World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Check for updates

Queue with preemptions and repeat or resumption of preempted service

P.K. Pramod *

Department of Mathematics, Panampilly Memorial Government College, Chalakudy, Kerala-680722, India.

World Journal of Advanced Engineering Technology and Sciences, 2026, 18(01), 058-065

Publication history: Received on 28 November 2025; revised on 05 January 2026; accepted on 07 January 2026

Article DOI: https://doi.org/10.30574/wjaets.2026.18.1.0008

Abstract

Here we consider a single server queueing model consisting of two queues-an infinite capacity queue of low priority customers and a finite capacity N of high priority customers. Customers join the system according to a MMAP. If the server is free, at the epoch of an arrival of a customer (low priority/ high priority) can immediately join for service. An (N +1) faces solid figure with the face marked 0 to N, is tossed at the beginning of the service of an ordinary customer. i^{th} face turns up with probability q_i ($0 \le i \le d$). This decides the maximum number of priority customer(s) allowed to be served during the service of the specified ordinary customer. During the service of a low priority customer pre-emption can take place by the arrival of a high priority customer. Then the preempted customer waits at the head of the low priority queue till either the high priority queue becomes empty or the maximum number of high priority customers permitted to be served, as per the outcome of the toss of solid object, whichever occurs first. The restart/ resumption of pre-empted customer's service takes place when the high priority queue becomes empty or the maximum number of high priority customer's service permitted during his effective service is realized. We introduce a threshold random variable which competes with the duration of each pre-emption; if this realizes before completion of preemption then the pre-empted customer has to get its service repeated; otherwise the service is resumed. Here the random variable corresponding to low priority customers service, high priority customers service and threshold random variable are all distinct and independent *PH* distributed. The system is analysed under stable regime. A few useful measures for system performance are obtained. These help in designing an efficient system. Numerical results are provided to illustrate the system performance.

Keywords: Priority Service; Pre-emptions;Phase Type Distribution; Exponential Distribution; MMAP; Threshold Clock.

1. Introduction

In this paper we consider interruption as the server processing a high priority customer. White and Christie [16] is the first reported work on queues with interruption. Subsequently Heathcote [8], Keilson [7], Gaver [5], Aissani and Artalejo [1], among others, analysed such queueing systems in continuous time. Discrete time queue with failure/interruption at the time when a new service start is a recent work by Atentia and Moreno [3]. (See also Alfa [2]). A detailed review on queues with service interruption could be found in Krishnamoorthy et.al. [11]. For motivation in the investigation of such queues one may refer to Krishnamoorthy et.al. [10] which, we believe, is the first work to give concrete conditions for resumption/repetition of an interrupted service, on removal of the interruption. Almost simultaneously Fiems et.al. [4] considered an interruption queueing model with arbitrarily distributed service time and interruption duration, the arrival constituting a Poisson process. They set a priori probability q for resumption of service; with complementary probability it is repeated. In all these no upper bound was set on the number of interruptions that a customer may undergo. This leads to impatience of waiting customers.

* Corresponding author: P.K. Pramod.

One of the objectives of this chapter is to generalize results in [11] and [12], were concerned with customers of the same priority. Though we would very well interpret an interruption as the server processing a high priority customer, the models were basically confined to single priority. Otherwise, questions like arrival process of priority customers, description of waiting space of such customers would arise. The present paper is concerned with a two-priority service system. In contrast to fixing N as the upper bound for number of pre-emptions of a low priority customer's service, the customer is given the option to choose the maximum number of pre-emptions he is willing to undergo, subject to a maximum of N. Nevertheless the customer who chooses to undergo preemptions closer to N (say > $N/2$) will be given incentives, which will not be available to those who do not opt for such length pre-emptions. Specifically, we assume that q_i, $0 \le i \le N$, is the probability of a low priority customer opting for maximum of i pre-emptions.

2. Mathematical Model

We consider a queueing model in which arrival of low priority and high priority customers occur according to MMAP with representation (D_0,D_1) of order r. The arriving customer is of low (high) priority with probability $p_1(p_2)$. If the server is idle, an arriving customer (low priority or high priority) can immediately join for service. During low priority customers service the arrival of high priority customer preempts him, provided the number of high priority customers served during his pre-emptions has not reached the maximum allowed by the outcome of the solid figure by his own initial choice, and the pre-empted low priority customer waits as the head of the infinity capacity queue of low priority customers. Subsequent high priority customers arriving during that period wait in the finite capacity (K) queue. An $N+1$ faced solid figure with markings $0,1,\ldots,N$, respectively is tossed at the beginning of a low priority customers service; let q_i be the probability that the tossing results in i, $(0 \le i \le N)$, then i is the maximum number of high priority customers allowed to be served during his service period. It may happen there is no priority customer present to be served during the effective service time of a low priority customer, even when the experimental outcome is i (≥ 1). The moment pre-emption takes place the threshold random clock starts ticking. The pre-empted customer gets its service repeated /resumed when the high priority queue becomes either empty or the number of high priority customers served during his service period reaches its maximum, whichever occurs first. When the pre-emption time exceeds a threshold random variable, the interrupted customer gets its service repeated on completion of pre-emption; else the service is resumed, that is it starts at the point where it got pre-empted. Duration of services of low and high priority customers are PH distributed random variables with representations (α,S) and (β,T), respectively; the threshold r.v is PH distributed with representation (δ,U). All these random variables are mutually independent. Write $S^0 = -S\underline{e}$, $T^0 = -T\underline{e}$ and $U^0 = -U\underline{e}$ where $\underline{e}$ is a column vector of 1's of appropriate order. Let $N_1(t)$, $N_2(t)$, $S(t)$, $B_1(t)$ and $B_2(t)$ denote, respectively, the number of low priority customers , high priority customers, status of server, maximum number of high priority customers permitted to be served during a low priority customers service and the number of high priority customers so far served , including the present one , if a high priority service is going on by preemption. When $S(t) = 0$, the server is busy with high priority service and a preempted low priority customer is waiting at the head of the queue; when $S(t) = 1$, the server is busy with high priority customer with no pre-empted customer waiting and $S(t) = 2$ stand for the server busy with low priority service. The process $X(t) = \{(N_1(t),N_2(t),S(t),S_1(t),S_2(t),S_3(t),M(t)),t \ge 0\}$; is a continuous time Markov chain (CTMC) which turns out to be LIQBD with n^{th} level given by $l(n) = \bigcup_{n=0}^{\infty} \psi(n,m,l), 0 \le \mathrm{m} \le K, l = 0,1,2.$ The subsets of $\psi(n,m,l)$ are defined as

$\{(n,m,0,j_1,j_2,i_1,i_2,i_3,i_4);1 \le j_1 \le N;1 \le j_2 \le j_1;1 \le i_1 \le a;1 \le i_2 \le b;0 \le i_3 \le c;1 \le i_4 \le r\}$, for $1 \le m \le K$,$\{(n,m,1,i_2,i_4);1 \le i_2 \le b;1 \le i_4 \le r\}$ and for $1 \le m \le K$, $\{(n,m,2,j_1,j_2,i_2,i_4);1 \le j_1 \le N;1 \le j_2 \le j_1;1 \le i_2 \le b;1 \le i_4 \le r\}$. The states in Ψ are listed in lexicographical order. The transitions among subsets $\psi(n,m,l)$, $l = 0,1,2$ are as follows:

For $1 \le m \le K$, $I_K \otimes I_{N(N+1)/2} \otimes Ia \otimes I_b \otimes I_{(c+1)} \otimes p1D1$, $I_K \otimes I_b \otimes p1D1$ and $I_{(N(N+1)/2)+Kd} \otimes I_a \otimes p_1D_1$ records transition rates to states in $\psi(n+1,m,0)$, $\psi(n+1,m,1)$ and $\psi(n+1,m,2)$ respectively, starting from states in $\psi(n+1,m,0)$, $\psi(n+1,m,1)$ and $\psi(n+1,m,2)$.

The matrix $D = (D_{1i})_{1 \times (N+1)}, i = 1,2,\ldots\ldots,N+1$ is a row vector with components $D_{11},D_{12},\ldots\ldots\ldots,D_{1,N+1}$; $D_{1i} = T^0 \otimes q_{i-1}e_i \otimes \alpha \otimes I_r$, $1 \le i \le N+1$; records transition rates at the beginning of low priority service on completion of a high priority priority customers service where e_i is a column vector of order i with 1 in the 1^{st} place and zero elsewhere.

The matrix $\mathrm{Q}=\left(Q_{1i}\right)_{1\times(N+1)}, i=1,2,\ldots,\mathrm{N}+1$ is a row vector with components $\mathrm{Q}=\left(Q_{1i}\right)_{1\times(N+1)}, i=1,2,\ldots,\mathrm{N}+1;$ $Q_{1i}=S^0\otimes q_{i-1}e_i\otimes\alpha\otimes l_t, 1\le i\le N+1$, records transition rates at the beginning of low priority service on completion of a low priority customers service in state 2 where e_i is a column vector of order i with 1 in the 1st place and zero elsewhere.

The matrix $B = (B_{1i})_{1\times(N+1)}$, $i = 1,2,\ldots\ldots,N+1$ is a row vector with components $B_{11}, B_{12}\ldots\ldots\ldots, B_{1,N+1}$; $B_{1i} = T^0 \otimes \begin{bmatrix} \alpha & e_j & e_j & . & . & e_j \end{bmatrix}^T \otimes I_r$, $1 \le i \le N + 1$; records transition rates corresponding to repeat/resumption of pre-empted low priority customer's service on completion of a high priority customer's service, where e_i is a column vector of order 1 in the i^{th} place and zero elsewhere.

For $1 \le m \le K$, the matrices $T^0 \otimes \beta \otimes I_r$, $I_b \otimes p_2D_1$ records transition rates to states in $\Psi(n,m-1,.)$ and $\Psi(n,m+1,.)$ respectively, from states in $\Psi(n,m,.)$; $S_2 \oplus D_0$ lists transition rates to states in $\Psi(n,m,.)$ from $\Psi(n,m,.)$ for $1 \le m \le K-1$ and $S_2 \oplus D_0 \oplus p_2D_1$ for $m = K$.

The infinitesimal generator of the Markov chain governing the system is given by

$$Q=\begin{bmatrix} C_0 & C_1 & 0 & 0 & 0 & 0 & 0 \\ C_2 & A_1 & A_0 & 0 & 0 & 0 & 0 \\ & A_2 & A_1 & A_0 & 0 & 0 & 0 \\ & & A_2 & A_1 & A_0 & 0 & 0 \\ & & & & . & . & . \end{bmatrix} \qquad \rightarrow (1)$$

The matrix $C_0 = \left[C_0^{(i,j)}\right]$ represents a square matrix of order (N+1)r(b+1) which corresponds to transition from i to j; $0 \le I, j \le K$. The matrices D_0, $\beta \otimes p_2D_1$, $T^0 \otimes I_r$ specifies elements of $C_0^{(0,0)}$, $C_0^{(0,1)}$ and $C_0^{(1,0)}$, respectively. $I_b \otimes p_2D_1$ provides the elements of $C_0^{(i,i+1)}, 1 \le i \le K+1$ and $T^0 \otimes \beta \otimes I_r$ lists the elements of $C_0^{(i,i-1)}, 2 \le i \le K$, while $S_2 \oplus D_0$ records the transitions $C_0^{(i,i)}, 1 \le i \le K-1$, $S_2 \oplus D_0 \oplus p_2D_1$ corresponds to the transition rates in $C_0^{(i,i)}$, $i = K$.

The only non zero block in C_1 are the transition from $\Psi(n,m,.)$, n=m=0 to $\Psi(1,m,2)$;

$\Psi(0,m,1)$ to $\Psi(1,m,1)$ and are denoted by $C^{(1)}$, $C_1^{(2)}$ respectively. Here $C_1^{(1)} = \left[C_1^{(11)} \quad [0]\right]$ where $C_1^{(11)} = \left[E_1 E_2 \ldots E_N\right]$ with $E_i = q_{i-1}e_i \otimes \alpha \otimes p_1D_1$. e_i is a $1 \times i$ row vector having 1 in the i^{th} place and zero elsewhere. The matrix $C_1^{(2)} = \begin{bmatrix} I_K \otimes I_b \otimes p_1D_1 & [0] \end{bmatrix}$ records arrival of a low priority customer when server's state is 2

The matrix $C_2 = \begin{bmatrix} [0] & C_2^{(1)} \end{bmatrix}^T$ where block [0] indicates no service completion of low priority customer during state 0 and 1 of the server. The matrix $e_{(N+1)(N+2)/2} \otimes S^0 \otimes I_r$ in $C_2^{(1)}$ records transition from $\Psi(n,0,2)$ to $\Psi(n,0,2)$, if $n = 1$ and $I_K \otimes e_N \otimes (S^0 \otimes \beta \otimes I_r)$ lists the transition rates to $\Psi(0,m,1)$ from $\Psi(1,m,2)$, $1 \le m \le K$.

The matrix A_0 records arrival of low priority customers in to the system where the only non zero elements are diagonal ones. The matrices $I_N \otimes I_{N(N+1)/2} \otimes I_a \otimes I_b \otimes I_{(c+1)} \otimes p_1D_1$, $IK \otimes Ib \otimes p1D1$ and $I((N+1)(N+2)/2)+Kd \otimes Ia \otimes p1D1$ in $A1$ lists the arrival of low priority customer within states $\Psi(n,m,0)$, $\Psi(n,m,1)$ and $\Psi(n,m,2)$.

The matrix A_2 in Q lists the service completion of low priority service. The matrices $A_2^{(1)} = \begin{bmatrix} [0] & I_K \otimes e_N \otimes S^0 \otimes \beta \otimes I_r \end{bmatrix}'$ and $A_2^{(2)} = \begin{bmatrix} e_{(N+1)(N+2)/2} \otimes B & [0] \end{bmatrix}$ where

$B = (B_{1i})^{T}_{1\times N+1}, i = 1,2,\ldots,N+1$ is a column vector of $B_{11}, B_{12}, \ldots, B_{1N+1}$; $B_{1i} = S^{0} \otimes q_{i-1} e_i \otimes \alpha \otimes I_r, 1 \le i \le N+1$ records transition rates in $\Psi(n-1,0,2)$, $\Psi(n-1,m,2)$ of A_2, starting from states in $\Psi(n,0,2)$ and $\Psi(n,m,2)$ respectively.

The matrix A_1 in Q records transition from $\Psi(n,m,l)$ to itself. The components in A_1 are A_{11}, A_{12}, A_{13}, A_{14}, A_{15} and A_{16} each of which records transition rates from states in $\Psi(n,m,0)$ to $\Psi(n,m,0)$; $\Psi(n,m,0)$ to $\Psi(n,m,1)$; $\Psi(n,m,1)$ to $\Psi(n,m,1)$; $\Psi(n,m,1)$ to $\Psi(n,m,2)$; $\Psi(n,m,2)$ to $\Psi(n,m,0)$ and $\Psi(n,m,2)$ to $\Psi(n,m,2)$.

(a) The matrix A_{11} is as follows: $I_{d(d+1)/2} \otimes I_a \otimes H$, where $H = G_1 \oplus G_2, G_1 = F \oplus D_0$,

$F = \begin{bmatrix} \bar{0} & 0 \\ U & U^0 \end{bmatrix}$, $G_2 = T \otimes I_{c+1}$ records transitions to $\Psi(n,m,0)$ from $\Psi(n,m,0)$,

$1 \le m \le N-1$, $I_{d(d+1)/2} \otimes I_a \otimes I_b \otimes I_{(c+1)} \otimes p_1 D_1$ records transition rates from $\Psi(n,m,0)$ to $\Psi(n,m+1,0)$, $1 \le m \le K-1$ and $I_{N(N+1)/2} \otimes I_a \otimes (H \oplus p_1 D_1)$ If $m = K$. $I_{N(N+1)/2} \otimes I_a \otimes T^{(0)} \otimes \beta \otimes I_{c+1} \otimes I_r$ records transition rates from $\Psi(n,m,0)$ to $\Psi(n,m-1,0)$, $2 \le m \le K$.

The matrix A_{12} records transitions in $\Psi(n,m,2)$ from $\Psi(n,m,0)$ and is as follows:

The matrix $\left[\ [0] \quad diag\left[\left(\ [0] \quad diag(I_j B)\ \right)\right]\ \right], 1 \le j \le N$, $diag\left[\left(\ [0] \quad diag(I_j B)\ \right)\right]$

denote a diagonal matrix whose i^{th} diagonal element is $([0] \quad diag(I_j B)\ )$ and $diag(I_j B)$ is a diagonal matrix whose j diagonal element is $I_j B$, I_j is the identity matrix of order j, records transition from $\Psi(n,1,0)$ to $\Psi(n,0,2)$. The matrix $diag[e_i B], 2 \le i \le K$ is a diagonal matrix with i^{th} diagonal element $e_i B$ where e_i is a column vector of order i with 1 in the i^{th} place and 0 elsewhere, records transition from $\Psi(n,m,2)$ to $\Psi(n,m-1,2)$, $2 \le m \le K$.

(c)The matrix A_{13} lists transition rates in $\Psi(n,i,1)$ to $\Psi(n,j,1)$, $1 \le i,j \le N$. $Diag[S \oplus D_0]$ records transition rates in $\Psi(n,m,1)$ from $\Psi(n,m,1)$, $1 \le m \le N-1$ and $S \oplus D_0 \oplus p_1 D_1$ if $m = N$. $diag[I_b \otimes p_2 D_1], 1 \le i \le N-1$ lists transition rates in $\Psi(n,m+1,1)$ from $\Psi(n,m,1)$ and $diag[T^0 \otimes \beta \otimes I_r]$ corresponds to transition rates to $\Psi(n,m-1,1)$ from $\Psi(n,m,1)$, $2 \le i \le K$.

(d)The matrix A_{14} records transition rates in $\Psi(n,i,2)$ from $\Psi(n,j,1)$, $1 \le i,j \le K$ where the transition to $\Psi(n,0,2)$ from $\psi(n,1,1)$ is $A_{14}^{(1)} = \left[W_{11} W_{12} \ldots W_{1(N+1)}\right]$ and $W_{1i} = T^0 \otimes q_{i-1} e_i \otimes \alpha \otimes I_r$ where e_i is a row vector of order i with 1 in the 1^{st} place and zero elsewhere and other transition in A_{14} are [0] block matrices.

(e)The matrix A_{15} in A_1 records transition rates in $\Psi(n,1,0)$ from $\Psi(n,0,2)$ and is described as follows: $A_{15} = \left[[0] \quad A_{15}^{(1)}\right]^T$ and $A_{15}^{(1)} = diag(e_i(I_a \otimes \beta \otimes \bar{\delta} \otimes p_2 D_1)), 1 \le i \le N$.

(f)The matrix A_{16} records transition rates to $\Psi(n,j,2)$ from $\Psi(n,i,2)$, $0 \le i,j \le K$, where the matrix $I_{(N+1)(N+2)/2} \otimes (S \oplus D_0)$ lists transition rates in $\Psi(n,0,2)$ from $\Psi(n,0,2)$. $I_N \otimes (S \oplus D_0)$ lists transition rates within $\Psi(n,m,2)$ for $1 \le m \le K-1$ and $I_N \otimes (S \oplus D_0)$ if m=K while $I_N \otimes I_a \otimes p_2 D_1$ records transition in $\Psi(n,m+1,2)$ from $\Psi(n,m,2)$, $1 \le m \le K-1$.

3. Description of the phase type distribution for the services

The focus of this section is to describe the time it takes to process a job once it enters into the service facility. We assume that the service times are of phase type with representation given by (α, S) of order a. The services are subject to pre-emptions. When the current service is pre-empted for the first time, counting clocks, which counts the number of priority customers served during his service period, pre-emption time (service time of current high priority customer), and threshold clock, respectively, will simultaneously be started. The pre-emption clock and threshold clock are of phase type with representations given by, respectively, (β, T) of order b, and (δ, U) of order c. Once the high priority queue becomes empty/the number of high priorities during the customer's service period reaches its maximum allowed level, whichever occurs first, the service of the pre-empted job will begin again. The service will resume (from the phase where the service got interrupted) or repeat (like a new service) depending on whether the interruption clock expired before the threshold clock or not. In addition, if the number of pre-emptions during the customers service reaches its maximum allowed level, then the service of the current job will not be pre-empted anymore once the service begins

again for this job. On the other hand, if the pre-emption clock expires before the number of pre-emption reaches its maximum, the counting end temporarily and will resume from this phase should there be an another interruption for the existing job. For the job under service, the number of interruptions will be tracked and when this number attains a pre-specified threshold value, $N < \infty$, no further interruptions are allowed.

Following the procedure indicated in the paper titled "A note on characterizing interruptions with Phase-type distributions" by A.Krishnamoorthy, P.K Pramod and S.R. Chakravarthy [12] we will be able to compute the phase type distribution governing the effective service rate. Mean of this phase type distribution can be computed in the usual manner. Thus the system is stable if and only if arrival rate less than effective service rate.

3.1. Stability Condition

Next we examine the system stability. We can anticipate that a very strong condition is needed here for the same since a service can get interrupted several times. What is needed is that the rate of drift to any lower level from a given level should be higher than that to a higher level. This means that the Markov chain is stable iff

$$\Pi A_0 e < \Pi A_2 e \qquad (2)$$

where Π is the unique solution to $\Pi A = 0, \Pi e = 1$ where $A = A_0 + A_1 + A_2$. The above condition implies that the arrival rate should be less than the effective service rate (reciprocal of the expected time to completely serve a customer).

3.2. Stationary Distribution

Denote by x the stationary vector of $X(t)$, and partition x in to sub vectors $x(n,m,l), 0 \le n \le \infty, 0 \le m \le K, l = 0,1,2.$ and satisfying the condition xQ=0 and xe=1. The vectors x(0) and x(1) are obtained by solving the equations

$$x(0)C_0 + x(1)C_2 = 0,\ x(0)C_1 + x(1)(A_1 + RA_2) = 0 \qquad (3)$$

subject to the normalizing condition

$$x(0)e + x(1)(I - R)^{-1}e = 1 \qquad (4)$$

where R is the minimal non-negative solution to the matrix equation

$$A_0 + RA_1 + R^2A_2 = 0. \qquad (5)$$

From these results, we obtain some interesting measures which helps in design of the system. Some of them are as follows:

- The mean number of units in the system, $E_s = \sum_{n=0}^{\infty} n\ x(n).$
- The mean number of high priority units in the system $= \sum_{n=0}^{\infty}\sum_{m=0}^{K} mx(n,m)$
- The fraction of time the server is idle $= \sum_{l=1}^{r} x(0,0,l)$
- Fraction of time the low priority customer is pre-empted $= \sum_{n=1}^{\infty}\sum_{m=1}^{K} x(n,m,0)$
- Fraction of time the server is busy with high priority customer (with no pre-empted customer in the system) $= \sum_{n=1}^{\infty}\sum_{m=1}^{K} x(n,m,1)$
- Fraction of time the server is busy with high priority customer (with pre-empted customer in the system)

$$= \sum_{n=1}^{\infty}\sum_{m=1}^{K} x(n,m,2)$$

- Fraction of time the customer is busy with low priority customer

$$= \sum_{n=1}^{\infty}\sum_{j_1=0}^{N}\sum_{j_2=0}^{j_1}\sum_{i=1}^{a}\sum_{l=1}^{r} x_{n,0,2,j_1,j_2,i,l} + \sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\sum_{i_1=1}^{N}\sum_{i=1}^{a}\sum_{l=1}^{r} x_{n,m,2,i_1,i,l}$$

- Effective service rate of low priority customer

$$= \sum_{n=1}^{\infty}\sum_{j_1=0}^{N}\sum_{j_2=0}^{j_1}\sum_{i=1}^{a}\sum_{l=1}^{r} x_{n,0,2,j_1,j_2,i,l} S_i^{\,0} + \sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\sum_{i_1=1}^{N}\sum_{i=1}^{a}\sum_{l=1}^{r} x_{n,m,2,i_1,i,l}\, S_i^{\,0}$$

4. Numerical Results

For the above input parameters we have plotted 4 graphs in Fig. 1. These represent variations in the mean number of customers in the system against increasing the value of the probability of customers of that priority.

K=3, N=2, a=b=c=r=2; $D_0 = \begin{bmatrix} -6.5 & 0.25 \\ 0.25 & 0.75 \end{bmatrix}, D_1 = \begin{bmatrix} 6.0 & 0.25 \\ 0.25 & 0.25 \end{bmatrix}, S = \begin{bmatrix} -12.0 & 6.0 \\ 6.0 & -12.0 \end{bmatrix},$

$T = \begin{bmatrix} -12.0 & 3.0 \\ 3.0 & -12.0 \end{bmatrix}, U = \begin{bmatrix} -12.0 & 8.0 \\ 8.0 & -12.0 \end{bmatrix},$

$S^0 = \begin{bmatrix} 6.0 & 6.0 \end{bmatrix}', T^0 = \begin{bmatrix} 9.0 & 9.0 \end{bmatrix}', U^0 = \begin{bmatrix} 4.0 & 4.0 \end{bmatrix}',$
$\alpha = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}, \beta = \begin{bmatrix} 0.3 & 0.7 \end{bmatrix}, \delta = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}.$

Depending on the outcome of the toss (the decision of the low priority customers to allow none, one or two priority customers to be served during its effective service time), we have the four graphs given in Fig. 1. (i) $q_0 = 1; q_1 = q_2 = 0$ (ii) $q_0 = 0.8, q_1 = 0.2, q_2 = 0$ (iii) $q_0 = 0.6; q_1 = q_2 = 0.2$ (iv) $q_0 = q_1 = 0.33; q_2 = 0.34$. At $p_1 = 0.2$ all the above result in almost the same mean number of customers; at $p_1 = 1.0$ (no high priority customer turns up) all these have the same value, which is not surprising.

Fig. 2 provides the fraction of time the sever is busy with high priority customers with increasing value of p_1 in the four cases indicated in Fig. 1 . As expected when $p_1 = 1$ (when all arrivals are of low priority), the fraction of time the server is busy serving low priority, turns out to be the maximum. Similarly the fraction of time the server is busy serving high priority customers decrease with increase in p_1 value (see Fig. 2).

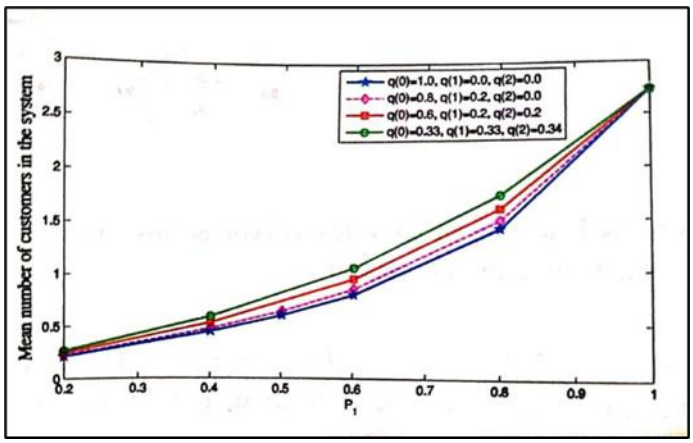

Figure 1 p_1 versus Mean Number of Customers in the System

An unexpected behavior of effective service rate of low priority customer versus p_1, when $p_1 > 0.8$ is seen in fig 3 in the case $q_0 = 1.0$ and $q_0 = 0.6$, $q_1 = q_2 = 0.34$. It decreasing for increasing p_1 in the range (0.8,1.0). The four graphs in Fig.4 are on expected lines.

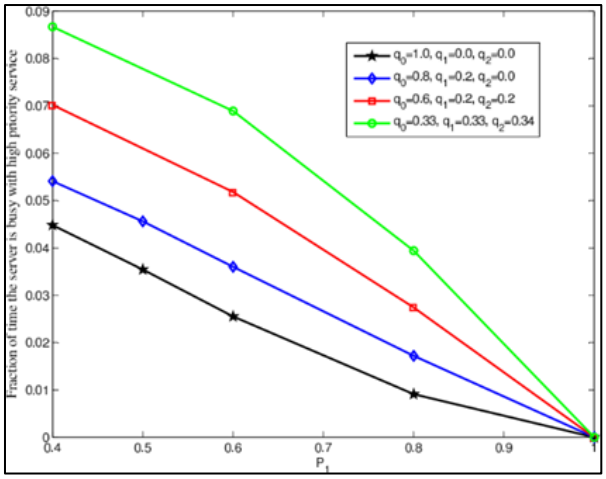

Figure 2 p_1 versus Fraction of time the server is busy with high priority service

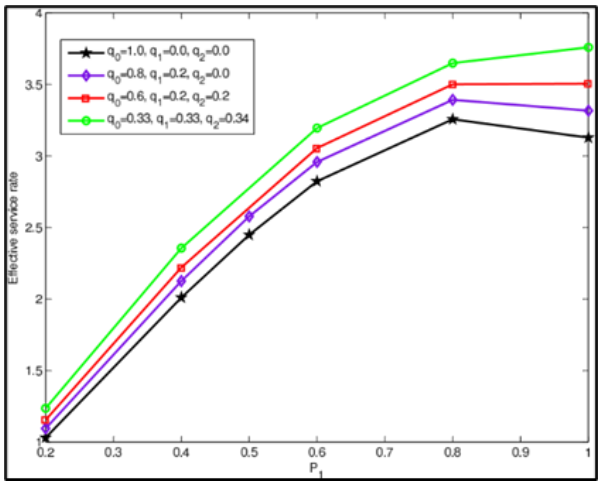

Figure 3 p_1 versus Effective service rate of low priority customers

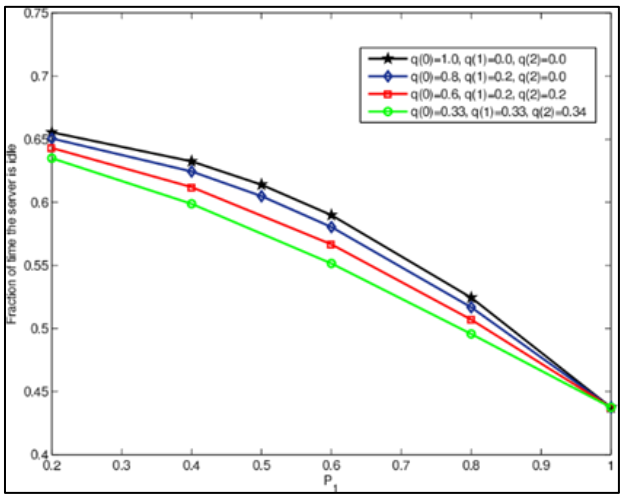

Figure 4 p_1 versus fraction of time the server is idle

5. Conclusion

In this paper we investigated a continuous time queueing system where interruption as the server processing a high priority customer. The low priority customer's service is pre-empted subject to some conditions. The "maximum number of pre-emptions" and "pre-emption clock" controls the system. The purpose of introducing threshold clock is to decide whether the service to be repeated or resumed on completion of interruption. Such Queueing systems are very common in real world and their modelling will help to improve their performance. Some numerical illustrations are provided for the measures that are investigated.

Compliance with ethical standards

Disclosure of conflict of interest

Since there was no financial assistance from any agency for this research work, the authors declare no conflict of interest.

References

[1] Aissani A and Artalejo J R 1998 On the single server retrial queue subject to breakdowns, *Queueing Systesm: Theory and Applications*, v30 n, 3-4, 309-321.

[2] Alfa A S 2007 Discrete time queues and matrix-analytic methods, *TOP, Springer publication*, 147-185.

[3] Atencia I and Moreno P 2006 A Discrete-Time Geo/G/1 retrial queue with the server subject to starting failures, *Annals of OR*, 141:85-107.

[4] Fiems D, Maertens T and Brunee H 2008 Queueing systems with different types of interruptions, *European Journal of Operations Research*, 188 (3), 838-845.

[5] Gaver D 1962 A waiting line with interrupted service including priority, *Journal of Royal Statistical Society*, B24, 73-90 .

[6] Ibe O and Trivedi K S 1990 Two Queues with alternating service and server breakdown, *Queueing Systems*, Vol.7(3-4), 253-268.

[7] Julian Keilson 1962 Queues Subject to Service Interruptions, *The Annals of Mathematical Statistics*.

[8] Heathcote, C R 1959 The time-dependent problem for a queue with preemptive priorities, *Operations Research*, 670-680.

[9] Karlin S and Taylor H E 1975 A first course in Stochastic Processes, 2nd edition, *Elsevier*.

[10] Krishnamoorthy A, Pramod P K and Deepak T G 2009 On a queue with interruptions and repeat or resumption of service. *Nonlinear Analysis, Theory, Methods and Applications*, 71 (12), 1673-1683.

[11] Krishnamoorthy A, Pramod P K and Chakravarthy S R 2014 A survey on queues with interruptions, *TOP* , 22,290-320.

[12] Krishnamoorthy A, Pramod P K and Chakravarthy S R 2013 A Note on Characterizing Service Interruptions with Phase-Type Distribution, *Stochastic Analysis and Applications*,Taylor and Francis, 31 (4), 671-683.

[13] Marcus M and Minc H 1964 A Survey of matrix theory and matrix inequalities, *Allyn and Bacon*, Boston, MA.

[14] Neuts M F 1981 Matrix-Geometric solutions in stochastic models: An algorithmic approach, *The Johns Hopkins University Press*, Baltimore, MD.

[15] Tewfik Kernane 2009 A Single Server Retrial Queue with Different Types of Server Interruptions,*HAL Open Science*.

[16] White H and Christie L S 1958 Queuing with Preemptive Priorities or with Breakdown, *Operational Research*. Vol. 6, pp. 79-95. .