Polling in a closed network *

Eitan Altman
INRIA
Centre Sophia Antipolis
06565 Valbonne Cedex, France

Uri Yechiali
Department of Statistics and OR
Tel-Aviv University
Tel-Aviv 69978, Israel

July 1993
Revised: December 1993

Abstract

We consider a closed queueing network with a fixed number of customers, where a single server moves cyclically between N stations, rending service in each station according to some given discipline (Gated, Exhaustive, or the Globally Gated regime). When service of a customer (message) ends in station j, it is routed to station k with probability P_{jk}. We derive explicit expressions for the probability generating function and the moments of the number of customers at the various queues at polling instants, and calculate the mean cycle duration and throughput for each service discipline. We then obtain the first moments of the queues' length at an arbitrary point in time. A few examples are given to illustrate the analysis. Finally, we address the problem of optimal dynamic control of the order of stations to be served.

Keywords: Polling systems, Routing, Closed network.

1 Introduction

In polling systems that have been studied in the literature (e.g. [1, 3, 4, 6, 7, 8, 9]), one usually asserts independent Poisson arrivals. However, there are situations where arrivals strongly

*Supported by a Grant from the France-Israel Scientific Cooperation (in Computer Science and Engineering) between the French Ministry of Research and Technology and the Israeli Ministry of Science and Technology, Grant Number 3321190.

depend on the departure process. One encounters such systems when an attempt is made to model and analyze interactive communications via Local Area Networks whose operation is based on token ring protocols. In such cases, a message is sent from a station only after it receives a message from another station. Each station can transmit only when it receives the token (typically a predetermined sequence of bits).

This leads us to consider, investigate and analyze closed polling systems, where a fixed number of M customers are routed among N stations, with no exogenous arrivals or outside departures. A single server moves cyclically between the queues, serving during each visit to a queue some customers, whose number is determined by the service discipline. We focus on the Gated, Exhaustive and Globally Gated regimes. Unlike many systems analyzed in the literature, an (endogenous) arrival occurs only when a service is completed in some queue, and the served customer is routed to another queue. The routing is governed by some transition probability matrix P.

An open polling system has been recently studied in [6] where each station receives an independent stream of exogenous Poisson arrivals, and a customer, after being served in some queue, is either routed to another queue or leaves the network. There seem to be several basic differences between the two models; in [6] the total arrival intensity into each queue is only a function of the exogenous arrival rates and the routing matrix P, whereas in our model the arrival densities depend on P, as well as on the mean service times and walking times. In addition, in our model the server's utilization depends also on the walking times, which is not the case in the open system. Finally, in the model studied here, the number of customers in the system is always the same and fixed, whereas in the open network this number fluctuates.

We derive explicit expressions for the probability generating function (PGF) and the moments of the number of customers at polling instants, and calculate the mean cycle duration, throughput and utilization for each service discipline. We then obtain the first moments of the queues' length at an arbitrary instant. A few examples are given to illustrate the analysis.

It is remarkable that, for a closed system, explicit expressions can be obtained for the PGF and second moments of the number of customers in the system. This is in contrast to the situation in open polling systems, where usually only implicit equations are obtained for the corresponding PGF, and where second moments are expressed as an (implicit) solution of a set of linear equations.

After introducing the model in Section 2, we analyze in Section 3 the Gated discipline.

In Section 4 we study a mixed regime where some queues are served exhaustively, while the others are served according to the Gated discipline. A special pleasant feature of the closed polling system is that by a simple transformation of the transition probability matrix P, the mixed regime can be obtained as a special case of the Gated discipline.

Finally, we analyze in Section 5 the Globally Gated regime, and then address the problem of optimal dynamic control of the order of stations to be served.

2 The model

A closed-network queueing system consists of N stations (queues), M customers (jobs, programs, files, packets, etc.) and a single server who moves through the channels, typically in a cyclic order (say 1,2,3,...,N-1,N,1,2,...). Upon completion of a visit at channel i, the server incurs a random switch-over time, D_i, having mean d_i, second moment $d_i^{(2)}$ and Laplace-Stieltjes Transform (LST) $d_i^*(\bullet)$. We shall denote $d = \sum_{i=1}^N d_i$ the total expected switch-over time in a cycle. The server resides in each queue according to the service discipline (e.g. Gated, Exhaustive, Globally-Gated), and the service times of individual customers at queue i ($i = 1, 2, ..., N$) are i.i.d. random variables (RVs) all distributed as a RV B_i, having mean b_i, second moment $b_i^{(2)}$, and (LST) $b_i^*(\bullet)$. Upon completion of service at queue i, a customer is routed (instantaneously) to queue j with probability P_{ij} ($\sum_{j=1}^N P_{ij} = 1$ for every i), and joins the queue there. We assume, for simplicity, that the matrix $P = \{P_{ij}\}$ is irreducible, but allow P to be periodic, which implies that a steady state distribution need not necessarily exist. The stationary probabilities will be understood as the Cezaro limit of the finite time probabilities. Thus, the stationary distribution of a Markov Chain whose transition probabilities are $\hat{P}$ is given by the limit of $t^{-1}\sum_{s=1}^t \hat{P}^s$, as t goes to infinity.

3 The Gated Discipline

In this Section we analyze the Gated service discipline by which, in every visit to a queue, the server serves only the customers present at the polling instant. New arrivals to the queue while being attended by the server will wait for the next visit.

Let X_i^j be the number of customers under the stationary regime at queue j when queue

i is polled (instant where the server enters queue i), and let $\underline{X_i} = (X_i^1, X_i^2, ..., X_i^N)$. We have

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

where $A_i^j = A_i^j(X_i^i)$ is the number of customers (out of X_i^i) that are routed from queue i to queue j during the server's visit in queue i. The A_i^j have a multinomial distribution:

$$P\left(A_i^j(X_i^i) = m_j, j = 1, 2, ..., N \,\middle|\, X_i^i\right) = \frac{X_i^i!}{\prod_{j=1}^N m_j!} \prod_{j=1}^N P_{ij}^{m_j} \tag{2}$$

where $\sum_{j=1}^N m_j = X_i^i$. Let $\underline{A_i} = (A_i^1, A_i^2, ..., A_i^N)$.

The probability generating function

Let $G_i(\underline{z}) = G_i(z_1, z_2, ..., z_N) = E\left[\prod_{j=1}^N z_j^{X_i^j}\right]$ be the probability generating function (PGF) of the number of customers at the various queues at polling instants of queue i. Then

$$\begin{aligned} G_{i+1}(\underline{z}) &= E\left\{E\left[\left(\prod_{\substack{j=1 \\ j \neq i}}^N z_j^{X_i^j + A_i^j}\right) z_i^{A_i^i} \,\middle|\, \underline{X}_i\right]\right\} = E\left\{\prod_{\substack{j=1 \\ j \neq i}}^N z_j^{X_i^j} E\left[\left(\prod_{j=1}^N z_j^{A_i^j(X_i^i)}\right) \middle|\, \underline{X}_i\right]\right\} \\ &= E\left\{\prod_{\substack{j=1 \\ j \neq i}}^N z_j^{X_i^j} \left(\sum_{j=1}^N P_{ij} z_j\right)^{X_i^i}\right\} \end{aligned}$$

The last equality follows since

$$\begin{aligned} E\left[\left(\prod_{j=1}^N z_j^{A_i^j(X_i^i)}\right) \middle|\, \underline{X}_i\right] &= \sum_{\sum_{j=1}^N m_j = X_i^i} \prod_{j=1}^N z_j^{m_j} P\left[A_i^j(X_i^i) = m_j, j = 1, 2, ..., N\right] \qquad (3) \\ &= \sum_{\sum_{j=1}^N m_j = X_i^i} \frac{X_i^i!}{\prod_{j=1}^N m_j!} \prod_{j=1}^N (P_{ij} z_j)^{m_j} = \left(\sum_{j=1}^N P_{ij} z_j\right)^{X_i^i} \end{aligned}$$

Thus,

$$G_{i+1}(\underline{z}) = G_i(z_1, z_2, ..., z_{i-1}, \sum_{j=1}^N P_{ij} z_j, z_{i+1}, ..., z_N) \tag{4}$$

The first moments

Let $f_i^j = E(X_i^j)$. Then by differentiating (4) with respect to z_j and evaluating at $\underline{z} = 1$, we derive

$$f_{i+1}^j = \left.\frac{\partial G_{i+1}(\underline{z})}{\partial z_j}\right|_{\underline{z}=\underline{1}} = \begin{cases} f_i^j + P_{ij} f_i^i & i \neq j \\ \\ P_{ij} f_i^i & i = j \end{cases} \tag{5}$$

Result (5) simply states that the expected number of customers at queue j ($j \neq i$), when queue $i+1$ is polled, equals the expected number of customers at j when queue i was polled, plus the expected number of customers routed from i to j during the visit of the server at queue i (Q_i). For $i = j$, the term f_i^i disappears in the expression for f_{i+1}^i, as all f_i^i customers are being served during the visit of queue i.

Summation (over j) of equations (5) yields

$$\sum_{j=1}^{N} f_{i+1}^j = \sum_{\substack{j=1 \\ j \neq i}}^{N} f_i^j + \sum_{j=1}^{N} P_{ij} f_i^i = \sum_{j=1}^{N} f_i^j, \qquad i = 1, 2, ..., N \tag{6}$$

Together with the condition

$$\sum_{j=1}^{N} f_i^j = M, \qquad i = 1, 2, ..., N$$

a solution for f_i^i can be derived as follows. Using (5) we rewrite (for $i = 1$):

$$\begin{aligned} f_2^1 &= P_{11} f_1^1 \\ f_3^1 &= f_2^1 + P_{21} f_2^2 = P_{11} f_1^1 + P_{21} f_2^2 = \sum_{k=1}^{2} P_{k1} f_k^k \\ &\;. \\ &\;. \\ &\;. \\ f_N^1 &= f_{N-1}^1 + P_{N-1,1} f_{N-1}^{N-1} = \sum_{k=1}^{N-1} P_{k1} f_k^k \\ f_1^1 &= f_N^1 + P_{N,1} f_N^N = \sum_{k=1}^{N} P_{k1} f_k^k \end{aligned} \tag{7}$$

In a similar manner,

$$f_i^i = \sum_{k=1}^{N} P_{ki} f_k^k, \qquad i = 1, 2, ..., N \tag{8}$$

Equation (8) indicates that when the server enters queue i, he finds there all those customers that were routed to that queue since the server's last polling instant. In matrix notation, with $\underline{f} = (f_1^1, f_2^2, ..., f_N^N)$, equation (8) can be rewritten as

$$\underline{f} = \underline{f}P, \quad \text{or} \quad \underline{f}(I - P) = \underline{0}.$$

Consider now the stationary probability vector $\underline{\pi} = (\pi_1, \pi_2 ...., \pi_N)$ of the transition matrix P. π is given as the unique solution of $\underline{\pi} = \underline{\pi}P$, $\sum_{j=1}^{N} \pi_j = 1$. That is, $\underline{f}$ admits a multiplicative solution of $\underline{\pi}$,

$$\underline{f} = c\underline{\pi} \qquad (f_i^i = c\pi_i). \tag{9}$$

For a complete knowledge of the f_i^i, it is just left to calculate c. From (5),

$$\begin{aligned}
f_1^N &= P_{NN} f_N^N \\
&\;\;. \\
&\;\;. \\
&\;\;. \\
f_1^j &= \sum_{k=j}^{N} P_{kj} f_k^k \qquad j = N-1, N-2, ..., 2 \\
&\;\;. \\
&\;\;. \\
&\;\;. \\
f_1^1 &= \sum_{k=1}^{N} P_{k1} f_k^k
\end{aligned} \tag{10}$$

Summing over j in (10), interchanging the order of summation and substituting $f_k^k = c\pi_k$, yields:

$$M = \sum_{j=1}^{N} f_1^j = \sum_{j=1}^{N} \sum_{k=j}^{N} P_{kj} f_k^k = \sum_{k=1}^{N} f_k^k \sum_{j=1}^{k} P_{kj} = c \sum_{k=1}^{N} \pi_k \sum_{j=1}^{k} P_{kj}$$

From which,

$$c = \frac{M}{\sum_{k=1}^{N} \pi_k \sum_{j=1}^{k} P_{kj}} \tag{11}$$

It follows from (9),(10) and (11) that

$$f_1^l = \frac{\sum_{k=l}^N P_{kl}\pi_k}{\sum_{k=1}^N \pi_k \sum_{j=1}^k P_{kj}} M$$

Similarly, we write, for $i = 1, 2, ..., N$,

$$f_i^l = \frac{\sum_{k=l}^{i-1} P_{kl}\pi_k}{\sum_{k=1}^N \pi_k \sum_{j=1}^k P_{kj}} M \qquad (l = 1, 2, ..., N) \tag{12}$$

where we use a cyclic summation implying that, if $i \leq l$, then $\sum_{k=l}^{i-1} a_k = \sum_{k=l}^N a_k + \sum_{k=1}^{i-1} a_k$ for any real numbers $a_1, ..., a_N$.

The expected cycle time $E[C]$ is readily given by

$$E[C] = \sum_{i=1}^N b_i f_i^i + \sum_{i=1}^N d_i = c \sum_{i=1}^N \pi_i b_i + d. \tag{13}$$

Finally, we denote by Λ the throughput of the system, which is the rate at which customers are served (or rerouted). It is given by

$$\Lambda^{Gated} = \frac{\sum_{k=1}^N f_k^k}{E[C]} = \frac{c}{E[C]}.$$

This follows from the fact that the expected number of customers served during a cycle is $\sum_{k=1}^N f_k^k$.

Below are two examples where $\underline{f}$ and $E[C]$ are calculated.

Example 1: Suppose $P_{i,i+1} = 1$ for $i < N$ and $P_{N1} = 1$. That is, all customers in a given queue are routed (after being served) to the next queue. This results in $\pi_i = 1/N, \ i = 1, 2, ..., N$ (note that since the Markov Chain induced by P is periodic, π is not equal to the limiting distribution of the Markov Chain, since such a limit distribution does not exist). Substituting in (11) yields

$$c = \frac{M}{\sum_{k=1}^N \left(\frac{1}{N}\right) \sum_{j=1}^k P_{kj}} = \frac{M}{\frac{1}{N}\cdot 0 + \frac{1}{N}\cdot 0 + ... + \frac{1}{N}P_{N1}} = MN.$$

Thus, $f_i^i = c\pi_i = M$. That is, eventually all M customers will accumulate in one queue and then be routed from one queue to the next. Now,

$$E[C] = MN\left(\sum_{i=1}^N \frac{1}{N} b_i\right) + d = \sum_{i=1}^N (Mb_i) + d,$$

since in each queue the server serves all M customers. The throughput for this example is

$$\Lambda_1^{Gated} = \frac{MN}{M\sum_{i=1}^{N} b_i + d}.$$

Example 2: Suppose $P_{ij} = 1/N$ for every i, j. That is, customers are routed uniformly out of each queue, so that (here too) $\pi_i = 1/N$, $i = 1, 2, ..., N$ (this time P is aperiodic, and therefore π equals to the limiting distribution). Now,

$$\sum_{k=1}^{N} \pi_k \sum_{j=1}^{k} P_{kj} = \frac{1}{N} \sum_{k=1}^{N} \left(\sum_{j=1}^{k} P_{kj} \right) = \frac{1}{N} \sum_{k=1}^{N} \frac{k}{N} = \frac{N+1}{2N}.$$

Thus, $c = \frac{2N}{N+1}M$, and

$$f_i^i = c\pi_i = \frac{2M}{N+1}, \qquad i = 1, 2, ..., N. \tag{14}$$

Using (13), we get

$$E[C] = \frac{2M}{N+1} \left(\sum_{i=1}^{N} b_i \right) + d.$$

To illustrate result (14), consider first the case $N = 1$. Then clearly, $f_1^1 = M$. For $M = 2$, $f_i^i = \frac{2}{3}M$, $i = 1, 2$. Indeed, when the server polls queue 1 he finds there (on the average) $2M/3$ customers. At that instant $f_1^2 = M/3$. When the server leaves queue 1 and moves to queue 2, the expected number of customers he finds there is

$$f_2^2 = f_1^2 + P_{12} f_1^1 = \frac{M}{3} + \frac{1}{2}\left(\frac{2}{3}M\right) = \frac{2}{3}M.$$

For $N = 3$, $P_{ij} = 1/3$, and $f_i^i = \frac{2M}{3+1} = M/2$. Thus, if $f_1^1 = M/2$ then $f_1^2 = M/3$ and $f_1^3 = M/6$. Indeed, from (10) we get

$$\begin{aligned} f_1^2 &= P_{22} f_2^2 + P_{32} f_3^3 = \frac{1}{3}\frac{M}{2} + \frac{1}{3}\frac{M}{2} = \frac{M}{3} \\ f_1^3 &= P_{33} f_3^3 = \frac{1}{3}\frac{M}{2} = \frac{M}{6} \end{aligned}$$

As for the throughput, we have

$$\Lambda_2^{Gated} = \frac{MN}{M\sum_{i=1}^{N} b_i + \frac{d(N+1)}{2}} < \Lambda_1^{Gated} \qquad \text{(for } N \geq 2\text{)}$$

Explicit Expression for $G_i(\underline{z})$

We now derive explicit expressions for $G_i(\underline{z})$. Suppose that there is a single customer in the system (i.e. $M = 1$). For such a case, let $\hat{\pi}_l(i)$ denote the steady state probability that the single customer resides in channel l when channel i is polled. Then, by setting $M = 1$ in equation (12) we have

$$\hat{\pi}_l(i) = \frac{\sum_{k=l}^{i-1} P_{kl}\pi_k}{\sum_{k=1}^{N} \pi_k \sum_{j=1}^{k} P_{kj}}$$

This follows because, when $M = 1$, $f_i^l = E(X_i^l) = \hat{\pi}_l(i) = P\{X_i^l = 1, X_i^j = 0 \ for \ j \neq l\}$. Thus, for $M = 1$, $G_i(\underline{z}) = \sum_{j=1}^{N} \hat{\pi}_j(i) z_j$.

We now make use of the following observation by Resing [5]. Consider the state of the original system (arbitrary M) at a polling instant of channel i. Then the location of each individual customer is independent of the location of the other $M - 1$ customers, and its distribution can be obtained by considering the system with $M = 1$. Using the above it readily follows that for an arbitrary M, the PGF is given by

$$G_i(\underline{z}) = \left(\sum_{j=1}^{N} \hat{\pi}_j(i) z_j \right)^M \tag{15}$$

Remark: $\hat{\pi}_j(i)$ can be obtained directly by using the following approach. For each $i = 1, 2, ..., N$ consider an $N \times N$ probability transition matrix $P(i)$ constructed from the unit matrix I by replacing its ith row by $(P_{i1}, P_{i2}, ..., P_{iN})$, the ith row of the original routing matrix P. Let $\hat{P}(i) = P(i)P(i+1)....P(N)P(1)...P(i-1)$. $\hat{P}_{kj}(i)$, the (k,j)th element of the matrix $\hat{P}(i)$, has the following interpretation: If a customer is found in channel k when channel i is polled, then, with probability $\hat{P}_{kj}(i)$, it will be found in channel j at the next time that channel i is polled. Now, the probability vector $\underline{\hat{\pi}}(i)$ is the solution of the set of equations $\underline{\hat{\pi}}(i) = \underline{\hat{\pi}}(i)\hat{P}(i), \quad \sum_{j=1}^{N} \hat{\pi}_j(i) = 1$.

Second Moments

Let

$$f_i(j,k) \overset{\text{def}}{=} \left. \frac{\partial^2 G_i(\underline{z})}{\partial z_j \partial z_k} \right|_{\underline{z}=\underline{1}} .$$

Then, for $j = k$, $f_i(j,k) = f_i(j,j) = E[X_i^j(X_i^j - 1)]$. Otherwize, $f_i(j,k) = E[X_i^j X_i^k]$. From (15) it follows that

$$f_i(j,k) = M(M-1)\hat{\pi}_j(i)\hat{\pi}_k(i)$$

Number of customers at an arbitrary moment

Let $F^*(\underline{z})$ be the PGF of the number of customers at different queues at an arbitrary moment (in the stationary regime). As in Sidi et. al. [6] (eq. (3.10)) $F^*(\underline{z})$ can be expressed as

$$F^*(\underline{z}) = \frac{1}{E[C]} \left[\sum_{i=1}^{N} [f_i^i b_i F^*(\underline{z}| \text{ service period i}) + d_i F^*(\underline{z}| \text{ switch-over period i})] \right] . \tag{16}$$

In order to obtain $F^*(\underline{z})$ we introduce the PGFs of the following quantities in stationary regime. Denote the number of customers at an arbitrary switch-over instant, service beginning instant and service completion instant at queue i by $\bar{F}_i(\underline{z})$, $V_i(\underline{z})$ and $\bar{V}_i(\underline{z})$ respectively. Following Eisenberg [4] and Sidi et. al. [6], we have

$$G_i(\underline{z}) + f_i^i \bar{V}_i(\underline{z}) = \bar{F}_i(\underline{z}) + f_i^i V_i(\underline{z}) \tag{17}$$

$$\bar{V}_i(\underline{z}) = V_i(\underline{z}) \sum_{j=1}^{N} P_{ij} z_j / z_i \tag{18}$$

$$G_i(\underline{z}) = \bar{F}_{i-1}(\underline{z}) \tag{19}$$

(18) follows from the fact that at the end of a service there is one customer less in the queue being served, but one more in the queue to which this customer was routed (see also Sidi et. al. [6] eq. (3.6)). We obtain

$$F^*(\underline{z}| \text{ switch-over period i}) = G_{i+1}(\underline{z}) \tag{20}$$

$$F^*(\underline{z}| \text{ service period i}) = V_i(\underline{z}) = \frac{[G_{i+1}(\underline{z}) - G_i(\underline{z})] z_i / f_i^i}{\sum_{j=1}^{N} P_{ij} z_j - z_i} \tag{21}$$

(where the first equality in (21) follows since there are no arrivals to any queue during a service period). Substituting in (16) we get

$$F^*(\underline{z}) = \frac{1}{E[C]}\left[\sum_{i=1}^{N}\left\{b_i z_i \frac{[G_{i+1}(\underline{z}) - G_i(\underline{z})]}{\sum_{j=1}^{N} P_{ij} z_j - z_i} + d_i G_{i+1}(\underline{z})\right\}\right]. \tag{22}$$

The moments of the queues' length at an arbitrary moment can now be obtained by differentiating $F^*(\underline{z})$.

Sojourn times, throughputs to each queue and the system's utilization

Let T_i be the expected sojourn time of an arbitrary customer at queue i (total time from arrival to departure). Let L_i be the expected length of queue i at an arbitrary moment at the stationary regime (it is obtained by differentiating (22)). Let λ_i be the throughput to queue i, i.e. the rate at which customers arrive to that queue (from other queues). By Little's Law, we have $E[T_i] = E[L_i]/\lambda_i$. It thus remains to obtain λ_i. The latter equals the expected total number of customers arriving to queue i during a cycle divided by the expected cycle duration, i.e.

$$\lambda_i^{Gated} = \frac{\sum_{k=1}^{N} f_k^k P_{ki}}{E[C]} = \frac{c\sum_{k=1}^{N} \pi_k P_{ki}}{E[C]} = \frac{c\pi_i}{E[C]} = \pi_i \Lambda^{Gated}.$$

Hence

$$E[T_i] = \frac{E[L_i]E[C]}{c\pi_i}.$$

By Little's law, the expected sojourn time of an arbitrary customer at an arbitrary queue is

$$E[T] = \frac{M}{\Lambda^{Gated}} = \frac{ME[C]}{c} = E[C]\sum_{k=1}^{N}\pi_k \sum_{j=1}^{k} P_{kj}$$

The utilization of the system is given by

$$\rho^{Gated} = \sum_{i=1}^{N} \lambda_i^{Gated} b_i = \frac{c}{E[C]}\sum_{i=1}^{M}\pi_i b_i = \Lambda^{Gated}\sum_{i=1}^{M}\pi_i b_i,$$

where ρ is the proportion of time that the server is busy. Using (13), we have, as can be expected from general balance arguments (see [1]),

$$E[C] = \frac{d}{1-\rho^{Gated}}.$$

4 Mixture of Gated and Exhaustive regimes

We analyze in this section the case where some queues are served exhaustively, while the others are served according to the Gated discipline. A special pleasant feature of the closed polling system is that most of the performance measures for this mixed regime can be obtained as a special case of the Gated regime, by a simple transformation of the transition probabilities P. Indeed, assume that there is a set of stations E which are served according to the Exhaustive service, whereas the other stations are served according to the Gated discipline. Consider the following new matrix $\tilde{P}$ of transition probabilities:

$$\tilde{P}_{ij} = \begin{cases} \frac{P_{ij}}{1-P_{ii}} & j \neq i, i \in E \\ 0 & j = i, i \in E \\ P_{ij} & \quad i \notin E \end{cases}$$

(This holds for $N > 1$ and $P_{ii} < 1$. When $N = 1$ and the single station is served exhaustively, then trivially, $X_1^1 = M$.) It follows that the PGF and the distribution of the cycle duration of the mixed discipline polling system are the same as the PGF and the distribution of the cycle duration when the Gated policy is used in all queues with transition probabilities $\tilde{P}$.

As an illustration, consider Example 2 from Section 3, and assume now that all stations are served exhaustively. Then,

$$\tilde{P}_{ij} = \begin{cases} \frac{1/N}{1-1/N} = \frac{1}{N-1} & j \neq i \\ 0 & j = i \end{cases}$$

Hence,

$$\pi_i = 1/N, \ i = 1, ..., N, \qquad c = \frac{M}{\sum_{k=1}^{N} N^{-1} \sum_{j=1}^{k} \tilde{P}_{kj}} = 2M, \qquad f_i^i = \frac{2M}{N}.$$

For $N = 2$, $f_i^i = M$, $i = 1, 2$. (A queue is left only when it is empty, whereupon all customers are waiting in the other queue). For $N = 3$, $f_1^1 = 2M/3$, $f_1^2 = M/3$ and $f_1^3 = 0$. It is easily seen that for $N \geq 2$ we get the same values f_1^j, $j = 1, ..., N-1$ when the Exhaustive policy is used, as we would get if there were only $N - 1$ stations and the Gated policy were used in all queues.

5 The Globally-Gated Regime

The (cyclic) Globally Gated (GG) service regime, recently introduced by Boxma, Levy and Yechiali [3], possesses two attractive properties. (i) it brings the polling system closer to the (fair) First Come First Served discipline (as opposed to the regular Gated or Exhaustive disciplines), and (ii) it enables one to obtain closed form results for cycle time, moments of waiting times and other performance measures. Under this service policy, <u>all</u> gates are closed (globally) at the moment the server polls queue 1, and during the coming cycle only those customers "captured" (present) at Q_j at the start of the cycle, will be served at Q_j.

The probability generating function

Let $G^n(\underline{z})$ be the PGF of the state of the system (number of customers) $\underline{X}_1(n) = (X_1^1(n), X_1^2(n), ..., X_1^N(n))$ at the start of the nth cycle. Then

$$\begin{aligned}
G^{n+1}(\underline{z}|\underline{X}_1(n)) &= E\left[\prod_{j=1}^{N} z_j^{X_1^j(n+1)}\middle|\underline{X}_1(n)\right] = E\left[\prod_{j=1}^{N} z_j^{\sum_{i=1}^{N} A_i^j(X_1^i(n))}\middle|\underline{X}_1(n)\right] \\
&= E\left[\prod_{j=1}^{N}\left(\prod_{i=1}^{N} z_j^{A_i^j(X_1^i(n))}\right)\middle|\underline{X}_1(n)\right] = E\left[\prod_{i=1}^{N}\prod_{j=1}^{N} z_j^{A_i^j(X_1^i(n))}\middle|\underline{X}_1(n)\right] \\
&= \prod_{i=1}^{N} E\left(\prod_{j=1}^{N} z_j^{A_i^j(X_1^i(n))}\middle|\underline{X}_1(n)\right) = \prod_{i=1}^{N}\left(\sum_{j=1}^{N} P_{ij} z_j\right)^{X_1^i(n)}
\end{aligned}$$

where the last equality follows from the same arguments as (3). Hence we get

$$G^{n+1}(\underline{z}) = G^n\left(\sum_{j=1}^{N} P_{1j}z_j, \sum_{j=1}^{N} P_{2j}z_j, ..., \sum_{j=1}^{N} P_{Nj}z_j\right).$$

In a stationary regime $G(\underline{z}) = G^n(\underline{z}) = G^{n+1}(\underline{z})$ (and if a limiting distribution exists, then $G(\underline{z}) = \lim_{n\to\infty} G^n(\underline{z})$). Hence, we have

$$G(\underline{z}) = G(\underline{s}(\underline{z})) \tag{23}$$

where $\underline{s}(\underline{z}) = (s_1(\underline{z}), s_2(\underline{z}), ..., s_N(\underline{z}))$ with $s_j(\underline{z}) = \sum_{k=1}^{N} P_{jk}z_k$.

Next we express the LST of the cycle duration in terms of $G(\bullet)$. We have

$$E\left[e^{-\omega C}\middle| X_1^1, X_1^2, ..., X_1^N\right] = E\left[e^{-\left\{\omega \sum_{j=1}^{N}\left(\sum_{k=1}^{X_1^j} B_{jk} + D_j\right)\right\}}\middle| X_1^1, X_1^2, ..., X_1^N\right]$$

$$= \prod_{j=1}^{N}[b_j^*(\omega)]^{X_1^j}\prod_{j=1}^{N} d_j^*(\omega)$$

where, for every $j = 1, 2, ..., N$, B_{jk} are i.i.d. and are distributed like B_j. Thus

$$c^*(\omega) = E\left[e^{-\omega C}\right] = E\left\{E\left[e^{-\omega C}\middle|\underline{X}_1\right]\right\} \tag{24}$$

$$= G(b_1^*(\omega), b_2^*(\omega), ..., b_N^*(\omega))\prod_{j=1}^{N} d_j^*(\omega)$$

First moments

By differentiating (23) we get

$$f_1^i = \sum_{j=1}^{N}\frac{\partial G(\underline{s}(\underline{z}))}{\partial s_j(\underline{z})}\frac{\partial s_j(\underline{z})}{\partial z_i}\Bigg|_{\underline{z}=1} = \sum_{j=1}^{N} f_1^j P_{ji} \tag{25}$$

so that, $\underline{f}_1 = \underline{f}_1 P$. Together with $\sum_{k=1}^{N} f_1^k = M$, the solution for $\underline{f}_1$ is $f_1^i = c_{GG}\pi_i$, $i = 1, 2, ..., N$, with $c_{GG} = M$. Thus,

$$f_1^i = M\pi_i. \tag{26}$$

The mean cycle time is given by

$$E[C] = \sum_{i=1}^{N} f_1^i b_i + d = M\sum_{i=1}^{N}\pi_i b_i + d \tag{27}$$

(the first equality can be obtained by differentiating (24)). It should be noted that in the GG case, f_i^i is the expected number of customers present at Q_i when the server polls this queue, but only f_1^i of which are served in the current cycle. We have,

$$f_i^i = f_1^i + \sum_{k=1}^{i-1} f_1^k P_{ki} = M\left[\pi_i + \sum_{k=1}^{i-1}\pi_k P_{ki}\right]. \tag{28}$$

Next, we express the throughput of the system. As the expected number of customers served during a cycle is $\sum_{k=1}^{N} f_1^k = M$, we readily have,

$$\Lambda^{GG} = \frac{M}{E[C]} = \frac{M}{M\sum_{k=1}^{N}\pi_k b_k + d}$$

For Example 1 of Section 3 ($P_{i,i+1} = 1$), as well as for Example 2 ($P_{i,j} = 1/N$), $\pi_i = 1/N$, and thus,

$$f_1^i = \frac{M}{N}, \qquad E[C] = \sum_{i=1}^{N} \frac{M}{N} b_i + d.$$

However, for Example 1, by using (28), $f_i^i = M/N + P_{i-1,i}(M/N) = 2M/N$ and $\sum_{i=1}^{N} f_i^i = 2M$, whereas for Example 2, $f_i^i = M/N + (M/N)([i-1]/N) = M(N+i-1)/N^2$, and $\sum_{i=1}^{N} f_i^i = (M/N)([3N-1]/2)$. The difference is that in Example 1 all customers at a given queue are routed to the next queue where they will be served in the next cycle, whereas in Example 2, customers are routed evenly to all queues.

In general, comparing throughputs, it follows that the throughput under the Gated discipline, Λ^{Gated}, is larger than that under the Globally-Gated discipline, Λ^{GG}. Indeed, from Eq. (11) and (13),

$$\Lambda^{Gated} = \frac{c}{E[C]} = \frac{M}{\sum_{k=1}^{N} \pi_k [M b_k + d \sum_{j=1}^{k} P_{kj}]} \geq \frac{M}{\sum_{k=1}^{N} \pi_k [M b_k + d]} = \Lambda^{GG}.$$

Explicit Expression for $G(\underline{z})$

Consider our closed system in steady state at the start of a cycle. Then following Resing's observation [5], the locations of the M different customers are independent and identically distributed, and their distributions are independent of M. Thus, suppose $M = 1$. Then $\pi_j = P(X_1^j = 1, X_1^i = 0 \ for \ all \ i \neq j)$, and

$$\begin{aligned} G(\underline{z})|_{M=1} &= \sum_{k=1}^{N} E\left[\prod_{j=1}^{N} z_j^{X_1^j} \,\middle|\, X_1^k = 1; X_1^i = 0, \text{ for all } i \neq k \right] P(X_1^k = 1; X_1^i = 0, \text{ for all } i \neq k) \\ &= \sum_{k=1}^{N} z_k \pi_k \end{aligned}$$

Then for arbitrary M,

$$G(\underline{z}) = \left[\sum_{k=1}^{N} z_k \pi_k \right]^M \tag{29}$$

It is now easy to check that (29) satisfies equation (23). Also, result (26) can be readily derived by differentiating (29).

Second Moments

By differentiating (29) twice, we obtain

$$f_1(j,k) = E(X_1^j X_1^k) = M(M-1)\pi_j \pi_k \tag{30}$$

To express the second moment of the cycle time we use (24) and (29) and write

$$c^*(\omega) = \left[\sum_{k=1}^{N} b_k^*(\omega)\pi_k\right]^M \left[\prod_{j=1}^{N} d_j^*(\omega)\right] \tag{31}$$

(from which equation (27) can readily be verified). By differentiating (31) twice we obtain

$$E[C^2] = \left.\frac{\partial^2 c^*(\omega)}{\partial \omega^2}\right|_{\omega=0} = 2d\sum_{j=1}^{N} f_1^j b_j + \sum_{j=1}^{N}\sum_{k=1}^{N} b_j b_k f_1(j,k) + \sum_{j=1}^{N} f_1^j b_j^{(2)} + \sum_{j=1}^{N} d_j^{(2)} + \sum_{j=1}^{N} d_j \sum_{\substack{k=1 \\ k\neq j}}^{N} d_k.$$

Number of customers at an arbitrary moment

Let $F^*(\underline{z})$, $\bar{F}_i(\underline{z})$, $V_i(\underline{z})$ and $\bar{V}_i(\underline{z})$ be as in Section 3. (16) is used again to evaluate $F^*(\underline{z})$. We obtain exactly the same expression for $F^*(\underline{z})$ as for the Gated discipline, i.e. Eq. (22). In fact Eq. (18), (19), (20) and (22) remain unchanged, whereas in (16), (17) and (21) f_i^i should be replaced by f_1^i. The moments of the queues' length at an arbitrary moment can now be obtained by differentiating $F^*(\underline{z})$.

Sojourn times, throughputs to individual queues and the utilization

With the same notation as in Section 3 we have

$$\lambda_i^{GG} = \frac{\sum_{k=1}^{N} f_1^k P_{ki}}{E[C]} = \frac{M\sum_{k=1}^{N} \pi_k P_{ki}}{E[C]} = \frac{M\pi_i}{E[C]}$$

$$E[T_i] = \frac{E[L_i]}{\lambda_i^{GG}} = \frac{E[L_i]E[C]}{M\pi_i}$$

Finally, by Little's law, the expected sojourn time of an arbitrary customer is

$$E[T] = \frac{M}{\Lambda^{GG}} = \frac{ME[C]}{M} = E[C]$$

The utilization is given by

$$\rho^{GG} = \sum_{i=1}^{N} \lambda_i^{GG} b_i = \sum_{i=1}^{N} \frac{M\pi_i b_i}{E[C]} = \Lambda^{GG} \sum_{i=1}^{N} \pi_i b_i$$

which can also be derived from (27). Again, similarly to the Gated case,

$$E[C] = \frac{d}{1-\rho^{GG}}.$$

Dynamic Optimization Under the GG Regime

We investigate a dynamic optimization problem, where the objective is to minimize the long-run expected weighted holding costs in the different queues. We relax the constraint of a fixed cylic movement of the server and allow it to select a new Hamiltonian tour at the beginning of each new cycle. Under such Hamiltonian polling scheme, switching times between any pair of stations depend on both stations and not only on the station which is left by the server. We consider a star configuration [2] where it takes S_j units of time ($ES_j = s_j$) to switch into station j and D_j units of time ($ED_j = d_j$) to switch out of the station. A key feature to this dynamic optimization problem is the fact that under the Globally Gated regime, both the Hamiltonian tour duration and the number of customers at the various queues at the end of a cycle (tour) are independent of the order of visits of the various queues. Both quantities are in fact functions of the state of the system $\underline{X}_1 = (X_1^1, X_1^2, ..., X_1^N)$ at the start of a cycle. It thus follows that minimizing expected total weighted holding costs incurred during each cycle individually, minimizes the long-run expected weighted cost. We assume that each customer pays a holding cost of c_i per time unit when in queue i. The total expected cost Z during a cycle (starting with $\underline{X}_1$) is composed of two components: (i) the total (weighted) costs incurred by all M customers in each and every queue before they are served and routed, and (ii) the total weighted costs incurred by all customers after their routing. Thus, following a Hamiltonian tour that visits the channels in the order $1, 2, ..., N$, we have

$$\begin{aligned} E[Z|\underline{X}_1] &= \sum_{k=1}^{N} c_k \left[X_1^k \left\{ \sum_{j=1}^{k-1} (s_j + X_1^j b_j + d_j) + s_k \right\} + \sum_{m=1}^{X_1^k} m b_k \right] \qquad (32) \\ &+ \sum_{k=1}^{N} \left[\sum_{i=1}^{N} P_{ki} c_i \left\{ \sum_{m=1}^{X_1^k} \left(E[C|\underline{X}_1] - \left[\sum_{j=1}^{k-1} (s_j + X_1^j b_j + d_j) + s_k \right] - m b_k \right) \right\} \right] \end{aligned}$$

The first term of (32) results from the fact that in queue k, all X_1^k customers first wait for the server to arrive and switch in, which occurs at time $t = \sum_{j=1}^{k-1} (s_j + X_1^j b_j + d_j) + s_k$, and then the mth customer (out of X_1^k) waits an expected $m b_k$ units of time until he is routed (upon his service completion) to some other station. As for the second term of (32), observe that a customer of queue k is routed to queue i with probability P_{ki}, and then the above mth customer resides in queue i for the remaining of the cycle, namely for $E[C|\underline{X}_1]-\left[\sum_{j=1}^{k-1}(s_j+X_1^jb_j+d_j)+s_k\right]-mb_k$ units of time (in expectation), incurring cost at a rate c_i. Thus,

$$\begin{aligned} E[Z|\underline{X}_1] &= \sum_{k=1}^{N} X_1^k(c_k-\sum_{i=1}^{N}P_{ki}c_i)\sum_{j=1}^{k-1}(s_j+X_1^jb_j+d_j) \qquad (33)\\ &+ \sum_{k=1}^{N}\left[c_k\sum_{m=1}^{X_1^k}mb_k+\sum_{i=1}^{N}P_{ki}c_i\left\{\sum_{m=1}^{X_1^k}(E[C|\underline{X}_1]-mb_k)\right\}\right]\\ &+ \sum_{k=1}^{N}X_1^ks_k\left(c_k-\sum_{i=1}^{N}P_{ki}c_i\right)\end{aligned}$$

Now, as only the first term in (33) depends on the order of visits, the expected total cost per cycle is minimized by the visit order that minimizes the expression

$$\sum_{k=1}^{N}X_1^k\hat{c}_k\sum_{j=1}^{k-1}(s_j+X_1^jb_j+d_j), \quad \text{where} \quad \hat{c}_k=c_k-\sum_{i=1}^{N}P_{ki}c_i. \qquad (34)$$

If we partition the N queues into three sets: $\{+\}\overset{\text{def}}{=}\{k:\hat{c}_k>0\}$, $\{0\}\overset{\text{def}}{=}\{k:\hat{c}_k=0\}$ and $\{-\}\overset{\text{def}}{=}\{k:\hat{c}_k<0\}$, (clearly $\{+\}\cup\{0\}\cup\{-\}=\{1,2,...,N\}$), then the optimal visit order is given by a $c\mu$-type rule as follows: arrange all queues belonging to the $\{+\}$ set in an increasing order of the index $(s_j+X_1^jb_j+d_j)/(X_1^j\hat{c}_j)$. Then arrange the queues comprising the $\{0\}$ set in any arbitrary order, and finally arrange the remaining queues (belonging to the $\{-\}$ set) also in an increasing order of $(s_j+X_1^jb_j+d_j)/(X_1^j\hat{c}_j)$. The optimal Hamiltonian tour is to first visit the queues of $\{+\}$ set (in the order specified above), then to visit the $\{0\}$-queues (in any order), and finally to visit the queues comprising the set $\{-\}$ following the above index rule. The proof of the above follows by an extension of a standard interchange argument, usually applied to the $\{+\}$ set only (see e.g. Boxma, Levy and Yechiali [3]) to the current case where all three sets $\{+\},\{0\}$ and $\{-\}$ may be non empty.

References

[1] E. Altman, P. Konstantopoulos and Z. Liu, "Stability, Monotonicity and Invariant Quantities in General Polling Systems", *Queuing Systems* **11** (special issue on Polling Models, Eds. H. Takagi and O. Boxma), pp. 35-57, 1992.

[2] D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, 1987.

[3] O. J. Boxma, H. Levy and U. Yechiali, "Cyclic Reservation Schemes for Efficient Operation of Multiple-Queue Single-Server Systems", *Annals of Operations Research* **35**, pp. 187-208, 1992.

[4] M. Eisenberg, "Queues with Periodic Service and Changeover Time", *Operations Research* **20**, pp. 440-451, 1972.

[5] J. Resing, Private Communication, 1993.

[6] M. Sidi, H. Levy and S. W. Fuhrmann, "A queueing network with a single cyclically roving server", *Queuing Systems* **11**, (special issue on Polling Models, Eds. H. Takagi and O. Boxma), pp.121-144, 1992.

[7] H. Takagi, *Analysis of Polling Systems*, The MIT Press, 1986.

[8] H. Takagi, "Queueing Analysis of Polling Models: An Update", in *Stochastic Analysis of Computer and Communication Systems* (H. Takagi, Ed.) pp. 267-318, North Holland, 1990.

[9] U. Yechiali, "Analysis and Control of Polling Systems", in *Performance Evaluation of Computer and Communications Systems* (L. Donantiello and R. Nelson, Eds.) pp. 630-650, Springer-Verlag, 1993.