Priorities in $M/G/1$ Queue with Server Vacations

Offer Kella and Uri Yechiali
Department of Statistics, Tel Aviv University, Tel Aviv 69978, Israel

The $M/G/1$ queue with single and multiple server vacations is studied under both the preemptive and non-preemptive priority regimes. A unified methodology is developed to derive the Laplace-Stieltjes transform and first two moments of the waiting time W_k of a class-k customer for each of the four models analyzed. The results are given a probabilistic representation involving mean residual lifetimes.

1. INTRODUCTION

We present a unified methodology for the study of waiting times in the $M/G/1$ queue with several classes of customers and with single or multiple server vacations under both the preemptive and non-preemptive priority regimes. Four models are analyzed simultaneously, and in each case we derive the Laplace-Stieltjes transform (LST) and the first two moments of the waiting time W_k of a class-k customer, assuming order-of-arrival service within classes. Our methodology is based on the observation that each model may be viewed as a special version of the basic single-class *nonpriority* $M/G/1$ queue with *multiple*-server vacations. Employing this approach, we obtain new results concerning the two vacation models under the preemptive-resume regime, as well as known results concerning the two vacation models under the non-preemptive service discipline. This approach has already been successfully used to obtain the LST and first two moments of W_k for the *many*-server non-preemptive priority $M/M/c$ queue with same mean service time for all classes (see Kella and Yechiali [12]).

Several authors have studied processes resembling the so-called "vacation" models. Gaver [8], Keilson [11], and earlier authors as well studied the $M/G/1$ model that allowed the server to be interrupted. Gaver defined interruptions as "the elements that prevent the continuous service of arrivals," and considered Poisson-type interruptions that are caused either by a machine breakdown or by the appearance of high-priority customers. Observing that "a busy period generated by high-priority class of customers acts as an interruption in low-priority service," he analyzed the $M/G/1$ queue with compound Poisson arrival and obtained the LST and associated moments of the busy period, as well as the generating function of the number of customers in system, for both preemptive and postponable (non-preemptive) service disciplines.

Cooper [2] was the first to use the term "vacation" and to define the vacation-type disciplines of "exhaustive service" and "gated service." He studied the single-class $M/G/1$ queue with multiple identical (but not necessarily independent) server vactions and obtained the LST and mean waiting time of an arbitrary

Naval Research Logistics, Vol. 35, pp. 23–34 (1988)
 CCC 0028-1441/88/010023-12$04.00

customer for the case of service in order of arrival. Heyman [10] termed server vacations as a "blocking process" and studied the expected delay of a class-k customer in a non-preemptive priority regime for "two specific blocking processes—one representing the use of the server for a potentially unlimited number of postponable jobs" (i.e., multiple vacations), "the other representing server repair" (i.e., single vacation). He derived the mean waiting time EW_k in the multiple-vacations case, but his derivation of EW_k for the single-vacation case (Eq. (7) in [10]) contains a flaw.

Levy and Yechiali [15] further studied the single- and multiple- (*iid*) vacation models in the nonpriority single-class $M/G/1$ queue. They derived the generating functions of the number of customers in system for each of the two vacation models and were the first to obtain the LST and moments of W_k for the single-vacation case. Their result (21) corrects Heyman's equation (7). Levy and Yechiali's explicit formulas for the LST of W_k in the multiple-vacation case and for the mean number of customers present (Eq. (36) and (35) in [15]) may be viewed as special cases of Cooper's [2] equations (18) and (20). Scholl and Kleinrock [16] also treated the $M/G/1$ queue with multiple-server vacations and gave additional results concerning waiting times under the first-come first-served, random order of service, and non-preemptive last-come first-served disciplines. Shanthikumar [17] analyzed the two $M/G/1$ vacation models with several classes of customers and *non-preemptive* priority service discipline. Using level-crossing arguments he obtained the LST and the first two moments of W_k, and gave a recursive relation for calculating higher moments. Shanthikumar [18] further utilized the level-crossing analysis to present a conservation identity for $M/G/1$ queues with server vacations. Levy and Kleinrock [14] studied the $M/G/1$ queue with a start-up delay and showed its similarities to the multiple-vacation $M/G/1$ system. Recently, Doshi [5] provided a methodological overview of various stochastic processes modeled as queueing systems with server vacations. He indicates the relationship between priority queueing models and vacation models and goes over a variety of techniques used to study them.

Several authors (e.g., Fuhrmann [6], Fuhrmann and Cooper [7], Doshi [4], Gelenbe and Iasnogorodski [9]) have studied the single-class $M/G/1$ and $GI/G/1$ queues with server vacations concentrating on "decomposition" results, i.e., on the phenomenon that "under fairly general conditions the waiting time of an arbitrary customer, in steady state, is distributed as the sum of two independent random variables: one corresponding to the waiting time without vacations and the other to the stationary forward recurrence time of the vacation" (Doshi [4]). These recent studies extend previous results for the $M/G/1$ queue by Cooper [2], who first obtained the decomposition result, and by Levy and Yechiali [15], who first identified the second term as the forward recurrence time.

In this article we develop a unified comprehensive framework for the analysis of priority-$M/G/1$ queues with server vacations. We study four models:

(1) NPMV: Nonpreemptive multiple vacations.
(2) NPSV: Nonpreemptive single vacation.
(3) PRMV: Preemptive-resume multiple vacations.
(4) PRSV: Preemptive-resume single vacation.

General results for all four models (and eventually others) are derived in Section 2. Steady-state probabilities are calculated in Section 3. The Laplace-Stieltjes transform of the waiting time, W_k of a class-k customer and its first two moments are calculated in Section 4. Finally, a unified probabilistic representation related to the first two moments of W_k is presented in Section 5.

2. DEFINITIONS, NOTATION, AND GENERAL RESULTS

We consider a priority-$M/G/1$ queue with n classes of customers, where the Poisson arrival rate of customers of class i is λ_i ($i = 1,2, \ldots ,n$), their service times are V_i's, and customers of class i have a priority (preemptive or non-preemptive) over customers of class j iff $i < j$. In addition, the server from time to time takes a so-called vacation. Two vacation models are considered.

In the *multiple-vacation* variant "the server works continuously as long as there is at least one customer in the system. When the server finishes serving a customer and finds the system empty, it goes away for a random length of time, U, called *vacation*. At the end of the vacation the server returns and begins to serve those customers, if any, who have arrived during the vacation. If the server finds no customers waiting at the end of a vacation, it immediately takes another vacation, and continues in this manner until it finds at least one waiting customer upon return from a vacation" (Cooper [3]). In the *single-vacation* case the server takes exactly one vacation at the end of each busy period. That is, if upon return from a vacation there are no new customers in the system, the server stays *idle* until the first arrival of a new customer, and only then it starts a (regular) busy period. Obviously, if customers arrive during a vacation, the server starts serving them as soon as its vacation terminates.

Our goal is to derive expressions for the LST and moments of the waiting time of an arbitrary class-k customer in each of the four models specified above. For the purpose of analysis, it is convenient to group the higher-priority and lower-priority classes into two distinct sets, defined by the following indices (see Conway, Maxwell and Miller [1]): (a) The index noting customers which are prior to (above) class-k customers, i.e., with priority index smaller than k. (b) The index noting customers which are inferior to (below) class-k customers, i.e., with priority index greater than k.

Thus, define $\lambda_a = \sum_{i=1}^{k-1} \lambda_i$, $\lambda_b = \sum_{i=k+1}^{n} \lambda_i$, $\lambda = \sum_{i=1}^{n} \lambda_i$, and let V_a and V_b denote the service times of class-a and class-b customers, respectively. Let $G_i(\cdot)$ denote the cumulative distribution function (CDF) of the service time V_i of class-i customers. Then, the CDF's of V_a and V_b are, respectively,

$$G_a(\cdot) = \sum_{i=1}^{k-1} \frac{\lambda_i}{\lambda_a} G_i(\cdot)$$

and

$$G_b(\cdot) = \sum_{i=k+1}^{n} \frac{\lambda_i}{\lambda_b} G_i(\cdot).$$

Also, let $\rho_i = \lambda_i EV_i$, $\rho_a = \lambda_a EV_a = \sum_{i=1}^{k-1} \rho_i$, $\rho_b = \lambda_b EV_b = \sum_{i=k+1}^{n} \rho_i$, $\rho = \sum_{i=1}^{n} \rho_i$, $\sigma_0 = 0$, $\sigma_j = \sum_{i=1}^{j} \rho_i$, $1 \le j \le n$. Note that $\sigma_n = \rho$, $\sigma_{k-1} = \rho_a$ and

$\rho - \sigma_k = \rho_b$. We will often interchange ρ_a with σ_{k-1} and ρ_b with $\rho - \sigma_k$. We also assume that the system is unsaturated, i.e., $\rho < 1$.

For the analysis in the sequel we require certain observations and results. Denote by θ_a the length of time from a moment a class-a customer enters service and no other class-a customers are present, until the first moment when there are no class-a customers in the system. Clearly θ_a is the duration of a busy period in a *standard* $M/G/1$ queue with arrival rate λ_a and service times V_a. Consequently, the LST of θ_a and its mean are given by (Cooper [3], p. 230)

$$\tilde{\theta}_a(s) = \tilde{V}_a(s + \lambda_a - \lambda_a\tilde{\theta}_a(s)), \qquad E\theta_a = EV_a/(1 - \rho_a), \tag{1}$$

where $\tilde{X}(s) \equiv E[e^{-sX}]$ is the LST of a random variable X.

Let V_{ak} denote the length of time from the moment a class-k customer enters service (clearly, no class-a customer is present) until the first moment after his service completion when there are no class-a customers in the system. It is easy to see that V_{ak} is a delay cycle, with delay V_k, in a standard $M/G/1$ queue with class-a customers only. That is, V_{ak} is the length of time the server is continuously busy in an $M/G/1$ queue with arrival rate λ_a and service time V_a, where the server starts with a service of duration V_k (= the delay) of a class-k customer (no type-a customers are present initially), and continues with service of type-a customers only, until none of them is present. Hence, the LST and mean of V_{ak} are given by (see Conway, Maxwell and Miller [1])

$$\tilde{V}_{ak}(s) = \tilde{V}_k(s + \lambda_a - \lambda_a\tilde{\theta}_a(s)), \qquad EV_{ak} = EV_k/(1 - \rho_a). \tag{2}$$

Observe that V_{ak} may also represent the time from a service initiation of a class-k customer until the first moment another class-k customer (if present) *may* enter service. (This duration is called "completion time" by Gaver [8].) Therefore, we consider V_{ak} as a *generalized* service time of a class-k customer, and set $\rho_{ak} \equiv \lambda_k EV_{ak} = \rho_k/(1 - \rho_a)$.

Similarly to the definition of V_{ak} we define two key *delay cycles:*

(i) T_{ak} cycle = a delay cycle for which the delay is T (no class-a or class-k customers are waiting in line initially), and the customers served thereafter are from classes 1 to k (i.e., types a and k) only. The cycle terminates as soon as no more customers of type a or k are present. Clearly, $E[T_{ak}$ cycle$] = ET/(1 - (\rho_a + \rho_k))$.

(ii) T_a cycle = a delay cycle starting with a delay T and the customers being served thereafter are from type a only (i.e., classes $1,2, \ldots ,k - 1$). T_a is the length of time from the beginning of the delay T (where no type-a customers are waiting in queue) until the first moment thereafter that a class-k customer may enter service. Similarly to (1) and (2) we have

$$\tilde{T}_a(s) = \tilde{T}(s + \lambda_a - \lambda_a\tilde{\theta}_a(s)), \qquad E[T_a] = ET/(1 - \rho_a). \tag{3}$$

In our models, each of the variables U, V_a, V_k, or V_b may serve as a delay T, generating a T_a delay cycle, which itself constitutes the initial phase in a T_{ak} delay cycle.

It is important to see that whenever the server is *not idle* the system is within some T_{ak} delay cycle. We just have to distinguish between the various cycles: U cycle, V_a cycle, V_k cycle and V_b cycle. A U cycle is a T_{ak} delay cycle with a delay U. Such a cycle starts with a regular server vacation U, continues for a period of time where only type-a customers are being served (this is the duration of the corresponding T_a cycle), and ends with a period of time where customers of both types k and a are being served. (Clearly, the duration between two consecutive services of class-k customers is V_{ak}.) Similarly, V_a cycle, V_k cycle, or V_b cycle is a T_{ak} delay cycle with delay V_a, V_k, or V_b, respectively.

We are now in a position to present the *main idea* of our analysis. Consider an arbitrary class-k customer C_k who arrives during some T_{ak} cycle. As pointed out above, the initial phase of this T_{ak} cycle is a T_a delay cycle, and the time intervals between two consecutive services of class-k customers are V_{ak}. Hence, as far as waiting times are considered, C_k may view the process as a *nonpriority* $M/G/1$ queue with *multiple* server vacations, where the arrival rate is λ_k, service times are V_{ak} (yielding traffic intensity $\rho_{ak} = \lambda_k EV_{ak}$), and the "vacation period" opening the T_{ak} cycle is T_a.

This key observation enables us to bring into the analysis the following known results concerning the multiple-vacation, nonpriority $M/G/1$ queue. It has been shown by Cooper [2] and by Levy and Yechiali [15] that for the multiple-vacation nonpriority $M/G/1$ queue with arrival rate λ_0, service time V_0, vacation duration U_0, and traffic intensity $\rho_0 = \lambda_0 EV_0$, the LST and first two moments of the waiting time W of an arbitrary customer are given by

$$\tilde{W}(s) = \frac{(1 - \rho_0)(1 - \tilde{U}_0(s))}{[\lambda_0 \tilde{V}_0(s) - \lambda_0 + s]EU_0}, \tag{4a}$$

$$EW = \frac{\lambda_0 EV_0^2}{2(1 - \rho_0)} + \frac{EU_0^2}{2EU_0}, \tag{4b}$$

$$EW^2 = \frac{\lambda_0 EV_0^2}{1 - \rho_0} EW + \frac{\lambda_0 EV_0^3}{3(1 - \rho_0)} + \frac{EU_0^3}{3EU_0}. \tag{4c}$$

Using the above key observation together with Eqs. (4) yields

$$E[e^{-sW_k}|T_{ak} \text{ cycle}] = \frac{(1 - \rho_{ak})(1 - \tilde{T}_a(s))}{[\lambda_k \tilde{V}_{ak}(s) - \lambda_k + s]ET_a}, \tag{5a}$$

$$E[W_k|T_{ak} \text{ cycle}] = \frac{\lambda_k EV_{ak}^2}{2(1 - \rho_{ak})} + \frac{ET_a^2}{2ET_a}, \tag{5b}$$

$$E[W_k^2|T_{ak} \text{ cycle}] = \frac{\lambda_k EV_{ak}^2}{1 - \rho_{ak}} E[W_k|T_{ak} \text{ cycle}] + \frac{\lambda_k EV_{ak}^3}{3(1 - \rho_{ak})} + \frac{ET_a^3}{3ET_a}. \tag{5c}$$

From the LST of T_a in (3) one can readily obtain

$$\begin{aligned} ET_a^2 &= \frac{ET^2}{(1 - \rho_a)^2} + \frac{\lambda_a EV_a^2 ET}{(1 - \rho_a)^3}, \\ ET_a^3 &= \frac{ET^3}{(1 - \rho_a)^3} + \frac{3\lambda_a ET^2 \cdot EV_a^2}{(1 - \rho_a)^4} + \frac{\lambda_a ET \cdot EV_a^3}{(1 - \rho_a)^4} + \frac{3(\lambda_a EV_a^2)^2 ET}{(1 - \rho_a)^5}. \end{aligned} \tag{6}$$

Furthermore, EV_{ak}^2 and EV_{ak}^3 can be obtained by substituting V_k in place of T in (6) (since then $V_{ak} = T_a$).

Making this substitution and inserting in (5) the expression for $\tilde{T}_a(s)$ from (3), and the expression for $\tilde{V}_{ak}(s)$ from (2), one gets (after some calculations)

$$E[e^{-sW_k}|T_{ak}\text{ cycle}] = \frac{(1-\rho_a-\rho_k)[1-\tilde{T}(s+\lambda_a-\lambda_a\tilde{\theta}_a(s))]}{[\lambda_k\tilde{V}_k(s+\lambda_a-\lambda_a\tilde{\theta}_a(s))-\lambda_k+s]ET}, \tag{7a}$$

$$E[W_k|T_{ak}\text{ cycle}] = \frac{\lambda_k EV_k^2+\lambda_a EV_a^2}{2(1-\rho_a)(1-\rho_a-\rho_k)} + \frac{ET^2}{2(1-\rho_a)ET}, \tag{7b}$$

$$E[W_k^2|T_{ak}\text{ cycle}] = \left[\frac{\lambda_k EV_k^2+\lambda_a EV_a^2}{(1-\rho_a)(1-\rho_a-\rho_k)} + \frac{\lambda_a EV_a^2}{(1-\rho_a)^2}\right] \times E[W_k|T_{ak}\text{ cycle}] + \frac{\lambda_k EV_k^3+\lambda_a EV_a^3}{3(1-\rho_a)^2(1-\rho_a-\rho_k)} + \frac{ET^3}{3(1-\rho_a)^2 ET}. \tag{7c}$$

It should be emphasized that results (5) and (7) hold for *both* the non-preemptive and preemptive-resume regimes as the variables V_{ak}, T_{ak} cycle, and T_a are the *same* for both queue disciples.

From the point of view of a class-k customer any point in time is either within some T_{ak} cycle (where $T = U, V_a, V_k$, or V_b), or within a *nondelay* time X_0 in which an arriving class-k customer enters service immediately upon arrival. Hence, for $T = U, V_a, V_k, V_b$,

$$\begin{aligned} \tilde{W}_k(s) &= \sum_T P[T_{ak}\text{ cycle}]\cdot E[e^{-sW_k}|T_{ak}\text{ cycle}] + P[X_0], \\ EW_k &= \sum_T P[T_{ak}\text{ cycle}]\cdot E[W_k|T_{ak}\text{ cycle}], \\ EW_k^2 &= \sum_T P[T_{ak}\text{ cycle}]\cdot E[W_k^2|T_{ak}\text{ cycle}], \end{aligned} \tag{8}$$

where $P[T_{ak}$ cycle$]$ is the probability that the system is within a specific T_{ak} cycle and $P[X_0]$ is the probability that the server is within a nondelay period.

Thus, in order to complete the calculation of $\tilde{W}_k(s)$, EW_k, and EW_k^2, all that remains to do is to evaluate in each model the steady-state probabilities $P[T_{ak}$ cycle$]$ for $T = U, V_a, V_k, V_b$, and the probability $P[X_0]$. It is convenient to use the following notation:

$$\Pi_a = P[V_a\text{ cycle}],\qquad \Pi_k = P[V_k\text{ cycle}],\qquad \Pi_b = P[V_b\text{ cycle}],$$
$$\Pi_u = P[U\text{ cycle}],\qquad \Pi_0 = P[X_0].$$

Note that some of the probabilities may *vanish* in certain cases, as some of the cycles may become irrelevant.

3. CALCULATION OF THE STEADY-STATE PROBABILITIES

There are certain relations between the above probabilities that are general to all cases. It is clear that $\Pi_b = 0$ for the two *preemptive-resume* models, since in these cases an arriving customer from clases $1, \ldots, k$ preempts any class-b customer at service. Thus V_b cycle is irrelevant.

In the *non-preemptive* cases, we have $\Pi_b = \rho_b/(1 - \rho_a - \rho_k) = \rho_b(1 - \sigma_k)$, where $\sigma_k = \rho_a + \rho_k$. This follows since each arriving class-b customer generates a V_b cycle whose mean duration is $E[V_b \text{ cycle}] = EV_b/(1 - \rho_a - \rho_k)$, and the mean number of V_b cycles in a unit of time is λ_b. Define P_u = the probability that the server is on vacation, and P_0 = the probability that the server is idle, but *not* on vacation.

From Levy and Yechiali [15], in all cases $P_0 + P_u = 1 - \rho$. Since $E[U \text{ cycle}] = EU/(1 - \rho_a - \rho_k)$, it is clear that $\Pi_u = P_u/(1 - \rho_a - \rho_k) = (1 - \rho - P_0)/(1 - \sigma_k)$.

It is obvious that for the *multiple*-vacations cases $P_0 = 0$, while in the *single*-vacation cases it has been shown by Levy and Yechiali that

$$P_0 = \frac{(1 - \rho)\tilde{U}(\lambda)}{\tilde{U}(\lambda) + \lambda EU}.$$

Furthermore, it is easy to see that in both non-preemptive cases $\Pi_0 = P_0$ where in the *preemptive-resume* cases $\Pi_0 = P_0 + \rho_b$.

It follows that in the *non-preemptive* cases

$$\begin{aligned}\Pi_b + \Pi_u + \Pi_0 = \rho_b/(1 - \sigma_k) + (1 - \rho - P_0)/(1 - \sigma_k) + P_0 \\ = 1 - P_0\sigma_k/(1 - \sigma_k),\end{aligned}$$

and in the *preemptive-resume* cases

$$\Pi_b + \Pi_u + \Pi_0 = 0 + \frac{1 - \rho - P_0}{1 - \sigma_k} + P_0 + \rho_b = 1 - \frac{(P_0 + \rho_b)\sigma_k}{1 - \sigma_k}.$$

Therefore, in all four models

$$\Pi_a + \Pi_k = 1 - (\Pi_b + \Pi_u + \Pi_0) = \frac{\sigma_k \Pi_0}{1 - \sigma_k}.$$

Now, for $j = a,k$, let $A_j = [\lambda_j/(\lambda_a + \lambda_k)][EV_j/(1 - \rho_a - \rho_k)]$. As the expected length of a V_j cycle is $EV_j/(1 - \rho_a - \rho_k)$, it follows that the probability of the system being within a V_j cycle given that it is within a V_a cycle or a V_k cycle is $A_j/(A_a + A_k) = \rho_j/(\rho_a + \rho_k)$. Thus, unconditioning, we finally have $\Pi_j = [\rho_j/(\rho_a + \rho_k)](\Pi_a + \Pi_k) = \rho_j \Pi_0/(1 - \sigma_k)$, $j = a,k$. We summarize the above results in Table 1.

4. FORMULAS FOR $\tilde{W}_k(s)$, EW_k, AND EW_k^2

Using Eqs. (7) and (8), together with the steady-state probabilities appearing in Table 1, we obtain explicit formulas for each of the four models studied in the preceding sections. The results concerning the NPMV and NPSV models (Section 4.1 and 4.2 below) have been obtained by Shanthikumar [17] using level-crossing arguments. The results concerning the preemptive-resume cases (Sections 4.3 and 4.4 below) are *new*.

Table 1. Steady-state probabilities.

	Π_u	Π_0	Π_a	Π_k	Π_b	P_0
NPMV	$\frac{1-\rho}{1-\sigma_k}$	0	0	0	$\frac{\rho_b}{1-\sigma_k}$	0
NPSV	$\frac{1-\rho}{1-\sigma_k}\frac{\lambda EU}{\tilde{U}(\lambda)+\lambda EU}$	$\frac{(1-\rho)\tilde{U}(\lambda)}{\tilde{U}(\lambda)+\lambda EU}$	$\frac{\Pi_0\rho_a}{1-\sigma_k}$	$\frac{\Pi_0\rho_k}{1-\sigma_k}$	$\frac{\rho_b}{1-\sigma_k}$	$\frac{(1-\rho)\tilde{U}(\lambda)}{\tilde{U}(\lambda)+\lambda EU}$
PRMV	$\frac{1-\rho}{1-\sigma_k}$	ρ_b	$\frac{\Pi_0\rho_a}{1-\sigma_k}$	$\frac{\Pi_0\rho_k}{1-\sigma_k}$	0	0
PRSV	$\frac{1-\rho}{1-\sigma_k}\frac{\lambda EU}{\tilde{U}(\lambda)+\lambda EU}$	$\frac{(1-\rho)\tilde{U}(\lambda)}{\tilde{U}(\lambda)+\lambda EU}+\rho_b$	$\frac{\Pi_0\rho_a}{1-\sigma_k}$	$\frac{\Pi_0\rho_k}{1-\sigma_k}$	0	$\frac{(1-\rho)\tilde{U}(\lambda)}{\tilde{U}(\lambda)+\lambda EU}$

4.1 NPMV

$$\tilde{W}_k(s) = \frac{\frac{1-\rho}{EU}[1 - \tilde{U}(s + \lambda_a - \lambda_a\tilde{\theta}_a(s))] + \lambda_b[1 - \tilde{V}_b(s + \lambda_a - \lambda_a\tilde{\theta}_a(s))]}{\lambda_k\tilde{V}_k(s + \lambda_a - \lambda_a\tilde{\theta}(s)) - \lambda_k + s},$$

$$EW_k = \frac{\sum_{i=1}^{n} \lambda_i EV_i^2 + (1-\rho)(EU^2/EU)}{2(1-\sigma_k)(1-\sigma_{k-1})},$$

$$EW_k^2 = \left[\left(\frac{\sum_{i=1}^{k}\lambda_i EV_i^2}{1-\sigma_k} + \frac{\sum_{i=1}^{k-1}\lambda_i EV_i^2}{1-\sigma_{k-1}}\right) EW_k + \frac{\sum_{i=1}^{n}\lambda_i EV_i^3 + (1-\rho)(EU^3/EU)}{3(1-\sigma_k)(1-\sigma_{k-1})}\right] \frac{1}{1-\sigma_{k-1}}.$$

4.2 NPSV

$$\tilde{W}_k(s) = \frac{[(1-\rho)/(\tilde{U}(\lambda) + \lambda EU)][\lambda(1 - \tilde{U}(s + \lambda_a - \lambda_a\tilde{\theta}_a(s))) + \tilde{U}(\lambda)(\lambda_a(1 - \tilde{\theta}_a(s)) + s)]}{\lambda_k\tilde{V}_k(s + \lambda_a - \lambda_a\tilde{\theta}_a(s)) - \lambda_k + s} + \frac{\lambda_b[1 - \tilde{V}_b(s + \lambda_a - \lambda_a\tilde{\theta}_a(s))]}{\lambda_k\tilde{V}_k(s + \lambda_a - \lambda_a\tilde{\theta}_a(s)) - \lambda_k + s},$$

$$EW_k = \frac{\sum_{i=1}^{n}\lambda_i EV_i^2 + (1-\rho)[\lambda EU/(\tilde{U}(\lambda) + \lambda EU)](EU^2/EU)}{2(1-\sigma_k)(1-\sigma_{k-1})},$$

$$EW_k^2 = \left[\left(\frac{\sum_{i=1}^{k}\lambda_i EV_i^2}{1-\sigma_k} + \frac{\sum_{i=1}^{k-1}\lambda_i EV_i^2}{1-\sigma_{k-1}}\right) EW_k + \frac{\sum_{i=1}^{n}\lambda_i EV_i^3 + (1-\rho)[\lambda EU/(\tilde{U}(\lambda) + \lambda EU)](EU^3/EU)}{3(1-\sigma_k)(1-\sigma_{k-1})}\right] \times \frac{1}{1-\sigma_{k-1}}.$$

4.3 PRMV

$$\tilde{W}_k(s) = \frac{((1-\rho)/EU)[1-\tilde{U}(s+\lambda_a-\lambda_a\tilde{\theta}_a(s))] + \rho_b[\lambda_a(1-\tilde{\theta}_a(s))+s]}{\lambda_k\tilde{V}_k(s+\lambda_a-\lambda_a\tilde{\theta}(s))-\lambda_k+s},$$

$$EW_k = \frac{\sum_{i=1}^{k}\lambda_i EV_i^2 + (1-\rho)(EU^2/EU)}{2(1-\sigma_k)(1-\sigma_{k-1})},$$

$$EW_k^2 = \left[\left(\frac{\sum_{i=1}^{k}\lambda_i EV_i^2}{1-\sigma_k} + \frac{\sum_{i=1}^{k-1}\lambda_i EV_i^2}{1-\sigma_{k-1}}\right)EW_k + \frac{\sum_{i=1}^{k}\lambda_i EV_i^3 + (1-\rho)(EU^3/EU)}{3(1-\sigma_k)(1-\sigma_{k-1})}\right]\frac{1}{1-\sigma_{k-1}}.$$

4.4 PRSV

$$\tilde{W}_k(s) = \frac{[\lambda(1-\rho)/(\tilde{U}(\lambda)+\lambda EU)][1-\tilde{U}(s+\lambda_a-\lambda_a\tilde{\theta}_a(s))] + [(1-\rho)\tilde{U}(\lambda)/(\tilde{U}(\lambda)+\lambda EU)+\rho_b][\lambda_a(1-\tilde{\theta}_a(s))+s]}{\lambda_k\tilde{V}_k(s+\lambda_a-\lambda_a\tilde{\theta}_a(s))-\lambda_k+s}$$

$$EW_k = \frac{\sum_{i=1}^{k}\lambda_i EV_i^2 + (1-\rho)[\lambda EU/(\tilde{U}(\lambda)+\lambda EU)](EU^2/EU)}{2(1-\sigma_k)(1-\sigma_{k-1})},$$

$$EW_k^2 = \left[\left(\frac{\sum_{i=1}^{k}\lambda_i EV_i^2}{1-\sigma_k} + \frac{\sum_{i=1}^{k-1}\lambda_i EV_i^2}{1-\sigma_{k-1}}\right)EW_k + \frac{\sum_{i=1}^{k}\lambda_i EV_i^3 + (1-\rho)[\lambda EU/(\tilde{U}(\lambda)+\lambda EU)](EU^3/EU)}{3(1-\sigma_k)(1-\sigma_{k-1})}\right] \times \frac{1}{1-\sigma_{k-1}}.$$

Note that by setting $\lambda_k = \lambda_0$, $\lambda_a = \lambda_b = 0$, $V_k = V_0$, and $U = U_0$ in the expressions for $\tilde{W}_k(s)$, EW_k, and EW_k^2 developed in this section for the NPMV case, one readily obtains the corresponding results for the *single-class* multiple vacation model [i.e., Eqs. (4a)–(4c) above]. Making the same substitutions in the expressions derived for the NPSV case yields the corresponding results obtained by Levy and Yechiali [15] for the single-class single-vacation variant.

5. PROBABILISTIC REPRESENTATION OF THE RESULTS

The results obtained in Section 4 for the first two moments of W_k may be given a unifying probabilistic representation, using the notion of residual lifetime.

Define R_k as the remaining *net* service time of the customer being served upon arrival of a class-k customer (provided that the former is not preempted by the latter), *or* as the residual time of a vacation. Then, in all four models above, the first two moments of W_k may be written as follows:

$$EW_k = \frac{ER_k}{(1 - \sigma_k)(1 - \sigma_{k-1})}, \tag{9}$$

$$EW_k^2 = \left[\left(\frac{\sum_{i=1}^{k} \lambda_i EV_i^2}{1 - \sigma_k} + \frac{\sum_{i=1}^{k-1} \lambda_i EV_i^2}{1 - \sigma_{k-1}}\right) EW_k + \frac{ER_k^2}{(1 - \sigma_k)(1 - \sigma_{k-1})}\right] \frac{1}{1 - \sigma_{k-1}}. \tag{10}$$

This follows since, in the *non-preemptive* models,

$$ER_k = \sum_{i=1}^{n} \rho_i \frac{EV_i^2}{2EV_i} + P_u \frac{EU^2}{2EU}, \tag{11}$$

$$ER_k^2 = \sum_{i=1}^{n} \rho_i \frac{EV_i^3}{3EV_i} + P_u \frac{EU^3}{3EU}, \tag{12}$$

while in the preemptive-resume models the summations in (11) and (12) are only up to k, as class-b customers are preempted by customers of classes k and a.

The above expressions for ER_k and ER_k^2 are self explanatory as ρ_i ($i = 1,2, \ldots, n$) or P_u is the probability that at a moment of arrival of a class-k customer the server is serving a type-i customer or it is on vacation, respectively. $EV_i^2/(2EV_i)$ and $EV_i^3/(3EV_i)$ are the mean residual service time of a class-i customer and its second moment, respectively. Similarly, $EU^2/(2EU)$ and $EU^3/(3EU)$ are the first and second moments of the residual time of a vacation.

As pointed out by Heyman [10], Levy and Yechiali [15], Doshi [5], and others, vacation durations may be interpreted as service times of lowest-priority customers who are always available for service. Under this interpretation—and considering the *non-preemptive* cases—Eq. (9) is equivalent to Eq. (3.31) in Kleinrock ([13], p. 121), where ER_k replaces W_0, the average delay to a newly arriving type-k customer due to the customer found in service. It is also easy to check that the conservation law regarding mean waiting times, i.e., $\sum_{k=1}^{n} \rho_k EW_k = \rho ER_k/(1 - \rho)$, holds naturally in these cases (see Eq. (3.16) in Kleinrock [13]).

REFERENCES

[1] Conway, R. W., Maxwell, W. L., and Miller, L. W., *Theory of Scheduling,* Addison-Wesley, Reading, MA, 1967.

[2] Cooper, R. B., "Queues Served in Cyclic Order: Waiting Times," *Bell System Technical Journal,* **49,** 399–413 (1970).

[3] Cooper, R. B., *Introduction to Queueing Theory,* 2nd ed., North-Holland, Amsterdam, 1981.

[4] Doshi, B. T., "A Note on Stochastic Decomposition in $GI/G/1$ Queue with Vacations or Set-Up Times," *Journal of Applied Probability,* **22,** 419–728 (1985).

[5] Doshi, B. T., "Queueing Systems with Vacations—A Survey," *Queueing Systems,* **1,** 29–66 (1986).
[6] Fuhrmann, S. W., "A Note on the $M/G/1$ Queue with Server Vacations," *Operations Research,* **32,** 1368–1373 (1984).
[7] Fuhrmann, S. W. and Cooper, R. B., "Stochastic Decomposition in the $M/G/1$ Queue with Generalized Vacations," *Operations Research,* **33,** 1117–1129 (1985).
[8] Gaver, D. P. Jr., "A Waiting Line with Interrupted Service, including Priorities," *Journal of the Royal Statistical Society, Series B,* **24,** 73–90 (1962).
[9] Gelenbe, E., and Iasnogorodski, R., "A Queue with Server of Walking Type (Autonomous Service)," *Annales de l'Institut Henri Poincaré, Section B: Calcul des Probabilités et Statistique, 16,* 63–73 (1980).
[10] Heyman, D. P., "A Priority Queueing System with Server Interference," *SIAM Journal of Applied Mathematics,* **17,** 74–82 (1969).
[11] Keilson, J., "Queues Subject to Service Interruptions," *Annals of Mathematical Statistics,* **33,** 1314–1322 (1962).
[12] Kella, O. and Yechiali, U., "Waiting Times in the Non-Preemptive Priority $M/M/c$ Queue," *Stochastic Models (Communications in Statistics),* **1,** 257–262 (1985).
[13] Kleinrock, L., *Queueing Systems,* Wiley, New York, 1976, Vol. II.
[14] Levy, H. and Kleinrock, L., "A Queue with Starter and a Queue with Vacations: Delay Analysis by Decomposition," *Operations Research,* **34,** 426–436 (1986).
[15] Levy, Y. and Yechiali, U., "Utilization of Idle Time in an $M/G/1$ Queueing System," *Management Science,* **22,** 202–211 (1975).
[16] Scholl, M. and Kleinrock, L., "On the $M/G/1$ Queue with Rest Periods and Certain Service-Independent Queueing Disciplines," *Operations Research,* **31,** 705–719 (1983).
[17] Shanthikumar, J. G., "Analysis of Priority Queues with Server Control," *Opsearch,* **21,** 183–192 (1984).
[18] Shanthikumar, J. G., "Level Crossing Analysis of Priority Queues and a Conservation Identity for Vacation Models," *Technical Report,* University of California, Berkeley, December 1986.

Manuscript received September 6, 1985
Revised manuscript received October 16, 1986
Revised manuscript received March 13, 1987
Accepted March 23, 1987