IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 June 2023, accepted 28 June 2023, date of publication 3 July 2023, date of current version 7 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3291392

RESEARCH ARTICLE

Waiting Time in a General Active Queue Management Scheme

ANDRZEJ CHYDZINSKI
Department of Computer Networks and Systems, Silesian University of Technology, 44-100 Gliwice, Poland
e-mail: andrzej.chydzinski@polsl.pl

This work was supported by the National Science Centre, Poland, under Grant 2020/39/B/ST6/00224.

ABSTRACT We derive the waiting time in a queueing scheme, in which an arriving job can be denied service with probability relative to the queue size. Such scheme is a generalization of the tail-drop queue, in which the job is denied service when the buffer (waiting room) is full, and can be found in computer networking, call centers and other everyday life applications of queueing systems. To make the model very general, we use an arrival process which enables shaping arbitrary the job interarrival time distribution and interarrival time autocorrelation, as well as general distribution of the service time and job rejection probabilities. For such model, we prove theorems on the waiting time in the transient case, i.e. as a function of time, as well as in the stationary case. Theoretical results are illustrated via numerical examples, in which the dependence of the behaviour of the system on various parameters is depicted. Among other things, it is demonstrated that the assumed job rejection mechanism may induce rather unexpected waiting times if combined with strong autocorrelation of the arrival process.

INDEX TERMS Active queue management, waiting time, workload, transient characteristic.

I. INTRODUCTION

We analyze a queue, in which every arriving job (customer) can be rejected, i.e. denied access to the queue and service. Such a job leaves the system unserved and never returns. What is more, the decision whether a job is allowed to the system or not, is probabilistic, and probability of rejection depends on the queue size upon this job arrival.

The most important area of application of such systems is networking. Algorithms in which the packets arriving to a router's buffer are deleted with probability growing with the queue size have been known and studied via simulations for a long time (see e.g. [1], [2], [3], [4], [5], [6], [7], [8], [9]). Recently, these algorithms were implemented in a networking device and studied in a real network of a university, [10]. The main reason why these algorithms are postulated is the necessity to eliminate high buffer occupancies (bufferbloat), typical in contemporary networks, [11], [12].

Networking is not necessarily the only area of application of the queueing model with rejection probability based on the queue size. For instance, it can be used for modelling a call center, which exploits an answering machine to inform a new caller about the number of callers waiting for the service before him (quite common nowadays). It can be conjectured that probability that a new caller leaves the queue immediately, without service, is a function of the size of the queue ahead of him. In fact, the same reasoning can be applied to any everyday life queue, if only a customer can see the size of the queue upon arrival.

The queueing scheme described above can be perceived as a generalization of the tail-drop queueing scheme, which is well known and used in many computer or electronic systems. In the tail-drop scheme, a buffer of a limited capacity, N, is used to store jobs/tasks/packets before service. When the buffer becomes full, a newly arriving job is rejected. It is easy to see that the tail-drop scheme is a special case of the scheme described above – the rejection probability is 0 when the queue size is below N, and 1, when the queue size is N.

To make the model considered herein general, we assume that the job arrival process can have correlated interarrival times. Such autocorrelation is typical in networking, [13], [14], but can be also found in other applications of queueing

The associate editor coordinating the review of this manuscript and approving it for publication was Tiago Cruz.

systems, including everyday life queues. For instance, a cafe next to a train station may experience an autocorrelated traffic, caused by train arrivals or departures. When the autocorrelation is present in a real system, it is absolutely crucial to take it into account in its model. It has been demonstrated that performance predictions based on the model with omitted autocorrelation can be wrong by several orders of magnitude, even if all other parameters of the model are accurate.

When characterizing any queueing system, one of the most important characteristic is the mean waiting time, i.e. the mean time spent in a queue by a job, before entering service.

Therefore, we derive herein the mean waiting time in the model described above. Both the transient and the stationary case are solved. Namely, we first derive the time-dependent characteristic, i.e. the mean waiting time assuming that a hypothetical job arrives at arbitrary time t (Theorem 1). Then, letting $t \to \infty$, we derive the mean waiting time in the stationary regime (Theorem 2). Pay attention, that having the solution for an arbitrary t, we may use small values of t to study the evolution of the system just after it has been activated, i.e. study the influence of its initial state on the short-time operation of the system.

To illustrate theoretical results, we present numerical examples with diversified parameters, including different autocorrelations of the arrival process, job rejection probabilities and initial states of the system. In these examples we can see how the mentioned parameters influence the operation of the queue.

As the model of the arrival process we use the Markov-modulated Poisson process, [15]. It combines superb modelling capabilities with moderate analytical difficulties. In particular, using this process we can mimic accurately any practically useful shape of the autocorrelation function, together with any practically useful shape of the interarrival time distribution (see, e.g. [16]).

Finally, both the distribution of the service time and the function assigning rejection probabilities are general and can have arbitrary forms. These make the model as general as possible.

Mathematical approach herein exploits regeneration points in the evolution of the system. They enable to formulate a system of integral equations using the total probability law. This method can be used not only to derive the classic performance parameters, e.g. the mean queue size and waiting time, but also some less frequently used characteristics, like the duration of the overflow period, or the burst ratio (see [17], [18]).

The remaining part of the article is organized as follows. Section II outlines the related work. In Section III, the model of the queue is specified, together with the model of the arrival process and its main characteristics. Then, the main results of the paper are shown in Section IV. Firstly, the time-dependent mean waiting time is derived and presented in Theorem 1. Then, as a corollary, the stationary mean waiting time is obtained in Theorem 2. In Section V, numerical examples are gathered. Three parameterizations of MMPP are used with different rejection probabilities and initial conditions to illustrate the waiting time evolution. Finally, conclusions are presented in Section VI.

II. RELATED WORK

As far as the author knows, the results of this article are new.

Queueing systems with job rejection probability based on the queue size have been studied mathematically since the beginning of the century. The great majority of work has been devoted to models with simple Poisson arrivals, [19], [20], [21], [22], [23], [24], [25], [26]. Under such assumption, various characteristics were obtained, including the distribution of the queue size, [19], [20], [21], [23], [24], response time, [25], time between two accepted jobs, [20], busy period, [26] and other. Most papers were devoted to stationary analysis, but some dealt with the transient characteristics as well, [22], [26], [27]. There were very few papers published with different than Poisson arrival process models (see [27], [28], where the renewal process is used).

Now, it should be stressed that none of papers [19], [20], [21], [22], [23], [24], [25], [26], [27], [28] takes into account autocorrelation of the arrival process. As already mentioned, autocorrelation is important for several reasons and ignoring it may produce optimistically erroneous predictions, with an error of several orders of magnitude.

Finally, autocorrelation of traffic in a queue with probabilistic rejections was taken into account in [29] and [30]. However, these papers were devoted to different queueing characteristics, namely the queue size distribution, [29], and the burst ratio, [30].

III. MODEL OF THE SYSTEM

The following queueing model is analyzed herein. Jobs arrive to the service station according to the Markov-modulated Poisson process, which is defined below. At the service station, they are being served in the arrival order. The distribution of the service time is general with distribution function $F(t)$.

An arriving job, if allowed, joins the queue of jobs waiting for service. A job is allowed to join the queue with probability $d(n)$, where n is the number of jobs present in the system upon the new job arrival. That is, with probability $1-d(n)$, the new job is rejected - it leaves the system unserved, immediately after arrival.

The capacity of the system is finite and equal to N. Namely, if upon a job arrival there are N jobs in the system, the new job is rejected with probability 1.

By $X(t)$ the system occupancy (queue size) at the time t will be denoted. The service position is included in the queue size, $X(t)$, if occupied. By M we denote the mean duration of the service time, while the system load is:

$$\rho = \lambda M, \tag{1}$$

where λ is the rate of the Markov-modulated Poisson process, given below in (5).

We assume that a new service begins at $t = 0$ if $X(0) > 0$, what makes the time origin to be the service completion time. This is a technical assumption, which does not cause any loss of generality of the model.

It is clear that the presented model generalizes the tail-drop queueing scheme. In the tail-drop model we have simply $d(n) = 0$ if $n < N$ and $d(n) = 1$ if $n = N$. In the model analyzed here, function $d(n)$ may have an arbitrary form.

Now, the Markov-modulated Poisson process (MMPP), [15], is defined using an auxiliary modulating process, i.e. a continuous-time Markov chain. The modulating process has m states $\{1, \ldots, m\}$. Its rate matrix is denoted by Q, while the modulating state at time t by $J(t)$. The arrivals in an MMPP happen according to the time-inhomogeneous Poisson process, such that the temporary arrival rate at time t is $\lambda_{J(t)}$. Hence, to parameterize an MMPP, we need m arrival rates, $\lambda_1, \ldots, \lambda_m$, in addition to matrix Q. In calculations, it is often convenient to use these rates in the form of a square matrix:

$$\Lambda = diag[\lambda_1, \ldots, \lambda_m]. \tag{2}$$

The main characteristic of an MMPP is its total rate (intensity), denoted by λ. To calculate λ, the stationary distribution of the modulating chain, π, is needed. It can be computed using the set of linear equations:

$$\begin{cases} \pi Q = [0, \ldots, 0], \\ \pi \cdot 1 = 1, \end{cases} \tag{3}$$

where

$$1 = [1, \ldots, 1]^T. \tag{4}$$

Then, the rate of an MMPP is:

$$\lambda = \pi \Lambda 1. \tag{5}$$

The interarrival time density in an MMPP is expressed by matrix exponential. Namely, we have:

$$g(t) = \left[g_{i,j}(t)\right]_{i,j=1,\ldots,m} = De^{-Dt}D^{-1}\Lambda, \tag{6}$$

where

$$g_{i,j}(t) = P\{\tau_{k+1} - \tau_k \in \mathrm{d}t, J(\tau_{k+1}) = j | J(\tau_k) = i\}, \tag{7}$$

$$D = \Lambda - Q, \tag{8}$$

P denotes probability and τ_k is the k-th arrival time. The variance, V, of the interarrival time in an MMPP equals:

$$V = \frac{2}{\lambda}\pi \Lambda D^{-3}\Lambda 1 - \frac{1}{\lambda^2}. \tag{9}$$

In this paper, an important role is played by the autocorrelation function. The k-lag autocorrelation of an MMPP is equal to:

$$R(k) = \frac{1}{\lambda V}\pi \Lambda D^{-2}\Lambda \left((D^{-1}\Lambda)^{k-1} - 1\pi\Lambda/\lambda\right) D^{-2}\Lambda 1. \tag{10}$$

MMPP enables fitting both the interarrival time distribution and the autocorrelation function to observed arrival process. Several methods for fitting matrices Q and Λ were proposed, see e.g. [16], [31], [32], [33], and [34].

IV. WAITING TIME

Let $W_{n,i}(t)$ be the mean time that a job that arrived hypothetically to the system at time t, and was allowed to join the queue, would spend in the system before service, under assumptions $X(0) = n$ and $J(0) = i$. It is easy to see that $W_{n,i}(t)$ is equal to the amount of unfinished work in the system at time t.

$W_{n,i}(t)$ depends on the initial system occupancy, n, and the initial modulating state, i, for every t. Thus it is a transient, time-dependent characteristic.

Define the following Laplace transform:

$$w_{n,i}(s) = \int_0^\infty e^{-st}W_n(t)\mathrm{d}t. \tag{11}$$

Both $W_{n,i}(t)$ na $w_{n,i}(s)$ will be also used in vector forms:

$$W_{n,i}(t) = \left[W_{n,1}(t), \ldots, W_{n,m}(t)\right]^T, \tag{12}$$

$$w_{n,i}(s) = \left[w_{n,1}(t), \ldots, w_{n,m}(s)\right]^T. \tag{13}$$

Denote by $A_{n,k,i,j}(v)$ the probability that in a system with suspended service, k jobs were allowed to join the queue in time interval $(0, v)$ and it was $J(v) = j$, under assumptions $X(0) = n$ and $J(0) = i$. (It will be shown later, how $A_{n,k,i,j}(v)$ can be computed).

Let us assume now that the system is initially non-empty, $X(0) = n > 0$. Conditioning on the end of the first service time, v, we obtain the following system of integral equations for $W_{n,i}(t)$:

$$W_{n,i}(t) = \sum_{j=1}^{m}\sum_{k=0}^{N-n}\int_0^t A_{n,k,i,j}(v)W_{n+k-1,j}(t-v)\mathrm{d}F(v)$$
$$+\sum_{j=1}^{m}\sum_{k=0}^{N-n}A_{n,k,i,j}(t)\int_t^\infty \big(M(n+k-1)+v-t\big)\mathrm{d}F(v), \tag{14}$$

where $n = 1, \ldots, N$, $i = 1, \ldots, m$. Indeed, if the end of the first service time happens before t, than with probability $A_{n,k,i,j}(v)$ there are $n + k - 1$ jobs in the queue at time v and the modulating state at time v is j. Counting from time v, the new, conditional value of the mean waiting time is $W_{n+k-1,j}(t-v)$. Therefore, summing up by all possible values of j, k and v, we obtain the first summand of (14). If the end of the first service time happens after t, than with probability $\sum_{j=1}^{m} A_{n,k,i,j}(t)$ there are $n+k$ jobs in the queue at time t, and one of these jobs is under service with the residual service time of $v - t$. Hence, to calculate the mean waiting time, we need to sum up $n + k - 1$ complete service times, and one incomplete service time of length $v - t$. Again, summing up by all possible values of k and v we arrive at the second summand of (14).

Integration by parts in (14) gives:

$$W_{n,i}(t)=\sum_{j=1}^{m}\sum_{k=0}^{N-n}\int_0^t A_{n,k,i,j}(v)W_{n+k-1,j}(t-v)\mathrm{d}F(v)$$
$$+M\sum_{j=1}^{m}\sum_{k=0}^{N-n}A_{n,k,i,j}(t)(n+k-1)(1-F(t))$$
$$+\sum_{j=1}^{m}\sum_{k=0}^{N-n}A_{n,k,i,j}(t)g(t), \qquad (15)$$

with

$$g(t)=\int_0^\infty(1-F(v+t))\mathrm{d}v. \qquad (16)$$

Administering the Laplace transform to both sides of (15) yields:

$$w_{n,i}(s)=\sum_{j=1}^{m}\sum_{k=0}^{N-n}c_{n,k,i,j}(s)w_{n+k-1,j}(s)$$
$$+M\sum_{j=1}^{m}\sum_{k=0}^{N-n}(n+k-1)r_{n,k,i,j}(s)$$
$$+\sum_{j=1}^{m}\sum_{k=0}^{N-n}h_{n,k,i,j}(s), \quad n=1,\ldots,N, \;\; i=1,\ldots,m. \qquad (17)$$

where

$$c_{n,k,i,j}(s)=\int_0^\infty e^{-sv}A_{n,k,i,j}(v)\mathrm{d}F(v), \qquad (18)$$
$$r_{n,k,i,j}(s)=\int_0^\infty e^{-sv}A_{n,k,i,j}(v)(1-F(v))\mathrm{d}v, \qquad (19)$$
$$h_{n,k,i,j}(s)=\int_0^\infty e^{-sv}A_{n,k,i,j}(v)g(v)\mathrm{d}v, \qquad (20)$$

Then, (17) can be simplified to:

$$w_n(s)=\sum_{k=0}^{N-n}C_{n,k}(s)w_{n+k-1}(s)+M\sum_{k=0}^{N-n}(n+k-1)R_{n,k}(s)1$$
$$+\sum_{k=0}^{N-n}H_{n,k}(s)1, \qquad n=1,\ldots,N, \qquad (21)$$

where

$$C_{n,k}(s)=\left[c_{n,k,i,j}(s)\right]_{i,j=1,\ldots,m}, \qquad (22)$$
$$R_{n,k}(s)=\left[r_{n,k,i,j}(s)\right]_{i,j=1,\ldots,m}, \qquad (23)$$
$$H_{n,k}(s)=\left[h_{n,k,i,j}(s)\right]_{i,j=1,\ldots,m}. \qquad (24)$$

Now we can analyze the situation, where the queue is empty at the time origin. Conditioning on the time of the first event in the arrival process, v, which can be either an arrival of a job, or a change of the modulating state, we have the equation:

$$W_{0,i}(t)=\sum_{j=1}^{m}\int_0^t p_{i,j}(\lambda_i-Q_{i,i})e^{-(\lambda_i-Q_{i,i})v}W_{0,j}(t-v)\mathrm{d}v$$
$$+\left(1-d(0)\right)\sum_{j=1}^{m}\int_0^t\Lambda_{i,j}e^{-(\lambda_i-Q_{i,i})v}W_{1,j}(t-v)\mathrm{d}v$$
$$+d(0)\sum_{j=1}^{m}\int_0^t\Lambda_{i,j}e^{-(\lambda_i-Q_{i,i})v}W_{0,j}(t-v)\mathrm{d}v, \qquad (25)$$

for $i=1,\ldots,m$, where:

$$p_{i,j}=\begin{cases}Q_{i,j}/(\lambda_i-Q_{i,i}), & \text{if } i\neq j,\\ 0, & \text{if } i=j.\end{cases} \qquad (26)$$

Indeed, with intensity $p_{i,j}(\lambda_i-Q_{i,i})e^{-(\lambda_i-Q_{i,i})v}$, the modulating state changes at time v from i to j, without an arrival of a job. If it happens before t, then the new, conditional value of the mean waiting time is $W_{0,j}(t-v)$. This gives the first summand of (25). With intensity $\Lambda_{i,j}e^{-(\lambda_i-Q_{i,i})v}$, a new job arrives at v. If it happens before t and the new job is accepted, then the conditional value of the mean waiting time is $W_{1,j}(t-v)$. This gives the second summand of (25). If the arriving job is rejected, then the conditional value of the mean waiting time is $W_{0,j}(t-v)$, which gives the third summand of (25). We do not have to take into account the situation, when the first job arrives after t, because the mean waiting time at t is then 0.

Administering the Laplace transform to both sides of (25) yields:

$$w_{0,i}(s)=\sum_{j=1}^{m}\frac{(1-d(0))\Lambda_{i,j}}{\lambda_i-Q_{i,i}+s}w_{1,j}(s)$$
$$+\sum_{j=1}^{m}\frac{(\lambda_i-Q_{i,i})p_{i,j}+d(0)\Lambda_{i,j}}{\lambda_i-Q_{i,i}+s}w_{0,j}(s). \qquad (27)$$

Denoting:

$$U_n(s)=\left[\frac{(\lambda_i-Q_{i,i})p_{i,j}+d(n)\Lambda_{i,j}}{\lambda_i-Q_{i,i}+s}\right]_{i,j=1,\ldots,m}, \qquad (28)$$
$$V_n(s)=\left[\frac{(1-d(n))\Lambda_{i,j}}{\lambda_i-Q_{i,i}+s}\right]_{i,j=1,\ldots,m}, \qquad (29)$$

from (27) we get:

$$w_0(s)=V_0(s)w_1(s)+U_0(s)w_0(s). \qquad (30)$$

As we can see, (21) and (30) constitute a system of linear equations with respect to $w_n(s)$, $n=0,\ldots,N$. After some easy algebra, its solution can be presented in the following explicite form.

Theorem 1: The transform of the mean waiting time at time t in a queue fed by MMPP and with rejection probabilities $d(n)$ equals:

$$w(s)=(B(s)-I)^{-1}y(s), \qquad (31)$$

where:

$$w(s) = [w_0(s), \ldots, w_N(s)], \quad y(s) = \begin{bmatrix} y_0(s) \\ \vdots \\ y_N(s) \end{bmatrix}, \tag{32}$$

$$y_0(s) = [0, \ldots, 0]^T, \tag{33}$$

$$y_i(s) = -M\sum_{k=0}^{N-i}(i+k-1)R_{i,k}(s)\mathbf{1} - \sum_{k=0}^{N-i} H_{i,k}(s)\mathbf{1}, \quad i=1, \ldots, N, \tag{34}$$

$$B(s) = [B_{i,j}(s)]_{i,j=0,\ldots,N},$$

$$B_{i,j}(s) = \begin{cases} C_{i,j+1-i}(s), & \text{if } i = 1, \ldots, N,\ j=i-1, \ldots, N-1, \\ U_0(s), & \text{if } i = j = 0, \\ V_0(s), & \text{if } i = 0, j = 1, \\ \mathbf{0}, & \text{otherwise}, \end{cases} \tag{35}$$

and $\mathbf{0}$ is an $m \times m$ matrix of zeros.

Note that matrices $U_0(s)$ and $V_0(s)$, occurring in the theorem above, are easy to compute directly from parameters of the model. On the other hand, matrices $C_{i,j}(s)$, $R_{i,j}(s)$ and $H_{i,j}(s)$ depend on probabilities $A_{n,k,i,j}(v)$, defined at the beginning of this section. Fortunately, these probabilities can be calculated using a result of [29] (see Theorem 1, page 107). Namely, it was proven that:

$$\overline{A}_{n,0}(s) = E_n(s), \tag{36}$$

$$\overline{A}_{n,k}(s) = G_n(s)G_{n+1}(s)\ldots G_{n+k-1}(s)E_{n+k}(s), \quad k \geq 1, \tag{37}$$

where

$$\overline{A}_{n,k}(v) = \left[\int_0^\infty e^{-st} A_{n,k,i,j}(v)\mathrm{d}v\right]_{i,j=1,\ldots,m}, \tag{38}$$

$$G_k(s) = (I - U_k(s))^{-1}V_k(s), \tag{39}$$

$$E_k(s) = (I - U_k(s))^{-1}Z(s), \tag{40}$$

$$Z(s) = diag\left[\frac{1}{\lambda_1 - Q_{1,1} + s}, \ldots, \frac{1}{\lambda_m - Q_{m,m} + s}\right]. \tag{41}$$

Finally, derivation of the stationary mean waiting time from Theorem 1 poses no problem. In particular, it is known that the limit of a function $f(t)$ as $t \to \infty$ is the same as the limit of $s\overline{f}(s)$ as $s \to 0+$, where $\overline{f}(s)$ is the Laplace transform of $f(t)$. Therefore, we get the following result.

Theorem 2: The stationary mean waiting time in a queue fed by MMPP and with rejection probabilities $d(n)$ equals:

$$W = W_{0,1}(\infty) = \lim_{s\to 0+} [(B(s) - I)^{-1}y(s)]_1, \tag{42}$$

where $y(s)$ and $B(s)$ are given in (34) and (35), respectively, while $[\cdot]_1$ is the first entry of a column vector.

Note that the stationary W does not depend on the initial queue size and the initial state of the modulating chain. Therefore, any other n and i could have been taken instead of 0 and 1 in (42).

To use Theorem 1 in practice, the Laplace transform should be inverted. All the numerical examples presented herein were obtained using the inversion method of [35], which is fast and accurate. Theorem 2 can be used directly, without inversion.

V. NUMERICAL EXAMPLES

If not stated otherwise, the following MMPP will be used in this section:

$$Q = \begin{bmatrix} -0.04751 & 0.02823 & 0.01928 \\ 0.02101 & -0.04231 & 0.02130 \\ 0.03952 & 0.03219 & -0.07171 \end{bmatrix}, \tag{43}$$

$$\Lambda = \begin{bmatrix} 0.04906 & 0 & 0 \\ 0 & 0.29804 & 0 \\ 0 & 0 & 3.88622 \end{bmatrix}, \tag{44}$$

of rate $\lambda = 1$.

The justification for choosing this parametrization of MMPP is that it produces a positive autocorrelation of traffic, which can be expected in networking. The autocorrelation provided by matrices (43) and (44) is moderate. It will be used as default. However, it can be easily strengthen or weakened multiplying Q by a positive number, if needed. This will be done in Section V-C, where in addition to parameterization (43) and (44), two other parameterizations will be considered, of much stronger and weaker autocorrelations, respectively.

If not stated otherwise, the following rejection probability function will be used:

$$d(n) = \begin{cases} 0, & \text{if } \ n < 16, \\ \frac{n}{16} - 1, & \text{if } \ 16 \leq n < 32, \\ 1, & \text{if } \ n \geq 32. \end{cases} \tag{45}$$

This will be altered in Section V-B, where five other, non-linear rejection probability functions will be used in addition to (45).

Finally, the distribution of the service time will be hyperexponential with parameters: (0.25, 0.75), (4.0, 0.8). It can be verified easily that this distribution has the mean of $M = 1$ and a moderate standard deviation of 1.17. Therefore, the system will be fully saturated: $\rho = \lambda M = 1$.

A. IMPACT OF THE INITIAL STATE OF THE QUEUE

In Fig. 1, the mean waiting time is depicted as a function of time. The initial modulating state is 1 in every case, but different initial queue sizes are used, from 0 to 32.

As we could expect, the transient evolution of the system depends strongly on the initial size of the queue. However, after about 80s, the waiting time reaches the stationary value.

For comparison, in Fig. 2 the same initial queue sizes are used, but with the initial modulating state of 3. As we can see, different modulating state made the transient evolution quite different. However, the time of convergence to the stationary value is more or less the same in Figs. 1 and 2.

The impact of the initial modulating state on the evolution of the system can be studied further in Figs. 3 and 4. Namely, in Fig. 3 the mean waiting time is depicted as a function of time for all three modulating states, but unaltered initial queue

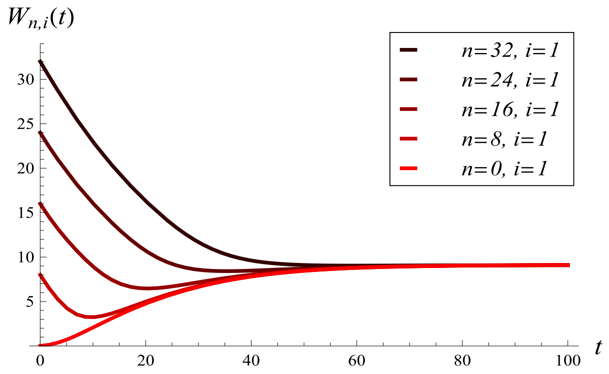

FIGURE 1. The mean waiting time for different initial queue sizes and $i = 1$.

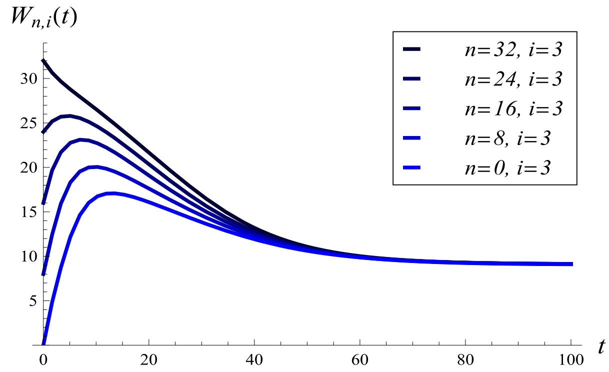

FIGURE 2. The mean waiting time for different initial queue sizes and $i = 3$.

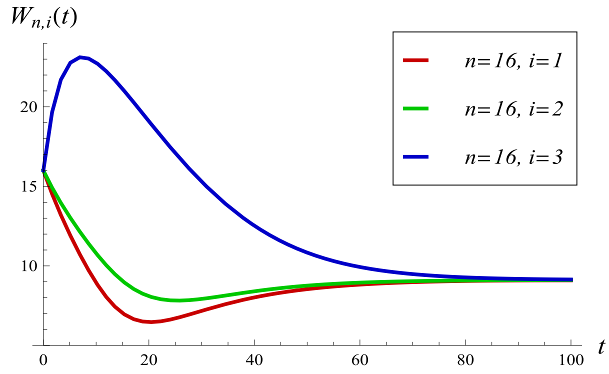

FIGURE 3. The mean waiting time for different initial modulating states and $n = 16$.

size of 16. Similarly, in Fig. 4, the initial queue size of 32 is used in combination with different modulating states.

As we can observe in Figs. 1-4, the influence of initial modulating state on the transient evolution is stronger when the initial queue size is small or moderate, and weaker, when the initial queue size is high.

Finally, it is worth mentioning, that in many cases the transient evolution of the waiting time is non-monotonic. See, for instance, the curves for $n = 8$ and $n = 16$ in Fig. 1, almost all curves in Fig. 2, and all curves in Fig. 3. Moreover, both a local minimum or a local maximum can occur (Fig. 3).

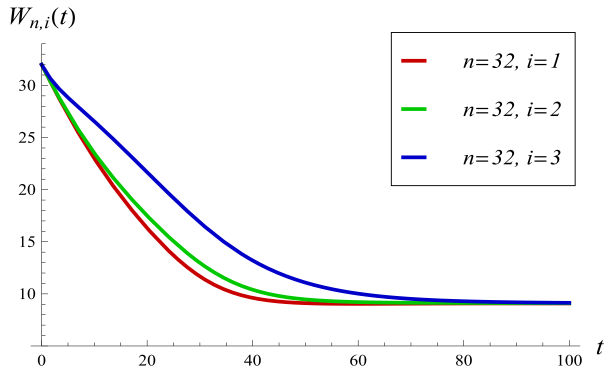

FIGURE 4. The mean waiting time for different initial modulating states and $n = 32$.

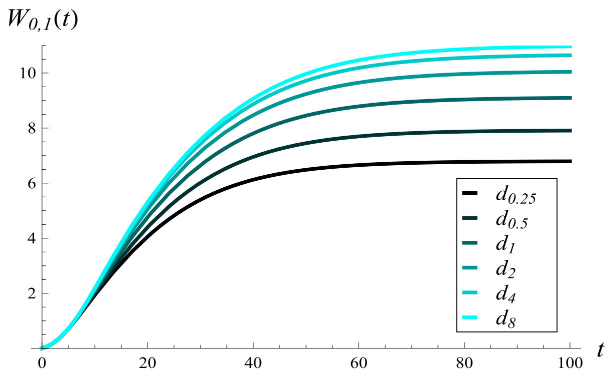

FIGURE 5. The mean waiting time for different rejection probability functions and $n = 0, i = 1$.

B. IMPACT OF DROP PROBABILITIES

In the section above, the rejection probability function was not altered and had always the form of (45). In this section, we will use the following function instead:

$$d_x(n) = (d(n))^x, \tag{46}$$

where $x > 0$ is a parameter and $d(n)$ is given in (45). Parameter x has an easy interpretation – the smaller x, the quicker the system will start rejecting jobs and the more of them will be rejected. An vice versa – a large x is equivalent to small rejection probabilities.

In Figs. 5, 6 and 7, the mean waiting time is depicted for six different rejection probability functions, namely $d_{0.25}$, $d_{0.5}$, d_1, d_2, d_4 and d_8. The difference between Figs. 5, 6 and 7 is that Fig. 5 was obtained for the initial system state $n = 0$ and $i = 1$, Fig. 6 for the initial state $n = 16$ and $i = 2$, while Fig. 7 for the initial state $n = 32$ and $i = 3$.

As we can see in these figures, the convergence to the stationary value is slightly quicker, when a strong rejection function is used (e.g. $d_{0.25}$), and slightly slower, when a weak function is applied (e.g. d_8). The differences, however, are

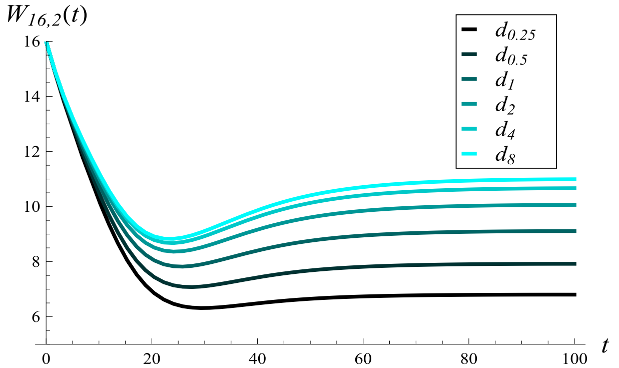

FIGURE 6. The mean waiting time for different rejection probability functions and $n = 16$, $i = 2$.

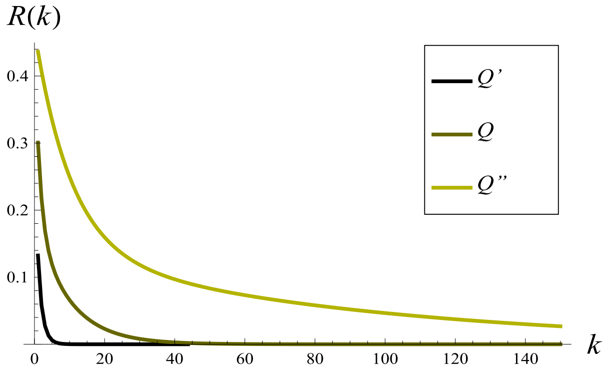

FIGURE 8. Autocorrelation function for traffic parameterized by matrices Q, Q' and Q''.

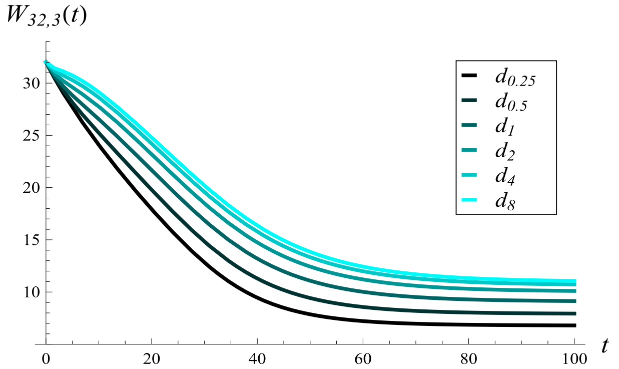

FIGURE 7. The mean waiting time for different rejection probability functions and $n = 32$, $i = 3$.

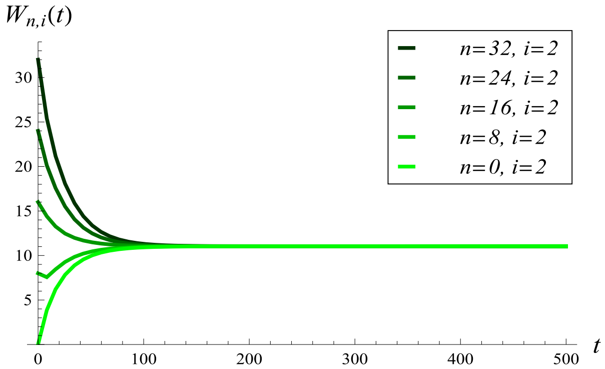

FIGURE 9. The mean waiting time for weak autocorrelation (Q'), several initial queue sizes and $i = 2$.

not profound – in all cases the stationary value is reached somewhere between 60s and 100s.

Moreover, the evolution of the system in each figure is different, depending on the initial state of the queue, no matter which d_x function was used.

Finally, we can see that for high values of x, the curves change less and less. This can be easily explained by the fact that if $x \to \infty$, then the rejection scheme approaches the tail-drop scheme, i.e. the rejection probability becomes 0 for every $n < 32$ and 1 for $n = 32$. Therefore, the curves in Figs. 5, 6 and 7 converge to the tail-drop curve, when x grows.

C. IMPACT OF AUTOCORRELATION

So far, only the parameterization of the MMPP given in (43) and (44) was used. We will alter this in this section, to obtain different strengths of the autocorrelation. Namely, the following two other parameterizations will be used:

MMPP': $Q' = 10Q$, $\Lambda' = \Lambda$,

MMPP": $Q'' = Q/10$, $\Lambda'' = \Lambda$,

together with function $d(n)$ from (45).

In Fig. 8, the autocorrelation functions for matrices Q' and Q'' are depicted, and accompanied by the original autocorrelation for Q. As we can see, autocorrelation for Q'' is strong and long term, while for Q' is weak and short term. Autocorrelation for the original Q is moderate, somewhere between the two.

In Fig. 9 the mean waiting time is presented as a function t for weak autocorrelation (Q'), several initial queue sizes and $i = 2$. Similarly, in Fig. 10 the mean waiting time is presented for strong autocorrelation (Q''), several initial queue sizes and $i = 2$.

When comparing Fig. 9 with Fig. 10, the first striking observation is the time of convergence to the stationary value. In the case of strong autocorrelation, the time of convergence is about 5 times longer (approx. 500s versus 100s). This was to be expected - the long term autocorrelation should cause such an effect.

An interesting and rather surprising observation can be made if we compare stationary waiting times for the three MMPP parameterizations considered so far. Namely, the stationary waiting times are 10.9, 9.2 and 6.9 for matrices Q', Q and Q'', i.e. for weak, moderate and strong autocorrelation, respectively. Note that other parameters of the system are not altered (the load, distribution of the service time and rejection probabilities).

This observation contradicts a naive belief that the stronger positive autocorrelation of traffic is, the worse all queueing performance characteristics are. In our case, the stronger

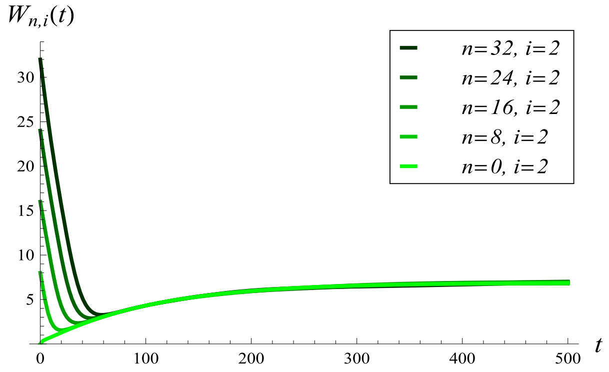

FIGURE 10. The mean waiting time for strong autocorrelation (Q''), several initial queue sizes and $i = 2$.

autocorrelation, the better (shorter) stationary waiting time. This effect is, naturally, of great significance in networking, where we can always expect autocorrelated traffic.

The observed phenomenon can be explained by the interaction between traffic autocorrelation and rejection function, $d(n)$. It is true that a strong, positive autocorrelation drives the queue size to grow. In a system without job rejections, we indeed would have higher queue sizes and waiting times for stronger autocorrelations. In our system, however, there are job rejections caused by $d(n)$. Moreover, function $d(n)$ rejects the more jobs, the longer the queue is. Therefore, the overall number of rejected jobs grows when autocorrelation gets stronger. In effect, the carried load of a system with the rejection mechanism decreases, even if the offered load, ρ, remains unaltered.

In other words, the reduced stationary waiting time for strongly correlated traffic comes with the price of increased number of rejected jobs.

VI. CONCLUSION

In this paper, an analysis of the waiting time was carried out in a queueing scheme, in which an arriving job can be denied service with probability relative to the queue size. Such queueing scheme can be found in computer networking, call centers and other customer service systems. It is also a generalization of the commonly used tail-drop scheme. Therefore, all the results presented herein can be used for the tail-drop queue as well, by applying function $d(n)$=0 for $n < N$ and 1 otherwise.

A general model of the queue was used, with the arrival process of arbitrary interarrival time distribution and interarrival time autocorrelation, arbitrary distribution of the service time and job rejection probabilities. For this model, two theorems on the mean waiting time were proven - in the transient and stationary regime, respectively.

These theorems were illustrated via numerical examples, in which the dependence of the transient behaviour of the system on various parameters was depicted. In particular, it was shown how the initial state of the queue and the form of the rejection probability function influence the transient evolution of the mean waiting time.

A special consideration was given to autocorrelation of traffic. It was shown that strong autocorrelation may increase significantly the time of convergence to the stationary value. It was also demonstrated that strong autocorrelation may cause the mean stationary waiting time to be shorter, if compared with a weak autocorrelation case. This contradicts a naive belief that the stronger autocorrelation of traffic, the worse all queueing performance characteristics. Such a simplification is clearly not valid in queues with the rejection mechanism.

Both of these observations are important in networking, where we do have traffic autocorrelation. Namely, when the rejection mechanism is used, we can expect long non-stationary periods of operation of the system, as well as reduced waiting times, at the price of increased number of rejected jobs.

REFERENCES

[1] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Trans. Netw.*, vol. 1, no. 4, pp. 397–413, Aug. 1993.

[2] S. Athuraliya, S. H. Low, V. H. Li, and Q. Yin, "REM: Active queue management," *IEEE Netw.*, vol. 15, no. 3, pp. 48–53, May 2001.

[3] V. Rosolen, O. Bonaventure, and G. Leduc, "A RED discard strategy for ATM networks and its performance evaluation with TCP/IP traffic," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 29, no. 3, pp. 23–43, Jul. 1999.

[4] K. Zhou, K. L. Yeung, and V. O. K. Li, "Nonlinear RED: A simple yet efficient active queue management scheme," *Comput. Netw.*, vol. 50, no. 18, pp. 3784–3794, Dec. 2006.

[5] D. R. Augustyn, A. Domański, and J. Domańska, "A choice of optimal packet dropping function for active queue management," in *Computer Networks* (Communications in Computer and Information Science), vol. 79. Berlin, Germany: Springer-Verlag, 2010, pp. 199–206.

[6] J. Domańska, D. R. Augustyn, and A. Domański, "The choice of optimal 3-rd order polynomial packet dropping function for NLRED in the presence of self-similar traffic," *Bull. Polish Acad. Sci., Tech. Sci.*, vol. 60, no. 4, pp. 779–786, Dec. 2012.

[7] C. Feng, L. Huang, C. Xu, and Y. Chang, "Congestion control scheme performance analysis based on nonlinear RED," *IEEE Syst. J.*, vol. 11, no. 4, pp. 2247–2254, Dec. 2017.

[8] S. Patel and Karmeshu, "A new modified dropping function for congested AQM networks," *Wireless Pers. Commun.*, vol. 104, no. 1, pp. 37–55, Jan. 2019.

[9] A. Giménez, M. A. Murcia, J. M. Amigó, O. Martínez-Bonastre, and J. Valero, "New RED-type TCP-AQM algorithms based on beta distribution drop functions," *Appl. Sci.*, vol. 12, no. 21, p. 11176, Nov. 2022.

[10] M. Barczyk and A. Chydzinski, "AQM based on the queue length: A real-network study," *PLoS ONE*, vol. 17, no. 2, Feb. 2022, Art. no. e0263407.

[11] J. Gettys and K. Nichols, "Bufferbloat: Dark buffers in the Internet," *Queue*, vol. 9, no. 11, pp. 1–15, 2011.

[12] V. G. Cerf, "Bufferbloat and other Internet challenges," *IEEE Internet Comput.*, vol. 18, no. 5, pp. 79–80, Sep./Oct. 2014.

[13] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Netw.*, vol. 2, no. 1, pp. 1–15, Feb. 1994.

[14] M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: Evidence and possible causes," *IEEE/ACM Trans. Netw.*, vol. 5, no. 6, pp. 835–846, Dec. 1997.

[15] W. Fischer and K. Meier-Hellstern, "The Markov-modulated Poisson process (MMPP) cookbook," *Perform. Eval.*, vol. 18, no. 2, pp. 149–171, Sep. 1993.

[16] P. Salvador, R. Valadas, and A. Pacheco, "Multiscale fitting procedure using Markov modulated Poisson processes," *Telecommun. Syst.*, vol. 23, pp. 1–2, pp. 123–148, 2003.

[17] A. Chydzinski, "Duration of the buffer overflow period in a batch arrival queue," *Perform. Eval.*, vol. 63, nos. 4–5, pp. 493–508, May 2006.

[18] A. Chydzinski, D. Samociuk, and B. Adamczyk, "Burst ratio in the finite-buffer queue with batch Poisson arrivals," *Appl. Math. Comput.*, vol. 330, pp. 225–238, Aug. 2018.

[19] T. Bonald, M. May, and J.-C. Bolot, "Analytic evaluation of RED performance," in *Proc. INFOCOM*, 2000, pp. 1415–1424.

[20] W. M. Kempa, "On main characteristics of the M/M/1/N queue with single and batch arrivals and the queue size controlled by AQM algorithms," *Kybernetika*, vol. 47, no. 6, pp. 930–943, 2011.

[21] O. Tikhonenko and W. M. Kempa, "The generalization of AQM algorithms for queueing systems with bounded capacity," in *Proc. Int. Conf. Parallel Process. Appl. Math.*, in Lecture Notes in Computer Science, vol. 7204, 2012, pp. 242–251.

[22] W. M. Kempa, "A direct approach to transient queue-size distribution in a finite-buffer queue with AQM," *Appl. Math. Inf. Sci.*, vol. 7, no. 3, pp. 909–915, May 2013.

[23] O. Tikhonenko and W. M. Kempa, "Performance evaluation of an M/G/n-type queue with bounded capacity and packet dropping," *Int. J. Appl. Math. Comput. Sci.*, vol. 26, no. 4, pp. 841–854, Dec. 2016.

[24] O. Tikhonenko and W. M. Kempa, "Erlang service system with limited memory space under control of AQM mechanizm," in *Proc. Int. Conf. Comput. Netw.*, in Communications in Computer and Information Science, vol. 718, 2017, pp. 366–379.

[25] A. Chydzinski, M. Barczyk, and D. Samociuk, "The single-server queue with the dropping function and infinite buffer," *Math. Problems Eng.*, vol. 2018, pp. 1–12, Oct. 2018.

[26] A. Chydzinski, "Non-stationary characteristics of AQM based on the queue length," *Sensors*, vol. 23, no. 1, p. 485, Jan. 2023.

[27] W. M. Kempa, "Time-dependent queue-size distribution in the finite GI/M/1 model with AQM-type dropping," *Acta Electrotechnica et Informatica*, vol. 13, no. 4, pp. 85–90, 2013.

[28] W. Hao and Y. Wei, "An extended $GI^X/M/1/N$ queueing model for evaluating the performance of AQM algorithms with aggregate traffic," in *Proc. Int. Conf. Netw. Mobile Comput.*, in Lecture Notes in Computer Science, vol. 3619, 2005, pp. 395–414.

[29] P. Mrozowski and A. Chydzinski, "Queues with dropping functions and autocorrelated arrivals," *Methodol. Comput. Appl. Probab.*, vol. 20, no. 1, pp. 97–115, Mar. 2018.

[30] A. Chydzinski and B. Adamczyk, "On the influence of AQM on serialization of packet losses," *Sensors*, vol. 23, no. 4, p. 2197, Feb. 2023.

[31] L. Deng and J. W. Mark, "Parameter estimation for Markov modulated Poisson processes via the EM algorithm with time discretization," *Telecommun. Syst.*, vol. 1, no. 1, pp. 321–338, Dec. 1993.

[32] T. Rydén, "An EM algorithm for estimation in Markov-modulated Poisson processes," *Comput. Statist. Data Anal.*, vol. 21, pp. 431–447, Apr. 1996.

[33] T. Yoshihara, S. Kasahara, and Y. Takahashi, "Practical time-scale fitting of self-similar traffic with Markov-modulated Poisson process," *Telecommun. Syst.*, vol. 17, nos. 1–2, pp. 185–211, 2001.

[34] L. N. Singh and G. R. Dattatreya, "A novel approach to parameter estimation in Markov-modulated Poisson processes," in *Proc. IEEE Emerg. Technol. Conf. (ETC)*, Richardson, TX, USA, Oct. 2004, pp. 1–6.

[35] V. Zakian, "Numerical inversion of Laplace transform," *Electron. Lett.*, vol. 5, pp. 120–121, Mar. 1969.

ANDRZEJ CHYDZINSKI received the M.Sc. degree (Hons.) in applied mathematics and the Ph.D. and D.Sc. degrees in computer science from the Silesian University of Technology, in 1997, 2002, and 2008, respectively. Currently, he is a Full Professor with the Silesian University of Technology and the Head of the Department of Computer Networks and Systems. He acted as the project leader in six large scientific projects, founded by research agencies, and participated in several other scientific projects as a researcher. He has authored two scientific monographs and about 120 peer-reviewed articles, including publications in many leading scientific journals. His current research interests include computer networking, in particular performance evaluation, network modeling, queueing theory, virtualization of networks, and discrete-event simulations. He received five awards for outstanding conference papers and a prestigious award from POLITYKA, a well-recognized Polish magazine. In 2015, he received the Professor title from the President of the Republic of Poland.