1992

O.J. Boxma, J.A. Weststrate, U. Yechiali

A globally gated polling system with server interruptions, and applications to the repairman problem

Department of Operations Research, Statistics, and System Theory Report BS-R9208 April

A Globally Gated Polling System with Server Interruptions, and Applications to the Repairman Problem

O.J. Boxma
CWI
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands;
Faculty of Economics, Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

J.A. Weststrate
Faculty of Economics, Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

U. Yechiali
Department of Statistics, School of Mathematical Sciences
The Raymond and Beverly Sackler Faculty of Exact Sciences
Tel-Aviv University, Tel-Aviv 69978, Israel

A repair crew is responsible for the maintenance and operation of *N* installations. The crew has to perform a collection of preventive maintenance tasks at the various installations. The installations may break down from time to time, generating corrective maintenance requests which have priority over the preventive maintenance tasks. We formulate and analyze this real-world problem as a single-server multi-queue polling model with Globally Gated service discipline and with server interruptions. We derive closed-form expressions for the Laplace-Stieltjes Transform and the first moment of the waiting time distributions of the preventive and corrective maintenance requests at the various installations, and obtain simple and easily implementable static and dynamic rules for optimal operation of the system. We further show that, for the so-called elevator-type polling scheme, mean waiting times of preventive maintenance jobs at all installations are *equal.*

1985 Mathematics Subject Classification: 60K25, 68M20.
Keywords and Phrases: Repairman problem, polling model, globally gated service, waiting times.
Note: This report will be submitted for publication.

1. Introduction

The standard polling system is a single-server multiple-queue configuration in which the server cyclically moves from queue to queue, providing some service at each non-empty queue according to a local service discipline. A common local service discipline is the *gated* service discipline, viz.: when the server finds *n* customers at a queue upon his arrival, he serves exactly those *n* customers before moving to the next queue. Recently, Boxma, Levy and Yechiali [4] introduced a *global* service discipline for polling systems, the *Globally Gated* (GG) service discipline. According to this discipline the server moves cyclically along the queues, using the instant of cycle-beginning as a Globally Gated reference point of time: when he reaches a queue he serves there only (and all) customers who were present at that queue at the beginning of the cycle. This discipline can easily be implemented by marking all customers with a time stamp denoting their arrival instant. In its nature the GG discipline resembles the regular gated discipline, the difference being that the gating mechanism is applied to all queues at the same time. The GG discipline leads to a simpler mathematical analysis than those for polling systems with regular gated service disciplines - or any other known service

Report BS-R9208
ISSN 0924-0659
CWI
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

discipline; closed-form expressions can be derived for the delay distributions in the various queues. As a result, the operation of the polling system by the GG discipline is easy to control and optimize. The GG discipline is also very 'fair' with respect to FCFS considerations, in the sense that all arrivals in one cycle are served in the next cycle.

In this paper we study a (cyclic) Globally Gated polling system in which the server is subject to interruptions that occur according to a Poisson process (and may form a queue of their own). If an interruption occurs when the server is serving a customer, several interruption-handling disciplines are possible. We consider two such disciplines: (i) the preemptive resume discipline, in which the server abandons its present job immediately upon interruption and resumes serving this job as soon as the interruption is over and no other interruptions are present, and (ii) the non-preemptive discipline in which the server finishes serving the present job before the interruption is handled. If the interruption occurs while the server is switching from one station to another, the server reacts according to the preemptive resume discipline. During an interruption the server is not available for service at any of the stations.

We use this polling system with server interruptions to model and analyze a maintenance process in which the repairman is occupied with both preventive maintenance requests (the customers) and corrective maintenance requests (the interruptions, caused by breakdowns) generated by a group of various installations. Exploiting the fact that the mathematically pleasing properties of the GG discipline are not lost in the case of Poison interruptions, we derive closed-form expressions for the Laplace-Stieltjes Transforms (LSTs) and first moments of the waiting time distributions of both corrective and preventive maintenance requests. Furthermore, we obtain easily implementable optimal rules for static or dynamic control of the system.

The structure of the paper is as follows. In Section 2 we describe the polling model and the associated maintenance process. Section 3 is devoted to cycle time analysis (length of a preventive maintenance tour), and Section 4 to the analysis of queue length distributions at various polling epochs. In Section 5 we derive the LSTs of the waiting time distributions of preventive maintenance requests and of corrective maintenance requests under the preemptive resume and the non-preemptive interruption disciplines. Section 6 is devoted to the derivation of optimal rules (both static and dynamic) for the order in which the repairman visits the installations. It is further shown that in the case of the so-called elevator-type polling scheme, mean waiting times of preventive maintenance jobs at all installations are *equal.*

The analysis in Sections 2-6 concerns the situation in which an idle server keeps switching from queue to queue. Section 7 briefly indicates how one can handle the situation in which the server stays idle at the first installation whenever a cycle terminates with an empty system. Section 8 contains some conclusions and suggestions for further research.

2. The maintenance process and the queueing model

The maintenance process

The maintenance process which we wish to model can be described as follows. There is a number of installations which generate two types of maintenance requests: preventive and corrective. A single repairman (or a crew) is assigned to fulfill the maintenance requirements of all the installations. The preventive maintenance requirements are described by (so-called) maintenance packages. The repairman visits the installations in a cyclic order to perform preventive maintenance as described by the maintenance packages. Naturally, it takes the repairman some time to move from one installation to the next, and to perform administrative duties. At the beginning of each tour the repairman is assigned all preventive maintenance requirements at the various installations and subsequently he will handle only these preventive maintenance requirements during the coming tour. This enables us (see Section 6) to consider control policies in which successive maintenance tours do not have the same visit order of the installations. For each new tour, the repairman can dynamically determine the optimal visit order of the installations - with respect to the mean total waiting cost incurred during the coming cycle - based on the preventive maintenance requirements at each installation at the beginning of the cycle.

The repairman will keep performing the preventive maintenance jobs, and will continue his 'walk' along the installations unless there is a breakdown at one of the installations. In such a case, the breakdown interrupts the regular preventive maintenance tour and the repairman has to move to the broken installation to restore its operation, which has been stopped as a result of the critical failure. If a breakdown occurs during a 'walking-time' of the repairman between the installations, corrective maintenance gets priority with no delay. If a breakdown occurs when the repairman is performing preventive maintenance, then there are two priority policies: (i) corrective maintenance gets priority with no delay (preemptive resume), or (ii) the repairman first finishes his present preventive maintenance job, and only then travels to the broken installation. It is assumed that the time to go to a broken installation and the travel times between the broken installations during an interruption period are incorporated in the repair time. When all installations restore operation, the repairman resumes his preventive maintenance tour from the state at which it was interrupted.

The queueing model

We consider a model consisting of N infinite-buffer queues, $Q_1, \dots, Q_N$, and a single server. Customers arrive at the various queues according to independent Poisson processes. The arrival rate to Q_i is λ_i. The service time distribution of customers at Q_i is $B_i(.)$, with first and second moments denoted by β_i and $\beta_i^{(2)}$ respectively, and with LST $\tilde{B}_i\{.\}$. The offered load to Q_i is $\rho_i = \lambda_i \beta_i$, so that the total system load is $\rho := \sum_{i=1}^{N} \rho_i$. The server moves among the queues in a strictly cyclic order; when leaving Q_i and before entering the next queue the server incurs a switchover period whose duration is a random variable S_i with distribution $S_i(.)$, first and second moment s_i and $s_i^{(2)}$ respectively, and LST $\tilde{S}_i\{.\}$. All switchover periods are assumed to be mutually independent. The total switchover time in a cycle is $S = \sum_{i=1}^{N} S_i$ with first two moments s and $s^{(2)}$, and LST given by $\tilde{S}\{\omega\} = \prod_{i=1}^{N} \tilde{S}_i\{\omega\}$, Re $\omega \geq 0$.

In conjunction with the maintenance process, the server represents the repairman; the queues represent the installations, the customers are the preventive maintenance requests, and the switchover times are the travel times between installations.

The service regime used by the server is the Globally Gated discipline which is being operated as follows: At the cycle-beginning, namely, when the server reaches Q_1, all customers present at $Q_1, \dots, Q_N$ are marked. During the coming cycle (i.e., the visits of queues $Q_1, \dots, Q_N$) the server serves all marked customers (but no others). Customers who meanwhile arrive at the various queues will have to wait until being marked at the next cycle-beginning, and will be served at the next cycle.

The server is interrupted from time to time according to an independent Poisson process with intensity μ. The time duration of a single interruption is a random variable **R** having distribution function $R(.)$ with first and second moment r and $r^{(2)}$ respectively, and LST $\tilde{R}\{.\}$. These single interruptions correspond to the occurence of corrective maintenance requests (breakdowns). Note that interruptions may occur sequentially, resulting in an *interruption period,* **P**, distributed as a busy period in an M/G/1 queue with LST $\tilde{P}\{.\}$, and with mean p and second moment $p^{(2)}$ given by (cf. Cohen [6]: p. 251):

$$p = \frac{r}{1-\mu r}, \tag{2.1}$$

$$p^{(2)} = \frac{r^{(2)}}{(1-\mu r)^3}. \tag{2.2}$$

As usual in the polling literature, it is assumed that the server keeps switching from one queue to another even when there are no customers present in the system. This assumption may not be fully realistic for a repairman in a maintenance system; such a repairman would probably return to his home base (say, Q_1) and wait there for new arrivals (maintenance requests) to occur. However, one can argue that due to the usually high load of maintenance systems, the probability that the repairman finds an empty system at a cycle-beginning is small, so that the results for the present system give useful insight into the behaviour of a system in which the repairman waits idling when there is no work. The case in which the server waits idling at Q_1 when the system is empty is briefly considered in Section 7.

Since at every cycle the server serves all the work that arrived during the previous cycle, the necessary and sufficient condition for ergodicity is $\rho+\mu r<1$; this can be proved in the same way as the ergodicity condition for the ordinary Globally Gated polling system is proved in Boxma, Levy and Yechiali [4].

All arrival, service time and switching processes and the interruption process are independent stochastic processes. We assume that the system is in equilibrium state.

3. The cycle time (length of a maintenance tour)

For the polling model with interruptions that was described in Section 2, we now derive the LST and first two moments of the cycle time of the server; in the next sections, the cycle time distribution will play a crucial role in the analysis of queue lengths and waiting times. First define:

Y_k := length of time during which the server is serving customers from Q_k during one cycle;

V_k := Y_k + the total time the server is busy with interruption periods originating during Y_k.

As a preparation for the cycle time analysis we derive the LST of V_k.

$$\tilde{V}_k\{\omega\} := E\{e^{-\omega \mathrm{V}_k}\} \tag{3.1}$$

$$= \sum_{m=0}^{\infty} E\{e^{-\omega \mathrm{V}_k} \,|\, m \textit{ interruptions during } \mathrm{Y}_k\}\, Pr\{m \textit{ interruptions during } \mathrm{Y}_k\}$$

$$= \sum_{m=0}^{\infty} \int_{t=0}^{\infty} e^{-\omega t}[\tilde{P}\{\omega\}]^m e^{-\mu t}\frac{(\mu t)^m}{m!} d_t Pr\{\mathrm{Y}_k \le t\} = \tilde{Y}_k\{\omega+\mu(1-\tilde{P}\{\omega\})\},$$

with $\tilde{A}\{.\}$ denoting the LST of a random variable A.

Let $\mathrm{X}_1, \ldots, \mathrm{X}_N$ denote the queue lengths at the start of an arbitrary cycle (where a cycle is defined as the time between two successive visits of the server to Q_1). Denoting by C the length of a cycle, and using the reasoning leading to (3.1),

$$E\{e^{-\omega \mathrm{C}} \,|\, \mathrm{X}_1, \ldots, \mathrm{X}_N\} = \tilde{S}\{\omega+\mu(1-\tilde{P}\{\omega\})\} \prod_{i=1}^{N} \tilde{B}_i\{\omega+\mu(1-\tilde{P}\{\omega\})\}^{\mathrm{X}_i}, \quad \mathrm{Re}\ \omega \ge 0. \tag{3.2}$$

In its turn, the length of a cycle determines the joint queue length distribution at the beginning of the next cycle:

$$E\{z_1^{\mathrm{X}_1} \cdots z_N^{\mathrm{X}_N} \,|\, \mathrm{C}=t\} = \exp[-\sum_{i=1}^{N} \lambda_i(1-z_i)t], \quad |z_i| \le 1, \quad i=1,\ldots,N. \tag{3.3}$$

Unconditioning we find, for $|z_i| \le 1$, $i=1,\ldots,N$,

$$E\{z_1^{\mathrm{X}_1} \cdots z_N^{\mathrm{X}_N}\} = \tilde{C}\{\sum_{i=1}^{N} \lambda_i(1-z_i)\}, \tag{3.4}$$

where

$$\tilde{C}\{\omega\} = E\{e^{-\omega \mathrm{C}}\}, \quad \mathrm{Re}\ \omega \ge 0.$$

From (3.2) and (3.4),

$$\tilde{C}\{\omega\} = \tilde{S}\{\omega+\mu(1-\tilde{P}\{\omega\})\}\, E\left[\prod_{i=1}^{N} \tilde{B}_i\{\omega+\mu(1-\tilde{P}\{\omega\})\}^{\mathrm{X}_i}\right] = \tag{3.5}$$

$$= \tilde{S}\{\omega+\mu(1-\tilde{P}\{\omega\})\}\ \tilde{C}\left\{\sum_{i=1}^{N}\lambda_i(1-\tilde{B}_i\{\omega+\mu(1-\tilde{P}\{\omega\})\})\right\}.$$

Now, for Re $\omega\geq 0$, let

$$\phi(\omega) := \omega+\mu(1-\tilde{P}\{\omega\}); \tag{3.6}$$

$$\delta(\omega) := \sum_{i=1}^{N}\lambda_i\left[1-\tilde{B}_i\{\phi(\omega)\}\right].$$

Equation (3.5) can then be written as $\tilde{C}\{\omega\} = \tilde{S}\{\phi(\omega)\}\tilde{C}\{\delta(\omega)\}$. Define recursively:

$$\delta^{(0)}(\omega) := \omega, \tag{3.7}$$

$$\delta^{(n)}(\omega) := \delta(\delta^{(n-1)}(\omega)), \quad n=1,2,\dots .$$

Applying (3.5) iteratively we obtain, for every $M=1,2,\dots$,

$$\tilde{C}\{\omega\} = \tilde{C}\{\delta^{(M)}(\omega)\}\prod_{m=0}^{M-1}\tilde{S}\{\phi(\delta^{(m)}(\omega))\}.$$

As in Boxma, Levy and Yechiali [4] it can be shown that

$$\lim_{M\to\infty}\delta^{(M)}(\omega) = 0,$$

and that the following limit exists:

$$\lim_{M\to\infty}\prod_{m=1}^{M}\tilde{S}\{\phi(\delta^{(m-1)}(\omega))\}.$$

Hence

$$\tilde{C}\{\omega\} = \prod_{m=0}^{\infty}\tilde{S}\{\phi(\delta^{(m)}(\omega))\}, \quad \text{Re } \omega\geq 0. \tag{3.8}$$

Differentiating (3.5) once and twice yields,

$$EC = \rho(1+\mu p)EC + (1+\mu p)s,$$

$$EC^2 = \rho\mu p^{(2)}EC + (1+\mu p)^2[\sum_{j=1}^{N}\lambda_j\beta_j^{(2)}]EC$$
$$+ \rho^2(1+\mu p)^2EC^2 + 2\rho(1+\mu p)^2 sEC + \mu p^{(2)}s + (1+\mu p)^2 s^{(2)}.$$

Substituting (2.1) and (2.2) in the above expressions and reordering terms we obtain the following closed-form expressions for the first two moments of the cycle time:

$$EC = \frac{s}{1-\rho-\mu r}, \tag{3.9}$$

(in particular, if $\mu=0$ then $EC=s/(1-\rho)$ as in Boxma, Levy and Yechiali [4]), and

$$EC^2 = \frac{1}{(1-\mu r)^2-\rho^2}[s^{(2)} + 2\rho s EC + EC\sum_{j=1}^{N}\lambda_j\beta_j^{(2)}] + \frac{\mu r^{(2)}}{(1-\mu r)[(1-\mu r)^2-\rho^2]}[\rho EC + s]. \tag{3.10}$$

Introducing C_p and C_r, the past and residual time, respectively, of a cycle, we can write for Re $\omega_p \geq 0$, Re $\omega_r \geq 0$ (cf. Cohen [6]: p. 113):

$$E[e^{-\omega_p C_p - \omega_r C_r}] = \frac{1}{EC}\,\frac{1}{\omega_p-\omega_r}\,[\tilde{C}\{\omega_r\} - \tilde{C}\{\omega_p\}]. \tag{3.11}$$

It follows in particular that the LST and the mean value of C_p and C_r are:

$$E\{e^{-\omega C_p}\} = E\{e^{-\omega C_r}\} = \frac{1-\tilde{C}\{\omega\}}{\omega EC},$$

$$EC_p = EC_r = \frac{EC^2}{EC} = \frac{1}{1+\rho-\mu r}\left[\frac{s^{(2)}}{2s} + \frac{\rho s}{1-\rho-\mu r} + \frac{1}{2(1-\rho-\mu r)}(\sum_{j=1}^{N}\lambda_j\beta_j^{(2)} + \mu r^{(2)})\right]. \tag{3.12}$$

In maintenance terms, one can view the cycle time as the time of a complete tour of the repairman along all the installations. Thus, (3.8), (3.9) and (3.10) respectively, reflect the LST, the mean and the second moment of the time it takes for the repairman to do the required preventive maintenance at each installation during one cycle and restore the breakdowns occuring during this period.

4. Queue lengths of preventive maintenance requests

In Formula (3.4) we have expressed the generating function of the joint queue length distribution at cycle beginnings (the instants at which the server polls Q_1) in terms of the LST of the cycle time distribution. In the polling literature, the most studied queue length related quantity is the generating function of the joint queue length distribution at epochs that the server polls some queue Q_i. The GG discipline allows us to obtain a much more explicit expression for this generating function than in any other known discipline. To derive such an expression, we first introduce some notation. Y_{ij} denotes the number of customers at Q_j at the instant that the server visits Q_i, $i,j=1,...,N$. $A_i[T]$ denotes the number of Poisson arrivals at Q_i during a period of length T, and $V_i(n)$ denotes the total service time of n customers at Q_i. Restricting ourselves for the moment to the case in which there are no interruptions ($\mu=0$), we can write:

$$E[z_1^{Y_{i1}} z_2^{Y_{i2}} \cdots z_N^{Y_{iN}} \,|\, X_1, \ldots, X_N] \tag{4.1}$$

$$= E[z_1^{A_1[\sum_{n=1}^{i-1}(V_n(X_n)+S_n)]}\; z_2^{A_2[\sum_{n=1}^{i-1}(V_n(X_n)+S_n)]} \cdots z_{i-1}^{A_{i-1}[\sum_{n=1}^{i-1}(V_n(X_n)+S_n)]}$$

$$z_i^{X_i+A_i[\sum_{n=1}^{i-1}(V_n(X_n)+S_n)]} \cdots z_N^{X_N+A_N[\sum_{n=1}^{i-1}(V_n(X_n)+S_n)]} \,|\, X_1, \ldots, X_N]$$

$$= \prod_{j=1}^{i-1} \tilde{B}_j^{X_j}(\sum_{n=1}^{N}\lambda_n(1-z_n)) \prod_{j=1}^{i-1} \tilde{S}_j(\sum_{n=1}^{N}\lambda_n(1-z_n)) \prod_{j=i}^{N} z_j^{X_j}\,.$$

Hence, using (3.4), for $|z_1|\leq 1,...,|z_N|\leq 1$,

$$E[z_1^{Y_{i1}} z_2^{Y_{i2}} \cdots z_N^{Y_{iN}}] = \prod_{j=1}^{i-1} \tilde{S}_j(\sum_{n=1}^{N}\lambda_n(1-z_n))\, \tilde{C}\left[\sum_{j=1}^{i-1}\lambda_j(1-\tilde{B}_j(\sum_{n=1}^{N}\lambda_n(1-z_n))) + \sum_{j=i}^{N}\lambda_j(1-z_j)\right], \tag{4.2}$$

with $\tilde{C}(.)$, the LST of the cycle time distribution, being given in (3.8).

If interruptions (either preemptive or non-preemptive) may occur, it easily follows (cf. (3.1) and (3.6)) that (4.2) remains valid when in its righthand side $\tilde{S}_j(\omega)$ is replaced by $\tilde{S}_j(\phi(\omega))$ and $\tilde{B}_j(\omega)$ is replaced by $\tilde{B}_j(\phi(\omega))$ (with $\omega=\sum_{n=1}^{N}\lambda_n(1-z_n)$). Note that $i=1$ yields the generating function of the joint distribution of $\mathbf{X}_1,\ldots,\mathbf{X}_N$ (cf. (3.4)).

From (4.2), in the case *with* server interruptions, we find:

$$E\mathbf{Y}_{im} = \lambda_m(1+\mu p)\sum_{j=1}^{i-1} s_j + \lambda_m E\mathbf{C}[(1+\mu p)\sum_{j=1}^{i-1}\rho_j + I(m\geq i)]. \tag{4.3}$$

We also derive the covariance of $\mathbf{Y}_{im}$ and $\mathbf{Y}_{in}$ from (4.2). In view of the complexity of the expression we first give the result for the case without interruptions:

$$cov(\mathbf{Y}_{im},\mathbf{Y}_{in}) = \lambda_m\lambda_n[Var\{\sum_{k=1}^{i-1}\mathbf{S}_k\} + E\mathbf{C}\sum_{k=1}^{i-1}\lambda_k\beta_k^{(2)} \tag{4.4}$$

$$+ Var(\mathbf{C})[\sum_{k=1}^{i-1}\rho_k + I(m\geq i)][\sum_{k=1}^{i-1}\rho_k + I(n\geq i)]] > 0,$$

$I(.)$ denoting an indicator function. For $i=1$, as $\mathbf{Y}_{1m}=\mathbf{X}_m$ and $\mathbf{Y}_{1n}=\mathbf{X}_n$, we find $cov(\mathbf{X}_m,\mathbf{X}_n) = \lambda_m\lambda_n Var(\mathbf{C})$.

For the case *with* server interruptions, using (4.2) with the modifications of its righthand side as described above, we derive:

$$cov(\mathbf{Y}_{im},\mathbf{Y}_{in}) = \lambda_m\lambda_n\,(1+\mu p)^2\,[Var\{\sum_{k=1}^{i-1}\mathbf{S}_k\} + E\mathbf{C}\sum_{k=1}^{i-1}\lambda_k\beta_k^{(2)}] \tag{4.5}$$

$$+ \lambda_m\lambda_n\,Var(\mathbf{C})[(1+\mu p)\sum_{k=1}^{i-1}\rho_k + I(m\geq i)][(1+\mu p)\sum_{k=1}^{i-1}\rho_k + I(n\geq i)]$$

$$+ \lambda_m\lambda_n\mu p^{(2)}[E\mathbf{C}\sum_{k=1}^{i-1}\rho_k + \sum_{k=1}^{i-1}s_k] > 0,$$

where (see (2.1) and (2.2)) $1+\mu p = 1/(1-\mu r)$ and $p^{(2)} = r^{(2)}/(1-\mu r)^3$.

Very similar to (4.1) we can derive the joint distribution of the queue lengths $\mathbf{Y}_{11},\cdots,\mathbf{Y}_{NN}$ successively found by the server in a cycle, *when arriving at* $Q_1,\cdots,Q_N$. In the case of no interruptions (μ=0) we have:

$$E[z_1^{\mathbf{Y}_{11}} z_2^{\mathbf{Y}_{22}} \cdots z_N^{\mathbf{Y}_{NN}} \,|\, \mathbf{X}_1,\cdots,\mathbf{X}_N] \tag{4.6}$$

$$= E[z_1^{\mathbf{X}_1} z_2^{\mathbf{X}_2+A_2[V_1(\mathbf{X}_1)+\mathbf{S}_1]} z_3^{\mathbf{X}_3+A_3[V_1(\mathbf{X}_1)+\mathbf{S}_1+V_2(\mathbf{X}_2)+\mathbf{S}_2]} \cdots z_N^{\mathbf{X}_N+A_N[V_1(\mathbf{X}_1)+\mathbf{S}_1+\cdots+V_{N-1}(\mathbf{X}_{N-1})+\mathbf{S}_{N-1}]} \,|\, \mathbf{X}_1,\cdots,\mathbf{X}_N]$$

$$= \prod_{j=1}^{N-1}\tilde{B}_j^{\mathbf{X}_j}(\sum_{n=j+1}^{N}\lambda_n(1-z_n))\prod_{j=1}^{N-1}\tilde{S}_j(\sum_{n=j+1}^{N}\lambda_n(1-z_n))\, z_1^{\mathbf{X}_1} z_2^{\mathbf{X}_2}\cdots z_N^{\mathbf{X}_N}.$$

Hence, using (3.4), for $|z_1|\leq 1,\cdots,|z_N|\leq 1$,

$$E[z_1^{\mathbf{Y}_{11}} z_2^{\mathbf{Y}_{22}} \cdots z_N^{\mathbf{Y}_{NN}}] = \prod_{j=1}^{N-1}\tilde{S}_j(\sum_{n=j+1}^{N}\lambda_n(1-z_n))\,\tilde{C}\Bigg[\sum_{j=1}^{N}\lambda_j\Bigg(1-z_j\tilde{B}_j(\sum_{n=j+1}^{N}\lambda_n(1-z_n))\Bigg]\Bigg]. \tag{4.7}$$

If interruptions may occur, (4.7) remains valid when in the righthand side $\tilde{S}_j(\omega)$ is replaced by $\tilde{S}_j(\phi(\omega))$ and $\tilde{B}_j(\omega)$ is replaced by $\tilde{B}_j(\phi(\omega))$.

$cov\,(\mathbf{Y}_{mm},\mathbf{Y}_{nn})$ can be obtained from (4.7). We give the formula in the case without interruptions; the case with interruptions leads to a similar formula.

$$cov(\mathbf{Y}_{mm},\mathbf{Y}_{nn}) = \lambda_m\lambda_n[Var\{\sum_{k=1}^{m-1}\mathbf{S}_k\} \tag{4.8}$$

$$+ Var(\mathbf{C})[\sum_{k=1}^{m-1}\rho_k + 1][\sum_{k=1}^{n-1}\rho_k + 1] + E\mathbf{C}[\beta_m + \sum_{k=1}^{m-1}\lambda_k\beta_k^{(2)}]] > 0, \quad m \le n.$$

5. Waiting times of preventive and corrective maintenance requests

The waiting time of a customer (preventive maintenance request) is defined as the time between his arrival at a queue and the time instant at which his service starts (for the first time). Consider an arbitrary customer K at Q_k in the underlying polling system. His waiting time $\mathbf{W}_k$, as defined above, is composed of

(i) a residual cycle time $\mathbf{C}_r$,

(ii) the service times of all customers who arrive at $Q_1,...,Q_{k-1}$ during the cycle in which K arrives,

(iii) the switchover times of the server from Q_1 to $Q_2, \ldots, Q_{k-1}$ to Q_k,

(iv) the service times of all customers who arrive at Q_k during the past part, $\mathbf{C}_p$, of the cycle in which K arrives,

(v) the time during which the server is interrupted within the time periods described in (ii), (iii) and (iv).

Note that the waiting time of a preventive maintenance request - as defined above, excluding interruption periods because of corrective maintenance - is independent of the interruption-handling policy being employed; that is, it is (probabilistically) the same under the preemptive resume and the non-preemptive service disciplines.

We now calculate the distribution of $\mathbf{W}_k$, k=1,...,N. Using the above described five-term decomposition of the waiting time of an arbitrary customer at Q_k (k=1,...,N) and the reasoning leading to (3.1), we can write for Re $\omega \ge 0$:

$$E\{e^{-\omega\mathbf{W}_k}\} = \prod_{i=1}^{k-1}\tilde{S}_i\{\phi(\omega)\}E\left\{\exp\left[-\sum_{i=1}^{k}\lambda_i[1-\tilde{B}_i\{\phi(\omega)\}]\mathbf{C}_p\right]\exp\left[-\sum_{i=1}^{k-1}\lambda_i[1-\tilde{B}_i\{\phi(\omega)\}] + \omega]\mathbf{C}_r\right]\right\}.$$

Using (3.11) it follows that for Re $\omega \ge 0$:

$$E\{e^{-\omega\mathbf{W}_k}\} = \prod_{i=1}^{k-1}\tilde{S}_i\{\phi(\omega)\}\,\frac{1}{E\mathbf{C}}\,\frac{\tilde{C}\left\{\sum_{j=1}^{k}\lambda_j[1-\tilde{B}_j\{\phi(\omega)\}]\right\} - \tilde{C}\left\{\sum_{j=1}^{k-1}\lambda_j[1-\tilde{B}_j\{\phi(\omega)\}]+\omega\right\}}{\omega-\lambda_k(1-\tilde{B}_k\{\phi(\omega)\})}. \tag{5.1}$$

It should be noted that the LST's of the waiting and sojourn time distributions at Q_k have the same simple relationship to the generating function of the queue length distribution in Q_k at customer departure epochs from Q_k as in the ordinary M/G/1 queue. Thus (5.1) immediately yields not only the generating function of the queue length distribution for Q_k at customer departure epochs, but also at customer arrival epochs (by a standard up-and-down-crossing argument) and at arbitrary epochs (by the PASTA property).

Differentiating (5.1) with respect to ω and evaluating the resulting expression at ω=0 we obtain the following expression for the mean waiting time of an arbitrary customer at Q_k, k=1,2,...,N:

$$E\mathbf{W}_k = E\mathbf{C}_r[1 + \frac{1}{1-\mu r}[2\sum_{j=1}^{k-1}\rho_j + \rho_k]] + \frac{1}{1-\mu r}\sum_{j=1}^{k-1}s_j, \tag{5.2}$$

with EC_r given by (3.12). There are no essential difficulties in deriving explicit expressions for higher moments of the waiting time distributions.

Remark 5.1
If the polling model consists of a *single* queue without switchover times, then EW_1 coincides with the mean waiting time of low-priority customers in an M/G/1 queue with two priority levels and non-preemptive priority (cf. Cohen [6], Section III.3.8).

It is obvious from (5.2) that $EW_1 < EW_2 < \cdots < EW_N$. In particular,

$$EW_{k+1} - EW_k = \frac{1}{1-\mu r}[(\rho_{k+1}+\rho_k)EC_r + s_k]. \tag{5.3}$$

Using (5.2) and (3.12) we obtain the pseudoconservation law for this model:

$$\sum_{k=1}^{N} \rho_k EW_k = \frac{\rho}{1-\mu r}\left[\frac{1}{2(1-\rho-\mu r)}[\sum_{i=1}^{N} \lambda_i \beta_i^{(2)} + \mu r^{(2)}] + \frac{s^{(2)}}{2s} + \frac{\rho s}{1-\mu r-\rho}\right] + \frac{1}{1-\mu r}\sum_{k=1}^{N}\rho_k \sum_{j=1}^{k-1} s_j. \tag{5.4}$$

Formula (5.4) could have also been obtained without explicitly determining the individual mean waiting times, by using the standard derivation of pseudoconservation laws in polling models with switchover times (see Boxma [3], and see in particular Boxma, Levy and Yechiali [4] for the Globally Gated case, with $\mu=0$).

The waiting time of a corrective maintenance request
The waiting time of a corrective maintenance request is defined as the time span between the occurence of a breakdown of an installation and the arrival of the repairman. The determination of the waiting time of a corrective maintenance request is dependent on whether the repairman applies a preemptive resume discipline, or he follows a non-preemptive procedure. We shall determine the LST and the first moment of the waiting time distribution of a corrective maintenance request for both disciplines. Recall that if the server is switching we assume that the corrective maintenance request gets priority with no delay. The breakdowns occur according to a Poisson process as described in Section 2.

Define:
W^p_{corr} := waiting time of a corrective maintenance request under the preemptive resume discipline;
W^{np}_{corr} := waiting time of a corrective maintenance request under the non-preemptive discipline.

The preemptive resume discipline
If the repairman reacts immediately to a breakdown, the waiting time of corrective maintenance requests can be described as the waiting time in an M/G/1 queue with arrival intensity μ and service times distributed as **R**. We can write (cf. Cohen [6]:Section II.4.5):

$$EW^p_{corr} = \frac{\mu r^{(2)}}{2(1-\mu r)}, \tag{5.5}$$

and

$$E\{e^{-\omega W^p_{corr}}\} = (1-\mu r)\frac{\omega}{\omega-\mu(1-\tilde{R}\{\omega\})}, \quad \text{Re}\ \omega \geq 0. \tag{5.6}$$

The non-preemptive discipline
Recall that, under the non-preemptive discipline, if a breakdown occurs somewhere in the system when the repairman is performing a preventive maintenance job, he will first finish the preventive maintenance job and only then travel to the broken installation. The repairman will resume operating on preventive tasks only

when there are no other corrective maintenance jobs. When the system is in a period where the repairman is at the i-th installation or deals with corrective maintenance jobs generated while he was attached to Q_i, then a corrective maintenance job's waiting time is identical to the waiting time of a customer in an M/G/1 queue with (multiple) vacations (cf. Levy and Yechiali [10], Doshi [7]) where the arrival rate is μ, service times are distributed as **R**, and vacation durations have distribution $B_i(.)$.

When the system is in a period where the repairman is walking or is occupied with corrective jobs connected to the walking times, then a corrective maintenance request's waiting time is the same as that of a customer in a regular M/G/1 queue with arrival rate μ and service times distributed as **R**. This is true following the assumption that corrective maintenance jobs have preemptive priority over walking times.

From Levy and Yechiali [10] we obtain the following expressions for the first moment and LST of the waiting time distribution in an M/G/1 vacation queue in which the arrival rate is denoted by μ, the first and second moment of the service time distribution are respectively denoted by r and $r^{(2)}$, the first and second moment of the distribution of the vacation period are respectively denoted by β and $\beta^{(2)}$, and the LST of the distribution of the vacation period is denoted as $\tilde{B}\{.\}$:

$$EW_{M/G/1}^{vacation} = \frac{\mu r^{(2)}}{2(1-\mu r)} + \frac{\beta^{(2)}}{2\beta}, \tag{5.7}$$

$$E\{e^{-\omega W_{M/G/1}^{vacation}}\} = E\{e^{-\omega W_{M/G/1}}\}\,\frac{1-\tilde{B}\{\omega\}}{\beta\omega}, \quad \mathrm{Re}\ \omega \geq 0, \tag{5.8}$$

where $E\{\exp[-\omega W_{M/G/1}]\}$ is the LST of the waiting time distribution in the underlying M/G/1 queue with no vacations.

Conditioning on the occurence of the above periods we can write:

$EW_{corr}^{np} = EW_{corr}^{np}$ I(repairman is walking or is dealing with corrective maintenance generated during a walk period)} +

$\sum_{i=1}^{N} E\{W_{corr}^{np}$ I(repairman is at the i-th installation or is dealing with corrective maintenance generated while he was serving at the i-th installation)},

with I(A) being the indicator function of the event A.

It is easy to see that:

$\alpha_i = Pr$\{repairman is at the i-th installation or is dealing with corrective maintenance generated while he was serving at the i-th installation\} $= \rho_i/(1-\mu r)$, $i=1,2,...,N$,

Pr\{repairman is walking or is busy with corrective maintenance generated during a walking period\} $= 1 - \sum_{i=1}^{N} \alpha_i = (1-\mu r-\rho)/(1-\mu r)$.

Hence we obtain:

$$EW_{corr}^{np} = \frac{1-\mu r-\rho}{1-\mu r}\,\frac{\mu r^{(2)}}{2(1-\mu r)} + \sum_{i=1}^{N} \frac{\rho_i}{1-\mu r}\,[\frac{\mu r^{(2)}}{2(1-\mu r)} + \frac{\beta_i^{(2)}}{2\beta_i}] = \frac{\mu r^{(2)} + \sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\mu r)}. \tag{5.9}$$

Indeed, since walking times do not defer corrective maintenance jobs, the waiting time of a corrective job is the same as that of a highest priority class job in an M/G/1 queue with N+1 classes and non-preemptive service discipline (see Kella and Yechiali [9]). The other classes are formed by the preventive maintenance jobs in queues $Q_1, \ldots, Q_N$.

The LST of W_{corr}^{np} is obtained using the same argument as in (5.9), and applying (5.6):

$$E\{e^{-\omega \mathbf{W}^{np}_{corr}}\} = \frac{1-\mu r-\rho}{1-\mu r}\frac{(1-\mu r)\omega}{\omega-\mu(1-\tilde{R}\{\omega\})} + \sum_{i=1}^{N}\frac{\rho_i}{1-\mu r}\frac{(1-\mu r)\omega}{\omega-\mu(1-\tilde{R}\{\omega\})}\frac{1-\tilde{B}_i\{\omega\}}{\beta_i\omega} \tag{5.10}$$

$$= \frac{(1-\mu r-\rho)\omega + \sum_{i=1}^{N}\lambda_i(1-\tilde{B}_i\{\omega\})}{\omega-\mu(1-\tilde{R}\{\omega\})}.$$

Note that switchover times do not appear in (5.9) and (5.10). This is due to the fact that the interruption handling mechanism during a switchover period is preemptive.

Another way to obtain (5.10) is to use the concept of a $\mathbf{B}_i$-delay cycle (cf. Kella and Yechiali [9]) in an M/G/1 queue with arrival rate μ, service times distributed as $\mathbf{R}$, and a delay $\mathbf{B}_i$ with distribution $B_i(.)$; that is, at the beginning of each busy period there is a delay of length $\mathbf{B}_i$ before actual service to customers can be started. We have

$$E\{e^{-\omega \mathbf{W}^{np}_{corr}} \mid \textit{system is within a } \mathbf{B}_i\textit{–delay cycle}\} = \frac{(1-\mu r)\omega}{\omega-\mu(1-\tilde{R}\{\omega\})}\frac{1-\tilde{B}_i\{\omega\}}{\beta_i\omega}. \tag{5.11}$$

Result (5.10) is readily obtained by using (5.6) and (5.11) and by observing the following: the probability that the system is within a $\mathbf{B}_i$-delay cycle equals $\alpha_i = \rho_i/(1-\mu r)$, and the probability that the repairman is walking or busy with corrective maintenance requests generated during the walking period equals $1-\rho/(1-\mu r)$.

The pseudoconservation law for the mean waiting times (including those of corrective maintenance requests) is obtained by combining (5.4) and (5.9):

$$\mu r E\mathbf{W}^{np}_{corr} + \sum_{k=1}^{N}\rho_k E\mathbf{W}_k = \frac{\mu r+\rho}{2(1-\mu r-\rho)}[\sum_{i=1}^{N}\lambda_i\beta_i^{(2)} + \mu r^{(2)}] \tag{5.12}$$

$$+\frac{\rho}{1-\mu r}[\frac{s^{(2)}}{2s} + \frac{\rho s}{1-\mu r-\rho}] + \frac{1}{1-\mu r}\sum_{k=1}^{N}\rho_k\sum_{j=1}^{k-1}s_j.$$

Remark 5.2
Define for $k=1,...,N$:

$\hat{\mathbf{W}}_k^p :=$ the time during which an arbitrary Q_k customer is in the system without receiving service, under the preemptive resume interruption-handling discipline;

$\hat{\mathbf{W}}_k^{np} :=$ the time during which an arbitrary Q_k customer is in the system without receiving service, under the non-preemptive interruption-handling discipline.

Clearly $E\hat{\mathbf{W}}_k^{np} = E\mathbf{W}_k$, but

$$E\hat{\mathbf{W}}_k^p = E\mathbf{W}_k + E\{\textit{time during which the service of a customer at } Q_k \textit{ is interrupted}\}$$

$$= E\mathbf{W}_k + \beta_k\mu p = E\mathbf{W}_k + \beta_k\frac{\mu r}{1-\mu r},\quad k=1,...,N. \tag{5.13}$$

Combining these results with (5.5) and (5.9), it follows that

$$\left[\mu r E\mathbf{W}^{np}_{corr} + \sum_{k=1}^{N}\rho_k E\hat{\mathbf{W}}_k^{np}\right] - \left[\mu r E\mathbf{W}^{p}_{corr} + \sum_{k=1}^{N}\rho_k E\hat{\mathbf{W}}_k^{p}\right] = \frac{\mu r}{1-\mu r}\left[\sum_{k=1}^{N}\rho_k[\frac{\beta_k^{(2)}}{2\beta_k} - \beta_k]\right]. \tag{5.14}$$

The righthand side of (5.14) does not involve any switchover times, because the server behaves the same under both disciplines when corrective maintenance occurs during a switchover period. Note that the righthand side of (5.14) becomes zero if the service times at all queues are negative exponentially distributed. Indeed, as can be seen from Theorem 6.2 of Gelenbe and Mitrani [8], the weighted sum of mean waiting times of *all* types of customers satisfies a conservation law in this particular case. Formula (5.14) is an interesting deviation from the theory of conservation laws; for a case *without* conservation, it gives a simple expression for the difference between the weighted sums of mean waiting times. (5.14) and (5.12) together yield an expression for a weighted sum of the mean waiting times in the case of a preemptive resume priority.

The expression in the righthand side of (5.14) can be explained in the following way. Observe that the total mean amount of work of waiting requests (corrective plus preventive) is the same under the preemptive and non-preemptive disciplines. In the non-preemptive (resp. preemptive) case, the mean amount of work of all waiting requests can be expressed as the mean number of waiting requests times their mean (resp. mean residual) service times. Using Little's formula, in the non-preemptive case the mean amount of work of all waiting requests equals

$$\mu r E\mathbf{W}_{corr}^{np} + \sum_{k=1}^{N} \rho_k E\hat{\mathbf{W}}_k^{np}.$$

In the preemptive case though, a fraction $\rho_k \mu p = \rho_k \mu r/(1-\mu r)$ of the time Q_k contains an interrupted customer. This interrupted customer has mean residual service time $\beta_k^{(2)}/2\beta_k$ instead of β_k. Summation over k of the resulting difference yields the righthand side of (5.14).

6. Optimal visit orders

Our objective in this section is to derive rules for the optimal operation of the polling (maintenance) system as described above. Let c_i represent the cost of a customer (i.e., preventive maintenance job) being delayed one unit of time in Q_i. Thus, the mean waiting cost of a customer at Q_i is $c_i E\mathbf{W}_i$.

We are interested in minimizing the waiting cost of an arbitrary customer in the model: $\sum_{n=1}^{N} (\lambda_n/\lambda) c_n E\mathbf{W}_n$ with $\lambda := \sum_{n=1}^{N} \lambda_n$. Such a minimization will determine the *static* order in which the server visits the various queues every cycle. However, a policy in which the order of visits may change from one cycle to the next - following the dynamic evolution of the system - is also of interest and will be discussed in the sequel.

Static optimization

From (5.2) it follows that

$$\sum_{k=1}^{N} \lambda_k c_k E\mathbf{W}_k = E\mathbf{C}_r \sum_{k=1}^{N} \lambda_k c_k [1 + \frac{1}{1-\mu r}[2\sum_{j=1}^{k-1} \rho_j + \rho_k]] + \frac{1}{1-\mu r} \sum_{k=1}^{N} \lambda_k c_k \sum_{j=1}^{k-1} s_j, \tag{6.1}$$

in which the only term that depends on the order of the queues is

$$\frac{1}{1-\mu r} \sum_{k=1}^{N} \lambda_k c_k \sum_{j=1}^{k-1} [2E\mathbf{C}_r \rho_j + s_j].$$

Using a standard interchange argument one can easily show that the optimal ordering of the queues is obtained by arranging them in an increasing order of

$$u_j = [2E\mathbf{C}_r \rho_j + s_j]/\lambda_j c_j. \tag{6.2}$$

Dynamic optimization

In applications in which the queue lengths can be evaluated at the cycle-beginnings (as in the maintenance problem of this paper), the GG policy can be used to dynamically control and optimize the system (see Browne and Yechiali [5]). If we consider the costs incurred during a cycle by the customers present at its initiation *together* with the cost incurred by the new arrivals between two cycle-beginnings, the long run minimal cost can be achieved by optimizing each cycle individually. This is true because the length of a cycle is independent of the order in which the queues are visited. The mean total waiting cost incurred during the coming cycle is:

$$\frac{1}{1-\mu r}[\sum_{k=1}^{N} c_k X_k \sum_{j=1}^{k-1} (X_j\beta_j+s_j) + \sum_{k=1}^{N} c_k\beta_k \sum_{j=1}^{X_k-1} j] + \sum_{k=1}^{N} c_k\lambda_k E\{\mathrm{C}(X_1,\ldots,X_N)^2\}/2, \tag{6.3}$$

with X_i the number of customers present at Q_i at a cycle beginning, and with $\mathrm{C}(X_1,\ldots,X_N)$ the length of such a cycle.

It can be shown that the optimal order for the next cycle is determined by *increasing* values of the indices

$$u_j = [X_j\beta_j + s_j]/c_jX_j. \tag{6.4}$$

Observe that Formulas (6.3) and (6.4) correct misprints in Subsection 4.1.3 of Boxma, Levy and Yechiali [4]. It is interesting to note that the presence of interruptions does not change the optimal order of visits. When $s_j \to 0$, the index rule approaches the well known $c\mu$ rule.

The elevator-type polling scheme

A static visit order that can be employed is the so-called elevator polling (see Altman, Khamisy and Yechiali [1]), in which the server first visits the installations in the order $1,2,\ldots,N-1,N$ ('up' or clockwise direction), then reverses its direction, visiting the installations in the order $N,N-1,\ldots,2,1$ ('down' or counter-clockwise direction), then changes its direction again, and so on. An immediate advantage of such a visit scheme is that it saves the switchover time S_N from Q_N to Q_1.

Under the GG service regime the marking of customers takes place each time the server changes direction and starts a new cycle, whether it is a down or an up cycle. If we assume that the switchover time, S_i, to move from Q_i to Q_{i+1}, has the same distribution as the switchover time in the opposite direction from Q_{i+1} to Q_i, then all up and down cycles have the same distribution as C (with the modification that $\mathrm{S}_N=0$). As a result, the combination of elevator polling with GG regime yields a 'fair' system in which the mean waiting times of customers (preventive maintenance jobs) at all installations are the same - even though the traffic parameters at the various installations may differ. To see this, observe that the actual waiting time of an arriving customer at Q_k depends on whether, upon that customer's arrival, the server is in its up or down direction. However, as all cycles are probabilistically the same, the probability of 'catching' an up or a down cycle is 0.5. We can therefore write:

$$E\mathrm{W}_k = 0.5E\,[\mathrm{W}_k \,|\, cycle\ N\to 1] + 0.5E\,[\mathrm{W}_k \,|\, cycle\ 1\to N]. \tag{6.5}$$

Similar to the arguments leading to (5.2) we have:

$$E\,[\mathrm{W}_k \,|\, cycle\ N\to 1] = E\mathrm{C}_r[1 + \frac{1}{1-\mu r}(2\sum_{j=1}^{k-1}\rho_j + \rho_k)] + \frac{1}{1-\mu r}\sum_{j=1}^{k-1} s_j. \tag{6.6}$$

Changing direction and indices, it readily follows that

$$E\,[\mathrm{W}_k \,|\, cycle\ 1\to N] = E\mathrm{C}_r[1 + \frac{1}{1-\mu r}(2\sum_{j=k+1}^{N}\rho_j + \rho_k)] + \frac{1}{1-\mu r}\sum_{j=k}^{N-1} s_j. \tag{6.7}$$

Combining (6.5), (6.6) and (6.7) results in

$$EW_k = EC_r[1 + \frac{\rho}{1-\mu r}] + \frac{1}{2(1-\mu r)} \sum_{j=1}^{N-1} s_j. \tag{6.8}$$

It follows that *mean waiting times are the same in each installation,* just as in the case without interruptions that was studied by Altman, Khamisy and Yechiali [1].

Formula (6.8) can be easily understood by observing that an average mean waiting time consists of a mean residual cycle time (before the start of the cycle in which the arriving customer is served), half the mean total switchover time (including interruptions), and the mean amount of work (including interruptions) done in the cycle in which the arriving customer is served - but before his own service. Using a balancing argument and the symmetry of the up and down cycles, we can write this last term as $0.5(EC_p + EC_r)\rho/(1-\mu r)$.

As in Altman, Khamisy and Yechiali [1], one can study the measure of variation, $|\Delta_k|$, in the waiting times incurred at Q_k, where

$$\Delta_k = E[W_k \,|\, cycle\ N \to 1] - E[W_k \,|\, cycle\ 1 \to N].$$

From (6.6) and (6.7) it can easily be derived that $\Delta_k = \sum_{j=1}^{k-1} a_j - \sum_{j=k+1}^{N} a_j - s_k/(1-\mu r)$, where $a_j = (2\rho_j EC_r + s_j)/(1-\mu r)$. It turns out that $\Delta_1 < 0$, $\Delta_N > 0$ and that Δ_k is increasing in k. Thus, one is interested in arranging the installations so as to *minimize* the largest value of $|\Delta_k|$. Such a goal is achieved when installation 1 has the largest value of all loads ρ_j, and installation N the second largest load (or vice versa).

7. The case of the dormant server

So far it has been assumed that the server keeps switching from one queue to another even when there are no customers present in the system. In the present section we briefly consider the case in which the server remains idle in Q_1 when $Q_1, \cdots, Q_N$ are all empty at the end of a cycle. We refer to the forthcoming report of Borst [2] for a detailed exposition of this problem, and other polling problems with a non-moving idle server, in the case *without* interruptions.

Suppose that all queues are idle at the end of the n-th cycle; this event has probability $\int_0^\infty \exp[-\Lambda t]\, dPr\{C_n < t\} = \tilde{C}_n(\Lambda)$, with $\Lambda := \sum_{i=1}^{N} \lambda_i$ and C_n denoting the length of the n-th cycle. We say that an *idle period* starts at the end of such a cycle. A new cycle can only start after a customer has arrived at one of the N queues. There are two possibilities: (i) a customer arrives at one of the N queues while the server is idle, waiting at Q_1 (probability z, to be determined below), or (ii) one or more customers arrive at $Q_1, \cdots, Q_N$ while the server is in an interruption period, performing corrective maintenance (probability $1-z$). In the latter case the n+1-st cycle only starts when the interruption period is completed. In either case the n+1-st cycle consists of (possibly extended) switchover periods and of (possibly extended) services of the customer(s) present at the beginning of this cycle. Note that, with $q = \mu/(\mu+\Lambda)$= Pr{in an idle period, an interruption occurs before an arrival at one of the queues} and with $\tilde{P}(\lambda)$ being the probability of no arrivals during an interruption period,

$$z = \sum_{k=0}^{\infty} (q\tilde{P}(\lambda))^k (1-q) = \frac{1-q}{1-q\tilde{P}(\Lambda)}. \tag{7.1}$$

Now we can write:

$$\tilde{C}_{n+1}(\omega) = \tilde{S}(\phi(\omega))\, [\tilde{C}_n(\delta(\omega)) - \tilde{C}_n(\Lambda)] \tag{7.2}$$

$$+\tilde{S}(\phi(\omega))\,\tilde{C}_n(\Lambda)\,[z\sum_{i=1}^{N}\frac{\lambda_i}{\Lambda}\tilde{B}_i(\phi(\omega))+(1-z)\{\tilde{P}(\delta(\omega))-\tilde{P}(\Lambda)\}].$$

The first term in the righthand side of (7.2) concerns the case in which the n-th cycle contains at least one arrival at $Q_1,\cdots,Q_N$, whereas the last term concerns the case of no such arrival. Introducing

$$F(\omega):=1-z(1-\frac{\delta(\omega)}{\Lambda})-(1-z)\{\tilde{P}(\delta(\omega))-\tilde{P}(\Lambda)\},$$

and taking the limit for $n\to\infty$, writing $\tilde{C}(\omega)=\lim_{n\to\infty}\tilde{C}_n(\omega)$ (note that $\rho+\mu r<1$), we can rewrite (7.2) into:

$$\tilde{C}(\omega)=\tilde{S}(\phi(\omega))[\tilde{C}(\delta(\omega))-\tilde{C}(\Lambda)F(\omega)]. \tag{7.3}$$

We solve (7.3) iteratively, similar to (3.5). M iteration steps lead to:

$$\tilde{C}(\omega)=\prod_{m=0}^{M}\tilde{S}(\phi(\delta^{(m)}(\omega)))\,\tilde{C}(\delta^{(M+1)}(\omega)-\tilde{C}(\Lambda)\sum_{m=0}^{M}F(\delta^{(m)}(\omega))\prod_{h=0}^{m}\tilde{S}(\phi(\delta^{(h)}(\omega))). \tag{7.4}$$

It can be shown that the sums and product in (7.4) converge and that $\tilde{C}^{(M+1)}(\omega)\to\tilde{C}(0)=1$ for $\rho+\mu r<1$; hence

$$\tilde{C}(\omega)=\prod_{m=0}^{\infty}\tilde{S}(\phi(\delta^{(m)}(\omega)))-\tilde{C}(\Lambda)\sum_{m=0}^{\infty}F(\delta^{(m)}(\omega))\prod_{h=0}^{m}\tilde{S}(\phi(\delta^{(h)}(\omega))). \tag{7.5}$$

Substitution of $\omega=\Lambda$ in (7.5) leads to the determination of $\tilde{C}(\Lambda)$, and hence to that of $\tilde{C}(\omega)$:

$$\tilde{C}(\Lambda)=\frac{\prod_{m=0}^{\infty}\tilde{S}(\phi(\delta^{(m)}(\Lambda)))}{1+\sum_{m=0}^{\infty}F(\delta^{(m)}(\Lambda))\prod_{h=0}^{m}\tilde{S}(\phi(\delta^{(h)}(\Lambda)))}. \tag{7.6}$$

Cycle time moments can again be easily determined, and LST's (cq. moments) of the waiting time distributions can be expressed in the LST (cq. moments) of the cycle time distribution in a similar way as was done in Section 5; cf. also Borst [2]. As an example we obtain the mean cycle time, by differentiating both sides of (7.3) w.r.t. ω and substituting $\omega=0$:

$$EC=\frac{s}{1-\mu r-\rho}[1-\tilde{C}(\Lambda)\tilde{P}(\Lambda)(1-z)]+\frac{\rho}{1-\mu r-\rho}\tilde{C}(\Lambda)[\frac{z}{\Lambda}+\frac{r}{1-\mu r}(1-z)]. \tag{7.7}$$

Here the interruption period's LST $\tilde{P}(\Lambda)$ is determined as the unique zero of $x=\tilde{R}(\Lambda+\mu(1-x))$ on (0,1), cf. Cohen [6]: p. 250.

8. Conclusions

We have studied a repairman-type problem modelled as a Globally Gated polling system with server interruptions, while providing a detailed analysis of cycle times, queue lengths and waiting times. The polling model is used as a vehicle to analyze and optimize a real maintenance process in which a single repairman is handling two types of maintenance, viz. preventive and corrective, generated by various installations. We have derived the LST and first moment of the waiting time distributions of preventive and corrective maintenance jobs for the two cases where corrective maintenance gets priority according to the preemptive resume discipline or according to the non-preemptive discipline. We have also derived rules for the static, as well as dynamic, optimal operation of such a maintenance system.

From a mathematical point of view, the Globally Gated service discipline is much simpler than any other service discipline in polling systems studied in the literature. It appears to be amenable to a very detailed analysis, thus yielding much insight into the queueing behaviour of systems that operate under this, or a similar, discipline. The Globally Gated discipline can be viewed as a reasonably realistic service discipline for modeling maintenance processes with a traveling repair crew. We mention the following topics for further research, that might further enhance the applicability of the model for maintenance situations:
(i) the corrective maintenance process depends on the amount of preventive maintenance in the system;
(ii) the k-th installation generates its own stream of corrective maintenance requests with intensity μ_k;
(iii) other priority disciplines with respect to corrective maintenance are in effect.

References

1. Altman, E., Khamisy, A., Yechiali, U. [1992]. On elevator polling with globally gated regime. *To appear in Queueing Systems 11, Special Issue on Polling Models.*

2. Borst, S.C. [1992]. CWI Report BS-R92xx, in preparation.

3. Boxma, O.J. [1989]. Workloads and waiting times in single-server queues with multiple customer classes. *Queueing Systems 5,* 185-214.

4. Boxma, O.J., Levy, H., Yechiali, U. [1992]. Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. *To appear in Annals of Operations Research.*

5. Browne, S., Yechiali, U. [1989]. Dynamic priority rules for cyclic-type queues. *Advances in Applied Probability 21,* 432-450.

6. Cohen, J.W. [1982]. *The Single Server Queue.* (North-Holland, Amsterdam; 2nd edition).

7. Doshi, B.T. [1986]. Queueing systems with vacations - A survey. *Queueing Systems 1,* 29-66.

8. Gelenbe, E., Mitrani, I. [1980]. *Analysis and Synthesis of Computer Systems.* (Academic Press, New York).

9. Kella, O., Yechiali, U. [1988]. Priorities in M/G/1 queues with server vacations. *Naval Research Logistics 35,* 23-34.

10. Levy, Y., Yechiali, U. [1975]. Utilization of idle time in an M/G/1 queueing system. *Management Science 22,* 202-211.