Chapter 4
M/M/c Queue

4.1 Introduction

In Sect. 4.2 below we prove a useful general result (which we call Theorem B) about SP transitions. This theorem facilitates the analysis of *transient* distributions of state variables, and will be applied variously in the sequel.

Sections 4.3–4.5 explain the sample-path structure and dynamics of the generalized M/M/c queue. In the generalized model the SP can make a *transition* between disjoint state-space *sets* (called *pages* or *sheets*). Geometrically, sheets are analogous to a package of sheets of paper, cards in a deck, or pages of a book. They form a discrete number of disjoint subsets of the state space, not connected by a continuous segment of the sample path. (We can also model complex *single-server* queues using the ***method of sheets*** (aka ***method of pages***) (see, e.g., [39, 53, 93]). The *method of sheets* provides LC with great flexibility to analyze different types of stochastic models: e.g., queues; dams (see Sect. 11.8); inventories; production-inventories; actuarial risk models, replacement models; models in the natural sciences, etc.

Sections 4.6.1–4.6.9 develop equations for transient and steady-state pdfs of wait in the generalized M/M/c model. Sections 4.7–4.12 provide steady-state analyses of M/M/c variants using LC. In particular, Sect. 4.8 derives known results for the standard M/M/c queue as a special case of the generalized model. The remaining Sections of the Chapter study variants of M/M/c queues. All Sections provide empirical background for potentially novel applications of LC.

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_4

4.2 Theorem B for Transient Analysis

We state and prove Theorem B. This straightforward theorem facilitates the transient analyses of a variety of stochastic models.

4.2.1 Theorem B

We first give a fundamental generalization of Theorems 3.1 and 3.2 of Chap. 3, which is useful for LC derivations of integro-differential equations for transient distributions in general.

Let $\{X(t)\}_{t\geq 0}$ denote a sample path of a general stochastic process with state space $\boldsymbol{S}$. Let $\boldsymbol{A}$, $\boldsymbol{B}$ be arbitrary measurable subsets of $\boldsymbol{S}$. Denote the transient probability $P(X(t) \in \boldsymbol{A})$ at instant t by $P_t(\boldsymbol{A})$, $t \geq 0$. Let $P_{t_1,t_2}(\boldsymbol{A}, \boldsymbol{B})$ be the joint probability $P(X(t_1) \in \boldsymbol{A},\ X(t_2) \in \boldsymbol{B})$ at instants $t_1, t_2 \geq 0$. Let $\mathcal{I}_t(\boldsymbol{A})$ be the number of SP *entrances* and $\mathcal{O}_t(\boldsymbol{A})$ the number of SP *exits* of $\boldsymbol{A}$, during $(0, t)$ (see Fig. 2.7). Assume the derivatives

$$\frac{\partial}{\partial t} P_t(\boldsymbol{A}),\ \frac{\partial}{\partial t} E(\mathcal{I}_t(\boldsymbol{A})),\ \frac{\partial}{\partial t} E(\mathcal{O}_t(\boldsymbol{A}))$$

exist for all $t > 0$. Both

$$\frac{\partial}{\partial t} E(\mathcal{I}_t(\boldsymbol{A})) > 0 \text{ and } \frac{\partial}{\partial t} E(\mathcal{O}_t(\boldsymbol{A})) > 0$$

hold wherever the derivatives exist, since $\mathcal{I}_t(\boldsymbol{A})$ and $\mathcal{O}_t(\boldsymbol{A})$ are counting processes which increase (wide sense, i.e., not strictly; they may be step functions) as t increases. The following useful result holds.

Theorem 4.1 *Theorem B* (P.H. Brill, 1983)

$$E(\mathcal{I}_t(\boldsymbol{A})) = E(\mathcal{O}_t(\boldsymbol{A})) + P_t(\boldsymbol{A}) - P_0(\boldsymbol{A}) \tag{4.1}$$

$$\frac{\partial}{\partial t} E(\mathcal{I}_t(\boldsymbol{A})) = \frac{\partial}{\partial t} E(\mathcal{O}_t(\boldsymbol{A})) + \frac{\partial}{\partial t} P_t(\boldsymbol{A}). \tag{4.2}$$

Proof We give two proofs in order to develop intuition about the result.

Proof 1: This proof is similar to that of Theorems 3.1 and 3.2 in Sect. 3.2.3 above. We make the correspondence:

$$\boldsymbol{A} \leftrightarrow (-\infty, x], \quad \mathcal{I}_t(\boldsymbol{A}) \leftrightarrow \mathcal{D}_t(x), \quad \mathcal{O}_t(\boldsymbol{A}) \leftrightarrow \mathcal{U}_t(x),$$
$$P_t(\boldsymbol{A}) \leftrightarrow F_t(x), t \geq 0, \quad P_{t_1,t_2}(\boldsymbol{A}, \boldsymbol{A}) \leftrightarrow F_{t_1,t_2}(x, x), t_1, t_2 \geq 0.$$

SP down- and upcrossings of level x are entrances and exits of sets (Definitions 2.2–2.5). Note that

$$\begin{aligned}
\mathcal{I}_t(\boldsymbol{A}) - \mathcal{O}_t(\boldsymbol{A}) = +1 &\iff X(0) \in \boldsymbol{A}^c,\ X(t) \in \boldsymbol{A},\\
\mathcal{I}_t(\boldsymbol{A}) - \mathcal{O}_t(\boldsymbol{A}) = -1 &\iff X(0) \in \boldsymbol{A},\ X(t) \in \boldsymbol{A}^c,\\
\mathcal{I}_t(\boldsymbol{A}) - \mathcal{O}_t(\boldsymbol{A}) = 0 &\iff X(0) \in \boldsymbol{A},\ X(t) \in \boldsymbol{A}\\
&\qquad\qquad \text{or } X(0) \in \boldsymbol{A}^c,\ X(t) \in \boldsymbol{A}^c.
\end{aligned}$$

We thus obtain the following values and corresponding probabilities:

$\mathcal{I}_t(\boldsymbol{A}) - \mathcal{O}_t(\boldsymbol{A})$	Probability
$+1$	$P_{0,t}(\boldsymbol{A}^c, \boldsymbol{A}) = P_t(\boldsymbol{A}) - P_{0,t}(\boldsymbol{A}, \boldsymbol{A})$
-1	$P_{0,t}(\boldsymbol{A}, \boldsymbol{A}^c) = P_0(\boldsymbol{A}) - P_{0,t}(\boldsymbol{A}, \boldsymbol{A})$
0	$1 - P_t(\boldsymbol{A}) - P_0(\boldsymbol{A}) + 2P_{0,t}(\boldsymbol{A}, \boldsymbol{A})$

Taking the expected value $E(\mathcal{I}_t(A) - \mathcal{O}_t(A))$ and then the derivative $\frac{\partial}{\partial t} E(\mathcal{I}_t(A) - \mathcal{O}_t(A))$ yields (4.1) and (4.2).

Proof 2: Fix $t \geq 0$. The probability of the sure event $\boldsymbol{S}$ is

$$P_t(\boldsymbol{S}) = P_t(\boldsymbol{A} \cup \boldsymbol{A}^c) = P_t(\boldsymbol{A}) + P_t(\boldsymbol{A}^c) = 1.$$

Consider $P_{t_1,t_2}(\boldsymbol{A}, \boldsymbol{S})$. Events $\{X(t_1) \in \boldsymbol{A}\}$ and $\{X(t_2) \in \boldsymbol{S}\}$ are independent for every $0 \leq t_1 \neq t_2$. Knowledge that $\{X(t_1) \in \boldsymbol{A}\}$ has occurred, does not effect the probability of event $\{X(t_2) \in \boldsymbol{S}\}$, which is $P_{t_2}(\mathbf{S}) = 1$, and vice versa. Similarly, the events $\{X(t_1) \in \boldsymbol{S}\}$ and $\{X(t_2) \in \boldsymbol{B}\}$ are independent. Note that $\boldsymbol{S} = \boldsymbol{A} \cup \boldsymbol{A}^c = \boldsymbol{B} \cup \boldsymbol{B}^c$. Hence

$$\left.\begin{aligned}
P_{t_1}(\boldsymbol{A}) &= P_{t_1,t_2}(\boldsymbol{A}, \mathbf{S}) = P_{t_1,t_2}(\boldsymbol{A}, \boldsymbol{B} \cup \boldsymbol{B}^c),\\
P_{t_2}(\boldsymbol{B}) &= P_{t_1,t_2}(\boldsymbol{S}, \boldsymbol{B}) = P_{t_1,t_2}(\boldsymbol{A} \cup \boldsymbol{A}^c, \boldsymbol{B}),
\end{aligned}\right\}$$

or

$$\left.\begin{aligned}
P_{t_1}(\boldsymbol{A}) &= P_{t_1,t_2}(\boldsymbol{A}, \boldsymbol{B}) + P_{t_1,t_2}(\boldsymbol{A}, \boldsymbol{B}^c),\\
P_{t_2}(\boldsymbol{B}) &= P_{t_1,t_2}(\boldsymbol{A}, \boldsymbol{B}) + P_{t_1,t_2}(\boldsymbol{A}^c, \boldsymbol{B}).
\end{aligned}\right\} \tag{4.3}$$

The possible values of $\mathcal{I}_t(\mathbf{A}) - \mathcal{O}_t(\mathbf{A})$ and corresponding joint probabilities at time points $t_1 = 0$ and $t_2 = t > 0$ are:

$\mathcal{I}_t(\mathbf{A}) - \mathcal{O}_t(\mathbf{A})$	Probability
0	$P_{0,t}(\boldsymbol{A}, \boldsymbol{A}) + P_{0,t}(\boldsymbol{A}^c, \boldsymbol{A}^c)$
$+1$	$P_{0,t}(\boldsymbol{A}^c, \boldsymbol{A})$
-1	$P_{0,t}(\boldsymbol{A}, \boldsymbol{A}^c)$

(4.4)

Taking the expected value of $\mathcal{I}_t(\mathbf{A}) - \mathcal{O}_t(\mathbf{A})$ in (4.4) yields

$$\begin{aligned} E\left(\mathcal{I}_t(\mathbf{A}) - \mathcal{O}_t(\mathbf{A})\right) &= P_{0,t}(\boldsymbol{A}^c, \boldsymbol{A}) - P_{0,t}(\boldsymbol{A}, \boldsymbol{A}^c) \\ &= P_{0,t}(\boldsymbol{A}^c, \boldsymbol{A}) + P_{0,t}(\boldsymbol{A}, \boldsymbol{A}) \\ &\quad - \left(P_{0,t}(\boldsymbol{A}, \boldsymbol{A}) + P_{0,t}(\boldsymbol{A}, \boldsymbol{A}^c)\right) \\ &= P_t(\boldsymbol{A}) - P_0(\boldsymbol{A}), \end{aligned}$$

which gives (4.1). Taking $\partial/\partial t$ in (4.1) yields (4.2). ■

Remark 4.1 Theorem B also applies to multi-dimensional processes with state space $\boldsymbol{S} \subseteq \mathbb{R}^n$, $n = 2, \ldots$, whose states are described by more than one continuous random variable. (**Note**: The symbol $\mathbb{R}$ denotes the set of real numbers.) We analyze two multi-dimensional inventory models in steady state, in Chap. 7.

4.3 Generalized M/M/c Queue

Customers arrive at an M/M/c queue in a Poisson stream at rate λ. There is one waiting line and c servers. Arrivals start service from the first available server, in order of arrival. We assume that for each arrival, the service time is exponentially distributed with rate selected from a nonempty set $\boldsymbol{\mu} = \{\mu_0, \ldots, \mu_J\}$ of $J + 1$ positive constants, depending on a *server-assignment policy* specified for the model; this allows service rates to be state dependent. Thus the standard M/M/c queue is a special case (see Sect. 4.8 below; p. 66ff in [84]).

For the generalized M/M/c queue we use a 'partition/synthesis' technique. We partition the state space into zero-wait and positive-wait states, and analyze the partitioned states to obtain 'partial' pdfs of the waiting time using LC. Then we synthesize those results to obtain the 'total' pdf of the waiting time, and related quantities of interest.

We next discuss: virtual wait; server workload; system configuration; the system point process (SP process); and give examples. (References for this section are [11] and [52], and others cited below.)

4.3.1 Virtual Wait and Server Workload

Notation 4.2 In the remainder of Sect. 4.3 we use two symbols for customers arriving to the system, depending on the context. (1) C (t) denotes a **would-**

be (potential) time-t arrival, $t \geq 0$. (2) $C_{a,t}$ denotes an **actual** time-t arrival, that arrives at t^-.

Let $C(t)$ be a would-be (potential) time-t arrival to the system, $t \geq 0$. Let $R_i(t)$ denote the (remaining) *workload* (in time units) at instant $t \geq 0$, of server i, $i = 1, \ldots, c$ (server numbering is arbitrary). Let $\{W(t)\}_{t \geq 0}$ be the *virtual wait process*. The random variable $W(t)$ is the would-be wait required by $C(t)$ measured from time t until the start of service of $C(t)$. Thus $W(t) = \min_{i=1,\ldots,c} \{R_i(t)\}$, $t \geq 0$. We assume: sample paths of $\{W(t)\}_{t \geq 0}$ and of $\{R_i(t)\}_{t \geq 0}$ are right continuous and have left limits; the model parameters are such that the steady state exists (condition relaxed for the transient analysis in Sects. 4.6.1–4.6.8 below).

Remark 4.2 In M/M/c ($c \geq 2$), virtual wait $\neq$ system workload. The system workload at time t is $\sum_{i=1}^{c} R_i(t)$.

Since $C_{a,t}$ is an *actual* time-t arrival to the system, that arrives at t^-, $C_{a,t}$'s wait is $W(t^-)$ before starting service, from some server i_t^*. If $R_j(t^-) = 0$ for some server j, then $W(t^-) = 0$, and i_t^* is one of the idle servers at t^-. For zero-wait arrivals, server i_t^* is selected from the idle servers according to the model's server-assignment policy (e.g., randomly, or by server number, etc.). If $R_i(t^-) > 0$, $i = 1, \ldots, c$, then $W(t^-) > 0$, and $C_{a,t}$ will start service from server i_t^* at instant $t + W(t^-)$ if i_t^* has the minimum workload among the c occupied servers at t^- (further explained in Sect. 4.4 below).

4.3.2 *Sample Paths of Workload and Virtual Wait*

In some models, sample paths of $R_i(t)$, $i = 1, \ldots, c$, are useful for the overall analysis. We now outline how to construct a sample path of each $R_i(t)$, $t \geq 0$, $i = 1, \ldots, c$. (Refer to Fig. 4.1, which depicts sample paths for a *special case* of generalized M/M/c with $c = 2$.) Without loss of generality, assume the system is empty at $t = 0$. Then $R_i(t) = 0$, $i = 1, \ldots, c$, from $t = 0$ until the first arrival instant (C_1). A new arrival (C_2) starts service from a server i^* and $R_{i^*}(\cdot)$ jumps upward to the ordinate $\underset{dis}{=} \text{Exp}_{\mu_{i^*}}$, where $\mu_{i^*} \in \boldsymbol{\mu}$. $R_{i^*}(\cdot)$, then decreases steadily with slope $= -1$ as service progresses.

Eventually all c servers become occupied simultaneously (just after C_2 arrives). Let $t_1 := \min\{t | \text{all } c \text{ servers are occupied}\}$. If the next customer C_τ arrives at time $\tau > t_1$ before any further service completion, then C_τ is the sole customer waiting to start service at time τ (C_3). C_τ's server will be i_τ^* if $R_{i_\tau^*}(\tau^-) = \min_{i=1,\ldots,c} \left\{R_i(\tau^-)\right\} := W(\tau^-)$ (virtual wait). The workload

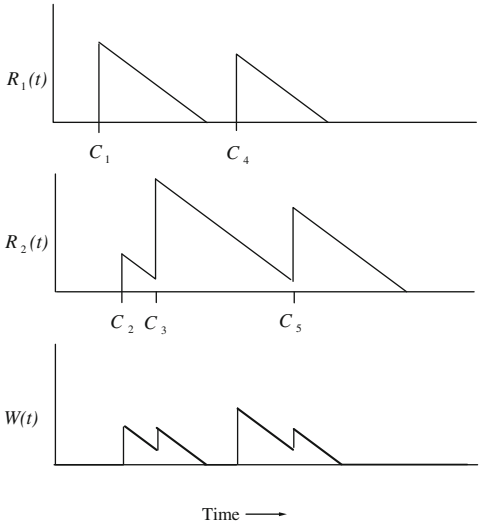

Fig. 4.1 Sample paths of $R_j(t)$, $j = 1, 2$, and $W(t)$, $t \geq 0$ in M/M/2. C_n, $n = 1, 2, \ldots$, denote customers at successive arrival instants

$R_{i_\tau^*}(\tau^-)$ jumps upward by C_τ's service time $s_{i_\tau^*} \underset{dis}{=} \text{Exp}_{\mu_{i_\tau^*}}$, where $\mu_{i_\tau^*} \in \boldsymbol{\mu}$. Thus $R_{i_\tau^*}(\tau) = R_{i_\tau^*}(\tau^-) + s_{i_\tau^*} = W(\tau^-) + s_{i_\tau^*}$. For *all other servers*, $R_i(\tau) = R_i(\tau^-)$, $i \neq i_\tau^*$. Subsequently $W(\tau) = \min_{i=1,\ldots,c}\{R_i(\tau)\}$.

The next arrival that "sees" *at least one idle server* (C_4), will cause the $\{W(t)\}_{t\geq 0}$ to evolve similarly. The next arrival that finds *all servers busy* will be assigned to that server which has minimum workload (C_5) and so forth. If arrivals find several customers waiting in line, the dynamics are similar to the case '*all servers busy*' (described in Sect. 4.4 below).

4.3.3 Distinguishable Servers

When tracking *server workloads*, we regard the servers as distinguishable (Fig. 4.1). However, we are often interested in the statistical properties of the entire system, rather than the processing of each individual customer, or the action of a particular server. Here we analyze the system by constructing a sample path of $\{W(t)\}_{t\geq 0}$ generated according to the model's prescribed probability laws for the service and interarrival times, and operational policies.

Suppose we can keep track of the c server workloads in continuous time. Then we could assign a '*ticket*' to each arrival, which points to its up-coming server, identified because it has the minimum workload at the arrival instant.

This procedure distributes *theoretical* waiting lines to the c servers, although there is only one physical waiting line in the waiting room.

4.3.4 Indistinguishable Servers

In many M/M/c models, it is not necessary to construct sample paths of the server workloads $\{R_i(t)\}_{t\geq 0}, i = 1, \ldots, c$, in order to construct a sample path of $\{W(t)\}_{t\geq 0}$. It suffices to regard the servers as *indistinguishable*. Then it is not necessary to track individual server workloads. To analyze important statistical properties of the model, it is sufficient to track ***directly*** the virtual wait $W(t) := \min_{i=1,\ldots,c}\{R_i(t)\}$. Thus we utilize what we call the ***system configuration*** (Sect. 4.4).

4.4 System Configuration

In generalized M/M/c, assume a 'system manager' knows the up-coming target server i_t^* to be occupied at instant $t + W(t^-)$ by a would-be time-t arrival, denoted by C(t) (i.e., the manager knows the server having minimum workload at time t). Let $\boldsymbol{M}(t) :=$ *system configuration* at time t. The process $\{\boldsymbol{M}(t)\}_{t\geq 0}$ tracks the service rates of the $c-1$ servers *other than* i_τ^*. We assume that the model specifies $J+1$ possible exponential service rates: $\boldsymbol{\mu} = \{\mu_0, \mu_1, \ldots, \mu_J\}$. Each arrival is assigned a service rate selected from the set $\boldsymbol{\mu}$. Recall that if t is not an arrival instant, sample-path right continuity implies $W(t^-) = W(t)$.

Definition 4.1 The system configuration $\boldsymbol{M}(t)$ is a $J+1$ vector of **server occupancy numbers** $m_j \geq 0$, namely $\boldsymbol{M}(t) = (m_0, \ldots, m_J)$, where $m_j :=$ **number of servers having service rate** $\mu_j \in \mu$, *among the* $c-1$ *servers other than* i_t^* *at* $t + W(t^-)$.

Definition 4.2 The set of all possible configurations is denoted by $\boldsymbol{M} = \{\boldsymbol{m} | \boldsymbol{m} = (m_0, \ldots, m_J)\}$.

For each configuration $\boldsymbol{m} \in \boldsymbol{M}$, $0 \leq \sum_{j=0}^{J} m_j \leq c-1$; C($t$) would start service at instant $t + W(t^-)$ and would be assigned a service rate $\mu_t(W(t^-), \boldsymbol{m}) \in \boldsymbol{\mu}$, which nay be a function of three variables: t, $W(t^-)$, and $\boldsymbol{m}$. That is

$$(t, W(t^-), \boldsymbol{m}) \rightarrow \mu_j \in \boldsymbol{\mu}, \text{ for some } j = 0, \ldots, J.$$

Remark 4.3 In various models, the service rate $\mu_t(\cdot,\cdot)$ may also depend on other variables as well. It may be selected randomly from the set $\boldsymbol{\mu}$. Additionally, the number of possible service rates in $\boldsymbol{\mu}$ may be countable.

4.4.1 Inter Start-of-service Depart Time $\mathcal{S}_t$

In generalized M/M/c a *key random variable* is the time-t *'look-ahead'* ***inter start-of-service depart time,*** denoted by $\mathcal{S}_t$. For example, let the state $(W(t^-), M(t^-))$ be $(x, \boldsymbol{m})$ when customer $\mathrm{C}_{a,t}$ arrives. Then $\mathrm{C}_{a,t}$'s required wait before starting service is $x \geq 0$ and the configuration is $\boldsymbol{m}$. If $0 \leq \sum_{i=0}^{J} m_j \leq c-1$ then $x = 0$ and $\mathrm{C}_{a,t}$ starts service immediately by one of the idle servers. If $\sum_{i=0}^{J} m_j = c-1$ there are two possibilities for x. If $x = 0$ then $\mathrm{C}_{a,t}$ starts service immediately by the ***single idle server***. If $x > 0$ then $\mathrm{C}_{a,t}$ waits time x before starting service by the first available server thereafter. Just after $\mathrm{C}_{a,t}$ starts service at time $t+x$ all c servers will be occupied.

Definition 4.3 The ***inter start-of-service depart time*** $\mathcal{S}_t$ is the time measured from $t+x$ (start of $\mathrm{C}_{a,t}$'s service time) until the first departure from the system after $t+x$. In other words $\mathcal{S}_t :=$ ***time from the start of service of*** $\mathrm{C}_{a,t}$ ***until the first departure from the system thereafter***.

Importantly,

$$\mathcal{S}_t \underset{dis}{=} \min\left\{\mathrm{Exp}_{m_0\mu_0}, \ldots, \mathrm{Exp}_{m_J\mu_J}, \mathrm{Exp}_{\mu_t(W(t^-),\boldsymbol{m})}\right\},$$

which is the *minimum* of $\sum_{i=0}^{J} m_j + 1 \,(= c)$ independent exponential r.v.s. Among these, m_j servers have rate μ_j, $j = 0, \ldots, J$, and one server has rate $\mu_t(x, \boldsymbol{m})$ (assigned to $\mathrm{C}_{a,t}$). Thus $\mathcal{S}_t \underset{dis}{=} \mathrm{Exp}_{\nu_t}$ where $\nu_t = \sum_{j=0}^{J} m_j\mu_j + \mu_t(x, \boldsymbol{m})$.

An important aspect of the forgoing definition of system configuration and use of S_t when $W(t^-) > 0$, is that once all c servers are occupied, the probability distribution of S_t is independent of future arrivals to the system. That is, the set of active servers functions like a separate sub-system until the first departure thereafter, mindful of the memoryless property of the exponential service times. Although the concept '*system configuration*' may appear 'different', it is straightforward to apply when developing model equations for the pdf of the waiting time in complex M/M/c queues (see, e.g., pp. 80–97 in [11], and also in [38, 52, 53]).

4.4.2 Number of Configurations

Let $(W(t), M(t)) = (x, \boldsymbol{m})$ (see Definition 4.1 above). Looking ahead to $t + W(t)$, assume $\boldsymbol{m} = (m_0, \ldots, m_J)$ is such that

$$\sum_{j=0}^{J} m_j = k, 0 \le k \le c-1.$$

The servers are considered to be *indistinguishable* (as in subsequent models in this monograph, unless otherwise noted). We track only the *number of servers occupied* with service rate $\mu_j \in \boldsymbol{\mu}$, $j = 0, \ldots, k$.

The number of possible configurations such that exactly k servers are occupied, is the number of non-negative integer solutions of the equation

$$m_0 + \cdots + m_J = k.$$

It is the same as the number of ways of distributing k *indistinguishable* balls in $J+1$ *distinguishable* cells, namely $\binom{J+k}{J} = \binom{J+k}{k}$ (see Lemma, p. 36, Chap. II in [73]). Thus, the total number of possible configurations is

$$\sum_{k=0}^{c-1} \binom{J+k}{J} = \binom{J+c}{J+1} = \binom{J+c}{c-1}. \tag{4.5}$$

The first equality in (4.5) is readily proved by induction.

Example 4.1 Consider an M/M/c queue with $c = 3$ and $J = 2$, so that $\boldsymbol{\mu} = \{\mu_0, \mu_1, \mu_2\}$. If a potential arrival C(t) finds the system **empty**, then $(W(t), M(t)) = (0, (0, 0, 0))$; thus $m_i = 0$, $i = 0, 1, 2$. C(t) would wait zero and "see" zero servers occupied, (0, 0, 0). The number of solutions of $m_0 + m_1 + m_2 = 0$ is $\binom{J+0}{J} = \binom{2}{2} = 1$. C($t$) would wait $W(t^-) = 0$ and start service from one the three unoccupied servers, per the server-assignment policy.

If C(t) would find **one** customer in the system (one occupied server), then $W(t^-) = 0$ and the configuration that C(t) would "see" is one of **three** possible vectors

$$M(t^-) \in \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}.$$

The number of solutions of $m_0 + m_1 + m_2 = 1$ is $\binom{J+1}{J} = \binom{3}{2} = 3$. C($t$) would start service from one of the two unoccupied servers, per the server-assignment policy.

If C(t) would find **two** customers in the system, then $W(t^-) = 0$ and the configuration that C(t) would "see" is one of **six** possible vectors

$$M(t^-) \in \{(2,0,0),(0,2,0),(0,0,2),(1,1,0),(1,0,1),(0,1,1)\}.$$

The number of solutions of $m_0 + m_1 + m_2 = 2$ is $\binom{J+2}{J} = \binom{4}{2} = 6$. C($t$) would start service from the **one unoccupied** server.

If C(t) would find **three or more** customers in the system, then all **three** servers would be occupied at t^-. The look-ahead configuration that C(t) would "see" **just before start of service** at $t + W(t^-)$ is also one of **six** possible vectors

$$M(t) \in \{(2,0,0),(0,2,0),(0,0,2),(1,1,0),(1,0,1),(0,1,1)\}.$$

The six possible configurations are the *same* as when C(t) sees **two occupied** servers. This is because a configuration tracks the service-rate occupancies of those servers **other than** C(t)'s eventual server. Customer C(t) would wait a **positive time** and start service at $t + W(t^-)$ from some server i_t^*. We "look ahead" to the **start of service instant** $t + W(t^-)$ **and assign rate** $\mu_t(W(t^-), M(t^-))$ *to* i_t^*. The random variable $M(t)$ tracks the service-rate occupancies of the two servers other than i_t^* at $t + W(t^-)$. (The look-ahead idea is not new. For example, it is tacitly assumed for the virtual wait in the standard M/G/1 queue, where we increase the virtual wait by a service time at an *arrival* instant, although the service is not started until the end of the waiting time. The generalized M/M/c generalizes the M/G/1 look-ahead idea to the start-of-service time.)

At instant t, the state $(W(t), M(t))$ conveys sufficient information to determine the probabilities of the m_j $(j = 0,\ldots,J)$ occupied servers that will have the minimum service time among all the occupied servers at time $t + W(t^-)$. These probabilities depend on the Markovian property. We shall illustrate this more fully in Example 4.2 below.

The total number of possible configurations is

$$\sum_{k=0}^{c-1} \binom{J+k}{J} = \sum_{k=0}^{2} \binom{J+k}{J} = \binom{J+c}{J+1} = \binom{5}{3} = 10.$$

4.4.3 Border States

In Example 4.1 the zero-wait state $\{(0, \boldsymbol{m})\}$ is a ***border*** state if

$$\boldsymbol{m} \in \{(2,0,0),(0,2,0),(0,0,2),(1,1,0),(1,0,1),(0,1,1)\}.$$

Definition 4.4 We call the state $(W(t), M(t))$ a **border state** if $W(t) = 0$ and $M(t))$ is such that $\sum_{j-0}^{J} m_j = c - 1$. A border state is a discrete zero-wait state in a **boundary** separating other discrete zero-wait states and a set of continuous positive-wait states.

In the above definition, the *other* zero-wait states are *non-border* states such that $0 \leq \sum_{j-0}^{J} m_j < c - 1$. Border states communicate with continuous positive-wait states in one step: at arrival instants (zero-wait $\rightarrow$ positive-wait); or at departure instants (positive-wait $\rightarrow$ zero-wait). When the SP moves on a path *from* a *non-border* zero-wait state *to* a continuous positive-wait state the path must pass through a border state at an arrival instant. In the opposite direction, *from* a positive-wait state to a non-border zero-wait state, the path must pass through a border state at a departure instant.

We denote the set of border states by $\boldsymbol{S}_b$, and the set of border configurations by $\boldsymbol{M}_b$. Thus

$$\begin{aligned} \boldsymbol{S}_b &= \left\{(0, \boldsymbol{m}) \,|\, \textstyle\sum_{j=0}^{J} m_j = c - 1\right\}, \\ \boldsymbol{M}_b &= \{\boldsymbol{m}|\, (0, \boldsymbol{m}) \in \boldsymbol{S}_b\} = \left\{\boldsymbol{m}|\, \textstyle\sum_{j=0}^{J} m_j = c - 1\right\}. \end{aligned} \tag{4.6}$$

4.4.4 The Next Configuration

Consider an actual arrival $\mathrm{C}_{a,t}$ at instant t. $\mathrm{C}_{a,t}$ "sees" configuration $M(t^-)$. Just *after* the arrival the configuration is $M(t)$. Either $M(t) = M(t^-)$ or $M(t) \neq M(t^-)$. We illustrate by example how to compute the probability mass function of $M(t)$.

Example 4.2 Consider Example 4.1 for M/M/c with $c = 3$, $J = 2$. Suppose $\mathrm{C}_{a,t}$ arrives when the wait is $W(t^-)$ and the configuration is $(m_0, m_1, m_2) = (1, 1, 0)$. The state is

$$(W(t^-), M(t^-)) = (0, (1, 1, 0)).$$

Suppose that $\mathrm{C}_{a,t}$ is assigned service rate μ_0, i.e., $\mu_t(W(t^-), (1, 1, 0)) = \mu_0$.

At instant $t + 0$, **just after** $\mathrm{C}_{a,t}$ **starts service**, there will be **two** servers with rate μ_0 since $m_0 = 1$. There will be *one* server with rate μ_1, since $m_1 = 1$. The inter-start-of-service-depart time $\mathcal{S}_t \underset{dis}{=} \mathrm{Exp}_{2\mu_0+\mu_1}$.

We now compute the probability distribution of the **next configuration** at instant $t + W(t)$. Thus,

$$\begin{aligned} P(M(t) &= (2,0,0)) \\ &= P(\text{rate-}\mu_1 \text{ server finishes first}) \\ &= \frac{\mu_1}{2\mu_0 + \mu_1}, \end{aligned}$$

$$\begin{aligned} P(M(t) &= (1,1,0)) \\ &= P(\text{rate-}\mu_0 \text{ server finishes first}) \\ &= \frac{2\mu_0}{2\mu_0 + \mu_1}. \end{aligned}$$

Importantly

$$P(M(t) = (2,0,0)) + P(M(t) = (1,1,0)) = 1.$$

The only two possible configurations for $M(t)$ are $(2,0,0)$ and $(1,1,0)$, independent of whether $W(t^-) = 0$ or $W(t^-) > 0$ (illustrated below in Example). No other configuration is possible for $M(t)$ once the arrival at t^- has been assigned rate μ_0. Knowledge of their probabilities is sufficient to analyze the sample path to write down model equations using LC.

Remark 4.4 The service mechanism can be generalized considerably. We can expand the domain of $\mu_t(w, \boldsymbol{m})$ to include: type or priority class of $C_{a,t}$; type of customer replaced by $C_{a,t}$ in server i_t^*; type of any customer followed by $C_{a,t}$ into service; identity of server i_t^* (e.g., server number or unique property); number of customers in the system or waiting for service at the arrival or start of service instant of $C_{a,t}$; various types of bounds on the virtual wait; reneging indices; blocked and cleared customers, etc. (see, e.g., [38] for an effective definition of $M(t)$ due to L. Green; [42]; also see [39, 44, 53]; and others).

Other generalizations may incorporate: a non-homogeneous Poisson arrival process with intensity λ_t, or a Poisson arrival rate $\lambda(W(t), M(t))$ which is a function of the current state $(W(t), M(t))$; or various Markov arrival processes.

4.5 System Point Process

We now discuss the ***system point process*** and the geometry of its state space (see Fig. 4.2).

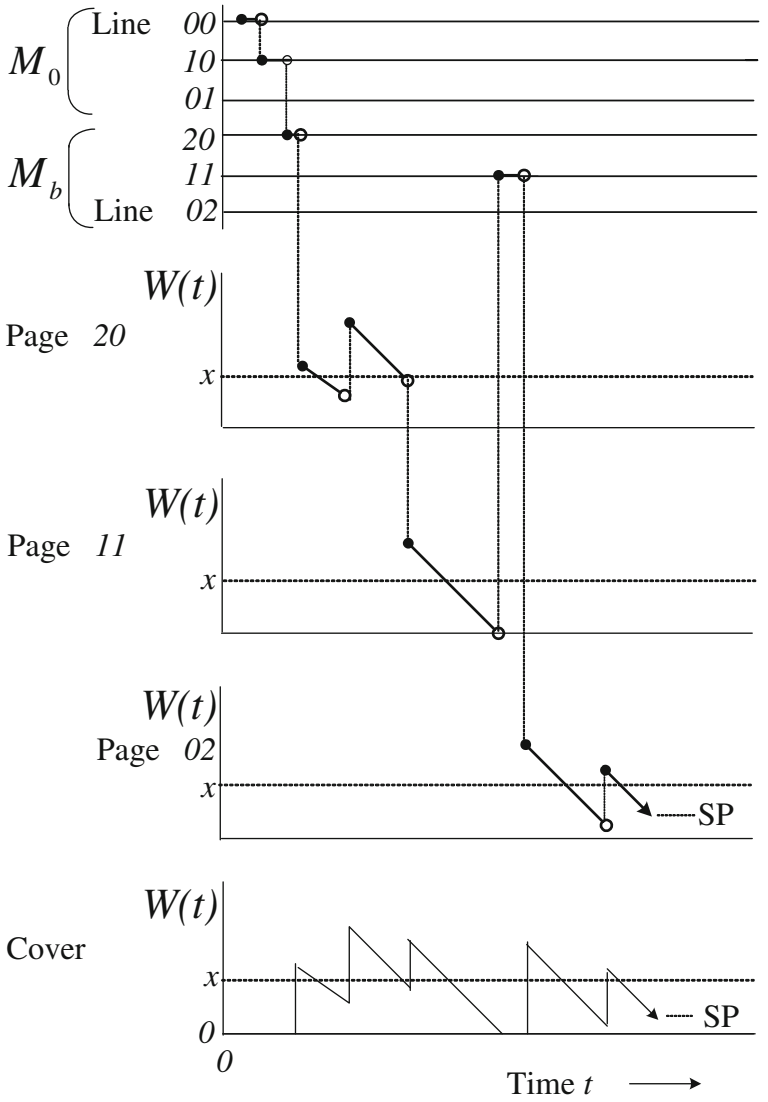

Fig. 4.2 Sample path of SP process $\{W(t), M(t)\}_{t\geq 0}$ for Example 4.4 with random assignment of service rates independent of state at arrival instants ($c = 3$, $J = 1$). The space $\boldsymbol{T} \times \boldsymbol{S}$ has 6 lines for zero-wait states, and 3 pages (sheets) for positive-wait states (pages 20, 11, 02). The cover is the projection of the sample path from all lines and pages onto one non-negative planar quadrant

We call $\{W(t), M(t)\}_{t\geq 0}$ the *system point (SP) process*. Its nomenclature derives from the fact that the SP traces out a sample path as the system evolves over time. The SP process for M/M/c queues is a generalization (with

exponential service times) of the virtual wait process for M/G/1 queues. State variable $W(t) :=$ *virtual wait*; state variable $M(t) :=$ *system configuration at time t*. Random variable $M(t)$ is discrete. The SP process is a Markov process (Sect. 4.5.7 below).

We partition the *state space* $\boldsymbol{S}$ into three disjoint state-space sets $\boldsymbol{S}_0$, $\boldsymbol{S}_b$, $\boldsymbol{S}_1$. $\boldsymbol{S}_0$ contains the zero-wait states that are non-boundary states. $\boldsymbol{S}_b$ contains the zero-wait states that are boundary states. $\boldsymbol{S}_1$ contains the positive-wait states. The states in $\boldsymbol{S}_0 \cup \boldsymbol{S}_b$ are atoms (see Sect. 2.4.9 for 'atom'). The states in $\boldsymbol{S}_1$ are points in a continuum, e.g., $(x, \boldsymbol{m})$, $x > 0$, and $\boldsymbol{m} = (m_{0,\ldots,m_J})$. Specifically,

$$\begin{aligned} \boldsymbol{S}_0 &= \{(0, \boldsymbol{m}) | 0 \leq \textstyle\sum_{j=0}^{J} m_j \leq c-2\}, \\ \boldsymbol{S}_b &= \{(0, \boldsymbol{m}) | \textstyle\sum_{j=0}^{J} m_j = c-1\}, \\ \boldsymbol{S}_1 &= \{(x, \boldsymbol{m}) | x > 0, \textstyle\sum_{j=0}^{J} m_j = c-1\}. \end{aligned} \tag{4.7}$$

Note that $\boldsymbol{S} = \boldsymbol{S}_0 \cup \boldsymbol{S}_b \cup \boldsymbol{S}_1$, and $\boldsymbol{S}_0 \cap \boldsymbol{S}_b = \boldsymbol{S}_0 \cap \boldsymbol{S}_1 = \boldsymbol{S}_b \cap \boldsymbol{S}_1 = \phi$, the empty set. The corresponding sets of system configurations are

$$\begin{aligned} \boldsymbol{M}_0 &= \{\boldsymbol{m} | (0, \boldsymbol{m}) \in \boldsymbol{S}_0) = \{\boldsymbol{m} | 0 \leq \textstyle\sum_{j=0}^{J} m_j \leq c-2\}; \\ \boldsymbol{M}_b &= \{\boldsymbol{m} | (0, \boldsymbol{m}) \in \boldsymbol{S}_b) = \{\boldsymbol{m} | \textstyle\sum_{j=0}^{J} m_j = c-1\}; \\ \boldsymbol{M}_1 &= \{m | (x, \boldsymbol{m}) \in \boldsymbol{S}_1\} = \{\boldsymbol{m} | x > 0, \textstyle\sum_{j=0}^{J} m_j = c-1\}; \end{aligned} \tag{4.8}$$

thus $\boldsymbol{M}_b = \boldsymbol{M}_1$ (see Sect. 4.4 above).

$\boldsymbol{W(t^-) = 0}$ An arrival C(t) would "see" $W(t^-) = 0$ if and only if the state at time t^- is in $\boldsymbol{S}_0 \cup \boldsymbol{S}_b$. C($t$) would then wait zero and start service from some server, say i_t^*, at time t. Geometrically, we associate a distinct horizontal line $\boldsymbol{T} \times (0, \boldsymbol{m})$ with each state $(0, \boldsymbol{m}) \in \boldsymbol{S}_0 \cup \boldsymbol{S}_b$ where $\boldsymbol{T}$ is the time axis $[0, \infty)$. We call the line $\boldsymbol{T} \times (0, \boldsymbol{m})$ "*line* $\boldsymbol{m}$" (e.g., Fig. 4.2).

$\boldsymbol{W(t^-) > 0}$ C(t) would "see" $W(t^-) > 0$ if and only if the state is in $\boldsymbol{S}_1$. C(t) would wait time $W(t^-)$ and start service from some server, say i_t^*, at time $t + W(t^-)$. Geometrically, we associate the quadrant of the plane $\boldsymbol{T} \times (0, \infty)$ with each set of continuous states $(x, \boldsymbol{m}) \in \boldsymbol{S}_1$, $x > 0$. We call the positive quadrant $\boldsymbol{T} \times ((0, \infty), \boldsymbol{m})$ ***sheet*** (or ***page) m*** (e.g., Fig. 4.2).

Sample Path Diagram The LC analyst draws (or visualizes) a plot of $W(t)$ versus t on page $\boldsymbol{m}$ while the system is in the state corresponding to configuration $\boldsymbol{m}$. In a diagram, we may place the zero-wait *border* lines (corresponding to the states in $\boldsymbol{S}_b$) $\boldsymbol{T} \times (0, \boldsymbol{m})$, $(0, \boldsymbol{m}) \in \boldsymbol{S}_b$, alongside the zero-wait *non-border* lines for states $(0, \boldsymbol{m}) \in \boldsymbol{S}_0$; or else at level 0 of the corresponding sheets for the positive-wait states $\boldsymbol{S}_1$ having the same configurations $\boldsymbol{m}$. There is a one-to-one correspondence between sheets and states in $\boldsymbol{S}_b$ (Fig. 4.2).

4.5.1 Sample Path of SP Process

A sample path of $\{W(t), M(t)\}_{t \geq 0}$ is a piecewise right-continuous function of t having left limits. It has a finite number of jumps during finite time intervals (see Sect. 2.2 and Definition 2.1 in Chap. 2). We plot a sample path within a Cartesian product space $\boldsymbol{T} \times \boldsymbol{S} = \boldsymbol{T} \times (\boldsymbol{S_0} \cup \boldsymbol{S_b} \cup \boldsymbol{S_1})$. The direction of time is from left to right. It is useful to envisage each Cartesian product $\boldsymbol{T} \times (0, \boldsymbol{m})$, $(0, \boldsymbol{m}) \in \boldsymbol{S_0} \cup \boldsymbol{S_b}$ as a ***line***; and each quadrant $\boldsymbol{T} \times ((0, \infty), \boldsymbol{m})$, $\boldsymbol{m} \subseteq \boldsymbol{M}_1$, as a sheet (or page in a ***book***).

Description of a Sample Path
Assume that the system starts empty. The SP moves among the zero-wait lines, jumping from line to line at arrival and departure instants, eventually reaching a zero-wait *border* line (often placed at level zero of some sheet $\boldsymbol{m} \in \boldsymbol{M}_b$). Eventually the SP jumps from line $\boldsymbol{m} \in \boldsymbol{M}_b$, to a positive level on some sheet '$\boldsymbol{k}$', at an arrival instant. It then moves steadily with slope -1 on sheet $\boldsymbol{k}$. It is possible that either $\boldsymbol{m} = \boldsymbol{k}$, or $\boldsymbol{m} \neq \boldsymbol{k}$, depending on the probabilities governing the motion of the SP. (See Fig. 4.2 and Example 4.4 in Sect. 4.5.5 for a detailed example.) Other clarifying examples are in the author's Ph.D. thesis (Fig. 4.3, p. 79, Chap. 4 in [11]; and in [52]).

At an arrival instant while the SP is on sheet $\boldsymbol{k}$, the SP may jump to another sheet, say $\boldsymbol{m}'$, and move steadily with slope -1 on sheet $\boldsymbol{m}'$ for a positive time. Otherwise the SP may jump, and stay on the same sheet $\boldsymbol{k}$. On each sheet it moves downward with slope -1. If the SP hits level 0 from above on page $\boldsymbol{k}$ before the next arrival, it starts moving immediately on the border line $\boldsymbol{k}$ (no customers waiting, $c - 1$ servers occupied).

If the SP is in a state in $\boldsymbol{S}_b \cup \boldsymbol{S}_1$ having configuration $\boldsymbol{m}$ at some arrival instant, it makes a jump ending either on page $\boldsymbol{m}$ or on some page $\boldsymbol{k} \neq \boldsymbol{m}$. If $\boldsymbol{k} \neq \boldsymbol{m}$, the SP is said to make an $\boldsymbol{m} \rightarrow \boldsymbol{k}$ *transition*. This may be an upward

jump from a border line $\boldsymbol{m}$, or from sheet $\boldsymbol{m}$ to sheet $\boldsymbol{k}$, at an arrival instant. Generally, $\boldsymbol{m} \to \boldsymbol{k}$ transitions do not give rise to 'typical' level crossings as in M/G/1 models that have exactly one 'page'. However, $\boldsymbol{m} \to \boldsymbol{m}$ transitions from a border line $\boldsymbol{m}$ or from a point on sheet $\boldsymbol{m}$ to a higher point on sheet $\boldsymbol{m}$, are similar to SP jumps as discussed for models with a single sheet (Sect. 2.3 in Chap. 2).

Remark 4.5 In some model variants, an $\boldsymbol{m} \to \boldsymbol{k}$ transition may be a **parallel jump**. That is, the SP makes a jump from a level y on page $\boldsymbol{m}$ to level y on page $\boldsymbol{k} \neq \boldsymbol{m}$, at an arrival instant. For example, in an M/G/1 queue, we may utilize a modified configuration $M(t) = n$, where n is the number of customers waiting for service, and the virtual wait is unchanged at some arrival instants. Such parallel jumps occur in M/G/1 or M/M/c queues with bulk service (see, e.g., [93]).

4.5.2 A Metaphor for Sample Path and SP Motion

The SP motion over the state space is like the motion of the tip of a pen writing out a *single-page* history of the system over time. The writing takes place in a book of transparent pages all the same size. The cover is also transparent. The pen moves from left to right, and never overlays what has been written already. After writing flat or sloped lines on a page for a random amount of time, the pen jumps to a different page, and continues writing. The pen jumps in this manner at random time points from page to page. The next page is selected by a random mechanism depending on where it is presently. The *entire* history up to an instant in time can be seen only by holding all the pages one behind the other, like pages in a book, and viewing the projected history on the ***cover***. The projected history on the cover is invariant with respect to shuffling or mixing the pages, which change their relative positions. An analyst that views an arbitrary page in isolation, sees only local segments of the history specific to that page (see Fig. 3.2, p. 49 and Fig. 4.3, p. 79 in [11]).

The global history is like the *total* sample path of the SP process over the state space $\boldsymbol{S_0} \cup \boldsymbol{S_b} \cup \boldsymbol{S_1}$. The local histories on various pages are like sample-path segments due to sojourns on the 'lines' and 'sheets' of the state space. On the cover, SP motion on all the lines occurs at level 0. That is, when all the lines are projected onto the cover, they are placed at level zero—to form a single "line 0".

We may think of the overall method as having several steps.

1. **Partition** the **Time-State space** into mutually exclusive and exhaustive lines and sheets.
2. **Analyze** the sample-path segments on the lines and sheets using LC methods.
3. **Project** the sample-path segments from the lines and sheets onto the 'cover' of the 'book'. Analyze the projected path on the cover using LC.
4. **Combine** all the LC results with a normalizing condition. Construct the model equations (usually Volterra integral equations of the second kind with parameter for the pdf's of interest) and derive probability distributions of the model.

The LC method utilizes statistical properties of the local path segments on the lines and sheets. It also uses statistical properties of the projected path on the cover. It employs the one-step communication properties among the lines and sheets (at successive arrivals and/or departures) to construct a sample path. Basic LC theorems apply to each page $\boldsymbol{m} \in \boldsymbol{M}$. Jumps out of, and into lines and sheets, follow rate-conservation laws.

Equations for PDF of Wait
We use sample-path structure, and transition rates into and out of state-space sets, to construct (by inspection of the sample path) integro-differential and differential equations in a *transient* analysis. Similarly, we construct integral equations and algebraic equations in a *steady-state* analysis. These are equations for the joint pdf and/or cdf of *wait and system configuration*. We can also derive equations for the marginal (total) pdf and cdf of wait, or for the probabilities of the system configurations.

Remark 4.6 The author originally had the idea for partitioning the state space, visualizing the positive-wait states over time in separate quadrants, and having a 'system point' moving on the quadrants over time, from an analogy with Riemann sheets and diagrams of winding numbers in complex variable theory (see, e.g., Sect. 3.4, p. 137ff in [96]). My Ph.D. thesis used the term 'sheets'. The term 'pages' was introduced soon after. I also thought of using the term 'cards', analogous to boxes of computer cards for data and programs, which were widely in use in 1974. Then, the state space could be pictured like a box or deck of rectangular cards. Such ('IBM') cards had been ubiquitous until personal computers became common in the 1980s.

4.5.3 Notation: Probabilities and Distributions

Transient Probabilities and Distributions
We denote: the zero-wait probabilities by

$$P_t(0, \boldsymbol{m}) = P(W(t) = 0, M(t) = \boldsymbol{m}), (0, \boldsymbol{m}) \in \boldsymbol{S}_0 \cup \boldsymbol{S}_b;$$

the mixed joint cdf of $(W(t), M(t))$ by

$$\begin{aligned} F_t(x, \boldsymbol{m}) &= P(W(t) \le x, M(t) = \boldsymbol{m}) \\ &= P_t(0, \boldsymbol{m}) + P(0 < W(t) \le x, M(t) = \boldsymbol{m}) \\ &= P_t(0, \boldsymbol{m}) + \int_{y=0}^{x} f_t(y, \boldsymbol{m}) dy, \ x \ge 0, t \ge 0, \\ &\qquad (0, \boldsymbol{m}) \in \boldsymbol{S}_0 \cup \boldsymbol{S}_b, (x, \boldsymbol{m}) \in \boldsymbol{S}_1, \end{aligned}$$

where $P(0 < W(t) \le x, \boldsymbol{M}(t) = \boldsymbol{m}) = P(\phi) = 0$ if $x = 0$; and the mixed joint pdf of $(W(t), M(t))$ is

$$f_t(x, m) = \frac{\partial}{\partial x} F_t(x, \boldsymbol{m}), x > 0, t \ge 0, (x, \boldsymbol{m}) \in \boldsymbol{S}_1,$$

wherever $\frac{\partial}{\partial x} F_t(x, \boldsymbol{m})$ exists.

We *assume*:

1. $F_t(x, \boldsymbol{m})$ and $f_t(x, \boldsymbol{m})$ are right continuous in x for every $t \ge 0, \boldsymbol{m} \in \boldsymbol{M}_1$.
2. $\frac{\partial}{\partial t} F_t(x, \boldsymbol{m})$ and $\frac{\partial}{\partial t} f_t(x, \boldsymbol{m})$, $t > 0$, $x \ge 0$, exist and are finite for every $\boldsymbol{m} \in \boldsymbol{M}_1$.

Let $P_0(t) := P(W(t) = 0)$ be the marginal probability of a zero wait at t. Then

$$\begin{aligned} P_0(t) &= \sum_{(0,\boldsymbol{m}) \in \boldsymbol{S}_0 \cup \boldsymbol{S}_b} P_t(0, \boldsymbol{m}) \\ &= \sum_{(0,\boldsymbol{m}) \in \boldsymbol{S}_0} P_t(0, \boldsymbol{m}) + \sum_{(0,\boldsymbol{m}) \in \boldsymbol{S}_b} P_t(0, \boldsymbol{m}), t \ge 0. \end{aligned}$$

The transient marginal cdf of wait $P(W(t) \leq x)$ is

$$\begin{aligned} F_t(x) &= \sum_{(0,\boldsymbol{m})\in \boldsymbol{S}_0} P_t(0,\boldsymbol{m}) + \sum_{(0,\boldsymbol{m})\in \boldsymbol{S}_b} F_t(x,\boldsymbol{m}) \\ &= \sum_{(0,\boldsymbol{m})\in \boldsymbol{S}_0\cup \boldsymbol{S}_b} P_t(0,\boldsymbol{m}) + P(0 < W(t) \leq x) \\ &= P_0(t) + P(0 < W(t) \leq x) \\ &= P_0(t) + \int_{y=0}^{x} f_t(y)dy,\ x \geq 0, t \geq 0. \end{aligned}$$

Note that $P_t(0,\boldsymbol{m}) = F_t(0,\boldsymbol{m})$ for $(0,\boldsymbol{m}) \in \boldsymbol{S}_b$.

(Recall the definitions of $\boldsymbol{M}_b$ and $\boldsymbol{S}_b$ in (4.6), and $\boldsymbol{M}_b = \boldsymbol{M}_1$, which is the set of system configurations for positive-wait states.)

The transient marginal pdf of $W(t)$ is

$$f_t(x) = \frac{\partial}{\partial x} F_t(x) = \sum_{\boldsymbol{m}\in \boldsymbol{M}_1} f_t(x,\boldsymbol{m}), x > 0, t \geq 0.$$

A potential (would-be) arrival C(t) would find the system configuration to be $\boldsymbol{m} \in \boldsymbol{M}_0 \cup \boldsymbol{M}_b$ with probability $P_t(0,\boldsymbol{m})$. C(t) would find the configuration to be $\boldsymbol{m} \in \boldsymbol{M}_1$ with probability $F_t(\infty,\boldsymbol{m})$. The normalizing condition for fixed $t \geq 0$, is

$$\begin{aligned} F_t(\infty) &= \sum_{\boldsymbol{m}\in \boldsymbol{M}_0} P_t(0,\boldsymbol{m}) + \sum_{\boldsymbol{m}\in \boldsymbol{M}_1} F_t(\infty,\boldsymbol{m}) \\ &= \sum_{\boldsymbol{m}\in \boldsymbol{M}_0\cup \boldsymbol{M}_b} P_t(0,\boldsymbol{m}) + \sum_{\boldsymbol{m}\in \boldsymbol{M}_1} \int_{y=0}^{\infty} f_t(y,\boldsymbol{m})dy \\ &= \sum_{(0,\boldsymbol{m})\in \boldsymbol{S}_0\cup \boldsymbol{S}_b} P_t(0,\boldsymbol{m}) + \sum_{\boldsymbol{m}\in \boldsymbol{M}_1} \int_{y=0}^{\infty} f_t(y,\boldsymbol{m})dy = 1. \end{aligned}$$

Steady-State Probabilities and Distributions

We denote the steady-state zero-wait probabilities, pdfs and cdfs of wait by dropping the subscript t in the immediately foregoing notation for the transient quantities.

4.5.4 Configuration Just After an Arrival

Example 4.3 below demonstrates the probability of a system configuration just after an arrival. Assume that an actual customer $C_{a,t}$ arrives and finds the state to be $(W(t^-), \boldsymbol{M}(t^-)) = (x, \boldsymbol{m})$. The service rate assigned to $C_{a,t}$ is $\mu_t(x, \boldsymbol{m}) \in \boldsymbol{\mu}$. Recall that sample paths are right continuous and have left limits.

Example 4.3 Consider Example 4.2 in Sect. 4.4.4, where $c = 3$, $J = 2$. Let each arrival receive a service rate selected **with equal probability** from the set $\boldsymbol{\mu} := \{\mu_0, \mu_1, \mu_2\}$. Then

$$P(C_{a,t} \text{ starts service at } t + W(t^-) \text{ with service rate } \mu_i) = \frac{1}{3}, i = 0, 1, 2,$$

independent of t and $W(t^-)$. Assume $(W(t^-), M(t^-)) = (x, (2, 0, 0))$, $x > 0$, *just before* $C_{a,t}$ arrives. Then $C_{a,t}$ will wait a positive time x. Looking ahead to time $t + W(t^-)$, two other occupied servers will have service rates μ_0 ($m_0 = 2$, $m_1 = m_2 = 0$) when $C_{a,t}$ starts service at $t + W(t^-)$, in the just-vacated server. **Question**: What is the configuration $M(t)$ *just after* $C_{a,t}$ arrives? It can be either $(2, 0, 0)$, $(1, 1, 0)$, or $(1, 0, 1)$. The probabilities for $M(t)$ are:

$$\begin{aligned}
P(M(t) = (2,0,0)) &= P(\mu_t(x,(2,0,0)) = \mu_0) \cdot 1 \\
&\quad + P(\mu_t(x,(2,0,0)) = \mu_1) \cdot \tfrac{\mu_1}{2\mu_0+\mu_1} \\
&\quad + P(\mu_t(x,(2,0,0)) = \mu_2) \tfrac{\mu_2}{2\mu_0+\mu_2} \\
&= \tfrac{1}{3}\left(1 + \tfrac{\mu_1}{2\mu_0+\mu_1} + \tfrac{\mu_2}{2\mu_0+\mu_2}\right). \\
P(M(t) = (1,1,0)) &= P(\mu_t(x,(2,0,0)) = \mu_1) \cdot \tfrac{2\mu_0}{2\mu_0+\mu_1} \\
&= \tfrac{1}{3} \cdot \tfrac{2\mu_0}{2\mu_0+\mu_1}. \\
P(M(t) = (1,0,1)) &= P(\mu_t(x,(2,0,0)) = \mu_2) \cdot \tfrac{2\mu_0}{2\mu_0+\mu_2} \\
&= \tfrac{1}{3} \cdot \tfrac{2\mu_0}{2\mu_0+\mu_2}.
\end{aligned}$$

Thus

$$\begin{aligned}
&P(M(t) = (2,0,0)) + P(M(t) = (1,1,0)) + P(M(t) = (1,0,1)) \\
&= \tfrac{1}{3}\left(1 + \tfrac{\mu_1}{2\mu_0+\mu_1} + \tfrac{\mu_2}{2\mu_0+\mu_2}\right) + \tfrac{1}{3} \cdot \tfrac{2\mu_0}{2\mu_0+\mu_1} + \tfrac{1}{3} \cdot \tfrac{2\mu_0}{2\mu_0+\mu_2} = 1.
\end{aligned}$$

The resulting virtual wait at time t is

$$W(t) = W(t^-) + \mathcal{S}_t = x + \mathcal{S}_t,$$

where $\mathcal{S}_t$ is the inter start-of-service-depart time, distributed as a mixture

$$\mathcal{S}_t \underset{dis}{=} \begin{cases} \text{Exp}_{3\mu_0}, \\ \text{Exp}_{2\mu_0+\mu_1}, \\ \text{Exp}_{2\mu_0+\mu_2}, \end{cases} \text{ with probability } 1/3 \text{ each.}$$

(see Sect. 4.4.1). The sample path will have a jump whose size is distributed as $\mathcal{S}_t$ at instant t (see Fig. 4.2).

4.5.5 Sample Path of SP Process Revisited

We first describe a typical sample path of the virtual wait in Example 4.4 wherein $J = 1$ and $\boldsymbol{\mu} = \{\mu_0, \mu_1\}$, to facilitate exposition. If $J > 1$, sample-path construction would be similar, but with more lines and pages (sheets) in the product space $\boldsymbol{T} \times \boldsymbol{S}$ (see Fig. 4.2). Next, Example 4.5 discusses the general nature of a typical sample path with reference to Example 4.4 and then we outline the mechanics of a *specific* sample path in Example 4.4, based on the M/M/3 queue in Example 4.3 above.

Example 4.4 Consider M/M/c with $c = 3$, $J = 1$. (Here we take $J = 1$ for exposition.) A typical sample path of the virtual wait is given in Fig. 4.2).

Arrivals are assigned an exponential service rate from $\boldsymbol{\mu} = \{\mu_0, \mu_1\}$ with equal probability 1/2. (In general the probabilities can be, e.g., p_0, $p_1 = 1 - p_0$.) The total number of possible configurations is $\binom{J+c}{J+1} = \binom{4}{2} = 6$. The full set of configurations is

$$\boldsymbol{M} = \{00, 10, 01, 11, 20, 02\}.$$

We write (2, 0) as 20 when $m_0 = 2, m_1 = 0$, indicating that 2 servers are occupied with rate μ_0; and similar notation for the other system configurations.

Example 4.5 General Nature of Sample Path with reference to Example 4.4. The state space consists of: (1) six discrete points for the zero-wait states $(0, \boldsymbol{m})$, $\boldsymbol{m} \in \boldsymbol{M}_0 \cup \boldsymbol{M}_b$. Thus $\boldsymbol{M}_0 = \{00, 10, 01\}$ and $\boldsymbol{M}_b = \boldsymbol{M}_1 = \{11, 20, 02\}$; (2) three intervals $((0, \infty), \boldsymbol{m})$, $\boldsymbol{m} \in \boldsymbol{M}_1$. The three border states are $(0, \boldsymbol{m})$, $\boldsymbol{m} \in \boldsymbol{M}_b$.

Arrival Waits Zero. Assume an arrival "sees" state $(0, m)$, $m \in \boldsymbol{M}_0$. The SP moves horizontally at time-rate 1 on a line $\boldsymbol{T} \times (0, \boldsymbol{m})$, $\boldsymbol{m} \in \boldsymbol{M}_0$. If the next arrival occurs before a departure, the SP jumps to a line $\boldsymbol{T} \times (0, \boldsymbol{m}')$, $\boldsymbol{m}' \in \boldsymbol{M}_0 \cup \boldsymbol{M}_b$, where

$$m_0' + m_1' = m_0 + m_1 + 1,$$

because there is one more occupied server. If a departure occurs before an arrival, the SP jumps to a line $\boldsymbol{T} \times (0, \boldsymbol{m}'')$, $\boldsymbol{m}'' \in \boldsymbol{M}_0$, where

$$m_0'' + m_1'' = m_0 + m_1 - 1,$$

because there is one less occupied server. If $\boldsymbol{m} = (0, 0)$, the state can change only due to an arrival.

If an arrival finds the system to be in state $(0, m)$, $m \in \boldsymbol{M}_b$ the SP jumps to a sheet $\boldsymbol{T} \times ((0, \infty), \boldsymbol{k})$, $\boldsymbol{k} \in \boldsymbol{M}_1$. Configuration $\boldsymbol{k}$ is determined by the service rate assigned to the new arrival, and which server finishes first after the new arrival starts service. Denote the service time of an arrival $\mathrm{C}_{a,t}$ by s_t. Then $s_t \underset{dis}{=} \mathrm{Exp}_{\mu_t}$ where

$$P(\mu_t = \mu_i) = \frac{1}{2}, i = 0, 1;$$

(see Fig. 4.2).

To fix ideas, let the SP be on the border line $\boldsymbol{T} \times (0, 20)$ at arrival instant t. Thus $\langle W(t^-), M(t^-) \rangle = \langle 0, 20 \rangle$. $\mathrm{C}_{a,t}$ starts service upon arrival in the one idle server and is assigned either rate μ_0 or μ_1 with probability 1/2 each. Let $\mathcal{S}_t$ denote the time from the start-of-service of $\mathrm{C}_{a,t}$ until the first departure from the system thereafter. (Since all 3 servers are busy $\mathcal{S}_t$ is independent of any future arrivals that join the waiting line.)

Case 1: Let us assume the service time s_t has been assigned rate μ_0. Then $\mathcal{S}_t \underset{dis}{=} \mathrm{Exp}_{3\mu_0}$ because $\mathcal{S}_t = \min\{3 \text{ i.i.d. } \mathrm{Exp}_{\mu_0}\mathrm{s}\}$. The SP jumps upward an amount $\mathcal{S}_t$. The virtual wait at time t is

$$W(t) = W(t^-) + \mathcal{S}_t = 0 + \mathcal{S}_t = \mathcal{S}_t.$$

At instant $t + W(t^-) + \mathcal{S}_t$ $(= t + \mathcal{S}_t)$, one of the three occupied servers completes service. The service rate of each of the resulting two occupied servers at $t + \mathcal{S}_t$ must be μ_0. By the **look-ahead process**, the configuration at t is $M(t) := M(t^-) = 20$. In this scenario the configuration remains the same as when the test customer arrived. Geometrically, at instant t, the look-ahead process impels the SP to jump from *line* 20 to *page* 20, at a height $\underset{dis}{=}$

$\text{Exp}_{3\mu_0}$ (see Fig. 4.2). In Case 1, the SP jumps from line 20 to page 20 at ordinate $\underset{dis}{=} \text{Exp}_{2\mu_0+\mu_1}$, resulting in $W(t) \underset{dis}{=} \text{Exp}_{2\mu_0+\mu_1}$ and $M(t) = 20$.

Case 2. Let us assume that s_t has been assigned rate μ_1. Then $\mathcal{S}_t \underset{dis}{=} \text{Exp}_{2\mu_0+\mu_1}$. At $t + \mathcal{S}_t$ one of the three servers completes service. The service rates of the other two still-occupied servers at $t + \mathcal{S}_t$ are either: (a) both μ_0 with probability $\frac{\mu_1}{2\mu_0+\mu_1}$ (the rate-μ_1 server finishes first), or (b) μ_0 and μ_1 with probability $\frac{2\mu_0}{2\mu_0+\mu_1}$ (a rate-μ_0 server finishes first).

In Case 2(a) at instant t, the SP jumps from line 20 to page 20 at an ordinate $\underset{dis}{=} \text{Exp}_{2\mu_0+\mu_1}$. Thus $W(t) \underset{dis}{=} \text{Exp}_{2\mu_0+\mu_1}$ and $M(t) = 20$. In Case 2(b) at instant t, the SP jumps from line 20 to page 11 at ordinate $\underset{dis}{=} \text{Exp}_{2\mu_0+\mu_1}$, resulting in $W(t) \underset{dis}{=} \text{Exp}_{2\mu_0+\mu_1}$ and $M(t) = 11$.

Arrival Waits a Positive Time. Assume $\text{C}_{a,t}$ arrives when the state is $(x, 20)$, $x > 0$ (SP is at ordinate x on page 20). If the service-rate assignment policy assigns $s_t \underset{dis}{=} \text{Exp}_{\mu_0}$, the SP jumps upward an amount $\text{Exp}_{3\mu_0}$, and moves with slope -1 steadily on page 20. If the service-rate assignment policy assigns $s_t \underset{dis}{=} \text{Exp}_{\mu_1}$, the SP can end up on either page 20 or page 11 at t. The SP jumps upward to $W(t) \underset{dis}{=} W(t^-) + \text{Exp}_{2\mu_0+\mu_1}$ and moves with slope -1 steadily on page 20, with probability $\frac{\mu_1}{2\mu_0+\mu_1}$. The SP jumps upward to ordinate $W(t) \underset{dis}{=} W(t^-) + \text{Exp}_{2\mu_0+\mu_1}$ on page 11, with probability $\frac{2\mu_0}{2\mu_0+\mu_1}$.

If the SP descends to the bottom of page 20 and hits level 0 from above in a continuous manner before a new arrival occurs, it immediately enters border line 20, and continues its motion along line 20.

4.5.6 A Specific Sample Path

We expound further on a possible realization of the SP motion as it traces out the sample path, with reference to Fig. 4.2. Assume that initially the system is empty. The SP moves on line 00. Arrival 1 (C_1) sees an empty system. The server-assignment policy assigns C_1 service rate μ_0. The SP jumps to, and moves on, line 10. C_2 arrives before C_1 completes service and is also assigned rate μ_0. At C_2's arrival the SP jumps to line 20. C_3 arrives while both C_1 and

C_2 are in service. C_3 receives rate μ_1. The SP jumps to an ordinate $\text{Exp}_{2\mu_0+\mu_1}$, and if the rate-$\mu_1$ customer finishes first among the three customers in service, the resulting configuration is again 20. The probability of this event is $\frac{\mu_1}{2\mu_0+\mu_1}$, due to the memoryless property of exponential variates. This explains why at C_3's arrival instant the SP jumps to page 20.

Just before C_4 arrives the SP is descending at slope -1 on page 20. C_4 is assigned service rate μ_0. The SP jumps upward an amount $\text{Exp}_{3\mu_0}$. It remains on page 20. That is, whichever server finishes first, the two remaining active service rates will be μ_0, resulting in configuration 20. C_5 arrives when the SP is on page 20. C_5 is assigned rate μ_1. Suppose a server with rate μ_0 finishes first. The probability of this event is $\frac{2\mu_0}{2\mu_0+\mu_1}$. The SP jumps upward by $\text{Exp}_{2\mu_0+\mu_1}$. It simultaneously makes a $20 \rightarrow 11$ transition from page 20 to page 11, since the two remaining occupied servers have rates μ_0 and μ_1 when the first service ends. The configuration changes immediately from 20 to 11.

No new arrivals occur prior to the completion of the first rate-μ_0 customer. The SP descends on page 11 with slope -1 and hits level 0 from above, exactly when the first rate-μ_0 customer finishes service. The system now presents a zero wait to a potential arrival. When the SP hits level 0, it enters border line 11 (in Fig. 4.2 it jumps to line 11). C_6 arrives, and starts service immediately. C_6 is assigned rate μ_1. The SP jumps to page 02, with probability $\frac{\mu_0}{\mu_0+2\mu_1}$ (μ_0-rate service finishes first), the next configuration is 02.

The system continues to evolve. The SP continues to trace a sample path on the lines and pages according to the probability laws of the model. The sample path gives us a precise picture of the evolving system over time. Construction of the sample path goes hand in hand with understanding the model dynamics, and writing the model equations by inspection.

Remark 4.7 In Sect. 4.8 below we develop the **steady-state** theory. We will then return to Example 4.3, and formulate the balance equations for the zero-wait probabilities $P(0, \boldsymbol{m}), \boldsymbol{m} \in \boldsymbol{M} \equiv \boldsymbol{M}_0 \cup \boldsymbol{M}_b$; integral equations for the 'partial' pdfs of wait $f(x, m), x > 0, \boldsymbol{m} \in \boldsymbol{M}_1$, and for the total pdf $\{P_0, f(x), x > 0\}$.

4.5.7 SP Process Is Markovian

We outline a proof that the SP process is a Markov process. For $t \geq 0$, let $(x, \boldsymbol{m})_t := \textit{event}\ \{(W(t), M(t)) = (x, \boldsymbol{m})\}$. It is required to show that for $x, y \geq 0, \boldsymbol{m}, \boldsymbol{k} \in \boldsymbol{M}$,

$$P((y, \boldsymbol{k})_{t+h}|(x, \boldsymbol{m})_t, (W(u), M(u)_{0\le u<t}) \\ = P((y, \boldsymbol{k})_{t+h}|(x, \boldsymbol{m})_t), t \ge 0, h > 0. \tag{4.9}$$

Formula (4.9) states that the probability of event $(y, \boldsymbol{k})_{t+h}$ given that event $(x, \boldsymbol{m})_t$ occurred, is independent of the history $(W(u), M(u))_{0\le u<t}$. We sketch the proof in two steps: (1) **zero-wait states**; (2) **positive-wait states**.

For a Poisson (or non-homogeneous Poisson) process, the probability of more than one event occurring in $(t, t + h)$ is $o(h)$ (e.g., Definition 5.3, p. 314, and Definition 5.4, p. 339 in [125]).

Zero-Wait Non-border States

Assume state $(0, \boldsymbol{m})_t \in \left\{(0, \boldsymbol{m})|0 \le \sum_{j=0}^{J} m_j \le c - 2\right\}$ ($\boldsymbol{m} \in \boldsymbol{M}_0$, SP $\in \boldsymbol{S}_0$ at time t).

No Departure or Arrival in $(t, t + h)$ The state remains $(0, \boldsymbol{m})$ in $(t, t + h)$ iff no arrival or departure occurs during $(t, t + h)$, or an event with probability $o(h)$ occurs. Thus

$$P((x, \boldsymbol{m})_{t+h}|(x, \boldsymbol{m})_t) = 1 - \left(\lambda + \sum_{j=0}^{J} m_j \mu_j\right) h + o(h),$$

which is independent of $(W(u), M(u))_{0\le u<t}$.

Arrival in $(t, t + h)$ Possibly there is an arrival during $(t, t + h)$. The next configuration will have the form

$$\boldsymbol{m}_{L^+} := (m_0, \ldots, m_L + 1, \ldots, m_J),$$

for some $L \in \{0, \ldots, J\}$. Then

$$P((0, \boldsymbol{m}_{L^+})_{t+h}|(0, \boldsymbol{m})_t) \\ = (\lambda h + o(h)) \cdot P(\mu_t((0, \boldsymbol{m})) = \mu_L) \\ = \lambda h P(\mu_t((0, \boldsymbol{m})) = \mu_L) + o(h), L \in \{0, \ldots, J\} \cdot \tag{4.10}$$

Formula (4.10) is the probability that there is an arrival during $(t, t + h)$ assigned service rate μ_L, which is independent of the history given by $(W(u), M(u))_{0\le u<t}$. Note that

$$\sum_{L=0}^{J} P(\mu_t((0, \boldsymbol{m}) = \mu_L) = 1.$$

Departure in $(t, t + h)$ Possibly there is a departure during $(t, t + h)$. Let configuration

$$\boldsymbol{m}_{L^-} := (m_0, \ldots, \theta_L \cdot (m_L - 1), \ldots, m_J), L \in \{0, \ldots, J\},$$

where

$$\theta_L = \begin{cases} 1 \text{ if } m_L \geq 1, \\ 0 \text{ if } m_L = 0. \end{cases}$$

Assume $\boldsymbol{m} \neq (0, \ldots, 0)$. Then

$$P((0, \boldsymbol{m}_{L-})_{t+h} | (0, \boldsymbol{m})_t) = (m_L \cdot \mu_L)h + o(h), \tag{4.11}$$

which is the probability of a rate-μ_L departure during $(t, t+h)$ (*rate-μ_L service finishes first)*. Expression (4.11) is independent of the history $(W(u), M(u)), 0 \leq u < t$. Note that $\left(\sum_{L=0}^{J} m_L \mu_L\right) h + o(h)$ is the probability of a departure during $(t, t+h)$.

Zero-Wait Border States

Consider zero-wait border states $\left\{(0, \boldsymbol{m})_t | \sum_{j=0}^{J} m_j = c - 1\right\}$ $(\boldsymbol{m} \in \boldsymbol{M}_b, (0, \boldsymbol{m}) \in \boldsymbol{S}_b)$.

No Arrival in $(t, t+h)$ If no arrival or departure occurs, or only a departure occurs, during $(t, t+h)$, the Markov property follows similarly as for the zero-wait non-border states given above.

Arrival in $(t, t+h)$ Possibly there is an arrival during $(t, t+h)$. In *this* case, the SP jumps to a *positive level on a sheet (page)*. Let configuration

$$\begin{aligned} \boldsymbol{k} &:= (m_0, \ldots, m_L + 1, \ldots, m_R - 1, \ldots, m_J) \\ &= (k_0, \ldots, k_J), \end{aligned}$$

for some $L, R \in \{0, \ldots, J\}$. Thus $\sum_{j=0}^{J} k_j = \sum_{j=0}^{J} m_j = c - 1$. Let

$$\nu_L = \sum_{j=0}^{J} m_j \mu_j + \mu_L.$$

The probability that the SP jumps to sheet $\boldsymbol{k}$ during $(t, t+h)$ and is in state-space interval $((y, y+dy), \boldsymbol{k})_{y>0}$ at $t+h$, is

$$\begin{aligned} &P((W(t+h), M(t+h)) \in ((y, y+dy), \boldsymbol{k}) | (0, \boldsymbol{m})_t) \\ &= (\lambda h + o(h)) \cdot P(\mu_t(0, \boldsymbol{m}) = \mu_L) \cdot \frac{m_R \mu_R}{\nu_L} \cdot \nu_L \cdot e^{-\nu_L \cdot y} dy \\ &= \lambda h \cdot P(\mu_t(0, \boldsymbol{m}) = \mu_L) \cdot m_R \mu_R \cdot e^{-\nu_L y} dy + o(h), L \in \{0, \ldots, J\}, \end{aligned}$$

which is independent of the history $(W(u), M(u))_{0\le u<t}$. The right side is the probability that there is an arrival in $(t, t+h)$, which is assigned service rate μ_L, and a rate-μ_R *service* finishes first among the occupied servers, at a time in the state space interval $(y, y+dy)$.

Positive-Wait States
Arrival In $(t, t+h)$ Given $(x, \boldsymbol{m})_t$, $x > 0$, where $\sum_{j=0}^{J} m_j = c-1$, there may be an arrival during $(t, t+h)$. Let

$$\boldsymbol{k} = (m_0, \ldots, m_L + 1, \ldots, m_R - 1, \ldots, m_J).$$

Reasoning as for zero-wait border states, we obtain

$$\begin{aligned} &P((W(t+h), M(t+h) \in ((x+y, x+y+dy), \boldsymbol{k})|(x, \boldsymbol{m})_t)) \\ &= \lambda h \cdot P(\mu_t(0, \boldsymbol{m}) = \mu_L) \cdot m_R \mu_R \cdot e^{-\nu_L(y-x)} dy + o(h), \end{aligned}$$

which is independent of $(W(u), M(u))_{0\le u<t}$.

Virtual Wait in $(0, h)$
Consider the case where all servers are occupied, no customers are waiting and $W(t) \in (0, h)$, where h is "small". Assume a server completes service before a new arrival occurs. Given $(x, \boldsymbol{m})_t$, $0 < x < h$, $\sum_{j=0}^{J} m_j = c-1$, we obtain

$$P((0, \boldsymbol{m})_{t+h}|(x, \boldsymbol{m})_t) = 1 - \lambda x + o(x).$$

The SP hits level 0 from above in a continuous manner at $t+x$. It immediately enters border line $\boldsymbol{m}$ corresponding to the border state $(0, \boldsymbol{m})$, and continues its motion in the direction of Time. This is independent of the past history prior to t.

The above cases cover all possible situations. Formula (4.9) follows in each case, implying that the SP process has the Markov property.

4.5.8 Departures from Positive-Wait States

We examine the departure rates during a sojourn on a sheet (page). (Table 4.1 describes the symbols in Fig. 4.3.)

Suppose the SP is at a positive level on page $\boldsymbol{m} \in \boldsymbol{M}_1$ ($\sum_{j=0}^{J} m_j = c-1$ and all c servers are occupied, including the last arrival). The occupancy number of service rate μ_j among the $c-1$ servers, not occupied by the last arrival, is m_j, $j \in \{0, \ldots, J\}$.

The single remaining server, which *is occupied by the last arrival*, may have an arbitrary service rate $\mu^* \in \boldsymbol{\mu}$. Assume μ^* does not match a positive

Table 4.1 Description of Symbols in Fig. 4.3

Symbol	Description
τ_n	Arrival instant
C_{τ_n}	Customer that arrives at τ_n
σ_n	Start of service instant of C_{τ_n}
$\mathcal{S}_{\tau_n}$	$\sigma_{n+1} - \sigma_n =$ inter start-of-service depart time

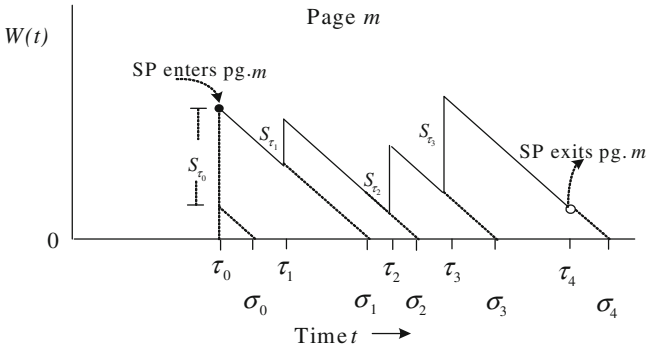

Fig. 4.3 SP sojourn on page m. Departure rate may differ on intervals (τ_0, σ_0), (σ_0, σ_1), (σ_1, σ_2), (σ_2, σ_3), (σ_3, τ_4). At instants σ_0, σ_1, σ_2, arrivals C_{τ_0}, C_{τ_1}, C_{τ_2} start service. Just after departure instants $\sigma_0 + \mathcal{S}_{\tau_0}$, $\sigma_1 + \mathcal{S}_{\tau_1}$, $\sigma_2 + \mathcal{S}_{\tau_2}$, the remaining $c - 1$ servers will have server occupancies $\boldsymbol{m} = (m_0, \ldots, m_J)$

component in configuration $\boldsymbol{m}$. In order for the SP to *remain* on page $\boldsymbol{m}$ just after that arrival, the rate-μ^* server must complete service first among the c occupied servers (see Fig. 4.3).

While the SP is on page $\boldsymbol{m}$, the system exponential departure rate will, in general, differ during inter-departure intervals. These possibly different exponential departure rates have no effect on the *Markov property* of the SP process. The configurations are determined *at arrival instants* (i.e., earlier when service rates are assigned) (Fig. 4.3).

4.6 Transient Analysis of Generalized M/M/c

Sections 4.6.1–4.6.6 develop LC relations and definitions leading to the formulation of integro-differential equations for the transient time-t pdf of the virtual wait, in Sect. 4.6.8 below. This development is based on the author's

working papers [21, 23]. The transient analysis complements the results in [52], which focuses on the generalized M/M/c queue in steady-state. Section 4.6.7 derives the steady-state integral equations by letting $t \to \infty$ in the transient equations. Section 4.7 serves as a brief tutorial on writing steady-state model equations using LC and sample paths.

4.6.1 Transient PDF of Wait and Downcrossings

We next determine relationships between the transient pdf of wait and sample-path transitions. Let $\mathcal{D}_t(x, \boldsymbol{m}) :=$ number of sample-path downcrossings of level x on page $\boldsymbol{m} \in \boldsymbol{M}_1$ during $[0, t]$. Let

$$\mathcal{D}_t(x) = \sum_{\boldsymbol{m} \in \boldsymbol{M}_1} \mathcal{D}_t(x, \boldsymbol{m})$$

denote the total number of downcrossings of level x on all pages during $[0, t]$. Theorem 4.3 connects the instantaneous rate of change of the expected number of downcrossings of level x in $[0, t]$, to the time-t transient pdf of wait at level x.

Theorem 4.3 For each configuration $\boldsymbol{m} \in \boldsymbol{M}_1$,

$$\frac{\partial}{\partial t} E(\mathcal{D}_t(x, \boldsymbol{m})) = f_t(x, \boldsymbol{m}), x > 0, t > 0, \tag{4.12}$$

$$\frac{\partial}{\partial t} E(\mathcal{D}_t(0, \boldsymbol{m})) = f_t(0^+, \boldsymbol{m}) \ (= f_t(0, \boldsymbol{m})), t > 0, \tag{4.13}$$

$$\frac{\partial}{\partial t} E(\mathcal{D}_t(x)) = f_t(x), x > 0, t > 0, \tag{4.14}$$

$$\frac{\partial}{\partial t} E(\mathcal{D}_t(0)) = f_t(0^+) \ (= f_t(0)), t > 0. \tag{4.15}$$

Proof Fix state-space level $x > 0$. Consider instants t and $t + h$, where $t > 0$, and $h > 0$ is small. To prove (4.12) and (4.13) for page $\boldsymbol{m}$, we develop a table similar to (3.10) in Chap. 3 for the M/G/1 queue, and proceed as in the proof of (3.8) and (3.9). Formulas (4.14) and (4.15) follow from the definitions of $\mathcal{D}_t(x)$ and the total pdf $f_t(x), x > 0$. ■

Corollary 4.1

$$E(\mathcal{D}_t(x, \boldsymbol{m})) = \int_{s=0}^{t} f_s(x, \boldsymbol{m})ds, \tag{4.16}$$

$$E(\mathcal{D}_t(0, \boldsymbol{m})) = \int_{s=0}^{t} f_s(0^+, \boldsymbol{m})ds, \tag{4.17}$$

$$E(\mathcal{D}_t(x)) = \int_{s=0}^{t} f_s(x)ds, \tag{4.18}$$

$$E(\mathcal{D}_t(0)) = \int_{s=0}^{t} f_s(0^+)ds. \tag{4.19}$$

Proof Integrating both sides of (4.12), (4.13), (4.14) and (4.15) with respect to s over the interval $[0, t]$ and applying the initial conditions

$$E(\mathcal{D}_0(x, \boldsymbol{m})) = E(\mathcal{D}_0(x)) = 0, x \geq 0,$$

yield (4.16), (4.17), (4.18) and (4.19), respectively. ■

4.6.2 Steady-State PDF of Wait and Downcrossings

Corollary 4.2 below connects the SP limiting downcrossing rate as $t \to \infty$ and the steady-state pdf of wait, at a state-space level. It is analogous to Corollary 3.2 for M/G/1. It also demonstrates the equality of the limit of the instantaneous rate of change of the expected number of downcrossings in $[0, t]$, and the limit of the average downcrossing rate over $[0, t]$.

Let $\boldsymbol{S}_{\boldsymbol{m}} = ([0, \infty), \boldsymbol{m})$, $\boldsymbol{m} \in \boldsymbol{M}_1$. The results below apply to each page $\boldsymbol{T} \times \boldsymbol{S}_{\boldsymbol{m}}, \boldsymbol{m} \in \boldsymbol{M}_1$ as well as to the "book" $\boldsymbol{T} \times \left(\cup_{\boldsymbol{m} \in \boldsymbol{M}_1} \boldsymbol{S}_{\boldsymbol{m}}\right)$.

Corollary 4.2 *Assume the following limits exist*

$$\lim_{t \to \infty} f_t(x, \boldsymbol{m}) \equiv f(x, \boldsymbol{m}), x \in \boldsymbol{S}_{\boldsymbol{m}}, \boldsymbol{m} \in \boldsymbol{M}_1.$$

Then

$$\lim_{t \to \infty} \frac{\partial}{\partial t} E(\mathcal{D}_t(x, \boldsymbol{m})) = \lim_{t \to \infty} \frac{E(\mathcal{D}_t(x, \boldsymbol{m}))}{t} = f(x, \boldsymbol{m}), x > 0, \tag{4.20}$$

$$\lim_{t \to \infty} \frac{\partial}{\partial t} E(\mathcal{D}_t(0, \boldsymbol{m})) = \lim_{t \to \infty} \frac{E(\mathcal{D}_t(0, \boldsymbol{m}))}{t} = f(0^+, \boldsymbol{m}) \equiv f(0, \boldsymbol{m}), \tag{4.21}$$

$$\lim_{t\to\infty}\frac{\partial}{\partial t}E(\mathcal{D}_t(x)) = \lim_{t\to\infty}\frac{E(\mathcal{D}_t(x))}{t} = f(x), x > 0, \tag{4.22}$$

$$\lim_{t\to\infty}\frac{\partial}{\partial t}E(\mathcal{D}_t(0)) = \lim_{t\to\infty}\frac{E(\mathcal{D}_t(0))}{t} = f(0^+) \equiv f(0). \tag{4.23}$$

Proof In (4.20), (4.21), (4.22) and (4.23), the equalities of the left-most terms to the pdfs on the right, follow by letting $t \to \infty$ in (4.12), (4.13), (4.14) and (4.15), respectively. The equalities of the middle terms to the pdfs on the right, follow by dividing both sides of (4.16), (4.17), (4.18) and (4.19) by $t > 0$ and letting $t \to \infty$. ■

4.6.3 SP $\boldsymbol{m} \to \boldsymbol{k}$ Transitions

Before discussing the relationship between the transient pdf of wait and SP upcrossings, we define SP $\boldsymbol{m} \to \boldsymbol{k}$ transitions. We say that the SP makes an $\boldsymbol{m} \to \boldsymbol{k}$ transition at instant t_0 if it *exits* state-space set $\boldsymbol{S_m}$ and *enters* state-space set $\boldsymbol{S_k}$ at t_0. That is, the SP exits $([0, \infty), \boldsymbol{m})$ and enters $([0, \infty), \boldsymbol{k})$ at t_0. If $\boldsymbol{m} = \boldsymbol{k}$, then an $\boldsymbol{m} \to \boldsymbol{k}$ transition maintains the SP on page $\boldsymbol{m}$ at t_0. Similar remarks apply to zero-wait lines $\boldsymbol{m}, \boldsymbol{k} \in \boldsymbol{M}_0$, or line $\boldsymbol{m} \in \boldsymbol{M}_b$ and $\boldsymbol{S_k}$ (see Sects. 2.4.3, 2.4.4 for definitions of entrance and exit).

$\boldsymbol{m} \to \boldsymbol{k}$ Upcrossing of a Level Consider $\boldsymbol{S_m}, \boldsymbol{S_k}$. Fix level $x > 0$. An $\boldsymbol{m} \to \boldsymbol{k}$ upcrossing of level x occurs at instant t_0 if the SP exits set $([0, x), \boldsymbol{m})$ and enters set $((x, \infty), \boldsymbol{k})$ at t_0. That is, the SP makes both an $\boldsymbol{m} \to \boldsymbol{k}$ transition and an upcrossing of level x at t_0. Thus the SP moves instantaneously (not in Time) from page $\boldsymbol{m}$ to page $\boldsymbol{k}$ and from a level below x to a level above x. Viewed from the "cover" of the "book", the upcrossing of level x resembles an "ordinary" upcrossing of x by a sample path of the virtual wait in the M/G/1 queue (see Fig. 4.2). Similar definitions apply to line $\boldsymbol{m}$ and $\boldsymbol{S_k}$ (page $\boldsymbol{k}$).

$\boldsymbol{m} \to \boldsymbol{k}$ Parallel Transition In some variants of the M/M/c queue, the SP may make "parallel" transitions. The SP makes an $\boldsymbol{m} \to \boldsymbol{k}$ *parallel* transition at t_0 if it exits $\boldsymbol{S_m}$ from a level y and enters $\boldsymbol{S_k}$ at the *same level* y, at t_0. SP parallel transitions can also occur in variants of *single-server* queues (e.g., queues with bulk service [20, 93]) and in other stochastic models. The concepts of system configuration, pages (sheets), cover, $\boldsymbol{m} \to \boldsymbol{k}$ transitions, etc., are useful in analyzing many other stochastic models.

4.6.4 SP $\boldsymbol{m} \to \boldsymbol{K}$ Upcrossings Viewed from "Cover"

Let

$$\mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k}), \boldsymbol{m}, \boldsymbol{k} \in \boldsymbol{M}_1$$

denote the number of SP $\boldsymbol{m} \to \boldsymbol{k}$ upcrossings of level x during $[0, t]$. Denote the *total* number of upcrossings of level x during $[0, t]$ (as viewed from the "cover" of the "book") by

$$\mathcal{U}_t(x) = \sum_{\boldsymbol{m}, \boldsymbol{k} \in \boldsymbol{M}_1} \mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k}). \tag{4.24}$$

In (4.24) $\mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k})$ will be positive only if $\boldsymbol{m}$, $\boldsymbol{k}$ are such that page $\boldsymbol{k}$ is accessible from page $\boldsymbol{m}$ in one step at an arrival instant (considering lines $\boldsymbol{m}$ and $\boldsymbol{k}$ as zero-levels of pages $\boldsymbol{m}$, $\boldsymbol{k}$ respectively). For an $\boldsymbol{m} \to \boldsymbol{k}$ upcrossing of level x to occur, the "target" page $\boldsymbol{k}$ can be either page $\boldsymbol{m}$ itself ($\boldsymbol{k} = \boldsymbol{m}$) or a different page ($\boldsymbol{k} \neq \boldsymbol{m}$).

4.6.5 Number of Types of $\boldsymbol{m} \to \boldsymbol{K}$ Upcrossings

A *type* of $\boldsymbol{m} \to \boldsymbol{k}$ upcrossing is an ordered pair $(\boldsymbol{m}, \boldsymbol{k})$. The *total* number of possible types of $\boldsymbol{m} \to \boldsymbol{k}$ upcrossings depends on how many pages communicate in one step at arrival instants. An upper bound on the total number of possible $\boldsymbol{m} \to \boldsymbol{k}$ upcrossings is

$$\begin{aligned} &\textit{number of ordered pairs } (\boldsymbol{m}, \boldsymbol{k}) \ = (\textit{number of configurations in } \boldsymbol{M}_1)^2 \\ &= \binom{J+c-1}{c-1}^2 = \binom{J+c-1}{J}^2. \end{aligned}$$

This maximum number $\binom{J+c-1}{c-1}^2$ is realized only if *all* $\binom{J+c-1}{c-1}$ pages communicate in one step. In that case, there are $\binom{J+c-1}{c-1}$ ways to select the "source" page $\boldsymbol{m}$ and $\binom{J+c-1}{c-1}$ ways to select the "target" page $\boldsymbol{k}$ (with replacement).

Example 4.6 Consider an M/M/c queue with $c = 3$ and $J = 1$, as in Example 4.4 (see Fig. 4.2). The set of configurations corresponding to pages is $\boldsymbol{M}_1 = \{20, 11, 02\}$. Here $\binom{J+c-1}{c-1} = \binom{3}{2} = 3$. An upper bound on the number of types of $\boldsymbol{m} \to \boldsymbol{k}$ transitions (ordered pairs $(\boldsymbol{m}, \boldsymbol{k})$) is $3^2 = 9$. This maximum can be realized only if all configurations in $\boldsymbol{M}_1$ communicate with each other in one step. This will depend on the probabilities governing the

evolution of the states over time. In the present example, configurations 20 and 02 do not communicate in one step (at an arrival instant). There are *seven* possible **types** of one-step transitions, namely,

$$\{20 \to 20, 20 \to 11, 11 \to 20, 11 \to 11, 11 \to 02, 02 \to 11, 02 \to 02\}.$$

Transition types $20 \to 02$ and $02 \to 20$ are not possible.

The Probability $p_t(z, \boldsymbol{m} \to \boldsymbol{k})$ We denote the probability that page $\boldsymbol{k}$ is accessible in one step from level z on page $\boldsymbol{m}$ at an arrival instant t, by $p_t(z, \boldsymbol{m} \to \boldsymbol{k})$. Thus for each $\boldsymbol{m} \in \boldsymbol{M}_1$

$$\sum_{\boldsymbol{k} \in \boldsymbol{M}_1} p_t(z, \boldsymbol{m} \to \boldsymbol{k}) = 1.$$

Usually, for fixed z, there is some $\boldsymbol{k}$ for which $p_t(z, \boldsymbol{m} \to \boldsymbol{k}) = 0$. Then page $\boldsymbol{k}$ is not accessible from level z on page $\boldsymbol{m}$ in one step. If such inaccessibility applies for all $(z, \boldsymbol{m})$, $z \geq 0$, then page $\boldsymbol{k}$ is not accessible from page $\boldsymbol{m}$ in one step. This is the case in Example 4.6: for $\boldsymbol{m} = 20$ and $\boldsymbol{k} = 02$,

$$p_t(z, 20 \to 02) = p_t(z, 02 \to 20) = 0, z \geq 0;$$

so, pages $\boldsymbol{m}$ and $\boldsymbol{k}$ do not communicate in one step.

4.6.6 Transient PDF of Wait and Upcrossings

If a time-t arrival $\mathrm{C}_{a,t}$ finds the state to be $(z, \boldsymbol{m})$, then $\mathrm{C}_{a,t}$ is assigned a service rate $\mu_t(z, \boldsymbol{m}) \in \boldsymbol{\mu}$. We assume that $\mu_t(z, \boldsymbol{m})$ is a right continuous with respect to both z and t. In Theorem 4.4 we use the fact that $\boldsymbol{M}_1 = \boldsymbol{M}_b = \left\{\boldsymbol{m} | \sum_{j=0}^{J} m_j = c - 1\right\}$ (defined in Sect. 4.4.3).

Theorem 4.4 For $\boldsymbol{m}, \boldsymbol{k} \in \boldsymbol{M}_1$, the instantaneous rate of change of the expected number of $\boldsymbol{m} \to \boldsymbol{k}$ upcrossings in $[0, t]$ is given by

$$\frac{\partial}{\partial t} E(\mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k})) = \lambda \int_{z=0}^{x} p_t(z, \boldsymbol{m} \to \boldsymbol{k}) e^{-\nu_t(z, \boldsymbol{m})(x-z)} dF_t(z, \boldsymbol{m}), x \geq 0, t \geq 0, \quad (4.25)$$

where

$$\nu_t(z, \boldsymbol{m}) = \sum_{j=0}^{J} m_j \mu_j + \mu_t(z, \boldsymbol{m}).$$

Proof Fix level $x > 0$ on page $\boldsymbol{m}$, and time $t > 0$. Examination of a sample path on page $\boldsymbol{m}$ over the time interval $(t, t+h)$, $h > 0$, leads to the non-zero values of $\mathcal{U}_{t+h}(x, \boldsymbol{m}, \boldsymbol{k}) - \mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k})$, and corresponding probabilities in (4.26) below. We omit $\mathcal{U}_{t+h}(x, \boldsymbol{m}, \boldsymbol{k}) - \mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k}) = 0$, which contributes 0 to $E(\mathcal{U}_{t+h}(x, \boldsymbol{m}, \boldsymbol{k}) - \mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k}))$. We omit negative values, because $\{\mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k})\}_{t \geq 0}$ is a counting process implying $\mathcal{U}_{t+h}(x, \boldsymbol{m}, \boldsymbol{k}) - \mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k}) \geq 0$.

$\mathcal{U}_{t+h}(x, \boldsymbol{m}, \boldsymbol{k}) - \mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k})$	Probability
$+1$	$\lambda h P_0(t) p_t(0, \boldsymbol{m} \to \boldsymbol{k}) e^{-\nu_t(0,\boldsymbol{m})x}$ $+\lambda h \int_h^x p_t(z, \boldsymbol{m} \to \boldsymbol{k}) e^{-\nu_t(z,\boldsymbol{m})(x-z)} f_t(z) dz + o(h)$
≥ 2	$o(h)$.

(4.26)

In (4.26), taking the expected value of $\mathcal{U}_{t+h}(x, \boldsymbol{m}, \boldsymbol{k}) - \mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k})$, dividing by $h > 0$ and letting $h \downarrow 0$, yields

$$\begin{aligned} \frac{\partial}{\partial t} E(\mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k})) &= \lambda \cdot P_0(t) \cdot p_t(0, \boldsymbol{m} \to \boldsymbol{k}) \cdot e^{-\nu_t(0,\boldsymbol{m})x} \\ &\quad + \lambda \int_{z=0}^{x} p_t(z, \boldsymbol{m} \to \boldsymbol{k}) \cdot e^{-\nu_t(z,\boldsymbol{m})(x-z)\cdot} f_t(z) dz \\ &= \lambda \int_{z=0}^{x} p_t(z, \boldsymbol{m} \to \boldsymbol{k}) \cdot e^{-\nu_t(z,\boldsymbol{m})(x-z)} \cdot dF_t(z, \boldsymbol{m}), \end{aligned} \tag{4.27}$$

which is the same as (4.25). ■

Corollary 4.3 *For* $\boldsymbol{m}, \boldsymbol{k} \in \boldsymbol{M}_1$,

$$\begin{aligned} &E(\mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k})) \\ &= \lambda \int_{s=0}^{t} \int_{z=0}^{x} p_s(z, \boldsymbol{m} \to \boldsymbol{k}) \cdot e^{-\nu_s(z,\boldsymbol{m})(x-z)} \cdot dF_s(z, \boldsymbol{m}) ds, \, x \geq 0, t \geq 0. \end{aligned} \tag{4.28}$$

Proof In (4.25) change the variable from t to s on both sides, integrate with respect to s over the interval $[0, t]$, and apply the initial condition $E(\mathcal{U}_0(x, \boldsymbol{m}, \boldsymbol{k})) = 0$. This yields (4.28). ■

Corollary 4.4 *Consider the "cover". For $x \geq 0, t \geq 0$,*

$$\frac{\partial}{\partial t} E(\mathcal{U}_t(x)) = \lambda \sum_{\boldsymbol{m},\boldsymbol{k} \in \boldsymbol{M}_1} \int_{z=0}^{x} p_t(z, \boldsymbol{m} \to \boldsymbol{k}) \cdot e^{-\nu_t(z,\boldsymbol{m})(x-z)} \cdot dF_t(z, \boldsymbol{m}) \tag{4.29}$$

$$\frac{\partial}{\partial t} E(\mathcal{U}_t(0)) = \lambda \sum_{\boldsymbol{m},\boldsymbol{k} \in \boldsymbol{M}_1} p_t(0, \boldsymbol{m} \to \boldsymbol{k}) \cdot F_t(0, \boldsymbol{m}). \tag{4.30}$$

Proof We define $\mathcal{U}_t(x), x \geq 0$ in (4.24). Equations (4.29) and (4.30) follow by setting $x > 0$, and $x = 0$, respectively, in (4.27), and applying (4.24). (The sample path viewed from the cover is the projection of the sample-path segments from all pages onto a single sheet.) ■

Corollary 4.5 *For $\boldsymbol{m}, \boldsymbol{k} \in \boldsymbol{M}_1$ and $x \geq 0, t \geq 0$,*

$$E(\mathcal{U}_t(x)) = \lambda \sum_{\boldsymbol{m},\boldsymbol{k}} \int_{s=0}^{t} \int_{z=0}^{x} p_s(z, \boldsymbol{m} \to \boldsymbol{k}) \cdot e^{-\nu_s(z,\boldsymbol{m})(x-z)} \cdot dF_s(z, \boldsymbol{m}) ds,$$

$$E(\mathcal{U}_t(0)) = \lambda \sum_{\boldsymbol{m},\boldsymbol{k}} \int_{s=0}^{t} p_s(0, \boldsymbol{m} \to \boldsymbol{k}) \cdot F_s(0, \boldsymbol{m}) ds.$$

Proof In (4.29) and (4.30) change t to s and integrate with respect to s on $[0, t]$. Then apply the initial condition $\mathcal{U}_0(x) = 0, x \geq 0$. ■

4.6.7 Steady-State PDF of Wait and Upcrossings

Corollary 4.6 below proves

$$\lim_{t \to \infty} \frac{\partial}{\partial t} E(\mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k})) = \lim_{t \to \infty} \frac{E(\mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k}))}{t},$$

by relating both limits to the steady-state pdf of wait. Let

$$p(z, \boldsymbol{m} \to \boldsymbol{k}), \ \nu(z, \boldsymbol{m}), \ F(z, \boldsymbol{m}), \text{ and } f(z, \boldsymbol{m})$$

be the limiting values of

$$p_t(z, \boldsymbol{m} \to \boldsymbol{k}), \ \nu_t(z, \boldsymbol{m}), \ F_t(z, \boldsymbol{m}), \ f_t(z, \boldsymbol{m}),$$

respectively, as $t \to \infty$ (for definition of: $p_t(z, \boldsymbol{m} \to \boldsymbol{k})$ see Sect. 4.6.5; $\nu_t(z, \boldsymbol{m})$ see formula 4.25).

Corollary 4.6 *For $\boldsymbol{m}, \boldsymbol{k} \in \boldsymbol{M}_1$ and $x \geq 0$,*

$$\begin{aligned}\lim_{t\to\infty} \frac{\partial}{\partial t} E(\mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k})) &= \lim_{t\to\infty} \frac{E(\mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k}))}{t} \\ &= \lambda \int_{z=0}^{x} p(z, \boldsymbol{m} \to \boldsymbol{k}) \cdot e^{-\nu(z,\boldsymbol{m})(x-z)} \cdot dF(z, \boldsymbol{m}) \\ &= \lambda p(0, \boldsymbol{m} \to \boldsymbol{k}) \cdot e^{-\nu(0,\boldsymbol{m})x} P_0 \\ &\quad + \lambda \int_{z=0}^{x} p(z, \boldsymbol{m} \to \boldsymbol{k}) \cdot e^{-\nu(z,\boldsymbol{m})(x-z)} \cdot f(z, \boldsymbol{m}) dz. \end{aligned} \tag{4.31}$$

Proof The equality

$$\lim_{t\to\infty} \frac{\partial}{\partial t} E(\mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k})) = \lambda \int_{z=0}^{x} p(z, \boldsymbol{m} \to \boldsymbol{k}) \cdot e^{-\nu(z,\boldsymbol{m})(x-z)} \cdot dF(z, \boldsymbol{m}),$$

follows by letting $t \to \infty$ on both sides of (4.25). The equality

$$\lim_{t\to\infty} \frac{E(\mathcal{U}_t(x, \boldsymbol{m}, \boldsymbol{k}))}{t} = \lambda \int_{z=0}^{x} p(z, \boldsymbol{m} \to \boldsymbol{k}) \cdot e^{-\nu(z,\boldsymbol{m})(x-z)} \cdot dF(z, \boldsymbol{m})$$

is obtained upon dividing both sides of (4.28) by $t > 0$, letting $t \to \infty$, and using L'Hôpital's rule (e.g., Theorem 9, p. 179 in [137]; and many Calculus texts). Equation (4.31) then follows. ■

The next corollary relates the limits

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

for the expected *total number* of upcrossings in $[0, t]$, to the steady-state *total* probability distribution of wait.

Corollary 4.7 *For $x \geq 0$,*

$$\begin{aligned}&\lim_{t\to\infty} \frac{\partial}{\partial t} E(\mathcal{U}_t(x)) = \lim_{t\to\infty} \frac{E(\mathcal{U}_t(x))}{t} \\ &= \lambda \sum_{\boldsymbol{m},\boldsymbol{k}\in\boldsymbol{M}_1} \int_{z=0}^{x} p(z, \boldsymbol{m} \to \boldsymbol{k}) \cdot e^{-\nu(z,\boldsymbol{m})(x-z)} \cdot dF(z, \boldsymbol{m}) \\ &= \lambda \sum_{\boldsymbol{m},\boldsymbol{k}\in\boldsymbol{M}_1} p(0, \boldsymbol{m} \to \boldsymbol{k}) \cdot e^{-\nu(0,\boldsymbol{m})x} P_{0,\boldsymbol{m}} \\ &\quad + \lambda \sum_{\boldsymbol{m},\boldsymbol{k}\in\boldsymbol{M}_1} \int_{z=0}^{x} p(z, \boldsymbol{m} \to \boldsymbol{k}) \cdot e^{-\nu(z,\boldsymbol{m})(x-z)} \cdot f(z, \boldsymbol{m}) dz. \end{aligned} \tag{4.32}$$

Proof The result (4.32) follows from (4.31) and the definition of $\mathcal{U}_t(x)$ in (4.24). ■

4.6.8 Equations for Transient PDF of Wait

We derive the transient model equations for the generalized M/M/c model. These equations comprise: (1) $\binom{J+c-1}{c-1}$ integro-differential equations for the partial pdfs $f_t(x, \boldsymbol{m}), x > 0, \boldsymbol{m} \in \boldsymbol{M}_1$; (2) $\binom{J+c-1}{c-1}$ differential equations for the zero-wait probabilities $P_t(0, \boldsymbol{m}), \boldsymbol{m} \in \boldsymbol{M}_1 (= \boldsymbol{M}_b)$; (3) $\binom{J+c-1}{c-2}$ differential equations for the zero-wait probabilities $P_t(0, \boldsymbol{m}), \boldsymbol{m} \in \boldsymbol{M}_0$; (4) one equation for the normalizing condition. Also $\boldsymbol{M}_0 = \left\{\boldsymbol{m} | 0 \le \sum_{i=0}^{J} m_j \le c-2\right\}$ (see definition in formula 4.8).

We also derive the model equations for the *total* transient mixed pdf of wait $\{P_0(t), f_t(x)\}_{x>0, t\ge 0}$ (cover of book—see definition in formula 4.5.3).

Formula (4.1) and especially (4.2) of Theorem B (Sect. 4.2.1) play important roles in these derivations. In Theorem B we take the set $\boldsymbol{A}$ to be an interval in the state space having one of its boundaries equal to x.

Equations for Partial Transient PDFs of Wait

Before stating Theorem 4.5, we introduce/review some definitions.

Definition 4.5 page $\boldsymbol{i} := T \times ((0, \infty), \boldsymbol{i})$ where system configuration $\boldsymbol{i} \in \boldsymbol{M}_1 (= \boldsymbol{M}_b)$—technically page $\boldsymbol{i}$ excludes line $\boldsymbol{i} := T \times (0, (0, \boldsymbol{i}))$, which may be separately depicted, or appended to the bottom of page $\boldsymbol{i}$, in geometric figures (see, e.g., Fig. 4.2); $J_t^{(0,x)}(\boldsymbol{k}, \boldsymbol{m}) :=$ *number of* $k \to m$ *jumps that* ***start*** *in state set* $((0, x), \boldsymbol{k})$ *during* $[0, t]$ and end in $((0, \infty), \boldsymbol{m})$; $J_t^0(\boldsymbol{k}, \boldsymbol{m}) :=$ *number of* $k \to m$ *jumps that* ***start*** *in state set* $(0, \boldsymbol{k})$ *during* $[0, t]$ and end in $((0, \infty), \boldsymbol{m})$; $\mathcal{U}_t(x, \boldsymbol{k}, \boldsymbol{m}) :=$ *number of SP* $\boldsymbol{k} \to \boldsymbol{m}$ *transitions that start in* $([0,x)\ , k)$ and ***jump-upcross*** *level* x *during* $[0, t]$ (start in $((0, x), \boldsymbol{k})$ or in $(0, \boldsymbol{k})$ during $[0, t]$ and end in $((x, \infty)\ , \boldsymbol{m})$).

Theorem 4.5 (1) The integro-differential equations for $f_t(x, \boldsymbol{m}), \boldsymbol{m} \in \boldsymbol{M}_1$, are

$$f_t(x, \boldsymbol{m}) + \lambda \sum_{\boldsymbol{k} \ne \boldsymbol{m}} \int_{z=0}^{x} p_t(z, \boldsymbol{k}, \boldsymbol{m})(1 - e^{-\nu_t(z, \boldsymbol{m})(x-z)}) f_t(z, \boldsymbol{k}) dz \quad (4.33)$$

$$+ \lambda \sum_{\boldsymbol{k} \ne \boldsymbol{m}} p_t(0, \boldsymbol{k}, \boldsymbol{m})(1 - e^{-\nu_t(0, \boldsymbol{k})(x-z)}) P_t(0, \boldsymbol{k})$$

$$= \frac{\partial}{\partial t} F_t(x, \boldsymbol{m}) - \frac{\partial}{\partial t} P_t(0, \boldsymbol{m}) + f_t(0, \boldsymbol{m})$$

$$+\lambda \int_{z=0}^{x} p_t(z, \boldsymbol{m}, \boldsymbol{m}) e^{-\nu_t(z,\boldsymbol{m})(x-z)}) f_t(z, \boldsymbol{m}) dz$$
$$+\lambda \sum_{\boldsymbol{k} \neq \boldsymbol{m}} \int_{z=0}^{x} p_t(z, \boldsymbol{m}, \boldsymbol{k}) f_t(z, \boldsymbol{m}) dz, x \geq 0, t \geq 0,$$

where configuration $\boldsymbol{k} \in \boldsymbol{M}_1$.
(2) The differential equation for $P_t(0, \boldsymbol{m})$, $\boldsymbol{m} \in \boldsymbol{M}_1$, is

$$f_t(0, \boldsymbol{m}) + \lambda \sum_{\boldsymbol{k}} p_t(0, \boldsymbol{k}, \boldsymbol{m}) P_t(0, \boldsymbol{k})$$
$$= \frac{\partial}{\partial t} P_t(0, \boldsymbol{m}) + \left(\lambda + \sum_{j=0}^{J} m_j \mu_j \right) P_t(0, \boldsymbol{m}) \tag{4.34}$$

where $\boldsymbol{k}$ is such that $\sum_{j=0}^{J} k_j = c - 2$.
(3) The differential equations for $P_t(0, \boldsymbol{m})$, $\boldsymbol{m} \in \boldsymbol{M}_0$, are

$$\lambda \sum_{\boldsymbol{r} \neq \boldsymbol{m}} p_t(0, \boldsymbol{r}, \boldsymbol{m}) P_t(0, \boldsymbol{r}) + \sum_{\boldsymbol{s} \neq \boldsymbol{m}} s_j \mu_j p_t(0, \boldsymbol{s}, \boldsymbol{m}) P_t(0, \boldsymbol{s})$$
$$= \frac{\partial}{\partial t} P_t(0, \boldsymbol{m}) + \left(\lambda + \sum_{j=0}^{J} m_j \mu_j \right) P_t(0, \boldsymbol{m}), \tag{4.35}$$

where state $(0, \boldsymbol{m})$ is accessible in one step from state $(0, \boldsymbol{r})$ at an arrival instant, and in one step from $(0, \boldsymbol{s})$ at a departure instant. That is,

$$\sum_{j=0}^{J} m_j = \sum_{j=0}^{J} r_j + 1 = \sum_{j=0}^{J} s_j - 1.$$

(4) The normalizing condition is

$$\sum_{\boldsymbol{m} \in \boldsymbol{M}_0 \cup \boldsymbol{M}_1} P_t(0, \boldsymbol{m}) + \sum_{\boldsymbol{m} \in \boldsymbol{M}_1} \int_{x=0}^{\infty} f_t(x, \boldsymbol{m}) dx = 1. \tag{4.36}$$

Proof **(1)** We derive (4.33) by applying Theorem B (Sect. 4.2.1).
Choose $\boldsymbol{A}$. In (4.1) and (4.2), choose $\boldsymbol{A} := ((0, x), \boldsymbol{m})$ (i.e., $\boldsymbol{A}$ is open interval $(0, x)$ on page $\boldsymbol{m}$). The measure of set $\boldsymbol{A}$ at time t is

$$P_t(\boldsymbol{A}) = F_t(x, \boldsymbol{m}) - F_t(0, \boldsymbol{m}) = F_t(x, \boldsymbol{m}) - P_t(0, \boldsymbol{m}).$$

Entrance rate into $\boldsymbol{A}$. The SP can *enter* $\boldsymbol{A}$ by: (i) downcrossing level x on page $\boldsymbol{m}$; (ii) making a $\boldsymbol{k} \to \boldsymbol{m}$ ($\boldsymbol{k} \neq \boldsymbol{m}$) upward jump starting in $((0, x), \boldsymbol{k})$, *that* ***ends in*** $((0, x), \boldsymbol{m})$; (iii) making a jump that *starts from state* (*0,k*) ($\boldsymbol{k} \in \boldsymbol{M}_1$) (sometimes located *at level* 0 on page $\boldsymbol{k}$ in figures), and *ends* ***in*** $((0, x), \boldsymbol{m})$.

The number of SP entrances into set $\boldsymbol{A}$ during $[0, t]$ is

$$\mathcal{I}_t(\boldsymbol{A}) = \mathcal{D}_t(x, \boldsymbol{m}) + \sum_{\boldsymbol{k} \neq \boldsymbol{m} \in \boldsymbol{M}_1} J_t^{(0,x)}(\boldsymbol{k}, \boldsymbol{m}) + \sum_{\boldsymbol{k} \in \boldsymbol{M}_1} J_t^0(\boldsymbol{k}, \boldsymbol{m}) - \sum_{\boldsymbol{k} \in \boldsymbol{M}_1} \mathcal{U}_t(x, \boldsymbol{k}, \boldsymbol{m}). \tag{4.37}$$

In (4.37) the algebraic sum

$$\sum_{\boldsymbol{k} \neq \boldsymbol{m} \in \boldsymbol{M}_1} J^{(0,t)}(\boldsymbol{k}, \boldsymbol{m}) + \sum_{\boldsymbol{k} \in \boldsymbol{M}_1} J_t^0(\boldsymbol{k}, \boldsymbol{m}) - \sum_{\boldsymbol{k} \in \boldsymbol{M}_1} \mathcal{U}_t(x, \boldsymbol{k}, \boldsymbol{m}) \tag{4.38}$$

= (number of SP jumps that start in $([0, x), \boldsymbol{k})$ on any pages or zero-wait lines $k \in \boldsymbol{M}_1$, and end in $((0, \infty), \boldsymbol{m})$ – (number of such jumps that end in $((x, \infty), \boldsymbol{m})$ on page m during $[0, t]$)). Thus, (4.38) is the number of SP entrances into $((0, x), \boldsymbol{m})$ during $[0, t]$, due to jumps that start below x on pages or lines outside of $\boldsymbol{T} \times ((0, x), \boldsymbol{m})$ and end in $((0, x), \boldsymbol{m})$. Therefore $\mathcal{I}_t(\boldsymbol{A})$ is the *total* number of SP entrances into $((0, x), \boldsymbol{m})$ from all sources in one step during $[0, t]$.

Taking expected values and then $\frac{\partial}{\partial t}$ in (4.37) yields

$$\frac{\partial}{\partial t} E\left(\mathcal{I}_t(A)\right) = \frac{\partial}{\partial t} E\left(\mathcal{D}_t(x, \boldsymbol{m})\right) + \sum_{\boldsymbol{k} \neq \boldsymbol{m}} \frac{\partial}{\partial t} E\left(J_t^{(0,x)}(\boldsymbol{k}, \boldsymbol{m})\right) + \sum_{\boldsymbol{k} \in \boldsymbol{M}_1} \frac{\partial}{\partial t} E\left(J_t^0(\boldsymbol{k}, \boldsymbol{m})\right) - \sum_{\boldsymbol{k} \in \boldsymbol{M}_1} \frac{\partial}{\partial t} E\left(\mathcal{U}_t(x, \boldsymbol{k}, \boldsymbol{m})\right). \tag{4.39}$$

Exit Rate of set $\boldsymbol{A}$. The SP can *exit* set $\boldsymbol{A}$ by: (i) hitting level 0 on page $\boldsymbol{m}$ from above in a continuous fashion, (i.e., exiting $((0, x), \boldsymbol{m})$ and simultaneously entering state $(0, \boldsymbol{m})$); (ii) starting in $((0, x), \boldsymbol{m})$ at an arrival instant and making an $\boldsymbol{m} \to \boldsymbol{k}$ (including $\boldsymbol{m} \to \boldsymbol{m}$) *upcrossing* of level x, ending in $((x, \infty), \boldsymbol{k})$, $\boldsymbol{k} \in \boldsymbol{M}_1$; (iii) starting in $((0, x), \boldsymbol{m})$ at an arrival instant instant and making an $\boldsymbol{m} \to \boldsymbol{k}$ ($\boldsymbol{k} \neq \boldsymbol{m}$) jump-transition that ends *below* x on any page $\boldsymbol{k} \neq \boldsymbol{m}$, i.e., in $((0, x), \boldsymbol{k})$, $\boldsymbol{k} \in \boldsymbol{M}_1$, $\boldsymbol{k} \neq \boldsymbol{m}$

The total number of exits from set $\boldsymbol{A}$ during $[0, t]$ is

$$\mathcal{O}_t(\boldsymbol{A}) = \mathcal{D}_t(0,\boldsymbol{m}) + \sum_{\boldsymbol{k}\in\boldsymbol{M}_1} \mathcal{U}_t(x,\boldsymbol{m},\boldsymbol{k}) + \sum_{\boldsymbol{k}\neq\boldsymbol{m}\in\boldsymbol{M}_1} J_t^{(0,x)}(\boldsymbol{m},\boldsymbol{k}) - \sum_{\boldsymbol{k}\neq\boldsymbol{m}\in\boldsymbol{M}_1} \mathcal{U}_t(x,\boldsymbol{m},\boldsymbol{k}). \tag{4.40}$$

Explanation of (4.40). On the right side, $\mathcal{D}_t(0,m)$ is the number of exits from $\boldsymbol{A}$ during $[0,t]$ by downcrossing level 0 on page $\boldsymbol{m}$ (entering $(0,\boldsymbol{m})$). The term $\sum_{\boldsymbol{k}\in\boldsymbol{M}_1}\mathcal{U}_t(x,\boldsymbol{m},\boldsymbol{k})$ is the number of SP jump exits from $([0,x),\boldsymbol{m})$ during $[0,t]$ that *upcross* level x on any page $\boldsymbol{k}$ (including $\boldsymbol{k}=\boldsymbol{m}$). $\sum_{\boldsymbol{k}\neq m\in\boldsymbol{M}_1} J_t^{(0,x)}(\boldsymbol{m},\boldsymbol{k})$ is the number jump-exits that start in $((0,x),\boldsymbol{m})$ and end in $((0,\infty),\boldsymbol{k})$ for any $\boldsymbol{k}\neq\boldsymbol{m}$. Term $-\sum_{\boldsymbol{k}\neq m\in\boldsymbol{M}_1}\mathcal{U}_t(x,\boldsymbol{m},\boldsymbol{k})$ cancels the extra number of jump-exits from $([0,x),\boldsymbol{m})$ during $[0,t]$ that *upcross* level x on any page $\boldsymbol{k}\neq\boldsymbol{m}$, i.e., ending in $((x,\infty),\boldsymbol{k})$.

Taking expected values and then $\frac{\partial}{\partial t}$ in (4.40) results in

$$\begin{aligned}&\frac{\partial}{\partial t}E\left(\mathcal{O}_t(\boldsymbol{A})\right)\\ &= \frac{\partial}{\partial t}E\left(\mathcal{D}_t(0,m)\right) + \sum_{\boldsymbol{k}\in\boldsymbol{M}_1}\frac{\partial}{\partial t}E\left(\mathcal{U}_t(x,\boldsymbol{m},\boldsymbol{k})\right)\\ &\quad + \sum_{\boldsymbol{k}\neq m}\frac{\partial}{\partial t}E\left(\mathcal{U}_t^{(0,x)}(x,\boldsymbol{m},\boldsymbol{k})\right) - \sum_{\boldsymbol{k}\neq m}\frac{\partial}{\partial t}E\left(\mathcal{U}_t(x,\boldsymbol{m},\boldsymbol{k})\right).\end{aligned} \tag{4.41}$$

Integro-differential Equation: We substitute in (4.41) from (4.12), (4.13), (4.25). This yields the integro-differential equation (4.33).

(2) We derive (4.34) by letting set $\boldsymbol{A}=(0,\boldsymbol{m})$ in Theorem B, and substituting formulas from Section 4.6.1 relating downcrossings and the transient distribution of wait, as in the proof of (1).

(3) We derive (4.35) in a similar manner as in (2).

(4) The final equation is the normalizing condition

$$\sum_{\boldsymbol{m}\in\boldsymbol{M}_0\cup\boldsymbol{M}_1} P_t(0,\boldsymbol{m}) + \sum_{\boldsymbol{m}\in\boldsymbol{M}_1}\int_{x=0}^{\infty} f_t(x,\boldsymbol{m})dx = 1.$$

■

Remark 4.8 In practice we can derive an equivalent set of model equations by letting set $\boldsymbol{A}=((x,\infty),\boldsymbol{m})$, $x>0$, in Theorem B (instead of substituting $((0,x],\boldsymbol{m})$). This choice of $\boldsymbol{A}$ may simplify the derivation of the model equations for $f_t(x,\boldsymbol{m})$. We would then consider SP jumps that start below and end above level x. This would yield terms of the form $e^{-\nu_t(z,\boldsymbol{m})(x-z)}$

rather than $(1 - e^{-\nu_t(z,\boldsymbol{m})(x-z)})$ in the integrands. In real-world applications, writing the integro-differential equations is much simpler than it may seem at this point. Some practice on a few simple models will quickly establish the method. It is very intuitive.

Remark 4.9 We can generalize the model upon replacing λ by λ_t, depending on t. The arrival stream would then be a non-homogeneous Poisson process. This generalization holds because the developments in the foregoing sections involving λ are essentially the same if λ_t is substituted for λ.

Model Equations for Total Transient PDF

In the following theorem, we utilize the previously defined equivalent notation $F_t(0, \boldsymbol{m}) \equiv P_t(0, \boldsymbol{m}), \boldsymbol{m} \in \boldsymbol{M}_1, F_t(0) \equiv P_0(t), f_t(0) \equiv f_t(0^+)$.

Theorem 4.6 For the total pdf of wait $\{P_0(t),\ f_t(x)\}_{x>0}$, as viewed from the 'cover', the following integro-differential and differential equations hold :

$$\begin{aligned} f_t(x) &= \frac{\partial}{\partial t} F_t(x) + \lambda \sum_{\boldsymbol{m} \in \boldsymbol{M}_1} \int_{z=0}^{x} e^{-\nu_t(z,\boldsymbol{m})(x-z)} dF_t(z, \boldsymbol{m}) \\ &= \frac{\partial}{\partial t} F_t(x) + \lambda \sum_{\boldsymbol{m} \in \boldsymbol{M}_1} P_t(0, \boldsymbol{m}) e^{-\nu_t(z,\boldsymbol{m})x} \\ &\quad + \lambda \sum_{\boldsymbol{m} \in \boldsymbol{M}_1} \int_{z=0}^{x} e^{-\nu_t(z,\boldsymbol{m})(x-z)} f_t(z, \boldsymbol{m}) dz, x > 0, t \geq 0, \quad (4.42) \\ f_t(0) &= \frac{\partial}{\partial t} P_0(t) + \lambda \sum_{\boldsymbol{m} \in \boldsymbol{M}_1} P_t(0, \boldsymbol{m}), t \geq 0. \quad (4.43) \end{aligned}$$

Proof In Theorem B (Sect. 4.2.1), consider the set

$$\boldsymbol{A} = \left(\cup_{\boldsymbol{m} \in \boldsymbol{M}_0 \cup \boldsymbol{M}_1} (0, \boldsymbol{m})\right) \cup \left(\cup_{\boldsymbol{m} \in \boldsymbol{M}_1} ((0, x], \boldsymbol{m}), x > 0\right).$$

Set $\boldsymbol{A}$ includes all $\binom{J+c}{c-1}$ zero-wait states $\left\{(0, \boldsymbol{m}) | 0 \leq \sum_{j=0}^{J} m_j \leq c - 1\right\}$, as well as all positive-wait states $\left\{(y, \boldsymbol{m}) | \sum_{j=0}^{J} m_j = c - 1, y \in (0, x]\right\}$.

Every SP *entrance* into $\boldsymbol{A}$ must occur from above at level x. Therefore all entrances are due to (continuous) SP downcrossings of level x. Every *exit* out of $\boldsymbol{A}$ must be due to a jump starting below level x on a page and ending at a level above level x on some page. Therefore all SP exits from set $\boldsymbol{A}$ are due to upcrossings of level x.

Thus

$$\mathcal{I}_t(A) = \mathcal{D}_t(x),\ \mathcal{O}_t(A) = \mathcal{U}_t(x),$$
$$E\left(\mathcal{I}_t(A)\right) = E\left(\mathcal{D}_t(x)\right),\ E\left(\mathcal{O}_t(A)\right) = E\left(\mathcal{U}_t(x)\right),$$
$$\frac{\partial}{\partial t}E\left(\mathcal{I}_t(A)\right) = \frac{\partial}{\partial t}E\left(\mathcal{D}_t(x)\right),\ \frac{\partial}{\partial t}E\left(\mathcal{O}_t(A)\right) = \frac{\partial}{\partial t}E\left(\mathcal{U}_t(x)\right).$$

We then substitute these expressions into formulas (4.14), (4.15), (4.29) and (4.30). This substitution yields the integro-differential equation (4.42) and the differential equation (4.43). ■

The normalizing condition

$$P_0(t) + \int_{x=0}^{\infty} f_t(x)dx = 1,$$

is used along with (4.42), (4.43) to solve for the unknown time-t zero-wait probabilities and positive-wait pdfs.

When it is not feasible to obtain an analytical solution, we can use numerical, simulation or approximation techniques to solve for the transient zero-wait probabilities and positive-wait pdfs.

4.6.9 Equations for Steady-State PDF of Wait

We obtain the model equations for the steady-state pdf of wait by letting $t \to \infty$ in (4.34)–(4.36). All quantities subscripted by t have limits as $t \to \infty$. We denote the limits utilizing the same notation, omitting subscript t. If stability holds, then

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

This corresponds to the cdf $F(x, \boldsymbol{m})$ being independent of t.

Theorem 4.7 The integral equation for the steady-state pdf $f(x, m)$, $\boldsymbol{m} \in \boldsymbol{M}_1$, is

$$\begin{aligned}
&f(x, \boldsymbol{m}) + \lambda \textstyle\sum_{\boldsymbol{k}\neq\boldsymbol{m}\in\boldsymbol{M}_1} \int_{z=0}^{x} p(z, \boldsymbol{k}, \boldsymbol{m})(1 - e^{-\nu(z,\boldsymbol{m})(x-z)}) f(z, \boldsymbol{k})dz \\
&\quad + \lambda \textstyle\sum_{\boldsymbol{k}\in\boldsymbol{M}_1} p(0, \boldsymbol{k}, \boldsymbol{m})(1 - e^{-\nu(0,\boldsymbol{k})(x-z)}) P(0, \boldsymbol{k}) \\
&= f(0, \boldsymbol{m}) \\
&\quad + \lambda \textstyle\int_{z=0}^{x} p(z, \boldsymbol{m}, \boldsymbol{m}) e^{-\nu(z,\boldsymbol{m})(x-z)}) f(z, \boldsymbol{m})dz \\
&\quad + \lambda \textstyle\sum_{\boldsymbol{k}\neq\boldsymbol{m}\in\boldsymbol{M}_1} \int_{z=0}^{x} p(z, \boldsymbol{m}, \boldsymbol{k}) f(z, \boldsymbol{m})dz, x \geq 0.
\end{aligned} \tag{4.44}$$

Proof We obtain (4.44) by letting $t \to \infty$ in (4.33). ■

Theorem 4.8 The model equation for the total steady-state pdf is

$$\begin{aligned} f(x) = {} & \lambda \textstyle\sum_{\boldsymbol{m} \in \boldsymbol{M}_1} P(0, \boldsymbol{m}) e^{-\nu(z,\boldsymbol{m})x} \\ & + \lambda \textstyle\sum_{\boldsymbol{m} \in \boldsymbol{M}_1} \int_{z=0}^{x} e^{-\nu(z,\boldsymbol{m})(x-z)} f(z, \boldsymbol{m}) dz, \; x > 0. \end{aligned} \tag{4.45}$$

Proof Let $t \to \infty$ in (4.42). ■

Remark 4.10 In practice, it is often more efficient to derive balance equations for SP exit/entrance rates with respect to the state-space sets $((x, \infty), \boldsymbol{m})_{\boldsymbol{m} \in \boldsymbol{M}_1}$, $x > 0$, rather than with respect to the state-space sets $((0, x), \boldsymbol{m})_{\boldsymbol{m} \in \boldsymbol{M}_1}, x > 0$. The derived equations will be equivalent, no matter which state-space sets are employed for rate balance.

Interpretation of Equations in Theorem 4.7 for Sheets
We now interpret (4.44) in terms of rate balance across levels and between pages. This interpretation gives LC power for deriving steady-state model equations by inspecting a typical sample path, in a vast array of complex stochastic models.

In (4.44) the *left* side is the SP *entrance rate* into $((0, x), \boldsymbol{m})$. The term $f(x, \boldsymbol{m})$ is the SP downcrossing rate of level x on page $\boldsymbol{m}$. The term

$$\lambda \sum_{\boldsymbol{k} \neq \boldsymbol{m} \in \boldsymbol{M}_1} \int_{z=0}^{x} p(z, \boldsymbol{k}, \boldsymbol{m})(1 - e^{-\nu(z,\boldsymbol{m})(x-z)}) f(z, \boldsymbol{k}) dz$$

is the rate at which the SP enters composite state $((0, x), \boldsymbol{m})$ due to jumps at arrival instants that originate in $((0, x), \boldsymbol{k})$ on pages $\boldsymbol{k} \neq \boldsymbol{m}$. The term

$$\lambda \sum_{\boldsymbol{k} \in \boldsymbol{M}_1} p(0, \boldsymbol{k}, \boldsymbol{m})(1 - e^{-\nu(0,\boldsymbol{k})(x-z)}) P(0, \boldsymbol{k})$$

is the rate at which the SP enters composite state $((0, x), \boldsymbol{m})$ due to jumps that originate at level $(0, \boldsymbol{k})$ on any zero-wait line $\boldsymbol{k} \in \boldsymbol{M}_1$). These three terms exhaust the possible paths by which the SP can enter $((0, x), \boldsymbol{m})$.

The *right* side of (4.44) is the SP *exit rate* of $((0, x), \boldsymbol{m})$. The term $f(0, \boldsymbol{m})$ is the rate at which the SP exits $((0, x), \boldsymbol{m})$ and simultaneously enters the zero-wait boundary state $(0, \boldsymbol{m})$, due to downcrossings of level 0. The term

$$\lambda \int_{z=0}^{x} p(z, \boldsymbol{m}, \boldsymbol{m}) e^{-\nu(z,\boldsymbol{m})(x-z)}) f(z, \boldsymbol{m}) dz$$

is the rate at which the SP exits $((0, x), \boldsymbol{m})$ and simultaneously enters $([x, \infty), \boldsymbol{m})$ due to jumps at arrival instants. The term

$$\lambda \sum_{\boldsymbol{k} \neq \boldsymbol{m}} \int_{z=0}^{x} p(z, \boldsymbol{m}, \boldsymbol{k}) f(z, \boldsymbol{m}) dz$$

is the rate at which the SP exits $((0, x), \boldsymbol{m})$ and simultaneously enters any page $\boldsymbol{k} \neq \boldsymbol{m}$. These three terms exhaust the possible paths by which the SP can exit $((0, x), \boldsymbol{m})$.

Thus equation (4.44) is a rate-balance equation of the form:

$$\textbf{\textit{Rate into}} \ ((0, x), \boldsymbol{m}) = \textbf{\textit{Rate out of}} \ ((0, x), \boldsymbol{m}),$$

which is a well-known principle for stochastic processes with discrete states, e.g., birth-death processes.

Interpretation of Equation for Total PDF

We now provide an LC interpretation of (4.45). We may view the LC analysis of the sheets as a *dissection* of the states of the model (into a partition). The total equation is like a *synthesis* , i.e., reconstruction of the parts into a single whole. This idea helps to derive model equations in complex models directly from sample-path considerations. It utilizes LC ideas for the sheets, lines and the 'cover'.

In (4.45) the *left* term $f(x)$ is the total downcrossing rate of level x, on all pages. On the right side, the term $\lambda \sum_{\boldsymbol{m} \in \boldsymbol{M}_1} P(0, \boldsymbol{m}) e^{-\nu(z, \boldsymbol{m})x}$ is the total rate at which the SP upcrosses level x at arrival instants, due to jumps starting at level 0. The term

$$\lambda \sum_{\boldsymbol{m} \in \boldsymbol{M}_1} \int_{z=0}^{x} e^{-\nu(z, \boldsymbol{m})(x-z)} f(z, \boldsymbol{m}) dz$$

is the total rate at which the SP upcrosses level x at arrival instants, starting from levels in $(0, x)$ on all pages $\boldsymbol{m} \in \boldsymbol{M}_1$. We form Eq. (4.45) by rate balance, with respect to level x

$$\textbf{\textit{Downcrossing rate}} = \textbf{\textit{Upcrossing rate}}.$$

The normalizing condition is

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

which too has an LC interpretation. That is, multiply both sides by λ. This yields

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

On the left side, λP_0 is the rate at which the SP makes jumps at arrival instants out of zero-wait states. The term $\lambda \int_{x=0}^{\infty} f(x)dx$ is the rate at which the SP makes jumps at arrival instants from positive-wait states. The right side λ is the total rate at which the SP makes jumps at arrival instants. The left and right sides are equal.

4.6.10 Discussion of Rate Balance in Complex Models

The rate-balance interpretation provides the analyst with a powerful technique for constructing model equations for steady-state distributions in very complex models. The method is straightforward, intuitive, and relatively easy.

1. Select a state-space interval with boundary x.
2. Express the SP entrance and exit rates of the interval algebraically in terms of the unknown probability of the interval and/or unknown pdf at x.
3. Apply rate balance to construct an integral equation (or other type of balance equation) for the probability and/or pdf at x.
4. Repeat (1)–(3) for every sub-partition of the state space as required, to form a complete system of Volterra integral equations of the second kind (as above), plus other relevant equations, depending on the model.
5. Write the normalizing condition.
6. Solve the entire system of equations simultaneously for the probabilities and pdfs of the model. This can be done analytically, numerically, by approximation, or by LC estimation (see Chap. 9).

Remark 4.11 The author realized in 1974 that the steady-state model equations discussed here, are really **rate-balance equations**. Originally, these steady-state equations had been derived by starting with Lindley recursions, analogous to those described for M/G/1 in Sect. 1.2 of Chap. 1. The derivation for M/M/c queues started, however, with more complex Lindley recursions.

4.7 Example of Steady-State Equations

This Section serves as a brief tutorial on writing steady-state model equations using LC and sample paths. We derive model equations for the steady-state pdf of wait in the *specific* M/M/c queue with $c = 3$ and $J = 1$, discussed in Example 4.4 in Sect. 4.5.5, with a sample path in Fig. 4.2. There are two possible service rates: $\boldsymbol{\mu} = \{\mu_0, \mu_1\}$. We make a slight generalization for the *service-rate* assignment policy. For each arrival, the rates $\{\mu_0, \mu_1\}$ are assigned with probabilities $\{\alpha_0, \alpha_1\}$, $\alpha_0 + \alpha_1 = 1$ (instead of $1/2$ each). Our present example reduces to Example 4.4 if $\alpha_0 = \alpha_1 = 1/2$.

We use α_0, α_1 to make it easier to follow the intuitive derivation of the model equations, since α_0, α_1 appear explicitly in the equations.

The set of possible configurations is $\boldsymbol{M}_0 \cup \boldsymbol{M}_1 = \{(m_0, m_1)\}$, where m_j denotes the number of servers occupied by customers with service rate μ_j, $j = 0, 1$. From the definition of system configuration (Sect. 4.4),

$$0 \leq \sum_{j=0}^{1} m_j \leq c - 1 = 2.$$

We abbreviate (m_0, m_1) as $m_0 m_1$. There are *six* possible configurations (same as in Example 4.4):

$$\boldsymbol{M}_0 \cup \boldsymbol{M}_1 = \{00, 10, 01, 20, 11, 02\}, \tag{4.46}$$

where

$$\boldsymbol{M}_0 = \{00, 10, 01\}, \qquad \boldsymbol{M}_1 = \{20, 11, 02\}.$$

When an arrival finds more than one server idle, it immediately occupies one of them in accordance with a *server-assignment* rule, and starts service at rate μ_i with probability α_i, $i = 0, 1$.

First we will derive the equations for the zero-wait states (atoms). These are represented in the virtual-wait diagram by the six lines $\boldsymbol{T} \times (0, \boldsymbol{m})$, $\boldsymbol{m} \in \boldsymbol{M}_0 \cup \boldsymbol{M}_1$ (Fig. 4.2).

Next we will derive the integral equations for the pdfs of the positive-wait states (continuous states). These states are represented by pages $\boldsymbol{T} \times ((0, \infty), m)$, $\boldsymbol{m} \in \boldsymbol{M}_1$ (Fig. 4.2). Fix level $x > 0$. For the equation corresponding to $\boldsymbol{m} \in \boldsymbol{M}_1$, the left side is the SP exit rate (*out of*) state-space interval $((x, \infty), \boldsymbol{m})$, and the right side is the SP entrance rate (*into*) $((x, \infty), \boldsymbol{m})$. (We use interval (x, ∞) instead of $(0, x)$, since (x, ∞) results in simpler (equivalent) equations.) Since $\boldsymbol{M}_1 = \{20, 11, 02\}$, there are three pages, three pdfs, and three corresponding integral equations.

Remark 4.12 To summarize, the zero-wait states are $(0, m), \boldsymbol{m} \in \boldsymbol{M}_0 \cup \boldsymbol{M}_1$. The positive-wait states we use for the derivation, are composite states $((x, \infty), \boldsymbol{m}), \boldsymbol{m} \in \boldsymbol{M}_1$. We could use alternative state-space intervals having a fixed level-x boundary , such as $((0, x), \boldsymbol{m})$ or $((x, a), \boldsymbol{m})$, where constant $a > x$, or $((x, bx), \boldsymbol{m}), b > 1$, etc. For different interval selections we would derive a different, but equivalent set of model equations. A creative choice of state-space interval may simplify the derivation and final form of the equations. It may lead to new identities or insights about the model. It may also suggest easier ways to obtain solutions of the equations.

The configurations for the zero-wait states are given in $\boldsymbol{M}_0 \cup \boldsymbol{M}_1$ and for the pages in $\boldsymbol{M}_1$, in (4.46) above; (see also Fig. 4.2).

We now derive the model equations; a detailed explanation follows immediately after.

4.7.1 Equations for Zero-Wait States

Notation 4.9 State $(0, m_0 m_1)$ means $m_0 + m_1$ servers are occupied: m_i serve at rate μ_i, $i = 0, 1$. $P_{m_0 m_1} := P(\text{system is in state } (0, m_0 m_1))$.

Using the principle for discrete states ***rate out = rate in***, we obtain the equations for the zero-wait states, as in (4.47). A detailed explanation follows below.

State	Rate out		Rate in
$(0, 00)$	λP_{00}	$=$	$\mu_0 P_{10} + \mu_1 P_{01}$
$(0, 10)$	$(\lambda + \mu_0) P_{10}$	$=$	$\lambda\alpha_0 P_{00} + 2\mu_0 P_{20} + \mu_1 P_{11}$
$(0, 01)$	$(\lambda + \mu_1) P_{01}$	$=$	$\lambda\alpha_1 P_{00} + 2\mu_1 P_{02} + \mu_0 P_{11}$
$(0, 20)$	$(\lambda + 2\mu_0) P_{20}$	$=$	$\lambda\alpha_0 P_{10} + f_{20}(0^+)$
$(0, 11)$	$(\lambda + \mu_0 + \mu_1) P_{11}$	$=$	$\lambda\alpha_1 P_{10} + \lambda\alpha_0 P_{01} + f_{11}(0^+)$
$(0, 02)$	$(\lambda + 2\mu_1) P_{02}$	$=$	$\lambda\alpha_1 P_{01} + f_{02}(0^+)$

(4.47)

Explanation for Discrete States $(0, \boldsymbol{m}), \boldsymbol{m} \in \boldsymbol{M}_0 \cup \boldsymbol{M}_b$
In (4.47) the first three equations are derived as in a "bubble" diagram for discrete-state continuous-time Markov chains, using *rate out = rate in.* The last three equations are derived similarly, except for the terms $f_{20}(0^+)$, $f_{11}(0^+)$, $f_{02}(0^+)$. These are the exit rates from $((0, \infty), 20)$, $((0, \infty), 11)$, and $((0, \infty), 02)$ into discrete states $(0, 20)$, $(0, 11)$, $(0, 02)$ respectively. At instants of these exits, the SP simultaneously enters the corresponding line $\boldsymbol{T} \times (0, 20)$, $\boldsymbol{T} \times (0, 11)$, or $\boldsymbol{T} \times (0, 02)$. It continues its motion.

4.7.2 Equations for States $((x,\infty),m)$, $m \in M_1$

We now derive the model equations for pages $\boldsymbol{m} \in \boldsymbol{M}_1$. Detailed explanations follow immediately after Eq. (4.51) below.

Rate balance of rates out (left side) and in (right side) for composite state $((x,\infty),20)$, $x>0$, result in the equation

$$\begin{aligned}
&f_{20}(x)+\lambda\alpha_1\frac{2\mu_0}{2\mu_0+\mu_1}\int_{y=x}^{\infty} f_{20}(y)dy\\
&=\lambda\left(\alpha_0 e^{-3\mu_0 x}+\alpha_1\frac{\mu_1}{2\mu_0+\mu_1}e^{-(2\mu_0+\mu_1)x}\right)P_{20}\\
&\quad+\lambda\alpha_0\frac{\mu_1}{2\mu_0+\mu_1}e^{-(2\mu_0+\mu_1)x}P_{11}+\lambda\alpha_0\int_{y=0}^{x} e^{-3\mu_0(x-y)}f_{20}(y)dy\\
&\quad+\lambda\alpha_0\frac{\mu_1}{2\mu_0+\mu_1}\int_{y=0}^{x} e^{-(2\mu_0+\mu_1)(x-y)}f_{11}(y)dy\\
&\quad+\lambda\alpha_0\frac{\mu_1}{2\mu_0+\mu_1}\int_{y=x}^{\infty} f_{11}(y)dy. \qquad (4.48)
\end{aligned}$$

Rate balance for composite state $((x,\infty),11)$, $x>0$, gives the equation

$$\begin{aligned}
&f_{11}(x)+\lambda\alpha_1\frac{\mu_0}{\mu_0+2\mu_1}\int_{y=x}^{\infty} f_{11}(y)dy+\lambda\alpha_0\frac{\mu_1}{2\mu_0+\mu_1}\int_{y=x}^{\infty} f_{11}(y)dy\\
&=\lambda\left(\alpha_1\frac{2\mu_1}{\mu_0+2\mu_1}e^{-(\mu_0+2\mu_1)x}+\alpha_0\frac{2\mu_0}{2\mu_0+\mu_1}e^{-(2\mu_0+\mu_1)x}\right)P_{11}\\
&\quad+\lambda\alpha_1\frac{2\mu_0}{2\mu_0+\mu_1}e^{-(2\mu_0+\mu_1)x}P_{20}+\lambda\alpha_0\frac{2\mu_1}{\mu_0+2\mu_1}e^{-(\mu_0+2\mu_1)x}P_{02}\\
&\quad+\lambda\alpha_1\frac{2\mu_1}{\mu_0+2\mu_1}\int_{y=0}^{x} e^{-(\mu_0+2\mu_1)(x-y)}f_{11}(y)dy\\
&\quad+\lambda\alpha_0\frac{2\mu_0}{2\mu_0+\mu_1}\int_{y=0}^{x} e^{-(2\mu_0+\mu_1)(x-y)}f_{11}(y)dy\\
&\quad+\lambda\alpha_1\frac{2\mu_0}{2\mu_0+\mu_1}\int_{y=0}^{x} e^{-(2\mu_0+\mu_1)(x-y)}f_{20}(y)dy\\
&\quad+\lambda\alpha_0\frac{2\mu_1}{\mu_0+2\mu_1}\int_{y=0}^{x} e^{-(\mu_0+2\mu_1)(x-y)}f_{02}(y)dy\\
&\quad+\lambda\alpha_1\frac{2\mu_0}{2\mu_0+\mu_1}\int_{y=x}^{\infty} f_{20}(y)dy+\lambda\alpha_0\frac{2\mu_1}{\mu_0+2\mu_1}\int_{y=x}^{\infty} f_{02}(y)dy. \qquad (4.49)
\end{aligned}$$

Rate balance for composite state $((x, \infty), 02)$, $x > 0$, gives the equation

$$\begin{aligned} & f_{02}(x) + \lambda\alpha_0 \frac{2\mu_1}{\mu_0 + 2\mu_1} \int_{y=x}^{\infty} f_{02}(y)dy \\ &= \lambda \left(\alpha_1 e^{-3\mu_1 x} + \alpha_0 \frac{\mu_0}{\mu_0 + 2\mu_1} e^{-(\mu_0+2\mu_1)x} \right) P_{02} \\ &\quad + \lambda\alpha_1 \frac{\mu_0}{\mu_0 + 2\mu_1} e^{-(\mu_0+2\mu_1)x} P_{11} + \lambda\alpha_1 \int_{y=0}^{x} e^{-3\mu_1(x-y)} f_{02}(y)dy \\ &\quad + \lambda\alpha_1 \frac{\mu_0}{\mu_0 + 2\mu_1} \int_{y=0}^{x} e^{-(\mu_0+2\mu_1)(x-y)} f_{11}(y)dy \\ &\quad + \lambda\alpha_1 \frac{\mu_0}{\mu_0 + 2\mu_1} \int_{y=x}^{\infty} f_{11}(y)dy. \end{aligned} \tag{4.50}$$

The normalizing condition is

$$\begin{aligned} & P_{00} + P_{10} + P_{01} + P_{20} + P_{11} + P_{02} \\ & + \int_{x=0}^{\infty} [f_{20}(x) + f_{11}(x) + f_{02}(x)]\, dx = 1. \end{aligned} \tag{4.51}$$

Explanation of Equations for States $((x, \infty), \boldsymbol{m})$, $\boldsymbol{m} \in \boldsymbol{M}_1$

Left Side of Equation (4.48) In (4.48), the left side represents the SP *exit* rate out of $((x, \infty), 20)$. There are two routes by which the SP can exit this composite state: (1) downcrossing level x on page 20; (2) jumping to page 11 pursuant to an arrival that is assigned rate μ_1. The term $f_{20}(x)$ is the downcrossing rate of level x on page 20.

The term

$$\lambda\alpha_1 \frac{2\mu_0}{2\mu_0 + \mu_1} \int_{y=x}^{\infty} f_{20}(y)dy$$

is the rate at which the SP jumps to page 11 at arrival instants. In this expression, $\lambda f_{20}(y)dy$ is the rate at which arrivals find the SP in interval $(y, y + dy)$ on page 20. The term α_1 is the probability that an arrival gets assigned rate μ_1, resulting in two servers having rate μ_0 and one server having rate μ_1 just after the arrival starts service. The term $\frac{2\mu_0}{2\mu_0+\mu_1}$ is the probability that a rate-μ_0 customer finishes first, causing the SP to jump to page 11. The SP cannot jump to page 02 if an arrival finds the configuration to be 20. The sum of the two terms on the left of side of (4.48) is the *exit* rate of the SP out of (x, ∞) on page 20.

Right Side of Equation (4.48) The right side of (4.48) is the SP *entrance* rate into $((x, \infty), 20)$. The fist term

$$\lambda\left(\alpha_0 e^{-3\mu_0 x} + \alpha_1 \frac{\mu_1}{2\mu_0 + \mu_1} e^{-(2\mu_0+\mu_1)x}\right) P_{20}$$

is the entrance rate into $((x, \infty), 20)$ due to arrivals that find the state to be $(0, 20)$. In it, the product λP_{20} is the rate at which arrivals find the state to be $(0, 20)$. The arrival does not wait, and immediately starts service from the one available server. The term $\alpha_0 e^{-3\mu_0 x}$ is the product of two probabilities: α_0, that the arrival is assigned rate μ_0; $e^{-3\mu_0 x}$, that the minimum of three independent service times, each having rate μ_0, exceeds x.

The term

$$\alpha_1 \frac{\mu_1}{2\mu_0 + \mu_1} e^{-(2\mu_0+\mu_1)x}$$

is the product of three probabilities: α_1, that the arrival is assigned rate μ_1; $\mu_1/(2\mu_0 + \mu_1)$, that the minimum of three service times, two having rate μ_0 and one having rate μ_1, is the rate μ_1; $e^{-(2\mu_0+\mu_1)x}$, that the minimum of the three service times exceeds x. Both terms result in the SP landing above x on page 02. The entire term is the rate at which the SP moves from level 0 on page 20 to interval (x, ∞) on page 20.

The term

$$\lambda\alpha_0 \frac{\mu_1}{2\mu_0 + \mu_1} e^{-(2\mu_0+\mu_1)x} P_{11}$$

is the rate at which arrivals find the state to be $(0, 11)$ (rate λP_{11}), are assigned service rate μ_0 (probability α_0), the minimum service time is a rate-μ_1 service (probability $\mu_1/(2\mu_0 + \mu_1)$), and the minimum exceeds x (probability $e^{-(2\mu_0+\mu_1)x}$). This is the rate at which the SP moves from discrete level 0 on page 11 to (x, ∞) on page 20.

The term

$$\lambda\alpha_0 \int_{y=0}^{x} e^{-3\mu_0(x-y)} f_{20}(y)dy$$

is the rate at which arrivals find the state to be $(y, 20)$, $y \in (0, x)$, are assigned service rate μ_0 (probability α_0), and the minimum of three service times each having rate μ_0 exceeds $x - y$ (probability $e^{-3\mu_0(x-y)}$) integrated over all $y \in (0, x)$. This is the rate at which the SP moves from $(0, x)$ on page 20 to (x, ∞) on page 20 (makes $20 \to 20$ upcrossings of x).

The term

$$\lambda\alpha_0 \frac{\mu_1}{2\mu_0 + \mu_1} \int_{y=0}^{x} e^{-(2\mu_0+\mu_1)(x-y)} f_{11}(y)dy$$

is the rate at which arrivals find the state to be in $((y, y + dy), 11)$, $y \in (0, x)$ (factor $\lambda f_{11}(y)dy$), are assigned service rare μ_0, the rate-μ_1 service ends first, and the minimum of three exponential r.v.s (two having rate μ_0 and one rate μ_1) exceeds $x - y$, integrated over all $y \in (0, x)$. This is the rate at which the SP moves from $(0, x)$ on page 11 to (x, ∞) on page 20 (makes $11 \to 20$ upcrossing of x).

The term

$$\lambda\alpha_0 \frac{\mu_1}{2\mu_0 + \mu_1} \int_{y=x}^{\infty} f_{11}(y)dy$$

is the rate at which arrivals find the state to be in $((y, y + dy), 11)$, $y > x$, are assigned service rate μ_0, the rate-μ_1 service finishes first, and the minimum of three exponential service times (two having rate μ_0 and one having rate μ_1) has any value in (x, ∞). This is the rate at which the SP moves from (x, ∞) on page 11 to (x, ∞) on page 20 (makes $11 \to 20$ transition, from and to, points above x).

Integral Equations (4.49) and (4.50)
We derive integral equations (4.49) and (4.50) for the pdfs $f_{11}(x)$ and $f_{02}(x)$ (pages 11 and 02), in a similar manner as for $f_{20}(x)$ above.

Normalizing Condition
The normalizing condition (4.51) ensures that the sum of all zero-wait and positive-wait probabilities is 1.

4.8 Standard M/M/c: Steady-State Analysis

We analyze the standard M/M/c queue as a special case of the generalized M/M/c queue developed in Sects. 4.3–4.7. It is instructive to derive known results for M/M/c using LC. *The standard M/M/c queue is one of the first models the author analyzed in 1974, to validate the LC method* (see pp. 37–39 in [11]).

We assume the number of servers is $c \geq 2$, each customer receives the same exponential service rate μ, and $\lambda < c\mu$. Using the notation of Sect. 4.3, we have here $J = 0$, $\boldsymbol{\mu} = \{\mu_0\} := \{\mu\}$. A system configuration has one component m_0, which can take values in $\{0, 1, \ldots, c - 1\}$. The virtual wait process is denoted as $\{W(t)\}_{t \geq 0}$.

In this model, a system configuration is a scalar $m_0 :=$ *number of customers in the other servers just after an arrival starts service*. Thus $m_0 \in \{0, 1, \ldots, c - 1\}$. Equivalently m_0 is the number of *other* occupied servers at a start of service instant. The set of all configurations, $\boldsymbol{M} = \boldsymbol{M}_0 \cup \boldsymbol{M}_1$, has size $\binom{J+c}{c-1} = \binom{0+c}{c-1} = \binom{c}{1} = c$ (see Sect. 4.4.2). That is,

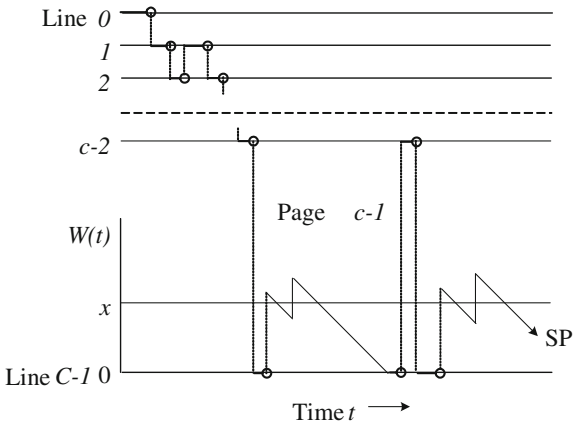

Fig. 4.4 Sample path of $\{W(t)\}_{t\geq 0}$ in *standard* M/M/c queue. There are c lines and *one* page. Line $c-1$ is at level 0 of page $c-1$

$$\boldsymbol{M}_0 = \{0, 1, \ldots, c-2\},\ \boldsymbol{M}_1 = \{c-1\}.$$

(Recall that $\boldsymbol{M}_1 = \boldsymbol{M}_b$, the set of 'border' configurations.)

A sample path of $\{W(t)\}_{t\geq 0}$ has c *lines* for the zero-wait states $(0, j)$, $j = 0, \ldots, c-1$, and *one page* (*sheet*) for the composite state $((0, \infty), c-1)$ (Fig. 4.4). Line $c-1$, the *border* line corresponding to state $(0, c-1)$, is usually placed as the bottom line of page $c-1$, but is arbitrarily located among the other 0-wait states in Fig. 4.4. This does not change the analysis because rate balance across level $x > 0$ (*downcrossing rate = upcrossing rate*) is equivalent to rate balance between sets $((x, \infty), c-1)$ and $([0, x], c-1)$ i.e., *rate from* $((x, \infty), c-1)$ *into* $([0, x], c-1)$ = *rate from* $([0, x], c-1)$ *into* $((x, \infty), c-1)$.

Denote the zero-wait probabilities as P_n, $n = 0, \ldots, c-1$, the pdf of wait as $f(x)$, $x > 0$, and the steady-state cdf of wait by $F(x)$, $x \geq 0$. Then

$$F(x) = \sum_{n=0}^{c-1} P_n + \int_0^x f(x)dx,\ x \geq 0,$$

$$F(0) = \sum_{n=0}^{c-1} P_n.$$

4.8.1 Equations for Steady-State PDF of Wait

We derive model equations for the steady-state pdf of wait, and give further explanations in Sect. 4.8.2 below.

Zero-Wait States
For the zero-wait states (atoms) the model equations are (using *rate out* = *rate in*)

$$\begin{aligned}
\lambda P_0 &= \mu P_1 \\
(\lambda + \mu) P_1 &= \lambda P_0 + 2\mu P_2 \\
(\lambda + 2\mu) P_2 &= \lambda P_1 + 3\mu P_3 \\
&\cdots \\
(\lambda + (c-2)\mu) P_{c-2} &= \lambda P_{c-3} + (c-1)\mu P_{c-1} \\
(\lambda + (c-1)\mu) P_{c-1} &= \lambda P_{c-2} + f(0^+).
\end{aligned} \tag{4.52}$$

The term $f(0^+)$ in the last equation in (4.52) connects the *pdf of a continuous random variable* (*waiting time*) with the *probabilities of atoms* (states $(0, c-1)$ and $(0, c-2)$). This observation (and other examples) led the author to coin the term "*border state*" (i.e., state $(0, c-1)$ in the present context).

Positive-Wait States
For the composite state $((0, \infty), c-1)$ (the single page) the model equation is

$$f(x) = \lambda P_{c-1} e^{-c\mu x} + \lambda \int_{y=0}^{x} e^{-c\mu(x-y)} f(y) dy, \; x > 0. \tag{4.53}$$

Composite state $((0, \infty), c-1)$ is accessible in one step at an arrival instant, only from the border state $(0, c-1)$. The normalizing condition is

$$F(\infty) = \sum_{n=0}^{c-1} P_n + \int_{y=0}^{\infty} f(x) dx = 1. \tag{4.54}$$

4.8.2 Explanation of Equations (4.52) and (4.53)

Linear Equations (4.52)
Equation (4.52) are rate-balance equations, which equate SP rates out of, and into, the discrete zero-wait states $(0, n)$, $n = 0, \ldots, c-1$. The term $f(0^+)$ $(:= f(0))$ is the SP downcrossing rate of level 0, i.e., the SP *entrance rate into state* $(0, c-1)$ from above. (In sample-paths, line $c-1$ may be equally placed at level 0 of page $(c-1)$. If it is placed separately as in Fig. 4.4, we can still imagine it to be *at level 0* of the page with respect to SP motion.)

Integral Equation (4.53)
To derive the positive-wait integral equation (4.53) consider composite state $((x, \infty), c-1)$ on the (single) page (Fig. 4.4). We equate the SP *exit rate*

(i.e., downcrossing rate of level x) to the *entrance rate* (i.e., 'upcrossing' rate of level x starting from line $c-1$ thought of as being at the bottom of the page). The downcrossing rate of level x is $f(x)$ (see Corollary 4.2 in Sect. 4.6.2).

The SP entrance rate into $((x,\infty),c-1)$ is from two sources:
(1) Entrances are generated by jumps due to arrivals when the state is the border state $(0,c-1)$, starting from level 0 of the page and ending above level x on the page. Since there is only one page, the only access to $((x,\infty),c-1)$ in one step from a zero-wait state is from state $(0,(c-1))$, i.e., line $c-1$ in the sample path. The SP entrance rate from this source is $\lambda P_{c-1}\cdot P(\mathcal{S}>x)$, where P_{c-1} is the limiting probability of state $(0,c-1)$, and $\mathcal{S}$ is the *inter start-of-service depart time*. (See Sect. 4.4.1 for a discussion of *inter start-of-service depart time*.) Random variable $S \underset{dis}{=} \text{Exp}_{c\mu}$, since there would be c customers with rate μ in service just after such an arrival starts service, and $S :=$ *minimum of* c *i.i.d.* Exp_μ *random variables*. Thus, $P(\mathcal{S}>x)=e^{-c\mu x}$. This gives the term $\lambda P_{c-1}e^{-c\mu x}$ in (4.53).
(2) Entrances into $((x,\infty),c-1)$ are generated by jumps due to arrivals when the state is a continuous state $(y,c-1)$, $y\in(0,x)$. Such jumps start at level y and end above level x. Just after such an arrival begins service (y after its arrival), all c servers will be occupied and $S \underset{dis}{=} \text{Exp}_{c\mu}$, independent of any new arrivals to the system. The SP will enter $((x,\infty),c-1)$ with probability $e^{-c\mu(x-y)}$. This leads to Eq. (4.53).

4.8.3 Solution of Equations

We first solve (4.53). Differentiating both sides with respect to x and solving the resulting first-order differential equation, gives

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

where A is a constant. Letting $x\downarrow 0$, we get the initial condition

$$f(0)=A=\lambda P_{c-1} \tag{4.55}$$

since $f(0)$ $(:=f(0^+))$ is the SP downcrossing rate of level 0, and λP_{c-1} is the "upcrossing" rate of level 0 (rate of egress from $(0,c-1)$ above). (Equivalently, $f(0)$ is the exit rate out of $((0,\infty),c-1)$ and λP_{c-1} is the entrance rate into $((0,\infty),c-1)$.) Thus $A=\lambda P_{c-1}$ and

$$f(x)=\lambda P_{c-1}e^{-(c\mu-\lambda)x},\ x>0. \tag{4.56}$$

Note that the condition (4.55) is itself a rate-balance equation for the rates out of, and into, $((0, \infty), c-1)$.

Next, from (4.52) and (4.56) we obtain

$$\begin{aligned} P_n &= \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} P_0, n = 0, \ldots, c-1, \\ P_{c-1} &= \left(\frac{\lambda}{\mu}\right)^{c-1} \frac{1}{(c-1)!} P_0. \end{aligned} \tag{4.57}$$

Substituting (4.57) into (4.56) gives

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

The normalizing condition (4.54) is

$$\left(\sum_{n=0}^{c-1} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}\right) P_0 + \lambda \left(\frac{\lambda}{\mu}\right)^{c-1} \frac{1}{(c-1)!} P_0 \int_{x=0}^{\infty} e^{-(c\mu-\lambda)x} dx = 1.$$

This gives the well-known value

$$P_0 = \frac{1}{\sum_{n=0}^{c-1} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} + \left(\frac{\lambda}{\mu}\right)^c \frac{c\mu}{c!(c\mu-\lambda)}}. \tag{4.58}$$

The cdf of wait is

$$\begin{aligned} F(x) &= P_0 + \int_{y=0}^{x} \lambda P_{c-1} e^{-(c\mu-\lambda)y} dy \\ &= P_0 \left(1 + \lambda \left(\frac{\lambda}{\mu}\right)^{c-1} \frac{1}{(c-1)!(c\mu-\lambda)} \left(1 - e^{-(c\mu-\lambda)x}\right)\right), x \geq 0. \end{aligned} \tag{4.59}$$

Boundedness of PDF of Wait

From (4.56) $f(x) < \lambda, x > 0$, since $P_{c-1} < 1$ and $e^{-(c\mu-\lambda)x} < 1$ $(c\mu - \lambda > 0)$.

4.8.4 CDF and PDF of Wait Geometrically

It is insightful and intuitive to derive the steady-state cdf and pdf of wait geometrically, directly from sample path properties. This derivation bypasses model equation (4.53). A similar geometric derivation for the cdf of wait in the M/M/1 queue is given in Sect. 3.5.6.

Consider level $x > 0$ on the single page (Fig. 4.4). Rate balance across level x applies the principle

$$\textit{Upcrossing rate of } x = \textit{Downcrossing rate of } x = f(x).$$

Equivalently, in symbols

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

or

$$\lim_{t\to\infty} \frac{E(\mathcal{U}_t(x))}{t} = \lim_{t\to\infty} \frac{E(\mathcal{D}_t(x))}{t} = f(x).$$

The sojourn time above level $x > 0$ on the page, initiated by each upcrossing of x, is := *busy period of a standard* $\mathrm{M}_\lambda/\mathrm{M}_{c\mu}/1$ queue with arrival rate λ and *service rate* $c\mu$ because, when the SP is on the page, all c servers are occupied and each is serving at rate μ. Thus the inter start-of-service depart time (see Definition 4.3 in Sect. 4.4.1) $\mathcal{S} \underset{dis}{=}$ *size of each jump ending on the page* $\underset{dis}{=} \mathrm{Exp}_{c\mu}$. Moreover, by the memoryless property, excess jumps above level x are $\underset{dis}{=} \mathrm{Exp}_{c\mu}$.

Let a_x denote an SP sojourn time above x. Then $a_x \underset{dis}{=}$ *busy period in* $\mathrm{M}_\lambda/\mathrm{M}_{c\mu}/1$ $(\lambda < c\mu)$. Thus

$$E(a_x) = \frac{1}{c\mu - \lambda}, \tag{4.60}$$

independent of x, since the expected value of the busy period in $M_\lambda/M_{c\mu}/1$ is $1/(c\mu - \lambda)$.

Let $d_x :=$ *inter-downcrossing time at level* $x \geq 0$. Since level-x downcrossings are regenerative points, similarly as in Sect. 3.4.15 we have

$$E(d_x) = 1/f(x). \tag{4.61}$$

The renewal reward theorem (Sect. 3.79), now yields

$$\frac{E(a_x)}{E(d_x)} = \lim_{t\to\infty} \frac{\text{time } W(\cdot) \in (x,\infty) \text{ during } (0,t)}{t} = 1 - F(x),$$

$$\frac{1/(c\mu - \lambda)}{1/f(x)} = 1 - F(x),$$

or

$$\frac{f(x)}{1 - F(x)} = c\mu - \lambda, x > 0, \tag{4.62}$$

equivalent to the differential equation

$$\frac{\frac{d}{dx}(1 - F(x))}{1 - F(x)} = -(c\mu - \lambda),$$

$$\frac{d}{dx} \ln(1 - F(x)) = -(c\mu - \lambda),$$

with solution

$$1 - F(x) = A \cdot e^{-(c\mu-\lambda)x},$$

where A is a constant, evaluated by letting $x \downarrow 0$, and yielding the cdf of wait

$$F(x) = 1 - (1 - F(0))e^{-(c\mu-\lambda)x}, x \geq 0, \tag{4.63}$$

where $F(0) = P(\text{zero wait})$. Taking $dF(x)/dx$, $x > 0$, in (4.63) gives the pdf of wait

$$f(x) = (1 - F(0))(c\mu - \lambda)e^{-(c\mu-\lambda)x}, x > 0. \tag{4.64}$$

We next employ the equations in (4.57) to get

$$F(0) = \sum_{n=0}^{c-1} P_n = P_0 \sum_{n=0}^{c-1} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}. \tag{4.65}$$

Note that $f(0) = \lambda P_{c-1}$., i.e., the SP entrance rate into state $(0, c-1)$ from above (downcrossing rate of level 0) is equal to the SP exit rate from state $(0, c-1)$ at arrival instants. Letting $x \downarrow 0$ In (4.64) yields

$$f(0) = (1 - F(0))(c\mu - \lambda) = \lambda P_{c-1}. \tag{4.66}$$

From (4.66) and (4.57)

$$F(0) = 1 - \frac{\lambda}{c\mu - \lambda} P_{c-1} = 1 - \frac{\lambda}{c\mu - \lambda} \left(\frac{\lambda}{\mu}\right)^{c-1} \frac{1}{(c-1)!} P_0. \tag{4.67}$$

Substituting the value of $F(0)$ from (4.65) into (4.67) and solving for P_0 gives (4.58). The upshot is two different ways to determined P_0; and two different, equivalent formulas for $f(x), x > 0$: (4.56) and (4.64).

Remark 4.13 Another way to obtain the second equality in (4.66) is to note that the SP expected sojourn time above 0 is

$$E(a_0) = E(\text{busy period of } M_\lambda/M_{c\mu}/1) = \frac{1}{c\mu - \lambda}.$$

The proportion of time the SP spends above level 0 is therefore

$$\lim_{t\to\infty} \frac{E(\mathcal{U}_t(0))}{t} \cdot \frac{1}{c\mu - \lambda} = \lambda P_{c-1} \cdot \frac{1}{c\mu - \lambda} = 1 - F(0).$$

Busy Period in M/M/c
Note that a_0 is equal to a busy period in M/M/c, denoted by $\mathcal{B}_{c-1,c}$, defined as the time measured from an arrival instant when the state is $(0, c-1)$ until the first departure instant thereafter that leaves the system in state $(0, c-1)$ again. (The arrival increases the number in the system to c. The departure decreases the number to $c-1$.) Since $a_x \equiv a_0, x \geq 0$,

$$E\left(\mathcal{B}_{c-1,c}\right) = E\left(a_0\right) = E\left(a_x\right) = \frac{1}{c\mu - \lambda}, x \geq 0. \tag{4.68}$$

We also call $\mathcal{B}_{c-1,c}$ a $[c-1, c]$ busy period.

4.8.5 PMF of Number in the System

We use the foregoing pdf of wait (4.56) to derive $P_n, n = c, c+1, \ldots$. This approach is the reverse order of the usual derivation, which first derives the pmf (probability mass function) of the number-in-system using a birth-death analysis. It then obtains the pdf of wait by conditioning on the number in the system when there is an arrival. The method we apply here utilizes partly birth-death analysis and partly LC. It provides a different perspective on the M/M/c model.

Due to Poisson arrivals, $P_n = a_n = d_n$, where a_n, d_n are the steady-state probabilities of n units in the system just before an arrival, and just after a departure, respectively (in this Section). Reasoning as for M/M/1 (see Sect. 3.5.3), we get

$$P_n = d_n = P(n - c \text{ arrivals during a waiting time}), n = c, c+1, \ldots .$$

Substituting from (4.56) and (4.57)

$$\begin{aligned} P_n &= \int_{x=0}^{\infty} \frac{e^{-\lambda x}(\lambda x)^{n-c}}{(n-c)!} f(x)dx \\ &= \left(\frac{\lambda}{\mu}\right)^{n-c+1} \frac{1}{c^{n-c+1}} P_{c-1} \int_{x=0}^{\infty} c\mu e^{-c\mu x} \frac{(c\mu x)^{n-c}}{(n-c)!} dx \\ &= \left(\frac{\lambda}{\mu}\right)^{n} \frac{1}{c^{n-c}c!} P_0, n = c, c+1, \ldots . \end{aligned}$$

In summary, we obtain the well-known formulas (e.g., p. 67 in [84])

$$\left.\begin{aligned} P_0 &= \frac{1}{\sum_{n=0}^{c-1} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} + \left(\frac{\lambda}{\mu}\right)^c \frac{c\mu}{c!(c\mu-\lambda)}} \\ P_n &= \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} P_0, n = 0, \ldots, c-1, \\ P_n &= \left(\frac{\lambda}{\mu}\right)^n \frac{1}{c^{n-c}} \frac{1}{c!} P_0, n = c, c+1, \ldots . \end{aligned}\right\} \tag{4.69}$$

The probability that all servers are occupied is

$$\begin{aligned} \sum_{n=c}^{\infty} P_n &= P(\text{wait} > 0) = \int_{x=0}^{\infty} f(x)dx \\ &= \lambda P_{c-1} \int_{x=0}^{\infty} e^{-(c\mu-\lambda)x} dx = \frac{\lambda}{c\mu-\lambda} P_{c-1} \\ &= \frac{\lambda \left(\frac{\lambda}{\mu}\right)^{c-1} \frac{1}{c!}}{c\mu-\lambda} P_0. \end{aligned} \tag{4.70}$$

The probability that there is at least one idle server is

$$\sum_{n=0}^{c-1} P_n = P(\text{wait} = 0) = 1 - \frac{\lambda \left(\frac{\lambda}{\mu}\right)^{c-1} \frac{1}{c!}}{c\mu-\lambda} P_0. \tag{4.71}$$

4.8.6 Inter-downcrossing and Sojourn Times

Consider d_x, a_x, b_x $(x \geq 0)$, respectively: time between successive SP downcrossings of level x; sojourn time above x initiated by an upcrossing of x; sojourn time at or below x initiated by a downcrossing of x. Formula (4.61)

shows

$$E(d_x) = \frac{1}{f(x)} = \frac{e^{(c\mu-\lambda)x}}{\lambda P_{c-1}}, x \geq 0; \tag{4.72}$$

formula (4.60) shows that, independent of x,

$$E(a_x) = \frac{1}{c\mu - \lambda}.$$

Note that

$$\lim_{t\to\infty} \frac{(\textit{time that the SP is above } x \textit{ during } (0,t))}{t} = 1 - F(x);$$

by the renewal reward theorem (Sect. 3.4.9),

$$\frac{E(a_x)}{E(d_x)} = \frac{E(a_x)}{1/f(x)} = 1 - F(x),$$
$$E(a_x) = \frac{1 - F(x)}{f(x)} = \frac{1}{c\mu - \lambda}., x > 0,$$

The last equality above corroborates formula (4.62) when solving for $f(x)$ geometrically in Sect. 4.8.4. Also, we can validate (4.62) in Sect. 4.8.4 using

$$\frac{1 - F(x)}{f(x)} = \frac{\int_{y=x}^{\infty} f(y)dy}{\lambda P_{c-1} e^{-(c\mu-\lambda)x}}$$
$$= \frac{\int_{y=x}^{\infty} \lambda P_{c-1} e^{-(c\mu-\lambda)y} dy}{\lambda P_{c-1} e^{-(c\mu-\lambda)x}} = \frac{1}{c\mu - \lambda}.$$

Note that $F(x) = \lim_{t\to\infty}(\textit{time the SP spends at or below } x \textit{ during } (0,t))/t$. Each instant that the SP downcrosses $x \geq 0$ is a regenerative point, due to the memoryless property of the interarrival times. From the renewal reward theorem (i.e., the theory of regenerative processes, e.g., [134]) $E(b_x)/E(d_x) = F(x)$, implying

$$E(b_x) = \frac{F(x)}{f(x)} = \frac{1 - (1 - F(0))e^{-(c\mu-\lambda)x}}{\lambda P_{c-1} e^{-(c\mu-\lambda)x}}$$
$$= \frac{e^{(c\mu-\lambda)x}}{\lambda P_{c-1}} - \frac{(1 - F(0))}{\lambda P_{c-1}}$$
$$= \frac{1}{\lambda \left(\frac{\lambda}{\mu}\right)^{c-1} \frac{1}{(c-1)!} P_0} \left(e^{(c\mu-\lambda)x} - (1 - P_0 \sum_{n=0}^{c-1} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}) \right)$$

$$= \frac{e^{(c\mu-\lambda)x}-1}{\lambda\left(\frac{\lambda}{\mu}\right)^{c-1}\frac{1}{(c-1)!}P_0} + \frac{\sum_{n=0}^{c-1}\left(\frac{\lambda}{\mu}\right)^n\frac{1}{n!}}{\lambda\left(\frac{\lambda}{\mu}\right)^{c-1}\frac{1}{(c-1)!}}. \tag{4.73}$$

Remark 4.14 From (4.60), when the SP upcrosses x, it next downcrosses x after a time a_x where $E(a_x)$ is independent of x. By contrast, (4.73) implies when the SP downcrosses level x, it next upcrosses x after a time b_x where $E(b_x)$ grows exponentially with increasing x.

The foregoing results for

$$d_x,\ a_x,\ b_x, E(d_x),\ E(a_x), E(b_x)$$

generalize analogous results for M/M/1 (Sect. 3.5.7).

4.9 M/M/c/c and Standard M/M/c Queues

The M/M/c/c queue is a special case of M/M/c/k, in which an upper limit k is placed on the number of customers allowed in the system at any time (see, e.g., Sect. 2.5, p. 76ff in [84]). Here, we develop a relationship between M/M/c/c and the standard M_λ/M_μ/c queue. By a judicious choice of parameters for M/M/c/c, the pdf of the virtual wait in the two models have identical forms. However, the jump structure of the sample path of M/M/c/c is much simpler than that of the corresponding M/M/c model, for positive values of the virtual wait. This jump structure makes it much easier to derive the pdf of the virtual wait in M/M/c/c. The point of this exercise is to obtain the pdf of wait in the parameter-modified M/M/c/c queue, which can be derived in one line, without having to solve an integral equation (as in M/M/c), and the derived pdf is the same as in M/M/c. This relationship suggests a broader prospect. For a given complex model, can we identify a related model having the same solution form, that can be solved more easily?

The M/M/c/c queue is usually analyzed using a birth-death analysis. Here, we employ an LC approach. Consider an M/M/c/c queue where the service time for each customer that enters the system has exponential rate $\mu - \frac{\lambda}{c} > 0$. (We choose $\lambda < c\mu$ because our related model is a standard M_λ/M_μ/c queue in equilibrium.)

In M/M/c/c all *actual* waits are 0 – there is no waiting line. In a queue where blocking is possible, we shall define the virtual wait as the time that a potential arrival *would* wait to start service, if it were not blocked and cleared.

Thus the virtual wait is not 0 for every arrival. In M/M/c/c, customers that arrive when the *virtual* wait is positive, are blocked and cleared from the system. In both models, the virtual wait is positive if and only if all c servers are occupied.

For M/M/c/c, consider the 'system point' process $\{W(t), M(t)\}_{t\geq 0}$, where $W(t)$ is the virtual wait and $M(t) \in \{0, \ldots, c-1\}$ is the system configuration at time t. $M(t)$ is the number of occupied servers at instant t^-, if there is an idle server at t^-. We denote the c *discrete states* by $\{(0,0), \ldots, (0, c-1)\}$. Thus $M(t) = n$ if n *other* servers are occupied when a customer joins the system and starts service, $n = 0, \ldots, c-1$. Denote the steady-state probability of $(0, n)$ as P_n, $n = 0,\ldots, c-1$. Denote the *positive* virtual-wait states as $\{(x, c-1), x \in (0, \infty)\}$.

4.9.1 Sample Path of $\{W(t), M(t)\}_{t\geq 0}$

Consider a sample path of $\{W(t), M(t)\}_{t\geq 0}$ (Fig. 4.5). Without loss of generality, assume the system starts empty. The SP is on line 0 at $t = 0$. As the system evolves, the SP moves among the lines until $c-1$ of the servers are occupied, just as in a standard $M_\lambda/M_{\mu-\frac{\lambda}{c}}/c$ model. In Fig. 4.5 we situate line $c-1$ at level 0 of the page; this layout makes it easier to depict SP exchanges between line $c-1$ and the virtual-wait positive states.

Suppose a customer arrives when $c-1$ servers are occupied. The arrival joins the system and starts service in the one free server. All c servers are busy just after the arrival starts service. The configuration is $c-1$, since $c-1$ *other* servers are occupied just after the arrival instant. Each of the c servers has service time $\underset{dis}{=} \text{Exp}_{\mu-\frac{\lambda}{c}}$ once the arrival starts service, due to the memoryless property of exponential service times. The SP jumps to ordinate $y \in (0, \infty)$ on the page, where $y \underset{dis}{=} \text{Exp}_{c\mu-\lambda}$ which is distributed as the minimum of c i.i.d. exponential r.v.s each distributed with rate $\mu - \frac{\lambda}{c}$.

The SP descends at rate 1 (slope $= -1$), until it makes a continuous hit of level 0 from above. New arrivals are blocked and cleared, and have no effect on the sample path during this descent. Once the SP hits level 0, it continues its motion among the states $(0,0), \ldots, (0, c-1)$, until it makes another jump out of state $(0, c-1)$ onto the page.

All upward jumps that end on the page start at level 0. Hence the jump structure for M/M/c/c is much simpler than that of the standard M/M/c queue, in which jumps that end on the page may start at any point in $[0, \infty)$.

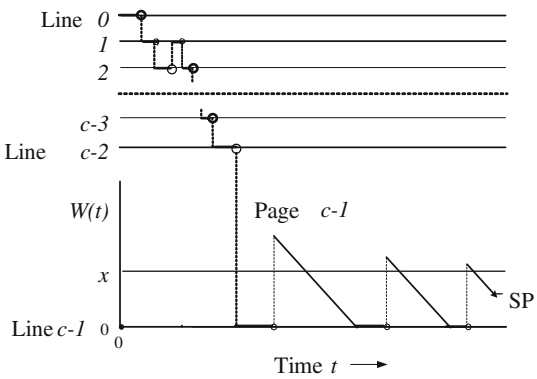

Fig. 4.5 Sample path of $\{W(t)\}_{t\geq 0}$ in $M/M/c/c$ queue. All jumps ending in $(0, \infty)$ begin at level 0 (state $(0, c-1)$) with size $\underset{dis}{=} \text{Exp}_{c\mu-\lambda}$

4.9.2 PDF of Virtual Wait

Denote the pdf of the virtual wait as $f_{c-1}(x) \equiv f(x)$, $x > 0$. To derive the pdf of $f(x)$, fix level $x > 0$. The SP downcrossing rate of x is $f(x)$. Since all SP jumps ending on the page start from state $(0, c-1)$ at arrival instants, and all jumps sizes are $\underset{dis}{=} \text{Exp}_{c\mu-\lambda}$, the upcrossing rate of x is $\lambda P_{c-1} e^{-(c\mu-\lambda)x}$. Balancing SP rates out of and into set $((x, \infty), c-1)$ yields

$$f(x) = \lambda P_{c-1} e^{-(c\mu-\lambda)x}, x > 0. \tag{4.74}$$

Remark 4.15 Formula (4.74) has precisely the same **form** as the steady-state pdf of wait in the standard M_λ /M_μ/c queue given by (4.56), except that P_{c-1} has a **different value**. For the $\text{M}_\lambda/\text{M}_{\mu-\lambda/c}$/c/c queue, formula (4.74) is derived "instantly" from observing a sample path of the virtual wait. There is no need to solve an integral equation, as in M/M/c. In M/M/c/c, the pdf formula for $f(x)$ is inherently a model equation. This is the main relationship between the two models we discuss here. The result for $\text{M}_\lambda/\text{M}_{\mu-\lambda/c}$/c/c allows us to write the **form** of the pdf of wait in $\text{M}_\lambda/\text{M}_\mu$ /c immediately.

4.9.3 Non-blocking States

The rate-balance equations for the non-blocking states $(0, 0), \ldots, (0, c-1)$ are the same as in (4.52) for $\text{M}_\lambda/\text{M}_\mu$/c, with $\mu - \frac{\lambda}{c}$ substituted for μ. Thus in M/M/c/c

$$P_n = \left(\frac{\lambda}{\mu - \frac{\lambda}{c}}\right)^n \frac{1}{n!} P_0, n = 0, \ldots, c-1,$$

so that

$$P_{c-1} = \left(\frac{\lambda}{\mu - \frac{\lambda}{c}}\right)^{c-1} \frac{1}{(c-1)!} P_0.$$

The normalizing condition is

$$\left(\sum_{n=0}^{c-1} \left(\frac{\lambda}{\mu - \frac{\lambda}{c}}\right)^n \frac{1}{n!}\right) P_0 + \int_{x=0}^{\infty} f(x)dx = 1.$$

Applying (4.74) gives

$$\left(\sum_{n=0}^{c-1} \left(\frac{\lambda}{\mu - \frac{\lambda}{c}}\right)^n \frac{1}{n!}\right) P_0 + \lambda \left(\frac{\lambda}{\mu - \frac{\lambda}{c}}\right)^{c-1} \frac{1}{(c-1)!} P_0 \int_{x=0}^{\infty} e^{-(c\mu-\lambda)x} dx = 1,$$

$$P_0 = \frac{1}{\sum_{n=0}^{c-1} \left(\frac{\lambda}{\mu - \frac{\lambda}{c}}\right)^n \frac{1}{n!} + \lambda \left(\frac{\lambda}{\mu - \frac{\lambda}{c}}\right)^{c-1} \frac{1}{(c-1)!} \frac{1}{c\mu - \lambda}}$$

$$= \frac{1}{\sum_{n=0}^{c} \left(\frac{\lambda}{\mu - \frac{\lambda}{c}}\right)^n \frac{1}{n!}}. \checkmark$$

4.9.4 Blocking Time T_B

Let T_B denote the time from the instant the system gets blocked (all c servers occupied) until the first instant that it becomes unblocked thereafter (at which $c-1$ servers are occupied). We call T_B the blocking time.

The pdf of the virtual wait in $M_\lambda/M_{\mu-\lambda/c}/c/c$ is the same as the pdf of $\mathcal{S}$ (inter start-of-service depart time) when an arrival "sees" state $(0, c-1)$. Also, $\mathcal{S} \underset{dis}{=} T_B$.

Then $E(T_B) = \text{Exp}_{c\mu-\lambda} = 1/(c\mu - \lambda)$. Let P_c denote the *proportion* of time the system is blocked. Then

$$P_c = \int_{x=0}^{\infty} f(x)dx = \lambda P_{c-1} \int_{x=0}^{\infty} e^{-(c\mu-\lambda)x} dx$$

$$
\begin{aligned}
&= \lambda \left(\frac{\lambda}{\mu - \frac{\lambda}{c}} \right)^{c-1} \frac{1}{(c-1)!} P_0 \int_{x=0}^{\infty} e^{-(c\mu - \lambda)x} dx \\
&= \left(\frac{\lambda}{\mu - \frac{\lambda}{c}} \right)^{c} \frac{1}{c!} P_0 \\
&= \frac{\left(\frac{\lambda}{\mu - \frac{\lambda}{c}} \right)^{c} \frac{1}{c!}}{\sum_{n=0}^{c} \left(\frac{\lambda}{\mu - \frac{\lambda}{c}} \right)^{n} \frac{1}{n!}}.
\end{aligned}
$$

P_c is the probability that a right-truncated Poisson variate (truncated at c), has value c. It is the classical *Erlang-B loss formula* for the $M_\lambda/M_{\mu - \frac{\lambda}{\mu}}$/c/c queue (see, e.g., p. 82, Sect. 2.6 in [84]).

Note that the blocking time is a $[c-1, c]$ busy period, denoted by $\mathcal{B}_{c-1,c}$, so that $T_B \underset{dis}{=} \mathcal{B}_{c-1,c}$. From Remark 4.8.4, $E\left(\mathcal{B}_{c-1,c}\right) = \frac{1}{c\mu - \lambda}$.

Remark 4.16 Suppose that in the M/M/c/c model the servers were numbered $1, \ldots, c$. Let the service rates assigned to arrivals depend on which server is occupied, say rates $\nu_i, i = 1, \ldots, c$. Assume $\sum_{i=1}^{c} \nu_i = c\mu - \lambda > 0$, where μ, λ are the parameters of a stable M_λ/M_μ/c queue. Then the distribution of T_B would be the same as in (4.74). So this specialized M/M/c/c model can also be used as a "companion" model to obtain the pdf of wait in the M_λ/M_μ/c queue.

4.9.5 Discussion

We can derive formula (4.74) for $f(x)$ geometrically as in Sect. 4.8.4. Let $F(x), x \geq 0$, be the cdf of the virtual wait. We get

$$
\begin{aligned}
\frac{d}{dx} \ln(1 - F(x) &= \frac{-1}{E(B_{c-1,c})} = -(c\mu - \lambda), \\
F(x) &= 1 - (1 - F(0))e^{-(c\mu - \lambda)x}, x \geq 0, \\
f(x) &= (c\mu - \lambda)(1 - F(0))e^{-(c\mu - \lambda)x}.
\end{aligned} \tag{4.75}
$$

Comparing (4.74) and (4.75) shows that

$$
\lambda P_{c-1} = (c\mu - \lambda)(1 - F(0)) = (c\mu - \lambda) P_c, \tag{4.76}
$$

where P_c is the probability of c units in the system.

In M/M/c/c an arrival enters the system iff the virtual wait is 0. Thus $F(0) = P(\text{an arrival enters the system})$. Hence $(1 - F(0)) = P(\text{an arrival is blocked and cleared}) = P_c$. Equation (4.76) is precisely the balance equation that would appear in a birth-death analysis of the system.

4.10 M/M/c in Which Zero-Wait Customers Get Special Service

Consider an M/M/c ($c \geq 2$) queue with arrival rate λ, in which zero-wait customers get service rate μ_0, and positive-wait customers get service rate μ_1 ($\neq \mu_0$). Thus, the assigned service rate is *state-dependent*. We derive below the steady-state pdf of wait, distribution of the number-in-system, and related model characteristics.

Denote the state of the system as $\{W(t), \boldsymbol{M}(t)\}_{t\geq 0}$, where $W(t) \geq 0$ is the virtual wait and $M(t)$ is the system configuration. Thus

$$\boldsymbol{M}(t) = (m_0, m_1), 0 \leq m_0 + m_1 \leq c - 1,$$

where m_j is the number of occupied servers operating at rate μ_j, $j = 0, 1$. In the notation of Sect. 4.4, integer $J = 1$. The number of zero-wait states is the total number of non-negative integer solutions for m_0, m_1 in the c equations

$$m_0 + m_1 = k, k = 0, \ldots, c - 1,$$

which is, since $J = 1$,

$$\sum_{k=0}^{c-1} \binom{J+k}{J} = \binom{J+c}{J+1} = \binom{c+1}{2}$$
$$= \frac{c(c+1)}{2} = 1 + 2 + \cdots + c.$$

From Sect. 4.4.2, $\boldsymbol{M}_0 = \left\{(0, \boldsymbol{m}) | 0 \leq \sum_{j=0}^{J} m_j \leq c - 2\right\}$, which contains $\frac{(c-1)c}{2}$ configurations. Set $\boldsymbol{M}_b = \left\{\boldsymbol{m} | \sum_{j=0}^{J} m_j = c - 1\right\}$ comprises the discrete *boundary* states, and contains $\binom{J+c-1}{J} = \binom{c}{1} = c$ configurations. (Note that $\boldsymbol{M}_b = \boldsymbol{M}_1$.)

Zero-wait Probabilities Let $P_{m_0 m_1}$ denote the steady-state probability that an arrival "sees" m_j rate-μ_j customers in service, $j = 0, 1$, and waits

zero before starting service. $P_{m_0 m_1}$ is the steady-state probability of state $(0, (m_0, m_1))$.

There are c positive-wait *pages (sheets)*, one for each configuration in $\boldsymbol{M}_b$, where

$$\boldsymbol{M}_b = \{(c-1, 0), (c-2, 1), \ldots, (1, c-2), (0, c-1)\}.$$

Positive-wait PDFs Let $f_{\boldsymbol{m}}(x)$, $x > 0$, denote the steady-state pdf of the virtual wait when the occupancies of the *other* $c-1$ servers will be $\boldsymbol{m} \in \boldsymbol{M}_b$ at start of service ('look-ahead' property of virtual wait).

4.10.1 Equations for Probabilities of Zero-Wait States

The $\frac{c(c+1)}{2}$ zero-wait states, having configurations in $\boldsymbol{M}_0 \cup \boldsymbol{M}_b$, viz.,

$$(0, (m_0, m_1)), 0 \le m_0 + m_1 \le c-1,$$

give rise to $\frac{c(c+1)}{2}$ linear equations for their probabilities, using the principle *rate out = rate in*, as in (4.77)–(4.79) below.

First consider states $(0, \boldsymbol{m})$, $\boldsymbol{m} \in \boldsymbol{M}_0$. For $m_0 = m_1 = 0$, (empty system) there is one equation:

$$\lambda P_{00} = \mu_0 P_{10} + \mu_1 P_{01}. \tag{4.77}$$

For states $(0, (m_0, m_1))$, $1 \le m_0 + m_1 \le c-2$, there are $\frac{(c-1)c}{2} - 1$ equations, each of the form

$$\begin{aligned}(\lambda + m_0\mu_0 + m_1\mu_1) P_{m_0 m_1} = \; & \lambda P_{(m_0-1)m_1} \\ & + (m_0+1)\mu_0 P_{(m_0+1)m_1} \\ & + (m_1+1)\mu_1 P_{m_0(m_1+1)}.\end{aligned} \tag{4.78}$$

For states $(0, (m_0, m_1)) \in \boldsymbol{M}_b$, there are c equations, each of the form

$$(\lambda + m_0\mu_0 + m_1\mu_1) P_{m_0 m_1} = \lambda P_{(m_0-1)m_1} + f_{m_0 m_1}(0). \tag{4.79}$$

In (4.79) the term $f_{m_0 m_1}(0)$ $(= f_{m_0 m_1}(0^+))$ is the rate at which the SP enters border state $(0, (m_0, m_1))$ due to left continuous hits of level 0 from above on page $m_0 m_1$.

4.10.2 Equations for PDF of Positive-Wait States

There are c Volterra integral equations for the positive-wait states. Consider composite state $((x, \infty), \boldsymbol{m})$, $x > 0$, on page $\boldsymbol{m} \in \boldsymbol{M}_b$. For positive-wait states $(y, m_0 m_1)$, $y > 0$, $m_0 + m_1 = c - 1$. We first specify the SP exit and entrance rates of the pertinent composite states in the state space. Then we will write the equations.

Rate Out of $((x, \infty), m_0 m_1)$

Because (m_0, m_1) is a configuration, $m_0 + m_1 = c - 1$. The SP rate out of $((x, \infty), m_0 m_1)$ is

$$f_{m_0 m_1}(x) + \lambda \frac{m_0 \mu_0}{m_0 \mu_0 + (m_1 + 1)\mu_1} \int_{y=x}^{\infty} f_{m_0 m_1}(y) dy. \tag{4.80}$$

Explanation of Terms in (4.80)

The first term $f_{m_0 m_1}(x)$ is the SP downcrossing rate of level x on page $m_0 m_1$. The second term

$$\lambda \frac{m_0 \mu_0}{m_0 \mu_0 + (m_1 + 1)\mu_1} \int_{y=x}^{\infty} f_{m_0 m_1}(y) dy$$

is the rate of arrivals when the state is $(y, m_0 m_1)$, $y > x$ (being assigned service rate μ_1 thereby adding one rate-μ_1 occupied server *upon start of service*); and a rate-μ_0 service completes first thereafter. At the arrival instant the SP jumps to level $y + \text{Exp}_{m_0\mu_0 + (m_1+1)\mu_1}$ on page $(m_0 - 1, m_1 + 1)$ (i.e., page $(m_0 - 1, c - m_0)$). If $m_0 = 0$, the SP would be on page $(0, c - 1)$. The only exit from the page would be via a downcrossing of level 0. All arrivals would be assigned service rate μ_1 and cause the SP to jump upward but remain on page $(0, c - 1)$; the second term in (4.80) would equal 0 if $m_0 = 0$.

Rate into $((x, \infty), m_0 m_1)$

The SP rate into $((x, \infty), m_0 m_1)$ is

$$\begin{aligned}
&\lambda \frac{(m_0 + 1)\mu_0}{(m_0 + 1)\mu_0 + m_1 \mu_1} e^{-((m_0+1)\mu_0 + m_1\mu_1)x} P_{m_0 m_1} \\
&+ \lambda \frac{(m_1 + 1)\mu_1}{m_0 \mu_0 + (m_1 + 1)\mu_1} e^{-(m_0\mu_0 + (m_1+1)\mu_1)x} P_{m_0 - 1, m_1 + 1} \\
&+ \lambda \frac{(m_0 + 1)\mu_0}{(m_0 + 1)\mu_0 + m_1 \mu_1} \int_{y=x}^{\infty} f_{m_0+1, m_1-1}(y) dy \\
&+ \lambda \frac{(m_0 + 1)\mu_0}{(m_0 + 1)\mu_0 + m_1 \mu_1} \int_{y=0}^{x} e^{-((m_0+1)\mu_0 + m_1\mu_1)(x-y)} f_{m_0+1, m_1-1}(y) dy
\end{aligned}$$

$$+\lambda\frac{(m_1+1)\mu_1}{m_0\mu_0+(m_1+1)\mu_1}\int_{y=0}^{x} e^{-(m_0\mu_0+(m_1+1)\mu_1)(x-y)} f_{m_0m_1}(y)dy. \tag{4.81}$$

where we have inserted a comma in subscripts like $m_0 - 1, m_1 + 1$, for clarity.

Explanation of Terms in (4.81)

The term

$$\lambda\frac{(m_0+1)\mu_0}{(m_0+1)\mu_0+m_1\mu_1}e^{-((m_0+1)\mu_0+m_1\mu_1)x}P_{m_0m_1}$$

is the rate at which the SP jumps at arrival instants from level 0 on page m_0m_1 into $((x,\infty), m_0m_1)$. At arrival instants customers are assigned service rate μ_0 (wait $= 0$), resulting in $(m_0 + 1)$ rate-μ_0 and m_1 rate-μ_1 customers in service. A rate-μ_0 service finishes first with probability

$$\frac{(m_0+1)\mu_0}{(m_0+1)\mu_0+m_1\mu_1},$$

in which case the SP jumps to page m_0m_1. SP jumps from level 0 over level x have probability $e^{-((m_0+1)\mu_0+m_1\mu_1)x}$ since $\mathcal{S} \underset{dis}{=} \text{Exp}_{(m_0+1)\mu_0+m_1\mu_1}$.

The term

$$\lambda\frac{(m_1+1)\mu_1}{m_0\mu_0+(m_1+1)\mu_1}e^{-(m_0\mu_0+(m_1+1)\mu_1)x}P_{m_0-1,m_1+1}$$

is the rate at which the SP jumps at arrival instants, from level 0 on page $(m_0 - 1, m_1 + 1)$ into $((x,\infty), m_0m_1)$. The arriving customer is assigned service rate μ_0 (wait $= 0$), resulting in m_0 rate-μ_0 and $(m_1 + 1)$ rate-μ_1 customers in service. If a rate-μ_1 service finishes first thereafter, the SP jumps to page m_0m_1; the probability is

$$\frac{(m_1+1)\mu_1}{m_0\mu_0+(m_1+1)\mu_1}.$$

SP jumps from level 0 upcross level x with probability $e^{-(m_0\mu_0+(m_1+1)\mu_1)x}$ since the inter-start-of-service depart time $\mathcal{S} \underset{dis}{=} \text{Exp}_{m_0\mu_0+(m_1+1)\mu_1}$.

The term

$$\lambda\frac{(m_0+1)\mu_0}{(m_0+1)\mu_0+m_1\mu_1}\int_{y=x}^{\infty} f_{m_0+1,m_1-1}(y)dy$$

is the rate at which the SP jumps at arrival instants, out of (x, ∞) on page $(m_0 + 1, m_1 - 1)$ into $((x, \infty), m_0 m_1)$. The arriving customer is assigned service rate μ_1 (wait > 0) resulting in $(m_0 + 1)$ rate-μ_0 and m_1 rate-μ_1 customers in service just after the start of service of the arrival. If a rate-μ_0 service finishes first, the SP jumps to page $m_0 m_1$; this has probability

$$\frac{(m_0 + 1)\mu_0}{(m_0 + 1)\mu_0 + m_1\mu_1}.$$

A jump $\mathcal{S}$ of any size will cause such a jump to enter $((x, \infty), m_0 m_1)$ since the start of the jump is already above level x.

The term

$$\lambda \frac{(m_0 + 1)\mu_0}{(m_0 + 1)\mu_0 + m_1\mu_1} \int_{y=0}^{x} e^{-((m_0+1)\mu_0 + m_1\mu()x-y)} f_{m_0+1, m_1-1}(y)dy$$

is the rate at which the SP jumps upward at arrivals, out of

$$((0, x),(m_0 + 1, m_1 - 1)) \text{ into } ((x, \infty), m_0 m_1).$$

That is, the SP makes a $(m_0 + 1, m_1 - 1) \to (m_0 m_1)$ upcrossing of level x. An arrival is assigned service rate μ_1 (wait > 0). Just after the arrival starts service there are $m_0 + 1$ rate-μ_0 and m_1 rate-μ_1 customers in service. The probability that a rate-μ_0 service finishes first is

$$\frac{(m_0 + 1)\mu_0}{(m_0 + 1)\mu_0 + m_1\mu_1},$$

causing the SP to jump to page $m_0 m_1$. Starting at level $y < x$ the SP will upcross level x if $\mathcal{S} > x - y$; since $\mathcal{S} \underset{dis}{=} \text{Exp}_{(m_0+1)\mu_0 + m_1\mu_1}$, this event has probability $e^{-((m_0+1)\mu_0 + m_1\mu_1)(x-y)}$.

The term

$$\lambda \frac{(m_1 + 1)\mu_1}{m_0\mu_0 + (m_1 + 1)\mu_1} \int_{y=0}^{x} e^{-(m_0\mu_0 + (m_1+1)\mu_1)(x-y)} f_{m_0 m_1}(y)dy$$

is the rate at which the SP jumps at arrival instants from $((0, x), m_0 m_1)$ upward into $((x, \infty), m_0 m_1)$, i.e., it upcrosses level x on page $m_0 m_1$. Arrivals are assigned service rate μ_1 (wait > 0). Just after the arrival starts service there are m_0 rate-μ_0 and $(m_1 + 1)$ rate-μ_1 customers in service. A rate-μ_1 service ends first with probability

$$\frac{(m_1+1)\mu_1}{m_0\mu_0+(m_1+1)\mu_1}.$$

causing the SP to jump to page m_0m_1. If the SP starts at level y it will upcross level x provided $\mathcal{S} > x - y$; since $\mathcal{S} \underset{dis}{=} \text{Exp}_{m_0\mu_0+(m_1+1)\mu_1}$, this event has probability $e^{-(m_0\mu_0+(m_1+1)\mu_1)(x-y)}$.

Writing Equations for Positive-Wait States

The model equation for the positive-wait states on page m_0m_1 is written by using the principle of rate balance with respect to set $((x,\infty), m_0m_1)$, *exit rate* = *entrance rate*. Equating exit rate (4.80) and entrance rate (4.81) gives

$$\begin{aligned}
&f_{m_0m_1}(x) + \lambda\frac{m_0\mu_0}{m_0\mu_0+(m_1+1)\mu_1}\int_{y=x}^{\infty} f_{m_0m_1}(y)dy \\
&= \lambda\frac{(m_0+1)\mu_0}{(m_0+1)\mu_0+m_1\mu_1}e^{-((m_0+1)\mu_0+m_1\mu_1)x}P_{m_0m_1} \\
&\quad + \lambda\frac{(m_1+1)\mu_1}{m_0\mu_0+(m_1+1)\mu_1}e^{-(m_0\mu_0+(m_1+1)\mu_1)x}P_{m_0-1,m_1+1} \\
&\quad + \lambda\frac{(m_0+1)\mu_0}{(m_0+1)\mu_0+m_1\mu_1}\int_{y=x}^{\infty} f_{m_0+1,m_1-1}(y)dy \\
&\quad + \lambda\frac{(m_0+1)\mu_0}{(m_0+1)\mu_0+m_1\mu_1}\int_{y=0}^{x} e^{-((m_0+1)\mu_0+m_1\mu_1)(x-y)} f_{m_0+1,m_1-1}(y)dy \\
&\quad + \lambda\frac{(m_1+1)\mu_1}{m_0\mu_0+(m_1+1)\mu_1}\int_{y=0}^{x} e^{-(m_0\mu_0+(m_1+1)\mu_1)(x-y)} f_{m_0m_1}(y)dy.
\end{aligned} \tag{4.82}$$

Equation for "Cover"

The total probability of a zero wait is

$$P_0 = \sum_{\boldsymbol{m}\in\boldsymbol{M}_0\cup\boldsymbol{M}_b} P_{\boldsymbol{m}} = \sum_{0\le m_0+m_1\le c-1} P_{m_0m_1}. \tag{4.83}$$

The total pdf of wait is

$$f(x) = \sum_{\boldsymbol{m}\in\boldsymbol{M}_1} f_{\boldsymbol{m}}(x) = \sum_{m_0+m_1=c-1} f_{m_0m_1}(x), x > 0. \tag{4.84}$$

Let $x > 0$ be fixed. The total SP downcrossing rate of x is $f(x)$. The total SP upcrossing rate of x due to jumps starting from level 0 at arrival instants, is

$$\lambda \sum_{m_0+m_1=c-1} e^{-((m_0+1)\mu_0+m_1\mu_1)x} P_{m_0m_1}.$$

The total SP upcrossing rate of x due to jumps starting from levels $y \in (0, x)$ at arrival instants, is

$$\lambda \sum_{m_0+m_1=c-1} \int_{y=0}^{x} e^{-(m_0\mu_0+(m_1+1)\mu_1)(x-y)} f_{m_0m_1}(y)dy.$$

Rate balance across level x gives the model equation for the cover,

$$\begin{aligned} f(x) = \lambda & \sum_{m_0+m_1=c-1} e^{-((m_0+1)\mu_0+m_1\mu_1)x} P_{m_0m_1} \\ & + \lambda \sum_{m_0+m_1=c-1} \int_{y=0}^{x} e^{-(m_0\mu_0+(m_1+1)\mu_1)(x-y)} f_{m_0m_1}(y)dy. \end{aligned} \tag{4.85}$$

Normalizing Condition
The normalizing condition $P_0 + \int_{x=0}^{\infty} f(x)dx = 1$ can be expressed as

$$\sum_{0\le m_0+m_1\le c-1} P_{m_0m_1} + \sum_{m_0+m_1=c-1} \int_{x=0}^{\infty} f_{m_0m_1}(x)dx = 1. \tag{4.86}$$

4.10.3 Solution of Model Equations

In Sect. 4.11 below, we formulate and solve the foregoing M/M/2 model with zero-wait customers receiving exceptional service, whose solution illustrates relevant SPLC ideas and related insights. A more general solution procedure of a two-server M/M/2 queue where service time depends on waiting time in *a general manner* is detailed in Chap. 4 of [11].

4.11 M/M/2: Zero-Waits Get Special Service

M/M/2/(μ_0, μ_1) , $(0, (0, \infty))$

To fix ideas and clarify the system dynamics of M/M/c with special service for zero-wait customers, we formulate the model with $c = 2$ servers. We discuss

the solution for the zero-wait probabilities and the positive-wait pdfs. We denote the model by M/M/2/(μ_0, μ_1) , $(0, (0, \infty))$. This notation indicates that 0-wait arrivals get service rate μ_0 and $(0, \infty)$-wait arrivals get service rate μ_1; diagrammatically, $\mu_0 \leftrightarrow$ 0-wait, $\mu_1 \leftrightarrow (0, \infty)$-wait.

There are only three zero-wait states in $\boldsymbol{M}_0 \cup \boldsymbol{M}_b$ (compare Sect. 4.10),

$$\{(0, m_0 m_1)\} = \{(0, 00), (0, 10), (0, 01)\}.$$

Denote the steady-state probabilities of the zero-wait states by P_{00}, P_{10}, P_{01} respectively.

For example, state (0, 10) indicates that an arrival would wait 0 and would "see" a rate-μ_0 customer being served by the *other* server. The arrival would be assigned rate μ_0 since it waits 0. There would then be two rate-μ_0 customers in service. The inter start-of-service depart time $\mathcal{S}$ would be $\underset{dis}{=} \text{Exp}_{2\mu_0}$.

There are only two zero-wait states such that $m_0 + m_1 = 1$ (both border states). Denote the pdfs of the positive-wait states $(x, 10)$, $(x, 01)$, by $f_{10}(x)$, $f_{01}(x)$, $x > 0$, respectively. A would-be arrival that finds the state $(x, 10)$, $x > 0$, for example, would wait x before service, and be assigned service rate μ_1 (wait > 0). Just after its start of service, it would have a rate-μ_0 customer as neighbor in the other server. Inter start-of-service depart time $\mathcal{S} \underset{dis}{=} \text{Exp}_{\mu_0+\mu_1}$. The rate-$\mu_1$ customer would finish service first with probability $\frac{\mu_1}{\mu_0+\mu_1}$, leaving the rate-$\mu_0$ customer in service ($m_0 m_1 = 10$). The rate-μ_0 customer would finish service first with probability $\frac{\mu_0}{\mu_0+\mu_1}$, leaving the rate-$\mu_1$ customer in service ($m_0 m_1 = 01$).

If an arrival "sees" state $(x, 01)$, $x > 0$, $\mathcal{S}$ would be $\underset{dis}{=} \text{Exp}_{2\mu_1}$. The first customer to complete service would have rate μ_1 with certainty. The customer remaining in service just after that service completion would have service rate μ_1 ($m_0 m_1 = 01$).

A sample-path diagram of the virtual wait process $\{W(t)\}_{t \geq 0}$, has three lines and two pages (Fig. 4.6).

The total (marginal) probability of a zero wait is

$$P_0 = P_{00} + P_{10} + P_{01}.$$

The total pdf of wait is

$$f(x) = f_{10}(x) + f_{01}(x), x > 0.$$

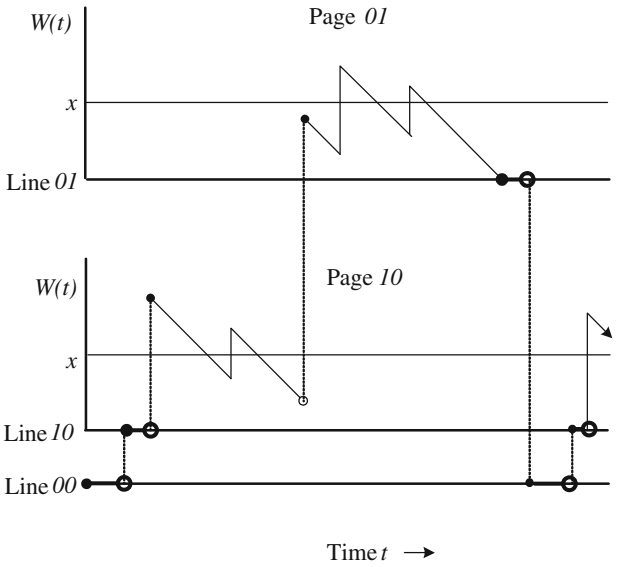

Fig. 4.6 Sample path of virtual wait in M/M/2/(μ_0, μ_1) , $(0, (0, \infty))$. Lines for states $(0, 10)$, $(0, 01)$ are at level 0 of corresponding pages. Line for state $(0, 00)$ is isolated. The SP can enter state $(0, 01)$ only by downcrossing level 0 on page 01 (See Fig. 4.10.)

4.11.1 Model Equations

Zero-Wait States

Applying *SP exit rate = SP entrance rate* for the zero-wait states $(0, 00)$, $(0, 10)$, $(0, 01)$ gives, respectively,

$$\begin{aligned} \lambda P_{00} &= \mu_0 P_{10} + \mu_1 P_{01}, \\ (\lambda + \mu_0) P_{10} &= \lambda P_{00} + f_{10}(0), \\ (\lambda + \mu_1) P_{01} &= f_{01}(0). \end{aligned} \tag{4.87}$$

In (4.87), the terms $f_{10}(0)$, $f_{01}(0)$ (same as $f_{10}(0^+)$, $f_{01}(0^+)$) are the rates at which the SP hits level 0 from above on pages 10 and 01 respectively. Immediately following such hits, the SP moves on lines 10 and 01 respectively.

Positive-Wait States

Applying *SP exit rate = SP entrance rate* for $((x, \infty), 10)$ (on page 10) yields the integral equation

$$f_{10}(x) + \lambda \int_{y=x}^{\infty} \frac{\mu_0}{\mu_0 + \mu_1} f_{10}(y)dy$$
$$= \lambda P_{10} e^{-2\mu_0 x} + \lambda \frac{\mu_1}{\mu_0 + \mu_1} P_{01} e^{-(\mu_0+\mu_1)x}$$
$$+ \lambda \frac{\mu_1}{\mu_0 + \mu_1} \int_{y=0}^{x} e^{-(\mu_0+\mu_1)(x-y)} f_{10}(y)dy, \; x > 0. \tag{4.88}$$

When formulating equation (4.88), note that the SP cannot jump directly from a positive-wait state on page 01 into set $((x, \infty), 10)$. An arrival that "sees" state $(y, 01)$, $y > 0$, will be assigned rate μ_1 and start service after a wait y; its neighbor in the other server will also have service rate μ_1 (because $m_0 m_1 = 01$). The random variable $\mathcal{S}$ will be distributed as $\text{Exp}_{2\mu_1}$, and the remaining customer in service just after the first departure thereafter, will have rate μ_1. At the arrival instant, the SP will start a jump at level y on page 01, which ends at level $y + \mathcal{S} \underset{dis}{=} y + \text{Exp}_{2\mu_1}$, also on page 01. The configuration remains 01 just after the arrival. The only exit route from page 01 is via a downcrossing of level 0 (continuous hit of 0 from above—see Fig. 4.6).

Now we balance the SP exit and entrance rates for $((x, \infty), 01)$ (page 01), giving integral equation

$$f_{01}(x) = \lambda \frac{\mu_0}{\mu_0 + \mu_1} P_{01} e^{-(\mu_0+\mu_1)x}$$
$$+ \lambda \int_{y=0}^{x} e^{-2\mu_1(x-y)} f_{01}(y)dy$$
$$+ \lambda \frac{\mu_0}{\mu_0 + \mu_1} \int_{y=0}^{x} e^{-(\mu_0+\mu_1)(x-y)} f_{10}(y)dy$$
$$+ \lambda \frac{\mu_0}{\mu_0 + \mu_1} \int_{y=x}^{\infty} f_{10}(y)dy. \tag{4.89}$$

When formulating (4.89), note that the SP can exit $((x, \infty), 01)$ only by downcrossing level x. Also, the SP cannot enter $((x, \infty), 01)$ from state $(0, 10)$ at arrivals, since all jumps that start from line 10 (corresponding to state $(0, 10)$) must end on page 10, at an ordinate $\underset{dis}{=} \text{Exp}_{2\mu_0}$.

The equation for the total pdf is

$$f(x) = f_{10}(x) + f_{01}(x),$$

as viewed from the "cover", the result of projecting sample-path segments on pages 10 and 01 onto a single sheet. An integral equation for $f(x)$ is obtained by balancing the SP *total* down- and upcrossing rates of level $x > 0$. This is equivalent to equating the exit and entrance rates for the state-space set

$$((x, \infty), 10) \cup ((x, \infty), 01).$$

The resulting equation is

$$\begin{aligned} f(x) = {} & \lambda P_{10} e^{-2\mu_0 x} + \lambda P_{01} e^{-(\mu_0+\mu_1)x} \\ & + \lambda \int_{y=0}^{x} e^{-(\mu_0+\mu_1)(x-y)} f_{10}(y) dy \\ & + \lambda \int_{y=0}^{x} e^{-2\mu_1(x-y)} f_{01}(y) dy, \ x > 0. \end{aligned} \tag{4.90}$$

Equation (4.90) can also be derived by summing the corresponding sides of (4.88) and (4.89). However, it is intuitive and instructive to interpret equation (4.90) as total SP rate-balance across level $x > 0$.

The normalizing condition is

$$P_{00} + P_{10} + P_{01} + \int_{x=0}^{\infty} f_{10}(x) dx + \int_{x=0}^{\infty} f_{01}(x) dx = 1,$$

or

$$P_0 + \int_{x=0}^{\infty} f(x) dx = 1. \tag{4.91}$$

4.11.2 Solution of Equations

Equation (4.88) is an integral equation in $f_{10}(x)$, which is not confounded by the presence of $f_{01}(x)$; so we utilize it to obtain the functional form of $f_{10}(x)$. Applying differential operator $\langle D \rangle \langle D + \mu_0 + \mu_1 \rangle$ to both sides of (4.88) leads to the second order differential equation

$$\begin{aligned} & f''_{10}(x) + (\mu_0 + \mu_1 - \lambda) f'_{10}(x) - \lambda \mu_0 f_{10}(x) \\ & = 2\lambda \mu_0 (\mu_0 - \mu_1) P_{10} e^{-2\mu_0 x}, \ x > 0. \end{aligned} \tag{4.92}$$

with solution

$$f_{10}(x) = C_{10}e^{ax} + C_{10}^{1}e^{bx} + \lambda K_{10}P_{10}e^{-2\mu_0 x},\ x > 0,$$

where a and b are the roots of the auxiliary quadratic equation, namely

$$a = \frac{1}{2}\left(\lambda - \mu_0 - \mu_1 - \sqrt{\lambda^2 + 2\lambda\mu_0 - 2\lambda\mu_1 + \mu_0^2 + 2\mu_0\mu_1 + \mu_1^2}\right) < 0,$$
$$b = \frac{1}{2}\left(\lambda - \mu_0 - \mu_1 + \sqrt{\lambda^2 + 2\lambda\mu_0 - 2\lambda\mu_1 + \mu_0^2 + 2\mu_0\mu_1 + \mu_1^2}\right) > 0,$$
$$K_{10} = \frac{2(\mu_0 - \mu_1)}{\lambda + 2\mu_0 - 2\mu_1},$$

and C_{10}, C_{10}^{1} are constants of integration. A necessary condition for system stability is $\lim_{x\to\infty} f_{10}(x) = 0$, which implies $C_{10}^{1} = 0$ (since $b > 0$). Thus the functional form of $f_{10}(x)$ is

$$f_{10}(x) = C_{10}e^{ax} + \lambda K_{10}P_{10}e^{-2\mu_0 x},\ x > 0, \tag{4.93}$$

where C_{10} is a constant to be determined.

The term K_{10} will be undefined if $\lambda + 2\mu_0 - 2\mu_1 = 0$. If $\lambda + 2\mu_0 - 2\mu_1 \neq 0$ and $\mu_0 - \mu_1 \neq 0$, then K_{10} may be positive or negative. If $\mu_0 - \mu_1 = 0$ the model reduces to a standard M/M/c queue with $c = 2$ (Sect. 4.8); the computed distribution of wait should then match that of a standard M/M/2 queue. (We will utilize this property later as a mild check on the correctness of the present solution.)

We obtain the functional form of $f_{01}(x)$ by substituting the expression for $f_{10}(x)$ (4.93) into (4.90). Since

$$f_{01}(x) = f(x) - f_{10}(x),$$

this substitution gives the integral equation

$$\begin{aligned} f_{01}(x) = {} & \lambda(1 - K_{10})P_{10}e^{-2\mu_0 x} + \lambda P_{01}e^{-(\mu_0+\mu_1)x} - C_{10}e^{ax} \\ & + \lambda\int_{y=0}^{x} e^{-(\mu_0+\mu_1)(x-y)}(C_{10}e^{ay} + \lambda K_{10}P_{10}e^{-2\mu_0 y})dy \\ & + \lambda\int_{y=0}^{x} e^{-2\mu_1(x-y)} f_{01}(y)dy. \end{aligned} \tag{4.94}$$

The first integral term in (4.94) is

$$\lambda \int_{y=0}^{x} e^{-(\mu_0+\mu_1)(x-y)}(C_{10}e^{ay} + \lambda K_{10}P_{10}e^{-2\mu_0 y})dy$$
$$= \frac{\lambda C_{10}}{\mu_0+\mu_1+a}e^{ax} - \frac{\lambda^2 K_{10}P_{10}e^{-2\mu_0 x}}{\mu_0-\mu_1}$$
$$- \left(\frac{\lambda C_{10}}{\mu_0+\mu_1+a} - \frac{\lambda^2 K_{10}P_{10}}{\mu_0-\mu_1}\right)e_1^{-(\mu_0+\mu_1)x}.$$

Thus (4.94) is equivalent to the integral equation

$$\begin{aligned} f_{01}(x) =& H_{01}C_{10}e^{ax} + \lambda B_{01}P_{10}e^{-2\mu_0 x} \\ &+ D_{01}e^{-(\mu_0+\mu_1)x} \\ &+ \lambda \int_{y=0}^{x} e^{-2\mu_1(x-y)} f_{01}(y)dy, \end{aligned} \tag{4.95}$$

where

$$H_{01} = \frac{\lambda}{\mu_0+\mu_1+a} - 1,$$
$$B_{01} = 1 - K_{10} - \frac{\lambda K_{10}}{\mu_0-\mu_1},$$
$$D_{01} = \lambda P_{01} - \frac{\lambda C_{10}}{\mu_0+\mu_1+a} + \frac{\lambda^2 K_{10}P_{10}}{\mu_0-\mu_1}.$$

Applying the differential operator $\langle D + 2\mu_1 \rangle$ to both sides of (4.95) yields the differential equation for $f_{01}(x)$,

$$\begin{aligned} f'_{01}(x) + (2\mu_1 - \lambda)f_{01}(x) =& (2\mu_1 + a)H_{01}C_{10}e^{ax} \\ &+ 2\lambda(\mu_1-\mu_0)B_{01}P_{10}e^{-2\mu_0 x} \\ &+ (\mu_1-\mu_0)D_{01}e^{-(\mu_1+\mu_0)x}. \end{aligned} \tag{4.96}$$

whose solution is

$$\begin{aligned} f_{01}(x) =& \frac{2\lambda(\mu_1-\mu_0)}{2\mu_1-\lambda-2\mu_0}B_{01}P_{10}e^{-2\mu_0 x} \\ &+ \frac{\mu_1-\mu_0}{\mu_1-\lambda-\mu_0}D_{01}e^{-(\mu_1+\mu_0)x} \end{aligned}$$

$$+\frac{2\mu_1+a}{2\mu_1-\lambda+a}H_{01}C_{10}e^{ax}$$
$$+C_{01}e^{-(2\mu_1-\lambda)x}, \tag{4.97}$$

where C_{01} is a constant of integration to be determined (see Sect. 4.11.4).

4.11.3 Stability Condition

Consider the functional forms of $f_{10}(x)$ and $f_{01}(x)$ in (4.93) and (4.97). In the exponents, all the coefficients of x are negative except possibly the coefficient $-(2\mu_1-\lambda)$ in $e^{-(2\mu_1-\lambda)x}$ of (4.97). A necessary condition for stability is that

$$f_{10}(\infty)=f_{01}(\infty)=f(\infty)=0;$$

implying $-(2\mu_1-\lambda)<0$, equivalent to $\lambda<2\mu_1$. That is, the arrival rate must be less than the system departure rate when both servers are occupied by positive-wait customers, regardless how large x is. Thus, for stability, if the waiting time is large and customers are arriving, then the mean inter-arrival time should exceed the mean inter-departure time. This ensures that the waiting time will return to zero in a finite time.

4.11.4 Determination of Constants

A complete solution for the distribution of wait requires the values of five unknown constants

$$P_{00},\ P_{10},\ P_{01},\ C_{10},\ C_{01},$$

which we obtain from five independent equations.

In (4.93) letting $x\downarrow 0$ to obtain $f_{10}(0)$, and referring to (4.87) gives

$$C_{10}+\lambda K_{10}P_{10}=(\lambda+\mu_0)P_{10}-\lambda P_{00}. \tag{4.98}$$

In (4.97) letting $x\downarrow 0$ to obtain $f_{01}(0)$ gives

$$f_{01}(0)=\frac{2\lambda(\mu_1-\mu_0)}{2(\mu_1-\mu_0)-\lambda}B_{01}P_{10}$$
$$+\frac{\mu_1-\mu_0}{\mu_1-\mu_0-\lambda}D_{01}$$
$$+\frac{2\mu_1+a}{2\mu_1+a-\lambda}H_{01}C_{10}+C_{01}. \tag{4.99}$$

Substituting $f_{01}(0)$ from (4.99) into (4.87) gives

$$C_{01} = (\lambda + \mu_1)P_{01} - \frac{2\lambda(\mu_1 - \mu_0)}{2(\mu_1 - \mu_0) - \lambda} B_{01}P_{10} - \frac{\mu_1 - \mu_0}{\mu_1 - \mu_0 - \lambda} D_{01} - \frac{2\mu_1 + a}{2\mu_1 + a - \lambda} H_{01}C_{10}. \tag{4.100}$$

We get another independent equation by substituting the functional form

$$f_{10}(x) = C_{10}e^{ax} + \lambda K_{10}P_{10}e^{-2\mu_0 x}$$

into the integral equation (4.88) and equating the coefficients of corresponding exponential terms on both sides after evaluating the integral (different exponentials are linearly independent—see, e.g., Sect. 3.3, pp. 99ff and p. 205 in [10]). The coefficient of $e^{-(\mu_0+\mu_1)x}$ on the right side of (4.88) must be 0. This yields the linear equation

$$\frac{\lambda\mu_1}{\mu_0 + \mu_1} P_{01} - \frac{1}{\mu_0 + \mu_1 + a} - \frac{\lambda K_{10}}{\mu_1 - \mu_0} P_{10} = 0. \tag{4.101}$$

The normalizing condition is

$$1 = P_{00} + P_{10} + P_{01} + \frac{C_{10}}{(-a)} + \frac{\lambda K_{10}P_{10}}{2\mu_0} + \frac{\lambda(\mu_1 - \mu_0)}{\mu_0(2(\mu_1 - \mu_0) - \lambda)} B_{01}P_{10} + \frac{\mu_1 - \mu_0}{(\mu_1 + \mu_0)(\mu_1 - \mu_0 - \lambda)} D_{01} + \frac{2\mu_1 + a}{(-a)(2\mu_1 + a - \lambda)} H_{01}C_{10} + \frac{1}{2\mu_1 - \lambda} C_{01}. \tag{4.102}$$

We now have a set of five equations to solve for the five constants: from (4.87)

$$\lambda P_{00} = \mu_0 P_{10} + \mu_1 P_{01},$$

and (4.98), (4.100), (4.101), (4.102).

Remark 4.17 In the derivation of the functional forms of $f_{10}(x)$, $f_{01}(x)$ the expressions

$$\mu_1 - \mu_0, \quad 2\mu_1 - \lambda - 2\mu_0, \quad \mu_1 - \lambda - \mu_0, \quad 2\mu_1 - \lambda + a$$

appear in various denominators. If any of these four expressions were equal to 0, the functional forms would have to be modified. The five equations used

to solve for the constants in the present model would have to be modified accordingly. In this monograph we emphasize the system-point level-crossing approach to derive model equations, and various techniques to solve them. However, there are many techniques to solve systems of integral equations, requiring additional study, outside the scope of the present volume. We give **numerical** solutions of the equations in several examples below.

Remark 4.18 It would be interesting to explain the appearance of the immediately above expressions in the denominators. Does the system reduce to a particular queueing model when a denominator is equal to 0? For example, when $\mu_1 - \mu_0 = 0$, the M/M/2/(μ_0, μ_1) , $(0, (0, \infty))$ system reduces to a standard M/M/2 model. In M/M/2/(μ_0, μ_1) , $(0, (0, \infty))$ the only criterion necessary for stability is $\lambda < 2\mu_1$. What do these exceptional denominators mean with regard to physical models?

Another question is how to select a set of linearly independent equations to solve for the constants. Once a set of equations is derived, it can be checked for independence using matrix methods. But this amounts to trial and error. Is there a way to derive five independent equations directly? Taking derivatives may be the answer to this question.

Example 4.7 We first give a mild numerical check on the five equations by letting $\mu_1 - \mu_0 = 0$. In this case M/M/2/(μ_0, μ_1) , $(0, (0, \infty))$ reduces to a standard M/M/2 queue. We arbitrarily take

$$\lambda = 1,\ \mu_0 = 1.5,\ \mu_1 = 1.5.$$

Then $a = -2.581139$. The solution for the constants is

$$C_{10} = 0.0,\ P_{01} = .133333,\ P_{10} = .20,\ C_{01} = .333333,\ P_{00} = .50.$$

We compare this solution with that of the standard M/M/2 queue with $\lambda = 1$, $\mu = 1.5$. In M/M/2, the probability of an empty system is $P_0 = 0.5$. The probability of 1 customer in the system *is indeed* $P_1 = 0.33333$. The values match P_{00} and $P_{10} + P_{01}$ in M/M/2/(μ_0, μ_1) , $(0, (0, \infty))$ model, as expected.

Also, in M/M/2/(μ_0, μ_1) , $(0, (0, \infty))$, we see from (4.97) that

$$\begin{aligned} f_{01}(x) &= C_{01} e^{-(2\mu_1 - \lambda)x} \\ &= \lambda P_1 e^{-(2\mu_1 - \lambda)x} \\ &= 1 \cdot (0.33333) e^{-2x},\ x > 0, \end{aligned}$$

since $\mu_1 - \mu_0 = 0$ and $C_{10} = 0$.

Example 4.8 Let $\lambda = 1$, $\mu_0 = 1.1$, $\mu_1 = 2.21$. These values preclude that any of the four above-mentioned denominators is 0. We get $a = -2.715136$. We solve the equations and obtain

$$P_{00} = .417715,\ P_{10} = 0.339103,\ P_{01} = 0.0202270,$$
$$C_{01} = 0.022818,\ C_{10} = -0.322655.$$

The functions $f_{10}(x), x > 0$, and $f_{01}(x), x > 0$, are linear combinations of exponentials,

$$f_{10}(x) = -0.322655e^{-2.715136x} + 0.617056e^{-2.2x},$$
$$f_{01}(x) = 0.505784e^{-2.2x} + 0.067831e^{-3.31x}$$
$$- 0.531504e^{-2.715136x} + 0.022818e^{-3.42x}.$$

We substitute the values of $P_{00}, P_{10}, P_{01}, f_{10}(x), f_{01}(x)$ into the normalizer (4.91), and obtain 1; it checks.

The partial pdfs of wait $f_{10}(x)$, $f_{01}(x)$ and total pdf of wait $f(x)$ are depicted in Figs. 4.7, 4.8, and 4.9 respectively.

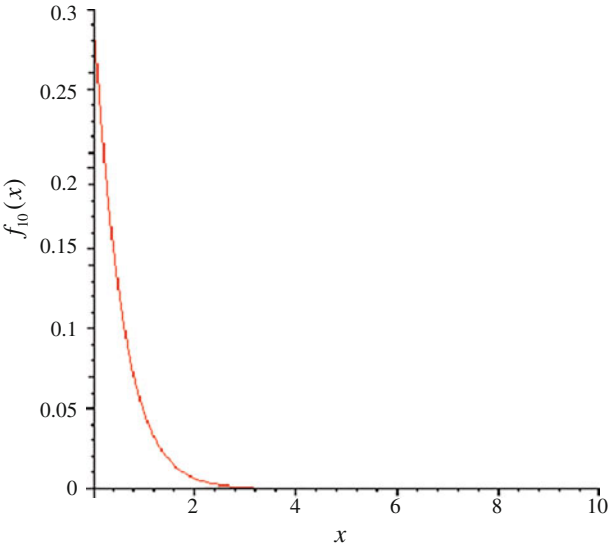

Fig. 4.7 Partial pdf of wait $f_{10}(x)$ in M/M/2/(μ_0, μ_1), $(0, (0, \infty))$. $\lambda = 1$, $\mu_0 = 1.1$, $\mu_1 = 2.21$

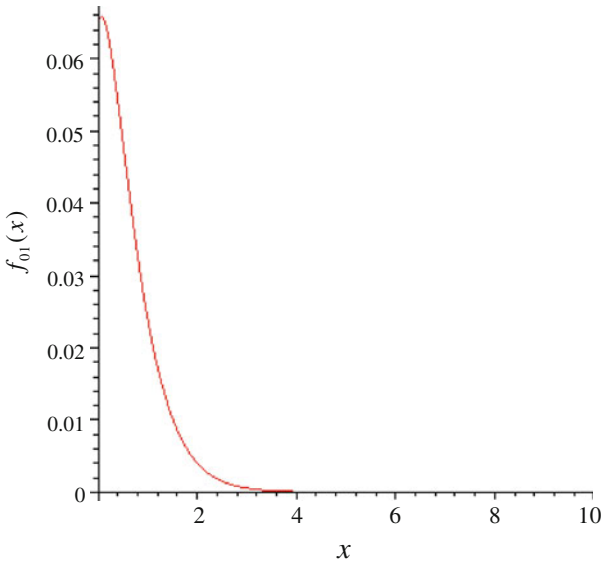

Fig. 4.8 Partial pdf of wait $f_{01}(x)$ in M/M/2/(μ_0, μ_1), $(0, (0, \infty))$. $\lambda = 1$, $\mu_0 = 1.1$, $\mu_1 = 2.21$

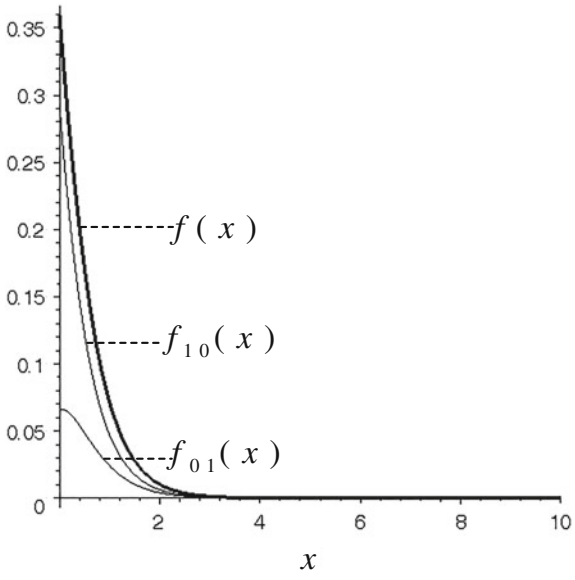

Fig. 4.9 Total pdf of wait $f(x) = f_{10}(x) + f_{01}(x)$ in M/M/2/(μ_0, μ_1), $(0, (0, \infty))$. $\lambda = 1$, $\mu_0 = 1.1$, $\mu_1 = 2.21$

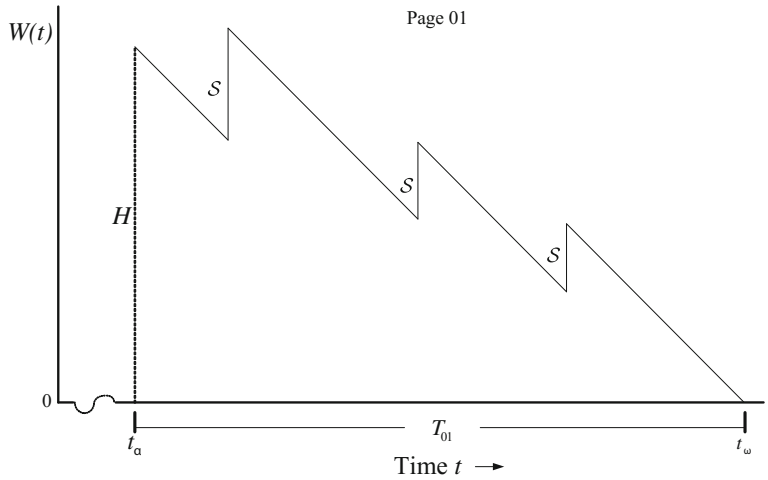

Fig. 4.10 T_{01} := sojourn on page 01. t_α := start of T_{01}, t_ω := end of T_{01}. $\mathcal{S} \underset{dis}{=} \text{Exp}_{2\mu_1}$. (See Fig. 4.6.)

4.11.5 Expected Sojourn Time on a Page

Consider page 01. The SP can enter page 01 from discrete state (0, 01) or from page 10, due to a jump at an arrival (Fig. 4.6). It cannot enter directly from state (0, 10) at an arrival instant, since zero-wait arrivals are assigned rate μ_0 resulting in both servers being occupied with rate-μ_0 customers; so any SP jump to a positive level must end on page 10.

In a sojourn on page 01, the first inter start-of-service depart time will be $\underset{dis}{=} \text{Exp}_{\mu_0+\mu_1}$; any other inter start-of-service depart times that follow while on page 01 will be $\underset{dis}{=} \text{Exp}_{2\mu_1}$. While the SP is on page 01, each departure will leave a rate-μ_1 customer in the neighboring occupied server. Given that the SP enters page 01, its source state was (0, 01) with probability (using Bayes' rule)

$$q = \frac{P_{01}}{P_{01} + \int_{y=0}^{\infty} f_{10}(y)dy}.$$

Its source was composite state ((0, ∞), 10) with probability

$$1 - q = \frac{\int_{y=0}^{\infty} f_{10}(y)dy}{P_{01} + \int_{y=0}^{\infty} f_{10}(y)dy}.$$

Let H denote the height above level 0 (ordinate) at which the SP enters page 01 (see Fig. 4.10). A sojourn on page 01 starts at level H, where

$$E(H|\text{source is } (0,01)) = \frac{1}{\mu_0+\mu_1},$$

and

$$E(H|\text{source is level } y \text{ on page } 10) = y + \frac{1}{\mu_0+\mu_1},\ y>0,$$

since the size of a jump from either source onto page 01 is $\underset{dis}{=} \text{Exp}_{\mu_0+\mu_1}$. Thus

$$E(H) = \frac{1}{\mu_0+\mu_1}\cdot q + \left(\int_{y=0}^{\infty}\left(y+\frac{1}{\mu_0+\mu_1}\right) f_{10}(y)dy\right)\cdot(1-q)\,.$$

From (4.93) $f_{10}(y)$ is given by

$$f_{10}(y) = C_{10}e^{ay} + \lambda K_{10}P_{10}e^{-2\mu_0 y},\ y>0,$$

and thus

$$\begin{aligned}
E(H) &= \frac{1}{\mu_0+\mu_1}\cdot q \\
&\quad + \left(\int_{y=0}^{\infty}\left(y+\frac{1}{\mu_0+\mu_1}\right)\left(C_{10}e^{ay}+\lambda K_{10}P_{10}e^{-2\mu_0 y}\right)dy\right)\cdot(1-q) \\
&= \frac{1}{\mu_0+\mu_1}\cdot q + \left(\frac{1}{4}\left(4C_{10}\mu_0^2\mu_1 + 4C_{10}\mu_0^3\right.\right. \\
&\quad + 3\lambda K_{10}P_{10}a^2\mu_0 + \lambda K_{10}P_{10}a^2\mu_1 \\
&\quad \left.\left. - 4C_{10}a\mu_0^2\right)/\left(a^2\mu_0^2(\mu_0+\mu_1)\right)\right)\cdot(1-q).
\end{aligned} \tag{4.103}$$

Let T_{01} denote a sojourn time on page 01, i.e., the time from SP entrance until the first exit from page 01 thereafter. The only possible exit is due to a downcrossing of level 0 (Fig. 4.6). Thus

$$T_{01} = H + \sum_{i=1}^{N_H}\mathcal{B}_i$$

where N_H is the number of arrivals during time H and $\mathcal{B}_i$ represents a busy period of an *M/M/1 queue* with service rate $2\mu_1$, since both servers are busy with rate-μ_1 customers. (See Sect. 3.4.12 and Fig. 3.6.) The expected busy period is obtained from (3.120) with $2\mu_1$ substituted for μ. Thus

$$E(\mathcal{B}_i) = \frac{1}{2\mu_1-\lambda},\ i=1,\ldots,N_H.$$

The r.v.s N_H and $\mathcal{B}_i$, $i = 1, \ldots, N_H$ are independent, since the $\mathcal{B}_i$s are i.i.d. each distributed as an $M_\lambda/M_{2\mu_1}/1$ busy period. The expected sojourn time on page 01 is

$$E(T_{01}) = E(H) + E\left(\sum_{i=1}^{N_H} \mathcal{B}_i\right) = E(H) + E(N_H)E(\mathcal{B}_i)$$
$$= E(H) + \lambda E(H)\frac{1}{2\mu_1 - \lambda} = \frac{E(H)}{1 - \lambda/(2\mu_1)}, \tag{4.104}$$

where $E(H)$ is given in formula (4.103). It is noteworthy that T_{01} is distributed as the *busy period* in an $M_\lambda/M_{2\mu_1}/1$ queue in which zero-wait arrivals obtain special service $\underset{dis}{=} H$, and positive-wait arrivals get service rate $2\mu_1$. This structure of T_{01} illustrates an interesting application, and the versatility, of the M/G/1 queue where zero-wait arrivals get exceptional service (see Sect. 3.6.1).

Example 4.9 In Example 4.8 with $\lambda = 1$, $\mu_0 = 1.1$, $\mu_1 = 2.21$, we obtain

$$q = .111216, \qquad 1 - q = .888784, \qquad E(H) = 0.151416.$$

The expected sojourn time on page 01 is $E(T_{01}) = 0.195689$.

Remark 4.19 Various questions arise regarding Example 4.9. What is the *proportion* of time that the SP spends circulating on page 01, page 10, or in the zero-wait states? Can this question be answered for a general M/M/c/(μ_0, μ_1), $(0, (0, \infty))$ queue with $c > 2$? If yes, then it would be straightforward to determine P_{00}. This would facilitate solving for all the zero-wait probabilities and the partial pdfs of wait.

4.12 M/M$_i$/c with Reneging

Consider an M/M/c queue, with $c \geq 2$ *distinguishable* servers having fixed exponential service rates μ_i, $i = 1, \ldots, c$. Thus, the queue has ***heterogeneous servers***. This model is denoted by M/M$_i$/c. Using the notation for the generalized M/M/c model (Sects. 4.3, 4.4 and 4.5), let $\{W(t), M(t)\}_{t \geq 0}$ denote the *system point process*, where $W(t) :=$ *virtual wait* at time t and $M(t) :=$ *system configuration* at time t (see Sect. 4.5). The set of possible exponential service rates is $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_c\}$. A new arrival receives one of those service rates, depending on which server it engages. We assume the

μ_is are distinct. When some or all of the μ_is are equal, the analysis is similar with slight modification.

Assume zero-wait arrivals start service immediately (no balking). In general, the zero-wait server-assignment policy is arbitrary. When formulating equations for the zero-wait probabilities in a specific model, however, we must specify a zero-wait server-assignment policy (see Sect. 4.12.7 below).

4.12.1 Staying Function

Let $\{\tau_n\}_{n=1,2,\ldots}$ be the arrival times of customers $C_n, n = 1, 2, \ldots$, respectively. Then $W(\tau_n^-) \equiv W_n :=$ *required wait* before start of service of C_n.

Define

$$\theta_n = \begin{cases} 1 \text{ if } C_n \text{ stays for a full service} \\ 0 \text{ if } C_n \text{ reneges while waiting for service} \end{cases}, n = 1, 2, \ldots.$$

With respect to the steady-state statistical properties of the waiting time, this model is equivalent to one in which customers balk from joining the system at arrival instants, depending on their required wait before service, i.e., on their arrival-point W_ns. (See a sample path of $\{W(t)\}_{t\geq 0}$ in Fig. 3.24 for a similar M/G/1 model with reneging.)

We define the *staying function* $\overline{R}(\cdot)$ similarly as in Sect. 3.13.1. For each $y \geq 0$, define the *conditional* probabilities

$$\overline{R}(y) \equiv P(\theta_n = 1 | W_n = y), \quad R(y) \equiv P(\theta_n = 0 | W_n = y),$$

independent of $n = 1, 2, \ldots$. Note that $\overline{R}(0) = 1$, and $\overline{R}(y) + R(y) = 1$, $y \geq 0$.

For each $y \geq 0$, given $W_n = y$, θ_n has a Bernoulli distribution (e.g., p. 26 in [125]). The staying function $\overline{R}(y)$ is the conditional probability of an arrival staying for a full service, given $W_n = y$. Its complement $R(y)$ is the probability of an arrival reneging while in the waiting line, given $W_n = y$.

Using the foregoing definition, $1 - \overline{R}(y)$, $y \geq 0$, is not necessarily a cdf.

We assume: $\overline{R}(0) = 1$; $\overline{R}(y)$, $y \geq 0$, is monotone decreasing in the wide sense (i.e., not strictly monotone—it may be non-increasing); $\overline{R}(y)$, $y > 0$, is *bounded from below* by 0. $\overline{R}(y)$ may be continuous or piecewise continuous; it may be a step function.

Due to boundedness from below and monotonicity, $\lim_{y\to\infty} \overline{R}(y)$ exists. Let

$$\lim_{y\to\infty} \overline{R}(y) = L, 0 \le L \le 1.$$

If $\overline{R}(y) \equiv 1$, $y \ge 0$, the model reverts to a standard M/M$_i$/c queue with no reneging; in that case L= 1 (see Sect. 3.13 and Theorem 3.8.)

4.12.2 System Configuration

The set of possible system configurations is

$$\boldsymbol{M} = \boldsymbol{M}_0 \cup \boldsymbol{M}_1 = \{\boldsymbol{m}|(m_{1,}m_{2,}\ldots,m_c)|0 \le \sum_{i=1}^{c} m_i \le c-1\},$$

where $m_i = \begin{cases} 1 \text{ if server } i \text{ is occupied} \\ 0 \text{ if server } i \text{ is idle} \end{cases}$, *just after* a start of service in some server, since the configuration represents the service rates of those servers *other than* the one just occupied.

There are $\binom{c}{j}$ configurations in which exactly j servers are occupied (i.e., $\sum_{i=1}^{c} m_i = j$). The total number of configurations in $\boldsymbol{M}$ is

$$\sum_{j=0}^{c-1} \binom{c}{j} = 2^c - 1.$$

The number of configurations in $\boldsymbol{M}_0 := \{\boldsymbol{m}|0 \le \sum_{i=1}^{c} m_i \le c-2\}$, is $2^c - 1 - c$. The number of configurations in $\boldsymbol{M}_1 := \{\boldsymbol{m}|\sum_{i=1}^{c} m_i = c-1\}$ (border configurations), is c. (Recall that $\boldsymbol{M}_1 = \boldsymbol{M}_b$.)

4.12.3 State of System and Sample Path

State of System

Denote the state of the system as $\{W(t), M(t)\}_{t\ge 0}$, where $W(t) \ge 0 :=$ *virtual wait*, and $M(t) \in \boldsymbol{M} :=$ *system configuration*, at instant t.

Sample Path

Consider a sample path of $\{(W(t), M(t))\}_{t\ge 0}$. A sample-path diagram has $2^c - 1$ lines corresponding to the zero-wait states $(0, m)$, $m \in \boldsymbol{M}$ (i.e., $W(t) = 0$); and c sheets (pages) corresponding to the positive-wait states (y, m), $y > 0$ (i.e., $W(t) > 0$). (See Fig. 4.11 for the special case c = 2.)

Assume the system starts empty at $t = 0$. Initially, arriving customers wait 0, complete service and depart. Eventually customers in service accumulate until $c - 1$ servers are occupied. Concurrently the SP moves among the $2^c - 1 - c$ lines for the non-border zero-wait states. It resides on each such line for an exponentially distributed time, making transitions from line to line. Various states unfold until the SP ends up on one of the c *border* lines.

All zero-wait arrivals stay for full service (no balking). Assume that a new arrival C_τ finds $c - 1$ servers occupied (SP on a border line). Then C_τ waits 0, and starts service in the single idle server. At τ^- the configuration is some $\boldsymbol{m} \in \boldsymbol{M}_b$. At instant τ all c servers are occupied. The SP jumps at instant τ to one of the c sheets, depending on which service will finish first. The probability that server k will finish first is μ_k/μ where $\mu := \mu_1 + \cdots + \mu_c$. The SP will be at a height $\underset{dis}{=} \text{Exp}_\mu$, since the inter start-of-service depart time $\mathcal{S}$ is the *minimum* of c independent exponentially distributed r.v.s with rates $\mu_1, \ldots, \mu_c$, due to the memoryless property.

Let $\boldsymbol{m}_{\bar{i}}$ denote a *border* configuration such that the rate-μ_i server (i.e., server i) is *idle* (see Remark 4.20). In configuration $\boldsymbol{m}_{\bar{i}}$, $m_j = 1$, if $j \neq i$, and $m_i = 0$, i.e.,

$$m_1 + \cdots + m_{i-1} + 0 + m_{i+1} + \cdots + m_c = c - 1.$$

At time τ the SP will end up at a positive height on page $\boldsymbol{m}_{\bar{k}}$ with probability μ_k/μ, $k = 1, \ldots, c$.

Remark 4.20 We use the notation $\bar{\imath}$ to shorten the representation of $\boldsymbol{m}$ if c is large. If c is small, e.g. $c = 3$, we can use notation like 100, 010, 001, 110, 101, 011. If $c = 2$, we can use 01, 10—see Sect. 4.12.8.

4.12.4 Zero-Wait Probabilities

Let P_n, $n = 0, \ldots, c - 1$ denote the steady-state probability of n customers in the system at an arbitrary point in time. Let $P_{n,\boldsymbol{m}}$ denote the probability that there are n customers in the system and the configuration is $\boldsymbol{m} \in \boldsymbol{M}$. There are $\binom{c}{n}$ configurations such that $\sum_{i=1}^{c} m_i = n$. Let

$$\boldsymbol{M}_n = \{\boldsymbol{m} | \sum_{i=1}^{c} m_i = n\}.$$

Thus

$$P_n = \sum_{\boldsymbol{m} \in \boldsymbol{M}_n} P_{n,\boldsymbol{m}}, n = 0, \ldots, c-1.$$

Due to Poisson arrivals P_n is the probability that an arrival waits 0 and "sees" n other customers in service just before it starts service (using PASTA, e.g., [145]).

Remark 4.21 For the **zero-wait** states, a configuration specifies the service rates in the servers at an arbitrary time point. Due to Poisson arrivals, this is the same as the service rates **just before** an arrival. It is also the same as the service rates in the **other** servers **just after** an arrival starts service in an available server.

The probability of a zero wait is denoted by $F(0)$, where

$$F(0) = \sum_{n=0}^{c-1} P_n = \sum_{n=0}^{c-1} \sum_{\boldsymbol{m} \in \boldsymbol{M}_n} P_{n,\boldsymbol{m}}. \tag{4.105}$$

4.12.5 Positive-Wait PDF and CDF

For the **positive-wait** states, a configuration defines the service rates in the **other** servers **just after** start of service.

Let $f_{\boldsymbol{m}}(x), x > 0$, denote the *partial* pdf of wait for page $\boldsymbol{m} \in \boldsymbol{M}_b$. Denote the marginal pdf for the *cover* as

$$f(x) = \sum_{\boldsymbol{m} \in \boldsymbol{M}_1} f_{\boldsymbol{m}}(x), x > 0.$$

The *total* pdf of wait is $\{\boldsymbol{P}_0, f(x), x > 0\}$. The cdf of wait is $F(x) = F(0) + \int_{y=0}^{x} f(y)dy, x \geq 0$, where $F(0)$ is defined in (4.105). The normalizing condition is

$$\lim_{x \to \infty} F(x) = F(0) + \int_{y=0}^{\infty} f(y)dy = 1. \tag{4.106}$$

4.12.6 Equations for Positive-Wait PDFs

A key assumption of this model is that each positive-wait arrival reneges from the waiting line with probability $R(y)$, and stays for complete service

with probability $\overline{R}(y)$ (=1 − $R(y)$), where $y \geq 0$ is the required wait before service.

Equation for Total PDF $f(x)$ We first derive an integral equation for $f(x)$, the total pdf of wait of stayers (who wait and receive a full service), namely,

$$f(x) = \lambda P_{c-1} e^{-\mu x} + \lambda \int_{y=0}^{x} e^{-\mu(x-y)} \overline{R}(y) f(y) dy, \, x > 0, \qquad (4.107)$$

directly using the sample path, as follows (see Fig. 4.11).

Explanation of Equation (4.107) On the left side, $f(x)$ is the *total SP downcrossing rate* of level x over all c sheets, *projected onto the "cover"*. On the right side, since all zero-wait arrivals stay for full service ($\overline{R}(0) = 1$), the term $\lambda P_{c-1} e^{-\mu x}$ is the total SP upcrossing rate of level x due to jumps starting at level 0 (i.e., line 0) of any of the c sheets (from border states $\{(0, \boldsymbol{m}_{\bar{i}})\}$, $i = 1, \ldots, c$), at arrival instants. These jumps have size $\mathcal{S} \underset{dis}{=} \text{Exp}_{\mu}$ ($\underset{dis}{=} \min_{i=1,\ldots,c} \left\{\text{Exp}_{\mu_i}\right\}$). The term $\lambda \int_{y=0}^{x} e^{-\mu(x-y)} \overline{R}(y) f(y) dy$ is the rate at which the SP upcrosses level x due to jumps starting at levels $y \in (0, x)$, on any page, at arrival instants of stayers. The right side is, therefore, the total SP upcrossing rate of level x. Rate balance across x yields (4.107).

Comparing (4.107) with Eq. (3.211) implies that the solution of (4.107) is

$$f(x) = \lambda P_{c-1} e^{-\left(\mu x - \lambda \int_{y=0}^{x} \overline{R}(y) dy\right)}, \, x > 0, \qquad (4.108)$$

where $\mu = \sum_{i=1}^{c} \mu_i$ and $P_{c-1} = \sum_{i=1}^{c} P_{c-1,\boldsymbol{m}_{\bar{i}}}$.

Equations for Partial PDFs $f_{\bar{i}}(x), x > 0, i = 1, \ldots, c$ We now obtain integral equations for the pdfs $f_{\bar{i}}(x), x > 0$, on the c sheets (see Fig. 4.11); they are

$$\begin{aligned} & f_{\bar{i}}(x) + \lambda(1 - \frac{\mu_i}{\mu}) \int_{y=x}^{\infty} \overline{R}(y) f_{\bar{i}}(y) dy \\ & = \lambda \frac{\mu_i}{\mu} P_{c-1} e^{-\mu x} + \lambda \frac{\mu_i}{\mu} \int_{y=0}^{x} e^{-\mu(x-y)} \overline{R}(y) f(y) dy \\ & \quad + \lambda \frac{\mu_i}{\mu} \int_{y=x}^{\infty} \overline{R}(y) \left(f(y) - f_{\bar{i}}(y)\right) dy, \, i = 1, \ldots, c. \end{aligned} \qquad (4.109)$$

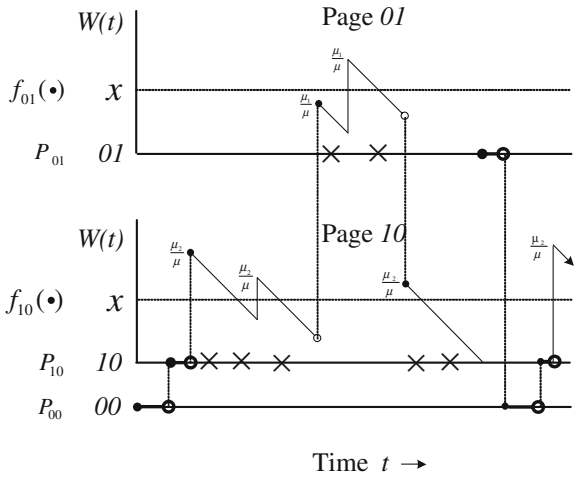

Fig. 4.11 Sample path of $\{W(t), M(t)\}_{t\geq 0}$ in M/M$_i$/2, where zero-wait arrivals join, and positive-wait arrivals may renege from the waiting line. Times marked $\times$ indicate arrivals that renege (do not contribute to the limiting pdf of wait)

Explanation of Equation (4.109) On the left side, $f_{\bar{i}}(x)$ is the SP *exit* rate from composite state $((x, \infty), \bar{i})$ due to SP downcrossings of level x; term $\lambda(1 - \frac{\mu_i}{\mu}) \int_{y=x}^{\infty} \overline{R}(y) f_{\bar{i}}(y) dy$ is the SP rate of jumps *out of* $((x, \infty), \bar{i})$ into the composite states $((x, \infty), \bar{j})$, $j \neq i$, on other sheets. On the right side, the first two terms are SP *entrance* rates into $((x, \infty), \bar{i})$ due to jumps starting at level-0 border states, and jumps starting at levels $y \in (0, x)$ on any sheet, respectively (recall $f(y) = \sum_{i=1}^{c} f_{\bar{i}}(y)$). The third term is the SP *entrance* rate into $((x, \infty), \bar{i})$ due to jumps starting in $\cup_{j\neq i}((x, \infty), \bar{j})$. Rate balance of SP exits and entrances of $((x, \infty), \bar{i})$ yields (4.109).

Solution of Equation (4.109) We obtain the solution of (4.109) in terms of the solution for $f(x)$, which is given in formula (4.108), using the following Proposition.

Proposition 4.1 *The partial pdf is given by*

$$f_{\bar{i}}(x) = \frac{\mu_i}{\mu} f(x), x > 0, i = 1, \ldots, c. \tag{4.110}$$

Proof Substitute $\frac{\mu_i}{\mu} f(x)$ for $f_{\bar{i}}(x)$ in Eq. (4.109), and cancel like terms. The proposition is true if and only if the following is an identity:

$$\frac{\mu_i}{\mu} f(x) + \lambda \int_{y=x}^{\infty} \frac{\mu_i}{\mu} \overline{R}(y) f(y)) dy$$
$$= \lambda \frac{\mu_i}{\mu} P_{c-1} e^{-\mu x} + \lambda \frac{\mu_i}{\mu} \int_{y=0}^{x} e^{-\mu(x-y)} \overline{R}(y) f(y) dy$$
$$+ \lambda \frac{\mu_i}{\mu} \int_{y=x}^{\infty} \overline{R}(y) f(y) dy, x > 0, \quad (4.111)$$

if and only if

$$f(x) = \lambda P_{c-1} e^{-\mu x} + \lambda \int_{y=0}^{x} e^{-\mu(x-y)\overline{R}(y)} f(y) dy, x > 0, \quad (4.112)$$

is an identity. Equation (4.112) is identical to Eq. (4.107). Hence the Proposition is true. ■

Exponential Staying Function
Consider an exponential staying function, $\overline{R}(x) := e^{-rx}, r > 0, x \geq 0$. (Note that $0 < e^{-rx} \leq 1$, and is strictly decreasing on $(0, \infty)$, satisfying the definition of staying function.) The total pdf $f(x)$ is now obtained by substituting e^{-ry} for $\overline{R}(y)$ in (4.108), which substituted into (4.110), gives

$$f_{\bar{i}}(x) = \lambda \frac{\mu_i}{\mu} e^{\frac{\lambda}{r}} P_{c-1} e^{-\mu x - \frac{\lambda}{r} e^{-rx}}, x > 0, i = 1, \ldots, c. \quad (4.113)$$

We shall solve an M/M$_i$/2 model using $\overline{R}(x) := e^{-rx}, r > 0, x \geq 0$, in Sect. 4.12.8 below.

4.12.7 Equations for Zero-Wait Probabilities

Assume that the zero-wait server assignment policy is: arrivals that find k available servers, $1 \leq k \leq c$, get served by a particular available server with probability $1/k$. (Other policies are also viable, e.g., the arrival gets served by the lowest-numbered available server, or by the fastest-available service rate, etc.) Using the principle *SP exit rate = SP entrance rate* for the zero-wait states, we obtain the equations (notation explained below)

$$
\begin{aligned}
(\lambda+\mu-\mu_i)P_{c-1,\bar{i}} &= f_{\bar{i}}(0)+\tfrac{\lambda}{2}\textstyle\sum_{j\in \boldsymbol{J}_i}^{c} P_{c-2,\overline{ij}},\ i=1,\ldots,c,\\
(\lambda+\mu-\mu_i-\mu_j)P_{c-2,\overline{ij}} &= \mu_j P_{c-1,\bar{i}}+\mu_i P_{c-1,\bar{j}}\\
&\quad +\tfrac{\lambda}{3}\textstyle\sum_{k\in J_{ij}}^{c} P_{c-3,\overline{ijk}},\ j=1,\ldots,c,\\
&\cdots\\
(\lambda+\mu_i)P_{1,i} &= \textstyle\sum_{k\neq i=1}^{c}\mu_k P_{2,ik}+\dfrac{\lambda}{c}P_{00},\ i=1,\ldots,c,\\
\lambda P_{00} &= \textstyle\sum_{i=1}^{c}\mu_i P_{1,i}.
\end{aligned}
\tag{4.114}
$$

Notation in equations (4.114) In the first c equations, the index j of the sum takes values in $\boldsymbol{J}_i := \{j \,|\, j=0,\ldots,c,\ j\neq i\}$, and the subscript $\overline{ij}$ means both servers$i$ and j are idle. In the second set of $\binom{c}{2}$ equations, the index k of the sum takes values in $\boldsymbol{J}_{ij}=\{k|k=0,\ldots,c,k\neq i,k\neq j\}$, and the subscript $\overline{ijk}$ means all three servers i, j and k are idle. The row of dots "$\cdots$" indicates similar rate balance equations for $P_{c-3,\overline{ijk}}$, …, $P_{2,\cdot}$. In the second last equation, for $P_{1,i}$, on the right side $P_{2,ik}$ denotes the probability of two units in the system, in servers i and k having service rates μ_i and μ_k respectively.

We solve Eq. (4.114) explicitly in Sect. 4.12.8 below for M/M$_i$/2, in order to convey some characteristics of the solution.

4.12.8 Solution for M/M$_i$/2 with Reneging

Notation When there is a small number of servers we can use an alternative, perhaps more familiar notation. If $c=2$, there are two sheets corresponding to configurations $\overline{1}$ and $\overline{2}$, which we now replace by 01 and 10 respectively. Thus, configuration 01 means server 1 is available and server 2 is occupied; configuration 10 means server 2 is available and server 1 is occupied.

Applying formula (4.113), the partial pdfs of wait are now denoted by

$$
\begin{aligned}
f_{10}(x) &= \lambda\tfrac{\mu_2}{\mu}e^{\frac{\lambda}{r}}P_1 e^{-\mu x-\frac{\lambda}{r}e^{-rx}},\ x>0,\\
f_{01}(x) &= \lambda\tfrac{\mu_1}{\mu}e^{\frac{\lambda}{r}}P_1 e^{-\mu x-\frac{\lambda}{r}e^{-rx}},\ x>0.
\end{aligned}
\tag{4.115}
$$

The marginal ("total") pdf of wait is

$$
f(x)=f_{10}(x)+f_{01}(x)=\lambda e^{\frac{\lambda}{r}}P_1 e^{-\mu x-\frac{\lambda}{r}e^{-rx}},\ x>0. \tag{4.116}
$$

The zero-wait probabilities are $P_{1,\bar{i}}$, $i = 1, 2$, and P_{00}; using the alternative notation we have

$$\begin{aligned} P_1 &= P_{1,\bar{2}} + P_{1,\bar{1}} = P_{10} + P_{01}, \\ P_0 &= P_{00} + P_{1,\bar{2}} + P_{1,\bar{1}} = P_0 + P_{10} + P_{01} \\ &= P_{00} + P_1. \end{aligned}$$

The rate-balance equations for the zero-wait probabilities are

$$\begin{aligned} (\lambda + \mu_1)P_{10} &= \tfrac{\lambda}{2}P_{00} + f_{10}(0), \\ (\lambda + \mu_2)P_{01} &= \tfrac{\lambda}{2}P_{00} + f_{01}(0), \\ \lambda P_{00} &= \mu_1 P_{10} + \mu_2 P_{01}. \end{aligned} \tag{4.117}$$

Substituting for $f_{10}(0)$, $f_{01}(0)$ from (4.115), we rewrite the equations in (4.117) as

$$\begin{aligned} (\lambda + \mu_1)P_{10} &= \tfrac{\lambda}{2}P_{00} + \lambda\tfrac{\mu_2}{\mu}P_1, \\ (\lambda + \mu_2)P_{01} &= \tfrac{\lambda}{2}P_{00} + \lambda\tfrac{\mu_1}{\mu}P_1, \\ \lambda P_{00} &= \mu_1 P_{10} + \mu_2 P_{01}. \end{aligned} \tag{4.118}$$

The solution of (4.118) in terms of P_{00} is

$$\begin{aligned} P_{01} &= \tfrac{\lambda}{2\mu_2}P_{00}, \\ P_{10} &= \tfrac{\lambda}{2\mu_1}P_{00}, \\ P_1 &= \tfrac{\lambda(\mu_1+\mu_2)}{2\mu_1\mu_2}P_{00} = \tfrac{\lambda\mu}{2\mu_1\mu_2}P_{00}. \end{aligned} \tag{4.119}$$

The normalizing condition

$$P_{00} + P_1 + \int_{x=0}^{\infty} f(x)dx = 1,$$

yields

$$P_{00} = \left(1 + \frac{\lambda(\mu_1+\mu_2)}{2\mu_1\mu_2} + \frac{\lambda(\mu_1+\mu_2)}{2\mu_1\mu_2}\lambda e^{\frac{\lambda}{r}}\int_{x=0}^{\infty} e^{-\mu x - \frac{\lambda}{r}e^{-rx}}dx\right)^{-1}. \tag{4.120}$$

The analytic solution comprises the results in (4.120), (4.119), (4.116) and (4.115).

Example 4.10 We present a numerical example for the M/M$_i$/2 queue with reneging allowed from the waiting line (see Fig. 4.12). Let

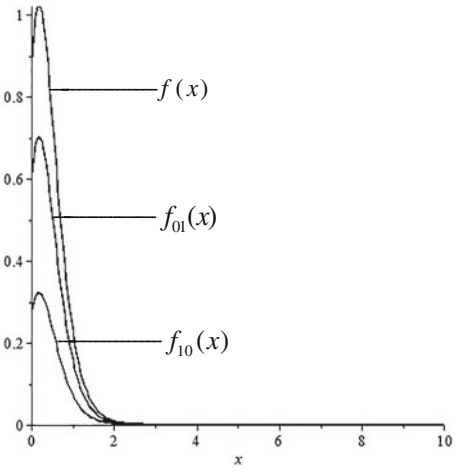

Fig. 4.12 Plot of $f(x)$, $f_{01}(x)$ $(= f_{\overline{1}}(x))$, $f_{10}(x)$ $(= f_{\overline{2}}(x))$, in Example 4.10

$$\lambda = 5.2,\ \mu_1 = 2.4,\ \mu_2 = 1.1,\ \mu = \mu_1 + \mu_2 = 3.5,\ r = 2.1.$$

Then

$$\int_{x=0}^{\infty} e^{-\mu x - \frac{\lambda}{r}e^{-rx}} dx = 0.074741,$$

and

$$\begin{aligned}
P_{00} &= 0.049059, \quad P_{01} = 0.115958, \quad P_{10} = 0.053147,\\
P_1 &= 0.169105,\\
F(0) &= P_{00} + P_1 = 0.218164,\\
F(\infty) &= F(0) + \lambda e^{\frac{\lambda}{r}} P_1 \int_{x=0}^{\infty} e^{-\mu x - \frac{\lambda}{r}e^{-rx}} dx\\
&= 0.218164 + 0.781836 = 1.0,\\
f(x) &= \lambda e^{\frac{\lambda}{r}} P_1 e^{-\mu x - \frac{\lambda}{r}e^{-rx}} = 10.461 \cdot e^{-3.5x - 2.476e^{-2.1x}},\\
f_{01}(x) &= \frac{\mu_1}{\mu} f(x) = 7.173 \cdot e^{-3.5x - 2.476e^{-2.1x}},\\
f_{10}(x) &= \frac{\mu_2}{\mu} f(x) = 3.288 \cdot e^{-3.5x - 2.476e^{-2.1x}}.
\end{aligned}$$

Remark 4.22 In M/M$_i$/c with reneging from the waiting line allowed, we can generalize the staying function $\overline{R}(x)$, $x \geq 0$. For example, $\overline{R}(x)$ may depend on the server that would be occupied by an arrival, i.e., on the system configuration at the arrival instant. We may then use the notation $\overline{R}_{\overline{i}}(x)$. Thus $\overline{R}_{\overline{i}}(x)$ may depend on, not only customer required wait before service, but also

on customer attraction or aversion to the "target" server. A natural question arises. Can this model be modified to study attraction or aversion in natural processes such as: electrically charged particles approaching an electrically charged or magnetized environment; asteroids approaching a planet; particles adhering or falling away from a surface; laser pulses affecting cells containing certain chemicals in biological or medical applications; etc.?

4.12.9 Stability Condition

Consider the M$_\lambda$/M$_i$/c ($c \geq 2$) queue with heterogeneous servers having rates $\mu_1, \ldots, \mu_c$ in which reneging depending on required wait is allowed before service begins. Let the staying function $\overline{R}(x), x \geq 0$, be monotone decreasing (includes non-increasing), let $\overline{R}(0) = 1$ (no balking upon arrival), and assume $0 \leq \overline{R}(x) \leq 1$, $x \geq 0$. Let $L = \lim_{x\to\infty} \overline{R}(x)$, which exists by monotonicity and boundedness. The ideas in Theorem 3.8 also apply in the M/M$_i$/c environment, as follows.

Theorem 4.10 In M$_\lambda$/M$_i$/c ($c \geq 2$) with reneging from the waiting line allowed, as described immediately above, a necessary and sufficient condition for stability is

$$\lambda < \frac{\mu}{L} \text{ if } 0 < L \leq 1,$$
$$\lambda < \infty \text{ if } L = 0,$$

where $\mu = \sum_{i=1}^{c} \mu_i$.

Proof The proof is similar to that of Theorem 3.8. The alternative proof given there, Remark 3.31 and Fig. 3.28 also apply for the present M/M$_i$/c queue with reneging, upon substituting $\mu = \sum_{i=1}^{c} \mu_i$. ■

4.13 Discussion

We can use LC to analyze a vast array of additional M/M/c models. We mention only a few examples.

LC has been applied to M/M/c queues in which customers receive simultaneous service from a random number of servers. The original source for such queueing models is the Ph.D. thesis of L. Green [81, 82]. An LC analysis, motivated by the work of L. Green, is given in [38].

LC has been applied to M/M/c with bounded system time (wait + service). An arrival balks upon arrival if its system time would exceed an upper bound K, e.g., a system manager informs an arrival of the current expected system time (see Example 1, p. 44 in [52]). This generalizes variant 2 of the M/G/1 model discussed above in Sect. 3.6. It is straightforward to apply LC to analyze an M/M/c model analogous to variant 1 in Sect. 3.16. In that model customers renege from *service* if their *age* in the system *(elapsed system time)* reaches K. Similar remarks apply to M/M/c where the actual waits are bounded by K (as in variant 3 in Sect. 3.16). In that case the workload can exceed K. We can develop an expression for the tail of the steady-state pdf of workload, from its integral equation.

LC can be used to analyze a variety of M/M/c queues with server vacations; priorities; and many others.