공학석사 학위논문

Server Operating Policies for Appointment System: A Simulation Approach

2012년 8월

서울대학교 대학원

산업공학과

백 민 정

Abstract

Server Operating Policies for Appointment System: A Simulation Approach

Baek Minjeong
Department of Industrial Engineering
College of Engineering
Seoul National University

As customers today expect timely services without delays, service offering firms have adopted various strategies to reduce their waiting times. As one of the most widely used ways to offer a timely service, an appointment system for customer arrival has been often adopted in many companies. Recent studies dealing with the appointment system have discussed about the determination of number of servers or the appointment scheduling for customers' arrivals through optimization, heuristic, or simulation-based approaches. In this study, we suggest a novel approach to server operating policies for the appointment system by introducing a new type of servers

defined as standby servers in addition to the typical type of regular servers. Standby servers perform regular servers' overtime work for the customer so that the regular server can provide the timely service to the next customer. We developed two different simulation models that illustrate the service offering process in our focal company: one is for the case under the base policy using only regular servers while the other is the mixed policy integrating both types of servers. Then, under various operating conditions corresponding to the changing environmental factors, we compared the performance of each policy measured by total cost including server operating cost and the delay penalty cost.

Keywords: Appointment system, standby servers, customer waiting time, service time

Student Number: 2010-23377

Contents

List of Tables

List of Figures

Chapter 1 Introduction

The service sector has become increasingly important in the economies of developed and developing countries. At the same time, fierce competition among service offering firms forced them to differentiate themselves in various dimensions to succeed. Timely service is one of the competitive priorities for successful service in today's environment (Davis and Heineke, 1998; Hensley and Sulek, 2007).

Customers today are more constrained by time than ever before and constantly look for shorter waiting time for services. There are various ways to reduce service delay such as giving customers advance notice of expected waiting time, speeding up pre-process waiting time, and guaranteeing quick service to customers (Nie, 2000). Service managers are able to increase the overall satisfaction with these operational practices.

As one way to offer timely service, appointment system for customer arrival is often adopted in many service providing companies. The logic behind this is that customers who arrive early into the service system are contentedly willing to wait until the scheduled time (Jones and Peppiatt, 1996; Bielen and Demoulin,

2007). However, appointment system itself can be troublesome due to the service time variability. As a result, service offering firms with appointment system often find it difficult to decide how far apart to schedule appointments (Maister, 1985). Moreover, the problem of adjusting schedule is usually related to the problem of adjusting capacity (Javel and Riopel, 2010). For example, companies need to increase the size of capacity by hiring more staffs or machines in order to make appointment interval further with each other unless they want customers to wait too long. However, large capacity size means higher operating costs. Thus, service managers need to carefully decide the appointment schedules to minimize possible service delays while taking into account those operating costs.

This study explores two different types of policies to manage service delays under appointment system with the constant interappointment time. First policy is to utilize every server as a regular server. Under this policy, every server is assigned with the equal number of customers to provide services. Thus, total number of servers decides the interappointment time. Second policy is to set aside a portion of servers as standby servers. Under this policy, only regular servers are assigned with the customer appointments while standby servers are assigned with the regular servers' overwork.

The important difference between these two polices is the type of

service delay customers experience. Under the first policy, customers can experience service delay when the server is not ready at the appointed time. On the other hand, under the second policy, customers who need disproportionately long service time can experience service delay when they are redirected to the standby servers after they receive a service for a certain amount of time from the regular servers.

While there have been abundant research on appointment scheduling and various related server operating polices, less is known about the impact of standby servers' presence. Here we have taken simulation approach to analyze the impacts of standby servers through a practical case of a service providing company.

The remainder of the paper is organized as follows. Section 2 reviews existing literature on related issues and gives direction of our study. In section 3, we model company's operating environments regarding appointment scheduling. Section 4 describes simulation experiment and presents the experimental results. Section 5 summarizes the research findings and suggests the directions for future investigations.

Chapter 2 Literature Review

We present a survey of the relevant research literature under two separate heads: appointment system and research methods.

2.1 Appointment system

In service operation context, an appointment system is a system in which customers are assigned to the specific timepoints at which the service firm starts to offer its services. The purpose of an appointment system is to provide timely services to customers. (Edward and Jr. P.H., 1976). It is used in many service providing firms to increase utilization of resources and decrease waiting time of customers. Several issues about appointment system have been studied extensively in various contexts such as in transportation (Sabria and Daganzo, 1989; Gupta, 2011), manufacturing (Wang, 1993; Elhafsi, 2002), call center environment (Pichitlamken, et al., 2003), etc.

Most of literatures about appointment system are mainly focused on appointment rules in the medical industry. For example, there is a bulk of literature dealing with the scheduling rules for patients, doctors, or ORs in a health care organization (Ho and Lau, 1992; Christie and Levary, 1998; Lawrence and Rachel, 2003; Gupta and

Denton, 2008; Vermeulen, et al., 2009; Najmuddin, et al., 2010; Denton and Gupta, 2003).

For the last 60 years, various appointment rules have been suggested. One of the most commonly discussed objectives of appointment rules is to decide the interarrival times between customers. They range from single-block appointments in which all customers arrive at the beginning of the service on the one extreme to individual appointments, in which all customers are given unique appointment times on the other. Most of the policies are modifying or combining these two systems (Wijewickrama and Takakuwa, 2005). Variable block sizes with fixed intervals (Rising, et al., 1973; Liao, et al.,1993) and fixed block sizes with variable intervals (Vanden and Dietz, 2000; Wang, 1997) have been studied. This study deals with the appointment system with a constant interarrival time as used in Bailey (1952) and Welch (1964). Also, the block size for each arrival is one here.

2.2 Server operating policies

Service organizations have to size the number of servers so that they can efficiently match server capacity to customer demand. Server sizing problem in an appointment system has been studied mostly in the medical center context. For example, various types of

performances of the service were measured through patients' waiting time or resource idle time in order to decide the optimal size of surgeons, ORs, nurses, etc. (May, et al., 2000; Elliott, 1990; Wright, et al, 2006).

Most of the times, the server size directly affects how many customers' can get timely services. In a way, the problem of server sizing in a service operating organization can be viewed as a newsvendor problem. As the newsvendor model seeks to find the optimal resource quantity to satisfy customer demand considering resourcing cost, the server sizing problem aims to optimize the number of servers to provide timely services to the customers considering server operating cost. Those problems need to take into account the uncertainty factors in the service system such as daily demand for the service, service time, customer no-shows, punctuality, etc. Server sizing problems can be framed by the company's server operating policy. Server operating policy refers to the conditions such as type of servers (e.g. skilled labor versus unskilled labor), working hour of each server, way of allocating servers to the different kinds of customers, etc.

For example, companies can adopt server operating policies related to the server operating time such as overtime or undertime, supplemental part-time, shift work, or split shifts to manage

fluctuating demands (Mabert and Watts, 1982; Bracken, et al., 1985; Bechtold, 1991). According to the policy adopted, the optimal server size can be derived. Service firms also can utilize servers for different purposes. Standby server and scheduled server suggested by Mok and Shanthikumar (1987) is one example for that. The role of the standby server is to be "standby" in case the service time for a customer who has been served by the scheduled server becomes longer than a certain amount of time. When that happens, one of the standby servers takes charge of the rest of the service. The purpose of this policy is to eliminate the customers' waiting time from the first place.

2.3 Research methods

The majority prior research has used analytical methods and simulation methodologies to tackle the appointment system problem. A lot of works adopted analytical methods based on queuing theory (Jansson, 1966; Pegden and Rosenshine, 1990; Stein and Côté, 1994). However, these models are restricted to using Erlang or Exponential service times to make them tractable (Cayirli and Veral, 2003). Also, they have difficulty in capturing the complexity of the system.

Many researches about appointment system used simulation techniques as well. Simulation is able to model various decision variables and environmental factors in the complex queuing systems.

It is possible to evaluate the performance of the system such as service delay time or server idle time to understand the impact of those factors on the system as well. However, it does not give the optimum value of the problem and researchers have to pre-determine all possible solutions (Klassen and Yoogalingam, 2009).

In this study, we will test a number of server operating options while evaluating the impacts of standby servers. Resources (servers) will be allocated either as standby servers or regular servers as opposed to the standby servers. Lognormal service times will be used to evaluate the performance of the polices under uncertainty. Lognormal times have been found empirically (e.g., Cayirli, et al., 2006; Klassen and Rohleder, 1996; O'Keefe, 1985). Also, Customer are assigned with appointed schedules with a fixed interappointment time.

The service we are going to deal with in this study is the one with an appointment system. We will analyze the system under two different server operating policies, one with the standby servers and the other without them. We will call the policy with the standby servers ‘mixed policy’ throughout this article and the one without the standby servers ‘base policy.’ Servers who are not standby servers will be called as regular servers.

Under mixed policy, regular servers are supported by standby servers, whose job is to finish the regular servers overwork. To the best of our knowledge, only one study has dealt with this type of server under non-appointment system (Mok and Shanthikumar, 1987). They have dealt with the queuing system with standby servers who should serve stochastically arriving customers.

In this study, we will analyze the service system where customers arrive to the system at the appointed time and compare the performance under base policy and mixed policy through simulation approach.

Chapter 3 The Model

3.1 Description of the service environment

The company studied in this research is a B2C furniture manufacturer who provides furniture delivery and installment services. When a customer places an order, the company not only delivers the ordered items to the customer's place but also provides installment services such as assembly of the knockdown items, rearrangement of the existing furniture, etc. In order to provide those services, it has been utilizing "delivery teams", each of which consists of one truck and two workers.

Every morning, 30 delivery teams are assigned with the delivery schedules in the distribution center where they load the ordered furniture on the trucks. The schedules contain places and time to visit on that day. Since the company does not want the delivery teams to waste their time on traveling around long distances, they put each team in charge of the orders that are near to each other. This usually results in each team being responsible for about 10 orders a day in a specified area.

A typical day of a delivery team begins with visiting a customer scheduled for the first place. Since delivery teams are not late for the

first stops, installment service usually starts on time for the first scheduled customers. After finishing installment of the ordered furniture in the first customer's place, they move to the second place, install the ordered furniture, and move to the third customer, and so on. The work of the delivery teams end when they finish the installment services for the last customer of the day.

Time to install the furniture varies among customers since customers' requirements over installment are different from each other. Delivery teams have no prior information regarding these installment times before they arrive to the customers' places. Time to travel between two places has much less variability and its amounts are negligible compared to the installment time. The company has been scheduling in a way that the visiting intervals between two consecutive places are equal, given its operating hours. For example, the visiting interval for a delivery team responsible for ten orders with ten operating hours is one hour.

Before the delivery teams leave the distribution center with the loaded furniture, the company makes a phone call to every customer in order to notify them the visiting time according to the schedule. Customers wait for the delivery teams at the appointed time at the appointed place, which is usually their home. Delivery teams can either arrive on time or be late for the appointment because of the

previous orders. The company has been receiving complaints from the customers because of delivery teams' late show-ups. Now it is considering increasing the number of delivery teams and choosing between two different server operating policies to utilize them, which we will explain with more details in 3.2.

3.2 Formulation

Assume that the customer j of the server i starts to wait punctually at the appointed time A_{ij}. Then, the installment service with the appointment system can be modeled as a queuing system where server i's visit to customer j at A_{ij} can be interpreted as customer j's arrival at A_{ij} requesting installment services, as shown in Figure 1.

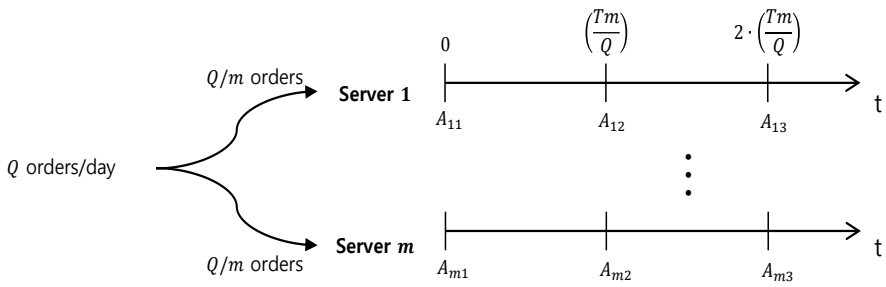

Figure 1 Order allocation for regular servers

Customers' appointment times are assigned by the following rule.

$$A_{ij} = \begin{cases} 0 & \text{for } j = 1 \\ \left(\frac{Tm}{Q}\right) \cdot (j-1) & \text{for } j = 2, \dots, \left(\frac{Q}{m}\right) \end{cases} \quad (1)$$

for all i, where T is the operating hours, m is the number of servers, and Q is the total number of orders for one day. This rule implies that the company assigns every server with an equal amount of orders, which is the total number of orders divided by the number of servers. Description of appointment scheduling is shown in Figure 2.

The company is considering two different server operating polices; 'base policy' and 'mixed policy.' The difference between two policies is the existence of the 'standby servers' as opposed to the 'regular servers' and how they respond to the long service times. In base policy, all servers are operated as regular servers while in mixed policy, servers are operated either regular servers or standby servers. Their detailed descriptions are shown below.

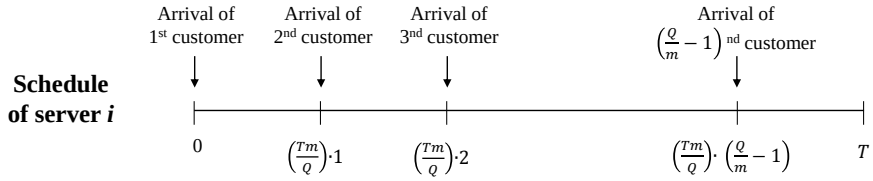

Figure 2 Appointment scheduling for each server

Base policy: regular servers only

First policy is the base policy, which is currently adopted in our focal company. This policy is a good benchmark since it is remained one of the best policies for appointment scheduling over the past sixty years (Wijewickrama and Takakuwa, 2005) and commonly adopted in many server offering firms.

In this policy, every server is utilized as a regular server. Assume that the number of regular servers is m. Then, each server is assigned with $\left(\frac{Q}{m}\right)$ orders. Regular servers can leave for the next customer only if it has fully finished the installment services for the previous customers. This means that the customers might have to wait for the delivery teams until their delivery team finishes the services for all previous customers.

Let t_{ij}, b_{ij} and e_{ij} be, respectively, the length of installment service time[1], the time at which service begins, and the time at which service ends for customer j of the server i. Since t_{ij} is stochastic,

[1] We will assume the time to travel between customers' places to be zero.

so are b_{ij} and e_{ij}. Without loss of generality, $A_{i1} = b_{i1} = 0$, . For $j > 1$, we have

$$b_{ij} = \max\left(A_{ij},\ e_{i(j-1)}\right),\ e_{ij} = b_{ij} + t_{ij}\ . \tag{2}$$

Customer j of server i's "start delay" is then

$$d_{ij} = \max\left(0, b_{ij} - A_{ij}\right). \tag{3}$$

If a server is ready at the customer's place at the appointed time, d_{ij} is 0. On the other hand, if a server shows up at the customer's place later than the appointed time, d_{ij} is higher than 0. An example for the possible start delays under the base policy are described in Figure 3.

Mixed policy: regular servers with standby servers

Under mixed policy, some of the servers are set aside as standby servers to respond to the contingency events. A contingency event is that installment service time for a certain customer becomes longer than a specific amount of time. As in the base policy, orders are assigned only to the regular servers which means visiting interval is $\left(\frac{Tm}{Q}\right)$. Upon contingency, regular servers stop its service at $\left(\frac{Tm}{Q}\right)$ and

let one of the standby servers finish the rest of their services so that they can leave for the next customer. We will call this action of regular servers to "toss." The residual installment service time tossed from the regular server i for customer j, r_{ij} is then defined as

$$r_{ij} = t_{ij} - \left(\frac{Tm}{Q}\right). \tag{4}$$

Meanwhile, the standby server who was called goes to the customer's place where contingency has occurred and finish the rest of the service if it is available. At every $A_{\cdot j}$, residual services, if there is any, are tossed to the standby servers with an oreder from the regular server 1 to the regular server m. However, there are cases when all standby servers are busy processing previous contingency events. This means that the customers might have to wait for the standby servers to come until they are available.

Let n be the number of standby servers. We define total length of tossed services, which standby server k has to process, at the moment regular server i tossed the residual service of customer j as

$$R_{ijk} = \begin{cases} R_{i(j-1)k} + r_{ij} & \text{if } d_{i(j-1)} = R_{i(j-1)k} \\ R_{i(j-1)k} & \text{else} \end{cases} \tag{5}$$

for $i = 2, \dots, m$ and for all j and k. Also,

$$R_{1jk} = \begin{cases} 0 & \text{if } j = 1 \\ \max\left(R_{m(j-1)k} - \left(\frac{Tm}{Q}\right), 0\right) & \text{else} \end{cases} \quad (6)$$

for all k.[2]

Customer j of server i's "restart delay", which is the amount of time until one of the standby servers starts to begin customer i's residual service tossed by server j, is then

$$d_{ij} = \begin{cases} 0 & \text{if } r_{ij} = 0 \\ \min\left(R_{ij1}, R_{ij2}, \dots, R_{ijn}\right) & \text{if } r_{ij} > 0. \end{cases} \quad (7)$$

If there was no residual service tossed by the regular server from the first place, d_{ij} is 0. On the other hand, if a customer has to wait for the standby server to finish the residual services, d_{ij} is higher than 0. An example for the possible restart delays under the mixed policy are described in Figure 4.

[2] When multiple regular servers toss their residual services at the same time, we assume that the services are tossed from regular server 1 to regular server *m* in order.

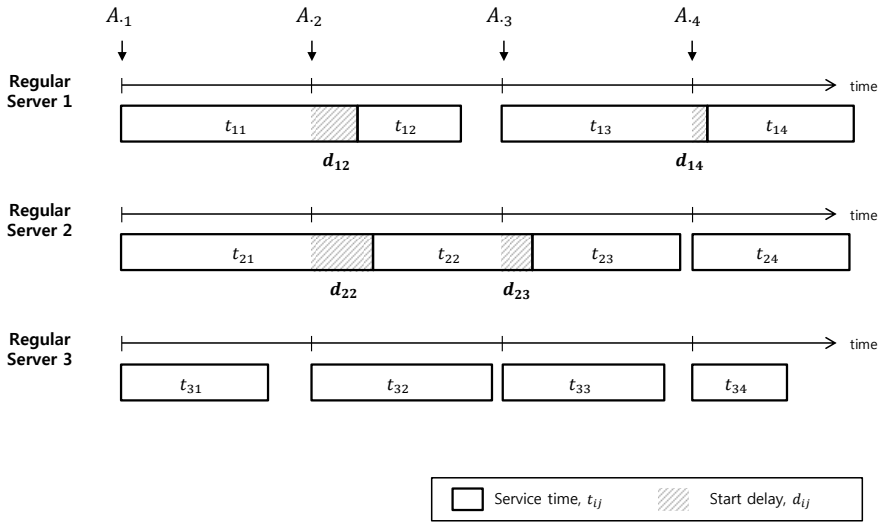

Figure 3 Service delays under base policy

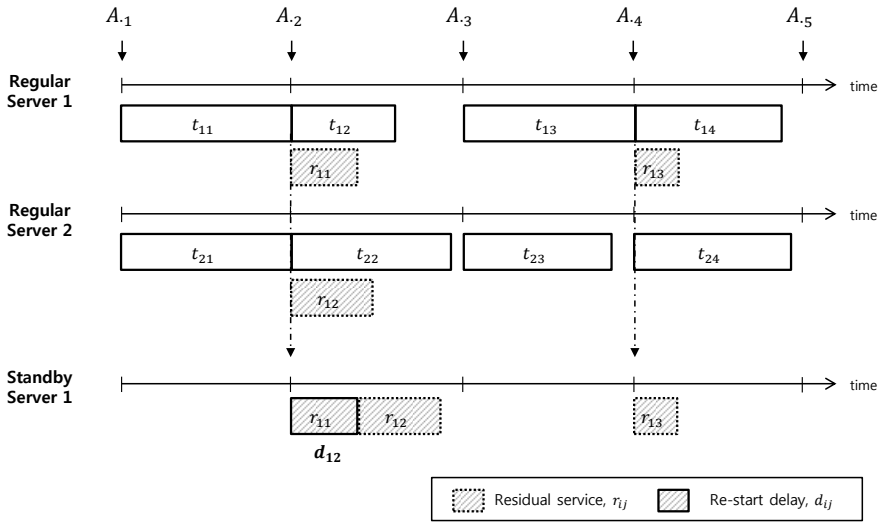

Figure 4 Service delays under mixed policy

3.3 Performance measures

The primary performance measure used to compare two server operating policies is the combined cost of server operating cost and service delay penalty cost. While customers' waiting time and servers' idle time are often the most concerned factors, we use server operating cost instead of costs derived from servers' idle time. The reason we use server operating cost instead of servers' idle time is based on the employee welfare company needs to meet. Even if a server is idle during the day, the company has to pay the wages for the full one-day. This imposes the company to carefully decide the number of servers beforehand considering demands.

Formally, total cost which is a performance measure used here is

$$\mathrm{TC} = c_o \cdot (m+n) + c_p \cdot \sum_i \sum_j d_{ij} \tag{8}$$

where c_o is operating cost to utilie one server a day and c_p is the penalty cost per hour incurred due to service delays.

Chapter 4 Simulation Experiment

The company wants to choose between base policy and mixed policy. In addition, the optimal number of servers needs to be decided here so that the company can minimize customers' waiting time while keeping server operating cost as low as possible. Since service operation under each policy with a given number of servers derives different performance of the system, the problem in this study is different from the simple resource-sizing problem which can be solved with the analytical approach based on queuing theory. Also, existence of standby servers hinders us from analytically study service operation under mixed policy.

Simulation experiments can give us an idea of which server operating option under which policy is more effective in reducing the total cost. We are also going to see how changes in server operating strategies can affect the total length of service delays.

4.1 Data collection

Data was collected via interviewing company administrators, workers consisting of servers, and other clerical personnel, with the addition of observation. General situations of the service operations regarding

daily demands, number of servers, steps involved in appointment scheduling, distribution of service time, costs related to servers and service delays represent the data collected from these sources.

Operating conditions

The company is currently utilizing delivery teams to deal with an approximate 300 daily demands. This study made the simplifying assumption that its daily demands are fixed as 300. Operating cost to utilize one delivery team, which is comprised of one truck and two workers, is $15 per hour. Official operating hours for those delivery teams is 10 hours a day. As for the penalty cost, we estimated the unit penalty cost for service delay by measuring how much a company is prepared to pay for the improved service quality. The service manager in the company is now willing to invest in up to 3 more servers to minimize service delays. Currently, total service delay is about 6 hours a day. This means the company is willing to pay $450 by hiring three more additional servers for a six-hour reduction in waiting time.

Table 1 shows the current environmental factors and their values. Those will be used as input data for simulation later.

Table 1 Input data

Number of orders (Q)	300 (per day)
Server operating hours (T)	10 (hours per day)
Server operating cost (c_o)	150 ($ per day)
Delay penalty cost (c_p)	75 ($ per hour)

Service time distribution

Service time to install furniture in customers' places varies. In computing service time distributions for our model, we utilize the information that the average service time is around 50 minutes. We also collected a small number of data coinciding with the above information. The service managers estimate total delay time in one day at 6 hours.

The company also expects the service time distribution to follow the similar form of normal distribution given that

In this study, the lognormal distribution with the mean of 50 minutes and the standard deviation of 10 minutes is used to generate the service time. To validate this assumption, we ran a mock simulation and checked the total service delay generated from our assumed distribution is around 6 hours.

Server operating options

Since the company currently is adopting the base policy, all 30 servers are operated as regular servers. The number of servers can be increased up to 33 servers from the current level of 30 servers. Possible options for server operating strategy is shown in table 2

Table 2 Server operating options

Experiment unit number	**Server operating policy**	**Number of regular servers (m)**	**Number of standby servers (n)**
1	Base	30	0
2	Mixed	30	1
3	Mixed	30	2
4	Mixed	30	3
5	Base	31	0
6	Mixed	31	1
7	Mixed	31	2
8	Base	32	0
9	Mixed	32	1
10	Base	33	0

4.2 The simulation model and analysis

This study employs a discrete event simulation technique and aims to identify the best server operating strategy under realistic environmental factors. The simulation models were built using Microsoft Excel spreadsheets combined with VBA (Visual Basics for Applications).

Under the specified conditions in 4.1., we ran 100 experiments for 10 operating options to calculated average total service delays and their related costs.

As for the cases when the number of orders is not exactly divided by the number of regular servers, we tried to allocate the orders to the regular servers as much as even among servers. For example, 300 orders were allocated to the 32 regular servers so that 12 servers deliver 10 orders and 20 servers deliver 9 orders. To randomly generate service times in our models, we used the Inverse Transform Method (Ross, 2002)

4.3 Results: Evaluation of scheduling polices

Table 3 shows results for each option in total cost and delay time. Results from the model reveal that the company needs to keep its current server operating strategy, which is to utilize all 30 servers as regular servers with no additional delivery teams. Total cost derived

from each policy is shown in Figure 4. It shows all possible options under mixed policy result in higher costs than the option (30,0) of base policy.

The reason for this can be explained in two respects. One is the effect of the service time distribution on two different kinds of service delays. Under the mixed policy, services are tossed from the regular servers to the standby servers only when the service time is longer than the interappointment time, which is $\left(\frac{Tm}{Q}\right)$. If there are too many services tossed by the regular servers at every $A_{\cdot j}$, a long queue of residual services is formed for standby servers leading to longer restart delays. Under the base policy, on the other hand, when the service time of a certain customer is longer than $\left(\frac{Tm}{Q}\right)$, the next customer's waiting time will be, at most, the amount of residual service. This means that switching into mixed policy with additional severs from 30 servers can result in longer delays. For example, server operating options, (30,1), (30,2), and (31,1) produce longer delays than (30,0). Thus, in order for the mixed policy to be effective to reduce the service delay, the number of the standby servers should be enough to handle the number of orders tossed by the regular servers. In our case with the service time distribution of 50 minutes mean and 10 minutes standard deviation, expected number of tossed services can be

calculated as follows.

$$m \cdot \Phi\left(\frac{\ln\left(\frac{10 \cdot m}{300}\right) - (-0.20193)}{0.19804}\right) \quad . \tag{9}$$

Table 3 shows the expected number of tossed services at each appointment time for different *m*.

Table 3. Expected number of tossed services

m	**E(number tossed services)**
30	4.619
31	3.657
32	2.855
33	2.201

Second, the ratio of server operating cost to delay penalty cost influences on the total cost under each server operating options. This is somewhat obvious. In Figure 5, we are able to see option (32,0), (31,2) incur longer total delay than (30,0) does. These options cannot be the best option which minimizes the total cost, because of the higher server operating cost than (30,0). Despite the fact that the company is willing to add 3 more servers, simulation results indicate it

is better to keep its current level of capacity size. This reveals that the company does not value the delay penalty cost highly enough to actually make a change in their operating system. It might as well keep 30 as add more servers.

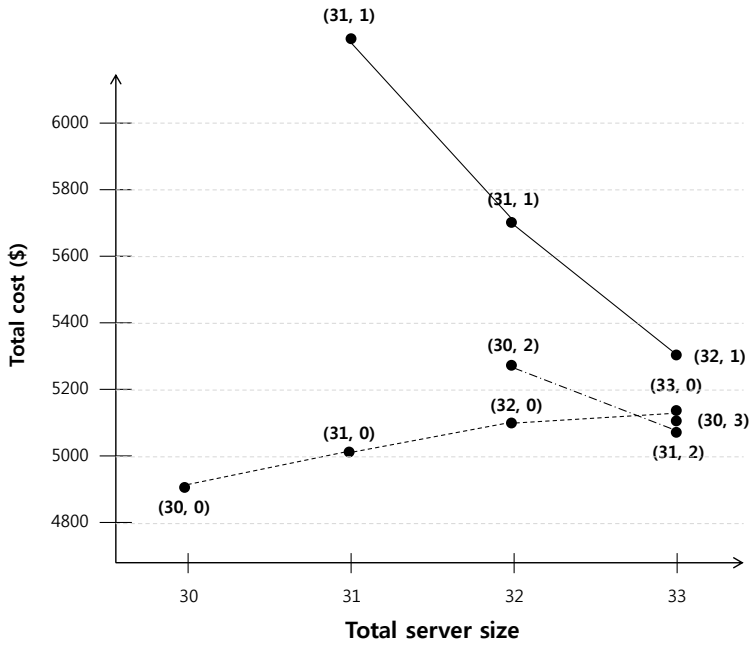

Figure 5 Total cost for each server operating option

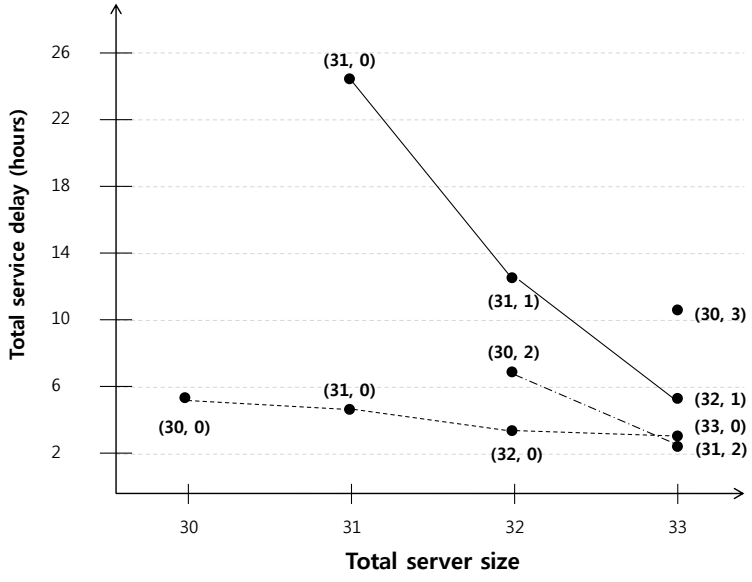

Figure 6 Total service delay for each server operating option

4.4 Sensitivity of the performance measures to the service time distribution

We now look at how the performance measures are affected by a change in the service time distribution. From the simulation experiment, we observed that the operating option resulting with the lowest total cost varies along with the changes in mean (μ) and standard deviation (σ) of the service time. μ was adjusted from 45 minutes to 59 minutes while σ was adjusted from 1 minute to 15 minutes. To better capture the impact of the service time distribution, delay penalty cost (c_p) was set to \$200 per hour. We ran 100 experiments for each combination. Table 3 shows the best server operating option under 165 different operating conditions.

Results show that mixed policy is prior to the base policy within a certain boundary of μ and σ. As mentioned above, we expect the relationship between expected number of tossed services and the number of standby servers together has a decisive effect on the effectiveness of the policy.

Table 4 Server operating option with the lowest total cost under various σ and μ

μ \ σ	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
45	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	**(30,2)**	**(31,2)**	**(31,2)**	**(31,2)**	**(31,2)**	(33,0)
46	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	**(30,2)**	**(31,2)**	**(31,2)**	**(31,2)**	**(31,2)**	**(31,2)**	(33,0)
47	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	**(31,2)**	**(31,2)**	**(31,2)**	**(31,2)**	(33,0)	(33,0)
48	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	**(31,2)**	**(31,2)**	**(31,2)**	**(31,2)**	(33,0)	(33,0)	(33,0)
49	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	**(31,2)**	**(31,2)**	**(31,2)**	**(31,2)**	(33,0)	(33,0)	(33,0)
50	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	**(31,2)**	**(31,2)**	**(31,2)**	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)
51	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	**(31,2)**	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)
52	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)
53	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)
54	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)
55	(30,0)	(30,0)	(30,0)	(30,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)	(33,0)

$c_o = 150,\ c_p = 200, T = 10, Q = 300$

Table 5. Total service delay under various server operating options and σ

σ \ Policy	(30,0)	(30,1)	(30,2)	(30,3)	(31,0)	(31,1)	(31,2)	(32,0)	(32,1)	(33,0)
1	**0**	0	0	0	0	0	0	0	0	0
2	**0**	0	0	0	0	0	0	0	0	0
3	0.005	**0**	0	0	0.004	0	0	0.002	0	0
4	0.064	0.011	**0**	0	0.057	0.001	0	0.029	0	0.006
5	0.342	0.136	0.018	0.010	0.265	0.039	**0.002**	0.113	0.022	0.039
6	0.830	0.764	0.133	0.110	0.606	0.245	**0.029**	0.373	0.169	0.171
7	1.754	1.830	0.629	0.435	1.201	0.991	**0.119**	0.868	0.372	0.452
8	2.736	5.923	1.745	1.221	2.19	2.199	**0.422**	1.557	0.849	0.879
9	4.565	12.683	3.326	3.147	3.357	4.683	**1.192**	2.154	2.783	1.433
10	5.851	24.431	6.395	5.835	4.910	12.541	**2.332**	3.613	4.975	2.372
11	8.261	73.706	14.231	8.919	6.548	32.758	3.868	5.208	15.291	**3.303**
12	10.327	104.360	17.566	15.015	8.214	77.488	8.305	7.357	41.340	**5.099**
13	12.992	232.715	33.803	30.587	11.124	128.019	17.999	8.796	72.515	**6.834**
14	15.446	295.906	58.069	55.925	14.124	210.866	27.473	11.032	103.132	**8.073**
15	19.111	457.686	107.387	86.568	15.095	309.311	54.471	13.129	194.693	**10.121**

$T = 10, Q = 300, \mu = 50$

Table 6. Total service delay under various server operating options and μ

Policy / μ	(30,0)	(30,1)	(30,2)	(30,3)	(31,0)	(31,1)	(31,2)	(32,0)	(32,1)	(33,0)
45	2.725	3.822	0.783	0.682	2.013	1.626	0.248	1.443	0.874	0.907
46	2.980	4.845	1.162	0.773	2.126	1.909	0.371	1.824	1.123	1.225
47	3.442	6.240	2.105	1.248	2.600	3.093	0.576	2.133	2.011	1.372
48	4.394	10.761	3.056	2.130	3.524	5.084	0.853	2.667	2.195	1.558
49	5.136	20.096	4.252	3.764	3.911	10.564	1.926	2.980	4.992	1.821
50	5.851	24.431	6.395	5.835	4.910	12.541	2.332	3.613	4.975	2.372
51	8.022	44.580	9.265	8.503	6.016	25.709	3.257	4.580	9.382	2.687
52	8.702	88.432	15.508	15.265	7.111	46.186	6.225	5.279	17.292	3.570
53	10.841	133.983	22.533	25.518	8.731	79.464	9.732	6.613	22.228	4.139
54	13.043	271.037	49.505	49.374	11.192	150.445	18.136	8.131	43.494	5.650
55	16.586	386.094	70.409	84.132	13.739	221.491	32.199	10.534	108.077	16.585
56	19.989	702.266	168.334	158.982	15.830	393.622	67.099	10.922	184.381	7.328
57	25.078	1004.577	261.393	258.404	19.125	708.930	163.205	13.702	304.905	9.235
58	32.133	1286.36	456.361	545.537	23.354	1037.371	291.988	17.525	453.810	11.232
59	38.210	1900.267	689.005	752.713	28.161	1598.56	406.146	21.718	781.064	13.537

$T = 10, Q = 300, \sigma = 10$

The expected number of tossed services at time $A_{.j}$ is calculated as follows.

$$\begin{aligned} &\sum_{i} \mathrm{P}\left(t_{i(j-1)} > A_{i(j-1)}\right) \\ &= \sum_{i} \mathrm{P}\left(t_{i(j-1)} > \left(\frac{Tm}{Q}\right)\right) \\ &= m \cdot \Phi\left(\frac{\ln\left(\frac{Tm}{Q}\right) - \mu'}{\sigma'}\right) \end{aligned} \tag{3}$$

where μ' and σ' are the mean and standard deviation, respectively, of the service time's natural logarithm.

If the number of tossed residual services is much higher than the number of standby servers, long queue of residual services will be formed. On the other hand, if regular servers finish almost all services within a time without leaving much of residual services to be processed, standby servers idle time will increase.

4.5 Sensitivity of the performance measures to the cost ratio

In the next experiment, we modify the assumptions of our simulation model by changing the cost ratio of server operating cost to delay penalty cost. We observed that the operating option resulting with the lowest total cost varies along with the changes the operating cost (c_o) and delay penalty cost (c_p). c_o was adjusted from \$100 to \$200 interval while c_p was adjusted from \$0 to \$200 in an interval of \$25. We set μ and σ as 50 minutes and 10 minutes here. Table 5 shows the best server operating option under 45 different operating conditions.

If we look at the Table 4, possible best options can be (30,0), (31,0), (32,0), and (31,2). If c_o is disproportionately large, (30,0) is likely to be the best option, while c_p is, (31,2) is likely to be the best option. This means in the region where the mixed policy produces the lowest service delay, lower the cost ratio, more likely the mixed policy is prior to the base policy.

Table 7 Server operating option with the lowest total cost under various c_o and c_p

c_o \ c_p	**0**	**25**	**50**	**75**	**100**	**125**	**150**	**175**	**200**
100	(30,0)	(30,0)	(30,0)	(30,0)	**(31, 2)**	**(31, 2)**	**(31, 2)**	**(31, 2)**	**(31, 2)**
125	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	**(31, 2)**	**(31, 2)**	**(31, 2)**	**(31, 2)**
150	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	**(31, 2)**	**(31, 2)**	**(31, 2)**
175	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	**(31, 2)**	**(31, 2)**	**(31, 2)**
200	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	**(31, 2)**	**(31, 2)**

$T = 10, Q = 300, \mu = 50, \sigma = 10$

Chapter 5 Conclusion

In this paper, we compared the performance of two server operating polices under the appointment system. One is to operate every single server as a regular server and the other is to set aside a portion of servers as standby servers. A simulation model was derived to examine number of different server operating options under two different policies. Results show that the mixed policy with standby servers can be effective when the number of standby servers is enough to deal with the tossed residual services as well as when there are enough tossed residual services to utilize the standby servers. We also analyzed the cost condition under which the service provider can benefit from adopting the mixed policy with the standby servers as well. Even if the mixed policy produces lower service delay, when the cost ratio of server operating cost to delay penalty cost is disproportionately high, service managers are better to adopt base policy over mixed policy.

The findings of the current study are not easy to generalize as the data source is restricted to the focal company. Therefore, it would be worthwhile to examine the impact of the server operating policies with the presence of standby servers by using more empirical data.

Our immediate extension that we plan to pursue is the differentiation of cost factors such as cost difference between regular and standby servers, and penalty cost difference between start delays and restart delays. Another extension would be the inclusion of other types of variability (e.g., daily demand, types of services). It would be also interesting to assess the performance of the appointment system with other types of measures such as server idle time and out-of-operating hours. Customers behavior such as punctuality or no-shows for the service can be considered as well. Incorporating the impact of those factors on service firm's server operating policy might be a promising direction for future research.

Bibliography

Bailey, N. T. J. (1952). "A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times." *Journal of the Royal Statistical Society*. Series B (Methodological): 185-199.

Bechtold, S. E. (1991). “Optimal work-rest schedules with a set of fixed-duration rest periods.” *Decision Sciences*, **22**(3): 157-170.

Bielen, F. and N. Demoulin (2007). "Waiting time influence on the satisfaction-loyalty relationship in services." *Managing Service Quality* **17**(2): 174-193.

Bracken, J., Calkins, J., Sanders, J., & Thesen (1985). A. “A strategy for adaptive staffing of hospitals under varying environmental conditions.” *Health Care Manageaent Review* **10**(4): 43-53.

Cayirli, T. and E. Veral (2003). "Outpatient scheduling in health care: a review of literature." *Production and Operations Management* **12**(4): 519-549.

Cayirli, T., E. Veral, et al. (2006). "Designing appointment scheduling systems for ambulatory care services." *Health Care Management Science* **9**(1): 47-58.

Christie, P. M. J. and R. R. Levary (1998). "The use of simulation in planning the transportation of patients to hospitals following a disaster." *Journal of medical systems* **22**(5): 289-300.

Davis, M. M. and J. Heineke (1998). "How disconfirmation, perception and actual waiting times impact customer satisfaction." *International Journal of Service industry Management* **9**(1): 64-73.

Denton, B. and D. Gupta (2003). "A sequential bounding approach for optimal appointment scheduling." *IIE transactions* **35**(11): 1003-1016.

Edward E. Madden and Jr. P.H. (1976). "A Manual Centralized Outpatient Appointment System." *Hospital Topics* **54**(3): 48-52.

Elhafsi, M. (2002). "Optimal leadtimes planning in serial production systems with earliness and tardiness costs." *IIE transactions* **34**(3): 233-243.

Elliott, N. W. (1990). "Models for determining estimated start times and case orderings in hospital operating rooms." *IIE transactions* **22**(2): 143-150.

Gupta, D. and B. Denton (2008). "Appointment scheduling in health care: Challenges and opportunities." *IIE transactions* **40**(9): 800-819.

Gupta, S. (2011). "A Framework to Span Airport Delay Estimates Using Transient Queuing Models." *Operation Research Center*.

Hensley, R. L. and J. Sulek (2007). "Customer satisfaction with waits in multi-stage services." *Managing Service Quality* **17**(2): 152-173.

Ho, C. J. and H. S. Lau (1992). "Minimizing total cost in scheduling

outpatient appointments." *Management Science*: 1750-1764.

Jansson, B. (1966). "Choosing a good appointment system-a study of queues of the type (d, m, 1)." *Operations Research*: 292-312.

Javel, Y., D. Riopel, et al. (2010). "Measurement of clients' satisfaction on appointment assignment." *International Journal of Mathematics in Operational Research* **2**(5): 634-655.

Jones, P. and E. Peppiatt (1996). "Managing perceptions of waiting times in service queues." *International Journal of Service industry* Management 7(5): 47-61.

Klassen, K. J. and T. R. Rohleder (1996). "Scheduling outpatient appointments in a dynamic environment." *Journal of Operations Management* **14**(2): 83-101.

Klassen, K. J. and R. Yoogalingam (2009). "Improving performance in outpatient appointment services with a simulation optimization approach." *Production and Operations Management* **18**(4): 447-458.

Lawrence, W. R. and R. Rachel (2003). "Scheduling doctors' appointments: optimal and empirically-based heuristic policies." *IIE transactions* **35**(3): 295-307.

Liao, C. J., C. D. Pegden, et al. (1993). "Planning timely arrivals to a stochastic production or service system." *IIE transactions* **25**(5): 63-73.

Mabert, V. A. and C. A. Watts (1982). "A simulation analysis of tour-shift construction procedures." *Management Science* **28**(5): 520-532.

Maister, D. H. (1985). "The psychology of waiting lines." *The service encounter* **1**: 13-23.

May, J. H., D. P. Strum, et al. (2000). "Fitting the Lognormal Distribution to Surgical Procedure Times." *Decision Sciences* **31**(1): 129-148.

Mok, S. K. and J. G. Shanthikumar (1987). "A transient queueing model for Business Office with standby servers." *European journal of operational research* **28**(2): 158-174.

Najmuddin, A., I. Ibrahim, et al. (2010). "A simulation Approach: Improving patient waiting time for multiphase patient flow of obstetrics and gynecology department (O and G Department) in local specialist centre." *WSEAS Trans. Math* **10**: 778-790.

Nie, W. (2000). "Waiting: integrating social and psychological perspectives in operations management." *Omega* **28**(6): 611-629.

O'Keefe, R. M. (1985). "Investigating outpatient departments: implementable policies and qualitative approaches." *Journal of the Operational Research Society:* 705-712.

Pegden, C. D. and M. Rosenshine (1990). "Scheduling arrivals to queues." *Computers & operations research* **17**(4): 343-348.

Pichitlamken, J., Deslauriers, A., L'Ecuyer, P., Avramidis, A.N. (2003). “Modelling and simulation of a telephone call center.” *Simulation Conference, 2003. Proceedings of the 2003 Winter,* **2**: 1805-1812.

Rising, E. J., R. Baron, et al. (1973). "A systems analysis of a university-health-service outpatient clinic." *Operations Research*: 1030-1047.

Sabria, F. and C. F. Daganzo (1989). “Approximate expressions for queuing systems with scheduling arrivals and established service order.” *Transportation Science* **23**(3): 159-165.

Stein, W. E. and M. J. Côté (1994). "Scheduling arrivals to a queue." *Computers & operations research* **21**(6): 607-614.

Vanden Bosch, P. M. and D. C. Dietz (2000). "Minimizing expected waiting in a medical appointment system." *IIE transactions* **32**(9): 841-848.

Vermeulen, I. B., S. M. Bohte, et al. (2009). "Adaptive resource allocation for efficient patient scheduling." *Artificial intelligence in medicine* **46**(1): 67-80.

Wang, P. P. (1993). "Static and dynamic scheduling of customer arrivals to a single-server system." *Naval Research Logistics (NRL)* **40**(3): 345-360.

Wang, P. P. (1997). "Optimally scheduling N customer arrival times for a single-server system." *Computers & Operations Research* **24**(8): 703-716.

Welch, J. (1964). "Appointment systems in hospital outpatient departments." *Operational Research Quarterly* **15**(3): 224-232.

Wijewickrama, A. and S. Takakuwa (2005). “Simulation analysis of appointment scheduling in an outpatient department of internal medicine”, *Simulation Conference, 2005. Proceedings of the 2005 Winter*: 2264-2273.

Wright, P. D., K. M. Bretthauer, et al. (2006). "Reexamining the nurse scheduling problem: Staffing ratios and nursing shortages." *Decision Sciences* **37**(1): 39-70.

S. M. Ross (2002). Simulation. Academic Press, San Diego.

초 록

예약제 서비스 하에서의 서버 운용 정책에 관한 시뮬레이션 접근법

오늘날 고객들이 서비스를 받는 데에 있어서 시간이라는 요소를 중요하게 생각함에 따라, 서비스 기업들은 다양한 전략을 통해 고객들의 대기 시간을 줄이고자 노력하고 있다. 적시에 서비스를 제공하기 위하여 널리 쓰이고 있는 방법 중에 하나는 예약제를 도입하여 고객들의 도착시간을 정해두는 것이다. 예약제를 다루는 최근의 많은 연구들은 최적화 기법, 휴리스틱 기법, 시뮬레이션 기법 등을 통하여, 예약제 하에서 몇 명의 서버를 두는 것이 고객의 대기시간을 가장 적게 할 수 있는지에 대하여 다루어왔다. 본 연구에서는 예약제 하에서 기존의 일반서버 뿐만이 아니라, 대기서버라는 새로운 종류의 서버형태를 도입한 서버운영정책을 제시하고 기존의 정책과의 성능을 비교하고자 한다. 대기서버는 일반서버가 다음 예약고객에게 제 시간에 서비스를 제공할 수 있도록, 현 고객에 대한 초과근무량을 처리한다. 본 연구는 대상기업에서의 서비스 제공 프로세스를 대기서버가 있는 경우와 없는 경우의 두 정책을 두 가지의 다른 시뮬레이션 모델로 개발한다. 또한, 다양한 환경적 요소

들이 변함에 따라 둘 중 어떤 정책이 우위를 점하는지를 분석하여, 대상기업이 고객의 대기시간으로부터 발생한 비용과 서버 운용비용을 고려할 때 어떠한 정책을 채택해야 하는지를 제시한다.

키워드: 예약제 서비스, 대기서버, 고객 대기 시간, 서비스 시간
학번: 2010-23377