Queueing model with non-reliable server and threshold-based recovery

Dmitry Efrosinin
Dept. of Prob. Theory and Math. Statistics
Peoples' Friendship University of Russia
117198 Moscow
Russia
dmitriy_e@mail.ru

Olga Semenova
ZAO RPC "INSET"
Staroalekseevskaja str., 5-215
129626 Moscow
Russia
olgasmnv@gmail.com

Abstract

In this paper we consider a queueing system $M/M/1$ with non-reliable server. When the server is in normal state, the service error (or failure) occurs according to a Poisson process. In the error state the server requires to be repaired at a repair facility with exponential repair time but according to the threshold policy it can be done only if the number of customers in the system reaches some prespecified threshold level $q \geq 1$. We perform a steady-state analysis of the corresponding continuous-time Markov chain and calculate optimal threshold level to minimize the long-run average losses given cost structure.[1]

1 Introduction

In the present paper we deal with the optimal control of non-reliable server in an $M/M/1$ queue. While serving a customer, an error or failure may occur subjecting to service interruption and transition to an error state. We set up a control threshold $q \geq 1$ which specifies a number of customers waiting in the queue. A repair policy for such system is a threshold-based control: when the system is in error state the recovery can be performed only if the number of customers in the queue reaches a given threshold level q.

Threshold-based policies have been investigated by many authors in applications to queues where the service rates, number of servers, server activation or vacation were under control. In (Efrosinin 2008), a multi-server heterogeneous system $MAP/PH/K$ is presented where a proof is given that the policy minimizing the mean number of customers in the system is of threshold type, i.e. the server $1 \leq j \leq K$ must be activated only if the number of customers in the queue exceeds a threshold level q_j. Threshold policy to control the service rate in multi-server queueing system with homogeneous servers was investigated in (Semenova and Dudin 2007). Wang (2003) considered a queueing system $M/M/1$ with removable and non-reliable server controlled under N-policy that turns the server on whenever the number of waiting customers reaches a given threshold level N and turns it off otherwise. Heyman and Sobel (1984) investigated an $M/G/1$ queueing system with vacations and threshold start-up policy when server ends a vacation only when the queue length reaches a given threshold.

The problem considered in this paper differs from previous works since it studies the queueing system with controllable recovery. We first develop analytical steady-state results for threshold policy q and next we construct the long-run average cost per unit time to calculate an optimal policy q^*.

2 Model

We consider M/M/1 queueing system with non-reliable server. Customers arrive to the system accordingly to a Poisson process with parameter λ. Service times are exponentially distributed with parameter μ.

[1]The research was supported by the Russian Foundation for Basic Research in the framework of 08-07-90102 project.

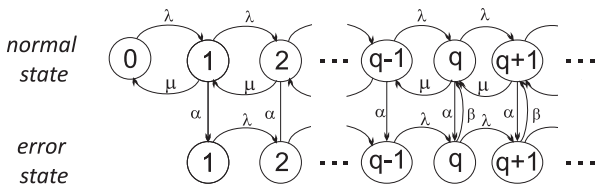

Figure 1: The state-transition-rate diagram for the threshold policy q

The server can be either in normal state or in an error state. While serving a customer in a normal state, server enters an error state after the period of time exponentially distributed with parameter α, stops the service process and waits for a repair. The repair time is exponentially distributed with parameter β and is started only when the number of customers in the system reaches the fixed level q ($q \geq 1$). When the repair time expires, the server gets back to a normal state and resumes processing the customer whose service was interrupted by an error.

In Section 3, we provide a steady-state analysis by means of probability generating functions, and in Section 4 we briefly run through optimization problem.

3 Steady-state analysis

Denote by $N(t)$ the number of customers in the system and by $D(t)$ the state of server at time t that takes the value 0 when the server is in an error state, and 1 when it is in normal state. The stochastic process $\{X(t)\}_{t\geq 0} = \{N(t), D(t)\}_{t\geq 0}$ is a continuous-time Markov chain with state space $E = \{x = (n, d) | n \in \mathbb{N}_0, d \in \{0, 1\}\}$. We define the stationary probabilities
$\pi_0(n) = \lim_{t\to\infty} \mathbb{P}[N(t) = n, D(t) = 0]$ that there are n customers in the system and the server is in error state, $n \geq 1$;
$\pi_1(n) = \lim_{t\to\infty} \mathbb{P}[N(t) = n, D(t) = 1]$ that there are n customers in the system and the server is in normal state, $n \geq 0$.

The state-transition-rate diagram for the threshold-based policy q is shown in Figure 1, and the system of balance equations is given as follows:

$$(\lambda + \beta I_{\{n\geq q\}})\pi_0(n) = \lambda\pi_0(n-1)I_{\{n>1\}} + \alpha\pi_1(n), \qquad n \geq 1, \quad (1)$$

$$(\lambda + (\mu + \alpha)I_{\{n>0\}})\pi_1(n) = \lambda\pi_1(n-1)I_{\{n>0\}} + \mu\pi_1(n+1) + \beta\pi_0(n)I_{\{n\geq q\}}, \quad n \geq 0, \quad (2)$$

where $I_{\{A\}} = 1$ if A is true, and $I_{\{A\}} = 0$ otherwise. To solve equations (1)–(2), we use the probability generating functions (for $|z| \leq 1$)

$$P_1(z) = \sum_{n=1}^{q-1} z^n\pi_0(n),\ P_2(z) = \sum_{n=q}^{\infty} z^n\pi_0(n),\ P_3(z) = \sum_{n=1}^{q-1} z^n\pi_1(n),\ P_4(z) = \sum_{n=q}^{\infty} z^n\pi_1(n). \quad (3)$$

Theorem 1. *The generation functions $P_i(z)$, $i = \overline{1,4}$, defined in (3) satisfy the following relations:*

$$P_1(z) = \frac{\alpha\lambda z - (\lambda(\lambda z - \mu)\Gamma(q-1) + \mu(\alpha - \lambda z + \mu)\Gamma(q))z^q}{\lambda(\lambda z^2 - (\alpha + \lambda + \mu)z + \mu)}\pi_1(0),$$

$$P_2(z) = \frac{(\lambda(\lambda z - \mu)\Gamma(q-1) + \mu(\alpha - \lambda z + \mu)\Gamma(q))z^q}{\lambda^2 z^2 - \lambda(\alpha + \beta + \lambda + \mu)z + (\beta + \lambda)\mu}\pi_1(0),$$

$$P_3(z) = \frac{\lambda(1-z)z + (\lambda z\Gamma(q-1) - \mu\Gamma(q))z^q}{\lambda z^2 - (\alpha+\lambda+\mu)z + \mu}\pi_1(0),$$
$$P_4(z) = \frac{(-\lambda^2 z\Gamma(q-1) + (\beta+\lambda)\mu\Gamma(q))z^q}{\lambda^2 z^2 - \lambda(\alpha+\beta+\lambda+\mu)z + (\beta+\lambda)\mu}\pi_1(0),$$

where $\Gamma(n) = g_1\gamma_1^n + g_2\gamma_2^n$, $n = \overline{1,q}$, $\gamma_{1,2} = \frac{\alpha+\lambda+\mu \pm \sqrt{(\alpha+\lambda+\mu)^2 - 4\lambda\mu}}{2\mu}$, $g_1 = \frac{\gamma_2(\gamma_1-1)}{\gamma_1-\gamma_2}$, $g_2 = \frac{\gamma_1(1-\gamma_2)}{\gamma_1-\gamma_2}$.

The probability $\pi_1(0)$ *is a unique solution of the normalizing equation* $\pi_1(0) + \sum_{i=1}^{4} P_i(1) = 1$ *and satisfies*

$$\pi_1(0) = \frac{\alpha\lambda(\beta\mu - (\alpha+\beta)\lambda)}{\beta\mu(-\lambda(\mu-\lambda)\Gamma(q-1) + \mu(\alpha-\lambda+\mu)\Gamma(q))}.$$

Theorem 2. *The system is stable if and only if* $\rho = \frac{\lambda}{\mu}\left(1 + \frac{\alpha}{\beta}\right) < 1$, $\beta > 0$ *and* $\mu > 0$.

Now we define the probabilities p_1 that the server is in error state and can not be repaired (the number of customers $N(t) < q$), p_2 that the server is in error state and is repaired, p_3 that the server is in normal state with $N(t) < q$, and p_4 that the server is in normal state with $N(t) \geq q$. Note that $p_i = P_i(1)$, $i = 1, 2, 4$, and $p_3 = P_3(1) + \pi_1(0)$. From Theorem 1 we have

$$p_1 = \frac{F(q) - \alpha\lambda}{F(q)}C_1, \quad p_2 = C_2, \quad p_3 = \frac{\lambda(\alpha + G(q))}{F(q)}C_1, \quad p_4 = \frac{H(q)}{F(q)}C_2,$$

where $F(q) = -\lambda(\mu-\lambda)\Gamma(q-1) + \mu(\alpha-\lambda+\mu)\Gamma(q)$,

$G(q) = -\lambda\Gamma(q-1) + \mu\Gamma(q)$, $H(q) = -\lambda^2\Gamma(q-1) + (\beta+\lambda)\mu\Gamma(q)$, $C_1 = 1 - \frac{\alpha+\beta}{\alpha}C_2$, $C_2 = \frac{\alpha\lambda}{\beta\mu}$.

The mean number of customers in the system for the given threshold q is defined by

$$\bar{N}(q) = \bar{N}_1(q) + \bar{N}_2(q) + \bar{N}_3(q) + \bar{N}_4(q), \tag{4}$$

where $\bar{N}_i(q) = \frac{d}{dz}P_i(z)\big|_{z=1}$, $i = \overline{1,4}$. By using Theorem 1, the equation (4) results in

$$\begin{aligned}\bar{N}(q) = &-\frac{\alpha-\lambda+\mu}{\alpha}C_1 + \frac{\alpha+\beta+\mu-\lambda}{\alpha}\frac{C_2^2}{C_1} + \lambda\mu C_1\frac{1}{F(q)} + \\ &+ \frac{-\lambda\alpha(\mu-\lambda) + \alpha\lambda\mu C_1 q - \lambda\mu(\alpha-\lambda+\mu)C_1}{\alpha\mu}\frac{G(q)}{F(q)} + \\ &+ \frac{\alpha C_1 C_2 q + (\alpha+\beta+\mu-\lambda)C_2^2}{\alpha C_1}\frac{H(q)}{F(q)} - \frac{\lambda^2(\mu-\lambda)}{\mu}\frac{\Gamma(q-1)}{F(q)} + \frac{\mu-\lambda}{\mu}q.\end{aligned}$$

4 Optimal threshold policy

Denote by $\bar{B}$ the mean regeneration cycle that represents the time period starting from the arrival of a customer to the empty system to the next arrival to the empty system,

$$\bar{B} = \bar{B}_1 + \bar{B}_2 + \bar{B}_3 + \bar{B}_4 = \frac{F(q)}{\alpha\lambda^2 C_1},$$

where the expected length of the error period when server is not recovered, the period of server recovery, the normal period with $N(t) < q$ and normal period with $N(t) \geq q$ are denoted by $\bar{B}_i$,

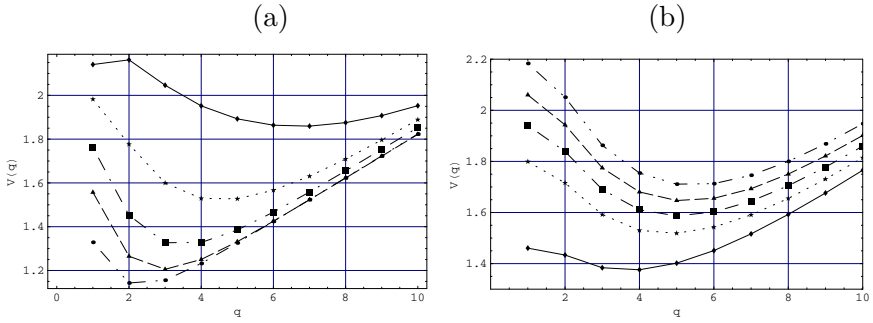

Figure 2: Long-rung average cost $V(q)$ versus q and α (a) and β (b)

$i = \overline{1,4}$, respectively. These values can be computed using the long-run fraction of time for the corresponding periods as $\bar{B}_i = p_i \bar{B}$, namely

$$\bar{B}_1 = \frac{F(q) - \alpha\lambda}{\alpha\lambda^2},\ \bar{B}_2 = \frac{1}{\alpha\lambda^2}\frac{C_2}{C_1}F(q),\ \bar{B}_3 = \frac{\alpha + G(q)}{\alpha\lambda},\ \bar{B}_4 = \frac{1}{\alpha\lambda^2}\frac{C_2}{C_1}H(q).$$

Since $\frac{\bar{B}_2}{\bar{B}}$ does not depend on the decision parameter q, we omit the cost for repair time. To find an optimal threshold q, we consider the following cost structure:

$$\begin{aligned}\bar{V}(q) &= c_h\bar{N}(q) + \frac{c_1\bar{B}_1 + c_3\bar{B}_3 + c_4\bar{B}_4}{\bar{B}} + \frac{c_r}{\bar{Y}} = \\ &= c_h\bar{N}(q) + c_1C_1\left(1 - \frac{\alpha\lambda}{F(q)}\right) + c_3\frac{C_1\lambda}{F(q)}(\alpha + G(q)) + c_4\frac{C_2H(q)}{F(q)} + c_r\frac{\alpha^2\lambda^2C_1}{\mu F(q)}(\gamma_1^q - \gamma_2^q),\end{aligned}$$

where c_h is the holding cost per unit time for each customer in the queue (including one being served), c_1 is the error cost per unit time when the server is in error state and is not repaired, c_3 (c_4) is the operating cost per unit time for the server in normal state with the number of customers $N(t) < q$ ($N(t) \geq q$), and c_r is a fixed cost for switching on a repair facility. Here the value $\bar{Y}$ denotes the mean period between two successive visits of the state $(q, 0)$, where the repair starts. After some algebra we obtain $\bar{Y} = \frac{\mu\bar{B}}{\alpha(\gamma_1^q - \gamma_2^q)}$. Setting $\frac{d}{dq}\bar{V}(q) = 0$, we can numerically calculate the optimal threshold level q^*. If q^* is not an integer, we take the nearest to q^* positive integer value q as a solution. Figure 2 shows the influence of α and β on $V(q)$. We observe that the optimal policy q^* decreases with increasing parameter α and decreasing parameter β.

References

[1] Efrosinin, D. (2008). *Controlled queueing systems with heterogeneous servers. Dynamic optimization and monotonicity properties*. Saarbrücken: VDM Verlag.

[2] Semenova, O.V. and Dudin, A.N. (2007). $M/M/N$ queueing system with controlled service mode and disasters. *Automatic Control and Computer Science, 41, 350-357.*

[3] Wang, K.-H. (2003). Optimal control of a removable and non-reliable server in an $M/M/1$ queueing system with exponential startup time. *Mathematical methods of operations research 58, 29–39.*

[4] Heyman, D.P. and Sobel, M.J. (1984). *Stochastic models in operations research*, Vol. II: Stochastic Optimization. New York: McGraw-Hill.