

When Does Forced Idle Time Improve Performance in Polling Models?

Robert B. Cooper • Shun-Chen Niu • Mandyam M. Srinivasan

Department of Computer Science and Engineering, Florida Atlantic University, Boca Raton, Florida 33431-0991

School of Management, The University of Texas at Dallas, Richardson, Texas 75083-0688

Management Science Program, College of Business Administration, The University of Tennessee,
Knoxville, Tennessee 37996-0562

Sarkar and Zangwill (1991) showed by numerical examples that reduction in setup times can, surprisingly, actually *increase* work in process in some cyclic production systems (that is, reduction in switchover times can increase waiting times in some polling models). We present, for polling models with exhaustive and gated service disciplines, some explicit formulas that provide additional insight and characterization of this anomaly. More specifically, we show that, for both of these models, there exist simple formulas that define for each queue a critical value z^* of the mean total setup time z per cycle such that, if $z < z^*$, then the expected waiting time at that queue will be *minimized* if the server is forced to idle for a constant length of time $z^* - z$ every cycle; also, for the symmetric polling model, we give a simple explicit formula for the expected waiting time and the critical value z^* that minimizes it.

(Polling Models; Cyclic Production Systems; Setup Times; Switchover Times; Waiting Times; Variance Paradox; Vacation Models; Decomposition)

1. Introduction

In their paper "Variance Effects in Cyclic Production Systems," Sarkar and Zangwill (1991) gave numerical examples that show that, in some cases, reduction in setup times (or in service times) can, surprisingly, actually *increase* work in process in cyclic production systems (or equivalently, waiting times in polling models). Although they offered an intuitive "explanation" for this anomaly, they did not characterize it in precise mathematical terms, which makes it difficult to say more than the following: The intuitively "obvious" fact that reducing overhead always increases efficiency is false.

Subsequently, Zangwill (1992) published in *Interfaces* "The Limits of Japanese Production Theory," in which he claimed that examples like those in Sarkar and Zangwill (1991) "expose a flaw" in Japanese production theory (which stresses the benefits of reducing overhead). This led to some heated criticism resulting in three articles that appeared in *Interfaces* (Duenyas 1994, Ger-

chak and Zhang 1994, and McIntyre 1994), together with a rejoinder by Zangwill and Sarkar (1994). These articles were prefaced by a statement from the Editor-in-Chief of *Interfaces* saying that Zangwill's (1992) article "alone drew more response than all other articles combined" during his editorship.

Other recent papers that have addressed this issue (the effects of switchover and/or setup times on waiting times in polling models) include Duenyas and Van Oyen (1994, 1996), Federgruen and Katalan (1996a, b), Gupta and Srinivasan (1996a, b), Olsen (1996a, b), Righter and Shanthikumar (1995), Samaddar and Kaul (1995), Srinivasan and Gupta (1996), and Van Oyen (1997).

Our objective here is not to enter the debate about Japanese production theory, but rather to cast some light on this startling anomaly by giving some explicit formulas that characterize it. That is, our purpose here is to develop simple explicit formulas for the expected waiting time as a function of the mean and variance of

the setup (switchover) times in standard polling models with either exhaustive or gated service discipline; this will provide additional insight into the counterintuitive phenomenon uncovered by Sarkar and Zangwill's numerical examples.

Our strategy is to consider the insertion of "forced idle time" instead of the reduction of existing setup times. Then, we explicitly show when it is possible to reduce the expected waiting time by inserting a forced idle time before or during a setup; and we specify the exact amount of forced idle time to be added to the sum of the setup times per cycle to minimize the waiting time at a given queue. (We also observe that the amount of this forced idle time will, in general, be different depending on the queue for which it is desired to minimize the expected waiting time, and that there is no single value of this sum that will simultaneously minimize the expected waiting time at every queue.)

More precisely, we give easily-computed formulas that define for each queue a critical value z^* of the mean total setup time z per cycle such that, if $z < z^*$, then the expected waiting time at that queue can be minimized by forcing the server to be idle each cycle for a time equal to $z^* - z$; and, for the symmetric polling model, we give simple closed-form formulas for the expected waiting time and the critical value z^* that yields its (global) minimum value.

In §2 we define the model, state our results, and give some numerical illustrations; and in §3 we give the details of the proofs.

2. The Model and the Optimal Setup Times

A single server serves in cyclic order a sequence of N infinite-capacity $M/G/1$ queues. Queue k ($k = 1, 2, \dots, N$) receives Poisson arrivals at rate λ_k , and has a service-time distribution with mean service time b_k and second moment $b_k^{(2)}$. We assume that a setup time Z_k , with mean $E[Z_k]$ and variance $V[Z_k]$, is incurred prior to each visit of the server to queue k . The arrival times, service times, and setup times are all mutually independent, and the service discipline is nonbiased (i.e., at each queue, the next customer selected for service does not depend on that customer's service time).

It may be helpful to think of Z_k as the sum $Z_k = R_k + S_k$, where R_k is the *switchover time* (the time required for the server to travel from queue $k - 1$ to queue k) and S_k is the *setup time* (the time required after the server arrives at queue k to prepare, or "setup," before beginning to serve any customers at queue k). In our model, we assume that a setup time will be required whether or not there are customers waiting when the server arrives. Then, in effect, the switchover times and the setup times can be treated as being the same; and because of the context of this paper, we will henceforth refer to Z_k as the setup time.

We will first limit our discussion to two standard service disciplines: exhaustive service and gated service. Extensions to more general service disciplines will be discussed in §3. With exhaustive service, the server leaves queue k only when there are no customers remaining in queue k ; and with gated service, the server closes a "gate" behind the waiting customers when its setup at queue k is complete and leaves queue k upon completion of service of all the customers in front of the gate.

Let $\rho_k \equiv \lambda_k b_k$, $\rho_T \equiv \rho_1 + \dots + \rho_N$, and $z \equiv E[Z_1] + \dots + E[Z_N]$. We assume that $\rho_T < 1$ (which guarantees stability) and that the system is in statistical equilibrium.

Let W_k be the waiting time (from arrival epoch to start of service) of a customer at queue k . We will show that for each k , $E[W_k]$ is convex in z ; and we will derive explicit formulas for the *optimal* value of z , denoted by $z^*(k)$, that minimizes $E[W_k]$. We will also derive similar results for the special case of a *symmetric* polling model, where each queue has the same arrival rate, the same service-time distribution, and the same setup-time distribution. For this latter case, we will suppress the queue index k in our notation; thus, for instance, $E[W]$ denotes the expected waiting time at any queue. We now formally state our main results.

THEOREM 1. *If the service discipline is exhaustive, then*

$$z^*(k) = \frac{1 - \rho_T}{1 - \rho_k} \sqrt{g_k}, \tag{1}$$

where g_k is a constant given by

$$g_k = V[Z_k] + \sum_{i=1}^N \frac{\Gamma_i^{(k)}}{\rho_i^2} V[Z_i]. \tag{2}$$

The coefficients $\{\Gamma_i^{(k)} : 1 \leq i \leq N\}$ in (2) are defined by

$$\Gamma_i^{(k)} = \sum_{c=0}^{\infty} (\gamma_{i,c}^{(k)})^2, \tag{3}$$

where: (i) For $k = 1$, the constants $\{\gamma_{i,c}^{(1)} : 1 \leq i \leq N; 0 \leq c < \infty\}$ are determined by the recursion

$$\gamma_{i,c}^{(1)} = \frac{\rho_i}{1 - \rho_i} \left[\sum_{j=i+1}^N \gamma_{j,c}^{(1)} + \sum_{j=1}^{i-1} \gamma_{j,c-1}^{(1)} \right], \tag{4}$$

with the initial condition $\gamma_{1,-1}^{(1)} = 1$ and $\gamma_{i,-1}^{(1)} = 0$ for $1 < i \leq N - 1$; and (ii) for $1 < k \leq N$, the constants $\{\gamma_{i,c}^{(k)} : 1 \leq i \leq N; 0 \leq c < \infty\}$ are determined by the same recursion (4) after renumbering the queues so that queue k becomes queue 1, queue $k + 1$ becomes queue 2, and so on.

In the special case of a symmetric polling model, the expected waiting time is given by the explicit formula

$$E[W] = \frac{\rho_T}{1 - \rho_T} \frac{b^{(2)}}{2b} + \frac{N}{2} \frac{V[Z]}{z} + \frac{z}{2} \frac{1 - \rho_T/N}{1 - \rho_T}, \tag{5}$$

and the minimizing value $z^*(k) = z^*$ is the same for all queues, given by

$$z^* = N \sqrt{V[Z] \frac{1 - \rho_T}{N - \rho_T}}. \tag{6}$$

THEOREM 2. If the service discipline is gated, then

$$z^*(k) = (1 - \rho_T) \sqrt{g_k}, \tag{7}$$

where g_k is a constant given by

$$g_k = \sum_{i=1}^N \frac{\Gamma_i^{(k)}}{\rho_i^2} V[Z_{i+1}]. \tag{8}$$

The coefficients $\{\Gamma_i^{(k)} : 1 \leq i \leq N\}$ in (8) are defined by

$$\Gamma_i^{(k)} = \sum_{c=0}^{\infty} (\gamma_{i,c}^{(k)})^2, \tag{9}$$

where: (i) For $k = 1$, the constants $\{\gamma_{i,c}^{(1)} : 1 \leq i \leq N; 0 \leq c < \infty\}$ are determined by the recursion

$$\gamma_{i,c}^{(1)} = \rho_i \left[\sum_{j=i+1}^N \gamma_{j,c}^{(1)} + \sum_{j=1}^i \gamma_{j,c-1}^{(1)} \right], \tag{10}$$

with the initial condition $\gamma_{1,-1}^{(1)} = 1$ and $\gamma_{i,-1}^{(1)} = 0$ for $1 < i \leq N$; and (ii) for $1 < k \leq N$, the constants $\{\gamma_{i,c}^{(k)} : 1 \leq i \leq N; 0 \leq c < \infty\}$ are determined by the same recursion (10)

after renumbering the queues so that queue k becomes queue 1, queue $k + 1$ becomes queue 2, and so on.

In the special case of a symmetric polling model, the expected waiting time is given by the explicit formula

$$E[W] = \frac{\rho_T}{1 - \rho_T} \frac{b^{(2)}}{2b} + \frac{N}{2} \frac{V[Z]}{z} + \frac{z}{2} \frac{1 - \rho_T/N}{1 - \rho_T}, \tag{11}$$

and the minimizing value $z^*(k) = z^*$ is the same for all queues, given by

$$z^* = N \sqrt{V[Z] \frac{1 - \rho_T}{N + \rho_T}}. \tag{12}$$

The theorems say that whenever $z^*(k)$ is larger than z , we can minimize the expected waiting time at queue k by inserting a deterministic value, equal to $z^*(k) - z$, to any of the individual setup times. Furthermore, since the waiting-time distribution is invariant with respect to deterministic shifts in setup times so long as the net shift is zero (Cooper et al. 1996, Srinivasan et al. 1995, Federgruen and Katalan 1996a, Righter and Shanthikumar 1995), this additional amount may be allocated in any manner among the individual setup times.

Finally, we note that it is easy to determine the optimal value z^* that minimizes $\sum_{k=1}^N \rho_k E[W_k]$, using the pseudo-conservation laws (Ferguson and Aminetzah 1985, Watson 1984, Boxma and Groenendijk 1987). For a number of service disciplines, the pseudo-conservation laws give simple closed-form formulas that express $\sum_{k=1}^N \rho_k E[W_k]$ in terms of the input parameters. In the case of exhaustive service, one has

$$\sum_{k=1}^N \rho_k E[W_k] = \frac{\rho_T}{2(1 - \rho_T)} \sum_{k=1}^N \lambda_k b_k^{(2)} + \frac{\rho_T E[(\sum_{k=1}^N Z_k)^2]}{2z} + \frac{z}{2(1 - \rho_T)} \left(\rho_T^2 - \sum_{k=1}^N \rho_k^2 \right). \tag{13}$$

The value z^* that minimizes the right-hand side of (13) can be obtained by differentiation, which yields

$$z^* = \sqrt{\left(\sum_{k=1}^N V[Z_k] \right) \frac{\rho_T(1 - \rho_T)}{\sum_{k=1}^N \rho_k(1 - \rho_k)}}. \tag{14}$$

(Similar results can be easily obtained for other service disciplines covered by the pseudo-conservation laws.)

To illustrate our theorems, we will first consider the 2-queue example given in Sarkar and Zangwill (1991) (p. 448). Their example assumes that the service times are constant and the service discipline is exhaustive, with the following parameters:

$$\lambda_1 = \lambda_2 = 5, b_1 = b_2 = 0.02, b_1^{(2)} = b_2^{(2)} = 0.0004,$$

$$E[Z_1] = 2, E[Z_2] = 3, V[Z_1] = 0, V[Z_2] = 3992.$$

(The above value of $V[Z_2]$ corrects an unimportant arithmetic error in Sarkar and Zangwill 1991.) With these parameters, Equation (2.2.8) in Sarkar and Zangwill (1991) yields $E[W_1] = 440.96$ and $E[W_2] = 363.07$. From Theorem 1, we arrive at $z^*(1) = 62.407$, for which the corresponding expected waiting times are $E[W_1] = 70.210$ and $E[W_2] = 63.970$. Thus, in this example, a twelfold increase in the sum of the mean setup times results in reductions in (both) expected waiting times to about a sixth of their original values. A similar calculation for queue 2 yields $z^*(2) = 56.589$, and this results in the corresponding expected waiting times $E[W_1] = 70.550$ and $E[W_2] = 63.665$, offering about the same order of magnitude in improvements.

An important observation regarding the theorems is that the optimal value $z^*(k)$ in both (1) and (7) are explicit nondecreasing functions of g_k , and hence of the variances of the individual setup times (cf. (2) and (8)). The theorems, therefore, can be helpful in answering the question: When does forced idle time lead to reduced waiting time? In the above example, the variance $V[Z_2]$ is several orders of magnitude greater than $E[Z_2]$, and this is the driving force behind the dramatic reduction in the expected waiting times. To observe a reduction in the expected waiting time, it is, in fact, sufficient to have a much smaller value of $V[Z_2]$ (while holding other parameters constant). For example, if $V[Z_2] = 30$, then the optimal value $z^*(1)$ is 5.41, resulting in a reduction of $E[W_1]$ from 6.11 to 6.09 (and, incidentally, a corresponding increase for $E[W_2]$ from 5.52 to 5.55). Thus, the counterintuitive anomaly can exist for more-realistic values of the setup-time variances.

In the remainder of this section, we will consider the symmetric polling model in more detail, since in this case the formulas are much simpler; and we will also restrict ourselves to exhaustive service. To illustrate the effect of the insertion of a fixed forced idle time (allo-

cated uniformly across all queues), we assume that at each queue, the setup time is a sum of two components: a variable amount Z , which can be thought of as the "original" setup time, and a constant value δ , which can be thought of as an incremental forced setup time. Thus, if we denote the overall setup time at each queue as Z_δ , then

$$Z_\delta = Z + \delta. \tag{15}$$

With these assumptions, we have $z = NE[Z]$, $E[Z_\delta] = E[Z] + \delta$, and $V[Z_\delta] = V[Z]$; and Equation (5) can be written as

$$E[W] = \frac{\rho_T}{1 - \rho_T} \frac{b^{(2)}}{2b} + \frac{1}{2} \left[\frac{V[Z]}{E[Z] + \delta} + (E[Z] + \delta) \frac{N - \rho_T}{1 - \rho_T} \right]. \tag{16}$$

Observe that the first term on the right-hand side of (16) is the expected waiting time that would prevail if there were no setup times; and the second term, which equals the incremental expected waiting time caused by the presence of the setup times (the setup-time delay), does not depend on the distribution of service times except for its mean. The value of δ that minimizes this setup-time delay is given by

$$\delta^* = \left(\sqrt{V[Z] \frac{1 - \rho_T}{N - \rho_T}} - E[Z] \right)^+, \tag{17}$$

where $(a)^+ = \max(0, a)$. Therefore, the anomaly (inserting forced idle time decreases expected waiting time) will occur if and only if

$$V[Z] > \frac{N - \rho_T}{1 - \rho_T} (E[Z])^2. \tag{18}$$

The right-hand side of (18) is minimized when $N = 1$ (in which case the polling model reduces to a vacation model where the "vacation" is the setup time). Then the condition (18) becomes $V[Z] > (E[Z])^2$; hence in the symmetric polling model with exhaustive service, the anomaly cannot occur unless the original setup time Z is "more variable" than exponential. Condition (18) clearly confirms (see (2)) that the essential factor in this anomaly is the variance-to-mean ratio of the setup times, and the variance of the service times is irrelevant.

That is, the insertion of a fixed forced idle time δ to each setup time increases the mean setup time but not its variance; and this reduction in the variance-to-mean ratio of the setup times can, intuitively, make the cycle times less variable relative to the mean, thus “smoothing” the cycle times and thereby reducing the waiting times.

As a specific numerical example, suppose there are $N = 4$ queues with constant service times with mean value 1 ($b = b^{(2)} = 1$), and server utilization $\rho_T = 0.8$; and suppose that the setup times are described by (15) with $E[Z] = 0.25$ and $V[Z] = 4$. Then (17) gives $\delta^* = 0.25$. Therefore, if there were no setup times, the expected waiting time would be

$$\frac{\rho_T}{1 - \rho_T} \frac{b^{(2)}}{2b} = 2;$$

the setup-time delay caused by the variable component Z alone is

$$\frac{1}{2} \left[\frac{V[Z]}{E[Z] + 0} + (E[Z] + 0) \frac{N - \rho_T}{1 - \rho_T} \right] = 10;$$

and the optimal setup-time delay, which is attained when $\delta = \delta^* = 0.25$, is

$$\frac{1}{2} \left[\frac{V[Z]}{E[Z] + 0.25} + (E[Z] + 0.25) \frac{N - \rho_T}{1 - \rho_T} \right] = 8.$$

Therefore, if the setup times were in fact given by Z alone, the effect of the setup times on the expected waiting times could be reduced by $(10 - 8)/10 = 20\%$, and the overall expected waiting time could be reduced by $(12 - 10)/12 = 16.7\%$, by forcing the server to be idle for $N\delta^* = 1$ unit of time (that is, 1 mean service time) each cycle, allocated arbitrarily throughout the cycle, whether or not there are customers waiting for service!

3. Proofs

Our proofs rely on a number of results derived in several previous papers. In the interest of brevity, we will draw freely from these papers, and provide specific references whenever an existing result is invoked.

We begin with the expected-value version of the decomposition theorem for $M/G/1$ queues with generalized vacations (Fuhrmann and Cooper 1985; Propositions 3 and 4, pp. 1125 and 1126):

$$E[W_k] = \frac{\lambda_k b_k^{(2)}}{2(1 - \rho_k)} + \frac{E[K_k]}{\lambda_k}, \tag{19}$$

where the first term on the right-hand side of (19) is the Pollaczek-Khintchine formula for the expected waiting time in an ordinary $M/G/1$ queue and K_k is the number of waiting customers at queue k as seen by a randomly-selected customer who arrives at queue k during a “vacation” from queue k . If we interpret the intervisit time (defined as the time interval from the instant at which the server leaves queue k until the instant at which it next completes its setup at queue k) as the vacation from queue k , then it is easily shown (Srinivasan et al. 1995; Equation (36), p. 163) that

$$E[K_k] = \frac{E[X_k(X_k - 1)] - E[T_k(T_k - 1)]}{2E[U_k]}, \tag{20}$$

where T_k is the number of waiting customers at queue k at the start of a vacation from queue k , U_k (possibly dependent on T_k) is the number of customers that arrive during the ensuing vacation, and $X_k = T_k + U_k$ (the number of waiting customers at queue k when the server next completes its setup at queue k).

Observe that the evaluation of (20) depends only on the marginal distributions of (T_k, X_k) . It follows that we can (and will), for convenience, reinterpret X_k and T_k as the numbers of waiting customers at the start and, respectively, the end of the same visit of the server to queue k . To unify the discussion, we will further assume that

$$T_k = \sum_{j=1}^{X_k} A_{kj}, \tag{21}$$

where $\{A_{kj} : j \geq 1\}$ is, independent of X_k , an i.i.d. sequence of nonnegative integer-valued random variables. In the case of exhaustive service, $A_{kj} = 0$ for all j ; and for gated service, A_{kj} is, for each j , distributed as the number of arrivals during a single service at queue k . Let A_k denote an instance of A_{kj} .

By conditioning on X_k , it follows easily from (21) that

$$E[T_k(T_k - 1)] = (E[A_k])^2 E[X_k(X_k - 1)] + E[A_k(A_k - 1)] E[X_k]; \tag{22}$$

and hence

$$\begin{aligned}
 &E[X_k(X_k - 1)] - E[T_k(T_k - 1)] \\
 &= [1 - (E[A_k])^2]E[X_k(X_k - 1)] \\
 &\quad - E[A_k(A_k - 1)]E[X_k]. \tag{23}
 \end{aligned}$$

From (21), we also have

$$E[U_k] = E[X_k] - E[T_k] = (1 - E[A_k])E[X_k]. \tag{24}$$

Substitution of (23) and (24) into (20) yields

$$\begin{aligned}
 E[K_k] &= (1 + E[A_k]) \frac{E[X_k(X_k - 1)]}{2E[X_k]} \\
 &\quad - \frac{E[A_k(A_k - 1)]}{2(1 - E[A_k])}; \tag{25}
 \end{aligned}$$

and therefore, from (19),

$$\begin{aligned}
 E[W_k] &= \frac{\lambda_k b_k^{(2)}}{2(1 - \rho_k)} + \frac{1 + E[A_k]}{\lambda_k} \frac{E[X_k(X_k - 1)]}{2E[X_k]} \\
 &\quad - \frac{1}{\lambda_k} \frac{E[A_k(A_k - 1)]}{2(1 - E[A_k])}. \tag{26}
 \end{aligned}$$

It is well known (Takagi 1986; Equation (4.10a), p. 73) that in the exhaustive case, $E[X_k] = \lambda_k E[C_k](1 - \rho_k)$, where $E[C_k]$ denotes the expected cycle time with respect to queue k , given by $E[C_k] = z/(1 - \rho_T)$, independent of k ; and for gated service, $E[X_k] = \lambda_k E[C_k]$. Thus, it remains for us to calculate $E[X_k(X_k - 1)]$.

The term $E[X_k(X_k - 1)]$ can be calculated using the descendant-sets method discussed in Konheim et al. (1994). The idea behind this method is to decompose the variable X_k at a polling epoch at queue k as an infinite sum of independent populations of "descendants" generated by "originator customers" who arrived during all of the past setup times. For exhaustive service, this method leads to

$$\begin{aligned}
 &E[X_k(X_k - 1)] \\
 &= (E[X_k])^2 + \lambda_k^2 V[Z_k] \\
 &\quad + \lambda_k^2 \sum_{i=1}^N \frac{\Gamma_i^{(k)}}{\rho_i^2} \left\{ \lambda_i b_i^{(2)} \frac{z}{1 - \rho_T} + V[Z_i] \right\} \tag{27}
 \end{aligned}$$

(Srinivasan et al. 1995; Equation (52), p. 167), where the $\Gamma_i^{(k)}$ s are given by (3). Substituting (27) into (26) (with $A_k = 0$) and rearranging terms yields

$$\begin{aligned}
 &E[W_k] \\
 &= \frac{1}{2} \left[z \frac{1 - \rho_k}{1 - \rho_T} + g_k \left(z \frac{1 - \rho_k}{1 - \rho_T} \right)^{-1} + \frac{h_k}{1 - \rho_k} \right], \tag{28}
 \end{aligned}$$

where g_k is given by (2) and

$$h_k = \lambda_k b_k^{(2)} + \sum_{i=1}^N \frac{\Gamma_i^{(k)}}{\rho_i^2} \lambda_i b_i^{(2)}. \tag{29}$$

(See Equations (2.2.5) and (2.2.7) in Sarkar and Zangwill 1991 for a similar expression, in terms of the solutions of a system of N linear equations.) Note that Equation (28) is convex in z . Differentiating (28) with respect to z and equating the resulting expression to zero yields (1).

Formula (26) appears to be new and useful. It gives a unified representation (in terms of X_k) of the expected waiting time for *any* service discipline satisfying assumption (21) (see Fuhrmann 1992, §5, for a related discussion of this assumption). In addition to the exhaustive and gated disciplines, other examples include binomially gated (Levy 1989) (every waiting customer at a polling epoch has a fixed probability of receiving service during a server visit) and round-robin with exponentially distributed service times and with a fixed service quantum (Fuhrmann 1981). Our proofs in §3 can be easily adapted to derive optimal-setup-times results similar to Theorems 1 and 2 for other members of the class of service disciplines defined by (21).

Formula (26) is also helpful in clarifying the "origin" of Sarkar and Zangwill's anomaly. Notice that the only term in (26) that depends on the parameter z appears in the second term as

$$\frac{E[X_k(X_k - 1)]}{2E[X_k]}, \tag{30}$$

which, interestingly, assumes the form of the expected forward-recurrence time of a discrete-time renewal process with its interevent times distributed as X_k . It is intuitive (and it can be easily shown by an induction over cycles) that X_k , being the number of waiting customers at queue k when the server next completes its setup at queue k , is a nondecreasing function of the individual setup times. It follows that both the numerator and the denominator in (30) are nondecreasing functions of z . Hence, an increase in z does not necessarily guarantee

an increase in $E[W_k]$ (unless X_k is a constant); and, since X_k is closely tied to the cycle time C_k , this observation provides a precise formalization of Sarkar and Zangwill's (1991) "explanation."

To prove (5), we will rely on a recent *decomposition theorem* for polling models (Cooper et al. 1996; Equation (1)):

$$E[W_k] = E[W_k^0] + \frac{z}{2} \frac{1 - \rho_k}{1 - \rho_T}, \tag{31}$$

where W_k^0 is the waiting time in the "corresponding" exhaustive service model in which the setup times are zero and the service times have second moment

$$x_k^{(2)} = b_k^{(2)} + V[Z_k] \left(\frac{\lambda_k z}{1 - \rho_T} \right)^{-1}. \tag{32}$$

The $M/G/1$ conservation law (Kleinrock 1976; Equation (3.16), p. 114) provides the following expression for $E[W_k^0]$:

$$\sum_{k=1}^N \rho_k E[W_k^0] = \frac{\rho_T}{1 - \rho_T} \sum_{k=1}^N \rho_k \frac{x_k^{(2)}}{2b_k};$$

and since $E[W_k^0]$ depends on the service-time distributions only through their first two moments (e.g., Srinivasan et al. 1995; Equation (43), p. 165)) and hence is the same for all k in a symmetric model, we suppress the index k and obtain

$$E[W^0] = \frac{\rho_T}{1 - \rho_T} \frac{x^{(2)}}{2b}. \tag{33}$$

Substitution of (32) and (33) into (31) yields (5). (Formula (5) was originally derived by Hashida 1972, and again by Fuhrmann 1985 via a simple vacation-decomposition argument; our proof requires only that the first two moments of the service times and the variances of the setup times be symmetric.)

The minimizing value of z^* is easily obtained by differentiating with respect to z in (5); then (6) gives the positive value of z that satisfies $dE[W]/dz = 0$.

We now turn to the gated case. Clearly, we have $E[A_k] = \rho_k$; and it is easily shown that

$$\frac{1}{\lambda_k} \frac{E[A_k(A_k - 1)]}{2(1 - E[A_k])} = \frac{\lambda_k b_k^{(2)}}{2(1 - \rho_k)}.$$

Hence, from (26),

$$E[W_k] = \frac{1 + \rho_k}{\lambda_k} \frac{E[X_k(X_k - 1)]}{2E[X_k]} \tag{34}$$

(in agreement with Takagi 1986; Equation (5.47b), p. 111). Using the descendant-sets method, it can be shown (Konheim et al. 1994; Equation (3.13), p. 1249) that

$$E[X_k(X_k - 1)] = (E[X_k])^2 + \lambda_k^2 \sum_{i=1}^N \frac{\Gamma_i^{(k)}}{\rho_i^2} \left\{ \lambda_i b_i^{(2)} \frac{z}{1 - \rho_T} + V[Z_{i+1}] \right\}, \tag{35}$$

where the $\Gamma_i^{(k)}$ s are given by (9). It follows from (34) and (35) that

$$E[W_k] = \frac{1 + \rho_k}{2} \left[\frac{z}{1 - \rho_T} + g_k \left(\frac{z}{1 - \rho_T} \right)^{-1} + h_k \right], \tag{36}$$

where g_k is given by (8) and

$$h_k = \sum_{i=1}^N \frac{\Gamma_i^{(k)}}{\rho_i^2} \lambda_i b_i^{(2)}. \tag{37}$$

Again, (36) is convex in z , and this immediately leads to (7).

The proofs of (11) and (12) are similar to those of (5) and (6), so we omit the details.¹

¹ Research supported in part by the National Science Foundation under grants DMI-9500216, 9500040, 9500471. Research also supported in part by a Summer Research Grant from the School of Management, The University of Texas at Dallas.

References

Boxma, O. J., W. P. Groenendijk 1987. Pseudo-Conservation Laws in Cyclic-Service Systems. *J. of Applied Probability* **24** 4 949-964.
 Cooper, R. B., S.-C. Niu, M. M. Srinivasan 1996. A Decomposition Theorem for Polling Models: The Switchover Times are Effectively Additive. *Oper. Res.* **44** 4 629-633.
 Duenyas, I. 1994. The Limitations of Sub-Optimal Policies. *Interfaces* **24** 5 77-84.
 —, M. P. Van Oyen 1994. Can Setup Reduction Result in Worse Performance? Working paper.
 —, — 1996. Heuristic Scheduling of Parallel Heterogeneous Queues with Set-ups. *Management Sci.* **42** 814-829.
 Federgruen, A., Z. Katalan 1996a. The Impact of Setup Times on the Performance of Multi-Class Service and Production Systems. *Oper. Res.* **44** 6 989-1001.
 —, — 1996b. The Stochastic Economic Lot Scheduling Problem: Cyclical Base-Stock Policies with Idle Times. *Management Sci.* **42** 6 783-796.

- Ferguson, M. J., Y. J. Aminetzah 1985. Exact Results for Nonsymmetric Token Ring Systems. *IEEE Trans. on Communications* COM-33 223–231.
- Fuhrmann, S. W. 1981. Performance Analysis of a Class of Cyclic Schedules. Technical Report, AT&T Bell Laboratories.
- 1985. Symmetric Queues Served in Cyclic Order. *Oper. Res. Letters* 4 3 139–144.
- 1992. A Decomposition Result for a Class of Polling Models. *Queueing Systems* 11 109–120.
- , R. B. Cooper 1985. Stochastic Decompositions in the $M/G/1$ Queue with Generalized Vacations. *Oper. Res.* 33 5 1117–1129.
- Gerchak, Y., Z. Zhang 1994. The Cheaper/Faster-Yet-More-Expensive Phenomenon: Are Zangwill's 'Paradoxes' Indeed Paradoxical? *Interfaces* 24 5 84–87.
- Gupta, D., M. M. Srinivasan 1996a. The Variance Paradox and Its Implications for Japanese Production Theory. *Interfaces* 26 4 69–77.
- , — 1996b. Polling Systems with State-Dependent Setup Times. *Queueing Systems* 22 403–423.
- Hashida, O. 1972. Analysis of Multiqueue. *Rev. Electrical Communication Laboratories, Nippon Telephone and Telegraph Public Corporation*. 20 189–199.
- Kleinrock, L. 1976. *Queueing Systems, Volume 2: Computer Applications*. John Wiley & Sons, New York.
- Konheim, A. G., H. Levy, M. M. Srinivasan 1994. Descendant Set: An Efficient Approach for the Analysis of Polling Systems. *IEEE Trans. on Communications* 42 1245–1253.
- Kuehn, P. J. 1979. Multiqueue Systems with Nonexhaustive Cyclic Service. *Bell System Technical J.* 58 671–698.
- Levy, H. 1989. Analysis of Cyclic-Polling Systems with Binomial-Gated Service. *Performance of Distributed and Parallel Systems*, Elsevier Science Publishers B. V., North-Holland, Amsterdam, 127–139.
- McIntyre, B. 1994. A Comment on Zangwill's 'The Limits of Japanese Production Theory.' *Interfaces* 24 5 87–89.
- Olsen, T. 1996a. Asymptotics for Polling Models with Increasing Setups. Working Paper.
- 1996b. Approximations for the Waiting Time Distribution in Polling Models With and Without State-Dependent Setups. Working Paper.
- Righter, R., J. G. Shanthikumar 1995. Multi-Class Production Systems with Setup Times. Working Paper.
- Samaddar, S. and T. Kaul 1995. Effects of Setup and Processing Time Reductions on WIP in the JIT Production Systems. *Management Sci.* 41 7 1263–1265.
- Sarkar, D., W. I. Zangwill 1989. Expected Waiting Time for Nonsymmetric Cyclic Queueing Systems—Exact Results and Applications. *Management Sci.* 35 12 1463–1474.
- , — April 1991. Variance Effects in Cyclic Production Systems. *Management Sci.* 37 4 443–453.
- Srinivasan, M. M., D. Gupta 1996. When Should a Roving Server Be Patient? *Management Sci.* 42 3 437–451.
- , S.-C. Niu, R. B. Cooper 1995. Relating Polling Models with Zero and Nonzero Switchover Times. *Queueing Systems* 19 149–168.
- Takagi, H. 1986. *Analysis of Polling Systems*. MIT Press, Cambridge, MA.
- Van Oyen, M. P. 1997. Monotonicity of Optimal Performance Measures for Polling Systems. *Probability in the Engineering and Informational Sciences* 11 219–228.
- Watson, K. S. 1984. Performance Evaluation of Cyclic Service Strategies—A Survey. *Performance '84*, Elsevier Publishers, New York.
- Zangwill, W. I. 1992. The Limits of Japanese Production Theory. *Interfaces* 22 5 14–25.
- , D. Sarkar 1994. Response to Comments on Our Work by Duenyas, by Gerchak and Zhang, and by McIntyre. *Interfaces* 24 5 90–94.

Accepted by Linda V. Green; received June 19, 1996. This paper has been with the authors 1 month for 1 revision.