DYNAMIC SERVER ROUTING IN BINOMIAL-GATED, BINOMIAL-EXHAUSTIVE, BERNOULLI-GATED, BERNOULLI-EXHUSTIVE AND MIXED POLLING SYSTEMS

Sid Browne and **Uri Yechiali**

Graduate School of Business
Columbia University
New York, N.Y. 10027

Department of Statistics
Tel Aviv University
Tel Aviv 69978, Israel

Abstract

Polling systems have been the subject of extensive research in recent years with the analyses focusing on evaluating performance measures of fixed-template routing schemes under the Exhaustive, Gated or Limited service disciplines. Optimal server routing procedures were only recently studied and dynamic policies derived (Browne and Yechiali [l988], [l989]) for systems where either all channels are of the Exhaustive type, or all channels follow the Gated regime.

Recently various probabilistic (yet static) service disciplines – the Binomial-Gated and the Binomial-Exhaustive – were proposed to help deal with the control of polling systems by assigning different service proportions to distinct channels. In this work we introduce two new variations of service regimes – the Bernoulli-Gated and the Bernoulli-Exhaustive – and consider the problem of *dynamically* control the server's transitions from one channel to another. We extend our previous results on optimal *dynamic* server routing policies to cover *all* four systems, as well as systems with mixed sets of channels. The policies derived are easy to implement – being of Dynamic Allocation type (Gittins index), they are "fair" in the sense of preserving the cyclic nature of the polling systems, and they are applicable and make sense even in the case of non-stable systems.

August 1989
Revised June 1990

1. Introduction

Research on polling systems has previously focused on evaluating and computing performance measures for fixed-template routing schemes under three main service disciplines: the Exhaustive, Gated, or Limited service regimes. Most of the results have been summarized in a book (Takagi [5]), and in surveys (Watson [7], Takagi [6]).

Recently, the study of optimal *control procedures* for polling systems has been broached in the literature. The authors, in previous works ([2], [3]), considered the problem of optimal *dynamic* server routing in fully Gated or fully Exhaustive systems and found the optimal policies to be of simple index-form that allowed direct implementation for the adaptive control of the systems. For the *static* control problem H. Levy [4] introduced and analyzed a new service policy, called *Binomial-Gated*, where, as usual, a single server attends cyclically K distinct channels, each of which being of the $M/G/1$ - queue type, but if the server finds m_i customers present in queue i, he serves there only a *random* number of customers, N_i, where N_i is Binomially distributed with parameters m_i and p_i. That is, according to this policy, the server renders service (on the average) to only a *fraction* p_i of the customers present at the moment he enters queue i. This discipline was introduced to allow for pseudo-prioritization of the stations in that the higher priority queues will be assigned higher p_i's, helping to reduce response times for higher priority customers.

Another type of "fractional service" discipline is the so-called Binomial-Exhaustive. This policy was suggested by W. P. Groenendijk and presented by O. Boxma [1] who derived pseudo-conservation laws for the two Binomial-type disciplines, as well as for various other regimes. This policy limits the attendance time of the server at queue i to only N_i *busy periods*, where, as in the Binomial-Gated case, N_i is Binomially distributed with parameters m_i and p_i, such that $E[N_i|m_i] = m_i p_i$.

In this paper we introduce two new service regimes which are also of the "fractional"-type – the Bernoulli-Gated and the Bernoulli-Exhaustive. These policies may be visualized as follows: each time the server reaches channel i a coin with probability of "success" p_i is flipped and the server enters the channel – to serve either regular Gated or regular Exhaustive, respectively, – only if the outcome is successful. These schemes preserve the random nature of the Binomial-Gated and the Binomial-Exhaustive disciplines, but in cases where switch-in or switch-out times are involved, it reduces those 'overhead' losses by allowing the server to make his decision *before* entering the channel.

We provide complete probabilistic analyses for the procedures described above, and blend the

static and *dynamic control aspects* by extending our previous results on *optimal dynamic* polling schemes to cover *all* four systems, as well as systems with mixed service disciplines. The main characteristic of our dynamic polling scheme is that it minimizes in each round the expected length of the new cycle to be traversed by the server. According to these policies the server can visit each channel *only once* in a cycle, but the *order* in which the queues are visited may change from one cycle to another, depending on the *state* of the system at the beginning of the cycle. This, in effect, changes priorities among the various queues in response to the dynamic evolution of the process. By considering dynamic routing in the Binomial or Bernoulli type systems the server is free in each new cycle to optimize his path, while the priorities of customers may be dealt with by the choice of the "success" probabilities. For further discussion of the implications of policies where the server dynamically optimizes his path, the reader is referred to Browne and Yechiali [l989].

Another important characteristic of our dynamic routing policies is that they are *meaningful* even if the entire system is *not stable* as long as the rate of work flowing to each *individual* channel is less than unity. This last condition implies that the duration of *each* visit of the server to a given channel is *finite* with probability 1, so that *every* cycle is completed with probability 1. Thus, minimizing anew each cycle-time optimizes in some sense the performance of the system.

In section 2 we describe the model, and in section 3 we calculate the Laplace-Stieltjes transform and mean cycle time under the Binomial-Gated regime. The analysis leads to an index-form type of optimal route that minimizes the mean length of any given cycle that starts with an arbitrary state-vector $(n_1, n_2, ..., n_k)$ of customers present in the various queues. In section 4 we analyze the Bernoulli-Gated scheme, while switching times are introduced in section 5. In sections 6 and 7 the Binomial-Exhaustive and the Bernoulli-Exhaustive regimes, respectively, are studied, and in section 8 polling systems with mixed sets of channels are considered. We show that, in all models discussed above, the *same principle* determines the optimal dynamic polling policies of the server.

2. The Model

A single server attends (polls) sequentially K channels (queues) where queue i $(1 \leq i \leq K)$ is of the M/G$_i$/l type with Poissonian arrival rate λ_i, and service requirements V_i possessing probability distribution function $G_i(\cdot)$, mean $E(V_i)$, and Laplace-Stieltjes Transform (LST) $\widetilde{V}_i(\cdot)$. Consider first the Binomial-Gated and the Binomial-Exhaustive disciplines. Suppose that the server finds m_i customers when he enters queue i, and let $N_i(m_i)$ be a Binomial random variable with parameters

m_i and p_i. Then, according to the Binomial-Gated (BG) policy the server resides in channel i until he serves $N_i(m_i)$ *customers*, while according to the Binomial-Exhaustive (BE) policy he stays there for $N_i(m_i)$ *busy periods*. That is, under the BE policy, when the server exits channel i he leaves behind him $m_i - N_i(m_i)$ waiting customers, whereas under the BG policy he leaves behind him $m_i - N_i(m_i) + A_i$ customers, where A_i is the number of *new* arrivals to channel i during the visit time of the server.

The Bernoulli-Gated and Bernoulli-Exhaustive disciplines differ from their Binomial counterparts in that the decision whether to serve customers in channel i or not, is probabilistically made *before* the server enters the channel. With probability p_i he enters the queue, and with probability $1-p_i$ he skips it. When the decision is to enter and render service, then, according to the Bernoulli-Gated (BRG) regime, service is completed only to those m_i customers present at the moment of decision, whereas according to the Bernoulli-Exhaustive (BRE) scheme, the server resides at queue i for m_i busy periods.

One important aspect of the distinction between the Binomial regimes and the Bernoulli schemes becomes evident when switching times are involved. In the Bernoulli schemes those "overhead" costs are saved if the decision is *not* to enter the channel at the current cycle.

3. Minimizing Cycle Time under the Binomial-Gated Policy

Suppose that at time 0 the state of the system is $(n_1, n_2, ..., n_K)$, where n_i is the number of customers present in queue i. Suppose also that the server visits the channels following the order (policy) $\pi_0 = (1, 2, \ldots, K)$, and the service discipline is Binomial-Gated. Let X_j be the server's sojourn time in channel j if he finds there m_j customers upon entering the queue. Then, the LST of X_j is given by

$$\begin{aligned}\widetilde{X}_j(s|m_j) \equiv E[\exp\{-sX_j\}|m_j] &= \sum_{m=0}^{m_j} E[\exp\{-s(\sum_{k=1}^{m} V_{jk})\}]P[N_j(m_j) = m] \\ &= \sum_{m=0}^{m_j} [\widetilde{V}_j(s)]^m \binom{m_j}{m} p_j^m (1-p_j)^{m_j-m} = [p_j\widetilde{V}_j(s) + (1-p_j)]^{m_j} \equiv D_j^{m_j}(s)\end{aligned} \tag{1}$$

where V_{jk} are distributed like V_j. Clearly, $E(X_j|m_j) = m_j p_j E(V_j)$. Under policy $\pi_0 = (1, 2, 3, \ldots, K)$ the exit time of the server from channel $j-1$ is $S_{j-1} \equiv \sum_{i=1}^{j-1} X_i$, so that the number of customers present when the server enters channel j is $m_j = n_j + A_j(S_{j-1})$, where $A_j(S_{j-1})$ is the number of customer arrivals to channel j during the time interval $(0, S_{j-1}]$. Thus,

$$\widetilde{X}_j(s|n_j + A_j(S_{j-1})) = [p_j\widetilde{V}_j(s) + (1-p_j)]^{n_j + A_j(S_{j-1})} .$$

Since $A_j(S_{j-1})$ is a Poisson random variable we obtain

$$\begin{aligned}\widetilde{X}_j(s|S_{j-1}) &= D_j^{n_j}(s)\sum_{n=0}^{\infty} D_j^n(s)\exp\{-\lambda_j S_{j-1}\}(\lambda_j S_{j-1})^n/n! \\ &= D_j^{n_j}(s)\exp\{-\lambda_j p_j(1-\widetilde{V}_j(s))S_{j-1}\}\end{aligned}$$

Finally, unconditioning on S_{j-1}, we have

$$\widetilde{X}_j(s) = [p_j\widetilde{V}_j(s) + (1-p_j)]^{n_j}\,\widetilde{S}_{j-1}(\lambda_j p_j(1-\widetilde{V}_j(s)))\ . \tag{2}$$

¿From Eq. (2) it readily follows that

$$E(X_j) = n_j p_j E(V_j) + b_j p_j E(S_{j-1}) \tag{3}$$

where $b_j \equiv \lambda_j E(V_j)$ is the average amount of work flowing to channel j per unit time. As in Browne and Yechiali [l989], by adding $Z_{j-1} = E(S_{j-1})$ to both sides of Eq.(3) we obtain a system of difference equations

$$Z_j - (1+p_j b_j)Z_{j-1} = n_j p_j E(V_j)\ , \quad (Z_0 = 0) \tag{4}$$

whose solution is

$$Z_j = \sum_{i=1}^{j} p_i n_i E(V_i)[\prod_{r=i+1}^{j}(1+p_r b_r)], \quad (j = 1,2,\ldots,K). \tag{5}$$

Note that $p_i n_i E(V_i)$ is the expected sojourn time of the server in queue i due to the original n_i customers present at time 0. During that period of time one expects $\lambda_{i+1}p_i n_i E(V_i)$ new arrivals to channel $i+1$, but only a fraction p_{i+1} of them will be served, requiring $p_{i+1}b_{i+1}p_i n_i E(V_i)$ time. Thus, the total expected delay in channels i and $i+1$ caused by the original n_i customers in queue i will be $p_i n_i E(V_i)(1+p_{i+1}b_{i+1})$. Proceeding in this manner it follows that the total expected delay caused to the cycle by the n_i initial customers in channel i is $p_i n_i E(V_i)[\prod_{r=i+1}^{K}(1+p_r b_r)]$. Therefore, the expected total cycle time, following policy π_0, is the sum of the expected delays caused by all initial customers present at the start of the cycle

$$Z_K \equiv C(\pi_0) = \sum_{i=1}^{K} p_i n_i E(V_i)[\prod_{r=i+1}^{K}(1+p_r b_r)]\ . \tag{6}$$

Define $a_i \equiv p_i n_i E(V_i)$, and $\alpha_i \equiv p_i b_i$. a_i is the initial expected processing time requirement at channel i, called its *core*, while α_i is the expected *growth* in service requirement at channel i for every unit time delay in performing service to channel i. Thus,

$$C(\pi_0) \equiv \sum_{i=1}^{K} a_i[\prod_{r=i+1}^{K}(1+\alpha_r)] \tag{7}$$

Similarly, if the server polling sequence is determined by the policy $\pi = (\pi(1), \pi(2), ..., \pi(K))$, then the mean cycle length is

$$C(\pi) = \sum_{i=1}^{K} a_{\pi(i)}[\prod_{r=i+1}^{K}(1+\alpha_{\pi(r)})] \ . \tag{8}$$

Applying an interchange argument we have shown in [2] that Eq. (8) is minimized if the channels are visited following a sequence determined by ordering the channels via *increasing* values of a_i/α_i. We therefore conclude

Theorem 1. *Suppose that at time 0 the state of the system is $(n_1, n_2, ..., n_k)$. Then, for the Binomial-Gated policy, the cycle time is minimized if the server visits the channels in an order determined by increasing values of n_i/λ_i.*

Proof: $a_i/\alpha_i = p_i n_i E(V_i)/(p_i b_i) = n_i/\lambda_i$. Q.E.D.

Remark. It is interesting to note that the optimal policy is *independent* of the p_i's and $E(V_i)$'s, and it is the *same* as the optimal policy for the *regular* Gated policy (see[2]).

4. Cycle Time Under the Bernoulli-Gated Scheme

Consider now the Bernoulli-Gated service discipline. If m_j customers are present at channel j when the server reaches the station then his sojourn time there is

$$X_j = \begin{cases} \sum_{k=1}^{m_j} V_{j,k}, & \text{with probability } p_j \\ 0, & \text{otherwise} \end{cases}$$

Therefore, the LST of X_j is derived as follows:

$$\widetilde{X}_j(s|m_j) = p_j[\widetilde{V}_j(s)]^{m_j} + (1-p_j) \equiv \widehat{D}_j^{m_j}(s)$$

Hence,

$$\widetilde{X}_j(s|n_j + A_j(S_{j-1})) = p_j[\widetilde{V}_j(s)]^{n_j + A_j(S_{j-1})} + (1-p_j)$$

$$\widetilde{X}_j(s|S_{j-1}) = p_j[\widetilde{V}_j(s)]^{n_j} \cdot \sum_{n=0}^{\infty}[\widetilde{V}_j(s)]^n \cdot e^{-\lambda_j S_{j-1}} \frac{(\lambda_j S_{j-1})^n}{n!} + (1-p_j)$$

$$= p_j[\widetilde{V}_j(s)]^{n_j} \cdot e^{-\lambda_j S_{j-1}(1-\widetilde{V}_j(s))} + (1-p_j)$$

Thus,

$$\widetilde{X}_j(s) = p_j \left[\widetilde{V}_j(s)\right]^{n_j} \widetilde{S}_{j-1}\left(\lambda_j(1-\widetilde{V}_j(s))\right) + (1-p_j), \tag{9}$$

from which it readily follows that

$$EX_j = p_j n_j E(V_j) + p_j b_j E(S_{j-1}) \ . \tag{10}$$

Writing, as in Eq. (4), $Z_j - (1+p_j b_j) Z_{j-1} = p_j n_j E V_j$, we get that Eqs. (5) and (6) hold in this case as well, with the *same core* $a_i = p_i n_i E(V_i)$ and *growth rate* α_i. That is, the *same* order of visits – by increasing values of n_i/λ_i – minimizes the cycle time under the Bernoulli-Gated regime.

5. Switching Times

The above analyses need be only slightly modified to account for switching times. Assume that a direct switch from station i to station j takes time $\theta_i + T_j$, where θ_i is the time to *switch out* of queue i and T_j is the time to *switch into* channel j (T_i and θ_i are independent of each other and of X_j, T_j and θ_j for all $j \neq i$). Let Y_j denote the total server occupation time with channel j during one cycle, so that now the exit time from channel j is $S_j = \sum_{i=1}^{j} Y_i$ with mean $Z_j = E(S_j)$. Assuming that the customers are gated only *after* the server switches into a channel, then, for the Binomial-Gated,

$$Y_j = T_j + \sum_{k=1}^{N_j(m_j)} V_{jk} + \theta_j, \quad \text{where} \quad m_j = n_j + A_j(S_{j-1} + T_j) \ .$$

Hence,

$$\begin{aligned}
\widetilde{Y}_j(s|T_j, m_j) &= \widetilde{\theta}_j(s) \exp\{-sT_j\} E_{N_j}[\widetilde{V}_j(s)]^{N_j(m_j)} = \widetilde{\theta}_j(s) \exp\{-sT_j\} D_j^{m_j}(s) \ , \\
\widetilde{Y}_j(s|T_j, S_{j-1}) &= \widetilde{\theta}_j(s) \exp\{-sT_j\} D_j^{n_j}(s) E_{A_j}[D_j^{A_j}(s)] \\
&= \widetilde{\theta}_j(s) D_j^{n_j}(s) \exp\{-sT_j\} \exp\{-\lambda_j(S_{j-1}+T_j) p_j (1-\widetilde{V}_j(s))\} \ ,
\end{aligned}$$

so that

$$\widetilde{Y}_j(s) = \widetilde{\theta}_j(s)[p_j \widetilde{V}_j(s) + (1-p_j)]^{n_j} \widetilde{S}_{j-1}(\lambda_j p_j (1-\widetilde{V}_j(s))) \widetilde{T}_j(s + \lambda_j p_j (1-\widetilde{V}_j(s))) \ , \tag{11}$$

and

$$E(Y_j) = p_j n_j E(V_j) + p_j b_j E(S_{j-1}) + (1+p_j b_j) E(T_j) + E(\theta_j) \ . \tag{12}$$

Upon identifying $p_i n_i E(V_i) + (1+p_i b_i) E(T_i) + E(\theta_i)$ as the "core", a_i, and $p_i b_i$ as the "growth rate", α_i, we can write for π_0,

$$Z_K = \sum_{i=1}^{K} [p_i n_i E(V_i) + (1+p_i b_i) E(T_i) + E(\theta_i)][\prod_{r=i+1}^{K} (1+p_r b_r)] \ . \tag{13}$$

¿From our previous principles we obtain

Theorem 2. *The order of visits that minimizes cycle time in a Binomial-Gated policy with switching times is determined by an increasing order of*

$$\frac{p_i n_i E(V_i) + (1 + p_i b_i) E(T_i) + E(\theta_i)}{p_i b_i} \tag{14}$$

Now, for the Bernoulli-Gated with switching times and routing policy π_0, suppose that the coin is flipped *after* leaving channel $j-1$, and *before* entering station j. Then,

$$Y_j = \begin{cases} T_j + \left[\sum_{k=1}^{n_j + A_j(S_{j-1}+T_j)} V_{jk} \right] + \theta_j \ , & \text{with probability } p_j \\ 0 \ , & \text{otherwise} \end{cases}$$

Assuming, as before, that the customers are gated only *after* the server switches into a channel, then

$$\begin{aligned} \widetilde{Y}_j(s \mid T_j, S_{j-1}) &= p_j \left[\widetilde{\theta}_j(s) e^{-sT_j} [\widetilde{V}_j(s)]^{n_j} \sum_{n=0}^{\infty} [\widetilde{V}_j(s)]^n e^{-\lambda_j(S_{j-1}+T_j)} \frac{[\lambda_j(S_{j-1}+T_j)]^n}{n!} \right] + (1 - p_j) \\ &= p_j \left[\widetilde{\theta}_j(s) [\widetilde{V}_j(s)]^{n_j} e^{-sT_j} e^{-\lambda_j(S_{j-1}+T_j)(1-\widetilde{V}_j(s))} \right] + (1 - p_j) \\ &= p_j \left[\widetilde{\theta}_j(s) [\widetilde{V}_j(s)]^{n_j} e^{(-s+\lambda_j(1-\widetilde{V}_j(s)))T_j} \cdot e^{-\lambda_j(1-\widetilde{V}_j(s))S_{j-1}} \right] + (1 - p_j) \end{aligned}$$

Unconditioning, we obtain

$$\widetilde{Y}_j(s) = p_j \left[\widetilde{\theta}_j(s) [\widetilde{V}_j(s)]^{n_j} \widetilde{T}_j(s + \lambda_j(1 - \widetilde{V}_j(s))) \widetilde{S}_{j-1}(\lambda_j(1 - \widetilde{V}_j(s))) \right] +)(1 - p_j) \tag{15}$$

$$EY_j = p_j[E\theta_j + n_j E(V_j) + (1 + b_j)E(T_j) + b_j E(S_{j-1})] \tag{16}$$

Thus,

$$Z_j - (1 + p_j b_j) Z_{j-1} = p_j[n_j E(V_j) + (1 + b_j)E(T_j) + E(\theta_j)] \ , \tag{17}$$

which results in arranging the channels in increasing order of

$$\frac{n_j E(V_j) + (1 + b_j)E(T_j) + E(\theta_j)}{b_j} \tag{18}$$

It is interesting to note that the policy dictated by Eq. (18) is identical to the optimal policy derived for the *pure* Gated regime (see [2]). Note also that the (small) difference between result (14) and policy (18) is due to the fact that in the derivation of Eq. (14) the server switches *with probability 1* to channel j and only *then* the value of the random variable $N_j(m_j)$ is realized, whereas in the derivation of Eq. (18) the coin is flipped *before* the server switches into the channel. Thus, while the growth rate $p_j b_j$ is *identical* for the Binomial-Gated and the Bernoulli-Gated regimes, the cores are *different*. For the former the core is $a_i = E(T_i) + p_i[n_i EV_i + b_i E(T_i)] + E(\theta_i)$, whereas for the latter the core is $p_i[ET_i + n_i E(V_i) + b_i E(T_i) + E(\theta_i)]$.

6. The Binomial-Exhaustive Policy

Consider now the Binomial-Exhaustive regime where the server, if he finds m_i customers in queue i, stays there until the queue length is depleted by $N_i(m_i)$ customers (i.e., for $N_i(m_i)$ *busy periods*), where $N_i(m_i)$ is Binomially distributed with parameters m_i and p_i. This is the Binomial-generalization of the Exhaustive class of disciplines.

Suppose first that there are no switching times. Then, using the same notation as for the Binomial-Gated case, we derive

$$\widetilde{X}_j(s \mid m_j) = \sum_{m=0}^{m_j} [\widetilde{B}_j(s)]^m \binom{m_j}{m} p_j^m (1-p_j)^{m_j - m} = [p_j \widetilde{B}_j(s) + (1-p_j)]^{m_j} \equiv R^{m_j}(s)$$

where B_j is the length of a regular busy period in an $M/G_j/1$ queue, and $\widetilde{B}_j(s)$ is its LST with mean $E(B_j) = E(V_j)/(1-b_j)$. Under policy π_0 the number of customers present in channel j when the server enters the channel is $m_j = n_j + A_j(S_{j-1})$. Hence,

$$\widetilde{X}_j(s \mid S_{j-1}) = R_j^{n_j}(s) \cdot \exp\{-\lambda_j p_j (1 - \widetilde{B}_j(s)) S_{j-1}\}$$

from which we derive

$$\widetilde{X}_j(s) = [p_j \widetilde{B}_j(s) + (1-p_j)]^{n_j} \widetilde{S}_{j-1}(\lambda_j p_j (1 - \widetilde{B}_j(s))) \tag{19}$$

$$E(X_j) = \frac{n_j p_j E(V_j)}{1-b_j} + \frac{p_j b_j E(S_{j-1})}{1-b_j} \; . \tag{20}$$

We can now identify $p_j n_j E(V_j)/(1-b_j)$ as the "core" of channel j, and $p_j b_j/(1-b_j)$ as its "growth rate". Correspondingly, it is immediate that the expected cycle length has the evaluation

$$Z_K = \sum_{i=1}^{K} \left(\frac{n_i p_i E V_i}{1-b_i} \right) \left[\prod_{r=i+1}^{K} \left(1 + \frac{p_r b_r}{1-b_r} \right) \right] \; , \tag{21}$$

and that the *optimal policy* is to once again order the channels in an increasing order of n_i/λ_i, which is *identical* to the optimal policy for the Binomial-Gated and again independent of p_i and $E(V_i)$.

When switching times are incurred, utilizing previous notation, we can readily modify the above by observing that Y_j, the server's occupation time with channel j, can be written as

$$Y_j = T_j + \sum_{k=0}^{N_j(m_j)} B_{jk} + \theta_j$$

where $m_j = n_j + A_j(S_{j-1} + T_j)$, and B_{jk} are distributed like B_j. Hence,

$$\widetilde{Y}_j(s \mid T_j, S_{j-1}) = \widetilde{\theta}_j(s) R_j^{n_j}(s) \exp\{-sT_j\} \exp\{-\lambda_j(S_{j-1} + T_j) p_j (1 - \widetilde{B}_j(s))\}$$

so that

$$\widetilde{Y}_j(s) = \widetilde{\theta}_j(s)[p_j \widetilde{B}_j(s) + (1 - p_j)]^{n_j} \widetilde{S}_{j-1}(\lambda_j p_j (1 - \widetilde{B}_j(s))) \widetilde{T}_j(s + \lambda_j p_j (1 - \widetilde{B}_j(s))) \tag{22}$$

and

$$E(Y_j) = p_j n_j E(V_j)/(1 - b_j) + [p_j b_j/(1 - b_j)] E(S_{j-1}) + [1 + p_j b_j/(1 - b_j)] E(T_j) + E(\theta_j) \ . \tag{23}$$

As before, this leads to a mean cycle time

$$Z_K = \sum_{i=1}^{K} \{[p_i n_i E(V_i) + (1 - b_i + p_i b_i) E(T_i) + (1 - b_i) E(\theta_i)]/(1 - b_i)\} \left[\prod_{r=i+1}^{K} \left(\frac{1 + p_r b_r}{1 - b_r} \right) \right] \ . \tag{24}$$

We conclude

Theorem 3. *The optimal sequence of visits by the server is determined by arranging the queues in an increasing order of*

$$\frac{p_i n_i E(V_i) + (1 - b_i + p_i b_i) E(T_i) + (1 - b_i) E(\theta_i)}{p_i b_i}$$

7. The Bernoulli-Exhaustive Scheme

In this case, if the server enters channel j and finds m_j customers, he resides there for m_j busy period. As before, the decision whether to enter or not is governed by a Bernoulli trial with probability of success p_i. As $m_j = n_j + A(S_{j-1})$, then, without switching times, we have

$$X_j = \begin{cases} \sum_{k=1}^{m_j} B_{jk} \ , & \text{with probability } p_j \\ 0 \ , & \text{otherwise} \end{cases}$$

with LST

$$\widetilde{X}_j(s \mid n_j + A(S_{j-1})) = p_j (\widetilde{B}_j(s))^{n_j} (B_j(s))^{A(S_{j-1})} + (1 - p_j)$$

$$\widetilde{X}_j(s \mid S_{j-1}) = p_j (\widetilde{B}_j(s))^{n_j} e^{-\lambda_j S_{j-1}(1 - \widetilde{B}_j(s))} + (1 - p_j) \ .$$

Finally,

$$\widetilde{X}_j(s) = p_j (\widetilde{B}_j(s))^{n_j} \widetilde{S}_{j-1}(\lambda_j (1 - \widetilde{B}_j(s))) + (1 - p_j) \ , \tag{25}$$

$$EX_j = p_j[n_j E(B_j) + \lambda_j E(B_j)E(S_{j-1})] = \frac{p_j}{1-b_j}\,(n_j EV_j + b_j E(S_{j-1}) \tag{26}$$

so that

$$Z_j - \left(1 + \frac{p_j b_j}{1-b_j}\right) = \frac{p_j n_j EV_j}{1-b_j}\ .$$

Identifying

$$a_j = \frac{p_j n_j EV_j}{1-bj} \quad \text{and} \quad \alpha_j = \frac{p_j b_j}{1-b_j}\ ,$$

the optimal order of visits is determined by increasing values of $a_i/\alpha_i = n_i/\lambda_i$, *exactly* as in the case for the Binomial-Exhaustive regime *without* switching times.

If we take into account switching times, we write

$$Y_j = \begin{cases} T_j + \sum_{k=1}^{n_j + A_j(S_{j-1}+T_j)} B_{jk} + \theta_j\ , & \text{with probability } p_j \\ 0\ , & \text{otherwise} \end{cases}$$

so that

$$\widetilde{Y}_j(s) = p_j\left[\widetilde{\theta}_j(s)[\widetilde{B}_j(s)]^{n_j}\widetilde{T}_j(s+\lambda_j(1-\widetilde{B}_j(s)))\widetilde{S}_{j-1}(\lambda_j(1-\widetilde{B}_j(s)))\right] + (1-p_j) \tag{27}$$

and

$$EY_j = p_j\left[E\theta_j + n_j\frac{EV_j}{1-b_j} + (1+\frac{b_j}{1-b_j})ET_j + \frac{b_j}{1-b_j}ES_{j-1}\right]\ . \tag{28}$$

Setting

$$a_j = p_j\left[\frac{n_j EV_j + ET_j + (1-b_j)E\theta_j}{1-b_j}\right]\ , \quad \text{and} \quad \alpha_j = \frac{p_j b_j}{1-b_j}\ ,$$

the optimal sequence is determined by the index

$$\frac{a_j}{\alpha_j} = \frac{n_j EV_j + ET_j + (1-b_j)E\theta_j}{b_j}$$

which is *identical* to the case with (fully) Exhaustive regime.

8. Mixed Sets of Channels

Our representation of the cycle times for the above four service disciplines in terms of cores (a_i) and growth rates (α_i) allows us to immediately solve for cases with *Mixed* channels, where the service discipline is not common for all channels, but rather, some channels require a pure Exhaustive regime, others - a pure Gated mode, and others – one form or another of "fractional-type". In addition, some channels may require switch-in or switch-out times or both. We then have

Theorem 4. *The mean cycle time is minimized if the channels are arranged by increasing values of a_i/α_i, where, if a channel is Binomial-Exhaustive, then*

$$a_i = [p_i n_i E(V_i) + (1 - b_i + p_i b_i)E(T_i) + (1 - b_i)E(\theta_i)]/(1 - b_i)$$

$$\alpha_i = p_i b_i/(1 - b_i)$$

whereas if it is Binomial-Gated,

$$a_i = p_i n_i E(V_i) + (1 + p_i b_i)E(T_i) + E(\theta_i)$$

$$\alpha_i = p_i b_i \ .$$

If a channel is Bernoulli-Gated, then

$$a_i = p_i[n_i E(V_i) + (1 + b_i)E(T_i) + E(\theta_i)]$$

$$\alpha_i = p_i b_i \ ,$$

whereas, if it is Bernoulli-Exhaustive,

$$a_i = p_i[n_i EV_i + ET_i + (1 - b_i)E\theta_i]/(1 - b_i)$$

$$\alpha_i = p_i b_i/(1 - b_i) \ .$$

Proof: Imitating the previous derivations, the expected cycle time under π_0 for *any* Mixed set of channels is given by

$$C(\pi_0) = \sum_{i=1}^{K} a_i \left[\prod_{r=i+1}^{K} (1 + \alpha_r) \right] \ ,$$

from which it follows that ordering by *increasing* values of a_i/α_i *minimizes* the expected cycle time.

Q.E.D.

9. Conclusion

We have derived optimal *dynamic* polling schemes for several service disciplines and for sets with mixed channels. Our methods take into account the *dynamic* and *stochastic* evolution of the process, while at the same time maintain a degree of fairness among the various channels by visiting each queue *once in every cycle*. The underlying principle in these dynamic policies is to look – at each decision epoch – at the "core" accumulated in each queue (that is, at the amount of work waiting for the server if he enters the queue at that moment), and to look at the "growth rate" of work associated with each queue. The ratio "core"/"growth" determines for every queue its position in the sequence of visits to be performed next by the server. The principle is applicable to any mix of service-disciplines of the queues and it is easy to implement regardless if one is able to calculate mean waiting times or not.

References

[1] O.J. Boxma, "Workloads and Waiting Times in Single-Server Systems with Multiple Customer Classes", to appear in *Queueing Systems*, 1990.

[2] S. Browne and U. Yechiali, "Dynamic Priority Rules for Cyclic-Type Queues", *Adv. App. Prob.* **21** (2) (1989) 432-450.

[3] S. Browne and U. Yechiali, "Dynamic routing in Polling Systems", in *Teletraffic Science*, M. Bonatti (ed), Elsevier Sciences Pub. (1988) 1455-1466.

[4] H. Levy, "Optimization of Polling Systems via Binomial Service", Technical Report 102/88, Dept. of Computer Science, Tel Aviv University (1988).

[5] H. Takagi, *Analysis of Polling Systems*, MIT Press (1986).

[6] H. Takagi, "Queueing Analysis of Polling Models", *ACM Comp. Surveys* **20** (l) (1988) 5-29.

[7] K.S. Watson, "Performance Evaluation of Cyclic Service Strategies – a Survey", in *Performance '84*, E. Gelenbe (ed), North Holland (1984) 521-533.