Priority tandem queueing model with admission control

Chesoong Kim, Sergey Dudin

Abstract

A two-stage multi-server tandem queue with two types of processed customers is analyzed. The input is described by the Marked Markovian Arrival Process (MMAP). The first stage has an infinite number of servers while the second stage has a finite number of servers. The service time at the both stages has an exponential distribution. Priority customers are always admitted to the system. Non-priority customers are admitted to the system only if the number of busy servers at the second stage does not exceed some pre-assigned threshold. Queueing system's behavior is described in terms of the multi-dimensional asymptotically quasi-Toeplitz continuous time Markov chain. It allows to exploit a numerically stable algorithm for calculation of the stationary distribution of the queueing system. The loss probability at the both stages of the tandem is computed. An economic criterion of the system operation is optimized with respect to the threshold. The effect of control on the main performance measures of the system is numerically demonstrated.

1. Introduction

Many problems in routing, flow control, bandwidth allocation and memory management in telecommunication networks, capacity planning and equipment order in manufacturing systems, optimal managing operation of emergency services, etc. can be solved with help of queueing theory which is an important branch of applied probability theory and operations research. Tandem queues represent the important bridge between the theory of queueing systems and theory of queueing networks. The theory of tandem queues is developed rather well, for references see, e.g., the fundamental book Gnedenko and Konig (1983) and surveys Balsamo, Persone, and Inverardi (2003), Perros (1989). In classical tandem queues, servers of a tandem work independently of each other. After the completion of the service at one stage of the tandem, a customer moves to the next stage. Some kind of dependency occurs only in the systems with a finite intermediate buffers with blocking the server operation in the case when the buffer at the next stage of a tandem is full or in the systems with feedback where a customer can return for the service at the previous stages. Last years, the models where the operation of the different stages of a tandem depends somehow on the situation at another stages where investigated, e.g., the models with flexible servers, see, e.g., Ahn, Duenyas, and Lewis (2002), models with the moving server, see, e.g., Grassmann and Tavakoli (2002), Irvani, Posner, and Bazacott (1997), models with coupled processors, see, e.g., Leeuwaarden and Resing (2005), Resing and Ormeci (2003), etc. In our paper we also consider the model when operation of the stages is dependent, namely we assume that the admission of customers at the first stage depends on the number of busy servers at the second stage.

The overwhelming majority of results in the theory of tandem queues is obtained for the case when the arrivals are described by a stationary Poisson process while it is widely recognized now that the flows of information in the modern communication and manufacturing networks cannot be well approximated by such a process because they are bursty and correlated. Direction of the investigation of tandem queues where the arrivals are described by the Markovian Arrival Process (MAP), which suits for description of correlated flows, is rather new, see papers Breuer, Dudin, Klimenok, and Tsarenkov (2004), Gomez-Corral (2002a, 2002b, 2002c, 2004), Klimenok, Breuer, Tsarenkov, and Dudin (2005, 2007) where several tandem queues with single server stages are dealt with. The multi-server tandem queueing system considered in this paper also assumes that the arrival process can be correlated. More, in contrast to all existing papers in tandem queues, we assume that the arrival process can be both correlated, i.e., inter-arrival times are dependent, and heterogeneous, i.e., there are two different types of customers arriving to the system. Note that we suggest here that customers are differentiated only at the arrival epoch. Both types of customers are treated equally after admitted to the system. This tandem system suits for modeling many real-world systems with heterogeneous customers, in which the service is provided in two stages. The service of a customer at
the first stage is a preliminary service, e.g., data base user authentication, car or patient diagnostics, half-finished product heating, etc. The service of a customer at the second stage is basic. It assumes data retrieval, car reparation, patient surgery, repousse performing, etc. The problem of optimal design and customers admission control in these systems can be solved based on algorithmic tool presented in this paper.

The considered model is also appropriate for modeling mobile communication networks. To attract potential subscribers, some providers of mobile communication service can offer several first (e.g., 5–15) seconds of a conversation free of charge. According to the government regulations in Russia since January, 2006, all providers have to provide at least 3 first seconds for free. Some providers allow more long periods of a free call, e.g., companies Elvis-Telecom, Trustintel (Russia) provide 6 free seconds, companies Bitel and Mobitel (Kyrgyzstan) provide 5 free seconds, company Babilon-M (Tadjikistan) and voice over IP mobile phone provider Busta (UK) provide 10 free seconds, etc. Of course, such a marketing solution can attract potential users of the network, because calls become cheaper or free of charge at all. But at the same time it can create essential problems to the provider. For example, some clients make excessive use of possibility to have a free conversation. Just before the free time expires, they stop conversation and make a new call. So, they do not pay for their calls at all and create problems for access to the network of other clients who are ready to pay but meet the busy servers. The problem of enhancement of the system operation in this situation (the problem of choosing optimal duration of a free conversation time in particular) is very complicated. One of the possible ways for solving this problem, which is offered in Kim, Klimenok, and Dudin (2006), is to split, physically or virtually, the process of customers’ service into two-stages (the free phase of the call is the first stage and the conversation, which is paid by a client, is the second phase) and to control the access to the first stage depending on the current situation at the second stage. The shortcoming of the solution presented in Kim et al. (2006) is the lack of the subscribers differentiation. In this paper, this shortcoming is partially overcome and the model with two classes of customers is investigated.

The model under study describes also operation of mobile networks without providing free service. The first system of a tandem consists of two sequential multi-server systems. We assume that it is finite, say \( N \), and the number of the servers at the first stage is infinite. According to the recent results presented in Leskela and Resing (2007), the infinite intermediate buffer is allowed and the problem of optimal design and customers admission control in these systems can be solved based on algorithmic tool presented in this paper.

The customers arrive to the system according to the MMAP (Marked Markovian Arrival Process). The notion of the MMAP and its detailed description are given, e.g., in He (1996). The MMAP is the essential generalization of the MAP (Markovian Arrival Process). Such an arrival process was introduced as a versatile Markovian point process (VMPP) by Neuts in the 70s. The original development of the VMPP contained extensive notations; however these notations were simplified greatly in Lucantoni (1991) and ever since this process bears the name Markovian Arrival Process (MAP). The class of MMAPs includes many input flows considered previously, such as stationary Poisson (\( M \)), Erlangian (\( E_r \)), Hyper-Markovian (\( HM \)), Phase-Type (\( PH \)), Markov Modulated Poisson Process (MMPP). Generally speaking, the MAP is correlated, so it is ideal to model correlated and bursty traffic in modern telecommunication networks. For more information about the MAP and related research see, e.g., Chakravarthy (2001), Lucantoni (1991). The main distinction of the MMAP comparing to the MAP is that the customers in the MAP are homogeneous, having the same requirements to the service process. Customers in the MMAP can be heterogeneous and having different types.

The arrival of customers in the MMAP are directed by the random process \( v_r,t \geq 0 \). The process \( v_r,t \geq 0 \), is an irreducible continuous time Markov chain with the state space \( \{0,1,\ldots,W\} \). The sojourn time of this chain in state \( v \) is exponentially distributed with the positive finite parameter \( \lambda_v \). When the sojourn time in the state \( v \) expires, with probability \( p^{(v)}_{\nu} \) the process \( v \) jumps to the state \( v' \) without generation of customers, \( v,v'=\emptyset,\emptyset \neq v' \), and with probability \( p^{(v)}_{\nu} \), the process \( v \) jumps to the state \( v' \) with generation of a customer of type \( r \). \( r=1,2, v'=\emptyset,\emptyset \neq v' \).

The notation \( v'=\emptyset \) means that the parameter \( v \) takes the values in the set \( \{0,1,\ldots,W\} \). The behavior of the MMAP is completely characterized by the matrices \( D_0,\bar{D}_1^{(1)},\bar{D}_1^{(2)},r=1,2 \), defined by their entries \( (\bar{D}_1^{(r)})_{v,v'}=\lambda_v p^{(v)}_{\nu} \). The process \( v \) jumps to the state \( v' \) without generation of customers, \( v=\emptyset,\emptyset \neq v' \), and with probability \( p^{(v)}_{\nu} \), the process \( v \) jumps to the state \( v' \) with generation of a customer of type \( r \). \( r=1,2, v'=\emptyset,\emptyset \neq v' \).

The matrix \( D(1)=D_0+\bar{D}_1^{(1)}+\bar{D}_1^{(2)} \) represents the generator of the process \( v_r,t \geq 0 \). The average arrival rate \( \lambda \) is defined by

\[
\lambda = \theta(D(1)^{(1)} + D(1)^{(2)}),
\]

where \( \theta \) is the invariant vector of a stationary distribution of the Markov chain \( v_r,t \geq 0 \). The vector \( \theta \) is the unique solution to the system

\[
\theta D(1) = 0, \quad \theta e = 1.
\]
Here $e$ is a column-vector of appropriate size consisting of 1’s and $0$ is a row-vector of appropriate size consisting of zeroes.

The average arrival rate $\lambda^{(r)}$ of the type $r$ customers is defined by $\lambda^{(r)} = \theta D^{(r)} e$, $r = 1, 2$. The squared coefficient of variation $\nu^{(r)}$ of inter-arrival times for the type $r$ customers is given by

$$\nu^{(r)} = \frac{2\theta (-D_0 - D^{(r)} - 1)}{\lambda^{(r)}} e - \left( \frac{1}{\lambda^{(r)}} \right)^2 \tilde{r}, \quad \tilde{r} \neq r, \quad \tilde{r}, \quad r = 1, 2.$$

The coefficient of correlation $C^{(r)}$ of two successive intervals between type $r$ customer arrivals is computed by

$$C^{(r)} = \frac{\theta (-D_0 + D^{(r)} - 1)}{\lambda^{(r)}} D^{(r)} (D_0 + D^{(r)} - 1) e - \left( \frac{1}{\lambda^{(r)}} \right)^2 \nu^{(r)} e - \left( \frac{1}{\lambda^{(r)}} \right)^2, \quad \tilde{r} \neq r.$$

We assume that type two customers are priority customers and they are always admitted into the system upon arrival. The admission of non-priority type 1 customers to the system is controlled by means of the parameter $M, M = \sum \lambda / \mu$. If at the epoch of a customer arrival to the first stage of tandem the number of customers at the second stage is less than the threshold $M$, then the customer is admitted to the system and starts the service at the first stage. Otherwise, it leaves the system without service (is lost). In the practical situations in telecommunications it means that the customer’s call is sent to some overflow route or is dropped.

After the admission of customers of both types to the system, service procedures for all customers are the same.

After completing the service at the first stage, the customer leaves the system forever with probability $q, 0 \leq q < 1$. With complementary probability $1 - q$, the customer moves for the service at the second stage of tandem. If all servers of this stage are busy, the customer leaves the system (is lost). Otherwise, it is processed by a server at the second stage and then leaves the tandem system.

We do not have an intention to prove that the considered here threshold admission strategy is optimal in some more wide class of strategies, say, class of all Markovian strategies. Such a strategy seems quite reasonable and easy implementable in real-world systems. Its effectiveness will be demonstrated below by means of the numerical examples.

We assume that the quality of the system operation is evaluated by means of the following cost criterion (average reward per unit time):

$$J(M) = a e^{\text{out}} - c_1 e^{P_{\text{loss}}^{(1)}} - c_2 e^{P_{\text{loss}}^{(2)}} - d(N - M + 1), \quad (1)$$

where $e^{\text{out}}$ is the rate of the flow of the customers, which get successful service at both stages of tandem, $a$ is the average profit obtained by the system from the service of one customer, $P_{\text{loss}}^{(r)}$ is the loss probability of a customer who wishes to get the service at the $r$th stage of the system, $c_r$ is the charge of the system when a customer is lost at the $r$th stage of the system, $r = 1, 2$, $d$ is the charge paid for reservation of one server at the second stage per time unit. This cost criterion is quite natural, e.g., in the case when the queueing system models the behavior of the mobile communication network described in the introduction. The main goal of the management by the system is to reach, by means of the proper choice of the threshold $M$, the maximal throughput of the system, which is characterized by rate $e^{\text{out}}$. At the same time, evidently the system manager should try to minimize probability of customers loss to fit the imposed requirements to the Quality of Service in the system. Probabilities of a customer loss at the first stage $P_{\text{loss}}^{(1)}$ (blocking probability) and at the second stage $P_{\text{loss}}^{(2)}$ (dropping probability) in the cost criterion may have different weights because it is usually assumed in mobile communication networks that the dropping of a call (interruption of a connection) is more essential disadvantage than blocking the call (rejection to establish a connection).

The problem of maximizing of the cost criterion (1) is not trivial because the increase of the threshold $M$ evidently causes the increase of the rate $e^{\text{out}}$ and decrease of the probability $P_{\text{loss}}^{(1)}$. But simultaneously the increase of $M$ can cause the sharp increase of the probability $P_{\text{loss}}^{(2)}$. So, the parameter $M$ should be carefully chosen depending on the system parameters and costs.

Note that the main contribution of the analysis presented in this paper is the algorithm for computing the key performance measures of the system under any fixed value of the threshold $M$. Once these measures will be computed, we can solve the problem of the optimal choice of the threshold $M$ for a variety of cost criteria, not only for criterion in the form (1). E.g., in the last experiment in 4 we will use another criterion which assumes that the loss of a high priority customer is more undesirable than the loss of a low priority customer.

It is worth to note also, that in the presented analysis we assume that all parameters of the system, namely, the number $N$ of the servers at the second stage, the matrices $D^{(k)}$, $k = 1, 2, r = 1, 2$, defining the arrival pattern, the service intensities $\mu_{k} k = 1, 2$, are known in advance. Of course, these parameters should be first identified for application of our model to some real-world system optimization. If our model will be applied in the logical or technical design of a new system, these parameters can come from the estimations of the experts, from requirements to the system parameters and characteristics, and from the known statistics for the analogous systems. If our model should be applied to enhance the operation of the existing system, the parameters can come from the statistical inference based on observation of the arrival and service process.

Note also that, although the service intensities $\mu_{k} k = 1, 2$, are assumed to be fixed, actually the numerical examples in the corresponding section below illustrate an effect of variation of these intensities. So, the dual problem of the optimal choice of these intensities can be solved based on our results. The service intensities $\mu_{k} k = 1, 2$, can be changed in real communication networks by means of the direct restriction of the service time as well as by means of managing the appropriate tariff policies.

The problem of the suitable choice of the cost coefficients (in our case coefficients $a, c_1, c_2, d$) in the cost criterion always play crucial role in successful implementation of optimization. We assume here that in our model the cost coefficients can come from the experts in the real-world object to which the model will be applied. Again, our numerical results illustrate the impact of the cost coefficients $c_1$ and $c_2$. So, these results can be useful for selecting the optimal values of these costs.

We can mention also that if there is no idea how to choose the cost coefficients $c_1$ and $c_2$, the problem of optimization can be easily modified as follows. Instead of criterion (1) we will maximize the throughput of the system under constraints imposed on probabilities of a customer loss at the first and the second stage of a tandem queue. E.g., on the basis of ITU – T requirements for land mobile services, see, ITU (1996), the maximal acceptable values for probabilities $P_{\text{loss}}^{(1)}$ (the new call blocking probability) and $P_{\text{loss}}^{(2)}$ (the call dropping probability) can be chosen as $10^{-3}$ and $5 \times 10^{-4}$, respectively.

3. Stationary distribution of the system states

Our goal is to calculate the main performance measures of this tandem system and to show the possibility of maximization of cost criterion (1) via an appropriate choice of the threshold $M$.

To this end, we consider the three-dimensional process

$$\tilde{z}_t = \{i_t, k_t, v_t\}, \quad t \geq 0,$

with the state space

$$i_t \geq 0, \quad k_t \equiv 0, N, \quad v_t \equiv 0, W.$$
where \( i_0 \) is the number of customers at the first stage, \( k_i \) is the number of customers at the second stage, \( v_i \) is the state of the directing process of the arrival process at the moment \( t \).

It is obvious that the process \( \xi_t, t \geq 0 \), is an irreducible regular continuous time Markov chain.

Let us enumerate the states of this Markov chain in the lexicographic order and refer to \((i,k,v)\) as the macro-state consisting of \((W+1)\) states \((i,k,v), v=0,\overline{W}\).

Introduce the following notation:

- \( \overline{W}=W+1, K=(N+1)(W+1), \)
- \( E_t=\begin{pmatrix} I_t & O_{(N+1-i)\times \overline{W}} \\ O_{(N+1-i)\times (N+1-j)} & O_{(N+1-j)\times \overline{W}} \end{pmatrix}, \quad I_t=0, N, \)
- \( E=\begin{pmatrix} 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix}, \)
- \( \tilde{E}=\begin{pmatrix} 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & -1 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & N-N \end{pmatrix} \)
- \( I_k \) is an identity matrix of dimension \( R \),
- \( Q_{a,b} \) means a zero matrix having \( a \) rows and \( b \) columns,
- \( \otimes \) is the symbol of Kronecker product of matrices, see, e.g., Graham (1981).

Let \( Q \) be the generator of the Markov chain \( \xi_t, t \geq 0 \), with blocks \( Q_{i,j} \) consisting of the intensities \((Q_{i,j})_{(i,k),j} \) of the Markov chain \( \xi_t, t \geq 0 \), transitions from the macro-state \((i,k)\) to the macro-state \((j,k')\), \( k', k=0,K \). The diagonal entries of the matrix \( Q_{i,j} \) are negative and the modulus of the diagonal entry \((Q_{i,j})_{i,j} \) defines the total intensity of leaving the corresponding state \((i,k,v)\) of the Markov chain. The block \( Q_{i,j}, i,j=0,K, \) has the dimension \( K \times K \).

**Theorem 1.** The generator \( Q \) of the Markov chain \( \xi_t, t \geq 0 \), has the three block diagonal structure \( Q=(Q_{i,j})_{i,j=0} \).

Non-zero blocks \( Q_{i,j} \) have the following form:

\[
Q_{i,j}=\begin{pmatrix} q_{i} \mu_{i} E_{i} + \mu_{1} (I_{N-1} - E_{N}) + (1-q_{i}) \mu_{i} E^{t} & \otimes I_{\overline{W}} \\ I_{(N+1-i)\times \overline{W}} & O_{(N+1-j)\times \overline{W}} \end{pmatrix}, \quad i \geq 1, \\
Q_{i,j}=\begin{pmatrix} E_{i} \otimes (D_{i}^{(1)} + D_{i}^{(2)}) + (I_{N-1} - E_{i}) \otimes D_{i}^{(2)} & I_{(N+1-i)\times \overline{W}} \\ I_{(N+1-i)\times \overline{W}} & O_{(N+1-j)\times \overline{W}} \end{pmatrix}, \quad i \geq 0.
\]

A proof of the theorem is implemented by means of the analysis of all transitions of the three-dimensional Markov chain \( \xi_t, t \geq 0 \), during the interval of an infinitesimal length and rewriting the generator of the chain in the block matrix form.

Let us denote

\[
\pi(i,k,v)=\lim_{t\to\infty} P[i_t=i, k_t=k, v_t=v], \quad i \geq 0, \quad k=0,K, \quad v=0,\overline{W},
\]

stationary probabilities of the Markov chain \( \xi_t, t \geq 0 \). It can be shown that, due to the customers loss at the second stage and the infinite number of servers at the first stage, the stationary probabilities \( \pi(i,k,v) \) exist for any choice of the system parameters. Let us enumerate probabilities \( \pi(i,k,v) \) in the lexicographic order and form row-vectors \( \pi \) of these probabilities corresponding to the value \( i \) of the first component of the Markov chain, \( i \geq 0 \).

It is well-known that the probability vectors \( \pi \) satisfy the following system of linear algebraic equations: \( (\pi_0, \pi_1, \pi_2, \ldots) Q=0, \)

\[
(\pi_0, \pi_1, \pi_2, \ldots, e)=1 \text{ which are called the equilibrium equations.}
\]

This system of equations is infinite. The problem of its solving is non-trivial because some coefficients in the truncated equations have modulus greater than the coefficients of the finite system which is obtained as the result of a truncation. Fortunately, the class of continuous time multi-dimensional asymptotically quasi-Toeplitz Markov chains (AQTCM) was recently introduced in Klimenok and Dudin (2006), where ergodicity conditions and the effective and numerically stable algorithm for computing the vectors \( \pi_i, i \geq 0 \), of stationary probabilities for the AQTCM are developed. The algorithm was derived in Klimenok and Dudin (2006) based on consideration of the jump Markov chain, see, Asmussen (2003), and substitution of the equilibrium equations by another system of equations by means of constructing the family of the censored Markov chains, see, e.g., Kemeni, Snell, and Knapp (1966).

It can be verified that the Markov chain \( \xi_t, t \geq 0 \), describing the tandem queue under study belongs to the class of AQTCM. So, we can use results from Klimenok and Dudin (2006) for calculation of the stationary probability vectors \( \pi, i \geq 0 \).

The algorithm presented in Klimenok and Dudin (2006) does not require three block diagonal structure of the generator while we have such a structure in the generator defined in Theorem 1. It allows us to simplify the algorithm from Klimenok and Dudin (2006) for application to Markov chain \( \xi_t, t \geq 0 \). The scheme of the simplified algorithm is given in the following statement.

**Theorem 2.** The probability vectors \( \pi_i, i \geq 0 \), are computed as follows:

\[
\pi_i=\pi_0 F_i, \quad i \geq 1,
\]

where the non-negative matrices \( F_i, i \geq 1 \), are calculated using the formulae:

\[
F_0=I_{N+1}, \quad F_i=\prod_{l=1}^{i} Q_{i+1,l}[-(Q_{i+2,l}+Q_{i+1,l}G_i)]^{-1}, \quad i \geq 1,
\]

the vector \( \pi_0 \) is computed as the unique solution of the system

\[
\pi_0(Q_{0,0}+Q_{0,1}G_0)=0, \quad \pi_0 \sum_{i=0}^{\infty} F_i e=1,
\]

where the matrices \( G_i \) are computed using recursion:

\[
G_i=[-(Q_{i+1,i}+Q_{i+2,i}G_{i+1})]^{-1}Q_{i+1,i}, \quad i \geq 0.
\]

**Theorem 2** defines the simple algorithmic way for calculation of the stationary distribution \( \pi_i, i \geq 0 \), of the Markov chain \( \xi_t, t \geq 0 \). Numerical stability of this algorithm is explained by the fact that all involved recurrent calculations do not contain a subtraction operation. The only difficulty in realization of the algorithm is the following. Recursion (5) is the backward one. So, we need to know the matrix \( G_{\infty} \) to start calculations. It follows from the definition of the AQTCM and Theorem 1 that the sequence of the matrices \( G_{\infty} \) tends, when \( i \) approaches infinity, to the matrix \( G \) where
Eventually we arrive to the following procedure for solving the backward recursion (5). Some value $i_0$ is assumed being sufficiently great. The matrices $G_n$, $i > i_0$ are assumed to be equal to the matrix $G$ which is defined by formula (6). The matrices $G_n$, $i = i_0 - 1$, $i_0 - 2, \ldots, 0$ are calculated recursively from (5). As a measure of control that the chosen $i_0$ is actually sufficiently great, the norm of the matrix $G_n - G$ or $G_n - G_{i_0-1}$ can be computed and compared with some pre-assigned accuracy level, e.g., with so called computer epsilon. So, the problem of computing the vectors $\pi_i$, $i > 0$, can be considered solved.

Having these vectors been computed, we can calculate the main performance measures of the tandem system including the measures involved into the cost criterion $J(M)$:

- The probability $P^{(1)}_{\text{loss}}$ that an arbitrary customer will be lost at the first stage of the system is computed by

$$P^{(1)}_{\text{loss}} = \frac{1}{\lambda} \sum_{i=0}^{\infty} \pi_i (1 - \lambda \Delta_t) \otimes D_i^{(1)} \mathbf{e}.$$ 

Sure, this customer can be only type 1 customer.

- The probability $P^{(1)}_{\text{res}}$ that an arbitrary type 1 customer will be lost at the first stage of the system is computed by

$$P^{(1)}_{\text{res}} = \frac{1}{\lambda} \sum_{i=0}^{\infty} \pi_i (1 - \lambda \Delta_t) \otimes D_i^{(1)} \mathbf{e}.$$ 

- The average number $N^{(1)}$ of busy servers at the first stage is computed by

$$N^{(1)} = \sum_{i=0}^{\infty} \pi_i \mathbf{1}.$$ 

- The intensity $\lambda^{(1)}_{\text{out}}$ of flow of customers, which get the service at the first stage of the system, is computed by

$$\lambda^{(1)}_{\text{out}} = N^{(1)} \mu_1.$$ 

- The average number $N^{(2)}$ of busy servers at the second stage of the system is computed by

$$N^{(2)} = \sum_{i=0}^{\infty} \pi_i (\mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots, \mathbf{N})^t.$$ 

- The intensity $\lambda^{(2)}_{\text{out}}$ of flow of customers, which get the service at the second stage of the system, is computed by

$$\lambda^{(2)}_{\text{out}} = N^{(2)} \mu_2.$$ 

- The probability $P^{(2)}_{\text{loss}}$ that an arbitrary customer will be lost at the second stage of the system (it leaves the system after the service at the first stage due to the occupancy of all servers at the second stage) is computed by

$$P^{(2)}_{\text{loss}} = 1 - \frac{\lambda^{(2)}_{\text{out}}}{\lambda^{(2)}_{\text{out}} (1 - q)}.$$ 

- The intensity $\gamma_1$ of the flow of non-priority customers, which get the service at the second stage of the system, is computed by

$$\gamma_1 = (\lambda^{(1)}_{\text{out}} - \lambda^{(2)}_{\text{out}}) (1 - q) (1 - P^{(2)}_{\text{loss}}).$$ 

- The intensity $\gamma_2$ of the flow of priority customers, which get the service at the second stage of the system, is computed by

$$\gamma_2 = \lambda^{(2)} (1 - q) (1 - P^{(2)}_{\text{loss}}).$$ 

- The intensity $\beta^{(1)}_1$ of the flow of non-priority customers, which are rejected at the second stage of the tandem, is computed by

$$\beta^{(1)}_1 = \lambda^{(1)} - \lambda^{(1)}_{\text{res}}.$$ 

- The intensity $\beta^{(2)}_2$ of the flow of priority customers, which are rejected at the second stage of the tandem, is computed by

$$\beta^{(2)}_2 = \lambda^{(2)} (1 - q) P^{(2)}_{\text{res}}.$$ 

4. Optimization problem and numerical examples

As it was mentioned above, our goal is to find the value $M^*$ of the threshold $M$, $0 < M < N$, which provides the maximal value of the cost criterion (1). Analytical results presented in the previous section allow to compute the value of the cost criterion (1) for any fixed value of $M$, $0 < M < N$. So, the problem of finding the optimal value of $M$ can be easily solved, e.g., by means of enumeration.

The goals of the following five numerical experiments is to show effectiveness of the control and dependencies of the optimal value of the cost criterion (1) on different parameters of the system and cost coefficients. The sixth experiment is made for the modified cost criterion which additionally differentiates the losses of high and low priority customers and the profit obtained from the service of high and low priority customers.

In the first experiment we illustrate the necessity to take carefully into consideration the correlation in the arrival process instead of approximation of this process in terms of the stationary Poisson arrival process.

Let the number $N$ of servers at the second stage be given by 30. Intensities of service at the first and the second stages are equal to $\mu_1 = 30$ and $\mu_2 = 0.08$, respectively. The probability $q$ is equal to 0.3. Let the profit obtained by the system from the service of one customer is $a = 200$, the charge of the system when a customer is lost at the first stage of the system $c_1 = 20$, the charge of the system when a customer is lost at the second stage of the system $c_2 = 200$, the charge paid for reservation of one server at the second stage per unit of time is equal to zero, i.e., $d = 0$.

Let us consider the MMAP which is defined by the matrices $D_0$, $D_1^{(1)}$ and $D_2^{(1)}$ be given by

$$D_0 = \begin{pmatrix} 20 & 0 \\ 0.008 & 0.558 \end{pmatrix}; \quad D_1^{(1)} = \begin{pmatrix} 5 & 0.12 \\ 0.04 & 0.32 \end{pmatrix};$$

$$D_2^{(1)} = \begin{pmatrix} 5 & 0.17 \end{pmatrix}.$$ 

This MMAP has coefficient of correlation between the successive inter-arrival times equal to 0.4, and the coefficient of variation $c^{2\text{out}} = 10.5$. The average arrival rate $\lambda \approx 5.1$. The average intensity of non-priority customers arrival $\lambda_1 = 3.6$. The average arrival rate of the priority customers $\lambda_2 \approx 1.5$.

We will compare the dependence of the cost criterion $J(M)$ on the threshold $M$ in the model having this MMAP as arrival process with the dependence of the cost criterion $J(M)$ on the threshold $M$ in the model where arrivals of non-priority customers are described by the Poisson process with rate $\lambda_1 = 3.6$ and arrivals of priority customers are described by the Poisson process with rate $\lambda_2 = 1.5$. These arrival processes have coefficient of correlation between the successive inter-arrival times equal to 0 and coefficient of variation equal to 1.

Fig. 1 presents the dependence of the cost criterion $J(M)$ on the threshold $M$ for MMAP and Poisson processes of priority and non-priority customers. Optimal values of $M$ for these curves are equal to 20 and 27 respectively. Optimal values of the cost criterion are equal to 94.5 for the MMAP and 366 for the Poisson process while the values of the cost criterion when no channel reservation is made at the second stage (i.e., $M = N = 30$) are equal.
to –61 and 270, correspondingly. Although arrival rates of the MMAP and the stationary Poisson processes coincide, one can see that the dependencies of the cost criterion on the threshold M are quite different. This figure evidently confirms that the optimal channel reservation at the second stage does make sense and gives essential profit. Another conclusion from this figure is that the correlation in the input flow essentially impacts the value of the optimal threshold and the optimal value of the cost criterion. The assumption that the arrival processes are the Poisson processes implies too optimistic results and can lead for incorrect servers reservation and decreasing of the possible profit or even increasing the loss.

In the second experiment we illustrate the dependence of the cost criterion (1) on the threshold M and parameters \( q, \mu_1, \mu_2 \). Let the number \( N \) of servers at the second stage be given by 20. The matrices \( D_0, D_1^{(1)} \) and \( D_2^{(1)} \), which define the MMAP, are given by

\[
D_0 = \begin{pmatrix} -16.46 & 0 \\ 0 & -16.51 \end{pmatrix}; \quad D_1^{(1)} = \begin{pmatrix} 6.51 & 6.53 \\ 6.53 & 6.48 \end{pmatrix}; \quad D_2^{(1)} = \begin{pmatrix} 1.75 & 1.67 \\ 1.7 & 1.8 \end{pmatrix}.
\]

The average arrival rate \( \lambda \approx 16.5 \). The average intensity of non-priority customers arrival \( \lambda_2 \approx 13 \). The average arrival rate of priority customers \( \lambda_1 \approx 3.5 \). The cost coefficients are the same as in the previous experiment.

Fig. 2 illustrates the shape of the dependence of the cost criterion on the threshold M and probability q, variable in the interval \( q \in [0,0.99] \).

It follows from Fig. 2, that the maximum of the cost criterion \( J(M) \) is equal to 886, when the value of the threshold \( M = 19 \) and \( q = 0.6 \). The dependence of the difference \( J(M) = J(M^*) - J(N) \) between the optimal value \( J(M^*) \) of the cost criterion and the value of the cost criterion \( J(N) \) of the system without channel reservation on the probability \( q \) is presented on Fig. 3. One can see, that the optimal channel reservation can ensure the essential profit.

Let us now fix the intensity of service at the second stage \( \mu_2 = 15 \) and let probability \( q \) be equal to 0.3. Fig. 4 shows the dependence of the cost criterion \( J(M) \) on the threshold M and intensity of service at the first stage \( \mu_1 \) variable in the interval \( \mu_1 \in [2,20] \). Intuitively, one might argue that increasing of value \( \mu_1 \) leads to the increasing of the cost criterion. Another conclusion from this figure is that the value of the cost criterion (1) is very sensitive to the value of the threshold M for any value of the service intensity \( \mu_1 \).

Now, we fix intensity of service at the first stage \( \mu_1 = 100 \). Fig. 5 shows the dependence of the cost criterion \( J(M) \) on the threshold M and the intensity \( \mu_2 \) of service at the second stage variable in the interval \( \mu_2 \in [0.2,1] \). One can see that the value of the intensity \( \mu_2 \) essentially effects the value of the cost criterion.

In the third experiment we illustrate the dependence of the optimal value \( J(M^*) \) of the cost criterion and the dependence of the optimal value \( M^* \) of the threshold on system parameters and cost coefficients. Firstly, we show the dependence of the optimal cost criterion \( J(M^*) \) on service intensities \( \mu_1 \) and \( \mu_2 \) variable in
the interval \( \mu_1 \in [2,30], \) and \( \mu_2 \in [0.2,1] \). Other parameters are the same as in the previous experiment. This dependence is given on Fig. 6.

The optimal value \( J(M^*) \) of the cost criterion is equal to 2277 when \( \mu_1 = 30 \) and \( \mu_2 = 1 \). The increase of the service intensities evidently implies the grow of the optimal value of the cost criterion (1).

Fig. 7 illustrates the dependence of the optimal cost criterion \( J(M^*) \) on the number \( N \) of servers at the second stage variable in the interval \([1,20]\) and on the intensity \( \lambda_2 \) of the priority flow when both flows are described by the stationary Poisson arrival process. The intensity \( \lambda_1 \) is assumed to be equal to 1.

The maximal value of the optimal cost criterion \( J(M^*) \) is equal to 549.65. It is achieved when \( N = 20 \) and \( \lambda_2 = 4 \). The monotone increasing of \( J(M^*) \) when \( N \) increases is evident while the existence of the optimal value of \( \lambda_2 \) is quite interesting fact.

Fig. 8 illustrates the dependence of the optimal value \( M^* \) of the threshold on the service intensities \( \mu_1 \) and \( \mu_2 \). The rest of the system parameters are the same as in the first experiment.

The dependence of the optimal value \( M^* \) on both the service intensities \( \mu_1 \) and \( \mu_2 \) looks almost monotone. However, this monotony is violated for small values of \( \mu_1 \) and big values of \( \mu_2 \).

Fast service at the second stage in combination with slow service of customers at the first stage (and slow customers arrival from the first stage to the second one) makes reservation of servers meaningless \( (M^* = N = 20) \). However, when \( \mu_1 \) increases reservation of one server at the second stage \( (M^* = N - 1 = 19) \) is required for the best operation of the system.

Figs. 9 and 10 illustrate the dependence of the optimal value \( M^* \) of the threshold on the cost coefficients \( c_1 \) and \( c_2 \). The service intensities \( \mu_1 \) and \( \mu_2 \) are assumed to be equal to 2 and 0.2 correspondingly. Fig. 9 gives the dependence for the \( \text{MMAP} \) defined in the first experiment. Recall that this \( \text{MMAP} \) has the coefficient of correlation between the successive inter-arrival times equal to 0.4.

Fig. 10 gives this dependence for another \( \text{MMAP} \), which has the same average rate of both types of customers arrival but the coefficient of correlation equal to –0.1. This \( \text{MMAP} \) is defined by the matrices

\[
D_0 = \begin{pmatrix} -3.76 & 0 \\ 0 & -8 \end{pmatrix}, \quad D_1^{(1)} = \begin{pmatrix} 0 & 2.93 \\ 5 & 0 \end{pmatrix}, \quad D_1^{(2)} = \begin{pmatrix} 0 & 0.83 \\ 3 & 0 \end{pmatrix}
\]

One can conclude that the variation of the value of \( M^* \) is more essential (and, consequently, intuitive solution of the optimization problem is more uncertain) and less monotone in the case of the positive correlation in the arrival process.

In the fourth experiment we consider the following problem. As it was mentioned in Introduction, the tandem queuing system under study can be applied for parameters tuning in mobile communication networks where, to attract potential users of a network, several first seconds of conversation are provided free of charge. Let us assume that the number of non-priority customers of the system and their activity are strongly correlated with the length of the service interval provided for free. The average service time at the first stage of tandem is equal to \( \mu_1^{-1} \). Let us assume that the intensity \( \lambda_1 \) of non-priority customers arrival is defined by the relation

\[
\lambda_1 = h\mu_1^{-1},
\]

where \( h \) is some parameter which can be evaluated from experimental data or Gallup polls.

Let the number of servers \( N \) be equal to 20, \( q = 0.3 \), the intensity of priority customers arrival, which is assumed being insensitive with respect to the length of the interval provided for free, \( \lambda_2 = 1.5 \), intensity of the second stage duration \( \mu_2 = 0.5 \). Fig. 11 rep-
resents the dependence of the optimal cost criterion \( J(M) \) on the intensity \( l_1 \) for three different values of the parameter \( h \).

One can observe that the value of the cost criterion \( J(M) \) quickly increases with the increase of the intensity \( l_1 \) (it is explained by the fact that very small value of \( l_1 \) implies too high intensity of non-priority customers what, in turn, causes high loss probability for both types of customers), then it reaches the maximum and then it quickly decreases (it is explained by the decrease of the intensity of non-priority customers arrival and decrease of the throughput of the system). So, the optimal value of the cost criterion \( J(M) \) is very sensitive with respect to the parameter \( l_1 \). Note that the optimal value of the cost criterion is almost the same for different values of the parameter \( h \), but the corresponding optimal values of the parameter \( l_1 \) are quite different. The importance of results presented in this paper stems from our ability to predict the optimal value of the parameter \( l_1 \), which provides the maximum to the cost criterion \( J(M) \), under any fixed set of system and cost parameters.

\[ \text{Fig. 8. Dependence of } M^* \text{ on the service intensities } \mu_1 \text{ and } \mu_2. \]

\[ \text{Fig. 9. Dependence of } M^* \text{ on the cost coefficients } c_1 \text{ and } c_2 \text{ in the case of the positive correlation in the arrival process.} \]

\[ \text{Fig. 10. Dependence of } M^* \text{ on the cost coefficients } c_1 \text{ and } c_2 \text{ in the case of the negative correlation in the arrival process.} \]

\[ \text{Fig. 11. Dependence of } J(M^*) \text{ on the intensity } l_1, \text{ which is inverse to the length of the interval provided for free, for different values of the parameter } h. \]
In the fifth experiment we consider the problem of choosing the minimal number \( N^* \) of servers at the second stage which is sufficient to provide (under the optimal choice of the threshold \( M \)) the value of the loss probability at the second stage less than the pre-assigned number \( e \). Fig. 12 presents the dependence of the number \( N^* \) on the service intensity \( \mu_2 \) for three different values of the \( e \). It is quite evident that the minimal number \( N^* \) of servers decreases when the service rate increases. However, only the results of calculations can help to answer the question about the optimal choice of the number \( N^* \) under any fixed set of system and cost parameters and the parameter \( e \) which characterizes the tolerance of the system with respect to the customers loss at the second stage.

Figs. 13 and 14 present the dependence of the number \( N^* \) on the probability \( q \) and intensity \( \lambda_2 \) of priority customers arrival correspondingly.

In this paper, we concentrate on the problem of customers admission control and try to provide better conditions for high priority customers by means of restriction of low priority customers admission when the number of idle servers at the second stage of the system is small. After the customers are admitted to the system, they are not differentiated. This is a reasonable politics and namely this politics is considered here.

If one would like to provide additional privilege for high priority customers, e.g., by means of reservation of a part of the servers at the second stage only for these customers, he should separately observe the number of high priority and low priority customers currently processed at the first stage of the system. This leads to necessity of consideration of the multi-dimensional Markov chain having two denumerable components. Such the chains hardly can be investigated analytically. Technique applied in our paper could be used for such a model if we restrict by some finite bound the number of non-priority customers which can be simultaneously processed at the first stage. We do not consider such a system here.

However, the presented above results can be easily adjusted to the situation when we implicitly provide additional privilege for high priority customers by differentiating the system charges when high and low priority customers are lost. Such a differentiation leads to the suitable change of the threshold \( M \). In criterion (1), we distinguished only the losses of customers at the first and second stages. Let us consider another criterion which additionally differentiates the losses of high and low priority customers and the profit obtained from the service of high and low priority customers, namely, criterion

\[
J(M) = a_1\gamma_1 + a_2\gamma_2 - c^{(1)}_1\beta^{(1)}_1 - c^{(2)}_2\beta^{(2)}_2 - c_3\beta_2,
\]

where \( a_1 \) is the average profit obtained by the system from the service of one non-priority customer, \( \gamma_1 \) is the intensity of flow of non-priority customers, which get the service at the second stage of the system, \( a_2 \) is the average profit obtained by the system from the service of one priority customer, \( \gamma_2 \) is the intensity of flow of priority customers, which get the service at the second stage of the system, \( c^{(1)}_1 \) is the charge of the system when a non-priority customer is lost at the \( r \)-th stage of the system, \( r = 1, 2 \), \( \beta^{(1)}_1 \) is the intensity of flow of non-priority customers, which are rejected at the \( r \)-th stage of the tandem, \( r = 1, 2 \), \( c^{(2)}_2 \) is the charge of the system when a priority customer is lost at the second stage of the tandem, \( \beta_2 \) is the intensity of flow of priority customers, which are rejected at the second stage of the tandem.

In the sixth experiment we illustrate the dependence of the new cost criterion on the threshold \( M \). We take the same parame-
ters of the model as in the first experiment except we set the service intensity at the second stage $\mu_2 = 0.1$. We fix the following values of the cost coefficients:

$$a_1 = 100, \quad a_2 = 300, \quad c_1^{(1)} = 20, \quad c_1^{(2)} = 100, \quad c_2 = 300,$$

i.e., we assume that the cost of a priority customer loss is three times higher than the cost of a non-priority customer loss at the second stage.

Fig. 15 shows the dependence of the new cost criterion on the threshold $M$ for the stationary Poisson and the MMAP arrival processes. Again, we see the different behavior of the two curves although the corresponding rates of priority and non-priority customers coincide. The optimal value of the threshold in the case of the stationary Poisson arrival process is equal to 28 and the optimal value of the cost criterion is equal to 440.3. The optimal value of the threshold in the case of the MMAP arrival process is equal to 22 and the optimal value of the cost criterion is equal to 271.1. The values of the cost criterion when there is no channel reservation at all (i.e., $M = N = 30$) are equal to 403 and 167.3, correspondingly.

The relative profit, which gives the optimal admission strategy comparing to the system without channels reservation is equal to 9.25% and 62%, respectively.

One can see that the profit is quite good, especially in the case of the MMAP which is the bursty and correlated arrival process typical for the modern telecommunication networks.

5. Conclusion

A multi-server priority tandem queuing system with admission control is considered. The stationary state distribution and performance characteristics are calculated under the fixed threshold strategy. The dependence of the cost criterion, which includes the rate of the flow of the customers that get successful service at the both stages of tandem, the profit obtained by the system from the service of one customer, the loss probability of a customer who wishes to get the service at each stage of the system, the charge of the system when a customer is lost at each stage of the system, and the charge paid for reservation of one server at the second stage per time unit, is derived. The cost criterion optimization problem is solved numerically. Numerical results giving insight into the system behavior are presented. In particular, these results evidently demonstrate high effectiveness of the considered strategy of control.

Acknowledgements

This work was supported by the Korea Research Foundation Grant Funded by the Korean Government (MOEHRD) (KRF-2006-21-D00569). This work was also supported by Grant No. R01-2006-000-10668-0 from the Basic Research Program of the Korea Science and Engineering Foundation.

References


