RAIRO-Oper. Res. 51 (2017) 417–432
DOI: 10.1051/ro/2016032

RAIRO Operations Research
www.rairo-ro.org

PRIORITY QUEUE WITH BATCH ARRIVAL, BALKING, THRESHOLD RECOVERY, UNRELIABLE SERVER AND OPTIMAL SERVICE

Madhu Jain[1]

Abstract. The threshold policy for the restoration of an unreliable server in a service system with bulk input and balking is investigated. The arriving customers in the queueing system are classified into two categories *i.e.* priority and ordinary customers. The priority customers are assumed to join the system in groups according to Poisson process. The ordinary customers join the system singly and require the essential service as well as optional service on demand and only a limited number of customers can wait in the queue when the server is busy. The service times of both types of customers and life time as well as repair time of the server are governed by the exponential distribution. When the server fails during the service of the ordinary customer, the repair is done following a threshold recovery rule according to which the repair of the failed server is started only when at least q ordinary customers are accumulated in the system. In case of failure while rendering the service to the priority customers, the server is immediately sent for the repair. The matrix geometric method (MGM) has been used to establish the queue size distribution and other performance indices. To validate the suggested MGM approach, numerical simulation is carried out by taking an illustration.

Mathematics Subject Classification. 60K25, 60K30, 68M20, 90B22, 90B25.

Received October 16, 2014. Accepted September 17, 2015.

1. Introduction

In many congestion situations, a preferential treatment in rendering service is given to some individual priority class customers. Such priority based service rule can be realized in many real world queueing problems including the communication congestion scenarios. The priority mechanism is an invaluable scheduling policy to serve the customers based on pre-specified rule and has been analyzed by many researchers working in the area of queueing theory. According to the priority rule, the customers of different classes are allowed to receive different quality of service and are studied in two broad categories (i) preemptive priority and (ii) non-preemptive priority. In preemptive priority queue, the service of a lower priority customer is interrupted on the arrival of high priority customer. According to non-preemptive priority rule, if a high priority customer joins the system, it is served after the completion of ongoing service in case when the server is already busy in servicing a lower priority customer. In the present investigation, we study a single server queueing system with two-classes of customers; the class one (*i.e.* priority) customers have preemptive priority over the class two (*i.e.* ordinary priority) customers.

Keywords. Bulk queue, priority, unreliable server, threshold recovery, balking function, optional service, matrix geometric method, queue size.

[1] Department of Mathematics, IIT Roorkee, Roorkee, India. drmadhujain.iitr@gmail.com

Article published by EDP Sciences

In recent past, important contributions on preemptive priority queue are due to Avi-Itzhak and Naor [1], Chang [3], Miller [25], and many others. A two-class single server queueing system with state dependent arrivals and preemptive priority discipline was considered by Bitran and Caldentey [2]. The queueing analysis of priority queue with two-classes and with K-classes of jobs was presented by Groenevelt *et al.* [11] and Derbala [5], respectively. Drekic and Woolford [8] analyzed a two-class single server preemptive priority queueing model with balking for low priority customers. Kamoun [21] analyzed a non-preemptive priority queueing system with correlated Markovian interruptions. Mokaddis *et al.* [26] studied a queueing system with single vacation and three classes of customers. Both preemptive (resume and repeated) and non-preemptive priority queueing models were discussed by Walraevens *et al.* [34]. Papier *et al.* [29] suggested an emerging method to improve the profit and to better serve high priority customers in a queueing system where the customers can choose between classic and premium services and premium service is priced above the classic service.

In queueing literature, several research articles can be found in which Markov queueing system with unreliable server has been studied but a little attention has been paid towards the analysis of priority queues with unreliable server. The early work on unreliable server with two priority classes was done by White and Christie [35]. They have used the method of generating function to establish the steady state distribution. Liu *et al.* [23] suggested the optimal N-policy for the unreliable server M/G/1 retrial queue with preemptive resume, vacation and feedback. Jain [14] provided the transient analysis of unreliable server priority queueing model for the machining systems supported by standbys. Numerical simulation was carried out to facilitate the transient performance indices by using Runge−Kutta method. Vadivu *et al.* [32] and Vinayak *et al.* [33] investigated the multi-server priority queue with retrial attempts and unreliable server.

In literature on the queueing modeling of unreliable server queues as cited above, the concept of the immediate repair of the failed server is taken in account. However, if there are a very few customers in the system, the concept of urgent repair is not much viable due to economic reason. To tackle such situation of the unreliable server queueing system, the repair process can be delayed according to the threshold recovery policy in which the repair can be started only when a minimum number of customers say $q(<1)$ or more are present in the system. The concept of threshold recovery policy was first time introduced by Efrosini and Semenova [9] to analyze the M/M/1 retrial queueing system with constant retrial rate and un-reliable server. They have proposed that the repair of the failed server can only be started when the queue length is build up upto a certain level, *i.e.* only when $q \geq 1$ or more customers are accumulated in the system. Efrosini and Winkler [10] provided the performance results for M/M/1 retrial queueing system with constant retrial rate, un-reliable server and threshold recovery policy. Purohit *et al.* [30] analyzed the threshold recovery policy for the finite queue with state dependent arrival rates. Jain and Bhagat [16] studied the threshold recovery policy for the finite capacity and finite population retrial queueing models by incorporating some realistic features namely geometric arrivals, second optional service, and impatient customers. They also facilitated the cost analysis to determine the optimal cost of the system. By considering the threshold recovery policy, Yang *et al.* [37] developed a time-dependent machine repair model with server vacations. More recently, Jain and Bhagat [18] presented a double orbit model to study the threshold recovery policy for the unreliable server retrial queue with priority.

There is vast literature on the bulk input queueing models in different frameworks [4]. The batch arrival priority queueing systems have drawn the attention of a few researchers [13,22]. Metwally and Zaki [24] studied priority queueing system with bulk arrivals and operating under N-policy and single vacation. A batch input Markovian priority queueing system with phase service was considered by Zhao *et al.* [38]. Jain and Bhargava [19] examined an unreliable server bulk arrival queue having two classes of non-preemptive priority subscribers and established various performance indices in closed form by using the supplementary variable approach. Thillaigovindan and Kalyanaraman [31] studied a queueing system in which the customers of type 1 arrive in batches and the customers of type 2 arrive singly according to Poisson processes. Dimitriou [6,7] investigated the concept of priority rule in the retrial queue with negative arrivals and service interruption due to server breakdown by considering some more noble features namely multiple vacations and state dependent arrivals, respectively.

For the analysis of queueing models, matrix geometric method (MGM) has been used to analyze Markov processes which have a particular (lower or upper) Hassenberg structure [27]. This method can be applied to the specific type of queueing problems whose coefficient matrix can be decomposed into two parts, the initial portion and the repetitive portion. Matrix geometric approach was first employed to study the priority queues by Halfin and Segal [12]. Jain and Agrawal [15] studied the $M^X/M/1$ queueing system with multiple types of breakdowns by applying matrix geometric approach. Further, Jain and Jain [20] investigated an unreliable server queue with working vacation by using the matrix geometric method. The matrix geometric approach to analyze a finite capacity queue with two phase service was used by Padma *et al.* [28]. Xu and Wang [36] developed the fluid model for the M/M/c queue with working vacation and vacation interruption and obtained probability distribution by considering the matrix geometric structure of the Laplace transform of stationary buffer contents.

There are plenty of applications of queueing model with priority for example ordinary and emergency patients at hospitals, executive and economy classes of customers for air ticket reservation, priority and ordinary customers at call centers, *etc.* In the present investigation we analyze priority queueing model by incorporating many realistic features namely (i) unreliable server (ii) threshold recovery (iii) optional service (iv) balking (v) batch input (vi) state dependent rates, *etc.* The motivation of present work lies in its potential applicability in cellular radio network operating under new call bounding scheme to deal with two type of traffic *i.e.* new and handoff calls. In cellular architecture of wireless communication system, the geographical area is divided into microcells; each microcell has one base station (BS) which transmits the (i) new calls which are originated one by one according to Poisson process in that particular cell and (ii) handover calls which enter in group according to Poisson process in the target cell from the neighboring cell due to mobility of the users. Once the connection of the users is established, it should be continued till completion of the calls thereby handover calls are the priority class traffic and has preemptive priority over the new calls. Also to give priority to handover calls, only a limited number of new calls (say N) are allowed in the buffer whereas there is no limitation of the handover calls in the system. More specifically in the context of fast moving users, the admission of new calls in a cell is considered as single arrival but 'group mobility' of the users which is commonly noticed due to movement of mobile users in vehicle *e.g.* in taxi, bus or train, the handover calls which are the priority calls, arrive according to the *bulk arrival* process. The service channel is subject to breakdown while providing the service of new as well as handover calls. In case of failure of connection due to channel breakdown while rending service to new calls, it is repaired only if a sufficient (say q) number of new calls are accumulated; however in case of service of handover calls, it is immediately repaired in order to reestablish the connection of ongoing call. While channel is busy, the new calls may be discouraged and balk with some probability instead of joining the queue. When there are many handover calls in the waiting lines, the server (*i.e.* channel) may switch over to faster rate to reduce the dropping probability of the handoffs as handover calls are delay sensitive; this can be done by splitting the existing channel into two channels which is known as sub-rating scheme or allocating more bandwidth to the channel after a certain threshold load.

The rest of the paper is structured as follows. The model description by stating the requisite assumptions and notations is given in Section 2. Section 3 provides the steady state equations and mathematical analysis to obtain the queue size distribution by using matrix geometric approach. The performance measures are established in terms of probabilities in Section 4. By taking numerical illustrations, the sensitivity analysis is presented in Section 5. Finally the silent features and further extension of the model developed are outlined in Section 6.

2. Model description

Consider M^X/M/1 priority queueing model with server breakdown, balking and optional service. Two types of customers namely priority and non-priority (*i.e.* ordinary) customers arrive to seek service by an unreliable server. The capacity of ordinary customers is finite (N) whereas there is no restriction on the capacity of the priority customers.

(i) Input process: The priority customers arrive according to Poisson fashion in batches of maximum size 'B'. The probability mass function of batch size X is $C_k = \Pr(X = k)$. Let $\Lambda(\Lambda')$ be the mean arrival rate for the priority customers when the server is busy (broken down). The ordinary priority customers arrive singly in Poisson fashion with state dependent rate λ_l. The ordinary customer may balk from the system depending upon the number of customers present in the system. In case when the server is busy in rendering the optional service, the customers join the system with rate λ. The effective arrival rate of the non-priority customers is given by

$$\lambda_{l,m} = \begin{cases} \lambda_l b_m, & 0 \leq m \leq N \\ 0, & m > N. \end{cases}$$

(ii) Service pattern: The service time of the priority customer is exponentially distributed with rate μ_h. If there is no customer of priority class, the server renders essential service to the ordinary customers according to exponential distribution. After completing the essential service with probability p, the server may also provide second phase service with probability p if the ordinary customer demands for it. The time of optional phase service is exponentially distributed with rate μ_0. The service rate of essential service of the ordinary customer depends upon of the number of ordinary customers present in the system. To cope up with the high load of ordinary customers when the number of the ordinary customers reaches a threshold level K, the server moves to faster service rate. The service rate of ordinary customers is given by

$$\mu_{l,m} = \begin{cases} \mu_l, & 0 \leq m \leq K \\ \mu_{l'}, & K < m \leq N. \end{cases}$$

(iii) Server breakdown and repair: The server is unreliable and is subject to breakdown. The life time and repair time of the server are exponentially distributed. The server may breakdown while rendering service to the priority customers by rate α and is repaired with rate β. The server may also fail with rate α_1 while rendering the essential service to the non-priority customers; it is repaired with rate β_1 by the repairman only when at least q customers are present in the system, *i.e.* if there are less than q customers in the system, the repair of the failed server is delayed and started only when some more customers join the system and the queue length builds up to the threshold recovery level q. It is assumed that during the optional phase service, the server cannot breakdown.

3. The analysis

The queueing model is formulated as Markov process with state space $\{E = (m, n, i) : i = 1, 2, 3; 0 \leq m \leq N$ and $n \geq 0\}$. The index m and n represent the number of ordinary and priority customers in the system, respectively, and $i = 1, 2, 3$ denote that the server is busy in rendering essential service of priority/ordinary server, optional service of ordinary server and broken down state, respectively. The steady state probabilities of the server being busy in rendering essential and optional services are represented by $P_{m,n,1}$ and $P_{m,n,2}$ $\forall$ $0 \leq m \leq N, n \geq 0$. When the server is in broken down state, the steady state probabilities is denoted by $P_{m,n,3}$ $\forall$ $0 \leq m \leq N, n \geq 0$. Corresponding to threshold recovery and finite capacity (N), we denote the indicator function as:

$$u_{m,q} = \begin{cases} 1, & q \leq m \leq N \\ 0, & 0 \leq m \leq q - 1 \end{cases}$$

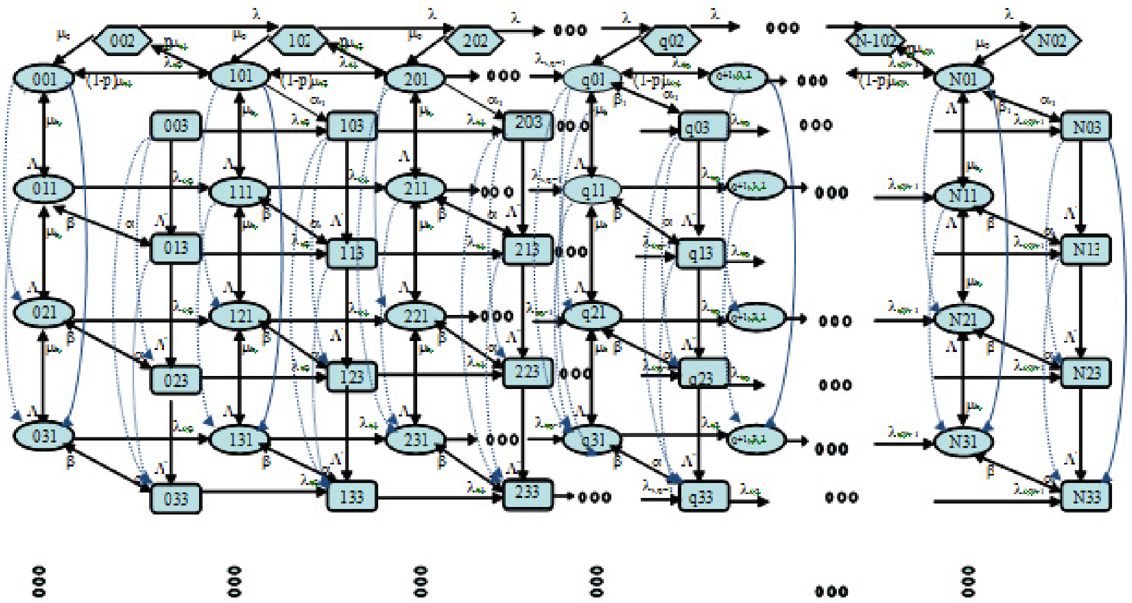

Figure 1. Transition rate diagram.

and

$$\delta_{m,q} = \begin{cases} 1, & m = N \\ 0, & 0 \le m \le N-1. \end{cases}$$

Also denote $g_n = \min(n, B)$, $C_0 = 1$.

For different system states, the steady state Chapman−Kolmogorov equations are constructed by considering the appropriate transition rates (see Fig. 1) as follows:

$$(\lambda_{l,0} + \Lambda + \alpha_1)P_{0,0,1} = (1-p)\mu_{l,1}P_{1,0,1} + \mu_h P_{0,1,1} + \mu_0 P_{0,0,2} \tag{3.1}$$

$$\begin{aligned} \{(1-\delta_{m,M})\lambda_{l,m} + \Lambda + \alpha_1 + \mu_{l,m}\}P_{m,0,1} &= \mu_{l,m+1}(1-p)P_{m+1,0,1} + \mu_h P_{m,1,1} \\ +\mu_0 P_{m,0,2} + \mu_{m,q}\beta_1 P_{m,0,3} + \lambda_{l,m-1}P_{m-1,0,1}, & \qquad 1 \le m \le N \end{aligned} \tag{3.2}$$

$$(\lambda + \mu_0)P_{0,0,2} = p\mu_{l,1}P_{1,0,1} \tag{3.3}$$

$$\{(1-\delta_{m,N})\lambda + \mu_0\}P_{m,0,2} = p\mu_{l,m+1}P_{m+1,0,1} + \lambda P_{m-1,0,2}, 1 \le m \le N \tag{3.4}$$

$$(\lambda_{l,0} + \Lambda')P_{0,0,3} = \alpha P_{0,0,1} \tag{3.5}$$

$$\{(1-\delta_{m,N})\lambda_{l,m} + \Lambda' + u_{m,q}\beta_1\}P_{m,0,3} =\alpha_1 P_{m,0,1} + \lambda_{l,m-1}P_{m-1,0,3}, 1 \le m \le N \tag{3.6}$$

$$(\lambda_{l,0}+\Lambda+\mu_h+\alpha)P_{0,n,1}=\mu_h P_{0,n+1,1}+\beta P_{0,n,3}+\Lambda\sum_{k=1}^{g_n}C_k P_{0,n-k,1}, n\geq 1 \tag{3.7}$$

$$\begin{aligned}\{(1-\delta_{m,N})\lambda_{l,m}+\Lambda+\mu_h+\alpha\}P_{m,n,1}&=\mu_h P_{m,n+1,1}+\beta P_{m,n,3}\\ +\Lambda\sum_{k=1}^{g_n}C_k P_{m,n-k,1}+\lambda_{l,m-1}P_{m-1,n,1}, n\geq 1, &\qquad 1\leq m\leq N\end{aligned} \tag{3.8}$$

$$\{\lambda_{l,0}+\Lambda'+\beta\}P_{0,n,3}=\alpha P_{0,n,1}+\Lambda'\sum_{k=1}^{g_n}C_k P_{0,n-k,3},\ g_n=\min(n,B), n\geq 1 \tag{3.9}$$

$$\begin{aligned}&\{(1-\delta_{m,N})\lambda_{l,m}+\Lambda'+\beta\}P_{m,n,3}=\alpha P_{m,n,1}+\sum_{k=1}^{g_n}C_k\Lambda' P_{m,n-k,3}+\lambda_{l,m-1}P_{m-1,n,3},\\ &g_n=\min(n,B),\qquad 1\leq m\leq N\end{aligned} \tag{3.10}$$

The probabilities associated with different states are determined by using the matrix geometric method proposed by Neuts [28]. For the analysis purpose, we consider an irreducible generator matrix Q in terms of coefficients of equations (3.1)−(3.10). The generator matrix **Q** in a block-partitioned form is as follows:

$$\mathbf{Q}=\begin{bmatrix}
\mathbf{D}_0 & \mathbf{M} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots\ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots\\
\mathbf{M}_1 & \mathbf{D} & \mathbf{M} & \mathbf{0} & \mathbf{0} & \ldots\ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots\\
\mathbf{M}_2 & \mathbf{M}_1 & \mathbf{D} & \mathbf{M} & \mathbf{0} & \ldots\ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots\\
\mathbf{M}_3 & \mathbf{M}_2 & \mathbf{M}_1 & \mathbf{D} & \mathbf{M} & \ldots\ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots\\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\
\mathbf{M}_B & \mathbf{M}_{B-1} & \mathbf{M}_{B-2} & \mathbf{M}_{B-3} & \ldots & \ldots\ldots & \mathbf{M}_1 & \mathbf{D} & \mathbf{M} & \mathbf{0} & \mathbf{0} & \ldots\\
\mathbf{0} & \mathbf{M}_B & \mathbf{M}_{B-1} & \mathbf{M}_{B-2} & \mathbf{M}_{B-3} & \ldots\ldots & \mathbf{M}_2 & \mathbf{M}_1 & \mathbf{D} & \mathbf{M} & \mathbf{0} & \ldots\\
\mathbf{0} & \mathbf{0} & \mathbf{M}_B & \mathbf{M}_{B-1} & \mathbf{M}_{B-2} & \ldots\ldots & \mathbf{M}_3 & \mathbf{M}_2 & \mathbf{M}_1 & \mathbf{D} & \mathbf{M} & \ldots\\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots
\end{bmatrix}$$

where $\mathbf{D_0}$, $\mathbf{D}$, $\mathbf{M}$, $\mathbf{M_k}$ ($k=1,2,\ldots,$ B) are the square matrices of order $3(N+1)$ and given by

$$\mathbf{D}_0=\begin{bmatrix}\mathbf{A}_1 & \mathbf{S}_2 & \mathbf{S}_3\\ \mu_{\mathbf{0}}\mathbf{I} & -(\lambda+\mu_0)I & \mathbf{0}\\ \mathbf{S}_1 & \mathbf{0} & \mathbf{A}_0\end{bmatrix}, \mathbf{D}=\begin{bmatrix}\mathbf{A}_3 & \mathbf{0} & \alpha I\\ \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \beta I & \mathbf{0} & \mathbf{A}_2\end{bmatrix}, \mathbf{M}=\begin{bmatrix}\mu_h I & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$$\mathbf{M_k}=\begin{bmatrix}\mathbf{L}_k & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & \mathbf{F}_k\end{bmatrix}, \text{Ł}_k=\mathbf{C}_k\boldsymbol{\Lambda}\mathbf{I},\ \mathbf{F}_k=\mathbf{C}_k\Lambda'\mathbf{I}$$

$$\mathbf{S}_1 = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \beta_1 \mathbf{I}_{N+1-q} \end{bmatrix}$$

$$\mathbf{S}_2 = \begin{bmatrix} 0 & p\mu_{l,1} & 0 & \dots & 0 \\ 0 & 0 & p\mu_{l,2} & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & p\mu_{l,N} \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

$$\mathbf{S}_3 = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & \alpha_1 & 0 & \dots & 0 \\ 0 & 0 & \alpha_1 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & \alpha_1 \end{bmatrix}$$

$$\mathbf{A}_0 = \begin{bmatrix} -(\lambda_{l,0} + \Lambda') & \lambda_{l,0} & \dots & 0 & 0 & \dots & 0 \\ 0 & -(\lambda_{l,0} + \Lambda') & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_{l,q-2} & 0 & \dots & \dots \\ \dots & \dots & \dots & -(\lambda_{l,q-1} + \Lambda') & \lambda_{l,q-1} & \dots & \dots \\ \dots & \dots & \dots & \mathbf{0} & -(\lambda_{l,q} + \Lambda' + \beta_1) & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \lambda_{l,N-1} \\ 0 & 0 & 0 & 0 & 0 & \dots & -(\lambda_{l,N} + \Lambda' + \beta_1) \end{bmatrix}$$

$\mathbf{A}_1 =$

$$\begin{bmatrix} -(\lambda_{l,0} + \Lambda) & \lambda_{l,0} & 0 & \dots & 0 \\ (1-p)\lambda_{l,1} & -(\lambda_{l,1} + \Lambda + \alpha_1 + \mu_{l,1}) & \lambda_{l,1} & \dots & 0 \\ 0 & (1-p)\lambda_{l,2} & -(\lambda_{l,2} + \Lambda + \alpha_1 + \mu_{l,2}) & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \lambda_{l,N-1} \\ 0 & 0 & 0 & (1-p)\lambda_{l,N} & -(\lambda_{l,N} + \Lambda + \alpha_1 + \mu_{l,N}) \end{bmatrix}$$

$$\mathbf{A}_2 = \begin{bmatrix} -(\lambda_{l,0} + \Lambda' + \beta) & \lambda_{l,0} & 0 & \dots & 0 \\ 0 & -(\lambda_{l,1} + \Lambda' + \beta) & \lambda_{l,1} & \dots & 0 \\ 0 & 0 & -(\lambda_{l,2} + \Lambda' + \beta) & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \lambda_{l,N-1} \\ 0 & 0 & 0 & \dots & -(\lambda_{l,N} + \Lambda' + \beta) \end{bmatrix}$$

$$\mathbf{A}_3 = \begin{bmatrix} -(\lambda_{l,0} + \Lambda + \alpha + \mu_h) & \lambda_{l,0} & 0 & \dots & 0 \\ 0 & -(\lambda_{l,1} + \Lambda + \alpha + \mu_h) & \lambda_{l,1} & \dots & 0 \\ 0 & 0 & -(\lambda_{l,2} + \Lambda + \alpha + \mu_h) & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \lambda_{l,N-1} \\ 0 & 0 & 0 & \dots & -(\lambda_{l,N} + \Lambda + \alpha + \mu_h).. \end{bmatrix}$$

The stability condition can be easily established by following the Theorem 3.1.1 of Neuts [27]. If the CTMC is ergodic *i.e.* irreducible and positive recurrent, then the queueing system will be stable. For the vector process of priority queueing system, the generator matrix $\mathbf{Q}$ is given by

$$\mathbf{Q} = \mathbf{M} + \mathbf{D} + \sum\nolimits_{k=1}^{B} \mathbf{M_k} \tag{3.11}$$

Let $\boldsymbol{\pi}$ be the steady state probability vector associated with matrix $\mathbf{Q}$, then the invariant probability vector $\boldsymbol{\pi}$ exists if and only if $\boldsymbol{\pi}\sum_{k=1}^{B} \mathbf{M}_k\mathbf{e} < \boldsymbol{\pi}\mathbf{Me}$ where $\mathbf{e}$ is a column unit vector of suitable dimension.

The matrix equation for the probability vector can be written as $\boldsymbol{\pi}$ $\mathbf{Q}$= 0, and $\boldsymbol{\pi}\mathbf{e}$=1.

To solve equation (3.11) with transition rate matrix $\mathbf{Q}$, let us partition probability vector $\boldsymbol{\pi}$ as $\boldsymbol{\pi}$ =[$\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_j, \ldots$] where $\mathbf{P}_0 = [P_{001}, P_{101}, \ldots, P_{N01}, P_{002}, P_{102}, \ldots, P_{N02}, P_{003}, P_{103}, \ldots, P_{N03}]$; $\mathbf{P}_n = [P_{0,n,1}, P_{1,n,1}, \ldots, P_{N,n,1}, P_{0,n,3}, P_{1,n,3}, \ldots, P_{N,n,3}]$; $n \geq 1$
To compute the state probabilities, we use the matrix geometric approach. For this purpose, we need to evaluate the rate matrix $\mathbf{R}$ that is the minimal non-negative solution of the matrix equation

$$\mathbf{M} + \mathbf{RD} + \sum\nolimits_{k=1}^{B} \mathbf{M_k R^{k+1}} = \mathbf{0}. \tag{3.12}$$

Now, we obtain a general equation for R as follows:

$$\mathbf{R_{n+1}} = -\left[\mathbf{M} + \sum_{k=1}^{B} \mathbf{M_k R_n^{k+1}}\right]\mathbf{D^{-1}}. \tag{3.13}$$

After evaluating R, we can evaluate the invariant probability vector recursively by using the relation

$$\mathbf{P_{n+1}} = \mathbf{P_1 R^n}, n \geq 0 \tag{3.14}$$

or

$$\mathbf{P_n} = \mathbf{P_0 R_n}, n \geq 0. \tag{3.15}$$

To obtain the boundary probability vectors [$\mathbf{P_0}, \mathbf{P_1}$], we define the matrix (*cf.* Thm. 1.5.1 of Neuts [27])

$$\mathbf{Q(R)} = \begin{bmatrix} \mathbf{D}_0 & \mathbf{M} \\ \sum_{k=1}^{B} \mathbf{M_k R^{k-1}} & \mathbf{D} + \sum_{k=1}^{B} \mathbf{M_k R^k} \end{bmatrix}.$$

The normalization condition is given by:

$$(\mathbf{P_0} + \mathbf{P_1}(\mathbf{I} - \mathbf{R})^{\mathbf{-1}})\mathbf{e} = 1. \tag{3.16}$$

4. Performance indices

By knowing the measures of performance of the queueing system, the designers/organizers can get insight about the optimal strategy to improve the concerned system by reducing the average queue length, average delay, *etc.* By controlling the suitable parameters, decision makers can also come up with some strategy to reduce the balking behavior of the customers. The steady state probabilities determined by MGM can be further used to establish various performance indices as follows:

(I) Average queue length

(a) Expected number of priority customers in the system is

$$\text{Ł}_1 = \Sigma_{m=0}^{N}\Sigma_{n=0}^{\infty} n(P_{m,n,1} + P_{m,n,3}) \tag{4.1}$$

(b) Expected number of ordinary priority customers in the system is

$$\text{Ł}_2 = \Sigma_{m=0}^{N}\Sigma_{n=0}^{\infty} m(P_{m,n,1} + P_{m,n,3}) + \Sigma_{i=0}^{N} m P_{m,0,2} \tag{4.2}$$

(c) Expected number of customers in the system when the server is in breakdown state, is

$$\text{Ł}_d = \Sigma_{n=0}^{\infty}\Sigma_{n=0}^{N}(m+n)P_{m,n,3} \tag{4.3}$$

(d) Expected number of customers in the system is

$$\text{Ł} = L_1 + L_2 \tag{4.4}$$

(II) Expected waiting time

(a) The carried load which is the effective arrival rate, is evaluated by using

$$\begin{aligned}\lambda_{\text{eff}} &= \Sigma_{m=0}^{N-1}\Sigma_{n=0}^{\infty}(\lambda_{l,m} + \Lambda E\{X\})P_{m,n,1} + \Sigma_{m=0}^{N-1}\lambda P_{m,0,2} + \Sigma_{m=0}^{N-1}\Sigma_{n=0}^{\infty}(\lambda_{l,m} + \Lambda' E\{X\})P_{m,n,3}\\ &+ \Sigma_{m=0}^{N-1} E\{X\}\Lambda' P_{N,n,3}\end{aligned} \tag{4.5}$$

(b) Expected waiting time can be determined by using Little's formula given by

$$W = \frac{L}{\lambda_{\text{eff}}}\cdot \tag{4.6}$$

(III) Throughput and expected delay

(a) The system throughput is determined by

$$TP = \Sigma_{m=1}^{N}\mu_{l,m}P_{m,0,1} + \Sigma_{m=0}^{N}\Sigma_{n=1}^{\infty}\mu_h P_{m,n,1} + \Sigma_{m=0}^{N}\mu_0 P_{m,0,2} \tag{4.7}$$

(b) The mean delay is obtained by using

$$D_L = \frac{L}{TP}\cdot \tag{4.8}$$

(VI) Reliability indices

(a) The availability of the server is determined by

$$A = \Sigma_{m=0}^{N}P_{m,0,1} + \Sigma_{m=0}^{N}\Sigma_{n=0}^{\infty}P_{m,n,1} + \Sigma_{m=0}^{N}P_{m,0,2}. \tag{4.9}$$

(b) The failure frequency of the server is obtained as:

$$F_f = \Sigma_{m=0}^{N}\alpha_1 P_{m,0,1} + \Sigma_{m=0}^{N}\Sigma_{n=1}^{\infty}\alpha P_{m,n,1}. \tag{4.10}$$

(V) Long run probabilities and cost function

(a) The long run probability that the server is in broken down state is

$$P_D = \Sigma_{n=0}^{\infty}\Sigma_{m=0}^{N} P_{m,n,3}. \tag{4.11}$$

(b) The long run probability of the server being busy is

$$P_B = \Sigma_{m=1}^{N} P_{m,0,1} + \Sigma_{m=0}^{N}\Sigma_{n=1}^{\infty} P_{m,n,1} + \Sigma_{m=0}^{N} P_{m,0,2}. \tag{4.12}$$

(c) To frame the cost function, we consider the following cost factors corresponding to different activities:

C_b: Cost per unit time when the server is busy;
C_h: Holding cost per unit time of the customers present in the system;
Cr: Repair cost incurred per unit time for a broken down server;
C_d: Cost incurred per unit time on the server when in broken down state but the repair is started not yet as threshold level of the number of ordinary customers is not reached

The total cost per unit time incurred on the system is formulated as function of threshold recovery parameter as:

$$TC(q) = C_b P_B(t) + C_h L + C_r\beta + C_d P_D(t). \tag{4.13}$$

5. Illustration and sensitivity analysis

The priority queueing models with unreliable server have enormous applications in day to day as well industrial set up. The prediction of throughput and delay seems to be of enormous utility for the system engineers and designers to determine the buffer capacity for the ordinary customers. To correlate the applicability of our model, we cite the example of two types of traffic in the cellular radio network due to origination of new calls who arrive singly and bulk arrivals of priority calls due to group mobility of the mobile users due to movement of some vehicle from the coverage area of base station of the neighboring cell to the targeted cell. The problem of reducing the dropping of calls in cellular radio system can be easily tackled by implementation of priority in the allocation of channel to handover calls over the new calls. In order to minimize the capacity wastage while keeping the number of new calls below a given tolerance level by bounding scheme, the provision of limited buffer for new calls is realistic assumption. The concept of facilitating the optional additional services to the new calls when no handover calls (*i.e.* priority class customers) are present in the system seems to noble concept in earning more profit to the service providers. Moreover, the additional feature of the threshold recovery of the broken down channel after accumulation of sufficient workload of new calls enhances the system capacity by reducing the idle time of the channel in case of low traffic of new calls. The priority queueing model developed has its vital utility in cellular radio network in order to provide managerial insights to the system designers and decision makers to reduce the dropping of handoffs and improve the throughput, Moreover, the concept of threshold recovery enhances the practical applicability of our model. That is because it happens in daily life activities due to the fact that the repairman is called upon or visited when jobs of failed units are accumulated so as to save the time and money.

Now by keeping the applicability in cellular radio network, we perform numerical simulation by assuming the following two functions for the joining probability of the ordinary calls:

(a) Fractional Balking Function (FBF):

$$b_m = \frac{1}{m+1}, 0 \le m \le N \tag{5.1}$$

TABLE 1. Performance measures by varying different parameters.

		L		TP		W		DL	
N	(μ_l, μ_h, μ_0)	EBF	FBF	EBF	FBF	EBF	FBF	EBF	FBF
7	1,3,2	5.98	5.83	7.52	5.01	18.02	12.49	0.70	0.89
	2,3,2	5.78	5.75	9.57	6.73	17.13	12.27	0.61	0.77
	3,4,3	4.65	4.62	10.25	8.12	15.73	11.79	0.51	0.66
	3,5,3	4.25	4.54	11.13	9.32	15.35	10.25	0.42	0.55
	3,5,4	4.12	4.27	13.46	11.13	14.42	10.15	0.37	0.42
	3,5,5	3.94	3.62	16.43	12.52	13.26	9.42	0.35	0.30
9	1,3,2	7.52	7.89	7.77	5.12	20.33	16.01	0.79	1.56
	2,3,2	7.31	7.66	10.46	7.71	19.60	17.13	0.69	0.91
	3,4,3	7.30	7.39	12.83	11.18	18.72	15.38	0.61	0.85
	3,5,3	6.08	6.15	13.77	12.07	16.27	14.99	0.52	0.59
	3,5,4	5.61	5.72	15.16	14.03	16.62	12.34	0.48	0.52
	3,5,5	5.35	5.27	17.09	14.84	14.19	10.25	0.38	0.44

(b) Exponential Balking Function (EBF):

$$b_m = \mathrm{e}^{-\sigma m}, (\sigma > 0). \tag{5.2}$$

For computing the numerical results, the coding of the computer program is done in software MATLAB. For the illustration purpose, the default system parameters are fixed as follows; $N = 11$, $B = 5$, $K = 6$, $H = 7$, $p = 0.3$, $\Lambda = 0.3$, $\Lambda' = 0.15$, $\lambda_h = 0.02$, $\lambda_l = \lambda = 0.01$, $\alpha = \alpha_1 = 2$, $\beta = \beta_1 = 3$, $q = 2$, $\sigma = 0.001$, $\mu_h = 8$, $\mu_l = 9$, $\mu_l' = 13.5$, $C_b = 2$ 0U, $C_h = 15$ U, $C_r = 25$ U, $C_b = 10$ U. By varying the values of q from 1 to $N = 11$, we compute the TC(q) and obtain the minimum total cost $TC = 224.94$ U at $q = 2$. For computation of various performance indices given in equations (4.1)−(4.7), we take $q = 2$ and obtain $L = 2.21$, $TP = 12.05$, $W = 2.0$.

By knowing the measures of performance of the queueing system, the designers/organizers can get insight about the optimal strategy to improve the concerned system by reducing the average queue length, average delay, *etc.* By controlling the suitable parameters, decision makers can also come up with some solution to reduce the balking behavior of the customers.

In order to facilitate the sensitivity analysis we display the numerical results in Table 1 and Figures 2−8 by varying different parameters. Tables 1 summarizes the results for different values of N with default parameters chosen as $q = 2$, $p = 0.3$, $\Lambda = 0.3$, $\Lambda' = 0.15$, $\lambda_h = 0.02$, $\lambda_l = \lambda = 0.01$, $\alpha = \alpha_1 = 2$, $\beta = \beta_1 = 3$, $q = 2$, $\sigma = 0.001$, $\mu_h = 8$, $\mu_l = 9$, $\mu_l' = 13.5$, $B = 5$, $K = 6$, $H = 7$, $N = 11$. The numerical results are computed by considering the exponential balking function (EBF) and the fractional balking function (FBF) separately for each case. From Table 1, it is observed that the system queue length (L), waiting time (W) and delay (DL) of the customers in the system are decreasing function of the service rates, while throughput (TP) of the system tends to increase on increasing the service rates. This is a quite common phenomenon seen in day-to-day life, where the better service enhances the throughput TP of the system and decrement in the waiting time or delay. The numerical results obtained for the FBF case are lesser than the EBF corresponding to L, TP and W, while a reverse effect is seen with the delay (DL).

Figures 2 and 3 depict the throughput (TP) of the system by varying service rate μ_h of the priority customers corresponding to different values of N and K, respectively. As N and K increase, there is increment in throughput as can be noticed from Figures 2 and 3, respectively. It is seen from the graphs, that for the lower value of service rate, the system throughput has slight variation, but thereafter it increases remarkably. Also, there is a significant increment in the values of TP in case of FBF as compared to EBF.

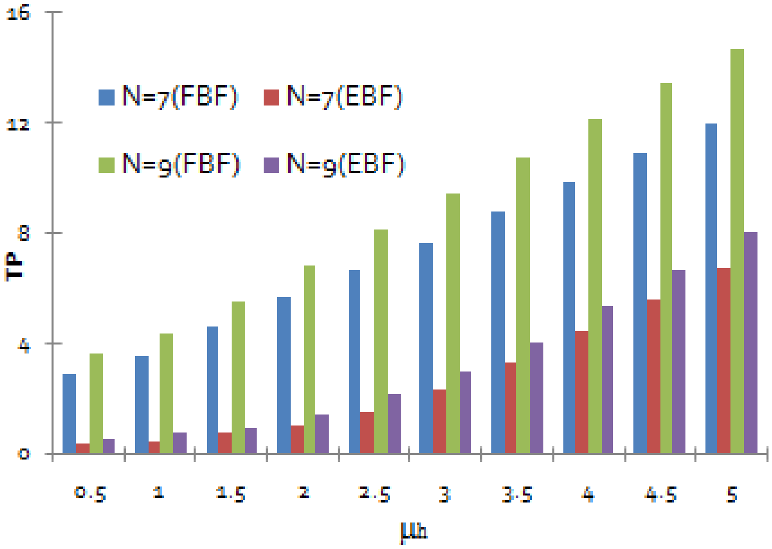

FIGURE 2. Effect of μ_h on TP by varying N.

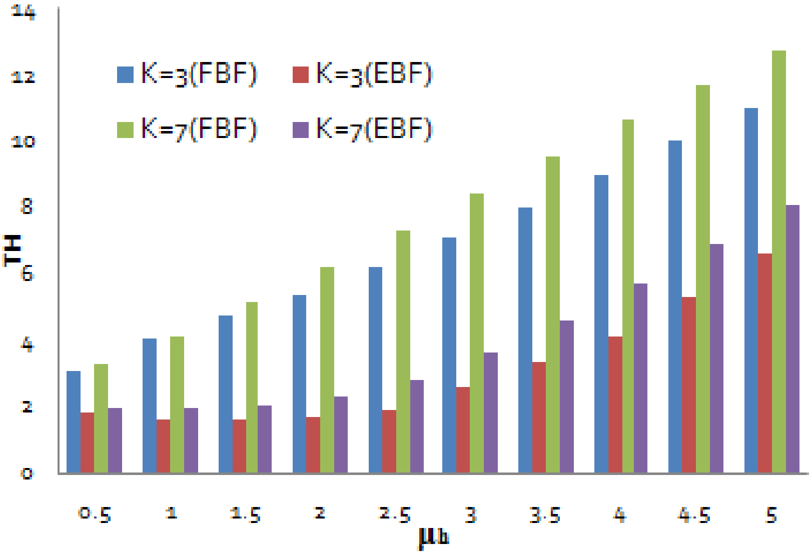

FIGURE 3. Effect of μ_h on TP by varying K.

Figures 4 and 5 show the pattern for the queue length L for different values of N and K. The discrete (continuous) lines correspond to FBF (EBF) case. As expected, L reveals the increasing trend for the increasing values of Λ in both figures. It is also observed that the system queue length increases (decreases) by increasing the parameter $N(K)$ for both exponential and fractional balking functions.

The queue length for different values of N by varying λ_l, α and β are depicted in Figures 6–8, respectively. It is seen from Figure 6 that the queue length (L) increases for the increasing values of arrival rate (λ_l) of ordinary

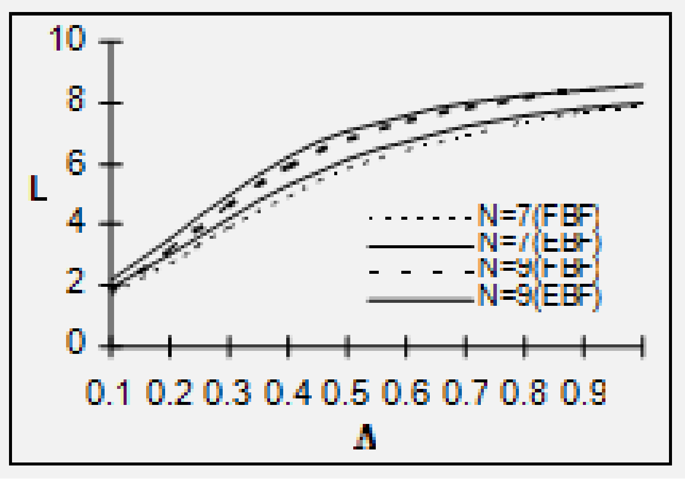

Figure 4. Average queue length (L) *vs.* Λ by varying N.

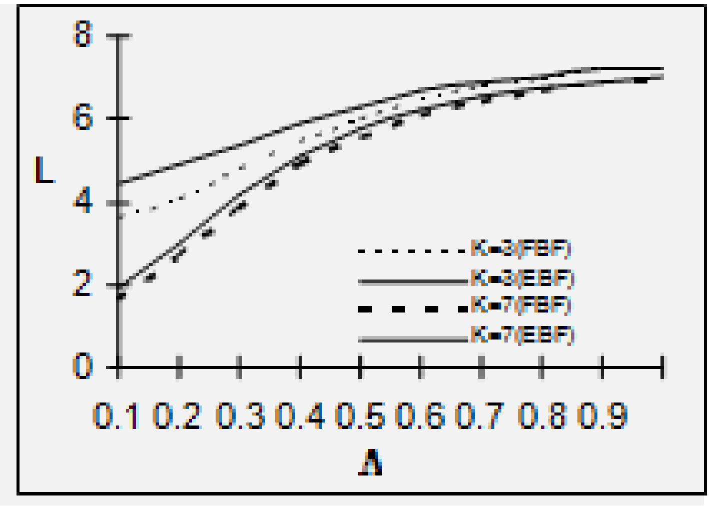

Figure 5. Average queue length (L) *vs.* Λ by varying K.

customers. The system queue length seems to increase sharply with respect to parameter N for the exponential balking function whereas it reveals almost constant value in case of fractional balking function.

Figures 7 and 8 depict the trends of the queue length for different values of N by varying α and β respectively. There is almost linear pattern of queue length (L) for increasing values of failure rate (α) of the server. On the contrary, a gradual decreasing trend in the queue length is noticed for the increasing value of the repair rate (β); the effect seems to more prominent for the higher value of N.

From the numerical results summarized in the form of table and graphs, we overall conclude that the failure rates as well as repair rates have significant impact on the queue length and TP of the system. The effects of the parameters N as well as K are quite noticeable for different system indices. As expected, an increment in

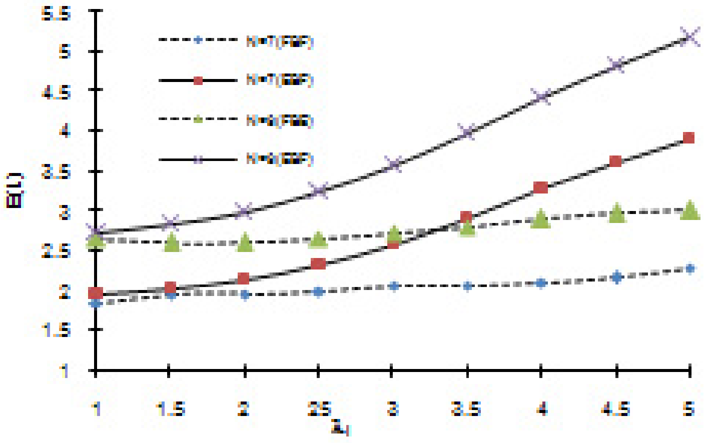

FIGURE 6. Effect of λ_l on L for different value of N.

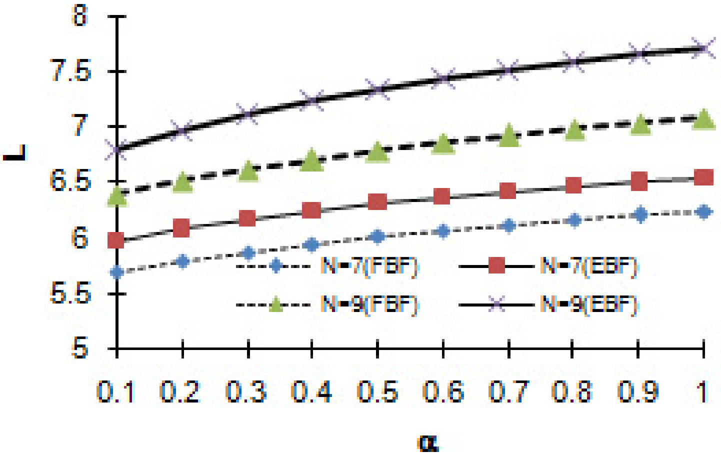

FIGURE 7. Effect of α on L for different value of N.

the service rate of priority customers results in buildup of the throughput to the reasonably good extent. The numerical results show the lower values of various system characteristics for exponential balking function as compared to fractional balking function.

The determination of optimal threshold recovery parameters by minimizing the total cost incurred on different activities can provide an insight to the system organizer to start of the repair of broken down server. For the cited applicability in cellular radio network, the service provider can handle the breakdown of channel and its repair based on the optimal threshold workload of new calls in the network. It is worth noting that delayed repair may prove economical in terms of both time and efforts. Based on above numerical simulation, we infer that the optimal threshold parameters and other system characteristics can be used to optimize the system capacity by enhancing the throughput. As far as managerial implication is concerned, the proposed model can be easily and effectively applied in the design and dimensioning of cellular radio networks, in particular when there is high group mobility of the users.

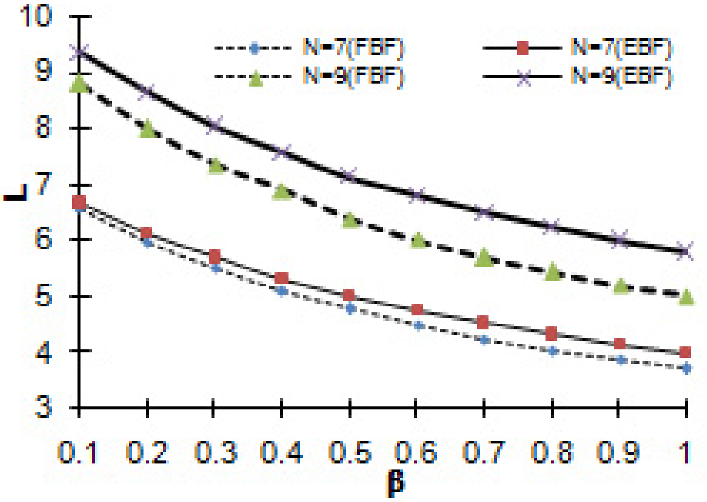

Figure 8. Effect of β on L for different value of N.

6. Conclusion

The priority queueing model with batch arrivals is of a considerable interest in particular when the priority is given to the batch arrivals. Due to incorporation of balking behavior, the optional service to the ordinary customers, unreliable server and threshold recovery, our model depicts many real time congestion situations. Such scenarios are commonly seen in computer systems and communication networks where our model can be easily implemented to improve the grade of service. In case of wireless communication system operating under cellular architecture, there may be two types of traffic *i.e.* new and handoff attempts. The handoff calls are of busty type due to group mobility of the users due to the movement of vehicles. The new originating calls may arrive singly and can wait in buffer of limited size in case when the channel is busy. The channel may be unreliable and subject to breakdown and repair. The numerical results and sensitivity analysis facilitated provide insight how the system can be made more efficient by controlling the sensitive parameters. The bulk service as well as server vacation are some important concepts which can be further incorporated in the model studied but the analysis will become more tedious.

References

[1] B. Avi-Itzhak and P. Naor, On a problem of preemptive priority queueing. *Oper. Res.* **9** (1961) 664–672.

[2] G. Bitran and R. Caldentey, Two class priority system with state dependent arrivals. *Queueing Syst., Theory Application* **40**(4) (2000) 355–382.

[3] W. Chang, Preemptive priority queues. *Oper. Res.* **13** (1965) 820–827.

[4] M.L. Chaudhary and G.C. Templeton, A First Course in Bulk Queues. John Wiley and Sons, New York (1983).

[5] A. Derbala,Priority queueing in an operating system. *Comput. Oper. Res.* **32** (2005) 229-238.

[6] I. Dimitriou, A mixed priority retrial queue with negative arrivals, unreliable server and multiple vacations. *Appl. Math. Modell.* **37** (2013) 1295–1309.

[7] I. Dimitriou, A preemptive resume priority retrial queue with state dependent arrivals, unreliable server and negative customers. *TOP* **21** (2013) 542–571.

[8] S. Drekic and D.G. Woolford, A preemptive priority queue with balking. *Eur. J. Oper. Res.* **164** (2005) 387–401.

[9] D.V. Efrosinin and O.V. Semenova, An M/M/1 system with an unreliable device and a threshold recovery policy. *J. Commun. Technology Electron.* **55** (2010) 1526.

[10] D. Efrosinin and A. Winkler, Queueing system with a constant retrial rate, non-reliable server and threshold-based recovery. *Eur. J. Oper. Res.* **210** (2011) 594–605.

[11] R. Groenevelt, G. Koole and P. Nain, On the bias vector of a two class preemptive priority queue. *Math. Methods Oper. Res.* **55** (2002) 107–120.

[12] S. Halfin and M. Segal, A priority queueing model for a mixture of two types of customers. *SIAM J. Appl. Math.* **23** (1972) 369–379.

[13] A.G. Hawkes, Time dependent solution of a priority queue with bulk arrivals. *Oper. Res.* **13** (1965) 586–596.

[14] M. Jain, Transient analysis of machining systems with service interruption, mixed standbys and priority. *Int. J. Math. Oper. Res.* **5** (2013) 604–625.

[15] M. Jain and P.K. Agrawal, Optimal policy for bulk queue with multiple types of server breakdown. *Int. J. Oper. Res.* **4** (2009) 35–54.

[16] M. Jain and A. Bhagat, Finite population retrial queueing model with threshold recovery, geometric arrivals and impatient customers. *J. Inf. Oper. Manage.* **3** (2012) 162–165.

[17] M. Jain and A. Bhagat, Transient analysis of retrial queues with double orbits and priority customers. *Proc. 8th International Conference on Queuing Theory and Network Applications,Taichung, Taiwan* (2013) 235–240.

[18] M. Jain and A. Bhagat, Double orbit finite retrial queues with priority customers and service interruption. *Appl. Math. Comput.* **253** (2015) 324–344.

[19] M. Jain and C. Bhargava, Bulk arrival retrial queue with unreliable server and priority subscribers. *Int. J. Oper. Res.* **5** (2008) 242–259.

[20] M. Jain and A. Jain, Working vacations queueing models with multiple types of server breakdowns. *Appl. Math. Model.* **34** (2010) 1–13.

[21] F. Kamoun, Performance analysis of a non-preemptive priority queueing system subjected to a correlated Markovian interruption process. *Comput. Opear. Res.* **35** (2008) 3969–3988.

[22] G.V. Krishna Reddy, R. Nadarajan and P.R. Kandasamy, A non-preemptive priority mutiserver queueing system with general bulk service and heterogeneous arrivals. *Comput. Oper. Res.* **20** (1993) 447–453.

[23] Z. Liu, J. Wu and G. Yang, An M/G/1 retrial G-queue with preemptive resume and feedback under N-policy subject to the server breakdowns and repairs. *Comput. Math. Appl.* **58** (2009) 1792–1807.

[24] S.A. Metwally and B.M. Zaki, Head-of-the-line priority discipline for the system with N-policy and single vacation. *J. the Egyptian Math. Soci.* **13** (2005) 159–172.

[25] R.D. Miller, Computation of steady state probabilities for M/M/1 priority queues. *Oper. Res.* **29** (1981) 945–958.

[26] G.S. Mokaddis, S.A. Metwally and B.M. Zaki, On a batch arrival queue with priority and single vacation. *Int. J. Inform. Manage. Sci.* **20** (2009) 519–534.

[27] M.F. Neuts, Matrix Geometric Solutions in Stochastic Models-An Algorithmic Approach, Dover Publications, New York (1981).

[28] Y. Padma, A.R. Reddy and L.V. Rao, Matrix geometric approach for M/M/C/N queue with two phase service. *Int. J. Eng. Sci. Adv. Technol.* **2** (2012) 166–175.

[29] F. Papier and U.W. Thonemann, Capacity rationing in rental systems with two customer classes and batch arrivals. *Omega* **39**(1) (2011) 73-85.

[30] G. N.Purohit, M. Jain and S. Rani, M/M/1 retrial queue with constant retrial policy, unreliable server, threshold based recovery and state dependent arrival rates. *Appl. Math. Sci.* **6** (2012) 1837–1846.

[31] N. Thillaigovindan and R. Kalyanaraman, A feedback retrial queueing system with two types of arrivals. *Proc. 6th Int, Conf. Queueing Theory and Network Applications* (2011) 177–181.

[32] A.S. Vadivu, R. Vanayak, S. Dharmaraja and R. Arumuganathan, Performance analysis of voice over internet protocol via non-Markovian loss system with preemptive priority and server breakdown. *OPSERACH* **5** (2014) 50–75.

[33] R. Vanayak, S. Dharmaraja and R. Arumuganathan, On the study of simultaneous service by random number of servers with retrial and preemptive priority. *Int. J. Oper. Res.* **20** (2014) 68–90.

[34] J. Walraevens, D. Fiems and H. Bruneel, Performance analysis of priority queueing systems in discrete time. *Network Perform. Eng.* **5233** (2011) 203–232.

[35] H. White and L. Christie, Queueing with preemptive priorities or with breakdown. *Oper. Res.* **356** (1958) 79–96.

[36] X. Xu, and L. Wang, Transient analysis of retrial queues with double orbits and priority customers. *Proc. 8th International Conference on Queuing Theory and Network Applications, Taichung, Taiwan* (2013) 311–315.

[37] D.H. Yang, C.H. Yen and Y.C. Chiang, Numerical analysis for time-dependent machine repair model with threshold recovery policy and server vacations, *Proceedings of the International MultiConference of Engineers and Computer Scientists,* Hong Kong (2013).

[38] J.A. Zhao, B. Li, X.R. Cao and I. Ahmad, A matrix analytic solution for the D-BMAP/PH/1 priority queue. *Queueing Syst., Theory Application* **53** (2006) 127–145.