Packet dropping characteristics in a queue with autocorrelated arrivals

Andrzej Chydzinski, Robert Wojcicki, Grzegorz Hryn

Abstract—This paper provides a detailed description of the packet dropping process connected with the buffer overflows in a network node. Namely, we show the formulas for the most important loss characteristics, both in the transient and the stationary regime and then illustrate them via numerical examples. In order to make it possible to obtain the dropping characteristics for strongly autocorrelated arrivals, the Markov-modulated Poisson process is used as a traffic model.

Index Terms—teletraffic modeling, finite-buffer queue, MMPP, packet losses

I. INTRODUCTION

Packet switched networks play today a prominent role because they are resilient to breaks in network links, make good use of bandwidth and are easier to deploy and maintain than other networks. Packet dropping is common in packet switched networks. It is caused by the limited buffering space in network devices and can seriously influence the performance of the network. Describing the packet loss process is an important task as it enables better network design in terms of buffer sizing and management, congestion control mechanisms, protocols etc.

Calculations of packet loss characteristics can be carried out using simulation or analysis. Both these approaches have their advantages and disadvantages. For simulation we may build a more accurate model but obtaining simulation results is sometimes difficult (due to rare events) or time-consuming.

In this paper we present an analysis of the packet loss process in a finite-buffer queue whose arrival process is given by the Markov-modulated Poisson process (MMPP). The MMPP was chosen due to its ability to mimic the complex statistical behaviour of recorded traffic traces. As was shown in [2], using the MMPP we can match not only the basic parameters of the traffic (mean rate, variance, higher moments) but also the shape of the marginal distribution and the autocorrelation function.

The main achievement presented herein is a closed-form formula for the transform of the average number of losses in \((0, t]\) interval. To the best of the authors’ knowledge, this result is new. Using it we can easily obtain both the transient and stationary number of packet losses as well as the loss ratio, a commonly used QoS parameter.

There is a vast amount of literature devoted to the applications of the MMPP [3]-[7], parameter fitting [2],[8]-[10] and its queueing behaviour [11]-[15] (these are just examples, the complete bibliography is much longer than that), but relatively little has been reported regarding the loss process in a finite-buffer queue fed by the MMPP. Typically, only the stationary loss ratio was studied using exact and approximate techniques [13], [16], [17].

The layout of this paper is as follows. The next section describes the arrival process, the queueing model and introduces the notation used throughout the paper. The major part of the paper (Section 3) then follows, presenting formulas for the number of packet losses in the transient and stationary regimes with proofs and comments. Section 4 then shows a numerical example that uses the MMPP parameterization based on an IP traffic trace file. The paper concludes in Section 5.

II. ARRIVAL PROCESS AND QUEUEING SYSTEM

The Markov-modulated Poisson process is constructed by varying the arrival rate of the Poisson process according to an \(m\)-state continuous-time Markov chain (for instance, [12]). When the Markov chain is in state \(i\), arrivals occur according to a Poisson process of rate \(\lambda_i\). The MMPP is usually parameterized by two \(m \times m\) matrices: \(Q\) and \(\Lambda\). In this parameterization, \(Q\) is the infinitesimal generator of the continuous-time Markov chain and \(\Lambda\) is a diagonal matrix that has arrival rates \((\lambda_1, \ldots, \lambda_m)\) on its diagonal and zeroes elsewhere.

The average rate of the MMPP can be calculated as

\[
\lambda = \pi \cdot (\lambda_1, \ldots, \lambda_m)\tau.
\]

where \(\pi\) is the stationary vector for \(Q\), namely

\[
\pi Q = (0,\ldots,0),
\]

\[
\pi \cdot 1 = 1,
\]

and

\[
1 = (1,\ldots,1)\tau.
\]

Assuming that a packet arrival occurs at \(t = 0\) and denoting by \(T_n, n = 1,2,\ldots\) the packet arrival times, the interarrival times

\[
T_n = t_n - t_{n-1},
\]

have the following properties. The transform of the distribution of \(T_1\) has the form:

\[
E e^{-sT_1} = (sI - Q + \Lambda)^{-1}\Lambda,
\]
while the joint transform of $T_1, \ldots, T_n$:

$$E\left[ \exp\left(-\sum_{k=1}^{n} s_k T_k \right) \right] = \prod_{k=1}^{n} \left( s_k I - Q + \Lambda \right)^{-1} \Lambda,$$

where $I$ denotes an $m \times m$ identity matrix. Moreover, we have

$$ET_k = t! \left( \Lambda - Q \right)^{-1} \Lambda,$$

$$E(T_1 - ET_1) (T_{k+1} - ET_{k+1}) = \left( \Lambda - Q \right)^{-2} \Lambda \left[ I - \left( \Lambda - Q \right)^{-1} \Lambda \right] (\Lambda - Q)^{2} \Lambda.$$

The $k$-th moment of the interarrival time in the stationary regime is equal to

$$m_k(T_i) = k! \ p \ (\Lambda - Q)^{-1} \Lambda \cdot 1, \ \ p = \frac{1}{\lambda \pi \Lambda}.$$ 

In particular, the variance of the interarrival has the following form:

$$Var = m_2(T_i) - \frac{1}{\lambda^2}.$$ 

Finally, the $k$-lag autocovariance in the stationary regime equals to

$$Cov(k) = \rho \ (\Lambda - Q)^{-2} \Lambda \left[ (\Lambda - Q)^{-1} \Lambda \right]^{k-1} - 1 \rho \cdot (\Lambda - Q)^{-2} \Lambda.$$ 

By $J(t)$ we will denote the state of the modulating Markov chain at time $t$ and by $P_{i,j}(n,t)$ the counting function for the MMPP, namely

$$P_{i,j}(n,t) = P(N(t) = n, J(t) = j | N(0) = 0, J(0) = i),$$

where $P(\cdot)$ is the probability and $N(t)$ is the total number of arrivals in $(0, t]$. Then, the generating function

$$P^{*}(z,t) = \sum_{n=0}^{\infty} P(n,t) z^n$$

can be explicitly presented in the form:

$$P^{*}(z,t) = e^{(Q^{-(1-z)} \Lambda s)} \cdot |z| \leq 1. \ \ (1)$$

In what follows, the crucial role will be played by the sequences of $m \times m$ matrices $A_k(s), \overline{D}_k(s)$ defined as:

$$A_k(s) = [a_{k,i,j}(s)]_{i,j}, \ \ \overline{D}_k(s) = [\overline{d}_{k,i,j}(s)]_{i,j}, \ \ (2)$$

where

$$a_{k,i,j}(s) = \int_{0}^{\infty} e^{-st} P_{i,j}(k,t) dF(t),$$

$$\overline{d}_{k,i,j}(s) = \int_{0}^{\infty} e^{-st} P_{i,j}(k,t)(1 - F(t))dt,$$

and $F(t)$ is the service time distribution function. From the practical point of view it is important that $A_k(s)$ and $\overline{D}_k(s)$ can be computed effectively by means of the well-known uniformization method (see, for instance, [11], [18]).

Also, we will be using frequently the following matrix-form characteristics of the MMPP:

$$Z(s) = \frac{\left( \lambda_i - Q_{ii} \right) p_{ij}}{s + \lambda_i - Q_{ii}},$$

$$E(s) = \left[ \frac{\Lambda_{ij}}{s + \lambda_i - Q_{ii}} \right]_{i,j},$$

where

$$p_{ij} = \begin{cases} 0 & \text{if } i = j, \\ Q_{ij}/(\lambda_i - Q_{ii}) & \text{if } i \neq j, \end{cases} \ \ (3)$$

and $Q_{ij}, \Lambda_{ij}$ denote elements of the matrices $Q$ and $\Lambda$, respectively.

Moreover, the following $m \times m$ matrices will be of use:

$$\overline{A}_n(s) = \sum_{k=n}^{\infty} A_k(s),$$

$$B_n(s) = A_{n+1}(s) - \overline{A}_{n+1}(s)(\overline{A}_0(s))^{-1},$$

$$R_0(s) = 0,$$

$$R_1(s) = A_0^{-1}(s),$$

$$R_{k+1}(s) = A_0^{-1}(s)(R_k(s) - \sum_{j=0}^{k} A_{j+1}(s)R_{k-j}(s)), \ k \geq 1,$$

As for the queuing model, we deal herein with a single server queue fed by an MMPP. The service time is distributed according to a distribution function $F(\cdot)$, which is not further specified, and the standard independence assumptions are made. The buffer size (system capacity) is finite and equal to $b$ (including service position). This means that if a packet at its arrival finds the buffer full, it is blocked and lost. We assume also that the time origin corresponds to a departure epoch.

### III. Number of Losses

Let $X(t)$ denote the queue size at time $t$ (including service position, if occupied). Let $L(t)$ be the number of losses in $(0, t]$ and $\Delta_{n,i}(t)$ be its average value provided $X(0) = n$ and $J(0) = i$, namely:

$$\Delta_{n,i}(t) = E(L(t) | X(0) = n, J(0) = i).$$

Moreover, let $\delta_{n,i}(s)$ denote the Laplace transform of $\Delta_{n,i}(t)$

$$\delta_{n,i}(s) = \int_{0}^{\infty} e^{-st} \Delta_{n,i}(t)dt,$$

and $\delta_n(s)$ be the column vector:

$$\delta_n(s) = (\delta_{n,1}(s), \ldots, \delta_{n,m}(s))^T.$$

**Theorem 1.** The Laplace transform of the average number of losses in $[0, t]$ in the MMPP/GI/b queue has the form:

$$\delta_n(s) = \sum_{k=0}^{b-n} R_{b-n-k}(s)v_k(s)$$

$$+ \left( R_{b-n+1}(s)A_0(s) + \sum_{k=0}^{b-n} R_{b-n-k}(s)B_{k}(s) \right) M_0^{-1}(s)y_0(s), \ \ n = 0, \ldots, b, \ \ (4)$$

$$\overline{\Lambda}_n(s) = \sum_{k=n}^{\infty} A_k(s),$$

$$B_n(s) = A_{n+1}(s) - \overline{A}_{n+1}(s)(\overline{A}_0(s))^{-1},$$

$$R_0(s) = 0,$$

$$R_1(s) = A_0^{-1}(s),$$

$$R_{k+1}(s) = A_0^{-1}(s)(R_k(s) - \sum_{j=0}^{k} A_{j+1}(s)R_{k-j}(s)), \ k \geq 1,$$
where

\[ v_k(s) = \bar{A}_{k+1}(s)(\bar{A}_0(s))^{-1}c_b(s) - c_{b-k}(s), \]

\[ c_k(s) = \frac{1}{s} \sum_{i=b-k}^{\infty} (i - b + k)A_i(s) \cdot 1 \]

\[ + \sum_{i=b-k}^{\infty} (i - b + k)\bar{D}_i(s) \cdot 1, \]

\[ y_b(s) = E(s) \sum_{k=0}^{b-1} R_{b-1-k}(s)v_k(s) \]

\[ -(I - Z(s)) \sum_{k=0}^{b} R_{b-k}(s)v_k(s), \]

\[ M_b(s) = (I - Z(s))[R_{b+1}(s)A_0(s) + \sum_{k=0}^{b} R_{b-k}(s)B_k(s)] \]

\[ -E(s)[R_b(s)A_0(s) + \sum_{k=0}^{b-1} R_{b-1-k}(s)B_k(s)]. \]

**Proof** of Theorem 1. Assuming \( 1 \leq X(0) \leq b \) and conditioning on the first departure time we obtain the following system of integral equations for \( n = 1, \ldots, b: \)

\[
\Delta_{n,i}(t) = \sum_{j=1}^{n} \sum_{k=0}^{b-n-1} \int_{0}^{t} \Delta_{n+k-1,j}(t-u)P_{i,j}(k,u)dF(u)
\]

\[
+ \sum_{j=1}^{n} \sum_{k=b-n}^{\infty} \int_{0}^{t} (k - b + n + \Delta_{b-1,j}(t-u))P_{i,j}(k,u)dF(u)
\]

\[
+ (1 - F(t)) \sum_{j=1}^{n} \sum_{k=b-n}^{\infty} (k - b + n)P_{i,j}(k,t), \quad (5)
\]

The first summand in (5) corresponds to the case where the first departure time \( u \) is before \( t \) and the buffer does not get full by the time \( u \). This means that the number of arrivals in \( [0, u] \) must not be greater than \( b - n - 1 \). The second summand corresponds to the case where the first departure time \( u \) is before \( t \) and the buffer gets full by the time \( u \). In this case \( k \geq b - n \) packets arrive in \( [0, u] \) and \( k - b + n \) of them are lost. Finally, the last summand corresponds to the case where the first departure time is after \( t \). Probability of this event is equal to \( 1 - F(t) \) and the average number of lost packets is then equal to \( \sum_{j=1}^{n} \sum_{k=b-n}^{\infty} (k - b + n)Q_{i,j}(k,t) \).

If \( X(0) = 0 \) then conditioning on the first event time in the MMPP (packet arrival or change of the modulating state) we have:

\[ \Delta_{0,i}(t) = \sum_{j=1}^{(n)} \int_{0}^{t} \Delta_{0,j}(t-u)(\lambda_i - Q_{i,j})P_{i,j}(t-u)dF(u) \]

\[ + \sum_{j=1}^{(n)} \int_{0}^{t} \Delta_{1,j}(t-u)\Lambda_{i,j}e^{-(\lambda_i - Q_{i,j})u}du. \quad (6) \]

We may now apply the Laplace transform to both sides of (5) and (6). After that, utilizing matrix notation we arrive at:

\[ \delta_n(s) = \sum_{k=0}^{b-n-1} A_k(s)\delta_{n+k-1}(s) \]

\[ + \sum_{k=b-n}^{\infty} A_k(s)\delta_{b-1}(s) + c_n(s), \quad n = 1, \ldots, b, \]

\[ \delta_0(s) = Z(s)\delta_0(s) + E(s)\delta_1(s). \]

Then, substituting \( \delta_n(s) = \delta_{b-n}(s) \) we have:

\[ \sum_{k=0}^{n} A_k(s)\delta_{n-k}(s) - \delta_n(s) = \psi_n(s), \quad n = 0, \ldots, b - 1, \]

\[ \delta_b(s) = Z(s)\delta_b(s) + E(s)\delta_{b-1}(s), \]

where

\[ \psi_n(s) = A_{n+1}(s)\delta_0(s) - \sum_{k=n+1}^{\infty} A_k(s)\delta_1(s) - c_{b-n}(s). \]

Applying Lemma 3.2.1 [19] with slight change in notation we conclude that the solution of the system (9) has the form:

\[ \tilde{\delta}_n(s) = R_{n+1}(s)C(s) + \sum_{k=0}^{n} R_{n-k}(s)\psi_k(s), \]

where \( C(s) \) is a column vector that does not depend on \( n \).

Now we are reduced to finding unknown \( C(s) \), \( \tilde{\delta}_0(s) \) and \( \tilde{\delta}_1(s) \). Substituting \( n = 0 \) in (11) we obtain

\[ C(s) = A_0(s)\tilde{\delta}_0(s), \]

while substituting \( n = 0 \) in (9) we have

\[ \tilde{\delta}_0(s) = \sum_{k=0}^{\infty} A_k(s)\tilde{\delta}_1(s) + c_b(s). \]

Substituting \( n = b \) and, subsequently, \( n = b - 1 \) into (11) and then applying condition (10) we obtain

\[ \tilde{\delta}_b(s) = M_b^{-1}(s)y_0(s), \]

and this finishes in fact the proof of Theorem 1.

**Corollary 1**: The stationary number of packet losses in time unit (i.e. \( \lim_{t \to \infty} \Delta_{n,i}(t)/t \)) is equal to

\[ \lim_{s \to 0^+} s^2 \delta_{b,1}(s), \]

where \( \delta_{b,1}(s) \) is the first element of vector \( \delta_{b,1}(s) \) given in (4).

As the stationary characteristics do not depend on initial state of the system, we can use any initial queue size and any modulating state instead of \( b \) and \( 1 \) in formula (15). In practice,
\( X(0) = b \) is the best choice as in this case (4) reduces to its simplest form, namely:

\[
\delta_b(s) = M_b^{-1}(s) g_b(s).
\]

Similarly, we can obtain the loss ratio \((LR)\), a very popular QoS parameter. We simply have:

\[
LR = \lim_{s \to 0^+} \frac{s^2 \delta_b(s)}{s^2 \lambda}.
\] (16)

Naturally, using Theorem 1 we can also obtain the transient number of losses. To accomplish that, an algorithm for the Laplace transform inversion has to be applied (see, for example, [20]).

### IV. Example 1

In this example we will use the system parametrization from [13]. The service time is constant and equal to 1 ms, \( b = 50 \), and the MMPP parameters are given by:

\[
Q = \begin{bmatrix}
-8.4733 \cdot 10^{-4} & 8.4733 \cdot 10^{-4} \\
5.0201 \cdot 10^{-6} & -5.0201 \cdot 10^{-6}
\end{bmatrix},
\]

\((\lambda_1, \lambda_2) = (1.0722, 0.48976)\).

This parameterization has the following properties. The stationary vector of the modulating Markov chain is:

\[
\pi = (0.00589, 0.99411),
\]

while the average rate of the traffic:

\[
\lambda = \pi \cdot (\lambda_1, \lambda_2)^T = 0.49319.
\]

This gives the link utilization of 49.319%.

Now we can obtain numerical values. Firstly, using (16) we can compute the stationary loss ratio:

\[
LR = 5.7993 \cdot 10^{-4}
\]

and the mean number of packet losses per 1ms:

\[
\lambda \cdot LR = 2.8602 \cdot 10^{-4}.
\] (17)

Secondly, we can evaluate the impact of the initial queue size and modulating phase on the short-time behaviour of the loss process. In Fig. 1 the function \( \Delta_{n,i}(t)/t \), representing the mean number of losses per time unit, is depicted for three different initial queue sizes and \( i = 1 \). After about 50-100s all the curves converge to the stationary value \((2.8602 \cdot 10^{-4})\), but for high \( X(0) \) a very high number of losses may be observed shortly after the system begins to operate.

In Figs. 2 and 3 the impact of the initial phase, \( i \), on the function \( \Delta_{n,i}(t)/t \), is shown for initially empty and full buffer. As we can see, for the higher initial arrival rate \((\lambda_1)\), the convergence to the steady state may be slow.
Basic characteristics of the original sample and its MMPP model are shown in Table I. It is important that the autocorrelation function properly matches the original sample on several time scales (see Fig. 5 in [21]).

<table>
<thead>
<tr>
<th></th>
<th>mean packet</th>
<th>arrival rate, $\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>original traffic</td>
<td>13.940</td>
<td>71732</td>
</tr>
<tr>
<td>MMPP</td>
<td>13.941</td>
<td>71729</td>
</tr>
</tbody>
</table>

It is assumed that $b = 120$ and the queue is served at a constant rate of 98689.5 pkts/s which gives the link utilization of 72.7%.

Using (16) we obtain the stationary loss ratio:

$$LR = 1.1044 \cdot 10^{-2},$$

and the mean number of packet losses per 1ms:

$$\lambda \cdot LR = 0.79219.$$  

Both of these values are relatively high, taking into account the moderate value of link utilization. This is caused, naturally, by the autocorrelated structure of the arrival process.

As for the transient characterization of the loss process, in Fig. 4 the mean number of losses per time unit is depicted for five different initial queue sizes and $J(0) = 4$. All the curves converge to 0.79219, but, as we can see, this convergence may be slow, especially for high initial queue sizes.

In Figs. 5, 6, 7 the impact of the initial phase, $i$, on the function $\Delta_n(t)/t$, is depicted for initial buffer occupancy of 0, 50%, and 100%, respectively. As we can see, the higher the initial arrival rate, the slower the convergence to the steady state (see curves for $i = 4$ or $i = 2$). What is more interesting is that the function may not only be nonmonotonic, but it may also have more than one extremum (for instance, Fig. 7, $i = 3$). This is probably caused by the structure of the transition matrix $Q$.

Finally, in Fig. 8 the stationary loss ratio as a function of the buffer size for two link utilizations: 72.7% and 99% is shown. As we can see, in the case of autocorrelated arrival process and high link utilization this function may decrease very slowly and even a large buffer does not eliminate losses completely.

VI. CONCLUSIONS

We presented a study on the packet loss process in a finite-buffer queue whose arrival process is a Markov-modulated Poisson process. The MMPP was chosen due to its ability to model the complex statistical behaviour of network traffic.

The main result, which is the Laplace transform of the mean number of losses in $[0, t]$, enables quick calculations of the stationary characteristics of the loss process and, by means of an inversion algorithm, also enables the calculations of the transient measures.
As these losses are common in packet networking and the ability to compute the characteristics of the loss process is helpful in network design, we believe that this study is of practical importance.

The results presented herein are devoted to the average number of losses. It is easily seen that they can be extended to the probability distribution instead of the average value. For this purpose, it is sufficient to use the function

$$\Delta_{n,i}(t, l) = P(L(t) = l | X(0) = n, J(0) = i)$$

instead of $\Delta_{n,i}(t)$, and obtaining an analog of Theorem 1 for the transform of $\Delta_{n,i}(t, l)$ is straightforward.

ACKNOWLEDGMENT

This work was supported in part by MNiSW under grant N N516 381134.

REFERENCES


Fig. 6. The mean number of losses per 1ms as a function of time for initial system occupancy of 50% and five different states of the modulating chain $J(0) = i = 1, \ldots, 5$. Example 2.

Fig. 7. The mean number of losses per 1ms as a function of time for initially full system and five different states of the modulating chain $J(0) = i = 1, \ldots, 5$. Example 2.

Fig. 8. Stationary loss ratio versus the buffer size for two link utilizations: 72.7% and 99%. Example 2.

Andrzej Chydzinski received his MS (in mathematics) and PhD (in computer science) degrees from Silesian University of Technology, Gliwice, Poland, in 1997 and 2002, respectively. He is currently an Assistant Professor in the Institute of Computer Sciences of this university. His academic and professional interests include modeling and simulation of computer networks and queueing systems. Dr. Chydzinski has an established record of publications in these fields, with more than fifty published papers.

Robert Wojcicki is an Assistant Professor in the Institute of Computer Sciences of Silesian University of Technology, Gliwice, Poland. He received his MS (in computer science) in 1995 and PhD (in computer science) in 2005, both from this university. His research area is primarily focused on traffic management for the Transmission Control Protocol/Internet Protocol (TCP/IP) in optical networks. He authored and co-authored a dozen of conference and journal papers.

Grzegorz Hryn received his MSc degree in 1995 and PhD in 2005, both from Silesian University of Technology, Gliwice, Poland. He is an Assistant Professor in the Institute of Computer Sciences of this university. His PhD thesis dealt with web content adaptation to the client’s requirements and connection capacity. His recent interests include Publish/Subscribe architectures with notions of load balancing and load shedding, high availability, fail-awareness and dependability. Dr. Hryn is an author or co-author of fifteen publications in journals and conference proceedings.