing delay performance under light traffic load while maintaining good delay performance under heavy traffic load, especially for a network scenario with large propagation delays.

The model proposed in the paper is for the AAPN deployed in the backbone of national or large metropolitan networks where the propagation delay is quite large. The accuracy of the model may decrease in the LAN environment due to the assumption that all slots are sent out through unreserved timeslots when $\rho < 0.5$. Figure 3 shows that more reserved timeslots are used for transmission in the region of $\rho < 0.5$ with a smaller propagation delay. It therefore reminds us that the propagation delay in LAN can be ignored since the scheduling delay Δ dominates the delay performance which can thus be modeled by classic queuing theories.

The conclusion in this paper can be used to design scheduling algorithms. For example, a Birkhoff-von Neumann decomposition based timeslot allocation algorithm is discussed in [10] where the residual bandwidth is allocated in a round-robin way to gain good delay performance. Furthermore, the residual bandwidth can also be allocated randomly with a weight proportional to the average traffic to the particular destination (average over some recent time period) to adapt to non-uniform traffic.

The current FFR reserves bandwidth for incoming traffic and tries to take advantage of unreserved timeslots. The reserved timeslots can only be used to send the slots that reserve them even if they have been sent out through the unreserved timeslots, which implies that these timeslots cannot be dedicated to others in this case. Hence, one of our future works is to study an enhanced version of FFR, in which a reserved timeslot can become unreserved if its associated slot has already been send out.

Acknowledgment

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and industrial and government partners, through the Agile All-Photonic Networks (AAPN) Research Network. Dr. Trevor J. Hall holds a Canada Research Chair in Photonic Network Technology at the University of Ottawa and is grateful to the Canada Research Chairs Program for their support. The authors are also indebted to Dr. Sofia Paredes and Dr. Jun Zheng for their insightful comments

and careful reading of the manuscript.

References

1. G. v. Bochmann, M.J. Coates, T. Hall, L. Mason, R. Vickers and O. Yang, “The Agile All-Photonic Network: An architectural outline”, Proc. Queen's University Biennial Symposium on Communications, 2004, pp.217-218
2. L.G. Mason, A. Vinokurov, N. Zhao and D. Plant, "Topological Design and Dimensioning of Agile All Photonic Networks", accepted to "Computer Networks", special issue on Optical Networking, edited by Prof. Harry Perros
3. Y. Tamir and G. Frazier, “High performance multiqueue buffers for VLSI communication switches”, In Proceedings of 15th International Symposium on Computer Architecture (ISCA), May/June 1988, pp. 343--354
4. N. McKeown, “The iSLIP Scheduling Algorithm for Input-Queued Switches”, IEEE/ACM Transactions On Networking, vol. 7, April 1999, pp.188--201
5. E. Leonardi, M. Mellia, F. Neri, and M.A. Marsan, “Bounds on delays and queue size averages and variances in input-queued cell-based switches”, Proc. of the IEEE INFOCOM, Anchorage, USA, April 2001,pp.1095-1103
6. D. Shah, and M. Kopikare, “Delay bounds for approximate Maximum weight matching algorithms for input-queued switches”, Proc. of the IEEE INFOCOM, New York, USA, June 2002, pp. 1024-1031
7. M. Maier, M. Scheutzow, M. Reisslein, and A. Wolisz, "Wavelength Reuse for Efficient Transport of Variable-Size Packets in a Metro WDM Network," in Proc., IEEE INFOCOM, vol. 3, June 2002, pp. 1432-1441
8. N. Kamiyama, "A Large-Scale AWG-Based Single-Hop WDM Network Using Couplers With Collision Avoidance," IEEE/OSA JLT, vol. 23, no. 7, July 2005, pp. 2194-2205
9. A. Bianco, E. Leonardi, M. Mellia, and F. Neri, "Network Controller Design for SONATA - A Large-Scale All-Optical Passive Network," IEEE JSAC, vol. 18, no. 10, Oct. 2000, pp. 2017-2028
10. C. Peng, G. v. Bochmann and T. J. Hall, “Quick Birkhoff-von Neumann Decomposition Algorithm for Agile All-Photonic Network Cores”, accepted by 2006 IEEE International Conference on Communications (ICC 2006), Istanbul, Turkey, June 2006