

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

Pedram Sahba, Barış Balcıoğlu, Dragan Banjevic

University of Toronto, Department of Mechanical and Industrial Engineering
5 King's College Rd., Toronto, ON M5S 3G8, CANADA,
pedram@mie.utoronto.ca, baris@mie.utoronto.ca, banjev@mie.utoronto.ca

Abstract

In this paper, we study a queueing system serving multiple classes of customers with an unreliable server. Each class is a finite-calling population. The server is subject to operation-independent disruptions (OID) and requires a setup time before resuming interrupted service or picking up a new customer. We refer to this case as the OID $M/G/1//N$ queue. We analyze the busy period and derive the steady-state system size distributions at departure/arrival epochs and arbitrary time epochs. We introduce the residual augmented process completion times (APCT) conditioned on the number of customers in the system to obtain the system time distribution. We also study the single-class ODD $M/G/1//N$ queue where ODD refers to the server being subject to operation-dependent disruptions.

Keywords and Phrases: Server disruptions, operation-independent disruptions, operation-dependent disruptions, process completion time, busy period analysis

1 Introduction

In this paper, we analyze an $M/G/1//N$ queuing system with an unreliable server serving m finite-source populations/customer classes indexed by $i = 1, \dots, m$. Each population i consists of N_i customers (type i 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 modeling telecommunication or computer networks (e.g., Sztrik and Gál, 1990, Almasi and Sztrik, 2004), 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 or fail from time to time; jobs generated by the terminals cannot be processed until the CPU is recovered. 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, again modeled as a single server (see, e.g., Wang, 1990). It is possible that the repair facility is unavailable from time to time; this increases the wait times of failed machines in the repair shop.

We assume that customers from different classes are served according to the preemptive-resume priority discipline. We consider setup times prior to picking up the next customer or resuming the service of an interrupted customer. The server can be disrupted from time to time whether it is idle, under setup or serving a customer. We define the times between interruptions as the times between the end of an interruption and the next interruption, and refer to them as the ON periods. We assume that ON periods and times between customer arrivals are exponentially distributed (possibly with different arrival rates for different customer types). The distinctive nature of our model is its capacity to include multiple classes and setup times. Another major contribution is our assumption that service, setup and down times (also referred to as OFF periods) are random variables (r.v.s) with general distributions.

In our problem, the server can be disrupted/interrupted even when it is idle or being setup. Therefore, we refer to this queueing system as the OID $M/G/1//N$ queue where

OID stands for the “operation-independent disruptions” indicating that the server can be disrupted at any time other than during an OFF period. If one assumes that the characteristics of times to interruptions and down times experienced by an idle server differ from when it is set up or serves customers, we have the ODD $M/G/1//N$ queue where ODD stands for “operation-dependent disruptions”. We present the analysis of the ODD $M/G/1//N$ queue in Section 2.6. Note that we adopt the definitions of OID and ODD from Altıok (1997, p. 85).

Queueing models with unreliable servers have been widely studied since the seminal paper by White and Christie (1957). 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. Gaver (1962), Avi-Itzhak and Naor (1963), and Thiruvengadam (1963) extend this model assuming that service times and OFF periods have general distributions. In his analysis, Gaver (1962) 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 it may experience. Avt-Itzhak and Naor (1963) and Thiruvengadam (1963) consider both operation-dependent and operation-independent ON periods. Mitrany and Avi-Itzhak (1968) and Neuts and Lucantoni (1979) study the multi-server $M/M/c$ queues with random breakdowns. For $M/G/1$ queues with operation-independent ON times, Federgruen and Green (1986) derive bounds and approximations for the mean waiting time, probability of delay and steady-state system size distribution when ON and OFF periods are general i.i.d. r.v.s. Federgruen and Green (1988) 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 Wang, Cao and Li (2001), Atencia, Bouza, and Moreno (2008), and Fiems, Maertens, and Bruneel (2008). Balcioğlu, Jagerman, and Altıok (2007) design an accurate approximation 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 Stecke and Aronson, 1985, and Haque and Armstrong, 2007, for an extensive bibliography on the MIP) with unreliable servers. Wang (1990) analyzes the $M/M/1//N$ queue with an unreliable server. For both operation-dependent and operation-independent interruptions, Wang assumes exponential ON and OFF periods. Wang and Kuo (1997) extend this model assuming exponential operation-independent ON periods, Erlangian service times and Erlangian OFF periods. Chakravarthy and Agrawal (2003) generalize the results of Wang and Kuo by considering phase-type distributions for service times and OFF periods.

As the literature review suggests, using non-exponential distributions for underlying r.v.s in these queueing systems is challenging. Incorporating non-exponential times between arrivals of customers in systems with finite-calling population is not analytically tractable. Nor is incorporating a non-exponential ON period distribution in exact analyses of systems experiencing operation-dependent and operation-independent server interruptions (except in $M/G/1$ systems with phase-type ON periods as in Federgruen and Green, 1988 and Balcioğlu, Jagerman, and Altıok, 2007). Similar difficulties arise for general service time and OFF period distributions. Among the three papers that are relatively closest to our problem, (Wang, 1990, Wang and Kuo, 1997, and Chakravarthy and Agrawal, 2003) two have successfully incorporated either Erlang distribution (Wang and Kuo, 1997) or phase-type distributions (Chakravarthy and Agrawal, 2002) for both r.v.s. The 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.

In Section 2, we consider the OID $M/G/1//N$ queue serving a single finite-calling population. We redefine the process completion time r.v., this time including setup times, and obtain its Laplace Transform (LT) in Section 2.1. This is followed in Section 2.2 by the busy period analysis of the system. Here, we derive its LT and the mean length of busy period. This enables us to obtain the steady-state system size distribution at departure/arrival and arbitrary time epochs in Section 2.3. For the probabilities at arbitrary time epochs, we need the LT of residual time left until the departure of the first customer from the system. This is derived in Section 2.4. After completing the analysis for a single finite-population, we design a recursive method to include multiple classes in the OID $M/G/1//N$ queue in Section 2.5. The single-class ODD $M/G/1//N$ queue is analyzed in Section 2.6.

As a special case, we propose an alternative solution for the OID $M/M/1//N$ queue in Section 3. With exponential service times, this method considers general OFF times; however, setup times are assumed to be negligible. Its extension to multiple classes still requires the busy period analysis from the OID $M/G/1//N$ queue, as explained in Section 3.1. We summarize our conclusions in Section 4. All proofs appear in Appendix A.

2 The OID $M/G/1//N$ Queue

In this section, we analyze a queueing system with an unreliable single server serving a finite population of N customers (the case with multiple finite populations is studied in Section 2.5). The times between the completion of a customer's service and the next arrival follow an exponential distribution with rate λ . The actual service times – in the absence of disruptions and excluding setup times – are independent and identically distributed (i.i.d.) r.v.s with an LT, $\tilde{b}(s)$. Whether there are customers in the system or not, from time to time the server is subject to interruptions (e.g., failures). In other words, the server is subject to “operation-independent” interruptions; this differentiates the problem from those where a server can be interrupted only when it is serving a customer. When an interruption occurs, the server becomes unavailable or “down”. The length of down times (e.g., repair times) are i.i.d. r.v.s,

denoted by D ; this follows a general continuous distribution $F(y) = \int_0^y f(u)du$ with density function $f(y)$, and has a LT $\tilde{f}(s)$. Letting $\bar{F}(y) = 1 - F(y)$, its first moment will be denoted by $E[D] = \int_0^\infty \bar{F}(y)dy$ and its hazard rate function by $\beta(y) = f(y)/\bar{F}(y)$. During down times, no additional failures occur and the times between the end of an interruption/down time and the next interruption are exponentially distributed r.v.s with rate α . If the interruption occurs during a service time, the customer under service is preempted; service resumes from the point of interruption once the server is ready to serve again. During the service time of a customer, the server may have no interruption, or it may have one or more. Each time a customer attempts to be served (for the first time or after an interruption) the server undergoes a setup/loading time which is denoted by the i.i.d. r.v. U with a density function $g(y)$ that is independent of both D and the (remaining) service time r.v. Interruptions can occur during setup time. At the end of the ensuing down time, a new setup time is generated from the same distribution until one is not interrupted. Only then does the server start serving or resuming serving the customer.

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.

We employ three stochastic processes to characterize the state of the system at time t : $R(t)$ equals 0 if the server is up, and 1 if it is down; $W(t) \in \{0, 1, \dots, N\}$ is the number of customers not yet in the queueing system; $V(t)$ is the elapsed time since the server went down. The elapsed time since the last setup time started 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 customers in the queueing system at time t , which is $N - W(t)$, because it is easier to express the state dependent arrival rates via $W(t)$ in our derivations. In the rest of the paper, we denote the mean for any r.v. X by $E[X]$.

2.1 The Process Completion Time with Setup Times

The process completion time (PCT) r.v. (Gaver, 1962), denoted by C , is the total time a customer spends on the server; this includes the actual service time plus possible OFF periods it may experience. The literature on the PCT ignores setup times, and considers OFF periods to be down times during which the server cannot experience new interruptions. We extend this model by incorporating setup times during which a customer has to wait. In our problem, C is the elapsed time between the instant a customer's first setup time begins and the instant the same customer departs from the system. This means that if interruptions occur, once the subsequent down time is over, a customer waits for an uninterrupted setup time before it resumes its service. 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.

Let $C(U, Z)$ be the r.v. denoting the PCT as a function of the setup time U and the remaining service time Z r.v.s (Z can be also the service time of the customer finding the server idle) and y the time until the next interruption. Then,

$$C(U, Z) = \begin{cases} U + Z, & \text{if } y \geq U + Z, \\ y + D + C(U, Z - (y - U)), & \text{if } U \leq y < Z + U, \\ y + D + C'(U, Z), & \text{if } 0 \leq y < U, \end{cases}$$

where $C'(U, Z)$ is identically distributed as $C(U, Z)$. Given that $U = u$ and $Z = z$, the LT of $C(U = u, Z = z)$, $\tilde{c}(s|u, z)$ is given by

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

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

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

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

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

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

we obtain the following solution

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

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

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

Similarly, when we remove the condition on u , we obtain the LT of C as

$$\tilde{c}(s) = \int_0^\infty \tilde{b} \left(s + \alpha - \frac{\alpha\tilde{f}(s)}{e^{(s+\alpha)u} + \frac{\alpha}{s+\alpha}(1 - e^{(s+\alpha)u})\tilde{f}(s)} \right) g(u)du.$$

Note that when there is no setup time, from the equations given above we arrive at the LT found in the literature (e.g., Altıok, 1997, p. 94)

$$\tilde{c}(s) = \tilde{b}(s + \alpha - \alpha\tilde{f}(s)). \quad (1)$$

2.2 Busy Period Analysis for the OID $M/G/1//N$ Queue

We define a busy period as an interval starting with either an interruption or a customer's arrival when the server is idle and ending as soon as the server is idle again. Thus, each busy period starts with an "initial delay" either in the form of a down time after an interruption or a PCT after a customer's arrival. Let $p_N^D(n)$ ($p_{N-1}^C(n)$) be the probability of having $0 \leq n \leq N$ ($0 \leq n \leq N-1$) customers present at the end of a down time (PCT) initiating a busy period in the $M/G/1//N$ system. Unlike the systems with constant customer arrival rates, in this analysis, state dependent arrival rates must be taken into account.

Before presenting the following Theorem, we define $P_N^D(n|d)$ ($P_{N-1}^C(n|c)$) as the probability of having n customers in the $M/G/1//N$ system at the end of the down time (PCT)

initiating a busy period given that $D = d$ ($C = c$). And, $\tilde{P}_N^D(n, s) = \int_0^\infty P_N^D(n|y)e^{-sy}f(y)dy$
($\tilde{P}_{N-1}^C(n, s) = \int_0^\infty P_{N-1}^C(n|z)e^{-sz}c(z)dz$).

Theorem 1 *The LT $\tilde{P}_N^D(n, s)$ is given by*

$$\tilde{P}_N^D(0, s) = \tilde{f}(s + N\lambda), \quad (2)$$

$$\tilde{P}_N^D(n, s) = \sum_{i=N-n}^N (-1)^{i-(N-n+1)} \binom{N}{i} \binom{i}{N-n} (\tilde{f}(s) - \tilde{f}(s + i\lambda)), \quad 0 < n < N, \quad (3)$$

$$\tilde{P}_N^D(N, s) = \sum_{i=1}^N (-1)^{i-1} \binom{N}{i} (\tilde{f}(s) - \tilde{f}(s + i\lambda)). \quad (4)$$

Note that Theorem 1 can be adjusted to obtain $\tilde{P}_{N-1}^C(n, s)$ (see the proof of Corollary 2).

The following Corollary directly follows from Theorem 1 since $P_N^D(n) = \tilde{P}_N^D(n, 0)$.

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

$$P_N^D(0) = \tilde{f}(N\lambda), \quad (5)$$

$$P_N^D(n) = \sum_{i=N-n}^N (-1)^{i-(N-n+1)} \binom{N}{i} \binom{i}{N-n} (1 - \tilde{f}(i\lambda)), \quad 0 < n < N, \quad (6)$$

$$P_N^D(N) = \sum_{i=1}^N (-1)^{i-1} \binom{N}{i} (1 - \tilde{f}(i\lambda)). \quad (7)$$

Similarly,

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

$$P_{N-1}^C(0) = \tilde{c}((N-1)\lambda), \quad (8)$$

$$P_{N-1}^C(n) = \sum_{i=N-1-n}^{N-1} (-1)^{i-(N-n)} \binom{N-1}{i} \binom{i}{N-1-n} (1 - \tilde{c}(i\lambda)), \quad 0 < n < N-1, \quad (9)$$

$$P_{N-1}^C(N-1) = \sum_{i=1}^{N-1} (-1)^{i-1} \binom{N-1}{i} (1 - \tilde{c}(i\lambda)). \quad (10)$$

If there are no customers waiting for service at the end of an initial delay, the busy period ends; otherwise, it continues until the server is idle. If there are n customers present in the system at the end of an initial delay ($1 \leq n \leq N$ if the initial delay is a down time, and $1 \leq n \leq N - 1$ if it is a PCT), in addition to the initial delay, the busy period consists of n sub-cycles. Each sub-cycle starting with i customers in the system ($1 \leq i \leq n$) is the time it takes until $i - 1$ customers remain in this system and is identical in distribution to the busy period of an $M/G/1/(N - i + 1)$ system (the queuing system with the same underlying stochastic processes but serving a finite population of $N - i + 1$ customers) initiated by a PCT (see Shanthikumar and Sumita, 1985, for a similar approach analyzing the $M/G/1/N$ queue without setup times and interruptions). Therefore, if there are n customers in the $M/G/1/N$ system at the end of an initial delay, the first (last) sub-cycle is identical in distribution to the busy period initiated by a PCT in the $M/G/1/(N - n + 1)$ system ($M/G/1/N$ system). If we denote the length of the busy periods initiated by a down time and a PCT in an $M/G/1/j$ system by T_j^D and T_j^C , and denote their LT's by $\tilde{h}_j^D(s)$ and $\tilde{h}_j^C(s)$, respectively, for the $M/G/1/N$ system, we have

$$T_N^D = \begin{cases} D, & \text{if there are no arrivals at the end of } D, \\ D + \sum_{j=N-n+1}^N T_j^C, & \text{if } 0 < n \leq N \text{ arrivals at at the end of } D, \end{cases}$$

$$T_N^C = \begin{cases} C, & \text{if there are no arrivals at the end of } C, \\ C + \sum_{j=N-n+1}^N T_j^C, & \text{if } 0 < n \leq N - 1 \text{ arrivals at the end of } C, \end{cases}$$

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

$$\tilde{h}_N^D(s) = \tilde{f}(s + N\lambda) + \sum_{n=1}^N \tilde{P}_N^D(n, s) \prod_{j=N-n+1}^N \tilde{h}_j^C(s), \quad (11)$$

$$\tilde{h}_N^C(s) = \tilde{c}(s + (N - 1)\lambda) + \sum_{n=1}^{N-1} \tilde{P}_{N-1}^C(n, s) \prod_{j=N-n+1}^N \tilde{h}_j^C(s).$$

Solving the above equation for $\tilde{h}_N^C(s)$ we get

$$\tilde{h}_N^C(s) = \frac{\tilde{c}(s + (N - 1)\lambda)}{1 - \sum_{n=1}^{N-1} \tilde{P}_{N-1}^C(n, s) \prod_{j=N-n+1}^{N-1} \tilde{h}_j^C(s)}.$$

Since a busy period starts either with an interruption or a customer arrival, the LT of the length of the busy period r.v. T_N in the $M/G/1//N$ system is

$$\tilde{h}_N(s) = \frac{\alpha}{\alpha + N\lambda} \tilde{h}_N^D(s) + \frac{N\lambda}{\alpha + N\lambda} \tilde{h}_N^C(s). \quad (12)$$

Then, the mean length of the busy period is

$$E[T_N] = \frac{\alpha}{\alpha + N\lambda} E[T_N^D] + \frac{N\lambda}{\alpha + N\lambda} E[T_N^C], \quad (13)$$

where

$$\begin{aligned} E[T_N^D] &= -\left. \frac{d\tilde{h}_N^D(s)}{ds} \right|_{s=0} = E[D] + \sum_{n=1}^N E[T_n^C] \sum_{j=N-n+1}^N P_N^D(j), \\ E[T_N^C] &= -\left. \frac{d\tilde{h}_N^C(s)}{ds} \right|_{s=0} = \frac{E[C] + \sum_{n=2}^{N-1} E[T_n^C] \sum_{j=N-n+1}^{N-1} P_{N-1}^C(j)}{P_{N-1}^C(0)}. \end{aligned} \quad (14)$$

Note that the times between two busy periods follow an exponential distribution with rate $\alpha + N\lambda$. By invoking the renewal theorem, the fraction of time the server is idle and up is $(1 + E[T_N](\alpha + N\lambda))^{-1}$, and the fraction of time the server is up is $(1 + \alpha E[D])^{-1}$. Thus, the fraction of time the server is in-service is $(1 + \alpha E[D])^{-1} - (1 + E[T_N](\alpha + N\lambda))^{-1}$.

2.3 System Size Distribution in the OID $M/G/1//N$ Queue

In this section, we obtain the steady-state probabilities of having i customers out of the system at departure/arrival epochs in Section 2.3.1, and then provide the system size distribution at an arbitrary instant in Section 2.3.2.

2.3.1 System Size Distribution at Arrival/Departure Epochs

We start our analysis by studying the embedded Markov chain of the number of customers left in the system after a customer departs. Let $p_{i,j}$ 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, $0 < i < N$, in the queue, the steady-state probability of the next departure leaving j customers behind ($j = i - 1, \dots, N - 1$) 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 $M/G/1/(N - i + 1)$ system, and can be obtained by invoking Corollary 2. Any other transition from i , $0 < i < N$, is not possible. When the last departure leaves an empty system, consider that the next arrival can find the server down, or up and idle. If it finds the server down, in steady-state, the arrival waits for the residual down time until the setup time starts. We denote this r.v. by D_R . Following Fiems, Maertens, and Bruneel (2008), the LT of D_R can be found as

$$\tilde{f}_R(s) = \frac{N\lambda(N\lambda - s) + N\lambda\alpha(\tilde{f}(s) - \tilde{f}(N\lambda))}{(N\lambda + \alpha - \alpha\tilde{f}(N\lambda))(N\lambda - s)}, \quad (15)$$

with

$$\tilde{f}_R(N\lambda) = \lim_{s \rightarrow N\lambda} \tilde{f}_R(s) = \frac{N\lambda(1 - \alpha\tilde{f}'(N\lambda))}{N\lambda + \alpha - \alpha\tilde{f}(N\lambda)}.$$

Only then does the PCT of the customer arriving during a down time start. In order for such a customer to leave j customers behind ($j = 0, 1, \dots, N - 1$), there should be j arrivals during the interval $L = D_R + C$, with an LT of $\tilde{l}(s) = \tilde{f}_R(s)\tilde{c}(s)$, and a mean of

$$E[L] = -\frac{d\tilde{l}(s)}{ds}\Big|_{s=0} = E[D_R] + E[C]. \quad (16)$$

Using Corollary 2 by substituting $\tilde{l}(s)$ for $\tilde{c}(s)$, $P_{N-1}^L(j) = p_{0,j}$ ($j = 0, 1, \dots, N - 1$) can be obtained. In summary, we have

$$p_{i,j} = \begin{cases} P_{N-1}^L(j), & i = 0, \quad 0 \leq j \leq N - 1, \\ P_{N-i}^C(j - i + 1), & 1 \leq i < N, \quad i - 1 \leq j \leq N - 1, \\ 0, & \text{otherwise.} \end{cases}$$

Now that we have $p_{i,j}$, we can construct the $N \times N$ transition probability matrix \mathbf{P} . From $\mathbf{\Pi} = \mathbf{\Pi P}$ and $\sum_{i=1}^N \pi_i = 1$, we can solve for the $1 \times N$ vector $\mathbf{\Pi} = [\pi_N, \pi_{N-1}, \dots, \pi_1]$. Here, π_i is the steady-state probability of having i customers (including the departing customer) out of the queueing system at departure instants (or equivalently having $N - i$ customers

left behind in the queueing system). Note that since this is an ergodic Markov chain, π_i is also the steady-state probability that an arrival finds $N - i$ customers in the system.

2.3.2 System Size Distribution at an Arbitrary Instant

In this section, we obtain \bar{P}_i , the steady-state probability of having i customers out of the system.

Lemma 1 *With $E[T_N]$ as the mean length of a busy period,*

$$\bar{P}_N = \frac{N\lambda + \alpha - \alpha \tilde{f}(N\lambda)}{N\lambda(1 + E[T_N](\alpha + N\lambda))}. \quad (17)$$

To obtain the entire distribution, we introduce the ‘‘augmented PCT’’ (APCT) r.v. denoted by \hat{C} , which is the PCT for all customers (i.e. $\hat{C} = C$) except those arriving as the first customers during a down time that initiates a busy period. In the latter case, the APCT is the residual down time such customers wait plus their PCT, that is $\hat{C} = L$. Then, the residual APCT r.v. \hat{C}_R with $\hat{c}_R(x)$ as its density function is the time left until the departure of the first customer (who may be waiting for the down time that initiates a busy period, or is in service, or is preempted) in the system. It is known that $P(\hat{C}_R = 0) = \bar{P}_N$, i.e., the probability that there are no customers in the queueing system, but we define $\hat{c}_R(0) = \lim_{x \rightarrow 0} \hat{c}_R(x)$.

Let $\hat{C}_R(t)$ denote the residual APCT at time t and

$$P_i(t, x)dx = P\{W(t) = i, x < \hat{C}_R(t) < x + dx\}, \quad 0 \leq i \leq N - 1,$$

denote the joint probability distribution of having i customers out of the queueing system at time t ($W(t) = i$), and that the residual APCT of the customer who is preempted or currently receiving service is 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 $P_{-1}(t, x)$ and its limiting probability are 0,

$$\begin{aligned} P_{N-1}(t + \Delta t, x) &= (1 - (N - 1)\lambda\Delta t)P_{N-1}(t, x + \Delta t) + N\lambda\Delta t\bar{P}_N(t)l(x) \\ &\quad + P_{N-2}(t, 0)c(x)\Delta t + o(\Delta t), \\ P_i(t + \Delta t, x) &= (1 - i\lambda\Delta t)P_i(t, x + \Delta t) + (i + 1)\lambda\Delta tP_{i+1}(t, x + \Delta t) \\ &\quad + P_{i-1}(t, 0)c(x)\Delta t + o(\Delta t), \quad 0 \leq i \leq N - 2, \end{aligned}$$

where $\bar{P}_N(t)$ is the probability of having N customers out of the system at time t . Here $l(x)$, and $c(x)$ are the density functions of the r.v.s L and C , respectively, and $c(x)\Delta t = P(x \leq C \leq x + \Delta t)$. Re-arranging the equations given above, we obtain

$$\begin{aligned} \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial x}\right) P_{N-1}(t, x) &= -(N - 1)\lambda P_{N-1}(t, x) + N\lambda\bar{P}_N(t)l(x) + P_{N-2}(t, 0)c(x), \\ \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial x}\right) P_i(t, x) &= -i\lambda P_i(t, x) + (i + 1)\lambda P_{i+1}(t, x) + P_{i-1}(t, 0)c(x), \quad 0 \leq i \leq N - 2. \end{aligned}$$

Letting $P_i(x) = \lim_{t \rightarrow \infty} P_i(t, x)$, if we take the limit of the equations given above as $t \rightarrow \infty$,

$$\frac{d}{dx} P_{N-1}(x) = (N - 1)\lambda P_{N-1}(x) - N\lambda\bar{P}_N l(x) - P_{N-2}(0)c(x), \quad (18)$$

$$\frac{d}{dx} P_i(x) = i\lambda P_i(x) - (i + 1)\lambda P_{i+1}(x) - P_{i-1}(0)c(x), \quad 0 \leq i \leq N - 2. \quad (19)$$

Observe that $P_i(x)$ is the density function of the residual APCT and i customers are out of the queueing system. When $i = 0$, integrating both sides of Eq. (19) gives

$$\begin{aligned} P_0(\infty) - P_0(0) &= -\lambda\bar{P}_1 \\ P_0(0) &= \lambda\bar{P}_1. \end{aligned}$$

Recursively, we can show that

$$P_i(0) = (i + 1)\lambda\bar{P}_{i+1}, \quad 0 \leq i \leq N - 1. \quad (20)$$

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

$$\begin{aligned}
\pi_{i+1} &\equiv P\{i \text{ customers out of the system} | \text{a departure is about to occur}\} \\
&= \frac{P_i(0)}{\widehat{c}_R(0)} = \frac{P_i(0)}{\sum_{i=0}^{N-1} P_i(0)}, \quad 0 \leq i \leq N-1, \\
\pi_i &= \frac{i\lambda\bar{P}_i}{\sum_{i=1}^N i\lambda\bar{P}_i}, \quad 1 \leq i \leq N, \\
\bar{P}_i &= \frac{N\bar{P}_N}{i\pi_N} \pi_i, \quad 1 \leq i \leq N.
\end{aligned} \tag{21}$$

Using Eq. (21) together with Lemma 1, we derive the solution for \bar{P}_i , which is also the steady-state probability of having $N-i$ customers in the system. Although we do not need it for our computations, Eq. (21) also helps us obtain $\widehat{c}_R(0) = N\lambda\bar{P}_N/\pi_N$.

The following theorem provides an alternative solution. Before presenting it, we introduce the conditional residual APCT, given that there are i customers out of the system. By definition, its density function is (the LT $\tilde{c}_{R|i}(s)$ is obtained in Section 2.4)

$$\widehat{c}_{R|i}(x) = \frac{P_i(x)}{\bar{P}_i}. \tag{22}$$

Theorem 2 *There is a recursive relationship between the steady-state probabilities \bar{P}_i so that*

$$\bar{P}_{N-1} = \frac{N}{(N-1)} \frac{1 - \tilde{l}((N-1)\lambda)}{\tilde{c}((N-1)\lambda)} \bar{P}_N, \tag{23}$$

$$\bar{P}_i = \frac{(i+1)\bar{P}_{i+1}}{i\tilde{c}(i\lambda)} (1 - \tilde{c}_{R|i+1}(i\lambda)), \quad 0 < i \leq N-2. \tag{24}$$

2.4 The Conditional Residual Augmented Process Completion Time

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

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

$$\begin{aligned} \widehat{c}_{R|N-1}(x) &= \frac{(N-1)\lambda e^{(N-1)\lambda x}}{1 - \widetilde{l}((N-1)\lambda)} \left\{ \widetilde{c}((N-1)\lambda) \int_x^\infty e^{-(N-1)\lambda u} l(u) du \right. \\ &\quad \left. + (1 - \widetilde{l}((N-1)\lambda)) \int_x^\infty e^{-(N-1)\lambda u} c(u) du \right\}, \end{aligned} \quad (25)$$

$$\widehat{c}_{R|i}(x) = i\lambda e^{i\lambda x} \int_x^{+\infty} e^{-i\lambda u} \left(\widetilde{c}(i\lambda) \frac{\widehat{c}_{R|i+1}(u)}{1 - \widetilde{c}_{R|i+1}(i\lambda)} + c(u) \right) du, \quad 0 < i \leq N-2. \quad (26)$$

And,

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

$$\widetilde{c}_{R|N-1}(s) = \frac{(N-1)\lambda}{s - (N-1)\lambda} \frac{\widetilde{c}((N-1)\lambda) \left(1 - \widetilde{l}(s)\right) - \widetilde{c}(s) \left(1 - \widetilde{l}((N-1)\lambda)\right)}{1 - \widetilde{l}((N-1)\lambda)}, \quad (27)$$

$$\widetilde{c}_{R|i}(s) = \frac{i\lambda}{s - i\lambda} \left(\widetilde{c}(i\lambda) \frac{1 - \widetilde{c}_{R|i+1}(s)}{1 - \widetilde{c}_{R|i+1}(i\lambda)} - \widetilde{c}(s) \right), \quad 0 < i \leq N-2, \quad (28)$$

$$\widetilde{c}_{R|0}(s) = \frac{\overline{P}_1 \lambda (1 - \widetilde{c}_{R|1}(s))}{\overline{P}_0 s}. \quad (29)$$

The following Theorem is presented without a proof since its proof is, in principle, the same as Theorem 2.2.2 in Kerner (2008); Kerner exploits Theorem 1 by van Doorn and Regterschot (1988).

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

Recall from Section 2.3.1 that, in steady-state an arrival finds $N - i$ customers in the system with probability π_i . Using Theorem 5, the system time of such a customer is the residual APCT of the customer on the server plus the sum of $N - i$ PCT's of the customers waiting in the queue and the new arrival; this has the LT of

$$\widetilde{w}_i(s) = \widetilde{c}_{R|i}(s) \widetilde{c}^{N-i}(s), \quad 1 \leq i \leq N-1.$$

With probability π_N , the customer finds no customers in the system and its system time is L . By the law of total probability, the LT of the system time of a customer is given by

$$\widetilde{w}(s) = \sum_{i=1}^{N-1} \pi_i \widetilde{w}_i(s) + \pi_N \widetilde{l}(s).$$

2.5 The Multi-class OID $M/G/1//N$ Queue

In this section, we consider m finite-source populations/customer classes indexed by $i = 1, \dots, m$ served by a single unreliable server. Each population i consists of N_i customers (type i customer). The times between the completion of a type i customer's service and the next arrival at the queueing system follow an exponential distribution with rate λ_i . The actual service times of customers – in the absence of disruptions and excluding setup times – are i.i.d. r.v.s with an LT, $\tilde{b}_i(s)$. The assumptions made concerning U , D and times between the end of an interruption/down time and the next interruption following an exponential distribution with rate α remain valid. Customer classes are prioritized as class 1 to m from highest to lowest. Since preemptive-resume priority policy is used, a class i customer can be serviced only during the periods the server is not allocated to higher priority classes 1 to $i-1$. The busy period for class i customers and the distribution of the number of type i customers in the system can be found from the single class $M/G/1//N_i$ queue with an “effective” interruption rate, from the point of view of type i customers, as $\alpha_i = \alpha + \sum_{n=1}^{i-1} N_n \lambda_n$ (to be used instead of α) and using the LT of the busy period for type $i-1$ customers from Eq. (12) instead of $\tilde{f}(s)$. Obviously, the LT of the busy period for type $i-1$ customers can be found by applying this procedure recursively, starting by analyzing class 2 in the $M/G/1//N_2$ queue.

2.6 The ODD $M/G/1//N$ Queue

In the ODD $M/G/1//N$ queue, the server may experience different disruptions depending on the state of the server, i.e., the interruption process is altered or halted when the server becomes idle. If the failure rate and the LT transform of the length of down times during the idle state change to α_I and $\tilde{f}_I(s)$, different from those of the busy state, only the time between busy periods and the down times that initiate a busy period are affected. In contrast, the process completion time remains unchanged. In this case, Eqs. (11), (12), (13), (15), and (17) must be obtained using α_I and $\tilde{f}_I(s)$ instead of α and $\tilde{f}(s)$. In a special case, when

$\alpha_I = 0$, a busy period can only be initiated by a PCT, and in our derivation of the system size distribution, no residual down time should be taken into account (i.e., $D_R = 0$).

Unlike in the OID $M/G/1//N$ queue, a high-priority class is not completely independent of the lower-priority classes in the ODD $M/G/1//N$ queue. This prevents us from using a similar approach as presented in Section 2.5 for the ODD case. To explain the dependency of a high-priority class on the low-priority class, consider a system with two classes. If the failure rate, or the fleet size of the low-priority class increases, or its repair rate decreases, the amount of time the server allocates to this fleet tends to increase. This tends to decrease the amount of time the server stays idle in both the OID and ODD systems. However, in the ODD $M/G/1//N$ queue, this also decreases (increases) the probability that the system experiences interruptions with rate α_I (α). Consequently, the performance measures concerning the high-priority class in the ODD $M/G/1//N$ queue are affected by the characteristics of the low-priority class.

If one is only interested in the system size distribution when $\alpha_I = 0$ in the single-class ODD system, as an alternative, the PCT r.v. can be used as the service time r.v. in an $M/G/1//N$ queue without interruptions, as analyzed by Gupta and Srinivasa Rao (1996); this will be referred to as the $M/PCT/1//N$ queue. Note that the $M/PCT/1//N$ queue should have the same λ and N as the original ODD $M/G/1//N$ queue; additionally, $\tilde{c}(s)$ in Eq. (1) should be used instead of the service time LT in the algorithm by Gupta and Srinivasa Rao. This way, the probability of having i customers in the $M/PCT/1//N$ equals that of the original ODD $M/G/1//N$ queue.

3 The OID $M/M/1//N$ Queue

In this section, we analyze the single class $M/M/1//N$ queue. Our assumptions here differ from those in Section 2 in that there is no setup time and actual service times in the absence of disruptions are exponentially distributed with rate μ . Note that due to the memoryless property of the exponential service times, the remaining service times of interrupted

customers are also exponentially distributed with rate μ . We introduce

$$P_{i,0}(t) = Pr\{W(t) = i, R(t) = 0\}, \quad 0 \leq i \leq N,$$

which is the probability that there are i customers out of the queueing system and the server is up at time t . Let also

$$P_{i,1}(t, y)dy = Pr\{W(t) = i, R(t) = 1, y \leq V(t) \leq y + dy\}, \quad 0 \leq i \leq N,$$

be the probability that there are i customers out of the queueing system, the server is down at time t , and the length of time since the server went down is in the interval $[y, y + dy]$.

By considering the transitions between states at time t , we have

$$\frac{d}{dt}P_{N,0}(t) = -(N\lambda + \alpha)P_{N,0}(t) + \mu P_{N-1,0}(t) + \int_0^\infty P_{N,1}(t, y)\beta(y)dy, \quad (30)$$

$$\begin{aligned} \frac{d}{dt}P_{i,0}(t) &= -(i\lambda + \mu + \alpha)P_{i,0}(t) + (i+1)\lambda P_{i+1,0}(t) + \mu P_{i-1,0}(t) \\ &\quad + \int_0^\infty P_{i,1}(t, y)\beta(y)dy, \quad 0 \leq i \leq N-1, \end{aligned} \quad (31)$$

and

$$\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial y} + N\lambda + \beta(y)\right)P_{N,1}(t, y) = 0, \quad (32)$$

$$\begin{aligned} \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial y}\right)P_{i,1}(t, y) &= -(i\lambda + \beta(y))P_{i,1}(t, y) \\ &\quad + (i+1)\lambda P_{i+1,1}(t, y), \quad 0 \leq i \leq N-1. \end{aligned} \quad (33)$$

Letting $P_{i,0} = \lim_{t \rightarrow \infty} P_{i,0}(t)$ and $P_{i,1}(y) = \lim_{t \rightarrow \infty} P_{i,1}(t, y)$ for $0 \leq i \leq N$, if we take the limit as $t \rightarrow \infty$ in Eqs. (30)-(33), we obtain

$$(N\lambda + \alpha)P_{N,0} = \mu P_{N-1,0} + \int_0^\infty P_{N,1}(y)\beta(y)dy, \quad (34)$$

$$(i\lambda + \mu + \alpha)P_{i,0} = (i+1)\lambda P_{i+1,0} + \mu P_{i-1,0} + \int_0^\infty P_{i,1}(y)\beta(y)dy, \quad 0 \leq i \leq N-1, \quad (35)$$

$$\frac{d}{dy}P_{N,1}(y) = -(N\lambda + \beta(y))P_{N,1}(y), \quad (36)$$

$$\frac{d}{dy}P_{i,1}(y) = -(i\lambda + \beta(y))P_{i,1}(y) + (i+1)\lambda P_{i+1,1}(y), \quad 0 \leq i \leq N-1, \quad (37)$$

and the boundary equation is

$$P_{i,1}(0) = \alpha P_{i,0}, \quad 0 \leq i \leq N. \quad (38)$$

We introduce the following to be used in Theorems to follow:

$$\mathcal{Q}_{N-1} = \frac{N\lambda + \alpha - \alpha \tilde{f}(N\lambda)}{\mu}, \quad (39)$$

$$\mathcal{Q}_{i-1} = \frac{(i\lambda + \mu + \alpha)\mathcal{Q}_i - (i+1)\lambda\mathcal{Q}_{i+1} - \alpha \sum_{j=i}^N \mathcal{Q}_j \zeta_{i,j}}{\mu}, \quad 1 \leq i \leq N-1. \quad (40)$$

$$\zeta_{i,i} = \tilde{f}(i\lambda), \quad 0 \leq i \leq N, \quad (41)$$

$$\zeta_{i,j} = \frac{j}{j-i} \zeta_{i,j-1} - \frac{i+1}{j-i} \zeta_{i+1,j}, \quad 0 \leq i \leq N. \quad (42)$$

$$\mathcal{D}_i = \mathcal{D}_{i+1} - \sum_{j=i}^N \mathcal{Q}_j \zeta_{i,j} + \mathcal{Q}_i, \quad 1 \leq i \leq N-1, \quad (43)$$

Theorem 6 *The steady-state probability that there are i customers out of the system is*

$$P_{i,0} = \frac{P_{N,1}(0)}{\alpha} \mathcal{Q}_i, \quad 1 \leq i \leq N-1, \quad (44)$$

$$P_{i,1} = \frac{P_{N,1}(0)}{i\lambda} \mathcal{D}_i, \quad 1 \leq i \leq N-1, \quad (45)$$

Corollary 3 *The probability density function of having i customers out of the system and an elapsed down time of y is*

$$P_{i,1}(y) = e^{-i\lambda y} \bar{F}(y) P_{N,1}(0) \sum_{j=i}^N \binom{j}{i} \mathcal{Q}_j (1 - e^{-\lambda y})^{j-i}, \quad 0 \leq i \leq N. \quad (46)$$

Theorem 7 *The probability that there are no customers in the system when the server is up is given by*

$$P_{N,0} = \frac{P_{N,1}(0)}{\alpha} = ((1 + \alpha E[D]) \sum_{i=0}^N \mathcal{Q}_i)^{-1}. \quad (47)$$

3.1 The Multi-class OID $M/M/1//N$ Queue

If the single unreliable server attends to multiple finite-source populations as described in Section 2.5, with the difference of no setup time and type i customers requiring exponential service times with rate μ_i , the analysis presented above can be used only for class 1 customers. For type $i > 1$ customers, we need to obtain the LT of the busy period of type $i-1$ customers to be used as the LT of down time r.v. in the OID $M/M/1//N_i$ queue with α_i given in Section 2.5 to be the effective interruption rate. For this purpose, the result of the busy period analysis from Section 2.2 can be used.

4 Conclusions

In this paper, we provide a method to obtain the exact steady-state system size distribution and busy period analysis of the $M/G/1//N$ queue with an unreliable server subject to operation-independent interruptions. We expand the classical queueing model by incorporating a general OFF period, service and setup time distributions, and considering multiple classes of customers. However, including non-exponential distributions to model times between customer arrivals and/or times between interruptions remains challenging and is an open research question. We provide an alternative analysis of the $M/M/1//N$ queue with exponential service times where setup times are negligible. We also analyze the single-class $M/G/1//N$ queue where the distributions of times to interruptions and down times change depending on if the server is in its idle or busy period.

Acknowledgements

This work was supported in part by Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

- Almasi, B. and J. Sztrik. 2004. “Reliability investigations of heterogeneous terminal systems using MOSEL”, *Journal of Mathematical Sciences*, Vol. 123, No. 1, 3795–3801.
- Altıok, T. 1997. *Performance Analysis of Manufacturing Systems*, Springer-Verlag, New York, NY.
- 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.
- 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.
- Balcioğlu, 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.
- 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.
- van Doorn E. A. and G. J. K. Regterschot. 1988. “Conditional PASTA”, *Operations Research Letters*, Vol. 7, No. 5, 229–232.
- Federgruen, A. and L. Green. 1986. “Queueing systems with service interruptions”, *Operations Research*, Vol. 34, No. 5, 752–768.
- Federgruen, A. and L. Green. 1988. “Queueing systems with service interruptions II”, *Naval Research Logistics*, Vol. 35, 345–358.
- 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.
- Gaver, D. P. 1962. “A waiting line with interrupted service, including priorities”, *Journal of the Royal Statistical Society*, Vol. 24, No. 1, 73–90.
- Gupta, U.C. and T.S.S. Srinivasa Rao. 1996. “Computing the steady state probabilities in

- $\lambda(n)/G/1/K$ queue”, *Performance Evaluation*, Vol. 24, 265–275.
- 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.
- Kerner, Y. 2008. “The Conditional Distribution of the Residual Service Time in the $M_n/G/1$ Queue,” *Stochastic Models*, Vol. 24, 364–375.
- 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.
- 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.
- 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.
- 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.
- 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.
- Thiruvengadam, K. 1963. “Queueing with breakdown”, *Operations Research*, Vol. 11, 62–71.
- 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.
- 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.
- 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.
- White, H. and L. Christie. 1958. “Queueing with preemptive priorities or with breakdown”,

Appendix A Proofs

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

Lemma A.1 *During the down time initiating a busy period in the $M/G/1//N$ system, the time-to-arrival r.v. $T_{N,n}$ of the n th customer has the following cumulative distribution function:*

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

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

$$T_{N,n} = \sum_{i=N-n+1}^N T_i,$$

where T_i follows an exponential distribution with rate $i\lambda$. Let $\tilde{h}_{N,n}(s)$ be the LT of $T_{N,n}$, then

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

Using

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

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

$$\tilde{h}_{N,n}(s) = (N - n + 1)\lambda \sum_{i=N-n+1}^N (-1)^{i-(N-n+1)} \binom{N}{i} \binom{i}{N-n+1} \frac{1}{i\lambda + s},$$

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

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

$$P_N^D(0|d) = P\{T_1 > d\} = 1 - H_{N,1}(d) = e^{-N\lambda d}, \quad (\text{A.3})$$

and for $0 < n < N$

$$\begin{aligned} P_N^D(n|d) &= P\{T_n < d < T_{n+1}\} = H_{N,n}(d) - H_{N,n+1}(d) \\ &= \sum_{i=N-n}^N (-1)^{i-(N-n+1)} \binom{N}{i} \binom{i}{N-n} (1 - e^{-i\lambda d}), \end{aligned} \quad (\text{A.4})$$

and finally,

$$P_N^D(N|d) = P\{T_N < d\} = H_{N,N}(d) = \sum_{i=1}^N (-1)^{i-1} \binom{N}{i} (1 - e^{-i\lambda d}). \quad (\text{A.5})$$

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

Proof. Corollary 2. The fundamental difference between a down time initiating a busy period and a PCT initiating a busy period are the following. The PCT has a different distribution from the down time, and at the beginning of the PCT, $N - 1$ customers are not yet in the queueing system. Therefore, Lemma A.1 and Theorem 1 can be employed for an $M/G/1/(N - 1)$ system where the down time has the same distribution as the PCT, and Eqs. (8)-(10) can be obtained. ■

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

$$\bar{P}_N = \lim_{t \rightarrow \infty} P\{(W(t) = N) \cap R(t) = 0\} + \lim_{t \rightarrow \infty} P\{(W(t) = N) \cap R(t) = 1\}. \quad (\text{A.6})$$

The probability of having no customers in the system and the server being up (as discussed at the end of Section 2.2) is

$$\lim_{t \rightarrow \infty} P\{(W(t) = N) \cap R(t) = 0\} = \frac{1}{1 + E[T_N](\alpha + N\lambda)}.$$

Observe that only during the down time which initiates a busy period can the server be down while no customer exists in the system; the average time the system remains empty during such a down time is given by

$$\int_0^\infty \left(\int_0^y t N \lambda e^{-N\lambda t} dt + y \int_y^\infty N \lambda e^{-N\lambda t} dt \right) f(y) dy = \frac{1 - \tilde{f}(N\lambda)}{N\lambda}.$$

Since the fraction of time the system is in a busy period initiated by a down time is

$$\frac{\alpha E[T_N^D]}{1 + E[T_N](\alpha + N\lambda)},$$

the fraction of time the server is empty and down is

$$\lim_{t \rightarrow \infty} P \{(W(t) = N) \cap R(t) = 1\} = \frac{\alpha \int_0^\infty \left(\int_0^y t N \lambda e^{-N\lambda t} dt + y \int_y^\infty N \lambda e^{-N\lambda t} dt \right) f(y) dy}{1 + E[T_N](\alpha + N\lambda)}.$$

The summation of these in Eq. (A.6) gives \bar{P}_N in Lemma 1. ■

Proof. Theorem 2. After substituting $P_{N-2}(0) = (N-1)\lambda\bar{P}_{N-1}$ from Eq. (20) into Eq. (18) and multiplying both sides by $e^{-(N-1)\lambda x}$, eventually, we have

$$\frac{d}{dx} (e^{-(N-1)\lambda x} P_{N-1}(x)) = -N\lambda e^{-(N-1)\lambda x} \bar{P}_N l(x) - (N-1)\lambda e^{-(N-1)\lambda x} \bar{P}_{N-1} c(x).$$

Integrating both sides gives

$$-e^{-(N-1)\lambda x} P_{N-1}(x) = -N\lambda \bar{P}_N \int_x^\infty e^{-(N-1)\lambda u} l(u) du - (N-1)\lambda \bar{P}_{N-1} \int_x^\infty e^{-(N-1)\lambda u} c(u) du. \quad (\text{A.7})$$

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

$$P_{N-1}(0) = N\lambda \bar{P}_N \tilde{l}((N-1)\lambda) + (N-1)\lambda \bar{P}_{N-1} \tilde{c}((N-1)\lambda).$$

The equation above together with Eq. (20) for $P_{N-1}(0)$ gives Eq. (23).

Similarly, by multiplying both sides of Eq. (19) by $e^{-i\lambda x}$, and skipping similar steps as in the first part of the proof, we arrive at

$$P_i(x) = e^{i\lambda x} \left(\int_x^\infty \lambda e^{-i\lambda u} (i+1) P_{i+1}(u) du + i\lambda \bar{P}_i \int_x^\infty e^{-i\lambda u} c(u) du \right). \quad (\text{A.8})$$

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

$$P_i(0) = \int_0^\infty \lambda e^{-i\lambda u} (i+1) P_{i+1}(u) du + i\lambda \bar{P}_i \int_0^\infty e^{-i\lambda u} c(u) du.$$

Note that by the definition given in Eq. (22), $\tilde{P}_{i+1}(s) = \bar{P}_{i+1} \hat{c}_{R|i}(s)$, which together with Eq. (20), leads us to

$$P_i(0) = (i+1)\lambda \bar{P}_{i+1} = (i+1)\lambda \bar{P}_{i+1} \tilde{c}_{R|i+1}(i\lambda) + i\lambda \bar{P}_i \tilde{c}(i\lambda),$$

from which Eq. (24) follows. ■

Proof. Theorem 3. Eq. (25) follows directly by substituting Eq. (23) in Eq. (A.7). Eq. (26), which is the same as Eq. (2) in Kerner (2008), is obtained by substituting Eq. (24) in Eq. (A.8). ■

Proof. Theorem 4. After multiplying both sides of Eq. (18) with e^{-sx} and integrating, we have

$$\begin{aligned} \int_0^\infty e^{-sx} dP_{N-1}(x) &= (N-1)\lambda \int_0^\infty e^{-sx} P_{N-1}(x) dx - N\lambda \bar{P}_N \int_0^\infty e^{-sx} l(x) dx \\ &\quad - P_{N-2}(0) \int_0^\infty e^{-sx} c(x) dx, \\ s\tilde{P}_{N-1}(s) - P_{N-1}(0) &= (N-1)\lambda \tilde{P}_{N-1}(s) - N\lambda \bar{P}_N \tilde{l}(s) - P_{N-2}(0) \tilde{c}(s), \\ \tilde{P}_{N-1}(s) &= \frac{N\lambda \bar{P}_N (1 - \tilde{l}(s)) - (N-1)\lambda \bar{P}_{N-1} \tilde{c}(s)}{s - (N-1)\lambda}. \end{aligned}$$

Note that for the last equation above, we used Eq. (20). After multiplying both sides of Eq. (23) by λ , we re-arranged it to express $N\lambda \bar{P}_N$. When this is substituted in the last equation above, we get

$$\tilde{P}_{N-1}(s) = \frac{(N-1)\lambda \bar{P}_{N-1} \left(\tilde{c}((N-1)\lambda) (1 - \tilde{l}(s)) - \tilde{c}(s) (1 - \tilde{l}((N-1)\lambda)) \right)}{(1 - \tilde{l}((N-1)\lambda))(s - (N-1)\lambda)}.$$

Dividing the equation given above by \bar{P}_{N-1} according to Eq. (22) gives Eq. (27). Similarly, Eq. (28) can be found by starting with Eq. (19) and is the same as Eq. (4) in Kerner (2008). When $i = 0$, multiplying both sides of Eq. (19) by e^{-sx} , integrating the results, and then using Eq. (22), gives

$$\begin{aligned} \tilde{P}_0(s) &= \frac{\lambda(\bar{P}_1 - \tilde{P}_1(s))}{s} \\ &= \frac{\lambda \bar{P}_1 (1 - \tilde{c}_{R|1}(s))}{s}. \end{aligned}$$

Dividing this equation by \bar{P}_0 according to Eq. (22) gives Eq. (29). ■

Proof. Theorem 6. If we divide both sides of Eq. (36) by $e^{-N\lambda y - \int_0^y \beta(x) dx} P_{N,1}(0)$ and Eq.

(37) by $e^{-i\lambda y - \int_0^y \beta(x) dx} P_{N,1}(0)$, we get

$$\frac{d}{dy} \left(\frac{e^{N\lambda y + \int_0^y \beta(x) dx} P_{N,1}(y)}{P_{N,1}(0)} \right) = 0, \quad (\text{A.9})$$

$$\frac{d}{dy} \left(\frac{e^{i\lambda y + \int_0^y \beta(x) dx} P_{i,1}(y)}{P_{N,1}(0)} \right) = \frac{(i+1)\lambda e^{i\lambda y + \int_0^y \beta(x) dx} P_{i+1,1}(y)}{P_{N,1}(0)}, \quad 0 \leq i \leq N-1, \quad (\text{A.10})$$

which are first order differential equations. Next, recalling that $\bar{F}(y) = e^{-\int_0^y \beta(x) dx}$, we define

$$Q_i(y) = \frac{P_{i,1}(y)}{e^{-i\lambda y} \bar{F}(y) P_{N,1}(0)}, \quad 0 \leq i \leq N, \quad (\text{A.11})$$

and solve Eqs. (A.9) and (A.10) as

$$Q_N(y) = 1, \quad (\text{A.12})$$

$$Q_i(y) = Q_i(0) + (i+1)\lambda \int_0^y Q_{i+1}(x) e^{-\lambda x} dx, \quad 0 \leq i \leq N-1. \quad (\text{A.13})$$

Considering the definition given in Eq. (A.11), and employing Eqs. (34), (35) and (38), we obtain

$$Q_{N-1}(0) = \frac{N\lambda + \alpha - \alpha \int_0^\infty Q_N(y) e^{-N\lambda y} f(y) dy}{\mu}, \quad (\text{A.14})$$

$$Q_{i-1}(0) = \frac{(i\lambda + \mu + \alpha)Q_i(0) - (i+1)\lambda Q_{i+1}(0) - \alpha \int_0^\infty Q_i(y) e^{-i\lambda y} f(y) dy}{\mu}, \quad 1 \leq i \leq N-1. \quad (\text{A.15})$$

For simplicity, we define

$$\mathcal{Q}_i = Q_i(0), \quad (\text{A.16})$$

$$\mathcal{B}_i = \int_0^\infty Q_i(y) e^{-i\lambda y} f(y) dy. \quad (\text{A.17})$$

In order \mathcal{Q}_i and \mathcal{B}_i to be finite, we have to show that $Q_i(y)$ is finite for all $i = 0, \dots, N$, which is proved in the following Lemma.

Lemma A.2 $\lim_{y \rightarrow \infty} Q_i(y) = Q_i(\infty)$ exists and is finite. We also have $Q_i(y) \leq Q_i(\infty)$.

Proof. Lemma A.2. From Eq. (A.11), $Q_i(y) \geq 0$ and from Eq. (A.13), we see that $Q_i(y)$ is increasing in y . Let $Q_i(\infty) = \lim_{y \rightarrow \infty} Q_i(y)$. Then, $Q_i(y) \leq Q_i(\infty)$, $0 \leq i \leq N-1$. If we take the limit as $y \rightarrow \infty$ in Eq. (A.13),

$$Q_i(\infty) \leq Q_i(0) + (i+1)Q_{i+1}(\infty), \quad 0 \leq i \leq N-1.$$

Starting with $Q_N(\infty) = 1$ (due to Eq. A.12) and using induction from the above equation, we see that $Q_i(\infty)$ is finite for all $i = 0, \dots, N$. ■

Let $\Phi_i(s) = \int_0^\infty Q_i(y)e^{-sy}dy$ be the LT of the function $Q_i(y)$. In this case, the LT's of $Q_N(y)$ and $Q_i(y)$ from Eq.s (A.12) and (A.13) will be

$$\Phi_N(s) = \frac{1}{s}, \quad (\text{A.18})$$

$$\Phi_i(s) = \frac{1}{s}Q_i + (i+1)\frac{\lambda}{s}\Phi_{i+1}(\lambda+s), \quad 0 \leq i \leq N-1. \quad (\text{A.19})$$

Starting from Eq. (A.18) and using the recursive formula in Eq. (A.19), we establish

$$\Phi_i(s) = \sum_{j=i}^N \binom{j}{i} \frac{(j-i)!\lambda^{(j-i)}}{s(\lambda+s)\cdots((j-i)\lambda+s)} Q_j, \quad 0 \leq i \leq N-1. \quad (\text{A.20})$$

Using

$$\frac{k!\lambda^k}{s(\lambda+s)\cdots(k\lambda+s)} = \sum_{j=0}^k (-1)^j \binom{k}{j} \frac{1}{j\lambda+s},$$

Eq. (A.20) can be rewritten as

$$\Phi_i(s) = \sum_{j=i}^N \binom{j}{i} Q_j \sum_{l=0}^{j-i} (-1)^l \binom{j-i}{l} \frac{1}{l\lambda+s}, \quad 0 \leq i \leq N-1. \quad (\text{A.21})$$

Observe that $(l\lambda+s)^{-1}$ on the right hand side of Eq. (A.21) is the LT of $e^{-l\lambda y}$. Using this, when we invert $\Phi_i(s)$, we obtain

$$\begin{aligned} Q_i(y) &= \sum_{j=i}^N \binom{j}{i} Q_j \sum_{l=0}^{j-i} (-1)^l \binom{j-i}{l} e^{-l\lambda y} \\ &= \sum_{j=i}^N \binom{j}{i} Q_j \sum_{l=0}^{j-i} \binom{j-i}{l} (-e^{-\lambda y})^l \\ &= \sum_{j=i}^N \binom{j}{i} Q_j (1 - e^{-\lambda y})^{j-i}, \quad 0 \leq i \leq N-1. \end{aligned} \quad (\text{A.22})$$

Substituting Eq. (A.22) in Eq. (A.17), we have

$$\mathcal{B}_i = \sum_{j=i}^N Q_j \zeta_{i,j}, \quad 0 \leq i \leq N, \quad (\text{A.23})$$

where

$$\zeta_{i,j} = \binom{j}{i} \int_0^\infty (1 - e^{-\lambda y})^{j-i} e^{-i\lambda y} f(y) dy, \quad j \geq i. \quad (\text{A.24})$$

This leads to Eqs. (41) and (42). Together with Eq. (A.23) as defined in Eq. (A.17), Eq. (A.16) gives Eqs. (39) and (40).

We define $\mathcal{D}_i = i\lambda \int_0^\infty (P_{i,1}(y)/P_{N,1}(0))dy$. Noting from Eq. (A.13) that $dQ_i(y) = (i+1)\lambda Q_{i+1}(y)e^{-\lambda y}$, if we rewrite Eq. (A.17) as $\mathcal{B}_i = -\int_0^\infty Q_i(y)e^{-i\lambda y}d\bar{F}(y)$, integration yields

$$\mathcal{B}_i = Q_i(0) + (i+1)\lambda \int_0^\infty Q_{i+1}(y)e^{-(i+1)\lambda y}\bar{F}(y)dy - i\lambda \int_0^\infty Q_i(y)e^{-i\lambda y}\bar{F}(y)dy.$$

Considering Eq. (A.11) for \mathcal{D}_i , the above given equation gives Eqs. (43) and (45). With Eq. (38) and the definitions given in Eqs. (A.11) and (A.16), we obtain Eq. (44). ■

Proof. Corollary 3. Substituting Eq. (A.11) in Eq. (A.22), we arrive at Eq.(46) ■

Proof. Theorem 7. By definition $\sum_{i=0}^N \bar{P}_i = \sum_{i=0}^N (P_{i,0} + \int_0^\infty P_{i,1}(y)dy) = 1$, which by using Eq. (38), becomes $\sum_{i=0}^N (P_{i,0}(0)/\alpha + \int_0^\infty P_{i,1}(y)dy) = 1$. If we divide this equation by $P_{N,1}(0)$, we have

$$P_{N,1}(0) = \left(\frac{1}{\alpha}S_N(0) + \int_0^\infty S_N(y)dy\right)^{-1}, \quad (\text{A.25})$$

where $S_N(y) = \sum_{i=0}^N P_{i,1}(y)/P_{N,1}(0)$.

Summing up Eqs. (36) and (37), we obtain the first order differential equation

$$\frac{d}{dy}S_N(y) = -\beta(y)S_N(y),$$

that has a solution of $S_N(y) = S_N(0)e^{-\int_0^y \beta(x)dx} = S_N(0)\bar{F}(y)$. Substituting this in Eq. (A.25) and using the fact that $E[D] = \int_0^\infty \bar{F}(y)dy$ gives us

$$P_{N,1}(0) = (S_N(0)\left(\frac{1}{\alpha} + E[D]\right))^{-1}$$

Considering Eqs. (A.16) and (A.11), $S_N(0) = \sum_{i=0}^N Q_i$; after substituting it in the equation given above, and using the boundary condition in Eq. (38) we obtain Eq. (47). ■