ICIC Express Letters
Volume **18**, Number **4**, April **2024**

 ISSN 1881-803X
pp. **385–391**

AN ANALYTICAL ANALYSIS OF MARKOVIAN QUEUEING MODEL WITH AN ALTERNATING SERVER AND STATE-DEPENDENT ALTERNATING PRIORITY POLICY

Doo Il Choi[1] and Dae-Eun Lim[2,*]

[1]Department of Applied Mathematics
Halla University
28 Halla University-gil, Wonju-si, Gangwon-do 26404, Korea
dichoi@halla.ac.kr

[2]Department of Industrial Engineering
Kangwon National University
1 Kangwondaehak-gil, Chuncheon-si, Gangwon-do 24341, Korea
*Corresponding author: del@kangwon.ac.kr

Received July 2023; accepted September 2023

Abstract. *In the dynamic landscape of modern technological systems, efficient queuing models play a pivotal role, especially in sectors such as telecommunications and data management. This study introduces a novel alternating server model with a state-dependent alternating priority policy. Specifically, we consider a Markovian queueing model with two queues, where the priority of service is determined by the length of each queue. A threshold is assigned to one of the queues, granting it priority when its length exceeds the threshold, while the other queue holds priority otherwise. Moreover, we derive a set of balance equations and present the joint queue length distribution in the form of a probability generating function, significantly advancing our understanding of queuing system dynamics. Delving into practical applications, our model proves particularly relevant in data centers, communication systems, and logistics networks. By precisely calculating performance measures such as mean queue length and waiting time, our study provides actionable insights for system optimization, directly influencing operational efficiency and user experience.*

Keywords: Alternating server, Polling system, Congestion control, Markovian queueing model, Priority, Threshold policy, State-dependent service, Stochastic model, Queue length, Steady-state distribution

1. **Introduction.** Polling systems employ a single server to process input streams from multiple sources based on a predefined set of rules. These systems have various real-world applications, including data centers [1], communication systems [2], and logistics systems [3]. In this context, we focus on an alternating server with two queues, drawing parallels to vehicles passing through an intersection [4] (see Figure 1), governed by right-of-way priority administered by traffic signals or other traffic control devices. If an intersection is regarded as a server, the time for a car to completely exit the intersection represents the service time.

Polling systems, studied extensively over time, include analyses of various service policies [5]. Takács [6] examined an $M/G/1$ model with two queues and an exhaustive policy in which a server continues its service until the number of customers in the attending queue reduces to zero; subsequently, the server moves to another queue. When a queue is being served, it can be considered to have priority over other queues. Eisenberg [7] considered a similar system but assumed the number of queues to be M (≥ 2). Sykes [8] and Boxma and Groenendijk [9] additionally considered switching time in Takács' analysis [6],

DOI: 10.24507/icicel.18.04.385

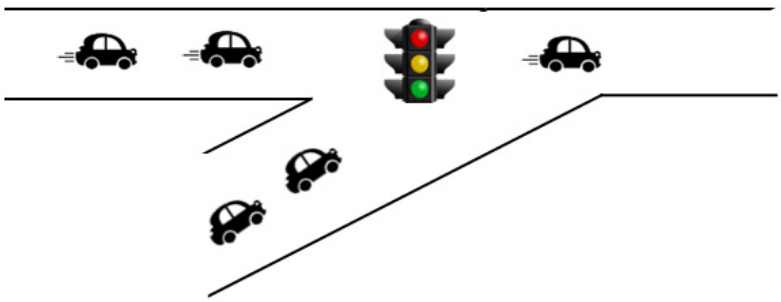

FIGURE 1. Two input streams at the intersection

which represents the time taken to move from one queue to another, assuming a general distribution.

In another study, Eisenberg [10] considered an $M/G/1$ model with two queues and an alternating service discipline, where a server serves only one customer in the current queue before moving to another. If a queue is empty, the server continues its service until a new customer arrives. Boon and Winands [11] analyzed a K-limited polling system with two queues in which the server switches to another queue after serving K customers. When there are fewer than K customers, the server continues serving until the queue size reaches zero. Ozawa [12] extended the K-limited polling system by considering mixed disciplines. Winands et al. [13] incorporated the setup time into Ozawa's model [12].

Recently, Avrachenkov et al. [14] and Perel and Yechiali [15] introduced new switching policies based on the queue length for a finite $M/M/1$ queuing model. The server then switches the size of the other unserved queue implying that each queue has its own threshold value. We investigate a different model from those of Avrachenkov et al. [14] and Perel and Yechiali [15]; the queue sizes are assumed to be infinite and only one queue has a threshold, and the allocation of priority to a particular queue is determined based on whether the queue size exceeds the threshold.

Polling systems have been applied to evaluating the performance of various systems, including transportation, road management, and production systems. In particular, polling systems have been widely used in the field of telecommunication systems [16,17]. Various studies continue to emerge on the development and evaluation of efficient operational strategies for tele- or data-communication systems using the latest technology [18-20].

Our study makes significant contributions by providing an analytical solution for a simple yet significant $M/M/1$ model. Notably, polling systems are known to be challenging for obtaining exact solutions. Our study differs from previous studies in that we provide an exact solution that has not been well studied. Previous studies have focused on modeling complicated systems using approximate solutions rather than exact solutions to real-world problems. The remainder of this paper is organized as follows. Section 2 outlines the necessary notations, and Section 3 presents the analytical solution. Finally, Section 4 presents the conclusion of our study.

2. **Problem Statement and Preliminaries.** In this section, we introduce the assumptions and define the mathematical notations for the system analysis. As depicted in Figure 2, our model involves two separate queues to accommodate Type-1 and Type-2 customers. Each queue has an infinite capacity. The arrival of Type-1 and Type-2 customers follows independent Poisson processes with rates λ_1 and λ_2, respectively. Type-1 (or Type-2) customers are queued in Queue I (or Queue II). The customers in each queue are served on a first-come first-service basis, and the service priority of each queue is determined by the length of Queue I. Specifically, we set a threshold L for Queue I. The server is initially idle and starts serving an arriving customer regardless of the type. If the length of Queue I is less than threshold L, then service priority is given to Queue II. However, if the length

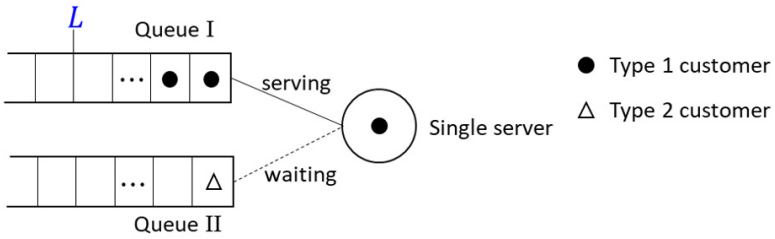

FIGURE 2. Example of an alternating server with two queues

of Queue I reaches the threshold L, service priority is given to Queue I until Queue I becomes empty. As soon as Queue I becomes empty, service priority is assigned to Queue II. This process is repeated. If one of the buffers is empty, customers from the order buffer are served, regardless of their type. The service time of all customers regardless of type is independent and identically distributed by the exponential distribution with rate μ. The switching time for the server to change the queue is assumed to be zero.

The mathematical notations are defined as follows. Let $N_1(t)$ and $N_2(t)$ be the lengths of Queues I and II at time t, respectively. Furthermore, we introduce the following notation:

$$\xi(t) = \begin{cases} 0, & \text{if the server is idle at time } t, \\ 1, & \text{if the server is busy at time } t. \end{cases}$$

$$R(t) = \begin{cases} 1, & \text{if the service priority is given to Queue I at time } t, \\ 2, & \text{if the service priority is given to Queue II at time } t. \end{cases}$$

To derive the joint queue length distribution of Queues I and II, we begin by defining the following:

$$\begin{aligned} &p(m,n) = \lim_{t\to\infty} \Pr\{N_1(t) = m, N_2(t) = n, \xi(t) = 1, R(t) = 1\}, \\ &q(m,n) = \lim_{t\to\infty} \Pr\{N_1(t) = m, N_2(t) = n, \xi(t) = 1, R(t) = 2\}, \\ &m \geq 0,\ n \geq 0, \\ &p_0 = \lim_{t\to\infty} \Pr\{\xi(t) = 0\}. \end{aligned}$$

$(\lambda_1 + \lambda_2) < \mu$ is assumed for the stability of the system. Considering that a stable system has limiting probabilities, the following balance equations can be established:

$$\begin{aligned} &(\mu + \lambda_1 + \lambda_2)p(m,n) \\ &= \lambda_1 p(m-1,n) + \lambda_2 p(m,n-1) + \mu p(m+1,n), \quad m \geq L+1,\ n \geq 0. \qquad (1) \\ &(\mu + \lambda_1 + \lambda_2)p(L,n) \\ &= \lambda_1 p(L-1,n) + \lambda_2 p(L,n-1) + \mu p(L+1,n) + \lambda_1 q(L-1,n), \quad m = L,\ n \geq 0. \qquad (2) \\ &(\mu + \lambda_1 + \lambda_2)q(m,n) \\ &= \lambda_1 q(m-1,n) + \lambda_2 q(m,n-1) + \mu q(m,n+1), \quad 1 \leq m < L,\ n \geq 1. \qquad (3) \\ &(\mu + \lambda_1 + \lambda_2)q(m,0) \\ &= \lambda_1 q(m-1,0) + \mu q(m,1) + \mu q(m+1,0)1_{\{m\neq L-1\}}, \quad 1 \leq m < L,\ n = 0. \qquad (4) \\ &(\mu + \lambda_1 + \lambda_2)q(0,n) \\ &= \lambda_2 q(0,n-1) + \mu q(0,n+1) + \mu p(1,n), \quad m = 0,\ n \geq 1. \qquad (5) \\ &(\mu + \lambda_1 + \lambda_2)q(0,0) \\ &= (\lambda_1 + \lambda_2)p_0 + \mu q(0,1) + \mu q(1,0) + \mu p(1,0), \quad m = 0,\ n = 0. \qquad (6) \\ &(\lambda_1 + \lambda_2)p_0 = \mu q(0,0). \qquad (7) \end{aligned}$$

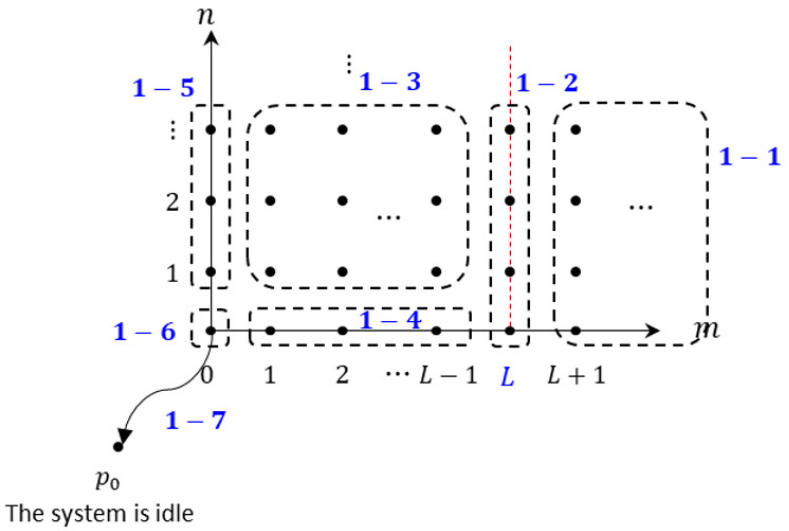

FIGURE 3. State space and its corresponding balance equations

The state space and corresponding balance equations for each state are illustrated in Figure 3. The indicator function $1_{\{condition\}}$ is equal to 1 if 'condition' is true, or 0 if it is false. The normalization condition yields

$$\sum_{m=1}^{\infty}\sum_{n=0}^{\infty} p(m,n) + \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} q(m,n) + p_0 = 1, \tag{8}$$

where $p(0,n) = 0$, $p(m,-1) = 0$ and $q(m,-1) = 0$.

3. **Main Results.** In this section, the set of balance equations in Section 2 is solved using probability-generating functions. We define the following probability-generating function for the queue length:

$$G_n(z) = \sum_{m=1}^{\infty} p(m,n) z^m, \quad |z| \le 1,$$

$$G(z,w) = \sum_{m=1}^{\infty} G_n(z) w^n,$$

$$H(w) = \sum_{m=1}^{\infty} p(1,n) w^n,$$

$$F_m(w) = \sum_{m=1}^{\infty} q(m,n) w^n, \quad |w| \le 1,\ 0 \le m < L.$$

From Equations (1) and (2), we obtain

$$(\mu + \lambda_1 + \lambda_2) G_n(z) = \lambda_1 z G(z) + \lambda_2 G_{n-1}(z) + \frac{\mu}{z} G(z) - \mu p(1,n) + \lambda_1 z^L q(L-1,n). \tag{9}$$

By multiplying w^n and summing over n in Equation (9), we obtain

$$(\mu + \lambda_1 + \lambda_2) G(z,w) = \lambda_1 z G(z,w) + \lambda_2 w G(z,w) + \frac{\mu}{z} G(z,w) - \mu H(w) + \lambda_1 z^L F_{L-1}(w).$$

That is,

$$\left\{\lambda_1 z - (\mu + \lambda_1 + \lambda_2) + \lambda_2 w + \frac{\mu}{z}\right\} G(z,w) = \mu H(w) - \lambda_1 z^L F_{L-1}(w).$$

The aforementioned equation can be rewritten into the following equation:

$$\frac{1}{z}\{z - z_*(w)\}\{z - \tilde{z}(w)\} G(z,w) = \mu H(w) - \lambda_1 z^L F_{L-1}(w), \tag{10}$$

where

$$z_*(w) = \frac{1}{2\lambda_1}\left\{\mu+\lambda_1+\lambda_2-\lambda_2 w-\sqrt{(\mu+\lambda_1+\lambda_2-\lambda_2 w)^2-4\lambda_1\mu}\right\},$$
$$\tilde{z}(w) = \frac{1}{2\lambda_1}\left\{\mu+\lambda_1+\lambda_2-\lambda_2 w+\sqrt{(\mu+\lambda_1+\lambda_2-\lambda_2 w)^2-4\lambda_1\mu}\right\}.$$

In Equation (10), putting $z = z_*(w)$, we obtain

$$H(w) = \frac{\lambda_1}{\mu} z_*(w)^L F_{L-1}(w). \tag{11}$$

By substituting $H(w)$ into Equation (10), we obtain

$$G(z,w) = \frac{\lambda_1 z}{\tilde{z}(w)-z} F_{L-1}(w) \sum_{k=0}^{n-1} z_*(w)^{n-1-k} z^k. \tag{12}$$

From Equations (3) and (4), we obtain

$$\begin{aligned}(\mu+\lambda_1+\lambda_2)F_m(w) = {} & \lambda_1 F_{m-1}(w)+\lambda_2 w F_m(w)+\frac{\mu}{w}F_m(w)-\frac{\mu}{w}q(m,0)\\ & +\mu q(m+1,0)1_{\{m\neq L-1\}}, \quad 1\le m<L.\end{aligned}$$

That is,

$$\begin{aligned}&\frac{1}{w}\left\{\lambda_2 w^2-(\mu+\lambda_1+\lambda_2)w+\mu\right\}F_m(w)\\ &= -\lambda_1 F_{m-1}(w)+\frac{\mu}{w}q(m,0)-\mu q(m+1,0)1_{\{m\neq L-1\}}.\end{aligned}$$

We can rewrite the aforementioned equation into the following equation:

$$\frac{1}{w}(w-w_*)(w-\tilde{w})F_m(w) = -\lambda_1 F_{m-1}(w)+\frac{\mu}{w}q(m,0)-\mu q(m+1,0)1_{\{m\neq L-1\}}, \tag{13}$$

where $w_* = \frac{1}{2\lambda_2}\left\{\mu+\lambda_1+\lambda_2-\sqrt{(\mu+\lambda_1+\lambda_2)^2-4\lambda_2\mu}\right\}$ and $\tilde{w} = \frac{1}{2\lambda_2}\Big\{\mu+\lambda_1+\lambda_2 +\sqrt{(\mu+\lambda_1+\lambda_2)^2-4\lambda_2\mu}\Big\}$.

In addition, using Equations (5) and (6), we obtain the following equation using Equation (7)

$$\begin{aligned}&\frac{1}{w}\left\{\lambda_2 w^2-(\mu+\lambda_1+\lambda_2)w+\mu\right\}F_0(w)\\ &= \frac{\mu}{w}q(0,0)-\mu H(w)-\mu q(1,0)-(\lambda_1+\lambda_2)p_0\\ &= (\lambda_1+\lambda_2)p_0\left(\frac{1}{w}-1\right)-\mu H(w)-\mu q(1,0), \qquad (14)\\ &\frac{1}{w}(w-w_*)(w-\tilde{w})F_0(w)\\ &= (\lambda_1+\lambda_2)p_0\left(\frac{1}{w}-1\right)-\mu H(w)-\mu q(1,0)\\ &= (\lambda_1+\lambda_2)p_0\left(\frac{1}{w}-1\right)-\lambda_1 z_*(w)^L F_{L-1}(w)-\mu q(1,0), \qquad (15)\end{aligned}$$

$$\frac{1}{w}(w-w_*)(w-\tilde{w})F_1(w) = -\lambda_1 F_0(w)+\frac{\mu}{w}q(1,0)-\mu q(2,0), \tag{16}$$

$$\frac{1}{w}(w-w_*)(w-\tilde{w})F_2(w) = -\lambda_1 F_1(w)+\frac{\mu}{w}q(2,0)-\mu q(3,0), \tag{17}$$

$$\vdots$$

$$\frac{1}{w}(w-w_*)(w-\tilde{w})F_{L-2}(w) = -\lambda_1 F_{L-3}(w) + \frac{\mu}{w}q(L-2,0) - \mu q(L-1,0), \quad (18)$$

$$\frac{1}{w}(w-w_*)(w-\tilde{w})F_{L-1}(w) = -\lambda_1 F_{L-2}(w) + \frac{\mu}{w}q(L-1,0). \quad (19)$$

By substituting $F_0(w)$ from Equation (15) into Equation (16), $F_1(w)$ can be obtained by $F_{L-1}(w)$. Based on Equation (19) derived from Equation (17), we obtain $F_{L-1}(w)$. Finally, for all $0 \le m \le L-1$, $F_m(w)$ is obtained by p_0, $q(1,0), q(2,0), \ldots, q(L-1,0)$. However, assuming $w = w_*$ in the equations from (15) to (19), $q(m,0)$ is obtained by p_0 and $F_m(w_*)$, $1 \le m \le L-1$, and p_0 is obtained by the normalization condition (8). Therefore, we obtain $F_m(w)$ for all m, $0 \le m \le L-1$. Finally, the joint probability generating function $G(z,w)$ is obtained by substituting $F_{L-1}(w)$ into Equation (12).

4. **Conclusions.** Using the joint probability-generating function $G(z,w)$, we can obtain various performance measures, including the mean queue length and waiting time. This yields an exact solution to an $M/M/1$ polling model with two queues and a state-dependent alternating-priority policy.

As a follow-up to this study, further research could involve:

- Analyzing a more general model by extending the service time from an exponential distribution to a general distribution. This would involve analyzing an $M/G/1$ queuing model with a more realistic service distribution.
- Considering the relevant cost factors associated with system operation to determine an optimal threshold level.

Acknowledgment. This work was supported by a research grant from 2023 Halla University. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1I1A3A04037238). The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] A. Jolles, E. Perel and U. Yechiali, Alternating server with non-zero switch-over times and opposite-queue threshold-based switching policy, *Performance Evaluation*, vol.126, pp.22-38, 2018.

[2] S. C. Borst, O. J. Boxma and H. Levy, The use of service limits for efficient operation of multistation single-medium communication systems, *IEEE/ACM Transactions on Networking*, vol.3, no.5, pp.602-612, 1995.

[3] E. M. M., Winands, I. J. B. F. Adan, G. J. Van Houtum and D. G. Down, A state-dependent polling model with k-limited service, *Probability in the Engineering and Informational Sciences*, vol.23, no.2, pp.385-408, 2009.

[4] M. A. Boon, I. J. Adan, E. M. Winands and D. G. Down, Delays at signalized intersections with exhaustive traffic control, *Probability in the Engineering and Informational Sciences*, vol.26, no.3, pp.337-373, 2012.

[5] A. Shausan and A. Vuorinen, Thirty-six years of contributions to queueing systems: A content analysis, topic modeling, and graph-based exploration of research published in the QUESTA Journal, *Queueing Systems*, vol.104, pp.3-18, 2023.

[6] L. Takács, Two queues attended by a single server, *Operations Research*, vol.16, no.3, pp.639-650, 1968.

[7] M. Eisenberg, Queues with periodic service and changeover time, *Operations Research*, vol.20, no.2, pp.440-451, 1972.

[8] J. S. Sykes, Simplified analysis of an alternating-priority queuing model with setup times, *Operations Research*, vol.18, no.6, pp.1182-1192, 1970.

[9] O. J. Boxma and W. P. Groenendijk, Two queues with alternating service and switching times, *Centrum voor Wiskunde en Informatica*, 1987.

[10] M. Eisenberg, Two queues with alternating service, *SIAM Journal on Applied Mathematics*, vol.36, no.2, pp.287-303, 1979.

[11] M. A. A. Boon and E. M. M. Winands, Heavy-traffic analysis of K-limited polling systems, *Probability in the Engineering and Informational Sciences*, vol.28, no.4, pp.451-454, 2014.
[12] T. Ozawa, Alternating service queues with mixed exhaustive and K-limited services, *Performance Evaluation*, vol.11, no.3, pp.165-175, 1990.
[13] E. M. M. Winands, I. J. B. F. Adan and G. J. van Houtum, A two-queue model with alternating limited service and state-dependent setups, *Proc. of Analysis of Manufacturing Systems-Production Management*, Zakynthos, pp.200-208, 2005.
[14] K. Avrachenkov, E. Perel and U. Yechiali, Finite-buffer polling systems with threshold-based switching policy, *TOP*, vol.24, no.3, pp.541-571, DOI: 10.1007/s11750-015-0408-6, 2016.
[15] E. Perel and U. Yechiali, Two-queue polling systems with switching policy based on the queue that is not being served, *Stochastic Models*, vol.33, no.3, pp.430-450, 2017.
[16] M. A. Boon, R. D. van der Mei and E. M. Winands, Applications of polling systems, *Surveys in Operations Research and Management Science*, vol.16, no.2, pp.67-82, 2011.
[17] V. Vishnevsky and O. Semenova, Polling systems and their application to telecommunication networks, *Mathematics*, vol.9, no.2, 117, 2021.
[18] H. Guo, W. Li, J. Lin, J. Zhou, Q. Xun, S. Zhan, Y. Huang, Y. Wang and Q. Cheng, A scalable asynchronous traffic shaping mechanism for TSN with time slot and polling, *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, Miami, FL, USA, pp.1-5, 2023.
[19] X. Wang, Z. Yang and H. Ding, Application of polling scheduling in mobile edge computing, *Axioms*, vol.17, no.7, 709, 2023.
[20] Z. Arefian, M. R. Khayyambashi and N. Movahhedinia, Delay reduction in MTC using SDN based offloading in fog computing, *PLoS ONE*, vol.18, no.5, e0286483, 2023.