COMMUN. STATIST.-STOCHASTIC MODELS, 5(3), 515-529 (1989)

AN EXTREMAL PROPERTY OF THE FIFO DISCIPLINE VIA AN ORDINAL VERSION OF $L=\lambda W$

Shlomo Halfin

Bell Communications Research
Morristown, NJ 07960

Ward Whitt

AT&T Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

We apply the relation $L=\lambda W$ to prove that in great generality the long-run average number of customers in a queueing system at an arrival epoch is equal to the long-run average number of arrivals during the period a customer spends in the system. This relation can be regarded as an ordinal version of $L=\lambda W$, arising when we measure time solely in terms of the number of arrivals that occur. We apply the ordinal version of $L=\lambda W$ to obtain a conservation law for single-server queues. We use this conservation law to establish an extremal property for the FIFO service discipline: We show that in a G/GI/1 system the FIFO discipline minimizes (maximizes) the long-run average sojourn time per customer among all work-conserving disciplines that are non-anticipating with respect to the service times (may depend on completed service times, but not remaining service times) when the service-time distribution is NBUE (NWUE). Among the disciplines in this class are round robin, processor sharing and shortest expected remaining processing time.

Key Words: queues, conservation laws, Little's law, work-conserving service disciplines, stochastic comparisons, new better than used in expectation.

1. Introduction

In this paper we establish a conservation law that can be viewed as an ordinal version of the relation $L=\lambda W$, due to Little [10] and Stidham [14]. (We briefly review $L=\lambda W$ in Section 2.) Our main result, Theorem 2, states that in great generality (the same kind of conditions as for $L=\lambda W$) *the long-run average number of customers in the system at an arrival epoch coincides with the long-run average number of arrivals during a customer's sojourn time* (the interval between arrival and departure). We may either count or not count the arriving customer, but we do the same for both averages.

For the FIFO (first-in first-out) discipline, our result is known, being any easy consequence of two facts. First, with the FIFO discipline, the customers left behind by a departing customer are just those customers that arrived during the departing customers sojourn time; see Keilson and Servi [9] for recent applications. Second, the departure-point averages coincide with the arrival-point averages, as can be seen from upcrossing and downcrossing arguments; e.g., see (9) below and p. 112 of Franken, König, Arndt and Schmidt [5]. However, we do *not* restrict attention to the FIFO discipline here.

We regard our conservation law as an ordinal version of $L=\lambda W$, because it is an analog of the standard version that arises *when we do not measure or observe time*. As with the standard version, customers are indexed in order of arrival, but we do not consider the actual arrival epochs. Similarly, customers depart some time after they arrive, but we do not consider the actual departure epochs. Instead, we measure time in terms of the arrival indices, so that the waiting time is replaced by the number of arrivals during a customer's sojourn time. In Section 3 we show that the ordinal version of $L=\lambda W$ can be derived from the standard version simply by making this interpretation.

In Section 4 we apply this ordinal version of $L=\lambda W$ to obtain a conservation law for single-server queues; see (6)-(10). In Section 5, we apply this conservation law to establish some useful comparisons between congestion

measures with different service disciplines. Theorem 3 shows that the FIFO discipline minimizes (maximizes) the expected sojourn time per customer in a G/GI/1 model, among all work-conserving disciplines that are non-anticipating with respect to the service times, when the service-time distribution is NBUE (NWUE), i.e., new better (worse) than used in expectation. In this model, the service times are i.i.d. (independent and identically distributed) and independent of the arrival process, but the arrival process can be general, e.g., non-renewal. *Work-conserving* means, first, that the server works at unit rate whenever there is work to do and, second, that the discipline does not affect the arrival times and service requirements; see p. 418 of Heyman and Sobel [8]. Since the discipline can affect the number of customers in the system, we cannot have a limited number of queue positions; the discipline could then affect the arrivals (customers that enter). On the other hand, we could have a finite capacity for work (service time), because the workload process is independent of the discipline within this class. *Non-anticipating with respect to the service times* means that the discipline can depend on the amount of completed service time of each customer, but not on the remaining service time. *All* work-conserving disciplines are of this form when the service requirements are not known upon arrival, but instead are only realized through service. Work-conserving disciplines such as round robin (RR) and processor sharing (PS) that do not depend on the service times in any way are of course included. A discipline that uses the completed service times is shortest expected remaining processing time (SERPT), where we compute the expectation conditional on the completed service time of each customer and we use processor sharing when there are ties. One might conjecture that SERPT always minimizes the long-run average sojourn time within this class of disciplines; this is so for NBUE service-time distributions, but *not* in general, even with a Poisson arrival process; see Remark 5.5.

Of course, disciplines such as last-in first-out (LIFO) and random order of service (ROS) for which the distribution of the equilibrium number of customers in the system is the same as far FIFO have the *same* expected sojourn time as FIFO (by virtue of $L = \lambda W$). A useful comparison with FIFO is

established for preemptive service-sharing disciplines such as RR, PS and SERPT. Consistent with intuition, FIFO tends to be good for low-variability service-time distributions, but bad for high-variability service-time distributions. With FIFO, more customers suffer from the exceptionally long service times. Since the exponential distribution is both NBUE and NWUE, the expected sojourn time is *independent of the discipline* in this class when the service-time distribution is exponential.

The stochastic comparisons under NBUE and NWUE conditions here extend previous results by Yamazaki and Sakasegawa for the special case of a GI/GI/1 queue (renewal arrival process) with a PS discipline having a finite number of service positions. Such systems were also analyzed by Rege and Sengupta [12] and Avi-Itzhak and Halfin [2]-[4].

2. Review of $L=\lambda W$

The standard $L=\lambda W$ framework is a sequence of ordered pairs of random variables $\{(A_k, D_k): k=1, 2, ...\}$, where $0 \le A_k \le A_{k+1}$ and $A_k \le D_k$ for all k; see Stidham [14], [15] and Glynn and Whitt [6], [7]. We usually *interpret* A_k and D_k as the arrival and departure epochs of customer k, but this is not necessary. Indeed, we will propose something different in the proof of Theorem 2.

Related quantities of interest are defined in terms of an indicator variable $1_k(t)$, which is 1 if $A_k \le t \le D_k$ and 0 otherwise. The number of customers in the system at time t and the waiting time of the k^{th} customer are

$$N(t) = \sum_{k=1}^{\infty} 1_k(t), \ t \ge 0, \quad \text{and} \quad W_k = D_k - A_k = \int_0^{\infty} 1_k(t)dt, \ k \ge 1. \qquad (1)$$

We can now state a version of Little's law, $L=\lambda W$; e.g., see Theorem 2 of Glynn and Whitt [6]. Throughout this paper, limits are understood to be with probability one (w.p.1).

Theorem 1. (Little's Law). *Assume that* $k^{-1}A_k \to \lambda^{-1}$ *as* $k \to \infty$, *where* $0 < \lambda < \infty$. *Then the following are equivalent:*

(*i*) $k^{-1}\sum_{j=1}^{k} W_j \to w \quad as \quad k \to \infty$

(*ii*) $t^{-1}\int_0^t N(s)ds \to \lambda w \quad as \quad t \to \infty \quad and \quad k^{-1} W_k \to 0 \quad as \quad k \to \infty.$

3. The Ordinal Version of L=λW

We now cease to keep track of time. Instead, we measure time in terms of arrival indices, so that the waiting time becomes the number of arrivals during a customer's sojourn time. As in Section 2, we assume that the customers are totally ordered, but we allow multiple arrivals at the same time; i.e., the k^{th} customer has arrival epoch A_k with $A_k \le A_{k+1}$. In the framework of Section 2, let X_k be the number of customers with indices greater than k that arrive while customer k is in the system, and let N_k^a be the number of customers with indices less than k that are in the system at the arrival epoch of customer k. (The qualification on the indices is included to cope with multiple events occurring at the same time.)

Here is our ordinal version of $L=\lambda W$.

Theorem 2. *The following are equivalent:*

(*i*) $k^{-1}\sum_{j=1}^{k} X_j \to x \quad as \quad k \to \infty;$

(*ii*) $k^{-1}\sum_{j=1}^{k} N_j^a \to x \quad and \quad k^{-1} X_k \to 0 \quad as \quad k \to \infty.$

A direct proof of Theorem 2 is not difficult (see Remark 3.1 below), but it essentially amounts to reproving a variant of Theorem 1. To show the connection to Theorem 1, we directly apply Theorem 1.

Proof. We put this problem in the framework of Section 2 by letting $A_k = k$ and letting D_k be the index of the last arrival during customer k's sojourn time (which is taken to be k if there are no other arrivals during this time). Then

$W_k = D_k - A_k = X_k$ is the number of arrivals with indices greater than k during the sojourn time of customer k. For t not an integer, $N(t)$ defined by (1) is the number of customers in the system at arrival epoch $[t]$ that are still present at arrival epoch $[t]+1$, where $[t]$ is the integer part of t. Thus, $N(t)$ only counts arrivals before $[t]+1$. Consequently, for any integer k,

$$k^{-1} \int_0^k N(s)ds = k^{-1} \sum_{j=1}^{k} N_j^a , \tag{2}$$

i.e., the average number of customers with smaller indices in the system upon arrival.

We are now ready to apply Theorem 1. Since $A_k = k$, the initial assumption is trivially satisfied with $\lambda=1$. Obviously, $k^{-1} \int_0^k N(s)ds \to x$ as $k \to \infty$ for integers k if and only if $t^{-1} \int_0^t N(s)ds \to x$ as $t \to \infty$, so $t^{-1} \int_0^t N(s)ds \to x$ as $t \to \infty$ is equivalent to $k^{-1} \sum_{j=1}^{k} N_j^a \to x$ as $k \to \infty$. Thus, Theorem 1 produces Theorem 2 with this interpretation. □

Remarks. (3.1) A direct proof starts by letting $Y_{ij} = 1$ if $j > i$ and customer j arrives while customer i is still in the system. Then let

$$X_k = \sum_{j=1}^{\infty} Y_{kj} = \sum_{j=k+1}^{\infty} Y_{kj} \quad \text{and} \quad N_k^a = \sum_{i=1}^{\infty} Y_{ik} = \sum_{i=1}^{k-1} Y_{ik} . \tag{3}$$

(3.2) Theorem 2 can also be proved by applying the general version of $L = \lambda W$ in Glynn and Whitt [7], but it is not any easier. In that context, Theorem 2 is interesting because it is treated by a two-dimensional cumulative input function $F(s, t)$ generated by a double sum.

(3.3) A central-limit-theorem version of Theorem 2 can be established, just as the central-limit-theorem version of Theorem 1 in Glynn and Whitt [6].

4. Relations for a Single-Server Queue

Consider a single-server queueing system in which work is processed at unit rate whenever work is present and the server is working. We can trivially express the time spent in the system by customer k, say T_k, in terms of five components:

S_k – service time of customer k,

W_k^a – work in system (in service time) just prior to arrival of customer k,

W_k^s – work (in service time) to arrive during customer k's sojourn time, (4)

I_k – idle time of the server during customer k's sojourn time,

W_k^d – work (in service time) just after the departure of customer k.

For each $k \geq 1$,

$$T_k = S_k + W_k^a + W_k^s + I_k - W_k^d . \tag{5}$$

We can obtain corresponding relations among long-run averages and expected steady-state values. In particular, we consider expected steady-state values, i.e., in a stationary process framework, as in Franken et al. [5]. Henceforth, assume that there is appropriate ergodicity as well as stationarity so that all long-run averages exist and coincide with expected stationary values. For the quantities in (5), this means the customer-stationary or synchronous version (Palm distribution). After dropping the index k, we obtain

$$E(T) = E(S) + E(W^a) + E(W^s) + E(I) - E(W^d) . \tag{6}$$

Furthermore, if the service times are i.i.d. and independent of the arrival process, (a G/GI/1 model), then we can express $E(W^s)$ as

$$E(W^s) = E(X)E(S) , \tag{7}$$

where X is the stationary number of arrivals during a customer's sojourn time. We can then apply Theorem 2 to obtain

$$E(X) = E(N^a) , \tag{8}$$

where N^a is the stationary number of customers in the system just prior to an arrival. We also have

$$E(N^a) = E(N^d), \tag{9}$$

where N^d is the stationary number of customers in the system just after a departure. (However, we must be careful about ties. For example, (9) is valid if we stipulate that N_k^a counts all customers that arrive at epoch A_k and have arrival indices *less* than k, while N_k^d counts all customers that depart at epoch D_k and have arrival indices *greater* than k. Then (9) follows by relating the upcrossings to the downcrossings.)

Finally, we can obtain an expression for $E(N)$, where N is the time-stationary number of customers in the system at an arbitrary time and λ is the arrival rate (assumed to be well defined, with $0<\lambda<\infty$). From Theorem 1,

$$E(N) = \lambda E(T). \tag{10}$$

5. Work-Conserving Disciplines Non-Anticipating with Respect to the Service Times

Now we assume that the G/GI/1 queue of Section 4 has a work-conserving discipline; i.e., the server is never idle when a customer is in the system, and the discipline does not affect the arrival epochs or the amount of service given to any customer. Hence, $I_k=0$ for all k in (5). As usual, let the traffic intensity be $\rho=\lambda E(S)<1$. Under all the assumptions above,

$$\begin{aligned} E(T) &= E(S)+E(W^a)+E(W^s) - E(W^d) \\ &= ES + E(W^a) + E(N^a)E(S) - E(W^d) \\ &= ES + E(W^a) + [E(N^d)E(S) - E(W^d)]. \end{aligned} \tag{11}$$

Of course, if the discipline is FIFO, then $E(T)=E(S)+E(W^a)$ and W^d is distributed as the random sum of N^d i.i.d. *full* service times, so that

$$E(W^d) = E(N^d)E(S). \tag{12}$$

Otherwise W^d is the sum of N^d *residual* service times, which can be quite complicated. However, we can use (11) to obtain useful comparisons for work-conserving disciplines that are non-anticipating with respect to the service

times. In particular, we can establish inequalities with NBUE and NWUE (new better or worse than used) service-time distributions. Recall that a service time S is NBUE (NWUE) if

$$E(S-t \mid S>t) \leq (\geq)\, E(S) \quad \text{for all} \quad t; \tag{13}$$

e.g., see Stoyan [16]. We summarize the conclusions in the following theorem.

Theorem 3. (*a*) *If the G/GI/1 queue of Section 4 has a work-conserving discipline, then*

$$E(T) = E(T;\ \text{FIFO}) + [E(N^d)E(S) - E(W^d)]. \tag{14}$$

(*b*) *If, in addition, the discipline is non-anticipating with respect to the service times, then*

$$E(T) \leq (\geq)\, E(T;\ \text{FIFO}) \quad \text{and} \quad E(N) \leq (\geq)\, E(N;\ \text{FIFO}) \tag{15}$$

whenever the service-time distribution is NWUE (NBUE), with equality holding when the service-time distribution is exponential.

Proof. (a) Apply (11) and (12). (b) Note that W^d is the sum of N^d residual service times, say S_i'. Let H^d represent the history of the system at the departure epoch associated with N^d and W^d, which includes the completed service times of all customers in the system. Under NBUE

$$E(W^d) = E\left[\sum_{i=1}^{N^d} S_i'\right] = E\left[E\left[\sum_{i=1}^{N^d} S_i' \mid H^d\right]\right]$$

$$\leq E\left[E\left[\sum_{i=1}^{N^d} (ES) \mid H^d\right]\right] = E(N^d)E(S). \quad \square$$

Remarks. (5.1) For the special case of a renewal arrival process, exponential service times and a PS discipline (the GI/M/1/∞/PS model), Theorem 3(b) is covered by Theorem 2 of Ramaswami [11]. Note that Ramaswami's result shows for the special case of exponential service times that, although the first moment of T does not depend on the discipline, the second moment (and thus the full distribution) of T *does* depend on the discipline.

(5.2) For GI/M/1, GI/E_k/1 and GI/D/1 queues, stronger stochastic comparisons consistent with Theorem 3(b) have been established in Section 3 of Shanthikumar and Sumita [13].

(5.3) For the special case of a Poisson arrival process and a PS discipline, Theorem 3(b) is an elementary consequence of known results; i.e.,

$$E(T;\ \mathrm{M/G/1,\ PS}) = E(T;\ \mathrm{M/M/1,\ FIFO}). \tag{16}$$

For queues with common ρ, $E(T;\mathrm{M/M/1,\ FIFO}) \le (\ge) E(T;\mathrm{M/G/1,\ FIFO})$ whenever $c_s^2 \ge (\le) 1$, where c_s^2 is the squared coefficient of variation of the service-time distribution. It is well known that S NBUE (NWUE) implies that $c_s^2 \le 1 (\ge 1)$.

(5.4) It is not difficult to see that the NBUE (NWUE) condition in Theorem 3(b) cannot be replaced by $c_s^2 \le 1$ (≥ 1) for general GI/GI/1 systems (non-Poisson arrival processes). A proof can be obtained by using a light-traffic argument as in Whitt [17] with a batch Poisson arrival process with geometric batch sizes, which is a renewal process. By making the Poisson arrival rate small, almost all busy periods consist of a single batch. By making the mean batch size small, batch sizes of size greater than k are asymptotically negligible compared to batch sizes of size k. All disciplines are the same for batches of size 1, so it suffices to consider batches of size 2. Let S_1 and S_2 be the service times of the two customers in such a batch. The expected sojourn time per customer during such a special busy period, say $E(T')$, with FIFO is

$$E(T';\ \mathrm{FIFO}) = S_1 + S_2/2$$

and, with PS, is

$$E(T';\ \mathrm{PS}) = \min\{S_1, S_2\} + \frac{\max\{S_1, S_2\} - \min\{S_1, S_2\}}{2}.$$

By considering specific distributions for S_i, e.g., two-point distributions, we see that $c_s^2 = 1$ is not the borderline between FIFO and PS.

(5.5) Note that FIFO coincides with SERPT for NBUE distributions and LERPT (longest) for NWUE distributions. Hence, Theorem 3(b) also

established extremal properties of SERPT and LERPT for subclasses of service-time distributions. However, it is not difficult to see that SERPT does *not* minimize $E(T)$ for *all* service-time distributions. For example, let $P(S=0)=0.9=1-P(S=11)$. Obviously, for $10<t<11$,

$$E(S-t \mid S>t) = 11-t<1.1 = E(S),$$

so that when a customer has been in the system for t in this range, SERPT does not give a new arrival a negligible amount of service to see if the zero service time is realized, as it should. A formal proof can be developed by using Poisson arrivals and light traffic, as in [17]. □

We now apply PASTA (Poisson arrival processes see time averages) as in Wolff [18] to obtain another consequence of (11). For the external exogenous arrival process considered here, we automatically have Wolff's *lack of anticipation assumption* (LAA): For each t, $\{A(t+u)-A(u): u\geq 0\}$ is independent of $\{N(s): 0\leq s\leq t\}$, where $A(t)$ is the arrival counting process.

Theorem 4. *(a) In an M/GI/1/∞ queue with a work-conserving discipline,*

$$E(N^a) = E(N) = \left[\frac{\rho}{1-\rho}\right]\left[1+\frac{E(W^a)-E(W^d)}{E(S)}\right]. \tag{17}$$

Proof. PASTA implies $E(N^a)=E(N)$ in (11). □

For an M/GI/1 model, we can also determine when $E(N)$ is the same as for an M/M/1 model with the FIFO discipline.

Corollary. In an M/GI/1/∞ queue with a work-conserving discipline

$$E(N) = \frac{\rho}{1-\rho} \equiv E(N;\ \text{M/M/1, FIFO}) \tag{18}$$

if and only if $E(W^a) = E(W^d)$.

Remarks. (5.6) In an M/GI/1 FIFO queue, $E(N^a)=E(N^d)$, but we typically do not have $E(W^a)=E(W^d)$, because W^d is the sum of N^d full service times, while W^a is the sum of N^a service times, one of which is residual (if $N^a\geq 1$).

(5.7) Let V be the steady-state workload at an arbitrary time. By PASTA, the distribution of W^a coincides with the distribution of V in the M/GI/1 model. Thus, $E(W^a)=E(W^d)$ if and only if $E(V)=E(W^d)$. Hence, departures see time averages (in the limited sense of these means) if and only if the moment $E(N)$ satisfies (18).

6. A Heavy Traffic Limit and An Approximation

In this final section, we apply our results to the case of a GI/GI/1 queue with a discipline such that at most m customers in the system can have received partial service at any time. For this model let

$$N^d = N_s^d + N_q^d$$

where N_s^d is the number of customers that have received partial service and N_q^d is the number of customers that have not received any service, among those N^d customers in the system at a departure epoch. Then

$$E(N^d)E(S) - E(W^d) = E(N_s^d)E(S) - E(W_s^d), \tag{19}$$

where W_s^d is the remaining work of the N_s^d customers that have received partial service.

For FIFO it is well known that

$$\lim_{\rho \to 1} (1-\rho)E(T) = E(S)(c_a^2 + c_s^2)/2, \tag{20}$$

where c_a^2 and c_s^2 are the squared coefficients of variation of the interarrival times and service times, respectively; see p. 196 of Asmussen [1]. To relate our discipline to FIFO, we assume the following regularity condition:

$$\sup_{t \geq 0} E(S-t \mid S>t) \leq K < \infty. \tag{21}$$

Theorem 5. *Consider a GI/GI/1 queue with a work-conserving discipline that is non-anticipating with respect to the service times and permits at most m customers in the system to have received partial service. If* (21) *holds, then*

$$\lim_{\rho \to 1} \frac{E(T)}{E(T;\ \text{FIFO})} = 1.$$

Proof. By (14), (19) and (21),

$$| E(T) - E(T; \text{FIFO}) | = | E(N_s^d)E(S) - E(W_s^d) | \leq mK,$$

so that the conclusion follows from (20). □

From (14) and (19), it follows that if one has good estimates or approximations for $E(N_s^d)$ and $E(W_s^d)$, then one obtains good estimates or approximations for $E(T)-E(T; \text{FIFO})$. For example, it should be much more efficient to estimate $E(N_s^d)$ and $E(W_s^d)$ by simulation than to estimate $E(T)-E(T; \text{FIFO})$ directly by simulation, especially if ρ is near 1.

A specific example is an M/G/1 queue with m "service positions", where the customers occupying the service positions are served according to a processor-sharing discipline. Variants of this model have been considered by Avi-Itzhak and Halfin [2]-[4] and Yamazaki and Sakasegawa [19]-[21]. In [2] the following approximation is proposed:

(1) N_s^d is approximately distributed as $\min(m-1, X)$; where X is the state of the equally loaded M/M/1 queue.

(2) The amount of residual work of a customer in a service position at a departure is approximately distributed as the stationary excess or residual lifetime of the service time.

These approximations are exact for the extreme cases $m=1$ and $m=\infty$. The resulting approximation is

$$E(T) \approx E(T; \text{FIFO}) + \frac{\rho-\rho^m}{1-\rho} \frac{1-c_s^2}{2} E(S). \tag{22}$$

Approximation (22) and variations could be used more generally.

As noted in the introduction, Yamazaki and Sakasegawa [19]-[21] derived for this model comparisons to the FIFO discipline. Their results are in agreement with the Theorem 3. Moreover, (22) is also proposed by them in (4.7) on p. 983 of [19].

ACKNOWLEDGMENTS

We thank Genji Yamazaki for bringing to our attention his related work in [19] and [21]. We thank Volker Schmidt and an anonymous referee for helpful comments.

REFERENCES

[1] ASMUSSEN, S. 1987. *Applied Probability and Queues,* Wiley, New York.

[2] AVI-ITZHAK, B. and S. HALFIN, 1987. Server sharing with a limited number of service positions and symmetric queues. *J. Appl. Prob.* 24, 990-1000.

[3] AVI-ITZHAK, B. and S. HALFIN, 1988. Expected response times in a non-symmetric time sharing queue with a limited number of service positions. *Proceedings of the 12th International Teletraffic Congress*, Torino, Italy, June 1988.

[4] AVI-ITZHAK, B. and S. HALFIN, 1988. Response times in M/M/1 time sharing schemes with limited number of service positions. *J. Appl. Prob.* 25, 579-595.

[5] FRANKEN, P., D. KÖNIG, U. ARNDT, and V. SCHMIDT, 1981. *Queues and Point Processes*, Akademie-Verlag, Berlin (and Wiley, Chichester, 1982).

[6] GLYNN, P. W. and W. WHITT, 1986. A central-limit-theorem version of $L=\lambda W$. *Queueing Systems: Theory and Applications*, 1, 191-215.

[7] GLYNN, P. W. and W. WHITT, 1989. Extensions of the queueing relations $L=\lambda W$ and $H=\lambda G$. *Opns. Res.*, to appear.

[8] HEYMAN, D. P. and M. J. SOBEL, 1982. *Stochastic Models in Operations Research,* Vol. I, McGraw-Hill, New York.

[9] KEILSON, J. and L. SERVI, 1988. A distributional form of Little's law. *Opns. Res. Letters*, 7, 219-222.

[10] LITTLE, J. D. C., 1961. A proof of the queueing formula: $L=\lambda W$. *Opns. Res.* 9, 383-387.

[11] RAMASWAMI, V., 1984. The sojourn time in the GI/M/1 queue with processor sharing. *J. Appl. Prob.* 21, 437-442.

[12] REGE, K. M. and B. SENGUPTA, 1985. Sojourn time distribution in a multiprogrammed computer system. *AT&T Tech. J.* 64, 1077-1090.

[13] SHANTHIKUMAR, J. G. and U. SUMITA, 1987. Convex ordering of sojourn times in single-server queues: extremal properties of FIFO and LIFO service disciplines. *J. Appl. Prob.* 24, 734-748.

[14] STIDHAM, S., Jr., 1974. A last word on $L = \lambda W$. *Opns. Res.* 22, 417-421.

[15] STIDHAM, S., Jr., 1982. Sample – path analysis of queues. In *Applied Probability – Computer Science: The Interface*, eds. R. Disney and T. J. Ott, Birkhauser, Boston.

[16] STOYAN, D., 1983. *Comparison Methods for Queues and Other Stochastic Models,* Wiley, New York.

[17] WHITT, W., 1986. Deciding which queue to join: some counterexamples. *Opns. Res.* 34, 55-62.

[18] WOLFF, R. W., 1982. Poisson arrivals see time averages. *Opns. Res.* 30, 223-231.

[19] YAMAZAKI, G. and H. SAKASEGAWA, 1986. The effect of multiplicity in limited processor-sharing systems. *Trans. Information Process Soc. Japan* 27, 979-987. (in Japanese)

[20] YAMAZAKI, G. and H. SAKASEGAWA, 1987. An optimal design problem for limited processor sharing systems. *Management Sci.* 33, 1010-1019.

[21] YAMAZAKI, G. and H. SAKASEGAWA, 1987. Limited Processor-Sharing Discipline in GI/GI/1 Models. Discussion Paper No. 321, University of Tsukuba, Sakura, Ibaraki 305, Japan.

Shlomo Halfin and Ward Whitt

An extremal property of the FIFO discipline via an ordinal version of L = lambda W

Received: 6/1/1988

Revised: 2/25/1989

Accepted: 3/1/1989

Recommended by Soren Asmussen, Editor