Queueing Models and
Service Management
Vol. 1, No. 1, page 001-016, 2018

QMSM


# Single Server, Multi-Class Queues with Markovian Arrival and Ambiguity of Class Determination

A. Krishnamoorthy[1,*], V. Vishnevsky[2], Dhanya Shajin[3] and A. S. Manjunath[4]

[1]Department of Mathematics
CMS College, Kottayam-686001, India.

[2]V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences
65 Profsoyuznaya Street, Moscow-117997, Russia.

[3]Department of Mathematics
Amrita School of Arts and Sciences
Amrita University, Kochi-682024, India.

[4]Department of Mathematics
Government Victoria College Palakkad-678001, India.

*(Received April 2017; accepted October 2017)*

**Abstract:** In this paper we introduce a queueing system manned by a single server who serves multiple class $(n)$ of customers. The customers arrive according to a Markovian arrival process and form a single queue. At the time when taken for service the customer of class $i$ may be taken for service of class $j$ (ambiguity in the determination of class of required service) with probability $p_{ij}$, $1 \le j \le n$. Service time in class $i$ is of phase type distributed with representation $(\gamma_i^{(i)}, T_i^{(i)})$ of order $m_i^{(i)}$, $1 \le i \le n$. If a customer of class $i$ is taken for class $j$ service initially then on completion of service there he is taken to class with probability $\eta_{jk}^{(i)}$, $1 \le k \le n$. A timer starts at the beginning of service of a customer. If the timer realizes before the customer is identified of the required (correct) service, this customer is instantly sent out of the system without getting correct service. On the other hand if the timer does not realize before identification of required service, then the customer is taken for service in that class and completes service successfully. In the case when the customer is provided the required class of service right from the very beginning, he leaves the system after completing this service. In the last case timer plays no role in the service time of such customers. In the case of such customers no ambiguity arises on the type (class) of service required. We analyze the above system to derive the expected time a customer spends with the server. Then we use it to derive the stability condition and the resulting system state distribution. Useful performance indices are computed. Numerical illustrations are provided to have a glimpse of the system performances. Some examples from real life situation are cited, as motivation for the study of the above mentioned model. Case of arbitrarily distributed service time is also considered. An application in telecommunication is indicated.

**Keywords**: Ambiguity on class determination, desired/undesired services, Markovian arrival process, multi-class queue, random clock (timer).

* Corresponding author
Email : achyuthacusat@gmail.com

## 1. Introduction

So far queueing theory literature concentrated on single / multi-server problems with the service requirement of customers exactly known. In most of the cases it was the same type of service that was rendered and so uncertainty did not arise (see for example, Gross and Harris [5]). Even when cases of different types of services were considered the exact requirement of type of service was assumed known. In multi - server cases, the servers may serve at different rates and hence only independence of service time duration, but not identical service distributions, were assumed.

For the classical multi-class queues one may refer to Atar et al. [1], Harrison and Zeevi [6], Kelly [7], Righter [11], and Sharif et al. [12]. None of these discuss ambiguity in determination of class of service to be offered. In retrial queueing context, Avarchenkov et al. [2] and Falin [4] consider multi-class customers. These authors assume that the exact class to which each customer belongs, is known to the customer as well as service system and hence no ambiguity on the nature of service arose.

Ambiguity in determination of service needed of a customer in a multi-class queue (whether single server or multi server / single channel or distinct lines) is quite common. However, that ambiguity has not so far been discussed in literature. There are several real life situations where neither the customer(s) nor the server(s) would be aware of the exact requirement of service the customer needs. This leads ***either*** the customer being served rendered handicaped (incapable of receiving the right service required) since undesired service was already provided for too long a time ***or*** the customer initially receives undesired service and after a while moves to required service, thus escapes getting incapacitated. In this paper we consider a single server queue providing $n$ distinct services with uncertainty in the type of service required by the multi - class customers.

We can have examples from medical services and also from repair facilities for the model under consideration. The patients queueing up at a physician's clinic is an apt situation of the model. The patient(s) tells the physician of the symptoms and the latter makes certain inferences. However, visible symptoms (or even certain tests) may not reveal the ailment, with the result that a wrong diagnosis is arrived at. The resulting course of medication could harm the patient beyond repair and the patient may turn unfit for any further medication. On the other hand, even when a wrong diagnosis is made and the patient is accordingly provided medication, at a later stage the correct diagnosis could be arrived, before the patient becomes unfit for the right medication. This case leads to the patient getting cured completely. In the other extreme case, the patient is diagnosed correctly at the very beginning of service and so get right medication.

Another example arises in telecommunication (protocol IEEE 802.11DCF). When a massage originates, the server could be idle and hence could be transmitted immediately. On the other hand if the server is busy at the time when the message originates, then it goes through a sequence of contention windows. This part we call the undesirable service. If the process of going through contention windows does not end up with meeting the server idle before the timer expires, then the message losses its significance and hence discarded. This example is explained in detail in Section 5 (special case).

Now we describe the occurrence of ambiguity into our service system. Assume that $n$ distinct services are offered by the server. Neither the server nor the customers are exactly aware of the service requirement of the latter. Introduce the probability vector $(p_1,\ldots,p_i,\ldots,p_n)$ in which the $i^{th}$ component $p_i$ stands for probability that a customer belongs to type $i$ when taken for service, $i=1,2,\ldots,n$. However, he may be diagnosed as requiring the $j^{th}$ type service with probability $p_{ij}$, $1\le j\le n$. Thus a customer, conditioned on requiring $i^{th}$ service, starts getting type $j$ service with probability $p_{ij}$. Obviously $p_{ii}$ is the probability of being taken for service of type $i$ under condition of service requirement of type $i$. Define the matrix $\mathsf{P}=\left(p_{ij}\right)_{i,j=1,2,\ldots,n}$ where $(i,j)^{th}$ entry $p_{ij}$ is as defined above. We also introduce another transition probability matrix $\Gamma^{(i)}=\left(\eta_{jk}^{(i)}\right)$ which is an $n\times n$ matrix governing transitions among undesired service types until finally the customer in service turns out to be not responsive to further service (realization of a timer) consequent to which he quits without getting required service or after some amount of time in undesired type of service, the customer transits to the desired service (before realization of the timer). In the latter case, the customer gets the desired service and then leaves the system. Thus we have a random clock (timer) governing service process if the customer starts getting service in undesired type:

a) The realization of this timer before identification of correct type of service required to the customer, renders the customer unresponsive to further service.
b) If timer does not realize until the customer is diagnosed of the exact service requirement, then he leaves the system completing the desired service.
c) If the customer directly gets into desired service (correct diagnosis right at the beginning of his service) then the timer does not play any role in his service.

The service time in different types of service are independent PH distributed random variables. The clock time (timer) has exponential distribution.

We analyze the above described model to extract condition for system stability and several useful system performance characteristics.

**NOTE:** In the sequel we interchangeably use desired / correct / required service; undesired / incorrect service; class / type of service.

***List of notations and abbreviations used***

- $CTMC$ : Continuous time Markov chain.
- $LIQBD$: Level independent quasi-birth and death process.
- $MAP$ : Markovian arrival process.
- $I$ : Identity matrix of appropriate order.
- $O$: Zero matrix of appropriate order.
- $0$ : Column vector of $0^{'}$s with appropriate order.
- $e$: Column vector of $1^{'}$s with appropriate order.

First we compute the response time (expected time spent in service - leaving the system without getting desired service / undesired to desired and complete service / right at the beginning get desired service and leave the system).

This paper is arranged as follows. In Section 2 the response time of a customer (leaving the system without getting required service / start with undesired service and before timer realization moves to required service / start at the very beginning of service in required class and leave the system) is derived. Having done that, in Section 3 we investigate the stability of the system. For the stable system we derive state distribution of the system. Performance indices of the system are given in Section 4. Some special cases are considered in Section 5. Performance measures are numerically illustrated in Section 6. We consider different inter-arrival time distributions to investigate numerically their effect on the mean number of customers in the system. Finally we conclude the paper indicating plan of future extension of the model described.

## 2. Response Time

Consider the continuous time Markov chain $\{S(t), t \geq 0\} = \{(N_1(t), N_2(t), N_3(t), N_4(t)), t \geq 0\}$ where

$$N_1(t): \quad \textit{required type of service at time } t$$

$$N_2(t): \quad \textit{mode of service at time } t$$

$$= \begin{cases} 1 & \textit{directly in correct service} \\ 2 & \textit{in incorrect service} \\ 3 & \textit{in correct service, beginning with incorrect type} \end{cases}$$

$$N_3(t): \quad \textit{type of service being provide at time } t$$

$$N_4(t): \quad \textit{phase of service at time } t$$

The Markov chain $\{S(t), t \geq 0\}$ has the state space

$$\{(i,1,j,k), 1 \leq i \leq n; j = i; 1 \leq k \leq m_i^{(i)}\} \bigcup$$
$$\{(i,2,j,k), 1 \leq i \leq n; 1 \leq j \neq i \leq n; 1 \leq k \leq m_j^{(i)}\} \bigcup$$
$$\{(i,3,j,k), 1 \leq i \leq n; j = i; 1 \leq k \leq m_i^{(i)}\}.$$

- The probability that a customer selected for service requires the $i^{th}$ type of service is $p_i$ so that $p_1 + p_2 + ... + p_n = 1$.
- A customer is selected for service $j$ but his desired (required) service is $i$, with probability $p_{ij}$ and $\sum_{j=1}^{n} p_{ij} = 1$. Thus we have the initial service chosen according to the entries in the matrix

$$= \begin{pmatrix} p_{11} & \cdots & p_{1i} & \cdots & p_{1n} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ p_{i1} & \cdots & p_{ii} & \cdots & p_{in} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ p_{n1} & \cdots & p_{ni} & \cdots & p_{nn} \end{pmatrix}.$$

- A random threshold clock (timer) is set, which follows exponential distribution with mean rate $\zeta_i$ when required service is $i$, so that the customer is pushed out of the system if the clock expires before service completion in undesired (incorrect) service.
- For $i = 1,2, ..., n, (\boldsymbol{\gamma}_i^{(i)}, T_i^{(i)})$ of order $m_i^{(i)}$ gives the PH-representation for the duration of the required service time distribution when the service of a customer starts in the required type (when it is assumed to be $i$). Let $T_i^{0(i)}$ be such that $T_i^{(i)} e + T_i^{0(i)} = 0$. Let $\mu_i' = -\boldsymbol{\gamma}_i^{(i)} (T_i^{(i)})^{-1} e$ be the mean of this PH-representation.
- $(\boldsymbol{\xi}^{(i)}, U_i)$ of order $\sum_{\substack{j=1 \\ j \neq i}}^{n} m_j^{(i)}$ gives the PH-representation for the duration of the incorrect service time distribution when the service of a customer starts in incorrect service class, given that type $i$ is the required service. The $j^{th}$ component of the initial probability vector $\boldsymbol{\xi}^{(i)} = \left(\beta_j^{(i)} \boldsymbol{\gamma}_j^{(i)}\right), 1 \leq j \leq n, j \neq i$, is the probability that a customer enters to incorrect type of service $j$ when actual requirement is $i$ where $\sum_{\substack{j=1 \\ j \neq i}}^{n} p_{ij} = 1 - p_{ii}$, $\sum_{\substack{j=1 \\ j \neq i}}^{n} \beta_j^{(i)} = 1$, and $\boldsymbol{\gamma}_j^{(i)} e = 1, 1 \leq j \neq i \leq n$. While in service in undesired type the rate vector of loss is given by $U_i^{0^*}$ and the rate vector of getting into correct service mode is given by $U_i^0$. Note that $U_i e + U_i^0 + U_i^{0^*} = 0$. Denote by $\hat{\mu}_{inc}^{(i)} = -\boldsymbol{\xi}^{(i)} (U_i)^{-1} e$ the mean of this PH-representation; then the corresponding rate is $\mu_{inc}^{(i)} = 1/\hat{\mu}_{inc}^{(i)}$. The transition matrix $U_i$ is of the form

$$U_i = \begin{pmatrix} T_1^{(i)} - \zeta_i I & T_{12}^{(i)} & \cdots & T_{1\,i-1}^{(i)} & T_{1\,i+1}^{(i)} & \cdots & T_{1\,n}^{(i)} \\ T_{21}^{(i)} & T_2^{(i)} - \zeta_i I & \cdots & T_{2\,i-1}^{(i)} & T_{2\,i+1}^{(i)} & \cdots & T_{2\,n}^{(i)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ T_{i-1\,1}^{(i)} & T_{i-1\,2}^{(i)} & \cdots & T_{i-1}^{(i)} - \zeta_i I & T_{i-1\,i+1}^{(i)} & \cdots & T_{i-1\,n}^{(i)} \\ T_{i+1\,1}^{(i)} & T_{i+1\,2}^{(i)} & \cdots & T_{i+1\,i-1}^{(i)} & T_{i+1}^{(i)} - \zeta_i I & \cdots & T_{i+1\,n}^{(i)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ T_{n1}^{(i)} & T_{n\,2}^{(i)} & \cdots & T_{n\,i-1}^{(i)} & T_{n\,i+1}^{(i)} & \cdots & T_n^{(i)} - \zeta_i I \end{pmatrix}.$$

having absorption rates

$$U_i^0 = \begin{pmatrix} T_{1i}^{(i)} \\ T_{2i}^{(i)} \\ \vdots \\ T_{i-1\,i}^{(i)} \\ T_{i+1\,i}^{(i)} \\ \vdots \\ T_{n\,i}^{(i)} \end{pmatrix}, \; U_i^{0^*} = \begin{pmatrix} \zeta e \\ \zeta e \\ \vdots \\ \zeta e \\ \zeta e \\ \vdots \\ \zeta e \end{pmatrix},$$

where $T_{j\,k}^{(i)} = \eta_{jk}^{(i)} T_j^{0(i)} \boldsymbol{\gamma}_k^{(i)}$, $k \neq i$ and $T_{j\,i}^{(i)} = \eta_{ii}^{(i)} T_j^{0(i)}$. Note that $T_j^{(i)} e + T_j^{0(i)} = 0$.

$(\boldsymbol{\gamma}_j^{(i)}, T_j^{(i)})$ of order $m_j^{(i)}$ gives the PH-representation for the duration of the service time distribution when the service of a customer starts in $j^{th}$, $1 \leq j \neq i \leq n$, type of service when $i$ is the type of service required. In the last row we write the last element as 1 to indicate that if service of type $i$ is offered at some instant of time, then it does not go to any other type other than leave the system on completion of that service. After service completion in the $j^{th}$, $1 \leq j(\neq i) \leq n$ service mode he goes to $k^{th}$, $1 \leq k(\neq j) \leq n$ service mode with probability $\eta_{jk}^{(i)}$. It may be noted that $j$ could be visited at a future time.

$$\Gamma^{(i)} = \begin{pmatrix} 0 & \eta_{12}^{(i)} & \cdots & \eta_{1i-1}^{(i)} & \eta_{1i+1}^{(i)} & \cdots & \eta_{1n}^{(i)} & \eta_{1i}^{(i)} \\ \eta_{21}^{(i)} & 0 & \cdots & \eta_{2i-1}^{(i)} & \eta_{2i+1}^{(i)} & \cdots & \eta_{2n}^{(i)} & \eta_{2i}^{(i)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \eta_{i-11}^{(i)} & \eta_{i-12}^{(i)} & \cdots & 0 & \eta_{i-1i+1}^{(i)} & \cdots & \eta_{i-1n}^{(i)} & \eta_{i-1i}^{(i)} \\ \eta_{i+11}^{(i)} & \eta_{i+12}^{(i)} & \cdots & \eta_{i+1i-1}^{(i)} & 0 & \cdots & \eta_{i+1n}^{(i)} & \eta_{i+1i}^{(i)} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \eta_{n1}^{(i)} & \eta_{n2}^{(i)} & \cdots & \eta_{ni-1}^{(i)} & \eta_{ni+1}^{(i)} & \cdots & 0 & \eta_{ni}^{(i)} \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

We have the following results whose proofs are omitted.

**Lemma 2.1.** *The probability that a customer who started with incorrect service leave the system without getting desired service, assumed to be* $i$*, due to clock realization, is* $p_{inc}^{clock} = \xi^{(i)}(-U_i)^{-1} U_i^{0^*}$.

**Lemma 2.2.** *The probability that a customer who started with undesired service subsequently moves to desired service (assumed to be* $i$*) before clock realization, is* $p_{inc}^{correct} = \xi^{(i)}(-U_i)^{-1} U_i^0$.

Note that by assumption the probability of starting in correct class (assumed to be $i$) and so leave system on completion of this service is PH distributed with representation $(\gamma_i^{(i)}, T_i^{(i)})$ of order $m_i^{(i)}$. From the above discussions we have the following

**Lemma 2.3.** *The duration of the service time of a customer whose desired service is* $i$*, can be modeled as a PH-distribution with representation* $(\boldsymbol{\alpha}_i, S_i)$ *of order* $m^{(i)} = 2m_i^{(i)} + \sum_{\substack{j=1 \\ j \neq i}}^{n} m_j^{(i)}$*, where*

$$\boldsymbol{\alpha}_i = (\boldsymbol{\alpha}_1^{(i)}, \boldsymbol{\alpha}_2^{(i)}, 0), \tag{1}$$

and

$$S_i = \begin{pmatrix} T_i & 0 & 0 \\ 0 & U_i & U_i^0 \otimes \boldsymbol{\gamma}^{(i)} \\ 0 & 0 & T_i \end{pmatrix}, \tag{2}$$

with

$$\boldsymbol{\alpha}_1^{(i)} = p_{ii}\boldsymbol{\gamma}^{(i)}, \boldsymbol{\alpha}_2^{(i)} = (1-p_{ii})\boldsymbol{\xi}^{(i)}.$$

Let $S_i^0$ be such that $S_i e + S_i^0 = 0$ where $S_i^0$ is given by $S_i^0 = \begin{pmatrix} T_i^{0(i)} \\ U_i^{0^*} \\ T_i^{0(i)} \end{pmatrix}$. Lemma 2.1 to 2.3 lead to

**Theorem 2.4.** *The unconditional service time (assuming any one of the $n$ service is the required one) of a customer can be modeled as a PH-distribution with representation $(\boldsymbol{\alpha}, S)$ of order $m = \sum_{i=1}^{n} m^{(i)}$, where*

$$\boldsymbol{\alpha} = (p_1\boldsymbol{\alpha}_1, p_2\boldsymbol{\alpha}_2, \ldots, p_n\boldsymbol{\alpha}_n), \tag{3}$$

and

$$S = \begin{pmatrix} S_1 & & & \\ & S_2 & & \\ & & \ddots & \\ & & & S_n \end{pmatrix}, \tag{4}$$

Let $S^0$ be such that $Se + S^0 = 0$ and $S^0 = \begin{pmatrix} S_1^0 \\ S_2^0 \\ \vdots \\ S_n^0 \end{pmatrix}$. Thus the expected service time of a customer is $E_{st} = -\boldsymbol{\alpha} S^{-1} e$. Now we consider the response time of a customer, conditioned on his required service type is $i$, for $1 \le i \le n$.

**Lemma 2.5.** *The response time of a customer who started with undesired service and leave the system without getting desired service due to clock realization is* $E_{T_1}^{(i)} = \left( \dfrac{\zeta_i}{\zeta_i + \mu_{inc}^{(i)}} \right) \xi^{(i)} (-U_i)^{-1} e.$

**Lemma 2.6.** *The response time of a customer who started with incorrect service and subsequently moved to desired service before realization of the timer is*

$$E_{T_2}^{(i)} = \left( \frac{\mu_{inc}^{(i)}}{\zeta_i + \mu_{inc}^{(i)}} \right) \xi^{(i)} (-U_i)^{-1} e + \gamma_i^{(i)} (-T_i^{(i)})^{-1} e.$$

**Lemma 2.7.** *The response time of a customer who was selected for the desired service at the very beginning of his service is* $E_{T_3}^{(i)} = \gamma_i^{(i)} (-T_i^{(i)})^{-1} e.$

Combining Lemma 2.5, 2.6 and 2.7, we have the following theorem.

**Theorem 2.8.** *The response time of a customer when his desired service is $i$, is*

$$E_T^{(i)} = (1-p_{ii}) \left\lfloor E_{T_1}^{(i)} + E_{T_2}^{(i)} \right\rfloor + p_{ii} E_{T_3}^{(i)}.$$

**Corollary 2.9.** *The unconditional response time of a customer in the system is* $E_T = \sum_{i=1}^{n} p_i E_T^{(i)}$. Now we go for the complete model description and its mathematical analysis.

## 3. Mathematical Model

Customers arrive to the single server system according to a *MAP* (Markovian arrival process) to form a single waiting line. In a *MAP* (see Chakravarthy [3]), the customers arrival is directed by an irreducible *CTMC* (continuous time Markov chain) $\{\phi(t), t \geq 0\}$ with the state space $\{1,2,...,r\}$. The transition intensities of the Markov chain $\{\phi(t), t \geq 0\}$ which are accompanied by arrival of $k$ customers, are described by the matrices $D_k, k = 0,1$. Vector $\boldsymbol{\eta}$ of the stationary distribution of the process $\{\phi(t), t \geq 0\}$ is the unique solution to the system of equations

$$\boldsymbol{\eta}(D_0 + D_1) = \boldsymbol{\eta} D = 0 \text{ and } \boldsymbol{\eta} e = 1 \tag{5}$$

The fundamental rate $\lambda$ of the $MAP(D_0, D_1)$ described above, is given by $\lambda = \boldsymbol{\eta} D_1 e$.

*The ambiguity in identification of required service does not arise from forming a single waiting line by customers of all classes. Rather this is due to ignorance of the customers and / server to identify the class to which each customer belongs.*

Service time in each type of service follows phase type distribution. The service time distribution in different types are different. A timer with exponentially distributed duration starts the moment the service of a customer begins. Further details of the service time process were given in Section 2.

Let $N(t)$ be the number of customers in the system, $S(t)$ the phase of service (as defined in Section 2) and $A(t)$ the phase of arrival process at time $t$. With these the process $\{(N(t), S(t), A(t)), t \geq 0\}$ is a continuous time Markov chain with state space

$$\{(0,a) : 1 \leq a \leq r\} \bigcup \{(\ell, b, a) : \ell \geq 1, 1 \leq b \leq m, 1 \leq a \leq r\}.$$

Thus the infinitesimal generator of this *CTMC* is an *LIQBD* and is of the form

$$Q = \begin{pmatrix} D_0 & A_{01} & & & \\ A_{10} & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{6}$$

where $A_{01} = \alpha \otimes D_1, A_{10} = S^0 \otimes I_r, A_0 = I_m \otimes D_1, A_1 = S \oplus D_0, A_2 = S^0 \alpha \otimes I_r$.

We are now in a position to investigate the stability of the system.

### *3.1. Stability condition*

Consider $A (= A_0 + A_1 + A_2)$, the generator matrix of the Markov chain corresponding to the phase changes. Then

$$A = (S + S^0 \alpha) \oplus D, \tag{7}$$

Let $\pi = (\pi_1, \pi_2, ..., \pi_n)$ be the steady-state probability vector of $(S + S^0 \alpha)$. Then

$$\pi(S + S^0 \alpha) = 0 \text{ and } \pi e = 1. \tag{8}$$

From the relation $\pi(S + S^0 \alpha) = 0$ we have

$$\pi_i S_i + p_i \left[ \sum_{j=1}^{n} \pi_j S_j^0 \right] \alpha_i = 0, \ 1 \leq i \leq n. \tag{9}$$

Multiplying (9) by $e$ on right hand side when $i = 1$, we get

$$\sum_{j=2}^{n}\pi_j S_j^0 = \frac{1-p_1}{p_1}\pi_1 S_1^0. \tag{10}$$

Putting (10) in (9) when $i = 2,3,...,n$, yields

$$\pi_i S_i^0 = -\frac{p_i}{p_1}\pi_1 S_1 e,\ 2 \le i \le n. \tag{11}$$

Substituting (11) in (9) when $i = 1$, gives

$$\pi_1(S_1 + S_1^0\alpha_1) = 0. \tag{12}$$

This implies that for an arbitrary constant $c$

$$\pi_1 = c\alpha_1(-S_1)^{-1}. \tag{13}$$

Substituting for $\pi_1$ in (11), we have

$$\pi_i = c\frac{p_i}{p_1}\alpha_i(-S_i)^{-1},\ 2 \le i \le n. \tag{14}$$

From the normalizing condition $\pi e = 1$, the value of $c$ is computed as

$$c = \left[\theta_1' + \sum_{j=2}^{n}\frac{p_j}{p_1}\theta_j'\right]^{-1}, \tag{15}$$

where $\theta_i' = \alpha_i(-S_i)^{-1}e,\ 1 \le i \le n.$

Now from (5) and (8) we get the steady state probability vector of $A$ as

$$\hat{\phi} = \pi \otimes \boldsymbol{\eta}.$$

**Theorem 3.1.** *The stability of the system is given by the relation*

$$\hat{\phi}(I_m \otimes D_1)e < \hat{\phi}(S^0\alpha \otimes I_r)e. \tag{16}$$

**Proof**. The queueing system under study with the $LIQBD$ type generator given in (6), is stable if and only if (see Neuts [10])

$$\hat{\phi}A_0 e < \hat{\phi}A_2 e. \tag{17}$$

This amounts to saying that the left drift rate is higher than the rate of drift to the right.

The left drift rate is $\hat{\phi}(I_m \otimes D_1)e$ and the right drift rate is $\hat{\phi}(S^0\alpha \otimes I_r)e$ (the fundamental arrival rate less than the reciprocal of the mean response time).

**NOTE:** In terms of response time that we evaluated in the previous section, we can also write the stability condition as $\lambda < \frac{1}{E_T}$.

***3.2. Steady-state probability vector***

A brief outline for the computation of the stationary probability vector of the system state is as follows. Let $\mathbf{x}$ denote the steady-state probability vector of the generator $Q$. Then

$$\mathbf{x}Q = 0 \ \text{ and } \ \mathbf{x}e = 1. \tag{18}$$

Assuming that the stability condition (16) holds and partitioning $\mathbf{x}$ as $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, ...)$, we obtain

$$\mathbf{x}_n = \mathbf{x}_1 R^{n-1},\ n \ge 1, \tag{19}$$

where $R$ is the minimal non negative solution to the matrix quadratic equation

$$R^2 A_2 + R A_1 + A_0 = O. \tag{20}$$

The two boundary equations involving $\mathbf{x}_0$ are

$$\mathbf{x}_0 D_0 + \mathbf{x}_1 A_{10} = 0, \tag{21}$$

$$\mathbf{x}_0 A_{01} + \mathbf{x}_1 [A_1 + R A_2] = 0. \tag{22}$$

From the normalizing condition we have

$$\mathbf{x}_0 [I + V(I-R)^{-1}] e = 1, \tag{23}$$

where

$$V = -A_{01} [A_1 + R A_2]^{-1}. \tag{24}$$

We have

$$\mathbf{x}_0 = \left(\mathbf{x}_0(a), 1 \le a \le r\right),$$

and

$$\begin{aligned} \mathbf{x}_\ell &= \left(\mathbf{x}_\ell(b,a), 1 \le b \le m, 1 \le a \le r\right) \\ &= \left(\mathbf{x}_\ell(i,1,j,k,a),\ 1 \le i \le n,\ j = i, 1 \le k \le m_i^{(i)}\right) \\ &\quad \bigcup \left(\mathbf{x}_\ell(i,2,j,k,a),\ 1 \le i \le n, 1 \le j \ne i \le n, 1 \le k \le m_j^{(i)}\right) \\ &\quad \bigcup \left(\mathbf{x}_\ell(i,3,j,k,a),\ 1 \le i \le n,\ j = i, 1 \le k \le m_i^{(i)}\right). \end{aligned}$$

Next we shall briefly discuss important performance characteristics of the system under study.

## 4. Performance Measures

- Probability that the server is idle:

$$P_0 = \sum_{a=1}^{r} \mathbf{x}_0(a)$$

- Probability that the server is busy with a customer who started service in desired type at the very beginning:

$$P_1 = \sum_{\ell=1}^{\infty} \sum_{i=1}^{n} \sum_{k=1}^{m_i^{(i)}} \sum_{a=1}^{r} \mathbf{x}_\ell(i,1,i,k,a)$$

- Probability that the server is busy in undesired type of service:

$$P_2 = \sum_{\ell=1}^{\infty} \sum_{i=1}^{n} \sum_{j=1, j\ne i}^{n} \sum_{k=1}^{m_j^{(i)}} \sum_{a=1}^{r} \mathbf{x}_\ell(i,2,j,k,a)$$

- Probability that the server is busy with the service of a customer in desired type, whose service started in an undesired type

$$P_3 = \sum_{\ell=1}^{\infty} \sum_{i=1}^{n} \sum_{k=1}^{m_i^{(i)}} \sum_{a=1}^{r} \mathbf{x}_\ell(i,3,i,k,a)$$

- Expected number of customers in the system is:

$$N_C = \sum_{\ell=1}^{\infty} \sum_{b=1}^{m} \sum_{a=1}^{r} \ell \mathbf{x}_\ell(b,a)$$

- Probability of customers leaving with desired service completion, starting in incorrect service type:

$$p_{cs} = \sum_{i=1}^{n} p_i (1 - p_{ii}) \xi^{(i)} (-U_i)^{-1} U_i^0$$

• Probability that a customer is lost (leaving the system without getting correct service):

$$p_{loss} = \sum_{i=1}^{n} p_i (1 - p_{ii}) \xi^{(i)} (-U_i)^{-1} U_i^{0^*}$$

• Rate at which customers leave successfully after being selected (at the very beginning) in correct service:

$$R_{CS} = \sum_{i=1}^{n} p_i p_{ii}$$

## 5. Special Case

Assume that $i^{th}$ is the required type of service of a customer. Suppose the service process is such that, if service started in some incorrect type $j \neq i$, which we denote by $j_1$ then from $j_1$ the customer progressively moves to $j_2$ then to $j_3$ and so on finally to $j_{n-1}$, all undesired for that customer and $j_1, j_2, ..., j_{n-1}$ are all of which are distinct and different from $i$. In between the timer may realize and the customer goes out or desired type of service is identified and the customer completes service before he leaves the system.

Under this assumption $(\boldsymbol{\xi}^{(i)}, U_i)$ of order $\sum_{k=1}^{n-1} m_{j_k}^{(i)}$ gives the PH-representation for the duration of the incorrect service time distribution when the service of a customer starts in incorrect service mode. The initial probability vector $\boldsymbol{\xi}^{(i)} = \left(\boldsymbol{\gamma}_{j_1}^{(i)}, 0, ..., 0\right)$ is the probability vector of a customer entering to undesired service whose actual requirement was $i$. The rate vector of loss is given by $U_i^{0^*}$ and the rate vector of getting into correct service mode is given by $U_i^0$. Note that $U_i e + U_i^0 + U_i^{0^*} = 0$. Let $\mu_{inc}^{'(i)} = -\boldsymbol{\xi}^{(i)} (U_i)^{-1} e$ be the mean of this PH-representation. The transition matrix $U_i$ is of the form

$$U_i = \begin{pmatrix} T_{j_1}^{(i)} - \zeta_i I & T_{j_1 j_2}^{(i)} & \cdots & T_{j_1 j_{n-2}}^{(i)} & T_{j_1 j_{n-1}}^{(i)} \\ & T_{j_2}^{(i)} - \zeta_i I & \cdots & T_{j_2 j_{n-2}}^{(i)} & T_{j_2 j_{n-1}}^{(i)} \\ & & \ddots & \vdots & \vdots \\ & & & T_{j_{n-2}}^{(i)} - \zeta_i I & T_{j_{n-2} j_{n-1}}^{(i)} \\ & & & & T_{j_{n-1}}^{(i)} - \zeta_i I \end{pmatrix},$$

with the two absorbing vectors

$$U_i^0 = \begin{pmatrix} T_{j_1 i}^{(i)} \\ T_{j_2 i}^{(i)} \\ \vdots \\ T_{j_{n-2} i}^{(i)} \\ T_{j_{n-1} i}^{(i)} \end{pmatrix}, U_i^{0^*} = \begin{pmatrix} \zeta_i e \\ \zeta_i e \\ \vdots \\ \zeta_i e \\ \zeta_i e \end{pmatrix},$$

to class $i$ (correct diagnosis of required service) or the random clock realization and

$$T_{j_\ell j_k}^{(i)} = \eta_{j_\ell j_k}^{(i)} T_{j_\ell}^{0(i)} \boldsymbol{\gamma}_{j_k}^{(i)}, \ and \ T_{j_\ell i}^{(i)} = \eta_{j_\ell i}^{(i)} T_{j_{ell}}^{0(i)}.$$

$(\boldsymbol{\gamma}_{j_k}^{(i)}, T_{j_k}^{(i)})$ of order $m_{j_k}^{(i)}$ gives the PH-representation for the duration of the service time distribution when the service of a customer starts in $j_k^{th}$ service mode. Let $T_{j_k}^{0(i)}$ be such that $T_{j_k}^{(i)}e + T_{j_k}^{0(i)} = 0.$

After service completion of the $j_\ell^{th}$ service mode he goes to $j_k^{th}, \ell+1 \le k \le n-1, 1 \le \ell \le n-2$ service mode with probability $\eta_{j_\ell j_k}^{(i)}$ and corresponding transition probability matrix is of the form

$$\Gamma^{(i)} = \begin{pmatrix} 0 & \eta_{j_1 j_2}^{(i)} & \eta_{j_1 j_3}^{(i)} & \cdots & \eta_{j_1 j_{n-2}}^{(i)} & \eta_{j_1 j_{n-1}}^{(i)} & \eta_{j_1 i}^{(i)} \\ & 0 & \eta_{j_2 j_3}^{(i)} & \cdots & \eta_{j_2 j_{n-2}}^{(i)} & \eta_{j_2 j_{n-1}}^{(i)} & \eta_{j_2 i}^{(i)} \\ & & 0 & \cdots & \eta_{j_3 j_{n-2}}^{(i)} & \eta_{j_3 j_{n-1}}^{(i)} & \eta_{j_3 i}^{(i)} \\ & & & \ddots & \vdots & \vdots & \vdots \\ & & & & 0 & \eta_{j_{n-2} j_{n-1}}^{(i)} & \eta_{j_{n-2} i}^{(i)} \\ & & & & & 0 & \eta_{j_{n-1} i}^{(i)} \\ & & & & & & 1 \end{pmatrix}.$$

If $i^{th}$ $(i = 1,2,...,n)$ is required type of service and that is the one initially chosen then the transition probability matrix $\mathsf{P}$ is of the form

$$\mathsf{P} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}.$$

Thus the response time distribution is $PH(\gamma_i^{(i)}, T_i^{(i)})$ for $1 \le i \le n$ when $i$ is the desired type of service. The entries 1 along the diagonal indicate that the class of service of such customers does not change; they leave the system on getting desired service right from the beginning of service.

For cases III and IV discussed below we define $\eta_{j_{i_\ell} j_{i_{\ell+1}}}$ as 1 and $g_{j_{i_{\ell+1}}}(u) = 1$ if transition $j_{i_\ell} \to j_{i_{\ell+1}}$ does not take place. Here $g_k(u)$ is the probability density function of service completion in the interval $(u, u+du)$ in class $k$.

### *5.1. Arbitrary distribution for service time*

Suppose $i^{th}$ is the required type of service $1 \le i \le n$, and the service time in each type is arbitrarily distributed.

(a) The first case we have is one in which after being served in one or more undesired classes, that customer is diagnosed correctly of the desired service and gets into it before timer realization. Note that all states $j_{i_1}, j_{i_2}, ..., j_{i_{n-1}}$ need not necessarily be visited by the customer.

$j_{i_1} \to j_{i_2} \to ... \to j_{i_k} \to i$ and $j_{i_1} < j_{i_2} < ... < j_{i_k}$

$$\int_0^t \sum_{\substack{k=1 \\ \forall \, permutations \, of \{1,2,...,i-1,i+1,...,n\}}}^{n-1} p_{ij_{i_1}} \eta_{j_{i_1} j_{i_2}} ... \eta_{j_{i_k} i} e^{-\xi u} (g_{j_{i_1}} * g_{j_{i_2}} * ... * g_{j_{i_k}} * g_i)(u)\, du \, .$$

(b) The second case deals with the scenario in which the required service is not diagnosed before timer realization. Thus the customer starts getting service in undesired type and then keep moving to other undesired types. The timer realizes before correct identification of the type of required service, thereby forcing the customer to exit from the system. This leads to: $j_{i_1} \to j_{i_2} \to ... \to j_{i_k} \not\to i$

$$\int_0^t \sum_{\substack{k=1 \\ \forall\, permutations\ of\, \{1,2,\ldots,i-1,i+1,\ldots,n\}}}^{n-1} p_{ij_{i_1}} \eta_{j_{i_1} j_{i_2}} \ldots \eta_{j_{i_{k-1}} j_{i_k}} (g_{j_{i_1}} * g_{j_{i_2}} * \ldots * g_{j_{i_k}})(u)\, \zeta\, e^{-\zeta u} du\,.$$

There are cases where customers may have a main service and thereafter none, one or more (atmost $n-1$) optional services where, in each state the service time is arbitrarily distributed independently of each other.

$g_i$ is the service time density in the essential service (if class $i$ is the essential service) and for optional services labelled $1,2,\ldots,i-1,i+1,\ldots,n$, the service time densities are $g_1,\ldots,g_{i-1},g_{i+1},\ldots,g_n$; after essential service, the customer chooses one of these as his initial optional service with probability $p_j, j=1,\ldots,i-1,i+1,\ldots,n$ and then goes to higher states according to a Markov chain rule $(p_{jk})_{(n-1)\times(n-1)}$ with $p_{jk}=0$ for $k \le j$ and $p_{jk}>0$ for $k>j$. In this case the total service time distribution is given by

$$\int_0^t \sum_{\substack{k=1 \\ \forall\, permutations of\, \{1,2,\ \ldots,i-1,i+1,\ldots,n\}}}^{n-1} p_{ij_{i_1}} \eta_{j_{i_1} j_{i_2}} \ldots \eta_{j_{i_{k-1}} j_{i_k}} (g_{j_{i_1}} * g_{j_{i_2}} * \ldots * g_{j_{i_k}})(u) du.$$

Madan [8] and Medhi [9] discuss a single server queue with a second optional service for each customer. In our case there are $n-1$ optional services; the customers can opt none, one, ..., upto $n-1$ optional services.

### *5.2. The model applied to telecommunication*

Suppose that when a message is generated, it starts service getting in undesired type. This can be interpreted as the message encountering a busy server at the time of its generation. Then it goes through contention windows checking at the end of each window whether the server is busy. If the server is found to be idle before timer realization successful transmission of the message takes place (desired service part).

On the other hand if timer realizes before the server is found idle, then the message does not get successful transmission and is discarded. The realization of timer before successful transmission of the message can be regarded as it being lost significance, especially when it is of emergency nature; so any further retransmission attempt is dropped. For certain protocols (eg. IEEE 802.11 DCF) the above description looks quite apt.

## 6. Numerical Illustrations

In this section we provide numerical illustration of the system performance with variation in values of underlying parameters.

We fix parameters $n=4, (p_1, p_2, p_3, p_4) = (0.1, 0.2, 0.3, 0.4)$,

$$\mathsf{P} = \begin{pmatrix} 0.7 & 0.1 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.1 & 0.2 \\ 0.1 & 0.2 & 0.6 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.7 \end{pmatrix},$$

$$\gamma_j^{(i)} = \begin{pmatrix} 1 & 0 \end{pmatrix}, T_j^{(i)} = \begin{pmatrix} -9.5i & 9.5i \\ 0 & -9.5i \end{pmatrix},\ for\ 1 \le i, j \le n$$

$$\eta_{12}^{(i)} = 0.3, \eta_{13}^{(i)} = 0.5, \eta_{14}^{(i)} = 0.3, \ for \ i = 2,3,4$$
$$\eta_{21}^{(i)} = 0.2, \eta_{23}^{(i)} = 0.5, \eta_{24}^{(i)} = 0.3, \ for \ i = 1,3,4$$
$$\eta_{31}^{(i)} = 0.2, \eta_{32}^{(i)} = 0.3, \eta_{34}^{(i)} = 0.5, \ for \ i = 1,2,4$$
$$\eta_{41}^{(i)} = 0.2, \eta_{42}^{(i)} = 0.3, \eta_{43}^{(i)} = 0.5, \ for \ i = 1,2,3$$
$$\beta^{(1)} = \begin{pmatrix}0.3 & 0.4 & 0.3\end{pmatrix}, \beta^{(2)} = \begin{pmatrix}0.2 & 0.2 & 0.6\end{pmatrix},$$
$$\beta^{(3)} = \begin{pmatrix}0.5 & 0.3 & 0.2\end{pmatrix}, \beta^{(4)} = \begin{pmatrix}0.8 & 0.1 & 0.1\end{pmatrix}$$

Write $\zeta_i = \zeta$, $1 \le i \le n$. The $MAP(D_0, D_1)$ is described by the matrices for $i$, $1 \le i \le i \le n$.

$$D_0 = \begin{pmatrix} -5.0111 & 5.0111 & 0 \\ 0 & -5.0111 & 0 \\ 0 & 0 & -1128.75 \end{pmatrix}, \ D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 4.96099 & 0 & 0.05011 \\ 11.2875 & 0 & 1117.4625 \end{pmatrix}.$$

### *6.1. Effect of $\zeta$ on the system performance*

We note from Table 1 that higher the value of $\zeta$, that is, higher rate of timer realization, higher is the probability of server being idle. The measure $P_1$ is not much affected by the timer realization rate since it gives probability of server busy with customer who directly gets into desired service. $P_2$ value decreases with increase in value of $\zeta$ as to be expected; so is the case with $P_3$ since timer realizes before identification of desired type of service, when $\zeta$ increases. The value of $p_{cs}$ decrease with increasing value of $\zeta$ since, starting with undesired type of service, identification of desired type service can not be realized with faster realization of the timer. Thus customer loss probability, without identification of desired service when started in undesired service type, increases with faster rate of realization of timer.

Table 1. Effect of $\zeta$.

| $\zeta$ | $P_0$ | $P_1$ | $P_2$ | $P_3$ | $P_{cs}$ | $P_{loss}$ |
|---|---|---|---|---|---|---|
| 0.1 | 0.1909 | 0.2267 | 0.4642 | 0.1182 | 0.3388 | 0.0112 |
| 0.2 | 0.2001 | 0.2289 | 0.4553 | 0.1156 | 0.3282 | 0.0218 |
| 0.3 | 0.2093 | 0.2310 | 0.4465 | 0.1132 | 0.3183 | 0.0317 |
| 0.4 | 0.2186 | 0.2328 | 0.4378 | 0.1107 | 0.3088 | 0.0412 |
| 0.5 | 0.2278 | 0.2345 | 0.4293 | 0.1083 | 0.2999 | 0.0501 |
| 0.6 | 0.2370 | 0.2360 | 0.4209 | 0.1060 | 0.2914 | 0.0586 |
| 0.7 | 0.2462 | 0.2374 | 0.4127 | 0.1037 | 0.2834 | 0.0666 |

### *6.2. Effect of arrival process*

In this section we investigate the effect of the timer on the number of customers in the system, with arrival process having the exponential / Erlang of order 2 / hyper exponential distributions (which are special cases of $MAP$. Thus for the arrival process, we consider the following five sets of values for $D_0$ and $D_1$ as follows.

**1. Exponential (** $EXP$ **)**

$$D_0 = (-5), \ D_1 = (5).$$

**2. Erlang (** $ERL$ **)**

$$D_0 = \begin{pmatrix} -10 & 10 \\ 0 & -10 \end{pmatrix},\ D_1 = \begin{pmatrix} 0 & 0 \\ 10 & 0 \end{pmatrix}.$$

**3. Hyper-exponential (** $HYP$ **)**

$$D_0 = \begin{pmatrix} -9.5 & 0 \\ 0 & -0.95 \end{pmatrix},\ D_1 = \begin{pmatrix} 8.55 & 0.95 \\ 0.855 & 0.095 \end{pmatrix}.$$

**4.** $MAP$ **with negative correlation (** $MAP^-$**)**

$$D_0 = \begin{pmatrix} -5.0111 & 5.0111 & 0 \\ 0 & -5.0111 & 0 \\ 0 & 0 & -1128.75 \end{pmatrix},\ D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0.05011 & 0 & 4.96099 \\ 1117.4625 & 0 & 11.2875 \end{pmatrix}.$$

**5.** $MAP$ **with positive correlation (** $MAP^+$ **)**

$$D_0 = \begin{pmatrix} -5.0111 & 5.0111 & 0 \\ 0 & -5.0111 & 0 \\ 0 & 0 & -1128.75 \end{pmatrix},\ D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 4.96099 & 0 & 0.05011 \\ 11.2875 & 0 & 1117.4625 \end{pmatrix}.$$

The above $MAP$ processes will be normalized so as to have a specific arrival rate. However, these are qualitatively different in that they have different variance and correlation structures. The first three arrival processes, namely, $EXP$, $ERL$ and $HYP$ have zero correlation for two successive inter-arrival times. The arrival processes labeled $MAP^-$ and $MAP^+$ respectively, have negative and positive correlation for two successive inter-arrival times with values -0.48891 and 0.48891. The standard deviation of the inter-arrival times of these five arrival processes are, respectively, 0.2, 0.1414, 0.4489, 0.2819 and 0.2819.

From Table 2 we note that for exponential, Erlang, hyper exponential and $MAP$ with negative correlation the expected number of customers in the system has considerably low values compared to that corresponding to $MAP$ with positive correlation. This discrepancy may be attributed to the fact that with positive correlation inter-arrival time gets considerably reduced and thus the system size increases much faster.

Table 2. Effect of arrival process on $N_C$

| $\xi$ | $EXP$ | $ERL$ | $HYP$ | $MAP^-$ | $MAP^+$ |
|---|---|---|---|---|---|
| 0.25 | 1.2960 | 1.2640 | 1.4886 | 1.3651 | 46.3492 |
| 0.35 | 1.2676 | 1.2336 | 1.4526 | 1.3346 | 45.1494 |
| 0.45 | 1.2409 | 1.2049 | 1.4189 | 1.3059 | 43.9779 |
| 0.55 | 1.2157 | 1.1779 | 1.3871 | 1.2789 | 42.8349 |

## 7. Conclusion

We analyzed a multi-class ($n$ classes), single server, single waiting line queueing system. Arrival process forms a $MAP$. Service time distribution in class $i$ is phase type distributed with representation $(\gamma_i^{(i)}, T_i^{(i)})$ of order $m_i^{(i)}$. A customer, assumed to belong to class $i$, is taken for service in class $j$ at the time when taken for service, with probability $p_{ij}$, $1 \le j \le n$. A timer starts ticking the moment a customer is taken for service. Service in wrong class can lead to permanent damage to the customer and hence he may have to leave the system if correct identification of required class of service is not made before the timer realization. The stability of this system is analyzed on computing the response time of customers. The system state distribution is derived for the stable system. Performance measures were computed and numerically illustrated. Further different arrival processes were considered to find out their impact on the average number of customers in the system. In a future work we propose to introduce $n$ servers, one each for serving customers assumed to belong to a certain class. Distinct queues will be assumed for such class of customers. However, the ambiguity in deciding the class to which a customer belongs is still assumed (a class $i$ customer may join class $j$ waiting line). Analysis of such system is highly challenging.

## Acknowledgment

The authors thank the reviewers for the comments on the earlier version of this paper, that helped in improving its presentation. Support from funding agencies, as indicated below, is gratefully acknowledged: (i) A. Krishnamoorthy: Department of Science and Technology, Government of India under grant INT/RUS/RSF/P-15; (ii) V. Vishnevsky: Russian Science Foundation and the Department of Science and Technology (India) via grant No. 16-49-02021 for the joint research project by the V. A. Trapeznikov Institute of Control Science and the CMS College Kottayam; and (iii) Dhanya Shajin: Kerala State Council for Science, Technology & Environment: 001-07/PDF/2016/KSCSTE.

## References

[1] Atar, R., Mandelbaum, A., & Reiman, M. I. (2004). Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. *Annals of Applied Probability*, 14(3), 1084-1134.

[2] Avarchenkov, K., Morozov, E., & Steyaert, B. (2016). Sufficient stability conditions for multi-class constant retrial rate systems. *Queueing Systems*, 82, 149-171.

[3] Chakravarthy, S. R. (2001). The batch Markovian arrival process: a review and future work, A. Krishnamoorthy et al. (Eds.). *Advances in Probability Theory and Stochastic Processes*, Notable Publications, NJ (Inc.) 21-49.

[4] Falin, G. I. (1988). On a multi-class batch arrival retrial queue. *Advances in Applied Probability*, 20, 483-487.

[5] Gross, D., & Harris, C. M. (1988). Fundamentals of Queueing Theory. John Wiley and Sons, New York.

[6] Harrison, J. M., & Zeevi, A. (2004). Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. *Operations Research*, 243-257,

[7] Kelly, F. P. (1991). Effective bandwidth of multi-class queues. *Queueing Systems*, 9(1), 5-15,

[8] Madan, K. C. (2000). An M/G/1 queue with second optional service. *Queueing systems*, 34, 37-46.

[9] Medhi, J. (2002). A single server Poisson input queue with a second optional channel. *Queueing systems*, 42, 239-242.

[10] Neuts, M.F. (1981). Matrix-Geometric Solutions in Stochastic Models, Johns Hopkins University Press, Baltimore.

[11] Righter, R. (2000). Expulsion and scheduling control for multiclass queues with heterogeneous servers. *Queueing Systems*, 34, 289-300.

[12] Sharif, A. B., Stanford, D. A., Taylor, P., & Ziedins, I. (2014). A multi-class multi-server accumulating priority queue with application to health care. *Operations Research for Health Care*, 3(2), 73-79.