韓國經營科學會誌
第33권 第2號
2008年 6月

고객 도착 정보를 이용한 서비스 차별화에 관한 점근적 분석*

최 기 석**

Asymptotic Analysis on Service Differentiation with Customer Arrival Information*

Ki-Seok Choi**

■ Abstract ■

An interesting problem in capacitated supply chains is how to guarantee customer service levels with limited resources. One of the common approaches to solve the problem is differentiating service depending on customer classes. High-priority customers receive a better service at the cost of low-priority customers' service level. One of common criteria to determine a customer's priority in practice is whether he has made a reservation before arrival or not. Customers with a reservation usually receive service based on the time the reservation was made. We examine the effect this advance information of customer arrival has on customer service levels. We show the trade-offs between the leadtime and other system characteristics such as the proportion of high-priority customers, when the service level is high. We also suggest how to differentiate service using the asymptotic ratio of the service levels for both types of customers.

Keyword : Service Level, Service Differentiation, Customer Arrival Information, Asymptotic Analysis, Priority Queue

1. Introduction

In order to provide the same level of service to all customers, a service provider needs to improve his service quality by increasing capacity or reducing service time variance. With a given capacity limit, the service provider sacrifices the system utilization to improve customer service

논문접수일 : 2007년 12월 15일 논문게재확정일 : 2008년 05월 29일
논문수정일(1차 : 2008년 03월 10일)
* 이 연구는 한국외국어대학교 교내학술연구비의 지원에 의해 이루어진 것임.
** 한국외국어대학교 산업경영공학부 조교수

level. This can be easily demonstrated by a simple example. Suppose the service provider is committed to his customers that the probability a customer order is delivered within a target leadtime t_0 is at least $(1-\beta)\times 100\%$. This service provider's customer service level is specified by (t_0, β) and this type of customer service level measure is called the *fill rate*, defined as the proportion of customers whose service request have been satisfied within a certain amount of time. Assumed that customer orders arrive according to a Poisson process with rate λ, the time to process one order has exponential distribution with rate μ, and the customer orders are processed one at a time in a first-in-first-out manner, the example is represented by an $M/M/1$ queue. The total time in the system R has a steady-state distribution given by

$$\Pr\{R \le t\} = 1 - e^{-(\mu-\lambda)t} \quad \text{for} \quad t \ge 0.$$

When $\mu = 1$, the target leadtime is at least

$$t_0 = \frac{-\log\beta}{1-\rho}$$

where $\rho = \lambda/\mu = \lambda$ is the system utilization. Once the customer service level is decided, it determines the system utilization; $\rho = 1 + (\log\beta)/t_0$. Thus, it is impossible to improve customer service level and system utilization at the same time in this case.

One way to achieve the objective of improving service level and utilization together with limited capacity is to give different priorities to customers. While the service provider cannot improve service level for all the customers, he may differentiate service and provide a part of customers with a better service level than the other customers. In return, the service provider may charge more service fee for a high service level. Express service in transportation and expediting orders in manufacturing systems are examples of differentiating service based on the customer priority.

In this paper, as a criterion for customer priority, we consider availability of advance arrival information. Customers are classified into two groups depending on whether they notify their arrival in advance. When a customer with advance notice finally arrives to the system, he has a higher priority than the waiting customers who have arrived without advance notice. However, it is assumed that even a high-priority customer cannot be served prior to the customers who arrived at least K time units earlier than him. Within the same priority group, the customers are served in the order of their arrivals. More specifically, the priority scheme can be described as follows. The high-priority customers notify their arrival K time units earlier. When they eventually come to the system to be served, their position in the waiting queue is decided not by the time they arrive but by the time they notified their arrival. The high-class customer has priority over the low-priority customers who have arrived earlier but after the high-class customer made his arrival notification. A common method for a customer to notify his arrival is to make a reservation. In the paper, we refer to high-priority customers as those who have made a reservation and low-priority customers as those who have not.

There is little existing literature concerning service differentiation with advance arrival information. One of notable research streams is the study of queueing systems with scheduled arri-

vals. Scheduled arrivals are related to our study in the sense the service order is determined by the time on the arrival schedule not by the time when a customer actually arrives. Mercer [8] considers a queueing system in which customers are scheduled to arrive during equal time intervals and must arrive within the scheduled interval. He analyzes the waiting time distribution when the service time is exponential. Sabria and Daganzo [9] consider a model where late customers still join the system and the service may take place in an order different from that of arrivals. They provide light traffic approximations of the expected waiting time for general lateness and service distributions. Doi et al. [3] study a queueing model with general scheduled interarrival times and exponential distributions for delay and service time. The customer arriving after the next scheduled arrival must leave the system immediately. For this model, they provide the steady-state distribution of waiting time.

While the scheduled arrivals are useful to represent practical problems such as berth space scheduling at a sea port, little has been known about the system performance such as customer waiting time for the cases where a late customer can get service even after other customers have arrived. Besides, all the customers are required to make a schedule before arrival. In this paper, we consider a service policy which has not been considered in the scheduled arrival literature before. The service system accepts every arriving customer, whether he has set up an arrival schedule before the arrival or not. But, the customer arriving without a schedule will receive a low priority of service. Under this policy, we provide a method to differentiate the service level for high- and low-priority customers through the asymptotic analysis of waiting time distributions.

Another research stream related to service level guarantee is concerning quotation of leadtime. Wein [12] studies due-date setting and priority sequencing problems in a multi-class queueing system. He compares a number of policies through simulation when the objective is to minimize the weighted average of the time between an order arrival and the quoted due date with a constraint on a long-run average tardiness. Duenyas and Hopp [4] analyze a model where the customer demand may depend on the quoted leadtime. They prove the optimality of different control policies of accepting orders to maximize profits when the orders are served in a first-come-first-served manner. They also give conditions under which the earliest-due-date policy is optimal. Hopp and Sturgis [6] suggest a method for quoting manufacturing due dates to achieve a target customer fill rate. Using simulation, they show that their method of determining leadtimes as a function of work in process predicts accurately the leadtime.

In the above literature, the estimated leadtime is known to a customer when he places an order. The quoted leadtimes change dynamically even for the customers in the same class. In this paper, we analyze customer service levels with a constant leadtime for each customer class, with which a target order fill rate is guaranteed over a long-term range.

The remainder of the paper is organized as follows. In the next section, we describe our two-priority class model. In Section 3, we provide the analysis for approximate service levels. Section 4 demonstrates the asymptotic results and trade-offs regarding service levels using numerical examples. We end the paper with concluding remarks.

2. Two-Priority Class Model

We consider a continuous time system where a single server with limited capacity serves customer orders. It includes production systems operating under a make-to-order policy and other general service systems that can be modelled by a single-server queueing system. In our model, customers are classified into two priority classes depending on whether they have made a reservation before arriving to the system. We use the following notations to explain the two-priority class system.

A_n : interarrival time between the n-th and $(n+1)$-th arrivals

p : probability that a customer has a reservation

λ : total customer arrival rate

K : time interval between making a reservation and arriving to the system

O_n : service order of the n-th arriving customer

R_n : leadtime of the n-th arriving customer

W_n : waiting time in the queue of the n-th arriving customer

B'_n : service time of the n-th departing customer

W'_n : waiting time in the queue of the n-th service position

For a random variable X, ϕ_X denotes its cumulant generating function : $\phi_X(\theta) = \log E[e^{\theta X}]$. Let A and B denote random variables following the same distribution as customer interarrival time and service time, respectively. We define γ as follows;

$$\phi_B(\gamma) + \phi_A(-\gamma) = 0. \quad (1)$$

The solution to the above equation exists for almost all commonly used distributions [5].

The interarrival time sequence $\{A_n : n = 0, 1, \cdots\}$ is assumed to be i.i.d. and follow an exponential distribution with rate λ. Each incoming customer has the identical and independent probability $0 < p < 1$ that he has a reservation. Thus, the customers with and without a reservation arrive according to Poisson processes with rate $p\lambda$ and $(1-p)\lambda$, respectively.

A customer arriving with a reservation is assumed to make the reservation $K < +\infty$ time units before the arrival. The service starting time for the customer is determined based on the reservation time. In other words, he is served as if he arrived at the reservation time. Thus, the customer with a reservation has a priority over non-reservation customers who have arrived before him but after his reservation time. However, we assume that the customer under service is not interrupted by new arrivals of customers even with a reservation (non-preemptive service policy).

The two-priority scheme can be viewed in another way. Assume that the service provider divides his customers into two classes and differentiates service depending on a certain criteria. For example, a customer may have a higher priority and receive a better service if he pays more service fee. The service provider offers a better service to high-priority customers by adjusting service orders. In this case, high-priority customers can save waiting time as much as K time units. The waiting position of high-priority customers is adjusted that he can save waiting time as much as K time units. The service of low-priority customers may be postponed due to high-priority customer arrivals.

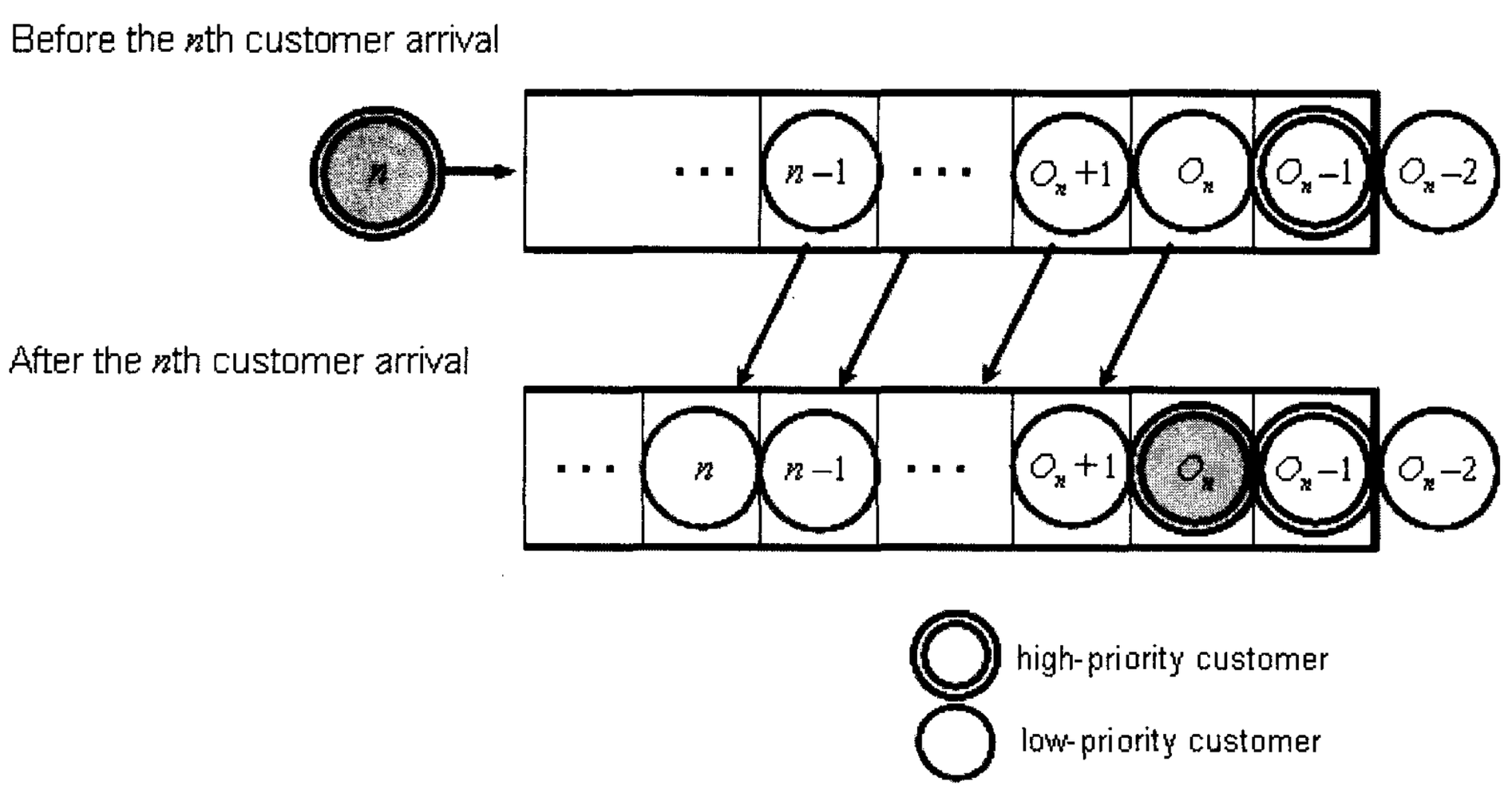
[Figure 1] explains the service order adjust-

ment due to a high-priority customer arrival. A newly arriving high-priority customer, customer n represented by a shaded double circle, may catch up with some low-priority customers waiting in the queue; the service order O_n is earlier than the arrival order n. And the service starting time for low-priority customers represented by a single circle is delayed to serve customer n first. However, the new arrival does not affect the order of another high-priority customer in service order $O_n - 1$ even if the customer has arrived within K time units of the new arrival. We assume that in the same class the customers are served according to their arrival orders.

We use H and L to denote the set of arrival indices of high- and low-priority customers, respectively; $L = \{1, 2, \cdots\} \setminus H$. If $n \in H$, then the n-th arriving customer is in the high-priority class. Otherwise (i.e. $n \in L$), the n-th arriving customer is in the low-priority class.

As mentioned above the service order O_n, which also means the order of departing the system, highly depends on the customer priority. If $n \in H$ then $O_n \leq n$. Otherwise, $O_n \geq n$. Note that the service order O_n is decided upon the arrival for high-priority customers but not for low-priority customers. The low-priority customer's service order depends on the future arrivals of high-priority customers. The service order O_n for $n \in L$ is decided once the customer starts to get served or K time units have past since he arrived.

We use a couple of notations related to service orders for the ease of analysis. For the customer in service order n, we denote his service time by B'_n (i.e. the service time of the n-th arriving customer is B'_{O_n}). We assume that the service time sequence $\{B'_n : n = 1, 2, \cdots\}$ is i.i.d. and independent of the interarrival time sequence $\{A_n : n = 0, 1, \cdots\}$. We denote the waiting time in service position n by W'_n. A service position is a place in the queue where a customer is waiting before getting served. The n-th service position is opened as soon as the n-th customer arrives. It is decided by the customer priority which cus-



[Figure 1] Arrival of a High-Priority Customer

tomer will be served at the newly opened position. However, no matter who eventually takes the position, the n-th position is to get served W'_n times after it is opened. We have the following equation for the waiting time;

$$W'_n = \max_{1 \le k \le n-1}\left\{0, \sum_{i=k}^{n-1}(B'_i - A_i)\right\} \quad (2)$$

with $W'_1 = 0$. Equation (2) is similar to the definition of the waiting time in a regular single-server FIFO queueing system, except that $\{B'_n\}$ is used instead of $\{B_n\}$, a common notation for service times.

Once a high-priority customer catches up one or more low-priority customers, it means he takes a service position which has opened earlier than he arrives (i.e. $O_n < n$) and his waiting time in the queue (W_n) is shorter than that of the service position (W'_{O_n}). Since the O_n-th service position opened when the O_n-th customer arrived, the difference between W_n and W'_{O_n} is equal to the interarrival time between the O_n-th and n-th customers ($\sum_{i=O_n}^{n-1} A_i$). In a similar manner, we can get an expression for waiting time for low-priority customers. In summary,

$$W_n = \begin{cases} W_{O_n} - \sum_{i=O_n}^{n-1} A_i & \text{for } n \in H, \\ W_n + \sum_{i=n}^{O_n-1} B'_i & \text{for } n \in L. \end{cases} \quad (3)$$

3. Service Level Approximations

In this paper, customer service level is measured by the unfill rate, defined to be the probability that a customers's leadtime in steady state is longer than a target leadtime;

$$\Pr_H\{R > t\} := \lim_{n\to\infty} \Pr\{R_n > t | n \in H\}$$

for high-priority customers and

$$\Pr_L\{R > t\} := \lim_{n\to\infty} \Pr\{R_n > t | n \in L\}$$

for low-priority customers.

The leadtime, time elapsed from when a customer arrives until he departures the system, consists of the waiting time in the queue and the service time;

$$R_n = W_n + B'_{O_n}.$$

We assume that the service time is invariant to customer priorities. A customer's priority affects the waiting time in the queue and in turn the leadtime. For the n-th arriving customer, the service order O_n is not later than the arrival order n if the customer is in high-priority class. The high priority over other customers in the queue shortens the customer's waiting time W_n and reduces the leadtime R_n. On the other hand, the service order of a low-priority customer cannot be sooner than the arrival order. Compared with a single-customer-class case, the two-priority queue holds low-priority customers longer in the waiting queue and prolongs their leadtime.

3.1 High-Priority Class

In this subsection, we provide an asymptotic result on the unfill rate of high-priority class customers through the analysis of the waiting time in the queue. In Equation (3), the waiting time W_n is expressed in terms of the service order O_n. In this subsection, we define another notation related to high-priority customer service level. For $n \in H$,

$$N_n := n - \left| \left\{ 1 \le k < n : \sum_{i=k}^{n-1} A_i \le K,\ k \in L \right\} \right|. \qquad (4)$$

A high-priority customer can catch up with only the low-priority customers who have arrived at most K time units earlier than he does. The last term in (4) denotes the number of those low-priority customers. If any such low-priority customers are waiting in the queue, they are served after the high-priority customer. Thus, N_n means the earliest service position that the n-th arriving customer could take when he is in the high-priority class. In other words, for O_n for $n \in H$;

$$N_n \le O_n \le n. \qquad (5)$$

However, the new arrival of a high-priority customer does not change the service order of other high-priority customers. It also does not interrupt the customer under service even if he is in low-priority class. Thus, to get an explicit expression of the service order O_n for $n \in H$, we need to consider not only how many low-priority customers have arrived within the past K time units but also how many low- and high-priority customers have been still waiting in the queue when the n-th customer arrives. Since it would involve additional notations and, above all, it is not required in the development of this paper, the equation for O_n is not given explicitly.

The leadtime R_n of high-priority customers is closely related to the waiting time of service position N_n. When there are many waiting customers in the queue, it is likely that the newly arriving high-priority customer catches up with almost all the low-priority customers who have arrived within K time units. In this case, the service order O_n would be close to N_n. The following result relates the leadtime R_n to the waiting time of service position N_n. All the proofs in this paper can be found in Appendix.

Lemma 1. For $n \in H$, let

$$\overline{R_n} := W'_{N_n} - \sum_{i=N_n}^{n-1} A_i + B'_{O_n} \qquad (6)$$

and

$$\overline{d_n}(t) := e^{\gamma t}\left(\Pr\{R_n > t\} - \Pr\{\overline{R_n} > t\}\right). \qquad (7)$$

If there exists $\gamma > 0$ satisfying Equation (1), then $\overline{d_n}(t)$ converges to 0 uniformly in n as $t \to \infty$.

The first term on the right-hand side of Equation (6) is the waiting time of service position N_n. The second term represents the time elapsed after the service position N_n has opened until the n-th customer arrives to the system. Thus, if $O_n = N_n$, the difference of the two terms $(W'_{N_n} - \sum_{i=N_n}^{n-1} A_i)$ is the waiting time in the queue and $\overline{R_n}$ is equal to the leadtime R_n. If O_n differs from N_n (i.e. $O_n > N_n$), it means that the n-th arriving high-priority customer cannot proceed to the earliest service position N_n. This can happen only when the service position N_n is already served or currently under service. Thus, the waiting time of the newly arriving customer is zero or equal to the remaining service time of the customer under service. In this case, the probability that the customer's leadtime is longer than the target leadtime t would be small.

Lemma 1 suggests a way to relate a customer's leadtime to a random variable $\overline{R_n}$, which is a function of the service position N_n rather than the service order O_n. Using Lemma 1, we obtain the following result on the unfill rate of high-priority customers.

Theorem 1 : If there exists $\gamma > 0$ satisfying Equation (1), then the following equation holds for some constant C that does not depend on K or p;

$$\lim_{t\to\infty} e^{\gamma t}\Pr_H\{R>t\} = Ce^{-\gamma(1-p)K}. \quad (8)$$

Theorem 1 suggests the following approximation of the unfill rate for high-priority customers;

$$\Pr_H\{R>t\} \approx Ce^{-\gamma(t+(1-p)K)}. \quad (9)$$

Note that $(1-p)K$ in the exponent on the right-hand side is mean interarival time $(1/\lambda)$ multiplied by $(1-p)\lambda K$, the expected number of low-priority customers who have arrived within K time units before a high-priority customer arrives. This number is equal to the average number of service positions that a high-priority customer can advance in a long waiting queue. More arrivals of low-priority customers in K time units decrease the approximate unfill rate in Equation (9), and thus, improve the high-priority customer service level. This observation is consistent with the intuition that high-priority customers would receive a better service when the proportion of high-priority customers is small (i.e. large $1-p$) and a customer's priority is valid over a long period (i.e. large K).

3.2 Low-Priority Class

For the analysis of low-priority class customer service level, we first introduce a new variable $\{M_n\}$, which is similar to $\{N_n\}$ for high-priority class customers. For the n-th arriving customer who has low priority (i.e. $n \in L$),

$$M_n := n + \left|\left\{k > n : \sum_{j=n}^{k-1} A_j \le K,\ k \in H\right\}\right|. \quad (10)$$

The meaning of M_n is the last service position that the n-th arriving customer could take when he is in the low-priority class. The newly arriving customer has a lower priority than the high-priority customers who arrive at most K time units later. Such high-priority customers are served prior to the low-priority customer. However, high-priority customer arrivals do not delay the service of the low-priority customer after the low-priority customer starts to be served.

The leadtime R_n of low-priority customers is closely related to the waiting time of service position n. When there are many waiting customers in queue so that the newly arriving low-priority customer should wait more than K time units in the queue, all the high-priority customers who arrive within next K time units will be served prior to the low-priority customer. The following result relates the leadtime R_n to the waiting time of service position M_n.

Lemma 2 : For $n \in L$, let

$$\widehat{R}_n := W'_n + \sum_{i=n}^{M_n-1} B'_i + B'_{O_n} \quad (11)$$

and

$$\hat{d}_n(t) := e^{\gamma t}\left(\Pr\left\{\widehat{R}_n > t\right\} - \Pr\{R_n > t\}\right). \quad (12)$$

If there exists $\gamma > 0$ satisfying Equation (1) and $\phi_B(2\gamma) < \infty$, then $\hat{d}_n(t)$ converges to 0 uniformly in n as $t \to \infty$.

The first term on the right-hand side of Equation (11) is the waiting time of service position n. The second term represents the time elapsed after the service position n has opened until the

n-th customer starts to receive service. From Equation (3), $\widehat{R}_n$ is equal to the leadtime R_n if $O_n = M_n$. Otherwise, the difference between $\widehat{R}_n$ and R_n is the sum of service times of the customers served in service position O_n through $M_n - 1$; $\widehat{R}_n - R_n = \sum_{i=O_n}^{M_n-1} B_i$.

Lemma 2 suggests a way to relate a low-priority customer's leadtime to a random variable $\widehat{R}_n$, which is a function of the service position M_n rather than the service order O_n. Using Lemma 2, we obtain the following asymptotic result on the unfill rate of low-priority customers.

Theorem 2 : If there exists $\gamma > 0$ satisfying Equation (1) and $\phi_B(2\gamma) < \infty$, then the following equality holds with the same constant C as in Equation (8);

$$\lim_{t\to\infty} e^{\gamma t}\Pr_L\{R>t\} = Ce^{\gamma pK}.$$

Theorem 2 suggests the approximation of the unfill rate for low-priority customers;

$$\Pr_L\{R>t\} \approx Ce^{-\gamma(t-pK)}. \quad (13)$$

Note that pK in the exponent on the right-hand side is mean interarival time ($1/\lambda$) multiplied by $p\lambda K$, the expected number of high-priority customers who arrive within K time units after a low-priority customer has arrived. This number is equal to the average number of service postponements that a low-priority customer would experience in a long waiting queue. More arrivals of high-priority customers in K time units increase the approximate unfill rate in Equation (13), and thus, deteriorate the low-priority customer service level. This observation is consistent with the intuition that low-priority customer would receive a better service when the proportion of high-priority customers is small (i.e. small p) and a customer's priority is valid only over a short period (i.e. small K).

4. Trade-Offs

In the previous section, Theorems 1 and 2 suggest the approximation of the unfill rate for high- and low-priority customers as in Equation (9) and (13). The set of target leadtime and parameter K and p which lead to target customer service levels (i.e. $\Pr_H\{R>t_H\} = \delta_H$ and $\Pr_L\{R>t_L\} = \delta_L$) can be approximated by the solution to $C\exp(-\gamma(t_H + (1-p)K)) = \delta_H$ and $C\exp(-\gamma(t_L - pK)) = \delta_L$. These equations reduce to

$$t_H = -(1-p)K + \log(C/\delta_H)/\gamma$$

for high-priority and (14)

$$t_L = pK + \log(C/\delta_L)/\gamma$$

for low-priority customers.

From the equations above, we can observe that there exists a liner relationship between the target leadtime (t_H, t_L) and either parameter p or K. For example, an increase of the proportion of high-priority customers by Δp extends the target leadtime t_H by $K\Delta p$ for the same high-priority customer service level. It also results in the same increment in t_L for low-priority customers. Thus, when the fill rates are high, varying the target leadtime and the high-priority-customer proportion according to the linear trade-off rule

$$\Delta t_H = K\Delta p \quad \text{and} \quad \Delta t_L = K\Delta p \quad (15)$$

results in little change in the fill rates.

We demonstrate how well the linear trade-off approximates the relationship between t and p. We first set $p=0.1$ and run a simulation to identify the target leadtime for which the fill rate is 80% or higher. We then get the target leadtime, t_H and t_L for different values of p according to the linear trade-offs in Equation (15). For each combination of (p, t_H, t_L), we conduct simulation to observe the actual fill rates for high- and low-priority customers. Plotting the fill rates generates a curve for each customer type. When the curve is nearly a flat line the linear trade-offs work well.

In [Figure 2], we test three service time distributions with different variations. We study the trade-offs between the target leadtime and the proportion of high-priority customers when the fill rate is around 80%, 85%, 90%, 95% and 99%. All cases have $\lambda=0.9$, $E[B]=1$, $K=5$. We use c_B^2 to denote the squared coefficient of variation for service time B.

The observed fill rate curves in [Figure 2] are close to flat for both high- and low-priority customers when the fill rate is 90% or higher, regardless of service time distributions. The linear trade-off approximates very well the relationship between leadtime t and high-priority customer proportion p which yield the same fill rate. Even for lower fill rates (80% and 85%), the linear trade-off still seems to work well when service time has a large variation ($c_B^2 \geq 1$ as in (a) and (c) of [Figure 2]). Since Theorem 1 and 2 are limiting results with respect to the target leadtime t, the trade-offs based on the results generally hold better for low-priority customers, who experience longer leadtime.

The approximation in Equation (14) leads to another linear trade-off involving the target leadtime t. An increase of the advance notice period K reduces the target leadtime for the high-priority customers and extends it for low-priority customers;

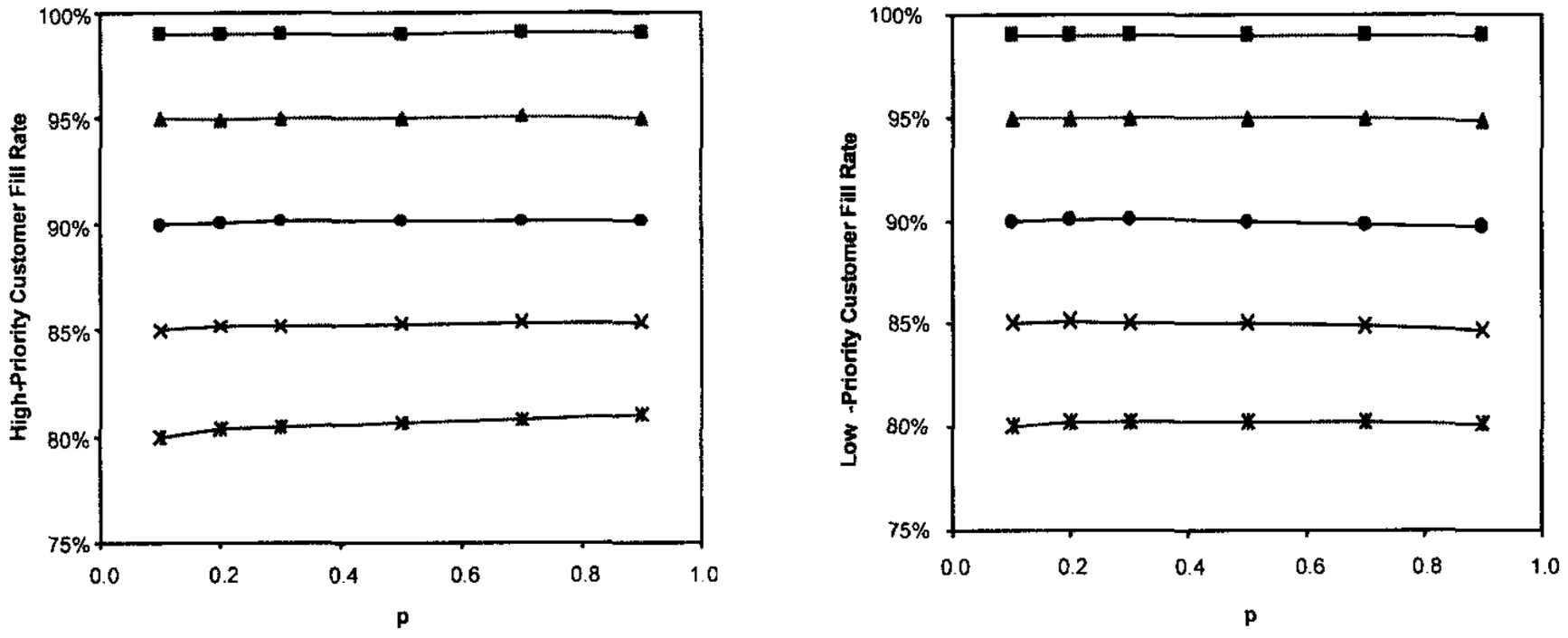
$$\Delta t_H = -(1-p)\Delta K \quad \text{and} \quad \Delta t_L = p\Delta K. \tag{16}$$

We test the linear trade-off in Equation (16) in a similar way. We first set $K=1$ and run a simulation to identify the target leadtime for which the fill rate is 80% or higher. We then calculate the target leadtime, t_H and t_L for different values of K according to the linear trade-offs in Equation (16).

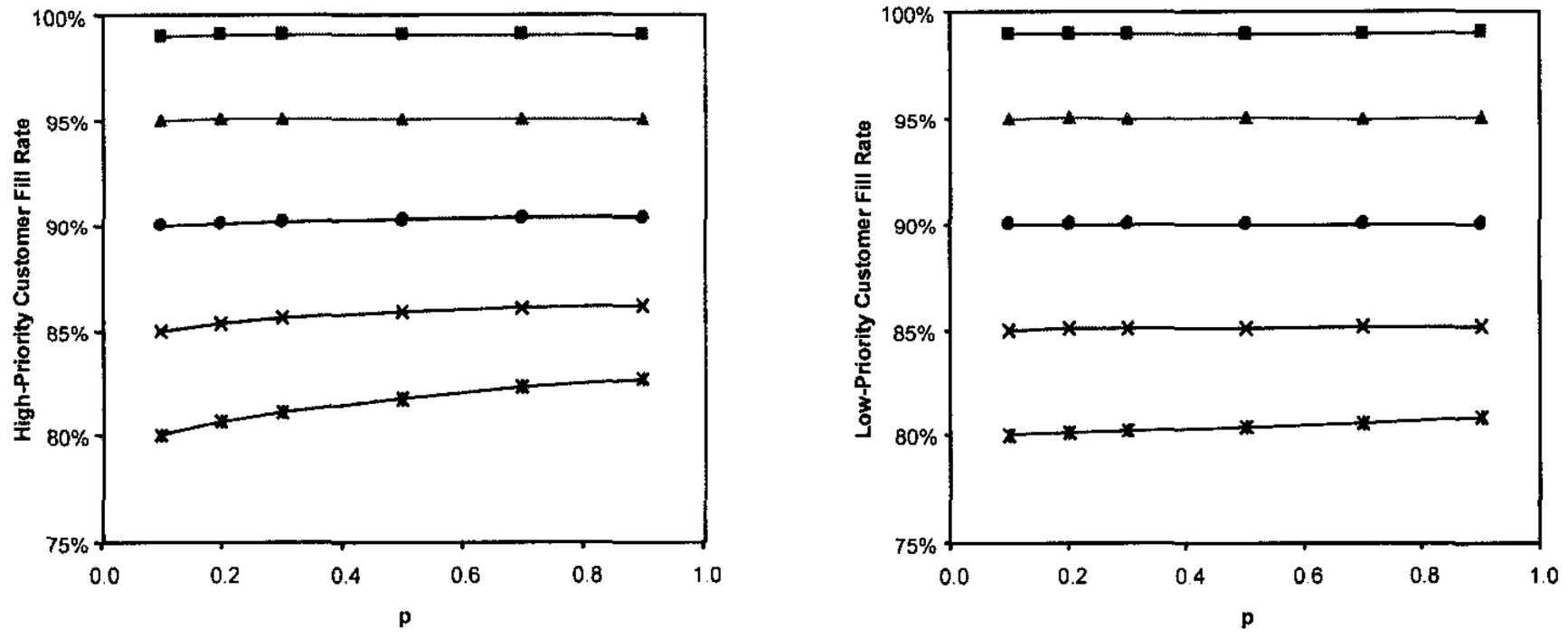
In [Figure 3], the observed fill rate curves are close to flat when the fill rate is 90% or higher and service time has a large variation ($c_B^2 \geq 1$). When the fill rate is low or the service time variation is small ((b) of [Figure 3]), the approximate trade-off in Equation (16) does not work well. In such cases, the leadtime for a certain target service level is relatively short. In the range of short leadtime, the liner trade-off rule would not work well because it is based on the asymptotic results which hold for long leadtime. In (b) of [Figure 3], the trade-off does not work well especially for high-priority customers because an increase in K reduces the target leadtime for high-priority additionally. On the other hand, as K increases so does the target leadtime for low-priority customers. [Figure 3] shows that the linear trade-off works well for low-priority customers with all distributions when the fill rate is 90% or higher.

Now we consider trade-offs between service levels for high-and low-priority customers. Without improving service time (increasing service rate and/or reducing service time variation), the service

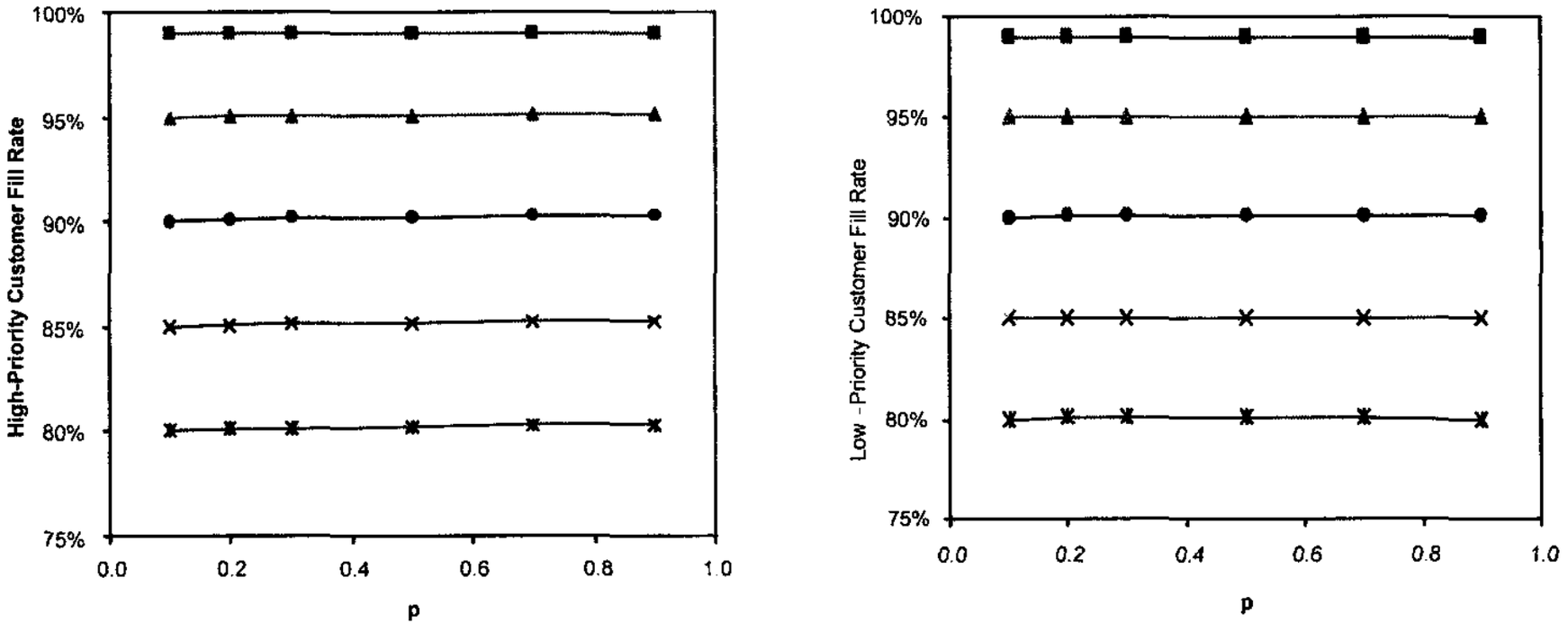
provider can improve the service level for one customer class only at the cost of the service level for the other customer class. Using the asymptotic results on the unfill rates, we examine how the service provider can control the service level difference between the two customer classes. From Theorem



(a) Exponential Service Time



(b) Gamma Service Time with $c_B^2 = 0.5$

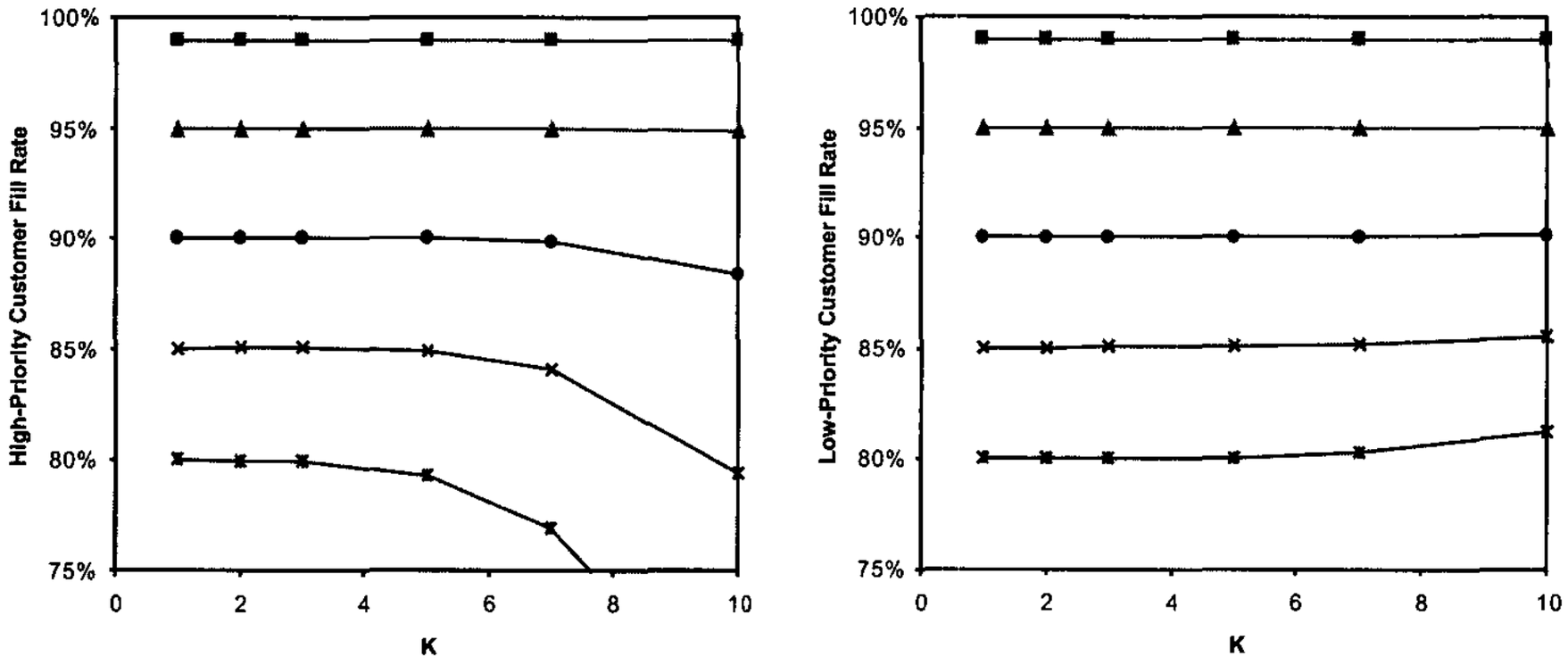


(c) Gamma Service Time with $c_B^2 = 2$

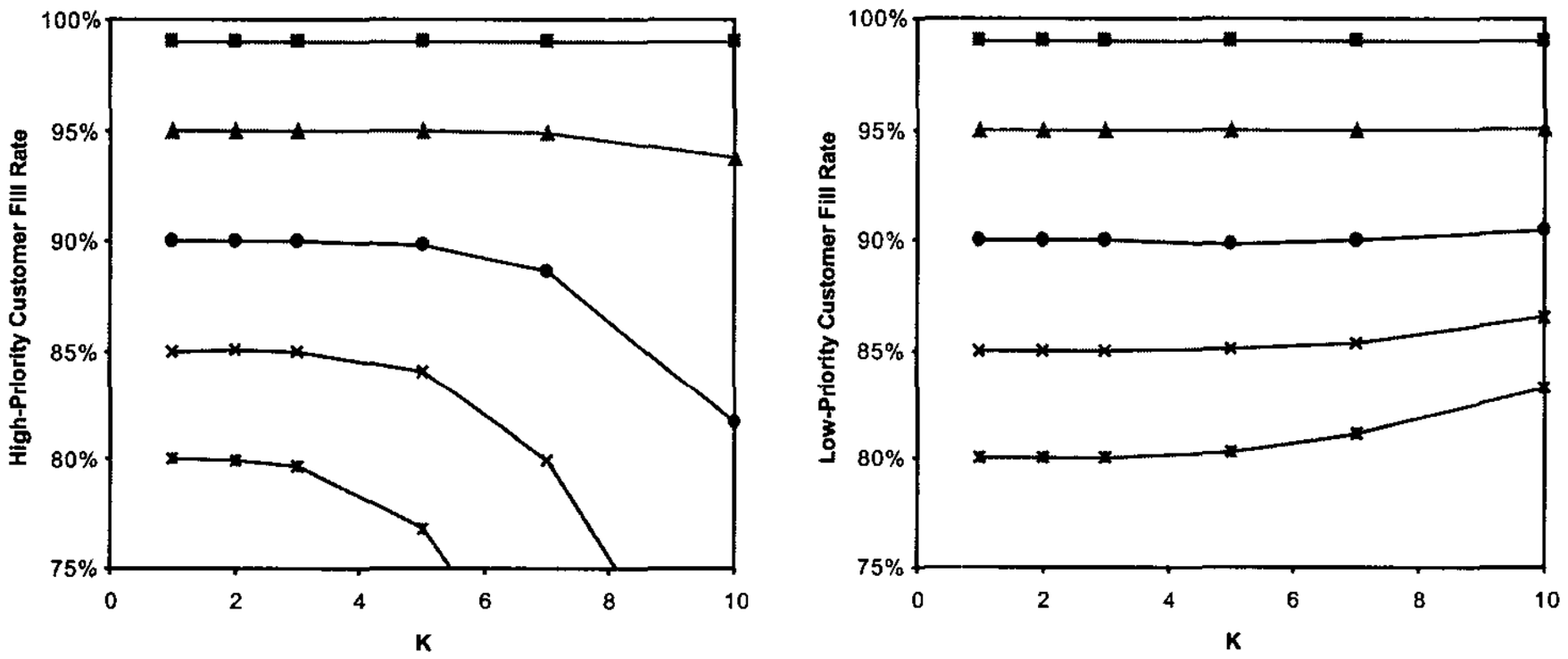
[Figure 2] Trade-off between Target Leadtime and p

1 and 2, we can immediately get the following result regarding the ratio of the unfill rates.

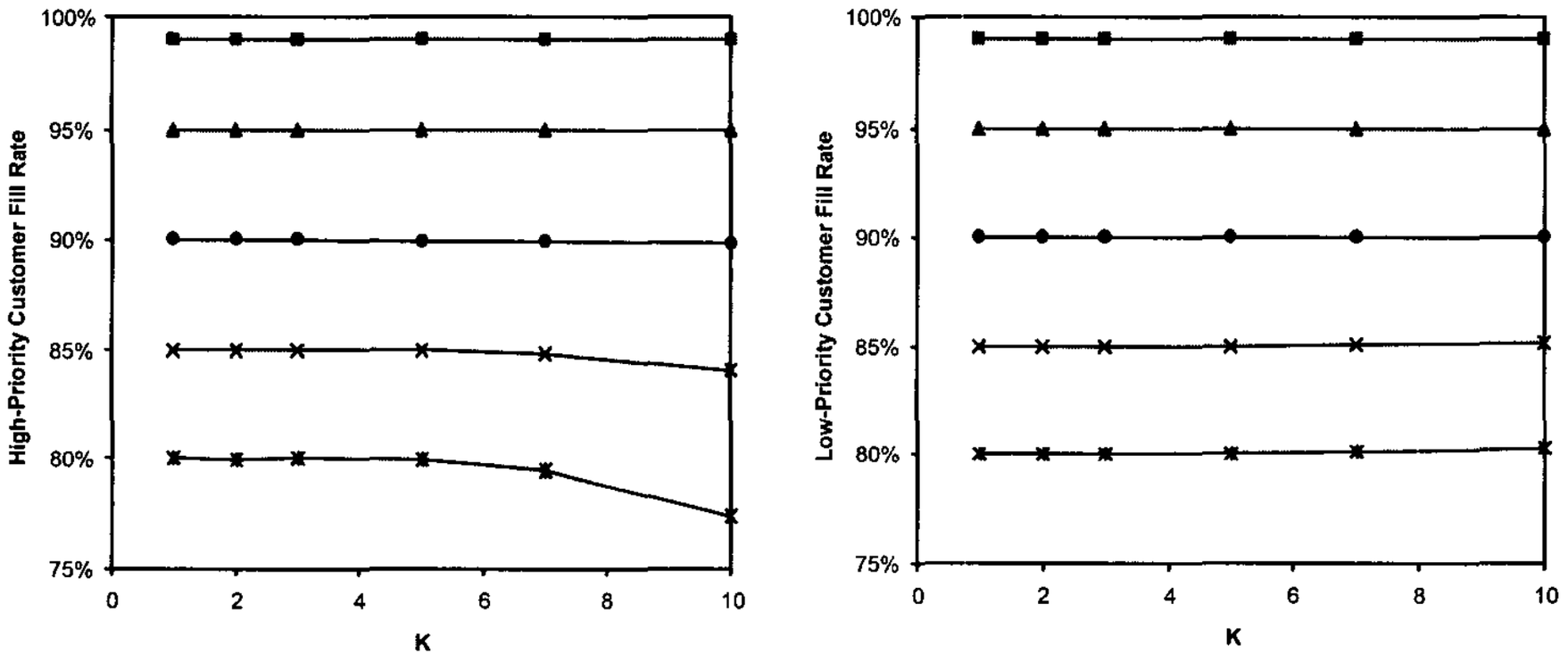
Corollary 1. If there exists $\gamma > 0$ satisfying Equation (1), then



(a) Exponential Service Time



(b) Gamma Service Time with $c_B^2 = 0.5$



(c) Gamma Service Time with $c_B^2 = 2$

[Figure 3] Trade-off between Target Leadtime and K

$$\lim_{t\to\infty}\frac{\Pr_H\{R>t\}}{\Pr_L\{R>t\}}=e^{-\gamma K}. \quad (17)$$

The right-hand side of Equation (17), called the asymptotic ratio, gives a hint on how much better service, compared with the low-priority customers, the high-priority customers receive when the fill rates for both type customers are high. For the same target leadtime, the unfill rate would be lower for high-priority customers than for low-priority customers. As expected the asymptotic ratio is no more than 1, i.e. the fill rate for high-priority customers is no less than the fill rate for low-priority customers. The asymptotic ratio becomes 1 only when $K=0$, which indicates that there is virtually no priority difference between customers. Note that the asymptotic ratio $e^{-\gamma K}$ does not include p, the proportion of high-priority customers. The trade-offs in Equation (15) show that the change in p has the same impact on the target leadtime for both customer classes. Corollary 1 implies that the proportion of high-and low-priority customers has little influence on the relative service levels between customer classes when the fill rate is high.

We conduct numerical experiments to verify the results on the ratio of the unfill rates using $\lambda=0.9$, $\mathrm{E}[B]=1$, $K=5$ and two different combinations of customers ($p=0.3, 0.8$). [Fighre 4] plots the fill rates from simulation for both customer classes and the actual and asymptotic ratio of the unfill rates in Equation (14). Regardless of p, the actual ratio of the unfill rates approaches to the asymptotic ratio, $e^{-\gamma K}$. In all cases, the gap between the two ratios becomes small when the target leadtime t is long enough to make the fill rate higher than 80%. As in [Figure 2], the asymptotic results work better for service time with a large variability. When $c_B^2<1$ as in (b) of [Figure 4], it becomes rare to observe the leadtime is longer than the target leadtime, which makes it inefficient to verify an asymptotic result through simulation.

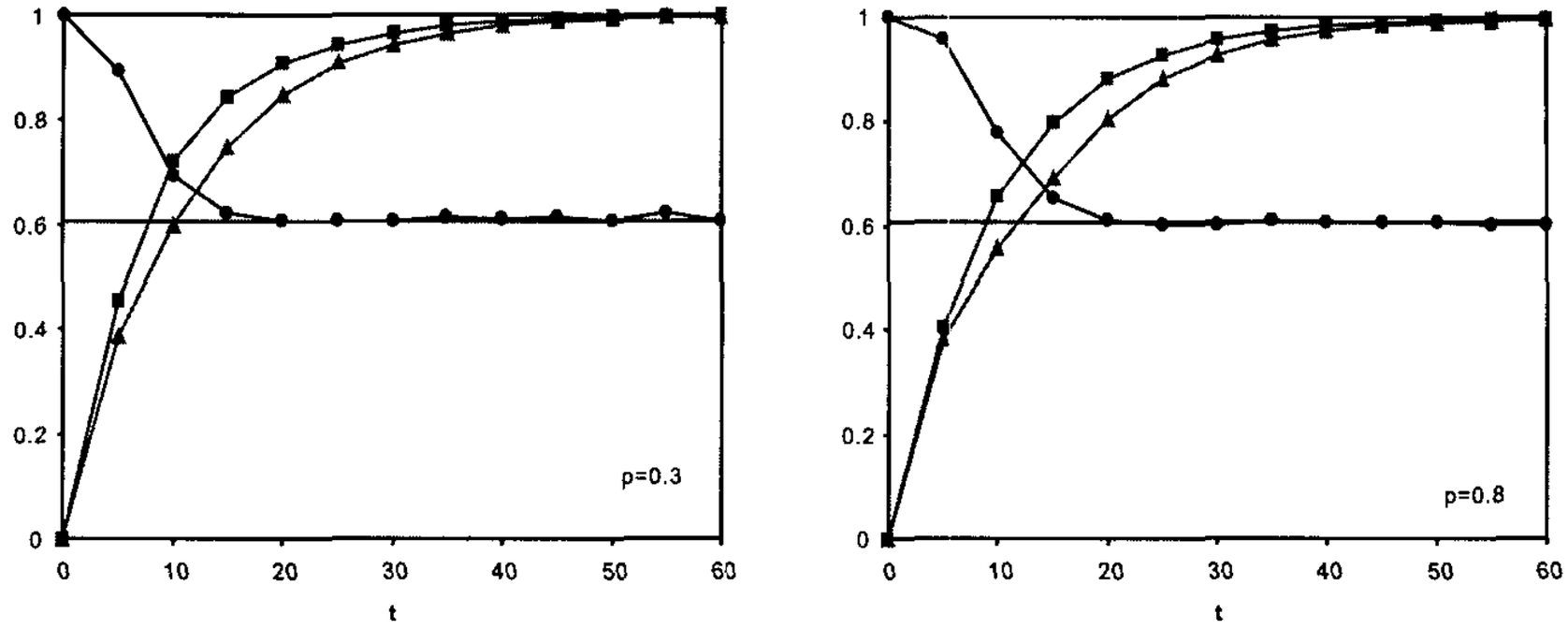
Corollary 1 suggests that the service provider can control the relative service levels between customer classes through parameter K. By numerical examples we demonstrate how well the asymptotic ratio approximates the actual ratio between unfill rates as changing K. We use the same parameter values as in [Figure 3] except for K. The service time is assumed to follow an exponential distribution ($\gamma=0.1$). <Table 1> lists the experiment results such as δ_H and δ_L (unfill rates for high- and low-priority customers, re-

〈Table 1〉 Asymptotic Ratio of Unfill Rates

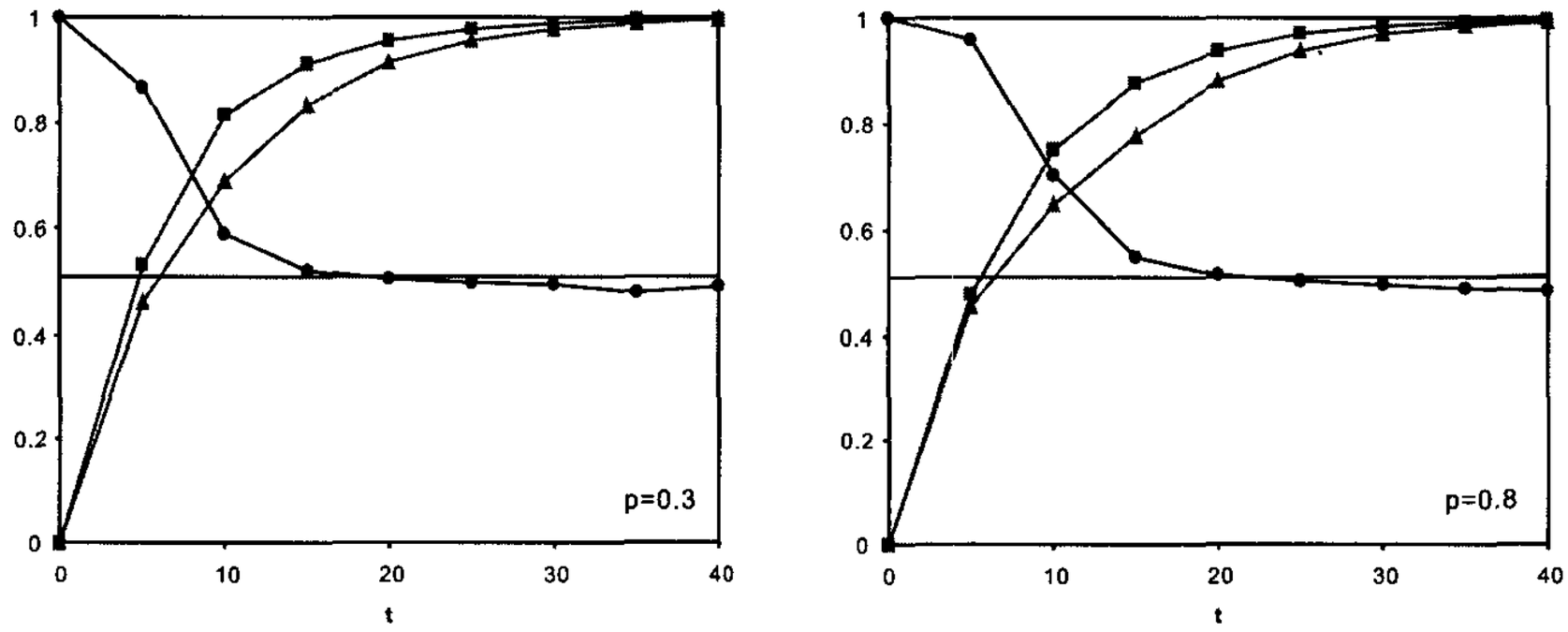
K	$t=20$			$t=25$			$t=30$			$e^{-\gamma K}$
	δ_H	δ_L	δ_H/δ_L	δ_H	δ_L	δ_H/δ_L	δ_H	δ_L	δ_H/δ_L	
1	12.81	14.06	0.911	7.63	8.31	0.918	4.48	4.89	0.916	0.905
2	11.90	14.50	0.821	7.09	8.58	0.826	4.18	5.06	0.825	0.819
3	11.06	14.93	0.741	6.59	8.86	0.744	3.88	5.23	0.743	0.741
5	9.55	15.84	0.603	5.69	9.45	0.602	3.35	5.58	0.601	0.607
7	8.37	16.69	0.501	4.90	10.05	0.488	2.92	5.96	0.490	0.497
10	7.11	17.51	0.406	3.99	10.89	0.366	2.37	6.55	0.361	0.368

주) δ_H and δ_L are in percentage.

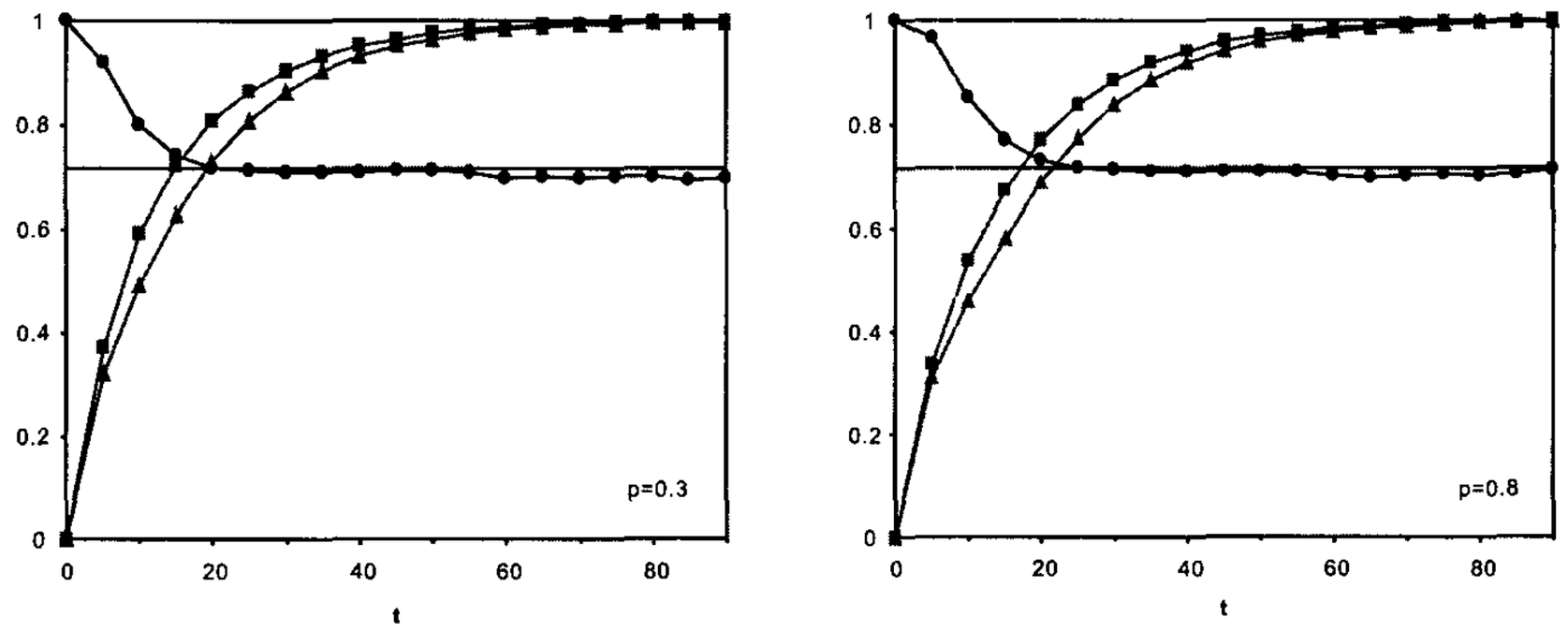
spectively) and the ratio δ_H/δ_L for several target leadtime ($t = 20, 25, 30$). We choose the target leadtime that the fill rate for both customer types is more than 80%. In most cases, the observed ratio is very close to the asymptotic ratio. When the leadtime is shorter, for example $t = 20$ in



(a) Exponential Service Time



(b) Gamma Service Time with $c_B^2 = 0.5$



(c) Gamma Service Time with $c_B^2 = 2$

[Figure 4] Asymptotic Ratio of Unfill Rates

<Table 1>, the gap between δ_H/δ_L and $e^{-\gamma K}$ increases as K increases. For a longer target leadtime, the asymptotic result approximates effectively the relative service level difference between customer classes. Thus, in order to lower the unfill rate for high-priority customers to, say, 60% of the unfill rate for low-priority customers, the service provider can find a proper value for the parameter K using the asymptotic result; $K = -\log(\delta_H/\delta_L)/\gamma = -\log(0.6)/0.1 \approx 5.1$. The numerical results in <Table 1> show that the actual service level ratio is close to 60% when $K=5$.

5. Concluding Remarks

In this paper, we have addressed how to manage customer service levels in a two-priority system where the customers with advance arrival notice have high priority. The priority scheme utilizes how early the advance arrival notice is available. Another characteristic of the two-priority system is the proportion of high- and low-priority customers. We have shown analytically the impact of those system characteristics on the unfill rates for both customer classes. We have demonstrated theoretically and numerically the linear trade-offs between the target leadtime and two system characteristics (p and K) when the fill rate is high. Using the results in this paper, the service provider may answer strategic questions such as how many high-priority customers he can accept and what kind of leadtime he can provide for each type of customers.

In our priority model, a customer's high priority is valid only over the low-priority customers who have arrived within a certain amount of time before the high-priority customer's arrival. When the priority is valid for a long period (large K), our priority model behaves similarly to the classical two-priority queue where a high-priority customer is served before all low-priority customers in the queue. The asymptotic waiting time for low-priority customers is hard to compute even with a Poisson arrival stream and exponential service time in the classical priority queue [1, 11]. Our study provides useful asymptotic results when the service provider has the flexibility to control customer service levels for different groups by adjusting the priority-assigning scheme.

In our approach, the service provider needs to keep track of the arrival time of low-priority customers in order to decide the service position of a high-priority customer. It may not be practical or impossible to keep the track of customer arrival times in some systems, for example, packet routers in telecommunication networks. An alternative priority scheme in such cases is to specify the maximum number of low-priority customers that a high-priority customer can catch up with. All the asymptotic results in this paper still remain valid with small modifications.

참 고 문 헌

[1] Abate, J. and W. Whitt, "Asymptotics for *M/G/1* Low-Priority Waiting-Time Tail Probabilities," *Queueing Systems*, Vol.25, No.1-4(1997), pp.173-223.

[2] Asmussen, S., *Probability and Queues*, Wiley, New York, 1987.

[3] Doi, M., Y. Chen, and H. Osawa, "A Queueing Model in which Arrival Times are Scheduled," *Operations Research Letters*, Vol. 21, No.5(1997), pp.249-252.

[4] Duenyas, I. and W.J. Hopp, "Quoting Customer Lead Times," *Management Science*,

Vol.41, No.1(1995), pp.43-57.

[5] Glasserman, P. and Y. Wang, "Leadtime-Inventory Trade-offs in Assemble-to-Order Systems," *Operations Research*, Vol.46, No.6(1998), pp.858-871.

[6] Hopp, W.J. and M.L.R. Sturgis, "Quoting Manufacturing Due Dates Subject to a Service Level Constraint," *IIE Transactions*, Vol.32, No.9(2000), pp.771-784.

[7] Kendall, M., *Advanced Theory of Statistics*, Vol.II, 5th Ed., Oxford, New York, 1987.

[8] Mercer, A., "Queues with Scheduled Arrivals : A Correction, Simplification and Extension," *Journal of the Royal Statistical Society, Series B*, Vol.35, No.1(1973), pp.104-116.

[9] Sabria, F. and C.F. Daganzo, "Approximate Expressions for Queueing Systems with Scheduled Arrivals and Established Service Order," *Transportation Science*, Vol.23, No. 3(1989), pp.159-165.

[10] Siegmund, D., *Sequential Analysis : Tests and Confidence Intervals*, Springer, New York, 1985.

[11] Subramanian, V.G. and R. Srikant, "Tail Probability of Low-Priority Waiting Times and Queue Lengths in *MAP/GI*/1 Queues," *Queueing Systems*, Vol.34, No.1-4(2000), pp.215-236.

[12] Wein, L.M., "Due-Date Setting and Priority Sequencing in a Multiclass *M/G*/1 Queue," *Management Science*, Vol.37, No.7(1991), pp. 834-850.

〈Appendix〉

Proof of Lemma 1. First, we show that if $O_n > N_n$ for $n \in H$, then

$$W_n \le \sum_{i=N_n}^{O_n-1} B'_i. \tag{18}$$

If $O_n > N_n$, it means the N_n-th service position started to get served before the n-th customer arrives. Before getting served, the n-th customer needs to wait at most for the customers in service position N_n to O_{n-1} finish their service. Thus, (18) holds.

Now, we show that for $n \in H$

$$0 \le \Pr\{R_n > t\} - \Pr\{\overline{R_n} > t\} \le \Pr\left\{\sum_{i=1}^{n-N_n+1} \overline{B_i} > t\right\} \tag{19}$$

where $\{\overline{B_n}\}$ is i.i.d. and has the same distribution as $\{B'_n\}$.

$$\begin{aligned}
&\Pr\{R_n > t\} \\
&= \Pr\left\{W'_{O_n} - \sum_{i=O_n}^{n-1} A_i + B'_{O_n} > t\right\} \\
&= \Pr\left\{O_n = N_n,\ W'_{O_n} - \sum_{i=O_n}^{n-1} A_i + B'_{O_n} > t\right\} \\
&\quad + \Pr\left\{O_n > N_n, W'_{O_n} - \sum_{i=O_n}^{n-1} A_i + B'_{O_n} > t\right\} \\
&= \Pr\left\{W'_{N_n} - \sum_{i=N_n}^{n-1} A_i + B'_{O_n} > t\right\} \\
&\quad + \Pr\left\{O_n > N_n, W'_{O_n} - \sum_{i=O_n}^{n-1} A_i + B'_{O_n} > t\right\} \\
&\quad - \Pr\left\{O_n > N_n, W'_{N_n} - \sum_{i=N_n}^{n-1} A_i + B'_{O_n} > t\right\}
\end{aligned} \tag{20}$$

From Equation (18), we have that

$$W_n \le \sum_{i=N_n}^{O_n-1} B'_i \quad \text{if} \quad O_n > N_n, \tag{21}$$

Thus, we can get bounds on the difference of the last two terms of Equation (20).

$$\begin{aligned}
0 &\le \Pr\left\{O_n > N_n, W'_{O_n} - \sum_{i=O_n}^{n-1} A_i + B'_{O_n} > t\right\} \\
&\quad - \Pr\left\{O_n > N_n, W'_{N_n} - \sum_{i=N_n}^{n-1} A_i + B'_{O_n} > t\right\} \\
&= \Pr\left\{O_n > N_n, t - \left(W'_{O_n} - \sum_{i=O_n}^{n-1} A_i\right) < B'_{O_n}\right. \\
&\le t - \left.\left(W'_{N_n} - \sum_{i=N_n}^{n-1} A_i\right)\right\} \quad \text{(using Equation (21))} \\
&\le \Pr\left\{O_n > N_n, B'_{O_n} > t - \left(W'_{O_n} - \sum_{i=O_n}^{n-1} A_i\right)\right\} \\
&\le \Pr\left\{O_n > N_n, \sum_{i=N_n}^{O_n} B'_i > t\right\} \quad \text{(using Equation (18))} \\
&\le \Pr\left\{O_n > N_n, \sum_{i=N_n}^{n} B'_i > t\right\} \le \Pr\left\{\sum_{i=1}^{n-N+1} \overline{B_i}\right\}
\end{aligned}$$

From Equation (19), once we show that

$$\lim_{t\to\infty} e^{\gamma t} \Pr\left\{\sum_{i=1}^{n-N_n+1} \overline{B_i} > t\right\} = 0 \quad \text{uniformly in } n,$$

we will prove that $\overline{d_n}(t)$ defined in Equation (7) converges to 0 uniformly in n as $t \to \infty$.

From the Markov's inequality,

$$\begin{aligned}
\Pr\left\{\sum_{i=1}^{n-N_n+1} \overline{B_i} > t\right\} &\le e^{-\theta t} \mathrm{E}\left[\exp\left(\theta \cdot \sum_{i=1}^{n-N_n+1} \overline{B_i}\right)\right] \\
&= e^{-\theta t} \mathrm{E}\left[e^{\phi_B(\theta)(n-N_n+1)}\right]
\end{aligned}$$

for an arbitrary $\theta \ge 0$. From the definition of N_n in Equation (4), we have that

$$n - N_n \overset{D}{=} \min\{n, \overline{N}\}$$

where $\overline{N}$ is a Poisson random variable whose definition is given in Equation (24) below. Thus, for $\theta > \gamma$ with finite $\phi_B(\theta)$,

$$e^{\gamma t}\Pr\left\{\sum_{i=1}^{n-N_n+1}\overline{B_i} > t\right\} \le e^{-(\theta-\gamma)t}\mathrm{E}\left[e^{\phi_B(\theta)(\overline{N}+1)}\right] \to 0$$

as $t \to \infty$.

Since the right-hand side of the above inequality does not depend on n, $\overline{d_n}(t)$ converges to 0 uniformly in n as $t \to \infty$.

Proof of Theorem 1. Let $\overline{R}$ denote a random variable whose distribution is the same as the steady-state distribution of $\{\overline{R_n} : n \in H\}$. From the definition of $\overline{R_n}$ and $\overline{d_n}(t)$, we have

$$\lim_{t\to\infty} e^{\gamma t}\Pr_H\{R > t\} = \lim_{t\to\infty} e^{\gamma t}\Pr\{\overline{R} > t\} + \lim_{t\to\infty}\lim_{n\to\infty}\overline{d_{n(t)}}.$$

Since it has been shown in Lemma 1 that $\overline{d_n}(t)$ converges to 0 uniformly in n as $t \to \infty$,

$$\lim_{t\to\infty}\lim_{n\to\infty}\overline{d_n}(t) = \lim_{n\to\infty}\lim_{t\to\infty}\overline{d_n}(t) = 0.$$

Thus, in order to complete the proof, we need to show that for some constant C

$$\lim_{t\to\infty} e^{\gamma t}\Pr\{\overline{R} > t\} = Ce^{-\gamma(1-p)K}. \tag{22}$$

Let $X_k = B'_k - A_k$ and define $\{S_n\}$ as

$$S_0 = 0 \text{ and } S_n = \sum_{k=1}^{n} X_k. \tag{23}$$

If $\lambda\mathrm{E}[B'_1] < 1$, then W'_n converges weakly to a random variable W' which has the same distribution as $\max_{k\ge 0}\{S_k\}$ (Asmussen [2]). Let $\{\overline{A_n}\}$ be an i.i.d. sequence with the same distribution as $\{A_n\}$. Using the result in the proof of Theorem 1 of Glasserman and Wang [5], we conclude $(W'_{N_n}, n - N_n)$ converges in distribution to $(W', \overline{N})$ with

$$\overline{N} \overset{D}{=} \left|\left\{k \ge 0 ; \sum_{i=1}^{k}\overline{A_i} \le K,\ k \in \overline{L}\right\}\right| \tag{24}$$

and W' is independent of $\overline{N}$ and $\{\overline{A_n}\}$. For explaining the new notations in Equation (24), we suppose a new system that operates under the same rule as the original two-priority system. The sequence of interarrival times $\{\overline{A_n}\}$ is i.i.d. and has the same distribution as the original sequence $\{A_n\}$. In the new system, the customers also have the same probability $1-p$ of being in low-priority class $\overline{L}$.

From the above argument with new notations, we have that

$$\Pr\{\overline{R} > t\} = \Pr\left\{W' - \sum_{i=1}^{\overline{N}}\overline{A_i} + \overline{B} > t\right\}$$

where $\overline{B}$ is a random variable having the same distribution as B'_n and independent of W', $\overline{N}$ and $\{\overline{A_n}\}$.

Using $\{S_n\}$ in Equation (23), we define $\tau = \inf\{n \ge 1 : S_n > T\}$ with $T = t + \sum_{i=1}^{\overline{N}}\overline{A_i} - \overline{B}$. Then,

$$\Pr\left\{W' - \sum_{i=1}^{\overline{N}}\overline{A_i} + \overline{B} > t\right\} = \Pr\left\{\max_{n\ge 0} S_n > t + \sum_{i=1}^{\overline{N}}\overline{A_i} - \overline{B}\right\} = \Pr\{\tau < \infty\}.$$

We use exponential twisting (see Chapter XII of

Asmussen [2] for relevant background). Specifically we use gamma-twisting of $\{X_n\}$ and $\overline{B}$ and $(-\gamma)$-twisting of $\{\overline{A}\}$, and denote the use of a twisted measure in computing expectations by $\tilde{\mathrm{E}}$. Note that τ is a stopping time for $\{X_i\}$ and $\overline{N}$ is a stopping time for $\{A_i\}$. From Theorem XII.4.1 of Asmussen [2], we have that

$$\Pr\{\tau<\infty\}=\tilde{\mathrm{E}}\left[\prod_{i=1}^{\tau}e^{-\gamma X_i+\phi_X(\gamma)}\cdot \prod_{i=1}^{\overline{N}}e^{\gamma\overline{A}_i+\phi_A(-\gamma)}\cdot e^{-\gamma\overline{B}+\phi_B(\gamma)};\tau<\infty\right]$$

where the semicolon inside the expectation indicates that the expectation is evaluated over the event after the semicolon. Using the definition of γ, we reduces the above equation further ;

$$\Pr\{\tau<\infty\}=\tilde{\mathrm{E}}\left[\exp\left\{-\gamma S_\tau+\gamma\sum_{i=1}^{\overline{N}}\overline{A}_i-\beta\overline{N}-\gamma\overline{B}+\beta\right\};\tau<\infty\right]$$

$$=e^{-\gamma t}e^{\beta}\tilde{\mathrm{E}}\left[e^{-\gamma(S_\tau-T)-\beta\overline{N}};\tau<\infty\right]$$

where $\beta=\phi_B(\gamma)$. Since a cumulant generating function is convex (Kendall [7]) and $\phi_X(0)=0$,

$$\tilde{\mathrm{E}}[X_i]=\mathrm{E}\left[e^{\gamma X_1}X_1\right]=\phi'_X(\gamma)>0,$$

and thus the event of $\{\tau<\infty\}$ has probability one. The random variable T is independent of $\{S_n\}$ and $T\to\infty$ as $t\to\infty$. From Corollary 8.33 of Siegmund [10], we have that

$$C_1:=\lim_{t\to\infty}\tilde{\mathrm{E}}\left[e^{-\gamma(S_\tau-T)}\right]=\tilde{\mathrm{E}}\left[e^{-\gamma Z}\right]$$

where Z is a ladder variable. Since the distribution of Z is independent of $\{\overline{A_n}\}$, we have that

$$\lim_{t\to\infty}e^{\gamma t}\Pr\left\{W'-\sum_{i=1}^{\overline{N}}\overline{A_i}+\overline{B}>t\right\}$$

$$=C_1e^{\beta}\tilde{\mathrm{E}}\left[e^{-\beta\overline{N}}\right]=C\tilde{\mathrm{E}}\left[e^{-\beta\overline{N}}\right]$$

with $C:=C_1e^{\beta}$.

After $(-\gamma)$-twisting, $\{\overline{A}\}$ has an exponential distribution with mean $1/(\lambda+\gamma)$. It means that under the twisted measure $\overline{N}$ has a Poisson distribution with mean $(\lambda+\gamma)(1-p)K$. Thus,

$$\tilde{\mathrm{E}}\left[e^{-\beta\overline{N}}\right]=\sum_{k=0}^{\infty}e^{-\beta k}\frac{e^{-(\lambda+\gamma)(1-p)K}((\lambda+\gamma)(1-p)K)^k}{k!}$$

$$=\exp\left(-(\lambda+\gamma)(1-p)K(1-e^{-\beta})\right)$$

$$=\exp(-\gamma(1-p)K)$$

and we finally have Equation (22).

Proof of Lemma 2. First, we show that if $O_n<M_n$ for $n\in L$, then

$$W'_n+\sum_{i=n}^{O_n-1}B'_i<K. \tag{25}$$

If $O_n<M_n$, there exists at least one high-priority customer who arrives within K time units after the n-th customer arrived but cannot catch up with him. This means he started to get served before the high-priority customer arrives. His waiting time in the queue must have been less than K ($W_n<K$). Thus, from Equation (3), the above inequality (25) holds.

Now, We first show that for $n\in L$

$$0\le\Pr\left\{\widehat{R}_n>t\right\}-\Pr\left\{R_n>t\right\}$$

$$\le\Pr\left\{\sum_{i=1}^{M_n-n}\overline{B_i}>(t-K)/2\right\} \tag{26}$$

where $\{\overline{B_n}\}$ is i.i.d. and has the same distribution as service time $\{B'_n\}$.

$$\begin{aligned}
&\Pr\{R_n > t\} \\
&= \Pr\left\{W'_n + \sum_{i=n}^{O_n-1} B'_i + B'_{O_n} > t\right\} \\
&= \Pr\left\{O_n = M_n, W'_n + \sum_{i=n}^{O_n-1} B'_i + B'_{O_n} > t\right\} \\
&\quad + \Pr\left\{O_n < M_n, W'_n + \sum_{i=n}^{O_n-1} B'_i + B'_{O_n} > t\right\} \\
&= \Pr\left\{W'_n + \sum_{i=n}^{M_n-1} B'_i + B'_{O_n} > t\right\} \\
&\quad + \Pr\left\{O_n < M_n, W'_n + \sum_{i=n}^{O_n-1} B'_i + B'_{O_n} > t\right\} \\
&\quad - \Pr\left\{O_n < M_n, W'_n + \sum_{i=n}^{M_n-1} B'_i + B'_{O_n} > t\right\}
\end{aligned} \tag{27}$$

We can get bounds on the difference of the last two terms of Equation (27).

$$\begin{aligned}
0 &\le \Pr\left\{O_n < M_n, W'_n + \sum_{i=n}^{O_n-1} B'_i + B'_{O_n} \right. \\
&\quad \left. + \sum_{i=O_n+1}^{M_n-1} B'_i + B'_{O_n} > t\right\} \\
&\quad - \Pr\left\{O_n < M_n, W'_n + \sum_{i=n}^{O_n-1} B'_i + B'_{O_n} > t\right\} \\
&= \Pr\left\{O_n < M_n, \left(t - W'_n - \sum_{i=n}^{O_n-1} B'_i - \sum_{i=O_n+1}^{M_n-1} B'_i\right)/2 \right. \\
&\quad \left. < B'_{O_n} \le t - W'_n - \sum_{i=n}^{O_n-1} B'_i\right\} \\
&\le \Pr\left\{O_n < M_n B'_{O_n} > \left(t - K - \sum_{i=O_n+1}^{M_n-1} B'_i\right)/2\right\} (\text{using (25)}) \\
&\le \Pr\left\{\sum_{i=n}^{M_n-1} B'_i > (t-K)/2\right\} = \Pr\left\{\sum_{i=1}^{M_n-n} \overline{B_i} > (t-K)/2\right\}
\end{aligned}$$

From Equation (26), once we show that

$$\lim_{t\to\infty} e^{\gamma t}\Pr\left\{\sum_{i=1}^{M_n-n} \overline{B_i} > (t-K)/2\right\} = 0 \quad \text{uniformly in } n,$$

we will prove that $\hat{d}_n(t)$ defined in Equation (12) converges to 0 uniformly in n as $t\to\infty$.

From the Markov's inequality,

$$\Pr\left\{\sum_{i=1}^{M_n-n} \overline{B_i} > (t-K)/2\right\} \le e^{-\theta t/2} e^{K/2} \mathrm{E}\left[e^{\phi_B(\theta)(M_n-n)}\right]$$

for an arbitrary $\theta \ge 0$. From the definition of M_n in Equation (10), we have that

$$M_n - n \overset{D}{=} \overline{M}$$

where $\overline{M}$ is a Poisson random variable, whose definition is given in Equation (29) below. Thus, for $\theta > 2\gamma$ with finite $\phi_B(\theta)$,

$$\begin{aligned}
&e^{\gamma t}\Pr\left\{\sum_{i=1}^{M_n-n} \overline{B_i} > (t-K)/2\right\} \\
&\le e^{-(\theta/2-\gamma)t} e^{K/2} \mathrm{E}\left[e^{\phi_B(\theta)\overline{M}}\right] \to 0 \quad \text{as } t\to\infty.
\end{aligned}$$

Since the right-hand side of the above inequality does not depend on n, $\hat{d}_n(t)$ converges to 0 uniformly in n as $t\to\infty$.

Proof of Theorem 2. Let $\hat{R}$ denote a random variable whose distribution is the same as the steady-state distribution of $\{\hat{R}_n : n\in L\}$. From the definition of $\hat{R}_n$ and $\hat{d}_n(t)$, we have that

$$\lim_{t\to\infty} e^{\gamma t}\Pr_L\{R > t\} = \lim_{t\to\infty} e^{\gamma t}\Pr\{\hat{R} > t\} - \lim_{t\to\infty}\lim_{n\to\infty} \hat{d}_n(t).$$

Since it has been shown in Lemma 2 that $\hat{d}_n(t)$ converges to 0 uniformly in n as $t\to\infty$,

$$\lim_{t\to\infty}\lim_{n\to\infty} \hat{d}_n(t) = \lim_{n\to\infty}\lim_{t\to\infty} \hat{d}_n(t) = 0.$$

Thus, in order to complete the proof, we need to show that for some constant C

$$\lim_{t\to\infty} e^{\gamma t}\Pr\{\hat{R}>t\} = Ce^{\gamma(1+\gamma/\lambda)pK}. \tag{28}$$

Using a similar method and notations as in the proof of Theorem 1, we can show that $(W'_n, M_n - n)$ converges in distribution to $(W', \overline{M})$ with

$$\overline{M} \overset{D}{=} \left|\left\{k \geq 0 ; \sum_{i=1}^{k} \overline{A_i} \leq K,\ k \in \overline{H}\right\}\right| \tag{29}$$

and W' is independent of $\overline{M}$. Thus, we have that

$$\Pr\{\hat{R}>t\} = \Pr\left\{W' + \sum_{i=1}^{\overline{M}+1} \overline{B_i} > t\right\}$$

where $\{\overline{B_n}\}$ is i.i.d. with the same distribution as $\{B'_n\}$ and independent of W', and $\{\overline{A_n}\}$.

Using $\{S_n\}$ in Equation (23), we define $\tau' = \inf\{n \geq 1 : S_n > T'\}$ with $T' = t - \sum_{i=1}^{\overline{M}+1} \overline{B_i}$. Then,

$$\Pr\left\{W' + \sum_{i=1}^{\overline{M}+1} \overline{B_i} > t\right\}$$

$$= \Pr\left\{\max_{n\geq 0} S_n > t - \sum_{i=1}^{\overline{M}+1} \overline{B_i}\right\} = \Pr\{\tau' < \infty\}.$$

With γ-twisting of $\{X_n\}$ and $\{\overline{B_n}\}$, we have that

$$\Pr\{\tau' < \infty\} = \tilde{\mathrm{E}}\left[\prod_{i=1}^{\tau'} e^{-\gamma X_i + \phi_X(\gamma)} \cdot \prod_{i=1}^{\overline{M}+1} e^{-\gamma \overline{B_i} + \phi_B(\gamma)} ; \tau' < \infty\right].$$

Using the definition of γ, we reduce the above equation further;

$$\Pr\{\tau' < \infty\} = \tilde{\mathrm{E}}\left[\exp\left\{-\gamma S_{\tau'} - \gamma \sum_{i=1}^{\overline{M}+1} \overline{B_i} + \gamma(\overline{M}+1)\right\} ; \tau' < \infty\right]$$

$$= e^{-\gamma t} e^{\beta} \tilde{\mathrm{E}}\left[e^{-\gamma(S_{\tau'} - T') + \beta \overline{M}} ; \tau' < \infty\right]$$

where $\beta = \phi_B(\gamma)$. With the twisted measure $\tilde{\mathrm{E}}[X_i] = \mathrm{E}[e^{\gamma X_1} X_1] = \phi'_X(\gamma) > 0$, and thus the event of $\{\tau' < \infty\}$ has probability one. The random variable T' is independent of $\{S_n\}$ and $T' \to \infty$ as $t \to \infty$.

From Corollary 8.33 of Siegmund [10], we have that

$$C_1 = \lim_{t\to\infty} \tilde{\mathrm{E}}\left[e^{-\gamma(S_{\tau'} - T')}\right] = \tilde{\mathrm{E}}\left[e^{-\gamma Z}\right]$$

where Z is a ladder variable. Since the distribution of Z is independent of $\{\overline{B_n}\}$, we have that

$$\lim_{t\to\infty} e^{\gamma t}\Pr\left\{W' + \sum_{i=1}^{\overline{M}+1} \overline{B_i} > t\right\} = C_1 e^{\beta} \tilde{\mathrm{E}}\left[e^{-\beta \overline{M}}\right] = C\tilde{\mathrm{E}}\left[e^{-\beta \overline{M}}\right]$$

with $C := C_1 e^{\beta}$. Note that the constant is the same as in Theorem 1.

Since $\overline{M}$ is independent of both $\{X_n\}$ and $\{\overline{B_n}\}$,

$$\tilde{\mathrm{E}}\left[e^{\beta \overline{M}}\right] = \mathrm{E}\left[e^{\beta \overline{M}}\right]$$

$$= \sum_{k=0}^{\infty} e^{\beta k} \frac{e^{-\lambda pK}(\lambda pK)^k}{k!} = \exp\left(-\lambda pK(1 - e^{\beta})\right)$$

$$= \exp(-\gamma pK)$$

and we finally have Equation (28).