Dynamic Priority Protocols for Packet Voice

THOMAS M. CHEN, STUDENT MEMBER, IEEE, JEAN WALRAND, MEMBER, IEEE, AND DAVID G. MESSERSCHMITT, FELLOW, IEEE

***Abstract*—Since the reconstruction of continuous speech from voice packets is complicated by the variable delays of packets through the network, delay variability is an important measure of performance for voice packet networks. A *dynamic priority* protocol designed to minimize the variability of packet delays is proposed. The protocol allows the priority of a packet to vary with time. Two examples of dynamic priorities are studied by means of queueing analysis and simulations. Some optimal properties are proven. Simulations indicate that under certain conditions, dynamic priorities could effectively reduce the delay variability.**

I. Introduction

SPEECH communications have historically been handled by analog, circuit-switched networks such as the public telephone network. However, packet switching is becoming increasingly attractive for voice as well as data for a number of reasons [1]. First, the telephone system is in the process of evolving into the integrated services digital network (ISDN) which is an all-digital network offering integrated user access to a wide range of voice and data services [2], [3]. Implementation of ISDN will involve packet switching as well as circuit switching facilities [4], [5]. Second, packet switching technology has matured over two decades of development of packet-switched data networks (e.g., ARPANET) [6]–[8]. Recently, *fast* or *wide-band* packet switching has emerged as a promising long-term technology for integrated services [9]–[15].

Moreover, packet switching offers several potential advantages in terms of performance. One advantage is efficient utilization of channel capacity, particularly for "bursty" traffic. Although not as bursty as interactive data, speech exhibits some burstiness in the form of talkspurts [16]. Average talkspurt durations depend on the sensitivity of the speech detector, but it is well known that individual speakers are active only about 35–45 percent in typical telephone conversations. By sending voice packets only during talkspurts, packet switching offers a natural way to multiplex voice calls as well as voice with data. Another advantage is that call blocking can be a function of the required average bandwidth rather than the required peak bandwidth. In addition, packet switching is flexible. For example, packet voice is capable of supporting point-to-multipoint connections and priority traffic. Furthermore, since packets are processed in the network, network capabilities in traffic control, accounting, and security are enhanced.

However, packet voice is not without difficulties. Continuous speech of acceptable quality must be reconstructed from voice packets that experience variable delays through the network. The reconstruction process involves compensating for the variable delay component by imposing an additional delay. Hence, packets should be delivered with low average delay and delay variability. This paper studies a dynamic priority queueing discipline designed to minimize delay variability. The effect of delay variability on speech reconstruction is discussed in Section II. The concept of dynamic priorities is described in Section III. Two examples of dynamic priorities are studied by means of queueing analysis and simulations. Some optimal properties are proven in Section IV, and simulation results are discussed in Section V.

II. Packet Voice

A. Packet Voice Protocols

Generally speaking, the network protocols developed for packet-switched data are not suitable for speech due to the different natures of speech and data. Unlike data, the nature of speech is subjective and conversational. Speech can tolerate a certain amount of distortion (e.g., compression, clipping) but is sensitive to end-to-end delay. Although the exact amount of maximum tolerable delay is subject to debate, it is generally accepted to be in the approximate range of 100–600 ms. The public telephone network, for example, has a maximum specification of 600 ms [17].

In order to minimize packetization and storage delays, it has been proposed that voice packets should be relatively short, on the order of 200–700 bits, and generally contain less than 10–50 ms of speech [18]–[22]. Network protocols should be simplified to shorten voice packet headers (e.g., on the order of 4–8 bytes), although timestamps and sequence numbers are likely needed. The need for timestamps is discussed in the next section. Since a certain amount of distortion is tolerable, error detection, acknowledgments, and retransmissions are unnecessary in networks with low-error rates. Flow control can be exercised end-to-end by blocking calls and throttling the speech encoder. In addition, network switches can possibly discard packets under heavy traffic conditions. In

Manuscript received December 5, 1988; revised March 5, 1989. This work was supported by a grant from Pacific Bell with a matching grant from the University of California MICRO Program.

The authors are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720.

IEEE Log Number 8928088.

0733-8716/89/0600-0632$01.00 © 1989 IEEE

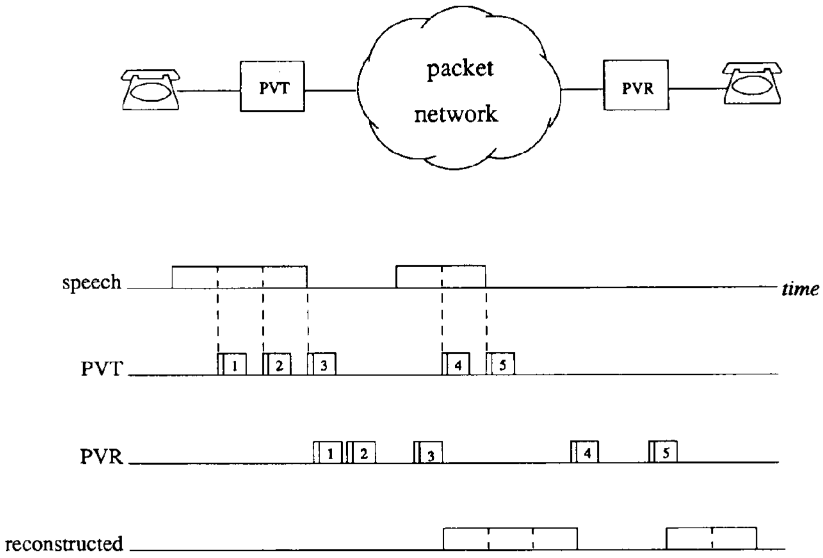


Fig. 1. Packet voice.

this case, embedded coding has been proposed whereby speech quality degrades gracefully with the loss of information [23], [24].

B. Speech Reconstruction

Although average delays and delay variability are reduced by the combination of simplified protocols, short packets, and fixed routing, delay variability could be significant in a long-haul network [25]. The variable delay component is primarily due to queueing delays which tend to increase with the traffic load. If delay variability is significant, an important problem is the reconstruction of continuous speech of acceptable quality from voice packets [25], [26]. As shown in Fig. 1, packets are generated at regular intervals during talkspurts at the *packet voice transmitter* (PVT). The reconstruction process at the *packet voice receiver* (PVR) must compensate for the variable delay component by adding a controlled delay before playing out each packet. This is constrained by some value, $D_{\max}$, the specified maximum allowable end-to-end delay, and P_{loss}, the specified maximum percentage of packets that can be "lost" or miss playout. In addition to buffering voice packets, it might be desirable for the PVR to attempt to detect lost packets and recover their information.

There are two basic approaches to the reconstruction process [25]–[27]. In the *null timing information* (NTI) scheme, reconstruction does not use timing information (i.e., timestamps) to determine packet delays through the network. The PVR adds a fixed delay D (e.g., 40 ms) to the first packet of each talkspurt, as shown in Fig. 2. If D_0 is the transit delay of a first packet through the network and D_g is a packet generation time (assumed to be constant), then the total delay of the first packet from entry into the network to playout is

$$D_t = D_0 + D. \tag{1}$$

Subsequent packets in the talkspurt are played out at intervals of D_g after the first packet, and therefore sequence numbers are required to indicate the relative positions of packets in the talkspurt. If a packet is not present at the PVR at its playout time, it is considered "lost." The choice of D involves a tradeoff. Increasing D reduces the percentage of lost packets but increases total end-to-end delays and the size of the queue at the PVR. D cannot be too large due to the constraint from $D_{\max}$ nor too small due to P_{loss}.

Since D_0 is random, the silence intervals between talkspurts are not reconstructed accurately. In the example shown in Fig. 3, let d and d' denote the values of D_0 for the talkspurts preceding and following a silence interval. Suppose that d and d' are identically distributed with variance σ^2 and have some positive correlation r. Then the error in the length of the reconstructed silence is $\epsilon = d' - d$ and has variance var $(\epsilon) = 2\sigma^2(1 - r)$ which is directly proportional to the variance of packet delays. Evidently, the NTI scheme would be adequate only if a small delay variance could be guaranteed. Also, since the scheme depends on the first packet of each talkspurt, the loss of a first packet might cause confusion at the PVR.

If delay variability can be significant, a more elaborate reconstruction process is necessary. In the *complete timing information* (CTI) approach, the reconstruction process uses full timing information in the form of timestamps to accurately determine each packet's delay through the network, denoted D_v. As shown in Fig. 4, the PVR adds a controlled delay D_r so that the total entry-to-playout delay

$$D_t = D_v + D_r \tag{2}$$

is as uniform as possible for all packets. In addition to timestamps, sequence numbers are also desirable for detecting lost packets.

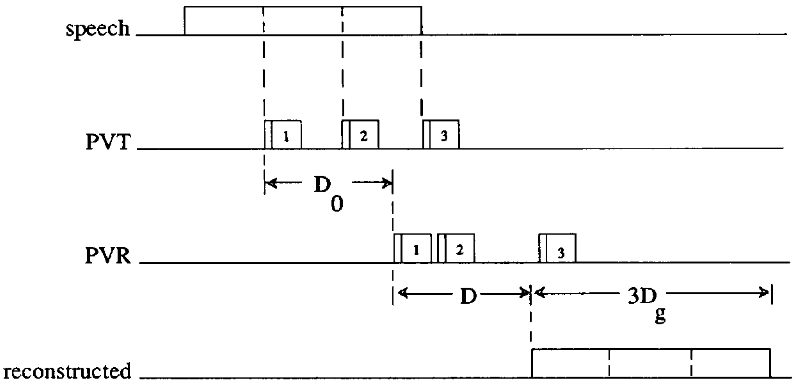


Fig. 2. NTI reconstruction scheme.

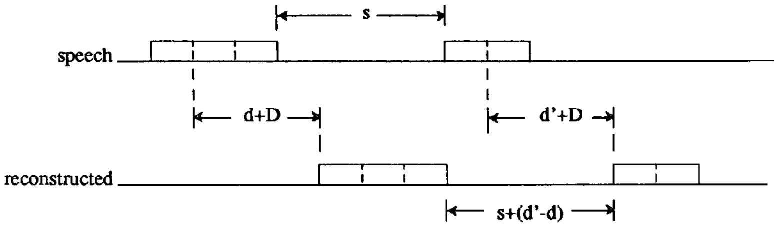


Fig. 3. Reconstruction of silences in NTI scheme.

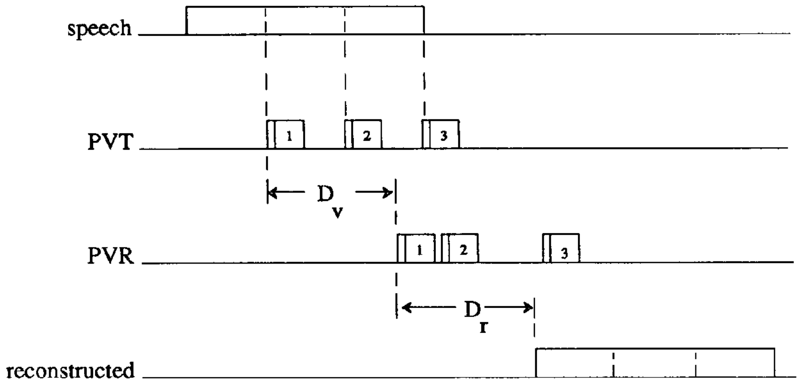


Fig. 4. CTI reconstruction scheme.

There are various choices for the format of the timestamp field. The most obvious choice is a global timestamp but this requires precise synchronization of both PVT and PVR to a global clock. A second choice is to encode the relative time between consecutive packets, but this means there is an unknown constant end-to-end delay. A large timestamp field is also required because the time between packets could be long. Another problem is the possible confusion created by lost packets, although this could be remedied by the use of sequence numbers.

Finally, the timestamp can indicate the delay that a packet has accumulated in transit so far [25]. In this case, the timestamp might be more appropriately called a *delay-stamp*. A packet is generated with a delay-stamp initialized to zero. Each node increments the stamp by the amount of time that the packet has spent in that node, possibly including propagation delays along links as well. At the PVR, the delay-stamp reveals the total transit delay (and hence generation time) of each packet. Alternatively, a packet can be generated with its maximum allowable delay, and each node decrements the stamp. The delay-stamp avoids the need for synchronization of clocks at both ends. A disadvantage of the delay-stamp is the extra processing and accounting required for each packet at each node.

Assuming that packet delays can be determined, the next task of the PVR is to choose the target playout times, i.e., the desired value of D_t. The choice of D_t involves a tradeoff. It should be small for subjective reasons and to reduce the queue at the PVR. On the other hand, the percentage of late or "lost" packets should be less than P_{loss}, and therefore D_t should be large (up to $D_{\max}$). The ob-

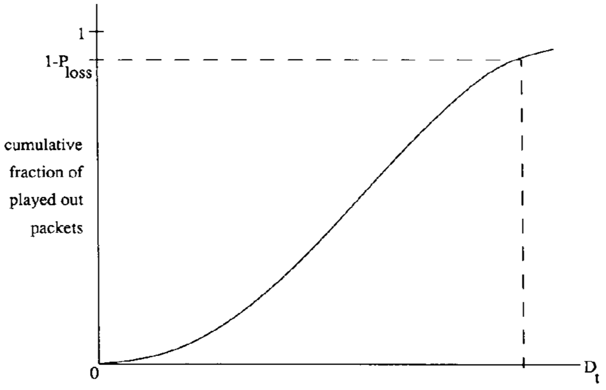


Fig. 5. Tradeoff in delay and lost packets.

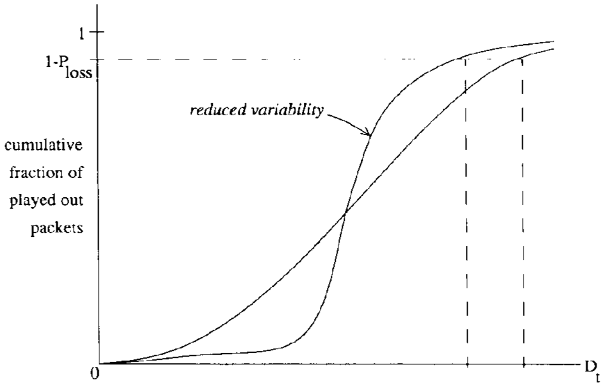


Fig. 6. Effect of reduced variability on tradeoff.

vious choice for D_t is the value at which the percentage of lost packets is exactly P_{loss}, as shown in Fig. 5.

III. Dynamic Priority Protocols

As seen from the previous section, delay variability complicates the speech reconstruction process and it is therefore undesirable. Reduced variability also implies a steeper tradeoff curve and hence a smaller desired D_t for the same average packet delay, as shown in Fig. 6. Although delay variability is somewhat reduced when overall delays are reduced and the actual amount of delay variability expected in fast packet networks is unknown at this time, the choice of queueing discipline allows another degree of control over delay variability. The *first-come-first-serve* (FCFS) discipline has usually been assumed although there is no particular reason to believe it is optimal for voice packet networks.

This paper studies a *dynamic priority* discipline for minimizing delay variability. The concept of dynamic priority is presented here in an intuitive manner. Under the usual FCFS discipline, some packets experience long transit delays while other packets experience much shorter delays. That is, some packets find long queues at each node while other packets find short queues. In order to control the variability of queueing delays to a greater degree, a queueing discipline different than the usual FCFS discipline is required. This queueing discipline must be capable of giving preference to those packets that are determined to be more "urgent," where urgency depends on factors such as a packet's present position, route, age, and expiration time. It necessarily involves time-varying priorities.

Such a queueing discipline was first described by Jackson [28]–[30]. In Jackson's "dynamic priority" queueing model, a customer arrives at a queue with an "urgency number" u. The priority of a customer in the queue is the sum of its urgency number and its waiting time in the queue so far. The arrival joins the queue behind those customers with priority greater than or equal to u and ahead of those with priority less than u. This model recognized the possibility that a customer's "urgency" (and hence priority) increases with its waiting time. In this sense, the priority is "dynamic." This model has been analyzed by Goldberg [31], [32].

A different version of dynamic priority was proposed by Kleinrock [33]. In Kleinrock's model, customers belong to one of P classes. The priority of a class k customer who arrived at instant τ at a queue is defined to be $b_k(t - \tau)$ at any time $t \geq \tau$ where $0 \leq b_1 \leq b_2 \leq \cdots \leq b_P$. Priority increases linearly with waiting time beginning at zero upon arrival, and the rate of increase is different for each class. Kleinrock's model has subsequently been elaborated to include linearly decreasing priorities [34]–[37] and concave increasing priorities [38].

Another dynamic priority discipline, called *head-of-the-line with priority jumps* (HOL-PJ), was proposed by Lim and Kobza [39] for packet switching multiple classes of delay-sensitive traffic. They assume C classes of packets with separate queues for each class. Packets are FCFS within each queue, while queue i has nonpreemptive priority over queue j if $i < j$. However, packets have a maximum limit on the queueing delay at each queue. When the maximum queueing delay is exceeded, that packet moves to the end of the next higher priority queue. Thus, the queueing delay at each node will be limited by a bound depending on the class of the packet.

Here we consider "dynamic priorities" to mean all time-varying priority disciplines. "Delay dependent priority" is a less precise term because it implies that priority is a function only of waiting time (and not of other factors such as route or traffic conditions). Compared to the usual *static* priority or *head-of-line* (HOL) discipline where priorities remain constant, the dynamic priority discipline permits a more general application to queueing networks. In this paper, two examples of dynamic priority disciplines are studied: *oldest-customer-first* (OCF) and *earliest-deadline-first* (EDF).

IV. Queueing Analysis

A. OCF Discipline

Under the FCFS discipline, some packets experience short transit delays because they find short queues at each node, while other packets find long queues and experience long delays. Intuitively, it seems that this situation could be improved by giving preference to those packets which have already spent more time in the network. This is the reasoning behind an *oldest-customer-first* (OCF) discipline which defines the priority of a customer (packet) to be equal to its ''age'' or the time it has spent so far in the network. It implicitly assumes that the urgency of a packet increases with its transit time in the network. As shown in Fig. 7, the age of a packet arriving to an OCF queue is determined from its timestamp (e.g., delay-stamp) and compared to the ages of packets already in the queue. The arrival is inserted behind older packets and ahead of younger packets.

Since the OCF discipline assigns higher priority to older packets, they should expect shorter waiting times in queue. The proposition below proves that Poisson arrivals with i.i.d. ages experience waiting times that are negatively correlated with their ages.

Proposition 1: Consider a work-conserving, nonpreemptive M/G/1 queue with P classes of customers. The arrival process of class k customers is Poisson with mean rate λ_k and independent of other arrivals. A class k customer arrives with random age α which is i.i.d. according to a probability distribution function $P_k(\alpha)$. The service times of class k customers are i.i.d. according to a distribution $B_k(x)$. If all customers are serviced on an OCF basis, then a customer's waiting time in the queue is negatively correlated with its age upon arrival.

Proof: Appendix A.

Whereas an assumption of Poisson arrivals might be questionable in a voice packet network, it is reasonable to assume deterministic service times. Voice packets are generated at constant intervals during talkspurts. Hence, voice packets are equal length except possibly the last packet in each talkspurt and those packets shortened by flow control measures. Assuming average talkspurts of 1.2 s and packet generation times of 50 ms, less than 4 percent of the packets occur at the end of a talkspurt. If we further assume that flow control is exercised infrequently, it implies that nearly all voice packets will be equal length. For equal length packets, the theorem below proves certain optimal properties of the OCF discipline.

Theorem 1: In a work-conserving, nonpreemptive queue with equal service times, the OCF discipline

a) minimizes the maximum age of departures

b) maximizes the minimum age of departures

c) minimizes the sample variance of ages of departures.

Proof: Appendix B.

Notice that this is a sample path property and does not depend on any assumptions made about the arrival process. Given *any* arrival process of equal length packets at a queue, Theorem 1 implies that OCF is the optimal work-conserving, nonpreemptive discipline in minimizing the variability of ages of departing packets in the sense of minimum, maximum, and sample variance. Unfortunately, the implications of the theorem do not seem to extend straightforwardly to the case of arbitrary service times or to general networks of queues due to the complex interactions between queues in a network.

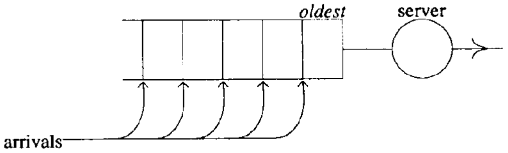


Fig. 7. OCF queue.

When service times are arbitrary, it would seem that the OCF discipline should still minimize the maximum age of departures; however, a counterexample demonstrates that this is not necessarily true. Suppose customers A and B are in the queue at time $t = 0$ with respective ages 1 and 0 and service times 2 and 1. Customer C arrives at time $t = 1$ with age 4 and service time 1. Under the OCF policy, they are serviced in the order A–C–B; the maximum departure age is 6 (customer C). However, the order B–C–A results in a maximum departure age of only 5 (customer C). The reason is that an arrival of relatively great age has a larger effect on the maximum departure age. With *a priori* knowledge of such an arrival, the order of service can be arranged so that the arrival is given service as promptly as possible. Evidently, the OCF discipline is not the optimal *noncausal* discipline (i.e., using knowledge of future events) for the case of arbitrary service times although it still might be the optimal *causal* discipline.

It is useful to consider a system of two M/M/1 queues in tandem, as shown in Fig. 8, with no intermediate arrivals or departures. If these queues are FCFS and stationary, it is well known that a packet's sojourn times in each queue would be independent [40]. If these queues are OCF, it seems at first that the sojourn times in each queue would be negatively correlated, but this would be incorrect. In this particular example, the OCF tandem queues behave as FCFS queues because a packet cannot overtake its predecessors. This behavior is generalized in the proposition below.

Proposition 2: Customers in the same ''stream'' (i.e., following the same fixed end-to-end path in the network) cannot overtake each other if the path consists only of FCFS and/or OCF single-server queues.

Proof: Customers in the same stream clearly cannot overtake each other in a FCFS queue. In an OCF queue, a customer can overtake a predecessor only if it is older, but all of its predecessors enter the system earlier. Hence, a customer is never older than its predecessors, and customers in the same stream cannot overtake each other in an OCF queue.

This nonovertaking property offers a way to conceptualize the operation of an OCF queue in a network with fixed routing. Within a stream, packets maintain their sequential order. The OCF queue compares the ages of the packets at the head of each stream and passes the stream with the oldest packet until another stream becomes older. Thus, the OCF discipline does not effect the sequential

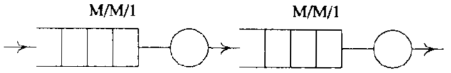


Fig. 8. M/M/1 queues in tandem.

order of packets in each stream, but it does effect the priority of different streams relative to each other.

B. EDF Discipline

Without more information, it would be reasonable to assume that a packet's urgency increases with its time in the network. However, urgency should also depend on many other factors. For example, a packet far from its destination is clearly more urgent than a packet of the same age that is close to its destination. Besides age and route length, urgency could also depend on expiration times and the traffic conditions at other queues. All available information should be used in determining a packet's priority.

Suppose that a node is given information such as each packet's age, route, expiration time, and traffic conditions at other nodes. With this information, it can determine the desired departure time, or "deadline," of each packet from that node (using terminology from scheduling theory [41]). As an example, suppose that a packet arrives at time t with age a, and the estimated delay along its remaining route in the network is d. For a maximum allowable entry-to-exit delay $D_{\max}$, its deadline might be determined as $t + (D_{\max} - a - d)$.

An *earliest-deadline-first* (EDF) discipline giving service to the packet waiting in queue with the earliest deadline is then a generalization of the OCF discipline. It assumes implicitly that a packet's urgency increases with the imminence of its deadline. Its similarity to the OCF discipline is apparent. The theorem below proves some optimal properties of the EDF discipline in terms of "lateness." Lateness is defined as the difference between a packet's actual departure time and its deadline [41].

Theorem 2: In a work-conserving, nonpreemptive queue with equal service times, the EDF discipline

a) minimizes the maximum lateness
b) maximizes the minimum lateness
c) minimizes the sample variance of lateness.

Proof: Appendix C.

An advantage of the EDF discipline is its applicability to integrated packet networks (a similar idea was proposed independently in [39]). For example, in a packet network carrying voice and data traffic, the voice packets could have stricter expiration times which would result in stricter deadlines (and hence higher priority under EDF). A practical disadvantage is the expensive processing and sorting necessary at each queue. We have not addressed the important issue of practicality in this study. Indeed, it is possible that implementation of the OCF or EDF discipline would be too costly or that buffering would be minimal in fast packet networks as suggested in [42]. This issue requires further investigation.

V. Simulations

A. Simulation Model

In order to study the effectiveness of dynamic priorities on network performance, the queueing model shown in Fig. 9 was simulated (based on the simulation model in [27]). In this model, arrivals of test packets are assumed to be Poisson with rate λ, although the packet process from an actual voice source is not Poisson. All test packets flow through five queues in tandem representing a virtual circuit. Each queue has infinite buffer capacity. In addition, interfering traffic is modeled by "transit" packets which enter the system at the ith queue as a Poisson process with rate λ_i independently of test packets. After the ith queue, each transit packet continues with probability q_i or departs the system with probability $1 - q_i$.

Test packets enter the system with zero ago, but transit packets are assumed to enter the system with a random age (from previous travel). For the lack of a better model, the ages of transit packets entering the system at the ith queue are assumed to be i.i.d. exponential with mean a_i. The service times of all packets are assumed to be deterministic and equal to μ^{-1}.

B. Simulation Results

The behavior of the simulation model under the OCF and FCFS disciplines are compared. The age of test packets (but not transit packets) is observed. Let d_i^{FCFS} and d_i^{OCF} denote the age of test packets departing from queue i in the FCFS and OCF models. The variance and 99th percentile of d_i are denoted as var (d_i) and $(d_i)_{0.99}$. We used the method of replications [43]–[45]. For each set of parameters, the simulation run was independently replicated eight times. If X_j is the measurement of variable X on the jth replication, then the observed simulation value of X is the average

$$\hat{X} = \tfrac{1}{8} \sum_{j=1}^{8} X_j. \tag{3}$$

Since the replications are identical and independent, the $\{X_j\}$ should be i.i.d. The 95 percent confidence interval of $\hat{X}$ is the usual $\hat{X} \pm (s/\sqrt{8})t_{7,0.975}$ where

$$s^2 = \tfrac{1}{7} \sum_{j=1}^{8} (X_j - \hat{X})^2 \tag{4}$$

is the usual sample variance and $t_{7,0.975}$ is the 97.5 percentile of the t-distribution with 7 degrees of freedom [43]–[45].

For simplicity, we set $a_1 = \cdots = a_5 = a$ and $q_1 = \cdots = q_5 = 0.3$ (as in [27]). In order to derive reasonable parameter values, consider that fast packet switches are designed to handle up to 50 000 calls simultaneously [9]. If the test traffic represents the packets associated with one call, the transit traffic should be much heavier than the test traffic. We assume the ratio is some large number, say $\lambda_1 = 10^3\lambda$. For convenience, we suppose further that the traffic intensity is the same at all queues; then we have the condition $\lambda_2 = \cdots = \lambda_5 = 0.7\lambda_1$. In this case, the

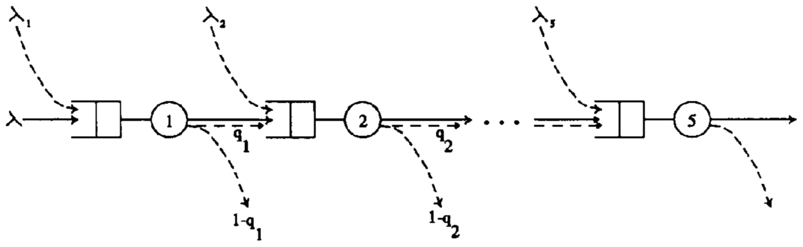


Fig. 9. Simulation model.

traffic intensity at every queue is

$$\rho = (10^3 + 1)\frac{\lambda}{\mu}. \quad (5)$$

Suppose that the speech encoding rate is 32 kbits/s and packets are 1 kbit. If the speech activity is 40 percent, then a single call has an average rate of 12.8 packets/s. Assuming a node can handle approximately 50 000 such packet streams simultaneously, the service rate must be about $\mu = 6 \cdot 10^5$ packets/s; we assume $\mu = 5 \cdot 10^5$ for each queue.

We must remark that such a high ratio of transit traffic to test traffic creates a practical difficulty. It takes a long time to run a simulation for even a few test packets (i.e., observations). For example, the observation of 10^3 test packets would involve the simulation of about 10^6 transit packets at each queue. Therefore, for a reasonably lengthy simulation run, we must settle for a relatively few number of test packets.

For each simulation run, the initial ten test packets were deleted in order to reduce the effect of bias due to initial conditions [43]-[45]. After the initial ten test packets, 100 test packets were observed in each run. Although this is a small number of observations, each simulation run was actually lengthy in simulated time and the system was likely to be observed in steady state.

Simulation values of var (d_i) and $(d_i)_{0.99}$ are listed in Tables I and II with their 95 percent confidence intervals. Note that under the OCF discipline, test packets arrive at the first queue with zero age and hence have lowest priority. As they travel further and accumulate age, their priorities increase relative to the transit traffic. Thus, it is expected that the performance of OCF should be worse compared to FCFS in the first queue and gradually improve as the test packets travel through more queues. This behavior is evident from Tables I and II. The OCF discipline has the effect of improving the delays of packets along longer paths at the expense of packets along shorter paths.

For $a = \mu^{-1}$, OCF becomes advantageous after the first or second queue. After the fifth queue, the delay variance is reduced by as much as 70 percent and the 99th pecentile is reduced by as much as 40 percent in heavy traffic. For $a = 5\mu^{-1}$, OCF becomes advantageous only after the third or fourth queue and only in heavy traffic. After the fifth queue, the delay variance is reduced by as much as 65 percent and the 99th percentile is reduced by as much as 30 percent in heavy traffic. It should be remarked, however, that these results are dependent on the model parameters and the assumptions made about the transit traffic (because most of the network traffic is transit). In general though, the results corroborate what we might have expected. It can be seen that the effectiveness of the OCF discipline 1) increases with the length of the path, 2) increases with increasing ρ, and 3) decreases with increasing a.

VI. Conclusions

Dynamic priorities have been studied by means of queueing theory and simulations. Optimal properties of the OCF and EDF disciplines have been proven which suggest that they may theoretically be effective in reducing the variability of packet delays. Simulation results of the OCF discipline indicate that the OCF discipline is most effective under conditions of long routes and heavy traffic. These are the conditions when delay variability is most likely to be significant. Under OCF, the delays of packets along long routes are improved at the expense of packets along short routes. More complex and realistic simulations, including simulations of the EDF discipline, are needed however.

This paper has not addressed the important issue of practicality. It is possible that buffering will be minimal in fast packet networks and hence delay variability will not be a significant issue. Even if delay variability is an issue, the increased processing and sorting necessary at each queue for the OCF or EDF disciplines could be too costly or could create additional delays impairing their effectiveness. It is also possible that other similar disciplines would be more simply implemented. The issue of practicality remains to be studied.

Appendix A

Proof of Proposition 1: Let $\overline{W}_{k\alpha}$ denote the expected waiting time in queue of a class k customer with age α upon arrival. The approach is to derive an implicit expression for $\overline{W}_{k\alpha}$ and show it is a decreasing function of α. The method for deriving $\overline{W}_{k\alpha}$ is similar to the analysis of time-dependent priorities by Kleinrock [33].

Under the OCF discipline, an arrival must wait for the customer found in service, customers already in the queue with greater age, and customers of greater age who arrive while he is still waiting in the queue and the queue ahead of him. Let x_k be the service time of a class k customer

TABLE I
SIMULATION RESULTS (ACTUAL VALUES HAVE BEEN MULTIPLIED BY 10^{14})

a	ρ	$var(d_1^{FCFS})$	$var(d_1^{QCF})$	$var(d_2^{FCFS})$	$var(d_2^{QCF})$	$var(d_3^{FCFS})$
μ^{-1}	.15	26 ±22%	48 ±36%	41 ±26%	58 ±35%	63 ±36%
	.30	66 ±15%	128 ±34%	153 ±16%	183 ±35%	248 ±21%
	.45	212 ±23%	288 ±21%	373 ±24%	403 ±12%	561 ±19%
	.60	412 ±15%	716 ±16%	807 ±14%	839 ±15%	1278 ±16%
	.75	1359 ±25%	1761 ±15%	3064 ±18%	2204 ±26%	4527 ±13%
	.90	8203 ±27%	9478 ±31%	17703 ±11%	11588 ±23%	27276 ±16%
$5\mu^{-1}$	.15	26 ±22%	47 ±42%	41 ±26%	67 ±38%	63 ±36%
	.30	66 ±15%	207 ±31%	153 ±16%	321 ±33%	248 ±21%
	.45	212 ±23%	407 ±30%	373 ±24%	660 ±21%	561 ±19%
	.60	412 ±15%	1583 ±16%	807 ±14%	2148 ±16%	1278 ±16%
	.75	1359 ±25%	3970 ±21%	3064 ±18%	4669 ±15%	4527 ±13%
	.90	8203 ±27%	14135 ±15%	17703 ±11%	16026 ±21%	27276 ±16%

a	ρ	$var(d_3^{QCF})$	$var(d_4^{FCFS})$	$var(d_4^{QCF})$	$var(d_5^{FCFS})$	$var(d_5^{QCF})$
μ^{-1}	.15	69 ±34%	89 ±33%	85 ±31%	115 ±32%	102 ±23%
	.30	214 ±30%	308 ±16%	239 ±27%	408 ±14%	267 ±25%
	.45	478 ±10%	792 ±23%	518 ±15%	997 ±24%	557 ±16%
	.60	909 ±12%	1799 ±16%	939 ±12%	2121 ±9%	955 ±13%
	.75	2392 ±21%	5817 ±13%	2405 ±18%	7432 ±10%	2482 ±16%
	.90	13114 ±21%	38799 ±13%	13637 ±21%	48173 ±14%	14058 ±19%
$5\mu^{-1}$	.15	105 ±35%	89 ±33%	129 ±30%	115 ±32%	142 ±25%
	.30	393 ±28%	308 ±16%	475 ±28%	408 ±14%	522 ±24%
	.45	870 ±11%	792 ±23%	971 ±16%	997 ±24%	1024 ±14%
	.60	2265 ±15%	1799 ±16%	2355 ±13%	2121 ±9%	2367 ±11%
	.75	4548 ±17%	5817 ±13%	4309 ±18%	7432 ±10%	4254 ±17%
	.90	16196 ±24%	38799 ±13%	16535 ±22%	48173 ±14%	16791 ±17%

and W_0 be the expected remaining service time of a customer found in service by an arrival. For an M/G/1 queue, it is known that the expected remaining service time is $E(x_k^2)/2E(x_k)$ if a class k customer is found in service by an arrival. The probability of finding a class k customer in service is $\lambda_k E(x_k)$, and therefore

$$W_0 = \sum_{k=1}^{P} \frac{\lambda_k E(x_k^2)}{2}. \tag{A.1}$$

Now, let $\overline{M}_{j\alpha}$ denote the expected number of class j customers that arrive and join the queue ahead of the considered customer while he is waiting in the queue. The considered customer, say customer C, arrives with age α and waits an average time of $\overline{W}_{k\alpha}$. His priority is shown as a function of time in Fig. 10. A later customer with age $\alpha' \le \alpha$ will not affect the customer C. A later customer with age $\alpha' \in (\alpha, \alpha + \overline{W}_{k\alpha}]$, will queue ahead of C if he arrives in a time interval of duration $\alpha' - \alpha$. A later customer with age $\alpha' > \alpha + \overline{W}_{k\alpha}$ can arrive in a time interval of duration $\overline{W}_{k\alpha}$ and queue ahead of C. Since class j customers with ages between α' and $\alpha' + d\alpha'$ have average arrival rate $\lambda_j p_j(\alpha')\, d\alpha'$,

TABLE II
SIMULATION RESULTS (ACTUAL VALUES HAVE BEEN MULTIPLIED BY 10^6)

a	ρ	$(d_1^{FCFS})_{.99}$	$(d_1^{OCF})_{.99}$	$(d_2^{FCFS})_{.99}$	$(d_2^{OCF})_{.99}$	$(d_3^{FCFS})_{.99}$	$(d_3^{OCF})_{.99}$	$(d_4^{FCFS})_{.99}$	$(d_4^{OCF})_{.99}$	$(d_5^{FCFS})_{.99}$	$(d_5^{OCF})_{.99}$
μ^{-1}	.15	5 ±7%	6 ±16%	7 ±8%	8 ±12%	10 ±12%	10 ±9%	13 ±12%	13 ±7%	15 ±13%	15 ±6%
	.30	6 ±8%	9 ±18%	10 ±9%	11 ±13%	14 ±10%	13 ±11%	17 ±7%	15 ±10%	20 ±9%	18 ±9%
	.45	9 ±17%	11 ±14%	13 ±13%	13 ±9%	17 ±9%	16 ±6%	21 ±11%	19 ±5%	25 ±9%	21 ±5%
	.60	13 ±12%	14 ±8%	18 ±14%	16 ±7%	23 ±15%	19 ±4%	29 ±12%	22 ±6%	33 ±8%	24 ±5%
	.75	21 ±18%	22 ±17%	34 ±19%	27 ±22%	40 ±12%	33 ±18%	45 ±8%	36 ±14%	55 ±8%	39 ±11%
	.90	45 ±24%	50 ±22%	72 ±13%	59 ±19%	91 ±13%	65 ±16%	111 ±12%	68 ±15%	123 ±9%	74 ±12%
$5\mu^{-1}$	.15	5 ±7%	6 ±25%	7 ±8%	9 ±18%	10 ±12%	12 ±13%	13 ±12%	14 ±9%	15 ±13%	16 ±7%
	.30	6 ±8%	10 ±22%	10 ±9%	14 ±16%	14 ±10%	16 ±12%	17 ±7%	19 ±11%	20 ±9%	21 ±8%
	.45	9 ±17%	12 ±22%	13 ±13%	15 ±13%	17 ±9%	20 ±6%	21 ±11%	23 ±6%	25 ±9%	25 ±5%
	.60	13 ±12%	20 ±11%	18 ±14%	24 ±7%	23 ±15%	26 ±7%	29 ±12%	29 ±7%	33 ±8%	32 ±6%
	.75	21 ±18%	30 ±12%	34 ±19%	34 ±8%	40 ±12%	37 ±7%	45 ±8%	39 ±7%	55 ±8%	42 ±6%
	.90	45 ±24%	58 ±16%	72 ±13%	64 ±17%	91 ±13%	70 ±14%	111 ±12%	76 ±13%	123 ±9%	82 ±9%

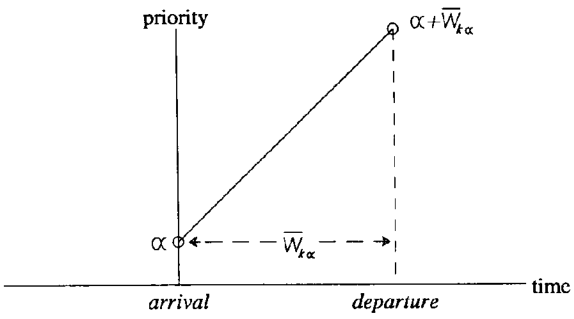


Fig. 10. Priority of customer C.

$$\overline{M}_{j\alpha} = \int_{\alpha}^{\alpha+\overline{W}_{k\alpha}} \lambda_j(\alpha'-\alpha)p_j(\alpha')\,d\alpha' + \sum_{\alpha+\overline{W}_{k\alpha}}^{\infty} \lambda_j \overline{W}_{k\alpha} p_j(\alpha')\,d\alpha'$$

$$= \int_{\alpha}^{\alpha+\overline{W}_{k\alpha}} \lambda_j(\alpha'-\alpha)p_j(\alpha')\,d\alpha' + \int_0^{\infty} \lambda_j \overline{W}_{k\alpha} p_j(\alpha')\,d\alpha' - \int_0^{\alpha+\overline{W}_{k\alpha}} \lambda_j \overline{W}_{k\alpha} p_j(\alpha')\,d\alpha'$$

$$= \int_{\alpha}^{\alpha+\overline{W}_{k\alpha}} \lambda_j(\alpha'-\alpha)p_j(\alpha')\,d\alpha' + \lambda_j \overline{W}_{k\alpha} - \lambda_j \overline{W}_{k\alpha} P_j(\alpha+\overline{W}_{k\alpha}) \quad \text{(A.2)}$$

where $p_j(\cdot)$ and $P_j(\cdot)$ are the age probability density and distribution functions for class j customers.

Let $\overline{N}_{j\alpha}$ denote the mean number of class j customers with age greater than α found in queue by customer C upon arrival. For convenience, suppose that C arrives at time 0. He finds a customer in queue if that customer arrived with age $\alpha' < \alpha$ at some time $-t$ (where $t > \alpha - \alpha'$) and waited longer than t, or arrived with age $\alpha' > \alpha$ at time $-t$ (where $t > 0$) and waited longer than t. The expected number of the first type of customers with ages between α' and $\alpha' + d\alpha'$ is

$$\int_0^{\infty} \Pr\{W_{j\alpha'} > t\}\,dt\,\lambda_j p_j(\alpha')\,d\alpha'.$$

The expected number of the second type is

$$\int_{\alpha-\alpha'}^{\infty} \Pr\{W_{j\alpha'} > t\}\,dt\,\lambda_j p_j(\alpha')\,d\alpha'.$$

$\overline{N}_{j\alpha}$ is the sum of both types of customers:

$$\overline{N}_{j\alpha} = \int_0^{\alpha}\int_{\alpha-\alpha'}^{\infty} \Pr\{W_{j\alpha'} > t\}\,dt\,\lambda_j p_j(\alpha')\,d\alpha' + \int_0^{\infty}\int_0^{\infty} \Pr\{W_{j\alpha'} > t\}\,dt\,\lambda_j p_j(\alpha')\,d\alpha'$$

$$= \int_0^{\infty}\int_0^{\infty} \Pr\{W_{j\alpha'} > t\}\,dt\,\lambda_j p_j(\alpha')\,d\alpha' - \int_0^{\alpha}\int_0^{\alpha-\alpha'} \Pr\{W_{j\alpha'} > t\}\,dt\,\lambda_j p_j(\alpha')\,d\alpha'$$

$$= \int_0^{\infty} \overline{W}_{j\alpha'}\lambda_j p_j(\alpha')\,d\alpha' - \int_0^{\alpha}\int_0^{\alpha-\alpha'} \cdot \Pr\{W_{j\alpha'} > t\}\,dt\,\lambda_j p_j(\alpha')\,d\alpha'$$

$$= \lambda_j \overline{W}_j - \int_0^{\alpha}\int_0^{\alpha-\alpha'} \Pr\{W_{j\alpha'} > t\}\,dt\,\lambda_j p_j(\alpha')\,d\alpha' \quad \text{(A.3)}$$

where $\overline{W}_j$ is the average unconditional waiting time in queue of a class j customer.

By Little's formula,

$$\overline{W}_{k\alpha} = W_0 + \sum_{j=1}^{P} E(x_j)(\overline{M}_{j\alpha} + \overline{N}_{j\alpha}). \quad (A.4)$$

This is unfortunately an implicit expression where $\overline{W}_{k\alpha}$ appears on both sides. However, differentiation with respect to α on both sides gives

$$\frac{d\overline{W}_{k\alpha}}{d\alpha} = \sum_{j=1}^{P} E(x_j)\left[\frac{d\overline{M}_{j\alpha}}{d\alpha} + \frac{d\overline{N}_{j\alpha}}{d\alpha}\right] \quad (A.5)$$

where

$$\frac{d\overline{M}_{j\alpha}}{d\alpha} = \lambda_j \frac{d\overline{W}_{k\alpha}}{d\alpha}(1 - P_j(\alpha + \overline{W}_{k\alpha})) - \lambda_j(P_j(\alpha + \overline{W}_{k\alpha}) - P_j(\alpha)) \quad (A.6)$$

and

$$\frac{d\overline{N}_{j\alpha}}{d\alpha} = -\int_0^{\alpha} \Pr\{W_{j\alpha'} > \alpha\} \lambda_j p_j(\alpha')\, d\alpha' \quad (A.7)$$

(using the formula

$$\frac{d}{dx}\int_{a(x)}^{b(x)} f(x, t)\, dt = \frac{db(x)}{dx} f(x, b) - \frac{da(x)}{dx} f(x, a) + \int_{a(x)}^{b(x)} \frac{\partial f(x, t)}{\partial x}\, dt).$$

After rearranging,

$$\frac{d\overline{W}_{k\alpha}}{d\alpha} = -\frac{\sum_{j=1}^{P} E(x_j)\lambda_j\left(P_j(\alpha + \overline{W}_{j\alpha}) - P_j(\alpha) + \int_0^{\alpha} \Pr\{W_{j\alpha'} > \alpha\}\lambda_j p_j(\alpha')\, d\alpha'\right)}{1 - \sum_{j=1}^{P} E(x_j)\lambda_j[1 - P_j(\alpha + \overline{W}_{k\alpha})]}. \quad (A.8)$$

Notice that the numerator is positive. In the denominator, $\Sigma_{j=1}^{P} \lambda_j E(x_j) < 1$ for stability so the denominator is also positive. This implies $d\overline{W}_{k\alpha}/d\alpha$ is negative and therefore, $\overline{W}_{k\alpha}$ is a decreasing function of α for any k. It follows that the waiting time of any class customer with arrival age α is negatively correlated with α.

Appendix B

Proof of Theorem 1: By contradiction. Statement a) is proven first; b) and c) are proven in the same way with minor modifications.

Assume there exists a dicipline denoted S that is different from OCF and results in a smaller maximum departure age than OCF. By definition, there exists a time T_0 when there are two customers (say, customers A and B) in the queue with respective ages α_A and α_B at time T_0 and $\alpha_A > \alpha_B$ (i.e., customer A is older than B), but customer B is given service instead of A. If service times are T, customer B departs with age

$$\delta_B = \alpha_B + T. \quad (B.1)$$

Since the system is nonpreemptive and work-conserving, customer A is given service at some later time $T_0 + nT$ where n is a positive integer. Customer A departs with age

$$\delta_A = \alpha_A + (n + 1)T. \quad (B.2)$$

Now consider a discipline denoted S' that is exactly like S except that customer A is given service at T_0 and B is given service at $T_0 + nT$ (i.e., A and B are switched exactly). Since service times are equal, all other customers are unaffected by this switch. Under S',

$$\delta'_A = \alpha_A + T$$
$$\delta'_B = \alpha_B + (n + 1)T. \quad (B.3)$$

Note that

$$\delta'_A < \delta_A$$
$$\delta'_B < \delta_A \quad (B.4)$$

and therefore,

$$\max(\delta'_A, \delta'_B) < \max(\delta_A, \delta_B). \quad (B.5)$$

Since all other customers are unaffected, the maximum age of departures under S' must be less than or equal to the maximum age of departures under S. This process can be continued until all violations of the OCF policy are removed. Hence, the maximum departure age under OCF is less than or equal to that under S, leading to a contradiction of the assumption.

In order to prove b) assume there exists a discipline S different than OCF that results in a larger minimum departure age than OCF. Instead of (B.4) and (B.5), note from (B.1)-(B.3) that

$$\delta'_A > \delta_B$$
$$\delta'_B > \delta_B \quad (B.6)$$

and therefore,

$$\min(\delta'_A, \delta'_B) > \min(\delta_A, \delta_B). \quad (B.7)$$

The remainder of the proof of b) follows in the same way as before.

In the proof of statement c), the sample variance under S for a population of any size $k > 1$ is

$$s^2(\delta) = \frac{1}{k-1}\sum_{i=1}^{k}\left(\delta_i - \frac{1}{k}\sum_{j=1}^{k}\delta_j\right)^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\delta_i)^2 - \frac{1}{k(k-1)}\left(\sum_{i=1}^{k}\delta_i\right)^2 \quad (B.8)$$

where δ_i is the age of customer i upon departure. From (B.1)–(B.3), note that

$$\delta_A + \delta_B = \delta_A' + \delta_B' \tag{B.9}$$

and

$$\sum_{i=1}^{k} \delta_i = \sum_{i=1}^{k} \delta_i' \tag{B.10}$$

because all other customers are unaffected. Also, by writing

$$\begin{aligned}(\delta_A)^2 + (\delta_B)^2 &= (\alpha_A + T)^2 + (\alpha_B + T)^2 \\ &\quad + 2nT(\alpha_A + T) + n^2T^2 \\ (\delta_A')^2 + (\delta_B')^2 &= (\alpha_A + T)^2 + (\alpha_B + T)^2 \\ &\quad + 2nT(\alpha_B + T) + n^2T^2\end{aligned} \tag{B.11}$$

it becomes clear that

$$(\delta_A)^2 + (\delta_B)^2 > (\delta_A')^2 + (\delta_B')^2. \tag{B.12}$$

Consequently, the first term in (B.8) is larger under S so

$$s^2(\delta) > s^2(\delta'). \tag{B.13}$$

The remainder of the proof of c) follows in the same way as before.

Appendix C

Proof of Theorem 2: Analogous to Theorem 1. Assume there exists a discipline denoted S that is different than EDF that results in a smaller maximum lateness than EDF. By definition, there exists a time T_0 when there are two customers (say, A and B) in the queue with respective deadlines d_A and d_B and $d_A < d_B$ (i.e., customer A's deadline is earlier), and B is given service instead of A. If service times are T, customer B has lateness

$$l_B = T_0 + T - d_B. \tag{C.1}$$

Since the system is nonpreemptive and work-conserving, customer A is given service at some later time $T_0 + nT$ where n is a positive integer. Customer A has lateness

$$l_A = T_0 + (n + 1)T - d_A. \tag{C.2}$$

Now consider a discipline denoted S' that is exactly like S except that customer A is given service at T_0 and B is given service at $T_0 + nT$ (i.e., A and B are switched exactly). Since service times are equal, all other customers are unaffected by this switch. Under S',

$$\begin{aligned}l_A' &= T_0 + T - d_A \\ l_B' &= T_0 + (n + 1)T - d_B.\end{aligned} \tag{C.3}$$

Note that

$$\begin{aligned}l_A' &< l_A \\ l_B' &< l_A\end{aligned} \tag{C.4}$$

and therefore,

$$\max\,(l_A', l_B') < \max\,(l_A, l_B). \tag{C.5}$$

Since all other customers are unaffected, the maximum lateness under S' must be less than or equal to the maximum lateness under S. This process can be continued until all violations of the EDF policy are removed. Hence, the maximum lateness under EDF is less than or equal to that under S, leading to a contradiction of the assumption.

The proofs of b) and c) are analogous to the proof of Theorem 1. The proof for b) depends on noticing

$$\begin{aligned}l_A' &> l_B \\ l_B' &> l_B.\end{aligned} \tag{C.6}$$

For c), notice

$$l_A + l_B = l_A' + l_B' \tag{C.7}$$

and

$$(l_A)^2 + (l_B)^2 > (l_A')^2 + (l_B')^2. \tag{C.8}$$

Hence,

$$s^2(l) > s^2(l'). \tag{C.9}$$

Acknowledgment

The authors are grateful to the anonymous reviewers for their suggestions and remarks.

References

[1] T. Chen and D. Messerschmitt, "Integrated voice/data switching," *IEEE Communications*, vol. 26, pp. 16–26, June 1988.

[2] I. Dorros, "Telephone nets go digital," *IEEE Spectrum*, pp. 48–53, Apr. 1983.

[3] ——, "ISDN," *IEEE Communications*, pp. 16–19, Mar. 1981.

[4] M. Decina, "Progress towards user access arrangements in integrated services digital networks," *IEEE Trans. Communications*, vol. COM-30, pp. 2117–2130, Sept. 1982.

[5] M. Decina and A. Roveri, "ISDN: integrated services digital network—Architectures and protocols," in *Advanced Digital Communications*, K. Feher, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[6] A. Tanenbaum, *Computer Networks.* Englewood Cliffs, NJ: Prentice-Hall, 1981.

[7] L. Roberts, "The evolution of packet switching," *Proc. IEEE*, vol. 66, pp. 1307–1313, Nov. 1978.

[8] P. Green, Jr., "An introduction to network architectures and protocols," *IEEE Trans. Commun.*, vol. COM-28, pp. 413–424, Apr. 1980.

[9] J. Turner, "Design of an integrated services packet network," *IEEE J. Select. Areas Commun.*, vol. SAC-4, pp. 1373–1380, Nov. 1986.

[10] G. Luderer *et al.*, "Wideband packet technology for switching systems," in *Proc. ISS '87*, pp. B6.1.1–B6.1.7.

[11] R. Muise *et al.*, "Experiments in wideband packet technology," in *1986 Zurich Seminar Dig. Commun.*, pp. D4.1–D4.5.

[12] P. Kirton *et al.*, "Fast packet switching for integrated network evolution," in *Proc. ISS '87*, pp. B6.2.1–B6.2.7.

[13] W. Standish and S. Sistla, "Network switching in the 1990's," in *Proc. ISS '87*, pp. C5.1.1–5.1.9.

[14] J. Turner, "New directions in communications (or which way to the information age?)," *IEEE Communications*, vol. 24, pp. 8–15, Oct. 1986.

[15] Special Issue on Switching Systems for Broadband Networks, *IEEE J. Select. Areas Commun.*, vol. SAC-5, Oct. 1987.

[16] P. Brady, "A model for generating on-off patterns in two-way communications," *Bell Syst. Tech. J.*, vol. 48, pp. 2445–2472, Sept. 1969.

[17] Bell Telephone Labs, *Engineering and Operations in the Bell System*, AT&T Bell Labs, 1984.

[18] M. Listanti and F. Villani, "An X.25-compatible protocol for packet voice communications," *Comput. Commun.*, vol. 6, pp. 23–31, Feb. 1983.

[19] J. Forgie and A. Nemeth, "An efficient packetized voice/data net-

work using statistical flow control,'' in *Proc. ICC '77*, pp. 38.2.44-38.2.48.

[20] B. Gold, ''Digital speech networks,'' *Proc. IEEE*, vol. 65, pp. 1636-1658, Dec. 1977.

[21] C. Weinstein and J. Forgie, ''Experience with speech communications in packet networks,'' *IEEE J. Select. Areas Commun.*, vol. SAC-1, pp. 963-980, Dec. 1983.

[22] D. Minoli, ''Optimal packet length for packet voice communication,'' *IEEE Trans. Commun.*, vol. COM-27, pp. 607-611, Mar. 1979.

[23] T. Bially *et al.*, ''A technique for adaptive voice flow control in integrated packet networks,'' *IEEE Trans. Commun.*, vol. COM-28, pp. 325-333, Mar. 1980.

[24] ——, ''Voice communication in integrated digital voice and data networks,'' *IEEE Trans. Commun.*, vol. COM-28, pp. 1478-1489, Sept. 1980.

[25] W. Montgomery, ''Techniques for packet voice synchronization,'' *IEEE J. Select. Areas Commun.*, vol. SAC-1, pp. 1022-1028, Dec. 1983.

[26] G. Barberis and D. Pazzaglia, ''Analysis and optimal design of a packet voice receiver,'' *IEEE Trans. Commun.*, vol. COM-28, pp. 217-227, Feb. 1980.

[27] T. Suda *et al.*, ''Performance evaluation of a packetized voice system—Simulation study,'' *IEEE Trans. Commun.*, vol. COM-32, pp. 97-102, Jan. 1984.

[28] J. Jackson, ''Some problems in queueing with dynamic priorities,'' *Nav. Res. Quart.*, vol. 7, pp. 235-247, Sept. 1960.

[29] ——, ''Waiting-time distributions for queues with dynamic priorities,'' *Nav. Res. Quart.*, vol. 19, pp. 31-36, Mar. 1962.

[30] ——, ''Queues with dynamic priority discipline,'' *Management Sci.*, vol. 8, pp. 18-34, Oct. 1961.

[31] H. Goldberg, ''Analysis of the earliest due date scheduling rule in queueing systems,'' *Math. Oper. Res.*, vol. 2, pp. 145-154, May 1977.

[32] ——, ''Jackson's conjecture on earliest due date scheduling,'' *Math. Oper. Res.*, vol. 5, pp. 460-466, Aug. 1980.

[33] L. Kleinrock, *Queueing Systems, Vol. 2: Computer Applications.* New York: Wiley, 1976.

[34] J. Hsu, ''A continuation of delay-dependent queue disciplines,'' *Oper. Res.*, vol. 18, pp. 733-738, 1970.

[35] J. Kanet, ''A mixed delay-dependent queue discipline,'' *Oper. Res.*, vol. 30, pp. 93-96, Jan.-Feb. 1982.

[36] U. Bagchi, ''A note on linearly decreasing, delay-dependent non-preemptive queue disciplines,'' *Oper. Res.*, vol. 32, pp. 952-957, July-Aug. 1984.

[37] U. Bagchi and R. Sullivan, ''Dynamic, non-preemptive priority queues with general, linearly increasing priority function,'' *Oper. Res.*, vol. 33, pp. 1278-1298, Nov.-Dec. 1985.

[38] A. Netterman and I. Adiri, ''A dynamic priority queue with general concave priority functions,'' *Oper. Res.*, vol. 27, pp. 1088-1100, Nov.-Dec. 1979.

[39] Y. Lim and J. Kobza, ''Analysis of a delay-dependent priority discipline in a multi-class traffic packet switching node,'' in *Proc. INFOCOM '88*, pp. 9A.4.1-9A.4.1.10.

[40] E. Reich, ''Waiting times when queues are in tandem,'' *Ann. Math. Statist.*, vol. 28, pp. 768-773, 1957.

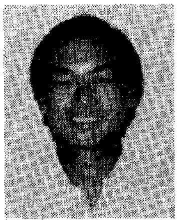
[41] R. Conway, W. Maxwell, and L. Miller, *Theory of Scheduling.* Reading, MA: Addison-Wesley, 1967.

[42] J. Hui and E. Arthurs, ''A broadband packet switch for integrated transport,'' *IEEE J. Select. Areas Commun.*, vol. SAC-5, pp. 1264-1273, Oct. 1987.

[43] G. Fishman, *Concepts and Methods in Discrete Event Digital Simulation.* New York: Wiley, 1973.

[44] J. Kleijnen, *Statistical Tools for Simulation Practitioners.* New York: Marcel Dekker, 1987.

[45] A. Law, ''Statistical analysis of simulation output data,'' *Oper. Res.*, vol. 31, pp. 983-1029, Nov. 1983.



Thomas M. Chen (S'89) received the B.S. and M.S. degrees in electrical engineering in 1984 from the Massachusetts Institute of Technology, Cambridge.

He is currently pursuing the Ph.D. degree in electrical engineering at the University of California, Berkeley. He has worked on image compression at the IBM San Jose Research Laboratory and ISDN at Pacific Bell.

Mr. Chen is a member of Eta Kappa Nu and Tau Beta Pi.

Jean Walrand (S'71-M'74-M'80) was born in Belgium. He received the Ingénieur Civil degree in electrical engineering from the Université de Liège, Belgium, in 1974 and the Ph.D. degree in electrical engineering from the University of California, Berkeley, in 1979.

He served as an Assistant Professor with the School of Electrical Engineering, Cornell University, Ithaca, NY, from 1979 until 1981 when he joined the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, where is now an Associate Professor. He is the author of *An Introduction to Queueing Networks* (Prentice-Hall, 1988) and of *Communication Networks: A First Course* (New York: McGraw-Hill). His research interests are in communication networks, queueing systems, stochastic control, and stochastic processes.

Dr. Walrand served as an Associate Editor of IEEE TRANSACTIONS ON AUTOMATIC CONTROL and is presently Associate Editor of *Probability in the Engineering and Informational Sciences*, of *Queueing Systems: Theory and Applications*, and of *Systems and Control Letters.*

David G. Messerschmitt (S'65-M'68-SM'78-F'83) received the B.S. degree from the University of Colorado, Boulder, in 1967, and the M.S. and Ph.D. degrees from the University of Michigan, Ann Arbor, in 1968 and 1971, respectively.

He is a Professor of Electrical Engineering and Computer Sciences, University of California, Berkeley. From 1968 to 1977 he was a member of the Technical Staff and later Supervisor at Bell Laboratories, Holmdel, NJ, where he did systems engineering, development, and research on digital transmission and digital signal processing (particularly relating to speech processing). His current research interests include applications of digital signal processing, digital communications (on the subscriber loop and fiber optics), architectural approaches to dedicated-hardware digital signal processing (with a current emphasis on video compression applications), and computer-aided design of communications and signal processing systems. He has published over 100 papers and two books and has 10 patents. Since 1977 he has also served as a consultant to a number of companies.

Dr. Messerschmitt is a member of Eta Kappa Nu, Tau Beta Pi, Sigma Xi, and has several best paper awards. He has served as a Senior Editor of the *Communications Magazine*, as Editor for Transmission of the TRANSACTIONS ON COMMUNICATIONS, and as a member of the Board of Governors of the Communication Society. He has also organized and participated in a number of short courses and seminars devoted to continuing engineering education.