# Distribution of the Workload in Multiclass Queueing Systems with Server Vacations

**Hideaki Takagi**
*IBM Research, Tokyo Research Laboratory, No. 36 Kowa Building,
5-19 Sanban-cho, Chiyoda-ku, Tokyo 102, Japan*

**Tetsuya Takine**
*Department of Applied Mathematics and Physics, Faculty of Engineering,
Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606, Japan*

**Onno J. Boxma**
*Centre for Mathematics and Computer Science, P. O. Box 4079,
1009 AB Amsterdam, The Netherlands*

The steady-state workload at an arbitrary time is considered for several single-server queueing systems with nonpreemptive services for multiple classes of customers (arriving according to Poisson processes) and server vacation (switchover) times. The distribution of the workload at an arbitrary point during the vacation period is obtained for systems with setup times, and for polling systems with exhaustive, gated, or globally gated service disciplines. From the stochastic decomposition property, this workload is added to the workload in the corresponding M/G/1 system without vacations to give the workload at an arbitrary time in vacation systems. Dependence of the workload distribution on the vacation parameters is studied.

## 1. INTRODUCTION

The *workload*, also called the *backlog*, *unfinished work*, or *work in system*, in a queueing system is defined as the sum of the remaining service times of all customers in the system. This article is concerned with the distribution of the steady-state workload in several single-server queueing systems with $P$ classes of customers (arriving according to independent Poisson arrival processes) and server vacation (switchover) times. Note that the workload indicates the system-wide congestion, while the queue size and the waiting time are interesting for customers of each class. Throughout the article, we assume that service times and vacation times are independent random variables. We focus on nonpreemptive service disciplines that use only information about the customer class in selecting the customer to serve. It is also assumed that the service discipline and the vacation process do not affect the amount of service time given to any customer.

*Naval Research Logistics*, Vol. 39, pp. 41–52 (1992)
CCC 0894-069X/92/010041-12$04.00

For a broad category of multiclass queueing systems with server vacations, including the above-mentioned systems considered in this paper, Boxma and Groenendijk [2] (see also Boxma [1]) established the following *work decomposition* result. The steady-state workload $U$ is distributed as the sum of the steady-state workload $U_{M/G/1}$ in the corresponding $M/G/1$ system without vacations and the steady-state workload $Y$ in the original system at an arbitrary time during a vacation period:

$$U \overset{\text{distr}}{=} U_{M/G/1} + Y. \tag{1}$$

Furthermore, $U_{M/G/1}$ and $Y$ are independent. Let $\lambda_p$, $b_p$, $b_p^{(2)}$, and $B_p^*(s)$ be the Poisson arrival rate, the mean, the second moment, and the Laplace-Stieltjes transform (LST) of the distribution function (DF) for the service time, respectively, of a customer of class $p$, where $p = 1, 2, \ldots, P$. In addition, if $U^*(s)$, $U_{M/G/1}^*(s)$, and $Y^*(s)$ denote the LST of the DF for $U$, $U_{M/G/1}$, and $Y$, respectively, we have, for $s \geq 0$,

$$U^*(s) = U_{M/G/1}^*(s)Y^*(s), \tag{2}$$

where

$$U_{M/G/1}^*(s) = \frac{(1-\rho)s}{s - \lambda + \lambda B^*(s)}, \tag{3}$$

$$\lambda \triangleq \sum_{p=1}^{P} \lambda_p; \qquad \rho \triangleq \sum_{p=1}^{P} \lambda_p b_p; \qquad B^*(s) \triangleq \sum_{p=1}^{P} \frac{\lambda_p}{\lambda} B_p^*(s). \tag{4}$$

The purpose of this paper is to give $Y^*(s)$ for systems with setup times (Section 3) and for polling systems (Section 4), thus obtaining the LST of the DF for the workload in those systems by (2). For general systems with nonpreemptive service, the evaluation of the mean workload $E[U]$ leads to the so-called *pseudoconservation law* with respect to the traffic-intensity-weighted sum of the mean waiting times for each class of customers (see Section 2), which has been studied for several polling systems. However, as far as the authors know, no results have been published for the *distribution* of the workload. Our results are summarized in Theorems 1 and 2. Concluding remarks are given in Section 5.

## 2. PSEUDOCONSERVATION LAWS

In this section, we first present a general form of the pseudoconservation law. It is followed by examples of systems with setup times and polling systems.

For multiclass systems with a nonpreemptive service discipline that does not distinguish customers on the basis of their service times [7, Sec. 6.2], we have

$$E[U] = \frac{\lambda b^{(2)}}{2} + \sum_{p=1}^{P} \rho_p E[W_p], \tag{5}$$

where

$$\rho_p \overset{\Delta}{=} \lambda_p b_p; \qquad b^{(2)} \overset{\Delta}{=} \sum_{p=1}^{P} \frac{\lambda_p}{\lambda} b_p^{(2)} \tag{6}$$

and $E[W_p]$ is the mean waiting time of a customer of class $p$. It follows from (1) and (3) that

$$\sum_{p=1}^{P} \rho_p E[W_p] = \frac{\rho \lambda b^{(2)}}{2(1-\rho)} + E[Y], \tag{7}$$

which is called the pseudoconservation law. For systems without vacations, we have $E[Y] \equiv 0$, and (7) reduces to Kleinrock's *conservation law* [8, Sec. 5.2, 9] as this intensity-weighted sum of the mean waiting times can never change no matter how the service discipline may use the class information. For systems with vacations, $E[Y]$ depends on the structure of the vacation mechanism.

In a system with setup times, a setup time $S_p$ is required prior to the busy period when a customer of class $p$ arrives during an idle period in the system, where $p = 1, 2, \ldots, P$. During the busy period, scheduling of classes to serve is arbitrary, for example, first-come-first-served (FCFS), nonpreemptive priority, or exhaustive-service polling [6]. Many priority queues with setup times are studied by Takagi [11], who shows, for example, that

$$E[W_p]_{\text{FCFS}} = \frac{(1-\rho)E[S_p]}{1+\lambda E[S]} + \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{\lambda E[S^2] + 2\sum_{k=1}^{P} \rho_k E[S_k]}{2(1+\lambda E[S])}, \tag{8}$$

where $b^{(2)}$ is defined in (6), and

$$E[S^i] \overset{\Delta}{=} \frac{1}{\lambda} \sum_{p=1}^{P} \lambda_p\, E[(S_p)^i], \qquad i = 1, 2, \ldots \tag{9}$$

is the $i$th moment of the setup time aggregated over all classes. For the nonpreemptive priority system with setup times in which class 1 has the highest priority and class $P$ the lowest, we have

$$E[W_p]_{\text{nonpreemptive priority}} = \frac{\lambda b^{(2)}}{2(1-\rho_{p-1}^{+})(1-\rho_p^{+})}$$

$$+ \frac{1-\rho}{(1+\lambda E[S])(1-\rho_{p-1}^{+})} \left[ E[S_p] + \frac{\lambda E[S^2] + 2\sum_{k=1}^{P} \rho_k E[S_k]}{2(1-\rho_p^{+})} \right], \tag{10}$$

where $\rho_p^{+} \overset{\Delta}{=} \Sigma_{k=1}^{p} \rho_k$. Both (8) and (10) satisfy (7) with

$$E[Y]_{\text{setup times}} = \frac{\rho \lambda E[S^2] + 2\sum_{p=1}^{P} \rho_p E[S_p]}{2(1+\lambda E[S])}. \tag{11}$$

In the next section, we derive the LST $Y^*(s)$ of the DF for $Y$ in a system with setup times.

Pseudoconservation laws are studied extensively for *polling systems* [1, 2, 10]. In a polling system, customers of each class are served in cyclic order with finite switchover times. Let $R_p^*(s)$, $r_p$, and $r_p^{(2)}$ be the LST of the DF, the mean, and the second moment, respectively, for the server switchover time from class $p$ to class $p$ mod $P + 1$, where $p = 1, 2, \ldots, P$. (Hereafter, all class indices in polling systems should read in a similar cyclic fashion, although not shown explicitly.) Switchover times are assumed to be independent. In the *exhaustive service* system, the server continues to serve each class until there are no customers of that class in the system. In the *gated service* system, the server continues to serve only those customers of each class that were waiting when the server started its service to that class, while those customers that arrive during this service period are served in the next round. The mean workload $E[Y]$ at an arbitrary time during the switchover times in these polling systems is given by

$$E[Y]_{\text{exhaustive}} = \frac{\rho R^{(2)}}{2R} + \frac{R\left(\rho^2 - \sum_{p=1}^{P} \rho_p^2\right)}{2(1 - \rho)}, \tag{12}$$

$$E[Y]_{\text{gated}} = \frac{\rho R^{(2)}}{2R} + \frac{R\left(\rho^2 + \sum_{p=1}^{P} \rho_p^2\right)}{2(1 - \rho)}, \tag{13}$$

where

$$R = \sum_{p=1}^{P} r_p; \qquad R^{(2)} = \sum_{p=1}^{P} (r_p^{(2)} - r_p^2) + R^2. \tag{14}$$

Note that $R$ and $R^{(2)}$ are the mean and the second moment of the sum of switchover times over one round. In the *globally gated service* system, which was recently proposed and analyzed by Boxma, Levy, and Yechiali [3], in each cycle of the server, only those customers that were found in the system at the start of the cycle (namely, when the server visited class 1) are served. In this system, we have

$$E[Y]_{\text{globally gated}} = \frac{\rho R^{(2)}}{2R} + \frac{R\rho^2}{2(1 - \rho)} + \sum_{p=2}^{P} \rho_p \sum_{j=1}^{p-1} r_j. \tag{15}$$

In this case, $E[Y]$ depends on the ordering of classes.

REMARK: For $P = 1$ (a single-class vacation model), if the waiting time of a customer is independent of the part of the arrival process that occurs after the customer's arrival time (as in the FCFS system), the LST $W^*(s)$ of the DF for the waiting time can be expressed also in the decomposition form [5]

$$W^*(s) = W^*_{M/G/1}(s)\, \chi(1 - s/\lambda), \tag{16}$$

where $\chi(z)$ is the probability generating function of the number of customers in the system at an arbitrary time during a vacation period, and

$$W^*_{M/G/1}(s) = \frac{(1-\rho)s}{s - \lambda + \lambda B^*(s)} = U^*_{M/G/1}(s). \tag{17}$$

Since each customer makes a contribution of its service time to the workload in system during the vacation period, it follows that

$$Y^*(s) = \chi[B^*(s)]. \tag{18}$$

Hence the relation between $U^*(s)$ and $W^*(s)$ is given by (2), (16), (17), and (18). In particular, we get

$$E[W] = E[W]_{M/G/1} + \frac{E[Y]}{\rho}, \tag{19}$$

which leads to

$$E[U] - \rho E[W] = \frac{\lambda b^{(2)}}{2}; \quad E[W] - E[U] = \frac{(1-\rho)E[Y]}{\rho} > 0. \tag{20}$$

## 3. SYSTEMS WITH SETUP TIMES

In a multiclass system with setup times, a vacation period consists of an idle period, exponentially distributed with mean $1/\lambda$, and a setup time $S_p$ if a customer of class $p$ arrives first during the idle period. During a vacation period, the system is in the idle period with probability

$$\frac{1/\lambda}{1/\lambda + E[S]} = \frac{1}{1 + \lambda E[S]}. \tag{21}$$

During the vacation period, the system is in the setup time $S_p$ with probability

$$\frac{E[S]}{1/\lambda + E[S]} \times \frac{\lambda_p E[S_p]}{\lambda\, E[S]} = \frac{\lambda_p E[S_p]}{1 + \lambda E[S]}, \quad p = 1, 2, \ldots, P. \tag{22}$$

Note that the probability generating function of the number of customers that arrive from the beginning of the setup time $S_p$ till an arbitrary time during $S_p$ is given by

$$\frac{1 - S_p^*(\lambda - \lambda z)}{\lambda E[S_p](1 - z)}. \tag{23}$$

Therefore, the LST of the DF for the workload at an arbitrary time during the setup time $S_p$ is given by

$$B_p^*(s)\frac{1 - S_p^*[\lambda - \lambda B^*(s)]}{\lambda E[S_p][1 - B^*(s)]}. \tag{24}$$

Hence we obtain

$$Y^*_{\text{setup times}}(s) = \frac{1}{1 + \lambda E[S]} \times 1 + \sum_{p=1}^{P} \frac{\lambda_p E[S_p]}{1 + \lambda E[S]} B^*_p(s) \frac{1 - S^*_p[\lambda - \lambda B^*(s)]}{\lambda E[S_p][1 - B^*(s)]}$$

$$= \frac{1 - \sum_{p=1}^{P} (\lambda_p/\lambda) B^*_p(s) S^*_p[\lambda - \lambda B^*(s)]}{(1 + \lambda E[S])[1 - B^*(s)]}, \tag{25}$$

from which we get the mean in (11), and the second moment

$$E[Y^2]_{\text{setup times}} = \frac{\lambda^2 b^{(2)} E[S^2]}{2(1 + \lambda E[S])}$$

$$+ \frac{\rho^2 \lambda E[S^3] + 3\rho \sum_{p=1}^{P} \rho_p E[S_p^2] + 3 \sum_{p=1}^{P} \lambda_p b_p^{(2)} E[S_p]}{3(1 + \lambda E[S])}. \tag{26}$$

Note that the LST of the DF for the workload in the system at an arbitrary time is given by (2) using (3) and (25). Thus we have established the following theorem.

THEOREM 1: The LST of the DF for the workload in a multiclass system with setup times at an arbitrary time in equilibrium is given by

$$U^*(s) = \frac{(1 - \rho)s}{s - \lambda + \lambda B^*(s)} \cdot \frac{1 - \sum_{p=1}^{P} (\lambda_p/\lambda)\, B^*_p(s) S^*_p\, [\lambda - \lambda B^*(s)]}{(1 + \lambda E[S])[1 - B^*(s)]}. \qquad \square$$

## 4. POLLING SYSTEMS

For polling systems, we can express $Y^*(s)$ only in terms of a certain function for which the functional equation is known. Let us focus on a moment at which the server completes the service of class $k$, and denote by $U^*_k(s)$ the LST of the DF for the workload $U_k$ in the system at that moment. An arbitrary point in time during switchover times falls in the switchover time from class $k$ to $k + 1$ with probability $r_k/R$, and the LST of the DF for the workload that is newly brought to the system before the arbitrary point during this switchover time is given by

$$\frac{1 - R^*_k[\lambda - \lambda B^*(s)]}{\lambda r_k[1 - B^*(s)]}. \tag{27}$$

Since this workload and $U_k$ are independent, we have

$$Y^*(s) = \sum_{k=1}^{P} \frac{r_k}{R} \frac{1 - R^*_k[\lambda - \lambda B^*(s)]}{\lambda r_k[1 - B^*(s)]} U^*_k(s). \tag{28}$$

Therefore, it remains to determine $U_k^*(s)$ for individual polling systems. Here we consider polling systems with exhaustive, gated, and globally gated service discipline.

In the exhaustive service system, we can express $U_k^*(s)$ in terms of the joint LST of the DF of the successive *station times*. The concept of a station time is introduced by Ferguson and Aminetzah [4] (they call it a *terminal service time*). In the exhaustive service system, the station time $\omega_p$ for class $p$ is defined as the time interval between the successive instants when the server leaves class $p - 1$ and class $p$. In other words, $\omega_p$ consists of the switchover time from class $p - 1$ to class $p$ and the following service period of class $p$. The joint LST of the distributions of $P$ successive station times $\omega_{k-P+1}, \omega_{k-P+2}, \ldots, \omega_k$ is defined by

$$\Omega_k^*(s_1, \ldots, s_P) \overset{\Delta}{=} E\left[\exp\left(-\sum_{i=1}^{P} \omega_{k-P+i}s_i\right)\right] \tag{29}$$

Ferguson and Aminetzah [4] show that it satisfies the equation

$$\begin{aligned}\Omega_k^*(s_1, \ldots, s_P) = R_{k-1}^*(s_P + \lambda_k[1 - \Theta_k^*(s_P)]) \cdot \Omega_{k-1}^*(0, s_1 + \lambda_k[1 - \Theta_k^*(s_P)],\\ s_2 + \lambda_k[1 - \Theta_k^*(s_P)], \ldots, s_{P-1} + \lambda_k[1 - \Theta_k^*(s_P)]),\end{aligned} \tag{30}$$

where $\Theta_k^*(s)$ is the LST of the DF for the length of a busy period in an $M/G/1$ system consisting only of customers of class $k$. It satisfies the equation

$$\Theta_k^*(s) = B_k^*[s + \lambda_k - \lambda_k\Theta_k^*(s)]. \tag{31}$$

In order to find $U_k^*(s)$, we study the set of customers of each class in the system when the server completes the service of class $k$. Because of exhaustive service, there are no customers of class $k$ in the system at that moment. Customers of class $k - 1$ in the system are those that arrived during the station time $\omega_k$. Customers of class $k - 2$ in the system are those that arrived during the station times $\omega_{k-1} + \omega_k$, and so on. Finally, customers of class $k - P + 1$ in the system are those that arrived during $\omega_{k-(P-2)} + \omega_{k-(P-3)} + \cdots + \omega_k$. In other words, only those customers of class $j$, $k - P + 1 \le j \le k - P + i - 1$, that arrived during the station time $\omega_{k-P+i}$ remain in the system when the server leaves class $k$, where $i = 2, 3, \ldots, P$. Hence, using the definition of the joint LST $\Omega_k^*(s_1, \ldots, s_P)$ of the distributions of the station times $\omega_{k-P+1}$, $\omega_{k-P+2}, \ldots, \omega_k$ in (29), we get

$$U_k^*(s) = \Omega_k^*(0, e_{2;k}(s), \ldots, e_{P;k}(s)), \tag{32}$$

where

$$e_{i;k}(s) \overset{\Delta}{=} \sum_{j=k-P+1}^{k-P+i-1} \lambda_j[1 - B_j^*(s)], \qquad i = 2, \ldots, P. \tag{33}$$

Substituting (32) into (28) we get

$$Y^*_{\text{exhaustive}}(s) = \frac{\sum_{k=1}^{P} \{1 - R^*_k[\lambda - \lambda B^*(s)]\}\Omega^*_k(0, e_{2;k}(s), \ldots, e_{P;k}(s))}{R\lambda[1 - B^*(s)]}, \quad (34)$$

where $\Omega^*_k(s_1, \ldots, s_P)$ is the solution to (30).

From the mean station times given by Ferguson and Aminetzah [4], we can derive $E[Y]_{\text{exhaustive}}$ in (12) from (34), as shown in the Appendix. It is also possible to calculate the second moment $E[Y^2]_{\text{exhaustive}}$ from (34) in terms of $\xi_{pq} \triangleq \text{cov}[\omega_p, \omega_q]$; the set of equations for $\{\xi_{pq}; p, q = 1, 2, \ldots, P\}$ is also given in Ferguson and Aminetzah [4]. Note that the first moment $E[Y]_{\text{exhaustive}}$ in (12) depends only on the total switchover time. It turns out, however, that the second moment $E[Y^2]_{\text{exhaustive}}$ depends on the individual switchover times.

In the gated service system, the station time $\omega_p$ for class $p$ is defined as the time interval between the successive instants when the server visits class $p$ and class $p + 1$. The joint LST of the distributions of $P$ successive station times $\omega_{k-P+1}, \omega_{k-P+2}, \ldots, \omega_k$ is again defined by (29). It satisfies the equation [4]

$$\begin{aligned} \Omega^*_k(s_1, \ldots, s_P) &= R^*_k(s_P) \\ &\quad \cdot \Omega^*_{k-1}(\lambda_k[1 - B^*_k(s_P)], s_1 + \lambda_k[1 - B^*_k(s_P)], \ldots, \\ &\qquad s_{P-1} + \lambda_k[1 - B^*_k(s_P)]) \quad (35) \end{aligned}$$

In order to find $U^*_k(s)$, we first consider the set of customers of each class in the system when the server starts the service of class $k$. From the definitions of the gated service and its associated station time, those customers of class $j$, $(k - 1) - P + 1 \leq j \leq (k - 1) - P + i$, that arrived during the station time $\omega_{(k-1)-P+i}$ remain in the system at that moment, where $i = 1, 2, \ldots, P$. In particular, the number of customers of class $k - P = k \bmod P$ in the system equals the number of customers of class $k$ that arrived during the entire cycle time $\omega_{k-P} + \omega_{k-P+1} + \cdots + \omega_{k-1}$. The service period of class $k$ consists of the service times of this number of customers, and those customers that arrive during this service period are added by the time when the server completes the service of class $k$. Therefore, the LST of the DF for the workload in the system when the server leaves class $k$ is given by

$$U^*_k(s) = \Omega^*_{k-1}(f_{1;k}(s), f_{2;k}(s), \ldots, f_{P;k}(s)), \quad (36)$$

where

$$\begin{aligned} f_{i;k}(s) &\triangleq \lambda_k\{1 - B^*_k[\lambda - \lambda B^*(s)]\} \\ &\quad + \sum_{j=k-P+1}^{k-P+i-1} \lambda_j[1 - B^*_j(s)], \qquad i = 1, \ldots, P. \quad (37) \end{aligned}$$

However, using Eq. (35) for $\Omega_k^*(s_1, \ldots, s_P)$, this can be rewritten as

$$U_k^*(s) = \frac{\Omega_k^*(g_{1;k}(s), g_{2;k}(s), \ldots, g_{P;k}(s))}{R_k^*[\lambda - \lambda B^*(s)]}, \tag{38}$$

where

$$g_{i;k}(s) \triangleq \sum_{j=k-P+1}^{k-P+i} \lambda_j[1 - B_j^*(s)], \qquad i = 1, \ldots, P-1$$

$$\triangleq \lambda[1 - B^*(s)], \qquad i = P. \tag{39}$$

Substituting (38) into (28), we obtain

$$Y_{\text{gated}}^*(s) =$$

$$\frac{\sum_{k=1}^{P} \{1 - R_k^*[\lambda - \lambda B^*(s)]\} \, \Omega_k^*(g_{1;k}(s), g_{2;k}(s), \ldots, g_{P;k}(s))/R_k^*[\lambda - \lambda B^*(s)]}{R\lambda[1 - B^*(s)]}, \tag{40}$$

where $\Omega_k^*(s_1, \ldots, s_P)$ is the solution to (35). As in the exhaustive service system, we can derive $E[Y]_{\text{gated}}$ in (13) from (40).

In the globally gated service system, let $C^*(s)$ be the LST of the DF for the length of a polling cycle, that is, the time interval between two successive starts of service for class 1. The functional equation for $C^*(s)$ is given by [3]

$$C^*(s) = C^*[\lambda - \lambda B^*(s)] \prod_{j=1}^{P} R_j^*[\lambda - \lambda B^*(s)]. \tag{41}$$

It is clear that the workload in the system when the server completes the service to class $k$ consists of the following parts: the workload brought by those customers that arrived during the service periods of class 1, 2, . . . , $k$, the workload of customers of class $k+1, \ldots, P$ that are still present since the beginning of the cycle, and the workload brought by those customers that arrived during the switchover times since the beginning of the cycle. Hence we get

$$U_k^*(s) = C^*\left[\sum_{j=1}^{k} \lambda_j[1 - B_j^*(\lambda - \lambda B^*(s))] + \sum_{j=k+1}^{P} \lambda_j[1 - B_j^*(s)]\right]$$

$$\times \prod_{j=1}^{k-1} R_j^*[\lambda - \lambda B^*(s)]. \tag{42}$$

Substituting (42) into (28), we obtain

$$Y^*_{\text{globally gated}}(s) = \frac{\sum_{k=1}^{P} \{1 - R_k^*[\lambda - \lambda B^*(s)]\}}{R\lambda[1 - B^*(s)]} \times C^*\left[\sum_{j=1}^{k} \lambda_j[1 - B_j^*(\lambda - \lambda B^*(s))] + \sum_{j=k+1}^{P} \lambda_j[1 - B_j^*(s)]\right] \prod_{j=1}^{k-1} R_j^*[\lambda - \lambda B^*(s)], \tag{43}$$

where $C^*(s)$ is the solution to (41). Differentiation of (43) readily leads to $E[Y]$ in (15).

Our results for polling system can be summarized as the following.

THEOREM 2: The LST of the DF for the workload in a polling system at an arbitrary time in equilibrium is given by

$$U^*(s) = \frac{(1 - \rho)s}{s - \lambda + \lambda B^*(s)} \cdot Y^*(s),$$

where

$$Y^*(s) = \frac{\sum_{k=1}^{P} \{1 - R_k^*[\lambda - \lambda B^*(s)]\}\Omega_k^*(0, e_{2;k}(s), \ldots, e_{P;k}(s))}{R\lambda[1 - B^*(s)]}$$

for the exhaustive service model, where $\Omega_k^*(s_1, \ldots, s_P)$ is the solution to (30),

$$Y^*(s) = \frac{\sum_{k=1}^{P} \{1 - R_k^*[\lambda - \lambda B^*(s)]\}\Omega_k^*(g_{1;k}(s), g_{2;k}(s), \ldots, g_{P;k}(s))/R_k^*[\lambda - \lambda B^*(s)]}{R\lambda[1 - B^*(s)]}$$

for the gated service model, where $\Omega_k^*(s_1, \ldots, s_P)$ is the solution to (35), and

$$Y^*(s) = \frac{\sum_{k=1}^{P} \{1 - R_k^*[\lambda - \lambda B^*(s)]\}}{R\lambda[1 - B^*(s)]} \times C^*\left[\sum_{j=1}^{k} \lambda_j[1 - B_j^*(\lambda - \lambda B^*(s))] + \sum_{j=k+1}^{P} \lambda_j[1 - B_j^*(s)]\right] \prod_{j=1}^{k-1} R_j^*[\lambda - \lambda B^*(s)]$$

for the globally gated service model, where $C^*(s)$ is the solution to (41). □

## 5. CONCLUDING REMARKS

The workload decomposition in (2) for single-server multiclass systems (Boxma and Groenendjik [2] and Boxma [1]) was initially derived in an attempt to interpret, unify, and generalize the pseudoconservation laws of Ferguson and Aminetzah [4] and of Watson [12]. Calculating the mean workload from (2) indeed easily leads to those conservation laws, as special cases of a much more general pseudoconservation law. However, (2) has until now not been exploited to obtain insight into the workload *distribution* of single-server multiclass systems with vacations (switchover times).

In the present paper, we have obtained the LST of the DF for the workload in the $M/G/1$ system with $P$ classes and server vacations for the cases of (a) setup times, and (b) cyclic service of the classes with exhaustive, gated, or globally gated service, respectively. For case (a), an explicit expression for the workload LST has been derived; for case (b), the workload LST has been expressed in terms of a certain function for which the functional equation is known [the functional equations (30), (35), and (41), respectively]. Workload moments can then be obtained. Perhaps more importantly, our results may be used to analyze the tail behavior of the workload distribution by studying the poles of the workload LST and identifying the pole with the largest real part. Such an analysis could be useful, as often information about mean values of waiting times or workload is not sufficient for judging the performance of a system.

## ACKNOWLEDGMENT

We would like to thank Professor J. W. Cohen, Dr. W. P. Groenendijk, and Mr. J. A. Weststrate for their comments on an early draft of the paper.

## APPENDIX

### Derivation of (12) from (34)

Ferguson and Aminetzah [4] show that

$$E[\omega_k] = r_{k-1} + \frac{\rho_k R}{1-\rho}. \tag{A1}$$

Thus, from (34) we have

$$\begin{aligned} E[Y]_{\text{exhaustive}} &= \sum_{k=1}^{P} \frac{r_k}{R}\left[\sum_{i=2}^{P} E[\omega_{k-P+i}] \sum_{j=k-P+1}^{k-P+i-1} \rho_j + \rho\frac{r_k^{(2)}}{2r_k}\right] \\ &= \frac{1}{R}\sum_{k=1}^{P} r_k \sum_{i=2}^{P} r_{k-P+i-1} \sum_{j=k-P+1}^{k-P+i-1} \rho_j \\ &\quad + \frac{1}{1-\rho}\sum_{k=1}^{P} r_k \sum_{i=2}^{P} \rho_{k-P+i} \sum_{j=k-P+1}^{k-P+i-1} \rho_j + \frac{\rho}{2R}\sum_{k=1}^{P} r_k^{(2)}. \end{aligned} \tag{A2}$$

However, after some manipulation, we get

$$\sum_{k=1}^{P} r_k \sum_{i=2}^{P} r_{k-P+i-1} \sum_{j=k-P+1}^{k-P+i-1} \rho_j = \frac{\rho}{2}\left(R^2 - \sum_{p=1}^{P} r_p^2\right) \tag{A3}$$

and

$$\sum_{k=1}^{P} r_k \sum_{i=2}^{P} \rho_{k-P+i} \sum_{j=k-P+1}^{k-P+i-1} \rho_j = \frac{R}{2}\left(\rho^2 - \sum_{p=1}^{P} \rho_p^2\right) \tag{A4}$$

Substituting (A3) and (A4) into (A2), we get (12).

## REFERENCES

[1] Boxma, O.J., "Workloads and Waiting Times in Single-Server Systems with Multiple Customer Classes," *Queueing Systems*, **5**, 185–214 (1989).

[2] Boxma, O.J., and Groenendijk, W.P., "Pseudo-Conservation Laws in Cyclic-Service Systems," *Journal of Applied Probability*, **24**, 949–964 (1987).

[3] Boxma, O.J., Levy, H., and Yechiali, U., "Cyclic Reservation Schemes for Efficient Operation of Multiple-Queue Single-Server Systems," *Annals of Operations Research* (to be published).

[4] Ferguson, M.J., and Aminetzah, Y.J., "Exact Results for Nonsymmetric Token Ring Systems," *IEEE Transactions on Communications*, **COM-33**, 223–231 (1985).

[5] Fuhrmann, S.W., and Cooper, R.B., "Stochastic Decompositions in the *M/G/*1 Queue with Generalized Vacations," *Operations Research*, **33**, 1117–1129 (1985).

[6] Fuhrmann, S.W., and Moon, A., "Queues Served in Cyclic Order with an Arbitrary Start-Up Distribution," *Naval Research Logistics*, **37**, 123–133 (1990).

[7] Gelenbe, E., and Mitrani, I., *Analysis and Synthesis of Computer Systems*, Academic Press, London, 1980.

[8] Kleinrock, L., *Communication Nets: Stochastic Message Flow and Delay*, McGraw-Hill, New York, 1964, reprinted by Dover Publications, Inc., New York, 1972.

[9] Kleinrock, L., "A Conservation Law for a Wide Class of Queueing Disciplines," *Naval Research Logistics Quarterly*, **12**, 181–192 (1965).

[10] Takagi, H., "Queueing Analysis of Polling Models: An Update," in *Stochastic Analysis of Computer and Communication Systems*, H. Takagi (Ed.), Elsevier, Amsterdam, 1990, pp. 267–318.

[11] Takagi, H., "Priority Queues with Setup Times," *Operations Research*, **38**, 667–677 (1990).

[12] Watson, K.S., "Performance Evaluation of Cyclic Service Strategies—A Survey," in *Performance '84*, E. Gelenbe (Ed.), Elsevier, Amsterdam, 1985, pp. 521–533.

Manuscript received January 7, 1991
Accepted June 6, 1991