Threshold start-up control policy for polling systems

Yavuz Günalay [a] and Diwakar Gupta [b]

[a] *Faculty of Business Administration, Bilkent University, Ankara 06533, Turkey*
E-mail: gunalay@bilkent.edu.tr
[b] *Michael G. DeGroote School of Business, McMaster University, Hamilton, Ontario L8S 4M4, Canada*
E-mail: guptad@mcmaster.ca

Received 27 June 1997; revised 31 March 1998

A threshold start-up policy is appealing for manufacturing (service) facilities that incur a cost for keeping the machine (server) on, as well as for each restart of the server from its dormant state. Analysis of single product (customer) systems operating under such a policy, also known as the N-policy, has been available for some time. This article develops mathematical analysis for multiproduct systems operating under a cyclic exhaustive or globally gated service regime and a threshold start-up rule. It pays particular attention to modeling switchover (setup) times. The analysis extends/unifies existing literature on *polling models* by obtaining as special cases, the continuously roving server and patient server polling models on the one hand, and the standard $M/G/1$ queue with N-policy, on the other hand. We provide a computationally efficient algorithm for finding aggregate performance measures, such as the mean waiting time for each customer type and the mean unfinished work in system. We show that the search for the optimal threshold level can be restricted to a finite set of possibilities.

Keywords: polling models, threshold start-up control, dormant/patient server, descendant sets, globally gated service, queueing theory

1. Introduction

Consider a manufacturing facility that can be modeled by a single server (e.g., a system with a single bottleneck station) and produces M different types of products (or parts). A natural and easy way to implement production control regime for such systems is the cyclic production policy in which a batch of type i products always follows a batch of type $i-1$. The underlying queueing model – a single server cycling around a fixed route and providing service to queued customers en route – is an extremely versatile model that can be used to study a variety of telecommunications, manufacturing, and service systems. For example, the server could represent a token needed to transmit/receive messages in a local area network, a transportation device like an automated guided vehicle, a robot attending to several different tasks, or else, a walking repairman. Such models are called polling models since the server typically "polls" each queue to determine the number waiting at that queue. Spurred by their

© J.C. Baltzer AG, Science Publishers

numerous applications, there have been many recent advances in the analysis of polling systems and research continues at a feverish pace.

This article is concerned with the analysis of polling systems having a threshold start-up control policy. Under this policy, once the server becomes dormant (that happens whenever the system is empty), it would restart only when the number of new arrivals to the system reaches a critical value. A threshold start-up control regime is relevant when there are costs, such as wages and energy costs, that are paid only when the server is available, but not when it is dormant. The optimal threshold is simply 1, if there are no additional start-up costs. However, in general, there might also be fixed start-up costs, such as a power surge, requirement of additional personnel, or a one-time setup. In that case, it makes sense to choose the threshold carefully to realize the least costly balance between start-up and waiting costs. For single-customer-class systems with a removable server this threshold start-up policy is widely known as the N-policy (Heyman and Sobel [11, section 7-2]). In the context of polling systems, we observe that the mean unfinished work in system does not change monotonically with respect to the threshold level N. Thus, even in the absence of explicit start-up costs, higher threshold levels might improve the system performance in terms of the mean unfinished work in system. This happens on account of the server changeover time necessary to set up for each different customer class.

In addition to start-up, shut-down, and cycling rules, a polling model must also specify how many jobs of any one type are processed after each new setup. Some commonly studied policies include: *exhaustive* – in which the server empties the queue before moving on to another, *gated* – here the server processes only those customers that it found waiting upon polling the station, and *globally gated* – in which *all* gates are closed at once when the server polls the home station and only those customers that are in front of their gates are served in the succeeding cycle. In this paper, we develop a computationally efficient method to calculate the mean waiting time of customers when either exhaustive (E) or globally gated (GG) service policy is in effect. Following previous literature, e.g., Resing [14], it is easy to show that the analysis of systems with a gated (G) service regime, or with some E and some G stations, is similar to that of systems with E service policy. For this reason, these variations are not discussed. We show that the search for an optimal N can be restricted to a finite set of candidates, considerably reducing the necessary computational effort.

There is a large body of literature dealing with polling systems in which the server never stops, i.e., $N = 0$. Such models have lately been labeled continuously roving server models and they are reviewed extensively by Takagi [18–20]. If the server stops whenever the system is empty and restarts as soon as a new customer arrives, i.e., if threshold $N = 1$, we obtain a special instance of our problem. For the E and G service regimes, this model has been studied in two recent articles: Eisenberg [7], and Srinivasan and Gupta [16]. This special case has been called the *stopping server* regime by Eisenberg, and the *patient server* regime by Srinivasan and Gupta. Similar special cases, i.e., $N = 0$ and $N = 1$, of the model with GG service discipline were studied by Boxma et al. [4] and Borst [2], respectively. Our work can be seen as a generalization

of these recent studies to any arbitrary threshold level. We provide a new unification of literature on polling models: by setting $N = 0$, our model and the entire analysis reduces to the standard continuously roving server model, and by setting it equal to 1, a similar match occurs with the patient server model. Similarly, upon setting the number of customer classes, $M = 1$, and accounting for some modeling differences, we obtain the mean waiting time under the well known N-policy for $M/G/1$ queues. Such transparent relationships build bridges, where none existed before, and provide new insights, similar in spirit to the recent work of Srinivasan et al. [17].

The analysis contained in this article uses the descendant sets method [12,16]. It is similar, for E and G disciplines, to a recent article by Srinivasan and Gupta [16], and for the GG discipline to the work by Boxma et al. [4] and Borst [2]. For this reason, we provide only a sketch of our arguments leading to the mean customer waiting times and the pseudo-conservation law. Interested readers can find these details in the technical report by Günalay and Gupta [9].

For the E service discipline, we use the simplicity of a symmetric model to find a critical threshold level $\overline{N}$ beyond which the system performance measure (pseudo-conservation law) never improves over what we can obtain by setting $N = 0$. It is also easy to show that this result holds, in essence, even when instead of the pseudo-conservation law, which uses a weight equal to station load for each station, we were to use arbitrary weights (holding costs). The system performance with $N = 0$ threshold is called the *base* performance level. Although the asymmetric case is more complex, and therefore an explicit expression for $\overline{N}$ difficult to find, we can both argue its existence and obtain an explicit upper bound. The GG service regime leads to simple enough expressions that we do find an explicit expression for $\overline{N}$ even for asymmetric models. Thus, in each case, the optimum threshold can be obtained by searching in a finite interval, $[0, \overline{N}]$.

We include several examples to illustrate the effect of system parameters on $\overline{N}$, and the optimal threshold N^*. Since we use the same data sets, we are also able to compare the effectiveness of E and GG strategies, in a limited fashion. For the E service regime, N^* and the system performance measure are quite sensitive to the mean arrival rate and the variance of the switchover times. Performance under the GG service regime is relatively insensitive to changes in N. When threshold is set at its optimum value in each case, the E policy always outperforms the GG policy in terms of the pseudo-conservation law; our overall system performance measure. However, as figures 1 and 2 demonstrate, if we choose a high threshold level, performance under the E policy can deteriorate rapidly and this can easily make it much worse than the GG policy.

Our numerical analysis serendipitously revealed another interesting fact. Performance under the GG policy is very sensitive to the order in which queues are visited; especially for asymmetric models. In fact, the effect of order of visitation appears to be far greater than the effect of threshold N. This is in direct contrast with the E policy for which the order of visitation has a very small impact on overall system performance. Seeing its importance for design of GG controlled systems, we have

obtained a simple sequencing rule which minimizes the pseudo-conservation law for a given set of system parameters. The optimal sequence is independent of N.

The plan of this article is as follows. We discuss details of the model and notation in section 2. The analysis needed to find the distribution of queue lengths at polling instants is presented in section 3. Section 4 contains explicit expressions for the mean station waiting times and the pseudo-conservation law for both E and GG service regimes. Numerical experiments and insights are presented in section 5. The paper concludes in section 6 with a simple rule for determining the optimum order of visitation for the GG service discipline.

2. Model description and notation

There are M customer (product) types, served by a single server (machine), that join their own queues (stations) upon arrival. Arrivals at station i are governed by an independent Poisson process of rate λ_i. The server visits stations in a cyclic order. Without loss of generality (w.l.o.g.), this order is assumed to be $1, 2, \ldots, M, 1, \ldots$. We denote the work load at a station by $\rho_i = \lambda_i E[B_i]$, the total system load by $\rho = \sum_{i=1}^{M} \rho_i$, the overall arrival rate by $\Lambda = \sum_{i=1}^{M} \lambda_i$, and the probability that an arbitrary arrival is type-i by $p_i = \lambda_i/\Lambda$. Similarly, the time to switch from station i to station $i+1$ is denoted by R_i, the total switchover time in a cycle by $R_T = \sum_{i=1}^{M} R_i$, service time of type-i customers by B_i, and the busy period generated by a type-i service by Θ_i.

Some notational conventions used in this article are as follows. For a random variable A, we use $A(t)$, $\widetilde{A}(s)$, $E[A]$ and $E[A^2]$ to denote the cumulative distribution function, the Laplace–Stieltjes transform (LST), the mean, and the second moment, respectively. When A is discrete then $A(z) \stackrel{\triangle}{=} E[z^A]$ denotes its probability generating function (PGF). Single and double prime notation is used to denote, respectively, first and second derivative with respect to z. The notation **n** (or **z**) represents a $1 \times M$ vector of n_i's (or z_i's). Thus, **1** and **0** denote a vector in which all n_i, $i = 1, \ldots, M$, equal to 1 and 0, respectively. Parameter N is used in parenthesis to emphasize the dependence of a performance measure on the threshold level. However, this notation is suppressed until section 5, i.e., until we explicitly consider calculation of optimal threshold levels.

Under service regime E, the server checks the status of queues at all stations whenever it finishes its work at any station (thus, ready to switch to the next station); this instant is called a *server-departure epoch*. This should not be confused with the actual instance at which the server leaves a station to begin its switch to the next station. The latter is called a *switch point* and this set of observation instants is a subset of server-departure epochs as explained below.

We define a *switch point* as the time instant when the server starts a switchover period. Switch points are either server-departure instants that observe at least one nonempty queue in the system, or instants at which the server is reactivated and

commences a switchover immediately. If all queues are empty at a server-departure epoch from some station, the server becomes idle and remains at that station without registering a switch point. It lies in this dormant mode until the polling system is populated by exactly N customers. Service always resumes from the same station where the server had stopped. Therefore, the server switches immediately after an idle period only if none of the N arrivals occur to the station where it is idling. When, on the other hand, service resumes at the same station where the server was idling, the server must register another server-departure instant upon emptying that station's queue, following which it will either switch or commence another idle period.

After passage of an appropriate switchover time, the server arrives at the next station (ready to serve customers waiting at that station), and this instant is called a *polling instant*. Note that the restart of service at station i following an idle period at that station is not a polling instant. For that purpose, we define *station beginning* instants as observation epochs that are either polling instants or instants at which the server is reactivated following an idle period at that same station. Thus, there is a one-to-one correspondence between the following pairs: station-beginning and server-departure instants, and polling instants and switch points.

In this article, we are concerned with stationary (steady state) behavior of polling systems with threshold startup control policy. All performance measures, i.e., queue lengths and waiting time distributions defined in the article therefore pertain to stationary characteristics. Conditions under which stationary distributions exist are discussed later in this section. We use $f_i(n_1, \ldots, n_M)$ to denote the stationary joint probability that the server polls station-i and finds n_j customers at station j, $j = 1, \ldots, M$. The corresponding PGF is $f_i(z_1, \ldots, z_M)$. Similarly, $k_i(z_1, \ldots, z_M)$, $g_i(z_1, \ldots, z_M)$ and $h_i(z_1, \ldots, z_M)$ denote partial PGFs of the stationary probability distribution of queue lengths, and either a type-i station beginning epoch, or a server-departure instant, or a switch point, respectively. Furthermore, we use uppercase letters in PGFs to indicate that the event is conditioned on the station type, e.g., $F_i(\mathbf{z})$ is the PGF of queue lengths given that it is a station i polling instant, and it is calculated as $F_i(\mathbf{z}) = f_i(\mathbf{z})/f_i(\mathbf{1})$, $i = 1, \ldots, M$.

The various concepts introduced in the previous two paragraphs apply also to the GG service model with some minor modifications, which we shall discuss next. Under the GG service regime, the server checks the status of queues only when it arrives at the *home base*, which we choose to be station 1, w.l.o.g. Thus, at a station 1 polling (system polling) instant the server becomes dormant if it finds the whole system empty, and stays dormant until a total of N customers accumulate in the system. The Nth arrival re-activates the server and it starts cycling immediately. In the cycle that follows an idle period exactly N customers are served. But, as in the E service discipline, this re-start instant at station 1 is not marked as another system polling instant. If the system is not empty at the system polling instant, then the server processes only those customers that are already in the system at the polling instant. Let $F(z_1, \ldots, z_M)$ denote the PGF of queue lengths at a system (station 1) polling instant. Notice that

unlike the E service discipline, we do not have a partial PGF, or $f(\cdot)$. Also, we do not need a subscript to denote the station index since the system is polled only at station 1.

For both E and GG disciplines, the server stops only when it finds the system empty at the appropriate observation epoch. However, depending on the threshold level, the number of *start-up states* can vary. Let $U(N)$ be the set of all states with exactly N customers in the system. Then,

$$U(N) = \{(n_1, \ldots, n_M) \in I_+^M\colon\ n_1 + \cdots + n_M = N\}. \tag{1}$$

We also define $U_i(N)$, a proper subset of $U(N)$, to be the set of states with n_i greater than zero, i.e.,

$$U_i(N) = \{(n_1, \ldots, n_M) \in U(N)\colon\ n_i > 0\}, \quad i = 1, \ldots, M. \tag{2}$$

Finally, $U_i^{\mathrm{c}}(N) \stackrel{\triangle}{=} U(N)\backslash U_i(N)$ represents the complement of $U_i(N)$.

Stability conditions

For threshold startup polling models with exhaustive, gated and globally gated service policies, the necessary and sufficient stability conditions are (i) $\rho < 1$, and (ii) $N < \infty$. It is also required that time length distributions B_i and R_i have finite first moments. That these conditions are both necessary and sufficient can be ascertained by following arguments similar to those presented in Altman et al. [1, lemma 3.1, proposition 3.2, corollary 3.3]. Although Atlman et al. do not consider threshold startup polling models, the main reason why their arguments still hold (at least for E, G and GG regimes) is that the number of customers served at each queue during a server visit to that queue monotonically increases to ∞ as the number of waiting customers goes to ∞. This holds even for server visits that occur immediately after a server idle period. Informally, this can be seen from relationships (29) and (35) in which the mean cycle lengths (average time elapsed between two server polling instants at the same station) are derived for E and GG regimes. Both these expressions are finite only when $\rho < 1$ and $N < \infty$. Notice that ϑ_i and $F(\mathbf{0})$ in (29) and (35) denote the probabilities of server idling and therefore lie in the interval $(0, 1)$.

3. Queue length distributions at polling instants

Let the *reference point* be an arbitrary polling instant of station 1, the time period between two successive polling instants of station 1 be a *cycle*, and Q_1 be the station 1 queue length at the reference point. Then, the "contribution" to Q_1 from a test customer $\mathcal{C}$ is a subset of Q_1 comprising of all *offsprings* of $\mathcal{C}$. The complete set of offsprings of a customer consists of itself, any customers that arrive during its service time, and all their offsprings. Let $L_{i,c}$ denote the contribution of a type-i customer that is served c cycles prior to the reference point, and let $L_{i,c}(z)$ represent its PGF. Similarly, let $R_{i,c}(z)$ denote the PGF of the contribution of arrivals during a switchover period from

station i to $i+1$, c cycles prior to the reference point. These PGF's depend on the service discipline, and procedures for calculating them are presented separately for each service regime.

Exhaustive service regime

We demonstrate our procedure, w.l.o.g., for station 1. Similar results for other stations can be obtained simply by rotating the station index. The contribution of a type-i customer that is served c cycles prior to the reference point $L_{i,c}$ is equal to the sum of the contributions of all customers (other than type-i) which arrive during its busy period. Therefore, its PGF $L_{i,c}(z)$ can be calculated recursively as follows (see, for example, [16]):

$$L_{i,c}(z) = \widetilde{\Theta}_i\left(\sum_{j=i+1}^{M}\left[\lambda_j - \lambda_j L_{j,c}(z)\right] + \sum_{j=1}^{i-1}\left[\lambda_j - \lambda_j L_{j,c-1}(z)\right]\right),$$
$$i = 1,\ldots,M,\ c \geqslant 0. \tag{3}$$

Similarly, $R_{i,c}$ is equal to the sum of the contributions of all customers which arrive during the type-i switchover period c cycles prior to the reference point, and its PGF $R_{i,c}(z)$ can be calculated as

$$R_{i,c}(z) = \widetilde{R}_i\left(\sum_{j=i+1}^{M}\left[\lambda_j - \lambda_j L_{j,c}(z)\right] + \sum_{j=1}^{i}\left[\lambda_j - \lambda_j L_{j,c-1}(z)\right]\right),$$
$$i = 1,\ldots,M,\ c \geqslant 0. \tag{4}$$

Recall that in expressions (3) and (4), the notation $\widetilde{\Theta}_i$ and $\widetilde{R}_i$ denotes the LST of Θ_i and R_i, respectively. The boundary conditions for (3) are as follows: $L_{1,-1}(z) = z$ and $L_{i,-1}(z) = 1$, for all $i > 1$. Let $p(\mathbf{n})$ represent the probability of observing state $\mathbf{n} \in U(N)$ at a start-up instant, and $u_{i,c}(\mathbf{n}, z)$ denote the PGF of the total contribution to Q_1 from this state when the server is at station i, c cycles prior to the reference point. Then we have

$$p(\mathbf{n}) = \frac{N!}{n_1!\cdots n_M!}\prod_{j=1}^{M} p_j^{n_j}, \tag{5}$$

and

$$u_{i,c}(\mathbf{n}, z) = \prod_{j=i}^{M} L_{j,c}(z)^{n_j}\prod_{j=1}^{i-1} L_{j,c-1}(z)^{n_j}, \quad i = 1,\ldots,M,\ c \geqslant 0. \tag{6}$$

Next, we present the main result of this section, which is obtained by allowing generalized start-up functions in equations (4)–(16) of Srinivasan and Gupta [16]. Details

of this derivation can be found in [9]. The PGF of the stationary distribution of the queue length of station 1 at a polling instant is

$$f_1(z,1,\ldots,1)=\Phi\left[\prod_{c=0}^{\infty}\prod_{i=1}^{M}R_{i,c}(z)-\sum_{c=0}^{\infty}\sum_{i=1}^{M}\vartheta_i J_{i,c}(z)E_{i,c}(z)\right], \tag{7}$$

where the following definitions have been used:

$$\Phi=f_1(\mathbf{1}), \tag{8}$$

$$\vartheta_i=g_i(\mathbf{0})/\Phi, \qquad i=1,\ldots,M, \tag{9}$$

$$J_{i,c}(z)=1-\sum_{\mathbf{n}\in U(N)}p(\mathbf{n})u_{i,c}(\mathbf{n},z), \quad i=1,\ldots,M,\ c\geqslant 0, \tag{10}$$

and

$$E_{i,c}(z)=\prod_{j=i}^{M}R_{j,c}(z)\prod_{l=0}^{c-1}\prod_{k=1}^{M}R_{k,l}(z), \quad i=1\ldots,M,\ c\geqslant 0. \tag{11}$$

The scaled empty system probabilities ϑ_i, $i=1,\ldots,M$, and the constant Φ can be calculated by following a method of solving M linear equations in as many unknowns. These linear equations have the following form:

$$\sum_{i=1}^{M}a_i^{(j)}\vartheta_i=b^{(j)}, \quad j=1,\ldots,M, \tag{12}$$

where

$$a_i^{(j)}=\sum_{c=0}^{\infty}J_{i,c}(\mathbf{0})E_{i,c}(\mathbf{0}), \quad i=1,\ldots,M, \tag{13}$$

and

$$b^{(j)}=\prod_{c=0}^{\infty}\prod_{i=1}^{M}R_{i,c}(\mathbf{0}). \tag{14}$$

The superscript j indicates that the reference point is a station j polling instant. We use $L_{i,c}(\mathbf{z})$ to denote the PGF of the joint contribution to *all* queues Q_k, $k=1,\ldots,M$, by a type-i customer served c cycles prior to the reference point. Similarly, $R_{i,c}(\mathbf{z})$, $i=1,\ldots,M$, $c\geqslant 0$, is defined as the PGF of the total joint contribution to all queues from a station i to $i+1$ switchover period, c cycles prior to the reference point. Then, setting $\mathbf{z}=\mathbf{0}$ and using the fact that the system is never empty at a polling instant, i.e., $f_i(\mathbf{0})=0$, $i=1,\ldots,M$, we obtain equation (12). After evaluating ϑ_i, we obtain Φ by summing up equation (8) for all i, and using the normalization $\sum_{i=1}^{M}g_i(\mathbf{1})=1$. This yields

$$\Phi=\frac{1}{M+\sum_{i=1}^{M}\vartheta_i(1-(1-p_i)^N)}. \tag{15}$$

Globally gated service regime

It is clear from the description of the GG model in section 2 that arrivals during cycle $c \geqslant 0$ are served in cycle $c-1$, unless they arrive during the idle period. Therefore, PGFs of contributions of arrivals during a service time and the sum of switchover periods per cycle can be written as follows:

$$L_{i,c}(z) = \widetilde{B}_i\left(\sum_{j=1}^{M}\left[\lambda_j - \lambda_j L_{j,c-1}(z)\right]\right), \quad i = 1,\ldots,M,\ c \geqslant 0, \tag{16}$$

and

$$R_{T,c}(z) = \widetilde{R}_T\left(\sum_{j=1}^{M}\left[\lambda_j - \lambda_j L_{j,c-1}(z)\right]\right), \quad c \geqslant 0. \tag{17}$$

Notice that in GG model, we are concerned only with the PGF of the total contributions of arrivals during the sum of all switchover times. This happens because arrivals during switchover times incurred c cycles prior to the reference point are always served in the cycle indexed $c-1$. The boundary conditions are the same as E service model, i.e., $L_{1,-1}(z) = z$ and $L_{j,-1} = 1$, for all $j > 1$.

The PGF of the sum of contributions to Q_1 from each customer present in the system at a polling instant c cycles prior to the reference point is denoted by $F_c(z)$, where

$$F_c(z) = F\big(L_{1,c}(z), L_{2,c}(z), \ldots, L_{M,c}(z)\big).$$

By definition, $F_c(z)$ must be equal to the sum of contributions from those customers that were present $c+1$ cycles ago, plus additional contributions from all those customers that arrive during switchover periods of the $(c+1)$st cycle. If the system is empty at the previous polling instant and the server restarts with system in state $\mathbf{n} \in U(N)$, then the PGF of the total contribution to Q_1 is simply $u_{1,c}(\mathbf{n}, z)$. Putting it all together, we get

$$F_c(z) = \big(F_{c+1}(z) - F(\mathbf{0})\big)R_{T,c+1}(z) + F(\mathbf{0})\sum_{\mathbf{n}\in U(N)} p(\mathbf{n})u_{1,c+1}(\mathbf{n}, z)R_{T,c+1}(z),$$
$$c \geqslant -1. \tag{18}$$

Setting $c = -1$ in equation (18) and then, writing $F_{-1}(z)$ in terms of $F_0(z)$ and $F_0(z)$ in terms of $F_1(z)$ and so on, we obtain an infinite recursion. Using arguments similar to the E service discipline models, we simplify this expression and obtain the following PGF of station 1 queue length at an arbitrary system polling instant

$$F(z,1,\ldots,1) = F_{-1}(z) = \left[\prod_{c=0}^{\infty} R_{T,c}(z) - F(\mathbf{0})\sum_{c=0}^{\infty} J_{1,c}(z)\prod_{m=0}^{c} R_{T,m}(z)\right], \tag{19}$$

where $J_{1,c}(z)$ is defined as in equation (10), and $F(\mathbf{0})$ is the empty system probability at a polling instant.

The empty system probability can be calculated as in the E service model. However, here the emptiness of the system matters only when it occurs at a system polling instant. We denote this probability by $F(\mathbf{0})$. It is obtained by first writing equation (19) for $F(\mathbf{z})$, and then setting $\mathbf{z} = \mathbf{0}$. Thus,

$$F(\mathbf{0}) = \frac{\prod_{c=0}^{\infty} R_{T,c}(\mathbf{0})}{1 + \sum_{c=0}^{\infty} J_{1,c}(\mathbf{0}) \prod_{m=0}^{c} R_{T,m}(\mathbf{0})}. \tag{20}$$

4. The mean waiting times

We use queue length distributions at polling instants to calculate the mean waiting times. Note that, for $M/G/1$ queues with $N > 1$ start-up threshold, Fuhrmann–Cooper decomposition applies to the distribution of number in the system, but not to the distribution of time in the system (see Cooper [6, exercise 12, part h, pp. 222–223] and Fuhrmann and Cooper [8]). However, as we show next, we can obtain expressions for mean waiting times and the pseudo-conservation law for both E and GG service models.

Exhaustive service regime

Let $\Pi_1(\mathbf{z})$ denote the PGF for the stationary distribution of queue lengths at a departure instant of a type-1 customer. If we treat each server departure instant from station 1 as the start of a server vacation, the following station 1 beginning instant as the end of that vacation, and apply the Stochastic Decomposition Theorem (Fuhrmann and Cooper [8, proposition 2]) for $M/G/1$ queues with server vacations, we obtain

$$\Pi_1(z, 1, \ldots, 1) = \frac{k_1(\mathbf{1}) - k_1(z, 1, \ldots, 1)}{k_1'(\mathbf{1})} \cdot \frac{(1 - \rho_1)\widetilde{B}_1(\lambda_1 - \lambda_1 z)}{\widetilde{B}_1(\lambda_1 - \lambda_1 z) - z}. \tag{21}$$

Differentiating equation (21) with respect to z and then setting $z = 1$, we get the expected queue length of station 1 at a customer departure instant, which is also the average queue length at station 1 at an arbitrary observation epoch (see, e.g., [6, pp. 186–188]). Then, using Little's law the mean waiting time of type-1 customers can be calculated as follows:

$$E[W_1] = \frac{k_1''(\mathbf{1})}{2\lambda_1 k_1'(\mathbf{1})} + \frac{\lambda_1 E[B_1^2]}{2(1 - \rho_1)}. \tag{22}$$

Recall that a station beginning instant is either a polling instant or an instant at which the server is reactivated following an idle period at the same station. Therefore, the PGF of the queue length at an arbitrary station 1 beginning instant is

$$k_1(z, 1, \ldots, 1) = f_1(z, 1, \ldots, 1) + g_1(\mathbf{0}) \sum_{\mathbf{n} \in U_1(N)} p(\mathbf{n}) z^{n_1}. \tag{23}$$

Substituting from equation (7), we can simplify $k_1(z,1,\ldots,1)$ as follows:

$$k_1(z,1,\ldots,1) = \Phi\left[\prod_{c=0}^{\infty}\prod_{i=1}^{M} R_{i,c}(z) - \sum_{c=0}^{\infty}\sum_{i=1}^{M}\vartheta_i J_{i,c}(z)E_{i,c}(z) + \vartheta_1 \sum_{\mathbf{n}\in U_1(N)} p(\mathbf{n})u_{1,-1}(\mathbf{n},z)\right], \tag{24}$$

where $u_{1,-1}(\mathbf{n},z) = z^{n_1}$ for $\mathbf{n}\in U(N)$.

Next, defining $\gamma_{i,c} = (\lambda_i/\lambda_1)L'_{i,c}(1)$, $i = 1,\ldots,M$, $c \geqslant -1$, then differentiating equation (24) with respect to z two times, and setting $z = 1$, we obtain

$$k'_1(\mathbf{1}) = \Phi\lambda_1(1-\rho_1)\left[\frac{\sum_{i=1}^{M} E[R_i] + (N\vartheta_i)/\Lambda}{1-\rho}\right], \tag{25}$$

and

$$k''_1(\mathbf{1}) = \Phi\lambda_1^2\left[\mathrm{Var}(R_M) + \frac{N(N-1)\vartheta_1}{\Lambda^2} + \sum_{i=1}^{M}\left(\frac{\Gamma_i}{\rho_i^2}\right)\left(\mathrm{Var}(R_{i-1}) + \lambda_i E\left[B_i^2\right]E[C] + \frac{N(N-1)\vartheta_i}{\Lambda^2}\right) + \left(\sum_{i=1}^{M} E[R_i]\left(\frac{1-\rho_1}{1-\rho}\right)\right)^2 + 2\sum_{i=1}^{M}\frac{N\vartheta_i}{\Lambda}\sum_{c=0}^{\infty}\left(\frac{\gamma_{i,c}t_{i,c}}{\rho_i}\right)\right], \tag{26}$$

where $t_{i,c} = E'_{i,c}(1)/\lambda_1$, $i = 1,\ldots,M$ and $c \geqslant 0$. That is,

$$t_{i,c} = \sum_{j=i+1}^{M} E[R_{j-1}]\frac{\gamma_{j,c}}{\rho_j} + \sum_{l=0}^{c-1}\sum_{k=1}^{M} E[R_{k-1}]\frac{\gamma_{k,l}}{\rho_k} + E[R_M], \tag{27}$$

and

$$\Gamma_i \triangleq \sum_{c=0}^{\infty}\gamma_{i,c}^2. \tag{28}$$

Notice that in equation (26), we have presented a much simplified form of the second factorial moment. Since the mean queue length at a station 1 beginning instant can also be written as $k'_1(\mathbf{1}) = \Phi\lambda_1(1-\rho_1)E[C]$, we have the following new definition for the average cycle length:

$$E[C] = \frac{\sum_{i=1}^{M} E[R_i] + (N\vartheta_i)/\Lambda}{1-\rho}. \tag{29}$$

Finally, substituting equations (25), (26) and (29) into relation (22), we obtain

$$E[W_1] = \frac{1-\rho_1}{2E[C]}\left[\frac{E[R_T]}{1-\rho}\right]^2 + \frac{\lambda_1 E[B_1^2] + [\mathrm{Var}(R_M) + N(N-1)\vartheta_1/\Lambda^2]/E[C]}{2(1-\rho_1)} + \sum_{i=1}^{M}\left(\frac{\Gamma_i}{\rho_i^2}\right)\frac{\lambda_i E[B_i^2] + [\mathrm{Var}(R_{i-1}) + N(N-1)\vartheta_i/\Lambda^2]/E[C]}{2(1-\rho_1)} + \sum_{i=1}^{M}\frac{N\vartheta_i}{\Lambda E[C]}\sum_{c=0}^{\infty}\frac{\gamma_{i,c}t_{i,c}}{\rho_i(1-\rho_1)}. \quad (30)$$

Another way of calculating $k_i''(\mathbf{1})$, $i = 1, \ldots, M$, is to differentiate functional relations for joint queue lengths twice and to then set $\mathbf{z} = \mathbf{1}$. This method requires the solution of M^3 equations in that many unknowns in order to get the M terms of interest (see, e.g., [18]). But the coefficient matrix of this equation set is sparse and symmetric, and upon carefully manipulating its rows, we are able to relate the weighted sum of the second derivatives of $k_i(\mathbf{z})$ to their first derivatives. The latter can be written explicitly in terms of system parameters. By substituting this weighted sum of $k_i''(\mathbf{1})$'s into the formula for the pseudo conservation law for E service models, we obtain

$$\sum_{j=1}^{M}\rho_j E[W_j] = \frac{\rho}{2(1-\rho)}\sum_{i=1}^{M}\left(\lambda_i E\left[B_i^2\right] + \frac{N(N-1)\vartheta_i/\Lambda^2}{E[C]}\right) + \frac{\rho E[R_T^2]}{2(1-\rho)E[C]} + \frac{E[R_T]}{2(1-\rho)}\left(\rho^2 - \sum_{i=1}^{M}\rho_i^2\right) + \sum_{i=1}^{M}\frac{N\rho_i Y_i}{1-\rho}, \quad (31)$$

where Y_i is defined as

$$Y_i = \sum_{j=1}^{i-1}\frac{E[R_j]}{\Lambda E[C]}\left(\sum_{k=i+1}^{M}\vartheta_k + \sum_{k=1}^{j}\vartheta_k\right) + \sum_{j=i+1}^{M}\frac{E[R_j]}{\Lambda E[C]}\sum_{k=i+1}^{j}\vartheta_k, \quad i = 1, \ldots, M. \quad (32)$$

The above expression is explicit if the (scaled) empty system probabilities ϑ_i, $i = 1, \ldots, M$, are calculated from (12). In equations (30) and (31) by putting $N = 0$ we get, respectively, the mean waiting time and the pseudo-conservation law for the system in which the server never stops [3]. Similarly, when $N = 1$ is substituted, we get the corresponding results for the patient server model [16]. Note that, Srinivasan and Gupta [16] define the switchover times differently: in their notation station $i \to i+1$ switchover time is denoted by R_{i+1}.

Globally gated service regime

Since the server stops upon finding the system empty at a system polling instant, a cycle might contain an idle period. We denote this idle period by the random

variable I. Also, we define the portion of the cycle time in which the server is busy serving customers or switching from one station to the next as the *busy segment* and denote it by the random variable S. Thus, $C = I + S$.

If the server becomes idle at a polling instant, it remains idle until N customers accumulate in the system. Therefore, I has a N-phase Erlang distribution with probability $F(\mathbf{0})$, and it is zero with probability $1 - F(\mathbf{0})$. The mean server idle period per cycle is

$$E[I] = \frac{NF(\mathbf{0})}{\Lambda}, \tag{33}$$

and higher moments can be calculated similarly.

Next, we use workload balancing arguments under stationary conditions to derive an expression for the mean cycle length. Under stationary conditions, the expected workload reduction during the busy segment of a cycle must be sufficient to offset the expected workload increase during the entire cycle (recall that $C = I + S$). Notice that this means that the workload reduction possible during the expected length of the cycle is strictly greater than the expected increase in workload during one cycle, which is the key condition for stability derived in Altman et al. [1, remark, p. 42]. Therefore, the busy segment must be equal to the sum of all switchover times (in a cycle) and the service times of all customers that arrive during an average cycle, i.e.,

$$E[S] = E[R_T] + \sum_{i=1}^{M} \lambda_i E[C] E[B_i]. \tag{34}$$

Using relations (33) and (34) we obtain

$$E[C] = \frac{E[R_T] + NF(\mathbf{0})/\Lambda}{1 - \rho}. \tag{35}$$

We further classify customers with respect to the server status at their arrival instant. A customer who finds the server idle is *class-I* and a customer who finds the server busy (with service or switching) is a *class-S* customer. Then, the mean waiting time of a type-i customer can be calculated as

$$E[W_i] = E\big[W_i^I\big] P_I + E\big[W_i^S\big] P_S, \tag{36}$$

where P_I and P_S are the probabilities that a customer belongs to class-I and class-S, respectively ($P_I + P_S = 1$). Similarly, $E[W_i^I]$ and $E[W_i^S]$ denote the mean waiting time of a type-i customer which belongs to class-I and class-S, respectively.

A class-I customer finds the system empty upon its arrival. Recall that $F(\mathbf{0})$ is the probability of finding the system empty at a system polling instant. However, P_I is the probability of finding the system empty at an arbitrary point in time and is equal to

$$P_I = \frac{NF(\mathbf{0})}{\Lambda E[C]}. \tag{37}$$

Now, we analyze the waiting time distribution of an arbitrary type-i customer with respect to its class. Consider a tagged type-i customer that also belongs to class-I, then its waiting time W_i^I is composed of the following periods: the remainder of the idle period, the delay due to the service and switchover times of station-j, $j = 1, \ldots, i-1$, and the delay due to the service times of type-i customers which arrive during the same idle period, but before the tagged customer. Therefore,

$$E\big[W_i^I\big] = \sum_{j=1}^{i-1} E[R_j] + \frac{N-1}{2\Lambda}\left(1 + 2\sum_{j=1}^{i-1}\rho_j + \rho_i\right). \tag{38}$$

In case the tagged type-i customer belongs to class-S, the next cycle has no idle period and the waiting time of such a customer, W_i^S, is composed of the following: residual length of the busy segment in which it arrives, total service time of all type-j, $j < i$, arrivals during the same residual length of the busy segment, total service time of all type-j, $j \leqslant i$, (including type-i customers who arrive before the tagged customer in the same cycle) arrivals during the portion of the busy segment that has already elapsed, and the total time necessary for the server to switch from station N to station-i. Therefore,

$$E\big[W_i^S\big] = \sum_{j=1}^{i-1} E[R_j] + \left(1 + 2\sum_{j=1}^{i-1}\rho_j + \rho_i\right)\frac{E[S^2]}{2E[S]}. \tag{39}$$

In continuously roving server ($N = 0$) case all customers belong to class-S, and hence the above relationship can also be obtained from equation (26) in Boxma et al. [4]. Substituting relations (37)–(39) into (36), we get

$$E[W_i] = \sum_{j=1}^{i-1} E[R_j] + \left(1 + 2\sum_{j=1}^{i-1}\rho_j + \rho_i\right)\left(\frac{N(N-1)F(\mathbf{0})}{2\Lambda^2 E[C]} + \frac{E[S^2]}{2E[C]}\right). \tag{40}$$

The distribution of the busy segment can be calculated using the PGF of joint queue lengths at system polling instants. Note that the queue length of any station at a system polling instant is equal to the number of arrivals to that station during the previous busy segment S. Therefore,

$$F(1, \ldots, 1, z_i, 1, \ldots, 1) = \widetilde{S}(\lambda_i - \lambda_i z_i), \quad i = 1, \ldots, M. \tag{41}$$

Thus, the second moment of the busy segment can be obtained by differentiating equation (19) twice, and by substituting it into equation (40)

$$E[W_i] = \sum_{j=1}^{i-1} E[R_j] + \left(1 + 2\sum_{j=1}^{i-1}\rho_j + \rho_i\right)\left(\frac{\Lambda^2 E[R_T^2] + N(N-1)F(\mathbf{0})}{2\Lambda(1+\rho)(\Lambda E[R_T] + NF(\mathbf{0}))} + \frac{\sum_{j=1}^{M}\lambda_j E[B_j^2] + 2\rho E[R_T]}{2(1-\rho^2)}\right), \quad i = 1, \ldots, M. \tag{42}$$

In equation (42) by setting $N = 0$ and $N = 1$ we can obtain results of Boxma et al. [4] for the continuously roving server model, and Borst [2] for the dormant server model, respectively. Also, from the same equation we can derive the pseudo-conservation law for the N-threshold polling systems with globally gated service regime as follows:

$$\sum_{i=1}^{M} \rho_i E[W_i] = \sum_{i=1}^{M} \rho_i \sum_{j=1}^{i-1} E[R_j] + \frac{\rho(\Lambda^2 E[R_T^2] + N(N-1)F(\mathbf{0}))}{2\Lambda(\Lambda E[R_T] + NF(\mathbf{0}))} + \frac{\rho \sum_{j=1}^{M} \lambda_j E[B_j^2] + 2\rho^2 E[R_T]}{2(1-\rho)}. \tag{43}$$

5. Numerical results

In this section, we set out to calculate the threshold level that minimizes the mean unfinished work in system. Although it is difficult to obtain an explicit expression for the optimum threshold (since empty system probabilities cannot be calculated explicitly) we do manage to show that there exists a (finite) critical threshold level $\overline{N}$ that bounds the optimum N^*. Thus, it suffices to enumerate system performance in the interval $[0, \overline{N}]$ only, in order to find the optimum threshold level. Recall that in previous sections we omitted parameter N and simply used W to denote the waiting time. However, since in this section we want to investigate the effect of the threshold level on the waiting time, we use notation $W(N)$ for the waiting time in a system with the threshold level N. We also show argument N for all other relevant notation (e.g., cycle length C, empty system probabilities ϑ_i's, etc.) to emphasize their dependence on the threshold level. Finally, $\overline{W(N)}$ is used to denote the mean unfinished work in system (also called the pseudo-conservation law) and this quantity is the primary measure of system performance in all numerical examples reported here.

We construct 7 examples – three for symmetric, and four for asymmetric systems, and examine the optimum threshold level for E and GG service regimes. All examples have 5 stations. Data sets I, II and III correspond to symmetric systems. Data sets I and II have $\Lambda = 0.1$, and exponential service times with mean $E[B] = 0.4$. In data set I, the switchover time has an exponential distribution with $E[R] = 1$, and in set II, the switchover time is either 1 with probability 0.9, or 100 with probability 0.1. Data set III has the same switchover time and service time distribution as set II, but the total arrival rate $\Lambda = 1$.

All asymmetric systems have the following common parameters: $\rho = 0.04$, $\lambda_i = 0.02$, $i = 1, \ldots, 5$, and service time distributions are exponential with station 1 having 75% of the total work load, i.e., $E[B_1] = 1.5$ and $E[B_j] = 0.125$, $j > 1$. For data set IV, R_i is either 1 with probability 0.9, or 100 with probability 0.1, for all i. For data set V, R_5 is either 20 or 95 with probabilities 0.8 and 0.2, and for data set VI, R_5 is either 64 or 130 with probabilities 0.75 and 0.25, respectively. R_i, $i < 5$, are identical in data sets IV, V and VI. Notice that $E[R_5]$ equals 10.9, 35 and 80.5

for data IV, V and VI, respectively, and Var(R_5) remains (almost) unchanged at 880. Data VII is a copy of data VI, with the only difference that R_5 and R_1 have been switched. Thus, while station 1 is still the heavily loaded station, $E[R_1] = 80.5$, and $E[R_j] = 10.9$, $j = 2, 3, 4, 5$.

Exhaustive service regime

Let $R_i = R$, $B_i = B$ and $\lambda_i = \lambda$, $i = 1, \ldots, M$, i.e., suppose we have a symmetric system. Then, $\lambda_i = \Lambda/M$, $\rho_i = \rho/M$ and $p_i = 1/M$. Furthermore, the empty system probabilities ϑ_i are same for all stations, and we denote them by $\vartheta(N)$, $N \geqslant 0$. Symmetry allows us to greatly simplify the analysis presented in the previous sections. For example, the mean cycle length becomes

$$E\big[C(N)\big] = \frac{M(\Lambda E[R] + N\vartheta(N))}{\Lambda(1-\rho)}. \tag{44}$$

Since there is only one empty system probability to find, equations (12) reduce to a single equation, and the threshold dependent terms $J_{i,c}(\mathbf{0})$, $i = 1, \ldots, M$, and $c \geqslant 0$, can be simplified further. The resulting equation is

$$\vartheta(N) = \frac{\prod_{i=1}^{M}\prod_{c=0}^{\infty} R_{i,c}(\mathbf{0})}{\sum_{i=1}^{M}\sum_{c=0}^{\infty}(1-\bar{\eta}_{i,c}^{N})E_{i,c}(\mathbf{0})}, \tag{45}$$

where $\bar{\eta}_{i,c}$ is the average of $L_{j,c}(\mathbf{0})$ terms, that is,

$$\bar{\eta}_{i,c} = (1/M)\left(\sum_{j=i}^{M} L_{j,c}(\mathbf{0}) + \sum_{j=1}^{i-1} L_{j,c-1}(\mathbf{0})\right). \tag{46}$$

The mean waiting time can now be obtained explicitly as shown below (after simplifications):

$$\begin{aligned} E\big[W(N)\big] = {} & \frac{(N-1+(M-1)\Lambda E[R])N\vartheta(N) + \Lambda^2(E[R^2] + (M-1)E[R]^2)}{2\Lambda(N\vartheta(N) + \Lambda E[R])} \\ & + \frac{\Lambda E[B^2] + \rho(M-1)E[R]}{2(1-\rho)}. \end{aligned} \tag{47}$$

Remark. By setting $M = 1$, we obtain the mean waiting time for the M/G/1 queue model operating under the N-policy (Heyman and Sobel, [10, pp. 444–447]). Note that for a single station model the switchover time is zero, i.e., $E[R] = 0$, and the system is empty at all server departure epochs, i.e., $\vartheta(N) = 1$. Upon substituting these variables in equation (47), we get

$$E\big[W(N)\big] = \frac{(N-1)}{2\Lambda} + \frac{\Lambda E[B^2]}{2(1-\rho)}. \tag{48}$$

Accounting for differences in notation, equation (48) is the same as equation (11-117a) of Heyman and Sobel [10, p. 445]. Similarly, upon setting $M = 1$ and $E[R_T] = 0$

in equation (42) of the GG service policy, we obtain equation (48). This makes sense since the GG service discipline behaves exactly like the E service discipline when $M = 1$; at every server departure instant from station 1, the server immediately restarts working at the same queue, unless the system is empty.

Liu et al. [13] have shown that for symmetric systems a patient server ($N = 1$) is superior to the continuously roving server protocol ($N = 0$) for minimizing unfinished work in system. From the explicit representation of expected work in system in equation (47), it is easy to confirm this result from our analysis as well. What is even more interesting that we can use the $N = 0$ model as a benchmark to find a range of values $1 \leqslant N \leqslant \overline{N}$ outside which $\overline{W(N)} > \overline{W(0)}$ (or equivalently $E[W(N)] > E[W(0)]$, because of symmetry).

Theorem 1. For symmetric systems with E service regime, there exists a $\overline{N} \geqslant 1$, such that $\overline{W(N)} > \overline{W(0)}$ for $N > \overline{N}$, where

$$\overline{N} = 1 + \left\lfloor \frac{\Lambda E[R^2]}{E[R]} \right\rfloor. \tag{49}$$

Proof. Let $\Delta(N) \stackrel{\triangle}{=} \overline{W(N)} - \overline{W(0)}$. Then using equation (47) we obtain

$$\Delta(N) = \frac{[(N-1)E[R] - \Lambda E[R^2]]N\rho\vartheta(N)}{2\Lambda E[R](N\vartheta(N) + \Lambda E[R])}. \tag{50}$$

Since $\vartheta(N) > 0$ for all $N \geqslant 0$, $\Delta(N) > 0$ if and only if $N > 1 + \Lambda E[R^2]/E[R]$. Thus, setting $\overline{N}$ to be 1 plus the integer floor of $\Lambda E[R^2]/E[R]$ completes the proof. □

Theorem 1 is useful since in order to find the optimum threshold level, N^*, it is now sufficient to enumerate $\overline{W(N)}$, in the interval $[1, \overline{N}]$. Furthermore, we strongly believe that the optimal N is the smallest positive integer for which the mean unfinished work in system increases upon increasing the threshold by 1. There are two lines of reasoning behind this belief. First, in all numerical experiments, $\overline{W(N)}$ has turned out to be convex in the threshold level N whenever $0 \leqslant N \leqslant \overline{N}$. Secondly, we have been able to prove certain properties of underlying functions that suggest convexity (though we lack formal proof). For example, we have been able to show that $\vartheta(N)$ is convex and strictly decreasing in N, and that $N\vartheta(N)$ is strictly increasing in N. In order to establish convexity of $E[W(N)]$ (and thus of $\overline{W(N)}$, in the symmetric case), we need to show that $N\vartheta(N)$ is concave in $[0, \overline{N}]$. Although, our analysis suggests that $N\vartheta(N)$ should increase in N with a decreasing rate, a formal proof eludes us since that requires an explicit expression for $\vartheta(N)$; a quantity that we can only find numerically.

Like their symmetric counterparts, asymmetric systems also have a critical threshold level $\overline{N}$ after which increasing the start-up threshold does not improve system performance. Notice that in this case, the optimum threshold can be 0 (see Srinivasan and

Gupta [16] for some examples). Unfortunately, it is difficult to find $\overline{N}$ in an explicit form similar to relationship (49). However, as the following theorem proves, there exists an upper bound for the critical threshold $\overline{N}$.

Theorem 2. For asymmetric systems with E service regime, there exists a $\overline{N} \geqslant 1$, such that $\overline{W(N)} > \overline{W(0)}$ for $N > \overline{N}$, where

$$\overline{N} \leqslant 1 + \left\lfloor \frac{\Lambda E[R_T^2]}{E[R_T]} \right\rfloor. \tag{51}$$

Proof. For the asymmetric model, using equations (31) and (32) the difference function $\Delta(N)$ can be written as

$$\Delta(N) = \frac{[(N-1)E[R_T] - \Lambda E[R_T^2]]N\rho \sum_{i=1}^{M} \vartheta_i(N)}{2\Lambda E[R_T](N \sum_{i=1}^{M} \vartheta_i(N) + \Lambda E[R_T])} + \frac{N \sum_{i=1}^{M} \rho_i Y_i}{(1-\rho)}. \tag{52}$$

Since $Y_i > 0$ for all $i = 1, \ldots, M$,

$$\Delta(N) > \frac{[(N-1)E[R_T] - \Lambda E[R_T^2]]N\rho \sum_{i=1}^{M} \vartheta_i(N)}{2\Lambda E[R_T](N \sum_{i=1}^{M} \vartheta_i(N) + \Lambda E[R_T])}. \tag{53}$$

Thus, $N \geqslant 1 + \Lambda E[R_T^2]/E[R_T]$ is sufficient to ensure that $\Delta(N) > 0$ (since $\vartheta_i(N) > 0$). Hence, the theorem is proved. □

The optimum threshold levels N^* and the critical threshold value $\overline{N}$ for each symmetric system data set are presented in table 1. Notice that the critical threshold level $\overline{N}$ increases with increasing variance of switchover time, and with increasing total arrival rate (as seen in equation (49)). However, at high arrival rates the mean waiting time appears to be robust with respect to the threshold level and small changes in N do not change $E[W(N)]$ (or $\overline{W(N)}$) significantly (see the third row of table 1).

Performance of systems corresponding to data sets IV, V, VI and VII, as measured by pseudo-conservation law is shown in figure 1. We observe that, asymmetric systems have a finite critical threshold value $\overline{N}$ and that $\overline{N}$ is influenced a great deal by the total arrival rate and the variance of switchover times. The critical threshold level $\overline{N}$ and, thus, N^* decrease when the mean switchover time to the heavily loaded station is large relative to the mean switchover times to low traffic stations. We also observe that

Table 1
Symmetric systems input data.

Data set	$\overline{N}$	N^*	$E[W(0)]$	$E[W(N^*)]$	$\overline{W(0)}$	$\overline{W(N^*)}$
I	1	1	3.100	2.292	0.124	0.092
II	10	3	68.638	56.148	2.746	2.246
III	92	45	82.513	81.875	33.005	32.750

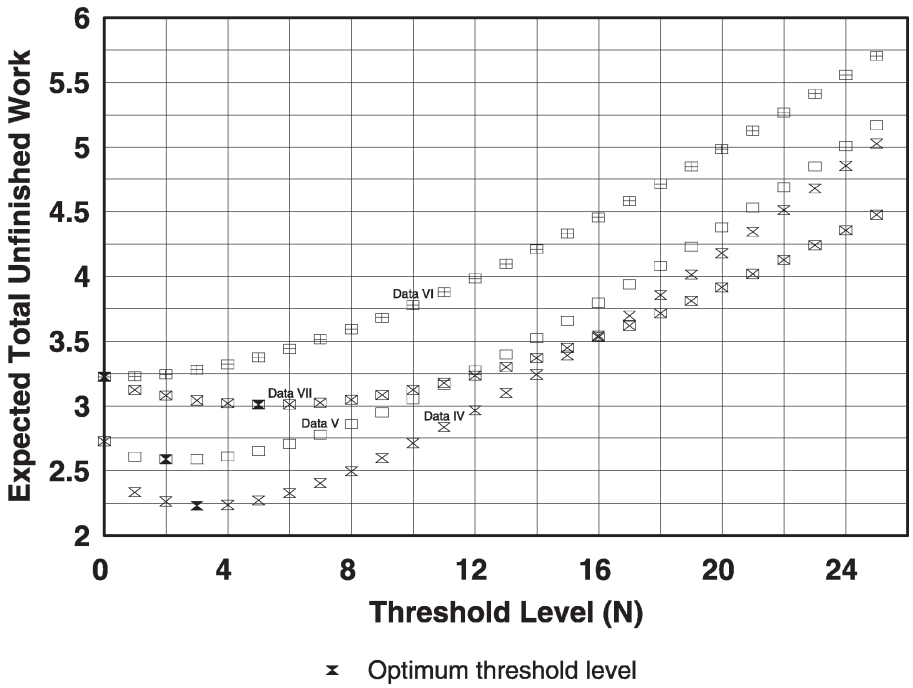

Figure 1. Exhaustive service regime.

the right-hand side of inequality (51) is not affected by this decrease in $\overline{N}$, implying that the upper bound is relatively more loose when both processing and switchover times are not balanced.

Globally gated service regime

Since in the GG service systems, there is only one unknown empty system probability, $F(\mathbf{0})$, the analysis of asymmetric systems is not any harder than symmetric ones. Furthermore, even in symmetric GG systems the mean waiting times differ from one station to the next. They depend significantly on the order of station visitation (sequence). In contrast, station mean waiting times in symmetric E service models are obviously not affected by the order of visitation, and furthermore, the impact of sequencing is small even in asymmetric models.

In a recent study, Borst [2] showed that for a fixed sequence the dormant server ($N = 1$) model dominates the continuously roving model in the sense of having a smaller mean unfinished work in system. Now, we extend his results to find the optimum threshold value N^*. Using $N = 0$ model as a benchmark, theorem 3 presents a range of threshold levels, $[1, \overline{N}]$, that contains the optimum threshold level. We omit the proof of this theorem, since it is similar to the exhaustive service case.

Table 2
Symmetric GG service models.

Data set	$\overline{N}$	N^*	$\overline{W(0)}$	$\overline{W(N^*)}$
I	1	1	0.209	0.144
II	13	5	5.962	5.176
III	138	68	50.445	50.444

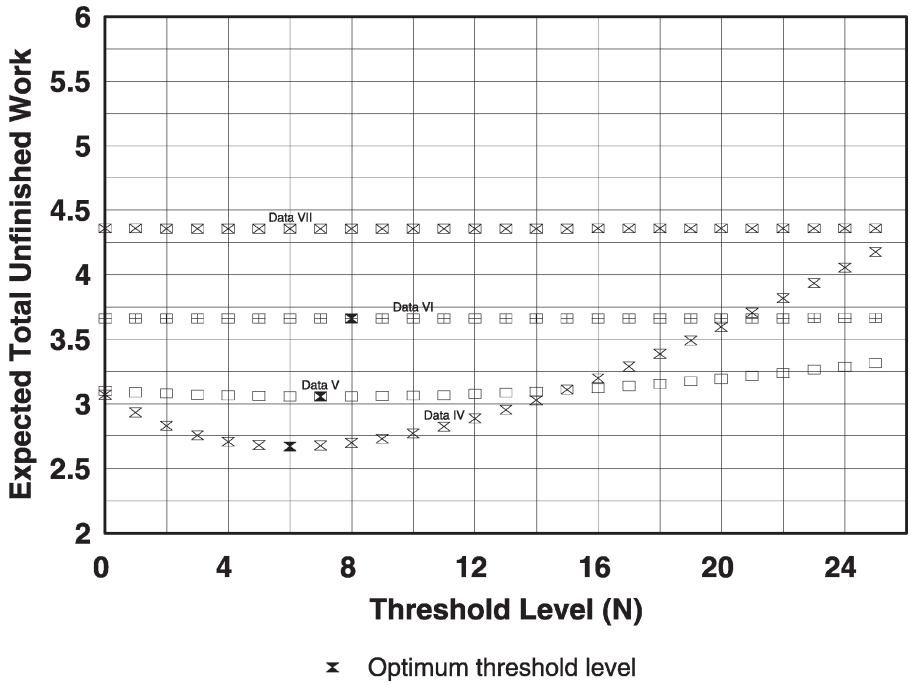

Figure 2. Globally gated service regime.

Theorem 3. For GG service systems, there exists a $\overline{N} \geqslant 1$, such that $\overline{W(N)} > \overline{W(0)}$ for $N > \overline{N}$, where

$$\overline{N} = 1 + \left\lfloor \frac{\Lambda E[R_T^2]}{E[R_T]} \right\rfloor. \tag{54}$$

For symmetric system examples (data sets I, II and III) optimum threshold level N^* and the critical value $\overline{N}$ are presented in table 2. Performance of asymmetric systems corresponding to data sets IV, V, VI and VII, is shown in figure 2.

Notice that, both N^* and $\overline{N}$ values are almost the same for E and GG service regimes. However, when optimum threshold values are used, the mean unfinished work in system for E service regime is considerably less than that of GG service model. Comparison of figures 1 and 2 shows that although E service model performs better

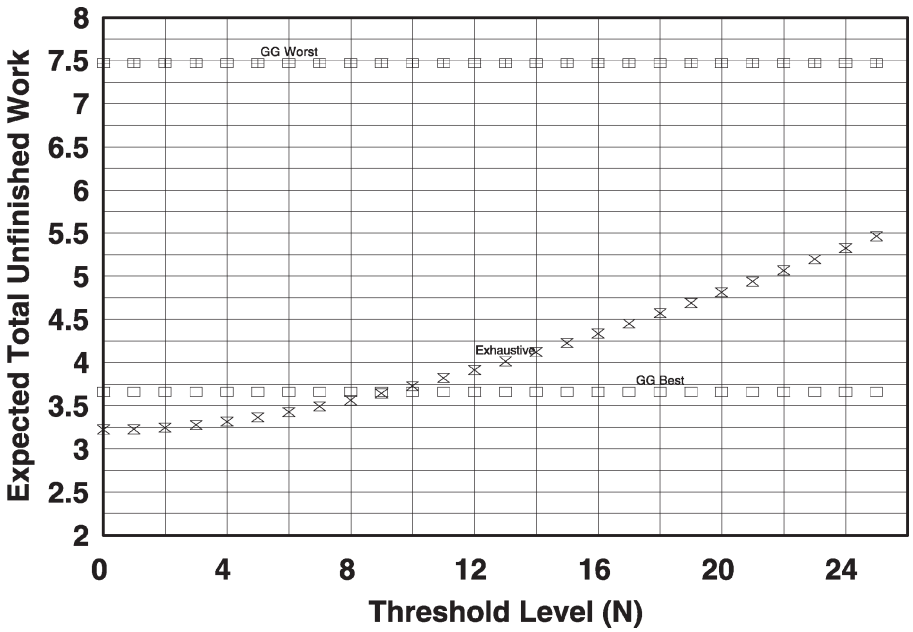

Figure 3. Sequence effect.

than GG service model when thresholds are chosen optimally, this advantage is quickly lost as N becomes large.

Even in asymmetric GG systems, the performance measure is not affected significantly by the threshold level. Furthermore, when switchover times are large, $\overline{W(N)}$ varies even less with N. When switchover times are interchanged (see data VI and VII), the performance measure is affected significantly by this change; even though the optimum threshold does not change much. This unusual effect led us to investigate the impact of the order of visiting stations. We generated all possible sequences for data VI and identified the best and worst sequences for the performance measure $\overline{W(N)}$. We found that sequence can affect the performance measure significantly and that its effect is independent of N. Figure 3 shows the performance measure for the best and worst sequences for the GG service regime. In contrast, the maximum spread between the best and the worst sequence for the E service regime was at most 4.3% for $N \leqslant 25$. Therefore, the third curve in figure 3, marked "exhaustive", represents $\overline{W(N)}$ that corresponds to the best sequence for each threshold level under the E policy. Note that in this case, the best sequence may be different for different values of N.

6. Extensions

The mean unfinished work in system is affected by the order in which the server visits stations. Our numerical studies show that the E service systems are relatively insensitive to the sequence in which stations are arranged. However, the sequence has a

large effect on the mean unfinished work in system for GG service models. Therefore, we next show how our analysis can be used to provide an optimal sequence.

We assume that there are no physical or user defined constraints in regard to ordering of stations in the system and let Ω denote the set of all possible sequences. Then Ω is the set of all possible permutations of the $1 \times M$ vector representing station sequences, and $|\Omega| = M!$. For a sequence $\omega \in \Omega$, the station index at location i, for $i = 1, \ldots, M$, is given by $\omega(i) = j$, $j = 1, \ldots, M$. Theorem 4, gives the ordering rule for generating the optimal sequence of stations.

Theorem 4. In a GG service system with N-threshold start-up policy the best sequence ω^* that minimizes $\overline{W(N)}$, for $N \geqslant 0$, satisfies the following condition:

$$\frac{E[R_{\omega^*(i)}]}{\rho_{\omega^*(i)}} \leqslant \frac{E[R_{\omega^*(i+1)}]}{\rho_{\omega^*(i+1)}}, \quad i = 1, \ldots, M-1. \tag{55}$$

Proof. Given that ω^* satisfies equation (55) we want to show that from any sequence $\omega \in \Omega$, and $\omega \neq \omega^*$, we can construct the sequence ω^* by a finite number of interchanges involving neighboring stations, such that at each interchange the objective function, i.e., $\overline{W(N)}$, improves. Since $\omega \neq \omega^*$, there exists at least one pair of stations that does not satisfy equation (55). Let $(k, k+1)$ be the first such pair in the sequence, for $k = 1, \ldots, M-1$, i.e.,

$$\frac{E[R_{\omega(k)}]}{\rho_{\omega(k)}} > \frac{E[R_{\omega(k+1)}]}{\rho_{\omega(k+1)}}. \tag{56}$$

By switching stations in kth and $(k+1)$th place in the sequence, we obtain a new sequence ω' such that $\omega'(i) = \omega(i)$, $i \neq k, k+1$, and $\omega'(k) = \omega(k+1)$ and $\omega'(k+1) = \omega(k)$. Let $\Delta = \overline{W(N,\omega)} - \overline{W(N,\omega')}$, where $\overline{W(N,\omega)}$ and $\overline{W(N,\omega')}$ denote the objective function of sequences ω and ω', respectively. From equation (43) we get

$$\Delta = \rho_{\omega(k+1)} E[R_{\omega(k)}] - \rho_{\omega(k)} E[R_{\omega(k+1)}]. \tag{57}$$

From equation (56), Δ is clearly positive. Thus, by switching stations in kth and $(k+1)$th positions, we improve the objective function. Proceeding with such interchanges, the number of stations which do not satisfy equation (55) decreases, and at each step the objective function improves. The sequence ω^* is reached after a finite number of interchanges [15]. Note that when $E[R_i]/\rho_i = E[R_j]/\rho_j$, for some i and j, ω^* is not a unique sequence. □

Fortunately, the best sequence for the GG service regime does not depend on the threshold level N. Thus, first finding the best sequence of stations and then searching for the optimum threshold level in the $[1, \overline{N}]$ range will give the minimum mean unfinished work in system.

Acknowledgements

Authors are grateful to Professor R.B. Cooper of Florida Atlantic University for his remarks on an earlier version of this article. This research has been supported in part by Natural Sciences and Engineering Research Council of Canada through research grant No. OGP0045904.

References

[1] E. Altman, P. Konstantopoulos and Z. Liu, Stability, monotonicity and invariant quantities in general polling systems, Queueing Systems 11 (1992) 35–57.
[2] S.C. Borst, A globally gated polling system with a dormant server, Probab. Engrg. Inform. Sci. 9 (1995) 239–254.
[3] O.J. Boxma and W.P. Groenendijk, Pseudo-conservation laws in cyclic-service systems, J. Appl. Probab. 24 (1987) 949–964.
[4] O.J. Boxma, J.A. Weststrate and U. Yechiali, A globally gated polling system with server interruptions, and applications to the repairman problem, Probab. Engrg. Inform. Sci. 7 (1993) 187–208.
[5] P.J. Burke, Delays in single-server queues with batch input, Oper. Res. 23 (1975) 830–833.
[6] R.B. Cooper, *Introduction to Queueing Theory*, 3rd ed. (CEE Press, 1990).
[7] M. Eisenberg, The polling system with a stopping server, Queueing Systems 18 (1994) 387–431.
[8] S.W. Fuhrmann and R.B. Cooper, Stochastic decompositions in the $M/G/1$ queue with generalized vacations, Oper. Res. 33 (1985) 1117–1129.
[9] Y. Günalay and D. Gupta, Threshold start-up control policy for polling systems, Working paper, McMaster University, Hamilton, Ontario, Canada (1996).
[10] D.P. Heyman and M.J. Sobel, *Stochastic Models in Operations Research, Vol. I: Stochastic Processes and Operating Characteristics* (McGraw-Hill, New York, 1982).
[11] D.P. Heyman and M.J. Sobel, *Stochastic Models in Operations Research, Vol. II: Stochastic Optimization* (McGraw-Hill, New York, 1984).
[12] A.G. Konheim, H. Levy and M.M. Srinivasan, Descendant set: An efficient approach for the analysis of polling systems, Trans. Commun. 42 (1993) 1245–1253.
[13] Z. Liu, P. Nain and D. Towsley, On optimal polling policies, Queueing Systems 11 (1992) 59–83.
[14] J.A.C. Resing, Polling systems and multitype branching processes, Queueing Systems 13 (1993) 409–426.
[15] W.E. Smith, Various optimizers for single-stage production, Naval Res. Logist. Quart. 3 (1956) 59–66.
[16] M.M. Srinivasan and D. Gupta, When should a server be patient?, Mangm. Sci. 42 (1996) 437–451.
[17] M.M. Srinivasan, S.C. Niu and R.B. Cooper, Relating polling models with nonzero and zero switchover times, Queueing Systems 19 (1995) 149–168.
[18] H. Takagi, *Analysis of Polling Systems* (MIT Press, Cambridge, MA, 1986).
[19] H. Takagi, Queuing analysis of polling models: An update, in: *Stochastic Analysis of Computer and Communication Systems*, ed. H. Takagi (Elsevier/North-Holland, 1990) pp. 267–318.
[20] H. Takagi, *Queueing Analysis of Polling Models: Progress in 1990–1993* (Institute of Socio-Economic Planning, University of Tsukuba, Japan, 1994).