<table>
<thead>
<tr>
<th>Title</th>
<th>Minimum delay scheduling in scalable hybrid electronic/optical packet switches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Wu, B; Yeung, KL</td>
</tr>
<tr>
<td>Citation</td>
<td>Proceedings of the Global Telecommunications Conference, 2006 (GLOBECOM 2006), San Francisco, CA, USA, 27 November - 1 December 2006</td>
</tr>
<tr>
<td>Issued Date</td>
<td>2006</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10722/99457">http://hdl.handle.net/10722/99457</a></td>
</tr>
<tr>
<td>Rights</td>
<td>©2006 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.</td>
</tr>
</tbody>
</table>
Minimum Delay Scheduling in Scalable Hybrid Electronic/Optical Packet Switches

Bin Wu and Kwan L. Yeung
Dept. of Electrical and Electronic Engineering
The University of Hong Kong
Pokfulam, Hong Kong
E-mail: {binwu, kyeung}@eee.hku.hk

Abstract—A hybrid electronic/optical packet switch consists of electronically buffered line-cards interconnected by an optical switch fabric. It provides a scalable switch architecture for next generation high-speed routers. Due to the non-negligible switch reconfiguration overhead, many packet scheduling algorithms are invented to ensure performance guaranteed switching (i.e. 100% throughput with bounded packet delay), at the cost of speedup. In particular, minimum delay performance can be achieved if an algorithm can always find a schedule of no more than $N$ configurations for any input traffic matrix, where $N$ is the switch size. Various minimum delay scheduling algorithms (MIN, $a^z$-SCALE and QLEF) are proposed. Among them, QLEF requires the lowest speedup bound. In this paper, we show that the existing speedup bound for QLEF is not tight enough. A new bound which is 10% lower than the existing one is derived.

Keywords—Minimum delay scheduling; performance guaranteed switching; reconfiguration overhead; speedup bound.

I. INTRODUCTION

The explosion of Internet traffic has led to ever-increasing demands for larger bandwidth and higher port density in the next generation routers. At present, most Internet backbone routers are based on a single-rack solution using a switched backplane. Typically, a standard telecommunication rack is of size 19 inches in width and 7 feet in height. It can accommodate 14-16 line-cards, with aggregate capacity up to 160 Gb/s. To further increase the capacity, multi-rack solution [1] is adopted, where line-cards in different racks are interconnected to/from the central electronic switch fabric by fibers. This architecture defines the 4th generation router, which can offer an aggregate capacity up to 10 Tb/s [2]. Since electronic switch fabric is used, O/E/O conversions are necessary at the central switch rack. As data is handled in the electronic domain, power consumption becomes the key constraint [3]. To solve these issues, the 5th generation router architecture is adopted as shown in Fig. 1, where a hybrid electronic/optical switch architecture is adopted. Compared to the 4th generation router, the central electronic switch fabric is replaced by an optical one. This not only removes the O/E/O conversions from the switch rack, but also reduces its power consumption. Following this way, the aggregate capacity can be up to 100 Tb/s [3-4].

However, the optical switch fabric needs a non-negligible amount of time to change its configurations, known as reconfiguration overhead. During this period, no packet can be transmitted across the switch. Reconfiguration overhead is due to three factors [5]. First, the optical fabric needs time to change its interconnection pattern, which can range from 10 ns to several milliseconds depending on the technology adopted [6-9]. Second, time is required to resynchronize the optical transceivers and the switch fabric in each reconfiguration. Finally, because the arriving time of optical signals varies, the clock and its phase have to be aligned, and extra clock margin has to be considered in order to avoid data loss.

To achieve performance guaranteed switching (i.e. 100% throughput with bounded packet delay) [10-13], the switch fabric must run faster to compensate for both the reconfiguration overhead and the scheduling inefficiency. The required speedup $S$ is defined as the ratio of the internal packet transmission rate to the external line-rate ($S \leq 1$).

Assume time is slotted and each time slot can accommodate one fixed-size packet. Several scheduling algorithms are proposed to achieve performance guaranteed switching [10-13]. They all adopt the same four-stage scheduling procedure as shown in Fig. 2. Stage 1 is for traffic accumulation. An $N \times N$ traffic matrix $C(T) = \{c_{ij}\}$ is obtained at the input buffers every $T$ time slots, where $N$ is the switch size. Each entry $c_{ij}$ denotes the number of packets arrived at input $i$ and destined to output $j$. As a common assumption [10-13], the entries in each line (i.e. row or column) of $C(T)$ sum to at most $T$. In Stage 2, a scheduling algorithm generates a schedule consisting of (at most) $N_S$ switch configurations in $H$ time slots. Each
configuration is denoted by a permutation matrix \( P_n = \{p^{(n)}_{ij}\} \) \((N \geq n \geq 1)\), where \( p^{(n)}_{ij} = 1 \) means that input port \( i \) is connected to output port \( j \) (In this case, we also say that \( P_n \) covers entry \((i, j)\)). A weight \( \phi_n \) is assigned to each \( P_n \) indicating the number of slots that \( P_n \) should be kept for packet transmission. The set of \( N \) configurations generated must cover \( C(T) \), i.e., \( \sum_{n=1}^{N} N \phi_n p^{(n)}_{ij} \geq c_{ij} \) for any \( i, j \in \{1, \ldots, N\} \). Then \( \sum_{n=1}^{N} \phi_n \) is the number of slots required to transmit all the packets in \( C(T) \). Let each reconfiguration take an overhead of \( \delta \) time slots. Accordingly, sending \( C(T) \) requires \( \delta N + \sum_{n=1}^{N} \phi_n \) \( \delta \) time slots. This is generally larger than the traffic accumulation time \( T \). Without speedup, 100% throughput is not possible. Stage 3 is for actual packet transmission, where the switch fabric is reconfigured according to the \( N \) configurations. At a speedup of \( S \), the slot size for a single packet transmission in Stage 3 is shortened by \( S \) times. Then 100% throughput is ensured by having
\[
\delta N + \frac{1}{S} \sum_{n=1}^{N} \phi_n = T. \tag{1}
\]
The values of \( N \) and \( \sum_{n=1}^{N} \phi_n \) in (1) are determined by the scheduling algorithm. Note that the total reconfiguration overhead time \( \delta N \) cannot be reduced by speedup and thus \( T = \delta N \). Finally, Stage 4 takes another \( T \) time slots to send the packets onto the output lines from the output buffers.

Rearranging (1), we have
\[
S = \frac{T}{1 - \delta N} \sum_{n=1}^{N} \phi_n = S_{\text{reconfigure}} \times S_{\text{schedule}}, \tag{2}
\]
where \( S_{\text{reconfigure}} \) and \( S_{\text{schedule}} \) are defined as
\[
S_{\text{reconfigure}} = \frac{T}{T - \delta N} \tag{3}
\]
\[
S_{\text{schedule}} = \frac{1}{T} \sum_{n=1}^{N} \phi_n \tag{4}
\]
\( S_{\text{reconfigure}} \) is the speedup factor to compensate for the idle time caused by reconfigurations, whereas \( S_{\text{schedule}} \) is the speedup factor to compensate for the scheduling inefficiency.

In Fig. 2, the packet delay is bounded by \( 2T + H \) slots where \( T = \delta N \). With a smaller \( N \), \( T \) and thus the packet delay bound can be reduced. But \( N \) must be no less than \( N \). Otherwise, the \( N \) configurations are not sufficient to cover every entry in \( C(T) \) \([10-13]\). Accordingly, scheduling algorithms that only require \( N \) configurations are called minimum delay scheduling algorithms. When \( N = N \), from (1) \( T \) can be made as close to its lower bound \( \delta N \) as possible by minimizing \( \sum_{n=1}^{N} \phi_n / S \). Generally, we hope to minimize \( S \) for a given packet delay (or equivalently a given \( T \)). In minimum delay scheduling, this translates to minimizing \( S_{\text{schedule}} \) according to (1) \rightarrow (4).

Recently, several minimum delay scheduling algorithms \([\text{MIN 10, d-SCALE 12 and QLEF 13}]\) have been proposed to minimize \( S_{\text{schedule}} \). Among them, QLEF (Quasi Largest-Entry-First) requires the lowest speedup bound. In this paper, we derive a new speedup bound for QLEF, which is 10% lower than the one in \([13]\). Because the same scheduling problem is also involved in other communication networks, such as SS/TDMA \([14]\), TWIN \([15]\) and wireless sensor networks \([16]\), our result also contributes to those systems.

The rest of the paper is organized as follows. In Section II, we review the QLEF algorithm \([13]\). The new speedup bound is derived in Section III. We conclude the paper in Section IV.

QLEF ALGORITHM

\textbf{Input:} 
An \( N \times N \) matrix \( C(T) \) with maximum line sum not more than \( T \).

\textbf{Output:} 
\( N \) configurations \( P_1, \ldots, P_N \) and weights \( \phi_1, \ldots, \phi_N \).

\textbf{Step 1: Initialization:}
Set \( 0 \rightarrow n \). Initialize \( P_1, \ldots, P_N \) to all-zero matrices and the \( N \times N \) reference matrix \( R = (r_{ij}) \) to all 1's.

\textbf{Step 2: Determine the first "half" configurations \( P_{\text{even}} \).}
\begin{enumerate}
\item Un-shadow \( C(T) \) and \( R \). Set \( 1 \rightarrow w \).
\item Select the largest entry \( c_{ij} \) in the not-yet-shadowed part of \( C(T) \).
\end{enumerate}

If \( \delta = 1 \), set \( P_{\text{even}} \)'s weight \( \phi_{ij} = c_{ij} \) and \( w = 0 \). Shadow the corresponding lines in both \( C(T) \) and \( R \), and set \( c_{ij} \) and \( r_{ij} \) to 0. Set \( 1 \rightarrow p^{(n)}_{ij} \) where \( p^{(n)}_{ij} \) is the entry \((i, j)\) of \( P_{\text{even}} \). Repeat this step until \( N \) \((2n+1)\) largest entries are selected.

\textbf{Step 3: Determine the second "half" configurations \( P_{\text{odd}} \).}
\begin{enumerate}
\item Un-shadow \( C(T) \) and \( R \). Find the largest entry \( c_{ij} \) in \( C(T) \) and set \( c_{ij} \) as the weight for all the subsequent configurations.
\item Find a maximum-size matching in the bipartite graph \( G \) and set the corresponding entries of \( P_{\text{odd}} \) to 1. These entries to 0 in \( C(T) \) and \( R \), and then set \( n + 1 \rightarrow n \). Repeat this step until \( n = N \).
\end{enumerate}
by \( P_{n+1} \). If \( c_{ij} \) is shadowed (see Fig. 4a) in the construction of \( P_k \) \((n \geq k \geq 1)\), then \( P_k \) is called an \( s \)-configuration of \( c_{ij} \). Otherwise, \( P_k \) is called a \( g \)-configuration of \( c_{ij} \).

Fig. 4b shows the conceptual QLEF scheduling procedure. We use a “scheduling trace” to represent the trend of \( c_{ij} \) values covered. It is usually a wave rather than a monotonically decreasing curve, although QLEF always selects the largest entry in the not-yet-shadowed part. Due to the shadowing operation, a large entry may be shadowed by an \( s \)-configuration and thus can only be covered \textit{later} by another configuration.

For a particular configuration \( P_{n+1} \), its weight \( \phi_{n+1} \) also appears as an entry in \( C(T) \) and is the \textit{first} selected-entry in \( P_{n+1} \). Therefore, in the following discussion, we treat \( \phi_{n+1} \) as an entry in \( C(T) \) rather than a weight. Among the \( n \) configurations constructed before \( P_{n+1} \) (i.e. \( P_1 \sim P_n \)), we assume that \( \Delta \) of them are \( g \)-configurations of \( \phi_{n+1} \) and the other \( n-\Delta \) configurations are its \( s \)-configurations, as in Fig. 4b.

\[ \phi_{n+1} \leq \frac{T}{M} \]  
\[ \phi_{n+1} \leq \frac{T}{N} \]  

On the other hand, because \( \phi_{n+1} \) is shadowed by \( n-\Delta \) \( s \)-configurations, from Lemma 2 in [13], we have

\[ \phi_{n+1} \leq \frac{\max_{0 \leq \Delta \leq n} \left( \frac{T}{N}, \frac{T}{M} \right)}{\min_{0 \leq \Delta \leq N-1} \left( \frac{T}{N-\Delta-1}, \frac{T}{n-\Delta+1} \right)} \]  

Combining (5) and (6), we can bound \( \phi_{n+1} \) as follows:

Formula (7) indicates that no matter what the value of \( \Delta \) \((0 \leq \Delta \leq n)\), the bound always holds because we have taken the worst case into consideration (i.e. the \textit{max} function for \( \Delta \)).

For the remaining \( N - \left\lfloor \frac{N}{2} \right\rfloor + 1 \) configurations constructed in Step 3 in Fig. 3, QLEF uses a small constant as the weights. According to QLEF, this constant is not larger than any weight of the first \( \left\lfloor \frac{N}{2} \right\rfloor - 1 \) configurations (since the weights are monotonically decreasing as shown by the dashed line in Fig. 4b). Therefore, it can be bounded by the weight of the last one among the first \( \left\lfloor \frac{N}{2} \right\rfloor - 1 \) configurations. That is

\[ \phi_{n+1} \leq \frac{n+1}{N} \leq \phi \left( \frac{N-1}{2} \right) \]  

After the \( N \) weights \( \phi_{n+1} \) are bounded by (7) and (8), we can calculate \( S_{\text{schedule}} \) bound in (4). Note that the key is to count the \textit{minimum} total number of LEs (i.e. \( M \) in (7)).

\section{Speedup Bound Formulation}

For entry \( \phi_{n+1} \), we consider its \( \Delta \) \( g \)-configurations and the other \( n-\Delta \) \( s \)-configurations (see Fig. 4b). In QLEF, all the selected-entries in the \( g \)-configurations are LEs. On the other hand, each \( s \)-configuration must cover one LE in the same line as \( \phi_{n+1} \). However, in addition to this LE, the \( s \)-configuration may also cover other LEs in different lines. Assume that a set of consecutive \( s \)-configurations \( \{P_x\} \) shadow \( \phi_{n+1} \), and \( P_y \) is the first \( g \)-configuration after \( \{P_x\} \). From Lemma 1 in the Appendix, the number of LEs covered by each \( P_x \in \{P_x\} \) is at least half of the number of the selected-entries in \( P_y \).

In Fig. 4b, the \( \Delta \) \( g \)-configurations and the \( n-\Delta \) \( s \)-configurations may queue in any order. From Lemma 2 in the Appendix, in order to minimize \( M \), the \( n-\Delta \) \( s \)-configurations should \textit{consecutively} locate at either the very beginning or the very end of the configuration sequence \( P_1 \sim P_x \).

\textbf{Case 1:} The \( n-\Delta \) \( s \)-configurations consecutively locate at the very end of the \( n \) configurations. In this case, all the selected-entries covered by the \( \Delta \) \( g \)-configurations are LEs, but the number of LEs covered by the \( n-\Delta \) \( s \)-configurations is trivial and is ignored when counting \( M \). So, \( M=(N-1)+(N-3)+\ldots+(N-2\Delta+1)\) \((N-\Delta)\). Note that none of the \( g \)-configurations shadows \( \phi_{n+1} \). So, these LEs reside in \( N-1 \) lines of \( C(T) \) instead of \( N \) lines. Replacing \( N \) in (7) by \( N-1 \), we have

Mathematically, this is equivalent to (10) below.
\[
\phi_{n+1} \leq \max_{0 \leq \sum \frac{N}{2}} \left[ \frac{T}{n-\Delta +1} \right] + \left. \frac{2n^2 - n + 1}{2N} \right. \Delta \left. - \frac{3(N-1)2 + 8(N-1)(n+2)}{4} \right. , \quad \phi_{n+1} \leq \left. \frac{T}{n-\Delta +1} \right. + \left. \frac{2n^2 + 2n + 2N}{2N} \right. \Delta \left. - \frac{3(N-1)2 + 8(N-1)(n+2)}{4} \right. . \quad (10)
\]

\[
\phi_{n+1} \leq \max_{0 \leq \sum \frac{N}{2}} \left[ \frac{T}{n-\Delta +1} \right] \text{ and } \Delta \left. - \frac{3(N-1)2 + 8(N-1)(n+2)}{4} \right. . \quad (11)
\]

Again, this is equivalent to

\[
\phi_{n+1} \leq \frac{T}{n-\Delta +1} \left. \frac{2n^2 - n + 1}{2N} \right. \Delta \left. - \frac{3(N-1)2 + 8(N-1)(n+2)}{4} \right. . \quad (12)
\]

Combining (10), (12) and (8), we get the bound for \( \phi_{n+1} \) as shown in formula (13). We can replace \( \phi_{n} \) in (4) by the bound in (13) to calculate the \( S_{\text{schedule}} \) bound.

**C. Results**

Fig. 5 shows the new speedup bound we derived for QLEF. As an example, the new \( S_{\text{schedule}} \) bound gives a gain of 11.29% over the existing bound in [13] for \( N = 200 \). Fig. 5b shows the \( S_{\text{schedule}} \) bound evolution for the minimum delay scheduling problem. Particularly, the bound \( S_{\text{schedule}}^{4(4+\log_2N)} \) is given in [10] for MIN, which is refined in [12] to produce the saw-toothed curve. A tighter bound is then provided by \( \alpha' \)-SCALE [12]. This is then followed by QLEF in [13]. In this paper we further push the QLEF speedup bound to be 10% lower.

**IV. Conclusion**

Hybrid electronic/optical packet switch provides a scalable switch architecture for huge-capacity backbone Internet routers. Because of the reconfiguration overhead of the optical switch fabric, packet delay is minimized by using \( N \) configurations (where \( N \) is the switch size) to schedule the traffic. However, this requires a very large speedup to achieve performance guaranteed switching. QLEF (Quasi Largest-Entry-First) is the most efficient minimum delay scheduling algorithm that gives the lowest speedup bound for a given packet delay. In this paper, we derived a new speedup bound for QLEF, which is 10% lower than the existing bound.

**APPENDIX**

**Lemma 1:** Assume that \( \{P_x\} \) is a set of consecutive s-configurations of \( \phi_{n+1} \), and \( P_x \) is the first g-configuration of \( \phi_{n+1} \) after \( \{P_x\} \). Then, any \( P_x \in \{P_x\} \) must cover at least \( h \) LEs, where \( h \) is half of the number of the selected-entries in \( P_x \).

**Proof:** Since \( P_x \) is a g-configuration of \( \phi_{n+1} \), any selected-entry \( a \) covered by \( P_x \) is an LE. Because \( P_x \) is constructed before \( P_y \) and \( a \) is not covered in \( P_x \), either 1) all the selected-entries covered by \( P_x \) are not smaller than \( a \), or 2) \( a \) is shadowed in \( P_x \) construction.

In case 1), all the selected-entries in \( P_x \) are LEs. Since \( P_x \) is constructed earlier than \( P_y \), it contains more selected-entries than \( P_y \). So, the number of LEs covered by \( P_x \) is larger than \( h \). In case 2), any \( a \) covered by \( P_x \) must have been shadowed in \( P_x \) construction. Since a selected-entry in \( P_x \) can shadow at most two smaller/equal selected-entries in \( P_y \) (in row and column, respectively), \( P_x \) must cover at least \( h \) LEs. Obviously, this is true for the first g-configuration \( P_y \) after \( \{P_x\} \).

**Lemma 2:** To minimize \( M \), all the s-configurations of \( \phi_{n+1} \) should be consecutively located at either the very beginning or the very end of the configuration sequence \( P_1 \sim P_n \).

**Proof:** In Fig. 6, let \( y \)-axis denote the number of selected-entries covered in each configuration, and \( x \)-axis denote the scheduling sequence. Without loss of generality, assume there are three sets of consecutive s-configurations of \( \phi_{n+1} \), denoted by \( A_1 \), \( A_2 \) and \( A_3 \) (others are g-configurations). Particularly, \( A_1 \) and \( A_2 \) contain \( x_1 \) and \( x_2 \) s-configurations respectively, and \( A_3 \) locates at the very end of the configuration sequence \( P_1 \sim P_n \).
The first s-configuration in $A_1$ covers $y_1$ selected-entries, and the first g-configuration after $A_1$ covers $(y_1-2x_1)$ selected-entries. Similarly, the first s-configuration in $A_2$ covers $y_2$ selected-entries, and the first g-configuration after $A_2$ covers $(y_2-2x_2)$ selected-entries. From Lemma 1, each s-configuration in $A_1$, $A_2$ covers at least $(v_1-2x_1)/2$, $(v_2-2x_2)/2$ LEs. We do not count any LEs in $A_3$ because there is no g-configuration after it. Although each s-configuration in $A_3$ covers one LE in the same line as $\phi_{n+1}$, it is trivial and is ignored when counting $M$.

We first consider $A_1$ and $A_2$. Since all selected-entries covered by g-configurations are LEs, minimizing $M$ is equivalent to maximizing the number of selected-entries in the shadowed areas in $A_1$ and $A_2$. That is

$$\max \left\{ \frac{x_1(y_1-2x_1)}{2} + \frac{x_1(x_1-1)}{2} \times 2 + \frac{x_2(y_2-2x_2)}{2} + \frac{x_2(x_2-1)}{2} \times 2 \right\}$$

or

$$\max \left\{ \frac{(y_1+2)x_1 + (y_2+2)x_2}{2} \right\}.$$ 

Let $\lambda$ be the number of g-configurations between $A_1$ and $A_2$. We have $y_2 = y_1 - 2(x_1 + \lambda)$. So, the above formula is equivalent to

$$\max \frac{(y_1+2)(x_1 + x_2) - 2x_1(x_1 + \lambda)}{2}.$$ 

To maximize the above formula for any given $x_1$ and $x_2$, it is necessary that $y_1$ takes the maximum possible value and $\lambda = 0$. This entails that $A_1$ and $A_2$ should be located consecutively at the very beginning of the configuration sequence $P_1 \sim P_n$. It is easy to see that this point can be generalized to the case where more sets of consecutive s-configurations are involved.

However, we still need to consider $A_3$ in Fig. 6. In fact, the $n$ configurations before $P_{n+1}$ may also locate as shown in Fig. 7, where the $\lambda$ g-configurations are in the middle and the $n-\lambda$ s-configurations are at the both ends (the $\beta$ s-configurations locate consecutively at the beginning of $P_1 \sim P_n$ as argued above). In Fig. 7, minimizing $M$ is equivalent to maximizing the number of selected-entries in the not-shadowed areas, i.e.

$$\max_{0 \leq \beta < n - \lambda} \left[ \frac{(n-\lambda-\beta)(n-2\lambda+1)}{2} + \frac{(n-\lambda-\beta)(n-\lambda-\beta-1)}{2} \times 2 \right]$$

or

$$\max_{0 \leq \beta < n - \lambda} \left\{ \beta^2 + \frac{4A-N+1}{\beta} + (n-\lambda)(N-n-\lambda) \right\}.$$ 

Because the above formula is a quadratic function of $\beta$, it can be maximized only if $\beta = 0$ or $\beta = n-\lambda$. From Fig. 7, obviously all the $n-\lambda$ s-configurations should be consecutively located at either the very beginning ($\beta = n-\lambda$) or the very end ($\beta = 0$) of the configuration sequence $P_1 \sim P_n$. The specific location is determined by the values of $N$, $n$ and $\Delta$.

REFERENCES


