Analysis of Discrete Time Queues via the Reversed Process

Roy D. Yates

February 25, 1994

Dept. of Electrical and Computer Engineering
Rutgers University
PO Box 909
Piscataway NJ 08855-0909
email: ryates@ece.rutgers.edu

ABSTRACT

A class of discrete-time $M/G/1$ queues, including both round robin and last come first served service, in which customers are subject to permutations is considered. These time slotted queues, analogous to the symmetric queues of Kelly, are analyzed by examination of the time reversed process. Product form stationary distributions are found for a type of doubly stochastic server of Schassberger [5] and for a Bernoulli arrival process queue model of Henderson and Taylor [2].

Keywords: Round robin queues, time reversibility, product form, insensitivity, cyclic service queues

1 Introduction

Cyclic service queues have been analyzed under a wide variety of models; see e.g. Takagi [6,7]. However, simple closed form solutions for the stationary distribution have been found in only a small number of cases by Daduna and Schassberger [1], Henderson and Taylor [2], and Schassberger [5]. The stationary distribution of several continuous time queues can be found by examination of the time reversed system. These queues, including the M/G/1 processor sharing queue and the M/G/1 last come first served (LCFS) queue, are characterized by Kelly [3] as having a symmetric service discipline in that the departure rate of customers from position i equals the arrival rate of customers to position i.

In this work, we describe a family of discrete time M/G/1 queues, which we will call permutation queues, that are analogous to the continuous time symmetric queue. These permutation queues can describe forms of round robin and LCFS service. One of these permutation queues is a special case of the doubly stochastic server of Schassberger [5] with the additional restrictions that only one customer is served per unit time and that the permutations of the queued customers occur deterministically. A second type of permutation queue has the additional restriction that only one event, either arrival or departure, is permitted each unit of time. In each case, we will identify the time reversed process and find a product form solution for the stationary distribution.

The form of this stationary distribution has been found by Daduna and Schassberger [1] in the specific case of round robin service and by Schassberger [5] for the doubly stochastic server. Similar stationary distributions also have been found by Henderson and Taylor [2], for a queue in which the server moves from position to position. As a result of choosing a somewhat more detailed state description that specifies the server's position, the distributions found in Henderson and Taylor [2] differ somewhat from those found here. Other variations on these permutation queues were studied by Yates [8]. The work presented here provides a significantly simpler proof of the stationary distribution. As a by-product of the reverse time analysis, we gain some insight into the conditions under which a discrete time queue has a well defined reverse time process.

We will derive the stationary distribution of these permutation queues by examination of the system under time reversal using the following result (Theorem 4.7.2) from Ross [4].

Theorem 1 *Consider an irreducible Markov chain with transition probabilities P_{uv}. If one can find nonnegative numbers π_u summing to unity and a transition probability matrix $[P^*_{uv}]$ such that for all states u and v,*

$$\pi_u P_{uv} = \pi_v P^*_{vu} \tag{1}$$

then the P^*_{uv} are the transition probabilities for the time reversed chain and the π_u are the stationary probabilities for both chains.

We will make use of Theorem 1 by constructing a queue for which it will be relatively easy to guess the behavior of the reverse time queue. This guess will specify a set of transition probabilities P^*_{vu}. We then will find π_u satisfying (1) to verify our guess.

2 Work Conserving Permutation Queues

In these queues, unlike those of Section 3, a unit of service is provided whenever the system is not empty. Time is divided into slots where one slot is the time needed to provide one unit of service. At the start of each slot, a customer of class c arrives with probability λ_c. No more than one customer can arrive in any slot and the probability of no arrival in a slot is $1-\lambda$ where $\lambda = \sum_{c=1}^{C} \lambda_c < 1$. Each class c customer has an independent service requirement consisting of an integer number of packets X_c. The random variable X_c has mean $\overline{X}_c$ and is characterized by probability functions $g_c(x) = P\{X_c = x\}$ and $\overline{G}_c(x) = P\{X_c > x\}$ such that $\overline{G}_c(0) = 1$. The mean service requirement of a customer is $\overline{X} = \sum_c (\lambda_c/\lambda)\overline{X}_c$. Given that a class c customer has already had w packets served, the conditional probability that the customer's service requirement will be fulfilled after its next packet is served will be denoted by

$$\begin{aligned} r_c(w) &= P\{X_c = w+1 | X_c > w\} \\ &= \frac{g_c(w+1)}{\overline{G}_c(w)} \end{aligned}$$

We define $Q_1, Q_2, \ldots$ such that Q_j is a j element permutation. Given $Q_1, Q_2, \ldots$, we define $Q_1^{-1}, Q_2^{-1}, \ldots$ as the set of inverse permutations. That is, given $(y_1, \ldots, y_j)$, we have $Q_j^{-1}Q_j(y_1, \ldots, y_j) = (y_1, \ldots, y_j)$. We let $(n, y_1, \ldots, y_n)$ represent the state of the queue, where n is the number of customers in the system and $y_i = (w_i, c_i)$ such that for the customer in position i, w_i is the number of packets already served and c_i is the class of the customer. Suppose the state of the queue at the start of slot t is $(n, y_1, \ldots, y_n)$. In the transition to the next slot, the following sequence of events

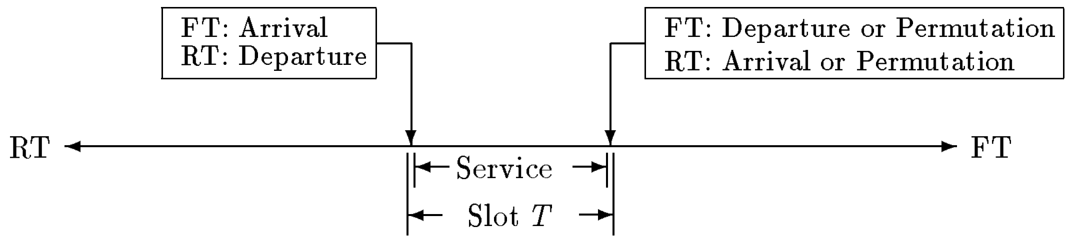

Figure 1: The corresponding sequences of events within a slot in forward time (FT) and reverse time (RT)

occurs:

1. With probability λ_c, a new customer arrives at the start of slot t and is inserted in position 1, shifting the existing customers to positions 2 through $n+1$. In this case, at time t^+, the instant after the arrival, the state of the queue will be $(n+1, (0, c), y_1, \ldots, y_n)$. However, if no arrival occurs, the state at time t^+ will remain $(n, y_1, \ldots, y_n)$.

2. The customer in position 1 of the queue has one packet served.

3. Following service, either the position 1 customer will depart if its service requirement has been met, or the customers of the queue are reordered using permutation Q_n. That is, if the queue state at time t^+ is $(n', y'_1, \ldots, y'_{n'})$, then the queue state at time $t+1$ will be either $(n'-1, y'_2, \ldots, y'_{n'-1})$ if a departure occurred or $(n', Q_{n'}((w'_1+1, c'_1), y'_2, \ldots, y'_{n'}))$ otherwise.

This sequence of events and the corresponding event sequence in the reverse time process are depicted in Figure 1. We call the above queue a late permutation queue in that permutations can occur at the end of each slot. We conjecture that the reverse time system is another permutation queue, to be called an early permutation queue, in which permutations can occur at the beginning of each slot. In addition, we shall see that the reverse time system uses the inverse permutations $Q_1^{-1}, Q_2^{-1}, \ldots$.

Theorem 2 *Assume that $\lambda \overline{X} < 1$, then the reverse time process is an early permutation queue, defined as follows. At the start of a slot, a class c customer arrives with probability λ_c. No more than one customer can arrive in any slot and the probability of no arrival in a slot is $1 - \lambda$ where $\lambda = \sum_{c=1}^{C} \lambda_c < 1$. Each class c customer has an independent service time distributed according to g_c. The state of the queue is $(n, y_1, \ldots, y_n)$ where n is the number of customers in the system and $y_i = (w_i, c_i)$ such that w_i is the residual service requirement in packets and c_i is the class of the customer in position i. Suppose the state of the queue at the start (in reverse time) of a slot equals $(n, y_1, \ldots, y_n)$. During this slot, the following sequence of events occurs:*

1. *Either with probability λ_c, a new customer arrives and is inserted in position 1 of the queue, shifting the existing customers to positions 2 through $n+1$, or with probability $1-\lambda$, the existing customers in the queue are reordered as $Q_n^{-1}(y_1, \ldots, y_n)$.*

2. *The position 1 customer has one packet served.*

3. *The position 1 customer departs if its service requirement has been met. If the position 1 customer departs, the remaining customers are shifted from positions $2, \ldots, n$ to $1, \ldots, n-1$.*

The stationary distribution for both early and late permutation queues is

$$\pi_{(n,y_1,\ldots,y_n)} = \pi_\phi \prod_{i=1}^{n} \frac{\lambda_{c_i}}{1-\lambda} \overline{G}_{c_i}(w_i) \tag{2}$$

where π_ϕ is the empty state probability.

Proof: Theorem 2 Given an arbitrary state u, we need to examine the set of transition probabilities $P_{u,v}$ to neighboring states v in the forward time chain and the corresponding transition probabilities $P^*_{v,u}$ from v to u in the reverse time chain. We will call u and v a state transition pair denoted by

$$u \; \underset{P^*_{v,u}}{\overset{P_{u,v}}{\rightleftarrows}} \; v$$

We will verify that for all neighboring states, the stationary distribution (2) satisfies (1). The four basic transitions that we must consider are:

- The customer in position 1 is served and departs. No new customer arrives.

  For the reverse transition, a new customer of class c_1 and service requirement w_1+1 must arrive to return to the state $(n, y_1, \ldots, y_n)$. In this instance, the transition pair is

$$(n, y_1, \ldots, y_n) \; \underset{\lambda_{c_1} g_{c_1}(w_1+1)}{\overset{(1-\lambda) r_{c_1}(w_1)}{\rightleftarrows}} \; (n-1, y_2, \ldots, y_n)$$

This transition pair implies that $\pi_{(n,y_1,\ldots,y_n)}$ must satisfy

$$\pi_{(n,y_1,\ldots,y_n)} = \frac{\lambda_{c_1}}{1-\lambda}\overline{G}_{c_1}(w_1)\pi_{(n-1,y_2,\ldots,y_n)} \tag{3}$$

It is easily verified that the distribution (2) satisfies (3).

We now describe the remaining three types of transition pairs. In all cases, it is easily verified that (2) satisfies the balance equation by repeated use of the identities $\overline{G}_c(0) = 1$ and $1 - r_{c_1}(w_1) = \overline{G}_{c_1}(w_1+1)/\overline{G}_{c_1}(w_1)$.

- The position 1 customer is served but does not depart. No new customer arrives.

  For the reverse transition, there must not be an arrival to ensure that the inverse permutation occurs and that the correct customer is served. This state transition pair corresponds to

$$(n, y_1, \ldots, y_n) \;\underset{1-\lambda}{\overset{(1-\lambda)[1-r_{c_1}(w_1)]}{\rightleftarrows}}\; (n, Q_n((w_1+1, c_1), y_2, \ldots, y_n))$$

- A new customer arrives, is served, and departs.

  The forward time transition corresponds to the arrival of a customer requiring one unit of service. In reverse time, there must also be an arriving customer requiring exactly one unit of service. The corresponding transition pair is

$$(n, y_1, \ldots, y_n) \;\underset{\sum_c \lambda_c g_c(1)}{\overset{\sum_c \lambda_c g_c(1)}{\rightleftarrows}}\; (n, y_1, \ldots, y_n)$$

- A new class c customer arrives, is served, but does not depart.

  For the reverse transition, the correct customer will be served and then depart iff no new customer arrives. The transitions are

$$(n, y_1, \ldots, y_n) \quad \underset{1-\lambda}{\overset{\lambda_c \overline{G}_c(1)}{\rightleftarrows}} \quad (n+1, Q_{n+1}((1,c), y_1, \ldots, y_n))$$

To complete the proof of Theorem 2, we show that the empty state is recurrent. Let N denote the number in the queue in equilibrium.

$$\begin{aligned} P\{N=n\} &= \sum_{c_1} \cdots \sum_{c_n} \sum_{w_1=1}^{\infty} \cdots \sum_{w_n=1}^{\infty} \pi_\phi \prod_{i=1}^{n} \frac{\lambda_{c_i}}{1-\lambda} \overline{G}_{c_i}(w_i) \\ &= \pi_\phi \left(\frac{\lambda(\overline{X}-1)}{1-\lambda} \right)^n \end{aligned}$$

Since $P\{N=0\} = \pi_\phi$, requiring $\sum_{n=0}^{\infty} P\{N=n\} = 1$ yields $\pi_\phi = (1-\lambda\overline{X})/(1-\lambda)$. As expected, the empty state is recurrent iff $\lambda\overline{X} < 1$. $\square$

We see that in order for the reverse process to have physical meaning, the discrete time queue must have certain properties. In particular, what Henderson and Taylor [2] call the welcoming quanta of work is required only if in reverse time, a customer immediately departs the system after its final packet is served. In addition, we see that if we were to permit multiple departures in the forward time system, the reverse time system would not have Bernoulli arrivals. As a result, we have restricted the permutation queue to serve one customer in each slot. By comparison, the doubly stochastic server of Schassberger [5] permits multiple customers to be served in the same slot; however, when two or more customers complete their service requirements at the same instant, each of those customers is forced to remain in the queue and start a fresh service requirement.

3 The One-Event Permutation Queue

The reverse time analysis can also be applied to a second queue considered by Henderson and Taylor [2]. This system does not require an arriving customer to have a packet served immediately upon arrival. However, in one slot, there can be either an arrival or departure but not both. As before, we let $(n, y_1, \ldots, y_n)$ represent the state of the queue. Suppose the state of the queue at the start of slot t is $(n, y_1, \ldots, y_n)$. In the transition to slot $t+1$, the following sequence of events occurs:

1. Either with probability λ_c, a class c customer arrives and is inserted in position 1 of the queue, or, with probability μ, $\mu \leq 1 - \sum_c \lambda_c$, the position 1 customer is given one unit of service. Following a unit of service, a customer will depart if its service requirement has been met.

2. The customers in the queue are permuted. That is, the queue state at the start of slot $t+1$ will be either $(n+1, Q_{n+1}((0, c), y_1, \ldots, y_n))$ if a class c arrival occurred, $(n, Q_n((w_1+1, c_1), y_2, \ldots, y_n))$ if a customer was served but did not depart, or $(n-1, Q_{n-1}(y_2, \ldots, y_n))$ if a customer was served and did depart.

We call this system a one-event late permutation queue. We emphasize that if a customer arrives at the start of a slot, then no service is provided in that slot. We now describe the reverse time system.

Theorem 3 *Assume $\lambda \overline{X} < \mu$, then the reverse time process is a one-event early permutation queue, defined as follows. If the state of the system at the start of a slot is $(n, y_1, \ldots, y_n)$, then the following sequence of events will occur.*

1. *The customers in the queue are reordered as $Q_n^{-1}(y_1, \ldots, y_n)$.*

2. *Either with probability λ_c, independent of the state of the queue, a class c arrival is inserted in position 1 of the queue and given one unit of service, or with probability μ, $\mu \leq 1 - \sum_c \lambda_c$, either the position 1 customer departs if its residual service requirement is zero or the position 1 customer is given one unit of service. Note that a customer does not depart in the slot in which its last packet is served.*

The stationary distribution of both early and late one-event permutation queues is described by

$$\pi_{(n,y_1,\ldots,y_n)} = \pi_\phi \prod_{i=1}^{n} \frac{\lambda_{c_i}}{\mu} \overline{G}_{c_i}(w_i) \tag{4}$$

Proof: Theorem 3 Just as in the proof of Theorem 2, we must examine the possible types of transitions. For each type of transition, we leave to the reader the verification that (4) satisfies the balance equation (1) for all transition pairs.

- A class c arrival occurs.

  In this case, no other action occurs in the forward time queue. In reverse time, after the inverse permutation, the position 1 customer departs with probability

μ. The corresponding transition pair is

$$(n, y_1, \ldots, y_n) \underset{\mu}{\overset{\lambda_c}{\rightleftarrows}} (n+1, Q_{n+1}((0,c), y_1, \ldots, y_n))$$

- There is no arrival. The position 1 customer is served and departs.

  For the reverse time transition, there must be a class c_1 arrival requiring $w_1 + 1$ units of service. The transition pair is

$$(n, y_1, \ldots, y_n) \underset{\lambda_{c_1} g_{c_1}(w_1+1)}{\overset{\mu r_{c_1}(w_1)}{\rightleftarrows}} (n-1, Q_{n-1}(y_2, \ldots, y_n))$$

- There is no arrival. The position 1 customer is served, but does not depart.

  In reverse time, after the inverse permutation, the customer in position 1 must be given a unit of service. The corresponding transition pair is

$$(n, y_1, \ldots, y_n) \underset{\mu}{\overset{\mu(1 - r_{c_1}(w_1))}{\rightleftarrows}} (n, Q_n((c_1, w_1+1), y_2, \ldots, y_n))$$

Lastly, we show that $\lambda \overline{X} < \mu$ implies the empty state of the one-event permutation queue is recurrent. Let M denote the number in the one-event permutation queue in equilibrium. In this case,

$$\begin{aligned} P\{M = m\} &= \sum_{c_1} \cdots \sum_{c_m} \sum_{w_1=0}^{\infty} \cdots \sum_{w_m=0}^{\infty} \pi_\phi \prod_{i=1}^{m} \frac{\lambda_{c_i}}{\mu} \overline{G}_{c_i}(w_i) \\ &= \pi_\phi \left(\frac{\lambda \overline{X}}{\mu} \right)^m \end{aligned}$$

Since $P\{M=0\} = \pi_\phi$, $\sum_{m=0}^{\infty} P\{M=m\} = 1$ implies $\pi_\phi = 1 - \lambda\overline{X}/\mu$. $\square$

In the one-event late permutation queue, an arriving customer is not provided a welcoming unit of service. In the corresponding reverse time system, the one-event early permutation queue, a customer that has completed service does not immediately depart but rather waits until the next slot in which it is in position 1 in the queue. In that slot, the customer departs with probability μ. In comparison, for one-event early permutation queue, new arrivals are immediately given a welcoming unit of service. Hence, in the one-event late permutation queue, a customer that completes its service requirement immediately departs.

We observe that when $\mu = 1 - \sum_c \lambda_c$, either an arrival occurs or a customer is served whenever the queue is busy. In this case, the work conserving permutation queues considered in Section 2 and the one-event permutation queues have the same form of stationary distribution. However, the work conserving queues and the one-event queues do not have the same state space. In the work conserving system, $w_i \geq 1$ for all i but in the one event system, it is possible that $w_i = 0$. The difference between the one-event permutation queues and the work conserving queues is that the work requirement of a customer in the one event permutation queue is effectively increased by one unit corresponding to the time slot which is devoted to acceptance of the arriving customer.

4 Properties of Permutation Queues

As detailed by Daduna and Schassberger [1] and Henderson and Taylor [2], the product form stationary distribution leads to a number of consequences. In particular, the number of type c customers in the queue depends on $\overline{X}_c$ but is otherwise insensitive to the service time distribution $\overline{G}_c(\cdot)$. In addition, the system time of a typical customer with service requirement τ is a linear function of τ.

We observe that in equilibrium in the reverse time system, the arrival process in future slots $t, t-1. \ldots$ is independent of the state of the queue at past times $t, t+1, \ldots$. For both early and late permutation queues, as well as the corresponding one-event queues, this implies

Corollary 1 *In equilibrium in the forward time system, the departure process in slots* $t, t-1, \ldots,$ *is independent of the state of the queue at times* $t, t+1, \ldots$.

By an appropriate choice of permutations $Q_1, Q_2, \ldots$, we can choose to represent either LCFS or round robin as a permutation queue. In particular, for round robin, we simply let Q_n describe the rotation of customer $i > 1$ to position $i-1$ and customer 1 to position n. For LCFS, the permutation Q_n is simply the identity function. That is, under LCFS, the customers are never reordered. Other disciplines can also be implemented. For example, we might choose to perform round robin among the customers in the first k positions. More unusual examples can also be constructed. We could choose $Q_1, Q_2, \ldots$ such that round robin is used when the number of customers in the system is even and LCFS is used when the number is odd. However, we emphasize that the permutations cannot depend on the classes of the customers in the queue.

The author wishes to acknowledge the numerous improvements suggested by the editors and referees. In addition, the author would like to thank Prof. Robert Gallager for his encouragement and supervision of this work at MIT in the Laboratory for

Information and Decision Systems. This work was supported by an AT&T Bell Laboratories fellowship and by National Science Foundation grant NSF-ECS-831698.

References

[1] H. Daduna and R. Schassberger. A discrete-time round-robin queue with Bernoulli input and general arithmetic service time distributions. *Acta Informatica*, 15:251–263, 1981.

[2] W. Henderson and P.G. Taylor. Insensitivity in discrete time queues with a moving server. *Queueing Systems*, 11:273–297, 1992.

[3] F. P. Kelly. *Reversibility and Stochastic Networks.* John Wiley & Sons, New York, 1979.

[4] S. M. Ross. *Stochastic Processes.* John Wiley & Sons, New York, 1983.

[5] R. Schassberger. The doubly stochastic server: A time sharing model. *Zeitschrift für Operations Research*, 25:179–189, 1981.

[6] H. Takagi. *Analysis of Polling Systems.* MIT Press, Cambridge, MA, 1986.

[7] H. Takagi. Queuing analysis of polling models. *ACM Computing Surveys*, 20(1):5–28, 1988.

[8] R. D. Yates. *High Speed Round Robin Queueing Networks.* PhD thesis, MIT Dept. of Electrical Engineering and Computer Science, 1990.

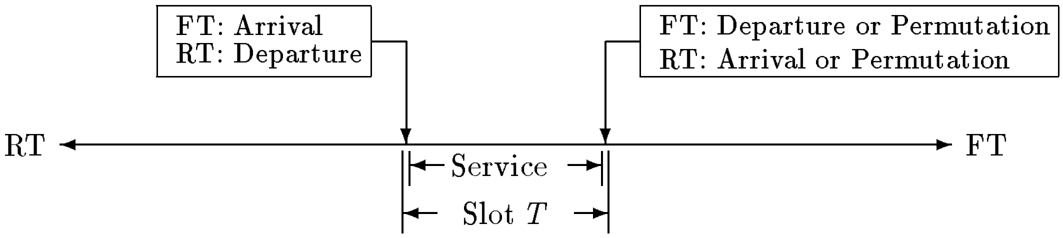

Figure 1: The corresponding sequences of events within a slot in forward time (FT) and reverse time (RT)