# Stationary analysis of a single queue with remaining service time dependent arrivals

**2 authors**, including:

Benjamin Legros
Ecole Centrale Paris

**20** PUBLICATIONS   **21** CITATIONS

Some of the authors of this publication are also working on these related projects:

Project   Routing in queue with abandonments View project

# Stationary analysis of a single queue with remaining service time dependent arrivals

Benjamin Legros[1] • Ali Devin Sezer[2]

[1]*Laboratoire Genie Industriel, CentraleSupélec Université Paris-Saclay*
*Grande Voie des Vignes, 92290 Chatenay-Malabry, France*

[2] *Middle East Technical University Institute of Applied Mathematics*
*Ankara, 06800, Turkey*

benjamin.legros@centraliens.net • devin@metu.edu.tr

July 31, 2017

**Abstract**

We study a generalization of the $M/G/1$ system (denoted by $rM/G/1$) with independent and identically distributed (iid) service times and with an arrival process whose arrival rate $\lambda_0 f(r)$ depends on the remaining service time $r$ of the current customer being served. We derive a natural stability condition and provide a stationary analysis under it both at service completion times (of the queue length process) and in continuous time (of the queue length and the residual service time). In particular, we show that the stationary measure of queue length at service completion times is equal to that of a corresponding $M/G/1$ system. For $f > 0$ we show that the continuous time stationary measure of the $rM/G/1$ system is linked to the $M/G/1$ system via a time change. As opposed to the $M/G/1$ queue, the stationary measure of queue length of the $rM/G/1$ system at service completions differs from its marginal distribution under the continuous time stationary measure. Thus, in general, arrivals of the $rM/G/1$ system do not see time averages. We derive formulas for the average queue length, probability of an empty system and average waiting time under the continuous time stationary measure. We provide examples showing the effect of changing the reshaping function on the average waiting time.

# 1    Introduction

The goal of the present note is the steady state analysis of a single server queueing system with iid service times and an arrival process whose rate is a function of the remaining service time of the current customer being served, if the server is busy, or a constant $\lambda_0$ otherwise. This is a generalization of the $M/G/1$ system. Because the arrival rate is allowed to depend on the remaining service time we will denote it by the notation '$rM/G/1$.' Arrival processes with remaining service time dependent rates can be used to model systems where customers can directly estimate the remaining service time by observing the amount of work that a server has to treat and use this information to decide whether to join the queue or not. This type of behavior occurs, for example, at checkout queues in supermarkets. A potential application area for $rM/\cdot/\cdot$ systems is call centers [1, 7] with inbound and outbound calls. Modern call centers call out customers to connect them with a server even when all servers are busy [17]; the decision to initiate an outbound call can use estimates of the remaining service time of the busy servers. New approaches to call center modeling also allow the control of the arrival process of inbound calls by postponing their routing to an agent or by giving incentives to callback later [13]; such approaches can make use of estimates of the remaining service time of servers. Generalizations of the $rM/G/1$ model may be useful in the analysis of these systems.

Queues with queue-length dependent and Markov modulated arrival or service time distributions have been studied in the literature, see, e.g., [6, 4, 12, 18]. The only works we are aware of allowing the arrival rate to depend on the remaining service time are [9, 10, 11]; these works study the remaining service time process (denoted by $U(t)$ in these works) when the arrival rate and the service rate of the arriving customer depends on $U$ ([10, 11] further contain two state Markov modulation whose transition rates depend on $U$). The analysis method used in these works is asymptotic approximation as arrival, service and transition rates are scaled by a parameter whose value is sent to $\infty$. In the current work, we study, within a narrower framework, the joint queue length and remaining service time distribution and our focus is on finding exact solutions.

2

To simplify exposition, we assume that the iid service times have a density, denoted by $g(\cdot)$. We further comment on this assumption in Section 6. The arrival process of customers is Poisson with constant arrival rate $\lambda_0$ if the system is empty or $\lambda_0 f(r)$ if the server is busy and the remaining service time of the customer being served is $r$. In the particular case where $f(r) = 1$ for $r \geq 0$, the system reduces to an $M/G/1$ queue. $f$ can be interpreted in two ways: if $f(r) \in (0,1)$, $r \in \mathbb{R}_+$, then $f(r)$ can be thought of as the probability that an arriving customer joins the queue after having observed the remaining service time $r$. $f$ can also be thought of as a control parameter that transforms / reshapes, the constant arrival rate $\lambda_0$ to optimize system performance. With this interpretation in mind, we will refer to $f$ as the "reshaping function" (the '$r$' in the abbreviation $rM/G/1$ refers also to "reshaping" of the arrival process). For the latter interpretation, a natural condition on a reshape function is that it doesn't change the overall average arrival rate to the system. In Proposition 10 of Subsection 4.2 the average arrival rate to an $rM/G/1$ system is computed to be $\alpha = \frac{\lambda_0}{1-\lambda_0(\bar{\nu}-\nu)}$, where $\nu = \int_0^\infty rg(r)dr$ is the average service length and $\bar{\nu} = \int_0^\infty F(r)g(r)dr$ with $F(r) = \int_0^r f(u)du$. Thus, under the assumption

$$\nu = \bar{\nu} \tag{1}$$

the average arrival rate of an $rM/G/1$ system remains $\lambda_0$. This assumption will be in force in Section 5 where we compare the average waiting times of a range of $rM/G/1$ systems with the same service time distribution and average arrival rate $\lambda_0$ but different reshape functions.

A natural framework for the study of the $rM/G/1$ queue is the piecewise deterministic processes (PDP) of [5]. Section 2 gives a construction of the $rM/G/1$ process as a piecewise deterministic Markov process based on this framework. The process is $X_t = (N_t, R_t)$; its first component represents the number of customers (i.e., queue length, including the customer being served) in the system the second component represents the remaining service time. Subsection 2.1 gives its generator and Subsection 2.2 derives the dynamics of the embedded random walk $\mathcal{N}$, which is the sequence of queue lengths observed at service completion times; Proposition 2 shows that the dynamics of $\mathcal{N}$ equals that of the embedded random walk (at service completion times) of an $M/G/1$ queue (whose state process is denoted by $\bar{X}$) with constant arrival rate $\lambda_0$ and with iid service times $\{\bar{\sigma}_k, k = 1, 2, 3, \}$ where

3

$\bar{\sigma}_k = F(\sigma_k)$, and $\{\sigma_k\}$ are the iid service times of the original $rM/G/1$ system. The stationary distribution of the $rM/G/1$ system at service completions (and arrivals) follows from this reduction; the details are given in Section 3. Proposition 3 derives the stability condition $\rho \doteq \lambda_0 \bar{\nu} < 1$, (15) gives the expected stationary queue length at service completions and (16) gives the stationary moment generating function of the queue length distribution at service completions.

As opposed to $M/G/1$ queues, the stationary distribution of queue length of an $rM/G/1$ system in continuous time does not equal its stationary distribution at service completions; therefore, for $rM/G/1$ queues, the continuous time stationary distribution and service measures based on it must be computed directly. Section 4 begins with the statement and recursive solution of the balance equation for the stationary distribution of the continuous time process $X$, which consists essentially of a sequence of linear ordinary differential equations (ODE) where $f$ serves as an $r$ dependent coefficient. Proposition 8 proves that the solution of the balance equation is indeed the stationary measure of the process $X$ under the stability assumption $\rho < 1$. The proof is based on the PDP framework of [5]. A number of further computations based on the continuous time stationary distribution is given in Section 4; in particular, Corollary 2 gives a simple formula for the stationary probability of an empty $rM/G/1$ system in continuous time and Proposition 9 gives a formula for the stationary expected queue length in continuous time. Proposition 10 of Subsection 4.2 gives the average arrival rate for the $rM/G/1$ system and finally (64) gives an explicit formula for the average sojourn time of a customer in an $rM/G/1$ system. In general, $f$ may take the value 0 and this may make $F$ noninvertible. For this reason, there is not, in general, a bijective correspondence between the continuous time stationary distribution of the $rM/G/1$ process $X$ and that of the $M/G/1$ process $\bar{X}$. However, for $f > 0$ a bijective correspondence can be established; this is treated in Subsection 4.4.

Section 5 gives two examples showing the impact of reshaping the arrival process on the average waiting time. We observe, as expected, that, for a given average arrival rate, the closer the customers arrive to the end of a service the shorter will be the average waiting time in the system. Section 6 points out directions for future research.

## 2 Dynamics of the process

The theory of piecewise-deterministic Markov Processes (PDP) of [5] provides the ideal mathematical framework for the analysis of the $rM/G/1$ queue. For the definition of the process we will use the PDP definition given in [5, page 57], which uses the following elements (all adopted from [5]): the state space of the process will be

$$E \doteq \bigcup_{k=0}^{\infty} E_k, E_0 = B(\mathbf{0}, \delta) \subset \mathbb{R}^2, \quad E_k \doteq \{k\} \times \mathbb{R}_+ = \{(k, r), r > 0\}, \quad k \in \{1, 2, 3, ...\},$$

where $\mathbf{0} = (0, 0) \in \mathbb{R}^2$ denotes the origin of $\mathbb{R}^2$ and $B(\mathbf{0}, \delta)$ denotes an open ball of radius $\delta < 1$; $\mathbf{0}$ represents the empty system (in [5] the letter $\zeta$ denotes the second component of $x \in E$, we use $r$ for the same purpose). The $rM/G/1$ process, $X_t = (N_t, R_t) \in E$, $t \geq 0$, will evolve, on each $E_k$ smoothly following the vector field $\mathfrak{X}_k : E_k \mapsto \mathbb{R}^2$

$$\mathfrak{X}_k(x) \doteq \begin{cases} (0, -1), & k > 0, \\ 0, & \text{otherwise,} \end{cases}$$

until it jumps. Let us denote the jump times of $X$ by the sequence $\{T_i, i = 1, 2, 3, ...\}$. The vector field $\mathfrak{X}_k$ defines the following trivial flow

$$\phi(t, (k, r)) = (k, r - t), \quad k > 0, \qquad \phi(t, \mathbf{0}) = \mathbf{0}; \tag{2}$$

the process $X$ follows this flow in between its jumps:

$$X_t = (N_{T_k}, \phi(t, X_{T_k})) = (N_{T_k}, R_{T_k} - (t - T_k)), T_k < t < T_{k+1}. \tag{3}$$

For $A \subset \mathbb{R}^2$, let $\partial A$ denote its boundary in the Euclidian topology. The exit boundary of the process is

$$\Gamma^* \doteq \cup_{k=0}^{\infty} \partial E_k = \partial B(\mathbf{0}, \delta) \cup (\cup_{k=1}^{\infty} \{(k, 0), k > 0\}).$$

For $x = (k, r) \in E$, define (following [5, page 57]) $t_*(x) \doteq \inf\{t > 0, \phi_k(t, r) \in \partial E_k\}$ where we use the convention that the infimum of the empty set is empty; $t_*(x)$ is the time when $X$ reaches $\Gamma^*$ if it doesn't doesn't jump until this happens. By definition ((2) and (3)) $X$

5

moves with unit speed toward the $k$-axis on each $E_k$, $k > 0$, therefore,

$$t_*(x) = r, x = (k, r) \in E, k > 0.$$

For $k = 0$, the process remains constant $\mathbf{0}$ until an arrival occurs, which implies $t_*(\mathbf{0}) = \infty$. Figure 1 shows an example sample path of $X$; the horizontal axis is the $k$-axis, showing the number of customers in the system and the vertical axis is the $r$-axis, showing the remaining service time of the current customer in service. The dynamics (3) means that $X$ travels with unit speed toward the $k$-axis in between its jumps. Two types of jumps are possible: either an arrival, which are jumps to the right or a service completion, which are jumps to the left occurring when $X$ hits the $k$-axis.

The jump dynamics are specified by the rate function $\boldsymbol{\lambda} : E \to \mathbb{R}_+$ and the transition measure $Q$. For the $rM/G/1$ system the jump rate function will be

$$\boldsymbol{\lambda}(k, r) \doteq \begin{cases} \lambda_0 f(r), & k > 0, \\ \lambda_0, & k = 0. \end{cases}$$

The transition measure $Q(\cdot, x)$, $x \in E \cup \Gamma^*$ for the $rM/G/1$ system will be as follows: $Q(\cdot, x)$ is the Dirac measure on $(k + 1, r)$ for $x = (k, r)$, $k > 0$ and $r > 0$ (represents an arrival to the busy system). For $(k, 0) \in \Gamma^*$, $k > 0$, $Q(\cdot, x)$ is the measure $g(r)dr$ on $E_{k-1}$ (represents the completion of a service and the start of another, this is exactly when the sample path $X$ hits the $k$-axis in Figure 1); $Q(\cdot, \mathbf{0})$ is the measure $g(r)dr$ on $E_1$ (represents an arrival to the empty system).

## 2.1 Generator of $X$

Let $\mathscr{E}$ denote the $\sigma$-algebra of Borel-measurable subsets of $E$. Let $\{T_n, n = 1, 2, 3, ..\}$ denote the jump times of $X$. For $h : E \times \mathbb{R}_+ \times \Omega \mapsto \mathbb{R}$, $h$ measurable, one writes $h \in L_1(X)$ if

$$\mathbb{E}\left[\sum_{i=1}^{\infty} |h(X_{T_i}, T_i, \omega)|\right] < \infty$$

and $h \in L_1^{\text{loc}}(X)$ if $h\mathbb{1}_{\{t < \sigma_n\}} \in L_1(X)$ for a sequence of stopping times $\sigma_n \nearrow \infty$. The characterization of the generator of $X$ given in the next paragraph uses these definitions.

The generator of any PDP process is derived explicitly in [5, Theorem (26.14), page 69];
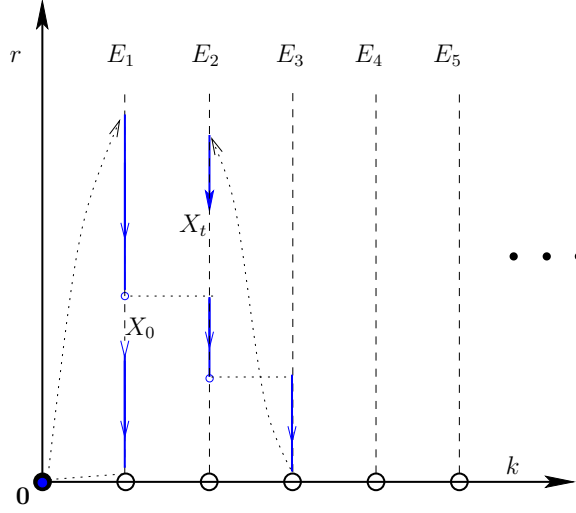
6

Figure 1: The state space and a sample path of $X$

for the $rM/G/1$ process $X$, it is given by the following operator:

$$\mathfrak{A}h(x) = \begin{cases} -\frac{d}{dr}h(x) + \lambda_0 f(r)\left(h(k+1,r) - h(k,r)\right), x = (k,r), k, r > 0, \\ +\lambda_0 g(r)(h(1,r) - h(\mathbf{0})), x = \mathbf{0}, \end{cases}$$

where $h \in \mathscr{D}(\mathfrak{A})$, the domain of $\mathfrak{A}$, consisting of measurable functions $h$ on $E \cup \Gamma^*$ satisfying:

1. for each $k > 0$, $h(k, \cdot)$ is absolutely continuous on $\mathbb{R}_+$,

2. $h(k,0) = \int h(k,r)g(r)dr, k > 0$ and

3. $\mathfrak{B}h \in L_1^{\mathrm{loc}}(X)$ where $\mathfrak{B}h$ is the process $t \mapsto (h(X_0) - h(X_{t-}))$.

## 2.2 Embedded random walk at service completion times

Let $S_k$ denote[1] the sequence of service completion times

$$S_1 \doteq \inf\{t : X_{t-} \in \Gamma^*\}, S_n \doteq \inf\{t > S_{n-1}, X_{t-} \in \Gamma^*\}, n > 1,$$

and define the process $(\mathcal{N}_n, \mathcal{R}_n) \doteq X_{S_n}$, the state of the system right after service completions. Let

$$\boldsymbol{p}(k,\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, 3, \ldots$$

---

[1]In [5], $S_k$ denotes the inter-jump times of the PDP, here they denote the successive times when the process hits the boundary of its state space.

7

denote the Poisson distribution with rate $\lambda$ and define

$$F(r) \doteq \int_0^r f(r-u)du = \int_0^r f(u)du.$$

**Proposition 1.** *The process $\{(\mathcal{N}_n, \mathcal{R}_n), n = 1, 2, 3, ...\}$ is a Markov chain with transition probabilities*

$$P(\mathcal{N}_{n+1} - \mathcal{N}_n = j | \mathcal{N}_n, \mathcal{R}_n) = \begin{cases} \boldsymbol{p}(j+1, F(\mathcal{R}_n)), & \mathcal{N}_n > 0, \\ \boldsymbol{p}(j, F(\mathcal{R}_n)), & \mathcal{N}_n = 0 \end{cases} \tag{4}$$

$$P(\mathcal{R}_{n+1} \in A | \mathcal{N}_{n+1}) = \begin{cases} \delta_0(A), & \mathcal{N}_{n+1} = 0, \\ \int_A g(r)dr, & \mathcal{N}_{n+1} > 0, \end{cases} \tag{5}$$

*where $\delta_0$ denotes the Dirac measure on 0.*

*Proof.* The definition of the process $X$ (or its strong Markov property) implies that $(\mathcal{N}, \mathcal{R})$ is a Markov chain. The jump distribution $Q$ determines where $X$ jumps after it hits $\Gamma^*$: $X$ jumps to $(N_{S_n-} - 1, \sigma_n)$ where $\sigma_n$ has density $g$, if $N_{S_n-} > 1$ or it jumps to $\boldsymbol{0} = (0,0)$ if $N_{S_n-} = 0$. This gives (5). To compute the conditional density of $\mathcal{N}_{n+1}$ given $(\mathcal{N}_n, \mathcal{R}_n)$ it suffices to compute that of

$$\mathcal{N}_{n+1} - \mathcal{N}_n = (N_{S_{n+1}} - N_{S_{n+1}-}) + (N_{S_{n+1}-} - N_{S_n}). \tag{6}$$

By the strong Markov property of $X$, the conditional distribution of $N_{S_{n+1}-} - N_{S_n}$ given $(\mathcal{N}_n, \mathcal{R}_n)$ is the same as that of $N_{S_1-} - N_0$ given $(N_0, R_0)$. The two cases $(N_0, R_0) = \boldsymbol{0}$ and $N_0, R_0 > 0$ are treated separately. Let us start with the latter: conditioned on $(X_0 = (N_0, R_0) = x = (k, r))$, $k > 0$, the dynamics of $X$ imply the following: $S_1 = t_*(x) = r$, and $R_t = r - t$ for $t \in [0, r)$. In the same time interval the $N$ process is Poisson with time dependent rate $\lambda_0 f(r-t)$. Therefore, conditioned on $X_0 = (k, r)$, $k > 0$, $N_{S_1-} - N_0$ has Poisson distribution with rate $F(r)$. Furthermore, for $k > 0$, one has $N_{S_1-} > 0$ and therefore, once again by the definition of the jump dynamics of $X$, $N_{S_1} - N_{S_1-} = -1$ (i.e., the customer whose service has just finished leaves the system). These and (6) imply

$$P(\mathcal{N}_{n+1} - \mathcal{N}_n = j | \mathcal{N}_n, \mathcal{R}_n) = \boldsymbol{p}(j+1, F(\mathcal{R}_n)), \mathcal{N}_n > 0. \tag{7}$$

The argument for the case $X_0 = \mathbf{0}$ is parallel and gives

$$P(\mathcal{N}_{n+1} - \mathcal{N}_n = j|(\mathcal{N}_n, \mathcal{R}_n) = \mathbf{0}) = \boldsymbol{p}(j, F(R_{\tau'_n})) \tag{8}$$

where $\tau'_n$ is the first jump time of $R$ after $S_n$. For $(\mathcal{N}_n, \mathcal{R}_n) = \mathbf{0}$, $\tau'_n$ will be a jump from state $\mathbf{0}$ (i.e., an arrival to the empty system) and by $X$'s definition $R_{\tau'_n}$'s density, given the whole history of $(\mathcal{N}, \mathcal{R})$, will again be $g$. This, (7) and (8) imply (4). $\qquad\square$

The process $\mathcal{N}$ itself is a Markov chain:

**Proposition 2.** $\mathcal{N}$ *is a Markov chain with transition matrix*

$$M = \begin{pmatrix} p(0) & p(1) & p(2) & p(3) & p(4) & \cdots \\ p(0) & p(1) & p(2) & p(3) & p(4) & \cdots \\ 0 & p(0) & p(1) & p(2) & p(3) & \cdots \\ 0 & 0 & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{9}$$

*where*

$$p(j) \doteq \int_0^\infty \boldsymbol{p}(j, \lambda_0 F(r)) g(r) dr. \tag{10}$$

*Proof.* The conditional distributions (4), (5) and the Markov property of the process $\{(\mathcal{N}_n, \mathcal{R}_n)\}$, imply that $\{\mathcal{N}_n, n = 0, 1, 2, 3, ..\}$ is a Markov chain; the distribution of its increments $\Delta \mathcal{N}_n = \mathcal{N}_{n+1} - \mathcal{N}_n$ are

$$P(\Delta \mathcal{N}_n = j|\mathcal{N}_n) = \mathbb{E}[\mathbb{E}[(\mathbb{1}_{\mathcal{N}_n = j}|\mathcal{N}_n, R_n)]\mathcal{N}_n] = \begin{cases} p(j+1), & \mathcal{N}_n > 0, \\ p(j), & \mathcal{N}_n = 0, \end{cases} \tag{11}$$

This implies that $M$ of (9) is the transition matrix of $\mathcal{N}$. $\qquad\square$

We now note the first connection between $rM/G/1$ and $M/G/1$ systems. That $\{\sigma_i\}$ is an iid sequence implies the same for $\bar{\sigma}_i \doteq F(\sigma_i)$. Then one can write (10) as

$$p(j) = \mathbb{E}[\boldsymbol{p}(j, \lambda_0 \bar{\sigma}_1)], j = 0, 1, 2, 3, , ...$$

and, by [16, Proposition 3.3.2, page 57], these are exactly the transition probabilities of the embedded random walk (at service completion times) of an $M/G/1$ system with constant

rate $\lambda_0$ and iid service time sequence $\{\bar{\sigma}_i, i = 1, 2, 3, ..\}$:

**Corollary 1.** *The dynamics at service completion times of the $rM/G/1$ system with arrival rate $\lambda_0 f(\cdot)$ and iid sequence of service times $\{\sigma_i, i = 1, 2, 3, ...\}$ is identical to the dynamics at service completion times of an $M/G/1$ system with constant arrival rate $\lambda_0$ and iid sequence of service times $\{\bar{\sigma}_i = F(\sigma_i), i = 1, 2, 3, ...\}$.*

The next section computes the stationary distribution of $\mathcal{N}$ under a natural stability assumption; before we move on, let us make the following observation:

**Remark 1.** Let $E_t$ denote the elapsed service time since the beginning of current service. If we replace the arrival rate from $\lambda_0 f(R_t)$ with $\lambda_0 f(E_t)$, conditioned on $\mathcal{R}_n = r$, the number of arrivals between the $n^{th}$ service completion and $(n+1)^{st}$ completion will be a Poisson random variable with rate $\lambda_0 \int_0^r f(u) du$, i.e., the same as that of the $rM/G/1$ system; therefore, the transition matrix $M$ of the embedded walk $\mathcal{N}$ remains unchanged if we replace the arrival rate $\lambda_0 f(R_t)$ with $\lambda_0 f(E_t)$. This implies that all of our computations concerning $\mathcal{N}$ above and in Section 3 below remain unchanged if the arrival rate process is changed from $\lambda_0 f(R_t)$ to $\lambda_0 f(E_t)$.

# 3 Stationary distribution at service completions or arrival times

A measure $q$ is the stationary measure of $\mathcal{N}$ if and only if it satisfies

$$q = qM. \tag{12}$$

We have seen in Corollary 1 that the dynamics of the $rM/G/1$ system at service completion times is identical to that of the $M/G/1$ system with constant arrival $\lambda_0$ and service times $\{\bar{\sigma}_i = F(\sigma_i)\}$, therefore (12) is also the balance equation of this $M/G/1$ system at its service completion times. The well known solution of this system is (see, e.g., [16, page 238] or [2, page 281])

$$q(j) = q(0)\bar{p}(j-1) + \sum_{i=1}^{j-1} q(i)\bar{p}(j-i), j = 1, 2, 3, ... \tag{13}$$

where $\bar{p}(j) \doteq \sum_{i=j+1}^{\infty} p(j)$. In particular, a (possibly degenerate) invariant distribution always exists and is uniquely defined as soon as $q(0)$ is fixed. By definition $q$ is nondegenarate if

and only if $\sum_{i=1}^{\infty} q(i) < \infty$, i.e., if $q$ is a finite measure on $\mathbb{N}$. [16, Proposition 10.3.1, page 239] gives precisely the condition for this to hold:

**Proposition 3.** $q$ of (13) defines a finite measure if and only if $-1 + \sum_n np(n) < 0$, i.e., if

$$\rho \doteq \lambda_0 \bar{\nu} = \lambda_0 \mathbb{E}[\bar{\sigma}_i] = \lambda_0 \mathbb{E}[F(\sigma_i)] = \lambda_0 \int_0^\infty F(r)g(r)dr < 1. \tag{14}$$

Then, under the stability condition (14), $q(0)$ can be chosen so that $\sum_{i=0}^{\infty} q(i) = 1$. and, with this choice, $q$ will be the unique stationary measure of the process $\mathcal{N}$. To determine the value of $q(0)$ for which $q$ is a proper probability measure, following [16, page 239], one sums both sides of (13) to get

$$\sum_{j=1}^{\infty} q(j) = q(0) \frac{\rho}{1 - \rho};$$

then, for $\sum_{i=0}^{\infty} q(i) = 1$ we must have

$$q(0) = 1 - \rho.$$

In the rest of this article, we will take $q(0) = 1 - \rho$ whenever the stability assumption (14) is made. Under these assumptions $q(0)$ is the stationary limit probability of an empty $rM/G/1$ queue right after service completions:

**Proposition 4.** The distribution of $\mathcal{N}_n$ converges in total variation norm to $q$. In particular,

$$\lim_{n \to \infty} P(\mathcal{N}_n = 0) = q(0) = 1 - \rho = 1 - \lambda \mathbb{E}[F(\sigma_1)] = 1 - \lambda_0 \int_0^\infty F(r)g(r)dr.$$

*Proof.* That $q$ is the stationary distribution of $\mathcal{N}$ follows from (12). $\mathcal{N}$ is strongly aperiodic; by [16, Proposition 10.3.1] it is positive when (14) holds. The convergence in total variation norm follows from these and [16, Theorem 13.3.1]. □

By Corollary 1, all results/computations for the $M/G/1$ queue at service completion times hold for the $rM/G/1$ queue. For example, the expected queue length at service completion times, is given by the Pollaczek-Khinchine formula [2, Equation (5.3), page 281]

$$\mathbb{E}_q[\mathcal{N}_1] = \sum_{k=1}^{\infty} kq(k) = \rho + \frac{\lambda_0^2 \mathbb{E}[F(\sigma_1)^2]}{2(1 - \rho)}, \tag{15}$$

11

where the subscript $q$ of $\mathbb{E}$ denotes that the Markov chain $\mathcal{N}$ is run in its stationary distribution, and the moment generating function of the stationary distribution is [2, (5.8), page 283]:

$$\mathbb{E}_q[e^{s\mathcal{N}_1}] = \sum_{k=1}^{\infty} e^{sk}q(k) = \frac{(1-\rho)(1-s)\psi_p(s)}{\psi_p(s) - s} \tag{16}$$

where $\psi_p$ is the moment generating function of the increments of $\mathcal{N}$:

$$\psi_p(s) = \mathbb{E}\left[e^{\lambda_0(1-s)\bar{\sigma}_1}\right] = \mathbb{E}\left[e^{\lambda_0(1-s)F(\sigma_1)}\right].$$

**Stationary distribution at arrival times**    Let $S_n^A$ be the sequence of arrival times to the system. Then $(\mathcal{N}_n^A, \mathcal{R}_n^A) = (N_{S_n^A} - 1, R_{S_n^A})$ is the embedded Markov chain of $X$ representing the state of the system just before arrivals. The fact that the queueing process $X$ changes in increments of 1 and $-1$ exactly at arrival and service completion times imply that, under the stability assumption (14), the process $\mathcal{N}^A$ will also have stationary distribution $q$, the stationary distribution of $\mathcal{N}$. For details of similar arguments we refer the reader to [2, Theorem 4.3, page 278] or [8, Section 5.3].

## 4    Stationary distribution in continuous time

One of the key properties of $M/G/1$ systems is that their stationary queue length distribution at service completion times is equal to the same distribution under their continuous time stationary measure. We will see in Corollary 2 below that the $rM/G/1$ system does not possess this property, hence the continuous time stationary measure and related performance measures (such as the average waiting time) for the $rM/G/1$ queue have to be computed separately. This is the goal of the present section. The following verification argument will give us the stationary distribution of $X$:

1. Derive the balance equation for the stationary distribution,

2. Solve the balance equation,

3. Invoke [5, Proposition (34.7), page 113] to show that the solution is indeed the stationary measure of $X$ (see Proposition 8 below).

12

For a measure $\mu$ on $E$ and $k \in \{1, 2, 3, ...\}$, we say that $\mu$ has density $m$ on $E_k$, if $\mu(A \cap E_k) = \int_0^\infty 1_A((k, r))m(r)dr$, for any measurable $A \subset E$. Define

$$\mathscr{M} \doteq \{\mu : \text{ is a measure on } E \text{ having density } m(k, \cdot) \text{ on } E_k, \ k = 1, 2, 3, ...\}.$$

The balance equation for the stationary distribution is

$$\mathfrak{A}^*(\mu)(x) = 0, x \in E. \tag{17}$$

where $\mathfrak{A}^*$ is the conjugate operator (acting on measures $\mu \in \mathscr{M}$) of the generator operator $\mathfrak{A}$:

$$\mathfrak{A}^*(\mu)(x) = \begin{cases} \frac{d}{dr}m(k, r) + \lambda_0 f(r)(m(k-1, r) - m(k, r)) + m(k+1, 0)g(r), k > 1, r > 0 \\ \frac{d}{dr}m(1, r) + \lambda_0\mu(\mathbf{0})g(r) + m(2, 0)g(r) - \lambda_0 f(r)m(1, r), r > 0 \\ m(1, 0) - \mu(\mathbf{0})\lambda_0. \end{cases} \tag{18}$$

The goal of this section is to show that (up to scaling) there is a unique solution $\mu^*$ to the balance equation (17) and this solution is the stationary measure of the continuous time $rM/G/1$ process $X$.

Keep $\mu^*(\mathbf{0}) > 0$ as a free parameter to be fixed below. The third line of (18) gives

$$\mathfrak{A}^*(\mu^*)(\mathbf{0}) = \mu^*(\mathbf{0})\lambda_0 - m^*(1, 0) = 0$$
$$m^*(1, 0) = \mu^*(\mathbf{0})\lambda_0. \tag{19}$$

The last equality and the second line of (18) imply that (17) reduces to the following equation for $m(1, \cdot)$:

$$\frac{d}{dr}m(1, r) + g(r)(m^*(1, 0) + m^*(2, 0)) - \lambda_0 f(r)m(1, r) = 0, r > 0. \tag{20}$$

The classical linear ODE theory implies that the unique solution of (20) vanishing at $\infty$ is

$$m^*(1, r) = (m^*(1, 0) + m^*(2, 0)) \int_r^\infty g(u)e^{(F(r) - F(u))\lambda_0}du, \tag{21}$$

13

Substituting $r = 0$ gives the following formula for $m^*(2, 0)$:

$$m^*(1, 0) = (m^*(1, 0) + m^*(2, 0)) \, p(0) \tag{22}$$

or

$$m^*(2, 0) \doteq \frac{1 - p(0)}{p(0)} m^*(1, 0) > 0, \tag{23}$$

where,

$$p(0) = \int_0^\infty g(r) e^{-F(r)\lambda_0} dr,$$

is the 0 increment probability of the embedded Markov chain $\mathcal{N}$, given in (10). That $m^*(2, 0) > 0$ implies $m^*(1, \cdot) > 0$. Next derive a second expression for $m^*(2, 0)$ by integrating both sides of (21) over $[0, \infty)$:

$$\int_0^\infty m^*(1, r) f(r) dr = (m^*(1, 0) + m^*(2, 0)) \int_0^\infty f(r) \int_r^\infty g(u) e^{(F(r) - F(u))\lambda_0} du dr$$
$$= (m^*(1, 0) + m^*(2, 0)) \frac{1}{\lambda_0} (1 - p(0)),$$

where we have used Fubini's theorem, $m(1, \cdot) > 0$ and the change of variable $s = F(r)$. The definition (23) of $m^*(2, 0)$ implies $m^*(1, 0) = \frac{p(0)}{1 - p(0)} m^*(2, 0)$; substituting this in the last line above gives

$$m^*(2, 0) = \lambda_0 \int_0^\infty m^*(1, r) f(r) dr \tag{24}$$

Formulas (19), (21) and (23) uniquely determine $m^*(1, \cdot)$ and $m^*(2, 0)$ given $\mu^*(\mathbf{0})$. For $k > 1$, (17) uses the first line of (18):

$$\frac{d}{dr} m(k, s) + \lambda_0 f(r)(m(k - 1, s) - m(k, s)) + m(k + 1, 0) g(r) = 0, r > 0. \tag{25}$$

The unique solution of this linear equation for $k = 2$ decaying at $\infty$ is

$$m^*(2, r) = m^*(3, 0) \int_r^\infty g(u) e^{(F(r) - F(u))\lambda_0} du + \lambda_0 \int_r^\infty e^{(F(r) - F(u))\lambda_0} m^*(1, u) f(u) du, \tag{26}$$

14

where $m^*(3,0)$ is yet to be determined. To determine it set $r = 0$ in the above display to get

$$m^*(3,0) = \frac{1}{p(0)} \left( m^*(2,0) - \lambda_0 \int_0^\infty e^{-F(u)\lambda_0} m^*(1,u) f(u) du \right); \qquad (27)$$

With this, $m^*(2,\cdot)$ and $m^*(3,0)$ are determined uniquely, given $\mu(\mathbf{0})$. (24) and the definition of $m^*(3,0)$ imply $m^*(3,0) > 0$, which in its turn implies $m^*(2,\cdot) > 0$.

Letting $r \to \infty$ in (26) gives $\lim_{r\to\infty} m^*(2,r) = 0$. This and the integration of (25) on $[0,\infty)$ gives

$$-m^*(2,0) + \lambda_0 \int_0^\infty m^*(1,r) f(r) dr - \lambda_0 \int_0^\infty m^*(2,r) f(r) dr + m^*(3,0) = 0.$$

This and (24) now imply a similar equation for $m^*(3,0)$:

$$m^*(3,0) = \lambda_0 \int_0^\infty m^*(2,r) f(r) dr. \qquad (28)$$

For $k > 2$, one solves (25) inductively, using $k = 2$ as the base case to get the following sequence of unique positive solutions of (25) vanishing at $\infty$:

$$m^*(k+1,0) \doteq \frac{1}{p(0)} \left( m^*(k,0) - \lambda_0 \int_0^\infty e^{-F(u)\lambda_0} m^*(k-1,u) f(u) du \right) \qquad (29)$$

$$m^*(k,r) \doteq m^*(k+1,0) \int_r^\infty g(u) e^{(F(r)-F(u))\lambda_0} du \qquad (30)$$
$$+ \lambda_0 \int_r^\infty e^{(F(r)-F(u))\lambda_0} m^*(k-1,u) f(u) du,$$

$r > 0$, and the solution satisfies

$$m^*(k+1,0) = \lambda_0 \int_0^\infty f(r) m^*(k,r) dr.$$

The last formulas are the extension of (26) and (27) to $k > 2$. Let us note the foregoing computations as a proposition:

**Proposition 5.** *Given $\mu^*(\mathbf{0}) > 0$, the balance equation* (17) *has a unique positive solution*

$\mu^*$ given by (19), (21), (23) *for $k = 1$ and* (26), *recursively, for $k \geq 2$. The solution satisfies*

$$m^*(k+1, 0) = \lambda_0 \int_0^\infty m^*(k, r) f(r) dr \qquad (31)$$

*for $k \geq 1$.*

The next proposition links the quantities $m^*(k, 0)$ to the stationary distribution of the embedded chain $\mathcal{N}$:

**Proposition 6.** *Let $\mu^* = (\mu^*(0), m^*(k, \cdot), k = 1, 2, 3, ...)$ be the unique solution (up to the choice of $\mu^*(0) > 0$) of the balance equation (17) derived in Proposition 5 above. Then the measure*

$$m^* \doteq (m^*(1, 0), m^*(2, 0), m^*(3, 0), \cdots)$$

*on $\mathbb{N}_+$ is $M$-invariant, i.e.,*

$$m^* M = m^* \qquad (32)$$

*and*

$$m^* = cq \qquad (33)$$

*for some $c > 0$ where $q$ is the stationary measure given in (13). In particular,*

$$\sum_{k=1}^\infty m^*(k, 0) < \infty \qquad (34)$$

*if the stability assumption (14) holds.*

*Proof.* By definition (9) of the matrix $M$ (32) is the following sequence of equations:

$$p(n)(m^*(1, 0) + m^*(2, 0)) + \sum_{k=2}^{n+1} m^*(k+1, 0) p(n+1-k) = m^*(n+1, 0), \qquad (35)$$

$n = 0, 1, 2, 3, ...$ For $n = 0$, (35) reduces to (22), which holds by definition. To prove (35) for $n > 0$, multiply both sides of (20) by $e^{-F(r)\lambda_0} \frac{(F(r)\lambda_0)^n}{n!}$, integrate from 0 to $\infty$ to get

$$0 = \int_0^\infty \frac{d}{dr} m^*(1, r) e^{-F(r)\lambda_0} \frac{(F(r)\lambda_0)^n}{n!} dr + (m^*(1, 0) + m^*(2, 0)) p(n)$$
$$- \lambda_0 \int_0^\infty m^*(1, r) e^{-F(r)\lambda_0} \frac{(F(r)\lambda_0)^n}{n!} f(r) dr.$$

Integration by parts on the first integral gives:

$$0 = (m^*(1,0) + m^*(2,0))p(n) - \lambda_0 \int_0^\infty e^{-F(r)\lambda_0} \frac{(F(r)\lambda_0)^{n-1}}{(n-1)!} m^*(1,r)f(r)dr. \qquad (36)$$

For $k = 2, ..., n+1$, multiply both sides of (25) by $e^{-F(r)\lambda_0} \frac{(F(r)\lambda_0)^{n+1-k}}{(n+1-k)!}$ and integrate by parts the first term to get

$$m^*(k+1,0)p(n+1-k) + \lambda_0 \int_0^\infty e^{-F(r)\lambda_0} \frac{(F(r)\lambda_0)^{n+1-k}}{(n+1-k)!} m^*(k-1,r)f(r)dr \qquad (37)$$

$$- \lambda_0 \int_0^\infty e^{-F(r)\lambda_0} \frac{(F(r)\lambda_0)^{n-k}}{(n+1-k)!} m^*(k,r)f(r)dr = 0.$$

Summing the last display over $k$, adding to the result (36) and finally noting (29) give (35) for $n > 0$. The Markov chain $\mathcal{N}$ is a constrained random walk on $\mathbb{Z}_+$ with iid increments and hence is obviously irreducible and will therefore have (up to scaling) a unique stationary distribution; (33) follows from this. (34) follows from (33) and Proposition 3. $\square$

Define
$$S(r) \doteq \sum_{k=1}^\infty m^*(k,r),$$

whose finiteness under the stability assumption follows from (31) and the previous proposition (see (34)); (31) also implies

$$\lambda_0 \int_0^\infty S(r)f(r)dr = S(0) - m^*(1,0) = S(0) - \lambda_0\mu^*(\mathbf{0}). \qquad (38)$$

Remember that $\mu^*(\mathbf{0})$ is still a free parameter. The next proposition computes $\int_0^\infty S(r)dr$ and $S(0)$ in terms of $\mu^*(\mathbf{0})$ and in terms of the system parameters.

**Proposition 7.** *Suppose the stability assumption* (14) *holds. Then*

$$S(0) = \lambda_0 \frac{\mu^*(\mathbf{0})}{1 - \rho}. \qquad (39)$$

*and*

$$\int_0^\infty S(r)dr = S(0)\nu. \qquad (40)$$

17

*Proof.* Summing the terms of the balance equation gives $S'(r) = -S(0)g(r)$, therefore,

$$S(r) = S(0)G(r), \tag{41}$$

where

$$G(r) \doteq P(\sigma_1 > r) = \int_r^\infty g(u)du. \tag{42}$$

Integrating both sides of (41) over $[0, \infty)$ gives (40). Next multiply both sides by $f(r)$ and integrate over $[0, \infty]$:

$$\int_0^\infty S(r)f(r)dr = S(0) \int_0^\infty f(r)G(r)dr = S(0)\bar{\nu},$$

where we have integrated by parts the middle integral. The last display and (38) imply

$$\lambda_0 S(0)\bar{\nu} = S(0) - \lambda_0 \mu^*(\mathbf{0})$$
$$S(0) = \lambda_0 \mu^*(\mathbf{0})\frac{1}{1-\rho},$$

which proves (39). $\qquad\qquad\square$

Let us fix the value for $\mu^*(\mathbf{0})$ to

$$\mu^*(\mathbf{0}) = 1 - \rho = q(0); \tag{43}$$

we will assume (43) whenever the stability assumption (14) is made. This implies by (39) and (41):

$$S(0) = \lambda_0, S(r) = \lambda_0 G(r). \tag{44}$$

A second implication is given in the next lemma.

**Lemma 1.** *Let $\mu^*(\mathbf{0})$ be fixed as in (43), i.e., we take $\mu^*(\mathbf{0}) = q(0)$. Then*

$$m^* = \lambda_0 q \tag{45}$$

*where $m^* = (m^*(1,0), m^*(2,0), ...)$ is as in Proposition 6. In particular,*

$$\sum_{k=1}^\infty km^*(k+1,0) = \lambda_0 \sum_{k=1}^\infty kq(k) = \lambda_0 \mathbb{E}_q[\mathcal{N}_1]. \tag{46}$$

18

*Proof.* We know by (33) that $m^* = cq$ for some $c > 0$. Because $q$ is a probability measure, this implies, $c = \sum_{k=1}^{\infty} m^*(k,0) = S(0)$, which equals $\lambda_0$, by (44). This proves (45); (46) follows from (45). □

With $\mu^*(\mathbf{0})$ fixed as in (43), the measure $\mu^*$ is determined uniquely via Proposition 5. Note

$$\mu^*(E) = \mu^*(\mathbf{0}) + \int_0^{\infty} S(r)dr = 1 - \rho + \lambda_0\nu. \tag{47}$$

where we have used (40), (39) and (43). Thus, in general, with $\mu^*(\mathbf{0})$ fixed as in (43), $\mu^*(E) \neq 1$- to get a proper probability measure, renormalize $\mu^*$:

$$\mu_1^* \doteq \mu^*/\mu^*(E).$$

Proposition 8 below proves that $\mu_1^*$ is the unique stationary measure of the $rM/G/1$ process $X$ under the stability assumption (14). The proof will require a subclass of functions in $\mathscr{D}(\mathfrak{A})$ that can separate measures in $\mathscr{M}$. The following lemma identifies such a class.

**Lemma 2.**

$$\mathscr{S} \doteq \{h \in \mathscr{D}(\mathfrak{A}), \sup_{x \in E} |h'(x)| < \infty, \sup_{x \in E} |h(x)| < \infty\}$$

*is a separating class of functions for measures in $\mathscr{M}$.*

*Proof.* For $\mu_1, \mu_2 \in \mathscr{M}$, $\mu_1 = \mu_2$ if and only if

$$\int_a^b m_{1,k}(r)dr = \int_a^b m_{2,k}(r)dr$$

for all $0 < a < b < \infty$ and $k > 0$ ($m_{1,k}$ and $m_{2,k}$ are densities of $\mu_1$ and $\mu_2$ on $E_k$). Define the standard mollifier

$$\eta(x) \doteq \begin{cases} C_\eta e^{\frac{1}{|x|^1 - 1}}, & |x| < 1 \\ 0, & |x| > 1, \end{cases}$$

where $C_\eta > 0$ is such that $\int_{-1}^1 \eta(x)dx = 1$. For any interval $(a,b)$, $0 < a < b$ define $h_n : E \to [0,1]$, $1/m < 1/2a$ as follows: for $x = (j,r) \in E$, $j < k$ $h_n(x) = 0$. For $x = (k,r)$

$$h_n(x) = n \int_{-1}^1 \eta(u/n)\mathbb{1}_{(a,b)}(u+r)dr, j = k. \tag{48}$$

For $j > k$ we proceed recursively:

$$h_n(j,0) = \int_0^\infty h_n(j-1,r)g(r)dr, \quad h_n(j,r) = h_n(j,0)n \int_{nr-1}^1 \eta(x/n)dx, \qquad (49)$$

where we write $h_n(j,0)$ instead of $h_n((j,0))$ to simplify notation. By its definition, $h_n \in \mathscr{S}$ and $\lim_{n\to\infty} h_n(k,r) = \mathbb{1}_{\{(k,r),r\in(a,b))\}}$ almost surely for any measure $\mu \in \mathscr{M}$, this and the bounded convergence theorem imply

$$\lim_{n\to\infty} \int_E h_n(x)\mu(dx) = \int_a^b m(k,r)dr.$$

This proves that functions of the form $h_n$ and (therefore the class $\mathscr{S}$ containing them) is a separating class for measures in $\mathscr{M}$.

It remains to show that $h_n \in \mathscr{D}(\mathfrak{A})$. The following three conditions for this are listed in Subsection 2.1: 1) $h_n$ must be absolutely continuous, this follows from its definition (48) and (49) 2) $h_n$ must satisfy $h_n((j,0)) = \int_0^\infty h_n(j-1,r)g(r)dr$; this again holds by definition and 3) $\mathfrak{B}h_n \in L_1^{loc}(X)$; this follows from the fact that $h_n$ is bounded. $\qquad \square$

**Proposition 8.** *If the stability assumption (14) holds then $\mu_1^*$ is the unique stationary measure of the process $X$. In particular if $X_0$ has distribution $\mu_1^*$, $X_t$ has the same distribution for all $t > 0$.*

*Proof.* The uniqueness follows from the uniqueness claim of Proposition 5. By [5, Proposition (34.7), page 113], $\mu_1^*$ is the stationary distribution of $X$ if

$$\int \mathfrak{A}h(x)\mu_1^*(dx) = 0; \qquad (50)$$

for a class of functions $h \in \mathscr{D}(\mathfrak{A})$ that forms a separating class for measures in $\mathscr{M}$ to which $\mu_1^*$ belongs; by Lemma 2 $\mathscr{S} \subset \mathscr{D}(\mathfrak{A})$ is such a class. Thus, to prove the proposition it suffices to prove (50) for $h \in \mathscr{S}$. By definition,

$$\int \mathfrak{A}h(x)\mu_1^*(dx) = \frac{1}{\mu^*(E)} \int \mathfrak{A}h(x)\mu^*(dx),$$

and one can directly work with the measure $\mu^*$ rather than the normalized $\mu_1^*$. For any $h \in \mathscr{S}$

$$\int_E \mathfrak{A}h(x)\mu^*(dx) = \lim_{N \to \infty} \sum_{k=1}^N \int_0^\infty \left( -\frac{dh}{dr}(k,r) + \lambda_0 f(r)(h(k+1,r) - h(k,r)) \right) m^*(k,r)dr$$
(51)

We begin by an integration by parts:

$$\sum_{k=1}^N \int_0^\infty \left( -\frac{dh}{dr}(k,r) + \lambda_0 f(r)(h(k+1,r) - h(k,r)) \right) m^*(k,r)dr$$

$$= \sum_{k=1}^N \int_0^\infty \left( \frac{dm^*}{dr}(k,r) - h(k,0)m^*(k,0) + \lambda_0 f(r)(h(k+1,r) - h(k,r)) \right) m^*(k,r)dr$$

$h(k,0) = \int_0^\infty g(r)h(k-1,r)dr$ because $h \in \mathscr{D}(\mathfrak{A})$, therefore,

$$= \sum_{k=1}^N \int_0^\infty \left( \frac{dm^*}{dr}(k,r) - g(r)h(k-1,r)m^*(k,0) + \lambda_0 f(r)(h(k+1,r) - h(k,r)) \right) m^*(k,r)dr$$

Rearrange the terms in the sum to factor out the common $h(k,r)$:

$$= \sum_{k=1}^N \int_0^\infty \left( \frac{dm^*}{dr}(k,r) - g(r)m^*(k+1,0) + \lambda_0 f(r)(m^*(k-1,r) - m^*(k,r)) \right) h(k,r)dr$$

$$+ \int_0^\infty m^*(N,r)f(r)\lambda_0 h(N+1,r)dr$$

Now $\mathfrak{A}^*\mu^* = 0$ implies

$$= \int m^*(N,r)f(r)\lambda_0 h(N+1,r)dr;$$

the last integral goes to 0 with $N$ because $\int m^*(N,r)f(r)dr \to 0$ and $h$ is bounded. Therefore, the limit on the right side of (51) is 0. This proves (50) and establishes that $\mu^*$ is the unique stationary distribution of the process $X$. □

The last proposition and Proposition 8 give

**Corollary 2.** *The stationary probability of an empty system in continuous time for a stable* $rM/G/1$ *queue is*

$$\mu_1^*(\mathbf{0}) = \frac{\mu^*(\mathbf{0})}{\mu^*(E)} = \frac{1 - \rho}{1 - \rho + \lambda_0 \nu}.$$

### 4.1 Expected queue length

As Corollary 2 demonstrates, the probability of a stable $rM/G/1$ being empty under its continuous time stationary distribution does not in general equal the same probability under its stationary distribution at service completion or arrival times:

$$P_{\mu_1^*}(N_t = 0) = \mu_1^*(\mathbf{0}) = \frac{1 - \rho}{1 - \rho + \lambda_0 \nu} \neq q(0) = 1 - \rho.$$

Thus, in general, the steady state queue length distribution of a stable $rM/G/1$ system in continuous time differs from the same distribution at service completion and arrival times. The following proposition gives a formula for $\mathbb{E}_{\mu_1^*}[N_1]$, under $\mu_1^*$, the expected queue length under the stationary distribution in continuous time; in general, this quantity will not equal $\mathbb{E}_q[\mathcal{N}_1]$, the expected queue length under the stationary distribution at service completion times.

**Proposition 9.** *The expected $rM/G/1$ queue length under its continuous time stationary measure equals*

$$\mathbb{E}_{\mu^*}[N_t] = \frac{1}{\mu^*(E)} \left( \lambda_0^2 \int_0^\infty \left( \int_0^x u f(u) du \right) g(x) dx + ((1 - \rho) + \mathbb{E}_q[\mathcal{N}_1]) \lambda_0 \nu \right) \tag{52}$$

*where $\mathbb{E}_q[\mathcal{N}_1]$ is the stationary mean queue length at service completion times whose formula is given in* (15).

*Proof.* The proof proceeds parallel to that of Proposition 7. Set

$$\varphi(r) \doteq \sum_{k=1}^\infty m_k^*(r) k; \tag{53}$$

by definition

$$\mathbb{E}_{\mu_1^*}[N_t] = \mathbb{E}_{\mu^*}[N_t]/\mu^*(E) = \frac{1}{\mu^*(E)} \int_0^\infty \varphi(r) dr. \tag{54}$$

Let us compute $\int_0^\infty \varphi(r) dr$. Multiply the first and the second lines of the balance equation (17) by $k$, $k = 1, 2, 3, 4, ...$ and sum over $k$ to get

$$\frac{d\varphi}{dr}(r) + \lambda_0 f(r) S(r) + \left( \mu^*(\mathbf{0}) \lambda_0 + \sum_{k=1}^\infty m_{k+1}^*(0) k \right) g(r) = 0, \tag{55}$$

22

where, as before,

$$S(r) = \sum_{k=1}^{\infty} m_k^*(r) = S(r) = \lambda_0 G(r);$$

the last equality follows from (44). (43) and (46) simplify the terms in paranthesis in (55) to

$$\left( \mu^*(\mathbf{0})\lambda_0 + \sum_{k=1}^{\infty} m_{k+1}^*(0)k \right) = \lambda_0 \left( (1-\rho) + \mathbb{E}_q[\mathcal{N}_1] \right),$$

where $\mathbb{E}_q[\mathcal{N}_1]$ is the stationary mean queue length at service completions. Then, the unique solution of (55) vanishing at $\infty$ is

$$\varphi(r) = \lambda_0^2 \int_r^{\infty} f(u)G(u)du + \left( (1-\rho) + \mathbb{E}_q[\mathcal{N}_1] \right) \lambda_0 G(r).$$

Integrating the last display over $r$ over $[0, \infty]$ yields

$$\int_0^{\infty} \varphi(r)dr = \lambda_0^2 \int_0^{\infty} \int_r^{\infty} f(u)G(u)dudr + \left( (1-\rho) + \mathbb{E}_q[\mathcal{N}_1] \right) \lambda_0 \nu$$

$$= \lambda_0^2 \int_0^{\infty} \left( \int_0^x u f(u)du \right) g(x)dx + \left( (1-\rho) + \mathbb{E}_q[\mathcal{N}_1] \right) \lambda_0 \nu.$$

This and (54) give (52). □

## 4.2 Average arrival rate

The random variable

$$\mathcal{A}_n \doteq \mathcal{N}_{n+1} - \mathcal{N}_n + 1,$$

represents the number of arrivals to the $rM/G/1$ system between the $n^{th}$ and $(n+1)^{st}$ service completions. It follows from (58) that its conditional distribution given in $\mathcal{N}_n$ is

$$P(\mathcal{A}_n = j | \mathcal{N}_n) = \begin{cases} p(j), & \mathcal{N}_n > 0 \\ p(j+1), & \mathcal{N}_n = 0, \end{cases} \tag{56}$$

It follows from the Markov property of $\mathcal{N}$ that $(\mathcal{N}, \mathcal{A})$ is a Markov chain and is stationary whenever $\mathcal{N}$ is, with stationary distribution

$$P(\mathcal{A}_\infty = j, \mathcal{N}_\infty = k) = \begin{cases} p(j)q(k), & k > 0 \\ p(j+1)q(0), & k = 0. \end{cases}$$

23

Then by the ergodic theorem for stable Markov chains

$$\lim_{n\to\infty} \frac{1}{n}\sum_{n=1}^{\infty} \mathcal{A}_n = q(0)\left(1 + \sum_{j=1}^{\infty} jp(j)\right) + (1 - q(0))\left(\sum_{j=1}^{\infty} jp(j)\right)$$

$$= q(0) + \rho = 1 - \rho + \rho = 1. \tag{57}$$

Define the interservice time

$$\tau_n \doteq S_{n+1} - S_n;$$

similar to the sequence $\mathcal{A}_n$ the distribution of $\tau_n$ is completely determined by $\mathcal{N}$ with the following conditional distribution:

$$P(\tau_n > t | \mathcal{N}_n) = \begin{cases} G(t), & \mathcal{N}_n > 0 \\ \int_0^\infty \lambda_0 e^{-\lambda_0 s} G((t-s)^+)ds, & \mathcal{N}_n = 0, \end{cases} \tag{58}$$

where the second distribution is the convolution of $g$ and the exponential distribution with rate $\lambda_0$ (this is the distribution of the sum of a service time and the first arrival time to the system). The process $(\mathcal{N}, \tau)$ is stable whenever $\mathcal{N}$ is, with the stationary distribution

$$P(\tau_\infty > t, \mathcal{N}_\infty = k) = \begin{cases} G(t)q(k), & k > 0, \\ \left(\int_0^\infty \lambda_0 e^{-\lambda_0 s} G((t-s)^+)ds\right) q(0), & k = 0. \end{cases}$$

The law of large numbers for Markov chains [16, Theorem 17.0.1, page 422] implies

$$\lim_{n\to\infty} \frac{1}{n}\sum_{n=1}^{\infty} \tau_n = \lim_{n\to\infty} \frac{1}{n} S_n = q(0)\left(\frac{1}{\lambda_0} + \nu\right) + (1 - q(0))\nu$$

$$= q(0)\frac{1}{\lambda_0} + \nu = (1 - \rho)\frac{1}{\lambda_0} + \nu. \tag{59}$$

**Proposition 10.** *Let $A_t$ denote the number of arrivals to an $rM/G/1$ queue up to time $t$. Then the ergodic average arrival rate to the $rM/G/1$ system equals*

$$\lim_{t\to\infty} \frac{A(t)}{t} = \alpha \doteq \frac{\lambda_0}{1 - \rho + \lambda_0\nu} = \frac{\lambda_0}{\mu^*(E)}. \tag{60}$$

*Proof.*

$$\lim_{n \to \infty} \frac{A_{T_n}}{T_n} = \frac{A_{T_n}/n}{T_n/n} = \frac{\lambda_0}{1 - \rho + \lambda_0 \nu} \tag{61}$$

follows from (57) and (59). For any other sequence $t_m \nearrow \infty$ we know that there exists a sequence $n_m$ with $T_{n_m} < t_m < T_{n_m+1}$. Borel Cantelli Lemma and that $\mathcal{A}_n$ has finite moments independent of $n$ imply

$$\lim_{n \to} \frac{\mathcal{A}_n}{T_n} = 0. \tag{62}$$

It follows from the monotonicity of $T_n$ and $A_n$ that

$$\frac{A_{T_{n_m}}}{T_{n_m+1}} \leq \frac{A_{t_m}}{t_m} \leq \frac{A_{T_{n_m+1}}}{T_{n_m}}.$$

This, (62) and (61) imply (60). $\qquad\square$

## 4.3   Average sojourn and waiting time

Let $\varsigma_k$ be the sojourn time (the total amount of time spent) of the $k^{th}$ customer arriving to the system. Little's law is the following statement

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \varsigma_k = \frac{\lim_{t \to} N_t/t}{\lim_{t \to \infty} A_t/t}. \tag{63}$$

The classical proof of this result outlined in [14] depends on the distribution of $X$ only to the following extent: that $N$ represents the number of customers in a single server queueing system and that the ergodic limits related to $N$ and $A$; the existence of the ergodic limits follow from the stationarity of $N$ (see, e.g., [3, Theorem 1.6.4, page 50]) and Proposition 10 above. Therefore, the classical proof requires no change for the current setup. For the $rM/G/1$ system, (63) and Proposition 10 give the following formula for the average sojourn time:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \varsigma_k = \frac{\mathbb{E}_{\mu_1^*}[N_t]}{\alpha}$$

$$= \lambda_0 \int_0^\infty \left( \int_0^x u f(u) du \right) g(x) dx + ((1 - \rho) + \mathbb{E}_q[\mathcal{N}_1]) \nu. \tag{64}$$

This gives the following formula for the average waiting time

$$\omega \doteq \lambda_0 \int_0^\infty \left( \int_0^x u f(u) du \right) g(x) dx + ((1 - \rho) + \mathbb{E}_q[\mathcal{N}_1]) \, \nu - \nu \qquad (65)$$

$$= \lambda_0 \int_0^\infty \left( \int_0^x u f(u) du \right) g(x) dx + (\mathbb{E}_q[\mathcal{N}_1] - \rho) \, \nu,$$

where $\mathbb{E}_q[\mathcal{N}_1]$ can be computed with formula (15).

## 4.4 Connection to $M/G/1$ queue in continuous time

By Corollary 1 we know that the embedded random walk $\mathcal{N}$ at service completions of the $rM/G/1$ queue has identical dynamics to that of an $M/G/1$ queue with constant rate $\lambda_0$ and sequence of service times $\{F(\sigma_1), F(\sigma_2), ....\}$- which implies that that these systems have the same stationary measures at service completions. Then a natural question is whether there is a similar correspondence between the continuous time stationary distributions. When $f$ takes the value 0 over a nonzero interval its integral $F$ becomes not-invertible. Because of this, in general, the continuous time stationary distribution of the $M/G/1$ system cannot completely be mapped to that of the $rM/G/1$ system (remember that $\sigma_i$ has density $g$; when $f = 0$ is allowed $F(\sigma_i)$ may have no density and $F(\sigma_i)$ may have compact support even when $\sigma$ takes values in all of $\mathbb{R}_+$). However, for $f > 0$ an exact mapping between the stationary measures is possible; the details follow.

Assuming $f > 0$ implies $F(r) = \int_0^r f(u) du$ is strictly increasing. Let $H$ denote its inverse function; that $F$ is differentiable implies the same for $H$ and the inverse function has the derivative

$$\frac{dH}{ds}(s) = \frac{1}{f(H(s))}. \qquad (66)$$

Define

$$\bar{g}(s) \doteq g(H(s)) \frac{dH}{ds}(s) = \frac{g(H(s))}{f(H(s))}, s > 0.$$

For $f > 0$, the change of variable formula of calculus implies that $F(\sigma_i)$ has density $\bar{g}$. The same formula allows one to rewrite the operator $\mathfrak{A}^*$ defining the balance equations of the

$rM/G/1$ system in the $s = F(r)$ variable thus:

$$\bar{\mathfrak{A}}^*(\mu)(x) = \begin{cases} \frac{d}{ds}m(k,s) + \lambda_0(m(k-1,s) - m(k,s)) + m(k+1,0)\bar{g}(s), k > 1, s > 0, \\ \frac{d}{ds}m(1,s) + \bar{g}(s)\lambda_0\mu(\mathbf{0}) + \bar{g}(s)m(2,0) - \lambda_0 m(1,s), s > 0, \\ \mu(\mathbf{0})\lambda_0 - m(1,0), \end{cases} \quad (67)$$

$\mu \in \mathcal{M}$. The equation

$$\bar{\mathfrak{A}}^*(\mu) = 0 \quad (68)$$

is the balance equation of the $M/G/1$ system with rate $\lambda_0$ and service density $\bar{g}$. Let us denote the continuous time process representing this $M/G/1$ system by $\bar{X}$ (which can be written in the PDP framework employed in Section 2). The relation between the solution of (68) and the solution of the balance equation (17) is given in the following proposition.

**Proposition 11.** *Assume $f > 0$. Let $\mu^*$ be the solution of (17) given in Proposition 5. Then $\bar{\mu}^* \in \mathcal{M}$ defined by $\bar{\mu}^*(\mathbf{0}) \doteq \bar{\mu}^*(\mathbf{0})$ and by the densities $\bar{m}^*(k,s) \doteq m^*(k, H(s))$ on $E_k$, $k = 1, 2, 3, ...$ solves (68) and does so uniquely up to the choice of $\bar{\mu}^*(\mathbf{0})$. Furthermore, if the stability condition 14 holds and $\bar{\mu}^*(\mathbf{0})$ is set to $1 - \rho$ we have $\bar{\mu}^*(E) = 1$ and $\bar{\mu}^*$ is the unique continuous time stationary measure of $\bar{X}$.*

*Proof.* $\mathfrak{A}^*(\mu^*) = 0 \Rightarrow \bar{\mathfrak{A}}^*(\bar{\mu}^*) = 0$ follows from the chain rule. The uniqueness claim follows from the linearity of (68). That $\bar{\mu}^*(E) = 1$ under the assumptions (14) and $\bar{\mu}^*(\mathbf{0}) = 1 - \rho$ follows from the following observation:

$$\bar{\mu}^*(E_k) = \int_0^\infty \bar{m}^*(k,s)ds = \int_0^\infty m^*(k,r)f(r)dr = q(k), k > 0;$$

the first equality follows from the change of variable $r = H(s)$ and the last equality follows from (31) and (45). That $\bar{\mu}^*$ is the stationary measure of $\bar{X}$ is proved exactly as in the proof of Proposition 8. $\qquad\square$

## 5 Illustration

Let us now observe the consequences of the results derived in the previous section over two examples. Figure 2a shows the average waiting times for three $rM/G/1$ systems with uniformly distributed service time on the interval $[0,1]$, as a function of the arrival rate $\lambda_0$.
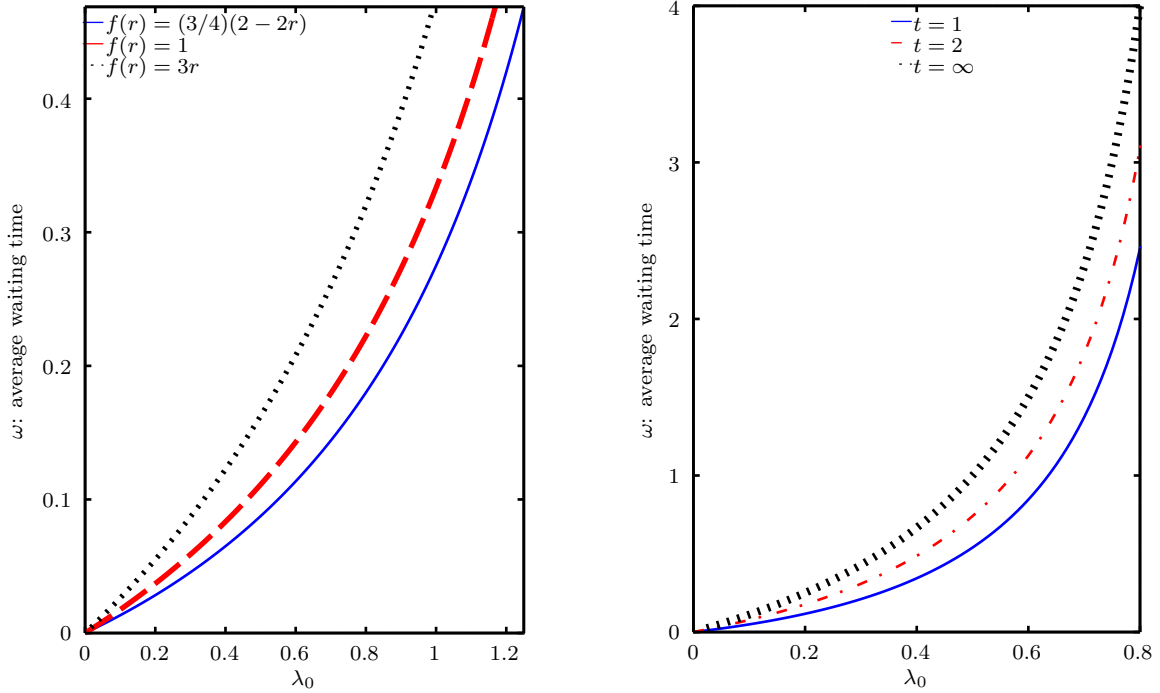
(a) $g(r) = \mathbb{1}_{(0,1)}(r)$          (b) $g(r) = e^{-r}$, $f(r) = \frac{1}{1-e^{-t}}\mathbb{1}_{(0,t)}(r)$

Figure 2: Average waiting time as a function of the average arrival rate for different reshaping functions

We consider three cases for the reshaping function $f$; increasing, constant and decreasing in $r$: $f(r) = \frac{3}{4}(2 - 2r)\mathbb{1}_{(0,1)}(r)$, $f(r) = \mathbb{1}_{(0,1)}(r)$ and $f(r) = 3r\mathbb{1}_{(0,1)}(r)$ (the constant case corresponds to the $M/U/1$ queue). All of these reshape functions $f$ satisfy (1), therefore they all have the same average arrival rate $\alpha = \lambda_0$, utilization $\rho = \lambda_0/2$ and empty system probability $\mu^*(\mathbf{0}) = q(0) = 1 - \rho$. Moreover for all of these reshape functions $f$ and the assumed system parameters, the formula (65) for the average waiting time has simple explicit expressions, for $f(r) = \frac{3}{4}(2 - 2r)\mathbb{1}_{(0,1)}(r)$,

$$\omega = \omega_1 = \frac{\lambda_0}{8} + \frac{3\lambda_0^2}{40(1 - \rho)},$$

for $f(r) = \mathbb{1}_{(0,1)}(r)$ (this is the $M/U/1$ case)

$$\omega = \omega_2 = \frac{\lambda_0}{6} + \frac{\lambda_0^2}{12(1 - \rho)}.$$

and for $f(r) = 3r\mathbb{1}_{(0,1)}(r)$,

$$\omega = \omega_3 = \frac{\lambda_0}{4} + \frac{9\lambda_0^2}{80(1-\rho)}.$$

We note $\omega_1 < \omega_2 < \omega_3$ for all $\lambda_0$ such that $\rho < 1$: i.e., pushing arrivals towards service completions (while keeping the average arrival rate constant) reduces average waiting times. Figure 2a shows the graphs of these functions as $\lambda_0$ varies.

Let us now consider an example in which the service time is exponentially distributed mean $\nu = 1$ and the reshaping function $f(r) = (1 - e^{-t})^{-1}\mathbb{1}_{0 \leq r \leq t}$; this function restricts arrivals to the last $t$ units of time of service and it satisfies (1). The average waiting time $\omega$ of (65) reduces for this case to

$$\omega(t, \lambda_0) = \lambda_0 \frac{1 - e^{-t}(t+1)}{1 - e^{-t}}\left(1 + \frac{\lambda_0^2}{(1 - e^{-t})(1 - \rho)}\right);$$

$\omega$ is increasing in $t$, i.e., once again, concentrating arrivals near service completions (while keeping the average arrival rate constant) reduces the average waiting time. Figure 2b, shows the graph of $\omega(t, \cdot)$ for $t = 1, 2$ and $t = \infty$ (the last corresponds to an $M/M/1$ queue).

We conclude this section with the following observation from our second example: set $t = 3\nu = 3$ in the last example, i.e., we restrict arrivals to the interval $[0, 3\nu]$, where $\nu$ is the mean service time. For $\lambda_0 = 0.7$, the system's utilization is $\rho = 0.7$ and the corresponding average waiting time turns out to be $\omega(3, 0.7) = 1.6041$; the same waiting time for the same parameter values but without reshaping is $\omega(\infty, 0.7) = 1.84$. Thus, this not so heavy reshaping reduces average waiting time by 13%.

## 6    Conclusion

Let us comment briefly on possible future research. We have assumed that the service time distribution has a density $g$. The analysis at service completions doesn't depend on this assumption and the results of Subsection 2.2 and Section 3 continue to hold without change when $\sigma_i$ doesn't have a density. The analysis of Section 4 does make use of the assumption that $\sigma_i$ has a density but the resulting performance measure formulas (average queue length, probability of an empty system, average waiting time) remain meaningful even when $\sigma_i$ doesn't have a density and one expects these results to hold under general service distributions. One simple method of extending our analysis to the general case would be first

a smooth approximation of the given service distribution and then taking weak limits. The details of such an argument could be given in future work. The special case of a deterministic constant service time case can be directly handled by appropriate modifications of the balance equation and our arguments based on it.

A natural question is the convergence of the distribution of $X_t$ to the stationary distribution $\mu^*$. As one of the referees pointed out, one way to establish this with precise rates of convergence would be to apply the approach of [15] based on coupling (at the first hitting time to $\mathbf{0}$) and monotonicity arguments. Future research could attempt to give details of this.

In many situations, one may only have an estimate of the remaining service time (rather than the ability to directly observe it, as assumed in the current work). One possible future work is the modeling and analysis of such a setup. We think that, given the possible applications in call centers, another natural direction is the treatment of many servers. Instead of allowing the rate to depend directly on the remaining service times of all of the servers a possibility is to allow it to depend on a function of them (e.g., their minimum or an estimate of it). Finally, it may also be of interest to apply the approach used in the present article to models where the arrival and service rates depend on the queue length as well as the remaining service time.

# References

[1] O.Z. Akşin, M. Armony, and V. Mehrotra, *The Modern Call-Center: A Multi-Disciplinary Perspective on Operations Management Research*, Production and Operations Management **16** (2007), 665–688.

[2] Soren Asmussen, *Applied probability and queues*, vol. 51, Springer Science & Business Media, 2008.

[3] François Baccelli and Pierre Brémaud, *Palm probabilities and stationary queues*, vol. 41, Springer Science & Business Media, 2012.

[4] René Bekker, Sem C Borst, Onno J Boxma, and Offer Kella, *Queues with workload-dependent arrival and service rates*, Queueing Systems **46** (2004), no. 3-4, 537–556.

[5] Mark HA Davis, *Markov models & optimization*, vol. 49, CRC Press, 1993.

[6] Jewgeni Dshalalow, *On single-server closed queues with priorities and state dependent parameters*, Queueing Systems **8** (1991), no. 1, 237–253.

[7] N. Gans, G. Koole, and A. Mandelbaum, *Telephone Call Centers: Tutorial, Review, and Research Prospects*, Manufacturing & Service Operations Management **5** (2003), 73–141.

[8] Leonard Kleinrock, *Queueing Systems, Theory*, vol. I, A Wiley-Interscience Publication, 1975.

[9] Charles Knessl, BJ Matkowsky, Zeev Schuss, and Charles Tier, *Busy period distribution in state-dependent queues*, Queueing Systems **2** (1987), no. 3, 285–305.

[10] _____ , *A markov-modulated m/g/1 queue i: stationary distribution*, Queueing Systems **1** (1987), no. 4, 355–374.

[11] _____ , *A markov-modulated m/g/1 queue ii: Busy period and time for buffer overflow*, Queueing Systems **1** (1987), no. 4, 375–399.

[12] Dinesh Kumar, Li Zhang, and Asser Tantawi, *Enhanced inferencing: Estimation of a workload dependent performance model*, Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2009, p. 47.

[13] Benjamin Legros, Oualid Jouini, and Ger Koole, *Optimal scheduling in call centers with a callback option*, Performance Evaluation **95** (2016), 1–40.

[14] John DC Little and Stephen C Graves, *Little's law*, Building intuition, Springer, 2008, pp. 81–100.

[15] Robert B Lund, Sean P Meyn, Richard L Tweedie, et al., *Computable exponential convergence rates for stochastically ordered markov processes*, The Annals of Applied Probability **6** (1996), no. 1, 218–237.

[16] Sean P Meyn and Richard L Tweedie, *Markov chains and stochastic stability, second edition*, Cambridge University Press, 2012.

[17] Guodong Pang and Ohad Perry, *A logarithmic safety staffing rule for contact centers with call blending*, Management Science **61** (2014), no. 1, 73–91.

[18] D Perry, W Stadje, S Zacks, et al., *A duality approach to queues with service restrictions and storage systems with state-dependent rates*, Journal of Applied Probability **50** (2013), no. 3, 612–631.