ADMISSION CONTROL FOR FINITE CAPACITY QUEUEING MODEL WITH GENERAL RETRIAL TIMES AND STATE-DEPENDENT RATES

Madhu Jain and Sudeep Singh Sanga
Indian Institute of Technology Roorkee - 247 667, India

(Communicated by Ryan Loxton)

Abstract. The finite state dependent queueing model with $F$-policy is investigated by considering the general retrial attempts. On arrival in the system, if the job finds the server engaged, it is forced to enter into the retrial orbit. After a random period of time, the job from the retrial orbit re-attempts for the service. According to $F$-policy, as the system attains its full capacity, the arrivals are restricted to join the system until the number of jobs comes down to the prefixed threshold value $'F'$. The supplementary variable corresponding to the remaining retrial time is used to frame the governing equations which are solved by using Laplace-Stieltjes transform and then applying the recursive method. Special models for machine repair and time-sharing queue are deduced by setting the state dependent rates. Several system indices are obtained explicitly which are further used to facilitate the sensitivity analysis by considering a numerical illustration. A cost function is constructed and minimized for evaluating the optimal threshold parameter and optimal service rate.

1. Introduction. In our routine life, the formation of queues can be seen everywhere such as at ATMs, bank counters, railway reservation windows, cafeteria, and at many other places. The formation of queues and consequently delay in the service is the major problem for both customers as well as system organizers. In order to maintain the smooth functioning of the service systems where long queues are built up, the arrivals/service should be controlled [24] and this can be done by implementing the optimal admission control policy. Optimal $F$-policy was first introduced by Gupta [13] to restrict the customers from an entry in the Markovian queueing system when the system capacity is exhausted. So far as the controlling of the arrivals in the finite capacity system is concerned, $F$-policy is quite useful to control the congestion of the jobs and helps in reducing the lost jobs in particular when the system capacity is full. In $F$-policy, the admission of the jobs is to be restricted due to the limitation of buffer size and it can be compensated by providing an opportunity to join the system again by pushing the jobs in the retrial orbit. According to the $F$-policy, as soon as the system reaches to its full capacity $K$ (i.e., when no more jobs can be accommodated), no further jobs are permitted to join in the system until enough jobs get served so that the number of jobs in the system drops to a threshold level $'F'$ ($0 \leq F < K - 1$).

2020 Mathematics Subject Classification. 60K25, 68M20, 90B22.

Key words and phrases. Finite queue, general retrial, $F$-policy, supplementary variable, sensitivity analysis, cost optimization.
The F-policy can be employed to resolve the issue of controlling of the arrivals in the queueing system so as to avoid the loss of revenue and inconvenience to the jobs. Based on F-policy, non-Markovian M/G/1/K and G/M/1/K queueing models were investigated by Wang et al. [36, 35]. The performance analysis of unreliable M/M/2/K queueing model by incorporating both F-policy and N-policy was developed by Jain et al. [20]. In recent years, some researchers have contributed to the controlling of the admission of jobs in the system based on F-policy [14, 6, 39, 7, 15, 18]. Yeh et al. [41] developed a finite capacity single server Markovian queueing model under the $\langle p - F \rangle$ policy and established the steady-state analytical formulae for various performance indices. The machine repair problem was investigated by Jain and Sanga [16, 17] by considering the arrivals of failed machines to the repair facility as per F-policy.

In many queueing models, the arrival and service rates may be dependent on the queue length of the system. The importance of the state dependent queue may be realized in many day-to-day queueing scenarios including machine repair systems and time-sharing systems. The state-dependent queueing model may also be applicable to the queueing system in which the decision-maker can facilitate the additional servers on observing the long queue based on threshold policy. In queueing literature, a few researchers have developed the state dependent queueing models under different assumptions. Massey [29] considered the time-varying rates to explore the performance of a queueing system and discussed the application of the developed queueing model in a telecommunication system. A single server queueing system in which the arrival and service rates depend on the system states, was analyzed by Adan and Kulkarni [1]. An important work on M/M/1 retrial queue was done by Parthasarathy and Sudhesh [31] by considering the state dependent rates. Lee [27] developed the state-dependent stochastic networks using the birth-death process and established the different moments and stability properties of the system. A single server bulk queueing model using threshold policy was studied by Banerjee and Gupta [3] by including the features of controlling the arrivals and batch service schedule. Kumara and Dharsana [26] analyzed the congestion problem by developing a single server Markovian model with queue size dependent arrival rate and impatient customers. Recently, an M/M/1 queueing model with queue size dependent service rate was studied by Rodrigues [33].

Based on arrival and or service rates, the state dependent queueing models can be further classified into different categories including the queueing models with the additional removable server, queue with discouragement, machine interference models, time-sharing models, etc. The failure of machines is a major difficulty not only for the users but also the loss of revenue to the organizers/manufactures in the concerned machining system. In such cases, to overcome the problems of machine failures and delay in production, many researchers contributed towards finite population models which also dealt with machine repair problems under different assumption [37, 34, 28, 23]. Sometimes, it is seen that due to less workload, the server may remain idle most often which is the wastage of revenue as well as time. The concept of time sharing is also used in a few scenarios including computer communication networks, manufacturing and production systems, etc. Many researchers have paid their attention to time-sharing models for computer systems [2, 21, 25]. The online optimization issues of machining system used for cloud computing were examined by Chandrasekaran et al. [5]. In recent years, for manufacturing–remanufacturing systems, the time sharing machining system was studied by Flapper et al. [11]. Jain
et al. [19] investigated the machine repair problem with two types of spares along with reneging and controlled rates by applying the concept of time sharing. Jain et al. [22] investigated a time-shared machine repair problem with mixed spares. In this investigation, they have considered that the caretaker of failed machines operates under $N$-policy.

In several real-life day-to-day as well as in industrial/business queueing scenarios, the jobs arriving in the system may be compelled to leave the service area and move to the retrial pool in the case when the server is busy. The jobs from the retrial pool can try again and again for the service after a random period of time so as to avail the service. The applications of queueing models with retrial orbit can be seen in many places/organizations including the call centers, telecommunication networks, mobile communication system, and service/manufacturing systems, etc. To highlight the practical utility of finite population retrial queues, we cite the queue formed at the shopping center wherein the arriving jobs on finding the busy server, may wait in the retrial pool and return back after some time with the hope that the server becomes free. The supplementary variable technique (SVT) developed by Cox [10] to analyze the M/G/1 queueing model has been extensively implemented by several researchers to study the non-Markovian queueing system in different frameworks. Some researchers have contributed to the analysis of retrial queueing models in different structures [8, 32]. An M/G/1 retrial queueing model was investigated by Moreno [30] by considering general retrial times. The author also presented the condition of ergodicity for the system and established analytical results for the stationary distribution and other performance measures. Gao et al. [12] developed an M/G/1 queueing model by considering the general distributed retrial times, working vacations and interruptions due to server breakdown. They have obtained the stationary state probability distribution by using the supplementary variable method. To obtain the queue size distribution and probability generating function (PGF) of the joint distributions of the queue size, M/G/1 queue with general distributed retrials times and Bernoulli vacation was dealt by Choudhury and Ke [9]. A stochastic comparison of Markov chains was proposed by Boualem et al. [4] for the study of single server queue with retrial times as general distributed. In the recent work of Yang et al. [38], the unreliable server retrial queue with general distributed retail attempts was studied by employing the supplementary variable method to establish several performance measures.

After literature survey, it is found that based on $F$-policy, no research work has been done on the optimal control of the arrivals of the jobs for the finite time sharing/machine repair queueing model with general repeated attempts. In this paper, we develop a finite queueing model in generic set up by considering many realistic features such as admission control policy, general retrial times and state-dependent arrival and service processes. To analyze the retrial model under $F$-policy, this article is arranged in different sections. Section 2 presents a model description of the concerned problem. In Section 3, equations for the non-Markovian model are framed by using supplementary variable corresponding to remaining of retrial time. Some special models deduced from our study are given in Section 4. Various system indices and the cost structure are established in Section 5. By taking the appropriate illustration, numerical experiment and sensitivity analysis of machine repair problem, as well as time sharing model and cost optimization, are presented in Section 6. Finally, Section 7 presents the conclusion of the investigation done.
2. Model description. For the queueing scenario with admission control according to $F$-policy, we consider a single server finite capacity (say K) queueing model with general retrial attempts. The service discipline for rendering the service to the jobs follows the first-come-first-served (FCFS) rule. The formation of the model is based on certain assumptions which are outlined as follows:

- The jobs join the system according to Poisson fashion with parameter $\lambda$.
- If the arriving job finds the server free, the job gets served according to an exponential distribution with rate $\mu$.
- If the server is occupied then the arriving job joins the retrial pool. From the orbit, the job re-attempts for the service with general distributed retrial time having probability distribution $G(x)$ ($x \geq 0$) with $G(0) = 0$, the probability density function $\gamma(x)$, and mean retrial time $1/\gamma$.
- When the system attains its full capacity, then setup time is required to stop the arriving jobs from joining the queue; the time required for the setup is assumed to exponentially distributed with mean $1/\varepsilon$.
- Once the system becomes full, the arrivals are restricted from joining the system. The further admission of jobs in the system is permitted when the number of jobs in the system ceases to a prefixed threshold value $F$ ($0 \leq F < K - 1$).

For developing the state-dependent model for the retrial queueing system, we denote the state-dependent arrival and service rates by $\lambda_n$ and $\mu_n$, respectively. The supplementary variable ($U$) is used corresponding to the remaining retrial time of the jobs while residing in the retrial pool. At the time $\tau$, $N(\tau)$ denotes the number of jobs present in the system. The status of the server at a time $\tau$ is denoted by $Y(\tau)$. To formulate the mathematical model, the random variable $Y(\tau)$ is defined as follows:

$$Y(\tau) = \begin{cases} 
0, & \text{the jobs are compelled to join the retrial pool on finding the server being busy}, \\
1(2), & \text{the server is occupied and the jobs are permitted (not permitted) to enter in the system}. 
\end{cases}$$

(1)

The system state probabilities at time epoch $\tau$ are as follows:

$$\begin{align*}
P_{0,0}(\tau) &= \Pr\{Y(\tau) = 0, N(\tau) = 0\} \\
P_{0,n}(x, \tau) dx &= \Pr\{Y(\tau) = 0, N(\tau) = n, x < U(\tau) \leq x + dx\}, \\
& \quad x \geq 0, 1 \leq n \leq K - 1 \\
P_{1,n}(\tau) &= \Pr\{Y(\tau) = 1, N(\tau) = n\}, 0 \leq n \leq K - 1 \\
P_{2,n}(\tau) &= \Pr\{Y(\tau) = 2, N(\tau) = n\}, F \leq n \leq K - 1
\end{align*}$$

(2)

We denote $P_{0,n}(\tau) = \Pr\{Y(\tau) = 0, N(\tau) = n\} = \int_0^\infty P_{0,n}(x, \tau) dx$, $1 \leq n \leq K - 1$.

At steady state, i.e., when $\tau \to \infty$, we define:

$$\begin{align*}
P_{0,0} &= \lim_{\tau \to \infty} P_{0,0}(\tau), \\
P_{0,n}(x) &= \lim_{\tau \to \infty} P_{0,n}(x, \tau), 1 \leq n \leq K - 1, \\
P_{1,n} &= \lim_{\tau \to \infty} P_{1,n}(\tau), 0 \leq n \leq K - 1, \\
P_{2,n} &= \lim_{\tau \to \infty} P_{2,n}(\tau), F \leq n \leq K - 1
\end{align*}$$

3. Governing equations and queue size distribution. To establish the steady state probabilities of the system state space, the governing equations for three levels (i.e., when $Y(\tau) = i = 0, 1, 2$) of the non-Markovian model are constructed by
introducing the supplementary variable corresponding to remaining of retrial time. The in-flows and out-flows of system states \((n, i)\) are depicted in the transition diagram shown in Figure 1.

![State transition diagram](image)

FIGURE 1. State transition diagram. \((n : \text{number of jobs present in the system}, i : \text{status of the server})\)

Now by using the probability arguments, Chapman-Kolmogorov equations are formulated as follows:

(i) For \(i = 0\):

\[
\lambda_0 P_{0,0} = \mu_1 P_{1,0}
\]

\[
- \frac{d}{dx} P_{0,n}(x) = -\lambda_n P_{0,n}(x) + \mu_{n+1} P_{1,n} \gamma(x), \quad 1 \leq n \leq K - 1.
\]

(ii) For \(i = 1\):

\[
(\lambda_1 + \mu_1) P_{1,0} = \lambda_0 P_{0,0} + P_{0,1}(0)
\]

\[
(\lambda_{n+1} + \mu_{n+1}) P_{1,n} = \lambda_n P_{1,n-1} + \lambda_n P_{0,n} + P_{0,n+1}(0), \quad 1 \leq n \leq F - 2
\]

\[
(\lambda_F + \mu_F) P_{1,F-1} = \lambda_{F-1} P_{1,F-2} + \lambda_{F-1} P_{0,F-1} + P_{0,F}(0) + \mu P_{2,F}
\]

\[
(\lambda_{n+1} + \mu_{n+1}) P_{1,n} = \lambda_n P_{1,n-1} + \lambda_n P_{0,n} + P_{0,n+1}(0), \quad F \leq n \leq K - 2
\]

\[
(\varepsilon + \mu_K) P_{1,K-1} = \lambda_{K-1} P_{1,K-2} + \lambda_{K-1} P_{0,K-1}
\]

(iii) For \(i = 2\):

\[
\mu P_{2,n+1} = \mu P_{2,n+2} = \varepsilon P_{1,K-1}; \quad F \leq n \leq K - 3
\]

Define Laplace-Stieltjes transform (LST) of \(\gamma(x)\) and \(P_{0,n}(x)\) by \(\gamma^*(\theta)\) and \(P_{0,n}^*(\theta)\), respectively.

Also, \(P_{0,n}^*(\theta) = P_{0,n} \gamma^*(\theta)\)

The symbols \(S_n\) and \(R_F\) are used for the brevity of notations and are defined as follows:

\[
S_n = \varepsilon \left( \prod_{i=n+2}^{K-1} \lambda_i \right) + \varepsilon \sum_{i=n+3}^{K} \left( \prod_{i=i}^{K-1} \lambda_i \right) \left( \prod_{j=n+2}^{i-1} \mu_j \gamma^*(\lambda_{j-1}) \right) + \left( \prod_{j=n+2}^{K} \mu_j \gamma^*(\lambda_{j-1}) \right)
\]

\[
R_F = \mu_1 \left( \prod_{j=1}^{F-1} \mu_{j+1} \gamma^*(\lambda_j) \right) \left( \varepsilon \left( \prod_{i=F+1}^{K-1} \lambda_i \right) + \mu_{F+1} \gamma^*(\lambda_F) S_F \right).
\]
Theorem 3.1. The steady state queue size distribution for the state dependent retrial model operating under $F$-policy is given by

\[
P_{0,n} = \begin{cases} 
\frac{\lambda_0}{\mu_1} \left( \frac{1}{\gamma^*(\lambda_1)} - 1 \right) P_{0,0}, & n = 1, \\
\left( \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}} \right) \left( \frac{n-1}{\prod_{j=1}^{n-1} \gamma^*(\lambda_j)} \right) \left( \frac{1}{\gamma^*(\lambda_n)} - 1 \right) P_{0,0}, & 1 < n \leq F - 1, \\
\left( \prod_{i=0}^{n} \lambda_i \right) \frac{\mu_{n+1}(1 - \gamma^*(\lambda_n))}{\lambda_n} \frac{S_n}{R_F} P_{0,0}, & F \leq n \leq K - 2, \\
\left( \prod_{i=0}^{n} \lambda_i \right) \frac{\mu_{n+1}(1 - \gamma^*(\lambda_n))}{\lambda_n} \frac{1}{R_F} P_{0,0}, & n = K - 1.
\end{cases}
\] (12)

\[
P_{1,n} = \begin{cases} 
\frac{\lambda_0}{\mu_1} P_{0,0}, & n = 0, \\
\left( \prod_{i=0}^{n} \frac{\lambda_i}{\mu_{i+1}} \right) \left( \frac{1}{\prod_{j=1}^{n} \gamma^*(\lambda_j)} \right) P_{0,0}, & 1 \leq n \leq F - 1, \\
\left( \prod_{i=0}^{n} \lambda_i \right) \frac{S_n}{R_F} P_{0,0}, & F \leq n \leq K - 2, \\
\left( \prod_{i=0}^{n} \lambda_i \right) \frac{1}{R_F} P_{0,0}, & n = K - 1.
\end{cases}
\] (13)

\[
P_{2,n} = \left( \prod_{i=0}^{K-1} \lambda_i \right) \frac{\varepsilon}{\mu_{RF}} P_{0,0}, & F \leq n \leq K - 1.
\] (14)

Proof. Taking LST on both sides of (4), we obtain

\[(\lambda_n - \theta) P_{0,n}^*(\theta) = \mu_{n+1} P_{1,n} \gamma^*(\theta) - P_{0,n}(0), \quad 1 \leq n \leq K - 1.\] (15)

Using (15), we have

\[
P_{1,n} = \frac{1}{\mu_{n+1} \gamma^*(\lambda_n)} P_{0,n}(0) \quad \text{and} \quad P_{1,n} = \frac{\lambda_n}{\mu_{n+1}(1 - \gamma^*(\lambda_n))} P_{0,n}. \] (16 a-b)

Using (3),

\[
P_{1,0} = \frac{\lambda_0}{\mu_1} P_{0,0}. \] (17)

Now (17) and (5) yield

\[
P_{0,1}(0) = \frac{\lambda_0 \lambda_1}{\mu_1} P_{0,0}. \] (18)

Using (16a) and (18), we obtain

\[
P_{1,1} = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2 \gamma^*(\lambda_1)} P_{0,0}. \] (19)

Further, using (16b) and (19), we get

\[
P_{0,1} = \frac{\lambda_0}{\mu_1} \left( \frac{1}{\gamma^*(\lambda_1)} - 1 \right) P_{0,0}. \] (20)
Setting $n = 1$ in (6) and using (17), (19) and (20), the probability $P_{0,2}(0)$ is obtained as

$$P_{0,2}(0) = \frac{\lambda_0 \lambda_1 \lambda_2}{\mu_1 \mu_2 \gamma^*(\lambda_1)} P_{0,0}. \tag{21}$$

From (21), (16a) and (16b), we obtain

$$P_{1,2} = \frac{\lambda_0 \lambda_1 \lambda_2}{\mu_1 \mu_2 \mu_3 \gamma^*(\lambda_1) \gamma^*(\lambda_2)} P_{0,0} \quad \text{and} \quad P_{0,2} = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2 \gamma^*(\lambda_1)} \left( \frac{1}{\gamma^*(\lambda_2)} - 1 \right) P_{0,0}. \tag{22-23}$$

In general, we obtain

$$P_{1,n} = \left( \prod_{i=0}^{n} \frac{\lambda_i}{\mu_{i+1}} \right) \frac{1}{\left( \prod_{j=1}^{n} \gamma^*(\lambda_j) \right)} P_{0,0}, \quad 1 \leq n \leq F - 1 \tag{24}$$

$$P_{0,n} = \left( \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}} \right) \left( \frac{1}{\prod_{j=1}^{n-1} \gamma^*(\lambda_j)} \right) \left( \frac{1}{\gamma^*(\lambda_n)} - 1 \right) P_{0,0}, \quad 1 < n \leq F - 1. \tag{25}$$

From (7)-(9) and using (24) and (25), we obtain

$$P_{1,n} = \left( \prod_{i=0}^{n} \lambda_i \right) \frac{S_n}{R_F} P_{0,0}, \quad F \leq n \leq K - 2, \tag{26}$$

$$P_{0,n} = \left( \prod_{i=0}^{n} \lambda_i \right) \frac{\mu_n + 1 - \gamma^*(\lambda_n)}{\lambda_n} \frac{S_n}{R_F} P_{0,0}, \quad F \leq n \leq K - 2, \tag{27}$$

$$P_{1,K-1} = \left( \prod_{i=0}^{K-1} \lambda_i \right) \frac{1}{R_F} P_{0,0}, \quad \tag{28}$$

$$P_{0,K-1} = \left( \prod_{i=0}^{K-1} \lambda_i \right) \frac{\mu_K (1 - \gamma^*(\lambda_{K-1}))}{\lambda_{K-1}} \frac{1}{R_F} P_{0,0}. \tag{29}$$

Now, using (28) in (10), result given in (14) is obtained. Also, $P_{0,0}$ can be determined using normalizing condition given by

$$\sum_{n=0}^{K-1} P_{0,n} + \sum_{n=0}^{K-1} P_{1,n} + \sum_{n=F}^{K-1} P_{2,n} = 1 \tag{30}$$

Remark 1. It should be noted that when $F = K - 1$, we come across M/M/1/K model with state dependent rates and general retrial. In this case, $F$-policy is not taken into account as such $\varepsilon = 0, S_F = 1$.

4. Special models. In this section, some special models are deduced by setting suitable parameter values for the state dependent rates. First of all, by setting state dependent arrival rates, we consider a finite population model for machine repair problem (MRP) and its particular case when the control of arrivals is not taken into consideration. Then after, by setting the state dependent service rate time sharing model is discussed.
4.1. **Finite population model.** In the present scenario of modern life style, machines are needed to perform day-to-day as well as specific jobs. It is noticed that the unexpected failures of machines have an adverse impact on the system efficiency/availability and also increases the production cost and downtime of the system. In this sub-section, the machine repair model which is a finite population model is presented as follows:

By setting \( \lambda_n = (K-n)\lambda \) and \( \mu_n = \mu \), equations (12), (13) and (14) yield the queue size distribution. Thus, for finite population model for MRP, we get

\[
P_{0,n} = \begin{cases} 
\frac{K\lambda}{\mu} \left( \frac{1}{\gamma^*(K-1)\lambda} - 1 \right) P_{0,0}, & n = 1, \\
\frac{K\lambda}{\mu} \left( \frac{1}{\gamma^*((K-i)\lambda)} - 1 \right) P_{0,0}, & 1 < n \leq F - 1, \\
\left( \prod_{i=0}^{n} ((K-i)\lambda) \right) \left( \frac{1 - \gamma^*(K-n)\lambda}{(K-n)\lambda} \right) S_n R_F P_{0,0}, & F \leq n \leq K - 2, \\
\left( \prod_{i=0}^{n} ((K-i)\lambda) \right) \frac{1 - \gamma^*(K-n)\lambda}{(K-n)\lambda} \frac{1}{R_F} P_{0,0}, & n = K - 1.
\end{cases}
\]

(31)

\[
P_{1,n} = \begin{cases} 
\frac{K\lambda}{\mu} P_{0,0}, & n = 0, \\
\frac{K\lambda}{\mu} \left( \prod_{i=1}^{n} ((K-i)\lambda) \right) P_{0,0}, & 1 \leq n \leq F - 1, \\
\left( \prod_{i=0}^{n} ((K-i)\lambda) \right) \frac{1}{R_F} P_{0,0}, & F \leq n \leq K - 2, \\
\left( \prod_{i=0}^{n} ((K-i)\lambda) \right) \frac{1}{R_F} P_{0,0}, & n = K - 1.
\end{cases}
\]

(32)

\[
P_{2,n} = \left( \prod_{i=0}^{n} ((K-i)\lambda) \right) \frac{1}{R_F} P_{0,0}, & F \leq n \leq K - 1.
\]

(33)

In particular, when \( F \)-policy is not taken into account, so that \( \mu_n = \mu \), \( \lambda_n = (K-n)\lambda \), Equations (12) and (13) yield

\[
P_{0,n} = \begin{cases} 
\frac{K\lambda}{\mu} \left( \frac{1}{\gamma^*(K-1)\lambda} - 1 \right) P_{0,0}, & n = 1, \\
\frac{K\lambda}{\mu} \left( \frac{1}{\gamma^*((K-i)\lambda)} - 1 \right) P_{0,0}, & 2 \leq n \leq K - 1.
\end{cases}
\]

(34)

\[
P_{1,n} = \begin{cases} 
\frac{K\lambda}{\mu} P_{0,0}, & n = 0, \\
\frac{K\lambda}{\mu} \left( \prod_{i=1}^{n} ((K-i)\lambda) \right) P_{0,0}, & 1 \leq n \leq K - 1.
\end{cases}
\]

(35)

Equations (34) and (35) provide the same results as obtained by Yang and Chang [40].

4.2. **Time-sharing model with state-dependent arrival rates.** The time-sharing system involves the sharing of source among many tasks by means of parallel operations or allocating a very small quantum of time to each task in round robin fashion. In case of single server time sharing queueing system, the arriving jobs
may wait in the queue for service for a small pre-specified time duration; if they do not get served within this duration, these jobs have to join the end of the queue. When the jobs again join the server for service, the same rule of time-sharing is again applied until they get served. If some jobs are already present for the service, the arriving job has to join the retrial orbit.

The state-dependent time sharing model is formulated by setting $\mu_n = \frac{\mu}{n}$ and $\lambda_n = \frac{\lambda}{n+1}$ in (12)-(14). The queue length distribution for the time-sharing system with state-dependent rates is obtained as

$$P_{0,n} = \begin{cases} \frac{\lambda}{\mu} \left( \frac{1}{\gamma^* \left(\frac{\lambda}{2}\right)} \right) P_{0,0}, & n = 1, \\ \left(\frac{\lambda}{\mu}\right)^n \frac{1}{\prod_{j=1}^{n-1} \gamma^* \left(\frac{\lambda}{j+1}\right)} \left(\frac{\lambda}{n+1}\right) \left(1 - \gamma^* \left(\frac{\lambda}{n+1}\right)\right) P_{0,0}, & 1 < n \leq F - 1, \\ \frac{\lambda^{n+1}}{(n+1)!} \frac{\mu}{\lambda} \left(1 - \gamma^* \left(\frac{\lambda}{n+1}\right)\right) S_n \frac{P_{0,0}}{R_F}, & F \leq n \leq K - 2, \\ \frac{\lambda^{n+1}}{(n+1)!} \frac{\mu}{\lambda} \left(1 - \gamma^* \left(\frac{\lambda}{n+1}\right)\right) \frac{1}{R_F} P_{0,0}, & n = K - 1. \end{cases}$$

(36)

$$P_{1,n} = \begin{cases} \frac{\lambda}{\mu} P_{0,0}, & n = 0, \\ \left(\frac{\lambda}{\mu}\right)^{n+1} \frac{1}{\prod_{j=1}^{n} \gamma^* \left(\frac{\lambda}{j+1}\right)} P_{0,0}, & 1 \leq n \leq F - 1, \\ \frac{\lambda^{n+1}}{(n+1)!} S_n \frac{P_{0,0}}{R_F}, & F \leq n \leq K - 2, \\ \frac{\lambda^{n+1}}{(n+1)!} \frac{1}{R_F} P_{0,0}, & n = K - 1. \end{cases}$$

(37)

$$P_{2,n} = \frac{\lambda^K}{K! \mu R_F} P_{0,0}, \quad F \leq n \leq K - 1.$$

(38)

In particular case when the arrival rate is constant, by setting $\lambda_n = \lambda$ in (36)-(38), we get the results for $F$-policy general retrial model with constant arrival rate.

5. Performance prediction. To analyze the queueing characteristics of the concerned retrial service system and to make the model applicable to the real-time situation, it is beneficial to establish various system indices and cost analysis.

5.1. Performance indices. The queueing model developed in the previous section for a single server finite model with general retrial attempts under admission control according to $F$-policy is analyzed by deriving some system indices as follows:
(i) The average number of jobs in the system and in the queue respectively, are

\[ E[N_s] = \sum_{n=0}^{K-1}nP_{0,n} + \sum_{n=0}^{K-1}(n+1)P_{1,n} + \sum_{n=F}^{K-1}(n+1)P_{2,n} \]

and

\[ E[N_q] = \sum_{n=0}^{K-1}nP_{1,n} + \sum_{n=F}^{K-1}nP_{2,n}. \]  

(39a - b)

(ii) The average number of jobs in the orbit is

\[ E[N_R] = \sum_{n=0}^{K-1}nP_{0,n} \]  

(40)

(iii) The system throughput is

\[ TP = \sum_{n=0}^{K-1}\mu_{n+1}P_{1,n} + \mu \sum_{n=F}^{K-1}P_{2,n} \]  

(41)

(iv) The status of the server can be represented by the probability of the server being free \((P_I)\) and being engaged \((P_{SB})\) in rendering service, respectively. Thus we obtain

\[ P_I = \sum_{n=0}^{K-1}P_{0,n} \text{ and } P_{SB} = \sum_{n=0}^{K-1}P_{1,n} + \sum_{n=F}^{K-1}P_{2,n}. \]  

(42a - b)

5.2. Cost function. For the retrial queueing system operating under \(F\)-policy, the organizers may be interested in the optimal service rate that optimizes the total system cost. To formulate the cost function, the cost components associated with different activities are used. To evaluate the threshold parameter \((F)\) and service rate that optimize the cost function \(TC(F,\mu)\), we formulate the total cost per unit time for operating the system as follows:

\[ TC(F,\mu) = C_IP_I + CBP_{SB} + CHE[N_q] + \mu CF + COE[N_R] \]  

(43)

where

\[ C_I : \text{Cost associated with the server per unit time during idle state} \]
\[ C_B : \text{Cost of the server per unit time when he is busy in rendering the service} \]
\[ C_H : \text{Holding cost of each job residing in the system} \]
\[ C_F : \text{Cost involved per unit time in rendering the service to the job} \]
\[ C_O : \text{Cost spent on each customer while residing in the orbit.} \]

The cost function given in \(43\) is highly non-linear and complex, therefore, its explicit analytical solution may not possible however numerical methods can be easily employed. To get optimal decision parameters \((F,\mu)\), (i) direct search method is used to determine optimal threshold parameter \((F^*)\) which is discrete decision parameter and then (ii) quasi-Newton method is used to evaluate optimal service rate \((\mu^*)\) which is continuous decision parameter.
5.2.1. Direct search method to evaluate ($F^*$). To control the admission of jobs in the system, we find the optimal threshold parameter ($F^*$) so as to optimize the system cost given in (43). It is noticed that the threshold parameter ‘$F$’ (0 ≤ $F$ < $K$ − 1) has integer values. Thus, direct search method based on a heuristic approach is used by successively substituting $F = 0, 1, 2, \ldots, K−2$ to compute cost function given in (43). The optimal threshold parameter ($F^*$) is evaluated by using the inequalities:

$$TC(F^* − 1, \mu) ≥ TC(F^*, \mu) \text{and } TC(F^* + 1, \mu) ≥ TC(F^*, \mu).$$

5.2.2. Quasi-Newton method to evaluate ($\mu^*$). After determining the optimal threshold parameter ($F^*$), we use the quasi-Newton method to determine optimal service rate ($\mu^*$) by minimizing the cost function $TC(F^*, \mu)$. The algorithmic steps of the quasi-Newton method are as follows:

**Inputs:** $K, F^*, \lambda, \mu_0, \gamma, \varepsilon, C_I, C_B, C_H, C_F, C_O$, and tolerance $\varepsilon_0$ of $\left[\frac{\partial TC(F^*, \mu)}{\partial \mu}\right]$.  

**Output:** Approximate optimal solution of service rate ($\mu$) and total cost per unit time ($TC$) as $\mu^*$, $TC(F^*, \mu^*)$.

**Step 1:** Compute $TC(F^*, \mu_0)$ for an initial trial solution $\mu_0$ for $\mu$.

**Step 2:** Evaluate the cost gradient $\nabla TC(\mu_0) = \frac{\partial TC}{\partial \mu} |_{\mu=\mu_0}$ and the Hessian $H(\mu_0)$

$$H(\mu_0) = \frac{\partial^2 TC}{\partial \mu^2} |_{\mu=\mu_0}.$$

**Step 3:** Find the new solution $\mu_{j+1} = \mu_j - [H(\mu_j)]^{-1} \nabla TC(\mu_j), j = 0, 1, 2, \ldots$.

**Step 4:** Set $j = j + 1$, and repeat the step until $\left|\frac{\partial TC}{\partial \mu_j}\right| < \varepsilon_0$ where $\varepsilon_0 = 10^{-6}$ is the tolerance.

**Step 5:** Find the approximate minimum $\mu^*$ and corresponding total cost function $TC(F^*, \mu^*)$.

**Step 6:** End.

6. Illustration and numerical results. $F$-policy state dependent retrial queueing model developed has applications in several real-time congestion problems including MRP and time-sharing system. To illustrate for MRP, we consider a computer repair shop in which finite number (say $K$) of computers can be repaired under a maintenance contract. The failed computers arrive for the repair job to the shop by following Poisson process with the rate $\lambda$. The repair job of a failed computer is done by the repairman following the exponential distribution with mean $1/\mu$. If the caretaker of the failed computer finds the repairman busy, the failed computers are put in the orbit; from the orbit, it can be sent again for the repair job; the retrial time is assumed to be generally distributed with mean $1/\gamma$. Due to the limited space of the shop, the arriving failed computers are not permitted to enter the shop as soon as the capacity of the shop becomes full. The failed computers are permitted to further join the shop only when the workload of failed computers reduces to a predefined level ‘$F$’ in terms of a number of failed computers. Before allowing the failed computers to enter the shop for repair, the setup time is required which is also assumed to be exponentially distributed with the rate $\varepsilon$.

The applicability of time sharing queueing models is quite prevalent in the multiplexed information and computing service system, operating systems, computer and communication system, cloud computing centers, etc. To be specific, we cite the illustration of a call center with a single server to serve the queries of arriving calls. The arriving calls contact the agent i.e. server of the call center to receive the service. If the agent is free at that time, then the call gets service immediately.
When the agent is busy, then the arriving call has to wait in the retrial orbit. After some random time, the call requests for service to the agent again. In the case when the number of calls accumulated in the call center reaches to the system capacity (i.e., $K$) of the call center, the arriving calls are not allowed to join until the number of calls in the system reduces to pre-fixed level (i.e., $F$).

The numerical simulation and cost optimization have been carried out for both machine repair problem (MRP) and the time-sharing model (TSM). The numerical experiment performed may be helpful to examine the effects of parameters on various performance measures and to determine the optimal threshold parameter and optimal service rate. The three distributions for the retrial time, have been considered. The Laplace- Stieltjes transform of $\gamma(x)$ for exponential ($\text{Exp}$), Erlang-3($E_3$), and deterministic ($D$) distribution are taken as $\gamma^*(\theta) = \frac{\gamma}{\theta + \gamma}$, $\gamma^*(\theta) = \left(\frac{3\gamma}{\theta + 3\gamma}\right)^3$ and $\gamma^*(\theta) = e^{-\theta/\gamma}$, respectively.

The software ‘MATLAB’ is used to develop the computer program to compute the system indices and cost function. For the machine repair problem and the time-sharing model, we set the state dependent rates as (i) $\lambda_n = \lambda(K - n)$, $\mu_n = \mu$ and (ii) $\lambda_n = \lambda/(n + 1)$, $\mu_n = \mu/n$, respectively.

### 6.1. Numerical results for the machine repair model.

To validate the practical application of the computer repair shop model, we evaluate the performance indices numerically for exponentially distributed retrial time by setting the default parameters as $K = 7$, $F = 4$, $\lambda = 1$ unit/hour, $\mu = 2$ unit/hour, $\gamma = 0.5$ unit/hour, $\varepsilon = 0.5$ unit/hour. For the computer repair example, the average number of failed computers in the shop is obtained as 5.65.

#### 6.1.1. Sensitivity analysis for MRP.

To explore the sensitivity of the repair rate ($\mu$), failure rate ($\lambda$) and retrial rate ($\gamma$) with respect to the indices $E[N_S]$ and $TP$, the graphs are plotted in Figures 2(a-c) and 3(a-c), respectively. Some other performance indices have also been summarized in Tables 1-3 for varying values of these parameters. For the computation of system indices, the default parameters are fixed as $K = 7$, $F = 4$, $\lambda = 0.5$, $\mu = 8$, $\gamma = 0.5$, $\varepsilon = 1$.

Based on numerical experiments performed, we present the sensitivity of the parameters as follows:

- **Effect of $\mu$:** It is observed that as the repairman (server) repairs the failed machines with a faster rate as much the average number of failed machines decreases which also demonstrates the validity of analytical results. Also the status of the repairman (idle or busy) completely depends on the number of failed machines. From Table 1, it is noticed that as repair rate ($\mu$) increases, the average number of failed machines in the queue ($E[N_q]$) decreases. Also the probability of repairman being busy (idle) decreases (increases) by enhancing the $\mu$. The graph plotted in Figure 2(a) depicts that $E[N_S]$ lowers down as the service rate goes up. From Figure 3(a), it is clear that the throughput ($TP$) grows up as the service rate ($\mu$) speeds up which is the same as we expect.

- **Effect of $\lambda$:** For the MRP, if the repairman provides repair job to the machines with constant rate and the failure rate ($\lambda$) of machines increases, then $E[N_q]$ seems to increase at a faster pace. Table 2 displays the numerical results which demonstrate that by keeping the service rate constant, if the rate of the failure of machines increases, then the average queue size ($E[N_q]$) enhances.
The probability of the repairman being busy (idle) also seems to increase (decrease) by increasing the failure rate of the machines. From graphs shown in Figures 2(b) and 3(b), it is evident that $E[N_S]$ and $TP$ grow up as $\lambda$ increases.

- **Effect of $\gamma$:** Table 3 presents the negligible effect of the retrial rate on various performance indices. We observe that $E[N_q]$ decreases very slowly as the retrial rate increases. The probability of a repairman being busy (idle) remains almost constant as the retrial rate grows up. Figure 2(c) reveals the trends of $E[N_S]$ which decreases at a slow pace as the retrial rate ($\gamma$) increases. Figure 3(c) shows that the throughput ($TP$) of the system enhances with a very slow rate as $\gamma$ enhances.

### Table 1. Various performance measures for varying values of $\mu$ for MRP.

| $\mu$ | $E[N_1]$ | $P_I$ | $P_{ps}$ | TC   |
|-------|----------|-------|----------|------|
|       | Exp      | $E_1$ | $D$      | Exp  | $E_3$ | $D$      | Exp  | $E_5$ | $D$      |
| 1     | 3.164    | 3.110 | 3.060    | 0.303 | 0.335 | 0.354    | 0.697 | 0.665 | 0.646    |
| 2     | 2.043    | 2.013 | 1.964    | 0.492 | 0.543 | 0.569    | 0.508 | 0.457 | 0.431    |
| 3     | 1.482    | 1.477 | 1.436    | 0.580 | 0.643 | 0.672    | 0.420 | 0.357 | 0.328    |
| 4     | 1.155    | 1.166 | 1.135    | 0.633 | 0.699 | 0.730    | 0.367 | 0.301 | 0.270    |
| 5     | 0.940    | 0.963 | 0.941    | 0.671 | 0.735 | 0.768    | 0.329 | 0.265 | 0.232    |

### Table 2. Various performance measures for varying values of $\lambda$ for MRP.

| $\lambda$ | $E[N_1]$ | $P_I$ | $P_{ps}$ | TC   |
|-----------|----------|-------|----------|------|
|           | Exp      | $E_1$ | $D$      | Exp  | $E_3$ | $D$      | Exp  | $E_5$ | $D$      |
| 1         | 1.132    | 1.069 | 0.998    | 0.704 | 0.771 | 0.793    | 0.296 | 0.229 | 0.207    |
| 2         | 1.742    | 1.604 | 1.542    | 0.644 | 0.687 | 0.702    | 0.356 | 0.313 | 0.298    |
| 3         | 2.161    | 2.026 | 1.985    | 0.582 | 0.617 | 0.626    | 0.418 | 0.383 | 0.374    |
| 4         | 2.494    | 2.372 | 2.347    | 0.530 | 0.559 | 0.564    | 0.470 | 0.441 | 0.436    |
| 5         | 2.767    | 2.661 | 2.645    | 0.486 | 0.510 | 0.514    | 0.514 | 0.490 | 0.486    |

### Table 3. Various performance measures for varying values of $\gamma$ for MRP.

| $\gamma$ | $E[N_1]$ | $P_I$ | $P_{ps}$ | TC   |
|-----------|----------|-------|----------|------|
|           | Exp      | $E_1$ | $D$      | Exp  | $E_3$ | $D$      | Exp  | $E_5$ | $D$      |
| 0.5       | 0.581    | 0.633 | 0.628    | 0.745 | 0.799 | 0.830    | 0.255 | 0.201 | 0.170    |
| 0.6       | 0.562    | 0.627 | 0.637    | 0.733 | 0.784 | 0.817    | 0.267 | 0.216 | 0.183    |
| 0.7       | 0.544    | 0.617 | 0.643    | 0.723 | 0.771 | 0.805    | 0.277 | 0.229 | 0.195    |
| 0.8       | 0.527    | 0.606 | 0.645    | 0.714 | 0.760 | 0.795    | 0.286 | 0.240 | 0.205    |
| 0.9       | 0.510    | 0.593 | 0.644    | 0.706 | 0.750 | 0.784    | 0.294 | 0.250 | 0.216    |
6.1.2. Cost optimization for MRP. In machine repair problem, the system organizers may be interested to evaluate optimal threshold parameter ‘$F$’, optimal service rate ($\mu$) and corresponding minimal cost of the system. From the practical point of view, the system capacity $K$ can be treated as an upper bound to determine the threshold parameter ‘$F$’ which can be searched in the desired feasible region. For evaluating the total cost, the four cost sets given in Table 4, have been taken into account.

Table 4. Cost sets with different cost elements (in $) for MRP.

| Cost Set | $C_I$ | $C_B$ | $C_H$ | $C_F$ | $C_O$ |
|----------|------|------|------|------|------|
| I        | 30   | 30   | 50   | 70   | 40   |
| II       | 10   | 10   | 120  | 15   | 90   |
| III      | 15   | 5    | 120  | 15   | 90   |
| IV       | 20   | 20   | 100  | 15   | 90   |

Optimal threshold parameter ($F^*$): In order to compute the optimal threshold parameter and minimum cost for the MRP, a direct search method, based on a heuristic approach is applied. To determine $F^*$ and corresponding total cost $TC(F^*, \mu)$ for Cost Set - I, the default parameters are fixed as $K = 25$, $\lambda = 0.1$, $\mu = 1.5$, $\gamma = 0.5$, $\varepsilon = 1$, and then the total cost is computed by varying $F$ in
the feasible region from 0 to 23. The total cost function for different distributions (exponential (Exp), Erlang-3 (E3) and deterministic (D)) is computed and plotted in Figure 4 for varying values of $F$. From Figure 4, it is seen that the expected cost function is unimodal and convex in the feasible range (0, 23) with respect to the threshold parameter $F$. We indicate the minimum cost corresponding to the optimal value of the threshold parameter $F$ for Exp, E3 and D distributions of retrial time. Table 5 displays the optimal threshold parameter $F^*$ and corresponding optimal cost $TC(F^*, \mu)$ for different distributions for varying values of retrial rate as $\gamma = 0.3, 0.5$ and 0.7.

**Table 5.** Searching the optimal F for MRP for different $\gamma$.

| $\gamma$ | Exp  | E3   | D     |
|----------|------|------|-------|
| 0.3      | (12, 852.01) | (14, 898.49) | (15, 937.18) |
| 0.5      | (10, 821.20)  | (12, 848.52)  | (13, 874.16)  |
| 0.7      | (8, 806.92)   | (10, 821.04)  | (11, 836.66)  |

**Figure 4.** TC vs. F for different distributions (Exp, E3 and D) when $\gamma = 0.5$.

**Optimal service rate** ($\mu^*$): Since $\mu$ is a continuous decision parameter, we apply the quasi-Newton method to evaluate the optimal service rate ($\mu^*$). To compute the optimal service rate ($\mu^*$), we choose the threshold optimal parameter $F^*$ which is already obtained using direct search method (see Table 5). Using algorithmic steps of the quasi-Newton method given in Section 5.2, for the exponential distribution, we see in Table 6 that after performing 7 iterations, the optimal service rate ($\mu^*$) is attained; the corresponding minimum cost $TC(F^*, \mu^*)$ is $821.20$. The total minimum costs corresponding to the optimal threshold parameter ($F^*$) and optimal service rate ($\mu^*$) for different distributions (Exp, E3 and D) by considering $\gamma = 0.3, 0.5$ and 0.7 are recorded in Table 7.

From the results given in Table 7, we notice that the expected minimum cost corresponding to optimal service rate in case of the exponential distribution is least
as compared to other two distributions viz. Erlang-3 and deterministic for the retrial time. For \(Exp, E_3\) and \(D\) retrial time distributions, the surface graphs of cost function \(TC\) by varying parameters \('F' and 'µ'\) are depicted in Figures 5(a-c) respectively; these graphs reveal the convexity of \(TC\) with respect to \(F\) and \(µ\) both.

By setting the parameters \(K = 7, F = 4, \lambda = 0.5, \varepsilon = 1\) for the Cost Sets II, III, and IV, the surface graphs for the expected cost function \(TC(F, µ)\) are also shown in Figures 6(a-c)-8(a-c) for the exponential (\(Exp\)), Erlang-3(\(E_3\)) and deterministic (\(D\)) distributions, respectively. In these figures, the total cost \(TC\) is displayed with varying the values of \(µ\) from 2 to 16 and \(γ\) from 0.4 to 2. We can see that the cost functions are convex and unimodal in a feasible range of service rate.

Table 6. Searching the \(µ^*\) by quasi-Newton method for exponential distribution \((F^* = 10, γ = 0.5)\).

| Iterations | \(F^*\) | \(µ\) | \(TC(F^*, µ)\) | Max. tolerance |
|------------|--------|-------|----------------|---------------|
| 0          | 10     | 1     | 853.525        | 9.87E+01      |
| 1          | 10     | 2     | 842.70         | 5.69E+01      |
| 2          | 10     | 1.6344| 823.783        | 3.5E+01       |
| 3          | 10     | 1.0491| 821.311        | 8.28          |
| 4          | 10     | 1.5255| 821.207        | 2.27          |
| 5          | 10     | 1.4917| 821.199        | 6.35E-02      |
| 6          | 10     | 1.4990| 821.199        | 4.12E-04      |
| 7          | 10     | 1.4988| 821.199        | 1.02E-05      |

Table 7. Minimum cost \((F^*, µ^*, TC(F^*, µ^*))\) for \(γ = 0.3, 0.5\) and 0.7 for MRP.

| \(γ\) | \(Exp\)          | \(E_3\)          | \(D\)            |
|-------|-------------------|-------------------|-------------------|
| 0.3   | (12, 1.601, 851.05)| (14, 1.668, 897.73)| (15, 1.323, 936.78)|
| 0.5   | (10, 1.499, 821.20)| (12, 1.579, 847.89)| (13, 1.691, 872.05)|
| 0.7   | (8, 1.489, 806.90)| (10, 1.543, 820.77)| (11, 1.627, 834.89)|

Figure 5. Total cost \(TC\) for varying \(µ\) and \(F\) for three distribution by taking three cost sets-I for MRP.
Figure 6. Total cost TC for varying \( \mu \) and \( \gamma \) for exponential distribution by taking three cost sets-II, III and IV

Figure 7. Total cost TC for varying \( \mu \) and \( \gamma \) for Erlang-3 distribution by taking three cost sets-II, III and IV

Figure 8. Total cost TC for varying \( \mu \) and \( \gamma \) for deterministic distribution by taking three cost sets-II, III and IV

6.2. Numerical results for the time-sharing model (TSM). The analytical derivation of the system indices for the time-sharing model is done in Section 4.2. However, to understand the system behavior, the sensitivity of the parameters for different indices is required.
6.2.1. Sensitivity analysis for TSM. For the TSM, the pictorial representations of $E[N_S]$ and $TP$ are shown in Figures 9(a-c) and 10(a-c), respectively. The numerical results for the probability of the server being busy or idle and $E[N_q]$ are summarized in Tables 8-10 for varying values of $\mu$, $\lambda$ and $\gamma$, respectively. For the computation of system indices, the default parameters are fixed as $K = 7, F = 4, \lambda = 2.8, \mu = 10, \gamma = 1, \varepsilon = 1$.

- **Effect of $\mu$**: Numerical results for $E[N_q], P_I, P_{SB}$ and cost function ($TC$) for varying values of $\mu$ are summarized in Table 8. It is seen that as the service rate increases, $E[N_q]$ and the probability of the server being busy (idle) decreases (increases). From Figure 9(a), it is clear that $E[N_S]$ decreases as $\mu$ grows up. Figure 10(a) reveals that $TP$ goes up as $\mu$ speeds up.

- **Effect of $\lambda$**: Table 9 displays the numerical results of various performance indices for varying values of $\lambda$. It is noted that as $\lambda$ increases, $E[N_q]$ also increases. The probability of the server is idle (busy) seems to decrease (increase) by increasing the value of $\lambda$ and keeping the $\mu$ as constant. The trends shown in Figures 9(b) and 10(b) indicate that $E[N_S]$ and $TP$ of the system grow up as the arrival rate of the jobs increases.

- **Effect of $\gamma$**: Table 10 reveals that the retrial rate ($\gamma$) of the jobs has no significant effects on the system indices. As the retrial rate goes up, $E[N_q]$ seems to decrease with a very slower pace. From Table 10, it is noticed that as the retrial rate increases, the probability of the server being busy (idle) increases (decreases) with very slow rate. From Figure 9(c), it is clear that $E[N_S]$ goes up very slowly as the retrial rate ($\gamma$) increases. Figure 10(c) shows that the throughput of the system increases as $\gamma$ grows.

| $\mu$ | $E[N_q]$ | $P_I$ | $P_{SB}$ | $TC$ |
|-------|---------|-------|----------|------|
|       | Exp     | $E_1$ | Exp      | $E_1$ |
| 6     | 0.876   | 0.995 | 1.041    | 0.552 |
| 8     | 0.476   | 0.570 | 0.633    | 0.654 |
| 10    | 0.286   | 0.352 | 0.399    | 0.721 |
| 12    | 0.187   | 0.235 | 0.272    | 0.476 |
| 14    | 0.131   | 0.166 | 0.196    | 0.800 |

| $\lambda$ | $E[N_q]$ | $P_I$ | $P_{SB}$ | $TC$ |
|------------|----------|-------|----------|------|
|           | Exp      | $E_1$ | Exp      | $E_1$ |
| 2          | 0.107    | 0.124 | 0.138    | 0.800 |
| 4          | 0.783    | 0.948 | 1.014    | 0.613 |
| 6          | 1.726    | 1.843 | 1.829    | 0.493 |
| 8          | 2.340    | 2.357 | 2.303    | 0.426 |
| 10         | 2.737    | 2.699 | 2.632    | 0.380 |
Table 10. Various performance measures for varying values of $\gamma$ for TSM.

| $\gamma$ | $E[N_s]_{Exp}$ | $P_1$ | $P_{sl}$ | $TC$ |
| --- | --- | --- | --- | --- |
| 1     | 0.286 0.352 0.399 | 0.721 | 0.723 | 0.279 0.278 | 0.277 | 237.26 | 261.73 | 280.32 |
| 1.5   | 0.225 0.256 0.276 | 0.719 | 0.721 | 0.281 | 0.279 | 214.93 | 226.13 | 234.03 |
| 2     | 0.195 0.212 0.223 | 0.718 | 0.720 | 0.282 | 0.280 | 204.02 | 210.31 | 214.43 |
| 2.5   | 0.177 0.188 0.194 | 0.717 | 0.720 | 0.281 | 0.280 | 197.58 | 201.59 | 204.06 |
| 3     | 0.165 0.173 0.177 | 0.716 | 0.718 | 0.284 | 0.281 | 193.35 | 196.12 | 197.75 |

6.2.2. Cost optimization for TSM. The time-sharing model is explored to determine the optimal threshold parameter ($F^*$) and optimal service rate ($\mu^*$) and the corresponding minimum cost $TC(F^*, \mu^*)$ of the system. It is noticed that the system capacity $K$ can be used as an upper bound for the feasible search space of $F^*$. The combined direct search method and the quasi-Newton method are used to evaluate $F^*$ and $\mu^*$, respectively. For determining the total cost, the four cost sets have been taken into consideration as given in Table 11.
Table 11. Cost sets with different cost elements (in $) for TSM.

| Cost Set | CI | CB | CH | CF | CO |
|----------|----|----|----|----|----|
| I        | 30 | 40 | 120| 60 | 90 |
| II       | 10 | 10 | 120| 15 | 90 |
| III      | 15 | 5 | 120| 15 | 90 |
| IV       | 10 | 10 | 100| 15 | 110|

**Optimal threshold parameter** \( (F^*) \): To evaluate the optimal threshold parameter \( F \), the default parameters are set as \( K = 20, \lambda = 4, \gamma = 3, \varepsilon = 1 \). For cost set-I, by varying the value of \( F \) from 0 to 18, the total cost \( TC \) is computed. Table 12 provides the minimum cost of the system corresponding to the optimal threshold parameter for the exponential, Erlang-3 and deterministic distributed retrial time by taking \( \gamma = 1, 3, 5 \). The trend of \( TC \) by varying \( F \) shown in Figure 11 reveals that the optimal threshold parameter \( (F^*) \) lies in the feasible range \( 0 \leq F < K - 1 \) of \( F \).

Table 12. Searching the optimal \( F \) for TSM.

| \( \gamma \) | \( (F^*, TC(F^*, \mu)) \) | \( \text{Exp} \) | \( E_3 \) | \( D \) |
|-------------|-----------------|-----------|----------|--------|
| 1           | (2, 822.28)     | (1, 858.94) | (1, 872.35) |
| 3           | (5, 665.29)     | (5, 678.51) | (5, 687.12) |
| 5           | (6, 627.96)     | (6, 632.94) | (6, 635.92) |

**Figure 11.** TC vs. \( F \) for different distributions (\( \text{Exp}, E_3 \) and \( D \)) when \( \gamma = 3 \).

**Optimal service rate** \( (\mu^*) \): For the time-sharing system, the service can be controlled by the system developers so as to provide the better service but at minimum cost. For \( \gamma = 3, F^* = 5 \), to determine the optimal service rate \( (\mu^*) \), quasi-Newton
method is applied by considering the retrial times as exponential. It is noticed that after 5th iterations, the minimum cost $TC(F^*, \mu^*)$ is achieved at $659.19$ corresponding to $\mu^* = 7.271$ as can be seen from Table 13.

Now to perform a numerical experiment to determine $(\mu^*)$, we use the optimal threshold parameter $F^*$ which is given in Table 12 and then use the quasi-Newton method for exponential, Erlang-3 and deterministic distributions by taking $\gamma = 1, 3, 5$. The optimal values of parameters and corresponding minimum cost are shown in Table 14.

It is seen that the minimum cost for the exponentially distributed retrial time is less as compared to Erlang-3 and deterministic distributions for the retrial time. From Figures 12(a-c), it is clear that the cost function is the convex and the minimum value of the cost is achieved at the optimal threshold parameter and optimal service rate. From the graphs plotted in Figures 12(a-c), it is also observed that the exponential distribution gives the lowest value of the minimum cost in comparison to other distributions for retrial time.

For the cost sets II, III and IV, by fixing $K = 7, F = 4, \lambda = 2.8, \varepsilon = 1$, the surfaces graphs for the total cost function are shown in Figures 13(a-c), 14(a-c) and 15(a-c) by varying $\mu$ from 5 to 12 and $\gamma$ from 1 to 4. It is seen that the cost functions are convex for all the three distributions viz. exponential, Erlang-3 and deterministic distributions and the minimum value of the cost can be attained at an optimal service rate.

### Table 13. Searching the $\mu^*$ by quasi-Newton method for exponential distribution $(F^* = 5, \gamma = 3)$.

| Iterations | $F^*$ | $\mu$ | $TC(F^*, \mu)$ | Max. tolerance |
|------------|-------|-------|----------------|----------------|
| 0          | 5     | 8     | 665.286        | 1.58E+01       |
| 1          | 5     | 7     | 660.172        | 7.35           |
| 2          | 5     | 7.317 | 669.221        | 1.16           |
| 3          | 5     | 7.273 | 659.194        | 6.86E-02       |
| 4          | 5     | 7.271 | 659.194        | 7.04E-04       |
| 5          | 5     | 7.271 | 659.194        | 0              |

### Table 14. Minimum cost $(F^*\mu^*, TC(F^*, \mu^*))$ for $\gamma = 1, 3$ and 5 for TSM.

| $\gamma$ | $(F^*, \mu^*, TC(F^*, \mu^*))$ |
|----------|---------------------------------|
| 1        | $(2, 6.169, 803.45)$ $(1, 5.854, 809.85)$ $(1, 6.092, 820.80)$ |
| 3        | $(5, 7.271, 659.19)$ $(5, 7.341, 673.70)$ $(5, 7.378, 682.92)$ |
| 5        | $(6, 7.012, 615.23)$ $(6, 7.048, 621.22)$ $(6, 7.068, 624.76)$ |
Figure 12. Total cost $TC$ for varying $\mu$ and $F$ for three distributions by taking three cost sets- I for TSM

Figure 13. Total cost $TC$ for varying $\mu$ and $\gamma$ for exponential distribution by taking three cost sets- II, III and IV

Figure 14. Total cost $TC$ for varying $\mu$ and $\gamma$ for Erlang-3 distribution by taking three cost sets- II, III and IV
7. Conclusions and future scope. The state-dependent M/M/1/K queueing model with general retrial times that operates under $F$-policy is studied by using the supplementary variable corresponding to remaining retrial times. Some specific distributions viz. exponential, Erlang-3 and deterministic distributions are considered for the general retrial times for providing the computational results for various performance indices. The retrial queueing situations encountered in repairable machining environment and time-sharing environment can be seen at computers/communication networks, manufacturing/production lines, etc. Several day-to-day queueing scenarios at bank counters, post offices, railway stations etc. can also be analyzed by using the performance results derived in the present investigation dealing with a non-Markovian model under $F$-policy rule. The system indices measured by taking numerical illustration show the validity of the model in real time system. The performance analysis of the concerned model may be applicable to many real-world congestion problems to chalk out the admission control policy by controlling the arrivals of jobs in the system in particular when the capacity of the system is not sufficient for heavy traffic. Based on the cost analysis presented, the optimal service rate and the corresponding minimum expected cost incurred on the system can be determined. The model presented in this article can be further extended by incorporating the bulk input and/or unreliable server concepts.

Acknowledgement. We thank the Editors and the anonymous referees for constructive suggestions and feedback that help us to improve the presentation. The corresponding author Sudeep Singh Sanga is thankful to MHRD, Govt. of India for providing a financial grant to carry out the research work.

REFERENCES

[1] I. J. B. F. Adan and V. G. Kulkarni, Single-server queue with Markov-dependent inter-arrival and service times, Queueing Syst., 45 (2003), 113–134.
[2] I. Adiri and B. Avi-Itzhak, A time-sharing queue, Manage. Sci., 15 (1969), 639–657.
[3] A. Banerjee and U. C. Gupta, Reducing congestion in bulk-service finite-buffer queueing system using batch-size-dependent service, Perform. Eval., 69 (2012), 53–70.
[4] M. Boualem, N. Djellab and D. Aïssani, Stochastic bounds for a single server queue with general retrial times, Bull. Iran. Math. Soc., 40 (2014), 183–198.
[5] M. Chandrasekaran, M. Muralidhar and U. S. Dixit, Online optimization of multipass machining based on cloud computing, Int. J. Adv. Manuf. Technol., 65 (2013), 239–250.
[6] C.-J. Chang, F.-M. Chang and J.-C. Ke, Economic application in a Bernoulli $F$-policy queueing system with server breakdown, *Int. J. Prod. Res.*, **52** (2014), 743–756.

[7] C.-J. Chang and J.-C. Ke, Randomised controlling arrival for a queueing system with subject to server breakdowns, *Optimization.*, **64** (2015), 941–955.

[8] J. Chang and J. Wang, Unreliable M/M/1/1 retrial queues with set-up time, *Qual. Technol. Quant. Manag.*, **3703** (2017), 1–13.

[9] G. Choudhury and J.-C. Ke, An unreliable retrial queue with delaying repair and general retrial times under Bernoulli vacation schedule, *Appl. Math. Comput.*, **230** (2014), 436–450.

[10] D. R. Cox, The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables, *Math. Proc. Cambridge Philos. Soc.*, **51** (1955), 433.

[11] S. D. Flapper, J.-P. Gayon and L. L. Lim, On the optimal control of manufacturing and remanufacturing activities with a single shared server, *Eur. J. Oper. Res.*, **234** (2014), 86–98.

[12] S. Gao, J. Wang and W. W. Li, An M/G/1 retrial queue with general retrial times, working vacations and vacation interruption, *Asia-Pacific J. Oper. Res.*, **31** (2014), 1440006.

[13] S. M. Gupta, Interrelationship between controlling arrival and service in queueing systems, *Comput. Oper. Res.*, **22** (1995), 1005–1014.

[14] M. Jain, An $(m, M)$ machine repair problem with spares and state dependent rates: A diffusion process approach, *Microelectron. Reliab.*, **37** (1997), 929–933.

[15] M. Jain and A. Bhagat, Transient analysis of finite $F$-policy retrial queues with delayed repair and threshold recovery, *Nat. Acad. Sci. Lett.*, **38** (2015), 257–261.

[16] M. Jain and S. S. Sanga, Performance modeling and ANFIS computing for finite buffer retrial queue under $F$-policy, in *Proceedings of Sixth International Conference on Soft Computing for Problem Solving*, Patiala, India, 2017, 248–258.

[17] M. Jain and S. S. Sanga, Control $F$-policy for fault tolerance machining system with general retrial attempts, *Nat. Acad. Sci. Lett.*, **40** (2017), 359–364.

[18] M. Jain and S. S. Sanga, $F$-policy for M/M/1/K retrial queueing model with state-dependent rates, in *Performance Prediction and Analytics of Fuzzy, Reliability and Queuing Models* (eds. K. Deep, M. Jain and S. Salhi), Springer, Singapore, 2019, 127–138.

[19] M. Jain, S. S. Sanga and R. K. Meena, Control $F$-policy for Markovian retrial queue with server breakdowns, in *1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, New Delhi, India, 2016, 1–5.

[20] M. Jain, G. C. Sharma and V. Rani, M/M/R+r machining system with reneging, spares and interdependent controlled rates, *Int. J. Math. Oper. Res.*, **6** (2014), 655–679.

[21] M. Jain, G. C. Sharma and R. Sharma, Optimal control of (N, F) policy for unreliable server queue with multi-optional phase repair and start-up, *Int. J. Math. Oper. Res.*, **4** (2012), 152–174.

[22] M. Jain, G. C. Sharma and C. Shekhar, Processor-shared service systems with queue-dependent processors, *Comput. Oper. Res.*, **32** (2005), 629–645.

[23] M. Jain, C. Shekhar and S. Shukla, Queueing analysis of machine repair problem with controlled rates and working vacation under F-Policy, *Proc. Natl. Acad. Sci. India Sect. A Phys. Sci.*, **86** (2016), 21–31.

[24] J. C. Ke, C. H. Wu and Z. G. Zhang, Recent developments in vacation queueing models: A short survey, *Int. J. Oper. Res.*, **7** (2010), 3–8.

[25] J. Kim and B. Kim, The processor-sharing queue with bulk arrivals and phase-type services, *Perform. Eval.*, **64** (2007), 277–297.

[26] E. R. Kumara and S. Dharsana, Analysis of M/M/1 queueing system with state dependent arrival and detainment of retracted customers, *Malaya J. Mat.*, (2015), 89–98.

[27] C. Lee, On moment stability properties for a class of state-dependent stochastic networks, *J. Korean Stat. Soc.*, **40** (2011), 325–336.

[28] C.-D. Liou, Optimization analysis of the machine repair problem with multiple vacations and working breakdowns, *J. Ind. Manag. Optim.*, **11** (2014), 83–104.

[29] W. A. Massey, The analysis of queues with time-varying rates for telecommunication models, *Telecommun. Syst.*, **21** (2002), 173–204.

[30] P. Moreno, An M/G/1 retrial queue with recurrent customers and general retrial times, *Appl. Math. Comput.*, **159** (2004), 651–666.

[31] P. R. Parthasarathy and R. Sudhesh, Time-dependent analysis of a single-server retrial queue with state-dependent rates, *Oper. Res. Lett.*, **35** (2007), 601–611.

[32] T. Phung-Duc and K. Kawanishi, Multiserver retrial queue with setup time and its application to data centers, *J. Ind. Manag. Optim.*, **15** (2019), 15–35.
[33] J. Rodrigues, S. M. Prado, N. Balakrishnan and F. Louzada, Flexible M/G/1 queueing system with state dependent service rate, Oper. Res. Lett., 44 (2016), 383–389.

[34] K. H. Wang, Cost analysis of the M/M/R machine-repair problem with mixed standby spares, Microelectron. Reliab., 33 (1993), 1293–1301.

[35] K. H. Wang, C. C. Kuo and W. L. Pearn, Optimal control of an M/G/1/K queueing system with combined F policy and startup time, J. Optim. Theory Appl., 135 (2007), 285–299.

[36] K.-H. Wang, C.-C. Kuo and W. L. Pearn, A recursive method for the F-policy G/M/1/K queueing system with an exponential startup time, Appl. Math. Model., 32 (2008), 958–970.

[37] K.-H. Wang and B. D. Sivazlian, Cost analysis of the M/M/R machine repair problem with spares operating under variable service rates, Microelectron. Reliab., 32 (1992), 1171–1183.

[38] D.-Y. Yang, F.-M. Chang and J.-C. Ke, On an unreliable retrial queue with general repeated attempts and J optional vacations, Appl. Math. Model., 40 (2016), 3275–3288.

[39] D.-Y. Yang and P.-K. Chang, A parametric programming solution to the F-policy queue with fuzzy parameters, Int. J. Syst. Sci., 46 (2015), 590–598.

[40] D.-Y. Yang and Y.-D. Chang, Sensitivity analysis of the machine repair problem with general repeated attempts, Int. J. Comput. Math., 95 (2018), 1761–1774.

[41] C. Yeh, Y.-T. Lee, C.-J. Chang and F.-M. Chang, Analysis of a two-phase queue system with \( <p, F> \)- policy, Qual. Technol. Quant. Manag., 14 (2017), 178–194.

Received March 2018; revised February 2019.

E-mail address: madhufma@iitr.ac.in
E-mail address: sudeepiitroorkee@gmail.com