Probability in the Engineering and Informational Sciences, **15**, 2001, 35–55. Printed in the U.S.A.

LIMIT THEOREMS FOR POLLING MODELS WITH INCREASING SETUPS

TAVA LENNON OLSEN
Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, Michigan 48109-2117
E-mail: tlennon@engin.umich.edu

Multiclass single-server systems with significant setup times (polling models) are common in industry. This article considers asymptotics for polling models with increasing setup times. Two types of polling model are considered, namely (a) polling models with polling tables, exhaustive service, and deterministic setups, and (b) cyclic exhaustive service polling models with general setups under heavy traffic. It is shown that as the mean setup time increases to infinity, the scaled intervisit time for each queue (time between service of that queue) converges in probability to a constant. This, in turn, is shown to imply that scaled steady-state waiting time converges in distribution to either a uniform distribution or a simple discrete random variable multiplied by a uniform random variable as setups tend to infinity. These results lead to considerable insight into the behavior of systems with setups, and conclusions are drawn with respect to previous studies.

1. INTRODUCTION

In the manufacturing industry, there are many environments in which a variety of products are produced on a single machine that requires a significant setup or changeover time between different products (see Olsen [19] for a description of three such environments). For example, stamping machines require die changes. As the dies are extremely heavy and require an exact adjustment, the time to change dies in an efficient plant will be about half an hour. We have seen inefficient plants where the die changes are approximately 3 h. On the other hand, stamping times per part are on the order of 10–20 parts per minute. Therefore, in these environments, setup times

© 2001 Cambridge University Press 0269-9648/01 $12.50

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

can be seen to be extremely large when compared to processing times. This article presents asymptotic results when mean setup times are increased to infinity and processing times are held constant. Just as traditional heavy-traffic results (see, e.g., Harrison [15]), which are proved as server utilization converges to 100%, can lead to insights for systems in moderate traffic, this article shows that results when the mean setup time converges to infinity can lead to insights for systems with moderate setups.

Single-server systems where a setup time is incurred whenever the server changes product class are known as *polling models.* Polling models have been heavily studied in the queueing and telecommunications literature. The interested reader is referred to Takagi [22–24] for surveys of this literature and Levy and Sidi [17] for a survey of polling model applications. We are envisioning manufacturing as the main application of interest for this article because setup times tend to be small in telecommunication applications.

For any polling system, one must decide the order in which queues are visited and how long to serve a queue once the server is set up. This article considers systems where the order of service is determined by a *polling table.* The polling table specifies the fixed order in which the queues will be visited, with high-priority queues appearing multiple times in the table. When the server completes a class, it advances to the next class in the table and begins a setup. A common special case of this system (see [23]), known as *cyclic* service, is where each customer class appears exactly once in the table. The scheduling regime considered in this article for determining when to stop service of a class is a very commonly studied policy (see [23]), known as *exhaustive* service. Under exhaustive service, the server serves all waiting customers in a class (including those that arrive during the current service period) before moving on to the next class. Service within a class is in the same order in which the customers arrived (i.e., FIFO service). We consider systems where a setup is always incurred regardless of the state of the queue (i.e., empty or not) and idleness is not permitted. In practice, however, this last assumption is irrelevant because as setup times are increased to infinity, it will indeed be the case that queues will not be empty when polled and idleness will not be required.

Although we only discuss a specific policy for serving these systems, Section 4.3 discusses extensions to more general policies. In a manufacturing environment, polling table policies are more likely to be found in a make-to-stock environment than in a make-to-order environment. In a make-to-stock environment, Federgruen and Katalan [12] have shown that an exhaustive polling table policy is equivalent to a base-stock policy where the product classes are replenished in a fixed order. They provide optimal base-stock levels under this type of policy.

This article provides limit theorems for two types of models: (a) polling models with polling tables, exhaustive service, and deterministic setups; and (b) cyclic exhaustive service polling models with general setups under heavy traffic. It is shown that as the mean setup time increases to infinity, the scaled intervisit time for each queue (time between service of that queue) converges in probability to a constant. This, in turn, is shown to imply that the scaled steady-state waiting time converges

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

in distribution to either a uniform distribution or a simple discrete random variable multiplied by a uniform random variable as setups tend to infinity.

Note that this type of limit theorem is fundamentally different from that for systems without setup times where the traditional limiting distribution for the steady-state waiting time is exponential (e.g., see Harrison [15]). Not only does this limit provide insight into system behavior, but it also provides a very simple uniform-type approximation for the waiting time distribution for systems with moderate setups. A number of authors have provided approximations for the waiting time distribution (as opposed to just the mean) in polling systems. Coffman, Puhalskii, and Reiman [6] have a heavy-traffic approximation (described in Sect. 2) for cyclic polling systems, which was refined in Olsen [18] and found to be very accurate. Federgruen and Katalan [10] and Choudhury and Whitt [5] have provided numerical algorithms for approximating the waiting time distribution. The uniform-type approximation given here will not compete with these approximations for accuracy, but it does provide a simple (and easily implementable) "back-of-the-envelope" approximation. However, we believe that the approximation's more important application is to deepen the understanding of the behavior of systems with setups.

Before leaving the literature review on performance analysis, there is some recent literature with significant potential that should also be mentioned. Van der Mei and Levy [27] provide heavy-traffic approximations for the mean waiting time in cyclic polling models with general "branching-type" service disciplines (of which exhaustive service is a special case). These results are asymptotically exact as $\rho \uparrow 1$ and found to provide good approximations for reasonable utilization levels. Polling systems with polling tables are considered in van der Mei [26]. Van der Mei [25] provides similar heavy-traffic approximations for cyclic polling models with exhaustive and gated service (see [23]) for higher-order moments of waiting time (as opposed to just the mean). However, in [25], setup times are taken to be zero. It is quite likely that the authors will be able to extend their work to systems with setups and to use the derived moments to obtain full distributional approximations for the waiting time. The work also assumes Poisson arrivals but may be extendable to more general distributions.

This article is organized as follows. Section 2 further defines the model, introduces notation, and presents some preliminary results. Section 3 gives the main results of the article. Section 4 describes the insights that can be gained from the results of Section 3. Finally, Section 5 concludes the article.

2. MODEL DESCRIPTION

Each polling model consists of n customer types, or queues. Customers are assumed to arrive to each queue i, $1 \leq i \leq n$, according to independent Poisson processes with rate λ_i. Let $\lambda = \sum_{i=1}^{n} \lambda_i$ be the total arrival rate to the system. The service times for customers at queue i, $1 \leq i \leq n$, are independent and identically distributed (i.i.d.) according to the random variable B_i with mean b_i and variance σ_i^2. Let $\rho_i = \lambda_i b_i$ and $\rho = \sum_{i=1}^{n} \rho_i$ be the individual and total system utilizations (excluding setups), re-

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

spectively. For stability, we assume that $\rho < 1$ (see Takagi [22]). Finally, we define a variance term $\sigma^2 = \sum_{i=1}^{n} \lambda_i (b_i^2 + \sigma_i^2)$.

Server routing is defined by a polling table of length M ($M \geq n$). Each entry in the table is referred to as a *pseudostation.* Let m_i be the number of times queue i appears in the polling table, $1 \leq i \leq n$. Thus, cyclic service corresponds to a table with $M = n$ and $m_i = 1$ for all i, $1 \leq i \leq n$. A *cycle* is defined as the time to service all M pseudostations. Before service can commence at pseudostation j, $1 \leq j \leq M$, a setup time S_j is incurred with mean s_j and second moment $s_j^{(2)}$. Let S be the total setup time around a cycle and s its expected value ($=\sum_{j=1}^{M} s_j$). Define $a_j = s_j/s$, $1 \leq j \leq M$. In what follows, s will be increased to infinity with each a_j remaining constant. If c is the expected steady-state cycle time, then $c = s/(1-\rho)$ (see Boxma et al. [3]).

Let $T(j)$, $1 \leq j \leq M$, represent the queue corresponding to the jth pseudostation. Let $L(j)$ be the pseudostation that corresponds to the last visit to queue $T(j)$ prior to j. Note that $L(j)$ may be greater than j if none of pseudostations 1 through $j-1$ correspond to queue $T(j)$. If $m_{T(j)} = 1$, then $L(j) = j$. We define the *intervisit time* for pseudostation j as the time that elapses from the server leaving pseudostation $L(j)$ to finishing the setup at pseudostation j. The steady-state intervisit time for pseudostation j is represented by I_j. Define $l_j = E[I_j]/s$; then from Boxma et al. [3], it is easy to show that the l_j are independent of s (but not the a_k, $1 \leq k \leq M$). Define $I_i^{(k)}$ to be the kth intervisit time for queue i in the cycle, $1 \leq i \leq n$ and $1 \leq k \leq m_i$. Let $l_i^{(k)} = E[I_i^{(k)}]/s$. Then, the $l_i^{(k)}$ are independent of s.

Let V_j, $1 \leq j \leq M$, represent the steady-state *visit time* for pseudostation j; that is, the time the server spends serving pseudostation j. Then (see Boxma et al. [3]),

$$E[V_j] = \frac{\rho_{T(j)} E[I_j]}{1 - \rho_{T(j)}}, \quad 1 \leq j \leq M \tag{1}$$

and

$$E[I_j] = \sum_{k=L(j)+1}^{j-1} (s_k + E[V_k]) + s_j, \tag{2}$$

where a sum that is from j to k, where $j > k$, is assumed to be the sum from 1 to k plus the sum from j to M (we will assume this notation throughout the rest of this article). For notational convenience, we will also assume that index 0 equals M and index $M+1$ equals 1. Let $\bar{I}_i$, $1 \leq i \leq n$, be the steady-state intervisit time for queue i. Then (see Boxma et al. [3]),

$$E[\bar{I}_i] = \sum_{j:T(j)=i} \frac{E[I_j]}{m_i} = \sum_{k=1}^{m_i} \frac{E[I_i^{(k)}]}{m_i} = \frac{(1-\rho_i)c}{m_i} \quad \text{for } 1 \leq i \leq n. \tag{3}$$

If X is some random variable with probability density function (p.d.f.) $f_X(x)$ and finite expectation EX, then we define the *length-biased* random variable X^* (see, e.g., Baccelli and Brémaud [1]) as a random variable with p.d.f. $f_{X^*}(x) = x f_X(x)/EX$. See [1] for a more general definition if the random variable has no p.d.f. or infinite

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

expectation. Therefore, $\bar{I}_i^*$ represents the steady-state length-biased intervisit time, which may be thought of as the length of the intervisit time that is in progress at some random time when the system is in steady state. Note that $\bar{I}_i^*$ does not have the same distribution as $\bar{I}_i$ because the majority of time is spent in long cycles. If for each queue i, $1 \le i \le n$, $\bar{I}_i$ has p.d.f. $f_{\bar{I}_i}(x)$ and finite mean $E[\bar{I}_i]$ then the length-biased random variable $\bar{I}_i^*$ has p.d.f.

$$f_{\bar{I}_i^*}(x) = x f_{\bar{I}_i}(x)/E[\bar{I}_i]. \tag{4}$$

The *steady-state waiting time* at queue i, W_i, is defined as the time in steady state from a job's arrival at queue i until it commences service. This wait includes the wait for the server to return to serving queue i plus the wait while all jobs ahead of it in the queue are served.

The *decomposition property* for $M/G/1$ queues with generalized vacations of Fuhrmann and Cooper [13] states that for each queue i, $1 \le i \le n$, for the exhaustive service polling models considered in this article,

$$W_i \stackrel{\mathrm{d}}{=} Y_i + X_i,$$

where Y_i is the residual life of the intervisit time and X_i (independent of Y_i) is the waiting time in an $M/G/1$ queue with arrival rate λ_i and service time B_i. Note that X_i does not depend on s. Straightforward manipulation of distributions can be used to show that because the intervisit time is a nonnegative random variable, for all queues i, $1 \le i \le n$,

$$W_i \stackrel{\mathrm{d}}{=} \bar{I}_i^* U + X_i \tag{5}$$

where U is a uniform random variable (r.v.) on $[0,1]$ and $\bar{I}_i^*$ is the steady-state length-biased intervisit time for queue i. Note that $\bar{I}_i^*$, U, and X_i are mutually independent. Taking expected values yields

$$E[W_i] = \frac{E[\bar{I}_i^2]}{2E[\bar{I}_i]} + E[X_i] = \sum_{j:T(j)=i} \frac{E[I_j]^2(c_{I_j}^2 + 1)}{2(1-\rho_i)c} + E[X_i], \tag{6}$$

where $c_{I_j}^2$ is the squared coefficient of variation of the intervisit time for pseudostation j.

We take $\xrightarrow{P}$ to represent convergence in probability and $\Rightarrow$ to represent convergence in distribution. The following lemma is used to prove results in the next two sections.

LEMMA 1: *Let $\{X_t\}$ be a sequence of nonnegative random variables with distribution functions $\{F_t\}$. If $EX_t/t \to x$ and the squared coefficients of variation $c_t^2 \to 0$ as $t \to \infty$, then both $X_t/t \xrightarrow{P} x$ and $X_t^*/t \xrightarrow{P} x$ as $t \to \infty$.*

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

PROOF: We first note that $E[|X_t/t - x|] \to 0$ since, by Jensen's inequality,

$$\begin{aligned}(E[|X_t/t - x|])^2 &\le E[(X_t/t - x)^2] \\ &= E[(X_t/t - EX_t/t)^2] + (EX_t/t - x)^2 \\ &= (EX_t)^2 c_t^2/t^2 + (EX_t/t - x)^2.\end{aligned}$$

By the assumptions in the lemma, the right-hand side converges to zero. Therefore, convergence in probability of X_t/t to x follows by Markov's inequality.

Now, $E[X_t^*] = E[X_t^2]/E[X_t]$. Therefore, $\lim_{t\to\infty} E[X_t^*]/t = x$. Defining $1(A)$ as the indicator function of event A, we have that

$$\begin{aligned}E[|X_t^*/t - x|] &= E[(X_t^*/t - x)] + 2E[(x - X_t^*/t)1(X_t^*/t < x)] \\ &= E[(X_t^*/t - x)] + 2\int_0^{xt} \frac{(x - y/t)y}{E[X_t]}\, dF_t(y) \\ &\le E[(X_t^*/t - x)] + \frac{2xt}{E[X_t]\displaystyle\int_0^{xt} |x - y/t|\, dF_t(y)} \\ &\le E[(X_t^*/t - x)] + (2xt/E[X_t])E[|X_t/t - x|].\end{aligned}$$

Thus, $\lim_{t\to\infty} E[|X_t^*/t - x|] = 0$. Again, by Markov's inequality, the result follows. ■

3. LIMIT THEOREMS

This section presents the main results of this article, namely three limit theorems for polling models with increasing setups. Section 3.1 presents two limit theorems for systems with polling tables with deterministic setups. Section 3.2 presents a limit theorem for cyclic polling models with general setups under heavy traffic.

3.1. Polling Tables with Deterministic Setups

In this section we implicitly consider a sequence of polling models with deterministic setups in which only the total setup time s changes. In other words, we are considering a parametric family of polling models, each satisfying the description in Section 2 and differing only with respect to the total setup time s. The following lemma shows the deterministic nature of the system under large deterministic setups.

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

LEMMA 2: *In exhaustive polling models with polling tables, if setups are deterministic, then for all* j, $1 \le j \le M$, $I_j/s \xrightarrow{P} l_j$ *as* $s \to \infty$ *(with* $a_k = s_k/s$ *remaining constant for all* k, $1 \le k \le M$*).*

PROOF: From Lemma 1, it suffices to show that $c_{I_j}^2 \to 0$ as $s \to \infty$ for all j, $1 \le j \le M$. The key to this proof will be to show that

$$\sum_{i=1}^{n} \rho_i E[W_i] = \frac{\rho\sigma^2}{2(1-\rho)} + \frac{\rho \operatorname{Var}[S]}{2s} + \sum_{i=1}^{n} \sum_{j:T(j)=i} \frac{\rho_i E[I_j]^2}{2(1-\rho_i)c}. \tag{7}$$

Equation (7) will be proven below, but assume for the moment that it holds. Equation (6) implies that

$$\sum_{i=1}^{n} \rho_i E[W_i] = \sum_{i=1}^{n} \rho_i \left(\sum_{j:T(j)=i} \frac{E[I_j]^2 (c_{I_j}^2 + 1)}{2(1-\rho_i)c} + E[X_i] \right),$$

where $E[X_i] = \lambda_i b_i^{(2)}/2(1-\rho_i)$. Therefore, by (7),

$$\sum_{i=1}^{n} \rho_i \sum_{j:T(j)=i} \frac{E[I_j]^2 c_{I_j}^2}{2(1-\rho_i)c} = \frac{\rho\sigma^2}{2(1-\rho)} + \frac{\rho \operatorname{Var}[S]}{2s} - \sum_{i=1}^{n} \rho_i E[X_i].$$

Therefore, using $c = s/(1-\rho)$,

$$\sum_{i=1}^{n} \rho_i \sum_{j:T(j)=i} \frac{(1-\rho) l_j^2 c_{I_j}^2}{1-\rho_i} = \frac{\rho\sigma^2}{s(1-\rho)} + \frac{\rho \operatorname{Var}[S]}{s^2} - \sum_{i=1}^{n} \rho_i 2E[X_i]/s.$$

If setup times are deterministic, then $\rho \operatorname{Var}[S]/s^2 = 0$. Clearly, $\lim_{s\to\infty} \rho\sigma^2/s(1-\rho) = 0$ and $\lim_{s\to\infty} \sum_{i=1}^{n} \rho_i 2E[X_i]/s = 0$. Because $(1-\rho)\rho_i l_j^2/(1-\rho_i) > 0$, we have that $\lim_{s\to\infty} c_{I_j}^2 = 0$ for all j, $1 \le j \le M$. Therefore, it suffices to prove (7).

Let $z_{ij} = 1$ if none of pseudostations $i+1, i+2, \ldots, j$ correspond to queue $T(i)$, and 0 otherwise. Let $w_{ij} = \rho_{T(i)} z_{ij} \sum_{k=i+1}^{j} (s_k + E[V_k])$. Let $\bar{w}_j = \sum_{i=1}^{M} w_{ij}$. Let $x_{ij} = w_{ij} + \rho_{T(i)} z_{ij} s_{j+1}$ and $\bar{x}_j = \sum_{i=1}^{M} x_{ij} = \bar{w}_j + \rho s_{j+1}$. Then, $E[I_j] = x_{L(j),j-1}/\rho_{T(j)}$ and, using the previous definitions, a little algebra yields $E[V_j] = (\bar{w}_j - \bar{x}_{j-1})/(\rho - 1)$. From the psuedoconservation law for polling tables of Boxma et al. [3],

$$\begin{aligned} \sum_{i=1}^{n} \rho_i E[W_i] &= \frac{\rho\sigma^2}{2(1-\rho)} + \frac{\rho \sum_{i=1}^{M} s_i^{(2)}}{2s} + \sum_{i=1}^{M} \sum_{j=1}^{M} \frac{s_{j+1}}{s} w_{ij} \\ &= \frac{\rho\sigma^2}{2(1-\rho)} + \frac{\rho \operatorname{Var}[S]}{2s} + \sum_{j=1}^{M} \frac{s_{j+1}}{2s} (\bar{w}_j + \bar{x}_j). \end{aligned}$$

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

Now, $s_{j+1} = (\bar{x}_j - \bar{w}_j)/\rho$ so that

$$\begin{aligned}
\sum_{j=1}^{M} \frac{s_{j+1}}{2s} (\bar{w}_j + \bar{x}_j) &= \sum_{j=1}^{M} \frac{1}{2\rho s} (\bar{x}_j^2 - \bar{w}_j^2) \\
&= \sum_{j=1}^{M} \frac{1}{2\rho c} (\bar{x}_j^2 - \bar{w}_j^2) + \sum_{j=1}^{M} \frac{1}{2(1-\rho)c} (\bar{x}_{j-1}^2 - \bar{w}_j^2) \\
&= \frac{1}{2c} \sum_{j=1}^{M} \sum_{i=1}^{M} (x_{ij} + w_{ij}) s_{j+1} + \frac{1}{2c} \sum_{i=1}^{M} \sum_{j=1}^{M} (x_{i,j-1} + w_{ij}) E[V_j] \\
&= \frac{1}{2c} \sum_{i=1}^{M} \sum_{j=1}^{M} (x_{L(i),j} + w_{L(i),j}) s_{j+1} \\
&\quad + \frac{1}{2c} \sum_{i=1}^{M} \sum_{j=1}^{M} (x_{L(i),j-1} + w_{L(i,j)}) E[V_j].
\end{aligned}$$

Using the fact that $z_{L(i),j} = 0$ for $j \geq i$,

$$\begin{aligned}
\sum_{j=1}^{M} \frac{s_{j+1}}{2s} (\bar{w}_j + \bar{x}_j) &= \frac{1}{2c} \sum_{i=1}^{M} \sum_{j=L(i)+1}^{i-1} (x_{L(i),j} + w_{L(i),j}) s_{j+1} \\
&\quad + \frac{1}{2c} \sum_{i=1}^{M} \sum_{j=L(i)+1}^{i-1} (x_{L(i),j-1} + w_{L(i),j}) E[V_j] + x_{L(i),i-1} E[V_i] \\
&= \frac{1}{2c} \sum_{i=1}^{M} \sum_{j=L(i)+1}^{i-1} (x_{L(i),j}^2 - w_{L(i),j}^2)/\rho_{T(i)} \\
&\quad - \frac{1}{2c} \sum_{i=1}^{M} \sum_{j=L(i)+1}^{i-1} (x_{L(i),j-1}^2 - w_{L(i),j}^2)/\rho_{T(i)} + x_{L(i),i-1} E[V_i] \\
&= \sum_{i=1}^{M} \frac{1}{2c} (x_{L(i),i-1}^2/\rho_{T(i)} + x_{L(i),i-1} E[V_i]) \\
&= \sum_{i=1}^{M} \frac{1}{2c} \rho_{T(i)} E[I_i](E[I_i] + E[V_i]) \\
&= \sum_{i=1}^{M} \frac{\rho_{T(i)} E[I_i]^2}{2c(1-\rho_{T(i)})} \\
&= \sum_{i=1}^{n} \sum_{j:T(j)=i} \frac{\rho_i E[I_j]^2}{2(1-\rho_i)c}.
\end{aligned}$$

We have shown (7); therefore, the result follows. ■

Note that if the setup times are not deterministic but are i.i.d. over all queues, then $\lim_{n\to\infty} \rho \operatorname{Var}[S]/s^2 = 0$. Hence, the above result holds if we let $n \to \infty$ (which automatically implies $s \to \infty$). However, it is not entirely clear what is meant by

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

letting $n \to \infty$ in asymmetric polling models (polling models where not all queues are identical); thus, we have not stated this as a formal result.

THEOREM 1: *In exhaustive service polling models with polling tables, if setups are deterministic, U is a uniform r.v. on* $[0,1]$ *and* Z_i *are discrete-valued random variables,* $1 \le i \le n$, *independent of each other and U, that take values* $l_i^{(j)}$, $1 \le j \le m_i$, *with probability* $l_i^{(j)} / \sum_{k=1}^{m_i} l_i^{(k)}$, *then* $W_i/s \Rightarrow Z_i U$ *as* $s \to \infty$ (*with* $a_k = s_k/s$ *remaining constant for all k,* $1 \le k \le M$).

PROOF: Lemmas 1 and 2 imply that $\bar{I}_i^*/s \Rightarrow Z_i$ as $s \to \infty$. Therefore, because $X_i/s \xrightarrow{P} 0$ as $s \to \infty$, (5) yields the desired result. ■

COROLLARY 1: *In exhaustive service polling models with polling tables, if for each queue i,* $1 \le i \le n$, *the intervisit times* $I_i^{(k)}$ *are identically distributed,* $1 \le k \le m_i$, *setup times are deterministic, and U is a uniform r.v. on* $[0,1]$, *then* $W_i/s \Rightarrow (1 - \rho_i)/(m_i(1 - \rho))U$ *as* $s \to \infty$ (*with* $a_k = s_k/s$ *remaining constant for all k,* $1 \le k \le M$).

PROOF: If the intervisit times for queue i, $1 \le i \le n$, are identically distributed, then from (3), $E[I_i^{(k)}]/s = l_i^{(k)} = (1 - \rho_i)/m_i(1 - \rho)$ for $1 \le k \le m_i$ and $1 \le i \le n$. Therefore, the result follows from Theorem 1. ■

The above corollary implies that for cyclic service polling models (where $m_i = 1$ for each queue i), the waiting time distribution converges to a uniform distribution as setups get large. Although not exact, an intuitive explanation for this distribution is that it represents how far around the cycle the server is when an arrival occurs. Furthermore, the recommended method for constructing polling tables is to space the queues as evenly as possible (see, e.g., [4]) so that the above result should hold approximately for well-designed polling tables.

3.2. Limits for Cyclic Service Polling Models Under Heavy Traffic

Using heavy-traffic theory, Coffman et al. [6] (hereafter, CPR) have derived limiting distributions (as $\rho \uparrow 1$) for the waiting time in cyclic serve-to-exhaustion polling models. Our next result involves a more complicated sequence of system models in the spirit of that considered by CPR. Consider a single cyclic exhaustive service polling system with interarrival times, service times, and setup times forming mutually independent renewal sequences with finite second moments. Now, create a doubly-indexed family of polling models by scaling the total arrival rate, λ, and the total mean setup time around a cycle, s, such that the proportions λ_i/λ and s_i/s_j $(= a_i/a_j)$ remain fixed. First, for each fixed s, we let λ increase in such a way that ρ approaches 1 from below. We then invoke the limit theorem proved by CPR to obtain a family of diffusions indexed by s. Finally, we let $s \to \infty$.

As is frequently the case for these heavy-traffic limit theorems, CPR do not show that the steady-state distribution of the polling system converges to the steady-

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

state distribution of the limiting diffusion process. However, they hypothesize that this is the case.

THEOREM 2: *In the sequence of polling models described immediately above, if the steady-state distribution of the polling system converges in distribution to the steady-state distribution of the limiting diffusion process as* $\rho \uparrow 1$, *then* $(1-\rho)W_i/s \Rightarrow (1-p_i)U$ *as* $\rho \uparrow 1$ *and then* $s \to \infty$ (*where* $p_i = \lim_{\rho \uparrow 1} \rho_i$ *and* U *is a uniform r.v. on* $[0,1]$).

PROOF: The CPR limit theorems imply that (see [16] for details)

$$(1-\rho)W_i \Rightarrow U\hat{I}_i \quad \text{as } \rho \uparrow 1, \tag{8}$$

where U is a uniform $[0,1]$ random variable that is independent of $\hat{I}_i$ and

$$(1-\rho)I_i^* \Rightarrow \hat{I}_i \quad \text{as } \rho \uparrow 1.$$

Note that this corresponds to the same result as (5), but X_i is negligible under heavy traffic. The CPR results further imply that $\hat{I}_i$ is a gamma random variable with density $f(x) = \alpha_i^{\beta} x^{\beta-1} e^{-\alpha_i x}/\Gamma(\beta)$, scale parameter

$$\alpha_i = \frac{2\sum_{j=1}^{n}\sum_{k<j} p_j p_k}{\sigma^2(1-p_i)},$$

and shape parameter

$$\beta = 2\sum_{j=1}^{n}\sum_{k<j} p_j p_k s/\sigma^2 + 1.$$

Therefore,

$$\lim_{s\to\infty} E[\hat{I}_i/s] = \lim_{s\to\infty} \frac{\left(2\sum_{i=1}^{n}\sum_{j<i} p_i p_j/\sigma^2 + 1/s\right)\sigma^2(1-p_i)}{2\sum_{i=1}^{n}\sum_{j<i} p_i p_j} = 1-p_i$$

and

$$c_{\hat{I}_i}^2 = \frac{1}{1+2\sum_{i=1}^{n}\sum_{j<i} p_i p_j s/\sigma^2} \to 0 \quad \text{as } s \to \infty.$$

Thus, by Lemma 1, $\hat{I}_i/s \xrightarrow{P} 1-p_i$ as $s \to \infty$. Putting this together with (8) yields the desired result. ■

Note that for the special case of deterministic setups, Theorem 2 with "$\rho \uparrow 1$ and then $s \to \infty$" replace by "$s \to \infty$ and then $\rho \uparrow 1$" is implied by Corollary 1 by letting $\rho \uparrow 1$. Although the CPR result is only for systems with cyclic service, it appears that

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

it might be extended to systems with polling tables. The intuition for this result is that as ρ tends to 1, the variance of the setup times becomes negligible. However, as s tends to infinity, all other sources of variance also become negligible.

4. APPLICATIONS FOR THE LIMIT THEOREMS

Although the limit theorems of the previous section may be of academic interest, the question remains as to whether they are of any practical value. In this section, we investigate their applicability. In particular, Section 4.1 discusses applications with respect to system performance evaluation. Section 4.2 discusses applications with respect to the other authors' previous work. Finally, Section 4.3 investigates numerically whether these limit theorems may hold for more general settings than the specific models considered in this article. It also discusses possible ways of extending this work.

4.1. Insights on System Behavior

Putting the discussions in Section 2 together, intervisit times should be approximately constant when any of the following are true:

1. The total setup s is large and setups have low variance.
2. There are a large number of queues, each with similar and significant setup times.
3. The total setup s is large and the system is in heavy traffic (i.e., ρ is close to 1).

If any of the above three traits are true, then the waiting time distribution for each queue may be approximated by a uniform distribution for cyclic service systems or systems where the pseudostations for each queue have been evenly spaced on the polling table, resulting in similar intervisit times. In the traditional manner of adjusting heavy-traffic results to correspond to known exact results (e.g., see Whitt [28]), a better approximation for the waiting time is to use an estimate of $2E[W_i]$ as the scale factor for the uniform distribution for queue i, $1 \leq i \leq n$ rather than $E[I_i]$ (as suggested by the theorems). If the system is asymmetric, then, given the simplicity of the approximation, it would make more sense to use an explicit approximation for $E[W_i]$ (such as the very accurate one provided by Everitt [9]) rather than a more accurate but more complicated numerical algorithm. If any of the three above traits are true and pseudostations have not been evenly spaced, then the waiting time distribution may be approximated by a uniform distribution multiplied by a simple (known) random variable, as described in Theorem 1.

To illustrate the above trends, we simulated a system with identical queues, Poisson arrivals, and exponential service with mean of 1 min. Setup times are taken to be deterministic or exponential with mean of 5, 10, or 20 min. There are 10 or 20 queues in the system and ρ is taken to be 0.85, 0.9, or 0.95. The queues are served using cyclical exhaustive service. The following statistic was computed:

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

$$\sup_x |F^{\text{app}}(x) - F^{\text{sim}}(x)| \times 100,$$

where $F^{\text{sim}}(x)$ is the empirical distribution function computed from the simulation data and $F^{\text{app}}(x)$ is the approximating uniform distribution (using $2E[W_i]$ as the scale factor, as suggested above). As $F^{\text{sim}}(x)$ only changes at each of the waiting time points, this supremum need only be calculated at each x that corresponds to a jump in the empirical distribution (i.e., at each waiting time). This statistic is similar to the one used in Federgruen and Katalan [10] and can be seen to bear a relation to the Kolmogorov–Smirnov statistic.

Table 1 shows the results for the metric for the cases described above. It can be seen that as setup time increases, the metric decreases. When setup times are not deterministic, as ρ increases, the metric decreases. If setup times are deterministic then there is no apparent trend with respect to ρ. The metric is greater for exponential setups than for deterministic setups. The metric is smaller at $n = 20$ than $n = 10$. All these trends were predicted above.

Table 2 examines the same data as in Table 1 with respect to two more quantities of interest. Setup times have been scaled down by a factor of 10 to give larger values in the table. In the columns labeled scv, the squared coefficient of variation of the intervisit time is given. This can be seen to be very small in general and decreasing to zero in accordance with the three trends predicted above. The only exception to these trends is that the squared coefficient of variation increases in ρ for the smallest value of the setup ($s_i = 0.5$). The columns labeled % give the percentage error in predicting the 90th percentile of waiting time. The percentage error is defined as 100(Estimate − Simulated Value)/Estimate. As would be expected, the errors decrease in accordance with the trends described above.

As a final example of illustrating this distributional approximation, Figure 1 gives a histogram of the waiting time for the 10-queue model with $s_i = 20$, $\rho = 0.8$, and deterministic setups. It can be seen that the distribution does, indeed, look uniform until the upper percentiles are reached.

TABLE 1. Trends in Waiting Time Approximation

		$n = 10$			$n = 20$		
Setup Distn.	ρ	$s_i = 5$	$s_i = 10$	$s_i = 20$	$s_i = 5$	$s_i = 10$	$s_i = 20$
Det.	0.85	6.11	4.64	3.40	4.57	3.33	2.44
Det.	0.90	6.30	4.76	3.52	4.66	3.46	2.55
Det.	0.95	6.31	4.76	3.54	4.68	3.46	2.54
Exp.	0.85	7.06	5.99	5.38	5.51	4.45	3.99
Exp.	0.90	6.93	5.79	5.14	5.22	4.26	3.89
Exp.	0.95	6.77	5.45	4.58	5.06	4.11	3.46

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

TABLE 2. Trends in Waiting Time Approximation

		$n = 10$						$n = 20$					
		$s_i = 0.5$		$s_i = 1$		$s_i = 2$		$s_i = 0.5$		$s_i = 1$		$s_i = 2$	
Setup Distn.	ρ	%	scv	%	scv	%	scv	%	scv	%	scv	%	scv
Det.	0.85	−18.23	0.28	−11.06	0.17	−5.69	0.09	−10.88	0.17	−5.29	0.09	−2.21	0.04
Det.	0.90	−17.72	0.31	−10.28	0.18	−4.77	0.09	−10.17	0.17	−4.68	0.09	−1.36	0.04
Det.	0.95	−15.24	0.37	−9.06	0.19	−4.17	0.09	−8.63	0.18	−3.97	0.09	−1.24	0.04
Exp.	0.85	−18.39	0.28	−12.62	0.19	−7.75	0.11	−11.12	0.17	−6.37	0.10	−3.57	0.05
Exp.	0.90	−17.52	0.32	−10.84	0.19	−5.80	0.10	−10.17	0.17	−5.33	0.09	−2.24	0.05
Exp.	0.95	−15.81	0.37	−9.09	0.19	−4.77	0.09	−8.85	0.18	−4.40	0.09	−1.44	0.04

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

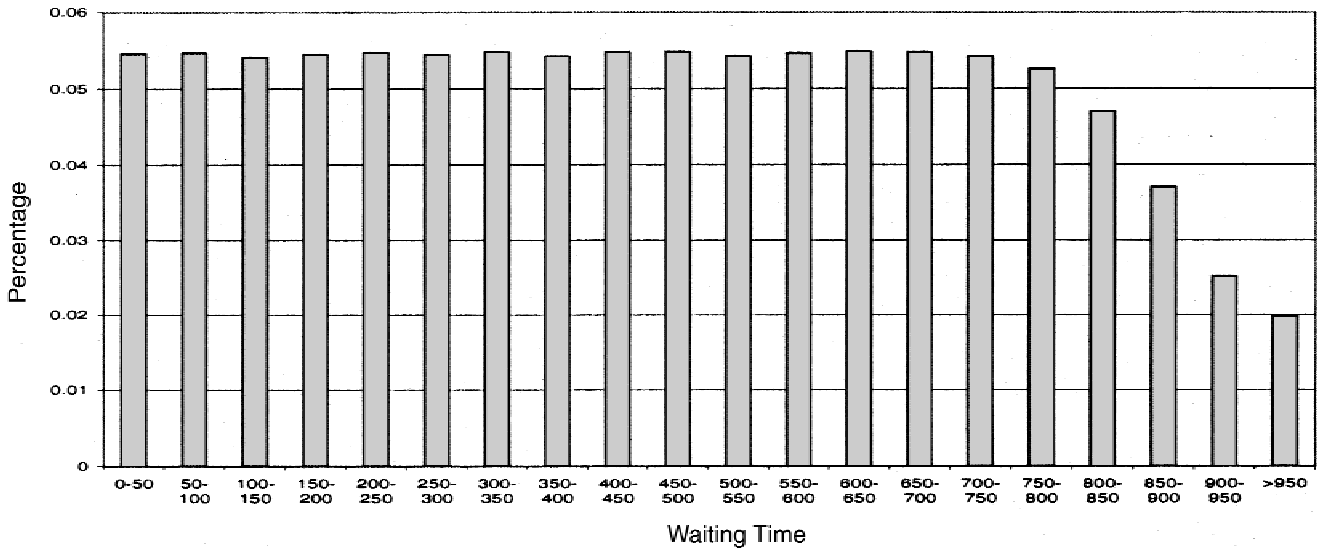


FIGURE 1. Histogram of waiting time distribution.

At this point, we feel it is prudent to remind readers that, as described in Section 1, a number of good approximations for the waiting time have already been developed. Therefore, this approximation should be used more as an intuition builder than an accurate approximation to the waiting time. As such, we do not feel that it is necessary to do extensive numerical testing of the accuracy of the approximation. Readers wanting a fairly accurate closed-form approximation are referred to the heavy-traffic approximation given in Olsen [18].

We now highlight further the insight that this uniform approximation yields into the behavior of the system. For example, suppose that some system satisfies the assumptions of Corollary 1. Then, if $P_i(p)$ is the pth percentile of the waiting time distribution at queue i (i.e., the number x such that $\text{Prob}(W_i \leq x) = p$), then $P_i(p) \approx l_i sp = (1-\rho_i)sp/(1-\rho)m_i$, $1 \leq i \leq n$. This shows that any percentile of the waiting time distribution is approximately proportional to s and is approximately proportional to $1/(1-\rho)$. This can aid in decision making when evaluating the worth of setup time reduction or the impact of increased demand.

As an example, Figures 2 and 3 examine these trends for a system with 20 identical queues, Poisson arrivals, exponential service with mean of 1 h, and deterministic setup times. Figure 2 has $\rho = 0.8$ and plots the simulated 95th percentile of waiting times versus setup times of mean 1–5 h. Figure 3 has mean setup time equal to 1 h and plots the simulated log of the 95th percentile of waiting time versus $\log(1/(1-\rho))$ for ρ ranging from 0.6 to 0.95 (logs are taken so that the scale is more readable—a linear trend with slope 1 indicates a linear trend in the original model). It can be seen in Figures 2 and 3 that the linear trends predicted by the approximation do appear to hold in the simulated systems.

4.2. Insights Provided for Previous Work

The limit theorems in this article provide insight into previous work by other authors. This section examines the impact of this work on an article by Boxma et al. [4] and another by Sarkar and Zangwill [21].

Boxma et al. [4] present an article on the scheduling of polling models with polling tables and exhaustive service. They show that the waiting time is minimized when $c_{I_i}^2$ is taken to be zero and perform their heuristic analysis assuming that this is the case. They also present a lower bound on the waiting time under their heuristic by taking I_i to be a constant. Therefore, our limit theorems suggest that not only should the performance of their heuristic improve as setup time increases but also that the performance of their lower bound should improve as well. Table 3 confirms that this is indeed the case.

Table 3 corresponds to Table 2 in Boxma et al. [4] except with exhaustive service. The system contains 13 queues with queues 2–13 being identical. The polling table contains between 13 and 19 pseudostations with only queue 1 being repeated (see [4] for the specific tables). Both service and setup times are deterministic. Service times have mean 1 for all queues and setup times have mean s_i as given in Table 3. The arrival rates to stations 2–13 are all equal to 0.02, whereas that of

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

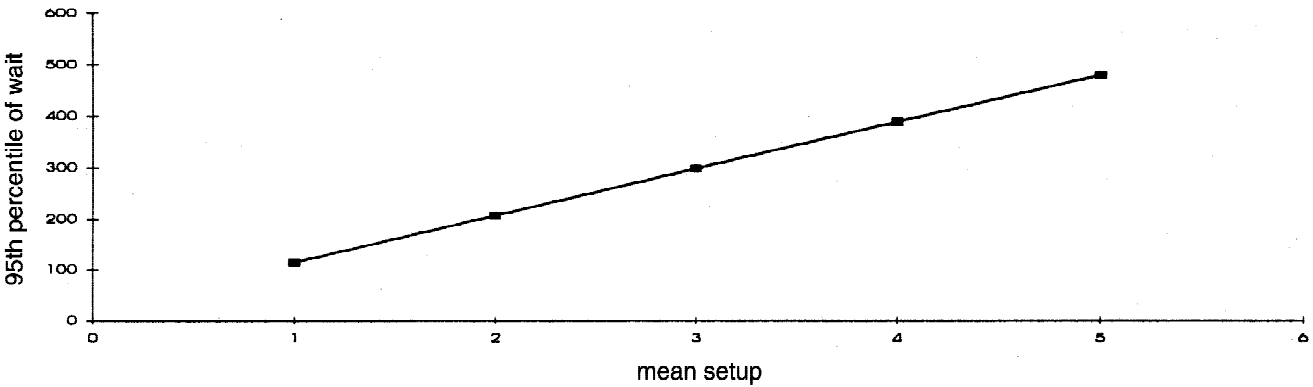


FIGURE 2. Effect of setup time on simulated 95th percentile of the waiting time.

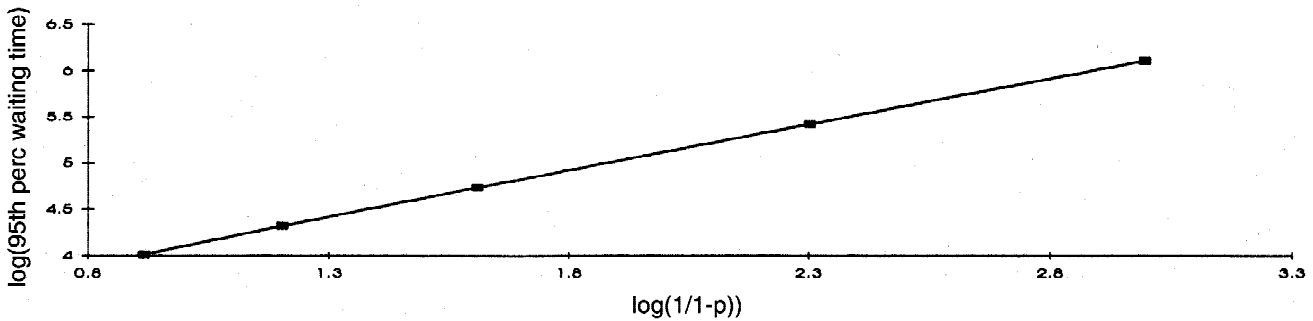


FIGURE 3. Effect of ρ on simulated 95th percentile of waiting time.

TABLE 3. Effect of Setup Times on Percentage Error for the Boxma et al. Bound

λ_1, λ_i	$s_i = 0.1$	$s_i = 0.5$	$s_i = 1$	$s_i = 2$	$s_i = 5$	$s_i = 10$
0.02, 0.02	15.9	3.7	2.0	1.0	0.4	0.2
0.075, 0.02	19.2	4.6	2.4	1.2	0.5	0.3
0.16, 0.02	23.9	6.2	3.3	1.7	0.8	0.4
0.26, 0.02	28.7	8.2	4.3	2.3	0.9	0.5
0.49, 0.02	39.2	13.3	7.3	3.9	1.6	0.8

station 1 ranges from 0.02 to 0.49, as shown. This model satisfies the assumptions of Theorem 1. The percentage error in estimating $\sum_{i=1}^{13} \lambda_i EW_i$ is given in the table. It can be seen that this error sharply decreases as the setup times increase.

Sarkar and Zangwill [21] provide examples where the decreasing mean setup time can increase the mean waiting time in the system. This phenomenon has been the subject of a heavy debate as to its relevance and general applicability (see, e.g., [7,8,11,14]). In the specific model presented in this article, our results provide an intuitive explanation for this phenomenon. As the setup time decreases, the coefficient of variation of the intervisit time increases. So, whereas decreasing mean setup time will decrease the mean intervisit time, expected waiting time may still increase.

4.3. Extensions to More General Models

This section examines whether the limit theorems described in this article might, in fact, apply to more general settings than those assumed above. We examine the variability in intervisit times under a dynamic scheduling rule using an industry dataset from a fence-making operation. Both the dataset and the scheduling rule are explained more fully in Olsen [19]. The dataset consists of 19 different products with an average mean service time of 0.1 h and a range from 0.02 to 0.27 h. The average mean setup time over all products is 2 h, with a range from 0.17 to 11.3 h. The arrival rates at the different queues are also diverse and range from 0.17% of the total arrival rate to 29.9% of the total arrival rate. We use the exponential distribution for both the setup and the service time distributions. The scheduling heuristic used chooses the next queue to serve as that with the most scaled age (both the scale factor and age are defined in [19]). Therefore, there is no prespecified order in which the queues are visited. Service is exhaustive and setups are only performed if there are customers present in the queue. Clearly, this is a much more dynamic scheduling rule than that used in Section 2.

Figure 4 plots the squared coefficient of variation of intervisit time versus a setup scale factor. A scale factor of 1 represents the setups from the original dataset, whereas a scale factor of 2 doubles the setups and a scale factor of 0.5 halves the setups. The squared coefficient of variation is plotted for three different values of ρ (denoted by u for utilization). The figure clearly shows that the squared coefficient

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

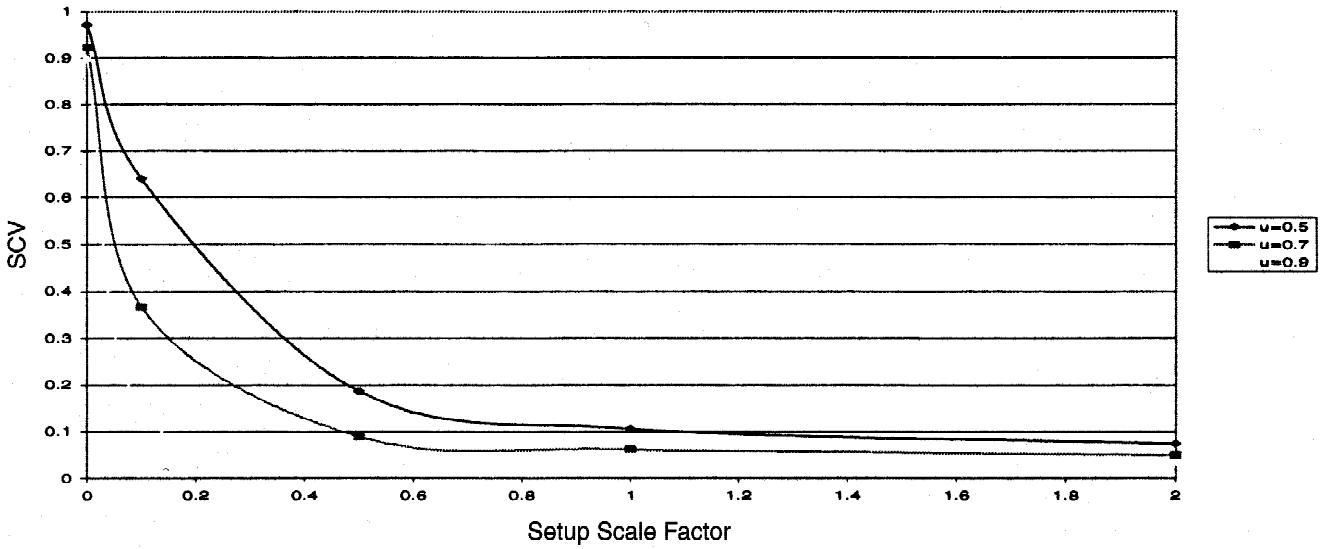


FIGURE 4. Effect of setup time on simulated 95th percentile of the waiting time.

of variation decreases as the setup time and ρ increase. In other words, the trends predicted by the above limit theorems carry over to this much more dynamic system. Note that when the setup scale factor is zero (so the system has no setups), then the squared coefficient of variation is close to one (a highly variable system) but having even moderate setups results in a large decrease in variability.

The intuition behind these results is that there appears to be a strong law-type effect occurring. Consider a periodic system with deterministic setups. It can be seen that

$$V_k \stackrel{\mathrm{d}}{=} \sum_{i=1}^{N_k} B_k^i$$

where V_k is the visit time for queue k, N_k is the number of customers that have arrived in the previous intervisit period, and B_k^i is the length of the busy period generated by the ith customer. Note that the B_k^i, $i \geq 1$, are i.i.d. random variables. If $N_k/s \to n_k$ a.s. as $s \to \infty$, then it is easy to show that the strong law of large numbers implies that $V_k/s \to c_k$ a.s. as $s \to \infty$, where c_k is some constant. Suppose that the previous appearance of queue $T(k)$ on the table corresponds to pseudostation j. Then, N_k is the number of Poisson arrivals in $\sum_{i=j+1}^{k-1}(s_i + V_i) + s_k$. If $V_i/s \to c_i$ a.s. as $s \to \infty$, then the result is proved. However, clearly, this is a circular argument that ignores the dependence between V_k and V_i.

Furthermore, there is no reason why a strong law effect as described above should not carry over to systems under different scheduling methods, so long as the setups perform a regulating effect. This appears to explain the phenomenon observed in Figure 4. Therefore, perhaps even in dynamic systems, similar limit theorems can be developed. This is left as the subject for future research. If true, then one obvious application is to use results for deterministic settings to approximate stochastic systems. Deterministic settings are frequently easier to analyze than their stochastic counterpart. This is similar to the approach that was used by Boxma et al. [4] to design effective polling tables. In [4], the system was lower bounded by its deterministic counterpart and then efficient queue visit frequencies were given for the deterministic system. These queue visit frequencies were then used to design polling tables. As illustrated in Section 4.2, such a method becomes more effective as setups increase.

Two questions remain: How broad a class of policies will these limit theorems hold for? How should such more general proofs proceed? In any generalization of these results, the setup times will need to perform a regulating effect in that one of the three traits given in Section 4.1 would need to hold. Resing [20] classifies polling systems as those which can be represented as multitype branching processes and those that cannot and explains why those that can are easier to analyze than those that cannot. A reasonable hypothesis is that our limit theorems may hold for systems that satisfy his Property 1 and, therefore, can be represented as multitype branching processes. Exhaustive and gated service disciplines fit into this framework but 1-limited service (see [23]) does not.

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

Key to our proof was the decomposition property of Fuhrmann and Cooper [13]. Bertsimas and Mourtzinou [2] have generalized this result and such generalizations may well serve as a starting point for proving our result in more general settings. With such a starting point, it would remain to prove that intervisit times converge to a constant and strong law results may well be the way to do this.

5. CONCLUSIONS

This article has presented limit theorems for (a) polling models with polling tables, exhaustive service, and deterministic setups and (b) cyclic exhaustive service polling models with general setups under heavy traffic. It was shown that as the setup time increases to infinity, the scaled intervisit time for each queue converges in probability to a constant. This, in turn, was shown to imply that the scaled waiting time converges in distribution to either the uniform distribution, or a simple random variable multiplied by a uniform random variable, as the mean setup time tends to infinity. Numerical testing revealed that these limits are relevant even for systems with moderate setups and appear to carry over to more general settings than the specific scheduling mechanism described in this article.

Acknowledgments

I am grateful to Peter Glynn and Mike Harrison for valuable advice on this topic.

References

1. Baccelli, F. & Brémaud, P. (1991). *Elements of queueing theory.* Berlin: Springer-Verlag.
2. Bertsimas, D. & Mourtzinou, G. (1999). Decomposition results for general polling systems and their applications. *Queueing Systems Theory and Applications* 31: 295–316.
3. Boxma, O., Groenendijk, W., & Weststrate, J. (1990). A pseudoconservation law for service systems with a polling table. *IEEE Transactions on Communications* COM-38(10): 1865–1870.
4. Boxma, O., Levy, H., & Weststrate, J. (1991). Efficient visit frequencies for polling tables: Minimization of waiting cost. *Queueing Systems Theory and Applications* 9(1–2): 133–162.
5. Choudhury, G.L. & Whitt, W. (1996). Computing transient and steady-state distributions in polling models by numerical transform inversion. *Performance Evaluation* 25: 267–292.
6. Coffman, E., Jr., Puhalskii, A., & Reiman, M. (1998). Polling systems in heavy-traffic: A Bessel process limit. *Mathematics of Operations Research* 23(2): 257–304.
7. Cooper, R.B., Niu, S., & Srinivasan, M.M. (1997). Setups in polling models: Does it make sense to set up if no work is waiting? Working paper.
8. Duenyas, I. (1994). The limitations of sub-optimal policies. *Interfaces* 24(5): 77–84.
9. Everitt, D. (1986). Simple approximations for token rings. *IEEE Transactions on Communications* COM-34(7): 719–721.
10. Federgruen, A. & Katalan, Z. (1994). Approximating queue size and waiting time distributions in general polling systems. *Queueing Systems Theory and Applications* 18: 353–386.
11. Federgruen, A. & Katalan, Z. (1996). The impact of setup times on the performance of multi-class service and production systems. *Operations Research* 44(6): 989–1001.
12. Federgruen, A. & Katalan, Z. (1996). The stochastic economic lot scheduling problem: Cyclical base stock policies with idle times. *Management Science* 42(6): 783–796.
13. Fuhrmann, S.W. & Cooper, R.B. (1985). Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Operations Research* 33(5): 1117–1129.

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press

14. Gupta, D. & Srinivasan, M.M. (1996). The variance paradox and its implications for Japanese production theory. *Interfaces* 26: 69–77.
15. Harrison, J.M. (1985). *Brownian motion and stochastic flow systems.* New York: Wiley.
16. Lennon, T.M. (1994). Response-time approximations for multi-server polling models, with manufacturing applications. Unpublished Ph.D. thesis, Stanford University.
17. Levy, H. & Sidi, M. (1990). Polling systems: Applications, modeling, and optimization. *IEEE Transactions on Communications* COM-38(10): 1750–1760.
18. Olsen, T.L. (1998). Approximations for the waiting time distribution in polling models with and without state-dependent setups. *Operations Research Letters* (to appear).
19. Olsen, T.L. (1999). A practical scheduling method for multi-class production systems with setups. *Management Science* 45(1): 116–130.
20. Resing, J.A.C. (1993). Polling systems and multitype branching processes. *Queueing Systems Theory and Applications* 13: 409–426.
21. Sarkar, D. & Zangwill, W. (1991). Variance effects in cyclic production systems. *Management Science* 37(4): 443–453.
22. Takagi, H. (1986). *Analysis of polling systems.* Cambridge, MA: The MIT Press.
23. Takagi, H. (1990). Queueing analysis of polling models: An update. In H. Takagi (ed.), *Stochastic analysis of computer and communication systems.* Amsterdam: Elsevier Science/North-Holland, pp. 267–318.
24. Takagi, H. (1997). Queueing analysis of polling models: Progress in 1990–1994. In J.H. Dshalalow (ed.), *Frontiers in queueing*. Boca Raton, FL: CRC Press, pp. 119–146.
25. van der Mei, R.D. (1999). Polling systems in heavy traffic: Higher moments of the delay. *Queueing Systems Theory and Applications* 31: 265–294.
26. van der Mei, R.D. (1999). Polling systems with periodic server routing in heavy traffic. *Stochastic Models* 15: 273–297.
27. van der Mei, R.D. & Levy, H. (1997). Polling systems in heavy traffic: Exhaustiveness of service policies. *Queueing Systems Theory and Applications* 27: 227–250.
28. Whitt, W. (1982). Refining diffusion approximations for queues. *Operations Research Letters* 1(5): 165–169.

https://doi.org/10.1017/S026996480115103X Published online by Cambridge University Press