

Discrete Time Queues with Priorities

Vinod Sharma*and N. D. Gangadhar
Laboratory of Telecommunication Technology,
Helsinki University of Technology,
P.O. Box 3000, FIN-02015 HUT, Finland.
Department of Electrical Engineering
Indian Institute of Science, Bangalore 560 012, India.
e-mail: vinod@keskus.hut.fi

Abstract

The different Quality of Service (QoS) requirements of various services that are to be supported by high speed networks call for priority based service and buffer management policies. Also, discrete time nature of some of these systems (*e.g.*, ATM networks) make discrete time models ideally suited for studying them. Motivated by these, we study a discrete queue with two priorities in this paper. The buffer could be finite or infinite and the arrivals could be Markov modulated. For the infinite buffer system only time priority is relevant. However for the finite buffer case we consider the following two cases. In case one both the traffic classes share the buffer on a first come basis. In case two each class has a reserved buffer space for itself. For all these cases, we compute the stationary average probability of loss, the stationary probability of loss of a typical packet, the distribution and moments of the stationary average delay and the stationary delay of a typical packet, for each class of traffic. First we provide the analysis for the i.i.d. traffic. Then we extend all the results to the case of Markov modulated arrivals.

1 Introduction

High speed networks are being designed to support services with different quality of service requirements (QoS). For instance, voice and video can tolerate a small packet loss but require the delays and delay jitter to be within certain bounds. On the other hand, data traffic (*e.g.*, file transfer) is more tolerant in its delay and delay jitter but accepts no loss.

To support different QoS requirements for different connections in a network, buffer and bandwidth management policies are to be in place at the switches. Various such policies have been proposed in literature (see Roberts et al.[11] for a survey). Two important mechanisms to meet the different QoS requirements are providing time and space priorities to different streams of traffic. For example, real time traffic which tolerates but small delay can be given time priority over the traffic that has good tolerance to delay. Similarly, traffic that has stringent loss requirements can be given space priority. This way it is possible to provide QoS guarantees to different services while at the same time achieving efficient utilization of bandwidth and buffer space.

Now we mention some related literature. A survey of results on discrete queueing systems is in Bruneel and Kim [3]. For a survey of different priority policies for traffic

*On leave from Dept. ECE, Indian Institute of Science, Bangalore, India.

management in high speed networks see Robert et al. [11]. A discrete time queue with two i.i.d. traffic streams and service time of one slot with multiple servers is studied in Lin and Silvester [9]. They consider the stationary probability of loss of packets from each class under complete sharing, complete sharing with pushout, partial buffer sharing space priorities and complete sharing with head of the line priority. The loss probabilities of various pushout schemes are compared in Kawahara *et al.* [6] for i.i.d. and Markov modulated Bernoulli arrivals. Takahashi and Hashida[15] analyze a discrete time queue with infinite buffer and i.i.d. arrivals from multiple priority classes (that may be correlated within a slot) under preemptive resume and head of the line priorities. They compute the generating function of the stationary mean waiting time distribution of packets from each class. Among the studies that have Markov modulated inputs are Garcia and Casal [5], Le Boudec [7] and Kawahara *et al.* [6] and Liao [8]. Garcia and Casals [5] performs an exact analysis for computing the probability of loss with Markov modulated Poisson processes, while Le Boudec [7] provides a numerical method for evaluating the loss probabilities in a discrete time queue with Markov modulated Bernoulli processes. Liao [8] considers a discrete queue with Markov modulated Poisson arrivals. He obtains the probability of loss of different classes. Other related work in literature includes the studies by Rubin and Tsai [12], the references there in, and Wagner [16]. Rubin and Tsai [12] obtain the generating function of the waiting time distributions for the infinite buffer queue and i.i.d. arrivals. Wagner considers a continuous time queue with Markov modulated arrivals and an infinite buffer. He obtains the generating function of the delays of the lower priority class. The delay moments and distributions of the two classes of traffic for a finite queue are not available in the above references. Recently, these results are obtained by us in Sharma and Gangadhar [14] for a discrete queue when space and time priority is considered.

In this paper we study a discrete time queue with two classes of traffic. We study infinite buffer as well as finite buffer queues. Also we will allow the arrival streams to be Markov modulated, and our methods can handle correlation among the streams. We use direct probabilistic arguments instead of transforms. Our system can model the output buffer of an ATM switch (with output queueing only) where two traffic streams are being multiplexed. We develop techniques to compute the stationary delays and probability of loss of packets from the two streams. Our methods work for more than two classes of traffic also except that the computations will, naturally, become more complex. We believe that the following results are new in this paper: (i) For the infinite buffer case the analysis of the priority system with Markov modulated arrivals. (ii) For the finite buffer case, in all the priority schemes we study, obtaining the distribution and moments of the stationary delays of the two classes. These are obviously of practical interest and usually are much more complicated to study than the probability of loss.

We clarify some of the notation used in this paper. We are considering a queue with batch arrivals. We will use various forms of stationary distributions. A distribution stationary at the slot boundaries is the *time stationary* distribution in the terminology of Brandt *et al.* [2]. For waiting times one can consider several different stationary distributions: the stationary distribution seen by the first packet of a batch of a class of packets, the average over the stationary distributions seen by different packets of a batch of a class (we will call this *mean stationary delay*) and the stationary waiting time seen by the packet stationary process (called the customer stationary in Brandt *et al.* [2]). The packet referenced at 0 in the customer stationary process for a batch arrival process in Brandt *et al.* [2] has position θ in its batch with the distribution given in (4) below. We call this packet as *typical* packet in this paper. The delay seen by this packet under stationarity is particularly important

and we will study it in this paper. Similarly, for a finite buffer queue, we can consider the following stationary probabilities of loss for a class: the stationary probability that the first (or the n th) packet of a batch of a class is lost, the average over the stationary probabilities of loss of different packets of a class (we will call this the *mean stationary* probability of loss and the stationary probability of loss of a *typical* packet. Since in this paper we are only concerned about the stationary probabilities, very often we will omit the word stationary.

Throughout the paper we will assume that within each class, the packets are served in FCFS order. We will denote by π and $\hat{\pi}$ in the different sections the stationary distributions of the Markov chains in those sections, *i.e.*, π and $\hat{\pi}$ in different sections will be different.

The paper is organized as follows. In Section 2 we consider the infinite buffer queue. We then study the finite buffer queue under complete sharing and partitioned buffer schemes in sections 3 and 4. We obtain the distributions of the stationary delays for both the streams for finite as well as infinite buffer queues for i.i.d. traffic. For a finite buffer queue we also compute the probability of loss of packets of each class. In section 5 we extend all these results to the Markov modulated arrival streams. Although the results in section 2 are available in Takahashi and Hashida [15], we provide them here for the following reasons. Our methods are different. Throughout the paper, we employ similar methods and hence they will be easiest to understand in this simplest setting. Also, in Section 5, we extend these results to the Markov modulated case which we believe are not available in the literature for the infinite buffer case also.

2 Infinite Buffer Queue

In this section we first describe the model, the notation and the assumptions which will be valid throughout the paper.

Consider a discrete time queue with time slotted into equal length slots, whose length is taken as a unit. Let $[k, k + 1)$ be the k th slot. In the k th slot, $X_k(i)$ packets from Class i , $i = 0, 1$, arrive at the queue. All the packets arriving during a slot will be in certain specific order and for a FCFS discipline (for each class), that ordering within a particular class will be honored. Each packet requires one slot of service. A packet that arrives in slot k is ready for service at the earliest at the beginning of slot $k + 1$. Till Section 4, we assume that $\{X_k(i), k \geq 0\}$ is i.i.d. and the traffic from different classes are independent (dependence of $X_k(0)$ and $X_k(1)$ can be allowed with minor modifications). In Section 5, we will extend all these results to Markov modulated $\{X_k(i), k \geq 0\}$.

The buffer length of the system is $N \leq \infty$. We consider the infinite buffer case in this section and the beginning of section 5. In the rest of the paper we analyze the finite buffer queue.

For the infinite buffer queue the queue length q_k at time k , evolves as

$$q_{k+1} = (q_k - 1)^+ + X_k(0) + X_k(1). \quad (1)$$

Class 1 traffic is given priority over Class 0 *i.e.*, a Class 0 packet is served only when there is no Class 1 packet in the queue. Also observe that from (1), if q_k is zero then the k th slot will go idle. This fact will be used in the analysis below. Within each class, the packets are served in a FCFS discipline. Then $\{q_k\}$ is a Markov chain. If $E[X_1(0)] + E[X_1(1)] < 1$, then it is aperiodic, irreducible, and ergodic. Hence it has a unique stationary distribution, say π . We make this assumption for the infinite buffer queue.

For computing the delay distribution of Class 1 packets, we consider the discrete queue with only Class 1 traffic. Then the stationary moments of the queue lengths $q_k(1)$ of Class 1 are (see Gangadhar [4])

$$\begin{aligned} \mathbb{E}[q_k(1)] &= \frac{\text{Var}[X_k(1)]}{2(1 - \mathbb{E}[X_k(1)])} + \frac{1}{2}\mathbb{E}[X_k(1)], & (2) \\ \mathbb{E}[(q_k(1))^2] &= \frac{1}{3(1 - \mathbb{E}[X_k(1)])} \cdot \left[\mathbb{E}[X_k(1)^3] + 3\mathbb{E}[q_k(1)](1 + \mathbb{E}[X_k(1)^2]) \right. \\ &\quad \left. - \mathbb{E}[X_k(1)](1 + 6\mathbb{E}[q_k(1)]) + 3\mathbb{E}[X_k(1)](\mathbb{E}[X_k(1)] - \mathbb{E}[X_k(1)^2]) \right] \end{aligned} \quad (3)$$

Now consider a typical packet of Class 1 to such a queue. Let it belong to a Class 1 batch arriving at time k (*i.e.*, in slot k) and let it be the n th packet in its batch. By *discrete* time PASTA, the distribution of the queue length this batch sees at time k is also π . Then the probability that its delay is $m \geq n$ slots is $\text{P}_\pi(q_k(1) = m - n + 1)$. The event that a *typical* packet of Class 1 is in the θ_1 -th position in its batch has distribution

$$\text{P}(\theta_1 = n) = \frac{\text{P}(X_k(1) = n)}{\text{P}(X_k(1) \geq n)}. \quad (4)$$

Thus we can calculate the distribution of the delay of a typical Class 1 packet (which equals $(q_k(1) - 1)^+ + \theta_1 - 1$ where θ_1 is independent of $q_k(1)$).

In order to obtain the distribution of delays of Class 0 packets we need to expand the state of the system to the vector $\{(q_k(0), q_k(1)), k \geq 0\}$, where $q_k(i)$ is the number of packets of Class i in the queue at time k . This is an irreducible, aperiodic, ergodic Markov chain under the above mentioned condition and let $\hat{\pi}$ be its unique stationary distribution.

Then the mean (average over all Class 0 packets) delay of Class 0 packets can be computed from the relation

$$\mathbb{E}[W](\mathbb{E}[X_1(1)] + \mathbb{E}[X_1(0)]) = \mathbb{E}[W(1)]\mathbb{E}[X_1(1)] + \mathbb{E}[W(0)]\mathbb{E}[X_1(0)], \quad (5)$$

where

$$\mathbb{E}[W(1)] = \sum_{m=1}^{\infty} \text{P}(X_k(1) = m | X_k(1) > 0) \frac{1}{m} \sum_{k=1}^m ((k-1) + \mathbb{E}[q_k(1)]) \quad (6)$$

and $\mathbb{E}[W]$ is obtained in the same way with $X_k(1)$ in (6) replaced by $X_k(0) + X_k(1)$ (since by Little's law, it does not depend on the work conserving service discipline).

One is also (in fact more) interested in the moments and distribution of the delay that a typical Class 0 packet experiences under stationarity. By *discrete* time PASTA, the distribution of the system that an arriving Class 0 batch observes is $\hat{\pi}$. However, the lower priority of Class 0 means that packets of Class 1 that arrive after this instant could also be served ahead of these packets. Consider a Class 0 packet at the n th position of a Class 0 batch, which arrives at time (say) k , (the probability that a typical packet is at position n in its batch can be obtained as in (4)). Let (n_0, n_1) be the state of the queue at time k . Then at time $k+1$, there will be $(n_1 - 1)^+ + X_k(1)$ Class 1 packets in the queue. Now consider a discrete time queue with only Class 1 traffic and let τ be the inter-visit time to the empty state of this queue. Let $\{\tau_k, k \geq 1\}$ be a sequence of i.i.d. random variables with the same distribution as τ . The distribution of τ and its moments are obtained in Gangadhar [4]. For example, the first two moments are given by

$$\mathbb{E}[\tau] = 1/(1 - \mathbb{E}[X_k(1)]), \quad (7)$$

$$\mathbb{E}[\tau^2] = \frac{\text{Var}[X_k(1)]}{(1 - \mathbb{E}[X_k(1)])^3} + \frac{1}{(1 - \mathbb{E}[X_k(1)])^2}. \quad (8)$$

Now, in our original queue, at time $k + \tau_1$, the number of Class 1 packets will be $n_1 - 1$, if $n_1 \geq 1$. Inductively we observe that the first time that the queue will not have any Class 1 packets will have the distribution of

$$k + \tau_1 + \tau_2 + \cdots + \tau_{(n_1-1)^+ + X_k(1)},$$

with $X_k(1)$ independent of $\{\tau_n\}$. This is also the first time when a packet of Class 0 will be transmitted after time k . Following the same argument, the n th packet of the batch considered above will be transmitted at time with the distribution of $k + \tau_1 + \tau_2 + \cdots + \tau_{(n_1+n_0-1)^+ + X_k(1) + n}$.

Therefore, from (4) and (7), the mean delay of the *typical* packet is

$$\begin{aligned} \mathbb{E}[D(0)] &= \sum_n \mathbb{P}(\theta_1 = n) \sum_{(n_0, n_1)} \hat{\pi}(n_0, n_1) \left((n_0 + n_1 - 1)^+ + n - 1 + \mathbb{E}[X_k(1)] \right) \mathbb{E}[\tau] \\ &= \frac{\mathbb{E}_{\hat{\pi}}[(q_k(0) + q_k(1) - 1)^+] + \mathbb{E}[\theta_1] + \mathbb{E}[X_1(1)] - 1}{1 - \mathbb{E}[X_k(1)]}. \end{aligned} \quad (9)$$

The second moment of its delay is

$$\begin{aligned} \mathbb{E}[(D(0))^2] &= \left(\mathbb{E}[D(0)] \right)^2 + \text{Var}[\tau] \mathbb{E}[(q_1(0) + q_1(1) - 1)^+ + \theta_1 - 1 + \mathbb{E}[X_1(1)]] \\ &= \mathbb{E}[D(0)]^2 + \frac{\text{Var}[X_k(1)]}{(1 - \mathbb{E}[X_k(1)])^2} \cdot \mathbb{E}[D(0)]. \end{aligned} \quad (10)$$

3 Finite Buffer Queue with Complete Sharing

In this section we consider the case when the buffer size N is finite and it is shared by the two classes. Now the queue length, q_k , satisfies

$$q_{k+1} = \min\{(q_k - 1)^+ + X_k(1) + X_k(0), N\}. \quad (11)$$

Then, if $\mathbb{E}[X_k(i)] < \infty$ and $\mathbb{P}(X_1(i) = 0) > 0$, $i = 1, 0$, the Markov chain has a unique aperiodic, irreducible set and forms an ergodic regenerative process. Thus it has a unique stationary distribution π . It may be possible to find another set of sufficient conditions for ergodicity. But the above conditions are weak enough and since this is not the focus of the paper, we do not elaborate. Similar comments will hold for ergodicity conditions in the later parts of the paper.

There are well known efficient methods available in the literature to find the stationary distributions of finite Markov chains, exploiting their special features. For example all the stationary distributions $\pi, \hat{\pi}$ that we have in this paper (finite as well as infinite) have $M/GI/1$ type transition matrices (block upper Hessenberg). For these matrices there are various efficient algorithms available to compute the stationary distributions (for recent references see [1] and Ramaswami [10]). Therefore we will assume that using these algorithms we can compute $\pi, \hat{\pi}$ required at various places in the paper. Using these distributions we provide methods to compute the distributions and moments of the waiting times and the packet loss probabilities for the two classes.

First we consider the probability of packet loss. For this we need to consider the distribution of arrivals in a slot. Let $p(n, m)$ be the probability that m Class 1 packets arrive ahead of the n th packet of a Class 0 batch in its slot of arrival. Since no class has space priority, the packets join the queue according to their order of arrival within a slot. On

arrival, according to the *discrete* time PASTA, a batch of Class i sees the system with distribution π , under stationarity. Consider the n th packet in a batch of Class 0 arriving at time k . Then the stationary probability that this packet is lost is

$$p_i(0) = \sum_{n_1=0}^N \sum_{m: m \geq N - (n_1 - 1) - n + 1} \mathbf{P}_\pi(q_k(1) + q_k(0) = n_1) \cdot p(n, m). \quad (12)$$

From this one can calculate the probability of loss of a typical Class 0 packet. Similarly, one can compute the probability of loss of the n th packet in a Class 1 batch and then of the *typical* packet of Class 1. Thus, in this case of time priorities, if the probability that m Class 0 packets arrive before the n th Class 1 packet in a slot is also $p(n, m)$, then the probability of packet loss of both the classes are the same. However, we know (and it can be seen below) that Class 1 has smaller delays.

Now we obtain the stationary distribution of delays of packets of the two classes of traffic. As before we expand the state representation of the system to $\hat{q}_k = (q_k(0), q_k(1))$, where $q_k(i)$ denotes the number of Class i packets in the queue at time k . The process $\{\hat{q}_k\}$ is a Markov chain and under the conditions stated at the beginning of this section, it has a unique stationary distribution, say $\hat{\pi}$. From *discrete* time PASTA, the stationary distribution that the first packet of a batch of Class 1 enters the queue and experiences a delay of n slots, $n < N$, is $\mathbf{P}_{\hat{\pi}}(q_k(1) = n)$. One can now also easily obtain the distribution of the delay faced by a typical packet in a batch of Class 1 as well as the mean stationary delay, $\mathbf{E}[W(1)]$.

Next consider the delays of Class 0 packets. The mean stationary delay, $\mathbf{E}[W(0)]$, of Class 0 can be computed using Little's law as follows: Let $\mathbf{E}[W]$ and $\mathbf{E}[W(1)]$ denote the mean stationary overall system delay and Class 1 delays. We have already computed the quantity $\mathbf{E}[W(1)]$. Then Little's law gives

$$\mathbf{E}[W] = \mathbf{E}[q] \left(\mathbf{E}[X_1(1)](1 - p(1)) + \mathbf{E}[X_1(0)](1 - p(0)) \right), \quad (13)$$

where $\mathbf{E}[q]$ is the mean overall queue length and $p(i)$ is the mean stationary probability of loss of Class i packets (which is obtained by averaging over the probabilities in Equation (12) as we show below). $\mathbf{E}[q]$ can be computed easily from π . $p(0)$ is given by

$$p(0) = \frac{1}{\mathbf{E}[X_k(0)]} \sum_{n,m} \mathbf{P}(X_k(0) \geq n) p(n, m) \mathbf{P}_{\hat{\pi}}((q_k(0) + q_k(1) - 1)^+ \geq N - n - m + 1). \quad (14)$$

Similarly one can compute $p(1)$. Then from

$$\begin{aligned} \mathbf{E}[W] & \left(\mathbf{E}[X_1(0)](1 - p(0)) + \mathbf{E}[X_1(1)](1 - p(1)) \right) \\ & = \mathbf{E}[W(0)] \mathbf{E}[X_1(0)](1 - p(0)) + \mathbf{E}[W(1)] \mathbf{E}[X_1(1)](1 - p(1)), \end{aligned} \quad (15)$$

we obtain $\mathbf{E}[W(0)]$.

Now we consider the distribution and moments of $D(0)$, the delay experienced by a typical Class 0 packet. By *discrete* time PASTA, on arrival, a Class 0 batch sees the queue with distribution $\hat{\pi}$. Let this batch sees $(\tilde{n}_0(1), \tilde{n}_1(1))$, packets on its arrival at time (say) k and let the typical packet be at position n in this batch. Let $(n_0(1), n_1(1))$ be the system state at time $k + 1$. Its distribution conditioned on the typical packet being admitted can be easily found. Denote by $\tau(n_0(1), n_1(1))$ the first time after $k + 1$ when there are no

Class 1 packets in the queue (observe that unlike for the infinite buffer case $\tau(n_0(1), n_1(1))$ depends upon $n_0(1)$ also). Hence, at time $k+1+\tau(n_0(1), n_1(1))$, the first among the Class 0 packets present at time $k+1$ will be served. Let $n'_0(1)$ be the number of Class 0 packets in the queue at time $k+1+\tau(n_0(1), n_1(1))$. Knowing the transition matrix of the Markov chain $\{\hat{q}_n, n \geq 0\}$, and its state $(n_0(1), n_1(1))$ at time k , it is easy to compute the joint distribution of $(\tau(n_0(1), n_1(1)), n'_0(1))$ (using for example the classical technique of taboo probabilities in the Markov chains). Next denote by $Y(n)$ the length of interval after which there will be a packet of Class 1 that enters the queue if, at time $k=0$, $(n, 0)$ is the state of the system. Then, the Class 0 packets will be served for $Y(n'_0(1))$ number of slots after time $k+1+\tau(n_0(1), n_1(1))$. The distribution of $Y(n'_0(1))$ depends only on $n'_0(1)$ and is otherwise independent of everything else. We will compute this distribution later on in this section. At time $k+1+\tau(n_0(1), n_1(1))+Y(n'_0(1))$, let the system state be $(n_0(2), n_1(2))$, where $n_1(2) \leq_{\text{st}} [X_1(1) \mid X_1(1) > 0]$ ($X \leq_{\text{st}} Y$ denotes $\mathbf{P}(X \geq n) \leq \mathbf{P}(Y \geq n)$, for all n). The next time Class 0 packets can be served is $k+1+\tau(n_0(1), n_1(1))+Y(n'_0(1))+\tau(n_0(2), n_1(2))$. Continuing in this fashion and defining

$$N_1 = \inf\{n : Y(n'_0(1)) + \dots + Y(n'_0(n)) > n_0(1) + n\}, \quad (16)$$

$$R = n_0(1) + n - \sum_{i=1}^{N_1-1} Y(n'_0(i)), \quad (17)$$

the waiting time of the tagged packet is

$$\begin{aligned} D(0) &\triangleq \tau(n_0(1), n_1(1)) + Y(n'_0(1)) + \dots + Y(n'_0(N_1 - 1)) + \tau(n_0(N_1), n_1(N_1)) + R. \\ &= n_0(1) + n - 1 + \tau(n_0(1), n_1(1)) + \dots + \tau(n_0(N_1), n_1(N_1)). \end{aligned} \quad (18)$$

Then its distribution is given by

$$\begin{aligned} \mathbf{P}(D(0) = w) &= \sum_{(n_0(1), n_1(1))} \sum_{\mathbf{S}} \mathbf{P}_{\hat{\pi}}(q_k(0) = n_0(1), q_k(1) = n_1(1)) \\ &\quad \cdot \mathbf{P}(\tau(n_0(1), n_1(1)) = m_1, q_{k+m_1}(0) = n'_0(1) \mid \\ &\quad \quad q_k(0) = n_0(1), q_k(1) = n_1(1)) \cdot \mathbf{P}(Y(n'_0(1)) = n_1) \\ &\quad \cdot \mathbf{P}(q_{k+m_1+n_1}(0) = n_0(2), q_{k+m_1+n_1}(1) = n_1(2) \mid \\ &\quad \quad q_{k+m_1}(0) = n'_0(1), Y(n'_0(1)) = n_1,) \\ &\quad \cdot \mathbf{P}(\tau(n_0(2), n_1(2)) = m_2, q_{k+m_1+n_1+m_2}(0) = n'_0(2) \mid \\ &\quad \quad q_{k+m_1+n_1}(0) = n_0(2), q_{k+m_1+n_1}(1) = n_1(2)) \dots \end{aligned} \quad (19)$$

where

$$w = T_r + n_0(1) + n - 1$$

with $T_r = \sum_{i=1}^r m_i$ and \mathbf{S} is the set of all $n_i(j)$, $n'_k(m)$, m_l and n_p such that $\sum_{i=1}^{r-1} n_i \leq n_0 + n < \sum_{i=1}^r n_i$. Then $N_1 = r$. From Equation (18) we can also compute the moments of $D(0)$. For example, the first two moments are given by (denoting the summands on the right hand side of Equation (19) by $\mathbf{F}(\{n_i(j), n'_k(m), m_l, n_p\})$),

$$\mathbf{E}[D(0)] = \sum_{\mathbf{A}} \left[\sum_{i=1}^r \mathbf{E}[\tau(n_0(i), n_1(i))] \right] \cdot \mathbf{F}(\{n_i(j), n'_k(m), m_l, n_p\}) + n_0(1) + n - 1 \quad (20)$$

$$\begin{aligned}
\mathbb{E} \left[(D(0))^2 \right] &= \sum_{\mathbf{A}} \left[\sum_{i=1}^r \mathbb{E} \left[\tau(n_0(i), n_1(i))^2 \right] + \right. \\
&\quad + \sum_i \sum_j \mathbb{E} \left[\tau(n_0(i), n_1(i)) \right] \cdot (n_0(1) + n - 1) \\
&\quad \left. \cdot \mathbf{F}(\{n_i(j), n'_k(m), m_l, n_p\}) + (n_0(1) + n - 1)^2, \right. \quad (21)
\end{aligned}$$

where \mathbf{A} is the set of all $n_i(j)$, $n'_k(m)$, m_l and n_p , and T_r is as defined above. Thus, to be able to compute the distribution and moments of the waiting time, we need the distributions and moments of $\tau(n_0(i), n_1(i))$ and $Y(n'_0(i))$. We have already mentioned the method to obtain the joint distribution of $(\tau(n_0(i), n_1(i)), n'_0(1))$. Below we provide efficient algorithms to compute the moments of $\tau(n_0(i), n_1(i))$ and the distribution of $Y(n'_0(i))$.

First consider $\tau(n_0, n_1)$. We need to consider for $n_1 > 0$ only. Then it satisfies

$$\begin{aligned}
\mathbb{E} [\tau(n_0, n_1)] &= \mathbf{1}_{\{n_1=1\}} \cdot \mathbb{P}(X_{k+1}(1) = 0) + \sum_{n=0}^{\infty} \sum_{m=1}^{\infty} \mathbb{P}(X_{k+1}(0) = n) \\
&\quad \cdot \mathbb{P}(X_{k+1}(1) = m) (1 + \mathbb{E} [\tau(n_0 + n', n_1 - 1 + m')]), \quad (22)
\end{aligned}$$

where n' and m' are truncated values of n and m such that the queue length does not exceed N (taking into consideration the ordering of new packets arriving in slot $k + 1$ – this actually requires knowledge of the joint distribution of the two arrival streams within a slot which we are assuming in this paper) and $n_0, n_1 = 1, \dots, N$. We can solve this set of equations for $\mathbb{E} [\tau(n_0, n_1)]$, $n_0, n_1 = 1, \dots, N$. The second moments $\mathbb{E} [(\tau(n_0, n_1))^2]$ satisfy

$$\begin{aligned}
&\mathbb{E} [(\tau(n_0, n_1))^2] \\
&= \mathbf{1}_{\{n_1=1\}} \cdot \mathbb{P}(X_k(1) = 0) + 2 \sum_{n=0}^{\infty} \mathbb{E} [\tau(n_0 + n', n_1 - 1 + m')] \\
&\quad + \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \mathbb{P}(X_k(0) = n) \mathbb{P}(X_k(1) = m) \left(1 + \mathbb{E} [(\tau(n_0 + n', n_1 - 1 + m'))^2] \right) \quad (23)
\end{aligned}$$

where n' and m' are as in (22). Similarly, we can compute the higher moments of $\tau(n_0(1), n_1(1))$.

Now consider $Y(n)$. Let $p(n)$ denote the probability that the first batch of Class 1 will be completely lost if the queue starts at $k = 0$ in the state $(n, 0)$. Then,

$$Y(n) = \sum_{k=0}^S Z_k, \quad (24)$$

where

$$\mathbb{P}(S = m) = p(N)^{m-1} (1 - p(N)) p(n), \quad \text{for } m > 0, \quad (25)$$

$\mathbb{P}(S = 0) = 1 - p(n)$, and $\{Z_i\}$ is an i.i.d. sequence with

$$\mathbb{P}(Z_i = m + 1) = \mathbb{P}(X_1(1) = 0)^m \mathbb{P}(X_1(1) > 0). \quad (26)$$

To compute the distribution of $Y(n)$, we further need to compute $p(n)$. For this we solve the following set of equations:

$$\begin{aligned}
p(n) &= \mathbb{P}(X_1(1) > 0) q(N - (n - 1)^+, 1) \\
&\quad + \sum_{m=0}^{\infty} \mathbb{P}(X_1(1) = 0) \mathbb{P}(X_1(0) = m) p((n - 1)^+ + m) \quad (27)
\end{aligned}$$

where $q(i, j)$ is the probability that in a slot i Class 0 packets arrive before the j th packet of Class 1.

The above results can be used to easily obtain the moments and the distribution of waiting times of an *admitted* Class i packet. This statement will hold in the rest of the paper also and we will not mention it again.

4 Finite Queue with Partitioning

Let (N_0, N_1) be a partition of the buffer length N . Buffer length N_i is reserved exclusively for Class i . Class 1 has higher time priority than Class 0. Let $\hat{q}_k = (q_k(0), q_k(1))$ denote the state of the system, with $q_k(i)$ being the queue length in the i th buffer. Let $\hat{\pi}$ be the stationary distribution of the Markov chain $\{\hat{q}_k\}$. Then, the probability that the n th packet of a Class 1 batch is lost is $P_{\hat{\pi}}((q_k(1) - 1)^+ \geq N_1 - n + 1)$, and the probability that its delay is m slots is $P_{\hat{\pi}}((q_k(1) - 1)^+ = m - n + 1)$.

Next consider the Class 0 packets. Consider a typical Class 0 packet and suppose it is the n th packet of a Class 0 batch. By discrete time PASTA, the probability that this packet is lost is

$$P_{\hat{\pi}}(q_k(0) \geq N_0 - n + 1, q_k(1) > 0) + P_{\hat{\pi}}((q_k(0) - 1)^+ \geq N_0 - n + 1, q_k(1) = 0). \quad (28)$$

Now consider its delay. Let the queue lengths at time k be (n_0, n_1) . At the end of the slot, the state will be

$$\left(\min\{(n_0 - \mathbf{1}_{\{n_1=0\}})^+ + X_k(0), N_0\}, \min\{(n_1 - 1)^+ + X_k(1), N_1\} \right)$$

Let us write this state to be (n'_0, n'_1) . Let τ be the inter-visit time to the epochs when there are no Class 1 packets in the system and let $\{\tau_1, \tau_2, \dots\}$ be i.i.d. with the distribution of τ . Also let $\tau(n)$ be the first time there will be no packets of Class 1 if at time 0 there are n Class 1 packets at time 0. Efficient algorithms for computing the distribution and moments of τ and $\tau(n)$ are available in Sharma and Gangadhar [13]. Then the typical packet considered above will be served at time $k + \tau(n'_1) + \tau_1 + \dots + \tau_{n_0 - \mathbf{1}_{\{n_1=0\}} + n - 1}$. Thus the distribution and the moments of the delay of the typical Class 0 packet can be computed.

5 Markov Modulated Arrivals

In many practical systems, the arrival streams have correlations and the independence assumption on the arrival traffic that we used until now in this paper do not hold. In this section we show that all our previous results can be easily extended to the case of Markov modulated arrival streams.

Let $\{M_n, n \geq 0\}$ be a finite state irreducible (hence ergodic) Markov chain with transition matrix $(p_{i,j})$. The distribution of the number of Class 0 and Class 1 arrivals in slot k , $X_k(0)$ and $X_k(1)$ respectively, depends upon the state M_k and is independent of everything else. (One could consider $X_k(0), X_k(1)$ modulated by two different independent Markov chains. But that would be a special case of our model). In the following subsections we present the changes that are required in the analysis from the i.i.d. case.

5.1 Infinite Buffer Queue

Consider the system in Section 2. In case of Markov modulated arrivals, $\{(q_k(0), q_k(1), M_k)\}$ forms a Markov chain. It is ergodic if $E_\pi[X_k(0)] + E_\pi[X_k(1)] < 1$. Let this condition hold for infinite buffer case and under this condition, let $\hat{\pi}$ be its stationary distribution. Efficient algorithms to compute $\hat{\pi}$ are available in Akar et al.[1] and Ramaswami [10].

Now consider the delays of Class 1. Since the Class 0 packets do not affect them, we can consider a single class queue with only Class 1 arrivals. To obtain the delays we need the distribution $\tilde{\pi}$ of the state of the system that a batch of Class 1 sees on arrival. Now discrete time PASTA does not hold. However $\tilde{\pi}$ can be obtained from $\hat{\pi}$ using the relation

$$\begin{aligned} \tilde{\pi}(A) &= P_{\hat{\pi}}((q_k(0), q_k(1), M_k) \in A | X_k(1) > 0) \\ &= \frac{P_{\hat{\pi}}(X_k(1) > 0 | (q_k(0), q_k(1), M_k) \in A) \hat{\pi} P_{\hat{\pi}}((q_k(0), q_k(1), M_k) \in A)}{P_{\hat{\pi}}(X_k(1) > 0)} \\ &= \frac{\sum_{r \in R} P(X_k(1) > 0 | M_k = r) P_{\hat{\pi}}((q_k(0), q_k(1), M_k) \in A)}{\sum_s P(X_k(1) > 0 | M_k = s) P_{\hat{\pi}}(M_k = s)}, \end{aligned} \quad (29)$$

where $R = \{r : (q_k(0), q_k(1), r) \in A\}$ (for delays of Class 1 we can replace the Markov chain $\{q_k(0), q_k(1), M_k\}$ by $\{q_k(1), M_k\}$). This relationship between $\tilde{\pi}$ and $\hat{\pi}$ is also obtained from the general relationship between time stationary and Palm probabilities of a stationary process available *e.g.*, in Brandt *et al.* [2]. This relationship holds for finite as well as infinite buffer systems. Using $\tilde{\pi}$, as shown in Section 3, we can compute the distribution and moments of the delay of a typical Class 1 packet.

Consider the delays of Class 0 (say of the n th packet, the typical packet, in a batch). From the relation analogous to (29) for Class 0, we have the distribution of the state of the system that a Class 0 batch sees on its arrival (say in slot k). Let $(\tilde{n}_0, \tilde{n}_1)$ be the state of the queues at time k and (n_0, n_1, i_0) be the state at time $k+1$. We can find its distribution conditioned on the typical packet being admitted. Let $k+1+\tau(i_0)$ be the first time that the system does not have Class 1 packets. Also, let the state of M_n at $k+1+\tau(i_0)$ be i_1 . We can easily compute the joint distribution of $(\tau(i_0), i_1)$ (say using the taboo probabilities). Now, given $(\tau(i_0), i_1)$, the distribution of the next time $(\tau(i_1), i_2)$ the queue becomes empty can be computed, and so on. The waiting time of the n th packet in the tagged Class 0 batch is

$$\tau(i_0) + \tau(i_1) + \cdots + \tau(i_T), \quad (30)$$

where $T = (\tilde{n}_1 + \tilde{n}_0 - 1)^+ + [X_k(1) | M_k = i_0] + n - 1$. The distribution and the moments of this delay can be computed once we have for $(\tau(i_0), i_1)$.

5.2 Finite Buffer with Complete Sharing

This section corresponds to Section 3. Again $\{(q_k(0), q_k(1), M_k)\}$ is a Markov chain, with a stationary distribution (say) $\hat{\pi}$.

The stationary distribution $\tilde{\pi}$ that a Class 1 batch sees on its arrival can be obtained from (29). Hence we can determine the probability of loss and distribution of delay of accepted Class 1 packets. Similarly we can determine the probability of loss of Class 0 packets.

By Little's law, we have Equations (13) and (15). Thus we can compute the mean delay of Class 0 packets. For computing the stationary delay distribution and moments of the delay of a typical Class 0 packet, we need to consider the quantity $\tau(n_0(1), n_1(1), i_1)$ which is the first time after $k+1$ when there are no Class 1 packets in the system if

at time $k + 1$ the state of the system is $(n_0(1), n_1(1), i_1)$. Let the state of the system at $k + 1 + \tau(n_0(1), n_1(1), i_1)$ be $(n'_0(1), 0, i'(1))$. We can compute the joint distribution of $\tau(n_0(1), n_1(1), i_1)$ and $(n'_0(1), 0, i'(1))$ using the taboo probability method from the transition matrix of the system Markov chain. Also, the length of time for which no Class 1 packet arrives, $Y(n'_0(1), i'_1)$ now depends on i'_1 also. With these changes (16) – (21) hold (with $\{M_k\}$ states appropriately filled in). The computational procedures for computing the distribution and moments of $\tau(n_0(j), n_1(j), i_j)$ can be modified for the Markov modulated case. For example (22) becomes

$$\begin{aligned} & \mathbb{E}[\tau(n_0, n_1, i_j)] \\ &= \mathbf{1}_{\{n_1=1\}} \cdot \mathbb{P}(X_k(1) = 0 \mid M_k = i_j) + \sum_{i_k} \sum_{n=0}^{\infty} \sum_{m=1}^{\infty} \mathbb{P}(X_k(0) = n \mid M_k = i_j) \\ & \quad \cdot \mathbb{P}(X_k(1) = m \mid M_k = i_j) (1 + \mathbb{E}[\tau(n_0 + n', n_1 - 1 + m', i_k)]) \cdot p_{i_j, i_k}, \end{aligned} \quad (31)$$

where $\{p_{i_j, i_k}\}$ are the transition probabilities of $\{M_n\}$ and n' and m' are as in (22). One can also adapt the procedure for computing $Y(n'_0(j), i'_j)$ by conditioning on i'_j . Now $p(n, i)$ denotes the probability that the first Class 1 batch that arrives after time 0 will be lost if the state of the system at time 0 is $(n, 0, i)$. Then

$$Y(n, i) = \sum_{l=0}^S Z_l(i_l), \quad (32)$$

where i_l is the state of the modulating Markov chain at the arrival instant of the l th lost Class 1 batch and $Z_l(i_l)$ depend only on i_l . Now the joint distribution of $(Z(i_l), i_l)$ can be determined from the transition matrix of the modulating Markov chain and $\mathbb{P}(X_k(1) = 0 \mid M_k = i_k)$. From this we can compute $p(n, i)$, when the initial state of the Markov chain $\{M_k\}$ is i . Then the distribution of S is

$$\begin{aligned} \mathbb{P}(S = 0) &= 1 - p(n, i) \\ \mathbb{P}(S = m) &= p(n, i) \left[\prod_{k=1}^{m-1} \mathbb{P}(Z_{k-1} = n, M_n = i_k \mid M_1 = i_{k-1}) p(N, i_k) \right] (1 - p(N, i_m)). \end{aligned}$$

5.3 Finite Buffer with Partitioning

This section corresponds to Section 4 where the buffer space is partitioned into N_0 and N_1 but Class 1 has higher time priority. Again, $\{(q_k(0), q_k(1), M_k)\}$ forms a Markov chain and let $\hat{\pi}$ be its stationary distribution. The probability of loss of a typical packet of either class is readily computable from the stationary distribution of the state that a batch from that Class sees on its arrival. This quantity can be determined as before from (29). Also from this, due to their higher time priority, the distribution of delay of Class 1 packets can be known.

In order to compute the delay of a typical Class 0 packet, we need to replace τ_n by $\tau_n(i_n)$ and $\tau(n)$ by $\tau(n, i)$, where i_n is the state of the modulating Markov chain at the beginning of the busy period of Class 1 packets corresponding to τ_{n+1} . The joint distribution of $(\tau(n_1, i_1), i_2)$ etc. can be computed using the taboo probability method from the transition matrix of the system Markov chain.

References

- [1] N. Akar, N.C. Oguz, and K. Sohraby. Matrix Geometric Solutions of M/GI/1-type Markov Chains: A Unifying Generalized State Space Approach. *IEEE J. Selet. Areas Commun.*, 16:626–639, 1998.
- [2] A. Brandt, P. Franken, and B. Lisek. *Stationary Stochastic Models*. John Wiley and Sons Ltd., Chichester, 1990.
- [3] H. Bruneel and B.G. Kim. *Discrete Time Models for Communication Systems Including ATM*. Kluwer Academic Publishers, Boston, 1993.
- [4] N. D. Gangadhar. Analysis of Discrete Time Queues with Applications to ATM based B-ISDNs. M.Sc. (Engg.) Thesis, Department of Electrical Engineering, Indian Institute of Science, Bangalore, India, Apr. 1995.
- [5] J. Garcia and O. Casals. Priorities in ATM Networks. NATO Advanced Research Workshop, June 1996.
- [6] K. Kawahara, K. Kitajima, T. Takine, and Y. Oie. Packet Loss Performance of Selective Cell Discard Schemes in ATM Networks. *IEEE JSAC*, 15(5):903–913, 1997.
- [7] J. Y. Le Boudec. An Efficient Solution Method for Markov Models of ATM Links with Loss Priorities. *IEEE JSAC*, 9(3), 1991.
- [8] K.-Q. Liao. Queueing Analysis of Partial Buffer Sharing with Markov Modulated Poisson Inputs. In *ITC-14*, pages 55–64. Elsevier, 1995.
- [9] A. Y.-M. Lin and Silvester J. Priority Strategies and Buffer Allocation Protocols for Traffic Control at an ATM Integrated Broadband Switching System. *IEEE JSAC*, 9(9):1524–1536, 1991.
- [10] V. Ramaswami. A Tutorial Overview of Matrix Analytic Methods: Some Extensions and New Results. in *Matrix Analytic Methods*, N.Y., pages –, 1996.
- [11] J. Roberts, U. Mocci, and J.(Ed.) Virtamo. *Broadband Network Teletraffic*. Springer Verlag, New York, 1996.
- [12] I. Rubin and Z. Tsai. Message Delay Analysis of Multiclass Priority TDMA, FDMA, and Discrete Time Queueing Systems. *Trans. Info. Theory*, IT-35:637–647, 1991.
- [13] V. Sharma and N. D. Gangadhar. Some Algorithms for Discrete Time Queues with Finite Capacity. *Queueing Systems*, 25:281–305, 1997.
- [14] V. Sharma and N.D. Gangadhar. A Finite Discrete Queue with Time and Space Priorities. In to be presented in 3rd IFIP TC6 workshop on Traffic management and Design of ATM Networks, London, pages –, 1999.
- [15] Y. Takahashi and O. Hashida. Delay Analysis of Discrete-Time Priority Queue with Stuctured Inputs. *Queueing Systems*, 8:149–164, 1991.
- [16] D. Wagner. Analysis of a Multi-server Model with Nonpreemptive Priorities and Non-renewal Input. In J. Labetoulle and J. W. Roberts, editors, *ITC-14*, pages 757–767. Elsevier, 1995.