Research Article

M/G/1 Queues with Postponed Interruptions

Hans (J. P. C.) Blanc

*Department of Econometrics and Operations Research, Tilburg University, P.O. Box 90153,
5000 LE Tilburg, The Netherlands*

Correspondence should be addressed to Hans (J. P. C.) Blanc, blanc@uvt.nl

Received 25 June 2012; Accepted 21 August 2012

Academic Editors: C. Proppe and L. Sacerdote

Copyright © 2012 Hans (J. P. C.) Blanc. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper studies the effect of postponing interruptions of the service in M/G/1 queueing systems until the end of the service in progress. The waiting and sojourn times in systems with postponement are compared with those in comparable systems without post-ponement, both for the case that interruptions can only occur during busy times of the server and for the case that interruptions are also possible during idle times of the server.

1. Introduction

Kelton et al. [1, Sc. 4.2.3] propose the following rule of thumb for dealing with failures of machines in the context of a simulation model. If the expected time to failure is large compared to the expected failure duration, postpone the failure until the end of the service in progress. They apply this rule to a machine which forms a node in a queueing network. The present paper compares the effect of this postponement against failures which preempt the service in progress on the time in system of the customers. For the simulation model in Kelton et al. [1, Sc. 4.2.3] the mean time at the node with the machine with failures is 8.65 ± 0.07 with postponement versus 8.81 ± 0.07 with preemption (95% confidence intervals from 100 replications of simulations with common random numbers of 500 days after 1 day warm up). A 95% paired-t confidence interval, cf. Kelton et al. [1, Sc. 6.4], of the difference between the time at the node with preemption minus that with postponement is 0.155 ± 0.046 , so the difference is significant at the 95% level. This means that the model with postponed failures underestimates the model with preemption by about 1.8%, partly because there will be somewhat (about 1%) less failures in the former model due to the postponement.

In the rest of this paper, we investigate the difference between postponement and preemption more systematically for M/G/1 queueing systems in which services may be interrupted because the server becomes unavailable during some time. For instance, the

server could be a machine that is subject to breakdowns, a clerk whose servicing is interrupted by phone calls, or a doctor who is called away from consultations for more urgent matters. The arrivals are generated by a Poisson process at rate λ . The service times are assumed to be a sequence of independent, identically distributed random variables with distribution $B(y)$, with moments $\beta_k, k = 1, 2, \dots$, and Laplace-Stieltjes transform (LST) $\beta(\zeta) \doteq \int_0^\infty e^{-\zeta y} dB(y)$. The interruptions or breakdowns are generated by an (interrupted) Poisson process at rate γ . The interruption or repair times are assumed to be a sequence of independent, identically distributed random variables with distribution $T(y)$, with moments $\theta_k, k = 1, 2, \dots$, and LST $\theta(\zeta) \doteq \int_0^\infty e^{-\zeta y} dT(y)$. Throughout it will be assumed that a preempted service is resumed when the service interruption is over. M/G/1 systems with service interruptions have been discussed by Gaver [2], Keilson [3], and Cohen [4, Sc. III.3]. More recent work on queueing systems with service interruptions (without postponement) includes Fiems et al. [5], who consider a mixture of preemptive resume and preemptive repeat interruptions, and Krishnamoorthy et al. [6], who consider a queue with interruptions in a matrix-analytic setting.

The organization of the rest of this paper is as follows. The imbedded Markov chain analysis for M/G/1 systems is shortly reviewed in Section 2. Section 3 considers the effect of postponing interruptions for the case that interruptions can only occur during active periods of the server. Section 4 contains a discussion of the same effect but for the case that interruptions can also occur when the server is idle. A conclusion can be found in Section 5.

2. Review of M/G/1 Analysis

Consider the standard M/G/1 system without interruptions. The analysis of this system is often based on the imbedded Markov chain of the number of customers present just after service completions, cf. for example, Cohen [4, Sc. II.4.2] and Gross and Harris [7, Sc. 5.1]. Let X_k denote the number of customers present just after the service completion of the k th customer, not including this k th customer. This sequence of random variables satisfies the following well-known recursion:

$$X_k = \max\{0, X_{k-1} - 1\} + A(B_k), \quad k = 1, 2, \dots \quad (2.1)$$

Here, $A(B_k)$ represents the number of arrivals during the service time B_k of the k th customer. Since the service times are assumed to be mutually independent, and since the arrivals occur according to a Poisson process, the random variables $A(B_k)$ are also mutually independent. Their probability generating function (PGF) is $\beta(\lambda[1 - z])$, $|z| \leq 1$. Recursion (2.1) leads to the following relation for the PGF of the stationary version X of the sequence X_k :

$$E\{z^X\} = \frac{1}{z} \beta(\lambda[1 - z]) [E\{z^X\} - (1 - z)\Pr\{X = 0\}], \quad |z| \leq 1. \quad (2.2)$$

Since $\Pr\{X = 0\} = 1 - \rho$, an M/G/1 system is stable if $\rho < 1$, and the PGF of the distribution of X follows as

$$E\{z^X\} = \frac{(1 - \rho)(1 - z)\beta(\lambda[1 - z])}{\beta(\lambda[1 - z]) - z}, \quad |z| \leq 1. \quad (2.3)$$

Next, let Q_k denote the number of customers waiting in the queue just after the start of the service of the k th customer, not including this k th customer. If $X_{k-1} > 0$, the service of the k th customer starts immediately after the departure of customer $k-1$ with X_{k-1} customers present in the system, so $Q_k = X_{k-1} - 1$ customers are waiting in the queue. If $X_{k-1} = 0$, the service of the k th customer starts after an idle period of the server with no customers waiting. Hence, these random variables are related by

$$Q_k = \max\{0, X_{k-1} - 1\}, \quad k = 1, 2, \dots \quad (2.4)$$

With (2.1) and (2.3), it readily follows that the stationary version Q of the sequence Q_k has a distribution with PGF

$$E\{z^Q\} = \frac{(1-\rho)(1-z)}{\beta(\lambda[1-z]) - z}, \quad |z| \leq 1. \quad (2.5)$$

Since there is a single server, the number of customers waiting in the queue, Q_k , must be equal to the number of arrivals during the first-come-first-served (FCFS) waiting time W_k of the k th customer. For a Poisson arrival process, the number of arrivals during W_k , $A(W_k)$, only depends on the length of W_k and the arrival rate λ : $A(W_k) = Q_k$. This implies for the stationary versions of these sequences that

$$E\{e^{-\lambda(1-z)W}\} = E\{z^Q\}, \quad |z| \leq 1. \quad (2.6)$$

This leads with (2.5) to the LST of the stationary FCFS waiting-time distribution

$$E\{e^{-\zeta W}\} = \frac{(1-\rho)\zeta}{\zeta - \lambda + \lambda\beta(\zeta)}, \quad \Re\zeta \geq 0. \quad (2.7)$$

3. Interruptions during Busy Periods Only

In this section it will be assumed that interruptions can only occur when the server is active. This means that the instants generated by the Poisson process with failure rate γ are only interpreted as a server interruption if the server is occupied with the service of a customer; they are ignored during idle periods and during interruption times. The analysis of this variant of the M/G/1 system can be reduced to that of a standard M/G/1 system with a modified service-time distribution. The total duration $\tilde{B}$ of a service, called the service-completion time, cf. Gaver [2], may include one or more interruption times. The number of interruptions during a service, $I(B)$, only depends on the length of the service time, and not on those of the interruption times. Given the service time $B = u$ and the number of interruptions $I(B) = k$, the total duration $\tilde{B}$ is the sum of u and of k independent interruption times, that is, for $k = 0, 1, 2, \dots$, $u > 0$,

$$E\{e^{-\zeta\tilde{B}} \mid I(B) = k, B = u\} = e^{-\zeta u} \theta^k(\zeta), \quad \Re\zeta \geq 0. \quad (3.1)$$

Because the number of interruptions during a service time of duration $B = u$ is Poisson-distributed with mean γu , it follows that

$$\tilde{\beta}(\zeta) \doteq E\left\{e^{-\zeta \tilde{B}}\right\} = \beta(\zeta + \gamma[1 - \theta(\zeta)]), \quad \Re \zeta \geq 0. \quad (3.2)$$

The first three moments of the distribution of the service completion time, including interruptions, are

$$\begin{aligned} \tilde{\beta}_1 &= \beta_1(1 + \gamma\theta_1), & \tilde{\beta}_2 &= \beta_2(1 + \gamma\theta_1)^2 + \gamma\beta_1\theta_2, \\ \tilde{\beta}_3 &= \beta_3(1 + \gamma\theta_1)^3 + 3\beta_2\gamma\theta_2(1 + \gamma\theta_1) + \beta_1\gamma\theta_3. \end{aligned} \quad (3.3)$$

Results for the standard M/G/1 system, cf. Section 2, are translated to the present variant with interruptions by replacing the LST $\beta(\zeta)$ and related quantities throughout by $\tilde{\beta}(\zeta)$ and corresponding quantities. For instance, the M/G/1 system with interruptions is stable if the modified load of the system, $\tilde{\rho} \doteq \lambda\tilde{\beta}_1$, is bounded by one so that

$$\tilde{\rho} \doteq \lambda\tilde{\beta}_1 = \lambda\beta_1(1 + \gamma\theta_1) < 1, \quad \text{or} \quad \rho = \lambda\beta_1 < \frac{1}{1 + \gamma\theta_1}. \quad (3.4)$$

The latter inequality states that the offered load ρ must be smaller than the probability that the server is available for service when interruptions are possible at any instant. Further, let W_F denote the waiting-time until the start of service under FCFS. The LST of this waiting time distribution becomes, cf. (2.7), as follows:

$$E\left\{e^{-\zeta W_F}\right\} = \frac{(1 - \tilde{\rho})\zeta}{\zeta - \lambda + \lambda\tilde{\beta}(\zeta)}, \quad \Re \zeta \geq 0. \quad (3.5)$$

The probability that the waiting time is zero is

$$\Pr\{W_F = 0\} = 1 - \tilde{\rho} = 1 - \lambda\beta_1(1 + \gamma\theta_1). \quad (3.6)$$

The mean waiting time until the start of service is

$$E\{W_F\} = \frac{\tilde{\rho}}{1 - \tilde{\rho}} \frac{\tilde{\beta}_2}{2\tilde{\beta}_1} = \frac{\rho(1 + \gamma\theta_1)}{1 - \rho(1 + \gamma\theta_1)} \left[(1 + \gamma\theta_1) \frac{\beta_2}{2\beta_1} + \frac{\gamma\theta_1}{1 + \gamma\theta_1} \frac{\theta_2}{2\theta_1} \right] \quad (3.7)$$

and the variance of the waiting times until the start of service is

$$\sigma^2\{W_F\} = \frac{\tilde{\rho}}{1 - \tilde{\rho}} \frac{\tilde{\beta}_3}{3\tilde{\beta}_1} + \left[\frac{\tilde{\rho}}{1 - \tilde{\rho}} \frac{\tilde{\beta}_2}{2\tilde{\beta}_1} \right]^2. \quad (3.8)$$

The time in system S_F of a customer is the independent sum of the waiting time until the start of service W_F and a service completion time $\tilde{B}$, so that

$$E\{S_F\} = E\{W_F\} + \tilde{\beta}_1, \quad \sigma^2\{S_F\} = \sigma^2\{W_F\} + \tilde{\beta}_2 - \tilde{\beta}_1^2. \quad (3.9)$$

The fraction of time the server is busy is equal to the offered load ρ and the fraction of time the server is idle is by the PASTA (Poisson Arrivals See Time Averages) property equal to $\Pr\{W_F = 0\}$, cf. (3.6). Hence, the fraction of time F_F the server is interrupted is

$$F_F = 1 - \rho - \Pr\{W_F = 0\} = \tilde{\rho} - \rho = \lambda\beta_1\gamma\theta_1. \quad (3.10)$$

Next, it will be assumed that interruptions can be postponed until the end of a service. Let X_k denote the number of customers left behind in the system at the end of the k th service, including a possible interruption. Then, this sequence of random variables forms an imbedded Markov chain and satisfies the recursion, cf. (2.1), as follows:

$$X_k = \max\{0, X_{k-1} - 1\} + A(B_k^+), \quad k = 1, 2, \dots \quad (3.11)$$

Here, the random variable B_k^+ represents the length of a service time and a possible interruption. With B the service time and C the first interruption occurrence after the start of this service, we have for the distribution of such a service completion time

$$\begin{aligned} E\{e^{-\zeta B^+}\} &= E\{e^{-\zeta B}(C > B)\} + E\{e^{-\zeta(B+C)}(C < B)\} \\ &= \int_0^\infty e^{-\zeta u} e^{-\gamma u} dB(u) + \theta(\zeta) \int_0^\infty e^{-\zeta u} (1 - e^{-\gamma u}) dB(u), \quad \Re \zeta \geq 0, \end{aligned} \quad (3.12)$$

so that

$$\beta^+(\zeta) \doteq E\{e^{-\zeta B^+}\} = \beta(\zeta + \gamma) + \theta(\zeta)[\beta(\zeta) - \beta(\zeta + \gamma)], \quad \Re \zeta \geq 0. \quad (3.13)$$

The first three moments of the distribution of the service completion time, including a possible, postponed interruption, are

$$\begin{aligned} \beta_1^+ &= \beta_1 + \theta_1[1 - \beta(\gamma)], & \beta_2^+ &= \beta_2 + 2\theta_1[\beta_1 + \beta'(\gamma)] + \theta_2[1 - \beta(\gamma)], \\ \beta_3^+ &= \beta_3 + 3\theta_1[\beta_2 - \beta''(\gamma)] + 3\theta_2[\beta_1 + \beta'(\gamma)] + \theta_3[1 - \beta(\gamma)]. \end{aligned} \quad (3.14)$$

This system is stable if

$$\rho^+ \doteq \lambda\beta_1^+ = \lambda(\beta_1 + \theta_1[1 - \beta(\gamma)]) < 1, \quad \text{or} \quad \rho = \lambda\beta_1 < \frac{1}{1 + \theta_1[1 - \beta(\gamma)]/\beta_1}. \quad (3.15)$$

From (3.11), it follows that the stationary distribution of the sequence X_k has a similar PGF as that of the standard M/G/1 system, cf. (2.3), but with $\beta(\zeta)$ replaced by $\beta^+(\zeta)$. Further,

although X_k is not necessarily related to a departure instant in the present system, relation (2.4) does hold for the present system, with Q_k the number of waiting customers at the start of the service of customer k . Hence, the stationary distribution of the sequence Q_k also has a similar PGF as that of the standard M/G/1 system, cf. (2.5), but with $\beta(\zeta)$ replaced by $\beta^+(\zeta)$. Let W_{FP} denote the waiting-time until the start of service under FCFS for this system with postponed interruptions. Since $A(W_{FP}) = Q_{FP}$ also holds for the present system, the LST of this waiting time distribution becomes that of the standard M/G/1 system (2.7) with $\beta(\zeta)$ replaced by $\beta^+(\zeta)$ as follows:

$$E\{e^{-\zeta W_{FP}}\} = \frac{(1 - \rho^+) \zeta}{\zeta - \lambda + \lambda \beta^+(\zeta)}, \quad \Re \zeta \geq 0. \quad (3.16)$$

The probability that the waiting time is zero is

$$\Pr\{W_{FP} = 0\} = 1 - \rho^+ = 1 - \lambda(\beta_1 + \theta_1[1 - \beta(\gamma)]). \quad (3.17)$$

The mean waiting time until the start of service is

$$E\{W_{FP}\} = \frac{\rho^+}{1 - \rho^+} \frac{\beta_2^+}{2\beta_1^+} = \frac{\rho\{\beta_2 + 2\theta_1[\beta_1 + \beta'(\gamma)] + \theta_2[1 - \beta(\gamma)]\}}{2\beta_1\{1 - \rho(1 + \theta_1[1 - \beta(\gamma)]/\beta_1)\}}, \quad (3.18)$$

and the variance of this waiting time is

$$\sigma^2\{W_{FP}\} = \frac{\rho^+}{1 - \rho^+} \frac{\beta_3^+}{3\beta_1^+} + \left[\frac{\rho^+}{1 - \rho^+} \frac{\beta_2^+}{2\beta_1^+} \right]^2. \quad (3.19)$$

The time in system S_{FP} of a customer is the independent sum of the waiting time until the start of service W_{FP} and an uninterrupted service time B so that, for example,

$$E\{S_{FP}\} = E\{W_{FP}\} + \beta_1, \quad \sigma^2\{S_{FP}\} = \sigma^2\{W_{FP}\} + \beta_2 - \beta_1^2. \quad (3.20)$$

The fraction of time this system is in an interrupted state is, cf. (3.17), as follows:

$$F_{FP} = 1 - \rho - \Pr\{W_{FP} = 0\} = \rho^+ - \rho = \lambda\theta_1[1 - \beta(\gamma)]. \quad (3.21)$$

Table 1 contains a numerical comparison of M/G/1 systems with preempting interruptions and postponed interruptions, where interruptions can only occur when the server is busy. In all cases, the mean service time is normalized to $\beta_1 = 1$. Case I is the base case, with offered load $\rho = \lambda = 0.8$, with exponentially distributed service times and interruption times, with failure rate $\gamma = 0.05$, and mean interruption time equal to the mean service time ($\theta_1 = 1$) so that the mean time to failure is more than 20 times the mean interruption time. The mean time in system is $(6.8125 - 6.6359) / 6.8125 \times 100\% = 2.6\%$ smaller when interruptions are postponed. Case II concerns less frequent longer interruptions. Less frequent longer interruptions are known to increase the delays of customers. But a decrease of the interruption rate by a factor

Table 1: Comparison of preempting versus postponed interruptions in busy times ($\beta_1 = 1$).

Case	I	II	III	IV	V	VI	VII	VIII	IX	X
λ	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.90	0.80	0.80
γ	0.05	0.01	0.25	0.05	0.05	0.10	0.01	0.05	1/19	0.50
θ_1	1.00	5.00	0.20	1.00	1.00	1.00	1.00	1.00	1.00	0.10
C_T^2	1.00	1.00	1.00	1.00	2.00	1.00	1.00	1.00	1.00	2.00
C_B^2	1.00	1.00	1.00	2.00	1.00	1.00	1.00	1.00	1.00	2.00
$\bar{\rho}$	0.8400	0.8400	0.8400	0.8400	0.8400	0.8800	0.8080	0.9450	0.8421	0.8400
ρ^+	0.8381	0.8396	0.8320	0.8372	0.8381	0.8727	0.8079	0.9429	0.8400	0.8234
$E\{W_F\}$	5.7625	6.7625	5.5625	8.5188	5.8875	8.7333	4.2921	18.859	5.8807	8.3063
$E\{W_{FP}\}$	5.6359	6.7136	5.1429	8.2558	5.7535	7.9481	4.2883	17.964	5.7375	7.1090
$E\{S_F\}$	6.8125	7.8125	6.6125	9.5688	6.9375	9.8333	5.3021	19.909	6.9333	9.3563
$E\{S_{FP}\}$	6.6359	7.7136	6.1429	9.2558	6.7535	8.9481	5.2883	18.964	6.7375	8.1090
$\sigma\{S_F\}$	6.8946	8.7556	6.6235	10.2903	7.0934	9.9847	5.3197	19.996	7.0193	10.0201
$\sigma\{S_{FP}\}$	6.7365	8.6472	6.1688	9.9813	6.9199	9.1211	5.3114	19.077	6.8425	8.7410
F_F	0.0400	0.0400	0.0400	0.0400	0.0400	0.0800	0.0080	0.0450	0.0421	0.0400
F_{FP}	0.0381	0.0396	0.0320	0.0372	0.0381	0.0727	0.0079	0.0429	0.0400	0.0234

5 leads to a reduction in the difference of the mean times in system to 1.3%. Case III concerns more frequent shorter interruptions. An increase of the interruption rate by a factor 5 leads to a rise in the difference of the mean times in system to 7.1%. Case IV concerns more variable service times, modeled by a gamma distribution with shape parameter 0.5. An increase of the squared coefficient of variation C_B^2 by a factor 2 leads to a rise in the difference of the mean times in system to 3.3%. Case V concerns more variable interruption times, also modeled by a gamma distribution with shape parameter 0.5. An increase of the squared coefficient of variation C_T^2 by a factor 2 leads to a slight rise in the difference of the mean times in system to 2.7%. Case VI concerns more frequent interruptions. An increase of the interruption rate by a factor 2 leads to a rise in the difference of the mean times in system to 9.0%. Case VII concerns less frequent interruptions. A decrease of the interruption rate by a factor 5 leads to a reduction in the difference of the mean times in system to 0.3%. Case VIII concerns a higher arrival rate. An increase of the load from 0.8 to 0.9 leads to a rise in the difference of the mean times in system to 4.7% while the mean times in system also strongly increase. Case IX concerns slightly more frequent interruptions to compensate for the reduction of interruptions due to the postponement of them; that is, F_{FP} at $\gamma = 1/19$ is the same as F_F at $\gamma = 1/20$. So, the values of the postponed interruptions model of case IX should be compared with the values of the preempting interruptions model of case I. After correction for the decrease of the number of actual interruptions, the difference of the mean times in system is still 1.1%. Finally, case X forms a kind of worst case scenario, with a rise in the difference of the mean times in system to 13.3%.

4. Interruptions during Idle and Busy Periods

In this section, it will be assumed that interruptions may also occur during idle periods. In this case, the probability p_A that the first customer of a busy period finds the server available, upon arrival is, since $\gamma/(\lambda + \gamma)$ is the probability that no customer arrives during a time

when the server is available and $\theta(\lambda)$ is the probability that no customer arrives during an interruption time is as follows:

$$p_A = \left[1 - \frac{\gamma}{\lambda + \gamma}\right] \sum_{j=0}^{\infty} \left[\frac{\gamma}{\lambda + \gamma} \theta(\lambda)\right]^j = \frac{\lambda}{\lambda + \gamma - \gamma\theta(\lambda)}. \quad (4.1)$$

The imbedded Markov chain $\{X_k, k = 0, 1, \dots\}$ of the number of customers left behind at departure instants satisfies the recurrence relations, for $k = 1, 2, \dots$, as follows:

$$X_k = \begin{cases} X_{k-1} - 1 + A(\tilde{B}_k), & \text{if } X_{k-1} > 0, \\ A(\tilde{B}_k), & \text{if } X_{k-1} = 0, \text{ w.p. } p_A, \\ A(\tilde{T}_k) - 1 + A(\tilde{B}_k), & \text{if } X_{k-1} = 0, \text{ w.p. } 1 - p_A. \end{cases} \quad (4.2)$$

Here, $A(\tilde{B}_k)$ denotes the number of arrivals during the k th service completion time (including interruption times), with PGF $\tilde{\beta}(\lambda[1-z])$, cf. (3.2), and $A(\tilde{T}_k)$ denotes the number of arrivals during the interruption time in which the k th customer arrives (if any), $k = 1, 2, \dots$. The PGF of the latter random variables is

$$E\{z^{A(\tilde{T}_k)}\} = \frac{\theta(\lambda[1-z]) - \theta(\lambda)}{1 - \theta(\lambda)}, \quad |z| \leq 1. \quad (4.3)$$

From the recursions (4.2), it follows by the procedure of Section 2 that the PGF of the stationary distribution of the number of customers in the system, X_I , for $|z| \leq 1$, is given by,

$$E\{z^{X_I}\} = \frac{(1 - \tilde{\rho})(1 - z)\tilde{\beta}(\lambda[1 - z])}{\tilde{\beta}(\lambda[1 - z]) - z} \frac{1 + \gamma\theta_1\hat{\theta}(\lambda[1 - z])}{1 + \gamma\theta_1}. \quad (4.4)$$

Here, $\hat{\theta}(\zeta) \doteq [1 - \theta(\zeta)]/[\zeta\theta_1]$ denotes the LST of the distribution of a residual interruption time. It turns out that the condition for stability is the same as that for systems without interruptions during idle periods, cf. (3.4). The LST of the stationary distribution of the waiting time W_I until the start of service under FCFS follows from (4.4) by the procedure of Section 2 as follows:

$$E\{e^{-\zeta W_I}\} = \frac{(1 - \tilde{\rho})\zeta}{\zeta - \lambda + \lambda\tilde{\beta}(\zeta)} \frac{1 + \gamma\theta_1\hat{\theta}(\zeta)}{1 + \gamma\theta_1}, \quad \Re\zeta \geq 0. \quad (4.5)$$

This LST is the product of the LST of the distribution of the waiting time W_F in systems without interruptions during idle periods, cf. (3.5), and the LST of the distribution of a residual interruption time at an arbitrary instant. Hence, it follows with (3.7) that

$$E\{W_I\} = E\{W_F\} + \frac{\gamma\theta_1}{1 + \gamma\theta_1} \frac{\theta_2}{2\theta_1} = \frac{1}{1 - \rho(1 + \gamma\theta_1)} \left[\rho(1 + \gamma\theta_1)^2 \frac{\beta_2}{2\beta_1} + \frac{\gamma\theta_1}{1 + \gamma\theta_1} \frac{\theta_2}{2\theta_1} \right]. \quad (4.6)$$

Observe that $E\{W_I\}$ does not vanish as $\rho \downarrow 0$, in contrast with $E\{W_F\}$, due to the possibly positive residual interruption time at an arrival instant. The variance of the waiting times can be computed from

$$\sigma^2\{W_I\} = \sigma^2\{W_F\} + \frac{\gamma\theta_1}{1+\gamma\theta_1} \frac{\theta_3}{3\theta_1} - \left[\frac{\gamma\theta_1}{1+\gamma\theta_1} \frac{\theta_2}{2\theta_1} \right]^2. \quad (4.7)$$

The probability that an arbitrary customer does not have to wait is the limit of the LST (4.5) as $\zeta \rightarrow \infty$

$$\Pr\{W_I = 0\} = \frac{1 - \tilde{\rho}}{1 + \gamma\theta_1} = \frac{1 - \rho(1 + \gamma\theta_1)}{1 + \gamma\theta_1}. \quad (4.8)$$

The time in system S_I is the independent sum of the waiting time W_I and the service completion time $\tilde{B}$, so that

$$E\{S_I\} = E\{W_I\} + \tilde{\beta}_1, \quad \sigma^2\{S_I\} = \sigma^2\{W_I\} + \tilde{\beta}_2 - \tilde{\beta}_1^2. \quad (4.9)$$

The fraction of time this system is in an interrupted state is, cf. (3.10), (4.8), given by

$$F_I = 1 - \rho - \Pr\{W_I = 0\} = \frac{\gamma\theta_1}{1 + \gamma\theta_1}. \quad (4.10)$$

This is intuitively clear, since the mean time that the server is interrupted is θ_1 , and the mean time between interruptions is $1/\gamma$.

Next, it will be assumed that interruptions can be postponed until the end of a service. Let X_k denote the number of customers left behind in the system at the end of the k th service, including a possible interruption. Then, this sequence of random variables forms an imbedded Markov chain and satisfies the recursion

$$X_k = \begin{cases} X_{k-1} - 1 + A(B_k^+), & \text{if } X_{k-1} > 0, \\ A(B_k^+), & \text{if } X_{k-1} = 0, \text{ w.p. } p_A, \\ A(\tilde{T}_k) - 1 + A(B_k^+), & \text{if } X_{k-1} = 0, \text{ w.p. } 1 - p_A. \end{cases} \quad (4.11)$$

Here, the random variable B_k^+ represents again the length of a service time and a possible interruption, and the PGF of $A(\tilde{T}_k)$ is again given by (4.3). From these recursions it follows by the procedure of Section 2 that the PGF of the stationary distribution of the number of customers in the system, X_{IP} , for $|z| \leq 1$, is given by,

$$E\{z^{X_{IP}}\} = \frac{(1 - \rho^+)(1 - z) \beta^+(\lambda[1 - z])}{\beta^+(\lambda[1 - z]) - z} \frac{1 + \gamma\theta_1 \hat{\theta}(\lambda[1 - z])}{1 + \gamma\theta_1}. \quad (4.12)$$

The number of customers Q_k waiting in the queue at the start of the k th service satisfies, like (2.4), the following:

$$Q_k = \begin{cases} X_{k-1} - 1, & \text{if } X_{k-1} > 0, \\ 0, & \text{if } X_{k-1} = 0, \text{ w.p. } p_A, \\ A(\check{T}_k) - 1, & \text{if } X_{k-1} = 0, \text{ w.p. } 1 - p_A. \end{cases} \quad (4.13)$$

With (4.11), it follows that $\beta^+(\lambda[1-z])E\{z^{Q_{IP}}\} = E\{z^{X_{IP}}\}$ for the stationary versions, so that (4.12) implies that

$$E\{z^{Q_{IP}}\} = \frac{(1-\rho^+)(1-z)}{\beta^+(\lambda[1-z]) - z} \frac{1 + \gamma\theta_1\hat{\theta}(\lambda[1-z])}{1 + \gamma\theta_1}, \quad |z| \leq 1. \quad (4.14)$$

Since $A(W_{IP}) = Q_{IP}$ holds as in the standard M/G/1 system, cf. (2.6), the LST of the FCFS waiting-time distribution follows as

$$E\{e^{-\zeta W_{IP}}\} = \frac{(1-\rho^+) \zeta}{\zeta - \lambda + \lambda\beta^+(\zeta)} \frac{1 + \gamma\theta_1\hat{\theta}(\zeta)}{1 + \gamma\theta_1}, \quad \Re\zeta \geq 0. \quad (4.15)$$

Like (3.16), this LST is the product of the LST of the distribution of the waiting time W_{FP} in systems without interruptions during idle periods and the LST of the distribution of a residual interruption time at an arbitrary instant. Hence, it follows like (4.6) and (4.7) that:

$$\begin{aligned} E\{W_{IP}\} &= E\{W_{FP}\} + \frac{\gamma\theta_1}{1 + \gamma\theta_1} \frac{\theta_2}{2\theta_1}, \\ \sigma^2\{W_{IP}\} &= \sigma^2\{W_{FP}\} + \frac{\gamma\theta_1}{1 + \gamma\theta_1} \frac{\theta_3}{3\theta_1} - \left[\frac{\gamma\theta_1}{1 + \gamma\theta_1} \frac{\theta_2}{2\theta_1} \right]^2. \end{aligned} \quad (4.16)$$

The probability that an arbitrary customer does not have to wait is

$$\Pr\{W_{IP} = 0\} = \frac{1 - \rho^+}{1 + \gamma\theta_1} = \frac{1 - \rho(1 + \theta_1[1 - \beta(\gamma)]/\beta_1)}{1 + \gamma\theta_1}. \quad (4.17)$$

The time in system S_{IP} of a customer is the independent sum of the waiting time until the start of service W_{IP} and a service time B so that

$$E\{S_{IP}\} = E\{W_{IP}\} + \beta_1, \quad \sigma^2\{S_{IP}\} = \sigma^2\{W_{IP}\} + \beta_2 - \beta_1^2. \quad (4.18)$$

The fraction of time this system is in an interrupted state is, cf. (3.21), (4.17), as follows:

$$F_{IP} = 1 - \rho - \Pr\{W_{IP} = 0\} = \frac{\gamma\theta_1}{1 + \gamma\theta_1} \left[1 - \rho + \rho \frac{1 - \beta(\gamma)}{\gamma\beta_1} \right]. \quad (4.19)$$

Table 2: Comparison of preempting versus postponed interruptions in up times ($\beta_1 = 1$).

Case	I	II	III	IV	V	VI	VII	VIII	IX	X
λ	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.90	0.80	0.80
γ	0.05	0.01	0.25	0.05	0.05	0.10	0.01	0.05	γ^*	0.50
θ_1	1.00	5.00	0.20	1.00	1.00	1.00	1.00	1.00	1.00	0.10
C_T^2	1.00	1.00	1.00	1.00	2.00	1.00	1.00	1.00	1.00	2.00
C_B^2	1.00	1.00	1.00	2.00	1.00	1.00	1.00	1.00	1.00	2.00
$\bar{\rho}$	0.8400	0.8400	0.8400	0.8400	0.8400	0.8800	0.8080	0.9450	0.8417	0.8400
ρ^+	0.8381	0.8396	0.8320	0.8372	0.8381	0.8727	0.8079	0.9429	0.8397	0.8234
$E\{W_I\}$	5.8101	7.0006	5.5720	8.5664	5.9589	8.8242	4.3020	18.907	5.9095	8.3134
$E\{W_{IP}\}$	5.6835	6.9517	5.1524	8.3035	5.8249	8.0390	4.2982	18.012	5.7692	7.1162
$E\{S_I\}$	6.8601	8.0506	6.6220	9.6164	7.0089	9.9242	5.3120	19.957	6.9616	9.3634
$E\{S_{IP}\}$	6.6835	7.9517	6.1524	9.3035	6.8249	9.0390	5.2982	19.012	6.7692	8.1162
$\sigma\{S_I\}$	6.9013	8.8873	6.6237	10.2948	7.1098	9.9934	5.3216	19.998	7.0043	10.0203
$\sigma\{S_{IP}\}$	6.7434	8.7806	6.1692	9.9860	6.9367	9.1306	5.3132	19.079	6.8310	8.7412
F_I	0.0476	0.0476	0.0476	0.0476	0.0476	0.0909	0.0099	0.0476	0.0496	0.0476
F_{IP}	0.0458	0.0472	0.0400	0.0450	0.0458	0.0843	0.0098	0.0456	0.0476	0.0318

From (4.18), (4.9), (4.6), (4.7), (4.16), (3.9), and (3.20), it follows that the absolute differences between the preempting and the postponing policies are the same for the cases with and without interruptions during idle periods.

$$E\{S_I\} - E\{S_{IP}\} = E\{S_F\} - E\{S_{FP}\}, \quad \sigma^2\{S_I\} - \sigma^2\{S_{IP}\} = \sigma^2\{S_F\} - \sigma^2\{S_{FP}\}. \quad (4.20)$$

Table 2 contains similar cases as Table 1 but for systems in which interruptions can also occur during idle periods. As relation (4.20) predicts, the absolute differences of the mean times in system without and with postponement are the same in both tables. Since the mean time in system is somewhat higher in systems in which interruptions can also occur during idle periods in comparison with that in corresponding systems without interruptions during idle periods, the relative differences of the mean times in system without and with postponement are somewhat smaller in Table 2. For case IX, the failure rate γ^* which makes the fraction of time in the interrupted state for the system with postponement, F_{IP} , equal to the fraction of time in the interrupted state for the system with preemption, $F_I = 0.0476$, at $\gamma = 0.05$, is $\gamma^* = 0.05217$, a bit smaller than the $1/19 \approx 0.05263$ in Table 1. This correction of the failure rate leads to difference of the mean times in system of 1.3% versus 2.6% in the base case I without this correction.

5. Conclusion

This paper has studied the differences in performance of queueing systems with interruptions which can be postponed until the end of the service in progress and comparable systems in which interruptions preempt the service in progress. It turns out that these differences are the same whether interruptions only occur during busy times of the server or also during idle times. The difference may only amount a few percentages in many cases, but may become quite large (well over 10%) if the failure rate increases at a fixed ratio of, say, 1 : 20 between the failure rate and the repair rate. And the difference further increases with the variability

in the service times and with the arrival rate. The difference is partly due to a reduction in the number of failures when failures are postponed and partly due to the feature that the customer in service is not delayed by a failure which is postponed.

References

- [1] W. D. Kelton, R. P. Sadowski, and N. B. Swets, *Simulation with Arena*, McGraw-Hill, New York, NY, USA, 5th edition, 2010.
- [2] D. P. Gaver, "A waiting line with interrupted service, including priorities," *Journal of the Royal Statistical Society, Series B*, vol. 24, pp. 73–90, 1962.
- [3] J. Keilson, "Queues subject to service interruption," *Annals of Mathematical Statistics*, vol. 33, pp. 1314–1322, 1962.
- [4] J. W. Cohen, *The Single Server Queue*, vol. 8 of *North-Holland Series in Applied Mathematics and Mechanics*, North-Holland, Amsterdam, The Netherlands, 2nd edition, 1982.
- [5] D. Fiems, T. Maertens, and H. Bruneel, "Queueing systems with different types of server interruptions," *European Journal of Operational Research*, vol. 188, no. 3, pp. 838–845, 2008.
- [6] A. Krishnamoorthy, P. K. Pramod, and T. G. Deepak, "On a queue with interruptions and repeat or resumption of service," *Nonlinear Analysis*, vol. 71, no. 12, pp. e1673–e1683, 2009.
- [7] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, Wiley Series in Probability and Statistics: Texts and References Section, John Wiley & Sons, New York, NY, USA, 3rd edition, 1998.