Cyclic Bernoulli Polling [*]

Eitan Altman
INRIA
Centre Sophia Antipolis
06565 Valbonne Cedex, France

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

November 1992

Abstract

We introduce, analyse and optimize the class of Bernoulli random polling systems. The server moves **cyclically** among N channels (queues), but **Change-over** times between stations are composed of **walking** times required to 'move' from one channel to another and **switch-in** times that are incurred **only** when the server actually enters a station to render service. The server uses a **Bernoulli** random mechanism to decide whether to serve a queue or not: upon arrival to channel i, it switches in with probability p_i, or moves on to the next queue (w.p. $1 - p_i$) without serving any customer (e.g. packet or job). The Cyclic Bernoulli Polling (CBP) scheme is independent of the service regime in any particular station, and may be applied to any service discipline. In this paper we analyse three different service disciplines under the CBP scheme: Gated, Partially Exhaustive and Fully Exhaustive. For each regime we derive expressions for (i) the generating functions and moments of the number of customers (jobs) at the various queues at polling instants, (ii) the expected number of jobs that an arbitrary departing job leaves behind it, and (iii) the LST and expectation of the waiting time of a customer at any given queue. The fact that these measures of performance can be explicitly obtained under the CBP is an advantage over all "parameterized" cyclic polling schemes (such as the k-limited discipline) that have been studied in the literature, and for which explicit measures of performance are hard to obtain. The choice of the p_i's in the CBP allows for fine tuning and optimization of performance measures, as well as prioritization between stations (this being achieved at a low

[*] 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.

computational cost). For this purpose, we develop a Pseudo-conservation law for a **mixed** system comprised of channels from all three service disciplines, and define a Mathematical Program to find the optimal values of the probabilities $\{p_i\}_{i=1}^{N}$ so as to minimize the expected amount of unfinished work in the system. Any CBP scheme for which the optimal p_i's are not all equal to one, yields a **smaller** amount of the expected unfinished work in the system than that in the standard cyclic polling procedure with equivalent parameters. We conclude by showing that even in the case of a single queue, it is not always true that $p_1 = 1$ is the best strategy, and derive conditions under which it is optimal to have $p_1 < 1$.

Keywords: Random Cyclic Polling, Walking times, Switch-in times, Optimization.

1 Introduction

Polling systems are used to model an abundant set of systems, such as computer networks, telecommunications or flexible manufacturing systems, repairman applications, and alike. In polling systems one is often interested in using acyclic visiting order of the server to the different channels. This enables flexible prioritization of the different queues which is desired either for optimization purposes, such as minimizing a weighted sum of waiting times in the different queues, or for obtaining fair service among the various queues. Acyclic visit order have been obtained by using a polling table (see [3], [6]), by using random polling ([7, 12, 13]), or by following a dynamic procedure derived by optimization consideration [8, 19]. Unfortunately, in many communication networks, modelled by a polling system (e.g. Local Area Networks based on a token ring protocol) the visit order has to remain cyclic, and the server can not choose in an arbitrary way which queue to visit next. In such cases one may still prioritize among the queues by using the following random access mechanism: when the server arrives to queue i it switches in to render service to jobs awaiting there with probability p_i, or it moves on to the next queue (with probability $1 - p_i$) without serving any jobs in that queue.

In this paper we study such a random cyclic visit mechanism which we call "Cyclic Bernoulli Polling" (CBP). Whenever the server attends a queue and renders service, either a gated regime, or a partially exhaustive scheme, or a fully exhaustive service discipline is assumed to be used. (These service disciplines will be explicitly defined in the sequal).

We assume that *switching* times between the queues are composed of two parts: *walking* times required to 'move' from one station to another, and *switch-in* times that are incurred only when the server enters a station to render service.

It is also possible to prioritize queues by following a simple cyclic visit order while giving service in each queue to only a partial number of the jobs there, according to some parameterized service discipline, e.g. the limited or the Bernoulli service discipline [16, 17, 19]. In that case, however, since the server visits each station on every hamiltonian tour, switch-in times are always incurred. In an effort to save such waisted times, we propose the above described CBP mechanism. An interesting feature of the cyclic Bernoulli polling is that it yields explicit expressions for quantities such as the expected waiting times in the different queues, which are not achievable in many mechanisms of partial service such as limited, Bernoulli or threshold service disciplines [1, 16, 17].

The paper is structured as follows: after introducing the model and notation, we present in Section 2 the evolution equations of the system for the case of Gated service discipline. We then derive implicit expressions for the generating functions of the number of jobs in the different queues at polling instants. This allows us to obtain the first moments of these quantities explicitly, as well as a set of N^3 linear equations to calculate their second moments. Based on these moments, we obtain formulae for the expected waiting times of jobs in the various queues. (The method of "station times", which enables in other models the calculation of the expected waiting times by solving considerably less than N^3 linear equations [9], is not applicable here. This follows from the fact that those station times do not form in our case a Markov Chain).

In Section 3 we analyse the Partially Exhaustive and Fully Exhaustive cases. In Section 4 we obtain an explicit expression for the total expected workload in the system under mixed strategies, and formulate a Mathematical Program to choose the optimal values of the probabilities $p_i,\ i = 1, ..., N$, so as to minimize that workload. Thus any CBP scheme for which the optimal p_i's are not all equal to one, yields a **smaller** amount of the expected unfinished work in the system than that in the standard cyclic polling procedure with equivalent parameters. We solve the Program explicitly for the case of a single queue, and show, surprisingly, that even in such a restricted case, it is not always optimal to choose $p_1 = 1$, a phenomenon which we explain by using an alternative avenue of analysis.

Model and Notation

Consider a polling system with N independent channels, where channel i $(i = 1, 2, ..., N)$ is modeled as an $M/G/1-$type queueing station. The jobs' arrival stream to station i is Poisson with rate λ_i, and service times are distributed as B_i, having Laplace-Stieltjes Transform (LST)

$b_i^*(s)$, and first and second moments b_i and $b_i^{(2)}$, respectively. $B_i(n)$ will represent the *total* service time of n jobs in station i. Denote by $\rho_i \overset{\text{def}}{=} \lambda_i b_i$, and by $\rho \overset{\text{def}}{=} \sum_{i=1}^N \rho_i$ the traffic offered to channel i, and to the system at large, respectively. Ω_i denotes a typical length of a standard M/G/1 queue busy period that starts with one job in queue i, and $\omega_i^*(s)$ denotes the LST of Ω_i. It is well known that $\omega_i \overset{\text{def}}{=} E[\Omega_i]$ satisfies $\omega_i = b_i[1-\rho_i]^{-1}$. Finally, let $\Omega_i(n)$ be the duration of n independent regular M/G/1 busy periods in queue i.

The time it takes between the end of service to the ith station and the polling instant at the next station is called the ith *walking time*, and is denoted by D_i. We assume that walking times are independent, with LST $d_i^*(s)$, and first and second moments d_i and $d_i^{(2)}$, respectively. Let $D = \sum_{i=1}^N D_i$ be the total walking time in a cycle, and denote by d, $d^{(2)}$ and $d^*(s)$ the expectation, second moment and LST of D, respectively.

The time it takes from the moment the server arrives at the ith station (i.e. the polling instant) till service can be started to jobs in that station is called the ith *switch-in time* and is denoted by R_i. We assume that the switch-in times are independent, with LST $r_i^*(s)$, and first and second moments r_i and $r_i^{(2)}$, respectively. These times, the walking times, the inter-arrival times and the service durations are mutually independent.

In a Cyclic Bernoulli Polling the server moves cyclically between the different queues, switching-in to actually give service in queue i with probability p_i, or moving on to the next channel with the complementary probability $1-p_i$. We analyse three service disciplines under the CBP scheme; (i) a **gated** regime, by which only jobs present in the queue at the momoent of server's arrival will be served. (However, before service starts in that queue, a switch-in time is required.) (ii) the **partially exhaustive** regime, where the server, upon finding n jobs, say, at the moment of arrival to a queue, and 'deciding' (by the Bernoulli mechanism) to switch in, stays there (after switch-in time is incurred) for the duration of n M/G/1 busy periods. Hence, if service is actually given to a queue, the number of jobs left there when the server leaves is equal to the number of jobs that arrived during the switch-in time. (iii) the **fully exhaustive** regime, where the server leaves a queue only after it is empty, where upon serving all the jobs that were present there before the switch-in time, plus all those that have arrived during its sojourn (switch-in and visit) time in that queue.

Let $\tilde{X}_n^j$ denote the number of jobs in station j at the nth time that a queue is polled. Without loss of generality, we shall assume that queue 1 is the first to be polled. It is easily seen that the vector $\{\tilde{X}_n^1, ..., \tilde{X}_n^N\}$, $n = 1, 2, ...$ is a Markov chain. We shall assume that this Markov chain is ergodic and $\tilde{X}_{i+kN}^j$, $1 \leq i, j \leq N$, $k = 0, 1, ...$ converges in distribution to a random variable X_i^j, denoting the number of jobs at station j at a polling instant to queue i when the system is in steady-state. (It can be seen from the expression for the expectation of X_i^i given below that $\rho < 1$ together with $p_i > 0$, $i = 1, .., N$ is a necessary condition for ergodicity. Using same technique as in [2], one can prove that these are also sufficient conditions for stability. A detailed analysis of the sufficient conditions for stability is however beyond the scope of this paper. We shall assume throughout that the polling system is ergodic.

and it seems also to be a sufficient one, see [2]).

Let $A_j(T)$ denote the typical *number* of arrivals to station j during a time interval of length T. Hence, $A_j(B_i(X_i^i))$, $A_j(D_i)$, and $A_j(R_i)$ denote, respectively, the number of arrivals to station j during the service of X_i^i jobs at, the walking time from, and the switch-in time to, station i.

2 The Gated Discipline

Evolution Equations

Let a_i be equal to one if the ith queue is to be served when the server arrives, and zero otherwise. a_i, $i = 1, 2, ..., N$ are independent random variables, with $E[a_i] = p_i$. Set $\bar{a}_i = 1 - a_i$.

Denote by $\bar{D}_i \overset{\text{def}}{=} D_i + a_i R_i$ the total switching time related to queue i, and by $\bar{D} = \sum_{i=1}^N \bar{D}_i$ the total switching times in a typical cycle. Set $\bar{d}_i$ and $\bar{d}_i^{(2)}$ as the first and second moments of $\bar{D}_i$, and set $\bar{d}$ and $\bar{d}^{(2)}$ as the first and second moments of $\bar{D}$.

The evolution of the *state* of the system (in steady state) is described by

$$X_{k+1}^i \overset{\text{d}}{=} \begin{cases} X_k^i + A_i\Big(a_k[R_k + B_k(X_k^k)] + D_k\Big) & i \neq k \\ \\ X_k^k \bar{a}_k + A_k\Big(a_k[R_k + B_k(X_k^k)] + D_k\Big) & i = k \end{cases} \tag{1}$$

where $1 \leq i, k \leq N$, and the symbol "$\overset{\text{d}}{=}$" means equality in distribution. (We shall understand $N + 1$ in the evolution equations above to be equal to 1).

Generating Functions

We define a set of multi-dimensional joint generating functions, describing the vector-state of the system at a polling instant of queue k.

Let $F_k(\underline{z}) \overset{\text{def}}{=} E\left[\prod_{i=1}^{N} z_i^{X_k^i}\right]$. Let $\tilde{d}_k \overset{\text{def}}{=} d_k^*(\sum_{i=1}^{N} \lambda_i(1-z_i))$ and define similarly $\tilde{b}_k$ and $\tilde{r}_k$. Using the evolution equations we obtain

$$F_{k+1}(\underline{z}) = E\left\{ z_k^{\bar{a}_k X_k^k + A_k\left(a_k[R_k+B_k(X_k^k)]+D_k\right)} \prod_{\substack{i=1\\ i\neq k}}^{N} z_i^{X_k^i + A_i\left(a_k[R_k+B_k(X_k^k)]+D_k\right)} \right\}$$

$$= E\left[\prod_{i=1}^{N} z_i^{X_k^i} e^{-\lambda_i(1-z_i)\left(a_k[R_k+B_k(X_k^k)]+D_k\right)} z_k^{-a_k X_k^k}\right]$$

$$= \tilde{d}_k E\left\{ r_k^*\left(a_k \sum_{i=1}^{N} \lambda_i(1-z_i)\right)\left(\prod_{i=1}^{N} z_i^{X_k^i}\right) E\left(e^{-a_k \sum_{i=1}^{N} \lambda_i(1-z_i) B_k(X_k^k)} \middle| X_k^k, a_k\right) z_k^{-a_k X_k^k}\right\}$$

$$= \tilde{d}_k E\left\{ r_k^*\left(a_k \sum_{i=1}^{N} \lambda_i(1-z_i)\right)\left(\prod_{i=1}^{N} z_i^{X_k^i}\right) b_k^*\left(a_k \sum_{i=1}^{N} \lambda_i(1-z_i)\right)^{X_k^k} z_k^{-a_k X_k^k}\right\}$$

$$= \tilde{d}_k E\left\{ r_k^*\left(a_k \sum_{i=1}^{N} \lambda_i(1-z_i)\right) F_k\left(z_1, z_2, ..., z_{k-1}, b_k^*\left(a_k \sum_{i=1}^{N} \lambda_i(1-z_i)\right) z_k^{\bar{a}_k}, z_{k+1}, ..., z_N\right)\right\}.$$

Thus

$$F_{k+1}(\underline{z}) = p_k \tilde{d}_k \tilde{r}_k F_k\left(z_1, z_2, ..., z_{k-1}, \tilde{b}_k, z_{k+1}, ..., z_N\right) + (1-p_k)\tilde{d}_k F_k(\underline{z}) \qquad (2)$$

We assume throughout the paper that F_k have first and second derivatives. We shall also assume that the system of implicit equations (and also (15) and (16)) have a solution.

Moments of number of jobs at polling instants

The first and second moment of the number of jobs at polling instants are obtained by differentiating the generating functions $F_k(\underline{z})$. We calculate

$$f_k(i) \overset{\text{def}}{=} \left.\frac{\partial F_k(\underline{z})}{\partial z_i}\right|_{\underline{z}=\{1,...,1\}}, \qquad f_k(l,i) \overset{\text{def}}{=} \left.\frac{\partial^2 F_k(\underline{z})}{\partial z_l \partial z_i}\right|_{\underline{z}=\{1,...,1\}}.$$

It follows that $f_k(i) = E[X_k^i]$, $f_k(l,i) = E[X_k^l X_k^i]$ if k, l, i are not all equal, and $f_i(i,i) = E[X_i^i(X_i^i - 1)]$.

By differentiating (2) we obtain the following set of $N+N(N-1)=N^2$ linear equations:

$$f_{k+1}(k)=\lambda_k\bar{d}_k+[p_k\rho_k+(1-p_k)]f_k(k), \qquad k=1,2,...,N \tag{3}$$

$$f_{k+1}(i)=\lambda_i\bar{d}_k+f_k(i)+p_k\lambda_i b_k f_k(k), \qquad i,k=1,2,...,N,\ i\neq k \tag{4}$$

(where $\bar{d}_k=d_k+p_k r_k$). Explicit expressions can now be derived for $f_k(k)$ following the method in [15]. From (4) we obtain for $i\neq k$

$$f_{k+1}(i)-f_k(i)=\lambda_i[\bar{d}_k+p_k b_k f_k(k)]$$

Summing from $k=j$ to $i-1$ we get

$$f_i(i)-f_j(i)=\lambda_i[\sum_{k=j}^{i-1}\bar{d}_k+\sum_{k=j}^{i-1}p_k b_k f_k(k)]$$

Substituting $j=i+1$ in the above expression, and using (3) for $f_{i+1}(i)$ we have

$$p_i(1-\rho_i)f_i(i)=\lambda_i\left[\sum_{k=1}^{N}\bar{d}_k+\sum_{\substack{k=1\\k\neq i}}^{N}p_k b_k f_k(k)\right] \tag{5}$$

Let $\bar{f}_i(i)\overset{\text{def}}{=}p_i(1-\rho_i)f_i(i)$. Then, from (5)

$$\bar{f}_i(i)=\lambda_i\left[\sum_{k=1}^{N}\bar{d}_k+\sum_{\substack{k=1\\k\neq i}}^{N}\frac{b_k}{1-\rho_k}\bar{f}_k(k)\right] \tag{6}$$

which is the *same* equation that satisfies $f_k(k)$ in the cyclic (nonrandom) *exhaustive* service discipline (see [15]). Hence

$$\bar{f}_i(i)=\frac{\lambda_i(1-\rho_i)\bar{d}}{1-\rho} \tag{7}$$

and thus

$$f_i(i)=\frac{\lambda_i\bar{d}}{p_i(1-\rho)} \tag{8}$$

That is, the expected number of jobs, present at queue i when it is polled, is $[p_i(1-\rho_i)]^{-1}$ times greater than that in a regular cyclic *exhaustive* regime for which $d=\bar{d}$, and is p_i^{-1} times greater than that in a regular cyclic *gated* regime (with $d=\bar{d}$).

The second moments are obtained by solving the following set of N^3 linear equations:

$$\begin{aligned}
f_{k+1}(i,l) &= p_k\{\lambda_i\lambda_l[d_k^{(2)}+2d_kr_k+r_k^{(2)}]+(d_k+r_k)[\lambda_l f_k(i)+\lambda_i f_k(l)] \\
&+ \quad f_k(k)\lambda_i\lambda_l[2(d_k+r_k)b_k+b_k^{(2)}]+f_k(i,l)+b_k\lambda_l f_k(k,i)+b_k\lambda_i f_k(k,l)+b_k^2\lambda_i\lambda_l f_k(k,k)\} \\
&+ (1-p_k)\{\lambda_i\lambda_l d_k^{(2)}-[\lambda_i f_k(l)+\lambda_l f_k(i)]d_k+f_k(i,l)\} \qquad k\neq i,\ i\neq l
\end{aligned} \tag{9}$$

$$\begin{aligned}
f_{k+1}(k,l) &= p_k\{\lambda_k\lambda_l[d_k^{(2)}+2d_kr_k+r_k^{(2)}]+(d_k+r_k)\lambda_k f_k(l) \\
&+ \qquad f_k(k)\lambda_k\lambda_l[2(d_k+r_k)b_k+b_k^{(2)}]+\lambda_k b_k f_k(k,l)+b_k^2\lambda_k\lambda_l f_k(k,k)\} \\
&+ (1-p_k)\{\lambda_k\lambda_l d_k^{(2)}-[\lambda_k f_k(l)+\lambda_l f_k(k)]d_k+f_k(k,l)\} \qquad k\neq l
\end{aligned}$$

$$\begin{aligned}
f_{k+1}(k,k) &= p_k\{\lambda_k^2[d_k^{(2)}+2d_kr_k+r_k^{(2)}]+f_k(k)\lambda_k^2[2(d_k+r_k)b_k+b_k^{(2)}]+b_k^2\lambda_k^2 f_k(k,k)\} \\
&+ (1-p_k)\{\lambda_k^2 d_k^{(2)}-2\lambda_k d_k f_k(k)+f_k(k,k)\}
\end{aligned}$$

For the efficient solution of the equations above, one may refer to [14].

Cycle duration

Let us define a cycle C to be the typical time in steady state between two consecutive arrival instants of the server to some given queue (say queue 1). Then the expected cycle duration E[C] in steady state satisfies

$$E[C]=\frac{\bar{d}}{1-\rho}$$

(see [2] Proposition 5.2 or [10] Proposition 3). An alternative derivation of $E[C]$ can be obtained through

$$E[C]=\bar{d}+\sum_{i=1}^{N}p_ib_if_i(i)$$

Waiting Times

Following Takagi [15], we define the following random variables (in system's steady-state):

$L_i(n) \overset{\text{def}}{=}$ number of jobs that the nth departing job from station i (counting from the moment that the station was last polled) leaves behind it, and

$L_i \overset{\text{def}}{=}$ number of jobs that an arbitrary departing job from station i leaves behind it.
Also, let T_i be the (random) number of jobs served in queue i in a typical cycle, C.

We shall obtain below explicit expressions for the expected waiting times in the different queues in steady state (at an arbitrary time) in terms of $f_i(i,i)$ and $f_i(i)$, and we shall express the LST of the waiting times (in steady state) in the different queues in terms of $F_i(z)$ (which are the solutions of the implicit equations (2)). To obtain this we use the moment generating function $Q_i(z) \overset{\text{def}}{=} E\left[z^{L_i}\right]$. As the distributions of number of jobs in the system at epochs of arrivals and epochs of departures are identical (see c.f. Kleinrock [11] pp. 232) then, by the well known PASTA (Poisson Arrivals See Time Averages) phenomenon (see c.f. Wolff [18]), $Q_i(z)$ also stands for the moment generating function of the number of jobs at station i in steady state regime at an arbitrary point in time. We have, as in Takagi [15] pp. 77-79, 109, and Khamisy et al [10]

$$Q_i(z) = \frac{E\left(\sum_{n=1}^{T_i} z^{L_i(n)}\right)}{E\left(T_i\right)} = \frac{p_i E\left(\sum_{n=1}^{X_i^i} z^{L_i(n)} \middle| a_i = 1\right)}{p_i E\left(X_i^i\right)} = \frac{E\left(\sum_{n=1}^{X_i^i} z^{L_i(n)}\right)}{E\left(X_i^i\right)}$$

With some abuse of notation, denote $F_i(z) = F_i(1,1,...,1,z,1,...,1)$ where z stands in the ith place. Set $\bar{b}_i \overset{\text{def}}{=} b_i^*(\lambda_i - \lambda_i z)$. The evaluation of the expression for $Q_i(z)$ is almost the same as in [15] p. 109. In our case we have $L_i(n) = X_i^i + A_i(R_i) - n + A_i(B_i(n))$ (for the case $a_i = 1$), whereas in [15] p. 109. $L_i(n) = X_i^i - n + A_i(B_i(n))$. Thus the result defers from the one in [15] by an extra term that expresses the number of jobs that arrived during the switch-in time. Hence

$$Q_i(z) = \frac{\bar{b}_i}{E(X_i^i)(z - \bar{b}_i)} \left\{ E\left[z^{X_i^i} - \bar{b}_i^{X_i^i} \right] \right\} \times r_i^*(\lambda_i(1-z)) \tag{10}$$

$$= \frac{p_i(1-\rho)\bar{b}_i}{\lambda_i \bar{d}(z - \bar{b}_i)} \left\{ F_i(z) - F_i(\bar{b}_i) \right\} \times r_i^*(\lambda_i(1-z)),$$

from which, by differentiation, we derive

$$E[L_i] = \rho_i + \frac{\left(E[(X_i^i)^2] - E[X_i^i]\right)(1+\rho_i)}{2E[X_i^i]} + \lambda_i r_i = \rho_i + \frac{f_i(i,i)(1+\rho_i)}{2f_i(i)} + \lambda_i r_i$$

where $f_i(i,i)$ is obtained by solving equations (9).

The LST and expectation of the waiting time W_i of an arbitrary job at queue i are obtained using the relations

$$W_i^*(\lambda_i - \lambda_i z) b_i^*(\lambda_i - \lambda_i z) = Q_i(z), \qquad \lambda_i E[W_i] + \lambda_i b_i = E[L_i]$$

This finally yields

$$W_i^*(s) = \frac{Q_i(1 - s/\lambda_i)}{b_i^*(s)} \tag{11}$$

$$E[W_i] = \frac{E[L_i]}{\lambda_i} - b_i \tag{12}$$

3 The Exhaustive Discipline

We analyse **two** versions of the exhaustive regime: (i) the **partially exhaustive** (PE) regime, where the server, upon switching into queue i (with probability p_i) stays there (after switch-in time is incurred) for the duration of X_i^i busy periods. In terms of *number* of jobs, this is equivalent to serving the X_i^i jobs that were present there before the switch-in time, plus all those that have arrived during the service in that queue, whereas jobs that arrive during the switch-in time are not served during the current visit; (ii) the **fully exhaustive** (FE) regime, where the server leaves a queue only after it is empty (where upon serving all the jobs that were present there before the switch-in time, plus all those that have arrived during its sojourn time in that queue). In this case, jobs that arrive during the switch-in time are served during the current visit.

With the same notation used in the previous section, the evolution of the *state* of the system (in steady state) is given by

$$\text{PE}: \quad X_{k+1}^i \overset{\text{d}}{=} \begin{cases} X_k^i + A_i\left(a_k[R_k + \Omega_k(X_k^k)] + D_k\right) & i \neq k \\ X_k^k \bar{a}_k + A_i\left(a_k R_k + D_k\right) & i = k \end{cases} \tag{13}$$

$$\text{FE}: \quad X_{k+1}^i \overset{\text{d}}{=} \begin{cases} X_k^i + A_i\left(a_k[R_k + \Omega_k(A_k(R_k) + X_k^k)] + D_k\right) & i \neq k \\ X_k^k \bar{a}_k + A_i\left(D_k\right) & i = k \end{cases} \tag{14}$$

where $1 \leq i, k \leq N$.

Let $F_k(\underline{z})$, $\tilde{r}_k$ and $\tilde{d}_k$ be as in the previous section, and define $\tilde{\omega}_k \overset{\text{def}}{=} \omega_k^* \left(\sum_{\substack{i=1 \\ i \neq k}}^N \lambda_i (1 - z_i) \right)$, and $\tilde{r}_k^E \overset{\text{def}}{=} r_k^* \left(\lambda_k - \lambda_k \tilde{w}_k + \sum_{\substack{i=1 \\ i \neq k}}^N \lambda_i (1 - z_i) \right)$. Using the evolution equations we obtain

$$\text{PE}: \quad F_{k+1}(\underline{z}) = E \left\{ z_k^{\bar{a}_k X_k^k + A_k(a_k R_k + D_k)} \prod_{\substack{i = 1 \\ i \neq k}}^N z_i^{X_k^i + A_i\left(a_k[R_k + \Omega_k(X_k^k)] + D_k\right)} \right\}$$

$$= E \left[\left(\prod_{i=1}^N z_i^{X_k^i} \right) e^{-\lambda_k(1-z_k)(a_k R_k + D_k)} \left(\prod_{\substack{i=1 \\ i \neq k}}^N e^{-\lambda_i(1-z_i)\left(a_k[R_k + \Omega_k(X_k^k)] + D_k\right)} \right) z_k^{-a_k X_k^k} \right]$$

$$= \tilde{d}_k E \left\{ r_k^* \left(a_k \sum_{i=1}^N \lambda_i (1 - z_i) \right) \left(\prod_{i=1}^N z_i^{X_k^i} \right) E \left(e^{-a_k \sum_{\substack{i=1 \\ i \neq k}}^N \lambda_i (1-z_i) \Omega_k(X_k^k)} \middle| X_k^k, a_k \right) z_k^{-a_k X_k^k} \right\}$$

$$= \tilde{d}_k E \left\{ r_k^* \left(a_k \sum_{i=1}^N \lambda_i (1 - z_i) \right) \left(\prod_{i=1}^N z_i^{X_k^i} \right) \omega_k^* \left(a_k \sum_{\substack{i=1 \\ i \neq k}}^N \lambda_i (1 - z_i) \right)^{X_k^k} z_k^{-a_k X_k^k} \right\}$$

$$= \tilde{d}_k E \left\{ r_k^* \left(a_k \sum_{i=1}^N \lambda_i (1 - z_i) \right) F_k \left(z_1, z_2, ..., z_{k-1}, \omega_k^* \left(a_k \sum_{\substack{i=1 \\ i \neq k}}^N \lambda_i (1 - z_i) \right) z_k^{\bar{a}_k}, z_{k+1}, ..., z_N \right) \right\}$$

Thus, we have for PE:

$$F_{k+1}(\underline{z}) = p_k \tilde{d}_k \tilde{r}_k F_k \left(z_1, z_2, ..., z_{k-1}, \tilde{\omega}_k, z_{k+1}, ..., z_N \right) + (1 - p_k) \tilde{d}_k F_k(\underline{z}) \tag{15}$$

For the FE case we write

$$\text{FE}: \quad F_{k+1}(\underline{z}) = E \left\{ z_k^{\bar{a}_k X_k^k + A_k(D_k)} \prod_{\substack{i = 1 \\ i \neq k}}^N z_i^{X_k^i + A_i\left(a_k[R_k + \Omega_k(A_k(R_k) + X_k^k)] + D_k\right)} \right\}$$

$$= E \left[\left(\prod_{i=1}^N z_i^{X_k^i} \right) e^{-\lambda_k(1-z_k) D_k} \left(\prod_{\substack{i=1 \\ i \neq k}}^N e^{-\lambda_i(1-z_i)\left(a_k[R_k + \Omega_k(A_k(R_k) + X_k^k)] + D_k\right)} \right) z_k^{-a_k X_k^k} \right]$$

$$= \tilde{d}_k E \left\{ exp \left(-R_k a_k \sum_{\substack{i=1 \\ i \neq k}}^N \lambda_i (1 - z_i) \right) \left(\prod_{i=1}^N z_i^{X_k^i} \right) E \left(e^{-a_k \sum_{\substack{i=1 \\ i \neq k}}^N \lambda_i (1-z_i) \Omega_k(A_k(R_k) + X_k^k)} \middle| X_k^k, a_k \right) z_k^{-a_k X_k^k} \right\}$$

$$= \tilde{d}_k E \left\{ exp \left(-R_k a_k \sum_{\substack{i=1 \\ i\neq k}}^{N} \lambda_i (1-z_i) \right) \left(\prod_{i=1}^{N} z_i^{X_k^i} \right) \omega_k^* \left(a_k \sum_{\substack{i=1 \\ i\neq k}}^{N} \lambda_i (1-z_i) \right)^{A_k(R_k)+X_k^k} z_k^{-a_k X_k^k} \right\}$$

$$= \tilde{d}_k E \left\{ exp \left(-R_k a_k \sum_{\substack{i=1 \\ i\neq k}}^{N} \lambda_i (1-z_i) - R_k \left[\lambda_k - \lambda_k \omega_k^* \left(a_k \sum_{\substack{i=1 \\ i\neq k}}^{N} \lambda_i (1-z_i) \right) \right] \right) \right.$$

$$\left. \times \; F_k \left(z_1, z_2, ..., z_{k-1}, \omega_k^* \left(a_k \sum_{\substack{i=1 \\ i\neq k}}^{N} \lambda_i (1-z_i) \right) z_k^{\bar{a}_k}, z_{k+1}, ..., z_N \right) \right\}$$

Thus, for FE:

$$F_{k+1}(\underline{z}) = p_k \tilde{d}_k \tilde{r}_k^E F_k \left(z_1, z_2, ..., z_{k-1}, \tilde{\omega}_k, z_{k+1}, ..., z_N \right) + (1-p_k) \tilde{d}_k F_k(\underline{z}) \tag{16}$$

With the same definitions of $f_k(i)$ and $f_k(l,i)$, as in the previous section, we get by differentiating (15) and (16) the following set of N^2 linear equations. For PE:

$$f_{k+1}(k) = \lambda_k \bar{d}_k + (1-p_k) f_k(k), \qquad k = 1, ..., N \tag{17}$$

$$f_{k+1}(i) = \lambda_i \bar{d}_k + f_k(i) + p_k \lambda_i \omega_k f_k(k), \qquad i,k = 1, ..., N, \; i \neq k \tag{18}$$

For FE:

$$f_{k+1}(k) = \lambda_k d_k + (1-p_k) f_k(k), \qquad k = 1, ..., N \tag{19}$$

$$f_{k+1}(i) = \lambda_i \bar{d}_k + f_k(i) + p_k \lambda_i \omega_k (\lambda_k r_k + f_k(k)), \qquad i,k = 1, ..., N, \; i \neq k \tag{20}$$

(where $\bar{d}_k = d_k + p_k r_k$). Following the same calculations as in the previous section, we obtain

$$\text{PE}: \quad p_i f_i(i) = \lambda_i \left[\sum_{k=1}^{N} \bar{d}_k + \sum_{\substack{k = 1 \\ k \neq i}}^{N} p_k \omega_k f_k(k) \right] \tag{21}$$

$$\text{FE}: \quad p_i(\lambda_i r_i + f_i(i)) = \lambda_i \left[\sum_{k=1}^{N} \bar{d}_k + \sum_{\substack{k = 1 \\ k \neq i}}^{N} p_k \omega_k (\lambda_k r_k + f_k(k)) \right] \tag{22}$$

Define $\bar{f}_i(i) \stackrel{\text{def}}{=} p_i f_i(i)$ for PE, and $\bar{f}_i(i) \stackrel{\text{def}}{=} p_i(\lambda_i r_i + f_i(i))$ for FE. Equations (21) or (22) yield again equation (6), from which, by the same argument as in the previous section, the explicit expression (7) for $\bar{f}_i(i)$ results. This leads to

$$\text{PE}: \quad f_i(i) = \frac{\lambda_i \bar{d}(1-\rho_i)}{p_i(1-\rho)} \tag{23}$$

$$\text{FE}: \quad f_i(i) = \frac{\lambda_i \bar{d}(1-\rho_i)}{p_i(1-\rho)} - \lambda_i r_i \tag{24}$$

It is clear from standard balance arguments that $E[C] = \bar{d}/(1-\rho)$ in the two exhaustive cases as well, and thus the interpretation of expressions (23) and (24) is straightforward.

Conclusion: *the expression for $f_i(i)$ for PE is exactly p_i^{-1} times larger than the one obtained in the standard exhaustive model, with purly cyclic service (for which $d = \bar{d}$). For FE it is further smaller by $\lambda_i r_i$.*

The second moments are obtained by solving the following set of N^3 linear equations. For PE:

$$\begin{aligned}
f_{k+1}(i,l) &= p_k \Bigg\{ \lambda_i \lambda_l [d_k^{(2)} + 2 d_k r_k + r_k^{(2)}] + (d_k + r_k)[\lambda_l f_k(i) + \lambda_i f_k(l)] \\
&+ f_k(k) \lambda_i \lambda_l \left[\frac{2(d_k + r_k) b_k}{1-\rho_k} + \frac{b_k^{(2)}}{(1-\rho_k)^3} \right] + f_k(i,l) \\
&+ \frac{b_k}{1-\rho_k} [\lambda_l f_k(k,i) + \lambda_i f_k(k,l)] + \frac{b_k^2 \lambda_i \lambda_l f_k(k,k)}{(1-\rho_k)^2} \Bigg\} \\
&+ (1-p_k) \left\{ \lambda_i \lambda_l d_k^{(2)} - [\lambda_i f_k(l) + \lambda_l f_k(i)] d_k + f_k(i,l) \right\} \qquad k \neq i, \ i \neq l
\end{aligned} \tag{25}$$

$$\begin{aligned}
f_{k+1}(k,l) &= p_k \left\{ \lambda_k \lambda_l [d_k^{(2)} + 2 d_k r_k + r_k^{(2)}] + (d_k + r_k) \lambda_k \left[f_k(l) + \frac{f_k(k) \lambda_l b_k}{1-\rho_k} \right] \right\} \\
&+ (1-p_k) \{ \lambda_k \lambda_l d_k^{(2)} - [\lambda_k f_k(l) + \lambda_l f_k(k)] d_k + f_k(k,l) \} \qquad k \neq l
\end{aligned}$$

$$\begin{aligned}
f_{k+1}(k,k) &= p_k \lambda_k^2 [d_k^{(2)} + 2 d_k r_k + r_k^{(2)}] \\
&+ (1-p_k) \{ \lambda_k^2 d_k^{(2)} - 2 \lambda_k d_k f_k(k) + f_k(k,k) \}
\end{aligned}$$

Similar equations are obtained for FE.

Next we compute the waiting times for PE. Define Y_i to be the number of jobs served in queue i during the visit of the server there. With the same definition of $Q_i(z)$ as in the previous section, we have

$$Q_i(z) = \frac{E\left(\sum_{n=1}^{T_i} z^{L_i(n)}\right)}{E\left(T_i\right)} = \frac{p_i E\left(\sum_{n=1}^{Y_i} z^{L_i(n)}\middle| a_i = 1\right)}{p_i E\left(Y_i | a_i = 1\right)} = \frac{(1-\rho_i) E\left(\sum_{n=1}^{Y_i} z^{L_i(n)}\middle| a_i = 1\right)}{E\left(X_i^i\right)} \tag{26}$$

The evaluation of the expression for $Q_i(z)$ is done similarly to the one in [15] p. 79, where we have, as in the previous section, an extra term that corresponds to the number of arrivals during the switch-in time. This term stems from the fact that those arriving jobs are seen by every leaving job, since they are not served in the current cycle. Hence, with $\bar{b}$ defined as above,

$$Q_i(z) = \frac{\bar{b}_i(1-\rho_i)}{E(X_i^i)(z-\bar{b}_i)} \left(E\left[z^{X_i^i}\right] - 1\right) \times r_i^*(\lambda_i(1-z)) \tag{27}$$

$$= \frac{p_i(1-\rho)\bar{b}_i}{\lambda_i \bar{d}(z-\bar{b}_i)} \left(F_i(z) - 1\right) \times r_i^*(\lambda_i(1-z)),$$

from which, by differentiation, we derive

$$E[L_i] = \rho_i + \frac{\lambda_i^2 b_i^{(2)}}{2(1-\rho_i)} + \frac{f_i(i,i)}{2f_i(i)} + \lambda_i r_i. \tag{28}$$

$f_i(i,i)$ is obtained by solving equations (25).

The LST and expectation of the waiting time W_i of an arbitrary job at queue i are obtained using the relations (11) and (12).

For FE, the first two equalities in (26) still hold. However, Y_i, given $a_i = 1$, is equal to the jobs served during $X_i^i + A_i(R_i)$ busy periods. Hence

$$E(Y_i | a_i = 1) = \frac{E(X_i^i) + \lambda_i r_i}{1 - \rho_i},$$

and instead of (27) we have (using result (24))

$$Q_i(z) = \frac{\bar{b}_i(1-\rho_i)}{[E(X_i^i) + \lambda_i r_i](z-\bar{b}_i)} \left(E\left[z^{X_i^i}\right] \times r_i^*(\lambda_i(1-z)) - 1\right) \tag{29}$$

$$= \frac{p_i(1-\rho)\bar{b}_i}{\lambda_i \bar{d}(z-\bar{b}_i)} \left(F_i(z) \times r_i^*(\lambda_i(1-z)) - 1\right),$$

Note that the reason for the term r_i^* in (29) is different than in (27). In the FE case it is due to the fact that the server stays in the queue $X_i^i + A_i(R_i)$ busy periods (rather than X_i^i as in the PE case). By differentiation, we obtain

$$E[L_i] = \rho_i + \frac{\lambda_i^2 b_i^{(2)}}{2(1-\rho_i)} + \frac{f_i(i,i) + 2f_i(i)\lambda_i r_i + \lambda_i^2 b^{(2)}}{2\left(f_i(i) + \lambda_i r_i\right)} \tag{30}$$

4 Pseudo-conservation law and optimization

We consider below a mixed system, where some queues may have the gated service discipline and others follow one of the exhaustive regimes. Of interest is the optimization problem of choosing the switch-in probabilities p_i, $i = 1, ..., N$, so as to **minimize the expected workload in the system,** $\sum_{i=1}^N b_i E[L_i] = \sum_{i=1}^N \rho_i E[W_i]$. To this end, we use the expression for the decomposition of the workload in polling systems given by Boxma [4] and Boxma and Groenendijk [5], known as pseudo-conservation laws. From these references

$$\sum_{i=1}^N \rho_i E[W_i] = \rho \frac{\sum_{i=1}^N \lambda_i b_i^{(2)}}{2(1-\rho)} + \rho \frac{\bar{d}^{(2)}}{2\bar{d}} + \frac{\rho^2 - \sum_{i=1}^N \rho_i^2}{2(1-\rho)} \bar{d} + \sum_{i=1}^N EM_i^{(1)} \tag{31}$$

where $EM_i^{(1)}$ is the expected unfinished work in the ith queue at an arbitrary instant of departure of the server from that queue. In our case, a departure instant from the ith queue is the time at which the server starts moving from that queue to the next one, regardless of whether service was actually given there or not. For the case of Gated service, with probability p_i the jobs present at such instant are those that arrived during the period comprised of the switch-in time plus the service time devoted to that queue. The expectation of this number is $\lambda_i(r_i + b_i f_i(i))$. In the Partially Exhaustive case, with probability p_i the jobs present at such instant are only those that arrived during the switch-in time, the expectation of which is $\lambda_i r_i$. For the FE case, this term is zero. With probability $1 - p_i$ the number found at departure instant is the same number of jobs found at the moment of server's arrival to the station (both in the gated and in the exhaustive case). Thus

$$EM_i^{(1)}(Gated) = b_i\left[p_i\lambda_i(r_i + b_i f_i(i)) + (1-p_i)f_i(i)\right] = \rho_i\left[p_i r_i - \frac{1-\rho_i}{1-\rho}\bar{d} + \frac{\bar{d}}{p_i(1-\rho)}\right] \tag{32}$$

$$EM_i^{(1)}(PE) = b_i\left[p_i\lambda_i r_i + (1-p_i)f_i(i)\right] = \rho_i\left[p_i r_i - \frac{1-\rho_i}{1-\rho}\bar{d} + \frac{\bar{d}(1-\rho_i)}{p_i(1-\rho)}\right] \tag{33}$$

$$EM_i^{(1)}(FE) = b_i(1-p_i)f_i(i) = \rho_i\left[-\frac{1-\rho_i}{1-\rho}\bar{d} + \frac{\bar{d}(1-\rho_i)}{p_i(1-\rho)} - r_i(1-p_i)\right] = EM_i^{(1)}(PE) - \rho_i r_i \tag{34}$$

Let G $(PE,\ FE)$ denote the set of queues that are served according to the Gated (Partially Exhaustive, Fully Exhaustive) discipline, respectively. Substituting (32) (33) and (34) in (31) yield the following conservation law:

$$\begin{aligned}\sum_{i=1}^{N}\rho_i E[W_i] &= \rho\frac{\sum_{i=1}^{N}\lambda_i b_i^{(2)}}{2(1-\rho)} + \rho\frac{\bar{d}^{(2)}}{2\bar{d}} + \frac{\rho^2-\sum_{i=1}^{N}\rho_i^2}{2(1-\rho)}\bar{d} \\ &+ \sum_{i\in G}\rho_i\left[p_i r_i - \frac{1-\rho_i}{1-\rho}\bar{d} + \frac{\bar{d}}{p_i(1-\rho)}\right] + \sum_{i\in PE}\rho_i\left[p_i r_i - \frac{1-\rho_i}{1-\rho}\bar{d} + \frac{\bar{d}(1-\rho_i)}{p_i(1-\rho)}\right] \\ &+ \sum_{i\in FE}\rho_i\left[-\frac{1-\rho_i}{1-\rho}\bar{d} + \frac{\bar{d}(1-\rho_i)}{p_i(1-\rho)} - r_i(1-p_i)\right] \end{aligned} \tag{35}$$

We wish to express the latter as a function of the parameters p_i. To do so, we note that

$$\bar{d} = \sum_{i=1}^{N}(d_i + p_i r_i)$$

$$\bar{d}^{(2)} = d^{(2)} + 2d\sum_{i=1}^{N}p_i r_i + \sum_{i=1}^{N}p_i r_i^{(2)} + \sum_{\substack{i=1\\ i\neq j}}^{N} p_i p_j r_i r_j$$

and hence

$$\begin{aligned} Z(\underline{p}) &\overset{\text{def}}{=} \sum_{i=1}^{N}\rho_i E[W_i] \qquad (36) \\ &= \rho\frac{\sum_{i=1}^{N}\lambda_i b_i^{(2)}}{2(1-\rho)} \\ &+ \rho\left(\frac{d^{(2)} + 2d\sum_{i=1}^{N}p_i r_i + \sum_{i=1}^{N}p_i r_i^{(2)} + \sum_{\substack{i=1\\ i\neq j}}^{N} p_i p_j r_i r_j}{2\sum_{i=1}^{N}(d_i + p_i r_i)}\right) \\ &+ \frac{\rho^2-\sum_{i=1}^{N}\rho_i^2}{2(1-\rho)}\sum_{i=1}^{N}(d_i + p_i r_i) \\ &+ \sum_{i\in G}\rho_i\left[p_i r_i + \frac{\sum_{j=1}^{N}(d_j + p_j r_j)}{(1-\rho)}\left(\frac{1}{p_i} - (1-\rho_i)\right)\right] \end{aligned}$$

$$+ \sum_{i \in PE} \rho_i \left[p_i r_i + \frac{\sum_{j=1}^{N}(d_j + p_j r_j)(1-\rho_i)}{(1-\rho)} \left(\frac{1}{p_i} - 1 \right) \right]$$

$$+ \sum_{i \in FE} \rho_i \left[p_i r_i + \frac{\sum_{j=1}^{N}(d_j + p_j r_j)(1-\rho_i)}{(1-\rho)} \left(\frac{1}{p_i} - 1 \right) - r_i \right]$$

Now, the optimization becomes the following Mathematical Program (see also Section 4 in [6]:

Find a vector $\underline{p} = \{p_1, p_2, ..., p_N\}$ that minimizes $Z(\underline{p})$ subject to $0 \leq p_i \leq 1, \ i = 1, ..., N$.

Since for every i,

$$\left[\frac{1}{p_i} - (1-\rho_i) \right] > \left[(1-\rho_i) \left(\frac{1}{p_i} - 1 \right) \right] > \left[(1-\rho_i) \left(\frac{1}{p_i} - 1 \right) - r_i \right]$$

it readily follows from Eq. (36) that for any fixed vector of switch-in probabilities $\{p_i\}$, and for **each** station, independently of the others, the expected workload when using PE is **smaller** than when using G; and this performance-measure is even smaller when applying the FE regime. This is also a direct consequence of (32), (33) and (34). As a result, the **best** performance among all choices of service disciplines in different stations and of switching probabilities is obtained when the Fully Exhaustive service regime is applied in all stations and the **optimal** switch-in probabilities obtained through the respective Mathematical Program are used.

Clearly, for all i satisfying $\rho_i > 0$, the optimal p_i has to be greater than zero. It seems reasonable to expect that for a queue with a low arrival rate we would get $p_i < 1$, so as to avoid the switch-in time to a queue that might be empty. This would allow the server to be more frequently available for queues with higher arrival rates. We could also expect that $p_i = 1$ for that queue i whose arrival rate λ_i, or whose ρ_i, is the highest. However, as we show in the sequal, even in the case when there is only a single queue, it is not always advantageous to have $p_1 = 1$.

Optimization of a single queue

In the case of a single queue, (36) reduces to

$$E[W_{gated}] = \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{d^{(2)} + 2drp + pr^{(2)}}{2(d+pr)} - d + \frac{rp + d}{p(1-\rho)} \tag{37}$$

$$= \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{d^{(2)} - 2d^2 + pr^{(2)}}{2(d+pr)} + \frac{rp + d}{p(1-\rho)}$$

and similarly

$$E[W_{PE}] = \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{d^{(2)} + 2drp + pr^{(2)}}{2(d+pr)} - d + \frac{rp+d}{p} \tag{38}$$

$$= \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{d^{(2)} - 2d^2 + pr^{(2)}}{2(d+pr)} + \frac{rp+d}{p}$$

$$E[W_{FE}] = \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{d^{(2)} + 2drp + pr^{(2)}}{2(d+pr)} - d + \frac{d}{p} \tag{39}$$

$$= \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{d^{(2)} - 2d^2 + pr^{(2)}}{2(d+pr)} + \frac{d}{p} = E[W_{PE}] - r$$

where the index 1 has been omitted. Set $q \overset{\text{def}}{=} p^{-1}$. The various $E[W]$ are now written as

$$E[W_{gated}] = const + \frac{d^{(2)}q - 2d^2q + r^{(2)}}{2(dq+r)} + \frac{1}{(1-\rho)}dq \tag{40}$$

and similarly

$$E[W_{PE}] = const + \frac{d^{(2)}q - 2d^2q + r^{(2)}}{2(dq+r)} + dq \tag{41}$$

Since $E[W_{FE}] = E[W_{PE}] - r$, $E[W_{FE}]$ it is clearly expressed by (41) too. The results described below for PE thus hold for FE too. Define $\hat{p}_{gated} \overset{\text{def}}{=} p$ that minimizes $E[W_{gated}]$ and $\hat{p}_{PE} \overset{\text{def}}{=} p$ that minimizes $E[W_{PE}]$. Further define

$$q^*_{gated} = \frac{1}{d}\left[\sqrt{\frac{r^{(2)}d + 2d^2r - d^{(2)}r}{2\frac{1}{(1-\rho)}d}} - r\right] \tag{42}$$

and

$$q^*_{exh} = \frac{1}{d}\left[\sqrt{\frac{r^{(2)}d + 2d^2r - d^{(2)}r}{2d}} - r\right] \tag{43}$$

proposition 1 *If q^*_{gated} is a real number and satisfies $q^*_{gated} \geq 1$ then $\hat{p}_{gated}$ is given by $\hat{p}_{gated} = [q^*_{gated}]^{-1}$. Otherwize $\hat{p}_{gated} = 1$. The same holds for the PE case.*

Proof: Consider first the Gated case. Looking for a real number q^* that achieves

$$\frac{\partial E[W_{gated}]}{\partial q} = 0 \tag{44}$$

we get

$$\frac{\partial E[W_{gated}]}{\partial q} = \frac{d^{(2)}r - 2d^2r - r^{(2)}d}{2(dq+r)^2} + \frac{1}{(1-\rho)}d$$

Condition (44) then yields

$$2(dq^* + r)^2 = \frac{r^{(2)}d + 2d^2r - d^{(2)}r}{\frac{1}{(1-\rho)}d}$$

and the proposition readily follows. The PE case is obtained similarly. ■

If we choose $p = 1$ (i.e. $q = 1$) we shall say that a "non-idling policy" is being used, whereas "idling policy" will stand for any choice of $p < 1$ (hence $q > 1$). In the following Corollaries we present conditions for the optimality of idling and non-idling policies.

corollary 2 *If the Gated discipline is used, then the optimal policy is idling iff both conditions are satisfied:*

(i) $r^{(2)}d + 2d^2r - d^{(2)}r > 0$, *and*

(ii) $\rho < 1 - \frac{2d(r+d)^2}{r^{(2)}d + 2d^2r - d^{(2)}r}$

If the PE discipline is used, then the optimal policy is idling iff

$$r^{(2)} > 2d^2 + 2r^2 + 2dr + d^{(2)}r/d$$

Proof: From (42) it is easily seen that $q^* > 1$ is equivalent to

$$r^{(2)}d + 2d^2r - d^{(2)}r > (d+r)^2\frac{2d}{1-\rho} \tag{45}$$

(note that (45) implies that the term inside the square-root is positive). (45) is easily seen to be equivalent to the conditions of the proposition. The PE case is obtained similarly. ■

corollary 3 *(i) In the case of gated service discipline, a sufficient condition for the optimal policy to be non-idling is that* $2r^2 + 2d^2 + 2dr \geq r^{(2)} - d^{(2)}r/d$. *(ii) Under either the Gated or the PE service discipline, if* R *is either deterministic or exponentially distributed, then the optimal policy is non-idling.*

Proof: If the sufficient condition is fulfilled then

$$1 < \frac{2d(r+d)^2}{r^{(2)}d + 2d^2r - d^{(2)}r}$$

hence by corollary 2, $\hat{p} < 1$ only if $\rho < 0$, which can never happen. (ii) then follows from (i) for the Gated case, and from Corollary 2 for the PE case, since $r^{(2)} = r^2$ in the deterministic case and $r^{(2)} = 2r^2$ in the exponential case. ■

Interpretation of corollary 2 and the conservation law:

It can be seen from (45) (and corollary 3) that $\hat{p}$ is less than one if the variance of R is large enough. To understand the fact that $p = 1$ may not be optimal, we present another viewpoint on the system (with a single queue), which allows the derivation of $E[W]$ in an alternative way.

Define a "generalized cycle" as the time between two consecutive visits of the server to the queue, at which it 'decides' to switch in and give service. A generalized cycle is thus composed of a switch-in period R, service of jobs (if the queue is not empty), and a geometric number of walking times, all distributed like D. A "generalized vacation" is then defined as $U \overset{\mathrm{d}}{=} \sum_{i=1}^{K} D^i + R$, where K is geometrically distributed with parameter p, and $D^i, i = 1, 2, ..., K$ are i.i.d. versions of D. It then follows that

$$E[U] = \frac{d}{p} + r \tag{46}$$

$$E[U^2] = \frac{d^{(2)}}{p} + \frac{2d}{p}\left[\frac{(1-p)d}{p} + r\right] + r^{(2)} \tag{47}$$

By standard balance arguments, the expected duration of a generalized cycle (for all three service disciplines) is given by

$$E[C_{generalized}] = \frac{E[U]}{1-\rho} = \frac{d+pr}{p(1-\rho)}$$

One can then use standard decomposition for obtaining the expected waiting time $E[W]$ (e.g. [4, 5]):

$$E[W] = \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{E[U^2]}{2E[U]} + \frac{E[M]}{\rho} \tag{48}$$

The second term above, $\frac{E[U^2]}{2E[U]}$, is the expected residual time of a generalized vacation; note that

$$\frac{E[U^2]}{2E[U]} = \frac{d^{(2)} - 2d^2 + r^{(2)}p}{2(d+pr)} + \frac{d}{p} \tag{49}$$

$E[M]$ is the total expected work at the departure instant of the server from the queue (after it decided to serve it). Hence

$$E[M_{gated}] = \rho^2 E[C_{generalized}] + r\rho, \qquad E[M_{PE}] = r\rho, \qquad E[M_{FE}] = 0 \tag{50}$$

(for the gated case, note that $\rho E[C_{generalized}]$ is the expected time that the server is busy during a generalized cycle, and $\rho^2 E[C_{generalized}]$ is thus the work that arrives to the system during the service time in a generalized cycle. The second term, $r\rho$, is the work that arrives during a switch-in time).

Substituting (49) and (50) into (48) yields the same expressions for $E[W]$ as obtained in (37), (38) and (39).

When the variance of the switch-in time R is large enough, it can be seen from (49) that by taking $p < 1$ we may diminish the expected residual time of a generalized vacation (in comparison to the case where $p = 1$), and hence diminish $E[W]$ if ρ is not too large (see (48)). This (partially) explains the conditions presented in corollary 2 for $\hat{p} < 1$.

References

[1] E. Altman, A. Khamisy, U. Yechiali, “Threshold Service Policies in Polling Systems”, preprint.

[2] E. Altman, P. Konstantopoulos, 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.

[3] J. E. Baker, I. Rubin, “Polling with a General-Service Order Table”, *IEEE Transactions on Communications*, **35**, pp. 283-288, 1987.

[4] O. J. Boxma, “Workloads and Waiting Times in Single-Server Systems with Multiple Customer Classes”, *Queuing Systems*, **5**, pp. 185-214, 1989.

[5] O. J. Boxma, W. P. Groenendijk, “Pseudo-Conservation Laws in Cyclic-Service Systems”, *Journal of Applied Probability,* **24**, pp. 949-964, 1987.

[6] O. J. Boxma, ”Analysis and Optimization of Polling Systems”, in *Queueing Performance and Control of ATM* (J. W. Cohen and C. D. Pack, Eds.), North Holland, pp. 173-183, 1991.

[7] O. J. Boxma, J. A. Westrate, “Waiting Times in Polling Systems with Markovian Server Routing”, in *Messung, Modellierung und Bewertung von Rechensystemen und Netzen*, (G. Stiege and J. S. Lie, Eds.), Springer, pp. 89-104, 1989.

[8] S. Browne, U. Yechiali, “Dynamic Priority Rules for Cyclic-Type Queues”, *Advances in Applied Probability*, **21**, pp. 432-450, 1989.

[9] M. J. Ferguson, Y. J. Aminetzah, “Exact Results for Nonsymmetric Token Ring Systems”, *IEEE Transactions on Communications,* **33**, No. 3, pp. 223-231, 1985.

[10] A. Khamisy, E. Altman, M. Sidi, “Polling Systems with Synchronization Constraints”, *Annals of OR*, special issue on Stochastic Modelling of Telecommunication Systems, Eds. P. Nain and K. W. Ross, pp. 231-267, 1992.

[11] L. Kleinrock, *Queuing Systems, Volume II: Computer Applications*, John Wiley, New York, 1976.

[12] L. Kleinrock, H. Levy, “The analysis of random polling systems”, *Operations Research*, **36**, pp. 716-732, 1988.

[13] H. Levy, “Analysis of Binomial-Gated Service”, in *Proceedings of Performance of Distributed and Parallel Systems*, (T. Hasegawa, H. Takagi, Y. Takahashi, Eds.), North-Holland, pp. 127-139, 1989.

[14] H. Levy, “Delay Computation and Dynamic Behavior of Non-symmetric Polling Systems”, *Performance Evaluation* **10**, pp. 35-51, 1989.

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

[16] H. Takagi, “Queuing Analysis of Polling Models: an Update”, in *Stochastic Analysis of Computer and Communications Systems*, (H. Takagi, Ed.), Elsevier Science Pub., pp. 267-318, 1990.

[17] Tedijanto, “Exact Results for the Cyclic Service Queue with a Bernoulli Schedule”, *Performance Evaluation*, **11**, pp. 107-115, 1990.

[18] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*, Prentice Hall, 1989.

[19] U. Yechiali, “Optimal Dynamic Control of Polling Systems”, in *Queueing, Performance and Control in ATM*, (J. W. Cohen and C. D. Pack, Eds.), North Holland, pp. 205-217, 1991.