An M/G/1 queueing model with gated random order of service

An M/G/1 queueing model with gated random order of service

Ronald Rietman
Philips Research Laboratories
Prof. Holstlaan 4
NL-5656 AA Eindhoven
The Netherlands
e-mail: ronald.rietman@philips.com

Jacques Resing
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: j.a.c.resing@tue.nl

Abstract

We analyse an $M/G/1$ queueing model with gated random order of service. In this service discipline there are a waiting room, in which arriving customers are collected, and a service queue. Each time the service queue becomes empty, all customers in the waiting room are put instantaneously and in random order into the service queue. The service times of customers are generally distributed with finite mean. We give two different derivations of various steady-state probabilities and of the bivariate Laplace–Stieltjes transforms of the joint distribution of the sojourn times in the waiting room and the service queue. The first derivation is based on a three-dimensional Markov model, where the random variables are the number of customers in the waiting room and in the service queue and the residual service time of the customer in service. The second derivation follows the line of reasoning of Avi-Itzhak and Halfin [4], is shorter and more elegant, but less straightforward than the first derivation.

1 The model

In this paper we consider an $M/G/1$ queueing model with a gated random order of service discipline. In this service discipline customers are first gathered in an unordered waiting room before they are put in random order in an ordered service queue at the moments that this latter queue becomes empty. Figure 1 shows a picture of the queueing system.

The model is motivated by a situation encountered in multi-access communication in cable networks. Cable networks are currently being upgraded to support bidirectional data transport. The system is thus extended with an “upstream” channel to complement the “downstream” channel that is already present. This upstream channel is shared among many stations so that contention resolution is essential for data transport. An efficient way to carry out the upstream data transport is via a request-grant mechanism. Stations request data slots in contention with other stations via contention trees. After a successful request, data transfer follows in reserved slots, not in contention with other stations.

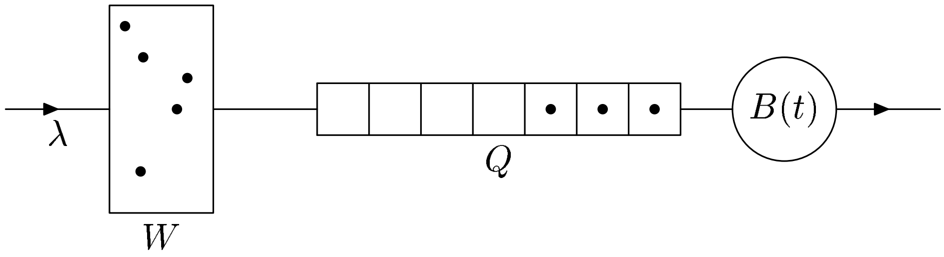

Figure 1: An $M/G/1$ queue with waiting room W and service queue Q.

There are two versions of the contention resolution mechanism via contention trees: the free access variant and the blocked access variant (see Mathys and Flajolet [7]). Essential features of the blocked access variant are

- requests competing in the same tree leave the tree in random order;
- new requests arriving when a tree is in progress have to wait until the current tree is resolved before they can be part of a tree itself.

Exactly these two features lead us to the study of the queueing model with a gated random order of service discipline. Here, customers in the service queue represent the requests currently competing in the tree. Customers in the waiting room represent the requests waiting until the current tree is resolved. Recently, motivated by the same application, a machine-repair model with gated random order of service discipline at the repair facility has been studied by Boxma, Denteneer and Resing [6].

A queueing model with two stages of waiting has been studied by Ali and Neuts [3] (see also Boxma and Cohen [5] for a related model). Essential differences between the models in [3, 5] and our model are that

- in [3, 5] the transfer of customers from the waiting room to the service queue takes some random transfer time $T > 0$, while in our model this transfer time is equal to 0;
- the feature that customers, after transfer, are put in a random order in the service queue is not captured in the models in [3, 5].

Gated $M/G/1$ queues are also studied in Avi-Itzhak and Halfin [4]. Although the authors' main interest in [4] is in the gated processor sharing case, they also discuss the gated FCFS, gated LCFS and gated random order of service case. In Section 4 of this paper we extensively come back to the approach used and the results obtained in [4].

For the $M/G/1$ queue with gated random order of service, we are particularly interested in the joint stationary distribution of the sojourn times of a customer in the waiting room and the service queue. In order to find this distribution, we first study the three-dimensional Markov process describing the number of customers in the waiting room, the number of customers in the service queue and the residual service time of the customer in service.

The arrival process of the customers at the waiting room is assumed to be a Poisson process with intensity λ. The service times of the customers are independently drawn from a distribution $B(t)$, for which we assume $B(0) = 0$, i.e., service times are strictly positive, and the first moment $\beta_1 = \int_0^\infty t \, \mathrm{d}B(t)$ is finite. The Laplace–Stieltjes transform of the service time distribution is denoted by

$$\beta(s) = \int_0^\infty \mathrm{e}^{-st}\,\mathrm{d}B(t). \tag{1}$$

In the sequel we assume that $\rho := \lambda\beta_1 < 1$ so that a steady state exists. As mentioned before, we are particularly interested in the joint distribution of the times S_1 and S_2 an arbitrary customer spends in steady state in the waiting room and in the service queue, respectively. We shall prove the following theorem:

Theorem 1 (Joint distribution of the sojourn times). *The steady-state joint sojourn time distribution*

$$F(x,y) = \Pr(S_1 \le x, S_2 \le y),$$

has a bivariate Laplace–Stieltjes transform

$$\phi(u,v) = \int_{0^-}^\infty \int_0^\infty \mathrm{e}^{-ux}\mathrm{e}^{-vy}\,\mathrm{d}_y\,\mathrm{d}_x F(x,y)$$

that is given by

$$\phi(u,v) = (1-\rho)\beta(v)\left[1 + \frac{\lambda}{u(1-\beta(v))}\int_{\beta(v)}^1 \sum_{\ell=1}^\infty (g_\ell(\xi) - g_\ell(\xi - u/\lambda))\,\mathrm{d}\xi\right],$$

where g_ℓ is the ℓ-fold iterate of the function $g(\xi) = \beta(\lambda - \lambda\xi)$, so $g_1(\xi) = g(\xi)$ and $g_{\ell+1}(\xi) = g(g_\ell(\xi))$ for $\ell \ge 1$.

The remainder of this paper is organized as follows. In Section 2 we study the joint steady-state distribution $P(k,n;x)$ of the number of customers, k,in the waiting room, the number of customers, n, in the service queue, and the residual workload, x, of the customer in service. From this result we derive the joint steady-state distribution $\pi(k,n)$ of the number of customers in the waiting room and service queue, and $P_{\mathrm{tot}}(k;x)$, the steady-state distribution of the number of customers in the waiting room and the total workload of the customers in the service queue. We also study the joint steady-state distribution $\pi_{\mathrm{s}}(k,n)$ of the number of customers in the waiting room and service queue immediately after a customer has completed his service, and we observe that $\pi_{\mathrm{s}}(k,n) \neq \pi(k,n)$. In Section 3 we use the result for $P_{\mathrm{tot}}(k;x)$ to calculate $\phi(u,v)$ and prove Theorem 1. We also compare the results for the LST $\phi(u,v)$ in the model with gated random order of service with the results in the model with gated FCFS and gated LCFS service discipline. Furthermore we give the Laplace–Stieltjes transforms $\phi_1(u)$ and $\phi_2(v)$ of the marginal distributions for the times spent in the waiting room and in the service queue, respectively. After these straightforward calculations, we present a more elegant derivation of the main results in Section 4, using a decomposition of the busy periods based on gate openings and generations, like in Avi-Itzhak and Halfin [4]. Finally, in Section 5 we evaluate our results in case the service time distribution is exponential.

2 Number of customers and workload

In order to describe the system we use the concept of *workload*: the workload of a customer at a certain instant of time is defined as the amount of time the server still needs to spend on serving that customer. For customers who have not yet received service, the workload equals their service time; for the customer who is in service the workload equals the residual service time, i.e., the service time minus the time already spent in service.

The state of the system can be described by a non-negative integer-valued random variable X_1, denoting the number of customers in the waiting room, a non-negative integer-valued random variable X_2, denoting the number of customers in the service queue, and the non-negative real-valued random variable W, denoting the workload of the customer who is currently receiving service. If the service queue is empty, i.e., $X_2 = 0$ then also $X_1 = 0$, since it is assumed that customers are transferred instantaneously from the waiting room to the service queue when the service queue is emptied. We also define W_{tot} as the total workload of the service queue. When $X_2 > 0$, W_{tot} is the sum of $X_2 - 1$ complete service times and W. Let

$$P(0,0;t) = \Pr(X_1 = 0, X_2 = 0 \text{ at time } t) \tag{2}$$

and, for $k \geq 0$ and $n \geq 1$,

$$P(k,n;x;t) = \Pr(X_1 = k, X_2 = n, W \leq x \text{ at time } t) \tag{3}$$

and their steady-state values

$$P(0,0) = \lim_{t\to\infty} P(0,0;t), \tag{4}$$

$$P(k,n;x) = \lim_{t\to\infty} P(k,n;x;t). \tag{5}$$

We define, for $-1 < \xi, \eta < 1$, the generating functions

$$Q(\xi,\eta;x) = \sum_{k=0}^{\infty}\sum_{n=1}^{\infty} \xi^k \eta^{n-1} P(k,n;x) \tag{6}$$

and the Laplace–Stieltjes transform

$$q(\xi,\eta;s) = \int_0^{\infty} \mathrm{e}^{-sx}\, \mathrm{d}Q(\xi,\eta;x). \tag{7}$$

Of particular interest is $\pi(k,n)$, the steady-state probability distribution of X_1 and X_2:

$$\pi(0,0) = P(0,0) \text{ and } \pi(k,n) = \lim_{x\to\infty} P(k,n;x). \tag{8}$$

We denote the generating function of the $\pi(k,n)$ by $Q(\xi,\eta)$:

$$Q(\xi,\eta) = \sum_{k=0}^{\infty}\sum_{n=1}^{\infty} \xi^k \eta^{n-1} \pi(k,n) = \lim_{x\to\infty} Q(\xi,\eta;x). \tag{9}$$

Remark that in the definitions of the generating functions $Q(\xi,\eta;x)$ and $Q(\xi,\eta)$ we omit the probabilities $P(0,0)$ and $\pi(0,0)$ and we have powers η^{n-1} instead of the usual η^n. Clearly,

$$\pi(0,0) = 1 - \rho, \tag{10}$$

because the average amount of work that enters the system per unit of time equals ρ and the server works at rate 1, i.e., one unit of workload per unit time, when the system is not empty. The other probabilities follow from the following theorem.

Theorem 2 (The function $q(\xi,\eta;s)$). *The function $q(\xi,\eta;s)$, defined by (7), is given by*

$$q(\xi,\eta;s)=\frac{\lambda(1-\rho)(\beta(s)-g(\xi))}{(\lambda(1-\xi)-s)(g(\xi)-\eta)}\left[1-\eta+\sum_{\ell=1}^{\infty}(g_\ell(\xi)-g_\ell(\eta))\right].$$

Here g_ℓ is the ℓ-fold iteration of the function g, as defined in Theorem 1.

Proof: First, we relate $P(k,n;x,t+\Delta)$ to $P(k,n;x;t)$ by considering what can happen in the interval $(t,t+\Delta)$. Including only terms of no higher order than Δ, at most one arrival in an interval $(t,t+\Delta)$ needs to be considered. It follows that

$$\begin{aligned}P(k,n;x;t+\Delta)&=(1-\lambda\Delta)[P(k,n;x+\Delta;t)-P(k,n;\Delta;t)\\&\qquad+P(k,n+1;\Delta;t)B(x)]\\&\quad+\lambda\Delta[P(k-1,n;x+\Delta;t)-P(k-1,n;\Delta;t)]\\&\quad+o(\Delta),\quad k\geq 1,n\geq 1,\end{aligned}\tag{11}$$

$$\begin{aligned}P(0,n;x;t+\Delta)&=(1-\lambda\Delta)[P(0,n;x+\Delta;t)-P(0,n;\Delta;t)\\&\qquad+(P(0,n+1;\Delta;t)+P(n,1;\Delta;t))B(x)]\\&\quad+o(\Delta),\quad n\geq 2,\end{aligned}\tag{12}$$

$$\begin{aligned}P(0,1;x;t+\Delta)&=(1-\lambda\Delta)[P(0,1;x+\Delta;t)-P(0,1;\Delta;t)\\&\qquad+(P(0,2;\Delta;t)+P(1,1;\Delta;t))B(x)]\\&\quad+\lambda\Delta P(0,0;t)B(x)+o(\Delta),\end{aligned}\tag{13}$$

$$P(0,0;t+\Delta)=(1-\lambda\Delta)[P(0,0;t)+P(0,1;\Delta;t)]+o(\Delta).\tag{14}$$

Taking the limits $\Delta\to 0$ and $t\to\infty$ and using $P(0,0)=\pi(0,0)=1-\rho$, it follows that

$$P(k,n;0)=0,\quad k\geq 0,n\geq 1,\tag{15}$$

$$\begin{aligned}&P'(k,n;x)-P'(k,n;0)+P'(k,n+1;0)B(x)\\&\qquad-\lambda P(k,n;x)+\lambda P(k-1,n;x)=0,\quad k\geq 1,n\geq 1,\end{aligned}\tag{16}$$

$$\begin{aligned}&P'(0,n;x)-P'(0,n;0)+[P'(0,n+1;0)+P'(n,1;0)]B(x)\\&\qquad-\lambda P(0,n;x)=0,\quad n\geq 2,\end{aligned}\tag{17}$$

$$\begin{aligned}&P'(0,1;x)-P'(0,1;0)+[\lambda(1-\rho)+P'(0,2;0)+P'(1,1;0)]B(x)\\&\qquad-\lambda P(0,1;x)=0,\end{aligned}\tag{18}$$

$$P'(0,1;0)=\lambda(1-\rho).\tag{19}$$

Here the prime denotes differentiation with respect to the third variable. In terms of the function $Q(\xi,\eta;x)$, defined in (6), the relations (15)–(19) imply

$$Q(\xi,\eta;0)=0\tag{20}$$

and

$$Q'(\xi,\eta;x)-\lambda(1-\xi)Q(\xi,\eta;x)=\mathcal{A}(\xi,\eta;x),\tag{21}$$

where

$$\begin{aligned}\mathcal{A}(\xi,\eta;x)&=Q'(\xi,\eta;0)\\&\quad-B(x)\left(\lambda(1-\rho)(1-\frac{1}{\eta})+\frac{1}{\eta}(Q'(\xi,\eta;0)+Q'(\eta,0;0)-Q'(\xi,0;0))\right).\end{aligned}\tag{22}$$

The differential equation (21) for $Q(\xi,\eta;x)$ with initial condition (20) is easily solved:

$$Q(\xi,\eta;x)=\mathrm{e}^{\lambda(1-\xi)x}\int_0^x \mathcal{A}(\xi,\eta;y)\mathrm{e}^{-\lambda(1-\xi)y}\,\mathrm{d}y. \tag{23}$$

Since $Q(\xi,\eta;\infty)=Q(\xi,\eta)$ must be finite for $-1<\xi,\eta<1$,

$$\int_0^\infty \mathcal{A}(\xi,\eta;y)\mathrm{e}^{-\lambda(1-\xi)y}\,\mathrm{d}y=0. \tag{24}$$

With partial integration and using (1) it follows that

$$\int_0^\infty B(y)\mathrm{e}^{-\lambda(1-\xi)y}\,\mathrm{d}y=\frac{g(\xi)}{\lambda(1-\xi)}, \tag{25}$$

where $g(\xi)=\beta(\lambda-\lambda\xi)$ as defined in Theorem 1, so (24) gives

$$\left(1-\frac{g(\xi)}{\eta}\right)Q'(\xi,\eta;0)=g(\xi)\lambda(1-\rho)\left(1-\frac{1}{\eta}\right)+\frac{g(\xi)}{\eta}(Q'(\eta,0;0)-Q'(\xi,0;0)). \tag{26}$$

Finiteness of $Q'(\xi,\eta;0)$ at $\eta=g(\xi)$ implies that

$$Q'(\xi,0;0)=Q'(g(\xi),0;0)-\lambda(1-\rho)(1-g(\xi)). \tag{27}$$

Iterating this equation L times we get

$$Q'(\xi,0;0)=Q'(g_L(\xi),0;0)-\lambda(1-\rho)\sum_{\ell=1}^{L}(1-g_\ell(\xi)), \tag{28}$$

where g_ℓ is the ℓ-fold iterate of g, as defined in Theorem 1. The function g is increasing on $(-\infty,1]$ and has a fixed point (for which $g(\xi)=\xi$) at $\xi=1$, since $\beta(0)=1$. Since $g'(1)=\rho<1$, this fixed point is stable and its domain of attraction includes the interval $(-\infty,1]$. As $e^x\geq 1+x$ for all real x,

$$g(\xi)=\int_0^\infty \mathrm{e}^{-(\lambda-\lambda\xi)x}\,\mathrm{d}B(x)\geq\int_0^\infty(1-(\lambda-\lambda\xi)x)\,\mathrm{d}B(x)=1-\rho(1-\xi), \tag{29}$$

so for $\xi\leq 1$

$$0\leq 1-g(\xi)\leq\rho(1-\xi)\text{ and }0\leq 1-g_\ell(\xi)\leq\rho^\ell(1-\xi). \tag{30}$$

Hence, for $\rho<1$, $\lim_{\ell\to\infty}g_\ell(\xi)=1$ and $\sum_{\ell=0}^\infty(1-g_\ell(\xi))$ converges. It then follows from (28) for $L\to\infty$ that for $\rho<1$

$$Q'(\xi,0;0)=Q'(1,0;0)-\lambda(1-\rho)\sum_{\ell=1}^{\infty}(1-g_\ell(\xi)) \tag{31}$$

and making use of $Q'(0,0;0)=P'(0,1,0)=\lambda(1-\rho)$, see (19), it follows that

$$Q'(\xi,0;0)=\lambda(1-\rho)(1+G(\xi)), \tag{32}$$

where $G(\xi)$ is defined as

$$G(\xi) = \sum_{\ell=1}^{\infty} (g_\ell(\xi) - g_\ell(0)). \tag{33}$$

Equation (26) now gives

$$Q'(\xi,\eta;0) = \frac{\lambda(1-\rho)g(\xi)}{g(\xi)-\eta}[1-\eta+G(\xi)-G(\eta)]. \tag{34}$$

so that $\mathcal{A}(\xi,\eta;x)$ is found, and using (24) we have

$$\begin{aligned} Q(\xi,\eta;x) &= \mathrm{e}^{\lambda(1-\xi)x} \int_0^x \mathcal{A}(\xi,\eta;y)\mathrm{e}^{-\lambda(1-\xi)y}\,\mathrm{d}y \\ &= -\mathrm{e}^{\lambda(1-\xi)x} \int_x^\infty \mathcal{A}(\xi,\eta;y)\mathrm{e}^{-\lambda(1-\xi)y}\,\mathrm{d}y. \end{aligned} \tag{35}$$

It is convenient to work with the LST of $Q(\xi,\eta;x)$, as defined in (7):

$$\begin{aligned} q(\xi,\eta;s) &= \int_0^\infty \mathrm{e}^{-sx}\,\mathrm{d}Q(\xi,\eta;x) \\ &= \int_0^\infty \mathrm{e}^{-sx}\mathcal{A}(\xi,\eta;x)\,\mathrm{d}x - \int_0^\infty \mathrm{e}^{-sx}\lambda(1-\xi)\mathrm{e}^{\lambda(1-\xi)x}\int_x^\infty \mathcal{A}(\xi,\eta;y)\mathrm{e}^{-\lambda(1-\xi)y}\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_0^\infty \mathrm{e}^{-sx}\mathcal{A}(\xi,\eta;x)\,\mathrm{d}x - \lambda(1-\xi)\int_0^\infty \mathcal{A}(\xi,\eta;y)\mathrm{e}^{-\lambda(1-\xi)y}\int_0^y \mathrm{e}^{(\lambda(1-\xi)-s)x}\,\mathrm{d}x\,\mathrm{d}y \\ &= -\frac{s}{\lambda(1-\xi)-s}\int_0^\infty \mathcal{A}(\xi,\eta;x)\mathrm{e}^{-sx}\,\mathrm{d}x, \end{aligned} \tag{36}$$

where in the last step (24) was used. Using (22) and (34) we find

$$\begin{aligned} q(\xi,\eta;s) &= \frac{1}{\lambda(1-\xi)-s}\Bigg[\beta(s)\Big[\lambda(1-\rho)(1-\frac{1}{\eta}) \\ &\qquad + \frac{1}{\eta}(Q'(\xi,\eta;0)+Q'(\eta,0;0)-Q'(\xi,0;0))\Big] - Q'(\xi,\eta;0)\Bigg] \\ &= \frac{\lambda(1-\rho)(\beta(s)-g(\xi))}{(\lambda(1-\xi)-s)(g(\xi)-\eta)}[1-\eta+G(\xi)-G(\eta)]. \end{aligned} \tag{37}$$

This completes the proof of Theorem 2. □

Since $Q(\xi,\eta;0)=0$, we have

$$Q(\xi,\eta) = \lim_{x\to\infty} Q(\xi,\eta;x) = \int_0^\infty \mathrm{d}Q(\xi,\eta;x) = q(\xi,\eta;0), \tag{38}$$

by definition of $q(\xi,\eta;s)$, so a direct corollary from Theorem 2 is

Corollary 3 (The function $Q(\xi,\eta)$). *The function $Q(\xi,\eta)$, defined in (9), is given by*

$$Q(\xi,\eta) = \frac{(1-\rho)(1-g(\xi))}{(1-\xi)(g(\xi)-\eta)}\left[1-\eta+G(\xi)-G(\eta)\right].$$

Next we will study the steady-state joint probability distribution of the occupancy of the waiting room and the total workload of the service queue,

$$P_{\mathrm{tot}}(k;x) = \Pr(X_1 = k, W_{\mathrm{tot}} \le x), \tag{39}$$

with transform

$$q_{\mathrm{tot}}(\xi;s) = E\left(\xi^{X_1} \mathrm{e}^{-sW_{\mathrm{tot}}}\right) = \sum_{k=0}^{\infty} \int_{0^-}^{\infty} \xi^k \mathrm{e}^{-sx}\, \mathrm{d}P_{\mathrm{tot}}(k;x). \tag{40}$$

Observe that $W_{\mathrm{tot}} = 0$ if $X_2 = 0$ and that

$$W_{\mathrm{tot}} = W + \sum_{i=1}^{X_2 - 1} Y_i \tag{41}$$

if $X_2 \ge 1$, where the Y_i are i.i.d. with distribution $B(\cdot)$. Therefore

$$\begin{aligned} q_{\mathrm{tot}}(\xi;s) &= \sum_{k=0}^{\infty} \int_{0^-}^{\infty} \xi^k \mathrm{e}^{-sx}\, \mathrm{d}P_{\mathrm{tot}}(k;x) \\ &= (1-\rho) \\ &\quad + \sum_{k=0}^{\infty} \sum_{n=1}^{\infty} \int_0^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} \xi^k \mathrm{e}^{-s(y_1 + \cdots + y_{n-1} + x)}\, \mathrm{d}B(y_1) \cdots \mathrm{d}B(y_{n-1})\, \mathrm{d}P(k,n;x) \\ &= 1 - \rho + q(\xi, \beta(s); s). \end{aligned} \tag{42}$$

Using the result (37) and noting that $\beta(s) = g(1 - s/\lambda)$ and that $G(g(z)) = G(z) + 1 - g(z)$, we have

Corollary 4 (The function $q_{\mathrm{tot}}(\xi;s)$). *The joint transform of the number of customers in the waiting room and the total workload in the service queue is given by*

$$q_{\mathrm{tot}}(\xi;s) = E\left(\xi^{X_1} \mathrm{e}^{-sW_{\mathrm{tot}}}\right) = (1-\rho)\left[1 + \frac{\lambda}{\lambda - \lambda\xi - s} \sum_{\ell=1}^{\infty} \left(g_\ell(1 - s/\lambda) - g_\ell(\xi)\right)\right]. \tag{43}$$

Finally, we consider $\pi_{\mathrm{s}}(k,n)$, the steady-state probability that immediately after the service completion of a customer the waiting room and the service queue contain k and n customers, respectively. Clearly, due to the PASTA property we have

$$\pi_{\mathrm{a}}(k,n) = \pi(k,n), \tag{44}$$

where $\pi_{\mathrm{a}}(k,n)$ is the steady-state probability that just before the arrival of a customer $X_1 = k$ and $X_2 = n$. Furthermore, a simple and well-known one-dimensional level-crossing argument shows that

$$\pi_{\mathrm{s}}(0,0) = \pi_{\mathrm{a}}(0,0) \tag{45}$$

$$\sum_{n=1}^{m} \pi_{\mathrm{s}}(m-n,n) = \sum_{n=1}^{m} \pi_{\mathrm{a}}(m-n,n), \qquad m \ge 1, \tag{46}$$

However, there is no reason why the individual probabilities $\pi_{\mathrm{s}}(k,n)$ should be equal to $\pi_{\mathrm{a}}(k,n)$ and hence to $\pi(k,n)$. In fact, it turns out that these probabilities are indeed different. This is expressed in the following theorem.

Theorem 5 (Joint distribution of customers right after service completion). *Let the steady-state probability that $X_1 = k$ and $X_2 = n$ immediately after a customer has completed his service and has left the system be denoted by $\pi_{\mathrm{s}}(k,n)$, then*

$$\pi_{\mathrm{s}}(0,0) = 1-\rho,$$
$$Q_{\mathrm{s}}(\xi,\eta) = \sum_{k=0}^{\infty}\sum_{n=1}^{\infty}\xi^k\eta^{n-1}\pi_{\mathrm{s}}(k,n) = \frac{1-\rho}{g(\xi)-\eta}\Big[1-g(\xi)+\sum_{\ell=1}^{\infty}(g_\ell(\xi)-g_\ell(\eta))\Big].$$

Proof: From the PASTA property and the level-crossing argument we have $\pi_s(0,0) = \pi(0,0) = 1-\rho$. Let A_m denote the probability that m new customers arrive during one service time, then

$$A_m = \int_0^\infty \mathrm{e}^{-\lambda x}\frac{(\lambda x)^m}{m!}\,\mathrm{d}B(x), \tag{47}$$

and so

$$\sum_{m=0}^{\infty} A_m z^m = \beta(\lambda-\lambda z) = g(z). \tag{48}$$

The probabilities $\pi_{\mathrm{s}}(k,n)$ satisfy the following equations which relate the probabilities according to what happens during the service time of one customer:

$$\pi_{\mathrm{s}}(k,n) = \sum_{m=0}^{k} A_{k-m}\pi_{\mathrm{s}}(m,n+1), \quad k\ge 1, n\ge 1 \tag{49}$$

$$\pi_{\mathrm{s}}(0,n) = A_0\pi_{\mathrm{s}}(0,n+1) + A_n\pi_{\mathrm{s}}(0,0) + \sum_{m=0}^{n} A_{n-m}\pi_{\mathrm{s}}(m,1), \quad n\ge 1 \tag{50}$$

$$\pi_{\mathrm{s}}(0,0) = A_0\pi_{\mathrm{s}}(0,1) + A_0\pi_{\mathrm{s}}(0,0). \tag{51}$$

Equations (49)-(51) imply that

$$\left(1-\frac{g(\xi)}{\eta}\right)Q_{\mathrm{s}}(\xi,\eta) = -\frac{g(\xi)}{\eta}Q_{\mathrm{s}}(\xi,0) + \frac{g(\eta)}{\eta}Q_{\mathrm{s}}(\eta,0) - \frac{1-g(\eta)}{\eta}\pi_{\mathrm{s}}(0,0). \tag{52}$$

This equation is similar to (26), the equation for $Q'(\xi,\eta;0)$, and it is solved in exactly the same way. Regularity of $Q_{\mathrm{s}}(\xi,\eta)$ at $\eta = g(\xi)$ implies that

$$Q_{\mathrm{s}}(\xi,0) = \frac{g_2(\xi)}{g(\xi)}Q_{\mathrm{s}}(g(\xi),0) - \frac{1-g_2(\xi)}{g(\xi)}\pi_{\mathrm{s}}(0,0), \tag{53}$$

and iterating this equation gives

$$Q_{\mathrm{s}}(\xi,0) = \frac{1}{g(\xi)}Q_{\mathrm{s}}(1,0) - \frac{1}{g(\xi)}\sum_{\ell=2}^{\infty}(1-g_\ell(\xi))\pi_{\mathrm{s}}(0,0). \tag{54}$$

Substituting $Q_{\mathrm{s}}(0,0) = \pi_{\mathrm{s}}(0,1) = \pi_{\mathrm{s}}(0,0)(1-g(0))/g(0)$, which follows from (51), gives

$$Q_{\mathrm{s}}(1,0) = \pi_{\mathrm{s}}(0,0)\sum_{\ell=1}^{\infty}(1-g_\ell(0)) \tag{55}$$

and hence, with $\pi_{\rm s}(0,0) = \pi(0,0) = 1-\rho$,

$$Q_{\rm s}(\xi,0) = \frac{1-\rho}{g(\xi)}\left[1 - g(\xi) + G(\xi)\right] \tag{56}$$

and (52) implies

$$Q_{\rm s}(\xi,\eta) = \frac{1-\rho}{g(\xi)-\eta}\left[1 - g(\xi) + G(\xi) - G(\eta)\right]. \tag{57}$$

This completes the proof of Theorem 5. □

Note that $Q_{\rm s}(\xi,\xi) = Q(\xi,\xi)$, consistent with the level crossing argument.

3 Sojourn times

Because of the PASTA property, just before the arrival of a new customer the variables $(X_1, W_{\rm tot})$ are distributed according to the functions $P_{\rm tot}(k;x)$. If the total workload of the service queue is $W_{\rm tot} = x$ at the moment a customer enters, that customer has to wait x time units until he is transferred from the waiting room to the service queue, together with the other customers in the waiting room. These other customers fall into two categories, those that were already present in the waiting room when our given customer entered and those that entered in the x time units after the arrival of our given customer. If the numbers of customers in these groups are k and r, respectively, the number of customers that are transferred to the service queue equals $k+r+1$. The probability that r customers enter the system in a time interval of length x is $\mathrm{e}^{-\lambda x}(\lambda x)^r/r!$. Since the order in which the $k+r+1$ customers end up in the service queue is random, our given customer is number ℓ in line, where $\ell = 1, 2, \ldots, k+r+1$, with probability $(k+r+1)^{-1}$ independent of ℓ. If he is number ℓ in line, he spends ℓ randomly drawn service times in the service queue until his service is completed.

Combining all this, the LST of the joint probability distribution $F(x,y)$ of S_1, the time a customer spends in the waiting room, and S_2, the time he subsequently spends in the service queue is given by

$$\begin{aligned}
\phi(u,v) &= E(\mathrm{e}^{-uS_1}\mathrm{e}^{-vS_2}) = \int_{0^-}^{\infty}\int_{0}^{\infty} \mathrm{e}^{-ux-vy}\mathrm{d}_y\mathrm{d}_x F(x,y) \\
&= (1-\rho)\beta(v) \\
&\quad + \sum_{k=0}^{\infty}\int_0^{\infty} \mathrm{e}^{-(u+\lambda)x}\sum_{r=0}^{\infty}\frac{(\lambda x)^r}{r!}\,\frac{1}{k+r+1}\sum_{\ell=1}^{k+r+1}(\beta(v))^{\ell}\,\mathrm{d}P_{\rm tot}(k;x)
\end{aligned} \tag{58}$$

We can now perform the summation over ℓ, giving $(1-(\beta(v))^{k+r+1})\beta(v)/(1-\beta(v))$, and use

$$\frac{1-z^{k+r+1}}{k+r+1} = \int_z^1 \xi^{k+r}\,\mathrm{d}\xi \tag{59}$$

for $z = \beta(v)$. Now the sum over r can be done and comparison with (42) gives

$$\phi(u,v) = \frac{\beta(v)}{1-\beta(v)}\int_{\beta(v)}^{1} E\left(\xi^{X_1}\mathrm{e}^{-(u+\lambda-\lambda\xi)W_{\rm tot}}\right)\mathrm{d}\xi. \tag{60}$$

We now use Corollary 4 with $s = u + \lambda - \lambda\xi$ and obtain

$$\phi(u,v) = (1-\rho)\beta(v)\left[1 + \frac{\lambda}{u(1-\beta(v))}\int_{\beta(v)}^{1}\sum_{\ell=1}^{\infty}(g_\ell(\xi) - g_\ell(\xi - u/\lambda))\,\mathrm{d}\xi\right]. \tag{61}$$

This proves Theorem 1. □

When we take the limit $v \to 0$ we obtain the LST of the marginal distribution $F_1(x) = \Pr(S_1 \le x)$:

$$\phi_1(u) = \phi(u,0) = (1-\rho)\left[1 + \frac{\lambda}{u}\sum_{\ell=1}^{\infty}(1 - g_\ell(1 - u/\lambda))\right] \tag{62}$$

and when we take the limit $u \to 0$ we obtain the LST of the marginal distribution $F_2(x) = \Pr(S_2 \le x)$:

$$\begin{aligned}\phi_2(v) = \phi(0,v) &= (1-\rho)\beta(v)\left[1 + \frac{\lambda}{1-\beta(v)}\int_{\beta(v)}^{1}\sum_{\ell=1}^{\infty}\frac{1}{\lambda}g'_\ell(\xi)\,\mathrm{d}\xi\right] \\ &= (1-\rho)\beta(v)\left[1 + \frac{1}{1-\beta(v)}\sum_{\ell=1}^{\infty}(1 - g_\ell(\beta(v)))\right].\end{aligned} \tag{63}$$

Note that the expressions for $\phi_1(u)$ and $\phi_2(v)$ are remarkably similar:

$$\phi_2(v) = \beta(v)\phi_1(\lambda - \lambda\beta(v)). \tag{64}$$

To understand this similarity it may help to compare the gated random order of service discipline to two other service disciplines, gated first come first served (GFCFS) and gated last come first served (GLCFS). These service disciplines have the same gate mechanism as before, but now the service order in the service queue is the same as (for GFCFS) or the reverse of (for GLCFS) the order of arrival in the waiting room. Note that the results of Section 2 on number of customers and workload remain valid for GFCFS and GLCFS. Furthermore, for both service disciplines the transform $\phi(u,v)$ can be obtained by minor modifications of (58). For GFCFS, the customer we consider is always number $k+1$ in line in the service queue, irrespective of r, so the summation over ℓ in (58) must be replaced by a single term $(\beta(v))^{k+1}$. This gives

$$\begin{aligned}\phi_{\mathrm{GFCFS}}(u,v) &= \beta(v)E(\beta(v)^{X_1}\mathrm{e}^{-uW_{\mathrm{tot}}}) \\ &= (1-\rho)\beta(v)\left[1 + \frac{\lambda}{\lambda - \lambda\beta(v) - u}\sum_{\ell=1}^{\infty}(g_\ell(1-u/\lambda) - g_\ell(\beta(v)))\right].\end{aligned} \tag{65}$$

For GLCFS, the customer we consider is always number $r+1$ in line in the service queue, irrespective of k. This means that we must replace the summation over ℓ by the single term $(\beta(v))^{r+1}$, so that

$$\begin{aligned}\phi_{\mathrm{GLCFS}}(u,v) &= \beta(v)E(\mathrm{e}^{-(u+\lambda-\lambda\beta(v))W_{\mathrm{tot}}}) \\ &= (1-\rho)\beta(v)\left[1 + \frac{\lambda}{\lambda - \lambda\beta(v) + u}\sum_{\ell=1}^{\infty}(1 - g_\ell(\beta(v) - u/\lambda))\right].\end{aligned} \tag{66}$$

Note that the marginal distributions for the three service disciplines are equal:

$$\phi_{\mathrm{GFCFS}}(u,0) = \phi_{\mathrm{GLCFS}}(u,0) = \phi_1(u), \tag{67}$$
$$\phi_{\mathrm{GFCFS}}(0,v) = \phi_{\mathrm{GLCFS}}(0,v) = \phi_2(v). \tag{68}$$

That $\phi_1(u)$ and $\phi_2(v)$ are the same for the three service disciplines can be understood without calculations: suppose that an outside observer is watching the system. This observer cannot distinguish between different customers, he can only record a set of entrance times, a set of gate opening times, and a set of service completion times. Then there is no way this observer can determine whether the service discipline is GROS, GFCFS or GLCFS, but this observer can determine the marginal distributions $\phi_1(u)$ from the observed distribution of entrance times and subsequent gate opening times and he can determine $\phi_2(v)$ from the observed distribution of gate opening times and subsequent service completion times. So $\phi_1(u)$ and $\phi_2(v)$ don't depend on whether the service discipline is GROS, GFCFS or GLCFS.

Equation (64) can most easily be explained by looking at the gated LCFS queue. Here, the waiting time of a customer in the service queue (excluding his own service time) is given by the amount of work arriving during his sojourn time in the waiting room. The LST of this amount of work is given by $\phi_1(\lambda - \lambda\beta(v))$.

We finish this section by noting that there is a relation between $\phi_{\mathrm{GFCFS}}(u,v)$ and the probabilities $\pi_{\mathrm{s}}(k,n)$.

Theorem 6 (Relation between ϕ_{GFCFS} and Q_{s}). *The following holds*

$$\begin{aligned} 1 - \rho + \eta Q_{\mathrm{s}}(\xi,\eta) = {} & \phi_{\mathrm{GFCFS}}(\lambda - \lambda\eta, \lambda - \lambda\xi) \\ & + \phi_{\mathrm{GFCFS}}(\lambda, \lambda - \lambda\eta) - \phi_{\mathrm{GFCFS}}(\lambda, \lambda - \lambda\xi). \end{aligned}$$

Proof: For the model with GFCFS service discipline, let $A_{n,k}$ denote the probability that n customers enter during the tagged customer's sojourn time in the waiting room, and that k customers enter during his sojourn time in the service queue. Then, in analogy with (48), it holds that

$$\sum_{n=0}^{\infty}\sum_{k=0}^{\infty} \eta^n \xi^k A_{n,k} = \phi_{\mathrm{GFCFS}}(\lambda - \lambda\eta, \lambda - \lambda\xi). \tag{69}$$

If $n > 0$, there are k customers in the waiting room and n customers in the service queue when the tagged customer leaves the system. If $n = 0$, the gate opens when the tagged customer had received service, so there are 0 customers in the waiting room and k customers in the service queue when the tagged customer leaves the system. This implies that

$$\pi_{\mathrm{s}}(0,n) = A_{n,0} + A_{0,n}, \qquad n > 0, \tag{70}$$
$$\pi_{\mathrm{s}}(k,n) = A_{n,k}, \qquad k > 0, n > 0, \tag{71}$$
$$\pi_{\mathrm{s}}(0,0) = A_{0,0}. \tag{72}$$

Multiplying the above equation with $\xi^k \eta^n$, summing over k and n, and making use of (69) the theorem follows. □

4 Gate openings and generations: an embedded Markov chain

In the previous sections we analysed the $M/G/1$ queue with gated random order of service using a state description, in which the customers in both the waiting room and the service queue are included. In [4] Avi-Itzhak and Halfin analysed the gated $M/G/1$ queue using an essentially one-dimensional approach: they consider the number of customers that pass through the gate when the service queue empties and furthermore consider the length of the time intervals between successive gate openings. Although they focus on the processor sharing service discipline, they also present some results for the random order of service discipline. In this section we briefly review their method and show how this one-dimensional model can be used to derive the LST of the sojourn times distribution $\phi(u,v)$ and the generating functions $Q(\xi,\eta)$ and $Q_{\mathrm{s}}(\xi,\eta)$.

The key ingredient of the analysis in [4] is the decomposition of the busy period in generations. The zeroth generation consists of one customer who enters in an idle system and who starts the busy period; the first generation consists of those customers who enter the waiting room while the customer in the zeroth generation is being served; the second generation consists of those customers who enter the waiting room while the first generation is in the service queue, and so on: the $j+1^{\mathrm{st}}$ generation consists of all customers who enter during the service time of the j^{th} generation. A busy period ends when the service queue is emptied and there are no customers in the waiting room.

If we denote the number of customers in the j^{th} generation by N_j and the total service time of that generation by T_j, it holds that

$$\Pr[N_0 = 1] = 1, \tag{73}$$

$$\Pr[N_{j+1} = n] = \int_0^\infty \frac{(\lambda t)^n}{n!} \mathrm{e}^{-\lambda t} \,\mathrm{d}\Pr[T_j \le t] \quad \text{for } j \ge 0, \tag{74}$$

and that

$$\Pr[T_j \le t] = \sum_{n=0}^{\infty} \Pr[N_j = n] B^{n*}(t). \tag{75}$$

It follows that the generating function for the number of customers in the j^{th} generation is

$$\sum_{n=0}^{\infty} z^n \Pr[N_j = n] = g_j(z) \tag{76}$$

and that the LST of the service time for the j^{th} generation is given by

$$\int_0^\infty \mathrm{e}^{-st} \,\mathrm{d}\Pr[T_j \le t] = g_j(\beta(s)) = g_{j+1}(1 - s/\lambda), \tag{77}$$

where the functions g_j are as defined in Theorem 1. This gives a nice interpretation of the functions g_j. Together with the LST for the idle time distribution

$$\int_0^\infty \mathrm{e}^{-st} \,\mathrm{d}\Pr[T_{\mathrm{idle}} \le t] = \frac{\lambda}{\lambda + s}, \tag{78}$$

they characterize the intervals between successive gate openings and the number of customers passing through the gate. In particular, the expected number of customers in a j^{th} generation is

$$E(N_j) = \rho^j, \tag{79}$$

the expected duration of the service time of a j^{th} generation is

$$E(T_j) = \lambda^{-1}\rho^{j+1}, \tag{80}$$

and the expected duration of an idle period is

$$E(T_{\text{idle}}) = \lambda^{-1} \tag{81}$$

4.1 Derivation of $\phi(u, v)$

The probability that a randomly picked customer belongs to a j^{th} generation is

$$\frac{E(N_j)}{\sum_{k=0}^{\infty} E(N_k)} = (1-\rho)\rho^j. \tag{82}$$

If $j = 0$, the customer goes directly to the service queue, this contributes $(1-\rho)\beta(v)$ to $\phi(u, v)$. When $j \geq 1$, the customer enters the waiting room during the service time of the preceding $j - 1^{\text{st}}$ generation. The probability that that service time lies in $[t, t + \mathrm{d}t)$ is $(t/E(T_{j-1}))\mathrm{d}\Pr[T_{j-1} \leq t]$, and the sojourn time in the waiting room is uniformly distributed over $(0, t)$. The probability that the particular j^{th} generation to which the chosen customer belongs has m customers equals the probability that $m - 1$ additional customers enter the system, i.e., $(\lambda t)^{m-1}\mathrm{e}^{-\lambda t}/(m-1)!$. When the m customers from this j^{th} generation are served in the service queue, the chosen customer will be number ℓ in line with probability $1/m$, and then that customer will spend ℓ service times in the service queue.

Combining the above, it follows that the contribution of the j^{th} generation with $j \geq 1$ to $\phi(u, v)$ is

$$(1-\rho)\rho^j \int_0^\infty \frac{1}{t}\int_0^t \mathrm{e}^{-ux}\,\mathrm{d}x \sum_{m=1}^{\infty} \frac{(\lambda t)^{m-1}}{(m-1)!}\mathrm{e}^{-\lambda t}\frac{1}{m}\sum_{\ell=1}^{m}\beta(v)^\ell\,\frac{\lambda t}{\rho^j}\,\mathrm{d}\Pr[T_{j-1} \leq t]. \tag{83}$$

Writing

$$\frac{1}{m}\sum_{\ell=1}^{m}\beta(v)^\ell = \frac{\beta(v)}{1-\beta(v)}\int_{\beta(v)}^{1}\xi^{m-1}\,\mathrm{d}\xi, \tag{84}$$

performing the sum over m and the integral over x, making use of (77) and finally summing over the generations j, we find $\phi(u, v)$ as in Theorem 1. If, instead of using (77) to do the integral over t, we do the integral over ξ, we find an alternative expression for $\phi(u, v)$ that for $v = u$ reduces to Equation (60) of [4] (note that there is an error in that formula: the factor t in the integrand should be t^{-1}).

4.2 Derivation of $\pi_s(k, n)$

Just like in the calculation of $\phi(u,v)$ we pick a random customer. With probability $(1-\rho)\rho^j$ that customer belongs to a j^{th} generation and the probability that that generation had m customers is

$$\frac{m \Pr[N_j = m]}{E(N_j)} = \rho^{-j} m \Pr[N_j = m]. \tag{85}$$

We now consider the state of the system when the chosen customer has completed service. When that customer is not the last of his generation, i.e., if he is number $\ell = 1, 2, \ldots, m-1$, each of which happens with probability $1/m$, there are $m - \ell$ customers in the service queue when he leaves the system and the number of customers in the waiting room equals the number of cutomers that entered during ℓ complete service times. If, on the other hand, our chosen customer is the last of his generation, which happens with probability $1/m$, all customers that entered during the m complete service times are transferred to the service queue when our customer completes service. It follows that

$$\pi_s(k,n) = \sum_{j=1}^{\infty}(1-\rho) \sum_{m=n+1}^{\infty} \Pr[N_j = m] \int_0^{\infty} \frac{(\lambda x)^k}{k!} e^{-\lambda x} \, dB^{(m-n)*}(x) \quad \text{for } k > 0 \tag{86}$$

$$\begin{aligned}\pi_s(0,n) &= \sum_{j=1}^{\infty}(1-\rho) \sum_{m=n+1}^{\infty} \Pr[N_j = m] \int_0^{\infty} e^{-\lambda x} \, dB^{(m-n)*}(x) \\ &\quad + \sum_{j=0}^{\infty}(1-\rho) \sum_{m=1}^{\infty} \Pr[N_j = m] \int_0^{\infty} \frac{(\lambda x)^n}{n!} e^{-\lambda x} \, dB^{m*}(x) \quad \text{for } n > 0\end{aligned} \tag{87}$$

$$\pi_s(0,0) = \sum_{j=0}^{\infty}(1-\rho) \sum_{m=1}^{\infty} \Pr[N_j = m] \int_0^{\infty} e^{-\lambda x} \, dB^{m*}(x). \tag{88}$$

Making use of (76), theorem 5 follows after straightforward calculations.

4.3 Derivation of $\pi(k, n)$

In order to calculate the steady-state probabilities $\pi(k,n)$, we pick a random point in time. The probability that it lies in an idle interval is

$$\frac{E(T_{\text{idle}})}{E(T_{\text{idle}}) + \sum_{k=0}^{\infty} E(T_k)} = (1-\rho) \tag{89}$$

and the probability that it lies in an interval in which a j^{th} generation is in service is

$$\frac{E(T_j)}{E(T_{\text{idle}}) + \sum_{k=0}^{\infty} E(T_k)} = (1-\rho)\rho^{j+1}. \tag{90}$$

The system is idle if and only if it is in state $(0,0)$, so

$$\pi(0,0) = 1 - \rho. \tag{91}$$

If the random point lies in generation j, the probability that that particular generation has m customers is $m \Pr[N_j = m]/E(N_j)$, and the probability that the random point lies in the service interval of the ℓ^{th} of these customers (so that $\ell-1$ customers have already been served and have left the system) is $\Pr[N_j = m]/E(N_j)$, for $\ell = 1, 2, \ldots, m$. The probability that the service time of this particular customer lies in $[t, t + \mathrm{d}t)$ is $(t/E(B))\mathrm{d}B(t)$. The chosen point in time will be uniformly distributed over this service time, so the time x that the particular customer has already spent in service is also uniformly distributed over $[0, t]$. The total time the server has already spent on the customers of this j^{th} generation is the sum of x and $\ell - 1$ complete service times. The customers that arrived during this period are in the waiting room. It now follows that $\pi(k, n)$ for $n > 0$ is given by

$$\sum_{j=0}^{\infty}(1-\rho)\rho^{j+1}\sum_{m=n}^{\infty}\int_0^{\infty}\int_0^{\infty}\int_0^{t}\frac{(\lambda(x+y))^k}{t\,k!}\mathrm{e}^{-\lambda(x+y)}\,\mathrm{d}x\,\mathrm{d}B^{(m-n)*}(y)\frac{t\Pr[N_j = m]}{E(B)E(N_j)}\,\mathrm{d}B(t). \quad (92)$$

The generating function $Q(\xi, \eta)$ is obtained by straightforward manipulations: summing over k, performing the integrals over x and y, interchanging the summations over m and n, and making use of (76). After these manipulations we obtain the result of Corollary 3.

5 Exponential service times

When the service time distribution is exponential, i.e., $B(t) = 1 - \mathrm{e}^{-\mu t}$, the expressions for $Q(\xi, \eta)$ and $\phi(u, v)$ become particularly simple. We first look at the number of customers and workload. When we substitute

$$\beta(s) = \frac{\mu}{\mu + s} \quad (93)$$

so that

$$g(\xi) = \frac{1}{1 + \rho(1 - \xi)}, \text{ with } \rho = \lambda/\mu, \quad (94)$$

into (37) we find that

$$q(\xi, \eta; s) = q(\xi, \eta; 0)\frac{\mu}{\mu + s}. \quad (95)$$

This implies

$$P(k, n; x) = \pi(k, n)(1 - \mathrm{e}^{-\mu x}), \quad (96)$$

which reflects, of course, the memoryless property of the exponential distribution. The residual workload of the customer in service has the same exponential distribution as his total workload. We now turn to the probabilities $\pi(k, n)$. Using $G(g(\xi)) = G(\xi) + 1 - g(\xi)$ the expression for $Q(\xi, \eta)$ as derived in Corollary 3 can be written in the form

$$Q(\xi, \eta) = \frac{(1 - \rho)(1 - g(\xi))}{1 - \xi}\left[1 + \frac{G(g(\xi)) - G(\eta)}{g(\xi) - \eta}\right]. \quad (97)$$

By iterating the function g from (94) it is easy to verify that

$$g_\ell(z) = \frac{1 - \rho^\ell - z\rho(1 - \rho^{\ell-1})}{1 - \rho^{\ell+1} - z\rho(1 - \rho^\ell)}. \quad (98)$$

Then we have

$$\frac{g_{\ell+1}(\xi)-g_\ell(\eta)}{g(\xi)-\eta}=\frac{\rho^\ell(1-\rho)^2(1+\rho-\xi\rho)}{(1-\rho^{\ell+2}-\xi\rho(1-\rho^{\ell+1}))(1-\rho^{\ell+1}-\eta\rho(1-\rho^\ell))}, \tag{99}$$

so

$$Q(\xi,\eta)=\sum_{\ell=0}^{\infty}\frac{\rho^{\ell+1}(1-\rho)^3}{(1-\rho^{\ell+2}-\xi\rho(1-\rho^{\ell+1}))(1-\rho^{\ell+1}-\eta\rho(1-\rho^\ell))}. \tag{100}$$

The series expansion in ξ and η is now straightforward. The probability $\pi(k,n)$ is the coefficient of $\xi^k\eta^{n-1}$ ($k\geq 0$, $n\geq 1$) in that series expansion. The result is

$$\pi(k,n)=\sum_{\ell=0}^{\infty}C_\ell\,(s_{\ell+1})^k\,(s_\ell)^{n-1}, \tag{101}$$

where

$$s_\ell=\frac{\rho(1-\rho^\ell)}{1-\rho^{\ell+1}}, \tag{102}$$

$$C_\ell=\frac{\rho^{\ell+1}(1-\rho)^3}{(1-\rho^{\ell+2})(1-\rho^{\ell+1})}. \tag{103}$$

Note that the $\ell=0$ term in (101) contributes only for $n=1$, because $s_0=0$. We derived the 'sum of products' form (101) for the $M/M/1$ case in [8] using a compensation method for two-dimensional Markov chains originally developed by Adan [1] (see also Adan, Wessels and Zijm [2]). In that preprint we also studied the large $k+n$ behaviour of $\pi(k,n)$.

Finally, we turn to the LST of the sojourn times distribution. The integral over ξ in (61) can be done analytically and after a little algebra we obtain

$$\phi(u,v)=(1-\rho)\frac{\mu}{\mu+v}-\frac{\mu^2}{uv}\sum_{\ell=0}^{\infty}\frac{C_\ell}{s_{\ell+1}}\ln\left(1-\frac{uvs_{\ell+1}}{(u+\mu(1-s_\ell))(v+\mu(1-s_{\ell+1}))}\right). \tag{104}$$

6 Conclusion

We have studied the steady-state behaviour of the $M/G/1$ queue with GROS, GFCFS and GLCFS service disciplines using two different approaches. The first approach makes use of the balance equations for $P(k,n;x)$, the second approach focuses on the properties of the time intervals between successive gate openings and of the number of customers that pass through the gate at the gate openings. The latter approach is more elegant, but quite subtle, and careful reasoning is required to derive the results that follow straightforwardly in the former approach.

References

[1] Adan, I.J.B.F. (1991) A compensation approach for queueing problems, Ph.D. thesis, Eindhoven University of Technology.

[2] Adan, I.J.B.F., Wessels, J., and Zijm, W.H.M. (1993). A compensation approach for two-dimensional Markov processes. *Adv. Appl. Probab.* **25**, 783–817.

[3] Ali, O.M.E. and Neuts, M.F. (1984) A service system with two stages of waiting and feedback of customers. *J. Appl. Probab.* **21**, 404–413.

[4] Avi-Itzhak, B. and Halfin, S. (1989) Response times in gated $M/G/1$ queues: the processor sharing case. *Queueing Systems* **4**, 263–279.

[5] Boxma, O.J. and Cohen, J.W. (1991) The $M/G/1$ queue with permanent customers. *IEEE J. Sel. Areas Commun.* **9**, 179–184.

[6] Boxma, O.J., Denteneer, D., and Resing, J.A.C. (2002) Some models for contention resolution in cable networks. In: Proceedings Networking 2002, Pisa, Italy.

[7] Mathys, P. and Flajolet, Ph. (1985). Q-ary collision resolution algorithms in random-access systems with free or blocked channel access. *IEEE Trans. Inform. Theory* **31**, 217–243.

[8] Resing, J.A.C. and Rietman, R. (2002). The $M/M/1$ queue with gated random order of service. Technische Universiteit Eindhoven SPOR-Report-2002-06.