Polling Systems with Zero Switch-Over Periods: A General Method for Analyzing the Expected Delay.

Hanoch Levy
AT&T Bell Laboratories
Crawfords Corner Road
Holmdel, NJ 07733

Leonard Kleinrock
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90024

June 26, 1987

ABSTRACT

Although the first polling systems to be analyzed were systems with zero switch-over periods (ZSOP), later research concentrated on studying polling systems with non-zero switch-over periods (NSOP). As a result, many variations of polling schemes (in particular, discrete-time variations) which have been analyzed for NSOP systems have not been analyzed for the corresponding ZSOP systems. In this note we propose a general approach for using the results that have been derived for NSOP systems and applying them to solve the corresponding ZSOP systems.

1. Introduction

We consider the problem of calculating the expected delay in polling systems with zero length switch-over periods. The models under consideration consist of a single server which serves N ***infinite independent*** queues. In these systems after completing the service of queue i, the server ***switches*** and starts serving queue j. The instant at which the server starts serving a queue is called a ***polling instant*** and the period during which the server switches from queue i to queue j is called a ***switch-over period*** (or ***reply interval*** or ***walking time***).

Although the first polling systems to be analyzed were systems with ***zero length switch-over periods (ZSOP),*** i.e., systems where all the switch-over periods are deterministically of zero length, later studies concentrated on studying polling systems with ***nonzero switch-over periods (NSOP),*** i.e., systems where not all of the switch over periods are of zero length. As a result, while the solutions for a large variety of NSOP polling models are available, this is not the case with ZSOP polling systems models -- solutions are available only for the continuous-time cyclic-polling system with exhaustive or gated service discipline (Cooper and Murray [1969], Cooper [1970], Humblet 1978]).

This situation raised the question of whether the solutions developed for NSOP systems can be extended and applied to ZSOP systems. This question was ***briefly*** addressed by several previous studies (Eisenberg [1972], Takagi [1986]) where it was stated that the method used to analyze the NSOP systems is not applicable for solving the ZSOP systems. However, since neither of those references comprehensively addressed the issue, the answer to this question remained unclear.

In this paper we address this question, concentrating on analyzing available methods which derive the expected delay in NSOP systems by solving a linear set of equations. While we agree with the previous statements that those methods cannot be applied ***directly*** to solve ZSOP systems, we show that by considering the limiting behavior of NSOP systems, and by applying a simple modification to the available equation sets one can easily create new equation sets from which the expected delay in ZSOP systems can be derived. As a result, we claim that for all practical purposes, ZSOP systems may be solved using the solutions for the appropriate NSOP systems, and additional analysis is not required.

The structure of this note is as follows: In Section 2 we review the details of the model and the relevant literature. In section 3 we concentrate on the discrete-time cyclic-polling exhaustive system and show how the expected delay in the ZSOP version of this model may be calculated using the method developed for solving the corresponding NSOP system. In Section 4 we list the models for which the treatment of Section 3 may be applied. In Section 5 we present numerical results and discuss the limitation of this method.

2. Description of the Models and Previous Work

The common denominator for all the models considered below is that a single server serves N infinite queues. The order in which the server serves these queues, i.e., the rules by which the next queue to be served (after the service of queue i has been completed) is determined, is called the ***polling order***. The rule by which the server decides to stop servicing a given queue is called the ***service policy***. The polling orders considered here are: 1) ***Cyclic*** - in which the next queue served after queue i is $i+1$ (modulo N), and 2) ***Memory-less random*** - in which the next queue to be served is queue j with probability p_j. The service policies considered are: 1) ***Exhaustive*** - in which the server stops serving queue i only when no more customers are left in the queue, 2) ***Gated*** - in which the server stops serving when he completeds the service of all customers found in queue i at the polling instant, and 3) ***Limited*** - in which the server serves at most one customer at a time and then switches to the next queue. Two time models are considered: ***Continuous*** time and ***discrete*** time. In the discrete-time model time is slotted, all times (service times and lengths of switch-over periods) are expressed in terms of the slot unit, and events (e.g., arrivals) occur at the slot boundary. Arrivals to each of the queues in the continuous-time system are assumed to be Poisson.

Previous Work

The main approach for analyzing polling systems has been to calculate the first two moments of the *buffer occupancy* (number of customers present at the system) at polling instants. Cooper and Murray [1969] and Cooper [1970] were the first to use it in analyzing the continuous-time cyclic-polling exhaustive and gated systems with zero reply intervals. A similar approach was later used by Eisenberg [1972] and Hashida [1972] (continuous-time cyclic-polling exhaustive and gated NSOP systems), Konheim and Meister [1974] and Swartz [1980] (discrete-time cyclic-polling exhaustive NSOP systems), Rubin and DeMoraes [1983] (discrete-time cyclic-polling NSOP systems), Nomura and Tsukamoto [1978] (continuous-time cyclic-polling limited-service NSOP system), Takagi [1985] (discrete-time cyclic-polling limited-service NSOP system), and Kleinrock and Levy [1985] (discrete-time random-polling exhaustive, gated and limited-service NSOP systems). Recently, a different approach for analyzing the system by looking at the *cycle times* was presented by Humblet [1978] and Ferguson and Aminetzah [1985]. This analysis was derived for the continuous-time cyclic-polling exhaustive and gated NSOP systems (Humblet's [1978] analysis can be used for ZSOP systems as well). Very recently Baker and Rubin [1987] used this approach to analyze systems with general-periodic polling. Lastly, a recent tutorial presenting the analysis of polling systems in an organized framework was written by Takagi [1986]. The main analysis approach taken in Takagi [1986] is the approach of calculating the moments of the buffer occupancy at polling instants.

3. The Discrete-Time Cyclic-Polling Exhaustive Service Policy

3.1 Model Description and Review of the Solution for NSOP Systems

In this section we consider a discrete-time model in which time is slotted with the slot size equal to the (fixed) service time of a customer, and is measured in slots; the time interval $[t, t+1]$ is called the tth slot. The polling order is cyclic and the service policy is exhaustive (i.e. - serve queue i until emptying it).

The arrival process at each queue is assumed to be independent of those at other queues and we denote by $X_i(t)$ the number of customers arriving to queue i during slot t; the arrivals are assumed to occur at the end of the slot. The sequence $\{X_i(t);\ t=0,\ 1,\ 2...\}$ is assumed to be an independent and identically distributed sequence of random variables whose mean and variance are: $\mu_i = E[X_i(t)]$, $\sigma_i^2 = Var[X_i(t)]$. The lengths of the switch-over periods from queue i to queue $i+1$ are assumed to be a sequence of independent and identically distributed random variables (denoted by S_i) with mean and variance r_i and δ_i^2 respectively.

We define the ***customer waiting time***[1], denoted by W, as the time spent by the customer in the queue, and the ***customer delay*** (or, ***sojourn time***), denoted by T, as the sum of the waiting time and the time spent in service (one unit in the discrete-time model and general in the continuous-time model). The solution for the expected customer delay in NSOP systems was derived by Konheim and Meister [1974] and Swartz [1980]. Below we follow that approach as presented by Takagi [1986]. The key for that method is to compute the first moments and the variance-covariance matrix of the number of customers present in the system at polling instants. More precisely, we let $L_i(t)$ be the number of customers present at queue i at time t, and $\tau_i(m)$ be the time at which the server becomes available to serve queue i (for the mth time). Then we denote:

$$F_i(z_1,\ \ldots,\ z_N) \triangleq E\left(\prod_{j=1}^{N} z_j^{L_j(\tau_i(m))} \right)$$

$$f_i(j) \triangleq \left. \frac{\partial F_i(z_1,\ \ldots,\ z_N)}{\partial z_j} \right|_{z_1=\cdots=z_N=1} \quad ; \quad f_i(j,k) \triangleq \left. \frac{\partial^2 F_i(z_1,\ \ldots,\ z_N)}{\partial z_j \partial z_k} \right|_{z_1=\cdots=z_N=1}$$

The expected delay observed in queue i can be calculated (see, e.g., Takagi [1986]) from $f_i(i)$ and $f_i(i,i)$:

1. Since arrivals are assumed to occur at the end of the slots, the customer waiting time (or delay) does not include the slot at which he arrived.

$$E(T_i) = \frac{f_i(i,i) + f_i(i)}{2\mu_i\, f_i(i)} + \frac{\sigma_i^2}{2\mu_i}\left(\frac{1}{1-\mu_i} - \frac{1}{\mu_i}\right) \tag{3.1}$$

The expressions for $f_i(j)$ are available in closed form:

$$f_i(i) = \frac{\mu_i(1-\mu_i)\sum_{k=1}^{N} r_k}{1-\sum_{k=1}^{N}\mu_k} \tag{3.2a}$$

$$f_i(j) = \mu_j\left(\sum_{k=j}^{i-1} r_k + \frac{\sum_{k=j+1}^{i-1}\mu_k \sum_{k=1}^{N} r_k}{1-\sum_{k=1}^{N}\mu_k}\right) \qquad j \neq i \tag{3.2b}$$

where all indices are modulo N.

Closed form expressions for $f_i(j,k)$ are not available, but these values may be computed by solving the following set of N^3 linear equations:

$$f_{i+1}(j,k) = \mu_j\mu_k(\delta_i^2 + r_i^2) + r_i\,\mu_k\, f_i(j) + r_i\,\mu_j\, f_i(k) + f_i(i)\,\mu_j\,\mu_k\left(\frac{2r_i}{1-\mu_i} + \frac{1}{(1-\mu_i)^2} + \frac{\sigma_i^2}{(1-\mu_i)^3}\right)$$
$$+ \frac{f_i(i,j)\,\mu_k + f_i(i,k)\,\mu_j}{1-\mu_i} + f_i(j,k) + \frac{f_i(i,i)\mu_j\,\mu_k}{(1-\mu_i)^2} \qquad i\neq j,\ i\neq k,\ j\neq k \tag{3.3a}$$

$$f_{i+1}(j,j) = \mu_j^2(\delta_i^2 + r_i^2) + r_i\,(\sigma_j^2 - \mu_j) + 2r_i\,\mu_j\, f_i(j) + f_i(j,j) + \frac{2f_i(i,j)\mu_j}{1-\mu_i} + \frac{f_i(i,i)\mu_j^2}{(1-\mu_i)^2}$$
$$+ f_i(i)\left\{\frac{\sigma_j^2-\mu_j}{1-\mu_i} + \mu_j^2\left(\frac{2r_i}{1-\mu_i} + \frac{1}{(1-\mu_i)^2} + \frac{\sigma_i^2}{(1-\mu_i)^3}\right)\right\} \qquad i\neq j \tag{3.3b}$$

$$f_{i+1}(i,k) = \mu_i\,\mu_k(\delta_i^2 + r_i^2) + r_i\,\mu_i\left(f_i(k) + \frac{f_i(i)\,\mu_k}{1-\mu_i}\right) \qquad i\neq k \tag{3.3c}$$

$$f_{i+1}(i,i) = \mu_i^2(\delta_i^2 + r_i^2) + r_i(\sigma_i^2 - \mu_i) \tag{3.3d}$$

where all indices are modulo N.

The computation of the expected delay, therefore, boils down to the solution of this equation set which can be carried out very efficiently using a simple iterative procedure (Levy [1986]). When all queues are identical we drop the subscript i from all parameters and the expected delay can be expressed explicitly:

$$E[T] = E[T_i] = \frac{\delta^2}{2r} + \frac{\sigma^2}{2\mu(1-N\mu)} + \frac{Nr(1-\mu)}{2(1-N\mu)} . \tag{3.4}$$

3.2 Server Behavior for System with Zero Switch-Over Periods

In the discrete-time system with zero switch-over periods, when the server completes serving queue i he is ***instantaneously*** ready to serve queue $i+1$. The server behavior, therefore, needs further specification in this case; the specific issue must be addressed is what service order does the server take at time $t+1$, if the system is ***empty*** at time t. In other words, the question is, which queue is first polled at time $t+1$ when the system is completely empty at time t.

For the cyclic system we assume that the first queue to be polled at time $t+1$ is randomly selected among the N queues with equal probability $(1/N)$ for each queue. Once the first queue to be polled at time $t+1$ is selected, the server then proceeds in cyclic order until it finds a queue which is non-empty at that time $(t+1)$. If no such queue is found the process (i.e., the random selection) repeats for time $t+2$ and so on.

For the memoryless random system we assume that the selection of the queue to be polled at time $t+1$ is done as in the case of non-empty system. This means that at time $t+1$ queue i is polled with probability p_i.

Obviously, one could propose alternative definitions for the polling orders at these instants. Nevertheless, the orders defined above seem to be natural and to follow the general philosophy of the two polling systems considered.

Note that the problem of which queue to poll next when the system becomes empty and the switch-over period is of zero length, ***does not*** arise in the context of the ***continuous-time*** system. The reason is that in this system, due to the continuous-time model and the assumption of Poisson arrivals, no two arrivals can occur at exactly the same time. Therefore, the server's choice in that system, when it becomes empty, is simple: select the first queue to become non-empty. Obviously, no tie-breaking rules are required in that case.

3.3 NSOP Equations May Not be Used Directly to Solve for ZSOP Systems

In a ZSOP system the switch-over periods are deterministically of zero length; Thus, we have

$$r_i = 0 \; ; \; \delta_i^2 = 0 \; ; \; i=1....N$$

It is straightforward to see that under these conditions equations (3.1), (3.2a-b) and (3.3a-d) may not be used directly to calculate $E[T_i]$. The reason is that from (3.2a-b) we get $f_i(j) = 0$ ($i,j=1,...N$) and, then, from (3.3a-d) we get $f_i(j,k) = 0$ ($i,j,k=1...N$). Therefore, (3.1) may not be used.

This problem has a simple physical interpretation. The variables $f_i(j)$ and $f_i(j,k)$ represent the moments of the number of customers present in the system at arbitrary instants of polling queue i. In a system where all switch-over periods are of zero length, then, when the system becomes empty, the server polls the queues infinitely many times at a single instant. This causes all these moments and the expected cycle time (which is a key for deriving (3.1)) to be equal to zero.

This property was mentioned by Takagi [1986] as the reason for the inapplicability of the above method for deriving the expected delay in the ZSOP system.

3.4 Using the Limit of the Expected Delay in the NSOP System

While the method described above is not suitable for analyzing ZSOP systems it is theoretically suitable for analyzing any NSOP system. In particular, it may be used to analyze NSOP systems where the length of the switch-over period is non-zero with probability p ($p>0$) and zero with probability $1-p$. The main idea presented in this subsection, is to consider a system where the switch-over periods are ***almost always*** of zero length. We call this a system with Almost-Zero Switch-Over Periods, or an AZSOP system. The use of an AZSOP system allows one to use the equations described above. A proper selection of the AZSOP system will lead to calculated delays that closely approximate the delay in the ZSOP system.

3.5 Proper Selection of the Almost-Zero Switch-Over Period

When constructing the AZSOP system the goal is to create a system whose behavior will approach the behavior of the ZSOP system. The arrival process parameters are therefore selected to be as in the ZSOP system and we concentrate on selecting the switch-over period parameters, by observing the relevant characteristics of the ZSOP system:

1) When the server finishes servicing queue i at time t, it is ready to serve any of the non-empty queues right at that time (t). If several queues are not empty, the queue to be served at t is selected according to the cyclic order.

2) When the system is empty at time t the server is ready to serve at time $t+1$ any of the non-empty queues. If several queues are non-empty at time $t+1$, the queue to be served next is determined as follows: First, the server randomly selects (with equal probability for each of the N queues) the first queue to be *polled* at time $t+1$. Then the server proceeds polling all queues, according to the cyclic order, until the first non-empty queue is found; This queue is selected to be served at time $t+1$.

To assure that the AZSOP system follows these two properties with high probability we choose the distribution of the switch-over period as follows:

$$P_r[S_i = 1] = p,$$
$$P_r[S_i = 0] = 1-p \triangleq q, \qquad (3.5)$$

and we let $p \rightarrow 0$.

The selection of a large mass at the origin provides for the first property, namely that the server will continuously serve the system as long as there are customers in the system. Concentrating the rest of the mass at 1 provides for the second property, namely, that when the system is empty at time t, the server becomes available to serve again at time $t+1$.

The equal probability selection stated in the second property is provided for as well. To verify this, assume that queue N is the one served until time t, when the system becomes completely empty. Then, the probability that queue i is polled first at time $t+1$ (denoted by q_i) is given by

$$q_i = q^{i-1}p + q^{i-1}q^N p + q^{i-1}(q^N)^2 + \cdots = q^{i-1}p\sum_{i=0}^{\infty}(q^N)^i = \frac{q^{i-1}p}{1-q^N}$$

It is easy to see that $\lim_{p \to 0} q_i = 1/N$. Thus, the probability of queue i being polled first at time $t+1$ approaches $1/N$, independently of the queue last being served at time t.

Since the selection of the switch-over period provides that the behavior of the AZSOP system will approach that of the ZSOP when $p \to 0$, it follows that the delay incurred in the AZSOP system approaches the delay in the ZSOP system.

The selection of our switch-over period parameters yields the following values for r_i and δ_i^2:

$$r_i = p \ ; \ \delta_i^2 = p(1-p) \tag{3.6}$$

which can be substituted into (3.2a-b) and (3.3a-d) to yield a solution for the expected delay in the AZSOP system. It is interesting to consider the symmetric system by substituting (3.6) into (3.4). The expected delay in the symmetric ZSOP system is derived by letting $p \to 0$ which yields:

$$E[T_i] = \frac{1}{2} + \frac{\sigma^2}{2\mu(1-N\mu)} \tag{3.7}$$

Note that when the arrival rate approaches zero the expected delay approaches $\frac{1}{2}+\frac{\sigma^2}{2\mu}$. For example, if the arrival distribution is Bernoulli, then $\lim_{\mu \to 0} E[T_i] = 1$, which as expected, equals one unit (the service time).

It is important to emphasize again that the proper selection of the switch-over period distribution is critical for correct results and it is not sufficient to select any arbitrary distribution whose moments approach zero. This issue may be best clarified by considering an alternative distribution:

$$Pr[S_i = k] = p$$
$$Pr[S_i = 0] = 1-p \tag{3.8}$$

Whose mean and variance are:

$$r_i = kp \; ; \; \delta_i^2 = k^2 p(1-p) \tag{3.9}$$

While the limits of r_i and δ_i^2 are identical to those of (3.6), the limit of their ratio is different: $\lim_{p \to o} \frac{\delta_i^2}{r} = k$. Thus, in the symmetric case the limiting expected delay of this system is:

$$E[T_i] = \frac{k}{2} + \frac{\sigma^2}{2\mu(1-N\mu)} \tag{3.10}$$

This is different from (3.7). This difference may be explained by examining the system behavior under this distribution and by observing that after the system becomes empty at time t the server is next ready to serve only at time $t+k$. This behavior actually reflects a system where the server takes vacations of length k whenever it goes idle (see Doshi [1986] for a survey of queueing systems with vacations).

3.6 Equation Modification to Avoid Numerical Problems

The method derived in Sections 3.4 and 3.5 is very stable (see Section 5 below). However, in extreme cases it may yield numerical problems which can be avoided using the method we outline below. The main approach (which is somewhat similar to that of Humblet [1978]) is to embed the limiting behavior into equations (3.2a-b) and (3.3a-d) directly. We do so by assuming that all switch-over periods are identical and $r_i = r$, $\delta_i^2 = \delta^2$, and by defining

$$g_i(j) \triangleq \lim_{p \to o} \frac{f_i(j)}{r} \quad ; \quad g_i(j,k) \triangleq \lim_{p \to o} \frac{f_i(j,k)}{r} \quad ; \quad i,j,k = 1,\ldots,N \tag{3.11}$$

We now substitute (3.11) into (3.2a-b) and (3.3a-d) and evaluate the limit of the equation when $p \to 0$ this yields:

$$g_i(i) = \frac{N\mu_i(1-\mu_i)}{1-\sum_{k=1}^{N}\mu_k} \tag{3.12a}$$

$$g_i(j) = \mu_j \left(i-j+\frac{N\sum_{k=j+1}^{i-1}\mu_k}{1-\sum_{k=1}^{N}\mu_k} \right) \qquad j \neq i \tag{3.12b}$$

$$g_{i+1}(j,k) = \mu_j\mu_k + g_i(i)\,\mu_j\,\mu_k \left(\frac{1}{(1-\mu_i)^2} + \frac{\sigma_i^2}{(1-\mu_i)^3} \right)$$
$$+\ \frac{g_i(i,j)\,\mu_k + g_i(i,k)\,\mu_j}{1-\mu_i} + g_i(j,k) + \frac{g_i(i,i)\mu_j\,\mu_k}{(1-\mu_i)^2} \qquad i\neq j,\ i\neq k,\ j\neq k \tag{3.13a}$$

$$g_{i+1}(j,j) = \mu_j^2 + (\sigma_j^2-\mu_j) + g_i(j,j) + \frac{2g_i(i,j)\mu_j}{1-\mu_i} + \frac{g_i(i,i)\mu_j^2}{(1-\mu_i)^2}$$
$$+\ g_i(i)\left\{ \frac{\sigma_j^2-\mu_j}{1-\mu_i} + \mu_j^2\left(\frac{1}{(1-\mu_i)^2} + \frac{\sigma_i^2}{(1-\mu_i)^3} \right) \right\} \qquad i\neq j \tag{3.13b}$$

$$g_{i+1}(i,k) = \mu_i\,\mu_k \qquad i\neq k \tag{3.13c}$$

$$g_{i+1}(i,i) = \mu_i^2 + \sigma_i^2 - \mu_i \tag{3.13d}$$

Once this linear equation set is solved we may derive the expected delay at the ZSOP system:

$$E[T_i] = \frac{g_i(i,i) + g_i(i)}{2\mu_i\, g_i(i)} + \frac{\delta_i^2}{2\mu_i}\left(\frac{1}{1-\mu_i} - \frac{1}{\mu_i} \right) \tag{3.14}$$

It is easy to see that the values of $g_i(i)$ and $g_i(i,i)$ are not identically zero (in contrast to the values of $f_i(i)$ and $f_i(i,i)$). Thus, the numerical problems, which may be encountered in using the approach described in Sections 3.4 and 3.5, are avoided here.

4. The Application of the Method to Other ZSOP Systems

4.1 Discrete-Time Cyclic-Polling Systems

The analysis of the discrete-time cyclic-polling *gated* NSOP system is similar to that of the exhaustive system: For the non-symmetric system a set of N^3 linear equations must be solved (see, e.g., Takagi [1986] equations (5.4a-b) and (5.5a-d)). For the symmetric system the expected delay is given by:

$$E[T_i] = \frac{\delta^2}{2r} + \frac{\sigma^2}{2\mu(1-N\mu)} + \frac{Nr(1+\mu)}{2(1-N\mu)} . \tag{4.1}$$

The treatment of the corresponding ZSOP system is, therefore, similar to the one described in Section 3. This implies that the solution may be achieved by either substituting $r_i = p$, $\delta_i^2 = p(1-p)$ (where p is very small) into (5.4a-b) and (5.5a-d) of Takagi [1986], and solving them or, by transforming the equations as done in Section 3.6. The expected delay in the symmetric system is obtained by considering the limits of equation (4.1). As expected, the result is identical to that of the exhaustive system (equation (3.7)).

The NSOP system with *limited service* has been analyzed only for the fully symmetric case and its expected delay is given by:

$$E[T_i] = \frac{\delta^2}{2r} + \frac{(1+Nr)\sigma^2}{2\mu(1-N\mu-Nr\mu)} + \frac{N\delta^2\mu}{2(1-N\mu-Nr\mu)} \tag{4.2}$$

The expected delay in the ZSOP system can be obtained by taking limits on this expression; Again, the result is identical to (3.7).

4.2 Discrete-Time Random-Polling Systems

The expected delay in the NSOP *random polling* systems has been derived in Kleinrock and Levy [1985]. The analysis approach is similar to that of the cyclic systems, i.e., the non-symmetric cases of the exhaustive and gated systems require the solution of N^2 linear equations (see in Kleinrock and Levy [1985], equations (17), (18), (19a-b) and (41) for the exhaustive system and equations (49), (51a-d) and (58) for the gated system), and the solution of symmetric cases for all three systems (exhaustive, gated and limited) are given in closed form. The latter are given in Kleinrock and Levy [1985] and can be obtained by adding $\frac{(N-1)r}{2(1-N\mu)}$ to (3.7) and to

(4.1) and $\frac{(N-1)r}{2(1-N\mu-Nr\mu)}$ to (4.2) to yield the expected delay in the exhaustive, gated and limited systems, respectively. The approach to handling the corresponding ZSOP systems is, therefore, identical to that of the cyclic systems. Note, that the expected delay in the symmetric ZSOP systems (take the limits on the additive terms, given above) are all identical to those of the corresponding cyclic systems.

4.3 Continuous-Time Systems

The expected delay in continuous-time cyclic ZSOP systems has been derived previously by Cooper and Murray [1969], Cooper [1970] and Humblet [1978]. For the sake of completeness, we next show how the method proposed above can be used for analyzing these systems as well.

The treatment of the continuous-time systems is similar to that of the discrete-time systems. The main difference is in the selection of AZSOP. Here, one wants to select a distribution which is all concentrated at $[0,\epsilon]$ and $\epsilon \to 0$. This would provide that once the system goes empty, the server will be able to serve the next arriving customer immediately when it arrives. For example, we may consider a distribution which is all concentrated at ϵ; In that case we have

$$r=\epsilon \ ; \ \delta^2=0 \ ; \ \frac{\delta^2}{r}=\lim_{\epsilon \to 0}\frac{\delta^2}{r}=0. \tag{4.3}$$

The equation sets for the non-symmetric NSOP cyclic systems can be found in Takagi [1986] ((4.10a-b) and (4.11a-d) for exhaustive, (5.31a-b) and (5,32a-c) for gated) where the notation used is λ_i (the Poisson rate to queue i) and b_i and $b_i^{(2)}$ (first and second moments of service time at queue i). The other notation used in that reference is identical to that of the discrete-time model. The treatment of these equations to yield the expected delay in the ZSOP systems is identical to that of Section 3, with the exception that the AZSOP is selected according to equation (4.3) and not according to equation (3.6).

The expected waiting time (limited to the time in queue and not including the time in service) in the NSOP symmetric systems is given in equations (4.33b), (5.46b) and (6.19) of Takagi [1986]. The limits of all these expressions (for the ZSOP systems) are given by:

$$E[W_i] = \frac{N\lambda b^{(2)}}{2(1-N\lambda b)} ,$$

which is, as expected, the expected waiting time in the M/G/1 system with the combined inputs of the polling system.

5. Numerical Considerations

The method described in Sections 3.4 and 3.5 requires the computation of two sets of variables which tend to zero, $\{f_i(i)\}$ and $\{f_i(i,i)\}$. As such, the method could become numerically unstable. However, careful examination of the equation sets considered reveals that the only algebraic operations used in the equations are additions and multiplications. As a result, if the equations are solved by an iterative scheme, the only operations required, during the whole solution procedure, are additions and multiplications (The only division operation required is at the end of the process when the expected delay is computed using Equation (3.1). However, this division is between two numbers of the same order of magnitude, and thus, does not introduce a numerical problem) . As such, the method is strongly immune against numerical errors. Nevertheless, when *extremely* small quantities of μ_i and σ_i^2 are used, the computer will reach its limit of precision and numerical errors will be encountered.

In contrast, the method described in Section 3.6 does not involve the computation of any small quantities. As such, the method (when implemented by an iterative scheme) is not expected to encounter special numerical problems.

Table 1: ρ=0.8		
p	Error (iterative)	Error (Eq. (3.4))
10^{-3}	$8.0\cdot10^{-3}$	$7.9\cdot10^{-3}$
10^{-5}	$8.0\cdot10^{-5}$	$7.9\cdot10^{-5}$
10^{-7}	$8.0\cdot10^{-7}$	$7.9\cdot10^{-7}$
10^{-9}	$3.0\cdot10^{-8}$	$7.9\cdot10^{-9}$
10^{-11}	$4.6\cdot10^{-6}$	$7.9\cdot10^{-11}$
10^{-13}	$2.0\cdot10^{-4}$	$7.9\cdot10^{-13}$
10^{-15}	$2.4\cdot10^{-2}$	$7.9\cdot10^{-15}$

We tested these methods in practice and compared their results with the closed-form results available for fully symmetric systems. In Table 1 we present the results obtained for a fully symmetric 10-station discrete-time cyclic-polling exhaustive system. The arrival parameters

Table 2: $\rho = 0.9$		
p	Error (iterative)	Error (Eq. (3.4))
10^{-3}	$8.9 \cdot 10^{-3}$	$8.9 \cdot 10^{-3}$
10^{-5}	$8.9 \cdot 10^{-5}$	$8.9 \cdot 10^{-5}$
10^{-7}	$8.9 \cdot 10^{-7}$	$8.9 \cdot 10^{-7}$
10^{-9}	$3.5 \cdot 10^{-8}$	$8.9 \cdot 10^{-9}$
10^{-11}	$4.3 \cdot 10^{-6}$	$8.9 \cdot 10^{-11}$
10^{-13}	$4.0 \cdot 10^{-4}$	$8.9 \cdot 10^{-13}$
10^{-15}	$4.1 \cdot 10^{-2}$	$8.9 \cdot 10^{-15}$

used are: $\mu_i = .08$, $\sigma_i^2 = .0736$ (the total system load is $\rho = 0.8$). We applied the method of Sections 3.4 and 3.5 by computing the values of Equation (3.2a-b) and then iterating on Equation (3.3a-d), where the input parameters for r_i and δ_i^2 are taken from Equation (3.6) and p varies from 10^{-3} to 10^{-15}. The iterative scheme was set to stop when the relative change was less than 10^{-14}. The relative error in the expected delay calculated by this method (in comparison to the exact result given in Equation (3.7)) is given in the second column of Table 1 (in the first column, the value of p is given). For comparison, we also give (third column) the relative error between Equation (3.7) and the result derived by substituting r_i and δ_i^2 into Equation (3.4). Similar results are presented in Table 2 for a similar system where $\mu_i = 0.09$ and $\delta_i^2 = 0.819$. The results show that for all practical purposes this numerical technique is very accurate and it can derive the expected delay to within accuracy of 10^{-7} (depending, obviously, on the computer and language used; The particular computer used here is VAX 11/785 and we used FORTRAN-77 with double precision arithmetic). The accuracy improves when p decreases, and numerical problems are encountered, as expected, only when very small values of p are used ($p < 10^{-8}$).

When this level of precision is not satisfactory, one can use the method described in Section 3.6. In our examination of this method we did not encounter any numerical problems. The relative error of this method, when applied to the two cases presented above (and to many others, including, e.g., cases with $\rho = 0.3$, 0.5 and with higher second moments δ_i^2), was consistently less than 10^{-14}. Note that the computation error, in this case, is a function of the number of iterations for which the method ran. For an analysis of the method accuracy as function of the number of iterations applied see Levy [1986] (That analysis was given for NSOP systems but applies for the method described in Section 3.6 as well).

Acknowledgements

The work of L. Kleinrock was supported in part by the Defense Advanced Research Projects Agency of the Department of Defense under contract MDA 903-82-C-0064. We would like to thank an anonymous referee of Kleinrock and Levy [1985] for bringing to our attention the problem of polling systems with zero switch-over periods.

REFERENCES

[1] J. E. Baker and I. Rubin 1987, "Polling with a General-Service Order Table," *IEEE Trans. Commun.* vol. COM-35, pp. 283-288, March 1987.

[2] R. B. Cooper 1970, "Queues Served in Cyclic Order: Waiting Times," *Bell Sys. Tech. J.*, vol. 49, pp. 399-413, Mar. 1970.

[3] R. B. Cooper and G. Murray 1969, "Queues Served in Cyclic Order," *Bell Sys. Tech. J.*, vol. 48, pp. 675-689, Mar. 1969.

[4] B. Doshi, "Queueing Systems with Vacations - A Survey," *Queueing Systems*, 1 (1986), pp. 29-66.

[5] M. Eisenberg, "Queues with periodic service and changeover time," *Oper. Res.*, Vol. 20, pp. 440-451, 1972.

[6] M. J. Ferguson and Y. J. Aminetzah 1985, "Exact results for nonsymmetric token ring systems," *IEEE Trans. Commun.* vol. COM-33, pp. 223-231, Mar. 1985.

[7] O. Hashida 1972, "Analysis of Multiqueue," Rev. Electrical Communication Lab. 20 (3,4), pp. 189-199, 1972.

[8] P. Humblet 1978, "Source coding for communication concentrators," Electron. Syst. Lab., Massachusetts Inst. Technol., Cambridge, ESL-R-798, Jan. 1978.

[9] L. Kleinrock and H. Levy 1985, "The Analysis of Random Polling Systems," To appear in *Operations Research*.

[10] A. G. Konheim and B. Meister 1974, "Waiting times and lines in system with polling," *J. Ass. Comput. Mach.*, vol. 21, pp. 470-490, 1974.

[11] H. Levy 1986, "Delay Computation and Dynamic Behavior of Non-Symmetric Polling Systems," AT&T Bell Laboratories, February 1986. Submitted for publication.

[12] M. Nomura and K. Tsukamoto 1978, "Traffic Analysis on Polling Systems," *The Transactions of the Institute of Electronics and Communication Engineering of Japan* J61-B(7), pp. 600-607 (July 1978).

[13] I. Rubin and L. F. DeMoraes 1983, "Message delay analysis for polling and token multiple-access schemes for local communication networks," *IEEE on selected areas in comm.* vol SAC-1, No. 5, Nov. 1983.

[14] G. B. Swartz 1980, "Polling in a loop system," *J. Ass. Comp. Mach.*, vol 27, pp. 42-59, 1980.

[15] H. Takagi 1985, "Mean Message Waiting Times in Symmetric Multi-Queue Systems with Cyclic Service," *Performance Evaluation*, Vol. 5, No. 4, pp. 271-277, 1985.

[16] H. Takagi 1986, "Analysis of Polling Systems," MIT Press, April 1986.