A queueing network with a single cyclically roving server

Moshe Sidi*
Electrical Engineering Department, Technion – Israel Institute of Technology, Haifa 32000, Israel

Hanoch Levy*
Computer Science Department, Tel-Aviv University, Tel-Aviv 69978, Israel

Steve W. Fuhrmann
IBM Research Division, Zurich Research Laboratory, Säumerstrasse 4, CH-8803 Rüschlikon, Switzerland

Received 13 August 1991; revised 30 December 1991

A queueing *network* that is served by a *single server* in a cyclic order is analyzed in this paper. Customers arrive at the queues from outside the network according to independent Poisson processes. Upon completion of his service, a customer may *leave* the network, be *routed* to another queue in the network or *rejoin* the same queue for another portion of service. The single server moves through the different queues of the network in a cyclic manner. Whenever the server arrives at a queue (polls the queue), he serves the waiting customers in that queue according to some service discipline. Both the gated and the exhaustive disciplines are considered. When moving from one queue to the next queue, the server incurs a switch-over period. This queueing network model has many applications in communication, computer, robotics and manufacturing systems. Examples include token rings, single-processor multi-task systems and others. For this model, we derive the generating function and the expected number of customers present in the network queues at arbitrary epochs, and compute the expected values of the delays observed by the customers. In addition, we derive the expected delay of customers that follow a specific route in the network, and we introduce pseudo-conservation laws for this network of queues.

Keywords: Polling systems, routing, queueing networks, single server.

*The work of this author was supported by the Bernstein Fund for the Promotion of Research and by the Fund for the Promotion of Research at the Technion.
*Part of this work was done while H. Levy was with AT&T Bell Laboratories.

1. Introduction

A queueing *network* that is served by a *single server* in a cyclic order is analyzed in this paper. Customers arrive at the queues from outside the network according to independent Poisson processes. Upon completion of his service, a customer may *leave* the network, be *routed* to another queue in the network or *rejoin* the same queue for another portion of service. The single server moves through the different queues of the network in a cyclic manner, and incurs switch-over times while moving between queues. The server either follows a gated discipline at each queue in the network or follows an exhaustive discipline at each queue in the network.

The introduction of *customer routing* considerably extends the modeling capability of polling systems, since in many applications customers require service at more than one facility of the system. In fact, networks of queues with a single server arise naturally in various models of computer, communication and robotics systems. One example is a local area network in which terminals are interconnected in either a physical or logical structure. Under the token ring protocol, a signal circulates around the ring and controls the network transmission. When the token comes to a terminal, that terminal transmits its messages, and no other terminal can transmit at the same time. When that terminal completes its transmission, the token proceeds to the next terminal. When one (or more) of the nodes of the ring is a file server, a request at some node for a file generates a job at the file server and this can be modeled only via customer routing. We elaborate on this application in section 8.

Another example in the context of a token ring is the modeling of a selective-repeat ARQ protocol, used to recover from transmission errors. In a selective-repeat ARQ protocol, a station that receives an erroneous message transmits a negative acknowledgement to the transmitting station to indicate that the message has to be retransmitted. To model this protocol via customer routing, each station is represented by two queues, one for messages that need to be sent out and one for negative acknowledgements to be sent back when erroneous messages are received. Packets transmitted to a station join its acknowledgement queue if their transmission fails, and a negative acknowledgement arriving to a station indicates that a message has to be retransmitted (so a message is generated).

Another typical example is a computer system where a single processor often has the responsibility of performing many distinct tasks, such as computation, sending information to memory, retrieve information, etc. Since the processor can only do one thing at a time, queues of tasks will be built up. Furthermore, a single application of the system may require more than one task to be performed, as in the case of getting input, adding it to retrieved data and print it. If different tasks are organized in different queues and these queues are served cyclically, then the above model enables the analysis of such a system. Other examples involving the use of polling models with customer routing can be found in Sarkar and Zangwill [26] and Levy and Sidi [20].

Networks of queues with a single server have been considered in the past. Klimov [15] studied the problem of dynamically allocating the server to the queues (without switch-over times) in such a way as to minimize some objective function. Several studies (Nair [22], Taube-Netto [35]) have been devoted to the queue-length analysis for the special case of tandem configurations. Another special case configuration (two queues feeding a third queue) has been considered by Katayama [12]. General networks where the queues are served by a single server according to a fixed priority discipline have been considered by Sidi and Segall [29] and by Simon [30]. Fuhrmann [8] earlier analyzed a class of polling models that includes the model in this paper as a special case. This work, however, was not published in the open literature. Moreover, the analysis in our paper is more comprehensive (in terms of our particular model) in several respects.

Polling systems in which customers always leave the system upon completion of their service (i.e., systems without customer routing) have been considered in numerous papers. The analysis approach used in this paper, the *buffer occupancy approach*, has been developed in the past for analyzing those systems. Cooper and Murray [5] and Cooper [4] used this approach to analyze continuous time cyclic systems with zero switch-over periods. Later, the method was used by Eisenberg [6], Hashida [10], Konheim and Meister [16], Swartz [31], Rubin and De Moraes [24], Levy [19] and others to analyze cyclic polling systems with non-zero switch-over periods, and by Kleinrock and Levy [14] for random order polling systems. Recently, it was used by Levy and Sidi [21] to analyze polling systems with correlated arrivals. Additional related work (regarding other variations of polling systems and different analysis approaches) may be found in the tutorials by Takagi [34] and Levy and Sidi [20].

The paper is organized as follows. In section 2, we introduce the model of the network under consideration. In section 3 and section 4, we analyze the gated and the exhaustive service disciplines, respectively. For both disciplines, we derive the first and second moments of the number of customers present in the queues at polling instants. Then, the expected number of customers present in the queues at arbitrary epochs is derived, and from Little's law, we obtain the expected customer delays. Section 5 is devoted to the computation of the expected delay of customers that follow specific routes in the network. A pseudo-conservation law for the network is derived in section 6. A similar law has been independently derived by Boxma [2]. We conclude the paper with a numerical example (section 7), an application of our model (section 8), and a discussion of the results (section 9). A summary of notation is provided in section 10.

2. Model description and preliminaries

Consider a polling system that consists of N infinite buffer queues that are served by a single server. The queues are indexed $1, 2, \ldots, N$ (and denoted $Q_1, Q_2, \ldots, Q_N$) and for simplicity of notation, all references to queue index are

implicitly assumed to be modulo N; this includes, e.g., sums such as $\sum_{k=j}^{i}$ that should be interpreted as $\sum_{k=j}^{N} + \sum_{k=1}^{i}$ if $i < j$.

The service order is cyclic: after completing servicing queue i, the server incurs a switch-over (walking) period (whose length is associated with queue i) and moves to serve queue $i+1$. Two *service disciplines* are considered: (i) *exhaustive service*, in which when the server polls queue i it will serve this queue until its buffer empties; (ii) *gated service*, in which when the server polls queue i it will serve all the customers it found in queue i at the polling instant. The period at which the server serves queue i is called a *service period* of queue i.

Customers arrive at queue i from outside the network according to a Poisson process with rate λ_i. These customers are called *Poisson* or *external customers*. The service time required by the customers that reside in queue i (*type-i customers*) is a random variable B_i (*type-i service*) with a general distribution, having LST $B_i^*(s)$, mean b_i and second moment $b_i^{(2)}$. All service times and arrival processes are assumed to be mutually independent.

Upon completion of his service at queue i, a customer leaves the network with probability $P_{i,0}$, or is routed to queue j, $j = 1, 2, \ldots, N$ with probability $P_{i,j}$. Note that this model can be used to capture feedback mechanisms ($P_{i,i} \neq 0$); in this case, a customer rejoins the queue he departs from (the exact placement is not important in the analysis we carry out in this paper). Obviously, we have $\sum_{j=0}^{N} P_{i,j} = 1$. The transition from queue i to queue j is assumed to take no time at all.

The total arrival rate of customers to queue i is denoted by γ_i. This total arrival process into queue i is composed of arrivals of external customers (λ_i), arrivals of customers that are routed to queue i from other queues and customers that complete their service at queue i and rejoin that queue. The total arrival rates γ_i, $i = 1, 2, \ldots, N$, may be calculated by solving the following set of linear equations:

$$\gamma_i = \lambda_i + \sum_{j=1}^{N} \gamma_j P_{j,i}, \qquad i = 1,2,\ldots,N. \tag{2.1}$$

The total rate of customers routed from queue i to queue j is given by $\gamma_{i,j} = \gamma_i P_{i,j}$. In the following, we exclude from the analysis degenerate queues for which $\gamma_i = 0$.

The offered load to queue i is defined as $\rho_i \triangleq \gamma_i b_i$ and the total network utilization is $\rho = \sum_{i=1}^{N} \rho_i$. The *total service time* of a customer is the total amount of service given to the customer during his presence in the network. The total service time consumed by a customer that enters the network at node i is denoted by $\tilde{B}_i$. The mean and the second moments of $\tilde{B}_i$ are denoted by $\tilde{b}_i$ and $\tilde{b}_i^{(2)}$, and can be calculated from the following sets of linear equations:

$$\tilde{b}_i = b_i + \sum_{j=1}^{N} \tilde{b}_j P_{i,j}, \qquad i = 1,2,\ldots,N \tag{2.2}$$

and

$$\tilde{b}_i^{(2)} = b_i^{(2)} + 2b_i \sum_{j=1}^{N} \tilde{b}_j P_{i,j} + \sum_{j=1}^{N} \tilde{b}_j^{(2)} P_{i,j}, \qquad i = 1,2,\ldots,N. \tag{2.3}$$

The switch-over period following the service of queue i is called a *type-i switch-over period.* Its duration is a random variable R_i, independent of the other network variables, and having a general distribution with LST $R_i^*(s)$, mean r_i and second moment $r_i^{(2)}$. We denote $r = \sum_{i=1}^{N} r_i$ and $r^{(2)} = E[(\sum_{i=1}^{N} R_i)^2]$.

Note that $\rho_i = \gamma_i b_i$ is the mean amount of type-i work entering the system (from outside) per time unit. In order for the system to be stable, $\rho = \sum_{i=1}^{N} \gamma_i b_i = \sum_{i=1}^{N} \lambda_i \tilde{b}_i$ should be smaller than one. This condition does not depend on the switch-over periods. The reason is that once the server arrives to a queue it is dedicated to serve all the work accumulated in the queue without any interruption. Thus, when the system becomes heavily loaded, the time wasted during switch-over periods is negligible compared to the time the server is busy in service.

In the analysis of the gated and the exhaustive disciplines, we use the buffer occupancy method. With this method, we are able to derive the generating functions (GF) of the number of customers present at the system at polling instants, and the first and the second moments of the queue size distributions at polling instants. From these, we derive the mean value for the queue distribution at arbitrary moments.

In the following, we let X_i^j denote the number of customers residing at queue j when queue i is polled and define:

$$F_i(z) = F_i(z_1, z_2, \ldots, z_N) \triangleq E\left[z_1^{X_i^1} z_2^{X_i^N} \ldots z_N^{X_i^N}\right], \tag{2.4}$$

$$f_i(j) \triangleq E[X_i^j], \tag{2.5}$$

$$f_i(j,k) \triangleq \begin{cases} E[X_i^j X_i^k], & j \neq k, \\ E[X_i^j (X_i^j - 1)], & j = k. \end{cases} \tag{2.6}$$

In section 3, we derive these quantities for the gated discipline and in section 4, we do the same for the exhaustive discipline. In addition, we obtain expressions for the expected number of customers in a queue at an arbitrary moment and the expected waiting time of a customer in a queue.

3. The gated discipline

Part of the analysis of the gated discipline appeared in Sidi and Levy [28]. In this section, we state the main results together with some new results that are derived here.

To describe the evolution of the queue lengths at polling instants, we introduce the following notation. The number of Poisson (external) customers arriving to

queue j during a switch-over period R_i is denoted by R_i^j. The number of Poisson customers arriving to queue j during a service period of queue i is denoted by A_i^j. The number of customers that join queue j after completing their service at queue i during a service period of queue i is denoted by T_i^j. The basic relation that holds in the polling network with gated service discipline is:

$$X_{i+1}^j = \begin{cases} X_i^j + A_i^j + T_i^j + R_i^j, & i \neq j, \\ A_i^i + T_i^i + R_i^i, & i = j. \end{cases} \tag{3.1}$$

The explanation of (3.1) is simple: With the gated service discipline, the number of customers at queue j when queue $i+1$ is polled ($j \neq i$) is the sum of the number of customers that were at queue j when queue i was polled, the number of customers that arrived to queue j during the service period of queue i (external arrivals and transiting customers) and the number of customers that arrived to queue j during a type-i switch-over period. The case $i = j$ does not contain the variable X_i^i since these customers are served during the service period of queue i.

We now derive the generating functions (GF) of the number of customers present at the system at polling instants and at arbitrary moments. We recall that $F_i(z) = F_i(z_1, z_2, \ldots, z_N)$ is the joint probability generating function of the number of customers present at the system at arbitrary polling instants of queue i. Similarly, denote the joint distribution GF at arbitrary switch-over instants (end of service periods), service beginning instants and service completion instants of queue i by $\overline{F}_i(z)$, $V_i(z)$ and $\overline{V}_i(z)$, respectively. Also, let $F^*(z)$ be the joint probability GF for the number of customers present at the system at arbitrary moments. Let $P_i(z) = P_{i,0} + \sum_{j=1}^{N} P_{i,j} z_j$.

For the GF of the number of customers present at the system at polling instants, we derive the following relation:

$$\overline{F}_i(z_1, z_2, \ldots, z_N) = F_i(z_1, \ldots, z_{i-1}, \xi_i, z_{i+1}, \ldots, z_N), \tag{3.2a}$$

$$F_{i+1}(z_1, z_2, \ldots, z_N) = F_i(z_1, \ldots, z_{i-1}, \xi_i, z_{i+1}, \ldots, z_N) R_i^* \left[\sum_{j=1}^{N} (\lambda_j - \lambda_j z_j) \right], \tag{3.2b}$$

where the ith argument is given by

$$\xi_i \triangleq P_i(z) B_i^* \left[\sum_{j=1}^{N} (\lambda_j - \lambda_j z_j) \right].$$

Note that $F_i(z)$ can be determined explicitly in terms of an infinite product as in Eisenberg [6].

To derive $F^*(z)$, we use the approach presented by Eisenberg [6]. Translating his eqs. (57), (58) and (59) to our notation and formulation, we obtain the following respective relations:

$$F_i(z) + f_i(i)\overline{V}_i(z) = f_i(i)V_i(z) + \overline{F}_i(z), \tag{3.3}$$

$$\overline{V}_i(z) = \left[B_i^*\left(\sum_{j=1}^{N} (\lambda_j - \lambda_j z_j) \right) \cdot P_i(z)/z_i \right] \cdot V_i(z), \tag{3.4}$$

$$F_i(z) = \overline{F}_{i-1}(z) \cdot R_i^*\left(\sum_{j=1}^{N} (\lambda_j - \lambda_j z_j) \right). \tag{3.5}$$

From these equations, we obtain an expression for the GF of the number of customers present at service beginning instants:

$$V_i(z) = \frac{(\overline{F}_i(z) - F_i(z))\dfrac{z_i}{f_i(i)}}{B_i^*\left(\sum_{j=1}^{N}(\lambda_j - \lambda_j z_j)\right) \cdot P_i(z) - z_i}. \tag{3.6}$$

Now, the GF of the number of customers present at arbitrary instants of service period i and arbitrary instants of switch-over period i are given by:

$$F^*(z|\text{service period } i) = V_i(z)\,\frac{1 - B_i^*\left(\sum_{j=1}^{N}(\lambda_j - \lambda_j z_j)\right)}{b_i \sum_{j=1}^{N}(\lambda_j - \lambda_j z_j)}, \tag{3.7a}$$

$$F^*(z|\text{switch-over period } i) = \overline{F}_i(z)\,\frac{1 - R_i^*\left(\sum_{j=1}^{N}(\lambda_j - \lambda_j z_j)\right)}{r_i \sum_{j=1}^{N}(\lambda_j - \lambda_j z_j)}, \tag{3.7b}$$

and finally:

$$F^*(z) = \frac{1}{c}\left[\sum_{i=1}^{N} [f_i(i) b_i F^*(z|\text{service period } i) + r_i F^*(z|\text{switch-over period } i)] \right]. \tag{3.8}$$

The expected number of customers at queue i when this queue is polled has been derived in Sidi and Levy [28] and is given by

$$f_i(i) = \gamma_i c = \frac{\gamma_i r}{1-\rho}, \qquad i = 1,2,\ldots,N, \tag{3.9}$$

where c is the expected duration of a cycle (the expected time between two consecutive polls of a queue) and is given by:

$$c = \frac{r}{1-\rho}. \tag{3.10}$$

The expected number of customers at queue k when queue j ($j \neq k$) is polled can be *directly* computed from (3.9) and the following relations: $f_{i+1}(j) = f_i(j) + f_i(i)(\lambda_j b_i + P_{i,j}) + \lambda_j r_i$, $1 \leq i, j \leq N$, $j \neq i+1$.

In Sidi and Levy [28], we also obtain a set of N^3 linear equations for the quantities $\{f_i(j,k)\}$. These equations can be obtained by differentiating (3.2b) or by a direct computation (see Sidi and Levy [27]). In Sidi and Levy [28], we also discuss an efficient approach for solving the set of N^3 linear equations.

Having obtained the first and second moments of the queue size distributions at polling instants, we can now derive the mean number of customers residing at the network queues at arbitrary moments. An approach which is commonly used in the analysis of polling systems is to derive this quantity by deriving the mean number of customers left behind by a departing customer. Due to the PASTA property (Poisson Arrivals See Time Averages) and the single-arrival single-departure property (at any instant there can be at most one arrival or departure), these two quantities usually coincide. This approach, however, cannot be used in our system. The reason is that transiting customers do not form a Poisson stream, and thus the PASTA property does not necessarily hold for this system. Therefore, we need to adopt an alternative approach for deriving the mean number of customers present in the system at arbitrary moments.

Let X_i^* be the number of customers residing at queue i at an arbitrary moment, and let Y_i be the duration of the service period in queue i. Then $E[X_i^*]$ can be calculated by conditioning on the specific period (service period or switch-over period of queue j) as follows:

$$E[X_i^*] = \frac{\sum_{j=1}^{N} E[X_i^* \mid \text{service period } j] \cdot E[Y_j] + \sum_{j=1}^{N} E[X_i^* \mid \text{switch-over period } j] \cdot E[R_j]}{E[C]}, \tag{3.11}$$

where C is the cycle length, $E[C] = c = r/(1-\rho)$, $E[R_j] = r_j$ and $E[Y_j] = f_j(j)b_j$.

The value of $E[X_{i/j}^*] = E[X_i^* \mid \text{service period } j]$ can be obtained by a direct analysis of the mean values of the random variables (see Sidi and Levy [27]) or by differentiating (3.7a):

$$E[X_{i/j}^*] = \begin{cases} \dfrac{\lambda_i[f_j(j)b_j^{(2)} + f_j(j,j)b_j^2]}{2f_j(j)b_j} + \dfrac{P_{j,i}f_j(j,j)}{2f_j(j)} + \dfrac{f_j(i,j)}{f_j(j)}, & i \neq j, \\ \dfrac{\lambda_i[f_i(i)b_i^{(2)} + f_i(i,i)b_i^2]}{2f_i(i)b_i} + \dfrac{P_{i,i}f_i(i,i)}{2f_i(i)} + \dfrac{f_i(i,i)}{2f_i(i)} + 1, & i = j. \end{cases} \tag{3.12}$$

The derivation of $E[X_i^* \mid \text{switch-over period } j]$ is done by differentiating (3.7b), yielding:

$$E[X_i^* \mid \text{switch-over period } j] = \begin{cases} f_j(i) + f_j(j)b_j\lambda_i + f_j(j)P_{j,i} + \dfrac{\lambda_i r_j^{(2)}}{2r_j}, & i \neq j, \\ f_i(i)b_i\lambda_i + f_i(i)P_{i,i} + \dfrac{\lambda_i r_i^{(2)}}{2r_i}, & i = j. \end{cases} \tag{3.13}$$

Now, as stated above, $E[X_i^*]$ can be derived from (3.11), (3.12) and (3.13). Finally, from $E[X_i^*]$, the mean sojourn time and mean waiting time at queue i can be simply calculated by using Little's law:

$$E[T_i] = E[X_i^*]/\gamma_i; \qquad E[W_i] = E[T_i] - b_i. \tag{3.14}$$

4. The exhaustive discipline

For the analysis of the exhaustive discipline, we keep the notation of the previous section. The basic relation that holds for the exhaustive discipline is

$$X_{i+1}^j = \begin{cases} X_i^j + A_i^j + T_i^j + R_i^j, & i \neq j, \\ R_i^i, & i = j, \end{cases} \tag{4.1}$$

and the explanation is similar to that of eq. (3.1).

To carry out the analysis of the number of customers present at arbitrary moments, it is convenient to assume that during a type-i service period, the type-i customers are served in an LCFS order. Moreover, we assume that when a type-i customer is routed back to queue i, he instantaneously joins the head of the queue and gets served immediately. Note that these assumptions do not affect the analysis of the number of customers present in the system, and they are used only to facilitate the presentation. Using these assumptions, a customer can be thought to have service time $\hat{B}_i$, which is the geometric sum (with parameter $P_{i,i}$) of independent random variables, each of which is distributed as B_i. When such a customer leaves queue i, he moves to queue j ($j \neq i$) with probability $\hat{P}_{i,j} = P_{i,j}/(1 - P_{i,i})$. Also, define $\hat{P}_{i,i} = 0$. The LST and first two moments of $\hat{B}_i$ are denoted by $\hat{B}_i^*(s)$, $\hat{b}_i$ and $\hat{b}_i^{(2)}$, and their values are:

$$\hat{B}_i^*(s) = (1 - P_{i,i})B_i^*(s)/[1 - P_{i,i}B_i^*(s)];$$

$$\hat{b}_i = \frac{b_i}{1 - P_{i,i}}: \qquad \hat{b}_i^{(2)} = \frac{b_i^{(2)}}{1 - P_{i,i}} + \frac{2P_{i,i}b_i^2}{(1 - P_{i,i})^2}.$$

In addition, let $\hat{\gamma}_i$ denote the total arrival rate at queue i. Under this formulation (namely, without counting the fed back customers as arrivals), we have:

$$\hat{\gamma}_i = \lambda_i + \sum_{j=1}^{N} \hat{\gamma}_j \hat{P}_{j,i}.$$

To derive the various generating functions, we follow the approach used in section 3 and use similar notation. We denote $\hat{P}_i(z) = \hat{P}_{i,0} + \sum_{j=1}^{N} \hat{P}_{i,j} z_j$. Now, let $G_i^*(z, s)$ be the joint probability generating function of the number of customers

served in a busy period at queue i and the LST for the length of the distribution function of the same busy period. This satisfies a well-known equation:

$$G_i^*(z,s) = z\hat{B}_i^*[s + \lambda_i - \lambda_i G_i^*(z,s)]. \tag{4.2}$$

Using $G_i^*(\cdot,\cdot)$, we derive the relation for the GF of the number of customers present at polling instants and switching instants:

$$\overline{F}_i(z_1, z_2, \ldots, z_N) = F_i(z_1, \ldots, z_{i-1}, \xi_i, z_{i+1}, \ldots, z_N), \tag{4.3a}$$

$$F_{i+1}(z_1, z_2, \ldots, z_N) = F_i(z_1, \ldots, z_{i-1}, \xi_i, z_{i+1}, \ldots, z_N) \cdot R_i^*\left[\sum_{j=1}^{N}(\lambda_j - \lambda_j z_j)\right], \tag{4.3b}$$

where the ith argument of $F_i(\cdot)$ is given by:

$$\xi_i = G_i^*\left[\hat{P}_i(z), \sum_{\substack{j=1 \\ j\neq i}}^{N}(\lambda_j - \lambda_j z_j)\right].$$

Now the derivation of $F^*(z)$ is similar to that of section 3, with the following simple change of variables: b_i changes to $\hat{b}_i$, $B_i^*(\cdot)$ to $\hat{B}_i^*(\cdot)$, γ_i to $\hat{\gamma}_i$ and $P_i(\cdot)$ to $\hat{P}_i(\cdot)$. So we finally have:

$$F^*(z|\text{service period } i) = \frac{(1-\lambda_i\hat{b}_i)(\overline{F}_i(z) - F_i(z))z_i}{f_i(i)\left(\hat{B}_i^*\left(\sum_{j=1}^{N}(\lambda_j - \lambda_j z_j)\right)\cdot \hat{P}_i(z) - z_i\right)} \frac{1 - \hat{B}_i^*\left(\sum_{j=1}^{N}(\lambda_j - \lambda_j z_j)\right)}{\hat{b}_i \sum_{j=1}^{N}(\lambda_j - \lambda_j z_j)}, \tag{4.4a}$$

$$F^*(z|\text{switch-over period } i) = \overline{F}_i(z)\frac{1 - R_i^*\left(\sum_{j=1}^{N}(\lambda_j - \lambda_j z_j)\right)}{r_i \sum_{j=1}^{N}(\lambda_j - \lambda_j z_j)}, \tag{4.4b}$$

$$F^*(z) = \frac{1}{c}\left[\sum_{i=1}^{N}[f_i(i)\hat{b}_i F^*(z|\text{service period } i) + r_i F^*(z|\text{switch-over period } i)]\right]. \tag{4.5}$$

The first moments of the buffer occupancy at polling instants can be derived from:

$$\frac{f_i(i)}{1 - b_i\lambda_i - P_{i,i}} = c\,\gamma_i, \tag{4.6}$$

in which the left-hand side represents the mean number of customers served in a service period of queue i and the right-hand side represents the mean number of type i customers entering queue i in a cycle. In addition, we have:

$$c = r + \sum_{i=1}^{N} \frac{f_i(i)b_i}{1 - b_i\lambda_i - P_{i,i}},$$

which, together with (4.6), can be solved to yield:

$$c = \frac{r}{1-\rho}; \qquad f_i(i) = \frac{r}{1-\rho}\,\gamma_i(1 - b_i\lambda_i - P_{i,i}). \tag{4.7}$$

To calculate the other buffer occupancy first moments, one may use a direct approach (Sidi and Levy [27]) or may differentiate (4.3b) to obtain:

$$f_{i+1}(j) = \begin{cases} r_i\lambda_i, & i = j, \\ r_i\lambda_j + f_i(j) + f_i(i)\dfrac{b_i\lambda_j + P_{i,j}}{1 - P_{i,i} - b_i\lambda_i}, & i \neq j. \end{cases} \tag{4.8}$$

Linear relations for the second moments of the buffer occupancy can be obtained by either a direct approach (Sidi and Levy [27]) or by differentiating (4.3b) to obtain:

$$\begin{aligned} f_{i+1}(j,k) &= f_i(j,k) + \lambda_j\lambda_k r_i^{(2)} + r_i\lambda_k f_i(j) + r_i\lambda_j f_i(k) \\ &\quad + [f_i(i,k) + f_i(i)r_i\lambda_k]\frac{b_i\lambda_j + P_{i,j}}{1 - \lambda_i b_i - P_{i,i}} + [f_i(i,j) + f_i(i)r_i\lambda_j]\frac{b_i\lambda_k + P_{i,k}}{1 - \lambda_i b_i - P_{i,i}} \\ &\quad + \frac{f_i(i)b_i[\lambda_j(1 - P_{i,i}) + \lambda_i P_{i,j}][P_{i,k}(1 - \lambda_i b_i) + P_{i,i}\lambda_k b_i]}{(1 - \lambda_i b_i - P_{i,i})^3} \\ &\quad + \frac{f_i(i)b_i[\lambda_k(1 - P_{i,i}) + \lambda_i P_{i,k}][P_{i,j}(1 - \lambda_i b_i) + P_{i,i}\lambda_j b_i]}{(1 - \lambda_i b_i - P_{i,i})^3} \\ &\quad + \frac{f_i(i)b_i^{(2)}[\lambda_j(1 - P_{i,i}) + \lambda_i P_{i,j}][\lambda_k(1 - P_{i,i}) + \lambda_i P_{i,k}]}{(1 - \lambda_i b_i - P_{i,i})^3} \\ &\quad + \frac{f_i(i,i)(b_i\lambda_j + P_{i,j})(b_i\lambda_k + P_{i,k})}{(1 - \lambda_i b_i - P_{i,i})^2}, \qquad i \neq j,\ i \neq k; \end{aligned} \tag{4.9a}$$

$$f_{i+1}(i,k) = \lambda_i\lambda_k r_i^{(2)} + r_i\lambda_i f_i(k) + f_i(i)\frac{b_i\lambda_k + P_{i,k}}{1 - \lambda_i b_i - P_{i,i}} r_i\lambda_i, \quad i \neq k; \tag{4.9b}$$

$$f_{i+1}(i,i) = \lambda_i^{(2)} r_i^{(2)}. \tag{4.9c}$$

We now derive the expected number of customers at arbitrary moments and the expected delays. The analysis approach and the notation used here are similar to those used in section 3. The value of $E[X^*_{i/j}]$ can be derived directly (see Sidi and Levy [27]) or by differentating (4.4a):

$$E[X^*_{i/j}] = \frac{\lambda_i \hat{b}_j^{(2)}}{2\hat{b}_j(1-\lambda_j\hat{b}_j)^2} + \frac{f_j(j,j)\lambda_i\hat{b}_j}{2f_j(j)(1-\lambda_j\hat{b}_j)} + \frac{f_j(i,j)}{f_j(j)}$$
$$+ \hat{P}_{j,i}\left\{\frac{\lambda_j^2\hat{b}_j^{(2)}}{2(1-\lambda_j\hat{b}_j)^2} + \frac{\lambda_j\hat{b}_j}{1-\lambda_j\hat{b}_j} + \frac{f_j(j,j)}{2f_j(j)(1-\lambda_j\hat{b}_j)}\right\}, \quad i \neq j, \quad (4.10a)$$

$$E[X^*_{i/i}] = 1 + \frac{\lambda_i\hat{b}_i^{(2)}}{2\hat{b}_i(1-\lambda_i\hat{b}_i)} + \frac{f_i(i,i)}{2f_i(i)}. \quad (4.10b)$$

The derivation of $E[X_i^* \mid \text{switch-over period } j]$ is similar to that of the gated system and the results are:

$$E[X_i^* \mid \text{switch-over period } j] = \begin{cases} f_j(i) + \dfrac{f_j(j)\hat{b}_j\lambda_i}{1-\lambda_j\hat{b}_j} + \dfrac{f_j(j)P_{j,i}}{(1-\lambda_j\hat{b}_j)(1-P_{i,i})} + \dfrac{\lambda_i r_j^{(2)}}{2r_j}, & i \neq j, \\ \dfrac{\lambda_i r_i^{(2)}}{2r_i}, & i = j. \end{cases} \quad (4.11)$$

Finally, we use the expression for the mean duration of the service period: $E[Y_i] = f_i(i)\hat{b}_i/(1-\lambda_i\hat{b}_i)$, and substitute (4.10a), (4.10b) and (4.11) into (3.11) to derive $E[X_i^*]$. From this value, we may now derive the mean waiting time and the mean sojourn time at Q_i, as done for the gated system, by using Little's result.

5. Path times

In the previous section, we derived the mean delay observed by the customers in the different queues. In this section, we are interested in the mean *path time* which is the mean amount of time spent by an arbitrary customer traversing a specific path. This can be used to derive the sojourn time of specific customers in the system. Our goal in this section is to present the analysis approach and thus we derive it only for the gated system. The analysis of the exhaustive system can be done in a similar way.

Consider a tagged customer which traverses the path $Q_{i_1}, Q_{i_2}, \ldots, Q_{i_M}$, where $i_k \in \{1, 2, \ldots, N\}$. Let $T_{i_1,\ldots,i_M}$ be the time spent by the customer from his departure

from Q_{i_1} until his departure from Q_{i_M}. To conduct the analysis, we need to trace the server movement through the system while the customer is on his path. The *server's path* is:

$$Q_{i_1}, Q_{i_1+1}, \ldots, Q_{i_1+k_1}, Q_{i_2}, Q_{i_2+1}, \ldots, Q_{i_2+k_2}, \ldots, Q_{i_{M-1}}, Q_{i_{M-1}+1}, \ldots, Q_{i_{M-1}+k_{M-1}}, Q_{i_M},$$

where $k_j = i_{j+1} - i_j - 1$ (for $j = 1, \ldots, M-1$).

Our approach is to compute the mean number of customers present in each of the queues in the polling instants of the server's path, assuming that the tagged customer is an arbitrary customer served at Q_{i_1}. Let $g_i(j)$ be the mean number of customers present at Q_j when Q_i is polled, conditioned on the fact that the tagged customer is present at Q_{i_1} when it is polled and that this customer traverses the path $Q_{i_1}, \ldots, Q_{i_M}$. Also, let $\bar{g}_{i_k}(i_k)$ $(k = 1, \ldots, M)$ be the mean number of customers which are present at Q_{i_k}, ahead of the tagged customer, when Q_{i_k} is polled.

To derive these values, we realize that since the tagged customer is an arbitrary customer visiting Q_{i_1}, the mean number of customers when Q_{i_1} is polled (and the tagged customer is there) is:

$$g_{i_1}(j) = \frac{f_{i_1}(i_1, j)}{f_{i_1}(i_1)}, \quad j = 1, \ldots, N, \quad j \neq i_1;$$

$$g_{i_1}(i_1) = \frac{f_{i_1}(i_1, i_1) + f_{i_1}(i_1)}{f_{i_1}(i_1)}. \tag{5.1}$$

We may now calculate recursively the other values of g. The first equation relates to polling instants of queues subsequent to queues on the customer path:

$$g_{i_m+1}(j) = g_{i_m}(j)\mathbf{1}(i_m \neq j) + g_{i_m}(i_m)b_{i_m}\lambda_j + [g_{i_m}(i_m) - 1]P_{i_m,j} + 1 \cdot \mathbf{1}(j = i_{m+1}) + r_{i_m}\lambda_j,$$

$$m = 1, \ldots, M; \quad j = 1, \ldots, N, \tag{5.2}$$

where $\mathbf{1}(A)$ is the indicator function, namely, its value is 1 if A holds true and zero otherwise. In (5.2), the first term represents the customers that were present at Q_j in the previous polling instant, the second term represents the customers which arrived (external arrivals) during the previous service period, the third and fourth terms represent the customers which transited from Q_{i_m} to Q_j, and the last term represents the customers which arrived during the previous switch-over period.

A similar equation is derived for all other queues on the server's path:

$$g_{i_m+l+1}(j) = g_{i_m+l}(j)\mathbf{1}(i_m + l \neq j) + g_{i_m+l}(i_m + l)b_{i_m+l}\lambda_j + g_{i_m+l}(i_m + l)P_{i_m+l,j} + r_{i_m+l}\lambda_j,$$

$$l = 1, \ldots, k_m; \quad m = 1, \ldots, M-1; \quad j = 1, \ldots, N. \tag{5.3}$$

The calculation of $\bar{g}_{i_m}(i_m)$ is done as follows:

$$\bar{g}_{i_1}(i_1) = \frac{g_{i_1}(i_1) - 1}{2}, \tag{5.4}$$

$$\bar{g}_{i_{m+1}}(i_{m+1}) = g_{i_m}(i_{m+1})\mathbf{1}(i_m \neq i_{m+1}) + \bar{g}_{i_m}(i_m)P_{i_m,i_{m+1}} + [\bar{g}_{i_m}(i_m) + 1]b_{i_m}\lambda_{i_{m+1}},$$

$$m = 1, \ldots, M-1. \tag{5.5}$$

Equation (5.4) simply reads that the customers ahead of the tagged customers are half of all the customers in the queue (minus himself). Equation (5.5) reads that the customers ahead of the tagged customer at $Q_{i_{m+1}}$ are those who were there when Q_{i_m} was polled, plus those who transited from Q_{i_m} before the tagged customer was served there, plus all the external arrivals that occurred at $Q_{i_{m+1}}$ during the service of Q_{i_m} and before the tagged customer transited to $Q_{i_{m+1}}$.

Using these variables, we may now compute $E[T_{i_1,\ldots,i_M}]$:

$$E[T_{i_1,\ldots,i_M}] = [g_{i_1}(i_1) - \bar{g}_{i_1}(i_1) - 1]b_{i_1}$$

$$+ \sum_{m=2}^{M-1} g_{i_m}(i_m)b_{i_m} + [\bar{g}_{i_M} + 1]b_{i_M} + \sum_{m=1}^{M-1}\sum_{l=1}^{k_m} g_{i_{m+l}}(i_m + l)b_{i_{m+l}}. \tag{5.6}$$

Note that $T_{i_1,\ldots,i_M}$ is the time spent by the customer along the path $Q_{i_2},\ldots,Q_{i_M}$ conditioned on the fact that he entered the path from Q_{i_1}. Similarly, denote by $T_{0,i_2,\ldots,i_M}$ the time on the path (from arrival to Q_{i_2} until departure from Q_{i_M}) of a customer who enters Q_{i_2} from the outside. To derive the mean value of this variable, we first compute the N values $E[T_{1,i_2,\ldots,i_M}], E[T_{2,i_2,\ldots,i_M}], \ldots, E[T_{N,i_2,\ldots,i_M}]$ and the value of $E[T_{i_2,\ldots,i_M}]$. Then $E[T_{0,i_2,\ldots,i_M}]$ can be computed from the formula

$$E[T_{i_2}] + E[T_{i_2,\ldots,i_M}] = \frac{\lambda_{i_2}}{\gamma_{i_2}} E[T_{0,i_2,\ldots,i_M}] + \sum_{m=1}^{N} \frac{\gamma_m P_{m,i_2}}{\gamma_{i_2}} E[T_{m,i_2,\ldots,i_M}], \tag{5.7}$$

which computes the mean time spent by customers who traverse the complete path $Q_{i_2},\ldots,Q_{i_M}$ in two different ways.

Finally, the total time spent in the system of a customer which enters at Q_{i_2}, goes along the path $Q_{i_2},\ldots,Q_{i_M}$ and then leaves the system is given by $E[T_{0,i_2,\ldots,i_M}]$.

Remark 5.1

Note that not all the N path time values $E[T_{k,i_2,\ldots,i_M}]$, $k = 1,\ldots,N$, need to be computed, but only those for which $P_{k,i_2} \neq 0$. In most practical cases, there are only a few such paths and in many of them, no such paths exist (i.e., the entrance

to Q_{i_2} is only from the outside). For example, consider the application discussed in section 8. In this case, we have:

$$E[T_{0,i_2,\ldots,i_M}] = E[T_{i_2}] + E[T_{i_2,\ldots,i_M}]. \tag{5.8}$$

Remark 5.2

Note that the computation of the path times can be done relatively cheaply. For a path of length k, the computation of the mean path time requires $O(Nk)$ operations.

6. A pseudo-conservation law

In Boxma and Groenendijk [3], a "pseudo"-conservation law has been derived for polling systems without routing. This law is a generalization of the laws previously derived by Ferguson and Aminetzah [7], Fuhrmann [9] and Watson [36]. The derivation of this law is based on a work decomposition theorem. Recently, Boxma [2] extended this theorem to much more general single-server systems with non-serving intervals. Here, we adapt theorem 2.1 of Boxma [2] to our network (with slight modifications of notation).

THEOREM 1

Consider a single-server cyclic-service network with gated or exhaustive disciplines. Suppose the network is ergodic. Then the amount of work H in this network at an arbitrary epoch is distributed as the sum of the amount of work $H_{M/G/1}$ in the "corresponding" $M/G/1$ system at an arbitrary epoch and the amount of work $H_{\text{switching}}$ in the network at an arbitrary epoch in a switching interval. In other words, $H \stackrel{D}{=} H_{M/G/1} + H_{\text{switching}}$, where $\stackrel{D}{=}$ stands for equality in distribution. Furthermore, $H_{M/G/1}$ and $H_{\text{switching}}$ are independent.

In this theorem, the "corresponding" $M/G/1$ is an $M/G/1$ system in which the arrival stream is identical to the *total* arrival stream of external customers into the polling system, and the service required by a customer entering at queue i is distributed according to $\tilde{B}_i$. The proof of the theorem appears in Boxma [2]. Note that the theorem also holds for systems with mixed strategies, which are systems where the different queues may have different service policies. The allowed service disciplines in the law derived by Boxma [2] are exhaustive, gated, limited-1, etc.

The expected amount of work in the corresponding $M/G/1$ queue is given by:

$$E[H_{M/G/1}] = \frac{\sum_{i=1}^{N} \lambda_i \tilde{b}_i^{(2)}}{2(1-\rho)}, \tag{6.1}$$

where the quantities $\tilde{b}_i^{(2)}$ are given in (2.3).

The expected value of the total amount of work in the network at an arbitrary epoch is

$$E[H] = \sum_{i=1}^{N} \tilde{b}_i \gamma_i E[W_i] + \sum_{i=1}^{N} \rho_i \left(\frac{b_i^{(2)}}{2b_i} + \tilde{b}_i - b_i \right), \tag{6.2}$$

where the quantities $\tilde{b}_i$ are given in (2.2). The reason for (6.2) is simple: The expected number of customers at an arbitrary epoch waiting to be served is $\gamma_i E[W_i]$ (Little's law). Each of these customers contributes $\tilde{b}_i$ to the mean total work. In addition, at an arbitrary epoch, a customer is served at queue i with probability ρ_i. The mean work he needs is his mean residual service time in queue i ($b_i^{(2)}/2b_i$) and the left-over of his total service time ($\tilde{b}_i - b_i$).

Using the fact that $E[H] = E[H_{M/G/1}] + E[H_{\text{switching}}]$, we obtain:

$$\sum_{i=1}^{N} \tilde{b}_i \gamma_i E[W_i] = \frac{\sum_{i=1}^{N} \lambda_i \tilde{b}_i^{(2)}}{2(1-\rho)} - \sum_{i=1}^{N} \gamma_i \left(\frac{b_i^{(2)}}{2} + (\tilde{b}_i - b_i) b_i \right) + E[H_{\text{switching}}]. \tag{6.3}$$

Remark

An expression similar to (6.3) has been independently derived in a different form in Boxma [2].

The derivation of $E[H_{\text{switching}}]$ follows along the same lines as in Boxma and Groenendijk [3], and the results are:

For the gated discipline:

$$\sum_{i=1}^{N} \tilde{b}_i \gamma_i E[W_i] = \frac{\sum_{i=1}^{N} \lambda_i \tilde{b}_i^{(2)}}{2(1-\rho)} - \sum_{i=1}^{N} \gamma_i \left(\frac{b_i^{(2)}}{2} + (\tilde{b}_i - b_i) b_i \right)$$

$$+ \rho \frac{r^{(2)}}{2r} + \frac{1}{1-\rho} \sum_{i=1}^{N} r_i \sum_{j=1}^{N} \lambda_j \tilde{b}_j \sum_{k=j}^{i} \rho_k + \frac{1}{1-\rho} \sum_{i=1}^{N} \sum_{j=1}^{N} \gamma_i P_{i,j} \tilde{b}_j \sum_{k=i}^{j-1} r_k. \tag{6.4}$$

For the exhaustive discipline:

$$\sum_{i=1}^{N} \tilde{b}_i \gamma_i E[W_i] = \frac{\sum_{i=1}^{N} \lambda_i \tilde{b}_i^{(2)}}{2(1-\rho)} - \sum_{i=1}^{N} \gamma_i \left(\frac{b_i^{(2)}}{2} + (\tilde{b}_i - b_i) b_i \right)$$

$$+ \rho \frac{r^{(2)}}{2r} + \frac{1}{1-\rho} \sum_{i=1}^{N} r_i \sum_{\substack{j=1 \\ j \neq i}}^{N} \lambda_j \tilde{b}_j \sum_{k=j+1}^{i} \rho_k + \frac{1}{1-\rho} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \gamma_i P_{i,j} \tilde{b}_j \sum_{k=i}^{j-1} r_k. \tag{6.5}$$

In the above, the sums $\sum_{k=j}^{i}$ and $\sum_{k=i}^{j-1}$ are cyclic sums.

Note that here, in contrast to the regular polling system (without routing), the value of the "conserved" quantity ($\sum_{i=1}^{N} \tilde{b}_i \gamma_i E[W_i]$) *does depend* on the polling order. Thus, we may conclude that the network performance is quite sensitive to the polling order.

We can also specify the above equations in the symmetric case to obtain the expected waiting time in a symmetric system. By a symmetric system, we mean that all quantities are identical in all queues (and these quantities are denoted with $'$). The probability that a customer completes its service is denoted by P_0 and $P_{i,j} = (1 - P_0)/N$. We obtain:

$$E[W'] = \frac{2b'(1-P_0) + Nb'^{(2)}\lambda' P_0}{2(1-N\rho')} - b'(1-P_0)$$

$$+ P_0^2 \frac{r^{(2)}}{2r} + \frac{r'(N \pm 1)(1 - P_0 + P_0 N\rho')}{2(1-N\rho')},$$

where the + and the − correspond to the gated and the exhaustive disciplines, respectively.

7. Examples

EXAMPLE 1: A GATED-SERVICE TANDEM

The results derived above can be simplified for special structure networks. In particular, let us consider a gated-service tandem network consisting of N queues in which $\lambda_1 = \lambda$, $\lambda_i = 0$ $(i \neq 1)$ and $P_{i,i+1} = 1$ $(i < N)$ and $P_{N,0} = 1$. Note that in this system, the server movement along the network is identical to that of the customers. For convenience of analysis, let us use the following notation: $B = \sum_{i=1}^{N} B_i$, $b = E[B]$, $b^{(2)} = E[B^2]$ and $\rho = \lambda b$. Also, let $R = \sum_{i=1}^{N} R_i$, $r = E[R]$ and $r^{(2)} = E[R^2]$.

We note that in this system, $X_i^j = X_j^j$ for all i and j and thus we drop the index from these variables and have $X = X_i^i$. Using the analysis given in section 3, we obtain:

$$E[X] = \lambda r/(1-\rho);$$

$$E[X^2] = \frac{\lambda^2}{(1-\rho)^2}\left(\frac{\lambda r(b^{(2)} - b^2)}{1-\rho} + r^{(2)} + \frac{2\rho r}{1-\rho}\right) + \frac{\lambda r}{(1-\rho)^3}, \tag{7.1}$$

and

$$E[X_1^*] = E[X^2 - X]/2E[X] + \rho_1 E[X^2 + X]/2E[X], \tag{7.2a}$$

$$E[X_i^*] = \rho_{i-1} E[X^2 - X]/2E[X] + \lambda r_{i-1} + \rho_i E[X^2 + X]/2E[X], \quad 2 \le i \le N. \tag{7.2b}$$

Finally, the mean sojourn time of a customer in the network is given by:

$$E[T] = \sum_{i=1}^{N} E[X_i^*]/\lambda = r - r_N + \frac{1}{\lambda}\left(\frac{\rho E[X^2]}{E[X]} + \frac{(1-\rho_N)E[X^2 - X]}{2E[X]}\right). \tag{7.3}$$

Note that queue N is playing a special role in the expression of the sojourn time. This implies that the arrangement of the queues in the tandem does not affect the mean sojourn time, with the only exception that the selection of the last queue does affect the performance.

EXAMPLE 2: TANDEM NETWORKS: A COMPARISON OF SERVICE DISCIPLINES AND VISIT ORDER

Consider a tandem network consisting of four queues ($N = 4$). Customers enter the tandem at queue 1 ($\lambda_1 > 0$), move to queue 2, queue 3, queue 4 and then leave the network. No external customers enter queues 2, 3 and 4 ($\lambda_2 = \lambda_3 = \lambda_4 = 0$). In the sequel, we compare various cyclic service orders and the gated versus the exhaustive service disciplines for this tandem network.

Consider first the following set of parameters: The mean service time and its second moment are 0.25 and 0.125, respectively, at all queues, and all switch-over times are deterministically of length 1 ($b_i = 0.25$, $b_i^{(2)} = 0.125$, $r_i = 1$, $r_i^{(2)} = 1$, $1 \le i \le 4$). We consider two cyclic orders for the server. In order (a), the server moves from queue 1 to queue 2 to queue 3 to queue 4 and back to queue 1. Order (b) is opposite to (a) (4 to 3 to 2 to 1 and back to 4). In fig. 1, we depict the *total* expected delay

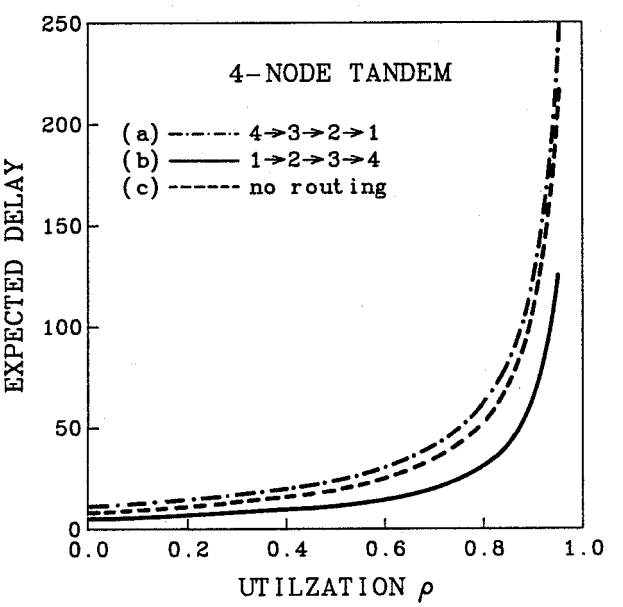

Fig. 1. Gated discipline. Expected delay versus utilization.

(network sojourn time) of the customers in the network as a function of ρ for the two orders when the gated service discipline is employed (the behavior for the exhaustive service discipline is similar). From this figure, we observe that when the server is moving in the same direction as the customers (order (a)), the network sojourn time is much smaller than when they move in opposite directions (order (b)). For very small arrival rate at queue 1, we see that the network sojourn time for order (a) is 6 and for order (b) it is 12. It is interesting to note that the ratio between the expected total delays of the two orders remains almost constant when the arrival rate increases. For completeness, we added in fig. 1 a third curve (c). This curve corresponds to the same network, except that customers are not routed from one queue to the other, but *external* customers arrive at each of the queues at the same rate as they arrived in the tandem. All queues are identical here, and we depict the expected delay of a customer, multiplied by four, to compare it with the total expected delay in the tandem.

We now assume that the server moves in the same direction as the customers and compare between the gated and the exhaustive service disciplines. The expected delays of both disciplines are depicted in curves (i) of fig. 2 for the set of parameters

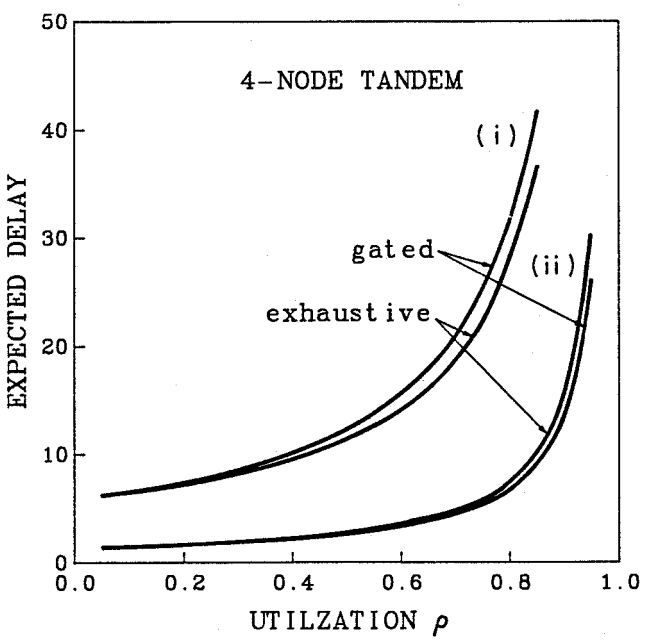

Fig. 2. Gated and exhaustive disciplines.
Expected delay versus utilization.

given above ($b_i = 0.25$, $b_i^{(2)} = 0.125$, $r_i = 1$, $r_i^{(2)} = 1$, $1 \le i \le 4$). For these parameters, the exhaustive discipline yields better performance than the gated discipline. The reason is that the switch-over times are large compared to the service times.

Another set of parameters that we consider is as follows: Mean service time at queue 1 is 0.7, at other queues 0.1. Second moment of service time at queue 1 is 0.49, at other queues 0.01. Switch-over times are deterministically of length 0.1. Curves (ii) of fig. 2 correspond to these parameters, and we observe that now the gated service discipline yields better performance than the exhaustive service discipline. The reason is that switch-over times are short and the mean service time at queue 1 is large compared to the other queues.

8. Modeling aspects

The customer routing feature considerably enhances the modeling and analysis capabilities of polling system models. Below, we demonstrate how the file-server application mentioned in the introduction can be modeled and precisely analyzed by this feature.

Consider a token ring network consisting of a single file-server and N workstations, in which the stations transmit file requests to the file-server, which in turn replies to the different stations by sending back the files they requested. The important performance measure here is the response time of a file-request, which is the time from the request generation by the workstation (being queued at that time at the output queue of the station, awaiting its turn to be transmitted) until the file arrives back at the station. Let the stream of requests generated by the station be Poisson of rate λ_i and the transmission time of such a request be a random variable B_i, $i = 1, \ldots, N$ (typically, B_i and B_j are identically distributed for all i and j). Let the transmission duration of a file be a random variable B (obviously, it is common practice that the mean value of B is orders of magnitude larger than the mean value of B_i). When a request from station i arrives at the file-server node, the appropriate file is retrieved and being placed in the output queue of the file-server for future transmission. Once the file-server node has its turn to transmit over the network (namely, when receiving the token), it will transmit all files queued in its output queue to their destination. Since the propagation (of the file transmission) is negligible in comparison to all queueing times and transmission times, one may assume that the time at which the file arrives to its destination is identical to the time at which the file-server completed its transmission.

We model this system by a polling model with customer routing. We have $N + 1$ queues: the first N represent the stations and the last one (indexed $N + 1$) represents the file-server. Customers arrive at Q_i ($1 \le i \le N$) at rate λ_i and their service time at that queue is B_i. Upon service completion, the customer is routed (with probability 1) to Q_{N+1}, in which it waits for service. The service time at Q_{N+1} (in which there are no external arrivals but only transitions) is a random variable B. Once served at Q_{N+1}, the customer leaves the system. Note that in this modeling we do not have to "route" the customer representing the file back to queue i (but rather let it "leave" the system); the reason is that our modeling concerns only the times spent by the different transactions in the system and thus, once the file is

transmitted, it is considered to be accepted and its time in the system is accounted for appropriately.

Using this model, one can calculate *precisely* the mean response time of a request generated at station i. This will be given by $E[T_{0,i,N}]$, which is the time in the system of a customer who enters Q_i, moves to Q_N and leaves the system.

Remark 8.1

For simplicity, we described in the model only file requests and their responses. However, note that one can add any other traffic pattern to the model (representing messages sent by the stations to each other); the analysis presented is obviously still valid for deriving the response time of file requests.

Remark 8.2

Similarly, the model can represent several file-servers (or other service centers) in the network. For simplicity, we may then mark the service centers by the indices $N+1, \ldots, N+K$. In this case, we will see that a customer from station i $(1 \le i \le N)$ will be routed to service center $N+k$ $(k = 1, \ldots, K)$ with probability $P_{i,N+k}$ and the response time of such a request will be $E[T_{0,i,N+k}]$. Since the model suggested here provides exact response times, it can be used to assess the quality of several design alternatives (in terms of the location of the service centers and the allocation of tasks among the centers) in order to optimize the system design.

9. Discussion

In this paper, we analyzed the performance of a queueing network that is served by a single cyclic server according to the gated or the exhaustive service disciplines. The analysis approach we use applies (after simple modifications) to most polling systems which have previously been analyzed under the assumption that customers always leave the system. These include networks with mixed service methods (i.e., in which the service is exhaustive for some queues and gated for others), networks with periodic polling order (service according to a polling table, Baker and Rubin [1]), networks with random polling (Kleinrock and Levy [14]), binomial-gated service (Levy [19]), networks with correlated arrivals (Levy and Sidi [21]), networks with zero switch-over periods, discrete time systems, etc. In contrast, it is an open question whether the limited-1 system (Nomura and Tsukamoto [23], Takagi [32], Fuhrmann [9] and Takagi [33]) can be analyzed via the conservation law; the reason is that the derivation of the conservation law for this system (without customer routing) used the assumption that all arrivals are Poisson, an assumption that cannot be used here.

An interesting question is whether the station-time analysis approach of regular polling systems (Humblet [11], Ferguson and Aminetzah [7], Baker and Rubin [1]

and Sarkar and Zangwill [25]) can be applied to this network. As of now, the quesion is open, but it seems that it is *inherently* difficult to apply that approach to our system. An alternative model in which *work* (rather than individual *customers*) is stochastically routed has been analyzed by Sarkar and Zangwill [26] via the station-time approach. That model differs from ours and they coincide only in the limit (when customer service time approaches zero while the utilization is kept constant).

10. Summary of notation

The following is a list of the notations frequently used in this paper:

$B_i, B_i^*(s)$: service time of a customer at queue i and its LST;

$b_i, b_i^{(2)}$: mean and second moment of B_i;

$R_i, R_i^*(s)$: duration of switch-over period from queue i and its LST;

$r_i, r_i^{(2)}$: mean and second moment of R_i;

$r, r^{(2)}$: mean and second moment of $\sum_{i=1}^{N} R_i$;

λ_i : external arrival rate of type-i customers;

γ_i : total arrival rate into queue i;

ρ_i : utilization of queue i; $\rho_i = \gamma_i b_i$;

ρ : system utilization $\sum_{i=1}^{N} \rho_i$;

$c = E[C]$: the expected cycle length;

X_i^j : number of customers in queue j when queue i is polled;

$X_i = X_i^i$: number of customers residing in queue i when it is polled;

$f_i(j)$: $E(X_i^j)$;

$f_i(j, k)$: $\begin{cases} E(X_i^j X_i^k), & j \neq k, \\ E[X_i^j)^2] - E(X_i^j), & j = k; \end{cases}$

X_i^* : number of customers residing in queue i at an arbitrary moment;

Y_i : the duration of a service period of queue i;

W_i, T_i : the waiting time and sojourn time of an arbitary customer at queue i;

$F^*(z_1, z_2, \ldots, z_N)$: GF of number of customers present at the queues at arbitrary moments;

$F_i(z_1, z_2, \ldots, z_N)$: GF of number of customers present at the queues at polling instants of queue i;

$\overline{F_i}(z_1, z_2, \ldots, z_N)$: GF of number of customers present at the queues at switching instants of queue i;

$V_i(z_1, z_2, \ldots, z_N)$: GF of number of customers present at the queues at service initiation instants at queue i:

$\overline{V}_i(z_1, z_2, \ldots, z_N)$: GF of number of customers present at the queues at service completion instants at queue i.

Acknowledgements

We would like to thank the anonymous referees for their helpful suggestions, and R. Ram for his remarks.

References

[1] J.E. Baker and I. Rubin, Polling with a general-service order table, IEEE Trans. Commun. COM-35(1987)283–288.
[2] O.J. Boxma, Workloads and waiting times in single-server systems with multiple customer classes, Queueing Systems 5(1989)185–214.
[3] O.J. Boxma and W.P. Groenendijk, Pseudo-conservation laws in cyclic queues, J. Appl. Prob. 24(1987)949–964.
[4] R.B. Cooper, Queues served in cyclic order: Waiting times, Bell. Syst. Tech. J. 49(1970)399–413.
[5] R.B. Cooper and G. Murray, Queues served in cyclic order, Bell. Syst. Tech. J. 48(1969)675–689.
[6] M. Eisenberg, Queues with periodic service and changeover time, Oper. Res. 20(1972)440–451.
[7] M.J. Ferguson and Y.J. Aminetzah, Exact results for nonsymmetric token ring systems, IEEE Trans. Commun. COM-33(1985)223–231-
[8] S.W. Fuhrmann, Performance analysis of a class of cyclic schedules, Bell Laboratories Technical Memorandum No. 81-59531-1 (1981).
[9] S.W. Fuhrmann, Symmetric queues served in cyclic order, Oper. Res. Lett. 4(1985)139–144.
[10] O. Hashida, Analysis of multiqueues, Rev. Electr. Commun. Lab. 20(1972)189–199.
[11] P. Humblet, Source coding for communication concentrators, Electron. Syst. Lab., MIT, Cambridge, ESL-R-798 (1978).
[12] T. Katayama, A cyclic service tandem queueing model, Report NTT Communication Switching Laboratories, Tokyo; also to appear in Queueing Systems and their Applications.
[13] L. Kleinrock, *Queueing Systems, Vol. 1: Theory* (Wiley Interscience, 1975).
[14] L. Kleinrock and H. Levy, The analysis of random polling systems, Oper. Res. 36(1988)716–732.
[15] G.P. Klimov, Time-sharing service systems, Theory Prob. Appl. 19(1974)532–551.
[16] A.G. Konheim and B. Meister, Waiting times and lines in systems with polling, J. ACM 21(1974) 470–490.
[17] H. Kushner, *Introduction to Stochastic Control* (Holt, Rinehard and Winston, 1971).
[18] H. Levy, Delay computation and dynamic behavior of non-symmetric polling systems, Perf. Eval. 10(1989)35–51.
[19] H. Levy, Binomial gated service: A method for effective operation and optimization of polling systems, IEEE Trans. Commun. COM-39(1991)1341–1350.
[20] H. Levy and M. Sidi, Polling systems: Applications, modeling and optimization, IEEE Trans. Commun. COM-38(1990)1750–1760.
[21] H. Levy and M. Sidi, Polling systems with simultaneous arrivals, IEEE Trans. Commun. COM-39(1991)823–827.
[22] S.S. Nair, A single server tandem queue, J. Appl. Prob. 8(1971)95–109.
[23] N. Nomura and K. Tsukamoto, Traffic analysis on polling systems, Trans. Inst. Electr. Commun. Eng. Japan J61-B(7) (1978) 600–607, in Japanese.

[24] I. Rubin and L.F. De Moraes, Message delay analysis for polling and token multiple-access schemes for local communication networks, IEEE J. Sel. Areas Comm. SAC-1(1983).
[25] D. Sarkar and W.I. Zangwill, Expected waiting time for nonsymmetric cyclic queueing systems – exact results and applications, Manag. Sci. 35(1989)1463–1474.
[26] D. Sarkar and W.I. Zangwill, Cyclic queues with interdependent work, Technical Report (April, 1989).
[27] M. Sidi and H. Levy, A queueing network with a single cyclically roving server, Technical Report, Department of Computer Science, Tel-Aviv University, Tel-Aviv, Israel (November 1988).
[28] M. Sidi and H. Levy, Customer routing in polling systems, *Performance'90*, Edinburgh (September 1990), pp. 319–323.
[29] M. Sidi and A. Segall, Structured priority queueing systems with applications to packet-radio networks, Per. Eval. 3(1983)264–275.
[30] B. Simon, Priority queues with feedback, J. ACM 31(1984)134–149.
[31] G.B. Swartz, Polling in a loop system, J. ACM 27(1980)42–59.
[32] H. Takagi, Mean message waiting time in a symmetric polling system, *Performance'84*, ed. E. Gelenbe (North-Holland, Amsterdam, 1985), pp. 293–302.
[33] H. Takagi, Mean message waiting times in symmetric multi-queue systems with cyclic service, Perf. Eval. 5(1985)271–277.
[34] H. Takagi, *Analysis of Polling Systems* (MIT Press, 1986).
[35] M. Taube-Netto, Two queues in tandem attended by a single server, Oper. Res. 25(1977)140–147.
[36] K.S. Watson, Performance evaluation of cyclic service straegies – a survey, in: *Performance'84*, ed. E. Gelenbe (North-Holland, Amsterdam, 1985), pp. 521–533.

Corrigendum

Correction to equation (5.6) in the paper: A queueing network with a single cyclically roving server

Moshe Sidi

Electrical Engineering Department, Technion–Israel Institute of Technology, Haifa 32000, Israel

Hanoch Levy

Computer Science Department, Tel-Aviv University, Tel Aviv 69978, Israel

Steve W. Fuhrmann

1 Independence Court, Morristown, NJ 07960, USA

Received 1 October 1993

R. Ram and A. Tripathi have pointed out that a term is missing in the expression given in [1] (eq (5.6)) for computing the expected path delay in a gated service system. The missing term is the contribution of the server switch-over times to the path delays. Equation (5.6) should therefore read:

$$E[T_{i_1,i_2,\ldots,i_M}] = [g_{i_1}(i_1) - \bar{g}_{i_1}(i_1) - 1]b_{i_1} + \sum_{m=2}^{M-1} g_{i_m}(i_m)b_{i_m} + [\bar{g}_{i_M} + 1]b_{i_M}$$
$$+ \sum_{m=1}^{M-1}\sum_{n=1}^{k_m} g_{i_m+n}(i_m+n)b_{i_m+n} + \sum_{m=1}^{M-1}\left[r_{i_m} + \sum_{n=1}^{k_m} r_{i_m+n}\right]. \quad (5.6)$$

We would like to thank R. Ram and A. Tripathi for pointing out the missing term.

Reference

[1] M. Sidi, H. Levy and S.W. Fuhrmann, A queueing network with a single cyclically roving server, Queueing Syst. 11 (1992) 121–144.