Performance Analysis of Optical Packet Switches Enhanced with Electronic Buffering

Zhenghao Zhang
Computer Science Department
Florida State University
Tallahassee, FL 32306, USA
zzhang@cs.fsu.edu

Yuanyuan Yang
Dept. Electrical & Computer Engineering
Stony Brook University
Stony Brook, NY 11794, USA
yang@ece.sunysb.edu

Abstract

Optical networks with Wavelength Division Multiplexing (WDM), especially Optical Packet Switching (OPS) networks, have attracted much attention in recent years. However, OPS is still not yet ready for deployment, which is mainly because of its high packet loss ratio at the switching nodes. Since it is very difficult to reduce the loss ratio to an acceptable level by only using all-optical methods, in this paper, we propose a new type of optical switching scheme for OPS which combines optical switching with electronic buffering. In the proposed scheme, the arrived packets that do not cause contentions are switched to the output fibers directly; other packets are switched to shared *receivers and converted to electronic signals and will be stored in the buffer until being sent out by* shared *transmitters. We focus on performance analysis of the switch, and with both analytical models and simulations, we show that to dramatically improve the performance of the switch, for example, reducing the packet loss ratio from* 10^{-2} *to close to* 10^{-6}*, very few receivers and transmitters are needed to be added to the switch. Therefore, we believe that the proposed switching scheme can greatly improve the practicability of OPS networks.*

1. Introduction

Optical networks have the potential of supporting ultra fast future communications because of the huge bandwidth of optics. In recent years, Optical Packet Switching (OPS) has attracted much attention since it is expected to have better flexibility in utilizing the huge bandwidth of optics than other types of optical networks [13], [6], [15], [14]. However, OPS in its current form is not yet practical and appealing enough to the service providers because typical OPS switching nodes suffer heavy packet losses due to the difficulty in resolving packet contentions. The common methods for contention resolution include wavelength conversion and all-optical buffering [6], [15], where wavelength conversion is to convert a signal on one wavelength to another wavelength, and all-optical buffering is to use Fiber-Delay-Lines (FDL) to delay an incoming signal for a specific amount of time proportional to the length of the FDL. Wavelength conversion is very effective but it alone cannot reduce the packet loss to an acceptable level, therefore buffering has to be used. However, FDLs are expensive and bulky and can provide only very limited buffering capacity. Thus the main challenge in designing an OPS switch is to find more practical methods to buffer the packets to resolve contention.

For this reason, we propose to use electronic buffers in OPS switches. In the proposed switch, the arrived packets that do not cause contentions are switched to the output fibers *directly*; other packets, called the "leftover packets," are switched to receivers and converted to electronic signals and will be stored in an electronic buffer until being sent out by transmitters. It is important to note that in this scheme not all packets need be converted to electronic signals; such conversion is needed only for packets that cause contentions. Therefore, the advantage of this scheme is that far less high-speed receivers and transmitters are needed compared to switches that convert every incoming packet to electronic signals, since if the traffic is random, it is likely that the majority of the arrived packets can leave the switch directly without having to be converted to electronic signals.

At a switching node in a wide area network, the arrived packets can be categorized into two classes, namely the "to-local packets" which are packets destined for this switching node, and the "non-local packets" which are packets destined for other switching nodes in the network and are only passing by. Also, there are some packets collected by this switching node from the attached local area networks that should be sent into the wide area network, which can be called the "from-local packets." Therefore, to receive the "to-local packets," the switch must be equipped with some receivers to which these packets can be routed to; similarly, to send the "from-local packets," the switch must be equipped with some transmitters that can be used to send the packets to the output fibers. (The receivers and transmitters are also referred to as the "droppers" and the "adders" in some optical networks, respectively.) In previous works on optical switches, the receivers and transmitters are only used for the to-local packets or the from-local packets. What we are proposing in this paper is to *open such resources to the non-local packets*, i.e., to allow the non-local packets to be received by the receivers and sent by the transmitters.

We are interested in finding how many more receivers and transmitters are needed in the switch to achieve acceptable performance in terms of packet loss ratio, delay, etc. With both analytical models and simulations, we will show that the new switch needs a relatively small number of receivers and even less number of transmitters to greatly improve the performance.

The rest of this paper is organized as follows. Section 2 describes some related works. Section 3 describes the operations of the switch. Section 4 studies the performance of the switch under Bernoulli traffic. Section 5 studies the performance of the switch under self-similar traffic. Finally, Section 6 concludes the paper.

2. Related Works

Optical Packet Switching has been studied extensively in recent years and many switch architectures have been proposed and analyzed. For example, [15], [16] considered all-optical switches with output buffer implemented by FDLs and gave analytical models for finding the relations between the size of the buffer and packet loss ratio. However, the results in [15], [16] show that to achieve an acceptable loss ratio, the load per wavelength channel has to be quite light if the number of FDLs is not too large. Switches with shared all-topical buffer have been proposed, for example, in [4], [5], in which all output fibers share a common buffer space implemented by FDLs. However, in this type of switches, to achieve an acceptable packet loss ratio, the number of FDLs is still large and is often no less than the number of input/output fibers, which increases the size of the switching fabric. To avoid the difficulty of all-optical buffering, the recent OSMOSIS project [1], [20], [2] proposed an optical switch with OEO conversion for every channel. In the OSMOSIS switch, all arriving signals are converted to electronic forms and stored in electronic buffer, and then they will be converted back to optical form before entering the switching fabric. The advantage of such an approach is that it needs no optical buffer and does not increase the size of the switching fabric; however, the disadvantage is the expected high cost since it needs high speed receivers, high speed electronic memories and high speed tunable transmitters for every channel. The switch proposed in this paper also uses electronic buffer, however, less buffer, transmitters and receivers are needed because they are *shared* by all channels.

3. Operations of the OPS Switch

3.1. Functionalities of the OPS Switch

We consider a switch with N input/output fibers where on each fiber there are k wavelengths. The switch has R receivers, T transmitters, and an electronic buffer of a very large size. As in [15], [16], the switch operates in a time slotted manner and receives packets of one time slot long at the beginning of time slots. Suppose at one time slot, among the packets arrived on the input fibers of the switch, V packets are to-local packets and H_i packets are non-local packets destined for output fiber i for $1 \leq i \leq N$. The switch will first send the to-local packets to the receivers. We assume the switch is capable of sending each to-local packet to a receiver if $V \leq R$; otherwise, R of the to-local packets will be sent to the receivers and the rest will be dropped. We also assume the switch is capable of sending all H_i packets to output fiber i (on some chosen wavelength) if $H_i \leq k$; otherwise, k packets will be sent. The number of packets that are leftover at output fiber i is $L_i = \max\{H_i - k, 0\}$. If there are still receivers available after receiving the to-local packets, i.e., if $V \leq R$, $\min\left\{\sum_{i=1}^{N} L_i, R - V\right\}$ non-local packets will be sent to the receivers and be converted to electronic signals to be stored in the buffer and others will be dropped. A random algorithm is used to determine which packets should be received and which packets should be dropped. Note that the switch will receive the to-local packets first because on average, the to-local packets have traveled longer distance than the non-local packets before reaching this node, therefore dropping to-local packets will waste more network resources than dropping non-local packets.

The from-local packets collected from the local area networks will also be first sent to the electronic buffer. There are N queues in the electronic buffer, where queue i stores the packets destined for output fiber i, including the leftover packets and the from-local packets. The switch will check each queue to see if there are some packets that can be sent to the output fibers. The number of available wavelengths at output fiber i is $F_i = \max\{k - H_i, 0\}$, thus the number of packets in queue i that can be sent to output fiber i is $C_i = \min\{q_i, F_i\}$, where q_i is the length of queue i. Therefore, the total number of packets that may be sent out is $\sum_{i=1}^{N} C_i$. However, since there are only T transmitters and one transmitter can be used to send only one packet, the number of packets that are actually sent out is $\min\left\{\sum_{i=1}^{N} C_i, T\right\}$. When $\sum_{i=1}^{N} C_i > T$, a random algorithm is used to select packets from the queues to ensure fairness to all queues.

3.2. A Realization of the OPS Switch

A possible realization of the switch is shown in Fig. 1. The composite signal coming from one input fiber will first be sent to a demultiplexer, in which signals on different wavelengths are separated from one another. The separated signal on one wavelength will then be sent to a *wavelength converter* to be converted to another wavelength if needed. The wavelength converters are *full range*, i.e., capable of converting a wavelength to any other wavelength. With full range wavelength converters, an incoming packet can be sent to any wavelength channel by converting the wavelength of

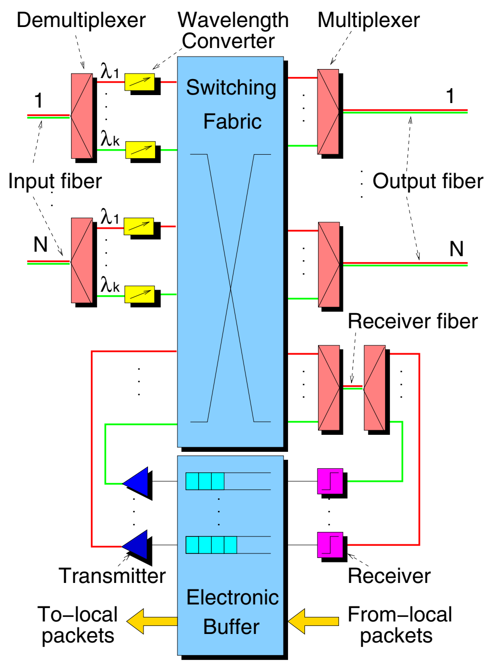

Figure 1. An OPS switch enhanced with electronic buffer.

the packet to the desired wavelength. After the wavelength conversion, the signal is then sent to a switching fabric, which is capable of sending the signal to one of the output fibers or to one of the receiver fibers shown in the right of the figure. Signals sent to the output fibers are combined into one composite signal by the multiplexer and then leave the switch. Signals sent to the receiver fibers are first combined by the multiplexer and then be demultiplexed into signals on separate wavelengths, and each of the demultiplexed signal will be sent to a receiver to be converted to electronic signals. Note that after each receiver fiber there can be at most k receivers, therefore if there are R receivers, there should be $[R/k]^+$ receiver fibers, where $[x]^+$ denotes the minimum integer greater than x. The to-local packets are all sent to the receivers. The non-local packets are sent to the output fibers whenever possible and the leftover ones are sent to the receivers and then to the electronic buffers. The packets stored in the buffer can be sent back to the switching fabric by the transmitters, which are *fast tunable lasers* that can be tuned to *any* wavelength. The switching fabric should be able to send any of the Nk signals from the input fibers to any of the $N + [R/k]^+$ output fibers, and receiver fibers should also be able to send any of the T signals from the transmitters to any of the N output fibers. Note that a simpler way seems to be sending the signals directly to the receivers without sending them to the receiver fibers to go through the multiplex/demultiplex process. However, there are several reasons for the current design choice and the most important one is that otherwise the switching fabric must be much larger, since it must be able to send an arriving packet to $N + R$ fibers instead of only $N + [R/k]^+$ fibers.

4. The Performance of the OPS Switch under Bernoulli Traffic

As mentioned earlier, we are interested in finding the number of receivers and transmitters needed for the switch to have acceptable performance measured by packet loss ratio and packet delay. In this section we will study the performance of the switch under Bernoulli traffic. Both analytical models and simulations will be used, and in our simulations, each point is obtained by running the program for 1,000,000 time slots. Analytical models are used in addition to simulations because they are usually much faster than simulations and can be more accurate when evaluating the likelihood of rare events such as packet loss with ratio under 10^{-6}. Analytical models and simulations can also be used to verify each other: if they match, it is likely that they are both correct since it is highly unlikely that they both went wrong in the same way.

We first introduce some notations and assumptions that will be used throughout this section. Under Bernoulli traffic, the probability that there is a packet arriving at an input wavelength channel in a time slot is the traffic load ρ and is independent of other time slots and other input wavelength channels. Let ρ_l and ρ_n be the arrival rate of to-local packets and non-local packets, respectively, where $\rho = \rho_l + \rho_n$. With probability ρ_n/ρ, an arrived packet is a non-local packet and with probability ρ_l/ρ, an arrived packet is a to-local packet. The destination of an arrived non-local packet is random. We assume that there are Q local ports that can send the from-local packets to the buffer, and the arrival rate of the from-local packets at a local port is ρ_{fl} where $Q\rho_{fl} = Nk\rho_l$, that is, the total arrival rate of the to-local packets and the from-local packets are the same.

For convenience, we will use $B(m, \sigma)$ to denote a Binomial distribution, that is, if a random variable X follows distribution $B(m, \sigma)$,

$$P(X = x) = \left(\begin{array}{c} m \\ x \end{array} \right) \sigma^x (1 - \sigma)^{m-x},$$

where $0 \leq x \leq m$. We will also use $M(m, \alpha, \beta)$ to denote a multinomial distribution, that is, if two random variables X and Y follow distribution $M(m, \alpha, \beta)$,

$$P(X = x, Y = y) = \frac{m!}{x!y!(m - x - y)!} \alpha^x \beta^y (1 - \alpha - \beta)^{(m-x-y)}$$

where x and y are non-negative integers and $x + y \leq m$.

4.1. The Minimum Number of Transmitters

We will first determine the minimum number transmitters needed to send the packets in the buffers. Regarding the buffer as a queuing system, the service rate should be no less than the arrival rate. The arrival rate to the system is $E(L) + Q\rho_{fl}$, where $E(L)$ is the average number of leftover packets and $Q\rho_{fl}$ is the average number of arrived from-local packets. The service rate, on the other hand, is no more than the number of transmitters, T. Therefore a lower bound of T is $E(L)+Q\rho_{fl}$. Note that $Q\rho_{fl}$ is determined by the traffic statistics of the local area network and at least this number of transmitters must be equipped in the switch only to send the from-local packets, thus we need only to derive $E(L)$.

Let L_i be the number of leftover packets destined for output fiber i where $1 \le i \le N$. L_1, L_2, $\ldots$, are random variables with the same distribution, although dependent upon each other. By probability theory, $E(L)=E(L_1+L_2+\cdots+L_N)=NE(L_1)$. Let H_1 be the number of non-local packets arrived for output fiber 1. Note that a packet can be sent out as long as there is some unoccupied wavelength channel on its destination fiber, since the wavelength of a packet can be converted to *any* other wavelength by a full range wavelength converter. Hence if $H_1 \le k$, no packet will be left over; otherwise, H_1-k packets will be left over. Therefore

$$E(L_1)=\sum_{h=k+1}^{Nk}(h-k)P(H_1=h)$$

where H_1 is a Binomial random variable $B(Nk,\rho_n/N)$, since the probability that there is a non-local packet arrived for output fiber 1 on an input wavelength channel is ρ_n/N, and there are totally Nk input channels.

The service rate is no more than T because it also depends on the destinations of the packets in the buffer and the destinations of the newly arrived non-local packets. For example, suppose $T=4$, $k=4$, and there are 4 packets in the buffer, all destined for output fiber 1. If there always arrive 4 non-local packets from the input fibers destined for output fiber 1, the packets in the buffer can never be sent out and therefore the service rate is 0. However, as confirmed by our simulations, as long as T is no less than $E(L)+Q\rho_{fl}$, the queues will be stable, i.e., will not grow to infinite size, which can be roughly explained as follows. Suppose the claim is not true, that is, the number of packets that should be buffered can be infinity when $T > E(L)+Q\rho_{fl}$. Then there must be one queue of infinite length in the buffer. Note that due to the symmetry of the traffic, the arrival rate to the queue is $(E(L)+Q\rho_{fl})/N$. The service rate of the queue is the minimum of T/N and $k(1-\rho_n)+E(L)/N$, where the former is the average number of transmitters used to send packets in this queue and the latter is the number of unoccupied wavelength channels on the output fiber. It can be easily verified that the service rate is larger than the arrival rate, therefore the length of the queue cannot stay at infinity.

In Fig. 2, $E(L)$ as a function of the arrival rate is shown for switches of two sizes when $\rho_n/\rho=0.9$, where the lines are obtained by simulations and the marks are obtained by analytical formulas. We can observe that $E(L)$ is remarkably small, for example, for the switch where $N=8$, $k=16$, when $\rho=0.8$, $E(L)$ is only slightly larger than 1. This is a very encouraging fact since it means that very few

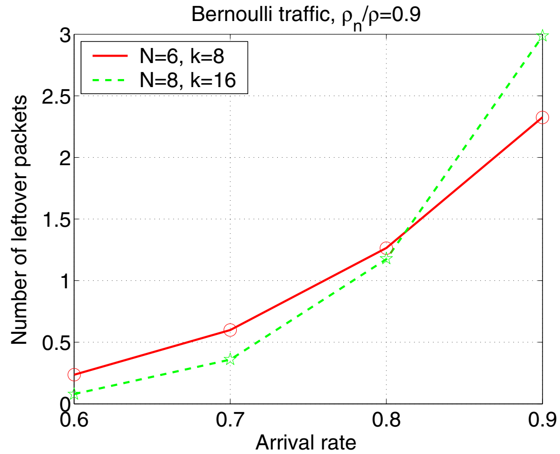

Figure 2. The average number of leftover packets for switches of two sizes when $\rho_n/\rho=0.9$ under Bernoulli traffic. The lines are obtained by simulations and the marks are obtained by analytical formulas.

transmitters are needed to be added to the switch to make sure that the non-local packets will not overflow the buffer.

4.2. The Minimum Number of Receivers

We next wish to find the minimum number of receivers needed to make sure that the packet loss probability is below a preset threshold.

As mentioned earlier, we assume that in case both to-local and non-local packets need to be received, to-local packets have higher priority, i.e., will be sent to receivers first, and the non-local packets can be sent to receivers only if there are some receivers left. The total number of arrived to-local packets, denoted by V, is a Binomial random variable $B(Nk,\rho_l)$. The packet loss probability (PLP) of to-local packets is thus

$$\sum_{v=R+1}^{Nk}(v-R)\left(\begin{array}{c}Nk\\ v\end{array}\right)\rho_l^v(1-\rho_l)^{Nk-v}/(Nk\rho_l)$$

where R is the total number of receivers. Fig. 3 shows the packet loss probability of to-local packets as a function of the number of receivers for switches of two sizes when $\rho_n/\rho=0.9$, where the lines are obtained by simulations and the marks are obtained by analytical formulas. We can see that to make the loss rate lower than an acceptable level, in general, a significant amount of receivers are needed. For example, for the switch where $N=8$, $k=16$, when $\rho=0.8$, to make the loss ratio close to 10^{-6}, at least 24 receivers are needed. Note that the switch *has* to have these number of receivers to receive to-local packets, and in the following, we will find how many more receivers are needed to be *added* to the switch to make sure that the loss ratio of the non-local packets is acceptably low.

To find the packet loss probability of non-local packets, we begin with the total number of packets arrived at the

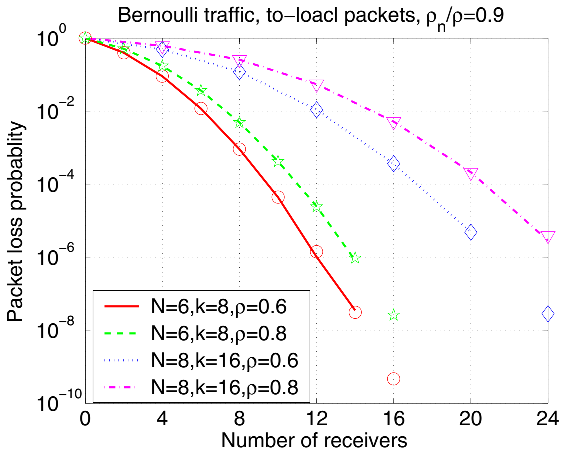

Figure 3. Packet loss probability of to-local packets for switches of two sizes when $\rho_n/\rho = 0.9$ under Bernoulli traffic. The lines are obtained by simulations and the marks are obtained by analytical formulas.

input fibers of the switch, including the to-local and the non-local packets, denoted as Y. Y is a Binomial random variable $B(Nk, \rho)$. Let the total number of non-local packets be S. The probability that if y packets arrive, there are s non-local packets is

$$P(S = s|Y = y) = \left(\begin{array}{c} y \\ s \end{array} \right) (\rho_n/\rho)^s (1 - \rho_n/\rho)^{y-s}$$

for $0 \leq s \leq y$. If there are R receivers, there will be $R-y+s$ left for the non-local packets. Let the probability that given $S = s$, there are l packets left over be written as $P(L = l|S = s)$. The packet loss probability (PLP) of the non-local packet is thus

$$\sum_{y=0}^{Nk} \sum_{s=0}^{y} \sum_{l=0}^{s} wP(L = l|S = s)P(S = s|Y = y)P(Y = y)/(Nk\rho_n)$$

where $w = \max\{0, l - R + y - s\}$.

It remains to find $P(L = l|S = s)$ to determine the packet loss probability. For convenience, the N output fibers can be considered as N boxes each with capacity k and the s packets can be considered as s balls, each to be randomly placed in one of the boxes. $P(L = l|S = s)$ is the probability that given there are s balls, l balls cannot be placed into their destination boxes because these boxes are full. The number of balls to be placed in box i is H_i and let $S_i = \sum_{j=1}^{i} H_j$. Define L_s^i as the number of leftover balls from box 1 to box i given $S_i = s$. Apparently, the p.m.f. of L_s^1 can be determined as:

$$P(L_s^1 = t) = \left\{ \begin{array}{ll} 1 & t = \max\{0, s-k\} \\ 0 & \text{otherwise} \end{array} \right.$$

The probability that L_s^i is a certain value, say, l, can be written as follows by conditioning on H_i:

$$P(L_s^i = l) = \sum_{h=0}^{s} P(L_{s-h}^{i-1} = l - z)P(H_i = h|S_i = s) \quad (1)$$

where $z = \max\{0, h-k\}$ which is the number of leftover balls of box i given there are h balls to be placed in this box. This equation holds since the total number of leftover balls from box 1 to box i is the number of leftover balls from box 1 to box $i-1$ plus the number of leftover balls of box i. This suggests an inductive way to analytically find the p.m.f. of L_s^i by starting with the p.m.f. of L_s^1, then use Eq. (1) to find L_s^i for larger i in each step. Note that

$$P(H_i = h|S_i = s) = \left(\begin{array}{c} s \\ h \end{array} \right) (1/i)^h (1 - 1/i)^{s-h},$$

and $P(L = l|S = s)$ is simply $P(L_s^N = l)$, by definition.

Fig. 4 shows the packet loss probability of non-local packets as a function of the number of receivers for switches of two sizes when $\rho_n/\rho = 0.9$, where the lines are obtained by simulations and the marks are obtained by analytical formulas. First note that our analytical results agree very well with the simulation results. It is very surprising to us to notice that very few receivers are needed to be added to the switch to greatly reduce the loss ratio of the non-local packets. For example, for the switch where $N = 8$, $k = 16$, when $\rho = 0.8$, if there is no receiver that is used to receive the non-local packets, the loss ratio is about 10^{-2}. However, the packet loss ratio is reduced to close to 10^{-6} when there are totally 24 receivers. Note that originally 24 receivers are needed to receive the to-local packets to make the loss ratio of the to-local packets close to 10^{-6}, thus, in this case, *no* receivers are needed to be added to the switch to reduce the loss ratio of the non-local packets from 10^{-2} to 10^{-6}! This is another very encouraging fact for supporting our new proposed scheme. The reason for this is that local packets and non-local packets all come from the input fibers of the switch, thus, when there are more local packets arrived, there will be less non-local packets that are left over, and vice versa. Sharing the same set of receivers can take advantage of this fact and therefore reduce the number of receivers.

4.3. Average Packet Delay

The average packet delay is harder to find because queues in the switch are not simple queues since they are interacting with each other by sharing the same set of transmitters. Intuitively, increasing the number of transmitters will reduce the packet delay. Therefore in this section we wish to find the relations between the number of transmitters and the packet delay. Since the size of the electronic buffer can be very large and the number of receivers have been chosen to guarantee a very low packet loss ratio, to simplify our study, we can assume that there is no packet loss in the switch and the packet delay is only determined by the number of inputs/outputs and the number of transmitters.

There are N queues in the buffer, one for each output fiber. Note that since the T transmitters are shared by all queues, the number of packets that can be sent out from a queue is

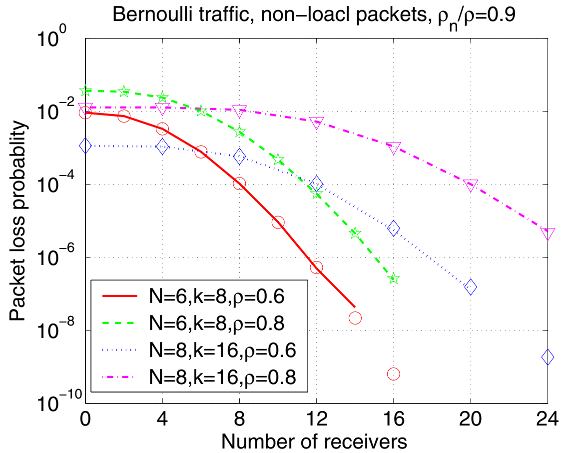

Figure 4. Packet loss probability of non-local packets for switches of two sizes when $\rho_n/\rho = 0.9$ under Bernoulli traffic. The lines are obtained by simulations and the marks are obtained by analytical formulas.

also determined by the number of available transmitters, i.e., the number of transmitters that are "left" by other queues. Due to this reason, a good analytical model must consider the N queues jointly. Since the input traffic is memoryless, one straightforward way to accurately model the N queues is to model them as an N-dimensional Markov chain, however, this will result in a state space growing exponentially with N and is thus not practical. Therefore we have used an approximation model to reduce the complexity. Our model is based on the idea of *aggregation* and finds the behavior of the N queues in an inductive way, whereas after step I $(1 \leq I \leq N)$ it will have found the behavior of I queues. To elaborate, note that at the first step when $I = 1$, the behavior of only one queue is easy to obtain. Suppose after step I, we have found the behavior of I queues when *aggregated* into a block, that is, the I queues will no longer be viewed as I separate queues but as a single component with I queues inside. We can now study $I + 1$ queues by regarding them as two components, that is, by regarding the first I queues as a block and queue $I + 1$ as a separate queue. Note that the behaviors of both components are known at this moment, therefore, the behavior of $I + 1$ queues can be found. After this we can *aggregate* the $I + 1$ queues into one block, then study $I + 2$ queues by regarding the first $I + 1$ queues as a block, and so on, until all queues have been aggregated. The advantage of this method is that it has polynomial complexity and can be much more accurate than other approaches.

The idea of aggregation was first introduced in [19] for switches with shared buffer where all queues share a common buffer. In a shared buffer switch, queues interact with each other through the fixed size buffer, for example, if some queues are long, i.e., occupying most of the buffer space, other queues must be short since the total buffer space is limited. The model given in this paper is also based on the idea of aggregation, however, the model is completely different from the one in [19] because the switch architectures are completely different and the ways the queues are interacting with each other are completely different.

In the following we describe the details of the model. The meanings of the symbols used in this model are summarized as follows:

- s : the number of non-local packets arrived for the first I outputs
- c : the number of from-local packets arrived for the first I outputs
- l : the number of leftover non-local packets among newly arrived non-local packets
- u : the number of packets currently stored in these I queues
- x : the number of packets in these I queues that *can* be sent to the output fibers

In addition, variables with a prime are used to denote corresponding values associated with queue $I+1$, for example, s' is the number of non-local packets arrived for output $I+1$, and so on.

Note that x is the number of packets that *can* be sent out and is not the number of packets that are actually sent out. For example, when $I = 1$, if there are 5 packets in queue 1 and there are $k-4$ non-local packets arrived for output fiber 1, the number of packets that can be sent out is 4. However, if there are only 3 transmitters, the number of packets that are actually sent out is 3. In our model, we assume that the transmitters are assigned to the queues according to a predetermined order, that is, they will be first used to send packets in queue 1, then the remaining transmitters will be used to send packets in queue 2, and then queue 3, etc. Thus, given x and x', $\min\{x, T\}$ packets are sent out among queue 1 to queue I and $\min\{x', \max\{T - x, 0\}\}$ packets are sent out in queue $I + 1$. This assignment strategy is not fair to all queues and is biased toward queues with low indices, however, it makes the analysis tractable and moreover, our simulations show that the packet delay under this assignment strategy is very close to that under a fairer random assignment strategy.

The behavior of a block containing I queues is described by a conditional probability written as $CT_I(x|s, l, u)$, which can be interpreted as the probability that there are x packets in the queues that can be sent out, given that there are currently u packets in the queues and there are s non-local packets arrived for the first I outputs and among them l are leftover. At the beginning when $I = 1$, note that $l = \max\{s - k, 0\}$, and the number of packets in queue 1 that can be sent out is $x = \min\{u, \max\{k - s, 0\}\}$. Thus $CT_1(x|s, l, u)$ is 1 for l and x satisfying these conditions, otherwise it is 0.

To study $I + 1$ queues, we will model them as a two-dimensional Markov chain (u, u'). Denote a generic initial state as (u_0, u'_0). First consider when s, s', c and c' are given. Given s non-local packets arrived for output 1 to output I, the probability that l packets are leftover is $P(L^I_s = l)$

which can be found by Eq. (1). The probability that there are x packets stored in queue 1 to queue I that can be sent out is $CT_I(x|s,l,u_0)$, which has been found in the previous step. Also note that given s', $l' = \max\{s'-k, 0\}$ and $x' = \min\{u'_0, \max\{k-s', 0\}\}$. Let

$$D(u'_1, s', c', u'_0, x) = \begin{cases} 1 & u'_1 = u'_0 - \min\{x', \max\{T-x, 0\}\} + l' + c' \\ 0 & \text{otherwise} \end{cases}$$

Thus, the transition rate from (u_0, u'_0) to another state denoted as (u_1, u'_1) when s, s', c and c' are given is

$$\Lambda(u_1, u'_1|u_0, u'_0, s, s', c, c') = \sum_{l,x} CT_I(x|s,l,u_0)P(L_s^I = l)D(u'_1, s', c', u'_0, x)$$

for all l and x satisfying $u_1 = u_0 - \min\{x, T\} + l + c$. The transition rate from (u_0, u'_0) to (u_1, u'_1) is thus

$$\sum_{s,s'c,c'} p_{s,s'}(s,s')p_{c,c'}(c,c')\Lambda(u_1, u'_1|u_0, u'_0, s, s', c, c')$$

where $p_{s,s'}(s,s')$ is the probability that there are s non-local packets arrived for output 1 to output I and s' non-local packets arrived for output $I+1$, and $p_{c,c'}(c,c')$ is the probability that there are c from-local packets arrived for output 1 to output I and c' from-local packets arrived for output $I+1$. $p_{s,s'}(s,s')$ and $p_{c,c'}(c,c')$ can be found according to the multinomial distribution. It can be verified that $p_{s,s'}(s,s')$ follows $M(Nk, \frac{I\rho_n}{N}, \frac{\rho_n}{N})$ and $p_{c,c'}(c,c')$ follows $M(Q, \frac{I\rho_{fl}}{N}, \frac{\rho_{fl}}{N})$ where Q is the number of local ports.

After obtaining the transition rate of the Markov chain, the steady state distribution, $\pi(u,u')$, can be found. We can then find the behavior of $I+1$ queues described by conditional probability $CT_{I+1}(x^*|s^*, l^*, u^*)$, where variables with superscript '*' denote values associated with $I+1$ queues defined similarly as those for I queues. We call (s, s', l, u, u') a "sub-state" of (s^*, l^*, u^*) if $s + s' = s^*$, $l = l^* - \max\{s'-k, 0\}$, and $u + u' = u^*$. Let $\Omega(x^*|s, s', l, u, u')$ be the probability that there are totally x^* packets from queue 1 to queue $I+1$ that can be sent out in sub-state (s, s', l, u, u'). Since there can be $x' = \min\{u', \max\{k-s', 0\}\}$ packets sent out from queue $I+1$, $\Omega(x^*|s, s', l, u, u')$ is simply the probability that there are $x^* - x'$ packets that can be sent out from queue 1 to queue I which is $CT_I(x^* - x'|s, l, u)$. Next, letting $P(s, s', l, u, u')$ be the probability of that the $I+1$ queues are in sub-state (s, s', l, u, u'), we have

$$\begin{aligned} P(s, s', l, u, u') &= P(l|s, s', u, u')P(s, s', u, u') \\ &= P(Ls_s^I = l)p_{s,s'}(s,s')\pi(u,u') \end{aligned}$$

Let $P(s^*, l^*, u^*)$ be the probability that the $I+1$ queues are in state (s^*, l^*, u^*). Clearly, $P(s^*, l^*, u^*) = \sum_i P(s_i, s'_i, l_i, u_i, u'_i)$ where $(s_i, s'_i, l_i, u_i, u'_i)$ denotes the i^{th} sub-state of (s^*, l^*, u^*). Then,

$$CT_{I+1}(x^*|s^*, l^*, u^*) = \sum_i \Omega(x^*|s_i, s'_i, l_i, u_i, u'_i)P(s_i, s'_i, l_i, u_i, u'_i)/P(s^*, l^*, u^*)$$

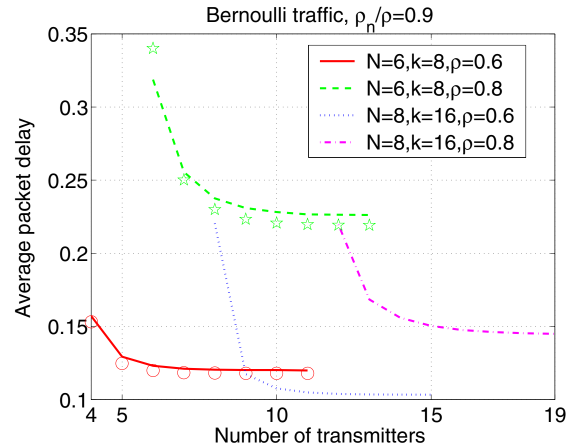

Figure 5. Packet delay for switches of two sizes when $\rho_n/\rho = 0.9$ under Bernoulli traffic. The lines are obtained by simulations and the marks are obtained by our analytical model.

After aggregating all N queues, the stationary distribution of the total number of packets in the buffer can be found, with which the average number of packets stored in the buffer can be found. The average packet delay can then be obtained by the Little's formula.

Fig. 5 shows the packet delay as a function of the number of transmitters for switches of two sizes when $\rho_n/\rho = 0.9$, in which the lines are obtained by simulations and the marks are obtained by our analytical model. First note that our analytical model agrees reasonably well with the simulations. We also found that when the number of transmitters is less than the minimum number of transmitters required in the switch, the packet delay becomes very long (not plotted in the figure); otherwise, the packet delay is relatively short. Another important observation is that the packet delay drops fastest at the beginning and almost ceases to drop when the number of transmitters further increases. This is because when there are enough number of transmitters, the packet delay will be mainly determined by the availability of wavelength channels on the output fibers. This suggests that not too many extra transmitters are needed to reduce the packet delay to close to the minimum level.

5. The Performance of the OPS Switch under Self-Similar Traffic

We have also studied the performance of the switch under self-similar traffic. Self-similar traffic is viewed as a more realistic traffic model because it has been shown by measurement studies that network traffic exhibits self-similarity and long range dependence [25]. We mainly used simulations in our study because unlike the Bernoulli traffic, self-similar traffic cannot be described in simple mathematical forms.

The self-similarity of traffic is described by the Hurst parameter, H, which takes value from 0.5 to 1. The larger the Hurst parameter, the more self-similar the traffic. It has been proved in [25] that self-similar traffic can be generated by aggregating a large number of independent on-off sources where the distributions of the on period and the off period follow heavy-tailed distributions such as the Pareto distribution. If a random variable X follows Pareto distribution $P(x_m, \alpha)$, $P(X < x) = 1 - (\frac{x_m}{x})^\alpha$, and the mean of X is $\frac{\alpha x_m}{\alpha - 1}$. In our simulations, the aggregated traffic of a total of 200 independent on-off sources is sent to an input fiber. The on period of an on-off source follows Pareto distribution $P(T_{on}, \alpha)$ while the off period follows Pareto distribution $P(T_{off}, \alpha)$, where T_{on} and T_{off} are constants and $\alpha = 3 - 2H$. In our simulations, T_{off} is fixed as 2.0 while T_{on} varies depending on the traffic load ρ. The on period of an on-off source represents the bursty traffic from one node to another node. A burst may have $N + 1$ possible destinations, that is, it can either go to one of the N output fibers or it can be a to-local burst. As a way to aggregate the traffic, for each input fiber of the switch, there are $N + 1$ queues which collect the bursts to the N output fibers plus the to-local bursts. Note that these queues are only for generating the self-similar traffic and are not part of the switch. At one time slot, a random algorithm is used to determine bursts in which queues can be sent to the input fiber.

In our simulations, each queue in the buffer of the switch may hold up to 1,000 packets. This size is chosen such that the packet loss is mainly caused by the lacking of receivers rather than by buffer overflow, since the high-speed receivers are harder to implement than the electronic memories. We show the results when $H = 0.7$ in Fig. 6 to Fig. 9 where each point is obtained by running the program for 10,000,000 time slots. Similar observations can be drawn as in Fig. 2 to Fig. 5, receptively. However, with the same number of transmitters and receivers, the switch under self-similar traffic has higher loss ratio and longer packet delay than those under Bernoulli traffic, especially when the traffic load is heavy. This is somewhat expected because self-similar traffic is much more "bursty" than Bernoulli traffic.

6. Conclusions

In this paper we have studied the performance of a new type of optical switch which combines optical switching with electronic buffering. In this switch not all optical packets need to be converted to electronic form and only those that cannot be sent to the output fibers due to contentions are converted by shared receivers to be stored in the buffer. We have shown with analytical models and simulations that the performance of the switch can be greatly improved by adding very few receivers and transmitters. We therefore believe that this switching scheme can greatly improve the practicability of OPS networks and should be used in future optical networks.

Acknowledgments

This research work was supported in part by the U.S. National Science Foundation under grant numbers CCR-

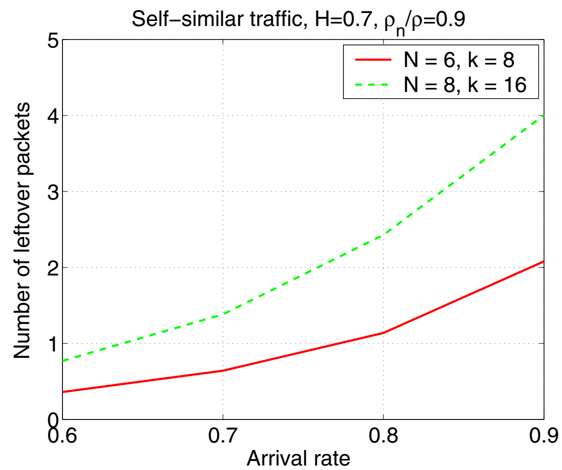

Figure 6. The average number of leftover packets for switches of two sizes when $\rho_n/\rho = 0.9$ under self-similar traffic.

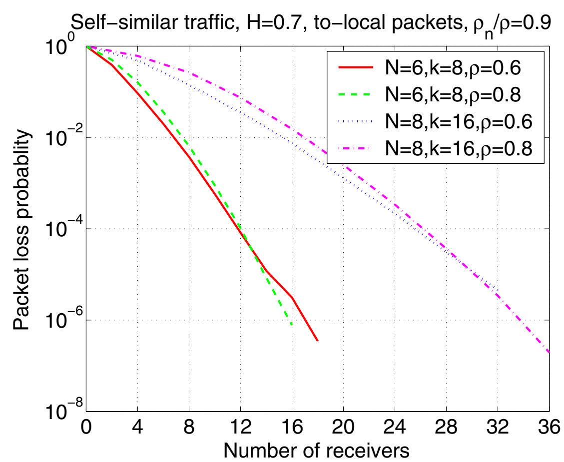

Figure 7. Packet loss probability of to-local packets for switches of two sizes when $\rho_n/\rho = 0.9$ under self-similar traffic.

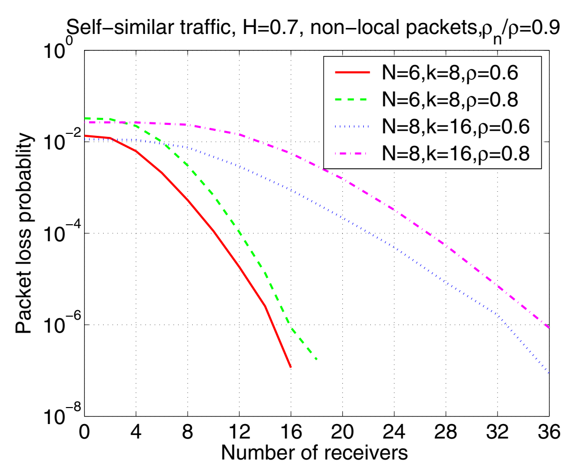

Figure 8. Packet loss probability of non-local packets for switches of two sizes when $\rho_n/\rho = 0.9$ under self-similar traffic.

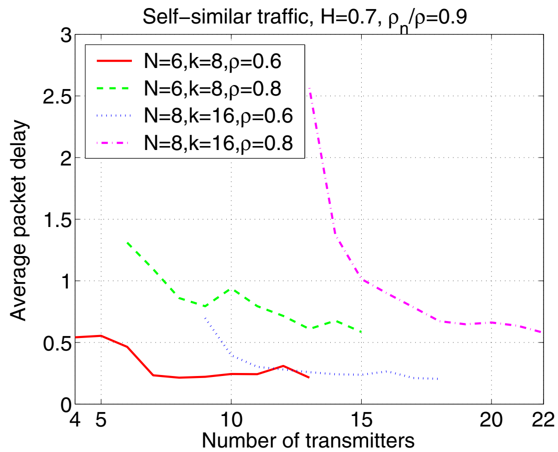

Figure 9. Packet delay for switches of two sizes when $\rho_n/\rho = 0.9$ under self-similar traffic.

0207999 and CCF-0744234.

References

[1] I. Iliadis and C. Minkenberg, "Performance of a speculative transmission scheme for scheduling-latency reduction," *IEEE/ACM Transactions on Networking*, vol. 16, no. 1, pp. 182-195, Feb. 2008.

[2] R.R. Grzybowski, B.R. Hemenway, M. Sauer, C. Minkenberg, F. Abel, P. Mller and R. Luijten, "The OSMOSIS optical packet switch for supercomputers: Enabling technologies and measured performance," *Proc. IEEE LEOS Photonics in Switching 2007*, San Francisco, CA, Aug. 2007.

[3] C. Minkenberg, et. al "Designing a crossbar scheduler for HPC applications," *IEEE Micro*, vol. 26, pp. 58-71, May-June 2006.

[4] G. Bendeli, et al., "Performance assessment of a photonic ATM switch based on a wavelength controlled fiber loop buffer", *OFC'96 Technical Digest*, pp. 106-107, OFC, 1996.

[5] D.K. Hunter, et al., "WASPNET: a wavelength switched packet network," *IEEE Communications Magazine*, vol. 37, no. 3, pp. 120-129, Mar. 1999.

[6] L. Xu, H.G. Perros and G. Rouskas, "Techniques for optical packet switching and optical burst switching," *IEEE Communications Magazine*, pp. 136-142, Jan. 2001.

[7] C. Qiao and M. Yoo, "Optical burst switching (OBS) - a new paradigm for an optical Internet," *Journal of High Speed Networks*, vol. 8, no. 1, pp. 69-84, 1999.

[8] R. Ramaswami and K.N. Sivarajan, *Optical Networks: A Practical Perspective*, 1st Ed., Academic Press, 2001.

[9] Z. Zhang and Y. Yang, "Optimal scheduling in WDM optical interconnects with arbitrary wavelength conversion capability," *IEEE Trans. Parallel and Distributed Systems*, vol. 15, no. 11, pp. 1012-1026, Nov. 2004.

[10] Z. Zhang and Y. Yang, "Optimal scheduling in buffered WDM packet switching networks with arbitrary wavelength conversion capability," *IEEE Transactions on Computers,* vol. 55, no. 1, pp. 71-82, Jan. 2006.

[11] R. Ramaswami and G. Sasaki, "Multiwavelength optical networks with limited wavelength conversion," *IEEE/ACM Trans. Networking*, vol. 6, pp. 744-754, Dec. 1998.

[12] "http://www.emcore.com/assets/fiber/ds00-306_EM.pdf."

[13] B. Mukherjee, "WDM optical communication networks:progress and challenges," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 10, pp. 1810-1824, Oct. 2000.

[14] T.S. El-Bawab and J.-D. Shin, "Optical packet switching in core networks: between vision and reality," *IEEE Communications Magazine*, vol. 40, no. 9, pp. 60-65, Sept. 2002.

[15] S.L. Danielsen, C. Joergensen, B. Mikkelsen and K.E. Stubkjaer, "Analysis of a WDM packet switch with improved performance under bursty traffic conditions due to tunable wavelength converters," *J. Lightwave Technology*, vol. 16, no. 5, pp. 729-735, May 1998.

[16] S.L. Danielsen, B. Mikkelsen, C. Joergensen, T. Durhuus and K.E. Stubkjaer, "WDM packet switch architectures and analysis of the influence of tunable wavelength converters on the performance," *J. Lightwave Technology*, vol. 15, no. 2, pp. 219-227, Feb. 1998.

[17] R. Van Caenegem, D. Colle, M. Pickavet, P. Demeester, J.M. Martinez, F. Ramos and J. Marti, "From IP over WDM to all-optical packet switching: economical view," *Journal of Lightwave Technology*, vol. 24, no. 4, pp. 1638-1645, Apr. 2006.

[18] R. Luijten, C. Minkenberg, R. Hemenway, M. Sauer and R. Grzybowski; "Viable opto-electronic HPC interconnect fabrics," *Proceedings of the ACM/IEEE Supercomputing Conference 2005*, pp: 18-18, Nov. 2005.

[19] Z. Zhang and Y. Yang, "A novel analytical model for switches with shared buffer," *IEEE/ACM Transactions on Networking*, vol. 15, no. 5, pp. 1191-1203, Oct. 2007.

[20] R. Hemenway, R. R. Grzybowski, C. Minkenberg and R. Luijten "Optical-packet-switched interconnect for supercomputer applications," *Journal of Optical Networking*, vol. 3, no. 12, pp. 900-913, Dec. 2004.

[21] J. Gripp, M. Duelk, J.E. Simsarian, A. Bhardwaj, P. Bernasconi, O. Laznicka and M. Zirngibl, "Optical switch fabrics for ultra-high-capacity IP routers," *Journal of Lightwave Technology*, vol. 21, no. 11, pp. 2839-2850, Nov. 2003.

[22] B. Ma, Y. Nakano and K. Tada, "Novel all-optical wavelength converter using coupled semiconductor optical amplifiers," *Lasers and Electro-Optics, 1998. CLEO 98*, pp. 477 - 478, May 1998.

[23] G. Grimment and D. Stirzaker, *Probability and Random Processes*, 3rd Edition, Oxford University Press, 2001.

[24] N. McKeown, "The iSLIP scheduling algorithm input-queued switch," *IEEE/ACM Trans. Networking*, vol. 7, pp. 188-201, Apr. 1999.

[25] W. Willinger, M.S. Taqqu, R. Sherman and D.V. Wilson, "Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Trans. Networking*, vol. 5, no. 1, pp. 71-86, Feb. 1997.