ORIGINAL ARTICLE

International Journal of Applied Mathematics in Control Engineering

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

# Research on Queuing Time Estimation for an M-type Multi-skill Call Center with Impatient Customers

**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 27 August 2026
Accepted 7 November 2026
Available online 8 November 2026

**Abstract**

Queue time notification refers to an operational feature in call centers that informs customers of their anticipated waiting duration through voice prompts. For service systems with invisible queues such as call centers, disclosing estimated waiting times carries particular practical importance. This paper addresses the waiting time computation problem for M type multi skill call centers with impatient customers. Adopting a Markov chain framework, we derive analytical expressions for queue waiting time. Numerical experiments are further conducted to quantify how queue time notification affects system staffing requirements.

**KEYWORDS**

multi-skill call center, impatient customer, Markov chain, waiting time calculation

## 1 | INTRODUCTION

Queue time notification represents a standard operational mechanism in modern call centers, whereby voice announcements deliver estimated waiting durations to waiting callers. This practice is especially valuable for service systems where end users cannot observe queue status directly. In such opaque queueing environments, customer impatience tends to be far more pronounced compared with systems with visible queues. Lacking visibility into queue length or service progress, callers experience mounting anxiety and frustration as waiting persists. By communicating expected waiting times, queue time notification substantially alleviates adverse customer sentiment and improves tolerance toward waiting periods. Prior literature has documented that waiting time information exerts a considerable influence on customer waiting psychology and enhances overall satisfaction. Motivated by this observation, this paper investigates N type and M type multi skill call centers embedded with queue time notification, and analyzes its system level impacts.

Multi skill call centers have attracted substantial research interest regarding their operational performance and service quality metrics. Gans et al. presented a comprehensive survey of multi skill call centers where each agent possesses no more than three skills, and introduced canonical configurations including V type, N type and M type architectures [1]. Specifically, an M type call center comprises two customer classes and three skill groups: one cross trained agent group capable of serving both customer classes, alongside two dedicated groups, each assigned exclusively to one customer class. Örmeci studied an M type setup where cross skilled agents operate with higher service efficiency; their findings suggested that prioritizing cross skilled agents constitutes an optimal control policy [2]. Dai and Huo developed an evaluation methodology for W type multi skill call centers and derived

Digital Object Identifiers: 10.62953/IJAMCE.755958

key performance indicators such as steady state probabilities using Markov process theory [3]. Zhang et al. performed simulation based comparisons of staffing demands under L type and V type routing strategies for multi skill call centers. Their simulation results indicated that the priority enabled V type routing policy minimizes staffing consumption under identical service level targets and agent occupancy constraints [4]. The present work focuses on deriving waiting time computation schemes for the M type multi skill call center configuration.

## 2 | M TYPE MULTI-SKILL CALL CENTER WITH IMPATIENT CUSTOMERS

### 2.1 | Model Description

We consider an M type multi skill call center queueing model featuring customer impatience. Two independent classes of calls arrive following Poisson processes with arrival rates $\lambda_1$ and $\lambda_2$, respectively. The system contains three agent groups with differentiated skill portfolios:

- Agent Group 1 handles only Class 1 calls. Service times obey exponential distributions with service rate $\mu_1$, and the group consists of $N_1$ agents.
- Agent Group 2 serves exclusively Class 2 calls. Service times are exponentially distributed with service rate $\mu_2$, with a total of $N_2$ agents.
- Agent Group 3 is cross skilled to process both Class 1 and Class 2 calls. Its service time also follows an exponential distribution with service rate $\mu_3$, and the group has $N_3$ agents.

Within each agent group, service rates remain homogeneous irrespective of call categories. Once a call is connected to an agent, service is completed in a single session with no inter agent call rerouting.

A skill based routing policy is implemented without inherent priority assigned to either call class. Upon arrival, Class 1 (Class 2) calls are first routed to Agent Group 1 (Agent Group 2). If all agents within the dedicated group are fully occupied, the incoming call is forwarded to Agent Group 3. When Group 3 also reaches full capacity, the call joins Queue 1 (Queue 2). Upon service completion, agents from dedicated Group 1 (Group 2) resume servicing backlogged Class 1 (Class 2) calls. When both queues hold pending calls, cross skilled agents in Group 3 select between the two call classes uniformly at random (probability $1/2$). The two queues operate independently with infinite buffer capacity; hence no calls are discarded due to queue space limitations. Each queue adheres to the first come first served (FCFS) service discipline.

Customer impatience is explicitly incorporated into the model: callers may abandon the system while waiting in queue. Arriving customers receive voice announcements of estimated waiting time. Every customer is assumed to have an intrinsic patience threshold. If the announced waiting estimate exceeds this threshold, the customer abandons the system permanently. Accounting for such abandonment behavior adds considerable complexity to waiting time calculations.

Each incoming customer is endowed with an individual patience limit. An arriving call receives immediate service provided that any qualified agent is idle. Conversely, when all agents are busy, the caller listens to the system generated waiting time notification. Should the announced duration surpass their personal patience threshold, the customer abandons the system with a certain probability; otherwise, the caller enters the queue and remains in the system until service is accomplished.

Let $n_1$ and $n_2$ denote the respective numbers of Class 1 calls in Queue 1 and Class 2 calls in Queue 2 upon a new call arrival. Our analysis concentrates on scenarios where every agent is occupied. Under such circumstances, the system broadcasts estimated waiting time to newly arrived calls. Customers depart permanently if the notified waiting time exceeds their patience. The patience threshold for Class 1 calls follows an exponential distribution with parameter $\theta_1$, representing the maximum tolerable waiting period for Class 1 customers. Analogously, Class 2 customers' patience thresholds follow an exponential distribution with parameter $\theta_2$. Owing to structural symmetry between the two call classes in our M type formulation, the abandonment probability for a newly arrived Class 1 call depends solely on $n_1$, denoted as $p^1_{bk}(n_1)$. Similarly, the abandonment probability for an incoming Class 2 call depends only on $n_2$, denoted as $p^2_{bk}(n_2)$.

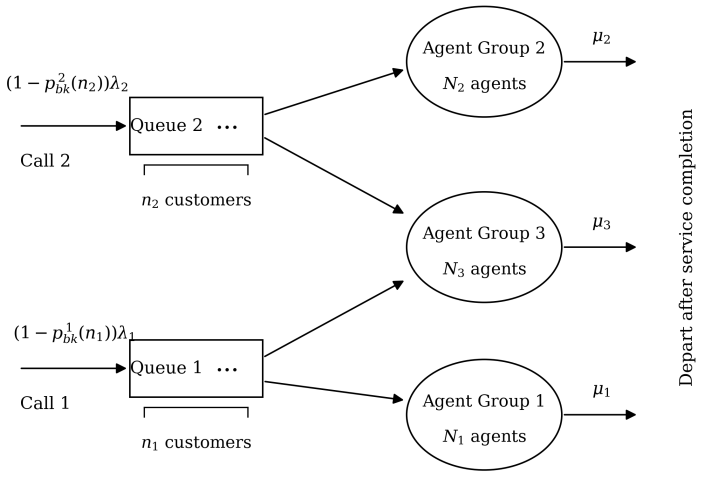



**FIGURE 1** M type call center model with impatient customers

Define $W_1(n_1,n_2)$ as the random waiting time for a newly arriving Class 1 call given $n_1$ Class 1 calls and $n_2$ Class 2 calls queued ahead of it. Likewise, $W_2(n_1,n_2)$ represents the random waiting time for an incoming Class 2 call. $f_{W_1}(\cdot)$ and $f_{W_2}(\cdot)$ stand for their respective probability density functions. Let $\hat{W}_1$ and $\hat{W}_2$ denote the waiting time estimates announced by the call center for Class 1 and Class 2 calls, which correspond to the expected actual waiting times for each call class. The abandonment probabilities triggered by waiting time announcements for Class 1 and Class 2 customers are presented in Eqs. (1) and (2):

$$p^1_{bk}(n_1) = P(T_1 < \hat{W}_1) = 1 - e^{-\theta_1 \hat{W}_1} \tag{1}$$

$$p^2_{bk}(n_2) = P(T_2 < \hat{W}_2) = 1 - e^{-\theta_2 \hat{W}_2} \tag{2}$$

Incoming Class 2 calls arrive at effective rate $(1-p^2_{bk}(n_2))\lambda_2$ and feed into Queue 2 with $n_2$ waiting calls. Queued Class 2 calls are served by dedicated Agent Group 2 ($N_2$ agents, service rate $\mu_2$) and cross skilled Agent Group 3 ($N_3$ agents, service rate $\mu_3$). Incoming Class 1 calls arrive at effective rate $(1-p^1_{bk}(n_1))\lambda_1$ and join Queue 1 containing $n_1$ waiting calls. Queued Class 1 calls receive service from Agent Group 1 ($N_1$ agents, service rate $\mu_1$) and cross skilled Agent Group 3. All calls exit the system once service finishes.

Since individual abandonment decisions are statistically independent, the aggregate call arrival process remains Poisson.

## 2.2 | Calculation of Customer Queue Waiting Time

Newly arrived calls obtain instant service without waiting time notifications as long as qualified idle agents are available. By contrast, when all eligible agents are busy, incoming calls are enqueued and notified of their projected waiting duration. This subsection characterizes waiting time distributions for both call classes and derives abandonment probabilities induced by waiting time alerts.

Under the adopted routing protocol, cross skilled agents from Group 3 pick Class 1 or Class 2 calls with equal probability $1/2$ whenever both queues are non empty. Agent Group 1 comprises $N_1$ agents with service rate $\mu_1$, and Agent Group 2 contains $N_2$ agents with service rate $\mu_2$. Subject to the FCFS rule, the waiting time for an incoming Class 1 (Class 2) call hinges only on the volume of preceding same class queued calls. It is independent of queue length for the opposite call class and of future incoming calls.

Consider a new Class 1 (Class 2) call arriving under full agent occupancy, with $n_1$ ($n_2$) same class calls already waiting in

Queue 1 (Queue 2). The incoming call cannot commence service until all preceding $n_1$ ($n_2$) calls have entered service and one service completion occurs. Per our modelling assumption, customers who have joined the queue will not subsequently abandon the system. Service times for Class 1 (Class 2) calls follow exponential distributions with mean $1/\mu_1$ ($1/\mu_2$). At this instant, $N_1+k_1$ agents concurrently serve Class 1 calls, yielding an aggregate service rate $(N_1+k_1)\mu_1$; correspondingly, $N_2+k_2$ agents serve Class 2 calls with combined service rate $(N_2+k_2)\mu_2$. Here $k_1$ ($k_2$) denotes the number of cross skilled agents from Group 3 currently serving Class 1 (Class 2) calls, with $0 \le k_1+k_2 \le N_3$. Accordingly, the waiting time for an incoming Class 1 (Class 2) call amounts to the sum of $n_1+1$ ($n_2+1$) independent and identically distributed exponential random variables with parameter $(N_1+k_1)\mu_1$ ($(N_2+k_2)\mu_2$), which follows an $(n_1+1)$ stage ($(n_2+1)$ stage) Erlang distribution.

For Class 1 calls, the probability density function of waiting time random variable $W_1$ is given by:

$$f_{W_1}(t) = \frac{(N_1+k_1)\mu_1\big((N_1+k_1)\mu_1 t\big)^{n_1}}{n_1!} e^{-(N_1+k_1)\mu_1 t}, \quad t \ge 0 \tag{3}$$

The corresponding mean waiting time reads:

$$E[W_1] = \frac{n_1+1}{(N_1+k_1)\mu_1} \tag{4}$$

Variance is computed as:

$$D[W_1] = \frac{n_1+1}{(N_1+k_1)^2\mu_1^2} \tag{5}$$

Let $\hat{W}_1$ represent the system announced estimated waiting time for Class 1 calls. The probability that Class 1 customers abandon upon receiving this notification is:

$$p_{bk}^1(n_1) = 1 - e^{-\theta_1 E[W_1]} = 1 - e^{-\frac{\theta_1(n_1+1)}{(N_1+k_1)\mu_1}} \tag{6}$$

For Class 2 calls, the probability density function of waiting time random variable $W_2$ takes the form:

$$f_{W_2}(t) = \frac{(N_2+k_2)\mu_2\big((N_2+k_2)\mu_2 t\big)^{n_2}}{n_2!} e^{-(N_2+k_2)\mu_2 t}, \quad t \ge 0 \tag{7}$$

Its mean waiting time is:

$$E[W_2] = \frac{n_2+1}{(N_2+k_2)\mu_2} \tag{8}$$

Variance:

$$D[W_2] = \frac{n_2+1}{(N_2+k_2)^2\mu_2^2} \tag{9}$$

$\hat{W}_2$ denotes the waiting time estimate announced for Class 2 calls. The resultant abandonment probability for Class 2 callers is:

$$p_{bk}^2(n_2) = 1 - e^{-\theta_2 E[W_2]} = 1 - e^{-\frac{\theta_2(n_2+1)}{(N_2+k_2)\mu_2}} \tag{10}$$

The foregoing analysis reveals that waiting time computation formulas are identical for M type models with infinite patience customers and those incorporating customer impatience. The core distinction lies in customer abandonment behavior, which emerges solely in the impatient customer setting.

## 3 | NUMERICAL ANALYSIS OF THE IMPACTS OF QUEUE TIME NOTIFICATION ON STAFFING REQUIREMENTS

Queue time notification reshapes effective arrival rates for the system. Upon receiving waiting time alerts, a fraction of impatient callers abandon the system, while others enter queues and await service completion. Reduced effective arrival rates may lower the required staffing level. To quantitatively evaluate this effect, we benchmark our proposed model against the baseline M

type multi skill call center staffing model reported in [5], referred to hereafter as Model 2. Model 2 implements a standard M type configuration without queue time notification functionality.

Numerical validations utilize field data collected from a logistics enterprise call center. Agent Group 1 processes service requests submitted by new customers, with arrival rate $\lambda_1$ and service rate $\mu_1$. Agent Group 2 handles service requests and inquiries from existing customers, with arrival rate $\lambda_2$ and service rate $\mu_2$. Cross skilled Agent Group 3 serves both customer segments with service rate $\mu_3$. Staffing requirements are evaluated across twelve time windows. Staffing costs for each agent group are measured in thousands of RMB, and waiting time quantities are reported in seconds. Table 1 contrasts optimal agent counts solved for Model 2 and our proposed model integrated with queue time notification.

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

| time periods | $\lambda_1$ | $\mu_1$ | $\lambda_2$ | $\mu_2$ | $\mu_3$ | The model without delay information | | | | The model with delay information | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $N_1$ | $N_2$ | $N_3$ | cost | $N_1$ | $N_2$ | $N_3$ | cost |
| 07:00–07:30 | 26 | 0.8 | 48 | 2.5 | 1.3 | 35 | 20 | 5 | 185 | 28 | 17 | 5 | 155 |
| 07:30–08:00 | 22 | 0.7 | 67 | 2.2 | 1.0 | 34 | 32 | 5 | 218 | 27 | 26 | 5 | 179 |
| 08:00–08:30 | 41 | 0.8 | 87 | 2.8 | 1.3 | 54 | 32 | 5 | 278 | 44 | 26 | 5 | 230 |
| 08:30–09:00 | 58 | 0.6 | 83 | 2.3 | 1.3 | 100 | 37 | 5 | 431 | 81 | 30 | 5 | 353 |
| 09:00–09:30 | 38 | 0.5 | 75 | 2.3 | 1.1 | 79 | 34 | 5 | 359 | 65 | 27 | 5 | 296 |
| 09:30–10:00 | 45 | 0.5 | 85 | 2.6 | 1.3 | 94 | 34 | 5 | 404 | 76 | 27 | 5 | 329 |
| 10:00–10:30 | 39 | 0.6 | 76 | 2.4 | 1.0 | 68 | 33 | 5 | 323 | 55 | 27 | 5 | 266 |
| 10:30–11:00 | 44 | 0.7 | 64 | 1.6 | 1.0 | 66 | 41 | 5 | 341 | 53 | 34 | 5 | 281 |
| 11:00–11:30 | 24 | 0.5 | 52 | 2.5 | 0.8 | 51 | 22 | 5 | 239 | 41 | 18 | 5 | 197 |
| 11:30–12:00 | 23 | 0.5 | 52 | 2.2 | 1.0 | 49 | 25 | 5 | 242 | 40 | 20 | 5 | 200 |
| 12:00–12:30 | 28 | 0.5 | 41 | 2.4 | 0.8 | 59 | 18 | 5 | 251 | 48 | 15 | 5 | 209 |
| 12:30–13:00 | 20 | 0.7 | 84 | 1.2 | 0.8 | 31 | 72 | 5 | 329 | 25 | 58 | 5 | 269 |

Numerical outputs from Table 1 demonstrate that queue time notification exerts a tangible influence on optimal staffing configurations. A proportion of callers abandon service upon receiving waiting time notifications, which reduces system load and reshapes optimal agent allocation. The primary operational goal for call centers is to minimize agent headcount while satisfying predefined service level constraints. Consequently, customer behavioral responses to waiting time information ought to be incorporated into system modelling. Integrating queue time notification within analytical models can prevent over staffing and mitigate unnecessary operational expenditures.

## 4 | CONCLUSIONS

This paper develops an M type multi skill call center queueing model that accounts for customer impatience. Drawing upon queueing theory principles, we derive closed form waiting time formulas for each call category. Since waiting time notifications alter customer abandonment decisions, modelling customer impatience bears important practical relevance. Waiting time analytical expressions are derived by means of Markov chain analysis. Numerical examples further quantify how queue time notification modulates staffing requirements. Informing arriving callers of their anticipated queueing duration enhances customer satisfaction and mitigates impatient abandonment behavior.

## 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, W.W.; supervision, C.L.; project administration, C.L. 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

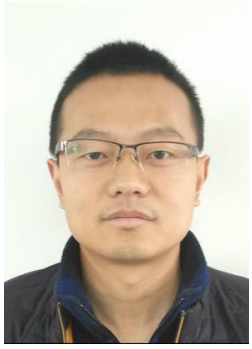
The authors declare no conflict of interest.

## REFERENCES

[1] Gans N, Koole G, Mandelbaum A. Telephone Call Centers: Tutorial, Review, and Research Prospects. Manufacturing & Service Operations Management 2003;5(2):79–141.
[2] Örmeci E. Dynamic Admission Control in a Call Center with One Shared and Two Dedicated Service Facilities. IEEE Transactions on Automatic Control 2004;49(7):1157–1161.
[3] Dai, T , Huo, J Z . An Evaluation Method for Multi-skill Call Centers Based on Markov Processes. Systems Engineering 2008;26(2):99–103.
[4] Zhang X Y , Zhang Y X , Huo J Z . Simulation Analysis of Staffing Requirements for L-type and V-type Routing Policies in Multi-skill Call Centers. Shanghai Management Science 2012;34(6):32–35.
[5] Li C, Wu W. Optimal Control of the Multi-skill Call Centers with Delay Information. International Journal of Applied Mathematics in Control Engineering 2026;9(1):142–153.



**Chunyan Li** obtained her Ph. D. 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.