Comparison of Buffering Strategies for Asymmetric Packet Switch Modules

Soung C. Liew, *Member, IEEE*, and Kevin W. Lu, *Member, IEEE*

Abstract—**This paper analyzes the performance of a class of asymmetric packet switch modules with channel grouping. The switch module considered has n inputs and m outputs. A packet destined for a particular output address (out of g) needs to access only one of the r available physical output ports; $m = gr$. The motivation for the study of these switch modules is that they are the key building blocks in many large multistage switch architectures. We concentrate on the performance of input-buffered and output-buffered switch modules under geometrically bursty traffic. A combination of exact derivation, numerical analysis, and simulation yields the saturation throughput of input-buffered switch modules and the mean delay of the input-buffered and output-buffered switch modules. Tables and formulas useful for traffic engineering are presented. Our results show that increasing the number of output ports per output address (r) can significantly improve switch performance, especially when traffic is bursty. An interesting observation is that although output-buffered switch modules have significantly better performance than input-buffered switch modules when there are equal numbers of input and output ports, this performance difference becomes significantly smaller when the switch dimensions are asymmetric.**

I. Introduction

RECENT research activities in asynchronous transfer mode (ATM) switching have progressed to the study of large switch architectures constructed of interconnections of smaller switch modules [1]–[4]. In many cases, the underlying switch modules are of asymmetric dimensions in that there are unequal numbers of input and output ports. In addition, channel grouping, the technique of allocating more than one output port to each output address, is often used to improve switch performance. To gain insight into the design of large switch architectures of this type, it is important to understand the performance of the individual switch modules thoroughly.

Toward this goal, this paper considers the performance of a general class of asymmetric packet switch modules illustrated in Fig. 1. There are hs input ports consisting of h input groups of s input ports each, and gr output ports of g output groups of r output ports each. To achieve acceptable performance with the overall switch architecture, it is necessary to choose the various parameters of the basic switch modules properly. The objective of this paper is to quantify the performance of these switch modules as a function of the switch dimensions, buffering strategies, and traffic characteristics.

Before proceeding further, for motivation, we give three examples of switch architectures in Figs. 2–4 which make use of the class of switch modules considered here. Fig. 2 is a modular nonblocking switch architecture proposed by Lee [1]. The first stage consists of Batcher-banyan switch modules of dimensions $n \times nk$ (i.e., with respect to the switch module in Fig. 1, s, $r \rightarrow 1$, $h \rightarrow n$, and $g \rightarrow nk$). The second-stage switch modules are statistical multiplexers of dimensions $k \times 1$. Fig. 3 is a general three-stage switch architecture proposed by Liew and Lu [2]. The dimensions of the first-stage, second-stage, and third-stage switch modules are $n \times m$ ($m > n$), $l \times l'$, and $m' \times n'$ ($m' > n'$), respectively. Here, a channel group of r (r') channels interconnects switch modules of adjacent stages. The structure is such that if r and r' were to be 1, there would be one and only one path between any input and output. For better performance, however, r, $r' > 1$ (in general) and packets have several alternative paths from their inputs to their destination outputs. Finally, Fig. 4 is a three-stage Clos switch architecture [4] that employs asymmetric switch modules at the two outer stages, and symmetric switch modules at the middle stage. There is no channel grouping internally, however. In all three schemes, asymmetric switch modules at the first stage result in internal line expansion which improves the performance of the overall switch architecture. It is also worth pointing out that the class of asymmetric switches we study here can also be used as

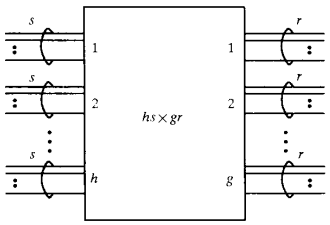


Fig. 1. The asymmetric switch module with channel grouping.

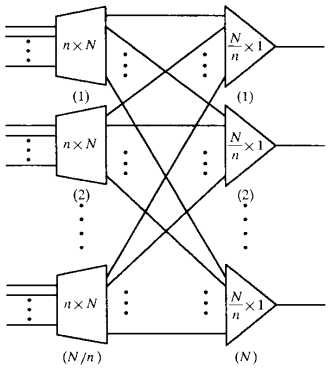
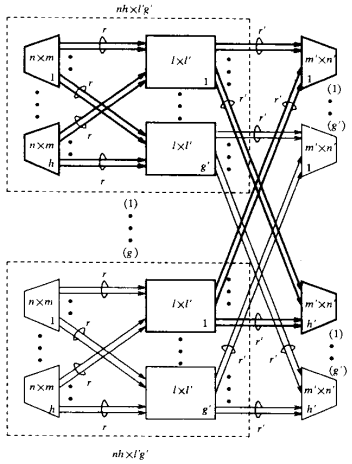


Fig. 2. Two-stage nonblocking modular switch architecture.

Manuscript received April 18, 1990; revised November 3, 1990.

The authors are with Bell Communications Research, Morristown, NJ 07960-1910.

IEEE Log Number 9142935.



$$N = n\,g\,h = n'g'h'$$

$$g' = \frac{m}{r} \qquad g = \frac{m'}{r}$$

$$h = \frac{l}{r} \qquad h' = \frac{l'}{r}$$

Fig. 3. General structure of a three-stage switch architecture.

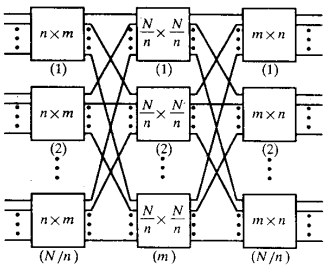


Fig. 4. Three-stage Clos switch architecture.

stand-alone concentrators or expanders rather than components of an overall switch.

Referring to Fig. 1, to the extent that packets at different input ports within the same group are uncorrelated, the switch module reduces to that shown in Fig. 5 in which $hs \rightarrow n$. For simplicity, this paper focuses on the structure shown in Fig. 5, assuming any correlations between packets of different input ports are small and negligible. An output group [2], [5] corresponds to an output address, and a packet can access any of the r output ports of its output address. In any given time slot, at most r packets can be cleared from a particular output group, one on each of the r output ports. Furthermore, we assume packets are destined for a particular output group (address) rather than a particular output port. That is, it does not matter which particular output port a packet accesses as long as the output port belongs to the correct output group. Reference [2] provides several designs of channel-grouping switch modules. It turns out that channel-grouping switch modules have smaller complexity (in terms of switch element counts) than ordinary switch modules of the same dimensions.

We focus on the performance of the input queueing and output queueing buffering strategies under geometric traffic. The

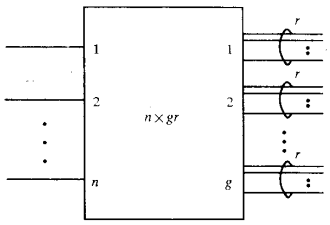


Fig. 5. The asymmetric switch module with uncorrelated inputs.

paper is a generalization of the work reported in [6]–[8] in two respects: the switch dimensions as well as the traffic characteristics have been generalized. With input queueing, an arriving packet enters a FIFO buffer on its input and waits for its turn to access its destination output. With output queueing, a logical FIFO buffer is allocated to each output group, and arriving packets destined for this output group are immediately placed into its FIFO. For simplicity, we assume infinite buffer queues for both input queueing and output queueing in our mean delay analysis.

The organization of this work is outlined as follows. Section II describes the geometric traffic model that we use to model bursty traffic. Section III investigates the maximum throughput and mean delay of input-buffered switch modules, and discusses the maximum throughput degradation due to head-of-line blocking under various settings. Section IV considers the mean delay of output-buffered switch modules. Finally, Section V concludes this work and discusses issues that deserve further attention.

II. Traffic Model

We consider ATM transport in which data streams are partitioned and transferred in cells (or packets) of fixed size. On a conceptual level, time is therefore divided into slots corresponding to the cell transmission time. For performance analysis, we assume synchronous switch operation in which cells arrive at the beginning of each time slot, and cells gaining access to their output lines are cleared by the end of each time slot. To quantify the traffic characteristics, we focus on the uniform geometrically bursty traffic model in which an input alternates between active and idle periods of geometrically distributed duration [9]. During the active period of an input, packets destined for the same output arrive at the input continuously in consecutive time slots (see Fig. 6). Termination of the active period is a renewal process, and it occurs with probability p after each active time slot. Thus, the probability that the active period (burst) will last for a duration of i time slots (consists of i packets) is

$$P(i) = p(1-p)^{i-1}, \qquad i \geq 1. \tag{1}$$

Note that we assume there is at least one cell in the burst. This geometric burst-length assumption yields a mean burst length of

$$l = E_B = \sum_{i=1}^{\infty} iP(i) = 1/p. \tag{2}$$

The idle period is also geometrically distributed and is characterized by another parameter q. The probability that an idle period lasts for j time slots is

$$Q(j) = q(1-q)^{j}, \qquad j \geq 0. \tag{3}$$

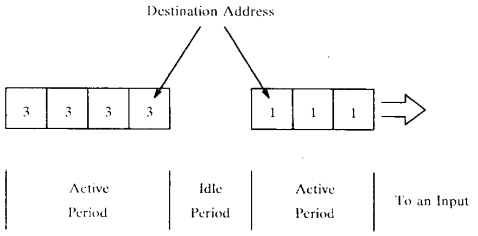


Fig. 6. Geometric packet arrivals to an input.

Unlike the duration of an active period, the duration of an idle period can be 0. The mean idle period is given by

$$E_I = \sum_{i=0}^{\infty} jQ(j) = (1 - q)/q. \tag{4}$$

Given p and q, the offered load ρ can be found by

$$\rho = E_B/(E_I + E_B). \tag{5}$$

For simplicity, we focus on uniform output destination distribution in which any burst has equal probability of being destined for any output group. In addition, there is no correlation between different bursts. Note that the uniform random traffic model discussed in [6] and [7] is a special case with $p = 1$ and $q = \rho$, i.e., the burst length is deterministic, and it is always one packet long.

III. Input Queueing

In this section, the maximum throughput of asymmetric input-buffered switch modules with channel grouping is obtained by numerical analysis, and the mean delay by simulation.

In an input-buffered switch, when there are multiple packets at the heads of input queues destined for the same output group, only up to r packets may access the output group. As a result, some inputs idle because the first packets in their queues are blocked. Meanwhile, subsequent packets in the queues, which may be destined for other available output groups, are also blocked because of the FIFO queueing discipline. This is often referred to as head-of-line blocking [6], [7], and it is well known that head-of-line blocking limits the maximum throughput of a symmetric input-buffer switch ($r = 1$, $n = g \to \infty$) to 0.586 under the uniform random traffic condition ($p = 1$). Although the maximum throughput when $r = 1$ can be derived in closed form (including when $p \neq 1$), a general closed-form solution is not possible when $r > 1$. Nevertheless, a similar approach could be taken to a point where the solution could be found by numerical analysis. The same analysis yields the throughput of the switch module as a concentrator ($n > g$) or an expansion network ($n < g$).

To find the maximum throughput, we consider the situation in which the input queues are saturated so that one can always find packets in every queue. In particular, there is always a packet at the head of each queue, waiting to access its destination. Only after this packet is cleared can the next packet move to the head of the queue.

We define the *free input queues* to be input queues with packets transmitted in the previous time slot. The subsequent packet in a free input queue immediately moves to the head, ready to access its output destination in the current time slot. This subsequent packet could be from the same burst as the cleared packet, in which case the destination remains the same, or it could be from a new burst, in which case it is equally likely to be destined for any output group. For our traffic model, the probability that the subsequent packet belongs to the same burst is given by $1 - p = (l - 1)/l$. Now, strictly speaking, for any finite buffer queue and $l > 1$, the mean burst length of bursts that arrive at the head of queue is not l, even though the mean burst length of the incoming traffic to the queue is l. This is because of the finite packet loss probability due to buffer overflow. For instance, if we overload the switch with a load of 1, then the mean burst length that arrives at the head of queue is actually $l\rho^*$, where ρ^* is the maximum allowable throughput. As far as analysis is concerned, the situation becomes even worse because the burst length is not strictly geometrically distributed after packets are dropped. Nonetheless, the queue would also saturate even if the offered load is just slightly over ρ^*, and in this situation the effective burst length would still be close to l and roughly geometrically distributed. For simplicity, therefore, we assume this to be the case for our saturation analysis. This assumption is further justified by later results which show that the maximum throughput approaches an asymptotic value very quickly as l increases; that is, the maximum throughput is not a strong function of l for moderate l values.

We now set up the framework for derivation of the maximum throughput. Consider a tagged output group i. Let A_j^i be the number of new bursts destined for output group i that move into the heads of free input queues in the beginning of time slot j. Note that under random uniform traffic ($l = 1$), A_j^i is also the number of packets destined for output group i since there is no distinction between bursts and packets. Under bursty traffic ($l > 1$), A_j^i does not include packet arrivals that belong to the same bursts as the packets just cleared. Let D_j^i be the number of bursts that terminate at the end of time slot j. Under bursty traffic, D_j^i is the number of departed packets minus departures which are subsequently replaced by packets of the same bursts. Let C_j^i be the number of head-of-line bursts that are destined for output group i at the beginning of time slot j, and let G_j^i be the number of head-of-line bursts left at the end of time slot j. Then,

$$C_{j+1}^i = G_j^i + A_{j+1}^i \tag{6}$$

where

$$G_j^i = C_j^i - D_j^i. \tag{7}$$

Note that C_j^i includes the bursts that are granted output access as well as the bursts that are blocked during time slot j.

By the assumption that packet output destinations are uniformly distributed across all output groups, all output groups face the same situation, and the superscript i can be dropped. The subscript j can also be dropped as the system approaches equilibrium. To simplify analysis, we will assume n, $g \to \infty$ while keeping a fixed value of g/n. This approximation is valid when n is large (e.g., $n \geq 16$). As in [8], it can be shown that $\lim_{n,g\to\infty} \Pr[A = k] = e^{-p\rho_0}(p\rho_0)^k/k!$, where $p\rho_0$ is the average arrival rate of new bursts, and ρ_0 the offered load per output group. For $n \to \infty$, there is no correlation between G and A, and the moment-generating functions of the parameters in (6) are related as follows:

$$C(z) = G(z)\,A(z) \tag{8}$$

where

$$A(z) = e^{-p\rho_0(1-z)}. \tag{9}$$

The key to finding the maximum throughput lies in the observation that the sum of the numbers of backlogged bursts over all output groups at the beginning of each time slot must be n, because there is always a head-of-line burst at each of the n input queues under the saturation condition. Since $C'(1)$ is the expected number of backlogged bursts per output group, we have

$$C'(1) = n/g. \tag{10}$$

The maximum throughput per output group can be found by equating $C'(1)$ obtained from the analysis based on (8) with n/g.

Let us define $P_i = \Pr[C = i]$. Then,

$$\begin{aligned} G(z) &= \sum_{i=0}^{\infty} G(z|C = i)P_i \\ &= P_0 + [p + (1-p)z]P_1 \\ &\quad + \left[p^2 + \binom{2}{1}p(1-p)z + (1-p)^2z^2\right]P_2 \\ &\quad \vdots \\ &\quad + \left[p^{r-1} + \binom{r-1}{1}p^{r-2}(1-p)z \right. \\ &\quad + \binom{r-1}{2}p^{r-3}(1-p)^2z^2 \\ &\quad \left. + \cdots + (1-p)^{r-1}z^{r-1}\right]P_{r-1} \\ &\quad + \sum_{j=0}^{\infty} z^j\left[p^r + \binom{r}{1}p^{r-1}(1-p)z \right. \\ &\quad + \binom{r}{2}p^{r-2}(1-p)z^2 \\ &\quad \left. + \cdots + (1-p)^r z^r\right]P_{r+j}. \end{aligned}$$

The above can be simplified to

$$\begin{aligned} G(z) &= \sum_{i=0}^{r-1} [p + (1-p)z]^i P_i \\ &\quad + z^{-r}[p + (1-p)z]^r\left[C(z) - \sum_{i=0}^{r-1} P_i z^i\right]. \end{aligned} \tag{11}$$

From (8) and (11), we obtain

$$C(z) = \frac{\sum_{i=0}^{r-1}\left\{z^r[p + (1-p)z]^i - z^i[p + (1-p)z]^r\right\}P_i}{z^r/A(z) - [p + (1-p)z]^r}. \tag{12}$$

The equilibrium probabilities P_i, $i = 1, \cdots, r-1$, can be obtained using a standard method described as follows. It can be shown by Rouche's Theorem [10] that the denominator of $C(z)$ has $r-1$ zeros, z_k, $k = 1, \cdots, r-1$, with magnitudes less than 1. Since $C(z)$ is a moment-generating function, it must be analytical for all $|z| < 1$, and therefore, z_k, $k = 1, \cdots, r-1$ must also be zeros of the numerator of $C(z)$. Thus, given z_k, $k = 1, \cdots, r-1$, we have $r-1$ linear equations relating r unknown P_k's. The normalization requirement $C(1) = 1$ gives us the other equation needed: $\sum_{i=0}^{r-1}(r-i)P_i = r - \rho_0$.

To summarize the above, the maximum throughput for an output group ρ_0 can be found numerically as follows. Starting with a guess of ρ_0, we first solve for z_k, $k = 1, \cdots, r-1$, with the following $r-1$ complex equations

$$A(z_k)^{1/r}[p + (1-p)z_k] = z_k\left(\cos\frac{2k\pi}{r} + i\sin\frac{2k\pi}{r}\right)$$
$$k = 1, \cdots, r-1. \tag{13}$$

The following r linear equations are then solved to find P_i.

$$\sum_{i=0}^{r-1}\left\{z_k^r[p + (1-p)z_k]^i - z_k^i[p + (1-p)z_k]^r\right\}P_i = 0$$
$$k = 1, \cdots r-1$$
$$\sum_{i=0}^{r-1}(r-i)P_i = r - \rho_0. \tag{14}$$

A new ρ_0 is found by

$$C'(1) = \frac{\rho_0(2r - p\rho_0) - r(r-1)(2-p) + \sum_{i=0}^{r-1}[r(r-1) - i(i-1)](2-p)P_i}{2(r-\rho_0)} = \frac{n}{g}. \tag{15}$$

The three steps are repeated with the new ρ_0, and the process is iterated until the solution converges to the desired accuracy. The maximum throughput per input is related to ρ_0 by

$$\rho^* = \frac{g}{n}\rho_0. \tag{16}$$

The above is the general method for finding ρ^*. Various specific cases described below are amenable to simpler analysis, and they are described as follows.

(r arbitrary, p = 1): For uniform random traffic ($p = 1$), the second step of the numerical iterations can be eliminated because it is not necessary to explicitly solve for P_i, $i = 1, \cdots, r-1$. In this case, instead of $2r-1$ zeros, there are only r zeros in the numerator of $C(z)$, and they are all equal to the roots of the denominator, $z = 1$ and z_k, $k = 1, \cdots, r-1$. We can directly express $C(z)$ in terms of z_k as follows.

$$C(z) = \frac{K(z-1)(z-z_1)\cdots(z-z_{r-1})}{z^r/A(z) - 1} \tag{17}$$

where $K = p(r-\rho_0)/(1-z_1)\cdots(1-z_{r-1})$ is a normalization constant found by setting $C(1) = 1$. Differentiating (17) with respect to z, and setting $z = 1$ yields

$$C'(1) = \frac{\rho_0(2r-\rho_0) - r(r-1)}{2(r-\rho_0)} + \sum_{k=1}^{r-1} \frac{1}{1-z_k}. \tag{18}$$

(r arbitrary, p → 0): In the limiting case when the average burst length $l \to \infty$ ($p \to 0$), (12) becomes

$$\lim_{p\to 0} C(z) = \frac{\sum_{i=0}^{r-1}(r-i)z^i P_i}{r-\rho_0 z}. \tag{19}$$

By definition, $C(z) = \Sigma_{i=0}^{\infty} P_i z^i$. Multiplying both sides of (19) by $(r-\rho_0 z)$, and equating the coefficient of z^i on the left side with that on the right side, we obtain

$$P_i = \begin{cases} \rho_0 P_{i-1}/i & \text{if } i < r \\ \rho_0 P_{i-1}/r & \text{if } i \geq r. \end{cases}$$

This simplifies to

$$P_i = \begin{cases} \rho_0^i P_0/i! & \text{if } i < r \\ \rho_0^i P_0/(r!r^{i-r}) & \text{if } i \geq r \end{cases} \tag{20}$$

where

$$P_0 = \frac{r-\rho_0}{\sum_{i=0}^{r-1}(r-i)\rho_0^i/i!} \tag{21}$$

which is obtained by normalizing $C(1) = 1$. It is remarkable that (20) is the exact result of the $M/M/r$ queue with λ/μ, the ratio of the arrival rate to the service rate, equal to ρ_0. This, however, does conform to the intuitive understanding that as $p \to 0$, the geometrically distributed burst length becomes exponentially distributed.

Differentiating $C(z)$ in (19) and setting $z = 1$, we obtain

$$C'(1) = \frac{\sum_{i=0}^{r-1} i(r-i)P_i + \rho_0}{(r-\rho_0)}. \tag{22}$$

Equating (22) with n/g gives us a polynomial of ρ_0, from which we can obtain ρ_0 numerically. For specific values of r and n/g listed below, $\rho^* = \rho_0 g/n$ can be obtained in closed form by solving for the roots of the corresponding second-order polynomials directly.

$$\rho^* = \begin{cases} \dfrac{g}{n}\Big/\left(1+\dfrac{g}{n}\right) & \text{if } r = 1 \\ 2\dfrac{g}{n}\left(\sqrt{\left(\dfrac{g}{n}\right)^2 + 1} - \dfrac{g}{n}\right) & \text{if } r = 2 \\ (\sqrt{162} - 6)/7 \approx 0.961 & \text{if } r = 3 \text{ and } g/n = 1. \end{cases} \tag{23}$$

For the interested readers, it turns out that ρ^* at $r \geq 4$ and $g/n = 1$ can be easily approximated. Consider $r = 4$ for an example. We know the corresponding ρ^* must be very close to 1 since ρ^* at $r = 3$ is already close to 1. Substituting $\rho^* = (1-\epsilon)$ into $C'(1) = 1$ and ignoring the second and higher order ϵ terms, we get $\epsilon = 0.007$. The resulting $\rho^* = 0.993$ agrees with the exact result to three decimal places. For general g/n and r, however, the numerical root-finding method is needed to find ρ^*.

(r = 1, p arbitrary): Finally, for $r = 1$ and arbitrary p, the numerator of $C(z)$ in (12) has only one root, $z = 1$, and ρ^* can be solved in closed form:

$$\rho^* = \frac{(1+g/n) - \sqrt{(1+g/n)^2 - 2pg/n}}{p}. \tag{24}$$

We are now ready to examine results generated by the above analysis. Table I(a) lists the maximum throughput per input port for various values of r and g/n under uniform random traffic. The column in which $g/n = 1$ corresponds to the special cases studied by [8] and [11]. For a given r, the maximum throughput increases with g/n because the load on each output group decreases with g/n. For a given g/n, the maximum throughput increases with r because each output group has more output ports for clearing packets. This is analogous to increasing the number of servers in a queueing system. As shown in the table, when g/n is fairly large (say, $g/n > 4$), there is less incentive to use channel grouping to increase the throughput because the throughput is already close to 1. When g/n is small (say, $g/n < 2$), however, the use of channel grouping can increase the throughput substantially. For concentrators ($g/n < 1$), increasing the number of output ports per output address from 1 to 2 approximately doubles the maximum throughput.

As the average burst length l increases (or p decreases), the maximum throughput degrades. As shown in Fig. 7, the maximum throughput in general approaches an asymptotic value rather quickly as l increases. In particular, the maximum throughput for $l > 5$ is essentially equal to the asymptotic value. Table I(b) lists the asymptotic maximum throughput as $l \to \infty$. As shown, the difference in maximum throughput between the two extreme cases of $p = 1$ and $p \to 0$ is very small. Furthermore, it can be seen that the qualitative results for uniform random traffic described above also hold here. In addition, it can be easily verified that for a fixed r (i.e., for a particular row in the table), the percentage change in maximum throughput by varying p from 1 to 0 is the greatest when $gr/n = 1$, i.e., when the switch dimensions are symmetric. For instance, for $r = 2$, $g/n = 1/2$ yields the greatest percentage change in maximum throughput.

If there were no head-of-line blocking, then the maximum allowable throughput per input would be min $(1, gr/n)$. This

TABLE I
MAXIMUM THROUGHPUT FOR AN INPUT QUEUE WITH q/n KEPT CONSTANT WHILE BOTH g AND $n \to \infty$

(a) $p = 1$

	$\frac{g}{n}$												
r	$\frac{1}{32}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	1	2	3	4	8	16	32
1	0.031	0.061	0.117	0.219	0.279	0.382	0.586	0.764	0.838	0.877	0.938	0.969	0.984
2	0.061	0.121	0.233	0.426	0.531	0.686	0.885	0.966	0.984	0.991	0.998	0.999	1.000
3	0.092	0.181	0.346	0.613	0.736	0.875	0.975	0.996	0.999	0.999	1.000	1.000	
4	0.123	0.241	0.457	0.768	0.875	0.959	0.996	1.000	1.000	1.000			
8	0.245	0.476	0.831	0.991	0.998	1.000	1.000						
16	0.487	0.878	0.999	1.000	1.000								
32	0.912	1.000	1.000										

(b) $p = 0$

	$\frac{g}{n}$												
r	$\frac{1}{32}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	1	2	3	4	8	16	32
1	0.030	0.059	0.111	0.200	0.250	0.333	0.500	0.667	0.750	0.800	0.889	0.941	0.970
2	0.061	0.117	0.221	0.390	0.481	0.618	0.828	0.944	0.974	0.985	0.996	0.999	1.000
3	0.091	0.176	0.328	0.565	0.678	0.823	0.961	0.994	0.998	0.999	1.000	1.000	
4	0.121	0.234	0.432	0.715	0.828	0.937	0.993	0.999	1.000	1.000			
8	0.241	0.460	0.791	0.987	0.997	1.000	1.000	1.000					
16	0.477	0.849	0.999	1.000	1.000								
32	0.891	1.000	1.000										

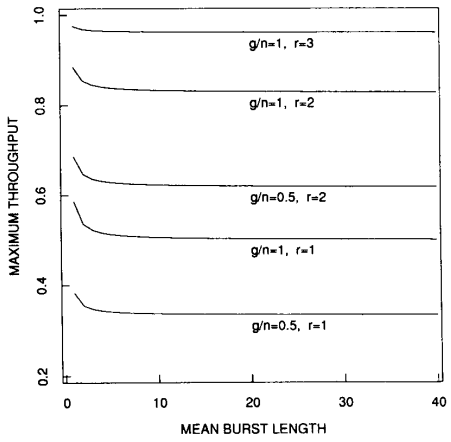


Fig. 7. Maximum throughput as a function of mean burst length.

is because we cannot load each input line with load greater than 1 or each output group with load greater than r. We can therefore define the degradation due to head-of-line blocking as

$$\Delta(r, g/n) = \min(1, gr/n) - \rho^*. \quad (25)$$

Since $\min(1, gr/n)$ is also the maximum throughput of the output-buffered switch module, $\Delta(r, g/n)$ can be interpreted as the throughput advantage of the output-buffered switch module over the input-buffered switch module. Table II(a) and (b) show the $\Delta(r, g/n)$ values for $p = 0$ and $p = 1$, respectively. It can be seen that for either a row (i.e., fixed r) or a column (i.e., fixed g/n), the degradation is the biggest when $gr = n$, and the degradation becomes progressively smaller as we deviate from this point. Thus, whenever the switch dimensions become asymmetric ($gr \neq n$), the throughput advantage of the output-buffered switch module over the input-buffered switch module diminishes. This can be explained intuitively as follows. When $gr < n$, both input queueing and output queueing are limited by the fact that there are fewer number of output ports than input ports, and head-of-line blocking is not the main limiting factor in input-buffered switch modules anymore. When $gr > n$, the maximum throughput of output-buffered switch modules is still limited by 1, while that of input-buffered switch modules improves because the detrimental effect of head-of-line blocking is alleviated by the fact that more head-of-line packets can be cleared now. The table also reveals that for a fixed number of output ports gr, decreasing the number of output addresses g while increasing the channel group size r also alleviates the head-of-line blocking effect and decreases the maximum throughput difference between the two buffering strategies.

As an example of the application of the above results, consider the two-stage switch architecture in Fig. 3. According to our results, the expanded Batcher-Banyan switch modules would have no significant throughput limitations if $N/n \geq 32$.

Analysis of the mean delay of input-buffered switch modules is difficult, so simulation is used here. For input queueing, a contention resolution scheme is needed in order to resolve conflicts when there are more than r packets destined for the same output group. Whereas the maximum throughput of input-buffered switch modules under geometric traffic is insensitive to the particular contention resolution scheme adopted (as long as no head-of-line packets are withheld from clearance when there are free destination output ports), the mean delay does depend on

TABLE II
INCREMENT OF MAXIMUM THROUGHPUT FOR AN OUTPUT QUEUE OVER AN INPUT QUEUE $\Delta(r, g/n) = \min(1, gr/n) - \rho^*$

(a) $p = 1$

	$\frac{g}{n}$												
r	$\frac{1}{32}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	1	2	3	4	8	16	32
1	0.000	0.002	0.008	0.031	0.054	0.118	**0.414**	0.236	0.162	0.123	0.062	0.031	0.016
2	0.002	0.004	0.017	0.074	0.136	**0.314**	0.115	0.034	0.016	0.009	0.002	0.001	0.000
3	0.002	0.007	0.029	0.137	**0.264**	0.125	0.025	0.004	0.001	0.001	0.000	0.000	
4	0.002	0.009	0.043	**0.232**	0.125	0.041	0.004	0.000	0.000	0.000			
8	0.005	0.024	**0.169**	0.009	0.002	0.000	0.000						
16	0.013	**0.122**	0.001	0.000	0.000								
32	**0.088**	0.000	0.000										

(b) $p = 0$

	$\frac{g}{n}$												
r	$\frac{1}{32}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	1	2	3	4	8	16	32
1	0.001	0.004	0.014	0.050	0.083	0.167	**0.500**	0.333	0.250	0.200	0.111	0.059	0.030
2	0.002	0.008	0.029	0.110	0.186	**0.382**	0.172	0.056	0.026	0.015	0.004	0.001	0.000
3	0.003	0.012	0.047	0.185	**0.322**	0.177	0.039	0.006	0.002	0.001	0.000	0.000	
4	0.004	0.016	0.068	**0.285**	0.172	0.063	0.007	0.001	0.000	0.000			
8	0.009	0.040	**0.209**	0.013	0.003	0.000	0.000	0.000					
16	0.023	**0.151**	0.001	0.000	0.000								
32	**0.109**	0.000	0.000										

the contention scheme. Our simulation experiments assume a random strategy in which a random input port, say Port_{top}, is chosen to have the highest priority at the beginning of each time slot. The priorities of the input ports for that time slot are then ordered in a cyclic manner: Port_{top}, $\text{Port}_{\text{top}+1 \pmod n}$, $\cdots$, $\text{Port}_{\text{top}+n-1 \pmod n}$. Fig. 8 shows the graphs of the mean delay versus the input offered load for various values of r and g/n, fixing n at 32, and l at 1 and 16. Simulation results show that for a given r and g/n, but $n > 32$, the mean delay is closely approximated by the results of $n = 2$. For all cases shown, enough packet statistics are collected so that the 95% confidence interval is no more than $\pm 6\%$ of the collected mean delay value.

As shown in the figure, for uniform random traffic ($l = 1$), the mean delay is rather low except for offered loads close to the maximum allowable throughput. This is, however, not the case for the bursty traffic ($l = 16$), where the maximum delay increases rather quickly as the offered load increases. Comparing cases to $r = 1$, $r = 2$, and $r = 4$, we also see that as traffic becomes bursty, the mean delay does not degrade as much for large r than for small r. To further illustrate this point, we plot in Fig. 9 the mean delay for various cases with $n = gr$ (the dotted lines) as a function of l for an offered load of 0.5 (the asymptotic maximum throughput when $n = g$, $r = 1$ and $p \to 0$). As shown, the slopes of the mean delay versus l curves decrease quite rapidly as r increases. Thus, in general, channel grouping improves the mean delay, as well as the maximum throughput, under bursty traffic conditions.

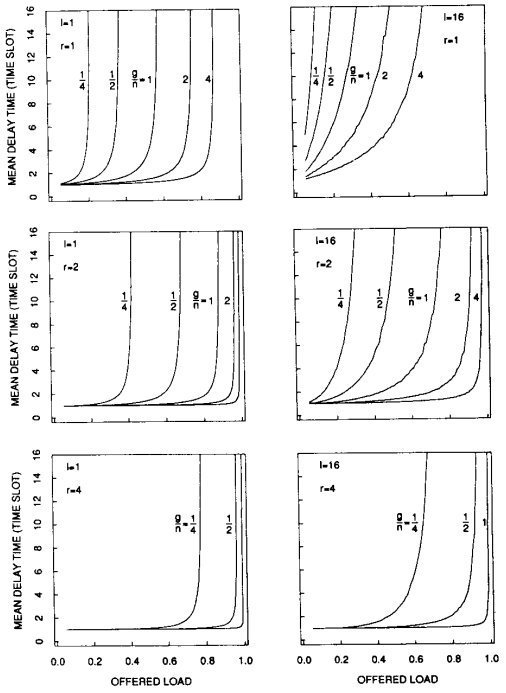


Fig. 8. Mean delay versus offered load of input-buffered switch modules.

IV. OUTPUT QUEUEING

For output queueing switch modules, we assume there is a single FIFO queue for each output group. Arriving packets destined for a given output group are immediately placed on the corresponding output queue. Unlike input queueing, there is no head-of-line blocking in output queueing, and the maximum throughput per output group is bounded by r. Except for cases

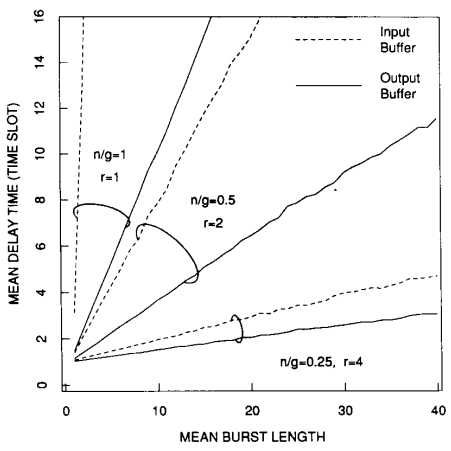


Fig. 9. Mean delay of input- and output-buffered switch modules as a function of the average burst length for an offered load of 0.5.

with $r > 1$ and $l > 1$ (i.e., $p < 1$), the mean delay of output-buffered switches can be obtained using the same framework for deriving the maximum throughput of input-buffered switches. In the following, we consider the three cases ($r \geq 1$, $p = 1$), ($r = 1, p < 1$), and ($r > 1, p < 1$), separately,

($r \geq 1, p = 1$): The situation faced by an output group is described precisely by the framework used to derive the maximum throughput of input-buffered switches, except that arriving packets are immediately presented to the output group for clearance, and do not have to first proceed to the heads of input queues. Unlike the derivation of the maximum throughput, however, ρ_0 is interpreted as the given offered load here.

We shall use the same notation as in the maximum throughput derivation. Consider a particular output group. Given an output offered load of ρ_0 per group, the expected number of backlogged packets for an output group at the beginning of a time slot is given by $C'(1)$ in (18). By Little's Law, the mean delay is

$$\bar{T} = C'(1)/\rho_0 = \frac{\rho_0(2r - \rho_0) - r(r-1)}{2(r - \rho_0)\rho_0} + \frac{1}{\rho_0}\sum_{k=1}^{r-1}\frac{1}{1 - z_k} \tag{26}$$

where z_k, $k = 1, \cdots, r-1$ are the $r-1$ roots of

$$\exp\left[\frac{-\rho_0(1 - z_k)}{r}\right] = z_k\left[\cos\left(\frac{2k\pi}{r}\right) + i\sin\left(\frac{2k\pi}{r}\right)\right] \qquad k = 1, \cdots, r-1. \tag{27}$$

($r = 1, p < 1$): When $l > 1$, a complete analysis involves a two-dimensional Markov chain which keeps track of the number of backlogged packets and the number of bursts with packets still arriving [9]. However, as shown below, simpler analysis is sufficient for deriving the mean delay.

The situation for $r = 1$ is closely related to the $M/G/1$ queue; bursts are analogous to customers and burst lengths to durations of service. As far as the mean delay of a packet is concerned, it does not matter whether we finish serving (clearing) packets of one burst before serving packets of the next burst, or serve packets of the backlogged bursts in an arbitrary order, since the unfinished work, or the number of remaining packets, is the same in either case. Without losing generality, we focus our attention on the former burst-by-burst service discipline.

There is a subtle difference between the $M/G/1$ queue and our situation, however. In the $M/G/1$ queue, when a customer arrives, it arrives in its totality, whereas, in our case, packets in a burst arrive in consecutive time slots. Nonetheless, the waiting time of a packet has the same distribution as that of a burst, and it can be obtained from $M/G/1$ analysis. To see this, consider the jth packet in a burst. Suppose that the waiting time of the burst (or the first packet), or the time the burst spends waiting in the queue before it is served, is W. Although the jth packet arrives $j - 1$ time slots later than the first packet, it is also served $j - 1$ time slots later than the first packet. Thus, the waiting time of the jth packet is also W. Notice that not only are the mean waiting times of the burst and its packets the same, the waiting times are also identically distributed under the burst-by-burst service discipline.[1]

By Little's Law, the mean delay of a burst is $C'(1)/p\rho_0$, where $C'(1)$is given by (15). Therefore, the mean burst or packet waiting time is $C'(1)/p\rho_0 - 1/p$, and the mean packet delay is

$$\bar{T} = C'(1)/p\rho_0 - 1/p + 1 = \frac{2 - p\rho_0}{2(1 - \rho_0)p} - 1/p + 1. \tag{28}$$

($r > 1, p < 1$): When $r > 1$ and $l > 1$, things become more complicated because the analogy between bursts and customers breaks down. To see this, consider the following. If there are fewer than r customers in an $M/G/r$ queue, then some of the servers are not active. For a switch, however, even if there are fewer than r backlogged bursts in the output queue, as long as there are at least r packets, all the r output lines would be active, and multiple packets from some bursts are served simultaneously. Nevertheless, this observation implies that $M/G/r$ analysis can be used to obtain an upper bound to the actual mean delay. That is, given an offered load ρ_0, the mean packet delay is upper-bounded by $C'(1)/p\rho_0 - 1/p + 1$, where $C'(1)$ is given by (15). This yields

$$\bar{T} < \frac{\rho_0(2r - p\rho_0) - r(r-1)(2-p) + \sum_{i=0}^{r-1}[r(r-1) - i(i-1)](2-p)P_i}{2(r - \rho_0)p\rho_0} - 1/p + 1 \tag{29}$$

where P_i, $i = 0, \cdots, r-1$, are obtained by solving (13) and (14).

A lower bound to the mean delay can be obtained by considering a modified system in which all packets in a burst are assumed to arrive simultaneously in the beginning of the burst. The basic idea is as follows. In the modified system, the arrival instants of all packets in a burst are shifted to the arrival instant

[1]This implies that the probability $P[W > b]$ obtained from $M/G/1$ analysis can be used as an upper bound for the packet loss probability of a finite buffer queue of b packets deep.

of the first packet. In contrast to the original system, the modified system allows all r output lines to be utilized even when the burst arrives at a queue with fewer than $r - 1$ backlogged packets. Consequently, the departures of some packets are also shifted to earlier instants in the modified system. After the mean delay of the modified system is found, it is easy to compensate for the extra delay due to the shift in arrival instants. It is, however, difficult to compensate for the shift in departure instants or else we would have found an exact solution. Nevertheless, by compensating only for the shift in arrival epochs, a lower bound to the mean delay of the original system is obtained.

In the following lower-bound analysis, instead of focusing on bursts, we focus on packets, and use C and A to denote the number of backlogged packets and the number of packet arrivals, respectively. We essentially have a $G/D/r$ system in which

$$C_{j+1} = \max(0, C_j - r) + A_j \tag{30}$$

where the moment-generating function of A_j is given by

$$\begin{aligned} A(z) &= \sum_{k=0}^{\infty} A(z|k)\, P[k \text{ bursts arrive}] \\ &= \sum_{k=0}^{\infty} \left[\frac{pz}{1-(1-p)z}\right]^k \frac{(p\rho_0)^k e^{-p\rho_0}}{k!} \\ &= \exp\left[\frac{-p\rho_0(1-z)}{1-(1-p)z}\right]. \end{aligned} \tag{31}$$

Using an analysis similar to that in the derivation of the maximum input-queueing throughput, we obtain

$$C(z) = A(z)\frac{\sum_{i=0}^{r-1}(z^r - z^i)P_i}{z^r - A(z)}. \tag{32}$$

This gives

$$\begin{aligned} C'(1) &= \frac{\rho_0(2r-\rho_0) + 2\rho_0(1-p)/p - r(r-1)}{2(r-\rho_0)} \\ &\quad + \sum_{k=1}^{r-1}\frac{1}{1-z_k} \end{aligned} \tag{33}$$

where z_k, $k = 1, \cdots, r-1$ are obtained by solving for the roots of

$$\exp\left\{\frac{-p\rho_0(1-z_k)}{[1-(1-p)z_k]r}\right\} = z_k\left[\cos\left(\frac{2k\pi}{r}\right) + i\sin\left(\frac{2k\pi}{r}\right)\right] \quad k = 1, \cdots, r-1. \tag{34}$$

The delay of a packet in this $G/D/r$ queue is then given by $C'(1)/\rho_0$.

We now compensate for the extra waiting time due to the earlier arrival assumption. Given a packet is in a burst of length k, its expected *extra* waiting time in the modified system is $(k-1)/2$. The probability that a packet is in a burst of length k is $k(1-p)^{k-1}p^2$. The expected extra waiting time is, therefore,

$$\sum_{k=1}^{\infty}\frac{(k-1)}{2}k(1-p)^{k-1}p^2 = (1-p)/p. \tag{35}$$

Thus, we have

$$\bar{T} > \frac{C'(1)}{\rho_0} - (1-p)/p. \tag{36}$$

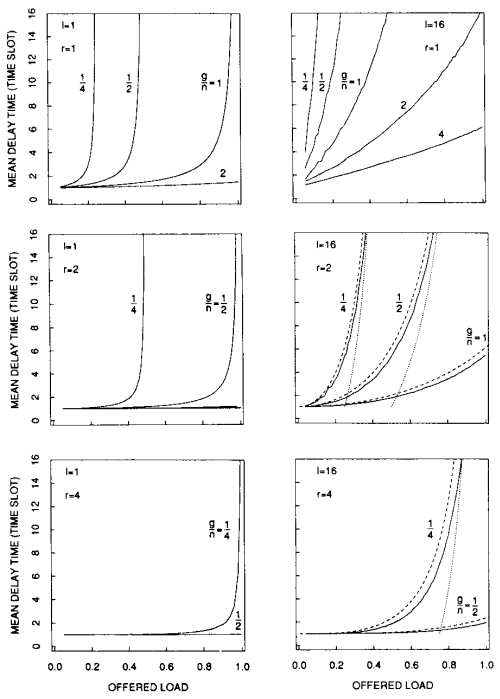


Fig. 10. Mean delay versus offered load of output-buffered switch modules.

It is worth pointing out that both the upper bound and lower bound described above become the exact solution when $p = 1$ or when $r = 1$.

Based on the above analysis, Fig. 10 shows the mean delay versus the offered input load ($g\rho_0/n$) for various values of r and g/n ratio, when $l = 1$ and when $l = 16$. Cases with ($r = 1$, 2, or 4, $l = 1$) and ($r = 1$, $l = 16$) are numerical results, whereas cases with ($r = 2$ or 4, $l = 16$) are simulation results. The figure also compares the analytical upper and lower bounds with the simulation results. Although the simulation assumes $n = 32$ and the analytical results assume $n \to \infty$, this slight discrepancy would not invalidate the following discussion, since the results are not very sensitive to n for $n > 16$.

As expected, comparing Fig. 10 with Fig. 8, the mean delay versus throughput performance of output queueing is uniformly better than that of input queueing for all cases. As in input queueing, however, bursty traffic tends to degrade the performance significantly. Also similar to input queueing is the fact that as the traffic becomes bursty, the mean delay does not degrade as much for large r than for small r. This point is further illustrated in Fig. 9, where we plot the mean delay for various cases with $n = gr$ as a function of l for an offered load of 0.5. As in input queueing, the slopes of the mean delay versus l curves decrease quite rapidly as r increases. It is also interesting to observe that for a fixed number of output ports gr, the difference in mean delay between input queueing and output queueing also decreases as r increases. So, channel grouping tends to decrease the performance gap between the two buffering strategies.

For cases with $r > 1$ and $l > 1$ shown in Fig. 10, the upper bound (dotted lines) is rather close to the exact solution when

$r = 2$, especially at regions of low mean delay. When $r = 4$, the upper bound is not very good at high mean delay. This shows that a switch with channel grouping has significant better delay performance than an $M/G/r$ queue when r is large. In contrast to the upper bound, the lower bound is poor when the mean delay is low. This is not surprising if we recall that the $G/D/r$ queue associated with the lower bound allows all r output lines to be utilized, even when a burst arrives when there are fewer than $r - 1$ packets in the queue. This artificial advantage occurs rather frequently when the offered load is low, but disappears when the buffer occupancy is high. In fact, the lower bound approximates the exact solution better than the upper bound at high mean delay.

V. Conclusions

This paper has quantified the throughput and mean-delay performance of a class of $n \times gr$ asymmetric packet switch modules with channel grouping at the outputs. These switch modules constitute the building blocks of many large switch architectures, and it is important to understand the performance of the switch modules in order to design the large switches properly.

Both input-buffered and output-buffered switch modules have been studied. It is shown that increasing the number of output ports per output address can significantly improve the delay-throughput performance of both buffering strategies, particularly when the ratio of the number of output addresses to the number input ports, g/n, is small. This agrees in principle with the idea originally propounded in the knockout switch [12], [3]. If we fix the line expansion ratio (gr/n), the performance is better for larger r. In other words, decreasing the number of output addresses while fixing the numbers of output and input ports improves the performance. However, reducing the number of output addresses implies reduced switching and, to the extreme that there is only one output address, no switching is performed. Thus, the result simply says one would perform switching to the extent that it is necessary. "Overswitching" not only degrades performance, but also increases switch complexity.

We have also shown that the mean delay performance of both buffering strategies degrades significantly as traffic becomes more bursty, although the maximum allowable throughput of the input-buffered switch module decreases only slightly. In general, however, channel grouping at the outputs tends to decrese the degradation in delay performance due to bursty traffic.

Although output queueing has uniformly better delay-throughput performance than input queueing for all switch dimensions, the advantage of output queueing over input queueing decreases as the switch dimensions become more and more asymmetric (for cases with $n < gr$ as well as $n > gr$). Intuitively, for $gr < n$, the performance limitation is mainly due to line concentration (i.e., fewer output ports than input ports). But this limitation applies to both input and output queueing switch modules. For $gr > n$, the effect of head-of-line blocking on input queueing switch modules is alleviated because of line expansion, and the performance approaches that of output queueing switch modules. In short, $n = gr$ is a special case in which the difference in performance between input and output queueing is the largest. The performance gap between the two buffering strategies also decreases when we increase r and decrease g while keeping gr constant. In fact, the largest performance gap is found in the previously studied case [6], [7] with $n = g$ and $r = 1$.

Finally, some research issues deserve further attention to extend the understanding of input and output queueing strategies in high-speed packet switches.

1) For simplicity, we have assumed that the traffic patterns on different input ports are uncorrelated. Strictly speaking, this is not true when the input ports are also grouped, as in the second stage of the switch architecture shown in Fig. 3. In fact, two packets of the same burst may arrive simultaneously on two input ports of the same group when switch modules with channel grouping are cascaded. It would be interesting to see how the results here need to be modified under this situation.

2) The study of nonuniform traffic distribution in which more packets are destined for some outputs than others also requires further attention. In particular, how would input-buffered and output-buffered switch modules compare with each other under nonuniform but geometrically distributed traffic?

3) When the burst length is not geometric, the maximum throughput of the input-buffered switch module would in general depend on the contention resolution scheme assumed. For instance, when the burst length is deterministic, the optimal strategy is the burst-by-burst service discipline in which we finish serving the packets of one burst before starting on the next burst. In fact, it can be shown that the maximum throughput in this case is the same as that of the uniform random traffic case, for arbitrary burst length l. It is interesting to investigate the sensitivity of our results to the particular bursty traffic model adopted.

Acknowledgment

The authors thank T. Lee for generously sharing his knowledge and expertise with them. This paper has also benefited much from comments by H. Lemberg and the anonymous reviewers.

References

[1] T. Lee, "A modular architecture for very large packet switches," *Conf. Rec., GLOBECOM '89*, vol. 3, pp. 1801–1809, 1989.

[2] S. C. Liew and K. W. Lu, "A 3-stage interconnection structure for very large packet switches," *Conf. Rec., ICC '90*, pp. 316.7.1–316.7.7, 1990.

[3] K. Y. Eng, M. J. Karol, and Y. S. Yeh, "A growable packet (ATM) switch architecture: Design principles and applications," *Conf. Rec., GLOBECOM '89*, pp. 32.2.1–32.2.7, 1990.

[4] H. Suzuki *et al.*, "Output-buffer switch architecture for asynchronous transfer mode," *Conf. Rec., ICC '89*, vol. 1, pp. 99–103, 1989.

[5] A. Pattavina, "Multichannel bandwidth allocation in a broadband packet switch," *IEEE J. Select. Areas Commun.*, vol. 6, no. 9, pp. 1489–1499, Dec. 1988.

[6] M. G. Hluchyj and M. J. Karol, "Queueing in high-performance packet switching," *IEEE J. Select. Areas Commun.*, vol. 6, no. 9, pp. 1587–1597, Dec. 1988.

[7] J. Y. Hui and E. Arthur, "A broadband packet switch for integrated transport," *IEEE J. Select. Areas Commun.*, vol. 5, no. 8, pp. 1264–1273, Oct. 1987.

[8] M. Y. Karol, M. G. Hluchyj, and S. Morgan, "Input versus output queueing on a space-division packet switch," *IEEE Trans. Commun.*, vol. 35, no. 12, pp. 1347–1356, Dec. 1987.

[9] A. Descloux, "Contention probabilities in packet switching networks with strung input processes," in *Proc. ITC 12*, 1988.

[10] L. Kleinrock, *Queueing Systems, Vol. 1: Theory*. New York: Wiley, 1975.

[11] Y. Oie *et al.*, "Effect of speedup in nonblocking packet switch," *Conf. Rec., ICC '89*, vol. 1, pp. 410–415, 1989.

[12] Y. Yeh, M. Hluchyj, and A. Acampora, "The knockout switch: A simple modular architecture for high-performance packet switching," *IEEE J. Select. Areas Commun.*, vol. 5, no. 8, pp. 1274–1283, Oct. 1987.

Soung C. Liew (S'84-S'87-M'87-M'88) was born in Malaysia in 1960. He received the S.B. degree in 1984, the S.M. and E.E. degrees in 1986, and the Ph.D. degree in electrical engineering in 1988, from the Massachusetts Institute of Technology, Cambridge.

From 1984 to 1988, he was a Research Assistant in the Local Communication Networks Group at the M.I.T. Laboratory for Information and Decision Systems, where he investigated fundamental problems in high-capacity fiber-optic networks. He was also a Teaching Assistant for a graduate course on data communication networks. In March 1988, he joined Bellcore, Morristown, NJ, where he is currently a Member of Technical Staff in the Network Systems Research Laboratory. His research interests include high-performance packet switch designs, ATM network control and modeling, and optical network architectures.

Dr. Liew is a member of the Sigma Xi and Tau Beta Pi fraternities.

Kevin W. Lu (S'81-M'85) received the B.S. degree in control engineering from the National Chiao Tung University, Taiwan, in 1979, and the M.S. and D.Sc. degrees in systems science and mathematics from Washington University, St. Louis, MO, in 1981 and 1984, respectively.

In August 1984, he joined Bellcore, Morristown, NJ, where he is currently a Member of Technical Staff in Applied Research. His research interests include modeling, analysis, and optimization for communications network systems and components. His current research activities are related to fiber-optic subscriber loop, network survivability, and broadband packet switches. He was Adjunct Professor at Rutgers Graduate School of Management, Newark, NJ, and Special Lecturer with the Department of Electrical Engineering at Columbia University, New York, NY, in 1989.

Dr. Lu is a member of Sigma Xi and has been active in the Optical Communications Committee of the IEEE Communications Society. He was the recipient of the Bellcore Award of Excellence in 1987 for his work on the technological and market obsolescence of telephone network equipment.