The Impact of Disruption Characteristics on the Performance of a Server

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 N customers with an unreliable server subject to disruptions even when idle. Times between server interruptions, service times, and times between customer arrivals are assumed to follow exponential distributions. The main contribution of the paper is to use general distributions for the length of server interruption periods/down times. Our numerical analysis reveals the importance of incorporating the down time distribution into the model, since their impact on customer service levels could be counterintuitive. For instance, while higher down time variability increases the mean queue length, for other service levels, can prove to be improving system performance. We also show how the process completion time approach from the literature can be extended to analyze the queueing system if the unreliable server fails only when it is serving a customer.

Keywords and Phrases: Queueing, server disruptions, operation-independent disruptions, operation-dependent disruptions, process completion time

1 Introduction

In this paper, we analyze an $M/M/1//N$ queueing system with an unreliable server. Such kinds of queueing models are popular for modeling telecommunication or computer networks or representing a repair shop. In telecommunication networks, (see, e.g., Sztrik and Gál, 1990), the finite number (N) of potential customers might correspond to active terminals generating jobs for the central processor unit (CPU) which might fail from time to time. In another setting, N can be the number of repairable identical machines fixed by the repair facility upon failure (see, e.g., Liang, Balcıoğlu, Svaluto, 2013). It is possible that the repair facility is not available from time to time, causing the wait times of failed machines in the repair shop to increase. The CPU or the repair facility can be modeled as the single server of the system. We assume that the finitely many customers are served on a first-come, first-served (FCFS) basis. The server can be disrupted from time to time whether it is idle or serving a customer. We define the times between interruptions as the times between the end of an interruption and the beginning of the next interruption. Since these are the times the server would be serving a customer (if any) or ready to serve, we call them the ON periods. We assume that the ON periods, service times and times between customer arrivals are exponentially distributed random variables (r.v.s) independent of one another. In contrast, the lengths of interruptions – the OFF periods during which the single server cannot serve customers – are independent and identically distributed (i.i.d.) general r.v.s (thus, the exponential ON and general i.i.d. OFF periods of the server form an alternating renewal process). General OFF times can naturally arise even when other r.v.s are exponentially distributed. For instance, if there is a higher priority class of customers (whose service and interarrival times are also exponentially distributed) which can preempt the service of a lower priority customer, from the point of view of the latter, the busy period of the higher priority class is an interruption with a phase-type distribution. The queue length distribution of the lower priority class in this specific example can be obtained using the method developed in this study. Thus, incorporating general OFF times in the $M/M/1//N$ queueing model is

the main contribution of the paper.

Note that in our problem, the server can be disrupted/interrupted even when it is idle. Therefore, we study the OID $M/M/1//N$ queue where OID stands for "operation-independent disruptions" indicating that the idle server can be disrupted, too. If we assume that the server can experience disruptions while it serves customers, results from the literature can be used to perform the steady-state analysis. In this case, during the idle periods of the server, the remaining time to disruption freezes until the server becomes busy again. In addition, service times need not be exponential. We discuss this model in Appendix B and refer to it as the ODD $M/G/1//N$ queue where ODD stands for "operation-dependent disruptions" indicating that the server does not experience disruptions when it is idle. Note that we adopt the definitions of OID and ODD from Altıok (1997, p. 85). Since our paper focuses on the OID $M/M/1//N$ queue, for the sake of simplicity, we refer to it as the $M/M/1//N$ queue.

Queueing models with unreliable servers have been widely studied in the literature. We consider a finite customer population queueing system; yet, before positioning our study among relevant work on finite-calling populations, we refer the reader to White and Christie (1958), Gaver (1962), Avi-Itzhak and Naor (1963), Thiruvengadam (1963), Mitrany and Avi-Itzhak (1968) and Neuts and Lucantoni (1979) who model systems with unreliable server(s) attending to infinite customer populations. When it comes to assuming general distributions for the underlying r.v.s in systems with homogeneous Poisson customer arrivals, we make note of the following. 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 same problem, this time assuming that ON periods are phase-type r.v.s. and provide an exact algorithm to obtain the steady-state system performance measures. In the $M/G/1$ queue, Tang (1997) assumes two types of ON periods: exponential ON periods when the server is idle and general i.i.d. ON periods when it is busy. Wang, Cao, and Li (2001) study the $M/G/1$ queue with no waiting space with

exponential operation-independent ON periods and general OFF periods. Atencia, Bouza, and Moreno (2008) consider batch arrivals at an $M/G/1$ retrial queue with no waiting space **and** with operation-dependent ON times. Lam, Zhang and Liu (2006) consider imperfect repairs for $M/M/1$ queues in which the server is subject to random failures (exponential ON periods) followed by exponentially distributed repair times (OFF periods). After each failure, the failure rate increases and the repair rate decreases. Fiems, Maertens, and Bruneel (2008) consider an $M/G/1$ queueing system with disruptive and non-disruptive server interruptions. Interruptions follow a Poisson process with two different rates depending on whether the server is busy or idle. The service restarts (resumes from the moment of interruption) after a disruptive interruption (non-disruptive interruption). Balcıoğlu, 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.

More related to our study are the papers that consider finite-calling populations, part of the *machine interference problem* (MIP) or, alternatively, the *machine repairperson problem* literature on unreliable servers. See Stecke and Aronson (1985) and Haque and Armstrong (2007) for an extensive bibliography on MIP. 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. We also see the application of the $M/M/1//N$ queue in modeling computer networks where N is the number of terminals served under various policies by the single server modeling the CPU that is subject to failures, e.g., Sztrik and Gál (1990), Almasi (1996), and Almasi and Sztrik (1993, 1998a, 1998b, 1999, and 2004). Wang and Kuo (1997) extend the model by Wang (1990) 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 incorporating phase-type distributions to model OFF periods and assuming that operation-independent ON periods are exponentially distributed. For systems with multiple unreliable servers (the $M/M/c//N$ queue), Wang (1993) considers cold-, warm-, and hot-standby spare machines. Assuming that each server is subject to random failure even when it is idle, the number of

servers and spares is optimized. Ke and Wang (1999) incorporate the possibility that a customer may balk and renege to Wang's (1994) model. Wang and Hsu (1995) allow each server to serve at either a slow or a fast rate in Wang's (1994) model. Liu and Cao (1995) model an $M/G/1//N$ system where the single server is composed of r unreliable components. Here, the server can serve only if all of the r components are functional. Operation-dependent ON periods for each component are independent exponential r.v.s while their OFF periods follow general distributions.

In a more recent study, Sahba, Balcıoğlu, and Banjevic (2013) present an $M/G/1//N$ model with general OFF period times. Observe that the model in this paper, due to exponential service time assumptions, is a special case of that by Sahba, Balcıoğlu, and Banjevic (2013). Before clarifying the new contributions made in this paper, based on our literature review we make the following observation. Using non-exponential distributions for underlying r.v.s in these queueing systems is challenging. For instance, especially for systems with finite-calling populations, non-exponential times between arrivals of customers are analytically intractable. Additionally, for systems experiencing operation-dependent and operation-independent server interruptions, non-exponential ON period distribution has not been incorporated in exact analyses (except in $M/G/1$ systems with phase-type ON periods as in Federgruen and Green, 1988 and Balcıoğlu, Jagerman, and Altıok, 2007). Similar difficulties arise for general service time and OFF period distributions. In two papers that are close to our problem, for both r.v.s either Erlang distribution (Wang and Kuo, 1997) or phase-type distributions (Chakravarthy and Agrawal, 2003) have been successfully incorporated. Both studies employ the matrix-analytic method to find the steady-state system size distribution, which can become computationally intensive if the structure of the phase-type distribution is complex. The queueing system we model in Section 2 is restrictive with its exponential service times assumption. Without this model, the general OFF times should be approximated by phase-type distributions and then the model of Chakravarthy and Agrawal (2003) be employed with possible approximation errors and computational difficulty. The model by Sahba, Balcıoğlu, and Banjevic (2013) can be exploited to analyze the problem of

this paper. However, their method derives from a busy period analysis whereas our method is completely different and developed using state transition equations. Our approach – as summarized in the algorithm in Section 3 – is more practical to implement and is significantly more efficient for exponentially distributed service times when compared to their method ($O(N^2)$ vs. $O(N^3)$). As in Katehakis, Smit, and Spieksma (2015a, 2015b) that provide alternative algorithms to obtain the steady-state distributions of quasi birth-and-death processes, algorithmic efficiency in problems similar in nature to the one we study here is crucial. Moreover, Sahba, Balcıoğlu, and Banjevic (2013) do not conduct a numerical analysis, which prevents us from observing the impact of OFF time distribution on the queue length statistics. Our numerical analysis also points out that while an algorithm can be computationally effective, it can suffer from computational inaccuracy problems inherent in coding languages currently available. We also address the ODD $M/G/1//N$ queue.

The rest of the paper is organized as follows. In Section 2, we present the exact steady-state analysis of the OID $M/M/1//N$ queue. The algorithm stemming from this analysis that is used to obtain the system performance measures is presented in Section 3. The numerical examples discussed in Section 4 demonstrate the importance of incorporating the down time distribution. We summarize our conclusions in Section 5. All proofs are included in Appendix A. In Appendix B, we summarize how the results in literature can be used to analyze the ODD $M/G/1//N$ queue.

2 The OID $M/M/1//N$ queue

In this section, we analyze a queueing system with an unreliable single server serving N customers. The times between the completion of a customer's service and the next arrival of the same customer at the queueing system follow an exponential distribution with rate λ. The service times are exponentially distributed with rate μ. Independent of whether there are customers in the system or not, the server is subject to interruptions (e.g., failures) from time to time. In other words, the server is subject to "operation-independent" interruptions

to differentiate it from problems where a server can be interrupted only when it is serving a customer (such as the model in Appendix B). The times between the end of an interruption and the beginning of the next interruption are exponentially distributed r.v.s with rate α. When an interruption occurs, the server becomes unavailable or "down". The lengths of down times (e.g., repair times) are i.i.d. r.v.s following a general continuous distribution with density function $f(y)$ and cumulative distribution function $F(y) = \int_0^y f(u)du$. Letting $\overline{F}(y) = 1 - F(y)$, its first moment will be denoted by $E[D] = \int_0^\infty \overline{F}(y)dy$ and its hazard rate function by $\beta(y) = f(y)/\overline{F}(y)$. According to this, when the server is not down, it is considered to be "up", which means that it is either serving a customer if there are any customers (and the server is considered to be "busy") or it is "idle" and ready to serve. Therefore, at any given time, the server is in one of the following three states: idle, busy or down. If the interruption happens when the server is busy, the customer being served is preempted and resumes service from the point of interruption when the down time is over. Due to the memoryless property of the exponential service times, the remaining service times are also exponentially distributed with rate μ.

It is possible that during the service time of a customer, the server may have none, or one or more interruptions. When the server is down, no additional interruptions can occur. The state of the system at time t is characterized by three stochastic processes: $R(t)$, which equals 0 if the server is up and 1 if it is down; $V(t)$, which is the elapsed time since the server has become down; and $W(t) \in \{0, 1, \ldots, N\}$, which is the number of customers that are not in the queueing system. We employ $W(t)$ instead of the stochastic process that gives the number of customers in the 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. We denote the steady-state probability of having i customers out of queueing system wit $\overline{P}_i$, $i = 0, 1, \ldots, N$. We introduce

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

which is the probability that there are i customers out of the queueing system and the server

is up at time t. Let

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

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 (and with $P_{-1,0}(t) = 0$), we have

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

$$\frac{d}{dt}P_{i,0}(t) = -(i\lambda + \mu + \alpha)P_{i,0}(t) + (i+1)\lambda P_{i+1,0}(t) + \mu P_{i-1,0}(t) + \int_0^\infty P_{i,1}(t,y)\beta(y)dy, \quad 0 \le i \le N-1, \tag{2}$$

and

$$(\frac{\partial}{\partial t} + \frac{\partial}{\partial y} + N\lambda + \beta(y))P_{N,1}(t,y) = 0, \tag{3}$$

$$(\frac{\partial}{\partial t} + \frac{\partial}{\partial y})P_{i,1}(t,y) = -(i\lambda + \beta(y))P_{i,1}(t,y) + (i+1)\lambda P_{i+1,1}(t,y), \ 0 \le i \le N-1. \tag{4}$$

In a stable system, we develop the model in steady-state by letting $P_{i,0} = \lim_{t\to\infty} P_{i,0}(t)$ and $P_{i,1}(y) = \lim_{t\to\infty} P_{i,1}(t,y)$ for $0 \le i \le N$. If we take the limit as $t \to \infty$ in Eqs. (1)-(4) (with $P_{-1,0} = 0$), we obtain

$$(N\lambda + \alpha)P_{N,0} = \mu P_{N-1,0} + \int_0^\infty P_{N,1}(y)\beta(y)dy, \tag{5}$$

$$(i\lambda + \mu + \alpha)P_{i,0} = (i+1)\lambda P_{i+1,0} + \mu P_{i-1,0} + \int_0^\infty P_{i,1}(y)\beta(y)dy, \quad 0 \le i \le N-1, \tag{6}$$

$$\frac{d}{dy}P_{N,1}(y) = -(N\lambda + \beta(y))P_{N,1}(y), \tag{7}$$

$$\frac{d}{dy}P_{i,1}(y) = -(i\lambda + \beta(y))P_{i,1}(y) + (i+1)\lambda P_{i+1,1}(y), \quad 0 \le i \le N-1, \tag{8}$$

and the boundary equations are

$$P_{i,1}(0) = \alpha P_{i,0}, \quad 0 \le i \le N. \tag{9}$$

Letting $\widetilde{f}(s) = \int_0^\infty e^{-sy} f(y)dy$ denote the Laplace transform of the length of server down time r.v., and with $\mathcal{Q}_N = 1$ (see the algorithm in Section 3), we introduce the following to be used in theorems:

$$\mathcal{Q}_{N-1} = \frac{N\lambda + \alpha - \alpha\widetilde{f}(N\lambda)}{\mu}, \tag{10}$$

$$\mathcal{Q}_{i-1} = \frac{(i\lambda + \mu + \alpha)\mathcal{Q}_i - (i+1)\lambda\mathcal{Q}_{i+1} - \alpha\sum_{j=i}^{N}\mathcal{Q}_j\zeta_{i,j}}{\mu}, \quad 1 \le i \le N-1. \tag{11}$$

$$\zeta_{i,i} = \widetilde{f}(i\lambda), \quad 0 \le i \le N, \tag{12}$$

$$\zeta_{i,j} = \frac{j}{j-i}\zeta_{i,j-1} - \frac{i+1}{j-i}\zeta_{i+1,j}, \quad 0 \le i \le N. \tag{13}$$

$$\mathcal{B}_i = \sum_{j=i}^{N}\mathcal{Q}_j\zeta_{i,j}, \quad 0 \le i \le N, \tag{14}$$

$$\mathcal{D}_i = \mathcal{D}_{i+1} - \mathcal{B}_i + \mathcal{Q}_i, \quad 1 \le i \le N-1, \tag{15}$$

with

$$\mathcal{D}_N = 1 - \tilde{f}(N\lambda). \tag{16}$$

We denote the steady-state probability that there are i customers out of the system and the server is down by $P_{i,1}$, the computation of which is given in the following Theorem along with that of $P_{i,0}$.

Theorem 1 *The steady-state probability that there are i customers out of the system is*

$$P_{i,0} = \frac{P_{N,1}(0)}{\alpha}\mathcal{Q}_i, \quad 0 \le i \le N, \tag{17}$$

$$P_{i,1} = \frac{P_{N,1}(0)}{i\lambda}\mathcal{D}_i, \quad 1 \le i \le N. \tag{18}$$

Corollary 1 *The probability density function of having i customers out of the system and an elapsed down time of y is*

$$P_{i,1}(y) = e^{-i\lambda y}\overline{F}(y)P_{N,1}(0)\sum_{j=i}^{N}\binom{j}{i}\mathcal{Q}_j(1-e^{-\lambda y})^{j-i}, \quad 0 \le i \le N. \tag{19}$$

Theorem 2 *The probability that there are no customers in the system when the server is up is*

$$P_{N,0} = \frac{P_{N,1}(0)}{\alpha} = ((1 + \alpha E[D]) \sum_{i=0}^{N} \mathcal{Q}_i)^{-1}. \tag{20}$$

We present the algorithm for computing $P_{i,0}, P_{i,1},\ i = 0, \cdots, N$ in the next section. With these, the probability that the server is down, $P_D = \sum_{i=0}^{N} P_{i,1}$, or serving a customer, $P_B = \sum_{i=1}^{N} P_{i,0}$, can be computed.

3 Algorithm

In this section, we present the algorithm to obtain the steady-state distribution of the number of customers in the queueing system and then, we show its complexity. Most of the notation used in the algorithm is introduced in the preceding section. After defining

$$Q_i(y) = \frac{P_{i,1}(y)}{e^{-i\lambda y}\overline{F}(y)P_{N,1}(0)}, \quad 0 \le i \le N, \tag{21}$$

$$\mathcal{Q}_i = Q_i(0), \quad 0 \le i \le N, \tag{22}$$

where $\overline{F}(y) = e^{-\int_0^y \beta(x)dx}$, next we present the algorithm.

1. **Initialization:**

   - $\mathcal{Q}_N = 1$ (From Eqs. 22 and 21).
   - $\zeta_{N,N} = \tilde{f}(N\lambda)$ (From Eq. 12).
   - $\mathcal{B}_N = \mathcal{Q}_N \zeta_{N,N} = \tilde{f}(N\lambda)$ (From Eq. 14).
   - $\mathcal{D}_N = 1 - \tilde{f}(N\lambda)$ (From Eq. 16).
   - $\mathcal{Q}_{N-1} = (N\lambda + \alpha - \alpha\mathcal{B}_N)/\mu$ (From Eq. 10).
   - Set $S = 1 + \mathcal{Q}_{N-1}$.
   - Set $i = N - 1$.

2. **Big Loop:** While $i > 0$

- $\zeta_{i,i} = \tilde{f}(i\lambda)$ (From Eq. 12).
- Set $j = i$ and $\mathcal{B}_i = 0$.
- **Small Loop:** While $j \leq N$
  - Set $\mathcal{B}_i = \mathcal{B}_i + \mathcal{Q}_j \zeta_{i,j}$ (From Eq. 14).
  - Set $j = j + 1$. Skip next line if $j = N + 1$.
  - $\zeta_{i,j} = \frac{j}{j-i}\zeta_{i,j-1} - \frac{i+1}{j-i}\zeta_{i+1,j}$ (From Eq. 13).
- $\mathcal{D}_i = \mathcal{D}_{i+1} - \mathcal{B}_i + \mathcal{Q}_i$ (From Eq. 15).
- $\mathcal{Q}_{i-1} = \{(i\lambda + \mu + \alpha)\mathcal{Q}_i - (i+1)\lambda\mathcal{Q}_{i+1} - \alpha\mathcal{B}_i\}/\mu$ (From Eq. 11).
- Set $S = S + \mathcal{Q}_{i-1}$.
- Set $i = i - 1$.

3. **Steady-state system size distribution:**

- $P_{N,1}(0) = [(1/\alpha + E[D])S]^{-1}$ (From Theorem 2).
- $P_{i,0} = \mathcal{Q}_i P_{N,1}(0)/\alpha$ for $i = 0, 1, \cdots, N$ (From Eq. 17).
- $P_{i,1} = \mathcal{D}_i P_{N,1}(0)/i\lambda$ for all $i = 1, \cdots, N$ (From Eq. 18).
- $P_{0,1} = 1 - \sum_{i=0}^{N} P_{i,0} - \sum_{i=1}^{N} P_{i,1}$.

Note that the number of operations performed in the initialization part of the algorithm is constant (and equals 7) irrespective of the value of N. In the last part for finding the steady-state system size distribution, the total number of operations performed is $2N$+3. In the Big Loop part, there are 6 operations that are not part of the Small Loop, that are executed $N-1$ times (for $i = N-1$ to – and including – 1). The 3 operations in the Small Loop are executed $((N-1)N/2) - 1$ times (when $i = N-1$, they are executed twice, when $i = N-2$, thrice and when $i = 1$, a total of $N-1$ times, the sum of which gives the result). Thus, the total number of operations performed is $(3N^2$+$19N$+$2)/2$, which is $O(N^2)$.

4 Numerical Experiment

In this section, we consider a single unreliable server that is attending to N customers. We explore how the variability in down times affects two performance measures, namely, (i) $\overline{P}_N = P_{N,0} + P_{N,1}$, which is the probability that all customers are out of the queueing system, and (ii) $E[N^O] = \sum_{i=0}^{N} i(P_{i,0} + P_{i,1})$, the average number of customers out of the system. The analyzed $M/M/1//N$ queueing system may represent a repair shop (where the repair crew or resources can become unavailable from time to time), and its customers may be broken machines. Then, N is the number of machines that a client company can be sending upon failure to be fixed at the repair shop. If the repair shop is a profit center, it would like to serve more customers (higher N). In return, the client may require $\overline{P}_N$, i.e., the proportion of time all N machines are functional (out of the repair shop) to be high. In other words, higher $\overline{P}_N$ indicates a higher service level provided to the client. Therefore, we obtain the maximum number of customers/machines (N) the server can serve while keeping $\overline{P}_N$ above certain targeted levels.

Note that $E[N^O]$ can be used as a secondary service level measure. It may even be a more important service level than $\overline{P}_N$; for example, if customers represent production plant machines that fail from time to time and are repaired by the unreliable server in our model, higher $E[N^O]$ (higher expected number of operational machines at the production plant) corresponds to a higher production rate.

In all the examples, the server becomes unavailable from time to time at rate $\alpha = 0.05$. We choose the distribution of the down time r.v. D from the following four distributions, each having a mean of $E[D] = 2$ and a squared-coefficient of variation (variance to squared-mean ratio) denoted by c_D^2:

- H2($a = 0.9, \gamma_1 = 10, \gamma_2 = 0.05236$) with $c_D^2 = 17.245$, and density function

$$f(x) = a\gamma_1 e^{-\gamma_1 x} + (1-a)\gamma_2 e^{-\gamma_2 x}.$$

- Gamma($\gamma = 0.1, k = 0.2$) with $c_D^2 = 5$, and density function

$$f(x) = \frac{\gamma(\gamma x)^{k-1} e^{-\gamma x}}{\Gamma(k)}.$$

- Exponential($\gamma = 0.5$) with $c_D^2 = 1$, and density function

$$f(x) = \gamma e^{-\gamma x}.$$

- and Erlang($\gamma = 2.5, k = 5$) with $c_D^2 = 0.2$ which is equivalent to Gamma$(2.5, 5)$.

We used Matlab as the implementation medium of the algorithm provided in Section 3. We determine the maximum N that the unreliable server can serve given a target level for $\overline{P}_N$, the proportion of time all N customers are out of the system. We consider $\lambda \in \{0.01, 0.05\}$ and increment N until the targeted value for $\overline{P}_N$ cannot be attained. Before presenting the results, we note that for small λ (such as 0.01), the algorithm is posed with computational inaccuracy problems that are inherent in coding languages available at the moment. As N is incremented, the algorithm starts yielding very small, yet negative, values for $\overline{P}_i$ for i close to 0 which are effectively zero as estimated by the simulation. In Table 1, this is observed beyond $N = 30, 25, 13, 12$ for H2, Gamma, Exponential, and Erlang down time distributions, respectively. When we simply assume these quite small negative probabilities to be 0, $E[N^O]$ obtained does not differ from the simulated estimates up to $N = 65, 65, 70, 65$ for H2, Gamma, Exponential, and Erlang down time distributions, respectively. If we continue incrementing N further, the round-off errors accumulate and the analytical results start deviating from those obtained via simulation. Thus, in Table 1, the values presented in bold face in the last column are obtained from the simulation.

In Table 1, we see that as the variability in down time decreases, the number of customers that can be served given a service level on $\overline{P}_N$ tends to increase. The only exception is when $\overline{P}_N \geq 0.2$ where H2 down time with the highest c_D^2 results in a higher N value than cases with less variable down time r.v.s.

In Table 2, we increase λ (compared to λ in Table 1), corresponding to shorter mean time of staying out of the queueing system for each customer (shorter mean time to failure

of machines). As expected, this has an adverse affect on the maximum N that can be served, given the service level constraints on $\overline{P}_N$. In this case, our algorithm does not suffer from computational inaccuracy problems and functions correctly up to $N = 360, 360, 355, 355$ for H2, Gamma, Exponential, and Erlang down time distributions, respectively.

Table 1: The maximum N that can be served when down time distribution changes and $\lambda = 0.01$.

Down Time Distribution	$\overline{P}_N \geq 0.9$	$\overline{P}_N \geq 0.8$	$\overline{P}_N \geq 0.7$	$\overline{P}_N \geq 0.5$	$\overline{P}_N \geq 0.2$
H2	5	13	23	43	**76**
Gamma	7	15	24	43	**75**
Exponential	8	16	25	44	**75**
Erlang	8	17	26	44	**75**

Table 2: The maximum N that can be served when down time distribution changes and $\lambda = 0.05$.

Down Time Distribution	$\overline{P}_N \geq 0.9$	$\overline{P}_N \geq 0.8$	$\overline{P}_N \geq 0.7$	$\overline{P}_N \geq 0.5$	$\overline{P}_N \geq 0.2$
H2	1	2	4	9	17
Gamma	1	3	5	9	17
Exponential	1	3	5	9	17
Erlang	1	3	5	9	17

For the examples presented in Table 1, we compute the mean number of customers out of the system and present them in Table 3. If we compare Tables 1 and 3, we see that when N is the same, lower down time variance increases $E[N^O]$ (i.e., shortens the mean queue length). Here the values presented in bold face in the last column are the mean

Table 3: Mean number of customers out of the system ($E[N^O]$) when down time distribution changes and $\lambda = 0.01$.

Down Time Distribution	$\overline{P}_N \geq 0.9$	$\overline{P}_N \geq 0.8$	$\overline{P}_N \geq 0.7$	$\overline{P}_N \geq 0.5$	$\overline{P}_N \geq 0.2$
H2	4.8730	12.6439	22.3015	41.5114	**70.6977**
Gamma	6.8819	14.7231	23.5053	41.8573	**70.987**
Exponential	7.891	15.7603	24.579	43.002	**71.4265**
Erlang	7.8969	16.7558	25.5789	43.0463	**71.5609**

$E[N^O]$ estimated from simulation for which the commercial software ARENA is used. The estimates are obtained from 20 replications each of which has 10,000 units of time as the warm-up period and 1,000,000 time units of time as the replication length. This leads to 0.01 as the 95% confidence interval half-widths for the estimated $E[N^O]$. Table 4 presents the mean number of customers out of the system for the cases presented in Table 2. When the two tables are compared, in cases with the same N, we again see that higher variability in down time increases the mean queue length in the system (reducing the mean number of customers out of the queueing system).

Table 4: Mean number of customers out of the system ($E[N^O]$) when down time distribution changes and $\lambda = 0.05$.

Down Time Distribution	$\overline{P}_N \geq 0.9$	$\overline{P}_N \geq 0.8$	$\overline{P}_N \geq 0.7$	$\overline{P}_N \geq 0.5$	$\overline{P}_N \geq 0.2$
H2	0.9097	1.8135	3.6012	7.9105	13.8914
Gamma	0.9298	2.7709	4.5823	8.0797	14.1217
Exponential	0.9404	2.8037	4.6379	8.1818	14.2820
Erlang	0.9432	2.8124	4.6531	8.2114	14.3355

5 Conclusions

In this paper, we provide a method to obtain the exact steady-state performance measures of the $M/M/1//N$ queue with an unreliable server in which the exponential ON and general i.i.d. OFF periods of the server form an alternating renewal process. Our contribution is to incorporate general OFF period distributions in this classical queueing model from the literature. The algorithm stemming from our model is easy to implement and provides an alternative to computationally difficult algorithms designed for cases with phase-type OFF period distributions. Yet, the practitioner should be cautious that with the computational accuracy levels of the available coding languages, the number of customers (N) the developed algorithm can handle may be limited for certain parameters. Thus, progress to increase computational accuracy should be considered and reflected in this and similar type of problems. Including non-exponential distributions to model times between customer arrivals and/or times between interruptions remains challenging, however. While our major contribution is for the problem in which the server can fail whether or not it has customers, we also demonstrate how the process completion time approach (used in the literature for the $M/G/1$ queue with an unreliable server) can be used for studying the $M/G/1//N$ queue with an unreliable server.

Acknowledgements

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

References

Almasi, B. 1996. "Response time for finite heterogeneous nonreliable queueing systems", *Computers and Mathematics with Applications*, Vol. 31, No. 11, 55–59.

Almasi, B. and J. Sztrik. 1993. "A queueing model for a non-homogeneous terminal system subject to BR992", *Computers and Mathematics with Applications*, Vol. 25, No. 4, 105-111.

Almasi, B. and J. Sztrik. 1998a. "A queueing model for a nonreliable multiterminal system with polling scheduling", *Journal of Mathematical Sciences*, Vol. 92, No. 4, 3974–3981.

Almasi, B. and J. Sztrik. 1998b. "The effects of service disciplines on the performance of a nonreliable terminal system", *Journal of Mathematical Sciences*, Vol. 92, No. 4, 3982–3989.

Almasi, B. and J. Sztrik. 1999. "Optimization problems on the performance of a nonreliable terminal system", *Computers and Mathematics with Applications*, Vol. 38, No. 3, 13–21.

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.

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.

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.

Federgruen, A. and L. Green. 1986. "Queueing systems with service interruptions", *Opera-*

tions 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.

Katehakis, M. N., L. C. Smit, and F. M. Spieksma. 2015a. "DES and RES processes and their explicit solutions", *Probability in the Engineering and Informational Sciences*, Vol. 29, 191–217.

Katehakis, M. N., L. C. Smit, and F. M. Spieksma. 2015b. "A comparative analysis of the successive lumping and the lattice path counting algorithms", *Journal of Applied Probability*, *forthcoming*.

Ke J.-C and K.-H Wang. 1999. "Cost analysis of the $M/M/R$ machine repair problem with balking, reneging, and server breakdowns", *Journal of the Operational Research Society*, Vol. 50, No. 2-3, 275–282.

Lam, Y., Y. L. Zhang and Q. Liu. 2006. "A geometric process model for $M/M/1$ queueing system with a repairable service station", *European Journal of Operational Research*, Vol. 168, No. 1, 100–121.

Liang, W. K., B. Balcıoğlu, R. Svaluto. 2013. "Scheduling policies for a repair shop problem", *Annals of Operations Research*, Vol. 211, 273–288.

Liu, B. and J. Cao. 1995. "A machine service model with a service station consisting of r unreliable units", *Microelectronics Reliability*, Vol. 35, No. 4, 683–690.

Mitrany, I. L. and B. Avi-Itzhak. 1968. "A many server queue with service interruptions", *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.

Sahba, P., B. Balcıoğlu, and D. Banjevic. 2013. "Analysis of the Finite-source Multi-class Priority Queue with an Unreliable Server and Setup Time," *Naval Research Logistics*, Vol. 60, 331–342.

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.

Tang, Y. H. 1997. "A single-server $M/G/1$ queueing system subject to breakdowns - Some reliability and queueing problems", *Microelectronics Reliability*, Vol. 37, No. 2, 315–321.

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. 1993. "Cost analysis of the $M/M/R$ machine-repair problem with mixed standby spares", *Microelectronics Reliability*, Vol. 33, No. 9, 1293–1301.

Wang , K.-H and L. Y. Hsu. 1995. "Cost analysis of the machine-repair problem with R non-reliable service stations", *Microelectronics Reliability*, Vol. 35, No. 6, 923–934.

Wang, K.-H and M.-Y. Kuo. 1997. "Profit analysis of the $M/Ek/1$ 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", *Operations Research*, Vol. 6, No. 1, 79–95.

Appendix A Proofs

Proof. Theorem 1. If we divide both sides of Eq. (7) by $e^{-N\lambda y-\int_0^y \beta(x)dx}P_{N,1}(0)$ and Eq. (8) by $e^{-i\lambda y-\int_0^y \beta(x)dx}P_{N,1}(0)$, we get

$$\frac{d}{dy}\left(\frac{e^{N\lambda y+\int_0^y \beta(x)dx}P_{N,1}(y)}{P_{N,1}(0)}\right) = 0, \tag{A.23}$$

$$\frac{d}{dy}\left(\frac{e^{i\lambda y+\int_0^y \beta(x)dx}P_{i,1}(y)}{P_{N,1}(0)}\right) = \frac{(i+1)\lambda e^{i\lambda y+\int_0^y \beta(x)dx}P_{i+1,1}(y)}{P_{N,1}(0)}, 0 \le i \le N-1, \tag{A.24}$$

which are first order differential equations. We solve Eqs. (A.23) and (A.24) using Eq. (21) as

$$Q_N(y) = 1, \tag{A.25}$$

$$Q_i(y) = Q_i(0) + (i+1)\lambda \int_0^y Q_{i+1}(x)e^{-\lambda x}dx, \quad 0 \le i \le N-1. \tag{A.26}$$

Considering the definition given in Eq. (21), and employing Eqs. (5), (6) and (9), we obtain

$$Q_{N-1}(0) = \frac{N\lambda + \alpha - \alpha\int_0^\infty Q_N(y)e^{-N\lambda y}f(y)dy}{\mu}, \tag{A.27}$$

$$Q_{i-1}(0) = \frac{(i\lambda+\mu+\alpha)Q_i(0) - (i+1)\lambda Q_{i+1}(0) - \alpha\int_0^\infty Q_i(y)e^{-i\lambda y}f(y)dy}{\mu}, 1 \le i \le N-1. \tag{A.28}$$

For simplicity, we define

$$\mathcal{B}_i = \int_0^\infty Q_i(y)e^{-i\lambda y}f(y)dy. \tag{A.29}$$

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

Lemma 1 $\lim_{y\to\infty} Q_i(y) = Q_i(\infty)$ *exists and is finite. We also have* $Q_i(y) \le Q_i(\infty)$.

Proof. Lemma 1. From Eq. (21), $Q_i(y) \geq 0$ and from Eq. (A.26), we see that $Q_i(y)$ is increasing in y. Let $Q_i(\infty) = \lim_{y\to\infty} Q_i(y)$. Then, $Q_i(y) \leq Q_i(\infty),\ 0 \leq i \leq N-1$. If we take the limit as $y \to \infty$ in Eq. (A.26),

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

Starting with $Q_N(\infty) = 1$ (due to Eq. A.25) and using induction from the above equation, we see that $Q_i(\infty)$ is finite for all $i = 0, \cdots, N$. ■

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 Eqs. (A.25) and (A.26) will be

$$\Phi_N(s) = \frac{1}{s}, \tag{A.30}$$

$$\Phi_i(s) = \frac{1}{s}\mathcal{Q}_i + (i+1)\frac{\lambda}{s}\Phi_{i+1}(\lambda+s), \quad 0 \leq i \leq N-1. \tag{A.31}$$

Starting from Eq. (A.30) and using the recursive formula in Eq. (A.31), we establish

$$\Phi_i(s) = \sum_{j=i}^{N} \binom{j}{i} \frac{(j-i)!\lambda^{(j-i)}}{s(\lambda+s)\cdots((j-i)\lambda+s)}\mathcal{Q}_j, \quad 0 \leq i \leq N-1. \tag{A.32}$$

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.32) can be rewritten as

$$\Phi_i(s) = \sum_{j=i}^{N}\binom{j}{i}\mathcal{Q}_j\sum_{l=0}^{j-i}(-1)^l\binom{j-i}{l}\frac{1}{l\lambda+s}, \quad 0 \leq i \leq N-1. \tag{A.33}$$

Observe that $(l\lambda+s)^{-1}$ on the right hand side of Eq. (A.33) is the LT of $e^{-l\lambda y}$. Using this, when we invert $\Phi_i(s)$, we obtain

$$\begin{aligned} Q_i(y) &= \sum_{j=i}^{N}\binom{j}{i}\mathcal{Q}_j\sum_{l=0}^{j-i}(-1)^l\binom{j-i}{l}e^{-l\lambda y} \\ &= \sum_{j=i}^{N}\binom{j}{i}\mathcal{Q}_j\sum_{l=0}^{j-i}\binom{j-i}{l}(-e^{-\lambda y})^l \\ &= \sum_{j=i}^{N}\binom{j}{i}\mathcal{Q}_j(1-e^{-\lambda y})^{j-i}, \quad 0 \leq i \leq N-1. \end{aligned} \tag{A.34}$$

Substituting Eq. (A.34) in Eq. (A.29), we have Eq. (14) where

$$\zeta_{i,j} = \binom{j}{i} \int_0^\infty (1 - e^{-\lambda y})^{j-i} e^{-i\lambda y} f(y) dy, \quad j \geq i. \tag{A.35}$$

This leads to Eqs. (12) and (13). Together with Eq. (14) as defined in Eq. (A.29), Eq. (22) gives Eqs. (10) and (11).

We define $\mathcal{D}_i = i\lambda \int_0^\infty (P_{i,1}(y)/P_{N,1}(0))dy$. Noting from Eq. (A.26) that $dQ_i(y) = (i+1)\lambda Q_{i+1}(y)e^{-\lambda y}$, if we rewrite Eq. (A.29) as $\mathcal{B}_i = -\int_0^\infty Q_i(y)e^{-i\lambda y} d\overline{F}(y)$, integration yields

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

Considering Eq. (21) for $\mathcal{D}_i$, the above given equation gives Eqs. (15) and (18). With Eq. (9) and the definitions given in Eqs. (21) and (22), we obtain Eq. (17). ■

Proof. Corollary 1. Substituting Eq. (21) in Eq. (A.34), we arrive at Eq.(19) ■

Proof. Theorem 2. By definition $\sum_{i=0}^N \overline{P}_i = \sum_{i=0}^N (P_{i,0} + \int_0^\infty P_{i,1}(y)dy) = 1$, which by using Eq. (9), 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

$$P_{N,1}(0) = (\frac{1}{\alpha} S_N(0) + \int_0^\infty S_N(y)dy)^{-1}, \tag{A.36}$$

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

Summing up Eqs. (7) and (8), we obtain the first order differential equation

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

that has a solution of $S_N(y) = S_N(0)e^{-\int_0^y \beta(x)dx} = S_N(0)\overline{F}(y)$. Substituting this in Eq. (A.36) and using the fact that $E[D] = \int_0^\infty \overline{F}(y)dy$ gives us

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

Considering Eqs. (22) and (21), $S_N(0) = \sum_{i=0}^N \mathcal{Q}_i$; after substituting it in the equation given above, and using the boundary condition in Eq. (9) we obtain Eq. (20). ■

Appendix B The ODD $M/G/1//N$ queue

In this section, we summarize how results from the literature can be easily used in analyzing the $M/G/1//N$ queue where unlike the model studied in this paper, the server can experience disruptions only if it has customers. We make the same assumptions and employ the same notations introduced in Section 2 for the underlying r.v.s with two differences. First, the actual service time of a job – in the absence of disruptions – is a general i.i.d. r.v. with the LT of $\tilde{b}(s)$. Second, while α still denotes the rate of the exponential times to interruption, the interruption process is halted when the server becomes idle until it becomes busy again. To handle this problem, one can use the process completion time (PCT) r.v. (Gaver, 1962) that is the total time a customer spends on the server including its actual service time plus possible OFF periods it may experience. Let C denote the PCT r.v. the LT of which is given by (e.g., Altıok, 1997, p. 94)

$$\tilde{c}(s) = \tilde{b}(s + \alpha - \alpha\tilde{f}(s)), \tag{B.37}$$

where $\tilde{f}(s)$ is the LT of the OFF periods.

The PCT r.v., C, includes all the information of ON and OFF periods. We can use it as the service time r.v. in an $M/G/1//N$ queue without interruptions analyzed by Gupta and Srinivasa Rao (1996), which will be referred to as the $M/PCT/1//N$ queue. Note that the $M/PCT/1//N$ queue has the same λ and N as in the original ODD $M/G/1//N$ queue, and additionally uses $\tilde{c}(s)$ from Eq. (B.37) as the LT of the service time. Using the algorithm by Gupta and Srinivasa Rao, one obtains P_i in the $M/PCT/1//N$ queue, which coincides with the probability of having i customers in the original ODD $M/G/1//N$ queue. With these probabilities, the expected system size and the probability of the server is idle in the ODD $M/G/1//N$ queue can be computed. The limitation of using the PCT approach is that we are unable to find the probability that the server is down or serving a customer, P_D and P_B, respectively.