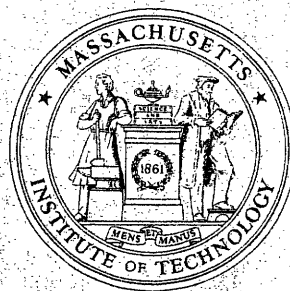


OPERATIONS RESEARCH CENTER

working paper



**MASSACHUSETTS INSTITUTE
OF TECHNOLOGY**

The Distributional Form of Little's Law
and the Fuhrmann-Cooper Decomposition

by

J. Keilson and L.D. Servi

OR 187-88

September 1988

ABSTRACT

For certain classes of customers, the ergodic number of customers in system (queue) and the ergodic time spent in system (queue) are related by a distributional form of Little's Law [6]. Classes for which Little's Law holds in distribution will be called LLD classes and will be said to have the LLD property. Because LLD classes have a simple characterization permitting their quick recognition, the LLD property is a powerful tool. For the M/G/1 system, for example, the Pollaczek-Khinchin distribution drops out in a few lines. Moreover, under simple conditions (cf. Theorem 2) the pgf of the ergodic number in queue is shown to have a structurally simple decomposition of the form of Fuhrmann and Cooper [2]. This decomposition is implemented for many queueing systems and is shown thereby to provide a degree of unification to queueing theory.

1. Introduction

In a previous paper, Keilson and Servi [6] showed that for certain "LLD" classes of customers (cf. Section 2), the ergodic number N_S of customers in system, i.e. in queue or in the service box, and the ergodic time T_S spent in system by customers of the class were related by a distributional form of Little's Law. Specifically it was shown that the two descriptive functions

$$\pi_S(u) = E[u^{N_S}] \text{ and } \alpha_{T_S}(s) = E(e^{-sT_S}),$$

are related by

$$(1.1) \quad \pi_S(u) = \alpha_{T_S}(\lambda - \lambda u).$$

Equation (1.1) is equivalent to the statement that N_S equals the number of Poisson arrivals at rate λ during an interval of duration T_S , i.e. to the equality in distribution

$$(1.2) \quad N_S \stackrel{d}{=} K_{\lambda T_S}$$

where K_θ is a Poisson variate of parameter θ . The prevalence of such LLD classes and the simple consequences of the law (1.1) were discussed in [6]. A corresponding result for the number in queue and the time in queue was also given.

LLD classes have a simple characterization permitting their quick recognition. It is then useful to observe that the LLD property provides an analytical tool of some power. For the M/G/1 system, for example, the Pollaczek-Khinchin distribution is found by simple algebra. Moreover, under the simple conditions of Theorem 2, the pgf of the ergodic number in queue for an LLD class has a decomposition into two structurally simple factors, the first being of the Pollaczek-Khinchin form. The decomposition of Theorem 2 unifies many of the results of queueing theory, and provides a quick derivation for other systems as well. Many of the results for priority queues fall out quickly, and acquire a simple structural form. The results for Head of the Line discipline are especially attractive.

Fuhrmann and Cooper [2] have demonstrated the decomposition of Theorem 2 under conditions close to that for LLD classes. They employ a somewhat longer and more

indirect argument and they provide little implementation for concrete cases of interest. Their theorem is given in Section 4.

2. LLD Classes

For ease of reference, the definition of an LLD class and the basic theorem of the earlier paper are given next in slightly different form.

Definition 2.1

Let an ergodic queueing system S have a subsystem S^* . Let C be a class of customers entering S^* such that:

- a) customers from C enter S^* in a Poisson stream of rate λ ;
- b) the customers of C in S^* leave S^* one at a time in order of arrival;
- c) for any time t , the entry process into S^* from C after time t and the time spent in S^* by any customer in C arriving before time t are independent.

The class C will then be called an LLD Class for S^* .

One then has the following theorem [6].

Theorem 1. Let C be an LLD class for S^* . Then the distributional form of Little's Law is valid for the ergodic number N_{S^*} of customers of C in S^* and the ergodic time T_{S^*} spent by customers of C in S^* , i.e. one has for $\pi_{S^*}(u) = E[N_{S^*}]$ and $\alpha_{T_{S^*}}(s) = E[T_{S^*}]$,

$$(2.1) \quad \pi_{S^*}(u) = \alpha_{T_{S^*}}(\lambda - \lambda u).$$

If the subsystem S^* is the queue and the service box then (2.1) reduces to (1.1). If S^* is simply the queue then one has

$$(2.2) \quad \pi_Q(u) = \alpha_{T_Q}(\lambda - \lambda u)$$

where $\pi_Q(u)$ is the p.g.f. of the number in the queue and $\alpha_{T_Q}(s)$ is the transform of the time in the queue.

Proposition 1. For a counting variate N and an associated time T governed by the distributional form of Little's Law given in (2.1) or (2.2) one has the following relations between the successive moments:

$$(2.2a) \quad E[N] = E[\lambda T]$$

$$(2.2b) \quad E[N^2] = E[(\lambda T)^2] + E[\lambda T]$$

$$(2.2c) \quad E[N^3] = E[(\lambda T)^3] + 3 E[(\lambda T)^2] + E[\lambda T]$$

$$(2.2d) \quad E[N^4] = E[(\lambda T)^4] + 6 E[(\lambda T)^3] + 7 E[(\lambda T)^2] + E[\lambda T]$$

$$(2.2e) \quad E[N^5] = E[(\lambda T)^5] + 10 E[(\lambda T)^4] + 25 E[(\lambda T)^3] + 15 E[(\lambda T)^2] + E[\lambda T]$$

Proof: Equations (2.2a) - (2.2e) were obtained by successive differentiation of (2.2) with the help of MACSYMA. ♦

Remark: Measurement of the mean, variance, skewness and kurtosis from an empirical distribution of N permits one to solve for the corresponding moments of T and to approximate the distribution of T by selecting it from some four parameter family of distributions.

For most of the examples in [6], the distributional form of Little's law was valid for both the number of customers of C in queue and the "number in system", i.e. the number in queue and service box.

The applications of interest are vacation models, priority service systems and cyclic service systems. In all of these systems one may visualize a server in a class as joining the queue for that class, entering the server box and initiating service. The service may be interrupted and resumes where left off or starts over. In this setting, the time in queue refers to the time from the arrival of the customer to the queue until the customer begins service. The *effective service time*, T_{EFF} , refers to the time from the beginning of service until the customer completes service. Until service is completed, the customer will be regarded as being in the *service box*, whether or not it is being served.

Theorem 2.

If the time in queue of a customer in class C is independent of the effective service time and both (1.1) and (2.2) are satisfied, then

$$(2.3) \quad \pi_Q(u) = \frac{(1-\rho_{EFF})(1-u)}{\alpha_{TEFF}(\lambda-\lambda u) - u} \pi_B(u) = \frac{1-\rho_{EFF}}{1 - \rho_{EFF} \alpha_{TEFF}^*(\lambda-\lambda u)} \pi_B(u)$$

where

$\pi_B(u)$ = the pgf of the ergodic number of class C customers in queue given that no class C customer is in the service box,

$\alpha_{TEFF}(s)$ = $E[e^{-sT_{EFF}}]$,

ρ_{EFF} = $\lambda E[T_{EFF}]$,

and

$$\alpha_{TEFF}^*(s) = \frac{1 - \alpha_{TEFF}(s)}{sE[T_{EFF}]}$$

Moreover,

$$(2.4) \quad N_Q \stackrel{d}{=} N_{M/G/1} + N_B$$

with $N_{M/G/1}$ and N_B independent. Here

N_Q = the ergodic number of class C customers in the queue,

$N_{M/G/1}$ = the ergodic number in the queue of an M/G/1 queue with an arrival rate λ and a service time T_{EFF} ,

N_B = the ergodic number of class C customers in the queue given that no class C customer is in the service box.

Proof:

To prove the theorem some additional notation is needed. Let

p_B = P [a customer is in the service box]

$\pi_{QB}(u)$ = the pgf of the ergodic number of class C customers in queue given that a class C customer is in the service box.

The number in the queue has a pgf of $\pi_{QB}(u)$ if a customer is in the service box and, if not, $\pi_B(u)$, i.e., $\pi_Q(u) = p_B \pi_{QB}(u) + (1-p_B) \pi_B(u)$. Using a similar argument, $\pi_S(u) = p_B u \pi_{QB}(u) + (1-p_B) \pi_B(u)$. Hence,

$$(2.5) \quad \pi_S(u) = \pi_Q(u)u + (1-p_B)(1-u)\pi_B(u).$$

If the time in queue is independent of the effective service time then $\alpha_{TS}(s) = \alpha_{TQ}(s)\alpha_{TEFF}(s)$. Therefore from (1.1) and (2.2),

$$(2.6) \quad \pi_S(u) = \pi_Q(u) \alpha_{TEFF}(\lambda - \lambda u).$$

Equating the right hand side of equations (2.5) and (2.6) and solving for $\pi_Q(u)$ gives $\pi_Q(u) = \pi_B(u)[(1-p_B)(1-u) / [\alpha_{TEFF}(\lambda - \lambda u) - u]]$. From $\pi_Q(1) = 1$ and L'Hôpital's Rule,

$$(2.7) \quad p_B = \rho_{EFF} = -\lambda \alpha_{TEFF}'(0)$$

and equation (2.3) then follows. From classical queueing theory (or from case 1 below) the first factor of the right hand side of (2.3) is $E[u^{NM/G/1}]$ so (2.4) follows. ♦

Remark: Equations (1.1) and (2.2), and hence (2.3) and (2.4), are true when the order of service is FIFO. However, for any service order which is not dependent on the individual service times, the distribution of the number in the queue is the same. Hence, (2.3) and (2.4) are valid for this larger class of service orders.

Corollary 1:

If the time in queue of a customer in class C is independent of the effective service time and both (1.1) and (2.2) are satisfied then the pgf of the number in the system is

$$(2.8) \quad \pi_S(u) = \frac{(1-\rho_{EFF}) \alpha_{TEFF}(\lambda - \lambda u)}{1 - \rho_{EFF} \alpha_{TEFF}(\lambda - \lambda u)} \pi_B(u),$$

the transform of the time in system is

$$(2.9) \quad \alpha_{TS}(s) = \frac{(1-\rho_{EFF}) \alpha_{TEFF}(s)}{1 - \rho_{EFF} \alpha_{TEFF}(s)} \pi_B([\lambda - s]/\lambda),$$

and the transform of the waiting time is

$$(2.10) \quad \alpha_Q(s) = \frac{1 - \rho_{EFF}}{1 - \rho_{EFF} \alpha_{TEFF}(s)} \pi_B([\lambda - s]/\lambda).$$

Proof: Equations (2.8)-(2.10) follow from (1.1), (2.2) and (2.3). ♦

3. Special cases.

CASE 1: M/G/1 Queue

Here $\alpha_{\text{TEFF}}(s) = \alpha_T(s)$ and $N_B = 0$ so that from (2.3)

$$(3.1) \quad \pi_Q(u) = \frac{(1-\rho)(1-u)}{\alpha_T(\lambda-\lambda u) - u} .$$

CASE 2: M/G/1 with vacations and exhaustive service

For this discipline, e.g., [1], an M/G/1 queue is served exhaustively. Then the server is inactive, (i.e., "on vacation") for a duration V . At the end of a vacation period another vacation period begins if the system is empty. Otherwise the queue is again served exhaustively. It is assumed that V is independent of the arrival process.

Again, $T_{\text{EFF}} \stackrel{d}{=} T$, the service time of the queue. If no service is in progress then the system must be on vacation. N_B is then equal to the number of customers that have arrived since the last vacation began. But the time at ergodicity since the last vacation began and the forward recurrence time V^* of a vacation time are equal in distribution. Hence $N_B \stackrel{d}{=} K_{\lambda V^*}$, and N_B has a pgf $\pi_B(u) = \alpha_V^*(\lambda-\lambda u)$ where $\alpha_V^*(s) = [1-\alpha_V(s)] / [E[V]s]$ and $\alpha_V(s) = E[e^{-sV}]$. Hence, from (2.3),

$$(3.2) \quad \pi_Q(u) = \frac{1-\rho}{1 - \rho \alpha_T^*(\lambda-\lambda u)} \alpha_V^*(\lambda-\lambda u) .$$

CASE 3: M/G/1 queue with two classes and preemptive priority.

Consider a schedule with two classes of Poisson traffic with arrival rates λ_1, λ_2 and customer service time transforms be $\alpha_{T1}(s), \alpha_{T2}(s)$ for the high priority and low priority traffic respectively. Let the transform of the high priority busy period be $\sigma_{\text{BP1}}(s)$. Suppose the high priority traffic preempts the low priority traffic. The cases of preempt resume, preempt-repeat and possible hybrid modes compatible with the LLD requirement are considered simultaneously.

The pgf of the number N_1 in queue of the high priority traffic has the classical Pollaczek-Khinchin form since low priority traffic is ignored. To find the pgf of N_2 , the number in queue of the low priority traffic, one needs the pgf $\pi_B(u)$ in Theorem 2. As we see next, for any preemptive case, one has

$$(3.3) \quad \pi_B(u) = 1 - \rho_1 + \rho_1 \sigma_{BP1}^*(\lambda_2 - \lambda_2 u)$$

where $\sigma_{BP1}^*(s) = [1 - \sigma_{BP1}(s)] / [E[T_{BP1}]s]$ is the transform of the backward recurrence time of the priority traffic busy period.

To see this let E_k be the event that no C_k customers are in the service box at ergodicity, $k = 1, 2$. Let A^c be the event not A . Then $E_2 = E_2 E_1 + E_2 E_1^c$ and

$$(3.4) \quad \begin{aligned} \pi_B(u) &= E[u^{N_2(\infty)} | E_2] \\ &= P[E_2 E_1 | E_2] E[u^{N_2(\infty)} | E_2 E_1] + P[E_2 E_1^c | E_2] E[u^{N_2(\infty)} | E_2 E_1^c] \end{aligned}$$

The events E_1 and E_2 are independent since $P[E_1 | E_2] = P[E_1]$. This is because the C_1 customers have preempt priority over C_2 customers and hence the C_2 customers are "invisible" to the C_1 customers. One then has $P[E_1 E_2] = P[E_1] P[E_2]$. One also verifies that the idle state $I = E_1 E_2$. Hence

$$A): \quad P[E_1 E_2 | E_2] = P[E_1] = 1 - \rho_1,$$

$$B): \quad P[E_2 E_1^c | E_2] = 1 - P[E_2 E_1 | E_2] = \rho_1,$$

and

$$C): \quad E[u^{N_2(\infty)} | E_2 E_1] = E[u^{N_2(\infty)} | I] = 1.$$

To see (3.3) from (3.4) and A),B),C) we need only establish that

$$D): \quad E[u^{N_2(\infty)} | E_2 E_1^c] = \sigma_{BP1}^*(\lambda_2 - \lambda_2 u),$$

the pgf of the number of Poisson arrivals of C_2 customers during the backwards recurrence time of a C_1 customer busy period.

The event $E_2E_1^C$ corresponds to a C_1 customer in service and no C_2 customer in the service box. Such an event is initiated only by the arrival of a C_1 customer to an idle server and the duration of the sojourn on the set $E_2E_1^C$ is a C_1 busy period. The result then follows from renewal theory applied to the recurrences of such initiations.

Using simple algebra one can show from (3.3) that

$$(3.5) \quad \pi_B(u) = \frac{(1-\rho_1)\zeta(u)}{\lambda_2 - \lambda_2 u}$$

where

$$(3.6) \quad \zeta(u) = \lambda_2 - \lambda_2 u + \lambda_1 - \lambda_1 \sigma_{BP1}(\lambda_2 - \lambda_2 u).$$

Case 3a Preempt - resume:

The effective service time of a low priority customer is the time from when it first leaves the queue until it completes service (after possibly one or more priority interruptions). One then has [3], [5],

$$(3.7) \quad \alpha_{T2EFF}(s) = \alpha_{T2}(s + \lambda_1 - \lambda_1 \sigma_{BP1}(s))$$

and from (2.7)

$$(3.8) \quad \rho_{2EFF} = \frac{\rho_2}{1 - \rho_1}.$$

Hence, from (2.3), (3.3), (3.7) and (3.8),

$$(3.9) \quad \pi_Q(u) = \frac{1 - \rho_{2EFF}}{1 - \rho_{2EFF} \alpha_{T2EFF}^*(\lambda_2 - \lambda_2 u)} [1 - \rho_1 + \rho_1 \sigma_{BP1}^*(\lambda_2 - \lambda_2 u)].$$

From (3.6),

$$(3.10) \quad \pi_Q(u) = \frac{(1 - \rho_1 - \rho_2)\zeta(u)}{\lambda_2 [\alpha_{T2}(\zeta(u)) - u]}$$

which is consistent with [7, equation (A.9)].

Using algebra and (3.10) one can also show that

$$(3.11) \quad \pi_Q(u) = \frac{1-\rho_1-\rho_2}{1 - (\rho_1-\rho_2)\alpha_{T12}^*(\zeta(u))}$$

where

$$\alpha_{T12}(s) = \frac{\lambda_1}{\lambda_1+\lambda_2} \alpha_{T1}(s) + \frac{\lambda_2}{\lambda_1+\lambda_2} \alpha_{T2}(s) \quad \text{and} \quad \alpha_{T12}^*(s) = \frac{1 - \alpha_{T12}(s)}{-\alpha_{T12}(0)s}$$

which is consistent with [8, equation (1.6)].

Case 3b Preempt-repeat

Here, whenever the interruption of low priority customer ends a *new* service time begins after the interruption. Two cases could be considered [4]. In one case after the interruption a new service times is randomly selected (Preempt-Repeat different). In the second case after the interruption the service time remains the same (Preempt-Repeat Identical).

First Preempt-Repeat different will be considered: The Laplace Transform of T_{2EFF} can be found either in [4, Chapter IV, equation (2.34)] or by using the simpler derivation found in the appendix. It is

$$(3.12) \quad \alpha_{T2EFF}(s) = \frac{\alpha_{T2}(s + \lambda_1)}{1 - \frac{\lambda_1}{s + \lambda_1} (1 - \alpha_{T2}(s + \lambda_1)) \sigma_{BP1}(s)}$$

Hence,

$$(3.13) \quad \rho_{2EFF} = \frac{\lambda_2(1 - \alpha_{T2}(\lambda_1))}{\lambda_1(1 - \rho_1) \alpha_{T2}(\lambda_1)}$$

Finding $\pi_B(u)$ and hence N_B is done using exactly the same argument as the Preempt-Resume case, i.e., $\pi_B(u) = p_A + (1 - p_A)\sigma_{BP1}^*(\lambda_2 - \lambda_2 u)$ where, from (2.7), $p_A = \rho_{2EFF}$ and ρ_{2EFF} is given in (3.13).

Therefore, for Preempt-Repeat different the pgf of N_Q is given by

$$(3.14) \quad \pi_{Q(u)} = \left[\frac{1 - \rho_{2EFF}}{1 - \rho_{2EFF} \alpha_{T_2}^*(\zeta(u))} \right] [p_A + (1 - p_A) \sigma_{BP1}^*(\lambda_2 - \lambda_2 u)]$$

where ρ_{EFF} , p_A and $\zeta(u)$ are given in (3.13), (2.7), and (3.6).

The distribution of the effective service time, T_{2EFF} , for preempt-repeat identical is obtained as follows. One evaluates the distribution for preempt-repeat different with the service time T_2 having a deterministic value x and then weights the result by the distribution of T_2 . From (3.14), one then has

$$(3.15) \quad \alpha_{T_{2EFF}}(s) = \int_0^{\infty} \frac{\exp(-(s + \lambda_1)x)}{1 - \frac{\lambda_1}{s + \lambda_1} (1 - \exp(-(s + \lambda_1)x)) \sigma_{BP1}(s)} dA_{T_2}(x)$$

where $A_{T_2}(x)$ is the cdf of the C_2 customer service time.

Case 4: M/G/1 queue with two classes and Head of Line priority

For this service discipline, there are again two classes of Poisson traffic with high and low priority. The lower priority C_2 traffic is interrupted by the higher priority C_1 traffic, but a low priority customer in progress completes service before C_2 service is interrupted.

Let $B(t)$, the combined backlog process for this schedule, be the total amount of C_1 and C_2 work in the system. Let $\beta(s)$ is the transform of the ergodic backlog $B(\infty)$. Then for preempt-resume discipline, the waiting time of C_2 is equal in distribution to the ergodic backlog with Poisson interruptions of rate λ_1 having iid durations equal to that of a C_1 busy period, i.e., $\alpha_{Q_2}(s) = \beta(s + \lambda_1 - \lambda_1 \sigma_{BP1}(s))$. From the distributional form of Little's law, the p.g.f. of the number in the queue is $\alpha_{Q_2}(\lambda_2 - \lambda_2 u) = \beta(\zeta(u))$ (where $\zeta(u)$ is defined in (3.6).

Backlog processes, however, are independent of the order of service. In particular $\beta(s)$ for preempt-resume and head-of-line (HOL) service disciplines are the same. For

$\alpha_{Q_2}(s)$ for both disciplines, moreover, this backlog is subject to the Poisson interruptions of C_1 busy periods. It follows that the *pgf. of the number of C_2 customers in the queue for the HOL and the preempt-resume service disciplines are the same* (and given in (3.10) and (3.11)).

Of course, the *effective* service time of a C_2 customer for a HOL has Laplace Transform $\alpha_{T_2}(s)$ (and not (3.7)) because the C_2 customers are not interrupted during service. Thus, the number of C_2 customers in the system is given by (2.6) where $\pi_Q(u)$ is (3.10) (or (3.11)) and $\alpha_{T_2EFF}(s)$ is $\alpha_{T_2}(s)$ and the time in the queue or the system follow from the distributional form of Little's Law, (2.1) [6].

The number of C_1 customers in the queue is found using an argument similar to that of case 3: As before, let E_k be the event that no C_k customers are in the service box at ergodicity. One needs the pgf $\pi_B(u) = E[u^{N_1(\infty)} | E_1]$. Again we have

$$(3.16) \quad \pi_B(u) = E[u^{N_1(\infty)} | E_1] \\ = P[E_1 E_2 | E_1] E[u^{N_1(\infty)} | E_1 E_2] + P[E_1 E_2^C | E_1] E[u^{N_1(\infty)} | E_1 E_2^C]$$

For $E_1 E_2$ one has $\text{Prob}[E_1 E_2 | E_1] = \text{Prob}[E_1 E_2] / \text{Prob}[E_1]$. But the idle state $I = E_1 E_2$ so $P[E_1 E_2] = 1 - \rho_1 - \rho_2$. Clearly, $P[E_1] = 1 - \rho_1$ so that $P[E_1 E_2 | E_1] = (1 - \rho_1 - \rho_2) / (1 - \rho_1)$. As before, $E[u^{N_2(\infty)} | E_1 E_2] = 1$.

For $E_1 E_2^C$, $\text{Prob}[E_1 E_2^C | E_1] = \text{Prob}[E_2^C | E_1] = 1 - P[E_2 | E_1] = \rho_2 / (1 - \rho_1)$. The event $E_1 E_2^C$ is equivalent to the event, E_2^C , that a C_2 service is under way with a service time distributed as T_2 . At the beginning of this service time, N_1 was zero, or C_1 service would have been started. At ergodicity the elapsed time since the C_2 service was initiated is distributed as the forward recurrence time for T_2 with transform $\alpha_{T_2}^*(s) = [1 - \alpha_{T_2}(s)] / [E[T_2] s]$.

Hence $E[u^{N_1(\infty)} | E_1 E_2^C] = \alpha_{T_2}^*(\lambda_1 - \lambda_1 u)$. A more formal and long winded argument based on semi-Markov processes reaches the same conclusion. From (3.16) we have finally

$$(3.17) \quad \pi_B(u) = \frac{1-\rho_1-\rho_2}{1-\rho_1} + \frac{\rho_2}{1-\rho_1} \alpha_{T_2}^*(\lambda_1-\lambda_1 u).$$

In HOL the C_1 are not interrupted so $\alpha_{T_{1EFF}}(s) = \alpha_{T_1}(s)$ and $\rho_{1EFF} = \rho_1$. Hence, from (2.3) and (3.17) the p.g.f. of the number of C_1 customers in the queue is given by

$$(3.18) \quad \pi_Q(u) = \frac{1-\rho_1}{1-\rho_1 \alpha_{T_2}^*(\lambda_1-\lambda_1 u)} \left[\frac{1-\rho_1-\rho_2}{1-\rho_1} + \frac{\rho_2}{1-\rho_1} \alpha_{T_2}^*(\lambda_1-\lambda_1 u) \right].$$

4. The Fuhrmann - Cooper Decomposition.

Fuhrmann and Cooper [2] have given an interesting and important decomposition for vacation models in a general setting similar to that of Theorem 2 but with somewhat different conditions listed below. The decomposition applies to M/G/1 vacation models which need not have exhaustive service. It could be applied for example to N-policies, cyclic service queues, M/G/1 queues with gated vacations, limited service queueing model, and static priority systems.

Theorem 3. (Fuhrmann and Cooper [2])

Let the following conditions hold.

1: Customers arrive to the system according to a Poisson Process of rate $\lambda > 0$ with identically distributed service times independent of each other, independent of the arrival process and independent of the vacation periods that precede it.

2: All customers are eventually served.

3: Customers are served in an order that is independent of the services times.

4: Service is non-preemptive. (Fuhrmann and Cooper observe that for preemptive service "it is a simple matter to appropriately 'inflate' the service times ... to account for the [preempt] interruptions". They give no details.)

5: The rules that govern when the server begins and ends vacations do not anticipate future jumps of the Poisson Arrival process.

Then the decomposition of Equation (2.3) holds with the summands independent.

If further

6: The number of customers that arrive during a vacation is independent of the number of customers present in the system when the vacation began

then

$$(4.1) \quad \pi_Q(u) = \left[\frac{(1-\rho)(1-u)}{\alpha_T(\lambda - \lambda u) - u} \right] \pi_{QV}(u) \alpha_{V^*}(\lambda - \lambda u).$$

This can be seen as follows, using the arguments of this paper: If no effective service time is in progress then the system must be on vacation. Hence, N_B equals the number in the system during an arbitrary instant in a vacation. But this equals in the number in the queue at the beginning of a vacation, N_{QV} , plus the number of arrivals during the vacation backward recurrent time, i.e., $N_B = N_{QV} + K_{\lambda V^*}$ so $\pi_B(u) = \pi_{QV}(u) \alpha_{V^*}(\lambda - \lambda u)$. Hence, (4.1) follows from (2.3).

Remark: Fuhrmann and Cooper do not require FIFO order of service because of their exclusive concern with queue population. As shown in the remark under Theorem 2, Theorem 2 has comparable requirements.

Acknowledgment: The authors wish to thank Hongtao Zhang for his succinct calculation of Proposition 1 via MACSYMA and Professor R. Larson for encouraging Proposition 1.

REFERENCES

- [1] Doshi, B. T, "Queueing Systems with Vacations - A Survey", *Queueing Systems*, Vol. 1, pp. 29-66, 1987.
- [2] Fuhrmann, S. W. and R.B. Cooper, "Stochastic Decompositions in the M/G/1 Queue with Generalized Vacations", *Operations Research*, Vol. 33, No. 5, September-October, 1985, pp. 1117-1129.
- [3] Gaver, D. P., "A Waiting Time with Interrupted Service, including Priorities, *Journal of Royal Statistical Society, Series B*, Vol 24, pp. 73-90.
- [4] Jaiswal, N. K., *Priority Queues*, Academic Press, New York, 1968.
- [5] Keilson, J., "Queues Subject to Service Interruptions", *Annals of Mathematical Statistics*, Vol. 33, No. 4, December, 1962, pp. 1314-1322.
- [6] Keilson, J. and L. D. Servi, "A Distributional Form of Little's Law", to appear in *Operations Research Letters*, Vol. 7, No. 5, 1988.
- [7] Keilson, J. and L. D. Servi, "The Preempt-Resume M/G/1 queue with Feedback and Clocked Schedules ", submitted to *Performance Evaluation*, 1988.
- [8] Keilson J. and U. Sumita, "Evaluation of the Total Time in System in a Preempt-Resume Priority Queue Via a Modified Lindley Process", *Advances in Applied Probability*, Vol. 15, 1984, pp. 840-856.

APPENDIX: Derivation of T_{EFF} for the Preempt-Repeat-Different discipline

For simplicity, assume that all of the random variables are absolutely continuous. In the case where they are not absolutely continuous an argument could be constructed using a limiting argument.

Let

t_1 be the time of the first interruption minus the time of the first C_2 service initiation. The interarrival times of interruptions are exponential distributed with a mean $1/\lambda_1$.

t_{BP1} be the duration of the busy period of a high priority customer. This equals the duration of an interruption. Let $s_{BP1}(x)$ and $\sigma_{BP1}(s)$ be the density function and Laplace transform, respectively, of t_{BP1} .

t_T be the service time of the low priority customer. This equals the time that the first low priority service would have ended if there were no interruptions. Let $a_{T2}(x)$ and $\alpha_{T2}(s)$ be its density function and Laplace transform, respectively, of t_{BP1} . Let $\bar{A}_{T2}(x) = \text{Prob} \{ t_T \geq x \}$.

$t_{2EFF(n)}$ be the duration of the interval from the first service initiation until *either* the end of a service time or until the end of the n^{th} interruption (whichever comes first). Let $a_{T2EFF(n)}(x)$ be its density function.

The key observation is that

$$(A.1) \quad t_{2EFF(n+1)} = \begin{cases} t_T & \text{if } t_T \leq t_1 \\ t_1 + t_{BP1} + t_{2EFF(n)} & \text{if } t_T > t_1 \end{cases}$$

Note that

$$(A.2) \quad \frac{d}{dx} \{ \text{Prob} [t_T \leq x \text{ and } t_T \leq t_1] \} = e^{-\lambda_1 x} a_{T2}(x)$$

and

$$(A.3) \quad \frac{d}{dx} \{ \text{Prob} [t_1 + t_{BP1} + t_{2EFF(n)} \leq x \text{ and } t_T > t_1] \} \\ = \lambda_1 e^{-\lambda_1 x} \bar{A}_{T2}(x) * s_{BP1}(x) * a_{T2EFF(n)}(x)$$

From (A.1) - (A.3),

$$a_{T2EFF(n+1)}(x) = e^{-\lambda_1 x} a_{T2}(x) + \lambda_1 e^{-\lambda_1 x} \bar{A}_{T2}(x) * s_{BP1}(x) * a_{T2EFF(n)}(x).$$

As n approaches infinity, $a_{T2EFF}(x) = e^{-\lambda_1 x} a_{T2}(x) + \lambda_1 e^{-\lambda_1 x} \bar{A}_{T2}(x) * s_{BP1}(x) * a_{T2EFF}(x)$
so (3.12) and (3.13) follow.