Analysis of the Finite-source Multi-class Priority Queue with an Unreliable Server and Setup Time

Pedram Sahba
University of Toronto, Department of Mechanical and Industrial Engineering
5 King's College Rd., Toronto, ON M5S 3G8, Canada,
pedram@mie.utoronto.ca
Barış Balcıoğlu
Sabancı University, Faculty of Engineering and Natural Sciences,
Orhanlı-Tuzla, 34956 Istanbul, Turkey,
balcioglu@sabanciuniv.edu
Dragan Banjevic
University of Toronto, Department of Mechanical and Industrial Engineering
5 King's College Rd., Toronto, ON M5S 3G8, CANADA,
banjev@mie.utoronto.ca

Abstract

In this paper, we study a queueing system serving multiple classes of customers. Each class has a finite-calling population. The customers are served according to the preemptive-resume priority policy. We assume general distributions for the service times. For each priority class, we derive the steady-state system size distributions at departure/arrival, and arbitrary time epochs. We introduce the residual augmented process completion times conditioned on the number of customers in the system to obtain the system time distribution. We then extend the model by assuming that the server is subject to operation-independent failures upon which a repair process with random duration starts immediately. We also demonstrate how setup times, which may be required before resuming interrupted service or picking up a new customer, can be incorporated in the model.

Keywords and Phrases: Multi-class finite-source populations, priority queues, process completion time, busy period analysis, operation-independent server disruptions

1 Introduction

In this paper, we analyze an $M/G/1//N$ queueing system with an unreliable server serving m finite-source populations/customer classes indexed by $k = 1, ..., m$. Each population k consists of N_k customers (type k customer). Such queueing models traditionally consider only a single finite-source population and a reliable server and, as such, are extensively studied in the literature. For instance, in the *machine interference problem* (MIP), N can be the number of machines in a fleet, each subject to failure; upon failure, they are repaired by the repair facility, modeled as a single server. The repair facility may be unavailable from time to time (see, e.g., [27]), thus increasing the wait times of failed machines in the repair shop. In modeling telecommunication or computer networks, e.g., [4, 24], the finite number (N) of potential customers might correspond to active terminals generating jobs for the central processor unit (CPU), which can be modeled as a single server. The CPU might be interrupted and become unavailable from time to time; jobs generated by the terminals cannot be processed until the CPU is recovered.

We assume that customers from different classes are served according to the preemptive-resume priority discipline. This setting can be modeled as a two-node closed queueing network where the second node hosts infinite-server groups. Customers departing from the single server queue occupies one of the infinite servers for an exponentially distributed amount of time (possibly with different rates for different customer types), and, then, re-enter the $M/G/1//N$ queue placed at the first node (see Figure 1 and deliberations on it in Section 2 for more clarification). There is a rich literature on closed queueing networks where one node hosts an infinite server group – as in our problem – capturing sojourn times of customers out of the queueing system while each one of the other nodes hosts a single server queue. In these studies, the focus is on the bottleneck single server system. We refer the reader to [18] for a single finite-source population, and [28] for multiple finite-source populations served under the first-come, first-served (FCFS) policy. Autonomous service at the bottleneck single server system where customers are served at random instants is considered for single and

multiple finite-source populations in [1] and [3], respectively. While exponential service times are assumed for bottleneck single server in [1, 3, 18, 28], extension of [1] with general service times can be found in [2].

In our problem, we first study the $M/G/1//N$ queue without considering server failures and setup times. A distinctive feature of our model is its capacity to include multiple classes of customers served under the preemptive-resume priority policy. Since preemptive-resume priority is used, the server becomes unavailable/disrupted for a class of customers because of arrival of higher priority customers. Such periods of interruptions end when all higher priority customers are cleared off the system. Preemptive-resume priority policy for finite-source populations is analyzed in [16] where the generating function of the queue length process is obtained. Assuming exponential service times for each class, a method to compute the steady-state distributions of the queue lengths is designed by [26] as an alternative to the computationally complex method in [16]. We also refer the reader to [21] that extends the results in [26]. In this study, we assume that service time random variables (r.v.s) have general distributions. We develop a recursive method to obtain the steady-state system size distribution, and the Laplace transform (LT) of the system time for each class in Sections 4.2 and 4.3, respectively.

After the analysis of the multi-class $M/G/1//N$ queue is completed, we consider having setup times prior to picking up the next customer or resuming the service of an interrupted customer. We also permit that the server can fail whether it is idle, under setup or serving a customer. A repair process starts immediately upon failures. We define the times between failures, or the ON periods, as the times between the end of one repair and the start of the next. We assume that ON periods are exponentially distributed. This implies that customers can experience "operation-independent disruptions (OID)" indicating that the server can be disrupted for them at any time – even when it is idle or being set up – except during the server's own OFF periods. If we assume that the characteristics of times between interruptions and down times experienced by an idle server differ from when it is serving customers, we arrive at the ODD $M/G/1//N$ queue where ODD stands for "operation-

dependent disruptions". Note that we adopt the definitions of OID and ODD from [5] (p. 85). Since our paper is on the $M/G/1//N$ queue with OID, for the sake of simplicity, we simply refer to it as the $M/G/1//N$ queue.

Queueing models with unreliable servers have been widely studied since the seminal paper by White and Christie [31]. Although the nature and the context of the problems analyzed vary considerably, the early body of work loosely revolves around two considerations: 1) whether the customer population is infinite or finite, and 2) whether the ON periods of the server(s) are operation-independent or operation-dependent.

We first summarize the papers that consider infinite populations. White and Christie assume operation-independent exponential ON periods in the $M/M/1$ queue. Assuming that OFF periods are also exponential r.v.s, they obtain the steady-state probability distribution of the time a customer spends in the system. In [7, 14, 25], this model is extended by assuming that service times and OFF periods have general distributions. In his analysis, Gaver [14] considers operation-dependent ON periods and assumes that the customer whose service is interrupted resumes its service from the moment of interruption once the OFF period is over. He introduces the *process completion time*, the total time a customer spends on the server including its actual service time plus possible OFF periods. Avt-Itzhak and Naor [7] and Thiruvengadam [25] consider both operation-dependent and operation-independent ON periods. The multi-server $M/M/c$ queues with random breakdowns are studied in [19, 20]. For $M/G/1$ queues with operation-independent ON times, bounds and approximations are derived in [12] for the mean waiting time, probability of delay and steady-state system size distribution when ON and OFF periods are general independent and identically distributed (i.i.d.) r.v.s. Federgruen and Green [13] revisit the problem, this time assuming that ON periods are phase-type r.v.s. They provide an exact algorithm to obtain the steady-state system performance measures. For the $M/G/1$ queue with interruptions, we also refer the reader to [6, 11, 30]. An accurate approximation is designed in [8] to obtain the mean waiting time in the $GI/D/1$ queue with operation-dependent phase-type ON and general OFF periods.

Next we note the papers that consider finite-calling populations, which are part of the MIP or alternatively the *machine repairperson problem* literature (see [15, 23] for an extensive bibliography on the MIP) with unreliable servers. The $M/M/1//N$ queue with an unreliable server is analyzed in [27] by assuming exponential ON and OFF periods for both operation-dependent and operation-independent interruptions. This model is extended in [29] by assuming exponential operation-independent ON periods, Erlangian service times and Erlangian OFF periods. The results in [29] are generalized by considering phase-type distributions for service times and OFF periods in [9].

As the literature review suggests, using non-exponential distributions for underlying r.v.s in these queueing systems is challenging. Neither incorporating non-exponential times between customer arrivals nor assuming non-exponential ON period distributions is analytically tractable in systems with a finite-calling population, whether these systems experience operation-dependent or operation-independent server disruptions (except in $M/G/1$ systems with phase-type ON periods as in [8, 13]). Similar difficulties arise for general service time and OFF period distributions. Among the three papers [9, 27, 29] that are relatively closest to our problem, two have successfully incorporated either Erlang distribution [29] or phase-type distributions [9] for both r.v.s. considering only a single finite population of customers to be served by the unreliable server. These studies employ the matrix-analytic method to find the steady-state system size distribution; this can be computationally intensive if the structure of the phase-type distribution is complex.

After outlining the problem in Section 2 without considering server failures and setup times, in Section 3 we conduct the busy period analysis of the system. Here, we derive its LT and the mean length of the busy period. This enables us to obtain the steady-state system size distribution at departure/arrival and arbitrary time epochs in Section 4. For the probabilities at arbitrary time epochs, we need the LT of the residual time left until the departure of the first customer in each class from the system. This is derived in Section 4.3. We summarize our conclusions in Section 5. All proofs appear in Appendix A. We include server failures in the model and redefine the process completion time r.v., this time including

setup times, and obtain its LT for each class in Appendix B.

2 Problem Definition

In this paper, we analyze a queueing system with a single server serving m finite-source populations/customer classes indexed by $k = 1, \ldots, m$. Each population k consists of N_k customers (type k customer). The times between the completion of a type k customer's service and the next arrival of the same customer at the queueing system follow an exponential distribution with rate λ_k. Customer classes are prioritized as class 1 to m from highest to lowest and customers are served according to the preemptive-resume priority policy. Therefore, if a "tagged" lower priority customer is preempted by a higher priority customer, the time until it resumes its service from the moment of preemption is a disruption for this tagged customer. The actual service times of type k customers – in the absence of disruptions – are i.i.d. r.v.s with an LT, $\widetilde{b}_k(s)$.

This problem can be represented as a two-node closed queueing network, a snapshot of which is given in Figure 1. According to this representation, one of the nodes is a single server system with two infinite capacity queues where service times are general i.i.d. r.v.s dependent on the customer type (with the LT $\widetilde{b}_k(s)$). Customers that are served according to the preemptive-resume priority policy depart from this node and high-priority type 1 (low-priority type 2) customers enter the infinite server group 1 (2) which is located at the other node of the network. Here, a type k customer stays for an exponentially distributed time with rate λ_k and is directed again to the queue reserved for its class at the single server node. In Figure 1, we have $m = 2$ finite-source populations with $N_1 = N_2 = 6$. In this snapshot, there are three type 1 customers in the $M/G/1//N$ queue at node 1, one of which is being served and two waiting in queue 1. Due to the preemptive-priority policy, type 2 customers have to wait until all type 1 customers are served. It is possible that all or some type 1 customers may have arrived at node 1 after the first type 2 customer in the queue. If this is the case, the service of the first type 2 customer was preempted, which will resume

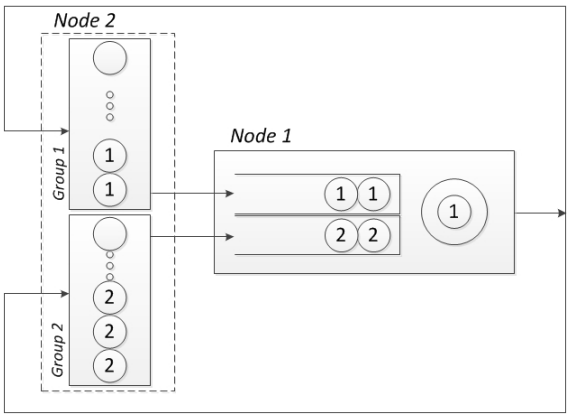

Figure 1: A two-node closed queueing network representation of the problem

from the moment of interruption only after no type 1 customers remain at node 1.

Since preemptive-resume priority policy is used, a class k customer can be serviced only during the periods the server is not allocated to higher priority classes 1 to $k-1$. The presence of lower priority classes $k+1$ to m does not have any impact on type k customers. In other words, from the point of view of type k customers, the server becomes unavailable/is disrupted with an "effective" interruption rate of $\alpha_k = \sum_{n=1}^{k-1} N_n \lambda_n$ for a random interruption period denoted by the r.v. D_k, $k = 2, \ldots, m$ due to the arrivals of the higher priority customers. The LT of the length of the interruption period D_k for type k customers, $\widetilde{f}_k(s)$, is obtained in Section 3. For class 1 customers, there are no such interruptions unless the server can break down from time to time, an extension which we discuss in Appendix B. Letting $\overline{F}_k(y) = 1 - F_k(y)$ where $F_k(y)$ is the distribution function of D_k, the first moment of D_k will be denoted by $E[D_k] = \int_0^\infty \overline{F}_k(y)dy$.

Due to these interruptions, instead of the actual service time, we need to consider the process completion time (PCT) r.v. [14], which is the total time a customer spends on the server; this includes the actual service time plus any possible interruption periods it may experience. In our problem, C_k (with a density function of $c_k(z)$) represents the PCT r.v. for a type k customer, and it is the elapsed time between the instant a type k customer's service begins and the instant the same customer departs from the system. If interruptions

occur, once the subsequent interruption period is over, the interrupted customer resumes its service from the moment of interruption. The LT of C_k is found in the literature (e.g., [16], p. 109)

$$\widetilde{c}_k(s) = \widetilde{b}_k(s + \alpha_k - \alpha_k \tilde{f}_k(s)), \tag{1}$$

where, as noted before, $\tilde{f}_k(s)$, the LT of D_k, is found using a recursive algorithm developed in Section 3. In the rest of the paper, we refer to the PCT for class k simply as the PCT.

We employ the following stochastic process to characterize the state of the system at time t: $R_k(t)$ equals 0 if the server is available, and 1 if it is unavailable/interrupted for class k; $W_k(t) \in \{0, 1, ..., N_k\}$ is the number of type k customers out of the queueing system. The elapsed time since the server became unavailable for class k is another stochastic process, but we do not need this information in our derivations. We do not use the stochastic process that gives the number of type k customers in the queuing system at time t, which is $N_k - W_k(t)$, because it is easier to express the state dependent arrival rates via $W_k(t)$ in our derivations. All performance measures investigated in this paper are steady-state performance measures. In the rest of the paper, we denote the mean for any r.v. X by $E[X]$.

3 Busy Period Analysis for Type 1 to Type k Customers

A busy period for type 1 to type k customers starts with either one of the following two events: *Event A:* A probable interruption for class k initiates the busy period if a customer of type 1 to $k-1$ arrives when there are no customers of type 1 to k in the system. *Event B:* The arrival of a type k customer initiates the busy period when there are no customers of type 1 to k in the system. Thus, each busy period starts with an "initial delay" either in the form of a probable interruption for type k customers (D_k) in case of *Event A*, or a PCT (C_k) in case of *Event B*. Due to the preemptive-priority policy, presence of customers of types $k+1$ to m in the system during the busy period is irrelevant from the point of view

of class k (or higher priority) customers. If there are no type k customers waiting for service at the end of an initial delay, the busy period of type 1 to type k customers ends; otherwise, it continues until the server clears all type k (or higher priority) customers from the system. In the remainder of this section and the related proofs given in Appendix A, we refer to the busy period for type 1 to type k customers simply as the busy period.

Let $p_{N_k}^{D_k}(n)$ ($p_{N_k-1}^{C_k}(n)$) be the probability of having $0 \leq n \leq N_k$ ($0 \leq n \leq N_k - 1$) type k customers present in the $M/G/1//N$ system at the end of an interruption (PCT) initiating a busy period. Unlike the systems with constant customer arrival rates, in this system, state dependent arrival rates must be taken into account.

Before presenting the following Theorem, we define $P_{N_k}^{D_k}(n|d)$ ($P_{N_k-1}^{C_k}(n|c)$) as the probability of having n type k customers at the end of the interruption (PCT) initiating a busy period given that $D_k = d$ ($C_k = c$). Further, $\widetilde{P}_{N_k}^{D_k}(n,s) = \int_0^\infty P_{N_k}^{D_k}(n|y)e^{-sy}f_k(y)dy$ ($\widetilde{P}_{N_k-1}^{C_k}(n,s) = \int_0^\infty P_{N_k-1}^{C_k}(n|z)e^{-sz}c_k(z)dz$).

Theorem 1 *The LT $\widetilde{P}_{N_k}^{D_k}(n,s)$ is given by*

$$\widetilde{P}_{N_k}^{D_k}(0,s) = \widetilde{f}_k(s + N_k\lambda_k), \tag{2}$$

$$\widetilde{P}_{N_k}^{D_k}(n,s) = \sum_{i=N_k-n}^{N_k} (-1)^{i-(N_k-n+1)} \binom{N_k}{i} \binom{i}{N_k-n} (\widetilde{f}_k(s) - \widetilde{f}_k(s+i\lambda_k)),\ 0 < n < N_k, \tag{3}$$

$$\widetilde{P}_{N_k}^{D_k}(N,s) = \sum_{i=1}^{N_k} (-1)^{i-1} \binom{N_k}{i} (\widetilde{f}_k(s) - \widetilde{f}_k(s+i\lambda_k)). \tag{4}$$

Note that Theorem 1 can be adjusted to obtain $\widetilde{P}_{N_k-1}^{C_k}(n,s)$ (see the proof of Corollary 2). The following Corollary directly follows from Theorem 1 since $P_{N_k}^{D_k}(n) = \widetilde{P}_{N_k}^{D_k}(n,0)$.

Corollary 1 *The steady-state probability of having n type k customers in the $M/G/1//N$*

system at the end of the interruption initiating a busy period is given by

$$\begin{aligned}
P_{N_k}^{D_k}(0) &= \widetilde{f}_k(N_k\lambda_k), \\
P_{N_k}^{D_k}(n) &= \sum_{i=N_k-n}^{N_k} (-1)^{i-(N_k-n+1)} \binom{N_k}{i} \binom{i}{N_k-n} (1-\widetilde{f}_k(i\lambda_k)),\ 0<n<N_k, \\
P_{N_k}^{D_k}(N) &= \sum_{i=1}^{N_k} (-1)^{i-1} \binom{N_k}{i} (1-\widetilde{f}_k(i\lambda_k)).
\end{aligned}$$

Similarly,

Corollary 2 *The steady-state probability of having n type k customers in the $M/G/1//N$ system at the end of the PCT initiating a busy period is given by*

$$P_{N_k-1}^{C_k}(0) = \widetilde{c}_k((N_k-1)\lambda_k), \tag{5}$$

$$\begin{aligned}
P_{N_k-1}^{C_k}(n) &= \sum_{i=N_k-1-n}^{N_k-1} (-1)^{i-(N_k-n)} \binom{N_k-1}{i} \binom{i}{N_k-1-n} (1-\widetilde{c}_k(i\lambda_k)), \\
&\qquad 0<n<N_k-1,
\end{aligned} \tag{6}$$

$$P_{N_k-1}^{C_k}(N_k-1) = \sum_{i=1}^{N_k-1} (-1)^{i-1} \binom{N_k-1}{i} (1-\widetilde{c}_k(i\lambda_k)). \tag{7}$$

In the remainder of this section, we employ "auxiliary" $M/G/1//N$ systems serving j type k customers, which we call the *auxiliary system j*, $j = 1, \ldots, N_k$. The $M/G/1//N$ system studied in this paper is referred to as the "original system". An *auxiliary system j* has the same underlying stochastic processes and serves the same finite populations as those of the original system except that the finite population k it serves consists of j (instead of N_k) customers. Accordingly, the original system is nothing but the *auxiliary system N_k*.

If there are $n > 0$ type k customers present in the original system at the end of an initial delay ($1 \le n \le N_k$ if the initial delay is an interruption, and $1 \le n \le N_k - 1$ if it is a PCT), in addition to the initial delay, the busy period for type k customers consists of n sub-cycles. Each sub-cycle starting with i type k customers in the original system ($1 \le i \le n$) is the time it takes until $i-1$ type k customers remain in the original system and is identical in distribution to the busy period in the *auxiliary system $N_k - i + 1$* initiated by a PCT (see [22] for a similar approach analyzing the single-class $M/G/1//N$ queue).

Let T_j be the length of the busy period (of type 1 to type k customers) in the *auxiliary system j*. Furthermore, in the *auxiliary system j*, we denote the length of the busy periods initiated by an interruption and a PCT by $T_j^{D_k}$ and $T_j^{C_k}$, and denote their LT's by $\widetilde{h}_j^{D_k}(s)$ and $\widetilde{h}_j^{C_k}(s)$, respectively. Recalling that the original $M/G/1//N$ system we analyze in this paper is the *auxiliary system* N_k, we have

$$T_{N_k}^{D_k} = \begin{cases} D_k, & \text{if there are no type } k \text{ customers at the end of } D_k, \\ D_k + \sum_{j=N_k-n+1}^{N_k} T_j^{C_k}, & \text{if } 0 < n \leq N_k \text{ type } k \text{ customers at at the end of } D_k, \end{cases}$$

$$T_{N_k}^{C_k} = \begin{cases} C_k, & \text{if there are no type } k \text{ customers at the end of } C_k, \\ C_k + \sum_{j=N_k-n+1}^{N_k} T_j^{C_k}, & \text{if } 0 < n \leq N_k - 1 \text{ type } k \text{ customers at the end of } C_k, \end{cases}$$

from which their LT's can be obtained using Theorem 1, respectively, as

$$\begin{aligned} \widetilde{h}_{N_k}^{D_k}(s) &= \widetilde{f}_k(s + N_k\lambda_k) + \sum_{n=1}^{N_k} \widetilde{P}_{N_k}^{D_k}(n,s) \prod_{j=N_k-n+1}^{N_k} \widetilde{h}_j^{C_k}(s), \\ \widetilde{h}_{N_k}^{C_k}(s) &= \widetilde{c}_k(s + (N_k-1)\lambda_k) + \sum_{n=1}^{N_k-1} \widetilde{P}_{N_k-1}^{C_k}(n,s) \prod_{j=N_k-n+1}^{N_k} \widetilde{h}_j^{C_k}(s). \end{aligned} \tag{8}$$

Solving the equation above for $\widetilde{h}_{N_k}^{C_k}(s)$ we get

$$\widetilde{h}_{N_k}^{C_k}(s) = \frac{\widetilde{c}_k(s + (N_k-1)\lambda_k)}{1 - \sum_{n=1}^{N_k-1} \widetilde{P}_{N_k-1}^{C_k}(n,s) \prod_{j=N_k-n+1}^{N_k-1} \widetilde{h}_j^{C_k}(s)}.$$

Since a busy period starts either with an interruption or a type k customer arrival when there are no type 1 to k customers in the system, the LT of the length of the busy period r.v. T_{N_k} for a type 1 to type k customer, $k = 1, \ldots, m$, in the original $M/G/1//N$ system is

$$\widetilde{h}_{N_k}(s) = \frac{\alpha_k}{\alpha_k + N_k\lambda_k} \widetilde{h}_{N_k}^{D_k}(s) + \frac{N_k\lambda_k}{\alpha_k + N_k\lambda_k} \widetilde{h}_{N_k}^{C_k}(s). \tag{9}$$

Then, the mean length of the busy period for type 1 to type k customers in the original system is

$$E[T_{N_k}] = \frac{\alpha_k}{\alpha_k + N_k\lambda_k} E[T_{N_k}^{D_k}] + \frac{N_k\lambda_k}{\alpha_k + N_k\lambda_k} E[T_{N_k}^{C_k}],$$

where

$$E[T_{N_k}^{D_k}] = -\frac{d\widetilde{h}_{N_k}^{D_k}(s)}{ds}|_{s=0} = E[D_k] + \sum_{n=1}^{N_k} E[T_n^{C_k}] \sum_{j=N_k-n+1}^{N_k} P_{N_k}^{D_k}(j),$$

$$E[T_{N_k}^{C_k}] = -\frac{d\widetilde{h}_{N_k}^{C_k}(s)}{ds}|_{s=0} = \frac{E[C_k] + \sum_{n=2}^{N_k-1} E[T_n^{C_k}] \sum_{j=N_k-n+1}^{N_k-1} P_{N_k-1}^{C_k}(j)}{P_{N_k-1}^{C_k}(0)}.$$

We conclude this section by observing that in the original system the interruption period for type $k > 1$ customers is the busy period of type 1 to type $k-1$ customers; in other words, $\widetilde{f}_k(s) = \widetilde{h}_{N_{k-1}}(s)$. Considering this, we present our recursive algorithm as follows:

Algorithm 1 *This algorithm explains how $\widetilde{f}_k(s)$ is obtained, $k = 1, \ldots, m$:*

Step 0. *For class 1 customers, since there is no interruption, $\widetilde{f}_1(s)$ is 1 (if the server experiences failures, $\widetilde{f}_1(s)$ is the LT of the repair time, see the extension in Appendix B).*

Step 1. *Use $\widetilde{f}_1(s)$ in Eqs.* (8) *and* (9) *to obtain $\widetilde{h}_{N_1}(s)$, which is the LT of the busy period for type 1 customers. To do this,*

- *Start by setting $\widetilde{h}_1^{C_1}(s) = \widetilde{b}_1(s)$, the LT of the busy period in the auxiliary system 1, i.e., $M/G/1//N$ queue with a single type 1 customer.*
- *Obtain $\widetilde{h}_j^{C_1}(s)$ in Eq.* (8) *recursively where $\widetilde{P}_{N_k-1}^{C_k}(n,s)$ is obtained from Theorem 1 (by making appropriate adjustments as in the proof of Corollary 2). When $j = N_1$, we have $\widetilde{h}_{N_1}^{C_1}(s)$.*
- *When the server does not experience failures, $\widetilde{h}_{N_1}(s) = \widetilde{h}_{N_1}^{C_1}(s)$. Otherwise, use Eq.* (9) *to obtain $\widetilde{h}_{N_1}(s)$. Set $\widetilde{f}_2(s) = \widetilde{h}_{N_1}(s)$, which is the LT of D_2, the interruption time for class 2 customers.*

Step k.*For classes $k > 1$, having $\widetilde{h}_{N_{k-1}}(s)$ from the earlier iteration, substitute $\widetilde{f}_k(s) = \widetilde{h}_{N_{k-1}}(s)$ in Eqs.*(8) *and* (9) *to obtain $\widetilde{h}_{N_k}(s)$.*

Note that the times between two busy periods of type 1 to type k customers follow an exponential distribution with rate $\alpha_k + N_k\lambda_k$. By invoking the renewal theorem, the fraction

of time there are no type 1 to k customers in the original system is $(1+E[T_{N_k}](\alpha_k+N_k\lambda_k))^{-1}$, and the fraction of time there are no type 1 to $k-1$ customers in the original system is $(1+\alpha_k E[D_k])^{-1}$. Thus, the fraction of time the server is in-service for type k customers is $(1+\alpha_k E[D_k])^{-1} - (1+E[T_{k,N_k}](\alpha_k+N_k\lambda_k))^{-1}$.

4 System Size Distribution for Type k Customers

In this section, we obtain the steady-state probabilities of having i type k customers out of the system at departure/arrival epochs in Section 4.1; we then provide the system size distribution of type k customers at an arbitrary instant in Section 4.2.

4.1 System Size Distribution at Arrival/Departure Epochs

In this section, in order to avoid unnecessary repetitions, we refer to "type k customer/arrival/ departure" simply as the "customer/arrival/departure" since other classes are not part of the discussion. Occasionally, we specifically use "type k customer/arrival/ departure" when we believe that the emphasis makes the explanation clearer. As in the previous section, the PCT stands for the PCT for type k customers.

We start our analysis by studying the embedded Markov chain of the number of type k customers left in the system after a type k customer departs. Let $p_{i,j}^k$ be the transition probability that the next departure leaves j customers in the system, given that the last departure left i customers. If the last departure left i customers, $0 < i < N_k$, in the system, the steady-state probability of the next departure leaving j customers behind ($j = i-1, \ldots, N_k-1$) is the probability of having $j-i+1$ arrivals during the PCT. This probability is the same as the steady-state probability of having $j-i+1$ customers at the end of the PCT that initiates a busy period in the *auxiliary system* $N_k - i + 1$ as introduced in Section 3, and can be obtained by invoking Corollary 2 in this system. Any other transition from i, $0 < i < N_k$, is not possible. After a type k departure leaves the original system empty of

type k customers, the next type k arrival can find the server unavailable/interrupted (serving a higher priority customer), or available (if there are lower priority customers being served, their services are preempted). If the server is found to be interrupted, in steady-state, this arrival waits for the residual interruption period before its service starts. We denote this r.v. by $D_{k,R}$. Following [11], the LT of $D_{k,R}$ can be found as

$$\widetilde{f}_{k,R}(s) = \frac{N_k\lambda_k(N_k\lambda_k - s) + N_k\lambda_k\alpha_k(\widetilde{f}_k(s) - \widetilde{f}_k(N_k\lambda_k))}{(N_k\lambda_k + \alpha_k - \alpha_k\widetilde{f}_k(N_k\lambda_k))(N_k\lambda_k - s)},$$

with

$$\widetilde{f}_{k,R}(N_k\lambda_k) = \lim_{s\to N_k\lambda_k}\widetilde{f}_{k,R}(s) = \frac{N_k\lambda_k\left(1 - \alpha_k\widetilde{f}'_k(N_k\lambda_k)\right)}{N_k\lambda_k + \alpha_k - \alpha_k\widetilde{f}_k(N_k\lambda_k)},$$

where $\widetilde{f}'_k(s)$ is the derivative of $\widetilde{f}_k(s)$ with respect to s. Only then does the PCT of the customer arriving during an interruption period start. In order for such a customer to leave j customers behind ($j = 0, 1, \ldots, N_k - 1$), there should be j arrivals during the interval $L_k = D_{k,R} + C_k$, with an LT of $\widetilde{l}_k(s) = \widetilde{f}_{k,R}(s)\widetilde{c}_k(s)$, and a mean of

$$E[L_k] = -\frac{d\widetilde{l}_k(s)}{ds}|_{s=0} = E[D_{k,R}] + E[C_k].$$

Using Corollary 2 by substituting $\widetilde{l}_k(s)$ for $\widetilde{c}_k(s)$, $P^{L_k}_{N_k-1}(j) = p^k_{0,j}$ ($j = 0, 1, \ldots, N_k - 1$) can be obtained. In summary, we have

$$p^k_{i,j} = \begin{cases} P^{L_k}_{N_k-1}(j), & i = 0,\ 0 \le j \le N_k - 1, \\ P^{C_k}_{N_k-i}(j - i + 1), & 1 \le i < N_k,\ i - 1 \le j \le N_k - 1, \\ 0, & \text{otherwise.} \end{cases}$$

Now that we have $p^k_{i,j}$, we can construct the $N_k \times N_k$ transition probability matrix $\mathbf{P}_k$. From $\mathbf{\Pi}_k = \mathbf{\Pi}_k\mathbf{P}_k$ and $\sum_{i=1}^{N_k}\pi_{k,i} = 1$, we can solve for the $1 \times N_k$ vector $\mathbf{\Pi_k}$= $[\pi_{k,N_k}, \pi_{k,N_k-1}, \ldots, \pi_{k,1}]$. Here, $\pi_{k,i}$ is the steady-state probability of having i type k customers (including the departing customer) out of the queueing system at departure instants (or equivalently having $N_k - i$ type k customers left behind in the queueing system). Since this is an ergodic Markov chain, $\pi_{k,i}$ is also the steady-state probability that an arrival finds $N_k - i$ type k customers in the system.

4.2 System Size Distribution at an Arbitrary Instant

In this section, we obtain $\overline{P}_{k,i}$, the steady-state probability of having i type k customers out of the system.

Lemma 1 *With $E[T_{N_k}]$ as the mean length of the busy period of type 1 to type k customers,*

$$\overline{P}_{k,N_k} = \frac{N_k\lambda_k + \alpha_k - \alpha_k\widetilde{f}_k(N_k\lambda_k)}{N_k\lambda_k(1 + E[T_{N_k}](\alpha_k + N_k\lambda_k))}.$$

To obtain the entire distribution, we introduce the "augmented PCT" (APCT) r.v. for type k customers denoted by $\widehat{C}_k$, which is the PCT for all type k customers (i.e. $\widehat{C}_k = C_k$) except for those arriving as the first type k customers during an interruption period that initiates a busy period. In the latter case, the APCT is the residual interruption period such customers wait plus their PCT, that is $\widehat{C}_k = L_k$. Then, the residual APCT r.v. $\widehat{C}_{k,R}$ with $\widehat{c}_{k,R}(x)$ as its density function is the time left until the departure of the first type k customer (that may be waiting for the interruption period that initiates a busy period, or is in service, or is preempted) in the system. It is known that $P(\widehat{C}_{k,R} = 0) = \overline{P}_{k,N_k}$, i.e., the probability that there are no type k customers in the queueing system, but we define $\widehat{c}_{k,R}(0) = \lim_{x\to 0}\widehat{c}_{k,R}(x)$.

Let $\widehat{C}_{k,R}(t)$ denote the residual APCT at time t and

$$P_{k,i}(t,x)dx = P\{W_k(t) = i, x < \widehat{C}_{k,R}(t) < x + dx\},\ 0 \le i \le N_k - 1,$$

denote the joint probability distribution of having i type k customers out of the queueing system at time t ($W_k(t) = i$), and the residual APCT of the customer (preempted or currently receiving service) being in the interval $[x, x+dx]$. Observe that from t to $t+\Delta t$, the residual APCT will decrease by Δt. Assuming that the probability of having more than one arrival is $o(\Delta t)$ and $P_{k,-1}(t,x)$ and its limiting probability are 0,

$$\begin{aligned}
P_{k,N_k-1}(t+\Delta t, x) &= (1-(N_k-1)\lambda_k\Delta t)P_{k,N_k-1}(t, x+\Delta t) + N_k\lambda_k\Delta t\overline{P}_{k,N_k}(t)l_k(x) \\
&\quad + P_{k,N_k-2}(t,0)c_k(x)\Delta t + o(\Delta t), \\
P_{k,i}(t+\Delta t, x) &= (1-i\lambda_k\Delta t)P_{k,i}(t, x+\Delta t) + (i+1)\lambda_k\Delta t P_{k,i+1}(t, x+\Delta t) \\
&\quad + P_{k,i-1}(t,0)c_k(x)\Delta t + o(\Delta t),\ 0 \le i \le N_k - 2,
\end{aligned}$$

where $\overline{P}_{k,N_k}(t)$ is the probability of having N_k type k customers out of the system at time t. Here $l_k(x)$, and $c_k(x)$ are the density functions of the r.v.s L_k and C_k, respectively, and $c_k(x)\Delta t = P(x \le C_k \le x + \Delta t)$. Re-arranging the equations given above, we obtain

$$\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial x}\right) P_{k,N_k-1}(t,x) = -(N_k-1)\lambda_k P_{k,N_k-1}(t,x) + N_k\lambda_k \overline{P}_{k,N_k}(t) l(x) + P_{k,N_k-2}(t,0)c_k(x),$$

$$\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial x}\right) P_{k,i}(t,x) = -i\lambda_k P_{k,i}(t,x) + (i+1)\lambda_k P_{k,i+1}(t,x) + P_{k,i-1}(t,0)c_k(x), \quad 0 \le i \le N_k - 2.$$

Letting $P_{k,i}(x) = \lim_{t\to\infty} P_{k,i}(t,x)$, if we take the limit of the equations given above as $t \to \infty$,

$$\frac{d}{dx}P_{k,N_k-1}(x) = (N_k-1)\lambda P_{k,N_k-1}(x) - N_k\lambda_k \overline{P}_{k,N_k} l(x) - P_{k,N_k-2}(0)c_k(x), \tag{10}$$

$$\frac{d}{dx}P_{k,i}(x) = i\lambda P_{k,i}(x) - (i+1)\lambda P_{k,i+1}(x) - P_{k,i-1}(0)c_k(x), \quad 0 \le i \le N_k - 2. \tag{11}$$

Observe that $P_{k,i}(x)$ is the density function of the residual APCT and i type k customers are out of the queueing system. When $i = 0$, integrating both sides of Eq. (11) gives

$$\begin{aligned} P_{k,0}(\infty) - P_{k,0}(0) &= -\lambda_k \overline{P}_{k,1} \\ P_{k,0}(0) &= \lambda_k \overline{P}_{k,1}. \end{aligned}$$

Recursively, we can show that

$$P_{k,i}(0) = (i+1)\lambda_k \overline{P}_{k,i+1}, \ 0 \le i \le N_k - 1. \tag{12}$$

Note that $P_{k,i}(0)$ is the probability that a type k customer is about to leave the server and there are i type k customers out of the queueing system. Then, using Bayes' theorem

$$\begin{aligned} \pi_{k,i+1} &\equiv P\{i \text{ type } k \text{ customers out of the system}|\text{a type } k \text{ departure is about to occur}\} \\ &= \frac{P_{k,i}(0)}{\widehat{c}_{k,R}(0)} = \frac{P_{k,i}(0)}{\sum_{i=0}^{N_k-1} P_{k,i}(0)}, \ 0 \le i \le N_k - 1, \\ \pi_{k,i} &= \frac{i\lambda_k \overline{P}_{k,i}}{\sum_{i=1}^{N_k} i\lambda_k \overline{P}_{k,i}}, \ 1 \le i \le N_k, \\ \overline{P}_{k,i} &= \frac{N_k \overline{P}_{k,N_k}}{i\pi_{k,N_k}} \pi_{k,i}, \ 1 \le i \le N_k. \end{aligned} \tag{13}$$

Using Eq. (13) together with Lemma 1, we derive the solution for $\overline{P}_{k,i}$, which is also the steady-state probability of having $N_k - i$ customers in the system. Eq. (13) also helps us obtain $\widehat{c}_{k,R}(0) = N_k\lambda_k\overline{P}_{N_k}/\pi_{k,N_k}$.

The following theorem provides an alternative solution. Before presenting it, we introduce the conditional residual APCT, given that there are i type k customers out of the system. By definition, its density function is (the LT $\widetilde{c}_{k,R|i}(s)$ is obtained in Section 4.3)

$$\widehat{c}_{k,R|i}(x) = \frac{P_{k,i}(x)}{\overline{P}_{k,i}}. \tag{14}$$

Theorem 2 *There is a recursive relationship between the steady-state probabilities $\overline{P}_{k,i}$ so that*

$$\overline{P}_{k,N_k-1} = \frac{N_k}{(N_k-1)}\frac{1-\widetilde{l}_k((N_k-1)\lambda_k)}{\widetilde{c}_k((N_k-1)\lambda_k)}\overline{P}_{k,N_k}, \tag{15}$$

$$\overline{P}_{k,i} = \frac{(i+1)\overline{P}_{k,i+1}}{i\widetilde{c}_k(i\lambda_k)}(1-\widetilde{c}_{k,R|i+1}(i\lambda_k)),\ \ 0 < i \le N_k - 2. \tag{16}$$

4.3 The System Time Distribution for Type k Customers

In this section, we obtain the LT $\widetilde{c}_{k,R|i}(s)$ of the conditional residual APCT of type k customers given that there are i type k customers out of the system.

Theorem 3 *There is a recursive relationship for $\widehat{c}_{k,R|i}(x)$ such that*

$$\begin{aligned}\widehat{c}_{k,R|N_k-1}(x) = {} & \frac{(N_k-1)\lambda_k e^{(N_k-1)\lambda_k x}}{1-\widetilde{l}_k((N_k-1)\lambda_k)}\{\widetilde{c}_k((N_k-1)\lambda_k)\int_x^\infty e^{-(N_k-1)\lambda_k u}l_k(u)du \\ & +(1-\widetilde{l}_k((N_k-1)\lambda_k))\int_x^\infty e^{-(N_k-1)\lambda_k u}c_k(u)du\},\end{aligned} \tag{17}$$

$$\widehat{c}_{k,R|i}(x) = i\lambda e^{i\lambda_k x}\int_x^{+\infty} e^{-i\lambda_k u}\left(\widetilde{c}_k(i\lambda_k)\frac{\widehat{c}_{k,R|i+1}(u)}{1-\widetilde{c}_{k,R|i+1}(i\lambda_k)} + c_k(u)\right)du,\ 0 < i \le N_k - 2. \tag{18}$$

And,

Theorem 4 *There is a recursive relationship for* $\widetilde{c}_{k,R|i}(s)$ *such that*

$$\widetilde{c}_{k,R|N_k-1}(s) = \frac{(N_k-1)\lambda_k}{s-(N_k-1)\lambda_k}\frac{\widetilde{c}_k((N_k-1)\lambda_k)\left(1-\widetilde{l}_k(s)\right)-\widetilde{c}_k(s)\left((1-\widetilde{l}_k((N_k-1)\lambda_k)\right)}{1-\widetilde{l}_k((N_k-1)\lambda_k))}, \tag{19}$$

$$\widetilde{c}_{k,R|i}(s) = \frac{i\lambda_k}{s-i\lambda_k}\left(\widetilde{c}_k(i\lambda_k)\frac{1-\widetilde{c}_{k,R|i+1}(s)}{1-\widetilde{c}_{k,R|i+1}(i\lambda_k)}-\widetilde{c}_k(s)\right),\ 0<i\le N_k-2, \tag{20}$$

$$\widetilde{c}_{k,R|0}(s) = \frac{\overline{P}_{k,1}}{\overline{P}_{k,0}}\frac{\lambda_k\left(1-\widetilde{c}_{k,R|1}(s)\right)}{s}. \tag{21}$$

The following Theorem is presented without a proof since its proof is, in principle, the same as that of Theorem 2.2.2 in [17] which exploits Theorem 1 in [10].

Theorem 5 *The conditional residual APCT of type* k *customers at an arrival epoch given that there are* i *type* k *customers out of the system has* $\widehat{c}_{k,R|i}(x)$ *as its density function.*

Recall from Section 4.1 that in steady-state a type k arrival finds N_k-i type k customers in the system with probability $\pi_{k,i}$. Using Theorem 5, the system time of such a customer is the residual APCT of the type k customer first in line plus the sum of N_k-i PCT's of the type k customers waiting behind it in the queue and the new arrival; this has the LT of

$$\widetilde{w}_{k,i}(s)=\widetilde{c}_{k,R|i}(s)\widetilde{c}_k^{N_k-i}(s),\ \ 1\le i\le N_k-1.$$

With probability π_{k,N_k}, the type k customer finds no type k customers in the system and its system time is L_k. By the law of total probability, the LT of the system time of a type k customer is given by

$$\widetilde{w}_k(s)=\sum_{i=1}^{N_k-1}\pi_{k,i}\widetilde{w}_{k,i}(s)+\pi_{k,N_k}\widetilde{l}_k(s).$$

5 Conclusions

In this paper, we develop a method to obtain the exact steady-state system size distribution and conduct the busy period analysis of the $M/G/1//N$ queue where multiple classes of customers are served according to the preemptive-resume priority policy. Eventually, we extend

the model to capture an unreliable server subject to operation-independent interruptions. We demonstrate how setup times that may be required before resuming interrupted service or picking up a new customer can be included in the PCT analysis. We assume general OFF period, service, and setup time distributions. Including non-exponential distributions to model times between customer arrivals and/or times between server interruptions remains challenging and is an open research question. In addition to the steady-state system size distribution obtained, we also provide the LT's for the PCT and system time for each class, and that of the busy period r.v. for class 1 to class k from which one can obtain the higher moments of the r.v.s of interest. This may help see the impact of the characteristics of the underlying r.v.s on system performance measures more clearly.

Acknowledgements

This work was supported in part by Natural Sciences and Engineering Research Council (NSERC) of Canada. The authors thank the two anonymous referees and the editors for their invaluable suggestions to improve the manuscript. The authors thank Dr. Elizabeth Thompson, PhD, who has proofread the manuscript.

References

[1] Abramov, V. M. 2000. "A large closed queueing network with autonomous service and bottleneck," *Queueing Systems*, Vol. 35, No. 1–4, 23–54.

[2] Abramov, V. M. 2001. "Some Results for Large Closed Queueing Networks with and without Bottleneck: Up- and Down-Crossings Approach," *Queueing Systems*, Vol. 38, No. 2, 149–184.

[3] Abramov, V. M. 2004. "A large closed queueing network containing two types of node and multiple customer classes: One bottleneck station," *Queueing Systems*, Vol. 48,

No. 1–2, 45–73.

[4] Almasi, B. and J. Sztrik. 2004. "Reliability investigations of heterogeneous terminal systems using MOSEL", *Journal of Mathematical Sciences*, Vol. 123, No. 1, 3795–3801.

[5] Altıok, T. 1997. *Performance Analysis of Manufacturing Systems*, Springer-Verlag, New York, NY.

[6] Atencia, I., G. Bouza, and P. Moreno. 2008. "An $M^{[X]}/G/1$ retrial queue with server breakdowns and constant rate of repeated attempts," *Annals of Operations Research*, Vol. 157, No. 1, 225–243.

[7] Avi-Itzhak, B. and P. Naor. 1963. "Some queueing problems with the service station subject to breakdown", *Operations Research*, Vol. 11, No. 3, 303–320.

[8] Balcıoğlu, B., D. L. Jagerman, and T. Altıok. 2007. "Approximate mean waiting time in a $GI/D/1$ queue with autocorrelated times to failures", *IIE Transactions*, Vol. 39, 985-996.

[9] Chakravarthy, S. R. and A. Agarwal. 2003. "Analysis of a machine repair problem with an unreliable server and phase type repairs and services", *Naval Research Logistics*, Vol. 50, No. 5, 462–480.

[10] van Doorn E. A. and G. J. K. Regterschot. 1988. "Conditional PASTA", *Operations Research Letters*, Vol. 7, No. 5, 229–232.

[11] Fiems, D., T. Maertens, and H. Bruneel. 2008. "Queueing systems with different types of server interruptions", *European Journal of Operational Research*, Vol. 188, No. 3, 838–845.

[12] Federgruen, A. and L. Green. 1986. "Queueing systems with service interruptions", *Operations Research*, Vol. 34, No. 5, 752–768.

[13] Federgruen, A. and L. Green. 1988. "Queueing systems with service interruptions II", *Naval Research Logistics*, Vol. 35, 345–358.

[14] Gaver, D. P. 1962. "A waiting line with interrupted service, including priorities", *Journal of the Royal Statistical Society*, Vol. 24, No. 1, 73–90.

[15] Haque, L. and M. J. Armstrong. 2007. "A survey of the machine interference problem", *European Journal of Operational Research*, Vol. 179, No. 2, 469-482.

[16] Jaiswal, N. K. 1968. *Priority queues*, Academic Press, New York.

[17] Kerner, Y. 2008. "The conditional distribution of the residual service time in the $M_n/G/1$ queue," *Stochastic Models*, Vol. 24, 364–375.

[18] Kogan, Y. and R. S. Lipster. 1993. "Limit non-stationary behavior of large closed queueing networks with bottlenecks," *Queueing Systems*, Vol. 14, No. 1–2, 33–55.

[19] Mitrany, I. L. and B. Avi-Itzhak. 1968. "A many server queue with service interruptions", *Management Science*, Vol. 25, 849–861. *Operations Research*, Vol. 16, No. 3, 628–638.

[20] Neuts, M. F. and D. M. Lucantoni. 1979. " A Markovian queue with N servers subject to breakdowns and repair", *Management Science*, Vol. 25, No. 9, 849–861.

[21] Sahba, P., B. Balcıoğlu, and D. Banjevic. 2012. "Spare Parts Provisioning for Multiple k-out-of-n : G Systems", *forthcoming in IIE Transactions*, DOI:10.1080/0740817X.2012.695102.

[22] Shanthikumar, J. G. and U. Sumita. 1985. "On the busy-period distributions of $M/G/1/K$ queues with state-dependent arrivals and FCFS/LCFS-P service disciplines", *Journal of Applied Probability*, Vol. 22, No. 4, 912–919.

[23] Stecke, K. E. and J. E. Aronson. 1985. "Review of operator/machine interference models", *International Journal of Production Research*, Vol. 23, No. 1, 129–151.

[24] Sztrik, J. and T. Gál. 1990. "A recursive solution of a queueing model for a multi-terminal system subject to breakdowns", *Performance Evaluation*, Vol. 11, No. 1, 1–7.

[25] Thiruvengadam, K. 1963. "Queueing with breakdown", *Operations Research*, Vol. 11, 62–71.

[26] Veran, M. 1984. "Exact analysis of a priority queue with finite source, In: Modelling and Performance Evaluation Methodology. Eds. F. Baccelli and G. Fayolle. *Proceedings of the International Seminar on Modelling and Performance Evaluation Methodology*, Paris, France, January 24-26, 1983, 371-390.

[27] Wang, K.-H. 1990. "Profit analysis of the machine-repair problem with a single service station subject to breakdowns", *Journal of the Operational Research Society*, Vol. 41, No. 12, 1153–1160.

[28] Wang, J. and K. W. Ross. 1994. "Asymptotic analysis for closed multiclass queueing networks in critical usage," *Queueing Systems*, Vol. 16, No. 1–2, 167–161.

[29] Wang, K.-H and Kuo, M.-Y. 1997. "Profit analysis of the $M/E_k/1$ machine repair problem with a non-reliable service station", *Computers and Industrial Engineering*, Vol. 32, No. 3, 587–594.

[30] Wang, J., J. Cao, and Q. Li. 2001. "Reliability analysis of the retrial queue with server breakdowns and repairs", *Queueing Systems*, Vol. 38, No. 4, 363–380.

[31] White, H. and L. Christie. 1958. "Queueing with preemptive priorities or with breakdown", *Operations Research*, Vol. 6, No. 1, 79–95.

Appendix A Proofs

Proof. Theorem 1. To prove Theorem 1, we need the following Lemma.

Lemma A.1 *During the interruption period initiating a busy period, the time-to-arrival r.v. $T_{k,N_k,n}$ of the nth type k customer has the following cumulative distribution function:*

$$H_{k,N_k,n}(t) = (N_k - n + 1) \sum_{i=N_k-n+1}^{N_k} (-1)^{i-(N_k-n+1)} \binom{N_k}{i} \binom{i}{N_k - n + 1} \frac{(1 - e^{-i\lambda_k t})}{i}. \quad \text{(A.1)}$$

Proof. Lemma A.1. Note that if an interruption initiates a busy period, at the beginning of the interruption, N_k type k customers are not yet in the queueing system. During the interruption period initiating a busy period, when $W_k(t) = N_k - n$, the time-to-arrival of the next type k customer is exponentially distributed with rate of $(N_k - n)\lambda_k$, and $T_{k,N_k,n}$ is the sum of n exponentially distributed r.v.s with rates of $N_k\lambda_k, (N_k-1)\lambda_k, \ldots,$ and $(N_k-n+1)\lambda_k$, i.e.,

$$T_{k,N_k,n} = \sum_{i=N_k-n+1}^{N_k} T_{k,i},$$

where $T_{k,i}$ follows an exponential distribution with rate $i\lambda_k$. Let $\widetilde{h}_{k,N_k,n}(s)$ be the LT of $T_{k,N_k,n}$, then

$$\begin{aligned} \widetilde{h}_{k,N_k,n}(s) &= \frac{N_k\lambda_k}{N_k\lambda_k + s} \frac{(N_k-1)\lambda_k}{(N_k-1)\lambda_k + s} \cdots \frac{(N_k-n+1)\lambda_k}{(N_k-n+1)\lambda_k + s}, \\ &= \frac{N_k! {\lambda_k}^n}{(N_k-n)!} \prod_{i=N_k-n+1}^{N_k} \frac{1}{i\lambda_k + s}. \end{aligned} \quad \text{(A.2)}$$

Using

$$\frac{N_k! {\lambda_k}^{n-1}}{(N_k-n+1)!} \prod_{i=N_k-n+1}^{N_k} \frac{1}{i\lambda_k + s} = \sum_{i=N_k-n+1}^{N_k} (-1)^{i-(N_k-n+1)} \binom{N_k}{i} \binom{i}{N_k - n + 1} \frac{1}{i\lambda_k + s},$$

in Eq. (A.2), we arrive at

$$\widetilde{h}_{k,N_k,n}(s) = (N_k - n + 1)\lambda_k \sum_{i=N_k-n+1}^{N_k} (-1)^{i-(N_k-n+1)} \binom{N_k}{i} \binom{i}{N_k - n + 1} \frac{1}{i\lambda_k + s},$$

the inversion of which gives Eq. (A.1). ■

To prove Theorem 1, given that $D_k = d$, and using Lemma A.1, we have

$$P_{N_k}^{D_k}(0|d) = P\{T_{k,1} > d\} = 1 - H_{k,N_k,1}(d) = e^{-N_k\lambda_k d}, \quad \text{(A.3)}$$

and for $0 < n < N_k$

$$\begin{aligned} P_{N_k}^{D_k}(n|d) &= P\{T_{k,n} < d < T_{k,n+1}\} = H_{k,N_k,n}(d) - H_{k,N_k,n+1}(d) \\ &= \sum_{i=N_k-n}^{N_k} (-1)^{i-(N_k-n+1)} \binom{N_k}{i} \binom{i}{N_k-n} (1 - e^{-i\lambda_k d}), \end{aligned} \tag{A.4}$$

and finally,

$$P_{N_k}^{D_k}(N_k|d) = P\{T_{k,N_k} < d\} = H_{k,N_k,N_k}(d) = \sum_{i=1}^{N_k} (-1)^{i-1} \binom{N_k}{i} (1 - e^{-i\lambda_k d}). \tag{A.5}$$

Taking the LT of Eqs.(A.3)-(A.5) yields Eqs. (2)-(4), respectively. ■

Proof. Corollary 2. The fundamental difference between an interruption period initiating a busy period and a PCT initiating a busy period are the following. The PCT has a different distribution from that of the interruption time, and at the beginning of the busy period initiated by PCT, $N_k - 1$ type k customers are not yet in the queueing system. Therefore, Lemma A.1 and Theorem 1 can be adjusted reflecting these differences and Eqs. (5)-(7) can be obtained. ■

Proof. Lemma 1. The probability of the system being empty of type k customers is

$$\overline{P}_{k,N_k} = \lim_{t\to\infty} P\left\{(W_k(t) = N_k) \cap R_k(t) = 0\right\} + \lim_{t\to\infty} P\left\{(W_k(t) = N_k) \cap R_k(t) = 1\right\}. \tag{A.6}$$

The probability of having no type k customers in the system and the server being available for class k (without any higher priority customers in the system, as discussed at the end of Section 3) is

$$\lim_{t\to\infty} P\left\{(W_k(t) = N_k) \cap R_k(t) = 0\right\} = \frac{1}{1 + E[T_{N_k}](\alpha_k + N_k\lambda_k)}.$$

Observe that only during the interruption period which initiates a busy period can the server be unavailable while no type k customer exists in the system; the average time the system remains empty of type k customers during such an interruption period is given by

$$\int_0^\infty \left(\int_0^y t N_k \lambda_k e^{-N_k\lambda_k t} dt + y \int_y^\infty N_k \lambda_k e^{-N_k\lambda_k t} dt \right) f_k(y) dy = \frac{1 - \widetilde{f}_k(N_k\lambda_k)}{N_k\lambda_k}.$$

For type k customers, the fraction of time the system is in a busy period initiated by an interruption is

$$\frac{\alpha_k E[T_{N_k}^{D_k}]}{1+E[T_{N_k}](\alpha_k+N_k\lambda_k)},$$

thus, the fraction of time the server is unavailable for and empty of type k customers is

$$\lim_{t\to\infty} P\left\{(W_k(t)=N_k)\cap R_k(t)=1\right\} = \frac{\alpha_k \frac{1-\widetilde{f}_k(N_k\lambda_k)}{N_k\lambda_k}}{1+E[T_{N_k}](\alpha_k+N_k\lambda_k)}.$$

The summation of these in Eq. (A.6) gives $\overline{P}_{k,N_k}$ in Lemma 1. ■

Proof. Theorem 2. After substituting $P_{k,N_k-2}(0)=(N_k-1)\lambda_k\overline{P}_{k,N_k-1}$ from Eq. (12) into Eq. (10) and multiplying both sides by $e^{-(N_k-1)\lambda_k x}$, eventually, we have

$$\frac{d}{dx}\left(e^{-(N_k-1)\lambda_k x}P_{k,N_k-1}(x)\right) = -N_k\lambda_k e^{-(N_k-1)\lambda_k x}\overline{P}_{k,N_k}l_k(x) - (N_k-1)\lambda_k e^{-(N_k-1)\lambda_k x}\overline{P}_{k,N_k-1}c_k(x).$$

Integrating both sides gives

$$\begin{aligned} -e^{-(N_k-1)\lambda_k x}P_{k,N_k-1}(x) &= -N_k\lambda_k\overline{P}_{k,N_k}\int_x^\infty e^{-(N_k-1)\lambda_k u}l_k(u)du \\ &- (N_k-1)\lambda_k\overline{P}_{k,N_k-1}\int_x^\infty e^{-(N_k-1)\lambda_k u}c_k(u)du. \end{aligned} \tag{A.7}$$

At $x=0$, Eq. (A.7) is

$$P_{k,N_k-1}(0) = N_k\lambda_k\overline{P}_{k,N_k}\widetilde{l}_k((N_k-1)\lambda_k) + (N_k-1)\lambda_k\overline{P}_{k,N_k-1}\widetilde{c}_k((N_k-1)\lambda_k).$$

The equation above together with Eq. (12) for $P_{k,N_k-1}(0)$ gives Eq. (15).

Similarly, by multiplying both sides of Eq. (11) by $e^{-i\lambda_k x}$, and skipping similar steps as in the first part of the proof, we arrive at

$$P_{k,i}(x) = e^{i\lambda_k x}\left(\int_x^\infty \lambda_k e^{-i\lambda_k u}(i+1)P_{k,i+1}(u)du + i\lambda_k\overline{P}_{k,i}\int_x^\infty e^{-i\lambda_k u}c_k(u)du\right). \tag{A.8}$$

For $x=0$, Eq. (A.8) is

$$P_{k,i}(0) = \int_0^\infty \lambda_k e^{-i\lambda_k u}(i+1)P_{k,i+1}(u)du + i\lambda_k\overline{P}_{k,i}\int_0^\infty e^{-i\lambda_k u}c_k(u)du.$$

Note that by the definition given in Eq. (14), $\widetilde{P}_{k,i+1}(s) = \overline{P}_{k,i+1}\widehat{c}_{k,R|i}(s)$, which together with Eq. (12), leads us to

$$P_{k,i}(0) = (i+1)\lambda_k \overline{P}_{k,i+1} = (i+1)\lambda_k \overline{P}_{k,i+1}\widetilde{c}_{k,R|i+1}(i\lambda_k) + i\lambda_k \overline{P}_{k,i}\widetilde{c}_k(i\lambda_k),$$

from which Eq. (16) follows. ■

Proof. Theorem 3. Eq. (17) follows directly by substituting Eq. (15) in Eq. (A.7). Eq. (18), which is the same as Eq. (2) in [17], is obtained by substituting Eq. (16) in Eq. (A.8). ■

Proof. Theorem 4. After multiplying both sides of Eq. (10) with e^{-sx} and integrating, we have

$$\begin{aligned}
\int_0^\infty e^{-sx} dP_{k,N_k-1}(x) &= (N_k-1)\lambda_k \int_0^\infty e^{-sx} P_{k,N_k-1}(x)dx - N_k\lambda_k \overline{P}_{k,N_k} \int_0^\infty e^{-sx} l_k(x)dx \\
&\quad - P_{k,N_k-2}(0) \int_0^\infty e^{-sx} c_k(x)dx, \\
s\widetilde{P}_{k,N_k-1}(s) - P_{k,N_k-1}(0) &= (N_k-1)\lambda_k \widetilde{P}_{k,N_k-1}(s) - N_k\lambda_k \overline{P}_{k,N_k}\widetilde{l}_k(s) - P_{k,N_k-2}(0)\widetilde{c}_k(s), \\
\widetilde{P}_{k,N_k-1}(s) &= \frac{N_k\lambda_k \overline{P}_{k,N_k}(1-\widetilde{l}_k(s)) - (N_k-1)\lambda_k \overline{P}_{k,N_k-1}\widetilde{c}_k(s)}{s-(N_k-1)\lambda_k}.
\end{aligned}$$

Note that for the last equation above, we used Eq. (12). After multiplying both sides of Eq. (15) by λ_k, we re-arranged it to express $N_k\lambda_k\overline{P}_{k,N_k}$. When this is substituted in the last equation above, we get

$$\widetilde{P}_{k,N_k-1}(s) = \frac{(N_k-1)\lambda_k \overline{P}_{k,N_k-1}\left(\widetilde{c}_k((N_k-1)\lambda_k)(1-\widetilde{l}_k(s)) - \widetilde{c}_k(s)\left(1-\widetilde{l}_k((N_k-1)\lambda_k)\right)\right)}{(1-\widetilde{l}_k((N_k-1)\lambda_k))(s-(N_k-1)\lambda_k)}.$$

Dividing the equation given above by $\overline{P}_{k,N_k-1}$ according to Eq. (14) gives Eq. (19). Similarly, Eq. (20) can be found by starting with Eq. (11) and is the same as Eq. (4) in [17]. When $i=0$, multiplying both sides of Eq. (11) by e^{-sx}, integrating the results, and then using Eq. (14), gives

$$\begin{aligned}
\widetilde{P}_{k,0}(s) &= \frac{\lambda_k(\overline{P}_{k,1} - \widetilde{P}_{k,1}(s))}{s} \\
&= \frac{\lambda_k \overline{P}_{k,1}(1-\widetilde{c}_{k,R|1}(s))}{s}.
\end{aligned}$$

Dividing this equation by $\overline{P}_{k,0}$ according to Eq. (14) gives Eq. (21). ■

Appendix B Incorporating Server Failures and the Process Completion Time Analysis with Setup Times for Type k Customers

One can easily incorporate server failures in the model studied where times to failures are exponentially distributed with a rate of α_1. This is the case in which the server is subject to "operation-independent" failures; this differentiates the problem from those where a server can fail only when it is serving a customer. Thus, the server can fail even when it is idle. When a failure occurs, the server becomes "down" (thus, unavailable), and a repair process starts at once. The length of each server down/repair time is an i.i.d. r.v., denoted by D_1; this follows a general continuous distribution $F_1(y) = \int_0^y f_1(u)du$ with density function $f_1(y)$, and has an LT $\widetilde{f}_1(s)$. Such failures can be easily included in the model by assuming a single highest priority customer with an arrival rate of α_1 and D_1 as its service time. Note that the process that counts the total number of failures forms a renewal process with inter-renewal times $X_1, X_2, \ldots$, where $X_i = D_i + Y_i$, D_i is the ith repair time, and Y_i follows an exponential distribution with rate α_1. Thus, α_1 is the interruption rate, and D_1 the interruption r.v. for class 1 customers. For classes $k > 2$, we adjust interruption rates as $\alpha_k = \alpha_1 + \sum_{n=1}^{k-1} N_n \lambda_n$.

In Algorithm 1, in Step 1, we use $\widetilde{f}_1(s)$ of D_1 in Eqs. (8) and (9). The rest of the algorithm follows in the same way but this time making use of $\alpha_k = \alpha_1 + \sum_{n=1}^{k-1} N_n \lambda_n$.

Next we discuss how we can incorporate setup times in the model. We can consider the possibility that each time the server attempts to serve a type k customer (for the first time or after an interruption), it undergoes a setup/loading time which is denoted by the i.i.d. r.v. U_k with a density function $g_k(y)$ that is independent of both D_k and the (remaining) service time r.v. Interruptions can occur during setup time. At the end of the ensuing interruption period, a new setup time is generated from the same distribution until one is not interrupted. Only then does the server start or resume serving the type k customer. If the server is interrupted during a setup time, the remaining service time of an interrupted

customer does not change. Only the amount of work done after an uninterrupted setup time reduces the remaining service time.

When the server is not down, it is considered to be "up", which means that it is either idle and ready to serve, or is being set up (and the server is considered to be "loading"), or is serving a customer (and the server is "in-service"). Therefore, at any given time, the server is in one of the following four states: idle, in-service, loading, or down.

Let $C_k(U_k, Z_k|y)$ be the r.v. denoting the PCT for a type k customer as a function of the setup time r.v., U_k, the service time r.v., Z_k, (Z_k can also be the remaining service time of an interrupted customer), and the time until the next interruption, y. Then,

$$C_k(U_k, Z_k|y) = \begin{cases} U_k + Z_k, & \text{if } y \geq U_k + Z_k, \\ y + D_k + C_k'(U_k, Z_k - (y - U_k)|y'), & \text{if } U_k \leq y < Z_k + U_k, \\ y + D_k + C_k'(U_k, Z_k|y'), & \text{if } 0 \leq y < U_k, \end{cases}$$

where $C_k'(U_k, Z_k|y')$, given U_k and Z_k, is i.i.d as $C_k(U_k, Z_k|y)$. This equation assures that an arrival seeing an up and idle server also undergoes a loading/setup period. For notational convenience, index k is removed in the following derivations. Given that $U = u$ and $Z = z$, the LT of $C(U = u, Z = z)$, $\widetilde{c}(s|u, z)$ is given by

$$\begin{aligned} \widetilde{c}(s|u,z) &= e^{-s(z+u)}e^{-\alpha(z+u)} + \alpha\widetilde{f}(s)\int_0^z e^{-(s+\alpha)(z+u-\omega)}\widetilde{c}(s|u,\omega)d\omega \\ &\quad + \tilde{f}(s)\widetilde{c}(s|u,z)\int_0^u \alpha e^{-(\alpha+s)y}dy, \end{aligned}$$

which, after being rearranged and by letting $\omega = z + u - y$, becomes

$$\begin{aligned} \widetilde{c}(s|u,z)e^{(s+\alpha)(z+u)} &= 1 + \alpha\widetilde{f}(s)\int_0^z e^{(s+\alpha)\omega}\widetilde{c}(s|u,\omega)d\omega \\ &\quad + \frac{\alpha}{\alpha+s}\left(e^{(s+\alpha)u} - 1\right)e^{(s+\alpha)z}\tilde{f}(s)\widetilde{c}(s|u,z), \end{aligned}$$

$$\left(e^{(s+\alpha)u} + \frac{\alpha}{s+\alpha}(1 - e^{(s+\alpha)u})\widetilde{f}(s)\right)e^{(s+\alpha)z}\widetilde{c}(s|u,z) = 1 + \alpha\widetilde{f}(s)\int_0^z e^{(s+\alpha)\omega}\widetilde{c}(s|u,\omega)d\omega.$$

After taking the derivative of both sides with respect to z,

$$\frac{\partial \ln e^{(s+\alpha)z}\widetilde{c}(s|u,z)}{\partial z} = \frac{\alpha\widetilde{f}(s)}{e^{(s+\alpha)u} + \frac{\alpha}{s+\alpha}(1 - e^{(s+\alpha)u})\widetilde{f}(s)},$$

we obtain the following solution

$$\widetilde{c}(s|u,z) = e^{-\left(s+\alpha-\frac{\alpha\widetilde{f}(s)}{e^{(s+\alpha)u}+\frac{\alpha}{s+\alpha}(1-e^{(s+\alpha)u})\widetilde{f}(s)}\right)z}.$$

If we remove the condition on z by integrating $\widetilde{c}(s|u,z)$ over all possible values of z, we obtain

$$\widetilde{c}(s|u) = \widetilde{b}\left(s+\alpha-\frac{\alpha\widetilde{f}(s)}{e^{(s+\alpha)u}+\frac{\alpha}{s+\alpha}(1-e^{(s+\alpha)u})\widetilde{f}(s)}\right).$$

Similarly, when we remove the condition on u, and reintroduce index k, we obtain the LT of C_k as

$$\widetilde{c}_k(s) = \int_0^\infty \widetilde{b}_k\left(s+\alpha_k-\frac{\alpha_k\widetilde{f}_k(s)}{e^{(s+\alpha_k)u}+\frac{\alpha_k}{s+\alpha_k}(1-e^{(s+\alpha_k)u})\widetilde{f}_k(s)}\right)g_k(u)du.$$

Note that when there is no setup time, from the equations given above we arrive at Eq. (1).