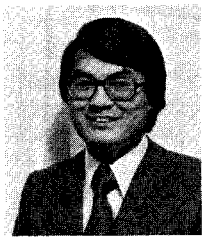
A Derivation of Response Time Distributions for a Multi-Class Feedback Queueing System

Simon S. Lam and A. Udaya Shankar
Department of Computer Sciences, University of Texas at Austin, Austin, TX 78712, U.S.A.

A single server queue with feedback and multiple customer types is analyzed. Arrival processes are independent Poisson processes. After receiving a quantum of service, a customer may depart or rejoin the end of the queue for more service. The number of quanta of service required by a customer of a specific type is a random variable having a general distribution with finite support. Each quantum of service is exponentially distributed. We derived the moment generating function of customer response time conditioned on the number of quanta of service required. An efficient algorithm for calculating the second-order statistics of the conditional response time is given. Numerical results are shown illustrating the variance of the conditional response time for different service requirement distributions and customer types. The performance of the FCFS and round-robin scheduling disciplines are compared.

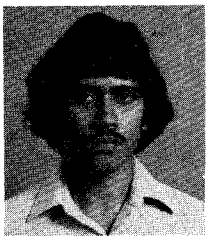
Keywords: Feedback Queue, Response Time Distribution, Round-Robin, Time-Sharing, Product-Form Queueing Network.



Simon S. Lam received the B.S.E.E. degree (with distinction) from Washington State University, Pullman, WA, in 1969, and the M.S. and Ph.D. degrees in Engineering from the University of California at Los Angeles, in 1970 and 1974, respectively.

From 1972 to 1974, he was with the ARPA Network project at UCLA and did research on satellite packet communication. From 1974 to 1977, he was a research staff member with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, where he worked on performance analysis problems of packet switching networks, SNA and satellite networks. Since September 1977, he has been with the University of Texas at Austin where he is now Associate Professor of Computer Sciences. His current research interests include modeling and analysis of computer systems, networks and communication protocols.

At the University of California at Los Angeles, he held a Phi Kappa Phi Fellowship from 1969 to 1970, and a Chancellor's Teaching Fellowship from 1969 to 1973. In 1975, he received (together with L. Kleinrock) the IEEE Leonard G. Abraham Award for the best paper in the field of Communications Systems. Simon is a member of Tau Beta Pi, Sigma Tau, Phi Kappa Phi, Pi Mu Epsilon, the Association for Computing Machinery and IEEE.



A. Udaya Shankar received the B. Tech. degree in Electrical Engineering from the Indian Institute of Technology, Kanpur, in 1976 and the M.S. degree in Computer Engineering from Syracuse University, Syracuse, NY, in 1978. Currently, he is working towards the Ph.D. degree in Electrical Engineering at the University of Texas at Austin and is a research assistant in the area of computer network protocols.

The authors would like to thank the anonymous referees for their constructive criticisms. This work was supported by National Science Foundation Grant No. ENG78-01803.

An early version of this paper appears in Proceedings Performance '80, Toronto, Canada, May 1980. (Copyright 1980, Association for Computing Machinery, Inc., by permission).

Performance Evaluation 1 (1981) 48–61

1. Introduction

A service facility using a round-robin scheduling discipline can be modeled as a feedback queue such as shown in Fig. 1. A single-server queue is considered with infinite waiting room and Q types of customers. The arrival process of type q customers is an independent Poisson process ($q = 1, 2, ..., Q$). Each new arrival joins the end of the queue. The customer at the head of the queue receives from the server a quantum of service which is an independent exponentially distributed random variable with mean $1/\mu$ seconds. After receiving a quantum of service, a customer may depart or rejoin the end of the queue for more service. The number of quanta of service required by a type q customer is a random variable with probability distribution $\{a_r^{(q)}, r = 1, 2, ..., R\}$ where R is finite and $a_r^{(q)}$ is the probability of a type q customer requiring exactly r quanta of service.

The queue length distribution of the above model is readily available since the feedback queue described is an open queueing network satisfying local balance [1]. The contribution of this paper is to characterize response time distributions of the different types of customers; specifically, we solved for the moment generating function of the conditional response time of customers requiring r quanta of service for $r = 1, 2, ..., R$.

1.1. Relationship to prior work

Time-sharing models were first studied by Kleinrock [2] who solved for the mean response time of a customer conditioning on his exact service requirement. He considered two cases: (a) constant quantume size Δ, and (b) the limiting case of $\Delta \to 0$ called processor-sharing. Customers are assumed to arrive according to a Poisson process. In case (a), the number of service quanta required by a customer is geometrically distributed. In case (b), the service requirements are characterized by an exponential distribution. (This is called the processor-sharing M/M/1 queue.) Kleinrock's conditional mean response time result was later shown to hold for a processor-sharing M/G/1 queue (i.e. service requirements characterized by a general distribution) by Sakata et al. [3]. Higher order response time statistics are much harder to get. The response time distribution for the processor-sharing M/M/1 queue was obtained by Coffman et al. [4]. The response time distribution for the constant quantum size case was obtained by Muntz [5] assuming exponentially distributed service requirements.

Our feedback queue model is different from the time-sharing models in that a service quantum in our model is exponentially distributed. (Our model can be used, however, to approximate processor-sharing by making $1/\mu$ very small relative to the mean service requirement.) Aside from the quantum size assumption, our model is more general than those of [4, 5] in two respects: (i) multiple types of customers, and (ii) the number of quanta of service required by customers of each type can have a general probability distribution with finite support. In terms of the exact amount of service required, service distributions that are admissible in our model are those with moment generating functions of the form

$$B_q^*(s) = \sum_{r=1}^{R} a_r^{(q)} \left(\frac{\mu}{s + \mu} \right)^r . \tag{1}$$

Our model is also different from the feedback queue model of Takács [6]. In his model, service quanta can have a general distribution. However, he considered one type of customers only and the number of quanta of service required by a customer is geometrically distributed; in other words, after each quantum of service, a

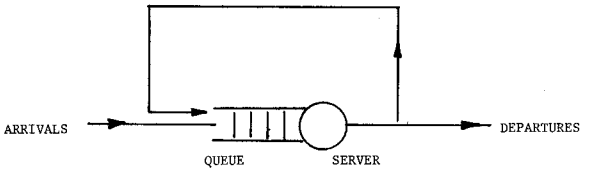


Fig. 1. A feedback queue model.

customer always departs with probability $(1-p)$ and rejoins the end of the queue with probability p (memoryless feedback behavior).

The original motivation of this work stems from our efforts to characterize the response time in a network of queues. For a network of FCFS queues that satisfies local balance, Wong [7] found the response time distribution of customers traversing loop-free paths. Our results in this paper represent efforts to understand the response time behavior along paths with loops in the simplest form of queueing networks satisfying local balance. Other attempts to characterize the response time behavior in queueing networks were made by Chow [8] and Yu [9].

1.2. Assumptions and definitions

Consider the following example of 2 types of customers. Type 1 customers arrive according to a Poisson process with rate α_1 customers per second. The number of quanta of service required by a type 1 customer has the probability distribution

$$a_r^{(1)} = \begin{cases} \frac{1}{100} & r = 1, 2, ..., 100 , \\ 0 & \text{otherwise} . \end{cases}$$

Type 2 customers arrive according to a Poisson process with rate α_2 customers per second. The number of quanta of service required by a type 2 customer has the probability distribution

$$a_r^{(2)} = \begin{cases} \frac{1}{10} & r = 1, 2, ..., 10 , \\ 0 & \text{otherwise} . \end{cases}$$

Using the properties of Poisson processes, the above model is equivalent to the following model with 100 types of customers. Type r customers ($r = 1, 2, ..., 100$) are defined to be those requiring exactly r quanta of service and they arrive according to a Poisson process with rate

$$\gamma_r = \begin{cases} 0.01\alpha_1 + 0.1\alpha_2 & r = 1, 2, ..., 10 , \\ 0.01\alpha_1 & r = 11, 12, ..., 100 , \\ 0 & \text{otherwise} . \end{cases}$$

Let R be the maximum number of service quanta required by any customer. We shall, without any loss of generality, consider the following model. There are R types of customers. The arrival process of the rth type is Poisson at rate γ_r customers per second. A type r customer requires exactly r quanta of service. It should be obvious that if we can derive response time distributions for this model, response time distributions for any model with Q customer types and service time requirements characterized by eq. (1) can be easily obtained.

Let t_r be the *response time* of attaining exactly r quanta of service; $r = 1, 2, ..., R$ and we define t_0 to be zero. We shall solve for its moment generating function

$$T_r^*(s) = E[e^{-st_r}] ,$$

where $E[\cdot]$ denotes the expectation of the function of random variable(s) inside the brackets.

We shall only consider steady-state results. For a single-server queue, stationarity is assured if the traffic intensity $\rho < 1$ where $\rho = \Sigma_{r=1}^{R} \gamma_r(r/\mu)$ (see Cohen [10]).

Customers in the queue are differentiated into R different *classes;* class k consists of all those customers in the queue with exactly k *more* quanta of service to go, where $k = 1, 2, ..., R$.

Let us follow the progress of a 'tagged' customer and introduce some more notation. Upon his initial arrival, the tagged customer finds n_k class k customers in the queue ($k = 1, 2, ..., R$). The system state thus found at an

arrival instant is denoted by $\boldsymbol{n} = (n_1, n_2, ..., n_R)$ and is described by the moment generating function

$$P^*(\boldsymbol{z}) = E[z_1^{n_1} z_2^{n_2} ... z_R^{n_R}] ,$$

where $\boldsymbol{z}$ is the shorthand notation for $(z_1, z_2, ..., z_R)$.

At the end of the tagged customer's rth quantum of service (given that he requires at least r quanta), let the system at that instant be denoted by $\boldsymbol{m}^{(r)} = (m_1^{(r)}, m_2^{(r)}, ..., m_R^{(r)})$ where $m_k^{(r)}$ is the number of customers who have exactly k more quanta of service to go. Define

$$M^{(r)} = \sum_{k=1}^{R} m_k^{(r)} .$$

In order to characterize $T_r^*(s)$, we shall need to first characterize the joint distribution of t_r and $\boldsymbol{m}^{(r)}$, which is described by

$$U_r^*(s, \boldsymbol{z}) = E[e^{-st_r} z_1^{m_1^{(r)}} z_2^{m_2^{(r)}} ... z_R^{m_R^{(r)}}] .$$

1.3. Summary of results

We derived a recursive equation relating $U_{r+1}^*(s, \boldsymbol{z})$ to $U_r^*(s, \boldsymbol{z})$ (Lemma 2). An explicit solution of $U_r^*(s, \boldsymbol{z})$ was found, from which $T_r^*(s)$ was obtained (Theorem 1). We then proved that the stationary distribution of $\boldsymbol{m}^{(r)}$, $r = 1, 2, ..., R$, is the same as that of $\boldsymbol{n}$ (Theorem 2). With this result, we solved for the mean value of t_r (Theorem 3); this last result is similar to the mean conditional response time result of processor-sharing models [2,3]. We also obtained an efficient algorithm to calculate recursively the second-order statistics of t_r (Theorem 4). Numerical results are shown in Section 3 to illustrate the behavior of response time variance for different service requirement distributions and to compare the performance of the FCFS and round-robin scheduling disciplines.

2. The analysis

Consider the sytem state $\boldsymbol{n} = (n_1, n_2, ..., n_R)$ at arrival instants. Recall that n_k is the number of class k customers with exactly k more quanta of service to go. [1] The aggregate arrival rate of customers to the kth class is

$$\lambda_k = \sum_{i=k}^{R} \gamma_i \tag{2}$$

since any new arrival who requires at least k quanta of service must enter and leave the kth class exactly once.

Lemma 1. *The moment generating function of* $\boldsymbol{n}$ *is*

$$P^*(\boldsymbol{z}) = \frac{1 - \rho}{1 - \sum_{k=1}^{R} \rho_k z_k} , \tag{3}$$

where

$$\rho_k = \lambda_k/\mu \qquad \textit{and} \qquad \rho = \sum_{k=1}^{R} \rho_k .$$

[1] Since the duration of a service quantum is assumed to be exponentially distributed, the remaining fraction of a quantum of the job found in service is counted as a full quantum.

Proof. Given Poisson arrival processes, the system state probabilities at an arrival instant are the same as system state probabilities at a random time instant [11]. With each quantum of service being exponentially distributed with the same mean $(1/\mu)$, we have an open queueing network that satisfies local balance [1]. Eq. (3) has been obtained by Reiser and Kobayashi [12]. (Q.E.D.)

Since each quantum of service is exponentially distributed, it has the moment generating function

$$B^*(s) = \frac{\mu}{s+\mu}. \tag{4}$$

A recursive solution of $U_r^*(s, \boldsymbol{z})$ is next given.

Lemma 2.

$$U_0^*(s, \boldsymbol{z}) = P^*(\boldsymbol{z}), \tag{5}$$

$$U_{r+1}^*(s, \boldsymbol{z}) = y_1(s, \boldsymbol{z})\, U_r^*(s, \boldsymbol{y}(s, \boldsymbol{z})), \qquad r \geqslant 0 \tag{6}$$

where

$$\boldsymbol{y}(s, \boldsymbol{z}) = (y_1(s, \boldsymbol{z}), y_2(s, \boldsymbol{z}), \ldots, y_R(s, \boldsymbol{z})),$$

$$y_1(s, \boldsymbol{z}) = B^*\left(s + \sum_{i=1}^{R} \gamma_i (1 - z_i)\right),$$

and

$$y_k(s, \boldsymbol{z}) = z_{k-1} y_1(s, \boldsymbol{z}) \qquad \text{for } 2 \leqslant k \leqslant R.$$

Proof. For $r = 0$, $t_0 = 0$ and $\boldsymbol{m}^{(0)} = \boldsymbol{n}$. This and the definition of $U_r^*(s, \boldsymbol{z})$ yield (5).

To show (6), consider the time period between t_r and t_{r+1} during which the server served $M^{(r)} + 1$ customers, where $M^{(r)} = \Sigma_{k=1}^{R} m_k^{(r)}$ and the extra one is for the tagged customer's $(r+1)$ st quantum. During the same time period, each class k customer became a class $(k-1)$ customer where $k = 2, 3, \ldots, R$. Furthermore, let $A_k(t)$ be the number of external new arrivals to class k during time $t (= t_{r+1} - t_r)$ according to a Poisson process of rate γ_k customers per second. We note that class R is an exception in that its $m_R^{(r+1)}$ customers are all new arrivals. Thus, conditioning on t_r and $\boldsymbol{m}^{(r)}$, we have

$$U_{r+1}^*(s, \boldsymbol{z}/t_r, \boldsymbol{m}^{(r)}) = E[\mathrm{e}^{-s(t+t_r)} z_1^{m_2^{(r)} + A_1(t)} z_2^{m_3^{(r)} + A_2(t)} \ldots z_R^{A_R(t)} / t_r, \boldsymbol{m}^{(r)}]$$

$$= \mathrm{e}^{-st_r} \left(\prod_{k=2}^{R} z_{k-1}^{m_k^{(r)}} \right) E[\mathrm{e}^{-st} z_1^{A_1(t)} z_2^{A_2(t)} \ldots z_R^{A_R(t)} / M^{(r)}].$$

The last quantity on the right-hand side is $(y_1(s, \boldsymbol{z}))^{M^{(r)}+1}$ because t is the sum of $M^{(r)} + 1$ independent identically distributed random variables with the moment generating function $B^*(s)$. The above equation can be rewritten as

$$U_{r+1}^*(s, \boldsymbol{z}/t_r, \boldsymbol{m}^{(r)}) = y_1(s, \boldsymbol{z}) \left\{ \mathrm{e}^{-st_r} y_1(s, \boldsymbol{z})^{m_1^{(r)}} \prod_{k=2}^{R} [z_{k-1} y_1(s, \boldsymbol{z})]^{m_k^{(r)}} \right\}.$$

Unconditioning on t_r and $\boldsymbol{m}^{(r)}$, (6) follows. (Q.E.D.)

Explicit solutions for $U_r^*(s, \boldsymbol{z})$ and $T_r^*(s)$ can now be shown.

Theorem 1. (i)

$$U_r^*(s, \boldsymbol{z}) = \frac{1-\rho}{P_r(s) - \sum_{k=1}^{R} Q_{k,r}(s)\, z_k} \qquad r \geqslant 0 \tag{7}$$

where $P_r(s)$ and $Q_{k,r}(s)$ are polynomials in s given by

$$\begin{bmatrix} P_r(s) \\ Q_{1,r}(s) \\ Q_{2,r}(s) \\ \vdots \\ Q_{R-1,r}(s) \\ Q_{R,r}(s) \end{bmatrix} = \begin{bmatrix} \left(1 + \frac{s}{\mu} + \rho_1\right) & -1 & 0 & 0 & \cdots & 0 \\ \gamma_1/\mu & 0 & 1 & 0 & & 0 \\ \gamma_2/\mu & 0 & 0 & 1 & & 0 \\ \vdots & \vdots & & & \ddots & \vdots \\ & & & & 1 & 0 \\ \gamma_{R-1}/\mu & 0 & 0 & \cdots & 0 & 1 \\ \gamma_R/\mu & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}^r \begin{bmatrix} 1 \\ \rho_1 \\ \rho_2 \\ \vdots \\ \rho_{R-1} \\ \rho_R \end{bmatrix}. \tag{8}$$

(ii)

$$T_r^*(s) = \frac{1-\rho}{P_r(s) - \sum_{k=1}^{R} Q_{k,r}(s)}. \tag{9}$$

Proof. (i) Because of (3) and (5), (7) holds for $r = 0$ with $P_0(s) = 1$ and $Q_{k,0}(s) = \rho_k$ for $1 \leqslant k \leqslant R$. Assuming that (7) holds for r, we use (6) and (4) to express $U_{r+1}^*(s, \boldsymbol{z})$ as follows

$$U_{r+1}^*(s, \boldsymbol{z}) = \frac{1}{1 + (s/\mu) + \sum_{i=1}^{R} (\gamma_i/\mu)(1 - z_i)} \cdot \frac{1-\rho}{P_r(s) - \dfrac{Q_{1,r}(s) - \sum_{k=1}^{R-1} Q_{k+1,r}(s)\, z_k}{1 + (s/\mu) + \sum_{i=1}^{R} (\gamma_i/\mu)(1 - z_i)}}$$

$$= (1-\rho)\left[\left\{\left(1 + \frac{s}{\mu} + \sum_{i=1}^{R} \frac{\gamma_i}{\mu}\right) P_r(s) - Q_{1,r}(s)\right\} - \sum_{k=1}^{R-1}\left[\frac{\gamma_k}{\mu} P_r(s) + Q_{k+1,r}(s)\right] z_k - \frac{\gamma_R}{\mu} z_R P_r(s)\right]^{-1}.$$

Thus, the form of (7) is maintained, and it is evident from the above that

$$\begin{bmatrix} P_{r+1}(s) \\ Q_{1,r+1}(s) \\ \vdots \\ Q_{R,r+1}(s) \end{bmatrix} = \begin{bmatrix} \left(1 + \frac{s}{\mu} + \rho_1\right) & -1 & 0 & \cdots & 0 \\ \gamma_1/\mu & 0 & 1 & & \vdots \\ \vdots & & & \ddots & 0 \\ & & & & 0 \\ \gamma_R/\mu & \cdots & & 0 & 0 \end{bmatrix} \begin{bmatrix} P_r(s) \\ Q_{1,r}(s) \\ \vdots \\ Q_{R,r}(s) \end{bmatrix}. \tag{10}$$

The recursion in (10) started at $r = 0$ yields (8).

(ii) (9) follows from (7) and $T_r^*(s) = U_r^*(s, \mathbf{1})$. (Q.E.D.)

For $r = 1, 2$ and 3, we show $U_r^*(s, z)$ below.

$$U_1^*(s, z) = \frac{1-\rho}{1 + (s/\mu) - \sum_{k=1}^{R} \rho_k z_k},$$

$$U_2^*(s, z) = \frac{1-\rho}{(1 + (s/\mu))^2 + (s/\mu)\rho_1 - \sum_{k=1}^{R} (\rho_k + (s/\mu^2)\gamma_k) z_k},$$

$$U_3^*(s, z) = (1-\rho)/\left\{\left(1+\frac{s}{\mu}\right)^3 + 2\left(\frac{s}{\mu}\right)^2 \rho_1 + \frac{s}{\mu}(\rho_2 + 2\rho_1 + \rho_1^2) - \sum_{i=1}^{R-1}\left\{\frac{\gamma_i}{\mu}\left[\left(1+\frac{s}{\mu}\right)^2 + \frac{s}{\mu}\rho_1\right] + \left[\rho_{i+1} + \frac{\gamma_{i+1}}{\mu}\frac{s}{\mu}\right]\right\} z_i \right.$$
$$\left. - \frac{\gamma_R}{\mu}\left[\left(1+\frac{s}{\mu}\right)^2 + \frac{s}{\mu}\rho_1\right] z_R \right\}.$$

From the above, we obtain $T_r^*(s)$ for $r = 1, 2$ and 3 by letting $z = \mathbf{1}$ in $U_r^*(s, z)$.

$$T_1^*(s) = \frac{1-\rho}{(1 + (s/\mu)) - \rho},$$

$$T_2^*(s) = \frac{1-\rho}{(1 + (s/\mu))^2 - \rho},$$

$$T_3^*(s) = \frac{1-\rho}{(1 + (s/\mu))^3 + \rho_1 (s/\mu)^2 - \rho}.$$

We note that the solutions for $U_r^*(s, z)$ and $T_r^*(s)$ become quite complex if one tries to solve for $P_r(s)$ and $Q_{k,r}(s)$ explicitly using the matrix eq. (8) when $r \geqslant 4$. In what follows, we turn our attention to finding the moments of t_r. To do so, we need the following result concerning the distribution of $\boldsymbol{m}^{(r)}$.

Theorem 2. *For any $r \geqslant 0$, $\boldsymbol{m}^{(r)}$ and $\boldsymbol{n}$ have the same stationary distribution. That is*

$$U_r^*(0, z) = E[z_1^{m_1^{(r)}} z_2^{m_2^{(r)}} \cdots z_R^{m_R^{(r)}}] = P^*(z). \tag{11}$$

Proof. By (5), (11) holds true for $r = 0$. Assume that (11) holds true for some r so that $U_r^*(0, z) = P^*(z)$. By (3), (6) and the induction hypothesis,

$$U_{r+1}^*(0, z) = y_1(0, z) \cdot \frac{1-\rho}{1 - \sum_{k=1}^{R} \rho_k y_k(0, z)} = \frac{1-\rho}{(1/y_1(0, z)) - \left(\rho_1 + \sum_{k=1}^{R-1} \rho_{k+1} z_k\right)}$$

$$= \frac{1-\rho}{1 + \sum_{i=1}^{R} (\gamma_i/\mu)(1 - z_i) - \rho_1 - \sum_{k=1}^{R-1} \rho_{k+1} z_k} = \frac{1-\rho}{1 - \sum_{k=1}^{R} \rho_k z_k}$$

which is $P^*(\boldsymbol{z})$. The last equality is obtained using the following relationships:

$$\rho_1 = \frac{\lambda_1}{\mu} = \sum_{i=1}^{R} \frac{\gamma_i}{\mu} \qquad \text{and} \qquad \rho_k = \frac{\gamma_k}{\mu} + \rho_{k+1} \qquad \text{for} \qquad 1 \leqslant k \leqslant R-1 \text{ . (Q.E.D.)}$$

The moments of t_r can be obtained from the moment generating function of t_r and $\boldsymbol{m}^{(r)}$ as follows.

$$\begin{aligned} E[t_r^n] &= (-1)^n \frac{\partial^n}{\partial s^n} U_r^*(s, \boldsymbol{z})\Big|_{s=0, z=1} \\ &= (-1)^n \frac{\partial^n}{\partial s^n} U_r^*(s, z, z, \ldots, z)\Big|_{s=0, z=1} . \end{aligned} \tag{12}$$

Theorem 3. *The conditional mean response time is*

$$E[t_r] = \frac{r/\mu}{1-\rho}. \tag{13}$$

Proof. Using (6) and (12), we have

$$\begin{aligned} E[t_{r+1}] &= -\frac{\partial}{\partial s} \{B^*(s + \lambda_1(1-z))\, U_r^*(s, \boldsymbol{y}(z, z, \ldots, z))\}\Big|_{s=0, z=1} \\ &= \frac{1}{\mu} - E\left[\frac{\partial}{\partial s} \{e^{-st_r} (B^*(s))^{M^{(r)}}\}\right]_{s=0} \\ &= \frac{1}{\mu} - \left\{-E[t_r] - E[M^{(r)}] \cdot \frac{1}{\mu}\right\}. \end{aligned}$$

By (11),

$$E[M^{(r)}] = \frac{\partial}{\partial z} P^*(z, z, \ldots, z)\Big|_{z=1} = \frac{\rho}{1-\rho}.$$

Substituting this into the above expression for $E[t_{r+1}]$, we have

$$E[t_{r+1}] = \frac{1/\mu}{1-\rho} + E[t_r]$$

which yields (13) by induction starting with $E[t_0] = 0$. (Q.E.D.)

Theorem 4. *The second-order statistics of the conditional response time can be found recursively using*

$$\mathrm{Var}(t_{r+1}) = \mathrm{Var}(t_r) + \frac{1-2\rho r}{\mu^2(1-\rho)^2} + \frac{2}{\mu} E[t_r M^{(r)}] , \tag{14}$$

$$E[t_{r+1} M^{(r+1)}] = \sum_{i=1}^{R} E[t_{r+1} m_i^{(r+1)}] \tag{15}$$

and

$$E[t_{r+1} m_i^{(r+1)}] = \frac{2\rho_i}{\mu(1-\rho)^2} + \frac{r\gamma_i}{\mu^2(1-\rho)} + \frac{\gamma_i}{\mu} E[t_r M^{(r)}] + E[t_r m_{i+1}^{(r)}] , \qquad 1 \leqslant i \leqslant R \tag{16}$$

where $\mathrm{Var}(t_r)$ *is the variance of* t_r *and* $E[t_r m_{R+1}^{(r)}]$ *is zero, with the initial condition*

$\mathrm{Var}(t_0) = 0$,

$E[t_0 m_i^{(0)}] = 0 \qquad$ *for* $1 \leqslant i \leqslant R$.

The above theorem is proved by taking derivatives of (6), using the moment generating properties of transforms; (11) and (13) are used to simplify the resulting expressions. (See Appendix for details of the proof.)

3. Numerical examples

The conditional mean response time result in Theorem 3 is analogous to results from analyses of a processor-sharing queue [2,3]. The mean response time

$$E[t_r] = \frac{r/\mu}{1-\rho}$$

of a type r job varies linearly as its (expected) service requirement r/μ.

We shall next illustrate the response time variance of a round-robin system for different service distributions. The recursive algorithm in Theorem 4 is applied to calculate the second-order statistics of t_r for the following customer types.

Type 1. The service requirements of customers have a coefficient of variation (CV) approximately equal to one. The probability of a customer requiring r quanta of service is given by the following truncated geometric distribution

$$a_r = \begin{cases} (1-p)\,p^{r-1} & r = 1, 2, ..., 99 , \\ p^{99} & r = 100 \end{cases}$$

where $p = 0.95$. The mean number of service quanta required is equal to

$(1 - p^{100})/(1 - p) = 19.88 \cong 20.$

Type 2. The service requirements of customers have a large CV. The probability of a customer requiring r quanta of service is

$$a_r = \begin{cases} \frac{80}{99} & r = 1 , \\ \frac{19}{99} & r = 100 , \\ 0 & \text{otherwise} . \end{cases}$$

The mean number of service quanta required is 20.

Type 3. The service requirements of customers have a small CV. The probability of a customer requiring r quanta of service is

$$a_r = \begin{cases} \frac{1}{3} & r = 19, 20, 21 , \\ 0 & \text{otherwise} . \end{cases}$$

The mean number of service quanta required is 20.

The mean quantum size $(1/\mu)$ is assumed to be 0.05 second so that the mean service requirement is exactly equal to 1 second for type 2 and type 3 customers and approximately equal to 1 second for type 1 customers.

We shall first consider systems with a single customer type.

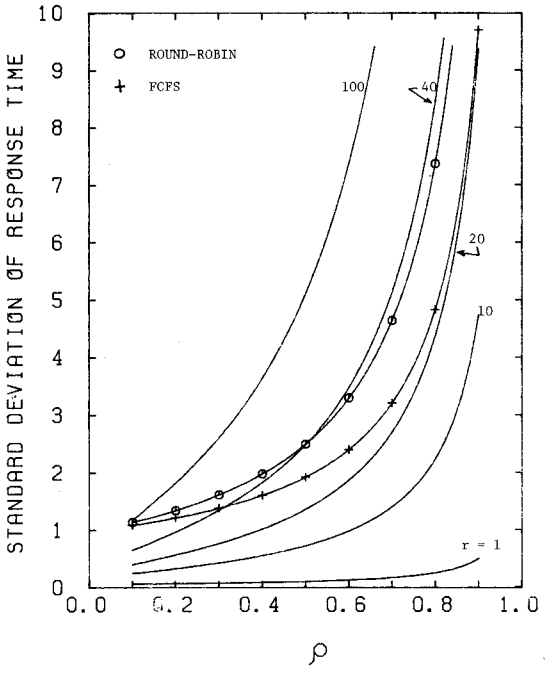


Fig. 2. Standard deviation of response time versus ρ for type 1 customers.

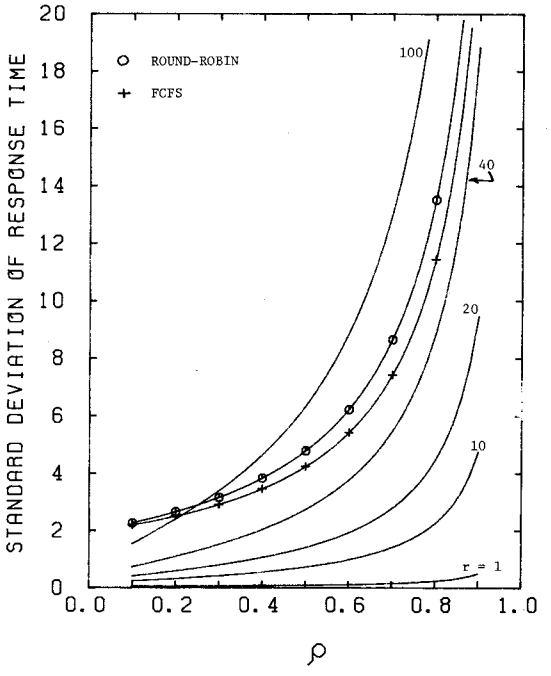


Fig. 3. Standard deviation of response time versus ρ for type 2 customers.

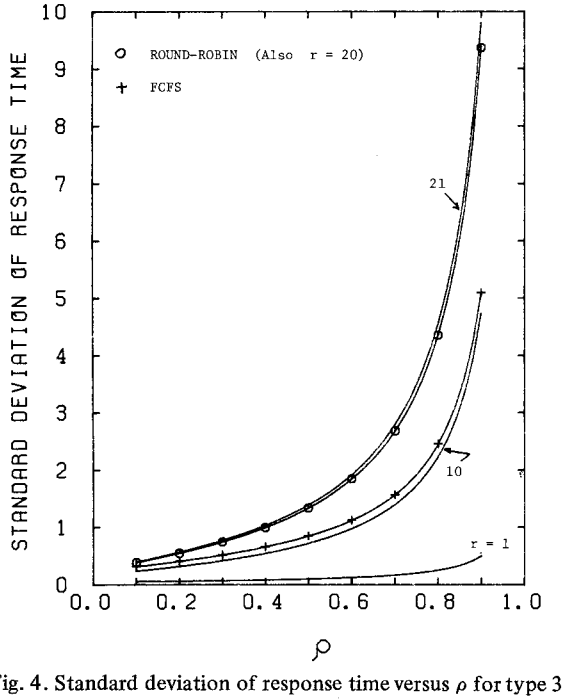


Fig. 4. Standard deviation of response time versus ρ for type 3 customers.

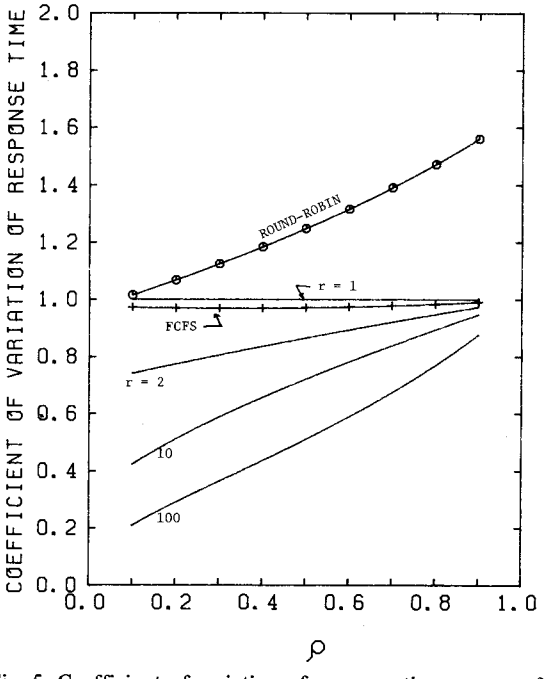


Fig. 5. Coefficient of variation of response time versus ρ for type 1 customers.

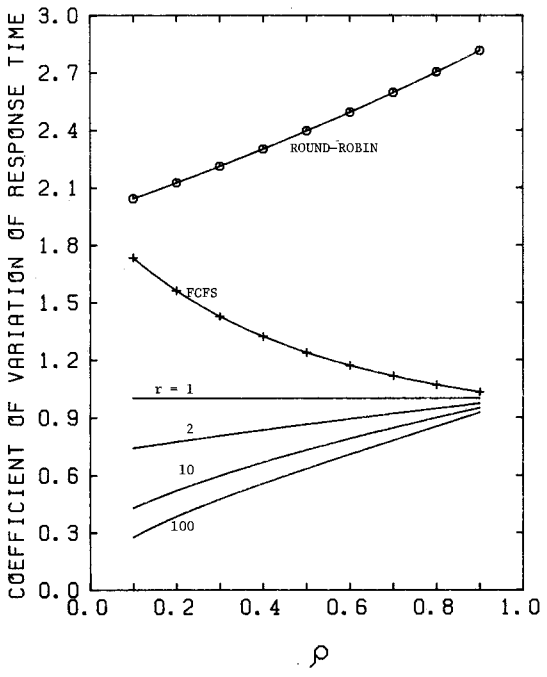


Fig. 6. Coefficient of variation of response time versus ρ for type 2 customers.

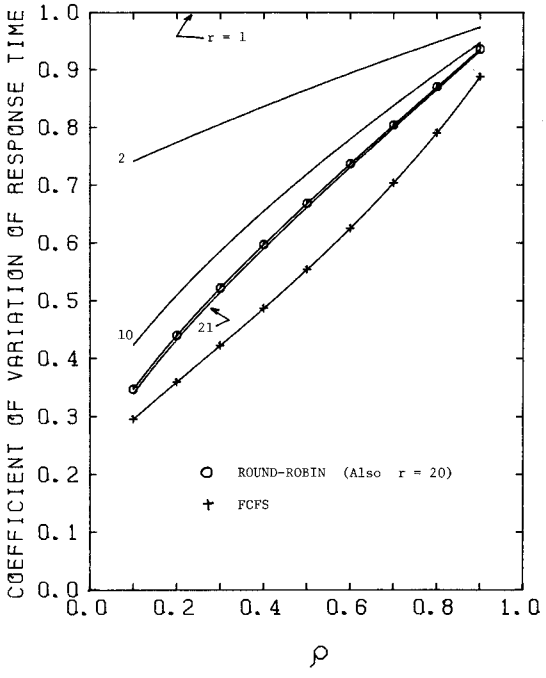


Fig. 7. Coefficient of variation of response time versus ρ for type 3 customers.

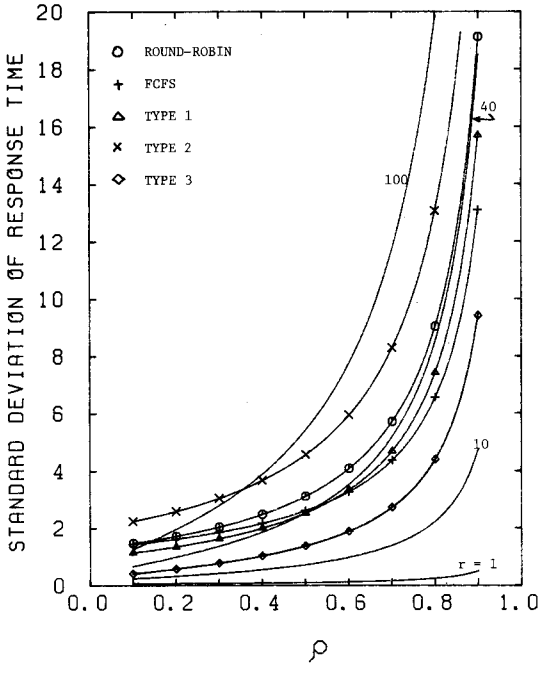


Fig. 8. Standard deviation of response time versus ρ for types 1–3 customers combined.

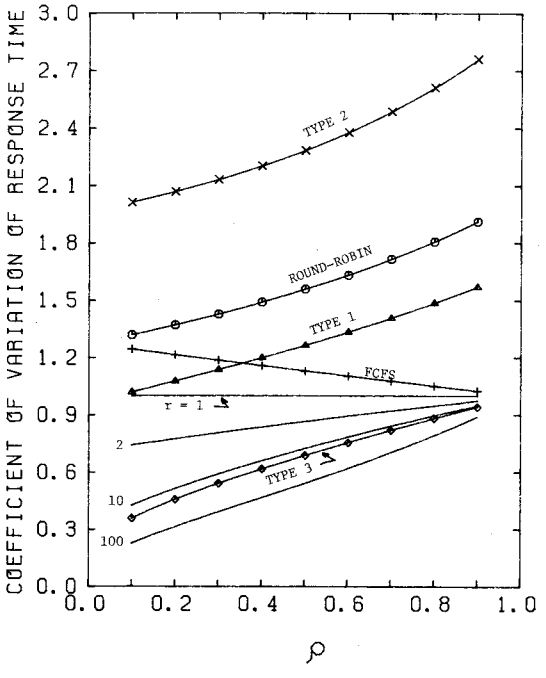


Fig. 9. Coefficient of variation of response time versus ρ for types 1–3 customers combined.

The standard deviation (SD) of t_r is plotted versus ρ for several values of r in Fig. 2 for a round-robin system serving only type 1 customers. For comparison, two additonal curves are also plotted. One is the SD of the response times of all customers in the round-robin system. The other is the SD of response times of all customers in a FCFS system (i.e., a queue with no feedback; a customer requiring r quanta of service gets all of them together). Corresponding results are shown in Figs. 3 and 4 for systems serving type 2 and type 3 customers respectively.

Note in Figs. 2, 3 and 4 that for all 3 customer types, the SD of the response times of all customers in a FCFS system is smaller than that in a round-robin system at any value of ρ.

For type 1 customers (i.e., service requirements with CV $\cong$ 1). Fig. 2 shows that the SD of t_r increases as r increases and is smaller than the SD of FCFS response times for small values of r. The exact crossover point is dependent upon the server utilization ρ under consideration. (Note that as ρ increases, the crossover point occurs at a small value of r.) Thus, the round-robin discipline benefits customers with small service requirements (small r) at the expense of customers with large service requirements (large r) in terms of the SD, as well as the mean value, of their response times.

Similar observations can be made in Fig. 3 for type 2 customers (i.e., service requirements with CV > 1). In fact, for ρ smaller than about 0.21 in Fig. 3, observe that the SD of t_r is smaller than the SD of FCFS for all r.

On the other hand for type 3 customers (i.e., service requirements with CV $\ll 1$), note in Fig. 4 that the $r = 1$ and $r = 10$ curves do not correspond to any customers that actually exit from the system after that many quanta of service. All customers require $r = 19$, 20, or 21 quanta of service. Hence the SD of FCFS response times is not only smaller than the response time SD of all customers in a round-robin system but also the SD of conditional response times of any substream of customers.

The CV of response times is plotted in Figs. 5–7 for systems serving each of the 3 customer types. We make the following observations:

(1) The CV of response times of all customers in a round-robin system is larger than that in a FCFS system in each case for any server utilization ρ.

(2) In a FCFS system, the response time CV is equal to (greater than, smaller than) one when the service requirement CV of customers is equal to (greater than, smaller than) one. In the case of 'greater than' ('smaller than'), the response time CV decreases (increases) to one as ρ is increased to one.

(3) In a round-robin system, the response time CV always increases as ρ increases. The CV of t_r (for $r \geqslant 2$) is less than 1 and increases to 1 as $\rho \uparrow 1$. The CV of t_r for $r = 1$ is unity. (The system behaves like an M/M/1 queue for those customers requiring exactly 1 service quantum.)

In conclusion, our observations indicate that the round-robin discipline provides better response time performance (both in terms of mean value and variance) than FCFS to customers with small service requirements at the expense of customers with large service requirements. The value of r at which this crossover takes place (for the SD of response times) is a function of the server utilization ρ. Also, for customers whose service requirements have a very small CV, the FGFS discipline is to be preferred.

Finally, to illustrate the generality of our model, we consider a system that handles all 3 types of customers together. The arrival process of each type of customers is assumed to have the same rate. The SD and CV of response times are plotted versus ρ in Figs. 8 and 9 respectively. The CV of the service requirements of all customers combined is larger than 1. In addition to the cases considered previously, the SD and CV curves of customers belonging to each of the 3 types are also plotted.

5. Conclusions

We analyzed a feedback queueing system with multiple customer types. The service requirements (in number of service quanta) of each type of customers have a general probability distribution with finite support. We assumed that service quanta are exponentially distributed. The contribution of this paper is the derivation of the

moment generating function of the conditional response time to achieve r quanta of service.

By assuming that each quantum of service is exponentially distributed, the multi-class feedback queue considered is an open queueing network satisfying local balance. Each type of jobs corresponds to customers following a fixed path. The key idea in our solution approach is to develop a recursive relationship between the response time of a path and the response time of the same path extended by one more transition. We hope that this solution approach can be generalized in the future to characterize the response time behavior in a network of queues.

Appendix

Proof of theorem 4.

$$E[t_{r+1}^2] = \frac{\partial^2}{\partial s^2}\,[y_1(s,z)\,U_r^*(s,z)]\Big|_{s=0,z=1} = \frac{\partial}{\partial s}\left\{\frac{\partial}{\partial s}\,\{y_1(s,1)\,E[e^{-st_r}(y_1(s,1))^{M^{(r)}}]\}\right\}_{s=0}$$

$$\left(\text{where } y_1(s,1) = \frac{1}{1+(s/\mu)}\right)$$

$$= \frac{\partial}{\partial s}\left\{-\frac{1}{\mu}(y_1(s,1))^2\,E[e^{-st_r}(y_1(s,1))^{M^{(r)}}]\right.$$

$$\left. + y_1(s,1)\,E\left[-t_r\,e^{-st_r}(y_1(s,1))^{M^{(r)}} + M^{(r)}\,e^{-st_r}y_1(s,1)^{M^{(r)}+1}\left(-\frac{1}{\mu}\right)\right]\right\}_{s=0}$$

$$= \left(-\frac{1}{\mu}\right)\left\{-E[t_r] + \left(-\frac{1}{\mu}\right)(E[M^{(r)}]+2)\right\} - E\left[t_r\left\{-t_r + \left(-\frac{1}{\mu}\right)(M^{(r)}+1)\right\}\right]$$

$$+ E\left[M^{(r)}\left(-\frac{1}{\mu}\right)\left\{-t_r + \left(-\frac{1}{\mu}\right)(M^{(r)}+2)\right\}\right]$$

$$= \left\{\frac{E[t_r]}{\mu} + \frac{E[M^{(r)}]}{\mu^2} + \frac{2}{\mu^2}\right\} + \left\{E[t_r^2] + \frac{E[t_rM^{(r)}]}{\mu} + \frac{E[t_r]}{\mu}\right\} + \left\{\frac{E[t_rM^{(r)}]}{\mu} + \frac{E[(M^{(r)})^2]}{\mu^2} + \frac{2E[M^{(r)}]}{\mu^2}\right\}$$

$$= \left\{\frac{2}{\mu^2} + \frac{2E[t_r]}{\mu} + \frac{3E[M^{(r)}]}{\mu^2} + \frac{E[(M^{(r)})^2]}{\mu^2}\right\} + \frac{2}{\mu}E[t_rM^{(r)}] + E[t_r^2]$$

(where the terms bracketted by { } can be evaluated using (11) and (13))

$$= \frac{2(1+r(1-\rho))}{\mu^2(1-\rho)^2} + \frac{2}{\mu}E[t_rM^{(r)}] + E[t_r^2]\,.$$

Substituting $E[t_r^2] = \mathrm{Var}(t_r) + (E[t_r])^2$ into the above yields (14).

The validity of (15) is obvious since $M^{(r+1)} = \Sigma_{i=1}^{R} m_i^{(r+1)}$. To solve for (16) below for $1 \leqslant i \leqslant R$, we shall interpret $m_{R+1}^{(r)}$ as zero.

$$E[t_{r+1}m_i^{(r+1)}] = -\frac{\partial^2}{\partial s\,\partial z_i}\,\{y_1(s,z)\,U_r^*(s,y(s,z))\}\Big|_{s=0,z=1}$$

$$= \left[-\frac{\partial}{\partial z_i}\left\{\left[\frac{\partial}{\partial s}y_1(s,z)\right]\cdot U_r^*(s,y(s,z)) + y_1(s,z)\cdot\right.\right.$$

$$\left.\left.\times\, E\left[z_1^{m_2^{(r)}}\cdots z_{R-1}^{m_R^{(r)}}\frac{\partial}{\partial s}\,\{e^{-st_r}(y_1(s,z))^{M^{(r)}}\}\right]\right\}_{s=0,z_j=1\,\text{for}\,j\neq i}\Bigg]_{z_i=1}$$

$$= \frac{\partial}{\partial z_i}\left\{\frac{1}{\mu}(y_1(z_i))^2 E[z_i^{m_{i+1}^{(r)}}(y_1(z_i))^{M^{(r)}}]\right.$$

$$\left. + y_1(z_i) E\left[z_i^{m_{i+1}^{(r)}}\left\{t_r(y_1(z_i))^{M^{(r)}} + M^{(r)}(y_1(z_i))^{M^{(r)}-1}\cdot\frac{1}{\mu}\cdot(y_1(z_i))^2\right\}\right]\right\}_{z_i=1}$$

$$\left(\text{where } y_1(z_i) = y_1(0,z)\Big|_{z_j=1 \text{ for} j\neq i} = \frac{1}{1+(\gamma_i/\mu)(1-z_i)}\right)$$

$$= \frac{2\gamma_i}{\mu^2} + \frac{1}{\mu}\left\{E[m_{i+1}^{(r)}] + E[M^{(r)}]\frac{\gamma_i}{\mu}\right\} + \frac{\gamma_i}{\mu}\left\{E[t_r] + \frac{1}{\mu}E[M^{(r)}]\right\} + \left\{\frac{\gamma_i}{\mu}E[t_r M^{(r)}] + \frac{\gamma_i}{\mu^2}E[M^{(r)}(M^{(r)}+1)]\right\}$$

$$+ \left\{E[t_r m_{i+1}^{(r)}] + \frac{1}{\mu}E[M^{(r)} m_{i+1}^{(r)}]\right\}$$

$$= \left\{\frac{2\gamma_i}{\mu^2} + \frac{1}{\mu}E[m_{i+1}^{(r)}] + \frac{3\gamma_i}{\mu^2}E[M^{(r)}] + \frac{\gamma_i}{\mu}E[t_r] + \frac{\gamma_i}{\mu^2}E[(M^{(r)})^2]\right\} + \frac{\gamma_i}{\mu}E[t_r M^{(r)}] + E[t_r m_{i+1}^{(r)}]$$

where the terms bracketted by { } can be evaluated using (11) and (13) to yield (16). (Q.E.D.)

References

[1] F. Baskett, K. Chandy, R. Muntz and F. Palacios, Open, closed and mixed networks of queues with different classes of customers, J. ACM 22 (1975).

[2] L. Kleinrock, Time-shared systems: A theoretical treatment, J.ACM 14 (April 1967).

[3] M. Sakata, S. Noguchi and J. Oizumi, An analysis of the M/G/1 queue under round-robin discipline, Operations Res. 19 (1971).

[4] E. Coffman, R. Muntz and H. Trotter, Waiting time distributions for processor-sharing systems, J.ACM 17 (January 1970).

[5] R. Muntz, Waiting time distribution for round-robin queueing systems, Proc. Symp. on Computer-Communications Networks and Teletraffic, Polytechnic Institute of Brooklyn (April 1972).

[6] L. Takács, A single-server queue with feedback, The Bell System Techn. J. (March 1963).

[7] J.W. Wong, Distribution of end-to-end delay in message-switched networks, Computer Networks 2 (1978).

[8] We-min Chow, The cycle time distribution of exponential cyclic queues, J.ACM (April 1980).

[9] P.S. Yu, Stochastic modelling of computer systems and networks, Ph.D. dissertation, Stanford University (1978).

[10] J.W. Cohen, The Single Server Queue (North-Holland, Amsterdam, 1969).

[11] L. Kleinrock, Queueing Systems, Vol. 1: Theory (Wiley-Interscience, New York, 1975) p. 118.

[12] M. Reiser and H. Kobayashi, Queueing networks with multiple closed chains: Theory and computational algorithms, IBM J. Res. and Develop. 19 (May 1975).