The Effects of Tuning Time in Bandwidth-Limited Optical Broadcast Networks

Murat Azizoğlu[†] Richard A. Barry[‡] Ahmed Mokhtar[†]

[†]Department of Electrical Engineering
University of Washington
Seattle, WA 98195

[‡]Lincoln Laboratory
Massachusetts Institute of Technology
Lexington, MA 02173

Abstract

We consider the effects of tuning delay in optical broadcast networks. We show that for off-line scheduling these effects are small even if the tuning time is as large as the packet duration. In particular, we consider scheduling of random traffic with tunable transmitters and fixed-tuned receivers. We provide a lower bound to the completion time of any off-line schedule with an arbitrary number of wavelengths. We then describe a near-optimal schedule which is based on the principle of having idle transmitters tune to wavelengths just-in-time to start their transmissions. Stability and capacity issues in the transmission of real-time traffic are considered. We show that the scheduling problem admits a single stable equilibrium point, and point out how the traffic capacity of a broadcast network can be reached. We also consider the implications in connection-oriented networks.

1 Introduction

We consider packet transmissions in an all-optical, wavelength division multiplexed (WDM) network with a broadcast star physical topology. Each of the N nodes in the network has a single tunable transmitter and a single fixed-tuned receiver. There are W ($\leq N$) wavelengths in the network. $K = N/W$ receivers share a wavelength. (We assume for simplicity that K is an integer.) For reasons that will become apparent, we say the network is *bandwidth limited* if $W \leq N/2$, so that each wavelength is shared by at least two receivers. Our main focus is on bandwidth-limited networks; however the case where each receiver is assigned a unique wavelength ($W = N$) will also be considered.

Any signal transmitted on any wavelength is received by all the nodes simultaneously (the propagation delays are negligible). Therefore, some form of coordination or scheduling is required to ensure that no two transmitters use the same wavelength at the same time [1]. This coordination is complicated by the fact that a transmitter needs time to tune from one wavelength to another. We define the normalized tuning delay δ as the time for a transmitter to tune, expressed in units of packet duration. The value of δ depends on the transmission rate, the packet size, and the laser tuning time. For instance, with a 1 Gbps rate, 1000 bit packets, and 1 μs tuning time, the normalized tuning delay $\delta = 1$. Advances in rapidly tunable lasers and optical filters will make δ smaller. Conversely, higher transmission speeds and smaller packet sizes increase δ.

We mainly consider the effects of tuning delay on bandwidth-limited networks supporting random traffic when $\delta \leq 1$. For results on deterministic traffic with larger δ, see [2, 3]. Our major contribution is to show that the inefficiency due to tuning delay can be eliminated through off-line scheduling, provided that the tuning times are shorter than the packet duration. We also consider scheduling circuit connections on a broadcast star with tunable transmitters and fixed-tuned receivers. This problem admits an identical formulation to the packet transmission case outlined above. We show that there is no throughput penalty associated with tuning delay as long as the tuning times are shorter than the time slots.

In Section 2, we introduce the mathematical model and summarize previous work. In Section 3, we provide a lower bound on the average time to transmit a set of packets. We also provide an upper bound by considering a schedule Pieris and Sasaki used for deterministic traffic [2]. In Section 4, we introduce a simple near-optimal schedule for random traffic. Section 5 relates the results to a real-time traffic situation, shows the inherent stability of the scheduling problem, and obtains the traffic capacity. The issue of time-slot assignment in connection-oriented broadcast networks is considered in Section 6. Conclusions are given in Section 7.

2 Traffic Model and Previous Work

We now present a general model which can be used for both connectionless and connection-oriented traffic. Consider first a packet transmission scenario. An off-line scheduler assigns a sequence of transmission times to head-of-line packets for all source-destination pairs. In particular, there is a random traffic matrix D with $d_{ij} = 1$ if the source-destination pair (i, j) has a packet to be scheduled and $d_{ij} = 0$ otherwise. We assume that the random variables $\{d_{ij}\}$ are independent and identically distributed (i.i.d.) with $\Pr(d_{ij} = 1) = p$. Thus the matrix D is composed

0743-166X/95 $04.00 © 1995 IEEE

of N^2 i.i.d. Ber(p) (Bernoulli distributed with parameter p) random variables. The parameter p models the buffer occupancy probability for each of the N^2 source-destination buffers.

Our goal is to investigate how the packet transmissions must be scheduled for a given set of parameters $\{N, W, \delta, p\}$ so that the average time to transmit D is minimized. We assume that the traffic matrix D is known to all nodes so that every node can determine its sequence of transmissions and tunings independently of other nodes.

The constraints on an admissible schedule are easily expressed in terms of the traffic matrix D:

$C1$) For $d_{ij} = 1$, let $t(i,j)$ be the starting time of transmission of the corresponding packet. Then

$$| t(i,j) - t(i',j') | \geq 1$$

for all (not necessarily distinct) rows i, i' and for all columns j, j' that belong to the same wavelength.

$C2$) Tuning constraint:

$$| t(i,j) - t(i,j') | \geq 1 + \delta$$

for all rows i and for all columns j and j' that belong to different wavelengths.

The constraint $C1$ stems from the fact that a wavelength can be used by one source destination pair at a time, and that each source can transmit at most one packet at a time. Constraint $C2$ includes the transmitter tuning delay and introduces an asymmetry between the rows and the columns of D.

While the above traffic model is described in terms of packet transmissions, it also applies to the case of connection-oriented networks such as IBM's Rainbow [4] and the MIT/AT&T/DEC Consortium's testbed network[1] [5]. Here the entries are viewed as sessions to be established between the nodes with the possibility of multiple connections per node. This point of view will be elaborated in Section 6.

The random traffic matrix model just described is more general than that considered by previous work on packet scheduling in networks. The case $p = 1$ corresponds to all-to-all packet transmission scenario considered by Pieris and Sasaki [2] and Aggarwal et.al. [3]. Here every node has exactly one packet to transmit to every other node. Pieris and Sasaki consider the case of tunable transmitters and fixed-tuned receivers and assume δ is a non-negative integer. They make the fundamental observation that for large tuning delays there exists an optimal number of wavelengths which optimally balances the wavelength concurrency with tuning delay. It is shown that any schedule should take a time of at least $N\sqrt{\delta}$ ($\delta \geq 1$), and two schedules are provided with clearance times of $N(\sqrt{\delta}+1)$ and $2N\sqrt{\delta}$ respectively. We will adopt one of these schedules in Section 3 to provide an upper bound in the case of random traffic. In this work, we also improve their lower bound to the clearance time, and generalize it for random traffic. Another difference in this paper is that we view the number of wavelengths as fixed and try to achieve optimal scheduling for all W.

Aggarwal et.al. [3] consider the case of tunable transmitters and tunable receivers and provide a lower bound of $N\sqrt{\delta/12}$ to the clearance time of an all-to-all transmission schedule, and an upper bound of $N(\sqrt{\delta}+0.5)$. These results assume $\delta \gg 1$, whereas we will be interested in the case $\delta \leq 1$.

The special case of $W = N$ wavelengths with an arbitrary traffic matrix D corresponds to the well-known scheduling problem in Satellite Switched Time Division Multiple Access (SS-TDMA). In SS-TDMA, d_{ij} is the number of time slots per frame that is needed for the source-destination pair (i,j), and the schedule corresponds to the design of a frame[2]. The N wavelengths are analogous to spatial diversity of N ground stations, and tuning corresponds to a reconfiguration of the on-board satellite switch [6]. However, there are important differences between optical WDM and SS-TDMA. First, changing the switch configuration corresponds to all users tuning simultaneously which is not a requirement in optical WDM. For instance, the scheduling algorithm that will be presented in Section 4 uses non-simultaneous tuning to achieve near-optimal clearance time. In SS-TDMA, δ typically is small and therefore the emphasis has been on minimizing the total transmission time [6, 7]. In this case polynomial-time algorithms for generating the optimal schedule are known. Gopal and Wong consider the other extreme $\delta \gg 1$, and show that finding the optimal SS-TDMA schedule for a given traffic matrix is NP-complete even in special cases [8]. In this paper, our interest is in pursuing average-case optimality when the traffic matrix is generated by a probability distribution.

3 Lower and Upper Bounds on the Clearance Time

As explained in the previous section, the traffic matrix D contains the packets to be transmitted in the current schedule. From the traffic matrix we define the $N \times W$ *collapsed traffic matrix* C as

$$c_{ij} = \sum_{k=(j-1)K+1}^{jK} d_{ik} \quad 1 \leq i \leq N, \;\; 1 \leq j \leq W$$

where $K = N/W$ is the number of receivers per wavelength. Thus c_{ij} represents the total number of packets from transmitter T_i to the receivers using wavelength λ_j. These entries are statistically independent and are Bin($N/W, p$) (Binomially distributed with parameters N/W and p).

Let us temporarily neglect the tuning delay, i.e., set $\delta = 0$. In this case, the results of [6] can be applied to determine the clearance time of the optimal schedule.

[1] These networks allow tunability at receivers unlike our model.

[2] This is analogous to our circuit-switched model in Section 6.

First, observe that for a given C the schedule will take a time at least

$$c_m \triangleq \max\left(\sum_j c_{ij}, \sum_i c_{ij}\right) .$$

c_m is the maximal line (row or column) sum of the matrix C. Second, Hall's theorem on Systems of Distinct Representations (SDRs) can be applied to express C as a sum of exactly c_m permutation matrices[3]. Since each such permutation matrix corresponds to conflict-free transmission of packets in C, the resulting schedule has an optimal clearance time of c_m. Polynomial-time algorithms to find optimal schedules are known [6].

When the collapsed traffic matrix C is generated according to the probability distribution described earlier, the optimal expected clearance time with $\delta = 0$, $\bar{T}^*(\delta = 0, p)$, is found as the expected value of the maximal line sum:

$$\bar{T}^*(\delta = 0, p) = E(c_m) \geq \max(E(c_r), E(c_c))$$

where c_r and c_c are the maximal row and column sums respectively, and the lower bound follows from the convexity of $\max(x, y)$ and Jensen's inequality [9].

Now let us return to the case of nonzero tuning delays and consider a simple idea that we will return to in the sequel. Suppose each packet is "padded" by δ time units to allow possible (and wasteful) tuning after each packet transmission. This padding results in a schedule that takes an average time of $\bar{T}^*(\delta = 0, p)(1 + \delta)$ to complete. Thus the optimal schedule has an average clearance time bounded by

$$(1 + \delta)E(c_m) \geq \bar{T}^*(\delta, p) \geq E(c_m) . \qquad (1)$$

For $\delta \leq 1$, the upper and lower bounds are within a factor of 2. One of the goals in the subsequent sections is to show that this potentially 100% inefficiency can be avoided by a more efficient schedule. Note that this is not an additional requirement, some form of scheduling is always necessary to avoid conflicts even without tuning delays.

A tighter lower bound than the one in (1) can be obtained by the following argument. Consider the ith row of C, and suppose there is a total of N_i packets distributed over K_i columns. Then transmitter T_i must spend a time N_i transmitting its packets and a time $K_i\delta$ tuning to K_i different wavelengths[4], hence the clearance time must be at least

$$T \geq N_i + K_i\delta \qquad i = 1, 2, \ldots, N$$

which implies

$$\bar{T}^*(\delta, p) \geq E\left[\max_{1 \leq i \leq N}(N_i + K_i\delta)\right] . \qquad (2)$$

[3] A permutation matrix is a 0-1 matrix with at most one nonzero entry per row and per column.

[4] The initial tuning of transmitters to appropriate wavelengths at the beginning of the schedule is included in the bound.

The N_i are i.i.d. Bin(N,p), while K_i are i.i.d. Bin(W, $[1-(1-p)^{N/W}]$). (The latter distribution follows from the fact that the occupancy events of W wavelengths are statistically independent. For $c_{ij} = 0$, all N/W corresponding entries must be zero.) If $W = N$, then $K_i = N_i$ and

$$\bar{T}^*(\delta, p) \geq (1 + \delta)E\left[\max_{1 \leq i \leq N} N_i\right] .$$

On the other hand, if $W < N$, N_i and K_i are correlated, obtaining the expectation in (2) appears to be a difficult task. In this case, we will use

$$\begin{aligned} \bar{T}^*(\delta, p) &\geq E\left[\max_{1 \leq i \leq N}(N_i + K_i\delta)\right] \\ &\geq E\left[\max_{1 \leq i \leq N} N_i\right] + E[K_{i^*}]\delta \\ &\geq E[c_r] + W[1 - (1 - p)^{N/W}]\delta \qquad (3) \end{aligned}$$

where K_{i^*} is the number of occupied wavelength groups in the row with maximal row sum. In the second line above, we have weakened the bound by considering the row that achieves maximal N_i, and in the third line we have used the fact that the maximally crowded row will have a larger expected wavelength occupancy than a typical row.

It is useful to define the expected value of the maximum of a set of i.i.d. Binomial random variables. Let X_i, $1 \leq i \leq L$, be i.i.d. Bin(M,ϵ), and define

$$f(M, \epsilon, L) \triangleq E\left[\max_{1 \leq i \leq L} X_i\right] .$$

It is easy to obtain $f(M, \epsilon, L)$ as

$$f(M, \epsilon, L) = M - \sum_{j=1}^{M}\left[\sum_{n=0}^{j-1}\binom{M}{n}\epsilon^n(1 - \epsilon)^{M-n}\right]^L .$$

It is also possible to obtain an approximation to $f(M, \epsilon, L)$, when M and L are both large, through a Chernoff bound as

$$f(M, \epsilon, L) \simeq M\epsilon + 2\sqrt{M\epsilon(1 - \epsilon)\ln L} .$$

Using this definition, we have from (3)

$$\bar{T}^*(\delta, p) \geq f(N, p, N) + W\left[1 - (1 - p)^{N/W}\right]\delta$$

for $W < N$, and

$$\bar{T}^*(\delta, p) \geq f(N, p, N)(1 + \delta)$$

for $W = N$. We will refer to this bound as the row lower bound. One can also obtain a column lower bound using the constraint that no more than one entry of a column of C can be transmitted simultaneously. Hence the schedule will take a time of at least the maximal column sum:

$$\bar{T}^*(\delta, p) \geq E(c_c) = f\left(\frac{N^2}{W}, p, W\right) .$$

Combining the row and column bounds, we have for $W < N$

$$\bar{T}^*(\delta,p) \geq \max\{f(N,p,N)+\bar{W}\delta, f(N^2/W,p,W)\} \quad (4)$$

with $\bar{W} = W[1-(1-p)^{N/W}]$, while for $W = N$

$$\bar{T}^*(\delta,p) \geq f(N,p,N)(1+\delta) \quad (5)$$

as the row bound is uniformly tighter than the column bound with N wavelengths.

The bounds in (4) and (5) will prove to be very important as we will find a schedule which achieves an average clearance time that is very close to the bounds.

At this point let us consider the special case of $p = 1$ in some detail. This is the all-to-all broadcast scenario that was analyzed by Pieris and Sasaki in [2]. The lower bound in (4)-(5) simplifies to

$$\bar{T}^*(\delta,p=1) \geq \max(N+W\delta, N^2/W) \quad (6)$$

for all W. This is a better lower bound than that given in [2] as $\max(W\delta, N^2/W)$. The value of W which minimizes the lower bound (6) is

$$W^*(\delta,p=1) = \frac{2N}{1+\sqrt{1+4\delta}}$$

which decreases from N to $0.62N$ as δ increases from 0 to 1. For $\delta \gg 1$, $W^* \simeq N/\sqrt{\delta}$, the same value obtained in [2]. Since our analytical development assumes that both W and N/W are integers, we must consider the cases $W = N/2$ and $W = N$ as potential minima. The lower bounds are $N(1+\delta)$ for $W = N$, and $2N$ for $W = N/2$ when $0 \leq \delta \leq 1$. Thus $W = N$ is the optimal wavelength setting. The same conclusion holds even if we allowed N/W to be non-integer: the schedule will take at least $2N$ time units when $N/2 \leq W < N$, since at least one channel will have $\lceil N/W \rceil = 2$ receivers. Therefore $W = N$ is the unique optimum for $p = 1$ and $\delta < 1$. (For $\delta = 1$, the same lower bound of $2N$ is achieved for $N/2 \leq W \leq N$.)

Figure 1 shows the lower bound for $p = 1$ and $N = 100$ as a function of W with $\delta = 0.1$ and $\delta = 1$. The conclusions reached above can also be confirmed from the curves. Note that the data points on these curves are computed only for integer N/W and are connected by straight lines only for presentation.

The lower bound is minimized either by $W = N$ or by $W = N/2$ for $p < 1$ as seen from Figure 2 which shows the normalized clearance time for $p = 0.1$ and $p = 0.5$. In fact, it can be shown from the definition of $f(\cdot)$ that $f(2N,p,N/2) \geq f(N,p,N) + Np$. That W^* is either N or $N/2$ for any value of p easily follows from this result.

While a lower bound to the average clearance time provides a limit to the efficiency of scheduling, it is necessary to assess its tightness before any conclusions can be drawn from such a bound. Therefore we would like to obtain an upper bound to $\bar{T}^*(\delta,p)$, preferably by using a simple scheduling algorithm. In this section, we consider a suboptimal schedule given in [2] for the case $p = 1$ and $\delta \gg 1$. This schedule, which will be called Pieris-Sasaki schedule, performs within a factor 2 of the lower bound in [2] when W is optimized. We generalize it for $p < 1$, and analyze it to obtain an upper bound to the clearance time. In the next section, we will present a better upper bound through a different scheduling algorithm.

Pieris-Sasaki schedule groups the transmitters into W groups $G_1, G_2, \ldots, G_W$ where G_i consists of transmitters $T_{(i-1)K+1}, T_{(i-1)K+2}, \ldots, T_{iK}$ $(K = N/W)$. Initially the transmitters in G_i are tuned to λ_i and sequentially transmit their packets to the receivers in λ_i. When all the transmissions in all W wavelengths are completed, the transmitters in G_i tune to $\lambda_{i\oplus 1}$ ($\oplus$ denotes modulo W addition defined over $\{1,2,\ldots,W\}$), and transmit their packets in $\lambda_{i\oplus 1}$. The schedule completes after W such tuning phases. The number of packets transmitted by a transmitter group on a wavelength is $\mathrm{Bin}(N^2/W^2, p)$. Since each phase of the algorithm involves a tuning and last until all of the W concurrent transmissions are completed, the average clearance time $\bar{T}_{PS}(\delta,p)$ is given by

$$\bar{T}_{PS}(\delta,p) = W\left(f\left(N^2/W^2,p,W\right)+\delta\right) \geq \bar{T}^*(\delta,p)\,. \quad (7)$$

This upper bound to the optimal clearance time is also shown in Figures 1 and 2 for various values of p and δ. It is seen that for $p = 1$ and $\delta \leq 1$, the upper bound and the lower bound are extremely close. In fact, in this case we have from (6) and (7)

$$K + \frac{\delta}{K} \geq \frac{\bar{T}^*(\delta,p=1)}{N} \geq \max\left(1+\frac{\delta}{K}, K\right)$$

where $K = N/W$. For $\delta \leq 1$, the ratio ρ of the upper bound to the lower bound is given by

$$\rho = \begin{cases} 1 & K = 1 \\ 1+\delta/K^2 & K \geq 2\,. \end{cases}$$

Thus the Pieris-Sasaki algorithm is optimal[5] when $p = 1$, $\delta \leq 1$ and $W = N$, and is within a factor of $1 + (W/N)^2\delta$ of the optimal performance when $W \leq N/2$. That is, for high load ($p \simeq 1$), and a limited number of wavelengths ($W \ll N$), this algorithm is nearly optimal.

However, this near-optimality is no longer attained when $p < 1$ as Figure 2 indicates. In fact, when $p \ll 1$ and W is large, the upper bound indicates an inferior performance than the lower bound. There are two potential sources for this discrepancy. First, the synchronized nature of Pieris-Sasaki algorithm is well suited for a full traffic matrix, but is less efficient for a sparse matrix. Second, the lower bound may not be tight for $p < 1$. In the next section, we will provide a different scheduling algorithm which achieves near-optimal results for all values of p. This will show that the lower bound is, in fact, tight.

[5]The optimal performance for $W = N$ and $p = 1$ could also be achieved by a schedule that employs padding.

4 Single Reservation Scheduling

For a random traffic matrix D, it is intuitively clear that an efficient scheduling algorithm must exploit the traffic information so as to avoid unnecessary tuning. We now describe an algorithm that is based on reserving idle transmitters to wavelengths that are close to completing the service of the active transmitters. The concurrency in the tuning and transmission events improves the efficiency. We call the algorithm the *Single Reservation Algorithm (SRA)* as it reserves at most one transmitter for each wavelength. The basic principle of SRA is to reserve an idle and unreserved transmitter for a wavelength λ_i, δ time units before the transmitter currently active in λ_i completes its transmission. Thus for $\delta < 1$, the reserved transmitter can tune to λ_i *just-in-time* to eliminate any dead-time in λ_i.

A reservation based real-time protocol, the MaTPi protocol, has been recently proposed for optical WDM with integer δ [10]. In MaTPi, a node with data reserves a time slot which is δ time slots in the future. SRA uses a different philosophy; it utilizes the global traffic information to reserve idle nodes, as explained below.

In the initial phase of the SRA, the transmitter which has the maximum number of packets in λ_1 tunes to λ_1. Of the remaining transmitters, the one with most packets in λ_2 tunes to λ_2 and so on. These transmitters then sequentially transmit all their packets in the tuned wavelength. The rest of the transmitters remain initially idle and unreserved. When the remaining transmission time in λ_i falls below δ, an idle transmitter with packets on λ_i is reserved. The reserved transmitter starts tuning to λ_i and transmits its packets as soon as its tuning is complete. The transmitter which was previously using λ_i joins the idle pool and becomes available for reservation. If there are more than one idle and unreserved transmitters that have packets for λ_i, the one with largest demand for λ_i is reserved. Conversely, if there are no idle transmitters with traffic on λ_i, λ_i remains unreserved and potentially unused until a reservation can be made. In the case of simultaneous reservations on two or more wavelengths, priority is given to the wavelength with lowest index[6]. The algorithm continues until the matrix is cleared.

It is difficult to model the performance of SRA analytically. Therefore we will resort to Monte Carlo simulations for evaluating the average clearance time. In Figures 1 and 2 we show the average clearance time of SRA for some sample values of δ and p. As suggested by these figures, we have observed that for $\delta \leq 1$ and $0 < p \leq 1$, the SRA clearance time is very close to the lower bound. We have simulated the cases $\delta = 0.1, 0.5, 1$, $p = 0.1, 0.5, 1$, and $W = 1, 2, 4, 5, 10, 20, 25, 50, 100$, for $N = 100$. The worst case discrepancy between the simulation and the lower bound was 30% and occurred when $\delta = p = 1$, and $W = 50$. In most cases the simulation performance was within 5% of the lower bound.

Note that the Single Reservation Algorithm will not be efficient for $\delta > 1$ as the reserved transmitters may not complete their tunings in time to avoid "dead times". This is particularly true when W is large and p is small, since in this case the idle pool will be small. Reserving multiple transmitters may be a better option to implement. Our goal in presenting SRA is not to provide a best possible scheduling algorithm; rather, it is to demonstrate that the lower bound presented earlier analytically captures the fundamental effects of tuning delay on performance in a random traffic situation. This way, one can reliably use the lower bound to derive insights into the effect of tuning delays on scheduling performance.

First, let us consider the all-to-all transmission case $p = 1$. Assuming the lower bound in (6) can be achieved, the scheduling penalty due to tuning delay is given by

$$A(\delta, p=1) \triangleq \frac{T^*(\delta, p=1)}{T^*(\delta=0, p=1)} = \begin{cases} 1+\delta & W = N \\ 1 & W \leq N/2 \end{cases}. \tag{8}$$

Thus if there are no constraints on the number of wavelengths, one would use $W = N$ wavelengths to minimize the clearance time and suffer a penalty factor of $1+\delta$. If $W \leq N/2$, as it would normally be the case in a large network, the clearance time would be larger than that with one wavelength per user; however there is no penalty due to the tuning delay. This conclusion has an important practical ramification: *In a bandwidth-limited network, the packet length can be made as short as the tuning time without any tuning penalty.*

Another relevant issue is the efficiency of the optimal schedule relative to a schedule with padding. Since padding achieves a clearance time of $T^*(\delta = 0, p = 1)(1+\delta)$, we define the improvement with the optimal schedule with $p = 1$ as

$$I(\delta, p=1) \triangleq \frac{(1+\delta)T^*(\delta=0, p=1)}{T^*(\delta, p=1)} = \frac{1+\delta}{A(\delta, p=1)} = \begin{cases} 1 & W = N \\ 1+\delta & W \leq N/2 \end{cases} \tag{9}$$

which means that with $W \leq N/2$, optimal scheduling will gain a factor of $1+\delta$ over padding. But if $W = N$, padding is optimal.

For $p < 1$ we define the scheduling penalty $A(\delta, p)$ and optimality improvement $I(\delta, p)$ similarly. The average clearance time with no tuning delay $\bar{T}^*(\delta = 0, p)$ is given by the expected value of the maximal line sum in C. Recall that the N row sums of C are the i.i.d. $\text{Bin}(N, p)$ while the W column sums are i.i.d. $\text{Bin}(N^2/W, p)$. However, since rows and columns are not independent, the exact computation of the average critical sum is intractable. Therefore we use the Jensen inequality to obtain the lower bound:

$$\bar{T}^*(\delta = 0, p) \geq \max\left(f(N, p, N), f(N^2/W, p, W)\right)$$

[6] This fixed priority implies that the rightmost columns of the traffic matrix are cleared later on the average. Fairness can be achieved by rotating the priority among the wavelengths.

to conservatively overestimate the penalty $A(\delta, p)$ and underestimate the improvement $I(\delta, p)$. The lower bound is expected to be uniformly tight since for $W < N$, the maximal line sum is a column sum with high probability and for $W = N$, the relative inaccuracy in the bound is at most $1 - f(N, p, N)/f(N, p, 2N)$ which is small when Np is large. Our numerical calculations show that the penalty and the improvement with respect to padding are still given by Equations (8) and (9) respectively for all values of p when $\delta \leq 1$. Therefore, in a bandwidth-limited network with $W \leq N/2$, there is no penalty due to tuning delay, and an improvement of $1 + \delta$ over padding. When $W = N$, the tuning delay penalty approaches $1 + \delta$ while improvement over padding vanishes ($I(\delta, p)$ becomes unity).

These conclusions are summarized in Figure 3 which shows the optimal clearance time $\bar{T}^*(\delta, p)$, the clearance time without any tuning delay $\bar{T}^*(\delta = 0, p)$, and the clearance time with padding $(1 + \delta)\bar{T}^*(\delta = 0, p)$ as functions of W. (We have used $f(N, p, L) \approx Np$ in the figure for clarity of presentation.)

An important implication of these results is that, in bandwidth-limited networks, introducing tunability at the receivers as well as the transmitters cannot improve the clearance time, as the performance is limited by the bandwidth and not by the tuning delay.

5 Real-Time Traffic

The off-line scheduling approach we have considered in the preceding sections must be embedded in a network with real-time packet traffic. In this setting, the packet stream will be stored in N^2 buffers, one per source-destination pair, and a single head-of-line packet will be cleared from each non-empty buffer per schedule[7].

This raises the question as to whether such a network reaches equilibrium and whether the equilibrium is stable. Let the packet arrival rate (normalized with respect to packet duration) per source-destination pair be λ. If λ is too large, the number of packets that arrive during a schedule will exceed the amount that can be cleared in the next schedule, the buffer occupancies will grow without bound, and the traffic input to the schedule converges to a full matrix ($p = 1$). Thus, there is a certain traffic capacity $C_0(\delta)$ beyond which input traffic rates cannot be supported. However, it is not clear that rates $\lambda \leq C_0(\delta)$ can be supported with stability, i.e., with a single distribution p on the traffic matrix and a steady state buffer occupancy. We now proceed to show that this is the case under optimal scheduling.

The aggregate input rate to the network is $N^2\lambda$ while the aggregate output rate is $N^2 p/\bar{T}^*(\delta, p)$. Thus, flow conservation dictates that in the steady state (if it exists)

$$\lambda = \frac{p}{\bar{T}^*(\delta, p)} . \qquad (10)$$

[7]Of course, we could allow traffic matrices with non-binary entries in scheduling.

For a given throughput λ, (10) must be solved for p. Let $g(p) \triangleq p/\bar{T}^*(\delta, p)$. For a single equilibrium point $g(p)$ must be monotonic on $(0, 1]$. That $g(p) \geq g(p/n)$ for integer n can be seen as follows. A suboptimal way to schedule a matrix C with probability distribution Ber(p) is to randomly decompose C into n matrices $C_1, C_2, \ldots, C_n$ each with probability distribution Ber(p/n) and to sequentially schedule these n matrices. Thus $\bar{T}^*(\delta, p) \leq n\bar{T}^*(\delta, p/n)$ which implies $g(p) \geq g(p/n)$. In general $g(p)$ is monotonically increasing with p.

As a result of this monotonicity, an equilibrium point exists as long as

$$\lambda \leq C_0(\delta) \triangleq \frac{1}{\bar{T}^*(\delta, p = 1)} = \frac{W}{N^2} \min\left(1, \frac{K^2}{K + \delta}\right)$$

where $C_0(\delta)$ is the scheduling capacity per source-destination pair. The overall *scheduling capacity* of the network is

$$\begin{aligned} C(\delta) &= N^2 C_0(\delta) = W \min\left(1, \frac{K^2}{K + \delta}\right) \\ &= \begin{cases} N/(1 + \delta) & W = N \\ W & W \leq N/2 \end{cases} \end{aligned}$$

where the last equality is valid for $\delta \leq 1$. With a large number of wavelengths the capacity is reduced by a factor $1 + \delta$. For $W \leq N/2$, the capacity is not affected by tuning delay (provided that $\delta \leq 1$).

Another implication of the monotonicity of $g(p)$ is that the equilibrium point is stable. A graphical perturbation analysis shows that small fluctuations in the input rate λ will be compensated by changes in p, rendering the equilibrium point (p, λ) stable.

The queueing delay of the underlying system can be found through a standard vacation model [11]. With Poisson packet arrivals, each source-destination pair sees an M/D/1 queueing system where the server serves a single packet from the buffer and takes a vacation of random duration T. In this limited-service server-vacation model, one needs the first two moments of the schedule clearance time in order to evaluate the average delay experienced by a packet. The first moment can be accurately approximated by the lower bound obtained in this paper. Similar techniques can also be developed for the second moment. Alternatively, the second moment can be estimated using numerical simulation.

6 Connection-Oriented Networks

We have found it convenient to describe the scheduling framework in this paper in terms of a connectionless network. The model and the results are equally applicable to a connection-oriented network with sessions. In this context the entry $d_{ij} = 1$ of the traffic matrix corresponds to a session to be established between transmitter T_i and receiver R_j. (Multiple connections between a transmitter and several receivers are allowed.) Since the topology is a broadcast star, a circuit connection corresponds to a wavelength and a time slot assignment in a Time Division

Multiplexed (TDM) frame. Since the wavelengths are preassigned to the receivers in a fixed manner, the scheduling problem becomes one of time slot assignment in the TDM frame subject to the same constraints in the packet switching case. (see Section 2). The goal is to find the frame of shortest duration that satisfies the tuning delay and transceiver constraints. This is equivalent to finding the minimum clearance time schedule when the packet duration is replaced by the duration of a time slot. Our results in Section 4 indicate that the time slot can be made as short as the tuning delay δ without any tuning penalty in a bandwidth-limited network.

One major difference between connectionless and connection-oriented traffic is the frequency at which the scheduling algorithm is executed. With packet transmissions, the schedule lasts only for one clearance time, so the execution time of the algorithm is important. On the other hand, with circuits the schedule is in effect for a longer time scale. The current time slot assignment will remain unchanged until one of the current sessions complete or until a new session request is received. Therefore, the time complexity of scheduling algorithms is less critical with connection-oriented traffic, a near-optimal algorithm may be attractive even with a high time complexity.

7 Conclusions

We have considered scheduling of random traffic in an optical WDM network in order to assess the effect of tuning delay on the performance. We first presented a lower bound on clearance time as a function of the network size, the number of available wavelengths, the statistical distribution of the traffic matrix, and the tuning delay δ. We then established that this bound captures the fundamental limits to the scheduling efficiency through numerical simulation of a scheduling algorithm based on advance reservation of the transmitters. As a consequence, we have shown that for tuning delays that are less than or equal to packet duration, the penalty due to tuning delay depends on the number of wavelengths, W, in the network. If there are as many wavelengths as there are users, the tuning delay δ causes an increase in the average clearance time by a factor $1 + \delta$. This optimal performance can be achieved by padding δ to each packet in the traffic matrix and by using well-known techniques for decomposing the given traffic matrix into a minimal number of permutation matrices. When the network has a limited number of available wavelengths, the practical situation for large N, we have reached a somewhat surprising conclusion that there is no penalty in the clearance time through optimal scheduling as long as $\delta \leq 1$. This means that through optimal scheduling one can eliminate the need for very rapidly tunable optical devices for packet switching. For instance, with a 100 Mbps transmission rate and 1000 bit packets, a 10 μs tuning delay is sufficient. This conclusion is in contrast with the previous conjecture that tuning delays must be very small relative to the packet size [12, pp. 274-275]. We have also shown that tunability only at one end (transmitters) is sufficient for attaining the best possible performance in bandwidth-limited networks.

The results in this paper apply not only to connectionless packet-switched traffic, but also to connection-oriented traffic where the scheduling is used to assign time slots to different sessions. Multiple sessions per node can be supported using the scheduling approach with slot lengths comparable to tuning delays.

We have also shown that scheduling results in a single stable equilibrium point with real-time traffic. A scheduling capacity has been introduced as the maximum traffic that can be carried by the network. It has been shown that the scheduling capacity is not affected by tuning delays for a network with limited number of wavelengths ($W \leq N/2$).

Large tuning delay of tunable devices has been viewed as a major impediment in establishing packet-switched and multi-session circuit-switched all-optical networks. It is likely that future networks will have a limited number of wavelengths relative to the number of nodes and a tuning delay comparable to the packet delay. The results of this paper indicate that efficient scheduling algorithms can provide near-optimal performance for a variety of services in such networks.

Acknowledgement

This research was supported by NSF under grant NCR-9309574.

References

[1] P. E. Green, *Fiber Optic Networks*. Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[2] G. R. Pieris and G. H. Sasaki, "Scheduling transmissions in WDM broadcast-and-select networks," *IEEE/ACM Transactions on Networking*, vol. 2, pp. 105–110, April 1994.

[3] A. Aggarwal, A. Bar-Noy, D. Coppersmith, R. Ramaswami, B. Schieber, and M. Sudan, "Efficient routing and scheduling algorithms for optical networks," Tech. Rep. RC 18967, IBM Research Report, June 1993.

[4] F. J. Janiello, R. Ramaswami, and D. G. Steinberg, "A prototype circuit-switched multi-wavelength optical metropolitan-area network," *Journal of Lightwave Technology*, vol. 11, pp. 777–782, May/June 1993.

[5] S. B. Alexander et. al., "A precompetitive consortium on wide-band all-optical networks," *Journal of Lightwave Technology*, vol. 11, pp. 714–735, May/June 1993.

[6] T. Inukai, "An efficient SS/TDMA time slot assignment algorithm," *IEEE Transactions on Communications*, vol. COM-27, pp. 1449–1455, October 1979.

[7] H. A. Choi and S. L. Hakimi, "Data transfers in networks," *Algorithmica*, no. 3, pp. 223–245, 1988.

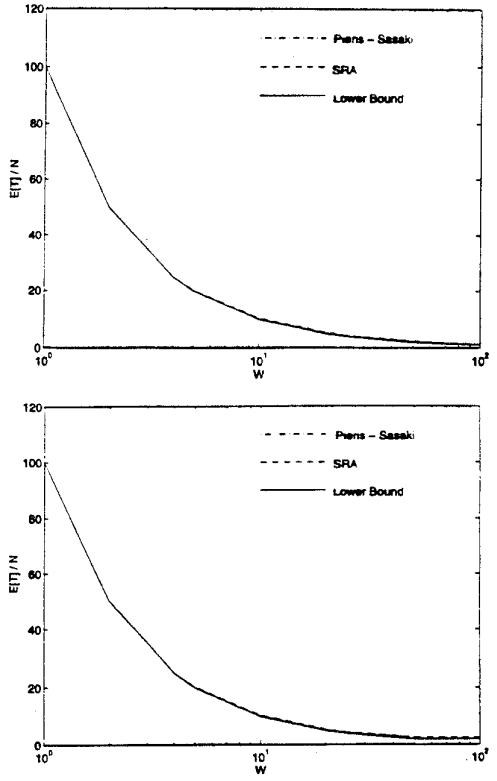

Figure 1: The lower bound, the Pieris-Sasaki upper bound, and the SRA performance for the clearance time with $N = 100$ and $p = 1$, (a) $\delta = 0.1$, (b) $\delta = 1$.

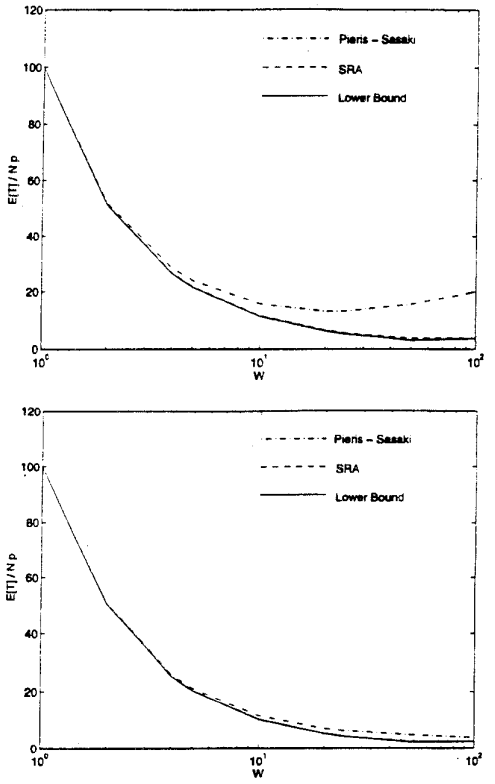

Figure 2: Same as Fig. 1 with $p < 1$ and $\delta = 1$, (a) $p = 0.1$, (b) $p = 0.5$.

[8] I. S. Gopal and C. K. Wong, "Minimizing the number of switchings in an SS/TDMA system," *IEEE Transactions on Communications*, vol. COM-33, pp. 497–501, June 1985.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, 1991.

[10] S. Tripandapani, J. S. Meditch, and A. K. Somani, "The MaTPi protocol: Masking tuning times through pipelining in WDM optical networks," in *Proceedings of IEEE Infocom '94*, 1994.

[11] D. P. Bertsekas and R. G. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice Hall, second ed., 1992.

[12] A. S. Acampora, *An Introduction to Broadband Networks*. New York: Plenum Press, 1994.

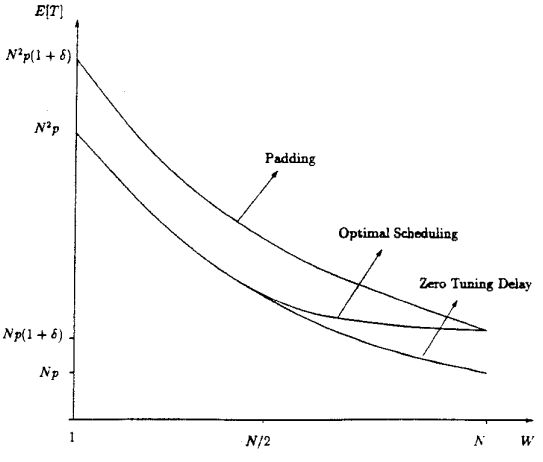

Figure 3: A comparison of the clearance times of the optimal schedule, the $\delta = 0$ schedule, and the schedule with padding.