Busy Period Analysis for $M/PH/1$ Queues with Workload Dependent Balking

Abstract

We consider an $M/PH/1$ queue with balking based on the workload. An arriving customer joins the queue and stays until served only if the system workload is below a fixed level at the time of arrival. We begin by considering a fluid model where the buffer content changes at a rate determined by an external stochastic process with finite state space. We derive systems of first-order linear different equations for both the mean and LST (Laplace-Stieltjes Transform) of the busy period in this model and solve explicitly. We obtain the mean and LST of the busy period in the $M/PH/1$ queue with workload dependent balking as a special limiting case of this fluid model. We illustrate the results with numerical examples.

1 Introduction

The balking queueing models incorporate the characteristics of impatience of the customer or a specific acceptance control policy. The service requirement of an arriving customer may not be (completely) accepted when the system is considerably congested. One natural measurement of the congestion of the system at time t is the workload, which is defined as the time it takes the server to empty the system, provided there are no arrivals after t. Various workload dependent balking queues can be found under the title "finite workload capacity", "dams" or "workload dependent arrival rate" (cf. Prabhu [23], Perry et al. [22], Perry and Stadje [20], Bekker et al. [3] and Bekker [2]). The workload dependent balking queues also have found applications in the study of clearing models (cf. Boxma et al. [5]).

In this paper, we consider queues with workload dependent balking that operates as follows: an arriving customer does not join the queue if and only if he/she sees the workload is at least a fixed amount b. We assume that once a customer decides

to join the queue, he/she stays in the system until service completion. We study the busy period which is the time until the system is empty. Precise formulation is given in Section 2.

If the service discipline is FCFS, then the workload ahead of an entering customer is exactly the time he/she waits before the service starts. The balking rule mentioned above is very natural in situations where the customers are impatient. For example, in call centers, typically a call-in customer who cannot be answered immediately is told how long a wait he/she faces before an operator is available. Then the customer can choose to wait (join the queue) or hang up (leave). The significance of customer impatience in practice and modeling of call centers has drawn much attention recently (cf. Koole and Mandelbaum [13], Garnett et al. [8], Whitt [25]). This model incorporates the right characteristics of the customer behavior. On the other hand, in some systems (e.g. communication systems), the information about amount of workload is often precisely available. What's more, from the view of control problems, the balking rule is actually a threshold type of customer acceptance/rejection policy based on the system workload. Such a policy is shown to be optimal under certain conditions in Johansen and Stidham [12]. This threshold type policy generates the model considered here.

Workload dependent balking queues have been extensively studied in the literature. Focusing on the steady-state distribution of the workload process, early papers include Hu and Zazanis [11], Gavish and Schweitzer [9] and Hokstad [10]. Perry and Asmussen [19] deals with the most general type of workload dependent balking models with several variations. Liu and Kulkarni [18] obtains explicit formulas for the steady-state workload distributions in an $M/PH/1$ setting via differential equations.

In comparison to the work devoted to the study of the steady-state distribution, we find fewer papers treating the busy periods. Perry and Asmussen [19] find the LST of the busy period for exponential service times via both differential equations and martingales. Further they identify the busy period with a first passage time in an associate birth and death process by coupling argument and extend the results to the case where b is not fixed but a random variable with exponential distribution. Perry et al. [21] gives closed-form expressions for the LST of the busy periods for $M/G/1$ queues with workload dependent balking as functions of the LSTs of certain stopping times. The paper further illustrates the duality between $M/G/1$ and $G/M/1$ and give similar formula for $G/M/1$ queues with workload dependent balking in terms of the same LSTs of the stopping times used in the $M/G/1$ formulas.

We use an alternative method to solve the first passage time problem via a fluid model whose net flow rate is governed by a stochastic process. We construct such a fluid model so that the $M/PH/1$ balking model is a limiting case of this fluid

model. For a general discussion of such a fluid model, we refer to the survey paper by Kulkarni [15]. Various methods are used to study the mean first passage times in fluid models (cf. Asmussen and Bladt [1], Chen and Samalam [6], Kulkarni and Narayanan [16], Boxma and Dumas [4] and Kulkarni and Tzenova [17]). Kulkarni and Tzenova [17] develops a differential-equation method which is more amendable for numerical algorithms. We slightly extend the method and apply it to obtain the mean and LST of the first passage time of the fluid model considered in this paper. The precise formulation of the fluid model is given in Section 3.

The rest of the paper is outlined as follows. In Section 2, we describe the balking queueing model and define the first passage time. We formulate the fluid model in Section 3 and give a general treatment of the first passage time of the fluid model in Section 4. In Section 5, we consider a special case of the fluid model and give the mean and LST in Theorem 4 and 5. In Section 6, we illustrate a construction to show that the balking model is a limiting case of the fluid model analyzed in Section 5. Then we take the limit of the results given in Theorem 4 and 5 and obtain explicit formulas for the mean and LST of the first passage time for the balking model (Theorem 6 and 7). In Section 7, we illustrate the usage of these formulas to compute the first passage time in the $M/M/1$ case and verify several known results. We present several numerical examples in Section 8.

2 The Balking Queueing Model

Here we consider an $M/PH/1$ queueing system where customers arrive according to a PP(λ) with iid phase type service requirements. Let S be a generic random variable representing the service time. We assume:

$$\Pr\{S > x\} = \alpha e^{Mx}\vec{1},$$

where $\vec{1}$ is a $n \times 1$ column vector with all coordinates equal to 1, the row vector $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$ has non-negative components and $\alpha\vec{1} = 1$, M is an $n \times n$ submatrix of the generator of an irreducible CTMC, with the following properties:

- M is invertible;
- M is diagonally dominant with all diagonal elements negative.

It is known that the above properties imply that all eigenvalues of M have negative real part.

Let $W(t)$ be the workload at time t which is defined as the time it takes the server to empty the system, provided there are no arrivals after t. Let $\{W(t), t \geq 0\}$ be the

workload process. In this paper we assume the following workload dependent balking model: an arriving customer at time t enters the system iff $W(t-) < b$, where b is a nonnegative constant.

Let $B = \min\{t \geq 0 : W(t) = 0\}$, i.e., the first passage time until the system is empty. In this paper, we compute the mean

$$m(x) = \mathrm{E}[B|W(0) = x] \tag{1}$$

and the LST

$$\psi(s, x) = \mathrm{E}[e^{-sB}|W(0) = x]. \tag{2}$$

We begin by considering a fluid model where the buffer content changes at a rate determined by an external stochastic process with finite state space. Later we shall show that the balking queueing model described above can be regarded as limiting case of a special case of such a fluid model.

3 The Fluid Model

Consider a fluid model with an infinite capacity buffer where the net flow rate is governed by a stochastic process $\{Z(t), t \geq 0\}$ on a finite state space $\mathcal{S} = \{0, 1, \cdots, n\}$ as follows: if $Z(t) = i$ the buffer level changes at rate r_i. For convenience, we assume that $r_i \neq 0, \forall i \in \mathcal{S}$. It will be clear later that this assumption is without loss of generality in the context of our application. Let $\mathcal{S}_- = \{i : r_i < 0\}$, $\mathcal{S}_+ = \{i : r_i > 0\}$, $n_- = |\mathcal{S}_-|$ and $n_+ = |\mathcal{S}_+|$.

Let $X(t)$ be the amount of fluid in the buffer. The dynamics of the buffer content process $X = \{X(t), t \geq 0\}$ is described by the following differential equation:

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = \begin{cases} r_i & \text{if} \quad Z(t) = i, X(t) > 0, \\ \max(r_i, 0) & \text{if} \quad Z(t) = i, X(t) = 0. \end{cases}$$

The X process is called a *fluid input-output process driven by the Z process* (cf. Kulkarni [15]). The driving process Z we consider here behaves, loosely speaking, as a CTMC whose infinitesimal generator matrix $Q(y)$ depends on the current buffer level y as follows:

$$Q(y) = \begin{cases} Q & \text{if} \quad y < b, \\ \bar{Q} & \text{if} \quad y \geq b, \end{cases}$$

where $Q = [q_{ij}]$ ($\bar{Q} = [\bar{q}_{ij}]$) is the generator of a CTMC. We assume that Q and $\bar{Q}$ have exactly one irreducible class. Such CTMCs have unique limiting distributions that are independent of their initial states.

Such a model is also known as a *feedback fluid queue*. A more rigorous treatment of such a model can be found in Scheinhardt et al. [24].

Clearly the joint process $\{(X(t), Z(t)), t \geq 0\}$ is a Markov process which is characterized by the matrices Q, $\bar{Q}$ and the diagonal matrix with net input rates $R = \text{diag}(r_0, r_1, \cdots, r_n)$. Let $p = [p_0, p_1, \cdots, p_n]$ be the solution to

$$pQ = 0, \quad p\vec{1} = 1$$

and $\bar{p} = [\bar{p}_0, \bar{p}_1, \cdots, \bar{p}_n]$ be the solution to

$$\bar{p}\bar{Q} = 0, \quad \bar{p}\vec{1} = 1.$$

Let

$$d = \sum_{i \in \mathcal{S}} p_i r_i$$

and

$$\bar{d} = \sum_{i \in \mathcal{S}} \bar{p}_i r_i.$$

It is easy to see that $\bar{d} < 0$ is a necessary condition for the joint process to be stable (cf. Kulkarni [15]). We assume this condition holds, i.e., the joint process is stable.

4 First Passage Time: the Fluid Model

In this section, we compute the mean and LST of the first passage time

$$T = \min\{t \geq 0 : X(t) = 0\}.$$

First we compute the following mean of T:

$$\pi_i(x) = \text{E}[T | Z(0) = i, X(0) = x]$$

and let

$$\pi(x) = [\pi_0(x), \pi_1(x), \cdots, \pi_n(x)]^{\text{T}}.$$

Theorem 1 *The vector $\pi(x)$ satisfies the following system of differential equation*

$$R\pi'(x) = \begin{cases} -\vec{1} - Q\pi(x), & 0 < x < b, \\ -\vec{1} - \bar{Q}\pi(x), & x > b, \end{cases} \tag{3}$$

with boundary conditions:

$$\begin{aligned} \pi_i(0) &= 0, \ i \in \mathcal{S}_-, \\ \pi(b-) &= \pi(b+). \end{aligned} \tag{4}$$

Proof: For $0 < x < b$, consider an infinitesimal time interval $[0, h]$ and use fist step analysis to obtain:

$$\pi_i(x) = h + (1 + q_{ii}h)\pi_i(x + r_i h) + \sum_{j \neq i} q_{ij} h \pi_j(x + r_i h), \ \forall i.$$

Divide both sides by h and let $h \to 0$. After some algebra, we get:

$$-r_i \pi_i'(x) = 1 + \sum_j q_{ij} \pi_j(x). \tag{5}$$

The system of differential equations is obtained by rearranging the terms and writing the $(n+1)$ equations in a matrix form. Same argument goes for the case $x > b$. The boundary conditions are obvious. ■

Notice that Equation (5) shows that it is possible to eliminate $\pi_j(x)$ if $r_j = 0$. Hence assuming $r_i \neq 0, \ \forall i \in \mathcal{S}$ is not a severe restriction.

Similar equations are derived and solved in Kulkarni and Tzenova [17] by using well-known techniques. We slightly extend the method that is used in Theorem 3.3 in Kulkarni and Tzenova [17] and apply it to the problem we consider here. First we need the following lemma.

Lemma 1 (From Theorem 11.5, Kulkarni [15]). *Suppose $\bar{Q}$ has exactly one irreducible class. There are $n+1$ (possibly repeated) eigenvalues of $-R^{-1}\bar{Q}$. When $\bar{d} < 0$, exactly n_+ have positive real parts, one is zero, and $n_- - 1$ have negative real parts.*

Denote the eigenvalues $\bar{\theta}_i$ as follows:

$$Re(\bar{\theta}_0) \leq Re(\bar{\theta}_1) \leq \cdots \leq Re(\bar{\theta}_{n_- - 2}) \leq Re(\bar{\theta}_{n_- - 1}) = 0 < Re(\bar{\theta}_{n_-}) \leq \cdots \leq Re(\bar{\theta}_n).$$

Let $\bar{v}_i$ be the right eigenvector corresponding to eigenvalue $\bar{\theta}_i$.

Similarly, we use the notation θ_i and v_i $(0, 1, ..., n)$, respectively, for the eigenvalues and right eigenvectors of $-R^{-1}Q$. But we do not order θ_i in the same fashion as $\bar{\theta}_i$, and we do not assume $d < 0$.

We give the main result in the following theorem. The proof is similar to that of Theorem 3.3 in Kulkarni and Tzenova [17] and is omitted here.

Theorem 2 *Let I be the $(n+1) \times (n+1)$ identity matrix. The solution to Equation (3) and (4) is given by:*

$$\pi(x) = \begin{cases} \sum_{j=0}^{n} a_j v_j e^{\theta_j x} - \frac{\vec{1}x}{d} + c, & 0 \leq x \leq b, \\ \sum_{j=0}^{n_- - 1} \bar{a}_j \bar{v}_j e^{\bar{\theta}_j x} - \frac{\vec{1}x}{\bar{d}} + \bar{c}, & x > b, \end{cases} \tag{6}$$

where c is any solution to the linear system

$$Qc = (R/d - I)\vec{1},$$

$\bar{c}$ is any solution to the linear system

$$\bar{Q}\bar{c} = (R/\bar{d} - I)\vec{1},$$

and the coefficients $\{a_j, 0 \le j \le n\}$ and $\{\bar{a}_j : 0 \le j \le n_- - 1\}$ are obtained by using the boundary conditions given in Equation (4).

Next, we compute the LST of the first passage time T defined as:

$$\phi_i(s, x) = \mathrm{E}[e^{-sT} | Z(0) = i, X(0) = x].$$

Let

$$\phi(s, x) = [\phi_0(s, x), \phi_1(s, x), \cdots, \phi_n(s, x)]^{\mathrm{T}}.$$

Similar to Theorem 1, we have the following theorem.

Theorem 3 *The vector $\phi(s, x)$ satisfies the following system of differential equation*

$$R\frac{\mathrm{d}\phi(s, x)}{\mathrm{d}x} = \begin{cases} (sI - Q)\phi(s, x), & 0 < x < b, \\ (sI - \bar{Q})\phi(s, x), & x > b, \end{cases} \tag{7}$$

with boundary conditions:

$$\begin{array}{c} \phi_i(s, 0) = 1, \ i \in \mathcal{S}_-, \\ \phi(s, b-) = \phi(s, b+). \end{array} \tag{8}$$

Proof: For $0 < x < b$, consider an infinitesimal time interval $[0, h]$ and use first step analysis to obtain:

$$\phi_i(s, x) = e^{-sh}(1 + q_{ii}h)\phi_i(s, x + r_i h) + \sum_{j \neq i} e^{-sh} q_{ij} h \phi_j(s, x + r_i h), \ \forall i.$$

Divide both sides by h and let $h \to 0$. After some algebra, we get:

$$-r_i \frac{\mathrm{d}\phi_i(s, x)}{\mathrm{d}x} + s\phi_i(s, x) = \sum_j q_{ij}\phi_j(s, x), \ \forall i.$$

The system of differential equations is obtained by rearranging the terms and writing the $(n+1)$ equations in a matrix form. Same argument goes for the case $x > b$. The boundary conditions are obvious. ■

It is more complicated to solve Equations (7) and (8). However, we shall see in the next section that for some special cases, explicit solutions can be obtained.

5 A Special Case of the Fluid Model

In this section we consider a special case of the fluid model whose R, Q and $\bar{Q}$ matrices are as given below. As we shall see, this helps us solve the first passage time problem for the balking model we describe at the beginning of this paper. Let M and α be the parameters of the phase type distribution and λ be the arrival rate as defined in Section 2. The model is parameterized by a real number $r > 0$. Suppose

$$\begin{gathered} r_0 = -1, r_i = r > 0 \quad \text{for } \ i = 1, 2, \cdots, n, \\ R = \text{diag}(r_0, r_1, \cdots, r_n), \end{gathered} \tag{9}$$

$$Q = \begin{bmatrix} -\lambda & \lambda\alpha \\ -Mr\vec{1} & Mr \end{bmatrix}, \tag{10}$$

$$\bar{Q} = \begin{bmatrix} 0 & 0 \\ -Mr\vec{1} & Mr \end{bmatrix}. \tag{11}$$

To understand the motivation behind this model consider two cases.

Case 1: The buffer content is less than b. Then the buffer content increases at rate r as long as the Z process is in the set $\{1, 2, .., n\}$ (we say it is "up"), and it decreases at rate 1 when the Z process is in state 0 (we say it is "down"). The Z process alternates between up and down periods. The down times are iid $\exp(\lambda)$ random variables, and the up times are iid with phase type distribution with parameters α and M.

Case 2: The buffer content is at least b. Then the buffer content increases at rate r as long as the Z process is up, and it decreases at rate 1 when the Z process is down. Once the Z process is down, it stays down until the buffer content drops below b and we switch to case 1 above.

Let $X_r(t)$ be the buffer content at time t in the fluid process described by the parameters Q, $\bar{Q}$, R given above. A typical sample path of the $\{X_r(t), t \geq 0\}$ process is shown in Figure 1.

Now define some special notation for this special case as follows:

$$T_r = \min\{t \geq 0 : X_r(t) = 0\},$$

$$\pi_i(x, r) = \text{E}[T_r | Z(0) = i, X(0) = x],$$
$$\pi(x, r) = [\pi_0(x, r), \pi_1(x, r), \cdots, \pi_n(x, r)]^{\text{T}},$$
$$\phi_i(s, x, r) = \text{E}[e^{-sT_r} | Z(0) = i, X(0) = x],$$
$$\phi(s, x, r) = [\phi_0(s, x, r), \phi_1(s, x, r), \cdots, \phi_n(s, x, r)]^{\text{T}},$$

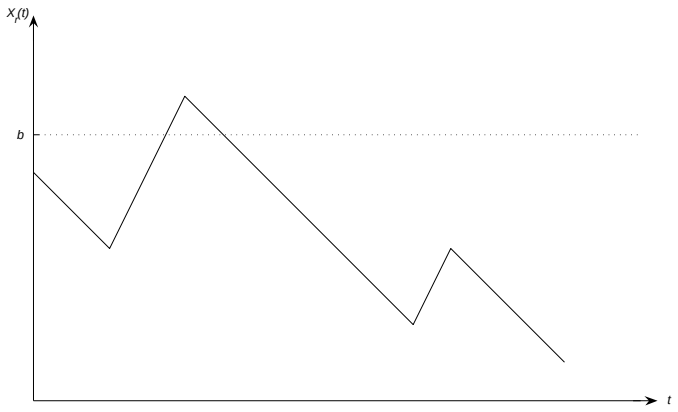

Figure 1: A Typical Sample Path of $X_r(t)$

$$A(s,r) = R^{-1}(sI - Q) = \begin{bmatrix} -\lambda - s & \lambda\alpha \\ M\vec{1} & -M + \frac{s}{r}I \end{bmatrix},$$

and

$$\bar{A}(s,r) = R^{-1}(sI - \bar{Q}) = \begin{bmatrix} -s & 0 \\ M\vec{1} & -M + \frac{s}{r}I \end{bmatrix}.$$

First we give an explicit formula for the mean first passage time $\pi(x,r)$ in the following theorem.

Theorem 4 *For R, Q and $\bar{Q}$ as specified in Equation (9), (10) and (11), respectively, the solution to Equation (3) and (4) is*

$$\pi(x,r) = \begin{cases} e^{A(0,r)x}(c - \int_0^x e^{-A(0,r)t}R^{-1}\vec{1}\mathrm{d}t), & 0 \le x \le b, \\ (x-b)\vec{1} + \pi(b,r), & x > b, \end{cases}$$

where

$$c = \begin{bmatrix} u_1 \\ \left[M\vec{1}, -M\right]e^{A(0,r)b} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ (1+1/r)\vec{1} + \left[M\vec{1}, -M\right]e^{A(0,r)b}\int_0^b e^{-A(0,r)t}R^{-1}\vec{1}\mathrm{d}t \end{bmatrix},$$

and u_1 is the first row of the identity matrix.

Proof: Substituting Q and $\bar{Q}$ in Equation (3), we get:

$$\frac{\mathrm{d}\pi(x,r)}{\mathrm{d}x} = -R^{-1}\vec{1} + A(0,r)\pi(x,r), \quad 0 < x < b, \tag{12a}$$

$$\frac{\mathrm{d}\pi(x,r)}{\mathrm{d}x} = -R^{-1}\vec{1} + \bar{A}(0,r)\pi(x,r), \quad x > b. \tag{12b}$$

It is well-known (cf. Finizo and Ladas [7]) that the solution to Equation (12a) is

$$\pi(x,r) = e^{A(0,r)x}(c - \int_0^x e^{-A(0,r)t}R^{-1}\vec{1}\mathrm{d}t),\ 0 < x < b, \tag{13}$$

where c is some constant vector to be determined. Notice that for $x > b$:

$$\pi_i(x,r) = \mathrm{E}[T_r|Z(0) = i, X(0) = x] = \mathrm{E}[x-b+T_r|Z(0) = i, X(0) = b] = x-b+\pi_i(b,r).$$

Then

$$\frac{\mathrm{d}\pi(x,r)}{\mathrm{d}x} = \frac{\mathrm{d}[(x-b)\vec{1} + \pi(b,r)]}{\mathrm{d}x} = \vec{1},\ x > b. \tag{14}$$

Consider the value $\lim_{x\downarrow b}\frac{\mathrm{d}\pi(x,r)}{\mathrm{d}x}$. From Eqution (14) and (12b), use the fact $\pi(b-,r) = \pi(b+,r)$, we get:

$$-R^{-1}\vec{1} + \bar{A}(0,r)\pi(b-,r) = \vec{1},$$

which reduces to:

$$[M\vec{1}, -M]\pi(b-,r) = (1+1/r)\vec{1}.$$

Using Equation (13) for $\pi(b-,r)$ we get:

$$[M\vec{1}, -M]e^{A(0,r)b}(c - \int_0^b e^{-A(0,r)t}R^{-1}\vec{1}\mathrm{d}t) = (1+1/r)\vec{1}.$$

Since $\pi_0(0,r) = 0$, the first component of c is 0. Writing this condition as an additional row of the equation above, we get:

$$\begin{bmatrix} u_1 \\ \big[M\vec{1}, -M\big]e^{A(0,r)b} \end{bmatrix} c = \begin{bmatrix} 0 \\ (1+1/r)\vec{1} + \big[M\vec{1}, -M\big]e^{A(0,r)b}\int_0^b e^{-A(0,r)t}R^{-1}\vec{1}\mathrm{d}t \end{bmatrix}.\blacksquare$$

Remark: Unfortunately, due to the singularity of $A(0,r)$, it is a little complicated to compute the integral $\int_0^x e^{-A(0,r)t}\mathrm{d}t$. There are several numerical methods for computing this integral. One is from the following observation. As we know $\pi(x,r) = \frac{\mathrm{d}\phi(s,x,r)}{\mathrm{d}s}|_{s=0}$, then

$$\int_0^x e^{-A(0,r)t}\mathrm{d}t = \lim_{s\to 0}\int_0^x e^{-A(s,r)t}\mathrm{d}t = \lim_{s\to 0}A^{-1}(s,r)(I - e^{-A(s,r)x}),$$

which is used in our numerical computations.

Next we compute an explicit formula for the LST of the first passage time $\phi(s,x,r)$. First we need the following lemma.

Lemma 2 *The matrix $\bar{A}(s,r)$ has an eigenvalue $-s$ with the corresponding right eigenvector:*

$$\bar{v}_0(r) = \begin{bmatrix} 1 \\ (M - s(1+1/r)I)^{-1}M\vec{1} \end{bmatrix}.$$

Proof: It is easy to verify that:

$$(-sI - \bar{A}(s,r)) \begin{bmatrix} 1 \\ (M - s(1+1/r)I)^{-1}M\vec{1} \end{bmatrix} = 0.\blacksquare$$

The explicit formula for $\phi(s,x,r)$ is given in the following theorem.

Theorem 5 *For R, Q and $\bar{Q}$ as specified in Equation (9), (10) and (11), respectively, the solution to Equation (7) and (8) is*

$$\phi(s,x,r) = \begin{cases} e^{A(s,r)x}c & 0 \le x \le b, \\ e^{-s(x-b)}e^{A(s,r)b}c & x > b, \end{cases}$$

where

$$c = ke^{-A(s,r)b}\bar{v}_0(r),$$

k is the scalar such that $u_1 c = 1$.

Proof: Substituting Q and $\bar{Q}$ in Equation (3), we get:

$$\frac{\mathrm{d}\phi(s,x,r)}{\mathrm{d}x} = A(s,r)\phi(s,x,r), \quad 0 < x < b, \tag{15a}$$

$$\frac{\mathrm{d}\phi(s,x,r)}{\mathrm{d}x} = \bar{A}(s,r)\phi(s,x,r), \quad x > b. \tag{15b}$$

It is well-known that the solution to Equation (15a) is

$$\phi(s,x,r) = e^{A(s,r)x}c, \ 0 < x < b, \tag{16}$$

where c is some constant vector to be determined. Notice that for $x > b$:

$$\phi_i(s,x,r) = \mathrm{E}[e^{-sT_r}|Z(0) = i, X(0) = x] = \mathrm{E}[e^{-s(x-b+T_r)}|Z(0) = i, X(0) = b] = e^{-s(x-b)}\phi_i(s,b,r).$$

Then

$$\frac{\mathrm{d}\phi(s,x,r)}{\mathrm{d}x} = \frac{\mathrm{d}[e^{-s(x-b)}\phi(s,b,r)]}{\mathrm{d}x} = -se^{-s(x-b)}\phi(s,b,r), \ x > b. \tag{17}$$

Consider the value $\lim_{x\downarrow b} \frac{\mathrm{d}\phi(s,x,r)}{\mathrm{d}x}$. From Equation (17) and (15b), use the fact $\phi(s,b-,r) = \phi(s,b+,r)$, we get:

$$\bar{A}(s,r)\phi(s,b-,r) = -s\phi(s,b-,r).$$

Using Equation (16) for $\phi(s,b-,r)$ we get

$$\bar{A}(s,r)e^{A(s,r)b}c = -se^{A(s,r)b}c,$$

which implies that $e^{A(s,r)b}c$ is a right eigenvector of $\bar{A}(s,r)$ corresponding to the eigenvalue $-s$, i.e.:

$$e^{A(s,r)b}c = k\bar{v}_0(r).$$

Finally, c is completely determined by the boundary condition that $\phi_0(s,0,r) = 1$. ■

6 First Passage Time: the Balking Queueing Model

In this section, we compute the mean and LST of the first passage time for the balking queuing model. First we give the following construction which indicates that the balking model is a limiting case of the fluid model we analyze in the previous section.

We construct the sample path of $\{W(t), t \geq 0\}$ in the balking model and the sample path of $\{X_r(t), t \geq 0\}$ in the fluid model on common probability space as follows. Without loss of generality, we assume the process $\{X_r(t), t \geq 0\}$ start in "down" time.

Let $\{U_i, i \geq 1\}$ be iid random variables with phase type distribution with parameter α and M; $\{D_i, i \geq 1\}$ be iid random variables with exponential distribution with parameter λ. We think of U_i as the service time of the i-th arriving customer (who may or may not balk) and D_i as the inter-arrival time between the i-th and $(i+1)$-st arriving customer. Then the sample path of the $\{W(t), t \geq 0\}$ process in the balking model is completely described by these two sequences and the parameter b. It is shown in Figure 2. Note that the second customer (arriving at time $D_1 + D_2$) finds the workload above b and hence balks.

Next we construct a sample path of the buffer content process $\{X_r(t), t \geq 0\}$ by using the same two sequences $\{U_i, i \geq 1\}$ and $\{D_i, i \geq 1\}$. It is shown in Figure 3. Here we use D_i as the i-th down time, and U_i/r as the i-th up time. Note that if the i-th downtime finishes while the buffer content is above b, we do not use the i-th uptime at all, and immediately start the next down time. This is equivalent to

having a null transition in the Z process from state 0 to 0. Such a transition occurs in Figure 3 at time $D_1 + \frac{U_1}{r} + D_2$.

Then, clearly,

$$\{X_r(t), t \geq 0\} \stackrel{a.s.}{\to} \{W(t), t \geq 0\} \text{ as } r \to \infty.$$

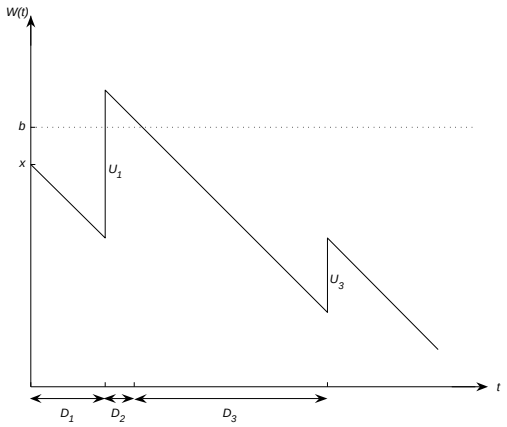

Figure 2: Construction of $W(t)$

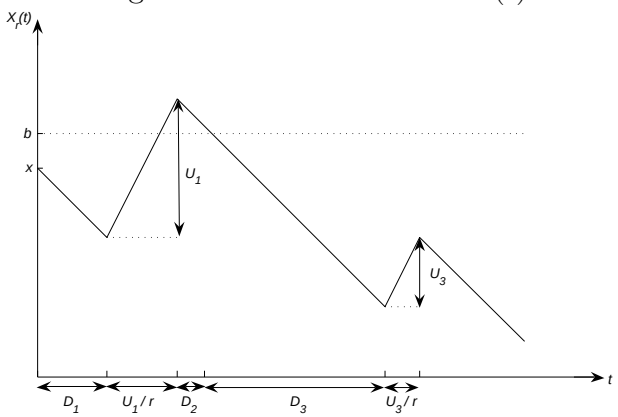

Figure 3: Construction of $X_r(t)$

Recall that $B = \min\{t \geq 0 : W(t) = 0\}$ and $T_r = \min\{t \geq 0 : X_r(t) = 0\}$. It follows from the preceding construction that

$$T_r \stackrel{a.s.}{\to} B$$

as $r \to \infty$.

Let

$$A(s) = \lim_{r\to\infty} A(s,r) = \begin{bmatrix} -\lambda - s & \lambda\alpha \\ M\vec{1} & -M \end{bmatrix},$$

$$\bar{A}(s) = \lim_{r\to\infty} \bar{A}(s,r) = \begin{bmatrix} -s & 0 \\ M\vec{1} & -M \end{bmatrix},$$

and

$$\bar{v}_0 = \lim_{r\to\infty} \bar{v}_0(r) = \begin{bmatrix} 1 \\ (M - sI)^{-1} M\vec{1} \end{bmatrix}.$$

It is clear that if we take the limit of the results given in Theorem 4 and 5, then the first component of the vector is the first passage time for the balking model. We summarize the results in Theorem 6 and 7. The proof is straightforward and hence is omitted.

Theorem 6 *The mean first passage time defined in Equation (1) is given by*

$$m(x) = \begin{cases} u_1 e^{A(0)x}(c + \int_0^x e^{-A(0)t} u_1^{\mathrm{T}} \mathrm{d}t), & 0 \le x \le b, \\ (x-b) + m(b), & x > b, \end{cases}$$

where

$$c = \begin{bmatrix} u_1 \\ \left[M\vec{1}, -M\right] e^{A(0)b} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \vec{1} - \left[M\vec{1}, -M\right] e^{A(0)b} \int_0^b e^{-A(0)t} u_1^{\mathrm{T}} \mathrm{d}t \end{bmatrix},$$

and u_1 is the first row of the identity matrix.

Theorem 7 *The LST of the first passage time defined in Equation (2) is given by*

$$\psi(s,x) = \begin{cases} u_1 e^{A(s)x} c & 0 \le x \le b, \\ u_1 e^{-s(x-b)} e^{A(s)b} c & x > b, \end{cases}$$

where

$$c = k e^{-A(s)b} \bar{v}_0,$$

k is the scalar such that $u_1 c = 1$.

Remark: Differential equations similar to (3) and (7) also hold for other first passage times. For example, suppose $T = \min\{t \ge 0 : X(t) = 0 \text{ or } X(t) = b\}$. This case is actually easier since the constant vector in the solution to Equation (3) is completely determined by the boundary conditions $\pi_i(0) = 0, i \in \mathcal{S}_-$ and $\pi_i(b) = 0, i \in \mathcal{S}_+$. Similarly, the constant vector in the solution to Equation (7) is completely determined by the boundary conditions $\phi_i(s,0) = 1, i \in \mathcal{S}_-$ and $\phi_i(s,b) = 1, i \in \mathcal{S}_+$.

7 Special Case: Exponential Service Times

In this section, we illustrate the results of the previous section with exponential service time with rate μ, and verify the known results.

The exponential distribution is simply a phase type distribution with parameters:

$$M = [-\mu], \alpha = [1].$$

Thus we have

$$A(s) = \begin{bmatrix} -s-\lambda & \lambda \\ -\mu & \mu \end{bmatrix}, \bar{A}(s) = \begin{bmatrix} 0 & 0 \\ -\mu & \mu \end{bmatrix}, \bar{v}_0 = \begin{bmatrix} 1 \\ \mu/(\mu+s) \end{bmatrix}.$$

Using Theorem 6, after tedious algebra, we get the following result for the mean first passage times

$$m(x) = \begin{cases} \frac{1}{(\mu-\lambda)^2}[\mu(\mu-\lambda)x - \frac{\lambda^2}{\mu}e^{-(\mu-\lambda)(b-x)} + \frac{\lambda^2}{\mu}e^{-(\mu-\lambda)b}], & 0 \le x \le b, \\ (x-b) + m(b), & x > b, \end{cases} . \quad (18)$$

Equation 18 is equivalent to the formula given in Proposition 3.1 in Perry and Asmussen [19].

Using Theorem 7, after simplification, we get the following formula which is consistent with Theorem 3.1 in Perry and Asmussen [19]:

$$\psi(s,x) = \begin{cases} \frac{\gamma_1 e^{\theta_1 x} - \gamma_2 e^{\theta_2 x}}{\gamma_1 - \gamma_2} & 0 \le x \le b, \\ e^{-s(x-b)}\psi(s,b) & x > b, \end{cases} \quad (19)$$

where

$$\theta_1 = \frac{(\mu - s - \lambda) + \sqrt{(s+\lambda-\mu)^2 + 4\mu s}}{2}, \quad (20)$$

$$\theta_2 = \frac{(\mu - s - \lambda) - \sqrt{(s+\lambda-\mu)^2 + 4\mu s}}{2}, \quad (21)$$

are the eigenvalues of $A(s)$ and

$$\gamma_i = (\mu - \theta_i - \frac{\lambda\mu}{s+\mu})e^{-\theta_i b}, \ i = 1, 2.$$

From Equation (18) and (19), by taking the limit $b \to \infty$, we obtain the mean and LST of the first passage time for the classical $M/M/1$ queueing model:

$$\lim_{b\to\infty} m(x) = \frac{\mu}{\mu-\lambda}x, \tag{22}$$

$$\lim_{b\to\infty} \psi(s,x) = e^{\theta_2 x}. \tag{23}$$

Equation (23) is exactly Equation (3.4) in Perry and Asmussen [19]. Inversion of the LST given by Equation (23) yields identical result given in Theorem 8 in Prabhu [23]. It is also worth noting that

$$-\left.\frac{\mathrm{d}e^{\theta_2 x}}{\mathrm{d}s}\right|_{s=0} = -\left. xe^{\theta_2 x}\frac{\mathrm{d}\theta_2}{\mathrm{d}s}\right|_{s=0} = \frac{\mu}{\mu-\lambda}x,$$

which matches with Equation (22).

8 Numerical Results

In this section, we illustrate our results with several numerical examples. We consider three different phase type service time distributions: exponential, Erlang and Hyper-exponential.

The exponential distribution is a phase type distribution with parameters:

$$M = [-\mu], \alpha = [1].$$

An Erlang distribution with parameter (n,μ) is a phase type distribution with parameters:

$$\alpha = (1, 0, \cdots, 0)_{1\times n},$$

and

$$M = \begin{pmatrix} -\mu & \mu & & & \\ & -\mu & \mu & & \\ & & \ddots & \ddots & \\ & & & -\mu & \mu \\ & & & & -\mu \end{pmatrix}_{n\times n}.$$

A hyper-exponential distribution (cf. Kulkarni [14]) is a phase type distribution with the following parameters:

$$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n),$$

and

$$M = \begin{pmatrix} -\mu_1 & & \\ & \ddots & \\ & & -\mu_n \end{pmatrix}.$$

Let τ and σ^2 be the mean and variance of the service time respectively. Let $\rho = \lambda\tau$. We summarize the parameters of the distributions we use in the numerical examples in the following list.

1. Exponential (exp): $\mu = 1$ ($\tau = 1$, $\sigma^2 = 1$);
2. 5-Erlang (erlang): $\mu = 5$ ($\tau = 1$, $\sigma^2 = 0.2$);
3. Hyper-exponential (hyper): $\mu_1 = 4, \mu_2 = 2, \mu_3 = 1, \mu_4 = 0.8, \mu_5 = 0.5$, $\alpha_1 = \cdots = \alpha_5 = 0.2$ ($\tau = 1$, $\sigma^2 = 1.75$).

All of them have mean service time of one. The variances are different, with 5-Erlang the smallest and hyper-exponential the largest. The balking threshold b is set to be 2 unless otherwise specified.

In addition to the plots of the mean of B, we also include the plots of the cdf and pdf of B which are defined as follows,

$$F(t, x) = \Pr\{B \le t | W(0) = x\},\ t \ge x,$$

and

$$f(t, x) = \frac{\mathrm{d}F(t, x)}{\mathrm{d}t},\ t > x.$$

Recall that $\psi(s, x) = \mathrm{E}[e^{-sB}|W(0) = x]$, then $F(t, x)$ and $f(t, x)$ can be calculated as the inverse Laplace transform of $\psi(s, x)/s$ and $\psi(s, x)$ respectively. Note that the distribution of B has a mass at $t = x$, since

$$\Pr\{B = x | W(0) = x\} = \Pr\{\text{no arrival during } [0, x]\} = e^{-\lambda x}.$$

We make different plots by varying the service time distributions, ρ or x as summarized in Table 1.

From Theorem 6, we have $m(x) = (x - b) + m(b)$ when $x > b$. The linearity is illustrated in all mean plots. Also, from Theorem 7, we have $\psi(s, x) = e^{-s(x-b)}\psi(s, b)$ when $x > b$. This is reflected as a shift of the pdf curve as shown in Figure 17.

Surprisingly, it is worth noting that as the variance of the service time becomes smaller, the mean $m(x)$ becomes larger. Figure 4, 5 and 6 indicate that $m_{\text{hyper}}(x) <$

Mean: $m(x)$	$\rho = 0.8$ $\rho = 1$ $\rho = 1.2$	exp, erlang, hyper	Figure 4 Figure 5 Figure 6
Mean: $m(x)$	$\rho = 0.8, 1, 1.2$	exp	Figure 7
Distribution: $F(t, x = 1, 2)$	$\rho = 0.8$ $\rho = 1$ $\rho = 1.2$	exp, erlang, hyper	Figure 8 Figure 9 Figure 10
Distribution: $F(t, x = 1)$	$\rho = 0.8, 1, 1.2$	erlang	Figure 11
Distribution: $F(t, x = 1, 2, 3)$	$\rho = 0.8$	erlang	Figure 12
Density: $f(t, x = 1, 2)$	$\rho = 0.8$ $\rho = 1$ $\rho = 1.2$	exp, erlang, hyper	Figure 13 Figure 14 Figure 15
Density: $f(t, x = 1)$	$\rho = 0.8, 1, 1.2$	erlang	Figure 16
Density: $f(t, x = 1, 2, 3)$	$\rho = 0.8$	erlang	Figure 17

Note:

1. Figure 8,9,10 13, 14 and 15 include two sets of three curves for x=1 and x=2.
2. $b = 2$.

Table 1: Plots Summary

$m_{\text{exp}}(x) < m_{\text{erlang}}(x), x > 0$. Moreover, in Figure 8,9 and 10, it is easy to see that $B_{\text{hyper}} <_{st.} B_{\text{exp}} <_{st.} B_{\text{erlang}}$.

Let $m(x, b)$ be the mean first passage time parameterized by the balking threshold b. Figure 18 numerically verifies the following obvious identity:

$$m(x^* + x, b) = m(x^*, b) + m(x, b - x^*),\ 0 \le x^* \le b,\ x \ge 0,$$

by using the exponential service time and $x^* = 1,\ b = 3$.

9 Conclusion

In this paper, we develop an alternative method to solve the first passage time problem for the $M/PH/1$ queues with workload dependent balking via fluid models. The two models are connected by the construction illustrated in Section 6. We use elementary techniques to solve the first passage time problem for the fluid model and obtain explicit solutions for the balking model. The method can also be applied to

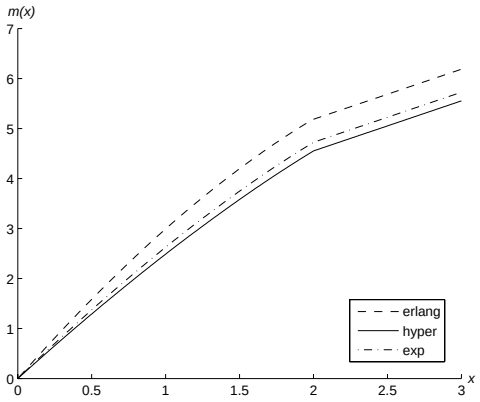

Figure 4: Mean Plot: $b = 2, \rho = 0.8$

the dam model where service requirement is truncated if the complete admission of an arriving customer causes the workload to go beyond a given level.

10 Acknowledgment

We thank David Perry and Wolfgang Stadje for bringing the first passage time problem to our attention. They had suggested using martingale methods to solve it. We found that the alternative method given here is numerically easier.

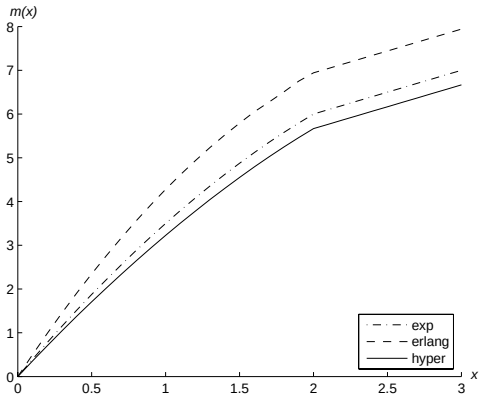

Figure 5: Mean Plot: $b = 2, \rho = 1$

References

[1] S. Asmussen and M. Bladt, *A sample path approach to mean busy periods for Markov-modulated queues and fluids*, Adv. Appl. Prob. **26** (1994), 1117–1121.

[2] R Bekker, *Finite-buffer queues with workload-dependent service and arrival rates*, Queue Syst. Theory Appl. **50** (2005), 231–253.

[3] R. Bekker, S. C. Borst, O. J. Boxma, and O. Kella, *Queues with workload-dependent arrival and service rates*, Queue Syst. Theory Appl. **46** (2004), 537–556.

[4] O. J. Boxma and V. Dumas, *The busy period in the fluid queue*, Tech. report, CWI Report PNA-R9718, November 1997.

[5] O. J. Boxma, D. Perry, and W. Stadje, *Clearing models for $M/G/1$ queues*, Queue Syst. Theory Appl. **38** (2001), 287–306.

[6] T. M. Chen and V. K. Samalam, *Time dependent behavior of fluid buffer models with Markov input and constant output rates*, SIAM J. Appl. Math. **55** (1995), 784–799.

[7] N. Finizo and G. Ladas, *Ordinary differential equations with modern applications*, Wadsworth, Belmont, CA, 1982.

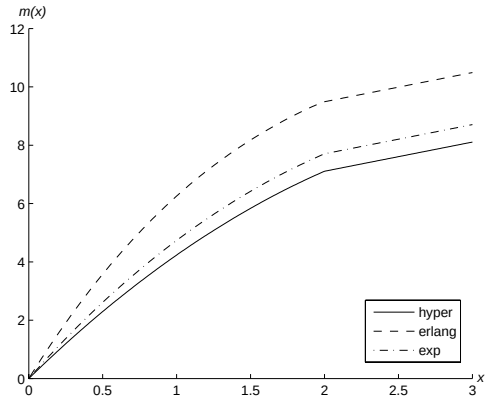

Figure 6: Mean Plot: $b = 2, \rho = 1.2$

[8] O. Garnett, A. Mandelbaum, and M. Reiman, *Designing a call center with impatient customers*, Manufacturing & Service Operations Management **4** (2002), 208–227.

[9] B. Gavish and P. J. Schweitzer, *The markovian queue with bounded waiting time*, Manage. Sci. **23** (1977), 1349–1357.

[10] P. Hokstad, *A single server queue with constant service time and restricted accessibility*, Manage. Sci. **25** (1979), 205–208.

[11] J. Q. Hu and M. A. Zazanis, *A sample path analysis of $M/GI/1$ queues with workload restrictions*, Queue Syst. Theory Appl. **14** (1993), 203–213.

[12] S. G. Johansen and S. Stidham, *Control of arrivals to a stochastic input-output system*, Adv. Appl. Prob. **12** (1980), 972–999.

[13] G. Koole and A. Mandelbaum, *Queueing models of call centers: an introduction*, Ann. Oper. Res. **113** (2002), 41–59.

[14] V. G. Kulkarni, *Modeling and analysis of stochastic systems*, Chapman and Hall, London, 1995.

[15] _____, *Frontiers in queueing: models and applications in science and engineering*, Probability and stochastics, pp. 321–338, CRC Press, Inc., 1997.

[16] V. G. Kulkarni and A. Narayanan, *First passage times in fluid models with application to two priority fluid systems*, IPDS'96, 1996.

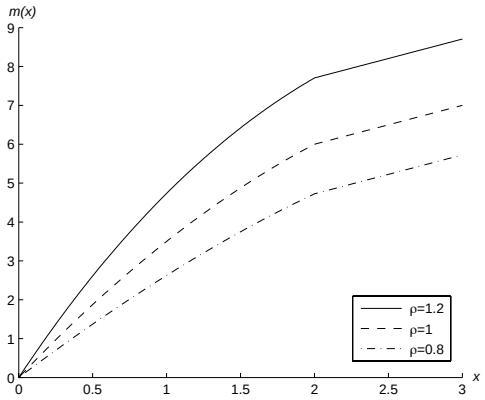

Figure 7: Mean Plot: Exponential Service Time, $b = 2$

[17] V. G. Kulkarni and E. Tzenova, *Mean first passage times in fluid queues*, Oper. Res. Letters **30** (2002), 308–318.

[18] L. Q. Liu and V. G. Kulkarni, *Explicit solutions for the steady state distributions in $M/PH/1$ queues with workload dependent balking*, Queue Syst. Theory Appl. **52** (2006), 251–260.

[19] D. Perry and S. Asmussen, *Rejection rules in the $M/G/1$ queue*, Queue Syst. Theory Appl. **19** (1995), 105–130.

[20] D. Perry and W. Stadje, *Duality of dams via mountain processes*, Oper. Res. Letters **31** (2003), 451–458.

[21] D. Perry, W. Stadje, and S. Zacks, *Busy period analysis for $M/G/1$ and $G/M/1$ type queues with restricted accessibility*, Oper. Res. Letters **27** (2000), 163–174.

[22] ———, *The $M/G/1$ queue with finite workload capacity*, Queue Syst. Theory Appl. **39** (2001), 7–22.

[23] N. U. Prabhu, *Stochastic storage processes: queues, insurance risk, dams, and data communication*, second ed., Applications of mathematics, pp. 123–124, Springer-Verlag, New York, 1998.

[24] W. Scheinhardt, N. van Foreest, and M. Mandjes, *Continuous feedback fluid queues*, Oper. Res. Letters **33** (2005), 551–559.

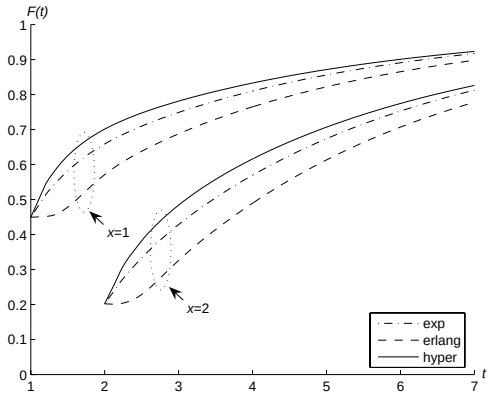

Figure 8: Distribution Plot: $b = 2, \rho = 0.8$ (two sets of three curves for x=1 and x=2)

[25] W. Whitt, *Engineering solution of a basic call-center model*, Manage. Sci. **51** (2005), 221–235.

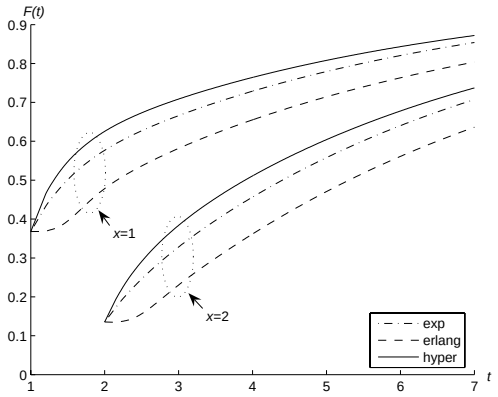

Figure 9: Distribution Plot: $b = 2, \rho = 1$ (two sets of three curves for x=1 and x=2)

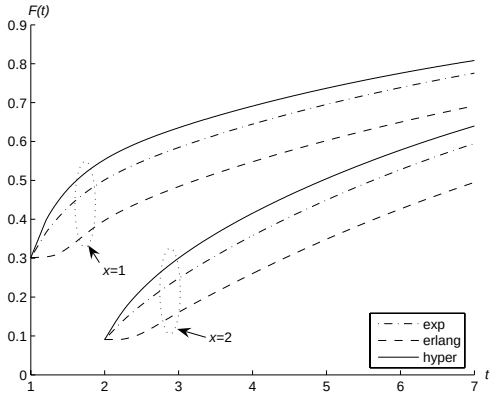

Figure 10: Distribution Plot: $b = 2, \rho = 1.2$ (two sets of three curves for x=1 and x=2)

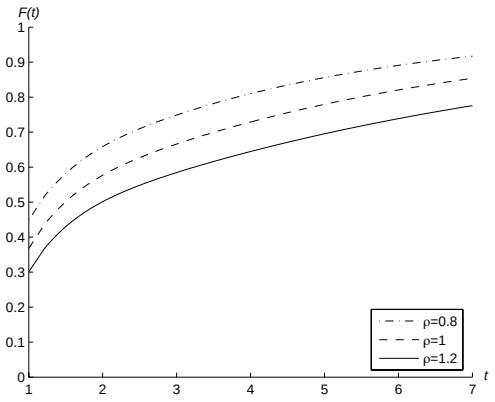

Figure 11: Distribution Plot: Erlang Service Time, $b = 2, x = 1$

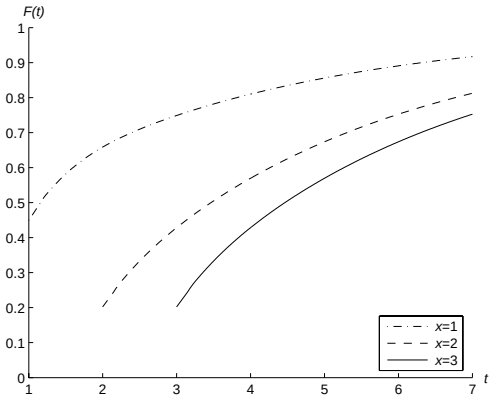

Figure 12: Distribution Plot: Erlang Service Time, $b = 2, \rho = 0.8$

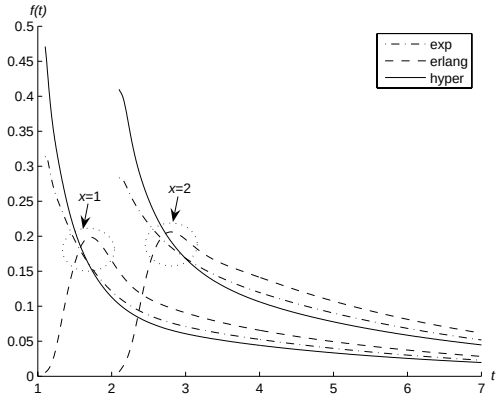

Figure 13: Density Plot: $b = 2, \rho = 0.8$, (two sets of three curves for x=1 and x=2)

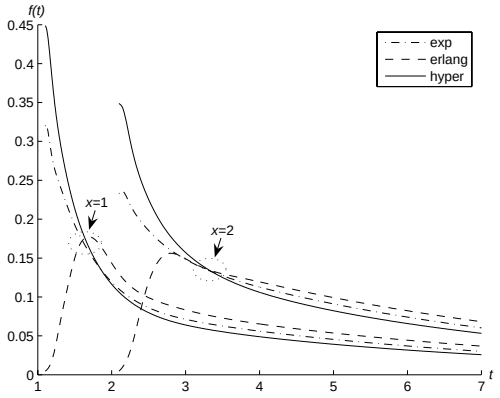

Figure 14: Density Plot: $b = 2, \rho = 1$ (two sets of three curves for x=1 and x=2)

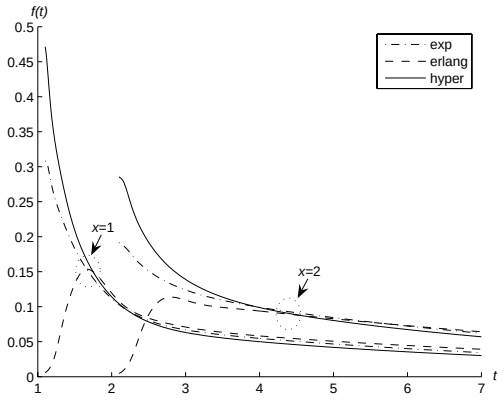

Figure 15: Density Plot: $b = 2, \rho = 1.2$ (two sets of three curves for x=1 and x=2)

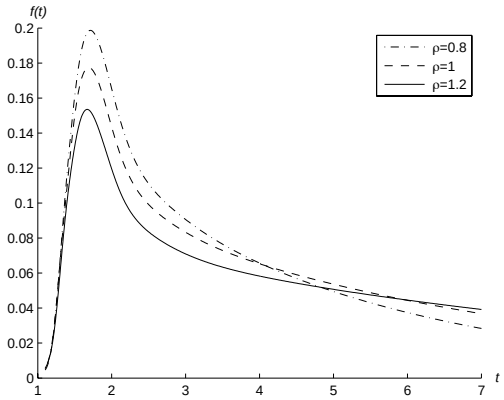

Figure 16: Density Plot: Erlang Service Time, $b = 2, x = 1$

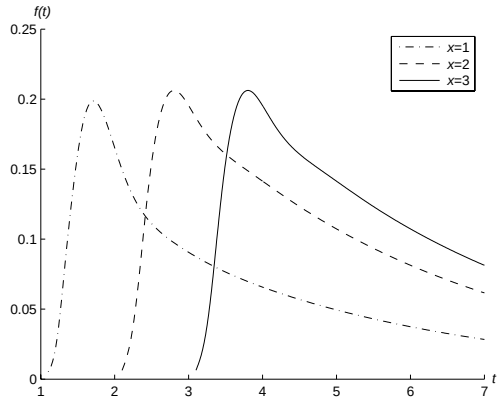

Figure 17: Density Plot: Erlang Service Time, $b = 2, \rho = 0.8$

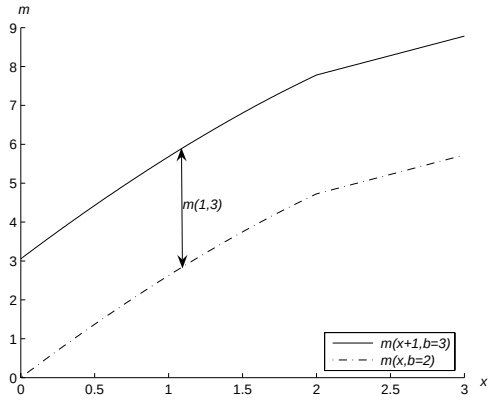

Figure 18: Mean Plot: Exponential Service Time, $\rho = 0.8$