2-layer Queuing Model for Online Customer Service with Variable Service Rates

Taotao Yu¹, Tao Dai¹,* and Xing Zhao¹
¹School of Donghua University, Shanghai, China

Abstract. We consider the Online Customer Service (OCS) platform where agents can serve multiple customers simultaneously. The paper constructs a 2-layer queuing model, including Message-layer and Customer-layer. Add up with the influence of the number of inquiries at Message-layer, a more complex and practical 2-layer queuing model with variable rates is derived. The analytical formulas were proved by a simulation experiment, in which the 3 key metrics calculated by our model can be approximately matched with simulation results.

1. Background
With the rapid growth of interactive online service for business transaction, online customer service (OCS) is gradually replacing the traditional telephone service as the mainstream instant service.

From the enterprise perspective, an OCS offers more convenience like high efficiency and less human resource, rich contents and diversified forms of communication, pre-routing and pertinent services, flexible communication, low environmental requirements and strong expansibility. Comprehensively, it has become a favored way of communicating in many situations because of the advantages showed in the context.

The quality of service is an important determinant of the success of OCS, and better service quality helps overcome some of the obstacles in providing effective chat service. Agents can set up the service speed voluntarily based on how busy they are, even deal with other work while serving. The specific service rate \( \mu_i \) is determined by \( m_i \). Here, let \( m_i \) denote the decision strategy of variable service rate point. In this paper, considering the effects of multitasking, we comprehensively formulate the key metrics of OCS and put forward a 2-layer queuing model for Online Customer Service with variable service rates.

![Figure 1. Variable service rate of OCS.](image_url)
2. Literature Review

The simplest and most widely used queuing model is the M/M/s queue, which is known as Erlang C as well. Based on the Erlang C, Garnet (2002) supplemented with the exponential distribution of customer patience and the busy signals M/M/N+M model, namely Erlang-A model [1]. Aktekin et al (2012) used the Markov Chain Monte Carlo method (MCMC) to perform Bayesian analysis on the Erlang-A queuing model [2]. Zeltyn and Mandelbaum (2005) summarized the Internet supplement and Baccelli-Hebuterne’s results on the M/M/n+G queue [3]. Janssen et al (2011) refined the number of CSR to ensure the stable operation of the system by extending the Erlang-C model and verified its effectiveness [4]. Galit et al (2014) analyzed the Erlang-R time-varying queuing model, which considers the reentry into the system [5]. Ding et al (2015) points out that the combination of redials and reconnects has not been considered when making staffing decisions [6]. Through setting threshold N customers in queuing system, Cheng and Wang (2005) consider the variable service rates [7]. Whitt (2008) showed that fluid models can be quite useful in approximating the performance measures of multi-server queues [8]. When it extends to the situation where there are multiple customers classes and a single server pool, Atar et al (2010,2011) proved the effectiveness of the fluid approximations in designing optimal control policies [9,10].

Similar to our paper, Luo and Zhang (2013) considered an OCS, in which one agent can serve multiple customers in parallel [11]. They choose to model such a center as a server pool consisting of many limited processor sharing (LPS) servers. Then characterize the underlying stochastic processes by establishing a fluid approximation in the many-server heavy-traffic regime. Tezcan and Zhang (2014) determine the efficient multitasking conditions, and calculated the staffing problem with impatient customers where the object is to minimize the number of the agents while providing a service level [12].

The relevant literature mainly focused on the queuing model in the one-to-one service mode of traditional call center or multitasking without variable service rates. While in reality, the service rate usually changes with the workforce. Which can not only reduce the waiting time, but also improve the utilization of the agents. The changeable service rates make the OCS manager adjust the service rates to achieve the optimal situation in the meanwhile. To balance the costs and the service quality, the decision strategy of when to change the rate is of great significance and necessary to be considered.

3. Model Construction

3.1. Model Assumptions

With reference to the related literature and considering the actual needs of this model, we put forward the following assumptions about parameters which is relevant to 2-layer queuing model with variable rates:

**Customer side:**
1) Customers are patient, no one will leave the system after entering;
2) Customer arrival is a Poisson process with a parameter of $\lambda$, whose interval time obeys the exponential distribution of the mean value of $1/\lambda$;
3) The time required for the customer to send each advisory message follows an exponential distribution of the mean of $\theta$;
4) The probability of the $i^{th}$ customer consulting is $p(i)$, then the average number of expected inquiries is $E(I) = \sum_{i} i^2 p(i) = A$;

**OCS side** (Let $m_0$ denote the decision strategy of variable service rates):
5) When the amount of inquiries to be solved is small (less than $m_0$), the time required for handling each inquiry is subject to the exponential distribution of the mean value of $\mu_1$;
6) Larger inquiries (larger than $m_0$) lead to a larger mean value of $\mu_2$ which also obeys the exponential distribution;
7) To ensure the quality of the service, we set up the biggest reception number for $m$ at the same time. People will wait in the queue when the total number of customers in the system is more than $m$. The model considers the case with no abandonments of customer.

8) The inquiry and reply between customers and CSRs are strictly "one to one" interaction process. Which means only when one receives the message will he/she continue to reply.

3.2. Model framework
In the steady state, the arrival follows a certain Poisson distribution. The customers ask the CSR for help as soon as they enter the system (with mean $\theta$ of exponential distribution). While the CSR needs to take time to provide the solution (with mean $\mu_1$ or $\mu_2$ of exponential distribution). Only when all the problems are solved will they leave.

3 key metrics of OCS queuing model: The same as Erlang queuing model, the online customer service queuing model also examines queuing performance indexes including the average waiting time of customers, the average total service time and the average length of queue. Moreover, the capacity of CSR needs to be considered. The customers who are beyond the capacity of CSR but less than the capacity of the system will wait in the queue.

The OCS queuing model was divided into Message-layer queuing model and Customer-layer queuing model. Furthermore, we take the variable service rate into consideration. Since the agent can control the service rate within their abilities, it is possible to properly improve the service rate when the inquiries are large. These flexibilities together with the simultaneous treatment are the main specificities of the OCS queuing model. Figure 2 is the illustration of the OCS 2-layer queuing model with service rates.

![Figure 2. 2-layer queuing model with variable rates.](image)

When the volume of inquiry is small, it means CSRs are free enough to finish the work. Hence, in this case, online CSRs can service at a lower service rate, other relevant work (like customer data analysis, customer relationship management...) can be arranged as well.

When it is larger but still within the service capacity, to fully guarantee the satisfaction of each customer and minimize the average waiting time for customers, the agents concentrate on the customers’ online inquiry to work at a higher service rate.

When the inquiries are out of our control, no matter how much the service rate is, customers’ inquiry still can't be satisfied. At this time, instead of accepting the entry and speeding up, customers who are beyond the capacity to wait until the capacity of the agent is met in avoid of low service quality.

3.2.1. Message-layer queuing model. Because of the service capacity, the amount of message to be replied is $[0, m]$. We use $P_i$ to denote the probability of state $i$ ($i=0,1,2,\ldots,m$) in the Message-layer queuing model. Assume that at some point, the number of messages to be answered by the customer is $c$, and the time required by the customer to send the message follows an exponential distribution of the mean of $\theta$. As a result, the frequency of new messages sent by the remaining customers is $(m-c)/\theta$. When the number of messages to be answered is more than a certain value $m_\alpha$, to avoid longer customer waiting time, the customer service reply speed is appropriately promoted. In detail, a
state of the Message-layer model is defined by the number of inquiries \( c \), \( m \geq c \geq 0 \). Then the transition rate from state \( c \) to \( c+1 \) is \( \frac{m-c}{\theta} \). While \( c \leq m_0 \), the transition rate from \( c \) to \( c-1 \) is \( \frac{1}{\mu} \). Otherwise, the transition rate is \( \frac{1}{\mu} \). The Markov chain of the model is depicted in Figure 3.

![Figure 3. Markov chain of Message-layer model.](image)

Now, we obtain the following equations of the Message-layer queuing model:

\[
\begin{align*}
mp_0/\theta &= p_1/\mu, \\
(m-1)p_1/\theta &= p_2/\mu, \\
&\vdots \\
(m-m_0+1)p_{m_0-1}/\theta &= p_{m_0}/\mu, \\
(m-m_0)p_{m_0}/\theta &= p_{m_0}/\mu, \\
&\vdots \\
p_{m_0}/\theta &= p_{m_0}/\mu.
\end{align*}
\]

It can be concluded by regularity that:

\[
\sum_{z=m_0}^{\infty} P_z = \sum_{z=m_0}^{\infty} P_{z} = \sum_{z=0}^{m} P_z = 1, \quad \text{in which} \quad \rho_1 = \frac{\mu_1}{\theta}, \quad \rho_2 = \frac{\mu_2}{\theta}, \quad \text{and}
\]

\[
P_0 = \left( \sum_{z=0}^{m_0} \frac{m!}{(m-i)! \rho_1^i \rho_2^{m_0-i}} + \sum_{i=m_0+1}^{\infty} \frac{m!}{(m-i)! \rho_1^{m_0} \rho_2^{i-m_0}} \right)^{-1}
\]

The average service time in Message-layer is:

\[
\mu = \sum_{z=m_0}^{\infty} P_z \mu_z + \sum_{z=0}^{m_0} P_z \mu_z
\]

Let \( c \) represent for the number of messages waiting for handling and \( m-c \) haven’t sent yet, therefore the effective arrival rate of new messages is:

\[
\lambda_{Message}^{Message} = (m-c)/\theta = P_{c-1}/\mu \frac{1}{\mu} = (1-P_0)/\mu
\]

The average number of inquiries wait to be replied is the part except for the messages being replied:

\[
l_{q_{Message}}^{Message} = c - (1-P_0) = m - (1-P_0)(1+\theta/\mu)
\]

Based on the Little’s and formula (4)(5), the average waiting time of each message including the time agents reply may be written as

\[
W(m) = l_{q_{Message}}^{Message}/\lambda_{Message}^{Message} + \mu = \mu l_{q_{Message}}^{Message} / (1-P_0) \frac{\mu m}{1-P_0} - \theta
\]

With \( W(m) \), the total average stay time per customer in the system (including the time customers send messages): \( T(m) = A \ast (\theta + W(m)) \) if the system has been able to ensure that there are just \( m \) customers. In the case, where \( A \), \( \mu \), and \( \theta \) are both known, then \( T \) and \( W \) are functions of \( m \).

3.2.2. Customer-layer queuing model. The number of customers in the system is constantly changing with the arrival of customers and the departure of the inquiry. The more customers, the faster CSR will serve. It’s concluded that the queuing model of Message-layer will affect that of Customer-layer.
Assuming that, the arrival rate is $\lambda$, and the capacity of the system is $N$ (Here $N$ is infinite, which means no one will abandon neither in the service nor in the queue.). Therefore, when the number of customers in the system is less than capacity $m$ (Here $m = N_0$), the service rate is $m/T(m)$ with the continuous change of customer. While it is more than $m$, the rate is constant $\{N_i/T(N_i)\}$. That is, a state of the Customer-layer queuing model is defined by the number of customers $N_i$, $N_i \geq 0$. Then the transition rate from state $N_i$ to $N_i+1$ is $\lambda$. While $N_i < N_0$ (or $m$), the transition rate from $N_i$ to $N_i-1$ is $m/T(m)$. Otherwise, the transition rate is $\{N_i/T(N_i)\}$. The specific transfer diagram is illustrated in Figure 4:

![Markov chain of Customer-layer model](image)

**Figure 4.** Markov chain of Customer-layer model.

Formally, the queuing model of Customer-layer can be formulated as follows: Let $PP_i$ denote the probability of state $i$ ($i=0,1,2,...,N$), given $\lambda$ and $m$ (or $N_0$),

$$
\lambda PP_i = PP_{i-1} \cdot \frac{1}{T(1)} \\
\lambda PP_i = PP_{i-1} \cdot \frac{2}{T(2)} \\
\lambda PP_i = PP_{m-1} \cdot \frac{m}{T(m)} \\
\lambda PP_i = PP_{N_0} \cdot \frac{N_0}{T(N_0)} \\
\lambda PP_i = PP_{N_i} \cdot \frac{N_i}{T(N_i)} \\
\lambda PP_i = PP_{N_{i+1}} \cdot \frac{N_i}{T(N_i)}
$$

Further induction $PP_i = \left\{ \begin{array}{ll}
\frac{\lambda}{i!} \prod_{d=1}^{i} T(d) & 1 \leq i \leq N_0 - 1 \\
q^{i-N_0} \ast PP_{N_0}, N_i \leq i \leq N
\end{array} \right.$ (7)

Let $q = \lambda T(N_0) / N_0$, in order to ensure that the system exists steady state, $q$ needs to meet $q < 1$, which is equivalent to $\lambda < N_0/T(N_0)$, and the regularity of the whole process decides that

$$
\sum_{i=0}^{\infty} PP_i = \sum_{i=0}^{N_0} PP_i - \sum_{i=0}^{N_0} PP_i = 1, \text{ we have:}
$$

$$
\sum_{i=0}^{\infty} PP_{N_i} = \frac{1 - q^{N_0-N_i+1}}{1-q} PP_{N_0} - \frac{1 - q^{N_0-N_i+1}}{1-q} PP_{N_i} = \frac{\lambda^{N_0}}{N_0!} \prod_{i=0}^{N_0} T(i) \left(1 - q^{N_0-N_i+1}\right) \prod_{i=0}^{N_0} T(i)
$$

$$
PP_i = \left\{ \begin{array}{ll}
\frac{\lambda^{N_0}}{i!} \prod_{d=1}^{i} T(d) & \lambda^{N_0} \left(1 - q^{N_0-N_i+1}\right) \prod_{i=0}^{N_0} T(i)
\end{array} \right. , \text{ with}
$$

$$
PP_0 = \left\{ \begin{array}{ll}
\frac{\lambda^{N_0}}{i!} \prod_{d=1}^{i} T(d) & \lambda^{N_0} \left(1 - q^{N_0-N_i+1}\right) \prod_{i=0}^{N_0} T(i)
\end{array} \right. , \text{ with}
$$

To sum up, when the number of customers in the system is constantly changing, the total average stay time is (in the 2-layer queuing model $m = N_0$):

$$
5
\[ T = \sum_{i=0}^{N_0} PP_i * T(i) + \sum_{i=N_0+1}^{N} PP_i * T(N_0) \]  \hspace{1cm} (9)

The average length of queue in the system, denoted by \( l_q \), may be written as:

\[ l_q = \sum_{i=0}^{N_0} i*PP_i + N_0* \sum_{i=N_0+1}^{N} PP_i \]  \hspace{1cm} (10)

The average customer waiting time for the reply in the system, defined as \( W_q \):

\[ W_q = \sum_{i=0}^{N_0} PP_i*W(i) + \sum_{i=N_0+1}^{N} PP_i*W(N_0) \]  \hspace{1cm} (11)

4. Numerical Experiments

In this section, the results of numerical experiments with a given \( m_0 \) is present to illustrate the accuracy of the model we proposed. The queuing model was solved by Lingo11.0 software, and the simulation was carried out by Flexsim7.5.4. The main purpose is to confirm our deductions of how OCS with variable service rates works and test the total average stay time.

We set the correlation coefficient \( \theta = 50s \), \( \mu_1 = 35s \), \( \mu_2 = 25s \), \( m_0 = 5 \). The capacity of CSR \( m = 10 \), and the capacity of the system is infinite, which means no one will abandon. The number of customer inquiry \( A = 3.09 \). The average time interval of customer arrival we test is \([80, 240]\) and the corresponding customer arrival rates are \([0.0042, 0.0125]\). After determining these coefficients, the model is further solved.

Under the circumstances, the results are present in the form of line chart. Firstly, we compare our model with 2-layer model with constant service rate. The line coloured blue and green represent refers to constant service rates \( \mu = 35 \) and \( 25 \) respectively. We next simulate the model with \( m = 10 \) and arrival rates ranging from 0.004 to 0.013, the theoretical value and simulation value are given in terms of the basic indicators of the queuing model.

**Figure 5. Comparison**

**Figure 6. Total time**

**Figure 7. Avg waiting length**

**Figure 8. Avg waiting time**
Figure 5 is the comparison of the queuing model. When the arrival rate is getting larger, our model has much shorter total waiting time, which can improve customer satisfaction effectively. The result shows that our queuing model is more applicable.

Figure 6, 7, 8 are the validation of the 2-layer queuing model. The analytical formulas were proved by a simulation experiment, in which the 3 critical metrics calculated by our model can be approximately matched with simulation results. As expected from our asymptotic results, the accuracy of the model is verified to some extent. Specifically, in Figure 7 and 8, when the arrival rate is low, the simulation value is slightly smaller than the theoretical value. The reason is that there will be a certain preheating time due to the low arrival rate. The Flexsim7.5.4 simulation software needs to take a time to warm up until the steady state, which results in the deviation.

5. Conclusion
Different from the relevant literature about OCS, we not only consider the customer behavior in the queuing model, but also research the processing of the request message, which makes the proposed queuing model more accurate.

The comparison shows that this model can reduce the waiting time effectively. By analyzing the results between the formula value and the simulation value, the accuracy of the analytical formula is verified. The model shows the hourly human cost of CSR as well, which provides an effective theoretical reference for the operation of online customer service platform of e-commerce.

Acknowledgments
This work was supported by NSFC No. 71872037.

References
[1] O. Garnet, A. Mandelbaum, and M. Reiman. Designing a Call Center with Impatient Customers. Manufacturing & Service Operations Management 4.3(2002):208-227.
[2] Aktekin, Tevfik, and R. Soyer. Bayesian analysis of queues with impatient customers: Applications to call centers. Naval Research Logistics 59.6(2012):441-456.
[3] Zeltyn, Sergey, and A. Mandelbaum. Call Centers with Impatient Customers: Many-Server Asymptotics of the M/M/ n, + G Queue. Queueing Systems 51.3-4(2005):361-402.
[4] Janssen, A. J. E. M., J. S. H. V. Leeuwaarden, and B. Zwart. Refining Square-Root Safety Staffing by Expanding Erlang C. INFORMS, 2011.
[5] Yom-Tov, G. B., & Mandelbaum, A. Erlang-r: a time-varying queue with reentrant customers, in support of healthcare staffing. Manufacturing & Service Operations Management, 16.2(2014), 283-299.
[6] Ding, S., et al. Fluid approximation of a call center model with redials and reconnects. Performance Evaluation 92.C(2015):24-39.
[7] Cheng, Feng, and J. F. Wang. The repairable M/G(M/G)/1 queueing system with two service speed and with exhaustive serve and multiple vacations. Chinese Journal of Engineering Mathematics 22.2(2005):249-254.
[8] Whitt, Ward. Fluid Models for Multiserver Queues with Abandonments. Operations Research 54.54(2008):37-54.
[9] Atar, R., Giat, C., & Shimkin, N. The cμ/0 Rule for Many-Server Queues with abandonment. Operations Research, 58.5(2010), 1427-1439.
[10] Atar, R., C. Giat, and N. Shimkin. On the asymptotic optimality of the cμ/0 rule under ergodic cost. Queueing Systems, 67.2(2011), 127-144.
[11] Luo, Jun, and J. Zhang. Staffing and Control of Instant Messaging Contact Centers. Operations Research 61.2(2013):328-343.
[12] Tezcan, Tolga, and J. Zhang. Routing and Staffing in Customer Service Chat Systems with Impatient Customers. INFORMS, 62.4(2014):943-956.