ELSEVIER

Computer Communications 20 (1997) 1292–1300

computer communications

An exact analysis of an asymmetric polling system with mixed service discipline and general service order

Lain-Chyr Hwang, Chung-Ju Chang*

Department of Communication Engineering and Center for Telecommunications Research, National Chiao Tung University, Hsinchu, Taiwan 300, Republic of China

Received 12 September 1996; revised 2 May 1997; accepted 5 May 1997

Abstract

This paper derives the exact mean waiting time for an asymmetric polling system with general service order and mixed service discipline. The mixed service discipline means that the service discipline of each stage (a turn in the service order sequence) for the same station can be gated or exhaustive. The general service order denotes that each station can be polled more than once in a polling cycle. We use the mean age and the mean excess of a cumulative time to obtain the mean waiting times for stages and stations, where the cumulative time for a stage is defined as the total arrival time period of all customers that are served at one visit of the server to the stage. The accuracy of our analysis is verified by comparisons with previously published results and simulation results. We also use a genetic algorithm (GA) to search for an optimal pattern of service order and service discipline for the asymmetrical polling system. The results of the paper can be applied to the design of computer communication networks with polling schemes. © 1997 Elsevier Science B.V.

Keywords: General service order; Genetic algorithm; Mean age; Mean excess; Mean waiting time; Mixed service discipline; Polling system

1. Introduction

Polling systems are widely applied in computer and communication systems. There are many kinds of polling systems, which employ different service orders and service disciplines. Takagi presented an excellent survey [1], where various polling systems were studied, further researches were proposed, and many references were listed. In previous research, exact solutions for polling systems with a unique service discipline of exhaustive or gated or limited one were yielded, and they were generally obtained by way of imbedded Markov chain analysis [2–5]. Previously we had derived the exact solution for the finite system with mixed service discipline and general service order [6] successfully. However, the computer algorithm requires many hours of CPU time.

In this paper, we study a polling system with general service order and mixed service discipline by way of exact approach, which is computationally efficient. Each station in the system may have multiple turns of polls and each stage (a turn in the service order sequence) may be independently assigned with gated or exhaustive service discipline (even stages corresponding to the same station). The general service order and mixed service discipline can make the system more flexible and able to meet levels of quality of service more easily for asymmetrical computer communication systems. It can support flexibility for designing an asymmetrical polling system, while the cyclic service order and single service discipline can only support an unique pattern.

Unlike the descendant set approach [7], which preformed analyses using probability generating function and the Laplace–Stieltjes transform, we propose an alternative approach that uses the mean age and the mean excess of a cumulative time to derive the exact mean waiting time. Here, the cumulative time for a stage is defined as the total arrival time period of all arrivals to the stage that will be completely served during the next visit of the server to the stage. (A detailed definition will be given in the next section.) Moreover, we use the concept of conditional mean to find a correlation between visit times and a correlation between walking time and visit time, which are key parameters for obtaining the second moments of the cumulative times. The analytical method is straightforward and can be easily followed, and the numerical algorithm is treatable. Sarkar and Zandwill [8] have used the aspect of mean to derive the mean waiting times for a polling system with

* Corresponding author. Tel: 886-3-5731923; fax: 886-3-5710116; e-mail:cjchang@cc.nctu.edu.tw

PII S0140-3664(97)00110-2

gated or exhaustive service discipline and cyclic service order. However, they only analyzed the cyclic service order, where they defined different renewal points for different service disciplines; it would be difficult to extend their approach to mixed service discipline and general service order.

To justify the exactness of our analysis, we use the same numerical examples as those of some previous papers [5,9], which considered systems with gated or exhaustive service discipline. The results show that they match perfectly. However, the systems in these previous papers are simply special cases of ours, and our method can be applied to more complex systems. We also provide several numerical examples for general systems with mixed service disciplines and compare our calculated results with simulations. The results show that our method is highly accurate. Finally, we use a genetic algorithm (GA) to find an optimal pattern of service order and service discipline for the asymmetric polling system. The results can be applied to the design of computer communication networks with polling schemes. Borst et al. [10] had also studied the optimization of the polling system by minimizing the waiting cost. However, the system they studied has limited service discipline and cyclic service order, and furthermore their analytical approach is approximate.

The paper is organized as follows. In Section 2 we perform the analysis to obtain the mean waiting times. The numerical algorithm and some numerical examples are presented in Section 3. Concluding remarks are given in Section 4.

2. Analysis

The polling system is assumed to have R stations and P stages (pseudostations). We let r and i denote the indexes of a station and a stage, respectively, and let r_i stand for the underlying station of stage i. The arrival process for station r is assumed to be an independent Poisson process with rate λ_r, The service time of a customer at station r, denoted by S_r, follows an independent general distribution with mean s_r and the second moment $s_r^{(2)}$. The walking time for stage i, denoted by U_i, follows an independent general distribution with mean u_i, the second moment $u_i^{(2)}$, and the total mean walking time of $u = \sum_{i=1}^{P} u_i$. Note that the walking time for stage i is defined as the time from the server's departure from stage i to the server's arrival at stage $i \oplus 1$, where the operation $i \oplus j$ ($i \ominus j$) equals $i + j$ ($i - j$) with modulo-P arithmetic and equals P if the remainder is zero. The traffic intensity for station r is denoted by ρ_r; $\rho_r = \lambda_r s_r$ and the total traffic intensity $\rho = \sum_{r=1}^{R} \rho_r$. The service discipline of each stage is either gated or exhaustive. Service disciplines can be different even for stages corresponding to the same station. If stage i is assigned the gated (exhaustive) service discipline, we refer to it as gated (exhaustive) stage i. We call a customer who receives service at stage i an i-customer.

We define here the *cumulative time* for stage i, denoted by C_i, as the cumulative arrival time period of all customers that are simultaneously served during one visit of the server to stage i. As shown in Fig. 1(a), C_i is given by

$$C_i = \begin{cases} \text{The time interval between } GG' \text{ for gated stages } b_i \text{ and } i, \\ \text{The time interval between } GE' \text{ for gated stage } b_i \text{ and exhaustive stage } i, \\ \text{The time interval between } EG' \text{ for exhaustive stage } b_i \text{ and gated stage } i, \\ \text{The time interval between } EE' \text{ for exhaustive stages } b_i \text{ and } i, \end{cases}$$

where b_i is the first stage before stage i that corresponds to the same station r_i and G (G') and E (E') are the beginning and the ending time, respectively, of the server's visit to stage b_i (i). Note that the cumulative time for stage i is dependent on the service disciplines assigned at stage b_i and stage i. Let c_i and $c_i^{(2)}$ be the mean and the second moment of C_i. We also define a determining time for stage i, denoted by D_i, as the time interval from the beginning of the cumulative time for stage i to the server's arrival epoch at stage i. As Fig. 1(a) shows, D_i is given by

$$D_i = \begin{cases} \text{The time interval } GG' \text{ for gated stage } b_i, \\ \text{The time interval } EG' \text{ for exhaustive stage } b_i \end{cases}$$

Note that D_i is only dependent on the service discipline of stage b_i. Let d_i and $d_i^{(2)}$ be the mean and the second moment of D_i. We call C_i the cumulative time for stage i because all arriving i-customers are accumulated during C_i and simultaneously served at the next visit of the server, and we call D_i the determining time for stage i because D_i determines the distribution of the visit time for stage i. The visit time for stage i is the time from the server's arrival at stage i to its departure.

Moreover, for gated stage i, we denote the age of C_i and the excess of C_i by C_i^{A} and C_i^{E}, respectively. C_i^{A} is the time interval from the beginning epoch of C_i to an arbitrary time epoch and C_i^{E} is the time interval from an arbitrary time epoch to the ending epoch of C_i.

The waiting time of an arbitrarily selected (labeled) i-customer is equal to the time from its arrival to the server's scan-instant at stage i plus the service time of the i-customers arriving before the labeled i-customer. As shown in Fig. 1(b), owing to the Poisson Arrivals See Time Averages (PASTA) property [11], the former time is C_i^{E} and the i-customers arriving before the labeled i-customer are those arriving during C_i^{A}. Consequently, the mean waiting time of a gated i-customer, denoted by w_i^{G}, can be expressed as

$$w_i^{\mathrm{G}} = E(C_i^{\mathrm{E}}) + \lambda_{r_i} E(C_i^{\mathrm{A}}) s_{r_i} = (1 + \rho_{r_i}) \cdot \frac{c_i^{(2)}}{2c_i} \tag{1}$$

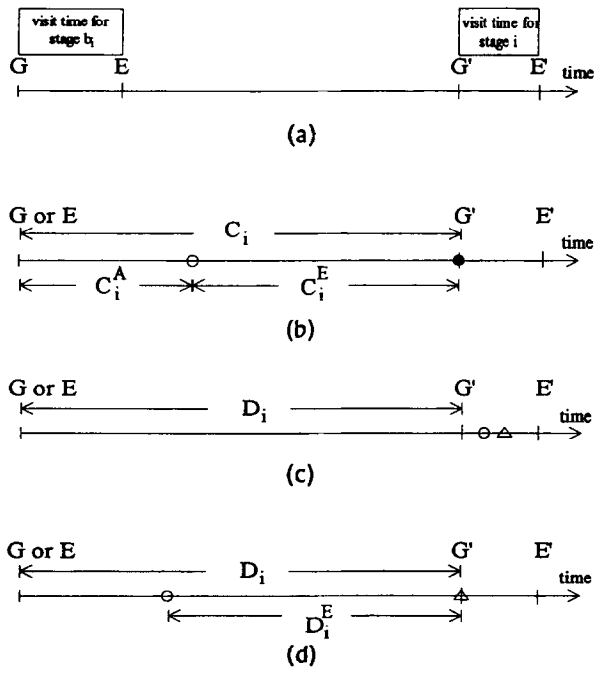

Fig. 1. Indication of time interval for cumulative time and determining time.

Note that both the mean age and the mean excess of a random variable X are given by [12] $E[X^2]/2E[X]$.

For exhaustive stage i, the waiting time of a labeled i-customer is equal to the time from its arrival to the service-beginning of the head-of-line (HOL) i-customer plus the service time of the i-customers in the waiting queue arriving before the labeled i-customer. Note that the HOL i-customer is the one seen by the labeled i-customer. Because of the property of PASTA, the former time is the residual service time of an i-customer S_{r_i} if the server is on service at stage i when the labeled i-customer arrives at stage i [see Fig. 1(c)] or the excess of D_i for exhaustive stage i, denoted by D_i^E, if the server is not on service at stage i when the labeled i-customer arrives at stage i [see Fig. 1(d)]. The mean residual service time of S_{r_i} is equal to $(s_{r_i}^{(2)})/(2s_{r_i})$; the mean excess of D_i is equal to $(d_i^{(2)})/(2d_i)$; and the probability that an i-customer finds the server on service at stage i when the i-customer arrives at the system is ρ_{r_i}. Therefore, the mean waiting time of an exhaustive i-customer, denoted by w_i^E, is given by

$$w_i^E = \rho_{r_i} \cdot \frac{s_{r_i}^{(2)}}{2s_{r_i}} + (1-\rho_{r_i}) \cdot \frac{d_i^{(2)}}{2d_i} + \lambda_{r_i} w_i^E s_{r_i}$$

or

$$w_i^E = \frac{\rho_{r_i}}{1-\rho_{r_i}} \cdot \frac{s_{r_i}^{(2)}}{2s_{r_i}} + \frac{d_i^{(2)}}{2d_i}$$

Because C_i is equal to $D_i + V_i$ for exhaustive stage i, where V_i denotes the visit time at stage i, with mean v_i, we can rewrite w_i^E as follows [9,13]:

$$w_i^E = (1-\rho_{r_i}) \cdot \frac{c_i^{(2)}}{2c_i} \tag{2}$$

Considering w_i^G in Eq. (1) and w_i^E in Eq. (2), we can generally express the mean waiting time for stage i, denoted by w_i, as

$$w_i = (1 + I_i \rho_{r_i}) \cdot \frac{c_i^{(2)}}{2c_i} \tag{3}$$

where I_i is 1 for gated stage i and -1 for exhaustive stage i. Note that c_i and $c_i^{(2)}$ in Eq. (3) are different for gated and exhaustive stage i. Consequently, the mean waiting time for station r of the system, denoted by $\bar{w}_r$ can be obtained by

$$\bar{w}_r = \sum_{\{i | r_i = r\}} \frac{\lambda_{r_i} c_i}{\lambda_{r_i} c} \cdot w_i = \frac{1}{2c} \sum_{\{i | r_i = r\}} (1 + \rho_{r_i} I_i) \cdot c_i^{(2)} \tag{4}$$

where $c \equiv \sum_{i=1}^{P} c_i$ is the mean whole cumulative time and is equal to $u/(1-\rho)$ because of $c = c\rho + u$.

To obtain the mean waiting time $\bar{w}_r$, we now need to find $c_i^{(2)}$, the second moment of the cumulative time for stage i corresponding to station r. Because the cumulative time C_i is dependent on the service disciplines of stage b_i and stage i, we first find $d_i^{(2)}$ of the determining time D_i, which is dependent only on the service discipline of stage b_i. Intuitively, D_i can be expressed as

$$D_i = \begin{cases} \sum_{j=b_i}^{i \ominus 1} (V_j + U_j), & \text{for gated stage } b_i, \\ \sum_{j=b_i}^{i \ominus 1} U_j + \sum_{j=b_i \oplus 1}^{i \ominus 1} V_j, & \text{for exhaustive stage } b_i, \end{cases} \tag{5}$$

where $\bigcap$ is defined as

$$\bigcap_{j=m}^{n} x_j = \begin{cases} \sum_{j=m}^{n} x_j, & \text{if } m \le n \\ \sum_{j=m}^{P} x_j + \sum_{j=1}^{n} x_j, & \text{if } m \ge n+1 \end{cases}$$

Then $d_i^{(2)}$ can be obtained by

$$d_i^{(2)} = \begin{cases} \bigcap_{j=b_i}^{i \ominus 1} (R_{jj} + u_j^{(2)}) + 2 \left[\bigcap_{j=b_i}^{i \ominus 2} \bigcap_{k=j \oplus 1}^{i \ominus 1} R_{jk} + \bigcap_{j=b_i}^{i \ominus 1} \bigcap_{k=j}^{i \ominus 1} v_j u_k + \bigcap_{j=b_i}^{i \ominus 2} \bigcap_{k=j \oplus 1}^{i \ominus 1} (\hat{R}_{jk} + u_j u_k) \right] \\ \quad \text{if } b_i \text{ is a gated stage} \\ \bigcap_{j=b_i}^{i \ominus 1} u_j^{(2)} + \bigcup_{j=b_i \oplus 1}^{i \ominus 1} R_{jj} + 2 \left[\bigcup_{j=b_i \oplus 1}^{i \ominus 2} \bigcap_{k=j \oplus 1}^{i \ominus 1} R_{jk} + \bigcap_{j=b_i \oplus 1}^{i \ominus 1} \bigcap_{k=j}^{i \ominus 1} v_j u_k + \bigcap_{j=b_i}^{i \ominus 2} \bigcap_{k=j \oplus 1}^{i \ominus 1} (\hat{R}_{jk} + u_j u_k) \right] \\ \quad \text{if } b_i \text{ is an exhaustive stage} \end{cases} \tag{6}$$

where

$$\bigcup_{j=m}^{n} x_j = \begin{cases} 0 & \text{if } m = n \oplus 1 \\ \sum_{j=m}^{n} x_j & \text{otherwise} \end{cases}$$

and R_{jj}, R_{jk} and $\hat{R}_{jk}$ are correlations defined as $R_{jj} \equiv E[V_j^2]$, $R_{jk} = E[V_j V_k | j \to k, \quad j \neq k] \equiv E[V_k V_j | j \to k, \quad j \neq k]$, and $\hat{R}_{jk} \equiv E[U_j V_k | j \to k]$. The $j \to k$ in the above definition means that stage j is served before stage k. If $j < k$, $j \to k$ denotes that stage j is served in the same cycle (from stage 1 to stage P) as stage k; if $j \geq k$, $j \to k$ indicates that stage j is served in the last cycle of stage k. Note that stages j and k are located between stages b_i and i, and $R_{jk} \neq R_{kj}$ and $\hat{R}_{jk} \neq \hat{R}_{kj}$.

We use the conditional mean to find the unknown correlations in Eq. (6). The second moment of the visit time, R_{jj}, can be found by following Ref. [13], see Appendix A:

$$R_{jj} = \begin{cases} \lambda_{r_j} d_j s_{r_j}^{(2)} + P_{rj}^2 d_j^{(2)}, \\ \text{for gated stage } j \\ \dfrac{s_{r_j}^{(2)}}{(1-\rho_{r_j})^3} \cdot \lambda_{r_j} d_j + \left[\dfrac{\rho_{r_j}}{1-\rho_{r_j}}\right]^2 \cdot d_j^{(2)}, \\ \text{for exhaustive stage } j \end{cases} \tag{7}$$

Additionally, $R_{jk} \equiv E[V_j V_k, \ |j \to k, \ j \neq k] = E[E(V_j V_k | j \to k, \ j \neq k, \ V_j, \ D_k)]$ can be obtained by

$$R_{jk} = \begin{cases} \rho_{r_k} E[V_j D_k | j \to k, \ j \neq k], \\ \text{for gated stage } k \\ \dfrac{\rho_{r_k}}{(1-\rho_{r_k})} E[V_j D_k | j \to k, \ j \neq k], \\ \text{for exhaustive stage } k \end{cases} \tag{8}$$

$E[V_j D_k | j \to k, \ j \neq k]$ in Eq. (8) is dependent on the service discipline of stage b_k. For a gated stage b_k, it is given by

$$E[V_j D_k | j \to k, \ j \neq k] = \begin{cases} \bigcap_{m=b_k}^{k\ominus 1} (R_{jm} + v_j u_m), & \text{for } b_k \in (j,k) \\ \bigcup_{m=b_k}^{j\ominus 1} (R_{mj} + \hat{R}_{mj}) + \bigcap_{m=j}^{k\ominus 1} (R_{jm} + v_j u_m), & \text{for } b_k \notin (j,k) \end{cases} \tag{9}$$

and for an exhaustive stage b_k, it is given by

$$E[V_j D_k | j \to k, \ j \neq k] = \begin{cases} \bigcap_{m=b_k}^{k\ominus 1} v_j u_m + \bigcap_{m=b_k\oplus 1}^{k\ominus 1} R_{jm}, \\ \text{for } b_k \in (j,k) \\ \bigcup_{m=b_k}^{j\ominus 1} \hat{R}_{mj} + \bigcup_{m=b_k\oplus 1}^{j} R_{mj} + \bigcap_{m=j}^{k\ominus 1} v_j u_m + \bigcup_{m=j\oplus 1}^{k\ominus 1} R_{jm}, \\ \text{for } b_k \notin (j,k) \end{cases} \tag{10}$$

where $(j,k) = \{j \oplus 1, j \oplus 2, \ldots, k\ominus 2, k\ominus 1\}$ and $b_k \in (j,k)$ means stage b_k is served after stage j and before stage k. Similarly, $\hat{R}_{jk}$ can be obtained by way of a similar approach [13] (see Appendix A).

3. Numerical examples

There are $2P^2$ unknown correlations in Eq. (6). As seen from Eqs. (7)–(10), an iterative algorithm is required to find the solutions of these correlations. The termination criterion for the iterative algorithm is here de fined as the absolute difference between two successive test values of less than 10^{-7}. We summarize the numerical algorithm for finding the mean waiting times as follows:

Numerical algorithm

Step 0: [Set system conditions]

(i) Set R, P, service order sequence, and service discipline of each stage.
(ii) Set λ_r, s_r and $s_r^{(2)}$ for all r, and obtain $\rho_r = \lambda_r s_r$ and $\rho = \sum_{r=1}^{R} \rho_r$.
(iii) Set u_i and $u_i^{(2)}$ for all i, and obtain $u = \sum_{i=1}^{P} u_i$ and $c = u/(1-\rho)$.
(iv) Set the termination criterion.

Step 1: [Obtain mean visit times and mean cumulative times]

(i) Initially, set $c_i = c/P$ and then $v_i = \rho_{r_i} c_i$ for all i.
(ii) Find d_i from Eq. (5) for all i.
(iii) Obtain a newer

$$c_i = \left\{ \begin{matrix} d_i & \text{for gated stage } i \\ d_i + v_i & \text{for exhaustive stage } i \end{matrix} \right\} \text{for all } i.$$

(iv) Obtain a newer $v_i = \rho_{r_i} c_i$ for all i.
(v) IF v_i does not satisfy the termination criterion for any i

GO TO (ii) in this *step*

END IF

Step 2: [Obtain correlations and the second moment of the cumulative times]

(i) Initially, $R_{jk} = v_j v_k$ and $\hat{R}_{jk} = u_j v_k$ for all j and k.
(ii) Find $d_i^{(2)}$ from Eq. (6) for all i.
(iii) Obtain a newer set of R_{jk} and $\hat{R}_{jk} = u_j v_k$ for all j and k, and all newer $d_i^{(2)}$ from Eq. (6).
(iv) IF $d_i^{(2)}$ does not satisfy the termination criterion for any i

GO TO (iii) in this *Step*

END IF

(v) Obtain

$$c_i^{(2)} = \begin{cases} d_i^{(2)}, \\ \text{for gated stage } i \\ \dfrac{1}{(1-\rho_{r_i})^2} d_i^{(2)} \dfrac{s_{r_i}^{(2)}}{(1-\rho_{r_i})^3} \lambda_{r_i} d_i, \\ \text{for exhaustive stage } i \end{cases}$$

Step 3: [Obtain the mean waiting times]

(i) $w_i = (1+\rho_{r_i} I_i)\,(c_i^{(2)})/(2c_i)$ for all i.
(ii) $\bar{w}_r = (1)/(2c) \sum_{\{i|r_i=r\}} (1+\rho_{r_i}) c_i^{(2)}$ for all r.
(iii) END

We may also construct sets of simultaneous equations to find the mean cumulative times and the correlations. Instead, we use the iterative schemes in the numerical algorithm, because it is easier to write iterative schemes than to construct the simultaneous equations and to solve them in the numerical program. The number of arithmetic operations to find all the P mean waiting times for general service order is $O(P^2)$ per iteration, which is the same as that in Konheim et al.'s paper [7]. Furthermore, the space used to save the correlations in our algorithm is $O(P^2)$, which is larger than that in their paper. However, the number of iterations by our algorithm is smaller than that in Ref. [7] which is similar to Choudhury's paper [5]. The numbers of iterations for Cases A and B in Table 1 of Choudhury's paper are 32 and 164 by Choudhury's method, while they are only 27 and 120 by our algorithm.

We first run all of the examples in Table 1 of Choudhury's paper [5] and Tables I–VII of Everitt's paper [9] to verify the correctness of our analysis. We also find that the results match perfectly. We next discuss an example of an asymmetrical polling system with general service order and mixed service discipline. This example is assumed to have six stations; the service time for every station is exponentially distributed with mean 1; the walking time for every stage is constant and equal to 0.1; and the arrival processes are Poisson processes with rate 0.2ρ for stations 1–4 and 0.1ρ for stations 5 and 6. The service order sequence is {1, 2, 3, 4, 5, 6, 3, 4}; odd stages adopt the exhaustive service discipline and even stages adopt the gated service discipline. Note that the asymmetrical polling system cannot be analyzed by Choudhury's and Everitt's algorithms. The analytical results are shown in Fig. 2. We find that they agree with the simulation results very well.

We also find that the single poll stations possess the same characteristics as in the cyclic case. For stations with the exhaustive service discipline, the mean waiting time of the heavy-load station (station 1) is less than that of the light-load station (station 5), and for stations with the gated service discipline, the mean waiting time of the heavy-load station (station 2) is greater than that of the light-load station (station 6). The first of these characteristics has been discussed previously [2,14,15], and it was concluded [14] that this is a result of the hogging property of the heavy-load station with the exhaustive service discipline. The hogging property of a station with the exhaustive service discipline means that the station will occupy the server longer, thereby reducing the mean waiting time for that station. The exhaustive service discipline serves a station until the station is empty, so the heavy-load station has a stronger hogging property. On the other hand, multiple poll stations (stations 3 and 4) have smaller mean waiting times than the others. From this example, we can infer that there exists an appropriate pattern of service order and service discipline if a performance criterion (or say, fitness function) is defined.

We use a genetic algorithm (GA) to search for an optimal

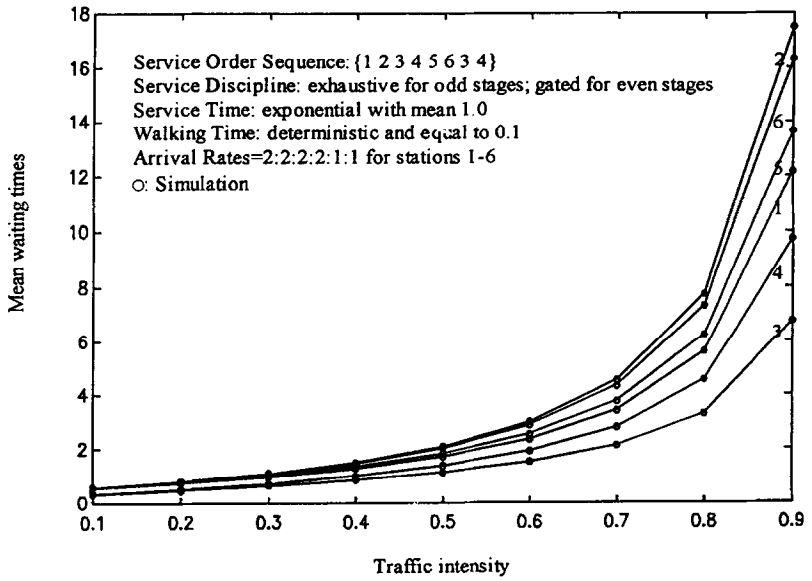

Fig. 2. Mean waiting times of customers for an asymmetric system with general service order sequence and mixed service discipline.

pattern of service order and service discipline for the system. Generally, there are three main types of search methods: calculus-based, enumerative, and random. Calculus-based methods are suitable for continuous and unimodal problems; enumerative methods are suitable for problems with a small search space; and random methods are suitable for discontinuous and multimodal problems. Goldbeg [16] has compared these methods. GAs, which combine the survival of the fittest with the innovative flair of a human search, are a form of random method. They are powerful, especially when the search space is not numerical.

As in biological evolution, GAs evolve generation by generation. In *generation n*, there is a population of m candidates denoted by $\{x_1^n, x_2^n, \ldots, x_m^n\}$. The candidates of the nth generation generate the candidates of the $(n+1)$st generation $\{x_1^{n+1}, x_2^{n+1}, \ldots, x_m^{n+1}\}$ via crossover and mutation. An objective function is defined to find the fitness of candidates. The fitness of the candidates in the nth generation will influence the production of candidates for the $(n+1)$st generation. The evolution in the GA will be terminated when an acceptable approximation is found, the number of searched candidates has reached a predetermined number, or some other reasonable criterion is satisfied. During the evolution, we find the optimal candidate x_{op}, which is defined to have the maximum value of fitness $f(x_{op})$. A detailed GA procedure was described by Hwang and Chang [6].

In this paper, we heuristically define a fitness function for a given candidate x_i^n in the nth generation, denoted by $f(x_i^n)$, as

$$f(x_i^n) = \left[\sum_{r=1}^{R} \frac{\lambda_r}{\lambda} \cdot \left(\tilde{w}_r - \sum_{s=1}^{R} \frac{\lambda_s}{\lambda} \cdot \tilde{w}_s \right) \right]^{-1}$$

As the equation implies, the fitness function $f(x_i^n)$ is used to find the optimal pattern so as to obtain a fair allocation of the mean waiting time for any individual station. The explicit parameters of the fitness function $f(x_i^n)$ are the mean waiting times, which are the performance measures for the candidate x_i^n, representing a pattern of service order and service discipline (the implicit parameters). Note that the service order sequence and the service discipline of each stage are coded into a binary string of genes. Here, we utilize the GAUCSD 1.4 developed at the University of California, San Diego [17], and adopt a predetermined number of searched candidates as our termination criterion.

The example system that we are going to design is assumed to have nine stations, where station 1 has much heavier traffic load than the other stations. It is a typical client/server network that has many ordinary user stations and a file server station. We assume the arrival rate of station 1 is eight times greater than the arrival rates of the other stations. The service time distribution S_r is exponentially distributed and the mean service time s_r is equal to 1, $1 \le r \le R$; the walking time U_i is deterministic and u_i is equal to 0.1, $1 \le i \le P$. Since the example system has one heavy-load station and eight light-load stations, here we consider only eight types of service order sequences that poll the heavy-load station from one to eight times. The eight sequences are as follows. Sequence A with $P = 9$, {1, 2, 3, 4, 5, 6, 7, 8, 9}; sequence B with $P = 10$, {1, 2, 3, 4, 5, 1, 6, 7, 8, 9}; sequence C with $P = 11$: {1, 2, 3, 4, 1, 5, 6, 7, 1, 8, 9}, and so on until sequence H with $P = 16$, {1, 2, 1, 3, 1, 4, 1, 5, 1, 6, 1, 7, 1, 8, 1, 9}. As the value of P increases from sequence A to sequence H, the service order sequence is arranged so that the number of stages between two consecutive polls of station 1 remains almost the same. We use the capital letter G (E) to denote the gated (exhaustive) service discipline of a stage and refer to the aggregation of G and E for stages in the service order as the pattern of the service discipline. For example, the pattern EGEEG means E, G, E, E, and G for stages 1, 2, 3, 4 and 5, respectively.

There are a total of 130 560 ($= \sum_{n=9}^{18} 2^n$) cases in an enumerative search. GAUCSD suggests only 230 cases to be searched; the efficiency is about 99.82%. The optimal patterns and the costs are shown in Table 1, where the cost function, denoted by f^{-1}, is the inverse function of the fitness function. The optimal pattern for fitness function f is "sequence A and GGGGGGGGG" for traffic intensifies below 0.5 and is "sequence B and GEEEEGEEEE" for traffic intensifies above 0.6. The optimal pattern uses the gated service discipline for all stages for traffic intensifies below 0.5; this is because the gated service discipline inherently distributes the mean waiting times more fairly than the exhaustive service discipline [15]. When the traffic intensity is higher, however, the mean waiting times for the heavy-load station increase at a faster rate. To decrease the mean waiting time for the heavy-load station, the system assigns the heavy-load station more polls. Because a small increase in the number of polls greatly decreases the mean waiting times, two polls are enough. On the other hand, the mean waiting times for light-load stations should be also decreased in order to make them only a little larger than the mean waiting time for the two-poll station and to make the system fairer, with a lower cost. For this reason, the

Table 1
Optimal pattern of service order and service discipline for the fitness function f

Traffic intensity	Optimal pattern of service order and service disicipline	Cost (f^{-1})
0.1	Sequence A, GGGGGGGGG	0.0013
0.2	Sequence A, GGGGGGGGG	0.0069
0.3	Sequence A, GGGGGGGGG	0.0199
0.4	Sequence A, GGGGGGGGG	0.0455
0.5	Sequence A, GGGGGGGGG	0.0931
0.6	Sequence B, GEEEEGEEEE	0.1749
0.7	Sequence B, GEEEEGEEEE	0.2878
0.8	Sequence B, GEEEEGEEEE	0.5300
0.9	Sequence B, GEEEEGEEEE	1.2964

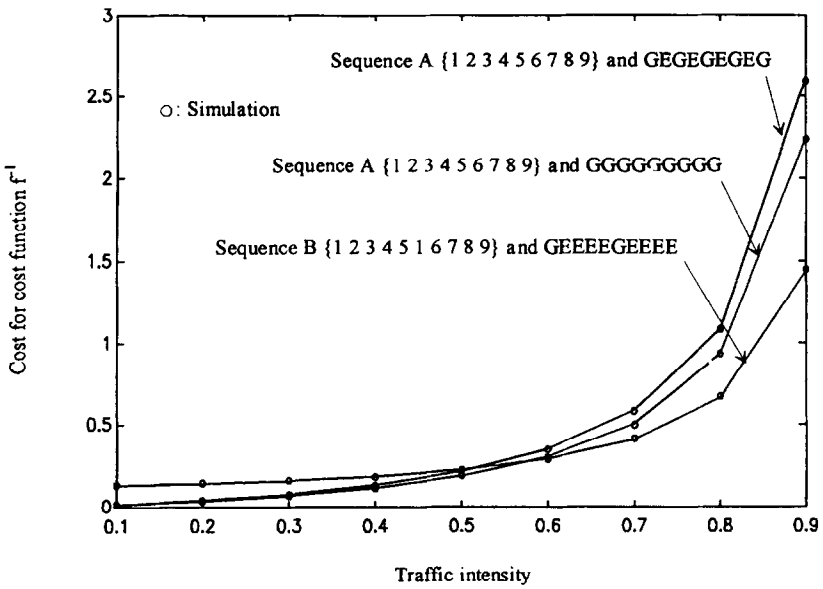

Fig. 3. Cost comparison for patterns of service order and service discipline, using the cost function f^{-1}.

system uses the exhaustive service discipline for the light-load stations.

We also plot the costs for the two optimal patterns and a randomly selected pattern "sequence A and GEGEGEGEG" vs. the traffic intensity in Fig. 3. The difference between these two optimal patterns is small below traffic intensity 0.6 and is more significant above 0.7. The cost increases about 40% if we use the pattern "sequence A and GGGGGGGGG" (cost = 0.7442) rather than the pattern "sequence B and GEEEEGEEEE" (cost = 0.5300) at traffic intensity 0.8. Moreover, the cost increases about 127% if we use the arbitrarily selected pattern "sequence A and GEGEGEGEG" (cost = 1.2039) rather than the optimal pattern "sequence B and GEEEEGEEEE" at traffic intensity 0.8. It is recommended that the optimal pattern "sequence B and GEEEEGEEEE" be adopted for all traffic intensifies.

4. Conclusions

We develop an analytical approach to derive the exact mean waiting time of a polling system with general service order and mixed (gated or exhaustive) service discipline. We first derive the mean age and the mean excess of a defined cumulative time to obtain formulae for the mean waiting times. Then we utilize the conditional mean to yield the correlations between visit times and between walking time and visit time to obtain the second moments of the cumulative times, which are key parameters in the formulae for the mean waiting times. Finally, an iterative algorithm is introduced to determine these necessary means and correlations. We use numerical examples to confirm the exactness of our approach and examine some characteristics of the polling system with general service order sequence and mixed service discipline. Furthermore, we heuristically define a fitness function and use a GA to find optimal patterns of service order and service discipline. The approach can be applied to complex polling systems.

Acknowledgements

This work is supported by National Science Council, Taiwan, Republic of China under contract number NSC 83-0404-E009-002.

Appendix A

The R_{jj} in Eq. (7) for exhaustive stage j can be found by the following equation:

$$E(V_i^2) = E[E(V_i^2|N_i)] = E[N_i \hat{s}_{r_i}^{(2)} + N_i(N_i-1)\hat{s}_{r_i}^2]$$
$$= \hat{s}_{r_i}^{(2)} E[N_i] + \hat{s}_{r_i}^2 E[N_i(N_i-1)] \quad \text{(A1)}$$

where N_i is the number of customers in station r_i when the server arrives at stage i and $\hat{s}_{r_i}$ and $\hat{s}_{r_i}^{(2)}$ are the mean and the second moment of the busy period of an M/G/1 queue on which station r_i is modeled. Define $D_i^*(s)$ as the LST of the CDF of D_i, and define $\hat{D}_i(z)$ as the pgf of the number of customers arriving during D_i. Because of the Poisson arrival process, we have

$$\hat{D}_i(z) = D_i^*(\lambda_{r_i} - \lambda_{r_i} z) \quad \text{(A2)}$$

Differentiating Eq. (A2) twice, we obtain

$$E[N_i(N_i-1)] = \lambda_{r_i}^2 d_i^{(2)}$$

Since $E[N_i] = \lambda_{r_i} d_i$, and

$$\hat{h}_{r_i}^{(2)} = \frac{h_{r_i}^{(2)}}{(1-\rho_{r_i})^3}, \text{ [18, p.20, (2.5)]}, \tag{A3}$$

after substituting Eq. (A3) into Eq. (A1), we obtain

$$R_{jj} = \frac{s_{r_j}^{(2)}}{(1-\rho_{r_i})^3} \cdot \lambda_{r_j} d_j + \left[\frac{\rho_{r_j}}{1-\rho_{r_j}}\right]^2 \cdot d_j^{(2)}$$

Similarly, using the approach above, we can find R_{jj} for gated stage j by

$$R_{jj} = E[N_j s_{r_j}^{(2)} + N_j(N_j-1)s_{r_j}^2] = \lambda_{r_j} d_j s_{r_j}^{(2)} + \rho_{r_j}^2 d_j^{(2)}$$

Using a similar method of above and in Eqs. (8)–(10), $\hat{R}_{jk}$ can be obtained by

$$\begin{cases} \hat{R}_{jk} = \rho_{r_k} E[U_j D_k | j \to k], & \text{for gated stage } k, \\ \hat{R}_{jk} = \dfrac{\rho_{r_k}}{1-\rho_{r_k}} E[U_j D_k | j \to k], & \text{for exhaustive stage } k \end{cases}$$

For a gated stage b_k, $E[U_j D_k | j \to k]$ is given by

$$E[V_j D_k | j \to k] = \begin{cases} \displaystyle\sum_{m=b_k}^{k-1} (\hat{R}_{jm} + u_j u_m), \\ \text{for } b_k \in (j,k) \\ \displaystyle\sum_{m=b_k}^{j-1} (v_m + u_j + u_m u_j) + v_j u_j + u_j^{(2)} \\ \quad + \displaystyle\sum_{m=j+1}^{k-1} (\hat{R}_{jm} + u_j u_m), \\ \text{for } b_k \notin (j,k) \end{cases}$$

and for an exhaustive stage b_k, $E[U_j D_k | j \to k]$ is given by

$$E[V_j D_k | j \to k,] \begin{cases} \displaystyle\sum_{m=b_k}^{k-1} u_j u_m + \sum_{m=b_k+1}^{k-1} \hat{R}_{jm}, \\ \text{for } b_k \in (j,k) \\ \displaystyle\sum_{m=b_k}^{j-1} u_m u_j + \sum_{m=b_k+1}^{j} v_m u_j + u_j^{(2)} \\ \quad + \displaystyle\sum_{m=j+1}^{k-1} (\hat{R}_{jm} + u_j u_m), \\ \text{for } b_k \notin (j,k) \end{cases}$$

References

[1] H. Takagi, Queueing analysis of polling model: an update, in: Stochastic Analysis of Computer and Communication Systems, Elsevier Science, North Holland, Amsterdam, 1990, pp. 267–318.

[2] M.J. Ferguson, Y.J. Aminetzah, Exact results for nonsymmetric token ring system, IEEE Transactions on Communications 33 (3) (1985) 223–231.

[3] M. Eisenberg, Queues with periodic service and changeover time, Operations Research 20 (1972) 440–451.

[4] J.E. Baker, I. Rubin, Polling with a general service order table, IEEE Transactions on Communications 35 (3) (1987) 283–288.

[5] G.L. Choudhury, Polling with a general service order table: gated service, Proceedings of the IEEE INFOCOM '90, pp. 268–276.

[6] L.C. Hwang, C.J. Chang, Optimal design of a finite-buffer polling network with mixed service discipline and general service order sequence, IEE Proceedings: Communications February (1995) 1–6.

[7] A.G. Konheim, H. Levy, M.M. Srinivasan, Descendant set: an efficient approach for the analysis of polling systems, IEEE Transactions on Communications. 42(2-4) (1994) 1245–1253.

[8] D. Saxkar, W.I. Zandwill, Expected waiting time for nonsynimetric cyclic queueing systems—exact results and applications, Management Science 35 (12) (1989) 1463–1474.

[9] D. Everitt, Simple approximations for token rings, IEEE Transactions on Communications 34 (7) (1986) 719–721.

[10] S.C. Borst, O.J. Boxma, H. Levy, The use of service limits for efficient operation of multistation single-medium communication systems, IEEE ACM Transactions on Communications 3 (5) (1995) 602–612.

[11] R.W. Wolff, Stochastic Modeling and the Theory of Queue, Prentice-Hall, Englewood Cliff, NJ, 1989.

[12] L. Kleinrock, Queueing Systems, Volume 1: Theory, Wiley, New York, 1975.

[13] L.C. Hwang, Polling schemes in computer communication networks, Ph.D. dissertation, National Chiao Tung University, Taiwan, 1994.

[14] W.Y. Jung, C.K. Un, Analysis of a finite buffer polling system with exhaustive service based on virtual buffering, IEEE Transactions on Communications 40(5) (1992) 860–862.

[15] G.L. Choundhury, H. Takagi, Comments on "Exact results for non-symmetric token ring systems", IEEE Transactions on Communications 38 (8) (1990) 1125–1127.

[16] D.E. Goldbeg, Genetic Algorithm in Search, Optimization and Machine Learning, Addison-Wesley, Reading, MA, 1986.

[17] N.N. Schraudolph, J.J. Grefenstette, A user's guide to GAucsa 1.4, Technical Report CSE DEP., UC San Diego, La Jolla, CA.

[18] H. Takagi, Queuing Analysis, A Foundation of Performance Evaluation, Volume 1 Vocation and Priority Systems, Part 1, Elsevier Science, North Holland, Amsterdam, 1991.

Lain-Chry Hwang was born in Taiwan on 26 December 1965. He received his B.S. degree in Electrical Engineering from the National Sun Yat-Sen University, Kaohsiung, Taiwan in 1988, his M.S. degree in Communication Engineering (1990) and his Ph.D. in Electronic Engineering (1994) from the National Chiao-Tung University, Hsinchu, Taiwan. From 1995 to 1996 he was with Telecommunications Laboratories, Chunghwa Telecom Co., Ltd, Taiwan as an associate researcher. There, he was involved in network planning of the intelligence network and personal communication services. In August 1996 he became associate professor in the department of Management Information Systems of the Kaohsiung Polytechnic Institute, Taiwan. His research interests include performance evaluation and computer communication networks.

Chung-Ju Chang was born in Taiwan in 1950. He received his B.E. and M.E. degrees in Electronics Engineering from the National Chiao-Tung University, Hsinchu, Taiwan in 1972 and 1976, respectively, and his Ph.D. in Electronic Engineering from the National Taiwan University in 1985. Between 1976 and 1988 he was with Telecommunications Laboratories, Directorate General of Telecommunications, Ministry of Communications, Republic of China, as a design engineer, supervisor, project manager, and then division director. There, he was involved in designing a digital switching system, RAX trunk tester, ISDN user–network interface, and ISDN service and technology trials in Science-Based Industrial Park. In the meantime he also acted as a science and technical advisor for the Minister of the Ministry of Communications from 1987 to 1989. In August 1988 he joined the Faculty of the Department of Computer Science, National Chiao-Tung University. He is currently a professor and also serves as an advisor for the Ministry of Education, Ministry of Transportation and Communication and Ministry of Economical Affairs. He was director of the Institute of Communication Engineering, National Chiao-Tung University from August 1993 to July 1995. His research interests include performance evaluation, ATM (asynchronous transfer mode) networks and PCS (personal communication service) networks. Dr Chang is a member of the Chinese Institute of Engineers (CIE) and IEEE.