J. Appl. Prob. Spec. Vol. **48A,** 209–213 (2011)

NEW FRONTIERS IN APPLIED PROBABILITY

A Festschrift for SØREN ASMUSSEN
Edited by P. GLYNN, T. MIKOSCH and T. ROLSKI

Part 4. Simulation

EXACT SIMULATION OF THE STATIONARY DISTRIBUTION OF THE FIFO M/G/c QUEUE

KARL SIGMAN, *Columbia University*

Department of Industrial Engineering and Operations Research, Columbia University, MC: 4704, New York, NY 10027, USA. Email address: karl.sigman@columbia.edu

APPLIED PROBABILITY TRUST
AUGUST 2011

https://doi.org/10.1239/jap/1318940466 Published online by Cambridge University Press

EXACT SIMULATION OF THE STATIONARY DISTRIBUTION OF THE FIFO M/G/c QUEUE

BY KARL SIGMAN

Abstract

We present an exact simulation algorithm for the stationary distribution of the customer delay D for first-in–first-out (FIFO) M/G/c queues in which $\rho = \lambda/\mu < 1$. We assume that the service time distribution $G(x) = \mathrm{P}(S \le x),\ x \ge 0$ (with mean $0 < \mathrm{E}(S) = 1/\mu < \infty$), and its corresponding equilibrium distribution $G_{\mathrm{e}}(x) = \mu \int_0^x \mathrm{P}(S > y)\,\mathrm{d}y$ are such that samples of them can be simulated. We further assume that G has a finite second moment. Our method involves the general method of dominated coupling from the past (DCFTP) and we use the single-server M/G/1 queue operating under the processor sharing discipline as an upper bound. Our algorithm yields the stationary distribution of the entire Kiefer–Wolfowitz workload process, the first coordinate of which is D. Extensions of the method to handle simulating generalized Jackson networks in stationarity are also remarked upon.

Keywords: Exact simulation; coupling from the past; processor sharing; queueing theory

2010 Mathematics Subject Classification: Primary 65C05
Secondary 60K25; 60J05; 68U20

1. Introduction

Consider a stable first-in–first-out (FIFO) M/G/c queue ($c \ge 2$) with arrival rate λ, Poisson arrival times $\{t_n : n \ge 1\}$, in which the independent and identically distributed (i.i.d.) service times $\{S_n\}$ are distributed as $G(x) = \mathrm{P}(S \le x),\ x \ge 0$, with tail denoted by $\bar{G}(x) = 1 - G(x)$, and finite mean $\mathrm{E}(S) = 1/\mu$.

Let $\boldsymbol{V}(t) = (V(1,t), V(2,t), \dots, V(c,t)),\ t \ge 0$, denote the *Kiefer–Wolfowitz workload vector* (see, for example, Chapter 12 of [1]).

At arrival epochs t_n with i.i.d. interarrival times $T_n = t_n - t_{n-1}$ ($t_0 := 0$), the vector $\boldsymbol{W}_n = \boldsymbol{V}(t_n-)$ satisfies the recursion

$$\boldsymbol{W}_n = R(\boldsymbol{W}_{n-1} + S_n \boldsymbol{e} - T_n \boldsymbol{f})^+, \qquad n \ge 1, \tag{1.1}$$

where $\boldsymbol{W}_n = (W_n(1), \dots, W_n(c))$, $\boldsymbol{e} = (1, 0, \dots, 0)$, $\boldsymbol{f} = (1, 1, \dots, 1)$, R places a vector in ascending order, and $(\cdot)^+$ denotes the positive part of each coordinate. Then $D_n = W_n(1)$ is the delay in queue (line) of the nth customer. (Equation (1.1) defines a Markov chain owing to the given i.i.d. assumptions.) With $\rho := \lambda/\mu < c$ (stability), it is well known that $\boldsymbol{W}_n$ converges in distribution to a proper stationary distribution. From Poisson arrivals see time averages (PASTA), this coincides with the time-stationary limiting distribution of $\boldsymbol{V}(t)$ as $t \to \infty$. Let π denote this stationary distribution. Our objective in the present paper is to provide a simulation algorithm for sampling exactly from π. Our method involves using an upper bound that can be simulated in reverse time in the spirit of a general *dominated-coupling-from-the-past*

https://doi.org/10.1239/jap/1318940466 Published online by Cambridge University Press

method (DCFTP method), as found, for example, in Definition 3 and Algorithm 4 of [4]. Such methods build upon the original *coupling-from-the-past method* (CFTP method) as introduced in [6]. (See also the introduction to *perfect sampling* in Chapter 4, Section 8 of [2].)

1.1. The upper bound

We will assume that $\rho < 1$; the system is *super stable*. (We will also assume that $\mathrm{E}(S^2) < \infty$ for reasons explained later.) Under the $\rho < 1$ condition, the corresponding single-server FIFO M/G/1 model is also stable, and, as is well known (we include a proof only for completeness) serves as a sample path upper bound for the workload.

Proposition 1.1. *Let $V_1(t)$ denote the total work done in the system at time t for the FIFO M/G/1 queue, and let $V_c(t) = \sum_{i=1}^{c} V(i,t)$ denote the total work in the system at time t for the corresponding FIFO M/G/c queue, where $V_1(0) = V_c(0) = 0$ and both are fed exactly the same input of Poisson arrivals and i.i.d. service times. Then*

$$\mathrm{P}(V_c(t) \le V_1(t) \textit{ for all } t \ge 0) = 1. \tag{1.2}$$

Proof. The work at time t is equal to the work that has arrived by time t minus the work processed by time t. The amount of work to arrive by time t,

$$\sum_{j=1}^{N(t)} S_j, \qquad t \ge 0,$$

where $N(t)$ is the number of arrivals by time t, is the same for both queues. But the rate at which the work is processed is 1 for the single-server queue, and between 1 and c for the c-server queue; hence, (1.2) holds.

1.2. Using the upper bound

If we were to start off with both the c-server and the single-server models empty at time $t = -\infty$ while feeding them exactly the same input, then both would have their stationary distributions at time 0 (via Loynes' lemma—see [5, Lemma 1]) and their workloads would be ordered at all times owing to Proposition 1.1. Moreover, if we walk backwards in time from the origin, and detect the first time $-\tau \le 0$ at which the single-server model is empty, then, from Proposition 1.1, the c-server model would be empty as well. We could then construct a sample of $\boldsymbol{V}(0)$ (having the stationary distribution π) by starting off empty at time $-\tau$ and using recursion (1.1) forwards in time from time $-\tau$ to 0. We now proceed to show how to accomplish this.

1.3. The algorithm

From the Pollaczek–Khintchine formula we know that the stationary workload for the M/G/1 queue can be written in distribution as

$$\sum_{j=1}^{Q} Y_j, \tag{1.3}$$

where the $\{Y_j\}$ are i.i.d. as the equilibrium distribution of service, with cumulative distribution function given by

$$G_{\mathrm{e}}(x) = \mu \int_0^x \mathrm{P}(S > y)\,\mathrm{d}y, \qquad x \ge 0,$$

and, independently, Q has a geometric distribution, $\mathrm{P}(Q = k) = \rho^k(1-\rho)$, $k \ge 0$.

https://doi.org/10.1239/jap/1318940466 Published online by Cambridge University Press

The workload for a single-server queue is invariant under changes of work-conserving disciplines, so here we will use processor sharing (PS) (see, for example, Section 5.7.3 of [7]). The point is that $\{V_1(t)\}$ has exactly the same sample paths under PS as it does under FIFO. Letting $Q(t)$ denote the number of customers in the system, and $Y_1(t), \dots, Y_{Q(t)}(t)$ denote the corresponding remaining service times, it is known that the stationary distribution of the PS M/G/1 queue, as $t \to \infty$, is of the form

$$(Q, Y_1, \dots, Y_Q), \tag{1.4}$$

where the Q and the $\{Y_j\}$ are as in (1.3). Actually, the $Y_j(t)$ can be modeled as either the age of the service times in service, or the remaining service times, and in either case

$$\{\boldsymbol{X}(t) \colon t \geq 0\} = \{(Q(t), Y_1(t), \dots, Y_{Q(t)}(t)) \colon t \geq 0\} \tag{1.5}$$

forms a Markov process (where we usually assume that the $Y_j(t)$ are in ascending order). If we assume that they are the remaining service times then it is also known that if the system is initially distributed as in (1.4), then the time reversal of $\{\boldsymbol{X}(t)\}$ is the same Markovian PS model in which arrivals are Poisson at rate λ, service times are i.i.d. as G, and the $Y_j(t)$ are now the ages (see, for example, [3, Section 4.2.2] and [7, pp. 278–280]).

The point here is that unlike the FIFO model, we can easily simulate the time reversal of the PS model, since it is yet again the same kind of queue, and we will use this fact as follows.

Algorithm 1.1. (*Simulating* $\boldsymbol{V}(0)$ *distributed as* π.)

1. Set $t = 0$ (time). Generate a vector $(Q, Y_1, \dots, Y_Q)$ distributed as the stationary distribution in (1.4), and set

$$\boldsymbol{X}(0) = (Q(0), Y_1(0), \dots, Y_{Q(0)}(0)) = (Q, Y_1, \dots, Y_Q).$$

If $Q = 0$ then stop simulating and set $\tau = 0$. Otherwise, continue to simulate (as a discrete-event simulation with i.i.d. interarrival times $T \sim \exp(\lambda)$ and i.i.d. service times $S \sim G$) the PS model in (1.5) in forward time $t \geq 0$ until time $\tau = \min\{t \geq 0 \colon Q(t) = 0\}$. The $Q > 0$ customers are served simultaneously at rate $r = 1/Q$ until the time of the *next event*: either a new arrival or a departure; reset t to be this new time.

If the next event is an arrival then generate a service time S for this customer distributed as G (keep a record of its value and place it in service), generate the next interarrival time T distributed as $\exp(\lambda)$, and reset $Q = Q + 1$ and set $r = 1/Q$.

If the next event is a departure then record this as the next departure time and record the service time of the customer associated with it, and reset $Q = Q - 1$. If $Q = 0$ then stop simulating and set $\tau = t$.

If $\tau > 0$ after stopping the simulation then let $t_1, \dots, t_k$ and $S_1, \dots, S_k$ respectively denote the $k \geq 1$ recorded departure times (in order of departure) and the associated service times that occurred up to time τ (with $t_k = \tau$ being the last such departure time). Define the interdeparture times $T_i = t_i - t_{i-1}$, $0 \leq i \leq k$, with $t_0 = 0$.

We view this as simulating the time reversal of the original PS model from time 0, into the past until time $t = -\tau$. In particular, owing to the discussion in Section 1.2, we have $\boldsymbol{V}(-\tau) = \boldsymbol{0}$. (See Remark 1.1 below for more details.)

2. We now construct $\boldsymbol{V}(0)$ as follows. If $\tau = 0$ then set $\boldsymbol{V}(0) = \boldsymbol{0}$. Otherwise, reset $(S_1, \dots, S_k) = (S_k, \dots, S_1)$ and $(T_1, \dots, T_k) = (T_k, \dots, T_1)$ (that is, place them in

https://doi.org/10.1239/jap/1318940466 Published online by Cambridge University Press

reverse order). (They have the *conditional* distribution of i.i.d. input given τ resulted in k departures, so they are no longer i.i.d.) Using $(S_1, \dots, S_k)$ and $(T_1, \dots, T_k)$ as the input, construct $\boldsymbol{W}_k$ (initializing with $\boldsymbol{W}_0 = \boldsymbol{0}$), by using (1.1) recursively from $n = 1$ up until $n = k$. Now set $\boldsymbol{V}(0) = \boldsymbol{W}_k$.

We assume that $\mathrm{E}(S^2) < \infty$ to ensure that $\mathrm{E}(\tau) < \infty$: τ (conditional on $\tau > 0$) has the stationary *excess* distribution of a busy period B for the M/G/1 queue; thus, $\mathrm{E}(\tau) = \rho\,\mathrm{E}(B^2)/2\,\mathrm{E}(B)$, and it is well known that $\mathrm{E}(B^2) < \infty$ if and only if $\mathrm{E}(S^2) < \infty$. (So if $\mathrm{E}(S^2) = \infty$ then although the algorithm still works, the expected length of time to completion is infinite.)

Remark 1.1. To simulate from the stationary distribution in (1.4), we first generate the geometric Q and then we assume that we can simulate from G_{e} so as to generate Q i.i.d. such copies of the Y_j. We actually need to simulate from the stationary *spread* distribution (having tail $\mathrm{P}(H > x) = \mu x \bar{G}(x) + \bar{G}_{\mathrm{e}}(x)$), or, equivalently, from the joint stationary distribution of age (A) and excess (E) given by $\mathrm{P}(A > x,\ E > y) = \bar{G}_{\mathrm{e}}(x + y)$; $H = A + E$ has the stationary spread distribution. This is because the Q customers at time 0 have service times distributed as H (not G) and we need to take that into account. If we directly simulate a copy H having the stationary spread distribution then by independently generating a U uniform over $(0, 1)$ we can use $A = UH$ and $E = (1 - U)H$. Thus, we could generate i.i.d. H_i and i.i.d. U_i (uniforms), and set $Y_i = U_i H_i$ while keeping a record of the whole service times H_i.

The point is that we will need, initially, the spread and age of each service time at time 0. After that, to simulate new arrivals, we will need only i.i.d. copies of $S \sim G$ and i.i.d. copies of $T \sim \exp(\lambda)$. We keep a record of each new customer's generated service time so that when this customer departs, we know their service time.

Remark 1.2. Exact simulation for a variety of other queueing models can be carried out in a similar fashion by using the PS model as an upper bound; we provide here an example. Consider a generalized Jackson network (with c single-server FIFO nodes), in which arrivals are Poisson at rate λ, service times at node i are i.i.d. as G_i, and routeing is general i.i.d. among customers. Letting $(S(1), i_1, \dots, S(l), i_l)$ denote the service times and route of a customer, we define $S = \sum_{i=1}^{l} S(i)$ as a 'service time' for the single-server PS model; we require the harsh condition that $\rho = \lambda\,\mathrm{E}(S) < 1$ to ensure stability of the PS model. Then the total work is sample path bounded above by the PS model.

Remark 1.3. Clearly, it is of interest to allow $1 \le \rho < c$, and still obtain an exact simulation algorithm in the same spirit as was done here for $\rho < 1$. As a 'first try', we might use, as an upper bound, the PS *random assignment* (RA) M/G/c, in which each server has its own PS queue, and arrivals are randomly assigned to a queue (probability $1/c$ to each). For then, each server is an independent PS M/G/1 queue, the time reversal is the same kind of model, and so on. However, the RA model does not serve as a *sample path* upper bound for the FIFO M/G/c as in Proposition 1.1; counterexamples exist (see [1, pp. 342–344] and [8]). (It does serve as a *stochastic* upper bound for each fixed time t.) The author is currently working on this for a future paper.

References

[1] Asmussen, S. (2003). *Applied Probability and Queues*, 2nd edn. Springer, New York.

[2] Asmussen, S. and Glynn, P. W. (2007). *Stochastic Simulation*. Springer, New York.

[3] Chen, H. and Yao, D. D. (2001). *Fundamentals of Queueing Networks*. Springer, New York.

[4] Kendall, W. (2004). Geometric ergodicity and perfect simulation. *Electron. Commun. Prob.* **9,** 140–151.

https://doi.org/10.1239/jap/1318940466 Published online by Cambridge University Press

[5] Loynes, R. M. (1962). The stability of a queue with non-independent interarrival and service times. *Math. Proc. Camb. Phil. Soc.* **58,** 497–520.

[6] Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms* **9,** 223–252.

[7] Ross, S. M. (1996). *Stochastic Processes*, 2nd edn. John Wiley, New York.

[8] Wolff, R. W. (1987). Upper bounds on work in system for multichannel queues. *J. Appl. Prob.* **24,** 547–551.

KARL SIGMAN, *Columbia University*

Department of Industrial Engineering and Operations Research, Columbia University, MC: 4704, New York, NY 10027, USA. Email address: karl.sigman@columbia.edu

https://doi.org/10.1239/jap/1318940466 Published online by Cambridge University Press