

Performance Analysis for Threshold-based N-Systems with Heterogeneous Servers

Le Anh Thu¹ ^a and Tuan Phung-Duc² ^b

¹Public Policy Program, VNU Vietnam Japan University, My Dinh Campus, Nam Tu Liem, Hanoi, Vietnam

²Department of Policy and Planning Sciences, Faculty of Engineering, Information and Systems, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

Keywords: Multi-skilled Servers, Threshold Policy, Matrix Analytic Method, Administrative Services.

Abstract: Driven by the need to develop methods for minimizing operational delays at public administration agencies, this paper considers problems involving routing and staffing in these agencies. We examine a threshold-based N-System of two queues with capacities $C_1 = \infty$ and $C_2 < \infty$, respectively. We use the matrix analytic method to obtain the steady-state probabilities, the performance measures, and the optimal threshold values in terms of the system parameters. Our numerical experiments reveal that the mean response time is sensitive to the stability condition, and the effectiveness of the threshold policy depends on the customer arrival rate.

1 INTRODUCTION


Almost everyone has waited for days or weeks to get an identity card, a driver's license, a visitor visa, a business registration certificate, etc. Waiting lines or queues are known as common phenomena in administrative services due to the inadequate resources in public administrations and rising demand for these services, typically in Immigration Department, Business Registration Office, etc. Queues exist mainly due to the limited resources of the system. Customer arrivals cannot be scheduled or controlled since the customers usually arrive randomly. Moreover, customer service times are independent random variables; some individuals take a short time, while others require a long period. It can be seen that queueing phenomena lead to three common problems: (i) Customer satisfaction declines due to the discomfort of spending hours in a crowded waiting room to access the services; (ii) The employees endure the overloaded work stress, which reduces the efficiency and quality of work; (iii) Worsening relationships between customers and staff, and leading to more disputes.


Waiting time has been identified as the critical factor influencing customer satisfaction, and consequently, decreasing delays has become a focus in optimizing the efficiency of public services (Osborne

et al., 2013). Allocating human resources based on staff capacity is an effective solution for optimizing staff performance and, as a result, reducing waiting time. The complexity of work in administrative services varies, and so do the qualifications of the staff. Experienced employees are advantageous in terms of performance in highly complex jobs; however, their performance tends to decline quickly in non-complex tasks when they become bored. Meanwhile, inexperienced employees are under pressure as their skills are not well-matched to their job duties (Hunter and Thatcher, 2007).

Motivated by these situations, we consider an N-design multi-server queueing system that serves two types of customers in two queues. This system employs two groups of servers: employees trained to handle low complexity tasks and more experienced employees who can handle all tasks. The switching policy of the model is based on the skills of two different types of employees and the thresholds of two queues. Similar configurations are found in various settings, including international call centers (single-language and multilingual servers), emergency medical departments (life-threatening injuries and others), etc. The problem of optimal allocation of customers between queues in queueing systems to minimize waiting time has received much attention. As for the routing and staffing issues, we refer to the survey paper by Gans et al. (2003).

The model used in this paper is related to the

^a  <https://orcid.org/0000-0002-1474-134X>

^b  <https://orcid.org/0000-0002-5002-4946>

stochastic service systems belonging to a class of models named "lane section" that was first proposed by Schwartz (1974). Stanford and Grassmann (1993) considered a similar model of N-design bilingual server system with both specialized and flexible servers. They presented an exact performance analysis to determine the minimum number of bilingual servers required. Due to the computational complexity of this method, only comparatively small systems can be solved. A closely related model is the paper by Li and Yue (2016). The authors examined the N-design call center with two types of users in which primary users have non-preemptive priority. The state-space division method has been employed to divide an infinite number of system states into several finite states and obtain the steady-state probability equation of the system. Shumsky (2004) presented an approximate analysis of a queueing model of the multi-skill call center in N-design, which has a fixed priority strategy. The approximate analysis provides reasonable accuracy while reducing the computational burden of large service centers.

A matrix analytic method has been successfully applied to the entire state space to obtain exact performance measures. Morozov et al. (2021) examined a modified Erlang loss system with two classes of customers, in which the primary users take precedence over secondary users. The authors assumed that the probability distribution of service time is the exponential distribution, then studied the model in depth by matrix analysis method to evaluate the influence of the input parameters on the secondary user performance. Perel and Yechiali (2017) studied a closely related system consisting of two non-identical M/M/1 queues controlled by a threshold-based switching policy. Jolles et al. (2018) expanded the model of Perel and Yechiali by adding a switchover time policy. In order to find the mean number of customers in each queue, the authors formulated the system as a Quasi-Birth-and-Death (QBD) process. Similar to the method in Latouche and Ramaswami (1999), they studied the steady-state behavior of the system and obtained the rate matrix by applying the matrix analytic method.

In this paper, we consider the staffing problem of the N-design model using the matrix-analytic method. Though queueing analysis has been used in public services, this is the first analytical result for public administration service models with multiple servers in N-design. The matrix-analytic method allows us to derive the stability condition and effects of the input parameters on the mean response time and users' performance. The advantage of this method is to provide systematically specific calculation formulas to

analyze more complicated models while not requiring complex data. Therefore, our model is suitable for providing evidence to evaluate administrative service performance and compare alternatives quickly.

The rest of the paper is structured as follows. In Section 2, we describe our model with a focus on the switching policy, while in Section 3, the matrix analytic method is applied to derive performance measures of the system in steady-state. In Section 4, we present the results of numerical experiments to show insights into the performance of our system and various phenomena that occur due to a result of changes in parameters. Section 5 concludes the paper.

2 MODEL DESCRIPTION

We consider an N-design system that serves two classes of customers with c_1 and c_2 servers, respectively. The arrival processes are assumed to be the Poisson processes with arrival rates λ_1 and λ_2 , respectively, and the servers have exponential distributions with mean $1/\mu_i$ for class- i customers, $i = 1, 2$. The server's switching policy is threshold-based.

We assume that number of customers in Queue 1 (Q_1) is not limited while Queue 2 (Q_2) can accommodate up to $N_{max} < \infty$ customers including the ones in service. Let C_i denote the capacity (the maximum number of customers accommodated in the system) of $Q_i, i = 1, 2$. We then have $C_1 = \infty$, and $C_2 = N_{max} < \infty$. If the two capacities C_1 and C_2 are infinite, a completely different approach is required to solve this problem. For Q_1 , the threshold level is $K \geq c_1$, while for Q_2 , it is $N, c_2 \leq N \leq N_{max}$. Denote by $Q_i(t)$ the number of customers in $Q_i, i = 1, 2$ at time t .

If a customer arrives at Q_2 and sees this queue already has N customers, this customer will transfer to Q_1 . At that time, if Q_1 already has K customers, Q_1 will not accept customers from Q_2 , and that customer will return to Q_2 . The system is illustrated in Figure 1.

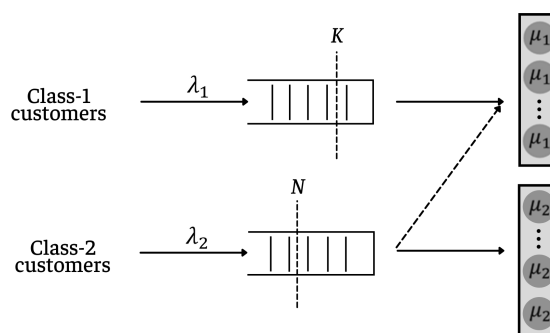


Figure 1: The N-design multi-server queueing system.

3 THE QBD PROCESS

In this section, we calculate the stationary distribution of the Markov process to obtain the corresponding stationary performance measures. The two-dimensional process $\{(Q_1(t), Q_2(t)), t \geq 0\}$ is a

continuous-time Markov Chain with the state space \mathbb{S} given by $\mathbb{S} = \{(i, j) \in \mathbb{N} \times \{0, 1, \dots, N_{max}\}\}$. The system can be formulated as a Quasi-Birth-and-Death process (QBD) with the infinitesimal generator Q given as

$$Q = \begin{pmatrix} B_0 & C & 0 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ A_1 & B_1 & C & 0 & 0 & \dots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & A_2 & B_2 & C & 0 & \dots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & A_{c_1} & B_{c_1} & C & 0 & \dots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & 0 & A_{c_1} & B_{K-1} & C & 0 & \dots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & 0 & 0 & A_{c_1} & B_K & C_K & 0 & \dots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & \dots & 0 & 0 & A_{c_1} & B_K & C_K & 0 & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \end{pmatrix},$$

where 0 is a $(N_{max+1}) \times (N_{max+1})$ zero matrix, and A_i, B_i, B_K, C, C_K are $(N_{max+1}) \times (N_{max+1})$ block matrices given by

$$A_i = \begin{pmatrix} \min(i, c_1)\mu_1 & 0 & 0 & 0 \\ 0 & \min(i, c_1)\mu_1 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \min(i, c_1)\mu_1 \end{pmatrix},$$

for $i < c_1$, and $A_i = A_{c_1}$ for $i \geq c_1$.

$$B_i = \begin{pmatrix} b_{i,0} & \lambda_2 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ \mu_2 & b_{i,1} & \lambda_2 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 2\mu_2 & b_{i,2} & \lambda_2 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & c_2\mu_2 & b_{i,c_2} & \lambda_2 & 0 & \dots & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & \dots & 0 & c_2\mu_2 & b_{i,N-1} & \lambda_2 & 0 & \dots & 0 \\ 0 & \dots & \dots & \dots & 0 & 0 & c_2\mu_2 & b_{i,N} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \dots & \dots & \dots & \dots & \dots & \dots & 0 & c_2\mu_2 & b_{i,N_{max}-1} & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & c_2\mu_2 & b_{i,N_{max}} \end{pmatrix},$$

for $i = 0, 1, 2, \dots, K - 1$, where $b_{i,n} = -[\lambda_1 + \lambda_2 + \min(i, c_1)\mu_1 + \min(n, c_2)\mu_2]$.

$$B_K = \begin{pmatrix} b_{i,0} & \lambda_2 & 0 & 0 & \dots & \dots & \dots & 0 \\ \mu_2 & b_{i,1} & \lambda_2 & 0 & \dots & \dots & \dots & 0 \\ 0 & 2\mu_2 & b_{i,2} & \lambda_2 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & c_2\mu_2 & b_{i,c_2} & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & 0 & c_2\mu_2 & b_{i,N_{max}-1} & \lambda_2 \\ 0 & \dots & \dots & \dots & \dots & 0 & c_2\mu_2 & b_{i,N_{max}} \end{pmatrix},$$

for $i = K, K + 1, K + 2, \dots, B_i = B_K$, where

$$b_{i,n} = -[\lambda_1 + \lambda_2 + \min(i; c_1)\mu_1 + \min(n; c_2)\mu_2],$$

for $n = 0, 1, \dots, N_{max} - 1$, and

$$b_{i,N_{max}} = -[\lambda_1 + \min(i; c_1)\mu_1 + \min(n; c_2)\mu_2].$$

$$C = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & \lambda_1 & 0 & \dots & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \vdots & 0 & \lambda_1 & 0 & \dots & 0 \\ 0 & \dots & \dots & 0 & \lambda_1 + \lambda_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & \dots & 0 & \lambda_1 + \lambda_2 \end{pmatrix},$$

where the diagonal element $C(i, i)$ is given by

$$C(i, i) = \lambda_1, \text{ for } i = 0, 1, \dots, N - 1,$$

$$C(i, i) = \lambda_1 + \lambda_2, \text{ for } i = N, N + 1, \dots, N_{max}.$$

$$C_K = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \lambda_1 \end{pmatrix}.$$

Let $M = A_{c_1} + B_K + C_K$, then

$$M = \begin{pmatrix} m_0 & \lambda_2 & 0 & 0 & \dots & \dots & 0 \\ \mu_2 & m_1 & \lambda_2 & 0 & \dots & \dots & 0 \\ 0 & 2\mu_2 & m_2 & \lambda_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & c_2\mu_2 & m_{c_2} & \lambda_2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & 0 & c_2\mu_2 & m_{N_{max}} \end{pmatrix},$$

where the diagonal elements of M is given by

$$m_i = -(\min(i, c_2)\mu_2 + \lambda_2), \text{ for } i = 0, 1, \dots, N_{max} - 1,$$

$$\text{and } m_{N_{max}} = -c_2\mu_2.$$

Let $\pi_M = (\pi_{M,i}; i = 1, 2, \dots, N_{max})$ be the stationary probability vector of the matrix M , i.e., $\pi_M M = 0$ and $\pi_M e = 1$, where e denotes the column vector of ones, whose dimension is determined upon context.

The stability condition of such a QBD, (see Theorem 1.7.1, Neuts (1994)) can be obtained by the condition

$$\pi_M C_K e < \pi_M A_{c_1} e.$$

This stability condition can be transformed as

$$\lambda_1 < c_1\mu_1. \tag{1}$$

Let $\pi(i, n) = P(Q_1(t) = i, Q_2(t) = n)$, for $i \in N$ and $n = 0, 1, \dots, N_{max}$ denote the stationary probability of the Markov chain.

We define

$$\pi_i^{(1)} = (\pi_{i,0}, \pi_{i,1}, \dots, \pi_{i,N_{max}}), \text{ for } i = 0, 1, 2, \dots,$$

$$\pi_n^{(2)} = (\pi_{0,n}, \pi_{1,n}, \pi_{2,n}, \dots), \text{ for } n = 0, 1, 2, \dots, N_{max}.$$

According to Matrix-analytic-method (Latouche and Ramaswami (1999); Phung-Duc et al. (2010)), we have

$$\pi_i^{(1)} = \pi_K^{(1)} R^{i-K}, \quad i > K, \tag{2}$$

$$\pi_i^{(1)} = \pi_{i-1}^{(1)} R^{(i)}, \quad i = K, K - 1, \dots, 1, \tag{3}$$

where R is the minimal non-negative solution of

$$C_K + RB_K + R^2 A_{c_1} = 0, \tag{4}$$

and

$$R^{(i)} = -C(B_i + R^{(i+1)} A_{i+1})^{-1}, \text{ for } i = K - 1, K - 2, \dots, 1, \tag{5}$$

given that

$$R^{(K)} = -C(B_K + R A_{c_1})^{-1}. \tag{6}$$

Then, π_0 is the solution of the following equations

$$\begin{aligned} \pi_0^{(1)} (B_0 + R^{(1)} A_1) &= 0, \\ \pi_0^{(1)} \left(I + \sum_{i=1}^{K-1} \prod_{j=1}^i R^{(j)} + \left(\prod_{j=1}^K R^{(j)} \right) (I - R)^{-1} \right) e &= 1, \end{aligned} \tag{7}$$

where we use I to denote the $(N_{max} + 1) \times (N_{max} + 1)$ identity matrix. The first and the second equation in (7) represent the boundary equations at level 0, and the normalization condition, respectively.

Let $E[L_i]$ denote the average number of customers in the system in Q_i , $i = 1, 2$. We then obtain the mean queue length of Q_1 and Q_2 as follows

$$\begin{aligned} E[L_1] &= \sum_{i=1}^{\infty} \pi_i^{(1)} i e \\ &= \sum_{i=1}^{K-1} \pi_i^{(1)} i e + \sum_{i=K}^{\infty} \pi_i^{(1)} i e \\ &= \sum_{i=1}^{K-1} \pi_i^{(1)} i e + \sum_{i=K}^{\infty} \pi_K^{(1)} R^{i-K} i e \\ &= \sum_{i=1}^{K-1} \pi_i^{(1)} i e + \pi_K^{(1)} \sum_{i'=0}^{\infty} R^{i'} (i' + K) e \\ &= \sum_{i=1}^{K-1} \pi_i^{(1)} i e + \pi_K^{(1)} \sum_{i'=0}^{\infty} R^{i'} K e + \pi_K^{(1)} \sum_{i'=0}^{\infty} R^{i'} i' e \end{aligned}$$

$$\begin{aligned}
 &= \sum_{i=1}^{K-1} \pi_i^{(1)} ie + \pi_K^{(1)} (I-R)^{-1} Ke + \pi_K^{(1)} R \sum_{j=1}^{\infty} R^{j-1} je \\
 &= \sum_{i=1}^{K-1} \pi_i^{(1)} ie + \pi_K^{(1)} (I-R)^{-1} Ke + \pi_K^{(1)} R [(I-R)^{-1}]^2 e \\
 &= \sum_{i=1}^{K-1} \pi_i^{(1)} ie + \pi_K^{(1)} [(I-R)^{-1} K + R [(I-R)^{-1}]^2] e,
 \end{aligned} \tag{8}$$

$$\begin{aligned}
 E[L_2] &= \sum_{n=0}^{N_{\max}} \pi_n^{(2)} ne \\
 &= \left(\sum_{n=0}^{K-1} \pi_n^{(1)} + \pi_K^{(1)} (I+R+R^2+\dots) \right) f \tag{9} \\
 &= \left(\sum_{n=0}^{K-1} \pi_n^{(1)} + \pi_K^{(1)} (I-R)^{-1} \right) f,
 \end{aligned}$$

where $f = (0, 1, 2, \dots, N_{\max})^T$.

Let $E[L]$ be the total average number of customers in the system in both queues, then

$$E[L] = E[L_1] + E[L_2]. \tag{10}$$

We obtain the mean number of busy servers, $E[S_i]$, in Q_i , $i = 1, 2$ as follows

$$\begin{aligned}
 E[S_1] &= \sum_{i=1}^{c_1-1} \pi_i^{(1)} ie + c_1 \sum_{j=c_1}^{\infty} \pi_j^{(1)} e \\
 &= \sum_{i=1}^{c_1-1} \pi_i^{(1)} ie + c_1 \sum_{j=c_1}^{K-1} \pi_j^{(1)} e + c_1 \sum_{j'=K}^{\infty} \pi_{j'}^{(1)} e \\
 &= \sum_{i=1}^{c_1-1} \pi_i^{(1)} ie + c_1 \sum_{j=c_1}^{K-1} \pi_j^{(1)} e + c_1 \sum_{j'=K}^{\infty} \pi_K^{(1)} R^{j'-K} e \\
 &= \sum_{i=1}^{c_1-1} \pi_i^{(1)} ie + c_1 \sum_{j=c_1}^{K-1} \pi_j^{(1)} e + c_1 \pi_K^{(1)} (I-R)^{-1} e \\
 &= \sum_{i=1}^{c_1-1} \pi_i^{(1)} ie + \left(\sum_{j=c_1}^{K-1} \pi_j^{(1)} + \pi_K^{(1)} (I-R)^{-1} \right) c_1 e.
 \end{aligned} \tag{11}$$

$$\begin{aligned}
 E[S_2] &= \sum_{i=1}^{c_2-1} \pi_i^{(2)} ie + c_2 \sum_{j=c_2}^{N_{\max}} \pi_j^{(2)} e \\
 &= \left(\sum_{n=0}^{K-1} \pi_n^{(1)} + \pi_K^{(1)} (I-R)^{-1} \right) g,
 \end{aligned} \tag{12}$$

where g is a $(N_{\max} + 1) \times 1$ column vector given by

$$g = (0, 1, 2, \dots, c_2 - 1, c_2, \dots, c_2)^T.$$

Denote by $E[T_i]$ the throughput of Q_i , $i = 1, 2$, that are

$$E[T_1] = E[S_1] \times \mu_1$$

$$\begin{aligned}
 &= \left(\sum_{i=1}^{c_1-1} \pi_i^{(1)} ie + c_1 \sum_{j=c_1}^{\infty} \pi_j^{(1)} e \right) \mu_1 \\
 &< c_1 \sum_{i=1}^{\infty} \pi_i^{(1)} e \mu_1 = c_1 \mu_1,
 \end{aligned} \tag{13}$$

$$E[T_2] = E[S_2] \times \mu_2. \tag{14}$$

Then the throughput of the system is given by

$$E[T] = E[T_1] + E[T_2]. \tag{15}$$

Furthermore, due to Little's law, we obtain the mean response time $E[R_i]$ in Q_i , $i = 1, 2$, respectively, as follows

$$E[R_i] = \frac{E[L_i]}{E[T_i]}, \quad \text{for } i = 1, 2. \tag{16}$$

The mean system response time is given by

$$E[R] = \frac{E[L]}{E[T]}. \tag{17}$$

For reference, we compare with a baseline model, i.e., two parallel queues without the threshold policy. In the absence of threshold policy ($K = 0$), our system becomes a system of an $M/M/c_1$ queue and an $M/M/c_2/N_{\max}$ queue.

According to Medhi (2002), the probability of zero customers in the system in Q_1 is calculated by

$$\pi_0^{(1)} = \left(\sum_{n=0}^{c_1-1} \frac{(\lambda_1/\mu_1)^n}{n!} + \frac{(\lambda_1/\mu_1)^{c_1}}{c_1!(1-\lambda_1/(c_1\mu_1))} \right)^{-1}.$$

The condition for the stability of Q_1 is $\lambda_1/(c_1\mu_1) < 1$. The mean number of customers in the system in Q_1 is given by

$$E[L_1] = \frac{\lambda_1}{\mu_1} + \frac{\rho_1}{1-\rho_1} C \left(c_1, \frac{\lambda_1}{\mu_1} \right), \tag{18}$$

where $\rho_1 = \frac{\lambda_1}{c_1\mu_1}$, and $C \left(c_1, \frac{\lambda_1}{\mu_1} \right) = \frac{(\lambda_1/\mu_1)^{c_1}}{c_1!(1-\lambda_1/(c_1\mu_1))} \pi_0^{(1)}$ is referred to as Erlang's C formula.

Then, the mean response time in Q_1 can be obtained by

$$E[R_1] = \frac{E[L_1]}{\lambda_1}. \tag{19}$$

According to Shortle et al. (2018), the probability of zero customers in the system in Q_2 is given by

$$\pi_0^{(2)} = \left[\sum_{n=0}^{c_2-1} \frac{\rho_2^n}{n!} + \left(\frac{\rho_2^{c_2}}{c_2!} \right) (N_{\max} - c_2 + 1) \right]^{-1},$$

for $\frac{\rho_2}{c_2} = 1$, and

$$\pi_0^{(2)} = \left[\sum_{n=0}^{c_2-1} \frac{\rho_2^n}{n!} + \left(\frac{\rho_2^{c_2}}{c_2!} \right) \left(\frac{1 - \left(\frac{\rho_2}{c_2} \right)^{N_{max}-c_2+1}}{1 - \frac{\rho_2}{c_2}} \right) \right]^{-1},$$

for $\frac{\rho_2}{c_2} \neq 1$, where $\rho_2 = \frac{\lambda_2}{\mu_2}$.

Denote by $P_{N_{max}}$ the blocking probability of Q_2 which means that Q_2 can satisfy at most N_{max} flow units, then

$$P_{N_{max}} = \frac{\pi_0^{(2)} \rho_2^{N_{max}}}{c_2^{N_{max}-c_2} c_2!}.$$

The mean number of customers in the system in Q_2 is calculated as

$$E[L_2] = \frac{\pi_0^{(2)} \rho_2^{c_2} \left(\frac{\rho_2}{c_2} \right)}{c_2! \left(1 - \frac{\rho_2}{c_2} \right)^2} \left[1 - \left(\frac{\rho_2}{c_2} \right)^{N_{max}-c_2+1} - \left(1 - \frac{\rho_2}{c_2} \right) (N_{max} - c_2 + 1) \left(\frac{\rho_2}{c_2} \right)^{N_{max}-c_2} \right] + \rho_2 (1 - P_{N_{max}}). \tag{20}$$

The mean system response time in Q_2 can be obtained by

$$E[R_2] = \frac{E[L_2]}{\lambda_2 (1 - P_{N_{max}})}. \tag{21}$$

We obtain the mean system response time $E[R]$ in the case without threshold policy as follows

$$E[R] = \frac{E[L_1] + E[L_2]}{\lambda_1 + \lambda_2 (1 - P_{N_{max}})}. \tag{22}$$

4 NUMERICAL INSIGHTS

This section presents several numerical experiments of the results obtained in Section 3 to find insights into the performance of our system. For fixed $\lambda_1 = 20$, $\lambda_2 = 30$, $\mu_1 = 8$, $\mu_2 = 12$, $c_1 = 4$, $c_2 = 3$, and $N_{max} = 50$, we show how the performance measures change according to the thresholds (K, N) . Under the same settings, we also compute these performance measures in the case without threshold policy ($K = 0$) using the classical M/M/c and M/M/c/m models.

Figure 2 reflects the changes in the mean response time of class-1 customers against Q_2 's threshold. The mean response time of class-1 customers decreases when N goes up to a specific value for a fixed K , then remains unchanged as N continues to increase. Meanwhile, Figure 3 shows the exact opposite trend in the response time of class-2 customers. It is noticeable that the mean amount of time that class-2 customers spend in the system depends more on N than on K .

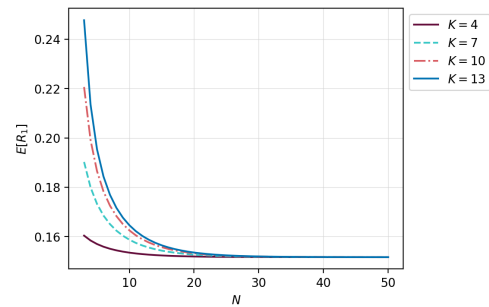


Figure 2: The mean response time of class-1 customers against the Q_2 's threshold.

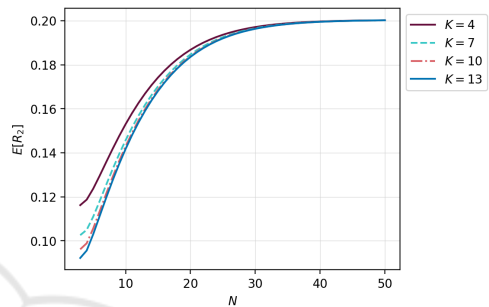


Figure 3: The mean response time of class-2 customers against the Q_2 's threshold.

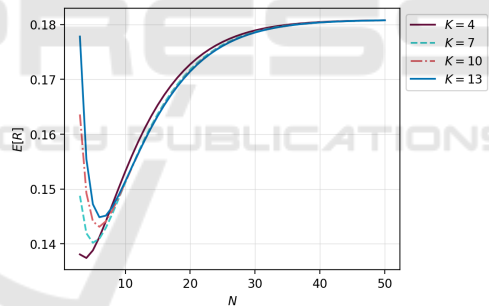


Figure 4: The mean system response time against the Q_2 's threshold.

Figure 4 indicates that the mean system response time $E[R]$ drops as N goes up to certain thresholds, then rises again sharply before remaining unchanged when N is at very high values. For small values of N , $E[R]$ goes down as N increases and K decreases. This occurs since the servers in Q_2 may remain idle even if there are waiting customers in Q_1 , including class-2 customers, leading to an increase in the mean system response time. In this experiment, $E[R]$ reaches the minimum at 0.1374 when the value of N is 4, and K is 4. Moreover, the mean response time of the system without threshold policy is 0.1808, which is higher than it is in the case of the optimal threshold policy.

For fixed $K = 4$ and $N = 3, 9$, we show the changes in the performance measure $E[R]$ against λ_1 and λ_2 ,

while all other parameters remain unchanged. We also illustrate the changes in these performance measures against λ_1 and λ_2 in the absence of threshold policy, thereby finding that a non-threshold policy is not optimal for our system. We recall that the stability condition in both cases with and without the threshold policy is given by $\lambda_1 < c_1\mu_1 = 32$.

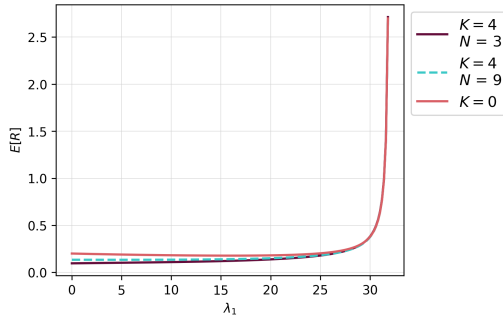


Figure 5: The mean system response time $E[R]$ against the arrival rate of class-1 customers ($\lambda_2 = 30$).

Figure 5 illustrates how the mean system response time $E[R]$ changes according to the arrival rate of class-1 customers. The mean system response time is large when class-1 customers arrive more frequently, especially as λ_1 is asymptotic to the value $c_1\mu_1$. The mean system response time is highly sensitive to these values of λ_1 , while changes in N and K at that time have no significant effect on $E[R]$. Therefore, as the arrival rate of class-1 customers approaches the stable threshold, increasing the number of servers in Q_1 is required to reduce the mean system response time.

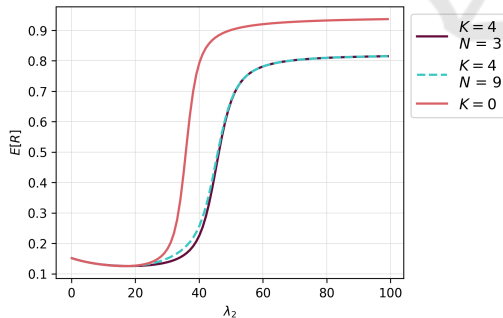


Figure 6: The mean system response time $E[R]$ against arrival rate of class-2 customers ($\lambda_1 = 20$).

Figure 6 shows the changes in the mean system response time $E[R]$ against λ_2 in both cases with and without the threshold policy. The performance measure $E[R]$ in these two cases shares the same trend when λ_2 changes. It can be seen that applying the threshold policy significantly reduces the mean system response time as the arrival rate of class-2 customers is large enough. If class-2 customers arrive more frequently, the mean system response time is

large, especially within a specific range of values of λ_2 . However, when λ_2 reaches a certain threshold, the mean system response time $E[R]$ will stop growing because the capacity of Q_2 is limited to N_{max} .

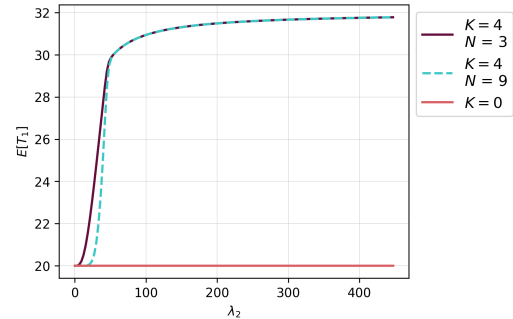


Figure 7: Throughput of Q_1 against arrival rate of class-2 customers ($\lambda_1 = 20$).

Figure 7 indicates the impact of the arrival rate of class-2 customers on the throughput of Q_1 . In the case of the threshold policy, the throughput $E[T_1]$ remains unchanged at λ_1 when λ_2 goes up to certain thresholds, then rises sharply as λ_2 continues to increase before remains stable at a value of $c_1\mu_1$ (see the condition (13)) when class-2 customers arrive at very high rates.

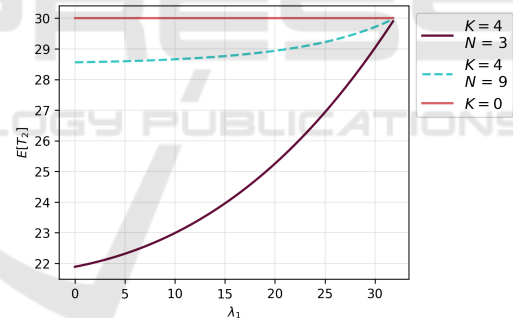


Figure 8: Throughput of Q_2 against arrival rate of class-1 customers ($\lambda_2 = 30$).

Figure 8 reflects the changes in the throughput of Q_2 against λ_1 when λ_2 is fixed. Under the threshold policy, the throughput of Q_2 closely approaches the value of λ_2 when λ_1 is asymptotic to the value of $c_1\mu_1$. For large values of N , the throughput of Q_2 is insensitive to the arrival rate of class-1 customers. Obviously, with the threshold policy, the throughput of Q_1 is greater than or equal to λ_1 , whereas the throughput of Q_2 is less than or equal to λ_2 because class-2 customers can transfer from Q_2 to Q_1 . In the absence of threshold policy, the throughputs $E[T_1]$ and $E[T_2]$ equal the arrival rates of class-1 and class-2 customers, respectively.

5 CONCLUSIONS

This paper has considered the routing and staffing problems of an administrative agency in an N-design model that serves two types of customers. Using the matrix analytic method, we have derived the steady-state probabilities and the performance measures. We then have determined the optimal threshold values according to the system parameters. We have found that the threshold policy is highly effective when the arrival rate of class-1 customers is low and the arrival rate of class-2 customers is high. When λ_1 approaches the critical value satisfying the stability condition or λ_2 is relatively small, increasing the number of servers combined with changing the threshold policy is the solution to reduce the mean system response time. As a result, we have provided a basis for reallocating resources when the customer arrival rate changes. Our findings could be used in decision-making, managing resources in administrative services, and related applications. It will be helpful to expand the analysis of our model to the case when both capacities are infinite.

ACKNOWLEDGEMENTS

The research of Tuan Phung-Duc was supported in part by JSPS KAKENHI Grant Numbers 18K18006, 21K11765.

REFERENCES

- Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141.
- Hunter, L. W. and Thatcher, S. M. (2007). Feeling the heat: Effects of stress, commitment, and job experience on job performance. *Academy of Management Journal*, 50(4):953–968.
- Jolles, A., Perel, E., and Yechiali, U. (2018). Alternating server with non-zero switch-over times and opposite-queue threshold-based switching policy. *Performance Evaluation*, 126:22–38.
- Latouche, G. and Ramaswami, V. (1999). *Introduction to matrix analytic methods in stochastic modeling*. SIAM.
- Li, C.-Y. and Yue, D.-Q. (2016). The staffing problem of the n-design multi-skill call center based on queuing model. In *Advances in computer science research, 3rd International Conference on Wireless Communication and Sensor Network*, volume 44, pages 427–432.
- Medhi, J. (2002). *Stochastic models in queueing theory*. Elsevier.

- Morozov, E. V., Rogozin, S., Nguyen, H., and Phung-Duc, T. (2021). Modified erlang loss system for cognitive wireless networks. *arXiv preprint arXiv:2103.03222*.
- Neuts, M. F. (1994). *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Courier Corporation.
- Osborne, S. P., Radnor, Z., and Nasi, G. (2013). A new theory for public service management? toward a (public) service-dominant approach. *The American Review of Public Administration*, 43(2):135–158.
- Perel, E. and Yechiali, U. (2017). Two-queue polling systems with switching policy based on the queue that is not being served. *Stochastic Models*, 33(3):430–450.
- Phung-Duc, T., Masuyama, H., Kasahara, S., and Takahashi, Y. (2010). A simple algorithm for the rate matrices of level-dependent qbd processes. In *Proceedings of the 5th international conference on queueing theory and network applications*, pages 46–52.
- Schwartz, B. L. (1974). Queuing models with lane selection: a new class of problems. *Operations Research*, 22(2):331–339.
- Shortle, J. F., Thompson, J. M., Gross, D., and Harris, C. M. (2018). *Fundamentals of queueing theory*, volume 399. John Wiley & Sons.
- Shumsky, R. A. (2004). Approximation and analysis of a call center with flexible and specialized servers. *OR Spectrum*, 26(3):307–330.
- Stanford, D. A. and Grassmann, W. K. (1993). The bilingual server system: A queueing model featuring fully and partially qualified servers. *INFOR: Information Systems and Operational Research*, 31(4):261–277.