## Title

Queueing Systems with Customer Abandonments and Retrials

## Author

DENG, SONG

## Publication Date

2013

# Queueing Systems with Customer Abandonments and Retrials

by

Song Deng

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering - Industrial Engineering & Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Assistant Professor Ying-Ju Chen , Co-chair
Professor Zuo-Jun “Max” Shen , Co-chair
Associate Professor Haiyan Huang

Spring 2013

**Queueing Systems with Customer Abandonments and Retrials**

## Abstract

Queueing Systems with Customer Abandonments and Retrials

by

Song Deng

Doctor of Philosophy in Engineering - Industrial Engineering & Operations Research

University of California, Berkeley

Assistant Professor Ying-Ju Chen , Co-chair

Professor Zuo-Jun "Max" Shen , Co-chair

In queueing theory, the phenomenon that customers get impatient and renege from the system when the waiting time exceeds their tolerance is called customer abandonments. For those abandoning customers, they may go back to the service system after some time. This is called customer retrials. In the modern design of service systems, the impact of customer abandonments and retrials on the system performance and queueing dynamics has been realized. In this thesis, we aim to characterize the stationary properties for queueing systems with both customer abandonments and retrials. We also apply some of those properties in a revenue management problem for queueing systems with customer abandonments.

First, we study a multi-server queueing system with both customer abandonments and retrials. By using RTA (retrials see time average) approximation, we characterize the stationary properties of such a queueing system. We also justify the appropriateness of RTA approximation in this model both theoretically and numerically.

Then, we extend the study to incorporate negative customers and unreliable servers. Two different models are considered. One is a multi-server retrial queueing system with both regular customers and negative customers. The other is a single-server retrial queue with negative customers and server interruptions. Closed-form expressions of some relevant performance measurements for both models are obtained.

For the last, we apply some of our findings in queues with abandonments and retrials to a revenue management problem for a queueing system with customer abandonments. By using a mechanism design framework, we identify the optimal price-lead time manus and scheduling policies for different scenarios.

To my parents and grandparents

# Contents

**Contents** **ii**

**1 Introduction** **1**

**2 Performance Analysis for Queues with Customer Abandonments and Retrials** **3**

2.1 Introduction . . . 3

2.2 Literature review . . . 5

2.3 Retrial queue analysis using RTA approximation . . . 6

2.4 Extensions to multi-class customers models . . . 15

2.5 Conclusion . . . 20

**3 Retrial Queueing System with Negative Customers and Unreliable Servers** **22**

3.1 Introduction . . . 22

3.2 Literature review . . . 23

3.3 Retrial queueing system with negative customer . . . 25

3.4 Retrial queueing system with negative customers and unreliable servers . . . 28

3.5 Conclusion . . . 31

**4 Revenue Management in Queue Systems With Customer Renege** **32**

4.1 Introduction . . . 32

4.2 Literature review . . . 33

4.3 Model description . . . 34

4.4 Structural results . . . 37

4.5 Summary . . . 39

**A Proofs** **40**

**Bibliography** **45**

## Acknowledgments

My deepest gratitude goes first to my advisors, Professor Ying-Ju Chen and Professor Zuo-Jun Max Shen, for their constant guidance and support. Their rigorous academic attitude and spirit of adventure benefit me a lot. Without their help, I would not be able to finish my Ph.D.

I would also like to thank my thesis committee member, Professor Haiyan Huang, who does excellent research in her field. It is a great pleasure for me to have her in my thesis committee and get her inspiring advice.

During my time in Berkeley, I got a lot of help from other graduate students in IEOR Department. We went to seminars, discussed research problems and had much fun together. It is them who make my life here so colorful. The time I spent with them will be a great memory for me.

For the last but not the least, I want to express my special thanks to my parents and grandparents, for everything they give me. Although I am away from home, I can always feel the support from them. Their support is the most important source of power in my research.

# Chapter 1

# Introduction

In the modern design of call centers, a crucial issue is to investigate the impact of customer abandonments and retrials on the queueing dynamics and system performance. In this thesis, we first study a multi-server queueing system with both customer abandonments and retrials. By using RTA (retrials see time average) approximation, we characterize the stationary properties of such a queueing system. In order to show the effectiveness of RTA approximation in our model, we analyze the difference between the average customer waiting time for the original queueing system and that for the model based on RTA approximation. Also, our numerical experiment and sensitivity analysis indicate that the model with this RTA approximation preserve all the structural properties of the original model, thereby justifying the appropriateness of our approach. Besides the case with homogeneous customers, we further incorporate the possibilities of heterogeneous abandonment rates or heterogeneous retrial rates; accordingly, customers may be given different priorities and effectively form different queues. Invoking the level crossing theorem, we obtain the detailed distribution for the virtual waiting time, based on which we derive some relevant performance measurements of these multi-class queueing systems.

In the second part of this thesis, we extend the study to incorporate negative customers and unreliable servers. we first study a multi-server queueing system with both regular customers and negative customers (which are considered as deletion signals). Stationary properties of such a queueing system is characterized by RTA approximation. We further extend the model to incorporate unreliable servers, where two kinds of server interruptions (server disasters and server vacations) are considered together. Using the methodology of stochastic decomposition, we obtain a closed-form expression for some relevant performance measurements by solving differential equations.

Nowadays, more and more attention is being given to the field of revenue management in queueing systems. In the third part of this thesis, we consider the revenue management problem in a $M/M/1/\infty$ queueing system with two types of customers with different valuation of service and different unit waiting cost. We also incorporate customer abandonments into consideration. By using a mechanism design framework and adapting the achievable region approach, we transform the scheduling control problem into a nonlinear program.

Then through a two-stage optimization technique, we identify the optimal price-lead time manus and scheduling policies for different scenarios. The optimality condition for strategic delay is also discussed.

The rest of this thesis is organized as follows. In Chapter 2, performance for queues with customer abandonments and retrials is analyzed. Our study in retrial queueing system with negative customers and unreliable servers is mentioned in Chapter 3. In Chapter 4, we continue our study to the revenue management problem for queueing systems with customer abandonments. Proofs of Lemmas and Propositions are included in Appendix.

# Chapter 2

# Performance Analysis for Queues with Customer Abandonments and Retrials

## 2.1 Introduction

In classical queueing theory, customers' patience is usually not taken into consideration. Even for cases when the assumption of infinite customers' patience is relaxed, it is also assumed that an impatient customer leaves the system without having the option to go back for service. Actually, the assumption about loss of customers who choose to leave the system is just a first-order approximation to the real situation. Usually such a customer returns to the system and tries to get service again after a random time. Real world examples can be easily found for returning customers. Take call centers as an example, incoming customers who find all customer representatives busy will hang up immediately or hold on to wait for the next available service slot. For those customers who choose to wait, when the waiting time exceeds their tolerance, they get impatient and renege from the system. This phenomenon is called *customer abandonment*. For those abandoning customers, they may leave the system forever, or go back to the service system after some time, which is called *customer retrial*. Similar behaviors may also be found for customers in the shopping mall. Customers who find a long waiting line may wish to do something else and return later on with the hope that the queue dissolves.

The importance of customers' abandonments and retrials is also discussed in some classic papers. Falin and Templeton (1997) emphasize that the standard queueing model that does not take retrial phenomenon into account cannot be applied in solving numbers of practically important problems. Furthermore, the impact of customers' abandonments and retrials on queueing systems performance can be found in many papers, such as Aguir et al. (2004), Wuchner et al. (2008), and etc. Nevertheless, most of previous research considers customer abandonments or retrials separately. To the best of our knowledge, only Mandelbaum et al. (2002) deal with queueing systems with both abandonments and retrials, where fluid model approximation is used. Consequently, only a few performance measurements are obtained, such as the average waiting time and the average queue length. For this chapter, we attempt to get a full characterization of dynamics for the queueing system with both customer abandonments and retrials.

In this chapter, we first study multi-server systems with customers who may renege from/abandon the system if their waiting time exceeds their patience limit, i.e., the maximum time they would be willing to wait for service. Also, abandoning customers may choose to rejoin the service system after some time. Our model consists of two nodes: a service node with several servers where customers wait to be served and a retrial pool (orbit) with infinite servers where customers wait to go back to the service node. Different from previous papers, we characterize the stationary behavior of the service node. To tackle this problem, we invoke the RTA (retrials see time average) approximation to separate the analysis of these two nodes, which subsequently allows us to get the performance measurements for the entire queueing system. The detailed description of this technique is given in Section 2.3. We verify the effectiveness of our algorithm both analytically and numerically. Also, we find that the modified model based on RTA approximation preserves all structural properties of the original model.

We also extend our study to the retrial queue with multi-class customers. In the real life, besides customers abandonments and retrials, we can find customers with different abandonment and retrial behaviors. For instance, emergency patients for whom reneging might refer to their deaths will have lower patience than non-emergency patients . Under these circumstances, the health-care management will most likely choose to give priority to the emergency patients over the others. Similarly, a customer contact center might offer an e-mailing option to its customers as an alternative medium to be reached. Customers may request in their e-mails to be called back or the type of service requested might be handled more easily by a customer representative calling them. Therefore, after serving high-priority impatient customers who place phone-calls, customer representatives can call the customers who have sent e-mail requests.

For our study in the multi-class customers model, two scenarios are considered. In the first scenario, we consider two types of customers who have different abandonment rates, where one type of customers have infinite patience and the others have finite patience. In this scenario, we assign the high priority to customers with finite patience. Consequently, two separate queues are introduced. We use *level crossing* techniques to get the density function for the virtual waiting time for customers in the queue with priority. Further, other performance measurements of the priority queue are obtained, such as average waiting time for customers being served, mean number of customers in the priority queue, and so on. The second scenario involves two type of customers with different retrial rates. In this scenario, we extend RTA approximation to the multi-types case. By using such an approximation, we separate the analysis of the service node from the orbit and get the stationary properties of the whole queueing system. With the detailed queueing length distribution and waiting time distribution obtained in our analysis, we can get the system performance evaluations under different settings, i.e., different numbers of servers or different priority rules. Further, we can use these results in the design of the queueing system.

The rest of this chapter is organized as follows. In Section 2.2, we review previous research papers briefly. In Section 2.3, we extend RTA approximation to our queueing model with both abandonments and retrials. We also prove the effectiveness of such extension both analytically and numerically. In Section 2.4, we continue our study to the case with multi-class customers, where two scenarios are considered. We summarize our findings in Section 2.5.

## 2.2 Literature review

Over the past couple of decades, there are many papers dealing with the multi-server model with customer abandonments. Movaghar (1998) derived the steady-state distribution of the number of customers in the system and the distribution of waiting time of M(n)/M/s queue with customer abandonments. Brandt and Brandt (2002) analyze the queueing model with state dependent exponential arrival and service rates as well as general abandonment rate. They derive equations for the state probabilities and waiting time distribution. For large scale queueing system, several performance measures in the limit as the arrival rates and servers go to infinity, are derived by Zeltyn and Mandelbaum (2005). In addition, Whitt (2006) performs the sensitivity of Erlang-A queueing model, demonstrating that performance measures in that model are sensitive to small changes in the arrival and service rates, but relatively insensitive to small changes in the abandonment rate.

In previous queueing literature, customer retrial has also been well studied. For the multi-server retrial model, explicit formulae for the main performance characteristics (stationary distribution, blocking probability and mean queue length) are absent when the number of servers is more than two, see Artalejo and Gómez-Corral (2008). Consequently, different approximation methods are introduced, such as direct truncation model, generalized truncation model, RTA (Retrial See time Average) approximation, and etc. For instance, Falin (1983) assumes that the retrial rate becomes infinite when the number of customers in orbit exceeds a level M. It means that, from the level M up, the system performs as an ordinary M/M/1 queue. Neuts and Rao (1990) develop a tractable approximation, which yields an infinitesimal generator that is essentially a quasi-birth-and-death (QBD) process with a large number of boundary states. Recently, Artalejo et al. (2005) deal with a multi-server retrial queueing model in which the number of active servers depends on the number of customers in the system. For a fixed choice of the threshold levels, the stationary distribution and various performance measures of the system are calculated. The optimum threshold level is also numerically computed in the case of equidistant connection levels. For detailed overviews of the main results and the bibliographical information about the retrial queues, see Artalejo (1999) and Artalejo (2010).

For the multi-class customers model, incorporating priority rule among impatient customer classes brings several challenges. Usually, analytical models are confined to problems with two classes of customers. For instance, Choi et al. (2001) analyze the M/M/1 queue with two classes of customers where class 1 customers have deterministic impatience time while class 2 customers are assumed to be infinitely patient. With class 1 customers having preemptive priority over class 2 customers, they obtain the joint distribution of the number of class 1 and class 2 customers in the system and the distribution of the total response time of class 2 customers. Subsequently, Brandt and Brandt (2004) extend the two class impatient customers model to a setting, where they allow the distribution of the high-priority customers abandonment time to be general. Also, high priority customers are given preemptive-resume priority over the other class in that model. By solving the balance equations for the partial probability generating functions of the detailed system state process, the joint queue length distribution and the Laplace transforms of the waiting and sojourn time for low-priority customers are derived. Recently, Iravani and Balcıoglu (2008) study a multi-server queueing model with two classes of customers, one has exponential distributed tolerance time and the other has infinite patience. Priority rule is also incorporated in that model, with the impatient customers having non-preemptive priority over the other. By using the level crossing method, they obtain the steady-state performance measures of the high-priority class. Factorial moments of the low-priority queue length of the multi-server system are also obtained.

In the field of multi-class queueing systems with customers' abandonments, simulation usually provides a good approach for complicated models. For instance, Ahghari and Balcioglu (2006) conduct an extensive simulation study on the customer contact center that provide different types of services to customers who place phone calls or send e-mail messages, with phone calls having priority over e-mail request. They show that strategies permitting pre-emptive-resume policies provide the best performance for phone calls. Among all the papers dealing with multi-class customers, with or without the consideration of priority rule, customer abandonments or retrials are omitted. In contrast, we consider a multi-class queueing model together with customer abandonments and retrials.

## 2.3 Retrial queue analysis using RTA approximation

In this section, we consider a multi-server queue where waiting customers may abandon and subsequently re-enter the service. We will get the stationary behavior of the above queueing system under appropriate approximation.

### Model description

Our model, depicted in the Figure 2.1, consists of two types of nodes: a service node with $c$ servers, and a retrial pool (orbit) with infinite servers, where customers effectively serve themselves. New customers arrive at the service node as a Poisson process with rate $\lambda$. The service node has a maximum capacity $N$. When there are already $N$ customers in the service node, new arrivals will be blocked. Blocked customers will leave the system.

In this section, we focus on the case with homogeneous customers, leaving the discussion on multi-type case to the next section. Arriving customers who find an idle server are taken into service with a duration which is exponentially distributed with rate $\nu$. Customers who find all servers busy join a queue and are served in a First-Come-First-Serve (FCFS) manner. Each waiting customer has the same maximum tolerance time which is exponentially distributed with rate $\gamma$; equivalently, each customer waiting in queue abandons at rate $\gamma$. An abandoning customer leaves the system forever with probability $1-\beta$ or joins the retrial pool with probability $\beta$. Customers in the retrial pool leave and subsequently enter the service node at rate $\mu$. Consequently, when there are $i$ customers in the orbit, the total retrial rate is $i\mu$. After customers complete their retrials, they are treated as new customers; thus, new customers and retrial customers are not distinguished. In our model, we assume the inter-arrival time, service time, inter-abandonment time and inter-retrial time to be exponential. The memoryless property brought by the exponential distribution facilitates our analysis, making the whole problem treatable. Exponential assumption is also commonly used in queueing literature, such as Iravani et al. (2008), Krishna et al. (2008), and etc.

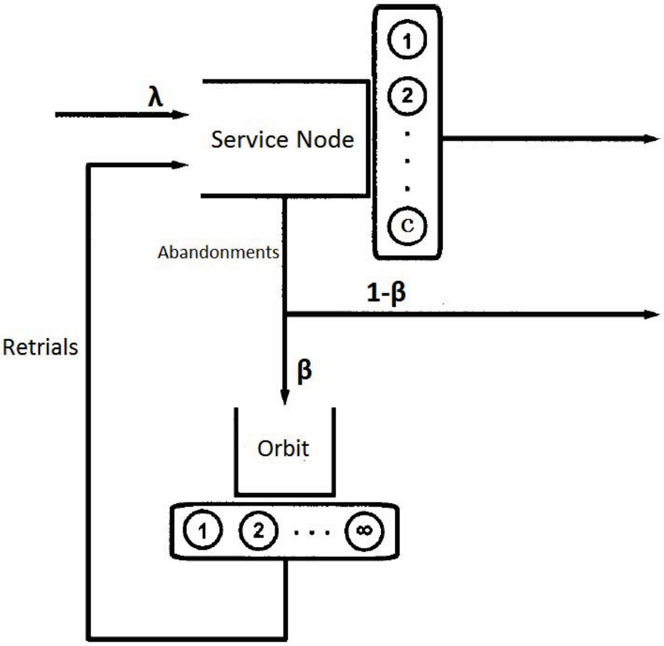

Figure 2.1: The multi-server queue with abandonments and retrials.

In order to describe the system, we need parameters shown in the following table.

| | |
|---|---|
| $\lambda$ | Arrival rate to the service node |
| $N$ | Maximum number of customers allowed in the service node |
| $\nu$ | Service rate of each service channel |
| $c$ | Total number of servers |
| $\gamma$ | Abandonment rate for customers in the main queue |
| $\beta$ | Probability that a customer joins the retrial orbit after abandonment |
| $\mu$ | Retrial rate for customers in the orbit |

## Stationary distribution analysis using RTA approximation

For the queueing system mentioned in the previous subsection, its state at time $t$ can be described as $\zeta(t) = \{C(t), M(t)\}$, where $C(t)$ is the number of customers in the main queue and $M(t)$ is the number of customers in the orbit. Obviously, $\{\zeta(t)|t \geq 0\}$ is a homogeneous continuous-time Markov chain with state space $\mathbb{E} = \{0, 1, 2...N\} \times \mathbb{N}$. Let $P_{i,j}$ be the stationary probability of state $(i, j)$, the balance equations for this Markov chain are as follows:

For $0 \leq i \leq c$:

$$P_{i,j}(\lambda + i\nu + j\mu) = \lambda P_{i-1,j} + (j+1)\mu P_{i-1,j+1} + (i+1)\nu P_{i+1,j};$$

for $c < i < N$:

$$\begin{aligned} P_{i,j}(\lambda + c\nu + j\mu + (i-c)\gamma) = {} & \lambda P_{i-1,j} + (j+1)\mu P_{i-1,j+1} + c\nu P_{i+1,j} + \\ & (i+1-c)\beta\gamma P_{i+1,j-1} + (i+1-c)(1-\beta)\gamma P_{i+1,j}; \end{aligned}$$

for $i = N$:

$$P_{i,j}(c\nu + (i-c)\gamma) = \lambda P_{i-1,j} + (j+1)\mu P_{i-1,j+1} + (j+1)\mu P_{i,j+1}.$$

Although the solution for the above equations will lead us to the stationary probabilities, in most cases characterizing the solution turns out to be a difficult task. Artalejo and Gómez-Corral (2008) indicate that when $c > 2$, we cannot find the explicit expression for $P_{i,j}$. Consequently, some approximations are needed to obtain $P_{i,j}$. We will use RTA (retrials see time average) approach. The retrials see time averages (RTA) approximation is similar to the well-known PASTA property. Roughly speaking, PASTA says that if the arrival process is Poisson, then the steady state probability that an arriving customer sees the system in state $i$ is the same as the limiting probability that the system is in state $i$. Applying this property into the retrial queueing system, we can expect that when the system is in steady state, the proportion of customers who see $i$ customers in the service node is the same as its limiting probability. This is the essence of RTA approximation. For previous papers dealing with retrial queues, RTA approximation has never been used in the model with both customers' abandonments and retrials. Here we will extend RTA approximation to such setting.

First, we need to make some additional notation. Define:

$$\nu_i = \min(i,c)\nu, \text{ and } \gamma_i = (i-c)^+\gamma.$$

Then, summing the above Kolmogorov equations over $j$, we get:

$$(L_{i-1}\mu + \lambda)P_{i-1} = (\nu_i + \gamma_i)P_i, \tag{2.1}$$

where $P_i = \sum_{j=0}^{\infty} P_{i,j}$ is the probability that there are $i$ customers in the service node, and $L_i = \sum_{j=0}^{\infty} jP_{i,j}$ is the expected number of customers in the orbit when there are $i$ customers in the service node. Based on RTA approximation, we have the following equation.

$$L_i = L, \, \forall i,$$

where $L$ is chosen in $R^+$. Considering the service node separately, we have the following equations:

$$(L\mu + \lambda)P_{i-1} = (\nu_i + \gamma_i)P_i, \text{ and } \sum_{i=0}^{N} P_i = 1.$$

Solving the above equations, we obtain the stationary distribution for the service node as follows:

$$P_0 = (1 + \sum_{i=1}^{N}(\prod_{j=1}^{i}\frac{L\mu+\lambda}{\nu_i+\gamma_i}))^{-1}, \tag{2.2}$$

$$P_i = \prod_{j=1}^{i}\frac{L\mu+\lambda}{\nu_i+\gamma_i}/(1 + \sum_{i=1}^{N}(\prod_{j=1}^{i}\frac{L\mu+\lambda}{\nu_i+\gamma_i})). \tag{2.3}$$

From the above expressions, we can see that $P_i$ is a function of $L$. In order get the value of $L$, we need to find another equation. By equating the flow rate into and the flow rate out of the subset $\mathbb{N} \times \{0, 1, 2..., j\} \subset E$ (recall that $E$ is the state space of the entire Markov Chain), we have:

$$(j+1)\mu\sum_{i=0}^{N} P_{i,j+1} = \beta\gamma\sum_{i=c}^{N}(i-c)P_{i,j}. \tag{2.4}$$

Summing (4) over $j$ and applying RTA assumption, we get:

$$\mu L = \beta\gamma\sum_{i=c}^{N}(i-c)P_i. \tag{2.5}$$

Plugging (2) and (3) to (5), it yields:

$$\mu L = \beta\gamma\sum_{i=c}^{N}(i-c)\prod_{j=1}^{i}\frac{L\mu+\lambda}{\nu_i+\gamma_i}/(1 + \sum_{i=1}^{N}(\prod_{j=1}^{i}\frac{L\mu+\lambda}{\nu_i+\gamma i})). \tag{2.6}$$

In order to make RTA approximation efficient for our model, we need a unique positive root when solving (6). Such uniqueness is established by the following lemma.

**Lemma 1.** *Equation (6) always has a unique positive root.*

Lemma 1 guarantees the usefulness of RTA approximation in this model. Now, with $L$, we are able to calculate some performance measurements for the queueing system.

**Proposition 2.** *The system performance can be characterized by means of the following quantities:*

*1.* The probability that all servers are idle:

$$P_0 = [1 + \sum_{i=1}^{N} (\prod_{j=1}^{i} \frac{L\mu + \lambda}{\nu_i + \gamma_i})]^{-1}. \tag{2.7}$$

*2.* The average number of customers in the service node*:*

$$C = \sum_{i=0}^{N} iP_i = \sum_{i=0}^{N} \frac{i \prod_{j=1}^{i} \frac{L\mu+\lambda}{\nu_i+\gamma_i}}{1 + \sum_{i=1}^{N} (\prod_{j=1}^{i} \frac{L\mu+\lambda}{\nu_i+\gamma_i})}. \tag{2.8}$$

*3.* The average waiting time for each customer in the main queue*:*

$$W = \frac{1}{L\mu + \lambda} \sum_{i=c}^{N} \frac{(i-c) \prod_{j=1}^{i} \frac{L\mu+\lambda}{\nu_i+\gamma_i}}{1 + \sum_{i=1}^{N} (\prod_{j=1}^{i} \frac{L\mu+\lambda}{\nu_i+\gamma_i})}. \tag{2.9}$$

*4.* The proportion of customers getting served*:*

$$P(s) = \frac{\nu}{\lambda} [\sum_{i=0}^{c} \frac{i \prod_{j=1}^{i} \frac{L\mu+\lambda}{\nu_i+\gamma_i}}{1 + \sum_{i=1}^{N} (\prod_{j=1}^{i} \frac{L\mu+\lambda}{\nu_i+\gamma_i})} + \sum_{i=c+1}^{N} \frac{c \prod_{j=1}^{i} \frac{L\mu+\lambda}{\nu_i+\gamma_i}}{1 + \sum_{i=1}^{N} (\prod_{j=1}^{i} \frac{L\mu+\lambda}{\nu_i+\gamma_i})} ]. \tag{2.10}$$

*5.* The proportion of customers getting blocked*:*

$$P(b) = P_N = \prod_{j=1}^{N} \frac{L\mu + \lambda}{\nu_i + \gamma_i} / (1 + \sum_{i=1}^{N} (\prod_{j=1}^{i} \frac{L\mu + \lambda}{\nu_i + \gamma_i}). \tag{2.11}$$

## Evaluation for the accuracy of RTA approximation

In the previous subsection, we obtain the stationary distribution for the main waiting queue by applying RTA approximation. Although the effectiveness of such approximation was proved for some other retrial queue settings in the previous literature, its usefulness is not guaranteed in our model. In order to check how this approximation works in our model, we need to compare the result obtained by RTA approximation with the actual system output, where we choose the average customer waiting time to make such a comparison. Nevertheless, the exact average customer waiting time for the actual system can not be calculated analytically. Alternatively, we find a pair of bounds for both the real system and the modified system by RTA approximation. By comparing these two bounds, we can see the difference between the result obtained by RTA approximation and the actual system output. Let **MO** denote the original model and **MR** denote the model using RTA approximation. First, we introduce two modified models which can serve as the upper bound and lower bound.

The first model is a M/M/c/N queue with customer abandonments only (**M1**). The setting of this model is the same as the original model (**MO**) except that we restrict $\beta = 0$ in **M1**. This means for those customers who choose to abandon, they will not join the orbit. All of them are forced to leave the system. In this model, since there is no orbit, we can focus our analysis on the main waiting queue. The balance equations are as follows:

$$\lambda P_i = (i+1)\nu P_{i+1} \text{ for } i < c,$$

$$\lambda P_i = (c\nu + (i+1-c)\gamma) P_{i+1} \text{ for } i \geq c,$$

$$\sum_{i=0}^{N} P_i = 1.$$

The second model is a M/M/c/N queue with an infinite retrial rate (**M2**). The setting of this model is the same as the original model (**MO**) except that we restrict $\mu = \infty$ in **M2**. This means for any customer who choose to retrial, he can get back to the service node immediately. In this model, due to the infinite retrial rate, the number of customers in the orbit is always zero. Therefore, we only pay attention to the main queue. In the absence of rearranged sequence due to retrials, the system can be considered as an M/M/c/N queue with customer abandonment rate $(1-\beta)\gamma$, the balance equation is quite similar to **M1**. Define $W_{M1}$, $W_{M2}$, $W_{MO}$ and $W_{MR}$ to be the average customer waiting time for the corresponding four models. We have the following Lemma.

**Lemma 3.** $|W_{MR} - W_{MO}| \leqslant W_{M2} - W_{M1}$.

To articulate the intuition behind Lemma 3, let us just take the service node into consideration. In this case, it is obvious that no matter how many customers in the service node and how many customers in the orbit, the customers' incoming rate to the service node in **M1** is less than **MO** and **MR**. Consequently, for the average customer waiting time, **M1** serves as the lower bound for both **MO** and **MR**. The situation for **M2** is just the opposite, this makes **M2** to be the upper bound for **MO** and **MR**.

Lemma 3 provides a way to estimate the difference of average customer waiting time between **MO** and **MR** by comparing the performance of **M1** and **M2**. From the definition of **M1** and **M2**, we know these are two Erlang-A queueing models with abandonment rate to be $\gamma$ and $(1-\beta)\gamma$. Whitt (2006) performs a complete sensitivity analysis of performance in the Erlang-A queueing model to changes in the model parameters. He shows that performance of the Erlang-A queueing model is insensitive to the change of abandonment rate. Consequently, the accuracy of RTA approximation in our model is quite impressive.

## Sensitivity analysis

After proving the effectiveness of RTA approximation in our model, we will explore how the system output changes with respect to the change of system parameters for both the original model and the model based on RTA approximation. We have the following proposition.

**Proposition 4.** *Suppose the average customer waiting times in the service node for the original model and the model based on the RTA approximation are $W_{MO}$ and $W_{MR}$, then $W_{MO}$ and $W_{MR}$ change according to the following properties:*

1. *$W_{MO}$ and $W_{MR}$ are increasing functions with respect to $\lambda$,*
2. *$W_{MO}$ and $W_{MR}$ are increasing functions with respect to $\beta$,*
3. *$W_{MO}$ and $W_{MR}$ are increasing functions with respect to $\mu$,*
4. *$W_{MO}$ and $W_{MR}$ are decreasing functions with respect to $\gamma$,*
5. *$W_{MO}$ and $W_{MR}$ are decreasing functions with respect to $\nu$.*

From the above proposition, we can see that the two models share the same monotonicity property as system parameters change, which shows the effectiveness of RTA approximation in this model once again.

## Numerical results

In Subsection 3.3, we justify the effectiveness of RTA approximation in our model analytically. In this subsection, we use simulation to get the the output of the actual system. Then we can make a numerical comparison between the result got by the model based on RTA approximation and actual system output. Here, two variables are used for such a comparison: one is the average waiting time in queue ($w$), and the other is the proportion of customers getting served ($P_s$). We choose five scenarios to check the robustness of RTA approximation in different settings. Especially, we will see if the RTA approximation can produce good result when the system size is large.

We test the accuracy of our algorithm in five different settings. Tables 1 - 5 show the performance of the RTA approximation in different traffic densities, abandonment rates, retrial rates, retrial probabilities, and system size. For the simulation results in the following tables, we run 5 replications, each of which generates around 30,000 arrivals, and then we take the mean value of the 5 replications. The run time for each simulation (including 5 replications) is about 30 minutes, while the calculation time for each situation for Tables 1 - 4 is less than 10 seconds and the calculation time for each case in Table 5 is less than one minute.

*Table 1*

$\gamma = 0.1,\ \beta = 0.5,\ \mu = 0.1,\ N = 100$

| | $w$ | | | $P_s$ | | |
|---|---|---|---|---|---|---|
| | RTA | Simulation | Difference | RTA | Simulation | Difference |
| $\lambda/c\nu = 1$ | 3.997 | 4.072 | 1.84% | 0.675 | 0.684 | 1.32% |
| $\lambda/c\nu = 0.5$ | 1.038 | 1.062 | 2.26% | 0.897 | 0.910 | 1.43% |
| $\lambda/c\nu = 0.25$ | 0.658 | 0.614 | 7.17% | 0.997 | 0.997 | 0 |
| $\lambda/c\nu = 0.2$ | 0.621 | 0.578 | 7.44% | 1 | 1 | 0 |

Table 1 shows the performance measures difference between the original model and the model based on RTA approximation in different traffic density. From this table we can see that RTA approximation works pretty well in heavy traffic situations ($\lambda/c\nu \geq 0.5$). Nevertheless, when the traffic density is low, the accuracy of such an approximation is not so satisfying. One possible reason for this phenomenon is that when the traffic density is low, the assumption that $L_i = L$ for all $i$ fails to hold. Since managers always want to set the system to high traffic density situation in order to make the best use of servers, the lack of accuracy of this algorithm in low traffic density situation should not be a big concern in practice.

*Table 2*

$\lambda = 0.1,\ \beta = 0.5,\ \mu = 0.1,\ N = 100,\ c = 2,\ \nu = 0.1$

| | $w$ | | | $P_s$ | | |
|---|---|---|---|---|---|---|
| | RTA | Simulation | Difference | RTA | Simulation | Difference |
| $\gamma = 0.2$ | 0.664 | 0.678 | 2.06% | 0.868 | 0.882 | 1.59% |
| $\gamma = 0.1$ | 1.038 | 1.062 | 2.26% | 0.897 | 0.910 | 1.43% |
| $\gamma = 0.05$ | 1.494 | 1.496 | 0.13% | 0.926 | 0.932 | 0.64% |
| $\gamma = 0.02$ | 2.130 | 2.122 | 0.38% | 0.958 | 0.966 | 0.83% |

The above table shows the comparison between the two models in different abandonment rates. It is shown that the RTA approximation works well in all kinds of abandonment rates. Also, we find that high abandonment rate will make the average waiting time in queue be lower. Nevertheless, a higher abandonment rate means higher loss of customers, which is not desired.

*Table 3*

$\lambda = 0.1,\ \beta = 0.5,\ \gamma = 0.1,\ N = 100,\ c = 2,\ \nu = 0.1$

| | $w$ | | | $P_s$ | | |
|---|---|---|---|---|---|---|
| | RTA | Simulation | Difference | RTA | Simulation | Difference |
| $\mu = 0.2$ | 1.040 | 1.066 | 2.44% | 0.898 | 0.902 | 0.44% |
| $\mu = 0.1$ | 1.038 | 1.062 | 2.26% | 0.897 | 0.910 | 1.43% |
| $\mu = 0.05$ | 1.037 | 1.042 | 0.48% | 0.896 | 0.896 | 0 |
| $\mu = 0.02$ | 1.036 | 1.038 | 0.19% | 0.896 | 0.894 | 0.22% |

From this table, we can see that RTA approximation works well in different retrial rates. Also, it is shown that a higher retrial rate will make the average waiting time higher. This is because a higher retrial rate results in a higher incoming rate for the retrial stream, so that the traffic density will get higher. However, in the case shown above, the probability that a customer will choose to abandon and then retrial is low. The effect of the retrial rate on the whole system performance is not so obvious.

*Table 4*

$\lambda = 0.1,\ \gamma = 0.1,\ \mu = 0.1,\ N = 0.1,\ c = 1,\ \nu = 0.125$

| | $w$ | | | $P_s$ | | |
|---|---|---|---|---|---|---|
| | RTA | Simulation | Difference | RTA | Simulation | Difference |
| $\beta = 0.2$ | 2.955 | 2.972 | 0.57% | 0.706 | 0.694 | 1.73% |
| $\beta = 0.4$ | 2.998 | 3.026 | 0.93% | 0.716 | 0.708 | 1.13% |
| $\beta = 0.6$ | 3.211 | 3.290 | 2.40% | 0.757 | 0.762 | 0.66% |
| $\beta = 0.8$ | 3.513 | 3.564 | 1.43% | 0.813 | 0.804 | 1.12% |

In the Table 4, we show the accuracy of RTA approximation for different $\beta$ values. We can see that the differences for $w$ and $P_s$ between the original model and RTA model are small. Besides, comparing different rows, we find that a higher probability to retrial ($\beta$ value) leads to worse system performance, e.g., a higher average waiting time. This is because a higher probability to retrial means more customers to retrial, which will make the traffic density larger. As a result, the average waiting time in queue will get worse. Nevertheless, when such a probability increases, the proportion of customers getting served is higher which is desired by managers.

*Table 5*

$N = 100,\ \gamma = 0.1,\ \mu = 0.1,\ \beta = 0.5,\ \nu = 0.1,\ \lambda/c\nu = 0.9$

| | $w$ | | | $P_s$ | | |
|---|---|---|---|---|---|---|
| | RTA | Simulation | Difference | RTA | Simulation | Difference |
| $c = 10$ | 0.997 | 1.012 | 1.48% | 0.937 | 0.947 | 1.06% |
| $c = 20$ | 0.939 | 0.920 | 2.07% | 0.965 | 0.971 | 0.62% |
| $c = 50$ | 0.660 | 0.654 | 0.92% | 0.986 | 0.982 | 0.41% |
| $c = 100$ | 0.342 | 0.348 | 1.72% | 0.993 | 0.998 | 0.50% |

The above table describes the performance of RTA approximation in different system sizes.

From this table we can see that even with a large system size, the result got by RTA approximation is closed to the actual system output. We also observe that as the system size grows, the whole system performance gets better, e.g., the average waiting time in queue is lower and the proportion of customers getting served is higher. From this observation, we can conclude that when the system size is big, we can try tougher staffing decision, which makes the traffic density even higher without worrying about failing to retain the service goal.

In this section, we complete our investigation over the queueing system with both abandonments and retrials for the single type customers case. In the next section, we will extend our study to multi-type customers scenarios.

## 2.4 Extensions to multi-class customers models

We now examine the queueing system with both abandonments and retrials to the case with multi-class incoming customers. Two scenarios are considered here: one is multi-class customers with different abandonment rates and the other is multi-class customers with different retrial rates.

### Multi-class retrial queue with different abandonment rates

We first study the model with two types of customers having different abandonment rates. Particularly, one type of customers have finite patience and the others have infinite patience, i.e., their abandonment rate is zero. Our model, depicted in Figure 2.2, contains three types of nodes. The first two are queue I and queue II, which share the same set of $c$ servers with exponential service time of rate $\nu$. The third one is the retrial pool (orbit). As mentioned before, it can be considered as a service system with infinite servers, where customers effectively serve themselves. Type-I customers who have finite patience enter queue I with Poisson arrivals of rate $\lambda_1$. The arrival process of type-II customers to queue II is also Poisson with rate $\lambda_2$. In our model, customers in queue I have non-preemptive priority over customers in queue II; thus the service for type-II customers cannot be interrupted by the arrival of a type-I customer. Each customer waiting in queue I has the same maximum tolerance time which is exponentially distributed with rate $\gamma$, meaning that each customer waiting in queue I abandons at rate $\gamma$. An abandoning customer joins queue II with probability $1-\beta$ or joins the retrial pool with probability $\beta$. Since customers in queue II have infinite patience, they do not renege. Customers in the retrial pool leave and subsequently enter queue I at rate $\mu$. In the service node, retrial customers and new customers are not distinguished.

For the above model, as we need a three-dimensional state space to fully describe the system, it is difficult to get the stationary distribution of queue I and queue II due to such a large state space. Here we choose to use *level crossing* techniques to get the virtual waiting time distribution. Nevertheless, in order to use such techniques, it is essential to separate the

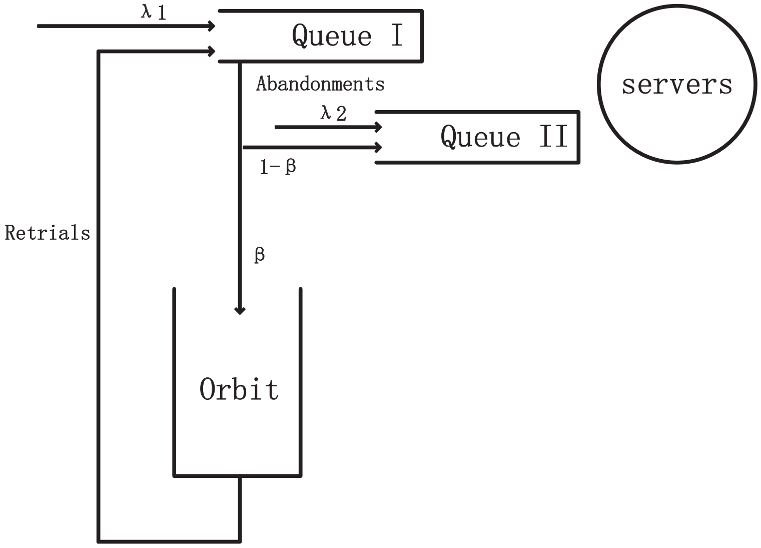

Figure 2.2: Retrial queue with two types of customers of different abandonment rates.

analysis of queue I and queue II with the retrial pool. Such a separation can be achieved by adopting RTA approximation mentioned before. In this model, without considering queue II, the sub-system with queue I and the retrial pool is just the same as the model mentioned in Section 2.3. Since customers in queue II do not abandon, it is reasonable to apply RTA approximation given that its effectiveness is proved comprehensively in the previous section. Just like the RTA approximation mentioned in Section 2.3, we assume that no matter how many customers are in queue I, the expected number of customers in the retrial pool remains a constant $L$.

Let $f_1(x)$ be the density function for the virtual waiting time of customers in queue I. We define the *active phase* to be the periods during which all servers are busy. Accordingly, $f_1(x)$ is defined for $x > 0$, which is the virtual waiting time given that we are in an active phase. In other words, we have $f_1(x) = f_1(x|AP)P(AP)$, where $P(AP)$ is the probability that the system is in an active phase. Let $P_j$ be the probability that $j$ servers are busy when $j \leq c-1$. Then the probability that an arriving customer finds all servers busy is

$$P(AP) = 1 - \sum_{j=0}^{c-1} P_j. \tag{2.12}$$

In order to get $f_1(x)$, we need the following lemma.

**Lemma 5.** *$f_1(x)$ can be obtained by solving the following equations.*

$$\begin{aligned} f_1(x) = {} & (\lambda_1+\lambda_2+L\mu)P_{c-1}e^{-c\nu x} + (\lambda_1+L\mu)\int_0^x e^{-c\nu(x-y)}e^{-\gamma y}f_1(y)dy + \lambda_2 P(AP)e^{-c\nu x} \\ & + (\lambda_1+L\mu)(1-\beta)e^{-c\nu x}\int_0^x (1-e^{-\gamma y})f_1(y)dy. \end{aligned} \tag{2.13}$$

$$\sum_{j=0}^{c-1}\frac{\rho^j}{j!}P_0 + \int_0^\infty f_1(y)dy = 1, \tag{2.14}$$

*where $\rho = (\lambda_1+\lambda_2+L\mu)/\nu$.*

Solving (2.13) and (2.14), we can get $f_1(x)$ and $P(AP)$ as functions of $L$. (Detailed calculation steps are included in the Appendix.) Here, we present the calculation of some queueing measures first and leave the discussion of the value of $L$ to the next part.

**Proposition 6.** *Given $f_1(x)$ and $P(AP)$, some performance measurements for the above queueing system can be calculated as follows.*

1. *The probability that customers in queue I choose to abandon and finally join queue II:*

$$P_c = (1-\beta)\int_0^\infty f_1(y)(1-e^{-\gamma y})dy.$$

2. *The conditional probability that a customer in queue I is served given that it does not abandon:*

$$P_s = \frac{1-P(AP)+\int_0^\infty f_1(y)e^{-\gamma y}dy}{1-P_c}.$$

3. *The waiting time density function of a customer who has been served without entering queue II:*

$$W_s(x) = \frac{1-P(AP)+\int_0^x f_1(y)e^{-\gamma y}dy}{(1-P_c)P_s}.$$

4. *Mean number of customers waiting in queue I:*

$$A = (\lambda_1+L\nu)\int_0^\infty xW_s(x)dx.$$

To apply Proposition 6, the value of $L$ is needed. We can use the same method mentioned in Section 2.3 to obtain such a value. Equating the total incoming rate and outgoing rate for the orbit, we have:

$$L\mu = A\gamma\beta.$$

Solving the above equation, we can get the value of $L$; afterwards, we can analyze properties of queue I following the approach mentioned above. In our study, we focus on the performance of queue I. Since queue II is considered to be a call-back queue, its performance measurements are omitted here.

### Multi-class retrial queue with different retrial rates

In this subsection, we consider the case with two types of customers who have different retrial rates involved. Similar to the model in Section 2.3, the model in this section also consists of two nodes: a service node with $c$ servers, and a retrial pool (orbit) with infinite servers. Two types of customers arrive at the service node as a Poisson process with rates $\lambda_1$ and $\lambda_2$. At the service node, both types of customers are treated equally, meaning that there is no priority rule between the two. Customers arriving to find an idle server are taken into service, service time for both types of customers is exponentially distributed with rate $\nu$. Customers who find all servers busy join a queue, where they are served in an FCFS manner, the maximum number of customers in the service node is $N$. Both types of waiting customers have the same maximum tolerance time which is also exponentially distributed with rate $\gamma$. An abandoning customer leaves the system forever with probability $1 - \beta$ or joins the retrial pool with probability $\beta$. In the retrial pool, type-I customers leave at rate $\mu_1$, while type-II customers leave at rate $\mu_2$. After customers complete their retrials, they are treated as new customers.

In order to fully describe the system, we need a four dimensional state space, $(i_1, i_2, j_1, j_2)$, where $i_1$ and $i_2$ are the numbers of type-I and type-II customers in the service node, $j_1$ and $j_2$ are the numbers of type-I and type-II customers in the retrial pool. Although we can write down the transition rate matrix for such a state space, it is not possible to solve the balance equations in most cases. Observing that two types of customers have completely the same properties in the service node. We have the following equations:

$$E[M_1] = \frac{\lambda_1'}{\lambda_1' + \lambda_2'} E[M] \text{ and } E[M_2] = \frac{\lambda_2'}{\lambda_1' + \lambda_2'} E[M],$$

where $E[M_1]$ and $E[M_2]$ are the average numbers of type-I and type-II customers in the service node, and $E[M] = E[M_1] + E[M_2]$ is the average number of total customers in the service node. $\lambda_1'$ and $\lambda_2'$ are the virtual arrival rates for type-I and type-II customers, which are the sum of new arrivals together with retrials. From the above equations, when calculating first moment, state space collapse can be achieved. In other words, we can reduce the original two dimensional problem for the service node into one, which serves as the basis for our further analysis.

Recall that we use the well known PASTA property as the basis for RTA approximation. Now we can extend the PASTA property into multi-class customers' case. When the arrival process is Poisson, the limiting probability that a retrial customer sees $i$ type-I (II) customers is the same as the *stationary* probability that there are $i$ type-I (II) customers in the service node. So, we can get the following equations.

$$N_{i_1} = E[N_1] \, \forall i \text{ , and } N_{i_2} = E[N_2] \, \forall i,$$

where $N_{i_1}$ and $N_{i_2}$ are the expected numbers of type-I and type-II customers in the orbit when there are $i$ customers in the service node. $E[N_1]$ and $E[N_2]$ are the average number of type I and II customers in the orbit. Let $E[N_1] = L_1$ and $E[N_2] = L_2$. We have

$$\lambda_1' = \lambda_1 + L_1\mu_1 \text{ , and } \lambda_2' = \lambda_2 + L_2\mu_2.$$

Now, we can use the same techniques mentioned in Section 2.3 to analyze queueing properties for the service node. The limiting probability that there are $i$ customers in total at the service node is given by

$$P_i = \prod_{j=1}^{i} \frac{\lambda_1 + L_1\mu_1 + \lambda_2 + L_2\mu_2}{\nu_i + \gamma_i} / (1 + \sum_{i=1}^{N} (\prod_{j=1}^{i} \frac{\lambda_1 + L_1\mu_1 + \lambda_2 + L_2\mu_2}{\nu_i + \gamma_i})). \tag{2.15}$$

We can calculate $E[M_1]$ and $E[M_2]$ as follows

$$E[M_1] = \frac{\lambda_1 + L_1\mu_1}{\lambda_1 + L_1\mu_1 + \lambda_2 + L_2\mu_2} \sum_{i=c}^{N} iP_i,$$

$$E[M_2] = \frac{\lambda_2 + L_2\mu_2}{\lambda_1 + L_1\mu_1 + \lambda_2 + L_2\mu_2} \sum_{i=c}^{N} iP_i.$$

Equating the incoming rate and outgoing rate for both type I and II customers in the orbit, we get:

$$L_1\mu_1 = \frac{\lambda_1 + L_1\mu_1}{\lambda_1 + L_1\mu_1 + \lambda_2 + L_2\mu_2} \sum_{i=c}^{N} (i - c)P_i\gamma\beta, \tag{2.16}$$

$$L_2\mu_2 = \frac{\lambda_2 + L_2\mu_1}{\lambda_1 + L_1\mu_1 + \lambda_2 + L_2\mu_2} \sum_{i=c}^{N} (i - c)P_i\gamma\beta. \tag{2.17}$$

Plugging (14) into (15) and (16) and solving the equations, we can get the value of $L_1$ and $L_2$. Some performance measures of the service node can be achieved.

**Proposition 7.** *Given $L_1$ and $L_2$, some performance measurements for the above queueing system can be calculated as follows.*

1. *The average waiting time for each customer in the service node:*

$$W = \sum_{i=c}^{N} \frac{(i-c)\prod_{j=1}^{i} \frac{\lambda_1+L_1\mu_1+\lambda_2+L_2\mu_2}{\nu_i+\gamma_i}}{1 + \sum_{i=1}^{N} (\prod_{j=1}^{i} \frac{\lambda_1+L_1\mu_1+\lambda_2+L_2\mu_2}{\nu_i+\gamma_i})} / \ (\lambda_1 + L_1\mu_1 + \lambda_2 + L_2\mu_2 \ ).$$

*2. The proportion of customers getting served:*

$$Ps = \frac{\nu}{\lambda}\left(\sum_{i=0}^{c} \frac{i\prod_{j=1}^{i}\frac{\lambda_1+L_1\mu_1+\lambda_2+L_2\mu_2}{\nu_i+\gamma_i}}{1+\sum_{i=1}^{N}(\prod_{j=1}^{i}\frac{\lambda_1+L_1\mu_1+\lambda_2+L_2\mu_2}{\nu_i+\gamma_i})} + \sum_{i=c+1}^{N} \frac{c\prod_{j=1}^{i}\frac{\lambda_1+L_1\mu_1+\lambda_2+L_2\mu_2}{\nu_i+\gamma_i}}{1+\sum_{i=1}^{N}(\prod_{j=1}^{i}\frac{\lambda_1+L_1\mu_1+\lambda_2+L_2\mu_2}{\nu_i+\gamma_i})}\right).$$

From Proposition 7, we can see that $W$ is increasing with respect to $\lambda_1$, $\lambda_2$, $\mu_1$, $\mu_2$ and $\beta$; but decreasing with respect to $\gamma$ and $\nu$. Actually, this fits our intuition quite well. Higher $\lambda_1$ and $\lambda_2$ mean larger incoming flows, with service capacity unchanged, the traffic density will get higher. As a result, the average waiting time will get higher. For $\beta$, a higher $\beta$ means more customers tend to retrial, which also results in a larger incoming flow as well as higher average waiting time. The impact of an increase in $\mu_1$ and $\mu_2$ is similar to $\beta$, which also results in a higher average waiting time. The situation for $\gamma$ is just the opposite, high value of $\gamma$ means more customers reneging from the system, which decreases the overall traffic density. As for $\nu$, the result is even more straight forward. With a higher service capacity, average waiting time will decrease when incoming flow is kept unchanged.

## 2.5 Conclusion

In this chapter, we study the stationary behavior of a queueing system with both abandonments and retrials. Using the RTA approximation, we separate the analysis of the service node and retrial pool, and subsequently derive the queueing dynamics. By finding the upper bound and lower bound for both the original model and RTA model, we prove that the difference between the result obtained by RTA approximation and the actual system output is small in the heavy traffic region. Additionally, our simulation results confirm the accuracy of our methodology. Further, the sensitivity analysis on various key parameters exhibits the same structural properties in both the original model and the model with approximation. We also derive the queueing behaviors in two kinds of queueing systems with multi-class customers. We first study the model involving two types of customers with different abandonment rates. In this model, two types of customers have separate queues, with one has priority over the other. Using level crossing techniques, we find the density function for the virtual average waiting time, by which we calculate other performance measures of the queue with priority. Finally, we consider the scenario where two types of customers with different retrial rates are introduced. By extending PASTA property and RTA approximation into multi-type case, we characterize the stationary behavior of the service node and get the mean number of both types of customers in the main queue.

Our analysis can be extended in several ways. As a natural extension, one can study the model with heterogeneous abandonment rates and heterogeneous retrial rates. This model is a combination of the two multi-class scenarios we have investigated. Another possible direction is to incorporate various sorts of priority rules in our multi-class models.

For example, in certain scenarios, the server may give equal priority to different classes of customers; in this case, the queue is no longer completely separated, and one has to keep track of the detailed sequence of customers awaiting in the same queue. This complicates the analysis because the state space explodes dramatically. While it is beyond the scope of this chapter, it remains a fruitful direction for future work.

# Chapter 3

# Retrial Queueing System with Negative Customers and Unreliable Servers

## 3.1 Introduction

In call centers, customer retrials can be commonly found. In a report regarding the recent development of call centers published by Global Industry Analysts, Inc., it is pointed out that around 10%-15% customers in call centers would choose to renege while waiting in the line. Among those customers, around 40% choose to redial after some time. Besides customer retrials, deletion signals and unreliable servers are also common in practice. In the year of 2009, the loss caused by failures of computer telephony integration systems is around 106 million USD.

As a motivating example of the queueing system with customer retrials, deletion signals (negative customers) as well as unreliable servers, consider a call center with computer telephony integration which consists of multiple trunk lines that connect calls to the center through the private automatic branch exchange. An arriving call that finds one customer service representative available will get served immediately. If all customer representatives are busy, the customer waits in queue until a customer representative becomes available. Some customers are not patient enough to wait until a customer representative becomes available. After getting impatient, some customers will retry to access the call center after random amount of time, this is called customer retrials. When serving existing customers, call centers also admit messages from other service sites, inviting customers to immigrate there. Customers are matched with single messages, and are sent to the sites where the messages had originated. Such process takes no time. In this case, such invitation messages can be considered as negative customers, which remove an ordinary (positive) customer upon arrival. In call centers, the computer telephony integration system is not always reliable. It would become non-functional due to virus or hardware failures, thereby violating the operation of the system. This phenomenon is called server disasters. Also, customer representatives in call centers can not work all the time. They need to take a rest sometimes, such as lunch break. Such rest usually happens at service completions, which is called server vocations.

In some classic papers, the impact of customer retrials, negative customers and unreliable servers on queueing system performance has also been discussed. Falin and Templeton (1997) emphasize that the standard queueing model that does not take retrial phenomenon into account cannot be applied in solving numbers of practically important problems. Artalejo (2000) provides a comprehensive survey on the impact of negative customers and disasters over queueing systems. It is shown by previous papers that for retrial queueing systems, incorporating negative customers or unreliable servers increases the complexity of queueing dynamics greatly. As a result, although much research has been done for the field of negative customers, customer retrials and unreliable servers, it is almost blank in queueing systems with these three phenomena. For previous papers on the performance measures of retrial queueing systems with negative customers and unreliable servers, results crucially rely on fluid models approximation or heavy traffic approximation. Different from these papers, we characterize the detailed queueing dynamics of the system in this chapter, which allows us to get high moments of performance measures.

In this chapter, we first study a multi-server queueing system with two types of customers, regular customers and negative customers. Regular customers may renege from the system if their waiting time exceeds their patience limit. Upon abandonment, customers may choose to rejoin the service system after some time. For negative customers, upon their arrival, they will remove some regular customers instantly. To tackle this problem, we invoke the RTA (retrials see time average) approximation. Subsequently, the performance measurements for the entire queueing system are derived.

We also extend our study to incorporate unreliable servers, where two kinds of server interruptions (server disasters and server vocations) are considered together. We derive differential equations regarding probability generating functions from the balance equations. Then, using the concept of stochastic decomposition, we get the generating function of the steady-state number of customers; this subsequently leads to the performance measurements for the entire queueing system.

The rest of this chapter is organized as follows. In Section 3.2, we review previous research papers briefly. In Section 3.3, we analyze the stationary behavior of a queueing system with customers abandonments, retrials as well as negative customers. In Section 3.4, we continue our study to the case with an unreliable server. We summarize our findings in Section 3.5.

## 3.2 Literature review

In the previous queueing literature, customer retrial has been well studied. For the multi-server retrial model, explicit formulae for the main performance characteristics are absent when the number of servers is more than two, see Artalejo and Gómez-Corral (2008). Consequently, different approximation methods are introduced, such as direct truncation model, generalized truncation model, RTA (Retrial See time Average) approximation, and etc. For instance, by applying PASTA property to the retrial queueing system, Artalejo (1995) assumes the average number of customers in the retrial pool to be a constant no matter how many customers in the queue. With this assumption, he got the average number of customers both in the queue and in the retrial pool. Recently, Artalejo et al. (2005) deal with a multi-server retrial queueing model in which the number of active servers depends on the number of customers in the system. For a fixed choice of the threshold levels, the stationary distribution and different performance measures of the system are calculated. The optimum threshold level is also numerically computed in the case of equidistant connection levels. For detailed overviews of the main results and the bibliographical information about the retrial queues, see Artalejo (2010).

Over the past couple of decades, many papers have studied negative customers. Applications of negative arrivals to queueing networks can be found in a survey by Artalejo (2000). Recently, the topic of negative customers was extended to a discrete-time queue systems. Atencia and Moreno (2004) present a stationary queue length distribution of the Geo/Geo/1 queue. That model has either negative customers or disasters under an assumption that an arriving customer is classified as a positive customer or a negative customer (disaster) with a certain probability. Park et al. (2009) extend the analysis to Geo/G/1 queue and get the stationary queue length as wellas sojourn time of that model.

On the other hand, many authors have investigated the systems with a repairable service system which has breakdowns or some other kinds of service interruptions. Takine and Sengupta (1997) study a queueing system with interruptions under the assumption that as soon as the service channel fails, it instantaneously undergoes repairs. Ke (2006) investigates some control policies for unreliable server systems. Recently, Choudhury and Tadj (2009) study an M/G/1 queue with a second optional service channel which is subject to random breakdowns. They derive the joint distributions of state of the server and queue size, and also obtain some important performance measures as well as reliability indices.

For queueing models with unreliable servers, incorporating customer retrials brings several challenges. Sherman and Kharoufeh (2006) analyze an unreliable M/M/1 retrial queue with infinite-capacity orbit and normal queue. Retrial customers repeatedly attempt to access the server at i.i.d. intervals until it is found functioning and idle. They provide stability conditions as well as several stochastic decomposability results. Adopting different settings, Li et al. (2006) consider a BMAP/G/1 retrial queue with a server subject to breakdowns and repairs, where the life time of the server is exponential and the repair time is general. By using the supplementary variable method, they apply the RG-factorization of a level-dependent continuous-time Markov chain of M/G/1 type and provide the stationary performance measures of the system.

## 3.3 Retrial queueing system with negative customer

### Model description

Our model consists of two types of nodes: a service node with $c$ servers, and a retrial pool (orbit) with infinite servers, where customers effectively serve themselves. Regular customers and negative customers arrive at the service node as Poisson process with rates $\lambda_1$ and $\lambda_2$. When one negative customer arrives at the system, $i$ regular customers are forced to leave the system with probability $q_i$, and this negative customer leaves. Such a process happens instantly. Note that when the number of regular customers in the system is less than $i$, all regular customers will be removed. The service node has a maximum capacity $K$. When there are already $K$ customers in the service node, new arrivals will be blocked. If regular customers arrive and find an idle server, they are taken into service with a duration which is exponentially distributed with rate $\nu$. Customers who find all servers busy join a queue and are served in a FCFS (First-Come-First-Serve) manner. Each waiting customer has the same maximum tolerance time which is exponentially distributed with rate $\gamma$; equivalently, each customer waiting in queue abandons at rate $\gamma$. An abandoning customer leaves the system forever with probability $1-\beta$ or joins the retrial pool with probability $\beta$. Customers in the retrial pool leave and subsequently enter the service node at rate $\mu$. Consequently, when there are $i$ customers in the orbit, the total retrial rate is $i\mu$. After customers complete their retrials, they are treated as new customers, i.e., new customers and retrial customers are not distinguished.

### Stationary distribution analysis

The above queueing system can be modeled as a continuous-time Markov chain with state $\zeta(t) = \{C(t), M(t)\}$, where $C(t)$ is the number of customers in the main queue and $M(t)$ is the number of customers in the orbit. Let $P_{i,j}$ be the stationary probability of state $(i, j)$, the balance equations for this Markov chain are as follows:

For $i = 0$:

$$P_{i,j}(\lambda_1 + j\mu) = (i+1)\nu P_{i+1,j},$$

For $0 < i \leq c$:

$$P_{i,j}(\lambda_1 + \lambda_2 + i\nu + j\mu) = \lambda P_{i-1,j} + (j+1)\mu P_{i-1,j+1} + ((i+1)\nu + \lambda_2)P_{i+1,j},$$

for $c < i < N$:

$$\begin{aligned}&P_{i,j}(\lambda_1 + \lambda_2 + c\nu + j\mu + (i-c)\gamma) = \lambda P_{i-1,j} +\\&(j+1)\mu P_{i-1,j+1} + (c\nu + \lambda_2)P_{i+1,j} + (i+1-c)\beta\gamma P_{i+1,j-1} + (i+1-c)(1-\beta)\gamma P_{i+1,j},\end{aligned}$$

for $i = N$:

$$P_{i,j}(c\nu + (i-c)\gamma) = \lambda P_{i-1,j} + (j+1)\mu P_{i-1,j+1} + (j+1)\mu P_{i,j+1}.$$

Define $L_i = \sum\limits_{j=0}^{\infty} jP_{i,j}$, which is the expected number of customers in the orbit when there are $i$ customers in the service node. Based on RTA (retrials see time average) approximation, we have: $L_i = L$, $\forall i$, where $L$ is chosen in $R^+$. With RTA approximation, we can separate the analysis of the service node apart from the analysis of the retrial pool. Here, we need some additional notations:

$$\nu_i = \min(i,c)\nu, \gamma_i = (i-c)^+\gamma, P_i = \sum_{j=0}^{\infty} P_{i,j},$$

where $P_i$ is the probability that there are $i$ customers in the service node. Considering the service node separately, we have the following balance equation:

$$P_i(\lambda_1 + L\mu) = P_{i+1}(\nu_{i+1} + \gamma_{i+1} + \lambda_2 q_1) + \sum_{j=i+2}^{k} P_j \lambda_2 q_{j-i}.$$

Define:

$$y_i = \frac{\nu_i + \gamma_i + \lambda_2 q_1}{\lambda_1 + L\mu}, F_i = \frac{\lambda_2 q_i}{\lambda_1 + L\mu}.$$

We have:

$$P_i = P_{i+1} y_{i+1} + \sum_{j=i+2}^{K} P_j F_{j-i}.$$

The above equation means that $P_i$ can be expressed as a function of $(P_{i+1}, P_{i+2}, ..., P_K)$. To simplify the expression for $P_i$, we define the following:

$$x(i,s) = \begin{cases} y_i \;\; if \;\; s=1 \\ F_s \;\; if \;\; s>1. \\ 1 \;\;\; if \;\; s=0 \end{cases}$$

Also, define $\mathbf{S_i}$ as a set of sequence $(s_K, s_{K-1}, ..., s_i)$ which satisfies $\sum\limits_{j=i}^{K} s_j = K-i, s_j \in N$. With the definition, we can get the expression for $P_i$ as follows:

$$P_i = P_K \sum_{S_i} \prod_{j=i}^{K} x(i + \sum_{t=i}^{j} s_t, s_j).$$

Also, we have the normalization condition: $\sum\limits_{i=0}^{K} P_i = 1$. Then we can get the stationary probabilities for this queueing system as follows:

$$P_i = \frac{\sum\limits_{S_i} \prod\limits_{j=i}^{K} x(i + \sum\limits_{t=i}^{j} s_t, s_j)}{\sum\limits_{i=0}^{K} \sum\limits_{S_i} \prod\limits_{j=i}^{K} x(i + \sum\limits_{t=i}^{j} s_t, s_j)}.$$

Given $P(i)$, the average number of customers waiting in queue can be calculated as: $E = \sum_{i=c}^{K}(i-c)P_i$. Then, equating the incoming rate to the orbit and the outgoing rate from the orbit, we have: $\gamma E = \mu L$. Solving this equation, we can get the value of $L$. Now, with $L$, we are able to calculate some performance measurements for the queueing system.

**Proposition 8.** *The system performance can be characterized as follows:*

*1.* The probability that all servers are idle:

$$P_0 = \frac{\sum\limits_{S_0}\prod\limits_{j=i}^{K} x(i+\sum\limits_{t=i}^{j} s_t, s_j)}{\sum\limits_{i=0}^{K}\sum\limits_{S_i}\prod\limits_{j=i}^{K} x(i+\sum\limits_{t=i}^{j} s_t, s_j)}.$$

*2.* The average number of customers in the service node*:*

$$E(Q) = \sum_{i=c}^{K}\frac{(i-c)\sum\limits_{S_i}\prod\limits_{j=i}^{K} x(i+\sum\limits_{t=i}^{j} s_t, s_j)}{\sum\limits_{i=0}^{K}\sum\limits_{S_i}\prod\limits_{j=i}^{K} x(i+\sum\limits_{t=i}^{j} s_t, s_j)}.$$

*3.* The average waiting time for each customer in the main queue*:*

$$E(w) = \frac{1}{L\mu+\lambda_1}\sum_{i=c}^{K}\frac{(i-c)\sum\limits_{S_i}\prod\limits_{j=i}^{K} x(i+\sum\limits_{t=i}^{j} s_t, s_j)}{\sum\limits_{i=0}^{K}\sum\limits_{S_i}\prod\limits_{j=i}^{K} x(i+\sum\limits_{t=i}^{j} s_t, s_j)}.$$

*4.* The proportion of customers getting blocked*:*

$$P(b) = \frac{1}{\sum\limits_{i=0}^{K}\sum\limits_{S_i}\prod\limits_{j=i}^{K} x(i+\sum\limits_{t=i}^{j} s_t, s_j)}.$$

From Proposition 8, we can find that $E(Q)$ is an increasing function with respect to $\lambda_1$, $\mu$ and $\beta$, it is also a decreasing function with respect to $\lambda_2$, $\nu$ and $\gamma$. This observation meets our intuition well. There is also an interesting observation that although we can not easily find the how the change of the average number of regular customers deleted by a negative customer ($\sum_{i=1}^{K} iq_i$) affect the value of $E(Q)$, we can find that there is no monotonic relationship existed through some numerical experiments. For example, when $\lambda_1 = 10$, $\lambda_2 =$

10, $\nu = 15$, $\mu = 5$, $\gamma = 5$, $\beta = 0.5$, consider two scenarios, one is $q_1 = q_2 = 0.5$, the other is $q_0 = 0.7$, $q_4 = 0.3$. Although the average number of regular customer deleted by a negative customer for scenario 2 is smaller than scenario 1, $E(Q)$ for scenario 2 is larger than scenario 1. One possible reason for this is that the distribution variance for scenario 2 is larger than scenario 1. In this sense, we can see that for negative customers deletion behavior, both the first order property and the second order property will affect the value of $E(Q)$.

## 3.4 Retrial queueing system with negative customers and unreliable servers

In this section, we take unreliable servers into consideration, i.e., the server in the queueing system would become dysfunctional randomly. We consider two kinds of server interruptions together. One could happen at any time, while the other only happens randomly at service completions. Once the server breaks down (no matter what kind of interruption happens), service to the current customer stops. After some random time, the server becomes functional again and the service resumes. Here, we allow different resume times for the two kinds of service interruptions.

### Model description

The model here consists of two types of nodes: a service node with a single server, and a retrial pool. Differently from the model mentioned in Section 3.3, we only consider a single server queueing system in this section. The primary reason for this setting is the tractability. Regular customers and negative customers arrive at the service node as a Poisson process with rates $\lambda_1$ and $\lambda_2$. For each negative customer, it would remove at most one regular customer in the system. Thus, when one negative customer arrives, if the system is not empty, one regular customer is forced to leave the system. Otherwise, the negative customer just leaves with no effect on the queueing system. Such *single deletion rule* can be found in many recent papers, such as Park et al. (2009), Chae et al. (2010), and so on. In this model, the server breaks down with rate $\theta$. Customers interrupted by a server failure do not leave the server; they will wait until the service resumes. The repair time is exponentially distributed with rate $\omega_1$. Upon each service completion, with probability $\alpha$, the server would enter vacation mode; in this case, the service to current customer is interrupted. The resume rate for the server in vacation mode is $\omega_2$. We also take customer abandonments and retrials into consideration. Customers renege with rate $\gamma$ and retrial with rate $\mu$.

### Stationary distribution analysis

In order to fully describe the system, we need a three-dimensional state space, $(i, j, k)$, where $i$ is the number of customers in the service node, and $j$ is the status of the server, $k$ is the number of customers in the orbit. However, by using retrials see time averages (RTA) approximation, state space collapse can be obtained. Assuming the number of customers in the orbit as a constant $L$, we can use a continuous-time Markov process $\{X(t), U(t)\}$ to describe the state of the system at time t, where $X(t)$ is the number of customers in the system at $t$, and $U(t)$ is the status of the system. If at time $t$, the system is experiencing an interruption, then $U(t)$ is equal to D (disaster); if at time $t$, the system is in vacation mode, then $U(t)$ is equal to V (vacation); otherwise, $U(t)$ is N (normal). Accordingly, we denote the steady-state probability of the system by $P_{i,D}$, $P_{i,V}$ and $P_{i,N}$.

The steady-state balance equations are given below

$$\begin{aligned}
&(\lambda_1 + L\mu + \lambda_2 + \nu + (i-1)\gamma + \theta)P_{i,N} = \\
&(\lambda_1 + L\mu)P_{i-1,N} + (\lambda_2 + (1-\alpha)\nu + i\gamma)P_{i+1,N} + \omega_1 P_{i,D} + \omega_2 P_{i,V}, \\
&(\lambda_1 + L\mu + (i-1)\gamma + \omega_1)P_{i,D} = (\lambda_1 + L\mu)P_{i-1,D} + i\gamma P_{i+1,D} + \theta P_{i,N}, \\
&(\lambda_1 + L\mu + (i-1)\gamma + \omega_2)P_{i,V} = (\lambda_1 + L\mu)P_{i-1,V} + i\gamma P_{i+1,V} + \alpha\nu P_{i,N}.
\end{aligned} \tag{3.1}$$

The boundary equations are

$$\begin{aligned}
&(\lambda_1 + L\mu + \theta)P_{0,N} = (\lambda_2 + \nu)P_{1,N} + \omega_1 P_{0,D} + \omega_2 P_{0,V}, \\
&(\lambda_1 + L\mu + \omega)P_{0,D} = \theta P_{0,N}, \\
&(\lambda_1 + L\mu + \omega)P_{0,V} = \alpha\nu P_{0,N}.
\end{aligned} \tag{3.2}$$

Let $G_N(z) = \sum_{i=0}^{\infty} z^i P_{i,N}$, $G_D(z) = \sum_{i=0}^{\infty} z^i P_{i,D}$, and $G_V(z) = \sum_{i=0}^{\infty} z^i P_{i,V}$ for $|z| \leq 1$. By stochastic decomposition, the generating function of the steady-state number of customers in the service node is given by $G(z) = G_N(z) + G_D(z) + G_V(z)$.

Multiplying both sides of (3.1) and (3.2) by $z^i$ and summing over all $i$'s yield the differential equations

$$\begin{aligned}
&((\lambda_1 + L\mu)(1-z) + (\lambda_2 + \nu - \gamma)(1 - \tfrac{1}{z}) + \theta)G_N(z) + \gamma(z-1)G'_N(z) \\
&+(r - \lambda_2 - (1-\alpha)\nu)(1 - \tfrac{1}{z})P_{0,N} = \omega_1 G_D(z) + \omega_2 G_V(z), \\
&((\lambda_1 + L\mu)(1-z) - \gamma(1 - \tfrac{1}{z}) + \omega_1)G_D(z) + \gamma(z-1)G'_F(z) + \gamma(1 - \tfrac{1}{z})P_{0,D} = \theta G_N(z), \\
&((\lambda_1 + L\mu)(1-z) - \gamma(1 - \tfrac{1}{z}) + \omega_2)G_V(z) + \gamma(z-1)G'_V(z) + \gamma(1 - \tfrac{1}{z})P_{0,V} = \alpha\nu G_N(z).
\end{aligned}$$

For above differential equations, there are six unknown variables. Thus, we still need three additional equations:

$$\begin{aligned}
&(\lambda_1 + L\mu + \omega)P_{0,D} = \theta P_{0,N}, \\
&(\lambda_1 + L\mu + \omega)P_{0,V} = \alpha\nu P_{0,N}, \\
&P_{0,N} = G_N(0),
\end{aligned}$$

where the first two equations are from the boundary conditions and the second one is from the property of probability generating function. Then, with the normalization equation

$G(1) = G_N(1) + G_D(1) + G_V(1) = 1$, the unique solution for $G_N(z)$ and $G_F(z)$ can be obtained as follows.

$$G_N(z) = \frac{\nu(\omega_1+\omega_2) - (\lambda_1+L\mu)(\omega_1+\theta) + \gamma\lambda_2}{\nu(\omega_1+\omega_2+\theta)} P(z)^c,$$

$$G_D(z) = \frac{(\lambda_1+L\mu)\{(1-z)[\omega_1+\theta+(\lambda_1+L\mu)(1-z)][\nu\omega_1-(\lambda_1+L\mu)(\omega_1+\theta z)]\}}{\nu(\omega_1+\theta)[\theta z-(\nu-(\lambda_1+L\mu)z)(1-z)]} P(z)^{c+1},$$

$$G_V(z) = \frac{(\lambda_1+L\mu)\{(1-z)[\omega_1+\theta+(\lambda_1+L\mu)(1-z)][\alpha\nu\omega_2-(\lambda_1+L\mu)\omega_2]\}}{(\alpha\nu^2+\nu\theta)[\theta z-(\alpha\nu-(\lambda_1+L\mu)z)(1-z)]} P(z)^c,$$

where

$$P(z) = \frac{\nu\omega_1-(\lambda_1+L\mu)(\omega_1+\theta)+\gamma\lambda_2}{\nu\omega_1-(\lambda_1+L\mu)(\omega_1+\theta z)+\gamma\lambda_2 z} + \frac{\omega_2\nu-\alpha\nu}{\omega_2\nu-\alpha\nu z},\ c = \frac{\omega_1+\omega_2+\theta+\alpha\nu}{\mu}.$$

Since the generating function of the steady-state number of customers can be calculated as $G(z) = G_N(z)+G_D(z)+G_V(z)$, we are able to calculate some performance measurements for the queueing system.

**Proposition 9.** *Given $G(z)$, some performance measurements for the above queueing system can be calculated as follows.*

*1.* The probability that all servers are idle:

$$P_0 = \left(\frac{\nu\omega_1-(\lambda_1+L\mu)(\omega_1+\omega_2+\theta)+\gamma\lambda_2}{\nu\omega_1-(\lambda_1+L\mu)\omega_2}\right)^{\frac{\omega_1+\theta}{\mu}} \cdot \frac{\nu\omega_1+\gamma\lambda_2}{(\omega_1+\omega_2)^2+(\omega_1+\omega_2)\alpha\nu}.$$

*2.* The average number of customers in the service node*:*

$$N = \frac{\nu\theta(\theta+\alpha\nu)+(\lambda_1+L\mu)(\omega_1+\omega_2+\theta)^2+\gamma\lambda_2}{\nu(\omega_1+\theta)[(\omega_1+\omega_2)^2+(\omega_1+\omega_2)\theta-(\lambda_1+L\mu)(\omega_2+\theta)]}$$

*3.* The average number of customers waiting*:*

$$Q = \frac{\nu\theta(\theta+\alpha\nu)+(\lambda_1+L\mu)(\omega_1+\omega_2+\theta)^2+\gamma\lambda_2}{\nu(\omega_1+\theta)[(\omega_1+\omega_2)^2+(\omega_1+\omega_2)\theta-(\lambda_1+L\mu)(\omega_2+\theta)]}.$$

To apply Proposition 9, the value of L is needed. We can use the same method mentioned in Section 3.3 to obtain such a value. Equating the total incoming rate and outgoing rate for the orbit, we have:

$$Q\gamma\beta = L\mu.$$

Solving the above equation, we can get the value of L.

From Proposition 9, we can see that $Q$ is increasing with respect to $\lambda_1$, $\theta$, $\omega_1$, $\alpha$ and $\omega_2$. It is decreasing with respect to $\lambda_2$. This observation is quite straight forward. A higher $\lambda_1$ means a larger incoming flow, with service capacity unchanged, the traffic density will get higher. As a result, the average number of waiting customers will get higher. While the situation for $\lambda_2$ is just the opposite. For $\theta$, higher $\theta$ means a larger proportion of time the system is down. In this sense, the virtual service rate will be lower. Then, the average number of waiting customers will get higher. In the same way, we can see that as $\omega_1$, $\alpha$ and $\omega_2$ increase, the average number of waiting customers will also increase.

## 3.5 Conclusion

In this chapter, we study the stationary behavior of a multi-server queueing system with two types of customers, regular customers and negative customers. Regular customers may renege and subsequently reenter the queueing system, while negative customers can be considered as deletion signals. Using the RTA approximation, we separate the analysis of the service node and retrial pool, and then derive the queueing dynamics. Further, we derive the queueing behaviors in the retrial queueing system with negative customers and unreliable servers, where two kinds of server interruptions (server disasters and server vocations) are considered together. Using stochastic decomposition, we get the generating function of the steady-state number of customers; this subsequently leads to the performance measurements for the entire queueing system. We also compare our results with come previous papers, finding that our results converge to the results got by those papers in extreme cases.

Our analysis can be extended in several ways. First, we assume that there is no processing time for negative customers. This assumption can be relaxed in further analysis by adopting deterministic or random processing time for negative customers. Then, one can assign different priority rules between regular customers and negative customers. For example, in certain scenarios, the server may give equal priority to regular customers and negative customers; in this case, one has to keep track of the detailed sequence of customers awaiting in the same queue. Another possible direction is to allow each negative customer to delete random number of regular customers for the model with unreliable servers. However, both directions would complicate the analysis a lot because the state space explodes dramatically. While it is beyond the scope of this chapter, it remains a fruitful direction for future work.

# Chapter 4

# Revenue Management in Queue Systems With Customer Renege

## 4.1 Introduction

As a motivating example of the chapter, consider a computer repair shop facing customers with different lead time sensitivities. Some customer just want their computers to be repaired, no matter how much time it takes. Some customers are more sensitive on the lead time between order placement and delivery since they have some urgent need for their computer, in this sense, they value speedy service more. For the second type of customers, if their wait time exceed some certain amount, they may leave and seek for other services, this phenomenon is called customer renege. For the owner of the computer repair shop, how to design a price-lead time menu to maximize its revenue from heterogeneous customers is a critical issue. This chapter will provide an answer for such question.

The above example is a good case for the revenue management problem in queueing systems with customers renege. Actually, similar cases can be easily found in real life. UPS and Fedex offer different price-lead time options to meet different customers' need, e.g., same-day, five days, and so on. Dell would give priority customer support to those who buy extra coverage plan. For recent years, there are many papers dealing with revenue management in queueing systems, like Maglaras (2006), Afeche (2004), etc.. However, none of these papers take customers renege into consideration. In queueing theory, it is emphasized that the standard queueing model that does not take customers renege into account cannot be applied in solving numbers of practically important problems. In this sense, this chapter is aiming to fill the gap by studying a revenue management problem in queueing systems with customers renege.

In this chapter, we consider a problem in the context of a queueing model with two types of customers, i.e., impatient customers and patient customers who have different waiting cost per unit time. Also, for impatient customers, they may renege from the system if their waiting time exceeds their patience limit. For both types of customers, their valuation of service and waiting cost per unit time are private information. The service provider's goal is to design a menu and corresponding scheduling policy to maximize revenue when facing this information asymmetry. By using a mechanism design framework and adapting the achievable region approach, we transform the scheduling control problem into a nonlinear program. Then through a two-stage optimization technique, we finally identify the optimal price-lead time manus and scheduling policies for different scenarios. We also discuss how the optimality conditions for strategic delay (see Afeche (2004)) change in the setting with customers renege.

The rest of the chapter is organized as follows. In Section 4.2, we review previous research papers briefly. In Section 4.3, our model and problem formulation are explained in detail. Then, we analyze the optimal price-lead time manus and scheduling policies and present some structural results in Section 4.4. Finally, our main findings will be summarized in Section 5.

## 4.2 Literature review

In this chapter, we adapt achievable region approach to transform the scheduling control problem into a nonlinear program. This approach is pioneered by Coffman and Mitrani (1980) which studies the problem of designing scheduling strategies when the demand on the system is known and waiting time requirements are prespecified. After that, Shanthikumar and Yao (1992) extend these results by introducing the concept of strong conservation laws and proving a powerful result about the achievable region when strong conservation laws hold. Different from papers which apply achievable region approach on performance measures that are expectations of steady-state random variable, Green and Stidham (2000) study optimal control for certain scheduling problems on a sample-path basis by using the property that strong conservation laws hold for performance measures at every time point on every sample path.

In the field of resource allocation using mechanism design, Mussa and Rosen (1978) consider a class of monopoly pricing problem involving a quality-differentiated spectrum of goods with the same type. They find that the optimal policy would reveal consumer preferences and assign different consumer types to different varieties of goods. After that, Rochet and Choné (1998) provide existence proofs and characterization results for the multidimensional version of the multi-product monopolist problem of Mussa and Rosen (1978), they also show that bunching is robust in these multidimensional screening problems, even with very regular distributions of types. Recently, Dana Jr and Yahalom (2008) generalize Mussa and Rosen's model by introducing an aggregate resource constraint, which induces an upwards distortion in product quality for high valuation consumers because the standard downward distortions in product quality for other consumers relax the resource constraint. The above three papers are quite close to this one, however, our setup differs the above three in two aspects: first, in our model, customer utility depends on overall demand because of capacity constraint, queueing and delay costs; second, service provider can control externalities through scheduling policy.

On the other hand, many authors work on incentive-compatible socially optimal pricing and scheduling. Van Mieghem (2000) studies the optimal prices and service quality grades that a firm provides to heterogeneous customers who measure quality by their experienced delay distributions. They also analyze multi-plan pricing, which offers all customers a menu with a choice of multiple rate plans. Ha (2001) derives incentive-compatible and socially optimal prices under customer-chosen service requirements, first-in-first-out (FIFO) service rule and processor sharing. Hsu et al. (2009) propose a resource allocation and pricing mechanism for a service system that serves multiple classes of jobs where class of service request is subject to a class-dependent quality of service (QoS) guarantee on the expected delay bound. They show that the pricing scheme with the QoS guarantee depends on the scheduling policy implemented and has different characteristics from that without the QoS guarantee.

Recently, there are also many research papers dealing with pricing operational decisions for queueing systems. Hassin and Haviv (2003) give a very good survey on different topics in this field, e.g., profit maximization in observable queues, incentive compatible prices in priority queues and so on. Besides, Çelik and Maglaras (2008) consider a make-to-order manufacturer that offers multiple products to a market of price and delay sensitive users. They derive near-optimal dynamic pricing, lead-time quotation, sequencing, and expediting policies using an approximating diffusion control problem. Some numerical results are also provided to show the value of joint pricing and lead-time control policies. Akan et al. (2012) study a congestible system serving multiple classes of customers with convex-concave delay cost functions. By using a novel fluid model, they investigate how such menus should be chosen dynamically to maximize welfare. They also show that the cost-balancing policy is socially optimal if the system manager have perfect information on customers type.

## 4.3 Model description

### Model primitives and problem formulation

In our model, we consider a service system facing two kinds of customer, patient customers and impatient customers. Here, we define impatient customers to be type-1 and patient customers to be type-2. The arrival rates for these two types of customers are $\Lambda_1$ and $\Lambda_2$. These two types of customers differ in two attributes, service valuation and delay cost rate. For the service valuation, it represents a customer's willingness to pay for immediate delivery of the service. While for the delay cost rate, it measures the cost per unit time between the placement and delivery of an order. Type-$i$ valuation $v_i$ is a random variables with c.d.f $F_i$ and p.d.f $f_i$. Type-$i$ delay cost rate is a deterministic number $c_i$. We assume w.o.l.g. that $c_1 > c_2$. For impatient (type-1) customers, their service would expire if the lead time exceeds some certain amount $T$. This means if a impatient customer waits for more than

$T$, even if she gets served eventually, this service would create no value to her. In this sense, all impatient customers would leave after waiting for time $T$, which is considered as customer abandonments. The service time for two kinds of customers is the same, which is exponentially distributed with rate $\mu$.

Upon each customer's arrival, she can not observe the current number of customers waiting for service. Instead, she is given a static menu of two options, $(p_1, W_1)$ (class-1 option) and $(p_2, W_2)$ (class-2 option). Here, $p_k(k = 1, 2)$ is job price, which measures the money a customer pays for getting the service; $W_k(k = 1, 2)$ is the expected lead time. We assume $p_1 > p_2$ and $W_1 < W_2$, meaning that if a customer pays more for getting the service, her expected waiting time would be shorter.

In this model, we assume all customers seek to maximize their expected utility. We also assume them to be risk neutral with respect to lead time uncertainty. When a type-$i$ customer chooses a class-$k$ option, her utility is given as $v_i - c_iW_k - p_k$. Adapting the concept in Mechanism Design, we can restrict attention to menus that target class-$i$ service to type-$i$ customers, meaning that type 1 customers would prefer class-1 option, while type 2 customers would prefer class-2 option. Such menus must satisfy the incentive-compatibility (IC) constraints $v_i - c_iW_i - p_i \geq v_i - c_iW_j - p_j$, $i \neq j$. Under an IC menu, when $v_i - c_iW_i - p_i < 0$, a type-$i$ customer would not choose any option in the menu, meaning that she would leave immediately after she sees the menu. In this sense, the actual arrival rate for type-$i$ customers is $\lambda_i = \Lambda_i \cdot \overline{F}_i(c_iW_i + p_i)$, $i = 1, 2$, where $\overline{F}_i = 1 - F_i$.

For the service provider, she knows the arrival rate $\Lambda_i$, the value distributions $F_i$, delay cost rates $c_i$ and the service time distribution. Her goal is to design a price-lead time menu and scheduling policy to maximize her expected revenue. Since the expect revenue for the service provider is $\lambda_1 p_1 + \lambda_2 p_2$, her goal becomes maximizing $\lambda_1 p_1 + \lambda_2 p_2$ under some constraints. Recall that $W_1$ and $W_2$ are the announced expected waiting time, in order to ensure the credibility of the service provider, $W_i = w_i^s$ should be satisfied, where $w_i^s$ is the realized class-$i$ mean steady-state delay given a scheduling policy s. For the value of $W_1$, if $W_1 > T$, no type-1 customer will join the service system since their service expires after time $T$. Because there is no incentive for the service provider to keep high type customers away, $W_1 \leq T$ should be satisfied. As for $W_2$, there are two possibilities. If $W_2 > T$, there is no incentive for a type-1 customer to choose class-2 option. So IC-1 constraint is not necessary. While when $W_1 \leq T$, IC-1 constraint should be considered.

In summary, the service provider solves a two-scenario optimizing problem to maximize her expected revenue.

**Scenario** 1

$$\begin{aligned} \max \quad & \lambda_1 p_1 + \lambda_2 p_2 \\ s.t. \quad & \lambda_i = \Lambda_i \cdot \overline{F}_i(c_iW_i + p_i), \quad i = 1, 2 \\ & c_2W_2 + p_2 \leq c_iW_1 + p_1 \\ & \lambda_1 + \lambda_2 < \mu \\ & W_1 \leq T \\ & W_2 > T \\ & W_i = w_i^s, \quad i = 1, 2 \end{aligned}$$

**Scenario** 2

$$\begin{aligned}
\max \quad & \lambda_1 p_1 + \lambda_2 p_2 \\
s.t. \quad & \lambda_i = \Lambda_i \cdot \overline{F}_i(c_i W_i + p_i), \quad i = 1, 2 \\
& c_1 W_1 + p_1 \leq c_i W_2 + p_2 \\
& c_2 W_2 + p_2 \leq c_i W_1 + p_1 \\
& \lambda_1 + \lambda_2 < \mu \\
& W_1 \leq T \\
& W_2 \leq T \\
& W_i = w_i^s, \;\; i = 1, 2
\end{aligned}$$

## Strategic delay and modified formulation

For the optimization problem mentioned in previous subsection, we are trying to find the optimal scheduling rule $s$. Actually, we can transform the problem into a the simpler optimization problem of finding $(W_1, W_2)$ in the corresponding achievable region $OA \triangleq \{(w_1^s, w_2^s) : s \in A\}$, where $A$ is the set of admissible scheduling policies. The admissible scheduling policies can be work conserving rules, like First-In-First-Out (FIFO), Last-In-First-Out (LIFO) or some strategic delay rule which is mentioned in Afeche (2004). In order to do this transformation, we have the following lemma.

**Lemma 10.** *There exists a policy $s \in A$ such that $W_i = w_i^s$, $i = 1, 2$ if and only if*

$$W_1 \geq \frac{1}{\mu - \lambda_1 \cdot \frac{\mu T}{1+\mu T}} \tag{4.1}$$

$$W_2 \geq \frac{1}{\mu - \lambda_2} \tag{4.2}$$

$$\frac{\lambda_1 T}{1+\mu T} \cdot W_1 + \frac{\lambda_2}{\mu} \cdot W_2 \geq \frac{1}{\mu} \cdot \frac{\lambda_1 \cdot \frac{\mu T}{1+\mu T} + \lambda_2}{\mu - \lambda_1 \cdot \frac{\mu T}{1+\mu T} - \lambda_2} \tag{4.3}$$

With this lemma, the constraint $W_i = w_i^s$, $i = 1, 2$ can be expressed as the above three inequalities. Further, the original problem is optimizing over $p_k (k = 1, 2)$, in order to make the solution process easier, we will transform it into an optimization problem over $\lambda_i (k = 1, 2)$ by making the following definition

$$v_i(\lambda_i) \triangleq \overline{F}_i^{-1}(\frac{\lambda_i}{\Lambda_i}), \; i = 1, 2$$

In this sense, the objective function is expressed as $\sum_{i=1}^{2} \lambda_i (v_i(\lambda_i) - c_i W_i)$, and IC constraints become $\frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} \geq W_1$ and $\frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} \leq W_2$. Then, we obtain the following equivalent problem.

**Scenario** 1

$$
\begin{aligned}
\max \quad & \sum_{i=1}^{2} \lambda_i(v_i(\lambda_i) - c_i W_i) \\
s.t. \quad & 0 \leq \lambda_i < \Lambda_i \;\; i = 1, 2 \\
& \lambda_1 + \lambda_2 < \mu \\
& W_1 \geq \frac{1}{\mu - \lambda_1 \cdot \frac{\mu T}{1+\mu T}} \\
& W_2 \geq \frac{1}{\mu - \lambda_2} \\
& W_1 \leq T \\
& W_2 > T \\
& \frac{\lambda_1 T}{1+\mu T} \cdot W_1 + \frac{\lambda_2}{\mu} \cdot W_2 \geq \frac{1}{\mu} \cdot \frac{\lambda_1 \cdot \frac{\mu T}{1+\mu T} + \lambda_2}{\mu - \lambda_1 \cdot \frac{\mu T}{1+\mu T} - \lambda_2} \\
& \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} \leq W_2
\end{aligned}
$$

**Scenario** 2

$$
\begin{aligned}
\max \quad & \sum_{i=1}^{2} \lambda_i(v_i(\lambda_i) - c_i W_i) \\
s.t. \quad & 0 \leq \lambda_i < \Lambda_i \;\; i = 1, 2 \\
& \lambda_1 + \lambda_2 < \mu \\
& W_1 \geq \frac{1}{\mu - \lambda_1 \cdot \frac{\mu T}{1+\mu T}} \\
& W_2 \geq \frac{1}{\mu - \lambda_2} \\
& W_1 \leq T \\
& W_2 \leq T \\
& \frac{\lambda_1 T}{1+\mu T} \cdot W_1 + \frac{\lambda_2}{\mu} \cdot W_2 \geq \frac{1}{\mu} \cdot \frac{\lambda_1 \cdot \frac{\mu T}{1+\mu T} + \lambda_2}{\mu - \lambda_1 \cdot \frac{\mu T}{1+\mu T} - \lambda_2} \\
& \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} \geq W_1 \\
& \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} \leq W_2
\end{aligned}
$$

## 4.4 Structural results

For the above optimization problem, the decision variables are $\lambda_i$ and $W_i$, in order to find the optimal solution, we adapt a two-step methodology. Detailed approach is as follows.
**STEP 1** Treat $\lambda_i$ as known, determine the optimal lead time $W_i$ for scenario 1, then do the same thing for scenario 2.
**STEP 2** Based on the results in step 1, find the arrival rates $\lambda_i^1$ and lead time $W_i^1$ which maximize the revenue in scenario 1, define $\Pi_1 = \sum_{i=1}^{2} \lambda_i^1(v_i(\lambda_i^1) - c_i W_i^1)$. Also, find the optimal $\lambda_i^2$ and $W_i^2$ for scenario 2, define $\Pi_2 = \sum_{i=1}^{2} \lambda_i^2(v_i(\lambda_i^2) - c_i W_i^2)$.
**STEP 3** Compare the value of $\Pi_1$ and $\Pi_2$, see which scenario yields a higher revenue.

Before solving the above optimization problem, we first consider the so-called first-best problem which is the original problem without IC constraints. For the first-best problem, we have the following Proposition.

**Proposition 11.** *For fixed $\lambda_1$ and $\lambda_2$, when IC constraints are not considered, the optimal solution (first-best optimal) for scenario 1 is*

$$\begin{aligned} W_1^{c\mu-1} &= \frac{1}{\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}} \\ W_2^{c\mu-1} &= \frac{\mu}{(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T})(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}-\lambda_2)}. \end{aligned}$$

*While the first-best optimal for scenario 2 is*

$$\begin{aligned} W_1^{c\mu-2} &= \frac{1}{\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}} \\ W_2^{c\mu-2} &= T \end{aligned}$$

From this proposition, we can see that if we don't consider IC constraints, the revenue maximization problems are exactly the same as delay cost minimization problems for both scenario 1 and 2. In this sense, the work conserving preemptive $c\mu$ scheduling rule is optimal for the first best problems of two scenarios, meaning that giving absolute priority to impatient customers (since they have higher unit waiting cost so that the $c\mu$ index for them is higher) produces highest revenue for the service provider if IC constraints are not considered.

If we incorporate the IC constraints, the optimization problems for both scenarios will be much more complicated. For the optimality of the second-best problems (first best problems with IC constraints, which is also the original problem), we have the following Proposition.

**Proposition 12.** *For fixed $\lambda_1$ and $\lambda_2$, the second-best optimal lead time is given as follows.*

1. *When $\frac{\mu}{(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T})(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}-\lambda_2)} < T$ and $\frac{v_1(\lambda_1)-v_2(\lambda_2)}{c_1-c_2} > \frac{\mu}{(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T})(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}-\lambda_2)}$, the second-best optimal lead time is*

$$W_1^{SD-2} = \frac{1}{\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}}, W_2^{SD-2} = \frac{\mu}{(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T})(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}-\lambda_2)}.$$

2. *When $\frac{\mu}{(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T})(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}-\lambda_2)} \geq T$ and $\frac{v_1(\lambda_1)-v_2(\lambda_2)}{c_1-c_2} > \frac{1}{\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}}$, the second-best optimal lead time is*

$$W_1^{SD-1} = \frac{1}{\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}}, W_2^{SD-1} = \frac{v_1(\lambda_1)-v_2(\lambda_2)}{c_1-c_2}.$$

3. *When $\frac{\mu}{(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T})(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}-\lambda_2)} < T$ and $\frac{v_1(\lambda_1)-v_2(\lambda_2)}{c_1-c_2} \leq \frac{\mu}{(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T})(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}-\lambda_2)}$, the second-best optimal lead time is*

$$W_1^{c\mu-1} = \frac{1}{\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}}, W_2^{c\mu-1} = \frac{\mu}{(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T})(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}-\lambda_2)}.$$

4. *When* $\frac{\mu}{(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T})(\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}-\lambda_2)} \geq T$ *and* $\frac{v_1(\lambda_1)-v_2(\lambda_2)}{c_1-c_2} \leq \frac{1}{\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}}$*, the second-best optimal lead time is*

$$W_1^{c\mu-2} = \frac{1}{\mu-\lambda_1\cdot\frac{\mu T}{1+\mu T}}, W_2^{c\mu-2} = T.$$

5. *When none of the condition 1-4 holds, there is no optimal solution.*

For the above proposition, $W_i^{SD-1}$ $W_i^{SD-2}$ correspond to some scheduling policy which is not work conserving. So, in some situations, it is optimal for the service provider to delay the service of patient customers intentionally, which is called *strategic delay* (see Afeche (2004)). Although by adapting those strategic delay scheduling policies, the aggregate delay cost rate will increase, impatient customers would have more incentives to buy the premium service, which gives extra revenue for the service provider.

## 4.5 Summary

In this chapter, we study a revenue management problem for a queueing system with two types of customers which have different valuation of service and different unit waiting time. The key difference between this chapter and previous research papers is that we assume the service for impatient customers would expire after time $T$. This brings customer abandonments to our model. By using a mechanism design framework and adapting the achievable region approach, we transform the scheduling control problem into a nonlinear program. Then through a two-stage optimization technique, we finally identify the optimal price-lead time manus and scheduling policies for different scenarios. We also discuss how the optimality conditions for strategic delay change in the setting with customers renege. In this sense, this chapter is an extension for Afeche (2010) by introducing customer abandonments.

# Appendix A

# Proofs

**Proof of Lemma 1.** Define: $x = L\mu + \lambda$, $a_i = \prod\limits_{j=1}^{i} \frac{1}{\nu_i+\gamma_i}$, $b_i = \beta\gamma(i-c)$. Then, (6) reduces to:

$$x - \lambda = \beta\gamma \sum_{i=c}^{N} (i-c) \frac{a_i x^i}{1 + \sum\limits_{i=1}^{N} a_i x^i}.$$

Making a transform on the above equation, we have:

$$-\lambda + (1-\lambda a_1)x + (a_2 - \lambda a_3)x^2 + ... + (a_{c-1} - \lambda a_c - b_c)x^c + ... + (a_{N-1} - \lambda a_N - b_N)x^N + a_N x^{N+1} = 0.$$

For the above equation, it is easy to see that the number of variations of sign in the coefficient sequence of $x^i$ is 1. So, by extended Descartes' rule of signs, there will be a unique positive root of $x$, which means that the positive root for L will also be unique. □

**Proof of Proposition 2.** Parts 1,2 and 5 are already derived in the main body of section 3. Here, we only prove parts 3 and 4.
For part 3, Let $EQ$ be the average number of customers waiting in the main queue. Apply the Little's Law to the main queue, we have:

$$W = \frac{EQ}{L\mu + \lambda} = \frac{1}{L\mu + \lambda} \sum_{i=c}^{N} (i-c) P_i = \frac{1}{L\mu + \lambda} \sum_{i=c}^{N} \frac{(i-c) \prod\limits_{j=1}^{i} \frac{L\mu+\lambda}{\nu_i+\gamma_i}}{1 + \sum\limits_{i=1}^{N} (\prod\limits_{j=1}^{i} \frac{L\mu+\lambda}{\nu_i+\gamma_i})}.$$

For part 4, in the long run, the proportion of customers getting served equals to the outgoing rate from the server over the incoming rate to the service node, so we have:

$$P(s) = \frac{\nu}{\lambda} [\sum_{i=0}^{c} \frac{i \prod\limits_{j=1}^{i} \frac{L\mu+\lambda}{\nu_i+\gamma_i}}{1 + \sum\limits_{i=1}^{N} (\prod\limits_{j=1}^{i} \frac{L\mu+\lambda}{\nu_i+\gamma_i})} + \sum_{i=c+1}^{N} \frac{c \prod\limits_{j=1}^{i} \frac{L\mu+\lambda}{\nu_i+\gamma_i}}{1 + \sum\limits_{i=1}^{N} (\prod\limits_{j=1}^{i} \frac{L\mu+\lambda}{\nu_i+\gamma_i})} ]. \square \tag{A.1}$$

**Proof of Lemma 3.** For MO, M1 and M2, we have:

$$(L_{i-1}\mu + \lambda)P_{i-1} = (\nu_i + \gamma_i)P_i \quad (MO),$$

$$\lambda P_{i-1} = (\nu_i + \gamma_i)P_i \qquad (M1),$$

$$\lambda P_{i-1} = (\nu_i + (1-\beta)\gamma_i)P_i \qquad (M2),$$

where $P_i$ is the stationary probability that there are $i$ customers in the service node. Also, we have: $L_{i-1}\mu = \beta\gamma_i$. So, we can easily get the following result: $P_i/P_{i-1}(M1) \leq P_i/P_{i-1}(MO) \leq P_i/P_{i-1}(M2)$. Further, we know that the average waiting time is an increasing function with respect to $P_i/P_{i-1}$. Consequently, we have:

$$W_{M1} \leqslant W_{MO} \leqslant W_{M2}.$$

Similarly, we can prove $W_{M1} \leqslant W_{MR} \leqslant W_{M2}$. So, we have: $|W_{MR} - W_{MO}| \leqslant W_{M2} - W_{M1}$. □

**Proof of Proposition 4.** Take part 1 as an example, proofs for other parts can be done in the same way. It is obvious that for the original model, a higher customer arrival rate will result in a higher average customer waiting time. So, the average customer waiting time $W_{MO}$ is increasing with respect to $\lambda$. For $W_{MR}$, we have: $(L\mu + \lambda)P_{i-1} = (\nu_i + \gamma_i)P_i$, and we know that $W_{MR}$ is an increasing function with respect to $P_i/P_{i-1}$. So, we only need to show that the value of $L$ increases with respect to $\lambda$.

Here, we are using the counter example to prove that $L$ is increasing with respect to $\lambda$. Suppose the value of $L$ does not change with respect to $\lambda$, we can conclude that the incoming flow to the orbit is greater than the outgoing flow. So, the orbit can not be in steady state, which makes a contradiction. Therefore, the value of $L$ can not keep the same. Then, suppose the value of $L$ decreases when $\lambda$ increases, we can conclude that the outgoing flow from the orbit is greater than its incoming flow, so the orbit can not be in steady state also, which makes another contradiction. Consequently, the value of $L$ can not decrease. Based on the above two claims, we can see that the value of $L$ is an increasing function with respect to $\lambda$. Then, we can get the conclusion that $W_{MR}$ is an increasing function with respect to $\lambda$ directly. □

**Proof of Lemma 5.** From Brill and Posner (1981), Brill (1979) and Brill and Posner (1977), the rate of downcrossing of level $x$ is $f_1(x)$. For the upcrossing of level $x$, there are four situations in total. The first situation is due to customers who find only one server idle upon arrival and initiate the active phase. In order for the virtual waiting time to upcross $x$, the time until the first service completion should be greater than $x$. Thus, such an upcrossing occurs with rate $(\lambda_1 + \lambda_2 + L\mu)P_{c-1}e^{-c\nu x}$.

The second situation is from customers who are patient enough to reach servers. The rate of such upcrossing is $(\lambda_1 + L\mu)\int_0^x e^{-c\nu(x-y)}e^{-\gamma y}f_1(y)dy$. In the third situation, consider one type-II customer arrives during an active phase, since type II customers will not increase the waiting time for type-I customer until they seize the first server becoming idle from the call back queue. In this situation, the next type-I customer' waiting time will be greater than $x$ only if the first service completion time for $c$ servers is greater than $x$. So such upcrossing rate is $\lambda_2 P(AP)e^{-c\nu x}$. The last situation involves customers initially joining queue I who later get inpatient and join queue II. Assume that the virtual waiting time upon this customer's arrival is at level $y$, we can find that the upcrossing rate is given by $(\lambda_1 + L\mu)(1-\beta)e^{-c\nu x}\int_0^x (1-e^{-\gamma y})f_1(y)dy$.

By equating the rate of downcrossings and the rate of upcrossings, we get

$$\begin{aligned} f_1(x) = {} & (\lambda_1 + \lambda_2 + L\mu)P_{c-1}e^{-c\nu x} + (\lambda_1 + L\mu)\int_0^x e^{-c\nu(x-y)}e^{-\gamma y}f_1(y)dy + \lambda_2 P(AP)e^{-c\nu x} \\ & + (\lambda_1 + L\mu)(1-\beta)e^{-c\nu x}\int_0^x (1-e^{-\gamma y})f_1(y)dy. \end{aligned}$$

For the normalizing condition, we have:

$$\sum_{j=0}^{c-1} P_j + \int_0^\infty f_1(y)dy = 1.$$

As we know that the number of busy servers is birth-death process, so the value of $P_j$ can be stated in terms of $P_0$. Let $\rho = (\lambda_1 + \lambda_2 + L\mu)/\nu$, the normalizing condition can be reduced to :

$$\sum_{j=0}^{c-1} \frac{\rho^j}{j!}P_0 + \int_0^\infty f_1(y)dy = 1. \square$$

**Calculation of $f_1(x)$ and $P(AP)$.** Taking the derivative of both sides in (2.13) with respect to $x$, we can get the first order differential equation:

$$\begin{aligned} & f_1'(x) = (\lambda_1 + L\mu)e^{-\gamma x}f_1(x) - c\nu[(\lambda_1 + \lambda_2 + L\mu)P_{c-1}e^{-c\nu x} + \lambda_2 P(AP)e^{-c\nu x} \\ & +(\lambda_1 + L\mu)\int_0^x e^{-c\nu(x-y)}e^{-\gamma y}f_1(y)dy + (\lambda_1 + L\mu)(1-\beta)e^{-c\nu x}\int_0^x (1-e^{-\gamma y})f_1(y)dy]. \end{aligned}$$

From equation (2.13), the above differential equation reduces to:

$$f_1'(x) = (\lambda_1 + L\mu)e^{-\gamma x}f_1(x) - c\nu f_1(x).$$

Solving the above equation, we get

$$f_1(x) = Ke^{\frac{\lambda_1 + L\mu}{\gamma}(1-e^{-\gamma x}) - c\nu x}, \tag{A.2}$$

where $K$ is a constant. In order to get the value of $K$, take Laplace transform on both sides of equation (2.13), we have

$$\begin{aligned} \widetilde{f}_1(s) = {} & \tfrac{1}{c\nu + s}[(\lambda_1 + \lambda_2 + L\mu)P_{c-1} + (\lambda_1 + L\mu)\widetilde{f}_1(s+\gamma) + \lambda_2 P(AP) \\ & + (\lambda_1 + L\mu)(1-\beta)(P(AP) - \widetilde{f}_1(\gamma))]. \end{aligned}$$

And we know $\lim_{s\to 0} \widetilde{f_1}(s) = P(AP)$, from the above equation, we can obtain

$$\widetilde{f_1}(\gamma) = \frac{-(\lambda_1+\lambda_2+L\mu)P_{c-1} + P(AP)(c\nu - \lambda_1 - \lambda_2 - L\mu + (\lambda_1 + L\mu)\beta)}{(\lambda_1 + L\mu)\beta}. \tag{A.3}$$

Setting $x = 0$ in (2.13) and (2) gives that

$$K = (\lambda_1+\lambda_2+L\mu)P_{c-1} + \lambda_2 P(AP) + (\lambda_1 + L\mu)(1-\beta)(P(AP) - \widetilde{f_1}(\gamma)). \tag{A.4}$$

From the normalizing condition and equation (2), (4), we can obtain

$$P_0 = \frac{1 - U[\lambda_1 + L\mu + \lambda_2 - \beta(\lambda_1 + L\mu) - \frac{1-\beta}{\beta}(c\nu - (\lambda_1 + L\mu + \lambda_2 - \beta(\lambda_1 + L\mu)))]}{UX + \sum_{j=0}^{c-1} \frac{\rho^j}{j!}},$$

where

$$U = \int_0^\infty e^{\frac{\lambda_1 + L\mu}{\gamma}(1 - e^{-\gamma x}) - c\nu x} dx,$$

and

$$X = \frac{(\lambda_1 + \lambda_2 + L\mu)\rho^{c-1}}{\beta(c-1)!} + \frac{(1-\beta)c\nu - (\lambda_1 + L\mu + \lambda_2 - \beta(\lambda_1 + L\mu)}{\beta} \sum_{j=0}^{c-1} \frac{\rho^j}{j!}.$$

With the value of $P_0$, we can compute $K$ and $P(AP)$ and obtain $f_1(x)$ with no difficulty. □

**Proof of Proposition 8** The average number of customers in the service node is

$$E(Q) = \sum_{i=c}^{K} (i-c)P_i = \sum_{i=c}^{K} \frac{(i-c)\sum\limits_{S_i} \prod\limits_{j=i}^{K} x(i + \sum\limits_{t=i}^{j} s_t, s_j)}{\sum\limits_{i=0}^{K} \sum\limits_{S_i} \prod\limits_{j=i}^{K} x(i + \sum\limits_{t=i}^{j} s_t, s_j)}.$$

For the average waiting time, applying Little's Law to the service node, we have

$$E(w) = \frac{E}{L\mu + \lambda_1} = \frac{1}{L\mu + \lambda_1} \sum_{i=c}^{K} \frac{(i-c)\sum\limits_{S_i} \prod\limits_{j=i}^{K} x(i + \sum\limits_{t=i}^{j} s_t, s_j)}{\sum\limits_{i=0}^{K} \sum\limits_{S_i} \prod\limits_{j=i}^{K} x(i + \sum\limits_{t=i}^{j} s_t, s_j)}.$$

In the long run, the proportion of customers getting blocked equals to the probability that the service node is in state $K$, so

$$P(b) = P_K = \frac{1}{\sum\limits_{i=0}^{K} \sum\limits_{S_i} \prod\limits_{j=i}^{K} x(i + \sum\limits_{t=i}^{j} s_t, s_j)}. \square$$

**Proof of Proposition 9** From the properties of probability generating function, the stationary probabilities can be calculated by the probability generating function.

$$P_k = \frac{G^{(k)}(0)}{k!}.$$

The probability that all servers are idle is

$$P_0 = G(0) = (\frac{\nu\omega_1 - (\lambda_1 + L\mu)(\omega_1 + \omega_2 + \theta) + \gamma\lambda_2}{\nu\omega_1 - (\lambda_1 + L\mu)\omega_2})^{\frac{\omega_1+\theta}{\mu}} \cdot \frac{\nu\omega_1 + \gamma\lambda_2}{(\omega_1 + \omega_2)^2 + (\omega_1 + \omega_2)\alpha\nu}.$$

Also, the average number of customers in the service node is

$$N = G'(1) = \frac{\nu\theta(\theta + \alpha\nu) + (\lambda_1 + L\mu)(\omega_1 + \omega_2 + \theta)^2 + \gamma\lambda_2}{\nu(\omega_1 + \theta)[(\omega_1 + \omega_2)^2 + (\omega_1 + \omega_2)\theta - (\lambda_1 + L\mu)(\omega_2 + \theta)]}$$

Since the average number of customers waiting equals the average number of customers in the system minus the average number of customers in the server, we have

$$Q = N - 1 \cdot P_1 = \frac{\nu\theta(\theta + \alpha\nu) + (\lambda_1 + L\mu)(\omega_1 + \omega_2 + \theta)^2 + \gamma\lambda_2}{\nu(\omega_1 + \theta)[(\omega_1 + \omega_2)^2 + (\omega_1 + \omega_2)\theta - (\lambda_1 + L\mu)(\omega_2 + \theta)]}. \square$$

# Bibliography

Afeche, P. (2004). Incentive-compatible revenue management in queueing systems: Optimal strategic delay and other delaying tactics. Technical report, Working paper.

Aguir, S., F. Karaesmen, O. Akşin, and F. Chauvet (2004). The impact of retrials on call center performance. *OR Spectrum 26*(3), 353–376.

Akan, M., B. ş Ata, and T. Olsen (2012). Congestion-based lead-time quotation for heterogenous customers with convex-concave delay costs: Optimality of a cost-balancing policy based on convex hull functions. *Operations Research 60*(6), 1505–1519.

Artalejo, J. (1995). A queueing system with returning customers and waiting line. *Operations Research Letters 17*(4), 191–199.

Artalejo, J. (1999). A classified bibliography of research on retrial queues: progress in 1990–1999. *Top 7*(2), 187–211.

Artalejo, J. (2000). G-networks: A versatile approach for work removal in queueing networks. *European Journal of Operational Research 126*(2), 233–249.

Artalejo, J. (2010). Accessible bibliography on retrial queues: progress in 2000-2009. *Mathematical and Computer Modelling 51*(9-10), 1071–1081.

Artalejo, J. and A. Gómez-Corral (2008). *Retrial queueing systems: a computational approach.* Springer Verlag.

Artalejo, J., D. Orlovsky, and A. Dudin (2005). Multi-server retrial model with variable number of active servers. *Computers & Industrial Engineering 48*(2), 273–288.

Atencia, I. and P. Moreno (2004). The discrete-time Geo/Geo/1 queue with negative customers and disasters. *Computers & Operations Research 31*(9), 1537–1548.

Brandt, A. and M. Brandt (2002). Asymptotic Results and a Markovian Approximation for the M (n)/M (n)/s+GI System. *Queueing Systems 41*(1), 73–94.

Brandt, A. and M. Brandt (2004). On the two-class M/M/1 system under preemptive resume and impatience of the prioritized customers. *Queueing Systems 47*(1), 147–168.

Brill, P. (1979). An embedded level crossing technique for dams and queues. *Journal of Applied Probability*, 174–186.

Brill, P. and M. Posner (1977). Level crossings in point processes applied to queues: single-server case. *Operations Research 25*(4), 662–674.

Brill, P. and M. Posner (1981). The system point method in exponential queues: a level crossing approach. *Mathematics of Operations Research*, 31–49.

Çelik, S. and C. Maglaras (2008). Dynamic pricing and lead-time quotation for a multiclass make-to-order queue. *Management Science 54*(6), 1132–1146.

Chae, K., H. Park, and W. Yang (2010). A GI/Geo/1 queue with negative and positive customers. *Applied Mathematical Modelling 34*(6), 1662–1671.

Choi, B., B. Kim, and J. Chung (2001). M/M/1 queue with impatient customers of higher priority. *Queueing systems 38*(1), 49–66.

Choudhury, G. and L. Tadj (2009). An M/G/1 queue with two phases of service subject to the server breakdown and delayed repair. *Applied Mathematical Modelling 33*(6), 2699–2709.

Coffman, E. and I. Mitrani (1980). A characterization of waiting time performance realizable by single-server queues. *Operations Research 28*(3-Part-II), 810–821.

Dana Jr, J. D. and T. Yahalom (2008). Price discrimination with a resource constraint. *Economics Letters 100*(3), 330–332.

Falin, G. (1983). Calculation of probability characteristics of a multiline system with repeat calls. *Moscow University Computational Mathematics and Cybernetics 1*, 43–49.

Falin, G. and J. Templeton (1997). *Retrial queues*, Volume 75. Chapman & Hall/CRC.

Green, T. C. and S. Stidham (2000). Sample-path conservation laws, with applications to scheduling queues and fluid systems. *Queueing Systems 36*(1-3), 175–199.

Ha, A. Y. (2001). Optimal pricing that coordinates queues with customer-chosen service requirements. *Management Science 47*(7), 915–930.

Hassin, R. J. and M. Haviv (2003). *To Queue or not to Queue: Equilibrium behavior in queuing systems*, Volume 59. Kluwer Academic Pub.

Hsu, V. N., S. H. Xu, and B. Jukic (2009). Optimal scheduling and incentive compatible pricing for a service system with quality of service guarantees. *Manufacturing & Service Operations Management 11*(3), 375–396.

Iravani, F. and B. Balcıoglu (2008). On priority queues with impatient customers. *Queueing Systems 58*(4), 239–260.

Ke, J. (2006). On M/G/1 system under NT policies with breakdowns, startup and closedown. *Applied Mathematical Modelling 30*(1), 49–66.

Li, Q., Y. Ying, and Y. Zhao (2006). A BMAP/G/1 retrial queue with a server subject to breakdowns and repairs. *Annals of Operations Research 141*(1), 233–270.

Maglaras, C. (2006). Revenue management for a multiclass single-server queue via a fluid model analysis. *Operations Research 54*(5), 914–932.

Mandelbaum, A., W. Massey, M. Reiman, A. Stolyar, and B. Rider (2002). Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems 21*(2), 149–171.

Movaghar, A. (1998). On queueing with customer impatience until the beginning of service. *Queueing Systems 29*(2), 337–350.

Mussa, M. and S. Rosen (1978). Monopoly and product quality. *Journal of Economic Theory 18*(2), 301–317.

Neuts, M. and B. Rao (1990). Numerical investigation of a multiserver retrial model. *Queueing systems 7*(2), 169–189.

Park, H., W. Yang, and K. Chae (2009). The Geo/G/1 queue with negative customers and disasters. *Stochastic Models 25*(4), 673–688.

Rochet, J.-C. and P. Choné (1998). Ironing, sweeping, and multidimensional screening. *Econometrica*, 783–826.

Shanthikumar, J. G. and D. D. Yao (1992). Multiclass queueing systems: Polymatroidal structure and optimal scheduling control. *Operations Research 40*(3-Supplement-2), S293–S299.

Sherman, N. and J. Kharoufeh (2006). An M/M/1 retrial queue with unreliable server. *Operations Research Letters 34*(6), 697–705.

Takine, T. and B. Sengupta (1997). A single server queue with service interruptions. *Queueing Systems 26*(3), 285–300.

Van Mieghem, J. A. (2000). Price and service discrimination in queuing systems: incentive compatibility of gc$\mu$ scheduling. *Management Science 46*(9), 1249–1267.

Whitt, W. (2006). Sensitivity of performance in the Erlang-A queueing model to changes in the model parameters. *Operations research 54*(2), 14–18.

Zeltyn, S. and A. Mandelbaum (2005). Call centers with impatient customers: many-server asymptotics of the M/M/n+ G queue. *Queueing Systems 51*(3), 361–402.