ANALYSIS OF THE MMPP/G/1/K QUEUE WITH A MODIFIED STATE-DEPENDENT SERVICE RATE

DOO IL CHOI¹, BOKEUN KIM², AND DAE-EUN LIM^{3†}

¹DEPARTMENT OF APPLIED MATHEMATICS, HALLA UNIVERSITY, WONJUSI 220-712, KOREA *E-mail address*: dichoi@halla.ac.kr

²SCM CONSULTING GROUP 2, SAMSUNG SDS, SEONGNAMSI 463-710, KOREA *E-mail address*: bokeun.kim@samsung.com

³DIVISION OF BUSINESS AND COMMERCE, BAEKSEOK UNIVERSITY, CHEONANSI 330-704, KOREA *E-mail address*: del@bu.ac.kr

ABSTRACT. We analyze the MMPP/G/1/K queue with a modified state-dependent service rate. The service time of customers upon service initiation is changed if the number of customers in the system reaches a threshold. Then, the changed service time is continued until the system becomes empty completely, and this process is repeated. We analyze this system using an embedded Markov chain and a supplementary variable method, and present the queue length distributions at a customer's departure epochs and then at an arbitrary time.

1. INTRODUCTION

In this paper, a finite capacity queueing system with a modified state-dependent service rate is analyzed. Queueing systems with finite buffers exist in a wide variety of applications such as computer systems, telecommunication networks, and production lines, among others. While operating systems with a queue, in which the arrivals at the systems and the service of customers (packets or lots) occur randomly, some customers may suffer long delays or be blocked. This can finally lead to a situation in which the delay requirements of users are not satisfied. Possible solutions to this problem is to control the arrival or the service rate. For the queueing model with queue length dependent arrival rate, refer to Choi et al. [3]. A variable service rate depending on the queue length considered in this paper operates as follows: when the number of customers in the system exceeds the threshold, the service rate is increased to a certain value to serve customers more quickly. The increased service rate continues until the

Received by the editors August 28 2014; Revised November 10 2014; Accepted in revised form November 17 2014; Published online November 25 2014.

²⁰⁰⁰ Mathematics Subject Classification. 90B22.

Key words and phrases. State-dependent service, Controllable service rate, Finite queue, Markov modulated Poisson process.

[†] Corresponding author.

system becomes empty for the first time, and then the service rate is reduced to its original value.

The scheme-variable service characteristics based on the state of the system have been extensively applied in real-world applications. For example, systems adapting service speed based on the queue length can be found in telecommunication systems such as that proposed by Choi et al. [4]. By applying a cell-discarding scheme for voice packets in ATM (Asynchronous Transfer Mode) networks, Choi et al. [4] analyzed the $M/G_1, G_2/1$ queue. Also, the applications in call centers can be found in Bekker et al. [1].

For analytical approaches, Choi et al. [4] obtained analytical expressions by means of integral representations and devised an asymptotic approximation for the system size distribution. Choi et al. [5] analyzed the $M^X/G/1$ model with queue-length dependent service times using Markov renewal theory and presented the queue length distributions including the transient distribution at time t and its limiting distributions. Also, the virtual waiting time distribution was presented. Choi et al. [4] and Choi et al. [5] aptly summarized previous work on queueing systems with queue-length dependent service times. These results dealt with one threshold policy, and the two thresholds policy can be found in Dudin [6], Nishimura and Jiang [7]. Nobel and Tijms [8], Zhernovyi and Zhernovyi [9, 10]. Specifically, Zhernovyi and Zhernovyi [9, 10] analyzed the finite queueing model with the two thresholds policy using the Korolyuk potential method. They gave the Laplace transform for the distribution of the number of customers during a the busy period, the distribution function for the busy period, the mean duration of the busy period, and the formula for the stationary distribution of the number of customers and other measures. Our model differs from previous works in that it considers a Markov-modulated Poisson process (MMPP) as the arrival process of customers. In many realistic situations, particularly in telecommunication systems, there is a correlation between the inter-arrival times of customers (or packets) and the degree of burstiness. It requires the use of correlated arrival models rather than models assuming Markoivan arrival streams [2]. The MMPP is used to model traffic streams with bursty characteristics and time correlations between inter-arrivals. For example, traffic such as voice and video in telecommunication networks has these properties. We claim that our model extends previous ones on queues with a state-dependent service rate with one threshold by considering the MMPP. The MMPP is assumed as the arrival process of customers that have not been considered; this assumption can be more suitably utilized in real-world problems such as telecommunication systems.

The remainder of this paper is organized as follows. Section 2 describes the mathematical model. In Section 3, queue length distributions upon departure and arbitrary epochs are presented. We first derive the queue length distribution at a customer's departure epochs using an embedded Markov chain. Next, the queue length distribution at an arbitrary time is obtained using a supplementary variable method. Other performance indices, such as the loss probability and the mean queue length, are also presented. Section 4 gives numerical examples and Section 5 concludes the paper.

2. MODEL DESCRIPTION

There are a single server and a buffer with finite capacity K including a customer in service. Customers arrived when the buffer is full are blocked and lost. The customers are served by first-come, first-served (FCFS) approach based on their arrival order. The service times of customers are different depending on queue length. Specifically, if the number of customers in the system is less than the threshold value L at the service initiation, the customers have the service time S_1 with distribution function G_1 , a mean of μ_1 and the Laplace transform $G_1^*(s)$. If the number of customers in the system is equal to or greater than the threshold value L at the service initiation, the customers have the service time S_2 with distribution function G_2 , a mean of μ_2 , and the Laplace transform $G_2^*(s)$. This service time (S_2) of customers continues until the system becomes empty. Then, the customers are served by the service time S_1 again until the number of customers in the system reaches the threshold L. We assume $\mu_2 \leq \mu_1$ because the faster services are required when there are relatively more customers in the buffer.

The arrival process of customers is assumed to follow an MMPP with representation (Q, Λ) . Here, the matrix Q is the infinitesimal generator matrix of an underlying Markov process J(t) with state space $\{1, 2, \dots, N\}$. And the matrix $\Lambda = \text{diag}(\lambda_i)$ is the arrival rate matrix. The stationary probability vector π of the underlying Markov process J(t) is given by solving the equation

$$\pi Q = 0, \qquad \pi e = 1,$$

where e and 0 are vectors of size N consist of all ones and zeroes, respectively. Let M(t) be the number of arrivals by Λ during the interval (0, t]. At this stage, we define the conditional probabilities:

$$p_{i,j}(n,t) = Pr\{M(t) = n, J(t) = j \mid M(0) = 0, J(0) = i\}, n \ge 0.$$

Then, the matrix P(n,t) is defined as $P(n,t) \triangleq (p_{i,j}(n,t))_{1 \le i,j \le N}$.

3. ANALYSIS

3.1. Queue length distribution at departure epochs. First, the queue length distribution at the departure epochs of customers is considered. We term the period in which the service time of customers is generated by the service time S_1 as the underload period and the period in which the service time of customers is generated by the service time S_2 as the overload period. As soon as the system becomes empty, the underload period is started. The overload period starts from when the number of customers in the system at service initiation reaches the threshold L to the instant when the system becomes empty.

Now we introduce the notations:

 τ_n = the *n*th customer's departure epoch, $n \ge 1, \tau_0 = 0$, N_n = the number of customers in the system at time τ_n +, $\xi_n = \begin{cases} 1, & \text{if the system is in the underload period at time } \tau_n + \\ 2, & \text{if the system is in the overload period at time } \tau_n +, \end{cases}$

 J_n = the state of the underlying Markov process at time τ_n + .

Then, the process $\{(N_n, \xi_n, J_n), n \ge 0\}$ form a Markov chain with finite state space in lexicographic order :

$$\{ (0,1,1), \cdots, (0,1,N), (0,2,1), \cdots, (0,2,N), \cdots, \\ (L-1,1,1), \cdots, (L-1,1,N), \cdots, (L,1,1), \cdots, (L,1,N), \cdots \\ (L+1,2,1), \cdots, (L+1,2,N), \cdots, (K-3,1,1), \cdots, (K-3,2,N), \cdots, \\ (K-2,1,1), \cdots, (K-2,2,N), (K-1,1,1), \cdots, (K-1,2,N) \}.$$

Note that if $N_n = 0$, then $\xi_n = 1$, and if $N_n \ge L$, then $\xi_n = 2$. We define the steady-state probability of the Markov chain $\{(N_n, \xi_n, J_n), n \ge 0\}$ as follows:

 $x_{k,r,j} = \lim_{n \to \infty} \Pr\{N_n = k, \xi_n = r, J_n = j\}, \quad 0 \le k < K, r = 1, 2, j = 1, 2, \cdots, N.$

Also, the vectors are defined.

$$x_{k,r} = (x_{k,r,1}, \cdots, x_{k,r,N})$$
$$x_k = (x_{k,1}, x_{k,2})$$
$$x = (x_0, x_1, \cdots, x_{K-1}).$$

Note that $x_{0,2} = 0$, $x_{k,1} = 0$ for $L \le k < K$. We introduce the following probability matrices:

$$\begin{split} A_n^r &= \int_0^\infty P(n,x) dG_r(x), r = 1,2\\ A_n^{'} &= \int_0^\infty P(0,t) dt \Lambda A_n^1 = (\Lambda - Q)^{-1} \Lambda A_n^1,\\ \overline{A}_n^r &= \sum_{k=n}^\infty A_k^r, \quad \overline{A}_n^{'} = \sum_{k=n}^\infty A_k^{'}. \end{split}$$

Also, the following matrices are introduced:

$$B_{k} = \begin{pmatrix} A'_{k} & 0\\ 0 & 0 \end{pmatrix}, \ 0 \le k \le L - 1, \qquad B'_{k} = \begin{pmatrix} 0 & A'_{k}\\ 0 & 0 \end{pmatrix}, \ k \ge L,$$
$$C_{k} = \begin{pmatrix} A^{1}_{k} & 0\\ 0 & A^{2}_{k} \end{pmatrix}, \quad C'_{0} = \begin{pmatrix} A^{1}_{0} & 0\\ A^{2}_{0} & 0 \end{pmatrix}, \quad C'_{k} = \begin{pmatrix} 0 & A^{1}_{k}\\ 0 & A^{2}_{k} \end{pmatrix},$$
$$D_{k} = \begin{pmatrix} 0 & 0\\ 0 & A^{2}_{k} \end{pmatrix}, \quad k \ge 0,$$

and

$$\overline{B}'_{n} = \begin{pmatrix} 0 & \overline{A}^{1}_{n} \\ 0 & 0 \end{pmatrix}, \quad \overline{C}'_{n} = \begin{pmatrix} 0 & \overline{A}^{1}_{n} \\ 0 & \overline{A}^{2}_{n} \end{pmatrix}, \quad \overline{D}_{n} = \begin{pmatrix} 0 & 0 \\ 0 & \overline{A}^{2}_{n} \end{pmatrix}.$$

298

Then, the transition probability matrix \overline{Q} of the Markov chain $\{(N_n, \xi_n, J_n), n \ge 0\}$ is given by

The steady-state probability vector x of the Markov chain $\{(N_n, \xi_n, J_n), n \ge 0\}$ is given by solving the equations:

$$\mathbf{x} \overline{Q} = \mathbf{x}, \quad \mathbf{x} \mathbf{e} = 1,$$

where $e = (1, 1, \dots, 1)^T$.

3.2. Queue length distribution at an arbitrary time. In this subsection, we derive the probability distribution of the queue length at an arbitrary time. Let N(t) and J(t) be the number of customers in the system and the state of the underlying Markov process at time t, respectively. In addition,

$$\xi(t) = \begin{cases} 1, & \text{if the system is in the underload period at time } t, \\ 2, & \text{if the system is in the overload period at time } t. \end{cases}$$

We define the stationary probabilities:

$$y_n(j) = \lim_{t \to \infty} \Pr\{N(t) = n, J(t) = j\}, \quad 0 \le n \le K.$$

 $y_n = (y_n(1), y_n(2), \cdots, y_n(N)).$

First, by the key renewal theorem, we have

$$y_0(j) = j$$
th component of $\left[\frac{1}{E}x_{0,1}(\Lambda - Q)^{-1}\right]$,

where $E = x_{0,1} \left[(\Lambda - Q)^{-1} e + \mu_1 \right] e + \sum_{n=1}^{L-1} x_{n,1} e \mu_1 + \sum_{n=1}^{K-1} x_{n,2} e \mu_2$ is the mean interdeparture time of customers.

Next, we derive the probabilities $y_n (n \ge 1)$ by using a supplementary variable method. Let \tilde{T} and \hat{T} be the elapsed and remaining service time for the customer in service, respectively.

Furthermore, we define the stationary joint probability distribution of the number of customers in the system and the remaining service time for the customer in service:

$$\alpha_r(n, j, x)dx = \lim_{t \to \infty} \Pr\{N(t) = n, \xi(t) = r, J(t) = j, x < \tilde{T} \le x + dx\},\$$

$$n \ge 1, \quad r = 1, 2,$$

and the Laplace transform of $\alpha_r(n, j, x)$

$$\alpha_r^*(n,j,s) = \int_0^\infty e^{-sx} \alpha_r(n,j,x) dx.$$

$$\alpha_r^*(n,s) = (\alpha_r^*(n,1,s), \cdots, \alpha_r^*(n,N,s)).$$

In order to derive the queue length distribution at an arbitrary time, the number of arrivals of customers during the elapsed service time should be obtained. Thus, we also define the following conditional probability $\beta_r(n, j_1, j_2, x)dx$ as follows:

$$\begin{split} \beta_r(n, j_1, j_2, x) dx &= \lim_{t \to \infty} \Pr\{ \text{ n arrivals of customers during \tilde{T}, $\xi(t) = r,} \\ J(t) &= j_2, x < \hat{T} \leq x + dx | J(\overline{t}) = j_1 \}, \quad n \geq 0, \ r = 1, 2, \end{split}$$

where \overline{t} is the service starting time of the customer serving at time t. We also define the Laplace transform $\beta_r^*(n, j_1, j_2, s)$ of $\beta_r(n, j_1, j_2, x)$ and matrix $\beta_r^*(n, s)$ with $\beta_r^*(n, j_1, j_2, s)$ as (j_1, j_2) -elements:

$$\beta_r^*(n, j_1, j_2, s) = \int_0^\infty e^{-sx} \beta_r(n, j_1, j_2, x) dx$$

$$\beta_r^*(n, s) = (\beta_r^*(n, j_1, j_2, s))_{1 \le j_1, j_2 \le N}, \quad r = 1, 2$$

By conditioning the queue length at last service completion epoch before time t, $\alpha_r^*(n,s)$ satisfy the following equations:

For $1 \leq n < K$,

$$\alpha_1^*(n,s) = \frac{\mu_1}{E} \left[x_{0,1}\beta_1^*(n-1,s) + \sum_{k=1}^{\min\{n,L-1\}} x_{k,1}\beta_1^*(n-k,s) \right],$$

$$\alpha_2^*(n,s) = \frac{\mu_2}{E} \left[\sum_{k=1}^n x_{k,2}\beta_2^*(n-k,s) \right].$$

Using the method introduced by Choi et al. [4], $\beta_r^*(n,s)$ is given as follows:

$$\beta_r^*(n,s) = \frac{1}{\mu_r} \left[\sum_{k=0}^n A_k^r R_{n-k}(s) - G_r^*(s) R_n(s) \right], \qquad r = 1, 2,$$

where $R_n(s) = (sI - \Lambda + Q)^{-1} \{\Lambda(\Lambda - sI - Q)^{-1}\}^n$. Substituting $\beta_r^*(n, s)(r = 1, 2)$ into above equations, and putting s = 0, we obtain the following stationary queue length probabilities at

an arbitrary time. For $1 \le n < K$,

$$y_{n} = \alpha_{1}^{*}(n,0) + \alpha_{2}^{*}(n,0)$$

$$= \frac{1}{E} \left[x_{0,1} \left[\sum_{k=0}^{n-1} A_{k}^{1}(Q-\Lambda)^{-1} \{\Lambda(\Lambda-Q)^{-1}\}^{n-1-k} - (Q-\Lambda)^{-1} \{\Lambda(\Lambda-Q)^{-1}\}^{n-1} \right] \right]$$

$$+ \sum_{k=1}^{\min\{n,L-1\}} x_{k,1} \left[\sum_{m=0}^{n-k} A_{m}^{1}(Q-\Lambda)^{-1} \{\Lambda(\Lambda-Q)^{-1}\}^{n-k-m} - (Q-\Lambda)^{-1} \{\Lambda(\Lambda-Q)^{-1}\}^{n-k} \right]$$

$$+ \sum_{k=1}^{n} x_{k,2} \left[\sum_{m=0}^{n-k} A_{m}^{2}(Q-\Lambda)^{-1} \{\Lambda(\Lambda-Q)^{-1}\}^{n-k-m} - (Q-\Lambda)^{-1} \{\Lambda(\Lambda-Q)^{-1}\}^{n-k} \right] \right].$$

and

$$y_K = \pi - \sum_{n=0}^{K-1} y_n.$$

Finally, we obtain the following performance measures using the stationary queue length distribution $\{y_n, n \ge 0\}$:

(a) The loss probability (P_{loss}) :

$$P_{\text{loss}} = \frac{y_K \Lambda e}{\sum_{k=0}^{K} y_k \Lambda e}$$

(b) The mean queue length:

$$M = \sum_{i=1}^{K} i y_i e.$$

(c) By Little's law, we obtain the mean waiting time in the system:

$$W = \frac{M}{\lambda^* (1 - P_{\text{loss}})},$$

where $\lambda^* = \pi \Lambda e$.

4. NUMERICAL RESULTS

In this section, we present numerical results on the effects of the modified state-dependent service rate on the mean waiting time and loss probability. We set the capacity of the buffer to K = 10 and the threshold value to L = 5. We assume that the arrivals of customers follow an MMPP with

$$Q = \begin{pmatrix} -q_{12} & q_{12} \\ q_{21} & -q_{21} \end{pmatrix}, \qquad \Lambda = \begin{pmatrix} \lambda_1 & 0 \\ others0 & \lambda_2 \end{pmatrix},$$

and $q_{12}(=q_{21}) = 0.1$ and $\lambda_2/\lambda_1 = 10$. The effective arrival rate λ^* for this MMPP is given by $\lambda^* = \pi \Lambda e$.

To investigate the effect of the modified state dependent service rate, we consider three cases. For the 'High' and 'Low' cases below, queueing systems have two service distributions S_1 and S_2 .

- 'without Threshold' case: The ordinary system without threshold values is assumed. Its service time distribution having mean 2 is assumed to follow hyper-exponential distribution with the probability density function $p\theta_1 e^{-\theta_1 t} + (1-p)\theta_2 e^{-\theta_2 t}$, where $p = 1/3, \theta_1 = 1/4$, and $\theta_2 = 1$.
- 'High' case: S_2 is a hyper-exponential variable with mean 0.5 ($p = 1/4, \theta_1 = 3, \theta_2 = 9/5$). S_1 is assumed to have the same distribution of the 'without Threshold' case (mean=2).
- 'Low' case: S_2 is a hyper-exponential variable with mean 1 ($p = 1/4, \theta_1 = 3/2, \theta_2 = 9/10$). S_1 is assumed to have the same distribution of the 'without Threshold' case (mean=2).

Fig. 1 and Fig. 2 show the mean waiting time and loss probability as a function of effective arrival rate, respectively. These figures show that the mean waiting time and the loss probability generally increase as the effective arrival rate increases. Furthermore our model outperforms the ordinary queueing system without threshold values. Also Fig.1 shows that the mean waiting times converge to certain values : (the capacity of the buffer) \times (mean service time). Figures 3 and 4 also present the mean waiting time and the loss probability in which all conditions are identical except the mean of the service time distribution of 'without Threshold' case is 4.

In Fig. 1 and Fig. 3, it is observed that the mean waiting time increases in the beginning but decreases later as the effective arrival rate increases. It can be interpreted as follows. As the effective arrival rate increases, the mean queue length also increases. When the mean queue length is larger than a threshold value, customers are more likely to be served by a increased service rate. As a result, the mean waiting time decreases.

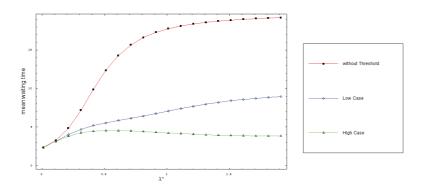


FIGURE 1. Mean waiting time over the λ^* (effective arrival rate)

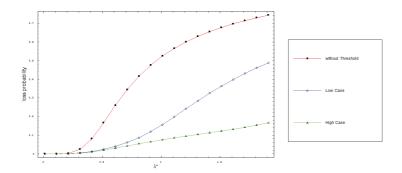


FIGURE 2. Loss probability over the λ^* (effective arrival rate)

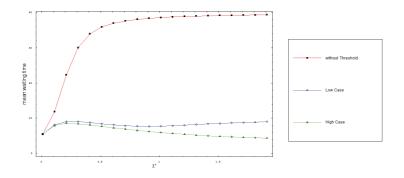


FIGURE 3. Mean waiting time over the λ^* (effective arrival rate)

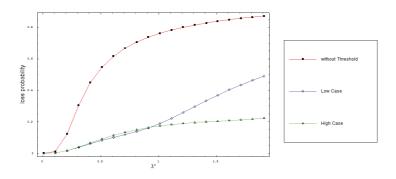


FIGURE 4. Loss probability over the λ^* (effective arrival rate)

5. CONCLUSION

In this paper, we analyzed an MMPP/G/1/K queueing system with queue length-dependent service rates. Several results including the queue length distributions, loss probability and mean queue length (mean waiting time) are presented. However, determining the optimal thresholds was not discussed. Determining an optimal threshold policy, such as the number of threshold values and specified threshold values, which minimize the long-run average cost with consideration of some reasonable cost factors can be suggested for future research. Holding cost per a customer, switching-over cost and operating cost for each service mode can be considered.

REFERENCES

- R. Bekker, G.M. Koole, B.F. Nielsen and T.B. Nielsen, *Queues with waiting time dependent service*, Queueing Systems, 68 (2011), 61–78.
- [2] H. Bruneel, D. Fiems, J. Walraevens and J. Wittevrongel, *Queueing models for the analysis of communication systems*, TOP, doi:10.1007/s11750-014-0330-3 (2014), 1–28.
- [3] D.I. Choi, T.-S. Kim and S. Lee, Analysis of an MMPP/G/1/K queue with queue length dependent arrival rates, and its application to preventive congestion control in telecommunication networks, European Journal of Operational Research, **187**.
- [4] D.I. Choi, C. Knessl and C. Tier, A queueing system with queue length dependent service times, with applications to cell discarding in ATM networks, Journal of Applied Mathematics and Stochastic Analysis, 12 (1999), 35–62.
- [5] B.D. Choi, Y.C. Kim, Y.W. Shin and C.E.M. Pearce, *The* $M^X/G/1$ *queue with queue length dependent service times*, Journal of Applied Mathematics and Stochastic Analysis, **14** (2001), 399–419
- [6] A. Dudin, Optimal Control for An $M^X/G/1$ queue with two operation modes, Probability in the Engineering and Informational Sciences, **11** (1997), 255–265.
- [7] S. Nishimura and Y. Jiang, An M/G/1 vacation model with two service modes, Probability in the Engineering and Informational Sciences, 9 (1995), 355–374.
- [8] R.D. Nobel and H.C. Tijms, Optimal control for an $M^X/G/1$ queue with two service modes, European Journal of Operational Research, **113** (1999), 610–619.
- [9] K.Y. Zhernovyi and Y.U. Zhernovyi, An $M^{\theta}/G/1/m$ System with Two-Threshold Hysteresis Strategy of Service Intensity Switching, Journal of Communications Technology and Electronics, **57** (2012), 1340–1349.
- [10] K.Y. Zhernovyi and Y.U. Zhernovyi, An $M^{\theta}/G/1$ System with Hysteretic Switching of the Service Intensity, Journal of Communications Technology and Electronics, **58** (2013), 602–612.