PERFORMANCE ANALYSIS AND OPTIMAL STAFFING OF TICKET QUEUES

LI XIAO

NATIONAL UNIVERSITY OF SINGAPORE

2015

PERFORMANCE ANALYSIS AND OPTIMAL STAFFING OF TICKET QUEUES

LI XIAO

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF DECISION SCIENCES

NATIONAL UNIVERSITY OF SINGAPORE

2015

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Li Xiao

Li Xiao

ACKNOWLEDGMENT

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Hanqin Zhang, for his constant support, guidance, and encouragement in my PhD journey. I feel extremely lucky to have this opportunity to work with him.

I am immensely indebted to Professor Susan Hong Xu and Professor David D. Yao. Without their professional guidance and advice, valuable suggestions, and critical comments, I could not complete this thesis. In particular, I am deeply grateful to Professor Susan Hong Xu, who passed away last year. Her rigorous attitude and passion for research will constantly inspire me in the future. I am also quite privileged to work with Professor Jeannette Song and Professor Paul H. Zipkin. Their rigorous attitude and tireless guidance are greatly valuable to me.

I would like to thank my thesis committee members, Professor Shuangchi He, Professors Jussi Keppo, and Professor Xueming Yuan, for their invaluable suggestions.

I learn a lot from many outstanding professors in Decision Science Department. I would like to thank Professor Chen Gongtao Lucy, Professor Chou Cheng Feng Mabel, Professor Lim Andrew, Professor Sim Melvyn, Professor Sun Jie, Professor Teo Chung Piaw, Professor Wang Tong, and

Professor Wu YaoZhong. I also would like to thank Professor Ang Soo Keng James, Professor Chou Fee Seng, and Professor Hum Sin Hoon, for giving me opportunities to work with them and experience various teaching duties.

I also would like to thank my friends in Decision Science, among them are: Qingxia Kong, Jin Qi, Zhuoyu Long, Xuchuan Yuan, Zhichao Zheng, Junfei Huang, Meilin Zhang, Rohit Nishant, Jeremy Chen, Zhi Chen, Sheng Zhao, Weijia Gu, Baiyu Li, Yini Gao, Shasha Han, Zhenzhen Yan, Guodong Lyu and Qinshen Tang.

Finally, I would like to thank my parents for their unconditional love. I also would like to thank my husband, Jing Xie, for his encouragement and support.

May, 2015

Singapore

Li XIAO

ABSTRACT

Ticket queues are popular in many service systems. Upon arrival, each customer is issued a numbered ticket and receives service on a first-come-first-served basis according to the ticket number. There is no physical queue; customers may choose to walk away and return later (before their numbers are called) to receive service. In this thesis, we study the problem of optimal staffing in such a system, where the staffing decision can only be based on ticket numbers, as opposed to the physical queue length in a traditional system. The thesis consists of two parts.

In the first part, we consider the system with two staffing levels (low and high). Using the renewal reward theorem, we first derive the long-run average cost (including customer delay and abandonment costs, server operating cost and cost for changing staffing levels), and then obtain the optimal staffing policy using the fractional programming. Moreover, with the help of random walk theory, we develop some approximations for the system performance measures, and then establish the asymptotical optimal staffing policy. The extensive numerical experiments show the asymptotical optimal policies perform very well.

The second part is devoted to the analysis of the system with more than two staffing levels. We use the fluid approximation approach to analyze the

system dynamics under the assumption that the customer arrival rate and service rate are very high. The optimal staffing policy for the fluid ticket queueing model can be determined by the optimal solution of EOQ model. Moreover, this optimal staffing policy for the fluid ticket queueing model is proved to be asymptotically optimal for our original ticket queue in the sense that its long-run average cost achieves the asymptotical lower bound.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Motivation

Ticket queues appear in hospitals, banks, retail stores, theme parks, government agencies, and many other service systems. Upon arrival, each customer is issued a numbered ticket. The ticket numbers are then called out in sequence whenever service becomes available, and the ticket holders receive service accordingly. Ticket queues have also been implemented in certain online services. One example is Dell's Internet customer service. The system issues each customer upon login a number and provides service following the natural (increasing) order of the numbers.

Compared with traditional queueing systems where there is a physical queue, ticket queues have many apparent advantages. Customers are freed from the physical discomfort of having to stand and wait in crowded queues. In fact, they have the option to walk away and return later (before their numbers are called) for service so as to make more productive usage of their waiting time. From the service provider's perspective, the absence of a physical queue reduces the pressure to provide adequate space capacity, alleviates over-crowding related problems, and makes it easier to manage the waiting area and customer flow.

On the other hand, ticket queues have a disadvantage of invisibility of abandonment customers to the system managers, as some ticket-holding customers may not return for their service (on time or at all). This disadvantage prevents system managers from obtaining information about exact queue lengths. However, for traditional queues, the more practical and widely adopted staffing policy is congestion-based staffing, where the number of servers is adjusted according to the queue length and current service level. Continuing along this line, the disadvantage in ticket queues makes the problem of optimal staffing more difficult for the system manager, since ticket number is the only information available to the managers. We study the problem of optimal staffing in such systems, where the staffing decision can only be based on ticket numbers, as opposed to the physical queue length in a traditional system.

Let's pursue the customer abandonment issue a bit further. Suppose the system manager records and updates the number of customers who are in service or have already received service up to time t, denoted $c(t)$. There are two other numbers the manager has ready access to: the last ticket number taken before t by an arriving customer, denoted $a(t)$; and the last ticket number called out (for a waiting customer to receive service) before t, denoted $b(t)$. We must have $a(t) \geq b(t) \geq c(t)$. Note that $a(t) - b(t)$ is the ticket queue-length, those waiting for service in the invisible ticket queue. The catch is, not every customer in the ticket queue may be present (at time t); indeed some may have walked away, and some may choose never to return; and $b(t) - c(t)$ exactly captures the number of no-show customers (cumulative up to time t). Thus, the system manager can use the ratio,

$(b(t) - c(t))/b(t)$, to estimate the abandonment rate of customers. In fact, realistically, the abandonment rate will depend on the number of servers in action — when customers observe more servers actively serving in the system they are less likely to walk away. Indeed, by associating the aforementioned ratio with the number of servers serving at time t, the manager can come up with estimates on abandonment rates that are server-dependent. Applying the abandonment rates to the ticket queue-length, $a(t) - b(t)$, the manager will have a solid grasp on the actual congestion level in the system, and will make staffing decision accordingly. This, in a nutshell, is the kind of staffing rule that we shall study in this paper.

Specifically, we will assume that the server-dependent abandonment rates are given; that is, we de-couple the estimation problem from the staffing problem, and focus on the latter instead. (Otherwise, the problem will be more complex, with the two problems, control and learning, intertwined; i.e., making staffing decisions while updating the estimates on customer abandonment rates.) Two types of costs are considered in our staffing decision: customer-related abandonment and delay costs, and service-related operating and changeover costs. (The last one refers to the cost associated with changing staffing levels.)

1.2 Literature Review

Existing studies on the optimal staffing for traditional queues can be classified into two categories. The first category assumes no customer abandonment. Yadin and Naor [36] investigate how to determine the service rate, based on the queue length, so as to minimize the long-run average total cost, including the operating cost, customer delay cost, and the service rate changeover cost. Using the Markov decision theory, Bell [7], and Gans and Zhou [14] consider multiserver queues and characterize the optimal policies of adjusting the number of working servers. When customer arrival rates change over time, Fu et al. [13] study the optimal staffing policy for a multiserver system with transient queueing effects. More recently, Zhang [38] studies the tradeoff between the expected queue length and the frequency of service capacity changeover. When the system has two capacity levels (low and high), the author develops fluid and diffusion approximations for the expected queue length, and then numerically illustrates the accuracy of these approximations and the effectiveness of the congestion-based staffing policy. If there is no customer abandonment, the ticket queue studied in this chapter will reduce to the traditional queue, and the optimal staffing problem studied in Zhang [38] will apply.

The second category takes into account customer abandonment in staffing decisions. Harrison and Zeevi [20] use fluid approximations to optimize the trade-off between the system cost and customer abandonment penalties for call centers with multiple customer classes and multiple server pools. Using diffusion approximation, the square-root staffing rule is studied by Garnett et al. [15] and Mandelbaum and Zeltyn [25] with/without constraints on the fraction of abandoning customers, average waiting time, and the probability of service delay. The study is recently refined by Zhang et al. [37], and extended by Pang and Perry [28] to different large-scale systems. When the abandonment and reneging probabilities are increasing and concave functions of the number of customers in the system (queue length), Armony et al. [4] establish certain properties of the queue length and abandonment process with respect to the service capacity, and then analyze the sensitivity of the optimal service capacity.

There is a rich body of literature on customer abandonment in traditional queueing systems. The earlier focus was on performance evaluation of queues with impatient customers; refer to Cox and Smith [12], Ancker and Gafarian (1962a, b), and Reynolds [29]. Later, Baccelli et al. [6], Gnedenko and Kovalenko [17], and Stanford [32] consider single-server queues with customer abandonment depending on the waiting time. Furthermore, the similar problem of customer abandonment depending on waiting-plus-service time is investigated by Gavish and Schweitzer [16], Hokstad [22], and Van Dijk [34]. More recently, Brown et al. [11], Mandelbaum and Shimkin [24], and Zohar et al. [39] develop statistical methods to estimate customer patience times. In ticket queues, information such as queue lengths, waiting times, and abandonment epochs in traditional queues becomes unavailable. Thus the methods reviewed in the literature above for characterizing customer abandonment behavior are not applicable. This explains why in this thesis we choose not to model directly customer abandonment behavior; instead, we base our staffing decision on the customer abandonment rate as observed by the system manager; and this parameter, as motivated earlier, is readily estimated by the ticket counts (along with a count of customers served and in service).

Specifically, our model is also related to the literature on the hysteretic optimal control in $M/M/1$ queue, where the change in service rate incurs set-up cost. In this study, it's not desirable to assign a service rate to a given queue length because of the set-up cost. The optimal value of service rate at any moment depends on previous history of the system, such as the queue length and previous service rate. Yadin and Naor [36] derive the stationary distribution of queue length given one hysteretic policy. Later, Lu and Serfozo [23] and Kitaev and Serfozo [19] build a Markov decision process and show that the optimal policy indeed is hysteretic policy, assuming that cost function are submodular and satisfy some additional technical conditions. Blackburn [10] takes into account customer balking and renege, and considers controlling an $M/M/1$ queue by turning the server on and off. Bell [8] study an $M/M/2$ queue with removable servers. Both Blackburn [10] and Bell [8] establish that optimal policy has hysteretic property, but they can only open or shut down servers instead of choosing service rate. Compared with existing literature, our study consider a more general problem and finds the asymptotic optimal policy.

To our knowledge, there are two papers studying ticket queues. The paper by Xu et al. [35] pursues an analytical study on ticket queues, where a single-server model is considered. A Markov chain analysis leads to the equilibrium distribution of the number of tickets in the system, along with numerical methods for performance evaluation. The analysis there shows the difficulties involved in deriving the analytic expressions for ticket queues, even just for a single-server model and without staffing control. Thus, the complexity in our model should come off as no surprise. Another paper by Jennings and Pender [18] compare ticket queueing system and standard queueing system. They conclude that the ticket queue and standard queue will perform asymptotically identically under heavy traffic condition.

1.3 Structure of the Thesis

In Chapter 2, we consider the system with two staffing levels (low and high). Using the renewal reward theorem and matrix analytic methods, we first derive the long-run average cost (including customer delay and abandonment costs, operating cost and cost for changing staffing levels), and then obtain the optimal staffing policy by the fractional programming. Moreover, with the help of random walk theory, we develop some approximations for the system performance measures, and then establish the asymptotical optimal staffing policy. The extensive numerical experiments show the asymptotical optimal policies perform very well.

In Chapter 3, we consider the system with more than two staffing levels. It is almost impossible to write an analytic expression of the long-run average cost. Instead, we use the fluid approximation approach to analyze the system dynamics under the assumption that the customer arrival rate and service rate are very large. After building the corresponding fluid model for ticket queues, we establish a connection between it and the EOQ model in inventory management. The optimal staffing policy for the fluid ticket queueing model can then be determined by the optimal solution of EOQ model. Moreover, the optimal staffing policy for the fluid ticket queueing model is proved to be asymptotically optimal for our original ticket queue.

In Chapter 4, we discuss several future research problems.

In Appendix, we derive the long-run average cost of the system with multiple servers and two staffing levels.

1.4 Notation

The following notation will be used throughout this thesis. $\mathsf{Pr}(A)$ denotes the probability of event A, I_A denotes the indicator of the event A, E denotes the expectation operator, and Var denotes the variance operator. For any real number x, let $x^+ = \max\{0, x\}$, $x^- = \max\{0, -x\}$, $\overline{x} = 1 - x$. We use to boldface uppercase characters to denote matrix, and use $\mathbf{I}$ to denote an identity matrix with the dimension being clear from the context. $D[0, \infty)$ denotes the space of functions defined on $[0, \infty)$, which are right continuous and have left-limits. A sequence of processes Z^n in $D[0, \infty)$ is said to converge *u.o.c.* to a process Z in $D[0, \infty)$, if Z^n converges to Z uniformly on any compact set on $[0, \infty)$ as $n \to \infty$.

2. MARKOV CHAIN ANALYSIS FOR TICKET QUEUES

In this chapter, we study the optimal staffing of the ticket queue with two staffing levels, based on information from the ticket counts only. We derive the optimal threshold to increase and decrease the staffing levels. The main contributions of the study are as follows:

- a Markov chain analysis for the ticket queue, with explicit analytical expressions derived for all major performance measures;
- a complete solution to the optimal staffing problem via fractional programming, along with key structural properties of the problem;
- sensitivity analysis with respect to abandonment rates and other cost parameters;
- random-walk approximations for system performance measures.

The chapter is organized as follows. Section 2.1 spells out the details of the mathematical model and the Markov chain analysis for the ticket queue with two staffing levels. Solutions to the optimal staffing policy and its properties are obtained in Section 2.2. Section 2.3 provides approximations based on random walk analysis. Numerical results including sensitivity analysis are given in Section 2.4. Concluding remarks are summarized in Section 2.5.

2.1 Formulation and Analysis

In the queueing system we consider, customers arrive according to a Poisson process with rate λ. Upon arrival, each customer will receive a numbered ticket with the ticket number running in increasing order. Customers are called to receive service according to the increasing order of the ticket numbers they hold. Assume the customer service requirements are iid (independent and identically distributed) exponential random variables, and independent of the arrivals. The system has two staffing levels, indexed by $i = 1, 2$, with service rates μ_i; and which staffing level to use to serve the customers is the main decision. Each staffing level may involve a single server or a group of multiple servers in parallel, but we will not model this level of granularity. Instead, we will assume at each staffing level i, the total output rate is equal to μ_i, a constant, unless the system is empty (in which case the output rate is zero). Thus, for ease of discussion, we shall refer to each staffing level i simply as server i, $i = 1, 2$.

A customer may abandon her ticket before her number is called for service (no show). If a customer shows up when her ticket number is called, the customer will immediately receive service from one of working servers. If the customer is a no-show, her number will be discarded and the next ticket number will be called. We use α_m $(m = 1, 2)$ to represent the abandonment probability of a ticket when m servers are in operation. That is, whenever one of the m servers (if there are m operating servers) is free to serve, she calls the next ticket number and that number has a probability of α_m to be associated with a no-show customer. Formally, we consider four cost components: (i) customer abandonment cost: each abandonment customer incurs cost r; (ii) adding one server cost (service capacity changeover cost or server setup cost, in the following "server setup cost" is used for the sake of simplicity): each server setup costs K; (iii) server operating cost: server-i operation costs per unit time $c_i, i = 1, 2$; (iv) customer delay cost: each delayed customer incurs cost h per unit time.

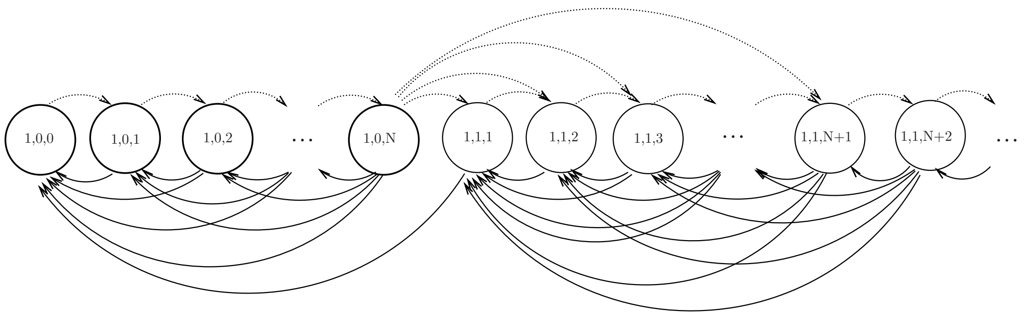


Fig. 2.1: States Transitions for $L = -1$

Our question is how to use ticket information to dynamically determine the staffing level of the ticket queue such that long-run average cost over the infinite time horizon is minimized. Let the binary variable $S_i(t)$ represent the working situation of server-i at time t. Namely, server-i is open at time t if $S_i(t) = 1$, and server-i is closed at time t if $S_i(t) = 0$. Thus, $S(t) = S_1(t) + S_2(t)$ is the number of open servers at time t. Let $Q(t)$ be the number of tickets in the system at time t, including the customers, if any, who are currently receiving service; that is, $Q(t)$ is the sum of the number of busy servers at time t and the difference between the number of the last issued ticket before time t and the maximum of the ticket numbers under service at time t. Then the number of uncalled tickets in queue at time t is $Q(t) - S(t)$. Denote state $(1, 0, 0)$ as the empty system with server-1 open. Starting from the initial state $(S_1(0), S_2(0), Q(0)) = (1, 0, 0)$, the system keeps only server-1 to handle arriving customers, and will add server-2 to handle the waiting customers when $Q(t)$ exceeds N. On the other hand, as soon as the number of tickets in the systems reduces to $L+1$ $(-1 \leq L < N)$ from $Q(t) = N+1$, the system will immediately shut down the server that has just finished the customer service to reset the number of open servers to one, and the system enters into state $(1, 0, L+1)$ or $(0, 1, L+1)$. If $L = -1$, the threshold for us to reset the number of operating servers to be one is zero. That is, only when the system becomes empty, we shut down one server from two operating servers, and to be specific (and without loss of generality), we will shut down server-2. Similarly, if $L = 0$, the threshold for us to shut down one operating server among two operating servers is one, that is, as long as one operating server gets idle, we shut it down, the system state transits from $(1, 1, 2)$ to one of $(1, 0, 1)$ and $(0, 1, 1)$.

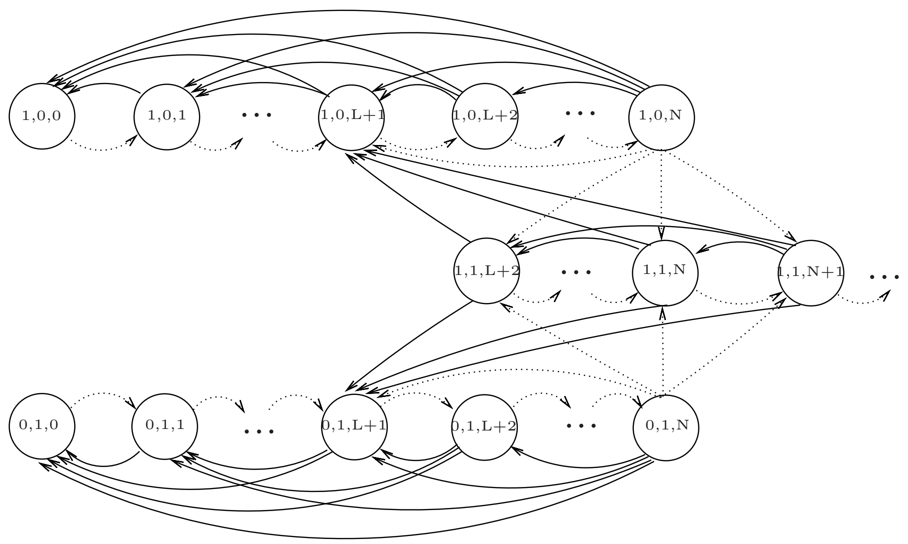


Fig. 2.2: States Transitions for $L \geq 0$

We need to determine the optimal threshold N (to add an operating server) and $L+1$ (to shut down one operating server) so as to minimize the expected long-run average cost. To avoid the trivial case, we assume

$$\frac{(1-\alpha_2)\lambda}{\mu_1+\mu_2} < 1. \tag{2.1}$$

2.1.1 Markov Chains

Due to exponential interarrivals and service times, $\{(S_1(t), S_2(t), Q(t)), t \geq 0\}$ is a Markov chain with the state space

$$\{(s_1, s_2, 0), (s_1, s_2, 1), \cdots, (s_1, s_2, N), (1, 1, n), s_1, s_2 = 0, 1 \text{ with} s_1 + s_2 = 1,$$
$$\text{and} \quad n \geq L+2\}.$$

The renewal reward theorem will be used to derive the expected long-run average cost. As the system operating cost and customer abandonment probability depend on the number of operating servers, we decompose the state space into two disjoint subspaces: the one-server region constituting the states when one server is open,

$$\{(1,0,0), (1,0,1)\cdots, (1,0,N); (0,1,0), (0,1,1)\cdots, (0,1,N)\},$$

and the two-server region containing the states when two servers are open, $\{(1,1,n), n \geq L+2\}$. Each cycle starts with state $(1, 0, L+1)$ and ends also with this state after the system visits state $(1, 0, N+1)$ for only one time.

Starting from this state, the system moves to either state $(1,0,L+2)$ or one of states $\{(1,0,n) : 0 \leq n \leq L\}$. From state $(1,0,n)$ with $1 \leq n \leq L$ (state $(1,0,L+2)$), the system then visits either state $(1,0,k)$ with $0 \leq k \leq n-1$ (state $(1,0,n)$ with $0 \leq n \leq L+1$) or state $(1,0,n+1)$ (state $(1,0,L+3)$) and so on. Of course, from state $(1,0,0)$, the system then moves to state (1,0,1) with probability one. According to the mechanism of our threshold policy, when the system moves to state $(1,0,N+1)$ from state $(1,0,N)$ due to a new arrival, it will immediately set up server-2, who will in turn call the first waiting customer for service. If this customer shows up, the system state changes to $(1,1,N+1)$; if she is a no show, her ticket will be discarded and the subsequent ticket number will be called, and so on. In general, suppose that the first n waiting tickets are discarded due to no shows and the $(n+1)$st ticket corresponds to a showing customer, $n = 0, 1, \ldots, N-L-1$, then the system moves to state $(1,1,N+1-n)$. If all the first $(N-L)$ waiting tickets correspond to no show customers, the system moves to state $(1,0,L+1)$, and server-1 that is originally busy is kept open while server-2 that is just opened will be shut down immediately. Consequently, our cycle is over. After moving to state $(1,1,N+1-n)$ with $n \leq N-L-1$, the system has two operating servers to handle customers. As soon as the number of tickets in the system drops down to $L+1$ due to a new service completion, the server that has just completed the service will be immediately closed, and the system state will change from $(1,1,L+2)$ to $(0,1,L+1)$ if server-1completes that service, and to $(1,0,L+1)$ if server-2 does. If the system state changes to $(1,0,L+1)$, the cycle is over. Otherwise, we start with $(0,1,L+1)$ to repeat the above. Figures 2.1 and 2.2 show the two groups of states and all possible transitions for $L = -1$ and $L > -1$ cases respectively, where dotted arcs represent the transitions triggered by customer arrivals, and solid arcs denote the transitions incurred by the service completions.

Let $c = c_1 + c_2$, $\mu = \mu_1 + \mu_2$, and

$$\hat{\mu}_i = \frac{\mu_i}{\overline{\alpha}_1}, \hat{\mu} = \frac{\mu}{\overline{\alpha}_2}, \rho_i = \frac{\lambda}{\mu_i}, \rho = \frac{\lambda}{\mu}, \overline{\alpha}_i = 1 - \alpha_i,$$
$$\beta_i = \lambda - \hat{\mu}_i, \beta = -\lambda + \hat{\mu}, \theta_i = \alpha_1 + \frac{1}{\rho_i}, i = 1, 2, .$$

Here θ_i reflects the traffic intensity in the one-server region. Namely, when $\theta_i \leq 1$, the traffic intensity is larger than or equal to one, and while $\theta_i > 1$, the traffic intensity is less than one. In view of (2.1), we can see that $\theta_i \leq 1$ is the more interesting case than $\theta_i > 1$, as it puts the traffic intensity ρ_i in the (higher) range of $[1/\overline{\alpha}_1,\ 2/\overline{\alpha}_2)$ with (2.1) holding. Let T_1 be the time interval that the system stays in the one-server region in a regenerative cycle. Similarly, let T_2 be the time interval that the system stays in the two-server region in a regenerative cycle. By the memoryless property of the exponential distribution, each regenerating cycle (the time interval between two entries to state $(1, 0, L + 1)$ and in which the system visits state $(1, 0, N + 1)$ only one time) is $T_1 + T_2$. First we consider T_1. When $L = -1$, the one-server region consists of only states when server-1 is open. By the Markov property of the process $(S_1(t), S_2(t), Q(t))$, T_1 can be considered as the absorbing time of the Markov chain $\{(S_1(t), S_2(t), Q(t)), t \geq 0\}$ with the state space $\{(1, 0, 0), (1, 0, 1), \cdots, (1, 0, N + 1)\}$, the absorbing state $(1, 0, N + 1)$, the generator $\mathbf{D}$

$$\begin{pmatrix}
-\lambda & \lambda & & & & & & \\
\mu_1 & -(\lambda+\mu_1) & \lambda & & & & & \\
\alpha_1\mu_1 & \overline{\alpha}_1\mu_1 & -(\lambda+\mu_1) & \lambda & & & & \\
\alpha_1^2\mu_1 & \alpha_1\overline{\alpha}_1\mu_1 & \overline{\alpha}_1\mu_1 & -(\lambda+\mu_1) & \lambda & & & \\
\vdots & \vdots & \vdots & & \ddots & & & \\
\alpha_1^{N-2}\mu_1 & \alpha_1^{N-3}\overline{\alpha}_1\mu_1 & \alpha_1^{N-4}\overline{\alpha}_1\mu_1 & \cdots & \cdots & -(\lambda+\mu_1) & \lambda & 0 \\
\alpha_1^{N-1}\mu_1 & \alpha_1^{N-2}\overline{\alpha}_1\mu_1 & \alpha_1^{N-3}\overline{\alpha}_1\mu_1 & \cdots & \cdots & \overline{\alpha}_1\mu_1 & -(\lambda+\mu_1) & \lambda \\
0 & 0 & 0 & \cdots & \cdots & 0 & 0 & 0
\end{pmatrix}, \tag{2.2}$$

and the initial distribution $\mathsf{Pr}((S_1(0), S_2(0), Q(0)) = (1, 0, L+1)) = 1$. When $L > -1$, the one-server region consists of the states when server-1 is open and possibly when server-2 is open. T_1 is equal to the above absorbing time plus the random number of the absorbing times of the Markov chain $\{(S_1(t), S_2(t), Q(t)), t \geq 0\}$ with the state space $\{(0, 1, 0), (0, 1, 1), \cdots, (0, 1, N+1)\}$, the absorbing state $(0, 1, N+1)$, the generator $\mathbf{D}$ with μ_1 replacing by μ_2, and the initial distribution $\mathsf{Pr}((S_1(0), S_2(0), Q(0)) = (0, 1, L+1)) = 1$. Let X represent this random number. We know during a regenerating cycle, the number of times to open server-2 is one, and the number of times to open server-1 is X, and

$$\mathsf{Pr}(X=k)=\begin{cases}\alpha_2^{N-L}+(1-\alpha_2^{N-L})\frac{\mu_2}{\mu}, & \text{if } k=0,\\ \left(\alpha_2^{N-L}+(1-\alpha_2^{N-L})\frac{\mu_1}{\mu}\right)^{k-1}(1-\alpha_2^{N-L})^2\frac{\mu_1\mu_2}{\mu^2}, & \text{if } k\geq 1.\end{cases}$$

It is direct to verify that

$$\mathsf{E}X=\frac{\mu_1}{\mu_2}. \tag{2.3}$$

With the help of $M/M/2$, the analysis for T_2 will be directly carried out.

2.1.2 Performance Measures

To get the system performance, we first compute the expected length of a regenerative cycle, $\mathsf{E}T_1+\mathsf{E}T_2$, and the expected cost per regenerative cycle including customer abandonment penalty, server operating cost, and customer delay cost.

We first look at T_1. Based on the above discussion, T_1 can be decomposed into two parts, namely, one-server region with server-1 open (write T_{11}), and one-server region with server-2 open (write T_{12}). Each of them is determined by the absorbing time of the Markov chain given by $(S_1(t), S_2(t), Q(t))$. Thus T_{11} and T_{12} can be represented by phase-type distributions. Using the phase-type distribution properties, we have:

Lemma 1. *The expected time interval for the system to use only one server* *during one regenerative cycle is given by*

$$\mathsf{E}T_1 = \begin{cases} \mathsf{E}T_{11} & \text{if } L = -1, \\ \mathsf{E}T_{11} + \mathsf{E}T_{12} & \text{if } L > -1, \end{cases} \tag{2.4}$$

where

$$\mathsf{E}T_{11} = \frac{\overline{\alpha}_1(N-L)}{\lambda\overline{\theta}_1} + \frac{\theta_1^{N+1} - \theta_1^{L+1}}{\lambda\rho_1\overline{\theta}_1^2}, \quad \mathsf{E}T_{12} = \frac{\mu_1}{\mu_2}\Big[\frac{\overline{\alpha}_1(N-L)}{\lambda\overline{\theta}_2} + \frac{\theta_2^{N+1} - \theta_2^{L+1}}{\lambda\rho_2\overline{\theta}_2^2}\Big].$$

Proof. First we look at $\mathsf{E}T_{11}$. Let $\widetilde{\mathbf{D}}_1$ be the $N \times N$ dimensional matrix obtained by removing the entries in generator $\mathbf{D}$ associated with state $(1, N+1)$:

$$\widetilde{\mathbf{D}}_1 = \begin{pmatrix} -\lambda & \lambda & & & & & \\ \mu_1 & -(\lambda+\mu_1) & \lambda & & & & \\ \alpha_1\mu_1 & \overline{\alpha}_1\mu_1 & -(\lambda+\mu_1) & \lambda & & & \\ \alpha_1^2\mu_1 & \alpha_1\overline{\alpha}_1\mu_1 & \overline{\alpha}_1\mu_1 & -(\lambda+\mu_1) & \lambda & & \\ \vdots & \vdots & \vdots & & \ddots & & \\ \alpha_1^{N-2}\mu_1 & \alpha_1^{N-3}\overline{\alpha}_1\mu_1 & \alpha_1^{N-4}\overline{\alpha}_1\mu_1 & \cdots & \cdots & -(\lambda+\mu_1) & \lambda \\ & & & & & & \\ \alpha_1^{N-1}\mu_1 & \alpha_1^{N-2}\overline{\alpha}_1\mu_1 & \alpha_1^{N-3}\overline{\alpha}_1\mu_1 & \cdots & \cdots & \overline{\alpha}_1\mu_1 & -(\lambda+\mu_1) \end{pmatrix}. \tag{2.5}$$

From the phase-type distribution theory,

$$\mathsf{E}T_1 = (\underbrace{0,\ldots,0}_{L+1},1,0,\ldots,0) \times (-\widetilde{\mathbf{D}}_1^{-1}) \times \mathbf{e}', \tag{2.6}$$

where $\mathbf{e}'$ is the transpose of the $N-$dimensional unit vector. It is direct to verify the inverse of $\widetilde{\mathbf{D}}_1$, denoted by $\widetilde{\mathbf{D}}_1^{-1} = (\widetilde{d}_{ij})_{N\times N}$, can be written as

$$\frac{-1}{\lambda\rho_1}\begin{pmatrix}
\rho_1+\sum_{i=0}^{N-1}\theta_1^i & \rho_1+\sum_{i=0}^{N-2}\theta_1^i & \rho_1+\sum_{i=0}^{N-3}\theta_1^i & \cdots & \rho_1+\sum_{i=0}^{2}\theta_1^i & \rho_1+\sum_{i=0}^{1}\theta_1^i & \rho_1+1 & \rho_1\\
\sum_{i=0}^{N-1}\theta_1^i & \rho_1+\sum_{i=0}^{N-2}\theta_1^i & \rho_1+\sum_{i=0}^{N-3}\theta_1^i & \cdots & \rho_1+\sum_{i=0}^{2}\theta_1^i & \rho_1+\sum_{i=0}^{1}\theta_1^i & \rho_1+1 & \rho_1\\
\theta_1\sum_{i=0}^{N-2}\theta_1^i & \sum_{i=0}^{N-2}\theta_1^i & \rho_1+\sum_{i=0}^{N-3}\theta_1^i & \cdots & \rho_1+\sum_{i=0}^{2}\theta_1^i & \rho_1+\sum_{i=0}^{1}\theta_1^i & \rho_1+1 & \rho_1\\
\theta_1^2\sum_{i=0}^{N-3}\theta_1^i & \theta_1\sum_{i=0}^{N-3}\theta_1^i & \sum_{i=0}^{N-3}\theta_1^i & \cdots & \rho_1+\sum_{i=0}^{2}\theta_1^i & \rho_1+\sum_{i=0}^{1}\theta_1^i & \rho_1+1 & \rho_1\\
\vdots & \vdots & \vdots & & \ddots & & & \\
\theta_1^{N-3}\sum_{i=0}^{2}\theta_1^i & \theta_1^{N-4}\sum_{i=0}^{2}\theta_1^i & \theta_1^{N-5}\sum_{i=0}^{2}\theta_1^i & \cdots & \sum_{i=0}^{2}\theta_1^i & \rho_1+\sum_{i=0}^{1}\theta_1^i & \rho_1+1 & \rho_1\\
\theta_1^{N-2}\sum_{i=0}^{1}\theta_1^i & \theta_1^{N-3}\sum_{i=0}^{1}\theta_1^i & \theta_1^{N-4}\sum_{i=0}^{1}\theta_1^i & \cdots & \theta_1\sum_{i=0}^{1}\theta_1^i & \sum_{i=0}^{1}\theta_1^i & \rho_1+1 & \rho_1\\
\theta_1^{N-1} & \theta_1^{N-2} & \theta_1^{N-3} & \cdots & \theta_1^2 & \theta_1 & 1 & \rho_1
\end{pmatrix}. \tag{2.7}$$

$\mathsf{E}T_{11}$ directly follows (2.6)-(2.7). Note that the expectation of T_{12} is $(1+\mathsf{E}X)$ multiplied by $\mathsf{E}T_{11}$ replacing θ_1 and ρ_1 by θ_2 and ρ_2, respectively. Hence $\mathsf{E}T_{12}$ can be obtained by (2.6)-(2.7) replacing θ_1 and ρ_1 by θ_2 and ρ_2, respectively.

□

Now we consider T_2. To determine the expectation of T_2, as mentioned in the above subsection, consider an auxiliary $M/M/2$ system in which customer arrivals follow a Poisson process with parameter $(1-\alpha_2)\lambda$, and the customer service times from two servers are different. Specifically, the service time from server-i is exponentially distributed with parameter μ_i, $i=1,2$. The initial number of customers in this $M/M/2$ system is $(1+j)$ ("1" represents the customer under service and j is the number of customers in queue who have not abandoned) with probability p_{j+1} given by

$$p_{j+1} = \tbinom{N-L^+}{j}\overline{\alpha}_2^j \alpha_2^{N-L^+-j}, \ j=0,\cdots,N-L^+. \tag{2.8}$$

For this $M/M/2$ system, let τ_j be the first passage time from state j to state $(j-1)$, where $j=2,\cdots,N+1$, and τ_{1i} the first passage time from state 1 with server-i busy to empty, $i=1,2$. Recall that when $L>-1$, there are one time to open server-2 and X times to open server-1 during a regenerating cycle, and when $L=-1$, there is only one time to open server-2. Moreover after each opening, the system evolves as $M/M/2$ described above. Thus we have

$$\mathsf{E}T_2 = \begin{cases} (1+\mathsf{E}X)\Big[p_2\mathsf{E}\tau_2 + p_3\left(\mathsf{E}\tau_2+\mathsf{E}\tau_3\right)+\cdots+p_{N-L+1}\sum\limits_{j=1}^{N-L}\mathsf{E}\tau_{j+1}\Big], & \text{if } L>-1,\\ p_1\mathsf{E}\tau_{11}+\sum_{j=1}^{N}p_{j+1}\Big(\frac{\mu_2}{\mu}\mathsf{E}\tau_{11}+\frac{\mu_1}{\mu}\mathsf{E}\tau_{12}+\sum\limits_{k=1}^{j}\mathsf{E}\tau_{k+1}\Big), & \text{if } L=-1. \end{cases} \tag{2.9}$$

By Lemma 1 in [27], we know that $\tau_2, \tau_3, \cdots, \tau_{N+1}$ have the same distribution with the mean

$$\mathsf{E}\tau_2 = \frac{1}{\mu - \overline{\alpha}_2\lambda}. \tag{2.10}$$

Going alone the line of the proof of Lemma 1 in [27], for τ_{11} and τ_{12}, we have

$$\mathsf{E}e^{-s\tau_{11}} = \frac{\lambda}{\lambda+\mu_1+s}\mathsf{E}e^{-s\tau_2}\Big(\frac{\mu_1}{\mu}\mathsf{E}e^{-s\tau_{12}} + \frac{\mu_2}{\mu}\mathsf{E}e^{-s\tau_{11}}\Big) + \frac{\mu_1}{\lambda+\mu_1+s}, \tag{2.11}$$

$$\mathsf{E}e^{-s\tau_{12}} = \frac{\lambda}{\lambda+\mu_2+s}\mathsf{E}e^{-s\tau_2}\Big(\frac{\mu_1}{\mu}\mathsf{E}e^{-s\tau_{12}} + \frac{\mu_2}{\mu}\mathsf{E}e^{-s\tau_{11}}\Big) + \frac{\mu_2}{\lambda+\mu_2+s}. \tag{2.12}$$

This, by taking derivative on both sides and letting $s = 0$, gives that

$$\mathsf{E}\tau_{11} = \frac{\mu(\lambda+\mu_2)(\mu+\lambda\alpha_2)}{\mu_1\mu_2(\mu+2\lambda)(\mu-\lambda\overline{\alpha}_2)},\ \mathsf{E}\tau_{12} = \frac{\mu(\lambda+\mu_1)(\mu+\lambda\alpha_2)}{\mu_1\mu_2(\mu+2\lambda)(\mu-\lambda\overline{\alpha}_2)}. \tag{2.13}$$

It follows from (2.3) and (2.8)-(2.13) that

Lemma 2. *The expected time interval between an open and a shutdown of the second server is given by*

$$\begin{aligned}\mathsf{E}T_2 =& \frac{(\mu+\lambda\alpha_2)[\alpha_2^N\mu_1(\mu_2-\mu_1)+\mu_1(\lambda+\mu_1)+\mu_2(\lambda+\mu_2)]}{\mu_1\mu_2(\mu+2\lambda)(\mu-\lambda\overline{\alpha}_2)}L^- \\ &+ \frac{\overline{\alpha}_2(N-L^+)}{\mu-\lambda\overline{\alpha}_2}\Big(1+\frac{\mu_1}{\mu_2}(1-L^-)\Big)\end{aligned}$$

Finally we compute the expected total cost in a regenerative cycle, which includes the server's setup and operation costs, the customer delay cost, and the customer abandonment cost. Clearly, the server's expected setup cost is K if $L = -1$ (as the second server is opened only once in each cycle), and $(1+\mathsf{E}X)K$ if $L > -1$ (as server-2 is opened only once and server-1 is opened

X times in each cycle), operation cost for server-1 is $c_1 \cdot (\mathsf{E}T_{11} + \mathsf{E}T_2)$, and operation cost for server-2 is $c_2 \cdot (\mathsf{E}T_{12} + \mathsf{E}T_2)$. For the customer delay cost, we consider two parts: one-server region and two-server region. Let C_1 denote the customer delay cost in the one-server region. For the one-server region, when $L = -1$, only the system states $(1, 0, n)$ with $2 \leq n \leq N$ may incur the customer delay cost; when $L > -1$, both the system states $(1, 0, n)$ and $(0, 1, n)$ with $2 \leq n \leq N$ may incur the customer delay cost. Moreover, if the system state is $(1, 0, n)$ or $(0, 1, n)$, the number of customers the system pays their delay cost (that is, the number of waiting customers) is a binomial random variable with mean $\overline{\alpha}_1(n-1)$. As the system sojourn time at each state $(1, 0, n)$ (or $(0, 1, n)$) with $1 \leq n \leq N$ is an exponential random variable with parameter $(\lambda + \mu_1)$ (or $(\lambda + \mu_2)$), thus to find the customer delay cost in the time intervals T_1, it suffices to find out the number of times for the system to visit each state $(1, 0, n)$ during T_1 and $(0, 1, n)$ during T_{12}. Based on this analysis, by the property of the Markov chains, we can prove

Lemma 3. *The customer delay cost in the one-server region is*

$$\begin{aligned} C_1 = & \frac{h\overline{\alpha}_1}{\lambda\rho_1}\Big[\frac{\theta_1^{L^+} - \theta_1^N}{\overline{\theta}_1^3} - \frac{1 + \rho_1\overline{\theta}_1}{2\overline{\theta}_1}\left((L^+)^2 - N^2\right) + \frac{2 + \overline{\theta}_1 + \rho_1\overline{\theta}_1^2}{2\overline{\theta}_1^2}(L^+ - N)\Big] \\ & + \frac{\mu_1}{\mu_2} \cdot \frac{h\overline{\alpha}_1}{\lambda\rho_2}\Big[\frac{\theta_2^{L^+} - \theta_2^N}{\overline{\theta}_2^3} - \frac{1 + \rho_2\overline{\theta}_2}{2\overline{\theta}_2}\left((L^+)^2 - N^2\right) \\ & \qquad + \frac{2 + \overline{\theta}_2 + \rho_2\overline{\theta}_2^2}{2\overline{\theta}_2^2}(L^+ - N)\Big](1 - L^-). \end{aligned}$$

Proof. First we consider the customer delay cost incurred by the period T_{11}.

For the Markov chain $\{(S_1(t), S_2(t), Q(t)), t \geq 0\}$ given in (2.2) starting with state $(1, 0, L+1)$, let $V(1, 0, i)$ be the number of visits to state $(1, 0, i)$ during T_1, $i = 0, \cdots, N$. In view of the definition of $\widetilde{\mathbf{D}}_1^{-1}$ given by (2.7). Let $f_{L+1,i}$ be the probability that the chain visits state $(1, 0, i)$ from state $(1, 0, L+1)$ and $f_{i,i}$ the probability that the chain revisits state $(1, 0, i)$, $i = 0, 1, \cdots, N$. From the theory of absorbing property of the first passage probability of the transient Markov chain, we know that $V(1, 0, i)$ is a geometric random variable with

$$\mathsf{Pr}\left(V(1, 0, i) = n\right) = f_{L+1,i} \times (1 - f_{ii}) \times (f_{ii})^{n-1}, \quad n = 1, 2, \cdots, i \neq L+1$$

$$\mathsf{Pr}\left(V(1, 0, L+1) = n\right) = (1 - f_{L+1,L+1}) \times (f_{L+1,L+1})^{n-1}, \ n = 1, 2, \cdots, i = L+1.$$

The first passage probabilities $f_{L+1,i}$ and f_{ii} can be computed by

$$\begin{aligned} f_{L+1,i} &= \frac{\sum_{n=1}^{\infty} p_{L+2,i+1}^{n}}{\sum_{n=0}^{\infty} p_{i+1,i+1}^{n}}, \ i \neq L+1, \\ f_{ii} &= \frac{\sum_{n=1}^{\infty} p_{i+1,i+1}^{n}}{\sum_{n=0}^{\infty} p_{i+1,i+1}^{n}}, \ i = 0, 1, \cdots, N. \end{aligned}$$

Here, p_{ij}^{n} is the n-step transition probability for the transition probability matrix given by

$$\mathbf{P} = (p_{ij})_{(N+1)\times(N+1)} = \frac{1}{\lambda + \mu}\widetilde{\mathbf{D}}_1 + \mathbf{I}.$$

Here $\widetilde{\mathbf{D}}_1$ is given by (2.5) and $\mathbf{I}$ is an identity matrix. This implies

$$(\mathbf{I}-\mathbf{P})^{-1} = (\lambda+\mu_1)\left(-\widetilde{\mathbf{D}}_1\right)^{-1}.$$

In view of (2.7), we have

$$f_{L+1,i} = \frac{\widetilde{d}_{L+2,i+1}}{\widetilde{d}_{i+1,i+1}} \quad \text{and} \quad f_{ii} = \frac{(\lambda+\mu_1)\widetilde{d}_{i+1,i+1}+1}{(\lambda+\mu_1)\widetilde{d}_{i+1,i+1}}.$$

We have

$$\mathsf{E}V(1,0,i) = -(\lambda+\mu_1)\widetilde{d}_{L+2,i+1}, \; i=0,\cdots,N.$$

This implies that the expected sojourn times in states $(1,i)$ is

$$\frac{1}{\lambda+\mu_1}\mathsf{E}V(1,0,i) = -\widetilde{d}_{L+2,i+1}, \; i=0,\cdots,N. \tag{2.14}$$

Given the ticket queue length i, the corresponding number of waiting customers follows a binomial distribution with mean $(1-\alpha_1)i$. Then, the customer delay cost in T_{11} is

$$h\left[(1-\alpha_1)(-\widetilde{d}_{L+2,3}) + 2(1-\alpha_1)(-\widetilde{d}_{L+2,4}) + \cdots + (N-1)(1-\alpha_1)(-\widetilde{d}_{L+2,N+1})\right]. \tag{2.15}$$

The customer delay cost in T_{12} can be obtained by (2.15) in which μ_1 is replaced by μ_2. Hence the proof of the lemma is completed. $\square$

Denote C_2 as the customer delay cost in the two-server region. When $L > -1$, there are $(1 + X)$ times to open the second server. According to the mechanism for us to use the second server and the memoryless property of exponential distributions, we know that after each opening, the system evolution follows $M/M/2$ system dynamics with the initial distribution of the number of customers given by (2.8). In other words, we have $(1 + X)$ two-server subregions in the two-server region, and each subregion has the same customer delay cost. We use $T_2^{(s)}$ denote a two-server subregion. Of course, $T_2^{(s)} = T_2$ if $L = -1$. Let $C_2^{(s)}$ be the customer delay cost during $T_2^{(s)}$. Note that there always exist at least L^+ waiting tickets in period $T_2^{(s)}$. Also there are N waiting tickets at the instant to open the second server. Note that the customer delay cost is independent of the service discipline as long as the system is work-conserving. Thus we keep the initial L^+ tickets never to be called during $T_2^{(s)}$. At the beginning, we first call the initial other $(N-L^+)$ tickets to get service, then we serve the customers who arrive during $T_2^{(s)}$. In view of this service arrangement, we can decompose the customer delay cost in period $T_2^{(s)}$ into three parts:

- $C_{21}^{(s)}$ is the expected customer delay cost incurred by arriving customers during $T_2^{(s)}$,
- $C_{22}^{(s)}$ is the expected customer delay cost incurred by the initial L^+ tickets during $T_2^{(s)}$,

- $C_{23}^{(s)}$ is the expected customer delay cost incurred by the initial $(N-L^+)$ tickets.

$C_2^{(s)}$ can be written as

$$C_2^{(s)} = C_{21}^{(s)} + C_{22}^{(s)} + C_{23}^{(s)}. \tag{2.16}$$

Recall that the number of waiting customers among the initial L^+ tickets is random and follows a binomial distribution with mean $\overline{\alpha}_2 L^+$. Hence,

$$C_{22}^{(s)} \quad = \quad h\overline{\alpha}_2 L^+ \times \mathsf{E}T_2^{(s)}. \tag{2.17}$$

Also the number of waiting customers among the initial $N - L^+$ tickets, denoted by Y, follows a binomial distribution with mean $\overline{\alpha}_2(N - L^+)$. C_{23} is just the delay cost of these Y customers. Since the waiting time of the ith customer in the sequence of Y customers is $(i-1)/\mu$, we have

$$C_{23}^{(s)} = h\mathsf{E}\Big(\sum_{i=1}^{Y} \frac{i-1}{\mu}\Big) = \frac{h}{\mu}\mathsf{E}\Big(\frac{Y(Y-1)}{2}\Big) = \frac{h}{2\mu}\overline{\alpha}_2^2(N-L^+)(N-L^+-1). \tag{2.18}$$

To get $C_{21}^{(s)}$, we again consider the auxiliary $M/M/2$ system. Let $Q_2(t)$ be the number of customers at time t, with the initial number of customers $Q_2(0) = 1 + Y$, where Y is the same random variable as one used by (2.18). We decompose $T_2^{(s)}$ into three periods denoted by $T_{21}^{(s)}, T_{22}^{(s)}$, and $T_{23}^{(s)}$, where $T_{21}^{(s)}$ is the first time at which the system can handle the customers arriving during $T_2^{(s)}$, $T_{22}^{(s)}$ is the first time at which the system has one idle server after $T_{21}^{(s)}$, and $T_{23}^{(s)}$ is the first passage time from state $(1,1,1)$ to state $(1,1,0)$ for $L=-1$. Clearly, if $L>-1$, we have $T_{23}^{(s)}=0$. Formally,

$$\begin{aligned}
T_{21}^{(s)} &= \inf\{t : \text{the number of the service completions by time } t \geq Y\},\\
T_{22}^{(s)} &= \inf\{t \geq 0 : Q_2(t)=1\} - T_{21}^{(s)},\\
T_{23}^{(s)} &= T_2^{(s)} - T_{21}^{(s)} - T_{22}^{(s)}.
\end{aligned}$$

Let W be the waiting time of a customer arriving during $T_2^{(s)}$. Then from Little's formula, we have

$$\begin{aligned}
C_{21}^{(s)} = h\mathsf{E}T_2^{(s)} \times \Big[\lambda\overline{\alpha}_2\mathsf{E}\Big(W\Big|\text{arriving during } (T_{21}^{(s)}+T_{22}^{(s)})\Big)\frac{\mathsf{E}(T_{21}^{(s)}+T_{22}^{(s)})}{\mathsf{E}T_2^{(s)}}\\
+ \lambda\overline{\alpha}_2\mathsf{E}\Big(W\Big|\text{arriving during } T_{23}^{(s)}\Big)\frac{\mathsf{E}T_{23}^{(s)}}{\mathsf{E}T_2^{(s)}}\Big].
\end{aligned} \tag{2.19}$$

From this we can get $C_{21}^{(s)}$.

Lemma 4. *The expected customer delay cost incurred by arriving customers during $T_2^{(s)}$ is given by*

$$C_{21}^{(s)} = \begin{cases} \frac{h\lambda\overline{\alpha}_2^2(N-L)}{\mu-\lambda\overline{\alpha}_2}\Big[\frac{\lambda\overline{\alpha}_2}{\mu(\mu-\lambda\overline{\alpha}_2)} + \frac{\overline{\alpha}_2(N-L)+1+\alpha_2}{2\mu}\Big], & \text{if } L>-1,\\ \frac{h\lambda\overline{\alpha}_2}{\mu-\lambda\overline{\alpha}_2}\Big[\frac{\overline{\alpha}_2^2N^2}{2\mu} + \overline{\alpha}_2 N\frac{\lambda\overline{\alpha}_2^2+(1+\alpha_2)\mu}{2\mu(\mu-\lambda\overline{\alpha}_2)} + \frac{\alpha_2^N\lambda\mu_1(\mu_2-\mu_1)+\lambda(\lambda\mu+\mu_1^2+\mu_2^2)}{\mu_1\mu_2(\mu+2\lambda)(\mu-\lambda\overline{\alpha}_2)}\Big], & \text{if } L=-1. \end{cases}$$

Proof. First, according to Theorem 1 of Omahen and Marathe (1978),

$$\mathsf{E}\Big(W\Big|\text{arriving during } (T_{21}^{(s)}+T_{22}^{(s)})\Big) = \frac{\lambda\overline{\alpha}_2}{\mu(\mu-\lambda\overline{\alpha}_2)} + \frac{\mathsf{E}(T_{21}^{(s)})^2}{2\mathsf{E}T_{21}^{(s)}}. \tag{2.20}$$

The Laplace-Stieltjes Transform of $T_{21}^{(s)}$ is given by

$$\mathsf{E}e^{-sT_{21}^{(s)}} = \sum_{i=0}^{N-L^+} \tbinom{N-L^+}{i}\alpha_2^{N-L^+-i}\overline{\alpha}_2^i\left(\frac{\mu}{\mu+s}\right)^i = \left(\frac{\mu+s\alpha_2}{\mu+s}\right)^{N-L^+}.$$

This implies

$$\mathsf{E}T_{21}^{(s)} = \frac{(N-L^+)\overline{\alpha}_2}{\mu},\quad \mathsf{E}\Big(T_{21}^{(s)}\Big)^2 = \frac{\overline{\alpha}_2^2(N-L^+)^2+(1-\alpha_2^2)(N-L^+)}{\mu^2}. \tag{2.21}$$

Using (2.20), we have

$$\mathsf{E}\Big(W\Big|\text{arriving during } (T_{21}^{(s)}+T_{22}^{(s)})\Big) = \frac{\lambda\overline{\alpha}_2}{\mu(\mu-\lambda\overline{\alpha}_2)} + \frac{\overline{\alpha}_2(N-L^+)+1+\alpha_2}{2\mu}. \tag{2.22}$$

By (2.9),

$$\begin{aligned}\mathsf{E}(T_{21}^{(s)}+T_{22}^{(s)}) &= \Big[p_2\mathsf{E}\tau_2 + p_3\left(\mathsf{E}\tau_2+\mathsf{E}\tau_3\right)+\cdots+p_{N-L+1}\sum_{i=1}^{N-L}\mathsf{E}\tau_{i+1}\Big] \\ &= \frac{\overline{\alpha}_2(N-L^+)}{\mu-\lambda\overline{\alpha}_2}.\end{aligned}$$

Hence the lemma for $L > -1$ follows from $T_{23}^{(s)} = 0$, (2.19) and (2.22).

Now we consider the case $L = -1$. According to the definition of $T_{23}^{(s)}$ and τ_{11} and τ_{12} in (2.9), we have

$$\begin{aligned}
&\mathsf{E}\Big(W\Big|\text{arriving during } T_{23}^{(s)}\Big)\\
=&\frac{\mu_1}{\mu}\mathsf{E}\Big(W\Big|\text{arriving during } \tau_{12}\Big) + \frac{\mu_2}{\mu}\mathsf{E}\Big(W\Big|\text{arriving during } \tau_{11}\Big).
\end{aligned} \tag{2.23}$$

Taking derivative with resect to "s" in (2.11)-(2.12), we have

$$\begin{aligned}
\mathsf{E}\tau_{11} &= \frac{1/(\lambda+\mu_1)}{1-\lambda\mu_2/[\mu(\lambda+\mu_1)]} + \frac{\lambda/(\lambda+\mu_1)}{1-\lambda\mu_2/[\mu(\lambda+\mu_1)]}\mathsf{E}\tau_2 + \frac{\lambda\mu_1/[\mu(\lambda+\mu_1)]}{1-\lambda\mu_2/[\mu(\lambda+\mu_1)]}\mathsf{E}\tau_{12},\\
\mathsf{E}\tau_{12} &= \frac{1/(\lambda+\mu_2)}{1-\lambda\mu_1/[\mu(\lambda+\mu_2)]} + \frac{\lambda/(\lambda+\mu_2)}{1-\lambda\mu_1/[\mu(\lambda+\mu_2)]}\mathsf{E}\tau_2 + \frac{\lambda\mu_2/[\mu(\lambda+\mu_2)]}{1-\lambda\mu_1/[\mu(\lambda+\mu_2)]}\mathsf{E}\tau_{11}.
\end{aligned}$$

Therefore $\mathsf{E}\Big(W|\text{ arriving during } \tau_{11}\Big)$ and $\mathsf{E}\Big(W|\text{ arriving during } \tau_{12}\Big)$ can be written as

$$\begin{aligned}
&\mathsf{E}\Big(W|\text{ arriving during } \tau_{11}\Big)\\
=&\frac{\lambda/(\lambda+\mu_1)}{1-\lambda\mu_2/[\mu(\lambda+\mu_1)]}\cdot\frac{\mathsf{E}\tau_2}{\mathsf{E}\tau_{11}}\cdot\mathsf{E}\Big(W|\text{ arriving during } \tau_2\Big)\\
&+\frac{\lambda\mu_1/[\mu(\lambda+\mu_1)]}{1-\lambda\mu_2/[\mu(\lambda+\mu_1)]}\cdot\frac{\mathsf{E}\tau_{12}}{\mathsf{E}\tau_{11}}\cdot\mathsf{E}\Big(W|\text{ arriving during } \tau_{12}\Big),
\end{aligned} \tag{2.24}$$

$$\begin{aligned}
&\mathsf{E}\Big(W|\text{ arriving during } \tau_{12}\Big)\\
=&\frac{\lambda/(\lambda+\mu_2)}{1-\lambda\mu_1/[\mu(\lambda+\mu_2)]}\cdot\frac{\mathsf{E}\tau_2}{\mathsf{E}\tau_{12}}\cdot\mathsf{E}\Big(W|\text{ arriving during } \tau_2\Big)\\
&+\frac{\lambda\mu_2/[\mu(\lambda+\mu_2)]}{1-\lambda\mu_1/[\mu(\lambda+\mu_2)]}\cdot\frac{\mathsf{E}\tau_{11}}{\mathsf{E}\tau_{12}}\cdot\mathsf{E}\Big(W|\text{ arriving during } \tau_{11}\Big).
\end{aligned} \tag{2.25}$$

By Theorem 2 of Omaben and Marathe (1978),

$$\mathsf{E}\Big(W|\text{ arriving during }\tau_2\Big)=\frac{1}{\mu-\lambda\overline{\alpha}_2}.$$

Hence, we have

$$\begin{aligned}\mathsf{E}\Big(W|\text{ arriving during }\tau_{11}\Big)&=\mathsf{E}\Big(W|\text{ arriving during }\tau_{12}\Big)\\&=\frac{\lambda}{(\mu+\lambda\alpha_2)(\mu-\lambda\overline{\alpha}_2)}.\end{aligned}\tag{2.26}$$

Recalling from (2.9) that

$$\mathsf{E}T_{23}^{(s)}\quad=\alpha_2^N\mathsf{E}\tau_{11}+(1-\alpha_2^N)\Big(\tfrac{\mu_1}{\mu}\mathsf{E}\tau_{12}+\tfrac{\mu_2}{\mu}\mathsf{E}\tau_{11}\Big),\tag{2.27}$$

we know that

$$\begin{aligned}&\mathsf{E}\Big(W|\text{ arriving during }T_{23}^{(s)}\Big)\\&=\Big(\alpha_2^N+(1-\alpha_2^N)\frac{\mu_2}{\mu}\Big)\frac{\mathsf{E}\tau_{11}}{\mathsf{E}T_{23}^{(s)}}\mathsf{E}\Big(W|\text{ arriving during }\tau_{11}\Big)\\&\quad+(1-\alpha_2^N)\frac{\mu_1}{\mu}\frac{\mathsf{E}\tau_{12}}{\mathsf{E}T_{23}^{(s)}}\mathsf{E}\Big(W|\text{ arriving during }\tau_{12}\Big).\end{aligned}\tag{2.28}$$

Combining (2.26)-(2.37) yields that

$$\begin{aligned}&\mathsf{E}T_{23}^{(s)} \times \mathsf{E}\Big(W|\text{ arriving during } T_{23}^{(s)}\Big)\\ =&\frac{\lambda}{\mu-\lambda\overline{\alpha}_2}\frac{\alpha_2^N\mu_1(\mu_2-\mu_1)+\mu_1(\lambda+\mu_1)+\mu_2(\lambda+\mu_2)}{\mu_1\mu_2(\mu+2\lambda)(\mu-\lambda\overline{\alpha}_2)}.\end{aligned}$$

The lemma for $L=-1$ directly follows from (2.19) and (2.22). □

Using (2.3), (2.16)-(2.18), Lemma 4, $C_2 = C_2^{(s)}$ if $L=-1$, and $C_2 = (1+\mathsf{E}X)C_2^{(s)}$, we get the customer delay cost for the two-server region.

Lemma 5. *The expected customer delay cost in the two-server region, C_2, is given by*

$$C_2 = \begin{cases} \frac{h\overline{\alpha}_2}{2(\mu-\lambda\overline{\alpha}_2)}\Big(\overline{\alpha}_2 N^2 + N\frac{\overline{\alpha}_2[(2+\overline{\alpha}_2)\lambda-\mu]}{\mu-\lambda\overline{\alpha}_2} + \frac{2\lambda^2\left[\alpha_2^N\mu_1(\mu_2-\mu_1)+\lambda\mu+\mu_1^2+\mu_2^2\right]}{\mu_1\mu_2(\mu+2\lambda)(\mu-\lambda\overline{\alpha}_2)}\Big) & \\ & \textit{if } L=-1; \\ \frac{h\overline{\alpha}_2^2\mu}{2\mu_2(\mu-\lambda\overline{\alpha}_2)}\Big(N^2-L^2+(N-L)\frac{\lambda(2+\overline{\alpha}_2)-\mu}{\mu-\lambda\overline{\alpha}_2}\Big) & \\ & \textit{if } L>-1. \end{cases}$$

To get the customer abandonment cost, for each cycle, we need find the expectation of the system idle time in one-server region, denoted by T_{10}, and the expectation of the one server idle in two-server region for $L=-1$, represented by T_{20}. Again by the properties of phase-type distributions, we have

Lemma 6. *For each cycle, the expectation of the system idle is given by*

$$\mathsf{E}T_{10} = \frac{\theta_1^{L^+} - \theta_1^N}{\lambda\rho_1\overline{\theta}_1} + \frac{L^-}{\lambda} + \frac{\mu_1}{\mu_2}\frac{\theta_2^{L^+} - \theta_2^N}{\lambda\rho_2\overline{\theta}_2}(1 - L^-), \tag{2.29}$$

and the expectation of one server idle is

$$\mathsf{E}T_{20} = \frac{\alpha_2^N\mu_1(\mu_2 - \mu_1) + \mu_2(\lambda + \mu_2) + \mu_1(\lambda + \mu_1)}{\mu_1\mu_2(\mu + 2\lambda)}. \tag{2.30}$$

Proof. In view of Lemma 3, we know that

$$\mathsf{E}T_{10} = \frac{1}{\lambda + \mu_1}\mathsf{E}V(1, 0, 0).$$

By (2.7)-(2.14), we have the lemma for $\mathsf{E}T_{10}$. Now consider $\mathsf{E}T_{20}$. Let $\tau_{1i}^{(0)}$ is the accumulative time for one server idle during $\tau_{1i}, i = 1, 2$. Then we have

$$\begin{aligned}
\mathsf{E}\tau_{11}^{(0)} &= \frac{1}{\lambda + \mu_1} + \frac{\lambda}{\lambda + \mu_1}\left[\frac{\mu_2}{\mu}\mathsf{E}\tau_{11}^{(0)} + \frac{\mu_1}{\mu}\mathsf{E}\tau_{12}^{(0)}\right], \\
\mathsf{E}\tau_{12}^{(0)} &= \frac{1}{\lambda + \mu_2} + \frac{\lambda}{\lambda + \mu_2}\left[\frac{\mu_2}{\mu}\mathsf{E}\tau_{11}^{(0)} + \frac{\mu_1}{\mu}\mathsf{E}\tau_{12}^{(0)}\right].
\end{aligned}$$

This gives that

$$\begin{aligned}
\mathsf{E}T_{20} &= \left[\alpha_2^N + \frac{\mu_2}{\mu}(1 - \alpha_2^N)\right]\mathsf{E}\tau_{11}^{(0)} + \frac{\mu_1}{\mu}(1 - \alpha_2^N)\mathsf{E}\tau_{12}^{(0)} \\
&= \frac{\alpha_2^N\mu_1(\mu_2 - \mu_1) + \mu_2(\lambda + \mu_2) + \mu_1(\lambda + \mu_1)}{\mu_1\mu_2(\mu + 2\lambda)},
\end{aligned}$$

which prove the lemmas for $\mathsf{E}T_{20}$. □

Without loss of generality, after making a cost normalization, we assume the cost per customer abandonment is one, i.e., $r = 1$. Using the results developed yet (Lemmas 1-3 and Lemmas 5-6), we can express the expected long-run average cost, denoted by $\mathcal{AC}(L, N)$, as

$$\begin{aligned}\mathcal{AC}(L,N) =& \frac{1}{\mathsf{E}T_1+\mathsf{E}T_2}\Big[\lambda\alpha_1(\mathsf{E}T_1-\mathsf{E}T_{10})+\lambda\alpha_2(\mathsf{E}T_2-L^-\times\mathsf{E}T_{20})+C_1+C_2\\ &+c_1(\mathsf{E}T_{11}+\mathsf{E}T_2)+c_2(\mathsf{E}T_{12}+\mathsf{E}T_2)+K+(1-L^-)\frac{\mu_1}{\mu_2}K\Big]\\ :=&\frac{f(L,N)}{g(L,N)},\end{aligned} \tag{2.31}$$

where

$$\begin{aligned}f(L,N) = {}& a(\theta_1^N-\theta_1^{L^+})+a_2(N^2-(L^+)^2)+a_1(N-L^+)+(a_0+a_e\alpha_2^N)L^-+K\\ &+(1-L^-)\frac{\mu_1}{\mu_2}\Big[a'(\theta_2^N-\theta_2^{L^+})+a_2'(N^2-(L^+)^2)+a_1'(N-L^+)+K\Big],\end{aligned} \tag{2.32}$$

$$\begin{aligned}g(L,N) = {}& b(\theta_1^N-\theta_1^{L^+})+b_1(N-L^+)+(b_0+b_e\alpha_2^N)L^-\\ &+(1-L^-)\frac{\mu_1}{\mu_2}\Big[b'(\theta_2^N-\theta_2^{L^+})+b_1'(N-L^+)\Big],\end{aligned} \tag{2.33}$$

$$a = \frac{\alpha_1\rho_1\hat{\mu}_1^2}{\beta_1^2}-\frac{\rho_1\hat{\mu}_1^2}{\beta_1^3}h+\frac{\hat{\mu}_1^2(1+\alpha_1\rho_1)}{\lambda\beta_1^2}c_1,\quad a_2=\frac{h}{2}\Big(\frac{\overline{\alpha}_1}{\beta_1}+\frac{\overline{\alpha}_2}{\beta}\Big), \tag{2.34}$$

$$a_1 = \frac{\lambda\alpha_1+c_1}{\beta_1}+\frac{\lambda\alpha_2+c}{\beta}-\frac{h}{2}\Big[\frac{\hat{\mu}_1(1+\alpha_1+\rho_1\overline{\alpha}_1^2)}{\beta_1^2}+\frac{\mu-\lambda(2+\overline{\alpha}_2)}{\beta^2}\Big], \tag{2.35}$$

$$a_0 = \frac{\big[\mu_1(\lambda+\mu_1)+\mu_2(\lambda+\mu_2)\big]\big[\lambda^2\alpha_2\beta+\beta c(\mu+\lambda\alpha_2)+h\lambda^2\big]}{\mu_1\mu_2\beta^2\overline{\alpha}_2(\mu+2\lambda)}+\frac{c_1}{\lambda}, \tag{2.36}$$

$$a_e = \frac{\lambda^2\alpha_2(\mu_2-\mu_1)+c(\mu_2-\mu_1)(\mu+\lambda\alpha_2)}{\mu_2\beta\overline{\alpha}_2(\mu+2\lambda)} + \frac{\lambda^2(\mu_2-\mu_1)}{\mu_2\beta^2\overline{\alpha}_2(\mu+2\lambda)}h, \tag{2.37}$$

$$a' = \frac{\alpha_1\rho_2\hat{\mu}_2^2}{\beta_2^2} - \frac{\rho_2\hat{\mu}_2^2}{\beta_2^3}h + \frac{\hat{\mu}_2^2(1+\alpha_1\rho_2)}{\lambda\beta_2^2}c_2, \quad a_2' = \frac{h}{2}\Big(\frac{\overline{\alpha}_1}{\beta_2}+\frac{\overline{\alpha}_2}{\beta}\Big), \tag{2.38}$$

$$a_1' = \frac{\lambda\alpha_1+c_2}{\beta_2} + \frac{\lambda\alpha_2+c}{\beta} - \frac{h}{2}\Big[\frac{\hat{\mu}_2(1+\alpha_1+\rho_2\overline{\alpha}_1^2)}{\beta_2^2} + \frac{\mu-\lambda(2+\overline{\alpha}_2)}{\beta^2}\Big], \tag{2.39}$$

$$b = \frac{1+\alpha_1\rho_1}{\lambda}\cdot\frac{\hat{\mu}_1^2}{\beta_1^2},\ b_1 = \frac{1}{\beta_1}+\frac{1}{\beta},\ b_0 = \frac{1}{\lambda} + \frac{(\mu+\lambda\alpha_2)\big[\lambda\mu+\mu_1^2+\mu_2^2\big]}{\beta\mu_1\mu_2\overline{\alpha}_2(\mu+2\lambda)} \tag{2.40}$$

$$b_e = \frac{(\mu_2-\mu_1)[\mu+\lambda\alpha_2]}{\beta\mu_2\overline{\alpha}_2(\mu+2\lambda)},\ b' = \frac{1+\alpha_1\rho_2}{\lambda}\cdot\frac{\hat{\mu}_2^2}{\beta_2^2}, \quad b_1' = \frac{1}{\beta_2}+\frac{1}{\beta}. \tag{2.41}$$

Our objective is to find L and N so as to minimize $\mathcal{AC}(L,N)$. That is,

$$\min_{L\geq -1, N\geq 0\vee L} \mathcal{AC}(L,N) = \min_{L\geq -1, N\geq 0\vee L} \frac{f(L,N)}{g(L,N)}. \tag{2.42}$$

2.2 Optimal Solution

To obtain the optimal thresholds of opening and closing the second server, we first look at some properties of the coefficients of the decision variables L and N in (2.42). The following relations follow immediately:

$$b > 0 \text{ and } b' > 0, \tag{2.43}$$

$$\theta_1 \leq 1 \Leftrightarrow \beta_1 \geq 0;\ \theta_2 \leq 1 \Leftrightarrow \beta_2 \geq 0, \tag{2.44}$$

$$b_1 > 0 \text{ and } a_2 > 0 \text{ if } \theta_1 \leq 1;\ b_1' > 0 \text{ and } a_2' > 0 \text{ if } \theta_2 \leq 1, \tag{2.45}$$

$$a > 0 \text{ if } \theta_1 > 1;\;\; a' > 0 \text{ if } \theta_2 > 1, \tag{2.46}$$

Also note that for $i = 1, 2$

$$\theta_i > \ (\leq) \, 1 \quad \text{if and only if} \quad \rho_i < \ (\geq) \, \frac{1}{1-\alpha_1}; \tag{2.47}$$

whereas

$$\beta > 0 \quad \text{if and only if} \quad \frac{\lambda}{\mu} < \frac{1}{1-\alpha_2}.$$

As $N \to +\infty$, we have, from (2.34)-(2.41) and (2.43)-(2.46),

$$\frac{f(L,N)}{g(L,N)} \to \begin{cases} +\infty, & \text{if } \theta_1, \theta_2 \leq 1, \\ \frac{a'}{b'} \mathsf{I}_{\{L \geq 0\}} + \infty \mathsf{I}_{\{L=-1\}} > 0, & \text{if } \theta_1 \leq 1 < \theta_2, \\ \frac{a}{b} > 0, & \text{if } \theta_2 \leq 1 < \theta_1, \\ \frac{a'}{b'} \mathsf{I}_{\{L \geq 0, \mu_1 < \mu_2\}} + \frac{a}{b} \Big(\mathsf{I}_{\{L \geq 0, \mu_1 > \mu_2\}} + \mathsf{I}_{\{L=-1\}} \Big) & \\ + \frac{\mu_2 a + \mu_1 a'}{\mu_2 b + \mu_1 b'} \mathsf{I}_{\{L \geq 0, \mu_1 = \mu_2\}} > 0, & \text{if } \theta_1 > 1, \theta_2 > 1. \end{cases} \tag{2.48}$$

The first limit above takes into account $a_2 > 0$ and $a_2' > 0$. The limit in (2.48) implies that to solve the minimization problem in (2.42), we only need to consider $(L, N) \in [-1, L_0] \times [0 \vee L, N_0]$ for some pre-specified sufficient large L_0 and N_0. This can enable us to use the standard fractional programming techniques to solve problem (2.42). For simplicity, we shall write $\min_{L,N}$ below, in lieu of $\min_{L \in [-1, L_0], N \in [0 \vee L, N_0]}$.

2.2.1 Fractional Programming

Formally, the optimal policy to (2.42) can be solved as follows:

$$\min_{L,N}[f(L,N) - xg(L,N)] := \Psi(x), \tag{2.49}$$

along with a line search

$$\Psi(x) = 0. \tag{2.50}$$

To see this, let x^* be the solution to the equation in (2.50), and (L^*, N^*) be the corresponding minimizer in (2.49), i.e., with $x = x^*$. Then,

$$x^* = \frac{f(L^*, N^*)}{g(L^*, N^*)} \leq \frac{f(L, N)}{g(L, N)}, \qquad \text{for } N \geq L \geq -1,$$

where the first equality follows from $\Psi(x^*) = 0$, and the second inequality is due to:

$$0 = f(L^*, N^*) - x^* \cdot g(L^*, N^*) \leq f(L, N) - x^* \cdot g(L, N).$$

Note, here we implicitly use that $g(L, N)$ $(= \mathsf{E}T_1 + \mathsf{E}T_2) > 0$. In addition, we need $g(L, N) < \infty$ for any feasible $(L, N) \in [-1, L_0] \times [0 \vee L, N_0]$, which certainly holds.

Below we go into more details about solving the two problems in (2.49) and (2.50). First, note that $\Psi(x)$ is strictly decreasing in x. To see this, consider $x_1 < x_2$ and let $(\widetilde{L},\widetilde{N})$ and $(\widehat{L},\widehat{N})$ be the two corresponding minimizers of (2.49). Then,

$$\begin{aligned}\Psi(x_1) &= f(\widetilde{L},\widetilde{N}) - x_1 g(\widetilde{L},\widetilde{N}) > f(\widetilde{L},\widetilde{N}) - x_2 g(\widetilde{L},\widetilde{N}) \\ &\geq f(\widehat{L},\widehat{N}) - x_2 g(\widehat{L},\widehat{N}) = \Psi(x_2),\end{aligned}$$

where the first (strict) inequality is due to $g(L,N) > 0$. Hence, the solution to (2.50) uniquely exists. Next, consider the minimization problem in (2.49). Define $f_1(y) = (a - xb)\theta_1^y + a_2 y^2 + (a_1 - xb_1)y$ and $f_2(y) = (a' - xb')\theta_2^y + a_2' y^2 + (a_1' - xb_1')y$. Then, (2.49) can be written as

$$\begin{aligned}\min_{L,N} \Big\{ & f_1(N) - f_1(L^+) + (a_0 - xb_0 + a_e\alpha_2^N - xb_e\alpha_2^N)L^- + K \\ & +(1 - L^-)\frac{\mu_1}{\mu_2}\Big[f_2(N) - f_2(L^+) + K\Big]\Big\}. \qquad (2.51)\end{aligned}$$

The second derivative with respect to N of the objective function above is:

$$\begin{aligned}&(a - xb)(\ln\theta_1)^2\theta_1^N + 2a_2 + \frac{\mu_1}{\mu_2}(1 - L^-)\big[(a' - xb')(\ln\theta_2)^2\theta_2^N + 2a_2'\big] \\ &\quad + L^-(a_e - xb_e)(\ln\alpha_2)^2\alpha_2^N, \qquad (2.52)\end{aligned}$$

and the second derivative with respect to L of the objective function is:

$$-(a - xb)(\ln\theta_1)^2\theta_1^L - 2a_2 - \frac{\mu_1}{\mu_2}(1 - L^-)\big[(a' - xb')(\ln\theta_2)^2\theta_2^L + 2a_2'\big]. \quad (2.53)$$

There are two steps to find optimal N^* and L^*. The first step is to find the optimal (L_1^*, N_1^*) given $L \geq 0$; and the second step, to find the optimal (L_2^*, N_2^*) given $L = -1$ (here $L_2^* = -1$). Then, we compare $\mathcal{AC}(L_1^*, N_1^*)$ and $\mathcal{AC}(L_2^*, N_2^*)$ to find the optimal (L^*, N^*). For $L \geq 0$, there are four cases:

(i) $\theta_1 < 1, \theta_2 < 1$; in which case $a_2 > 0, a_2' > 0$ (see (2.45)).

(i-a) Suppose $a - xb \geq 0$, and $a' - xb' \geq 0$. Then, by (2.52), the objective function in (2.51) is strictly convex with respect to N; hence, the solution N_1^* uniquely exists. Similarly, the objective function in (2.51) is strictly concave with respect to L. Thus, in view of $L \leq N$, the optimal L_1^* should be 0.

(i-b) Suppose $a - xb < 0$, and $a' - xb' < 0$. Then, the objective function in (2.51) is, with respect to N, either convex, provided

$$(a - xb)(\ln \theta_1)^2 + 2a_2 + \frac{\mu_1}{\mu_2}\Big[(a' - xb')(\ln \theta_2)^2 + 2a_2'\Big] \geq 0, \quad (2.54)$$

(since the first (negative) term in (2.52) becomes less negative as N increases); or it starts with a concave piece, followed by a convex piece, with switch over at $N = \widetilde{N}$, where $\widetilde{N}$ is unique since $(a - xb)\theta_1^N(\ln \theta_1)^2 + 2a_2 + \frac{\mu_1}{\mu_2}\big[(a' - xb')\theta_2^N(\ln \theta_2)^2 + 2a_2'\big]$ is increasing in N here. Hence, the optimal solution N_1^* is either 0 or the minimal point of the convex piece of $f_1(y) + \frac{\mu_1}{\mu_1} f_2(y)$. Following the same argument of (i-a), the optimal solution L_1^* is either 0 or the maximal point of the concave piece of $f_1(y) + \frac{\mu_1}{\mu_1} f_2(y)$.

(i-c) Suppose $a - xb < 0$, and $a' - xb' \geq 0$.

(i-c-1) Suppose $\mu_1 \geq \mu_2$. The convexity of objective function in

(2.51) may have three cases. The first case is that the objective function in (2.51) is convex, then the optimal solutions N_1^* and L_1^* are similar to (i-a); the second one is that the objective function in (2.51) starts with a concave piece, followed by a convex piece, and the optimal solutions N_1^* and L_1^* are similar to (i-b); the last case is that the objective function in (2.51) starts with a convex piece, switches to a concave piece, and switches to a convex piece. In the last case, solution N_1^* is the minimal point of the lower convex piece, and solution L_1^* is either 0 or maximal point of the concave piece.

(i-c-2) Suppose $\mu_1 < \mu_2$. Then the objective function in (2.51) either is convex or starts with a concave piece, followed by a convex piece, and the optimal solutions for N and L is similar to (i-a) and (i-b), respectively.

(i-d) Suppose $a - xb \geq 0$, and $a' - xb' < 0$. This case is completely similar to (i-c).

(ii) $0 < \theta_1 < 1 < \theta_2$; in which case $a_2 > 0$ (see (2.45)).

(ii-a) Suppose $a - xb \geq 0$, and $a' - xb' \geq 0$. Similar to (i-c-1), we can obtain the optimal solutions N_1^* and L_1^*.

(ii-b) Suppose $a - xb < 0$, and $a' - xb' < 0$. Then the convexity of objective function in (2.51) may have three cases. The first case is that the objective function in (2.51) is concave, then the optimal solution N_1^* is N_0, and the optimal L_1^* is either the maximal point of $f_1(y)+\frac{\mu_1}{\mu_2}f_2(y)$ or 0; the second one is that the objective function in (2.51) starts with a convex piece, followed by a concave piece, and the optimal N_1^* is either N_0 or the minimal point of the convex piece, the optimal L_1^* is either the maximal point of the concave piece of $f_1(y) + \frac{\mu_1}{\mu_2} f_2(y)$ or 0; the last case is that the objective function in (2.51) starts with a concave piece, switches to a convex piece, and switches to a concave piece. In the last case, solution N_1^* is either N_0 or the minimal point of the convex piece, and optimal L_1^* is the maximal point of the higher concave piece or 0.

(ii-c) Suppose $a - xb < 0$, and $a' - xb' \geq 0$. This case is similar to (i-c-2).

(ii-d) Suppose $a - xb \geq 0$, and $a' - xb' < 0$. Then the objective function in (2.51) either is concave or starts with a convex piece, followed by a concave piece, and the optimal solutions N_1^* and L_1^* are similar to (ii-b).

(iii) $0 < \theta_2 < 1 \leq \theta_1$; in which case $a_2' > 0$ (see (2.45)). This case is completely similar to (ii).

(iv) $\theta_1 \geq 1, \theta_2 \geq 1$.

(iv-a) Suppose $a - xb \geq 0$, and $a' - xb' \geq 0$. Then the objective function in (2.51) either is convex or starts with a concave piece, followed by a convex piece, and the optimal solutions N_1^* and L_1^* are similar to (i-b).

(iv-b) Suppose $a - xb < 0$, and $a' - xb' < 0$. Then the objective function in (2.51) either is concave or starts with a convex piece, followed by a concave piece, and the optimal solutions N_1^* and L_1^* are similar to (ii-b).

(iv-c) Suppose $a - xb < 0$, and $a' - xb' \geq 0$.

(iv-c-1) Suppose $\mu_1 \geq \mu_2$. Then the convexity of objective function in (2.51) may have three cases. Namely, concave; starting with a convex piece followed by a concave piece; starting with a concave piece, switching to a convex piece, and switching to a concave piece. The optimal solutions N_1^* and L_1^* can be obtained by the approach discussed in (ii-b).

(iv-c-2) Suppose $\mu_1 < \mu_2$. Then the convexity of objective function in (2.51) may have three cases. Namely, convex; starting with a concave piece followed by a convex piece; starting with a convex piece, switching to a concave piece, and switching to a convex piece. The optimal solutions N_1^* and L_1^* are similar to (i-c-1).

(iv-d) Suppose $a - xb \geq 0$, and $a' - xb' < 0$. This case is completely similar to (iv-c).

For $L = -1$, because $0 \leq \alpha_2 \leq 1$ there are two cases:

(v) $\theta_1 < 1$. The solution is similar to (i).

(vi) $\theta_1 \geq 1$. The solution is similar to (iii).

2.2.2 Properties

In this subsection we look at some properties that help us to understand how the setup cost can affect the optimal policy.

Proposition 7. *The optimal threshold to open the second server N^* is increasing and the optimal threshold to shut down the second server L^* is decreasing in K. Furthermore, the cycle length between two consecutive actions to open server-2 is increasing in K.*

Proof. First, write the objective function as

$$\mathcal{AC}(L,N,K) = \frac{1}{g(L,N)}\Big[\Big(f(L,N)-K-\frac{\mu_1}{\mu_2}(1-L^-)K\Big)+K+\frac{\mu_1}{\mu_2}(1-L^-)K\Big].$$

By noticing that $(f(L,N) - K - \frac{\mu_1}{\mu_2}(1 - L^-)K)$ does not contain K, for $\widetilde{K} \geq K$,

$$\mathcal{AC}(L,N,\widetilde{K}) - \mathcal{AC}(L,N,K) = \frac{1}{g(L,N)}\Big(1+\frac{\mu_1}{\mu_2}(1-L^-)\Big)(\widetilde{K}-K). \quad (2.55)$$

Thus for $\widetilde{K} \geq K$,

$$\begin{aligned}
&\Big[\mathcal{AC}(L,N+1,\widetilde{K}) - \mathcal{AC}(L,N+1,K)\Big] - \Big[\mathcal{AC}(L,N,\widetilde{K}) - \mathcal{AC}(L,N,K)\Big] \\
&\quad = \frac{g(L,N)-g(L,N+1)}{g(L,N)g(L,N+1)}\Big(1+\frac{\mu_1}{\mu_2}(1-L^-)\Big)(\widetilde{K}-K), \\
&\Big[\mathcal{AC}(L+1,N,\widetilde{K}) - \mathcal{AC}(L+1,N,K)\Big] - \Big[\mathcal{AC}(L,N,\widetilde{K}) - \mathcal{AC}(L,N,K)\Big]
\end{aligned}$$

$$= \Big[\frac{g(L,N)-g(L+1,N)}{g(L,N)g(L+1,N)}\Big(1+\frac{\mu_1}{\mu_2}(1-L^-)\Big)+\frac{1}{g(L+1,N)}\frac{\mu_1}{\mu_2}L^-\Big](\widetilde{K}-K),$$

where

$$\begin{aligned} & g(L,N+1)-g(L,N) \\ & \quad = -b\theta_1^N\overline{\theta}_1+b_1-L^-b_e\alpha_2^N\overline{\alpha}_2-\frac{\mu_1}{\mu_2}(1-L^-)\Big(b'\theta_2^N\overline{\theta}_2-b_1'\Big), \end{aligned} \tag{2.56}$$

$$\begin{aligned} & g(L+1,N)-g(L,N) \\ & \quad = \begin{cases} b\theta_1^L\overline{\theta}_1-b_1+\frac{\mu_1}{\mu_2}\Big(b'\theta_2^L\overline{\theta}_2-b_1'\Big), & \text{if } L\geq 0, \\ -\Big(b_0+b_e\alpha_2^N\Big)+\frac{\mu_1}{\mu_2}\Big(b'(\theta_2^N-1)+b_1'N\Big), & \text{if } L=-1. \end{cases} \end{aligned} \tag{2.57}$$

We want to show the objective function $\mathcal{AC}(L,N,K)$ is submodular in (N,K) and supermodular in (L,K). Consequently, from the monotone and antitone properties associated with minimizing submodular and supermodular functions (refer to [33]), we know the optimal solution N^* is increasing in K and L^* is decreasing in K. This, in turn, implies the desired result,

$$g(L^*(K+1),N^*(K+1))\geq g(L^*(K+1),N^*(K))\geq g(L^*(K),N^*(K)).$$

So, we next show $g(L,N+1)-g(L,N)\geq 0$, as $g(L+1,N)-g(L,N)\leq 0$

is completely analogous by

$$b_0 + b_e\alpha_2^N = \frac{(\mu+\lambda\alpha_2)\big(\mu_1\mu_2\alpha_2^N + \mu_1^2(1-\alpha_2^N) + \lambda\mu + \mu_2^2\big)}{\beta\mu_1\mu_2\overline{\alpha}_2(\mu+2\lambda)} \geq 0, \text{ and } b' \geq 0,$$

under (2.1). To do so, it suffices to show

$$-b\overline{\theta}_1 + b_1 - L^- b_e\alpha_2^N\overline{\alpha}_2 + \frac{\mu_1}{\mu_2}(1-L^-)\Big(-b'\overline{\theta}_2 + b_1'\Big) \geq 0, \tag{2.58}$$

in both cases of $\theta_i \geq 1$ and $\theta_i < 1$, $i \in \{1,2\}$, as evident from (2.56) (Note $b > 0$ and $b' > 0$.) Making use of equations (2.34)-(2.41), we can write

$$-b\overline{\theta}_1 + b_1 = \frac{\rho_1+1}{\lambda\rho_1} + \frac{1}{\beta}, \; -b'\overline{\theta}_2 + b_1' = \frac{\rho_2+1}{\lambda\rho_2} + \frac{1}{\beta}.$$

For $L \geq 0$, (2.58) is true; for $L = -1$, if $b_e \leq 0$, (2.58) is also true, otherwise $b_e > 0$, and we show $-b\overline{\theta}_1 + b_1 - b_e\alpha_2^N\overline{\alpha}_2 \geq 0$. It's sufficient to show $-b\overline{\theta}_1 + b_1 - b_e\overline{\alpha}_2 \geq 0$ because it becomes less negative as N increases. By using again equations (2.34)-(2.41), we have

$$-b\overline{\theta}_1 + b_1 - b_e\overline{\alpha}_2 = \frac{\rho_1+1}{\lambda\rho_1} + \frac{\mu_1\mu + \lambda\big(\mu_2(1+\overline{\alpha}_2) + \alpha_2\mu_1\big)}{\beta\mu_2(\mu+2\lambda)} \geq 0.$$

□

Proposition 8. *Assume that $\lambda\overline{\alpha}_1/\mu_i \geq 1$ with $i = 1, 2$. There exists a finite*

K_0 such that when the setup cost K goes beyond K_0, the optimal threshold to shut down the second server, L^*, would be -1, that is, the optimal threshold to shut down one of two operating servers is for the system to become empty.

Proof. For each setup cost K, let $N^*(K)$ and $L^*(K)$ be the optimal thresholds to open and shut down the second server, respectively. First note that $\lambda\overline{\alpha}_1/\mu_i \geq 1$ is equivalent to $\theta_i \leq 1$. By the monotonicity of $N^*(K)$ and $L^*(K)$ given by Proposition 7, to prove the proposition, it suffices to show that there exists a $K_0 \geq 1$ such that for $N \geq N^*(K_0)$ and $L \leq L^*(1)$,

$$\frac{f(L+1,N)}{g(L+1,N)} \geq \frac{f(L,N)}{g(L,N)}. \tag{2.59}$$

We rewrite $f(L,N)/g(L,N)$ as

$$\frac{f(L,N)}{g(L,N)} := \frac{f_1(L,N) + \frac{\mu_1}{\mu_2} f_2(L,N)}{g_1(L,N) + \frac{\mu_1}{\mu_2} g_2(L,N)},$$

where

$$\begin{aligned} f_1(L,N) &= a(\theta_1^N - \theta_1^{L^+}) + a_2(N^2 - (L^+)^2) + a_1(N - L^+) + K \\ &\quad + (a_0 + a_e\alpha_2^N)L^-, \\ f_2(L,N) &= a'(\theta_2^N - \theta_2^{L^+}) + a_2'(N^2 - (L^+)^2) + a_1'(N - L^+) + K, \\ g_1(L,N) &= b(\theta_1^N - \theta_1^{L^+}) + b_1(N - L^+) + (b_0 + b_e\alpha_2^N)L^-, \end{aligned}$$

$$g_2(L,N) = b'(\theta_2^N - \theta_2^{L^+}) + b_1'(N - L^+).$$

To show (2.59) is true for $N \geq N^*(K_0)$ and $L \leq L^*(1)$, it's sufficient to prove the following four inequalities are true,

$$\frac{f_1(L+1,N)}{g_1(L+1,N)} \geq \frac{f_1(L,N)}{g_1(L,N)}, \quad \frac{f_2(L+1,N)}{g_2(L+1,N)} \geq \frac{f_2(L,N)}{g_2(L,N)}, \tag{2.60}$$

$$\frac{f_1(L+1,N)}{g_2(L+1,N)} \geq \frac{f_1(L,N)}{g_2(L,N)}, \quad \frac{f_2(L+1,N)}{g_1(L+1,N)} \geq \frac{f_2(L,N)}{g_1(L,N)}. \tag{2.61}$$

We first look at inequality $f_1(L+1,N)/g_1(L+1,N) \geq f_1(L,N)/g_1(L,N)$ in (2.60). After a simplification, this is equivalent to show that

$$\begin{aligned}&\Big[b\theta_1^L\overline{\theta}_1 - b_1\Big] \cdot \Big[a(\theta_1^N - \theta_1^L) + a_2(N^2 - L^2) + a_1(N-L) + K\Big] \\ &\quad \leq \Big[b(\theta_1^N - \theta_1^L) + b_1(N-L)\Big] \times \Big[a\theta_1^L\overline{\theta}_1 - a_2(2L+1) - a_1\Big]. \end{aligned} \tag{2.62}$$

In view of $b\theta_1^L\overline{\theta}_1 - b_1 < 0$ when $\theta_1 < 1$, it is sufficient to show that there exists a K_1 for $N \geq N^*(K_1)$ and $L \leq L^*(1)$,

$$\begin{aligned}K \geq &- a(\theta_1^N - \theta_1^L) - a_2(N^2 - L^2) - a_1(N-L) \\ &+ \frac{1}{b\theta_1^L\overline{\theta}_1 - b_1} \times \Big[b(\theta_1^N - \theta_1^L) + b_1(N-L)\Big] \times \Big[a\theta_1^L\overline{\theta}_1 - a_2(2L+1) - a_1\Big]. \end{aligned} \tag{2.63}$$

Hence, if we can find a finite $\widetilde{K}_1$ and an upper bound for the right-hand side of (2.63) on the region $\{(N, L) : N \geq N^*(\widetilde{K}_1), L \leq L^*(1)\}$, then setting K_1 just to be the maximum between $\widetilde{K}_1$ and this upper bound, we have (2.63) for $K \geq K_1$. In the remain of the proof, we identify a $\widetilde{K}_1$ and build an upper bound on the right-hand side of (2.63) on the region $\{(N, L) : N \geq N^*(\widetilde{K}_1), L \leq L^*(1)\}$. Note that, by again $\theta_1 < 1$ and the monotonicity of $L^*(\cdot)$,

$$\begin{aligned}
&- a(\theta_1^N - \theta_1^L) - a_2(N^2 - L^2) - a_1(N - L) \\
&+ \frac{1}{b\theta_1^L\overline{\theta}_1 - b_1} \times \Big[b(\theta_1^N - \theta_1^L) + b_1(N - L)\Big] \times \Big[a\theta_1^L\overline{\theta}_1 - a_2(2L + 1) - a_1\Big] \\
&\leq |a| + \frac{b(|a| + |a_1|)}{b_1 - b\overline{\theta}_1} + \frac{a_2 b}{b_1 - b\overline{\theta}_1}(2L^*(1) + 1) \\
&\quad - \Big[a_2(N + L) + a_1 - \frac{b_1}{b\theta_1^L\overline{\theta}_1 - b_1}\Big(a\theta_1^L\overline{\theta}_1 - a_2(2L + 1) - a_1\Big)\Big](N - L) \\
&\leq |a| + \frac{b(|a| + |a_1|)}{b_1 - b\overline{\theta}_1} + \frac{a_2 b}{b_1 - b\overline{\theta}_1}(2L^*(1) + 1) \\
&\quad - a_2 \cdot N(N - L^*(1)) + \Big[|a_1| + \frac{(|a| + |a_1| + a_2(2L^*(1) + 1)) \cdot b_1}{b_1 - b\overline{\theta}_1}\Big] \cdot (N + 1).
\end{aligned} \tag{2.64}$$

Next we prove the non-positivity of the last expression in (2.64) when N is large enough. Let $\widetilde{N}_1$ be the solution (larger one) to the following quadratic equation of N

$$a_2 \cdot N(N - L^*(1)) - \Big[|a_1| + \frac{(|a| + |a_1| + a_2(2L^*(1)+1)) \cdot b_1}{b_1 - b\overline{\theta}_1}\Big] \cdot (N+1)$$
$$= |a| + \frac{b(|a| + |a_1|)}{b_1 - b\overline{\theta}_1} + \frac{a_2 b}{b_1 - b\overline{\theta}_1}(2L^*(1)+1).$$

Let $\widetilde{K}_1$ be the solution given by $\widetilde{N}_1 = N^*(\widetilde{K}_1)$. For $N > \widetilde{N}_1$ $(:= N^*(\widetilde{K}_1))$, we have

$$a_2 \cdot N(N - L^*(1)) - \Big[|a_1| + \frac{(|a| + |a_1| + a_2(2L^*(1)+1)) \cdot b_1}{b_1 - b\overline{\theta}_1}\Big] \cdot (N+1)$$
$$\geq |a| + \frac{b(|a| + |a_1|)}{b_1 - b\overline{\theta}_1} + \frac{a_2 b}{b_1 - b\overline{\theta}_1}(2L^*(1)+1). \tag{2.65}$$

Combining (2.64)-(2.65) yields an upper bound for the right-hand side of (2.63). That is, for $N \geq N^*(\widetilde{K}_1)$, we have

$$- a(\theta_1^N - \theta_1^L) - a_2(N^2 - L^2) - a_1(N - L)$$
$$+ \frac{1}{b\theta_1^L\overline{\theta}_1 - b_1} \times \Big[b(\theta_1^N - \theta_1^L) + b_1(N-L)\Big] \times \Big[a\theta_1^L\overline{\theta}_1 - a_2(2L+1) - a_1\Big] \leq 0.$$

This implies (2.63) for $K \geq K_1 = \widetilde{K}_1 \vee 1$. We can follow the same procedure to prove the second inequality in (2.60) and another two inequalities in (2.61), and derive corresponding K_2, K_3, and K_4. Hence setting

$$K_0 = K_1 \vee K_2 \vee K_3 \vee K_4,$$

we have the proposition. $\square$

2.3 Random-Walk Method

In view of Proposition 8, with the help of the random-walk theory, this section devotes to develop a method to approximate $\mathsf{E}T_i$ and C_i with $L = -1$, and then to provide approximations for the expected long-run average cost. To the end, we first give some preliminary results on the random walks.

2.3.1 Preliminary Results

We consider a simple random walk

$$S_0 := 0, \qquad S_n := X_1 + \cdots + X_n,$$

where X_i's are i.i.d. random variables with

$$\mathsf{Pr}(X_i = 1) = p \;\text{ and }\; \mathsf{Pr}(X_i = -1) = \overline{p}.$$

Write

$$\gamma := 2p - 1 = \mathsf{E}X_i \;\text{ and }\; \sigma^2 := 1 - \gamma^2 = \mathsf{Var}(X_i).$$

Define the stopping time $T_{(-B,A)}$ by

$$T_{(-B,A)} = \min\{n : S_n \geq A \text{ or } S_n \leq -B\} \quad \text{with } A, B > 0.$$

Let Y_n's be nonnegative i.i.d. random variables such that $\{Y_i, i \geq n\}$ is independent of $\{X_1, \cdots, X_{n-1}\}$ and $\mathsf{E}Y_1 < \infty$. Then we have the following results.

Lemma 9. (Two Absorbing Barriers) *Assume that* $\gamma = \mathsf{E}(X_i) \neq 0$. (i) $\mathsf{E}T_{(-B,A)} = \frac{A[1-(\overline{p}/p)^{-B}]-B[(\overline{p}/p)^{A}-1]}{\gamma[(\overline{p}/p)^{A}-(\overline{p}/p)^{-B}]}$; (ii) $\mathsf{Pr}(S_{T_{(-B,A)}} = A) = \frac{(\overline{p}/p)^{B}-1}{(\overline{p}/p)^{A+B}-1}$; (iii) $\mathsf{Pr}(S_{T_{(-B,A)}} = -B) = \frac{(\overline{p}/p)^{A+B}-(\overline{p}/p)^{B}}{(\overline{p}/p)^{A+B}-1}$; (iv) $\mathsf{E}\Big(\sum_{i=1}^{T_{(-B,A)}} Y_i\Big) = \mathsf{E}T_{(-B,A)} \times \mathsf{E}Y_1$; (v) *For any constant* D, $\mathsf{E}\Big(\sum_{i=1}^{T_{(-B,A)}}(D + S_{i-1})Y_i\Big) = \Big[\Big(D + \frac{A-B}{2} - \frac{1}{2\gamma}\Big)\mathsf{E}T_{(-B,A)} + \frac{AB}{2\gamma}\Big] \cdot \mathsf{E}Y_1$.

Proof. The first three results directly follow from the random walk theory (see, for example, [31]). (iv), by noting that T is a stopping time for the sequence $\{Y_n, n \geq 1\}$, follows from Wald's equation. Now we show (v). Let $\mathcal{F}_n$ be the sigma field generated by $\{(X_i, Y_i), i = 1, \ldots, n\}$. Note both $\{S_n\}$ and $\{Y_n\}$ are adapted to the filtration $\{\mathcal{F}_n, n \geq 1\}$. Hence

$$\mathsf{E}\Big(\sum_{i=1}^{n} S_{i-1}Y_i \Big| \mathcal{F}_{n-1}\Big) = \sum_{i=1}^{n-1} S_{i-1}Y_i + \mathsf{E}(S_{n-1}Y_n|\mathcal{F}_{n-1}) = \sum_{i=1}^{n-1} S_{i-1}Y_i + S_{n-1}\mathsf{E}Y_n, \tag{2.66}$$

where we use the fact that Y_n is independent of $\mathcal{F}_{n-1}$. Taking expectation on both sides of (2.66), we have

$$\mathsf{E}\Big(\sum_{i=1}^{n} S_{i-1}Y_i\Big) = \mathsf{E}\Big(\sum_{i=1}^{n-1} S_{i-1}Y_i\Big) + (\mathsf{E}S_{n-1}) \times \mathsf{E}Y_n = \cdots = \mathsf{E}\Big(\sum_{i=1}^{n-1} S_i\Big) \times \mathsf{E}Y_1.$$

Hence

$$\begin{aligned}\mathsf{E}\Big(\sum_{i=1}^{T} S_{i-1}Y_i\Big) &= \mathsf{E}Y_1 \times \mathsf{E}\Big(\sum_{i=1}^{T-1} S_i\Big) = \mathsf{E}Y_1 \times \Big[\mathsf{E}\Big(\sum_{i=1}^{T} S_i\Big) - \mathsf{E}S_T\Big] \\ &= \mathsf{E}Y_1 \times \Big[\mathsf{E}\Big(\sum_{i=1}^{T} S_i\Big) - \gamma \times \mathsf{E}T\Big]. \end{aligned} \tag{2.67}$$

Now we consider $\mathsf{E}\Big(\sum_{i=1}^{T} S_i\Big)$. Write $X_i = \xi_i + \gamma$, where ξ_i's are i.i.d., and $\mathsf{E}\xi_i = 0$, $\mathsf{Var}(\xi_i) = \sigma^2$, we have

$$\begin{aligned}\sum_{i=1}^{T} S_i = \sum_{i=1}^{T}\sum_{n=1}^{i} X_n &= \sum_{n=1}^{T}(T-n+1)X_n = TS_T - \sum_{n=1}^{T}(n-1)(\xi_n + \gamma) \\ &= TS_T - \frac{\gamma}{2}T(T-1) - \sum_{n=1}^{T}(n-1)\xi_n.\end{aligned}$$

The last term above is a martingale. Hence

$$\mathsf{E}\Big(\sum_{i=1}^{T} S_i\Big) = \mathsf{E}(TS_T) - \frac{\gamma}{2}\mathsf{E}T^2 + \frac{\gamma}{2}\mathsf{E}T. \tag{2.68}$$

Applying optimal stopping theorem to the martingale $\{(S_n - n\gamma)^2 - n\sigma^2\}$

yields

$$\begin{aligned}
\mathsf{E}\Big(S_T - T\gamma\Big)^2 = \sigma^2\mathsf{E}T \quad &\Rightarrow \quad \mathsf{E}(S_T - T\gamma)^2 = \sigma^2\mathsf{E}T \\
&\Rightarrow \quad \mathsf{E}S_T^2 + \gamma^2\mathsf{E}T^2 - 2\gamma\mathsf{E}(S_T T) = \sigma^2\mathsf{E}T \\
&\Rightarrow \quad \mathsf{E}(S_T T) = \frac{\mathsf{E}S_T^2}{2\gamma} - \frac{1-\gamma^2}{2\gamma}\mathsf{E}T + \frac{\gamma}{2}\mathsf{E}T^2. \qquad (2.69)
\end{aligned}$$

Plug (2.69) into (2.68) and we have

$$\mathsf{E}\Big(\sum_{i=1}^{T} S_i\Big) = \frac{\mathsf{E}S_T^2}{2\gamma} + \frac{2\gamma^2 - 1}{2\gamma}\mathsf{E}T.$$

Since, by $\mathsf{E}S_T = \gamma \cdot \mathsf{E}T = A\pi_A - B(1-\pi_A)$,

$$\mathsf{E}S_T^2 = A^2\pi_A + B^2(1-\pi_A) = (A-B)\gamma\mathsf{E}T + AB,$$

we simplify $\mathsf{E}\Big(\sum_{i=1}^{T} S_i\Big)$ as

$$\mathsf{E}\Big(\sum_{i=1}^{T} S_i\Big) = \Big(\frac{A-B}{2} + \gamma - \frac{1}{2\gamma}\Big)\mathsf{E}T + \frac{AB}{2\gamma}.$$

Thus, by (2.67), we have

$$\mathsf{E}\Big(\sum_{i=1}^{T} S_{i-1}Y_i\Big) \quad = \quad \mathsf{E}Y_1 \times \Big[\Big(\frac{A-B}{2} + \gamma - \frac{1}{2\gamma}\Big)\mathsf{E}T + \frac{AB}{2\gamma} - \gamma\mathsf{E}T\Big]$$

$$= \mathsf{E}Y_1 \times \Big[\Big(\frac{A-B}{2} - \frac{1}{2\gamma}\Big)\mathsf{E}T + \frac{AB}{2\gamma}\Big].$$

This gives (v). Therefore, the lemma is proved. □

Now define an one-barrier stopping time $T_{(-B,\infty)}$ with negative drift as

$$T_{(-B,\infty)} = \inf\{n : S_n \le -B\} \ \text{ with } B > 0.$$

Similar to Lemma 9, we have the following result.

Lemma 10. (One Absorbing Barrier with Negative Drift) *Assume that* $\gamma < 0$. (i) $\mathsf{E}T_{(-B,\infty)} = -\frac{B}{\gamma}$; (ii) $\mathsf{E}\Big(\sum_{i=1}^{T_{(-B,\infty)}} Y_i\Big) = -\frac{B}{\gamma} \times \mathsf{E}Y_1$; (iii) *For any constant* D, $\mathsf{E}\Big(\sum_{i=1}^{T_{(-B,\infty)}}(D + S_{i-1})Y_i\Big) = \frac{B}{2\gamma}\Big[-2D + B + \frac{1}{\gamma}\Big] \cdot \mathsf{E}Y_1$.

Proof. Going along the line of the proof of Lemma 9, the lemma can be proved similarly. □

2.3.2 Random-Walk Approximations

To obtain approximations for the expected long-run average cost, we build a connection between the ticket queue given in Section 2.1 and the random-walk studied in the above subsection. The connection is characterized by an one-to-one mapping between the dynamics of the ticket position $Q(t)$ and the random walk. Formally, a customer arrival in the system will be considered to be a right-side-movement for the random walk, while a service completion in the system will be considered to be a left-side-movement for the random walk. As each service completion will deplete $1/\overline{\alpha}_i$ tickets on the average (when there are i working servers), we can consider the service rate to be $\hat{\mu}_i$ or $\hat{\mu}$ from the customer ticket perspective. Further, noting that only customer arrivals and service completions can change the dynamics of the ticket position, the expectation of the sojourn time is $1/(\lambda+\hat{\mu}_i)$ $(1/(\lambda+\hat{\mu}))$ for each system-nonempty state with server-i working (2 working servers), and $1/\lambda$ for the system-empty state. Thus, to approximate T_1, we just consider a random walk with -1 as a reflecting barrier and N as an absorbing barrier, and calculate how many steps for the random walk to be absorbed. Hence, by Lemma 9 with $p = \lambda/(\lambda + \hat{\mu}_1)$, we have

$$\begin{aligned} \mathsf{E}T_1 &\approx \sum_{k=0}^{\infty}(1 - \mathsf{Pr}(T_{(-1,N)} = N))^k \cdot \left[\frac{1}{\lambda} + \mathsf{E}T_{(-1,N)} \times \frac{1}{\lambda + \hat{\mu}_1}\right] \\ &= \frac{N+1}{\beta_1} - \frac{\hat{\mu}_1}{\beta_1^2}\left[1 - \frac{1}{(\overline{\alpha}_1\rho_1)^{N+1}}\right] := \mathsf{E}T_1^{rw}. \end{aligned} \tag{2.70}$$

To get the approximation of the expected system-empty time $\mathsf{E}T_0$ (see Lemma 6), note that the zero ticket-position, at which the original system visits each time in one regenerative cycle, just corresponds that the random walk moves to the reflecting barrier. Hence,

$$\mathsf{E}T_{10} \approx \sum_{k=0}^{\infty}(1 - \mathsf{Pr}(T_{(-1,N)} = N))^k \cdot \frac{1}{\lambda} = \frac{1}{\beta_1}\left[1 - \frac{1}{(\overline{\alpha}_1\rho_1)^{N+1}}\right] := \mathsf{E}T_{10}^{rw}. \tag{2.71}$$

Consider the approximation of the expected two-server region part $\mathsf{E}T_2$. As the ticket position increases to $(N+1)$, the system enters into the two-server region, and then leaves it until the system becomes empty. Thus the corresponding random-walk will be set to start with N. Furthermore, when the system has one ticket ($Q(t)=1$) in the two-server region, only one server works even the other server is in operating state. Specifically, the probability for server-i working is μ_{3-i}/μ. This observation indicates that it is necessary to modify our random walk's sojourn time in state 0 to $1/[\lambda+2\mu_1\mu_2/(\mu\overline{\alpha}_2)]$. When the random walk moves into state 0, it will move to -1 in exactly one step with probability $2\mu_1\mu_2/(\lambda\mu\overline{\alpha}_2+2\mu_1\mu_2)$ ($:=\pi_{(-1,1)}$). Thus, from Lemma 10 with $p=\lambda/(\lambda+\mu_2)$,

$$\begin{aligned}
\mathsf{E}T_2 &\approx \mathsf{E}T_{(-N,\infty)}\frac{1}{\lambda+\hat{\mu}}+\frac{1}{\lambda+2\mu_1\mu_2/(\mu\overline{\alpha}_2)} \\
&\quad +\sum_{k=1}^{\infty}k(1-\pi_{(-1,1)})^k\pi_{(-1,1)}\Big[\mathsf{E}T_{(-1,\infty)}\frac{1}{\lambda+\hat{\mu}}+\frac{1}{\lambda+2\mu_1\mu_2/(\mu\overline{\alpha}_2)}\Big] \\
&= \frac{N}{\beta}+\frac{\mu^2}{2\mu_1\mu_2\beta}:=\mathsf{E}T_2^{rw}. \qquad (2.72)
\end{aligned}$$

Note that the state 0 of the random walk corresponds to one server busy and the other one is idle but in operating state. Consequently, there is no customer abandonment in this case. Thus, when considering abandonment cost, we need to know how many times the random walk visits state 0 during T_2. Based on the above analysis, it is straightforward to see that the average number of times to visit state 0 is

$$\pi_{(-1,1)}+2(1-\pi_{(-1,1)})\pi_{(-1,1)}+\cdots=\frac{1}{\pi_{(-1,1)}}.$$

Hence,

$$\mathsf{E}T_{20} \approx \frac{1}{\lambda + 2\mu_1\mu_2/(\mu\overline{\alpha}_2)} \times \frac{1}{\pi_{(-1,1)}} = \frac{\mu\overline{\alpha}_2}{2\mu_1\mu_2} := \mathsf{E}T_{20}^{rw}. \tag{2.73}$$

For the customer delay cost, note that only non-abandonment customers get the delay cost payment. When i servers operate, we will pay $h(1-\alpha_i)\times k$ on the average if there are k tickets waiting to be called. Thus, $h(1-\alpha_1)\times(k-1)^+$ will be charged if the corresponding random walk moves at k when one server operates. Hence, from Lemma 9 with $D=0, A=N, B=N$,

$$\begin{aligned} C_1 \approx& \frac{h\overline{\alpha}_1}{1-\mathsf{Pr}(T_{(-1,N)}=N)}\Big[\Big(\frac{N-1}{2} - \frac{\lambda+\hat{\mu}_1}{2\beta_1}\Big)\mathsf{E}T_{(-1,N)} + \frac{N(\lambda+\hat{\mu}_1)}{2\beta_1}\Big]\frac{1}{\lambda+\hat{\mu}_1} \\ =& \frac{\overline{\alpha}_1}{2\beta_1}hN^2 - \frac{\mu_1(1+\overline{\alpha}_1\rho_1)}{2\beta_1^2}hN + \frac{\mu_1\lambda h}{\beta_1^3}\Big(1-\frac{1}{(\overline{\alpha}_1\rho_1)^N}\Big) := C_1^{rw}. \end{aligned} \tag{2.74}$$

Finally consider the approximation for the customer delay cost in the two-server region, C_2. There is no delay cost incurred when the ticket position is one or two ($Q(t)=1,2$). After the random-walk moves to 1, the system will incur the delay cost only when the random-walk moves to 2 in the next step. So the approximation will be decomposed into two parts: the delay cost for the period in which the random-walk will first move to 1 starting with N; and the delay cost for the period in which the random-walk first move to 1 starting with 2. By Lemma 10 with $D=N-1, B=N-1$ and $D=1, B=1$ respectively, the first part cost is given by

$$\frac{h\overline{\alpha}_2}{2}\cdot\frac{\hat{\mu}+\lambda}{\beta}\Big((N-1)+\frac{\hat{\mu}+\lambda}{\beta}\Big)\cdot\frac{N-1}{\hat{\mu}+\lambda},$$

and the second part is

$$\frac{h\overline{\alpha}_2}{2} \cdot \frac{\hat{\mu}+\lambda}{\beta}\Big(1+\frac{\hat{\mu}+\lambda}{\beta}\Big) \cdot \frac{1}{\hat{\mu}+\lambda}.$$

Note that the probability that starting with 1, the random walk reaches -1 before reaching 2 is $2/[\rho_1\rho_2\overline{\alpha}_2^2 + 2(\rho\overline{\alpha}_2+1)]$ $(:= \pi_{(-1,2)})$. Therefore,

$$\begin{aligned} C_2 &\approx \frac{h\overline{\alpha}_2}{2} \cdot \frac{\hat{\mu}+\lambda}{\beta}\Big((N-1)+\frac{\hat{\mu}+\lambda}{\beta}\Big) \cdot \frac{N-1}{\hat{\mu}+\lambda} \\ &\quad + \sum_{k=1}^{\infty} k(1-\pi_{(-1,2)})^k \pi_{(-1,2)} \cdot \frac{h\overline{\alpha}_2}{2} \cdot \frac{\hat{\mu}+\lambda}{\beta}\Big(1+\frac{\hat{\mu}+\lambda}{\beta}\Big) \cdot \frac{1}{\hat{\mu}+\lambda} \\ &= \frac{\overline{\alpha}_2}{2\beta} hN^2 + \frac{\mu}{2\beta^2}(3\overline{\alpha}_2\rho - 1)hN + \frac{\overline{\alpha}_2^2 \mu\rho_1\rho_2}{2\beta^2} h := C_2^{rw}. \end{aligned} \tag{2.75}$$

In view of (2.70)-(2.75), then our long-run average cost can be approximated by

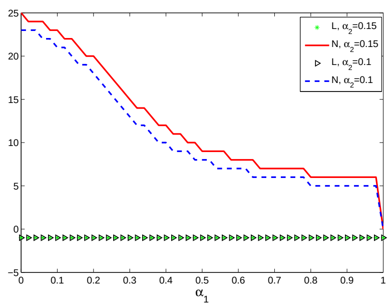
$$\begin{aligned} \mathcal{AC}(-1,N) &\approx \frac{1}{\mathsf{E}T_1^{rw} + \mathsf{E}T_2^{rw}} \Big[\lambda\alpha_1(\mathsf{E}T_1^{rw} - \mathsf{E}T_{10}^{rw}) + (\lambda\alpha_2 + c_2)\mathsf{E}T_2^{rw} \\ &\qquad - \lambda\alpha_2 \mathsf{E}T_{20}^{rw} + c_1(\mathsf{E}T_1^{rw} + \mathsf{E}T_2^{rw}) + C_1^{rw} + C_2^{rw} + K\Big] \\ &:= \mathcal{AC}^{rw}(-1,N). \end{aligned} \tag{2.76}$$

Following the fractional programming technique developed in Subsection 2.2.1, we can solve $\min_{N\geq 0} \mathcal{AC}^{rw}(-1,N)$. Let N^{rw*} be its solution. Compared with the exact analysis developed in Section 2.1, the random-walk method provides a unified and simpler approach to evaluate the system performance measures such as the expectations of one-server and two-server regions, the system cumulative idle times, and the customer delay cost. Of course, when $\alpha_1 = \alpha_2 = 0$, we know that the exact analysis and the random-walk approximation are same, that is, $\mathcal{AC}(-1, N) = \mathcal{AC}^{rw}(-1, N)$.

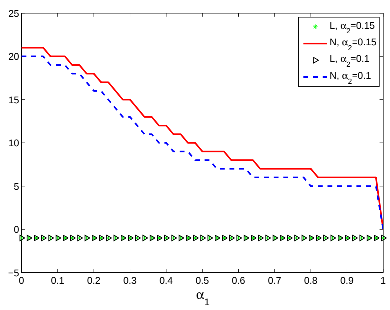
2.4 Numerical Studies

In this section we provide numerical results to show the sensitivity of the optimal policies with respect to the abandonment probabilities, the customer delay and operating costs, the efficiency of the approximations developed in Section 2.3, and the comparison with the results existing in the literature. First we look at the sensitivity.

2.4.1 Sensitivity



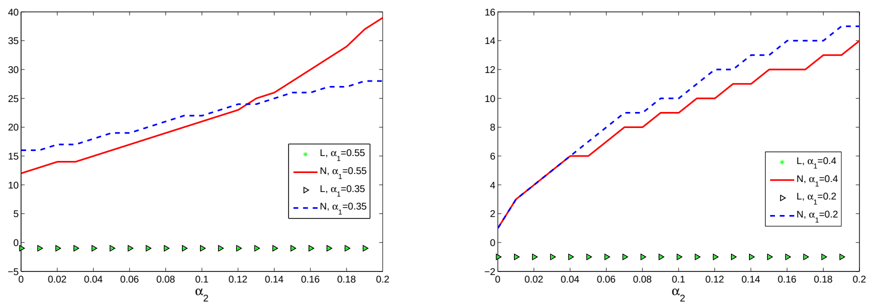
(a) $h = 0.6$



(b) $h = 0.8$

Fig. 2.3: α_1 Sensitivity Analysis

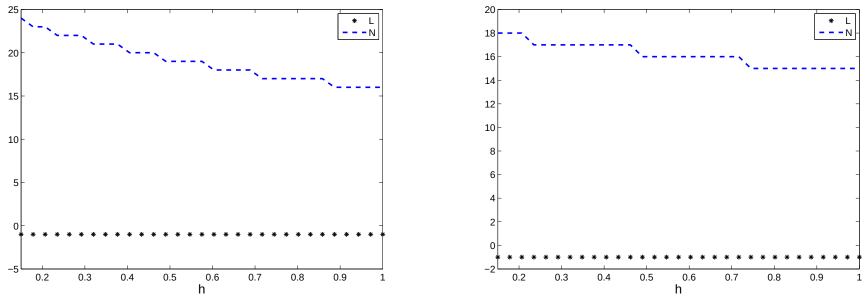
In Figure 2.3, we choose $(\lambda, \mu_1, \mu_2, K, c_1, c_2) = (160, 120, 100, 5, 0.12, 0.1)$. Figures 2.3 (a) and 2.3 (b) show that when the abandonment rate α_2 for the two-server region is smaller, the optimal threshold N^* of opening the second server is decreasing with respect to α_1. Furthermore, for each fixed one-server region abandonment rate α_1, the higher the abandonment rate α_2 is, the higher the optimal threshold of opening the second server is. The reason for this monotonicity is to reduce the customer abandonment cost by delaying opening the second server. However, for the optimal threshold L^* of shutting down the second operating server, with consideration of the setup cost already incurred, the system needs a longer cycle to consume the setup cost (that is, conservative to close the second operating server). Thus, L^* is very insensitive with respect to α_1. Compared Figure 2.3 (a) (customer delay cost $h = 0.6$) with Figure 2.3 (b) ($h = 0.8$), we can see when the customer delay cost gets higher, the second server will be opened earlier to reduced the delay cost.



(a) (μ_1, c_1)= (18, 0.19) (b) (μ_1, c_1)= (9, 0.09)

Fig. 2.4: α_2 Sensitivity Analysis

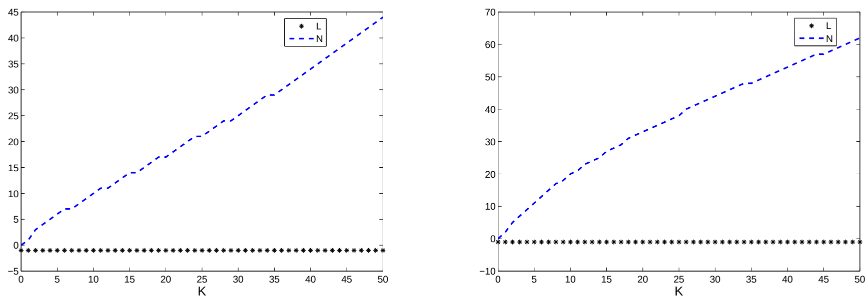
In Figure 2.4, we choose $(\lambda, \mu_2, h, K, c_2)$=(19,10,0.25, 10, 0.1). Figures 2.4 (a) and 2.4 (b) show the sensitivity about the abandonment rate α_2 in the two-server region. A comparison between two figures illustrates the higher the operating cost is, the later we put the second server into operation.



(a) $\alpha_2 = 0.15$ (b) $\alpha_2 = 0.05$

Fig. 2.5: h Sensitivity Analysis

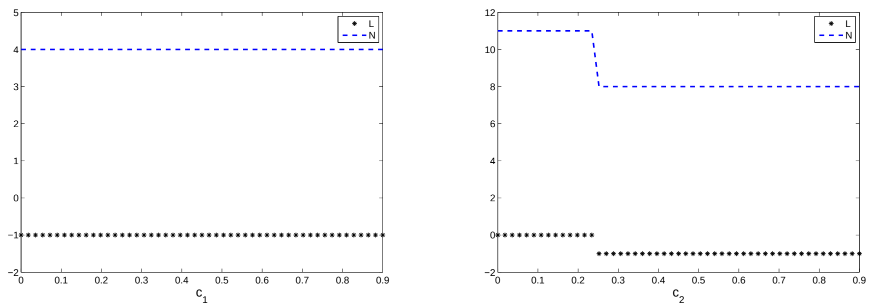
In Figure 2.5, we choose $(\alpha_1, \lambda, \mu_1, \mu_2, c_1, c_2, K)$=(0.4, 15, 13, 10,0.15,0.1, 10). Figures 2.5 (a) -2.5 (b) show that the optimal N^* decreases with respect to h. The reason is intuitive. The system could speed up the service rate by opening the second server earlier such that the customer delay cost can be reduced. Compared Figure 2.5 (b) with Figure 2.5 (a), we open the second server earlier to enjoy the lower abandonment cost from $\alpha_2 = 0.15$ to $\alpha_2 = 0.05$.



(a) $\alpha_2 = 0$ (b) $\alpha_2 = 0.2$

Fig. 2.6: K Sensitivity Analysis

In Figure 2.6 we choose $(\alpha_1, \lambda, \mu_1, \mu_2, h, c_1, c_2)$=(0.4,20, 15, 10, 0.25, 0.3, 0.2). Figures 2.6 (a) -2.6 (b) illustrate the monotonicity of the optimal threshold to open the second server with respect to the setup cost K, which is consistent with Proposition 7. The higher the setup cost K is, the higher the threshold to open the second server is. The figures shows the threshold to shut down the second server is not sensitive to increase the setup cost K. Compared Figure 2.6 (a) with Figure 2.6 (b), we open the second server earlier to enjoy the lower abandonment cost from $\alpha_2 = 0$ to $\alpha_2 = 0.2$.



(a) (μ_1, μ_2, c_2)=(8,10,0.2)

(b) (μ_1, μ_2, c_1)=(12.5,15,0.2)

Fig. 2.7: Sensitivity Analysis for Operation Cost

In Figure 2.7, we choose $(\alpha_1, \alpha_2, \lambda, h, K)$=(0.4,0.2,20,0.9,5). Figure 2.7 (a) shows that the optimal thresholds are very insensitive to the operation cost c_1. For Figure 2.7 (b), when c_2 increases, both L^* and N^* decrease. The reason for this follows from the utilization rate of server-2 with respect to its cost c_2. When we keep $L = -1$ and $N = 11$, the utilization rate of server-2 is given by $\mathsf{E}(T_{12} + T_2)/\mathsf{E}(T_{11} + T_{12} + T_2) = 0.6591$; and while the policy with $L = -1$ and $N = 8$, the utilization rate of server-2 is 0.3157. Thus if $c_2 \geq (\mu_2/\mu_1)c_1 = 0.24$, then server-2 is more expensive and the system enjoys its lower utilization; if $c_2 \leq (\mu_2/\mu_1)c_1 = 0.24$, then server-2 is cheaper and the system enjoys its higher utilization rate.

2.4.2 Accuracy of Random-Walk Approximation

In this subsection, we will make a comparison between the exact analysis given by Section 2.2 and the random-walk approximation given by Section 2.3.

In table 2.1, we choose $(\lambda, \mu_1, \mu_2, \alpha_2, c_1, c_2, h, K)$=(1,0.65,0.55,0,0.2,0.1,1, 50). In table 2.2, we choose $(\lambda, \mu_1, \mu_2, \alpha_2, c_1, c_2, h, K)$=(1,0.7,0.5,0.05,0.2,0.1, 0.4,20).

Tab. 2.1: Comparison between Exact and RW: I

α_1	Exact Analysis			*RW* Approx		
	L^*	N^*	cost	N^{rw*}	cost	error%
0	-1	3	6.0518	3	6.0518	0
0.2	-1	4	5.9354	4	5.9354	0
0.32	-1	4	5.8375	5	5.8544	0.29
0.36	-1	5	5.7970	6	5.8543	0.99
0.4	-1	5	5.7377	8	5.9161	3.11
0.42	-1	5	5.7073	12	6.1769	8.23

Tables 2.1-2.2 show that the customer abandonment rate at the one-server region makes a big impact on the accuracy of the random-walk approximation. If we look at the generator given by the Markov chain (see (2.2)) $\{(S_1(t), S_2(t), Q(t)), t \geq 0\}$, which is used to characterize the system state, and make a comparison with the random-walk method, we observe that the smaller the customer abandonment rate α_1 is, the more closer for the two generators corresponding the Markov chain (2.2) and the random-walk, re-

Tab. 2.2: Comparison between Exact and RW: II

α_1	Exact Analysis			*RW* Approx		
	L^*	N^*	cost	N^{rw*}	cost	error%
0	-1	4	2.2327	4	2.2327	0
0.2	-1	4	2.1938	5	2.1951	0.06
0.32	-1	5	2.1449	6	2.1530	0.37
0.36	-1	6	2.1247	7	2.1405	0.74
0.39	-1	6	2.1029	9	2.1496	2.22
0.42	-1	7	2.0771	228	2.3012	10.79

spectively. Hence, these two tables indicate that the higher the customer abandonment rate α_1 is, the more inaccurate the random-walk approximations incur.

In table 2.3, we choose $(\lambda, \mu_1, \mu_2, \alpha_2, c_1, c_2, h, K)$=(10,6.5,5.5,0,0.2,0.1,1,50). In table 2.4, we choose $(\lambda, \mu_1, \mu_2, \alpha_2, c_1, c_2, h, K)$=(100,70,50,0.05,0.2,0.1,0.4,20). Compared with Tables 2.1-2.2, Tables 2.3-2.4 have the higher arrival and service rates. Tables 2.3-2.4 show that the higher arrival rate and service rate can dilute the impact incurred by the customer abandonment rate. This can be explained by the law of the large number as the arrival and service rates get bigger and bigger, the mean of the arrivals (or service) plays a big role.

Tab. 2.3: Comparison between Exact and RW: III

α_1	Exact Analysis			*RW* Approx		
	L^*	N^*	cost	N^{rw*}	cost	error%
0	-1	11	13.0831	11	13.0831	0
0.2	-1	12	12.4240	11	12.4372	0.10
0.32	-1	13	11.6488	13	11.6488	0
0.36	-1	14	11.2646	14	11.2646	0
0.4	-1	17	10.7555	18	10.7722	0.15
0.42	-1	19	10.4215	25	10.5671	1.40

Tab. 2.4: Comparison between Exact and RW: IV

α_1	Exact Analysis			*RW* Approx		
	L^*	N^*	cost	N^{rw*}	cost	error%
0	-1	38	17.8058	.38	17.8058	0
0.2	-1	33	25.0262	32	25.0284	0.01
0.32	-1	25	31.2293	23	31.2730	0.14
0.36	-1	22	33.4578	20	33.5398	0.25
0.39	-1	21	35.1048	18	35.2360	0.37
0.42	-1	19	36.7045	17	36.8095	0.29

2.4.3 Comparison with Existing Results

Following our discussion in the introduction, our model also generalizes Zhang [38] when $L = -1$ and $\alpha_1 = \alpha_2 = 0$. So here we make a comparison with his result. As he does not consider operation cost, here we just let $\mu_1 = \mu_2$ and $c_1 = c_2 = 0$. Zhang [38] uses a fluid approximation for the one-server region and a diffusion approximation for the two-server region. As the exact analysis can more precisely capture the customer delay cost than just fluid approximation and diffusion approximation, the results obtained in Section 2.2 performs much better than Zhang [38].

In table 2.5, we choose $\lambda = 215$, $\mu_1 = 200$, $h = 0.25$. In table 2.6, we choose $\lambda = 205$, $\mu_1 = 190$, $h = 0.25$. Tables 2.5-2.6 show that the smaller the setup cost is, the bigger error Zhang [38] incurs. The reason is that a smaller setup cost will give a lower optimal threshold to open the second server. When the threshold of opening the second server becomes lower, the one-server region will get smaller, which consequently implies the time when we use one server will become shorter. It turns out inaccurate to use the fluid-model to approximate the original queueing model in the short time period even though the system evolves under the heavy traffic regime.

Tab. 2.5: Comparison between Exact and Existing Results: I

K	Exact Analysis		Zhang (2009)		
	N^*	cost	N^{z*}	cost	error%
0.1	7	1.0486	2	2.2246	112
1	19	2.9581	9	4.4254	49.6
10	45	8.5124	32	9.3543	9.9

Tab. 2.6: Comparison between Exact and Existing Results: II

K	Exact Analysis		Zhang (2009)		
	N^*	cost	N^{z*}	cost	error%
0.1	7	1.0290	2	2.1252	107
1	18	2.9231	9	4.2758	46.3
10	45	8.4573	32	9.2094	8.9

Tab. 2.7: Comparison between Exact and Existing Results: III

λ	Exact Analysis		Zhang (2009)		
	N^*	cost	N^{z*}	cost	error%
212	19	2.8321	8	4.8308	70.6
215	19	2.9581	9	4.4254	49.6
220	19	3.1611	11	3.9287	24.2
260	20	4.3905	17	4.4492	1.3
300	20	4.9420	19	4.9454	0.07

In table 2.7, we choose $\mu_1 = 200$, $h = 0.25$, $K = 1$. In table 2.8, we choose $\mu_1 = 110$, $h = 0.25$, $K = 1$. Tables 2.7-2.8 indicate that when the system cost parameters are fixed, the traffic intensity also impact the accuracy of the method proposed by Zhang [38]. Only when the traffic intensity becomes very high, Zhang [38] can perform better.

Tab. 2.8: Comparison between Exact and Existing Results: IV

λ	Exact Analysis		Zhang (2009)		
	N^*	cost	N^{z*}	cost	error%
120	15	2.3490	7	3.4933	48.7
130	15	2.7948	10	3.1135	11.4
140	15	3.1501	12	3.2494	3.2
170	14	3.7315	13	3.7477	0.44
190	12	3.9859	12	3.9859	0

Tab. 2.9: Comparison between Exact and Existing Results: V

h	Exact Analysis		Zhang (2009)		
	N^*	cost	N^{z*}	cost	error%
0.9	27	15.3640	13	24.1613	57.3
0.5	34	11.0284	18	15.5411	40.9
0.1	62	4.5541	43	5.0915	11.8

Tab. 2.10: Comparison between Exact and Existing Results: VI

h	Exact Analysis		Zhang (2009)		
	N^*	cost	N^{z*}	cost	error%
0.9	11	6.8697	6	8.8282	28.5
0.7	12	6.0320	7	7.5118	24.5
0.3	18	3.8862	12	4.3190	11.1

In table 2.9, we choose $\lambda = 210$, $\mu_1 = 200$, $K = 10$. In table 2.10, we choose $\lambda = 230$, $\mu_1 = 200$, $K = 1$. Tables 2.9-2.10 show that the customer delay cost also plays a big role in the approximation given by Zhang [38] regardless of the setup costs. Under either the higher setup cost ($K = 10$) or lower setup cost ($K = 1$), the higher the customer delay cost is, the more inaccurate the approximation proposed by Zhang [38]. The reason is that when the customer delay cost becomes higher, the system needs to use the second server to reduce the number of the waiting customers. In order to put the second server into use earlier, we need to pull down the threshold of opening the second server. This consequence again incurs a shorter period for the one-server region. Thus with the same reason shed by Tables 2.5-2.6, the method of Zhang [38] gives a big error when the customer delay cost increases.

2.5 Concluding Remarks

In this chapter we provide a study on the optimal staffing problem for a ticket queue with two staffing levels. The only information required to carry out the optimal policy is the ticket counts along with a count of customers served. Customer abandonment rates are assumed given, and as we outlined in the Introduction these rates can be readily estimated (also by simple counts of tickets and customers served). Thus, the optimal staffing rule is suitable for practical implementations.

The Markov chain and random walk analyses developed here can be readily extended to multiple staffing levels. To solve the optimal staffing problem in that more general setting, however, is a quite different matter. For instance, suppose there are $m > 2$ servers. We need to first address the issue, how many different staffing levels do we need to focus on? Only use either 1 server or m servers, or any other number of servers in between should also be considered? Only when this issue is resolved, then we can decide the corresponding thresholds upon which to switch up or down to the next staffing level. We will answer these questions in Chapter 3.

Another natural extension of this model is to allow customer abandonment rate to depend on both staffing levels and ticket queue length. When an arrival customer observes a shorter ticket queue, he is less likely to abandon and will choose to stay. On the other hand, when he observes a long ticket queue, he may choose to leave, but he can still come back later since his ticket occupies his position. Thus it's unclear whether long ticket queue will lead to a high abandonment rate. Another difficulty of this extension is: by incorporating ticket queue length into customer abandonment rate, we need to first address the problem of how to characterize system dynamics, which might be much more complex. We leave this extension for future study.

3. FLUID MODEL AND ASYMPTOTICS FOR TICKET QUEUES

In this chapter, we study the optimal staffing of the ticket queue with more than two staffing levels. Based on information from the ticket counts and previous service rate, we show that policy with two staffing levels is better than policy with multiple staffing levels, and the optimal threshold to change staffing level can be derived through the EOQ formula.

The main contributions of the study are as follows:

- Asymptotic optimal policy for staffing problem in ticket queues with customer abandonment;
- Simple structure of the asymptotic optimal policy;
- Fluid model for ticket queue with customer abandonment;
- Connection between EOQ model and fluid ticket queue.

This chapter is organized as follows. Section 3.1 introduces the details of the mathematical model. Section 3.2 derives the fluid model for the ticket queue. Analysis of the long-run average cost in the fluid model is given in Section 3.3, and the optimal staffing policy in the fluid model is given in Section 3.4. We show that the optimal policy derived from fluid model is asymptotic optimal in Section 3.5. Numerical results are given in Section 3.6. Concluding remarks are summarized in Section 3.7.

3.1 Problem Formulation

The queueing system has m identical servers available. Customers arrive according to a general renewal process with rate λ. Formally, the arrival time of the first customer is given by u_1/λ, and the time between the $(\ell-1)$st and ℓth customer arrivals for $\ell \geq 2$ is given by u_ℓ/λ, where $\{u_\ell : \ell \geq 1\}$ is a sequence of independently and identically distributed (iid) random variables with unit-mean. The number of the customer arrivals by time t, $A(t)$, is given by

$$A(t) = \max\left\{k : \frac{u_1}{\lambda} + \cdots + \frac{u_k}{\lambda} \leq t\right\}. \tag{3.1}$$

Upon arrival, each customer will receive a numbered ticket with the ticket number running in an increasing order to proceed to its service. The system has m servers and the number of operating servers, denoted by i, can be adjusted to any number in $\{1, \cdots, m\}$ immediately after an arrival or a service completion. Right after receiving its numbered ticket, the ticketed customer will immediately receives service if there is one idle server among the i operating servers. Otherwise, the ticketed customer has to wait to be called to receive service. The waiting customers are called to get service according to increasing order of their ticket numbers. A customer may abandon his ticket before his number is called for service (no show). If a customer shows up when his ticket number is called, the customer will immediately receive service from an available server among the operating servers. If the customer is a no-show, his number will be discarded and the next ticket number will be called. We use α_i to represent the no-show probability of a ticket when i $(i = 1, \cdots, m)$ servers are in operation. That is, whenever one of the i operating servers is free to serve customers, she calls the next ticket number and that number has a probability of α_i to be associated with a no-show customer. The customer service times are assumed to be iid random variables with rate μ. Namely, the first customer service time is s_1/μ, and the ℓth customer's service time is s_ℓ/μ, where $\{s_\ell : \ell \geq 1\}$ is a sequence of iid random variables with unit-mean.

Similar to two-level staffing policy case, we consider four cost components: (i) the abandonment cost: each no-show customer will incur cost r; (ii) the nonabandoned customer waiting cost: each delayed customer who will not abandon the system will incur cost h per unit time (iii) the server setup cost: each server setup will cost K (that is, K is applied to each server whenever one is *added* into service, but there is no cost to remove a server); and (iv) the server operating cost: i operating servers cost c_i per unit time.

Our question is how to use ticket information to dynamically determine the staffing level of the ticket queue that minimizes the system long-run average cost. To characterize ticket information, let $S(t)$ be the number of operating servers at time t, and let $Q(t)$ be the number of tickets in the system at time t, including the customers, if any, who are currently receiving service; that is, $Q(t)$ is the sum of the number of busy servers at time t, $S(t)$, and the difference between the number of the last issued ticket before time t and the maximum of the ticket numbers under service at time t. Then the number of uncalled tickets in queue at time t is $(Q(t) - S(t))^+$.

We first look at the system cost by time t. Let $c_{S(t)}$ be the operating cost incurred when staffing level is $S(t)$, that is, the unit operating cost incurred by $S(t)$ servers at time t. Following the way how to charge the server setup and operating costs, we have the cumulative operating cost up to time t,

$$\mathsf{E} \int_0^t c_{S(x)} \mathsf{d}x := \mathcal{O}(t), \tag{3.2}$$

and the cumulative setup cost up to time t

$$K \cdot \mathsf{E} \int_0^t \mathsf{I}_{\{S(x) > S(x-)\}} \mathsf{d}S(x) := \mathcal{S}(t). \tag{3.3}$$

We say the system to be in *i-server region* if there are exact i operating servers among m servers. Let $T_{ij}(t)$ be the total amount of time that server j is processing the customer service requirement when the system is in i-server region during $[0, t]$. It is straightforward to see that $\sum_{i=1}^m T_{ij}(t)$ is the total amount of time that server j is busy during $[0, t]$. Recall that adding one operating server is triggered by a customer arrival, and shutting down one operating server is triggered by a customer service completion from it.

Let $v(t)$ be the virtual waiting time, which is the amount of time a hypothetical customer would have to wait before its numbered ticket to be called upon arriving at time t. Hence, with $\tau_\ell = (1/\lambda) \sum_{\ell'=1}^{\ell} u_{\ell'}$, $v(\tau_\ell -)$ $(:= v_\ell)$ is the time that the ℓth arriving customer has to wait before its ticket gets a call. In order to describe the costs of customer abandonments and customer delay, we introduce m independent sequences of i.i.d binary random variables $\{z_{i\ell} : \ell \geq 1\}$ $(i = 1, \cdots, m)$ with

$$\mathsf{Pr}(z_{i\ell} = 0) = \alpha_i \text{ and } \mathsf{Pr}(z_{i\ell} = 1) = 1 - \alpha_i.$$

Suppose that the system makes the ℓth call for a ticket number among the waiting customers, and the system is being operated under i-server region. Then the called ticket will abandon if $z_{i\ell} = 0$ and show up if $z_{i\ell} = 1$. Thus, the total number of abandonments incurred by the customers who have arrived in the system by time t can be written as

$$\sum_{\ell=1}^{A(t)} \sum_{i=1}^{m} \Big(1 - z_{i\ell}\Big) \times \mathsf{I}_{\{S(\tau_\ell + v_\ell) = i, Q(\tau_\ell + v_\ell) > i\}} := \mathcal{R}(t). \tag{3.4}$$

Let $\{B_j(t) : t \geq 0\}$ $(j = 1, \cdots, m)$ be m independent and identical renewal processes with the same distribution of $\{B(t) : t \geq 0\}$ given by

$$B(t) = \max\Big\{\ell : \frac{s_1}{\mu} + \cdots + \frac{s_\ell}{\mu} \leq t\Big\}. \tag{3.5}$$

Then

$$\mathcal{D}_j(t) = B_j\Big(\sum_{i=1}^{m} T_{ij}(t)\Big) \tag{3.6}$$

is the number of customers who have departed from server j after receiving their service by time t. Let $\tau(t)$ be the arrival time of the customer who is the last one to start receiving service among the customers currently in service if $(Q(t) - S(t))^+ > 0$, and to be t if $(Q(t) - S(t))^+ = 0$. In view of

(3.4), the customer abandonment cost by time t is

$$r \cdot \mathsf{E}\mathcal{R}(\tau(t)), \tag{3.7}$$

and the customer delay cost by time t is

$$h \sum_{i=1}^{m} (1-\alpha_i) \mathsf{E} \int_0^t \left[A(x) - \mathcal{R}(\tau(x)) - \sum_{j=1}^{m} \mathcal{D}_j(x) - i \right]^+ \cdot \mathsf{I}_{\{S(x)=i\}} \mathsf{d}x := \mathcal{H}(t). \tag{3.8}$$

The system dynamics are given by

$$Q(t) = A(t) - \mathcal{R}(\tau(t)) - \sum_{j=1}^{m} \mathcal{D}_j(t). \tag{3.9}$$

Note that in (3.4), $\mathcal{R}(t)$ is the cumulative number of the abandonments counted from all the customers who have arrived in the system by time t. Among them, some of their tickets have been called out, and some have not been called out yet by time t. In contrast to (3.4), $\mathcal{R}(\tau(t))$ in (3.9) is the cumulative number of the abandonments counted from all the customers who have arrived in the system, and have been also called out by time t. Hence,

$$\mathcal{R}(t) - \mathcal{R}(\tau(t))$$

is the number of the abandonments from the customers who have already arrived in the system but their ticker numbers have not been called out yet by time t.

Based on the information only given by $Q(t)$, our objective is to dynamically determine $S(t)$ at any time t to minimize

$$\frac{r \times \mathsf{E}\mathcal{R}(\tau(T)) + \mathcal{H}(T) + \mathcal{O}(T) + \mathcal{S}(T)}{T} \tag{3.10}$$

over the time interval $[0,\ T]$ with large enough T. To avoid the trivial case, we assume there exists a m_0 with $1 < m_0 \leq m$ such that

$$\frac{(1-\alpha_{m_0+1})\lambda}{(m_0+1)\mu} < 1. \tag{3.11}$$

That is, the overall arrival traffic (after balking) can only be handled by $m_0 + 1$ or more servers working simultaneously.

Without loss of generality, after making a cost normalization, we assume the cost per customer abandonment is one, i.e., $r = 1$ in the remainder of the paper. The methodology that we use to study the above problem is fluid approximation. We consider a sequence of systems similar the one described above. For the nth system, the customer arrival rate is $n\lambda$, and the service rate is $n\mu$. Because in the fluid limit (letting n go to infinite), the jumps incurred by customer arrivals or service completion become negligible, this simple feature makes the above problem analytically tractable.

3.2 Fluid Approximation

This section describes the fluid approximation of our problem. Consider a sequence of systems as described in the previous section, indexed by $n \geq 1$. For the nth system, the arrival time of the first customer is given u_1/λ^n, and the time between the $(\ell - 1)$st and ℓth customer arrivals for $\ell \geq 2$ is given by u_ℓ/λ^n. The number of customers that arrived during $[0, t]$ is given by $\{A^n(t) : t \geq 0\}$ with

$$A^n(t) = \max\left\{k : \frac{u_1}{\lambda^n} + \cdots + \frac{u_k}{\lambda^n} \leq t\right\}.$$

The sequence of customer service times is given by $\{s_\ell/\mu^n : \ell \geq 1\}$ accordingly. Here the sequences of arrival rates $\{\lambda^n : n \geq 1\}$ and service rates $\{\mu^n : n \geq 1\}$ satisfy

$$\lim_{n\to\infty} \frac{\lambda^n}{n} = \lambda \text{ and } \lim_{n\to\infty} \frac{\mu^n}{n} = \mu \text{ with } \lambda \text{ and } \mu \text{ satisfying (3.11)} \qquad (3.12)$$

All the other processes associated with the nth network are appended with a superscript n. In order to make the problem analytically tractable, we impose convergence assumption on the arrival process, namely, with probability one, the following limit holds uniformly on compact sets of $[0, \infty)$:

$$\frac{A^n(t) - \lambda^n t}{n} \to 0 \text{ and } \frac{B^n(t) - \mu^n t}{n} \to 0 \text{ as } n \to \infty. \qquad (3.13)$$

Furthermore, we assume the independence between the customer abandonments and the customer arrivals and service times. Namely,

$$\{z_{i\ell} : \ell \geq 1\} \text{ is independent of } \{u_\ell : \ell \geq 1\} \text{ and } \{s_\ell : \ell \geq 1\}. \tag{3.14}$$

It follows from (3.9) that

$$\begin{aligned} Q^n(t) = & [A^n(t) - \lambda^n t] - \sum_{\ell=1}^{A^n(\tau^n(t))} \sum_{i=1}^{m} [(1 - z_{i\ell}) - \alpha_i] \times \mathsf{I}_{\{S^n(\tau_\ell^n + v_\ell^n) = i, Q^n(\tau_\ell^n + v_\ell^n) > i\}} \\ & - \sum_{j=1}^{m} \Big[B_j^n \Big(\sum_{i=1}^{m} T_{ij}^n(t) \Big) - \mu^n \sum_{i=1}^{m} T_{ij}^n(t) \Big] \\ & + \lambda^n t - \sum_{\ell=1}^{A^n(\tau^n(t))} \sum_{i=1}^{m} \alpha_i \times \mathsf{I}_{\{S^n(\tau_\ell^n + v_\ell^n) = i, Q^n(\tau_\ell^n + v_\ell^n) > i\}} - \mu^n \sum_{j=1}^{m} \sum_{i=1}^{m} T_{ij}^n(t). \end{aligned} \tag{3.15}$$

From the definition of $\tau^n(\cdot)$, we also have

$$\Big(Q^n(t) - S^n(t) \Big)^+ = A^n(t) - A^n(\tau^n(t)). \tag{3.16}$$

In view of the work-conserving property, we have that for each $i \in \{1, \cdots, m\}$,

$$\int_0^t \mathsf{I}_{\{S^n(x) = i, Q^n(x) \geq i\}} \mathsf{d} \Big(ix - \sum_{j=1}^{m} T_{ij}^n(x) \Big) = 0. \tag{3.17}$$

By (3.15)-(3.16), we get the fluid-scaled processes,

$$\begin{aligned}\frac{Q^n(t)}{n} =& \frac{A^n(t)-\lambda^n t}{n} - \frac{1}{n}\sum_{j=1}^{m}\Big[B_j^n\Big(\sum_{i=1}^{m}T_{ij}^n(t)\Big) - \mu^n\sum_{i=1}^{m}T_{ij}^n(t)\Big] \\ & - \frac{1}{n}\sum_{\ell=1}^{A^n(\tau^n(t))}\sum_{i=1}^{m}\Big[\Big(1-z_{i\ell}\Big)-\alpha_i\Big] \times \mathrm{I}_{\{S^n(\tau_\ell^n+v_\ell^n)=i, Q^n(\tau_\ell^n+v_\ell^n)>i\}} \\ & + \frac{\lambda^n}{n}t - \frac{\mu^n}{n}\sum_{j=1}^{m}\sum_{i=1}^{m}T_{ij}^n(t) \\ & - \frac{1}{n}\sum_{\ell=1}^{A^n(\tau^n(t))}\sum_{i=1}^{m}\alpha_i \times \mathrm{I}_{\{S^n(\tau_\ell^n+v_\ell^n)=i, Q^n(\tau_\ell^n+v_\ell^n)>i\}},\end{aligned} \tag{3.18}$$

and

$$\frac{\Big(Q^n(t)-S^n(t)\Big)^+}{n} = \frac{A^n(t)-\lambda^n t}{n} - \frac{A^n(\tau^n(t))-\lambda^n\tau^n(t)}{n} + \frac{\lambda^n}{n}(t-\tau^n(t)). \tag{3.19}$$

The equicontinuous property of $\{T_{ij}^n(\cdot) : n \geq 1\}$ follows from the fact that

$$0 < \sum_{i=1}^{m}\Big(T_{ij}^n(t) - T_{ij}^n(s)\Big) < (t-s) \text{ for all } t > s > 0,\ j = 1, \cdots, m, \text{ and } n \geq 1. \tag{3.20}$$

In order to get the fluid approximation, we first establish the following lemma.

Lemma 11. *Suppose that* (3.13)-(3.14) *hold. With probability one, for any subsequence of* $\{\tau^n : n \geq 1\}$ *with* $\tau^n = \{\tau^n(t) : t \geq 0\}$*, there exists a further* *subsequence* $\{\tau^{n_\ell} : \ell \geq 1\}$ *with* $n_\ell \to \infty$ *as* $\ell \to \infty$ *such that as* $\ell \to \infty$,

$$\tau^{n_\ell} \to \bar{\tau} \quad u.o.c.,$$

where $\bar{\tau} = \{\bar{\tau}(t) : t \geq 0\}$ *is Lipschitz continuous on* $[0, \infty)$.

Proof. First note that, by the definition of τ^n ($\tau^n(t)$ is the arrival time of the customer who is the last one to receive service in the nth system among the customers currently in service if $Q^n(t) > 0$, and to be t if $Q^n(t) = 0$), τ^n is nondecreasing and $0 \leq \tau^n(t) \leq t$ for all $t \geq 0$ and all $n \geq 1$. This observation gives that for any subsequence of $\{\tau^n : n \geq 1\}$ there exist a subsequence $\{\tau^{n_\ell} : \ell \geq 1\}$ and a nondecreasing function $\bar{\tau}$ defined on all rational numbers in $[0, \infty)$ with $0 \leq \bar{\tau}(t) \leq t$ such that

$$\tau^{n_\ell}(t) \to \bar{\tau}(t) \text{ as } \ell \to \infty \text{ for all rational numbers } t \geq 0. \tag{3.21}$$

Since $\bar{\tau}$ is a nondecreasing function on all rational numbers on $[0, \infty)$, it can be extended to all real numbers on $[0, \infty)$ in an obvious way: for any irrational real number $t > 0$, find a decreasing sequence of rational numbers t_ℓ such that $t_\ell \to t$ as $\ell \to \infty$ and then define $\bar{\tau}(t)$ to be the limit of $\bar{\tau}(t_\ell)$ as $\ell \to \infty$. If we can show that the process $\bar{\tau} = \{\bar{\tau}(t) : t \geq 0\}$ is Lipschitz continuous, then by a result in Resnick (2007) (which states that if a sequence of nondecreasing functions on $[0,\infty)$ converges to a continuous function on $[0,\infty)$ for all rational numbers, then the convergence is u.o.c.), we complete the proof.

Now we show that the Lipschitz continuity of $\bar{\tau}$. To the end, it suffices to show that there exists a constant C such that for any rational numbers $s,t\in[0,\infty)$ with $s\le t$,

$$\limsup_{n\to\infty}\Big(\tau^n(t)-\tau^n(s)\Big)\le C\times(t-s). \tag{3.22}$$

According to the definition of $\{T_{ij}(t) : t \ge 0\}$, for the nth system, the cumulative number of the customer service completion during time interval $(s,t]$ is given by

$$\sum_{j=1}^{m}\Big(B_j^n(\sum_{i=1}^{m}T_{ij}^n(t))-B_j^n(\sum_{i=1}^{m}T_{ij}^n(s))\Big). \tag{3.23}$$

By again the definition of $\tau^n(t)$, the service requirements of the customers who have arrived during time interval $[\tau^n(s),\tau^n(t)]$ but not abandoned either have been completed or have not been completed but started during time interval $(s,t]$. Note that the number of the customers who have arrived during time interval $(\tau^n(s), \tau^n(t)]$ but not abandoned is given by

$$\Big(A^n(\tau^n(t)) - A^n(\tau^n(s))\Big) - \sum_{\ell=A^n(\tau^n(s))+1}^{A^n(\tau^n(t))} \sum_{i=1}^{m} \Big(1 - z_{i\ell}\Big) \times \mathsf{I}_{\{S^n(\tau_\ell^n+v_\ell^n)=i, Q^n(\tau_\ell^n+v_\ell^n)>i\}}. \tag{3.24}$$

Among them, there are at most m customers who have started their service but haven't finished during time interval $(s, t]$, since at most m servers are in operation. Hence, we have

$$\begin{aligned}\sum_{j=1}^{m} \Big(\mathcal{D}_j(t) - \mathcal{D}_j(s)\Big) \geq & \Big(A^n(\tau^n(t)) - A^n(\tau^n(s))\Big) - m \\ & - \sum_{\ell=A^n(\tau^n(s))+1}^{A^n(\tau^n(t))} \sum_{i=1}^{m} \Big(1 - z_{i\ell}\Big) \times \mathsf{I}_{\{S^n(\tau_\ell^n+v_\ell^n)=i, Q^n(\tau_\ell^n+v_\ell^n)>i\}}.\end{aligned} \tag{3.25}$$

It follows from (3.20) that the sequences $\{T_{ij}^n : n \geq 1\}$ given by $T_{ij}^n = \{T_{ij}^n(t) : t \geq 0\}$ $(i, j = 1, \cdots, m)$ are equicontinuous. Therefore, by the Asoli-Arzela theorem (Royden 1988), any subsequences of $\{T_{ij}^n : n \geq 1\}$ have further convergent subsequences $\{T_{ij}^{n_\ell} : \ell \geq 1\}$ such that for $i, j = 1, \cdots, m$,

$$T_{ij}^{n_\ell} \to \bar{T}_{ij} \;\; \text{u.o.c. as } \ell \to \infty \tag{3.26}$$

with $\bar{T}_{ij} = \{\bar{T}_{ij}(t) : t \geq 0\}$ $(i, j = 1, \cdots, m)$ being increasing and Lipschitz continuous functions satisfying $0 < \sum_{i=1}^{n}\Big(\bar{T}_{ij}(t) - \bar{T}_{ij}(s)\Big) < (t-s)$ for all $s,t \in [0,\infty)$ with $s < t$. Using (3.12)-(3.13) and (3.26), we have that

$$\begin{aligned}
&\lim_{\ell\to\infty}\frac{1}{n_\ell}\sum_{j=1}^{m}\Big(B_j^{n_\ell}(\sum_{i=1}^{m}T_{ij}^{n_\ell}(t)) - B_j^{n_\ell}(\sum_{i=1}^{m}T_{ij}^{n_\ell}(s))\Big)\\
&\quad= \lim_{\ell\to\infty}\frac{1}{n_\ell}\sum_{j=1}^{m}\Big[\Big(B_j^{n_\ell}(\sum_{i=1}^{m}T_{ij}^{n_\ell}(t)) - \mu^{n_\ell}\sum_{i=1}^{m}T_{ij}^{n_\ell}(t)\Big)\\
&\qquad - \Big(B_j^{n_\ell}(\sum_{i=1}^{m}T_{ij}^{n_\ell}(s)) - \mu^{n_\ell}\sum_{i=1}^{m}T_{ij}^{n_\ell}(s)\Big)\Big] + \frac{\mu^{n_\ell}}{n_\ell}\sum_{j=1}^{m}\sum_{i=1}^{m}\Big(T_{ij}^{n_\ell}(t) - T_{ij}^{n_\ell}(s)\Big)\\
&\quad= \mu\sum_{j=1}^{m}\sum_{i=1}^{m}\Big(\bar{T}_{ij}(t) - \bar{T}_{ij}(s)\Big)\\
&\quad\le m\mu(t-s). \qquad (3.27)
\end{aligned}$$

Note that

$$\begin{aligned}
&\Big(A^n(\tau^n(t)) - A^n(\tau^n(s))\Big) - m\\
&\qquad - \sum_{\ell=A^n(\tau^n(s))+1}^{A^n(\tau^n(t))}\sum_{i=1}^{m}\Big(1 - z_{i\ell}\Big)\times \mathbf{1}_{\{S^n(\tau_\ell^n+v_\ell^n)=i,Q^n(\tau_\ell^n+v_\ell^n)>i\}}\\
&\quad= \Big(A^n(\tau^n(t)) - A^n(\tau^n(s))\Big) - m\\
&\qquad - \sum_{\ell=A^n(\tau^n(s))+1}^{A^n(\tau^n(t))}\sum_{i=1}^{m}\alpha_i\mathbf{1}_{\{S^n(\tau_\ell^n+v_\ell^n)=i,Q^n(\tau_\ell^n+v_\ell^n)>i\}}\\
&\qquad + \sum_{\ell=A^n(\tau^n(s))+1}^{A^n(\tau^n(t))}\sum_{i=1}^{m}(z_{i\ell} - (1-\alpha_i))\mathbf{1}_{\{S^n(\tau_\ell^n+v_\ell^n)=i,Q^n(\tau_\ell^n+v_\ell^n)>i\}}\\
&\quad\ge \Big(A^n(\tau^n(t)) - A^n(\tau^n(s))\Big) - m - \max_i \alpha_i\Big(A^n(\tau^n(t)) - A^n(\tau^n(s))\Big)
\end{aligned}$$

$$+ \sum_{\ell=A^n(\tau^n(s))+1}^{A^n(\tau^n(t))} \sum_{i=1}^{m} (z_{i\ell} - (1-\alpha_i)) \mathsf{I}_{\{S^n(\tau_\ell^n + v_\ell^n)=i, Q^n(\tau_\ell^n+v_\ell^n)>i\}}$$

$$= \Big(A^n(\tau^n(t)) - A^n(\tau^n(s))\Big)(1 - \max_i \alpha_i) - m$$
$$+ \sum_{\ell=A^n(\tau^n(s))+1}^{A^n(\tau^n(t))} \sum_{i=1}^{m} (z_{i\ell} - (1-\alpha_i)) \mathsf{I}_{\{S^n(\tau_\ell^n + v_\ell^n)=i, Q^n(\tau_\ell^n+v_\ell^n)>i\}}. \tag{3.28}$$

As $\{z_{i\ell} : \ell \geq 1\}$ $(i = 1, \cdots, m)$ are sequences of iid binary random variables, we have that for $i = 1, \cdots, m$,

$$Z_i^n \to 0 \text{ u.o.c. as } n \to \infty, \tag{3.29}$$

where $Z_i^n = \{Z_i^n(t) : t \geq 0\}$ with $Z_i^n(t) = \frac{1}{n}\sum_{\ell=1}^{\lfloor nt \rfloor}\Big(z_{i\ell} - (1-\alpha_i)\Big)$. By the random-time change theorem (see Billingsley, 2009)and (3.13), we have

$$\tilde{Z}_i^n \to 0 \text{ u.o.c. as } n \to \infty, \tag{3.30}$$

where $\tilde{Z}_i^n = \{\tilde{Z}_i^n(t) : t \geq 0\}$ with $\tilde{Z}_i^n(t) = \frac{1}{n}\sum_{\ell=1}^{A^n(t)}\Big(z_{i\ell} - (1-\alpha_i)\Big)$. In view of $\tau^n(t) \leq t$ and (3.30), we have that with probability one, for any $s, t \in [0, \infty)$ with $s \leq t$,

$$\sum_{\ell=A^n(\tau^n(s))+1}^{A^n(\tau^n(t))} \sum_{i=1}^{m} \Big(z_{i\ell} - (1-\alpha_i)\Big) \to 0 \text{ as } n \to \infty. \tag{3.31}$$

Using (3.6), (3.25), (3.13), and (3.21), there exists a subsequence $\{\tau^{n'_\ell} : \ell \geq$

$1\}$ of $\{\tau^{n_\ell} : \ell \geq 1\}$ given by (3.26) such that for any rational numbers $s, t \in [0, \infty)$ with $s \leq t$,

$$\begin{aligned}
&\frac{1}{n'_\ell}\Big(A^{n'_\ell}(\tau^{n'_\ell}(t)) - A^{n'_\ell}(\tau^{n'_\ell}(s))\Big) \\
&= \frac{1}{n'_\ell}\Big[\Big((A^{n'_\ell}(\tau^{n'_\ell}(t)) - \lambda^{n'_\ell}\tau^{n'_\ell}(t)\Big) - \Big(A^{n'_\ell}(\tau^{n'_\ell}(s)) - \lambda^{n'_\ell}\tau^{n'_\ell}(s)\Big)\Big] \\
&\quad + \frac{\lambda^{n'_\ell}}{n'_\ell}\Big(\tau^{n'_\ell}(t) - \tau^{n'_\ell}(s)\Big) \\
&\to \lambda\Big(\bar{\tau}(t) - \bar{\tau}(s)\Big) \quad \text{as } \ell \to \infty. \qquad (3.32)
\end{aligned}$$

Combining (3.27)-(3.28) and(3.31)-(3.32) yields that for any rational numbers $s, t \in [0, \infty)$ with $s \leq t$,

$$\bar{\tau}(t) - \bar{\tau}(s) \leq \frac{m\mu}{\lambda(1 - \max_i \alpha_i)}(t - s), \qquad (3.33)$$

which implies that (3.22). Therefore, the proof of the lemma is completed.

□

For $i = 1, \cdots, m$, define

$$R_i^n(t) = \frac{1}{n}\sum_{\ell=1}^{A^n(\tau^n(t))}\Big[(1 - z_{i\ell}) - \alpha_i\Big] \times \mathsf{I}_{\{S^n(\tau_\ell^n + v_\ell^n) = i, Q^n(\tau_\ell^n + v_\ell^n) > i\}}, \qquad (3.34)$$

and $R_i^n = \{R_i^n(t) : t \geq 0\}$.

Lemma 12. *Suppose that* (3.12)-(3.14) *hold. With probability one, as* $n \to$

∞, *for* $i = 1, \cdots, m$,

$$R_i^n \to 0 \quad u.o.c.$$

Proof. For each $i = 1, \cdots, m$, consider the sequence $\{X_i^n : n \geq 1\}$ given by

$$X_i^n = \sum_{\ell=1}^{n} \Big[(1 - z_{i\ell}) - \alpha_i\Big] \times \mathsf{I}_{\{S^n(\tau_\ell^n + v_\ell^n) = i, Q^n(\tau_\ell^n + v_\ell^n) > i\}}.$$

Let $\mathcal{F}_\ell^n$ be the σ-field generated by

$$\{(z_{1k}, \cdots, z_{m,k}), u_k, s_k : 1 \leq k \leq \ell - 1\}\ ,\ \{u_k : \ell \leq k \leq A^n(\tau_\ell^n + v_\ell^n) + 1\},$$
$$\text{and } \{S(t) : t \in [0, \tau_\ell^n)\}.$$

Then we know that for $\ell < \ell'$, $\Big[(1 - z_{i\ell}) - \alpha_i\Big] \times \mathsf{I}_{\{S^n(\tau_\ell^n + v_\ell^n) = i, Q^n(\tau_\ell^n + v_\ell^n) > i\}}$ is measurable with respect to $\mathcal{F}_{\ell'}^n$, and $\mathsf{I}_{\{S^n(\tau_\ell^n + v_\ell^n) = i, Q^n(\tau_\ell^n + v_\ell^n) > i\}}$ is measurable with respect to $\mathcal{F}_\ell^n$. Hence, by (3.14), we have that for $\ell < \ell'$,

$$\begin{aligned}
&\mathsf{E}\bigg(\Big[(1 - z_{i\ell}) - \alpha_i\Big] \times \mathsf{I}_{\{S^n(\tau_\ell^n + v_\ell^n) = i, Q^n(\tau_\ell^n + v_\ell^n) > i\}} \\
&\quad \times \Big[(1 - z_{i\ell'}) - \alpha_i\Big] \times \mathsf{I}_{\{S^n(\tau_{\ell'}^n + v_{\ell'}^n) = i, Q^n(\tau_{\ell'}^n + v_{\ell'}^n) > i\}}\bigg) \\
&\quad = \mathsf{E}\bigg[\mathsf{E}\bigg\{\bigg(\Big[(1 - z_{i\ell}) - \alpha_i\Big] \times \mathsf{I}_{\{S^n(\tau_\ell^n + v_\ell^n) = i, Q^n(\tau_\ell^n + v_\ell^n) > i\}} \\
&\qquad \times \Big[(1 - z_{i\ell'}) - \alpha_i\Big] \times \mathsf{I}_{\{S^n(\tau_{\ell'}^n + v_{\ell'}^n) = i, Q^n(\tau_{\ell'}^n + v_{\ell'}^n) > i\}}\bigg)\bigg|\mathcal{F}_{\ell'}^n\bigg\}\bigg] \\
&\quad = \mathsf{E}\bigg[\bigg(\Big[(1 - z_{i\ell}) - \alpha_i\Big] \times \mathsf{I}_{\{S^n(\tau_\ell^n + v_\ell^n) = i, Q^n(\tau_\ell^n + v_\ell^n) > i\}}
\end{aligned}$$

$$\times \mathsf{I}_{\{S^n(\tau^n_{\ell'}+v^n_{\ell'})=i,Q^n(\tau^n_{\ell'}+v^n_{\ell'})>i\}}\Big) \times \mathsf{E}\Big\{\Big[(1-z_{i\ell'})-\alpha_i\Big]\Big|\mathcal{F}^n_{\ell'}\Big\}\Big]$$

$$= \mathsf{E}\Big[\Big(\Big[(1-z_{i\ell})-\alpha_i\Big] \times \mathsf{I}_{\{S^n(\tau^n_{\ell}+v^n_{\ell})=i,Q^n(\tau^n_{\ell}+v^n_{\ell})>i\}}$$

$$\times \mathsf{I}_{\{S^n(\tau^n_{\ell'}+v^n_{\ell'})=i,Q^n(\tau^n_{\ell'}+v^n_{\ell'})>i\}}\Big) \times 0\Big]$$

$$= 0.$$

Thus, we have $\mathsf{E}\Big(X^n_i\Big)^2 \le n(1-\alpha_i)\alpha_i$. This, in turn, implies

$$\frac{1}{n}X^n_i \text{ converges to zero in probability.} \tag{3.35}$$

Define $Y^n_i = \{Y^n_i(t) : t \ge 0\}$ with

$$Y^n_i(t) = \frac{1}{n}\sum_{\ell=1}^{\lfloor nt\rfloor}\Big[(1-z_{i\ell})-\alpha_i\Big] \times \mathsf{I}_{\{S^n(\tau^n_{\ell}+v^n_{\ell})=i,Q^n(\tau^n_{\ell}+v^n_{\ell})>i\}}.$$

Consequently, by (3.35) and the Skorohod representation theorem, with probability one,

$$Y^n_i \to 0 \quad u.o.c. \text{ as } n \to \infty. \tag{3.36}$$

It follows from Lemma 11 and the random-time change theorem (see Billingsley, 2009) that with probability one,

$$R_i^n \to 0 \quad \textit{u.o.c.} \text{ as } n \to \infty. \tag{3.37}$$

Therefore, we have the lemma. □

Let $B^n = \{B^n(t) : t \geq 0\}$ with

$$B^n(t) = \frac{1}{n} \sum_{j=1}^{m} \Big[B_j^n \Big(\sum_{i=1}^{m} T_{ij}^n(t) \Big) - \mu^n \sum_{i=1}^{m} T_{ij}^n(t) \Big]. \tag{3.38}$$

Lemma 13. *Suppose that* (3.13)-(3.14) *hold. With probability one, as* $n \to \infty$,

$$B^n \to 0 \quad \textit{u.o.c.}$$

Proof. By (3.13), first we have that for any constant $C > 0$, with probability one,

$$B_C^n \to 0 \quad \text{u.o.c. as } n \to \infty,$$

where $B_C^n = \{B_C^n(t) : t \geq 0\}$ with $B_C^n(t) = (1/n) \sum_{j=1}^{m} \Big[B_j^n(Ct) - \mu^n Ct \Big]$. Then the lemmas directly follows from the fact that for $j = 1, \cdots, m$, and $t \in [0, \infty)$,

$$\sum_{i=1}^{m} T_{ij}^n(t) \leq t.$$

□

Define

$$L^n(t) = \frac{1}{n} \sum_{\ell=1}^{A^n(\tau^n(t))} \sum_{i=1}^{m} \alpha_i \times \mathsf{I}_{\{S^n(\tau_\ell^n + v_\ell^n) = i, Q^n(\tau_\ell^n + v_\ell^n) > i\}}. \tag{3.39}$$

Notice that, by the definition of z_{il}, (3.34), and (3.39), $n(L^n(t) + \sum_{i=1}^m R_i^n(t))$ is the number of customer abandonments between time 0 and $\tau^n(t)$ in the nth system.

Lemma 14. *Suppose that* (3.13)-(3.14) *hold. With probability one, for any subsequence of* $\{L^n : n \geq 1\}$ *with* $L^n = \{L^n(t) : t \geq 0\}$*, there exists a further subsequence* $\{L^{n_\ell} : \ell \geq 1\}$ *with* $n_\ell \to \infty$ *as* $\ell \to \infty$ *such that as* $\ell \to \infty$*,*

$$L^{n_\ell} \to \bar{L} \quad u.o.c.,$$

where $\bar{L} = \{\bar{L}(t) : t \geq 0\}$ *is Lipschitz continuous on* $[0, \infty)$*.*

Proof. Note that for all $t \geq 0$ and all $n \geq 1$,

$$\begin{aligned} L^n(t) &= \frac{1}{n} \sum_{\ell=1}^{A^n(\tau^n(t))} \sum_{i=1}^{m} \alpha_i \times \mathsf{I}_{\{S^n(\tau_\ell^n + v_\ell^n) = i, Q^n(\tau_\ell^n + v_\ell^n) > i\}} \\ &= \frac{1}{n} \int_0^{\tau^n(t)} \sum_{i=1}^{m} \alpha_i \times \mathsf{I}_{\{S^n(x + v^n(x)) = i, Q^n(x + v^n(x)) > i\}} \mathsf{d}A^n(x) \\ &\leq \frac{1}{n} \int_0^{\tau^n(t)} \max_i \alpha_i \mathsf{d}A^n(x) \\ &\leq \frac{A^n(\tau^n(t))}{n} \max_i \alpha_i \\ &\leq \frac{A^n(t)}{n} \max_i \alpha_i. \end{aligned} \tag{3.40}$$

As the integrand is nonnegative in $L^n(t)$, we know that L^n is nondecreasing. Hence for any subsequence of $\{L^n : n \geq 1\}$ there exist a subsequence $\{L^{n_\ell} : \ell \geq 1\}$ and a nondecreasing function $\bar{L}$ defined on all rational numbers in $[0,\infty)$ such that

$$L^{n_\ell}(t) \to \bar{L}(t) \quad \text{as } \ell \to \infty \text{ for all rational numbers } t \geq 0. \tag{3.41}$$

Similar to the proof of Lemma 11, we extend the domain (nonnegative rational numbers) of $\bar{L}$ to $[0,\infty)$. To prove the lemma, it suffices to show that there exists a constant $C > 0$ such that for all rational numbers $s, t \in [0,\infty)$ with $s \leq t$,

$$\bar{L}(t) - \bar{L}(s) \leq C \times (t - s). \tag{3.42}$$

To the end, by (3.40),

$$\begin{aligned} L^n(t) - L^n(s) &= \frac{1}{n}\int_{\tau^n(s)}^{\tau^n(t)} \sum_{i=1}^{m} \alpha_i \times \mathsf{I}_{\{S^n(x+v^n(x))=i, Q^n(x+v^n(x))>i\}} \mathsf{d}A^n(x) \\ &\leq \frac{1}{n}\Big(A^n(\tau^n(t)) - A^n(\tau^n(s))\Big) \times \max_i \alpha_i \\ &= \max_i \alpha_i \times \Big[\frac{A^n(\tau^n(t)) - \lambda^n \tau^n(t)}{n} - \frac{A^n(\tau^n(s)) - \lambda^n \tau^n(s)}{n} \\ &\qquad + \frac{\lambda^n}{n}\Big(\tau^n(t) - \tau^n(s)\Big)\Big]. \end{aligned}$$

(3.42) directly follows from (3.13) and (3.32)-(3.33) with

$$C = \frac{m\mu}{1 - \max_i \alpha_i} \times \max_i \alpha_i.$$

□

Define

$$Q^n = \left\{ \frac{Q^n(t)}{n} : t \geq 0 \right\} \text{ and } T^n = \left\{ \left(T^n_{ij}(t), i, j = 1, \cdots, m \right) : t \geq 0 \right\}. \tag{3.43}$$

With the help of Lemmas 11-14, we get the following fluid approximation. Since we are interested in the long-run average cost, in theorem 15, proposition 16, and theorem 18, we consider the system starting from empty state.

Theorem 15. *Suppose that* (3.13)-(3.14) *hold. With probability one, for any subsequence of* $\{(\tau^n, L^n, T^n, Q^n) : n \geq 1\}$, *there exists a further subsequence* $\{(\tau^{n_\ell}, L^{n_\ell}, T^{n_\ell}, Q^{n_\ell}) : \ell \geq 1\}$ *with* $n_\ell \to \infty$ *as* $\ell \to \infty$ *such that as* $\ell \to \infty$,

$$\left(\tau^{n_\ell}, L^{n_\ell}, T^{n_\ell}, Q^{n_\ell} \right) \to \left(\bar{\tau}, \bar{L}, \bar{T}, \bar{Q} \right) \quad u.o.c., \tag{3.44}$$

where $\bar{\tau} = \{\bar{\tau}(t) : t \geq 0\}, \bar{L} = \{\bar{L}(t) : t \geq 0\}$, *and* $\bar{T} = \{(\bar{T}_{ij}(t), i, j = 1, \cdots, m) : t \geq 0\}$ *are increasing and Lipschitz continuous on* $[0, \infty)$, *and* $\bar{Q} = \{\bar{Q}(t) : t \geq 0\}$ *is Lipschitz continuous on* $[0, \infty)$. *At the same time, the* *above limit satisfies*

$$\bar{Q}(t) = \lambda t - \bar{L}(t) - \mu \sum_{i=1}^{m} \sum_{j=1}^{m} \bar{T}_{ij}(t) = \lambda(t - \bar{\tau}(t)) \geq 0, \tag{3.45}$$

$$0 \leq \sum_{i=1}^{m} \left(\bar{T}_{ij}(t) - \bar{T}_{ij}(s) \right) \leq (t-s) \quad \textit{for all } t > s > 0 \textit{ and } j = 1, \cdots, m. \tag{3.46}$$

Proof. The convergence given by (3.44) and (3.45)-(3.46) directly follow from Lemmas 11-14 and (3.18)-(3.20). □

Similarly, using $Q^n(t)$ and $S^n(t)$, we can also write down the corresponding cost function $\mathcal{O}^n(t)$, $\mathcal{S}^n(t)$, $\mathcal{R}^n(\tau^n(t))$, and $\mathcal{H}^n(t)$ in nth system. Define

$$\mathcal{O}^n = \left\{ \frac{\mathcal{O}^n(t)}{n} : t \geq 0 \right\}, \quad \mathcal{S}^n = \left\{ \frac{\mathcal{S}^n(t)}{n} : t \geq 0 \right\}, \tag{3.47}$$

$$\mathcal{R}^n = \left\{ \frac{\mathcal{R}^n(\tau^n(t))}{n} : t \geq 0 \right\}, \quad \mathcal{H}^n = \left\{ \frac{\mathcal{H}^n(t)}{n} : t \geq 0 \right\}. \tag{3.48}$$

Proposition 16. *Suppose that* (3.13)-(3.14) *hold. With probability one, for any subsequence of* $\{(\mathcal{O}^n, \mathcal{S}^n, \mathcal{R}^n, \mathcal{H}^n) : n \geq 1\}$*, there exists a further subsequence* $\{(\mathcal{O}^{n_\ell}, \mathcal{S}^{n_\ell}, \mathcal{R}^{n_\ell}, \mathcal{H}^{n_\ell}) : \ell \geq 1\}$ *with* $n_\ell \to \infty$ *as* $\ell \to \infty$ *such that as* $\ell \to \infty$*,*

$$\left(\mathcal{O}^{n_\ell}, \mathcal{S}^{n_\ell}, \mathcal{R}^{n_\ell}, \mathcal{H}^{n_\ell} \right) \to \left(\bar{\mathcal{O}}, \bar{\mathcal{S}}, \bar{\mathcal{R}}, \bar{\mathcal{H}} \right) \quad \textit{u.o.c.}, \tag{3.49}$$

where $\bar{\mathcal{O}} = \{\bar{\mathcal{O}}(t) : t \geq 0\}, \bar{\mathcal{S}} = \{\bar{\mathcal{S}}(t) : t \geq 0\}$, $\bar{\mathcal{R}} = \{(\bar{\mathcal{R}}(\tau(t)) : t \geq 0\}$, *and* $\bar{\mathcal{H}} = \{\bar{\mathcal{H}}(t) : t \geq 0\}$ *are increasing and Lipschitz continuous on* $[0, \infty)$. *At the same time, the above limit satisfies*

$$\bar{\mathcal{O}}(t) = \int_0^t \sum_{i=1}^m c_i \times \mathsf{I}_{\{\bar{S}(x)=i\}} \mathsf{d}x, \tag{3.50}$$

$$\bar{\mathcal{S}}(t) = K \int_0^t \mathsf{I}_{\{\bar{S}(x)>\bar{S}(x-)\}} \mathsf{d}\bar{S}(x), \tag{3.51}$$

$$\bar{\mathcal{R}}(t) = \bar{L}(t), \tag{3.52}$$

$$\bar{\mathcal{H}}(t) = h \sum_{i=1}^m (1-\alpha_i) \int_0^t \bar{Q}(x) \mathsf{I}_{\{\bar{S}(x)=i\}} \mathsf{d}x. \tag{3.53}$$

The proof are similar as the proof in Theorem 15. We can further specify the expressions if the actions are given. We illustrate this in the following theorem.

Theorem 17. *For any fixed* $T > 0$, *we assume that each system uses* k *different staffing-level policy during* $[0, T)$. *More specifically, for the* n*th system,* i_ℓ *servers are put into operation during the time interval* $[t^n_{\ell-1}, t^n_\ell)$ *where* t^n_ℓ $(\ell = 1, \cdots, k)$ *are random and* $0 = t^n_0 < t^n_1 < \cdots < t^n_k = T$. *If with probability one,* $\lim_{n\to\infty} t^n_\ell = t_\ell$ *for* $\ell = 1, \cdots, k$, *then for* $t \in [t_{\ell-1}, t_\ell)$,

$$0 = \sum_{i=1}^m \int_0^t \mathsf{I}_{\{\bar{S}(x)=i, \bar{Q}(x)>0\}} \mathsf{d}\Big(ix - \sum_{j=1}^m \bar{T}_{ij}(x)\Big), \tag{3.54}$$

$$\bar{L}(t) = \lambda \sum_{i=1}^m \int_0^{\bar{\tau}(t)} \alpha_i \mathsf{I}_{\{\bar{S}(x)=i, \bar{Q}(x)>0\}} \mathsf{d}x, \tag{3.55}$$

$$\bar{\tau}(t) = \sum_{i=1}^{m} \int_0^t \mathsf{I}_{\{\bar{S}(x)=i,\bar{Q}(x)>0\}} \frac{\mu_i}{\lambda} \mathsf{d}x, \tag{3.56}$$

$$\bar{Q}(t) = \sum_{i=1}^{m} \int_0^t \mathsf{I}_{\{\bar{S}(x)=i,\bar{Q}(x)>0\}} \beta_i \mathsf{d}x, \tag{3.57}$$

where

$$\bar{S}(x) = i_\ell \quad \textit{for } x \in [t_{\ell-1}, t_\ell) \textit{ and } \ell = 1, \cdots, k.$$

Proof. Now we prove (3.54)-(3.57) of the theorem. Consider the subsequence $\{n_\ell : \ell \geq 1\}$ given by (3.44). When $t \in [0, t_1)$, by the positivity of $\bar{Q}(t)$, for large enough n_ℓ,

$$Q^{n_\ell}(t) > m. \tag{3.58}$$

This together with (3.17) gives that

$$\begin{aligned} 0 &= \int_0^t \mathsf{I}_{\{S^{n_\ell}(x)=i_1,Q^{n_\ell}(x)\geq i_1\}} \mathsf{d}\Big(i_1 x - \sum_{j=1}^{m} T^{n_\ell}_{i_1 j}(x)\Big) \\ &= \int_0^t \mathsf{d}\Big(i_1 x - \sum_{j=1}^{m} T^{n_\ell}_{i_1 j}(x)\Big). \end{aligned}$$

Hence,

$$\sum_{j=1}^{m} T^{n_\ell}_{i_1 j}(t) = i_1 t \quad \text{for } t \in [0, t_1),$$

which, by (3.44), implies that (3.54) holds for $t \in [0, t_1)$. For (3.55), by (3.40) and (3.58), for $t \in [0, t_1)$,

$$\begin{aligned} L^{n_\ell}(t) &= \frac{1}{n_\ell}\int_0^{\tau^{n_\ell}(t)} \sum_{i=1}^{m} \alpha_i \times \mathsf{I}_{\{S^{n_\ell}(x+v^{n_\ell}(x))=i, Q^{n_\ell}(x+v^{n_\ell}(x))>i\}} \mathsf{d}A^{n_\ell}(x) \\ &= \frac{1}{n_\ell}\int_0^{\tau^{n_\ell}(t)} \alpha_{i_1} \times \mathsf{I}_{\{S^{n_\ell}(x+v^{n_\ell}(x))=i_1, Q^{n_\ell}(x+v^{n_\ell}(x))>i_1\}} \mathsf{d}A^{n_\ell}(x) \\ &= \frac{1}{n_\ell}\int_0^{\tau^{n_\ell}(t)} \alpha_{i_1} \mathsf{d}A^{n_\ell}(x) \\ &= \frac{A^{n_\ell}(\tau^{n_\ell}(t))}{n_\ell}\alpha_{i_1} \to \lambda\alpha_{i_1}\bar{\tau}(t). \end{aligned}$$

This shows that (3.55) holds for $t \in [0, t_1)$. By Theorem 15, for $t \in [0, t_1)$, we have that

$$\begin{aligned} \bar{Q}(t) &= \lambda t - \bar{L}(t) - \mu \sum_{j=1}^{m} \bar{T}_{i_1 j}(t) \\ &= \lambda t - \lambda\alpha_{i_1}\bar{\tau}(t) - i_1\mu t \\ &= \lambda(t - \bar{\tau}(t)). \end{aligned} \tag{3.59}$$

This implies

$$\bar{\tau}(t) = \frac{\mu_{i_1}}{\lambda} t. \tag{3.60}$$

Plug (3.60) into (3.59), we have

$$\bar{Q}(t) = \beta_{i_1} t.$$

Repeating the above procedure, we can show that (3.54)-(3.57) hold for any $t \in [0, T)$. Hence we have the theorem. □

3.3 Analysis of the Long-Run Average Cost

Based on the information up to time t, our objective is to dynamically determine $\bar{S}(t)$ among $\{1, \cdots, m\}$ at any time t to minimize

$$\mathcal{AC}(T) := \frac{r \times \bar{\mathcal{R}}(T) + \bar{\mathcal{H}}(T) + \bar{\mathcal{O}}(T) + \bar{\mathcal{S}}(T)}{T} \tag{3.61}$$

for large enough T. Denote

$$\beta_i := \lambda - \mu_i = \frac{\lambda_i - i\mu}{1 - \alpha_i}, \; \mu_i = \frac{i\mu}{1 - \alpha_i}, \; \lambda_i = (1 - \alpha_i)\lambda, \qquad i = 1, \cdots, m. \tag{3.62}$$

We assume that

$$\beta_i \;\text{ is decreasing and convex on } [1, m]. \tag{3.63}$$

β_i can be viewed as the net input rate. It's natural to assume β_i is decreasing. In addition, we also assume β_i is convex in i. This implies that β_i decreases very fast with small i, but decreases very slowly with large i. That is, the initial added servers are more efficient at increasing the net input rate.

In view of Assumptions (3.11) and (3.63),

$$\beta_i > 0, \quad i \le m_0; \qquad \beta_i < 0, \quad i > m_0. \tag{3.64}$$

This, by Theorem 15, gives that for any given time interval $[s,t)$, if the fluid queue length $\bar{Q}$ is positive and the system is in i-server region, then

$$\bar{Q}(s) < \bar{Q}(t) \ \text{ for } i \le m_0, \text{ and } \quad \bar{Q}(s) > \bar{Q}(t) \ \text{ for } i > m_0. \tag{3.65}$$

We use the idea from the renewal reward theorem to solve the problem (3.61). The regenerative point is defined by $\bar{Q}(t) = 0$, that is, the points of the system empty. During each cycle, suppose we have k times of changing service regions, where staffing levels are denoted by $i_1, \ldots, i_k$, and the thresholds to switch the service region are sequentially given by $\bar{Q}_1, \cdots, \bar{Q}_k$. More concrete, starting with empty during each cycle, there are i_1 servers to process customer service requirements, and the queue length builds up. When the queue length first accumulates to $\bar{Q}_1$, we switch from i_1-server region to i_2-server region. When the queue length either builds up to (if $i_2 \le m_0$) or shrinks to (if $i_2 \ge m_0 + 1$) $\bar{Q}_2$, we change over to i_3-server region, and so on. Finally, the queue length starts with $\bar{Q}_{k-1}$ and system runs in i_k-server region, the cycle will be over as soon as the system becomes empty. Clearly, $i_1 \le m_0$, $i_k \ge m_0 + 1$, and $\bar{Q}_k = 0$.

The pairs $(i_\ell, \bar{Q}_\ell)$ $(\ell = 1, \cdots, k)$ and k are our decision variables to solve the problem (3.61). By (3.65), we have that

$$\bar{Q}_{\ell-1} < \bar{Q}_\ell \ \text{ for } i_\ell \leq m_0, \text{ and } \ \bar{Q}_{\ell-1} > \bar{Q}_\ell \ \text{ for } i_\ell \geq m_0 + 1.$$

We consider only stationary policies, which adopt same actions in each cycle, since cost will not be reduced by considering nonstationary policies. Thus, in the following, we will derive the average cost in fluid model, given a feasible policy $(i_\ell, \bar{Q}_\ell)$ $(\ell = 1, \cdots, k)$ in one cycle.

For the nth system given by Section 3.2, we repeat to use the above policy: the system starts with staffing level i_1, the staffing level will be switched from i_1 to i_2 when the queue length Q^n first reaches to $n\bar{Q}_1$. Then the ticket queue length either builds up to $n\bar{Q}_2$ (if $i_2 \leq m_0$), or reduces to $n\bar{Q}_2$ (if $i_2 > m_0$). If it first reaches $n\bar{Q}_2$ before reaching empty (which means this cycle ends), the staffing level will be switched from i_2 to i_3. This process continues until the staffing level is switched to i_k, and the system runs in i_k-server region until system becomes empty, i.e. reaches $n\bar{Q}_k = 0$. We call this policy $(i_\ell, n\bar{Q}_\ell)$ $(\ell = 1, \cdots, k)$. Then we have our results as follows.

Theorem 18. *For the nth system, we use policy $(i_\ell, n\bar{Q}_\ell)$ $(\ell = 1, \cdots, k)$. Denote $\delta_\ell = \bar{Q}_\ell - \bar{Q}_{\ell-1}$, where $\bar{Q}_0 = \bar{Q}_k = 0$. The fluid approximation*

$\left(\bar{\tau}, \bar{L}, \bar{T}, \bar{Q}\right)$ *given by* (3.44) *in Theorem* 15 *satisfies that, for* $\ell = 0, 1, \ldots,$

$$\bar{Q}(t) = \begin{cases} \beta_{i_1}\Big(t - \ell \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}}\Big), \qquad \textit{for } t \in [\ell \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}}, \ell \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}} + \frac{\delta_1}{\beta_{i_1}}), \\ \delta_1 + \beta_{i_2}\Big(t - \ell \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}} - \frac{\delta_1}{\beta_{i_1}}\Big) \\ \qquad \textit{for } t \in [\ell \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}} + \frac{\delta_1}{\beta_{i_1}}, \ell \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}} + \frac{\delta_1}{\beta_{i_1}} + \frac{\delta_2}{\beta_{i_2}}), \\ \quad \vdots \\ \sum_{j=1}^{k-1} \delta_j + \beta_{i_k}\Big(t - \ell \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}} - \sum_{j=1}^{k-1} \frac{\delta_j}{\beta_{i_j}}\Big) \\ \qquad \textit{for } t \in [\ell \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}} + \sum_{j=1}^{k-1} \frac{\delta_j}{\beta_{i_j}}, (\ell+1) \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}}). \end{cases} \tag{3.66}$$

$$\bar{L}(t) = \begin{cases} \ell \sum_{j=1}^{k} \frac{\alpha_{i_j}\mu_{i_j}\delta_j}{\beta_{i_j}} + \alpha_{i_1}\mu_{i_1}\Big(t - \ell \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}}\Big), \\ \qquad \textit{for } t \in [\ell \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}}, \ell \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}} + \frac{\delta_1}{\beta_{i_1}}), \\ \ell \sum_{j=1}^{k} \frac{\alpha_{i_j}\mu_{i_j}\delta_j}{\beta_{i_j}} + \frac{\alpha_{i_1}\mu_{i_1}\delta_1}{\beta_{i_1}} + \alpha_{i_2}\mu_{i_2}\Big(t - \ell \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}} - \frac{\delta_1}{\beta_{i_1}}\Big), \\ \qquad \textit{for } t \in [\ell \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}} + \frac{\delta_1}{\beta_{i_1}}, \ell \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}} + \frac{\delta_1}{\beta_{i_1}} + \frac{\delta_2}{\beta_{i_2}}), \\ \quad \vdots \\ \ell \sum_{j=1}^{k} \frac{\alpha_{i_j}\mu_{i_j}\delta_j}{\beta_{i_j}} + \sum_{j=1}^{k-1} \frac{\alpha_{i_j}\mu_{i_j}\delta_j}{\beta_{i_j}} + \alpha_{i_k}\mu_{i_k}\Big(t - \ell \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}} - \sum_{j=1}^{k-1} \frac{\delta_j}{\beta_{i_j}}\Big), \\ \qquad \textit{for } t \in [\ell \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}} + \sum_{j=1}^{k-1} \frac{\delta_j}{\beta_{i_j}}, (\ell+1) \sum_{j=1}^{k} \frac{\delta_j}{\beta_{i_j}}). \end{cases} \tag{3.67}$$

$$\bar{\tau}(t)=\begin{cases} \ell\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}+\frac{\mu_{i_1}}{\lambda}\Big(t-\ell\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}\Big), \\ \qquad\qquad \textit{for } t\in[\ell\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}},\ell\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}+\frac{\delta_1}{\beta_{i_1}}), \\ \ell\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}+\frac{\mu_{i_1}\delta_1}{\lambda\beta_{i_1}}+\frac{\mu_{i_2}}{\lambda}\Big(t-\ell\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}-\frac{\delta_1}{\beta_{i_1}}\Big), \\ \qquad\qquad \textit{for } t\in[\ell\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}+\frac{\delta_1}{\beta_{i_1}},\ell\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}+\frac{\delta_1}{\beta_{i_1}}+\frac{\delta_2}{\beta_{i_2}}), \\ \quad\vdots \\ \ell\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}+\sum_{j=1}^{k-1}\frac{\mu_{i_j}\delta_j}{\lambda\beta_{i_j}}+\frac{\mu_{i_k}}{\lambda}\Big(t-\ell\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}-\sum_{j=1}^{k-1}\frac{\delta_j}{\beta_{i_j}}\Big), \\ \qquad\qquad \textit{for } t\in[\ell\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}+\sum_{j=1}^{k-1}\frac{\delta_j}{\beta_{i_j}},(\ell+1)\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}). \end{cases} \tag{3.68}$$

$$\sum_{\ell=1}^{k}\sum_{j=1}^{m}\bar{T}_{i_\ell j}(t)=\begin{cases} \ell\sum_{j=1}^{k}\frac{\delta_j i_j}{\beta_{i_j}}+i_1(t-\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}), \\ \qquad \textit{for } t\in[\ell\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}},\ell\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}+\frac{\delta_1}{\beta_{i_1}}), \\ \ell\sum_{j=1}^{k}\frac{\delta_j i_j}{\beta_{i_j}}+\frac{i_1\delta_1}{\beta_{i_1}}+i_2(t-\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}-\frac{\delta_1}{\beta_{i_1}}), \\ \qquad \textit{for } t\in[\ell\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}+\frac{\delta_1}{\beta_{i_1}},\ell\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}+\frac{\delta_1}{\beta_{i_1}}+\frac{\delta_2}{\beta_{i_2}}), \\ \qquad\vdots \\ \ell\sum_{j=1}^{k}\frac{\delta_j i_j}{\beta_{i_j}}+\sum_{j=1}^{k-1}\frac{i_j\delta_j}{\beta_{i_j}}+i_k(t-\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}-\sum_{j=1}^{k-1}\frac{\delta_j}{\beta_{i_j}}), \\ \qquad \textit{for } t\in[\ell\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}+\sum_{j=1}^{k-1}\frac{\delta_j}{\beta_{i_j}},(\ell+1)\sum_{j=1}^{k}\frac{\delta_j}{\beta_{i_j}}). \end{cases} \tag{3.69}$$

Moreover, the long-run average cost incurred by the above fluid model is *equal to*

$$\Big[\sum_{\ell=1}^{k}\Big(h(1-\alpha_{i_\ell})\Big(\bar{Q}_\ell-\frac{\delta_\ell}{2}\Big)+\mu_{i_\ell}\alpha_{i_\ell}+c_{i_\ell}\Big)\frac{\delta_\ell}{\beta_{i_\ell}}+K(i_\ell-i_{\ell-1})^+\Big]\Big/\sum_{\ell=1}^{k}\frac{\delta_\ell}{\beta_{i_\ell}}. \tag{3.70}$$

Proof. For the nth system, let ξ^n be the first time of the queue length Q^n reaching $n\bar{Q}_1$, and ς^n the first time of the queue length reaching $n\bar{Q}_2$ after ξ^n. Define

$$\xi_0^n=\xi^n\wedge\frac{2\bar{Q}_1}{\beta_{i_1}}\quad\text{and}\quad\varsigma_0^n=\varsigma^n\wedge 2\Big(\frac{\bar{Q}_1}{\beta_{i_1}}+\frac{\bar{Q}_2-\bar{Q}_1}{\beta_{i_2}}\Big).$$

It follows from (3.19) that under the policy $(i_\ell, n\bar{Q}_\ell)$ ($\ell=1,\cdots,k$), for $t\in[0,\ \xi_0^n]$,

$$\begin{aligned}\frac{Q^n(t)}{n}=&\frac{A^n(t)-\lambda^n t}{n}-\frac{1}{n}\sum_{j=1}^{m}\Big[S_j^n\Big(T_{i_1j}^n(t)\Big)-\mu^n T_{i_1j}^n(t)\Big]\\&-\frac{1}{n}\sum_{\ell=1}^{A^n(\tau^n(t))}\Big[\Big(1-z_{i_1\ell}\Big)-\alpha_i\Big]\times \mathsf{I}_{\{Q^n(\tau_\ell^n+v_\ell^n)>i_1\}}\\&+\frac{\lambda^n}{n}t-\frac{\mu^n}{n}\sum_{j=1}^{m}T_{i_1j}^n(t)-\frac{1}{n}\sum_{\ell=1}^{A^n(\tau^n(t))}\alpha_{i_1}\times \mathsf{I}_{\{Q^n(\tau_\ell^n+v_\ell^n)>i_1\}}.\end{aligned} \tag{3.71}$$

Note that with probability one, $\{\xi_0^n : n\geq 1\}$ and $\{\varsigma_0^n : n\geq 1\}$ are bounded. Hence, for each $\omega\in\Omega$, there exists a subsequence of $\{\xi_0^n : n\geq 1\}$, called

$\{\xi_0^{n_\ell(\omega)}(\omega) : \ell \geq 1\}$ such that

$$\xi_0^{n_\ell(\omega)}(\omega) \to \bar{\xi}_0(\omega) \quad \text{as } \ell \to \infty.$$

By Theorem 15, we have that for $t \in [0,\ \xi_0(\omega)]$,

$$\begin{aligned}
&\Big(\tau^{n_\ell(\omega)}(t,\omega), L^{n_\ell(\omega)}(t,\omega), T^{n_\ell(\omega)}(t,\omega), Q^{n_\ell(\omega)}(t,\omega)\Big) \\
&\quad \to \Big(\bar{\tau}(t,\omega), \bar{L}(t,\omega), \bar{T}(t,\omega), \bar{Q}(t,\omega)\Big) \quad \text{as } \ell \to \infty, && (3.72) \\
&\bar{Q}(t,\omega) = \lambda t - \bar{L}(t,\omega) - \mu \sum_{j=1}^{m} \bar{T}_{i_1 j}(t,\omega), && (3.73) \\
&\bar{Q}(t,\omega) = \lambda\Big(t - \bar{\tau}(t,\omega)\Big), && (3.74) \\
&\bar{L}(t,\omega) \leq \lambda \alpha_{i_1} t, \quad \sum_{j=1}^{m} \bar{T}_{i_1 j}(t,\omega) \leq i_1 t. && (3.75)
\end{aligned}$$

The limit satisfies that for $t \in [0,\ \xi_0(\omega)]$,

$$\begin{aligned}
\bar{Q}_1 \geq \bar{Q}(t,\omega) &= \lambda t - \bar{L}(t,\omega) - \mu \sum_{j=1}^{m} \bar{T}_{i_1 j}(t,\omega) \\
&\geq \lambda t - \lambda \alpha_{i_1} t - \mu i_1 t \\
&= \Big(\lambda - \lambda \alpha_{i_1} - \mu i_1\Big) t. \qquad (3.76)
\end{aligned}$$

Hence, $\bar{Q}(t,\omega)$ is positive only except $t = 0$. By again Theorem 15, we have that

$$\bar{L}(t,\omega) = \lambda\alpha_{i_1}\bar{\tau}(t,\omega) \quad \text{and} \quad \sum_{j=1}^{m} \bar{T}_{i_1 j}(t) = i_1 t. \tag{3.77}$$

This, by (3.73)-(3.74),

$$\begin{aligned} \bar{Q}(t,\omega) &= \lambda t - \lambda\alpha_{i_1}\bar{\tau}(t,\omega) - \mu i_1 t \\ &= \lambda\Big(t - \bar{\tau}(t,\omega)\Big). \end{aligned} \tag{3.78}$$

This implies

$$\bar{\tau}(t,\omega) = \frac{\mu_{i_1}}{\lambda} t. \tag{3.79}$$

Plugging (3.79) into (3.78) yields that for $t \in [0,\ \xi_0(\omega)]$,

$$\bar{Q}(t,\omega) = \beta_{i_1} t. \tag{3.80}$$

By the first inequality of (3.76), we have that for $t \in [0,\ \xi_0(\omega)]$,

$$\bar{Q}_1 \geq \beta_{i_1} t,$$

which implies

$$\xi_0(\omega) \le \bar{Q}_1/\beta_{i_1}.$$

In view of the definitions of ξ^n and $\xi_0(\omega)$, we have that

$$\lim_{\ell\to\infty} \xi^{n_\ell(\omega)}(\omega) = \xi_0(\omega) < \frac{2\bar{Q}_1}{\beta_{i_1}}. \tag{3.81}$$

Therefore, for large enough n_ℓ,

$$[0,\ \xi_0^{n_\ell(\omega)}(\omega)] = [0,\ \xi^{n_\ell(\omega)}(\omega)].$$

Thus, replacing t by $\xi^{n_\ell(\omega)}(\omega)$ in (3.71), its right-hand side is $\bar{Q}_1$. Letting $n_\ell(\omega) \to \infty$, by (3.73), (3.78) and (3.81), we have that $\bar{Q}_1 = \beta_{i_1} \times \xi_0(\omega)$. This gives $\xi_0(\omega) = \frac{\bar{Q}_1}{\beta_{i_1}}$. Combining (3.72), and (3.77)-(3.79), we know that $(\bar{\tau}, \bar{L}, \bar{T}, \bar{Q})$ given by (3.68)-(3.69) holds for $t \in [0,\ \bar{Q}_1/\beta_{i_1}]$. Going along the similar line, we can prove the theorem for the other intervals. Here the details are omitted.

Now let's verify that the long-run average cost for the fluid model is characterized by (3.70). Let $\bar{T}_\ell$ denotes the time length of i_ℓ-server region in the fluid model. Then

$$\bar{T}_\ell = \frac{\delta_\ell}{\beta_{i_\ell}} = \frac{\bar{Q}_\ell - \bar{Q}_{\ell-1}}{\beta_{i_\ell}}. \tag{3.82}$$

The system dynamics of the fluid model in one cycle are shown in Figure 3.1.

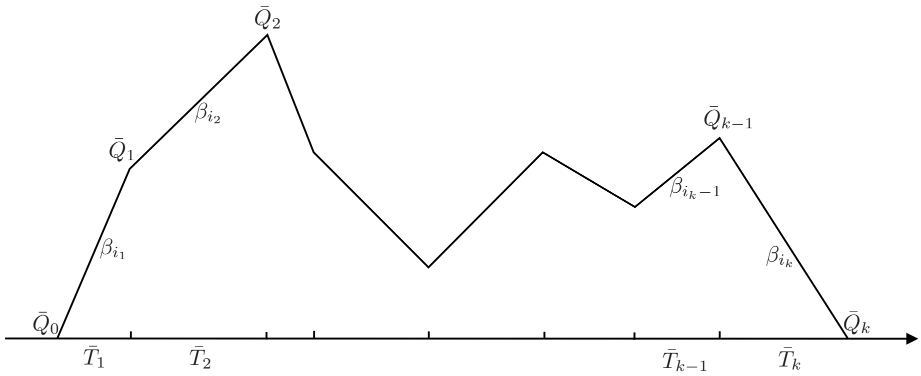


Fig. 3.1: System Dynamics

Denote one cycle length by $\bar{T}_c = \sum_{\ell=1}^{k} \bar{T}_\ell$. By Proposition 16 and the above analysis, the relevant cost in one cycle are:

- operating costs : $\mathcal{O}(\bar{T}_c) = \sum_{\ell=1}^{k} c_{i_\ell} \bar{T}_\ell$;
- setup cost: $\mathcal{S}(\bar{T}_c) = K \sum_{\ell=1}^{k} (i_\ell - i_{\ell-1})^+$, where $i_0 = 0$;
- customer abandonment costs :

$$\mathcal{R}(\bar{T}_c) = \lambda \sum_{\ell=1}^{k} \alpha_{i_\ell} (\bar{\tau}(\bar{T}_\ell) - \bar{\tau}(\bar{T}_{\ell-1})) = \sum_{\ell=1}^{k} \alpha_{i_\ell} \mu_{i_\ell} \bar{T}_\ell;$$

- customer delay cost: $\mathcal{H}(\bar{T}_c) = \frac{h}{2}\sum_{\ell=1}^{k}(1-\alpha_{i_\ell})(\bar{Q}_{\ell-1}+\bar{Q}_\ell)\bar{T}_\ell$.

By (3.82), the average cost during one cycle is

$$\Big[\sum_{\ell=1}^{k}\Big(h(1-\alpha_{i_\ell})\Big(\bar{Q}_\ell-\frac{\delta_\ell}{2}\Big)+\mu_{i_\ell}\alpha_{i_\ell}+c_{i_\ell}\Big)\frac{\delta_\ell}{\beta_{i_\ell}}+K(i_\ell-i_{\ell-1})^+\Big]\Big/\sum_{\ell=1}^{k}\frac{\delta_\ell}{\beta_{i_\ell}}.$$

For any given large T, we use $\mathcal{AC}^n(T)$ to denote the average cost incurred by nth system under policy $(i_\ell, n\bar{Q}_\ell)$ $(\ell = 1,\dots,k)$. Then

$$\begin{aligned}&\lim_{T\to\infty}\lim_{n\to\infty}\mathcal{AC}^n(T)\\ &=\Big[\sum_{\ell=1}^{k}\Big(h(1-\alpha_{i_\ell})\Big(\bar{Q}_\ell-\frac{\delta_\ell}{2}\Big)+\mu_{i_\ell}\alpha_{i_\ell}+c_{i_\ell}\Big)\frac{\delta_\ell}{\beta_{i_\ell}}+K(i_\ell-i_{\ell-1})^+\Big]\Big/\sum_{\ell=1}^{k}\frac{\delta_\ell}{\beta_{i_\ell}}.\end{aligned} \tag{3.83}$$

Consequently, we have the theorem. □

3.4 The Optimal Policy in the Fluid Model

In this section, we minimize the objective function (3.70) and find the optimal policy in fluid model. The constraints are:

$$\delta_\ell = \bar{Q}_\ell - \bar{Q}_{\ell-1} \ \text{ and } \ i_\ell \in [0,m], \qquad \ell = 1,\dots,k. \tag{3.84}$$

Note that we have relaxed the integer requirement on i_ℓ in the above constraints. This is consistent with the continuous nature of the fluid model. In terms of implementing the optimal solution, this should not be a problem. For instance, if $i_\ell = 2.5$, we can alternately use 2 and 3 servers in consecutive (regeneration) cycles. Also note that we do not require $\delta_\ell \geq 0$; $\bar{Q}_\ell$ could very well be less than $\bar{Q}_{\ell-1}$. However, do note that δ_ℓ and β_{i_ℓ} always have the same sign ($\bar{Q}_\ell < \bar{Q}_{\ell-1}$ means $i_\ell > m_0$; so, $\beta_{i_\ell} < 0$). Thus, $\delta_\ell/\beta_{i_\ell} \geq 0$, for all ℓ.

Another observation is this. The setup cost is lower-bounded by

$$\sum_{\ell=1}^{k} K(i_\ell - i_{\ell-1})^+ \geq \sum_{\ell=1}^{k} K(i_\ell - i_{\ell-1}) = K i_{k+1}.$$

We make further assumptions on the abandonment probability α_i. Namely,

$$\mu_i\alpha_i + c_i \quad \text{is increasing and convex,} \tag{3.85}$$

$$1 - \alpha_i \ \ \text{is increasing and convex.} \tag{3.86}$$

It's natural to assume $\mu_i\alpha_i + c_i$ and $1 - \alpha_i$ is increasing. In addition, we also assume them to be convex in i. This implies that marginal cost is increasing, which further imply the following results in Proposition 19.

Proposition 19. *Under Assumptions* (3.63) *and* (3.85)-(3.86)*,*

$$\min_{(i_\ell,\delta_\ell)} \Big[\sum_{\ell=1}^{k} \Big(h(1-\alpha_{i_\ell})\Big(\bar{Q}_\ell - \frac{\delta_\ell}{2}\Big) + \mu_{i_\ell}\alpha_{i_\ell} + c_{i_\ell} \Big) \frac{\delta_\ell}{\beta_{i_\ell}} + K(i_\ell - i_{\ell-1})^+ \Big] \Big/ \sum_{\ell=1}^{k} \frac{\delta_\ell}{\beta_{i_\ell}}$$

$$\geq \min_{i_1,i_2} \Big\{ \frac{(\mu_{i_2}\alpha_{i_2} + c_{i_2})\beta_{i_1} - (\mu_{i_1}\alpha_{i_1} + c_{i_1})\beta_{i_2}}{\beta_{i_1} - \beta_{i_2}} + \sqrt{2hK\frac{c^2}{c_\alpha} i_2} \Big\}, \tag{3.87}$$

where i_ℓ and δ_ℓ satisfy (3.84)*,* $i_1 \le m_0$*,* $i_2 \ge m_0 + 1$*,*

$$\frac{1}{c} = \frac{1}{\beta_{i_1}} - \frac{1}{\beta_{i_2}}, \quad \textit{and} \quad \frac{1}{c_\alpha} = \frac{1-\alpha_{i_1}}{\beta_{i_1}} - \frac{1-\alpha_{i_2}}{\beta_{i_2}}. \tag{3.88}$$

.

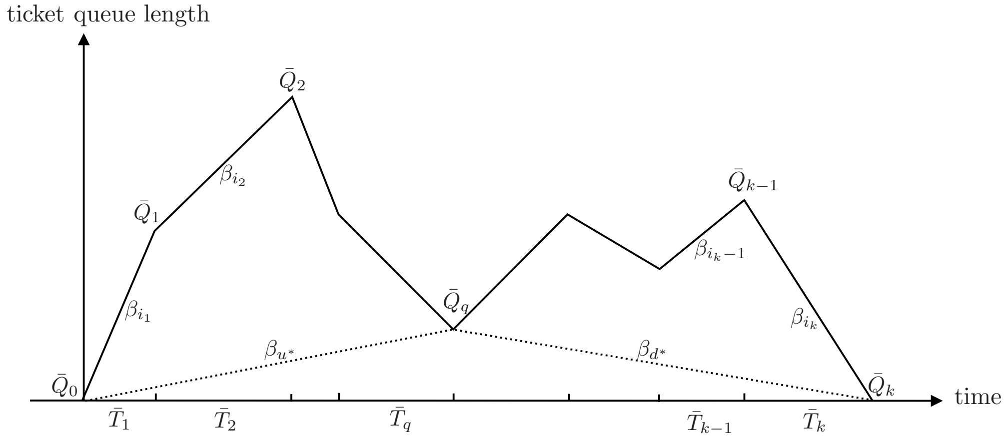


Fig. 3.2: Generating a Two-Piece Policy

Proof. Let $\bar{Q}_q$ represent the smallest positive $\bar{Q}_\ell$, i.e.,

$$\bar{Q}_q = \min\{\bar{Q}_1, \cdots, \bar{Q}_k\}.$$

We connect points $(0, 0)$ and $(\sum_{\ell=1}^{q} \bar{T}_\ell, \bar{Q}_q)$, $(\sum_{\ell=1}^{q} \bar{T}_\ell, \bar{Q}_q)$ and $(\sum_{\ell=1}^{k} \bar{T}_\ell, 0)$ (dotted line in Figure 3.2), then we derive a 2-piece policy. The first piece has slop β_{u^*} and the second piece has slop β_{d^*}, where

$$\beta_{u^*} = \frac{\sum_{\ell=1}^{q} \bar{T}_\ell \beta_{i_\ell}}{\sum_{\ell=1}^{q} \bar{T}_\ell} \ge 0, \text{ and } \beta_{d^*} = \frac{\sum_{\ell=q+1}^{k} \bar{T}_\ell \beta_{i_\ell}}{\sum_{\ell=q+1}^{k} \bar{T}_\ell} \le 0. \tag{3.89}$$

β_{d^*} is nonpositive because $\bar{Q}_q = \sum_{\ell=1}^{q} \bar{T}_\ell \beta_{i_\ell} = -\sum_{\ell=q+1}^{k} \bar{T}_\ell \beta_{i_\ell}$. Let

$$u = \frac{\sum_{\ell=1}^{q} i_\ell \bar{T}_\ell}{\sum_{\ell=1}^{q} \bar{T}_\ell} \text{ and } d = \frac{\sum_{\ell=q+1}^{k} i_\ell \bar{T}_\ell}{\sum_{\ell=q+1}^{k} \bar{T}_\ell}.$$

Taking into account the convexity of β_i with respect to i (Assumption (3.63)), we have

$$\beta_u \leq \frac{\sum_{\ell=1}^{q} \bar{T}_\ell \beta_{i_\ell}}{\sum_{\ell=1}^{q} \bar{T}_\ell} = \beta_{u^*} \text{ and } \beta_d \leq \frac{\sum_{\ell=q+1}^{k} \bar{T}_\ell \beta_{i_\ell}}{\sum_{\ell=q+1}^{k} \bar{T}_\ell} = \beta_{d^*}. \tag{3.90}$$

Hence, $u \geq u^*$ and $d \geq d^*$ follows from β_i decreasing in i (Assumption (3.63)).

Now we show that 2-piece cost is less than k-piece cost given by the left-hand side of (3.87). For the k-piece cost, the customer delay cost, without constant multiplier $h/2$, is

$$\begin{aligned}
&\sum_{\ell=1}^{k}(1-\alpha_{i_\ell})(\bar{Q}_{\ell-1}+\bar{Q}_\ell)\bar{T}_\ell \\
&\quad = \sum_{\ell=1}^{k} 2\bar{T}_\ell\Big(\bar{Q}_{\ell-1} + \frac{\bar{T}_\ell \beta_{i_\ell}}{2}\Big)(1-\alpha_{i_\ell}) \\
&\quad = \sum_{\ell=1}^{k} \bar{T}_\ell\Big(2\sum_{\ell'=1}^{\ell-1} \bar{T}_{\ell'}\beta_{i_{\ell'}} + \bar{T}_\ell \beta_{i_\ell}\Big)(1-\alpha_{i_\ell}) \\
&\quad = \sum_{\ell=1}^{k} \bar{T}_\ell (1-\alpha_{i_\ell}) \sum_{\ell'=1}^{\ell} \bar{T}_{\ell'}\beta_{i_{\ell'}} + \sum_{\ell=1}^{k} \bar{T}_\ell (1-\alpha_{i_\ell}) \sum_{\ell'=1}^{\ell-1} \bar{T}_{\ell'}\beta_{i_{\ell'}}.
\end{aligned} \tag{3.91}$$

For the 2-piece cost, the customer delay cost, without constant multiplier

$h/2$, is

$$\begin{aligned}
&\Big(\sum_{\ell=1}^{q}\bar{T}_\ell\Big)^2(1-\alpha_{u^*})\beta_{u^*}-\Big(\sum_{\ell=q+1}^{k}\bar{T}_\ell\Big)^2(1-\alpha_{d^*})\beta_{d^*}\\
&\quad=\sum_{\ell=1}^{q}\bar{T}_\ell(1-\alpha_{u^*})\cdot\sum_{\ell=1}^{q}\bar{T}_\ell\beta_{i_\ell}-\sum_{\ell=q+1}^{k}\bar{T}_\ell(1-\alpha_{d^*})\cdot\sum_{\ell=q+1}^{k}\bar{T}_\ell\beta_{i_\ell}\\
&\quad\le\sum_{\ell=1}^{q}\bar{T}_\ell(1-\alpha_{i_\ell})\cdot\sum_{\ell=1}^{q}\bar{T}_\ell\beta_{i_\ell}-\sum_{\ell=q+1}^{k}\bar{T}_\ell(1-\alpha_{i_\ell})\cdot\sum_{\ell=q+1}^{k}\bar{T}_\ell\beta_{i_\ell}\\
&\quad=\sum_{\ell=1}^{q}\bar{T}_\ell(1-\alpha_{i_\ell})\cdot\sum_{\ell=1}^{q}\bar{T}_\ell\beta_{i_\ell}+\sum_{\ell=q+1}^{k}\bar{T}_\ell(1-\alpha_{i_\ell})\cdot\sum_{\ell=1}^{q}\bar{T}_\ell\beta_{i_\ell}\\
&\quad=\sum_{\ell=1}^{k}\bar{T}_\ell(1-\alpha_{i_\ell})\cdot\sum_{\ell=1}^{q}\bar{T}_\ell\beta_{i_\ell}, \qquad (3.92)
\end{aligned}$$

where the inequality follows from $1-\alpha_i$ increasing and convexity with respect to i

$$1-\alpha_{u^*}\le 1-\alpha_u\le\frac{\sum_{\ell=1}^{q}\bar{T}_\ell(1-\alpha_{i_\ell})}{\sum_{\ell=1}^{q}\bar{T}_\ell}, \qquad (3.93)$$

$$1-\alpha_{d^*}\le 1-\alpha_d\le\frac{\sum_{\ell=q+1}^{k}\bar{T}_\ell(1-\alpha_{i_\ell})}{\sum_{\ell=q+1}^{k}\bar{T}_\ell}, \qquad (3.94)$$

see Assumption (3.86). Therefore, to prove the customer delay cost given by the k-piece policy is larger than the customer delay cost incurred by the 2-piece policy, it is sufficient to show that

$$\sum_{\ell=1}^{k}\bar{T}_\ell(1-\alpha_{i_\ell})\cdot\sum_{\ell=1}^{q}\bar{T}_\ell\beta_{i_\ell}$$

$$\leq \sum_{\ell=1}^{k} \bar{T}_\ell (1-\alpha_{i_\ell}) \sum_{\ell'=1}^{\ell} \bar{T}_{\ell'} \beta_{i_{\ell'}} + \sum_{\ell=1}^{k} \bar{T}_\ell (1-\alpha_{i_\ell}) \sum_{\ell'=1}^{\ell-1} \bar{T}_{\ell'} \beta_{i_{\ell'}} \tag{3.95}$$

After simplification, (3.95) is equivalent to

$$\begin{aligned}
&\sum_{\ell=1}^{q} \bar{T}_\ell (1-\alpha_{i_\ell}) \sum_{\ell'=\ell+1}^{q} \bar{T}_{\ell'} \beta_{i_{\ell'}} \\
&\quad \leq \sum_{\ell=q+1}^{k} \bar{T}_\ell (1-\alpha_{i_\ell}) \sum_{\ell'=q+1}^{\ell} \bar{T}_{\ell'} \beta_{i_{\ell'}} + \sum_{\ell=1}^{k} \bar{T}_\ell (1-\alpha_{i_\ell}) \sum_{\ell'=1}^{\ell-1} \bar{T}_{\ell'} \beta_{i_{\ell'}}.
\end{aligned} \tag{3.96}$$

We notice that $\sum_{\ell'=1}^{\ell-1} \bar{T}_{\ell'} \beta_{i_{\ell'}} \geq 0$ for any ℓ. For $\ell \leq q$, we have $\sum_{\ell'=1}^{q} \bar{T}_{\ell'} \beta_{i_{\ell'}} \leq \sum_{\ell'=1}^{\ell} \bar{T}_{\ell'} \beta_{i_{\ell'}}$, therefore $\sum_{\ell'=\ell+1}^{q} \bar{T}_{\ell'} \beta_{i_{\ell'}} \leq 0$; for $\ell > q$, we have $\sum_{\ell'=1}^{q} \bar{T}_{\ell'} \beta_{i_{\ell'}} \leq \sum_{\ell'=1}^{\ell} \bar{T}_{\ell'} \beta_{i_{\ell'}}$, therefore $\sum_{\ell'=q+1}^{\ell} \bar{T}_{\ell'} \beta_{i_{\ell'}} \geq 0$. Thus in (3.96), the left-hand side is negative and the right-hand side is positive, and (3.96) is true.

Next, consider the abandonment and operating cost.

$$\begin{aligned}
&(\mu_{u^*}\alpha_{u^*} + c_{u^*}) \sum_{\ell=1}^{q} \bar{T}_\ell + (\mu_{d^*}\alpha_{d^*} + c_{d^*}) \sum_{\ell=q+1}^{k} \bar{T}_\ell \\
&\quad \leq (\mu_u \alpha_u + c_u) \sum_{\ell=1}^{q} \bar{T}_\ell + (\mu_d \alpha_d + c_d) \sum_{\ell=q+1}^{k} \bar{T}_\ell \\
&\quad \leq \sum_{\ell=1}^{q} (\mu_{i_\ell}\alpha_{i_\ell} + c_\ell) \bar{T}_\ell + \sum_{\ell=q+1}^{k} (\mu_{i_\ell}\alpha_{i_\ell} + c_\ell) \bar{T}_\ell \\
&\quad = \sum_{\ell=1}^{k} (\mu_{i_\ell}\alpha_{i_\ell} + c_\ell) \bar{T}_\ell.
\end{aligned} \tag{3.97}$$

Here the first inequality follows from $\mu_i \alpha_i + c_i$ increasing in i, and the second inequality follows from convexity of $\mu_i\alpha_i + c_i$ in i, see (3.85). (3.97) implies that the abandonment and operation cost incurred by the k-piece policy is larger than the one given by the 2-piece policy.

Finally, we look at setup cost. We have

$$\sum_{\ell=1}^{k+1}(i_\ell - i_{\ell-1})^+K \geq \big[(u^*)^+ + (d^* - u^*)^+\big]K = d^*K.$$

The above inequality is true according to the definition of $\bar{Q}_q$. Thus, compared with the $(k+1)$-piece policy, we can get better-off when the 2-piece policy is implemented.

Now we prove the optimization problem for 2-piece policies can be written as the right-hand side of (3.87). As any 2-piece policy can be determined by three variables, namely, i_1, i_2 and $\bar{Q}_1$. That is, we need to decide what staffing level to start the system ($i_1 < m_0$), which threshold level for the queue length to switch another staffing level ($\bar{Q}_1$), and what staffing level to be used after switching (i_2). For the 2-piece policy with parameters $(i_1, i_2, \bar{Q}_1)$, the system average cost has the following three parts:

- average customer delay cost $= \frac{hc}{2c_\alpha}\bar{Q}_1$;
- average setup cost $= \frac{cK}{\bar{Q}_1} i_2$ (note $i_1 + (i_2 - i_1)^+ = i_2$ here);
- average abandonment and operating cost $= (\mu_{i_1}\alpha_{i_1} + c_{i_1})\frac{c}{\beta_{i_1}} - (\mu_{i_2}\alpha_{i_2} +$

$c_{i_2})\frac{c}{\beta_{i_2}}$;

where c and c_α are given in the proposition. The optimization problem can be written as

$$\min_{i_1,i_2,\bar{Q}_1}\left\{\frac{hc}{2c_\alpha}\bar{Q}_1+\frac{cK}{\bar{Q}_1}i_2+(\mu_{i_1}\alpha_{i_1}+c_{i_1})\frac{c}{\beta_{i_1}}-(\mu_{i_2}\alpha_{i_2}+c_{i_2})\frac{c}{\beta_{i_2}}\right\}. \tag{3.98}$$

We can first optimize $\bar{Q}_1$, and use $\bar{Q}^*$ to represent optimal $\bar{Q}_1$, i.e.,

$$\bar{Q}^*=\sqrt{\frac{2K}{h}c_\alpha i_2},$$

which implies that (3.98) is equivalent to

$$\min_{i_1,i_2}\left\{\sqrt{2hK\frac{c^2}{c_\alpha}i_2}+\frac{(\mu_{i_2}\alpha_{i_2}+c_{i_2})\beta_{i_1}-(\mu_{i_1}\alpha_{i_1}+c_{i_1})\beta_{i_2}}{\beta_{i_1}-\beta_{i_2}}\right\}.$$

This completes the proof of the theorem. □

Now let

$$(i_1^*,i_2^*)=\arg\min_{i_1,i_2}\left\{\sqrt{2hK\frac{c^2}{c_\alpha}i_2}+\frac{(\mu_{i_2}\alpha_{i_2}+c_{i_2})\beta_{i_1}-(\mu_{i_1}\alpha_{i_1}+c_{i_1})\beta_{i_2}}{\beta_{i_1}-\beta_{i_2}}\right\}, \tag{3.99}$$

$$\bar{Q}^*=\sqrt{\frac{2K}{h}c_\alpha^* i_2^*}\quad\text{with}\quad\frac{1}{c_\alpha^*}=\frac{1-\alpha_{i_1^*}}{\beta_{i_1^*}}-\frac{1-\alpha_{i_2^*}}{\beta_{i_2^*}}. \tag{3.100}$$

Corollary 1. *Assume that* $1-\alpha_i = a+bi$ *with* $a, b \geq 0$, *and* $\frac{i\mu}{a+bi}(1-a-bi)+c_i$ *is increasing and convex with respect to* i. *Then,*

$$(i_1^*, i_2^*) = (m_0, m_0+1). \tag{3.101}$$

Remark 1. *It is straightforward to see that if* $1-\alpha_i = a+bi$ *with* $a, b \geq 0$, *then* β_i *is is decreasing and convex on* $[1, m]$. *Thus, we know that the assumptions given by the corollary imply that* (3.63) *and* (3.86) *hold.*

Proof. To prove the corollary, it is sufficient to show that

$$\frac{(\mu_{i_2}\alpha_{i_2}+c_{i_2})\beta_{i_1}-(\mu_{i_1}\alpha_{i_1}+c_{i_1})\beta_{i_2}}{\beta_{i_1}-\beta_{i_2}} \quad \text{and} \quad \frac{c^2}{c_\alpha} \tag{3.102}$$

are increasing in i_2 and decreasing in i_1. First we consider the monotonicity of c^2/c_α in i_1 and i_2.

To get the increasing property of c^2/c_α in i_2, it suffices to show that

$$\frac{2}{c}\cdot\frac{\mathsf{d}c}{\mathsf{d}i_2} \geq \frac{1}{c_\alpha}\cdot\frac{\mathsf{d}c_\alpha}{\mathsf{d}i_2}. \tag{3.103}$$

Taking derivative with respect to i_2 on both sides of $\frac{1}{c_\alpha} = \frac{1-\alpha_{i_1}}{\beta_{i_1}} - \frac{1-\alpha_{i_2}}{\beta_{i_2}}$, we have

$$-\frac{1}{c_\alpha^2}\cdot\frac{\mathsf{d}c_\alpha}{\mathsf{d}i_2} = -\frac{1}{\beta_{i_2}^2}\cdot\Big(\beta_{i_2}b-(1-\alpha_{i_2})\frac{\mathsf{d}\beta_{i_2}}{\mathsf{d}i_2}\Big),$$

which implies

$$\frac{1}{c_\alpha} \cdot \frac{\mathsf{d} c_\alpha}{\mathsf{d} i_2} = \frac{c_\alpha}{\beta_{i_2}^2} \Big(\beta_{i_2} b - (1-\alpha_{i_2}) \frac{\mathsf{d}\beta_{i_2}}{\mathsf{d} i_2} \Big).$$

Similarly, taking derivative with respect to i_2 on both sides of $\frac{1}{c} = \frac{1}{\beta_{i_1}} - \frac{1}{\beta_{i_2}}$, we have

$$-\frac{1}{c^2} \cdot \frac{\mathsf{d} c}{\mathsf{d} i_2} = \frac{1}{\beta_{i_2}^2} \cdot \frac{\mathsf{d}\beta_{i_2}}{\mathsf{d} i_2},$$

which implies

$$\frac{1}{c} \cdot \frac{\mathsf{d} c}{\mathsf{d} i_2} = -\frac{c}{\beta_{i_2}^2} \cdot \frac{\mathsf{d}\beta_{i_2}}{\mathsf{d} i_2}.$$

Therefore (3.103) is equivalent to

$$\frac{\mathsf{d}\beta_{i_2}}{\mathsf{d} i_2} \cdot \Big(\frac{2}{c_\alpha} - \frac{1-\alpha_{i_2}}{c} \Big) \le -\frac{b\beta_{i_2}}{c}. \tag{3.104}$$

Substituting c and c_α we have

$$\frac{2}{c_\alpha} - \frac{1-\alpha_{i_2}}{c} = \frac{2(1-\alpha_{i_1}) - (1-\alpha_{i_2})}{\beta_{i_1}} - \frac{1-\alpha_{i_2}}{\beta_{i_2}}.$$

This implies that (3.104) is equivalent to

$$\frac{a\mu}{a+bi_2}\Big(\lambda - \frac{i_1\mu}{a+bi_1}\Big) + \Big(\frac{i_2\mu}{a+bi_2} - \lambda\Big)\frac{a\mu[(a+bi_1)^2 + b^2(i_1-i_2)^2]}{(a+bi_1)(a+bi_2)^2} \ge 0. \tag{3.105}$$

Since $i_2 > i_1$ and $\frac{i_2\mu}{a+bi_2} \geq \lambda \geq \frac{i_1\mu}{a+bi_1}$, (3.105) is true.

Next we consider the decreasing property of c^2/c_α in i_1. It suffices to show taht

$$\frac{2}{c} \cdot \frac{\mathsf{d}c}{\mathsf{d}i_1} \leq \frac{1}{c_\alpha} \cdot \frac{\mathsf{d}c_\alpha}{\mathsf{d}i_1}. \tag{3.106}$$

Taking derivative with respect to i_1 on both sides of $\frac{1}{c_\alpha} = \frac{1-\alpha_{i_1}}{\beta_{i_1}} - \frac{1-\alpha_{i_2}}{\beta_{i_2}}$, we have

$$\frac{1}{c_\alpha^2} \cdot \frac{\mathsf{d}c_\alpha}{\mathsf{d}i_1} = \frac{1}{\beta_{i_1}^2} \cdot \Big(- \beta_{i_1} b + (1-\alpha_{i_1})\frac{\mathsf{d}\beta_{i_1}}{\mathsf{d}i_1}\Big),$$

which implies

$$\frac{1}{c_\alpha} \cdot \frac{\mathsf{d}c_\alpha}{\mathsf{d}i_1} = \frac{c_\alpha}{\beta_{i_1}^2}\Big(- \beta_{i_1} b + (1-\alpha_{i_1})\frac{\mathsf{d}\beta_{i_1}}{\mathsf{d}i_1}\Big).$$

Similarly, taking derivative with respect to i_1 on both sides of $\frac{1}{c} = \frac{1}{\beta_{i_1}} - \frac{1}{\beta_{i_2}}$, we have

$$\frac{1}{c^2} \cdot \frac{\mathsf{d}c}{\mathsf{d}i_1} = \frac{1}{\beta_{i_1}^2} \cdot \frac{\mathsf{d}\beta_{i_1}}{\mathsf{d}i_1},$$

which implies

$$\frac{1}{c} \cdot \frac{\mathsf{d}c}{\mathsf{d}i_1} = \frac{c}{\beta_{i_1}^2} \cdot \frac{\mathsf{d}\beta_{i_1}}{\mathsf{d}i_1}.$$

Therefore (3.106) is equivalent to

$$\frac{\mathsf{d}\beta_{i_1}}{\mathsf{d}i_1} \cdot \Big(\frac{2}{c_\alpha} - \frac{1-\alpha_{i_1}}{c}\Big) \le \frac{-b\beta_{i_1}}{c}. \tag{3.107}$$

Substituting c and c_α we have

$$\frac{2}{c_\alpha} - \frac{1-\alpha_{i_1}}{c} = \frac{-2(1-\alpha_{i_2}) + (1-\alpha_{i_1})}{\beta_{i_2}} + \frac{1-\alpha_{i_1}}{\beta_{i_1}}.$$

This gives that (3.107) is equivalent to

$$\frac{i_2\mu}{a+bi_2} - \lambda + \Big(\lambda - \frac{i_1\mu}{a+bi_1}\Big)\Big[\frac{a+bi_1}{a+bi_2} + \frac{2b(i_2-i_1)}{a+bi_1}\Big] \ge 0. \tag{3.108}$$

Since $i_2 > i_1$ and $\frac{i_2\mu}{a+bi_2} \ge \lambda \ge \frac{i_1\mu}{a+bi_1}$, (3.108) is true.

Finally we consider the monotonicity of the first term in (3.102). Note that

$$\begin{aligned}
&\frac{(\mu_{i_2}\alpha_{i_2} + c_{i_2})\beta_{i_1} - (\mu_{i_1}\alpha_{i_1} + c_{i_1})\beta_{i_2}}{\beta_{i_1} - \beta_{i_2}} \\
&\quad = \mu_{i_1}\alpha_{i_1} + c_{i_1} + \frac{\big[(\mu_{i_2}\alpha_{i_2} + c_{i_2}) - (\mu_{i_1}\alpha_{i_1} + c_{i_1})\big]\beta_{i_1}}{\mu_{i_2} - \mu_{i_1}} \\
&\quad = \mu_{i_1}\alpha_{i_1} + c_{i_1} + \beta_{i_1}\frac{(\mu_{i_2}\alpha_{i_2} + c_{i_2}) - (\mu_{i_1}\alpha_{i_1} + c_{i_1})}{i_2 - i_1}\Big/\frac{\mu_{i_2} - \mu_{i_1}}{i_2 - i_1}.
\end{aligned} \tag{3.109}$$

By the increasing property and convexity of $\mu_i\alpha_i + c_i$, we have that for fixed

i_1,

$$\beta_{i_1}\frac{(\mu_{i_2}\alpha_{i_2}+c_{i_2})-(\mu_{i_1}\alpha_{i_1}+c_{i_1})}{i_2-i_1} \text{ is positive and increasing in } i_2. \tag{3.110}$$

By the increasing property and concavity of μ_i (as $\beta_i = \lambda - \mu_i$ is decreasing and convex), we know that for fixed i_1,

$$\frac{\mu_{i_2}-\mu_{i_1}}{i_2-i_1} \text{ is positive and decreasing in } i_2. \tag{3.111}$$

Combining (3.109)-(3.111) yields that for fixed i_1,

$$\frac{(\mu_{i_2}\alpha_{i_2}+c_{i_2})\beta_{i_1}-(\mu_{i_1}\alpha_{i_1}+c_{i_1})\beta_{i_2}}{\beta_{i_1}-\beta_{i_2}} \text{ is increasing in } i_2.$$

Finally consider the monotonicity of the first term of (3.102) in i_1. Similar to (3.109), we have

$$\begin{aligned}
&\frac{(\mu_{i_2}\alpha_{i_2}+c_{i_2})\beta_{i_1}-(\mu_{i_1}\alpha_{i_1}+c_{i_1})\beta_{i_2}}{\beta_{i_1}-\beta_{i_2}} \\
&\quad= \mu_{i_2}\alpha_{i_2}+c_{i_2}-\frac{\big[(\mu_{i_1}\alpha_{i_1}+c_{i_1})-(\mu_{i_2}\alpha_{i_2}+c_{i_2})\big]\beta_{i_2}}{\mu_{i_2}-\mu_{i_1}} \\
&\quad= \mu_{i_2}\alpha_{i_2}+c_{i_2}+\beta_{i_2}\frac{\big[(\mu_{i_2}\alpha_{i_2}+c_{i_2})-(\mu_{i_1}\alpha_{i_1}+c_{i_1})}{i_2-i_1}\Big/\frac{\mu_{i_2}-\mu_{i_1}}{i_2-i_1}.
\end{aligned}$$

Similar to (3.110)-(3.111), we can prove

$$\frac{(\mu_{i_2}\alpha_{i_2}+c_{i_2})\beta_{i_1}-(\mu_{i_1}\alpha_{i_1}+c_{i_1})\beta_{i_2}}{\beta_{i_1}-\beta_{i_2}}$$

is decreasing in i_1. Thus we have the corollary. □

3.5 Asymptotic Optimality

In Section 3.3 and Section 3.4, we only consider cyclical policy. That is, by using that policy, ticket queue length will reach system empty infinitely many times. We exclude policies who are not cyclical because they cannot be optimal. We illustrate this point in the following.

Suppose there exists one policy, after finite time, ticket queue length will never reach system empty. In figure 3.3, we use solid line to represent ticket queue length trajectory by using this policy. Based on that, we will generate a new policy, whose ticket queue length trajectory are represented by the dotted line. The dotted line hits system empty (i.e. ticket queue length 0) after finite time, say t_s. Then we show that dotted line incurs lower average cost. Suppose Q_s is the smallest ticket queue length among all positive ticket queue lengths. Suppose the first piece of solid line has slope β_{i_1}. The new policy represented by the dotted line is: in the first time interval $[0, \frac{Q_s}{\beta_{i_1}})$, set staffing level i_{m_0+1}; at time point $\frac{Q_s}{\beta_{i_1}}$, adjust staffing level from i_{m_0+1} to i_1; from time point $\frac{Q_s}{\beta_{i_1}}$ on, follow exactly same actions determined by initial policy. Compared with initial policy, this new policy have same server operation cost, server setup cost, and customer abandonment costs, except in the first interval $[0, \frac{Q_s}{\beta_{i_1}})$. So new policy and initial policy have same average server setup cost, average server operating cost, and average customer abandonment cost. But new policy can reduce average customer delay cost by at least $hQ_s(1-\alpha_1)$. Now we can start from time point t_s on, and continue same procedure to generate another new policy and find the next time point when ticket queue length hits system empty. Continuing along this line, we can find that optimal policy belongs to cyclical policies; or in other words, optimal policy hits ticket queue length empty infinitely many times.

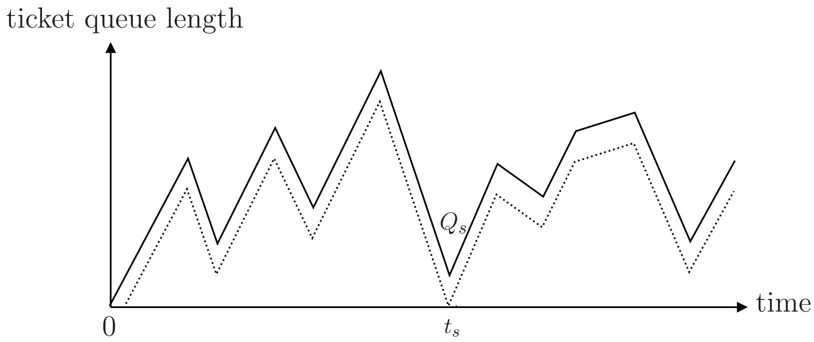


Fig. 3.3: Policy with No Cycle Feature

Consider the sequence of the system given by Section 3.2, a staffing policy sequence $\{\pi_*^n : n \geq 1\}$ is said to be *asymptotically optimal*, if for any feasible policy $\{\pi^n : n \geq 1\}$, we have

$$\lim_{T\to\infty} \lim_{n\to\infty} \mathcal{AC}^n_{\pi^n_*}(T) \leq \lim_{T\to\infty} \lim_{n\to\infty} \mathcal{AC}^n_{\pi^n}(T),$$

where

$$\mathcal{AC}^n_{\pi^n}(T) = \frac{r \times \mathcal{R}^n_{\pi^n}(\tau^n_{\pi^n}(T)) + \mathcal{H}^n_{\pi^n}(T) + \mathcal{O}^n_{\pi^n}(T) + \mathcal{S}^n_{\pi^n}(T)}{nT},$$
$$\mathcal{AC}^n_{\pi^n_*}(T) = \frac{r \times \mathcal{R}^n_{\pi^n_*}(\tau^n_{\pi^n_*}(T)) + \mathcal{H}^n_{\pi^n_*}(T) + \mathcal{O}^n_{\pi^n_*}(T) + \mathcal{S}^n_{\pi^n_*}(T)}{nT}.$$

For the nth system given by Section 3.2, we repeat to use the following policy: the system starts with staffing level i_1^*, the staffing level will be switched from i_1^* to i_2^* when the ticket queue length Q^n reaches to $n\bar{Q}^*$, and the i_2^* staffing level will be used until the system becomes empty, where i_1^*, i_2^* and $\bar{Q}^*$ are given by (3.99) and (3.100). We call this policy as 2-*piece* $(i_1^*, i_2^*, n\bar{Q}^*)$ *policy.* Define

$$\frac{1}{c^*} = \frac{1}{\beta_{i_1^*}} - \frac{1}{\beta_{i_2^*}}, \quad \frac{1}{c^*_\alpha} = \frac{1-\alpha_{i_1^*}}{\beta_{i_1^*}} - \frac{1-\alpha_{i_2^*}}{\beta_{i_2^*}}.$$

Then by Theorem 18 and Proposition 19, we derive our main result

Theorem 20. (Asymptotic Optimality) *Suppose that Assumptions* (3.63) *and* (3.85)-(3.86) *hold. If for the nth system, the* 2-*piece* $(i_1^*, i_2^*, n\bar{Q}^*)$ *is implemented, then the fluid approximation* $\left(\bar{\tau}, \bar{L}, \bar{T}, \bar{Q}\right)$ *given by* (3.44) *in Theorem*

15 *satisfies that*

$$\bar{\tau}(t) = \begin{cases} \frac{\ell\bar{Q}^*}{c^*} + \frac{\mu_{i_1^*}}{\lambda}\Big(t - \frac{\ell\bar{Q}^*}{c^*}\Big), & \textit{for} \ \ t \in [\frac{\ell\bar{Q}^*}{c^*}, \ \ \frac{\ell\bar{Q}^*}{c^*} + \frac{\bar{Q}^*}{\beta_{i_1^*}}), \\ \\ \frac{\ell\bar{Q}^*}{c^*} + \frac{\mu_{i_1^*}\bar{Q}^*}{\lambda\beta_{i_1^*}} + \frac{\mu_{i_2^*}}{\lambda}\Big(t - \frac{\ell\bar{Q}^*}{c^*} - \frac{\bar{Q}^*}{\beta_{i_1^*}}\Big), & \textit{for } t \in [\frac{\ell\bar{Q}^*}{c^*} + \frac{\bar{Q}^*}{\beta_{i_1^*}}, \frac{(\ell+1)\bar{Q}^*}{c^*}); \end{cases} \tag{3.112}$$

$$\bar{L}(t) = \begin{cases} \ell\Big(\frac{\alpha_{i_1^*}\mu_{i_1^*}\bar{Q}^*}{\beta_{i_1^*}} + \frac{\alpha_{i_2^*}\mu_{i_2^*}\bar{Q}^*}{\beta_{i_2^*}}\Big) + \alpha_{i_1^*}\mu_{i_1^*}\Big(t - \frac{\ell\bar{Q}^*}{c^*}\Big), \\ \qquad\qquad \textit{for} \ \ t \in [\frac{\ell\bar{Q}^*}{c^*}, \ \ \frac{\ell\bar{Q}^*}{c^*} + \frac{\bar{Q}^*}{\beta_{i_1^*}}), \\ \\ \ell\Big(\frac{\alpha_{i_1^*}\mu_{i_1^*}\bar{Q}^*}{\beta_{i_1^*}} + \frac{\alpha_{i_2^*}\mu_{i_2^*}\bar{Q}^*}{\beta_{i_2^*}}\Big) + \frac{\alpha_{i_1^*}\mu_{i_1^*}\bar{Q}^*}{\beta_{i_1^*}} + \alpha_{i_2^*}\mu_{i_2^*}\Big(t - \frac{\ell\bar{Q}^*}{c^*} - \frac{\bar{Q}^*}{\beta_{i_1^*}}\Big), \\ \\ \qquad\qquad \textit{for } t \in [\frac{\ell\bar{Q}^*}{c^*} + \frac{\bar{Q}^*}{\beta_{i_1^*}}, \frac{(\ell+1)\bar{Q}^*}{c^*}); \end{cases} \tag{3.113}$$

$$\sum_{\ell=1}^{2}\sum_{j=1}^{m}\bar{T}_{i_\ell^* j}(t) = \begin{cases} \ell\big(\frac{i_1^*\bar{Q}^*}{\beta_{i_1^*}} - \frac{i_2^*\bar{Q}^*}{\beta_{i_2^*}}\big) + i_1^*\Big(t - \frac{\ell\bar{Q}^*}{c^*}\Big), \\ \qquad\qquad \textit{for} \ \ t \in [\frac{\ell\bar{Q}^*}{c^*}, \ \ \frac{\ell\bar{Q}^*}{c^*} + \frac{\bar{Q}^*}{\beta_{i_1^*}}), \\ \\ \ell\big(\frac{i_1^*\bar{Q}^*}{\beta_{i_1^*}} - \frac{i_2^*\bar{Q}^*}{\beta_{i_2^*}}\big) + \frac{i_1^*\bar{Q}^*}{\beta_{i_1^*}} + i_2^*\Big(t - \frac{\ell\bar{Q}^*}{c^*} - \frac{\bar{Q}^*}{\beta_{i_1^*}}\Big), \\ \qquad\qquad \textit{for } t \in [\frac{\ell\bar{Q}^*}{c^*} + \frac{\bar{Q}^*}{\beta_{i_1^*}}, \frac{(\ell+1)\bar{Q}^*}{c^*}); \end{cases} \tag{3.114}$$

and

$$\bar{Q}(t) = \begin{cases} \beta_{i_1^*}\Big(t - \frac{\ell\bar{Q}^*}{c^*}\Big), & \textit{for } \ t \in [\frac{\ell\bar{Q}^*}{c^*}, \quad \frac{\ell\bar{Q}^*}{c^*} + \frac{\bar{Q}^*}{\beta_{i_1^*}}), \\ \\ \bar{Q}^* + \beta_{i_2^*}(t - \frac{\ell\bar{Q}^*}{c^*} - \frac{\bar{Q}^*}{\beta_{i_1^*}}) & \textit{for } t \in [\frac{\ell\bar{Q}^*}{c^*} + \frac{\bar{Q}^*}{\beta_{i_1^*}}, \frac{(\ell+1)\bar{Q}^*}{c^*}). \end{cases} \tag{3.115}$$

Moreover, the long-run average cost incurred by the above fluid model is equal to

$$\frac{(\mu_{i_2^*}\alpha_{i_2^*} + c_{i_2^*})\beta_{i_1^*} - (\mu_{i_1^*}\alpha_{i_1^*} + c_{i_1^*})\beta_{i_2^*}}{\beta_{i_1^*} - \beta_{i_2^*}} + \sqrt{2hK\frac{(c^*)^2}{c_\alpha^*}i_2^*}.$$

Hence, by Proposition 19*, the* 2*-piece* $(i_1^*, i_2^*, n\bar{Q}^*)$ *is an asymptotically optimal policy. In particular, If* $1-\alpha_i = a + bi$ *and assumptions in Corollary* 1 *hold, then* $(m_0, m_0+1, n\bar{Q}^*)$ *is an asymptotically optimal policy.*

3.6 Numerical Studies

In this section, we make extensive numerical experiments to show that the asymptotic policy established in the fluid model performs very well. To make direct comparisons, we compute optimal staffing levels and threshold through both Markov analysis and fluid analysis. For Markov analysis, we use long-run average cost expression, denoted by $\Pi(Q)$, given in Appendix. We use i_1^m, i_2^m, and Q^m to denote the optimal staffing levels and threshold derived through Markov analysis. For fluid analysis, we use formulas (3.99) and (3.100) to derive the optimal staffing levels i_1^*, i_2^*, and the optimal threshold

$\bar{Q}^*$. Assume $1 - \alpha_i = a + bi$ and $c_i = d \times i^2$. First we consider the situation $\alpha_{i_1} = \alpha_{i_2}$.

3.6.1 Same α_{i_1} and α_{i_2}

Case I: In table 3.1 - table3.2, we change operating cost c_i. In table 3.1, we choose $(\lambda, \mu, h, K, a, b, m)$=(40,10,2,25,0.85,0,7); in table 3.2, we choose $(\lambda, \mu, h, K, a, b, m)$=(40,10,2,25,0.45,0,7).

Tab. 3.1: Markov vs. Fulid: I(a)

d	m_0	ρ_{m_0}	ρ_{m_0+1}	Markov				Fluid			
				i_1^m	i_2^m	Q^m	$\Pi(Q^m)$	i_1^*	i_2^*	$\bar{Q}^*$	$\Pi(\bar{Q}^*)$
0.0025	3	1.13	0.85	2	4	17	35.94	3	4	18	43.74
0.25	3	1.13	0.85	2	4	17	39.46	3	4	18	47.20
25	3	1.13	0.85	3	4	46	369.10	3	4	18	393.39

Tab. 3.2: Markov vs. Fulid: I(b)

d	m_0	ρ_{m_0}	ρ_{m_0+1}	Markov				Fluid			
				i_1^m	i_2^m	Q^m	$\Pi(Q^m)$	i_1^*	i_2^*	$\bar{Q}^*$	$\Pi(\bar{Q}^*)$
0.0025	1	1.8	0.9	1	2	16	49.44	1	2	20	49.65
0.25	1	1.8	0.9	1	2	16	50.33	1	2	20	50.54
25	1	1.8	0.9	1	2	22	139.47	1	2	20	139.54
200	1	1.8	0.9	1	2	47	756.63	1	2	20	768.84

Case II: In table 3.3 - table 3.5, we change holding cost h. In table 3.3, we choose $(\lambda, \mu, d, K, a, b, m) = (40, 10, 0.5, 25, 0.6, 0, 7)$; in table 3.3, we choose $(\lambda, \mu, d, K, a, b, m) = (48, 10, 0.5, 25, 0.85, 0, 7)$; in table 3.3, we choose $(\lambda, \mu, d, K, a, b, m) = (50, 15, 0.5, 25, 0.45, 0, 7)$.

Tab. 3.3: Markov vs. Fulid: II(a)

h	m_0	ρ_{m_0}	ρ_{m_0+1}	Markov				Fluid			
				i_1^m	i_2^m	Q^m	$\Pi(Q^m)$	i_1^*	i_2^*	$\bar{Q}^*$	$\Pi(\bar{Q}^*)$
0.02	2	1.2	0.8	2	3	228	21.67	2	3	224	21.67
0.2	2	1.2	0.8	2	3	72	27.08	2	3	71	27.08
2	2	1.2	0.8	2	3	24	43.31	2	3	22	43.35
20	2	1.2	0.8	2	4	10	86.17	2	3	7	86.43

Tab. 3.4: Markov vs. Fulid: II(b)

h	m_0	ρ_{m_0}	ρ_{m_0+1}	Markov				Fluid			
				i_1^m	i_2^m	Q^m	$\Pi(Q^m)$	i_1^*	i_2^*	$\bar{Q}^*$	$\Pi(\bar{Q}^*)$
0.02	4	1.02	0.82	4	5	273	19.30	4	5	113	21.55
0.2	4	1.02	0.82	4	5	101	27.59	4	5	36	46.10
2	4	1.02	0.82	2	5	18	46.42	4	5	11	83.47
20	4	1.02	0.82	2	6	7	80.78	4	5	4	97.29

Tab. 3.5: Markov vs. Fulid: II(c)

h	m_0	ρ_{m_0}	ρ_{m_0+1}	Markov				Fluid			
				i_1^m	i_2^m	Q^m	$\Pi(Q^m)$	i_1^*	i_2^*	$\bar{Q}^*$	$\Pi(\bar{Q}^*)$
0.02	1	1.5	0.75	1	2	299	31.44	1	2	304	31.44
0.2	1	1.5	0.75	1	2	94	37.18	1	2	96	37.18
2	1	1.5	0.75	1	2	30	55.28	1	2	30	55.28
20	1	1.5	0.75	1	3	14	118.40	1	2	10	122.26

Case III: In table 3.6 - table 3.8, we change λ and μ while keeping λ/μ constant, and customer delay cost is smaller than server operation cost. In table 3.6, we choose $(h, d, K, a, b, m) = (0.2, 0.5, 25, 0.45, 0, 7)$; in table 3.7, we choose $(h, d, K, a, b, m) = (0.2, 0.5, 25, 0.65, 0, 8)$; in table 3.8, we choose $(h, d, K, a, b, m) = (0.2, 0.5, 25, 0.95, 0, 8)$.

Tab. 3.6: Markov vs. Fulid: III(a)

λ	μ	m_0	ρ_{m_0}	ρ_{m_0+1}	Markov				Fluid			
					i_1^m	i_2^m	Q^m	$\Pi(Q^m)$	i_1^*	i_2^*	$\bar{Q}^*$	$\Pi(\bar{Q}^*)$
5	1.5	1	1.5	0.75	1	2	31	6.78	1	2	30	6.77
50	15	1	1.5	0.75	1	2	94	37.18	1	2	96	37.18
500	150	1	1.5	0.75	1	2	298	303.03	1	2	304	303.04
5000	1500	1	1.5	0.75	1	2	945	2836.25	1	2	962	2836.30

Tab. 3.7: Markov vs. Fulid: III(b)

λ	μ	m_0	ρ_{m_0}	ρ_{m_0+1}	Markov				Fluid			
					i_1^m	i_2^m	Q^m	$\Pi(Q^m)$	i_1^*	i_2^*	$\bar{Q}^*$	$\Pi(\bar{Q}^*)$
8.8	1.5	3	1.27	0.95	2	4	14	15.47	3	4	23	15.64
88	15	3	1.27	0.95	2	4	45	47.24	3	4	73	48.26
880	150	3	1.27	0.95	2	4	172	344.31	3	4	232	347.26
8800	1500	3	1.27	0.95	3	4	798	3193.80	3	4	734	3194.10

Tab. 3.8: Markov vs. Fulid: III(c)

λ	μ	m_0	ρ_{m_0}	ρ_{m_0+1}	Markov				Fluid			
					i_1^m	i_2^m	Q^m	$\Pi(Q^m)$	i_1^*	i_2^*	$\bar{Q}^*$	$\Pi(\bar{Q}^*)$
8	1.5	5	1.01	0.84	5	6	51	19.32	5	6	12	21.95
80	15	5	1.01	0.84	5	6	113	31.26	5	6	39	40.85
800	150	5	1.01	0.84	5	6	290	95.43	5	6	125	119.17
8000	1500	5	1.01	0.84	5	6	772	544.68	5	6	394	585.92

Case IV: In table 3.9 - table 3.11, we also change λ and μ while keeping λ/μ constant, but customer delay cost is larger than server operating cost. In table 3.9, we choose $(h, d, K, a, b, m) = (2, 0.025, 25, 0.8, 0, 8)$; in table 3.10, we choose $(h, d, K, a, b, m) = (2, 0.025, 25, 0.8, 0, 8)$; in table 3.11, we choose $(h, d, K, a, b, m) = (2, 0.025, 25, 0.96, 0, 8)$.

Tab. 3.9: Markov vs. Fulid: IV(a)

λ	μ	m_0	ρ_{m_0}	ρ_{m_0+1}	Markov				Fluid			
					i_1^m	i_2^m	Q^m	$\Pi(Q^m)$	i_1^*	i_2^*	$\bar{Q}^*$	$\Pi(\bar{Q}^*)$
7.5	3.5	1	1.71	0.86	1	3	13	19.10	1	2	7	19.12
75	35	1	1.71	0.86	1	2	23	57.16	1	2	24	57.20
750	350	1	1.71	0.86	1	2	74	273.07	1	2	75	273.09
7500	3500	1	1.71	0.86	1	2	234	1880.10	1	2	236	1880.10

Tab. 3.10: Markov vs. Fulid: IV(b)

λ	μ	m_0	ρ_{m_0}	ρ_{m_0+1}	Markov				Fluid			
					i_1^m	i_2^m	Q^m	$\Pi(Q^m)$	i_1^*	i_2^*	$\bar{Q}^*$	$\Pi(\bar{Q}^*)$
5	1.5	2	1.33	0.89	2	4	10	11.23	2	3	6	18.17
50	15	2	1.33	0.89	2	3	18	44.07	2	3	20	44.19
500	150	2	1.33	0.89	2	3	59	199.62	2	3	63	199.79
5000	1500	2	1.33	0.89	2	3	194	1314.60	2	3	198	1314.70

Tab. 3.11: Markov vs. Fulid: IV(c)

λ	μ	m_0	ρ_{m_0}	ρ_{m_0+1}	Markov				Fluid			
					i_1^m	i_2^m	Q^m	$\Pi(Q^m)$	i_1^*	i_2^*	$\bar{Q}^*$	$\Pi(\bar{Q}^*)$
6.5	2	3	1.04	0.78	2	4	9	13.91	3	4	5	20.44
65	20	3	1.04	0.78	2	4	29	54.78	3	4	15	107.47
650	200	3	1.04	0.78	2	4	99	212.20	3	4	48	377.70
6500	2000	3	1.04	0.78	3	4	359	900.50	3	4	151	1222.20

3.6.2 Different α_{i_1} and α_{i_2}

Here we consider $\alpha_{i_1} \neq \alpha_{i_2}$. In table 3.12, we choose (h, d, K, a, b, m)=(0.05,1, 2,0.45,0.005,6); in table 3.13, we choose (h, d, K, a, b, m)=(0.05,1,2,0.7,0.005,6);

in table 3.14, we choose (h, d, K, a, b, m)=(0.05,1,2,0.25,0.005,8).

Tab. 3.12: Markov vs. Fulid: V(a)

λ	μ	m_0	ρ_{m_0}	ρ_{m_0+1}	Markov				Fluid			
					i_1^m	i_2^m	Q^m	$\Pi(Q^m)$	i_1^*	i_2^*	$\bar{Q}^*$	$\Pi(\bar{Q}^*)$
40	10	1	1.82	0.92	1	2	28	26.2	1	2	32	26.2
400	100	1	1.82	0.92	1	2	79	222.1	1	2	101	222.2
400	180	1	1.01	0.51	1	2	34	214.1	1	2	39	214.3
400	70	2	1.31	0.89	2	3	116	214.8	2	3	132	224.8

Tab. 3.13: Markov vs. Fulid: V(b)

λ	μ	m_0	ρ_{m_0}	ρ_{m_0+1}	Markov				Fluid			
					i_1^m	i_2^m	Q^m	$\Pi(Q^m)$	i_1^*	i_2^*	$\bar{Q}^*$	$\Pi(\bar{Q}^*)$
40	10	2	1.42	0.95	2	3	26	21.4	2	3	24	21.4
40	15	1	1.88	0.94	1	2	19	16.7	1	2	21	16.7
400	100	2	1.42	0.95	2	3	67	125.7	2	3	75	125.7
400	180	1	1.57	0.79	1	2	105	123.3	1	2	118	123.3
400	80	3	1.19	0.90	2	4	109	130.5	3	4	108	130.6
400	50	5	1.16	0.97	5	6	60	145.9	5	6	78	146.0

Tab. 3.14: Markov vs. Fulid: V(c)

λ	μ	m_0	ρ_{m_0}	ρ_{m_0+1}	Markov				Fluid			
					i_1^m	i_2^m	Q^m	$\Pi(Q^m)$	i_1^*	i_2^*	$\bar{Q}^*$	$\Pi(\bar{Q}^*)$
50	10	1	1.28	0.65	1	2	68	39.8	1	2	69	39.8
400	100	1	1.02	0.52	1	2	46	294.5	1	2	69	295.1
400	50	2	1.04	0.71	2	3	205	302.1	2	3	114	302.6
400	20	5	1.10	0.93	5	6	327	324.7	5	6	166	324.7

In summary, we find that fluid model performs well when μ and λ is large.

3.7 Concluding Remarks

In this chapter we study the optimal staffing policy for a ticket queue system with multiple staffing levels. We build a fluid model for the ticket system, and show that, changing staffing level once in each cycle is better than changing staffing level multiple times. Besides, the threshold to change staffing level is determined through the EOQ formula. Finally, we prove the above policy is asymptotical optimal.

4. FUTURE RESEARCH

There are several directions for follow-up research:

- **Incorporating the estimation of customer abandonments** . One candidate for follow-up research is what we alluded to in the Introduction: incorporating the estimation of customer abandonment rates into the staffing decision. Start off with initially assumed server-dependent abandonment rates, run the optimal staffing rule based on these rates, just like what we have done here. At the end of several cycles (the length of which has to do with the trade-off between learning and control), update the abandonment statistic (e.g., do a Bayesian update), and then repeat, until convergence (need to be established/justified).

- **Provide some information to customers**. Another aspect that we didn't mention in this study is: whether and when to provide some information to customers? In other words, when should the service provider make a delay announcement? And if so, what to announce? In the literature of delay announcement, there are two types of announcements. The first type of announcement is to be made upon customer arrival, and often an estimated duration of delay is announced, see Armony et al. [5]. The second type of announcement is to be made during customer waiting, and various levels of information will be given, such as the customer's waiting time or the customer's current position in the queue, see Allon and Bassamboo [1] and Mandelbaum and Zeltyn [26].

With more information, customers may change their decision about staying and abandoning, which will consequently affect the abandonment rate α_i. That is, α_i not only depends on the number of open servers, but also depends on other available information. The question is how to quantify the impact of additional information on customer decisions and the system performance measures. By incorporating those information, we need to find a way to modify our model in this more general setting.

APPENDIX

A. APPENDIX

In this appendix, we study a ticket queueing system with staffing policy (i_1, i_2, Q) with $i_1 < i_2$. The policy works this way: the system starts with staffing level i_1, the staffing level will be switched from i_1 to i_2 when the ticket queue length reaches to Q, and the i_2 staffing level will be used until the system becomes empty. Arrival process is Poisson process, and the service time follows exponential distribution with rate μ.

Using the idea in Chapter 2, we can derive the performance measure $\mathsf{E}T_1$, C_1, $\mathsf{E}T_2$, C_2, and further derive the long-run average cost expression, denoted $\Pi(Q)$. However, the expressions will become much more complex, because the transition matrix becomes more complex than before. In the numerical study of Chapter 3, we use $\Pi(Q)$ to find the optimal i_1, i_2, and Q, which are compared with the solution derived through fluid model.

We will use a new definition of cycle in this appendix: each cycle is the time duration between two consecutive entry to system empty after servers finish serving some customers. The definition of cycle will not affect the long-run average cost, but using this definition will slightly simplify our calculation here.

In the following, we computer $\mathsf{E}T_1$ and C_1 in A.1, and computer $\mathsf{E}T_2$ and C_2 in A.2. Then we will get the long-run average cost expression in A.3.

The notations we will use in this appendix include:

$$
\begin{aligned}
&\theta_{i_1} = \alpha_{i_1} + \frac{i_1\mu}{\lambda}, \quad \rho = \frac{\lambda}{\mu}, \\
&\beta_1 = \lambda - \frac{i_1\mu}{1-\alpha_{i_1}}, \quad \beta_2 = \frac{i_2\mu}{1-\alpha_{i_2}} - \lambda, \\
&\mu_1 = \frac{i_1\mu}{1-\alpha_{i_1}}, \quad \mu_2 = \frac{i_2\mu}{1-\alpha_{i_2}}.
\end{aligned}
$$

A.1 $\mathsf{E}T_1$ and C_1

In period T_1, i_1 servers are working. When the ticket queue length reaches Q, we add $i_2 - i_1$ servers. Before ticket queue length reaches either Q or 0, the transition rate matrix $\mathbf{D}_1$is

$$
\mathbf{D}_1 = \begin{bmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{13} & \mathbf{D}_{14} \end{bmatrix}.
$$

$\mathbf{D}_{11}$ is a $(i_1-2)\times(i_1-2)$ square matrix and $\mathbf{D}_{14}$ is a $(Q-i_1+1)\times(Q-i_1+1)$ square matrix.

$$
\mathbf{D}_{11} = \begin{bmatrix}
-(\lambda+\mu) & \lambda & & & \\
2\mu & -(\lambda+2\mu) & \lambda & & \\
& 3\mu & -(\lambda+3\mu) & \lambda & \\
& & \ddots & \ddots & \ddots \\
& & & (i_1-2)\mu & -(\lambda+(i_1-2)\mu)
\end{bmatrix},
$$

and

$$\mathbf{D}_{14} = \begin{bmatrix} -(\lambda+(i_1-1)\mu) & \lambda & & & & \\ i_1\mu & -(\lambda+i_1\mu) & \lambda & & & \\ i_1\mu\alpha_{i_1} & i_1\mu(1-\alpha_{i_1}) & -(\lambda+i_1\mu) & \lambda & & \\ \vdots & \vdots & \vdots & \ddots & \ddots & \\ i_1\mu\alpha_{i_1}^{Q-i_1-1} & i_1\mu\alpha_{i_1}^{Q-i_1-2}(1-\alpha_{i_1}) & i_1\mu\alpha^{Q-i_1-3}(1-\alpha_{i_1}) & \cdots & \cdots & -(\lambda+i_1\mu) \end{bmatrix}.$$

$\mathbf{D}_{12}$ is a $(i_1-2)\times(Q-i_1+1)$ matrix with only one nonzero element. $\mathbf{D}_{13}$ is a $(Q-i_1+1)\times(i_1-2)$ matrix with only one nonzero element.

$$\mathbf{D}_{12} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \\ \lambda & 0 & \cdots & 0 \end{bmatrix},$$

and

$$\mathbf{D}_{13} = \begin{bmatrix} 0 & \cdots & 0 & (i_1-1)\mu \\ 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Denote

- T_{11}: starting at 1, time duration of reaching either i_1-1 or 0;
- $\hat{T}_{11}$: starting at i_1-2, time duration of reaching either i_1-1 or 0;
- T_{14}: starting at i_1-1, time duration of reaching either i_1-2 or Q;

- π_1: starting at 1, probability of reaching $i_1 - 1$ before 0;
- π_2: starting at $i_1 - 1$, probability of reaching $i_1 - 2$ before Q;
- π_3: starting at $i_1 - 2$, probability of reaching $i_1 - 1$ before 0.

Period T_1 is the time of i_1-server region (here T_1 doesn't include idle time). Then T_1 can be written as

$$\begin{aligned}\mathsf{E}T_1 =& \mathsf{E}T_{11} + \mathsf{E}T_{14}\pi_1 \sum_{j=0}^{\infty} (\pi_2\pi_3)^j + \mathsf{E}\hat{T}_{11}\pi_1\pi_2 \sum_{j=0}^{\infty} (\pi_2\pi_3)^j \\ =& \mathsf{E}T_{11} + \mathsf{E}T_{14}\frac{\pi_1}{1-\pi_2\pi_3} + \mathsf{E}\hat{T}_{11}\frac{\pi_1\pi_2}{1-\pi_2\pi_3}. \end{aligned} \tag{A.1}$$

To calculate $\mathsf{E}T_{11}$ and $\mathsf{E}\hat{T}_{11}$, it suffices to know the inverse matrix of $\mathbf{D}_{11}$, which is denoted by

$$(\mathbf{D}_{11})^{-1} = (\bar{d}_{ij})_{(i_1-2)\times(i_1-2)}.$$

We have

$$\bar{d}_{ij} = \begin{cases} -\frac{i!v(i)}{\mu\rho^{i-j}j!} + \frac{i!v(i)v(j)}{\mu\rho^{i-1}(1+v(1))} & i \geq j, \\ -\frac{v(j)}{\mu} + \frac{i!v(i)v(j)}{\mu\rho^{i-1}(1+v(1))} & i < j, \end{cases}$$

where $v(i) = \sum_{k=i}^{i_1-2} \frac{k!}{i!\rho^{k-i+1}}$. By the definition of $\mathsf{E}T_{11}$ and $\mathsf{E}\hat{T}_{11}$, we have

$$\begin{aligned}\mathsf{E}T_{11} &= (1,0,\ldots,0)(-\mathbf{D}_{11}^{-1})\mathbf{e}' \\ &= \sum_{k=1}^{i_1-2} \frac{v(k)}{\mu(1+v(1))}, \end{aligned} \tag{A.2}$$

and

$$
\begin{aligned}
\mathsf{E}\hat{T}_{11} &= (0,\ldots,0,1)(-\mathbf{D}_{11}^{-1})\mathbf{e}' \\
&= \frac{(i_1-2)!}{\mu\rho^{i_1-1}}\sum_{k=1}^{i_1-2}\frac{\rho^k}{k!} - \frac{(i_1-2)!}{\mu\rho^{i_1-2}(1+v(1))}\sum_{k=1}^{i_1-2}v(k).
\end{aligned} \tag{A.3}
$$

Similarly, it suffices to know the inverse matrix of $\mathbf{D}_{14}$ to derive $\mathsf{E}T_{14}$, which is denoted by

$$(\mathbf{D}_{14})^{-1} = (d_{ij})_{(Q-i_1+1)\times(Q-i_1+1)}.$$

We have

$$
d_{ij} = \begin{cases}
B_j, & i=1, \text{ and } j=1,\ldots,Q-i_1+1,\\
(-c_3+c_1B_j)\sum_{k=0}^{Q-i_1+1-i}c_2^k, & i>1, \text{ and } j<i,\\
-\left(c_3\sum_{k=1}^{Q-i_1+1-j}c_2^k+c_4\right)c_2^{j-i}+c_1B_j\sum_{k=0}^{Q-i_1+1-i}c_2^k, & i>1, \text{ and } j\ge i,
\end{cases}
$$

where

$$
\begin{aligned}
&c_1 = \frac{-(1-\alpha_{i_1})\lambda+\alpha_{i_1}(i_1-1)\mu+\mu}{\lambda\alpha_{i_1}+i_1\mu},\ c_2=\frac{\lambda}{\lambda\alpha_{i_1}+i_1\mu},\\
&c_3 = \frac{1-\alpha_{i_1}}{\lambda\alpha_{i_1}+i_1\mu},\ c_4=\frac{1}{\lambda\alpha_{i_1}+i_1\mu},\\
&B_j = \frac{\left(c_4+c_3\sum_{k=1}^{Q-i_1+1-j}c_2^k\right)c_2^{j-1}}{c_1\sum_{k=1}^{Q-i_1}c_2^k-(\lambda+(i_1-1)\mu)\,c_4},\ j=1,\ldots,Q-i_1+1.
\end{aligned}
$$

By definition of $\mathsf{E}T_{14}$, we have

$$\begin{aligned}\mathsf{E}T_{14} =&(1,0,\dots,0)(-\mathbf{D}_{14}^{-1})\mathbf{e}'\\ =&\frac{1}{(1-\alpha_{i_1})\beta_1}\cdot\frac{\theta_{i_1}^{Q-i_1+1}-1+\frac{\beta_1}{\mu_1}(1-\alpha_{i_1})(Q-i_1+1)}{-\frac{i_1-1}{\rho}\theta_{i_1}^{Q-i_1}+\frac{(i_1-1)(1-\alpha_{i_1})}{i_1}+\frac{\beta_1}{\mu_1}}\end{aligned}\tag{A.4}$$

where ρ, θ_{i_1}, β_1 and μ_1 are defines at the beginning of the note.

Consider the embedded markov chain, we derive the following probabilities

$$\pi_1 =\frac{1}{1+v(1)},\tag{A.5}$$

$$\pi_2 =\frac{(i_1-1)\theta_{i_1}^{Q-i_1}-\frac{\lambda(i_1-1)}{\mu_1}}{(i_1-1)\theta_{i_1}^{Q-i_1}-\rho\big[\frac{\beta_1}{\mu_1}+\frac{(1-\alpha_{i_1})(i_1-1)}{i_1}\big]}\tag{A.6}$$

$$\pi_3 =\frac{1+\sum_{k=1}^{i_1-3}\frac{k!}{\rho^k}}{1+v(1)}.\tag{A.7}$$

Plugging (A.2)-(A.7) into (A.1), we could derive the expression of $\mathsf{E}T_1$. Note that when $i_1=2$ and $i_1=1$, $\mathsf{E}T_{11}=\mathsf{E}\hat{T}_{11}=0$ and $\pi_1=\pi_3=1$.

All the above approach applies to $i_1\geq 1$. But we should notice that, when $i_1\geq 2$, T_1 doesn't include idle time; when $i_1=1$, T_1 includes idle time. We delay the detailed discussion of special case $i_1=1$ to the long-run average cost section.

In i_1-server region, system incurs delay cost only when ticket queue length exceed i_1. That is, only over T_{14} delay cost is incurred. Denote delay cost in i_1-server region by C_1, delay cost over T_{14} by C_{14}, then we have

$$\begin{aligned}C_{14} =&(1,0,\ldots,0)(-\mathbf{D}_{14})^{-1}(0,0,1,2,\ldots,N-i_1-1)'(1-\alpha_{i_1})h\\ =&h\mu_1\frac{\frac{2\rho}{i_1}(\theta_{i_1}^{Q-i_1}-1)-\frac{\beta_1^2}{\mu_1^2}(1-\alpha_{i_1})(Q-i_1)^2+\frac{\beta_1}{\mu_1}\big[2+\frac{\beta_1}{\mu_1}(1-\alpha_{i_1})\big](Q-i_1)}{2\beta_1^2\big(\frac{i_1-1}{\rho}\theta_{i_1}^{Q-i_1}-\frac{\beta_1}{\mu_1}-\frac{(i_1-1)(1-\alpha_{i_1})}{i_1}\big)}\end{aligned} \tag{A.8}$$

and

$$C_1=\frac{\pi_1}{1-\pi_2\pi_3}C_{14}. \tag{A.9}$$

The probability of reaching Q can be written as

$$\begin{aligned}\pi =&\pi_1(1-\pi_2)\sum_{k=0}^{\infty}(\pi_2\pi_3)^k=\frac{\pi_1(1-\pi_2)}{1-\pi_2\pi_3}\\ =&\pi_1\cdot\frac{\frac{\lambda(i_1-1)}{\mu_1}-\rho\big[\frac{\beta_1}{\mu_1}+\frac{(i_1-1)(1-\alpha_{i_1})}{i_1}\big]}{(i_1-1)(1-\pi_3)\theta_{i_1}^{Q-i_1}-\rho\big[\frac{\beta_1}{\mu_1}+\frac{(i_1-1)(1-\alpha_{i_1})}{i_1}\big]+\frac{\lambda(i_1-1)}{\mu_1}\pi_3}.\end{aligned} \tag{A.10}$$

Based on (A.1)-(A.10), we have

$$\begin{aligned}\frac{\mathsf{E}T_1}{\pi}=&-\frac{\mu_1}{\rho\beta_1}\Big\{\theta_{i_1}^{Q-i_1}\Big[\frac{(i_1-1)(1-\pi_3)\mathsf{E}T_{11}}{\pi_1}-\frac{\rho\alpha_{i_1}+i_1}{(1-\alpha_{i_1})\beta_1}+(i_1-1)\mathsf{E}\hat{T}_{11}\Big]\\ &-(Q-i_1+1)\frac{\rho}{\mu_1}+\frac{\mathsf{E}T_{11}}{\pi_1}\Big[-\rho\big(\frac{\beta_1}{\mu_1}+\frac{(1-\alpha_{i_1})(i_1-1)}{i_1}\big)\\ &+\frac{\lambda(i_1-1)\pi_3}{\mu_1}\Big]+\frac{\rho}{(1-\alpha_{i_1})\beta_1}-\frac{\lambda(i_1-1)}{\mu_1}\mathsf{E}\hat{T}_{11}\Big\}\end{aligned}$$

$$\frac{C_1}{\pi}=-\frac{h\mu_1^2}{\beta_1^3}\Big\{\frac{\rho}{i_1}(\theta_{i_1}^{Q-i_1}-1)-\frac{\beta_1^2}{2\mu_1^2}(1-\alpha_{i_1})(Q-i_1)^2$$

$$+ \frac{\beta_1}{2\mu_1}\big[2 + \frac{\beta_1}{\mu_1}(1-\alpha_{i_1})\big](Q-i_1)\Big\}$$

A.2 $\mathsf{E}T_2$ and C_2

Now we calculate $\mathsf{E}T_2$ and delay cost C_2. After reaching Q, we add $i_2 - i_1$ servers and assign $Q-i_1$ tickets to these i_2-i_1 servers. Among $Q-i_1$ tickets, let Z be the real customers and it follows distribution

$$\mathsf{Pr}(Z=k) = \binom{Q-i_1}{k}(1-\alpha_{i_2})^k \alpha_{i_2}^{Q-i_1-k}, \quad k = 0,\ldots,Q-i_1. \tag{A.11}$$

Let τ_k be the first passage time from k to $k-1$, we have

$$\mathsf{E}\tau_k = \mathsf{E}\tau_{i_2} = \frac{1}{i_2\mu - \lambda(1-\alpha_{i_2})}, \quad \text{for any } k \geq i_2; \tag{A.12}$$

$$\begin{aligned}\mathsf{E}\tau_k &= \frac{1}{k\mu} + \frac{\lambda}{k\mu}\mathsf{E}\tau_{k+1} \\ &= \sum_{j=0}^{i_2-1-k}\frac{(k-1)!\rho^j}{\mu(k+j)!} + \frac{(k-1)!\rho^{i_2-k}}{(i_2-1)!}\frac{1}{i_2\mu-\lambda(1-\alpha_{i_2})}, \quad k = 1,2,\ldots,i_2-1.\end{aligned} \tag{A.13}$$

Therefore we write $\mathsf{E}T_2$ as

$$\begin{aligned}\mathsf{E}T_2 &= \mathsf{E}\sum_{j=1}^{i_1+Z}\tau_j, \\ &= \sum_{j=1}^{i_1}\mathsf{E}\tau_j + \sum_{j=1}^{i_2-i_1-1}\mathsf{Pr}(Z \geq j)\mathsf{E}\tau_{i_1+j} \\ &\quad + \mathsf{E}\tau_{i_2}\sum_{j=i_2-i_1}^{Q-i_1}\mathsf{Pr}(Z=j)(j-i_2+i_1+1).\end{aligned} \tag{A.14}$$

Plug (A.11) into (A.14) we derive

$$
\begin{aligned}
\mathsf{E}T_2 =& \frac{Q-i_1}{\beta_2} + \sum_{j=1}^{i_2-1} \mathsf{E}\tau_j - \frac{i_2-i_1-1}{(1-\alpha_{i_2})\beta_2} \\
& - \alpha_{i_2}^{Q} \sum_{k=0}^{i_2-i_1-2} \frac{(Q-i_1)!}{k!(Q-i_1-k)!}(1-\alpha_{i_2})^k \alpha_{i_2}^{-i_1-k}\Big[\sum_{j=i_1+1+k}^{i_2-1} \mathsf{E}\tau_j \\
& - \frac{i_2-i_1-(k+1)}{(1-\alpha_{i_2})\beta_2}\Big].
\end{aligned}
$$

To derive C_2, we decompose it into two parts:

$$
C_2 = C_{21} + C_{22}. \tag{A.15}
$$

C_{21} is the delay cost incurred by initial $Q - i_1$ tickets and C_{22} is delay cost incurred by new arrival in i_2-server region.

$$
\begin{aligned}
C_{21} =& \frac{h}{i_2\mu} \sum_{k=i_2-i_1+1}^{Q-i_1} \mathsf{Pr}(Z=k) \sum_{j=1}^{k-i_2+i_1} j \\
=& \frac{h}{2i_2\mu}\Big\{(Q-i_1)^2(1-\alpha_{i_2})^2 + (Q-i_1)(1-\alpha_{i_2})\big[-(1-\alpha_{i_2}) \\
& \qquad - 2(i_2-i_1-1))\big] + (i_2-i_1)(i_2-i_1-1) \\
& \qquad + \alpha_{i_2}^{Q} \sum_{k=0}^{i_2-i_1-2} \frac{(Q-i_1)!}{(Q-i_1-k)!k!}(1-\alpha_{i_2})^k \alpha_{i_2}^{-i_1-k}\big[-k(k-1) \\
& \qquad + (2k-i_2+i_1)(i_2-i_1-1)\big]\Big\}.
\end{aligned} \tag{A.16}
$$

We represent $\mathsf{E}T_2$ as

$$\mathsf{E}T_2 = \sum_{j=1}^{i_1} \mathsf{E}\tau_j + \sum_{j=1}^{i_2-i_1-1} \mathsf{E}\tau_{i_1+j}\mathsf{Pr}(Z \geq j) + \mathsf{E}T_{21},$$

where

$$\mathsf{E}T_{21} = \mathsf{E}\tau_{i_2} \sum_{j=i_2-i_1}^{Q-i_1} \mathsf{Pr}(Z = j)(j - i_2 + i_1 + 1).$$

Then we have

$$\begin{aligned} C_{22} =& h\mathsf{E}T_2 \times \lambda(1-\alpha_{i_2}) \left[\sum_{j=1}^{i_1} \mathsf{E}(W|\text{arriving during } \tau_j)\frac{\mathsf{E}\tau_j}{\mathsf{E}T_2} \right. \\ & + \sum_{j=1}^{i_2-i_1-1} \mathsf{E}(W|\text{arriving during } \tau_{i_1+j})\frac{\mathsf{E}\tau_{i_1+j}\mathsf{Pr}(Z \geq j)}{\mathsf{E}T_2} \\ & \left. + \mathsf{E}(W|\text{arriving during } T_{21})\frac{\mathsf{E}T_{21}}{\mathsf{E}T_2} \right]. \end{aligned}$$

Since we know

$$\begin{aligned} \mathsf{E}(W|\text{arriving during } \tau_k) &= \frac{\lambda}{k\mu}\frac{\mathsf{E}\tau_{k+1}}{\mathsf{E}\tau_k}\mathsf{E}(W|\text{arriving during } \tau_{k+1}), \\ & \qquad\qquad k = 1, \ldots, i_2 - 1, \\ \mathsf{E}(W|\text{arriving during } \tau_{i_2}) &= \frac{1}{i_2\mu - \lambda(1-\alpha_{i_2})}, \end{aligned}$$

we only need to calculate $\mathsf{E}(W|\text{arriving during } T_{21})$. Introducing delay T_{20}, which is the service time of real customers among Z tickets given that $i_2 - 1$ servers are used. Let X_1 be the exponentially distributed service time with mean $1/(i_2\mu)$, then

$$T_{20} = \sum_{k=0}^{Q-i_2} I_{\{Z=i_2-i_1+k\}}(k+1)X_1.$$

Therefore we rewrite T_{21} as

$$T_{21} = \min[t : (i_2-1) \text{ customers in system when delay } T_{20} \text{ commences at },$$
$$\text{time } 0^+ (i_2-1) \text{ customers in system at time } t, \text{ where } t \geq T_{20}].$$

By Theorem 1 of Omahen and Marathe (1978),

$$\mathsf{E}(W|\text{arriving during } T_{21}) = \frac{\lambda(1-\alpha_{i_2})}{i_2\mu(i_2\mu - \lambda(1-\alpha_{i_2}))} + \frac{\mathsf{E}T_{20}^2}{2\mathsf{E}T_{20}}. \quad \text{(A.17)}$$

The laplace-Stieltjes Transform of T_{20} is

$$\mathsf{E}e^{-sT_{20}} = \sum_{j=i_2-i_1}^{Q-i_1} \binom{Q-i_1}{j} \alpha_{i_2}^{Q-i_1-j}(1-\alpha_{i_2})^j \left(\frac{i_2\mu}{i_2\mu+s}\right)^{j-i_2+i_1+1}. \quad \text{(A.18)}$$

This implies

$$\mathsf{E}T_{20} = \frac{1}{i_2\mu} \sum_{j=i_2-i_1}^{Q-i_1} \frac{(Q-i_1)!}{j!(Q-i_1-j)!} \alpha_{i_2}^{Q-i_1-j}(1-\alpha_{i_2})^j (j-i_2+i_1+1), \quad \text{(A.19)}$$

$$\mathsf{E}T_{20}^2 = \frac{1}{(i_2\mu)^2} \sum_{j=i_2-i_1}^{Q-i_1} \frac{(Q-i_1)!}{j!(Q-i_1-j)!} \alpha_{i_2}^{Q-i_1-j}(1-\alpha_{i_2})^j (j-i_2+i_1+1)$$
$$\times (j-i_2+i_1+2).$$

(A.20)

Plug (A.19) and (A.20) into (A.17) we can get $\mathsf{E}(W|\text{arriving during } T_{21})$. Finally, we use

$$\mathsf{E}(W|\text{arriving during } \tau_j) = \mathsf{E}(W|\tau_j)$$

to simplify our notaion and write C_{22} as

$$\begin{aligned} C_{22} =& h\lambda(1-\alpha_{i_2})\Big\{\sum_{j=1}^{i_2-1}\mathsf{E}(W|\tau_j)\mathsf{E}\tau_j + \frac{(Q-i_1)^2}{2\beta_2\mu_2} + \frac{Q-i_1}{(1-\alpha_{i_2})\beta_2\mu_2}\big[-\frac{1-\alpha_{i_2}}{2} \\ &+ \frac{\lambda}{\beta_2} - (i_2-i_1-2)\big] + \frac{i_2-i_1-1}{(1-\alpha_{i_2})\beta_2 i_2\mu}\big(-\frac{\lambda}{\beta_2} + \frac{i_2-i_1-2}{2}\big) \\ &+ \alpha_{i_2}^Q \sum_{k=0}^{i_2-i_1-2}\frac{(Q-i_1)!}{(Q-i_1-k)!k!}\alpha_{i_2}^{-i_1-k}(1-\alpha_{i_2})^k O_k\Big\}, \end{aligned}$$

where

$$\begin{aligned} O_k =& -\sum_{j=k+1}^{i_2-i_1-1}\mathsf{E}(W|\tau_{i_1+j})\mathsf{E}\tau_{i_1+j} + \frac{\lambda(1-\alpha_{i_2})\mathsf{E}\tau_{i_2}}{i_2\mu(i_2\mu-\lambda(1-\alpha_{i_2}))}(i_2-i_1-(k+1)) \\ &+ \frac{\mathsf{E}\tau_{i_2}}{2i_2\mu}\big[-k(k-1) + (2k-i_2+i_1+1)(i_2-i_1-2)\big] \end{aligned}$$

with convention $\sum_{j=i_2-i_1}^{i_2-i_1-1}\mathsf{E}(W|\tau_{i_1+j})\mathsf{E}\tau_{i_1+j} = 0$. Now we write C_2 as

$$\begin{aligned} C_2 =& h\Big\{(Q-i_1)^2\frac{1-\alpha_{i_2}}{2\beta_2} + \frac{Q-i_1}{2\beta_2}\big[-(1-\alpha_{i_2}) - 2(i_2-i_1-1) + \frac{2\lambda}{\beta_2}\big] \\ &+ \frac{i_2-i_1-1}{\beta_2(1-\alpha_{i_2})}\big(\frac{i_2-i_1}{2} - \frac{\lambda}{\beta_2}\big) + \lambda(1-\alpha_{i_2})\sum_{j=1}^{i_2-1}\mathsf{E}(W|\tau_j)\mathsf{E}\tau_j \\ &+ \alpha_{i_2}^Q\sum_{k=0}^{i_2-i_1-2}\frac{(Q-i_1)!}{(Q-i_1-k)!k!}\alpha_{i_2}^{-i_1-k}(1-\alpha_{i_2})^k\big[\frac{1}{2i_2\mu}\big(-k(k-1) \\ &+ (2k-i_2+i_1)(i_2-i_1-1)\big) + \lambda(1-\alpha_{i_2})O_k\big]\Big\} \end{aligned}$$

A.3 Long-run Average Cost $\Pi(Q)$

For $i_1 \geq 2$, we directly have

$$\begin{aligned}\Pi(Q) =& \frac{\mathsf{E}T_1(\lambda\alpha_{i_1} + p_{i_1}) + C_1 + \pi[\mathsf{E}T_2(\lambda\alpha_{i_2} + p_{i_2}) + C_2 + i_2 K]}{1/\lambda + \mathsf{E}T_1 + \pi\mathsf{E}T_2} \\ =& \frac{\frac{\mathsf{E}T_1}{\pi}(\lambda\alpha_{i_1} + p_{i_1}) + \frac{C_1}{\pi} + \mathsf{E}T_2(\lambda\alpha_{i_2} + p_{i_2}) + C_2 + i_2 K}{\frac{1/\lambda + \mathsf{E}T_1}{\pi} + \mathsf{E}T_2} \\ :=& \frac{a\theta_{i_1}^{Q-i_1} + a_2(Q-i_1)^2 + a_1(Q-i_1) + a_0 + \alpha_{i_2}^Q \sum_{k=0}^{i_2-i_1-2} \frac{(Q-i_1)!}{(Q-i_1-k)!} A_k}{b\theta_{i_1}^{Q-i_1} + b_1(Q-i_1) + b_0 + \alpha_{i_2}^Q \sum_{k=0}^{i_2-i_1-2} \frac{(Q-i_1)!}{(Q-i_1-k)} B_k},\end{aligned} \tag{A.21}$$

where

$$\begin{aligned}
a =& \frac{\mu_1(\lambda\alpha_{i_1} + p_{i_1})}{\rho\beta_1}\Big[- \frac{(i_1-1)(1-\pi_3)\mathsf{E}T_{11}}{\pi_1} + \frac{\rho\alpha_{i_1} + i_1}{(1-\alpha_{i_1})\beta_1} - (i_1-1)\mathsf{E}\hat{T}_{11}\Big] - \frac{h\rho\mu_1^2}{i_1\beta_1^3}, \\
a_2 =& \frac{h(1-\alpha_{i_1})}{2\beta_1} + \frac{h(1-\alpha_{i_2})}{2\beta_2}, \\
a_1 =& \frac{\lambda\alpha_{i_1} + p_{i_1}}{\beta_1} + \frac{\lambda\alpha_{i_2} + p_{i_2}}{\beta_2} - \frac{h}{2}\Big[\frac{1-\alpha_{i_1}}{\beta_1} + \frac{2\mu_1}{\beta_1^2} + \frac{1-\alpha_{i_2}}{\beta_2} + \frac{2(i_2-i_1-1)}{\beta_2} - \frac{2\lambda}{\beta_2^2}\Big], \\
a_0 =& \frac{\mu_1(\lambda\alpha_{i_1} + p_{i_1})}{\rho\beta_1}\Big[\frac{\rho}{\mu_1} - \frac{\mathsf{E}T_{11}}{\pi_1}\big[- \rho\big(\frac{\beta_1}{\mu_1} + \frac{(1-\alpha_{i_1})(i_1-1)}{i_1}\big) + \frac{\lambda(i_1-1)\pi_3}{\mu_1}\big] \\
&- \frac{\rho}{(1-\alpha_{i_1})\beta_1} + \frac{\lambda(i_1-1)\mathsf{E}\hat{T}_{11}}{\mu_1}\Big] + \frac{h\rho\mu_1^2}{i_1\beta_1^3} + (\lambda\alpha_{i_2} + p_{i_2})\big(\sum_{j=1}^{i_2-1}\mathsf{E}\tau_j \\
&- \frac{i_2-i_1-1}{(1-\alpha_{i_2})\beta_2}\big) + \frac{h(i_2-i_1-1)}{\beta_2(1-\alpha_{i_2})}\big(\frac{i_2-i_1}{2} - \frac{\lambda}{\beta_2}\big) + h\lambda(1-\alpha_{i_2})\sum_{j=1}^{i_2-1}\mathsf{E}(W|\tau_j)\mathsf{E}\tau_j, \\
A_k =& \frac{(1-\alpha_{i_2})^k}{k!\alpha_{i_2}^{i_1+k}}\Big[- (\lambda\alpha_{i_2} + p_{i_2})\Big(\sum_{j=i_1+1+k}^{i_2-1}\mathsf{E}\tau_j - \frac{i_2-i_1-(k+1)}{(1-\alpha_{i_2})\beta_2}\Big) \\
&+ h\frac{-k(k-1) + (2k-i_2+i_1)(i_2-i_1-1)}{2i_2\mu} + h\lambda(1-\alpha_{i_2})O_k\Big], \\
b =& \frac{\mu_1}{\rho\beta_1}\Big[- \frac{(i_1-1)(1-\pi_3)}{\lambda\pi_1} - \frac{(i_1-1)(1-\pi_3)\mathsf{E}T_{11}}{\pi_1} + \frac{\rho\alpha_{i_1} + i_1}{(1-\alpha_{i_1})\beta_1} - (i_1-1)\mathsf{E}\hat{T}_{11}\Big],
\end{aligned}$$

$$
\begin{aligned}
b_1 &= \frac{1}{\beta_1} + \frac{1}{\beta_2},\\
b_0 &= \frac{\mu_1}{\rho\beta_1}\Big[\frac{\rho}{\mu_1} - \frac{\mathsf{E}T_{11}}{\pi_1}\big[-\rho\big(\frac{\beta_1}{\mu_1} + \frac{(1-\alpha_{i_1})(i_1-1)}{i_1}\big) + \frac{\lambda(i_1-1)\pi_3}{\mu_1}\big] - \frac{\rho}{(1-\alpha_{i_1})\beta_1}\\
&\quad + \frac{\lambda(i_1-1)\mathsf{E}\hat{T}_{11}}{\mu_1} + \frac{\rho}{\lambda\pi_1}\big[\frac{\beta_1}{\mu_1} + \frac{(1-\alpha_{i_1})(i_1-1)}{i_1}\big] - \frac{(i_1-1)\pi_3}{\mu_1\pi_1}\Big]\\
&\quad + \sum_{j=1}^{i_2-1}\mathsf{E}\tau_j - \frac{i_2-i_1-1}{(1-\alpha_{i_2})\beta_2},\\
B_k &= -\frac{(1-\alpha_{i_2})^k}{k!\alpha_{i_2}^{i_1+k}}\Big(\sum_{j=i_1+1+k}^{i_2-1}\mathsf{E}\tau_j - \frac{i_2-i_1-(k+1)}{(1-\alpha_{i_2})\beta_2}\Big).
\end{aligned}
$$

For $i_1 = 1$, we only have matrix $\mathbf{D}_{14}$ and $\pi_1 = \pi_3 = \pi = 1$ and $\pi_2 = 0$. In this case, T_1 includes idle time because of the structure of matrix $\mathbf{D}_{14}$, which includes transitions of all states before reaching N. Also notice that when $i_1 = 1$, $\mathbf{D}_{14}$ is exactly the same as matrix $\bar{\mathbf{D}}_1$ in paper 1. Thus we can derive expected idle time $\mathsf{E}T_0$ directly from there by replacing $Q+1$ by Q and θ by θ_1, which is

$$
\mathsf{E}T_0 = \frac{1-\theta_1^{Q-1}}{\lambda\rho(1-\theta_1)} + \frac{1}{\lambda} = \frac{\rho(1-\alpha_1) - \theta_1^{Q-1}}{\rho(1-\alpha_1)\beta_1}.
$$

Then, long-run average cost $\Pi(Q)$ can be written as

$$
\Pi(Q) = \frac{(\mathsf{E}T_1 - \mathsf{E}T_0)(\lambda\alpha_1 + p_1) + C_1 + \mathsf{E}T_2(\lambda\alpha_{i_2} + p_{i_2}) + C_2 + K}{\mathsf{E}T_1 + \mathsf{E}T_2}
$$

If we still use expression (A.21) , we only need to modify the following coefficients:

$$\begin{aligned}
a &= \frac{(\lambda\alpha_1 + p_1)\mu_1}{(1-\alpha_1)\beta_1^2} - \frac{h\rho\mu_1^2}{\beta_1^3},\\
a_0 &= -\frac{(\lambda\alpha_1 + p_1)\mu_1}{(1-\alpha_1)\beta_1^2} + \frac{h\rho\mu_1^2}{\beta_1^3} + (\lambda\alpha_{i_2} + p_{i_2})\big(\sum_{j=1}^{i_2-1} \mathsf{E}\tau_j - \frac{i_2-2}{(1-\alpha_{i_2})\beta_2}\big)\\
&\quad + \frac{h(i_2-2)}{\beta_2(1-\alpha_{i_2})}\Big(\frac{i_2-1}{2} - \frac{\lambda}{\beta_2}\Big) + h\lambda(1-\alpha_{i_2})\sum_{j=1}^{i_2-1} \mathsf{E}(W|\tau_j)\mathsf{E}\tau_j,\\
b_0 &= \frac{1}{\beta_1} - \frac{\mu_1}{(1-\alpha_1)\beta_1^2} + \sum_{j=1}^{i_2-1} \mathsf{E}\tau_j - \frac{i_2-2}{(1-\alpha_{i_2})\beta_2}.
\end{aligned}$$

BIBLIOGRAPHY

[1] Allon G, Bassamboo A (2011) The impact of delaying the delay announcements *Operations Research* 59(5):1198–1210.

[2] Ancker CJ, Gafarian AV (1963a) Some queuing problems with balking and reneging, I. *Operations Research* 11(1):88–100.

[3] Ancker CJ, Gafarian AV (1963b) Some queuing problems with balking and reneging, II. *Operations Research* 11(6):928–937.

[4] Armony M, Plambeck E, Seshadri S (2009) Sensitivity of optimal capacity to customer impatience in an unobservable $M/M/S$ queue (Why you shouldn't shout at the DMV). *Manufacturing & Service Operations Management* 11(1):19–32.

[5] Armony M, Shimkin N, Whitt W (2009) The impact of delay announcements in many-server queues with abandonment *Operations Research* 57(1):66–81.

[6] Baccelli F, Boyer P, Hebuterne G (1984) Single-server queues with impatient customers. *Advances in Applied Probability* 16(3):887–905.

[7] Bell CE (1975) Turning off a server with customers present: Is this

any way to run an $M/M/c$ queue with removable servers? *Operations Research* 23:571–574.

[8] Bell CE (1980) Optimal operation of an $M/M/2$ queue with removable servers. *Operations Research* 28(5):1189–1204.

[9] Billingsley P (2009) *Convergence of Probability Measures*. John Wiley & Sons.

[10] Blackburn D (1972) Optimal control of a single-server queue with balking and reneging *Management Science* 19(3):297–313.

[11] Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association* 100(1): 36–50.

[12] Cox DR, and Smith WL (1961) *Queues* (CRC Press).

[13] Fu MC, Marcus SI, Wang I (2000) Monotone optimal policies for a transient queueing staffing problem. *Operations Research* 48(2):327–331.

[14] Gans N, Zhou Y (2003) A call-routing problem with service-level constraints. *Operations Research* 51(2):255–271.

[15] Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing & Service Operations Management* 4(3):208–227.

[16] Gavish B, Schweitzer PJ (1977) The Markovian queue with bounded waiting time. *Management Science* 23(12):1349–1357.

[17] Gnedenko BV, Kovalenko I (1989) *Introduction to queueing theory* (Birkhauser Boston Inc).

[18] Jennings O, Pender J (2015) Comparisons of ticket and standard queues. *Working paper.*

[19] Kitaev M Y, Serfozo R F (1999) $M/M/1$ queues with switching costs and hysteretic optimal control, *Operations Research* 47(2):310–312.

[20] Harrison JM, Zeevi A (2005) A method for staffing large call centers based on stochastic fluid models. *Manufacturing & Service Operations Management* 7(1):20–36.

[21] Hall P, Heyde CC(1980) *Martingale Limit Theory and its Application.* New York: Academic Press.

[22] Hokstad P (1979) A Single Server Queue with Constant Service Time and Restricted Accessibility. *Management Science* 25(2):205–208.

[23] Lu FV, Serfozo RF (1984) $M/M/1$ queueing decision processes with monotone hysteretic optimal policies, *Operations Research* 32(5):1116–1132.

[24] Mandelbaum A, Shimkin N (2000) A model for rational abandonments from invisible queues. *Queueing Systems* 36(1-3):141–173.

[25] Mandelbaum A, Zeltyn S (2009) Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Operations Research* 57(5):1189–1205.

[26] Mandelbaum A, Zeltyn S (2013) Data-stories about (im) patient customers in tele-queues *Queueing Systems* 75(2-4):115–146.

[27] Omahen K, Marathe V (1978) Analysis and Applications of the Delay Cycle for the $M/M/c$ Queueing System. *Journal of the Association for Computing Machinery* 25(2):283–303.

[28] Pang G, Perry O (2014) A Logarithmic Safety Staffing Rule for Contact Centers with Call Blending. *Management Science* to appear.

[29] Reynolds JF (1968) A stationary solution of a multiserver queueing model with discouragement. *Operations Research* 16(1):64–71.

[30] Resnick S I (2007) *Extreme Values, Regular Variation, and Point Processes.* Springer.

[31] Ross SM (1966) *Stochastic Processes* (John Wiley & Sons, New York).

[32] Stanford RE (1979) Reneging phenomenal in single-server queues. *Mathematics of Operations Research* 4(1):162–178.

[33] Topkis D (1998) *Supermodularity and Complementarity* (Princeton University Press).

[34] Van Dijk NM (1990) Queueing systems with restricted workload: an explicit solution. *Journal of Applied Probability* 22(2):393–400.

[35] Xu S, Gao L, Ou J (2007) Service performance analysis and improvement for a ticket queue with balking customers. *Management Science* 53(2):971–990.

[36] Yadin M, Naor P (1967) On queueing systems with variable service capacities. *Naval Research Logistics Quarterly* 14(1):43–53.

[37] Zhang B, van Leeuwaarden J, Zwart B (2012) Staffing call centers with impatient customers: Refinements to many-server asymptotics. *Operations Research* 60(2):461–474.

[38] Zhang ZG (2009) Performance analysis of a queue with congestion-based staffing policy. *Management Science* 55(2):240–251.

[39] Zohar E, Mandelbaum A, Shimkin N (2002) Adaptive behavior of impatient customers in tele-queues: theory and empirical support. *Management Science* 48(4):566–583.