J. Soc. Korea Ind. Syst. Eng Vol. 40, No. 3 : 66-75, September 2017
https://doi.org/10.11627/jkise.2017.40.3.066
ISSN : 2005-0461(print)
ISSN : 2287-7975(online)

(N, n)-Preemptive Repeat-Different Priority Queues

Kilhwan Kim†

Department of Management Engineering, Sangmyung University

(N, n)-선점 재샘플링-반복 우선순위 대기행렬

김 길 환†

상명대학교 경영공학과

Priority disciplines are an important scheme for service systems to differentiate their services for different classes of customers. (N, n)-preemptive priority disciplines enable system engineers to fine-tune the performances of different classes of customers arriving to the system. Due to this virtue of controllability, (N, n)-preemptive priority queueing models can be applied to various types of systems in which the service performances of different classes of customers need to be adjusted for a complex objective. In this paper, we extend the existing (N, n)-preemptive resume and (N, n)-preemptive repeat-identical priority queueing models to the (N, n)-preemptive repeat-different priority queueing model. We derive the queue-length distributions in the M/G/1 queueing model with two classes of customers, under the (N, n)-preemptive repeat-different priority discipline. In order to derive the queue-length distributions, we employ an analysis of the effective service time of a low-priority customer, a delay cycle analysis, and a joint transformation method. We then derive the first and second moments of the queue lengths of high- and low-priority customers. We also present a numerical example for the first and second moments of the queue length of high- and low-priority customers. Through doing this, we show that, under the (N, n)-preemptive repeat-different priority discipline, the first and second moments of customers with high priority are bounded by some upper bounds, regardless of the service characteristics of customers with low priority. This property may help system engineers design such service systems that guarantee the mean and variance of delay for primary users under a certain bounds, when preempted services have to be restarted with another service time resampled from the same service time distribution.

Keywords : Queueing Systems, Priority Queues, Preemption, Preemptive Repeat Different, N-Policy

1. Introduction

Priority disciplines are an important scheme for service systems to differentiate their services for different classes of customers. For this reason, many telecommunication, computer, and production systems employ various priority disciplines. Priority queueing models are typically utilized to analyze the performance of systems under various priority disciplines, and there have been a number of studies on priority queueing models [9, 10, 12, 21, 22, 24].

There are two fundamental classical priority disciplines [23] : One is the preemptive priority discipline, under which customers with high priority can interrupt the service of customers with lower priority, and the other is the nonpreemptive priority discipline, under which, once the service of customers with low priority has been started, it is not interrupted even though there are customers with high priority arriving in the system during the service. However, these two priority

Received 12 July 2017; Finally Revised 14 August 2017; Accepted 16 August 2017
† Corresponding Author : khkim@smu.ac.kr

disciplines have their own drawbacks. Under the preemptive priority discipline, the quality of service (QoS) for customers with low priority may be severely degraded especially when preemption frequently occurs and the service interrupted has to be completely repeated. On the other hand, under the non-preemptive priority discipline, the QoS for customers with high priority may be severely degraded especially when the service time of customers with low priority is relatively long and significantly varies in length.

In order to overcome these drawbacks of the classical priority disciplines, several hybrid-type priority disciplines have been introduced [1~4, 6~8, 11, 13~16, 18~20, 22~24]. See [13~16] for a brief review on these hybrid-type priority disciplines. The (N, n)-preemptive priority discipline is one of these hybrid-type priority discipline [14]. Under this discipline, the decision on whether or not to preempt the service of customers with low priority depends on the number of customers with high priority present in the system. Once the service of customers with low priority has started, it is preempted only when the number of customers with high priority reaches or exceeds a certain threshold $N, N \geq 1$; the interrupted service is restored when the number of customers with high priority shrinks to another certain threshold $n, 0 \leq n \leq N-1$. The advantage of the (N, n)-preemptive priority discipline is that it is easier than other hybrid-type priority disciplines to control the QoS for customers with high priority in a certain bound, regardless of the arrival and service characteristics of customers with low priority [14].

This paper extends Kim's result [14] to a different variant of the (N, n) preemptive priority discipline. While Kim [14] assumes that the interrupted service is retained and resumed when the service is available again for the preempted low-priority customer (which will be called the (N, n)-preemptive resume (PR) priority discipline) or it is repeated with the identical service time (which will be called the (N, n)-preemptive repeat-identical (PRI) priority discipline), this paper assumes that the interrupted service of the preempted low-class customer is repeated with a different service time from the same service distribution, which will be called the (N, n)-preemptive repeat-different (PRD) priority discipline). The (N, n)-PRD discipline is a natural extension of the classical preemptive repeat-different discipline and its variant [6, 17, 25] to the (N, n)-preemptive discipline. There are two contributions which extended Kim's original (N, n)-preemptive models [18, 19] to discrete-time queueing models. However, these contributions only considered the models with geometric service times, where there are no differences between (N, n)-PR, PRD, and PRI disciplines due to the so-called memoryless property of the service time distribution. Here, we consider the models with general service-time distributions, and the types of the way of restoring interrupted services have different impact on the system overall performance.

This paper is organized as follows : In Section 2, we define the mathematical model considered in this paper; In Section 3, we analyze the effective service time structure of a low-priority customer. In Section 4, we derive the queue-length distributions for high-priority and low-priority customers. In Section 5, we present some numerical examples by using the result obtained in Sections 3 and 4. In Section 6, we conclude the paper with some comments.

2. Model

We consider the following M/G/1 queueing model under the (N, n)-preemptive repeat-different preemptive discipline : There are two classes of customers and class-$i, i=1,2$, customers arrive at the system according to a Poisson process with rate λ_i. The total arrival rate λ is defined as $\lambda = \lambda_1 + \lambda_2$. The service times S_i of class-i customers are independent and identically distributed, and follow an arbitrary general distribution. Let $S_i(x)$ denote the distribution function of S_i, and $S_i^*(\theta)$ denote the Laplace-Stieltjes transform (LST) of $S_i(x)$.

Under the (N, n)-PRD priority discipline, class-1 customers have priority over class-2 customers. If there are no preempted class-2 customers in the system, one of class-1 customers (if any) is first selected for the next service just after the service being process is completed. If there are no class-1 customers at the service completion time, one of class-2 customers (if any) can be selected for the next service. While the service of a class-2 customer is being processed, class-1 customers can preempt that service only when the number of class-1 customers reaches or exceeds a certain threshold $N, N \geq 1$. Then, the interrupted service of the preempted class-2 customer is completely restarted from the beginning with a different service time resampled from the same service time distribution $S_2(x)$, when the number of class-1 customers shrinks to another threshold $n, 0 \leq n \leq N-1$. Note that, when $N=1$ and $n=0$, the (N, n)-PRD priority discipline become identical to the classical preemptive repeat-different priority discipline. Also, when $N=\infty$, the (N, n)-PRD prio-

rity discipline become identical to the classical nonpreemptive priority discipline.

Let $\rho_i = \lambda_i E[S_i]$. We assume that the system is stable for class-1 customers. That is, $\rho_1 < 1$. Let also $\rho = \rho_1 + \rho_2$. Note that $\rho < 1$ is not the stability condition for class-2 customers because the service of class-2 customers can be repeated more than once. We assume that the system can be either stable or unstable for class-2 customers. We will assume the stability of the system for class-2 customers only when we derive the queue-length distribution of class-2 customers.

3. Analysis of Service Time Structure

The overall approach here is the same as that in [14]. Thus, the common derivation in both [14] and this study will not be repeated here, except when it is needed for the overall understanding of the system considered here. As with many studies on priority queueing models [5~7, 13, 14, 16], we employ a classical method in which the effective service time of a low-class customer is analyzed and the result obtained is plugged into a kind of M/G/1 vacation queueing models to derive the queue-length distributions of high- and low-class customers, respectively (see [23] for a brief review on the queue-length distribution of vacation queues). In this approach, three types of the effective service time of a class-2 customer are defined : the gross service time G of a class-2 customer is defined as the total time spent by the server for this class-2 customers (including all the service repetitions (if any)); the completion time C is defined as the time interval from the first service start time of the class-2 customer until the completion of that service; the occupation time R is defined as the time interval from the first service start time of the class-2 customer until the server is available for the next class-2 customer (if any).

We further divide the class-1 customers who arrived during G into two groups : some class-1 customers who arrive during G will preempt the service of the class-2 customer and complete their service before G ends. We will call those class-1 customers GP customers. Some other class-1 customers who arrive during G will not preempt the service of the class-2 customer and begin their service after G ends. We will call those class-1 customers GN customers. Let a^{GP} and a^{GN} denote the number of GP and GN customers during G, respectively. Let a^G denote the number of class-1 customers who arrive during G. Thus, $a^G = a^{GP} + a^{GN}$. We also define the following joint transform of G, a^{GP} and a^{GN} :

$$G(\theta; z, w) = E\left[e^{-G\theta} z^{a^{GP}} w^{a^{GN}}\right] \tag{1}$$

If we let $C^*(\theta)$ and $R^*(\theta)$ denote the LSTs of C and R, then we have from [14] :

$$C^*(\theta) = G(\theta; B^*(\theta), 1) \tag{2}$$

and

$$R^*(\theta) = G(\theta; B^*(\theta), B^*(\theta)) \tag{3}$$

where $B^*(\theta)$ is the LST of the standard M/G/1 busy period with S_1 as the corresponding service time and λ_1 as the corresponding arrival rate, and $B^*(\theta)$ is expressed as the following well-known equation (see [23]) :

$$B^*(\theta) = S_1(\theta + \lambda_1 - \lambda_1 B^*(\theta)) \tag{4}$$

Under the (N, n)-PRD discipline, each time the number of class-1 customers in the system reaches N during the service time of a class-2 customer, the service of the class-2 customer is preempted. This preempted service will be completely repeated at the next service attempt with a different service time resampled from the same service-time distribution $S_2(x)$, as soon as the number of class-1 customers shrinks to n.

Hence, if the service time S_2 of a class-2 customer at its first service attempt is shorter than the Erlang random variable E_N with parameters N and λ_1, the service of the class-2 customer will be completed at this service attempt. Otherwise, it will be preempted and repeated with a different service time resampled from the same distribution, when the number of class-1 customers shrinks to n. If the service time S_2 of a class-2 customer at each service attempt except the first service attempt is shorter than the Erlang random variable E_{N-n} with parameters $(N-n)$ and λ_1, the service of the class-2 customer will be completed at this service attempt. Otherwise, it will be preempted again. It is because, at the beginning of each service attempt except the first service attempt, there are already n class-1 customers in the system.

In order to take into consideration this dependency of the service completion of a class-2 customer on S_2, E_N and E_{N-n}, we first define a^{S_2} as the number of class-1 customers

who arrive during S_2 at an arbitrary service attempt. We next define the following two joint transforms for $r=1, 2, \cdots$:

$$S_2^*(\theta; r) = \Pr\left[a^{S_2}=r\right] E\left[e^{-\theta S_2} \mid a^{S_2}=r\right] \tag{5}$$

$$= \int_0^\infty \frac{(\lambda_1 x)^r}{r!} e^{-(\lambda_1+\theta)x} dS_2(x)$$

and

$$E_r^*(\theta) = \Pr\left[E_r < S_2\right] E\left[e^{-\theta E_r} \mid E_r < S_2\right]$$

$$= \int_0^\infty (1-S_2(x)) \frac{(\lambda_1 x)^{r-1}}{(r-1)!} \lambda_1 e^{-(\lambda_1+\theta)x} dx$$

where E_r denotes the Erlang random variable with parameters r and λ_1. By integrating by parts we have

$$E_1^*(\theta) = \frac{\lambda_1}{\lambda_1+\theta}\left(1-S_2^*(\theta; 0)\right)$$

$$E_r^*(\theta) = \frac{\lambda_1}{\lambda_1+\theta}\left(E_{r-1}^*(\theta) - S_2^*(\theta; r-1)\right),\ r \geq 2.$$

Thus we have

$$E_r^*(\theta) = \left(\frac{\lambda_1}{\lambda_1+\theta}\right)^r - \sum_{k=0}^{r-1}\left(\frac{\lambda_1}{\lambda_1+\theta}\right)^{r-k} S_2^*(\theta; k),\ r \geq 1, \tag{6}$$

which also yields

$$1-E_r^*(0) = \sum_{k=0}^{r-1} S_2^*(0; k),\ r \geq 1. \tag{7}$$

To derive the joint transform $G^*(\theta; z, w)$, we also define $\widetilde{G}$ to be the remaining gross service time at the beginning of the second service attempt of a class-2 customer, provided the service of the class-2 customer has not completed at its first service attempt. Let $a^{\widetilde{GP}}(a^{\widetilde{GN}})$ denote the numbers of class-1 customers who arrive during $\widetilde{G}$ and preempt (do not preempt) the service of the class-2 customer. The corresponding joint transform $\widetilde{G}^*(\theta; z, w)$ is defined as

$$\widetilde{G}^*(\theta; z, w) = E\left[e^{-\widetilde{G}\theta} z^{a^{\widetilde{GP}}} w^{a^{\widetilde{GN}}}\right].$$

Since the service order among class-1 customers does not affect the distribution of G, a^{GP} and a^{GN}, we will assume the Last-In-First-Out (LIFO) service order for class-1 customers in our derivation of the joint transformation $G^*(\theta; z, w)$. We now consider the following two events that can occur at the first service attempt of a class-2 customer. (Note that the service time of a class-2 customer is resampled from the same distribution function $S_2(x)$ at each service attempt) :

If $a^{S_2} \geq N$ (i.e., $E_N < S_2$) : In this case, at least N class-1 customers arrive during the first service attempt, and the service of the class-2 customer gets preempted immediately after the N^{th} class-1 customer arrives at the system. Since we assume the LIFO service order for class-1 customers, only the last $N-n$ class-1 customers are served before the preempted class-2 customer is restored again for its service. Also, the remaining n class-1 customers are not serviced until the service of the class-2 customer is completed. It is because, whenever the queue length of class-1 customers shrinks to n, the service of the class-2 customer is restored, and we assumed the LIFO service order for class-1 customers. Thus, we have

$$\Pr\left[a^{S_2} \geq N\right] \cdot E\left[e^{-G\theta} z^{a^{GP}} w^{a^{GN}} \mid a^{S_2} \geq N\right]$$

$$= E_N^*(\theta) z^{N-n} w^n \cdot \widetilde{G}^*(\theta; z, w).$$

If $a^{S_2}=r,\ 0 \leq r \leq N-1$: In this case, no preemption occurs and the service of the class-2 customer is completed at the first service attempt. Thus, all the r class-1 customers who have arrived during the first service attempt are GN customers, and we have

$$\Pr\left[a^{S_2}=r\right] \cdot E\left[e^{-G\theta} z^{a^{GP}} w^{a^{GN}} \mid a^{S_2}=r\right] = S_2^*(\theta; r) w^r.$$

Combining the two cases above, we have

$$G(\theta; z, w) = E_N^*(\theta) z^{N-n} w^n \cdot \widetilde{G}^*(\theta; z, w) \tag{8}$$

$$+ \sum_{r=0}^{N-1} S_2^*(\theta; r) w^r.$$

In order to derive the term $G^*(\theta; z, w)$, we consider the following two events at the second service attempt in a manner similar to the first service attempt.

If $a^{S_2} \geq N-n$ (i.e., $E_{N-n} < S_2$) : In this case, at least $N-n$ class-1 customers arrive during the second service attempt, and the service of the class-2 customer gets preempted immediately after the $(N-n)^{\text{th}}$ class-1 customer arrives at the system and it makes N class-1 customers in the system.

(Note that there are already n customers at the beginning of the second service attempt.) Since we assume the LIFO service order for class-1 customers, all the $(N-n)$ class-1 customers who arrive during the second service attempt are served before the preempted class-2 customer is restored again for its service. Also, the probability structure at the third service attempt is the same as that at the second service attempt, except that the two service attempts have been already tried, because the arrival process is a Poisson process and the service time is resampled from the identical distribution. Thus, we have

$$\Pr\left[a^{S_2} \geq N-n\right] \cdot E\left[e^{-\widetilde{G}\theta} z^{a^{\widetilde{GP}}} w^{a^{\widetilde{GN}}} \mid a^{S_2} \geq N-n\right]$$
$$= E^*_{N-n}(\theta) z^{N-n} \cdot \widetilde{G}^*(\theta; z, w).$$

If $a^{S_2} = r, 0 \leq r \leq N-n-1$: In this case, no preemption occurs and the service of the class-customer is completed at the second service attempt. Thus, all the r class-1 customers who have arrived during the second service attempt are GN customers, and we have

$$\Pr\left[a^{S_2} = r\right] \cdot E\left[e^{-\widetilde{G}\theta} z^{a^{\widetilde{GP}}} w^{a^{\widetilde{GN}}} \mid a^{S_2} = r\right] = S^*_2(\theta; r) w^r.$$

Combining the two cases, we have

$$\widetilde{G}^*(\theta; z, w) = E^*_{N-n}(\theta) z^{N-n} \cdot \widetilde{G}^*(\theta; z, w) + \sum_{r=0}^{N-n-1} S^*_2(\theta; r) w^r,$$

which yields

$$\widetilde{G}^*(\theta; z, w) = \frac{\sum_{r=0}^{N-n-1} S^*_2(\theta; r) w^r}{1 - E^*_{N-n}(\theta) z^{N-n}}. \tag{9}$$

Plug (9) in (8) we have

$$G(\theta; z, w) = \sum_{r=0}^{N-1} S^*_2(\theta; r) w^r + \frac{E^*_N(\theta) z^{N-n} w^n}{1 - E^*_{N-n}(\theta) z^{N-n}} \cdot \sum_{r=0}^{N-n-1} S^*_2(\theta; r) w^r. \tag{10}$$

From (2), (3) and (10), we have

$$C^*(\theta) = \sum_{r=0}^{N-1} S^*_2(\theta; r) + \frac{E^*_N(\theta) B^*(\theta)^{N-n}}{1 - E^*_{N-n}(\theta) B^*(\theta)^{N-n}} \cdot \sum_{r=0}^{N-n-1} S^*_2(\theta; r) \tag{11}$$

and

$$R^*(\theta) = \sum_{r=0}^{N-1} S^*_2(\theta; r) B^*(\theta)^r + \frac{E^*_N(\theta) B^*(\theta)^N}{1 - E^*_{N-n}(\theta) B^*(\theta)^{N-n}} \cdot \sum_{r=0}^{N-n-1} S^*_2(\theta; r) B^*(\theta)^r. \tag{12}$$

Differentiating (10) with respect to θ, z, and w, respectively, and letting $\theta = 0$, $z = 1$, and $w = 1$ gives

$$E[G] = \frac{1}{\lambda_1}\left(N - \sum_{r=0}^{N-1} (N-r) S^*_2(0, r)\right) + \left(1 - \sum_{r=0}^{N-1} S^{*(0;r)}_2\right) \frac{(N-n) - \sum_{r=0}^{N-n-1} (N-n-r) S^*_2(0, r)}{\lambda_1 \sum_{r=0}^{N-n-1} S^*_2(0; r)}$$

$$E[a^{GP}] = \frac{(N-n)\left(1 - \sum_{r=0}^{N-1} S^*_2(0; r)\right)}{\sum_{r=0}^{N-n-1} S^{*(0;r)}_2}$$

$$E[a^{GN}] = \sum_{r=0}^{N-1} r S^*_2(0; r) + \left(1 - \sum_{r=0}^{N-1} S^*_2(0; r)\right)\left\{n + \frac{\sum_{r=1}^{N-n-1} r S^*_2(0; r)}{\sum_{r=0}^{N-n-1} S^*_2(0; r)}\right\}$$

$$E[a^{GN}(a^{GN}-1)] = \sum_{r=0}^{N-1} r(r-1) S^*_2(0; r) + \left(1 - \sum_{r=0}^{N-1} S^{*(0;r)}_2\right) \times n(n-1) + 2n \frac{\sum_{r=1}^{N-n-1} r S^*_2(0; r)}{\sum_{r=0}^{N-n-1} S^*_2(0; r)} + \frac{\sum_{r=2}^{N-n-1} r(r-1) S^*_2(0; r)}{\sum_{r=0}^{N-1} S^*_2(0; r)}\Bigg\}$$

$$E[a^{GN}(a^{GN}-1)(a^{GN}-2)] = \sum_{r=0}^{N-1} r(r-1)(r-2) S^*_2(0; r) + \left(1 - \sum_{r=0}^{N-1} S^{*(0;r)}_2\right) \times \Bigg\{ n(n-1)(n-2) + 3n(n-1) \frac{\sum_{r=1}^{N-n-1} r S^*_2(0; r)}{\sum_{r=0}^{N-n-1} S^{*(0;r)}_2} + 3n \frac{\sum_{r=2}^{N-n-1} r(r-1) S^*_2(0; r)}{\sum_{r=0}^{N-1} S^*_2(0; r)} + \frac{\sum_{r=3}^{N-n-1} r(r-1)(r-2) S^*_2(0; r)}{\sum_{r=0}^{N-n-1} S^*_2(0; r)} \Bigg\}.$$

Differentiating (11) and (12) with respect to θ and letting $\theta=0$ gives

$$E[C]=\frac{(N-n)\left(1-\sum_{r=0}^{N-1}S_2^*(0;r)\right)}{\lambda_1(1-\rho_1)\sum_{r=0}^{N-n-1}S_2^*(0;r)}+\frac{1}{\lambda_1}\left(N-\sum_{r=0}^{N-1}(N-r)S_2^*(0,r)\right)$$
$$-\frac{\left(1-\sum_{r=0}^{N-1}S_2^*(0;r)\right)\sum_{r=0}^{N-n-1}(N-n-r)S_2^*(0,r)}{\lambda_1\sum_{r=0}^{N-n-1}S_2^*(0;r)}$$

and

$$E[R]=E[G]/(1-\rho_1)$$

In a similar way, expressions for $E[C^2]$, $E[R^2]$, and $E[R^3]$, which are needed for the moments of the class-2 queue length, can be calculated by taking the appropriate derivatives of the respective PGFs as well. (The expressions are omitted because they are too elaborate, but we can derive them using a computer algebra system such as Mathematica or Maxima. In order to show this, we demonstrate some figures of the second moment of the class-2 queue length in Section 5).

4. Queue-Length Distributions

Let ρ_G denote the total utilization factor of the server for class-2 customers. Then, it is expressed as $\rho_G=\lambda_2E[G]$ from Little's formula. Let ρ_T denote the total utilization factor for both class-1 and class-2 customers. Then it is expressed as $\rho_T=\rho_1+\rho_G$. Since the stability of the system is only assumed for class-1 customers (i.e., $\rho_1<1$), ρ_T may equal or exceed 1, and in that case, the system is unstable for class-2 customers.

Since the (N, n)-PRD priority queue has the same structure of the general cycle as those in the other (N, n)-preemptive priority queue (see [14]), we have the same PGF form of the queue-length distributions of both classes of customers, except that the distributions of G, C, and R are different. Thus, from [14], the PGF $\Pi_1(z)$ of the queue-length distribution of class-1 customers is expressed as :

$$\Pi_1(z)=V(z)\cdot\Pi_{1,M/G/1}(z), \tag{13}$$

where

$$V(z)=\max\left[0,\frac{1-\rho_T}{1-\rho_1}\right]+\min\left[1,\frac{\rho_G}{1-\rho_1}\right]$$
$$\times\left\{\frac{E[a^{GP}](z^n-z^N)}{E[a^G](N-n)(1-z)}+\frac{1-G^*(0,1,z)}{E[a^G](1-z)}\right\}$$

and

$$\Pi_{1,M/G/1}(z)=\frac{(1-\rho_1)(1-z)S_1^*(\lambda_1-\lambda_1z)}{S_1^*(\lambda_1-\lambda_1z)-z}.$$

We now derive the moments of the queue length of class-1 customers explicitly, which was omitted in [14]. If we let L_1 and $L_1^{(2)}$ be the first and second moments of the queue length of class-1 customers in the system in a steady state, we have from (13)

$$L_1=\lim_{z\to1}\Pi_{1'(z)}=V_1+L_{1,M/G/1}$$

$$L_1^{(2)}=\lim_{z\to1}\Pi_1^{(2)}(z)+\Pi_{1'(z)}=V_1^{(2)}+2V_1\cdot L_{1,M/G/1}+L_{1,M/G/1}^{(2)}$$

where

$$V_1=\frac{\lambda_2E[a^{GP}](N+n-1)+\lambda_2E[a^{GN}(a^{GN}-1)]}{2\lambda_1(1-\rho_1)}$$

$$V_1^{(2)}=\frac{\lambda_2E[a^{GP}]\{(N+n-1)(2(N+n)-1)-2Nn\}}{6\lambda_1(1-\rho_1)}$$
$$+\frac{\lambda_2E[a^{GN}(a^{GN}-1)(2a^{GN}-1)]}{6\lambda_1(1-\rho_1)}$$

and

$$L_{1,M/G/1}=\frac{\lambda_1^2E[S_1^2]}{2(1-\rho_1)}+\rho_1$$

$$L_{1,M/G/1}^{(2)}=\frac{\lambda_1^3E[S_1^3]}{3(1-\rho_1)}+\frac{\left(\lambda_1^2E[S_1^2]\right)^2}{2(1-\rho_1)^2}+\frac{3\lambda_1^2E[S_1^2]}{2(1-\rho_1)}+\rho_1.$$

Similarly, if we assume that $\rho_T<1$, then, from [14], the PGF $\Pi_2(z)$ of the queue-length distribution of class-2 customers is expressed as :

$$\Pi_2(z)=\frac{\lambda_1(1-B^*(\lambda_2-\lambda_2z))+\lambda_2(1-z)}{\lambda_2(1-z)/(1-\rho_1)} \tag{14}$$
$$\times\frac{(1-E[R])(1-z)C^*(\lambda_2-\lambda_2z)}{R^*(\lambda_2-\lambda_2z)-z}.$$

We also derive the moments of the queue length of class-2 customers explicitly, which was also omitted in [14]. From (14), we have

$$L_2 = \lim_{z\to 1} \Pi_{2'(z)} = \frac{\lambda_1\lambda_2 E[S_1^2]}{2(1-\rho_1)^2} + \frac{\lambda_2^2 E[R^2]}{2(1-\lambda_2 E[R])} + \lambda_2 E[C]$$

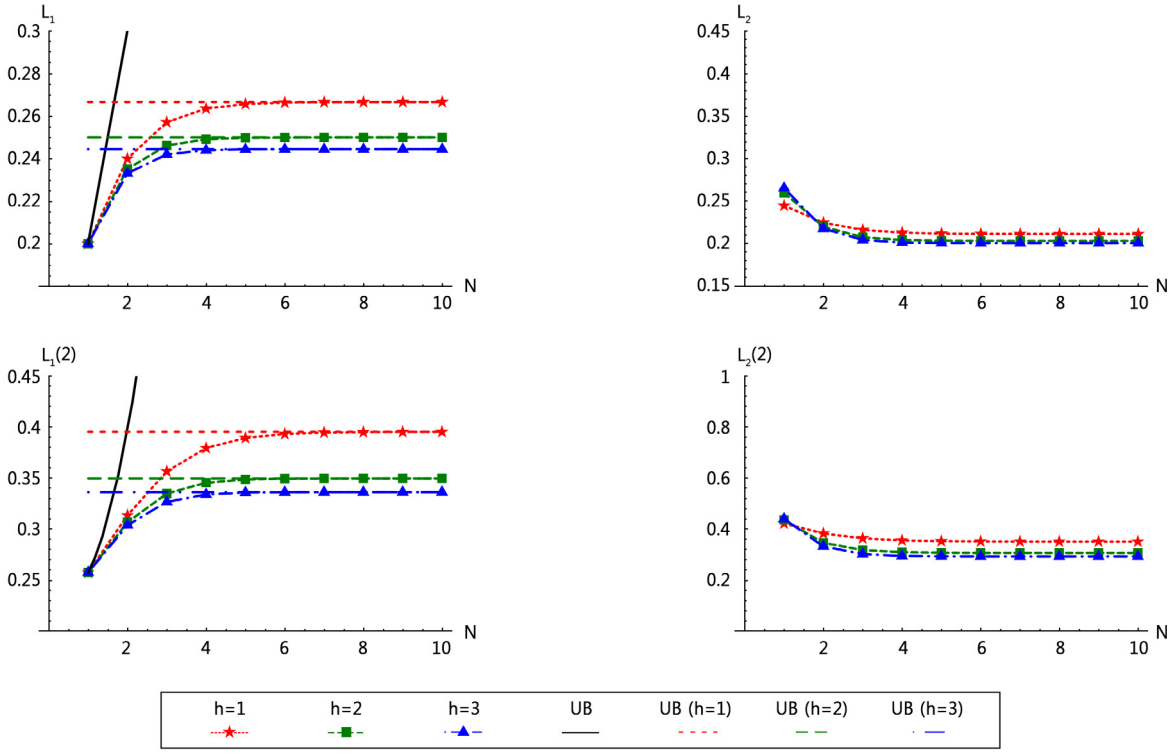
and

$$\begin{aligned} L_2^{(2)} &= \lim_{z\to 1} \Pi_2^{(2)}(z) + \Pi_{2'(z)} \\ &= \frac{\lambda_1\lambda_2^2 E[S_1^3]}{3(1-\rho_1)^3} + \left(\frac{\lambda_1\lambda_2 E[S_1^2]}{(1-\rho_1)^2}\right)^2 + \frac{\lambda_2^3 E[R^3]}{3(1-\lambda_2 E[R])} + \frac{(\lambda_2^2 E[R^2])^2}{2(1-\lambda_2 E[R])^2} \\ &\quad + \frac{\lambda_1\lambda_2 E[S_1^2]}{2(1-\rho_1)^2}\left(1 + \frac{\lambda_2^2 E[R^2]}{1-\lambda_2 E[R]} + 2\lambda_2 E[C]\right) \\ &\quad + \frac{\lambda_2^2 E[R^2](1+2\lambda_2 E[C])}{2(1-\lambda_2 E[R])} + \lambda_2^2 E[C^2] + \lambda_2 E[C]. \end{aligned}$$

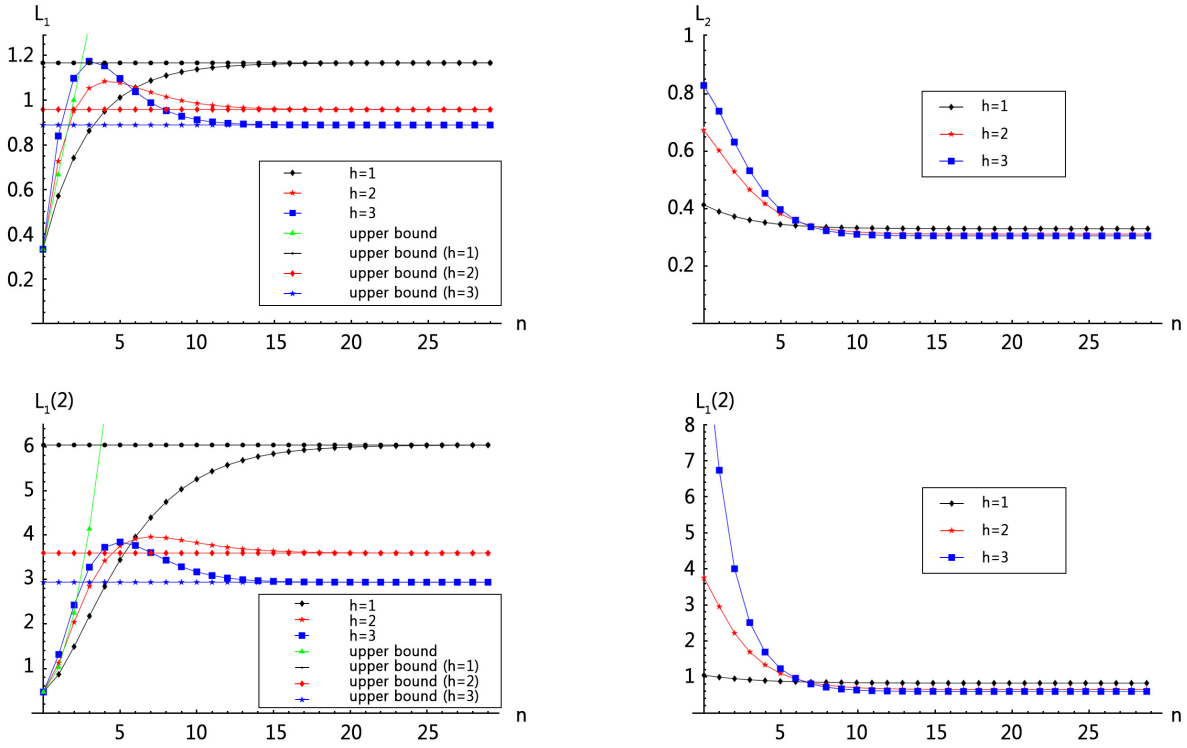
Even though we omit the proof, Lemma 1 and Theorem 2 in [14] on the upper bounds of L_1 and $L_1^{(2)}$ also hold for the (N, n)-PRD queueing model. Refer to the last two paragraphs in [14] for an intuitive explanation of why the same upper bounds hold for different types of (N, n)-preemptive priority queues.

5. Numerical Examples

In this section, we present a numerical example of the first and second moments of the queue lengths of class-1 and class-2 customers. <Figure 1> shows how the moments of the queue lengths are influenced as the upper threshold N changes, when the lower threshold $n=0$, for three different squared coefficients of variation (SCVs) of S_2 (which is denoted by h in <Figure 1>). Note that, if $N=1, n=0$, the (N, n)-PRD queue is identical to the classical PRD queue, and if $N=\infty$, the (N, n)-PRD queue is identical to the classical nonpreemptive queue. In <Figure 1>, we can see that the shapes of the first and second moments of the class-1 queue length in the (N, n)-PRD priority queue are very similar to those in other (N, n)-preemptive queues in [14]. Actually, the upper bounds of the moment of the class-1 queue length in this (N, n)-PRD example are exactly the same as those in the previous (N, n)-preemptive example in [14] because we use the identical arrival processes and the service time distributions for the both examples, in order to make it easy to compare the effect of N for different (N, n)-preemptive disciplines. There is a single remarkable difference between the shapes of the moments of the class-1 queue lengths between the (N, n)-PRD queue and the other (N, n)-preemptive



<Figure 1> The Moments of the Queue Lengths in the (N, n)-PRD Priority Queue ($n=0$)



<Figure 2> The Moments of the Queue Lengths in the (N, n)−PRD Priority Queue ($N-n=1$)

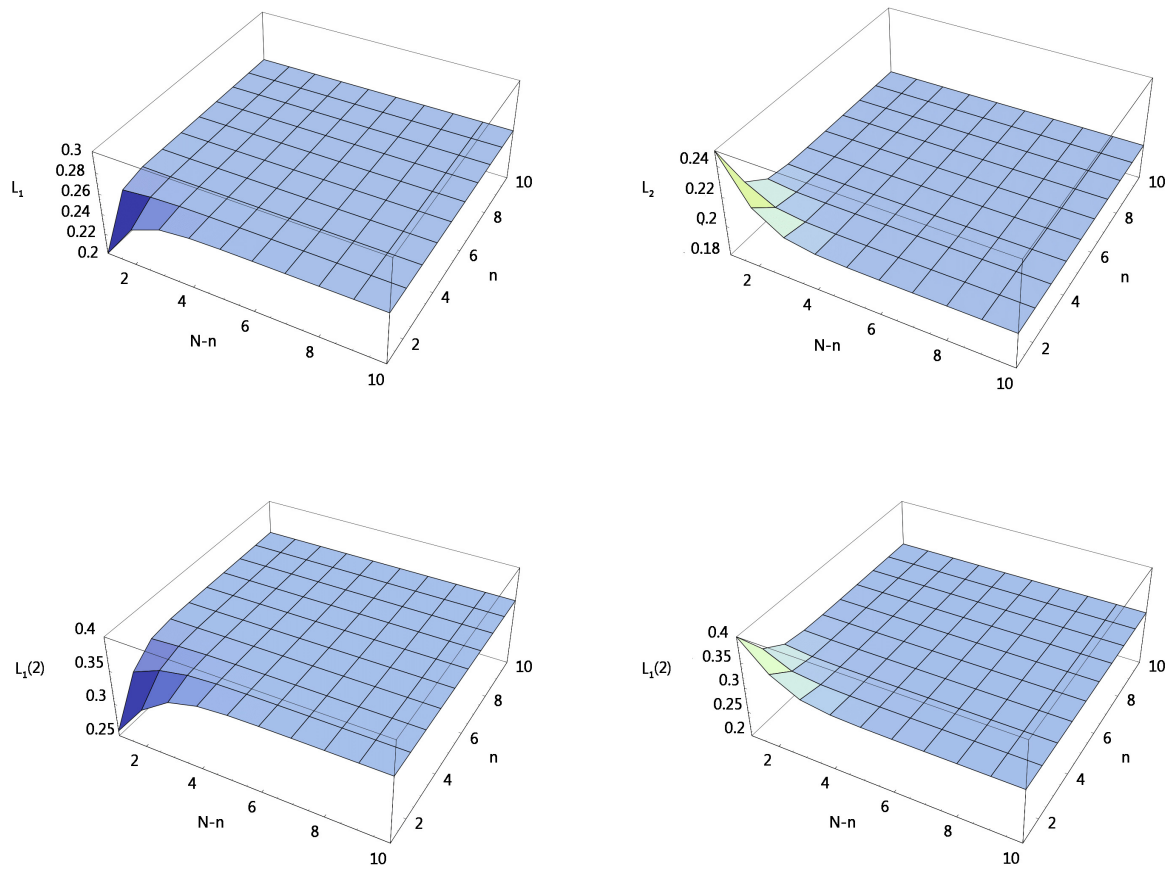
queues : While, for the other (N, n)-preemptive priority queues, the smaller the SCV is (i.e., the relative variance of the service time of class-2 customers is lower), the smaller are the first and second moments of the class-2 queue lengths regardless of the value of N, for the (N, n)-PRD priority queue, this tendency holds only for a sufficiently large N. For the (N, n)-PRD priority discipline, for several small values of N, the smaller the SCV is, the larger are the first and second moments of the class-2 queue length. The reason for this is that, for the case of a large variance of the service time of a class-2 customer, a long service time of a class-2 customer can have a chance to be replaced with a relatively shorter service time by preemption of class-1 customers, and this can positively affect the performance measures of class-2 customers.

In <Figure 2>, the first and second moments of the class-1 and class-2 queue lengths are shown as functions of the threshold n, given $N-n=1$, in the (N, n)-PRD priority queue, for three different SCVs (denoted by h in the figure) of the service time of a class-2 customer.

While, for a sufficiently large n, the shapes of the first and second moments of the class-1 queue length are very similar to the other (N, n)-preemptive priority queue in [14], for a small n, the shapes are different. Interestingly, for the cases of $h=2$ and 3, as n increases, the first and second moments of the class-1 queue length first increase, and then decrease, and finally converge to the first and second moment in the corresponding nonpreemptive queue. These tendencies are also different from the shapes of the first and second moments of the class-1 queue length in <Figure 1>, where the threshold N increase with $n=0$.

The reason of these rising and fall shapes of the first and second moments of the class-1 queue length in the (N, n)-PRD priority queue is that, for the case of $N-n=1$, as n increases, class-1 customers have less chance to preemption, and even if they preempt the service of a class-2 customer, the service of the class-2 customer has to be completely restarted, which is a very severely negative effect on the class-1 customers' performance. However, when n becomes too high, hence N also high, class-1 customers almost never preempt class-2 customers, which avoids the repetition of services of class-2 customers, therefore a better performance for class-1 customers.

In <Figure 3>, the first and second moments of the class-1 and and class-2 queue lengths are shown as functions of the threshold n and the difference of N and n in the (N, n)-PRD priority queue, for the case of $h=2$. For a given n, as $N-n$ increases, the first and second moments of the class-1 queue



<Figure 3> The Moments of the Queue Lengths in the (N, n)-PRD Priority Queue ($h=2$)

length increase, while the first and second moments of the class-2 queue length decrease. Also, for a given $N-n$, as n increases, the first and second moments of the class-1 queue length increase, while the first and second moments of the class-2 queue length decrease.

6. Conclusion

In this paper, we derived the queue-length distributions of high-class and low-class customers in the (N, n)-preemptive repeat-different queueing model, and presented a numerical example for this queueing model. As shown in the numerical example, the first and second moments of the class-1 queue length are bounded by the upper bounds, regardless of the characteristics of the service time of class-1 customers and the preemptive mode, as in the other (N, n)-preemptive priority queue in [14]. This property helps system engineers design such service systems that guarantee the mean and variance of delay for primary users under a certain bounds, when preempted services have to be restarted with another service time resampled from the same service time distribution. The property may be very useful, especially when the stochastic characteristic of low-priority customers easily varies or it cannot be easily determined.

Acknowledgement

This research was supported by a 2016 Research Grant from Sangmyung University.

References

[1] Adiri, I. and Domb, I., A Single Server Queueing System Working under Mixed Priority Disciplines, *Operations Research*, 1982, Vol. 30, No. 1, pp. 97-115.

[2] Adiri, I. and Domb, I., Mixing of Non-Preemptive and Preemptive Repeat Priority Disciplines, *European Journal of Operational Research*, 1984, Vol. 18, No. 1, pp. 86-97.

[3] Avi-Itzhak, B., Brosh, I., and Naor, P., On Discretionary Priority Queueing, *ZAMM-Journal of Applied Mathematics and Mechanics*, 1964, Vol. 44, No. 6, pp. 235-242.

[4] Cho, Y.Z. and Un, C.K., Analysis of the M/G/1 Queue under a Combined Preemptive/Nonpreemptive Priority Discipline, *IEEE Transactions on Communications*, 1993, Vol. 41, No. 1, pp. 132-141.

[5] Conway, R.W., Maxwell, W.L., and Miller, L.W., *Theory of scheduling*, Reading, MA : Addison-Wesley, 1967.

[6] Drekic, S. and Stanford, D.A., Reducing Delay in Preemptive Repeat Priority Queues, *Operations Research*, 2001, Vol. 49, No, 1, pp. 145-156.

[7] Drekic, S. and Stanford, D.A., Threshold-Based Interventions to Optimize Performance in Preemptive Priority Queues, *Queueing Systems*, 2000, Vol. 35, No. 1, pp. 289-315.

[8] Drekic, S., A Preemptive Resume Queue with an Expiry Time for Retained Service, *Performance Evaluation*, 2003, Vol. 54, No. 1, pp. 59-74.

[9] Dudin, A., Lee, M.H., Dudina, O., and Lee, S.K., Analysis of Priority Retrial Queue with Many Types of Customers and Servers Reservation as a Model of Cognitive Radio System, *IEEE Transactions on Communications*, 2017, Vol. 65, No. 1, pp. 186-199.

[10] Gao, S., A Preemptive Priority Retrial Queue with Two Classes of Customers and General Retrial Times, *Operational Research*, 2015, Vol. 15, No. 2, pp. 233-251.

[11] Gay, T.W. and Seeman. P.H., Composite Priority Queue, *IBM Journal of Research and Development*, 1975, Vol. 19, No. 1, pp. 78-81.

[12] Jouini, O. and Roubos, A., On Multiple Priority Multi-Server Queues with Impatience, *Journal of the Operational Research Society*, 2014, Vol. 65, No. 5, pp. 616-632.

[13] Kim, K. and Chae, K.C., Discrete-Time Queues with Discretionary Priorities, *European Journal of Operational Research*, 2010, Vol. 200, No. 2, pp. 473-485.

[14] Kim, K., (N, n)-Preemptive Priority Queues, *Performance Evaluation*, 2011, Vol. 68, No. 7, pp. 575-585.

[15] Kim, K., The Analysis of an Opportunistic Spectrum Access with a Strict T-preemptive Priority Discipline, *Journal of Society of Korea Industrial and Systems Engineering*, 2012, Vol. 35, No. 4, pp. 162-170.

[16] Kim, K., T-Preemptive Priority Queue and Its Application to the Analysis of an Opportunistic Spectrum Access in Cognitive Radio Networks, *Computers & Operations Research*, 2012, Vol. 39, No. 7, pp. 1394-1401.

[17] Lee, Y. and Lee, K.-S., Discrete-Time Queue with Preemptive Repeat Different Priority, *Queueing Systems*, 2003, Vol. 44. No. 4, pp. 399-411.

[18] Ma, Z., Hao, Y., Wang, P., and Cui, G., Analysis of the Geom/Geom/1 Queue under (N, n)-Preemptive Priority Discipline, *Journal of Information & Computational Science*, 2015, Vol. 12, No. 3, pp. 1029-1036.

[19] Ma, Z., Zheng, X., Xu, M., and Wang, W., Performance Analysis and Optimization of the (N, n)-Preemptive Priority Queue with Multiple Working Vacation, *ICIC Express Letters*, 2016, Vol. 10, No. 11, pp. 2735-2741.

[20] Paterok, M. and Ettl, M., Sojourn Time and Waiting Time Distributions for M/GI/1 Queues with Preemption-Distance Priorities, *Operations Research*, 1994, Vol. 42, No. 6, pp. 1146-1161.

[21] Sharif, A.B., Stanford, D.A., Taylor, P., and Ziedins, I., A Multi-Class Multi-Server Accumulating Priority Queue with Application to Health Care, *Operations Research for Health Care*, 2014, Vol. 3, No. 2, pp. 73-79.

[22] Stanford, D.A., Taylor, P., and Ziedins, I., Waiting Time Distributions in the Accumulating Priority Queue, *Queueing Systems*, 2014, Vol. 77, No. 3, pp. 297-330.

[23] Takagi, H., *Queueing Analysis, Volume 1 : Vacation and Priority Systems, Part 1*. Amsterdam : North-Holland, 1991.

[24] Walraevens, J., Maertens, T., and Bruneel, H., A Semi-Preemptive Priority Scheduling Discipline : Performance Analysis, *European Journal of Operational Research*, 2013, Vol. 224, No. 2, pp. 324-332.

[25] Walraevens, J., Steyaert, B., and Bruneel, H., A Preemptive Repeat Priority Queue with Resampling : Performance Analysis, *Annals of Operations Research*, 2006, Vol. 146, No. 1, pp. 189-202.

ORCID

Kilhwan Kim | http://orcid.org/0000-0002-0577-7906