Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2012, Article ID 831956, 18 pages
doi:10.1155/2012/831956

Research Article

A Unified Method of Analysis for Queues with Markovian Arrivals

Andrzej Chydzinski

Institute of Informatics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

Correspondence should be addressed to Andrzej Chydzinski, andrzej.chydzinski@polsl.pl

Received 24 June 2011; Revised 21 October 2011; Accepted 21 October 2011

Academic Editor: Angelo Luongo

We deal with finite-buffer queueing systems fed by a Markovian point process. This class includes the queues of type M/G/1/N, M^X/G/1/N, PH/G/1/N, MMPP/G/1/N, MAP/G/1/N, and BMAP/G/1/N and is commonly used in the performance evaluation of network traffic buffering processes. Typically, such queueing systems are studied in the stationary regime using matrix-analytic methods connected with M/G/1-type Markov processes. Herein, another method for finding transient and stationary characteristics of these queues is presented. The approach is based on finding a closed-form formula for the Laplace transform of the time-dependent performance measure of interest. The method can be used for finding all basic characteristics like queue size distribution, workload distribution, loss ratio, time to buffer overflow, and so forth. To demonstrate this, several examples for different combinations of arrival processes and characteristics are presented. In addition, the most complex results are illustrated via numerical calculations based on an IP traffic parameterization.

1. Introduction

Since the beginning of the 1990s, when the strong auto-correlation of the Internet traffic was discovered, a variety of processes have been developed or adapted for proper teletraffic modeling. For instance, fractional Brownian motion [1], chaotic maps [2], FARIMA [3], and multifractal wavelets [4] have been applied in wide range of tasks connected with performance evaluation of buffering processes, traffic predictability, congestion and admission control, buffer sizing, and so forth.

However, none of the aforementioned processes suits as well for the teletraffic modeling as the famous class N of Markovian point processes [5] or one of its well-known reparameterizations or subclasses (MMPP, MAP, BMAP, etc.). First of all, this is connected with the fact that N processes are analytically tractable. They are also easy to simulate, and a variety of parameter fitting procedures have been developed for them [6–12].

The only disadvantage of Markovian models used for teletraffic modeling is that they are not truly self-similar or long-range dependent. However, for practical purposes, it is typically enough to mimic the self-similarity over a few time scales. This can be easily accomplished using Markovian processes and, as shown in [9], the resulting model can be reliable in terms of its marginal distribution, autocovariance function, and queueing behaviour.

One of the main reasons for developing traffic models is finding their queueing performance characteristics. In this paper, we deal with the finite-buffer queue whose arrival process is given by a Markovian point process from the class N. So far such queueing systems have been solved typically in their stationary regime using matrix-analytic methods connected with M/G/1-type Markov chains [13–16] (this set of papers is not intended to be exhaustive, the literature devoted to the subject is vast).

Herein, a different, unified method for solving these queueing systems is described. Its main advantage is that it gives formulas for the characteristics of interest in a closed, easy-to-use form. It is devoted to computing transient characteristics, but the steady-state measures could also be obtained from the transient solutions. It can be used for all processes in the class N (e.g., Poisson processes, batch Poisson processes, phase-type renewal processes, MMPPs, MAPs, BMAPs) and for many queueing performance characteristics, including queue size, virtual waiting time, loss ratio, time to buffer overflow, buffer overflow period. To demonstrate this, three queueing systems with different combinations of arrival processes and characteristics of interest are solved using the proposed method. In every next queueing system, an arrival process of growing complexity is used. Namely, we will start with the Poisson arrival process and the queue size distribution and finish with the MAP arrival process and the workload distribution.

The remaining part of the paper is structured in the following way. In Section 2, a description of the proposed method is presented. In Sections 3, 4, and 5, three detailed examples of its applications are shown. In particular, in Section 3 a formula for the transient queue size distribution in the M/G/1/N system is proven and illustrated via numerical examples. In Section 4, a formula for the time to buffer overflow in the M^X/G/1/N queue is presented. In Section 5, a formula for the workload distribution in the MAP/G/1/N model is shown and illustrated via numerical example based on an IP traffic parameterization. Finally, remarks concluding the paper are gathered in Section 6.

2. Method

The proposed method can be sketched in the following three-step scheme.

(I) In the beginning, we apply the total probability formula with respect to the first departure time. This allows us to utilize the Markovian structure of the arrival process and develop a system of integral equations for the characteristic of interest.

(II) Then, by using the Laplace transform technique, we reduce the problem to a system of linear equations.

(III) In the next step the solution of the resulting system of equations is presented in a closed-form formula using recurrent sequences.

By means of the resulting formula, we can compute the steady-state characteristic at once using basic properties of the Laplace transform or we can compute the transient characteristic applying an inversion algorithm.

The third step in this scheme is based on the following lemma (for proof, see [17, page 201]).

Lemma 2.1. *Assume that $A_0, A_1, A_2, \ldots$ is a sequence of $m \times m$ matrices such that A_0 is nonsingular and $\psi_1, \psi_2, \ldots$ is a sequence of column vectors of size m. Then every solution of the system of equations*

$$\sum_{k=-1}^{n-1} A_{k+1} x_{n-k} - x_n = \psi_n, \quad n \geq 1, \tag{2.1}$$

has the form:

$$x_n = R_n c + \sum_{k=1}^{n} R_{n-k} \psi_k, \quad n \geq 1, \tag{2.2}$$

where c is a column vector that does not depend on n and the sequence R_k is defined to be

$$R_0 = \mathbf{0}, \qquad R_1 = A_0^{-1}, \qquad R_{k+1} = R_1 \left(R_k - \sum_{i=0}^{k} A_{i+1} R_{k-i} \right), \quad k \geq 1, \tag{2.3}$$

and $\mathbf{0}$ denotes the $m \times m$ matrix of zeroes.

It is easy to check that if the system (2.1) is indexed from 0, namely

$$\sum_{k=-1}^{n} A_{k+1} x_{n-k} - x_n = \psi_n, \quad n \geq 0, \tag{2.4}$$

then its every solution has the form

$$x_n = R_{n+1} c + \sum_{k=0}^{n} R_{n-k} \psi_k, \quad n \geq 0. \tag{2.5}$$

Now it is time to show how this method works in practice.

3. Poisson Arrivals and Queue Size Distribution

In the first example, we will find a formula for the transient queue length distribution in the M/G/1/N model, that is, for the system with Poisson arrivals (with intensity λ), general type of the service time distribution (given by distribution function $F(t)$), and finite capacity (the total number of customers in the system, including service position, must not exceed N).

To the best of our knowledge, a closed-form formula for the transient queue size in the M/G/1/N system has not been reported in the English literature yet. A transient solution for the infinite-buffer M/G/1 system can be found in [18, Section 1.7] and [19, Chapter 3].

The transient behavior of queueing systems depends on the initial buffer content. Herein the initial buffer occupancy is not further specified and can be zero or nonzero. It is

assumed that the time origin corresponds to a departure epoch. Thus, if the initial buffer content is non-zero, then the service begins at the time origin. Otherwise, the service begins at the first arrival time. The service time distribution can have any particular form, but for practical reasons we restrict this study to the class of service time distributions with explicit Laplace-Stieltjes transform.

Let $\mathbf{P}(\cdot)$ denote probability, $X(t)$ the queue size process, and

$$\widehat{\phi}_{nl}(t) = \mathbf{P}(X(t) = l \mid X(0) = n). \tag{3.1}$$

(I) We start from using the total probability formula with respect to the first departure time. For $0 < n \leq N$, we obtain

$$\begin{aligned}\widehat{\phi}_{nl}(t) &= \sum_{k=0}^{N-n-1} \int_0^t \widehat{\phi}_{n+k-1,l}(t-u) \frac{e^{-\lambda u}(\lambda u)^k}{k!} dF(u) \\ &\quad + \sum_{k=N-n}^{\infty} \int_0^t \widehat{\phi}_{N-1,l}(t-u) \frac{e^{-\lambda u}(\lambda u)^k}{k!} dF(u) + \rho_{nl}(t), \\ \rho_{nl}(t) &= (1 - F(t)) \cdot \begin{cases} 0 & \text{if } l < n, \\ \dfrac{e^{-\lambda t}(\lambda t)^{l-n}}{(l-n)!} & \text{if } n \leq l < N, \\ \displaystyle\sum_{k=N-n}^{\infty} \frac{e^{-\lambda t}(\lambda t)^k}{k!} & \text{if } l = N. \end{cases}\end{aligned} \tag{3.2}$$

The first sum in (3.2) describes the situation where the first departure time u is before t and there is no buffer overflow by the time u, which means that the number of arrivals in $(0, u]$ must be less than $N-n$. The second sum describes the situation where the first departure time u is before t and an overflow occurs by the time u. Finally, $\rho_{nl}(t)$ describes the case where the first departure time u is after t.

Using the total probability formula with respect to the first arrival time for the initially empty system $(X(0) = 0)$, we have

$$\widehat{\phi}_{0l}(t) = \int_0^t \widehat{\phi}_{1l}(t-u)\lambda e^{-\lambda u} du + \delta_{0l} e^{-\lambda t}, \tag{3.3}$$

where δ_{ij} is the Kronecker symbol, that is, $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

(II) In the second step, we apply the Laplace transform to both sides of (3.2) and (3.3). Therefore, for the transform

$$\phi_{nl}(s) = \int_0^{\infty} e^{-st} \widehat{\phi}_{nl}(t) dt \tag{3.4}$$

we obtain

$$\phi_{nl}(s) = \sum_{k=0}^{N-n-1} a_k(s)\phi_{n+k-1,l}(s) + \sum_{k=N-n}^{\infty} a_k(s)\phi_{N-1,l}(s) + r_{nl}(s), \qquad 0 < n \le N, \tag{3.5}$$

$$\phi_{0l}(s) = \frac{\lambda}{s+\lambda}\phi_{1l}(s) + \frac{\delta_{0l}}{s+\lambda}, \tag{3.6}$$

where

$$a_k(s) = \int_0^{\infty} \frac{e^{-(\lambda+s)t}(\lambda t)^k}{k!} dF(t), \qquad d_k(s) = \int_0^{\infty} \frac{e^{-(\lambda+s)t}(\lambda t)^k}{k!}(1 - F(t))dt,$$

$$r_{kl}(s) = \begin{cases} 0 & \text{if } l < k, \\ d_{l-k}(s) & \text{if } k \le l < N, \\ \dfrac{(1-f(s))}{s} - \displaystyle\sum_{i=0}^{N-n-1} d_i(s) & \text{if } l = N, \end{cases} \tag{3.7}$$

$$f(s) = \int_0^{\infty} e^{-st} dF(t).$$

Substituting $\varphi_n(s) = \phi_{N-n,l}(s)$, we get from (3.5), (3.6)

$$\sum_{k=0}^{n+1} a_k(s)\varphi_{n-k+1}(s) - \varphi_n(s) = \psi_n(s), \quad 0 \le n < N, \tag{3.8}$$

$$\varphi_N(s) = \frac{\lambda}{s+\lambda}\varphi_{N-1}(s) + \frac{\delta_{0l}}{s+\lambda}, \tag{3.9}$$

with

$$\psi_n(s) = a_{n+1}(s)\varphi_0(s) - \sum_{k=n+1}^{\infty} a_k(s)\varphi_1(s) - r_{N-n,l}(s). \tag{3.10}$$

(III) Now, the system (3.8) has exactly the same form as (2.4). Thus, its solution has the same form as (2.5), namely,

$$\varphi_n(s) = R_{n+1}(s)c(s) + \sum_{k=0}^{n} R_{n-k}(s)\psi_k(s), \tag{3.11}$$

where $c(s)$ does not depend on n and $R_k(s)$ is given in (2.3) for the sequence $a_k(s)$.

Now we only need to find unknown $c(s)$, $\varphi_0(s)$, and $\varphi_1(s)$. In order to find $c(s)$, we put $n = 0$ into (3.11) and get $c(s) = \varphi_0(s)/R_1(s)$. Putting $n = 0$ into (3.8) and observing that

$\sum_{k=0}^{\infty} a_k(s) = f(s)$ yield $\varphi_1(s) = (\varphi_0(s) - r_{Nl}(s))/f(s)$. Then, substituting these results into (3.11) we have

$$\varphi_n(s) = \varphi_0(s)c_n(s) + h_{nl}(s), \tag{3.12}$$

$$h_{kl}(s) = \sum_{i=0}^{k} R_{k-i}(s)\left[r_{Nl}(s)\left(1 - \frac{1}{f(s)}\sum_{j=0}^{i} a_j(s)\right) - r_{N-i,l}(s)\right], \tag{3.13}$$

$$c_k(s) = R_{k+1}(s)a_0(s) + \sum_{i=0}^{k} R_{k-i}(s)b_i(s), \tag{3.14}$$

$$b_k(s) = a_{k+1}(s) + \frac{1}{f(s)}\sum_{i=0}^{k} a_i(s) - 1. \tag{3.15}$$

To calculate $\varphi_0(s)$, we set $n = N$ and $n = N-1$ in (3.12) and use the boundary condition (3.9). This gives

$$\varphi_0(s) = \frac{\lambda h_{N-1,l}(s) - (s+\lambda)h_{Nl}(s) + \delta_{0l}}{(s+\lambda)c_N(s) - \lambda c_{N-1}(s)}. \tag{3.16}$$

Using (3.16) and (3.12) with $n' = N - n$, we obtain the final result.

Theorem 3.1. *The transform of the queue size distribution in the M/G/1/N system has the form*

$$\phi_{nl}(s) = c_{N-n}(s)\frac{\lambda h_{N-1,l}(s) - (s+\lambda)h_{N,l}(s) + \delta_{0l}}{(s+\lambda)c_N(s) - \lambda c_{N-1}(s)} + h_{N-n,l}(s), \quad 0 \le n \le N, \tag{3.17}$$

where $h_{kl}(s)$ and $c_k(s)$ are given in (3.13) *and* (3.14)*, respectively.*

Now we can obtain some numerical results.

3.1. Numerical Example

In this example, we will observe the transient queue size distributions and check how long it takes to stabilize the initially overflowed queue.

We assume that the system capacity is 20 (i.e., $N = 20$), the arrival rate is 1 (i.e., $\lambda = 1$), and the service time is constant and equal to 0.9. Therefore, we have $\rho = 0.9$—the traffic intensity is moderate. However, we assume that the system is initially full (i.e., $X(0) = 20$).

Using Theorem 3.1 and the Laplace transform inversion proposed in [20], we obtain the results depicted in Figure 1 and Table 1. In Figure 1 we can observe the queue size distribution after 10, 20, 50, 100, and 200 seconds of the system work and in the steady state (the thick curve). The distribution converges from shapes concentrated around 20 to the steady-state distribution. Shapes close to the steady-state are achieved after about 200 s of the system work.

As we can see, for high values of t, the distribution of the queue size reaches its maximum at $l = 1$. To explain this, we note first that $\rho < 1$. This causes that for growing t

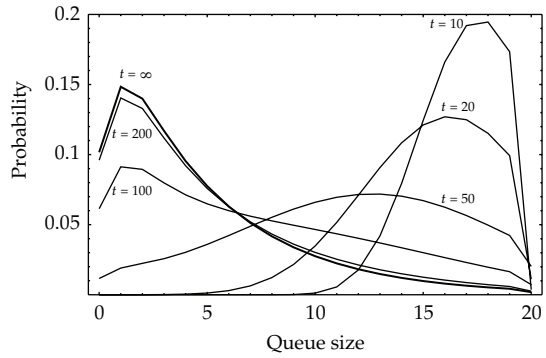

Figure 1: Queue size distributions in the M/D/1/20 system at different moments in time.

Table 1: The mean queue size and the standard deviation in the M/D/1/20 system at different moments in time.

	Mean queue size	Standard deviation
$t = 1$	19.08	0.29
$t = 2$	18.78	0.63
$t = 5$	18.23	1.23
$t = 10$	16.60	1.88
$t = 20$	14.98	2.92
$t = 50$	11.30	4.95
$t = 100$	7.19	5.33
$t = 200$	4.97	4.40
$t = 500$	4.62	4.13
$t = \infty$	4.62	4.13

the probability mass moves towards the level 0. However, the maximum is not at the level 0, which is connected with the fact that the level 0 is a reflecting barrier for the queue size process. Roughly speaking, the level 1 can be reached either from the level 2 (job departure) or from the level 0 (arrival of job to an empty system), but the level 0 can be reached from the level 1 only (job departure). Therefore, the probability at $l = 1$ is higher than at $l = 0$.

In Table 1 we can observe the convergence of the standard deviation to its steady-state value. We may notice that the standard deviation does not change monotonically and reaches a maximum for some $t \in (50, 200)$. This can be explained in the following way. As we start from a queue size of 20, for a small t the probability mass is concentrated around 20. Moreover, as we also have $N = 20$, the queue cannot get longer than 20. Therefore, for a small t, the distribution has only the left tail, and its variance is small. Now, for a large t, as explained before, the probability mass is concentrated around 1, the distribution has only the right tail and its variance is also relatively small. On the other hand, for moderate values of t, the probability mass is distributed more uniformly between 0 and 20, which results in a higher variance. Thus, at least one maximum is to be expected for moderate values of t.

4. Batch Poisson Arrivals and Time to Buffer Overflow

In the second example, we will find a formula for the distribution of the time to buffer overflow in the $M^X/G/1/N$ system. In this model, groups of customers arrive according to the Poisson process with rate λ. Sizes of consecutive groups are independent, identically distributed with discrete distribution $\{p_0, p_1, p_2, \ldots\}$, where $\sum_{i=0}^{\infty} p_i = 1$. The partial rejection scheme is assumed. This means that, in the case of insufficient remaining buffer capacity for all the customers included in an arriving group, only a part of it is accepted and the rest is lost. We assume again that the service time of one customer is distributed according to distribution function $F(t)$, which is not further specified.

We are interested in the distribution of the time to buffer overflow in this system, namely, in the distribution of τ_n, where τ_n is defined as follows:

$$\tau_n = \inf\{t > 0 : X(t) = N \mid X(0) = n\}, \tag{4.1}$$

and $X(t)$ denotes the number of customers in the system at time $t+$.

(I) Using the total probability formula with respect to the first departure epoch for initially nonempty system, $0 < X(0) < N$, we have

$$\begin{aligned} \mathbf{P}(\tau_n > t) = &\sum_{k=0}^{N-n-1} \int_0^t \mathbf{P}(\tau_{n+k-1} > t-u) m_k(u) dF(u) \\ &+ (1 - F(t)) \sum_{k=0}^{N-n-1} m_k(t), \quad 0 < n < N, \end{aligned} \tag{4.2}$$

where $m_k(t)$ denotes the probability that k customers arrive in interval $(0, t]$.

The first term in (4.2) describes the situation where the first departure time u is before t and there is no buffer overflow by the time u, which means that the number of arrivals in $(0, u]$ must not exceed $N - n - 1$. The second term describes the situation where the first departure time is after t and there is no buffer overflow by the time t. Naturally, the situation where an overflow occurs in interval $(0, t]$ is not taken into account now, as in this case we have $\mathbf{P}(\tau_n > t) = 0$.

If the system is initially empty, then conditioning on the first arrival epoch we get

$$\mathbf{P}(\tau_0 > t) = \sum_{k=0}^{N-1} p_k \int_0^t \mathbf{P}(\tau_k > t-u) \lambda e^{-\lambda u} du + e^{-\lambda t}. \tag{4.3}$$

(II) The Laplace transform applied to (4.2) and (4.3) reduces the problem to

$$\begin{aligned} l_n(s) &= \sum_{k=0}^{N-n-1} l_{n+k-1}(s) a_k(s) + \tilde{d}_{N-n}(s), \quad 0 < n < N, \\ l_0(s) &= \frac{\lambda}{s+\lambda} \sum_{k=0}^{N-1} p_k l_k(s) + \frac{1}{s+\lambda}, \end{aligned} \tag{4.4}$$

with

$$l_n(s) = \int_0^\infty e^{-st}\mathbf{P}(\tau_n > t)dt,$$
$$a_k(s) = \int_0^\infty e^{-st}m_k(t)dF(t), \qquad \widetilde{d}_k(s) = \sum_{i=0}^{k-1}\int_0^\infty e^{-st}m_i(t)(1-F(t))dt. \tag{4.5}$$

Substituting $l_n(s) = u_{N-n}(s)$ we obtain

$$\sum_{k=-1}^{n-1} u_{n-k}(s)a_{k+1}(s) - u_n(s) = \varphi_n(s), \quad 0 < n < N, \tag{4.6}$$

$$u_N(s) = \frac{\lambda}{s+\lambda}\sum_{k=0}^{N-1} p_k u_{N-k}(s) + \frac{1}{s+\lambda}, \tag{4.7}$$

where $\varphi_n(s) = u_1(s)a_n(s) - \widetilde{d}_n(s)$.

(III) Now, applying Lemma 2.1 the general solution of the system (4.6) has the form

$$u_n(s) = c(s)R_n(s) + \sum_{k=1}^{n}\varphi_k(s)R_{n-k}(s), \quad n > 0, \tag{4.8}$$

where $c(s)$ does not depend on n and $R_k(s)$ is given in (2.3) for $a_k(s)$ defined in (4.5).

Putting $n = 1$ in (4.8), we can observe that $c(s) = u_1(s)/R_1(s)$. Then, using condition (4.7) together with (4.8), we have

$$u_1(s) = \frac{v_N(s)}{w_N(s)},$$
$$v_N(s) = \frac{\lambda}{s+\lambda}\sum_{k=1}^{N} p_{N-k}\sum_{i=1}^{k}\widetilde{d}_i(s)R_{k-i}(s) - \sum_{k=1}^{N}\widetilde{d}_k(s)R_{N-k}(s) - \frac{1}{s+\lambda}, \tag{4.9}$$

$$w_N(s) = \frac{\lambda}{s+\lambda}\sum_{k=0}^{N} p_{N-k}\sum_{i=0}^{k}a_i(s)R_{k-i}(s) - \sum_{k=0}^{N}a_k(s)R_{N-k}(s). \tag{4.10}$$

Finally, rewriting (4.8) as

$$u_n(s) = u_1(s)\sum_{k=0}^{n}a_k(s)R_{n-k}(s) - \sum_{k=1}^{n}\widetilde{d}_k(s)R_{n-k}(s), \tag{4.11}$$

we arrive at the following theorem.

Theorem 4.1. *The transform of the time to buffer overflow in the $M^X/G/1/N$ system has the form*

$$l_n(s) = \frac{v_N(s)}{w_N(s)} \sum_{k=0}^{N-n} a_k(s) R_{N-n-k}(s) - \sum_{k=1}^{N-n} \tilde{d}_k(s) R_{N-n-k}(s), \quad 0 \le n < N, \tag{4.12}$$

where $v_N(s)$ and $w_N(s)$ are given in (4.9) *and* (4.10)*, respectively.*

To make this theorem useful, we have to be able to compute $a_k(s)$ and $\tilde{d}_k(s)$. Computing these coefficients is not very demanding and may be carried out, for instance, using generating functions. It is easy to check that

$$\begin{aligned} a(z,s) &= \sum_{k=0}^{\infty} z^k a_k(s) = f\big(s + \lambda(1 - p(z))\big), \quad p(z) = \sum_{k=0}^{\infty} z^k p_k, \\ \tilde{d}(z,s) &= \sum_{k=1}^{\infty} z^k \tilde{d}_k(s) = \frac{z\big[1 - f\big(s + \lambda(1 - p(z))\big)\big]}{(1-z)\big(s + \lambda(1 - p(z))\big)}. \end{aligned} \tag{4.13}$$

A very effective algorithm for generating function inversion can be found in [20]. Namely, if we have a generating function $q(z) = \sum_{k=0}^{\infty} q_k z^k$, then the original values of q_k can be restored as

$$q_k \approx \frac{1}{2klr^k} \left(a_0(k,l,r) + (-1)^k a_k(k,l,r) + 2 \sum_{j=1}^{k-1} (-1)^j Re\big(a_j(k,l,r)\big) \right), \tag{4.14}$$

where

$$a_j(k,l,r) = \sum_{n=0}^{l-1} e^{-\pi i n/l} q\Big(r e^{\pi i (n+lj)/lk}\Big), \tag{4.15}$$

while l and r are used to control the roundoff error. (Typically, we use $l = 1$, $r = 10^{-4/k}$.)

An alternative way of computing $a_k(s)$ and $\tilde{d}_k(s)$ is the uniformization technique [21]. Applying this technique to $a_k(s)$, we obtain

$$a_k(s) = \sum_{j=0}^{\infty} \gamma_j(s) K_{k,j}, \tag{4.16}$$

where

$$\gamma_j(s) = \frac{1}{j!} \int_0^{\infty} e^{-t(s+1)} t^j dF(t), \tag{4.17}$$

and $K_{k,j}$ can be computed as follows:

$$\begin{aligned} K_{0,0} &= 1, \\ K_{k,0} &= 0, \qquad k \ge 1, \end{aligned}$$

$$K_{0,j} = 0, \quad j \geq 1,$$
$$K_{k,j} = \sum_{i=0}^{k-1} K_{i,j-1}\, p_{k-i}. \tag{4.18}$$

Similarly, for $\widetilde{d}_k(s)$, we get

$$\widetilde{d}_k(s) = \sum_{i=0}^{k-1}\sum_{j=0}^{\infty} \delta_j(s) K_{i,j}, \tag{4.19}$$

with

$$\delta_j(s) = \frac{1}{j!}\int_0^{\infty} e^{-t(s+1)} t^j (1 - F(t))dt. \tag{4.20}$$

For the bibliography on other computational methods for τ_n, see [22].

4.1. Numerical Example

To demonstrate how (4.12) can be used in practice, let us assume that we have batch Poisson arrivals parameterized as follows:

$$p_1 = \frac{1}{6}, \qquad p_3 = \frac{1}{2}, \qquad p_7 = \frac{1}{3}, \qquad \lambda = \frac{1}{2}. \tag{4.21}$$

Therefore, the average batch size is equal to 4 and the total arrival rate is 2. We assume that the service time is constant and equal to $2/5$ (which gives $\rho = 4/5$) and the buffer size is 100.

Suppose we want to compute the average time to buffer overflow, starting from an empty buffer, that is,

$$\mathbf{E}\tau_0 = \int_0^{\infty} t\, d\mathbf{P}(\tau_0 < t). \tag{4.22}$$

It is easy to see that

$$\mathbf{E}\tau_0 = l_0(0). \tag{4.23}$$

The value of $l_0(0)$ can be computed using the uniformization technique. From (4.17) and (4.20), we obtain, respectively,

$$\gamma_j(0) = \frac{e^{-2/5}(2/5)^j}{j!},$$
$$\delta_j(0) = \frac{\Gamma(j+1,0) - \Gamma(j+1,2/5)}{j!}, \tag{4.24}$$

where $\Gamma(j,x)$ denotes the incomplete gamma function. Now, using (4.16) and (4.19) we can compute $a_k(0)$ and $\tilde{d}_k(0)$. Finally, applying (4.12), we get

$$\mathbf{E}\tau_0 = 82596.64. \tag{4.25}$$

As we have a moderate traffic intensity and a quite big buffer, a large time to buffer overflow was to be expected.

5. MAP Arrivals and Workload Distribution

In the third example, we will find a formula for the workload distribution in the MAP/G/1/N model, that is, for the model with MAP arrivals, general type of the service time distribution (given by distribution function $F(t)$), and finite capacity N.

The Markovian arrival process (MAP) is one of the most flexible arrival processes from the class N of Markovian processes. It enables a very precise fitting to network trace files in terms of not only the basic statistical parameters (mean, variance, higher moments) but also the shape of the marginal distribution and autocorrelation function. (For the newest, excellent parameter fitting procedures for MAP processes, see [11].)

The MAP is parametrized by two $m \times m$ matrices, D_0 and D_1, such that D_1 is nonnegative, D_0 has nonnegative off-diagonal elements, and negative diagonal elements and $D = D_0 + D_1 \neq D_0$ is an irreducible infinitesimal generator. We will use $J(t)$ to denote the state of the underlying Markov chain, $N(t)$ to denote the number of arrivals in $(0,t]$, and $P_{ij}(n,t)$ to denote the counting function, that is,

$$P_{ij}(n,t) = \mathbf{P}\big(N(t) = n, J(t) = j \mid N(0) = 0, J(0) = i\big). \tag{5.1}$$

We will also use intensities λ_i and probabilities $p_i(k,j)$, $k = 0,1$, defined as

$$\lambda_i = -(D_0)_{ii}, \qquad p_i(1,j) = \frac{1}{\lambda_i}(D_1)_{ij}, \quad 1 \le i,j \le m,$$
$$p_i(0,j) = \frac{1}{\lambda_i}(D_0)_{ij}, \quad 1 \le i,j \le m,\ j \neq i. \tag{5.2}$$

By the workload $V(t)$ we mean the length of time a job (packet) which arrives at time t waits before entering service. This is one of the most important characteristics from the practical point of view as it can be used to compute the queueing delay for packets or cells in

network devices. The workload in a MAP queue has been studied so far either in the infinite-buffer model [23] or in the steady state [24]. We assume herein that the workload of a blocked cell is zero.

We will study the workload using its Laplace transform

$$w_{n,i}(s_1,s_2)=\int_0^\infty e^{-s_2t}dt\int_0^\infty e^{-s_1x}\tilde{w}_{n,i}(x,t)dx,$$
$$\tilde{w}_{n,i}(x,t)=\mathbf{P}(V(t)>x\mid X(0)=n,J(0)=i), \tag{5.3}$$

in the column vector form

$$w_n(s_1,s_2)=(w_{n,1}(s_1,s_2),\dots,w_{n,m}(s_1,s_2))^T. \tag{5.4}$$

As previously, $X(t)$ denotes the number of customers in the system at time $t+$.

(I) As in the previous sections, we start from using the total probability formula with respect to the first departure moment. For $0<n\le N$, $1\le i\le m$, we obtain

$$\begin{aligned}\tilde{w}_{n,i}(x,t)=&\sum_{j=1}^{m}\sum_{k=0}^{N-n-1}\int_0^t\tilde{w}_{n+k-1,j}(x,t-u)P_{i,j}(k,u)dF(u)\\&+\sum_{j=1}^{m}\sum_{k=N-n}^{\infty}\int_0^t\tilde{w}_{N-1,j}(x,t-u)P_{i,j}(k,u)dF(u)\\&+\sum_{j=1}^{m}\sum_{k=0}^{N-n-1}P_{i,j}(k,t)\int_t^\infty\left(1-F^{(n+k-1)*}(x-u+t)\right)dF(u),\end{aligned} \tag{5.5}$$

where $F^{(k)*}$ is the k-fold convolution of the distribution function F with itself.

The first double sum in (5.5) describes the case where the first departure time u is before t and there is no buffer overflow by the time u. The second double sum describes the case where the first departure time u is before t and an overflow occurs by the time u, which means that the number of arrivals is equal to $N-n$ or more. The third double sum describes the case where the first departure time u is after t and there is no overflow by the time t. In this case, we have $\mathbf{P}(V(t)>x)=1-F^{(n+k-1)*}(x-u+t)$, where k is the number of arrivals in $(0,t]$.

If the system is initially empty, then for $1\le i\le m$ we get

$$\tilde{w}_{0,i}(x,t)=\sum_{j=1}^{m}\sum_{k=0}^{1}\int_0^t\tilde{w}_{k,j}(x,t-u)p_i(k,j)\lambda_ie^{-\lambda_iu}du. \tag{5.6}$$

(II) Applying transforms and matrix notation to (5.5) and (5.6) we obtain

$$w_n(s_1,s_2) = \sum_{k=0}^{N-n-1} A_k(s_2) w_{n+k-1}(s_1,s_2) + \sum_{k=N-n}^{\infty} A_k(s_2) w_{N-1}(s_1,s_2) + q_n(s_1,s_2), \quad 0 < n \le N, \tag{5.7}$$

$$w_0(s_1,s_2) = \sum_{k=0}^{1} Y_k(s_2) w_k(s_1,s_2), \tag{5.8}$$

where

$$\begin{aligned} A_k(s) &= \left[\int_0^\infty e^{-st} P_{i,j}(k,t)\,dF(t)\right]_{i,j}, \qquad Y_k(s) = \left[\frac{\lambda_i p_i(k,j)}{s+\lambda_i}\right]_{i,j}, \\ \overline{D}_k(s) &= \left[\int_0^\infty e^{-st} P_{i,j}(k,t)(1-F(t))\,dt\right]_{i,j}, \\ C_k(s_1,s_2) &= \left[\int_0^\infty e^{-s_2 t} P_{i,j}(k,t)\,dt \int_0^\infty e^{-s_1 x} d_x F(x+t)\right]_{i,j}, \\ q_n(s_1,s_2) &= \frac{1}{s_1}\sum_{k=0}^{N-n-1}\left[\overline{D}_k(s_2) - f^{n+k-1}(s_1) C_k(s_1,s_2)\right]\cdot \mathbf{1}, \quad \mathbf{1} = (1,\ldots,1)^T. \end{aligned} \tag{5.9}$$

Replacing $v_n(s_1,s_2) = w_{N-n}(s_1,s_2)$ and (5.7) gives

$$\begin{aligned} &\sum_{k=0}^{n+1} A_k(s_2) v_{n-k+1}(s_1,s_2) - v_n(s_1,s_2) = \Psi_n(s_1,s_2), \quad 0 \le n < N, \\ &\Psi_n(s_1,s_2) = A_{n+1}(s_2) v_0(s_1,s_2) - \sum_{k=n+1}^{\infty} A_k(s_2) v_1(s_1,s_2) - q_{N-n}(s_1,s_2). \end{aligned} \tag{5.10}$$

(III) We can see now that (5.10) has the same form as (2.4). Therefore, its solution is given in (2.5). Proceeding in the same way as in the previous sections, we arrive at the final result.

Theorem 5.1. *The transform of the workload distribution in the MAP/G/1/N system has the form*

$$w_n(s_1,s_2) = \sum_{k=0}^{N-n} R_{N-n-k}(s_2) h_k(s_1,s_2) + \left(\sum_{k=-1}^{N-n} R_{N-n-k}(s_2) B_k(s_2) + R_{N-n+1}(s_2)\right) M_N^{-1}(s_2) u_N(s_1,s_2), \quad 0 \le n \le N, \tag{5.11}$$

with

$$
\begin{aligned}
h_n(s_1,s_2) &= \overline{A}_{n+1}(s_2)\left(\overline{A}_0(s_2)\right)^{-1} q_N(s_1,s_2) - q_{N-n}(s_1,s_2),\\
\overline{A}_k(s) &= \sum_{i=k}^{\infty} A_i(s), \qquad B_k(s) = A_{k+1}(s) - \overline{A}_{k+1}(s)\left(\overline{A}_0(s)\right)^{-1},\\
M_N(s) &= \sum_{k=0}^{N+1} R_{N-k+1}(s)B_{k-1}(s) + R_{N+1}(s)\\
&\quad - \sum_{k=N-1}^{N} Y_{N-k}(s)\left[\sum_{i=0}^{k+1} R_{k-i+1}(s)B_{i-1}(s) + R_{k+1}(s)\right],\\
u_N(s_1,s_2) &= \sum_{k=N-1}^{N} Y_{N-k}(s_2)\sum_{i=0}^{k} R_{k-i}(s_2)h_i(s_1,s_2) - \sum_{k=0}^{N} R_{N-k}(s_2)h_k(s_1,s_2).
\end{aligned}
\tag{5.12}
$$

Note that matrices A_k, $\overline{D}_k$, and C_k can be computed effectively by means of the uniformization technique [21]. Using the elementary properties of the Laplace transform we can easily obtain the average workload in steady state—simply by calculating $\lim_{s1,s_2\to 0+} s_2 w_n(s_1,s_2)$. Putting $s_1 = 0+$ into $w_n(s_1,s_2)$ and inverting the result with respect to s_2 only, we can compute the transient average workload. Finally, using a two-dimensional inversion algorithm, we may obtain the shape of the workload distribution for an arbitrary t.

It is easy to check that the number of floating-point operations needed to compute (5.11) (time complexity) grows as $O(m^3N^2)$. This estimate is a consequence of the form of (5.11) and (2.3) and the fact that matrix multiplication and inversion are of $O(m^3)$ order. Thus, the approach proposed herein reduces the numerical complexity when comparing it to the brute-force solution of the system (5.7), which is of $O(m^3N^3)$ order.

5.1. Numerical Example for MAP Arrivals

For numerical purposes, we are going to utilize a parameterization of the MAP based on a recorded IP traffic sample. To accomplish that, the AMP-1138809025-1.tsh trace file, recorded at the AMP aggregation point run by the Passive Measurement and Analysis Project, has been used. Using an implementation of the EM algorithm [7] written for *Mathematica* environment, the following MAP parameterization was obtained:

$$
\begin{aligned}
D_0 &= \begin{bmatrix} -11188.00 & 145.53 & 845.35 & 816.76 \\ 173.27 & -4786.97 & 364.04 & 202.79 \\ 729.86 & 739.55 & -11958.43 & 236.94 \\ 191.44 & 791.87 & 105.75 & -11481.10 \end{bmatrix},\\
D_1 &= \begin{bmatrix} 2467.30 & 2907.27 & 1055.81 & 2949.98 \\ 1229.92 & 609.92 & 1178.33 & 1028.70 \\ 2273.49 & 2442.68 & 3166.51 & 2369.40 \\ 1932.71 & 1522.55 & 3761.29 & 3175.49 \end{bmatrix}.
\end{aligned}
\tag{5.13}
$$

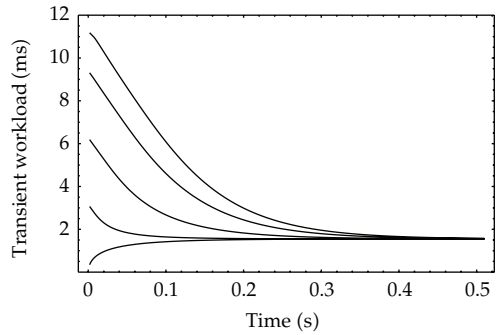

Figure 2: Mean workload versus time for the initial buffer occupancy of 0%, 25%, 50%, 75%, and 100%, counting from the bottom. $\rho = 0.95$, $J(0) = 1$, $N = 100$ pkts.

The average rate of the fitted MAP is

$$\Lambda = \pi \cdot D_1 \cdot \mathbf{1} = 7608.46\,\text{pkts/s}, \tag{5.14}$$

where $\pi = (0.18537, 0.40467, 0.20238, 0.20758)$ is the stationary distribution for the underlying Markov chain $J(\cdot)$ and $\mathbf{1} = (1, \ldots, 1)^T$.

It is assumed that the service time is constant and equal to d. Manipulating d we can easily obtain different traffic intensities $\rho = \Lambda d$.

In Figure 2, the mean workload as a function of time, $\mathbf{E}V(t)$, for the initial buffer occupancy of 0%, 25%, 50%, 75%, and 100% is depicted. The traffic intensity was set for $\rho = 0.95$, the initial phase for $J(0) = 1$, and the buffer size for 100 packets.

As we can see, no matter what the initial buffer occupancy was, the steady-state value (1.549 ms) was reached after about 0.5 s, which is equivalent to about 3800 packet arrivals.

In Figure 3, the stationary mean workload, $\lim_{t\to\infty}\mathbf{E}V(t)$, as a function of the buffer size, is shown for four traffic intensities, namely, 0.75, 0.90, 0.95, and 0.99. In each case, the curve becomes flat starting from some threshold value of the buffer size. For $\rho = 0.75$ this border buffer size is about 20, for $\rho = 0.90$ about 50, for $\rho = 0.95$ about 100, and for $\rho = 0.99$ about 500.

There is an obvious explanation of this behaviour of the workload—for large buffers the finite-buffer system is practically equivalent to the infinite-buffer one; thus, the constant workload observed for large buffers is equal to the infinite-buffer value. However, this behaviour is of some practical importance, especially when the border buffer size is known. Decreasing the buffer size below this border value, we can shorten the queueing delay of the system. The cost paid for this is a higher loss ratio, but it can be beneficial in some applications. In order to evaluate the tradeoff precisely, we have to know the loss ratio, which also can be computed using the method presented in this paper.

6. Conclusions

We presented a unified method for solving queues with Markovian arrivals. The most important features of this approach are the following

(i) it can be applied for finding both steady-state and transient characteristics;

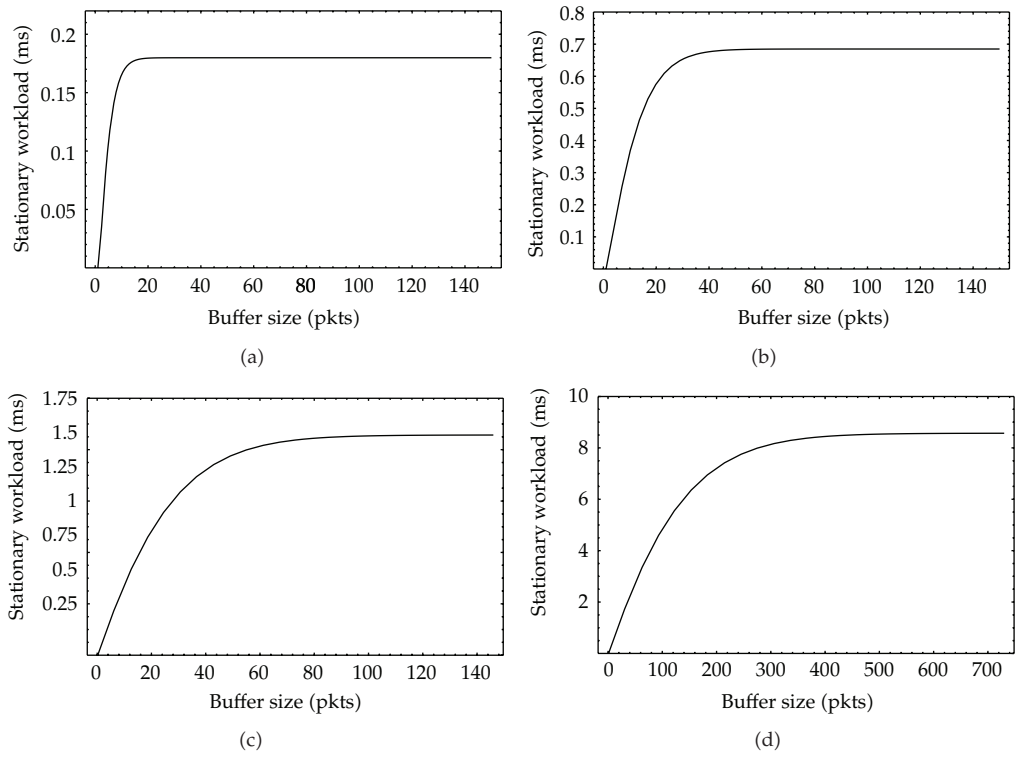

Figure 3: Stationary mean workload versus the buffer size for different traffic intensities. (a): $\rho = 0.75$. (b): $\rho = 0.90$. (c): $\rho = 0.95$. (d): $\rho = 0.99$.

(ii) it produces results in a closed, easy-to-use form;

(iii) reduced numerical complexity comparing to the brute-force solution;

(iv) it is suitable for computing many performance measures of finite-buffer queues, including queue size, workload, loss ratio, time to buffer overflow, buffer overflow period.

The main disadvantage of the method is that it cannot be used directly in solving infinite-buffer queues. This is connected with the necessity to invert the order of the system of equations (for instance, the substitution $\varphi_n(s) = \phi_{N-n}(s)$ in (3.5)), which cannot be carried out in the infinite-buffer model.

References

[1] I. Norros, "A storage model with self-similar input," *Queueing Systems. Theory and Applications*, vol. 16, no. 3-4, pp. 387–396, 1994.

[2] A. Erramilli, R. P. Singh, and P. Pruthi, "An application of deterministic chaotic maps to model packet traffic," *Queueing Systems. Theory and Applications*, vol. 20, no. 1-2, pp. 171–206, 1995.

[3] Y. Shu, Z. Jin, J. Wang, and O. W. Yang, "Prediction-based admission control using FARIMA models," in *Proceedings of the IEEE International Conference on Communications (ICC '00)*, vol. 3, pp. 1325–1329, New Orleans, La, USA, 2000.

[4] R. H. Riedi, M. S. Crouse, V. J. Ribeiro, and R. G. Baraniuk, "A multifractal wavelet model with application to network traffic," *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 992–1018, 1999.
[5] M. F. Neuts, "A versatile Markovian point process," *Journal of Applied Probability*, vol. 16, no. 4, pp. 764–779, 1979.
[6] S. H. Kang, Y. H. Kim, D. K. Sung, and B. D. Choi, "An application of Markovian arrival process (MAP) to modeling superposed ATM cell streams," *IEEE Transactions on Communications*, vol. 50, no. 4, pp. 633–642, 2002.
[7] A. Klemm, C. Lindemann, and M. Lohmann, "Modeling IP traffic using the batch Markovian arrival process," *Performance Evaluation*, vol. 54, no. 2, pp. 149–173, 2003.
[8] T. Rydén, "An EM algorithm for estimation in Markov-modulated Poisson processes," *Computational Statistics & Data Analysis*, vol. 21, no. 4, pp. 431–447, 1996.
[9] P. Salvador, R. Valadas, and A. Pacheco, "Multiscale fitting procedure using Markov modulated Poisson processes," *Telecommunication Systems*, vol. 23, no. 1-2, pp. 123–148, 2003.
[10] P. Buchholz, "An EM-Algorithm for MAP fitting from real traffic data," in *Computer Performance Evaluation Modelling Techniques and Tools*, P. Kemper and W. H. Sanders, Eds., vol. 2794 of *Lectures Notes on Computer Science*, pp. 218–236, Springer, 2003.
[11] G. Casale, E. Z. Zhang, and E. Smirni, "Trace data characterization and fitting for Markov modeling," *Performance Evaluation*, vol. 67, no. 2, pp. 61–79, 2010.
[12] G. Casale, E. Z. Zhang, and E. Smirni, "KPC-Toolbox: best recipes for automatic trace fitting using Markovian Arrival Processes," *Performance Evaluation*, vol. 67, no. 9, pp. 873–896, 2010.
[13] M. F. Neuts, *Structured Stochastic Matrices of M/G/1-Type and Their Applications*, vol. 5, Marcel Dekker, New York, N Y, USA, 1989.
[14] V. Ramaswami, "A stable recursion for the steady state vector in Markov chains of M/G/1 type," *Communications in Statistics. Stochastic Models*, vol. 4, no. 1, pp. 183–188, 1988.
[15] B. Meini, "Solving M/G/1-type Markov chains: recent advances and applications," *Communications in Statistics. Stochastic Models*, vol. 14, no. 1-2, pp. 479–496, 1998.
[16] A. Riska and E. Smirni, "Exact aggregate solutions for M/G/1-type Markov processes," in *Proceedings of the ACM SIGMETRICS Conference*, pp. 86–96, Marina del Rey, Calif, USA, 2002.
[17] A. Chydzinski, "Time to reach buffer capacity in a BMAP queue," *Stochastic Models*, vol. 23, no. 2, pp. 195–209, 2007.
[18] H. Takagi, *Queueing Analysis*, vol. 1, Elsevier Science, 1991.
[19] R. Hadianti, *Wiener-Hopf technique for the analysis of the time-dependent behavior of queues*, Ph.D. thesis, University of Twente, The Netherlands, 2007.
[20] J. Abate, G. L. Choudhury, and W. Whitt, "An introduction to numerical transform inversion and its application to probability models," in *Computational Probability*, W. Grassman, Ed., pp. 257–323, Kluwer Academic Publishers, Boston, Mass, USA, 2000.
[21] D. M. Lucantoni, "New results on the single server queue with a batch Markovian arrival process," *Communications in Statistics. Stochastic Models*, vol. 7, no. 1, pp. 1–46, 1991.
[22] S. Asmussen, M. Jobmann, and H.-P. Schwefel, "Exact buffer overflow calculations for queues via martingales," *Queueing Systems. Theory and Applications*, vol. 42, no. 1, pp. 63–90, 2002.
[23] D. M. Lucantoni, G. L. Choudhury, and W. Whitt, "The transient BMAP/G/1 queue," *Communications in Statistics. Stochastic Models*, vol. 10, no. 1, pp. 145–182, 1994.
[24] C. Blondia, "The N/G/1 finite capacity queue," *Communications in Statistics. Stochastic Models*, vol. 5, no. 2, pp. 273–294, 1989.