ANALYSIS OF $MAP/M/1/K$ TICKET QUEUE WITH USERS BALKING AND RENEGING AND SERVICE OF NO-SHOW USERS

Chesoong Kim
Department of Business Administration
Sangji University
26339, Wonju, Republic of Korea
Email: dowoo@sangji.ac.kr

Alexander Dudin, Sergei Dudin, Olga Dudina
Department of Applied Mathematics and
Computer Science
Belarusian State University
220030, 4 Nezavisimosti Ave., Minsk, Belarus
Email: dudin@bsu.by, dudins@bsu.by,
dudina@bsu.by

KEYWORDS

Impatient users; Ticket queue; Markovian arrival process; Admission control

ABSTRACT

The $MAP/M/1/K$ type ticket queue is analysed. Arriving users obtain a ticket for service but may balk the system with the probability depending on the queue of tickets ahead of him/her. Irrespective of joining or balking, the ticket issued for the user remains in the ticket buffer. Also, the waiting user can abandon the system during the waiting time, but his or her ticket remains in the buffer. Users are served in the order of their tickets. The server is unaware of the presence of users in the system and spends some time for his/her service even if the respective user already left the system. Ticket queuing systems describe the wide range of real world system. The stationary distribution of the considered system is computed. Numerical illustrations are presented.

I. Introduction

Queuing theory is widely used for the performance evaluation and optimization of various industrial, logistics, and telecommunications systems and communication networks. Classical queuing models assume that each incoming user is accepted into the system if there is at least one free space in the queue and is served in a certain order. There is also a sort of queuing systems, so-called "visible queues", in which an arriving user observes the length of a queue and makes a decision to join the queue or to balk even if the buffer is not full, see, for example, (Sun et al. 2018).

Also, in classical queuing models, it is assumed that accepted users always wait for their turn and will certainly be served. However, frequently in real systems, users may be impatient and leave the system after some waiting time if their service has not begun. The relevant literature is quite extensive, see, for example, (Dudin et et al. 2022; Garnett et al. 2002; Wang et al. 2010).

Another kind of practically important and interesting queues is so-called ticket queues. In such queues, each arriving user receives a numbered ticket (slip, token, etc.) and observes the number of the user being served, which is broadcast on a display panel. When the service of the user with the displayed number is finished, the system calls for the next number, i.e., the service is implemented according to the First In - First Out discipline. In a ticket queue, no physical queue is formed; a customer can only see his own ticket number and the number being served. Based on information about the difference between his/her number and the displayed number, the user decides whether to balk or wait for service. In contrast to the usual queues with balking users, the balking user in ticket queues leaves the system physically, but his number remains in the queue. However, the server does not have any information about which users are physically present and which have left the system. We call below the former users as active users and the latter ones as inactive users. During the active user's stay in the buffer, he or she may show impatience and depart from the system. Thus, he or she becomes an inactive user but his or her ticket remains in the queue.

Ticket queues are considered in the literature due to their high practical importance and many advantages over technologies requiring the physical presence of the user, see, e.g., (Xuet et al. 2007). The ticket queue technology has seen widespread use in financial institutions, government agencies, health care organizations, and retail stores, for more details and concrete examples, see, e.g., (Hanukov et al. 2021; Xuet et al. 2007). In those papers, good reviews of relevant research are presented and $M/M/1$ ticket queue is analysed.

In real practical systems, users' departure due to balking and/or impatience leads to an increase in their loss probability, which, in turn, negatively affects the revenue received by the system. In addition, the system

Communications of the ECMS, Volume 37, Issue 1,
Proceedings, ©ECMS Enrico Vicario, Romeo Bandinelli,
Virginia Fani, Michele Mastroianni (Editors) 2023
ISBN: 978-3-937436-80-7/978-3-937436-79-1 (CD) ISSN 2522-2414

may suffer reputation losses due to the dissatisfaction of users who left the system without receiving the service. Therefore, system managers should try to minimize the negative consequences associated with such user behavior. In ticket queues, such a behavior creates more problems for system management than in the classical queues and needs more careful analysis.

It is supposed in (Xuet et al. 2007) that an arriving user balks if the difference between his or her number and the number of the serviced user exceeds the predefined threshold. No reneging of the users (departure from the queue during waiting time) is supposed. The authors propose a Markov chain (MC) model of a ticket queue operation and develop effective tools for an approximate evaluation of the system's performance.

A similar model is described and analyzed in (Jennings and Pender 2016). But authors additionally allow users to resign from the queue. The paper contains the heavy traffic-inspired approximations for performance measures of the system and their comparison with the analogous measures of the corresponding classical queueing system.

In (Xiao et al. 2022), analysis provided in (Xuet et al. 2007) is extended for the use for managerial goals. It is assumed that there are two levels of server operation, distinguished by the service rate and abandonment probability, depending on the current level.

Advantages of analysis provided in our paper over the results from (Jennings and Pender 2016; Xiao et al. 2022; Xuet et al. 2007) are threefold.

- We do not assume that the service time of inactive users (no-show users) is negligible compared to the service time of active users. From our personal experience, namely the waste of time required to the service provider to realize that the called user will not show up is the main disadvantage of the ticket queue. This time includes the time until the waiting user reaches the server, while some users in real systems do not rush. Also, in many real-world systems, the call to a user is repeated if the user does not approach the server during the fixed time. All this makes the service to an inactive user far from negligible. The proper account of service time of inactive users is important to the use in managerial goals because this time causes irritation of users and loss of the throughput of the server.
- We assume that an arbitrary user balks with the probability arbitrarily dependent on the difference between the user's own number and the displayed number. Frequently considered in the literature, the threshold strategy is the particular case of our randomized strategy in which the balking probabilities are equal to zero when the difference does not exceed the threshold and equal to one otherwise.
- All papers mentioned assume the stationary Poisson arrival process. We consider the essentially more general model of Markov arrival process (MAP), see, e.g., (Chakravarthy 2022a, 2022b; Dudin et al. (2020); Lucantoni (1991)). This allows us to avoid the serious under-estimating of the required server throughput and buffer capacity in comparison with the stationary Poisson arrival process which ignores essential fluctuation of arrival rate in real world system.

Note that importance of account of non-negligible service time of inactive users is indicated in (Hanukov et al. 2021) where this time is referred to as calling time.

II. Mathematical model

We consider a single-server queuing system with a finite buffer of capacity K, the structure of which is displayed in Figure 1.

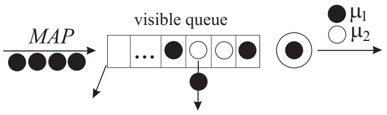

Fig. 1. Structure of the system

The MAP flow of users enters the system. This arrival flow is defined by the underlying process ν_t, $t \geq 0$, which is an irreducible MC with continuous-time and the finite state space $\{1, 2, \dots, W\}$, and the matrices D_0, D_1. The average user arrival intensity is denoted as λ and calculated as $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$ where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_W)$ is the invariant probability vector of the chain ν_t. It is defined as the only solution to the system $\boldsymbol{\theta}(D_0 + D_1) = \mathbf{0}$, $\boldsymbol{\theta}\mathbf{e} = 1$. Here and throughout this paper, $\mathbf{e}$ is a column vector of appropriate size consisting of units, and $\mathbf{0}$ is a row vector of appropriate size composed of zeros. More detailed descriptions of the MAP and formulas for finding its characteristics, e.g., the coefficients of correlation and variation, can be found in (Chakravarthy 2022a, 2022b; Dudin et al. 2020; Lucantoni 1991).

If an arriving user finds the server idle, he or she occupies it and starts processing. Otherwise, he/she occupies a free place in a buffer and waits until he or she will be picked up for service according to the First In – First Out discipline. If during the arrival epoch the buffer is full, the user leaves the system permanently. After joining the system, the user can observe the queue. We assume that the user decides that the queue length is too long for him/her and becomes an inactive customer with the probability q_k, $0 \leq q_k \leq 1$, where k is the number of users (active and inactive) in the system during the arrival epoch. With the complimentary probability $1 - q_k$, the arriving user will stay in the buffer as an active one.

An inactive user occupies a place in the buffer but does not require the full service, and the server will not gain profit from serving such a user. Note that the server cannot recognize whether or not the user is inactive before the start of his/her service. An arbitrary user's service time follows an exponential distribution with the parameter μ_1, $\mu_1 > 0$, if this user is active, and with the parameter μ_2, $\mu_2 \geq \mu_1$, if this user is inactive.

The active users staying in the buffer can be impatient. This means that, independently of other users and their own ticket number, each active user can be-come inactive after an exponentially distributed time with the parameter β, $\beta > 0$.

III. The process of system states and its stationary distribution

Let k_t, $k_t = \overline{0, K+1}$, be the number of users in the system (in the buffer and on the server); n_t, $n_t = \overline{0, \max\{0, k_t - 1\}}$, be the number of active users in the buffer; m_t be the state of the server: $m_t = 0$, if the server provides service to an active user, $m_t = 1$, if the server provides service to an inactive user; ν_t, $\nu_t = \overline{1, W}$, be the state of the underlying process of the MAP at time t, $t \geq 0$.

Unfortunately, the four-dimensional random process $\xi_t = \{k_t, n_t, m_t, \nu_t\}$, $t \geq 0$, is non-Markovian because the rate of service of the next user depends on whether this user is active or not. Therefore, to have the Markovian random process, it is necessary to supplement the introduced components with components defining the status of each user in the queue. The simplest way is to explicitly indicate the status of each user. In such a way, the component n_t can be eliminated. For $k_t \geq 1$, each of $k_t - 1$ users in the queue is marked by the number 1 if he/she is active and number 0 otherwise. The cardinality of this set of components, for each $k_t \geq 1$, is $2^{k_t - 1}$. Correspondingly, the size of the system equilibrium equations will be $W2^{K+1}$. This number can be large and solving this system becomes infeasible.

Another possible way to obtain the Markovian process via supplementing the process ξ_t is to supplement it by the indication of places in the buffer occupied by the active users and the number of inactive users in the queue staying in turn after any active user. Such a way of supplementing is discussed in (Hanukov et al. 2021; Xuet et al. 2007). However, as it is mentioned in (Xuet et al. 2007), the authors of (Xuet et al. 2007) did not succeed to manage computation for $K > 9$. Recall, that the queueing model considered in (Xuet et al. 2007) is significantly more simple for analysis than our model due to assumption about zero service time of inactive users and the parameter W in their study is equal to 1 while we admit an arbitrary finite W.

Therefore, we restrict ourselves to providing an approximate analysis. We make the following assumption. If the number of users in the buffer at an user service completion moment is equal to k, $k = \overline{1, K}$, and the number of active users is j, $j = \overline{1, k}$, then the next service will be provided to the active user with the probability $\frac{j}{k}$. This assumption can be omitted if the balking probability q_k does not depend on the number of users in the system k and when the users are picked up from the queue in random order, but not in the order of arrival.

Under this assumption, the process $\xi_t = \{k_t, n_t, m_t, \nu_t\}$, $t \geq 0$, becomes the MC.

Let the states of the MC ξ_t be enumerated in lexicographic order.

Theorem 1. The generator Q of the MC ξ_t, $t \geq 0$, has the following block-tridiagonal structure:

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & \dots & O & O & O \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & \dots & O & O & O \\ O & Q_{2,1} & Q_{2,2} & \dots & O & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ O & O & O & \dots & Q_{K,K-1} & Q_{K,K} & Q_{K,K+1} \\ O & O & O & \dots & O & Q_{K+1,K} & Q_{K+1,K+1} \end{pmatrix}$$

where

$$Q_{0,0} = D_0,\ Q_{1,1} = I_2 \otimes D_0 - \text{diag}\{\mu_1, \mu_2\} \otimes I_W,$$

$$Q_{k,k} = I_{2k} \otimes D_0 - I_k \otimes \text{diag}\{\mu_1, \mu_2\} \otimes I_W -$$

$$-\beta C_k \otimes I_{2W} + \beta C_k E_k^- \otimes I_{2W},\ k = \overline{2, K},$$

$$Q_{K+1,K+1} = I_{2(K+1)} \otimes (D_0 + D_1) - I_{K+1} \otimes \text{diag}\{\mu_1, \mu_2\} \otimes I_W -$$

$$-\beta C_{K+1} \otimes I_{2W} - \beta C_{K+1} E_{K+1}^- \otimes I_{2W},$$

$$Q_{0,1} = \begin{pmatrix} 1 & 0 \end{pmatrix} \otimes D_1,$$

$$Q_{k,k+1} = (1 - q_k) E_k^+ \otimes I_2 \otimes D_1 + q_k \tilde{E}_k \otimes I_2 \otimes D_1,\ k = \overline{1, K},$$

$$Q_{1,0} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \otimes I_W,$$

$$Q_{k,k-1} = H_k \tilde{E}_k^- \otimes \begin{pmatrix} \mu_1 & 0 \\ \mu_2 & 0 \end{pmatrix} \otimes I_W +$$

$$+(I - H_k)\hat{E}_k \otimes \begin{pmatrix} 0 & \mu_1 \\ 0 & \mu_2 \end{pmatrix} \otimes I_W,\ k = \overline{2, K+1}.$$

Here

$\otimes$ is the symbol of the Kronecker product of matrices;

$C_k = \text{diag}\{0, 1, 2, \dots, k-2, k-1\}$, $k = \overline{2, K+1}$;

diag$\{\dots\}$ means the diagonal matrix with the diagonal entries listed in the brackets;

E_k^-, $k = \overline{2, K+1}$, is a square matrix of size k with all zero entries except the ones $(E_k^-)_{l,l-1}$, $l = \overline{1, k-1}$, which are equal to 1;

E_k^+, $k = \overline{1, K}$, is a $k \times (k+1)$ matrix with all zero entries except the ones $(E_k^+)_{l,l+1}$, $l = \overline{0, k-1}$, which are equal to 1;

$\tilde{E}_k$, $k = \overline{1, K}$, is a $k \times (k+1)$ matrix with all zero entries except the ones $(\tilde{E}_k)_{l,l}$, $l = \overline{0, k-1}$, which are equal to 1;

$\tilde{E}_k^-$, $k = \overline{2, K+1}$, is a $k \times (k-1)$ matrix with all zero entries except the ones $(\tilde{E}_k^-)_{l,l-1}$, $l = \overline{1, k-1}$, which are equal to 1;

$\hat{E}_k$, $k = \overline{2, K+1}$, is a $k \times (k-1)$ matrix with all zero entries except the ones $(\hat{E}_k)_{l,l}$, $l = \overline{0, k-1}$, which are equal to 1;

$H_k = \text{diag}\{0, \frac{1}{k-1}, \frac{2}{k-1}, \dots, \frac{k-2}{k-1}, 1\}$, $k = \overline{2, K+1}$.

The theorem is proved by analyzing the intensities of all possible transitions of the MC ξ_t over an infinitesimal time interval. The block-tridiagonal form of the Q generator is easily explained by the fact that users come into the system and leave it one at a time.

If the system is empty (the buffer is empty and the server is idle), the behavior of the MC ξ_t is determined only by the process ν_t. The rates of its transitions to other states are given by non-diagonal entries of the matrix D_0, and the exit intensities from the corresponding states are determined up to sign by the diagonal elements of this matrix, hence $Q_{0,0} = D_0$.

Then let us explain the form of the block $Q_{k,k},\ k = \overline{1, K+1}$. This is a diagonal block of the generator, so all its diagonal entries are negative, and the moduli of these entries determine the exit intensities of the MC ξ_t from the corresponding states. The exit of the MC ξ_t from the current state is possible in the following cases:

1) The underlying process ν_t of users arriving leaves its current state. The corresponding transition intensities are determined up to a sign by the diagonal entries of the matrices $I_{2k} \otimes D_0$ if $k = \overline{1, K+1}$.

2) The user completes the service. The transition intensities are determined by the diagonal entries of the matrices $I_k \otimes \text{diag}\{\mu_1, \mu_2\} \otimes I_W$, $k = \overline{1, K+1}$.

3) An active user from the buffer leaves the system due to impatience. The corresponding intensities are given by the matrices $\beta C_k \otimes I_{2W}$, $k = \overline{2, K+1}$.

The non-diagonal entries of the matrix $Q_{k,k},\ k = \overline{1, K+1}$, determine the transition intensities of the MC ξ_t without changing the value k of the first component. These transitions are defined by the following entries:

1) non-diagonal entries of the matrices $I_{2k} \otimes D_0$, $k = \overline{1, K+1}$, when the underlying process ν_t makes a transition without generating a user.

2) entries of the matrices $\beta C_k E_k^- \otimes I_{2W}$, $k = \overline{2, K+1}$, when an active user becomes inactive due to impatience.

3) entries of the matrix $I_{2(K+1)} \otimes D_1$, if a user arrives to the system when the buffer is full (when $k = K+1$) and leaves the system without receiving a ticket.

As a result, we obtain the blocks $Q_{k,k},\ k = \overline{0, K+1}$, presented above.

The form of blocks $Q_{k,k+1}$, $k = \overline{0, K}$, is explained as follows. These blocks contain the transition rates of the MC ξ_t that lead to increase in the number of users in the system (in the buffer and in the server) by one. If $k = 0$ (the server is idle and there are no users in the buffer), these transitions happen when a user arrives at the system and starts servicing. The transition intensities of this event are determined by the entries of the matrix $\begin{pmatrix} 1 & 0 \end{pmatrix} \otimes D_1$. If $k = \overline{1, K}$, an increase in the number of users in the system from the value k to the value $k+1$ can occur when a new user enters the system. The transition intensities of this event are defined by the entries of the matrices $(1-q_k)E_k^+ \otimes I_2 \otimes D_1$, if the arriving user joins the buffer as an active customer and the entries of the matrices $q_k \tilde{E}_k \otimes I_2 \otimes D_1$ if the arriving user becomes an inactive one.

Now consider the blocks $Q_{k,k-1},\ k = \overline{1, K+1}$. These blocks contains of the transition intensities of the MC ξ_t from the state with the value k of the first component to the state with the value $k-1$ of this component. Such transitions are possible only in the case of a completed service. If the buffer is empty during the service completion epoch, the intensities of this event are given by the entries of the matrices $\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \otimes I_W$. Otherwise, the intensities of this event are given by the entries of the matrices $H_k \tilde{E}_k^- \otimes \begin{pmatrix} \mu_1 & 0 \\ \mu_2 & 0 \end{pmatrix} \otimes I_W$ if an active user starts service and the entries of the matrices $(I - H_k)\hat{E}_k \otimes \begin{pmatrix} 0 & \mu_1 \\ 0 & \mu_2 \end{pmatrix} \otimes I_W$ if an inactive user is chosen for service. Taking into account all these explanations, we obtain the formulas for the blocks $Q_{k,k-1},\ k = \overline{1, K+1}$, presented above.

It is obvious that the stationary probabilities of the system states $\pi(k, n, m, \nu)$, $k = \overline{0, K+1}$, $n = \overline{0, \max\{0, k-1\}}$, $m = 0, 1$, $\nu = \overline{1, W}$, exist for all possible values of the system parameters. Let us form the row vectors $\boldsymbol{\pi}_k$ of these probabilities enumerated in the lexicographic order of the components n, m, ν. It is well known that these vectors satisfy the following system of linear algebraic equations:

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{K+1})Q = \mathbf{0}, \ \ (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{K+1})\mathbf{e} = 1$$

where Q is the infinitesimal generator of the MC ξ_t, $t \geq 0$. To solve this system, we recommend using the efficient and numerically stable algorithm developed in (Dudin et al. 2021).

IV. Performance measures

The mean number of users in the system is calculated by the formula $L = \sum\limits_{k=1}^{K+1} k\boldsymbol{\pi}_k\mathbf{e}$.

The mean number of users in the buffer is computed by

$$N_{buf} = \sum_{k=2}^{K+1}(k-1)\boldsymbol{\pi}_k\mathbf{e}.$$

The mean number of active users in the buffer is computed by

$$N_{buf-act} = \sum_{k=2}^{K+1}\sum_{n=1}^{k-1} n\boldsymbol{\pi}(k,n)\mathbf{e}.$$

The mean number of inactive users in the buffer is computed by

$$N_{buf-inact} = \sum_{k=2}^{K+1}\sum_{n=0}^{k-2}(k-1-n)\boldsymbol{\pi}(k,n)\mathbf{e} =$$

$$N_{buf} - N_{buf-act}.$$

The average output rate of active users is defined as:

$$\lambda_{out-act} = \mu_1 \sum_{k=1}^{K+1}\sum_{n=0}^{k-1}\boldsymbol{\pi}(k,n,0)\mathbf{e}.$$

The average output rate of inactive users is defined as:

$$\lambda_{out-inact} = \mu_2 \sum_{k=1}^{K+1}\sum_{n=0}^{k-1}\boldsymbol{\pi}(k,n,1)\mathbf{e}.$$

The probability that an arbitrary moment the system is idle is computed as $P_{idle} = \boldsymbol{\pi}_0\mathbf{e}$.

The probability that an arrival user starts service upon arrival is computed as $P_{imm} = \frac{1}{\lambda}\boldsymbol{\pi}_0 D_1\mathbf{e}$.

The loss probability of an arbitrary active user from the buffer due to impatience is calculated by the formula:

$$P_{imp-loss} = \frac{\beta}{\lambda}\sum_{k=2}^{K+1}\sum_{n=1}^{k-1} n\boldsymbol{\pi}(k,n)\mathbf{e} = \frac{\beta N_{buf-act}}{\lambda}.$$

The loss probability of an arbitrary user upon arrival due to abandonment is calculated by the formula:

$$P_{balk} = \frac{1}{\lambda}\sum_{k=1}^{K} q_k \boldsymbol{\pi}_k (I_{2k} \otimes D_1)\mathbf{e}.$$

The loss probability of an arbitrary user upon arrival due to a full buffer is defined as:

$$P_{ent-busy-loss} = \frac{1}{\lambda}\boldsymbol{\pi}_{K+1}(I_{2(K+1)} \otimes D_1)\mathbf{e}.$$

The loss probability of an arbitrary customer is computed as

$$P_{loss} = 1 - \frac{\lambda_{out-act}}{\lambda} =$$

$$P_{imp-loss} + P_{ent-abad-loss} + P_{ent-busy-loss}.$$

The last expression can be used for an accuracy check during the debugging of the code and the computation of stationary probabilities and performance measures.

V. Numerical examples

We consider five arrival flows with the same average arrival intensity of users $\lambda = 3$ which is defined as follows.

The first arrival process coded as M is stationary Poisson. It is defined by the matrices $D_0 = (-3)$, $D_1 = (3)$ of size $W = 1$. It has coefficients of correlation c_{cor} and variation of successive inter-arrival times c_{var} equal to 0 and 1, correspondingly.

The second flow is defined by the matrices

$$D_0 = \begin{pmatrix} -1.20113 & 0.480452 & 0.720678 \\ 3.90367 & -208.396 & 204.492 \\ 3.90367 & 3.90367 & -810.763 \end{pmatrix},$$

$$D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 300.883 & 502.072 & 0 \end{pmatrix}.$$

It is the IPP flow with $c_{cor} = 0$ and $c_{var} = 4$.

The rest arrival flows are MAPs with the same variation $c_{var} = 4$ but different correlation coefficients. Denote by MAP_x the MAP flow with the coefficient of correlation equal to x.

$MAP_{0.1}$ is defined by the matrices:

$$D_0 = \begin{pmatrix} -7.96609 & 0.359372 & 0.359372 \\ 0.38932 & -1.49739 & 0.239582 \\ 0.419268 & 0.239582 & -0.958327 \end{pmatrix},$$

$$D_1 = \begin{pmatrix} 6.88797 & 0.239582 & 0.119791 \\ 0.269529 & 0.539059 & 0.0598954 \\ 0.149739 & 0.0299477 & 0.119791 \end{pmatrix}.$$

$MAP_{0.2}$ is defined by the matrices:

$$D_0 = \begin{pmatrix} -9.55774 & 0.362952 & 0.362952 \\ 0.30246 & -1.36107 & 0.272214 \\ 0.362952 & 0.332706 & -1.17959 \end{pmatrix},$$

$$D_1 = \begin{pmatrix} 8.58986 & 0.181476 & 0.060492 \\ 0.060492 & 0.635166 & 0.090738 \\ 0.060492 & 0.120984 & 0.30246 \end{pmatrix}.$$

$MAP_{0.3}$ is defined by the matrices:

$$D_0 = \begin{pmatrix} -15.4094 & 0.241244 & 0.211088 \\ 0.0874508 & -1.34493 & 0.120622 \\ 0.180933 & 0.241244 & -1.05544 \end{pmatrix},$$

$$D_1 = \begin{pmatrix} 14.6254 & 0.271399 & 0.0603109 \\ 0.0211088 & 1.00418 & 0.111575 \\ 0. & 0.150777 & 0.482487 \end{pmatrix}.$$

We assume that the intensity of impatience $\beta = 0.03$, the service rate of active users $\mu_1 = 5$, the service rate of inactive users $\mu_2 = 6$. The probabilities q_k are defined as $q_k = k/101$, $k = \overline{0, K}$. We vary the buffer capacity in the interval $[0; 50]$ with step 1.

Figures 2-6 illustrate the dependence of the average output intensity $\lambda_{out-act}$ of active users, the average output intensity $\lambda_{out-inact}$ of inactive users, the loss probabilities of an arbitrary user upon arrival due to full buffer $P_{ent-busy-loss}$ and due to abandonment P_{balk}, the loss probability of an arbitrary active user from the buffer due to impatience $P_{imp-loss}$ on the buffer size K for different arrival processes described above.

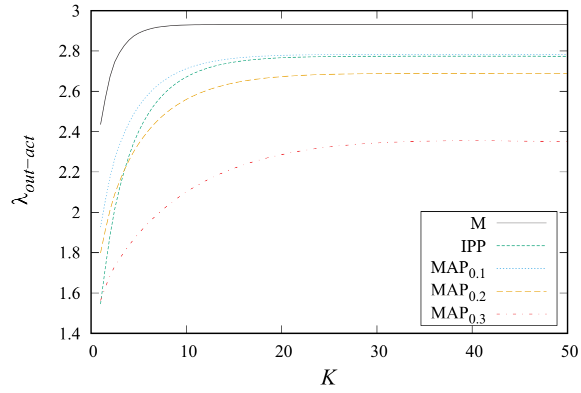

Fig. 2. Dependence of $\lambda_{out-act}$ on the buffer size K for different arrival processes

One may conclude from these figures that correlation has a profound effect on the performance measures of the system. The operation of the system became significantly worse with the growth of correlation. The assumption that the arrival flow is described by the stationary Poisson arrival process imposed in the cited papers leads to an overly optimistic prediction of the system's performance.

Let us assume that the quality of the system's operation is defined by the following economic criterion defining the revenue of the system:

$$E = E(K) = a\lambda_{out-act} - b\lambda_{out-inact} -$$

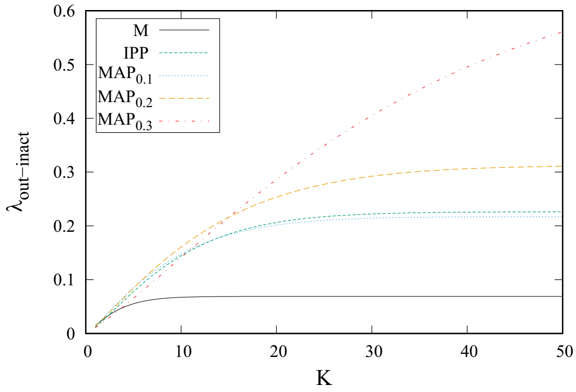

Fig. 3. Dependence of $\lambda_{out-inact}$ on the buffer size K for different arrival processes

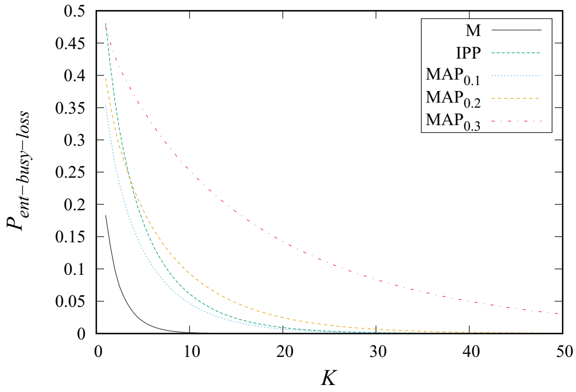

Fig. 4. Dependence of the probability $P_{ent-busy-loss}$ on the buffer size K for different arrival processes

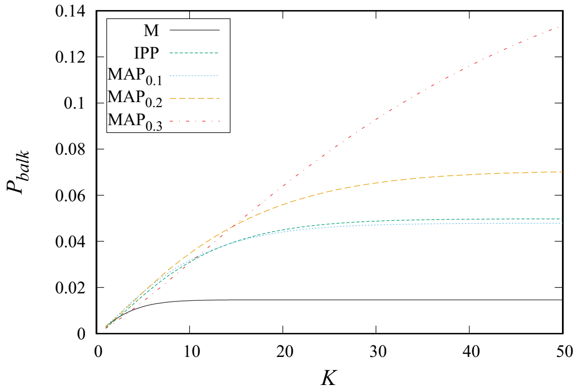

Fig. 5. Dependence of the probability P_{balk} on the buffer size K for different arrival processes

$$c_1\lambda P_{balk} - c_2\lambda P_{ent-busy-loss} - c_3\lambda P_{imp-loss} - dK$$

where a is a profit obtained by the system for servicing one active user, b is a system cost for servicing an inactive user, c_1, c_2 and c_3 are charges paid by the system for a customer loss due to balking, buffer overflow and impatience respectively, and d is the charge paid by the system for maintenance on one unit of buffer space. This economic criterion $E(K)$ defines the average profit obtained by the system per unit of time. Our aim is to optimally choose the buffer capacity K to maximize the average system profit.

We fix the following values of cost coefficients: $a = 5$, $b = 2$, $c_1 = 1/3$, $c_2 = 1/2$, $c_3 = 2/3$, $d = 0.02$.

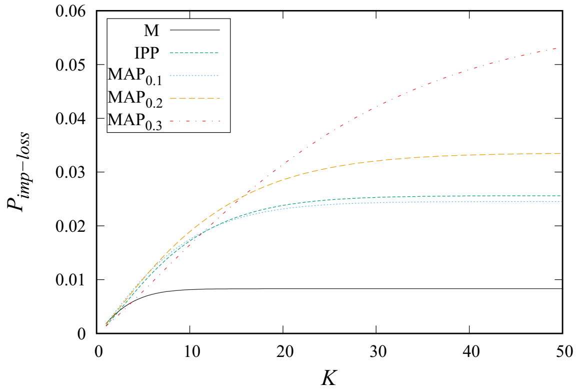

Fig. 6. Dependence of the probability $P_{imp-loss}$ on the buffer size K for different arrival processes

The dependence of the cost criterion $E(K)$ on the buffer size K for different arrival flows is shown in Figure 7.

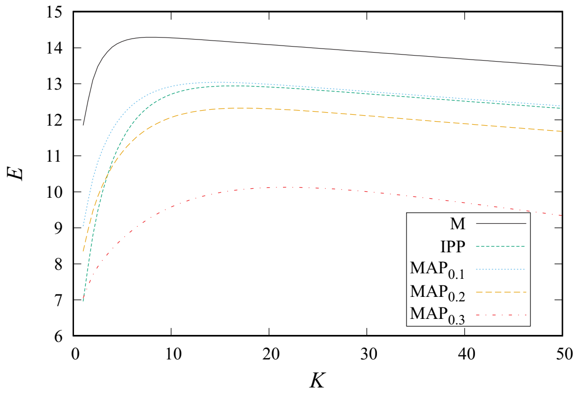

Fig. 7. Dependence of the probability E on the buffer size K for different arrival processes

The optimal values of E^* the cost criterion and K^* of the buffer capacity for different arrival flows having the same average arrival rate are presented in Table 1.

TABLE I: The optimal values of the cost criterion for different arrival processes

	M	IPP	$MAP_{0.1}$	$MAP_{0.2}$	$MAP_{0.3}$
E^*	14.2931	12.9421	13.0415	12.3241	10.1269
K^*	8	16	15	17	22

Based on Figure 7 and Table 1, it is possible to conclude the following:

- Choice of the proper capacity of the buffer allows the system to increase its revenue.
- The optimal capacity of the buffer essentially depends on the correlation in the arrival process (under the same mean arrival rate);
- The optimal capacity of the buffer increases as the correlation in the arrival process increases.
- The maximum revenue of the system decreases as the correlation increases.

VI. Conclusion

The approximate Markov model of $MAP/M/1/K$ type ticket queue is analysed. Arriving users can balk the system after receiving a ticket with the probability dependent on the current length of the queue of the tickets. An admitted user can balk immediately if he or she decides that the queue size is inappropriate and renege after some random amount of time. Tickets of users who left the system remain in the queue, and the server needs to waste time for their processing along with the service of users who did not balk or renege.

Presented numerical results illustrate the importance of account correlation in the arrival process and the possibility of an optimal choice of the size of buffer space. Results are planned to be extended to the multi-server systems, including systems with varying number of servers, systems with more general, than exponential, phase type distribution of service times, and systems with service rate dependent on the queue length.

VII. Acknowledgments

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2022R1I1A3067543)

REFERENCES

Chakravarthy, S.R. 2022a. *Introduction to Matrix-Analytic Methods in Queues 1: Analytical and Simulation Approach – Basics.* ISTE Ltd, London and John Wiley and Sons, New York.

Chakravarthy, S.R. 2022b. *Introduction to Matrix-Analytic Methods in Queues 2: Analytical and Simulation Approach – Queues and Simulation.* ISTE Ltd, London and John Wiley and Sons, New York.

Dudin, A.N.; V.I. Klimenok; and V.M. Vishnevsky. 2020. *The theory of queuing systems with correlated flows.* Springer Nature, ISBN 978-3-030-32072-0.

Dudin, A.; O. Dudina; S. Dudin; and K. Samouylov. 2021. "Analysis of Single-Server Multi-Class Queue with Unreliable Service, Batch Correlated Arrivals, Customers Impatience and Dynamical Change of Priorities." *Mathematics*, 9(1257) doi:10.3390/math9111257

Dudin, A.; O. Dudina; S. Dudin; and Y. Gaidamaka. 2022. "Self-Service System with Rating Dependent Arrivals". *Mathematics*, 10(3), 297.

Garnett, O.; A. Mandelbaum; and M. Reiman. 2002. "Designing a call center with impatient customers". *Manufacturing & Service Operations Management*, 4(3), 208-227.

Hanukov, G.; M. Hassoun; and O. Musicant. 2021. "On the benefits of providing timely information in ticket queues with balking and calling times". *Mathematics*, 9(21), 2753.

Jennings, O.B. and J. Pender. 2016. "Comparisons of ticket and standard queues". *Queueing Systems*, 84, 145-202.

Lucantoni, D. 1991. "New results on the single server queue with a batch Markovian arrival process", *Communication in Statistics-Stochastic Models*, 7, 1-46.

Sun, B.; A. Dudin; and S. Dudin. 2018. "Queueing system with impatient customers, visible queue and replenishable inventory". *Applied and Computational Mathematics*, 17(2), 161-174.

Wang, K.; N. Li; and Z. Jiang. 2010. "Queueing system with impatient customers: A review". In *Proceedings of 2010 IEEE international conference on service operations and logistics, and informatics*, 82-87.

Xiao, L.; S.H. Xu; D.D. Yao; and H. Zhang. 2022. "Optimal staffing for ticket queues". *Queueing Systems*, 102(1-2), 309-351.

Xu, S.H.; L. Gao; and J. Ou. 2007. "Service performance analysis and improvement for a ticket queue with balking customers". *Management Science*, 53(6), 971-990.

CHESOONG KIM earned his PhD in Engineering at the Department of Industrial Engineering at Seoul National University in 1993. He was a Visiting Scholar in the Department of Mechanical Engineering at the University of Queensland, Australia from September of 1998 to August of 1999. He was foreign scientist at the School of Mathematics & Statistics at Carleton University, Canada from July of 2003 to August of 2004. He was also a Visiting Professor in the Department of Industrial Engineering at the University of Washington, USA from August of 2004 to August of 2005. He had scientific visits to Belarusian State University, University of Debrecen and Azerbaijan National Academy of Sciences, respectively. He is currently Full Professor of Business Administration and Department of Industrial Engineering at Sangji University. His current research interests are in stochastic process, queueing theory with particular emphasis on computer and wireless communication network, queueing network modeling and their applications. He has published around 100 papers in internationally refereed journals.

ALEXANDER DUDIN earned PhD degree in Probability Theory and Mathematical Statistics in 1982 from Vilnius University and Doctor of Science degree in 1992 from Tomsk University. He is Head of Laboratory of Applied Probabilistic Analysis in Belarusian State University. He is author of more than 450 publications including 5 books and more than 130 papers in top level journals. He coedited more than ten volumes of the Springer series. In 2013, he received Scopus Award Belarus for outstanding contribution to the field of Mathematics. He is the Chairman of Belarusian Winter Workshops in Queueing Theory which are held since 1985 and the Chairman of IPC of the International conference named after A.F. Terpugov since 2014. He was invited for lecturing and research to USA, UK, Germany, France, the Netherlands, Japan, Korea, India, Russia, China, Italy, and Sweden.

SERGEI DUDIN was graduated from Belarusian State University in 2007. In 2010, he earned PhD degree in Belarusian State University in System Analysis, Control and Information Processing and works currently as leading scientific researcher of Research Laboratory of Applied Probabilistic Analysis in Belarusian State University. His main fields of interests are queueing systems with correlated arrival flows and controlled tandem models. He published the monograph and over 100 papers including more than 70 papers cited in Scopus.

OLGA DUDINA was graduated from Belarusian State University in 2007. In 2010, she earned PhD degree in Belarusian State University in Probability Theory and Mathematical Statistics. Her PhD dissertation got Award as the best dissertation of the year in Republic of Belarus in Natural Sciences. She works as leading scientific researcher of Research Laboratory of Applied Probabilistic Analysis in Belarusian State University. Her main fields of interests are queueing tandem queueing models with correlated arrival flows, non-markovian queueing systems. She published the monograph and over 100 papers including more than 65 papers cited in Scopus.