Probability in the Engineering and Informational Sciences, **13**, 1999, 489–496. Printed in the U.S.A.

MORE ON USING FORCED IDLE TIME TO IMPROVE PERFORMANCE IN POLLING MODELS

Erol A. Peköz
School of Management
Boston University
Boston, Massachusetts 02215

A recent interesting paper (Cooper, Niu, and Srinivasan [2]) shows how for some cyclic production systems reducing setup times can surprisingly increase work in process. There the authors show how in these situations the introduction of forced idle time can be used to optimize system performance. Here we show that introducing forced idle time at different points during the production cycle can further improve performance in these situations and also in situations where the suggestions from [2] yield no improvement.

1. INTRODUCTION AND SUMMARY

In a provocative *Interfaces* article, Zangwill [9] claims to have found examples which "expose a flaw" in Japanese production theory (which advocates reducing overhead), since results from Sarkar and Zangwill [4] show that for some cyclic production systems reducing setup times can actually increase work in process. This led to some heated criticisms resulting in several more *Interfaces* articles and a rejoinder (Zangwill and Sarkar [10]), which all appeared together prefaced by a statement from the editor-in-chief saying that Zangwill's article [9] "alone drew more response than all other articles combined" while he was editor.

Inspired by this debate was the article of Cooper, Niu, and Srinivasan [2] showing how the introduction of forced idle time can be used to optimize system performance. Continuing this work we show that introducing forced idle time at different points during the production cycle can yield further improvements, and may yield improvements in situations where the suggestions from [2] yield no improvements.

0269-9648/99 $12.50

The basic cyclic polling model we consider consists of a number of queues and a single server who visits them in a cyclic order in order to serve the customers. There is a large amount of literature on polling models, see Tagaki [5,6], Konheim and Levy [3], and the recent Van Der Mei and Levy [7] for some background and references.

In Section 2 we introduce the model, describe the strategy suggested in [2], introduce two new strategies which perform better, and state some theorems about the performance of the strategies. In Section 3, we give proofs of the results stated in Section 2.

2. MAIN RESULTS

Here we consider n $M/G/1$ queues and a single server who visits them in a predetermined fixed cyclic order. The switchover times of the server between queues are independent, identically distributed random variables denoted generically as Z. The service times for customers are independent and identically distributed random variables with first moment b and second moment $b^{(2)}$, and for each queue there is an independent Poisson process of arrivals with parameter λ. Customers leave the system immediately after being served. Below we let $\rho = n\lambda b$ and consider three service strategies.

Strategy 1. Under this strategy the server upon arrival to a queue begins serving customers and works until the queue is empty. Then the server remains idle for a deterministic δ time units, and at this point the server switches over to the next queue and repeats this process.

Strategy 2. Under this strategy, upon arrival to a queue the server serves customers until the queue is empty, then remains idle for a deterministic δ time units. At this point, the server checks the queue and, if there are any customers waiting, begins serving them and works until the queue is empty again. Then the server switches over to the next queue and repeats this process.

Strategy 3. While at a queue under this strategy, the server works whenever there are customers in the queue and waits idle for new arrivals whenever the queue is empty. Once the server has spent a total of δ time units waiting idle at a queue, the server switches over to the next queue and repeats this process.

Below we use W_i to denote a random variable with the stationary waiting time distribution (until beginning service) for a generic customer under service strategy i, and we assume the systems are stable. The first strategy is considered in [2, p. 1082], where the result

$$E[W_1] = \frac{\rho}{1-\rho}\,\frac{b^{(2)}}{2b} + \frac{1}{2}\left[\frac{\mathrm{Var}(Z)}{E[Z]+\delta} + (E[Z]+\delta)\,\frac{n-\rho}{1-\rho}\right] \tag{1}$$

is given for the stationary waiting time. This is minimized using a value of δ given by

$$\delta_1^* = \left(\sqrt{\operatorname{Var}(Z)\,\frac{1-\rho}{n-\rho}} - E[Z] \right)^+ . \tag{2}$$

The surprising fact from this is that δ_1^* can be positive in some situations, meaning that adding idle time can improve performance. The main result we give here is that Strategy 2 is better than Strategy 1, and that Strategy 3 is better than Strategy 2. This is summarized by the following theorems.

THEOREM 2.1: *The stationary waiting time for customers under the second service strategy is given by*

$$E[W_2] = E[W_1] - \frac{\delta}{E[Z]+\delta}\left[E[Z] + \delta\,\frac{n-1}{n-\rho}\right], \tag{3}$$

which is minimized using a value of δ given by

$$\delta_2^* = \left(\sqrt{\frac{\operatorname{Var}(Z) + 2(E[Z])^2\,\dfrac{1-\rho}{n-\rho}}{\dfrac{n-\rho}{1-\rho} - 2\,\dfrac{n-1}{n-\rho}}} - E[Z] \right)^+ . \tag{4}$$

THEOREM 2.2: *The stationary waiting time for customers under the third service strategy is given by*

$$E[W_3] = E[W_1] - \frac{\delta}{E[Z]+\delta}\left[E[Z] + \frac{\delta}{2}\left(1 + \frac{n-1}{n-\rho}\right)\right] \tag{5}$$

$$= E[W_2] - \left(\frac{\delta^2/2}{E[Z]+\delta}\right)\left(\frac{1-\rho}{n-\rho}\right), \tag{6}$$

which is minimized using a value of δ given by

$$\delta_3^* = \left(\sqrt{\frac{\operatorname{Var}(Z) + (E[Z])^2\,\dfrac{1-\rho}{n-\rho}}{\dfrac{n-\rho}{1-\rho} - \dfrac{n-1}{n-\rho} - 1}} - E[Z] \right)^+ . \tag{7}$$

Note 1: From these theorems when $\delta, E[Z] > 0$

$$E[W_3] < E[W_2] < E[W_1],$$

so performance under the third service strategy is strictly better than performance under the second, which is strictly better than performance under the first. Some algebra also shows that, when $\delta_1^* > 0$, we have $\delta_2^* > \delta_3^* > \delta_1^*$. Interestingly, this shows that the second strategy advocates the most idle time, and that the worst strategy, the first strategy, advocates the least idle time.

Note 2: Also as mentioned in [2, p. 1082], δ_1^* is positive if and only if the squared coefficient of variation of Z, $scv = \mathrm{Var}(Z)/(E[Z])^2$, is greater than $(n-\rho)/(1-\rho)$. Some more algebra also shows that both δ_3^* and δ_2^* are positive if and only if scv is greater than $(n+\rho-2)/(1-\rho)$. Since $(n-\rho)/(1-\rho) > (n+\rho-2)/(1-\rho)$, this shows that the second and third strategies may yield benefits in situations where the first strategy yields no benefits. In particular, when $n = 2$ the second and third strategies yield benefits for any distribution of Z if traffic is sufficiently light, whereas the first strategy would only yield benefits if the coefficient of variation of Z is larger than 2 and traffic is sufficiently light.

3. PROOFS OF RESULTS

We first need some definitions and a lemma.

Let U_i be a random variable distributed as the amount of work at an arbitrary epoch in a stationary cyclic-service system under strategy i.

Let V be a random variable distributed as the amount of work at an arbitrary epoch in a stationary $M/G/1$ queuing system where service times have first moment b and second moment $b^{(2)}$, and the Poisson arrival process has parameter $n\lambda$.

Let Y_i be a random variable whose distribution is the same as the conditional distribution of the amount of work at an arbitrary epoch in a stationary cyclic-service system under strategy i given that the server is not serving at that epoch.

Lemma 3.1: *With the above definitions*

$$U_i \overset{d}{=} V * Y_i, \qquad i = 1,2, \tag{8}$$

where $\overset{d}{=}$ *denotes equality in distribution and* $*$ *denotes convolution.*

Proof of Lemma 3.1: The case for $i = 1$ appears in Boxma and Groenendijk [1, Thm. 1]. The locations of the server idle times are not used anywhere in the argument given, so the same argument also applies to the second service strategy described above, thus establishing the lemma.

The details appear in [1, Thm. 1], but the essential argument can be summarized as follows. Consider the cyclic-service system under either strategy and on the same probability space define a modified LIFO single-queue $M/G/1$ having the same arrival process of work. By this we mean that when there is an arrival to any of the n queues in the cyclic-service system with a given service requirement, there also will be an arrival to the single-queue system with the same service requirement. We do, however, impose the following modification on the single-queue system: whenever the server in the cyclic-service system is not serving customers, the server in the single-queue system is forced (interrupting any service in progress) to be idle as well. In the single-queue system, customers which arrive during the forced idle period are served in LIFO fashion (after the idle period) until the server eventually gets to the customer whose service was interrupted, and, at this point, the server continues with the remaining portion of the unfinished service. Note that since the systems have the same arrival streams and work and are idle at the same times, the amount of work in both systems is equal at all times. It therefore suffices to study just the single-queue system.

At a given time while the server (in the single-queue system) is serving, label the customer who most recently arrived while the server was not serving as the current “primary” customer. When the server is not serving, no customer is labeled as the “primary” customer. Note that a chosen customer can spend some time as the “primary” customer; this can be interrupted while another customer is the “primary” customer, and then the chosen customer can resume as the “primary” customer.

For a given customer who spends time as a “primary” customer, consider the process of the work in the system which arrives after him. We next argue that this process, when viewed only while this given customer is the “primary” customer, evolves like the work process for the busy period of an $M/G/1$ LIFO queue. Due to the LIFO service policy whenever a given customer’s “primary” status is interrupted, the work in the system after the interruption will be the same as it was before the interruption. This is because whenever a service is interrupted, any new customers are served in LIFO fashion until the server returns to resume serving the customer whose service was interrupted. This means that for a randomly chosen time, the amount of work that has arrived since the current “primary” customer is distributed as the amount of work at a random time during the busy period of an $M/G/1$ LIFO queue.

Thus, at an arbitrarily chosen time while the server is serving, the amount of work in the system equals the amount of work present when the current “primary” customer arrived (which has the distribution of Y_i) plus any remaining work, which has the distribution V given V is positive, and these are independent. At an arbitrarily chosen time when the server is not serving, the amount of work in the system simply has distribution Y_i. Unconditioning then gives the lemma. ■

Proof of Theorem 2.1: Note that under either service strategy the probability at an arbitrary epoch that the server is serving is ρ. Letting L_i be the mean stationary number of customers waiting under service strategy i in a single one of the queues, we have, for $i = 1,2$,

$$E[U_i] = nbL_i + \rho \frac{b^{(2)}}{2b} \tag{9}$$

$$= \rho E[W_i] + \rho \frac{b^{(2)}}{2b}. \tag{10}$$

The second term in the first line above follows because, with probability ρ, a customer is being served, and the mean residual service time is $b^{(2)}/(2b)$. The second line follows using "Little's Result," $L_i = \lambda E[W_i]$. Using Eq. (10) and the lemma above in the form $E[U_i] = E[V] + E[Y_i]$, we obtain

$$E[W_2] = E[W_1] - (E[Y_1] - E[Y_2])/\rho. \tag{11}$$

Next, instead of the first strategy above we consider in its place the following strategy, which is identical with respect to customer waiting times: Upon arrival to a queue, the server first remains idle for a deterministic δ time units and then begins serving customers and works until the queue is empty. At this point the server switches over to the next queue and repeats this process.

We now decompose $E[Y_i]$ by writing it as the sum

$$E[Y_i] = E[C_i] + E[D_i] \tag{12}$$

where, for a random epoch when the server is not serving (under service strategy i), $E[C_i]$ is the mean amount of work in the system when the server most recently began switching queues, and $E[D_i]$ is the mean additional work in the system at this epoch.

Next, we claim that for a given δ the distribution of the time spent at a given queue is the same under either service strategy. This is most easily seen with the following argument. Suppose at time t_1 the server, under the modified first service strategy above, arrives at a queue to find k customers waiting. Let t_2 be the time after the initial idle period when there are once again k customers in the queue, and let t_3 be the time when the server begins switching to the next queue. Now consider a time t_1' in an independent system under the second service strategy above, when the server first arrives at a queue also to find k customers waiting. Let t_2' be the time when the queue next becomes empty, and t_3' be the time when the server begins switching to the next queue. The important things to note are that $t_2 - t_1$ has the same distribution as $t_3' - t_2'$, $t_3 - t_2$ has the same distribution as $t_2' - t_1'$, and all four differences are independent. Thus, $t_3 - t_1$ has the same distribution as $t_3' - t_1'$ under the same initial conditions. Thus, the distribution of time spent at a given queue is the same under either service strategy, and only depends on the number of customers present when the server arrives. This also means that under either strategy the systems have the same probabilistic behavior when viewed at the epochs when the server begins switching to the next queue, thus,

$$E[C_1] = E[C_2]. \tag{13}$$

Next, note that for the first service strategy

$$E[D_1] = \rho \frac{E[Z]}{E[Z] + \delta} (E[Z^2]/2E[Z]) + \rho \frac{\delta}{E[Z] + \delta} (E[Z] + \delta/2). \tag{14}$$

The first term follows because with probability $E[Z]/(E[Z] + \delta)$ the server is switching, the mean amount of time this has been in progress is $E[Z^2]/2E[Z]$, and the mean amount of work arriving to the system per unit time is ρ. The second term follows because with probability $\delta/(E[Z] + \delta)$ the server is waiting idle, and the mean amount of time since switching began is $E[Z] + \delta/2$.

Next, note that for the second service strategy

$$E[D_2] = \rho \, \frac{E[Z]}{E[Z]+\delta} \, (E[Z^2]/2E[Z]) \tag{15}$$

$$+ \, \frac{\delta}{E[Z]+\delta} \, (\rho\delta/2 - (\rho - \lambda b)(\delta + \delta\lambda T)), \tag{16}$$

where $T = b/(1-\lambda b)$ (see Wolff [8, p. 389]) is the expected length of a busy period at one of the queues. To justify term (15) and term (16), consider a randomly chosen epoch $\mathcal{E}$ where the server is not serving. As for Eq. (14), the first term (15) follows because with probability $E[Z]/(E[Z]+\delta)$ the server is switching at $\mathcal{E}$, and the mean amount of time this has been in progress is $E[Z^2]/2E[Z]$.

To understand the second term (16), let $\mathcal{I}_1$ be the time interval between when the server most recently began switching queues and $\mathcal{E}$, and let $\mathcal{I}_2$ be the time interval between $\mathcal{E}$ and the time the server next begins switching queues. With probability $\delta/(E[Z]+\delta)$ the server is waiting idle at $\mathcal{E}$. Since the system is stable, the mean net work (amount of arriving work minus amount of completed work) in the system during $\mathcal{I}_1$ plus the mean net work in the system during $\mathcal{I}_2$ must equal zero. Thus, to compute the mean net work during $\mathcal{I}_1$, we can subtract the amount of work that is expected to arrive during $\mathcal{I}_2$ from the amount of work that is expected to be completed during $\mathcal{I}_2$. Note that $\mathcal{I}_2$ has expected length $\delta/2 + \delta\lambda T$, since the expected remaining time until the idle period ends is $\delta/2$, the mean number of arrivals during the idle period is $\delta\lambda$, and each takes expected time T to handle.

We now consider separately the queue where the server is currently waiting idle and the remaining queues. For the current queue all work that arrives during $\mathcal{I}_2$ will be finished, plus any work that arrived during the portion of the idle period prior to $\mathcal{E}$, which has mean $\lambda b\delta/2$. For the remaining queues, the mean amount of work that arrives is $\rho - \lambda b$ multiplied by the expected length of the time interval. Thus, the net amount of work that leaves the system during $\mathcal{I}_2$ is

$$\lambda b\delta/2 - (\rho - \lambda b)(\delta/2 + \delta\lambda T) = \rho\delta/2 - (\rho - \lambda b)(\delta + \delta\lambda T),$$

explaining term (16).

Combining the equations we get

$$\begin{aligned} E[W_2] &= E[W_1] - (E[D_1] - E[D_2])/\rho \\ &= E[W_1] - \frac{\delta}{E[Z]+\delta}\left(E[Z] + \delta/2 + \frac{\rho - \lambda b}{\rho}\,(\lambda\delta T + \delta) - \delta/2\right) \\ &= E[W_1] - \frac{\delta}{E[Z]+\delta}\left[E[Z] + \delta\,\frac{n-1}{n-\rho}\right]. \end{aligned}$$

The theorem then follows by differentiating in order to find the minimum. ■

Proof of Theorem 2.2: For a fixed value of $k > 1$ we consider the following fourth strategy.

Strategy 4. The server upon arrival to a queue serves customers until the queue is empty, then remains idle for a deterministic δ/k time units. At this point the server checks the queue and, if there are any customers waiting, begins serving them and works until the queue is empty again. At this time the server remains idle again for a deterministic δ/k time units, and then begins again serving any new customers. The server repeats this a total of k times so that a total of δ time units are spent idle at this queue. Then the server switches over to the next queue and repeats this process.

As k approaches infinity, Strategy 4 approaches Strategy 3 and

$$\lim_{k\to\infty} E[W_4] = E[W_3].$$

The same proof of Lemma 3.1 applies to Strategy 4, and since $E[C_4] = E[C_2]$ we can use the same reasoning as in the proof of Theorem 2.1 to get

$$E[W_4] = E[W_1] - (E[D_1] - E[D_4])/\rho. \qquad \textbf{(17)}$$

We next obtain

$$E[D_4] = \rho \frac{E[Z]}{E[Z]+\delta}(E[Z^2]/2E[Z]) \qquad \textbf{(18)}$$

$$+ \frac{\delta}{E[Z]+\delta}\left(\rho\frac{\delta/2}{k} - \left(\frac{k+1}{2}\right)(\rho - \lambda b)\left(\frac{\delta}{k} + \frac{\delta}{k}\lambda T\right)\right), \qquad \textbf{(19)}$$

which follows using the same justification as the ones given above for (15) and (16) except now each idle period has length $\delta/2$ and a randomly chosen idle period is equally likely to be any of the k idle periods the server experiences at each queue, so the expected number of idle periods yet to be completed at the queue before switching is $(k+1)/2$. Using Eq. (17) and letting k approach infinity then gives Eq. (5) and some algebra gives (6). Minimizing by taking the derivative gives Eq. (7). ■

References

1. Boxma, O.J. & Groenendijk, W.P. (1987). Pseudo-conservation laws in cyclic-service systems. *Journal of Applied Probability* 24: 949–964.
2. Cooper, R.B., Niu, S., & Srinivasan, M.M. (1998). When does forced idle time improve performance in polling models? *Management Science* 44(8): 1079–1086.
3. Konheim, A.G. & Levy, H. (1994). Descendent set: An efficient approach for the analysis of polling systems. *IEEE Transactions on Communications* 42(2/3/4): 1245–1252.
4. Sarkar, D. & Zangwill, W.I. (1991). Variance effects in cyclic production systems. *Management Science* 37(4): 443–453.
5. Tagaki, H. (1986). *Analysis of polling systems*. Cambridge: MIT Press.
6. Tagaki, H. (1994). Queuing analysis of polling models: Progress in 1990–1993. In J.H. Dshalalow (ed.), *Frontiers in queuing models, methods and problems*. Boca Raton: CRC Press.
7. Van Der Mei, R.D. & Levy, H. (1998). Expected delay analysis of polling systems in heavy traffic. *Advances in Applied Probability* 30: 586–602.
8. Wolff, R. (1989). *Stochastic modeling and the theory of queues*. Englewood Cliffs: Prentice-Hall.
9. Zangwill, W.I. (1992). The limits on Japanese production theory. *Interfaces* 22(5): 14–25.
10. Zangwill, W.I. & Sarkar, D. (1994). Response to comments on our work by Duenyas, by Gerchak and Zhang, and by McIntyre. *Interfaces* 24(5): 90–94.