

A two-server queueing system with periodic and credit-based server availabilities*

Ioannis Stavrakakis

*Department of Electrical Engineering and Computer Science, University of Vermont,
Burlington, VT 05405, USA*

Email: stavraka@emba.uvm.edu

A discrete-time, two-server queueing system is studied in this paper. The service time of a customer (cell) is fixed and equal to one time unit. Server 1 provides for periodic service of the queue (period T). Server 2 provides for service only when server 1 is unavailable and provided that the associated service credit is nonzero. The resulting system is shown to model the queueing behavior of a network user which is subject to traffic regulation for congestion avoidance in high speed ATM networks. A general methodology is developed for the study of this queueing system, based on renewal theory. The dimensionality of the developed model is independent of T ; T increases with the network speed. The cell loss probabilities are computed in the case of finite capacity queue.

Keywords: Traffic regulation, ATM, loss probabilities, buffer behavior.

1. Introduction

Discrete-time queueing systems are extensively adopted for the modeling of distributed and centralized media access protocols in modern communication and computer systems. Such systems are naturally formulated for the description of resource allocation policies for slotted networks transmitting information organized into small units of fixed length (cells).

The queueing system considered in this paper may serve as a model of the buffer of a user of a high-speed Asynchronous Transfer Mode (ATM) network, which is subject to traffic regulation for network congestion avoidance. Network congestion avoidance (preventive control) is widely considered to be the most promising approach for traffic congestion management in the emerging high-speed ATM networks. An extensive survey of Congestion Avoidance Mechanisms (CAMs) – called bandwidth enforcement mechanisms or traffic regulators – developed for the implementation of such control may be found in [1]. The object function of a CAM is to control the flow of the user traffic to the network in a way that serious network congestion be avoided; no network state information is

*Research supported by the National Science Foundation under grant NCR-9011962.

assumed to be available to the users. Although a specific application of the studied queueing model is considered – namely a version of the Leaky Bucket CAM [1–11] – its functions seem to capture the two general objectives associated with the design of a CAM, as explained in section 3. Thus, the model could be useful for the study of other CAMs, whose design is based on a similar philosophy.

The queueing system is described in the next section. It is a discrete-time two-server queueing system with deterministic customer service time (one time unit or slot). One of the servers provides for periodic service to the customers (called cells in this paper). Unlike traditional multiple-server systems, the availability of the second server does not increase the potential total amount of service offered to the queue, but provides to the queue some of the unused capacity of the first server, which had become available to the queue in the past (accumulated service credit). Some limit is being imposed on the maximum amount of credit that can be accumulated. The two servers cannot operate simultaneously. It turns out that the availability of the second server depends on the queue occupancy history and the position of the current time slot within the time frame determined by the periodic service pattern of the first server.

In section 4, a methodology based on renewal theory is developed for the calculation of the cell loss probabilities, when the queue has finite capacity. It turns out that the cell loss probabilities can be computed in terms of the solution of linear equations. This performance measure has been considered in [4, 7, 8]. Poisson cell arrivals have been assumed in [4]. A special case of the Leaky Bucket described in terms of a single counter has been studied in [7] in terms of a $G/D/1$ queueing model with finite capacity and in [8] by using a fluid flow approximation. The general case of the Leaky Bucket under correlated arrivals and infinity queue capacity has been considered in [5, 6]. The queue occupancy distribution was derived based on matrix-geometric techniques and z -transforms (see section 4).

Finally, some numerical results, a comparison with the tail of the queue occupancy distribution under infinite capacity [6] and a final discussion on the applicability of the developed approach are presented in the last section.

2. The two-server queueing system

The discrete-time queueing system considered in this paper is shown in figure 1. All events are assumed to occur at the discrete-time instants of the system (discrete) time axis. The time unit (slot) is assumed to be equal to the fixed service time of the customers, which will be called cells (of information).

The source is assumed to generate cells according to a Markov Modulated Bernoulli (MMB) process, based on an underlying Markov chain with state space $S = \{0, 1\}$. Let $\pi(i)$ denote the steady state probability that the Markov chain is in state i , $i \in S$; let $p(i, j)$ denote the transition probability from state i to state j ,

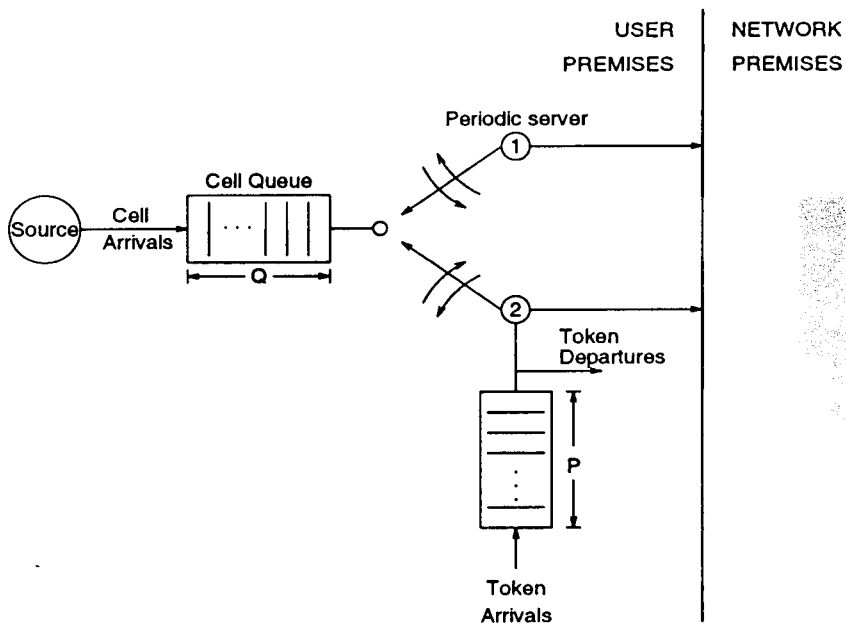


Fig. 1. The two-server queuing system.

$i, j \in S$. No cell is generated when the source is in state 0; one cell is generated with probability p , $0 \leq p \leq 1$, when the source is in state 1. Let λ denote the cell rate of the source. Clearly, $\lambda = \pi(1)p$. Let $\gamma = p(1, 1) - p(0, 1)$ be defined to be the burstiness coefficient of the cell source; $\gamma = 0$ corresponds to a Bernoulli source with rate λ . For $\gamma > 0$, a larger value of γ would result in larger clusters of packets. It is easy to establish that the Markov Modulated Bernoulli process can be completely described in terms of the triplet $\{\lambda, p, \gamma\}$. The cells are temporarily stored in the Cell Queue (CQ) of capacity Q . They are discarded (lost) if upon cell arrival the CQ is found to be full. Cell arrivals are declared at the beginning of the slots and they are assumed to leave the CQ momentarily at the beginning of their service slot.

Server 1 visits the CQ periodically every T time slots. It remains at the CQ for one time slot, providing service to one cell (assuming the CQ is non-empty). Then, it switches away from the CQ.

Unlike the guaranteed periodic availability of server 1, the availability of server 2 depends on the existence of service credit. One service credit unit (or token) is required for the service of one cell by server 2. Service tokens arrive to and depart from the credit (Token) Pool (TP) as described later. Let $[x]_k$ denote the content of entity x at the discrete-time instant k ; let \bar{t}_1 be the end of a slot at which server 1 switches away from the CQ. The following steps describe the service policy associated with server 2.

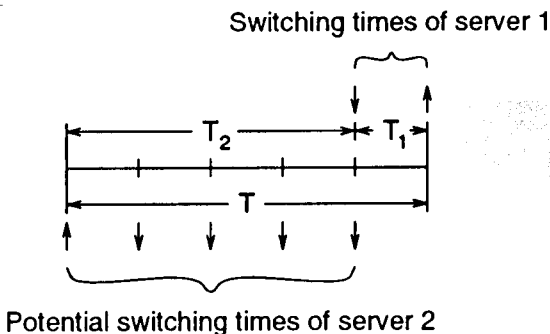


Fig. 2. The framed structure of time. Server 1 is always available to the CQ over T_1 . Server 2 can be available to the CQ over T_2 , provided that sufficient service credit is available.

- (a) If $[TP]_{\bar{t}_1} = 0$, server 2 remains away from the CQ. At time $\bar{t}_1 + T$ the value of $[TP]_{\bar{t}_1 + T}$ is re-examined. As it will be seen below, the content of the TP cannot increase before $\bar{t}_1 + T$.
- (b) If $[TP]_{\bar{t}_1} > 0$, server 2 switches to the CQ and remains there for up to $\min\{T - 1, [TP]_{\bar{t}_1}\}$ time slots, providing service to up to $\min\{T - 1, [TP]_{\bar{t}_1}\}$ cells. Then, it switches away from the CQ.

The (potential, for server 2) switching times of servers 1 and 2 are depicted in figure 2, where time is shown to be organized in frames of length T . Server 2 switches to the CQ at the beginning of the frame if $[TP] > 0$. It switches away from the CQ at the latest at the end of the $(T - 1)$ st slot of the frame or at the end of the slot at which no credit is left in the TP. Server 1 always switches to the CQ at the beginning of the T th slot of a frame and switches away from the CQ at the end of the frame.

One token (credit time) is generated at the end of a slot over which server 1 was idle, while at the CQ. That is, service tokens represent the unused capacity of server 1 which has been available to the CQ. These tokens are forwarded to the TP at the end of the generation slot, which coincides with the end of some frame. The TP capacity is assumed to be equal to P . Thus, the maximum service credit (unused capacity of server 1) which can become available to the CQ at any time slot is equal to P . This service credit is provided at a rate of one token per slot, provided that the CQ is non-empty and the slot is not the last of a frame. Recall that server 2 is never allowed to provide service to the CQ over the last slot of a frame. At the end of a slot which a cell has been served by server 2, a token leaves the TP (used token).

The two-server queueing system described above may be adopted for the modeling of the queueing behavior of an ATM (Asynchronous Transfer Mode) network user which is subject to traffic regulation for network congestion avoidance. This application is described in the next section, along with a rather non-traditional – but insightful – description of the function of a CAM.

3. Application to network congestion avoidance mechanisms – the leaky bucket case

As stated earlier, the purpose of a Congestion Avoidance Mechanism (CAM) is to regulate the traffic of a bursty source, so that the potential for network congestion (or stress) be reduced. At the same time, the traffic regulation should be controlled in a way that the induced stress (quality of service deterioration) at the user premises be acceptable. Apparently, there is a trade-off between the amount of the induced user stress and removed network stress, which is determined in a way that key performance indices (such as cell loss probabilities, network utilization, etc.) be maximized. The desired balance between user and network stress may be seen as the product of the impact of two separate functions implemented by a CAM: the primary and the secondary, as explained below.

The primary function of a CAM removes network stress by modulating the user traffic so that the resulting network stress be potentially minimized. Such a primary function is the one which would spread the cells uniformly in time with minimum time separation T . This periodic cell delivery to the network reduces the network stress by reducing the burstiness of the original source traffic. At the same time, the maximum cell rate delivered to the network is equal to $R_T = 1/T$; typically, R_T would be chosen to be greater than but close to the cell rate R of the regulated source. The periodic server 1 in the two-server queueing model presented in section 2 is responsible for the implementation of this primary function of a CAM.

The primary function of a CAM induces significant user stress as manifested by the increased intensity of the queueing problems at the user premises. The secondary function of a CAM is designed to alleviate the increased user stress, by transferring a controlled amount of it back to the network. This task is implemented by server 2 in the two-server queueing model. The amount of user stress transferred back to the network is controlled through the bounded credit. This credit is generated only when the user does not use the allocated periodic slots and, thus, after it induces additional stress reduction to the network. Thus, credit becomes available to the user following a period of reduced network stress. To prevent a temporary large increase of the network stress, due to the usage of a large amount of accumulated credit, the amount of credit is bounded. Notice that this credit-based user relief guarantees that the long term cell rate delivered to the network be bounded by R_T . This constraint is important for the design of efficient call acceptance policies. The token arrival process to the TP in the two-server queueing model is easily seen to be in accordance with the above described credit-based user relief.

In the sequel, the Leaky Bucket CAM is described; its primary and secondary functions are identified and the applicability of the two-server queueing model system to its study is easily established. One version of the Leaky Bucket CAM is shown in figure 3. It will be referred to as the double token pool Leaky Bucket (d-LB). The cell traffic delivered to the network by the user is controlled by means of tokens. Tokens are generated periodically with period T (slots). They are stored

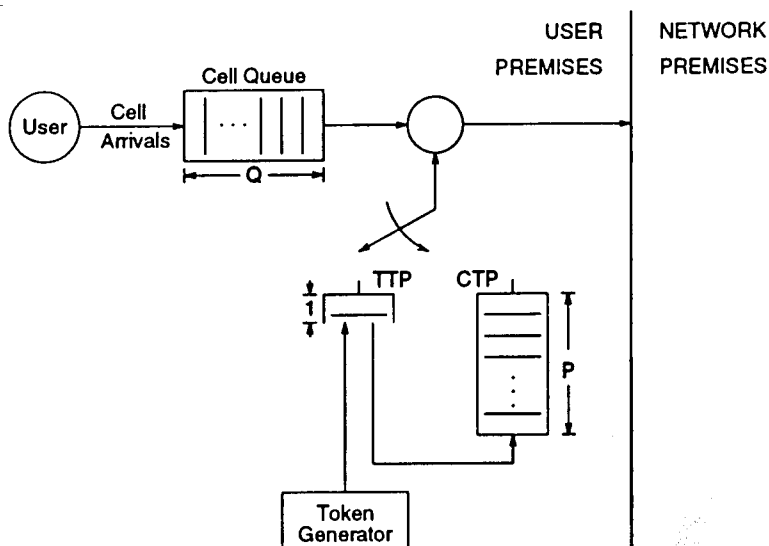


Fig. 3. The Leaky Bucket congestion avoidance mechanism.

in the Temporary Token Pool (TTP) or the Credit Token Pool (CTP) as determined by the token arrival protocol described below. The cells (of information) generated by the user are temporarily stored in the Cell Queue (CQ), before they are delivered to the network. A cell (token) that finds the CQ (TTP or CTP) full is discarded. In order for a cell to be delivered to the network, it must obtain a token from either the TTP or the CTP. The operation of the d-LB is completely described in terms of the token arrival and token release protocols; a token release enables the transmission of the cell at the head of the CQ.

Token arrival protocol: Tokens arrive to the TTP periodically with period T ; token arrivals are declared at the beginning of the token arrival slot. If a token is not released (used) over its arrival slot, it is forwarded to the CTP by the end of this slot. If the CTP is full, the token is discarded. The capacities of the TTP and CTP are equal to 1 and P , respectively.

Token release protocol: Let t be the time instant (beginning of the slot) when a cell is forwarded to the head of the CQ. Let $[x]_t$ denote the content of entity x at time t .

- If $[TTP]_t = 1$, the token is released from the TTP immediately (and, thus, the cell at the head of the CQ is delivered to the network).
- If $[TTP]_t = 0$ and $[CTP]_t = 0$, the token is released from the TTP as soon as it arrives.
- If $[TTP]_t = 0$ and $[CTP]_t > 0$, a token is released from the CTP.

The d-LB described above is virtually identical to the standard Leaky Bucket

(s-LB) which has been widely considered in the past. A single token pool is used for the token storage in the s-LB. This difference between the s-LB and d-LB is insignificant, though. The d-LB version of the Leaky Bucket will be assumed in any reference to the Leaky Bucket in the rest of the paper.

In view of the above discussion, it is easily established that the two-server queueing system described in the previous section can model exactly the queueing behavior of the CQ of the Leaky Bucket CAM. The tokens provided directly from the TTP correspond to the service provided by server 1 and implement the primary function. The tokens provided through the CTP correspond to the service provided by server 2 and implement the secondary function.

4. Analysis of the two-server queueing system

In this section the queueing system described in section 2 is analyzed. The study is focused on the derivation of the cell loss probabilities in the CQ. Notice that the cell loss probability is an important performance measure of the Leaky Bucket CAM which can be modeled by this two-server queueing system, as illustrated in section 3.

The evolution of the two-server queue (figure 1) can be completely described in terms of the Markov chain $\{(I_j, T_j, P_j, Q_j)\}_{j \geq 0}$. $\{I_j\}_{j \geq 0}$ is a Markov chain which describes the state of the cell source; let $|S^I|$ denote the cardinality of its state space. $\{T_j\}_{j \geq 0}$ is a Markov chain which determines whether server 1 will be available to the CQ in the current slot or not. It is a periodic Markov chain which indicates the position of the current slot within a frame. If the current slot is the last of the frame, then server 1 will become available to the CQ. $\{T_j\}_{j \geq 0}$ evolves according to the rule $T_j = j \bmod T + 1$; its state space is given by $S^T = \{1, 2, \dots, T\}$ with cardinality $|S^T| = T$. $\{P_j\}_{j \geq 0}$ is a process describing the number of tokens in the TP at the beginning of the current slot. Its state space is given by $S^P = \{0, 1, \dots, P\}$ with cardinality $|S^P| = P + 1$, where P is the TP capacity. Finally, $\{Q_j\}_{j \geq 0}$ is a process describing the number of cells in the CQ at the beginning of the current slot. Its state space is given by $S^Q = \{0, 1, \dots, Q\}$ with cardinality $|S^Q| = Q + 1$, where Q is the CQ capacity. The state space of $\{(I_j, T_j, P_j, Q_j)\}_{j \geq 0}$ has cardinality equal to $|S^I|T(P + 1)(Q + 1)$.

When the cell source is memoryless, the two-server system may be described in terms of the Markov chain $\{(T_j, P_j, Q_j)\}_{j \geq 0}$. A Markov chain of this type has been considered in [5] for the study of the queueing behavior of the standard Leaky Bucket when the CQ capacity is infinite. Matrix-geometric approach was followed for the derivation of the CQ occupancy distributions. Process $\{(T_j, P_j)\}_{j \geq 0}$ was defined as the process involved in the application of this analytic approach. The complexity of this approach is largely due to the need to solve non-linear matrix equations involving matrices of dimension $|S^T||S^P| \times |S^T||S^P| = T(P + 1) \times T(P + 1)$. The solution of these equations becomes a severe problem beyond relatively small values of T and P .

A similar approach has been followed in [6] for the study of the queueing behavior of the standard Leaky Bucket, when the CQ capacity is infinite and the cell source Markovian. The formulated four-dimensional Markov chain $\{(I_j, T_j, P_j, Q_j)\}_{j \geq 0}$ was reduced to a three-dimensional $\{(I_j, T_j, Q'_j)\}_{j \geq 0}$. The queue occupancy distribution of the resulting equivalent system was obtained using z-transform techniques. Finally, the queue occupancy distribution of the original system was computed recursively from that of the equivalent system. The calculation of boundary probabilities required by this technique as well as inversion of the z-transform, are the major complexity issues associated with this approach.

In this section, a methodology based on renewal theory is developed for the calculation of the cell loss probabilities, when the CQ has finite capacity. This performance measure has not been considered in the above mentioned past work for the analysis of the standard Leaky Bucket, where the CQ capacity has been assumed to be infinite. It turns out that the cell loss probabilities can be computed in terms of the solution of $|S|(P + Q + 1)$ linear equations; $|S|$ denotes the cardinality of the state space of the MMB process defined in section 2.

Let the balance process $\{B_j\}_{j \geq 0}$ be defined by

$$B_j = Q_j - P_j, \quad (1)$$

where $\{Q_j\}_{j \geq 0}$ and $\{P_j\}_{j \geq 0}$ are the CQ and TP occupancy processes, respectively. Let $S^B = \{i: i \in \mathbb{Z}, -P \leq i \leq Q\}$ denote the state space of $\{B_j\}_{j \geq 0}$, where \mathbb{Z} denotes the set of integer numbers. As it will become clear shortly, the balance process will be used as a means of reducing the dimensionality of the solution approach developed in this paper, in the same way that the deficit function reduces the dimensionality of the solution approach followed in [6]. Although the two quantities are used in a different manner and are technically differently defined, they are virtually identical. The following lemma will be needed for the proof of the theorem that follows.

LEMMA

Let time instant j mark the beginning of (the first slot of) a frame. Let $k, k \geq 0$ be the number of cell arrivals over this frame. Then at least $\min\{[TP]_j + 1, k\}$ cells will be served within this frame.

Proof

Notice that the maximum number of cell arrivals per slot, denoted by \hat{R}_c , is equal to one. The number of tokens that can be released per slot, denoted by \hat{R}_r , is equal to one, provided that tokens (either new or in the TP) exist. Since $\hat{R}_c \leq \hat{R}_r$, it is implied that, provided that tokens exist, they will be provided to the cells at a rate at least equal to the maximum cell arrival rate over a slot. Thus, if k cells arrive over a frame, at least k tokens, if $[TP]_j + 1 > k$, will be released over the same frame and,

thus, at least $\min\{[TP]_j + 1, k\}$ cells will be served before the end of the frame; $[TP]_j + 1$ is the total number of tokens, including the new token generated at the last slot of the frame. \square

A direct consequence of lemma 1 is stated below, using the terminology introduced in lemma 1.

COROLLARY 1

The CQ content cannot be increased between consecutive frame boundaries, unless the TP becomes empty. That is, if $Q_{j+\tau} > Q_j$ then $P_{j+\tau} = 0$.

Let $|a|^+ = a$ if $a \geq 0$ and zero otherwise; let $|a|^- = |a|$ if $a \leq 0$ and zero otherwise. The following theorem is important for the derivation of the state equations for the two-server queueing system. Its proof may be found in appendix A.

THEOREM 1

Let j denote the beginning of (the first slot of) a frame. Then,

$$Q_j = [CQ]_j = |B_j|^+ \quad \text{and} \quad P_j = [TP]_j = |B_j|^-. \quad (2)$$

That is, if B_j is positive, it is equal to the CQ content and implies that the TP is empty; if B_j is negative, its absolute value is equal to the TP content and implies that the CQ is empty; if B_j is zero, it implies that both the CQ and the TP are empty.

Let $\{(I_m, B_m)\}_{m \geq 0}$ be a process imbedded at the beginning of the frames. B_m denotes the value of the balance process at the beginning of the first slot of the m th frame. I_m denotes the state of the Markov chain of the cell source at the beginning of the last slot of the $(m-1)$ st frame; cell arrivals due to the visit to the I_m state occurred (if any) at the beginning of the last slot of the $(m-1)$ st frame. In view of theorem 1, it is easily established that $\{(I_m, B_m)\}_{m \geq 0}$ is a Markov chain imbedded at the frame boundaries; its state space is given by $S \times S^B = \{0, 1\} \times \{i; i \in Z, -P \leq i \leq Q\}$. From now on, the state of $\{(I_m, B_m)\}_{m \geq 0}$ will define the state of the (two-server queueing) system.

Let $\{M_m\}_{m \geq 0}$ be a sequence of frame boundaries (time instants) at which the process $\{(I_m, B_m)\}_{m \geq 0}$ visits a specific state, say state $(0, -P)$; this state is visited after a sufficiently long period of source inactivity which renders the CQ empty and the TP full. Clearly $\{M_m\}_{m \geq 0}$ is a renewal sequence. Let $\{X_m\}_{m \geq 0}$ denote the sequence of time intervals (in frames) between consecutive renewal points; that is, $X_m \doteq M_{m+1} - M_m$. Let W_m denote the number of cells lost over X_m . Clearly, $\{W_m\}_{m \geq 0}$ is a regenerative process with respect to the renewal process $\{M_m\}_{m \geq 0}$. Let $\bar{X} = E\{X_m\}$ and $\bar{W} = \{W_m\}$, where $E\{\cdot\}$ denotes the expectation

operator. The following theorem provides for the calculation of the loss probability.

THEOREM 2

For $\bar{X} < \infty$, the cell loss probability L is given by

$$L = \frac{\bar{W}}{\lambda T \bar{X}} \quad \text{with probability 1.} \quad (3)$$

Proof

This theorem is a direct application of the regeneration theorem [10]. It may be easily proved by invoking the strong law of large numbers as follows. Let $\{A_m\}_{m \geq 0}$ denote the total number of cells generated by the source over the m th frame. Since each of the $\{X_m\}_{m \geq 0}$, $\{W_m\}_{m \geq 0}$ and $\{A_m\}_{m \geq 0}$ are independent and identically distributed (i.i.d.) random variables and since $\bar{X} < \infty$, it is easily concluded that $\bar{W} < \infty$ and $\bar{A} = E\{A_m\} = \lambda T \bar{X} < \infty$, as well. The cell loss probability is then given by

$$L = \lim_{n \rightarrow \infty} \frac{\sum_{m=0}^n W_m}{\sum_{m=0}^n A_m} = \lim_{n \rightarrow \infty} \frac{\frac{1}{n+1} \sum_{m=0}^n W_m}{\frac{1}{n+1} \sum_{m=0}^n A_m} \stackrel{(*)}{=} \frac{E\{W_m\}}{E\{A_m\}} = \frac{\bar{W}}{\lambda T \bar{X}}$$

with probability 1; equality (*) is justified by the strong law of large numbers [12]. \square

Since the mean renewal cycle \bar{X} for a system with finite CQ capacity is upper bounded by the corresponding quantity for the infinite CQ capacity (and otherwise identical) system and the latter is bounded for $\lambda T < 1$, the following corollary is easily established.

COROLLARY 2

A sufficient condition for the validity of (3) for any CQ capacity Q is given by $\lambda T < 1$.

In the sequel, the quantities \bar{X} and \bar{W} are calculated. Let (i, n) , $(i, n) \in S \times S^B$ be the state of the system at the beginning of the current frame. Let $X(i, n)$ be a random variable denoting the length (in frames) between the beginning of the current frame and the next renewal instant from $\{M_m\}_{m \geq 0}$. Let $W(i, n)$ be a random variable denoting the number of cells lost over $X(i, n)$. From the definition of

$\{M_m\}_{m \geq 0}$, $\{X_m\}_{m \geq 0}$ and $\{W_m\}_{m \geq 0}$ turns out that

$$X(0, -P) = X_\infty, \quad W(0, -P) = W_\infty \quad (4)$$

and

$$\bar{X}(0, -P) = \bar{X}, \quad \bar{W}(0, -P) = \bar{W}, \quad (5)$$

where X_∞ (W_∞) denotes the generic random variable in the i.i.d. sequence $\{X_m\}_{m \geq 0}$ ($\{W_m\}_{m \geq 0}$), $\bar{X}(0, -P) = E\{X(0, -P)\}$ and $\bar{W}(0, -P) = E\{W(0, -P)\}$.

The rest of this section is focused on the calculation of $\bar{X}(0, -P)$ and $\bar{W}(0, -P)$. From the service policy of the two-server system and theorem 1, the following equations may be derived with respect to $X(i, n)$, $(i, n) \in S \times S^B$ (see appendix B).

For $i \in S$:

$$x(i, -P) = \begin{cases} 1 & \text{if } (i \xrightarrow{T} 0, a^T = 0 \text{ or } 1); \\ 1 + X(1, -P) & \text{if } (i \xrightarrow{T} 1, a^T = 0 \text{ or } 1); \\ 1 + X(j, -P - 1 + k), & \text{if } (i \xrightarrow{T} j, a^T = k), 2 \leq k \leq P + Q, j \in S; \\ 1 + X(j, Q) & \text{if } (i \xrightarrow{T-1} m, a^{T-1} = k'), (m \xrightarrow{1} j, a^1 = 1), \\ & P + Q \leq k' \leq M, m, j \in S; \\ 1 + X(j, Q - 1) & \text{if } (i \xrightarrow{T-1} m, a^{T-1} = k'), (m \xrightarrow{1} j, a^1 = 0), \\ & P + Q + 1 \leq k' \leq M, m, j \in S. \end{cases} \quad (6)$$

For $i \in S$:

$$x(i, -P + 1) =$$

$$\begin{cases} 1 & \text{if } (i \xrightarrow{T} 0, a^T = 0); \\ 1 + X(1, -P) & \text{if } (i \xrightarrow{T} 1, a^T = 0); \\ 1 + X(j, -P + k), & \text{if } (i \xrightarrow{T} j, a^T = k), 1 \leq k \leq P + Q - 1, j \in S; \\ 1 + X(j, Q) & \text{if } (i \xrightarrow{T-1} m, a^{T-1} = k'), (m \xrightarrow{1} j, a^1 = 1), \\ & P + Q - 1 \leq k' \leq M, m, j \in S. \\ 1 + X(j, Q - 1) & \text{if } (i \xrightarrow{T-1} m, a^{T-1} = k'), (m \xrightarrow{1} j, a^1 = 0), \\ & P + Q \leq k' \leq M, m, j \in S. \end{cases} \quad (7)$$

For $i \in S$ and $-P + 2 \leq n \leq Q$:

$$X(i, n) = \begin{cases} 1 + X(j, n + k - 1) & \text{if } (i \xrightarrow{T} j, a^T = k), 0 \leq k \leq Q - n, j \in S; \\ 1 + X(j, Q) & \text{if } (i \xrightarrow{T-1} m, a^{T-1} = k'), (m \xrightarrow{1} j, a^1 = 1), \\ & Q - n \leq k' \leq M, m, j \in S; \\ 1 + X(j, Q - 1) & \text{if } (i \xrightarrow{T-1} m, a^{T-1} = k'), (m \xrightarrow{1} j, a^1 = 0), \\ & Q - n + 1 \leq k' \leq M, m, j \in S, \end{cases} \quad (8)$$

where $i \xrightarrow{k} j$ denotes a transition of the Markov chain of the source from state i to state j in k steps; a^k denotes the number of cells generated in k consecutive slots; M denotes the maximum number of cell arrivals over $T - 1$. By applying the expectation operator to the above equations, the following system of linear equations is obtained:

$$\bar{X}(i_1, n_1) = b(i_1, n_1) + \sum_{i_2 \in S} \sum_{n_2 \in S^B} a(i_1, n_1, i_2, n_2) \bar{X}(i_2, n_2), \quad (9)$$

where $\bar{X}(i, n)$ denotes the expected value of $X(i, n)$, $i \in S$, $n \in S^B$. The coefficients $a(i_1, n_1, i_2, n_2)$ and the constants $b(i_1, n_1)$ for $i_1, i_2 \in S, n_1, n_2 \in S^B$, may be found in appendix B. The desired quantity \bar{X} is then computed as the value of the unknown $\bar{X}(0, -P)$ in (9).

Equations similar to those in (6)–(8) can be derived with respect to $W(i_1, n_1)$ for $i_1 \in S, n_1 \in S^B$; $W(i_1, n_1)$ denotes the number of cells lost over $X(i_1, n_1)$. These equations and the coefficients of the resulting system of linear equations with respect to $\bar{W}(i_1, n_1) = E\{W(i_1, n_1)\}$ for $i_1 \in S, n_1 \in S^B$ – which is of the form of that in (9) – are presented in appendix C. The desired quantity \bar{W} is then computed as the value of the unknown $\bar{W}(0, -P)$. Finally, the cell loss probability can be calculated from (3).

5. Numerical results and discussions

The analysis presented in the previous section is applied for the calculation of the cell loss probability induced at the user premises by the d-LB CAM described in section 3. The cell source is assumed to be Markov with burstiness coefficient γ equal to 0.0 (Bernoulli source), 0.4 and 0.8. The probability p that a cell is generated from the active state of the source is equal to 1. The resulting cell generation rate is equal to $\lambda = 0.08$. The period T of server 1 is chosen to be equal to 10, resulting in a maximum cell rate delivered to the network equal to 0.1. The cell loss probability L , for TP capacity P equal to 10, is shown in figure 4 as a function of the CQ capacity Q and for various values of the burstiness coefficient. As expected, L increases with γ and decreases with Q . Notice that the number of linear equations needed to be

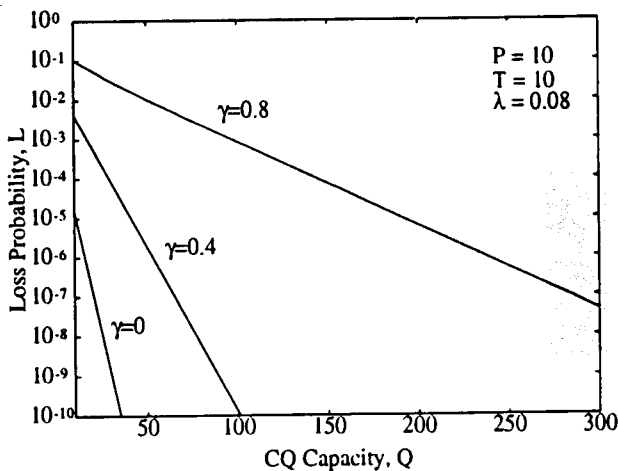


Fig. 4. The cell loss probability L as a function of the CQ capacity Q .

solved is less than $2(Q + P + 1) = 622$, for the CQ capacities shown in figure 4. Similar results are shown in figure 5 as a function of the TP capacity P and for various values of the CQ capacity Q .

A comparison between the cell loss probabilities when the CQ capacity is finite (derived in this paper) and the tail of the distribution of the queue occupancy process of the corresponding infinite CQ capacity system (derived in [6]) may be carried out by using figure 6. Due to analytical complexity, the latter quantity is sometimes used as an approximation of the former. The parameters of the system associated with figure 6 are identical to the system considered for the derivation

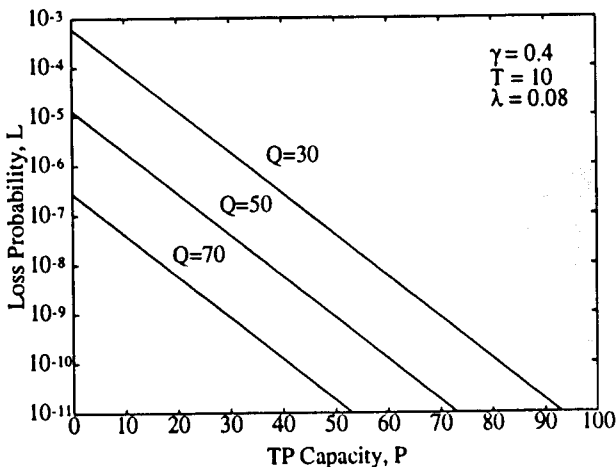


Fig. 5. The cell probability L as a function of the TP capacity P .

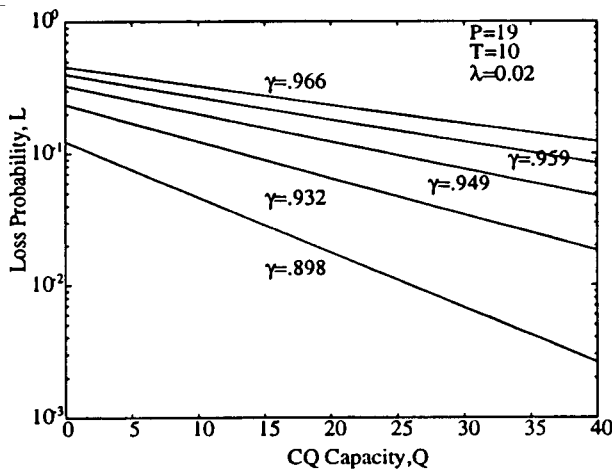


Fig. 6. The cell loss probability L as a function of the CQ capacity Q .

of the tail of the queue occupancy distribution shown in figure 2 of [6]. The burstiness coefficient γ is proportional to the average length of an on (state 1) period of the source; the values of γ shown in figure 6 correspond to average on periods of length 10, 15, 20, 25 and 30, which are considered in figure 2 in [6]. It is easy to establish that the results in [6] for the tail of the distribution underestimate the cell loss probabilities shown in figure 6, by about one order of magnitude. This behavior seems to be consistent with the expectation shaped by the study in [13] according to which the tail of the occupancy distribution seems to underestimate (overestimate) the loss probabilities in the corresponding finite queue system for light (heavy) traffic load.

Although a two-state Markov Modulated Bernoulli cell arrival process to the two-server queueing system has been assumed in this paper, the presented study is applicable under more general cell arrival processes, such as an $MMB(p_1, p_2, \dots, p_n)$ process; $MMB(p_1, p_2, \dots, p_n)$ denotes a Markov Modulated Bernoulli process based on an underlying n -state Markov chain which generates a cell with probability p_i when in state i . Under such a cell arrival process, the number of linear equations to be solved will increase by a factor of $n/2$ compared to the number solved under the $MMB(0, p)$ model considered in the paper.

As the network speed increases, the assumption that the source may deliver at most one cell may be trivially satisfied, since sources will become increasingly unable to deliver at the high network speed, due to processing time limitation. The general methodology is also applicable to the case of multiple cell arrivals per slot. In this case, the state space reduction approach through the consideration of the balance process seems not to be possible. As a result a significantly larger number of linear equations will have to be set up and solved ($2(P+1)(Q+1)$ versus $2(P+Q+1)$).

In addition to the relatively low numerical complexity associated with the

solution of linear equations, their dimensionality is independent of the period T of server 1. This is a very important characteristic since, as the network speed increases, the traffic load offered by a single source will decrease and T will increase. The methodologies for the derivation of the tail of the queue occupancy distribution (which could approximate the cell loss probability, as mentioned above) based on matrix-geometric techniques and z-transform inversion/boundary probabilities calculation [6] present increasing numerical complexity as T increases and become intractable for sufficiently large value of T . The same holds true for a study of the two-server queueing system based on a discrete-time Markov chain describing the system state at the slot boundaries [4].

Finally it should be noted that an alternative approach for the derivation of the cell loss probabilities, based on the system description in terms of the Markov chains $\{(I_m, B_m)\}_{m \geq 0}$ (section 4) imbedded at the frame boundaries, would present similar numerical complexity. In this case, the steady state probabilities ($\bar{\Pi}$) of this Markov chain and the conditional cell loss probabilities given a certain state, would need to be calculated to be used in the calculation of the cell loss probabilities. A potentially advantageous characteristic of the presented approach is that the calculation of $\bar{\Pi}$ is not necessary. When the state space of $\{(I_m, B_m)\}_{m \geq 0}$ is very large, potentially tight performance bounds may be derived by solving truncated versions of (9) and employing various bounding techniques [16,17]. It should also be noted that, by defining processes counting other events of renewal cycles, the probability of occurrence of these events may be readily calculated by simply modifying the constants in the system of linear equations (9). This approach has been used for the calculation of probabilities of joint events [14,15] and delays [16,17]. Other applications of the presented approach resulting in a system description in terms of linear equations as in (9) include the study of a class of protocols for multi-user communication protocols (in terms of both throughput and delay) [16], the approximate characterization of network traffic [14,15] and the delay analysis of infinite capacity queueing systems with priorities [17].

Finally, it should be noted that the dimensionality of the system in (9) may increase beyond the available computing resources due to the large value of the Q , P , the state space of the MMB cell arrival process or the allowance for multiple cell arrivals per slot. In this case, the increased computational complexity may be addressed by solving truncated versions of (9), as mentioned above, or by utilizing the structure of the transition matrix associated with (9) and employing computationally efficient algorithms [18].

Appendix A

Proof of theorem 1

At first it is easy to establish that (2) is satisfied at the beginning of a frame when the CQ is empty. When $Q_j = 0$, (1) implies that $B_j = -P_j$ and thus (2) is

satisfied. The empty CQ state may be considered to be the initial state of the system. Even if the actual initial CQ state is different, the empty CQ state may be assumed to be the "initial" state, since it will be reached in a finite horizon, assuming that λT is less than one; λ denotes the mean cell arrival rate per unit time (slot).

In the sequel, it is shown that if (2) is satisfied at the beginning of a frame (beginning of time slot j), then it will also be satisfied at the beginning of the frame which follows (beginning of time slot $j + T$). There are three possible cases which are considered below. Let $k \geq 0$ denote the number of cell arrivals over the frame starting with the j th slot.

(i) If $B_j = 0$ then (by assumption) $P_j = 0$ and $Q_j = 0$. Then,

$$B_{j+T} = \begin{cases} 0 & \text{if } k = 1, \text{ with } P_{j+T} = 0 \text{ and } Q_{j+T} = 0, \\ -1 < 0 & \text{if } k = 0, \text{ with } P_{j+T} = 1 \text{ and } Q_{j+T} = 0, \\ k - 1 > 0 & \text{if } k \geq 2, \text{ with } P_{j+T} = 0 \text{ and } Q_{j+T} = k - 1 \end{cases}$$

Recall that according to the two-server queueing model one token is generated at the beginning of the last slot of a frame. This token will be used over the last slot of the frame if CQ is non-empty at the beginning of the last slot (cell arrivals occur at the beginning of a slot), or else it will be transferred to the TP by the end of this slot.

(ii) If $B_j = n > 0$ then (by assumption) $P_j = 0$ and $Q_j = n$. Then,

$$B_{j+T} = \min \{n - 1 + k, Q\} \geq 0,$$

$$\text{with } P_{j+T} = 0 \text{ and } Q_{j+T} = \min \{n - 1 + k, Q\}.$$

Note that one cell will be served by using the single token arriving at the last slot of the frame and the rest will be moved to the CQ.

(iii) If $B_j = -n < 0$ then (by assumption) $P_j = n$ and $Q_j = 0$. Then, in view of lemma 1, $\min \{n + 1, k\}$ cells will be removed from the CQ leaving $|k - n - 1|^+$ cells in the CQ and $|k - n - 1|^-$ tokens in the TP. Thus (the min and max below impose the constraint that [CQ] and [TP] cannot exceed Q and P , respectively),

$$B_{j+T} = \min \{Q, \max \{-P, -n + k - 1\}\},$$

with

$$\begin{cases} B_{j+T} \leq 0, & P_{j+T} = |\max \{-P, -n + k - 1\}| \text{ and } Q_{j+T} = 0 & \text{if } k \leq n + 1 \\ B_{j+T} \geq 0, & P_{j+T} = 0 \text{ and } Q_{j+T} = \max \{Q, -n + k - 1\} & \text{if } k \geq n + 1. \end{cases}$$

From (i)–(iii) it is concluded that if (2) is satisfied at the beginning of a frame, then it will be satisfied at the beginning of the frame which follows as well.

It has been shown that (2) is satisfied at the beginning of a frame (call it frame 1) when the CQ is empty, which may be considered to be the initial state of a system. Then, (2) is satisfied at the beginning of frame 2 that follows, as shown by (i) and (iii). Finally, according to (i), (ii) and (iii), (2) will be satisfied for all frames which will follow frame 2 and the proof of the theorem is completed. \square

Appendix B

Some comments of the derivation of equation (6) are offered in this appendix; equations (7) and (8) can be explained similarly. It is assumed that a cell arriving to a full CQ in the last slot of a frame is not dropped, but occupies the position left by the cell served over this slot (by server 1). Let time instant m mark the beginning of the current frame. Note that B_m determines both P_m and Q_m (theorem 1). Let k be the number of cells arriving over the current frame; let j denote the state of Markov chain of the source at $m + T$ (beginning of the next frame).

Starting from state $(i, -P)$ – the TP is full, from theorem 1 – the system moves to a renewal point at $m + T$ if $j = 0$ and $k \leq 1$; in this case, $X(i, -P) = 1$.

If $j \neq 0$, then $m + T$ cannot be a renewal point. If $k \leq 1$, then the TP remains full and the system state at $m + T$ will be $(j, -P)$. The additional time (beyond the current frame) which will be required until the next renewal is reached, is given by $X(j, -P)$.

If $k \geq 2$, then $m + T$ cannot be a renewal point since $B_{m+T} > -P$ (at least one of the tokens needed will come from the TP and there will be no replacement within the frame). As above, the time until the first renewal will be equal to $X(j, r)$ in addition to the current frame length where $B_{m+T} = r$. From lemma 1, theorem 1 and their proofs, it can be established that $r = k - P - 1$ if $2 \leq k \leq P + Q$ (third equation in (6)). If $k \geq P + Q + 1$ then B_{m+T} will reach state Q or $Q - 1$ depending on whether an arrival occurs in the last slot or not (last two equations in (6)).

The coefficients and the constants of the system of linear equations (9) are derived below. Let $f^m(i, j, k)$ be the probability of the event $(i \xrightarrow{m} j, a^m = k)$, as defined in (4)–(6), for $m = 1, T - 1$ and T . Clearly,

$$f^1(i, j, k) = Pr\{i \xrightarrow{1} j, a^1 = k\} = p(i, j)g(i, k),$$

where

$$g(i, k) = Pr\{k \text{ cells are generated from state } i\}.$$

For $m > 1$,

$$\begin{aligned} f^m(i, j, k) &= \Pr\{(i \xrightarrow{m} j, a^m = k)\} \\ &= \sum_{i_1 \in S} p(i, i_1) \sum_{i_2 \in S} p(i_1, i_2) \cdots \sum_{i_{m-1} \in S} p(i_{m-1}, j) [g(i_1, \cdot) \\ &\quad \otimes \cdots \otimes g(i_{m-1}, \cdot) \otimes g(j, \cdot)](k), \end{aligned}$$

where $[g(i_1, \cdot) \otimes \cdots \otimes g(i_{m-1}, \cdot) \otimes g(j, \cdot)](k)$ denotes an m -fold convolution of $g(\cdot, \cdot)$ evaluated at k . Since $g(0, 0) = 1$, it is easy to establish that

$$[g(i_1, \cdot) \otimes \cdots \otimes g(i_{m-1}, \cdot) \otimes g(j, \cdot)](k) = \text{conv}(i_1 + \cdots + i_{m-1} + j, k),$$

where $\text{conv}(n, k)$ is the n -fold convolution of $g(\cdot, \cdot)$ evaluated at k , with $\text{conv}(0, 0) = 1$.

By applying the expectation operator to (6)–(8), the following equations are obtained with respect to $\bar{X}(i, n) = E\{X(i, n)\}$, $(i, n) \in S \times S^B$.

For $i \in S$:

$$\begin{aligned} \bar{X}(i, -P) &= 1 + \sum_{k=0}^1 f^T(i, 1, k) \bar{X}(1, -P) + \sum_{j=0}^1 \sum_{k=2}^{P+Q} f^T(i, j, k) \bar{X}(j, k - P - 1) \\ &\quad + \sum_{j=0}^1 \sum_{k=P+Q}^M \sum_{s=0}^1 f^{T-1}(i, s, k) f^1(s, j, 1) \bar{X}(j, Q) \\ &\quad + \sum_{j=0}^1 \sum_{k=P+Q+1}^M \sum_{s=0}^1 f^{T-1}(i, s, k) f^1(s, j, 0) \bar{X}(j, Q - 1). \end{aligned} \quad (\text{B.1})$$

For $i \in S$:

$$\begin{aligned} \bar{X}(i, -P + 1) &= 1 + f^T(i, 1, 0) \bar{X}(1, -P) + \sum_{j=0}^1 \sum_{k=1}^{P+Q-1} f^T(i, j, k) \bar{X}(j, k - P) \\ &\quad + \sum_{j=0}^1 \sum_{k=P+Q-1}^M \sum_{s=0}^1 f^{T-1}(i, s, k) f^1(s, j, 1) \bar{X}(j, Q) \\ &\quad + \sum_{j=0}^1 \sum_{k=P+Q}^M \sum_{s=0}^1 f^{T-1}(i, s, k) f^1(s, j, 0) \bar{X}(j, Q - 1). \end{aligned} \quad (\text{B.2})$$

For $i \in S$, $-P+2 \leq n \leq Q$:

$$\begin{aligned} \bar{X}(i, n) &= 1 + \sum_{j=0}^1 \sum_{k=0}^{Q-n} f^T(i, j, k) \bar{X}(j, n+k-1) \\ &+ \sum_{j=0}^1 \sum_{k=Q-n}^M \sum_{s=0}^1 f^{T-1}(i, s, k) f^1(s, j, 1) \bar{X}(j, Q) \\ &+ \sum_{j=0}^1 \sum_{k=Q-n+1}^M \sum_{s=0}^1 f^{T-1}(i, s, k) f^1(s, j, 0) \bar{X}(j, Q-1). \end{aligned} \quad (\text{B.3})$$

Equations (B.1)–(B.3) are easily seen to result in the $|S||S^B|$ -dimensional system of linear equations in (9). The coefficients of the unknown and constants are easily established from (B.1)–(B.3) and they are given by

For $i_1 \in S$:

$$b(i_1, n_1) = 1, \quad -P \leq n_1 \leq Q,$$

$$a(i_1 - P, 1, -P) = \sum_{k=0}^1 f^T(i_1, 1, k),$$

$$a(i_1, -P, i_2, n_2) = f^T(i_1, i_2, n_2 + P + 1), \quad -P + 1 \leq n_2 \leq Q - 2, i_2 \in S,$$

$$a(i_1, -P, i_2, Q) = \sum_{k=P+Q}^M \sum_{s=0}^1 f^{T-1}(i_1, s, k) f^1(s, i_2, 1), \quad i_2 \in S,$$

$$a(i_1, -P, i_2, Q-1) = \sum_{k=P+Q+1}^M \sum_{s=0}^1 f^{T-1}(i_1, s, k) f^1(s, i_2, 0) + f^T(i_1, i_2, Q+P), \quad i_2 \in S.$$

$$a(i_1 - P + 1, 1, -P) = f^T(i_1, 1, 0),$$

$$a(i_1, -P + 1, i_2, n_2) = f^T(i_1, i_2, n_2 + P), \quad -P + 1 \leq n_2 \leq Q - 2, i_2 \in S,$$

$$a(i_1, -P + 1, i_2, Q) = \sum_{k=P+Q-1}^M \sum_{s=0}^1 f^{T-1}(i_1, s, k) f^1(s, i_2, 1), \quad i_2 \in S,$$

$$a(i_1, -P + 1, i_2, Q-1) = \sum_{k=P+Q}^M \sum_{s=0}^1 f^{T-1}(i_1, s, k) f^1(s, i_2, 0) + f^T(i_1, i_2, P+Q-1), \quad i_2 \in S.$$

For $i_1 \in S, -P+2 \leq n_1 \leq Q$:

$$a(i_1, n_1, i_2, n_2) = f^T(i_1, i_2, n_2 - n_1 + 1), \quad n_1 - 1 \leq n_2 \leq Q - 2, i_2 \in S,$$

$$a(i_1, n_1, i_2, Q) = \sum_{k+Q-n_1}^M \sum_{s=0}^1 f^{T-1}(i_1, s, k) f^1(s, i_2, 1), \quad i_2 \in S,$$

$$a(i_1, n_1, i_2, Q-1) = \sum_{k+Q-n_1+1}^M \sum_{s=0}^1 f^{T-1}(i_1, s, k) f^1(s, i_2, 0) + f^T(i_1, i_2, Q - n_1),$$

$i_2 \in S.$

Appendix C

Similarly to the derivation of equations (6)–(8) the following equations may be derived with respect to $W(i_1, n_1), (i_1, n_1) \in S \times S^B$.

For $i \in S$:

$$W(i, -P) =$$

$$\left\{ \begin{array}{ll} 0 & \text{if } (i \xrightarrow{T} 0, a^T = 0 \text{ or } 1); \\ W(1, -P) & \text{if } (i \xrightarrow{T} 1, a^T = 0 \text{ or } 1); \\ W(j, -P - 1 + k), & \text{if } (i \xrightarrow{T} j, a^T = k), 2 \leq k \leq P + Q, j \in S \\ k - (P + Q) + W(j, Q) & \text{if } (i \xrightarrow{T-1} m, a^{T-1} = k), (m \xrightarrow{1} j, a = 1), \\ & P + Q \leq k \leq M, \quad m, j \in S; \\ k - (P + Q) + W(j, Q - 1) & \text{if } (i \xrightarrow{T-1} m, a^{T-1} = k), (m \xrightarrow{1} j, a^1 = 0), \\ & P + Q + 1 \leq k \leq M, \quad m, j \in S. \end{array} \right. \quad (\text{C.1})$$

For $i \in S$:

$$W(i, -P + 1) =$$

$$\left\{ \begin{array}{ll} 0 & \text{if } (i \xrightarrow{T} 0, a^T = 0); \\ W(1, -P) & \text{if } (i \xrightarrow{T} 1, a^T = 0); \\ 1 + W(j, -P + k) & \text{if } (i \xrightarrow{T} j, a^T = k), 1 \leq k \leq P + Q - 1, j \in S; \\ k - (P + Q - 1) + W(j, Q) & \text{if } (i \xrightarrow{T-1} m, a^{T-1} = k'), m \xrightarrow{1} j, a^1 = 1), \\ & P + Q - 1 \leq k' \leq M, \quad m, j \in S; \\ k - (P + Q - 1) + W(j, Q - 1) & \text{if } (i \xrightarrow{T-1} m, a^{T-1} = k'), (m \xrightarrow{1} j, a^1 = 0), \\ & P + Q \leq k' \leq M, \quad m, j \in S. \end{array} \right. \quad (\text{C.2})$$

For $i \in S$ and $-P + 2 \leq n \leq Q$:

$$W(i, n) = \begin{cases} W(j, n+k-1) & \text{if } (i \xrightarrow{T} j, a^T = k), 0 \leq k \leq Q-n, j \in S; \\ k' - (Q-n) + W(j, Q) & \text{if } (i \xrightarrow{T^{-1}} m, a^{T^{-1}} = k'), (m \xrightarrow{1} j, a^1 = 1), \\ & Q-n \leq k' \leq M, m, j \in S; \\ k' - (Q-n) + W(j, Q-1) & \text{if } (i \xrightarrow{T^{-1}} m, a^{T^{-1}} = k'), (m \xrightarrow{1} j, a^1 = 0), \\ & Q-n+1 \leq k' \leq M, m, j \in S. \end{cases} \quad (C.3)$$

Equation (C.3) may be explained in the following way. The total cell losses over $X(i, n)$, denoted by $W(i, n)$, is equal to those over the first frame of session $X(i, n)$, plus those over the session $X(\cdot, \cdot)$ initiated in the frame that follows. The latter cell losses are given by $W(\cdot, \cdot)$. The former cell losses are zero under the first condition and they are equal to $k - (Q - n)$ under any of the last two conditions. Equations (C.1) and (C.2) may be explained similarly.

By applying the expectation operator to (C.1)–(C.3) an $|S||S^B|$ -dimensional system of linear equations is obtained with respect to $\bar{W}(i, n)$, $(i, n) \in S \times S^B$; $\bar{W}(i, n)$ denotes the expected value of $W(i_1, n_1)$. It is easy to establish that the coefficients of the unknown, $a_w(i_1, n_1, i_2, n_2)$, are identical to $a(i_1, n_1, i_2, n_2)$ derived in appendix A, $i_1, i_2 \in S, n_1, n_2 \in S^B$. The constants $b_w(i_1, n_1)$, $(i_1, n_1) \in S \times S^B$, are given by

$$b_w(i_1, n_1) = \sum_{k=Q-n_1+1}^M \sum_{s=0}^1 [k - (Q - n_1)] f^{T^{-1}}(i_1, s, k), (i_1, n_1) \in S \times S^B.$$

References

- [1] J. Bae and T. Suda, Survey of traffic control schemes and protocols in ATM networks, Proc. IEEE 79 (Feb. 1991).
- [2] Congestion control in high speed networks, Special Issue, IEEE Commun. Mag. (Oct 1991).
- [3] B-ISDN: High performance transport, Special Issue, IEEE Commun. Mag. (Sept. 1991).
- [4] M. Sidi, W. Liu, I. Cidon and I. Gopal, Congestion control through input traffic rate regulation, IEEE Globecom'89 Conf. (1989).
- [5] H. Ahmadi, R. Guerin and K. Sohrawy, Analysis of a rate-based access control mechanism for high-speed networks, IEEE Globecom'90 Conf. (1990).
- [6] K. Sohrawy and M. Sidi, On the performance of bursty and correlated sources subject to Leaky Bucket rate-based access control scheme, IEEE Infocom'91 Conf. (1991).
- [7] E. Rathgeb, Modeling and performance comparisons at policing mechanisms for ATM networks, IEEE J. Sel. Areas Commun. SAC-9 (Apr. 1991).
- [8] M. Butto, E. Cavallero and A. Tonietti, Effectiveness of the "Leaky Bucket" policing mechanism in ATM networks, IEEE J. Sel. Areas Commun. SAC-9 (Apr. 1991).

- [9] W. Leland, Window based congestion management in broadband ATM networks: the performance of three access-control policies, *IEEE Globecom'89 Conf.* (1989).
- [10] J. Cohen, *On Regenerative Processes in Queueing Theory* (Springer, 1976).
- [11] K. Bala, I. Cidon and K. Sohraby, Congestion control for high speed packet switched networks, *Infocom'90 Conf.* (1990).
- [12] K. Chung, *A Course in Probability Theory* (Academic Press, 1974).
- [13] C. Bisdikian, J. Lew and A. Tantawi, On the approximation of the blocking probability of single server queues with finite buffer capacity, IBM Research Report, RC 17556, T.J. Watson Research Center (Jan. 1992).
- [14] I. Stavrakakis and D. Kazakos, On the approximation of the output process of multi-user random access communication networks, *IEEE Trans. Commun.* COM-38 (Feb. 1990).
- [15] I. Stavrakakis and D. Kazakos, Performance analysis of a star topology of interconnected networks under 2nd-order Markov network output processes, *IEEE Trans. Commun.* COM-38 (Oct. 1990).
- [16] I. Stavrakakis and D. Kazakos, A limited sensing protocol for multi-user packet radio systems, *IEEE Trans. Commun.* COM-37 (April, 1989).
- [17] I. Stavrakakis, A considerate priority queueing system with guaranteed policy fairness, *Infocom'92 Conf.* (1992).
- [18] J.Y. LeBoudec, An efficient solution method for Markov models of ATM links with loss priorities, *IEEE J. Sel. Areas in Commun.* SAC-9 (April 1991).