Queues with system disasters and impatient customers when system is down

Uri Yechiali

Department of Statistics and Operations Research
School of Mathematical Sciences
Tel Aviv University, Tel Aviv 69978, Israel
(uriy@post.tau.ac.il)

Abstract. Consider a system (e.g. a computer farm or a call center) operating as a M/M/c queue, where $c = 1$, or $1 < c < \infty$, or $c = \infty$. The system as a whole suffers disastrous breakdowns, resulting in the loss of all running and waiting sessions. When the system is down and undergoing a repair process, newly arriving customers become impatient: each individual customer activates a random-duration timer. If the timer expires before the system is repaired, the customer abandons the queue. We analyze this model and derive various quality of service measures: mean sojourn time of a *served* customer; proportion of customers served; rate of lost customers due to disasters; and rate of abandonments due to impatience.

Keywords queues, M/M/1, M/M/c, M/M/∞, failures, disasters, impatience, abandonments, sojourn times, quality of service.

1 Introduction

Consider a system (e.g. a call center or a computer farm) operating as a M/M/c queue. The system as a whole suffers random failures such that, when a failure occurs, *all* connections are cut and all existing requests are rejected and lost. The system then goes through a repair process whose duration is random. Meanwhile, while the system is down, the stream of newly arriving requests (customers) continues, but the customers become impatient: each customer 'activates' his own 'timer' with random duration T, such that, if the system is still down when the timer expires, the customer abandons the system never to return. Our goal is to calculate Quality of Service (QoS) measures: proportion of customers served; rate of customers rejected due to disasters; and rate of abandonments due to impatience when the system is down.

Models with customers impatience in queues have been studied by various authors in the past, where the source of impatience was either a long wait already experienced in the queue, or a long wait anticipated by a customer upon arrival (see e.g. [5], [9],[3],[7] and references there). Recently, we analyzed in [1] and [2] models with customers impatience when the server(s) is (are) on vacation and unavailable for service. The M/M/1, M/G/1, M/M/c and M/M/∞ queues were investigated. In the current study we extend the analysis to deal with the case where the system suffers random disasters, resulting in the *loss* of all customers present.

We first consider in section 2 the M/M/1 queue with Exponentially distributed life, repair and impatience times. We derive the PGFs of the queue sizes when the system is functioning and when it is down. The results depend on the solution of a certain differential equation. We then calculate the mean Sojourn Time of a *served* customer and derive Quality of Service (QoS) measures: proportion of customers served; rate of customers lost due to failures; and rate of abandonment due to impatience when the system is down.

In section 3 we study the case where failures may occur *only* when the system is functioning and serving customers. Similar QoS measures are derived. In section 4 we analyze the M/M/c queue and derive the corresponding conditional PGFs, mean queue sizes and sojourn times. QoS measures characterizing the system's effectiveness are calculated. Finally, in section 5, the M/M/∞ queue is investigated.

2 The M/M/1 queue with exponentially distributed life, repair and impatience times

2.1 The model

Customers arrive to a M/M/1-type queue according to a Poisson process with rate λ. Service times B are exponentially distributed with mean $1/\mu$. The system suffers disastrous breakdowns, occurring when the server is at its functioning phase, at a Poisson rate η. That is, the system 'life-time' is exponentially distributed with mean $1/\eta$. When the system fails, *all* customers present are rejected and *lost*. Upon failure, a repair process starts immediately. The repair time U is exponentially distributed with mean $1/\gamma$. Customers arriving while the system is down become *impatient*: Each customer activates an independent 'impatience timer' T, exponentially distributed with mean $1/\xi$, such that, if the repair process has not been completed by the time T expires, the customer abandons the system never to return.

The above process generates a two-dimensional continuous time Markov process as follows. Let J indicate the system's phase: $J = 1$ denotes the system is functioning and serving customers, while $J = 0$ indicates that the system is down, undergoing a repair process. Let L denote the number of customers in the system. Then the transition-rate diagram of the (J, L) process is depicted in Figure 1.

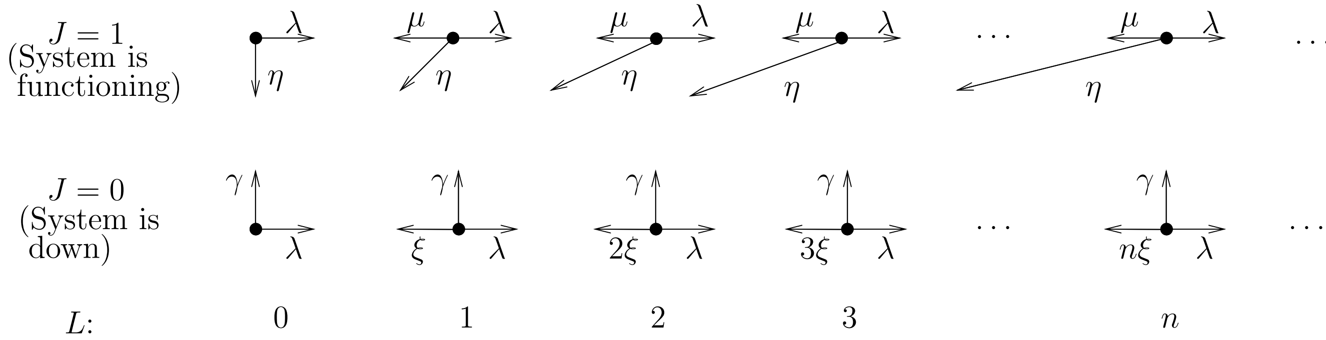

Figure 1: Transition-rate diagram

2.2 Balance equations and generating functions

Let $P_{jn} = P\{J = j, L = n\}$ $(j = 0, 1; n = 0, 1, 2, 3, ...)$ denote the system's steady-state probabilities. Let $P_{j\bullet} =: \sum_{n=0}^{\infty} P_{jn}$ be the probability that the system is in phase $j (j = 0, 1)$.

Then, the set of balance equations is given below:

$$\underline{J=0}\begin{cases} n = 0 & (\lambda+\gamma)P_{00} = \xi P_{01} + \eta \sum_{n=0}^{\infty} P_{1n} = \xi P_{01} + \eta P_{1\bullet} \\ n \geq 1 & (\lambda+\gamma+n\xi)P_{0n} = \lambda P_{0,n-1} + (n+1)\xi P_{0,n+1} \end{cases} \tag{2.1}$$

$$\underline{J=1}\begin{cases} n = 0 & (\lambda+\eta)P_{10} = \gamma P_{00} + \mu P_{11} \\ n \geq 1 & (\lambda+\mu+\eta)P_{1n} = \lambda P_{1,n-1} + \mu P_{1,n+1} + \gamma P_{0n} \end{cases} \tag{2.2}$$

Summing equations (2.1) over n yields

$$\gamma P_{0\bullet} = \eta P_{1\bullet} \tag{2.3}$$

Clearly, (2.3) can be obtained directly by employing a horizontal cut between the phases $J = 0$ and $J = 1$ in Figure 1.

Equation (2.3) together with $1 = P_{0\bullet} + P_{1\bullet}$ yield

$$P_{0\bullet} = \frac{\eta}{\gamma+\eta} \qquad P_{1\bullet} = \frac{\gamma}{\gamma+\eta} \tag{2.4}$$

Indeed, observing just the phase process itself, in which the system is either up or down, it is an alternating renewal process with the corresponding fractions of time $P_{1\bullet}$ and $P_{0\bullet}$.

For $j = 0, 1$, let $G_j(z) =: \sum_{n=1}^{\infty} P_{jn} z^n$ define the (conditional) Probability Generating Function (PGF) of phase j. Then, for each j, by multiplying correspondingly every equation for n by z^n and summing over n, we obtain, from (2.1) and (2.2), respectively

$$\xi(1-z)G_0'(z) = [\lambda(1-z)+\gamma]G_0(z) - \eta P_{1\bullet} \tag{2.5}$$

$$[(\lambda z-\mu)(1-z)+\eta z]\cdot G_1(z)=\gamma z G_0(z)-\mu(1-z)P_{10} \tag{2.6}$$

where $G_0'(z)=\frac{d}{dz}G_0(z)$.

Following the procedure in Altman and Yechiali [2004, [2005] it can be shown that the solution of the differential equation (2.5) is given by

$$G_0(z)=G_0(0)e^{\frac{\lambda}{\xi}z}\left[1-\frac{\int_0^z(1-s)^{\frac{\gamma}{\xi}-1}e^{-\frac{\lambda}{\xi}s}ds}{K}\right](1-z)^{-\frac{\gamma}{\xi}} \tag{2.7}$$

where $K=\int\limits_0^1(1-s)^{\frac{\gamma}{\xi}-1}e^{-\frac{\lambda}{\xi}s}ds$ and

$$G_0(0)=P_{00}=\eta P_{1\bullet}K/\xi\ . \tag{2.8}$$

Combining with eq. (2.1) for $n=0$, we get

$$P_{01}=\frac{\eta}{\xi}\left[\frac{(\lambda+\gamma)}{\xi}K-1\right]P_{1\bullet}\ . \tag{2.9}$$

Now, knowing P_{00} and P_{01}, any probability P_{0n}, for $n\geq 2$, can be calculated progressively by using equation (2.1), starting with $n=1$. From eq.(2.5), using $E[L_j]=\lim\limits_{z\to 1}G_j'(z)$, we get

$$E[L_0]=\frac{-\lambda G_0(1)+\gamma G_0'(1)}{-\xi}=\frac{\lambda P_{0\bullet}-\gamma E[L_0]}{\xi}\ ,$$

implying that

$$E[L_0]=\frac{\lambda P_{0\bullet}}{\gamma+\xi}\ . \tag{2.10}$$

Indeed, (2.10) states equality between the input rate to phase $J=0$, being $\lambda P_{0\bullet}$, and the output rate from that phase, $(\gamma+\xi)E[L_0]$. Differentiating equation (2.6) at $z=1$ we derive

$$(-\lambda+\mu+\eta)P_{1\bullet}+\eta E[L_1]=\gamma(P_{0\bullet}+E[L_0])+\mu P_{10}\ .$$

Using (2.3) we get

$$\eta E[L_1]+\mu(P_{1\bullet}-P_{10})=\lambda P_{1\bullet}+\gamma E[L_0] \tag{2.11}$$

Again, equation (2.11) equates the input and output rates of phase $J=1$.

To conclude the calculation of $E[L_1]$ we need P_{10}. This probability is obtained as follows. Let $g(z)=(\lambda z-\mu)(1-z)+\eta z$. Then equation (2.6) can be written as

$$g(z)G_1(z)=\gamma z G_0(z)-\mu(1-z)P_{10}\ .$$

The quadratic function $g(z)$ has exactly one root z_0 in $(0,1)$. (To see this, observe that $g(0)=-\mu$, $g(1)=\eta$, and $g(\infty)=-\infty$). Thus $g(z_0)=0$, implying that

$$P_{10}=\frac{\gamma z_0}{\mu(1-z_0)}G_0(z_0)\ . \tag{2.12}$$

where $z_0 = \frac{(\lambda+\mu+\eta)-\sqrt{(\lambda+\mu+\eta)^2-4\lambda\mu}}{2\lambda}$.

Thus, given P_{10}, $G_1(z)$ is completely determined by (2.6). Now, substituting (2.10) in (2.11) yields

$$\eta E[L_1] = \frac{\lambda}{\gamma+\xi}\left(\xi P_{1\bullet}+\gamma\right) - \mu\left(P_{1\bullet}-P_{10}\right) \tag{2.13}$$

Remark 1. One can view the system as a Markovian queue in 'random environment' (see Yechiali and Naor [10], Gupta, Scheller-Wolf, Harchol-Balter, and Yechiali [8], Byakal-Gursoy and Xiao [4], and D' Auria [6]), where the system alternates randomly between two phases: phase $J=0$, when it operates as an M(λ)/M(ξ)/∞ queue, and phase $J=1$, where it operates as an M(λ)/M(μ)/1 queue (or M/M/c in section 4, or M/M/∞ in section 5). However, there are two distinctions between the current process and the ones mentioned above: (i) here, when a switch from phase $J=1$ to phase $J=0$ occurs, it always brings the system to the same state, namely, to state $(0,0)$. (ii) all the above works assume work conservation, which is not the case in this work. This, in fact, leads to the quadratic function $g(z)$ (or $g_c(z)$ in the M/M/c case), rather than to a cubic $g(z)$, as is the case in [10] and [8].

2.3 Sojourn times

Let S denote the total Sojourn time of an arbitrary customer in the system, regardless of whether he has completed service or not. Then by Little's law,

$$E[S] = \frac{1}{\lambda}\left[E[L_0]+E[L_1]\right] . \tag{2.14}$$

Let S_{jn} be the total sojourn time in the system of a customer that *completes* service, given that he arrives when the system state is (j,n). Then, since future arrivals don't affect a customer's sojourn time, we have

$$E[S_{10}] = \frac{\mu}{\mu+\eta}\left(\frac{1}{\mu+\eta}\right) . \tag{2.15}$$

Clearly, $E[S_{10}] < \frac{1}{\mu}$. This follows since, if a service is completed before a failure , it should be a 'short' one. Also, when $\eta \to 0$, $E[S_{10}] \to \frac{1}{\mu}$.

Now, for $n \geq 1$,

$$E[S_{1n}] = \frac{\mu}{\mu+\eta}\left(\frac{1}{\mu+\eta}+E[S_{1,n-1}]\right) ,$$

implying that

$$E[S_{1n}] = \alpha + \beta E[S_{1,n-1}] \tag{2.16}$$

where $\alpha =: \frac{\mu}{(\mu+\eta)^2} = E[S_{10}]$, $\beta =: \frac{\mu}{\mu+\eta}$.

Iterating (2.16) we get

$$E[S_{1n}] = \alpha\sum_{k=0}^{n}\beta^k = \alpha\frac{1-\beta^{n+1}}{1-\beta} = \frac{\mu}{\eta(\mu+\eta)}\left[1-\left(\frac{\mu}{\mu+\eta}\right)^{n+1}\right] \tag{2.17}$$

Indeed, when $\eta \to 0$, $E[S_{1n}] \to \frac{n+1}{\mu}$.

We now turn to calculate $E[S_{0n}]$ for $n = 0, 1, 2, \ldots$

$$E[S_{00}] = \frac{\gamma}{\gamma+\xi}\left(\frac{1}{\gamma+\xi} + E[S_{10}]\right) \quad . \tag{2.18}$$

For $n \geq 1$,

$$E[S_{0n}] = \frac{\gamma}{\gamma+(n+1)\xi}\left(\frac{1}{\gamma+(n+1)\xi} + E[S_{1n}]\right) + \frac{(n+1)\xi}{\gamma+(n+1)\xi}\cdot\frac{n}{n+1}\left(\frac{1}{\gamma+(n+1)\xi} + E[S_{0,n-1}]\right)$$

where $n/(n+1)$ is the probability that, if impatience-dependent abandonment occurs, it is among one of the other n customers present. Thus, substituting (2.17),

$$(\gamma+(n+1)\xi)E[S_{0n}] = \frac{\gamma+n\xi}{\gamma+(n+1)\xi} + \gamma\alpha\frac{1-\beta^{n+1}}{1-\beta} + n\xi E[S_{0,n-1}] \tag{2.19}$$

Iterating (2.19) we derive

$$E[S_{0n}] = \frac{1}{\gamma+(n+1)\xi}\left[a_n + \sum_{k=0}^{n-1} a_k \prod_{j=k+1}^{n} c_j + \gamma\alpha\left(b_{n+1} + \sum_{k=0}^{n-1} b_{k+1} \prod_{j=k+1}^{n} c_j\right)\right] \tag{2.20}$$

where $c_j = \frac{j\xi}{\gamma+j\xi}$, $j = 0, 1, 2, \ldots$; $b_k = \frac{1-\beta^k}{1-\beta}$, $k \geq 1$; and $a_k = \frac{\gamma+k\xi}{\gamma+(k+1)\xi}$ $k = 0, 1, 2, \ldots$

Finally, the expected sojourn time of a customer that is *served* may be calculated using the expression

$$E[S_{(served)}] = \sum_{n=0}^{\infty} P_{0n}E[S_{0n}] + \sum_{n=0}^{\infty} P_{1n}E[S_{1n}] \tag{2.21}$$

Clearly, calculating $E[S_{(served)}]$ numerically, one has to use truncation.

2.4 Proportion of customers served

The system suffers from two types of losses: (i) rejected customers due to system's disastrous failures and (ii) abandonments of impatient customers during the repair phase. When the system is in state (j, n), $n \geq 1$, the rate of failure is η and then n customers are lost. Thus, the unit-time rate of lost customers, r, is given by

$$r = \sum_{n=1}^{\infty} \eta n P_{1n} = \eta E[L_1] \tag{2.22}$$

Similarly, the expected number of customers served per unit of time is $\mu(P_{1\bullet} - P_{10})$, implying that the proportion of customers *served* is

$$P_{(served)} = (P_{1\bullet} - P_{10})\,\frac{\mu}{\lambda} \tag{2.23}$$

Finally, the rate of *abandonment* due to impatience, $R_{(aband)}$, is given by

$$R_{(aband)} = \lambda - \eta E[L_1] - \mu(P_{1\bullet} - P_{10}) = \lambda P_{0\bullet} - \gamma E[L_0] = \xi E[L_0] \tag{2.24}$$

3 M/M/1: Failures occur only when the system is functioning and $L \geq 1$

Consider now the case where a functioning system (in phase $J = 1$) may fail *only* when it is operative, i.e. only when $L \geq 1$ and it serves customers. Then, the corresponding transition-rate diagram will look similar to Figure 1 with the exception that the only possible transition from state (1,0) is to state (1,1), a transition caused by an arrival of a new customer.

3.1 Balance equations and generating functions

With P_{jn} defined as before, the set of balance equations is:

$$J = 0 \begin{cases} n = 0 & (\lambda + \gamma)P_{00} = \xi P_{01} + \eta(P_{1\bullet} - P_{10}) \\ n \geq 1 & (\lambda + \gamma + n\xi)P_{0n} = \lambda P_{0,n-1} + (n+1)\xi P_{0,n+1} \end{cases} \tag{3.1}$$

$$J = 1 \begin{cases} n = 0 & \lambda P_{10} = \gamma P_{00} + \mu P_{11} \\ n \geq 1 & (\lambda + \mu + \eta)P_{1n} = \lambda P_{1,n-1} + \mu P_{1,n+1} + \gamma P_{0n} \end{cases} \tag{3.2}$$

Summing equations (3.1) over n yields

$$\gamma P_{0\bullet} = \eta(P_{1\bullet} - P_{10}) \tag{3.3}$$

Using (3.1) and (3.2) we derive the corresponding PGFs:

$$\xi(1-z)G_0'(z) = [\lambda(1-z) + \gamma]G_0(z) - \eta(P_{1\bullet} - P_{10}) \tag{3.4}$$

and

$$[(\lambda z - \mu)(1-z) + \eta z]G_1(z) = \gamma z G_0(z) + [\mu(z-1) + \eta z]P_{10} \tag{3.5}$$

The solution of (3.4) is given by equation (2.7) but this time

$$G_0(0) = P_{00} = \frac{\eta(P_{1\bullet} - P_{10})K}{\xi} \tag{3.6}$$

Combining with (3.1) for $n = 0$, and using (3.3), we get

$$P_{01} = \frac{\gamma}{\xi}\left[\frac{(\lambda + \gamma}{\xi}K - 1\right]P_{0\bullet} \tag{3.7}$$

From equation (3.4), using $E[L_j] = \lim_{z\to 1} G_j'(z)$, we derive

$$E[L_0] = \frac{\lambda P_{0\bullet} - \gamma E[L_0]}{\xi}$$

implying that

$$E[L_0] = \frac{\lambda P_{0\bullet}}{\gamma + \xi} \tag{3.8}$$

From (3.5),

$$(-\lambda + \mu + \eta)P_{1\bullet} + \eta E[L_1] = \gamma(P_{0\bullet} + E[L_0]) + (\mu + \eta)P_{10} \tag{3.9}$$

To calculate $E[L_1]$ we need P_{10}. We rewrite equation (3.5) as

$$g(z)G_1(z) = \gamma z G_0(z) + [\mu(z-1) + \eta z]P_{10}$$

where the quadratic equation $g(z)$ has a unique root z_0 in $(0, 1)$, given in Section 2.

Thus, $g(z_0) = 0$ implies that

$$P_{10} = \frac{\gamma z_0}{\mu(1 - z_0) - \eta z_0} G_0(z_0) \tag{3.10}$$

Using (3.3) and $P_{0\bullet} = 1 - P_{1\bullet}$ we have

$$P_{1\bullet} = \frac{\gamma + \eta P_{10}}{\eta + \gamma} \quad P_{0\bullet} = \frac{\eta(1 - P_{10})}{\eta + \gamma} \tag{3.11}$$

Now, given P_{10} and using (3.3), equation (3.9) reduces to equality between output and input rates of phase 1:

$$\mu(P_{1\bullet} - P_{10}) + \eta E[L_1] = \gamma E[L_0] + \lambda P_{1\bullet} \tag{3.12}$$

Using (3.8) and $P_{0\bullet} + P_{1\bullet} = 1$ we obtain $E[L_1]$:

$$E[L_1] = \frac{1}{\eta}\left[\frac{\gamma\lambda}{\gamma + \xi}(1 - P_{1\bullet})\right] - (\mu - \lambda)P_{1\bullet} + \mu P_{10} \tag{3.13}$$

The expected total number of customers in the system is

$$E[L] = E[L_0] + E[L_1] \tag{3.14}$$

3.2 Sojourn times

As in section 2, $E[S] = \frac{1}{\lambda}E[L]$.

The mean sojourn time of a customer that *completes* service is calculated similarly to Section 2. Indeed, equations (2.15) to (2.20) *hold here as well*. Thus, $E[S_{(served)}]$ is given by (2.21), *but* the P_{jn}'s are *different*.

3.3 Proportion of customers served

Equations (2.22), (2.23) and (2.24) hold true. However, $P_{10}, P_{1\bullet}$ and $E[L_1]$ are given by (3.10), (3.11) and (3.13), respectively.

4 The c-server case

4.1 The model, balance equations, PGFs and mean queue sizes

Consider now the case with $c \geq 1$ servers. The system alternates between phases 1 and 2 as described in Section 2.1. When the system fails, *all* servers stop working and *all* customers present are *lost*. It turns out that the proportions of time the system stays in the two phases are not effected by the number of servers in the system. That is, $P_{0\bullet}$ and $P_{1\bullet}$ are given by equation (2.4).

The balance equations for phase $J = 0$ are the *same* as those given by equations (2.1). The balance equations for $J = 1$ are:

$$\begin{aligned} n = 0 \quad & (\lambda + \eta)P_{10} = \gamma P_{00} + \mu P_{11} \\ c - 1 \geq n \geq 1 \quad & (\lambda + n\mu + \eta)P_{1n} = \lambda P_{1,n-1} + (n+1)\mu P_{1,n+1} + \gamma P_{0n} \\ n \geq c \quad & (\lambda + c\mu + \eta)P_{1n} = \lambda P_{1,n-1} + c\mu P_{1,n+1} + \gamma P_{0n} \end{aligned} \tag{4.1}$$

It follows that the PGF $G_0(z)$, the probabilities P_{00} and P_{01}, and $E[L_0]$ are given by equations (2.7),(2.8),(2.9) and (2.10), respectively. As before, with P_{00} and P_{01} given, one can calculate any probability P_{0n} progressively from (2.1)

The PGF of phase 1, $G_1(z)$, is derived as:

$$[(\lambda z - c\mu)(1 - z) + \eta z]G_1(z) = \gamma z G_0(z) - \mu(1 - z)\sum_{n=0}^{c}(c - n)P_{1n}z^n \ . \tag{4.2}$$

Let $g_c(z) = (\lambda z - c\mu)(1 - z) + \eta z$. Then, $g_c(z)$ has a unique root z_c in $(0, 1)$, where

$$z_c = \frac{(\lambda + c\mu + \eta) - \sqrt{(\lambda + c\mu + \eta)^2 - 4\lambda c\mu}}{2\lambda}$$

Thus, using (4.2), $g(z_c) = 0$ implies

$$0 = \gamma z_c G_0(z_c) - \mu(1 - z_c)\sum_{n=0}^{c}(c - n)P_{1n}z_c^n \ . \tag{4.3}$$

Equation (4.3) involves $(c + 1)$ unknown probabilities: P_{10} to P_{1c}. The first c equations in (4.1), for $c - 1 \geq n \geq 0$, give additional c equations in those $(c + 1)$ unknowns. Solving this set of $(c + 1)$ independent equations yields the required probabilities.

Finally, to get $E[L_1]$, we use (4.2) and obtain, similarly to (2.11),

$$\eta E[L_1] + c\mu P_{1\bullet} - \mu\sum_{n=9}^{c}(c - n)P_{1n} = \lambda P_{1\bullet} + \gamma E[L_0] \tag{4.4}$$

4.2 Sojourn times

We use the same notation as in previous sections. First we note that S_{0n} is independent of the number of servers, c. Thus, the calculations of $E[S_{0n}]$ are no different, implying that $E[S_{00}]$, as well as $E[S_{0n}]$ for $n \geq 1$, are given by (2.18) and (2.20), respectively. However, the calculations of $E[S_{1n}]$ depend on c and are more involved. We start with $n \geq c$:

$$E[S_{1n}] = \frac{c\mu}{c\mu + \eta}\left(\frac{1}{c\mu + \eta} + E[S_{1,n-1}]\right) \tag{4.5}$$

The term $E[S_{1,n-1}]$ in (4.5) follows since, when $n \geq c$, a customer *waits in line*, and a service completion advances him one position forward in the line. Therefore, similarly to (2.16),

$$E[S_{1n}] = \alpha_c + \beta_c E[S_{1,n-1}] \quad (n \geq c) \tag{4.6}$$

where,

$$\alpha_c = \frac{c\mu}{(c\mu + \eta)^2} \quad \text{and} \quad \beta_c = \frac{c\mu}{c\mu + \eta}$$

Thus, for $m \geq 0$

$$\begin{aligned} E[S_{1,c+m}] &= \alpha_c \frac{1 - \beta_c^{m+1}}{1 - \beta_c} + \beta_c^{m+1}[ES_{1,c-1}] \\ &= \frac{c\mu}{\eta(c\mu + \eta)}(1 - \beta_c^{m+1}) + \beta_c^{m+1} E[S_{1,c-1}] \end{aligned} \tag{4.7}$$

Now, for $n \leq c - 1$, each arrival is independent of the other customers present, leading to

$$E[S_{1n}] = \frac{\mu}{\mu + \eta}\left(\frac{1}{\mu + \eta}\right) \quad (n \leq c - 1) \tag{4.8}$$

Note: It can readily be checked that when $\eta = 0$ (no breakdowns) $E[S_{1n}] = 1/\mu$ for all $0 \leq n \leq c - 1$, and that $E[S_{1,c+m}] = \frac{m+1}{c\mu} + \frac{1}{\mu}$, for $m \geq 0$. Finally, $E[S_{(served)}]$, the expected sojourn time of a served customer is given (again) by (2.21), but with the values $\{P_{0n}\}$,$\{P_{1n}\}$, $\{E[S_{0n}]\}$ and $\{E[S_{1n}]\}$ derived in this section.

4.3 Proportion of customers served

The expected number of *rejected* customers per unit-time, due to failures, is

$$r_c = \eta E[L_1] \tag{4.9}$$

The expected number of customers *served* per unit-time is

$$E[Cust.\ Served] = \sum_{n=1}^{c-1} n\mu P_{1n} + \sum_{n=c}^{\infty} c\mu P_{1n} = c\mu P_{1\bullet} - \mu \sum_{n=0}^{c-1}(c - n)P_{1n} \tag{4.10}$$

Thus, using (4.4), the rate of abandonment is given by

$$R_{(aband)} = \lambda - r_c - E[Cust.\ Served] = \lambda P_{0\bullet} - \gamma E[L_0] = \xi E[L_0] \tag{4.11}$$

Remark 2. The scenario where the system can fail only when there is at least one customer present can be analyzed in a similar manner and will not be presented here.

5 The M/M/∞ queue

In this section we analyze the case where the underlying process is an M/M/∞ queue. Although more involved, it can be analyzed in a similar manner to the single-server and the multi-server cases treated in Sections 2 and 4, respectively. We present the scenario where a disaster can occur only when the system is rendering service.

5.1 Balance equations and PGFs

For $J = 0$, equations (2.1) hold here as well, implying that $G_0(z)$ and $G_0(0) = P_{00}$, are, again, given by (2.7) and (2.8), respectively. Moreover, the proportion of time the system resides in phase j $(j = 1, 2)$ is given, once more, by (2.4), while $E[L_0]$ is given by (2.10).

For $J = 1$ we have:

$$\underline{J=1}\begin{cases} n=0 & (\lambda+\eta)P_{10} = \gamma P_{00} + \mu P_{11} \\ n \geq 1 & (\lambda + n\mu + \eta)P_{1n} = \lambda P_{1,n-1} + (n+1)\mu P_{1,n+1} + \gamma P_{0n} \end{cases} \tag{5.1}$$

The above leads to the following differential equation for $G_1(z)$:

$$\mu(1-z)G_1'(z) = [\lambda(1-z)+\eta]G_1(z) - \gamma G_0(z) \tag{5.2}$$

Dividing by $\mu(1-z)$ and multiplying by $e^{-\frac{\lambda}{\mu}z}(1-z)^{\eta/\mu}$ leads to

$$e^{-\frac{\lambda}{\mu}z}(1-z)^{\frac{\eta}{\mu}}G_1'(z) - \left[\frac{\lambda}{\mu} + \frac{\eta}{\mu(1-z)}\right] e^{-\frac{\lambda}{\mu}z}(1-z)^{\frac{\eta}{\mu}}G_1(z) = -\frac{\gamma}{\mu(1-z)}e^{-\frac{\mu}{\lambda}z}(1-z)^{\eta/\mu}G_0(z)$$

That is,

$$\frac{d}{dz}\left[e^{-\frac{\mu}{\lambda}z}(1-z)^{\frac{\eta}{\mu}}G_1(z)\right] = -\frac{\gamma}{\mu}e^{-\frac{\mu}{\lambda}z}(1-z)^{\frac{\eta}{\mu}-1}G_0(z) \tag{5.3}$$

Integrating (5.3) from 0 to z leads to

$$e^{-\frac{\lambda}{\mu}z}(1-z)^{\eta/\mu}G_1(z) - G_1(0) = -\frac{\gamma}{\mu}\int_{s=0}^{z}(1-s)^{\eta/\mu-1}e^{-\frac{\lambda}{\mu}s}G_0(s)ds$$

Thus,

$$G_1(z) = G_1(0)e^{\frac{\lambda}{\mu}z}(1-z)^{-\frac{\eta}{\mu}} - \frac{\gamma}{\mu}e^{\frac{\lambda}{\mu}z}(1-z)^{-\frac{\eta}{\mu}}\int_{s=0}^{z}(1-s)^{\eta/\mu-1}e^{-\frac{\lambda}{\mu}s}G_0(s)ds \tag{5.4}$$

At $z = 1$ we have

$$G_1(1) = P_{1\bullet} = e^{\frac{\lambda}{\mu}}\left[G_1(0) - \frac{\gamma}{\mu}\int_{s=0}^{1}(1-s)^{\eta/\mu-1}e^{-\frac{\lambda}{\mu}s}G_0(s)ds\right]\lim_{z\to 1}(1-z)^{-\eta/\mu}$$

Since $P_{1\bullet} > 0$ and $\lim_{z\to 1}(1-z)^{-\eta/\mu} = \infty$, we must have that

$$G_1(0) = \frac{\gamma}{\mu}\int_{s=0}^{1}(1-s)^{\eta/\mu-1}e^{-\frac{\lambda}{\mu}s}G_0(s)ds =: \frac{\gamma}{\mu}K_\infty \tag{5.5}$$

where

$$K_\infty = \int_{s=0}^{1}(1-s)^{\frac{\eta}{\mu}-1}e^{-\frac{\lambda}{\mu}s}G_0(s)ds$$

Then, using (2.7) and (2.8), we get

$$K_\infty = P_{1\bullet}\frac{\eta}{\xi}\int_{s=0}^{1}(1-s)^{\frac{\eta}{\mu}-\frac{\gamma}{\xi}-1}e^{\left(\frac{\lambda}{\xi}-\frac{\lambda}{\mu}\right)s}\left[\int_{x=s}^{1}(1-x)^{\frac{\gamma}{\xi}-1}e^{-\frac{\lambda}{\xi}x}dx\right]ds \tag{5.6}$$

Clearly, K_∞ can easily be calculated for any set of parameters.

Finally, $E[L_1]$is derived from (5.2). Dividing both sides by $(z-1)$ we get

$$\mu E[L_1] = \lambda P_{1\bullet} + \lim_{z\to 1}\left\{\frac{\eta G_1(z)-\gamma G_0(z)}{1-z}\right\} = \lambda P_{1\bullet} + \gamma E[L_0] - \eta E[L_1]$$

Substituting (2.10) in the above leads to

$$(\mu+\eta)E[L_1] = \lambda\left(1-\frac{\xi}{\gamma+\xi}P_{0\bullet}\right) \tag{5.7}$$

5.2 Sojourn times and abandonment rate

There is no change in the derivations of $E[S_{0n}]$ for all n. Once more, $E[S_{00}]$ and $E[S_{0n}]$ for $n \geq 1$ are given by (2.18) and (2.20), respectively. $E[S_{1n}]$ is given by (4.8) for all n.

The abandonment rate is given by

$$R_{(\text{aband})} = \lambda - (\mu+\eta)E[L_1] \tag{5.8}$$

6 Conclusion

In this paper we've studied the M/M/1, M/M/c and M/M/∞ queues subject to system's disasters, resulting in the loss of all customers present. Then, when the system is down and undergoes a repair period, each waiting customer becomes *impatient* and *abandons* the system as soon as his actual waiting time *exceeds* his individual (random) waiting time threshold. We've derived important QoS performance measures such as the rate of lost customers due to failures, and rate of abandonments due to customers impatience when the system is down. Our model differs from the common scenarios of customers impatience (see e.g. [5] [9],[3],[7] and references there) by assuming that customers become impatient *only* when the system is down and they have no exact information on the time when it will become operative again. This situation reflects customers behavior in various real life service systems.

References

[1] E. Altman and U. Yechiali, Analysis of Customers' Impatience in Queues with Server Vacations. Queueing Systems. 52(2006) 261–279.

[2] E. Altman and U. Yechiali, Infinite-Server Queues with System's Additional Tasks and Impatience Customers. Technical Report, Tel Aviv University (April 2005).

[3] F. Baccelli, P. Boyer and G. Hebuterne, Single-Server Queues with Impatience Customers. Advances in Applied Probability, 16(1984) 887–905.

[4] M. Baykal-Gursoy and W. Xiao, Stochastic Decomposition in M/M/∞ Queues with Markov-Modulated Service Rates. Queueing Systems. 48(2004) 75–88.

[5] D.J. Daely, General Customer Impatience in the Queue GI/G/1. J. Applied Probability, 2(1965) 186–205.

[6] B. D' Auria, Stochastic Decomposition of the M/G/∞ Queue in Random Environment. Report 2005-045, Eurandom, The Netherlands.

[7] N. Gans, G. Koole and A. Mandelbaum, Telephone call centers: Tutorial, review, and research prospects. Manufacturing and Service Operations Management 5(2003) 79–141.

[8] V. Gupta, A. Scheller-Wolf, M. Harchol-Balter and U. Yechiali, Fundamental Characteristics of Queues with Fluctuating Load. Proceedings of ACM SIGMETRIC (2006).

[9] L. Takacs, A Single-Server Queue with Limited Virtual Waiting Time. J. Applied Probability 11(1974) 612–617.

[10] U. Yechiali and P. Naor, Queuing Problems with Heterogeneous Arrivals and Service. Operations Research 19(1971) 722–734.