A Queuing Model of the N-design Multi-skill Call Center with Impatient Customers

Chunyan Li$^{1,3}$ and Dequan Yue$^2$

$^1$Yanshan University, School of Economics and Management, Hebei Qinhuangdao
$^3$Zhijiang College of Zhejiang University of Technology, College of Science, Zhejiang Hangzhou 310024

$^2$Yanshan University, College of Science, Hebei Qinhuangdao 066004, ydq@ysu.edu.cn$^2$, llccyy1980@126.com$^3$

Abstract

This paper studies a queuing model of the N-design multi-skill call center. In this model, there are two types of customers and two server groups who have different skills. Group 1 consists of specialized servers who can only serve one type of customers, Group 2 consists of flexible servers who can serve both two types of customers. The customers waiting for the service in the queue may leave the system due to impatience. By dividing the system’s state space into several sets of states, we obtain the equilibrium equations and the steady-state probabilities of the system. We also obtain the computational formula of the service level and the computational procedure of the staffing problem.

Keywords: multi-skill; call center; queuing system; impatient customer; steady-state probabilities

1. Introduction

Call Center, also called Customer Service Center, is a kind of comprehensive information service system, it is based on the Computer Telephone Integration technology, and it takes full advantage of communication network and computer network. Call center is widely used in many areas and industries, such as communication, finance, e-business, emergency centers and many others. Call center enables customers to obtain a fast, exact and warm response from the organizations, and the organizations could provide many services via call center. For many companies, such as hotels, airlines, and credit card companies, call centers provide a primary link between customers and service provider. It is very important for increasing the competitiveness of companies, improving customer satisfaction and developing new customers.

Recent years, the call center industry has been rapidly expanded and it has been spread all over the worldwide. The total number of call centers is increasing substantially, and call centers become bigger and bigger. With the rapid development of the technology, the call centers also have other channels, for example, email, fax, internet instant message and so on. These makes the agents in a call center more busier and the agents have to master many skills. But, in a call center, any one agent could not master all the skills. Nowadays, many call centers are of multi-skill type rather than of single-skill type. In single-skill call centers, there is no distinction among the calls and all agents are supposed to handle all calls. However, in multi-skill call centers, the calls are different types and the agents can handle several types of calls, i.e., they are of multiple skills.

On the other hand, multi-skill call centers make the call center’s operation and evaluation more complicated. Due to the complexity of multi-skill call centers, some methods used to analyze single-skill call centers are no longer applicable in analyzing multi-skill call centers. So, it needs urgently new methods to assess accurately the service.
level of multi-skill call centers. There are many researches on the performance evaluation of multi-skill call centers. Gans et al. [1] comprehensively considered the multi-skill call center for the cases of less skill groups (less than or equal to three groups), and provided detailed information on V-design, N-design, M-design, W-design and so on. Cezik and Bhulai studied the cases of more call types and more skill groups of the multi-skill call centers. Atlason [3] presented an iterative cutting plane method for minimizing staffing costs in a single-skill call center system. Cezik et al. [2] generalized Atlason’s method to multi-skill call centers by using the linear programming and simulation method. Bhulai et al. [4] presented a method of solving the model, the main characteristic of the method is to use firstly a heuristic algorithm to calculate the number of agents in each skill group every period and then to use linear programming method to solve the shift scheduling problem such that the cost is minimized. This method reduces the number of times of simulation. At the same time, Gurvich, Aksin and Jouini studied the case that all calls have different types and the servers can serve all customers. Gurvich et al. [5] studied large-scale service systems with multiple classes of customers and fully flexible servers, they investigated the question that how many servers are needed and how to match them with customers. They found that a single-class staffing rule and an idle-server-based threshold priority control are asymptotically optimal in the many-server heavy-traffic limiting regime. Aksin [6] studied a model that captured the impact of shared information processing resources on phone center performance, where they considered the reneging behavior. Furthermore, Aksin [7] studied a series of related problems. Jouini et al. [8] formulated and analyzed a multi-class call center model with priorities and impatient customers, where anticipated delays may be announced upon arrival. An approximation based on a normal distribution was proposed. Chevalie et al. [9] discussed the staffing problem of a block model, there is no waiting queue in the queuing model, when all agents busy, the call arriving will be blocked. The evaluation of the service level is the proportion of the call not be blocked. There are two types agents: the specialized servers (only have one skill) and the flexible servers (own all of the skills). The conclusion is that 80% of the specialized servers and 20% of the flexible servers is a kind of ideal match. Shumsky[10] also discussed the match problem of the specialized servers and the flexible servers under two skills condition. In addition, Dai Tao et al. [11] discussed a method for performance evaluation of W-design multi-skill call center based on Markov process.

This paper studies N-design multi-skill call center with impatient customers, where there are two types of calls and two groups of servers who have different skills. We present a new method of solving the model by dividing the system’s state space into several sets of states. Then, we obtain the equilibrium equations and the steady-state probabilities of the system. We also obtain the computational formula of the service level and the computational procedure of the staffing problem.

2. System Model

In this paper, we studies a queuing model of the N-design multi-skill call center with impatient customers, where there are two types of calls and two groups of servers. The model is shown in Figure 1.

1) There are two types of calls (or customers), Call 1 and Call 2. The two types of calls arrive according to a Poisson process with rates $\lambda_1$ and $\lambda_2$, respectively. There are two queues, Queue 1 and Queue 2, which consist of customers of Call 1 and Call 2, respectively. We assume that the calls through the call center’s selection system can be accurately classified.
2) We assume that customers may leave the queue due to impatience. The impatience time is exponentially distributed with the means $\theta$.

3) There are two categories of servers, Group 1 with $N_1$ servers and Group 2 with $N_2$ servers. Group 1 is of specialized servers who can only serve customers of Call 1, while Group 2 of flexible servers who can serve customers of both Call 1 and Call 2. The service times of servers in Group 1 and 2 are all exponentially distributed with means $\mu_1$ and $\mu_2$, respectively.

4) The routing policy of the model is based on skills and the importance of the two different types of calls. It is assumed that Call 1 is important than Call 2. In other words, Call 1 has non-preemptive priority that Call 2. When a server in Group 2 completes his (her) service, if there are customers of Call 1 waited in Queue 1 this server will service a customer waited in Queue 1, and if there is no customers of Call 1 waited in Queue 1 this server will serve a customer waited in Queue 2. When a server in Group 1 completes his (or her) service, if there is a customer waited in Queue 1 he (she) will serves a customers of Call 1, otherwise he (she) will be free if Queue 1 is empty.

5) There are infinite waiting spaces for both queues. For the same type calls, they are served in First-come First-serviced (FCFS) discipline. The queues of Call 1 and Call 2 are independent of each other.

3. The Calculation of the Steady-state Probability

3.1. The Division of the State Space

We divide the state space by considering the relationship between the number of calls and the number of agents in each group. This division is different from the division given by Dai Tao et al. [11].

There are 7 state sets in the system, let $S_i (i=1, 2, ..., 7)$ denote the specific state set, define $S_1$ is the state set that the agents in Group 1 are the idle state ($n_1 < N_1$), and the agents in Group 2 are the idle state ($n_2 < N_2$) too. $S_2$ is the agents in Group 1 are the idle state ($n_1 < N_1$), but the agents in Group 2 are the just full state ($n_2 = N_2$). $S_3$ is the agents in Group 1 are the idle state ($n_1 < N_1$), but the agents in Group 2 are the busy state ($n_2 > N_2$). $S_4$ is the agents in Group 1 are the just full state ($n_1 = N_1$), but the agents in Group 2 are the idle state ($n_2 < N_2$). $S_5$ is the agents in Group 1 are the just full state ($n_1 = N_1$), and the agents in Group 2 are the just full state ($n_2 = N_2$). $S_6$ is the agents in Group 1 are the just full state ($n_1 = N_1$), but the agents in Group 2 are the busy state ($n_2 > N_2$). $S_7$ is the agents in
Group 1 are the busy state \((n_1 > N_1)\), and the agents in Group 2 are the busy state \((n_2 > N_2)\) too. The state transition diagram is shown in Figure 2.

![State Transition Diagram](image)

**Figure 2. The State Transition Diagram**

### 3.2. The Division of the State Space

In this subsection, we calculate the transition rates among different sets of states by means of Figure 2. Let \(P(S_i)\), \((i=1,2,\ldots,7)\) denote the steady-state probability of each state set, \(q(S_i\rightarrow S_j)\), \((i, j=1,2,\ldots,7)\) denote the state set transition rate. The queues in the state sets are independent of each other, there are only two cases that the state will be changed: arrivals of calls or leaves of calls.

1. The change of states due to arrivals of calls

Consider the state-transfer of agents in Group 1 for set \(S_1\). The state transfer diagram is shown in Figure 3.

![Diagram of State-transfer of the Agents in Group 1 for Set \(S_1\)](image)

**Figure 3. The Diagram of State-transfer of the Agents in Group 1 for Set \(S_1\)**

From Figure 3, for \(n_1 \leq N_1 - 2\), if a call arrives at the system then the set \(S_1\) will not be changed. For \(n_1 = N_1 - 1\), if a call arrives at the system then set of states will be changed from set \(S_1\) to set \(S_4\). The trigger of the transfer from set \(S_1\) to set \(S_4\) is due to arrivals of Call 1, and the arrive rate is \(\lambda_1\). Thus, we can obtain the transfer rate from set \(S_1\) to set \(S_4\) as follows:
Where $P(n_1=N_1-1)$ is the probability of that the number of Call 1 is $n_1=N_1-1$ for the agents in Group 1 for set $S_1$. Note that $n_1<N_1$ and $n_2<N_2$, and that the two queues are independent of each other, the results of M/M/c/c loss queuing system can be used to our analysis. Thus, we have

$$P(n=N-1) = \frac{1}{(N-1)!} \sum_{i=0}^{N-1} \frac{\lambda_i^{N-1}}{i!}. $$

A similar analysis can also get the rest of the seven state transfer rate caused by call arriving, as follows:

$q(S_0 - S_2) = \lambda_2$,
$q(S_2 - S_0) = P(n_0 = N_0 -1) \cdot \lambda_2 \cdot q(S_0 - S_2) = \lambda_2$,
$q(S_3 - S_0) = P(n_0 = N_0 -1) \cdot \lambda_2 \cdot q(S_3 - S_0) = \lambda_2$,
$q(S_3 - S_1) = \lambda_1$,
$q(S_0 - S_1) = \lambda_1$.

(2) The change of states due to leaves of calls:

Figure 4. The Diagram of State-transfer of the Agents in Group 2 for Set $S_2$

One leaves of calls due to the service completed, the state of the agents will from the just full state to the idle state. Consider the state-transfer of agents in Group 2 for set $S_2$. The state transfer diagram is shown in Figure 4.

From Figure 4, in the set $S_2$ that $n_2=N_2$, if a call completed the service then set of states will be changed from set $S_2$ to set $S_1$. The trigger of the transfer from set $S_2$ to set $S_1$ is due to service completed of call 2, and the service rate is $N_2 \mu_2$. Thus, we can obtain the transfer rate from set $S_2$ to set $S_1$ as follows:

$q(S_2 - S_1) = N_2 \mu_2$.

A similar analysis can also get another four state transfer rate caused by service completed, as follows:

$q(S_1 - S_0) = N_1 \mu_1$; $q(S_4 - S_3) = N_2 \mu_2$;
\[ q(S_3 - S_2) = N_1 \mu_k \quad q(S_6 - S_3) = N_1 \mu_k. \]

Another state change is from the busy state (there is a queue) to the just full state, due to the service completed or impatience. Consider the state-transfer of agents in Group 2 for set \( S_3 \). The state transfer diagram is shown in Figure 5.

**Figure 5. The Diagram of State-transfer of the Agents in Group 2 for Set \( S_3 \)**

From Figure 5, for \( n_2 \geq N_1 + 2 \), if a call leaves due to the service completed or impatience then the set \( S_3 \) will not be changed. For \( n_2 = N_1 + 1 \), if a call leaves from the system then set of states will be changed from set \( S_1 \) to set \( S_2 \). The trigger of the transfer from set \( S_1 \) to set \( S_2 \) is due to leaves of Call 2. Thus, we can obtain the transfer rate from set \( S_1 \) to set \( S_2 \) as follows:

\[ q(S_3 - S_2) = P(n_2 = N_2 + 1) \times (N_2 \mu_2 + \theta) : \]

Where \( P(n_2 = N_2 + 1) \) is the probability of that the number of Call 2 is \( n_2 = N_2 + 1 \) for the agents in Group 2 for set \( S_3 \). Note that \( n_1 < N_1 \), and \( n_2 > N_2 \), and that the two queues are independent of each other, the results of M/M/c+M queuing[21] system can be used to our analysis. Thus, we have

\[ P(n_2 = N_2 + 1) = \frac{\lambda^2}{(N_2 \mu_2 + \theta)} \cdot N_2! \cdot P_0, \]

where,

\[ P_0 = \left[ \sum_{j=0}^{N_1} \frac{\lambda^2}{j! \mu_2} \right] + \sum_{k=N_1+1}^{\infty} \left[ \sum_{j=N_1+1}^{\infty} \prod_{j-N_1+1}^{k} \frac{\lambda^2}{(N_2 \mu_2 + (j-N_2) \theta) N_2!} \right]^{-1}. \]

A similar analysis can also get another 2 state transfer rate, as follows:

\[ q(S_6 - S_3) = P(n_2 = N_2 + 1) \times (N_2 \mu_2 + \theta) ; \]

\[ q(S_7 - S_6) = P(n_1 = N_1 + 1) \times (N_1 \mu_1 + N_2 \mu_2 + \theta). \]

### 3.3. The Establishment of the Equilibrium Equation and the Calculation of the Steady-state Probability

By the former analysis we can get 16 state transition rates, thus we can obtain the equilibrium equations of system, as follows:

\[ P(S_1)[q(S_1 - S_2) + q(S_1 - S_4)] = P(S_2)q(S_2 - S_1) + P(S_4)q(S_4 - S_1) ; \]

\[ P(S_2)[q(S_2 - S_1) + q(S_2 - S_3)] = P(S_1)q(S_1 - S_2) + P(S_3)q(S_3 - S_2) ; \]
The probability of the calls not be serviced in state set $S_j$ for call 1 is \( P(S_j) \equiv P(S_j) q(S_2-S_1) + q(S_1-S_6) \).
\[
P(S_1) q(S_2-S_1) + q(S_1-S_6) = P(S_1) [q(S_1-S_2) + q(S_1-S_6)] + P(S_6) q(S_6-S_1) ;
\]
\[
P(S_2) q(S_2-S_1) + q(S_2-S_5) = P(S_2) [q(S_2-S_3) + q(S_2-S_5)] + P(S_5) q(S_5-S_2) ;
\]
\[
P(S_3) q(S_3-S_1) + q(S_3-S_6) = P(S_3) [q(S_3-S_4) + q(S_3-S_6)] + P(S_6) q(S_6-S_3) ;
\]
\[
P(S_4) q(S_4-S_1) + q(S_4-S_6) = P(S_4) [q(S_4-S_5) + q(S_4-S_6)] + P(S_6) q(S_6-S_4) ;
\]
\[
P(S_5) q(S_5-S_1) + q(S_5-S_6) = P(S_5) [q(S_5-S_6) + q(S_5-S_6)] + P(S_6) q(S_6-S_5) ;
\]
\[
P(S_6) q(S_6-S_1) + q(S_6-S_6) = P(S_6) [q(S_6-S_6) + q(S_6-S_6)] + P(S_6) q(S_6-S_6) ;
\]

and $\sum_{j=1}^{7} P(S_j) = 1$, that is the sum of the steady-state probability of all the state sets is equal to 1.

From the above equations we can obtain all of the steady-state probabilities $P(S_i)$, $i=1,2,\ldots,7$. Because the equations are linear, using MATLAB software to solve, the calculating speed can be guaranteed.

### 4. The Calculation of the Call Center’s Service Level

The service level of most of the call centers is defined as: the percentage of the serviced calls in a given fixed waiting time, denoted by $P/T$. Actually, the 80/20 principle is a general rule, that is to say within 20 seconds waiting time, 80% of the calls should be serviced. The service level can also be expressed as the probability of the calls which cannot be serviced in a given fixed waiting time. We can get the service level using the steady-state probabilities above.

Such as call 1, here we assume that the service level of call 1 is defined as the probability of the calls not be serviced in $T_1$ time, call 1 has a queue only happen in state sets $S_7$. Due to the importance of call 1 is very high, if a server in group 2 completed a service, he will choose the call 1 in the queue to serve immediately, so in the state set $S_7$, the service rate of call 1 is $N_1\mu_1+N_2\mu_2$, thus we can get, in $T_1$ time, the number of the calls could be served is an integer not greater than $T_1 \times (N_1\mu_1+N_2\mu_2)$. By the above analysis we can get the probability of call 1 cannot be serviced in $T_1$ time is:

\[
P^1_{\text{none}} = P(S_7) \times P_{S_7} (n_1 \geq N_1 + N_2 + [T_1 \times (N_1\mu_1 + N_2\mu_2)])
\]

where,
\[
P_{S_7} (n_1 = i) = \prod_{j=N_1+1}^{N_1+N_2} \left( \frac{1}{N_1\mu_1 + (j-N_1)\theta} \right)^{\frac{\lambda_1^n}{\mu_1^N_1}} N_1! \cdot P_0,
\]

and,
\[
P_0 = \left[ \sum_{j=0}^{N_1} \left( \frac{\lambda_1^n}{\mu_1^N_1} \right) j! + \sum_{k=N_1+1}^{\infty} \sum_{j=N_1+1}^{\infty} \frac{\lambda_1^n}{\mu_1^N_1} \left( \frac{\lambda_1^n}{\mu_1^N_1} \right)^j \frac{1}{N_1!} \right]^{-1}
\]

For call 2, it is more complicated than call 1, but the analysis method is the same, call 2 have a queue happens in the state set $S_3$, $S_6$ and $S_7$, therefore the probability of call 2 cannot be served in $T_2$ time is:
5. Staffing Problem

Assuming that the cost of the servers group 1 is \( C_1 \), and the cost of servers group 2 is \( C_2 \), in order to minimum the cost, we try to find the optimal number of servers \( N_1 \) and \( N_2 \), the model is as follows:

\[
\begin{align*}
\text{min} & \quad C_1 N_1 + C_2 N_2 \\
\text{s.t.} & \quad P^1_{\text{none}} \leq \alpha_1 \\
& \quad P^2_{\text{none}} \leq \alpha_2 \\
& \quad N_1, N_2 \in \mathbb{Z}^+
\end{align*}
\]

Here the constraint conditions denote that the probabilities of each type calls cannot be serviced are less than or equal to \( \alpha_1 \), \( \alpha_2 \), respectively. \( N_1, N_2 \) are unknown positive integers for solving.

6. Conclusions

This paper studies a queuing model of the N-design multi-skill call center with impatient customers. The system has two types of customers and two server groups. One group can only serve one type of calls, the other can serve both two types of calls. We divide the system’s state space into regions by using the number of customers waiting before each servers group and the number of servers in each group. We obtain the equilibrium equations and the steady-state probabilities of the system. We also obtain the computational formula of the service level and the computational procedure of the staffing problem.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (71071133), and in part by the Natural Science Foundation of Hebei Province, China (G2012203136).

References


Authors

**Chunyan Li**, she received her B.Sci. degree in information and computing science and the M.Sci. degree in operations research and cybernetics from Yanshan University, Qinhuangdao, China. Now she is studying the Ph.D. degree in management science and engineering Yanshan University, and she is a university lecturer at the Department of Science, Zhijiang College, Zhejiang University of Technology, Hangzhou, China. Her research interests include performance analysis of queueing systems and reliability analysis of repairable systems.

**Dequan Yue**, he received his B.Sci. degree in applied mathematics from Northeast Heavy Machinery Institute, Qiqihai, China, the M.Sci. degree in applied mathematics from Xi’an Electronic University of Science and Technology, Xi’an, China, and Dr.Sci. degree in operations research and cybernetics from the Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing, China. Now Dr. Yue is a professor at Department of Statistics, Yanshan University, Qinhuangdao, China. His research interests include performance analysis of queueing systems, reliability analysis of repair systems, performance evaluation and modeling of communication networks, and stochastic order and its applications. Dr. Yue has published over 50 papers in journals including Naval Research Logistics, Operations Research Letters, Computer Communications, Microelectronics and Reliability, Nonlinear Dynamics and Systems Theory, International Journal of Pure and Applied Mathematics, Journal of Systems Science and Complexity, and other journals.