Queueing Syst
DOI 10.1007/s11134-011-9266-6

Exact simulation of the stationary distribution of the FIFO M/G/c queue: the general case for $\rho < c$

Karl Sigman

Received: 1 May 2011 / Revised: 3 September 2011

Abstract We present an exact simulation algorithm for the stationary distribution of customer delay for FIFO M/G/c queues in which $\rho = \lambda/\mu < c$. In Sigman (J. Appl. Probab. 48A:209–216, 2011) an exact simulation algorithm was presented but only under the strong condition that $\rho < 1$ (super stable case). We only assume that the service-time distribution $G(x) = P(S \leq x)$, $x \geq 0$, with mean $0 < E(S) = 1/\mu < \infty$, and its corresponding equilibrium distribution $G_e(x) = \mu \int_0^x P(S > y)\, dy$ are such that samples of them can be simulated. Unlike the methods used in Sigman (J. Appl. Probab. 48A:209–216, 2011) involving coupling from the past, here we use different methods involving discrete-time processes and basic regenerative simulation, in which, as regeneration points, we use return visits to state 0 of a corresponding random assignment (RA) model which serves as a sample-path upper bound.

Keywords Exact simulation · Queueing theory · Multi-server queues · Regenerative processes

Mathematics Subject Classification (2000) 60K25 · 65C05 · 68U20 · 90B22 · 60J05 · 60K05

1 Introduction

Consider a first-in-first-out (FIFO) M/G/c queue ($c \geq 2$), with Poisson arrival times $\{t_n : n \geq 1\}$ at rate λ, in which the independent and identically distributed (iid) service times $\{S_n : n \geq 0\}$ are distributed as $G(x) = P(S \leq x)$, $x \geq 0$, with finite mean $E(S) = 1/\mu$.

K. Sigman (✉)
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, USA
e-mail: ks20@columbia.edu

Published online: 09 November 2011

With iid interarrival times $A_n = t_{n+1} - t_n$ ($t_0 \stackrel{\text{def}}{=} 0$), let $\mathbf{W}_n = (W_n(1), \ldots, W_n(c))$ denote the *Kiefer–Wolfowitz workload vector* (see for example, p. 341 in Chap. 12 of [1]). It satisfies the recursion

$$\mathbf{W}_{n+1} = R(\mathbf{W}_n + S_n\mathbf{e} - A_n\mathbf{f})^+, \quad n \geq 0, \tag{1}$$

where $\mathbf{W}_n = (W_n(1), \ldots, W_n(c))$, $\mathbf{e} = (1, 0, \ldots, 0)$, $\mathbf{f} = (1, 1, \ldots, 1)$, R places a vector in ascending order, and $^+$ takes the positive part of each coordinate. $D_n = W_n(1)$ is then the customer delay in queue (line) of the nth customer. Recursion (1) defines a Markov chain due to the given iid assumptions, and whenever $\mathbf{W}_n = \mathbf{0}$, the chain regenerates in the sense of a regenerative stochastic process with initial condition $\mathbf{W}_0 = \mathbf{0}$. The event $\mathbf{W}_n = \mathbf{0}$ is equivalent to "the nth arrival finds the entire system empty." With $\rho \stackrel{\text{def}}{=} \lambda/\mu < c$ (stability), it is well known that $\mathbf{W}_n$ converges in distribution to a proper stationary distribution. Let π denote this stationary distribution. Our objective in the present paper is to provide a simulation algorithm for sampling exactly from π.

In [6], an exact simulation algorithm was presented but only under the strong assumption that $\rho = \lambda/\mu < 1$. The method involves ideas/methods of *coupling from the past*, *dominated coupling from the past*. Here, we present a different algorithm that works for any $\rho < c$. It is based on general methods of simulating stationary distributions of regenerative processes. Also, we will work in discrete time (e.g., from arrival epochs) instead of continuous time as was the case in [6]. Our only assumption is that we can simulate from both G and its equilibrium distribution G_e.

The main idea involves using, as a sample-path upper bound, the random assignment (RA) model (for total number of customers in the system), and using its returns to state 0 (empty state) as regeneration points which also then serve as regeneration points for the FIFO model.

2 Preliminaries on regenerative simulation

Suppose one can simulate a non-delayed version of a positive recurrent regenerative process. Here we quickly present the basic method of simulating exactly from the stationary distribution of this process if one is able to simulate exactly from the equilibrium distribution (stationary excess distribution) of a cycle length. The result given below, presented in discrete time, with a simple proof provided for completeness, is from [3] (also see the more recent exposition in [2], Sect. 8, p. 420, in particular Proposition 8.4).

Suppose that $\{X_n : n \geq 0\}$ is a positive recurrent non-delayed discrete-time regenerative process, with iid cycle lengths generically denoted by T distributed as $F(n) = P(T \leq n)$, $n \geq 0$, with finite and nonzero mean $E(T) = 1/\lambda$. A generic length T cycle is thus $C = \{X_n : 0 \leq n < T\}$. From regenerative process theory, the (marginal) stationary distribution π is given by (expected value over a cycle divided by the expected cycle length)

$$\pi(\cdot) = \lambda E \sum_{n=0}^{T-1} I\{X_n \in \cdot\} = \lambda E \sum_{n=1}^{T} I\{X_n \in \cdot\}. \tag{2}$$

Proposition 2.1

1. *Suppose we can and do sequentially simulate iid copies of* $C = \{X_n : 0 \le n < T\}$ (*the first cycle*), *denoted by* $C_j = \{X_n(j) : 0 \le n < T_j\},\ j \ge 1$, *having iid cycle lengths* $\{T_j\}$ *distributed as* F.
2. *Suppose further that we can and do simulate* (*independently*) *one copy* T^e *distributed as the* (*discrete-time*) *equilibrium distribution of* F *having probability mass function* $P(T^e = n) = \lambda P(T \ge n),\ n \ge 1$. (*This is the stationary excess distribution for the underlying renewal process of regeneration times*.)
3. *Let* $\tau = \min\{j \ge 1 : T_j \ge T^e\}$.
4. *Use cycle* C_τ *to construct* $X^* = X_{T^e}(\tau)$ (*e.g., if* $T^e = n$ *and* $\tau = j$, *then* $X^* = X_n(j)$).

   Then the simulated random element X^* *is distributed as* π.

Proof Conditional on $T^e = n$, it holds that $\tau = \min\{j \ge 1 : T_j \ge n\}$, and thus C_τ simply has the distribution of a first cycle given that its length is greater than or equal to n:

$$P\big(X^* \in \cdot \mid T^e = n\big) = P(X_n \in \cdot \mid T \ge n) = \frac{P(X_n \in \cdot\ ,\ T \ge n)}{P(T \ge n)}.$$

Since $P(T^e = n) = \lambda P(T \ge n)$, we obtain

$$\begin{aligned} P(X^* \in \cdot) &= \sum_{n=1}^{\infty} \frac{P(X_n \in \cdot\ ,\ T \ge n)}{P(T \ge n)} \lambda P(T \ge n) \\ &= \lambda \sum_{n=1}^{\infty} P(X_n \in \cdot\ ,\ T \ge n) \\ &= \lambda E \sum_{n=1}^{T} I\{X_n \in \cdot\} \\ &= \pi(\cdot). \end{aligned}$$

□

Remark 2.1 Proposition 2.1 is also valid for continuous-time regenerative processes in which case T^e is a continuous random variable with probability density function $\lambda P(T > x),\ x \ge 0$. In the present paper, however, we are only using the discrete-time framework.

3 Random assignment model (RA) as a sample-path upper bound

Given a c-server queueing model, the *random assignment model* (*RA*) is the case when each of the c servers forms its own FIFO single-server queue, and each arrival to the system, independent of the past, randomly chooses queue i to join with probability $1/c,\ i \in \{1, 2, \ldots, c\}$. In the M/G/c case, we refer to this as the RA M/G/c model.

The following is a special case of Lemma 1.3, p. 342 in [1]. (Such results and others that are even more general are based on [4, 8], and [5].)

Lemma 3.1 *Let $Q_F(t)$ denote total number of customers in system at time $t \geq 0$ for the FIFO M/G/c queue, and let $Q_{RA}(t)$ denote total number of customers in system at time t for the corresponding RA M/G/c model in which both models are initially empty and fed exactly the same input of Poisson arrivals $\{t_n\}$ and iid service times $\{S_n\}$. Assume further that for both models the service times are used by the servers in the order in which service initiations occur (S_n is the service time used for the nth such initiation). Then*

$$P\big(Q_F(t) \leq Q_{RA}(t), \text{ for all } t \geq 0\big) = 1. \tag{3}$$

The importance of Lemma 3.1 is that it allows us to jointly simulate versions of the two stochastic processes $\{Q_F(t) : t \geq 0\}$ and $\{Q_{RA}(t) : t \geq 0\}$ while achieving a coupling such that (3) holds. In particular, whenever an arrival finds the RA model empty, the FIFO model is found empty as well. These consecutive epochs in time constitute regeneration points (for both models) due to the iid assumptions on the input. We explain how to use these facts to our advantage in the next section.

Remark 3.1 Under FIFO, customers enter service in the same order that they arrive and so assigning S_n for the nth service initiation is the same as assigning S_n to the nth arriving customer. For the RA model this is not so, since customers can enter service in a different order from their order of arrival. Reordering the service times, however, does not change the distribution of $\{Q_{RA}(t) : t \geq 0\}$ because of the iid assumptions.

4 Using regeneration points from the RA model for the FIFO model: simulating copies of a cycle C for the FIFO model

For the RA M/G/c model, let $\mathbf{Q}(t) = (Q_1(t), \ldots, Q_c(t))$, where $Q_i(t)$ denotes the number of customers in the ith queue at time t (including the customer in service, if any), and let $\mathbf{Q}_n = (Q_{1,n}, \ldots, Q_{c,n}) = \mathbf{Q}(t_n-)$ denote the number in system at the nodes as found by the nth arriving customer (to the entire RA model and not including themselves). We will simulate the discrete-time process $\mathbf{Q}_n$, starting empty, $\mathbf{Q}_0 = \mathbf{0}$, until it empties again. Consecutive visits of $\mathbf{Q}_n$ to the empty state $\mathbf{0}$ constitute positive recurrent regeneration points for the RA model. (See the discussion right after (7).) These also serve as positive recurrent regeneration points for the FIFO model due to Lemma 3.1. Any arrival finding the RA model empty, will find the FIFO model empty as well: If $Q_{RA}(t_n-) = 0$, then $Q_F(t_n-) = 0$ and hence $\mathbf{W}_n = \mathbf{0}$ (recall the Markov chain defined in (1)).[1] We will now proceed to take advantage of this so as to employ Proposition 2.1.

A generic cycle length T is defined by initializing $\mathbf{Q}_0 = \mathbf{0}$ and setting

$$T = \min\{n \geq 1 : \mathbf{Q}_n = \mathbf{0}\}. \tag{4}$$

This yields a generic cycle for the RA model. To generate a sample of T requires a standard discrete-event simulation of $\{\mathbf{Q}(t) : t \geq 0\}$, where the events are an *arrival*

[1] These are different than consecutive visits of the FIFO model to state $\mathbf{0}$, which would occur more often.

versus a *service completion*, and a service time S is generated only when it is needed for processing by a server to ensure that Lemma 3.1 applies. The sequentially generated input random variables (rv's) required are the iid service times $\{S_n : n \geq 0\}$ distributed as G, the iid interarrival times $\{A_n : n \geq 0\}$ distributed as exponential at rate λ, and the iid random selection rv's $\{U_n : n \geq 0\}$ distributed as the discrete uniform distribution over $\{1, 2, \ldots, c\}$. (If $U_n = i$, then the nth arrival joins the ith queue.)

At time $t_0 = 0$, the rv U_0 is generated, and a server is randomly selected according to U_0 and begins servicing a generated service time S_0 (e.g., the system is found empty at time $t_0 = 0$ by an initial customer who then starts the cycle). The number in system at queue U_0 is increased to 1. A_0 is then generated so as to schedule the next arrival. The simulation continues into the future analogously until an arriving customer finds the entire system empty, thus ending the RA cycle.

We do not simulate the FIFO model until the RA cycle is complete (because we want to efficiently use Step 3 of Proposition 2.1), at which time we use the input that was used for the RA cycle to construct the FIFO cycle for the workload vector in (1):

We store the T service times used in simulating T $((S_0, \ldots, S_{T-1}))$ as well as the T interarrival times $((A_0, \ldots, A_{T-1}))$ so they can be used to construct the FIFO cycle $C = \{\mathbf{W}_1, \ldots, \mathbf{W}_T\}$ by using recursion (1) with $\mathbf{W}_0 = \mathbf{0}$, from $n = 0$ up to $n = T - 1$.

5 Simulating a copy of T^e

To fully employ Proposition 2.1, we need to be able to simulate a copy of T^e. Here we show how to do this. The main idea is to take advantage of the basic fact that T^e has the stationary excess distribution (stationary forward recurrence time distribution) of the (discrete-time) renewal process of visits of the RA model to the empty state (this renewal process has iid cycle lengths distributed as T). Letting $A(n)$ denote the excess at time n, that is, the amount of time units starting from n until the next renewal occurs (it forms an aperiodic positive recurrent Markov chain), we have that $A(n) \Rightarrow T^e$ in distribution, as $n \to \infty$. But in our case, by the definition of the RA regeneration points, we also have that $A(n) = \min\{k \geq 1 : \mathbf{Q}_{n+k} = \mathbf{0}\}$. Thus, taking $n \to \infty$, we see that if we take a stationary version of $\{\mathbf{Q}_n : n \geq 0\}$, denoted by $\{\mathbf{Q}_n^* : n \geq 0\}$, then $T^e = \min\{n \geq 1 : \mathbf{Q}_n^* = \mathbf{0}\}$.

Since to obtain a stationary version of the RA model we only need to consider the RA model by itself, *independently* of the FIFO model (recall Step 2 of Proposition 2.1), we do not need to couple the service times as in Lemma 3.1; we can assign service times upon arrival. Also, because arrivals are Poisson, and we independently partition them into c independent Poisson processes each at rate λ/c, *we can simply treat each server as its own independent stable FIFO M/G/1 queue with Poisson arrivals at rate* $\lambda/c < \mu$. Moreover, we can model *workload* instead of number in system since they both empty together and thus share the same regeneration times.

The stationary workload distribution at each queue i is thus given by the classic *Pollaczek–Khintchine* formula. In terms of a random variable D, the stationary

distribution has representation

$$D = \sum_{j=1}^{L} Y_j, \tag{5}$$

where the $\{Y_j\}$ are iid distributed as the equilibrium distribution of service, with cumulative distribution function given by $G_e(x) = \mu \int_0^x P(S > y)\, dy,\ x \geq 0$, and independently L has a geometric distribution, $P(L = k) = \rho_1^k(1 - \rho_1),\ k \geq 0$, where $\rho_1 \stackrel{\text{def}}{=} \lambda/c\mu$ (see, for example, Theorem 5.7(b), p. 237, in [1]).

To put this to use: Letting $\mathbf{V}_n = (V_n(1), \ldots, V_n(i))$ denote workload (at each node) as found by the nth arriving customer to the RA model (from the rate λ Poisson process with arrival times $\{t_n\}$), we have, for each node $i \in \{1, 2, \ldots, c\}$,

$$V_{n+1}(i) = \big(V_n(i) + S_n I\{U_n = i\} - A_n\big)^+, \quad n \geq 0, \tag{6}$$

where here, S_n is the iid service time of the nth (Poisson rate λ) arriving customer, and independently $\{U_n : n \geq 0\}$ denotes an iid sequence of random variables with the discrete uniform distribution over $\{1, 2, \ldots, c\}$. ($I\{B\}$ is the indicator rv for the event B.)

$\{\mathbf{V}_n : n \geq 0\}$ forms a Markov process due to the iid assumptions on the input. Denote the corresponding continuous-time process (also Markov because of the Poisson arrivals) by $\mathbf{V}(t) = (V(t, 1), \ldots, V(t, i))$, where $V(t, i)$ denotes the workload at the ith node at time $t \geq 0$, and $\mathbf{V}_n = \mathbf{V}(t_n-),\ n \geq 1$. From *Poisson arrivals see time averages* (*PASTA*) (see [7], and Theorem 6.7, p. 218 in [1]), the limiting stationary distribution of $\mathbf{V}_n$, as $n \to \infty$, is identical with that of $\mathbf{V}(t)$, as $t \to \infty$. But the coordinates of $\mathbf{V}(t)$, namely $V(t, 1), \ldots, V(t, i)$, are iid copies of workload for the M/G/1 queue (as was pointed out in the beginning of this section). Thus, the *joint* time-stationary distribution of workload is given by

$$\big(D(1), \ldots, D(c)\big), \tag{7}$$

where the $D(i)$ here are iid distributed as D in (5).

We conclude that the stationary distribution for $\{\mathbf{V}_n : n \geq 0\}$ is the same as in (7) and thus *the proportion of arrivals who find the RA system empty is given by* $P(D = 0)^c = (1 - \rho_1)^c > 0$*; visits to the empty state constitute positive recurrent regeneration points;* $E(T) < \infty$.

With $\mathbf{V}_0 = \mathbf{0}$, we have an identically distributed version of a cycle length (4) given by $T = \min\{n \geq 1 : \mathbf{V}_n = \mathbf{0}\}$. $\{\mathbf{V}_n : n \geq 0\}$ is a Markov process, so if we start it off with $\mathbf{V_0}$ distributed as in (7), then the process will be a stationary version, denoted by $\{\mathbf{V}_n^* : n \geq 0\}$. As was explained in the beginning of this section, we conclude that $T^e = \min\{n \geq 1 : \mathbf{V}_n^* = 0\}$.

The above analysis immediately leads to the following algorithm for simulating T^e, where we only need to assume that we can simulate from both G and G_e.

Algorithm for simulating T^e

1. Initialize $\mathbf{V}_0 = (D(1), \ldots, D(c))$ as distributed as in (7).

2. Simulate sequentially $\{\mathbf{V}_n : n \geq 1\}$ using the recursion in (6) until time

$$T^e = \min\{n \geq 1 : \mathbf{V}_n = 0\}.$$

6 The algorithm

We only need to assume that we can simulate from both G and G_e.

Algorithm for simulating W distributed as the stationary distribution π for the stable FIFO M/G/c queue

1. Simulate a copy of T^e using the algorithm from Sect. 5. Set $k = T^e$.
2. Independently generate T using the method in Sect. 4.
3. If $T < k$, then go back to step (2).
4. Construct the FIFO cycle $C = \{\mathbf{W}_1, \ldots, \mathbf{W}_T\}$ as described (at the end) in Sect. 4. Set $\mathbf{W} = \mathbf{W}_k$.
5. Output $\mathbf{W}$.

Remark 6.1 Our assumption that we can simulate from G and G_e is not very restrictive in applications. G is a basic primitive of the model; one would not be able to simulate the M/G/c queue at all (even starting empty) if service times could not be simulated. One would typically know explicitly what G is (exponential, Erlang, phase-type, Pareto, etc.). One would also typically have significant further information about G, such as several of its moments, or reasonable bounds on them or bounds on G itself, in which case simulation from G_e is indeed possible, using, for example, acceptance-rejection methods. (See [2], Corollary 8.2 and Example 8.6, pp. 421–422 for further details.)

References

1. Asmussen, S.: Applied Probability and Queues, 2nd edn. Springer, New York (2003)
2. Asmussen, S., Glynn, P.W.: Stochastic Simulation. Springer, New York (2007)
3. Asmussen, S., Glynn, P.W., Thorisson, H.: Stationary detection in the initial transient problem. ACM Trans. Model. Comput. Simul. **2**, 130–157 (1992)
4. Foss, S.G.: Approximation of mutichannel queueing systems. Sib. Math. J. **21**, 132–140 (1980)
5. Foss, S.G., Chernova, N.I.: On optimality of the FCFS discipline in mutiserver queueing systems and networks. Sib. Math. J. **42**, 372–385 (2001)
6. Sigman, K.: Exact simulation of the stationary distribution of the FIFO M/G/c queue. J. Appl. Probab. **48A**, 209–216 (2011). Special Volume: New Frontiers in Applied Probability
7. Wolff, R.W.: Poisson arrivals see time averages. Oper. Res. **30**, 223–231 (1982)
8. Wolff, R.W.: Upper bounds on work in system for multi-channel queues. J. Appl. Probab. **14**, 547–551 (1987)