J. Appl. Prob. **45,** 568–574 (2008)
Printed in England

RELATING TIME AND CUSTOMER AVERAGES FOR QUEUES USING 'FORWARD' COUPLING FROM THE PAST

EROL A. PEKÖZ,* *Boston University*
SHELDON M. ROSS,** *University of Southern California*

Abstract

We give a new method for simulating the time average steady-state distribution of a continuous-time queueing system, by extending a 'read-once' or 'forward' version of the coupling from the past (CFTP) algorithm developed for discrete-time Markov chains. We then use this to give a new proof of the 'Poisson arrivals see time averages' (PASTA) property, and a new proof for why renewal arrivals see either stochastically smaller or larger congestion than the time average if interarrival times are respectively new better than used in expectation (NBUE) or new worse than used in expectation (NWUE).

Keywords: PASTA; coupling from the past; simulation; queues

2000 Mathematics Subject Classification: Primary 60K25; 68U20
Secondary 60J27; 60J2S; 60J10; 90B22

1. Introduction

In a queueing system the fraction of arrivals finding the queue in some state is not necessarily the same as the fraction of time the queue is in that state. For example, if arrivals occur in infrequent but heavy bursts, most customers can see a lot of congestion even though the system is usually empty. The celebrated 'Poisson arrivals see time averages' (PASTA) property of queueing theory, attributed to Wolff (1982), says that with Poisson arrivals the fraction of arrivals finding a queue in some state equals the fraction of time the queue is in that state.

There is a very large literature on the PASTA property. Wolff's (1982) article has nearly 300 citations listed in the *Web of Science* citation index, making it the the third most cited article in *Operations Research* since 1980 and among the top 10 most frequently cited articles in the history of the journal. For some entry points to this literature, see, for example, Melamed and Whitt (1990a), Melamed and Whitt (1990b), Melamed and Yao (1995), Brèmaud *et al.* (1992), Heyman and Stidham (1980), Köning and Schmidt (1980), (1981), Shanthikumar and Zazanis (1999), and the references therein.

In this article we present new proofs relating time averages and customer averages in queues by extending a version of the coupling from the past (CFTP) idea, first attributed to Propp and Wilson (1996), used to exactly sample the stationary distribution of a Markov chain. Propp and Wilson's original 1996 article in *Random Structures and Algorithms*, according to the

Received 22 November 2006; revision received 24 April 2007.
* Postal address: Department of Operations and Technology Management, Boston University, 595 Commonwealth Avenue, Boston, MA 02215, USA. Email address: pekoz@bu.edu
** Postal address: Department of Industrial and System Engineering, University of Southern California, Los Angeles, CA 90089, USA.

Web of Science citation index, has 162 citations and, thus, is the most frequently cited article in the history of that journal. The more recent article Wilson (2000) gives a 'read-once' or 'forward' CFTP technique for exactly sampling the stationary distribution of a discrete-time Markov chain. Other approaches for exact sampling in continuous time appear in Corcoran and Tweedie (2001), and Foss *et al.* (1998). Additional material on 'forward' CFTP can be found in Ross and Peköz (2007).

The organization of this article is as follows. In Section 2 we present our main result, which is an extension of the 'read-once' or 'forward' CFTP technique to continuous time. In Section 3 we apply this result to relating customer and time averages, and in particular, we give new proofs of the PASTA property and of why renewal arrivals see either stochastically smaller or larger congestion than the time average if interarrival times are respectively new better than used in expectation (NBUE) or new worse than used in expectation (NWUE).

2. Simulating the time average queue

Consider a queueing system having a renewal customer arrival process, and let $Q(t)$ be the state of the system at time t, immediately prior to any arrival which may occur at that time. Let $X_1, X_2, \ldots$ be independent and identically distributed customer interarrival times (denoted generically as X with cumulative distribution function F), and let $T_n = \sum_{i=1}^{n} X_i$ be the time of the nth customer arrival. We suppose that the state of the system encodes the amount of time each customer has been in service, along with their positions in the system, so that $Q(T_n)$ is a Markov chain, where $Q(T_n)$ is the state seen by arrival n—that is, it is the state of the system immediately before customer n arrives. Let S denote the state space of this Markov chain, and let $0 \in S$ correspond to the system being empty.

Suppose that, for some $p > 0$, we have

$$P_{s,0} \equiv \mathrm{P}(Q(T_{n+1}) = 0 \mid Q(T_n) = s) \geq p \quad \text{for all states } s.$$

This condition will hold if interarrival times are unbounded and the total amount of work allowed in the system is bounded. This condition will also hold in the special setting of 'exploding' customers, defined and used in the next section, where after the arrival of each customer, the system is instantly cleared of all customers with probability p (independent of all else) and enters state 0 immediately. It also holds in a finite-capacity system where the interarrival times are unbounded and the failure rate function of the service distribution is bounded (or, more generally, if there is a distribution G that, for all t, is stochastically larger than the distribution of the remaining service time of a customer who has already spent t time units in service).

We now construct a sequence $J_n,\ n \geq 1$, of independent and identically distributed Bernoulli random variables with parameter p, that is, such that whenever $J_n = 1$ then $Q(T_n) = 0$. To do so, let $U_i,\ i \geq 1$, be a sequence of independent uniform $(0, 1)$ random variables that is generated independent of the queueing system. Let

$$J_n = \max_{s \in S} I\left\{ Q(T_{n-1}) = s,\ Q(T_n) = 0,\ U_n < \frac{p}{P_{s,0}} \right\},$$

where we use the notation $I\{B\}$ for the indicator variable for the event B. That is, $J_n = 1$ if, for some state s, the state seen by arrival number $n - 1$ is s, the state seen by the next arrival is 0, and the corresponding uniform is less than $p/P_{s,0}$.

We say that the random variable Z taking values in S has the time average steady-state distribution if

$$\mathrm{P}(Z \in A) = \lim_{t\to\infty} \frac{1}{t} \int_0^t I\{Q(s) \in A\}\,\mathrm{d}s$$

for any set of states A.

Let Y be a random variable that is independent of all the other random variables and which has the equilibrium distribution for the renewal process. That is,

$$\mathrm{P}(Y \le y) = \frac{1}{\mathrm{E}[X]} \int_0^y \bar{F}(x)\,\mathrm{d}x,$$

where $\bar{F}$ is the tail distribution function. Note that the density function for Y is

$$g(x) = \frac{\bar{F}(x)}{\mathrm{E}[X]}.$$

Suppose that the system is empty at time 0 and say that a new cycle begins at time T_N, where $N = \min(n > 0\colon J_n = 1)$ denotes the number of arrivals in the first cycle. Also, let $Q_n(t)$ be a random variable, independent of all else, whose distribution is the conditional distribution of $Q(T_n + t)$ given that $T_{n+1} > T_n + t$ and $N \ge n$. Note that this means that $Q_n(t)$ is independent of both N and $Q(t)$, though its distribution is defined in terms of the distributions of $Q(t)$ and N.

Proposition 1. *With the above definitions, the variable $Q_{N-1}(Y)$ has the time average steady-state distribution.*

Remark 1. The variable $Q_{N-1}(Y)$ can be simulated as follows. Simulate the queue until a cycle ends, let $s = Q(T_{N-1})$, generate a random variable Y having the equilibrium distribution, start a completely independent simulation of the queueing system from state s at time 0 without any arrivals, and output the state of this independent simulation at time Y. Note that this is not the same as simply outputting the state $Q(T_{N-1} + Y)$, since $Q_{n-1}(t)$ is defined to be independent of all else (including N) and, thus, N does not indicate the end of a cycle in the system $Q_{n-1}(t)$.

Remark 2. The preceding result can be viewed as an extension to continuous time of the 'read-once' or 'forward' CFTP technique for discrete-time Markov chains given in Wilson (2000).

Proof of Proposition 1. Suppose that the system is empty at time 0 and say that a new cycle begins each time an exploding customer arrives. Let $N = \min(n > 0\colon J_n = 1)$, so that the first cycle is of length T_N. Letting

$$R_A = \int_0^{T_N} I\{Q(s) \in A\}\,\mathrm{d}s$$

be the amount of time spent in A during the first cycle, by the renewal reward theorem we have

$$\mathrm{P}(Z \in A) = \frac{\mathrm{E}[R_A]}{\mathrm{E}[T_N]}.$$

Because N is a stopping time for the sequence of interarrival times, Wald's equation gives

$$\mathrm{E}[T_N] = \mathrm{E}[X]\,\mathrm{E}[N] = \frac{\mathrm{E}[X]}{p}.$$

Now, let $Q_n(t)$ be a random variable whose distribution is the conditional distribution of $Q(T_n + t)$ given that $T_{n+1} > T_n + t$ and $N \geq n$. Then, with

$$F^{-1}(x) \equiv \inf\{t\colon F(t) \geq x\},$$

$M \sim$ geometric(p) independent of all else, and $U \sim U(0, 1)$,

$$\begin{aligned}
\mathrm{E}[R_A] &= \mathrm{E}\left[\int_0^{T_N} I\{Q(t) \in A\}\,\mathrm{d}t\right] \\
&= \mathrm{E}\left[\sum_{n=1}^{N}\int_0^{X_n} I\{Q(T_{n-1}+t) \in A\}\,\mathrm{d}t\right] \\
&= \mathrm{E}\left[\sum_{n=1}^{\infty} I\{N \geq n\}\int_0^{X_n} I\{Q(T_{n-1}+t) \in A\}\,\mathrm{d}t\right] \\
&= \sum_{n=1}^{\infty}\mathrm{E}\left[\int_0^{X_n} I\{Q(T_{n-1}+t) \in A\}\,\mathrm{d}t \mid N \geq n\right]\mathrm{P}(N \geq n) \\
&= \sum_n \mathrm{P}(N \geq n)\,\mathrm{E}\left[\mathrm{E}\left[\int_0^{X_n} I\{Q(T_{n-1}+t) \in A\}\,\mathrm{d}t \mid X_n,\ N \geq n\right] \,\middle|\, N \geq n\right] \\
&= \sum_n \mathrm{P}(N \geq n)\,\mathrm{E}\left[\mathrm{E}\left[\int_0^{F^{-1}(U)} I\{Q(T_{n-1}+t) \in A\}\,\mathrm{d}t \mid U,\right.\right. \\
&\qquad\qquad \left.\left. X_n = F^{-1}(U),\ N \geq n\right] \,\middle|\, N \geq n\right] \\
&= \sum_n \mathrm{P}(N \geq n)\int_0^1\int_0^{F^{-1}(x)} \mathrm{P}(Q_{n-1}(t) \in A)\,\mathrm{d}t\,\mathrm{d}x \\
&= \sum_n \mathrm{P}(N \geq n)\int_0^{\infty}\int_{F(t)}^1 \mathrm{P}(Q_{n-1}(t) \in A)\,\mathrm{d}x\,\mathrm{d}t \\
&= \sum_n \mathrm{P}(N \geq n)\int_0^{\infty} \bar{F}(t)\,\mathrm{P}(Q_{n-1}(t) \in A)\,\mathrm{d}t \\
&= \sum_n \mathrm{P}(N \geq n)\,\mathrm{E}[X]\int_0^{\infty} g(t)\,\mathrm{P}(Q_{n-1}(t) \in A)\,\mathrm{d}t \\
&= \sum_n \mathrm{P}(N \geq n)\,\mathrm{E}[X]\,\mathrm{P}(Q_{n-1}(Y) \in A) \\
&= \sum_n \frac{1}{p}\,\mathrm{P}(N = n)\,\mathrm{E}[X]\,\mathrm{P}(Q_{n-1}(Y) \in A) \\
&= \frac{\mathrm{P}(Q_{M-1}(Y) \in A)\,\mathrm{E}[X]}{p} \\
&= \frac{\mathrm{P}(Q_{N-1}(Y) \in A)\,\mathrm{E}[X]}{p},
\end{aligned}$$

and the result is proved. The last line follows because $Q_n(t)$ is defined to be independent of all else, including N.

3. Relating time and customer averages

We say that the random variable W taking values in S has the customer average steady-state distribution if

$$\mathrm{P}(W \in A) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I\{Q(T_i) \in A\}.$$

Using a parameter $p \geq 0$, suppose that each customer, independently of all else, with probability p is labeled as an exploding customer. Immediately after the arrival of such a customer, the system is instantly cleared of all customers and enters state 0. Now let J_n be the indicator variable equal to 1 if customer n is an exploding customer. If $J_n = 1$ then $Q(T_n)$ is the state seen by an exploding arrival (e.g. the system state immediately before the exploding arrival clears the system). Although our main interest is when $p = 0$, we will first consider the case when $p > 0$.

Although our next result can be deduced as a corollary from our preceding proposition, we give an alternative self-contained proof for simplicity.

Proposition 2. *If $p > 0$ then $Q(T_N)$ has the customer average steady-state distribution.*

Proof. For any $A \subset S$, let π_A be the steady state probability that the state of the Markov chain is in A. Imagine that an exploding customer arrived at time 0, and let a new cycle begin at each arrival that immediately follows an exploding customer. So the initial cycle begins when the first customer arrives and the next cycle begins when customer $N + 1$ arrives, where, as always, $N = \min(n > 0\colon J_n = 1)$. Let N_A denote the number of time periods the Markov chain is in state A during this cycle. That is,

$$N_A = \sum_{n=1}^{N} I\{Q(T_n) \in A\}.$$

If we suppose that a reward of 1 is earned each time the state of the chain lies in A then the renewal reward theorem yields

$$\pi_A = \frac{\mathrm{E}[N_A]}{\mathrm{E}[N]} = p\,\mathrm{E}[N_A].$$

Now, let I_k be the indicator variable for the event that a new cycle begins on the transition following the kth visit to the set of states A. Note that

$$\sum_{k=1}^{N_A} I_k = I\{Q(T_N) \in A\}.$$

Because $I_1, I_2, \ldots$ are independent and identically distributed and the event $\{N_A = n\}$ is independent of $I_{n+1}, I_{n+2}, \ldots$, it follows from Wald's equation that

$$\mathrm{P}(Q(T_N) \in A) = \mathrm{E}[N_A]\,\mathrm{E}[I_1] = p\,\mathrm{E}[N_A].$$

This completes the proof.

Remark 3. Proposition 2 is a version of the 'read-once' or 'forward' CFTP technique which appears in Wilson (2000), and allows perfect simulation of the stationary distribution of a discrete-time Markov chain. Some connections to the PASTA property were proposed and discussed there as well. A more general version of Proposition 2 is proven in Ross and Peköz (2007, Example 6.9).

Our next result relates time and customer averages for queues with exploding customers.

Corollary 1. *Suppose that* $p > 0$.

(i) *If* X *has an exponential distribution then* $W \stackrel{\mathrm{D}}{=} Z$, *where* '$\stackrel{\mathrm{D}}{=}$' *denotes equality in distribution. In other words, if arrivals are according to a Poisson process then the customer average steady-state distribution is the same as the time average steady-state distribution.*

(ii) *Let* $G(s)$, $s \in S$, *be a function of the system state having the property that, for some partial ordering, the value of* $G(Q(t))$ *decreases with respect to that partial ordering at all times at which customers do not arrive. Then* $G(W) \leq_{\mathrm{st}} G(Z)$ *or* $G(W) \geq_{\mathrm{st}} G(Z)$ *when the interarrival distribution* X *is NBUE or, respectively, NWUE. In other words, the customer average steady-state distribution of* $G(Q)$ *is stochastically smaller than or stochastically larger than the time average steady-state distribution of* $G(Q)$ *when the interarrival distribution is NBUE or, respectively, NWUE.*

Proof. It follows from Proposition 2 that the customer average steady-state distribution is the distribution of the system state as seen by exploding arrivals, which is distributed as the system state as seen by customers immediately preceding an exploding arrival after a random interarrival time during which there are no new arrivals. On the other hand, from Proposition 1, the time average steady-state distribution is the distribution of the system state as seen by customers immediately preceding an exploding arrival after a random time, distributed according to the equilibrium distribution, during which there are no new arrivals. In the exploding customers model, each interarrival time is independent of the type of customer (exploding or nonexploding). Then, because the equilibrium distribution is the interarrival distribution for a Poisson process, the PASTA result, part (i), is proven. Also, because the equilibrium distribution is stochastically smaller than F when F is NBUE and is stochastically larger when F is NWUE, part (ii) follows.

We finally present the main result of this section, which relates customer and time averages for general queues.

Corollary 2. *Corollary 1 holds if* $p = 0$, *under the stability condition that state* 0 *is positive recurrent.*

Proof. Fix some $A \subset S$, and, for some given value of p, let a_p and b_p denote the probability that the system is in the set of states A respectively under the time average and the customer average steady-state distributions. We will show that $\lim_{p\to 0} a_p = a_0$ and $\lim_{p\to 0} b_p = b_0$, which along with Corollary 1 gives the result.

Say a new cycle begins whenever a customer arrives to find the system in state 0. Let M_p be the total time the system is in the set of states A during a cycle, and let L_p be the length of a cycle. Clearly, $M_p \leq_{\mathrm{st}} L_p \leq_{\mathrm{st}} L_0$, and the stability condition gives $\mathrm{E}[L_0] < \infty$. By the renewal reward theorem we must have $a_p = \mathrm{E}[M_p]/\mathrm{E}[L_p]$, and we can then use the Lebesgue dominated convergence theorem to obtain $\lim_{p\to 0} \mathrm{E}[M_p] = \mathrm{E}[M_0]$ and $\lim_{p\to 0} \mathrm{E}[L_p] = \mathrm{E}[L_0]$. This gives $\lim_{p\to 0} a_p = a_0$.

The same argument can be applied to obtain $\lim_{p\to 0} b_p = b_0$ if we write $b_p = \mathrm{E}[M'_p]/\mathrm{E}[L'_p]$, where M'_p is the number of arrivals finding the system in the set of states A during a cycle, and L'_p is the number of arrivals during a cycle. This then establishes the corollary.

Remark 4. Corollary 2, proved using different methods, appears in Shanthikumar and Zazanis (1999) and Niu (1984).

Acknowledgements

We would like to thank José Blanchet and Rhonda Righter for their very insightful comments.

References

[1] Brèmaud, P., Kannurpatti, R. and Mazumdar, R. (1992). Event and time averages: a review. *Adv. Appl. Prob.* **24,** 377–411.
[2] Corcoran, J. N. and Tweedie, R. L. (2001). Perfect sampling of ergodic Harris chains. *Ann. Appl. Prob.* **11,** 438–451.
[3] Foss, S. G., Tweedie, R. L. and Corcoran, J. N. (1998). Simulating the invariant measures of Markov chains using backward coupling at regeneration times. *Prob. Eng. Inf. Sci.* **12,** 303–320.
[4] Heyman, D. P. and Stidham, S., Jr. (1980). The relation between customer and time averages in queues. *Operat. Res.* **28,** 983–994.
[5] Köning, D. and Schmidt, V. (1980). Stochastic inequalities between customer-stationary and time-stationary characteristic of queueing systems with point processes. *J. Appl. Prob.* **17,** 768–777.
[6] Köning, D. and Schmidt, V. (1981). Relationships between time- and customer-stationary characteristics of service systems. In *Point Processes and Queueing Problems*, eds P. Bartfai and J. Tomko, North-Holland, Amsterdam, pp. 181–225.
[7] Melamed, B. and Whitt, W. (1990). On arrivals that see time averages. *Operat. Res.* **38,** 156–172.
[8] Melamed, B. and Whitt, W. (1990). On arrivals that see time averages: a martingale approach. *J. Appl. Prob.* **27,** 376–384.
[9] Melamed, B. and Yao, D. D. (1995). The ASTA property. In *Advances in Queueing*, ed. J. H. Dshalalow, CRC, Boca Raton, FL, pp. 195–224.
[10] Niu, S. (1984). Inequalities between arrival averages and time averages in stochastic processes arising from queueing theory. *Operat. Res.* **32,** 785–795.
[11] Propp, J. and Wilson, D. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms* **9 ,** 223–252.
[12] Ross, S. and Peköz, E. A. (2007). *A Second Course in Probability.* ProbabilityBookstore.com, Boston, MA.
[13] Shanthikumar, G. and Zazanis, M. (1999). Inequalities between event and time averages. *Prob. Eng. Inf. Sci.* **13,** 293–308.
[14] Wilson, D. (2000). How to couple from the past using a read-once source of randomness. *Random Structures Algorithms* **16,** 85–113.
[15] Wolff, R. (1982). Poisson arrivals see time averages. *Operat. Res.* **30,** 223–231.