Transient analysis of a Markovian queue with deterministic rejection

Benjamin Legros

Ecole de Management de Normandie, Laboratoire Métis, 64 Rue du Ranelagh, 75016 Paris, France

benjamin.legros@centraliens.net

Abstract

We analyze the transient behavior of the M/M/1+D queue. Considering an Erlang distribution for customers' waiting time, we approximate the real system by a Markov chain. We obtain the Laplace Transform of the transient probabilities in the approximated model and the Laplace transform of the main performance measures for the real system. We next analyze the busy period of this queue. One interesting insight is that the busy period of the unstable M/M/s queue has a finite coefficient of variation.

Keywords: Transient analysis; performance evaluation; deterministic rejection; queueing model; Erlang approximation

1 Introduction

Most results in the queuing theory and its applications are for the stationary regime. They characterize the system when the time from initialization becomes very large which renders the impact of the initial conditions negligible. The popularity of the stationary analysis comes from its simplicity. By solving a set balance equations, the stationary performance measures of many classical queues (M/M/1, M/G/1, M/M/c, ...) are known explicitly and have relatively simple forms. In practice, the analysis of the stationary regime makes sense in some contexts. For instance in call centers, it is appropriate to assume that a system with constant parameters achieves a steady-state quickly within short-half hour or hour-intervals [15, 13].

Nevertheless, the stationary analyses are inappropriate in many situations if the time from initialization is not large enough. This is particularly the case when there is a definite closing time and when the service times are long. For instance, the number of patients seen by a physician during a working period is not sufficient to assume that a stationary regime is achieved. Even in call centers, the recent improvements in customers identification via data analysis reduce the value of the stationary analysis where customers are seen as a uniform flow. Therefore, the transient

analysis is highly valuable for a better understanding of queueing systems. However, due to the complexity of the transient regime, available results are usually restricted.

In this article, we consider a single server queue with infinite capacity, starting initially empty, a first-come-first-served discipline, an exponential service time with service rate μ and a Poisson arrival process with rate λ. In addition, we assume that a customer is automatically rejected if her actual waiting time reaches the deterministic threshold τ. This corresponds to Web applications where a timeout threshold is set by administrators [32] or call centers where customers are invited to be called back later at a given waiting time [24]. This queue is referred to as the M/M/1+D queue. To the best of our knowledge, the transient analysis of this queue hasn't yet been done. Note that our results can lead to the transient performance measures in the multi-server case or with different initial conditions. These extensions are presented in the Online Supplement.

The difficulty for the analysis of this queue is the presence of a non-exponential duration; the rejection time. The system therefore cannot be modeled by a simple Markov chain where a state of the system corresponds to the number of customers. To overcome this difficulty, we first approximate the waiting time of the first customer in line in the queue by an Erlang distribution as in [22] and [25]. This allows us to represent the system evolution by a Markov chain. As the parameters of the Erlang distribution tend to infinity, the approximated model converges to the real one. After writing the balanced equations, we introduce the z-transform of the transient probabilities. We next obtain an explicit solution for the Laplace transform of this function which in turn allows us to derive the Laplace transform of the relevant performance measures; the probability of an empty system, the probability of rejection, the expected waiting time and the probability of waiting more than a given threshold. Finally, with a similar approach, we analyze the *busy period* of the M/M/1+D queue. We deduce from this analysis that the busy period of the unstable M/M/1 queue has a finite coefficient of variation when $\lambda > \mu$.

Structure of the article. The remainder of this paper is structured as follows. We conclude this section with a literature survey. Section 2 explains the system modeling. Section 3 determines the explicit Laplace transform of the transient probabilities. Section 4 computes the performance measures of the real system. Section 5 illustrates the applicability of our results. Finally, Section 6 investigates the busy period of the M/M/1+D queue. In the Online Supplement, we present the multi-server case, the performance measures under different initial conditions and detailed proofs

for the main results.

Literature review. In the queueing literature, the analyses of queues under a *transient regime* have a long history. The M/M/1 queue is the first studied queue [23, 10, 11]. The transient queue-length distribution is explicitly known in terms of modified Bessel functions of the first kind. However, the complexity of the involved expression makes it complicated to obtain insights for this queue. Further investigations have therefore been devoted to a better understanding of this queue. For instance, [2] and [3] establish a transform factorization that facilitates developing approximations for the moments of the queue length. Several approaches for the analysis of the M/M/1 queue have been considered. We refer to [30] for a review of the main results for the computation of the performance measures of the M/M/1 queue. The most popular approach has been the one of of [10] involving generating functions for the partial differential equation. For instance, [28] apply this approach for the explicit performance measures of the M/M/1 queue with finite capacity. The extension from the M/M/1 to the M/G/1 queue has been extensively studied. [29] is the first to provide integral expressions of the performance measures for this queue. Later, [4] investigate the moments of this queue. A moment is characterized in terms of a differential equation involving lower moment functions and the time-dependent server occupation probability. Different variations of the M/G/1 have been studied. [14] determine an analytical expression of the probability distribution of the M/D/1/N queue initialized at an arbitrary deterministic state. [16] consider a particular M/G/1 queue with an Erlang service time distribution. [31] consider the M/G/1 retrial queue with disasters and service failures. [20] tackle the finite buffer M/G/1 queue with server vacations. In addition, the M/G/1 queue has been considered under a processor sharing discipline [21, 18]. For the multi-server setting, [19] evaluate the transient behavior of the M/M/s queue and show the implications of this analysis for simulations. Later, [26] obtain a solution for the M/M/s queue from which the stationary behavior can be easily derived. Including abandonment or rejection renders the performance evaluation difficult. Therefore, most studies of such queues have been done under stationary assumptions [27, 8, 9]. Considering the transient analysis, we mention [6] for the performance measures of the M/M/s+M queue and [7] for the study of its busy period.

2 Model description

We define a continuous time Markov chain in which we approximate the waiting time of the first customer in line (FIL) by an Erlang distribution with rate γ per phase. The total number of phases of this distribution is not known beforehand. This is determined by service completion times and the FIL rejection time. This non-traditional approach for the definition of the system state has been first proposed in [22] without abandonment and next extended in [25] with abandonment.

More precisely, we denote by x a state of the system, where $-1 \leq x \leq n$. State $x = -1$ corresponds to an empty system, State $x = 0$ corresponds to a busy server with an empty queue, and states with $0 < x \leq n$ correspond to a situation where the FIL is at phase time x. We choose x, n and γ such that $\frac{x}{\gamma} \stackrel{\Delta}{=} t$ and $\frac{n}{\gamma} \stackrel{\Delta}{=} \tau$. This ensures that as x, n, and γ tend to infinity, the random variable which represents the discretized waiting phase time of the FIL converges to a deterministic elapsing of time t ($0 \leq t \leq \tau$). This in turn leads to an exact analysis.

We now explain the transition structure of the Markov process. Assume that the FIL is in waiting phase x, for $x > 0$. Since the discipline of service is first-come-first-served, a service completion (see transition Type 3) results in removing the FIL from the queue. Moreover, since we consider a deterministic rejection time, a rejection (see transition Type 4) can also only happen to the FIL. Therefore, after a service completion or a rejection, the new FIL (if any) is in a lower waiting phase than the initial one. We denote by h the number of waiting phases between the FIL just before and the FIL just after a service completion or a rejection ($0 \leq h \leq x$). The transition probability, $r_{x,x-h}$, from state $x > 0$ to state $x - h$ can be found in [25] (Table 1, line 4) by $r_{x,x-h} = \left(\frac{\lambda}{\lambda+\gamma}\right)\left(\frac{\gamma}{\lambda+\gamma}\right)^h$ for $0 \leq h < x$, and by $r_{x,0} = \left(\frac{\gamma}{\lambda+\gamma}\right)^x$. The four possible transitions in the Markov chain are the following:

1. An arrival with rate λ with an empty queue ($x = -1, 0$), which changes the state to $x+1$. If initially $x = -1$, then the server becomes busy. Otherwise if initially $x = 0$, the FIL entity is created in phase time 1.

2. A phase increase with rate γ while the system is not empty and the FIL is not in waiting phase n ($0 < x < n$), which changes the state to $x+1$. The phase time of the FIL is increased by 1.

3. A service completion with rate $\mu r_{x,x-h}$ while the queue is not empty ($0 < x \leq n, 0 \leq h \leq x$),

which changes the state to $x-h$, that is, the new FIL is in phase time $x-h$ if $x-h>0$ or the queue is empty if $x-h=0$.

4. A phase increase with rate $\gamma r_{n,n-h}$ while the FIL in the queue is in waiting phase n ($x=n, 0\leq h\leq n$), which changes the state to $n-h$, that is, the new FIL is in waiting phase $n-h$ if $n-h>0$ or the queue is empty if $n-h=0$.

The transition structure defined above determines a Markov chain for which we are interested in the transient behavior. We denote by $\pi_x(t)$, the transient probability to be in state x at time $t\geq 0$ and assume that the system starts empty; $\pi_{-1}(0)=1$. In order to simplify the notations, we write π_x instead of $\pi_x(t)$ and denote by q the ratio $\frac{\gamma}{\lambda+\gamma}$. The differential-difference equations governing the phase time of the FIL are given in Equation (1) as

$$
\begin{aligned}
\frac{\partial \pi_{-1}}{\partial t} &= -\lambda\pi_{-1}+\mu\pi_0, \\
\frac{\partial \pi_0}{\partial t} &= -(\lambda+\mu)\pi_0+\lambda\pi_{-1}+\mu\sum_{k=1}^{n}q^k\pi_k+\gamma q^n\pi_n, \\
\frac{\partial \pi_1}{\partial t} &= -(\gamma+\mu q)\pi_1+\mu\sum_{k=1}^{n-1}(1-q)q^k\pi_{1+k}+\gamma(1-q)q^{n-1}\pi_n+\lambda\pi_0, \\
\frac{\partial \pi_x}{\partial t} &= -(\gamma+\mu q)\pi_x+\mu\sum_{k=1}^{n-x}(1-q)q^k\pi_{x+k}+\gamma(1-q)q^{n-x}\pi_n+\gamma\pi_{x-1}, \text{ for, } 2\leq x\leq n-1, \\
\frac{\partial \pi_n}{\partial t} &= -(\gamma+\mu)q\pi_n+\gamma\pi_{n-1}.
\end{aligned}
\tag{1}
$$

3 Laplace Transforms of the transient probabilities π_x

In Theorem 1, we provide explicit expressions of the Laplace Transform of the transient probabilities. To prove Theorem 1, we introduce the probability generating function, defined as $P(z,t)=\sum_{x=0}^{n}\pi_x z^x$. This function is related to the π_x's via $\pi_x=\frac{1}{x!}\frac{\partial^x P(z,t)}{\partial z^x}|_{z=0}$, for $0\leq x\leq n$. Using Equation (1), we determine the differential equation satisfied by $P(z,t)$. We define the Laplace Transform (LT) of a function $f(z,t)$ ($z\in\mathbb{C}$, $t\geq 0$) as follows:

$$
f^*(z,y)=\int_0^\infty e^{-yt}f(z,t)\,\mathrm{d}t,
$$

for $y\in\mathbb{C}$, with Re $(y)>0$. This allows us to express the LT of $P(z,t)$, denoted by $P(z,y)^*$, as a function of π_{-1}^*, π_0^*, and π_n^*. The zeros of the denominator of $P(z,y)^*$ are next used to express π_0^*,

and π_n^* as functions of π_{-1}^*. Finally, the first line of Equation (1) leads to the expression of π_{-1}^*.

Theorem 1 *We have*

$$\pi_{x}* = \frac{(y(1-q)+\lambda(1-z_1))\pi_0^* - (1-q)(1-y\pi_{-1}^*)}{\gamma(z_2-z_1)z_1^x} \tag{2}$$
$$+ \frac{-(y(1-q)+\lambda(1-z_2))\pi_0^* + (1-q)(1-y\pi_{-1}^*)}{\gamma(z_2-z_1)z_2^x},\ \textit{for } 0 < x \leq n,\ \textit{with}$$

$$\pi_0^* = \frac{1}{y}\frac{\lambda\gamma[(1-z_1)z_1^n - (1-z_2)z_2^n]}{[(\gamma-\lambda)(y+\lambda)+\gamma\mu-\gamma(y+\mu+\gamma)z_1]z_1^n - [(\gamma-\lambda)(y+\lambda)+\gamma\mu-\gamma(y+\mu+\gamma)z_2]z_2^n},\ \textit{and} \tag{3}$$

$$\pi_{-1}^* = \frac{1}{y}\left(1 - \frac{\lambda[(\gamma(1-z_1)-\lambda)z_1^n - (\gamma(1-z_2)-\lambda)z_2^n]}{[(\gamma-\lambda)(y+\lambda)+\gamma\mu-\gamma(y+\mu+\gamma)z_1]z_1^n - [(\gamma-\lambda)(y+\lambda)+\gamma\mu-\gamma(y+\mu+\gamma)z_2]z_2^n}\right), \tag{4}$$

where

$$z_1 = \frac{1}{2\gamma}[y+\gamma+q(\mu+\gamma)+\sqrt{(y+\gamma+q(\mu+\gamma))^2-4\gamma q(y+\mu+\gamma)}],\ \textit{and,}$$
$$z_2 = \frac{1}{2\gamma}[y+\gamma+q(\mu+\gamma)-\sqrt{(y+\gamma+q(\mu+\gamma))^2-4\gamma q(y+\mu+\gamma)}].$$

Proof. In order to derive $P(z,t)$, we multiply the x^{th} differential equation in Equation (1) by z^x $(0 \leq x \leq n)$. We subsequently sum up over all x to obtain a single differential equation leading after some algebra to

$$\frac{\partial(P(z,t)+\pi_{-1})}{\partial t} = -(1-z)\left[\gamma+\mu\frac{q}{q-z}\right]P(z,t) + (1-z)\mu\frac{q}{q-z}P(q,t) + (\gamma-\lambda)(1-z)\pi_0 \tag{5}$$
$$+\gamma\frac{(1-z)(q^{n+1}-z^{n+1})}{q-z}\pi_n.$$

From the second line of Equation (1), we get

$$\mu P(q,t) = \frac{\partial(\pi_0+\pi_{-1})}{\partial t} + (\lambda+\mu)\pi_0 - \gamma q^n\pi_n.$$

By replacing the expression of $P(q,t)$ in Equation (5), we obtain

$$\frac{\partial(P(z,t)+\pi_{-1})}{\partial t} = -(1-z)\left[\gamma+\mu\frac{q}{q-z}\right]P(z,t) + \frac{q(1-z)}{q-z}\frac{\partial(\pi_0+\pi_{-1})}{\partial t} + \frac{(1-z)(q\mu+\lambda z+\gamma(q-z))}{q-z}\pi_0 - \gamma\frac{z^{n+1}(1-z)}{q-z}\pi_n. \tag{6}$$

Applying the LT to Equation (6) and using $P(z,0)=\pi_0(0)=0$ and $\pi_{-1}(0)=1$, we obtain

$$P(z,y)^* = -\frac{\gamma(1-z)z^{n+1}\pi_n^* + (1-z)((\gamma-\lambda)z - q(y+\mu+\gamma))\pi_0^* + (1-q)z(1-y\pi_{-1}^*)}{\gamma z^2 - (y+q(\mu+\gamma)+\gamma)z + q(y+\mu+\gamma)}. \tag{7}$$

The denominator of $P(z,y)^*$ is a quadratic in z. It has two zeros, z_1 and z_2, as defined in Theorem 1. The values z_1 and z_2 are also zeros of the numerator of $P(z,y)^*$. This can be seen by multiplying $P(z,y)^*$ by its denominator in Equation (7). Therefore, we deduce that

$$\gamma(1-z_i)z_i^{n+1}\pi_n^* + (1-z_i)(\gamma z_i^2 - (y+q(\mu+\gamma)+\lambda)z_i)\pi_0^* + (1-q)z_i(1-y\pi_{-1}^*) = 0,$$

for $i=1,2$. These two equations allow us to derive π_n^* and π_0^* as functions of π_{-1}^*. One then may write

$$\pi_0^* = \frac{1-y\pi_{-1}^*}{y}\frac{(1-z_1)z_1^n - (1-z_2)z_2^n}{(1-z_1-\lambda/\gamma)z_1^n - (1-z_2-\lambda/\gamma)z_2^n}, \text{ and,}$$
$$\pi_n^* = \frac{1-y\pi_{-1}^*}{y}\frac{\lambda(z_2-z_1)}{\gamma[(1-z_1-\lambda/\gamma)z_1^n - (1-z_2-\lambda/\gamma)z_2^n]}.$$

The LT of the first line of Equation (1) is $y\pi_{-1}^* - 1 = -\lambda\pi_{-1}^* + \mu\pi_0^*$. This equation together with the expression of π_0^* given above leads to π_{-1}^* and π_0^* as given in Theorem 1. There remains to determine the other probabilities (as functions of y) using $\pi_x^* = \frac{1}{x!}\frac{\partial^x P(z,y)^*}{\partial z^x}|_{z=0}$. We rewrite $P^*(z,y)$ as

$$P^*(z,y) = -\frac{(1-z)z^{n+1}}{(z-z_1)(z-z2)}\pi_n^* + \frac{\gamma-\lambda}{\gamma}\pi_0^* + \frac{-z_1(y(1-q)+\lambda(1-z_1))\pi_0^* + (1-q)z_1(1-y\pi_{-1}^*)}{\gamma(z-z_1)(z_2-z_1)} + \frac{z_2(y(1-q)+\lambda(1-z_2))\pi_0^* - (1-q)z_2(1-y\pi_{-1}^*)}{\gamma(z-z_2)(z_2-z_1)}.$$

The x^{th} derivative of the term proportional with π_n^* evaluated in $z=0$ is equal to zero for $x \le n$ because $z_i \neq 0$, for $i=1,2$ and the x^{th} derivative of z^{n+1} at $z=0$ is zero for $x \le n$. Using

$\frac{\partial^x (z-z_i)^{-1}}{\partial z^x}|_{z=0} = -\frac{x!}{z_i^{x+1}}$, we deduce the expression of π_x^*, for $0 < x \leq n$. □

4 Performance measures of the real system

The LT of the performance measures for the real system are obtained in Theorem 2 by letting γ tend to infinity. We consider the probability of an empty system (or an idle server), π_{-1}, the proportion of customers who are rejected, denoted by P_R, the expected waiting time, denoted by $E(W)$, and the probability of waiting longer than a time threshold w, denoted by $P(W > w)$, with $0 < w < \tau$.

Theorem 2 *We have*

$$\pi_{-1}^* = \frac{1}{y}\left(1 - \frac{\lambda\left(y_1 - y_2 e^{-(y_1-y_2)\tau}\right)}{y_1(\lambda+\mu+y) - \lambda\mu - (y_2(\lambda+\mu+y) - \lambda\mu)e^{-(y_1-y_2)\tau}}\right),$$

$$P_R^* = \frac{e^{-\tau(y_1-\lambda)}}{y}\frac{\lambda(y_1-y_2)}{y_1(\lambda+\mu+y) - \lambda\mu - (y_2(\lambda+\mu+y) - \lambda\mu)e^{-(y_1-y_2)\tau}},$$

$$E(W)^* = \frac{\mu}{y^2}\frac{y_1 - (y+\mu) - (y_2 - (y+\mu))e^{-(y_1-y_2)\tau} + (\tau\lambda y/\mu - 1)(y_1-y_2)e^{-\tau(y_1-\lambda)}}{y_1(\lambda+\mu+y) - \lambda\mu - (y_2(\lambda+\mu+y) - \lambda\mu)e^{-(y_1-y_2)\tau}}, \text{ and,}$$

$$\begin{aligned} P(W > w)^* = {} & \frac{-y_1 + y + \mu + \lambda + \pi_{-1}^*((y+\lambda)(y_1 - y - \lambda) - y\mu)}{(y_1-\lambda)(y_1-y_2)}(e^{(\lambda-y_1)w} - e^{(\lambda-y_1)\tau}) \\ & - \frac{-y_2 + y + \mu + \lambda + \pi_{-1}^*((y+\lambda)(y_2 - y - \lambda) - y\mu)}{(y_2-\lambda)(y_1-y_2)}(e^{(\lambda-y_2)w} - e^{(\lambda-y_2)\tau}) \\ & + \frac{e^{-\tau(y_1-\lambda)}}{y}\frac{\lambda(y_1-y_2)}{y_1(\lambda+\mu+y) - \lambda\mu - (y_2(\lambda+\mu+y) - \lambda\mu)e^{-(y_1-y_2)\tau}}. \end{aligned}$$

Proof. Using a Taylor expansion of z_1, z_2 and z_i^n, as γ tends to infinity, we get

$$\begin{aligned} z_1 &= 1 + \frac{1}{2\gamma}\left(y + \mu - \lambda + \sqrt{(y+\lambda+\mu)^2 - 4\lambda\mu}\right) + o(1/\gamma) = 1 + \frac{y_1-\lambda}{\gamma} + o(1/\gamma), \\ z_2 &= 1 + \frac{1}{2\gamma}\left(y + \mu - \lambda - \sqrt{(y+\lambda+\mu)^2 - 4\lambda\mu}\right) + o(1/\gamma) = 1 + \frac{y_2-\lambda}{\gamma} + o(1/\gamma), \text{ and,} \\ z_i^n &= e^{(y_i-\lambda)\tau} + o(1/\gamma), \text{ for } i = 1, 2. \end{aligned}$$

We observe that y_1/λ and y_2/λ are the roots of the denominator of the Laplace transform of the generating function in an M/M/1 queue (e.g., see [17], Equation (2.57), p.99). This directly leads to the expression of π_{-1}^*.

At time t, a customer can be rejected only from state $x = n$. The probability to be in state

n is π_n, the number of rejected customers from this state during an interval of time $\mathrm{d}t$ is $\gamma \mathrm{d}t$, the number of arrivals during the same interval is $\lambda \mathrm{d}t$. Therefore, the proportion of customers who are rejected from the system can be obtained as $P_R = \lim_{\gamma \longrightarrow \infty} \frac{\gamma}{\lambda} \pi_n$.

Let us now consider the performance related to the waiting time in the queue. The embedded Markov chain at service initiations and rejection times is considered. In this way, we consider the virtual waiting time of a customer who would initiate a service or would be rejected at time t. Service initiations occur at μ-transitions from states $0 < x \leq n$. Rejections occur at γ-transitions from state $x = n$. The expected duration of a waiting phase is $1/\gamma$. Therefore, the virtual expected waiting time of served or rejected customers at time t is

$$E(W) = \lim_{\gamma \longrightarrow \infty} \left(\frac{\mu}{\lambda} \sum_{x=1}^{n} \frac{x}{\gamma} \pi_x + \frac{\gamma}{\lambda} \frac{n}{\gamma} \pi_n \right) = \lim_{\gamma \longrightarrow \infty} \left(\frac{\mu}{\lambda} \frac{\frac{\partial P(z,t)}{\partial z}|_{z=1}}{\gamma} + \frac{\gamma}{\lambda} \frac{n}{\gamma} \pi_n \right).$$

We therefore deduce the LT of the expected waiting time from

$$E(W)^* = \lim_{\gamma \longrightarrow \infty} \left(\frac{\mu}{\lambda \gamma} \frac{\partial P^*(z,y)}{\partial z}|_{z=1} + \tau \frac{\gamma}{\lambda} \pi_n^* \right).$$

We now consider the probability of waiting more than a time threshold w such that $0 < w < \tau$ irrespective if the customer is rejected or served; $P(W > w)$. We can decompose this probability depending if a customer is served or rejected;

$$P(W > w) = (1 - P_R) P(W > w | \text{Service}) + P_R P(W > w | \text{Rejection}).$$

Since rejections only occur after τ time units and $w < \tau$, we have $P(W > w|\text{Rejection}) = 1$. Let us now focus on served customers. Consider a customer served from waiting phase x ($0 < x \leq n$). This customer has stayed in the queue during x γ-phases. The probability that an Erlang distribution with x phases and rate γ per phase exceeds w is $e^{-\gamma w} \sum_{k=0}^{x-1} \frac{(\gamma w)^k}{k!}$. Therefore,

$$(1-P_R) P(W > w|\text{Service}) = \lim_{\gamma \longrightarrow \infty} \left(\frac{\mu}{\lambda} \sum_{x=1}^{n} \pi_x e^{-\gamma w} \sum_{k=0}^{x-1} \frac{(\gamma w)^k}{k!} \right) = \lim_{\gamma \longrightarrow \infty} \left(\frac{\mu e^{-\gamma w}}{\lambda} \sum_{x=0}^{n-1} \frac{(\gamma w)^x}{x!} \sum_{k=x+1}^{n} \pi_k \right).$$

From Equation (2), we observe that $\pi_x^* = \frac{A_1}{z_1^x} - \frac{A_2}{z_2^x}$, with $A_i = \frac{(y(1-q)+\lambda(1-z_i))\pi_0^* - (1-q)(1-y\pi_{-1}^*)}{\gamma(z_2 - z_1)}$,

for $i = 1, 2$. Moreover, $\sum_{k=x+1}^{n} \frac{1}{z_i^k} = \frac{z_i^{-x} - z_i^{-n}}{z_i - 1}$, for $i = 1, 2$. This leads to

$$((1 - P_R)P(W > w|\text{Service}))^* = \lim_{\gamma \longrightarrow \infty} \left(\frac{\mu e^{-\gamma w}}{\lambda} \sum_{i=1}^{2} (-1)^{i+1} \frac{A_i}{z_i - 1} \sum_{x=0}^{n-1} \left(\frac{(\gamma w / z_i)^x}{x!} - z_i^{-n} \frac{(\gamma w)^x}{x!} \right) \right).$$

By letting γ tend to infinity, we obtain $P(W > w)^*$. □

5 Comments and Numerical Illustration

The transient performance measures can be computed using a Laplace transform inversion. We use the speed up version of the Gaver-Stehfes algorithm presented in [12], page 144, equation (7.7), where a given function $f(t)$ is approximated by

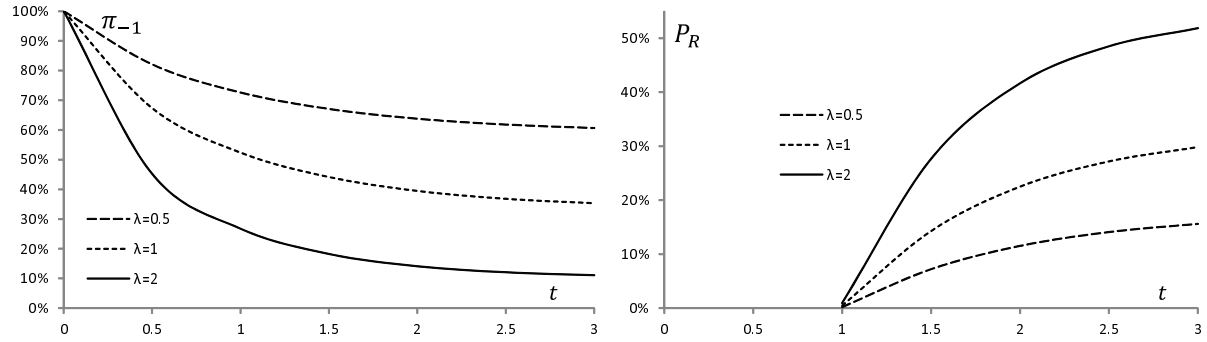
$$\frac{\ln(2)}{t} \sum_{n=1}^{N} K_n \cdot f^* \left(n \frac{\ln(2)}{t} \right),$$

where N is even and

$$K_n = (-1)^{n+\frac{N}{2}} \sum_{k=\left[\frac{n+1}{2}\right]}^{\min(n, N/2)} \frac{k^{N/2}(2k)!}{(N/2 - k)!k!(k-1)!(n-k)!(2k-n)!}.$$

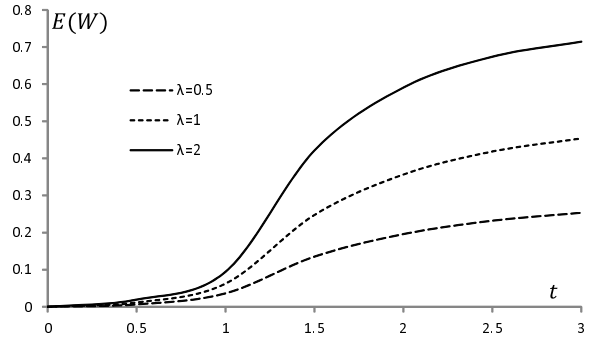
One difficulty to apply this formula in practice is to determine a sufficiently high value for N and a sufficiently high number of digits for the values of K_n in order to obtain a sufficiently accurate value for the function to invert. [1] investigated numerically the precision produced as a function of the parameter N. From extensive experimentation, they conclude that about $0.45 \times N$ significant digits are sufficient to obtain a relative error of the order of $10^{-0.45N}$. However, their result depends on the transform. In our case, the complexity of the formulas requires higher values for N in particular in the zone where the elapsed time since the origin is close to the rejection threshold. Alternative methods for numerical Laplace transform inversion can be found in [5].

In Figure 1, we derive the main performance measures as a function of the time elapsed since the origin. From this and other numerical experiments, we observe that the M/M/1+D queue reaches a close to stationary behavior quicker than the corresponding M/M/1 queue. The justification of this observation is related to the reduction of the waiting time variability with a low rejection threshold.

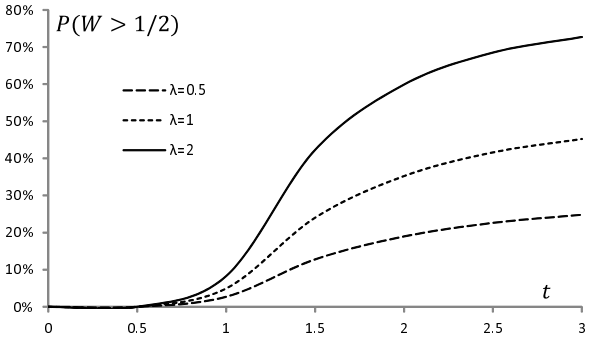


(a) π_{-1}

(b) P_R



(c) $E(W)$



(d) $P(W > 1/2)$

Figure 1: Numerical results ($\mu = 1$, $\tau = 1$)

The Laplace transforms of the performance measures allow us to compute their stationary expressions using the Final Value Theorem (e.g., see [12], Theorem 2.6, pages 40-41). These are obtained by computing the limit as y tends to 0 of the product of y with the LT of the wanted performance measure. This leads to $\pi_{-1}^{\infty} = \frac{1-a}{1-a^2e^{-\tau(\mu-\lambda)}}$, $P_R^{\infty} = \frac{a(1-a)e^{-\tau(\mu-\lambda)}}{1-a^2e^{-\tau(\mu-\lambda)}}$, $E(W)^{\infty} = \frac{1}{\mu}\frac{a(1-(1+a\tau(\mu-\lambda))e^{-\tau(\mu-\lambda)})}{(1-a)(1-a^2e^{-\tau(\mu-\lambda)})}$, and $P(W > w)^{\infty} = \frac{a(e^{-w(\mu-\lambda)}-ae^{-\tau(\mu-\lambda)})}{1-a^2e^{-\tau(\mu-\lambda)}}$, where $a = \lambda/\mu$.

6 Busy Period Analysis

A busy period is the time that elapses between two consecutive arrivals finding an empty system. In this section, we determine the mean and the LT of a busy period in an M/M/1+D queue. These results can be extended to the *full busy period* of the M/M/s+D queue defined as a period commencing when an arriving customer finds exactly one idle server and ending at the first departure epoch leaving behind exactly one idle server. We use the same state definition as in the previous sections. Let the random variable C_x be the time till the system is empty again if the FIL is in state x ($1 \leq x \leq n$) or if only one customer is in the system ($x = 0$). Since a busy period starts

when the first customer after an idle period arrives and it ends when the system is empty again, C_0 is the length of a busy period. Using the transition structure defined in Section 2, we get

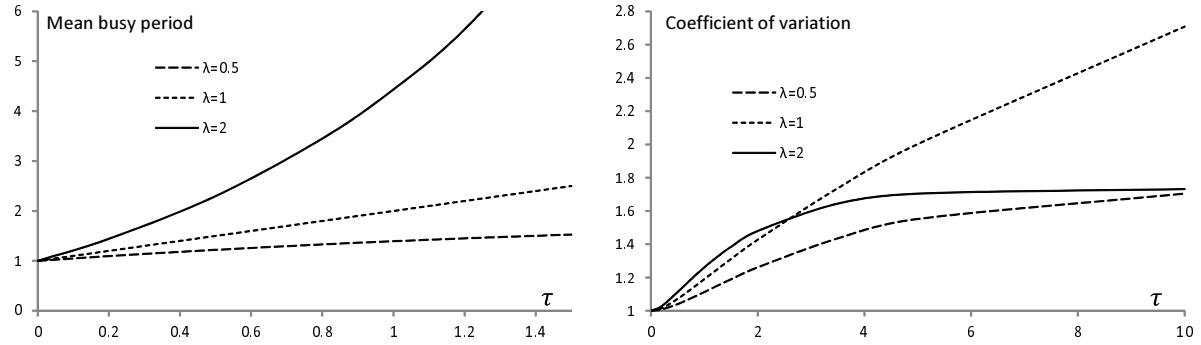
$$
\begin{aligned}
&C_0^*(y+\mu+\lambda) = \lambda C_1^* + \mu, \qquad (8)\\
&C_x^*(y+\mu+\gamma) = \gamma C_{x+1}^* + \mu q^x C_0^* + \mu(1-q)\sum_{k=1}^{x} q^{x-k} C_k^*, \text{ for } 1 \le x < n,\\
&C_n^*(y+\mu+\gamma) = (\mu+\gamma) q^n C_0^* + (\mu+\gamma)(1-q)\sum_{k=1}^{n} q^{n-k} C_k^*,
\end{aligned}
$$

In Theorem 3, we give the LT of C_0, denoted by C_0^*, the expected duration of the busy period, denoted by $E(C_0)$ and the variance of the busy period, denoted by $V(C_0)$. The proof follows a similar approach as the analysis of Section 3. The details are given in the Online Supplement.

Theorem 3 *We have*

$$
\begin{aligned}
C_0^* &= \mu\left[\lambda\left(1 - \frac{\mu(y_2-\mu+(\mu-y_1)e^{-\tau(y_1-y_2)})}{y_1(y_2-\mu)+y_2(\mu-y_1)e^{-\tau(y_1-y_2)}}\right)+\mu+s\right]^{-1},\\
E(C_0) &= \frac{\mu-\lambda e^{-\tau(\mu-\lambda)}}{\mu(\mu-\lambda)}, \text{ and,}\\
V(C_0) &= \frac{\mu^2(\mu+\lambda) - 2\lambda\mu(\mu-\lambda)(2+(\lambda+\mu)\tau)e^{-\tau(\mu-\lambda)} - \lambda^2(\mu+\lambda)e^{-2\tau(\mu-\lambda)}}{\mu^2(\mu-\lambda)^3}.
\end{aligned}
$$

In Figure 2, we illustrate the impact of the deterministic rejection threshold τ on the expected duration of the busy period and the coefficient of variation of the busy period (i.e., it is the ratio between the standard deviation and the expected duration of the busy period). As expected, these two measures increase with τ. As τ tends to infinity, we obtain the results for the M/M/1 queue. The expected duration tends to $\frac{1}{\mu-\lambda}$ if $\mu > \lambda$ and to infinity otherwise (instability). The coefficient of variation tends to $\sqrt{\frac{\lambda+\mu}{|\mu-\lambda|}}$ in all cases. This explains the relative position of the curves in Figure (2(b)) and provides an interesting property for the unstable M/M/1 queue. For low values of τ, the coefficient of variation increases with λ. For larger values of τ, either $\lambda < \mu$ and the system behaves close to a stable M/M/1 queue or $\lambda > \mu$ and most customers are rejected at τ time units. In both cases the coefficient of variation is controlled. The uncertainty on the duration of the busy period is maximized when $\lambda = \mu$. This may explain the relative position of this curve compared to the others.



(a) Expected duration of the busy period

(b) Coefficient of variation of the busy period

Figure 2: Numerical results ($\mu = 1$)

References

[1] J. Abate and P. Valkó. Multi-precision Laplace transform inversion. *International Journal for Numerical Methods in Engineering*, 60(5):979–993, 2004.

[2] J. Abate and W. Whitt. Transient behavior of the M/M/1 queue: Starting at the origin. *Queueing Systems*, 2(1):41–65, 1987.

[3] J. Abate and W. Whitt. Transient behavior of the M/M/1 queue via Laplace transforms. *Advances in Applied Probability*, 20(1):145–178, 1988.

[4] J. Abate and W. Whitt. Transient behavior of the M/G/1 workload process. *Operations Research*, 42(4):750–764, 1994.

[5] J. Abate and W. Whitt. A unified framework for numerically inverting Laplace transforms. *INFORMS Journal on Computing*, 18(4):408–421, 2006.

[6] R. Al-Seedy, A. El-Sherbiny, S. El-Shehawy, and S. Ammar. Transient solution of the M/M/c queue with balking and reneging. *Computers & Mathematics with Applications*, 57(8):1280–1285, 2009.

[7] S.I. Ammar, M.M. Helan, and F.T. Al Amri. The busy period of an M/M/1 queue with balking and reneging. *Applied Mathematical Modelling*, 37(22):9223–9229, 2013.

[8] M. Armony, N. Shimkin, and W. Whitt. The impact of delay announcements in many-server queues with abandonment. *Operations Research*, 57(1):66–81, 2009.

[9] R. Atar, C. Giat, and N. Shimkin. The cμ/θ rule for many-server queues with abandonment. *Operations Research*, 58(5):1427–1439, 2010.

[10] N.T.J. Bailey. A continuous time treatment of a simple queue using generating functions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 288–291, 1954.

[11] D.G. Champernowne. An elementary method of solution of the queueing problem with a single server and constant parameters. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 125–128, 1956.

[12] A. Cohen. *Numerical methods for Laplace transform inversion*, volume 5. Springer Science & Business Media, 2007.

[13] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

[14] J. Garcia, O. Brun, and D. Gauchard. Transient analytical solution of M/D/1/N queues. *Journal of Applied Probability*, 39(4):853–864, 2002.

[15] L.V. Green, P.J. Kolesar, and J. Soares. Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research*, 49(4):549–564, 2001.

[16] J. Griffiths, G.M. Leonenko, and J. Williams. The transient solution to M/Ek/1 queue. *Operations Research Letters*, 34(3):349–354, 2006.

[17] D. Gross and C.M. Harris. *Fundamentals of Queueing Theory*. Wiley series in probability and mathematical statistics, 1985. 2rd Edition.

[18] A. Jean-Marie and P. Robert. On the transient behavior of the processor sharing queue. *Queueing Systems*, 17(1-2):129–136, 1994.

[19] W.D. Kelton and A.M. Law. The transient behavior of the M/M/s queue, with implications for steady-state simulation. *Operations Research*, 33(2):378–396, 1985.

[20] W.M. Kempa. Transient workload distribution in the M/G/1 finite-buffer queue with single and multiple vacations. *Annals of Operations Research*, 239(2):381–400, 2016.

[21] M. Kitaev. The M/G/1 processor-sharing model: transient behavior. *Queueing Systems*, 14(3-4):239–273, 1993.

[22] G. Koole, B. Nielsen, and T. Nielsen. First in line waiting times as a tool for analysing queueing systems. *Operations Research*, 60(5):1258–1266, 2012.

[23] W. Ledermann and G. Edzard H. Reuter. Spectral theory for the differential equations of simple birth and death processes. *Philosophical Transactions of the Royal Society A*, 246(914):321–369, 1954.

[24] B. Legros, O. Jouini, and G. Koole. Optimal scheduling in call centers with a callback option. *Performance Evaluation*, 95:1–40, 2016.

[25] B. Legros, O. Jouini, and G. Koole. A uniformization approach for the dynamic control of queueing systems with abandonments. *Operations Research*, 66(1):200–209, 2017.

[26] P. Parthasarathy and M. Sharafali. Transient solution to the many-server poisson queue: a simple approach. *Journal of applied probability*, 26(3):584–594, 1989.

[27] J. Reed and A. Ward. Approximating the GI/GI/1+ GI queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Mathematics of Operations Research*, 33(3):606–644, 2008.

[28] O. Sharma and U. Gupta. Transient behaviour of an M/M/1/N queue. *Stochastic Processes and their Applications*, 13(3):327–331, 1982.

[29] L. Takaćs. The time dependence of a single-server queue with poisson input and general service times. *The Annals of Mathematical Statistics*, 33(4):1340–1348, 1962.

[30] M. Van de Coevering. Computing transient performance measures for the M/M/1 queue. *OR-Spektrum*, 17(1):19–22, 1995.

[31] J. Wang, B. Liu, and J. Li. Transient analysis of an M/G/1 retrial queue subject to disasters and server failures. *European Journal of Operational Research*, 189(3):1118–1132, 2008.

[32] W. Xiong and T. Altiok. An approximation for multi-server queues with deterministic reneging times. *Annals of Operations Research*, 172(1):143–151, 2009.