Chapter 3
M/G/1 Queues and Variants

3.1 Introduction

This chapter considers the virtual wait process (workload) and related concepts in the M/G/1 queue, and variants of the model. It first develops relationships between sample-path level crossings and the *time dependent* (transient) distribution of wait. These relationships provide sample-path quantities obtainable via simulation or computation, which can estimate the analytical transient pdf of wait. They also lead in Sect. 3.3 to an alternative proof of the basic LC theorem for the *steady-state* pdf of wait (Theorem 1.1 in Sect. 1.6), by taking limits as time tends to infinity. The relationships are also of inherent interest for general time-dependent methods of analysis.

Next, in Sect. 3.3.1, alternative *forms* of Eq. (1.8) are derived by using a different, but very useful LC interpretation of sample-path jumps. These equation forms facilitate the analysis of certain variants of M/G/1 queues such as M/Discrete/1 where the service time has a general discrete distribution (Sect. 3.11.3).

The remainder of the chapter gives LC analyses of M/M/1 and M/G/1 models in the steady state, which illustrate the effectiveness of LC in practice.

3.2 Transient Distribution of Wait

Consider an M/G/1 queue with Poisson arrival rate λ, positive service times with cdf $B(x)$, $x \geq 0$, and pdf $\frac{d}{dx}B(x) = b(x)$, where the derivative exists. Let $\overline{B}(x) \equiv 1 - B(x)$. Consider a sample path of the virtual wait $\{W(t)\}_{t\geq 0}$, and fix level $x > 0$ in the state space $\mathbf{S} = [0, \infty)$ (Figs. 2.1 and 3.1). Let $\mathcal{D}_t(x)$

P.H. Brill, *Level Crossing Methods in Stochastic Models*,
International Series in Operations Research & Management Science 250,
DOI 10.1007/978-3-319-50332-5_3

and $\mathcal{U}_t(x)$ denote the number of down- and upcrossings of level $x \geq 0$ during $(0, t)$, respectively. Both $\{\mathcal{D}_t(x)\}_{t\geq 0}$ and $\{\mathcal{U}_t(x)\}_{t\geq 0}$ are counting processes (e.g., p. 312 in [125]). The existence of $\frac{\partial}{\partial t}E(\mathcal{D}_t(x))$ and $\frac{\partial}{\partial t}E(\mathcal{D}_t(x))$ is important for the transient analysis, so we consider this property in Sects. 3.2.1 and 3.2.2.

3.2.1 Derivative $\partial E(\mathcal{D}_t(x))/\partial t, x \geq 0$

For economy of notation, we define $\mathcal{D}_t(0) \equiv \mathcal{D}_t(0^+) = \mathcal{H}_t^{a,c}(0)$ (number of left-limit continuous hits of 0 from above during $(0, t)) = \mathcal{I}_t(0)$ (number of SP entrances into $\{0\}$ during $(0, t)$) (see Sect. 2.4.11). (Here all downcrossings are continuous downcrossings).

Proposition 3.1 The partial derivative $\frac{\partial}{\partial t}E(\mathcal{D}_t(x))$, $x \geq 0$, exists and is positive for $t > 0$.

Proof The memoryless property of the exponential distribution implies that for each $x \geq 0$ $\{\mathcal{D}_t(x)\}_{t\geq 0}$ is a *delayed renewal* process (e.g., p. 466 in [125]; p. 197 in [99]), i.e., after each downcrossing of level x the future is a probabilistic replica of the whole process starting at time d_0. The delay d_0 depends on the initial wait $W(0) = x_0$. If $x_0 = x$, $d_0 = 0$. If $x_0 \neq x$, d_0 is the time from $t = 0$ to the first downcrossing of x. Starting from time d_0, let the level-x inter-downcrossing times be $d_1, d_2, \ldots$, where $d_k \underset{dis}{\equiv} d_1, k = 2, 3, \ldots$ (Fig. 3.1). Let $H_{d_0}(\cdot)$, $h_{d_0}(\cdot)$ denote the cdf and pdf of d_0, respectively. We prove the result when $d_0 > 0$; if $d_0 = 0$, the proof is similar.

The following well-known relationship (e.g., p. 423 in [125]; p. 167 in [99]) holds for $n = 1, 2, \ldots$, and $t > 0$:

$$\mathcal{D}_t(x) \geq n \iff d_0 + d_1 + \cdots + d_{n-1} \leq t.$$

$$\text{Thus } P(\mathcal{D}_t(x) \geq n) = P(d_0 + d_1 + \cdots + d_{n-1} \leq t).$$

Summing on both sides over $n = 1, 2, \ldots$, gives, by mutual independence of $\{d_i\}_{i=0,1,2,\ldots}$,

$$E(\mathcal{D}_t(x)) = \sum_{n=1}^{\infty} F_{d_0+d_1+\cdots+d_{n-1}}(t) = \sum_{n=1}^{\infty} \int_{s=0}^{t} F_{d_1}^{n-1}(t-s)h_{d_0}(s)ds,$$

where $F_{d_0+d_1+\cdots+d_{n-1}}(t)$, $t > 0$, is the cdf of $d_0 + d_1 + \cdots + d_{n-1}$ and $F_{d_1}^{n-1}(\cdot)$ is the $(n-1)$-fold self convolution of $F_{d_1}(\cdot)$. Since $\{\mathcal{D}_t(x)\}_{t\geq 0}$ is a delayed renewal process, $E(\mathcal{D}_t(x))$ is the renewal function. Thus $E(\mathcal{D}_t(x))$

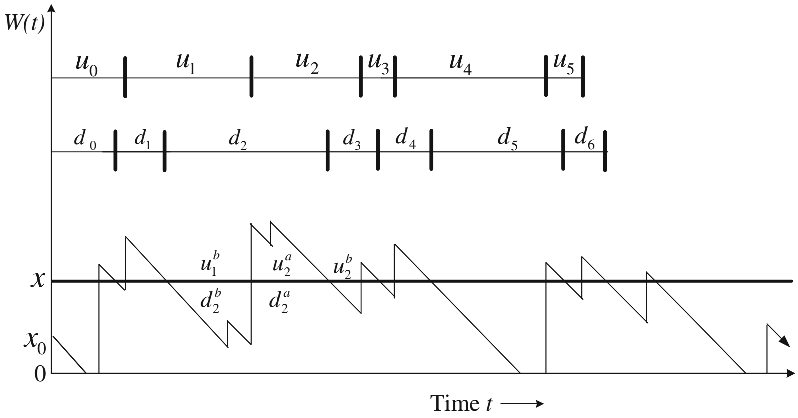


Fig. 3.1 Sample path of virtual wait in M/G/1 showing inter down- and upcrossing times for level x, $\{d_n\}_{n=1,2,\ldots}$, $\{u_n\}_{n=1,2,\ldots}$, and their components, e.g., d_2^b, d_2^a, u_2^a, u_2^b, etc.

is finite for all t, and the series $\sum_{n=1}^{\infty} F_{d_0+d_1+\cdots+d_{n-1}}(t)$ converges uniformly (e.g., p. 182 in [99]).

Since $F_{d_1}^{n-1}(0) = 0$, we obtain the derivative of the $(n-1)$-th summand as

$$\begin{aligned}&\tfrac{\partial}{\partial t}\int_{s=0}^{t} F_{d_1}^{n-1}(t-s)h_{d_0}(s)ds\\ &= \int_{s=0}^{t} \tfrac{\partial}{\partial t} F_{d_1}^{n-1}(t-s)h_{d_0}(s)ds + F_{d_1}^{n-1}(0)h_{d_0}(t)\\ &= \int_{s=0}^{t} f_{d_1}^{n-1}(t-s)h_{d_0}(s)ds,\end{aligned}$$

where $f_{d_1}^{n-1}(\cdot)$ is the pdf of the $(n-1)$-fold convolution of d_1. (Due to Poisson arrivals, d_i is a continuous random variable implying $d_0 + d_1 + \cdots + d_{n-1}$ is continuous.) If we assume the parameters of the M/G/1 queue are such that the series of derivatives $\left\{\int_{s=0}^{t} f_{d_1}^{n-1}(t-s)h_{d_0}(s)ds\right\}_{n=1,2,\ldots}$ also converges uniformly, then we can interchange the order of differentiation and summation (e.g., p. 317 in [6]), giving

$$\begin{aligned}\frac{\partial}{\partial t}E(\mathcal{D}_t(x)) &= \sum_{n=1}^{\infty}\frac{\partial}{\partial t}\int_{s=0}^{t} F_{d_1}^{n-1}(t-s)h_{d_0}(s)ds\\ &= \sum_{n=1}^{\infty}\int_{s=0}^{t} f_{d_1}^{n-1}(t-s)h_{d_0}(s)ds\ .\end{aligned}$$

Moreover, $\frac{\partial}{\partial t}E(\mathcal{D}_t(x)) > 0$ since both $f_{d_1}^{n-1}(t-s) > 0$ and $h_{d_0}(s) > 0$. (Alternatively, $\frac{\partial}{\partial t}E(\mathcal{D}_t(x)) > 0$ follows since $E(\mathcal{D}_t(x))$ is a non-decreasing function of t.) ■

Let $d_i = d_i^a + d_i^b$, where d_i^a is the time interval spent above x and d_i^b is the immediately preceding time interval spent below x by $W(\cdot)$ during d_i, $i = 1, 2, \ldots$.

3.2.2 Derivative $\partial E(\mathcal{U}_t(x))/\partial t$, $x \geq 0$

Consider a sample path of $\{W(t)\}_{t\geq 0}$. The process $\{\mathcal{U}_t(x)\}_{t\geq 0}$ is a "delayed" process. In general, however, $\{\mathcal{U}_t(x)\}_{t\geq 0}$ is not a renewal process. The delay u_0, is the time from $t = 0$ to the first (jump) *upcrossing* of x after time d_0. The subsequent level-x *inter-upcrossing* times starting at u_0 are denoted by $u_1, u_2, \ldots$ (Fig. 3.1). Let $u_i = u_i^a + u_i^b$ where u_i^a is the time interval spent above x and u_i^b is the immediately following time interval spent below x by $W(\cdot)$ during u_i. In general the random variables $\{u_i\}_{i=1,2,\ldots}$ are *not i.i.d.* (see Remark 3.1).

Remark 3.1 Let $\gamma_{x|y}$ denote the excess of a jump over level x given that the jump starts at level $y < x$ and initiates the interval u_i. Then $P(\gamma_{x|y} > z) = \frac{\overline{B}(x-y+z)}{\overline{B}(x-y)}$. Thus u_i^a depends on $x - y$, and $u_i^a \underset{dis}{=} \mathcal{B}_{\gamma_{x|y}}$, where $\mathcal{B}_{\gamma_{x|y}}$ denotes an M/G/1 busy period that starts with $W(0) \underset{dis}{=} \gamma_{x|y}$. (**Note**: The symbol "$\underset{dis}{=}$" means "is equal in distribution to", henceforth.) If $i \neq j$ then $u_i^a \underset{dis}{\neq} u_j^a$ $a.s.$ (almost surely, i.e., with probability 1) because the start-of-jump position y is a continuous random variable. Now $u_i^b \underset{dis}{\equiv} u_1^b$, $i = 2, 3, \ldots$, because at the start of u_i^b $\{W(t)\}_{t\geq 0}$ downcrosses x and the future evolution is a probabilistic replica of $\{W(t)\}_{t\geq 0}$ from the start time of d_1, due to Poisson arrivals (that is, due to the memoryless property of the exponential interarrival times).

Proposition 3.2 The partial derivative $\partial E(\mathcal{U}_t(x))/\partial t$, $x \geq 0$, exists and is positive for $t > 0$.

Proof The delay time u_0 is a continuous random variable (r.v.). The process $\{\mathcal{U}_t(x)\}_{t\geq 0}$ is a counting process, but is not a renewal process (Fig. 3.1). Let $H_{u_0}(\cdot)$, $h_{u_0}(\cdot)$ denote the cdf and pdf of u_0, respectively.

The following relationship, usually applied for a renewal process, also holds for a counting process even though the inter-occurrence times are not independent. Thus

$$\mathcal{U}_t(x) \geq n \iff u_0 + u_1 + \cdots + u_{n-1} \leq t, n = 1, 2, \ldots$$

$$P(\mathcal{U}_t(x) \geq n) = P(u_0 + u_1 + \cdots + u_{n-1} \leq t).$$

Summing on both sides over $n = 1, 2, \ldots$ gives

$$E(\mathcal{U}_t(x)) = \sum_{n=1}^{\infty} F_{u_0+u_1+\cdots+u_{n-1}}(t)$$
$$= \sum_{n=1}^{\infty} \int_{s=0}^{t} F_{u_1+\cdots+u_{n-1}}(t-s)h_{u_0}(s)ds$$

where $F_{u_1+\cdots+u_{n-1}}(t)$ is the cdf of $u_1 + \cdots + u_{n-1}$. Since u_i is continuous for each $i = 1, 2, \ldots$, the sum $u_0 + u_1 + \cdots + u_{n-1}$ is a continuous r.v. Taking $\frac{\partial}{\partial t}$ on both sides (differentiating under the integral) gives

$$\frac{\partial}{\partial t} E(\mathcal{U}_t(x)) = \sum_{n=1}^{\infty} \Big(\int_{s=0}^{t} \frac{\partial}{\partial t} F_{u_1+\cdots+u_{n-1}}(t-s)h_{u_0}(s)ds$$
$$+ F_{u_1+\cdots+u_{n-1}}(0)h_{u_0}(t) \Big)$$
$$= \sum_{n=1}^{\infty} \int_{s=0}^{t} f_{u_1+\cdots+u_{n-1}}(t-s)h_{u_0}(s)ds,$$

where $f_{u_1+\cdots+u_{n-1}}(\cdot)$ is the pdf of $u_1 + \cdots + u_{n-1}$, since $F_{u_1+\cdots+u_{n-1}}(0) = 0$. The rest of the proof is similar to that in Proposition 3.1. Positiveness follows since $E(\mathcal{U}_t(x))$ is an increasing function of t. ■

The derivatives $\frac{\partial}{\partial t} E(\mathcal{U}_t(x))$ and $\frac{\partial}{\partial t} E(\mathcal{D}_t(x))$ are fundamentally related to the transient cdf of $W(t)$ (Sects. 3.2.3–3.2.8).

Remark 3.2 Assume the service time S is exponentially distributed with mean $1/\mu$ as in M/M/1 (Sect. 3.5). Then for any sample path

$$P(\gamma_{x|y} > z) = \frac{\overline{B}(x-y+z)}{\overline{B}(x-y)} = \frac{e^{-(x-y+z)}}{e^{-(x-y)}} = e^{-\mu z},$$

which is independent of x, y, and $x - y$, so that $u_i^a \underset{dis}{\equiv} \text{Exp}_\mu$ (:= exponential r.v. with mean $1/\mu$). In that case $u_i^a \underset{dis}{\equiv} \mathcal{B}_{M/M/1}, i - 1, 2, \ldots$ where $\mathcal{B}_{M/M/1}$:= busy period of an M/M/1 queue with $S \underset{dis}{=} \text{Exp}_\mu$ (Sect. 3.5.6). Then $\{u_i\}_{i=1,2,\ldots}$ **is a renewal process**.

3.2.3 Level Crossings and Transient CDF of Wait

Denote the transient time-t cdf, pdf and probability of a zero wait respectively as

$$\begin{aligned} &F_t(x) = P(W(t) \leq x), x \geq 0, t \geq 0, \\ &f_t(x) = \tfrac{\partial}{\partial x} F_t(x), x > 0, t \geq 0, \text{ wherever } \tfrac{\partial}{\partial x} F_t(x) \text{ exists}, \\ &P_0(t) = F_t(0), t \geq 0. \end{aligned} \tag{3.1}$$

Define the joint cdf of $(W(t_1), W(t_2))$ as

$$F_{t_1,t_2}(x_1, x_2) = P(W(t_1) \leq x_1, W(t_2) \leq x_2), t_1 \neq t_2 \geq 0, x_1, x_2 \geq 0. \tag{3.2}$$

The marginal cdfs are

$$F_{t_1}(x_1) = F_{t_1,t_2}(x_1, \infty),\ \ F_{t_2}(x_1) = F_{t_1,t_2}(\infty, x_2), x_1, x_2 \geq 0 \tag{3.3}$$

Note that $\mathcal{D}_t(x) - \mathcal{U}_t(x) \in \{0, +1,\ -1\}$ for every $x \geq 0,\ t \geq 0,$ since down- and upcrossings of a fixed state-space level alternate in time (formulas (2.4) and (2.2)). The simple but useful Theorem 3.1 below connects $E(\mathcal{U}_t(x))$, $E(\mathcal{D}_t(x))$ and the transient cdf $F_t(x)$, by using (3.3) with $t_1 = 0$, $t_2 = t > 0$, and $x_1 = x_2 = x$ (Fig. 3.2).

In M/G/1, $\mathcal{D}_t(x) = \mathcal{D}_t^c(x)$ since all downcrossings are *left-continuous* (see Sect. 2.4.5). Also $\mathcal{U}_t(x) = \mathcal{U}_t^j(x)$, since all upcrossings are *jump* upcrossings.

Theorem 3.1 (P.H. Brill 1983) In the M/G/1 queue, for fixed $x \geq 0, t \geq 0$,

$$E(\mathcal{D}_t(x)) = E(\mathcal{U}_t(x)) + F_t(x) - F_0(x). \tag{3.4}$$

Proof Equation (3.4) holds when $t = 0$ because $\mathcal{D}_0(x) = \mathcal{U}_0(x) = 0, x \geq 0$. For $t > 0$, we compare possible sample path values of $\{W(s)\}_{0 \leq s \leq t}$ at $s = 0$ and $s = t$ with respect to level x, and relate the possible values to $F_0(x)$, $F_t(x)$ and $F_{0,t}(x, x)$ (Fig. 3.2). This procedure leads to the following values and probabilities for $\mathcal{D}_t(x) - \mathcal{U}_t(x)$:

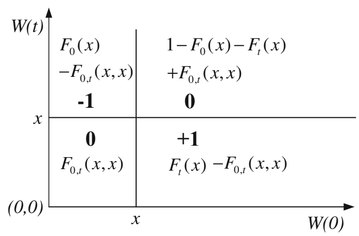


Fig. 3.2 Values of $\mathcal{D}_t(x) - \mathcal{U}_t(x)$ are $+1, 0, -1$, with probabilities shown in the finite and infinite sub-squares and two infinite rectangles of the $(W(0), W(t))$ plane

$\mathcal{D}_t(x) - \mathcal{U}_t(x)$	Probability
0	$1 - F_t(x) - F_0(x) + 2F_{0,t}(x,x)$
$+1$	$F_t(x) - F_{0,t}(x,x)$
-1	$F_0(x) - F_{0,t}(x,x)$

(3.5)

From (3.5) we obtain for fixed $x \geq 0$,

$$E(\mathcal{D}_t(x)) - E(\mathcal{U}_t(x)) = F_t(x) - F_0(x), t \geq 0, \tag{3.6}$$

equivalent to (3.4). ■

In (3.5) the term $\mathcal{D}_t(x) - \mathcal{U}_t(x) = 0$ contributes 0 to $E\left(D_t(x) - \mathcal{U}_t(x)\right)$; it is included for completeness. In further similar computations of expected value, terms with value 0 may be omitted.

Equation (3.6) leads to Theorem 3.2 below, which is fundamental for relating the transient probability distribution of wait and sample-path properties (see Remark 3.3 below).

3.2.4 Relating the Transient CDF and Level Crossings

Theorem 3.2 (P.H. Brill 1983) In the M/G/1 queue

$$\frac{\partial}{\partial t} E(\mathcal{D}_t(x)) = \frac{\partial}{\partial t} F_t(x) + \frac{\partial}{\partial t} E(\mathcal{U}_t(x)), t > 0, x \geq 0. \tag{3.7}$$

Proof Differentiating (3.4) with respect to t gives formula (3.7). (Existence of $\frac{\partial}{\partial t}(\cdot)$ of each term in (3.4) is considered in Sects. 3.2.1 and 3.2.2.) ■

Remark 3.3 Theorem 3.2 is a special case of the more general Theorem B (Theorem 4.1 in Sect. 4.2.1), which connects the sample-path marginal entrance rate and marginal exit rate of an arbitrary measurable set $\boldsymbol{A} \subset \boldsymbol{S}$ ($\boldsymbol{S}$:= state space) to $P_t(\boldsymbol{A})$ (:= probability of $\boldsymbol{A}$ at time t). In Theorem 3.2 $\boldsymbol{A} = [0, x]$.

3.2.5 Downcrossings and Transient PDF of Wait

Theorem 3.3 below shows that the sample-path quantity $\partial E(\mathcal{D}_t(x))/\partial t$ equals the analytic pdf $f_t(x), x \geq 0$, where $f_t(0) :\equiv f_t(0^+)$. We now briefly outline an important consequence, realizable by a computer program.

Simulate a finite number of independent sample paths of $\{W(s)\}_{s\geq 0}$ on $[0, t+h]$ (h small) to estimate $E(\mathcal{D}_t(x))$ and $E(\mathcal{D}_{t+h}(x))$, respectively. Then use $(E(\mathcal{D}_{t+h}(x)) - E(\mathcal{D}_t(x)))\,/\,h$ to estimate $\partial E(\mathcal{D}_t(x))/\partial t \approx f_t(x)$, $x \geq 0$. Adjust the values of t, x and h as needed to fit the particular model being considered.

Another important consequence is Corollary 3.2 below, which leads to an alternative proof of the crucial '*downcrossing*' part of the basic LC theorem for the *steady-state* pdf of wait (i.e., $\lim_{t\to\infty} E(\mathcal{D}_t(x))/t = f(x)$, in Theorem 1.1) in Chap. 1.

Theorem 3.3 *In the M/G/1 queue, for each* $t > 0$,

$$\frac{\partial}{\partial t} E(\mathcal{D}_t(x)) = f_t(x), x > 0, \tag{3.8}$$

$$\frac{\partial}{\partial t} E(\mathcal{D}_t(0)) = f_t(0). \tag{3.9}$$

Proof For the virtual wait $\{W(t)\}_{t\geq 0}$, fix state-space level $x > 0$. Consider the state-space triangular set $\triangle_{t,x,h} := \{(t, x+h), (t, x), (t+h, x)\}, t > 0$, where h is "small" (see Fig. 3.3). Examination of some possible sample paths $W(s)$ with respect to $\triangle_{t,x,h}$ leads to the possible values of $\mathcal{D}_{t+h}(x) - \mathcal{D}_t(x)$ and their corresponding probabilities given in the table in (3.10); a brief explanation follows immediately after Theorem 3.3.

$\mathcal{D}_{t+h}(x) - \mathcal{D}_t(x)$	Probability
$+1$	$F_t(x+h) - F_t(x) + o(h)$
-1	0, since $\mathcal{D}_t(x)$ increases with t
≥ 2	$o(h)$

(3.10)

Computing $E\,(D_{t+h}(x) - D_t(x))$ using (3.10), and dividing by h yields

$$\frac{E(\mathcal{D}_{t+h}(x)) - E(\mathcal{D}_t(x))}{h} = \frac{F_t(x+h) - F_t(x)}{h} + \frac{o(h)}{h}.$$

Letting $h \downarrow 0$ gives (3.8); then letting $x \downarrow 0$ yields (3.9). ■

Explanation of the Probabilities in (3.10)

Let the Poisson arrival rate be λ. Assume the pdf $b(x)$ of the common service time is bounded on $(0, \infty)$. In this paragraph we denote the event $\{\mathcal{D}_t(x+h) - \mathcal{D}_t(x) = 1\}$ by $\{Diff1\}$ for brevity. Consider $P(Diff1)$ for the three types of paths that enter and exit set $\triangle_{t,x,h}$ in Fig. 3.3. Employing the memoryless property of the exponential distribution, we get

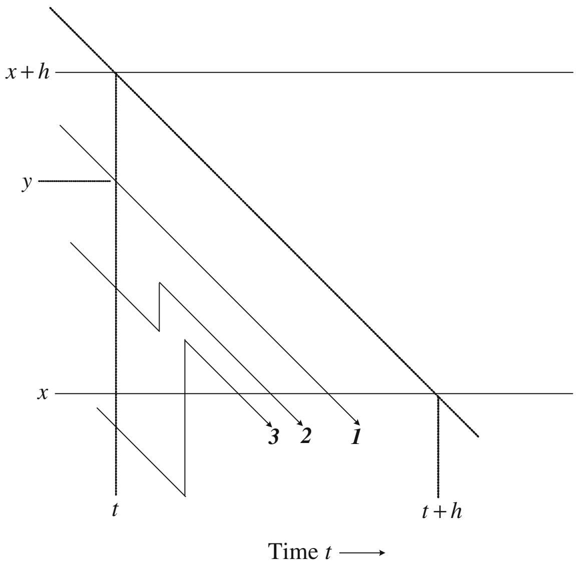


Fig. 3.3 Sample-path examples in triangle $\Delta_{t,x,h}$ resulting in $\mathcal{D}_t(x+h) - \mathcal{D}_t(x) = 1$. Probabilities are: P(path type 1) $= 1 - \lambda(y-x) + o(y-x)$; P(path type 2) $\leq o(h)$; P(path type 3) $\leq o(h)$

$$P(\mathit{Diff1}|\text{path 1}) = e^{-\lambda(y-x)} \to 1 \text{ as } h \downarrow 0 \text{ since } (y-x) < h,$$

$$P(\mathit{Diff1}|\text{path is type 2}) = [\lambda h + o(h)]\,[b(\cdot)h + o(h)] = o(h),$$

$$P(\mathit{Diff1}|\text{path is type 3}) < [\lambda h + o(h)]\,[b(\cdot)h + o(h)] = o(h),$$

where $\lambda h + o(h) = P$(an arrival occurs in $(0,h)$), and $b(\cdot)h + o(h) = P$(a service-time jump ends in an interval of size $< h$). Similar consideration of other possible paths implies that $P(\mathcal{D}_{t+h}(x) - \mathcal{D}_t(x) = n) = o(h)$ $(n = 2, 3, \ldots)$. Then $P(\mathcal{D}_{t+h}(x) - \mathcal{D}_t(x) \geq n) = o(h)$ results because a countable sum of $o(h)$s = $o(h)$.

Alternative Proof of Formula (3.8) for Perspective

We can write

$$\begin{aligned}
&E(D_{t+h}(x) - D_t(x)) = 1 \cdot P(D_{t+h}(x) - D_t(x) = 1) + o(h)\\
&= \int_{y=x}^{x+h} e^{-\lambda(y-x)} f_t(y)dy + o(h)\\
&= \int_{y=x}^{x+h} [1 - \lambda(y-x) + o(y-x)]\, f_t(y)dy + o(h)\\
&= F_t(x+h) - F_t(x) - \int_{y=x}^{x+h} [\lambda(y-x) - o(y-x)]\, f_t(y)dy + o(h)\\
&= F_t(x+h) - F_t((x) - h\left[\lambda(y^* - x) - o(y^* - x)\right] f_t(y^*)dy + o(h),\\
&\qquad\qquad x < y^* < x+h,
\end{aligned} \tag{3.11}$$

by the mean value theorem for integrals (e.g., p. 237 in [137]). Dividing both sides by h gives

$$\frac{E(D_{t+h}(x)-D_t(x))}{h}$$
$$= \frac{F_t(x+h)-F_t((x)}{h} - \lambda(y^* - x) + o(y^* - x)\cdot f_t(y^*) + \frac{o(h)}{h}.$$

Letting $h \downarrow 0$ leads to (3.8), because $h \downarrow 0$ implies: $(y^* - x) \downarrow 0$; $o(y^* - x) \downarrow 0$; $f_t(y^*) \to f_t(x)$; $o(h)/h \to 0$.

Corollary 3.1 For fixed $t > 0$,

$$E(\mathcal{D}_t(x)) = \int_{s=0}^{t} f_s(x)ds,\, x > 0, t > 0. \tag{3.12}$$
$$E(\mathcal{D}_t(0)) = \int_{s=0}^{t} f_s(0)ds,\, t > 0. \tag{3.13}$$

Proof Solving (3.8) for $E(\mathcal{D}_t(x))$ and (3.9) for $E(\mathcal{D}_t(0))$ by integrating with respect to t, and applying the initial condition $E(\mathcal{D}_0(x)) \equiv 0, x \ge 0$, gives (3.12) and (3.13), respectively. ■

3.2.6 *Alternative Proof of* $\lim_{t\to\infty} E(\mathcal{D}_t(x))/t = f(x)$

Starting from the transient analysis, Corollaries 3.2 and 3.3 below provide an alternative proof of the *downcrossing-rate* part of Theorem 1.1 in Sect. 1.6, Chap. 1, i.e., Eqs. (1.12) and (1.13). Let $\{P_0, f(x)\}_{x>0}$ denote the limiting (*steady-state*) mixed pdf of $\{W(t)\}_{t\ge0}$. We assume $\lambda E(S) < 1$ (condition for existence of steady state).

Corollary 3.2

$$\lim_{t\to\infty} \frac{\partial}{\partial t} E(\mathcal{D}_t(x)) = \lim_{t\to\infty} \frac{E(\mathcal{D}_t(x))}{t} = f(x), x > 0 \tag{3.14}$$
$$\lim_{t\to\infty} \frac{\partial}{\partial t} E(\mathcal{D}_t(0)) = \lim_{t\to\infty} \frac{E(\mathcal{D}_t(0))}{t} = f(0^+) \equiv f(0). \tag{3.15}$$

Proof Let $t \to \infty$ in (3.8) and (3.9) giving respectively

$$\lim_{t\to\infty} \frac{\partial}{\partial t} E(\mathcal{D}_t(x)) = \lim_{t\to\infty} f_t(x) = f(x), x > 0,$$
$$\lim_{t\to\infty} \frac{\partial}{\partial t} E(\mathcal{D}_t(0)) = \lim_{t\to\infty} f_t(0) = f(0).$$

In (3.12) and (3.13) divide both sides by $t > 0$, and let $t \to \infty$. Then

$$\lim_{t\to\infty} E(\mathcal{D}_t(x))/t = \lim_{t\to\infty} \frac{1}{t}\int_{s=0}^{t} f_s(x)ds = f(x), x > 0,$$
$$\lim_{t\to\infty} E(\mathcal{D}_t(0))/t = \lim_{t\to\infty} \frac{1}{t}\int_{s=0}^{t} f_s(0)ds = f(0).$$

Then (3.14) and (3.15) follow. ■

Corollary 3.3

$$\lim_{t\to\infty} \frac{\mathcal{D}_t(x)}{t} = f(x), x \geq 0\,(a.s.)\,. \tag{3.16}$$

Proof Since $\{\mathcal{D}_t(x)\}_{y\geq 0}$ is a renewal process due to Poisson arrivals, by the elementary renewal theorem,

$$\lim_{t\to\infty} \frac{E(\mathcal{D}_t(x))}{t} = \lim_{t\to\infty} \frac{\mathcal{D}_t(x)}{t}, x \geq 0\,(a.s.)\,.$$

Thus (3.16) follows from (3.14) and (3.15). ■

Corollary 3.4 gives an alternative perspective of set and rate balance (see Sect. 2.4.7) in Chap. 2.

Corollary 3.4

$$\lim_{t\to\infty} \frac{E(\mathcal{D}_t(x))}{t} = \lim_{t\to\infty} \frac{E(\mathcal{U}_t(x))}{t}, x \geq 0, \tag{3.17}$$
$$\lim_{t\to\infty} \frac{\mathcal{D}_t(x)}{t} = \lim_{t\to\infty} \frac{\mathcal{U}_t(x)}{t}, x \geq 0\ (a.s.)\,. \tag{3.18}$$

Proof $\mathcal{D}_t(x) - \mathcal{U}_t(x) \in \{0, +1, -1\}$, $t \geq 0$, $x \geq 0$, for all possible sample paths of $\{W(t)\}_{t\geq 0}$. Hence $-1 \leq \mathcal{D}_t(x) - \mathcal{U}_t(x) \leq +1$, and $-1 \leq E(\mathcal{D}_t(x)) - E(\mathcal{U}_t(x)) \leq +1$. Dividing by $t > 0$ and letting $t \to \infty$ gives (3.17) and (3.18). ■

Remark 3.4 Formulas (3.17) and (3.18) also state the principle of set balance for sets $[0, x)$ and $[x, \infty)$, $x \geq 0$. That is, the equation *sample-path exit rate from set* $[0, x)$ = *sample-path entrance rate into* $[0, x)$ holds. The same principle applies to set $[x, \infty)$. Moreover, **SP** motion contains the sample path as a subset; i.e., **SP** motion includes the "not-in-Time" state-space jumps (see Sect. 2.3 in Chap. 2). Hence the same principle applies to **SP** exits and entrances.

3.2.7 Upcrossings and Transient PDF of Wait

Theorem 3.4 below connects the sample-path quantity $\frac{\partial}{\partial t}E(\mathcal{U}_t(x))$ to the analytical transient mixed pdf $\{P_0(t),\ f_t(y)\}_{0<y<x}, t > 0$.

Theorem 3.4 In the M/G/1 queue with arrival rate λ and service time cdf $B(\cdot)$

$$\frac{\partial}{\partial t}E(\mathcal{U}_t(x)) = \lambda\overline{B}(x)P_0(t) + \lambda\int_{y=0}^{x}\overline{B}(x-y)f_t(y)dy \tag{3.19}$$

$$\frac{\partial}{\partial t}E(\mathcal{U}_t(0)) = \lambda P_0(t). \tag{3.20}$$

Proof We define $\mathcal{U}diff := \mathcal{U}_{t+h}(x) - \mathcal{U}_t(x)$ here for brevity. Let $x > 0, t > 0$, and small $h > 0$ be given. We examine possible SP upcrossings of level x in the state-space infinite rectangle $\sqcup := \{(t, t+h) \times (0, \infty)\}$ (Fig. 3.4). We consider possible SP entrances into $\sqcup$ at time t. Entrances that occur: above $x + h$ imply $\mathcal{U}diff = 0$; within $(x, x+h)$ imply $\mathcal{U}diff = 0$ ***or*** $P(\mathcal{U}diff = 1) = [\lambda h + o(h)]\,\overline{B}(x-y)f_t(y)dy$ for some $y \in (x-h, x)$; within $(0, x)$ imply $\mathcal{U}diff = 0$ ***or*** $P(\mathcal{U}diff = 1) = [\lambda h + o(h)]\,\overline{B}(x-y)f_t(y)dy$ for some $y \in (0, x-h)$; at level 0 imply $\mathcal{U}diff = 0$ ***or*** $P(\mathcal{U}diff = 1) = [\lambda h + o(h)]\,\overline{B}(x)P_t(0)$. For any $n \geq 2$, $P(\mathcal{U}diff = n) = o(h)$ because at least

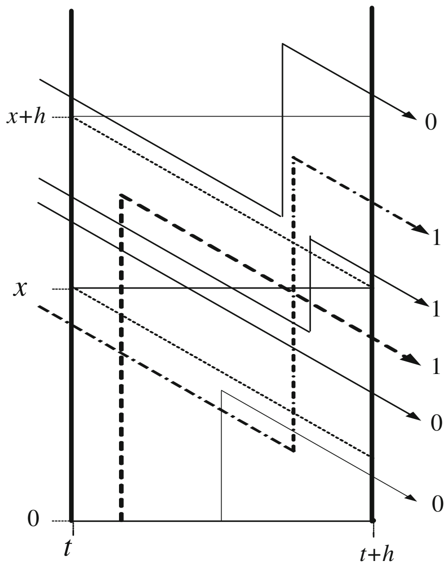


Fig. 3.4 Examples of sample-path (and SP) entrances and exits of set shaped like ⊔. Numbers at the ends of path segments are values of *Udiff*

n arrivals would be required during $(t, t+h)$ (equivalently during $(0, h)$ due to the memoryless property). Finally we get

$\mathcal{U}_{t+h}(x) - \mathcal{U}_t(x)$	Probability
$+1$	$[\lambda h + o(h)]\, P_0(t)\overline{B}(x)$ $+ [\lambda h + o(h)] \int_0^x \overline{B}(x-y) f_t(y) dy + o(h)$
≥ 2	$o(h)$.

(3.21)

In (3.21), the values $\mathcal{U}_{t+h}(x) - \mathcal{U}_t(x) \leq 0$ are omitted since $\{\mathcal{U}\ \mathit{diff} = 0\}$ does not affect $E(\mathcal{U}\mathit{diff})$, and negative values are not possible because $\mathcal{U}_t(x)$ is a counting process (non-decreasing). Utilizing (3.21) gives

$$E(\mathcal{U}_{t+h}(x) - \mathcal{U}_t(x)) = [\lambda h + o(h)]\, P_0(t)\overline{B}(x) + [\lambda h + o(h)] \int_{y=0}^x \overline{B}(x-y) f_t(y) dy + o(h).$$

Dividing both sides by h and taking limits as $h \downarrow 0$ gives (3.19). Letting $x \downarrow 0$ in (3.19) gives (3.20) since $\mathcal{U}_t(0) \equiv \mathcal{U}_t(0^+)$, and $\overline{B}(0) = 1$, since $\overline{B}(x)$ is right continuous. ■

Corollary 3.5 For fixed $t > 0$,

$$E(\mathcal{U}_t(x)) = \lambda \int_{s=0}^t \overline{B}(x) P_0(s) ds + \lambda \int_{s=0}^t \int_{y=0}^x \overline{B}(x-y) f_s(y) dy ds,\ x > 0, \tag{3.22}$$

$$E(\mathcal{U}_t(0)) = \lambda \int_{s=0}^t P_0(s) ds. \tag{3.23}$$

Proof Integrate over time from 0 to t in (3.19) and (3.20). The constants of integration are 0 because $E(\mathcal{U}_0(x)) = 0,\ x \geq 0$. ■

Corollary 3.6 If the steady state exists, then

$$\lim_{t\to\infty} \frac{\partial}{\partial t} E(\mathcal{U}_t(x)) = \lim_{t\to\infty} \frac{E(\mathcal{U}_t(x))}{t} = \lambda \overline{B}(x) P_0 + \lambda \int_0^x \overline{B}(x-y) f(y) dy, \tag{3.24}$$

$$\lim_{t\to\infty} \frac{\partial}{\partial t} E(\mathcal{U}_t(0)) = \lim_{t\to\infty} \frac{E(\mathcal{U}_t(0))}{t} = \lambda P_0. \tag{3.25}$$

Proof Note that

$$\lim_{t\to\infty} F_t(x) = F(x),\quad \lim_{t\to\infty} f_t(x) = f(x),\quad \lim_{t\to\infty} P_0(t) = P_0.$$

In (3.24) and (3.25), the results for

$$\lim_{t\to\infty}\frac{\partial}{\partial t}E(\mathcal{U}_t(x)) \text{ and } \lim_{t\to\infty}\frac{\partial}{\partial t}E(\mathcal{U}_t(0))$$

follow from (3.19) and (3.20) respectively. The results for

$$\lim_{t\to\infty}\frac{E(\mathcal{U}_t(x))}{t} \text{ and } \lim_{t\to\infty}\frac{E(\mathcal{U}_t(0))}{t}$$

follow from (3.22) and (3.23). ■

Corollary 3.6 completes the alternative transient-analysis derivation of Theorem 1.1, which seems to provide a more general perspective than the equilibrium-analysis approach of Sect. 1.6.

3.2.8 Integro-differential Equation for PDF of Wait

We apply LC to derive the ***Takács integro-differential equation*** for the transient probability distribution of wait, by utilizing Theorems 3.2, 3.3 and 3.4 above. (See Remarks 3.5, 3.6 and 3.7 below.)

Theorem 3.5 For an M/G/1 queue with arrival rate λ and service time cdf $B(\cdot)$, the transient distribution of the virtual wait satisfies the following equations for each $t > 0$:

$$f_t(x) = \frac{\partial}{\partial t}F_t(x) + \lambda\overline{B}(x)P_0(t) + \lambda\int_{y=0}^{x}\overline{B}(x-y)f_t(y)dy,\ x > 0, \tag{3.26}$$

$$f_t(0) = \frac{\partial}{\partial t}P_0(t) + \lambda P_0(t), \tag{3.27}$$

$$P_0(t) + \int_{y=0}^{\infty} f_t(y)dy = 1. \tag{3.28}$$

Proof The theorem follows by applying (3.7), substituting from (3.8), (3.9), (3.19), (3.20), and using (3.1). Equation (3.28) is the normalizing condition. ■

Remark 3.5 Equation (3.26) was derived by Takács in [139] by a different technique. Also, see formula (17), p. 87 in [140].

Remark 3.6 Minor extensions of the proofs in this section yield relationships and integro-differential equations for the transient pdf of wait in the important cases where the arrival rate and probability distribution of the service time are also time-dependent. In the formulas (3.26) and (3.27) we can replace λ by $\lambda(t)$ so that the arrival process is non-homogeneous Poisson; and $B(y)$ by $B_t(y)$ so that the service time is time-dependent (see Sect. 3.2.9 below). Equation (3.26) is called in the literature the Takács integro-differential equation (see [139]; formula (5.172), p. 227 in [104]).

Remark 3.7 The LC proofs of (3.26) and (3.27) have important ramifications. The relationship of both sides of (3.26) and (3.27) to $E(\mathcal{D}_t(x))$, $E(\mathcal{U}_t(x))$, $x \geq 0$, leads to techniques for **LC estimation of the transient distributions** by simulation of multiple independent sample paths (see Remark 9.2 in Sect. 9.2). **LC estimation (computation, approximation) for steady-state distributions is discussed in** Chap. 9. LC estimation is a form of non-parametric distribution (or density) estimation.

Example 3.1 below illustrates how transient sample-path quantities can be used to solve transient integro-differential equations numerically for analytical transient pdfs or transient probabilities.

Example 3.1 Assume $W(0) = x_0$ (≥ 0) so that $P_0(x_0) = 1$. Note that $P_0(s) = 0, 0 \leq s \leq x_0$. What is a point estimate of $P_t(0)$ for a finite time $t > x_0$? From Eqs. (3.27) and (3.9) we have the differential equation

$$\frac{\partial}{\partial t}P_0(t) + \lambda P_0(t) = \frac{\partial}{\partial t}E(\mathcal{D}_t(0)), t > x_0. \tag{3.29}$$

Using integrating factor $e^{\lambda t}$ in (3.29) and solving for $P_0(t)$ we get

$$\begin{aligned}\frac{d}{dt}\left(e^{\lambda t}P_0(t)\right) &= e^{\lambda t}\frac{d}{dt}E(\mathcal{D}_t(0)),\\ e^{\lambda t}P_0(t) &= \int_{s=x_0}^{t} e^{\lambda s}\left[\frac{d}{ds}E(\mathcal{D}_s(0))\right]ds + A\end{aligned}$$

where A is a constant. Letting $t \downarrow x_0$ implies $A = e^{\lambda x_0}P_0(x_0) = e^{\lambda x_0}$, and

$$P_0(t) = e^{-\lambda t}\left(\int_{s=x_0}^{t} e^{\lambda s}\left[\frac{d}{ds}E(\mathcal{D}_s(0))\right]ds\right) + e^{-\lambda(t-x_0)}, t > x_0. \tag{3.30}$$

We estimate the function $\frac{d}{ds}E(\mathcal{D}_s(0))$, $x_0 \leq s \leq t$ in (3.30) as follows. Select a partition on $[x_0, t]$ having small norm h such that $t - x_0 = \nu h, \nu \in \mathbb{N}^+$ (set of positive integers). E.g., if $t - x_0$ is rational or irrational select $h = 0.001\,(t - x_0)$ or $0.0002\,(t - x_0)$, etc. Simulate N independent sample paths of $W(s)$, $s \in [0, t + h]$, where N is large. Let $\mathcal{D}_{i,j}(0)$:= number of downcrossings of level 0 (left continuous hits from above) during time intervalls $[x_0 + (j-1)h, x_0 + jh]$, $j = 1, \ldots, \nu + 1$ for the ith sample path, $i = 1, \ldots, N$. Let $\overline{\mathcal{D}}_j(0) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{D}_{i,j}(0)$, $j = 1, \ldots, \nu + 1$. An estimate of $\frac{d}{ds}E(\mathcal{D}s(0))$ is the step function

$$\frac{\overline{\mathcal{D}}_j(0)}{h},\ x_0 + (j-1)\,h < s < x_0 + jh,\ j = 1, \ldots, \nu + 1.$$

Substituting $\frac{\overline{\mathcal{D}}_j(0)}{h}$ into (3.30), we get the point estimate of $P_0(t)$ as

$$\begin{aligned}\widehat{P}_0(t) &= \frac{e^{-\lambda t}}{h}\sum_{j=1}^{\nu+1}\overline{\mathcal{D}}_j(0)\int_{s=x_0+(j-1)h}^{x_0+jh} e^{\lambda s}ds + e^{-\lambda(t-x_0)} \\ &= \frac{e^{-\lambda t}}{\lambda}\sum_{j=1}^{\nu+1}\left(\overline{\mathcal{D}}_j(0)\cdot\frac{e^{\lambda(x_0+jh)} - e^{\lambda(x_0+(j-1)h)}}{h}\right) + e^{-\lambda(t-x_0)}. \quad (3.31)\end{aligned}$$

FORTRAN-programmed computations were carried out in the Masters project [120] to estimate $P_0(t)$ when $x_0 = 0$, using a special case of the method outlined in this example. The computations generally agreed with the known analytical value of $P_0(t)$, $t > 0$, computed from the analytic formula given in [140].

Remark 3.8 Concepts in Example 3.1 relate to renewal theory since downcrossings of level 0 occur at the ends of **busy cycles**, which are i.i.d. random variables forming a renewal process (see formula (5.1) p. 189 in [99]). This will be discussed further in Chap. 10.

3.2.9 PDF When Arrivals and Service Are Time Dependent

We very briefly revisit the *transient pdf* of wait in the M/G/1 queue in Theorem 3.5 above in Sect. 3.2.8. We can prove by a slight generalization of the proofs in Sect. 3.2, that the theory holds for models where the arrival rate λ and cdf of service time $B(x)$, $x > 0$, depend on time t. Denoting them by $\lambda(t)$ and $B_t(x)$, $x > 0$, respectively, we obtain

$$\begin{aligned}f_t(x) &= \frac{\partial}{\partial t}F_t(x) + \lambda(t)\,\overline{B}_t(x)P_0(t) \\ &\quad + \lambda(t)\int_{y=0}^{x}\overline{B}_t(x-y)f_t(y)dy,\ x > 0, \quad (3.32)\\ f_t(0) &= \frac{\partial}{\partial t}P_0(t) + \lambda(t)\,P_0(t).\end{aligned}$$

The solution of the differential equation for $P_0(t)$ in (3.32) is

$$P_0(t) = e^{-m(t)} \int_{s=0}^{t} e^{m(s)} f_s(0) ds + P_0(0) e^{-m(t)}, \tag{3.33}$$

where $m(t) = \int_{s=0}^{t} \lambda(s)\, ds$ and $P_0(0) = \begin{cases} 1 \text{ if } W(0) = 0, \\ 0 \text{ otherwise.} \end{cases}$

3.2.10 Steady-State PDF of Wait from Transient PDF

Equation (1.8) for the steady state distribution of wait, is now proved directly from the foregoing LC connections between sample paths and the transient distribution of wait. The next theorem gives two such proofs.

Theorem 3.6 For an M/G/1 queue with arrival rate λ and service time S having cdf $B(\cdot)$, where $\lambda E(S) < 1$, the steady state pdf of the virtual wait $\{P_0, f(x)\}_{x>0}$, is given by

$$f(x) = \lambda \overline{B}(x) P_0 + \lambda \int_0^x \overline{B}(x-y) f(y) dy, \, x > 0, \tag{3.34}$$

$$f(0) = \lambda P_0, \tag{3.35}$$

$$P_0 + \int_0^\infty f(y) dy = 1. \tag{3.36}$$

Proof Since $\lambda E(S) < 1$, the transient probability distribution converges to the steady state probability distribution, i.e., $\lim_{t\to\infty} F_t(x) = F(x)$, $\lim_{t\to\infty} f_t(x) = f(x)$, $\lim_{t\to\infty} P_0(t) = P_0$. Moreover

$$\lim_{t\to\infty} \frac{\partial}{\partial t} F_t(x) = 0, \ x \geq 0, \ \lim_{t\to\infty} \frac{\partial}{\partial t} P_0(t) = 0.$$

Then (3.34) and (3.35) follow from Theorem 3.5 by letting $t \to \infty$.

Alternatively, (3.34) and (3.35) follow from the principle of rate balance expressed in (3.17), (3.18), and substituting from (3.14), (3.15), (3.24), and (3.25). ■

Remark 3.9 For the M/G/1 queue with $\lambda E(S) < 1$, it is well known that

$$\lim_{t\to\infty} P(W(t) \leq x) = \lim_{n\to\infty} P(W_n \leq x), \ x \geq 0,$$

where W_n is the waiting time of the nth customer (arrival-point wait) (see [140]). Hence Eqs. (3.34)–(3.36) hold for the steady-state distributions of both the arrival-point wait and the virtual wait.

Remark 3.10 Using LC to derive (3.34)–(3.36) is useful because each algebraic term corresponds to a unique down- or upcrossing rate of level $x \geq 0$. This one-to-one correspondence enables the derivation of exact analytical integral equations for steady-state distributions of state variables in many complex stochastic models, intuitively and straightforwardly, using the sample path as a template. The idea is to construct a pertinent typical sample path of the stochastic model; then write the integral equation(s) by inspection using LC theorems and the principle of rate and/or set balance. The solution of the equation(s) is found with the aid of initial conditions (e.g., $f(0) = \lambda P_0$, $f'(0) = -\lambda P_0 b(0) + \lambda^2 P_0$). This procedure can save time and help the analyst focus on the model dynamics.

3.3 Steady-State Distribution of Wait

We begin with Example 3.2 below, which illustrates the derivation of the steady-state pdf of wait in an M/G/1 queue, where $G := \text{Erl}_{k,\mu}$ is the sum of k i.i.d. Exp_μ r.v.s. ($\text{Erl}_{k,\mu}$ denotes an Erlang r.v.; Exp_μ denotes an exponential r.v. with mean $1/\mu$.) In the M/$\text{Erl}_{k.\mu}$/1 queue $E(S) = k \cdot E(\text{Exp}_\mu) = k/\mu$.

Example 3.2 Consider the M/$\text{Erl}_{k.\mu}$/1 queue with arrival rate λ. Let $\mathcal{S}_k(x)$:= event {sum of k i.i.d. Exp_μs $\leq x$}, and $\mathcal{G}_k(x)$:= event {number of Poi_μ events in $(0, x)$ is $\geq k$}, where Poi_μ denotes a Poisson process with rate μ (see pp. 312–316 in [125]). Since $\mathcal{S}_k(x) \iff \mathcal{G}_k(x)$, we have $P(\mathcal{S}_k(x)) = P(\mathcal{G}_k(x))$, and cdf $B(x) = P(S \leq x) = P(\mathcal{S}_k(x)) = P(\mathcal{G}_k(x))$. Therefore

$$B(x) = P(\mathcal{G}_k(x)) = \sum_{i=k}^{\infty} \frac{e^{-\mu x}(\mu x)^i}{i!}, x > 0. \tag{3.37}$$

(See Sect. 2.3.2 for Exp_μ; Chap. 5 for Poi_μ, in [125].) Taking $\frac{d}{dx}$ in (3.37) readily shows that $b(x)$ ($:= \frac{d}{dx}B(x)$) is given by

$$b(x) = e^{-\mu x}\frac{(\mu x)^{k-1}\mu}{(k-1)!}, x > 0. \tag{3.38}$$

(Intuitively, (3.38) is equivalent to '$b(x)dx = P((k-1)$ Poi_μ events in $(0, x)$ and the kth event occurs at time $x)\, dx$'.) Since $\sum_{i=0}^{\infty} e^{-\mu x}(\mu x)^i/i! = 1$,

$$\overline{B}(x) = 1 - B(x) = e^{-\mu x}\left(\sum_{i=0}^{k-1} \frac{(\mu x)^i}{i!}\right), x \geq 0. \tag{3.39}$$

The condition for existence of the steady state is $\lambda E(S) < 1$ or $\lambda < \mu/k$.

Substituting $e^{-\mu x}\left(\sum_{i=0}^{k-1}\frac{(\mu x)^i}{i!}\right)$ for $\overline{B}(x)$ in (3.34), we obtain

$$\begin{aligned} f(x) &= \lambda P_0 e^{-\mu x}\left(\sum_{i=0}^{k-1}\frac{(\mu x)^i}{i!}\right) \\ &\quad + \lambda\int_{y=0}^{x} e^{-\mu(x-y)}\left(\sum_{i=0}^{k-1}\frac{(\mu(x-y))^i}{i!}\right) f(y)dy,\ x > 0. \end{aligned} \tag{3.40}$$

where $P_0 = 1 - \lambda E(S) = 1 - (\lambda k)/\mu$. The normalizing condition is

$$P_0 + \int_{x=0}^{\infty} f(x)dx = 1.$$

Case $k = 2$: We illustrate the solution when $k = 2$, which corresponds to the M/Erl$_{2,\mu}$/1 queue. From (3.40) we have

$$f(x) = \lambda P_0 e^{-\mu x}(1 + \mu x) + \lambda\int_{y=0}^{x} e^{-\mu(x-y)}(1 + \mu(x - y)) f(y)dy, x > 0. \tag{3.41}$$

Differentiating (3.41) with respect to x twice results in the second order differential equation

$$f''(x) + (2\mu - \lambda) f'(x) + (\mu^2 - 2\lambda\mu) f(x) = 0, x > 0,$$

with solution

$$f(x) = a_1 e^{r_1 x} + a_2 e^{r_2 x}, x > 0, \tag{3.42}$$

where a_1, a_2 are constants to be determined, and

$$r_1 = -\mu + \frac{\lambda}{2} - \frac{1}{2}\sqrt{\lambda^2 + 4\mu\lambda},\quad r_2 = -\mu + \frac{\lambda}{2} + \frac{1}{2}\sqrt{\lambda^2 + 4\mu\lambda},$$

are the solutions of the characteristic function

$$z^2 + (2\mu - \lambda)z + (\mu^2 - 2\lambda\mu) = 0.$$

Both $r_1 < 0$, $r_2 < 0$. The constants a_1, a_2 and P_0 can be determined from the initial condition $f(0) = \lambda P_0$, and the normalizing condition $P_0 + \int_{y=0}^{\infty} f(y)dy = 1$, giving

$$a_1 = \frac{r_1 r_2}{r_1 - r_2}(1 - P_0 + \frac{\lambda P_0}{r_2}),$$
$$a_2 = \lambda P_0 - a_1,$$
$$P_0 = 1 - \frac{2\lambda}{\mu}.$$

3.3.1 Alternative LC Equations for PDF of Wait

We now give two different forms of the basic integral equation (1.8) for the limiting pdf of wait in the M/G/1 queue (see Fig. 1.6). The alternative forms are useful due to their applicable LC interpretation. We can write (1.8) as

$$\begin{aligned} f(x) &= \lambda(1 - B(x))P_0 + \lambda \int_{y=0}^{x} (1 - B(x-y)) f(y) dy \\ &= \lambda \left(P_0 + \int_{y=0}^{x} f(y) dy \right) - \lambda \left(B(x) P_0 + \int_{y=0}^{x} B(x-y) f(y) dy \right) \end{aligned}$$

which gives *two* alternative forms of the LC equation:

$$f(x) = \lambda F(x) - \lambda \int_{y=0}^{x} B(x-y) dF(y), \ \ x \geq 0; \tag{3.43}$$

$$f(x) = \lambda F(x) - \lambda \int_{y=0}^{x} F(x-y) dB(y), \ \ x \geq 0, \tag{3.44}$$

noting that $F(x) = P_0 + \int_{y=0}^{x} f(y) dy$, and $F(\infty) = 1$. Formulas (3.43) and (3.44) have intuitive LC interpretations which help us write them immediately. Consider a sample path of the virtual wait (Fig. 1.4) and observe a one-to-one correspondence between the set of algebraic terms in the equations and a set of mutually exclusive and exhaustive sample-path crossings of level x, different from those depicted in Fig. 1.6.

In (3.43) and (3.44) the left side is the SP downcrossing rate of level x, as usual (see formula (3.14)). However, on the right side, $\lambda F(x)$ is the rate of *all* SP jumps that start in the state-space interval $[0, x]$. The second term subtracts off the rate of such jumps *that end below level* x (do not upcross x). Therefore the right side is precisely the total rate at which SP jumps upcross level x. Rate balance, (3.17) or (3.18), gives these equations directly. Note that (3.43) yields (3.44) by using the transformation $z = x - y$, $dz = -dy$, and integrating by parts.

Equations (3.43) and (3.44) are useful when analyzing variants of M/D/1 and M/Discrete/1 queues (Sects. 3.10 and 3.11); they help us derive the steady-state cdf $F(x)$ directly since $f(x) = F'(x)$. They are also useful in theoretical applications, such as in TAM (transform approximation method) [87, 129, 130]. The LC *intuitive* interpretations of (3.43) and (3.44) also suggest how to use LC to develop integral equations for the pdf of state variables in more general models.

Example 3.3 Consider the $M/U_{(0,c)}/1$ queue with arrival rate λ, where the service time $S \underset{dis}{=} U_{(0,c)}$, a uniformly distributed r.v. on $(0, c)$, $c > 0$, i.e.,

$$B(x) = \begin{cases} 0, x < 0, \\ \frac{x}{c}, 0 \leq x < c, \\ 1, x \geq c. \end{cases} \tag{3.45}$$

We assume: $\{W(t)\}_{t \geq 0}$ is unbounded, i.e., $0 \leq W(t) < \infty$; the steady state pdf $\{P_0, f(x)\}_{x>0}$ and cdf $F(x), x \geq 0$, exist. A necessary and sufficient condition for the steady state to exist is $\lambda E(S) < 1 \iff (\lambda c/2) < 1$. Then busy periods are finite $(a.s.)$, and $P_0 = 1 - \lambda E(S) = 1 - \lambda \frac{c}{2}$.

Solution Approach in Example 3.3
We first solve (3.52) for $f(x), 0 < x < c$; then we indicate the iteration on successive state-space intervals $[c, 2c), [2c, 3c), \ldots$. (In Sect. 3.10 we obtain a complete solution for the M/D/1 queue, using a similar technique.)

Substituting from (3.45) into (3.43) and using $F(x - c) = 0, 0 < x < c$, gives (3.46).

$$f(x) = \lambda F(x) - \lambda \int_{y=0}^{x} \frac{(x-y)}{c} dF(y), 0 < x < c, \tag{3.46}$$

$$f(x) = \lambda F(x) - \lambda F(x-c) - \lambda \int_{y=x-c}^{x} \frac{(x-y)}{c} dF(y), x \geq c. \tag{3.47}$$

The LC explanation of (3.46) is the same as for 3.43. In (3.47) on the right side, $\lambda F(x)$ is the total rate of jumps that start below x. The term $-\lambda F(x - c)$ subtracts off the rate of jumps that start at any $y \in [0, x - c)$, and thus cannot upcross level x. The term $-\lambda \int_{y=x-c}^{x} \frac{(x-y)}{c} dF(y)$ subtracts off the rate of jumps that start in $(x - c, x)$ but are too small to upcross x.

Differentiating (3.46) twice with respect to x results in the second order linear homogeneous differential equation

$$f''(x) - \lambda f'(x) + \frac{\lambda}{c} f(x) = 0, x \in (0, c). \tag{3.48}$$

with characteristic (also called "auxiliary") equation

$$r^2 - \lambda r + \frac{\lambda}{c} = 0,$$

having solution

$$r = \frac{1}{2}\left(\lambda \pm \sqrt{\lambda^2 - 4\frac{\lambda}{c}}\right). \tag{3.49}$$

This gives

$$f(x) = a_1 . e^{\frac{\lambda}{2}x} \cos\left(\frac{1}{2}\sqrt{\frac{4\lambda}{c} - \lambda^2} \cdot x\right) + a_2 \cdot e^{\frac{\lambda}{2}x} \sin\left(\frac{1}{2}\sqrt{\frac{4\lambda}{c} - \lambda^2} \cdot x\right), x \in (0, c), \tag{3.50}$$

where a_1, a_2 are constants to be determined. The *cos* and *sin* functions occur because we assumed that $\lambda < 2/c < 4/c$, so that the discriminant $\sqrt{\lambda^2 - 4\frac{\lambda}{c}}$ in (3.49) is a complex number (see Sect. 3.5, pp. 106–114 in [10]). In (3.46), applying the initial conditions $f(0) = \lambda P_0$, $f'(0) = \lambda^2 P_0 - \frac{\lambda P_0}{c}$ with $P_0 = 1 - \frac{\lambda c}{2}$, gives a_1, a_2 in (3.50) as

$$a_1 = \lambda(1 - \frac{\lambda c}{2}), \quad a_2 = \frac{(1 - \frac{\lambda c}{2})\lambda(\lambda - \frac{1}{c})}{\sqrt{\frac{4\lambda}{c} - \lambda^2}}.$$

We can iterate to solve for $f(x)$, $x \in [c, 2c)$, $x \in [2c, 3c)$, etc., by using (3.50). For $x \in [c, 2c)$, we have

$$f(x) = \lambda F(x) - \lambda \int_{y=c}^{x} \frac{(x-y)}{c} dF(y) - \lambda \int_{y=x-c}^{c} \frac{(x-y)}{c} f(y) dy - \lambda F(x-c), c \leq x < 2c. \tag{3.51}$$

We solve (3.51) by substituting $f(y)$ from (3.50) on the interval $(x - c, c)$ into the second integral in (3.51), Then use discontinuity at $x = c$, i.e., $f(c^+) - f(c^-) = -\lambda P_0$ (letting $x \downarrow c$ in (3.51), $x \uparrow c$ in (3.50), and subtracting). The computation of $f(x)$, $c < x < 2c$ by stepping upward from state-space interval $(0, c)$ to interval $[c, 2c)$ is iterated on intervals $[ic, (i+1)c)$, $i \geq 2$. (A similar discontinuity in the pdf $f(x)$ occurs at $x = D$ in the M/D/1 queue considered below in Sect. 3.12.)

Example 3.4 Now we assume a workload-bounded M/U$_{(0,c)}$/1 queue, i.e., $\{W(t)\}_{t\geq 0}$ is bounded at level $\boldsymbol{K} > 0$. To demonstrate the solution technique we let $\boldsymbol{K} := c$, and assume all service times that cause the virtual wait to exceed level c are truncated at level c. (See variant 1 in Sect. 3.16 and Fig. 3.33.)

The steady-state cdf $F(x)$ exists for all $\lambda > 0$ (see Sect. 2.1 in [25]). Substituting from (3.45) into (3.43) and using $F(x-c) = 0$, $0 < x < c$, gives

$$\begin{aligned} F'(x) &= \lambda F(x) - \lambda \int_{y=0}^{x} \tfrac{(x-y)}{c} dF(y) \\ &= \lambda F(x) - \lambda \int_{y=0}^{x} \tfrac{(x-y)}{c} f(y)dy - \lambda P_0 \tfrac{x}{c}, \; 0 < x < c, \end{aligned} \tag{3.52}$$

for the steady-state cdf $F(x)$.

Solution Approach for Example 3.4
Taking d/dx in (3.52) leads to the second order differential equation

$$F''(x) - \lambda F'(x) + \frac{\lambda}{c} F(x) = 0, \; 0 \leq x \leq c. \tag{3.53}$$

Assuming $\lambda > 4/c$, the solution of (3.53) is

$$F(x) = a_1 \cdot e^{r_1 x} + a_2 \cdot e^{r_2 x}, \; 0 \leq x \leq c, \tag{3.54}$$

where

$$r_1 = \frac{\frac{1}{2}\left(\lambda c + \sqrt{c^2\lambda^2 - 4c\lambda}\right)}{c}, \; r_2 = \frac{\frac{1}{2}\left(\lambda c - \sqrt{c^2\lambda^2 - 4c\lambda}\right)}{c},$$

and a_1 and a_2 are constants to be determined. Using the initial conditions

$$\begin{aligned} F(0) &= a_1 + a_2 = P_0, \\ F'(0^+) &= a_1 \cdot r_1 + a_2 \cdot r_2 = \lambda P_0, \end{aligned}$$

results in

$$a_1 = \frac{r_2 - \lambda}{r_1 - r_2} P_0, \quad a_2 = \frac{r_1 - \lambda}{r_1 - r_2} P_0. \tag{3.55}$$

From (3.54) we get

$$F(x) = P_0 \left(\frac{r_2 - \lambda}{r_1 - r_2} \cdot e^{r_1 x} + \frac{r_1 - \lambda}{r_1 - r_2} \cdot e^{r_2 x} \right), 0 \leq x \leq c; \tag{3.56}$$

by using the boundary condition $F(c) = 1$,

$$P_0 = \left[\frac{r_2 - \lambda}{r_1 - r_2} \cdot e^{r_1 c} + \frac{r_1 - \lambda}{r_1 - r_2} \cdot e^{r_2 c} \right]^{-1}.$$

Generalization When Workload Bound is Greater Than c
Suppose the workload bound $\boldsymbol{k}$ is such that $c < \boldsymbol{k} < 2c$. Define $F(x) := F_0(x) \cdot \boldsymbol{I}_{[0,c]}(x) + F_1(x) \cdot \boldsymbol{I}_{[c,\boldsymbol{k}]}(x)$, where $\boldsymbol{I}_{\boldsymbol{A}}(x) = 1$ if $x \in A$, and 0 if $x \notin A$. The corresponding pdfs are $f_i(x) = dF_i(x)/dx$, $i = 0, 1$. Thus $F_0(x) = a_0 \cdot e^{r_1 x} + b_0 \cdot e^{r_2 x}$, $0 \leq x \leq c$ as in (3.54). (Here a_0 and b_0 will have different values than a_1, a_2 in (3.55) because now $F_1(\boldsymbol{k}) = 1$.) An integral equation for $F_1(x)$, $c \leq x \leq \boldsymbol{k}$, is given in terms of $F_0(\cdot)$ and $f_0(\cdot)$ as

$$\begin{aligned} F_1'(x) =& \lambda F_1(x) - \lambda \int_{y=c}^{x} \frac{(x-y)}{c} f_1(y)dy - \lambda F_0(x-c) \\ & - \lambda \int_{y=x-c}^{c} \frac{(x-y)}{c} f_0(y)dy, c < x \leq k. \end{aligned} \tag{3.57}$$

If the bound $\boldsymbol{k} \in (jc, (j+1)c]$, $j \in \mathbb{N}^+$, we can iterate to solve for $F_{j+1}(x)$, $x \in [jc, k)$, $j = 1, 2, \ldots$, similarly as in (3.57). In the solution, we can use $F_{j+1}(jc^-) = F_j(jc^+)$ by continuity of the cdf at jc, $j = 1, 2, \ldots, \lfloor k/c \rfloor$ to facilitate solving for the constants. A related solution technique is applied for the M/D/1 queue in Sect. 3.12. When numerics are substituted for the parameters λ and c, the solution procedure can be programmed on a computer.

3.3.2 Relating System and Waiting Times Using LC

Let σ denote the *system time* in the M/G/1 queue. Denote the pdf and cdf of σ as $f_\sigma(x)$ and $F_\sigma(x)$, $x > 0$, respectively (see, e.g., Sect. 3.5.2). Then $\sigma = W_q + S$, where W_q is the wait before service and S is the common service time. The pdf and cdf of W_q are $f(x)$, $x > 0$, and $F(x)$, $x \geq 0$, respectively. We use LC interpretations of sample-path quantities to develop an analytical equation relating $f(x)$, $F(x)$ and $F_\sigma(x)$. This is an example where using LC interpretations of sample-path quantities can lead directly to analytical results, or to estimation methods for analytical quantities in particular models (see LC estimation in Chap. 9).

Peaks, Troughs and Downcrossings
A sample path of $\{W(s)\}_{s \geq 0}$ (Fig. 3.5) has a sequence of peaks (relative maxima) and troughs (relative 'minima', which are infima, i.e., greatest lower bounds, due to sample-path *right continuity*). A trough at level 0 is considered

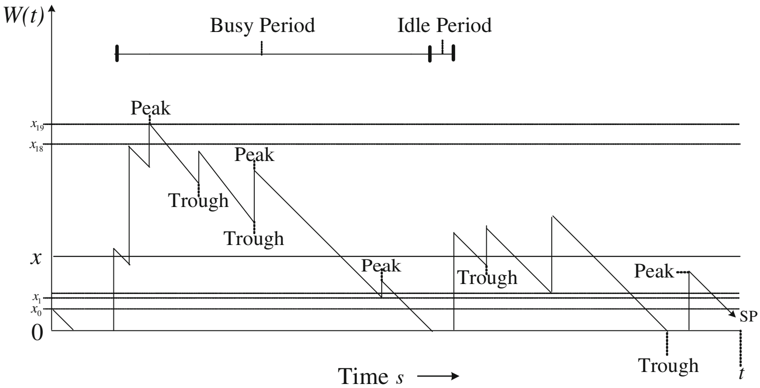


Fig. 3.5 Sample path of virtual wait $\{W(s)\}_{0\le s\le t}$ showing peaks, troughs, level x, and subset of the associated partition $\{0 = x_0 < x_1 < \cdots < x_{18} < x_{19}\}$ which depends on the realized sample path over $[0, t]$

to be an interval which starts at an instant the SP hits level 0 from above, and ends at the next instant the SP leaps (jumps upward) from level zero.

Fix time $t > 0$ and level $x \ge 0$. Let $\mathcal{P}_t^+(x)$, $\mathcal{T}_t^+(x)$ denote respectively the number of peaks and troughs strictly above x during $[0, t)$. When $s = t$ the point $(t, W(t))$ is a trough since $\frac{d}{ds}W(s) = -1 \cdot \boldsymbol{I}_{(0,\infty)}(W(s)) + 0 \cdot \boldsymbol{I}_{\{0\}}(W(s))$. Then $\mathcal{D}_t(x)$ (number of SP downcrossings of x during $(0, t)$), is a step function with respect to x, with constant integer values on subintervals of the partition $\{0 = x_0 < x_1 < \cdots x_{n-1} < x_n\}$, where x_j is the ordinate of a peak or trough ($j = 1, \ldots, n-1$) and x_n is the highest peak during $[0, t]$. Such a fixed partition exists for each realized sample path (Fig. 3.5). An LC interpretation leads to

$$\mathcal{D}_t(x) = \mathcal{P}_t^+(x) - \mathcal{T}_t^+(x), x > 0. \tag{3.58}$$

The values of $\mathcal{D}_t(x)$ in adjacent subintervals, (x_{j-1}, x_j) and (x_j, x_{j+1}), j = 1, 2, …, differ by ± 1, or 0 if S is a continuous random variable. If S has discontinuities, as in M/D/1 in which $S \equiv D$, then the values of $\mathcal{D}_t(x)$ in the two subintervals abutting on D will generally differ by more than 1; in this case, a difference >1 is the result of a discontinuity in the pdf of wait at $x = D$. Formula (3.58) can be useful when simulating sample paths for estimating state-space pdfs.

Equation Relating $f(x)$, $F(x)$ and $F_\sigma(x)$
Let $N_A(t)$ denote the number of arrivals during $(0, t)$. Assume $N_A(t) > 0$. Dividing (3.58) by $t > 0$, we obtain

$$\begin{aligned}\frac{\mathcal{D}_t(x)}{t} &= \frac{P_t^+(x)}{t} - \frac{T_t^+(x)}{t}\\ &= \frac{N_A(t)}{t}\cdot\frac{P_t^+(x)}{N_A(t)} - \frac{N_A(t)}{t}\cdot\frac{T_t^+(x)}{N_A(t)},\ t > 0. \end{aligned} \tag{3.59}$$

Note that $P_t^+(x)$ represents the number of *system times* greater than x in $(0, t)$. Also $\mathcal{T}_t^+(x)$ represents the number of *waiting times* greater than x in $(0, t)$. Taking limits of the terms on the right side of (3.59) as $t \to \infty$ yields

$$\lim_{t\to\infty}\frac{N_A(t)}{t} \underset{a.s}{=} \lambda,\quad \lim_{t\to\infty}\frac{P_t^+(x)}{N_A(t)} \underset{a.s}{=} 1 - F_\sigma(x),\quad \lim_{t\to\infty}\frac{T_t^+(x)}{N_A(t)} \underset{a.s}{=} 1 - F(x),$$

which provides two more alternative forms of the M/G/1 integral equation for the pdf of wait, namely

$$f(x) = \lambda(1 - F_\sigma(x)) - \lambda(1 - F(x)), \tag{3.60}$$

and

$$f(x) = \lambda F(x) - \lambda F_\sigma(x). \tag{3.61}$$

LC Interpretations of (3.60) and (3.61)
On the right side of (3.60) the first term is the rate of *all* jumps that *end above* level x (system time $> x$). The second term subtracts off the rate of those jumps that *start above* level x (wait $> x$). Thus, the right side is the rate of SP jumps that upcross x.

The LC interpretation of (3.61) is that the first term on the right side is the rate of all jumps that *start* in $[0, x]$ (wait $\le x$). The second term subtracts off the rate of those jumps that *end* at levels in $[0, x]$ (system time $\le x$). Thus the right side is the rate of SP jumps that upcross x. Equation (3.61) is equivalent to (3.43) since, by independence of S and W_q

$$\begin{aligned}F_\sigma(x) = P(S + W_q \le x) &= \int_{y=0}^{x} P(S \le x - y | W_q = y) dF(y)\\ &= \int_{y=0}^{x} B(x - y) dF(y).\end{aligned}$$

Remark 3.11 Equation (3.59) combines sample-path peaks and troughs and the key part of the basic LC theorem $\lim_{t\to\infty}\frac{\mathcal{D}_t(x)}{t} \underset{a.s.}{=} f(x)$, for a concrete derivation of integral equation (1.8) (same as (3.34)) in Fig. 1.6, based on LC interpretations of SP motion.

3.4 Waiting Time Properties in Steady State

We derive several familiar properties of the steady-state distribution of the waiting time before service starting from the basic LC integral equation (3.34). We let W_q := *wait before start of service*.

3.4.1 Probability of Zero Wait

In (3.34) integrate both sides with respect to x over $(0, \infty)$. This yields

$$1 - P_0 = \lambda P_0 \int_{x=0}^{\infty} \overline{B}(x)dx + \lambda \int_{x=0}^{\infty} \int_{y=0}^{x} \overline{B}(x-y)f(y)dydx;$$

interchanging the order of integration in the double integral leads to

$$\begin{aligned} 1 - P_0 &= \lambda P_0 E(S) + \lambda E(S)(1 - P_0), \\ P_0 &= 1 - \lambda E(S) = 1 - \rho. \end{aligned} \tag{3.62}$$

3.4.2 Pollaczek-Khinchine (P-K) Formula

In (3.34) multiply both sides by x and integrate with respect to x over $(0, \infty)$. We obtain

$$\int_{x=0}^{\infty} xf(x)dx = \lambda P_0 \int_{x=0}^{\infty} x\overline{B}(x)dx + \lambda \int_{x=0}^{\infty} \int_{y=0}^{x} x\overline{B}(x-y)f(y)dydx.$$

In the double integral, interchange the order of integration, write $x = x - y + y$, and simplify, giving

$$E(W_q) = \lambda P_0 \frac{E(S^2)}{2} + \lambda(1 - P_0)\frac{E(S^2)}{2} + \lambda E(W_q)E(S),$$

from which we obtain the well-known Pollaczek-Khinchine (P-K) formula

$$E(W_q) = \frac{\lambda E(S^2)}{2(1 - \lambda E(S))} = \frac{\lambda E(S^2)}{2\,(1-\rho)} = \frac{\lambda(var(S) + (E(S))^2)}{2P_0}, \tag{3.63}$$

where $var(S) := E(S^2) - (E(S))^2$. (See pp. 220–225 in [84] for a discussion and variations of the P-K formula.)

3.4.3 Expected Number in Queue and in System

Let N_q denote the number of customers waiting in the queue before service; let $L_q = E(N_q)$. From Little's formula $\boldsymbol{L = \lambda W}$ (see [110]), and formula (3.63), we get

$$L_q = \lambda E(W_q) = \frac{\lambda^2 E(S^2)}{2(1-\lambda E(S))} = \frac{\lambda^2 E(S^2)}{2(1-\rho)} = \frac{\lambda^2 E(S^2)}{2P_0}. \tag{3.64}$$

The expected number in the system is

$$L = L_q + L_s$$

where L_s denotes the expected number in service, given by

$$L_s = 1 \cdot (1 - P_0) + 0 \cdot P_0 = \lambda E(S) = \rho.$$

Thus

$$L = \frac{\lambda^2 E(S^2)}{2(1-\lambda E(S))} + \lambda E(S) = \frac{\lambda^2 E(S^2)}{2P_0} + \rho. \tag{3.65}$$

3.4.4 Laplace-Stieltjes Transform (LST) of a PDF

Before deriving the LST of $f(x)$, i.e., the pdf of W_q, we very briefly define the LST and related Laplace transform LT of a function. (See pp. 455–460 in [84] for a concise, clear introduction to the LST and LT.) The LST applies when the function has atoms or is continuous. The LT applies when the function is continuous. (Sect. 11.9 in Chap. 11 below presents an LC technique for estimating the LST and LT.)

LST
The Laplace-Stieltjes transform of $f(x)$ is

$$F^*(s) := E(e^{sW_q}) = \int_{x=0}^{\infty} e^{-sx} dF(x) = P_0 + \int_{x=0}^{\infty} e^{-sx} f(x)dx, s > 0. \tag{3.66}$$

The LST of $b(x)$, i.e., the pdf of the service time, is

$$B^*(s) := \int_{x=0}^{\infty} e^{-sx} dB(x) = \int_{x=0}^{\infty} e^{-sx} b(x)dx.$$

LT

The Laplace transform (LT) of $B(x)$, i.e., the cdf of the service time, is

$$\tilde{B}(s) := \int_{x=0}^{\infty} e^{-sx} B(x)dx.$$

Integrating $\tilde{B}(s)$ by parts shows that $B^*(s) = s\tilde{B}(s)$, $s > 0$.

In (3.34), the basic Volterra integral equation for $f(x)$, $x > 0$, we multiply both sides by e^{-sx} and integrate with respect to x over $(0, \infty)$, giving

$$\begin{aligned}\widetilde{f}(s) &= F^*(s) - P_0 = \int_{x=0}^{\infty} e^{-sx} f(x)dx \\ &= \lambda P_0 \int_{x=0}^{\infty} e^{-sx}\overline{B}(x)dx + \lambda \int_{x=0}^{\infty} e^{-sx} \int_{y=0}^{x} \overline{B}(x-y)f(y)dydx.\end{aligned} \tag{3.67}$$

In the double integral, express e^{-sx}as $e^{-sy} \cdot e^{-s(x-y)}$ and interchange the order of integration, giving

$$\begin{aligned}\widetilde{f}(s) = \lambda P_0 \int_{x=0}^{\infty} e^{-sx}\overline{B}(x)dx \\ +\lambda \int_{y=0}^{\infty} e^{-sy} f(y) \int_{x=y}^{\infty} e^{-s(x-y)}\overline{B}(x-y)dxdy\end{aligned} \tag{3.68}$$

Simplifying yields the well-known formula

$$\begin{aligned}\widetilde{f}(s) &= \frac{sP_0}{s - \lambda(1 - B^*(s))} \\ &= \frac{s(1 - \lambda E(S))}{s - \lambda(1 - B^*(s))} = \frac{1-\rho}{1 - \rho\left(\frac{1-B^*(s)}{sE(S)}\right)}, s > 0,\end{aligned} \tag{3.69}$$

(see p. 237 in [84]). Substituting $\rho := \lambda E(S)$ and expanding $\widetilde{f}(s)$ as a geometric series gives

$$\widetilde{f}(s) = (1-\rho)\sum_{k=0}^{\infty} \rho^k \left(\frac{1 - B^*(s)}{sE(S)}\right)^k. \tag{3.70}$$

3.4.5 *Series for PDF of W_q by Inverting $\widetilde{f}(s)$*

Let γ_S denote the limiting excess service time having pdf $g(x)$, $x > 0$. Generally $g(x) = \overline{B}(x)/E(S)$, $x \in (0, \infty) \cap$ (*domain* of S). (See, e.g., p. 193 in [99]; p. 453 in [125]; p. 317 in [143]; and others.) (In Chap. 10 below we use

LC to derive an analytical expression for $g(x)$, which is denoted as $f_\gamma(x)$ therein.) Then

$$\widetilde{g}(s) = \frac{1}{E(S)} \int_{x=0}^{\infty} e^{-sx}(1 - B(x))dx = \frac{1}{E(S)} \left(\frac{1}{s} - \int_{x=0}^{\infty} e^{-sx} B(x)dx \right)$$
$$= \frac{1}{E(S)} \left(\frac{1}{s} - \frac{B^*(s)}{s} \right) = \frac{1 - B^*(s)}{sE(S)},$$

which is raised to the power k in the series (3.70). Moreover, $(\widetilde{g}(s))^k$ is the LT of the kth self convolution of $g(x)$, which we denote by $g_{(k)}(x)$, with $g_{(0)}(x) \equiv 1$. Since the LT uniquely defines a function and conversely, we can write (3.70) as the series

$$f(x) = (1 - \rho) \sum_{k=0}^{\infty} \rho^k g_{(k)}(x), x > 0, \tag{3.71}$$

which is known as the Beneš series (see [8]). Due to its importance in queueing theory we give several additional references for (3.71): pp. 200–201 in [104]; p. 236 in [84]; Example 7.24, p. 453 in [125]; pp. 169–170 in [99] ; Theorem 18, p. 37 in [118]; and also see [32, 65]; Sect. 10.1.3 in Chap. 11 below. Section 3.17 shows that (3.71) is a special case of a more general series having a term by term level-crossing interpretation.

Probabilistic Interpretation of LT

Remark 3.12 Equations (3.67) and (3.69) can be interpreted as the probability that the waiting time in queue is less than an independent 'catastrophe' random variable, say $Y \underset{dis}{=} \text{Exp}_s$. That is, the wait in queue finishes before the catastrophe occurs with probability $F^*(s)$. This probabilistic interpretation is useful for deriving Laplace transforms of random variables associated with stochastic models (see, e.g., Sect. 7.2, p. 267ff in [104]; Sect. 3 in [41]; and also see [92, 126]; many major papers supervised by M. Hlynka, University of Windsor).

3.4.6 Another Look at System Time

Here we use the notation of Sect. 3.3.2. For an arbitrary arrival, $\sigma > x$ iff the arrival waits in queue $y \leq x$ and its service time exceeds $x - y$, or, the arrival waits in queue $> x$. Thus

$$
\begin{aligned}
1 - F_\sigma(x) &= P(\sigma > x) \\
&= P_0\overline{B}(x) + \int_{y=0}^{x} \overline{B}(x-y) f(y) dy + 1 - F(x) \\
&= \frac{f(x)}{\lambda} + 1 - F(x),
\end{aligned}
\tag{3.72}
$$

implying

$$f(x) = \lambda F(x) - \lambda F_\sigma(x),$$

which is the same as (3.61). If $f(x)$ is known, then $F(x)$ can be computed. Then $F_\sigma(x)$ and $F'_\sigma(x) \equiv f_\sigma(x)$ can be obtained.

3.4.7 Connecting PDFs of System and Waiting Times

We now give a new LC-derived equation connecting $f_\sigma(x)$ directly with $f(x)$. Consider a sample path of the virtual wait and fix level $x > 0$. We view the SP jumps at arrival instants from the *ends* of the jumps (rather than from the starts of the jumps). The level of the end of a jump represents the system time of the corresponding arrival.

The downcrossing rate of level x is given by

$$\lambda \int_{y=x}^{\infty} e^{-\lambda(y-x)} f_\sigma(y) dy,$$

since $\lambda f_\sigma(y)dy$ is the rate of SP jumps that *end* within a "dy" neighborhood about level $y > x$, and $e^{-\lambda(y-x)}$ is the probability that the next customer arrives more than $y - x$ later. Thus the time interval of duration $y - x$ is devoid of new arrivals and associated SP jumps. The SP descends with slope -1 to level x, making a left-continuous downcrossing of x. (In this scenario, the jumps that end 'at' y may start either below x or in state-space interval (x, y). The end level y is the system time of the associated arrival.)

By Theorem 1.1, another expression for the SP downcrossing rate of x is $f(x)$ (also equal to upcrossing rate of x). Hence we have the equation

$$f(x) = \lambda \int_{y=x}^{\infty} e^{-\lambda(y-x)} f_\sigma(y) dy. \tag{3.73}$$

Multiplying both sides of (3.73) by $e^{-\lambda x}$ and differentiating with respect to x yields

$$f_\sigma(x) = f(x) - \frac{f'(x)}{\lambda}, x > 0, \tag{3.74}$$

wherever $f'(x)$ exists. Thus, if $f(x)$ is known, $f_\sigma(x)$ can be found directly using (3.74).

Example 3.5 In M$_\lambda$/M$_\mu$/1, $f(x) = \lambda P_0 e^{-(\mu-\lambda)x}$, $x > 0$ (see (3.112) and (3.117) in Sect. 3.5.2). Substituting $f(x)$ into (3.74) yields

$$\begin{aligned} f_\sigma(x) &= (\mu - \lambda)\, e^{-(\mu-\lambda)x}, x > 0, \\ F_\sigma(x) &= \textstyle\int_{y=0}^{x} f_\sigma(y)dy = 1 - e^{-(\mu-\lambda)x}, x \geq 0, \end{aligned} \tag{3.75}$$

Example 3.6 In M/Erl$_{2,\mu}$/1, the continuous part of the pdf of wait is

$$f(x) = a_1 e^{r_1 x} + a_2 e^{r_2 x}, x > 0;$$

thus

$$f_\sigma(x) = a_1 e^{r_1 x} + a_2 e^{r_2 x} - \frac{a_1 r_1 e^{r_1 x} + a_2 r_2 e^{r_2 x}}{\lambda}, x > 0,$$

where $a_i, r_i, i = 1, 2$ are given in Example 3.2, Sect. 3.3.

3.4.8 Number in System Probability Distribution

We obtain the steady-state probability distribution of the number in the system in two ways: by conditioning on W_q, or conditioning on σ. Let $P_n, n = 0, 1, \ldots,$ denote the probability of n customers in the system at an arbitrary time point (P_n := proportion of time n are in the system). Let a_n, $d_n, n = 0, 1, \ldots,$ denote the steady-state probability of n in the system just before an arrival, and just after a departure, respectively (a_n := proportion of arrivals that "see" n; d_n := proportion of departures that leave n).

For the M/G/1 queue it is well known that $P_n = a_n$ due to Poisson arrivals, and generally $a_n = d_n$ (e.g., pp. 501–502 in [125]; see also in [145]).

Conditioning on W_q, we obtain

$$\begin{aligned} P_n &= \int_{y=0}^{\infty} P(n-1 \text{ arrivals during } y | W_q = y) f(y)dy \\ &= \int_{y=0}^{\infty} e^{-\lambda y} \frac{(\lambda y)^{n-1}}{(n-1)!} f(y)dy, n = 1, 2, \ldots . \end{aligned} \tag{3.76}$$

Equation (3.76) is consistent with $P_0 + \int_{y=0}^{\infty} f(y)dy = 1$ since the proportion of time the system presents a positive wait to a potential arrival is

$$\sum_{n=1}^{\infty} P_n = \int_{y=0}^{\infty} e^{-\lambda y} \sum_{n=1}^{\infty} \frac{(\lambda y)^{n-1}}{(n-1)!} \cdot f(y)dy$$
$$= \int_{y=0}^{\infty} e^{-\lambda y} e^{\lambda y} f(y)dy = \int_{y=0}^{\infty} f(y)dy = 1 - P_0.$$

Alternatively, conditioning on σ,

$$P_n = \int_{y=0}^{\infty} P(n \text{ arrivals during } y|\sigma = y) f_\sigma(y)dy$$
$$= \int_{y=0}^{\infty} e^{-\lambda y} \frac{(\lambda y)^n}{n!} f_\sigma(y)dy, n = 0, 1, \ldots, \tag{3.77}$$

which is also consistent with $P_0 + \int_{y=0}^{\infty} f(y)dy = 1$ since

$$\sum_{n=0}^{\infty} P_n = \int_{y=0}^{\infty} e^{-\lambda y} \left(\sum_{n=0}^{\infty} \frac{(\lambda y)^n}{n!} \right) \cdot f_\sigma(y)dy = \int_{y=0}^{\infty} f_\sigma(y)dy = 1.$$

If $f(\cdot)$, $f_\sigma(\cdot)$ are known for a particular M/G/1 model, either Eq. (3.76) or (3.77) can be applied to yield $\{P_n\}_{n=0,1,\ldots}$. Note that both a_n and d_n are also given by (3.76) or (3.77).

Interestingly

$$P_0 = \int_{y=0}^{\infty} e^{-\lambda y} f_\sigma(y)dy = \widetilde{f}_\sigma(\lambda), \tag{3.78}$$

the Laplace transform of $f_\sigma(\cdot)$. Using the probabilistic interpretation of the LT, formula (3.78) says that $P_0 = P(\sigma < Y)$ where Y is a an independent exponentially distributed "catastrophe" variable having rate λ (see Remark 3.12 in Sect. 3.4.4).

3.4.9 Renewal Reward Theorem: Statement

We state here the renewal reward theorem for easy reference, due to intermittent use in the sequel. The theorem applies generally to regenerative processes, although we state it here with respect to busy cycles in the standard M/G/1 queue. This brief section is based on the references in the Proof section immediately after Eq. (3.79) below.

Theorem Let R_n denote the amount of 'reward' earned during the busy cycle $\mathcal{C}_n$, where $\{R_n\}_{n=1,2,\ldots}$ are i.i.d. random variables. Assume $E(|R_1|) < \infty$, and let $R(t)$ denote the *total* reward earned during the time interval $(0, t)$, $t > 0$. Then $\{R(t)\}_{t \geq 0}$ is called the *renewal reward process*. The key result is

$$\frac{E(R_1)}{E(C)} = \lim_{t \to \infty} \frac{R(t)}{t} \text{ with probability 1.} \tag{3.79}$$

Proof Proofs of (3.79), and related material, are given in the following references: p. 41ff in [143]; p. 439ff in [125]; Proposition 3.4.1, p. 192 in [122]. ■

3.4.10 *Expected Busy Period in M/G/1*

Let $\mathcal{B}$ denote a busy period, $\mathcal{I}$ an idle period, and $\mathcal{C}$ a busy cycle. Then $\mathcal{C} = \mathcal{B} + \mathcal{I}$. The sequence $\{\mathcal{C}_n\}_{n=1,2,\ldots}$, where $\mathcal{C}_n \underset{dis}{=} \mathcal{C}$, forms a renewal process. Consider a sample path of the virtual wait $\{W(t)\}_{t \geq 0}$. $\{W(t)\}_{t \geq 0}$ is a regenerative process with respect to $\{\mathcal{C}_n\}_{n=1,2,\ldots}$. (For discussions on regenerative processes see, e.g., p. 447ff in [125]; p. 215ff in [122]; also see [132, 134], and others.)

Expected Busy Period

We now look at several ways to derive $E(\mathcal{B})$, for perspective.

{1} An expression for the (long-run) expected proportion of time that the sample path is in the state-space interval $(0, \infty)$ is $1 - P_0 = \rho := \lambda E(S)$. A different expression for the same proportion of time is

$$\lim_{t \to \infty} \frac{\mathcal{U}_t(0) E(\mathcal{B})}{t} = \lim_{t \to \infty} \frac{\mathcal{U}_t(0)}{t} E(\mathcal{B}) = \lambda P_0 E(\mathcal{B}),$$

since each exit of level 0 above (upcrossing of 0) initiates an independent busy period; moreover $\lim_{t \to \infty} \mathcal{U}_t(0)/t = \lambda P_0$. Equating these two different expressions gives

$$\lambda P_0 E(\mathcal{B}) = \lambda E(S),$$

$$E(\mathcal{B}) = \frac{E(S)}{P_0}. \tag{3.80}$$

{2} From the elementary renewal theorem (see, e.g., Proposition 7.1, p. 428 and Theorem 7.1, p. 432 in [125]), and LC theory,

$$E(\mathcal{C}) = \frac{1}{\text{downcrossing rate of level 0}} = \frac{1}{f(0)} = \frac{1}{\lambda P_0}. \tag{3.81}$$

In the *renewal reward theorem* let $R_n = \mathcal{B}_n$, where $\mathcal{B}_n$ is the busy period embedded in $\mathcal{C}_n$, $\mathcal{B}_n \underset{dis}{=} \mathcal{B}$, $n = 1, 2, \ldots$. Then $E(R_1) = E(\mathcal{B})$. Equation (3.79) gives

$$\frac{E(\mathcal{B})}{E(\mathcal{C})} = \frac{E(\mathcal{B})}{\frac{1}{\lambda P_0}} = \lim_{t\to\infty} \frac{\text{amount of time server is busy during } (0, t)}{t}$$
$$= \text{proportion of time workload is in } (0, \infty) = \rho = \lambda E(S).$$
$$E(\mathcal{B}) = \frac{\lambda E(S)}{\lambda P_0} = \frac{E(S)}{P_0},$$

which agrees with (3.80).
{3} Since $\mathcal{C} = \mathcal{B} + \mathcal{I}$,

$$E(\mathcal{B}) = E(\mathcal{C}) - E(\mathcal{I}) = \frac{1}{\lambda P_0} - \frac{1}{\lambda} = \frac{1 - P_0}{\lambda P_0} = \frac{E(S)}{P_0}.$$

{4} Intuitively $E(\mathcal{B})$ is the $(1 - P_0)$-th proportion of $E(\mathcal{C})$, i.e.,

$$E(\mathcal{B}) = (1 - P_0) \cdot E(\mathcal{C}) = \frac{1 - P_0}{\lambda P_0} = \frac{\lambda E(S)}{\lambda P_0} = \frac{E(S)}{P_0};$$

this is really a version of the renewal-reward-theorem method.

The appearance of P_0 in the denominator of (3.80) follows from the renewal reward theorem, or from $f(0) = \lambda P_0$ in Theorem 1.1, Corollary 1.1. The expression

$$E(\mathcal{B}) = \frac{1 - P_0}{\lambda P_0} \tag{3.82}$$

appears to be more fundamental than the expression $E(\mathcal{B}) = \frac{E(S)}{1-\lambda E(S)}$, since in some well-known variants of the standard M/G/1 queue, $P_0 \neq 1 - \lambda E(S)$ (e.g., if the workload has a positive barrier (see [25]; also Sects. 3.9 and 3.13 below).
{5} Busy periods and idle periods form an alternating renewal process. Hence

$$P_0 = \frac{E(\mathcal{I})}{E(\mathcal{B}) + E(\mathcal{I})} = \frac{\frac{1}{\lambda}}{E(\mathcal{B}) + \frac{1}{\lambda}} = 1 - \lambda E(S);$$

the last equality implies (3.82). This derivation also assumes the renewal reward theorem, so is similar to derivation {2}. However, it does not directly "explain" the appearance of P_0 in the denominator; derivation {2} does provide the explanation.

Remark 3.13 Formula (3.82) shows immediately that

$$E(\mathcal{B}) < \infty \text{ iff } 0 < P_0 \leq 1,$$

and equivalently

$$E(\mathcal{B}) = \infty \text{ iff } P_0 = 0.$$

The **stability condition** for the standard M/G/1 queue is $P_0 > 0$ (same as $\lambda E(S) < 1$). The queue is stable iff state $\{0\}$ is positive recurrent iff $\mathcal{B}$ is finite $(a.s.)$

Remark 3.14 Formula $E(\mathcal{B}) = \frac{1-P_0}{f(0)}$ is even more fundamental than $E(\mathcal{B}) = \frac{1-P_0}{\lambda P_0}$, since in some M/G/1 variants $f(0) \neq \lambda P_0$. For example $f(0) = \lambda P_0 B(K)$ in a workload-barrier M/G/1 queue with finite barrier $K > 0$, where a customer balks if its service time would cause the workload to overshoot the barrier (variant 2 of Sect. 3.16.3); in that case $E(\mathcal{B}) = \frac{1-P_0}{\lambda P_0 B(K)}$.

3.4.11 *Equation for $f(x)$ via Renewal Reward Theorem*

Consider $\{W(t)\}_{t\geq 0}$. Let $\{P_0, f(x)\}_{x>0}$ be the limiting pdf of wait in M/G/1. We have $f(x) = \lim_{t\to\infty} \mathcal{D}_t(x)/t$ by Theorem 1.1. We now apply the *renewal reward theorem* to derive the *right hand side* of Eq. (1.8), as a check on the upcrossing-rate interpretation in Theorem 1.1, and because the renewal reward theorem is useful for analyzing many complex models as well (see references following Eq. (3.79)). Let $\mathcal{C}$:= an M/G/1 *busy cycle*, and $A_{\mathcal{C}}$:= *number of arrivals during C* (same as number of SP jumps of the embedded busy period $\mathcal{B}$). Denote the customers served in $\mathcal{B}$ as $\{\mathrm{C}_i\}_{i=1,\ldots,A_{\mathcal{C}}}$. Let

$$\mathcal{U}_i(x) = \begin{cases} 1 \text{ if customer-}i\text{'s service jump upcrosses level } x, \; i = 1, \ldots, A_{\mathcal{C}}, \\ 0 \text{ otherwise.} \end{cases}$$

Assume we do not know the order of arrival of the C_i's. Conditioning on the starting levels of the SP jumps, we have

$$P(\mathcal{U}_i(x) = 1) = P(S > x | W_i = 0) P(W_i = 0) \\ + \int_{y=0}^{x} P(S > x - y | W_i = y) d\dot{y}, i = 1, \ldots, A_{\mathcal{C}}.$$

where the events $\{W_i = 0\}$ and $\{W_i = y\}_{y>0}$ are mutually exclusive and exhaustive. Thus

$$P(\mathcal{U}_i(x) = 1) = \overline{B}(x)P_0 + \int_{y=0}^{x} \overline{B}(x-y)f(y)dy, i = 1, \ldots, A_C;$$

$$E(\mathcal{U}_i(x) = \overline{B}(x)P_0 + \int_{y=0}^{x} \overline{B}(x-y)f(y)dy, i = 1, \ldots, A_C.$$

Since $E(A_\mathcal{C}) = 1/P_0$ (see 3.4.14 below), The number of upcrossings of x during $A_\mathcal{C}$ is,

$$\begin{aligned}
\mathcal{U}_\mathcal{C}(x) &= \sum_{i=1}^{A_\mathcal{C}} \mathcal{U}_i(x), x > 0, \\
E(\mathcal{U}_\mathcal{C}(x)) &= E(A_\mathcal{C})E(\mathcal{U}_i(x) \\
&= \frac{1}{P_0}\left(\overline{B}(x)P_0 + \int_{y=0}^{x} \overline{B}(x-y)f(y)dy\right) \\
&= \overline{B}(x) + \frac{1}{P_0}\int_{y=0}^{x} \overline{B}(x-y)f(y)dy.
\end{aligned}$$

Finally the renewal reward theorem implies

$$\begin{aligned}
\lim \frac{\mathcal{U}_t(x)}{t} = \frac{E(\mathcal{U}_\mathcal{C}(x))}{E(\mathcal{C})} &= \frac{\overline{B}(x) + \frac{1}{P_0}\int_{y=0}^{x} \overline{B}(x-y)f(y)dy}{1/(\lambda P_0)} \\
&= \lambda P_0 \overline{B}(x) + \lambda \int_{y=0}^{x} \overline{B}(x-y)f(y)dy;
\end{aligned}$$

rate balance across level x, viz., $\lim_{t\to\infty} \mathcal{D}_t(x)/t = \lim \mathcal{U}_t(x)/t$, yields Eq. (1.8).

3.4.12 Busy Period Structure in Standard M/G/1

The M/G/1 busy period $\mathcal{B}$ can be partitioned into a set of sub-busy periods, different from a classical partition (see pp. 206–211 and p. 220ff in [104]; also [140]). Direct observation of a sample path of $\{W(t)\}_{t\geq 0}$ in Fig. 3.6, leads to a partition of $\mathcal{B}$ which preserves the scale with respect to the time axis '$t \rightarrow$' and the ordinates $W(t)$ throughout $\mathcal{B}$. Suppose a customer arrives at t_A^- and $W(t_A^-) = y \geq 0$; the SP then has coordinates (t_A^-, y). The SP immediately jumps an amount $\underset{dis}{=} S$, ending at $(t_A, y + S)$. Let

$$t_y = \min\{t > t_A | W(t) = y\}.$$

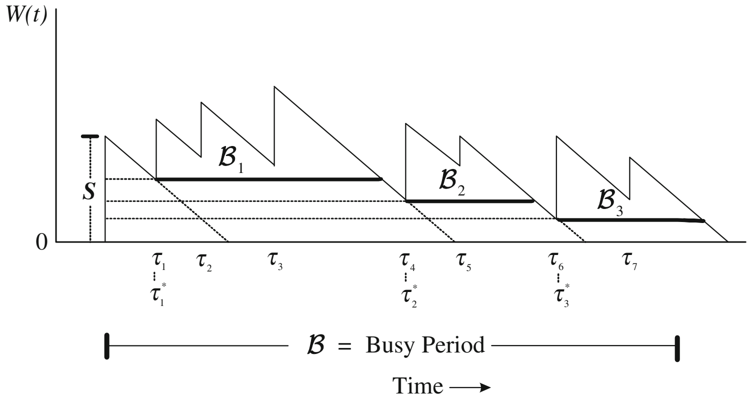


Fig. 3.6 Busy period $\mathcal{B} \underset{dis}{=} S + \sum_{i=1}^{N_S} \mathcal{B}_i$. $\mathcal{B}_i \underset{dis}{=} \mathcal{B}$, $i = 1, \ldots, N_S$. N_S = number of "tagged" (pseudo) arrivals in $\mathcal{B}$. Here $N_S = 3$. $N_S \underset{dis}{=}$ number of arrivals during S. Tagged arrival times are $\tau_1^* = \tau_1$, $\tau_2^* = \tau_4$, $\tau_3^* = \tau_6$. Tagged arrivals 1, 4, 6 during $\mathcal{B}$ initiate sub-busy periods $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$

The interval $\left(t_A, t_y\right)$ having length $t_y - t_A$ is a busy period $\mathcal{B}$ if $y = 0$; $\left(t_A, t_y\right)$ is a sub-busy period $\underset{dis}{=} \mathcal{B}$ if $y > 0$. The time interval $t_y - t_A$ is independent of y, since the initial SP jump at t_A is $\underset{dis}{=} S$. We utilize this partition of $\mathcal{B}$ to study its structure. (The foregoing definition of busy period is equivalent to the usual definition made for $y = 0$ only, e.g., [140]; see also p.10 and p. 102 in [84].)

Consider $\mathcal{B}$ within which $n \geq 1$ customers arrive. Denote their arrival times within $\mathcal{B}$ by $\tau_1 < \tau_2 < \cdots < \tau_n$, implying that τ_1 occurs within the initial service time S. Then $W(\tau_i^-) > 0, i = 1, 2, \ldots$. Define $\tau_1^* = \tau_1$ and $\tau_j^* = \min\{t > \tau_{j-1}^* | 0 < W(t) < W(\tau_{j-1}^*)\}$, $j = 2, \ldots, n$. Due to the memoryless property of the inter-arrival times and since $\frac{d}{dt}W(t) = -1$ $(W(t) > 0)$, the ordinates $\{W(\tau_j^{*-})\}_{j=1,\ldots,n}$ are distributed the same as n customer arrival times *during the first service time* S of $\mathcal{B}$. We call the customers that arrive at $\{\tau_j^*\}_{j=1,\ldots,n}$ "*tagged*" or "*pseudo*" *arrivals with respect to the initial* S *of* $\mathcal{B}$ (see Fig. 3.6).

Let N_S denote the number of *tagged arrivals* **during** $\mathcal{B}$. Then N_S is distributed as *the number of arrivals to the system during the* **service time** S. Tagged arrivals initiate their own sub-busy periods starting at $\left\{(\tau_n^{-*}, W(\tau_n^{-*}))\right\}_{n=1,\ldots,N_S}$ similar to $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$ depicted in Fig. 3.6 (where $\tau_1^* = \tau_1$, $\tau_2^* = \tau_4$, $\tau_3^* = \tau_6$). The tagged arrivals during $\mathcal{B}$ are customers 1, 4 and 6, which initiate $\mathcal{B}_1$, $\mathcal{B}_2$, $\mathcal{B}_3$, respectively. Note that $(\tau_n^{-*}, W(\tau_n^{-*}))_{n=1,\ldots,N_S}$ are strict descending ladder points [74] within $\mathcal{B}$. Then

$$\mathcal{B} = S + \sum_{i=1}^{N_S} \mathcal{B}_i, \tag{3.83}$$

where $\{\mathcal{B}_i\}_{i=1,2,\ldots}$ are i.i.d. sub-busy periods $\underset{dis}{=} \mathcal{B}$, and independent of N_S. Equation (3.83) is known, and is usually derived by different, but equivalent, reasoning (see Example 5.27, pp. 347–349 in Ross [125]). From (3.83), we obtain

$$E(\mathcal{B}) = E(S) + E(N_S)E(\mathcal{B}) = E(S) + \lambda E(S)E(\mathcal{B}) \tag{3.84}$$

which gives $E(\mathcal{B})$ as in (3.80).

Also, we can obtain (3.80) by recursively substituting for $\mathcal{B}_i$ in (3.83). This gives an infinite series of terms

$$\mathcal{B} \underset{dis}{=} S + \sum_{i=1}^{N_S} S_i + \sum_{i=1}^{N_S}\sum_{j=1}^{N_S} S_{i,j} + \sum_{i=1}^{N_S}\sum_{j=1}^{N_S}\sum_{k=1}^{N_S} S_{i,j,k} + \cdots$$

where $S_i, S_{i,j}, S_{i,j,k}, \ldots$, are $\underset{dis}{=} S$. Assuming $0 < \lambda E(S) < 1, \{P_0, f(x)\}_{x>0}$ exists and $\mathcal{B} < \infty$ $(a.s.)$. Then

$$\begin{aligned} E(\mathcal{B}) &= E(S) + \geq(E(S))^2 + \geq^2(E(S))^3 + \cdots \\ &= E(S) \cdot (1+\geq E(S)+(\geq E(S))^2 + \cdots) \\ &= \frac{E(S)}{1 - \lambda E(S)}. \end{aligned}$$

If $\lambda E(S) \geq 1$ it is possible for the busy period to be infinite. Then its mean and variance do not exist.

We compute the known formula for the variance of $\mathcal{B}$ assuming it exists from (3.83) and the definition $Var(\mathcal{B}) = E(\mathcal{B}^2) - (E(\mathcal{B}))^2$,for completeness; we intend to use the result for $E(\mathcal{B}^2)$ when discussing M/G/1 priority queues in Sect. 3.14 (see p. 349 in [125]).

To compute $E(\mathcal{B}^2)$, we first obtain a formula for $\mathcal{B}^2$ from (3.83) as

$$\mathcal{B}^2 = S^2 + 2S\sum_{i=1}^{N_S} \mathcal{B}_i + \left(\sum_{i=1}^{N_S} \mathcal{B}_i\right)^2.$$

Conditioning on $S = s$, gives the conditional expected value

$$E(\mathcal{B}^2|S = s) = s^2 + 2sE\left(\sum_{i=1}^{N_s} \mathcal{B}_i\right) + E\left(\left(\sum_{i=1}^{N_s} \mathcal{B}_i\right)^2\right).$$

In the second term on the right $\sum_{i=1}^{N_s} \mathcal{B}_i$ is a compound Poisson process with rate λ (see p. 346 in [125]). Thus

$$E\left(\sum_{i=1}^{N_s} \mathcal{B}_i\right) = \lambda s E(\mathcal{B}).$$

The third term on the right is

$$\begin{aligned} E\left(\left(\sum_{i=1}^{N_s} \mathcal{B}_i\right)^2\right) &= E\left(\sum_{i=1}^{N_s} \mathcal{B}_i^2 + \sum_{i \neq j=1}^{N_s} \mathcal{B}_i \mathcal{B}_j\right) \\ &= \lambda s E(\mathcal{B}^2) + E(N_s(N_s - 1)\mathcal{B}_i \mathcal{B}_j) \\ &= \lambda s E(\mathcal{B}^2) + E(N_s(N_s - 1))(E(\mathcal{B}))^2 \\ &= \lambda s E(\mathcal{B}^2) + (\lambda s)^2 (E(\mathcal{B}))^2. \end{aligned}$$

since

$$E(N_s(N_s - 1)) = \sum_{n=2}^{\infty} \frac{n(n-1)e^{-\lambda s}(\lambda s)^n}{n!} = (\lambda s)^2.$$

Thus

$$E(\mathcal{B}^2|S = s) = s^2 + 2\lambda s^2 E(\mathcal{B}) + \lambda s E(\mathcal{B}^2) + (\lambda s)^2 (E(\mathcal{B}))^2.$$

Unconditioning with respect to the service time distribution, substituting from (3.80) and simplifying yields

$$E(\mathcal{B}^2) = \frac{E(S^2)(1 + \lambda E(\mathcal{B}))^2}{1 - \lambda E(S)} = \frac{E(S^2)}{(1 - \lambda E(S))^3} = \frac{E(S^2)}{(1 - \rho)^3}, \tag{3.85}$$

where $\rho := \lambda E(S)$.

Since $Var(\mathcal{B}) = E(\mathcal{B}^2) - (E(B))^2$, from (3.80) and (3.85)

$$Var(\mathcal{B}) = \frac{Var(S) + \lambda(E(S))^3}{(1 - \lambda E(S))^3} = \frac{Var(S) + \lambda \rho^3}{P_0^3}. \tag{3.86}$$

3.4.13 Probability Distribution of the Busy Period

Starting from formula (3.83) above, we can proceed as on pp. 211–226 in [104] to derive $F_{\mathcal{B}}(y)$, $y > 0 :=$ the cdf of $\mathcal{B}$. Formula (5.169) on p. 226 in [104] gives an explicit expression for $F_{\mathcal{B}}(y)$, $y > 0$ as

$$F_{\mathcal{B}}(y) = \int_{s=0}^{y} \sum_{n=1}^{\infty} e^{-\lambda s} \frac{(\lambda s)^n}{n!} b_{(n)}(s) ds, \; y > 0, \tag{3.87}$$

where $b_{(n)}(s) :=$ the n-fold self convolution of $b(s)$. The paragraph following (5.169) therein observes that the "study of the busy period has really been the study of a transient phenomenon", which makes it more complicated than the analysis of a phenomenon in steady state.

3.4.14 Expected Number Served in Busy Period

Let $N_{\mathcal{B}} :=$ the number of customers served in a busy period. Let $A_C :=$ number of arrivals in a busy cycle. Then $N_{\mathcal{B}} = A_C$. Let $A(t)$ denote the number of arrivals to the system during time interval $(0, t)$. We get $E(N_{\mathcal{B}})$ by applying the renewal reward theorem; thus

$$\begin{aligned} \frac{E(N_{\mathcal{B}})}{E(\mathcal{C})} &= \frac{E(A_{\mathcal{C}})}{E(\mathcal{C})} = \lim_{t \to \infty} \frac{A(t)}{t} = \lambda, \\ E(N_{\mathcal{B}}) &= \lambda E(\mathcal{C}) = \lambda \frac{1}{\lambda P_0} = \frac{1}{P_0}. \end{aligned} \tag{3.88}$$

(See Exercise 17, p. 233 in [64].)

Another View for $E(N_{\mathcal{B}})$ using $N_{\mathcal{B}}$ as a Stopping Time

Let S_i, T_i denote the ith service and inter-arrival times during $\mathcal{B}$, respectively, $i = 1, 2, \ldots$. Then $N_{\mathcal{B}} = \min\{n | \sum_{i=1}^{n}(S_i - T_i) \leq 0\}$ is a *stopping time* for the sequence $\{(S_i - T_i)\}_{n=1,2,\ldots}$ (see, e.g., Exercise 13, p. 486, and pp. 678–679 in Ross [125]). Since $T_i \underset{dis}{\equiv} \text{Exp}_{\lambda}$, the excess inter-arrival time at the end of $\mathcal{B}$ is also distributed as Exp_{λ} due to the memoryless property. Hence $\sum_{i=1}^{N_{\mathcal{B}}}(S_i - T_i)$ ends a distance *below* 0, which is $\underset{dis}{=} \text{Exp}_{\lambda}$, implying

$$E\left(\sum_{i=1}^{N_{\mathcal{B}}}(S_i - T_i)\right) = -\frac{1}{\lambda}.$$

Applying Wald's equation (aka Wald's identity; see, e.g., p. 47ff in [122]) gives

$$E(N_{\mathcal{B}})\left(E(S)-\frac{1}{\lambda}\right)=-\frac{1}{\lambda}, \tag{3.89}$$

$$E(N_{\mathcal{B}})=\frac{1}{1-\lambda E(S)}=\frac{1}{P_0}. \tag{3.90}$$

We may also write $N_{\mathcal{B}}=\min\{n|\sum_{i=1}^{n}S_i\leq\sum_{i=1}^{n}T_i\}$. In this form it is seen that $N_{\mathcal{B}}$ is a stopping time for both sequences $\{S_i\}_{i=1,2,\ldots}$ and $\{T_i\}_{i=1,2,\ldots}$. That is, we observe the r.v.s in the order $S_1, T_1, S_2, T_2, \ldots$ and stop at n in both sequences when the stopping criterion ($\sum_{i=1}^{n}S_i\leq\sum_{i=1}^{n}T_i$) is first satisfied. Thus the event $\{N_{\mathcal{B}}=n\}$ is independent of $S_{n+1}, T_{n+1}, \ldots$. Moreover, since $\mathcal{B}=\sum_{i=1}^{N_{\mathcal{B}}}S_i$ where $S_i\underset{dis}{\equiv}S$, from (3.80) we have

$$E(\mathcal{B})=E(N_{\mathcal{B}})E(S)=\frac{E(S)}{1-\lambda E(S)},$$

which yields (3.90). (Interestingly, $E(S)\times E(N_{\mathcal{B}})$ is an intuitive way of thinking about $E(\mathcal{B})$.)

Note that $\mathcal{C}=\sum_{i=1}^{N_{\mathcal{B}}}T_i$ (one interarrival time precedes each arrival in a *busy cycle*). From $E(\mathcal{C})=1/(\lambda P_0)$ we have

$$E(\mathcal{C})=\frac{1}{\lambda P_0}=E(N_{\mathcal{B}})E(T)=(E(N_{\mathcal{B}}))\frac{1}{\lambda}, \tag{3.91}$$

which also gives (3.90).

We may write

$$N_{\mathcal{B}}=1+\sum_{i=1}^{N_S}N_{\mathcal{B}_i}$$

where $N_{\mathcal{B}_i}\underset{dis}{\equiv}N_{\mathcal{B}}$, and $N_S\equiv$ number of arrivals in the first service time of $\mathcal{B}$ (see Fig. 3.6; one sub-busy period for each arrival during the first service time). Then

$$E(N_{\mathcal{B}})=1+E(N_S)E(N_{\mathcal{B}})=1+\lambda E(S)E(N_{\mathcal{B}}),$$

again leading to (3.90).

In (3.90) if $P_0\lessapprox 1$ (close to 1) corresponding to a very low traffic intensity ρ, then $E(N_{\mathcal{B}})\gtrapprox 1$ (close to 1) meaning most customers in service are alone in the system.

The role of LC in this section, is that the downcrossing rate of level 0 (SP hit rate of 0 from above) is $f(0)$, which implies $E(\mathcal{C}) = \frac{1}{f(0)} = \frac{1}{\lambda P_0}$. Also, applying the stopping-time definition of a busy cycle just preceding (3.89), leads to (3.90).

3.4.15 Inter-Downcrossing Time of a State-Space Level

Consider a sample path of $\{W(t)\}_{t\geq 0}$ (Fig. 3.7). Let d_x denote the time between two successive downcrossings of level $x \geq 0$. Starting at the instant of the first downcrossing of state-space level x, d_x is an interval of a renewal process $\{\mathcal{D}_t(x)\}_{t\geq 0}$ due to exponential inter-arrival times. The renewal rate is $\lim_{t\to\infty} \frac{\mathcal{D}_t(x)}{t} = \lim_{t\to\infty} \frac{E(\mathcal{D}_t(x))}{t} = f(x)$ (see Corollary 3.2 in Sect. 3.2.5 above; Theorem 7.1, p. 432 in Ross [125]). Thus,

$$E(d_x) = \frac{1}{f(x)}, x \geq 0 \tag{3.92}$$

where $f(x)$ is the solution of (3.34) and (3.36).

Since $d_0 = \mathcal{C} :=$ busy cycle, $d_0 = \mathcal{B} + \mathcal{I}$ ($\mathcal{B} :=$ busy period; $\mathcal{I} :=$ idle period). Letting $x \downarrow 0$ in (3.92) gives

$$E(d_0) = \frac{1}{f(0)} = E(\mathcal{B}) + E(\mathcal{I}).$$

Thus, using method {3} in Sect. 3.4.10 we get $E(\mathcal{B})$ in (3.80).

3.4.16 Sojourn Below a Level of $\{W(t)\}_{t\geq 0}$

Let b_x denote a sojourn time below, or at, level $x \geq 0$ (Fig. 3.7). Assuming the queue is stable ($\rho < 1$), the proportion of time a sample path spends at or below x, is $\lim_{t\to\infty} E(\mathcal{D}_t(x))/t \cdot E(b_x) = f(x)E(b_x)$, *and is also equal to* the limiting cdf $F(x)$. Hence

$$E(b_x) = \frac{F(x)}{f(x)} \tag{3.93}$$

(see Remark 3.15 below). Letting $x \downarrow 0$, reduces (3.93) to the expected idle period

$$E(b_0) = \frac{F(0)}{f(0)} = \frac{P_0}{\lambda P_0} = \frac{1}{\lambda}.$$

Also, from (3.93)

$$\frac{d}{dx} \ln F(x) = \frac{1}{E(b_x)},$$

which leads to expressions for the cdf $F(x)$ and pdf $f(x)$ $(= F'(x))$ of wait in terms of $E(b_y)$, $0 < y < x$,

$$F(x) = P_0 e^{\int_{y=0}^{x} \frac{dy}{E(b_y)}}, x \geq 0, \tag{3.94}$$

$$f(x) = \frac{P_0}{E(b_x)} e^{\int_{y=0}^{x} \frac{dy}{E(b_y)}}, x > 0\,. \tag{3.95}$$

3.4.17 Sojourn Above a Level of $\{W(t)\}_{t \geq 0}$

Let a_x denote a sojourn time above level $x \geq 0$ (Fig. 3.7). Then $a_0 = \mathcal{B}$. By Theorem 1.1 in Sect. 1.6 the down- and upcrossing rates of level x are both equal to $f(x), x \geq 0$. The proportion of time that a sample path spends above x is $\lim_{t\to\infty} (\mathcal{U}_t(x) \cdot E(a_x))/t = \lim_{t\to\infty} (\mathcal{U}_t(x)/t) \cdot E(a_x) = f(x)E(a_x)$, *and is also equal to* $1 - F(x)$. Therefore

$$E(a_x) = \frac{1 - F(x)}{f(x)}, x \geq 0. \tag{3.96}$$

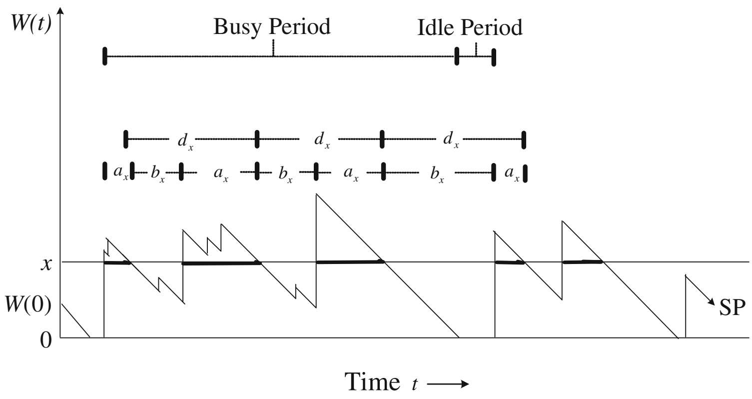


Fig. 3.7 Sample path of $\{W(t)\}_{t\geq 0}$ in M/G/1. Shows inter-downcrossing time d_x, sojourns a_x and b_x, busy and idle periods

Letting $x \downarrow 0$, in (3.96) gives $E(a_0) = (1 - P_0) / (\lambda P_0) = E(\mathcal{B})$.

Remark 3.15 Formula (3.96) can also be proved using the renewal reward theorem (Sect. 3.4.9), since $\{\mathcal{D}_t(x)\}_{t \geq 0}$ is a renewal process (starting from the first downcrossing of level x) since interarrival times are $\underset{dis}{=}$ Exp_λ having the memoryless property. Thus

$$\frac{E(a_x)}{E(d_x)} = \lim_{t \to \infty} \frac{\text{time spent above } x \text{during}(0, t)}{t} \underset{a.s.}{=} 1 - F(x),$$
$$E(a_x) = E(d_x) \cdot (1 - F(x)) = \frac{1 - F(x)}{f(x)}.$$

We can derive formula (3.93) for $E(b_x)$ similarly.

Proposition 3.3 below shows that *if* $E(a_x) \equiv E(\mathcal{B})$, $x \geq 0$, *then* the absolutely continuous part of $\{P_0, f(x)\}_{x>0}$ has an exponential form. Assume $\rho := \lambda E(S) < 1$.

Proposition 3.3 If $E(a_x) = E(\mathcal{B})$ for all $x \geq 0$, then the steady-state cdf of wait is $F(x) = 1 - \rho e^{-\frac{x}{E(\mathcal{B})}}$ and $\{P_0, f(x)\}_{x>0}$ is given by

$$P_0 = 1 - \rho, \quad f(x) = \lambda P_0 e^{-\frac{x}{E(\mathcal{B})}}, x > 0.$$

Proof If $E(a_x) \equiv E(\mathcal{B})$, $x \geq 0$, then from (3.96)

$$\frac{f(x)}{1 - F(x)} \equiv \frac{1}{E(\mathcal{B})}, x > 0, \tag{3.97}$$
$$\frac{d}{dx} \ln(1 - F(x)) \equiv -\frac{1}{E(\mathcal{B})}, x > 0.$$

Formula (3.97) is the *hazard rate (failure rate)* of the pdf of wait at x. (See Sect. 3.4.18 below.) Integration with respect to x yields

$$1 - F(x) = Ae^{-\frac{x}{E(\mathcal{B})}}, x > 0,$$

where A is a constant. Letting $x \downarrow 0$ gives

$$\begin{aligned} A &= 1 - F(0) = 1 - P_0 = \rho; \\ F(x) &= 1 - \rho e^{-\frac{x}{E(\mathcal{B})}}, x \geq 0. \end{aligned} \tag{3.98}$$

Differentiation of $F(x)$ in (3.98) with respect to $x > 0$ gives

$$f(x) = \frac{1}{E(\mathcal{B})}\rho e^{-\frac{x}{E(\mathcal{B})}} = \frac{1}{\frac{E(S)}{P_0}}\lambda E(S) \cdot e^{-\frac{x}{E(\mathcal{B})}} = \lambda P_0 e^{-\frac{x}{E(\mathcal{B})}}, x > 0. \tag{3.99}$$

■

Remark 3.16 The standard $M_\lambda/M_\mu/1$ queue satisfies the hypothesis of Proposition 3.3 because $S \underset{dis}{=} \text{Exp}_\mu$. All jumps that upcross level x have excess above $x \underset{dis}{=} \text{Exp}_\mu$ by the memoryless property, implying $a_x \underset{dis}{=} \mathcal{B}$, $x \geq 0$ (see Sect. 3.5.6).

3.4.18 Hazard Rate of PDF of Waiting Time

The term **hazard rate**, also called **failure rate**, is usually defined for positive continuous random variables in renewal theory, and failure time of components in reliability models (see, e.g., pp. 1–7 in [66]). In this monograph, we apply the 'hazard rate' to the pdf at x of waiting time (and other state variables, e.g., pdf at x of content of a dam in Sect. 6.2.12 in Chap. 6, etc.). In M/G/1 we may think of sojourn a_x as a 'lifetime' spent above level x. Thus, $\phi(x)E(a_x) = 1$, where $\phi(x) := f(x)/(1 - F(x))$, the hazard rate at x. Then $E(a_x)$ (E(lifetime above x)) varies inversely with $\phi(x)$. This idea fits the notion of failure rate in reliability models. Let X := *lifetime of a component* (also called *failure time*). The failure rate at lifetime x is the conditional pdf of lifetime given the lifetime exceeds x. Following pp. 1–4 in [66],

$$\begin{aligned}\phi(x) &= \lim_{\Delta x \downarrow 0} \frac{P(x < X \leq x + \Delta x | X > x)}{\Delta x} \\ &= \lim_{\Delta x \downarrow 0} \frac{P(x < X \leq x + \Delta x)}{\Delta x P(X > x)} \\ &= \lim_{\Delta x \downarrow 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} \frac{1}{(1 - F(x)} \\ &= \frac{f(x)}{(1 - F(x))}.\end{aligned}$$

For the pdf of waiting time, the dimension of $\phi(x)$ is the same as that of $f(x)$, viz., $1/[Time]$. In other stochastic models the dimension of $\phi(x)$ is the same as that of the pdf of the state variable.

3.4.19 Sojourn Above a Level and Distribution of Wait

Proposition 3.4 below relates $E(a_y)$, $y \in (0, x)$, to $F(x)$ and $f(x)$, $x > 0$. In general $E(a_y)$ varies with $y > 0$. (However, in M/M/1 $E(a_y) \equiv E(\mathcal{B})$, $y > 0$.)

Proposition 3.4 For the M/G/1 queue in equilibrium ($\rho < 1$),

$$F(x) = 1 - \rho \cdot e^{-\int_{y=0}^{x} \frac{1}{E(a_y)} dy}, \ x \geq 0. \tag{3.100}$$

$$f(x) = \frac{\rho}{E(a_x)} \cdot e^{-\int_{y=0}^{x} \frac{1}{E(a_y)} dy}, \ x > 0. \tag{3.101}$$

Proof Consider a sample path of $\{W(t)\}_{t \geq 0}$. The pdf $f(x)$ is the SP upcrossing (and downcrossing) rate of level x. Hence the long-run proportion of time $\{W(t)\}_{t \geq 0}$ spends above level x is

$$f(x)E(a_x) = 1 - F(x).$$

Thus

$$\frac{f(x)}{1 - F(x)} = \frac{1}{E(a_x)}, \ x > 0. \tag{3.102}$$

The term $f(x)/(1 - F(x))$ is the hazard rate of the waiting time at level x (see Sect. 3.4.18 above). From (3.102)

$$\frac{d}{dx} \ln(1 - F(x)) = -\frac{1}{E(a_x)}, \ x > 0.$$

Integrating with respect to x gives

$$1 - F(x) = Ae^{-\int_{y=0}^{x} \frac{1}{E(a_y)} dy},$$

where A is a constant. Letting $x \downarrow 0$, gives

$$A = 1 - F(0^+) = 1 - F(0) = 1 - P_0 = \rho.$$

Hence we obtain (3.100); (3.101) follows by taking $dF(x)/dx$ in (3.100). ■

Equivalence of Formulas for $F(x)$ in Terms of $E(b_x)$ and $E(a_x)$
We now check that the right sides of (3.100) and (3.94) are both equal to $F(x)$, $x > 0$, and therefore to each other. Thus

$$P_0 e^{\int_{y=0}^{x} \frac{f(y)dy}{F(y)}} = P_0 e^{\int_{y=0}^{x} d\ln F(y)} = P_0 e^{\ln\left(\frac{F(x)}{F(0)}\right)}$$
$$= P_0 \frac{F(x)}{F(0)} = F(x), x \geq 0,$$

and

$$1 - \rho e^{-\int_{y=0}^{x} \frac{f(y)dy}{1-F(y)}} = 1 - \rho e^{\int_{y=0}^{x} d\ln(1-F(y))}$$
$$= 1 - \rho e^{\ln\left(\frac{1-F(x)}{1-F(0)}\right)} = 1 - \rho\left(\frac{1-F(x)}{1-F(0)}\right)$$
$$= 1 - \rho\left(\frac{1-F(x)}{\rho}\right) = F(x), x \geq 0,$$

proving the equivalence of the two formulas.

3.4.20 Computing $F(x)$ via $E(a_x)$

Suppose we do not have an explicit formula for $F(x)$ in a particular M/G/1 model. We can compute $E(a_x)$ (reciprocal of hazard rate) either analytically or using simulation, and apply formula (3.100) to obtain an analytical formula for $F(x)$, or an estimate of $F(x)$.

Analytical We can get analytic expressions for $E(a_x)$ in some models. We know $E(a_x) = \mathcal{B}$ in M/M/1. In general, however, $E(a_x)$ may be difficult to compute analytically. Example 3.7 below computes $E(a_x)$ analytically in an $M_\lambda/Erl_{2,\mu}/1$ queue. ($Erl_{k,\mu}$:= Erlang random variable; see Gamma distribution in Table 2.2, p. 66 in [125].)

Example 3.7 In $M/Erl_{2,\mu}/1$ with arrival rate λ, $E(S) = 2/\mu$ and $\rho = \lambda \cdot \frac{2}{\mu} < 1$, consider a sample path of $\{W(t)\}_{t\geq 0}$ (see also Example 3.2 in Sect. 3.3). $S \underset{dis}{=} Erl_{2,\mu}$ is the sum of two i.i.d. Exp_μ random variables; we call these phase 1 and phase 2 respectively. Either $a_x \underset{dis}{=} \mathcal{B}$ for the **standard** $M_\lambda/Erl_{2.\mu}/1$ queue, or $a_x \underset{dis}{=} \mathcal{B}$ for the $M_\lambda/Erl_{2,\mu}/1$ queue **where zero-wait customers have** $S \underset{dis}{=} Exp_\mu$ (i.e., special (exceptional) service for 'zero-wait' arrivals), depending on the initial service-time phase that covers x. That is, a_x's initial SP upcrossing of x covers x either during phase 1 or during phase 2 of the $Erl_{2,\mu}$ service time. If phase 1 covers x, then the excess jump above $x \underset{dis}{=} Erl_{2,\mu}$, due to the memoryless property of Exp_μ. If phase 2 covers x, then the excess jump above $x \underset{dis}{=} Exp_\mu$. If phase 1 covers x, applying (3.82) we get

$E(\mathcal{B}) = \frac{2}{\mu-2\lambda}$. If phase 2 covers x then the initial $S \underset{dis}{=} Exp_{\mu}$; this results in an $M_{\lambda}/Erl_{2,\mu}/\,1$ with $E(a_x) = E(\mathcal{B}) = \frac{1}{\mu-2\lambda}$, because $a_x \underset{dis}{=} \mathcal{B}$ in which zero-wait customers receive "special" service Exp_{μ} different from the rest of the service times which are $Erl_{2,\mu}$s (see Sect. 3.6.1 below). Thus,

$$E(a_x) = p_1(x)\left(\frac{2}{\mu - 2\lambda}\right) + p_2(x)\left(\frac{1}{\mu - 2\lambda}\right),$$

where $p_i(x) = P$ (phase i of an SP jump covers x|SP upcrosses x), $i = 1, 2$. From (3.100)

$$F(x) = 1 - \rho \exp\left(-\int_{y=0}^{x} \frac{1}{p_1(y)\left(\frac{2}{\mu-2\lambda}\right) + p_2(y)\left(\frac{1}{\mu-2\lambda}\right)} dy\right). \tag{3.103}$$

In Example (3.2), Eq. (3.41) for M_{λ} /$Erl_{2,\mu}$/1 yields

$$\begin{aligned} p_1(x) &= \frac{\lambda\left(P_0 e^{-\mu x} + \int_{y=0}^{x} e^{-\mu(x-y)} f(y)dy\right)}{f(x)} \\ &= \frac{P_0 e^{-\mu x} + \int_{y=0}^{x} e^{-\mu(x-y)} f(y)dy}{P_0 e^{-\mu x}(1+\mu x) + \int_{y=0}^{x} e^{-\mu(x-y)}(1+\mu(x-y)) f(y)dy, x>0}, \\ p_2(x) &= \frac{P_0 e^{-\mu x}\mu x + \int_{y=0}^{x} e^{-\mu(x-y)}\mu(x-y) f(y)dy}{P_0 e^{-\mu x}(1+\mu x) + \int_{y=0}^{x} e^{-\mu(x-y)}(1+\mu(x-y)) f(y)dy, x>0}, \ x > 0. \end{aligned} \tag{3.104}$$

where $\{P_0, f(y)\}_{y\geq 0}$ is specified in (3.42).

Example 3.8 In Example 3.7, $S \underset{dis}{=} Erl_{2,\mu}$ and $E(B) = E(S)/P_0 = \frac{2/\mu}{1-\lambda(2/\mu)} = \frac{2}{\mu-2\lambda}$. Then $E(a_x) = p_1(x)\left(\frac{2}{\mu-2\lambda}\right) + p_2(x)\left(\frac{1}{\mu-2\lambda}\right)$ where $p_1(x) + p_2(x) = 1$ and $p_1(x) > 0$, $p_2(x) > 0$. Thus $E(a_x) < \frac{2}{\mu-2\lambda} = E(B)$.

Alternatively, we could *estimate* $p_1(x)$, $p_2(x)$, $x > 0$, from a simulated sample path of $\{W(t)\}_{t\geq 0}$. Then substitute the estimated values into (3.103) to estimate $F(x)$, $x > 0$. This **hybrid technique** combines estimated values from simulation and analytical results.

Simulation to Estimate $E(a_y)$

To estimate $E(a_y)$, $y \in [0, x]$, simulate a single sample path of $\{W(t)\}$, $0 \leq t \leq T_{sim}$, where T_{sim} is "large". We utilize $\{W(t)\}_{0\leq t\leq T_{sim}}$ to *estimate* $E(a_{y_j})$ where y_j is a level of a state-space partition of $[0, x]$: $0 = y_0 < y_1 < \cdots < y_N$, and choose the subintervals to be "small", e.g., $y_{j+1} - y_j \equiv h > 0$, $j = 0, \ldots, N-1$ (depending on the required accuracy). Take $N = \lfloor x/h \rfloor$ where

$\lfloor \alpha \rfloor$ denotes the greatest integer $\leq \alpha$. Suppose in the simulated sample path there are M_j sojourns above level y_j during $[0, T_{sim}]$; let their observed *values* be $a_{y_j,1}, \ldots, a_{y_j,M_j}$. Assume T_{sim} is sufficiently large so that each M_j is "large" enough for the required accuracy. Then estimate $E(a_{y_j})$ using

$$\widehat{E}(a_{y_j}) = \frac{1}{M_j} \sum_{i=0}^{M_j} a_{y_j,i},\ j = 0, 1, \ldots, N.$$

We can estimate the value of $\int_{y=0}^{x} \frac{1}{E(a_y)} dy$ in (3.100) by

$$\int_{y=0}^{x} \frac{1}{E(a_y)} dy \underset{est}{=} h \sum_{j=0}^{N} \frac{1}{\widehat{E}(a_{y_j})}.$$

(We consider LC estimation in Chap. 9.)

Intuitive Meaning of the Hazard Rate

Denote the hazard rate of wait at x by $\phi(x)$. From (3.102), a plausible estimate of $\phi(x)$ is

$$\widehat{\phi}(x) = \frac{1}{\widehat{E}(a_x)}. \tag{3.105}$$

By definition

$$\phi(x)dx = P(W_q \in (x, x + dx) | W_q > x) = \frac{P(x < W_q < x + dx)}{P(W_q > x)},$$

where W_q is the teady-state queue wait (see, e.g., p. 299 in [125]). Formula (3.102) suggests an intuitive meaning based on $\phi(x) = 1/E(a_x)$, i.e., $\phi(x)$ is large iff $E(a_x)$ is small, and $\phi(x)$ is small iff $E(a_x)$ is large. This suggests studying connections between hazard rates of state random variables, and their sample-path expected sojourn times with respect to state-space levels in related stochastic models. (See Sect. 3.4.18 for pertinent comments.)

3.4.21 Events During an Inter-downcrossing Time

Consider $\{W(t)\}_{t \geq 0}$. We derive formulas for E(number of SP downcrossings of an arbitrary level $x \geq 0$) during d_y, $y \geq 0$, and E(number of customer arrivals) during d_y, $y \geq 0$; see Sect. 3.4.15 and Fig. 3.1. (See also Sect. 3.5 below regarding the M/M/1 queue.)

Consider a sample path of $\{W(t)\}_{t\geq 0}$ for an M/G/1 queue with $\rho < 1$. Denote the steady-state pdf of wait by $\{P_0, f(x)\}_{x\geq 0}$. Fix level $y \geq 0$. Let $\mathcal{D}_{d_y}(x)$ denote the number of SP downcrossings of an **arbitrary** level $x \geq 0$ during a sample-path inter-downcrossing d_y.

Proposition 3.5

$$E(\mathcal{D}_{d_y}(x)) = \frac{f(x)}{f(y)}. \tag{3.106}$$

Proof Since $\{\mathcal{D}_t(y)\}_{t\geq 0}$ is a renewal process starting at the first downcrossing of y,

$$\frac{E(\mathcal{D}_{d_y}(x))}{E(d_y)} = \lim_{t\to\infty} \frac{\mathcal{D}_t(x)}{t} = f(x),$$

by Theorem 1.1 and the renewal reward theorem. Also, $E(d_y) = 1/f(y)$ by the elementary renewal theorem. Equation (3.106) follows. ■

Observe that for arbitrary x, $E(\mathcal{D}_{d_y}(x))/E(d_y)$ is invariant for all $y \geq 0$. For example, if $y = 0$ then $d_0 = \mathcal{C}$ (busy cycle), and $E(\mathcal{D}_{d_0}(x))/E(d_0) = E(\mathcal{D}_{d_0}(x))/(1/(\lambda P_0)) = f(x)$.

If $y = 0$ and $x \downarrow 0$ then $E(\mathcal{D}_{d_0}(0)) = E(\mathcal{D}_{\mathcal{C}}(0)) = 1$. Thus

$$E(\mathcal{D}_{d_0}(0))/(1/(\lambda P_0)) = 1/(1/(\lambda P_0)) = \lambda P_0 = f(0),$$

which is compatible with Theorem 1.1.

Let A_{d_y} := number of customer arrivals during d_y. Let $A(t)$:= number of customer arrivals during $(0, t)$.

Proposition 3.6

$$E(A_{d_y}) = \frac{\lambda}{f(y)}, \; y \geq 0. \tag{3.107}$$

Proof

$$\frac{E(A_{d_y})}{E(d_y)} = \frac{E(A_{d_y})}{1/f(y)} = \lim_{t\to\infty} \frac{A(t)}{t} = \lambda,$$

by Theorem 1.1 and the renewal reward theorem, resulting in (3.107). ■

Letting $y \downarrow 0$ in 3.106 gives

$$E(A_{d_0}) = E(A_{\mathcal{C}}) = E(N_B) = \frac{\lambda}{f(0)} = \frac{\lambda}{\lambda P_0} = \frac{1}{P_0},$$

which is an additional proof of Eq. (3.88).

3.4.22 Boundedness of PDF in Steady State

Why is it potentially useful to know that in the limiting pdf of wait $\{P_0, f(x)\}_{x>0}$, $f(x)$ is bounded by a finite quantity? Suppose we want to estimate $f(x)$ in an analytically intractable M/G/1 model by means of simulation of a sample path. It would be helpful to know this fact when writing a computer program for the simulation.

In the standard M/G/1 queue let the arrival rate be λ, let S have cdf $B(y)$, $y > 0$, and $\rho < 1$. Assume $B(\cdot)$ is absolutely continuous.

Proposition 3.7

$$f(x) \le \lambda, x > 0. \tag{3.108}$$

Proof (1) In the integral equation for $\{P_0, f(x)\}_{x>0}$ (1.8), repeated here for convenience,

$$f(x) = \lambda P_0 \overline{B}(x) + \lambda \int_{y=0}^{x} \overline{B}(x-y) f(y) dy, x > 0.$$

$\overline{B}(x) < 1$, $x > 0$ (for any cdf $H(\cdot)$, $0 \le H(x) \le 1$, where $H(x)$ is right-continuous and monotone increasing). Thus

$$f(x) < \lambda P_0 + \lambda \int_{y=0}^{x} f(y) dy = \lambda \left(P_0 + \int_{y=0}^{x} f(y) dy \right) = \lambda F(x) \le \lambda, x > 0.$$

(2) On the right side of the alternative form of the LC integral equation (3.43) (repeated here)

$$f(x) = \lambda F(x) - \lambda \int_{y=0}^{x} B(x-y) f(y) dy, x > 0.$$

the subtracted term is > 0. Thus

$$f(x) < \lambda F(x) \le \lambda.$$

(3) Consider a sample path of $\{W(t)\}_{t\geq 0}$. Let $\mathcal{D}_t(x)$ and $A(t)$ denote the number of SP downcrossings of level x, and number of arrivals to the system during $(0,t)$, respectively. Examination of the sample path yields $\mathcal{D}_t(x) < A(t), x \geq 0, t > 0$, $(a.s.)$. Hence

$$f(x) = \lim_{t\to\infty} \frac{\mathcal{D}_t(x)}{t} \leq \lim_{t\to\infty} \frac{A(t)}{t} = \lambda,$$

since $\{A(t)\}$ is a Poisson process with rate λ. ■

Example 3.9 In $M_\lambda/M_\mu/1$, $f(x) = \lambda P_0 e^{-(\mu-\lambda)x}, x > 0$, $P_0 = 1-\rho > 0$ (Sect. 3.5.1). Both $0 < P_0 < 1$ and $0 < e^{-(\mu-\lambda)x} < 1$, $x > 0$. Therefore $f(x) < \lambda P_0, x > 0$.

Inequality (3.108) also holds in: the workload-bounded M/G/1 queue (Sect. 3.16); the M/D/1 queue (Sect. 3.12); and others.

3.5 M/M/1 Queue

We now derive some steady-state results for the standard M/M/1 queue with FCFS (first come first served) discipline. Some well-known results are included to develop facility with LC and reinforce intuitive background. Let λ := arrival rate, service time $S \underset{dis}{=} \text{Exp}_\mu$, $\overline{B}(x) = e^{-\mu x}, x \geq 0$, $B(x) = 1 - e^{-\mu x}, x\dot{\geq}0$, $\rho := \lambda E(S) = \lambda/\mu < 1$.

3.5.1 Waiting Time PDF and CDF

Consider a sample path of $\{W(t)\}_{t\geq 0}$ (e.g., Fig. 3.5). From the basic LC integral equation (3.34), or Fig. 1.6 in Sect. 1.7, we get

$$f(x) = \lambda P_0 e^{-\mu x} + \lambda \int_{y=0}^{x} e^{-\mu(x-y)} f(y)dy,\ x > 0, \tag{3.109}$$

where $\{P_0, f(x)\}_{x>0}$ is the steady-state pdf of wait.

Differentiating both sides of (3.109) with respect to x, yields the ordinary differential equation

$$f'(x) + (\mu - \lambda) f(x) = 0, x > 0, \tag{3.110}$$

with solution

$$f(x) = Ae^{-(\mu-\lambda)x}, x > 0; \tag{3.111}$$

the constant A is determined by letting $x \downarrow 0$ in both (3.109) and (3.111). Thus $A = f(0^+) = \lambda P_0$, giving

$$f(x) = \lambda P_0 e^{-(\mu-\lambda)x}, x > 0, \tag{3.112}$$

where, for the standard M/G/1 (see e.g., Eq. (3.62))

$$P_0 = 1 - \rho = 1 - \frac{\lambda}{\mu}. \tag{3.113}$$

We may also compute P_0 by substituting (3.112) into the normalizing condition,

$$P_0 + \int_{x=0}^{\infty} f(x)dx = 1, \tag{3.114}$$

which yields (3.113) directly.

From (3.112) the cdf of wait is

$$F(x) = P_0 + \int_{y=0}^{x} \lambda(1-\rho)e^{-(\mu-\lambda)y}dy = 1 - \rho e^{-(\mu-\lambda)x}, x > 0. \tag{3.115}$$

3.5.2 System Time PDF and CDF

Let σ denote the system time, $f_\sigma(x)$ its pdf, $F_\sigma(x)$ its cdf, $x > 0$ (see Sect. 3.3.2). Since $\sigma = W_q + S$, we obtain

$$\begin{aligned}
P(\sigma > x) &= P(S > x|W_{q=0})P_0 + \textstyle\int_{y=0}^{x} P(S > x-y|W_q = y)f(y)dy \\
&\qquad + P(W_q > x) \\
&= P_0 e^{-\mu x} + \lambda P_0 \textstyle\int_{y=0}^{x} e^{-(\mu-\lambda)y} e^{-\mu(x-y)}dy + \int_{y=x}^{\infty} \lambda P_0 e^{-(\mu-\lambda)y}dy \\
&= \frac{P_0}{1-\frac{\lambda}{\mu}} e^{-(\mu-\lambda)x} = e^{-(\mu-\lambda)x}, x > 0.
\end{aligned} \tag{3.116}$$

We can also obtain (3.116) using Eq. (3.61) (or equivalently Eq. (3.72)). Thus $\sigma \underset{dis}{=} \text{Exp}_{\mu-\lambda}$, i.e.,

$$\begin{aligned}
f_\sigma(x) &= (\mu-\lambda)\, e^{(\mu-\lambda)x}, x > 0 \\
F_\sigma(x) &= 1 - e^{(\mu-\lambda)x}, x \geq 0.
\end{aligned} \tag{3.117}$$

Additionally, we can obtain $f_\sigma(x)$ directly in terms of $f(x)$ using (3.74), thus getting (3.117) similarly as in Example 3.5 in Sect. 3.4.7.

3.5.3 Number in System Probability Distribution

Let N denote the number of units in the M/M/1 system at an arbitrary time point in the steady state. Let $P(N = n) = P_n, n = 0, 1, \ldots$ (see Sect. 3.4.8). We obtain the distribution of N by conditioning on W_q, or on σ.

Conditioning on W_q, substitute $f(x)$ in (3.112) into (3.76), getting

$$\begin{aligned} P_n &= \int_{y=0}^{\infty} e^{-\lambda y} \frac{(\lambda y)^{n-1}}{(n-1)!} \lambda P_0 e^{-(\mu-\lambda) y} dy \\ &= P_0 \left(\frac{\lambda}{\mu}\right)^n \int_{y=0}^{\infty} e^{-\mu y} \frac{(\mu y)^{n-1}}{(n-1)!} \mu dy = P_0 \rho^n, n = 0, 1, \ldots , \end{aligned}$$

since the integrand $e^{-\mu y}(\mu y)^{n-1}\mu/(n-1)!$ is the pdf of $\text{Erl}_{n,\mu}$ (see formula (3.38) in Example 3.2, Sect. 3.3).

The normalizing condition $\sum_{n=0}^{\infty} P_n = 1$ yields $P_0 \sum_{n=0}^{\infty} \rho^n = 1$, whence $P_0 = 1 - \rho$, giving the well-known geometric distribution

$$P_n = P_0 (1 - P_0)^n = (1 - \rho)\rho^n, n = 0, 1, \ldots . \tag{3.118}$$

Conditioning on σ, substitute $f_\sigma(x)$ from (3.117) into (3.77) in Sect. 3.4.8, getting

$$\begin{aligned} P_n &= \int_{y=0}^{\infty} e^{-\lambda y} \frac{(\lambda y)^n}{n!} (\mu - \lambda) e^{-(\mu-\lambda) y} dy \\ &= \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) \int_{y=0}^{\infty} e^{-\lambda y} \frac{(\mu y)^n}{n!} \mu dy \\ &= \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n = (1 - \rho)\rho^n, n = 0, 1, \ldots, \end{aligned}$$

because the integrand $e^{-\lambda y}(\mu y)^n \mu/n!$ is the pdf of $\text{Erl}_{n+1,\mu}$; P_n so derived is consistent with (3.118).

E(**number in system**) The right tail probability is $P(N \geq n) = \rho^n, n = 0, \ldots$. Thus

$$E(N) = \sum_{n=1}^{\infty} P(N \geq n) = \sum_{n=1}^{\infty} \rho^n = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}, \tag{3.119}$$

which agrees with the M/G/1 result $E(N) = L = \frac{\lambda^2 E(S^2)}{2P_0} + \rho$ in (3.65), since $E(S^2) = 2/\mu^2$ when $S \underset{dis}{=} Exp_{\mu}$.

Remark 3.17 A classical way to derive P_n, $n = 0, 1, \ldots$, in M/M/1 is via birth and death processes (e.g., pp. 49–55 in [84]; Sect. 6.3, p. 374 and Example 6.14, p. 395 in [125]; and others). Using the birth-death derived values of P_n, the pdf $\{P_0, f(x)\}_{x>0}$ of wait is then derived by conditioning on N. Here, we reason in the opposite direction: first derive the pdf $\{P_0, f(x)\}_{x>0}$ or $f_{\sigma}(x)$, $x > 0$, then condition on W_q or on σ to derive the values of P_n, $n = 1, 2, \ldots$. Similar remarks apply to other exponential models, like multiple server M/M/c queues (Chap. 4).

3.5.4 Expected Busy Period

The $M_{\lambda}/M_{\mu}/1$ queue is an M_{λ}/G/1 queue having $E(S) = \frac{1}{\mu}$. Substituting $\frac{1}{\mu}$ into (3.80) in Sect. 3.4.10 gives the well-known result

$$E(\mathcal{B}) = \frac{E(S)}{P_0} = \frac{1}{\mu\left(1 - \frac{\lambda}{\mu}\right)} = \frac{1}{\mu - \lambda}. \tag{3.120}$$

3.5.5 CDF and PDF of Busy Period in M/M/1

Applying formula (3.87) in Sect. 3.4.13 we obtain, since the n-fold convolution of $b(y)$ $(:= \mu e^{-\mu y})$ is $b_{(n)}(y) \underset{dis}{=} \mathrm{Erl}_{n,\mu}(y)$. The cdf of $\mathcal{B}$ is

$$\begin{aligned}
F_{\mathcal{B}}(x) &= \int_{y=0}^{x} \sum_{n=1}^{\infty} \frac{e^{-\lambda y}\,(\lambda y)^{n-1}}{n!} \frac{e^{-\mu y}(\mu y)^{n-1}\mu}{(n-1)!} dy \\
&= \int_{y=0}^{x} e^{-(\lambda+\mu)y} \frac{\mu}{\sqrt{\lambda\mu}y} \sum_{n=1}^{\infty} \frac{\left(\sqrt{\lambda\mu}y\right)^{2n-1}}{n!\,(n-1)!} dy \\
&= \int_{y=0}^{x} e^{-(\lambda+\mu)y} \sqrt{\frac{\mu}{\lambda}} \frac{1}{y} \sum_{n=1}^{\infty} \frac{\left(\frac{2\sqrt{\lambda\mu}y}{2}\right)^{2n-1}}{n!\,(n-1)!} dy, \; x > 0,
\end{aligned}$$

which yields $F_{\mathcal{B}}(x)$ and pdf $f_{\mathcal{B}}(x)$ of $\mathcal{B}$ as

$$F_{\mathcal{B}}(x) = \int_{y=0}^{x} \frac{\sqrt{\mu/\lambda}e^{-(\lambda+\mu)y} I_1(2\sqrt{\lambda\mu}y)}{y} dy, \tag{3.121}$$

$$f_{\mathcal{B}}(x) = \frac{\sqrt{\mu/\lambda}e^{-(\lambda+\mu)x} I_1(2\sqrt{\lambda\mu}x)}{x}, x > 0, \tag{3.122}$$

where $I_1(z) :=$ *modified Bessel function of the first kind of order* 1 given by

$$I_1(z) = \sum_{n=1}^{\infty} \frac{(z/2)^{2n-1)}}{(n-1)!n!}$$

(see, e.g., pp. 101–102 in Gross et al. [84]).

3.5.6 Geometric Derivation of CDF and PDF of Wait

Consider a sample path of $\{W(t)\}_{t\geq 0}$ in M/M/1. Let $\mathcal{B}$ denote a busy period. Given that the SP upcrosses level x, the sojourn above x is $a_x \underset{dis}{=} \dot{\mathcal{B}}$, *independent* of $x \geq 0$, due to the memoryless property of Exp_μ (Fig. 3.8). (See Proposition 3.4 in Sect. 3.4.19; also paragraph following "Key Question" in Sect. 1.5.2.)

Substituting $E(\mathcal{B})$ for $E(a_x)$ in formulas (3.100) and (3.101), and applying (3.120) yields

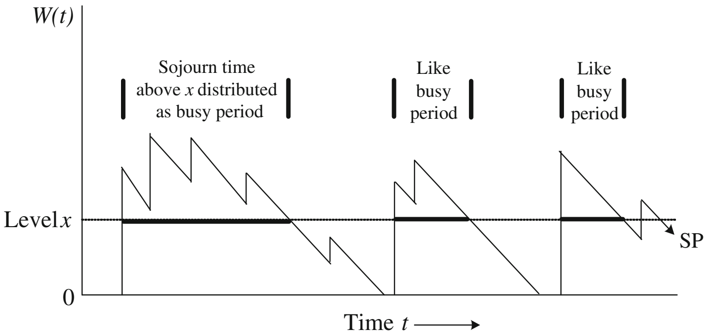


Fig. 3.8 Sample path of $\{W(t)\}_{t\geq 0}$ in $\text{M}_\lambda/\text{M}_\mu/1$ queue showing $a_x \underset{dis}{=} \mathcal{B}$. SP excess jumps above x are $\underset{dis}{=} \text{Exp}_\mu$

$$F(x) = 1 - \rho e^{-\frac{x}{E(\mathcal{B})}} = 1 - \rho e^{-(\mu-\lambda)x}, x \geq 0, \tag{3.123}$$
$$f(x) = \lambda(1-\rho)e^{-(\mu-\lambda)x} = \lambda P_0 e^{-(\mu-\lambda)x}, x > 0. \tag{3.124}$$

The M/M/1 model satisfies Proposition 3.3 in Sect. 3.4.17.

3.5.7 Inter-crossing Time of Level x

We now consider d_x, b_x, a_x, defined in Sects. 3.4.15, 3.4.16 and 3.4.17, respectively. We look at the time between SP successive upcrossings (inter-*upcrossing* time), and E(number of SP crossings of a level) during a busy cycle or during sojourns above or below an arbitrary level.

Inter-downcrossing Time of Level x

We have

$$d_x = b_x + a_x, \; E(d_x) = E(b_x) + E(a_x).$$

In M/M/1 the inter-arrival and service times are $\underset{dis}{=}$ Exp_λ and Exp_μ, respectively. For fixed $x \geq 0$, successive triplets $\{d_{x,n}, b_{x,n}, a_{x,n}\}_{n=1,2,...}$ form a sequence of i.i.d. random variables ($d_{x,n} \underset{dis}{=} d_x$, $b_{x,n} \underset{dis}{=} b_x$, $a_{x,n} \underset{dis}{=} a_x$). Thus $\left\{d_{x,n}\right\}_{n=1,2,...}$ forms a renewal process and $\{b_{x_n}, a_{x,n}\}_{n-1,2,...}$ forms an alternating renewal process. As in Sects. 3.4.15, 3.4.16 and 3.4.17,

$$E(d_x) = \frac{1}{f(x)}, \; E(b_x) = \frac{F(x)}{f(x)}, \; E(a_x) = \frac{1-F(x)}{f(x)}. \tag{3.125}$$

Since $a_x \underset{dis}{\equiv} \mathcal{B}$

$$E(a_x) = \frac{1}{\mu - \lambda}, x \geq 0, \tag{3.126}$$
$$E(d_x) = \frac{F(x)}{f(x)} + \frac{1}{\mu - \lambda}, x \geq 0. \tag{3.127}$$

Letting $x = 0$ in (3.127) gives $E(d_0) = E(\mathcal{C})$ where $\mathcal{C}$:= busy cycle = d_0. Thus

$$E(\mathcal{C}) = \frac{F(0)}{f(0)} + \frac{1}{\mu-\lambda} = \frac{P_0}{\lambda P_0} + \frac{1}{\mu - \lambda} = \frac{1}{\lambda(1-\rho)} = \frac{1}{\lambda P_0}, \tag{3.128}$$

which agrees with formula (3.81). We obtain $E(d_x)$ by substituting $f(x)$ from (3.124) into (3.127). Thus

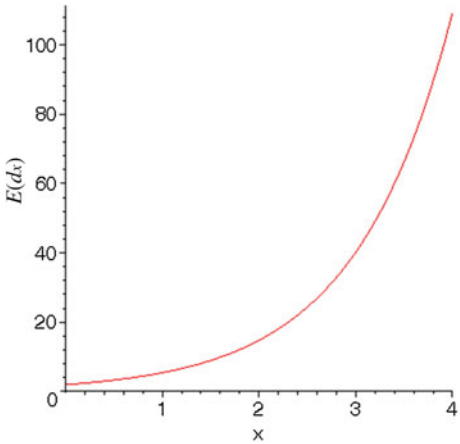


Fig. 3.9 Expected inter-downcrossing (or inter-upcrossing) time of level x, $E(d_x)$ (or $E(u_x)$) in M/M/1: $\lambda = 1.0$, $\mu = 2.0$, $\rho = 0.5$

$$E(d_x) = \frac{e^{(\mu-\lambda)x}}{\lambda(1-\rho)}, x \geq 0, \tag{3.129}$$

which increases exponentially with x (Fig. 3.9).

Inter-upcrossing Time of a Level
Denote the inter-upcrossing time of level x by u_x. Inspection of sample paths of $\{W(t)\}_{t\geq 0}$ indicates that $u_x \underset{dis}{=} d_x$ due to the memoryless property of both the inter-arrival and service times in M/M/1. Hence the plot of $E(u_x)$ versus x is identical to that of $E(d_x)$ versus x in Fig. 3.9.

3.5.8 Number of Crossings of a Level in a Busy Cycle

Denote the number of downcrossings of level $x \geq 0$ during $d_0 (= \mathcal{C})$ by $\mathcal{D}_{d_0}(x) (= \mathcal{D}_{\mathcal{C}}(x))$. Since $\mathcal{D}_t(x)$ is the number of downcrossings of x during time interval $(0, t)$, from the renewal reward theorem

$$\frac{E(\mathcal{D}_{d_0}(x))}{E(d_0)} = \lim_{t\to\infty} \frac{E(\mathcal{D}_t(x))}{t} = f(x) = \lambda P_0 e^{-(\mu-\lambda)x}, x \geq 0.$$

Hence,

$$\begin{aligned} E(\mathcal{D}_{d_0}(x)) &= \lambda P_0 e^{-(\mu-\lambda)x} \cdot E(d_0) = \lambda P_0 e^{-(\mu-\lambda)x} \cdot \tfrac{1}{\lambda P_0} \\ &= e^{-(\mu-\lambda)x}, x \geq 0. \end{aligned} \tag{3.130}$$

Since $\lambda < \mu$, $E(\mathcal{D}_{d_0}(x)) \leq 1$. From (3.130), $E(\mathcal{D}_{d_0}(x))$ decreases exponentially as x increases.

Let $\mathcal{U}_{d_0}(x)$:= number of upcrossings of level x during d_0. Since $\mathcal{D}_{d_0}(x) = \mathcal{U}_{d_0}(x)$, $x \geq 0$, formula (3.130) gives

$$E(\mathcal{D}_{d_0}(0)) = E(\mathcal{U}_{d_0}(0)) = \lim_{x \downarrow 0} e^{-(\mu-\lambda)x} = 1. \tag{3.131}$$

Equation (3.131) is intuitive, since during $\mathcal{C}$ the SP hits level 0 from above exactly once, and egresses from level 0 above (upcrosses 0) exactly once. The SP hit occurs at the end of the embedded $\mathcal{B}$. The SP egress occurs at the start of the embedded $\mathcal{B}$.

3.5.9 Downcrossings at Different Levels

From formula (3.106) in Sect. 3.4.21 for the M/G/1 queue, E(number of SP downcrossings of x) during an inter-downcrossing time d_y is given by

$$E(\mathcal{D}_{d_y}(x)) = \frac{f(x)}{f(y)}, x \geq 0, \tag{3.132}$$

which implies in M/M/1

$$E(\mathcal{D}_{d(y)}(x)) = e^{-(\mu-\lambda)(x-y)},\ x \geq 0,\ y \geq 0, \tag{3.133}$$

since $f(x) = \lambda P_0 e^{-(\mu-\lambda)x}$, $x \geq 0$. From (3.133)

$$E(\mathcal{D}_{d_y}(x)) \begin{cases} < 1 \text{ if } x > y, \\ = 1 \text{ if } x = y, \\ > 1 \text{ if } x < y. \end{cases} \tag{3.134}$$

In (3.134) $E(\mathcal{D}_{d_x}(x)) = e^{-(\mu-\lambda)(x-x)} = 1, x \geq 0$, in agreement with intuition, upon examining a sample path of $\{W(t)\}_{t \geq 0}$.

Proposition 3.8 For arbitrary state-space levels $x, y, y_1, y_2, \ldots, y_n$

$$E(\mathcal{D}_{d_y}(x)) = E(\mathcal{D}_{d_y}(y_1)) \cdot E(\mathcal{D}_{d_{y_1}}(y_2)) \cdots E(\mathcal{D}_{d_{y_{n-1}}}(y_n)) \cdot E(\mathcal{D}_{d_{y_n}}(x)) \tag{3.135}$$

Proof From (3.132) we obtain

$$E(\mathcal{D}_{d_y}(x)) = \frac{f(x)}{f(y)} = \frac{f(y_1)}{f(y)} \cdot \frac{f(y_2)}{f(y_1)} \cdots \frac{f(y_n)}{f(y_{n-1})} \cdot \frac{f(x)}{f(y_n)}, n = 1, 2, \ldots$$

which is equivalent to (3.135). ■

Remark 3.18 The results in (3.132) and (3.135) hold for the standard M/G/1 queue, since the proofs depend only on having a Poisson arrival process. In order to apply these formulas to a specific M/G/1 model, we must have a formula for $f(x)$. The pdf $f(x)$ is known analytically in many M/G/1 models, e.g., M/D/1, M/Erl$_{k,\mu}$/1 and variants; otherwise $f(x)$ can be approximated or estimated by numerical or simulation methods.

3.5.10 Number Served in a Busy Period

Equation (3.88) in Sect. 3.4.14 yields

$$E(N_{\mathcal{B}}) = \frac{1}{P_0} = \frac{1}{1-\rho}. \tag{3.136}$$

It follows that the number served in a k-busy period, starting with k customers in the system at time 0, is equal to k/P_0 (see Exercise 17, p. 233 in Cooper [64]).

Remark 3.19 Sect. 5.1.15 in Chap. 5 considers the number of system times above or below a state-space level x during a sojourn a_y, $y \geq 0$, and related quantities. The M/M/1 results are presented in Sect. 5.1.15 because they follow as special cases of related results for G/M/1 queues given in Sects. 5.1.13 and 5.1.14.

3.5.11 Relationship Between M/M/1 and M/M/1/1

The M/M/1/1 queue is an M/M/1 variant restricted to having *at most one* customer in the system at all $t \geq 0$. The second /1 in the notation M/M/1/1

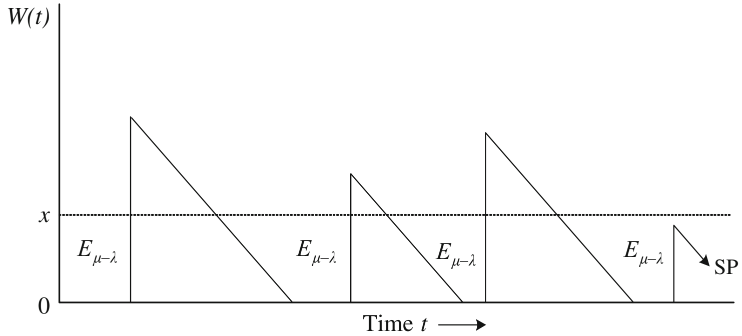


Fig. 3.10 Sample path of workload for $M_\lambda/M_{\mu-\lambda}/1/1$ queue with arrival rate λ and service rate $\mu - \lambda$. Blocked customers are cleared

refers to the *queue discipline*, namely: customers arriving when there is a customer in service, are blocked and cleared; customers arriving when the system is empty start service immediately. We compare the virtual wait process (same as workload) $\{W(t)\}_{t\geq 0}$ for M/M/1 (Fig. 3.8) and the *workload* process $\{W(t)\}_{t\geq 0}$ for M/M/1/1 (Fig. 3.10). (In M/M/1/1 all customers that get served wait 0.) The LC approach immediately connects the two models in steady-state. The cdf (3.123) and pdf (3.124) of *wait* (workload) in $M_\lambda/M_\mu/1$, are respectively *identical to* the steady-state cdf and pdf of *workload* in $M_\lambda/M_{\mu-\lambda}/1/1$ (arrival rate λ, ***service rate*** $\boldsymbol{\mu - \lambda}$).

This exact similarity of cdfs and pdfs is evident from a sample path of the workload $\{W(t)\}_{t\geq 0}$ in $M_\lambda/M_{\mu-\lambda}/1/1$ (Fig. 3.10). Fix level $x > 0$. The SP downcrossing rate of x is $f(x)$, as in Theorem 1.1. The SP upcrossing rate of x is $\lambda P_0 P(S > x) = \lambda P_0 e^{-(\mu-\lambda)x}$, since *all* SP jumps start at level 0, and are distributed as $\text{Exp}_{\mu-\lambda}$, where $\mu > \lambda$. In *both* M/M/1 and M/M/1/1, $E(\mathcal{B}) = \frac{1}{\mu-\lambda}$ and $P_0 = 1 - \lambda/\mu$. In $M_\lambda/M_{\mu-\lambda}/1/1$, the busy period $\mathcal{B}$ and the blocking time are identical, and are $\underset{dis}{=} (\mu - \lambda)e^{-(\mu-\lambda)x}, x > 0$. Also, the system times are both $\underset{dis}{=} (\mu - \lambda)e^{-(\mu-\lambda)x}, x > 0$. Although the expected busy periods are identical, their busy-period probability distributions are quite different—evident from formulas (3.121) and 3.122 involving Bessel functions. These probability distributions depend on the (different) jump structures of the $\{W(t)\}_{t\geq 0}$s. The $M_\lambda/M_{\mu-\lambda}/1/1$ workload has the same distribution as the wait (workload) in $M_\lambda/M_\mu/1$, namely

$$P_0 = 1 - \frac{\lambda}{\mu},\ f(x) = \lambda P_0 e^{-(\mu-\lambda)x}, x > 0.$$

A key point is that the pdf of workload $\{P_0, f(x)\}_{x>0}$ in $\text{M}_\lambda/\text{M}_{\mu-\lambda}/1/1$ is derived *immediately by inspection*, since *all* SP jumps start at level 0.

The foregoing relationship suggests re-examining integral equation (3.109). We substitute the $\text{M}_\lambda/\text{M}_{\mu-\lambda}/1/1$ solution into the integral in (3.109), i.e., $f(y) = \lambda P_0 e^{-(\mu-\lambda)y}$, and simplify. The immediate result is the solution for the $\text{M}_\lambda/\text{M}_\mu/1$ model $f(x) = \lambda P_0 e^{-(\mu-\lambda)x}$, $x > 0$, obtained while bypassing differential equation (3.110). This solution for $\text{M}_\lambda/\text{M}_{\mu-\lambda}/1/1$ "solves" integral equation (3.109) for $\text{M}_\lambda/\text{M}_\mu/1$.

This solution procedure suggests exploring conditions that facilitate solving for the pdf of state variables "by inspection" in more general models than M/M/1. The idea is to identify a "companion" or "isomorphic" model having a much simpler sample-path jump structure.

3.6 M/G/1: Service Time Depending on Wait

Consider an M/G/1 queue with arrival rate λ and service time depending on the wait before service, $S(W_q)$. Let $P(S(W_q) \leq x | W_q = y) = B_y(x)$, $x \geq 0$, $y \geq 0$, having pdf $b_y(x) = \frac{\partial}{\partial x} B_y(x)$, $x > 0$, $y \geq 0$, wherever the derivative exists. Let W_q have steady-state cdf $F(x)$, $x \geq 0$ and pdf $\{P_0, f(x)\}_{x>0}$ (assuming $\frac{d}{dx} F(x) = f(x)$ exists). We define $f(0) \equiv f(0^+)$ for convenience (does not add probability to P_0). A sample path of $\{W(t)\}_{t\geq 0}$ resembles that for the standard M/G/1 queue, except that the SP jump size (service time) generated by each arrival depends on the SP level at the start of the jump (actual wait).

Consider a fixed state-space level $x \geq 0$ in a sample path of $\{W(t)\}_{t\geq 0}$. The downcrossing rate of x is $f(x)$, by Theorem 1.1. The total *upcrossing* rate of x is

$$\lambda P_0 \overline{B}_0(x) + \lambda \int_{y=0}^{x} \overline{B}_y(x-y) f(y) dy; \; x > 0. \tag{3.137}$$

In (3.137) the term $\lambda P_0 \overline{B}_0(x)$ is the upcrossing rate of x by SP jumps at arrival instants when the system is empty. The term $\lambda \int_{y=0}^{x} \overline{B}_y(x-y) f(y) dy$ is the upcrossing rate of x by SP jumps at arrival instants when $\{W(t)\}_{t\geq 0}$ is at state-space levels $y \in (0, x)$. Rate balance across level x yields the integral equation for $f(x)$,

$$f(x) = \lambda P_0 \overline{B}_0(x) + \lambda \int_{y=0}^{x} \overline{B}_y(x-y) f(y) dy, \; x \geq 0. \tag{3.138}$$

As in the *standard* M/G/1 queue, letting $x \downarrow 0$ gives $f(0) = \lambda P_0 \overline{B}_0(0) = \lambda P_0$.

Integrating (3.138) on both sides with respect to x over $(0, \infty)$ gives

$$1 - P_0 = \rho_0 P_0 + \int_{y=0}^{\infty} \rho_y f(y)dy,$$
$$P_0 = \frac{1 - \int_{y=0}^{\infty} \rho_y f(y)dy}{1 + \rho_0}, \qquad (3.139)$$

where $\rho_y \equiv \lambda E(S_y)$, $y \geq 0$. (Eq. (3.139) is an implicit formula for P_0 since, from (3.138), $f(y)$ in the integral contains P_0. See Eq. (3.144) below for an explicit value for P_0 in the case where zero-wait customers receive special service.)

One way to deal with Eq. (3.138) is to partition the state space using $\{x_i\}_{i=0,\ldots,M+1}$, where integer $M \geq 0$, and

$$0 \equiv x_0 < x_1 < x_2 < \cdots < x_M < x_{M+1} \equiv \infty, \qquad (3.140)$$

as in the paper by Posner [117]. Denote the service time of a zero-wait customer as S_0, and of a y-waiting customer, $y \in (x_{i-1}, x_i)$, as S_i. Assume the service-time is S_0 for all arrivals who wait zero, and S_i for all arrivals who wait $y \in (x_{i-1}, x_i)$. Thus the cdf of service time is

$$\begin{aligned} &B_0(x),\, x > 0 \text{ for zero-wait arrivals,} \\ &B_j(x),\, x > 0 \text{ for all arrivals who wait } y \in (x_{i-1}, x_i),\, i = 1, \ldots, M+1. \end{aligned} \qquad (3.141)$$

Integral equation (3.138) can then be written

$$\begin{aligned} f(x) = \lambda P_0 \overline{B}_0(x) &+ \lambda \textstyle\sum_{i=1}^{j-1} \int_{y=x_{i-1}}^{x_i} \overline{B}_i(x-y) f(y)dy \\ &+ \lambda \textstyle\int_{y=x_{j-1}}^{x} \overline{B}_j(x-y) f(y)dy,\; x \in (x_{j-1}, x_j),\; j = 1, \ldots, M+1. \end{aligned} \qquad (3.142)$$

where $\sum_{i=1}^{0} \equiv 0$. In (3.142), for any fixed $x > 0$, the right side is the upcrossing rate of level x. Thus, we have constructed integral equation (3.142) in a fast, easy, intuitive, straightforward manner using LC.

Queues with service time depending on wait appear in the literature in, e.g., ([57, 58]). The single-server model was treated in the literature using Laplace transforms in [108], and by the embedded Markov chain technique using a Lindley recursion in [117], who obtained an explicit solution for $\{P_0, f(x)\}_{x>0}$.

Remark 3.20 Deriving (3.142) using the embedded Markov chain technique is "relatively" tedious and purely algebraic (see Sect. 1.3 in Chap. 1). The single-server model was generalized to multiple servers using the embedded Markov chain technique in [48, 49] (the *original* topic and methodology of the author's Ph.D. thesis). After my discovery of LC in 1974, the model solution was completely revised using LC in the Ph.D. thesis [11], which greatly simplified the derivation of the integral equations. An analysis of an M/M/2 model with service time depending on wait is given in [53]; a revised version appears in Sect. 4.11 below.

3.6.1 M/G/1: Zero-Wait Arrivals Get Special Service

A particular case of M/G/1 with service time depending on wait, which has many useful applications, is a model where the initial customer of each busy period receives special service; we set $M = 0, x_0 = 0, x_1 = \infty$ in the state-space partition (3.140). (e.g., see, [144]; also Example 3.7 in Sect. 3.4.19; the last division of this Section; Example 3.11 in Sect. 3.8.5 below).

The integral equation (3.142) reduces to

$$f(x) = \lambda P_0 \overline{B}_0(x) + \lambda \int_{y=0}^{\infty} \overline{B}_1(x-y) f(y) dy, x > 0. \tag{3.143}$$

Integrating (3.143) with respect to x over $(0, \infty)$, using $\int_{x=0}^{\infty} f(x)dx = 1 - P_0$, gives

$$P_0 = \frac{1 - \lambda E(S_1)}{1 - \lambda E(S_1) + \lambda E(S_0)} = \frac{1 - \rho_1}{1 - \rho_1 + \rho_0}. \tag{3.144}$$

A necessary condition for stability is $\rho_1 < 1$ (guarantees $P_0 > 0$ and $\{0\}$ is a positive recurrent state). (If $\rho_1 > 1$ then $1 - \rho_1 < 0$. We would then need $1 - \rho_1 + \rho_0 < 0$ to ensure that $P_0 > 0$, causing $|1 - \rho_1 + \rho_0| < |1 - \rho_1|$. But that would imply $P_0 > 1$ in (3.144), a contradiction. If $\rho_1 = 1$, then $P_0 = 0$, which would imply the queue is unstable.)

Multiplying both sides of (3.143) by x, and integrating for $x \in (0, \infty)$ gives a Pollaczek-Khinchine (P-K)-like result for the expected wait before service

$$E(W_q) = \frac{\lambda (E(S_0^2) + E(S_1^2))}{2(1 - \lambda E(S_1))}. \tag{3.145}$$

Expected Busy Period When $M = 0$ in Partition (3.140)

In this case there are two types of arrivals. Customers that don't have to wait (wait time = 0) have service time S_0. Customers that wait a positive time have service time S_1. We determine $E(\mathcal{B})$ where $\mathcal{B}$:= busy period when initial service is $\underset{dis}{=} S_0$ and all other services are $\underset{dis}{=} S_1$.

Method 1
The busy period is

$$\mathcal{B} = S_0 + \sum_{i=1}^{A_{S_0}} \mathcal{B}_{1,i}, \tag{3.146}$$

where $A_{S_0} \underset{dis}{=}$ number of arrivals, including pseudo arrivals during S_0, the initial service time of $\mathcal{B}$ (see Sect. 3.4.12 and Fig. 3.6 therein); the sub-busy periods $\{\mathcal{B}_{1,i}\}_{j=1,2,...}$ are i.i.d. r.v.s distributed as a busy period $\mathcal{B}_1$ in a standard M_λ/G/1 queue with service time S_1 (see Fig. 3.11). The $\mathcal{B}_{1,i}$s are independent of A_{S_0}. Taking the expected value in (3.146) gives

$$\begin{aligned} E(\mathcal{B}) &= E(S_0) + \lambda E(S_0)E(\mathcal{B}_1) = E(S_0) + \lambda E(S_0)\frac{E(S_1)}{1-\lambda E(S_1)} \\ &= \frac{E(S_0)}{1-\rho_1} = \frac{E(S_0)}{P_{0,1}}, \end{aligned} \tag{3.147}$$

where $P_{0,1} = P(\text{wait} = 0)$ in the standard M/G/1 with common service time S_1.

Method 2
Applying the LC-based result for the expected busy period in M/G/1 (3.82), and using P_0 in (3.144) we get (3.147) as follows:

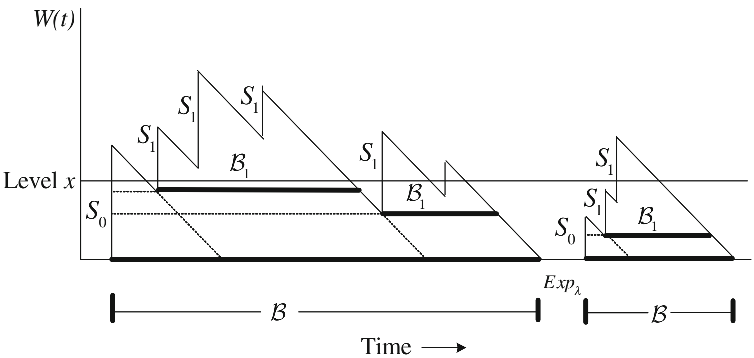


Fig. 3.11 $\mathcal{B}$'s are busy periods in M_λ/G/1 with zero-waits receiving service time $\underset{dis}{=} S_0$. $\mathcal{B}_1$'s are busy periods of M_λ/G/1 with all service times $\underset{dis}{=} S_1$, generated by pseudo arrivals during S_0

$$E(\mathcal{B}) = \frac{1-P_0}{\lambda P_0} = \frac{1-\frac{1-\lambda E(S_1)}{1-\lambda E(S_1)+\lambda E(S_0)}}{\lambda \frac{1-\lambda E(S_1)}{1-\lambda E(S_1)+\lambda E(S_0)}} = \frac{E(S_0)}{1-\lambda E(S_1)} = \frac{E(S_0)}{1-\rho_1} = \frac{E(S_0)}{P_{0,1}}.$$

We can now derive the expression for P_0 directly using the expression for $E(\mathcal{B})$. Thus

$$P_0 = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda}+E(\mathcal{B})} = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda}+\frac{E(S_0)}{1-\rho_1}} = \frac{1-\rho_1}{1-\rho_1+\rho_0}.$$

3.6.2 M/M/1: Zero-Wait Arrivals Get Special Service

We now derive the pdf $\{P_0, f(x)\}_{x>0}$ when service times are exponentially distributed with $B_0(x) = 1 - e^{-\mu_0 x}$, $B_1(x) = 1 - e^{-\mu_1 x}$. Substituting $e^{-\mu_0 x}$ for $\overline{B}_0(x)$ and $e^{-\mu_1 x}$ for $\overline{B}_1(x-y)$ in (3.143) and applying differential operator $\langle D+\mu_0\rangle \langle D+\mu_1\rangle$ (equivalent to differentiating twice with respect to x, followed by some algebra) yields a second order differential equation

$$\langle D+\mu_1-\lambda\rangle \langle D+\mu_0\rangle f(x) = 0,$$

with solution

$$f(x) = ae^{-(\mu_1-\lambda)x} + be^{-\mu_0 x}, \ x > 0, \tag{3.148}$$

provided $\mu_0 \neq \mu_1 - \lambda$ (if $\mu_0 = \mu_1 - \lambda$, $f(x)$ in the differential equation has a different solution; see, e.g., pp. 106–113 in [10]). Constants a, b are obtained from two independent initial conditions:

$$f(0) = \lambda P_0 \text{ and } f'(0) = -\mu_0 \lambda P_0 + \lambda f(0),$$

giving

$$a = \frac{-\lambda^2 P_0}{\mu_1-\mu_0-\lambda}, \ b = \frac{\lambda(\mu_1-\mu_0)P_0}{\mu_1-\mu_0-\lambda}, \ P_0 = \frac{1-\rho_1}{1-\rho_1+\rho_2}, \tag{3.149}$$

where $\rho_i = \lambda/\mu_i, i = 1, 2$. (See Example 3.12 in Sect. 3.17.3 for an alternative solution technique to derive $f(x), x > 0$.)

Expected Busy Period When Service Times Are Exponential
From Eq. (3.147),

$$E(\mathcal{B}) = \frac{\frac{1}{\mu_0}}{1-\frac{\lambda}{\mu_1}} = \frac{\mu_1}{\mu_0(\mu_1-\lambda)}.$$

Mild check on $E(\mathcal{B})$: If $\mu_0 = \mu_1 = \mu$ then $E(\mathcal{B}) = 1/(\mu - \lambda)$, as in the standard $M_\lambda/M_\mu/1$ queue.

Sojourn Above Level x When Service Times Are Exponential
Let γ_x denote the excess above x given $\{W(t)\}_{t\geq 0}$ upcrosses level x. Then $\gamma_x \underset{dis}{=} \text{Exp}_{\mu_0}$ or $\gamma_x \underset{dis}{=} \text{Exp}_{\mu_1}$, where $S_0 \underset{dis}{=} \text{Exp}_{\mu_0}$ and $S_1 \underset{dis}{=} \text{Exp}_{\mu_1}$. Then

$$E(a_x) = p_1(x)\frac{1/\mu_0}{1-\frac{\lambda}{\mu_1}} + p_2(x)\frac{1/\mu_1}{1-\frac{\lambda}{\mu_1}},$$

where $p_i(x) := P$(an upcrossing is due to the service time jump of a type-i arrival), $i = 1, 2$, and $p_1(x) + p_2(x) = 1$. If $1/\mu_0 < 1/\mu_1$ then $E(a_x) < E(\mathcal{B})$. If $1/\mu_0 > 1/\mu_1$ then $E(a_x) > E(\mathcal{B})$. Moreover (see derivation of (3.170) in Sect. 3.8.6 below)

$$p_1(x) = \frac{P_0 e^{-\mu_0 x}}{f(x)/\lambda}, \qquad p_2(x) = \frac{\int_{y=0}^{x} B(x-y)f(y)dy}{f(x)/\lambda},$$

where $\{P_0, f(x)\}_{x>0}$ is given in (3.148) and (3.149).

3.7 Expected Sojourn Above Level *x* in M/G/1

We derive $E(a_x)$ in M/G/1 with general service time S, where $a_x :=$ *sojourn by* $\{W(t)\}_{t\geq 0}$ *above a fixed level* $x \geq 0$. The derivation utilizes a connection with M/G/1 where *zero-wait customers receive special service* (Sect. 3.6.1). Consider a sample path of $\{W(t)\}_{t\geq 0}$. A sojourn a_x is initiated by the excess of an upcrossing of x. We derive a formula for $E(a_x)$ when S is a positive continuous random variable having pdf $b(y)$, $y > 0$, cdf $B(y)$, $y > 0$, and $\overline{B}(y) = 1 - B(y)$, $y \geq 0$. Let $\{P_0, f(x)\}_{x>0}$ be the limiting mixed pdf of $\{W(t)\}_{t\geq 0}$ as $t \to \infty$. From Theorem 1.1 in Chap. 1, $\{P_0, f(x)\}_{x>0}$ is determined by the equations

$$\begin{gathered} f(x) = \lambda P_0 \overline{B}(x) + \lambda \int_{y=0}^{x} \overline{B}(x-y) f(y)dy, \, x > 0, \\ P_0 + \int_{x=0}^{\infty} f(x)dx = 1. \end{gathered} \tag{3.150}$$

Let $\gamma_x :=$ *excess over* x, which initiates an a_x whenever $\{W(t)\}_{t\geq 0}$ upcrosses level x. The a_xs are i.i.d. random variables since they occur within regenerative cycles delimited by successive level-x downcrossings (one a_x per cycle) (see Fig. 3.12).

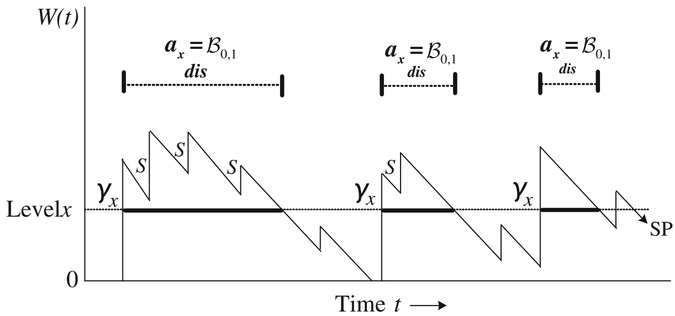


Fig. 3.12 γ_x := excess over level x. a_x := sojourn above level x. $\mathcal{B}_{0,1}$:= busy period when zero-wait service $\underset{dis}{=} \gamma_x$ and other services $\underset{dis}{=} S$

The *first* jump size of each a_x is $\underset{dis}{=} \gamma_x$. However, *during* a_x, all jump sizes are $\underset{dis}{=} S$. Thus, $a_x \underset{dis}{=}$ *busy period of an MG/1 queue where the first service is exceptional (special), denoted as* $\mathcal{B}_{0,1}$ in Fig. 3.12 (see formula (3.147) in Sect. 3.6.1).

Let $G_{\gamma_x}(z)$, $z > 0$, denote the cdf of γ_x.

Theorem 3.7 For fixed $x \geq 0$,

$$E(a_x) = \frac{E(\gamma_x)}{P_0} = \frac{\int_{z=0}^{\infty} \left(1 - G_{\gamma_x}(z)\right) dz}{P_0}$$
$$= \frac{\int_{z=0}^{\infty} \left[\lambda \int_{y=0}^{x} \overline{B}(x+z-y) dF(y)\right] dz}{f(x) P_0}. \qquad (3.151)$$

Proof We employ the equation

$$\left(1 - G_{\gamma_x}(z)\right) f(x) = \lambda P_0 \overline{B}(x+z) + \lambda \int_{y=0}^{x} \overline{B}(x+z-y) f(y) dy, x \geq 0, z > 0, \qquad (3.152)$$

where the LHS = $P(_{\gamma_x} > z | SP$ upcrosses level $x) \times$ (rate at which SP upcrosses level x), mindful that $f(x)$ is both the up- and downctossing rate of x (see Theorem 1.1). Thus the LHS is the upcrossing rate of level $x + z$ by the excess over x, of jumps starting below x. The RHS is a *different expression for* the upcrossing rate of level $x + z$ by jumps staring below x; all upcrosses of $x + z$ occur during the excess over x. Therefore Eq. (3.152) follows, implying

$$1 - G_{\gamma_x}(z) = \frac{\lambda P_0 \overline{B}(x+z) + \lambda \int_{y=0}^{x} \overline{B}(x+z-y) f(y) dy}{f(x)},$$

so that

$$E(\gamma_x) = \int_{z=0}^{\infty} \left(1 - G_{\gamma_x}(z)\right) dz = \frac{\int_{z=0}^{\infty} \left[\lambda \int_{y=0}^{x} \overline{B}(x+z-y) dF(y)\right] dz}{f(x)}.$$

Formula (3.151) follows since $E(a_x) = E(\mathcal{B}_{0,1}) = \frac{E(\gamma_x)}{P_0}$ by formula (3.147) in Sect. 3.6.1. ■

We can solve the equations in (3.150) for $\{P_0, f(x)\}_{x \geq 0}$ analytically, numerically or by simulation; or obtain an approximate solution. Theorem 3.7 then enables us to calculate $E(a_x)$, $x \geq 0$.

3.8 M/G/1 with Multiple Poisson Inputs

Customers arrive at a single-server system in N independent Poisson streams at rates λ_i, $i = 1, \ldots, N$, $\sum_{i=1}^{N} \lambda_i = \lambda$. Let the corresponding service times be S_i having cdf $B_i(x)$, $\overline{B}_i(x) = 1 - B_i(x)$, $x \geq 0$, and pdf $b_i(x) = \frac{d}{dx} B_i(x)$, $x > 0$, wherever the derivative exists. The service discipline is FCFS. The service time of an arbitrary arrival is S_i with probability λ_i/λ. Denote the steady-state pdf and cdf of the wait before service, W_q, by $\{P_0, f(x)\}_{x>0}$ and $F(x)$, $x \geq 0$, respectively.

Due to independent Poisson arrivals, we may view the system as an M/G/1 queue with arrival rate λ and service time

$$S = \begin{cases} S_1 \text{ with probability } \frac{\lambda_1}{\lambda}, \\ \cdots \\ S_N \text{ with probability } \frac{\lambda_N}{\lambda}. \end{cases} \tag{3.153}$$

S has a *mixture* probability distribution with mixture components S_i and mixture weights λ_i/λ (> 0) such that $\sum_{i=1}^{N} (\lambda_i/\lambda) = 1$. Hence the nth moment $S^n = S_i^n$ with probability λ_i/λ, $i = 1, \ldots, N$; $n = 1, 2, \ldots$. Thus

$$E(S) = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} E(S_i), \; E(S^2) = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} E(S_i^2). \tag{3.154}$$

Employing $\rho_i = \lambda_i E(S_i)$, $i = 1, \ldots, N$,

$$P_0 = 1 - \lambda E(S) = 1 - \sum_{i=1}^{N} \rho_i. \tag{3.155}$$

Stability

The system is stable iff every typical sample path of $\{W(t)\}_{t\geq 0}$ returns to state $\{0\}$ iff $P_0 > 0$, i.e.,

$$\sum_{i=1}^{N} \rho_i < 1. \tag{3.156}$$

3.8.1 Integral Equation for PDF of Wait

Sample paths of $\{W(t)\}_{t\geq 0}$ resemble those of the standard M/G/1 queue, except that the jump size due to an arrival is $\underset{dis}{=} S_i$ with probability λ_i/λ, having cdf $B_i(\cdot)$, $i = 1, \ldots, N$. Thus jumps $\underset{dis}{=} S_i$ occur at Poisson rate λ_i. By Theorem 1.1, for a fixed state-space level $x > 0$, the SP downcrossing rate is $f(x)$. The SP upcrossing rate for type i arrivals is

$$\lambda_i P_0 \overline{B}_i(x) + \lambda_i \int_{y=0}^{x} \overline{B}_i(x-y) f(y) dy, i = 1, \ldots, N.$$

Balancing the *total* SP down- and upcrossing rates of level x for all customer types, yields the integral equation for $f(x)$,

$$f(x) = \sum_{i=1}^{N} \lambda_i \left[P_0 \overline{B}_i(x) + \int_{y=0}^{x} \overline{B}_i(x-y) f(y) \right] dy,$$

or

$$f(x) = \lambda P_0 \left(\sum_{i=1}^{N} \frac{\lambda_i}{\lambda} \overline{B}_i(x) \right) + \lambda \int_{y=0}^{x} \left(\sum_{i=1}^{N} \frac{\lambda_i}{\lambda} \overline{B}_i(x-y) \right) f(y) dy. \tag{3.157}$$

Integral equation (3.157) is in the form of the analogous integral equation (3.34) for the pdf of wait in a *standard* M/G/1 queue with $\lambda = \sum_{i=1}^{N} \lambda_i$, and $\overline{B}(x) = \sum_{i=1}^{N} (\lambda_i/\lambda) \overline{B}_i(x)$, $\dot{x} > 0$.

3.8.2 Expected Wait Before Service

Since $E(S^2) = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} E(S_i^2)$, the Pollaczek-Khinchine (P-K) formula (3.63) gives the expected wait before service as

$$E(W_q) = \frac{\lambda E(S^2)}{2(1 - \lambda E(S))} = \frac{\lambda \sum_{i=1}^{N} (\lambda_i/\lambda)\, E(S_i^2)}{2(1 - \sum_{i=1}^{N} \rho_i)} = \frac{\sum_{i=1}^{N} \lambda_i E(S_i^2)}{2P_0}. \tag{3.158}$$

Alternatively, we can obtain $E(W_q)$ in (3.158) directly from (3.157) upon multiplying both sides by x, then integrating both sides with respect to $x \in (0, \infty)$, changing the order of integration in the double integral, and doing some algebra.

3.8.3 Expected Number in Queue

Let L_q = expected number of units in the queue before service in the steady state. Then by $\boldsymbol{L = \lambda W}$ (Little [110]) and (3.158)

$$L_q = \lambda E(W_q) = \frac{\lambda \sum_{i=1}^{N} \lambda_i E(S_i^2)}{2P_0}. \tag{3.159}$$

Denote the steady-state expected number of type i units in the queue by $L_{q,i}$. Let the wait of an arbitrary *type i customer* be $W_{q,i}$, and the wait of an *arbitrary customer* be W_q. Then $W_{q,i} \underset{dis}{=} W_q$, because the waiting time of any arrival depends only on the current workload at the arrival instant. Thus $E(W_{q,i}) = E(W_q)$, $i = 1, \ldots, N$, and by $\boldsymbol{L = \lambda W}$,

$$L_{q,i} = \lambda_i E(W_{q,i}) = \lambda_i E(W_q) = \frac{\lambda_i \sum_{i=1}^{N} \lambda_i E(S_i^2)}{2P_0}, i = 1, \ldots, N. \tag{3.160}$$

3.8.4 Expected Busy Period

Applying (3.82) and (3.155), the expected busy period is given by

$$E(\mathcal{B}) = \frac{1 - P_0}{f(0)} = \frac{1 - P_0}{\lambda P_0} = \frac{\sum_{i=1}^{N} \rho_i}{\lambda \left(1 - \sum_{i=1}^{N} \rho_i\right)}. \tag{3.161}$$

As a mild check on formula (3.161), let $S = S_i$ with probability $1/N$, $i = 1, \ldots, N$. Then $\lambda_i/\lambda \equiv 1/N$ so that $\sum_{i=1}^{N} (\lambda_i/\lambda) = \sum_{i=1}^{N} (1/N) = 1$, $\rho_i \equiv \lambda_i E(S_i) = (\lambda/N)\, E(S_i)$ and $\sum_{i=1}^{N} \rho_i = (\lambda/N) \sum_{i=1}^{N} E(S_i)$. The multiple Poisson input model reduces to a standard M/G/1 queue with arrival rate λ and $E(S) = \frac{1}{N} \sum_{i=1}^{N} E(S_i)$. From (3.161)

$$E(\mathcal{B}) = \frac{\frac{\lambda}{N} \sum_{i=1}^{N} E(S_i)}{\lambda \left(1 - \sum_{i=1}^{N} \rho_i\right)} = \frac{\frac{1}{N} \sum_{i=1}^{N} E(S_i)}{1 - \sum_{i=1}^{N} \rho_i} = \frac{E(S)}{P_0},$$

which is the formula for $E(\mathcal{B})$ for the standard M/G/1 queue.

3.8.5 M/M/1 with Multiple Poisson Inputs

To outline a solution technique for integral equation (3.157), we assume the service times are $\underset{dis}{=} \text{Exp}_{\mu_i}$ with $\overline{B}_i(x) = e^{-\mu_i x}$, $i = 1, 2, \ldots, N$. Then (3.157) becomes

$$f(x) = \sum_{i=1}^{N} \lambda_i \left[P_0 e^{-\mu_i x} + \int_{y=0}^{x} e^{-\mu_i (x-y)} f(y) dy \right], \; x > 0. \qquad (3.162)$$

We can apply the differential operator $\langle D + \mu_1 \rangle \ldots \langle D + \mu_N \rangle$ to Eq. (3.162), to derive and solve analytically an Nth order differential equation with constant coefficients for $f(x)$, using initial conditions to obtain the constants of integration.

The differential operator $\langle D + \mu_i \rangle$ is commutative with respect to exponential functions of the form $e^{\alpha x + \beta}$, where α and β are constants, i.e., for any permutation $(i_1, i_2, \ldots, i_N)$ of the numbers $(1, 2, \ldots, N)$

$$\begin{aligned} \langle (D + \mu_1) \cdots (D + \mu_N) \rangle e^{\alpha\alpha x + \beta} &= \langle D + \mu_1 \rangle \cdots \langle D + \mu_N \rangle \, e^{\alpha x + \beta} \\ &= \left\langle D + \mu_{i_1} \right\rangle \cdots \left\langle D + \mu_{i_N} \right\rangle e^{\alpha x + \beta} \\ &= \langle (D + \mu_{i_1}) \cdots (D + \mu_{i_N}) \rangle e^{\alpha x + \beta}. \end{aligned}$$

The commutativity property simplifies the transformation of an integral equation into a differential equation, when the kernel of any integral in the equation has an exponential form like $e^{-\mu_i (x-y)}$ in (3.162).

Expected Wait and Expected Number in Queue

If $S_i \underset{dis}{=} \text{Exp}_{\mu_i}$ then $E(S_i^2) = 2/\mu_i^2$, which substituted into (3.158) and (3.159) respectively, yield

$$E(W_q) = \frac{\sum_{i=1}^{N} \frac{\lambda_i}{\mu_i^2}}{1 - \sum_{i=1}^{N} \frac{\lambda_i}{\mu_i}}, \tag{3.163}$$

$$L_q = \frac{\lambda \sum_{i=1}^{N} \frac{\lambda_i}{\mu_i^2}}{1 - \sum_{i=1}^{N} \frac{\lambda_i}{\mu_i}}. \tag{3.164}$$

Similarly, substituting into (3.160), gives

$$L_{q,i} = \frac{\lambda_i \sum_{i=1}^{N} \frac{\lambda_i}{\mu_i^2}}{1 - \sum_{i=1}^{N} \frac{\lambda i}{\mu_i}}, \, i = 1, \ldots, N. \tag{3.165}$$

Two Customer Types

To illustrate the solution, we consider two distinct customer types, and derive $\{P_0, f(x)\}$. Set $N = 2$ in (3.162). Applying differential operator $\langle D + \mu_1 \rangle \langle D + \mu_2 \rangle$ to both sides, gives a second order differential equation

$$\left\langle D^2 + (\mu_1 + \mu_2 - \lambda) D + (\mu_1 \mu_2 - \mu_1 \lambda_2 - \mu_2 \lambda_1) \right\rangle f(x) = 0$$

having solution

$$f(x) = a e^{R_1 x} + b e^{R_2 x}, \tag{3.166}$$

where R_i, $i = 1, 2$ are the roots for z in the characteristic equation

$$z^2 + (\mu_1 + \mu_2 - \lambda) z + \mu_1 \mu_2 - \mu_1 \lambda_2 - \mu_2 \lambda_1 = 0.$$

Both roots are negative since the product $R_1 R_2 = \mu_1 \mu_2 - \mu_1 \lambda_2 - \mu_2 \lambda_1 > 0$ (equivalent to $1 - \rho_1 - \rho_2 > 0$, the stability condition), and $R_1 + R_2 = -(\mu_1 + \mu_2 - \lambda) < 0$. Constants a and b are determined by applying two independent initial conditions for $f(0) = a + b$ and $f'(0) = R_1 a + R_2 b$, obtained from (3.166) and also from (3.162), resulting in two equations for a, b:

$$\begin{aligned} a + b &= \lambda P_0, \\ R_1 a + R_2 b &= -(\mu_1 \lambda_1 + \mu_2 \lambda_2) P_0 + \lambda f(0) = -(\mu_1 \lambda_1 + \mu_2 \lambda_2 - \lambda^2) P_0. \end{aligned} \tag{3.167}$$

Thus $f(x)$ is given by (3.166) where $[a, b]$ is the solution of the two equations in (3.167):

$$\begin{aligned} a &= -P_0(R_2\lambda - \lambda^2 + \lambda_1\mu_1 + \lambda_2\mu_2)/(R_1 - R_2), \\ b &= P_0(R_1\lambda - \lambda^2 + \lambda_1\mu_1 + \lambda_2\mu_2)/(R1 - R2), \end{aligned} \tag{3.168}$$

and

$$\left.\begin{aligned} &P_0 = 1 - \tfrac{\lambda_1}{\mu_1} - \tfrac{\lambda_2}{\mu_2}, \\ &R_1 = \tfrac{-\boldsymbol{B}}{2} + \tfrac{\sqrt{\boldsymbol{B}^2 - 4\boldsymbol{AC}}}{2},\ R_2 = \tfrac{-\boldsymbol{B}}{2} - \tfrac{\sqrt{\boldsymbol{B}^2 - 4\boldsymbol{AC}}}{2}, \\ &\boldsymbol{A} = 1,\ \boldsymbol{B} = \mu_1 + \mu_2 - \lambda,\ \boldsymbol{C} = \mu_1\mu_2 - \mu_1\lambda_2 - \mu_2\lambda_1. \end{aligned}\right\} \tag{3.169}$$

Example 3.10 Consider a numerical example with $N = 2$, $\lambda_1 = 1$, $\lambda_2 = 0.5$, $\mu_1 = 3$, $\mu_2 = 2$. Then $P_0 = 0.4167$, $R_1 = -1.0$, $R_2 = -2.5$, $a = 0.555555$, $b = 0.069444$, and

$$f(x) = 0.555555\, e^{-1.0x} + 0.069444\, e^{-2.5x},\ x > 0.$$

A computational check shows that $F(\infty) = 1$, and $f(0) = \lambda P_0$, i.e.,

$$\begin{aligned} F(\infty) &= P_0 + \int_{x=0}^{\infty} f(x)dx \\ &= 0.4167 + \int_{x=0}^{\infty} (0.555555\, e^{-1.0x} + 0.069444\, e^{-2.5x})dx = 1, \\ f(0) &= a + b = \lambda P_0 = 0.625. \end{aligned}$$

3.8.6 Expected Sojourn Above Level x − E(a_x)

Let $p_i(x) := P$ (arrival type $i \mid \{W(t)\}_{t\geq 0}$ upcrosses x), $i = 1, \ldots N$. Using Bayes' formula, and Eq. (3.162),

$$\begin{aligned} p_i(x) &= \frac{P(\{W(t)\}_{t\geq 0} \text{ upcrosses } x \mid \text{arrival type } i) \cdot P(\text{arrival type } i)}{P(\{W(t)\}_{t\geq 0} \text{ upcrosses } x)} \\ &= \frac{\left(P_0\overline{B}_i(x) + \int_{y=0}^{x} \overline{B}_i(x-y)f(y)dy\right) \cdot (\lambda_i/\lambda)}{\sum_{i=1}^{N}\left[P_0\,(\lambda_i/\lambda)\,\overline{B}_i(x) + \int_{y=0}^{x} (\lambda_i/\lambda)\,\overline{B}_i(x-y)f(y)dy\right]} \\ &= \frac{P_0\,(\lambda_i/\lambda)\,\overline{B}_i(x) + \int_{y=0}^{x} (\lambda_i/\lambda)\,\overline{B}_i(x-y)f(y)dy}{f(x)/\lambda}. \end{aligned} \tag{3.170}$$

Let $\gamma_{S_i}(x)$:= excess above level x due to a type-i upcrossing of x. Sojourn $a_x \underset{dis}{=} \mathcal{B}_{(i)}$ with probability $p_i(x)$, i.e., a_x has a mixture distribution with components $\mathcal{B}_{(i)}$ and mixture probabilities $p_i(x)$. which is a busy period where zero-waits get service $S_0 \underset{dis}{=} \gamma_{S_i}(x)$, and positive-waits get service time S_1 (see formula (3.153)). Applying (3.147) in Sect. 3.6.1 and formula (3.161) gives

$$E(a_x) = \frac{E(\gamma_S(x))}{1 - \sum_{i=1}^N \rho_i}. \tag{3.171}$$

$E(a_x)$ in M/M/1 with Two Types of Poisson Inputs

Consider two types of input with $S_i \underset{dis}{=} \text{Exp}_{\mu_i}$, $i = 1, 2$. By the memoryless property $\gamma_{S_i}(x) \underset{dis}{=} \text{Exp}_{\mu_i}$, $i = 1, 2$, for all $x \geq 0$. If the upcrossing of x that initiates a_x is due to a type-i arrival then $a_x \underset{dis}{=} \mathcal{B}_i$, an M/G/1 busy period with *initial* service $S_0 = \text{Exp}_{\mu_i}$, and other service times a mixture of Exp_{μ_1} and Exp_{μ_2}. Thus

$$E(a_x|\text{excess} = \gamma_{S_i}(x)) = \frac{1}{\mu_i\left(1 - \sum_{i=1}^2 \rho_i\right)}, i = 1, 2,$$

and

$$E(a_x) = \sum_{i=1}^{2} p_i(x) \frac{1}{\mu_i\left(1 - \sum_{i=1}^2 \rho_i\right)}, \tag{3.172}$$

From (3.170) the probability of a type-i upcrossing of level x is

$$p_i(x) = \frac{P_0 (\lambda_i/\lambda) e^{-\mu_i x} + \int_{y=0}^{x} (\lambda_i/\lambda) e^{-\mu_i (x-y)} f(y)dy}{f(x)/\lambda}, i = 1, 2.$$

Example 3.11 Using the input values and $f(x)$, $x > 0$, in Example 3.10, Sect. 3.8.5, we compute

$$p_1(x) = \frac{0.27777\, e^{-3.x}(0.5\, e^{0.5x} + e^{2.x})}{0.55555\, e^{-x} + 0.06944\, e^{-2.5x}}, \qquad p_2(x) = 1 - p_1(x), x \geq 0.$$

Applying formula (3.172) and using $p_2(x) = 1 - p_1(x)$ gives

$$E(a_x) = p_1(x)\frac{1}{1.25} + p_2(x)\frac{1}{0.83333}.$$

Fig. 3.13 $E(a_x)$ versus x in Example 3.11. $E(a_0) = E(\mathcal{B})$

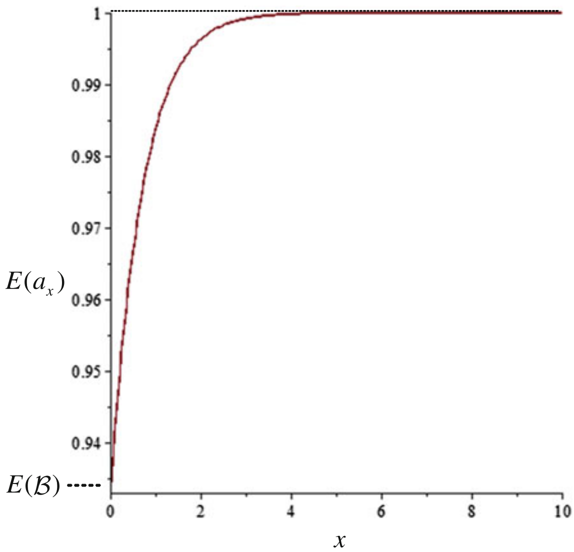
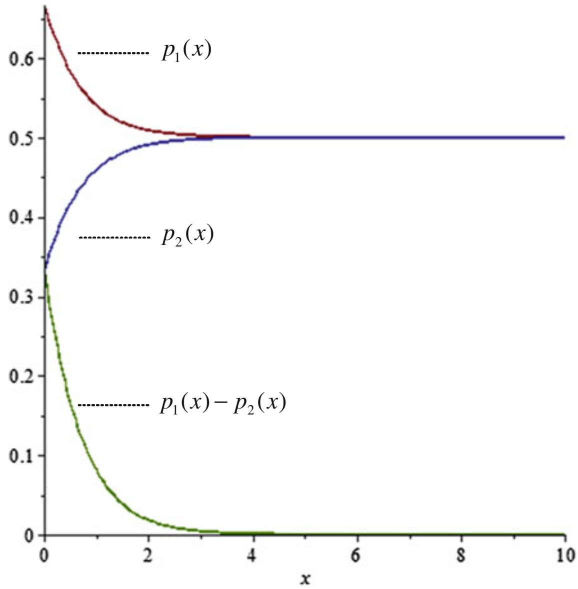


Fig. 3.14 Changes in $p_i(x)$, $i = 1, 2$ and $p_1(x) - p_2(x)$, $x > 0$, in Example 3.11



From (3.161) the expected value of the M/G/1 busy period with multiple Poisson inputs is $E(\mathcal{B}) = 0.9333$. We see from a plot that $E(a_x) > E(\mathcal{B}), x > 0$, $E(a_0) = E(\mathcal{B})$ and $\lim_{x\to\infty} E(a_x) = 1$ (Fig. 3.13). Moreover $\lim_{x\to\infty} p_i(x) = 0.5, i = 1, 2$ (Fig. 3.14). These observations are readily verified analytically. The growth of $E(a_x)$ as x increases is due to the fact that $E(\gamma_x) = 1/\mu_i$ independent of x, implying that the evolution of $E(a_x)$ on $(0, \infty)$ is determined by the $p_i(x)$s which do change as x increases. (In examples where $E(\gamma_x)$ depends on x the properties of $E(a_x)$ would be different.)

3.9 M/G/1: Wait-Number Dependent Service

Arrivals occur at Poisson rate λ. The queue discipline is FCFS. The service time is denoted as $S(N_q)$ where N_q := number of customers left waiting in the queue *just after* a start of service. Thus $N_q \in \{0, 1, \ldots\}$. For exposition, we assume there are two different types of service. Let

$$S(N_q) = \begin{cases} S_0, & \text{if } N_q = 0, \\ S, & \text{if } N_q = 1, 2, \ldots. \end{cases}$$

Let $P(S_0 \le x) = B_0(x)$, $\overline{B}_0(x) = 1 - B_0(x)$, $P(S \le x) = B(x)$, $\overline{B}(x) = 1 - B(x)$. Denote the steady-state wait before service as W_q having cdf $P(W_q \le x) = F(x)$ and mixed pdf $\{P_0, f(x)\}_{x>0}$ wherever $\frac{d}{dx}F(x)$ $(=f(x))$ exists.

We represent this M/G/1 queue by M/G(N_q)/1. The analysis utilizes the construction of a sample path of the ***virtual wait*** $\{W(t)\}_{t\ge 0}$ by applying the definition of virtual wait *literally*. The virtual wait $W(t)$ at instant t, is defined as the time that a potential (would-be) arrival at t must wait before starting service. The virtual wait is a continuous-state continuous-time process. Its value at any instant t is conditional on an arrival occurring at instant t.

Remark 3.21 In order to validate the LC method immediately after its discovery in 1974, the author applied LC to derive $\{P_0, f(x)\}_{x>0}$ in M/G(N_q)/1 (and in several other queueing models in the literature; and in multiple-server state-dependent queues in his original Ph.D. thesis topic, where solutions had been derived using Lindley recursions and embedded Markov chains [48, 49]). The author included an LC analysis of M/G(N_q)/1 in his Ph.D.

thesis (pp. 206–213 in [11]). The results agreed with the classically-based analysis of M/G(N_q)/1 in C.M. Harris's 1966 Ph.D. thesis [85] and in C.M. Harris's 1967 journal article [86]. (See also Example 5 in [51].)

3.9.1 Sample Path of $\{W(t)\}_{t\geq 0}$

Consider Fig. 3.15. The first customer C_1 arrives, initiates a busy period and receives a service time S_0, since zero customers are left behind in the queue when C_1 starts service. Customer C_2 arrives at t^- during C_1's service time and is allotted a "virtual" service time S, although C_2's actual service time is not known until later at C_2's start-of-service instant. The reason is that the virtual wait may be considered to be the answer to the following question asked a non-countably infinite number of times, i.e., at every instant $t \geq 0$: "***How long would a new arrival at instant*** t ***have to wait before its start-of-service instant?***" The answer to this question forces us to allot service time S to C_2 at its arrival instant. That is, a would-be new arrival *immediately after*

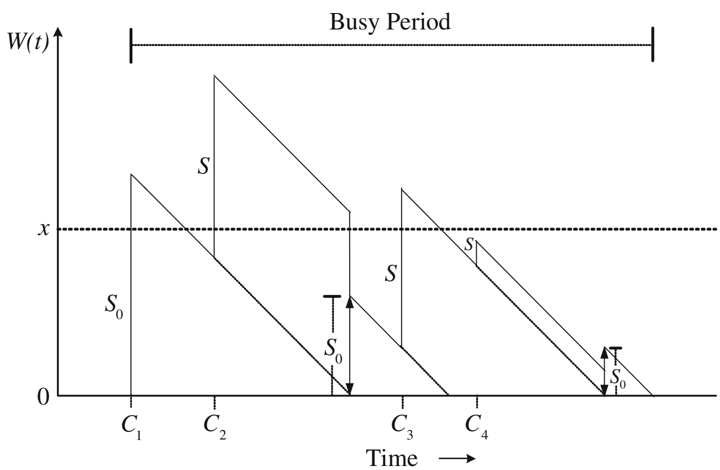


Fig. 3.15 Sample path of $\{W(t)\}_{t\geq 0}$ in M/G(N_q)/1 during a busy period. Shows jumps of size S_0 from level 0 and size S from positive levels. Illustrates a double jumps in the virtual wait $\{W(t)\}_{t\geq 0}$

C_2's arrival, would force C_2 to start service with at least one customer left waiting in the queue. In other words, if C_2 arrives at t^-, the virtual wait at t is the time that a would-be new arrival would have to wait before service.

Suppose, as depicted in Fig. 3.15, *zero* customers arrive during C_2's wait. Then at C_2's start-of-service instant, C_2 must receive an actual service time S_0. This cancels S assigned at C_2's arrival epoch, and substitutes an actual service time S_0. The SP jumps both downward to level 0, and upwards by an amount S_0, at the start-of-service instant of C_2. All SP upward jumps from level 0 are $\underset{dis}{=} S_0$, and all SP upward jumps from positive levels are $\underset{dis}{=} S$.

At instants like the start-of-service instant of C_2 the SP makes a *double* jump (for other examples of double jumps see Examples 2.2 and 2.3 in Sect. 2.3, and Figs. 2.3, 2.5 and 2.6 in Chap. 2).

Next we discuss and derive the steady-state distribution of the **virtual wait** (in contrast to workload).

3.9.2 Integral Equation for PDF of Virtual Wait

Consider a sample path of $\{W(t)\}_{t\geq 0}$ and fix level $x > 0$ in the state space (Fig. 3.15). The SP downcrossing rate of x has *two* components:

1. $f(x)$ by Theorem 1.1,
2. $\lambda\overline{B}(x)\tilde{f}(\lambda)$ due to SP *downward* jumps similar to that at the start-of-service instant of C_2. Here $\tilde{f}(s) := \int_{y=0}^{\infty} e^{-sy} f(y)dy, s > 0$, is the Laplace transform of $f(x)$, and $\tilde{f}(\lambda) = \tilde{f}(s)|_{s=\lambda}$. ($\tilde{f}(s)$ is also denoted by $\mathcal{L}_f(s)$ or other symbols; Sect. 3.4.4 briefly discusses the Laplace transform.)

In component 2, S must be greater than x in order for a downcrossing of x to occur at instants such as the *start of service* of C_2 in Fig. 3.15. The rate of such downcrossings is

$$\begin{aligned}\lambda P(S > x,\ &\text{and zero customers arrive in a waiting time})\\ &= \lambda P(S > x) P(\text{zero customers arrive in a waiting time})\\ &= \lambda P(S > x)\int_{y=0}^{\infty} e^{-\lambda y} f(y)dy = \lambda\overline{B}(x)\tilde{f}(\lambda),\end{aligned}$$

by independence of S and the arrival stream. The total downcrossing rate of x is

$$f(x) + \lambda\overline{B}(x)\tilde{f}(\lambda), x > 0. \tag{3.173}$$

The SP upcrossing rate of x has *three* components:

1. $\lambda\overline{B}_0(x)P_0$, due to arrivals when the system is empty,
2. $\lambda\int_{y=0}^{x}\overline{B}(x-y)f(y)dy$, due to arrivals when the virtual wait is $y\in(0,x)$,
3. $\lambda\overline{B}_0(x)\tilde{f}(\lambda)$, due to arrivals that must wait a positive time and have zero customers arrive behind them during their wait in queue. The total upcrossing rate is

$$\lambda\overline{B}_0(x)P_0+\lambda\int_{y=0}^{x}\overline{B}(x-y)f(y)dy+\lambda\overline{B}_0(x)\tilde{f}(\lambda). \qquad (3.174)$$

Rate balance across level x equates (3.173) and (3.174), leading to the integral equation for $f(\cdot)$,

$$\begin{aligned} f(x)= {} & \lambda\overline{B}_0(x)P_0+\lambda\int_{y=0}^{x}\overline{B}(x-y)f(y)dy \\ & +\lambda\left(\overline{B}_0(x)-\overline{B}(x)\right)\cdot\tilde{f}(\lambda),\ \ x>0. \end{aligned} \qquad (3.175)$$

3.9.3 Exponential Service

Assume $\overline{B}_0(x)=e^{-\mu_0x}$, $\overline{B}(x)=e^{-\mu x}$, $x>0$, and let $\rho_0=\frac{\lambda}{\mu_0}$, $\rho=\frac{\lambda}{\mu}$. Then (3.175) reduces to

$$\begin{aligned} f(x)= {} & \lambda e^{-\mu_0x}P_0+\lambda\int_{y=0}^{x}e^{-\mu(x-y)}f(y)dy \\ & +\lambda\left(e^{-\mu_0x}-e^{-\mu x}\right)\cdot\tilde{f}(\lambda),\ x>0. \end{aligned} \qquad (3.176)$$

Applying differential operator $\langle D+\mu_0\rangle\langle D+\mu\rangle$ to both sides of (3.176) yields the second order differential equation with constant coefficients

$$\langle D^2+(\mu_0+\mu-\lambda)D+\mu_0(\mu-\lambda)\rangle f(x)=0, \qquad (3.177)$$

with general solution

$$f(x)=ae^{-(\mu-\lambda)x}+be^{-\mu_0x},\ x>0, \qquad (3.178)$$

assuming $\mu_0\neq\mu-\lambda\neq0$. From the first term of formula (3.178), a necessary condition for stability is $\lambda<\mu$, since necessarily $f(\infty)=0$.

Using the initial condition $f(0) = \lambda P_0$, substituting $f(y)$ from (3.178) into (3.176), and equating coefficients of common exponents, we obtain the parameters in (3.178) as

$$P_0 = \frac{1-\rho}{1-\rho+\rho_0+\rho_0^2-\rho_0\rho}, \tag{3.179}$$

and

$$a = \frac{-\lambda\rho_0^2 P_0}{\rho_0-\rho-\rho_0\rho}, \ b = \frac{\lambda(1+\rho_0)(\rho_0-\rho)P_0}{\rho_0-\rho-\rho_0\rho}. \tag{3.180}$$

Expected Busy Period

The rate at which the SP makes left-continuous hits of level 0 from above is $f(0) = \lambda P_0$ (Fig. 3.15). Hence the expected busy period is, from (3.82),

$$E(\mathcal{B}) = \frac{1-P_0}{\lambda P_0} = \frac{\rho_0+\rho_0^2-\rho_0\rho}{\lambda(1-\rho)}. \tag{3.181}$$

As a mild check on $E(\mathcal{B})$, set $\rho_0 = \rho = \frac{\lambda}{\mu}$. Then the model reduces to a standard M/M/1 queue. Formula (3.181) reduces to $E(\mathcal{B}) = \frac{1}{\mu-\lambda}$, corresponding to $E(\mathcal{B})$ for the standard M/M/1 queue.

Distribution of Number in System

Applying formula (3.76) and using (3.178) and (3.180) we obtain the steady-state probability of n customers left in the system at *departure* instants,

$$\begin{aligned} d_n &= \int_{x=0}^{\infty} \frac{e^{-\lambda x}(\lambda x)^{n-1}}{(n-1)!} f(x)dx \\ &= \frac{\rho_0 \cdot \left(\rho_0^{n-1} - \rho\rho_0^{n-2} - \rho^n(1+\rho_0)^{n-1}\right)}{(\rho_0-\rho-\rho_0)(1+\rho_0)^{n-1}} P_0, \ n = 1, 2, \ldots, \end{aligned} \tag{3.182}$$

where P_0 $(=d_0)$ is given in (3.179). The values in (3.182) agree with the values of d_n obtained in the earlier works [85, 86].

3.9.4 Workload

In the standard M/G/1, $\{W(t)\}_{t\geq 0}$ is the same as the workload at instant t. In M/G(N_q)/1, the workload is not known at the instant just after an arrival, because the added service time is either S_0 or S depending on *future* arrivals during its wait before service. We can determine the probabilities of these two service times, which allows us to proceed with the analysis.

Consider the **workload process** which we designate $\{W_{wk}(t)\}_{t\geq 0}$. Then $W_{wk}(t)$:= amount of remaining work in the system at time t. Denote the steady-state pdf of $\{W_{wk}(t)\}_{t\to\infty}$ by $\left\{P_{0,wk}, g(x)\right\}_{x>0}$.

In order to construct a sample path, we ask the question immediately after an arrival when the actual workload is y: "**What is the workload just after the arrival**?". The answer logically causes the SP to make a jump of size S with probability $(1-e^{-\lambda y})$ (P(at least 1 arrival in time y)), or size S_0 with probability $e^{-\lambda y}$ (P(no arrivals in time y)). This leads to the upcrossing rate of level x as the right side of (3.183) below. The downcrossing rate of x is $g(x)$. Rate balance across level x gives

$$g(x) = \lambda\overline{B}_0(x)P_{0,wk} + \lambda\int_{y=0}^{x}\overline{B}(x-y)(1-e^{-\lambda y})g(y)dy + \lambda\int_{y=0}^{x}\overline{B}_0(x-y)e^{-\lambda y}g(y)dy. \tag{3.183}$$

If service time $S_0 \underset{dis}{=} \text{Exp}_{\mu_0}$, $S \underset{dis}{=} \text{Exp}_{\mu}$ then $\overline{B}_0(z) = e^{-\mu_0 z}$, $\overline{B}(z) = e^{-\mu z}$, $z > 0$, in (3.183). Applying $\langle D+\mu\rangle\,\langle D+\mu_0\rangle$ to the resulting integral equation yields a second order differential equation with a variable coefficient for $g(x)$

$$\left\langle D^2 + (\mu_0+\mu-\lambda)D + \mu_0(\mu-\lambda) - (\mu-\mu_0)\lambda e^{-\lambda x}\right\rangle g(x) = 0.$$

The solution is given by

$$g(x) = e^{\frac{1}{2}(-\mu-\mu_0+\lambda)x}\left(a\ \text{BesselJ}\left(-\frac{|-\lambda-\mu+\mu_0|}{\lambda}, \frac{2\sqrt{\mu_0-\mu}\,e^{-\frac{1}{2}\lambda x}}{\sqrt{\lambda}}\right) + b\ \text{BesselY}\left(-\frac{|-\lambda-\mu+\mu_0|}{\lambda}, \frac{2\sqrt{\mu_0-\mu}\,e^{-\frac{1}{2}\lambda x}}{\sqrt{\lambda}}\right)\right), \tag{3.184}$$

where a, b are constants to be determined using the initial conditions $g(0) = \lambda P_{0,wk}$, $g'(0) = -(\mu_0-\lambda)\lambda P_{0,wk}$; and BesselJ := first kind ($\nu = 1$), BesselY := second kind ($\nu = 2$), which satisfy Bessel's equation

$$xy'' + xy' + (-\nu^2 + x^2)y = 0, \nu = 1, 2,$$

(see Bessel functions in Maple 17 software). We solve for $P_{0,wk}$ using the normalizing condition $P_{0,wk} + \int_{x=0}^{\infty} g(x)dx = 1$. Due to the Bessel functions in (3.184), it is difficult to get analytic solutions for a, b and $P_{0,wk}$. However, one can obtain numerical solutions when the input parameters λ, μ, and μ_0 have numerical values.

3.10 M/D/1 Queue

The M/D/1 queue is a classical model in queueing theory, first analyzed by A.K. Erlang in 1909 [72].

Here we use LC to derive the cdf of the wait before service, $F(x), x \geq 0$, the mixed pdf $\{P_0, f(x)\}_{x>0}$, where $f(x) = dF(x)/dx, x > 0$, wherever the derivative exists. We also obtain the probability distribution of the number of customers in the system P_n, $n = 0, 1, 2, \ldots$, and related quantities.

The arrival stream is Poisson at rate λ. The service time for each customer is deterministic $S = D > 0$. The traffic intensity is $\rho = \lambda E(S) = \lambda D < 1$, implying stability. Consider the virtual wait $\{W(t)\}_{t\geq 0}$, (Fig. 3.16) and the actual waiting times $\{W_n\}_{n=1,2,\ldots}$. Denote $P(W_n \leq x)$ by $H_n(x), x \geq 0$ and $\lim_{n\to\infty} H_n(x) = H(x), x \geq 0$. Due to Poisson arrivals (e.g., [140])

$$F(x) \equiv \lim_{t\to\infty} P(W(t) \leq x) = \lim_{n\to\infty} P(W_n \leq x) = H(x), x \geq 0.$$

The $\{W(t)\}_{t\geq 0} \leftrightarrow \{W_n\}_{n=0,1,\ldots}$ connection ensures that a study of the virtual wait yields considerable information about both processes.

We define $f(x), x > 0$, to be right continuous, and for notational convenience $f(0) = f(0^+)$ which adds zero probability to $F(0)$. The probability

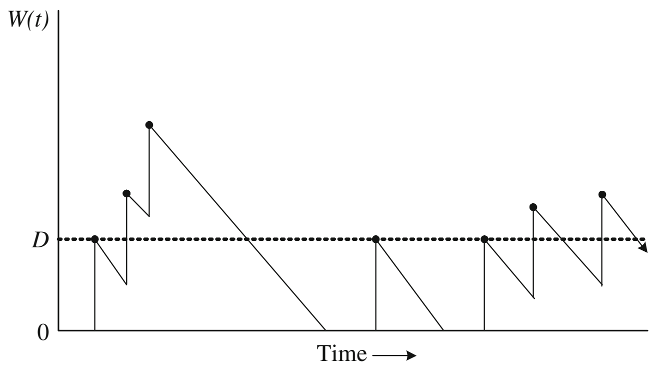


Fig. 3.16 Sample path of $\{W(t)\}_{t\geq 0}$ in M/D/1 queue. Black circles at peaks indicate right continuity

of a zero wait is $P_0 = F(0) = 1 - \rho = 1 - \lambda D$. The mixed pdf $\{P_0, f(x)\}_{x>0}$ is related to $F(x)$ by

$$F(x) = P_0 + \int_{y=0}^{x} f(y)dy, x \geq 0, \qquad F(\infty) = P_0 + \int_{y=0}^{\infty} f(x)dx = 1. \tag{3.185}$$

3.10.1 Properties of PDF and CDF of Wait

We use LC to derive three properties of $\{P_0, f(x)\}_{x>0}$ and a property of $F(x), x \geq 0$.

Proposition 3.9 For the M/D/1 queue: (1) $\{P_0, f(x)\}_{x>0}$ has exactly one atom, which is at $x = 0$; (2) $f(x)$ has a downward jump discontinuity of size λP_0 at $x = D$; (3) $f(x)$ is continuous for all $x > 0, x \neq D$.

Proof Consider sample paths of $\{W(t)\}_{t\geq 0}$ in Figs. 3.16 and 3.17.

1. State $\{0\}$ is an atom since a sample path spends a positive proportion of time in $\{0\}$ $(a.s.)$, namely $P_0 = 1 - \lambda D > 0$ (from (3.62) in Sect. 3.4). The state space $\boldsymbol{S} = [0, \infty)$ has no other atoms, since the proportion of time the SP spends in each state $x > 0$, is 0.
2. Consider state-space levels D and $D - \varepsilon$, $0 < \varepsilon < D$. Fix time $t > 0$. $\mathcal{T}_t^b(D)$ is the number of tangents of level D from below during $(0, t)$

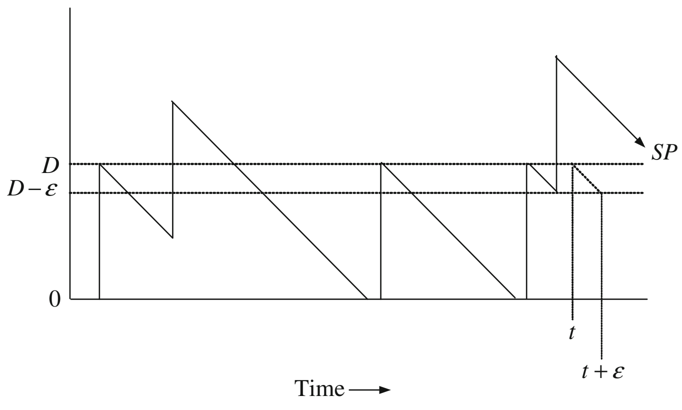


Fig. 3.17 Sample path of $\{W(t)\}_{t\geq 0}$ in M/D/1 showing levels D and $D - \varepsilon$ and instants $t, t + \varepsilon$. See Proposition 3.7, Proof, Part (2)

(see Fig. 2.13 (row 2, column 2) in Sect. 2.5; and Examples 2.4 and 2.5 in Sect. 2.4.10). We have

$$\mathcal{D}_{t+\varepsilon}(D-\varepsilon) = \sum_{j=1}^{\mathcal{D}_t(D)+\mathcal{T}_t^b(D)} I_j(D,\varepsilon); \tag{3.186}$$

where $I_j(\chi,\varepsilon) = 1$ if the jth downcrossing *or* tangent of level χ from below, is followed by a downcrossing of level $\chi - \varepsilon$ exactly ε time units later, and $I_j(\chi,\varepsilon) = 0$ otherwise. Due to the memoryless property $P(I_j(\chi,\varepsilon) = 1) = e^{-\lambda\varepsilon}, \chi > 0$. Set $\chi = D$; $I_j(D,\varepsilon)$ is independent of $\mathcal{D}_t(D) + \mathcal{T}_t^b(D)$, and $E(I_j(D,\varepsilon)) = e^{-\lambda\varepsilon}$, $j = 1, 2, \ldots$. Taking expected values on both sides of (3.186) gives

$$E(\mathcal{D}_{t+\varepsilon}(D-\varepsilon)) = E(\mathcal{D}_t(D) + \mathcal{T}_t^b(D))e^{-\lambda\varepsilon}. \tag{3.187}$$

By Corollary 3.2 of Theorem 3.3 in Sect. 3.2.5

$$\lim_{t\to\infty} \frac{E(\mathcal{D}_t(D))}{t} = f(D)\, and \, \lim_{t\to\infty} \frac{E(\mathcal{D}_t(D-\varepsilon))}{t} = f(D-\varepsilon).$$

Also, $\lim_{t\to\infty} \frac{E\left(\mathcal{T}_t^b(D)\right)}{t} = \lambda P_0$, due to the one-to-one correspondence between zero-wait arrivals and tangents of level D from below. Dividing both sides of (3.187) by t, writing $\frac{1}{t} = \frac{1}{t+\varepsilon} \cdot \frac{t+\varepsilon}{t}$ on the left side, and letting $t \to \infty$ gives

$$f(D-\varepsilon) = (f(D) + \lambda P_0)e^{-\lambda\varepsilon}.$$

Letting $\varepsilon \downarrow 0$ yields, since $f(D) = f(D^+)$,

$$f(\mathcal{D}^-) - f(D^+) = \lambda\mathcal{P}_0. \tag{3.188}$$

3. **Case** $x > D$. With probability 1, sample paths are not tangent to level x due to continuous inter-arrival times ($\underset{dis}{=}\text{Exp}_\lambda$). Let ε be $< (x - D)$ and small. Then

$$\mathcal{D}_{t+\varepsilon}(x-\varepsilon) = \sum_{j=1}^{\mathcal{D}_t(x)} I_j(x,\varepsilon) + \sum_{j=1}^{A_t(\varepsilon)} \nu_j(\varepsilon), \tag{3.189}$$

where $\nu_j(\varepsilon) = 1$ if an arrival occurs when $W(t) = \xi \in (x-\varepsilon-D, x-D)$ causing a jump ending at $\xi + D \in (x-\varepsilon, x)$. Note that $P(\nu_j(\varepsilon) = 1)$ $= \int_{x-\varepsilon-D}^{x-D} f(y)dy = \varepsilon f(\xi^*),\ \xi^* \in (x-D-\varepsilon, x-D)$. But $f(\xi^*) <$

λ, (see Proposition 3.7 in Sect. 3.4.22). So $P(\nu_j(\varepsilon) = 1) < \varepsilon\lambda$. Thus $E(\nu_j(\varepsilon)) < \varepsilon\lambda$, which tends to 0 as $\varepsilon \downarrow 0$.
Taking expected values in (3.189) and dividing both sides by t, gives

$$E\left(\mathcal{D}_{t+\varepsilon}(x-\varepsilon)\right) = E\left(\mathcal{D}_t(x)\right) \cdot e^{-\lambda\varepsilon} + E(A_t) \cdot E(\nu_j(\varepsilon))$$
$$\lim_{t\to\infty} \frac{E\left(\mathcal{D}_{t+\varepsilon}(x-\varepsilon)\right)}{t} = \lim_{t\to\infty} \frac{E\left(\mathcal{D}_t(x)\right) e^{-\lambda\varepsilon}}{t} + \lim_{t\to\infty} \frac{\lambda t E(\nu_j(\varepsilon))}{t}$$
$$f(x-\varepsilon) = f(x)e^{-\lambda\varepsilon} + \lambda E(\nu_j(\varepsilon)).$$

Letting $\varepsilon \downarrow 0$ gives $f(x^-) = f(x)$.

3. **Case** $0 < x < D$. If $0 < x < D$ then, similar to Eq. (3.186) with D replaced by x, and omitting $\mathcal{T}_t^b(x)$, we have

$$\mathcal{D}_{t+\varepsilon}(x-\varepsilon) = \sum_{j=1}^{\mathcal{D}_t(x)} I_j(D, \varepsilon).$$

Taking expected values on both sires, dividing by t, letting $t \to \infty$, then letting $\varepsilon \downarrow 0$, gives $f(x^-) = f(x)$. ■

Proposition 3.10 (1) $F(x), x \geq 0$, has a jump discontinuity at $x = 0$ of size P_0; (2) $F(x)$ is continuous for all $x > 0$.

Proof (1) $F(x)$ has a discontinuity at $x = 0$, since 0 is an atom having probability $F(0) = P_0$. (2) Fix $x > 0$ in the state space. Then x is not an atom (Proposition 3.9 Part (1)); therefore $P(\{x\}) = 0$. That is, x is not a point of increase in probability. Thus x is a point of continuity of $F(\cdot)$. ■

3.10.2 Integral Equation for PDF of Wait

Applying the alternative form of the basic LC integral equation (3.44) with $B(x-y) = 0$ if $x - y < D$ and $B(x-y) = 1$ if $x - y \geq D$, we immediately write an equation for $f(x)$ in terms of $F(\cdot)$, which is a differential equation for the cdf $F(x)$ since $f(x) = F'(x)$,

$$f(x) = \lambda F(x) - \lambda F(x-D), x > 0. \quad (3.190)$$

To explain (3.190) in terms of LC, consider a sample path of $\{W(t)\}_{t\geq 0}$ (Fig. 3.16). In (3.190) the left side $f(x)$ is the SP downcrossing rate of level x. SP jumps occur at rate λ, all upward of size D. On the right side of (3.190), the first term $\lambda F(x)$ is the rate of SP jumps that start in state set $[0, x]$. The second term, $-\lambda F(x-D)$ subtracts off the rate of jumps that start in $[0, x]$ and end below x, because jumps starting below $x-D$ cannot upcross x. Thus the right side is the upcrossing rate of x. Rate balance across level x then yields (3.190).

Remark 3.22 The properties in Proposition 3.9, and Eq. (3.190) are readily inferred intuitively upon considering a sample path (Fig. 3.16), and applying LC interpretations of transition rates. Such intuitive insights often lead to formal proofs as in Proposition 3.9.

3.10.3 Analytic Solution for CDF and PDF of Wait

CDF of Wait We give the solution of (3.190), for completeness. For $x \in [0, D)$, $F(x-D) \equiv 0$; thus $f(x) = \lambda F(x)$, or

$$F^{'}(x) - \lambda F(x) = 0,$$

having solution

$$F(x) = A_0 e^{\lambda x}, x \in [0, D)$$

where A_0 is a constant. Letting $x \downarrow 0$, gives the constant $A_0 = P_0 = 1 - \rho$. Thus

$$F(x) = P_0 e^{\lambda x}, x \in [0, D).$$

For $x \in [D, 2D)$, (3.190) is equivalent to

$$F^{'}(x) - \lambda F(x) = -\lambda P_0 e^{\lambda(x-D)}, x \in [D, 2D).$$

Multiplying both sides by the integrating factor $e^{-\lambda(x-D)}$ and then integrating both sides over $[D, x)$ yields the solution up to a constant

$$F(x) = -P_0 \lambda(x-D) e^{\lambda(x-D)} + A_1 e^{\lambda(x-D)}, x \in [D, 2D).$$

The constant A_1 is determined from the *continuity* of $F(x), x > 0$ (Proposition 3.10). Thus $F(D^-) = F(D)$, or $A_1 = P_0 e^{\lambda D}$ resulting in the solution

$$F(x) = P_0 \left(-\lambda(x-D) e^{\lambda(x-D)} + e^{\lambda x} \right), x \in [D, 2D).$$

Mathematical induction on (3.190) yields the classical formula for the cdf of wait originally derived in [72],

$$F(x) = P_0 \sum_{i=0}^{m} (-\lambda)^i \frac{(x - iD)^i}{i!} e^{\lambda(x-iD)}, \; x \in [m, (m+1)D), \; m = 0, 1, 2, \ldots \tag{3.191}$$

An alternative form of (3.191) is (e.g., p. 385 in [84]),

$$F(x) = P_0 \sum_{i=0}^{\lfloor x/D \rfloor} (-\lambda)^i \frac{(x - iD)^i}{i!} e^{\lambda(x-iD)}, \; x \geq 0, \tag{3.192}$$

where $\lfloor \alpha \rfloor$:= greatest integer $\leq \alpha$.

PDF of Wait The pdf $f(x)$ may be obtained by differentiating $F(x)$ with respect to x. More simply, we obtain $f(x)$ by substituting (3.191) into (3.190) giving

$$f(x) = \lambda P_0 e^{\lambda x}, 0 < x < D$$

and for $x \in [mD, (m+1)D)$, $m = 0, 1, 2, \ldots$,

$$\begin{aligned} f(x) &= \lambda P_0 \sum_{i=0}^{m} (-\lambda)^i \frac{(x-iD)^i}{i!} e^{\lambda(x-iD)} - \sum_{i=0}^{m-1} (-\lambda)^i \frac{(x-(i+1)D)^i}{i!} e^{\lambda(x-(i+1)D)} \\ &= \lambda P_0 (-\lambda)^m \frac{(x-mD)^m}{m!} e^{\lambda(x-mD)} \\ &\quad + \sum_{i=0}^{m-1} \frac{(-\lambda)^i}{i!} [(x - iD)^i e^{\lambda(x-iD)} - (x - (i+1)D)^i e^{\lambda(x-(i+1)D)}]. \end{aligned} \tag{3.193}$$

The pdf $f(x)$ in (3.193) has a discontinuity at $x = D$ (Proposition 3.9 Part (2)). That is $f(D^-) = \lambda P_0 e^{\lambda D}$, and $f(D^-) - f(D) = \lambda P_0$, illustrating that $f(x)$ has a downward jump of size λP_0 at $x = D$. Moreover $f(x)$ is continuous for all other $x > 0$ (see Fig. 3.18). In Fig. 3.18 there is a concave wave in $f(x)$ for $x \in [D, 2D)$, the waviness dampens to the right of $x = 2D$. The cdf $F(x)$,

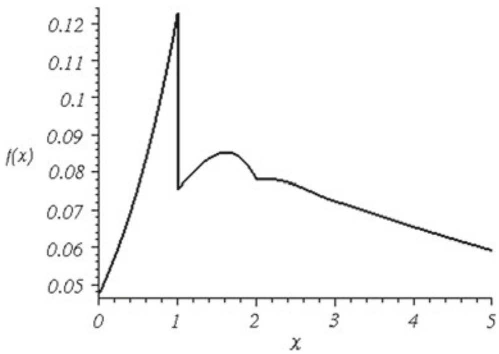


Fig. 3.18 PDF $f(x)$ of wait in M/D/1: $\lambda = 0.95$, $D = 1$, $\rho = 0.95$ (high traffic). Shows discontinuity and downward jump of size λP_0 at $x = D$; and extreme waviness in right neighborhood $[D, 2D)$

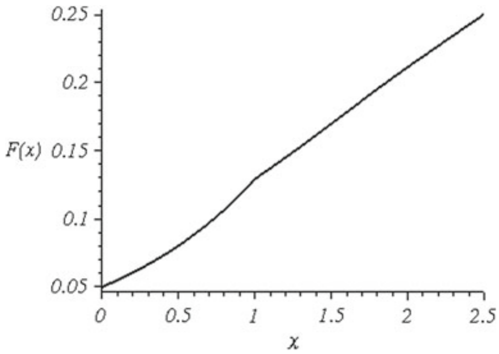


Fig. 3.19 CDF $F(x)$ of wait in M/D/1: $\lambda = 0.95$, $D = 1$. Shows continuity of $F(x), x > 0$; and decrease in slope of $F(x)$ at $x = D$

for the same example, is given in formula (3.191) and plotted in Fig. 3.19, where the continuity of $F(x)$, $x > 0$, and discontinuity of $\frac{d}{dx}F(x)|_{x=D}$ are evident.

Remark 3.23 An LC examination of a typical sample path of $\{W(t)\}_{t\geq 0}$ suggests an isomorphism: {sample-path properties} $\leftrightarrow$ {analytical properties of $f(x)$ and $F(x)$}.

3.10.4 Probability Distribution of Number in System

Let N be the number of customers in the system at an arbitrary time point and let W_q (≥ 0) be the wait before service, in the steady-state. Let $P_n := P(N = n)$. Consider a_n, d_n, the probabilities that the number of customers in the system is n just before an arrival, and just after a departure, respectively. Due to Poisson arrivals, $a_n = P_n = d_n$, $n = 0, 1, 2, \ldots$. Arrivals "see" n customers in the system iff $W_q \geq 0$ and $W_q \in ((n-1)D, nD]$, $n = 0, 1, 2, \ldots$. Thus

$$a_n = F(nD) - F((n-1)D) = P_n = d_n, n = 0, 1, 2, \ldots.$$

From (3.191)

$$\begin{aligned}
P_0 &= F(0) - F(-D) = F(0) = P_0 \\
P_1 &= F(D) - F(0) = P_0 e^{\lambda D} - P_0 = P_0(e^{\lambda D} - 1) \\
P_2 &= F(2D) - F(D) = P_0 e^{\lambda D}(-\lambda D + e^{\lambda D} - 1) \\
&\ldots \\
P_n &= F(nD) - F((n-1)\,D), n = 0, 1, 2, \ldots
\end{aligned} \tag{3.194}$$

The cdf of N is

$$P(N \leq n) = \sum_{i=0}^{n} P_i = F(nD), \ \ n = 0, 1, 2, \ldots, \tag{3.195}$$

where $F(nD)$ is computed using (3.191) or (3.192).

3.11 M/Discrete/1 Queue Aka M/D$_n$/1

We look at the M/D$_n$/1 queue, which is an M/G/1 queue with multiple Poisson inputs where the service times are discrete quantities $\{D_n\}_{n=1,2,\ldots}$ (also called an M/Discrete/1 queue). We study the wait before service W_q, and derive analytical properties of its cdf $F(x)$, $x \geq 0$, and pdf $\{P_0, f(x)\}_{x>0}$, where $f(x) = \frac{d}{dx}F(x)$, $x > 0$, wherever the derivative exists. Consider a typical sample path of the virtual wait $\{W(t)\}_{t\geq 0}$ (Fig. 3.20).

Customers arrive in a Poisson stream at rate λ at a single server. For each arrival,

$$P(S = D_i) = p_i, \quad \sum_{i=1}^{N} p_i = 1,$$

where $D_i > 0$, $i = 1, \ldots, N$, and N is a positive integer. Then $E(S) = \sum_{i=1}^{N} p_i D_i$. Without loss of generality, reorder the D_is if necessary, such that

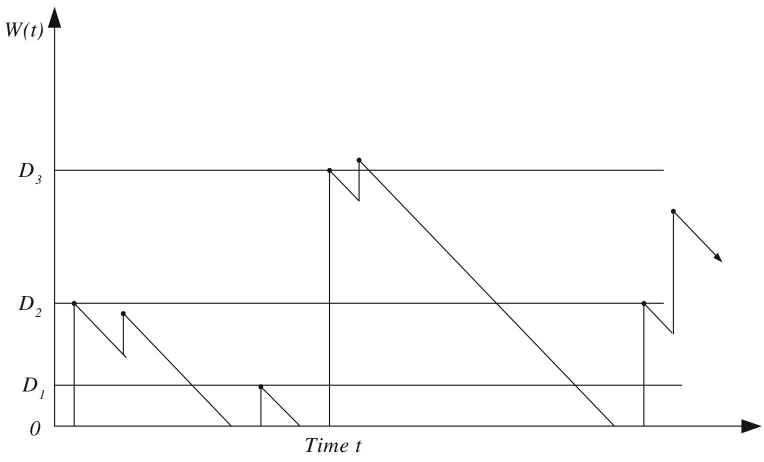


Fig. 3.20 Sample path of $\{W(t)\}_{t\geq 0}$ in $M/\{D_n\}/1$ queue with $N = 3$ service levels

$$0 \equiv D_0 < D_1 < \cdots < D_N < D_{N+1} \equiv \infty.$$

Customers that receive a service time D_i arrive at rate λp_i. The traffic intensity is $\rho = \lambda E(S) < 1$ (stability). Due to Poisson arrivals (e.g., [140]),

$$\lim_{t\to\infty} P(W(t) \leq x) = \lim_{n\to\infty} P(W_n \leq x),$$

where $\{W_n\}_{n=1,2,\ldots}$ is the process of actual (arrival-point) waits.

We define $f(x), x \geq 0$, as right continuous. The probability of a zero wait is

$$P_0 \equiv F(0) = 1 - \rho = 1 - \lambda \sum_{i=1}^{N} D_i p_i.$$

The cdf and pdf are related by

$$F(x) = P_0 + \int_{y=0}^{x} f(y)dy,\ F(\infty) = P_0 + \int_{y=0}^{\infty} f(x)dx = 1. \quad (3.196)$$

Remark 3.24 The arrival stream may be viewed in two distinct ways:

1. A homogeneous class of customers arrives at rate λ. For each arrival the service time S has a mixture probability distribution with components D_i and mixture probabilities (weights) p_i, $\sum_{i=1}^{N} p_i = 1$.
2. N classes of customers arrive in independent Poisson processes at rates $\lambda_i \equiv \lambda p_i$, $\sum_{i=1}^{N} p_i = 1$, and receive independent service times $D_i, i = 1, \ldots, N$, respectively. This way shows that M/D$_n$/1 is an M/G/1 queue with multiple Poisson inputs.

These two viewpoints yield the same steady-state distribution of wait, as reflected in the two equivalent forms for the traffic intensity $\rho = \lambda \left(\sum_{i=1}^{N} p_i D_i\right) = \sum_{i=1}^{N} \lambda_i D_i$, where $\lambda_i = \lambda p_i$ (see Sect. 5.3.4, p. 319 in [125]).

Remark 3.25 A similar analysis of the M/D$_n$/1 queue applies if $N = \infty$.

3.11.1 Properties of PDF and CDF of Wait

The steady-state distribution of wait has analytical properties given in Proposition 3.11.

Proposition 3.11 In the M/D$_n$ /1 queue, $\{P_0; f(x), x > 0\}$: (1) has exactly one atom which is at $x = 0$ (state $\{0\}$ is an atom); (2) has exactly $\boldsymbol{N}$ down-

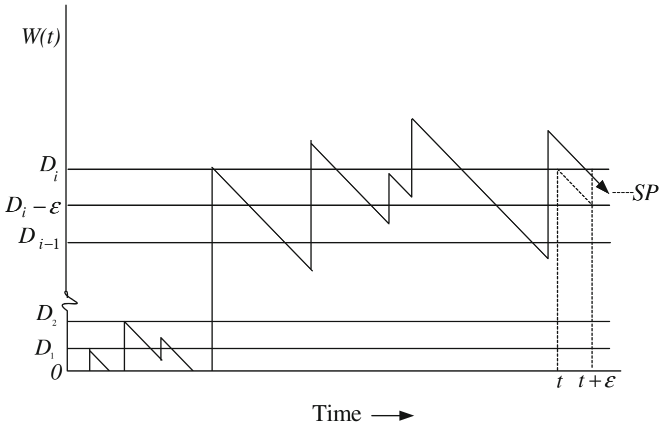


Fig. 3.21 Sample path in M/{D$_n$}/1 showing levels D_i, $D_i - \varepsilon$ and instants $t, t+\varepsilon$. See Proposition 3.11, Proof, Part (2)

ward jump discontinuities of sizes $\lambda p_i P_0$ at $x = D_i$, $i = 1, \ldots, N$; (3) is continuous for all $x > 0$, $x \neq D_i$, $i = 1, \ldots, N$.

Proof Check a typical sample path of $\{W(t)\}_{t\geq 0}$ (Fig. 3.20).

1. State {0} is an atom since a sample path spends a positive proportion of time in {0} (a.s.), namely $P_0 = 1 - \lambda \sum_{i=1}^{N} p_i D_i$. Each sojourn time in $\{0\} \underset{dis}{=} \text{Exp}_\lambda$. There are no other atoms in the state space, since the proportion of time that a sample path spends in each state $x > 0$, is 0.
2. The proof is similar to the proof of Part (2) in Proposition 3.9, Sect. 3.10, upon replacing D, $D - \varepsilon$ by D_i, $D_i - \varepsilon$; λ by λp_i; and where $\varepsilon \in (0, D_i - D_{i-1})$, $i = 1, \ldots, N$; (as in Fig. 3.21). Using similar reasoning as in Proposition 3.9 we obtain

$$f(D_i - \varepsilon) = (f(\mathcal{D}_i) + \lambda p_i P_0)e^{-\lambda\varepsilon}, i = 1, \ldots, N$$

where $\lambda p_i P_0$ is the rate at which the SP makes a tangent to level D_i from below, which is the same as the arrival rate of type-i customers when the system is empty (rate of SP jumps of size D_i from level 0). Letting $\varepsilon \downarrow 0$ results in

$$f(\mathcal{D}_i^-) - f(\mathcal{D}_i) = \lambda p_i P_0, i = 1, \ldots, N.$$

verifying downward jumps at D_i of size $\lambda p_i P_0$, $i = 1, \ldots, N$.

3. The proof is similar to the proof of Part (3) in Proposition 3.9. We thus obtain for $x > 0, x \notin \{D_i\}_{i=1,\ldots,N}$

$$f(x-\varepsilon) = f(x) \cdot e^{-\lambda\varepsilon}.$$

Letting $\varepsilon \downarrow 0$ yields $f(x^-) = f(x)$ so that x is a point of continuity. ■

Remark 3.26 From Part (2) of Proposition 3.11, the sum of the downward jumps at points of discontinuity of the pdf $f(x)$ is $\lambda P_0 \sum_{i=1}^N p_i = \lambda P_0$. This formula is the same as the size of the single downward jump in the pdf of wait in the M/D/1 model, independent of N.

Proposition 3.12 In the M/$\{D_n\}$/1 queue the steady-state cdf of wait $F(x)$, $x \geq 0$, has a single jump discontinuity at $x = 0$ of size P_0, and is continuous for all $x > 0$.

Proof $F(\cdot)$ has a jump discontinuity at level 0, since $\{0\}$ is an atom having probability $P_0 = F(0)$ (Proposition 3.11, Part (2)). Fix $x > 0$ in the state space. Then x is not an atom (Proposition 3.11, Part (3)). Hence x has probability 0. Thus x is a point of continuity of $F(\cdot)$. ■

3.11.2 *Expected Busy Period*

From (3.80) the expected busy period is

$$E(\mathcal{B}) = \frac{E(S)}{1-\lambda E(S)} = \frac{1-P_0}{\lambda P_0} = \frac{\sum_{i=1}^N D_i p_i}{1-\lambda\sum_{i=1}^N p_i D_i}.$$

Another way to compute P_0 is, letting $\mathcal{I}$ denote an idle period,

$$P_0 = \frac{E(\mathcal{I})}{E(\mathcal{I})+E(\mathcal{B})} = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda} + \frac{\sum_{i=1}^N p_i D_i}{1-\lambda\sum_{i=1}^N p_i D_i}} = 1 - \lambda\sum_{i=1}^N p_i D_i.$$

3.11.3 *Integral Equation for PDF of Wait*

The alternative form of the LC integral equation for M/G/1 (3.44) leads immediately to an "integral" equation for the pdf $f(x)$ (*differential equation* for cdf $F(x)$),

$$f(x) = \lambda F(x) - \lambda \sum_{i=1}^{N} p_i F(x - D_i)$$
$$= \lambda F(x) - \sum_{i=1}^{N} \lambda_i F(x - D_i), x > 0. \quad (3.197)$$

To explain (3.197) consider a virtual-wait sample path (Fig. 3.20). In (3.197), the left side $f(x)$ is the downcrossing rate of level x. SP jumps occur at rate $\lambda = \sum_{1=1}^{N} \lambda_i$; having size D_i with probability $p_i = \lambda_i/\lambda$. On the right side, the first term $\lambda F(x)$ is the rate at which SP jumps start in state-space set $[0, x]$. The second term, $-\lambda \sum_{i=1}^{N} F(x - D_i) p_i$, subtracts off the rate of those jumps which start in state set $[0, x]$ and end *below* level x. SP jumps of size D_i that start below $x - D_i$, cannot upcross level x. Thus the right side is the sample-path upcrossing rate of x. Rate balance across level x gives (3.197).

3.11.4 Solution for CDF of Wait

Differential equation (3.197) for $F(x)$ is solvable. However the form of $F(x)$ differs in the state-space intervals

$$[0, D_1),\ [D_1, 2D_1),$$
$$\ldots, [j_{11} D_1, D_2),\ [D_2, (j_{11}+1)D_1),\ [(j_{11}+1)D_1, (j_{11}+2)D_1),$$

etc., where $j_{11} = \left\lfloor \frac{D_2}{D_1} \right\rfloor$ (greatest integer $\le \frac{D_2}{D_1}$). At D_3 in the state space, we need to consider $j_{12} = \left\lfloor \frac{D_3}{D_1} \right\rfloor$ and $j_{22} = \left\lfloor \frac{D_3}{D_2} \right\rfloor$, etc. This makes the solution procedure complex. We must keep track of the positions in the state space of the break points where the functional form changes, by considering the relative sizes of $D_1, D_2, \ldots, D_N$. Section 3.11.5 discusses another approach to solve for $F(x)$, $x \ge 0$.

3.11.5 Alternative Approach for CDF of Wait

We can obtain a solution for $F(x)$, $x \ge 0$, using a "specialized" M/D$_n$/1 queue. Assume, without loss of *computational accuracy*, that all D_is are rational numbers. (Rationals can approximate irrational numbers arbitrarily closely.). Let

$$D_1 = k_1 D,\, D_2 = k_2 D, \ldots, D_N = k_N D,$$

where $D = \gcd\{D_1, \ldots, D_N\}$ (gcd denotes greatest common divisor); and $0 < k_1 < k_2 < \cdots < k_N$ are positive integers.

Consider an M/D$_n$/1 queue where $D_i = iD$, $i = 1, \ldots, N$. We call this model an M/$\{iD\}$/1 queue. It is somewhat easier to obtain an analytical solution for the cdf and pdf of wait in M/$\{iD\}$/1 than in M/D$_n$/1. Once a solution for M/$\{iD\}$/1 is obtained, then adjust the *arrival rates* for customers that get service times $k_i D$ $(=D_i)$ so that they correspond to those of the original M/D$_n$/1 queue. The arrival rates for intermediate service time values $\{iD | iD \neq D_i, i = 1, \ldots, N\}$ are set to 0 in that solution. The resulting cdf for M/$\{iD\}$/1 is equal to $F(x)$, $x \geq 0$, for the original M/D$_n$ /1 model (i.e., the solution of (3.197)).

Thus M/$\{iD\}$/1 (where $D = \gcd\{D_1, \ldots, D_N\}$) may be considered as equivalent M/D$_n$/1. Also, it is more amenable analytically and computationally. We next examine the M/$\{iD\}$/1 queue.

3.12 M/{*iD*}/1 Queue

We analyze the M/$\{iD\}$/1 queue, mindful of its close relationship to M/D$_n$ /1 (Sect. 3.11.5).

In M/$\{iD\}$/1 there are N types of arrivals at Poisson rates λ_i, $i = 1, \ldots, N$, where N is a positive integer. Customers of type i receive a service time $S = iD$, where $D > 0$ is fixed. Equivalently, customers arrive at Poisson rate λ and get $S = iD$ with probability p_i, $\sum_{i=1}^{N} p_i = 1$. Thus $\lambda p_i \equiv \lambda_i$. The expected service time is $E(S) = \sum_{i=1}^{N} iDp_i$. Assume $\lambda E(S) < 1$ (stability). Let P_0 denote the steady-state probability that the system is empty. Then

$$P_0 = 1 - \lambda E(S) = 1 - \lambda \sum_{i=1}^{N} iDp_i = 1 - \sum_{i=1}^{N} iD\lambda_i. \tag{3.198}$$

The M/D/1 queue is a special case of M/$\{iD\}$/1 with $N = 1$. The M/$\{iD\}$/1 queue is a special case of M/$\{$D$_n\}$/1, with $D_n = k_n D$, $n = 1, \ldots, N$, k_n is an integer in the set $\{1, \ldots, N\}$, and $D = \gcd\{D_1, \ldots, D_N\}$ (gcd := greatest common divisor). Paradoxically, M/$\{iD\}$/1 may also be considered as a *generalization* of M/D$_n$ /1! (Sect. 3.11.5).

3.12.1 Integral Equation for CDF of Wait

Let W_q denote the wait before service in the steady state, having cdf $F(x) \equiv P(W_q \le x), x \ge 0$ and pdf $f(x) = \frac{d}{dx}F(x), x > 0$, wherever the derivative exists. We apply the 'alternative LC' Eq. (3.43) (see also Eq. (3.190) for the M/D/1 queue) relating $f(x)$ and $F(x)$ of wait to obtain

$$f(x) = \lambda F(x) - \lambda \sum_{i=1}^{N} F(x - iD)p_i = \lambda F(x) - \sum_{i=1}^{N} \lambda_i F(x - iD), x > 0. \tag{3.199}$$

Consider the virtual wait process $\{W(t)\}_{t \ge 0}$ (similar to Fig. 3.20). To explain (3.199) the left side is the sample path downcrossing rate of x. On the right side, the term $\lambda F(x)$ is the rate of jumps that start at levels in $[0, x]$. The term $-\sum_{i=1}^{N} \lambda_i F(x - iD)$ subtracts off the rate of jumps that start at levels in $[0, x]$ and *end below* x. For example, $\lambda_i F(x - iD)$ is the rate of type-i jumps of size iD that do not upcross x, since they start below $x - iD$. Hence, the right side is the upcrossing rate of x. Equation (3.199) results by rate balance across level x.

3.12.2 Recursion for CDF of Wait

We now outline a procedure to solve (3.199) recursively for $F(x)$, $x \in [mD, (m+1)D), m = 0, 1, 2, \ldots$. Let

$$F(x) \equiv F_m(x), \;\; f(x) \equiv f_m(x), \;\; x \in [mD, (m+1)D), m = 0, 1, 2, \ldots$$

and $F_{-k}(x) \equiv 0$ if k is a positive integer (see Figs. 3.22 and 3.23). Then write (3.199) as

$$f_m(x) = \lambda F_m(x) - \sum_{i=1}^{N} \lambda_i F_{m-i}(x - iD),$$
$$x \in [mD, (m+1)D), \;\; m = 0, 1, 2, \ldots \tag{3.200}$$

First, let us consider the state-space interval $[0, D)$. The cdf $F(x - D) = 0$ if $x - D < 0$. For $x \in [0, D)$, Eq. (3.200) reduces to

$$f_0(x) = \lambda F_0(x), x \in [0, D),$$

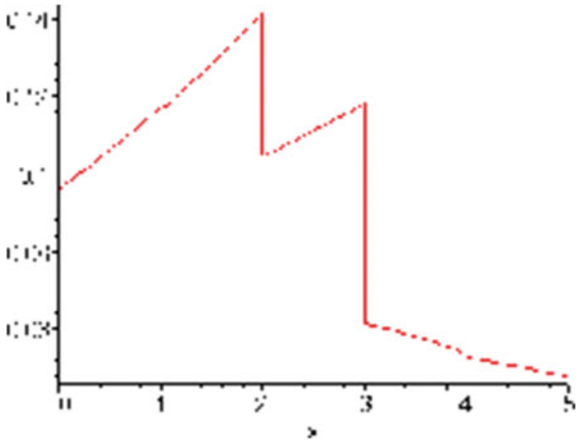


Fig. 3.22 PDF of wait in $M/\{iD\}/1$ queue: $D = 1.0$, four arrival types ($N = 4$), $\lambda = 0.2$, $p_1 = p_4 = 0.01$, $p_2 = 0.39$, $p_3 = .59$. Downward jumps at $x = 1, 2, 3, 4$

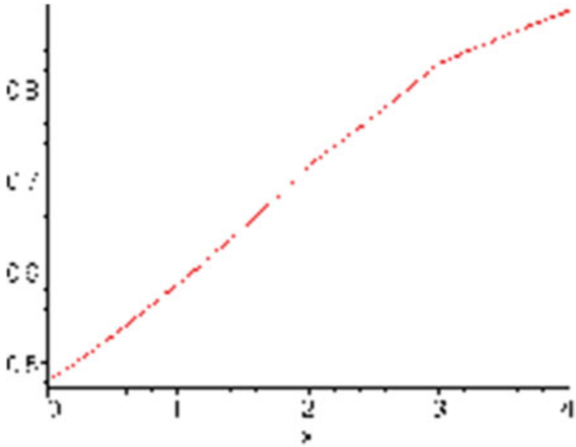


Fig. 3.23 CDF of wait in $M/\{iD\}/1$ queue. $D = 1.0$, $N = 4$, $\lambda = 0.2$, $p_1 = p_4 = 0.01$, $p_2 = 0.39$, $p_3 = 0.59$. The slope decreases abruptly at $x = 1, 2, 3, 4$

or differential equation

$$F_0'(x) = \lambda F_0(x), x \in (0, D),$$

with solution, using the initial condition $F(0) = P_0$,

$$F_0(x) = P_0 e^{\lambda x}, x \in [0, D).$$

Next, on interval $[D, 2D)$, Eq. (3.200) reduces to

$$f_1(x) = \lambda F_1(x) - F_0(x - D)\lambda_1, x \in [D, 2D),$$
$$\text{or } F_1'(x) = \lambda F_1(x) - P_0 e^{\lambda(x-D)}\lambda_1, x \in [D, 2D).$$

The last equation is a differential equation in $F_1(x)$, which is readily solved up to a constant, by using continuity $F_0(D^-) = F_1(D)$, resulting in

$$F_1(x) = P_0\left(e^{\lambda x} + \lambda_1(D-x)e^{-\lambda(D-x)}\right), x \in [D, 2D).$$

Imagine extending the domain of $F_0(x)$ to $[0, \infty)$. The last equation can then be written as

$$F_1(x) = F_0(x) + P_0\lambda_1(D-x)e^{-\lambda(D-x)}, x \in [D, 2D).$$

Similarly we obtain recursively

$$F_2(x), x \in [2D, 3D), \qquad F_3(x), x \in [3D, 4D), \qquad F_4(x), x \in [4D, 5D),$$

where we extend the domains of $F_m(x)$ to $[m, \infty)$, $m = 0, 1, \ldots$. The recursive formulas in (3.201) below summarize the values of $F(x)$ on state-space interval $[0, 5D)$ by specifying the corresponding functions on intervals $[0, D)$, …, $[4D, 5D)$, respectively.

$$\begin{aligned}
F_0(x) =&\ P_0 e^{\lambda x},\\
F_1(x) =&\ F_0(x) + P_0\lambda_1(D-x)e^{-\lambda(D-x)},\\
F_2(x) =&\ F_1(x) + P_0\left(\lambda_2(2D-x) + \tfrac{\lambda_1^2(2D-x)^2}{2!}\right)e^{-\lambda(2D-x)},\\
F_3(x) =&\ F_2(x) + P_0\Big(\lambda_3(3D-x) + \lambda_2\lambda_1(3D-x)^2\\
&+\tfrac{\lambda_1^3(3D-x)^3}{3!}\Big)e^{-\lambda(3D-x)},\\
F_4(x) =&\ F_3(x) + P_0\Big(\lambda_4(4D-x) + \lambda_3\lambda_1(4D-x)^2 + \tfrac{\lambda_2^2(4D-x)^2}{2!}\\
&+\tfrac{\lambda_2\lambda_1^2(4D-x)^3}{3!} + \tfrac{\lambda_1^4(4D-x)^4}{4!}\Big)e^{-\lambda(4D-x)}.
\end{aligned} \tag{3.201}$$

The recursion (3.201) can be continued indefinitely. The general solution appeared in an article in 2005 by J.F. Shortle and P.H. Brill (see [128]), and is stated below in Sect. 3.12.3.

3.12.3 Solution for CDF and PDF of Wait

Using mathematical induction, it can be shown that an analytical solution of the *indefinitely extended* recursion (3.201) for the cdf of W_q is

$$F_m(x) = P_0 \sum_{i=0}^{m} e^{-\lambda(iD-x)} \sum_{\mathcal{L}\in\mathcal{P}(i)} \frac{(iD-x)^{|\mathcal{L}|}}{H(\mathcal{L})} \prod_{j\in\mathcal{L}} \boldsymbol{\lambda}_j, \qquad x \in [mD, (m+1)D),\ m = 0, 1, \ldots, \tag{3.202}$$

where: $\mathcal{P}(i)$ is the set of partitions of integer i; $\mathcal{L}$ is a partition in $\mathcal{P}(i)$; $r_1 > r_2 > \cdots > r_d$ are the distinct integers in $\mathcal{L}$ with multiplicities $n_1, \ldots, n_d$, respectively; $H(\mathcal{L}) \equiv n_1! n_2! \cdots n_d!$; $|\mathcal{L}| = n_1 + n_2 + \cdots + n_d$; $\prod_{j\in\mathcal{L}} \boldsymbol{\lambda}_j \equiv \lambda_{r_1}^{n_1} \lambda_{r_2}^{n_2} \cdots \lambda_{r_d}^{n_d}$. Also, if $i = 0$, then

$$\sum_{\mathcal{L}\in\mathcal{P}(0)} \frac{(iD-x)^{|\mathcal{L}|}}{H(\mathcal{L})} \prod_{j\in\mathcal{L}} \lambda_j \equiv 1.$$

The pdf of wait is $f_m(x) = F_m'(x)$. Differentiating (3.202) with respect to x, gives for $x \in (mD, (m+1)D)$, $m = 0, 1, 2, \ldots$,

$$f_m(x) = P_0 \sum_{i=0}^{m} e^{-\lambda(iD-x)} \sum_{\mathcal{L}\in\mathcal{P}(i)} (\lambda(iD-x) - |\mathcal{L}|) \frac{(iD-x)^{|\mathcal{L}|-1}}{H(\mathcal{L})} \prod_{j\in\mathcal{L}} \boldsymbol{\lambda}_j. \tag{3.203}$$

As a mild check on the cdf of W_q in M/$\{iD\}$/1 given in (3.202), we obtain from it the cdf of W_q in M/D/1 (formula (3.191)), namely

$$F_m(x) = P_0 \sum_{i=0}^{m} e^{-\lambda(iD-x)} \frac{(iD-x)^i}{i!} \lambda^i = P_0 \sum_{i=0}^{m} (-\lambda)^i \frac{(x-iD)^i}{i!} e^{-\lambda(iD-x)}, \qquad x \in [mD, (m+1)D), m = 0, 1, \ldots.$$

To explain, the latter M/D/1 formula it results since: (1) $\lambda_1 = \lambda$ and $\lambda_i = 0$, $i > 1$; (2) for each i, the only partition in $\mathcal{P}(i)$ that contributes positive terms is that of i 1s, $\{1, \ldots, 1\}$; (3) each i yields one such partition with $n_1 = i$, $H(\mathcal{L}) = i!$, and $\prod_{j\in\mathcal{L}} \boldsymbol{\lambda}_j = \lambda^i$.

Remark 3.27 In [128], formula (3.202) was derived by inversion of the Laplace transform of wait (see Eq. (3.69). The inversion procedure is at least as involved as the foregoing LC derivation. Moreover, it also requires the induction step. The advantages of the LC approach are: (1) the analysis prior to the induction step is intuitive and completely in the time domain; (2) the effect on the solution, due to the discontinuities in $f(x)$, and the continuity of

$F(x)$, is clear using LC; (3) since LC emphasizes sample paths, it enhances intuitive understanding of the model dynamics, and suggests new avenues for research.

3.13 M/G/1: Wait Related Reneging/Balking

We analyze an M/G/1 queue in which arrivals either: (1) join the system and stay for full service, or (2) balk from joining the system or renege from the waiting line, depending on their estimated (approximate) required arrival-point wait and on their staying resolve (e.g., patience). We assume that the arrivals to the system occur according to a Poisson process at rate λ, from a homogeneous source.

Let $\{W(t)\}_{t\geq 0}$ denote the virtual wait process, and τ_n the arrival time of customer C_n, $n = 1, 2, \ldots$ (Fig. 3.24). Let the service time be

$$S_n = \begin{cases} S \text{ if } C_n \text{ obtains full service,} \\ 0 \text{ if } C_n \text{ balks/reneges before starting service} \end{cases}, n = 1, 2, \ldots,$$

where S has cdf $B(x)$, $x > 0$, and $\overline{B}(x) = 1 - B(x)$, $x \geq 0$, independent of n. The arrival-point waiting time $W(\tau_n^-)$ (:=W_n) is the *required wait* before service of C $_n$, $n = 1, 2, \ldots$. We assume that a system manager informs C_n

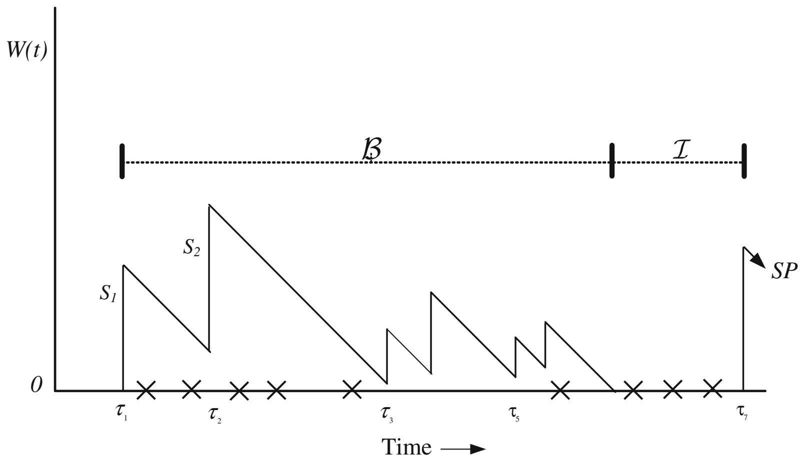


Fig. 3.24 M/G/1 with wait-dependent reneging: busy period $\mathcal{B}$, idle period $\mathcal{I}$; stayers arrive at τ_n, $n = 1, 2, \ldots$; balkers arrive at $\times$

at time τ_n^-, the *estimated* (*approximate)* waiting time W_n. Some arriving customers will balk immediately upon arrival. Others will wait hoping that the approximate wait is higher than the true wait, or that their patience will endure the true wait, or joining has high personal priority. For example, a bus terminal continuously displays electronically the (approximate) wait until the next bus departure; a doctor's office informs an arriving patient about the (approximate) required wait to see the doctor; a telephone answering service informs the caller about the (approximate) wait for the next available agent; etc. If the customers are mechanical devices needing service, the manager may accept or reject entrance to the system, according to the (approximate) required wait before service.

Define, for $n = 1, 2, \ldots$,

$$\theta_n = \begin{cases} 1 \text{ if } C_n \text{ stays and receives full service,} \\ 0 \text{ if } C_n \text{ balks from joining and is cleared.} \end{cases} \tag{3.204}$$

Since the customer source is homogeneous, we define the common random variable $\theta \equiv \theta_n$. Thus θ is a Bernoulli random variable taking the value 1 (***stay***), or 0 (***balk***).

Our aim here is to determine the steady-state mixed pdf of wait denoted by $\{P_0, f(x)\}_{x\geq 0}$, where $f(x)$:= pdf of customers who join and wait for service; and related quantities.

3.13.1 The Staying Function $\overline{R}(y)$, $y \geq 0$

For each $y \geq 0$, we define the common *conditional* probabilities

$$\begin{aligned} \overline{R}(y) &:= P(\theta = 1 | W_n = y), \, y \geq 0, \\ R(y) &:= P(\theta = 0 | W_n = y), \, y \geq 0, \end{aligned} \tag{3.205}$$

independent of $n = 1, 2, \ldots$. From (3.204)

$$\overline{R}(y) + R(y) = 1, \, y \geq 0; \tag{3.206}$$

$P(C_n \text{ stays} | W_n = y) = \overline{R}(y)$; $P(\mathrm{C}_n \text{ balks} | W_n = y) = R(y)$.

3.13.2 Sample Path of $\{W(t)\}_{t\geq 0}$

The r.v., $W(t)$, is the required wait until service of a would-be time-t arrival. Consider a sample path of $\{W(t)\}_{t\geq 0}$ (Fig. 3.24) and an arrival at τ_n^-. If $W_n = y$ then the SP jump size $\underset{dis}{=}$ S having cdf $B(\cdot)$ with probability $\overline{R}(y)$, and jump size = 0 with probability $R(y)$. A would-be arrival at time τ_n just after a balker/reneger arrives (and is cleared), also would have a required wait y until service. This implies $W(\tau_n) = W(\tau_n^-) = y$ if $y > 0$. The sample path would be continuous with slope -1 at τ_n (such t_ns are denoted by $\times$ in Fig. 3.24).

Integral Equation for $\{P_0, f(x)\}_{x\geq 0}$
An integral equation for $\{P_0, f(x)\}_{x\geq 0}$ is (see, e.g., Fig. 1.6 in Sect. 1.7, and Eq. (3.34) in Sect. 3.2.10)

$$f(x) = \lambda\overline{R}(0)P_0\overline{B}(x) + \lambda\int_0^x \overline{B}(x-y)\overline{R}(y)f(y)dy, x > 0, \quad (3.207)$$

with normalizing condition $P_0 + \int_0^\infty f(x)dx = 1$. The left and right sides of Eq. (3.207) are equal to the sample path down- and upcrossing rates of level x, respectively. The upward jump sizes are related to $\overline{B}(\cdot)$ on the right side. Jumps occur at rates that customers stay for service. These rates are state-dependent, viz., $\lambda\overline{R}(y)$, $y \geq 0$. The pdf on the left side is a time-average pdf, the pdf on the right side is the arrival-point pdf at arrival instants; their equality is addressed below. We first briefly discuss the system dynamics with respect to the state-dependence.

E**(Idle Period) and State Dependence**
Consider an idle period $\mathcal{I}$ (Fig. 3.24). When $y = 0$ arrivals enter the system at Poisson rate $\lambda\overline{R}(0)$, implying $E(\mathcal{I}) = 1/\left(\lambda\overline{R}(0)\right)$. Viewed alternatively, customers arrive at Poisson rate λ; at each arrival instant P(the customer stays for service) = $\overline{R}(0)$, and P(the customer balks) = $R(0)$. When $y = 0$, the number of balks until the next stay is distributed as a geometric random variable where a start of service is a *success* and a balk is a *failure*. Thus, E(number of arrivals until a start of service) = $1/\overline{R}(0)$ (see, e.g., p. 37 in [125]). The expected time between arrivals is $1/\lambda$. By independence of the interarrival times and random variable θ, $E(\mathcal{I}) = (1/\lambda)\cdot\left(1/\overline{R}(0)\right) = 1/\left(\lambda\overline{R}(0)\right)$, which agrees with taking $\lambda\overline{R}(0)$ to be the Poisson rate of *stayers*. A similar argument holds for any fixed arrival-point wait $y > 0$.

Equality of Time-Average and Arrival-Point PDFs
In Sect. 8.4.2, Chap. 8, we use the *embedded* LC method to show that the time-average pdf $\{P_0, f(x)\}_{x>0}$ is identical to the limiting arrival-point pdf

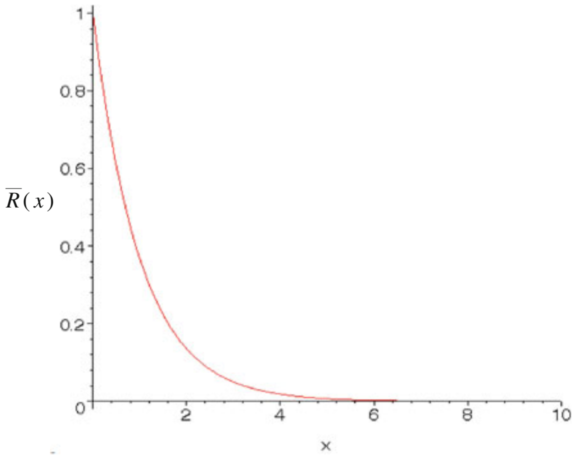


Fig. 3.25 Staying function $\overline{R}(x) = e^{-rx}$. ($r$ = 1 in the diagram); limit $L = 0$. $R(x) = 1 - \overline{R}(x) = 1 - e^{-rx}$., $x \geq 0$, the cdf of an Exp_r random variable

as $n \to \infty$, denoted by $\left\{P_{\iota,0},\ f_\iota(x)\right\}_{x>0,\ldots}$, because both pdfs satisfy integral equation (3.207) and the same normalizing condition.

Form of the Staying Function $\overline{R}(\cdot)$
We assume $\overline{R}(y)$, $y \geq 0$, is a monotone, piecewise continuous, non-increasing function (decreasing in the wide sense), with range a subset of [0, 1]. (See Figs. 3.25, 3.26 and 3.27.) Then $\lim_{y\to\infty} \overline{R}(y) := L \in [0, 1]$ exists. If $\overline{R}(y) \equiv 1$, $y \geq 0$, then $L = 1$, and there would be no reneging or balking; each arrival would receive full service. The model would be a standard M/G/1 queue.

Remark 3.28 In a more general model, $\overline{R}(y)$ may be an arbitrary function such that $\overline{R}(y) \in [0, 1]$, $y \geq 0$, is not necessarily monotone. In that case, the

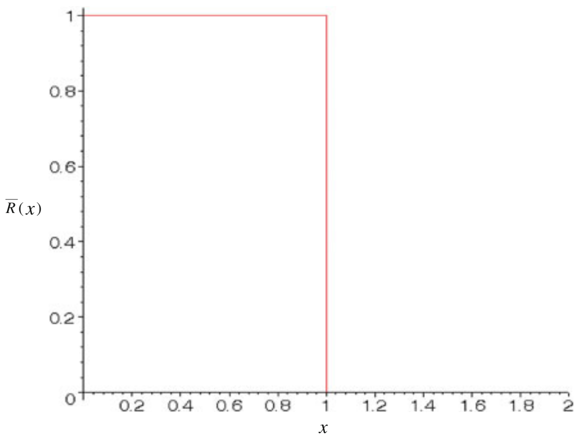


Fig. 3.26 Staying function $\overline{R}(x) = 1$, $0 < x < 1$, $\overline{R}(x) = 0$, $x \geq 1$. Limit $L = 0$. $R(x) = 1 - \overline{R}(x)$, is the cdf of a deterministic r.v.

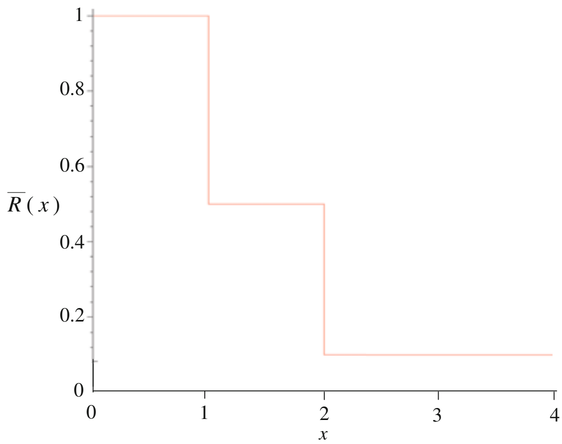


Fig. 3.27 $\overline{R}(x) = 1$, $x < 1$, $\overline{R}(x) = 0.5$, $1 \le x < 2$, $\overline{R}(x) = 0.1$, $x \ge 2$. $R(x)$ is not a cdf. Limit $L > 0$

presented analysis applies as well. However, the stability condition would be slightly modified (see Theorem 3.8 and Remark 3.31 below).

Interestingly, the renege/balk M/G/1 queue where $\overline{R}(x) = 1 \cdot \boldsymbol{I}_{[0,1)}(x) + 0 \cdot \boldsymbol{I}_{[1,\infty)}(x)$ is essentially the same as M/G/1 with a threshold at level 1 denoted as Variant 3 (with $K = 1$) in Sect. 3.16.6 below.

Proportion of Customers that Get Full Service

Denote by q_S the proportion of arrivals that are stayers. Then $q_S := P$(an arbitrary arrival gets full service), namely

$$q_S = \overline{R}(0)P_0 + \int_{y=0}^{\infty} \overline{R}(y) f(y) dy. \tag{3.208}$$

The proportion of customers that balk upon knowing their actual or approximate required wait before service is

$$q_B = 1 - q_S = R(0)P_0 + \int_{y=0}^{\infty} R(y) f(y) dy.$$

3.13.3 M/M/1: Wait Dependent Reneging/Balking

We now study the particular case where the service times of stayers are $\underset{dis}{=} \text{Exp}_{\mu}$, with $\overline{B}(x) = e^{-\mu x}$, $x \ge 0$. Then (3.207) becomes

$$f(x) = \lambda P_0 \overline{R}(0) e^{-\mu x} + \lambda \int_{y=0}^{x} e^{-\mu(x-y)} \overline{R}(y) f(y) dy. \tag{3.209}$$

Applying differential operator $\langle D + \mu \rangle$ to both sides of (3.209) yields the first order differential equation

$$\begin{aligned} \langle D + \mu \rangle\, f(x) &= \lambda \overline{R}(x) f(x), \\ f'(x) + (\mu - \lambda \overline{R}(x)) f(x) &= 0, \\ \frac{f'(x)}{f(x)} = \frac{d \ln f(x)}{dx} &= -(\mu - \lambda \overline{R}(x)). \end{aligned}$$

Integration on both sides of the last equation with respect to x, followed by exponentiation gives

$$f(x) = A e^{-\left(\mu x - \lambda \int_{y=0}^{x} \overline{R}(y) dy\right)}, \; x > 0, \tag{3.210}$$

where A is a constant. Letting $x \downarrow 0$ in (3.209) and (3.210) implies

$$f(0) = A = \lambda P_0 \overline{R}(0).$$

From LC, $f(0)$ is the SP *entrance* rate into $\boldsymbol{T} \times \{0\}$ (i.e., into level 0) from above. The term $\lambda P_0 \overline{R}(0)$ is the SP *exit* rate of level 0 above (i.e., into state-space interval $(0, \infty)$). The resulting pdf of wait is

$$f(x) = \lambda P_0 \overline{R}(0) e^{-\left(\mu x - \lambda \int_{y=0}^{x} \overline{R}(y) dy\right)}, \; x > 0. \tag{3.211}$$

The normalizing condition $P_0 + \int_{x=0}^{\infty} f(x) dx = 1$ leads to

$$P_0 = \frac{1}{1 + \lambda \overline{R}(0) \int_{x=0}^{\infty} e^{-\left(\mu x - \lambda \int_{y=0}^{x} \overline{R}(y) dy\right)} dx}. \tag{3.212}$$

3.13.4 M/M/1: Reneging/Balking-Stability Condition

In the $M_\lambda/M_\mu/1$ queue, Theorem 3.8 below gives a necessary and sufficient condition relating λ and μ, such that $\{P_0, f(x)\}_{x>0}$ exists (stability).

Theorem 3.8 In $M_\lambda/M_\mu/1$ with wait-time dependent reneging/balking assume the staying function $\overline{R}(x)$, $x \geq 0$, is monotone non-increasing and piecewise continuous. Let $L = \lim_{x\to\infty} \overline{R}(x)$. A necessary and sufficient condition for stability is

$$\begin{aligned} 0 < \lambda < \tfrac{\mu}{L} \quad &\text{if } 0 < L \le 1, \\ 0 < \lambda < \infty \quad &\text{if } L = 0. \end{aligned} \tag{3.213}$$

Proof (Adapted from [90]) By the hypothesis $1 \ge \overline{R}(a) \ge \overline{R}(b) \ge 0$ whenever $a < b$; hence $\lim_{x\to\infty} \overline{R}(x) := L \in (0, 1]$ exists (see, e.g., Problem *8, p. 119, in Chap. 8, in [137]). Stability holds iff the discrete state $\{0\}$ is positive recurrent iff $0 < P_0 \le 1$. Let

$$I := \int_{x=0}^{\infty} e^{-\mu x + \lambda \int_{y=0}^{x} \overline{R}(y)dy} dx,$$

in the denominator of (3.212). Stability holds iff

$$I < \infty. \tag{3.214}$$

We now show that the condition (3.214) is equivalent to the condition (3.213) above. We have $L \le \overline{R}(x)$, $x \ge 0$, because L is the greatest lower bound (i.e., *glb*, *infimum*) of the range of $\overline{R}(\cdot)$. Hence

$$\begin{aligned} \lambda L x \quad &= \quad \lambda \int_{y=0}^{x} L dx \le \lambda \int_{y=0}^{x} \overline{R}(y)dy \\ &\Longleftrightarrow \quad e^{-\mu x + \lambda L x} \le e^{-\mu x + \lambda \int_{y=0}^{x} \overline{R}(y)dy} \\ &\Longleftrightarrow \quad \int_{x=0}^{\infty} e^{-(\mu - \lambda L)x} dx \le I. \end{aligned} \tag{3.215}$$

For a given $\varepsilon > 0$ there exists $M_\varepsilon > 0$ such that $\overline{R}(x) < \varepsilon + L$ for $x > M_\varepsilon$. Thus

$$\begin{aligned} \lambda \int_{y=0}^{x} \overline{R}(y)dy &< \lambda \int_{y=0}^{M_\varepsilon} \overline{R}(y)dy + \lambda \int_{y=M_\varepsilon}^{x} (\varepsilon + L)\, dy \\ &= C_1 + \lambda (\varepsilon + L)\, x, \;\; x > M_\varepsilon \end{aligned}$$

$$\begin{aligned} &\Longrightarrow \; e^{-\mu x + \lambda \int_{y=0}^{x} \overline{R}(y)dy} < C_2 e^{-\mu x + \lambda(\varepsilon + L)x}, \;\; x > M_\varepsilon \\ &\Longrightarrow \; \int_{x=M_\varepsilon}^{\infty} e^{-\mu x + \lambda \int_{y=0}^{x} \overline{R}(y)dy} dx < C_2 \int_{x=M_\varepsilon}^{\infty} e^{(-\mu + \lambda L + \lambda\varepsilon)x} dx \\ &\Longrightarrow \; I < C_3 + C_2 \int_{x=M_\varepsilon}^{\infty} e^{(-\mu + \lambda L + \lambda\varepsilon)x} dx, \end{aligned} \tag{3.216}$$

where C_1, C_2, C_3 are positive constants. Combining inequalities (3.215) and (3.216) gives

$$\int_{x=0}^{\infty} e^{-(\mu-\lambda L)x} dx \le I < C_3 + C_2 \int_{x=M_\varepsilon}^{\infty} e^{(-\mu+\lambda L+\lambda\varepsilon)x} dx. \tag{3.217}$$

In (3.217), if $I < \infty$ then

$$\int_{x=0}^{\infty} e^{-(\mu-\lambda L)x} dx < \infty \iff \mu - \lambda L > 0. \tag{3.218}$$

If $\mu - \lambda L > 0$ then choose ε so that $-\mu + \lambda L + \lambda\varepsilon < 0$, i.e., $\varepsilon < \frac{\mu-\lambda L}{\lambda}$. Then

$$\int_{x=M_\varepsilon}^{\infty} e^{(-\mu+\lambda L+\lambda\varepsilon)x} dx < \infty \implies I < \infty. \tag{3.219}$$

The stability condition (3.213) is equivalent to (3.218) and (3.219). ■

Remark 3.29 To shed additional perspective on the stability condition (3.213), consider the exponent in the integrand of

$$I \equiv \int_{x=0}^{\infty} e^{-\mu x+\lambda \int_{y=0}^{x} \overline{R}(y)dy} dx.$$

The function μx is linear with slope $\mu > 0$. The function of x, $\int_{y=0}^{x} \overline{R}(y)dy$, $x > 0$, is positive and increasing with slope $\frac{d}{dx}\int_{y=0}^{x} \overline{R}(y)dy = \overline{R}(x)$, $x > 0$. If $\overline{R}(x)$, $x > 0$, is *strictly* decreasing and differentiable, then $\int_{y=0}^{x} \overline{R}(y)dy$ is concave since $\frac{d^2}{dx^2}\int_{y=0}^{x} \overline{R}(y)dy = \frac{d}{dx}\overline{R}(x) < 0$, $x > 0$. Additionally, $\lim_{x\to\infty} \frac{d}{dx}\int_{y=0}^{x} \overline{R}(y)dy = \lim_{x\to\infty} \overline{R}(x) = L$. We compare the graphs of μx and $\lambda \int_{y=0}^{x} \overline{R}(y)dy$, $x > 0$ in Fig. 3.28.

If $L > 0$ then there exists $M \ge 0$ such that $\mu x - \lambda \int_{y=0}^{x} \overline{R}(y)dy > 0$ for all $x \ge M$ iff $\mu - \lambda L > 0$ iff $\lambda < \mu/L$. If $L = 0$, there exists $M \ge 0$ such that $\mu x - \lambda \int_{y=0}^{x} \overline{R}(y)dy > 0$ for all $x \ge M$ iff $\mu \ge \lambda \cdot 0$. In that case λ can assume any positive value, i.e., $\lambda \in (0, \infty)$.

Remark 3.30 If $\overline{R}(x)$ is piecewise continuous, we can obtain similar perspective as in Remark 3.29.

Another Look at Theorem 3.8

We provide an alternative verification of the stability condition, in order to clarify the intuition behind the result. Consider an *optimization problem*

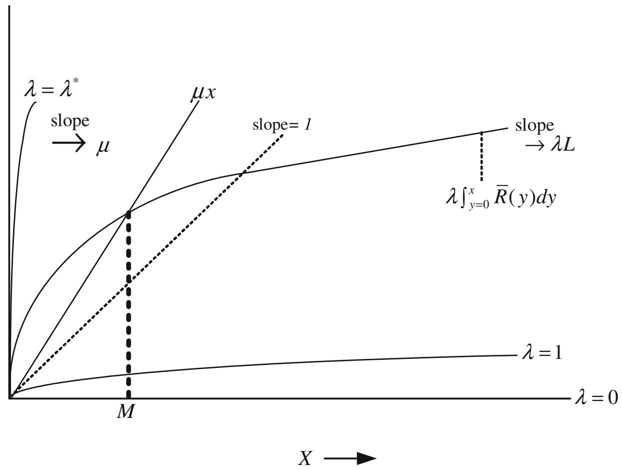


Fig. 3.28 Functions μx and $\lambda \int_{y=0}^{x} \overline{R}(y)dy$, indicating M such that $\mu x - \lambda \int_{y=0}^{x} \overline{R}(y)dy > 0$ for $x \geq M$. Indicates range $0 < \lambda < \lambda^*$ such that stability holds. The system is stable for λ if $\lambda \int_{y=0}^{x} \overline{R}(y)dy$ intersects and remains below μx thereafter

where λ is the decision variable. We shall derive a range $0 < \lambda < \lambda^*$ for which there exists $M \geq 0$ such that $\mu x - \lambda \int_{y=0}^{x} \overline{R}(y)dy > 0$ for all $x \geq M$, thereby making the system stable (see Fig. 3.28). The value λ^* is the solution of the following optimization problem **P**. (Note that $\mu > 0$, $L \geq 0$.)

Problem **P**	
Maximize	λ
such that	$\mu - \lambda L \geq 0$
subject to	$\lambda > 0.$

The solution of problem **P** is readily seen to be

$$\lambda^* = \begin{cases} \frac{\mu}{L} \text{ if } L > 0, \\ \infty \text{ if } L = 0, \end{cases}$$

which is the same result as in Theorem 3.8.

Remark 3.31 The stability condition given in Theorem 3.8 was originally proved in [16] together with a theorem in which $\overline{R}(y)$, $y \geq 0$ may be other than monotone non-increasing. That proof is based on the fact that

$$\int_{x=0}^{\infty} e^{-\mu x+\lambda \int_{y=0}^{x} \overline{R}(y)dy} dx = \int_{x=0}^{\infty} e^{-\mu x} \cdot e^{\lambda \int_{y=0}^{x} \overline{R}(y)dy} dx$$

is the Laplace transform of $e^{\lambda \int_{y=0}^{x} \overline{R}(y)dy}$ evaluated at the parameter μ. A sufficient condition for the Laplace transform to be finite is that $e^{\lambda \int_{y=0}^{x} \overline{R}(y)dy}$ is of exponential order. Let $\overline{L} = \limsup_{x\to\infty} \overline{R}(x)$. A *sufficient* condition for stability is

$$\lambda < \frac{\mu}{\overline{L}} \text{ if } \overline{L} > 0,$$
$$\lambda < \infty \text{ if } \overline{L} = 0.$$

3.13.5 M/M/1: Reneging/Balking-Exponential $\overline{R}(\cdot)$

We illustrate the M/G/1 model by taking G(·) := Exp$_\mu$. Let $\overline{B}(x) = e^{-\mu x}$, $x \geq 0$, and $\overline{R}(y) = e^{-ry}$, $y > 0$, $r > 0$. Thus $\overline{R}(y)$ is monotone decreasing and $L = \lim_{y\to\infty} \overline{R}(y) = 0$ in the notation of Sect. 3.13.3. Also, $\overline{R}(0) = 1$, so that all zero-wait customers join the system.

Equation (3.209) becomes

$$f(x) = \lambda P_0 e^{-\mu x} + \lambda \int_{y=0}^{x} e^{-\mu(x-y)} e^{-ry} f(y)dy. \tag{3.220}$$

Substituting e^{-ry} for $\overline{R}(y)$ in (3.211) gives the pdf of wait $\{P_0, f(x)\}_{x>0}$ as

$$f(x) = \lambda P_0 e^{-\mu x+\frac{\lambda}{r}(1-e^{-rx})} = \lambda e^{\lambda/r} P_0 e^{-\mu x-\frac{\lambda}{r}e^{-rx}}, x > 0. \tag{3.221}$$

Substituting (3.221) into (3.212) yields

$$P_0 = \frac{1}{1+\lambda e^{\lambda/r} \int_{x=0}^{\infty} e^{-\mu x-\frac{\lambda}{r}e^{-rx}} dx} = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda}+e^{\lambda/r} \int_{x=0}^{\infty} e^{-\mu x-\frac{\lambda}{r}e^{-rx}} dx}. \tag{3.222}$$

In the denominator of P_0 the term $\int_{x=0}^{\infty} e^{-\mu x-\frac{\lambda}{r}e^{-rx}} dx < 1/\mu < \infty$ for *every* trio of positive numbers $\{\lambda, \mu, r\}$, since the integrand $e^{-\mu x-\frac{\lambda}{r}e^{-rx}} < e^{-\mu x}$, $x \geq 0$. Thus $P_0 > 0$ for all positive $\{\lambda, \mu, r\}$. In particular $P_0 > 0$ for *every* arrival rate $\lambda > 0$. This adds credence to Theorem 3.8 above when $\lim_{x\to\infty} \overline{R}(x) = L = 0$.

Expected Busy Period $E(\mathcal{B})$

In the standard M/G/1 queue, $E(\mathcal{B}) = E(S)/(1 - \lambda E(S))$. However, in M/G/1 with balking $P_0 \neq 1 - \lambda E(S)$. Hence, we use the more fundamental formula for $E(\mathcal{B})$ in terms of P_0. From (3.82) and (3.222),

$$E(\mathcal{B}) = \frac{1 - P_0}{f(0)} = \frac{1 - P_0}{\lambda P_0}$$

$$= e^{\frac{\lambda}{r}} \int_{x=0}^{\infty} e^{-\mu x - \frac{\lambda}{r} e^{-rx}} dx = \int_{x=0}^{\infty} e^{-\mu x + \frac{\lambda}{r}(1 - e^{-rx})} dx. \tag{3.223}$$

We can infer formula (3.223) immediately since P_0 in (3.222) has the form

$$P_0 = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda} + E(\mathcal{B})} = \frac{E(\mathcal{I})}{E(\mathcal{I}) + E(\mathcal{B})},$$

and $E(\mathcal{I}) = 1/\lambda$, because all zero-wait customers stay for service if $\overline{R}(y) = e^{-ry},\ y \geq 0$.

3.13.6 M/M/1: Reneging/Balking and Standard M/M/1

Assume $\lambda < \mu$ (stability condition for standard M/M/1). In (3.223), $(1 - e^{-rx}) < rx, x > 0$ and $(1 - e^{-r \cdot 0}) = 0$. Letting subscript 'b' represent M/M/1 with reneging/balking, and subscript 's' the standard M/M/1, we have, since $(1 - e^{-rx})/r < x, x > 0$,

$$E(\mathcal{B}_{\rm b}) = \int_{x=0}^{\infty} e^{-\mu x + \frac{\lambda}{r}(1 - e^{-rx})} dx < \int_{x=0}^{\infty} e^{-(\mu - \lambda)x} dx = \frac{1}{\mu - \lambda} = E(\mathcal{B}_{\rm s}).$$

In (3.222), we again apply the inequality

$$\int_{x=0}^{\infty} e^{-\mu x + \frac{\lambda}{r}(1 - e^{-rx})} dx < \frac{1}{\mu - \lambda},$$

which gives

$$P_{\rm b,0} > \frac{1}{1 + \lambda \cdot \frac{1}{\mu - \lambda}} = 1 - \frac{\lambda}{\mu} = P_{\rm s,0}.$$

The comparisons for $E(\mathcal{B})$ and P_0 are intuitive. In the reneging/balking model, the arrival rate of customers that increase workload is $\lambda \overline{R}(y),\ y \geq 0$. In the standard model, it is $\lambda > \lambda \overline{R}(y),\ y > 0$.

3.13.7 M/M/1: Reneging/Balking-Number in System

Let P_n, a_n, d_n denote the steady-state probabilities of n stayers in the system at an arbitrary time point, just before an arrival and just after a departure, respectively. Then $a_n = d_n = P_n$, $n = 0, 1, 2, \ldots$, (see Sect. 8.2.2, p. 500 in [125]); and P_0 is given in (3.222). Furthermore, since $\overline{R}(y) = e^{-ry}$, $y \geq 0$,

$$\begin{aligned} d_n &= \int_{x=0}^{\infty} \left(e^{-\lambda\int_0^x \overline{R}(y)dy}\right) \frac{\left(\lambda\int_0^x \overline{R}(y)dy\right)^{n-1}}{(n-1)!} f(x)dx \\ &= \lambda P_0 \int_{x=0}^{\infty} \frac{(\frac{\lambda}{r}(1-e^{-rx}))^{n-1}}{(n-1)!} e^{-\mu x} dx, n = 1, 2, \ldots. \end{aligned} \tag{3.224}$$

(see Eq. (3.76) in Sect. 3.4.8).

In formula (3.224), $\lambda\overline{R}(y)$ $(=\lambda e^{-ry})$ is the arrival rate of stayers when the required wait is y.

Remark 3.32 We outline a derivation of (3.224) using an approximation of $\overline{R}(x)$ by a step function. Let $[0,\Omega)$ be a large waiting-time interval in the state space. Partition $[0,\Omega)$ into m subintervals $\Delta_i = [x_i, x_{i+1})$, $i = 0, \ldots, m-1$, where $x_0 = 0$, $x_m = \Omega$. We then approximate $\overline{R}(y)$ by $\overline{R}(y) \equiv \overline{R}(x_i)$, $y \in \Delta_i$. Thus the arrival rate of stayers is a constant $\lambda\overline{R}(x_i)$ if the required wait $y \in [x_i, x_{i+1})$ at an arrival instant. The probability that $n-1$ stayers arrive during an individual required wait $y \in \Delta_i$ is approximately

$$\frac{e^{-\lambda\overline{R}(x_i)x_i'}(\lambda\overline{R}(x_i)x_i')^{n-1}}{(n-1)!}$$

where $x_i' \in \Delta_i$. The probability that $n-1$ stayers arrive during $(0,\Omega)$ is approximately the Riemann sum

$$\sum_{i=0}^{m-1} \frac{e^{-\lambda\overline{R}(x_i)x_i'}(\lambda\overline{R}(x_i)x_i')^{n-1}}{(n-1)!} f(x_i'')\Delta_i$$

where $x_i'' \in \Delta_i$. Let $m \to \infty$ and $\Delta_i \downarrow 0$, $i = 0, \ldots, m-1$. Then $x_i, x_i', x_i'' \to x$ and

$$\begin{aligned} &\lim_{\substack{m\to\infty \\ \Delta_i\downarrow 0}} \sum_{i=0}^{m-1} e^{-\lambda\overline{R}(x_i)x_i'} \frac{(\lambda\overline{R}(x_i)x_i')^{n-1}}{(n-1)!} f(x_i'')\Delta_i \\ &= \int_{x=0}^{\Omega} e^{-\lambda\overline{R}(x)x} \frac{(\lambda\overline{R}(x)x)^{n-1}}{(n-1)!} f(x)dx. \end{aligned}$$

Letting $\Omega \to \infty$ implies (3.224), where $f(x)$ is given by (3.221).

3.13.8 Proportion of Customers Served

In M/M/1 with wait-time dependent reneging/balking and $\overline{R}(y) = e^{ry}$, $y \geq 0$, from (3.208), (3.221) and (3.222), the proportion of customers that get complete service is

$$q_S = P_0 + \int_{x=0}^{\infty} e^{-rx} f(x)dx = \frac{1 + \lambda e^{\frac{\lambda}{r}} \int_{x=0}^{\infty} e^{-\mu x - \frac{\lambda}{r} e^{-rx} - rx} dx}{1 + \lambda e^{\frac{\lambda}{r}} \int_{x=0}^{\infty} e^{-\mu x - \frac{\lambda}{r} e^{-rx}} dx}. \tag{3.225}$$

The proportion of customers that renege from the waiting line is $1 - q_S$.

In the expressions for P_0, $E(\mathcal{B})$, and q_S the integrals do not have closed forms. They can be evaluated using series expansion or numerical methods, for given values of λ, μ, and r.

3.14 M/G/1 with Priorities

Assume N types of customers arrive at a single-server system at independent Poisson rates $\lambda_i, i = 1, \ldots, N$. We denote the type-$i$ service time as S_i having cdf $B_i(x), x > 0, \overline{B}_i(x) = 1 - B_i(x), x \geq 0$, and pdf $b_i(x), x > 0$. We assume type 1 ($i = 1$) has the highest priority, type 2 the next highest, …, and type N ($i = N$) the lowest priority. The service discipline is FCFS within priority classes. The priority discipline is non-preemptive, i.e., any customer that starts service is allowed to complete it without interruption. The customer at the head of the highest-priority line, among all waiting customers, will start service immediately after the next service completion.

Denote the steady-state pdf and cdf of wait before service of a type i customer, by $\{P_0, f_i(x)\}_{x>0}$, and $F_i(x)$, $x \geq 0$ respectively. The probability P_0 of a zero wait, is independent of customer type.

3.14.1 Two Priority Classes

For exposition we consider two priority classes, so that $N = 2$. We will confirm the well-known stability condition, $\lambda_1 E(S_1) + \lambda_2 E(S_2) < 1$, using an

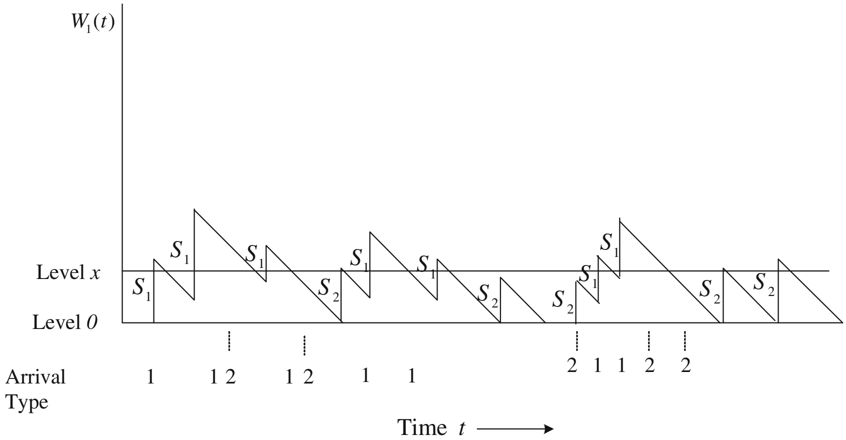


Fig. 3.29 Sample path of virtual wait for high priority type-1 arrivals. Low priority type-2 arrivals that must wait, start service at the end of a $\mathcal{B}_1$ or a $\mathcal{B}_{21}$ (Fig. 3.30) busy period. All type 2 jumps start at level 0

LC approach. Let $\{W_1(t)\}_{t\geq 0}$ be the virtual wait process for *type*-1 *customers*; a sample path is shown in Fig. 3.29. Fix level $x > 0$ in the state space.

3.14.2 Integral Equation for $\{P_0, f_i(x)\}_{x>0}$

From the sample path, we construct the integral equation

$$f_1(x) = \lambda_1 \overline{B}_1(x)P_0 + \lambda_2 \overline{B}_2(x)P_0 + \lambda_1 \int_{y=0}^{x} \overline{B_1}(x-y) f_1(y)dy$$
$$+\lambda_2(1-P_0)\overline{B}_2(x). \tag{3.226}$$

To explain (3.226), the left side $f_1(x)$ is the SP downcrossing rate of x (as in Theorem 1.1 in Chap. 1). On the right side, the terms $\lambda_1 \overline{B}_1(x)P_0$ and $\lambda_2 \overline{B}_2(x)P_0$ are respectively the SP upcrossing rates of x due to type-1 and type-2 arrivals, when the system is empty. The term $\lambda_1 \int_{y=0}^{x} \overline{B_1}(x-y) f_1(y)dy$ is the upcrossing rate of x due to type-1 arrivals that wait a positive time $y \in (0,x)$. The term $\lambda_2(1-P_0)\overline{B}_2(x)$ is the upcrossing rate of x due to type-2 arrivals that wait positive times before they start service. The first-in-line of such type 2s must wait *until the end* of a type 1 busy period to start service. Any other such type 2s wait longer before they start service. Those type 2s can start service only when the type-1 virtual wait hits level 0. The corresponding SP jumps of size S_2 start at level 0. The *long-run rate* at which such type 2s start service is $\lambda_2(1-P_0)$ since all type 2s must eventually get served in a finite time, due to stability.

3.14.3 Stability Condition

Integrate both sides of (3.226) with respect to x on $(0,\infty)$. Since $\int_{x=0}^{\infty} f_1(x)dx = 1-P_0$, and $\int_{x=0}^{\infty} \overline{B}_i(x)dx = E(s_i)$ some algebra yields

$$P_0 = 1 - \lambda_1 E(S_1) - \lambda_2 E(S_2) = 1 - \rho_1 - \rho_2, \tag{3.227}$$

where $\rho_i = \lambda_i E(S_i)$, $i = 1, 2$. For stability, we must have $0 < P_0 < 1$, or

$$0 < \rho_1 + \rho_2 < 1, \tag{3.228}$$

which implies both $\rho_1 < 1$ and $\rho_2 < 1$.

3.14.4 Expected Wait of High Priority Customers

We confirm the known formula for the expected wait of type-1 customers using (3.226). Denote the wait in queue before service of an arbitrary type-1 arrival by $W_{q,1}$. Multiplying both sides of (3.226) by x and integrating on $(0,\infty)$ with respect to x, the left side becomes $\int_0^{\infty} x f_1(x)dx = E(W_{q1})$; the right side results in the equation

$$E(W_{q1}) = \left(\lambda_1 \frac{E(S_1^2)}{2} + \lambda_2 \frac{E(S_2^2)}{2}\right) P_0 + \lambda_1 E(S_1) E(W_{q1})$$
$$+\lambda_1(1-P_0)\frac{E(S_1^2)}{2} + \lambda_2(1-P_0)\frac{E(S_2^2)}{2}.$$

Simplifying yields the familiar result (e.g., p. 545 in [125])

$$E(W_{q1}) = \frac{\lambda_1 E(S_1^2) + \lambda_2 E(S_2^2)}{2(1-\rho_1)}. \tag{3.229}$$

3.14.5 Equation for PDF of Wait of Type-2 Customers

Let $\{W_2(t)\}_{t\geq 0}$ be the virtual wait process of type-2 customers. Let $W_{q,2}$ be the steady-state wait. Denote the pdf of $W_{q,2}$ by $\{P_0, f_2(x)\}_{x>0}$, for which we now develop an integral equation.

Preliminaries
Let $\mathcal{B}_1$ denote a an M/G/1 type-1 *busy period*, consisting of type-1s only, having cdf $\boldsymbol{B}_1(x)$, $x > 0$ and $\overline{\boldsymbol{B}}_1(x) = 1 - \boldsymbol{B}_1(x)$, $x \geq 0$. We let $\mathcal{B}_{2,1}$ denote a busy period in which the first service is type 2, and all subsequent services are type 1 (Fig. 3.30). Let random variable $N_{S_{2,1}}$ denote the number of strict descending ladder points that occur in a sample path of a $\mathcal{B}_{2,1}$ busy period. Then $N_{S_{2,1}}$ has the same distribution as the number of type-1 customers that arrive during a type-2 *service time* S_2. Thus we have

$$\mathcal{B}_{2,1} \underset{dis}{=} S_2 + \sum_{i=1}^{N_{S_{2,1}}} \mathcal{B}_{1,i}, \tag{3.230}$$

where the $\mathcal{B}_{1,i}$s are i.i.d. random variables distributed as an M/G/1 type-1 busy period $\mathcal{B}_1$ independent of $N_{S_{2,1}}$. Equation (3.230) follows due to the memoryless property of the type-1 inter-arrival times ($\underset{dis}{=} \text{Exp}_{\lambda_1}$). (A related discussion of busy period structure is given above in Sect. 3.4.12.)

We illustrate the meaning of $N_{S_{2,1}}$ in Fig. 3.30, with $N_{S_{21}} = 3$. There are three type-1 sub-busy periods in $B_{2,1}$. There are four vertical gaps, each distributed as an inter-arrival time, separating and bordering on these three sub-busy periods. The basic observation is that the sum of the four gaps is equal to S_2.

From (3.80)

$$E(\mathcal{B}_1) = \frac{E(S_1)}{1 - \lambda_1 E(S_1)}. \tag{3.231}$$

Taking expected values in (3.230) we obtain

$$\begin{aligned} E(\mathcal{B}_{2,1}) &= E(S_2) + \lambda_1 E(S_2) E(\mathcal{B}_1) \\ &= E(S_2) + \lambda_1 E(S_2) \frac{E(S_1)}{1 - \lambda_1 E(S_1)} \\ &= \frac{E(S_2)}{1 - \lambda_1 E(S_1)} = \frac{E(S_2)}{1 - \rho_1}. \end{aligned} \tag{3.232}$$

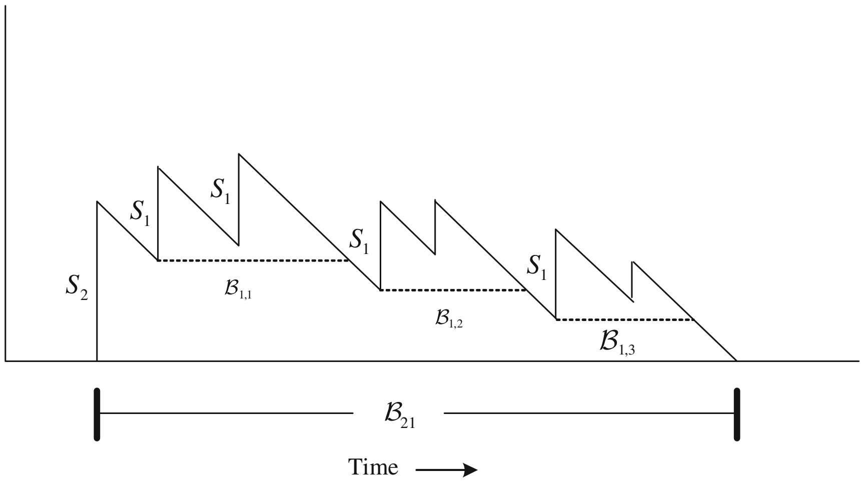


Fig. 3.30 Busy period $\mathcal{B}_{2,1}$. Initial jump is a type 2 service S_2. Each subsequent jump is a type 1 service S_1. $\mathcal{B}_{1,j}$, $j = 1, 2, \ldots$, are M/G/1 type 1 busy periods

Remark 3.33 $E(\mathcal{B}_{2,1})$ is the same as the expected busy period in an M/G/1 queue in which zero-waiting customers receive exceptional service. Thus we can obtain (3.232) immediately as a special case of (3.147).

Let $\boldsymbol{B}_{2,1}(x)$ denote the cdf of $\mathcal{B}_{2,1}$, and $\overline{\boldsymbol{B}}_{2,1}(x) = 1 - \boldsymbol{B}_{2,1}(x)$, $x \geq 0$. Consider a sample path of the virtual wait of type-2 customers $\{W_2(t)\}_{t\geq 0}$ (Fig. 3.31). The sample path illustrates that *type*-2 customers may view the model as a queue with server vacations (see Sect. 3.15). When a type 1 arrives to an empty system, the server vacation is $\mathcal{B}_1$. When a type 2 arrives, the server vacation consists of $N_{S_{21}}\mathcal{B}_1$s. By (3.230), type-2 generated SP jumps are $\underset{dis}{=} \mathcal{B}_{2,1}$.

Integral Equation for $f_2(x)$
We now construct the integral equation

$$f_2(x) = \lambda_1 \overline{\boldsymbol{B}}_1(x) P_0 + \lambda_2 \overline{\boldsymbol{B}}_{2,1}(x) P_0 + \lambda_2 \int_{y=0}^{x} \overline{\boldsymbol{B}}_{2,1}(x-y) f_2(y) dy. \tag{3.233}$$

In (3.233) the left side $f_2(x)$ is the sample-path downcrossing rate of level x (Theorem 1.1 in Chap. 1). On the right side the term $\lambda_1 \overline{\boldsymbol{B}}_1(x) P_0$ is the SP upcrossing rate of x due to type-1 arrivals when the system is empty. A potentially arriving type-2 customer, immediately after the initial type-1 starts service, would wait a type-1 busy period before starting service. The term $\lambda_2 \overline{\boldsymbol{B}}_{2,1}(x) P_0$ is the SP upcrossing rate of x due to type-2 arrivals when the

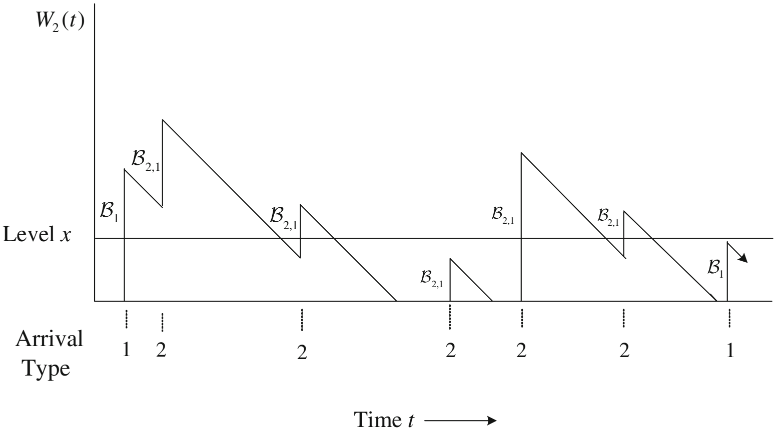


Fig. 3.31 Sample path of virtual wait for *low priority, type 2 arrivals*. High priority type 1's that arrive when the system is empty generate jumps distributed as $\mathcal{B}_1$ busy periods. All type 2 arrivals generate jumps distributed as $\mathcal{B}_{2,1}$ busy periods (see Fig. 3.30). All type 1's that must wait, are counted in the $\mathcal{B}_{2,1}$ jumps

system is empty. A potentially arriving type-2 customer, immediately after the type 2 starts service, would wait a busy period $\mathcal{B}_{2,1}$ before starting service. It is possible that $\mathcal{B}_{2,1}$ consists of the initial type-2 service only. Possibly no type 1s arrive during the initial service time. Generally, $\mathcal{B}_{2,1}$ includes an additional run of $N_{S_{2,1}}$ $\mathcal{B}_1$s (Fig. 3.30). The term $\lambda_2 \int_{y=0}^{x} \overline{\boldsymbol{B}}_{2,1}(x-y) f_2(y)dy$ is the upcrossing rate of x due to type-2 arrivals that must wait a positive time $y \in (0, x)$. A would-be type-2 customer that arrives immediately after such a type-2 arrival, would face an additional wait equal to $\mathcal{B}_{2,1}$ before starting service.

The three terms on the right of (3.233) account for all arrivals to the system. The type 2s are counted in the last two terms; they include all type 2s that wait ≥ 0. The type 1s are counted in all three terms. The type 1s that wait zero are counted in the first term. The type 1s that wait a positive time are counted in all three terms.

Both Types Have the Same P_0

We test for consistency of integral equations (3.233) and (3.226), by checking whether they give the same value of P_0. It is required to show that (3.227) results from (3.233). We integrate both sides of (3.233) with respect to x on $(0, \infty)$. Simplification gives

$$\begin{aligned}1 - P_0 &= \lambda_1 E(\mathcal{B}_1) P_0 + \lambda_2 E(\mathcal{B}_{21}) P_0 + \lambda_2 E(\mathcal{B}_{21})(1 - P_0) \\ &= \lambda_1 E(\mathcal{B}_1) P_0 + \lambda_2 E(\mathcal{B}_{21}).\end{aligned}$$

Substituting for $E(\mathcal{B}_1)$, $E(\mathcal{B}_{21})$ from (3.231), (3.232) respectively we obtain

$$1 - P_0 = \lambda_1 \frac{E(S_1)}{1 - \lambda_1 E(S_1)} P_0 + \lambda_2 \frac{E(S_2)}{1 - \lambda_1 E(S_1)},$$

or

$$P_0 = 1 - \lambda_1 E(S_1) - \lambda_2 E(S_2) = 1 - \rho_1 - \rho_2,$$

which is identical to (3.227) QED.

3.14.6 Expected Wait of Type-2 Customers

We obtain the expected wait $E(W_{q,2})$ by multiplying integral equation (3.233) by x on both sides and integrating with respect to x on $(0, \infty)$. Some algebra gives

$$\begin{aligned} E(W_{q2}) = &\ \lambda_1 \tfrac{E(\mathcal{B}_1^2)}{2} P_0 + \lambda_2 \tfrac{E(\mathcal{B}_{21}^2)}{2} P_0 \\ &+ \lambda_2 \tfrac{E(\mathcal{B}_{21}^2)}{2} (1 - P_0) + \lambda_2 E(\mathcal{B}_{21}) E(W_{q2}) \end{aligned}$$

or

$$E(E(W_{q2})) = \frac{\lambda_1 E(\mathcal{B}_1^2) P_0 + \lambda_2 E(\mathcal{B}_{21}^2)}{2(1 - \lambda_2 E(\mathcal{B}_{21}))}.$$

Substituting from (3.85), (3.227) and (3.232) gives

$$E(W_{q,2}) = \frac{\left(\lambda_1 \frac{E(S_1^2)}{(1-\rho_1)^3} (1 - \rho_1 - \rho_2) + \lambda_2 E(\mathcal{B}_{21}^2)\right) \cdot (1 - \rho_1)}{2\,(1 - \rho_1 - \rho_2)}. \quad (3.234)$$

The term $\lambda_2 E(\mathcal{B}_{2,1}^2)$ in the numerator of (3.234) is

$$\begin{aligned} \lambda_2 E(\mathcal{B}_{2,1}^2) = &\ \lambda_2 E\left(\left(S_2 + \sum_{i=1}^{N_{S_{2,1}}} \mathcal{B}_{1,i}\right)^2\right) \\ = &\ \lambda_2 E(S_2^2) + 2\lambda_2 E\left(S_2 \sum_{i=1}^{N_{S_{21}}} \mathcal{B}_{1,i}\right) + \lambda_2 E\left(\left(\sum_{i=1}^{N_{S_{21}}} \mathcal{B}_{1,i}\right)^2\right). \end{aligned}$$

We condition on $N_{S_{2,1}} = n$, $S_2 = s$ in the last two terms. Then $N_{S_{2,1}}$ is a Poisson random variable with parameter $\lambda_1 s$. We then carry out some algebra, and "uncondition". This procedure yields

$$\lambda_2 E(\mathcal{B}_{2,1}^2) = \lambda_2 E(S_2^2) + 2\lambda_2 E(S_2^2)\frac{\rho_1}{1-\rho_1}$$
$$+\lambda_2(\lambda_1 E(S_2)E(\mathcal{B}_1^2) + \lambda_1^2 (E(\mathcal{B}_1))^2 E(S_2^2)).$$

Substituting from (3.85) into the last equation gives

$$\lambda_2 E(\mathcal{B}_{2,1}^2) = \lambda_2 E(S_2^2) + 2\lambda_2 E(S_2^2)\tfrac{\rho_1}{1-\rho_1}$$
$$+\rho_2\lambda_1 \tfrac{E(S_1^2)}{(1-\rho_1)^3} + \lambda_2 \tfrac{\rho_1^2}{(1-\rho_1)^2} E(S_2^2). \tag{3.235}$$

Substituting the expression in (3.235) for $\lambda_2 E(\mathcal{B}_{2,1}^2)$ in the numerator of (3.234) gives

$$\text{coefficient of } E(S_1^2) = \frac{\lambda_1}{(1-\rho_1)},$$
$$\text{coefficient of } E(S_2^2) = \frac{\lambda_2}{(1-\rho_1)}.$$

Hence

$$E(W_{q2}) = \frac{\frac{\lambda_1}{(1-\rho_1)} E(S_1^2) + \frac{\lambda_2}{(1-\rho_1)} E(S_2^2)}{2\,(1-\rho_1-\rho_2)}$$
$$= \frac{\lambda_1 E(S_1^2) + \lambda_2 E(S_2^2)}{2(1-\rho_1)\,(1-\rho_1-\rho_2)}, \tag{3.236}$$

which agrees with the result in the literature (e.g., p. 545 in [125]).

Remark 3.34 We have used LC to derive $E(W_{q,1})$ from the integral equation for $f_1(x)$/, and $E(W_{q,2})$ from the integral equation for $f_2(x)$. The importance of this approach is that we essentially have an analytic solution for the pdfs and cdfs of wait of both priority classes. The LC analysis is in the time domain without use of transforms. Integral equations (3.226), (3.233) can be solved analytically in some cases; or else numerically. The LC analysis highlights conceptual properties of the priority queue that are in common with queues having: (1) service time depending on wait, (2) multiple Poisson inputs, (3) server vacations. In addition, the exercise of constructing the sample paths of wait for the different priority classes, leads to an intuitive understanding of the model dynamics.

3.14.7 Exponential Service

We now solve for $\{P_0, f_1(x)\}_{x>0}$ in an M/M/1 queue with two priority types. Here $S_i \underset{dis}{=} Exp_{\mu_i}, i = 1, 2$. Substituting $\overline{B}_i(x) = e^{-\mu_i x}$ into (3.226) gives an integral equation for $f_1(x)$,

$$\begin{aligned} f_1(x) = \ & \lambda_1 e^{-\mu_1 x} P_0 + \lambda_2 e^{-\mu_2 x} P_0 + \lambda_1 \int_{y=0}^{x} e^{-\mu_1(x-y)} f_1(y) dy \\ & + \lambda_2(1 - P_0) e^{-\mu_2 x}. \end{aligned} \quad (3.237)$$

We apply differential operator $\langle D + \mu_1 \rangle \langle D + \mu_2 \rangle$ to both sides of (3.237), obtaining the second order differential equation

$$\langle D + \mu_2 \rangle \langle D + \mu_1 - \lambda \rangle f_1(x) = 0,$$

with solution

$$f_1(x) = ae^{-(\mu_1 - \lambda_1)x} + be^{-\mu_2 x}, x \geq 0, \quad (3.238)$$

where constants a, b are to be determined.

Letting $x \downarrow 0$ in (3.237) and (3.238) yields

$$a + b = \lambda_1 P_0 + \lambda_2. \quad (3.239)$$

Taking $\frac{d}{dx}$ on both sides of (3.237) and letting $x \downarrow 0$ gives

$$f_1'(0) = -\lambda_1 \mu_1 P_0 + \lambda_1^2 P_0 + \lambda_1 \lambda_2 - \lambda_2 \mu_2. \quad (3.240)$$

Taking $\frac{d}{dx}$ in (3.238), letting $x \downarrow 0$, and equating to (3.240) gives

$$-(\mu_1 - \lambda_1)a - \mu_2 b = -\lambda_1 \mu_1 P_0 + \lambda_1^2 P_0 + \lambda_1 \lambda_2 - \lambda_2 \mu_2. \quad (3.241)$$

We use (3.238) and the condition $P_0 + \int_{x=0}^{\infty} f_1(x) dx = 1$ to obtain

$$P_0 + \frac{a}{\mu_1 - \lambda_1} + \frac{b}{\mu_2} = 1. \quad (3.242)$$

We now solve the system of three Eqs. (3.239), (3.241), (3.242) for P_0, a, b to obtain

$$P_0 = \frac{(\mu_2\mu_1 - \mu_2\lambda_1 - \mu_1\lambda_2)}{\mu_2\mu_1}, \tag{3.243}$$

$$a = \frac{\lambda_1(\mu_2\mu_1^2 + 2\mu_2\mu_1\lambda_1 + \mu_2^2\mu_1 - \mu_2\lambda_1^2 - \mu_2^2\lambda_1 + \mu_1^2\lambda_2 - \mu_1\lambda_2\lambda_1)}{(-\mu_1 + \lambda_1 + \mu_2)\mu_2\mu_1}, \tag{3.244}$$

$$b = \frac{\lambda_2(\mu_2 - \mu_1)}{(-\mu_1 + \lambda_1 + \mu_2)}. \tag{3.245}$$

Check on the Values of P_0, a, b
We conduct a mild check (indicated by ✓) on the values of P_0, a, b. Set $\lambda_2 = 0$. The model reverts to a standard $M_{\lambda_1}/M_{\mu_1}/1$ queue in which $f(x)$ and P_0 are given in (3.112) and (3.113), respectively.

Substituting $\lambda_2 = 0$ in (3.243), (3.244) and (3.245) yields: $P_0 = 1 - \lambda_1/\mu_1$; $a = \lambda_1\,(1 - \lambda_1/\mu_1)$; $b = 0$. ✓

3.15 M/G/1 with Server Vacations

There are many M/G/1 server-vacation models. During a server vacation the server is not available to serve customers. For example, vacations may start after each service completion, or when the server becomes idle, or both. (See, e.g., Problems 9.2 and 9.6, pp. 420–422 in [143], and see also [39] in which consecutive vacations are connected by a Markov chain.)

Here we apply LC to a basic M/G/1 server-vacation model. Let the arrival rate be λ and service time be S having cdf $B(x)$, $x > 0$. Assume that after each service completion the server goes on vacation for a time U having cdf $V(x)$, $x > 0$. During U the server may be doing required work after each service. For example, a doctor updates a record after seeing each patient, a bank teller does required paper work after serving each customer, an auto service manager fills out forms after receiving a car for service. Consider the virtual wait process $\{W(t)\}_{t\geq 0}$ (Fig. 3.32).

Denote the complementary cdf of $S + U$ by $\overline{B * V}(x)$. An integral equation for the steady-state pdf of wait $\{P_0, f(x)\}_{x\geq 0}$ is

$$f(x) = \lambda P_0\overline{B * V}(x) + \lambda\int_{y=0}^{x}\overline{B * V}(x - y)f(y)dy, x \geq 0. \tag{3.246}$$

In (3.246) the left side $f(x)$ is the SP downcrossing rate of level x. On the right side $\lambda P_0\overline{B * V}(x)$ is the SP upcrossing rate of level x, starting from

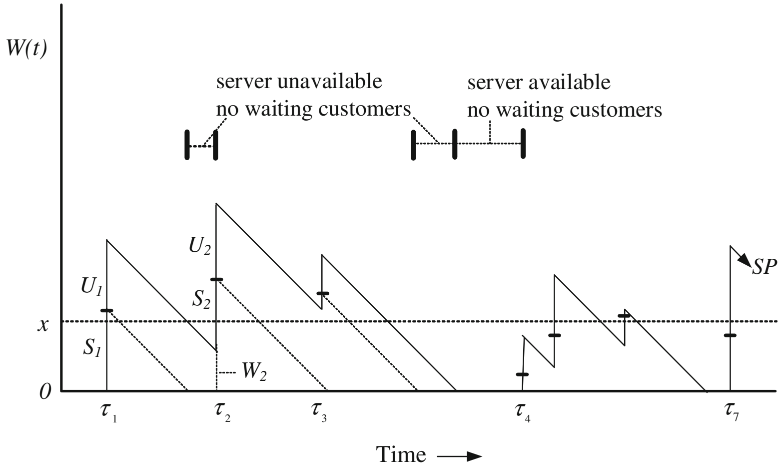


Fig. 3.32 Sample path of $\{W(t)\}_{t\geq 0}$ in M/G/1 queue with a server vacation after each service completion

level 0. The term $\lambda \int_{y=0}^{x} \overline{B * V}(x-y)f(y)dy$ is the SP upcrossing rate of level x, starting from levels in the state-space interval $(0, x)$.

Comparing (3.246) and (3.34) indicates that the server-vacation and standard M/G/1 models are equivalent with regard to the integral equation for the pdf of queue wait; only the "service time" cdfs differ.

3.15.1 Probability of Zero Wait

Since the queue behaves like $M_\lambda/G/1$ with common service time $S+U$ with respect to the customer wait until service, then

$$P_0 = 1 - \lambda E(S+U) \tag{3.247}$$

provided $\lambda E(S+U) < 1$.

3.15.2 Expected Busy and Idle Period

Define the idle period $\mathcal{I}$ as the time interval when the server is available to start service and no customers are waiting. Then $E(I) = 1/\lambda$. Let $\mathcal{B}_s$:= time that the server is busy serving customers, $\mathcal{B}_u$:= time that server is "on vacation", during a "busy period" $\mathcal{B}$, where $\mathcal{B} = \mathcal{B}_s + \mathcal{B}_u$. Then $\mathcal{B}$ is distributed as a regular busy period in a standard M$_\lambda$/G/1 queue with service time $S + U$. Applying (3.247)

$$E(\mathcal{B}) = \frac{1 - P_0}{\lambda P_0} = \frac{\lambda E(S+U)}{\lambda(1 - \lambda E(S+U))}. \tag{3.248}$$

Given the server is "busy", the pairs $\{S_i, U_i\}, i = 1, 2, \ldots$, form an alternating renewal process (Fig. 3.32). During a "busy" period, the proportion of time the server is busy serving customers $= \frac{E(S)}{E(S)+E(U)}$; "on vacation" = $\frac{E(U)}{E(S)+E(U)}$. Thus

$$E(\mathcal{B}_s) = \frac{E(S)}{E(S) + E(U)} \cdot E(\mathcal{B}),\ E(\mathcal{B}_u) = \frac{E(U)}{E(S) + E(U)} \cdot E(\mathcal{B});$$

from (3.248)

$$E(\mathcal{B}_s) = \frac{E(S)}{1 - \lambda E(S+U)},\ E(\mathcal{B}_u) = \frac{E(U)}{1 - \lambda E(S+U)}.$$

3.15.3 Number in System

Let d_n denote the probability of n customers in the system *just after the server returns from vacation*. Then (see Eq. (3.76) in Sect. 3.4.8)

$$d_n = \int_{x=0}^{\infty} \frac{e^{-\lambda x}(\lambda x)^{n-1}}{(n-1)!} f(x) dx.$$

Let a_n denote the probability that an arrival "sees" n customers in the system. Then $a_n = d_n = P_n$ due to Poisson arrivals, P_n is the long-run proportion of time there are n customers in the system.

3.15.4 M/M/1 with Server Vacations $\underset{dis}{=} Exp_\nu$

Let $\overline{V}(x) = e^{-\nu x}$, $\overline{B}(x) = e^{-\mu x}$, $x \geq 0$. Assume $\nu \neq \mu > 0$. Then

$$\overline{B * V}(x) = P(S + V > x) = \frac{\mu e^{-\nu x} - \nu e^{-\mu x}}{\mu - \nu}, x \geq 0,$$

and (3.246) reduces to

$$\begin{aligned} f(x) = &\ \lambda P_0 \tfrac{\mu e^{-\nu x} - \nu e^{-\mu x}}{\mu - \nu} \\ &+ \lambda \tfrac{1}{\mu-\nu} \int_{y=0}^{x} \left(\mu e^{-\nu(x-y)} - \nu e^{-\mu(x-y)}\right) f(y) dy, \; x \geq 0. \end{aligned} \tag{3.249}$$

In (3.249), applying differential operator $\langle D + \nu\rangle\langle D + \mu\rangle$ to both sides results in the differential equation

$$f''(x) + (\nu + \mu - \lambda) f'(x) + (\nu\mu - \lambda\mu - \lambda\nu) f(x) = 0,$$

with solution

$$f(x) = c_1 e^{R_1 x} + c_2 e^{R_2 x}, \; x \geq 0,$$

where roots R_1, R_2 are the (negative) roots of the characteristic equation

$$z^2 + (\nu + \mu - \lambda)z + (\nu\mu - \lambda\mu - \lambda\nu) = 0.$$

Applying the initial conditions $f(0) = \lambda P_0$, $f'(0) = \lambda^2 P_0$, and the normalizing condition $P_0 + \int_{y=0}^{\infty} f(x)dx = 1$ yields

$$c_1 = \lambda P_0 \frac{\lambda - R_2}{R_1 - R_2}, \qquad c_2 = -\lambda P_0 \frac{-R_1 + \lambda}{R_1 - R_2}, \qquad P_0 = \frac{c_1 R_2 + c_2 R_1 + R_1 R_2}{R_1 R_2}.$$

Busy Period

The expected values of $\mathcal{B}$, $\mathcal{B}_s$, $\mathcal{B}_u$ are

$$E(\mathcal{B}) = \frac{\frac{1}{\mu} + \frac{1}{\nu}}{1 - \lambda\left(\frac{1}{\mu} + \frac{1}{\nu}\right)}, \qquad E(\mathcal{B}_s) = \frac{\frac{1}{\mu}}{\frac{1}{\mu} + \frac{1}{\nu}} E(\mathcal{B}), \qquad E(\mathcal{B}_u) = \frac{\frac{1}{\nu}}{\frac{1}{\mu} + \frac{1}{\nu}} E(\mathcal{B}).$$

Number in System

The probability that the server finds n in the system just after a vacation is for $n = 1, 2, \ldots,$

$$\begin{aligned} d_n &= \int_{x=0}^{\infty} \frac{e^{-\lambda x}(\lambda x)^{n-1}}{(n-1)!} \left(c_1 e^{R_1 x} + c_2 e^{R_2 x}\right) dx \\ &= \frac{1}{\lambda}\left(\left(\frac{\lambda}{\lambda - R_1}\right)^n c_1 + \left(\frac{\lambda}{\lambda - R_2}\right)^n c_2\right), \end{aligned}$$

where $R_i, c_i, i = 1, 2$ are given above. The probability that an arrival "sees" n customers in the system is $a_n = d_n = P_n$.

3.16 M/G/1 with Bounded Workload

We look at three M/G/1 variants with a finite barrier $K > 0$ on the virtual wait (workload) process. These and related models (e.g., risk models with a dividend barrier in actuarial science) have been discussed widely in the literature (e.g., [78]; Example 5.5.2, p. 213 and Exercise 9.9, p. 423 in [143], and also in [25]; M/M/c queues with bounded wait in Example 1, p. 44 in [52], and also in [54, 79, 100]; and others). They are also useful in the proof of Proposition 9.1 in Sect. 9.4, Chap. 9 on level crossing estimation. As $K \to \infty$, variants 1–3 tend to a standard M/G/1 queue with infinite waiting buffer, under mild conditions. We illustrate this property with M/M/1 in Sects. 3.16.2, 3.16.4, and 3.16.6. In all three variants, we denote the arrival rate by λ; the requested full service time for each arrival, by S having cdf $B(x)$, $x > 0$, $\overline{B}(x) = 1 - B(x)$, $x \geq 0$; and the virtual wait (workload) process as $\{W_K(t)\}_{t \geq 0}$.

3.16.1 Variant 1

All customers join the system, and all waiting times (before start of service) are in $[0, K)$. Each arrival gets either full service S, or truncated service if S causes $\{W_K(t)\}_{t \geq 0}$ to exceed K, i.e., customers in service must *renege if and when* their total system time reaches K. We define the service time S_K due the level-K barrier, in terms of S as follows. If a customer must wait $y \geq 0$ then $S_K = \min(S, K - y)$. Thus for all customers, *wait* + *service time* $\leq K$. Consider a sample path of $\{W_K(t)\}_{t \geq 0}$ (Fig. 3.33). Let the mixed pdf

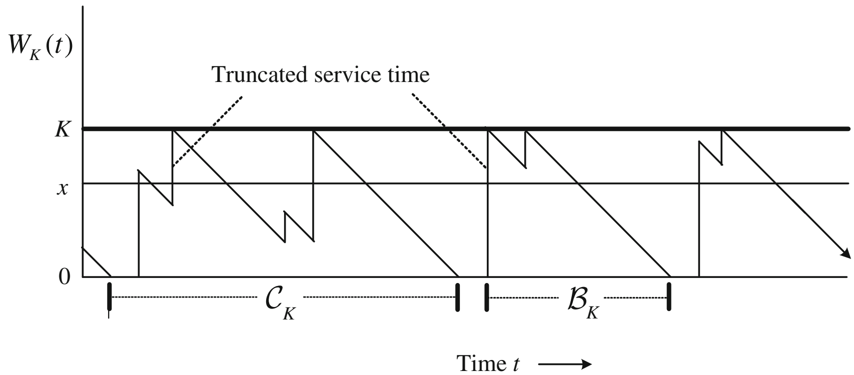


Fig. 3.33 Variant 1. Sample path of $\{W_K(t)\}_{t \geq 0}$ in M/G/1 with bounded workload. $\mathcal{C}_K$:= busy cycle, $\mathcal{B}_K$:= busy period

of wait be $\left\{P_{K,0}, f_K(x)\right\}_{x>0}$. Rate balance across level x gives immediately Eq. (3.250) for $f_K(x)$, where the left and right sides are the SP down- and upcrossing rates, respectively:

$$f_K(x) = \lambda P_{K,0}\overline{B}(x) + \lambda \int_{y=0}^{x} \overline{B}(x-y) f_K(y)dy, 0 < x < K, \quad (3.250)$$

$$P_{K,0} + \int_{y=0}^{K} f_{K,0}(x)dx = 1. \quad (3.251)$$

If $K \in (0, \infty)$ then $\left\{P_{K,0}, f_K(x)\right\}_{x>0}$ exists for all values of $\lambda > 0$ (see Sect. 2.1 in [25]). Also, [25] gives the pdf of S_K and shows the important result that $E(S_K) = \left(1 - P_{K,0}\right)/\lambda$, equivalently $P_{K,0} = 1 - \lambda E(S_K)$. (Interestingly, this is similar to $P_0 = 1 - \lambda E(S)$ in the standard *no-barrier* M/G/1 queue in steady state.) If there exists $M > 0$ such that $P_{K,0} > 0$ for all $K > M$, and we assume $\lambda E(S) < 1$ then $\left\{P_{K,0}, f_K(x)\right\}_{x>0} \to \{P_0, f(x)\}_{x>0}$ in the standard no-barrier M/G/1, since Eqs. (3.250), (3.251) would converge to Eqs. (3.34)–(3.36).

3.16.2 Variant 1: M/M/1 Model

In the $M_\lambda/M_\mu/1$ model $\overline{B}(x) = e^{-\mu x}$; the solution of (3.250) and (3.251) is

$$\left.\begin{aligned} f_K(x) &= \lambda P_{K,0} e^{-(\mu-\lambda)x}, 0 < x < K, \\ P_{K,0} &= \frac{\mu - \lambda}{\mu + e^{-(\mu-\lambda)K}}. \end{aligned}\right\} \quad (3.252)$$

If we assume $\lambda < \mu$ so that the no-barrier M/M/1 is stable, and let $K \to \infty$, then $P_{K,0} \to 1 - \lambda/\mu$ and the domain $(0, K)$ of $f_K(\cdot)$, tends to $(0, \infty)$. This results in the solution for the standard no-barrier $M_\lambda/M_\mu/1$ queue (see formulas (3.112) and (3.113)).

3.16.3 Variant 2

Upon arrival customers *balk and are cleared* if their system times would exceed K. We assume that the workload $W_K(t^-)$ and the service time S of a would-be time-t arrival are known to a "system manager" by some means. A time-t arrival joins the system only if $W_K(t^-) + S < K$. We define the service time S_K due the level-K barrier in terms of S as follows. If a customer must wait $y \geq 0$ then

$$S_K = \begin{cases} S \text{ if } y + S \le K, \\ 0 \text{ if } y + S > K. \end{cases}$$

Customers that are allowed to join receive full service S, and depart upon completing service. Consider a sample path of $\{W_K(t)\}_{t\ge 0}$ (Fig. 3.34). We obtain via LC the integral equation for $f_K(x)$:

$$\begin{aligned} f_K(x) &= \lambda P_{K,0}\left(\overline{B}(x) - \overline{B}(K)\right) \\ &\quad + \lambda \int_{y=0}^{x} \left(\overline{B}(x-y) - \overline{B}(K-y)\right) f_K(y)dy, \, 0 < x < K, \end{aligned} \tag{3.253}$$

with normalizing condition $P_{K,0} + \int_{y=0}^{K} f_K(x)dx = 1$. In (3.253), the term $\overline{B}(x) - \overline{B}(K) = P(x < S < K)$ and the term $\overline{B}(x-y) - \overline{B}(K-y)$ = $P(x - y < S < K - y)$. Using the technique in [25] for Variant 1, we can also find in Variant 2, the pdf of S_K and show that $E(S_K) = (1 - P_{K,0})/\lambda$.

3.16.4 Variant 2: M/M/1 Model

In the M$_\lambda$/M$_\mu$/1 queue with $\overline{B}(x) = e^{-\mu x}$, we obtain immediately the solution of (3.253) for $\left\{P_{K,0},\ f_K(x)\right\}_{x\in(0,K)}$ as a special case of the M/M/c queue with bounded system time. (In Example 1, p. 44 in [52], we set *number of servers* = 1.) We get

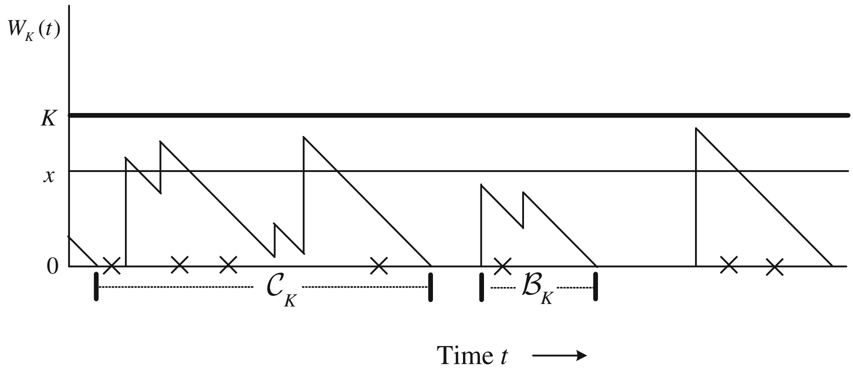


Fig. 3.34 Variant 2. Sample path of $\{W_K(t)\}_{t\ge 0}$ in M/G/1 with bounded workload. '×' indicates arrivals who balk because $wait + S > K$. $\mathcal{C}_K$:= busy cycle, $\mathcal{B}_K$:= busy period

$$\left.\begin{aligned} f_K(x) &= \lambda e^{\rho\beta} P_{K,0} e^{\mu(\rho-1)x}(1-\beta e^{\mu x})e^{-\mu\beta e^{\mu x}}, 0<x<K, \\ P_{K,0} &= \frac{1}{1+\lambda e^{\rho\beta}\int_{x=0}^{K} e^{\mu(\rho-1)x}(1-\beta e^{\mu x})e^{-\mu\beta e^{\mu x}}dx}, \end{aligned}\right\} \quad (3.254)$$

where $\rho = \lambda/\mu$, $\beta = e^{-\mu K}$. The solution in (3.254) checks with the single-server Markovian result obtained in [78], and is more complex than the solution (3.252) for variant 1.

If $K \to \infty$ then $\beta \downarrow 0$. Additionally, if $\lambda < \mu$ then (3.254) becomes

$$f(x) = \lambda P_0 e^{-(\mu-\lambda)x}, x > 0, \qquad P_0 = 1 - \frac{\lambda}{\mu},$$

as in the standard no-barrier M/M/1 queue.

3.16.5 Variant 3

All arrivals that "see" a *wait* $< K$ join the system and receive full service S. Some of these service tines will cause jumps that upcross level K (Fig. 3.35). (In variant 3 we call level K a *threshold* rather than a barrier, because sample-path conditions switch at level K.) Arrivals that "see" a wait $> K$, are blocked from joining, and are cleared. (Effectively, they balk upon arrival. With respect to the arrival-point waiting time Variant 3 is identical to M/G/1 with reneging/balking and having a *staying function*

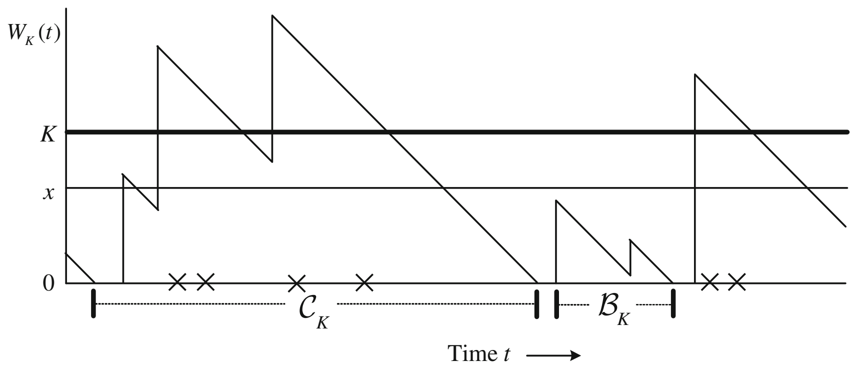


Fig. 3.35 Variant 3. Sample path of $\{W_K(t)\}_{t\geq 0}$ in M/G/1 with threshold at level K. '$\times$' indicates arrivals who balk because $W_K(\cdot) > K$ upon their arrival. $\mathcal{C}_K$:= busy cycle, $\mathcal{B}_K$:= busy period

$\overline{R}(y) = 1 \cdot \boldsymbol{I}_{[0,k)}\ (W_\iota) + 0.\boldsymbol{I}_{[k,\infty)}\ (W_\iota)$, where W_ι := arrival-point wait ... see Fig. 3.35 and Sect. 3.13, which analyzes the renege/balk M/G/1 queue. We define the service time S_K due to the level-K threshold, in terms of S as follows. If a customer must wait y then

$$S_K = \begin{cases} S \text{ if } y \in [0, K)\,, \\ 0 \text{ if } y \in [K, \infty)\,, \end{cases}$$

which may be written as $S_K = S \cdot \boldsymbol{I}_{[0,k)}\ (W_\iota) + 0.\boldsymbol{I}_{[k,\infty)}\ (W_\iota)$, where $\boldsymbol{I}_{\boldsymbol{A}}(\cdot)$ is the characteristic function of set $\boldsymbol{A}$.

We denote the mixed pdf of wait as $\left\{P_{K,0},\ f_{K,i}(x)\right\}_{i=0,1}$ where the domain of $f_{K,0}(x)$ is $(0, K)$ and the domain of $f_{K,1}(x)$ is $[K, \infty)$. Using LC we can write integral equations for $f_{K,i}(x)$, $i = 1, 2$, by inspection of Fig. 3.35, as follows.

$$\left.\begin{aligned} f_{K,0}(x) &= \lambda P_{K,0}\overline{B}(x) + \lambda \textstyle\int_{y=0}^{x} \overline{B}(x-y) f_{K,0}(y)dy,\ x \in (0, K)\,, \\ f_{K,1}(x) &= \lambda P_{K,0}\overline{B}(x) + \lambda \textstyle\int_{y=0}^{K} \overline{B}(x-y) f_{K,0}(y)dy,\ x \in [K, \infty)\,, \\ P_{K,0} &+ \textstyle\int_{y=0}^{K} f_{K,0}(x)dx + \int_{x=K}^{\infty} f_{K,1}(x)dx = 1. \end{aligned}\right\} \tag{3.255}$$

We infer from Fig. 3.35 and Theorem 1.1, the continuity condition at K

$$f_{K,1}(K^+) = f_{K,0}(K^-), \tag{3.256}$$

noting $\lim_{t\to\infty} D_t(x)/t = \lim_{t\to\infty} D_t(x^-)/t$, and there are no SP tangents at level K. (Contrast this property with that at level D in M/D/1 where there is a *discontinuity*; see Proposition 3.9 Part (2) **in** Sect. 3.10.1.)

Expected Sojourn Above Level K

Let γ_K := *excess of a jump* over level K, a_K := *sojourn* above level K. Then $a_K = \gamma_K$, and $E(a_K) = E(\gamma_K)$. Let $F_{\gamma_K}(z) := P(\gamma_K \le z)$, $z > 0$. Two different expressions for $\lim_{t\to\infty} \mathcal{U}_t(K+z)/t$ are

$$\left(1 - F_{\gamma_K}(z)\right) f_{K,0}(K^-)$$

$$\text{and} \qquad \lambda P_{K,0}\overline{B}(K+z) + \lambda \int_{y=0}^{K} \overline{B}(K+z-y) f_{K,0}(y)dy.$$

In the first expression $f_{K,0}(K^-)$ is the upcrossing rate (also the downcrossing rate) of level K, and $1 - F_{\gamma_K}(z)$ is the upcrossing rate of level $K + z$ given the SP upcrosses level K. The second term is the upcrossing rate of level $K + z$ due to upward jumps that start in $[0, K)$. Thus

$$
\begin{aligned}
1 - F_{\gamma_K}(z) &= \frac{\lambda P_{K,0}\overline{B}(K+z) + \lambda \int_{y=0}^{K} \overline{B}(K+z-y) f_{K,0}(y)dy}{f_K, 0(K^-)} \\
&= \frac{\lambda P_{K,0}\overline{B}(K+z) + \lambda \int_{y=0}^{K} \overline{B}(K+z-y) f_{K,0}(y)dy}{\lambda P_{K,0}\overline{B}(K) + \lambda \int_{y=0}^{K} \overline{B}(K-y) f_{K,0}(y)dy}
\end{aligned}
$$

and

$$
\begin{aligned}
E(a_K) &= E(\gamma_K) = \int_{z=0}^{\infty} \left(1 - F_{\gamma_K}(z)\right) dz \\
&= \int_{z=0}^{\infty} \left[\frac{\lambda P_{K,0}\overline{B}(K+z) + \lambda \int_{y=0}^{K} \overline{B}(K+z-y) f_{K,0}(y)dy}{\lambda P_{K,0}\overline{B}(K) + \lambda \int_{y=0}^{K} \overline{B}(K-y) f_{K,0}(y)dy} \right] dz.
\end{aligned}
\tag{3.257}
$$

Using the technique in [25] for Variant 1, we can also find in Variant 3, the pdf of S_K and show that $E(S_K) = (1 - P_{K,0})/\lambda$.

3.16.6 Variant 3: M/M/1 Model

Setting $\overline{B}(x) = e^{-\mu x}$ in (3.255), and solving by converting to differential equations, gives

$$
\left.
\begin{aligned}
f_{K,0}(x) &= \lambda P_{k,0} e^{-(\mu-\lambda)x}, \, x \in (0, K)\,, \\
f_{K,1}(x) &= \lambda P_{k,0} e^{-(\mu x-\lambda K)}, \, x \in [K, \infty)\,, \\
P_{K,0} &= \frac{1}{1 + \frac{\lambda}{\mu-\lambda}(1 - e^{-(\mu-\lambda)K}) + \frac{\lambda}{\mu} e^{-(\mu-\lambda)K}}.
\end{aligned}
\right\}
\tag{3.258}
$$

In (3.258) if $x > (\lambda K)/\mu$ then $\mu x - \lambda K > 0$ and $\int_{x=K}^{\infty} f_{K,1}(x)dx$ is finite. If additionally $\lambda < \mu$ then as $K \to \infty$ the denominator of $P_{K,0} \to \frac{1}{1+\lambda/(\mu-\lambda)}$ = $1 - \lambda/\mu$, which is P_0 in the no-threshold M/M/1 queue. Also $f_{K,0}(x) \to \lambda P_0 e^{-(\mu-\lambda)x}$, $x \in (0, \infty)$ which is $f(x)$, $x > 0$ in the no-threshold M/M/1 queue.

From (3.257), $E(a_K) = \int_{z=0}^{\infty} e^{-\mu z} dz = 1/\mu$.

3.17 Generalized Beneš Series for PDF of Wait

In this Section we use LC to generalize the Beneš series for the pdf of wait in M/G/1 (see formula (3.71) in Sect. 3.4.5). We use LC, the busy-period structure (Fig. 3.6 in Sect. 3.4.12), the multiplicative structure (Fig. 3.36), and the ***renewal reward theorem*** (see references following Eq. (3.79)) to develop a *series* for the pdf of wait (W_q). Combining LC and the renewal reward theorem facilitates creating more general series for the pdf of W_q in MG/1 variants as well. We illustrate the more generalized series in an M/G/1 model where zero-wait arrivals receive exceptional service (see Sect. 3.6.1).

3.17.1 Model Description

The arrival rate is λ. Zero-wait arrivals (initiators of busy periods) receive service time S_0. Positive-wait arrivals receive service time S_1 ($\neq_{dis} S_0$). We denote: the cdf of S_i as $B_i(x)$, $x > 0$, $\overline{B}_i(x) = 1 - B_i(x)$, $x \geq 0$, $i = 0, 1$; the steady-state pdf of wait as $\{P_0, f(x)\}_{x>0}$; the limiting excess of S_i as γ_{S_i};; *the pdf* of γ_{S_i} by $g_i(x)$, $x \in (0, \infty) \cap (domain\ of\ S_i)$. It is well known that $g_i(x) = (1/E(S_i))\,\overline{B}_i(x)$, $i = 0, 1$ (see Example 7.24, p. 453 in [125] ; formula (6.2), p. 193 in [99]). Also $\rho_i := \lambda E(S_i)$, $i = 0, 1$.

Examining a busy period of the virtual wait process $\{W(t)\}_{t\geq 0}$ (Fig. 3.36) and applying LC rate balance across level x (>0), yields Eq. (3.143) of Sect. 3.6.1 (repeated here for handy reference)

$$f(x) = \lambda P_0 \overline{B}_0(x) + \lambda \int_{y=0}^{x} \overline{B}_1(x-y) f(y) dy, x > 0. \tag{3.259}$$

Integrating both sides of (3.259) with respect to $x \in (0, \infty)$ and simplifying leads to formula (3.144) for P_0, whose form implies $P_0 \in (0, 1)$ iff $\rho_1 < 1$.

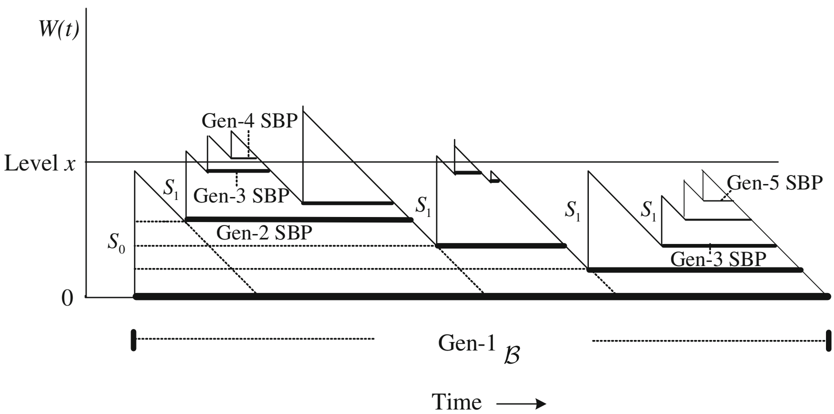


Fig. 3.36 Multiplicative structure of $\mathcal{B}$ of $\{W(t)\}_{t\geq 0}$ for Beneš series analysis. Each arrival generates the initial jump of a $\mathcal{B}$ or sub-busy period (SBP). Initial jumps of all busy/sub-busy periods account for all arrivals (Gen := generation)

3.17.2 Applying the Renewal Reward Theorem

Consider the gen-1 (abbreviation for generation-1) busy period $\mathcal{B}$ in Fig. 3.36. Fix level $x > 0$. Let: $E(\mathcal{U}_{\text{gen-}k}(x)) := E$(number of upcrossings of x by gen-k initiated jumps in $\mathcal{B}$); $E(\mathcal{U}_{\text{gen-}k,t}(x)) := E$(number of upcrossings of x by gen-k initiated jumps *during* $(0, t)$). Then

$$E(\mathcal{U}_{\text{gen-1}}(x)) = 1 \times \overline{B}_0(x);\ E(\mathcal{C}) = 1/(\lambda P_0)$$

(for $E(\mathcal{C})$ see formula (3.81) in Sect. 3.4.10). All jumps during $\mathcal{C}$ (busy cycle) occur during its embedded $\mathcal{B}$. By the *renewal reward theorem*, the *long-run* upcrossing rate of x due to gen-1 busy period initiated jumps is

$$\begin{aligned} \frac{\lim_{t\to\infty}\mathcal{U}_{\text{gen-1},t}(x)}{t} &= \frac{E(\mathcal{U}_{\text{gen-1}}(x))}{E(\mathcal{C})} \\ = \frac{\overline{B}_0(x)}{1/(\lambda P_0)} &= \lambda P_0 \overline{B}_0(x) = P_0\rho_0\overline{g}_0(x). \end{aligned} \tag{3.260}$$

A similar analysis of gen-2 sub-busy period initiated upcrossings of x gives

$$E(\mathcal{U}_{\text{gen-2}}(x)) = \lambda E(S_0)\int_{y=0}^{x} \overline{B}_1(x-y)g_0(y)dy,$$

since E(*number of gen-2 jumps during C – same as number in* $\mathcal{B}$) = $\lambda E(S_0)$. Thus

$$\begin{aligned} E(\mathcal{U}_{\text{gen-2}}(x)) &= (1/\mu_1)\lambda E(S_0)\int_{y=0}^{x} \mu_1\overline{B}_1(x-y)g_0(y)dy \\ &= (1/\mu_1)\rho_0\left(g_{1(1)} * g_0\right)(x). \end{aligned}$$

By the *renewal reward theorem*

$$\begin{aligned} \frac{\lim_{t\to\infty}\mathcal{U}_{\text{gen-2},t}(x)}{t} &= \frac{E(\mathcal{U}_{\text{gen-2}}(x))}{E(\mathcal{C})} \\ = \frac{(1/\mu_1)\rho_0\left(g_{1(1)}*g_0\right)(x)}{1/(\lambda P_0)} &= P_0\rho_0\rho_1\left(g_{1(1)} * g_0\right)(x), \end{aligned} \tag{3.261}$$

where $\left(g_{1(1)} * g_0\right)(x) = \int_{y=0}^{x} g_{1(1)}(x-y)g_0(y)dy$.

Similarly, for gen-3 sub-busy period initiated upcrossings of x,

$$E(\mathcal{U}_{\text{gen-3}}(x)) = (\lambda E(S_1))(\lambda E(S_0)) \int_{y=0}^{x} \overline{B}_1(x-y)\left(g_{1(1)} * g_0\right)(y)dy$$
$$= (1/\mu_1)(\lambda E(S_1))(\lambda E(S_0)) \int_{y=0}^{x} \mu_1 \overline{B}_1(x-y)\left(g_{1(1)} * g_0\right)(y)dy$$
$$= (1/\mu_1)\rho_0\rho_1\left(g_{1(2)} * g_0\right)(x);$$

the factor $\lambda E(S_1)$ occurs because each gen-2 sub-busy period initiated jump is $\underset{dis}{=} S_1$—the initial service time of a gen-3 sub-busy period. By the renewal reward theorem

$$\frac{\lim_{t\to\infty}\mathcal{U}_{\text{gen-3},t}(x)}{t} = \frac{E(\mathcal{U}_{\text{gen-3}}(x))}{E(\mathcal{C})}$$
$$= \frac{(1/\mu_1)\rho_0\rho_1\left(g_{1(2)} * g_0\right)(x)}{1/(\lambda P_0)} = P_0\rho_0\rho_1^2\left(g_{1(2)} * g_0\right)(x),\ x > 0.$$

Similar reasoning for gen-k sub-busy period initiated upcrossings of x, yields

$$\frac{\lim_{t\to\infty}\mathcal{U}_{\text{gen-}k,t}(x)}{t} = \frac{E(\mathcal{U}_{\text{gen-}k}(x))}{E(\mathcal{C})}$$
$$= P_0\rho_0\rho_1^{k-1}\left(g_{1(k-1)} * g_0\right)(x),\ k = 1, 2, \ldots, \tag{3.262}$$

where $g_{1(k-1)}(\cdot)$ is the $(k-1)$-*fold self-convolution of* $g_1(\cdot)$, and $g_{1(0)} \equiv 1$.

The principle of rate balance across level x gives

$$\frac{\lim_{t\to\infty}\mathcal{D}_t(x)}{t} = \sum_{k=1}^{\infty}\frac{\lim_{t\to\infty}\mathcal{U}_{\text{gen-}k,t}(x)}{t},\ x > 0,$$
$$f(x) = P_0\rho_0\sum_{k=1}^{\infty}\rho_1^{k-1}\left(g_{1(k-1)} * g_0\right)(x),\ x > 0, \tag{3.263}$$

upon applying formula (3.262). In (3.263) the right side is the total upcrossing rate of level x; term 1 is the upcrossing rate of x due to gen-1 busy period initiated jumps, and term k is the upcrossing rate of x due to gen-$(k-1)$ sub-busy period initiated jumps, $k = 2, 3, \ldots$.

3.17.3 LC Equation for $\{P_0, f(x)\}_{x\geq 0}$ via a Series

In (3.263) term k is the SP upcrossing rate of level x due to the gen-k busy/sub-busy period initiated jumps, where $\left(g_{1(0)} * g_0\right)(x) \equiv g_0(x)$. From Fig. 3.36

every arrival is the initiator of some gen-1 busy period or some gen-k sub-busy period, $k = 2, 3, \ldots$ Hence, the initial jumps of all the gen-k busy/sub-busy periods, $k = 1, 2, \ldots$, *account for all arrivals to the system*. In (3.263) the left and right sides are the SP down- and upcrossing rates of level x, respectively. Hence, (3.263) is an *alternative* way of viewing the LC balance equation for $f(x)$. Due to the geometric factors ρ_1^{k-1}, $k = 1, 2, \ldots$, ($\rho_1 < 1$), the series converges geometrically fast, to $f(x)$. Formula (3.263) is a series solution of the standard Volterra integral equation for the pdf given by (3.259). Moreover, because (3.263) is the sum of gen-k initiated upcrossing rates of level x, (3.263) is an alternative LC equation for $\{P_0, f(x)\}_{x>0}$. (In fact, the right side of (3.259) is the series expansion of the integral in (3.259)). Interestingly, we now have a geometric/physical interpretation of each term via LC. By computing or approximating the convolutions $\left(g_{1(k-1)} * g_0\right)(x)$, $k = 1, 2, \ldots$, we can quickly estimate $f(x)$ by summing the first N appropriate terms of (3.263).

In the standard M/G/1 queue, $g_0(x) \equiv g_{1(k-1)(x)}$ and the series (3.263) simplifies to the well-known Beneš series (3.71) (see [8]; formula (5.111), p. 201 in [104]).

Example 3.12 In M/M/1 where zero-wait arrivals get exceptional service $g_i(y) = \mu_i e^{-\mu_i y}$, and $E(S_i)$ =$1/\mu_i$, $\rho_i = \lambda/\mu_i$, $i = 0, 1$. Then

$$g_0(y) = e^{-\mu_0 y}\mu_0 \equiv g_{1(0)}, \qquad g_{1(k-1)}(y) = \frac{e^{-\mu_1 y}\,(\mu_1 y)^{k-2}\,\mu_1}{(k-2)!}, k = 2, 3, \ldots,$$

so that

$$\begin{aligned} &\left(g_{1(k-1)} * g_0\right)(x) \\ &= \begin{cases} e^{-\mu_0 x}\mu_0, \ k = 1, \\ \int_{y=0}^{x} \frac{e^{-\mu_1 (x-y)}(\mu_1 (x-y))^{k-2}\mu_1}{(k-2)!} \cdot e^{-\mu_0 y}\mu_0 dy, \ k = 2, 3, \ldots \end{cases} \end{aligned} \tag{3.264}$$

where $k-2 := (k-1)-1$ (see formula (3.39) for the pdf of $\mathrm{Erl}_{k,\mu}$). Substituting from (3.264) into (3.263) gives the first term of the series as

$$P_0\rho_0 e^{-\mu_0 x}\mu_0 = P_0\lambda e^{-\mu_0 x}.$$

The sum of the subsequent terms of the series is

$$P_0\rho_0 \sum_{k=2}^{\infty} \rho_1^{k-1} \int_{y=0}^{x} \frac{e^{-\mu_1(x-y)}(\mu_1(x-y))^{k-2}\mu_1}{(k-2)!} \cdot e^{-\mu_0(y)}\mu_0 dy$$
$$= P_0\rho_0 e^{-\mu_1 x} \int_{y=0}^{x} \lambda\mu_0 e^{\mu_1 y} \cdot \sum_{k=2}^{\infty} \frac{(\lambda(x-y))^{k-2}}{(k-2)!} \cdot e^{-\mu_0 y} dy$$
$$= P_0\rho_0 e^{-\mu_1 x} \int_{y=0}^{x} \lambda\mu_0 e^{\mu_1 y} \cdot e^{\lambda(x-y)} \cdot e^{-\mu_0 y} dy$$
$$= P_0\rho_0\lambda\mu_0 e^{-(\mu_1-\lambda)x} \int_{y=0}^{x} e^{(\mu_1-\lambda-\mu_0)y} d\dot{y}$$
$$= P_0\lambda^2 \frac{e^{-\mu_0 x} - e^{-(\mu_1-\lambda)x}}{\mu_1-\lambda-\mu_0}.$$

Summing all the terms gives

$$f(x) = P_0\lambda e^{-\mu_0 x} + P_0\lambda^2 \frac{e^{-\mu_0 x} - e^{-(\mu_1-\lambda)x}}{\mu_1-\lambda-\mu_0}$$
$$= P_0 \left(\frac{-\lambda^2}{\mu_1-\lambda-\mu_0} e^{-(\mu_1-\lambda)x} + \frac{\lambda(\mu_1-\mu_0)}{\mu_1-\lambda-\mu_0} e^{-\mu_0 x} \right),$$
$$P_0 = \frac{\mu_1-\lambda}{\mu_1-\lambda-\mu_0},$$

which is identical to formulas (3.148) and (3.149) in Sect. 3.6.1, which were obtained by converting an integral equation to a differential equation, solving the latter, and then using initial conditions to obtain the constants of integration.

The foregoing example illustrates important properties of the level crossing method.

1. We can partition the sample-path jumps of $\{W(t)\}_{t\geq 0}$ into subsets, such as jumps that initiate generation-k sub-busy periods, in order to obtain new views of the queueing kinetics directly from the structure of the sample path. In this Section the partition into gen-k jumps results in a generalization of the Beneš series for M/G/1.
2. Once the convolutions in the series are specified, it is straightforward in many cases to derive the pdf $f(x)$, $x > 0$. Comparing the above example with the solution method for $f(x)$, $x > 0$ in Sect. 3.6.1 shows that the LC-derived generalized Beneš series approach is more straightforward, and computes the coefficients of $e^{-(\mu_1-\lambda)x}$ and $e^{-\mu_0 x}$ directly without resorting to differential equations and using initial conditions.

3.17.4 Brief Discussion

We have indicated how to apply LC to derive transient and steady-state properties of the waiting time in several M/G/1 and M/M/1 queues, emphasizing steady-state results. Many of the LC-derived properties have been obtained in the literature by different methods, but some properties and results given in Chap. 3 are new.

A vast array of additional models and variants have been analyzed using LC, since 1976. For example, M/G/1 queues with Markov-generated server vacations [39] generalizes the standard M/G/1 server-vacation model. The vacation time following a service completion depends on the length of the immediately preceding *vacation*, via a Markov chain. Such dependency arises in many situations. A teller in a bank may do paper work following each service. After the next service completion, the paper work required may depend on the quality and quantity of the paperwork completed during the preceding vacation. Similar remarks apply to workers who write a report after completing a service, e.g., medical practitioners after seeing a patient; dentists after treating a patient; repairmen after completing a job; salesmen after completing a sale; and so forth.

Variants of the M/G$^{(a,b)}$/1 queue with bulk service were analyzed using LC in [20, 93]. The model utilizes a two-dimensional state $\{W(t), M(t)\}_{t\geq 0}$ where $W(t)$ is the virtual wait. Random variable $M(t)$ is discrete; it represents at time t, the number of 'customers in the waiting line confirmed' to form a service group, where $M(t) \in \{a, \ldots, b\}$ and b is the maximum service-group size. $M(t)$ is called the *system configuration*, which is explained for M/M/c queues in Sects. 4.4 and 4.5 in Chap. 4. The idea of *system configuration* was introduced by the author in [11] (see also Sect. 2 in [52]). System configurations are very useful in many stochastic models, by giving the LC method much flexibility for modelling various situations (see, e.g., the effective system configuration due to L. Green in [38]). A system configuration introduces sufficient detailed information, to make a model Markov ian. Creating a useful system configuration requires thinking through the system dynamics carefully.