Volume 11, Number 3
Fall 1998

ISSN 1048-9533

J.A.M.S.A.

Journal of Applied Mathematics and Stochastic Analysis

Special Issue Dedicated to RYSZARD SYSKI

EDITED by Jewgeni H. Dshalalow
Florida Institute of Technology, Melbourne, U.S.A.

North Atlantic Science Publishing Company

Journal of Applied Mathematics and Stochastic Analysis, 11:3 (1998), 355-368.

SOME REFLECTIONS ON THE RENEWAL-THEORY PARADOX IN QUEUEING THEORY[1]

ROBERT B. COOPER
Florida Atlantic University
Department of Computer Science and Engineering
Boca Raton, FL 33431-0991 USA

SHUN-CHEN NIU
The University of Texas at Dallas
School of Management, P.O. Box 830688
Richardson, TX 75083-0688 USA

MANDYAM M. SRINIVASAN
The University of Tennessee
Management Science Program, College of Business Admin.
Knoxville, TN 37996-0562 USA

(Received October, 1997; Revised January, 1998)

The classical renewal-theory (waiting time, or inspection) paradox states that the length of the renewal interval that covers a randomly-selected time epoch tends to be longer than an ordinary renewal interval. This paradox manifests itself in numerous interesting ways in queueing theory, a prime example being the celebrated Pollaczek-Khintchine formula for the mean waiting time in the M/G/1 queue. In this expository paper, we give intuitive arguments that "explain" why the renewal-theory paradox is ubiquitous in queueing theory, and why it sometimes produces anomalous results. In particular, we use these intuitive arguments to explain decomposition in vacation models, and to derive formulas that describe some recently-discovered counterintuitive results for polling models, such as the reduction of waiting times as a consequence of forcing the server to set up even when no work is waiting.

Key words: Inspection Paradox, Renewal-Theory Paradox, Waiting-Time Paradox, M/G/1 Queues, Vacation Models, Polling Models, Waiting Times, Decomposition.

AMS subject classifications: 60K25, 60K30, 90B22.

[1]Research supported in part by the National Science Foundation under grants DMI-9500216, 9500040, 9500471.

Printed in the U.S.A. ©1998 by North Atlantic Science Publishing Company

1. Introduction

Queueing theory was born in the early 1900s, when A.K. Erlang first addressed the questions raised by telephone engineers who were trying to understand the effects of the randomness of telephone traffic. In 1960, Professor Syski published his classic book, *Introduction to Congestion Theory in Telephone Systems* [28], which summarized in its 742 pages (just about) everything that had been done in the field of teletraffic (and queueing) theory. The book is a masterful exposition of an area in which sophisticated mathematics and real applications intersect; and it has had a profound effect on its field, and on those of us who work in it. In this paper, we reflect on some of the surprising insights into reality provided by the mathematical analysis of randomness, as exemplified by Professor Syski's expository and theoretical contributions over the last 40 years.

Queueing theory is interesting because it is based on a simple model of a simple reality (customers arrive at random, wait if necessary for the server to become available, hold the server for a random length of time, then leave) that exposes some surprising behavior. For example, it is common sense that if the arrival rate increases, then so will the waiting time. But it is the antithesis of common sense to assert that waiting times can be arbitrarily large even when the utilization of the server (the fraction of time that the server is busy) is arbitrarily small (assuming, of course, that the server never sits idle when there is work waiting to be done). But this is easily shown to be true: Consider the waiting time W in the ordinary M/G/1 queue. The celebrated Pollaczek-Khintchine formula relates the mean waiting time to the basic parameters of the model (the arrival rate λ of the Poisson arrival process, and the mean and variance of the service time S) as follows:

$$E(W) = \frac{\rho}{1-\rho}\frac{1}{2}\left(E(S) + \frac{V(S)}{E(S)}\right), \tag{1}$$

where

$$\rho = \begin{cases} \lambda E(S) & \text{if } \lambda E(S) < 1 \\ 1 & \text{if } \lambda E(S) \geq 1 \end{cases} \tag{2}$$

and $E(\cdot)$ and $V(\cdot)$ denote the mean and variance, respectively.

The quantity ρ defined in (2) equals the utilization of the server. As (1) clearly shows, for any positive value of ρ, no matter how small, $E(W)$ can be arbitrarily large (because the variance-to-mean ratio $V(S)/E(S)$, often called the "index of dispersion" in statistics, on the right-hand side of (1) can arbitrarily large). This is a truly remarkable and counterintuitive insight.

Furthermore, the probability that a customer will have to wait at all is $P(W > 0) = \rho$ (the fraction of customers who find the server busy equals the fraction of time that the server is busy, by PASTA; see Wolff [35]), which is *insensitive* to the particular form of the service-time distribution function. Thus, of two fundamental performance measures for M/G/1, one of them (How often do customers wait?) does not depend in any way on the variability of the service times, whereas the other measure (How long do the customers wait?) is quite sensitive to this variability.

The culprit here is, of course, the term

$$E(Y) = \frac{1}{2}\left(E(S) + \frac{V(S)}{E(S)}\right) = \frac{E(S^2)}{2E(S)}, \tag{3}$$

which is the mean value of the time interval Y from a random interruption of a renewal process whose interrenewal intervals are iid random variables S_1, S_2, ... until the end of the interrupted interval. The fact that $E(Y) > E(S)/2$ (unless the Ss are constant) is the famous renewal-theory (length-biasing) paradox, or *waiting-time paradox*, which has been widely used to demonstrate that "surprising phenomena can occur in waiting time problems" (Takács [29, p. 10]); likewise, it has caused some grief among researchers who overlooked its presence.

In this paper, we discuss some recently-discovered counterintuitive behavior in polling models (in which a single server switches from queue to queue, "polling" each queue to see whether there are customers waiting), and we argue that their strange behaviors are further examples of the hidden effects of the waiting-time paradox. In particular, we discuss the fact that in certain polling models, forcing the server to remain idle even when there is work waiting to be done can actually *decrease* the waiting times. We describe this phenomenon and we argue that it is, in fact, a manifestation of the waiting-time paradox. Along the way, we try to give some insight into why (3) is ubiquitous in queueing theory, and why it has been so troublesome. Since this is an "explanation" paper, we will not give mathematical details nor worry about mathematical precision in the presentation.

2. The Renewal-Theory (Waiting Time) Paradox

Let $S_1, S_2, \ldots$ be a sequence of iid interevent times of a renewal process, and let T be an arbitrary random point that "picks out" (that is, *interrupts*, or is *covered by*) one of the Ss. More formally, let $n(T)$ be the number of the Ss that are contained in the interval $[0, T)$; then, $I \equiv S_{n(T)+1}$ is the particular S that is "selected" by this sampling procedure. Assume for the moment that the Ss are discrete. Then, the probability that the variable I is of length t is approximately (assuming that T is "large enough" so that, for all practical purposes, the origin is infinitely distant from T) given by

$$P(I = t) \approx \frac{tn_t}{\sum_{i=1}^{n(T)+1} S_i},$$

where n_t denotes the number of $S_1, S_2, \ldots, S_{n(T)+1}$ that are of length t. This expression shows, for $n(T)$ "large enough",

$$P(I = t) = \frac{t(n(T)+1)P(S = t)}{(n(T)+1)E(S)} = \frac{tP(S = t)}{E(S)}; \tag{4}$$

that is, the likelihood of sampling an I of length t is proportionally biased by t.

Of course, this "derivation" of (4) is a casuistry, but the physical argument can be formalized to provide a rigorous proof; and it is easy to see that (4) remains correct when we pass from discrete time to continuous time. That is, in general,

$$dF_I(t) = \frac{t\,dF_S(t)}{E(S)}. \tag{5}$$

Now, taking expectations in (5) yields

$$E(I) = \frac{E(S^2)}{E(S)} = E(S) + \frac{V(S)}{E(S)}, \tag{6}$$

and (3) follows from symmetry, because $E(Y) = E(I)/2$. This is standard fare in

renewal theory (see, e.g., Heyman and Sobel [19], Ross [24], and Wolff [36, 37].

An interesting feature of (6) (and (3)) is that the right-hand side is not necessarily an increasing function of $E(S)$; that is, it is possible for the ratio $E(S^2)/E(S)$ to have a negative "slope". For example, suppose $S = X + u$, where u is a nonnegative constant and X is a nonnegative random variable that does not depend on the value of u. Then

$$E(S) + \frac{V(S)}{E(S)} = E(X) + u + \frac{V(X)}{E(X) + u}$$

is convex in u and will have a minimum at $u = u^* = \sqrt{V(X)} - E(X)$ if $\sqrt{V(X)} \geq E(X)$ (if $\sqrt{V(X)} < E(X)$, let $u^* = 0$). That is, suppose that an arbitrary customer (the "tagged" customer) interrupts a renewal process $\{S_1, S_2, \ldots\}$ and waits a time from his arrival epoch until the end of the interrupted interval (think of T as the tagged customer's arrival epoch, and Y as his waiting time); and suppose that $S = X + u$ (as described above), where X is "more variable than exponential" (that is, X has a coefficient of variation $\sqrt{V(X)}/E(X)$ that is greater than unity). Then his mean waiting time $E(Y)$ will *decrease* as u (and therefore as $E(S)$) is increased from zero, reaching a minimum when $u = u^* = \sqrt{V(X)} - E(X)$.

Now suppose that customers arrive according to a Poisson process, and let A_S be the number of arrivals during an interval of length S. Consider a tagged customer whose arrival epoch T "picks out" one of the Ss that comprise the renewal process, and let A_I be the number of arrivals (in addition to the tagged customer) that arrive during the I that contains T. Then $E(A_I)/2$ is the mean number of customers who arrive during the tagged customer's I but before (and after) T. (That is, the tagged customer belongs to a batch of size $A_I + 1$ and, on average, $E(A_I)/2$ of those customers have preceded him and $E(A_I)/2$ followed him.) Thus,

$$E(A_I) = \int_0^\infty \lambda t \, dF_I(t) = \frac{\lambda E(S^2)}{E(S)},$$

where the second equality follows from (5).

Also, straightforward calculations give $E(A_S) = \lambda E(S)$ and

$$E(A_S^2) = \int_0^\infty E(A_S^2 \mid S = t) dF_S(t) = \int_0^\infty E(A_t^2) dF_S(t) = \lambda E(S) + \lambda^2 E(S^2);$$

and combination of these equations gives, for the mean number of customers preceding (following) the tagged customer in his batch,

$$\frac{E(A_I)}{2} = \frac{E[A_S(A_S - 1)]}{2E(A_S)}. \tag{7}$$

As pointed out by Burke [3], several eminent probabilists (and, likely, many less-eminent probabilists as well) have published incorrect results because they mistakenly ignored the size-bias of a batch that contains a particular customer. This is an easy mistake to make, because in this case the error-detecting feature of queueing theory fails. That is, if one implicitly makes an incorrect assumption, then the subsequent analysis usually produces a result that is clearly wrong on its face (like a negative probability); but the oversight alluded to here often produces a "reasonable", but wrong, answer.

This phenomenon is also discussed by Whitt [34], who cites his earlier paper Whitt

[33], where "it is shown for a large class of queueing systems in which customers arrive in batches that the delay distribution of the last customer in a batch to enter service is a function of the batch-size distribution whereas the delay distribution of an arbitrary customer is the same function of the associated batch-size stationary-excess distribution."

An apparently similar phenomenon appears in a computer-science analysis of hash-structured files by Cooper and Solomon [9]. (This paper was written in response to an invitation by Professor Syski as a "Contribution to the Special Issue on Teletraffic Theory and Engineering in Memory of Félix Pollaczek (1892-1981).") Here the quantities of interest are the length of a *successful search* (the number of accesses required to locate a record in a "chain" of records) and the length of an *unsuccessful search* (the number of accesses required to search to the end of the chain and affirm that the record sought is not on the chain). Numerical calculations made from formulas that describe these values showed that, in some cases, "the average length of an unsuccessful search which [goes to] the end of the chain is less than the average length of a successful search which goes only to the 'middle' of the chain .... Apparently, this is an example of 'batch-biasing,' a discrete version of the famous waiting-time (or inspection) paradox, so dear to the hearts of queueing theorists."

Returning to the continuous-time case, one might wonder why $E(W)$ in (1) can be written simply

$$E(W) = \frac{\rho}{1-\rho}E(Y), \tag{8}$$

where Y can now be interpreted as the amount of service remaining when a service time (in a renewal process whose intervals are service times) is interrupted by an arriving customer. Every arrival epoch occurs either when the server is idle (in which case $W = 0$ for the arriving customer) or busy; thus, every customer who waits "picks out" a "long" service time, so it is not surprising that the effect of the renewal-theory paradox (clearly, the server idle-times are irrelevant) is manifested in the formula for the mean waiting time. But why so simply?

Consider the M/G/1 queue with the queue discipline LIFO Preemptive-Resume. (Each arrival begins service immediately, bumping the customer in service, if any, back to the head of the queue. When a customer who has been preempted eventually resumes service, his service continues from where it left off when it was last preempted.) Since all we know about a waiting customer is that his service time has been interrupted at least once, it is intuitively clear that each customer in the queue must have the same distribution of remaining service time, independent of the number of customers in the queue; and moreover, it would be surprising if this common distribution of remaining service time were not given by the forward-recurrence time Y of the service times.

Now, the total amount of remaining service time is the amount of time that the arriving customer would have to wait if the queue discipline were FIFO (instead of LIFO Preemptive-Resume). Then, (8) will follow if we can show that $\rho/(1-\rho)$ is the mean number of customers present at an arrival epoch in the M/G/1 LIFO Preemptive-Resume queue.

To this end, let π_j be the equilibrium probability that an arriving customer finds j other customers present when he arrives, and let P_j be the corresponding equilibrium probability for an arbitrary time epoch. Then, equating the rate (transitions per unit time) up from any state to the rate down from the state above it, we have

$$\lambda\pi_{j-1} = \mu_j P_j \qquad (j = 1, 2, \ldots) \tag{9}$$

where λ is the (Poisson) arrival rate and μ_j is the "aggregate" service completion rate when there are j customers present. But, by our previous observation that each customer present at an arbitrary point in time has the same distribution of remaining service time regardless of queue length, it follows that $\mu_j = \mu$ (independent of j). Also, since the server can never be idle when there is work waiting, therefore, $P_0 = 1 - \rho$. Finally, invoking PASTA (that is, $\pi_j = P_j$), it is easy to see that (9) defines a geometric distribution, whose mean is $\rho/(1-\rho)$; and (8) is "proved". (This argument provides a basis for a derivation of the famous formula of Beneš [1] for the distribution of waiting times in the M/G/1 FIFO queue. See Kelly [20] and Niu [23] for rigorous proofs, and Cooper and Niu [5] for an intuitive argument similar to the one given here.) Observe also that this argument shows, in passing, that the state probabilities $\{P_j\}$ for the M/G/1 LIFO Preemptive-Resume queue are *insensitive* to the form of the service-time distribution; thus, arguably, this is a fundamental queue discipline, our democratic preference for FIFO notwithstanding.

Based on an intuitive understanding of the renewal-theory paradox, we have "derived" (8). Now, using (8) together with the elementary fact that $E(W) = \rho E(W \mid W > 0)$, it is easy to verify that

$$E(W \mid W > 0) = E(W) + E(Y). \tag{10}$$

Equation (10) is a remarkable decomposition. It says that the mean waiting time for customers who do not begin service immediately at their arrival epoch is the sum of the overall mean waiting time $E(W)$, which includes the zero waiting time of the first customer served in a busy period, and an amount $E(Y)$ that equals "half" of the mean of the length-biased service time of this customer. (Equation (10), in fact, can be extended to GI/G/1; see Li and Niu [22], Proposition 1.)

3. Vacation Models

Now consider the M/G/1 (*multiple-*) *vacation model* with *exhaustive service*. In this model, the server continues to serve until there is no more work to be done (exhaustive service), and then it becomes unavailable for service (goes on "vacation") for a random length of time, after which it returns to see whether there are any customers waiting; if so, then it resumes working, otherwise it takes another vacation, and so on.

Every arrival epoch must occur either during a service time or during a vacation, so it is reasonable to expect that the effects of the renewal-theory paradox will appear twice; each customer arrives during either a "long" service time or a "long" vacation. But surprisingly, this double length-biasing effect manifests itself in a very simple way. Let W_V be the waiting time of an arbitrary customer in this vacation model. Then, as is well known,

$$E(W_V) = \frac{\rho}{1-\rho} E(Y_S) + E(Y_V), \tag{11}$$

where Y_S is the forward-recurrence time of the renewal process constructed from the service times, and Y_V is defined similarly with respect to the vacation times. (Significantly, the vacations need *not* be mutually independent for (11) to be valid.)

From (11) we see that the "interesting feature" observed in the discussion following (6) applies here: Because of the term $E(Y_V)$ on the right-hand side of (11), it is possible that longer vacations can lead to shorter waiting times. In particular, as the

example discussed here shows, it is possible that *the mean waiting time in a vacation model can be decreased by extending every vacation by a constant amount.* (Note that the first term on the right-hand side of (11) does not produce a similar effect, because that effect is "canceled" by the factor $\rho = \lambda E(S)$.)

Observe that in the vacation model, $P(W_V > 0) = 1$, so $E(W_V) = E(W_V \mid W_V > 0)$. In light of this observation and equation (8), equation (11) can be written

$$E(W_V \mid W_V > 0) = E(W_0) + E(Y_V), \tag{12}$$

where W_0 is the waiting time in the corresponding M/G/1 queue *without* vacations. With the present notation, equation (10) can be written

$$E(W_0 \mid W_0 > 0) = E(W_0) + E(Y_S). \tag{13}$$

The similarity between (12) and (13) is striking.

In ordinary M/G/1 queues, the effect of the unavailability of the server to the first blocked customer of the busy period contributes an amount $E(Y_S)$ to the mean waiting time of the blocked customers. Apparently, in the vacation counterpart, the same role is played by the vacation time during which the first blocked customer of a busy period arrives. That is, the *vacation time* that is interrupted by the first blocked customer of a busy period in a vacation model plays the same role as the *service time* that is interrupted by the first blocked customer of a busy period in the corresponding ordinary (no vacations) M/G/1 queue. With hindsight, this is obvious, and the waiting-time decomposition(11) (or, equivalently, (12)) is now "explained." (For rigorous proofs of the decompositions (12) and (13), see Doshi [10] and Li and Niu [22]; these results, in fact, extend to GI/G/1.)

Equation (11) is an example of the well-known decomposition phenomenon of some vacation models. Under certain conditions (weaker than those assumed here), the FIFO waiting time in the vacation model is distributed as the sum of two independent random variables: one random variable is the waiting time in the corresponding ordinary M/G/1 queue (i.e., without vacations), and the other random variable relates only to the characteristics of the vacations; and the successive vacations need not be independent. Indeed, other quantities, such as workload or total number of customers, sometimes enjoy similar decompositions. These properties make vacation models useful adjuncts in the analysis of polling models, to which we will return shortly. The fact of decomposition in the case of exhaustive service was originally uncovered by Skinner [26] and Cooper [4]. The fact that the vacation term in this decomposition is a forward-recurrence time was first recognized by Levy and Yechiali [21]. The more general conditions under which decompositions occur in the M/G/1 queue were given in Fuhrmann and Cooper [17]. Extensions from waiting times to workloads were given by Boxma and Groenendijk [2]. A comprehensive survey of vacation models is given in Doshi [11], and detailed textbook treatments are given in Takagi [31] and Wolff [37].

4. Polling Models

In a *polling model*, N queues are served by a single server that travels from queue to queue in a prescribed sequence, "polling" each queue to determine whether to provide service to the polled queue. We will restrict our attention to the classical polling model in which the server polls the queues in cyclic order, and continues to serve a

queue until it is empty (exhaustive service). The *switchover time* is the time it takes for the server to travel from one queue and poll the next one; and the *setup time* is the time it takes for the server to prepare itself (set up) before beginning to serve a polled queue. In isolation, each queue would be an ordinary M/G/1 queue; but linked together through the sharing of a single server, they interact in complicated ways, and their performance measures are strongly dependent on each other.

Polling models have important applications in telecommunications (where switchover times are important) and manufacturing (where setup times are important), and consequently a huge literature has grown up since the mid 1960s, when polling models were first studied. (See Takagi [30, 32] for comprehensive surveys.) Recent interest has centered on the effects of the "dead time" (switchover and setup times) during which the server is not serving the customers. From the viewpoint of a particular queue, when the server is away serving the other queues (or switching, or setting up), the server can be imagined to be on "vacation"; so vacation models are useful adjuncts in the study of polling models.

Counterintuitive behavior (like a decrease in waiting times with an increase in switchover or setup times) has recently attracted much attention. Here we argue that this paradoxical behavior is another manifestation of the renewal-theory paradox, and we discuss it in the context of decomposition in vacation models.

In a much-discussed paper, Sarkar and Zangwill [25] gave numerical examples that showed increases in mean waiting times when the mean switchover times were decreased. (In their model, setup and switchover times are in effect the same. We will refer to the dead times as switchover times unless it is necessary to distinguish between switchover times and setup times.) They correctly associated this phenomenon with the renewal-theory paradox: "These examples are another manifestation of the aforementioned 'inspection paradox' or 'random incidence' from renewal theory. By reducing average setup or processing times we make the expected lengths of ordinary cycles smaller. If the variances are not reduced proportionately, however, there exist cycles of longer length. But the probability that a random arrival will fall in this large interval depends on the ratio of larger to smaller intervals. Thus, by making the lengths of ordinary cycles smaller, we increase the probability of arrival into a larger cycle. As a result, arrivals wait longer, and this explains the nonintuitive outcomes of the examples."

Here, we summarize some results from subsequent work of ours that elucidates this observation. In Cooper, Niu, and Srinivasan [6] we showed that mean waiting times in some polling models exhibit a decomposition with respect to switchover times that is similar to (but not exactly the same as) the decomposition (11) in M/G/1 vacation models. Specifically, we showed that, for the N-queue cyclic-service polling model with exhaustive service, the mean waiting time $E(W_i)$ for an arbitrary customer in queue i ($i = 1,\ldots,N$) is given by

$$E(W_i) = E(W_i^0) + \frac{r}{2}\frac{1-\rho_i}{1-\rho}, \tag{14}$$

where W_i^0 is the waiting time in a "corresponding" zero-switchover-times model (the server stops traveling whenever the system becomes empty, and advances instantaneously to the queue at which the next arrival occurs), ρ_i is the server utilization at queue i, $\rho = \rho_1 + \ldots + \rho_N$ is the total server utilization, and r is the total mean switchover time per cycle. By definition, the "corresponding" zero-switchover-times model is the same as the original model, except that the switchover times are zero and the second moment $x_i^{(2)}$ of the service times at queue i is given by

$$x_i^{(2)} = b_i^{(2)} + \delta_{i-1}^2 \frac{1-\rho}{\lambda_i r}, \tag{15}$$

where λ_i and $b_i^{(2)}$ are, respectively, the arrival rate and second moment of the service times at queue i, and δ_{i-1}^2 is the variance of the original model's switchover times as the server switches from queue $i-1$ to queue i. Observe that if the switchover times are constant, then $\delta_{i-1}^2 = 0$ and the corresponding zero-switchover-times model has the same parameters as the original model. This version of the theorem (i.e., for constant switchover times) was first proved by Fuhrmann [15]. But the variability of the switchover times is the crucial factor underlying the anomalous behavior we are discussing here.

The theorem embodied in (14) and (15) says that the expected waiting time in the general-switchover-times model (the model with nonzero switchover times) decomposes into a sum of two terms; one term is the mean waiting time in the "corresponding" zero-switchover-times model, and the other is a simple term that depends only on the server utilizations and the sum (but not the individual values) of the mean switchover times. (See also Srinivasan, Niu, and Cooper [27], where (14) is extended to a relation between the waiting-time distributions.) This is reminiscent of the ordinary M/G/1 vacation decomposition, but the polling model does not satisfy the sufficient conditions for decomposition given in Fuhrmann and Cooper [17]. Our proof of (14) and (15) is straightforward (based on the equations given by Ferguson and Aminetzah [13] for the general-switchover-times model), but no simple, intuitive "explanation" (like the "derivation" of the vacation decomposition (13) by analogy with (12)) has been forthcoming.

In Cooper, Niu, and Srinivasan [7], we applied this theorem to the *symmetric* polling model, in which, by definition, all queues have the same arrival rate, the same service-time distribution, and the same switchover-time distribution. Then, clearly, the "corresponding" zero-switchover-times polling model is simply an ordinary M/G/1 queue. Hence, $E(W_i^0) = E(W^0)$ is given by the Pollaczek-Khintchine formula (1),

$$E(W^0) = \frac{\rho}{1-\rho} \frac{E(S^2)}{2E(S)},$$

where $E(S^2) = x^{(2)}$ given by (15); and after some trivial algebraic simplification, (14) becomes

$$E(W) = \frac{\rho}{1-\rho} \frac{b^{(2)}}{2b} + \frac{1}{2}\left(\frac{1-\rho/N}{1-\rho} E(R) + \frac{V(R)}{E(R)}\right), \tag{16}$$

where b is the mean service time and R denotes a switchover time.

Observe that the second term on the right-hand side of (16) is "almost" equal to the mean forward-recurrence time of the renewal process constructed from the switchover times; that is, it assumes essentially the same form as (3) (and (6)). Indeed, if $N = 1$ then (16) is precisely the vacation-decomposition result (11), with the switchover times being the vacations. Therefore, the symmetric polling model behaves essentially like a vacation model, and shares with it the "interesting feature" discussed following (6). As explained earlier, it is possible, because of the renewal-theory paradox, that longer vacations in a vacation model can lead to shorter waiting times. In a polling model, this anomaly translates into: *longer switchover times can lead to shorter waiting times*; the essential factor is the variance-to-mean ratio of the switchover times, and the variance of the service times is irrelevant.

Thus, the counterintuitive effect observed by Sarkar and Zangwill can indeed be attributed to the renewal-theory paradox: Customers are more likely to arrive during

long setup (switchover) times than during short ones; and reducing the variance-to-mean ratio of the setup times (by, for example, increasing each setup time by a constant value) can produce a reduction in mean waiting time (also see Fuhrmann [15, 16] for a discussion of the length-biasing effect in polling models). And conversely, cutting setup times (which, according to Sarkar and Zangwill is a commonsense goal of manufacturing engineers) may cause the work-in-process to increase. Although our argument applies, strictly speaking, only to the symmetric case, it is clear that, qualitatively, the same forces are at work in the general case, and hence similar counterintuitive results will be observed (as Sarkar and Zangwill did, in fact, observe in other, nonsymmetric examples).

As we have indicated, in the context of manufacturing there is much interest in systems that can be described by polling models with setup times. It is just commonsense that it is a waste of time to set up at a queue if no work is waiting at that queue. But polling models in which the decision to set up depends on the state of the queue being polled (State-Dependent setups, SD) are much more difficult to analyze mathematically than their counterparts in which setups are always performed, whether or not work is waiting at the polled queue (State-Independent setups, SI). For this reason, and consistent with the commonsense view that "empty" setups are wasted setups, the SI model has been proposed to provide upper bounds for the waiting times in the corresponding SD model (Ferguson [12]).

But, one might argue, if decreasing the setup times in an SI model can increase the waiting times, perhaps eliminating altogether some setup times (when there is no work waiting in the polled queue) might have a similar deleterious effect. In fact, Gupta and Srinivasan [18] showed numerically, via exact analysis, that the SI model does *not* necessarily provide an upper bound for the mean waiting time in its corresponding SD counterpart. In Cooper, Niu, and Srinivasan [8], we investigated this question further, and for symmetric polling models with $N = 2$, we proved a surprisingly simple comparison theorem, given as equation (18) below, that characterizes the difference in the mean waiting times in the SD and the SI models.

Denote the dead time (defined earlier as the switchover and setup times during which the server is not serving any customers) between two successive queues as V, and assume that V is the sum of a switchover time R and a possibly-absent setup time Z (we assume that both R and Z are not identically zero), i.e.,

$$V = R + \delta Z \tag{17}$$

where δ is the indicator function of the event that the server sets up when it polls a queue (at the expiration of the preceding switchover time). Thus, in the SD model, $\delta = 1$ whenever at least one customer is waiting in the queue when it is polled; whereas in the SI model, we always have $\delta = 1$. By assumption, Z is independent of both R and δ; but clearly, in the SD model, R and δ are dependent.

Let $\widehat{R}$ be a switchover time during which (i.e., given that) no customers arrive at the queue to which the server is switching. It is easily seen that $E(\widehat{R}) = E(\lambda R e^{-\lambda R})/E(e^{-\lambda R})$, where λ is the arrival rate at each queue. Let W^D denote the waiting time in the two-queue symmetric SD polling model, and let W^I be its counterpart in the corresponding SI model. Then,

$$E(W^D) - E(W^I) = c\left\{[E(R) + E(Z)]\left[1 - \frac{E(\widehat{R})}{E(R)}\right] + \left[\frac{V(R)}{E(R)} - \frac{V(Z)}{E(Z)}\right]\right\}, \tag{18}$$

where c is a known positive number.

Equation (18) completely characterizes the sign of $E(W^D) - E(W^I)$ in terms of only the distribution of R and the first two moments of Z. In particular, it has the following easily-verified consequences: (*i*) If the switchover times are constant and the setup times are constant, then $E(W^D) = E(W^I)$; (*ii*) if the switchover times are variable and the setup times are constant, then $E(W^D) > E(W^I)$; and (*iii*) if the switchover times are constant and the setup times are variable, then $E(W^D) < E(W^I)$. It is remarkable that these consequences of (18) depend only on whether the switchover times and setup times are constant or variable, and do not depend on the server utilization or the service-time distribution. Furthermore, only statement (*iii*) agrees with commonsense intuition.

The explicit formula (18), together with statements (*i*)-(*iii*) above, clearly demonstrates that the relative variance-to-mean ratios of the switchover times and the setup times, that is, the difference

$$\frac{V(R)}{E(R)} - \frac{V(Z)}{E(Z)}$$

is the primary determinant of the sign of $E(W^D) - E(W^I)$. In other words, the culprit behind this counterintuitive behavior is, again, the renewal-theory paradox.

These results were derived for the special case $N = 2$. We conjecture that similar qualitative statements can be made for the more-general polling model with $N > 2$ queues (but it seems unlikely that a formula as "clean" as (18) will result).

Our (rigorous) proof of (18) is quite detailed, which obscures the underlying role of the renewal-theory paradox; therefore, in keeping with the spirit of this paper, we conclude with a derivation of (18) via a direct, intuitive argument based on the selection-bias effects of the renewal-theory paradox.

Consider first the SD model. Let L^D be the *total* (i.e., both queues included) number of customers left behind by a (randomly-selected) departing customer in this model. It is well known that $E(L^D)$ also equals the time-average total number of customers in the system. Therefore, from Little's law,

$$E(W^D) = \frac{E(L^D)}{2\lambda}. \tag{19}$$

Now, define (following Fuhrmann [14]) *each* dead time as a vacation, and take the viewpoint of a randomly-selected customer *who arrives during a vacation*. Call this customer the "tagged customer," and let K^D be the total number of customers in the SD model as seen by tagged customer. Then, from Proposition 3 of Fuhrmann and Cooper [17], we have (with $=^d$ denoting equality in distribution)

$$L^D =^d K^D + L^M, \tag{20}$$

where L^M is independent of K^D and is distributed as the total number of customers left behind by a departing customer in a (corresponding) standard M/G/1 queue. Since $E(L^M)$ is well known, our remaining task is to compute $E(K^D)$. To this end, we will use a selection-bias argument.

Let T^D be the total number of waiting customers at both queues at a server-departure epoch, equally likely to be from either queue; and let A_R and A_Z be the total number of Poisson arrivals (at rate 2λ) during an R and a Z, respectively. We now argue that

$$E(K^D) = \frac{E(A_R)}{E(A_R) + E(\delta)E(A_Z)}\left\{E(T^D) + \frac{E[A_R(A_R-1)]}{2E(A_R)}\right\} \tag{21}$$

$$+\frac{E(\delta)E(A_Z)}{E(A_R)+E(\delta)E(A_Z)}\left\{E(T^D \mid \delta=1)+E(A_R \mid \delta=1)+\frac{E[A_Z(A_Z-1)]}{2E(A_Z)}\right\}.$$

Observe that K^D can be split into the sum of (i) the total number of waiting customers in the system at the server-departure epoch that precedes the tagged customer's arrival epoch (i.e., the preceding T^D), and (ii) the number of customers who arrive to both queues prior to the tagged customer's arrival epoch but within the same vacation. To compute $E(K^D)$, the expected counts in (i) and (ii) need to be corrected for selection bias. We consider two cases, corresponding to whether the tagged customer arrives in the "R portion" or (whenever applicable) in the "Z portion" of a vacation:

Case 1: Clearly, the probability for the tagged customer to arrive during an R is given by

$$\frac{E(A_R)}{E(A_R+\delta A_Z)}=\frac{E(A_R)}{E(A_R)+E(\delta)E(A_Z)}.$$

From the viewpoint of the tagged customer, since T^D and R are independent, there is no length bias for the customer count T^D at the preceding server-departure epoch; this leads to the first term $E(T^D)$ in the first pair of braces. There is, however, a length bias for the number of customers who arrived in R prior to the tagged customer; this length-biased expected count is, from (7), given by $E[A_R(A_R-1)]/[2E(A_R)]$. Putting these observations together yields the first term on the right-hand side of (21).

Case 2: The probability for the tagged customer to arrive during a Z is given by

$$\frac{E(\delta A_Z)}{E(A_R+\delta A_Z)}=\frac{E(\delta)E(A_Z)}{E(A_R)+E(\delta)E(A_Z)}.$$

From the viewpoint of the tagged customer, the total count in (ii) can be further split into the sum of the number A_R of customers who arrived during the preceding R, and the number of customers who arrived prior to the tagged customer's arrival epoch but within the Z selected by the tagged customer. Observe that the tagged customer can land in a Z only if $\delta=1$ in a vacation in progress; and that $\delta=1$ in a vacation if and only if $T^D+\widehat{A}_R>0$, where $\widehat{A}_R$ denotes the number of arrivals at the *next* queue during the R portion of that vacation. Therefore, both T^D and A_R (customers counted in $\widehat{A}_R$ constitute a subset of those counted in A_R) are biased by the event that $\delta=1$; this leads to the terms $E(T^D \mid \delta=1)$ and $E(A_R \mid \delta=1)$ in the second pair of braces in (21). Finally, the length-biased expected number of customers who arrived in Z prior to the tagged customer is, again, given by $E[A_Z(A_Z-1)]/[2E(A_Z)]$. These observations together lead to the second term on the right-hand side of (21).

Starting from the relations (19), (20), and (21), it is now straightforward to derive an explicit formula for $E(W^D)$. Moreover, our derivation can also be repeated, almost verbatim, to produce a parallel formula for $E(W^I)$ in the SI model. Combining these two formulas then leads, after some algebra, to equation (18). (See Cooper, Niu, and Srinivasan [8] for these remaining details).

5. Conclusion

Using intuitive arguments, we have tried to explain why the length-biasing property of renewal theory is ubiquitous in queueing theory, and why some recently-uncovered

surprising behavior in polling models can be attributed to its effects. Our reflections in this festschrift paper for Professor Syski are part of a larger appreciation of the power of mathematical reasoning in general, and queueing theory in particular, to uncover and describe commonplace phenomena whose very existence is not at first clear and which, once discovered, seem to defy common sense.

Acknowledgment

We thank Hideaki Takagi for his thoughtful and constructive comments on the manuscript.

References

[1] Beneš, V.E., On queues with Poisson arrivals, *The Annals of Mathematical Statistics* **28** (1957), 670-677.

[2] Boxma, O.J. and Groenendijk, W.P., Pseudo-conservation laws in cyclic-service systems, *J. of Applied Probability* **24** (1987), 949-964.

[3] Burke, P.J., Delays in single-server queues with batch input, *Operations Research* **23** (1975), 830-833.

[4] Cooper, R.B., Queues served in cyclic order: Waiting times, *Bell System Tech. J.* **49** (1970), 399-413.

[5] Cooper, R.B. and Niu, S.-C., Beneš's formula for M/G/1-FIFO 'explained' by preemptive-resume LIFO, *J. of Applied Probability* **23** (1986), 550-554.

[6] Cooper, R.B., Niu, S.-C. and Srinivasan, M.M., A decomposition theorem for polling models: The switchover times are effectively additive, *Operations Research* **44** (1996), 629-633.

[7] Cooper, R.B., Niu, S.-C. and Srinivasan, M.M., When does forced idle time improve performance in polling models?, *Management Science* (1998), to appear.

[8] Cooper, R.B., Niu, S.-C. and Srinivasan, M.M., Setups in polling models: Does it make sense to set up if no work is waiting?, (1997), submitted.

[9] Cooper, R.B. and Solomon, M.K., Teletraffic theory applied to the analysis of hash-structured files, *Intern. J. of Electronics and Commun.* **47** (1993), 336-341.

[10] Doshi, B.T., A note on stochastic decomposition in a GI/G/1 queue with vacations or set-up times, *J. of Applied Prob.* **22** (1985), 419-428.

[11] Doshi, B.T., Single-server queues with vacations, *Stoch. Anal. of Comp. and Commun. Systems*, (ed. by H. Takagi), Elsevier, North-Holland (1990).

[12] Ferguson, M.J., Mean waiting times for a token ring with station-dependent overheads, In: *Local Area and Multiple Access Networks* (ed. by R.L. Pickholtz), Computer Science Press (1986), 43-67.

[13] Ferguson, M.J. and Aminetzah, Y.J., Exact results for nonsymmetric token ring systems, *IEEE Trans. on Commun.* **COM-33** (1985), 223-231.

[14] Fuhrmann, S.W., Symmetric queues served in cyclic order, *Operations Research Letters* **4** (1985), 139-144.

[15] Fuhrmann, S.W., A decomposition result for a class of polling models, *Queueing Systems, Theory and Appl.* **11** (1992), 109-120.

[16] Fuhrmann, S.W., On approximating mean waiting times in polling models with half of the mean cycle times, *Teletraffic Cont. for the Info. Age*, Proceedings,

ITC 15 (ed. by V. Ramaswami and P.E. Wirth) **Vol. 2a**, Elsevier (1997), 265-273.

[17] Fuhrmann, S.W. and Cooper, R.B., Stochastic decompositions in the M/G/1 queue with generalized vacations, *Operations Research* **33** (1985), 1117-1129.

[18] Gupta, D. and Srinivasan, M.M., Polling systems with state-dependent setup times, *Queueing Systems, Theory and Appl.* **22** (1996), 403-423.

[19] Heyman, D.P. and Sobel, M.J., *Stochastic Models in Operations Research, Vol. 1*, Mc-Graw Hill, New York 1982.

[20] Kelly, F.P., *Reversibility and Stochastic Networks*, Wiley, Chicester 1979.

[21] Levy, Y. and Yechiali, U., Utilization of idle time in an M/G/1 queueing system, *Management Science* **22** (1975), 202-211.

[22] Li, J. and Niu, S.-C., The waiting-time distribution for the GI/G/1 queue under the *D*-policy, *Prob. in the Eng. and Info. Sciences* **6** (1992), 287-308.

[23] Niu, S.-C., Representing workloads in GI/G/1 queues through the preemptive-resume LIFO queue discipline, *Queueing Systems, Theory and Appl.* **3** (1988), 157-178.

[24] Ross, S.M., *Stochastic Processes*, Wiley, New York 1983.

[25] Sarkar, D. and Zangwill, W.I., Variance effects in cyclic production systems, *Management Science* **37** (1991), 443-453.

[26] Skinner, C.E., A priority queueing system with server-walking time, *Operations Research* **15** (1967), 278-285.

[27] Srinivasan, M.M., Niu, S.-C. and Cooper, R.B., Relating polling models with zero and nonzero switchover times, *Queueing Systems, Theory and Appl.* **19** (1995), 149-168.

[28] Syski, R., *Introduction to Congestion Theory in Telephone Systems* (1960), Oliver and Boyd, Edinburgh, Second edition, North-Holland, Amsterdam 1986.

[29] Takács, L., *Introduction to the Theory of Queues*, Oxford University Press, New York 1962.

[30] Takagi, H., Queueing analysis of polling models: An update, In: *Stochastic Analysis of Comp. and Commun. Systems* (ed. by H. Takagi), North-Holland, Amsterdam 1990.

[31] Takagi, H., *Queueing Analysis, Vol. 1: Vacation and Priority Systems, Part 1*, North-Holland, Amsterdam 1991.

[32] Takagi, H., Queueing analysis of polling models: Progress in 1990-1994, Chapter 5 in *Frontiers in Queueing: Models and Applications in Science and Engineering* (ed. by J.H. Dshalalow), CRC Press, Boca Raton, Florida (1997), 119-146.

[33] Whitt, W., Comparing batch delays and customer delays, *Bell System Tech. J.* **62** (1983), 2001-2009.

[34] Whitt, W., The renewal-process stationary-excess operator, *J. of Applied Prob.* **22** (1985), 156-167.

[35] Wolff, R.W., Poisson arrivals see time averages, *Operations Research* **30** (1982), 223-231.

[36] Wolff, R.W., Sample-path derivations of the excess, age, and spread distributions, *J. of Applied Prob.* **25** (1988), 432-436.

[37] Wolff, R.W., *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, Englewood Cliffs, New Jersey 1989.