Hindawi
Journal of Applied Mathematics
Volume 2022, Article ID 5834258, 20 pages
https://doi.org/10.1155/2022/5834258

Research Article

(N, n)-Preemptive-Priority M/G/1 Queues with Finite and Infinite Buffers

Kilhwan Kim

Department of Management Engineering, Sangmyung University, Republic of Korea

Correspondence should be addressed to Kilhwan Kim; khkim@smu.ac.kr

Received 8 November 2021; Revised 22 February 2022; Accepted 26 February 2022; Published 22 March 2022

Academic Editor: Anum Shafiq

In this paper, we analyze an M/G/1 priority queueing model with finite and infinite buffers under the (N, n)-preemptive priority discipline, under which preemption decisions are made based on the number of high-priority customers. This priority queueing model can be used for the performance analysis of communication systems accommodating delay- and loss-sensitive packets simultaneously. To analyze the proposed model, we extend the method of delay cycle analysis and develop a queue length version of it for finite-buffer queues. Throughout our analysis, we demonstrate that by the proposed method the analysis of the complex priority queueing model can be reduced to that of simple delay cycles, so two different preemption modes of the queueing model can be dealt with in a unified way. The numerical study reveals that adjusting the decision variables N and n allows us to fine-tune system performance for different classes of customers, and N operates as a primary control variable, regardless of the preemption mode and service-time distributions.

1. Introduction

In this paper, we consider an M/G/1 priority queueing model with a finite buffer for high-priority customers and an infinite buffer for low-priority customers. We call this an M/G/1/(K, ∞) priority queue. The priority queue operates under the (N, n)-preemptive priority discipline. This is a flexible priority discipline in which the decision whether to preempt a low-priority customer in service is made based on the number of high-priority customers in the system.

In general, priority queueing models are extensively used for the analysis of communication and computer systems [1–6]. Priority queueing models with finite and infinite buffers have been studied mainly for the analyses of communication systems used to service heterogeneous streams of traffic simultaneously [7–12]. In such a system, there is loss-sensitive traffic as well as delay-sensitive traffic arriving at the system. To accommodate heterogeneous service needs of these two types of traffic, the loss-sensitive traffic is offered space priority with an effectively infinite buffer, while the delay-sensitive traffic is given time priority with the privilege of getting serviced earlier than the loss-sensitive traffic. However, all the studies mentioned above are restricted to priority queueing models with finite and infinite buffers under the nonpreemptive priority discipline. Although the nonpreemptive priority discipline is more straightforward to implement than the preemptive priority discipline, there is a major drawback that the service quality for high-priority customers may degrade severely when the service time of a low-priority customer is long or varies significantly. For this reason, there have been recent studies on preemption-based scheduling methods for communication systems that accommodate both high-volume multimedia traffic and delay-sensitive traffic [2–6]. In this line of thought, [13] has recently analyzed an M/G/1/(K, ∞) priority queueing model under the preemptive-resume priority discipline as well as under the nonpreemptive one. To do this, [13] developed a delay cycle analysis of single-class finite-buffer M/G/1 queues and combined it with the traditional delay cycle analysis of infinite-buffer M/G/1 priority queues. [14] also has studied M/G/1(K, ∞) priority queues under the preemptive-repeat-different priority discipline as well as under the preemptive-repeat-identical discipline.

In this study, we further extend the method proposed in [13] to a more sophisticated M/G/1/(K, ∞) priority queueing model under the (N, n)-preemptive priority discipline. The (N, n)-preemptive priority discipline was proposed in [15] to deal with the drawbacks of the classical nonpreemptive and preemptive priority disciplines. As mentioned in [15], both nonpreemptive and preemptive disciplines are extreme cases with respect to preemption conditions. During servicing of a low-priority customer, the former never allows high-priority customers to preempt the service, regardless of how many high-priority customers are waiting for their services in the system, while the latter always does preempt the service on the arrival of high-priority customers. To refine both extreme conditions for preemption, several preemption-based dynamic priority disciplines have been proposed [15–19], and the (N, n)-preemptive priority discipline is one of such priority disciplines. Under the (N, n)-preemptive priority discipline, high-priority customers can preempt a low-priority customer in service when the number of high-priority customers in the system reaches N, $N \geq 1$. Then, the preempted service of the low-priority customer is restored when the number of high-priority customers drops to n, $0 \leq n \leq N-1$. [15] also showed that, in the M/G/1 priority queue with infinite buffers for both high- and low-priority customers under the (N, n)-preemptive priority discipline, the performance measures of both classes of customers can be balanced and fine-tuned by adjusting the thresholds N and n. While [15] dealt with infinite-buffer M/G/1 queueing models under the (N, n)-preemptive resume and (N, n)-preemptive repeat-identical priority disciplines, [20] studied an infinite-buffer M/G/1 queueing model under the (N, n)-preemptive repeat-different priority disciplines. There are also studies on discrete-time queueing models with geometric service times under the (N, n)-preemptive priority discipline [21, 22]. However, all these studies are limited to priority queueing models with infinite buffers for both high- and low-priority customers. To the best of the author's knowledge, there are no studies on priority queueing models with finite and infinite buffers under any preemption-based dynamic priority discipline, including the (N, n)-preemptive priority discipline, rather than the classical nonpreemptive and preemptive disciplines. Only work on priority queueing models with finite and infinite buffers under a sophisticated priority rule other than classical priority disciplines is [11, 12], where an arriving customer can determine whether to join a finite high-priority queue or an infinite low-priority queue. However, these queueing models employ a nonpreemptive dynamic priority discipline without preemption, so they are different from our study because we consider a preemption-based dynamic priority discipline here. As mentioned above, since the classical nonpreemptive and preemptive priority disciplines are rigid and static, a more sophisticated preemptive-based dynamic priority discipline needs to be applied to service systems with finite and infinite buffers to balance service performance between the classes.

In this study, the delay cycle analysis of finite-buffer M/G/1 queues proposed in [13], where the waiting time distribution of the delay cycle is derived, is extended to develop a queue length version of the delay cycle analysis of finite-buffer M/G/1 queues. The reason for this is that the derivation of queue length distributions is more straightforward under the (N, n)-preemptive priority discipline. This is due to the nature of the discipline depending on the queue length of high-priority customers. Using the proposed queue length version of the delay cycle analysis of finite-buffer M/G/1/K queues, we then analyze the structure of the effective service time of a low-priority customer under two different preemption modes of the (N, n)-preemptive disciplines: the preemptive resume (PR) and preemptive repeat-identical (PRI) modes (see Section 2). Then, combining the developed finite-buffer delay cycle analysis with the traditional delay cycle analysis of infinite-buffer priority queues, we derive the customer loss probability and queue length distribution of high-priority customers and the queue length distribution of low-priority customers for M/G/1/(K, ∞) priority queues under the (N, n)-PR and the (N, n)-PRI disciplines.

Furthermore, we demonstrate by numerical examples that we can balance and fine-tune system performance measures between high-priority and low-priority customers by adjusting the decision variables of N and n. The impact of the thresholds N and n on the system performance is different between the preemption modes and between the service-time distributions of a low-priority customer (see Section 7). However, regardless of the preemption mode, the threshold N can be used as a primary control variable to tune the system performances between the classes of customers, while the threshold n as a secondary control variable to fine-tune the balance between the classes.

Compared to the previous studies [13–15, 20], there are a couple of distinguishable contributions in this article. Unlike [15, 20], where infinite-buffer priority queues are considered, this study deals with a priority queue with finite and infinite buffers. Thus, the service-time structure of a low-priority customer is different from that in [15, 20], and the analysis of the service-time structure is more involved (see the first and second paragraphs in Section 4). With this new analysis, we can analyze the effect of the finite buffer size on the system performance under the (N, n)-preemptive priority discipline. On the other hand, unlike [13, 14], where the waiting-time version of the delay cycle analysis is used, this study develops a queue length version of the delay cycle analysis of finite-buffer M/G/1 queues, because it is more suitable for the analysis of finite-buffer priority queues with sophisticate preemption decisions made based on queue lengths (see Section 4).

In summary, the novelty of this study is twofold. First, to the best of the author's knowledge, this is the first study dealing with a queueing model with finite and infinite buffers under a preemption-based dynamic priority discipline other than the classical priority disciplines. Second, the method of delay cycle analysis of finite-buffer M/G/1 queue is extended to a queue length version of it. The proposed method is helpful for the analysis of complex priority queueing models with finite and infinite buffers because it can be so easily combined with the traditional delay cycle analysis of infinite-buffer queues that the analysis of a complex priority queueing model with finite and infinite buffers can be reduced to

the analysis of several simple finite- and infinite-buffer delay cycles. As a result, we can analyze similar priority queues in a unified manner by exploiting a common delay cycle structure. We will demonstrate this point throughout our analysis.

2. Model and Notation

In this paper, we consider the following M/G/1 priority queueing model with finite and infinite buffers. There are two classes of customers, namely, class 1 and class 2, arriving at the system. We assume that class-i customers arrive at the system according to a Poisson process with rate λ_i, $i = 1, 2$. The service times S_i of class-i customers follow an i.i.d. general distribution $S_i(t)$, $t \geq 0$. The Laplace-Stieltjes transform (LST) of $S_i(t)$ is denoted by $S_i^*(s)$, $\Re(s) \geq 0$. While there is an infinite buffer for class-2 customers, there is a finite buffer of size K, $K \geq 0$, for class-1 customers. Hence, when a class-1 customer is in service, there can be at most $K + 1$ class-1 customers in the system (one in service and at most K in the buffer). However, when a class-2 customer is in service, there can be at most K class-1 customers in the system (only in the buffer).

Class-1 customers are given priority over class-2 customers according to the following (N, n)-preemptive priority rule. Once the service of a class-2 customer has been started, preempting the service is not allowed if the number of class-1 customers in the system is less than a certain threshold N, $N \geq 1$. As soon as the number of class-1 customers reaches N, the class-2 customer service is immediately preempted, and a service period for class-1 customers begins. The server then works continuously for class-1 customers until the number of class-1 customers drops to another threshold n, $0 \leq n \leq N - 1$. As soon as the number of class-1 customers reaches n, the preempted service of the class-2 customer is restored in one of the following two modes: the preemptive resume (PR) and preemptive repeat-identical (PRI) modes. Under the PR mode, the preempted service is resumed when it is restored, whereas under the PRI mode, it is completely restarted with the original service time of the class-2 customer. Notice that, when $N \geq K + 1$, no preemption occurs so that the (N, n)-preemptive priority discipline reduces to the nonpreemptive priority discipline, and that when $N = 1$ and $n = 0$, preemption occurs on every arrival of class-1 customers so that the (N, n)-preemptive priority discipline reduces to the classical preemptive priority discipline. To exclude trivial cases, we assume $0 \leq n < N \leq K$ throughout the paper. Since we only consider the loss probability and the queue length distributions, we assume that the service order between customers of the same class is one of impartial service orders such as the FCSF (first come first serve), LCFS (last come first serve), or ROS (random order of service) orders. Under any impartial service order between customers of the same class, the customer loss probability and the queue length distributions are unaffected by the service order between customers of the same class.

In this paper, we derive the queue length distributions of both classes and the customer loss probability of class 1. Let $\Pi_i(z)$ and L_i, $i = 1, 2$, denote the PGF (probability generating function) and mean of the number of class-i customers in the system in a steady state. Let also P^b denote the loss probability of class-1 customers in a steady state due to no vacancy in the buffer.

We also introduce a standard notation with respect to a random variable Y which denotes the length of a certain time period. Y can be one of the random variables in Table 1; the definitions of which will be given when they are first used in the subsequent sections. Throughout the paper, $A(Y)$, $A^a(Y)$, and $A^b(Y)$ denote the numbers of class-1 customers who arrive, who arrive and enter the system, and who arrive but are lost due to no vacancy in the buffer during the time period Y, respectively. Thus, $A(Y) = A^a(Y) + A^b(Y)$. Let $Y^*(s)$ denote the LST of Y, and $Y^*(s, z, w)$ denote the following joint transform:

$$Y^*(s, z, w) = E\left[e^{-sY} z^{A^a(Y)} w^{A^b(Y)}\right], \Re(s) \geq 0, |z|, |w| \leq 1. \quad (1)$$

Hence, $Y^*(s) = Y^*(s, 1, 1)$ from the definition. We also define the partial transform $Y^*(s; j)$ as

$$Y^*(s; j) = P\{A(Y) = j\} \cdot E\left[e^{-sY} \mid A(Y) = j\right]. \quad (2)$$

Hence, $Y^*(s) = \sum_{j=0}^{\infty} Y^*(s; j)$ and $Y^*(0; j) = P\{A(Y) = j\}$ from the definition. In addition, we define $\Pi_Y(z)$ and L_Y to be the PGF and mean of the number of class-1 customers during the Y period, P_Y^b to be the loss probability of a class-1 customer during the Y period, and π_Y to be the probability that a class-1 customer arrives during an Y period in a steady state.

3. A Queue Length Version of the Delay Cycle Analysis of Finite-Buffer M/G/1 Queues

Let $\Theta_{(k)}$ denote the length of the standard busy period, starting with only one customer and terminating when the system becomes empty of customers, in the M/G/1/$k + 1$ queue where the buffer size is k, the service time is S_1, and the arrival rate is λ_1. Then, the joint transform of $\Theta_{(k)}$, $A^a(\Theta_k)$, and $A^b(\Theta_k)$ can be obtained by the recursive equation (4) in [13], and the customer loss probability $P^b_{\Theta_{(k)}}$ during the $\Theta_{(k)}$ period is also given by the equation (5) in [13]. Hence, we assume that the moments of $\Theta_{(k)}$, $A^a(\Theta_k)$, and $A^b(\Theta_k)$, and the customer loss probability $P^b_{\Theta_{(k)}}$ are given throughout this paper.

In order to derive the queue length distribution, we now extend the delay cycle analysis of single-class finite-buffer M/G/1 queues proposed in [13], where the waiting time distribution is derived rather than the queue length distribution. The reason why the queue length distribution is considered here is that the derivation of queue length distributions is more straightforward under the (N, n)-preemptive priority discipline because the queue length plays a main role in the decision of whether to preempt.

Table 1: Random variables representing a certain time period.

Symbol	Definition
S_i	The service time of a class-i customer, $i = 1, 2$
G	The gross service time of a class-2 customer
C	The completion time of a class-2 customer
R	The occupation time of a class-2 customer
I	The idle period
$\Theta_{(k)}$	The standard busy period starting with a single customer in the M/G/1/$k+1$ queue
$\Theta_{(k;a,b)}$	The busy period starting with a customers and terminating as soon as the queue size reaches b in the M/G/1/$k+1$ queue, $0 \le b \le a \le k$

Let $\Pi_{\Theta_{(k)}}(z)$ denote the PGF of the number of customers during the $\Theta_{(k)}$ period in a steady state. Also, let $\Theta_{(k;a,b)}$ denote the busy period starting with a customers and terminating when the queue size reaches b, $0 \le b \le a \le k$, in the M/G/1/$k+1$ queue with λ_1 and S_1 as its arrival rate and service time. Hence, $\Theta_{(k)}$ is identical to $\Theta_{(k;1,0)}$. Observe that the $\Theta_{(k)}$ period can be viewed as a delay cycle with the first service time S_1 as the initial delay and the remaining part of $\Theta_{(k)}$ as the delayed busy period of the delay cycle (see Section 3 in [13]). Consider also the number $A(S_1) = j$, $j \ge 0$, of customers who arrive at the system during the initial delay S_1. If $A(S_1) = j < k$, the delayed busy period following the initial delay S_1 starts with j customers and lasts until the system is empty of customers. Thus, the delay cycle $\Theta_{(k)}$ period is decomposed into the initial delay of S_1 and the delayed busy period of $\Theta_{(k;j,0)}$ when $A(S_1) = j < k$. If $A(S_1) \ge k$, the delayed busy period following the initial delay S_1 starts with k customers, due to the finite buffer, and lasts until the system is empty of customers. Thus, the delay cycle $\Theta_{(k)}$ period is decomposed into the initial delay of S_1 and the delayed busy period of $\Theta_{(k;k,0)}$ when $A(S_1) \ge k$. Therefore, if we let $\Pi_{S_1,(k)}(z)$ denote the PGF of the number of customers during the initial delay S_1, and $\Pi_{\Theta_{(k;a,b)}}(z)$ denote the PGF of the number of customers during the $\Theta_{(k;a,b)}$ period, we have [13]

$$\Pi_{\Theta_{(0)}}(z) = \Pi_{S_1,(0)}(z), \tag{3}$$

$$\begin{aligned}\Pi_{\Theta_{(k)}}(z) = {} & \frac{E[S_1]}{E\left[\Theta_{(k)}\right]}\Pi_{S_1,(k)}(z) + \sum_{j=1}^{k-1}\frac{P\{A(S_1)=j\}E\left[\Theta_{(k;j,0)}\right]}{E\left[\Theta_{(k)}\right]}\Pi_{\Theta_{(k;j,0)}}(z) \\ & + \frac{P\{A(S_1)\ge k\}E\left[\Theta_{(k;k,0)}\right]}{E\left[\Theta_{(k)}\right]}\Pi_{\Theta_{(k;k,0)}}(z), k = 1, 2, \cdots.\end{aligned} \tag{4}$$

We now derive the term $\Pi_{S_1,(k)}(z)$. Let S_1^E denote the elapsed service time of S_1 and $A(S_1^E)$ and the number of customers who arrive during S_1^E. Then, from PASTA and renewal theory, we have

$$P\left\{A\left(S_1^E\right) = j\right\} = \frac{P\{A(S_1) > j\}}{\lambda_1 E[S_1]} = \frac{\sum_{h=j+1}^{\infty} S_1^*(0;h)}{\lambda_1 E[S_1]}, j = 0, 1, \cdots. \tag{5}$$

Observe that there are $\min\{A(S_1^E), k\}$ plus one customers in the system during the the initial delay S_1 in a steady state. More specifically, $\min\{A(S_1^E), k\}$ customers who arrive and enter the system during S_1^E are in the buffer, and one customer who initiates the $\Theta_{(k)}$ period is in service. Thus, $\Pi_{S_1,(k)}(z)$ is given by

$$\begin{aligned}\Pi_{S_1,(k)}(z) &= z\left[\sum_{j=0}^{k-1} P\left\{A\left(S_1^E\right) = j\right\}z^j + P\left\{A\left(S_1^E\right) \ge k\right\}z^k\right] \\ &= \frac{\sum_{j=0}^{k-1} z^{j+1}\sum_{h=j+1}^{\infty} S_1^*(0;h) + z^{k+1}\sum_{j=k}^{\infty}\sum_{h=j+1}^{\infty} S_1^*(0;h)}{\lambda_1 E[S_1]}, k = 0, 1, \cdots.\end{aligned} \tag{6}$$

We now argue that the $\Theta_{(k;a,b)}$ period can be decomposed into the $\Theta_{(k-a+1)}$, $\Theta_{(k-a)}$, …, $\Theta_{(k-b)}$ periods, $0 \le b \le a \le k$. Without loss of generality, we can assume the LCFS service order when considering the length of a busy period and the queue length distribution during that busy period. From the definition of $\Theta_{(k;a,b)}$, there are a customers in the system at the beginning of the $\Theta_{(k;a,b)}$ period. Consider the jth customer among those initial a customers in the order of their arrivals (we assume that ties are broken arbitrarily). Under the LCFS order, the jth customer, $b+1 \le j \le a$, can begin his/her service when there are only j customers in the system. More specifically, in the order of their arrival, the 1st, 2nd, ⋯, $(j-1)$th, jth customers who were initially present in the system. Thus, the $\Theta_{(k;a,b)}$ period can be decomposed into consecutive periods initiated by the jth customers, $j = a, a-1, \cdots, b+1$, in the reverse order of their arrivals, that starts from the beginning of the jth customer's service and lasts until the number of customers in the system drops to $j-1$ for the first time. Note that these periods initiated by the jth customer, $b+1 \le j \le a$, are stochastically identical to the standard busy periods $\Theta_{(k-j+1)}$, except for $j-1$ more customers in the buffer, because the buffer of size k is already

taken up by $j-1$ customers who arrive earlier than the jth customer so that the effective buffer size is $k-j+1$. Thus, the length of the $\Theta_{(k;a,b)}$ period is expressed as

$$\Theta_{(k;a,b)} = \sum_{j=b+1}^{a} \Theta_{(k-j+1)} = \sum_{h=k-a+1}^{k-b} \Theta_{(h)}, \tag{7}$$

and the LST of $\Theta_{(k;a,b)}$ is given by [13]

$$\Theta^*_{(k;a,b)}(s) = \prod_{h=k-a+1}^{k-b} \Theta^*_h(s) \tag{8}$$

for $0 \le b \le a \le k$. Note that, when $a = b$, $\Theta^*_{k;a,a}(s) = 1$ from the convention of the product. Also, from the same decomposition argument used for deriving (8), the PGF $\Pi_{\Theta_{(k;a,b)}}(z)$ of the number of customers during the $\Theta_{(k;a,b)}$ is given by

$$\begin{aligned}\Pi_{\Theta_{(k;a,b)}}(z) &= \sum_{j=b+1}^{a} \frac{E\left[\Theta_{(k-j+1)}\right]}{E\left[\Theta_{(k;a,b)}\right]} \Pi_{\Theta_{(k-j+1)}}(z)\, z^{j-1} \\ &= \sum_{h=k-a+1}^{k-b} \frac{E\left[\Theta_{(h)}\right]}{E\left[\Theta_{(k;a,b)}\right]} \Pi_{\Theta_{(h)}}(z)\, z^{k-h}\end{aligned} \tag{9}$$

for $0 \le b \le a \le k$.

Plugging the expression of $\Pi_{\Theta_{(k;a,b)}}(z)$ into (4) and rearranging the terms leads to

$$\Pi_{\Theta_{(0)}}(z) = \Pi_{S_1,(0)}(z) = z, \tag{10}$$

$$\Pi_{\Theta_{(k)}}(z) = \frac{E[S_1]\Pi_{S_1,(k)}(z) + \sum_{h=1}^{k-1} E\left[\Theta_{(h)}\right]\Pi_{\Theta_{(h)}}(z) z^{k-h} \sum_{j=k-h+1}^{\infty} S_1^*(0;j)}{S_1^*(0;0) E\left[\Theta_{(k)}\right]}, \quad k = 1, 2, \cdots. \tag{11}$$

Note that $\Pi_{\Theta_{(k)}}(z)$ is different from the PGF of the number of customers in the M/G/1/$k+1$ queue, where we have to consider the number of customers during the idle period as well as that during the busy period $\Theta_{(k)}$. Thus, if we let $\Pi_{(k)}$ denote the PGF of the number of customers in the M/G/1/k+1 queue, then we have

$$\Pi_{(k)} = \frac{1/\lambda_1 + E\left[\Theta_{(k)}\right]\Pi_{\Theta_{(k)}}(z)}{1/\lambda_1 + E\left[\Theta_{(k)}\right]}. \tag{12}$$

Furthermore, when $k = 0$ and 1, we can obtain the PGF of the number of customers for the M/G/1/1 and M/G/1/2 queues as follows:

$$\begin{aligned}\Pi_{(0)} &= \frac{1 + \lambda_1 E[S_1] z}{1 + \lambda_1 E[S_1]}, \\ \Pi_{(1)} &= \frac{S_1^*(0;0) + z(1 - S_1^*(0;0)) + z^2(S_1^*(0;0) - 1 + \lambda_1 E[S_1])}{S_1^*(0;0) + \lambda_1 E[S_1]},\end{aligned} \tag{13}$$

which correspond the established results of the M/G/1/K queues [23], pp. 206–208.

Let $L_{\Theta_{(k)}}$ denote the mean number of customers during the $\Theta_{(k)}$ period. Then, from (6), (10), and (11), we have

$$L_{\Theta_{(0)}} = \Pi_{\Theta_{(0)}}{}'(1) = L_{S_1,(0)} = 1, \tag{14}$$

$$\begin{aligned}L_{\Theta_{(k)}} &= \Pi_{\Theta_{(k)}}{}'(1) \\ &= \frac{E[S_1] L_{S_1,(k)} + \sum_{h=1}^{k-1} E\left[\Theta_{(h)}\right]\left\{L_{\Theta_{(h)}} + (k-h)\right\} \sum_{j=k-h+1}^{\infty} S_1^*(0;j)}{S_1^*(0;0) E\left[\Theta_{(k)}\right]}, \quad k = 1, 2, \cdots,\end{aligned} \tag{15}$$

where

$$L_{S_1,(k)} = \Pi_{S_1,(k)}{}'(1) = \frac{\sum_{g=0}^{k} \sum_{j=g}^{\infty} \sum_{h=j+1}^{\infty} S_1^*(0;h)}{\lambda_1 E[S_1]}, \quad k = 0, 1, \cdots. \tag{16}$$

Notice that $L_{\Theta_{(K)}}$ is different from the mean number of customers $L_{(k)}$ in the M/G/1/$k+1$ queue with λ_1 and S_1 as the arrival rate and service time because we need to consider the mean number of customers during the idle period. Thus, $L_{(k)}$ is given by

$$L_{(k)} = \frac{\lambda_1 E\left[\Theta_{(k)}\right]}{1 + \lambda_1 E\left[\Theta_{(k)}\right]} \cdot L_{\Theta_{(k)}}. \tag{17}$$

The PGF $\Pi_{\Theta_{(k)}}(z)$ and mean $L_{\Theta_{(K)}}$ of the number of customers in M/G/1/$k+1$ queues can be derived with other established methods such as the embedded Markov chain technique and the supplementary variable technique. However, there are two reasons for developing the queue length version of the delay cycle analysis of M/G/1/$k+1$ queues. One is that the proposed delay cycle analysis of finite-buffer M/G/1/$k+1$ queues can be combined with the well-established delay cycle analysis of infinite-buffer M/G/1 priority queues so readily that the complicated analysis of M/G/1/(K, ∞) queues under a sophisticated priority discipline can rely on the well-established terminology and approaches (see Section 5). The other is that, as seen in (11) and (15), the PGF $\Pi_{\Theta_{(k)}}(z)$ and mean $L_{\Theta_{(K)}}$ of the number of customers are expressed in recursive equations only with their own terms. As a result, as the buffer size K and the thresholds N and n increase or decrease by one, changes in the system performance can be recursively calculated. This makes it

straightforward to perform a numerical study of M/G/1/(K, ∞) priority queues because there is no need to repetitively calculate the probabilities of the number of customers in the system for all values of K, N, and n (see Section 7).

4. Service-Time Structure

In the (N, n)-preemptive priority queue, three different random variables can be viewed as the service time of a class-2 customer: the gross service time G, the completion time C, and the occupation time R. The gross service time G is defined to be a total of service effort on the class-2 customer from the server to complete the class-2 customer service. If we assume the (N, n)-preemptive resume discipline, the gross service time is identical to the original service time S_2. However, if we assume the (N, n)-preemptive repeat discipline, the gross service time may be different from the original service time because the service is completely repeated after every preemption. The completion time C is defined to be a time from the beginning of the first service attempt of the class-2 customer until the class-2 customer service is completed. Note that C is composed of G and zero or more class-1 busy periods, which will be called service interruption periods, that begins when the number of class-1 customers reaches N and the class-2 customer service get preempted, and ends when the number of class-1 customers drops to n and the class-2 customer service get restored. The occupation time R is defined as a time from the beginning of the first service attempt of the class-2 customer until the server is available for the next class-2 customer (if any). R is composed of C and zero or one class-1 busy period which begins if there are class-1 customers at the end of C and ends when the system becomes empty of class-1 customers. Under the classical preemptive priority discipline, R is identical to C because it is impossible to service class-2 customers when there are any class-1 customers in the system. However, under the (N, n)-preemptive priority discipline, it is possible to service a class-2 customer even when there are class-1 customers in the system, as long as the number of class-1 customers in the system is less than N. See Figure 1 in [15] for the detailed structure of the gross, complete and occupation times under the (N, n)-preemptive queue with infinite buffers for both classes, the overall structure of which is similar to the structure of G, C, and R under the (N, n)-preemptive queue with finite and infinite buffers.

However, due to the limited buffer for class-1 customers, a service interruption period of class-2 customers by class-1 customers after service preemption and a time from the service completion of the class-2 customer until the server is available to another class-2 customer (if any) are different from those in the (N, n)-preemptive priority queue with infinite buffer for both classes. This causes the derivation of the distributions of C and R to be more involved than those in the case of infinite buffer for both classes, where every class-1 customer who arrive during the gross service time G can be viewed to initiate an identical, standard M/G/1 busy period during C or R if we assume the LCFS order between class-1 customers. This makes it straightforward to derive the distribution of C and R from the distribution of G [15], see Section 3. However, this is not the case for the finite-buffer case.

Let M denote the number of preemption occurrences during the gross service time G of a class-2 customer. Since under any impartial service order between customers of the same class, the lengths of G, C, R, and M are unaffected by the service order between customers of the same class, we can assume the LCFS order between class-1 customers without loss of generality. Assuming the LFCS order between class-1 customers, define $A(G^P)$ to be the number of class-1 customers who arrive during G and are serviced before the service completion of the class-2 customer, and $A(G^N)$ the number of class-1 customers who arrive during G but are serviced after the service completion of the class-2 customer. Therefore, the number $A(G)$ of class-1 customers who arrive during G is $A(G^P) + A(G^N)$. Observe that under the (N, n)-preemptive priority discipline and the LCFS order within the same class, class-1 customers who arrive during a service interruption period are serviced or lost during the same service interruption period because $n < N$. Thus, when the class-2 customer service is completed, he/she leaves behind only $A(G^N)$ class-1 customers when departing from the system.

While the joint transform $G^*(s, z, w) = E[e^{-sG} z^{A(G^P)} w^{A(G^N)}]$, $\Re(s) \geq 0, |z|, |w| \leq 1$, was utilized in [15], a partial transform of G and M conditioning on $A(G^N)$ will be employed here, because customers who arrive during G do not initiate their own standard, identical M/G/1 busy periods any more in the finite buffer case. Since the class-2 customer service is preempted whenever the number of class-1 customers reaches N, there can be at most $N-1$ class-1 customers after the completion of the class-2 customer service. Given $A(G^N) = j$, $0 \leq j \leq N-1$, define the partial transform of G and M to be

$$\tilde{G}_j^*(s, u) = P\{A(G^N) = j\} E[e^{-sG} u^M \mid A(G^N) = j], \Re(s) \geq 0, |u| \leq 1. \tag{18}$$

Observe that, since the LCFS order is assumed between class-1 customers, every service interruption period after preemption by class-1 customers consumes exactly $(N-n)$ class-1 customers from $A(G)$ class-1 customers who arrive during G. Thus, we have

$$A(G^P) = M(N-n), \tag{19}$$

which leads to the following relationship [15]

$$G^*(s, z, w) = \sum_{j=0}^{N-1} \tilde{G}_j^*(s, z^{N-n}) w^j. \tag{20}$$

The completion time is composed of the gross service time and service interruption periods by class-1 customers, and under the (N, n)-preemptive priority discipline, a service interruption period starts with N class-1 customers and ends when the number of class-1 customers drops to n

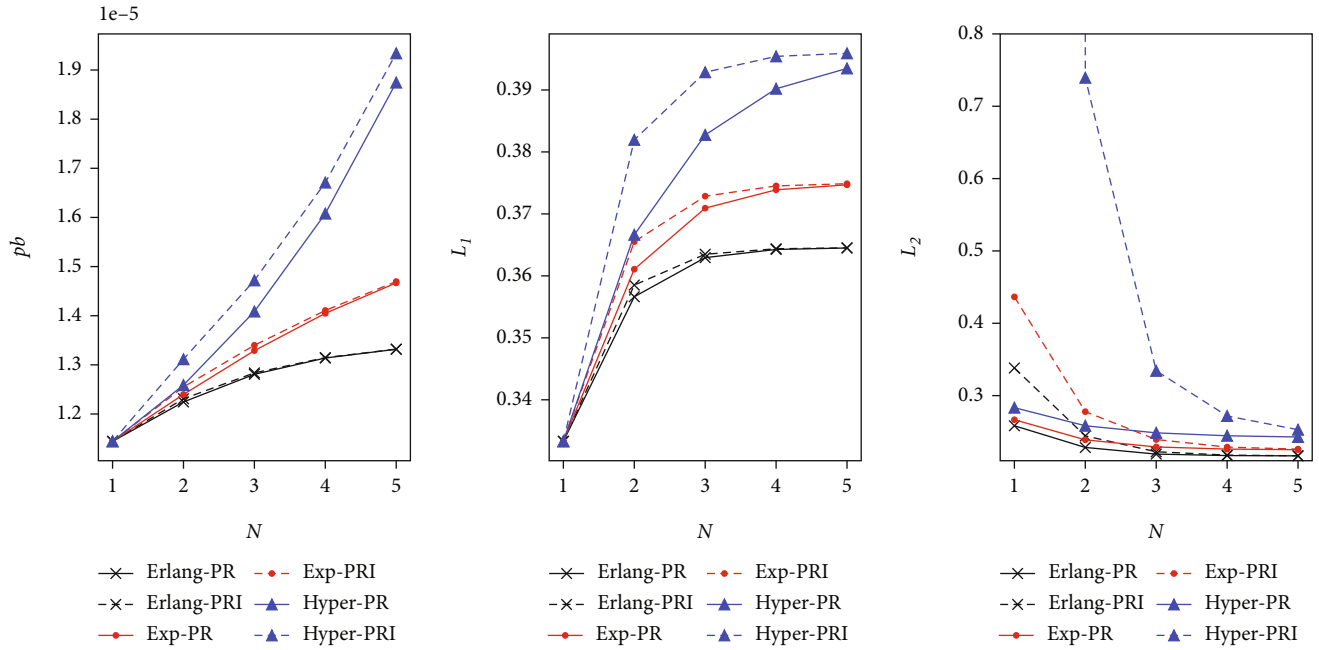

Figure 1: Performance measures for various values of N.

. Thus, if we let $\Theta_{(K;N,n)}$ denote a busy period that begins with N customers and ends as soon as the number of customers drops to n, in the M/G/1/$K+1$ queue with λ_1 and S_1 as the arrival rate and service time, then a service interruption period during C is identical to $\Theta_{(K;N,n)}$. Thus, the LST of the completion time C is expressed as

$$C^*(s) = E\left[e^{-sC}\right] = \sum_{j=0}^{N-1} \tilde{G}_j^*\left(s, \Theta^*_{(K;N,n)}(s)\right), \tag{21}$$

where $\Theta_{(K;N,n)^*(s)} = E\left[e^{-s\Theta_{(K;N,n)}}\right]$ according to our standard notation. Similarly, the occupation time is composed of C and the following class-1 busy period initiated by $A(G^N)$ class-1 customers, which is identical to the $\Theta_{(K;A(G^N),0)}$ period. Thus, the LST of R can be expressed as

$$R^*(s) = E\left[e^{-sR}\right] = \sum_{j=0}^{N-1} \tilde{G}_j^*\left(s, \Theta^*_{(K;N,\mathrm{n})}(s)\right) \cdot \Theta^*_{(K,j,0)}(s). \tag{22}$$

Differentiating (20), (21), (22), and (8) and setting $s=0$, $z=1$, and $w=1$ gives

$$E[G] = \sum_{j=0}^{N-1} g_j, \tag{23}$$

$$E[A(G)] = E\left[A\left(G^P\right)\right] + E\left[A\left(G^N\right)\right] = \sum_{j=0}^{N-1} \left(m_j(N-n) + \tilde{G}_j^*(0,1)j\right), \tag{24}$$

$$E[C] = \sum_{j=0}^{N-1} \left(g_j + m_j E\left[\Theta_{(K;N,n)}\right]\right), \tag{25}$$

$$E[R] = \sum_{j=0}^{N-1} \left(g_j + m_j E\left[\Theta_{(K;N,n)}\right] + \tilde{G}_j^*(0,1) E\left[\Theta_{(K;j,0)}\right]\right), \tag{26}$$

$$E\left[R^2\right] = \sum_{j=0}^{N-1} \left[g_j^{(2)} + 2g_j^{(1,1)} E\left[\Theta_{(K;N,n)}\right] + m_j^{(2)} E\left[\Theta_{(K;N,n)}\right]^2 + m_j E\left[\Theta^2_{(K;N,n)}\right]\right., \tag{27}$$

$$\left. + 2\left(g_j + m_j E\left[\Theta_{(K;N,n)}\right]\right) E\left[\Theta_{(K;j,0)}\right] + \tilde{G}_j^*(0,1) E\left[\Theta^2_{(K;j,0)}\right]\right], \tag{28}$$

where

$$g_j = P\left\{A\left(G^N\right) = j\right\} E\left[G \mid A\left(G^N\right) = j\right] = -\left.\frac{\partial \tilde{G}_j^*(s,u)}{\partial s}\right|_{s=0,u=1},$$

$$g_j^{(2)} = P\left\{A\left(G^N\right) = j\right\} E\left[G^2 \mid A\left(G^N\right) = j\right] = \left.\frac{\partial^2 \tilde{G}_j^*(s,u)}{\partial s^2}\right|_{s=0,u=1},$$

$$g_j^{(1,1)} = P\left\{A\left(G^N\right) = j\right\} E\left[G \cdot M \mid A\left(G^N\right) = j\right] = -\left.\frac{\partial^2 \tilde{G}_j^*(s,u)}{\partial s \partial u}\right|_{s=0,u=1},$$

$$m_j = P\left\{A\left(G^N\right) = j\right\} E\left[M \mid A\left(G^N\right) = j\right] = \left.\frac{\partial \tilde{G}_j^*(s,u)}{\partial u}\right|_{s=0,u=1},$$

$$m_j^{(2)} = P\left\{A\left(G^N\right) = j\right\} E\left[M(M-1) \mid A\left(G^N\right) = j\right] = \left.\frac{\partial^2 \tilde{G}_j^*(s,u)}{\partial u^2}\right|_{s=0,u=1},$$

$$E\left[\Theta_{(K;a,b)}\right] = \sum_{h=K-a+1}^{K-b} E\left[\Theta_{(h)}\right],$$

$$E\left[\Theta^2_{(K;a,b)}\right] = \sum_{h=K-a+1}^{K-b} E\left[\Theta^2_{(h)}\right] + 2\sum_{h=K-a+1}^{K-b} \sum_{g=h+1}^{K-b} E\left[\Theta_{(h)}\right] E\left[\Theta_{(g)}\right], \tag{29}$$

for $0 \le b \le a \le K$, and $E[\Theta_h]$ and $E[\Theta_h^2]$, $0 \le h \le K$, are given in the Appendix in [13].

We now derive the $\tilde{G}_j^*(s,u)$ for both the (N,n)-PR and (N,n)-PRI disciplines.

4.1. (N,n)-Preemptive Resume Discipline. In order to derive $\tilde{G}_j^*(s,u)$ for the (N,n)-PR priority discipline, the approach used for deriving the joint transform $G^*(s,z,w)$ in [15] is adapted for $\tilde{G}_j^*(s,u)$. Under the (N,n)-PR priority discipline, the gross service time G is identical to the service time S_2 of the class-2 customer. Note also that, if $A(G) = A(S_2) < N$, then no preemption occurs during G, and that, if $n + m(N-n) \le A(S_2) \le N - 1 + m(N-n)$ for $m = 1, 2, \cdots$, then preemption occurs m times during G, $A(G^P) = m(N-n)$, and $A(G^N) = A(G) - A(G^P) = A(S_2) - m(N-n)$ because every preemption and the corresponding service-interruption period consumes $(N-n)$ class-1 customers from $A(G) = A(S_2)$ class-1 customers. Hence, for $m = 0, 1, \cdots$, we have

$$\left(M, A\left(G^N\right)\right) = \begin{cases} (0, A(S_2)), & \text{if, } 0 \le A(S_2) \le n-1, \\ (m, A(S_2) - m(N-n)), & \text{if, } n + m(N-n) \le A(S_2) \le N - 1 + m(N-m). \end{cases} \tag{30}$$

This leads to

$$\tilde{G}_j^*(s,u) = \begin{cases} S_2^*(s\,; j), & 0 \le j \le n-1, \\ \displaystyle\sum_{m=0}^{\infty} S_2^*(s\,; m(N-n)+j)\, u^m, & n \le j \le N-1, \end{cases} \tag{31}$$

for $0 \le j \le N-1$. From the complex number theory, for integers m, j, and l, $m \ge 0$ and $n \le j, l \le N-1$, we have the following identity:

$$\sum_{q=0}^{N-n-1} \left(e^{2q\pi i/(N-n)}\right)^{m(N-n)+l-j} = \begin{cases} N-n, & l = j, \\ 0, & l \ne j, \end{cases} \tag{32}$$

where $i = \sqrt{-1}$. Thus, we can rewrite $\tilde{G}_j^*(s,u)$ for $n \le j \le N-1$ as

$$\begin{aligned} \sum_{m=0}^{\infty} S_2^*(s\,; m(N-n)+j)u^m &= \sum_{m=0}^{\infty} \sum_{l=n}^{N-1} S_2^*(s\,; m(N-n)+l) \left(u^{1/(N-n)}\right)^{m(N-n)+l-j} \frac{\sum_{q=0}^{N-n-1} \left(e^{2q\pi i/(N-n)}\right)^{m(N-n)+l-j}}{N-n} \\ &= \frac{1}{N-n} \sum_{q=0}^{N-n-1} \sum_{m=0}^{\infty} \sum_{l=n}^{N-1} S_2^*(s\,; m(N-n)+l) \left(u^{1/(N-n)} e^{2q\pi i/(N-n)}\right)^{m(N-n)+l} \left(u^{1/(N-n)} e^{2q\pi i/(N-n)}\right)^{-j} \\ &= \frac{1}{N-n} \sum_{q=0}^{N-n-1} \left\{ S_2^*\left(s + \lambda_1 - \lambda_1 u^{1/(N-n)} e^{2q\pi i/(N-n)}\right) - \sum_{r=0}^{n-1} S_2^*(s\,; r) \left(u^{1/(N-n)} e^{2q\pi i/(N-n)}\right)^r \right\} \left(u^{1/(N-n)} e^{2q\pi i/(N-n)}\right)^{-j}, \end{aligned} \tag{33}$$

where the first equality holds from the identity above and the following identity that $(u^a)^b = u^{ab}$ for a complex number u, a rational numbers a, and an integer b; the second equality holds from the following identity that $u^b v^b = (uv)^b$ for complex numbers u and v and an integer b; and the third equality holds from $S_2^*(s + \lambda_1 - \lambda_1 u) = \sum_{r=0}^{\infty} S_2^*(s\,; r) u^r$. Hence, we finally obtain the $\tilde{G}_j^*(s,u)$ for the (N,n)-PR priority discipline as follows:

$$\tilde{G}_j^*(s,u)_{NnPR} = \begin{cases} S_2^*(s\,; j), & 0 \le j \le n-1, \\ \dfrac{1}{N-n} \displaystyle\sum_{q=0}^{N-n-1} \left\{ S_2^*\left(s + \lambda_1 - \lambda_1 u_q\right) - \sum_{r=0}^{n-1} S_2^*(s\,; r) u_q^r \right\} u_q^{-j}, & n \le j \le N-1, \end{cases} \tag{34}$$

where $u_q = u^{1/(N-n)} e^{2q\pi i/(N-n)}$.

Given the distribution of S_2, differentiating (34) and setting $s = 0$ and $u = 1$ gives g_j, $g_j^{(2)}$, $g_j^{(1,1)}$, m_j, and $m_j^{(2)}$ for the (N, n)-PR priority discipline, which, in turn, gives $E[G]$, $E[C]$, $E[R]$, and $E[R^2]$ from (23)–(28). In numerical examples in Section 7, we will demonstrate that these moments can be numerically obtained for a specific distribution of S_2.

4.2. (N, n)-Preemptive Repeat-Identical Discipline. In order to derive $\tilde{G}_j^*(s, u)$ for the (N, n)-PRI priority discipline, we introduce the following notation: Given $S_2 = x$ and $A(G^N) = j$, we define the following partial transform of G and M as

$$G_{j,x}^*(s, u) = P\{A(G^N) = j \mid S_2 = x\} E[e^{-sG} u^M \mid A(G^N) = j, S_2 = x]. \tag{35}$$

Due to preemption by class-1 customers, the class-2 customer may attempt his/her service multiple times under the (N, n)-preemptive priority discipline. We now partition G into portions at each service attempt, and define G_i to be a total of service effort on the class-2 customer from the server on the ith service attempt of the class-2 customer, and $A(G_i)$ to be the number of class-1 customers who arrive during G_i, $i = 1, 2, \cdots$. Hence, $G = \sum_{i=1}^{\infty} G_i$ and $A(G) = \sum_{i=1}^{\infty} A(G_i)$. In addition, let E_r denote an Erlang random variable with rate λ_1 and shape parameter r, $r = 1, 2, \cdots$, and $E_{r,x}^*(s)$ denote the following partial transform of E_r when $E_r < x$:

$$E_{r,x}^*(s) = P\{E_r < x\} E[e^{-sE_r} \mid E_r < x]. \tag{36}$$

Then, we have

$$E_{r,x}^*(s) = \left(\frac{\lambda_1}{\lambda_1 + s}\right)^r - \sum_{h=0}^{r-1} \left(\frac{\lambda_1}{\lambda_1 + s}\right)^{r-h} \frac{(\lambda_1 x)^h}{h!} e^{-(s+\lambda_1)x}, \tag{37}$$

for $r \geq 1$ (see (8) in [15]).

We now consider the following two events at the first service attempt of the class-2 customer when $S_2 = x$.

(1) Assume that r class-1 customers, $0 \leq r \leq N - 1$, arrive during the first service attempt. Then, the class-2 customer service is completed at the first service attempt and there is no preemption during G. Hence, $G = G_1 = x$, $M = 0$, and $A(G^N) = A(G_1) = r$. Also, since class-1 customers arrive according to a Poisson arrival with rate λ_1, we have

$$P\{A(G_1) = r \mid S_2 = x\} = \frac{(\lambda_1 x)^r}{r!} e^{-\lambda_1 x}. \tag{38}$$

Thus, for $0 \leq r \leq N - 1$, we have

$$P\{A(G^N) = j, A(G_1) = r \mid S_2 = x\} E[e^{-sG} u^M \mid A(G^N) = j, A(G_1) = r, S_2 = x] = \begin{cases} \dfrac{(\lambda_1 x)^r}{r!} e^{-(s+\lambda_1)x}, & j = r \\ 0, & \text{otherwise} \end{cases} \tag{39}$$

(2) Assume that N class-1 customers arrive during the first service attempt. Then, the class-2 customer service is immediately preempted as soon as the number of class-1 customers reaches N, and then after the service interruption period of $\Theta_{(K;N,n)}$, the service is freshly restarted with the identical service time x. Notice that, if $A(G_1) = N$, then $n \leq A(G^N) \leq N - 1$ because the class-2 customer service is preempted whenever the number of class-1 customers reaches N and is restored whenever it drops to n. Let G_i^+ and M_i^+ denote the remaining gross service time and the number of preemption occurrences after the end of the ith service attempt. Then, $G = G_1 + G_1^+$ and $M = M_1 + M_1^+$. Let also the partial transform $\tilde{H}_{j,x}^*(s, u)$ denote the following partial transform of G_1^+ and M_1^+:

$$\tilde{H}_{j,x}^*(s, u) = P\{A(G^N) = j \mid A(G_1) = N, S_2 = x\} E\left[e^{-sG_1^+} u^{M_1^+} \mid A(G^N) = j, A(G_1) = N, S_2 = x\right], \tag{40}$$

for $n \leq j \leq N - 1$. Given $S_2 = x$, the event that $A(G_1) = N$ is equivalent to the event that $E_N < x$, i.e., N class-1 customers arrive before the class-2 customer service is completed. Notice also that, given $E_N < x$, $G_1 = E_N$, $G = E_N + G_1^+$, and $M = 1 + M_1^+$, and that, given $E_N < x$, G_1^+, M_1^+, and $A(G^N)$ are independent of G_1 because the class-2 customer service is completely restarted at the next service attempt and the arrival process are Poisson processes. Hence, we have

$$\begin{aligned} &P\{A(G^N) = j, A(G_1) = N \mid S_2 = x\} E[e^{-sG} u^M \mid A(G^N) \\ &= j, A(G_1) = N, S_2 = x] = P\{A(G_1) = N \mid S_2 = x\} \cdot P\{A(G^N) \\ &= j \mid A(G_1) = N, S_2 = x\} \times E\left[e^{-s(E_N + G_1^+)} u^{1+M_1^+} \mid A(G^N)\right] \\ &= j, A(G_1) = N, S_2 = x] = P\{A(G_1) = N \mid S_2 = x,\} E[e^{-sE_N} \mid A(G^N) \\ &= j, A(G_1) = N, S_2 = x] \times u \cdot P\{A(G^N) = j \mid A(G_1) = N, S_2 = x\} \\ &E\left[e^{-sG_1^+} u^{M_1^+} \mid A(G^N) = j, A(G_1) = N, S_2 = x\right] = E_{N,x}^*(s) \cdot u \cdot \tilde{H}_{j,x}^*(s, u), \end{aligned} \tag{41}$$

for $n \leq j \leq N - 1$

Combining the above two cases, we have

$$\tilde{G}_{j,x}^*(s, u) = \sum_{r=0}^{N} P\{A(G^N) = j, A(G_1) = r \mid S_2 = x\}$$

$$E\left[e^{-sG}u^{M} \mid A\left(G^{N}\right)=j, A(G_1)=r, S_2=x\right] = \begin{cases} \dfrac{(\lambda_1 x)^j}{j!}e^{-(s+\lambda_1)x}, & 0 \le j \le n-1, \\ \dfrac{(\lambda_1 x)^j}{j!}e^{-(s+\lambda_1)x} + E^*_{N,x}(s)\cdot u \cdot \tilde{H}^*_{j,x}(s,u), & n \le j \le N-1. \end{cases} \tag{42}$$

We now derive the term $\tilde{H}_{j,x}(s,u)$ in the equation for $\tilde{G}^*_{j,x}(s,u)$. Given $A(G_1)=N$, consider the following two events at the second service attempt of the class-2 customer.

(1) Assume that the number $A(G_2)$ of class-1 customers who arrive on the second service attempt of the class-2 customer is less than $N-n$. Then, the class-2 customer service is completed at the second service attempt, so $A(G_1^+)=A(G_2)$, $M_1^+=0$ and $A(G^N)=n+A(G_2)$. Hence, for $0 \le r \le N-n-1$, we have

$$P\left\{A\left(G^{N}\right)=j, A(G_2)=r \mid A(G_1)=N, S_2=x\right\} \times E\left[e^{-sG_1^+}u^{M_1^+} \mid A\left(G^{N}\right)=j, A(G_2)=r, A(G_1)=N, S_2=x\right] = \begin{cases} \dfrac{(\lambda_1 x)^r}{r!}e^{-(s+\lambda_1)x}, & j=n+r, \\ 0, & \text{otherwise} \end{cases} \tag{43}$$

(2) Assume that $A(G_2)=N-n$. Then, on the second service attempt of the class-2 customer, preemption occurs as soon as N − n class-1 customers arrive during G_2. Given $S_2=x$, the event that $A(G_2)=N-n$ is identical to the event $G_2=E_{N-n}<x$. Hence, $G_1^+=E_{N-n}+G_2^+$, and $M_1^+=1+M_2^+$. Observe that G_2^+ and M_2^+ are stochastically equivalent to G_1^+ and M_1^+ because the service is restarted with the same service time and the arrival processes are assumed to be Poisson processes. Thus, we have

$$P\left\{A\left(G^{N}\right)=j, A(G_2)=N-n \mid A(G_1)=N, S_2=x\right\} \times E\left[e^{G_1^+}u^{M_1^+} \mid A\left(G^{N}\right)=j, A(G_2)=N-n, A(G_1)=N, S_2=x\right] = E^*_{N-n,x}(s)\cdot u \cdot \tilde{H}^*_{j,x}(s,u) \tag{44}$$

Combining the above two cases, we have for $n \le j \le N-1$

$$\tilde{H}^*_{j,x}(s,u) = \sum_{r=0}^{N-n} P\left\{A\left(G^{N}\right)=j, A(G_2)=r \mid A(G_1)=N, S_2=x\right\} \times E\left[e^{G_1^+}u^{M_1^+} \mid A\left(G^{N}\right)=j, A(G_2)=r, A(G_1)=N, S_2=x\right] = \frac{(\lambda_1 x)^{j-n}}{(j-n)!}e^{-(s+\lambda_1)x} + E^*_{N-n,x}(s)\cdot u \cdot \tilde{H}^*_{j,x}(s,u), \tag{45}$$

which leads to

$$\tilde{H}^*_{j,x}(s,u) = \frac{(\lambda_1 x)^{j-n}}{(j-n)!}e^{-(s+\lambda_1)x}\left[1-E^*_{N-n,x}(s)\cdot u\right]^{-1}. \tag{46}$$

Plugging the result of $\tilde{H}^*_{j,x}(s,u)$ into the equation for $\tilde{G}^*_{j,x}(s,u)$ and eliminating the condition $S_2=x$ on $\tilde{G}^*_{j,x}(s,u)$, we finally have the partial transform $\tilde{G}^*_j(s,u)$ for the (N,n)-PRI discipline as follows:

$$\tilde{G}^*_j(s,u)_{NnPRI} = \int_{x=0}^{\infty}\tilde{G}^*_{j,x}(s,u)\,dS_2(x) = \begin{cases} \displaystyle\int_{x=0}^{\infty}\frac{(\lambda_1 x)^j}{j!}e^{-(s+\lambda_1)x}dS_2(x), & 0 \le j \le n-1, \\ \displaystyle\int_{x=0}^{\infty}\frac{(\lambda_1 x)^j}{j!}e^{-(s+\lambda_1)x}dS_2(x) + \int_{x=0}^{\infty}\frac{(\lambda_1 x)^{j-n}}{(j-n)!}e^{-(s+\lambda_1)x}\cdot\frac{E^*_{N,x}(s)\cdot u}{1-E^*_{N-n,x}(s)\cdot u}dS_2(x), & n \le j \le N-1. \end{cases} \tag{47}$$

Given the distribution of S_2, differentiating (47) and setting $s=0$ and $u=1$ gives g_j, $g_j^{(2)}$, $g_j^{(1,1)}$, m_j, and $m_j^{(2)}$ for the (N,n)-PRI priority discipline, which, in turn, gives $E[G]$, $E[C]$, $E[R]$, and $E[R^2]$ from (23) to (28). In numerical examples in Section 7, we demonstrate that these moments can be obtained numerically.

5. Loss Probability and Queue Length Distribution of Class-1 Customers

While the system is always stable for class-1 customers due to the finite buffer, it can be unstable for class-2 customers. Since only one class-2 customer is serviced during an

occupation time, if $\rho_R = \lambda_2 E[R] < 1$, then the system is stable for class-2 customers.

Assuming first that the system is stable for class-2 customer, we derive the loss probability and the queue length distribution of class-1 customers. Note that if the system is stable for class-2 customers, then a class-1 customer may arrive at the system during an idle period, a busy period for class-1 customers, or a gross service time for a class-2 customer. Class-1 customers who arrive during an idle period or a gross service time initiate class-1 busy periods, which can be grouped into three types: A first type of class-1 busy periods is those that are initiated by a single class-1 customer who arrive during an idle period, and continue until there are no class-1 customers in the system. This type of class-1 busy periods is stochastically identical to the standard busy period in the M/G/1/$K+1$ queue with λ_1 and S_1 as its arrival rate and service time, which will be denoted by the $\Theta_{(K)}$ busy period. A second type of class-1 busy periods is those that are initiated by N class-1 customers who arrive during the gross service time of a class-2 customer and preempt of the class-2 customer in service, and continue until the number of class-1 customers drops to n. This type of class-1 busy periods will be denoted by the $\Theta_{(K;N,n)}$ busy period. A third type of class-1 busy periods is those that are initiated by $A(G^N)$ class-1 customers, $0 \le A(G^N) \le N-1$, who arrive during the gross service time but are not allowed to preempt the class-2 customer service, which begin on completing the class-2 customer service and continue until there are no class-1 customers in the system. This type of class-1 busy periods will be denoted by the $\Theta_{(K;j,0)}$ busy period, $1 \le j \le N-1$.

We derive the loss probability P^b that a class-1 customer who arrive at the system is lost due to no vacancy of the class-1 buffer when $\rho_R < 1$. Denote also π_I, π_G, $\pi_{\Theta_{(K)}}$, $\pi_{\Theta_{(K;N,n)}}$, and $\pi_{\Theta_{(K;j,0)}}$ to be the probabilities that the system is in the idle period, in the gross service time, in the $\Theta_{(K)}$ busy period, in the $\Theta_{(K;N,n)}$ busy period, and in the $\Theta_{(K;j,0)}$ busy period at an arbitrary time, respectively. Then, from Little's law and PASTA we have

$$\pi_I = 1 - \lambda_1\left(1 - P^b\right)E[S_1] - \lambda_2 E[G] = 1 - \rho_1 - \rho_G + \rho_1 P^b, \tag{48}$$

$$\pi_G = \lambda_2 E[G] = \rho_G, \tag{49}$$

where $\rho_1 = \lambda_1 E[S_1]$ and $\rho_G = \lambda_2 E[G]$. Note that the $\Theta_{(K)}$ busy period occurs at a rate of $\lambda_1\pi_I$, the $\Theta_{(K;N,n)}$ busy period at a rate of $\lambda_2 E[M] = \lambda_2 \sum_{j=0}^{N-1} m_j$, and the $\Theta_{(K;j,0)}$ busy period at a rate of $\lambda_2 P\{A(G^N) = j\} = \lambda_2 \tilde{G}_j^*(0,1)$, $0 \le j \le N-1$. Thus, from Little's law, we also have

$$\pi_{\Theta_{(K)}} = \lambda_1 \pi_I E\left[\Theta_{(K)}\right], \tag{50}$$

$$\pi_{\Theta_{(K;N,n)}} = \lambda_2 \sum_{j=0}^{N-1} m_j E\left[\Theta_{(K;N,n)}\right], \tag{51}$$

$$\pi_{\Theta_{(K;j,0)}} = \lambda_2 \tilde{G}_j^*(0,1) E\left[\Theta_{(K;j,0)}\right] \quad 1 \le j \le N-1. \tag{52}$$

From (48)–(52) and (26), we have

$$\pi_G + \pi_{\Theta_{(K;N,n)}} + \sum_{j=1}^{N-1} \pi_{\Theta_{(K;j,0)}} = \lambda_2 E[R] = \rho_R. \tag{53}$$

Furthermore, since

$$\pi_I + \pi_G + \pi_{\Theta_{(K)}} + \pi_{\Theta_{(K;N,n)}} + \sum_{j=1}^{N-1} \pi_{\Theta_{(K;j,0)}} = 1, \tag{54}$$

we have

$$1 - \rho_R = \pi_I + \pi_{\Theta_{(K)}} = \pi_I\left(1 + \lambda_1 E\left[\Theta_{(K)}\right]\right). \tag{55}$$

This, together with (48), leads to

$$1 - P^b = (1 - \rho_R) \cdot \frac{E\left[\Theta_{(K)}\right]}{\rho_1\left(1/\lambda_1 + E\left[\Theta_{(K)}\right]\right)} + \rho_R \cdot \frac{E[R] - E[G]}{\rho_1 E[R]}, \rho_R < 1, \tag{56}$$

where the term $E[\Theta_{(h)}]$, $0 \le h \le K$, can be obtained from Appendix A in [13], and the terms $E[G]$ and $E[R]$ from (23), (27), and the partial transform $\tilde{G}_j^*(s,u)$ that are derived for each priority discipline in Section 4. When $N = 1$ and $n = 0$, the (N,n)-preemptive priority discipline reduces to the classical preemptive priority discipline, and from (25) and (26), we have $E[R] = E[G] + \lambda_1 E[G]E[\Theta_{(K)}]$. Thus, from (56), the loss probabilities P_{PR}^b and P_{PRI}^b for the classical PR and PRI disciplines are given by

$$1 - P_{PR}^b = 1 - P_{PRI}^b = \frac{E\left[\Theta_{(K)}\right]}{\rho_1\left(1/\lambda_1 + E\left[\Theta_{(K)}\right]\right)}, \tag{57}$$

which is identical to the loss probability $P_{M/G/1/K+1}^b$ in the M/G/1/$K+1$ queue because class-2 customers do not affect the loss probability of class-1 customers under the classical preemptive priority disciplines. Also, this result corresponds to the following basic relationship in the M/G/1/$K+1$ queue:

$$P\{\text{server is busy in the M/G/1/K + 1queue}\} = \lambda_1\left(1 - P_{M/G/1/K+1}^b\right)$$

$$E[S] = \frac{E\left[\Theta_{(K)}\right]}{1/\lambda + E\left[\Theta_{(K)}\right]}. \tag{58}$$

We now derive the PGF $\Pi_1(z)$ of the number of class-1 customers at an arbitrary time when $\rho_R < 1$. Denote $\Pi_I(z)$,

$\Pi_G(z)$, $\Pi_{\Theta_{(K)}}(z)$, $\Pi_{\Theta_{(K;N,n)}}(z)$, and $\Pi_{\Theta_{(K;j,0)}}(z)$ to be the PGF of the number of class-1 customers at an arbitrary time when the system is in the idle period, the gross service time, the $\Theta_{(K)}$ busy period, the $\Theta_{(K;N,n)}$ busy period, and the $\Theta_{(K;j,0)}$ busy period, respectively. As there are no customers in the idle period, we have

$$\Pi_I(z) = 1. \tag{59}$$

Notice that, as we assume that $N \le K$, no class-1 customers who arrive during the gross service time are lost. Hence, from PASTA the number of class-1 customers at an arbitrary time during G is identical in distribution to the number of class-1 customers immediately before a class-1 arrival during G. During $GA(G)$ class-1 customers arrive, and $A(G^P) = M(N-n)$ customers preempt the class-2 customer in service and are serviced in a service interruption period, but $A(G^N)$ customers are serviced after the class-2 customer service is completed. Since assuming the LCFS service order among class-1 customers does not affect the distribution of the number of class-1 customers, we can assume the LCFS service order between class-1 customers. Observe that under the LCFS service order, the $A(G^N)$ class-1 customers observe $0, 1, \cdots, A(G^N)-1$ class-1 customers in the system on their arrivals, respectively. Similarly, $(N-n)$ class-1 customers are serviced on every service interruption period, and under the LCFS service order, they are the latest $(N-n)$ arrivals before each service preemption and observe $n, n+1, \cdots, N-1$ class-1 customers on their arrival, respectively. Thus, we have

$$\begin{aligned}
\Pi_G(z) &= \frac{E\left[A\left(G^P\right)\right]}{E[A(G)]} \sum_{r=n}^{N-1} \frac{z^r}{N-n} + \sum_{j=1}^{N-1} \frac{P\left\{A\left(G^N\right)=j\right\}\cdot j}{E[A(G)]} \sum_{r=1}^{j-1} \frac{z^r}{j} \\
&= \frac{E[M](N-n)}{\lambda_1 E[G]} \cdot \frac{z^n - z^N}{(N-n)(1-z)} + \sum_{j=1}^{N-1} \frac{P\left\{A\left(G^N\right)=j\right\}\cdot j}{\lambda_1 E[G]} \cdot \frac{1-z^j}{j(1-z)} \\
&= \frac{\sum_{j=0}^{N-1} m_j \cdot \left(z^n - z^N\right) + \sum_{j=1}^{N-1} \tilde{G}_j^*(0,1)\cdot\left(1-z^j\right)}{\lambda_1 E[G]\cdot(1-z)}.
\end{aligned} \tag{60}$$

In addition, $\Pi_{\Theta_{(K;N,n)}}(z)$ and $\Pi_{\Theta_{(K;j,0)}}(z)$ can be obtained from (9). Hence, the PGF $\Pi_1(z)$ is expressed as

$$\begin{aligned}
\Pi_1(z) &= \pi_I \Pi_I(z) + \pi_G \Pi_G(z) + \pi_{\Theta_{(K)}} \Pi_{\Theta_{(K)}}(z) \\
&\quad + \pi_{\Theta_{(K;N,n)}} \Pi_{\Theta_{(K;N,n)}}(z) + \sum_{j=1}^{N-1} \pi_{\Theta_{(K;j,0)}} \Pi_{\Theta_{(K;j,0)}}(z) \\
&= (1-\rho_R)\cdot \frac{1+\lambda_1 E\left[\Theta_{(K)}\right]}{1+\lambda_1 E\left[\Theta_{(K)}\right]} \Pi_{\Theta_{(K)}}(z) \\
&\quad + \frac{\lambda_2 \sum_{j=0}^{N-1} m_j}{\lambda_1} \left[\frac{z^n - Z^N}{1-z} + \lambda_1 \sum_{h=K-N+1}^{K-n} E\left[\Theta_{(h)}\right] \Pi_{\Theta_{(h)}}(z)\cdot z^{K-h}\right] \\
&\quad + \frac{\lambda_2 \sum_{j=1}^{N-1} \tilde{G}_j^*(0,1)}{\lambda_1} \left[\frac{1-z^j}{1-z} + \lambda_1 \sum_{h=K-j+1}^{K} E\left[\Theta_{(h)}\right] \Pi_{\Theta_{(h)}}(z)\cdot z^{K-h}\right], \rho_R < 1,
\end{aligned} \tag{61}$$

where $\Pi_{\Theta_{(h)}}(z)$ can be obtained from from (10) and (11), $E[\Theta_{(h)}]$ from Appendix A in [13], and m_j and $\tilde{G}_j^*(0,1)$ from $\tilde{G}_j^*(s,u)$ derived for each priority discipline in Section 4.

Remark 1. The PGF $\Pi_1(z)$ in (61) exhibits a decomposition form similar to one in the corresponding infinite-buffer M/G/1 queue. If $K \longrightarrow \infty$, then the terms $\Pi_{\Theta_{(h)}}$ in (61) converges to the PGF $\Pi_{\Theta_{(\infty)}}$ of the number of customers in the corresponding standard M/G/1 queue, and $E[\Theta_{(h)}]$ to $E[S_1]/(1-\rho_1)$, the mean length of the busy period in the corresponding standard M/G/1 queue (see Lemma 1 in [13]). Thus, if $K \longrightarrow \infty$, then (61) converges to the following well-know decomposition form of M/G/1 queues:

$$\begin{aligned}
\Pi_1(z) &= \left[1 - \rho_R + \frac{\lambda_2 \sum_{j=0}^{N-1} m_j}{\lambda_1(1-\rho_1)} \cdot \frac{z^n - Z^N}{1-z} + \frac{\lambda_2}{\lambda_1(1-\rho_1)} \cdot \frac{1 - \sum_{j=1}^{N-1} \tilde{G}_j^*(0,1) z^j}{1-z}\right] \\
&\quad \times \left[1 - \rho_1 + \rho_1 \Pi_{\Theta_{(\infty)}}(z)\right],
\end{aligned} \tag{62}$$

which corresponds to the results for the (N,n)-preemptive priority queue with infinite buffers for both classes (see (17) and (18) in [15]).

From (61), the mean number L_1 of class-1 customers is given by

$$\begin{aligned}
L_1 &= (1-\rho_R)\cdot \frac{\lambda_1 E\left[\Theta_{(K)}\right]}{1+\lambda_1 E\left[\Theta_{(K)}\right]} \cdot L_{\Theta_{(K)}} \\
&\quad + \frac{\lambda_2 \sum_{j=0}^{N-1} m_j}{\lambda_1} \left[\frac{(N-n)(n+N-1)}{2} + \lambda_1 \sum_{h=K-N+1}^{K-n} E\left[\Theta_{(h)}\right]\left(L_{\Theta_{(h)}} + K - h\right)\right] \\
&\quad + \frac{\lambda_2 \sum_{j=1}^{N-1} \tilde{G}_j^*(0,1)}{\lambda_1} \left[\frac{(j-1)j}{2} + \lambda_1 \sum_{h=K-j+1}^{K} E\left[\Theta_{(h)}\right]\left(L_{\Theta_{(h)}} + K - h\right)\right], \rho_R < 1.
\end{aligned} \tag{63}$$

We now consider the case when the system is unstable for class-2 customers, i.e., $\rho_R \ge 1$. In this case, the system cycle consists of only consecutive occupation times. Thus, we can obtain the loss probability of class-1 customers from the following simple relationship:

$$p\{\text{server is busy for class}-1\text{customers}\} = \lambda_1\left(1-P^b\right)E[S] = \frac{E[R]-E[G]}{E[R]}, \tag{64}$$

which leads to

$$1 - P^b = \frac{E[R]-E[G]}{\rho_1 E[R]}, \rho_R \ge 1. \tag{65}$$

Furthermore, since the $\Theta_{(K;N,n)}$ busy period occurs M times and the $\Theta_{(K;j.0)}$ busy period occurs $\tilde{G}_j^*(0,1)$ times

during a single occupation time, the PGF $\Pi_1(z)$ is given by

$$\begin{aligned}\Pi_1(z) &= \frac{E[G]}{E[R]} \cdot \Pi_G(z) + \frac{E[M]E\left[\Theta_{(K;N,n)}\right]}{E[R]} \cdot \Pi_{\Theta_{(K;N,n)}}(z) \\ &\quad + \sum_{j=1}^{N-1} \frac{\tilde{G}_j^*(0,1)E\left[\Theta_{(K;j,0)}\right]}{E[R]} \cdot \Pi_{\Theta_{(K;j,0)}}(z) \\ &= \frac{\sum_{j=0}^{N-1} m_j}{\lambda_1 E[R]} \left[\frac{z^n - Z^N}{1-z} + \lambda_1 \sum_{h=K-N+1}^{K-n} E\left[\Theta_{(h)}\right] \Pi_{\Theta_{(h)}}(z) \cdot z^{K-h}\right] \\ &\quad + \frac{\sum_{j=1}^{N-1} \tilde{G}_j^*(0,1)}{\lambda_1 E[R]} \left[\frac{1-z^j}{1-z} + \lambda_1 \sum_{h=K-j+1}^{K} E\left[\Theta_{(h)}\right] \Pi_{\Theta_{(h)}}(z) \cdot z^{K-h}\right], \rho_R \geq 1.\end{aligned} \tag{66}$$

This also leads to

$$\begin{aligned}L_1 &= \frac{\sum_{j=0}^{N-1} m_j}{\lambda_1 E[R]} \left[\frac{(N-n)(n+N-1)}{2} + \lambda_1 \sum_{h=K-N+1}^{K-n} E\left[\Theta_{(h)}\right]\left(L_{\Theta_{(h)}} + K - h\right)\right] \\ &\quad + \frac{\sum_{j=1}^{N-1} \tilde{G}_j^*(0,1)}{\lambda_1 E[R]} \left[\frac{(j-1)j}{2} + \lambda_1 \sum_{h=K-j+1}^{K} E\left[\Theta_{(h)}\right]\left(L_{\Theta_{(h)}} + K - h\right)\right], \rho_R \geq 1.\end{aligned} \tag{67}$$

6. Queue Length Distribution of Class-2 Customers

When considering the queue length distribution of class-2 customers, we assume that the system is stable for class-2 customers, i.e., $\rho_R < 1$. If the system is stable for class-2 customers, the system cycle consists of the idle period and the overall busy period, and the overall busy period is initiated by either a single class-1 customer or a single class-2 customer and ends when the system becomes empty of customers of both classes. We will call these two types of the overall busy periods initiated by a class-1 or -2 customer the type I and II busy periods, respectively.

Let Γ_I and Γ_{II} denote the number of class-2 customers who are serviced during a type I and a type II busy period, respectively. Let also $\Pi_{I,Q}(z)$ and $\Pi_{II,Q}(z)$ denote the PGF of the number of class-2 customers in the buffer just after the beginning of the occupation time of a class-2 customer during the type I and type II busy periods, respectively. From the point of view of class-2 customers, the type I busy period can be viewed as a delay cycle consisting of an initial delay of the $\Theta_{(K)}$ period, which is initiated by the class-1 customer arriving during the idle period and ends when the system becomes empty of class-1 customers, and the following delayed busy period in the corresponding M/G/1 queue with the occupation time R and λ_2 as its effective service time and arrival rate. Additionally, the type II busy period can be viewed as the standard busy period in the M/G/1/queue with R and λ_2 as its effective service time and arrival rate. Thus, from the well-known decomposition property of the delay cycle [24], Sec. 8-3, [25], Sec. 3-3 we have

$$\begin{aligned}E[\Gamma_I] &= \frac{\lambda_2 E\left[\Theta_{(K)}\right]}{1-\rho_R}, \\ E[\Gamma_{II}] &= \frac{1}{1-\rho_R}, \\ \Pi_{I,Q}(z) &= \frac{1-\Theta_{(K)}^*(\lambda_2 - \lambda_2 z)}{\lambda_2 E\left[\Theta_{(K)}\right](1-z)} \cdot \frac{(1-\rho_R)(1-z)}{R^*(\lambda_2 - \lambda_2 z) - z}, \\ \Pi_{II,Q}(z) &= \frac{(1-\rho_R)(1-z)}{R^*(\lambda_2 - \lambda_2 z) - z}.\end{aligned} \tag{68}$$

Note that the probabilities that the overall busy period is of type I and II are $\lambda_1/(\lambda_1+\lambda_2)$ and $\lambda_1/(\lambda_1+\lambda_2)$, respectively. Hence, if we let $\Pi_{2,Q}(z)$ denote the PGF of the number of class-2 customers in the buffer just after the beginning of the occupation time of a class-2 customer in a steady state, then we have

$$\Pi_{2,Q}(z) = \frac{\lambda_1/(\lambda_1+\lambda_2) \cdot E[\Gamma_I] \cdot \Pi_{I,Q}(z) + \lambda_2/(\lambda_1+\lambda_2) \cdot E[\Gamma_{II}] \cdot \Pi_{II,Q}(z)}{\lambda_1/(\lambda_1+\lambda_2) \cdot E[\Gamma_I] + \lambda_2/(\lambda_1+\lambda_2) \cdot E[\Gamma_{II}]}. \tag{69}$$

Since a class-2 customer leaves the system in the completion time C after the beginning of the occupation time, the PGF of the number of class-2 customers left behind by a class-2 customer who depart from the system is $\Pi_{2,Q}(z) \cdot C^*(\lambda_2 - \lambda 2z)$. Hence, from Burke's argument and PASTA the PGF $\Pi_2(z)$ of the number of class-2 customers at an arbitrary time is given by

$$\begin{aligned}\Pi_2(z) &= \Pi_{2,Q}(z) \cdot C^*(\lambda_2 - \lambda 2z) \\ &= \frac{(1-\rho_R)\left[\lambda_1\left\{1-\Theta_{(K)}^*(\lambda_2-\lambda_2 z)\right\} + \lambda_2(1-z)\right]C^*(\lambda_2-\lambda_2 z)}{\lambda_2\left(\lambda_1 E\left[\Theta_{(K)}\right]+1\right)\{R^*(\lambda_2-\lambda_2 z)-z\}}.\end{aligned} \tag{70}$$

Thus, from (70), the mean number L_2 of class-2 customers is given by

$$L_2 = \frac{\lambda_1\lambda_2 E\left[\Theta_{(K)}^2\right]}{2\left(\lambda_1 E\left[\Theta_{(K)}\right]+1\right)} + \frac{\lambda_2^2 E[R^2]}{2(1-\rho_R)} + \lambda_2 E[C], \tag{71}$$

where $E[\Theta_{(K)}]$ and $E[\Theta_{(K)}^2]$ can be obtained from Appendix A in [13], and $E[C]$, $E[R]$, and $E[R^2]$ from (25) to (28) and $\tilde{G}_j^*(s,u)$ derived for each priority discipline in Section 4.

7. Numerical Examples

In this section, we perform a numerical study on the effects of the threshold N and n, the buffer size K, the arrival rates λ_1 and λ_2, and the distributions of the service times S_1 and S_2

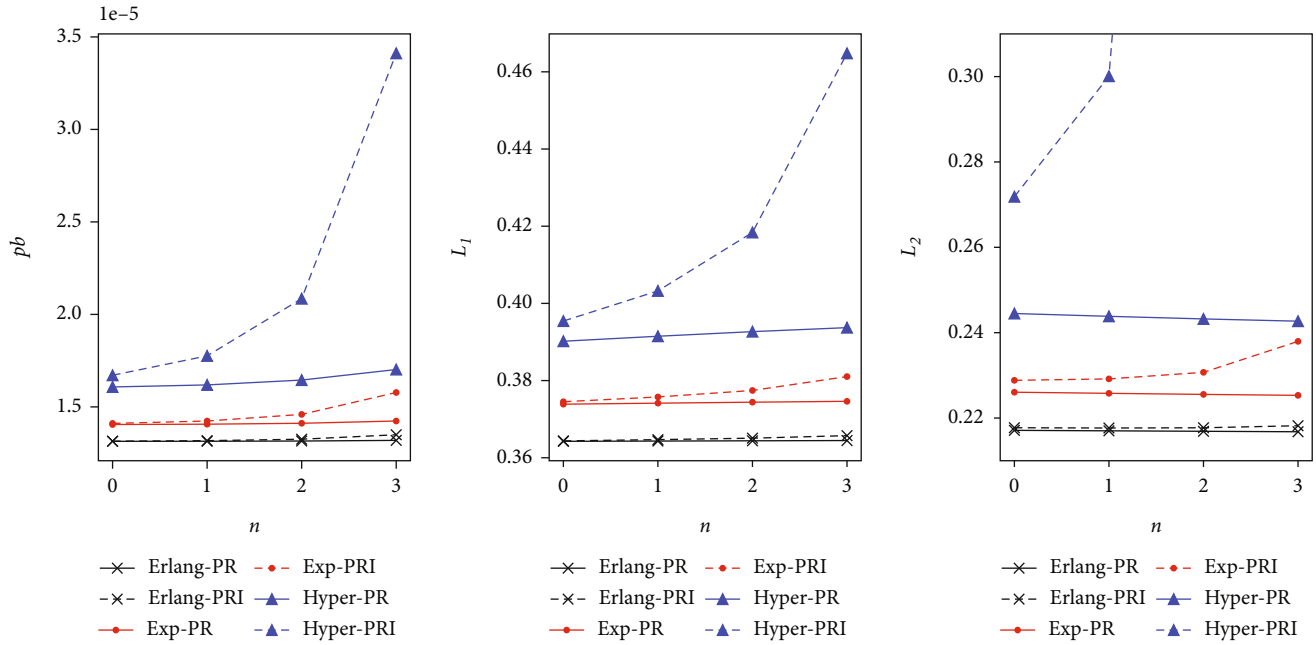

Figure 2: Performance measures for various values of n.

on performance measures in the (N, n)-preemptive priority M/G/1/(K, ∞) queue.

We first consider the effect of the threshold N on the performance measures. In order to do this, we use the following numerical example: The arrival rates λ_1 and λ_2 are set to 1 and 1/2, respectively. The service time S_1 follows an exponential distribution with rate 4. Hence, $E[S_1] = 1/4$, and $\rho_1 = 1/4$. Since the distribution of S_2 can severely affect the performance measures due to service repetition under the (N, n)-PRI discipline, we consider three different service distributions of S_2 with the same mean $E[S_2] = 1/4$: an exponential distribution with rate 4, an Erlang distribution with rate 8 and shape 2, an hyperexponential distribution with rates 8 and 2 and weights 2/3 and 1/3. The coefficients of variation (CVs) of the exponential, Erlang, hyperexponential distributions are 1, $1/\sqrt{2}$, and $\sqrt{2}$, respectively. Also, to see the effect of the threshold N, N varies from 1 to 5, while the threshold n is fixed at 0. The buffer size K for class-1 customers is set to 7.

Figure 1 shows the effect of N on the loss probability P^b, the mean queue length L_1 of class-1 customers, and the mean queue length L_2 of class-2 customers for the three different distributions of S_2 and the two preemption modes PR and PRI of the (N, n)-preemptive priority discipline. As seen in Figures 1(a) and 1(b), as N increases, P^b and L_1 increase for all the combinations of the distributions of S_2 and the preemption modes. This is because a larger N leads to less frequent preemption occurrences by class-1 customers, which is favorable to class-2 customer, but unfavorable to class-1 customers. For all the three distributions of S_2, P^b, and L_1 under the (N, n)-PRI discipline are higher than those under the (N, n)-PR discipline, except for $N = 1$. When $N = 1$ and $n = 0$, the (N, n)-PR and (N, n)-PRI disciplines reduce to the classical PR and PRI disciplines, where class-1 customers can preempt service for class-2 customers as soon as they arrive. Therefore, the loss probability and the mean queue length of class-1 customers are not affected by whether the preempted service of a class-2 customer is resumed or repeated. However, when $N > 1$, class-1 customers have to wait while a class-2 customer is in service and the number of class-1 customers is less than N. The differences in class-1 performance measures between these two disciplines for class-1 customers become more prominent for $N = 2$ or 3, and reduce for a larger N. This is because, for a sufficiently large N, preemption rarely occurs, and both (N, n)-PR and (N, n)-PRI disciplines converge to the nonpreemptive discipline. On the other hand, as seen in the left and center panels of Figure 1, the higher CV of S_2 is, the higher P^b and L_1 are for all values of N, except for $N = 1$. This is because, when $N > 1$, the (N, n)-preemptive priority discipline has nonpreemptible portion of the gross service time of a class-2 customer, and the variance of S_2 has negative impact on the performance measures of class-1 customers, as in the nonpreemptive priority discipline. Also, the higher CV of S_2 is, the more prominent the differences in class-1 performance measures between the (N, n)-PR and (N, n)-PRI disciplines are. The right panel of Figure 1 shows the effect of N on the mean queue length L_2 of class-2 customers. As N increases, L_2 decreases for all the combinations of the distributions of S_2 and the preemption modes. The effect of N on L_2 under the (N, n)-PRI discipline is more prominent than that under the (N, n)-PR discipline because a smaller N leads to more frequent preemption, the effect of which on L_2 is more amplified by service repetition under the (N, n)-PRI discipline. As with the performance measures of class-1 customers, the effect of N on L_2 is more prominent when the CV of S_2 is larger, and so is the

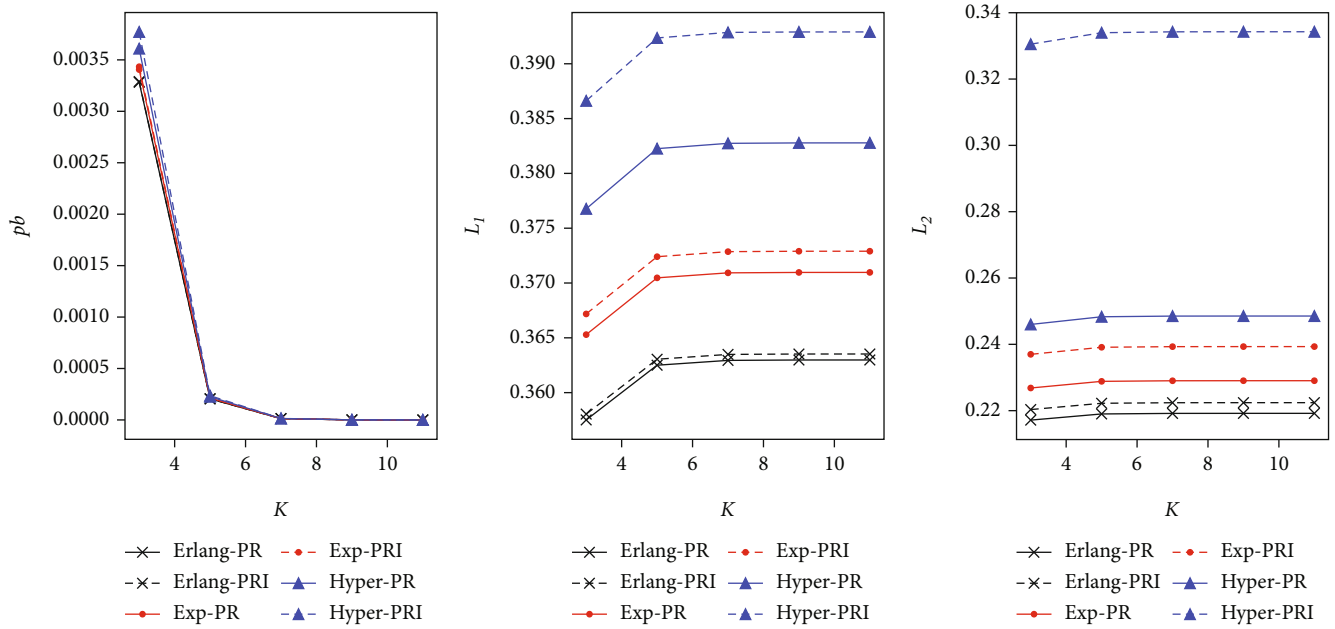

Figure 3: Performance measures for various values of K.

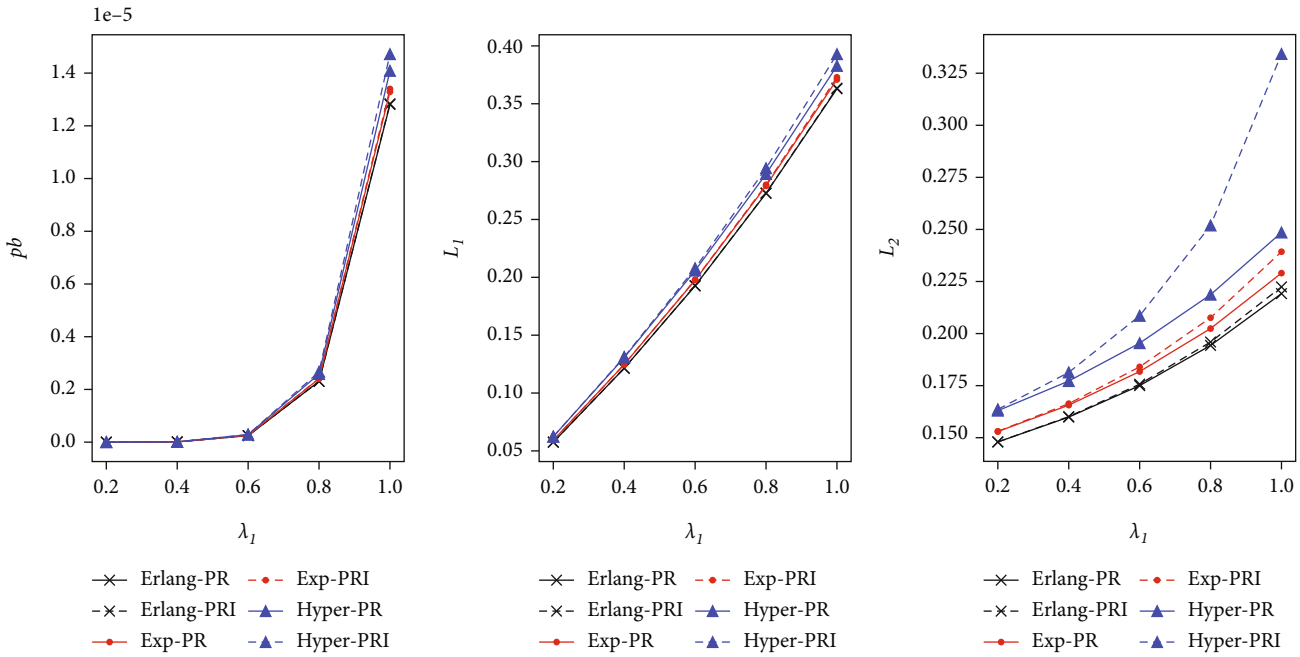

Figure 4: Performance measures for various values of λ_1.

differences in the effect of N between the (N, n)-PR and (N, n)-PRI disciplines.

We now consider the effect of the threshold n on the performance measures. In order to do this, we use the same numerical example as for the effect of N, except that, to see the effect of the threshold n, n varies from 0 to 3, while the threshold N is fixed at 4. Figure 2 shows the effect of n on P^b, L_1, and L_2 for the three different distribution of S_2 and the two preemption modes. As seen in the left and center panels of Figure 2, as n increases, P^b and L_1 under the (N, n)-PR discipline increase so slightly as the curves seem to be flat, and those under the (N, n)-PRI discipline increase more rapidly for all the distributions of S_2. This is partly because a larger n leads to a shorter busy period for class-1 customers, which is favorable to class-2 customer, but unfavorable to class-1 customers. On the other hand, for a larger n, the difference between N and n is smaller, which leads to more frequent occurrences of preemption. This alleviates the

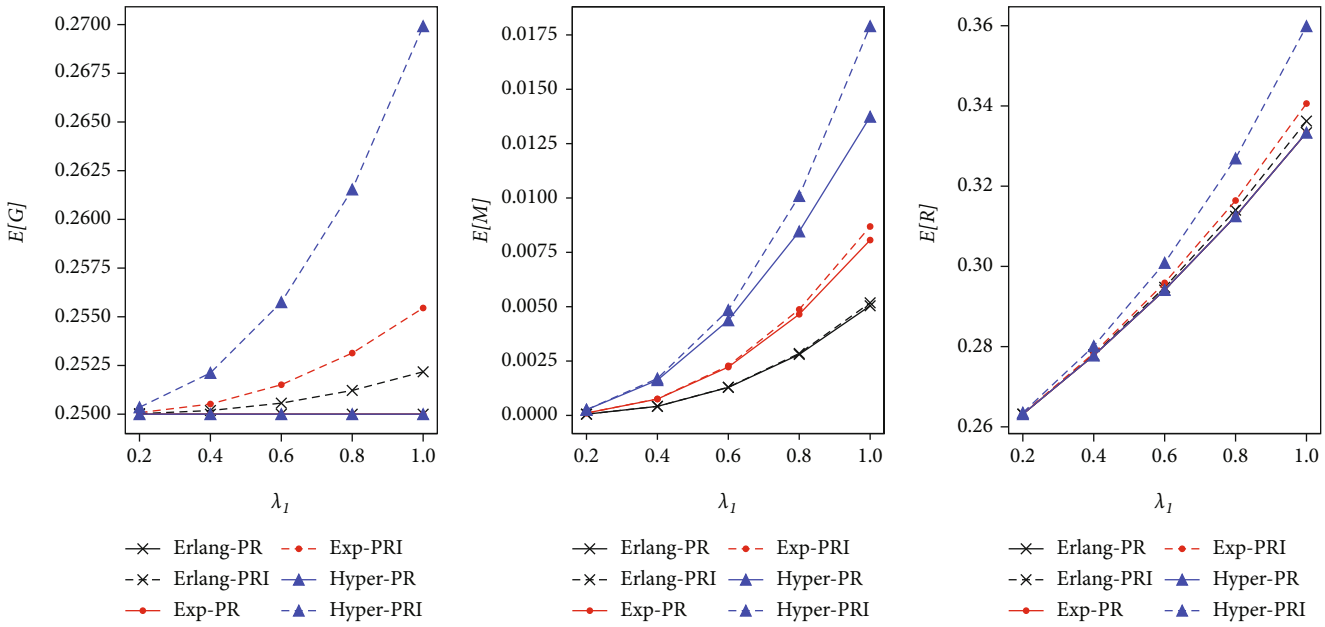

FIGURE 5: The expected gross service time, preemption occurrence, and occupation time of a class-2 customer for various values of λ_1.

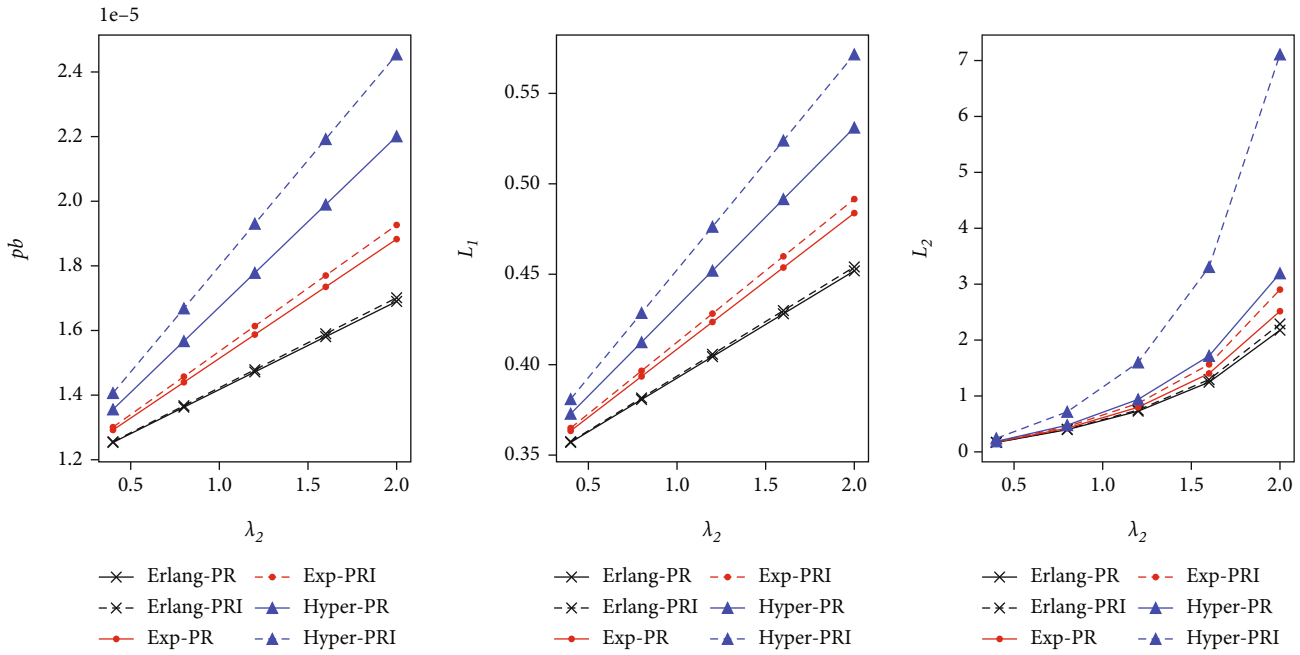

FIGURE 6: Performance measures for various values of λ_2.

negative effect of shorter class-1 busy periods caused by a larger n on the performance measures of class-1 customers under the (N, n)-PR discipline. Thus, the effect of n is almost flat for the (N, n)-PR cases. However, under the (N, n)-PRI discipline, this is not the case because more frequent occurrences of preemption result in a longer gross service time G due to service repetition, which intensifies the negative effect of a larger n on the class-1 performance measures. Also, the larger CV of S_2 is, the more prominent this negative effect is. Similarly, as seen in the right panel of Figure 2, as n increases, L_2 under the (N, n)-PR discipline decreases so slightly as the curves of L_2 seem to be flat because there is a mixed effect of shorter class-1 busy periods (service interruption time) and more frequent occurrences of preemption (occurrences of service interruptions) on the class-2 service performance. However, under the (N, n)-PRI discipline, more frequent occurrences of preemption result in a longer gross service time due to service repetition,

and the net result of shorter class-1 busy periods and more frequent occurrences of preemption becomes a prominent negative impact on the class-2 service performance, especially when the CV of S_2 is large.

We now consider the effect of the buffer size K on the performance measures. In order to do this, we use the same numerical example as for the effect of N, except that, to see the effect of K, K varies from 3 to 11, while the thresholds N and n are fixed at 3 and 0, respectively. Figure 3 shows the effect of K on P^b, L_1, and L_2 for the three different distribution of S_2 and the two preemption modes. As expected, the buffer size K has a great impact on the customer loss probability P^b, while it has a mild impact on the mean queue lengths of both classes. As seen in the left panel of Figure 3, the differences in the impact of K on P^b between the preemption modes and between the CVs of S_2 are not prominent. This is because class-1 customer loss occurs only during the class-1 busy period, which is basically not related to the service distribution of S_2 and the preemption mode of S_2, except for their effect on the probabilities $\pi_{\Theta_{(K)}}$, $\pi_{\Theta_{(K;N,n)}}$, and $\pi_{\Theta_{(K;j,0)}}$ of the type of a class-1 busy period. On the other hand, as K increases, the expected queue lengths of both classes increase, but are bounded by the corresponding values when the buffer size K is infinite when $\rho_1 < 1$.

We now consider the effect of the arrival rate λ_1 on the performance measures. In order to do this, we use the same numerical example as for the effect of N, except that, to see the effect of the arrival rate λ_1, λ_1 varies from 0.2 to 1.0, while the thresholds N and n are fixed at 3 and 0, respectively. Figure 4 shows the effect of λ_1 on P^b, L_1, and L_2 for the three different distribution of S_2 and the two preemption modes. As seen in the left and center panels of Figure 4, as λ_1 increases, the loss probability P^b rapidly soars for all the distributions of S_2 and the preemption modes, while the mean queue length L_1 rather linearly increases. This is because class-1 customer loss restricts the build-up of the class-1 queue. On the other hand, as λ_1 increases, the mean queue length L_2 rises (see the right panel in Figure 4). This is because a larger λ_1 causes a longer occupation time of a class-2 customer, which, in turn, leads to a large mean queue length of a class-2 customer. Furthermore, as in the effects of the thresholds N and n, the effect of λ_1 on L_2 becomes more prominent under the (N, n)-PRI discipline than under the (N, n)-PR discipline, and so does it for a larger CV of S_2. However, the differences in the effect of λ_1 on the class-1 performance measure P^b and L_1 between the CVs of S_2 and between the different preemption modes are less prominent than that on the class-2 performance measure L_2. The reason for this is the following: As seen in (56) and (63), P^b and L_1 are clearly related to the gross service time, the duration of which depends on the CV of S_2 and the preemption mode. Furthermore, as λ_1 increases, the gross service time increase more rapidly for a larger CV under the (N, n)-PRI discipline (see the left panel of Figure 5). However, the length of the busy period of class-1 customers increases far more rapidly than the gross service time for a sufficiently large K. This impact of a longer class-1 busy period, which is identical for all the CVs of S_2 and the preemption modes, dominates the impact of a longer gross service time, leading to the less prominent effect of λ_1 on P^b and L_1. Figure 5 confirms this explanation. Figure 5(a) shows that, as λ_1 increase, the differences in $E[G]$ between the CVs of S_2 and between the preemption modes become prominent, but the right panel shows that the difference in $E[R]$, which consists of $E[G]$ and the class-1 busy periods, is relatively less prominent than that of $E[G]$ because the portion of the class-1 busy periods in $E[R]$ is far bigger than that of $E[G]$.

We now consider the effect of the arrival rate λ_2 on the performance measures. In order to do this, we use the same numerical example as for the effect of N, except that, to see the effect of the arrival rate λ_2, λ_2 varies from 0.4 to 2.0, while the thresholds N and n are fixed at 3 and 0, respectively. Figure 6 shows the effect of λ_2 on P^b, L_1, and L_2 for the three different distribution of S_2 and the two preemption modes. As seen in the left and center panels of Figure 6, as λ_2 increases, the loss probability P^b and L_1 linearly increase for all the distributions of S_2 and the preemption modes. This is because P^b and L_1 are linear function of ρ_R as seen in (56) and (63), and ρ_R is the product of λ_2 and $E[R]$. Furthermore, the slopes of P^b and L_1 are steeper under the (N, n)-PRI discipline than under the (N, n)-PR discipline, especially when S_2 has a large CV. This is because $E[R]$ is larger under the (N, n)-PRI discipline than under the (N, n)-PR discipline when S_2 has a large CV. On the other hand, as λ_2 increases, the mean queue length L_2 rises as in an infinite-buffer M/G/1 queue (see Figure 6(c)).

Finally, we consider the effect of the CV of S_1 on the performance measures. In order to do this, we use the same numerical example as for the effect of N, except that, to see the effect of the CV of S_1, the CV of S_1 varies from 0.25 to 1.0, while $E[S_1]$ is fixed at 1/4 and the thresholds N and n at 3 and 0, respectively. For various CV values of S_1, we use Erlang distributions with the same mean. Figure 7 shows the effect of the CV of S_1 on P^b, L_1, and L_2 for the three different distribution of S_2 and the two preemption modes. As seen in the left and center panels of Figure 7, as the CV of S_1 increases, the loss probability P^b rapidly soar for all the distributions of S_2 and the preemption modes, while the mean queue length L_1 rises rather slowly. This is because the finite buffer size restricts the build-up of the class-1 queue. On the other hand, as the CV of S_1 increases, the mean queue length L_2 also rises (see the right panel in Figure 7) because a high CV of S_1 leads to a large variance of the length of the class-1 busy period, which results in a large variance of the occupation time, the effective service time of a class-2 customer. Figure 8 confirms this explanation. As the CV of S_1 increases, $E[G]$ and $E[R]$ are flat, but $E[R^2]$ increases.

In summary, the thresholds N and n, the buffer size K, the arrival rates λ_1 and λ_2, and the CVs of S_1 and S_2 all affect the performance measures P^b, L_1, and L_2. However, their effects are somewhat different from each other. One of the most prominent factors that have a significant effect on all the performance measures is the upper threshold N. More specifically, it has a more prominent effect on the blocking probability P^b and the mean queue length L_2 than on the

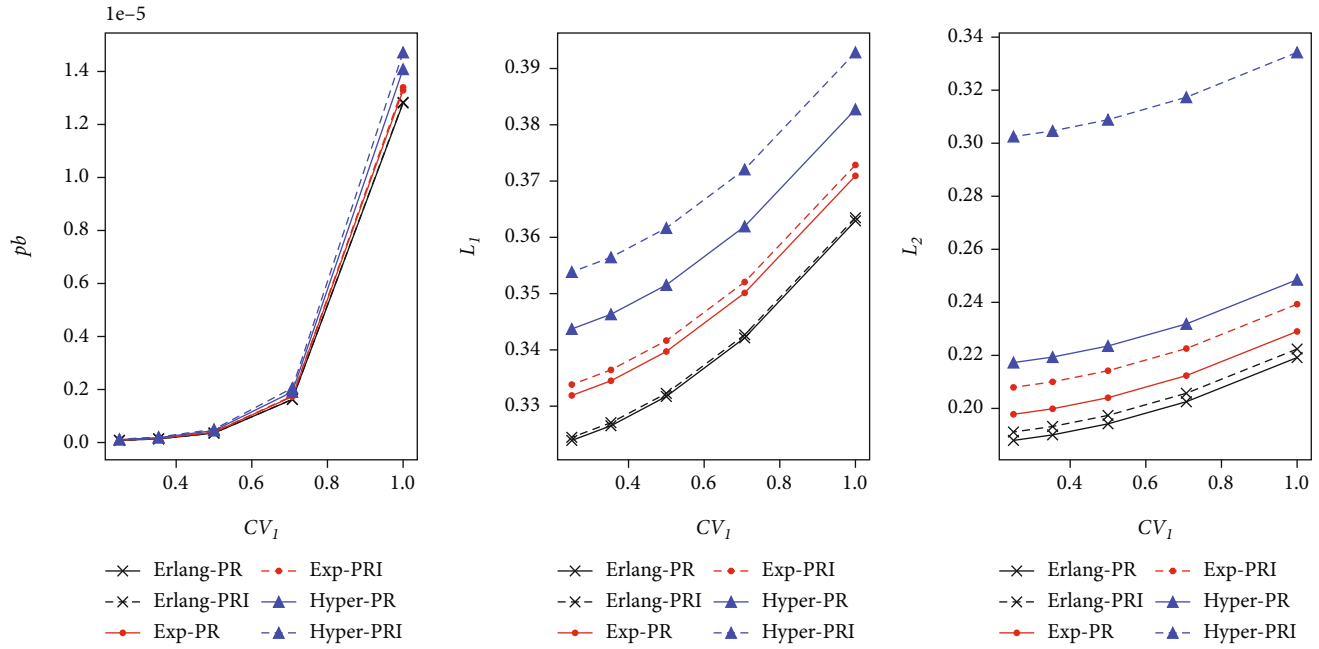

Figure 7: Performance measures for various values of CVs of S_1.

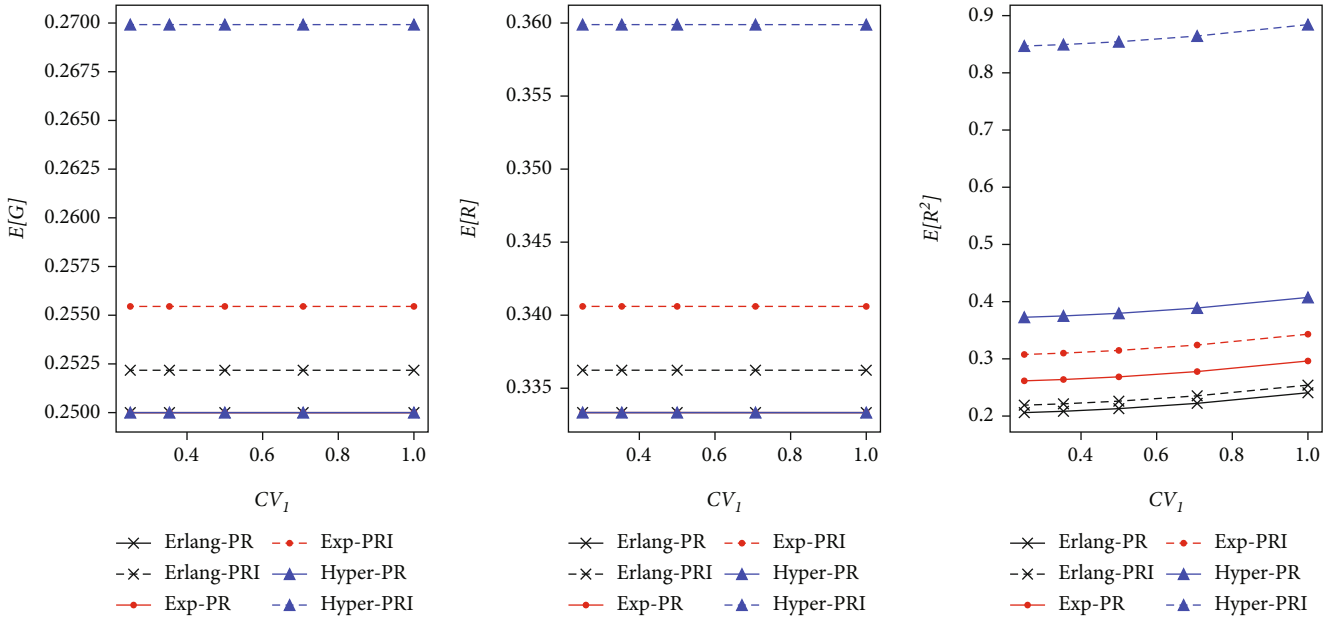

Figure 8: The expected gross service time and the first and second moments of the occupation time of a class-2 customer for various values of CVs of S_1.

mean queue length L_1 because of the finite buffer that restricts a build-up of the queue of class-1 customers. In addition, while the difference of the effects of N on P^b and L_1 between the two preemption modes is mild, those on L_2 is prominent, especially when the CV of S_2 is large. This is because N is related to the preemption frequency of class-2 customers, which has a significant effect on the effective service time of a class-2 customer under the (N, n)-PRI discipline when the CV of S_2 is large. Overall, this implies that the upper threshold N can operate as a primary control variable to govern the system performance because data traffic in communication systems generally has a heavy-tailed distribution of the service time, and data may have to be entirely retransmitted when a preemption occurs, due to a time bound of transmission. Compared to the upper threshold N, the lower threshold n has a mild effect on all the performance measure when the CV of S_2 is not large. This is because the lower threshold n has two mixed effects. One is shortening the length of a busy period of class-1 customers, which is favorable to class-2 customers. The other

is causing frequent occurrences of preemption during the service of a class-2 customer, which is unfavorable to class-2 customers. This implies that the lower threshold n may operate as a subordinate control variable to fine-tune the system performance when the CV of S_2 is small or mild. On the other hand, the buffer size K has a very significant effect on the blocking probability P^b for all the combinations of the service distributions of S_2 and the preemption modes, but it has a mild effect on the mean queue lengths L_1 and L_2. The arrival rates λ_1 and λ_2 and the CVs of S_1 and S_2 all have a significant effect on the performance measures as they increase, because an increase in them causes a larger mean and/or a larger variance of the traffic load. However, in many cases, these factors are exogenous, and we cannot control them.

From the point of a system operator, the thresholds N and n and the buffer size K are decision variables to tune the balance of system performance between the classes. While increasing K generally incurs an investment cost, adjusting the thresholds N and n does not. The effect of N on the system performance is prominent for all the distributions of S_2 and the preemption modes, while the effect of n is rather subtle under the (N, n)-PR discipline and can be negative under the (N, n)-PRI discipline. Therefore, regardless of the preemption mode, the threshold N can be used as a primary control variable to tune the system performances between the classes of customers, while the threshold n as a secondary control variable to fine-tune the balance between the classes only under the (N, n)-PR discipline.

8. Conclusions

In this paper, we analyzed an M/G/1 priority queueing model with finite and infinite buffers under the (N, n)-preemptive priority discipline. This study has two main contributions. From a practical point of view, this study will help system engineers tune their service systems for different classes of customers with heterogeneous requirements. Previous studies on priority queueing models with finite and infinite buffers have been restricted to those under the nonpreemptive priority or the classical preemptive priority disciplines. The nonpreemptive and preemptive priority disciplines are both static and rigid. When the service times of low-priority customers fluctuate, the nonpreemptive discipline may cause a severe degradation of service performance for high-priority customers. On the other hand, when preempted services of low-priority customers have to be repeated completely, the fluctuation of the arrival rates and the service times of high-priority customers may cause a severe degradation of service performance for low-priority customers. Compared to these classical priority disciplines, as shown in Section 7, the (N, n)-preemptive discipline is so flexible that whether to preempt a low-priority service is determined dynamically based on the queue length of high-priority customers, and we can adjust the thresholds N and n to balance the performance of high-priority customers and that of low-priority customers (see Section 7). This will help system engineers fine-tune their systems for different classes of customers with heterogeneous requirements, without any additional investment cost such as a cost to increase the buffer.

From a theoretical point of view, this study extended the delay cycle analysis of finite-buffer M/G/1 queues recently proposed in [13], and developed the queue length version of the finite-buffer delay cycle analysis. Delay cycle analysis is one of the standard methods to analyze M/G/1 priority queueing models. However, delay cycle analysis has been applied mostly to infinite-buffer M/G/1 priority queueing models, and few studies have used it to deal with finite-buffer priority queueing models because the delay cycle analysis of finite-buffer M/G/1 queues is more involved than that of infinite-buffer M/G/1 priority queues [13, 14]. One of the advantages of the proposed queue length version of the finite-buffer delay cycle analysis is that, when the priority discipline is a dynamic rule based on queue lengths, the analysis of the priority queue is more straightforward than the corresponding waiting-time version in [13, 14] (see Section 4). Another advantage is that, as in the waiting-time version of the delay cycle analysis of finite-buffer M/G/1 queues, the queue length version of the finite-buffer delay cycle analysis is so easily combined with the traditional delay cycle analysis of infinite-buffer queues that the analysis of a priority queueing model with finite and infinite buffers can be reduced to the analysis of several simple finite- and infinite-buffer delay cycles. As a result, similar M/G/1 priority queues with finite and infinite buffers can be analyzed in a unified manner by exploiting a common delay cycle structure among the similar priority disciplines. In fact, the two different preemption queues, the (N, n)-PR and (N, n)-PRI priority queues, was analyzed in a unified manner in this study (see Sections 4–6), and their performance measures were compared in Section 7. Therefore, it is expected that other priority queues with finite and infinite buffers that have a priority discipline based on queue lengths, such as M/G/1/(K, ∞) under the (N, n)-preemptive repeat different priority discipline or batch-arrival or/and discrete-time priority queues under the (N, n)-preemptive discipline, can be tackled with the proposed method.

Data Availability

The parameter input values used in the numerical study are described in Section 7 of the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] L. Guo, T. Yan, S. Zhao, and C. Jiang, "Dynamic performance optimization for cloud computing using M/M/m queueing system," *Journal of Applied Mathematics, vol.*, vol. 2014, pp. 1–8, 2014.

[2] W. Miao, G. Min, Y. Wu, and H. Wang, "Performance modelling of preemption-based packet scheduling for data plane in software defined networks," in *2015 IEEE international conference on Smart City/SocialCom/SustainCom (SmartCity)*, pp. 60–65, Chengdu, China, 2015.

[3] Z. Zhou, Y. Yan, S. Ruepp, and M. Berger, "Analysis and implementation of packet preemption for time sensitive networks," in *2017 IEEE 18th international conference on high performance switching and routing (HPSR)*, pp. 1–6, Campinas, Brazil, 2017.

[4] S. R. Pandey, M. Alsenwi, Y. K. Tun, and C. S. Hong, "A downlink resource scheduling strategy for URLLC traffic," in *2019 IEEE international conference on big data and smart computing (BigComp)*, pp. 1–6, Kyoto, Japan, 2019.

[5] W.-R. Wu, T.-H. Lin, and Y.-H. Lee, "An indicator-free eMBB and URLLC multiplexed downlink system with correlation-based SFBC," in *Proceedings of the 3rd international conference on telecommunications and communication engineering*, pp. 105–110, Tokyo, Japan, 2019.

[6] W. Yang, C.-P. Li, A. Fakoorian, K. Hosseini, and W. Chen, "Dynamic URLLC and eMBB multiplexing design in 5G new radio," in *2020 IEEE 17th annual consumer communications & networking conference (CCNC)*, pp. 1–5, Kuala Lumpur, Malaysia, 2020.

[7] K. Al-Begain, A. Dudin, A. Kazimirsky, and S. Yerima, "Investigation of the M2/G2/1/∞,N queue with restricted admission of priority customers and its application to HSDPA mobile systems," *Computer Networks*, vol. 53, no. 8, pp. 1186–1201, 2009.

[8] T. Demoor, J. Walraevens, D. Fiems, S. De Vuyst, and H. Bruneel, "Influence of real-time queue capacity on system contents in DiffServ's expedited forwarding per-hop-behavior," *Journal of Industrial and Management Optimization*, vol. 6, no. 3, pp. 587–602, 2010.

[9] J. Van Velthoven, B. Van Houdt, and C. Blondia, "The impact of buffer finiteness on the loss rate in a priority queueing system," in *European performance engineering workshop*, pp. 211–225, Budapest, Hungary, 2006.

[10] S. R. Chakravarthy, "A dynamic non-preemptive priority queueing model with two types of customers," in *International conference on mathematics and computing*, pp. 23–42, Varanasi, India, 2018.

[11] D. Babu, V. Joshua, and A. Krishnamoorthy, "Token based parallel processing retrial queueing system with a probabilistic joining strategy for priority customers," in *International conference on distributed computer and communication networks*, pp. 183–194, Moscow, Russia, 2020.

[12] D. Babu, V. C. Joshua, and A. Krishnamoorthy, "A queueing system with probabilistic joining strategy for priority customers," in *International conference on information technologies and mathematical modelling*, pp. 337–351, Tomsk, Russia, 2020.

[13] K. Kim, "Delay cycle analysis of finite-buffer M/G/1 queues and its application to the analysis of M/G/1 priority queues with finite and infinite buffers," *Performance Evaluation*, vol. 143, pp. 102133–102133, 2020.

[14] K. Kim, "M/G/1 preemptive priority queues with finite and infinite buffers," *Journal of the Society of Korea Industrial and Systems Engineering*, vol. 43, no. 4, pp. 1–14, 2020.

[15] K. Kim, "(N,n)-preemptive priority queues," *Performance Evaluation*, vol. 68, no. 7, pp. 575–585, 2011.

[16] S. Drekic and D. A. Stanford, "Reducing delay in preemptive repeat priority queues," *Operations Research*, vol. 49, no. 1, pp. 145–156, 2001.

[17] S. Drekic and D. A. Stanford, "A preemptive resume queue with an expiry time for retained service," *Performance Evaluation*, vol. 54, no. 1, pp. 59–74, 2003.

[18] K. Kim, "T-preemptive priority queue and its application to the analysis of an opportunistic spectrum access in cognitive radio networks," *Computers & Operations Research*, vol. 39, no. 7, pp. 1394–1401, 2012.

[19] T. E. Fahim, A. Y. Zakariya, and S. I. Rabia, "A novel hybrid priority discipline for multi-class secondary users in cognitive radio networks," *Simulation Modelling Practice and Theory*, vol. 84, pp. 69–82, 2018.

[20] K. Kim, "(N, n)-preemptive repeat-different priority queues," *Journal of Society of Korea Industrial and Systems Engineering*, vol. 40, no. 3, pp. 66–75, 2017.

[21] Z. Ma, Y. Hao, P. Wang, and G. Cui, "Analysis of the Geom/Geom/1 queue under (N, n)-preemptive priority discipline," *Journal of Information & Computational Science*, vol. 12, no. 3, pp. 1029–1036, 2015.

[22] Z. Ma, X. Zheng, M. Xu, and W. Wang, "Performance analysis and optimization of the (n, n)-preemptive priority queue with multiple working vacation," *ICIC Express Letters*, vol. 10, no. 11, pp. 2735–2741, 2016.

[23] H. Takagi, *Queueing Analysis, Volume 2: Finite Systems*, North-Holland, Amsterdam, 1993.

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

[25] L. Kleinrock, *Queueing Systems, Volume 2: Computer Applications*, Wiley, New York, 1976.