Single Server Queue with Server Breakdown Including Priority and Varying Rates

MADHU JAIN

Department of Mathematics, Indian Institute of Technology, Delhi,
Hauz Khas, New Delhi – 110016 (India)
e-mail: madhuj@maths.iitd.ernet.in

Abstract. A single server queueing system with service interruption due to failure of the major/minor service unit has been studied. The job’s inter-arrival time as well as the failure time and the repair time of the major/minor service unit are assumed to be exponentially distributed. The arrival of the jobs to the service facility consisting of major and minor units depends upon the state of the server which may be either operational (partially operational) mode with both units (only major unit) functioning well or in breakdown mode due to failure of the major unit or both units. The failure of service units may occur individually or due to some common cause. The repair rate of the major service unit is also affected with the state of the minor service unit. The repair of the major unit is given pre-emptive priority over the repair of the minor unit. The steady state queue size distribution for various states has been obtained by using generating function method. The average number of jobs in various states, server availability etc. have been derived explicitly.

Keywords: Queue size distribution, priority, interrupted service, generating functions, state dependent rates.

1. Introduction

In many real-life queueing systems, e.g. in computer systems, manufacturing systems and communication systems etc., the server is subject to breakdown[1]. In such the arrival rate of jobs may be influenced by the status of the server which alternates stochastically between operational and failed states. Yechiali and Naor[2], and Shogan[3] developed a single server queueing model with arrival rate depending upon operational or breakdown state of the server. Shanthikumar[4] investigated a single server queue with general service time and operation-dependent server failure. A single-server queueing system with general bulk service and arrival rate dependent on server breakdowns was studied by Jayaraman[5].
In some unreliable service systems, the server is also capable to provide service when there is minor failure in the system although with service rate slower than that when the system is operational in full capacity. This phenomenon is common in computer and communication systems where processors are subject to failure and repair and provide service with slower rate whenever one or more processors are down. Some researchers have dealt with single server queueing model with service interruption including priorities[6,7]. Recently Wartenhorst[8] studied the influence of machine breakdown and limited repair capacity on the performance of N parallel queueing system. Madan[9] analyzed a queueing system with two types of failures having pre-emptive priority to the repair of the major failures over the minor ones. The arrival rate of customers is assumed to be constant.

In this paper, we study a queueing system with a single server facility consisting of two units; major and minor. The server is capable of operating with the major unit only but with a slower rate than when both units are functioning well. The jobs’ arrival rate depends upon the state of the server whether operating (normally/partially) or in breakdown state. The major and the minor units of the system may fail individually or due to some common cause. By inclusion of common cause failure, the model deals with more realistic situations because there are certain external events (e.g., voltage in case of computer systems) which may affect both units simultaneously and may result in their failure. An analytical and explicit queue size distribution for various states by using the generating function method is developed. The steady state mean queue lengths for various states are also derived.

2. The Model and Assumptions

Consider a single server queueing model with interrupted service with the following characteristics:

The system may be in any one of the following states:

(i) Normal operational state (O,O) : Both units of service facility are functioning well.

(ii) Partially operational state (O,R) : The major unit is in operating mode whereas the minor unit is broken down and under repair.

(iii) Failed state (R,F) : Both units are broken down and the major unit is under repair.

(iv) Partially failed state (R,O) : The major unit is broken and is under repair whereas the minor unit is in operating mode. In this state, the system is also shut down.

Assumptions:

- The units are completely rejuvenated after each repair.
- The switch over time from one state to the other is perfect and instantaneous.
- The inter-arrival time, the service time, the failure time and the repair time are exponentially distributed.
- The repair of the minor unit is preempted whenever the major unit fails and the repairman starts repair of major unit. After completing the repair of the major unit, the repairman continues the repairing of the minor one at the same point where he left.
• The repairman repairs the major unit with a faster rate when the minor unit is also non-operative.
• The failure of both units is detected immediately and perfectly.
• The jobs at the service facility are served according to the first in, first out (FIFO) discipline.

The following notations are adopted for our model:

$\lambda_1(\lambda_2)$, $\lambda$  arrival rate of jobs when the system is partially (normally) operating and in failed state respectively.
$\mu(v)$  service rate when the system is operating normally (partially).
$\alpha_1(\alpha_2)$  failure rate the major (minor) unit.
$\alpha_c$  common cause failure rate.
$\beta_1(\beta'_1), \beta_2$  repair rate of the major unit when the system is failed partially (completely) and repair rate of the minor unit respectively.
$p_1(n), p_2(n)$  the steady state probability that there are $n$ jobs in the system when the system is operating partially and normally respectively.
$q_1(n), q_2(n)$  the steady state probability that there are no jobs in the system when the system is in partially-failed and completely-failed states respectively.

It should be noted that the system states (F,R), (R,R) and (F,F) are not possible due to the assumptions of preemptive priority to the repair of the major unit, single repairman and instantaneous switch over from failure to repair state respectively. Figure 1 represents the steady state transition due to failure and repair of the major/minor unit irrespective of the number of jobs present in the system.

Fig. 1. State transition diagram.
3. The Generating Function Method

The difference equations governing the model are:

\[
(\lambda_2 + \alpha_1 + \alpha_2 + \alpha_c) p_2 (0) = \mu p_2 (1) + \beta_1 q_1 (0) + \beta_2 p_1 (0) \quad (1)
\]

\[
(\lambda_2 + \mu + \alpha_1 + \alpha_2 + \alpha_c) p_c (n) = \lambda_2 p_2 (n-1) + p_2 (n+1) + \beta_1 q_1 (n) + \beta_2 p_1 (n), \quad (n \geq 1) \quad (2)
\]

\[
(\lambda_1 + \alpha_1 + \beta_2) p_1 (0) = \nu p_1 (1) + \alpha_2 p_2 (0) + \beta_1' q_2 (0) \quad (3)
\]

\[
(\lambda_1 + \nu + \alpha_1 + \beta_2) p_1 (n) = \lambda p_1 (n-1) + \nu p_1 (n+1) + \alpha_2 p_2 (n) + \beta_1' q_2 (n), \quad (n \geq 1) \quad (4)
\]

\[
(\lambda + \alpha_2 + \beta_1) q_1 (0) = \alpha_1 p_2 (0) \quad (5)
\]

\[
(\lambda + \alpha_2 + \beta_1') q_1 (n) = \lambda q_1 (n-1) + \alpha_1 p_2 (n) \quad (6)
\]

\[
(\lambda + \beta_1') q_2 (0) = \alpha_1 p_1 (0) + \alpha_2 q_1 (0) + \alpha_c p_c (0) \quad (7)
\]

\[
(\lambda + \beta_1') q_2 (n) = \lambda q_2 (n-1) + \alpha_1 p_1 (n) + \alpha_2 q_1 (n) + \alpha_c p_c (n) \quad (n \geq 1) \quad (8)
\]

We define the following generating functions

\[
P_i (z) = \sum_{n=0}^{\infty} p_i (n) z^n \quad i = 1, 2, \quad |z| \leq 1 \quad (9)
\]

\[
Q_i (z) = \sum_{n=0}^{\infty} p_i (n) z^n \quad i = 1, 2, \quad |z| \leq 1 \quad (10)
\]

Multiplying each of equations (1) - (8) by an appropriate power of \( z \) and summing for \( n = 0, 1, \ldots \) and then using (9) and (10), we get

\[
g_2 P_2 (z) = \beta_1 z Q_1 (z) + \beta_2 z P_1 (z) + \mu (z - 1) P_2 (0) \quad (11)
\]

\[
g_1 P_1 (z) = \alpha_2 z P_2 (z) + \beta_1' z Q_2 (z) + \nu (z - 1) P_1 (0) \quad (12)
\]

\[
h_1 Q_1 (z) = \alpha_1 P_2 (z) \quad (13)
\]

\[
h_2 Q_2 (z) = \alpha_c P_2 (z) + \alpha_1 P_1 (z) + \alpha_2 Q_1 (z) \quad (14)
\]

where

\[
g_1 = -\lambda_1 z^2 + (\lambda_1 + \nu + \alpha_1 + \beta_2) z - \nu \quad (15)
\]
Equation (13) can be rewritten as,
\[ Q_1(z) = \frac{\alpha_1}{h_1} P_2(z) \quad . \] (19)

Using equations (14) and (19), we get,
\[ Q_2(z) = \frac{\alpha_1}{h_2} P_1(z) + \left[ \frac{\alpha_1 \alpha_2}{h_1 h_2} + \frac{\alpha_c}{h_2} \right] P_2(z) \quad . \] (20)

From equations (12) and (20) we have,
\[ P_1(z) = \frac{\beta_1}{h_2} z P_2(z) \left[ \frac{\alpha_1 + \frac{\alpha_1 \alpha_2 \beta_1'}{h_1 h_2} + \frac{\alpha_c \beta_1'}{h_2}}{g_1 - \frac{\alpha_1 \beta_2' z}{h_2}} + \frac{\nu(v-1)p_1(0)}{g_1 - \frac{\alpha_1 \beta_2' z}{h_2}} \right] . \] (21)

Substituting values from equations (19) and (21) into equation (11), we find,
\[ P_2(z) = \frac{(z-1)h_1 \{h_2 \beta_2 v z p_1(0) + \mu(g_1 h_2 - \alpha_1 \beta_1' z)p_2(0)\}}{(h_1 g_2 - \alpha_1 \beta_1 z) \{h_2 g_1 - \alpha_1 \beta_2' z\} - \beta_2 z(\alpha_2 (h_1 h_2 + \alpha_1 \beta_1') + \alpha_c \beta_1' h_1}) . \] (22)

In equation eq. (22), \( p_2(z) \) is expressed in terms of two unknowns \( p_1(0) \) and \( p_2(0) \) and \( p_2(0) \) are determined, we can obtain expressions for \( Q_1(z) \), \( Q_2(z) \) and \( P_1(z) \) in terms of \( P_2(z) \) by using equations (19) - (21). Now first we determine the limiting values of these generating functions as \( z \to 1 \), in terms of \( \lim_{z \to 1} P_2(z) \) as follows;
\[ \lim_{z \to 1} Q_1(z) = \left\{ \frac{\alpha_1}{\alpha_2 + \beta_1} \right\} \lim_{z \to 1} P_2(z) \quad , \] (23)
\[ \lim_{z \to 1} P_1(z) = \left\{ \frac{\alpha_2 (\alpha_1 + \alpha_2 + \beta_1) + \alpha_c (\beta_1 + \beta_2)}{\beta_2 (\beta_1 + \alpha_2)} \right\} \lim_{z \to 1} P_2(z) \quad , \] (24)
\[ \lim_{z \to 1} Q_2(z) = \left\{ \frac{\alpha_1 \alpha_2 (\alpha_1 + \alpha_2 + \beta_1) + \alpha_c (\beta_1 + \beta_2)(\alpha_2 + \beta_1)}{\beta_1 \beta_2 (\alpha_2 + \beta_1)} \right\} \lim_{z \to 1} P_2(z) \quad . \] (25)

The normalizing condition is given by
\[ \lim_{z \to 1} [Q_1(z) + Q_2(z) + P_1(z) + P_2(z)] = 1 \quad , \] (26)
which provides,
\[
\lim_{z \to 1} P_2(z) = \frac{\beta_1' \beta_2 (\alpha_2 + \beta_1)}{\delta},
\]
where:
\[
\delta = \beta_1' \beta_2 (\alpha_1 + \beta_1) + \alpha_2 (\alpha_1 + \beta_1') (\alpha_1 + \alpha_2 + \beta_1 + \beta_2) + \alpha_c (\alpha_2 + \beta_1) (\alpha_1 + \beta_1' + \beta_2).
\]

Substituting the value of \(\lim_{z \to 1} P_2(z)\) from equation (27) in equations (23-25), we get,
\[
\lim_{z \to 1} Q_1(z) = \frac{\alpha_1 \beta_1' \beta_2}{\delta},
\]
\[
\lim_{z \to 1} R_1(z) = \frac{\beta_1' (\alpha_2 (\alpha_1 + \alpha_2 + \beta_1) + \alpha_c (\beta_1 + \alpha_2))}{\delta},
\]
\[
\lim_{z \to 1} Q_2(z) = \frac{(\alpha_1 \alpha_2 (\alpha_1 + \alpha_2 + \beta_1 + \beta_2) + \alpha_c (\alpha_1 + \beta_2) (\alpha_2 + \beta_1))}{\delta}.
\]

In the limiting case when \(z \to 1\), the numerator and denominator of equation (22) vanish. By applying L-Hospital rule, we compute \(\lim_{z \to 1} P_2(z)\) from equation (22) which after simplification gives
\[
\lim_{z \to 1} P_2(z) = \frac{\beta_1' \beta_2 (\alpha_2 + \beta_1) (\mu p_2(0) + \nu p_1(0))}{\beta_1' \beta_2 (\alpha_2 + \beta_1) \mu + \beta_1' \nu \theta - [\lambda l + \lambda_1 m + \lambda_2 n]},
\]
where,
\[
l = \beta_1' \beta_2 \alpha_1 + \beta_2 (\alpha_2 (\alpha_1 + \alpha_c) + \beta_1 \alpha_c) + \beta_1 \alpha_c \alpha_1 \theta,
\]
\[
m = \beta_1' \theta,
\]
\[
n = \beta_1' \beta_2 (\alpha_2 + \beta_1),
\]
\[
\theta = \alpha_2 (\alpha_1 + \alpha_2 + \alpha_c + \beta_1) + \beta_1 \alpha_c.
\]

From equations (27) and (32), we find
\[
\mu p_2(0) + \nu p_1(0) = \frac{\beta_1' \beta_2 (\alpha_2 + \beta_1) \mu + \beta_1' \nu \theta - [\lambda l + \lambda_1 m + \lambda_2 n]}{\delta},
\]
On comparing both sides, we have
\[
p_2(0) = \frac{\beta_1' \beta_2 (\alpha_2 + \beta_1) - \beta_2 (\lambda \alpha_1 + \lambda_2 \beta_1) + \lambda \alpha_c (\alpha_2 + \beta_1)) / \mu}{\delta}
\]
and
\[
p_1(0) = \frac{\beta_1' \theta - \{ (\lambda \alpha_1 + \lambda_1 \beta_1') \theta + \alpha_2 \beta_2 (\lambda \alpha_1 + \lambda_2 \beta_1') \} / \nu}{\delta}
\]
It is noted from equation (32) that the system is stable if,
\[
\frac{\lambda l + \lambda_1 m + \lambda_2 n}{\beta_1' \beta_2 (\alpha_2 + \beta_1) \mu + \beta_1' \nu \theta} < 1.
\]
From equations (38) and (39), we find

\[
\frac{b_1'(\lambda \alpha_1 + \lambda_2 \beta_1) + \lambda \alpha_c (\alpha_2 + \beta_1)}{b_1'\mu (\alpha_2 + \beta_1)} < 1
\]  
and

\[
\frac{b_1'(\lambda \alpha_1 + \lambda_2 \beta_1) + \lambda \alpha_c (\alpha_2 + \beta_1)}{b_1'\theta \nu} < 1
\]

4. Mean Queue Length

Let us denote the expected number of jobs in the system when the service channel is operating normally with both units and partially with only the major unit by \(L_{\text{ON}}\) and \(L_{\text{OP}}\) respectively. Also denote the expected number of jobs in the system when the service channel is in a failed state due to failure of both units and only the major unit by \(L_{\text{FC}}\) and \(L_{\text{FP}}\) respectively. We determine the expected number of jobs in the system for various states as follows:

\[
L_{\text{FP}} = \lim_{z \to 1} Q'_1(z) = \frac{\alpha_1}{h_1} \lim_{z \to 1} P_2 (z) + \frac{1}{h_1} \lim_{z \to 1} P'_2(z),
\]

\[
L_{\text{FC}} = \lim_{z \to 1} Q'_2(z) = \frac{\alpha_1}{b_1'} \left\{ \lambda \lim_{z \to 1} P_1 (z) + \beta_1 \lim_{z \to 1} P'_1(z) \right\},
\]

\[
+ \frac{\lambda}{b_1'} \left\{ \alpha_c + \frac{\alpha_1 \alpha_2}{h_1} + \frac{\alpha_1 \alpha_2 b_1'}{h_1^2} \right\} \lim_{z \to 1} P_2 (z)
\]

\[
+ \frac{\lambda}{b_1'} \left\{ \alpha_c + \frac{\alpha_1 \alpha_2}{h_1} \right\} \lim_{z \to 1} P'_2(z),
\]

\[
L_{\text{OP}} = \lim_{z \to 1} P'_1(z) = (\lambda (\alpha_1 \alpha_2 \beta_2 b_1' + h_1 w + \alpha_1 \theta) + \beta_1 ' \theta (\lambda_1 - \nu)) b_1 \lim_{z \to 1} P_2 (z)
\]

\[
+ \beta_2' \beta_2 h_1 \{ s \lim_{z \to 1} P'_2(z) + v \theta p_1(0) \},
\]

where

\[
w = \beta_2 (\alpha_1 \alpha_2 + \alpha_2 \alpha_c + \alpha_c \beta_1),
\]

\[
L_{\text{ON}} = \lim_{z \to 1} P'_2(z) = \frac{N''(1)D'(1) - N'(1)D''(1)}{[2D'(1)]^2},
\]

where

\[
D'(1) = \delta,
\]

\[
N'(1) = b_1' b_2 (\alpha_2 + \beta_1) \{ \mu p_2(0) + \nu p_1(0) \},
\]

\[
D''(1) = 2 \lambda [-(-\lambda_1 + \alpha_1 + \beta_2 + \nu) \theta + b_1' \alpha_1 (\alpha_1 + \alpha_2 + \alpha_c) - \beta_2 (-\lambda_2 + \alpha + \mu)] + \alpha_1 b_1 (\alpha_1 + \beta_2)
\]
Some Special Cases

(i) In case of identical arrival, failure and repair rates, i.e., when \( \lambda_1 = \lambda_2 = \lambda \), \( \alpha_1 = \alpha_2 = \alpha \), \( \beta_1 = \beta_2 = \beta \) equations (38) and (39) reduce to

\[
p_2(0) = \frac{\beta(\beta(\mu - \lambda) - \lambda \alpha)}{\mu((\alpha + \beta)^2 + \alpha^2 + \alpha_c(\alpha + \beta))}, \tag{48}
\]

\[
p_1(0) = \frac{v \beta \theta - \lambda(\alpha + \beta)^2(2\alpha + \alpha_c)}{v(\alpha + \beta)((\alpha + \beta)^2 + \alpha^2 + \alpha_c(\alpha + \beta))}, \tag{49}
\]

where

\[
\theta = \alpha(2\alpha + \beta) + \alpha_c(\alpha + \beta) .
\]

Equations (29) - (31) and (32) provide;

\[
\lim_{z \to 1} Q_1(z) = \frac{\alpha \beta^2}{(\alpha + \beta)((\alpha + \beta)^2 + \alpha^2 + \alpha_c(\alpha + 2\beta))}, \tag{50}
\]

\[
\lim_{z \to 1} P_1(z) = \frac{\beta(\alpha(2\alpha + \beta) + \alpha_c(\alpha + \beta))}{(\alpha + \beta)((\alpha + \beta)^2 + \alpha^2 + \alpha_c(\alpha + 2\beta))}, \tag{51}
\]

\[
\lim_{z \to 1} Q_2(z) = \frac{2\alpha^2 + \alpha_c(\alpha + \beta)}{(\alpha + \beta)^2 + \alpha^2 + \alpha_c(\alpha + 2\beta)}, \tag{52}
\]

\[
\lim_{z \to 1} P_2(z) = \frac{\beta^2}{\beta^2 + 2\alpha(\alpha + \beta) + \alpha_c(\alpha + 2\beta)}, \tag{53}
\]

(ii) When there is no common cause failure (i.e. \( \alpha_c = 0 \)), equations (47-49) and equations (50-53) of case (i) coincide with the results as obtained by Madan\[9\] (see his equations (34) and (33) respectively).

The fraction of time for which the server is operational (normally or partially) is given by

\[
A = \lim_{z \to 1} P_1(z) + \lim_{z \to 1} P_2(z)
\]

\[
= \frac{\beta_1^' \{\alpha_1 \alpha_2 + (\alpha_2 + \beta_1)(\alpha_2 + \alpha_c + \beta_2)\}}{\delta} \tag{54}
\]
The fraction of time for which the server is failed (completely or partially) is given by

\[ \bar{A} = 1 - A. \]  

(55)

**Conclusion**

An analytical and explicit solution for single server queue with varying rates and service interruption due to failure of the major as well as the major and the minor service units is developed. Our model generalizes Madan’s[9] model by incorporating common cause failure and arrival rates of jobs dependent upon server’s state. We have also included state dependent repair rate for the major unit so that the considered model can be fitted to more realistic situations where the repairman works with a faster rate in order to restart the interrupted repair of the minor unit.

**Acknowledgements**

This work is supported by University Grant Commission, New Delhi (India) under UGC career award project No. F. 10 - 29/93 (SA-III). The helpful comments and suggestions of referees are gratefully appreciated.

**References**


الطابور أحادي الخدمة يتعرض لحالات تعطل
وتشتمل على أسبقيّة ومعدلات متغيرة

مادهو جين
قسم الرياضيات، المعهد الهندسي للتكنولوجيا، دلهي - الهند

المستخلص. قام بدراسة نظام طابور أحادي الخدمة بخدمة تعطل
نتيجة لوجود خطأ ما في الوحدة الرئيسية أو الفرعية.

وقد تم افتراض أن الزمن بين وصول الوظائف وفترة الخدمة وكذلك
زمن العطل زمن القيام بأعمال الإصلاح للوحدة الرئيسية أو الفرعية
للخدمة كلها ذات توزيع أسي.

ويكون وصول الأعمال لإمكانية الخدمة من وحدات أساسية وآخرى
فرعية معتمدًا على حالة الخادم التي إذاً أن تكون فعالة (فعالة جزئيًا) مع
كون كلا الوحدتين (مع الوحدة الرئيسية فقط) تعمل جيدة، أو أن بها
عطل نتيجة لوجود خطأ في الوحدة الرئيسية أو في كلا الوحدتين. فخطأ
وحدة الخدمة ربما يحدث فردًا أو نتيجة لسبب مشترك.

وبطبيعة الحال سوف يكون معدل إصلاح الوحدة الرئيسية أيضًا متاثرًا
بحالة وحدة الخدمة الفرعية فإصلاح الوحدة الرئيسية يكون أعلى
بالأسبقيّة على إصلاح الوحدة الفرعية.

إن دراسة حالة التوزيع لحجم الطابور لمختلف الحالات قد تم الحصول
عليه باستخدام طريقة الدالة المولدة، وكذلك فإن متغيرات متوسط عدد
الأعمال في مختلف الحالات وتوافر الخدمة وما إلى ذلك قد تم اشتقاقها
في صورة صريحة.