論文 94-19-5-8

ATM 통신망에서의 지연 및 손실 우선순위를 갖는 다중화 알고리즘의 성능 평가

正會員 全 龍 熙*

Performance Evaluation of Multiplexing Algorithms with Both Delay and Loss Priorities in ATM Networks

Yong-Hee Jeon* *Regular Member*

요 약

광대역 종합정보 통신망 (B-ISDN)이 수용할 여러 서비스들은 다양한 지연, 지연 변화 및 셀 손실 확율 요구 사항을 가지고 있다. B-ISDN을 위한 적절한 제어 방법의 설계는 매우 중요하고 어려운 문제이다.

본 논문에서는, 그러한 다양한 요구사항을 만족하기 위하여, 지연 및 손실 우선 순위 모두를 가진 다중화 알고리즘 을 제안한다.

손실 우선 순위의 구현을 위하여, 음성 셀은 폐기 불가한 셀(즉, 높은 우선 순위의 셀)및 폐기 가능한 셀(즉, 낮은 우선 순위의 셀)로 생성됨을 가정하였다. 낮은 순위의 음성셀은 통신망내에서 혼잡이 발생하면 폐기될 수 있다.

본 셀 탈락 방법은 음성 및 데이타의 지연뿐만 아니라 셀 손실도 감소 시키는 것을 보여주고 있다. 이러한 부하 감소 방법은 광대역 종합정보 통신망의 이용율을 크게 개선 시킬 수 있을 것으로 기대된다.

ABSTRACT

The various services that a broadband integrated services digital network (B-ISDN) carries, have a wide range of delay, delay jitter and cell loss probability requirements. Design of appropriate control schemes for B-ISDN is an extremely important and challenging problem.

In this paper, we propose multiplexing algorithms with both delay and loss priorities in order to satisfy the diverse requirements.

* 효성여자대학교 전자계산학과
Hyosung Women's University
論文番號 : 93225
接受日字 : 1993年 11月 22日

For the implementation of cell loss priority, we assumed that voice cells are generated as non-discardable(i.e., high priority) and discardable (i.e., low priotity)cells. The low priority voice cell may be discarded inside the network if congestion occurs.

The cell dropping scheme is shown to reduce cell losses as well as delays for both voice and data. Such a load shedding scheme is expected to improve significantly utilization of B-ISDN.

I. INTRODUCTION

The Asynchronous Transfer Mode (ATM) enables integrated transport of multiple bit rate and bursty traffic types and realizes the gains of statistical multiplexing. ATM thus offers dynamic bandwidth allocation with a fine degree of granularity. The various services that a broadband ISDN (B-ISDN) network carries, have a wide range of delay, delay jitter and cell loss probability requirements (Quality of Service or QOS ruquirements). For example, a voice packet has more stringent delay requirements than data, while a voice packet can tolerate higher loss probability than a data packet. Thus, network operation in order to provide guaranteed levels of QOS with diverse requirements pose real challenge in traffic control problem in ATM networks.

Design of appropriate control schemes for B-ISDN is an extremely important and challenging problem. Some factors which make this a challenging problem are:1) the different time scales of dynamics in the network traffic, 2) different QOS requirements (QOSR) in several services carried in B-ISDN, 3) a small time scales of transient periods in the network, compared with propagation delay. Thus, the feedback from the network is usually outdated and any action the source takes may be too late to resolve congestion. Therefore, to guarantee a desired QOSR, a set of control procedures at the various levels of activity as described in ITU-TS Rec. I.371 [1], is required.

To accommodate diverse QOS requirements, we may nave either : 1) a single ATM cell transfer service mechanism based on network dimensioning in order to meet the most stringent QOS requirement imposed on a certain type of traffic, or 2) priority mechanism to discriminate cell transfer service based on the disparate QOS requirements. The first approach will result in poor utilization of the network resources and may provide QOS more than necessary for a set of traffic types. This approach also makes it difficult to find a proper control algorithm in order to meet the QOS requirements for each traffic type. The priority handling mechanism is more flexible approach and can take advantage of the diverse QOS requiremnents for each traffic type.

In this paper, we propose a mixed scheduling and buffer allocation algorithm in order to satisfy the desired QOS. Such an algorithm can be included in existing or new communication protocols. Typically, the QOSR in B-ISDN can be expressed in terms of *end-to-end* delay and *end-to-end* probability of cell loss. The basic problem is how to break down the end-to-end requirements to node-by-node requirements. In this paper, we consider the case of a single node only. In order to accommodate the diverse QOSR, we propose a mixed scheduling and buffer allocation algorithm. In order to satisfy delay requirements, we propose communication link scheduling algorithms. Buffer allocation algorithms can be effectively used for controlling loss requirements.

Section II describes priority control schems in B-ISDN and investigates some works done for them. In Section III, our traffic model is described. Our proposed priority control schemes are presented in Section IV. Section V depicts

the simulation model and results. Finally, conclusions are included in Section VI.

II. BACKGROUND

2.1 Priority Control Schemes in B-ISDN

The main performance parameters for QOSR in ATM network include cell delay, delay jitter and cell loss.

Our main concern for the delay in this paper is controllable queueing delay occurred inside ATM network nodes. The cell delay jitter is not specifically considered in this paper. The introduction of delay priority will drastically decrease the *end-to-end* delay jitter [2]. Routing can be decided to minimize the propagation delay. In this paper, we are also primarily concerned about the cell losses due to buffer overflows in ATM network nodes (i.e., ATM switch, multiplexer, concentrators). The cell loss probability thus can be defined as the ratio of the number of cell losses (due to buffer overflow) to the sum of the lost and successfully delivered cells.

As we have seen, the various services that a broadband ISDN network carries, have a wide range of delay and loss probability requirements. First-Come-First-Served (FCFS) transmission of cells, that is usually assumed in ATM multiplexing, is not an optimal service discipline from the voice user's perspective, since voice traffic can experience excessive delay when data traffic is heavy. Recent results [3] show that the FIFO scheme does not protect light users in the presence of overload, and it does not protect normal users, to extent that round robin does. Therefore, an efficient *priority service discipline* needs to be devised to guarantee a QOSR of each class for the priority services.

In an ATM network, even though a call is admitted to the network, the QOS requirements may not be guaranteed due to ATM's packer swithing nature. A *conservative* call admission control policy will minimize the probability of cell level congestion. However, this allows relatively low loading on the network and could result in higher level of connection blocking compared to a more *aggressive* call admission control policy. If an aggressive admission policy is adopted, cell level control functions such as link scheduling and buffer allocation algorithms are critical to ensure QOSR.

To meet the diverse delay and loss requirements of each traffic type, we can use priorities between and within service classes. The priority schemes can be used in two ways: one is to use a priority mechanism as a scheduling method, i.e., queueing discipline (we will call this as *priority scheduling*), and the other is to use it as a congestion control method (we will call this as *priority scheduling*). Thus, priority scheduling determines the order of cell transmission while priority discarding determines which cells are dropped when buffer overflow occurs (in push-out scheme) or when the total occupancy of buffer exceeds some threshold (in cell discarding schemes using thresholds). Thus, delay priority can reduce the cell delay jitter of the real-time services. Loss priority enables the network to reduce the loss of critical information of a (data) service.

A lot of studies have been performed in order to evaluate the performance of various time priority mechanisms (for example, see [4,5]). Delay priorities studied in the literature include 1) Head-of-the-line (HOL) policy; 2) limited policy; 3) gated policy; 4) round-robin policy; 5) alternate service policy; and 6) other less significant policies.

The HOL policy is the most common priority in service discipline. In this policy, customers queue according to priority groups and are strictly separated on the basis of groups to which they belong. All queued cells of a higher priority group are transmitted, before a lower priority cell is transmitted. A limited policy limits the maximum number of transmitted cell, to $k(k=1, 2, \cdots)$. In a gated

policy, each class transmits all the cells that were in its queue, when it is its turn to transmit. A round-robin policy transmits cells from each group periodically. It may be limited, gated or alternate service policy. In alternate service policy, the server serves a single cell from one queue, then one cell from the other queue.

Although the dorminant portion of a cell delay in an ATM network is the propagation delay, *a delay priority* discipline may reduce the cell delay jitter and protects the time-sensitive services against instant bursts of other traffic in the network. If the delay requirements of the services of a B-ISDN range from 1ms to a few tens of ms per node, it has been shown [6] that a delay priority discipline could potentially improve network utilization.

Typical *loss priority* mechanisms for systems with two priority classes may be classifed into [7, 8]:

- -Common buffer with pushout mechanism: Cells of both priorities share a common buffer. If the buffer is full and a high priority cell arrires, a cell with low priority (if any is available)will be pushed out and lost.
- -Partial buffer sharing: Low priority cells can only access the buffer if the total buffer occupancy is below a given threshold. High priority cells can access the whole buffer. Partial buffer sharing may be implemented with common buffer [7] or separate buffer. By adjusting the threshold, it is possible to adapt the system to various load situations.
- -Separate buffer scheme: For different priority classes, separate buffers are used. This mechanism is simple to implement. In this scheme, by accommodating each type of traffic in a separate queue, traffic enforcement functions may easily be exercised (i.e., selective cell discarding scheme). We can also limit the buffer size for delay-sensitive class (i.e., voice) in order to limit the maximum delay and larger buffer sizes may be assigned for loss-sensitive class (i.e., data).

2.2 Literature Survey and The Background for The Study

In [9], the authors compared four link scheduling policies: 1) First Come First Served(FCFS), 2) Head of the Line (HOL) Priority, in which real-time packets are given priority, 3) Minimum Laxity Threshold (MLT) Policy, and 4) Queue Length Threshold (QLT) Policy. Their results show that the FCFS policy causes relatively high losses for the real-time traffic, while providing relatively low message delays for the non-real time traffic. In their model, the real-time packets have a fixed real-time constraint (deadline). A real-time packet which is not transmitted by the end of its dealine, is assumed lost and removed from the buffer. They assumed an infinite buffer for non-real time traffic. Hence, the performance metric for the real-time traffic is the percentage of messages lost because of deadlines. The performance metric for the non-real time traffic is average delay.

The HOL policy is shown to reduce the percentage of real-time messages lost at the expense of higher delays for non-real time packets.

In the MLT policy, priority is given to the real-time traffic when the *minimum laxity* of the real-time packets is less than or equal to a threshold; otherwise priority is given to the non-real time traffic. The laxity of a cell is defined as the number of slots remaining before its deadline expires. In the QLT scheme, priority is given to the non-real time traffic when the number of non-real time packets in the queue is above a threshold; otherwise priority is given to the real-time packets.

The MLT may be difficult to implement in an ATM network, since it may require heavy processing at each switching node due to updating the laxity of each real-time packet in every time slot. However, the queue length is easier to work

with. Therefore, it is concluded that QLT is more practical than MLT due to its simpler implementation

Head-Of-Line with Priority Jumps (HOL-PJ) is proposed in [10]. In this scheme, each priority class forms its own queue. When a packet has spent a time in a queue, greater than the local delay limit for that queue, it jumps to the next higher priority queue. In [11], the the HOL priority control mechanism is chosen as the time priority control mechanism.

In the following section Ⅳ, our proposed schemes are presented. In [9], as we have seen, the authors assumed that the real-time packets have a local deadline at a multiplexing node. In our scheme, we don't assume that the real-time traffic has a local deadline. In multi-hop communication networks, even though a real-time traffic has been queued above a local deadline, this traffic can be better served at other nodes. Thus, there are still possibilities that the real-time traffic which exceeded a local deadline, can be reconstructed at the destination node.

Furthermore, our main concern for the non-real time is probability of cell loss, in contrast with the average delay assumed in [9]. In order to implement the loss priortity, we assumed that voice cells are generated as pairs of low and high priority cells, with the low priority ones being discardable.

Our motivation in this paper is to devise an effcient control mechanism for satisfying QOSR of each traffic, based on both delay and loss priority control mechanism. Priority control schemes in B-ISDN are fundamentally related with the QOSR problem and need further research since they are not fully understood. Furthermore, the domestic research results in this field is rare.

In the following section, we present our traffic model used for the simulation study, for voice and data arrival processes.

Ⅲ. TRAFFIC MODEL

A B-ISDN network will carry traffic from a variety of bursty sources. Such sources have different characteristics in terms of bandwidth and QOSR. To model such a network, we have to characterize the cell arrival process to an ATM node. Typically, cells originate from bursty sources and from other ATM nodes. Accurate modeling of the traffic in an ATM network becomes very complex and difficult. The Poisson arrival process was extensively used so far to model non-bursty sources for conventional telecommunication and data networks. It is not appropriate since it can not capture the burstiness which is commonly occured in ATM networks. In this paper, we use Interrupted Bernoulli Process (IBP) in our simulation study.

3.1 IBP Model

In order to capture the burstiness of the arrival process at each buffer, we model it as an IBP. This discrete time process is more suitable to model the current communication systems because the arrival and departure processes occur in a discrete time slot. In IBP process, the chain alternates between the *Active* and *Idle* states. The Active state corresponds to talkspurts(or data transmission), while the Idle one corresponds to silence duration (or no data transmission) in voice (data) traffic modeling. During the active state, cells are generated. No cells are generated during the idle period.

Given that the process is in active state, it will remain in this state with probability p or it will move to the idle state with probability 1-p in the next slot. If the process is in the idle state, it will be still in the idle state with probability q, or it will change to the active state with probability 1-q. (See Figure 1.) In general, if the process is in the active period, the slot will contain a cell with probability α. In this work, we assume that $\alpha = 1$.

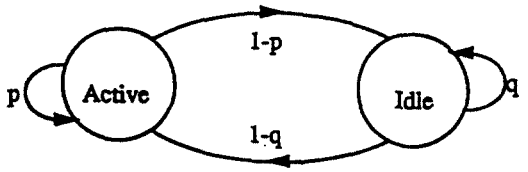

Figure 1. Two-State Markov Chain Diagram of IBP Process

For a geometrically distributed period (active state), arrivals occur according to a Bernoulli cess. This period is followed bv another period (i. e., idle state) which is also geometrically distributed, during which no arrivals occur.

The transition probability matrix of the IBP process is given by:

$$P=\begin{pmatrix} p & 1-p \\ 1-p & q \end{pmatrix} \quad (1)$$

Let π_A, π_I denote the steady-state probability of active and idle states, respectively.
From the steady-state equation $\pi P=\pi$, we can obtain

$$\pi_A = \frac{1-q}{2-p-q}$$

$$\pi_I = \frac{1-p}{2-p-q} \quad (2)$$

We note that π_A is the average bandwidth or average rate of arrivals λ. In a different interpretation, π_A is the probability that any slot is busy (i.e., it carries a cell). It is equal to the mean number of cells generated during the active period. This is equal to the mean length of the active period over the mean length of the silence and active period.

Let d be the interarrival time between successive cells. It can be shown that the generating function, G(z), of the probability distribution of the interarrival time is

$$G(z)\triangleq E\{z^d\} = \frac{z(p(1-zq)+z(1-p)(1-q))}{1-zq} = \frac{z(p+z(1-p-q))}{1-zq} \quad (3)$$

From this generating function we can obtain the mean interarrival time $E\{d\}$ and the second moment of the mean interarrival time $E\{d^2\}$ as follows.

$$E\{d\}=G'(z)|_{z=1} = \frac{2-p-q}{1-q}$$

$$E\{d^2\}=G'(z)|_{z=1}+G''(z)|_{z=1} = \frac{4-3(p+q)+q^2+pq}{(1-q)^2} \quad (4)$$

From the above equations, we can obtain the squared coefficient of variation of the interarrival time, C^2, as

$$C^2 = \frac{Var(d)}{[E\{d\}^2]} = \frac{(p+q)(1-p)}{(2-(p+q))^2} \quad (5)$$

In this paper, we will use the parameter C^2, as a measure of burstiness. Other measures of burstiness can be defined as [12, 13]:

$$\frac{\text{peak bandwidth}}{\text{average bandwdith}} \text{ or } \frac{\text{peak cell rate}}{\text{average cell rate}} \quad (6)$$

However, they are not accurate measures since, for example, two calls with similar peak and average rates have dissimilar traffic characteristics [14]. The squared coefficient of variation of the interarrival time has an advantage in that it readily produces approximations that capture the main qualitative behavior of variability (i.e., burstiness) [15].

Using the same approach as in [15], we approximated the aggregated voice and data arrivals from the large number of sources with two parameters, i.e., the average arrival rate and the squared coefficient of variation of the interarrival

times, C^2. Note that by varying p and q from (2), (5), we can alter the values of the mean arrival rate λ, and of C^2. In particular, we have

Table 1. C^2 in IBP Model

p, q	λ	C^2
$p \to 0, q \to 0$	$\lambda \to 0.5$	$C^2 \to 0$
$p \to 0, q \to 1$	$\lambda \to 0$	$C^2 \to 1$
$p \to 1, q \to 0$	$\lambda \to 1$	$C^2 \to 0$
$p \to 1, q \to 1$	$\lambda \to 0.5$	$C^2 \to \infty$

From Table 1, we can see that as $p \to 1$ and $q \to 1$, $C^2 \to \infty$. Thus, the traffic becomes very bursty. In this paper, the arrival rate λ and C^2 values are assumed given, in order to get the mean cell delay and the mean cell loss probabilities based on some traffic characteristics. With the λ and C^2 values for the arrival processes given, we can easily determine the p and q values as:

$$p = \frac{c^2 - 1 + 3\lambda - 2\lambda^2}{1 - \lambda + c^2}$$

$$q = \frac{1 - 2\lambda - p\lambda}{1 - \lambda} \qquad (7)$$

Ⅳ. THE PROPOSED PRIORITY CONTROL SCHEMES IN ATM MULTIPLEXING

Since the performance objectives for the various services in B-ISDN will be greatly different, we need link scheduling and buffer control mechanisms to satisfy these greatly different performance requirements. The (N1, N2) scheme we proposed in this paper is similar in spirit to the (T1, T2) priority scheme for the wide-band packet network proposed in [16], where a timer is set to limit the maximum transmission time for voice and data traffic.

4.1 Transmission Scheduling Algorithm

In this section, we propose a transmission scheduling algorithm for priority service discipline.

(N1, N2) Scheme: This algorithm falls in the category of *limited* servicing process. N1 (N2) is the maximum number of voice (data) cells which can be transmitted in each cycle. We assume non-preemptice service discipline. That is, once service for the lower priority class has started, higher class cells have to wait until the completion of lower priority traffic transmission before their transmission. Whenever a queue is exhausted, the service is switched to the other class. Thus each traffic type is allowed to use any available bandwidth that is otherwise allocated to the other class. The values of N1, N2 can be adjusted to control delay for voice and data. In (N1, N2) scheme with a voice controller, the threshold(s) is put to block the arriving lower priority voice cells when the voice (data) queue size is greater than or equal to Tv(Td).

In a fixed priority scheme, as long as the higher priority queue is not empty, cells in that class are served. When the higher priority queue becomes empty, lower priotity cells can be served. Most schemes studied so far in the literature assume exhaustive service duscipline for higher priority. This priority scheme will be advantageous for continuous bit rate traffic service since it will always have service priority. However, performance for the lower priority classes may become poor. When there is a large volume of high priority traffic, the delay for the lower priority classes may become intolerably large.

However, in (N1, N2) scheme, cells in lower priority class also have some chance to transmit even if there are higher priority cells in the queue. In this regards, (N1, N2) scheme is a flexible bandwidth allocation algorithm in that whenever one queue is exhausted the transmission is immediately moved over to the other queue if it has a cell waiting to be transmitted. The scheme guarantees bandwidth to voice and data in the proportion of their respective allocations, N1 and N2, i.e., a minimum bandwidth of $N1 \cdot C/(N1+N2)$ for the aggregate voice traffic and $N2 \cdot C/$

(N1+N2) for the aggregate data traffic, where C is the transmission rate of the link.

4.2 Buffer Allocation Algorithm

We propose a simple threshold policy to selectively drop lower priority voice cells. The lower priority cells are only admitted when the total occupancy of buffer is below the threshold(s). This threshold policy is simple to implement and has been shown to give near optimal performance[17].

Voice traffic becomes bursty when it is multiplexed using Digital Speech Interpolation(DSI) or other methods of compression. In DSI, speech activity detection (SAD) is done to detect talkspurts/silence duration. Cells are generated only during talkspurts. Such compression techniques cause voice traffic to become bursty.

In cell discarding (CD) schemes, pairs of cells are generated over two cell formation intervals: the first cell carries the more significant bits of all speech samples in the paired-cells and the second cell carries the less significant bits [18]. The first cell is identified as nondiscardable (i.e., high priority), and the second cell is marked as discardable (i.e., low priority) by appropriately setting the cell loss priority (CLP) field in the ATM header. When this bit is set (i.e., CLP=1), the cell may be discarded inside the network if congestion occurs. The packetization delay will be increased, since cells are now generated in pairs.

In this paper, we propose a voice/data multiplexer in which voice cells are selectively discarded during periods of congestion. All voice packets which are not discarded are served on a FIFO, not a priority basis; hence, both low and high priority cells experience the same average delay.

In the cell discarding scheme, voice quality is expected to degrade gracefully when overloads occur. However, the discarding of low priority cells contributes to improving the performance of high priority (i.e., CLP=0) cells. Such load shedding scheme is necessary in high-speed networks because of the burst nature of traffic. The load shedding provides the network with and inherent resiliency to load surges, and may have only minimal impacts on end services and applications if the load shedding is done selectively [19].

4.3 Multiplexing Algorithm With Both Delay and Loss Priorities

(N1, N2) Scheme with Cell Discarding : In this algorithm, we use both priority service discipline, based on (N1, N2) scheme in order to control mainly delay QOSR and buffer allocation algorithm using a selective cell discarding scheme, in order to control mainly loss QOSR. Therefore, the capacity of a link is allocated according to the (N1, N2) round-robin-like allocation mechansim. The allocation of buffer is done according to the buffer allocation algorithm, as we proposed in the previous section (i.e., threshold policy). In this scheme, we have two cases: 1) when the threshold is imposed only at the voice queue, 2) when the threshold is imposed at both voice and data queues. Voice cells are selectively discarded during periods of congestion. Voice cells are dropped based on its own backlog in case 1) and on backlogs from either its own one or data queue in case 2).

In the next section, using this algorithm, we will simulate an ATM voice and data multiplexer in which voice cells are selectively discarded during periods of congestion.

V. SIMULATION MODEL AND RESULTS

5.1 Model Description

The model proposed in this paper is too complicated to study analytically. Hence we rely on simulations. The simulation model chosen for the ATM/SONET voice and data multiplexing consists of two separate buffers for voice and data, and a single transmission line with or wothout a voice controller. (Figure 2)

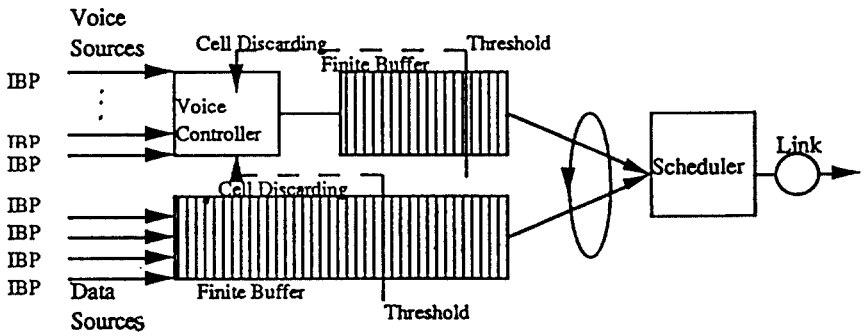

Figure 2. Statistical Multiplexing Schemes with Both Delay and Loss Priorities

In our model, we have simultaneous arrivals from multiple input ports. Therefore, we assume that the multiplexer is fast enough to handle cells even when all the input ports inject the maximum number of cells on each buffer. During the ith slot time, all arriving cells (say c) are admitted to the queue if there are c buffers available. Otherwise, arriving cells which can not find a free space are lost immediately.

For simplicity, a single node is modeled only. This single node model can be used as a good estimator of end-to-end objectives. We have discrete-time queueing systems with N input arrival processes. The server is assumed to be slotted, and the service time is constant and equal to 1 slot time. Service begins and ends at slot boundaries. Each arrival stream is also slotted with a slot equal to a server slot. Without loss of generality, we assume a synchronous operation with slot boundaries between input and output throughout the simulation study. This assumtion is reasonable due to the synchronous operation of SONET frame and the fixed cell size of ATM. We assumed the *come right* in service strategy [20,21]; namely, a cell finding the buffer empty is immediately served and removed from the buffer.

Our model can be characterized as follows:

-Two independent streams of voice and data arrive at each buffer, according to an N-IBP process with mean total arrival rate λ_1 and λ_2, respectively. We assumed here that N=8 for voice and N=4 for data (for the multiple input case).

-The ATM cell size is 53 bytes. The service time is constant (i.e., 2.83 microsec) since the cell length is fixed.

-The transmission link is slotted. One slot contains one cell. The link capacity is SONET STS-3c rate (i.e., 155.52 Mbps). The data rate is 149.76 Mbps.

-Burstiness: We assumed 20 as the C^2 value of voice. (In [15], it is shown that the value of C^2 = 18.1 for the packet arrival process due to a single voice source.) The C^2 values of data are assumed as 1, 20, 50, and 100, in order to investigate the effect of burstiness of data traffic type.

-The switchover time is assumed negligible for the separate buffer scheme, since the switching overhead time was considered the same as accessing one cell from the memory to the next.

-The queue discipline is FIFO in each buffer.

5.2 Results

5.2.1 Single Input Models for Voice and Data

We first compared the performance of three policies for voice and data multiplexing in which we have a single input for each voice and data, in order to check the necessity of separate buffer schemes. The three policies are: 1) First-In-First-Out (FIFO) scheme, 2) FIFO scheme with cell

discarding (CD), and 3) (N1, N2) scheme. In the FIFO statistical multiplexing scheme with single input for each voice and data, cells are served in the order of arrival. Cells arriving in the same slot are transmitted in random order. In FIFO scheme with cell discarding, we assumed that voice cells are generated with pairs of higher priority and lower priority cells as described in Section Ⅳ. To find out how the cell discarding scheme works in the FIFO scheme, a threshold is imposed on the buffer (we arbitrarily used 300 as the threshold value).

Figure 3 is provided to compare the mean cell delay of those three control schemes with a fixed voice load at 0.3 of total link load. The voice buffer size is assumed as 200 while data buffer size is assumed to have 1000. (95% confidence intervals for the simulations are calculated. Since they are too small to depict, they are not shown in Figure 3.)

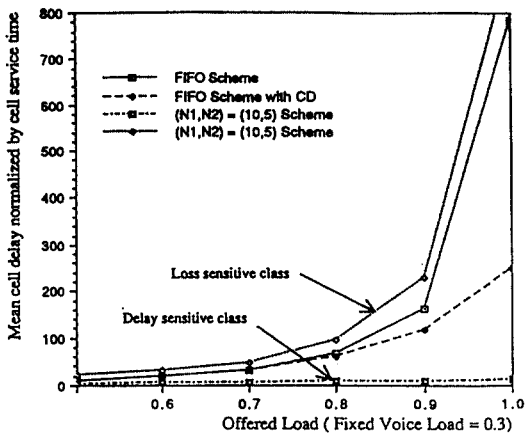

Figure 3. Mean Cell Delay Comparison, $C^2 = 50$, Fixed Voice Load = 0.3

In this figure, it is shown that the delay performance for the delay sensitive class in (N1, N2) scheme is much better than that of the FIFO scheme. This is because N1 is much larger than the data transmission allocation N2, in the (N1, N2) scheme. Figure 3 shows that the delay performance for the delay sensitive class is not much affected by the increased data traffic amount. Hence, in (N1, N2) scheme, voice traffic is well protected when data traffic causes overload whereas the voice delay rapidly increases under the FIFO scheme. In FIFO scheme, we also found that when the data traffic was more bursty the voice performance degradation was much worse. In FIFO scheme with cell discarding, the delay performance shows improvement as offered load increases.

5.2.2. Multiple Input Models for Voice and Data

5.2.2.1 (N1, N2) Scheme without CD

Table 2 depicts the effect of (N1, N2) value on voice delay without the cell discarding scheme.

Table 2. The effect of (N1, N2) value on voice cel performance (fixed voice load : 30%)

C^2	% Load	$(N_1, N_2)=(10, 5)$	
		Mean Delay	Max. Delay
50	60	19.26	622.60
	70	24.67	812.21
	80	29.64	826.36
	90	34.94	826.36
	100	36.47	826.36
100	60	19.34	605.62
	70	19.34	605.62
	80	29.09	815.04
	90	30.26	815.04
	100	31.60	817.87

C^2	% Load	$(N_1, N_2)=(10, 10)$		
		Mean Delay	Max. Delay	Cell Loss Prob.
50	60	28.30	891.45	0
	70	41.14	931.07	4.3×10^{-6}
	80	55.47	984.84	1×10^{-5}
	90	72.70	1129.17	2.3×10^{-5}
	100	78.04	1129.17	2.5×10^{-5}
100	60	30.87	1129.17	1.8×10^{-6}
	70	30.87	1129.17	1.8×10^{-6}
	80	56.35	1129.17	6.7×10^{-6}
	90	59.94	1129.17	6.7×10^{-6}
	100	63.99	1129.17	1.5×10^{-5}

This table clearly shows that the maximum voice cell delay becomes more than 1 msec with (N1, N2)=(10, 10) when C^2 = 100 for data. Table 2 also provides cell loss probabilities when (N1, N2) =(10, 10). There is no cell loss when (N1, N2)= (10, 5) in this case. Note that for fixed N1, N2 values, the maximum delay saturates when the offered load exceeds some threshold value; this is because in that case, maximum cell delay depends primarily on the buffer size.

5.2.2.2 (N1, N2) Scheme with CD

We experimented with the cell discarding scheme for two cases: 1) when the threshold is imposed only at the voice queue, 2) when the threshold is imposed at both voice and data queue, in order to investigate how we can trade off the performance achieved by each traffic class, by adjusting the threshold parameters.

A) Threshold only at the voice queue

Table 3 provides the effect of dropping low priority cells on voice delay when (N1, N2) = (10,5). We can see that for the same conditions, voice performance is considerably improved in terms of mean and maximum delay compared with the no cell discarding schemes in Table 2. The improvement is more noticeable in maximum delay.

Table 3. The effect of dropping low priority cells on voice cell delay (time: μsec, fixed voice load : 30%)

C^2	% Load	(50, 200)	
		Mean Delay	Maximum Delay
50	60	18.37	331.11
	70	23.27	390.54
	80	27.71	390.54
	90	32.60	393.37
	100	33.87	393.37
100	60	18.25	348.09
	70	18.25	348.09
	80	27.31	348.09
	90	28.41	353.75
	100	29.62	353.75

The probability of dropping low priority voice cells for the (N1, N2) scheme, when the *voice* load is fixed at 30% is provided in Table 4; it is provided in Table 5, when the *data* load is fixed at 30%. In these tables, Tv is threshold parameter at the voice buffer and K is the buffer size for voice.

We can observe a higher cell loss probability in Table 5 than Table 4, since voice traffic is significant. The cell loss probability decreases as the value of the threshold, Tv, increases; since the lower priorty voice cell will have less chance of being dropped in higher threshold value.

Table 6 provides the effect of threshold value Tv on mean and maximum voice cell delay.

As we expected, both the mean and maximum delay are lower, with lower threshold value. The difference of mean cell delays between the schemes with CD and without CD increases as load increases.

The effect of threshold value on mean data cell delay, with varying voice load is shown in Figure 4, when (N1, N2) = (10, 5).

In this figure, K represents the voice buffer size. This figure shows that there is much decrease in data delay with cell discarding when a significant source of traffic is voice. We also observed the decrease in cell loss probability for data with voice cell discarding.

B) Thresholds at both voice and data queue

Figure 5 shows that we have a big improvement in loss QOS requirments, when a threshold is imposed on the data queue as well as on the voice queue.

In this figure, C^2 = 100 for data and (N1, N2)= (10,5) are assumed. In this figure, K_1 represents the voice buffer size and K_2 represents the data buffer size. It is seen in this figure that the cell delay and loss QOSR for data can be significantly improved by imposing the threshold at data buffer so that data can influence cell dropping controller based on its own backlog. The threshold values at both voice and data buffers may be

Table 4. Probability of dropping low priority voice cells for (N1, N2) scheme ($(N_1, N_2)=(10, 5)$, data buffer size=1000, fixed voice load : 30%)

C^2	% Load	$(T, K)=(100, 200)$		$(T, K)=(50, 200)$	
		Cell Loss Pr.	95% C.I	Cell Loss Pr.	95% C.I
50	60	2.3×10^{-4}	$\pm1.1\times10^{-4}$	5.5×10^{-3}	$\pm5.8\times10^{-4}$
	70	3.5×10^{-4}	$\pm1.4\times10^{-4}$	7.6×10^{-3}	$\pm7.5\times10^{-4}$
	80	4.5×10^{-4}	$\pm1.8\times10^{-4}$	9.3×10^{-3}	$\pm8.8\times10^{-4}$
	90	6.1×10^{-4}	$\pm2.1\times10^{-4}$	1.1×10^{-2}	$\pm1.0\times10^{-3}$
	100	6.3×10^{-4}	$\pm2.1\times10^{-4}$	1.2×10^{-2}	$\pm1.1\times10^{-3}$
100	60	1.7×10^{-4}	$\pm9.9\times10^{-5}$	5.3×10^{-3}	$\pm5.4\times10^{-4}$
	70	1.7×10^{-4}	$\pm9.9\times10^{-5}$	5.3×10^{-3}	$\pm5.4\times10^{-4}$
	80	3.6×10^{-4}	$\pm1.6\times10^{-4}$	9.1×10^{-3}	$\pm8.3\times10^{-4}$
	90	4.4×10^{-4}	$\pm1.9\times10^{-4}$	9.5×10^{-3}	$\pm8.6\times10^{-4}$
	100	4.6×10^{-4}	$\pm1.9\times10^{-4}$	9.9×10^{-3}	$\pm8.7\times10^{-4}$

Table 5. Probability of dropping low priority voice cells for (*N1, N2*) scheme ($(N_1, N_2)=(10, 5)$, data buffer size=1000, fixed data load : 30%)

C^2	% Load	$(T, K)=(100, 200)$		$(T, K)=(50, 200)$	
		Cell Loss Pr.	95% C.I	Cell Loss Pr.	95% C.I
50	60	2.3×10^{-4}	$\pm1.1\times10^{-4}$	5.5×10^{-4}	$\pm5.8\times10^{-4}$
	70	9.5×10^{-4}	$\pm2.5\times10^{-4}$	1.5×10^{-2}	$\pm1.1\times10^{-3}$
	80	5.0×10^{-3}	$\pm5.1\times10^{-4}$	3.7×10^{-2}	$\pm1.8\times10^{-3}$
	90	2.1×10^{-2}	$\pm1.3\times10^{-3}$	7.8×10^{-2}	$\pm2.9\times10^{-3}$
	100	6.4×10^{-2}	$\pm2.8\times10^{-3}$	1.4×10^{-1}	$\pm4.4\times10^{-3}$
100	60	1.7×10^{-4}	$\pm9.9\times10^{-5}$	5.3×10^{-3}	$\pm5.4\times10^{-4}$
	70	1.1×10^{-3}	$\pm2.2\times10^{-4}$	1.5×10^{-2}	$\pm1.2\times10^{-3}$
	80	5.3×10^{-3}	$\pm5.4\times10^{-4}$	3.8×10^{-2}	$\pm1.9\times10^{-3}$
	90	2.3×10^{-2}	$\pm1.5\times10^{-3}$	8.1×10^{-2}	$\pm3.1\times10^{-3}$
	100	7.0×10^{-2}	$\pm3.2\times10^{-3}$	1.5×10^{-1}	$\pm4.8\times10^{-3}$

Table 6. The effect of dropping low priority voice cells on voice delay in (*N1, N2*) scheme ($(N_1, N_2)=(10, 5)$, fixed data load : 30%)

C^2	% Load	Without CD		With CD			
				(100,200)	(100,200)	(50,200)	(50,200)
		Mean	Max.	Mean	Max.	Mean	Max.
100	60	19.34	605.62	19.23	492.42	18.25	348.09
	70	30.87	815.04	30.31	543.36	26.94	365.07
	80	54.09	846.17	50.22	588.64	40.00	441.48
	90	101.13	846.17	82.73	693.35	56.36	486.76
	100	205.93	846.17	130.78	846.17	76.04	730.14

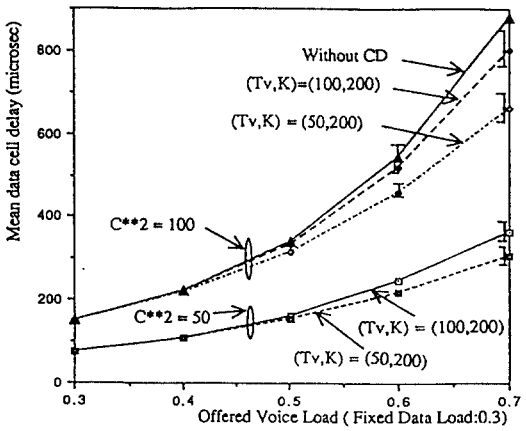

Figure 4. Effect of Threshold Value on Mean Data Cell Delay

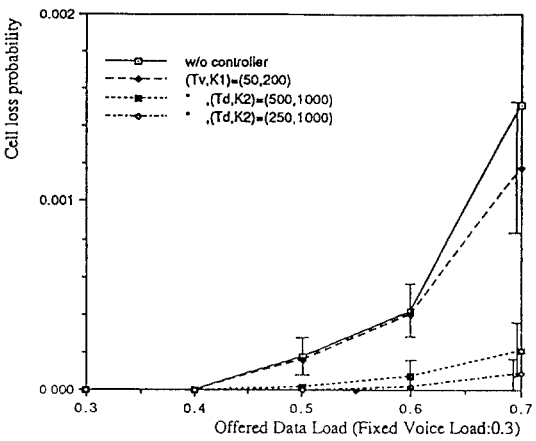

Figure 5. The Effect of Imposing Threshold on Data Buffer on Mean Data Cell Loss, (N1, N2) = (10,5), $C^2 = 100$

tuned to meet certain QOSR. For example, when voice traffic volume is very low relative to data, then threshold value at data buffer can be set to the (maximum) buffer size to protect voice traffic from data congestion.

VI. CONCLUSIONS

In this paper, we defined our desired QOSR. They are *delay* QOSR and *loss* QOSR. Typically, the QOSR is expressed in terms of end-to-end delay and loss in B-ISDN. However, to simplify our notation, we handled a single node only. Of course, if every individual node QOSR is satisfied, the end to-end QOSR will also be satisfied. We have proposed statistical multiplexing schemes with both delay and loss priorities. To implement both priorities, we have proposed separate buffers for delay sensitive and loss sensitive traffics such that we can efficiently control QOSR for the two types of traffic. The capacity of a link is shared between voice and data according to an (N1, N2) round robin-like allocation mechanism. The scheme is dynamic in that it allows the different traffic classes or VPs (Virtual Paths) /VCs(Virtual Channels) to share the bandwidth with a soft boundary. This scheme may also be used to combine constant-bit-rate (CBR) services with variable-bit-rate traffic, by adjusting the allocated bandwidth. In order to implement the loss priority, we assumed that voice cells are generated as pairs of low and high priority cells, with the low priotity ones being discardable. The buffer allocation strategy is based on a simple threshold policy.

Our simulation results show that the cell dropping controller, along with the (N1, N2) allocation scheme works effectively to meet the disparate QOSR for voice and data. The cell dropping scheme is shown to reduce cell losses as well as delays for both voice and data. It is found that the performance for data can significantly improved by imposing thresholds at both voice and data buffers.

In high-speed networks, the packet arrival processes are expected to be bursty. Such arrival processes are not adequately characterized by simple statistics, such as arrival rates. As a consequence, static algorihms that rely on knowledge of such paramenters are expected to perform poorly in a high-speed environment. We have therefore to resort to robust dynamic algorithms, that

do not ruquire knowledge of these statistics. Our algorithms based on the threshold parameters will be very efficient and very practical in such a high-speed network environment. They will also allow the network designer to explicitly tradeoff the performance achieved by each trffic class, by adjusting these threshold parameters.

REFERENCES

1. ITU-TS Rec. I.371: 'Traffic control and congestion control in B-ISDN', Geneva, June 1992.
2. Annie Gravey and Gerald Hebuturne, "Mixing time and loss priorities in a single server queue," Queueing, Performance and Control (ITC-13), pp. 47-52, Copenhagen, Denmark, 1991.
3. Samuel P. Morgan, "Queueing disciplines and passive congestion control in byte-stream networks," IEEE Trans. on Comm., vol. 39, no. 7, pp. 1097-1106, July 1991.
4. L. Kleinrock, Queueing Systems, Vol. 2. New York: John Wiley and Sons, 1976.
5. Hideaki Takagi (Editor), "Queueing analysis of polling models: An update," Stochastic Analysis of Computer and Comm. Systems, North-Holland, pp. 267-318, 1990.
6. D.P.K. Hsing, "Simulation and performance evaluation of ATM multiplexer using priority scheduling," IEEE J. on Selected Areas in Comm., pp. 418-427, April 1991.
7. H. Kroener, G. Hebuterne, P. Boyer, and A. Gravey, "Priority management in ATM switching nodes," IEEE J. on Selected Areas in Comm., pp. 418-427, April 1991.
8. Arthur Y.-M. Lin and John A. Silvester, "Priority queueing strategies and buffer allocation protocols at an ATM integrated broadband switching system," IEEE J. on Selected Areas in Comm., vol. 9, no, pp. 1524-1536, Dec. 1991.
9. Renu Chipalkatti, James F. Kurose, and Don Towsley, "Scheduling policies for real-time and non-real-time traffic in a statistical multiplexer," Conf. Proc. of INFOCOM, pp. 774-783, 1989.
10. Y. Lim and J. Kobza, "Analysis of a delay-dependent priority discipline in a multiclass traffic packet switching node," IEEE INFOCOM, pp. 889-898, 1988.
11. Jae Shin Jang, Byung Cheol Shin, and Kwon Chul Park, "Performance Analysis of ATM Switch with Priority Control Mechanisms," The J. of The Korean Inst. of Comm. Sciences, Vol. 18, No. 8, pp. 1190-1200, Aug. 1993.
12. C.A. Cooper and Kun I. Park, "Toward a broadband congestion control strtegy", IEEE Comm. Magazine, vol. 4, no. 3, May 1990.
13. G. Rigolio and F. Fratta, "Input rate regulation and bandwidth assignment in ATM networks: an integrated approach", ITC-13, Copenhagen, Denmark, pp. 123-128, 1991.
14. Duke Hong and Tatsuya Suda, "Congestion control and prevention in ATM networks", IEEE Network Magazine, vol 5, no. 4, pp. 10-16, July 1991.
15. K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," IEEE Journal on Selected Areas in Comm. (SAC), VOL sac-4, pp. 833-846, Sep. 1986.
16. K. Sriram, "Dynamic bandwidth allocation and congestion control schemes for voice and data multiplexing in Wideband packet technology," ICC, pp. 324.3.1, 1990.
17. K. Bala, I. Cidon, and K. Sohraby, "Congestion control for high speed packet switched networks," IEEE INFOCOM, pp. 520-526, 1990.
18. K. Sriram, R.S. McKinney, and M.H. Sherif, "Voice packetization and compression in broadband ATM networks," IEEE J. on Selected Areas in Comm., pp. 294-304, April 1991.
19. Adrian E. Eckberg, Bharat T. Doshi, and Richard Zoccolillo, "Controlling congestion in B-ISDN/ATM:Issues and strategies," IEEE Comm. Magazine, pp. 64-70, Sep. 1991.

20. J.F.Hayes, Modeling and Analysis of Computer Communications Networks. New York: Plenum, 1984.
21. Masayuki Murate, Yuji Oie, Tatsuya Suda, and Hideo Miyahara," Analysis of a discrete-time single server queue with bursty inputs for traffic control in ATM networks," IEEE J. on SAC, vol. 8, no. 3, pp. 447-458, April 1009.

全 龍 熙(Yong-Hee Jeon) 정회원
1953년 4월 27일생
1978년 2월 : 고려대학교 전기공학과 졸업(공학사)
1989년 8월 : 미국 노스캐롤라이나 주립大 Elec. and Computer Eng.(MS)
1992년 12월 : 미국 노스캐롤라이나 주립大 Elec. and Computer Eng.(Ph.D.)
1978년 1월 ~ 1978년 11월 : 삼성중공업(주) 근무
1978년 11월 ~ 1985년 7월 : 한국전력기술(주) 근무
1989년 1월 ~ 1992년 9월 : 미국 노스캐롤라이나 주립大 부설 CCSP(Center For Comm. & Signal Processing) 연구원
1992년 10월 ~ 1994년 2월 : 한국전자통신연구소 선임연구원
1994년 3월 ~ 현재 : 효성여자대학교 전자계산학과 교수
※주관심분야 : 컴퓨터네트워크, 광대역통신망, ATM 통신망 제어 및 성능 분석