$M/G/1$ Queues with workload-based balking

Liqiang Liu and Vidyadhar G. Kulkarni
Department of Statistics and Operations Research
University of North Carolina
Chapel Hill, N.C. 27599-3180

Abstract
We consider an $M/G/1$ 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 use level-crossing argument to derive an integral equation for the steady state workload distribution. We describe a procedure to solve the equation for general distributions and we solve the equation explicitly for several special cases of service time distributions, such as Phase type, Erlang, Exponential and Deterministic service times. We illustrate the results with numerical examples.

1 Introduction
We analyze a single server first-come first-served (FCFS) queueing system that operates as follows: an arriving customer joins the queue only if he/she sees that the workload in the system is less than a fixed amount $b$. We assume the customer knows the exact amount of the workload in the system at the time of arrival. We also assume that once a customer decides to join the queue, he/she stays in the system until service completion.

The workload at time $t$ is defined as the time it takes the server to empty the system, provided there are no arrivals after $t$. Since the service discipline is FCFS, from the prospective of an arriving customer, the workload can be also interpreted as the queuing time (the time he/she spends in the queue until being served) if he/she chooses to join the queue at all. The balking rule mentioned above is very natural in situations where the customers are unwilling to wait longer than $b$ unit of time for start of service. This happens in systems like call-centers, where usually a
call-in customer is told how long a wait he/she faces before an operator is available to answer the call. Then the customer has the choice to wait (join the queue) or hang up (leave). Although the waiting time information in call-centers is not precise, this model incorporates the right characteristics of the customer behavior. On the other hand, there are some systems, for instance, some communication systems, where the information about amount of workload is always precisely available. What’s more, from the view of control problems, the balking rule can be regarded as 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 [6]. This threshold type policy generates the model considered here.

The assumption of FCFS discipline is not indispensable in our analysis focusing on the workload process, since in general, the workload process is invariant under work-conserving service disciplines. Here we assume FCFS in which case the work content also represents the queuing time, and hence the model represents customers balking in face of long waits.

The only work on such queueing model we know is by Hu and Zazanis [5], who consider various types of restrictions on workload (i.e. balking rules). They obtain the steady state distribution of the workload process for the system with workload restrictions in terms of that of the corresponding queue without restrictions. This requires the unrestricted system to be stable so as to solve the restricted system. But intuitively the system with workload restriction is always stable. The inability to solve the restricted system when the corresponding unrestricted system is unstable is inevitable due to the “cut and paste” technique they use, since the basic idea of such method is to obtain the steady state distribution of “complicated” queueing models in terms of known steady state distribution of simpler models.

In this paper, we use a level-crossing argument, which essentially is a type of sample path analysis, to derive an equation satisfied by the steady state distribution. Cohen [1] introduces Level-Crossing Theory (LCT) for regenerative processes of the $GI/G/1$ type. Doshi [2] generalizes the theory to stationary dam process and presents applications of level-crossing analysis to many queueing systems, especially to single-server queues. Gavish and Schweitzer [3] use level-crossing analysis to study an $M/G/1$ system where arrivals are rejected if their waiting plus service times would exceed a fixed amount (note that the service time for each customer is known upon arrival, which is different from the model we consider here).

We successfully solve the steady-state equation and obtain specific results for different service time distributions. Hu and Zazanis [5] have more general results in the sense that they allow load dependent service rates and vacations while we only consider constant service rate in this paper. It is possible to extend our method
to handle such cases. We do not cover that in this paper in order to emphasize a basic idea by relatively simple models. Our analysis is more general since it does not require stability of the system without balking.

2 An $M/G/1$ Queue with Balking

Here we consider an $M/G/1$ FCFS system with balking based on the workload. The arrival process is Poisson with rate $\lambda$. The service times are iid with a general distribution with mean $\tau$ and a complementary cumulative distribution function (cdfs) $G(\cdot)$. That is, if $S$ is a generic service time random variable, then

$$\Pr\{S > x\} = G(x), \quad E(S) = \tau, \quad \text{Var}(S) = \sigma^2.$$ (2.1)

Thus, the traffic density is

$$\rho = \lambda \tau.$$ (2.2)

Let $\{W(t), t \geq 0\}$ be the workload or virtual waiting time process. $W(t)$ can be interpreted as the time needed to empty the system from time $t$ onwards if there are no arrivals after $t$. The balking rule is characterized by a constant $b$ as follows: A customer arriving at time $t$ joins the system (and stays until service completion) if $W(t) < b$, else he/she leaves and is lost. A typical sample path of $W(t)$ is shown in Figure 1. The arriving times are denoted by $T_i$. Note that the arrival at $T_4$ balks since $W(T_4) > b$.

Clearly $W(t)$ is a regenerative process that regenerates whenever it hits 0, if $\rho < \infty$ (when $\rho = \infty$, it regenerates when it hits $b$). Hence the process $\{W(t), t \geq 0\}$ has a limiting distribution. Let

$$F(x) = \lim_{t \to \infty} \Pr\{W(t) \leq x\},$$

$$\bar{W} = \lim_{t \to \infty} E(W(t)).$$

It is clear that the limiting distribution has a mass at 0, $c = F(0)$, and a density $f(x)$ for $x \geq 0$. We focus on computing $c$, $f(x)(x \geq 0)$ and $F(x)$ (for $x \geq 0$).

3 Equilibrium Distribution of the Workload Process

In this section, we derive an equation for $f(x)$ and $c$ in Theorem 1. We describe a procedure to find its solution in Theorem 2. We also discuss several limiting cases of parameters $b$ and $\lambda$. 


Theorem 1 The equilibrium probability density function (pdf) of the workload process of the $M/G/1$ queue with balking satisfies:

\begin{align}
  f(x) &= \lambda \int_0^{x \wedge b} f(u)G(x - u)du + c\lambda G(x), \\
  \int_0^\infty f(x)dx + c &= 1,
\end{align}

where $x \wedge b = \min(x, b)$.

Proof: We prove this by level-crossing argument. Suppose the process \{\(W(t), t \geq 0\}\} is stationary. Then, during interval \((t, t+h)\), the probability that the workload down-crosses level \(x\) is:

\[ [F(x + h) - F(x)](1 - \lambda h). \]

Thus this is also the expected number of down-crossings during \((t, t+h)\).
Similarly, the probability that the workload up-crosses level \( x \) is:

\[
\int_0^{x \wedge b} f(u)\lambda h Pr\{S \geq x - u\} du + Pr\{W = 0\} \lambda h Pr\{S \geq x\}, \quad (3.3)
\]

which, using our notations yields

\[
\int_0^{x \wedge b} f(u)\lambda h G(x - u) du + c\lambda h G(x). \quad (3.4)
\]

Thus this is also the expected number of up-crossings during \((t, t + h)\).

The level-crossing argument (cf. [2]) implies that (3.2) must equal to (3.4). Now divide both side by \( h \) and let \( h \to 0 \). We get Equation (3.1a). Equation (3.1b) is the normalizing equation. ■

Notice that the first term in the right hand side of Equation (3.1a) is just the convolution of \( f(x) \) and \( G(x) \) multiplied by \( \lambda \), when \( x \wedge b \) is replaced by \( x \). Let \( f_1(x) \) be the solution to

\[
f_1(x) = \lambda \int_0^x f_1(u) G(x - u) du + G(x), \quad x \geq 0. \quad (3.5)
\]

Let

\[
f_2(x) = \lambda \int_0^b f_1(u) G(x - u) du + G(x), \quad x \geq b \quad (3.6)
\]

The solution to Equation (3.1) is given in the following theorem.

**Theorem 2** The solution to (3.1) is:

\[
f(x) = \begin{cases} 
  c\lambda f_1(x) & x < b \\
  c\lambda f_2(x) & x \geq b 
\end{cases} \quad (3.7)
\]

where

\[
c = \left[ \lambda \int_0^b f_1(x) dx + \lambda \int_b^{\infty} f_2(x) dx + 1 \right]^{-1}. \quad (3.8)
\]

**Proof:** The solution is easy to verify by substitution. ■

From the above theorem, it is clear that a possible procedure to obtain \( f(x) \) is to find \( f_1(x) \) first, then compute \( f_2(x) \) by using Equation (3.6). By the normalizing equation (3.1b), after computing the integral, we are able to compute \( c \). This completes the computation of \( f(x) \). Obviously, one main step is to solve Equation (3.5) for \( f_1(x) \). One method is to use Laplace Transform (LT).
Let $G^*(s)$ be the Laplace Transform of $G(x)$, from (3.5), we get the LT of $f_1(x)$ (assuming its existence):

$$f_1^*(s) = \frac{G^*(s)}{1 - \lambda G^*(s)}. \quad (3.9)$$

To continue our procedure, we need the inverse LT of $f_1^*(s)$. A closed form inversion is possible if $G^*(s)$ is rational. However, in some cases, there is an alternative method to solve Equation (3.5). We demonstrate these in Section 4.

We can instantly obtain several interesting results from Theorem 2 under some limiting values of $b$ and $\lambda$. The first case is $b \to 0$. Under this regime, the system reduces to a normal $M/G/1/1$ model. From Theorem 2, as $b \to 0$,

$$c \to \frac{1}{1 + \rho}, \quad (3.10)$$

$$f(x) \to \frac{\lambda}{1 + \rho} G(x), \quad x \geq 0 \quad (3.11)$$

$$\bar{W} \to \frac{\lambda}{2(1 + \rho)} (\sigma^2 + \tau^2). \quad (3.12)$$

It is easy to verify that the above results coincide with the results of an $M/G/1/1$ queueing system.

For the case when $\lambda \to \infty$, a sample path of workload is illustrated in Figure 2. Obviously,

$$\lim_{\lambda \to \infty} f(x + b) = \lim_{\lambda \to \infty} \lim_{b \to 0} f(x), \quad x \geq 0.$$  

Using the fact above, the limiting result in this case can be quickly obtained from what we have in the previous case, simply let $\lambda \to \infty$. That is:

$$f(x) \to 0, \text{ when } 0 \leq x < b, \quad (3.13)$$

$$f(x) \to \frac{1}{\tau} G(x - b), \text{ when } x \geq b, \quad (3.14)$$

$$c \to 0, \quad (3.15)$$

$$\bar{W} \to b + \frac{1}{2\tau} (\sigma^2 + \tau^2). \quad (3.16)$$

When $b \to \infty$, in the limit the system reduces to a normal $M/G/1$ queue. The system is stable as long as $b < \infty$. For $\rho < 1$ (in which case $f_1^*(s)$ exists for $s \geq 0$),
Figure 2: A sample path of $W(t)$ when $\lambda \to \infty$

notice that,

$$\int_0^\infty f_1(x)dx = f_1^*(0),$$

$$G^*(0) = \tau,$$

$$\bar{W} = \left. \frac{df^*(s)}{ds} \right|_{s=0}.$$

From Theorem 2 and Equation (3.9), we get the following limiting results as $b \to \infty$:

$$c \to 1 - \rho,$$  \hspace{1cm} (3.17)

$$f(x) \to (1 - \rho)\lambda f_1(x),$$  \hspace{1cm} (3.18)

$$\bar{W} \to \frac{\lambda(\sigma^2 + \tau^2)}{2(1 - \rho)}.$$  \hspace{1cm} (3.19)

These are consistent with the usual $M/G/1$ results. Next, in Section 4 and 5, we will focus mainly on solving Equation (3.5) for several specific service time distributions,
by using the transform method or directly.

4 Rational $G^*(s)$ and $M/PH/1$ Queue with Balking

Suppose $G^*(s)$ is rational, i.e.,

$$G^*(s) = \frac{N(s)}{D(s)}, \quad (4.1)$$

where $D(s)$ is a $p$ degree polynomial in $s$, $N(s)$ is a polynomial in $s$ whose degree is less than $p$. Then

$$f^*(s) = \frac{N(s)}{D(s) - \lambda N(s)}. \quad (4.2)$$

Let $\theta_i (i = 1, 2, \cdots, p)$ be the roots to

$$D(s) - \lambda N(s) = 0. \quad (4.3)$$

We assume they are distinct. Then from general method of computing inverse LT (cf. [7]), we obtain a closed form expression for $f_1(x)$ as following:

$$f_1(x) = \sum_{i=1}^{p} A_i e^{\theta_i x}, \quad (4.4)$$

where

$$A_i = \lim_{s \to \theta_i} \frac{(s - \theta_i)G^*(s)}{1 - \lambda G^*(s)}, \quad i = 1, 2, \cdots, p. \quad (4.5)$$

Next, as a specific example, we apply our method to solve the queueing system with a common phase type distribution of service time (which has a rational $G^*(s)$). In addition, we give results for Erlang, hyper exponential and exponential distributions as three more special cases of the phase type distribution.

4.1 $M/PH/1$: Transform Method

For a phase type distribution with parameter $(\alpha, M)$ (cf. [8]), the complementary cdf $G(x)$ is given by

$$G(x) = \alpha e^{Mx} \bar{1}, \quad (4.6)$$

where $\bar{1}$ is a column vector with all coordinates equal to 1, $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$ is a non-negative row vector and $\alpha \bar{1} = 1$, $M$ is an $n$ by $n$ 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 mean and the second moment of a phase type distribution is given by

$$E(S) = -\alpha M^{-1} \vec{1} = \tau, \quad (4.7)$$
$$E(S^2) = 2\alpha M^{-2} \vec{1} = \sigma^2 + \tau^2. \quad (4.8)$$

It is also known that the above properties of $M$ imply that all eigenvalues of $M$ have negative real part. We need this condition in the computation of $f_2(x)$ later on. In this case, the LT of $G(x)$ is

$$G^*(s) = \alpha(sI - M)^{-1} \vec{1}, \quad (4.9)$$

where $I$ is an $n$ by $n$ identity matrix.

We follow the procedure described in Section 3. The main results for phase type distribution of service time is given in the following Theorem.

**Theorem 3** Let the service time distribution be a phase type distribution with complementary cdf given by (4.6). The equilibrium pdf of the workload process is given by

$$f(x) = \begin{cases} 
  c\lambda \sum_{i=1}^{n} A_i e^{\theta_i x} & 0 \leq x < b \\
  c\lambda \left[ e^{Mx} + \sum_{i=1}^{n} \lambda A_i e^{Mx} (\theta_i I - M)^{-1} (e^{(\theta_i I - M)b} - I) \right] \vec{1} & x \geq b
\end{cases} \quad (4.10)$$

where $\theta_i$ ($i = 1, 2, \cdots, n$) are $n$ distinct roots to Equation (4.3). $A_i$ are given by Equation (4.5), and $G^*(s)$ is given by Equation (4.9).

The probability that the system is empty is:

$$c = \left\{ \sum_{i=1}^{n} \lambda \frac{A_i}{\theta_i} (e^{\theta_i b} - 1) - \alpha \left[ \sum_{i=1}^{n} \lambda^2 A_i M^{-1} e^{Mb} (\theta_i I - M)^{-1} (e^{(\theta_i I - M)b} - I) \right] \vec{1} \right\}^{-1} - \lambda \alpha M^{-1} e^{Mb} \vec{1} + 1 \quad (4.11)$$
Proof: Note $p = n$. Take inverse LT of $f_1^*(s)$ and apply Theorem 2. ■

Remark: Writing Equation (4.3) as $1 - \lambda \alpha (sI - M)^{-1} \bar{I} = 0$, we see that when $\rho = 1$, i.e., when $-\lambda \alpha M^{-1} \bar{I} = 1$, $\theta_1 = 0$ is one of the roots. In this case, we replace the zero-dividing terms which appear in the above Theorem by the corresponding limits. These terms and their limits are: $\lim_{\theta_1 \to 0} (e^{\theta_1 b} - 1)/\theta_1 = b$ and $\lim_{\theta_1 \to 0} (e^{\theta_1 b} - \theta_1 b - 1)/\theta_1^2 = b^2/2$. Therefore, we do not give the results separately for the case where $\rho = 1$ from now on. Also note that computation of $c$ needs the integral $\int_b^\infty e^{Mx}$ to converge. This is guaranteed by the fact that all eigenvalues of $M$ have negative real part.

We now compute $F(x)$ and $\bar{W}$ by using

$$F(x) = c + \int_0^x f(u)du,$$

$$\bar{W} = \int_0^\infty [1 - F(x)]dx.$$  

A straight forward computation yields the following corollaries.

Corollary 1 (cdf) The equilibrium cdf of the workload process of the $M/PH/1$ queue with balking is:

$$F(x) = \begin{cases} 
    c + c\lambda \sum_{i=1}^n \frac{A_i}{\theta_i}(e^{\theta_i x} - 1), & 0 \leq x \leq b, \\
    F(b) + c\lambda M^{-1}(e^{Mb} - e^{Mb}) \\
    \times \left[I + \sum_{i=1}^n \lambda A_i(\theta_i I - M)^{-1}(e^{(\theta_i I - M)b} - I)\right] \bar{I}, & x > b.
\end{cases}$$

Corollary 2 (Mean workload in equilibrium) The expected value of workload in steady state of the $M/PH/1$ queue with balking is:

$$\bar{W} = b - bc - c\lambda \sum_{i=0}^n \frac{A_i}{\theta_i^2}(e^{\theta_i b} - \theta_i b - 1)$$

$$+ c\lambda M^{-2} e^{Mb} \left[I + \sum_{i=1}^n \lambda A_i(\theta_i I - M)^{-1}(e^{(\theta_i I - M)b} - I)\right] \bar{I}.$$  

It can be verified that the results for limiting parameters $b$ and $\rho$ are consistent with those given for general service time distribution in Section 3.
4.2 $M/PH/1$: Differential Equation Approach

So far, theoretically, the problem of limiting workload distribution for phase type distribution of service time is solved. However, in practice, it can be hard to compute $\theta_i$’s of Equation (4.3) by using $G^*(s)$ of Equation (4.9), for a general phase type distribution. We know $G^*(s)$ plays an essential role in the problem by giving $\theta_i$ and $A_i$. Therefore, we seek an alternative way to solve Equation (3.5) directly. We describe the method in Theorem 4.

First we introduce some notation that will be used in the statement and proof of Theorem 4. We consider the case where the service time has a phase type distribution with complementary cdf given by Equation (4.6).

Let $a_0, a_1, \ldots, a_n$ be the coefficients of the characteristic polynomial of $M$, i.e. :

$$\det(xI - M) = \sum_{j=0}^{n} a_j x^j. \quad (4.12)$$

Let

$$P(\theta) = \sum_{i=0}^{n} \left[ \alpha (a_i I + \lambda \sum_{j=0}^{i} a_j M^{j-i-1} I) \right] \theta^i \quad (4.13)$$

be an $n^{th}$ order polynomial in $\theta$. Let $\theta_1, \theta_2, \ldots, \theta_n$ be the roots of $P(\theta)$. We assume they are distinct. Note that if $-\alpha M^{-1}I = 1/\lambda$ (i.e. traffic density $\rho = -\lambda \alpha M^{-1}I = \rho$), then $\theta_1 = 0$ is one of the roots.

Let $\Theta$ be the Vandermonde matrix of $\theta_1, \theta_2, \ldots, \theta_n$, i.e.:

$$\Theta = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \theta_1 & \theta_2 & \cdots & \theta_n \\ \theta_1^2 & \theta_2^2 & \cdots & \theta_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ \theta_1^{n-1} & \theta_2^{n-1} & \cdots & \theta_n^{n-1} \end{pmatrix}. $$

Since all $\theta_i$ are distinct, $\Theta$ is invertible (cf. [4]).

Let $M_0 = I$, and define $M_j, j \geq 1$ recursively by:

$$M_j = MM_{j-1} + \lambda \alpha M_{j-1} I. \quad (4.14)$$

Let

$$m_i = \alpha M_i I \quad (i = 0, 1, \ldots, n - 1) \quad \text{and} \quad m = (m_0, m_1, \ldots, m_{n-1})^T. \quad (4.15)$$

With these notations, we are ready to state the following theorem.
Theorem 4 Suppose the service time has a phase type distribution with complementary cdf given by (4.6), and assume that the polynomial in (4.13) has \( n \) distinct roots, \( \theta_1, \theta_2, \cdots, \theta_n \). The equilibrium pdf of the workload process is given by

\[
 f(x) = \begin{cases} 
 c\lambda \sum_{i=1}^{n} A_{i} e^{\theta_i x} & 0 \leq x < b \\
 c\lambda \alpha \left[ e^{Mx} + \sum_{i=1}^{n} \lambda A_{i} e^{Mx} (\theta_i I - M)^{-1} (e^{(\theta_i I - M)b} - I) \right] \mathbf{1} & x \geq b 
\end{cases}
\]

(4.16)

where \( A = (A_1, A_2, \cdots, A_n)^T \) is given by:

\[
 \Theta A = m. 
\]

(4.17)

The probability that the system is empty is:

\[
 c = \left\{ \sum_{i=1}^{n} \lambda \frac{A_{i}}{\theta_i} (e^{\theta_i b} - 1) \\
 - \alpha \left[ \sum_{i=1}^{n} \lambda^2 A_{i} M^{-1} e^{Mb} (\theta_i I - M)^{-1} (e^{(\theta_i I - M)b} - I) \right] \mathbf{1} \\
 - \lambda \alpha M^{-1} e^{Mb} \mathbf{1} + 1 \right\}^{-1}.
\]

(4.18)

Proof: Plugging (4.6) in Equation (3.5), and simplifying, we get:

\[
 f_1(x) = \lambda ae^{Mx} \left[ \int_{0}^{x} f_1(u) e^{-Mu} du + I/\lambda \right] \mathbf{1}.
\]

(4.19)

Taking repeated derivatives of the equation above with respect to \( x \), we get (using \( f_1^{(i)}(x) \) to denote the \( i^{th} \) order derivative of \( f_1(x) \) with respect to \( x \)):

\[
 f_1^{(0)}(x) = \lambda \alpha \left[ M^0 e^{Mx} \left( \int_{0}^{x} f_1(u) e^{-Mu} du + I/\lambda \right) \right] \mathbf{1}, \\
 f_1^{(1)}(x) = \lambda \alpha \left[ M^1 e^{Mx} \left( \int_{0}^{x} f_1(u) e^{-Mu} du + I/\lambda \right) + f_1^{(0)}(x) I \right] \mathbf{1}, \\
 f_1^{(2)}(x) = \lambda \alpha \left[ M^2 e^{Mx} \left( \int_{0}^{x} f_1(u) e^{-Mu} du + I/\lambda \right) + M f_1^{(0)}(x) + f_1^{(1)}(x) I \right] \mathbf{1}, \\
 \vdots \\
 f_1^{(n)}(x) = \lambda \alpha \left[ M^n e^{Mx} \left( \int_{0}^{x} f_1(u) e^{-Mu} du + I/\lambda \right) + \sum_{j=0}^{n-1} M^j f_1^{(n-1-j)}(x) \right] \mathbf{1}.
\]

(4.20)
Now multiply the $i$-th equation above by coefficient $a_i$ in the characteristic polynomial of $M$ as defined in (4.12) and add. We get:

$$
\sum_{i=0}^{n} a_i f_1^{(i)}(x) = \lambda \alpha \left[ \sum_{i=0}^{n} a_i M^i e^{Mx} \left( \int_{0}^{x} f_1(u) e^{-Mu} du + I/\lambda \right) + \sum_{i=0}^{n} \sum_{j=0}^{i-1} a_i M^j f_1^{(i-1-j)}(x) \right] \vec{1}.
$$

(4.21)

By Cayley-Hamilton Theorem, we know

$$
\sum_{i=0}^{n} a_i M^i = 0.
$$

(4.22)

Using this and doing algebraic manipulations, Equation (4.21) can be simplified and rewritten as:

$$
\sum_{i=0}^{n} \left[ \alpha \left( a_i I + \lambda \sum_{j=0}^{i} a_j M^{i-j-1} \right) \right] f_1^{(i)}(x) = 0.
$$

(4.23)

This is simply an $n^{th}$ order differential equation with constant coefficients. Using standard methods of solving such equations (cf. [7]), we get the polynomial (4.13) and the solution $f_1(x) = \sum_{i=1}^{n} A_i e^{\theta_i x}$, where the constants $A_i$’s are to be determined by using the initial conditions.

The initial conditions can be found by plugging $x = 0$ in (4.20):

$$
f_1^{(j)}(0) = m_j \quad j = 0, 1, \cdots, n - 1,
$$

(4.24)

where $m_j$ are as defined in (4.15). This yields Equation (4.17) for the constant $A$.

The rest part of the theorem follows by applying Theorem 2. ■

### 4.3 Special Cases

Next, we consider three special cases of service time distribution: Erlang, hyper-exponential and exponential. These belong to the phase type distribution category, and have special parameters ($\alpha, M$). For these cases we can further simplify the results for a general phase type distribution.

#### 4.3.1 Erlang

An Erlang distribution with parameter $(n, \mu)$ is simply a phase type distribution with parameters:

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

(4.25)
\[ M = \begin{pmatrix} -\mu & \mu \\ -\mu & \mu \\ \vdots & \vdots \\ -\mu & \mu \end{pmatrix}_{n \times n}, \]

where we display only the non-zero entries of \( M \). We solve \( f_1(x) \) by transform. Using Equation (4.9), \( G^*(s) \) can be shown to be

\[ G^*(s) = \frac{(s + \mu)^n - \mu^n}{(s + \mu)^n s}. \tag{4.25} \]

Finding the roots to \( 1 - \lambda G^*(s) \) is equivalent to solving the following equation:

\[ \frac{1}{s}[(s - \lambda)(s + \mu)^n + \lambda \mu^n] = 0. \tag{4.26} \]

Note that the left hand side is actually an \( n \) degree polynomial in \( s \). It can be proved that there are exactly \( n \) distinct roots, \( \theta_1, \theta_2, \ldots, \theta_n \). So, writing \( f_1^*(s) \) as

\[ f_1^*(s) = \frac{G^*(s)}{1 - \lambda G^*(s)} \]

\[ = \frac{1}{s}[(s + \mu)^n - \mu^n] - \frac{1}{s}[(s - \lambda)(s + \mu)^n + \lambda \mu^n] \]

\[ = (s - \lambda)^{-1} \left[ \prod_{i=1}^{n} (s - \theta_i) - \prod_{i=1}^{n} (\lambda - \theta_i) \right] \]

\[ \prod_{i=1}^{n} (s - \theta_i), \quad \tag{4.27} \]

then computing \( A_i \) by Equation (4.5) yields

\[ A_i = \prod_{j \neq i} \frac{\lambda - \theta_j}{\theta_i - \theta_j}, \quad (i = 1, 2, \ldots, n). \]

Next, we compute the matrix exponential explicitly and give the result in Theorem 5. Before that, we introduce two more notations. We denote the well known incomplete gamma function by:

\[ \Gamma(n, x) = \int_x^\infty t^{n-1}e^{-t}dt = (n - 1)!e^{-x} \sum_{k=0}^{n-1} \frac{x^k}{k!}. \]

Let

\[ d_i = \frac{\mu}{\mu + \theta_i}, \quad (i = 1, 2, \ldots, n). \]
Theorem 5  The equilibrium pdf of the workload process of the \( M/E_n/1 \) queue with balking is:
\[
f(x) = c\lambda \sum_{i=1}^{n} A_i \theta_i x ,
\]
when \( 0 \leq x < b \); and
\[
f(x) = c\lambda^2 \sum_{i=1}^{n} \frac{A_i}{\theta_i(n-1)!} \left\{ d_i^n e^{\theta_i x} \left[ \Gamma(n, (\mu + \theta_i)x) - \Gamma(n, (\mu + \theta_i)(x-b)) \right] 
+ e^{\theta_i b} \Gamma(n, \mu(x-b)) - \Gamma(n, \mu x) \right\} + c\lambda \frac{\Gamma(n, \mu x)}{(n-1)!},
\]
when \( x \geq b \).
\[
c^{-1} = \lambda \sum_{i=1}^{n} \frac{A_i}{\theta_i} (e^{\theta_i b} - 1) + \frac{\lambda}{\mu(n-1)!} \left[ \Gamma(n+1, \mu b) - \mu b \Gamma(n, \mu b) \right] + 1 
+ \lambda^2 \sum_{i=1}^{n} \frac{A_i}{\mu \theta_i^2 (n-1)!} \left\{ (1 + \theta_i b) \Gamma(n, \mu b) - \mu d_i^n e^{\theta_i b} \Gamma(n, (\mu + \theta_i)b) 
- \theta_i \Gamma(n+1, \mu b) + \theta_i e^{\theta_i b} n! - \mu e^{\theta_i b} (1 - d_i^n)(n-1)! \right\}.
\]

4.3.2  Hyper-exponential

A hyper-exponential distribution (cf. [8]) is a phase type distribution with the following parameters:
\[
\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n),
\]
\[
M = \begin{pmatrix}
-\mu_1 \\
\cdots \\
-\mu_n
\end{pmatrix}.
\]
In this case, \( f_1(x) \) can be solved either by transform or directly. Here we only show some results by using Theorem 4 to solve \( f_1(x) \) directly.

The characteristic polynomial of \( M \) is:
\[
\sum_{i=0}^{n} a_i x^i = \prod_{i=1}^{n} (x + \mu_i).
\]
Then the coefficients \(a_i\) can be computed easily. It is possible to simplify \(P(\theta)\) in Equation (4.13) in terms of the moments of \(S\) as follows:

\[
P(\theta) = \sum_{i=0}^{n} b_i \theta^i,
\]

where

\[
b_i = a_i + \lambda \sum_{j=1}^{i+1} a_{i+1-j} (-1)^j \frac{E(S^j)}{j!}.
\]

Unfortunately, the initial conditions do not simplify, hence we keep the remaining results in terms of \(m\) of (4.15) and \(A\) of (4.17).

### 4.3.3 Exponential

Exponential distribution is the most special case of phase type distribution. It is also a special case of Erlang or hyper-exponential distributions. We give the simplified solution in the following theorem and skip the proof.

**Theorem 6** The equilibrium pdf of the workload process of the \(M/M/1\) queue with balking is:

\[
f(x) = \begin{cases} 
  c \lambda e^{-(\mu-\lambda)x} & 0 \leq x < b \\
  c \lambda e^{\lambda b} e^{-\mu x} & x \geq b,
\end{cases}
\]

where

\[
c = \begin{cases} 
  \frac{1-\rho}{1-\rho e^{-(\mu-\lambda)b}} & \rho \neq 1 \\
  \frac{1}{2+\rho b} & \rho = 1.
\end{cases}
\]

### 5 An Example of Non-rational \(G^*(s): M/D/1\)

As we mentioned before, it can be difficult to find the inverse LT of \(f_1^*(s)\) when \(G^*(s)\) is not rational. In this case, we try to solve Equation (3.5) directly. Here we give the solution when the service time is deterministic with mean \(\tau\), i.e.,

\[
G(x) = \begin{cases} 
  1 & 0 \leq x < \tau \\
  0 & x \geq \tau
\end{cases}
\]

(5.1)

In this case, the LT of \(f_1\) is given by

\[
f_1^*(s) = \frac{1-e^{s\tau}}{s - \lambda + \lambda e^{-s\tau}}.
\]

(5.2)
Computing its inverse is intractable. Hence we show how we solve (3.5) directly in this case.

First we partition \([0, +\infty)\) into intervals of length \(\tau\): \([0, \tau), [\tau, 2\tau), \cdots\) and denote them as \(I_0, I_1, \cdots\) respectively. Since \(G(x)\) is 1 when \(x \in I_0\) (or \(\lfloor \frac{x}{\tau} \rfloor = 0\)) and 0 elsewhere, we rewrite Equation (3.5) as follows:

\[
\begin{align*}
\lambda \int_0^x f_1(u)du + 1 &= f_1(x), & \text{when } x \in I_0, \\
\lambda \int_{x-\tau}^{k\tau} f_1(u)du + \lambda \int_x^{k\tau} f_1(u)du &= f_1(x), & \text{when } x \in I_k, k = 1, 2, \cdots.
\end{align*}
\] (5.3)

We solve these equations recursively and obtain \(f_1(x)\) for each interval. That is, we solve the first equation and get \(f_1(x) = e^{\lambda x}\) when \(x \in I_0\), and so on. Suppose

\[f_1(x) = Q_k(x)e^{\lambda(x-k\tau)}, \quad \text{when } x \in I_k, k = 0, 1, 2, \cdots,\] (5.4)

where \(Q_k(x)\) is a polynomial in \(x\) and \(Q_0(x) = 1\). Substituting (5.4) in (5.3) and take derivative with respect to \(x\), we get

\[Q'_k(x) = -\lambda Q_{k-1}(x-\tau), \quad k = 1, 2, \cdots.\] (5.5)

Therefore, \(Q_k(x)\) can be computed recursively as

\[Q_k(x) = -\lambda \int_0^x Q_{k-1}(u-\tau)du + B_k, \quad k = 1, 2, \cdots.\] (5.6)

The constant \(B_k\) can be computed by the fact that \(f_1(\tau^-) = f_1(\tau^+) + 1\) and \(f_1(x)\) is continuous at \(2\tau, 3\tau, \cdots\). We get

\[B_1 = e^\rho + \rho - 1,\]
\[B_k = Q_{k-1}(k\tau)e^\rho + \lambda \int_0^{k\tau} Q_{k-1}(u-\tau)du, \quad k = 2, 3, \cdots.\]

In the special case when \(b < \tau\), the computation is fairly easy, and we get

\[f(x) = \begin{cases} 
 c\lambda e^{\lambda x} & 0 \leq x < b \\
 c\lambda e^{\lambda b} & b \leq x < \tau \\
 c\lambda(e^{\lambda b} - e^{\lambda(x-\tau)}) & \tau \leq x < b + \tau \\
 0 & \text{elsewhere}
\end{cases}\] (5.7)

In this case, the probability that the system is empty is given by

\[c = \frac{1}{\tau \lambda e^{\lambda b} + 1}.\] (5.8)
6 Numerical Examples

In this section, we illustrate our results with several numerical examples. We consider three different service time distributions:

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 first set of figures, Figure 3, 4, 5 illustrate the shapes of $f(x)$ for different service time distributions, with $\rho = 0.8, 1, 1.2$ respectively and $b = 5$. Then, we pick curves for exponential service time from Figure 3, 4, 5 and put them together in Figure 6 to show how $f(x)$ is affected by the traffic density. As $\rho$ gets larger, the turn at $x = b$ becomes sharper. When $\rho \gg 1$ (in our experiment, $\rho = 10$ is large enough), $f(x)$ is almost 0 when $x < b$. Note that $f(x)$ is a decreasing function of $x$ for $x \geq b$. However, for $0 \leq x < b$, the density function can exhibit complex behavior. It may be increasing, decreasing, constant or non-monotonic.

The second set of figures, Figure 7, 8, 9 and 10, are about the expected workload in steady state ($\bar{W}$). In Figure 7, we fix $b = 5$ and compare $\bar{W}$ for different service time distributions against $\rho$ (since $\tau = 1$, $\rho = \lambda$). When $\rho$ gets larger, $\bar{W}$ clearly converges to the expected levels (see Equation (3.16)), to be specific, to 6, 5.6 and 6.375 for exponential, Erlang and hyper-exponential distributions.

In Figure 8, 9 and 10, for each graph, we fix $\rho$ and plot $\bar{W}$ against $b$ for different service time distributions. At $b = 0$, $\bar{W}$ starts from the value we expect (see Equation (3.10) and (3.12)), and then it increases as $b$ increases. When $\rho = 0.8$, the convergence of $\bar{W}$ to the theoretical level (see Equation (3.19)) is again clearly shown. On the other hand, $\bar{W}$ increases rapidly in $b$ when $\rho \geq 1 \ (\rho = 1, 1.2)$.

Finally, Figure 11 shows the probability that the system is empty in steady state ($c$). As expected, $c$ is decreasing in arrival rate ($\lambda$).
Figure 3: \( f(x) \) for different service time distributions, \( \rho = 0.8, b = 5 \)

Figure 4: \( f(x) \) for different service time distributions, \( \rho = 1, b = 5 \)
Figure 5: $f(x)$ for different service time distributions, $\rho = 1.2, b = 5$

Figure 6: $f(x)$ for different $\rho$, exponential service time, $b = 5$
Figure 7: $\bar{W}$ vs. $\rho$ for different service time distributions, $b = 5$

Figure 8: $\bar{W}$ vs. $b$ for different service time distributions, $\rho = 0.8$
Figure 9: $\bar{W}$ vs. $b$ for different service time distributions, $\rho = 1$

Figure 10: $\bar{W}$ vs. $b$ for different service time distributions, $\rho = 1.2$
Figure 11: \( c \) vs. \( \lambda \) for different service time distributions

References


