ORIGINAL ARTICLE

International Journal of Applied Mathematics in Control Engineering

Journal homepage: http://www.ijamce.com

Optimal Control of the Multi-skill Call Centers with Delay Information

Chunyan Li[1]* | Wei Wu[1]

[1]Zhejiang Financial College, Hangzhou, 310000, China; lichunyan@zfc.edu.cn (C.L)

Correspondence
Chunyan Li, Zhejiang Financial College, Hangzhou, 310000, China
Email: lichunyan@zfc.edu.cn

Article Info
Article history:
Received 21 July 2026
Accepted 30 September 2026
Available online 12 Octorber 2026

Abstract

In order to achieve optimal control of the call center system, one of the important operation management means of call centers is to provide delay information for customers to alleviate system congestion. This paper studies the multi-skilled call center with delay information, in which customers will be informed to wait a certain time upon their arrival, and analyzes the influence of queuing system with delay information. There are two types of calls in the model. For the model with infinite patience customers, the two types of calls calculation formulas of waiting time are given. For the model with impatience customers, a method based on Markov chains was applied to estimate how long customers need to wait, and the effects of the delay information on the staffing problem are also investigated through numerical examples.

KEYWORDS

Optimal control, call center, multi-skills, queuing model, delay information

1 | INTRODUCTION

With the continuous development of social economy and science and technology, market competition is becoming increasingly fierce, and the diversity and randomness of customer needs have led to a gradual shift in competition between enterprises from product competition to service competition. As one of the main links between enterprises and customers, call centers have become a focus of attention for managers and academia on how to exercise optimal control over call centers. Call Center, that is Customer Service Center can quickly, correctly and kindly provide customers with information consultation, business acceptance and complaints and other services [1]. In recent years, with the rapid development of call centers, they have gradually become larger, and there are more and more multi-skilled call centers. In a multi-skill call center, incoming calls are divided into different types, and agents can also have various skills. The complexity of call center skills makes the original theory supporting single-skill call center operation service level calculation unable to continue to be applied in multi-skill call centers. Therefore, research on multi-skilled call centers has received more and more attention. Gans et al. comprehensively introduced the situation that the number of skills possessed by an agent in a multi-skill call center is less than or equal to 3, and provided detailed information on V-design, N-design, M-design, W-design and so on [2]. Stanford and Grassman considered an N-design call center queuing model with a fixed priority policy [3]. Shumsky proposed an approximate analysis of this model [4].

Digital Object Identifiers: 10.62953/IJAMCE.808444

Provide delay information for customers is a new operation mode of the call center, which means that the call center informs customers of the estimated waiting time by means of voice prompts. Managers use this model to effectively regulate demand and alleviate system congestion issues. Informing customers of the waiting time they need has a special and important role in service systems where the queue cannot be seen, such as call center service systems. In such a service system where the queue is invisible, the impatience caused by waiting is much higher than that in the system where the queue is visible, and this impatience does not gradually decrease over time, and the customer has no way to estimate the length of the queue, or the degree of progress, so emotions such as impatience and anxiety will continue to increase with waiting. Providing queuing information prompts for customers can greatly avoid the occurrence of these situations, so that the waiting process can be accepted by customers. The queuing information prompt has a great influence on the customer's waiting psychology and can improve the customer's satisfaction.

Predicting effective delays in Markov queuing models involves transient analysis of birth-death processes, and many papers propose the use of birth-death processes to study the instantaneous behavior of queuing. But in general it is difficult to obtain analytical solutions, and Whitt introduced some results on prediction delays [5]. Jouini and Dallery studied the problem of predicting delay using a birth-death process [6]. Ibrahim and Whitt studied the delay prediction problem of a multi-waiter queuing system with time-varying arrival rate, where the number of waiters is also time-varying and customers can abandon the queue [7]. The next further research is to notify the predicted delay time, Whitt studied the impact of notification delay on the performance of single-skill call center [8]. Jouini et al. further studied some related works later, his research method is based on queuing analysis, and a different research method is based on flow model [9, 10]. Armony et al. used this method to study the effect of delay information on multi-waiter queuing with abandonment [11]. Gou and Zipkin further analyzed and compared the impact of different layers of delay information on the queuing system [12]. Yu Miao et al. considered a call center fluid approximation method for a single-skill call center waiting and prompting strategy [13]. Jouini et al. studied an approximate M(t)/M/s(t) call center queuing model, used Erlang distribution to approximate the delay distribution, proposed two delay estimates and verified them with simulation methods [14]. Notification delay time has a large impact on customer response, and there is a lot of research in this area. The earliest literature on the impact of delayed information on customers can be found in Naor [15]. Taylor's study pointed out that in an experiment involving flying flights, delay information directly affects customer service evaluation [16]. Allon et al. studied the customer's response when the call center provided delayed information and the decision-making problem of the company's response to the customer [17]. Some issues related to optimal control can be found in reference [18, 19, 20].

The above literature are all studied the traditional single-skill call centers. This paper further expands the research on the multi-skill call centers, it studies a N-design multi-skilled call center with delay information. The system has two types of calls, at the same time, it has 2 types of agent groups, each with different skills, using the knowledge of the two-dimensional Markov process and the birth-death process, the calculation formula of the average waiting time of phone 1 and phone 2 is given, and the numerical analysis is carried out. The impact of prompting delay information on the manpower requirement of the system is discussed.

This article consists of the following parts. In Section II, the model of N-design multi-skill call center with infinite patient customers is discussed. In Section III, further research the model of N-design multi-skill call center with impatient customers, the calculation formula of the average waiting time required when a call 1 is reached, and the average waiting time when a call 2 is reached are analyzed. In Section IV, the influence of delay information on the optimal agents number demand of N-design model is studied. In Section V, some conclusions of the article are summarized.

2 | THE MODEL OF N-DESIGN MULTI-SKILL CALL CENTER WITH INFINITE PATIENT CUSTOMERS

2.1 | System model

This section studies a N-design multi-skilled call center with delay information. The system has two types of calls. The arrivals of these two types of calls are independent of each other and obey the Poisson process with the arrival rates λ_1 and λ_2 respectively. At the same time, the system has 2 types of agent groups, each with different skills. Among them, agent group 1 can

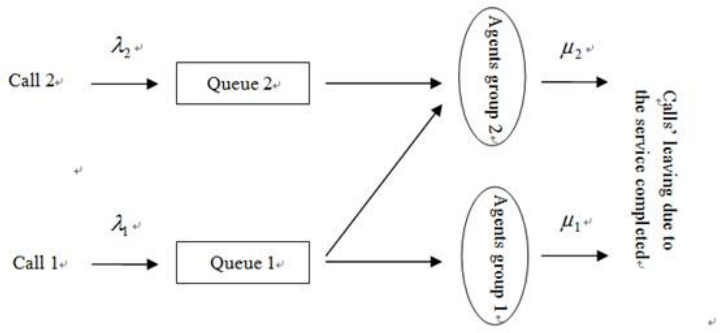


FIGURE 1 Call center model with infinite patience customers for N-design

only serve call 1, the service time is subject to exponential distribution, the service rate is μ_1, and the number of agents is N_1. The agent group 2 can serve the call 1 or the call 2, the service time also obeys the exponential distribution, the service rate is μ_2, and the number of agents is N_2. It is assumed that the same agent group has the same service rate for different types of calls, and at the same time, it is assumed that the calls received by the agents can be serviced satisfactorily at one time, that is, there will be no situation that calls are transferred between different agents.

The routing policy of this model is skill-based routing, while considering the importance of 2 different types of calls, this paper assumes that call 1 are more important than call 2, that is, call 1 has non-preemptive priority. When the call arrives, call 2 is undoubtedly assigned to the service of agent group 2, and for call 1, it is assigned to the service of agent group 1 first. When the agent group 1 is all busy, the service of agent group 2 is selected. If all are busy, enter queue 1 and wait in line. When the agent ends the service, the agent group 2 preferentially selects call 1 for service, and only selects call 2 for service when the queue of call 1 is empty, but agent group 1 only serves call 1. The queues of the 2 types calls are independent of each other. It is assumed that the queue space is infinite, i.e. calls will not be dropped by the system due to insufficient queue space. All calls of the same type are on a first-come, first-served discipline.

The impatient characteristics of customers are not considered in this model, so prompting customers for delay information will not affect the original model, so whether to inform customers of the queuing time they need to wait has no effect on the performance indicators of the system. But informing customers how long they need to wait can greatly reduce the uncertainty in the customer waiting process, thereby improving customer satisfaction.

2.2 | The calculation of the waiting time

Considering that a new call arrives in the system, there are two possible situations. One is that there is an idle waiter who can serve it in the system, then the new call will be served immediately without waiting. So the system does not need to provide it with any waiting information. Another situation is that when the call arrives, the agents in the system can provide services are all busy, so it must wait in the queue for service, and the system will prompt him with queuing information to inform him of the waiting time.

For modern call center systems, the state of the system is known when each new call arrives, and the state of the system is defined by the number of customers in the system. Let n_1 denote the number of call 1 in queue 1 seen by the newly arrived call 1, n_2 denote the number of call 2 in queue 2 seen by the newly arrived call 2, and let n_T denote the total number of customers in queue both in queue 1 and queue 2. The specific model is shown in Figure 1 below. At this time, there are n_1 customers in queue 1 and n_2 customers in queue 2.

Let's calculate the average waiting time of call 1 and call 2. Suppose the system agents are busy, and $n_T = n_1 + n_2 \geq 0$, there are calls waiting in the queue. The waiting time of call 1 and the waiting time of call 2 are studied separately. Because call 1 has non-preemptive priority, call 1 will occupy the agents of call 2 for service, so that the waiting time of call 2 is affected by the arrival of call 1, therefore, separate studies are required.

2.2.1 | The calculation of call 1 waiting time

Due to the routing policy of this model, call 1 has priority. When the agents in agent group 1 are all busy, the agents in agent group 2 can also serve call 1, and call 1 is preferentially selected for service. At this time, two agent groups serve call 1 at the same time. Agent group 1 has N_1 agents with a service rate of μ_1, and agent group 2 has N_2 agents with a service rate of μ_2. Whether it is the agent group 1 service completes one or the agent group 2 service completes one, the next call to select the service is call 1. The time that the newly arrived call 1 waits in the queue has nothing to do with the number of call 2 waiting in the queue, and it also has nothing to do with the call 1 that arrives later, because the system is first come, first served.

Assuming that when a new call 1 arrives, there are n_1 call 1 in queue 1 in the system, and all agents are in a busy state, then the newly arrived call 1 must wait for the n_1 calls in front of it to enter the service state, and When one is just finished serving, it can enter the state of being served. The service time of the system obeys the exponential distribution, and the mean is $1/(N_1\mu_1+N_2\mu_2)$, because at this time there are N_1+N_2 agents in the system serving call 1 at the same time, the service rate of N_1 agents is μ_1, and the service rate of N_2 agents is μ_2. Let $W_{n_1}^1$ represents the actual waiting time of call 1, that is, the waiting time of call 1 entering the system before it gets service. The waiting time $W_{n_1}^1$ is a random process of pure extinction. Therefore, it can be obtained that the waiting time of the newly arrived call 1 is the sum of n_1+1 independent and identically distributed exponential random variables with parameter $N_1\mu_1+N_2\mu_2$, which is an Erlang distribution of order n_1+1.

Set $\mu = N_1\mu_1+N_2\mu_2$, then the probability distribution density function of the waiting time $W_{n_1}^1$ of call 1 can be obtained as

$$f_1(t) = \frac{\mu(\mu t)^{n_1}}{n_1!}e^{-\mu t}, \quad t \geq 0 \tag{1}$$

The distribution function is

$$F_1(t) = 1 - e^{-\mu t}\sum_{i=0}^{n_1}\frac{(\mu t)^i}{i!}, \quad t \geq 0 \tag{2}$$

The average waiting time is

$$E[W_{n_1}^1] = \frac{n_1+1}{\mu} \tag{3}$$

The variance is

$$D[W_{n_1}^1] = \frac{n_1+1}{\mu^2} \tag{4}$$

2.2.2 | The calculation of call 2 waiting time

The calculation of the waiting time of call 2 is more complicated than that of call 1. Due to the priority of call 1, call 2 can only get service after call 1 has been serviced. Suppose a newly arrived call 2 enters the system. At this time, call 1 is queued in the system, there is n_1 calls in queue 1, call 2 is also queued, and there is n_2 calls in queue 2, then this call 2 before entering the service must wait for all n_T calls have been completely served, and the call 1 that may arrive during this period will also be serviced, and a call 2 has just been serviced. Since all calls are independent of each other, and the system can work continuously, it must be noted that the waiting time of call 2 does not depend on the service order of the customer in front of it. The call in front of call 2 can be divided into two parts, one part is the time it takes for call 1 to complete the service, the other part is the time it takes for call 2 to complete the service. Call 2 can only be served by agent group 2, while call 1 can be served by agent group 1 or agent group 2.

Let $W_{n_T}^2$ denote the actual waiting time of call 2, that is, the waiting time of call 2 after entering the system before getting service. Let $W_{n_2}^2$ denote the time to serve n_2 call 2, and $W_{n_1}^2$ denote the time to serve n_1 call 1 and call 1 arriving during this period, then the actual waiting time of call 2 is $W_{n_T}^2 = W_{n_1}^2 + W_{n_2}^2$. First, let's look at the service time $W_{n_2}^2$ of n_2 call 2. At this time, there are N_2 agents in the system to serve the call 2, the service rate is μ_2, the service time obeys an exponential distribution, and the mean is $1/(N_2\mu_2)$. From this, it can be obtained that the waiting time for service call 2 is the sum of n_2 independent and identically distributed exponential random variables with parameter $N_2\mu_2$, which is a n_2 order Erlang distribution.

Then the probability distribution density function of the waiting time $W_{n_2}^2$ of the second part of the call 2 can be obtained as

$$f_2(t) = \frac{N_2\mu_2(N_2\mu_2 t)^{n_2}}{n_2!} e^{-N_2\mu_2 t}, \quad t \geq 0 \tag{5}$$

The distribution function is

$$F_2(t) = 1 - e^{-N_2\mu_2 t} \sum_{i=0}^{n_2} \frac{(N_2\mu_2 t)^i}{i!}, \quad t \geq 0 \tag{6}$$

The average waiting time is

$$E[W_{n_2}^2] = \frac{n_2 + 1}{N_2\mu_2} \tag{7}$$

Next, let's calculate the waiting time $W_{n_1}^2$ of serving n_1 call 1 and the call 1 arriving during this period. Due to the non-preemption priority of call 1, there are $N_1 + N_2$ agents in the system serving call 1 at the same time. The service rate of N_1 servers is $N_1\mu_1$, the service rate of N_2 servers is $N_2\mu_2$, and at the same time the arrival rate of call 1 is λ_1. Because the service time of calls obeys the exponential distribution, which are independent of each other and have nothing to do with the calls that have been serviced, then the service rate of the system for call 1 is $N_1\mu_1 + N_2\mu_2$, at the same time, there are λ_1 more call 1 arriving, so the number of calls that can be served per unit time is $N_1\mu_1 + N_2\mu_2 - \lambda_1$. It is difficult to obtain the probability density function of the waiting time $W_{n_1}^2$, but the mean value of the waiting time $W_{n_1}^2$ is relatively easy to obtain. By summing the average service time of n_1 call 1, we can obtain

$$E[W_{n_1}^2] = \frac{n_1}{N_1\mu_1 + N_2\mu_2 - \lambda_1} \tag{8}$$

From the above analysis of the two parts, The average wait time of call 2 is

$$E[W_{n_T}^2] = E[W_{n_1}^2] + E[W_{n_2}^2] = \frac{n_1}{N_1\mu_1 + N_2\mu_2 - \lambda_1} + \frac{n_2 + 1}{N_2\mu_2} \tag{9}$$

3 | THE MODEL OF N-DESIGN MULTI-SKILL CALL CENTER WITH IMPATIENT CUSTOMERS

3.1 | System model

In this section, we further consider adding the customer impatience feature to the model, allowing customers to leave due to impatience while waiting in the queue. A multi-skilled N-design call center with delay information and impatient customers is studied here. The system has two types of calls and two types of agent groups. The description of the model is the same as that in Section 2.1, but here the impatient character of the customer is added. When customers enter the system, they will hear the queuing information provided by the system, that is, the waiting time. Each customer has a certain patience value. If the waiting time is longer than the customer's patience value, the customer will leave the system and will not come back. With the impatient feature of customers added, the calculation of the waiting time in the queue is more complicated.

Compared to the model with infinite patient customers in the previous section, the model in this section is more complex, because customer behavior may change, so it's no longer simple to calculate how long a customer will have to wait in the queue. In this section, we will consider the effect of notifying customers of the waiting time in the queue on customer abandonment behavior. When a customer is notified of the waiting time, the customer may feel that the waiting time is too long and immediately hang up and leave, or may find the wait time acceptable and wait in the queue for service. For the latter possibility, there are still two results. One is that the customer will never give up and wait until the service is accepted. The other is that the customer's patience will change, that is, although the customer has chosen to wait in the queue, he will still be impatient and leave during the waiting process. The customer's behavior will change according to the waiting time in the queue that is notified to the customer. This waiting time will be different from the waiting time expected by the customer, thus, it affects the customer's next decision, whether to leave or continue to wait. See Armony et al. [11] and Guo et al. [12] for some details on this.

There are many forms of queuing waiting information prompts. For example, according to the number of people in the queue,

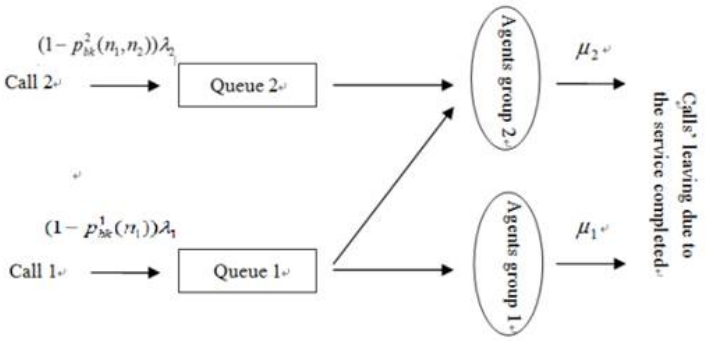


FIGURE 2 Call center model with impatient customers

providing the number of customers waiting in the queue. Or according to the waiting time, provide the waiting time the customers need to wait. The best is to tell a newly arriving customer the exact time he needs to wait, but this is random and impossible to know in advance. In practice it is natural to give an approximation of the actual waiting time for newly arriving customers, that is the estimated waiting time. The call center provides customers with an estimated waiting time with a reliability level of β, and the reliability level is the probability that the actual waiting time of queuing does not exceed the indicated waiting time of queuing.

Assuming that each arriving customer has his own patience value, when a new customer arrives, the customer will be served immediately if there is a agent available for him in the system. Otherwise, if all the agents are busy, when he hears the delay information provided by the system, if it goes beyond his patience value, he will choose to leave the system immediately with a certain probability, otherwise he will enter the system to queue for service. It is assumed here that after the customer chooses to enter the queue and wait in the queue, he will not leave the system. This assumption is also reasonable when the reliability of the system is high, because the estimated waiting time in this case is also relatively accurate. For the situation that may be impatient to leave after joining the queue, the literature of Jouini et al. [10] studied the situation of only one type of customer.

When a new call arrives, the number of call 1 in queue 1 is n_1, the number of call 2 in queue 2 is n_2, and let n_T represents the total number of calls in queue 1 and queue 2 in the system. Here we only focus on the situation when all the agents are busy. When a new call arrives, the system will notify it the waiting time in the queue. When the prompted waiting time is greater than the customer's patience value, the customer will choose to exit directly and never come back. Here, the patience value of call 1 is set as T_1, that is, the maximum time that call 1 is willing to wait, which obeys the exponential distribution with parameter θ_1. The patience value of call 2 is T_2, that is, the maximum time that call 2 is willing to wait, which obeys the exponential distribution with parameter θ_2. From the previous analysis, it can be seen that for the N-design multi-skill call center, the situation of call 1 and call 2 are completely different. call 1 has non-preemption priority, and Call 2 must wait until all queued call 1 have been served before being served. Therefore, the exit probability of the newly arrived call 1 is only related to n_1, set as $P_{bk}^1(n_1)$. And the exit probability of the newly arrived call 2 is related to n_1 and n_2, set as $P_{bk}^2(n_1,n_2)$.

Set $Y_{n_1}^1$ to be a random variable of wait time when call 1 arrives and there are n_1 call 1 in the front queue. Let $Y_{(n_1,n_2)}^2$ to be a random variable of the waiting time when call 2 arrives, and there are n_1 call 1 and n_2 call 2 in the front queue. Further let $G_{n_1}^1(t)$ and $G_{(n_1,n_2)}^2(t)$ be the probability density functions of $Y_{n_1}^1$ and $Y_{(n_1,n_2)}^2$ when $t \geq 0$, respectively. Let the estimated waiting times provided by the call center for arriving call 1 and call 2 be $w_{n_1}^1$ and $w_{(n_1,n_2)}^2$, respectively are their expectations of the actual waiting time. Then when the prompt message is heard, the probability of call 1 and call 2 leaving impatiently is:

$$P_{bk}^1(n_1) = P(T_1 < w_{n_1}^1) = 1 - e^{-\theta_1 w_{n_1}^1} \tag{10}$$

$$P_{bk}^2(n_1,n_2) = P(T_2 < w_{(n_1,n_2)}^2) = 1 - e^{-\theta_2 w_{(n_1,n_2)}^2} \tag{11}$$

The system model with delay information is shown in Figure 2:

It should be noted here that this paper assumes that the customer's decision to leave is independent of each other, so the customer's arrival process is still a Poisson process.

3.2 | The calculation of the waiting time

For a modern call center, when each new call arrives, the state of the system is known, and when there is a agent available in the system that can serve it, the new call will be served immediately without waiting. So the system does not need to provide it with any waiting information. When the call arrives, all the agents that the system can provide are in a busy state, so it must wait in the queue for service, and the system will prompt him with queuing information to inform it of the time it needs to wait. Here, the waiting time $Y^1_{n_1}$ and $Y^2_{(n_1,n_2)}$ of call 1 and call 2 are separately studied. Due to the setting of the system routing policy, the analysis of call 2 is much more complicated than that of call 1. First, the actual waiting time is calculated, and then the probability of calls leaving after hearing the prompt is further calculated.

3.2.1 | The calculation of call 1 waiting time

Due to the setting of the routing policy in this model, call 1 has priority. When the agents in group 1 are all busy, the agents in group 2 can also serve call 1, and call 1 is preferred for service. At this time, both agent groups serve call 1 at the same time, agent group 1 has N_1 agents, service rate is μ_1, and agent group 2 has N_2 agents, service rate is μ_2. Whether it is the agent group 1 service completes one or the agent group 2 service completes one, the next call to select is always call 1. The time that the newly arrived call 1 waits in the queue has nothing to do with the number of call 2 waiting in the queue, and it also has nothing to do with the call 1 that arrives later, because the system is first come, first served.

Assuming that when a new call 1 arrives, there is n_1 call 1 in queue 1 in the system, and all agents are in a busy state, then the newly arrived call 1 must wait for the call in front of it to enter the service state, and When one is just finished serving, it can enter the state of being served. Since the previous assumption of this model is that when customers enter the queue and wait in the queue, they will not leave, so the service time of the system obeys an exponential distribution, and the mean is $1/(N_1\mu_1+N_2\mu_2)$. Because there are N_1+N_2 agents in the system serving call 1 at the same time, the service rate of N_1 agents is μ_1, and the service rate of N_2 agents is μ_2. Therefore, it can be obtained that the waiting time of the newly arrived call 1 is the sum of n_1+1 independent and identically distributed exponential random variables with parameter $N_1\mu_1+N_2\mu_2$, which is an Erlang distribution of order n_1+1.

Set $\mu = N_1\mu_1+N_2\mu_2$, then the probability distribution density function of the random variable $Y^1_{n_1}$ of the waiting time of call 1 can be obtained as

$$G^1_{n_1}(t) = \frac{\mu(\mu t)^{n_1}}{n_1!}e^{-\mu t}, \quad t \geq 0 \tag{12}$$

The average waiting time is

$$E[Y^1_{n_1}] = \frac{n_1+1}{\mu} \tag{13}$$

Variance is

$$D[Y^1_{n_1}] = \frac{n_1+1}{\mu^2} \tag{14}$$

Then the estimated waiting time $w^1_{n_1} = E[Y^1_{n_1}]$ provided by the system to the call 1 can be obtained, and then the probability $P^1_{bk}(n_1) = P(T_1 < w^1_{n_1}) = 1-e^{-\theta_1 w^1_{n_1}}$ that the call 1 chooses to quit after hearing the waiting time prompted by the system can be calculated too.

3.2.2 | The calculation of call 2 waiting time

Let's analyze the waiting time of call 2. Due to the priority of call 1, call 2 can only get service after all calls 1 have been serviced. Suppose a newly arrived call 2 enters the system, at this time, call 1 is queued in the system, there are n_1 call 1 in queue 1, and call 2 is also queued, and there are n_2 call 2 in queue 2. Then, before this call 2 enters the service, it must wait for all n_T calls to be serviced, and the call 1 that may arrive during this period must also be serviced, and a call 2 has just been serviced

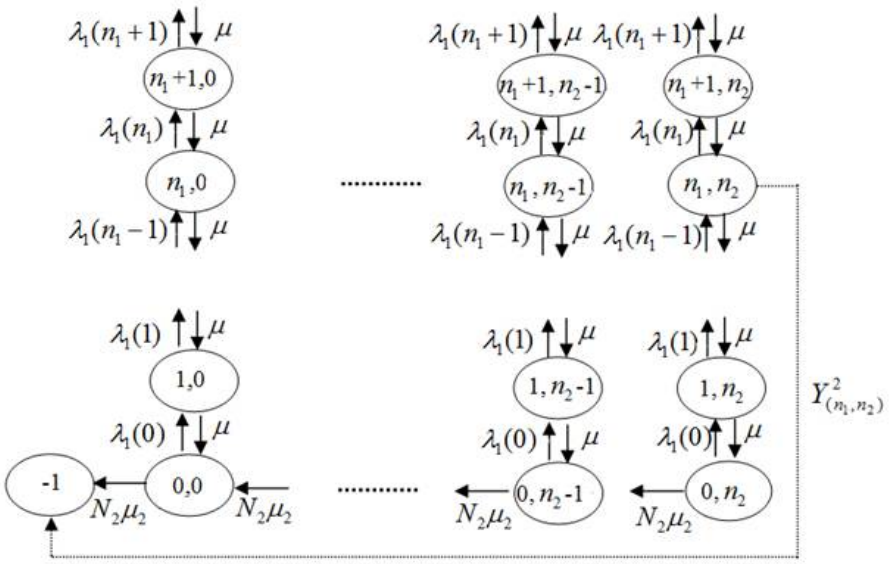


FIGURE 3 Diagram of the state-transition

before this call 2 can be serviced.

The waiting time random variable $Y^2_{(n_1,n_2)}$ of the newly arrived call 2 is independent of the later arriving call 2, because the system is first come, first served, but all arriving call 1 during the waiting period must be considered, because call 1 has priority, have to service call 1 first. Since it is assumed that the call enters the queue and waits in the queue and does not leave, only the call 1 that enters the queue during the waiting period is considered. So for the newly arrived call 2, the change of the queue in front of it is that as long as call 1 has not been served, the number of call 2's in front of it is always n_2, and it will not increase. After the service of call 1 is finished, the number of call 2 will continue to decrease as the service is completed. However, the number of call 1 is changing. When a call 1 arrives during the waiting period, the number will increase, and the number will decrease after serving one. At the same time, it should also be noted that after call 1 arrives, it will choose to leave or enter the queue to wait for service with a certain probability when it hears the prompt waiting information.

Based on the above explanation, suppose that the system is in a random state (m_1, m_2), which m_1 represents the number of calls 1 in queue 1, m_2 represents the number of calls 2 in queue 2, $m_1, m_2 \geq 0$, then the state transition process of the system can use a two-dimensional Markov Process representation. Also, suppose this Markov chain has an absorption state, denoted by (-1). When the system transitions to the (-1) state, both queues are empty, and one call 2 is serviced, that is, the newly arrived the (n_1+1)th call 2 is just ready to enter the service. When there are m_1 call in queue 1, since call 1 will leave with a certain probability after hearing the delay information, the arrival rate can be expressed as

$$\lambda_e(m_1) = \lambda_1\left(1 - P^1_{bk}(m_1)\right), \quad m_1 \geq 0 \tag{15}$$

Then the state transition diagram of the system at this time is shown in the following 3:

The transfer rate is

$$\begin{cases} q_{(m_1,m_2),(m_1+1,m_2)} = \lambda_1(m_1), & m_1, m_2 \geq 0 \\ q_{(m_1,m_2),(m_1-1,m_2)} = \mu, & m_1, m_2 \geq 0 \\ q_{(0,m_2),(0,m_2-1)} = N_2\mu_2, & m_2 \geq 0 \\ q_{(0,0),(-1)} = N_2\mu_2 \end{cases} \tag{16}$$

where $\mu = N_1\mu_1 + N_2\mu_2$.

From the state transition diagram, it can be seen that the random variable $Y^2_{(n_1,n_2)}$ of the waiting time of call 2 is the sum of the

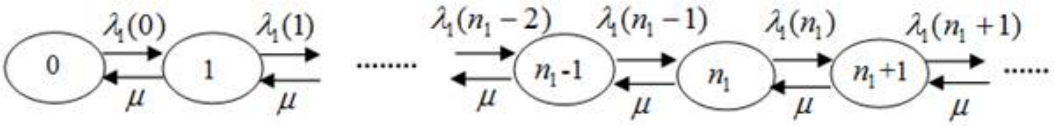


FIGURE 4 Diagram of the state-transition of call 1

waiting time from state (n_1, n_2) to state (-1), so the random variable $Y^2_{(n_1,n_2)}$ can be written as

$$Y^2_{(n_1,n_2)} = U(n_1) + V_{n_2} + \cdots + V_1 + V_0 \tag{17}$$

where $U(n_1)$ is a random variable of the time taken to service all calls 1. V_i, $i = 1,2,\cdots,n_2$, represents the random variable of the time used to serve each call 2, and V_0 represents the random variable of the time used to service another call 2, that is, the time it takes for the system to transition from state $(0,0)$ to state (-1).

According to the properties of Markov process, $U(n_1)$, V_i, $i = 1,2,\cdots,n_2$, V_0 these variables are independent of each other, and V_i, $i = 1,2,\cdots,n_2$, V_0 are independent and identically distributed. Let $E(Y^2_{(n_1,n_2)})$ represent the expectation of the random variable $Y^2_{(n_1,n_2)}$, that is, the average waiting time, then we can get

$$E(Y^2_{(n_1,n_2)}) = E(U(n_1)) + (n_2+1)E(V_0) \tag{18}$$

The following first calculates the expectation of the time $U(n_1)$ it takes to service call 1. In fact, the service process of call 1 is a process of birth and death, and its state transition process is shown in Figure 4:

Then the random variable $U(n_1)$ is the time it takes to transfer from state n_1 to state 0. Liu et al. [26] conducted a detailed study on this birth-death process and gave good results, which are directly quoted here. For the simplicity of the result representation, set $\delta_m = \lambda_1(m-1)$, $m \geq 1$, when $m = 0$, $\delta_0 = 1$, and set $\Phi_0 = 1$, then

$$\Phi_m = \frac{\prod_{j=0}^{m-1} \delta_j}{\mu^m}, \quad m \geq 1 \tag{19}$$

and we can obtain

$$E(U(n_1)) = \sum_{m=0}^{n_1} \frac{1}{\delta_m \Phi_m} \sum_{j=m+1}^{\infty} \Phi_j \tag{20}$$

Let's look at the expected time spent on service call 2. There are N_2 agents in the system who serves call 2, the service rate is μ_2, the service time obeys an exponential distribution, and the mean is $1/(N_2\mu_2)$. From this, it can be obtained the waiting time V_i, $i = 1,2,\cdots,n_2$, that service call 2, V_0 is n_2+1 independent and identically distributed exponential random variables with parameter $N_2\mu_2$, then we can get

$$E(V_0) = \frac{1}{N_2\mu_2} \tag{21}$$

In summary, the average waiting time of call 2 is

$$E(Y^2_{(n_1,n_2)}) = E(U(n_1)) + (n_2+1)E(V_0) = \sum_{m=0}^{n_1} \frac{1}{\delta_m \Phi_m} \sum_{j=m+1}^{\infty} \Phi_j + \frac{n_2+1}{N_2\mu_2} \tag{22}$$

From the above results, the estimated waiting time provided by the system to the call 2 can be further obtained as $w^2_{(n_1,n_2)} = E(Y^2_{(n_1,n_2)})$, and then the probability that the call 2 chooses to quit after hearing the waiting time prompted by the system can be calculated as:

$$P^2_{bk}(n_1,n_2) = P(T_2 < w^2_{(n_1,n_2)}) = 1 - e^{-\theta_2 w^2_{(n_1,n_2)}} \tag{23}$$

4 | THE INFLUENCE OF DELAY INFORMATION ON MANPOWER DEMAND OF N-DESIGN MODEL

The following is a calculation and analysis of the data of a logistics company's call center, and the results are shown in Table 1. The agent group 1 serves new customers, λ_1 and μ_1 represents the arrival rate and service rate of the new customer's shipment request. Agent group 2 serves old customers, λ_2 and μ_2 represents the arrival rate and service rate of old customers' sending requests and inquiries, while agent group 2 can also serve new customers and give priority to new customers. The data of 12 time periods were selected for the calculation of manpower requirements. It is assumed here that the cost parameters $C_1 = 3$ (thousand yuan), $C_2 = 4$ (thousand yuan), the unit of cost in the table is thousand yuan, and the time is $T_1 = 20$ (seconds), $T_2 = 30$ (seconds). The optimal staff calculation process is detailed in Article [21].

From the numerical results in Table 1, it can be further confirmed that the prompting of queuing information to customers will indeed affect the number of agents in the system, because some customers choose to leave the system after hearing the queuing information and do not accept the service, that is, the actual arrival rate of customers entering the system has changed, which has resulted in a change in the number of agents required by the system. The operation goal of the call center is to set the minimum number of agents under the premise of ensuring the customer service level. Therefore, it is necessary to consider the impact of customer behavior on the system, and it is necessary to take the queuing information prompt into the model.

TABLE 1 The comparison results of the optimal agents' number and cost

time periods	λ_1	μ_1	λ_2	μ_2	The model without delay information			The model with delay information		
					N_1	N_2	cost	N_1	N_2	cost
07:00–07:30	26	0.8	48	2.5	39	23	209	32	17	164
07:30–08:00	22	0.7	67	2.2	35	34	241	30	24	186
08:00–08:30	41	0.8	87	2.8	58	35	314	47	26	245
08:30–09:00	58	0.6	83	2.3	108	44	500	90	32	398
09:00–09:30	38	0.5	75	2.3	85	37	403	68	28	316
09:30–10:00	45	0.5	85	2.6	99	38	449	81	28	355
10:00–10:30	39	0.6	76	2.4	72	37	364	61	26	287
10:30–11:00	44	0.7	64	1.6	74	47	410	60	36	324
11:00–11:30	24	0.5	52	2.5	55	24	261	44	18	204
11:30–12:00	23	0.5	52	2.2	52	27	264	43	20	209
12:00–12:30	28	0.5	41	2.4	63	23	281	53	17	227
12:30–13:00	20	0.7	84	1.2	32	74	392	26	53	290

5 | CONCLUSIONS

This paper first studies the N-design multi-skill model with infinite patient customers. The system has two types of calls and two agent groups at the same time. Using the queuing model, when both queue 1 and queue 2 are queued in the system, the calculation formula of the average waiting time required when a call 1 is reached, and the average waiting time when a call 2 is reached are analyzed. Second, an N-design model with impatient customers is analyzed. The impatience here means that when customers enter the system, the system will provide them with delay information, to inform them of the time they need to wait in

the queue. However, each customer has a certain degree of patience. When the waiting time exceeds the customer's patience, then the customer will leave the system immediately and will not come back, otherwise the customer will enter the queuing system to wait for the service. Using the knowledge of the two-dimensional Markov process and the birth-death process, the calculation formula of the average waiting time of call 1 and call 2 is given. The numerical analysis is carried out, prompting customers with queuing information does indeed affect the number of servers in the system, as some customers may choose to leave the system and not accept service after hearing the delay information, resulting in a change in the number of servers required by the system. The operational goal of call centers is to set the minimum number of servers while ensuring customer service levels. Therefore, it is necessary to consider the impact of customer behavior on the system, delay information has an impact on the optimal control of the system.

AUTHOR CONTRIBUTIONS

Conceptualization, C.L. and W.W.; methodology, C.L.; software, W.W.; formal analysis, C.L.; investigation, W.W.; data curation, W.W.; writing—original draft preparation, C.L.; writing—review and editing, C.L.; visualization, Z.G.; supervision, C.L.; project administration, C.L.; funding acquisition, Z.G. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by Youth Foundation of Social Science and Humanity, China Ministry of Education (Project No. 22YJCZH038).

INSTITUTIONAL REVIEW BOARD STATEMENT

This study did not require ethical approval.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

Exclude this statement, because this study did not report any data.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

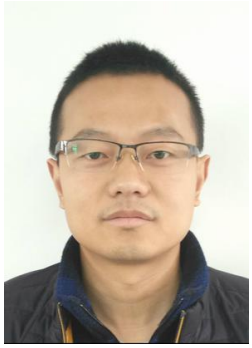
REFERENCES

[1] Hao J. Call center base on Internet. PhD thesis, Beijing Jiaotong University; 2004.
[2] Gans N, Koole G, Mandelbaum A. Telephone Call Centers: Tutorial, Review, and Research Prospects. Manufacturing & Service Operations Management 2003;5(2):79–141.
[3] Stanford DA, Grassmann WK. Bilingual Server Call Centres. In: Analysis of Communication Networks: Call Centres, Traffic and Performance Providence: American Mathematical Society; 2000.p. 31–47.
[4] Shumsky R. Approximation and Analysis of a Queuing System with Flexible and Specialized Servers. OR Spektrum 2004;26:307–330.
[5] Whitt W. Predicting Queuing Delays. Management Science 1999;45(6):870–888.
[6] Jouini O, Dallery Y. Moments of First Passage Times in General Birth-death Processes. Mathematical Methods of Operations Research 2008;68(1):49–76.

[7] Ibrahim R, Whitt W. Wait-Time Predictors for Customer Service Systems with Time-Varying Demand and Capacity. Operations Research 2010;59(5):1106–1118.
[8] Whitt W. Improving Service by Informing Customers about Anticipated Delays. Management Science 1999;45(2):192–207.
[9] Jouini O, Dallery Y, Nait-Abdallah R. Analysis of the Impact of Team-based Organizations in Call Centers Management. Management Science 2008;54:400–414.
[10] Jouini O, Dallery Y, Aksin OZ. Call Centers with Delay Information: Models and Insights. Manufacturing & Service Operations Management 2011;13(4):534–548.
[11] Armony M, Shimkin N, Whitt W. The Impact of Delay Announcements in Many-server Queues with Abandonment. Operations Research 2009;57(1):66–81.
[12] Guo P, Zipkin P. Analysis and Comparison of Queues with Different Levels of Delay Information. Management Science 2007;53(6):962–970.
[13] Yu M, Gong J, Kong F. Fluid Approximation of Call Center with Delay Information Strategy. Operations Research and Management Science 2020;29(12):118–124. In Chinese.
[14] Jouini O, Akşin OZ, Karaesmen F, Dallery Y. Call Center Delay Announcement Using a Newsvendor-like Performance Criterion. Production and Operations Management 2015;24(4):587–604.
[15] Naor P. The Regulation of Queue Size by Levying Tolls. Econometrica 1969;37(1):15–24.
[16] Taylor S. Waiting for Service: The Relationship between Delays and Evaluations of Service. Journal of Marketing 1994;58:56–69.
[17] Allon G, Bassamboo A, Gurvich I. "We Will Be Right with You": Managing Customers with Vague Promises. Operations Research 2011;59(6):1382–1394.
[18] Theodorou EA. Nonlinear Stochastic Control and Information Theoretic Dualities: Connections, Interdependencies and Thermodynamic Interpretations. Entropy 2015;17(5):3352.
[19] Abbasov AA, Mamedova ET, Mamtiev KK. On the Straight Line Method for Approximate Solution of Optimal Control Problems. Entropy 1987;13(6):450–465.
[20] Leike RH, Enßlin TA. Optimal Belief Approximation. Entropy 2017;19(8):402.
[21] Li C, Yue D. The Staffing Problem of the N-Design Multi-Skill Call Center Based on Queuing Model. In: 3rd International Conference on Wireless Communication and Sensor Networks (WCSN 2016); 2017. .



Chunyan Li obtained her PhD degree from the School of Economics and Management, Yanshan University, Qinhuangdao, China in 2017. She obtained her BS degree from the School of Science Yanshan University in 2003 and her MS degree from the School of Science Yanshan University in 2006. Her main research interests are in the areas of queueing models, operational research and so on.



Wei Wu is Master of Software Engineering, Beijing Institute of Technology, 2013. Currently working at Zhejiang Financial College, responsible for campus network management.