SINGLE SERVER QUEUEING MODEL WITH 
N-POLICY AND REMOVABLE SERVER

CHARAN JEET SINGH AND MADHU JAIN

ABSTRACT. This investigation deals with $M/E_k/1$ queueing system under $N$-policy, by using the method of entropy maximization. The queueing system consists of a removable server, who is controllable such that the organizer can turn the server on at the arrival approach of the $N^{th}$ unit, or turn off at the service completion of the last unit when the system becomes empty. The interarrival times and service times of the units are assumed to follow the negative exponential distribution and Erlang-$k$ type distribution, respectively. By using the principle of entropy, we obtain the formula for the approximate value of the probability distribution of the queue size in terms of known mean operational metrics. Other measures of performance are also discussed. Numerical illustration to validate the analytical results of the queue size distribution, is also given.

1 Introduction

In many applications of computer networks, telecommunications and manufacturing systems, queueing models have played an important role to predict the performance indices of interest. In an $N$-policy $M/E_k/1$ queueing system with a removable service station, the removable server is used as the threshold policy of the system which allows one to turn on and turn off the server, depending on the number of units present in the system. According to $N$-policy, the server starts service only when $N$ ($N \geq 1$) units are present in the system and is shut down (turned off) when no unit is present in the system; after the server is turned off, the server may not operate until again $N$ units are accumulated in the system.

In some cases, the classical approaches of queueing theory fail to give exact result for the queue size distribution of the congestion models in different frameworks. The complex queueing systems encountered in real
time systems such as computer networks, telecommunications, production and manufacturing systems etc., cannot be analyzed based upon the assumptions made by classical queueing modeling. In this context, the use of information theoretic techniques based on the principle of maximum entropy provides an elegant alternative to the stochastic queueing models in order to evaluate the approximate results using some known mean characteristics.


In the present investigation, the queueing model with single service station, which can turn on at an arrival epoch or off at service completion epoch based on a threshold policy has been investigated to obtain probability distribution by using maximum entropy principle. Early important work on queueing problems with removable service station under \(N\)-policy was due to Yadin and Naor [25], which turns the server on when the number of units in the system reaches a specific number \(N\) \((N \geq 1)\), and turns the server off when there is no unit in the system. Queueing models with a single-removable server have been investigated by several researchers, including Bell [2, 3], Heymann [6], and Kimura [9], etc. under various assumptions on the interarrival and service time distributions. Sivazlian and Stanfel [15] developed an \(M/M/1\) queueing model based on controllable rates of arrival units. Various performance measures along with cost analysis have been obtained by Wang et al. [22] for a queueing model with warm standby components and general repair station. Wang et al. [17] extended the model with additional assumptions for a repairable service station. A model on reliability and
sensitivity analysis of a repairable system with unreliable service stations was discussed by Wang et al. [21].

The classical $M/E_k/1$ queueing system was discussed by Gross and Harris [4]. The purpose of this study is to derive the steady state queue size distribution for an $M/E_k/1$ queueing model with a removable service station under steady state conditions based on the principle of maximum entropy which uses the information available for mean values and the probability of the server being in different states. Section 2 presents the model description and mathematical formulation. The probability distribution for the queue length of $M/E_k/1$ queueing model using principle of an maximum entropy is derived in Section 3. Particular cases are discussed in Section 4. In Section 5, other performance measures of an $M/E_k/1$ queueing model with removable service station are discussed. A numerical illustration of the queue size distribution is given in Section 6. Finally, applications of such model are outlined in Section 7.

2 The model description In this section, we consider a single removable server queueing system with the following assumptions (cf. Wang and Huang [18]):

(i) There is only one server who renders service to the arriving units.
(ii) The interarrival time of the units follows a negative exponential distribution.
(iii) The service time of the units follow an Erlang $k$-type distribution with mean $1/\mu$. Each stage is independent and identically exponential distributed with mean $1/(k\mu)$.
(iv) The server can serve only one unit at a time, and it takes zero set-up time to restart the service station.
(v) The units arriving at the service station form a single waiting line and are served in the order of their arrival.
(vi) If the service station is busy, then a newly arriving unit or waiting units must wait in the queue until the station is available.
(vii) As soon as the system becomes empty, the idle period starts.
(viii) When the server finds at least $N$ units waiting in the system, the server begins the service immediately until the system becomes empty again.
(ix) The next unit will not be taken for service until the previous unit has completed all the stages of the service.

To formulate the mathematical model for $M/E_k/1$ queueing model under $N$-policy, the notations used are as follows:
\( N \): The threshold level of queue size to turn on the server;
\( k \): The number of independent and identical exponential stages in
the Erlang distribution characterizing the service time;
\( \lambda \): The arrival rate of the units;
\( \mu \): The service rate of the units;
\( p_{0,n} \): Probability that there are \( n \) \((0, 1, \ldots, N - 1)\) units in the
system when the server is idle;
\( p_{1,n} \): Probability that there are \( n \) \((\geq 1)\) units in the
system when the server is busy;
\( p_n \): Probability that there are \( n \) units in the system;
\( E(I) \): Expected length of the idle period;
\( E(B) \): Expected length of the busy period;
\( E(C) \): Expected length of the busy cycle;
\( C_h \): Holding cost per unit time for each unit present in the system;
\( C_o \): Operating cost per unit time for the service station in operation;
\( C_s \): Start-up cost per unit time for activating the service station;
\( C_d \): Close down cost per unit time for turning off the service station;
\( C_p \): Cost per unit time for performing an auxiliary task by
the service station during idle period;
\( \rho \): Ratio of arrival rate to service rate, i.e., \( \rho = \lambda / \mu \).

### 3 Maximum entropy principle and queue size distribution

We evaluate the approximate result for the queue size distribution for
queueing model. First we outline briefly the principle of the maximum
entropy which is further employed to derive expression for the approximate
queue size distribution at steady state.

#### 3.1 The principle of maximum entropy

It is well established that if the arrival rate \( \lambda \) is given, the maximum entropy principle implies an
exponential interval arrival time (for the proof, see Jain [7]).

**Theorem.** If the expected number of units is \( L \), then according to maximum entropy principle, the probability distribution of the state \( S \) of the system is given by

\[
(1) \quad p_n = P(S = n) = \frac{L^n}{(1 + L)^{n+1}}; \quad n = 0, 1, \ldots
\]

**Proof.** We maximize the discrete countable entropy defined by

\[
(2) \quad H = - \sum_{n=0}^{\infty} p_n \log p_n,
\]
subject to

\[ 1 = \sum_{n=0}^{\infty} p_n, \]  
\[ L = \sum_{n=0}^{\infty} np_n. \]

By using the same approach as given by Guiasu [5], we obtain equation (1).

3.2 Queue size distribution for \( M/E_k/1 \) queueing model

Similar to Wang and Huang [18], we have analytic closed-form expressions of the expected number of units in an \( N \)-policy \( M/E_k/1 \) queueing system in steady state as follows:

- \( L_{\text{off}} \): The expected number of units in the system when the service station is turned off.
- \( L_{\text{on}} \): The expected number of units in the system when the service station is turned on.
- \( L \): The expected number of units in the system.

Now following Wang and Huang [18], we get

\[ L_{\text{off}} = \frac{(N-1)(1-\rho)}{2}, \]
\[ L_{\text{on}} = \frac{\rho(N+1-\rho N + r)}{2(1-\rho)}, \]
\[ L = \frac{(N-1)}{2} + \frac{\rho(r - \rho + 2)}{2(1-\rho)}. \]

Using equations (5)–(7) in equation (1), we obtain

\[ \hat{p}_{0,n} = \frac{2A^n}{(2 + A)^{n+1}}; \quad n = 0, 1, \ldots, N - 1, \]
\[ \hat{p}_{1,n} = 2(1 - \rho) \left[ \frac{(\rho + B)^n}{(2 - \rho + B)^{n+1}} \right]; \quad n \geq 1, \]
\[ \hat{p}_n = 2(1 - \rho) \left[ \frac{C^n}{(2(1 - \rho) + C)^{n+1}} \right], \]
where \( r = \rho/k \),
\[
A = (N - 1)(1 - \rho), \quad B = \rho(N + r - \rho N), \quad C = (N - N\rho - 1 + 3\rho + \rho r - \rho^2).
\]

When service station is turned off or turned on, the approximate results for the expected number of units in the system based on principle of maximum entropy can be obtained as
\[
\hat{L} = \sum_{n=0}^{N-1} np_{0,n} + \sum_{n=1}^{\infty} np_{1,n}.
\]

The exact result of the expected number of units in the system \((L)\) can be obtained by using equation (7).

4 Particular cases

Now we discuss some models which can be considered as special cases of our model.

(i) When \( N = 1, \ k = 1 \), the present model converts to classical \( M=M=1 \) queueing model.

(ii) When \( k = 1 \), the present model reduces to \( N \) policy single server markovian queueing model.

(iii) When \( k \to \infty \), this model converts to single server deterministic service queuing model under \( N \)-policy.

(iv) When \( k \to \infty, \ N = 1 \), the present model provides results for the classical single server deterministic service queueing model.

5 Performance measures

In this section, we discuss some performance measures of an \( M=E_k=1 \) queueing model under \( N \)-policy at equilibrium.

5.1 Probability measures for long run fraction of time

As in [12], we observe that results of a queueing system with steady state condition for performance measures such as the long-run fraction of time for the server being in different states is identical to an \( N \)-policy \( M=E_k=1 \) queueing system with removable servers [20]. There are three different stages of servers in this queueing model. The computation of long run fraction of time measures can be done by considering the duration of server states as follows:

(i) The idle period \((I)\) is the length of time during which the service station is turned off per cycle.
(ii) The busy period \((B)\) is the length of time during which the service station is turned on per cycle and the units are being served.

(iii) The busy cycle \((C)\) is the duration from the beginning of the last idle period to the beginning of the following next idle period.

Thus the busy cycle is the sum of the idle period and busy period, i.e.,

\[
E(C) = E(I) + E(B).
\]

Now we define steady state probabilities of the server’s states as

- \(P_I\) : Long run probability of the server being idle.
- \(P_B\) : Long run probability of the server being busy.

Using the results stated in [18], we obtain the long run fraction of time, for the server is in idle period, busy period and busy cycle, respectively, as

\[
\frac{E(I)}{E(C)} = 1 - \rho, \quad \frac{E(B)}{E(C)} = \rho, \quad E(C) = \frac{N}{\lambda(1-\rho)}.
\]

The number of cycles per unit time is

\[
\frac{1}{E(C)} = \frac{\lambda(1-\rho)}{N}.
\]

The probability that there is no unit in the system and no station is in service (i.e., the service station is turned off) is given by

\[
P_0 = \frac{1 - \rho}{N}.
\]

5.2 **Expected cost function per unit time** On the line of Pearn and Chang [12], we discuss the expected total cost function per unit time for the \(N\)-policy \(M/E_k/1\) queueing system with decision variable \(N\) by considering various cost elements as considered below:

(i) The holding cost \((C_h)\) can be treated as the penalty cost for delaying service to the units waiting in the system for service.
(ii) The operating cost \((C_o)\) is incurred by the operating service station to provide service for the units.

(iii) The start-up cost \((C_s)\) is incurred each time the service station starts a new operation, when the service station is in turned-off station.

(iv) The close down cost \((C_d)\) is incurred each time the operating service station is removed from the system.

(v) The performing cost \((C_p)\) per unit time of an auxiliary task by the service station.

The expected total cost function per unit time for one unit is given by

\[
TC(N) = C_hL_N + C_0 \frac{E(B)}{E(C)} + (C_s + C_d) \frac{1}{E(C)} + C_p \frac{E(I)}{E(C)}
\]

To determine the optimal value of decision variable \(N (N^*)\), we minimize the total expected cost function per unit time per customer.

6 Numerical illustration In this section, we present numerical examples to evaluate the queue size distribution by using the approximate formula based on principle of maximum entropy as well as exact analytical results derived in Section 3. By fixing the value of \(N\), the approximate queue lengths \(\bar{L}\) and exact queue length \(L\) are obtained using equations (11) and (7) respectively, for different values of traffic intensity \(\rho\) and the number of stages \(k\); the numerical results are summarized in Tables 1 and 2. We examine the effects of variation of parameters traffic density \(\rho\) and number of stages \(k\). We observe that

- For a fixed parameter \(N\) and for different values of \(\rho\) and \(k\), the queue length \(L\) is always greater than approximate queue length \(\bar{L}\).
- For fixed parameters \(N\) and \(k\), both exact queue length \(L\) and approximate queue length \(\bar{L}\) increase (decrease) as \(\rho\) increases (decreases).
- The approximate queue length is at par with the exact queue length as maximum errors do not exceed more than 18%.

7 Discussion In the present investigation, we have obtained the probability distribution for the number of units in the system for an \(N\)-policy \(M/E_k/1\) queueing model with removable service station, by using the maximum entropy principle (MEP). By maximizing Shannon’s
TABLE 1: The effect of number of traffic intensity on the queue length.

<table>
<thead>
<tr>
<th>k</th>
<th>ρ</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>L</td>
<td>4.061</td>
<td>5.065</td>
<td>6.375</td>
<td>10.520</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>4.892</td>
<td>5.366</td>
<td>6.389</td>
<td>10.590</td>
</tr>
<tr>
<td></td>
<td>% Error</td>
<td>16.9</td>
<td>5.6</td>
<td>.22</td>
<td>.66</td>
</tr>
<tr>
<td>5</td>
<td>L</td>
<td>4.042</td>
<td>4.990</td>
<td>6.132</td>
<td>9.529</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>4.873</td>
<td>5.292</td>
<td>6.153</td>
<td>9.726</td>
</tr>
<tr>
<td></td>
<td>% Error</td>
<td>17.1</td>
<td>5.7</td>
<td>.34</td>
<td>2.0</td>
</tr>
<tr>
<td>8</td>
<td>L</td>
<td>4.037</td>
<td>4.971</td>
<td>6.071</td>
<td>8.938</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>4.868</td>
<td>5.273</td>
<td>6.039</td>
<td>9.494</td>
</tr>
<tr>
<td></td>
<td>% Error</td>
<td>17.1</td>
<td>5.7</td>
<td>0.36</td>
<td>5.8</td>
</tr>
</tbody>
</table>

TABLE 2: The effect of number of stages (k) on the queue length for ρ = 0.5.

<table>
<thead>
<tr>
<th>k</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>5.190</td>
<td>5.065</td>
<td>5.023</td>
<td>5.002</td>
<td>4.990</td>
</tr>
<tr>
<td>L</td>
<td>5.489</td>
<td>5.366</td>
<td>5.325</td>
<td>5.304</td>
<td>5.292</td>
</tr>
<tr>
<td>% Error</td>
<td>5.4</td>
<td>5.6</td>
<td>5.7</td>
<td>5.7</td>
<td>5.8</td>
</tr>
</tbody>
</table>

entropy, we have found explicit formula for the queue size distribution subject to constraints expressed in terms of mean arrival rate, mean service rate and mean number of units in the system. In computer communication networks, such queueing models are very helpful in the heavy traffic situation in particular when the transmission of packets is started when some messages are accumulated and service is rendered in a number of stages.

Acknowledgements The authors are thankful to the learned referees and chief editor for their valuable comments and suggestions for the improvement of the paper. First author is thankful to University Grants Commission, New Delhi for financial support under Major Research Project No. 37-191/2009 (SR).
REFERENCES

11. E. Omey and S. van Gulck, *Note on the article: Maximum entropy analysis of the M^x/M/1 queueing system with multiple vacations and server breakdowns*, Comp. Ind. Eng. 54 (2008), 1078–1086.


Corresponding author: Charan Jeet Singh
Department of Mathematics,
Guru Nanak Dev University, Amritsar-143005
E-mail address: cjsmath@gmail.com

Department of Mathematics, Indian Institute of Technology,
Roorkee, Hardwar-247667
E-mail address: madhufma@iitr.ernet.in