MULTICLASS QUEUEING SYSTEMS IN HEAVY TRAFFIC: AN ASYMPTOTIC APPROACH BASED ON DISTRIBUTIONAL AND CONSERVATION LAWS

DIMITRIS BERTSIMAS and GEORGIA MOURTZINOU

Massachusetts Institute of Technology

(Received October 1993; revisions received January 1995, June 1995; accepted July 1995)

We propose a new approach to analyze multiclass queueing systems in heavy traffic based on what we consider as fundamental laws in queueing systems, namely distributional and conservation laws. Methodologically, we extend the distributional laws from single class queueing systems to multiple classes and combine them with conservation laws to find the heavy traffic behavior of the following systems: (a) $\Sigma GI/G/1$ queue under FIFO, (b) $\Sigma GI/G/1$ queue with priorities, (c) Polling systems with general arrival distributions. Compared with traditional heavy traffic analysis via Brownian processes, our approach gives new insight to the asymptotics used, solves systems that traditional heavy traffic theory has not fully addressed, and, more importantly, leads to closed form answers, which compared to simulation are very accurate even for moderate traffic.

The goal of the paper is to present a new approach for heavy traffic analysis of multiclass queueing systems. Starting with a new extension of distributional laws to multiple classes and combining them with conservation laws, we find the heavy traffic behavior of the following systems:

1. $\Sigma GI/G/1$ queue under the First-In-First-Out (FIFO) discipline, in which there are N general renewal processes in a single server queueing system that has a general service time distribution and uses the FIFO discipline. In this system we derive the joint distributions of the number of customers in the system and the waiting time distributions of the various classes.

2. $\Sigma GI/G/1$ queue in which the various classes have preemptive (or nonpreemptive) priorities. In this system we use conservation and distributional laws to find the expected number in the system from each class.

3. $\Sigma GI/G/1$ queue with changeover times and cyclic service, in which the server serves the various classes in a cyclic order, spending time d_{ij} when he moves from class i to class j (polling systems). In this system we derive the expected number in the system from each class.

For all the above systems our results lead to closed form expressions, which even in moderate traffic are very close to those obtained via simulation. We would also like to stress that our results are not identical with traditional heavy traffic results. In contrast with these results, our expressions yield the same numerical answers only for traffic intensities extremely close to one. For finite traffic intensities the two methods differ, with ours being closer to the exact answer in numerical experiments.

More importantly, we feel that our analysis illustrates the following general approach in the analysis of queueing systems: Start the analysis by defining the random variables of interest. Derive the laws that relate these random variables from general laws of queueing theory. In this way we have a complete description of the system, in the sense that we have a sufficient number of equations and unknowns. The only difficulty is that the complexity of the equations prevents us from solving them exactly. In heavy traffic, however, we can use asymptotic expansions to find asymptotically exact closed form expressions. Our approach has parallels in the physics tradition, in which there are fundamental laws that fully describe a physical system, and lead, using mathematical tools, to a complete solution to the quantities of interest.

We feel that the proposed approach gives a clear perspective of the physics of the system, since it starts with a complete description of the system for every traffic. Heavy traffic, then, is nothing more than solving the equations that describe the system asymptotically.

RELATED WORK

Multiclass queueing systems are used to model complex production and service systems with multiple types of customers which may differ in their arrival processes, service requirements, and cost or profit functions. As there are several important applications of the systems we consider in telecommunication, computer, transportation, and jobshop manufacturing systems, there is a huge literature in analyzing their performance.

Related to System 1 ($\Sigma GI/G/1$ under FIFO) Iglehart and Whitt (1970) prove heavy traffic limit theorems. In addition, Fendick, Saksena and Whitt (1989) prove heavy traffic limits for a more general $\Sigma GI/G/1$ under FIFO, where batch arrivals and dependencies are allowed. Our results can be seen as an alternative derivation of the heavy traffic behavior of the system, which leads to closed form expressions that are not identical with those obtained

Subject classification: Queues/limit theorems: multiclass queueing systems in heavy traffic. Queues/cyclic, priority. multiclass priority and polling systems.
Area of review. STOCHASTIC PROCESSES AND THEIR APPLICATIONS.

0030-364X/97/4503-0470 $05 00

in Iglehart and Whitt (1970), but compared with simulation results are very accurate.

Related to System 2 ($\Sigma GI/G/1$ with priorities) Whitt (1971) and Reiman and Simon (1990) prove heavy traffic limits. On the other hand, Gelenbe and Mitrani (1980), Federgruen and Groenevelt (1988a, 1988b) and Shantikumar and Yao (1992) derive conservation laws for expected performance measures. While conservation laws lead to explicit expressions for the performance of systems under priority policies for systems with Poisson arrivals, the performance for systems with general arrivals is not known. We find that the distributional laws lead to explicit expressions for the conservation laws in heavy traffic for systems with general arrivals and thus enable us to analyze the performance of priority policies.

System 3 (polling systems) has been extensively studied for the case of Poisson arrivals (see Takagi 1975 for a survey). Perhaps the most efficient algorithm for the analysis of polling systems with Poisson arrivals is due to Sarkar and Zangwill (1989), in which they analyze the system by solving a linear system of N equations in N unknowns. We generalize their work using distributional laws and derive the heavy traffic behavior of a polling system with general renewal arrivals. Recently, Coffman et al. (1993) proposed an alternative heavy traffic approach, via Brownian processes, for a polling system with two stations.

Regarding the methodological foundation of the paper, namely the distributional laws, Haji and Newell (1971) derive the distributional laws for an overtake free single class system, and for the case of Poisson arrivals Keilson and Servi (1988, 1990) found that the distributional laws have a very convenient form that can lead to complete solutions for some queueing systems.

The approach in the present paper has its origin in the work of Bertsimas and Nakazato (1995) and Bertsimas and Mourtzinou (1996), who give exact expressions for systems involving mixed generalized Erlang arrival distributions and asymptotically exact heavy traffic results for single class systems.

The present paper can be seen as the extension of the distributional laws and their applications to the multiclass case.

The rest of the paper is organized as follows. In Section 1, we develop the multiclass distributional law. In Sections 2, 3, and 4 we derive the heavy traffic behavior of the $\Sigma GI/G/1$ under FIFO, $\Sigma GI/G/1$ with priorities and polling systems respectively as applications of the distributional and conservation laws. Finally, in Section 5 we report numerical results, comparing our results with the traditional heavy traffic approach and simulation.

1. THE MULTICLASS DISTRIBUTIONAL LAW

In this section we first review the single class distributional law for systems with arbitrary renewal arrival processes, and then present a generalization of the distributional law in the multiclass case.

1.1. A Review of the Single Class Distributional Law

Consider a general queueing system, with a *single* stationary renewal arrival process of rate λ. As it will become apparent later, "the system" may correspond to either a single queue, or to a queue and a service facility. We assume that the system satisfies the following conditions:

Assumptions A.

A.1. *All arriving customers enter the system one at a time, remain in the system until served (there is no blocking, balking or reneging) and leave also one at a time.*

A.2. *The customers leave the system in the order of arrival (FIFO).*

A.3. *New arriving customers do not affect the time in the system for previous customers.*

Let $N_a(t)$ be the number of customers up to time t for the ordinary renewal process (where the time of the first interarrival time has the same distribution as the interarrival time). Let $N_a^*(t)$ be the number of customers up to time t for the equilibrium process (where the time of the first interarrival time is distributed as the forward recurrence time of the arrival process).

Then, given that they exist in steady-state, let D be the stationary time a customer spends in the system, and let C be the stationary number of the customers in the system, for a system that satisfies Assumptions A. Let also C^-, C^+ be the number in the system just before an arrival or just after a departure, respectively. We denote by $F_D(t) = P\{D \leq t\}$ the distribution function of D and by $G_C(z) = E[z^C]$ the generating function of C.

The single class distributional law can be stated as follows:

Theorem 1. (*Haji and Newell* 1971, *Bertsimas and Nakazato* 1995.) *For a single class system that satisfies Assumptions A.1–A.3, the stationary number of customers, C, and the stationary system time, D, are related in distribution by:*

$$C \overset{d}{=} N_a^*(D), \quad \textit{equivalently}$$

$$G_C(z) = \int_0^\infty K(z, t)\, dF_D(t), \tag{1}$$

where $K(z, t) \triangleq E[zN_a^*(t)] = \sum_{n=0}^\infty z^n P\{N_a^*(t) = n\}$.

Similar relations hold for the number of customers in the system just before an arrival or just after a departure. Namely,

$$C^- \overset{d}{=} C^+ \overset{d}{=} N_a(D) \quad \text{equivalently}$$

$$G_{C^-}(z) = G_{C^+}(z) = \int_0^\infty K_o(z, t)\, dF_D(t), \tag{2}$$

where $K_o(z, t) \triangleq E[z^{N_a(t)}] = \sum_{n=0}^\infty z^n P\{N_a(t) = n\}$.

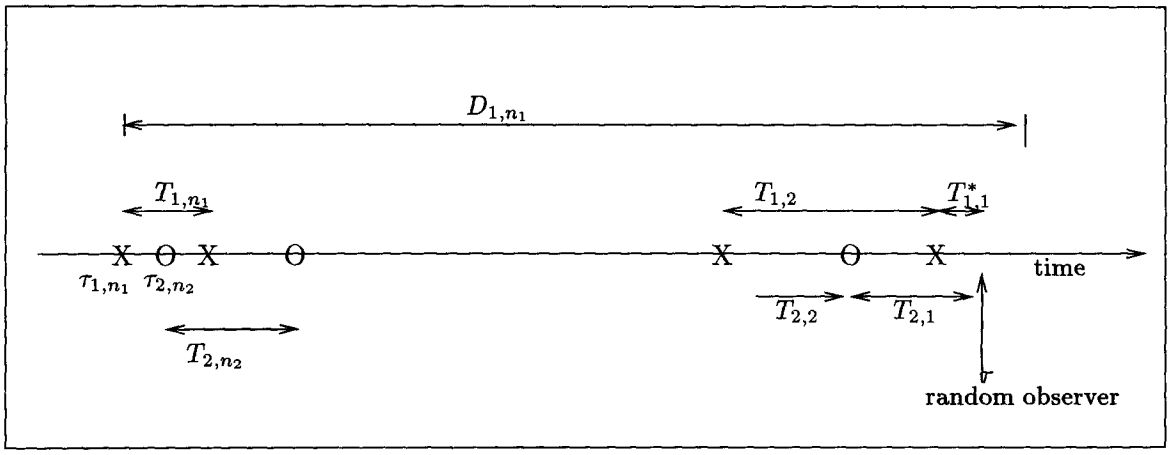

Figure 1. A possible observation scenario in the case of two customer classes.

1.2. The Multiclass Distributional Law

We now consider a general queueing system, with N stationary renewal arrival streams of rate λ_i. We allow different customers classes to have different service requirements and we assume that the system satisfies Assumptions A.1–A.3 and the following Assumption A.4.

Assumption A.4. *Arrival streams from different classes are mutually independent.*

Let $N_{a_i}(t)$, $N^*_{a_i}(t)$ be the number of customers up to time t for the ordinary and equilibrium renewal process of the ith class, respectively. Given that they exist in steady-state, let D_i be the stationary time spent in the system for class i customers and let C_i be the stationary number of class i customers in the system. Finally let $C \triangleq \sum_{i=1}^N C_i$, $F_{D_i}(t) \triangleq P\{D_i \leq t\}$ and $G_{C_1,\ldots,C_N}(z_1, \ldots, z_N) \triangleq E[z_1^{C_1} \ldots z_N^{C_N}]$.

The multiclass distributional law can be stated as follows:

Theorem 2. *For a multiclass queueing system that satisfies Assumptions A.1–A.4,*

$$G_{C_1,\ \ldots,C_N}(z_1, \ldots, z_N)$$

$$= 1 + \sum_{i=1}^{N} \int_0^{\infty} \int_0^{t} \prod_{\substack{j=1 \\ j \neq i}}^{N} K_j(z_j, x) dK_i(z_i, x)\, dF_{D_i}(t), \qquad (3)$$

with $K_i(z_i, t) \triangleq E[z_i^{N^*a_i(t)}] = \sum_{n=0}^{\infty} z_i^n P\{N^*_{a_i}(t) = n\}$.

Proof. Let τ be the time that an observer starts observing the system. Let τ_{i,n_i} be the arrival time of the n_ith customer of the ith class and D_{i,n_i} be his system time. Note that within each class, we number customers "looking backward" from the observation epoch, hence, the customer who is numbered 1 is the customer who arrived most recently. Therefore, τ_{i,n_i} and D_{i,n_i} are ordered in the reverse time direction. (See Figure 1).

Let $T^*_{i,1} \triangleq \tau - \tau_{i,1}$ for $i = 1, \ldots, N$, i.e., $T^*_{i,1}$ is distributed as the backward recurrence time (age) of the ith arrival process, and $T_{i,n_i} \triangleq \tau_{i,n_i-1} - \tau_{i,n_i}$, $n_i \geq 2$, i.e., T_{i,n_i} is the interarrival time of the ith arrival process. The key observation of the proof is that, because of Assumption A.2, for an observer to see, at the random observation epoch τ, at least n_i customers of the ith class in the system, where $n_i \geq 1$, the n_ith customer of the ith class, $i = 1, \ldots, N$, should still be in the system at time τ. Due to Assumptions A.1 and A.2 the event that the n_ith customer is in the system at time τ is equivalent to the event $\{D_{i,n_i} > \tau - \tau_{i,n_i}\}$. Thus, we obtain for $n_i \geq 1$, $i = 1, \ldots, N$ that

$$C_1 \geq n_1, \ldots, C_N \geq n_N \quad \text{if and only if}$$
$$D_{1,n_1} > \tau - \tau_{1,n_1}, \ldots, D_{N,n_N} > \tau - \tau_{N,n_N}. \qquad (4)$$

Therefore,

$$P\{C_1 \geq n_1, \ldots, C_N \geq n_N\}$$
$$= P\{D_{1,n_1} > \tau - \tau_{1,n_1}, \ldots, D_{N,n_N} > \tau - \tau_{N,n_N}\}.$$

We, then, condition on the type of the customer that arrived first to the system, i.e., the less recent customer and obtain:

$$P\{C_1 \geq n_1, \ldots, C_N \geq n_N\}$$
$$= \sum_{i=1}^{N} P\{\tau - \tau_{i,n_i} = \max_j(\tau - \tau_{j,n_j}),$$
$$D_{j,n_j} > \tau - \tau_{j,n_j} \quad \forall j = 1, \ldots, N\}$$
$$= \sum_{i=1}^{N} P\{\tau - \tau_{i,n_i} = \max_j(\tau - \tau_{j,n_j}),$$
$$D_{i,n_i} > \tau - \tau_{i,n_i}\},$$
$$\cdot P\{D_{j,n_j} > \tau - \tau_{j,n_j} \quad \forall j \neq i | \tau - \tau_{i,n_i}$$
$$= \max_j(\tau - \tau_{j,n_j}), D_{i,n_i} > \tau - \tau_{i,n_i}\}.$$

Since the discipline is FIFO (Assumption A.2),

$$P\{D_{j,n_j} > \tau - \tau_{j,n_j} \quad \forall j \neq i | \tau - \tau_{i,n_i}$$
$$= \max_j(\tau - \tau_{j,n_j}), D_{i,n_i} > \tau - \tau_{i,n_i}\} = 1,$$

meaning that given the event that the customer who arrived the first to the system, among the set of customers

$\{n_j, j = 1, \ldots, N\}$, is still in the system at time τ, all customers $\{n_j, j = 1, \ldots, N\}$ are in the system at time τ. Therefore,

$$P\{C_1 \geq n_1, \ldots, C_N \geq n_N\}$$
$$= \sum_{i=1}^{N} P\{\tau - \tau_{i,n_i} = \max_j(\tau - \tau_{j,n_j}) \text{ and } D_{i,n_i} > \tau - \tau_{i,n_i}\}.$$

Since the system is in steady-state, S_{i,n_i} is distributed as the steady-state system time S_i. Moreover, because of Assumption A.3, D_{i,n_i} and $\tau - \tau_{i,n_i}$ are independent. We further condition on D_i and obtain

$$P\{C_1 \geq n_1, \ldots, C_N \geq n_N\}$$
$$= \sum_{i=1}^{N} \int_0^\infty P\left\{\bigcap_{j\neq i}(\tau - \tau_{i,n_i} \geq \tau - \tau_{j,n_j}), \tau - \tau_{i,n_i} < t\right\} dF_{D_i}(t).$$

Conditioning next on $\tau - \tau_{i,n_i}$, introducing the notation

$$A_{i,n_i}(x) \triangleq P\{\tau - \tau_{i,n_i} \leq x\} = P\left\{T^*_{i,1} + \sum_{k=2}^{n_i} T_{i,k} \leq x\right\},$$

and using the independence of $\tau - \tau_{j,n_j}$ for all j (Assumption A.4) we obtain for $n_i \geq 1$, $i = 1, \ldots, N$

$$P\{C_1 \geq n_1, \ldots, C_N \geq n_N\}$$
$$= \sum_{i=1}^{N} \int_0^\infty \int_0^t \prod_{j\neq i}^{N} P\{\tau - \tau_{j,n_j} \leq x\}\, dA_{i,n_i}(x)\, dF_{D_i}(t)$$
$$= \sum_{i=1}^{N} \int_0^\infty \int_0^t \prod_{j\neq i}^{N} A_{j,n_j}(x)\, dA_{i,n_i}(x)\, dF_{D_i}(t). \quad (5)$$

We next consider the general case where the random observer, upon his arrival, does not see *any* customers from classes $k \in \mathcal{A} \subset \{1, \ldots, N\}$ in the system, and sees $n_i \geq 1$ customers from class $i \notin \mathcal{A}$. Similarly with relation (4), we obtain

$$\bigcap_{i\notin\mathcal{A}}(C_i \geq n_i) \quad \text{if and only if} \quad \bigcap_{i\notin\mathcal{A}}(D_{i,n_i} > \tau - \tau_{i,n_i}).$$

Thus, following the derivation of (5), we obtain, for $n_i \geq 1$, $i \notin \mathcal{A}$:

$$P\left\{\bigcap_{i\notin\mathcal{A}}(C_i \geq n_i)\right\}$$
$$= \sum_{i\notin\mathcal{A}} \int_0^\infty \int_0^t \prod_{\substack{j\notin\mathcal{A}\\ j\neq i}} A_{j,n_j}(x)\, dA_{i,n_i}(x)\, dF_{D_i}(t), \quad (6)$$

We next calculate $P\{C_1 = n_1, \ldots, C_N = n_N\}$ iteratively, based on (5) and (6) and using the fact that for $n_i \geq 0$,

$$P\{C_1 = n_1, \ldots, C_i = n_i, C_{i+1} \geq n_{i+1}, \ldots, C_N \geq n_N\}$$
$$= P\left\{\bigcap_{k\leq i-1}(C_k = n_k), \bigcap_{j\geq i}(C_j \geq n_j)\right\}$$
$$- P\left\{\bigcap_{k\leq i-1}(C_k = n_k), C_i \geq n_i + 1, \bigcap_{j\geq i+1}(C_j \geq n_j)\right\}.$$

Finally, we compute generating functions and, after some algebra, we find that:

$$G_{C_1,\ldots,C_N}(z_1, \ldots, z_N)$$
$$= 1 + \sum_{i=1}^{N} \int_0^\infty \int_0^t \prod_{\substack{j=1\\ j\neq i}}^{N} K_j(z_j, x)\, dK_i(z_i, x)\, dF_{D_i}(t),$$

with

$$K_i(z, t) \triangleq E[z^{N^*_{a_i}(t)}] = P\{T^*_{i,1} \geq t\} + \sum_{n=1}^{\infty} z^n \left\{P\left\{T^*_{i,1} + \sum_{j=2}^{n} T_{i,j} < t\right\} - P\left\{T^*_{i,1} + \sum_{j=2}^{n+1} T_{i,j} < t\right\}\right\}. \quad \square$$

Remarks.

1. Note that for the case of a single class (3) reduces to (1).
2. The generating function of the total number of customers, C, in the system can be found if we set $z_i = z$ in (3):

$$G_C(z) = 1 + \sum_{i=1}^{N} \int_0^\infty \int_0^t \prod_{\substack{j=1\\ j\neq i}}^{N} K_j(z, x)\, dK_i(z, x)\, dF_{D_i}(t). \quad (7)$$

We define as **overtake free multiclass queueing systems** those systems that satisfy Assumptions A.1–A.4 and therefore, satisfy multiclass distributional laws. These include

(a) $\Sigma GI/G/1$ queueing system under FIFO (where we can define "the system" to be either just the queue or the queue together with the server),

(b) $\Sigma GI/D/s$ queueing system under FIFO (where we can define "the system" to be either just the queue or the queue together with the s servers),

(c) $\Sigma GI/G/s$ queueing system under FIFO (where we must define the "the system" to be only the queue, since if the "system" is the queue together with the s servers overtaking can take place and therefore Assumption A.2 is violated),

(d) multiclass queueing systems with vacations (see Bertsimas and Mourtzinou 1996, and Keilson and Servi 1990) (where, once again, we can define we can define "the system" to be either just the queue or the queue together with the server).

1.3. Asymptotic Forms of the Kernels $K_o(z, t)$ and $K(z, t)$

The main contribution of our analysis so far is that we established a set of relationships between the distributions of the number of customers in the system and the system time for a class of systems that satisfy Assumptions A.1–A.4. These distributional laws relationships are expressed as integral relationships between the generating function

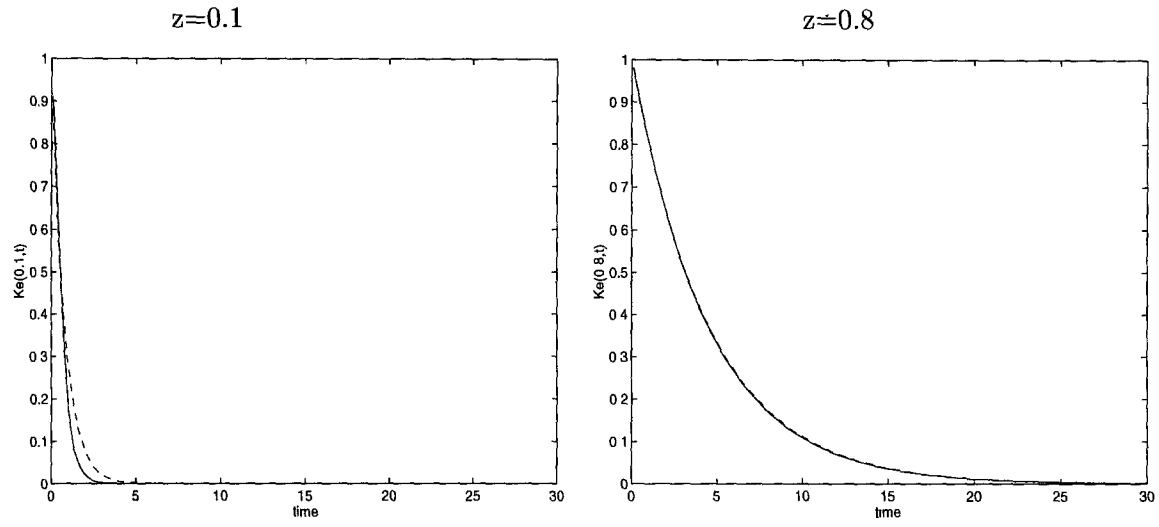

Figure 2. The function $K_e(z, t)$ for Erlang 16 arrivals.

of the number of customers in the system and the distribution of the system time. For example, for the single class system we have that:

$$G_C(z) = \int_0^\infty K(z, t)\, dF_D(t),$$

and for multiclass systems we also have

$$G_{C_1, \ldots, C_N}(z_1, \ldots, z_N) = 1 + \sum_{i=1}^{N} \int_0^\infty \int_0^t \prod_{\substack{j=1 \\ j \neq i}}^{N} K_j(z_j, x)\, dK_i(z_i, x)\, dF_{D_i}(t),$$

where the kernels $K(z, t)$ and $K_i(z, t)$ were defined in Theorem 1 and Theorem 2, respectively.

Since we are going to use these kernels extensively, we next compute them asymptotically as $t \to \infty$ and $z \to 1$ for general renewal processes. We use the notation that $h(x) \sim r(x)$ as $x \to a$ means that $\lim_{x \to a} h(x)/r(x) = 1$ and following the asymptotic approach introduced in Smith (1954) (see also Cox 1962, Ch. 4–6) we obtain (see Mourtzinou 1995):

Theorem 3. *For a renewal process with rate λ and square coefficient of variation c_a^2, asymptotically, as $t \to \infty$ and $z \to 1$:*

$$K(z, t) \sim e^{-tf(z)} \quad \text{and} \quad K_o(z, t) \sim \frac{f(z)}{\lambda(1 - z)} e^{-tf(z)},$$

where $f(z) \triangleq \lambda(1 - z) - \frac{1}{2}\lambda(1 - z)^2(c_a^2 - 1)$.

It is important to notice that the above relationships are exact for Poisson processes under any traffic intensity.

In order to check the accuracy of our asymptotic method, we next present some numerical results. In Figures 2 and 3 the solid line corresponds to the exact value

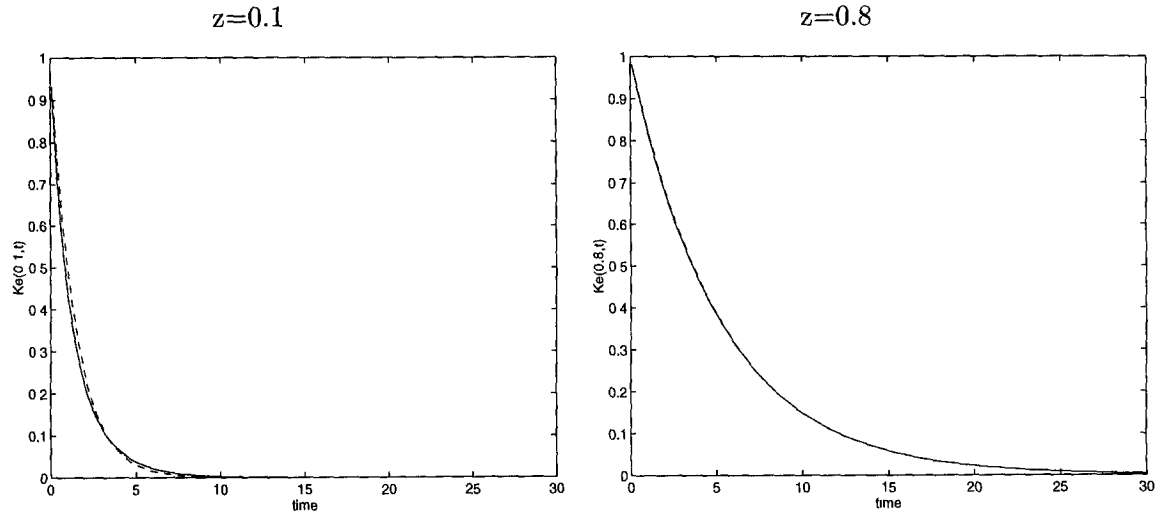

Figure 3. The function $K_e(z, t)$ for Hyperexponential arrivals with $c_a^2 = 1.5$.

of the kernel $K_e(z, t)$, obtained via numerical Laplace inversion, and the dashed line to the asymptotic expansion. To invert the Laplace transform of $K_e(z, t)$ we used the two algorithms in Hosono (1981) and in Abate and Whitt (1995) which we programmed in Matlab and we got exactly the same results.

Notice that our expansion is indeed asymptotically exact as $z \to 1$ and $t \to \infty$. Moreover, in all the cases we consider it is exact for $t > 20$.

We should point out that, according to the line of arguments in Mourtzinou (1995) Proposition 2.2, $-f(z)$ is the root of $1 - z\alpha(s) = 0$ as $z \to 1$, where $\alpha(s)$ is the Laplace transform of the interarrival distribution. In other words,

$$1 - z\alpha(-f(z)) = 0 \quad \text{as} \quad z \to 1. \tag{8}$$

We will use the above equation extensively in the sequel.

2. THE ΣGI/G/1 QUEUEING SYSTEM UNDER FIFO

Consider a $\Sigma GI/GI/1$ queue with N classes of customers. Class i customers arrive at the system according to an ordinary renewal process of rate λ_i, squared coefficient of variation $c_{a_i}^2$ and Laplace transform of the interarrival times $\alpha_i(s)$. Let X_i be the random variable corresponding to the service time of a class i customer. We denote with $E[X_i]$ and $c_{x_i}^2$ the mean and the squared coefficient of variation of X_i.

Let, also, X_i^* be the backward recurrence time (age) of the service time of a class i customer, i.e., if at a random epoch τ a class i customer is in the server, X_i^* corresponds to the amount of service time this customer has received up to time τ. Let $\rho_i \triangleq \lambda_i E[X_i]$ and $\rho \triangleq \sum_{i=1}^N \rho_i$.

Finally, let W_i be the time spent in the queue and S_i be the time spent in the queue and the server for class i customers, in steady-state. Let Q_i be the number of the ith class in the queue and L_i be the number of the ith class in the queue and the server, given that those quantities exist in steady-state. Denote, also by $Q(L)$ the steady-state number of all the customers in the queue (or queue and server).

Our goal for the rest of this section is to evaluate the performance of the multiclass $\Sigma GI/GI/1$ queue under FIFO. Since the system is multiclass we need to calculate the distribution functions of the individual L_i, Q_i, S_i, and W_i, for all $i = 1, 2, \ldots, N$, as well as the joint distribution of all the customers in the system or in the queue. Equivalently in the transform domain we need to calculate $\phi_{S_i}(s)$, $\phi_{W_i}(s)$, $G_{L_i}(z_i)$, $G_{Q_i}(z_i)$ as well as $G_{L_1, L_2, \ldots, L_N}(z_1, z_2, \ldots, z_N)$ and $G_{Q_1, Q_2, \ldots, Q_N}(z_1, z_2, \ldots, z_N)$.

Let us start with the individual quantities first. As the service policy is FIFO the multiclass as well as the individual distributional laws hold if we consider the "system" to be either the queue or the queue plus the service facility. Therefore we have that for all $i = 1, 2, \ldots, N$.

$$G_{L_i}(z_i) = \int_0^\infty K_i(z_i, t)\, dF_{S_i}(t), \quad \text{and}$$

$$G_{Q_i}(z_i) = \int_0^\infty K_i(z_i, t)\, dF_{W_i}(t), \tag{9}$$

with $K_i(z_i, t) \triangleq E[z^{N_{a_i}(t)}] = \sum_{n=0}^\infty z_i^n P\{N_{a_i}^*(t) = n\}$. Furthermore, we have that

$$S_i = W_i + X_i, \quad \text{for all} \quad i = 1, 2, \ldots, N. \tag{10}$$

Moreover, in order to complete the required number of equations to be able to form an adequate system (notice that we have $4N$ unknowns and only $3N$ equations so far) we prove the following theorem.

Theorem 4. *In a $\Sigma GI/GI/1$ queue under FIFO*

$$G_{L_i}(z_i) = (1 - z_i)\left[(1 - \rho) + \sum_{\substack{j=1 \\ j \neq i}}^N \rho_j \int_0^\infty K_i(z_i, t)\, dF_{W_j + X_i^*}(t)\right] + z_i G_{Q_i}(z_i), \tag{11}$$

where $K_i(z_i, t) \triangleq E[z^{N_{a_i}^*(t)}] = \sum_{n=0}^\infty z_i^n P\{N_{a_i}^*(t) = n\}$.

Proof. Denote by B_i the event that at the arrival epoch of a random observer the server is busy by a class i customer. By applying Little's law to the server we obtain: $P\{B_i\} = \rho_i$.

Conditioning on the state of the server at a random epoch, we have that:

$$G_{Q_1, \ldots, Q_N}(z_1, \ldots, z_N) = (1 - \rho) + \sum_{i=1}^N \rho_i E[z_1^{Q_1} \ldots z_N^{Q_N} | B_i], \tag{12}$$

$$G_{L_1, \ldots, L_N}(z_1, \ldots, z_N) = (1 - \rho) + \sum_{i=1}^N z_i \rho_i E[z_1^{Q_1} \ldots z_N^{Q_N} | B_i]. \tag{13}$$

Moreover, because of FIFO, if at a random observation time τ the server is busy servicing a class i customer (we call this customer the *tagged customer*) and there are n_j class j customer waiting in queue, those customers must have arrived after the arrival of the tagged customer (τ_1) and before τ. In other words, they must have arrived during the interval $W_i + X_i^*$ where W_i is the stationary waiting time and X_i^* is the age of the service time for the tagged customer.

Notice, however, that we start counting customers upon the arrival of the tagged customer, that is upon a *renewal epoch* of the ith process that constitutes a *random incidence* for the other arrival processes, due to Assumption A.4 (see Figure 4).

Consequently, we must have n_i renewals of the ith arrival process in $\tau - \tau_1$, where the time of the first renewal has the same distribution as the interarrival time and n_j renewals of jth arrival process ($j \neq i$) in the same interval, where the time of the first renewal has the same distribution as the forward recurrence interarrival time of the jth process.

Furthermore, due to FIFO and to the independence of the arrival processes, W_i, X_i^* and the arrival processes are independent, and therefore:

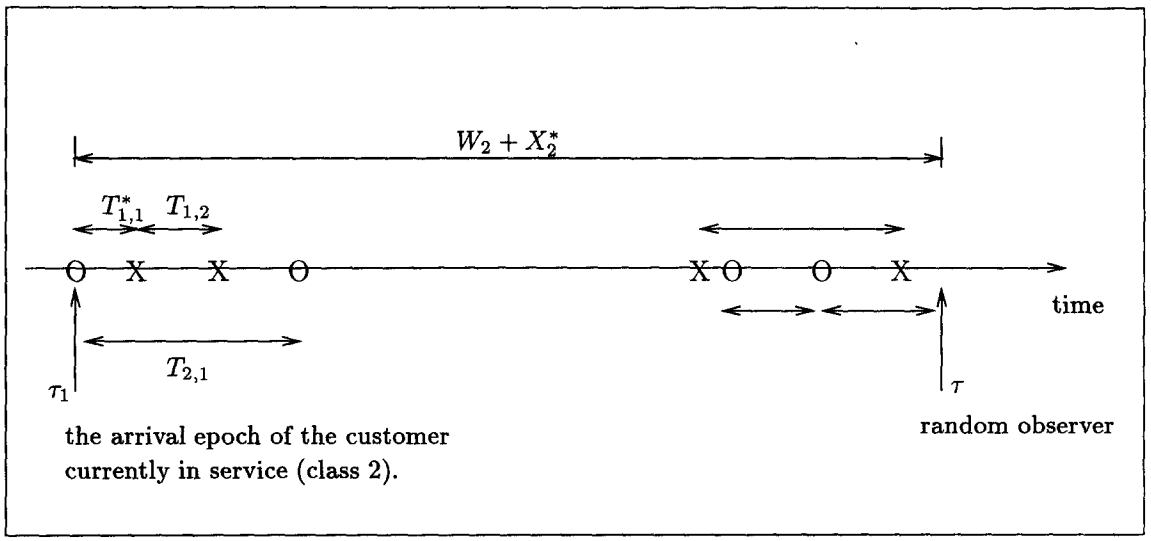

Figure 4. A possible observation scenario.

$$P\{Q_1 = n_1, \ldots, Q_N = n_N | B_i\}$$
$$= P\{N^*_{a_1}(W_i + X_i^*) = n_1, \ldots, N_{a_i}(W_i + X_i^*)$$
$$= n_i, \ldots, N^*_{a_N}(W_i + X_i^*) = n_N\}. \quad (14)$$

By taking z-transforms we have:

$$E[z_1^{Q_1} \ldots z_N^{Q_N} | B_i]$$
$$= \int_0^\infty K_{o,i}(z_i, t) \prod_{\substack{j=1 \\ j \neq i}}^{N} K_j(z_j, t)\, dF_{W_i + X_i^*}(t), \quad (15)$$

where for $i = 1, \ldots, N$, we define $K_i(z_i, t) \triangleq E[z^{N^*_{a_i}(t)}]$ and $K_{o,i}(z_i, t) \triangleq E[z^{N_{a_i}(t)}]$.

Substituting (15) into (12) and (13), we obtain

$$G_{Q_1, \ldots, Q_N}(z_1, \ldots, z_N)$$
$$= (1 - \rho) + \sum_{i=1}^{N} \rho_i \int_0^\infty K_{o,i}(z_i, t)$$
$$\cdot \prod_{\substack{j=1 \\ j \neq i}}^{N} K_j(z_j, t)\, dF_{W_i + X_i^*}(t), \quad (16)$$

$$G_{L_1, \ldots, L_N}(z_1, \ldots, z_N)$$
$$= (1 - \rho) + \sum_{i=1}^{N} z_i \rho_i \int_0^\infty K_{o,i}(z_i, t)$$
$$\cdot \prod_{\substack{j=1 \\ j \neq i}}^{N} K_j(z_j, t)\, dF_{W_i + X_i^*}(t). \quad (17)$$

As special case of the above relations we obtain for $i = 1, \ldots, N$ that

$$G_{Q_i}(z) = (1 - \rho) + \rho_i \int_0^\infty K_{o,i}(z, t)\, dF_{W_i + X_i^*}(t)$$
$$+ \sum_{\substack{j=1 \\ j \neq i}}^{N} \rho_j \int_0^\infty K_i(z, t)\, dF_{W_j + X_j^*}(t), \quad (18)$$

$$G_{L_i}(z) = (1 - \rho) + z\rho_i \int_0^\infty K_{o,i}(z, t)\, dF_{W_i + X_i^*}(t)$$
$$+ \sum_{\substack{j=1 \\ j \neq i}}^{N} \rho_j \int_0^\infty K_i(z, t)\, dF_{W_j + X_j^*}(t). \quad (19)$$

Combining (18) and (19) we complete the proof. □

Let us note that in the special case of a single class $GI/GI/1$ queue (16) and (17) have been proved in Lemoine (1974). Moreover, (11) together with (9) and (10) form an $4N \times 4N$ system of equations that completely characterizes the individual distributions of L_i, Q_i, S_i, and W_i in the multiclass $\Sigma GI/GI/1$ system under FIFO. In particular, the distributions of W_i's can be obtained as follows.

Theorem 5. *For a $\Sigma GI/GI/1$ queueing system under FIFO the distribution of W_i for all $i = 1, \ldots, N$ satisfy the following $N \times N$ system of integral equations*:

$$\int_0^\infty K_i(z_i, t)[dF_{W_i + X_i}(t) - z dF_{W_i}(t)]$$
$$= (1 - z_i)\left[(1 - \rho) + \sum_{\substack{j=1 \\ j \neq i}}^{N} \rho_j \cdot \int_0^\infty K_i(z_i, t)\, dF_{W_j + X_j^*}(t)\right], \quad (20)$$

where $K_i(z_i, t) \triangleq E[z_i^{N^*_{a_i}(t)}] = \sum_{n=0}^{\infty} z_i^n P\{N^*_{a_i}(t) = n\}$.

This is a system of integral equations that can not be solved analytically and therefore, motivated the following asymptotic approach.

2.1. Heavy Traffic Behavior of the ΣGI/G/1 Under FIFO

Our goal in this section is to examine the behavior of the $\Sigma GI/G/1$ under FIFO as $Q_i, W_i \to \infty$. For the rest of this paper we only consider systems in which either the interarrival or the service times are *nonarithmetic*. It is well known that for these systems there is a natural parameter ρ, the traffic intensity, such that as $\rho \to 1$, Q_i and $W_i \to \infty$ for all $i = 1, 2, \ldots, N$. Therefore, whenever we say that a system is operating under heavy traffic conditions, we mean that $\rho \to 1$ and therefore, $Q_i, W_i \to \infty$ for all $i = 1, 2, \ldots, N$.

Under heavy traffic conditions we prove the following theorem:

Theorem 6. *In a $\Sigma GI/GI/1$ system under FIFO, in heavy traffic, the Laplace transforms of the individual waiting times are given by*

$$\phi_{W_i}(s) = \frac{(1-\alpha_i(-s))}{1-\alpha_i(-s)\phi_{X_i}(s)+\rho_i(1-\alpha_i(-s))\phi_{X_i^*}(s)}\frac{(1-\rho)}{1-D(s)}, \tag{21}$$

where $\alpha_i(s)$ is the Laplace transform of class i interarrival times and

$$D(s) \triangleq \sum_{j=1}^{N} \frac{(1-\alpha_j(-s))\rho_j\phi_{X_j^*}(s)}{1-\alpha_j(-s)\phi_{X_j}(s)+\rho_j(1-\alpha_j(-s))\phi_{X_j^*}(s)}.$$

Proof. We start by noticing that under heavy traffic conditions, $Q_i \to \infty$ for all $i = 1, \ldots, N$. From the definition of $G_{Q_i}(z_i) \triangleq E[\exp(Q_i \log z_i)]$, we observe that the behavior of the distribution Q_i under heavy traffic conditions is associated with the behavior of its generating function for $-\log z_i$ near zero (see also Cox 1962, p. 14). Moreover, under heavy traffic conditions, $L_i \to \infty$ for all $i = 1, \ldots, N$. Hence, under heavy traffic conditions we are interesting in the asymptotic forms of (20) as $z_i \to 1$.

On the other hand, under heavy traffic conditions $W_i \to \infty$ for all $i = 1, \ldots, N$ and therefore all integrands in (20) vanish unless $t \to \infty$.

Hence, we take the asymptotic expansion of $K_i(z_i, t)$ around both $z \to 1$ and $t \to \infty$ and obtain, from Theorem 3, that for all $i = 1, 2, \ldots, N$,

$$[\phi_{X_i}(f_i(z_i)) - z_i]\phi_{W_i}(f_i(z_i)) \sim (1-z_i)\left[(1-\rho) + \sum_{\substack{j=1 \\ j\neq i}}^{N} \rho_j \phi_{X_j^*}(f_i(z_i))\phi_{W_j}(f_i(z_i))\right].$$

Multiplying both sides of the above equation with $\alpha_i(-f_i(z_i))$ and using the fact that $1 - z_i\alpha_i(-f_i(z_i)) = 0$ we get

$$[1-\alpha_i(-f_i(z_i))\phi_{X_i}(f_i(z_i))]\phi_{W_i}(f_i(z_i)) \sim (1-\alpha_i(-f_i(z_i))) \cdot \left[(1-\rho) + \sum_{\substack{j=1 \\ j\neq i}}^{N} \rho_j \phi_{X_j^*}(f_i(z_i))\phi_{W_j}(f_i(z_i))\right].$$

Setting for each i: $s \triangleq f_i(z_i)$, we obtain for $i = 1, \ldots, N$:

$$\phi_{W_i}(s)[1-\alpha_i(-s)\phi_{X_i}(s)] \sim (1-\alpha_i(-s))\left[(1-\rho) + \sum_{\substack{j=1 \\ j\neq i}}^{N} \rho_j \phi_{X_j^*}(s)\phi_{W_j}(s)\right].$$

The previous equations form a $N \times N$ linear system which can be solved by adding and subtracting $(1 - \alpha_i(-s))\rho_i\phi_{X_i^*}(s)\phi_{W_i}(s)$. We can then solve for each $\phi_{W_i}(s)$ as a function of $\sum_{j=1}^{N} \rho_j\phi_{X_j^*}(s)\phi_{W_j}(s)$, from which (21) follows. □

Having found the transforms of $\phi_{W_i}(s)$ we can obtain $\phi_{S_i}(s)$, $G_{L_i}(z_i)$ and $G_{Q_i}(z_i)$ for all $i = 1, \ldots, N$, via Equations (9)–(11). We can also obtain the asymptotic (heavy traffic) joint distribution $G_{L_1, \ldots, L_N}(z_1, \ldots, z_N)$ and $G_{Q_1, \ldots, Q_N}(z_1, \ldots, z_N)$ using (3) and also the distribution of the total number of customers in the queue if we set $z_i = z$ for all $i = 1, \ldots, N$ in the formula of the joint transform of $(Q_1, \ldots, Q_N)$.

It is interesting to notice that for Poisson arrival processes $\alpha_i(-s) = \lambda_i/(-s + \lambda_i)$, so we obtain, as it was expected (see Kleinrock 1975), the exact expression:

$$\phi_{W_i}(s) = \frac{1-\rho}{1-\sum_{j=1}^{N}\rho_j\phi_{X_j^*}(s)}, \quad i = 1, \ldots, N.$$

We next find closed form expressions for the expectations of the performance measures, since we will use them in the next section.

Proposition 1. *In a $\Sigma GI/G/1$ queue under FIFO in heavy traffic, for $i = 1, \ldots, N$*

$$E[W_i] \sim \frac{\sum_{j=1}^{N}\lambda_j E[X_j^2] + \rho_j E[X_j](c_{a_j}^2 - 1)}{2(1-\rho)} + \frac{1}{2}E[X_i](c_{a_i}^2 - 1). \tag{22}$$

Proof. From Little's law, $E[Q_i] = \lambda_i E[W_i]$. By differentiating (18) we obtain,

$$E[Q_i] = E[N_{a_i}(W_i + X_i^*)] + \sum_{\substack{j=1 \\ j\neq i}}^{N} E[N_{a_j}^*(W_i + X_i^*)] \sim \lambda_i \sum_{j=1}^{N} \rho_j(E[W_j] + E[X_j^*]) + \frac{1}{2}\rho_i(c_{a_i}^2 - 1).$$

Substituting $E[X_j^*] = E[X_j^2]/2E[X_j]$ and solving the resulting system we obtain (22). □

Table I
Systems Satisfying Strong Conservation Laws in Steady-State

System	Special	Characteristics	Performance Measure	Evaluation of b
$\Sigma M/G/1$	N-classes	nonpreemptive	$\rho_i E[W_i]$	Gelenbe and Mitrani (1980)
$\Sigma GI/M/1$	N-classes	preemptive	$\rho_i E[W_i]$	Theorem 8a
$\Sigma GI/G/1$	N-classes same service	nonpreemptive	$\rho_i E[W_i]$	Theorem 8b
$\Sigma GI/G/1$	2-classes	nonpreemptive	$\rho_i E[W_i]$	Theorem 8c

3. THE $\Sigma GI/G/1$ UNDER PRIORITY DISCIPLINES

So far we have only considered systems under the FIFO service discipline. There are, however, other service disciplines, in particular priority policies, that arise in practical situations and therefore it is interesting to develop a methodology to analyze performance under such policies.

Our goal in this section is to use conservation laws that have been developed in the last decade for multiclass queueing systems, together with the results of the previous section in order to analyze explicitly the performance of priority policies.

Let us first review the conservation laws. Consider a $\Sigma GI/G/1$ system, and denote by $E = \{1, 2, \ldots, N\}$ the set of all classes and by 2^E the set of all subsets of E. Let $\mathcal{U}$ to be the set of all *work conserving and nonanticipative* policies (for formal definitions of these policies see Heyman and Sobel 1982). For any policy $u \in \mathcal{U}$ and any class i, we let x_i^u be the performance measure of class i ($i \in E$) customers under policy u. We restrict our attention to performance measures which are expectations. We then define $x^u := (x_i^u)_{i \in E}$ to be the performance vector under policy u. Finally, for any given permutation π of the N elements of E, we let x_i^π denote the performance measure of class i under an absolute policy rule that assigns priorities to customer types according to the permutation π, i.e., type $\pi(1)$ has the highest priority, ..., type $\pi(N)$ has the lowest priority.

Then, the following is a summary of the relevant conservation laws results presented in Shantikumar and Yao (1992).

Theorem 7. *If a performance vector x satisfies* strong conservation laws, *i.e., if there exists a set function b: $2^E \to R_+$ such that $b(\emptyset) = 0$ satisfying*:

$$\sum_{i \in A} x_i^\pi = b(A) \quad \text{for all } \pi: \{\pi(1), \ldots, \pi(|A|)\} = A$$

and for all $A \subset E$;

and for any policy $u \in \mathcal{U}$,

$$\sum_{i \in A} x_i^u \geq b(A) \quad \text{for all } A \subset E \text{ and } \sum_{i \in E} x_i^u = b(E),$$

then the performance vector of an absolute priority policy π, $\{\pi(1), \ldots, \pi(N)\} = E$, is given by:

$$x_{\pi(1)}^\pi = b(\{\pi(1)\})$$
$$x_{\pi(2)}^\pi = b(\{\pi(1), \pi(2)\}) - b(\{\pi(1)\})$$
$$\vdots$$
$$x_{\pi(N)}^\pi = b(E) - b(\{\pi(1), \ldots, \pi(N-1)\}).$$

The major result about systems that satisfy conservation laws is that if we know the set function $b(\cdot)$ we are able to calculate the performance of priority policies; where $b(A)$ is the minimal performance $\Sigma_{i \in A} x_i^u$ over customer classes in a subset $A \subset E$ achieved by an absolute priority policy giving priority to classes in the set A over all other classes in $E - A$. Unfortunately the set functions $b(\cdot)$ (and therefore the performance of arbitrary policies) are only known for systems with Poisson arrivals (see, e.g., Gelenbe and Mitrani 1980). Our contribution in this section is to calculate the set function $b(\cdot)$ in heavy traffic for a variety of systems $\Sigma GI/G/1$ that satisfy conservation laws. We note that conservation laws hold even for multiserver systems (see Shantikumar and Yao 1992), but we only deal with $\Sigma GI/G/1$ in this paper.

In Table I below we summarize $\Sigma GI/G/1$ systems that satisfy conservation laws. Note that in the last three systems the set function $b(\cdot)$ is not known. We calculate the set function $b(\cdot)$, under heavy traffic conditions, in Theorem 8.

Recall that Q_i denotes the number of class i customers in the queue and W_i denotes the steady state waiting time of class i. Furthermore, we denote by ρ_i and $E[X_i]$ the traffic intensity and the mean service time, respectively, for the class i.

3.1. Evaluation of the Set Function $b(\cdot)$ in Heavy Traffic

In this section we evaluate the set function $b(\cdot)$ for the systems presented in Table I in heavy traffic. The idea of our derivation is that the set function $b(A)$ is insensitive to any change in the control policy as long as we are restricted to work conserving and nonanticipative policies that give priority to the classes in set A over these classes in $E - A$. The *distributional laws* enable us to evaluate the performance measures when the service discipline is FIFO. Therefore, we can assume the FIFO discipline within A and $E - A$ and then use the distributional laws in order to evaluate the set function $b(\cdot)$. In this way we will

be able to find $b(\cdot)$ in closed form in heavy traffic as a function of λ_i, $c_{a_i}^2$, $E[X_i]$, $E[X_i^2]$ and ρ_i for all i.

Theorem 8. *In a $\Sigma GI/G/1$ system with customer classes in $E = \{1, \ldots, N\}$, the value of the set function $b(A)$ is given as follows, for any $A \subset E$ that satisfies the heavy traffic condition (i.e., $\rho_A = \sum_{j \in A} \rho_j \to 1$):*

(a) *When preemption is allowed,*

$$b(A) \sim \frac{\rho_A \sum_{j \in A} \lambda_j E[X_j^2] + \sum_{j \in A} \rho_j E[X_j](c_{a_j}^2 - 1))}{2(1 - \rho_A)}. \tag{23}$$

(b) *If all customers have the same service requirement and preemption is not allowed,*

$$b(A) \sim \frac{\rho_A E[X^2] \sum_{i \in E} \lambda_i + E[X] \sum_{j \in A} \rho_j (c_{a_j}^2 - 1)}{2(1 - \rho_A)}. \tag{24}$$

(c) *If there are two customer classes having different service requirements and preemption is not allowed,*

$$b(A) \sim \frac{\rho_A \sum_{i \in E} \lambda_i E[X_i^2] + \sum_{j \in A} \rho_j E[X_j](c_{a_j}^2 - 1)}{2(1 - \rho_A)}. \tag{25}$$

Proof. Based on the previous discussion we have that for all $A \subset E$:

$$b(A) = \sum_{i \in A} \rho_i E[W_i], \tag{26}$$

where $E[W_i]$ is the mean waiting time of the ith class under a policy that gives priority (preemptive or nonpreemptive depending on the case considered) to the subset A and uses FIFO inside the sets A and $E - A$. We next evaluate $E[W_i]$ under different assumptions.

(a) If preemption is allowed, the customers in the set A are not influenced by customers in $E - A$. Hence, we can evaluate $E[W_i]$ by considering a $\Sigma GI/G/1$ system with classes just from A, where all customers are served under the FIFO discipline.

But in (22) we have evaluated $E[W_i]$ in heavy traffic for such a system. Substituting to (26) and rearranging terms we obtain (23).

(b) If all customers have the same service requirement X and preemption is not allowed, we need to find $E[W_i]$, $i \in A$, when we give non-preemptive priority to customers in A over customers in $E - A$ and within the set A we use FIFO. From Little's law we obtain:

$$E[Q_i] = \lambda_i E[W_i], \quad i \in E. \tag{27}$$

Let B^j the event that a random observer finds the server busy by a class j customer. Clearly, $P\{B^j\} = \rho_j$, $j \in E$. Then, conditioning on the class a random observer finds in service, we obtain

$$E[Q_i] = \sum_{j \in E} \rho_j E[Q_i | B^j], \quad i \in E. \tag{28}$$

First, consider the case where $i \in A$ and $j \in E - A$. Due to the fact that customers in A have priority over those in $E - A$, we know that when the service of the customer from class j was initialized there were no customers from class i present. Therefore, $E[Q_i|B^j]$ is exactly the expected number of class i that arrived *after* the initialization of the current service and *before* the arrival of the random observer. Moreover, the arrival of the random observer constitutes a random incidence for both the arrival process of class i and the current service time. Hence,

$$E[Q_i|B^j] \sim E[N_{a_i}^*(X^*)] = \lambda_i E[X^*], \quad i \in A, j \in E - A, \tag{29}$$

where $E[X^*]$ is the mean backward recurrence time (age) of the service time distribution.

Next, if we consider classes $i, j \in A$ we have from (14) that

$$E[Q_i|B^j] = E[N_{a_i}^*(W_j + X^*)] = \lambda_i (E[W_j] + E[X^*]), \quad i, j \in A, j \neq i, \tag{30}$$

$$E[Q_i|B^i] = E[N_{a_i}(W_i + X^*)] \sim \lambda_i (E[W_i] + E[X^*]) + \tfrac{1}{2}(c_{a_i}^2 - 1). \tag{31}$$

Using Equations (27)–(31) we obtain the following system of equations for $i \in A$:

$$E[W_i] - \sum_{j \in A} \rho_j E[W_j] \sim \rho E[X^*] + \tfrac{1}{2} E[X](c_{a_i}^2 - 1).$$

Solving the above system yields (24).

(c) If there are two customer classes with different requirements, and preemption is not allowed, we follow exactly the proof of case (b) above, but instead of Equations (29), (30), and (31) we use:

$$E[Q_i|B^j] = \lambda_i E[X_j^*], \quad i \in A, j \in E - A.$$

$$E[Q_i|B^j] = \lambda_i (E[W_j] + E[X_j^*]), \quad i, j \in A, j \neq i,$$

$$E[Q_i|B^i] \sim \lambda_i (E[W_i] + E[X_j^*]) + \tfrac{1}{2}(c_{a_i}^2 - 1).$$

Using the above equations we form a $|A| \times |A|$ system, which, once solved, yields (25). □

Remarks.

1. It is important to notice that in part (a) of Theorem 8 we only used the heavy traffic condition to obtain closed form expressions of $E[W_i]$. Alternatively, one can solve, numerically, the integral equations of Theorem 5, calculate the expectations substitute in (22) and obtain the *exact* formula for $b(A)$, under any traffic intensity. This is not true, however, for parts (b) and (c).

2. For the case of Poisson arrivals and under nonpreemption (24) and (25) are exact. Moreover, under preemption, Poisson arrivals, and exponential service times ($\Sigma M/M/1$), (23) is also exact.

3.2. Performance Analysis of Priority Policies

Consider a $\Sigma GI/G/1$ system that satisfies conservation laws under heavy traffic conditions, i.e., the total traffic intensity $\rho \to 1$. Suppose that an absolute priority policy π is used

that gives highest priority to class 1, then to class 2, etc. Then, from Theorem 7, $\rho_1 E[W_1] = b(\{1\})$, $\rho_i E[W_i] = b(\mathcal{N}_i) - b(\mathcal{N}_{i-1})$, where $\mathcal{N}_i = \{1, \ldots, i\}$.

In Theorem 8 we evaluated $b(\mathcal{N}_i)$ in heavy traffic, i.e., as long as $\rho_{\mathcal{N}_i} \to 1$. Hence, for classes i such that $\rho_{\mathcal{N}_{i-1}} \to 1$ using Theorems 7 and 8 we can obtain asymptotically exact formulae for $E[W_i]$. However, even if $\rho \to 1$, $\rho_{\mathcal{N}_k} \not\to 1$ for $k \ll N$, where N is the total number of classes. Hence, our method only provides an approximation for classes k such that $\rho_{\mathcal{N}_{k-1}} \not\to 1$.

For example, consider a $\Sigma GI/G/1$ system with four customer classes, where $\rho_1 = \frac{1}{3} - \frac{4}{n}$, $\rho_2 = \frac{2}{3} + \frac{1}{n}$ and $\rho_3 = \rho_4 = \frac{1}{n}$. Assume that we use an absolute priority policy $\pi = \{1, 2, 3, 4\}$. For such a system as $n \to \infty$, $\rho = \rho_{\{1,2,3,4\}} \to 1$ and also $\rho_{\{1,2\}}$, $\rho_{\{1,2,3\}} \to 1$. Hence, our method provides asymptotically exact results for $E[W_4]$ and $E[W_3]$ and approximate results for $E[W_1]$, $E[W_2]$. In Section 5 we illustrate that this approximation is quite effective as long as $\rho_i \geq 0.3$.

Note that in the case where preemption is allowed we established, as discussed the first remark after Theorem 8, an alternative numerical method of obtaining the exact formula of $b(\mathcal{N}_i)$ and hence of obtaining the *exact* performance analysis of a $\Sigma GI/G/1$ system under preemptive priorities.

4. POLLING SYSTEMS

In this section we consider the classical cyclic order polling system with general renewal arrival streams, independent service time distributions and a gated service strategy (see Takagi 1975). Polling systems are extensions of the $\Sigma GI/G/1$ queue, since a polling system is a $\Sigma GI/G/1$ in which the server follows a gated cyclic policy and there are changeover times when the server changes classes. Our contribution in this section is that we find in heavy traffic the performance of the mean waiting times and the cycle time by using extensively the distributional laws.

In Section 4.1 we introduce the model and our notation, while in Section 4.2 we analyze the model and construct a linear $N \times N$ system which once solved yields the expected performance measures.

4.1. Model Description and Notation

We consider a $\Sigma GI/G/1$ system, in which a single server is servicing N classes of customers in a cyclic order $1, \ldots, N$, $1, \ldots$ under a gated service discipline. One can visualize this process as if there were N queues, each corresponding to a different class, in a circle and the server services them cyclically in the following way: if there are N_{i-1} customers waiting in the $i-1$st queue when the server starts servicing this class, then the server processes all N_{i-1} customers, and after encountering a random delay, d_i it starts servicing the class i customers that are waiting in the ith queue. Notice that the class $i-1$ customers that arrive while the server is servicing the N_{i-1} customers, have to wait for the next visit of the server to the $i-1$st queue, i.e., for a full cycle to be completed. Traditionally these systems have been called *polling systems.*

We use the notation of Section 2 for the arrival processes and service time distributions. Let $\rho \triangleq \sum_{i=1}^{N} \rho_i < 1$ be the traffic intensity. Notice that for the gated cyclic policy the stability condition ($\rho < 1$) is independent of the changeover times (see Takagi 1975).

We also introduce the following additional notation:
T_i^k: the time that the server spends servicing the ith class in the kth visit;
θ_i^k: the station time, i.e., the time interval from the moment the server starts servicing class i until he starts servicing class $i+1$, during the kth visit;
C_i^k: the $(k-1)$st cycle with respect to class i, i.e., the time interval from the $(k-1)$st entrance to queue i until the kth entrance to queue i ($C_{N+1}^k = C_1^{k+1}$);
Δ_i^k: the intervisit time with respect to class i, i.e., the time between the end of the $(k-1)$st visit and the beginning of the kth visit to class i.
Furthermore, we let $\theta_i = \lim_{k\to\infty} \theta_i^k$, $C_i = \lim_{k\to\infty} C_i^k$, $\Delta_i = \lim_{k\to\infty} \Delta_i^k$.

4.2. Analysis of the Polling System

The departure point of our investigation is the following proposition.

Proposition 2. *In a $\Sigma GI/G/1$ polling system where the server is servicing customers cyclically using a gated policy, the expected waiting time of class i decomposes in heavy traffic as follows:*

$$E[W_i] \sim E[W_i^{GI/G/1}] + \frac{E[T_i \Delta_i]}{E[\Delta_i]} + \frac{E[(\Delta_i)^2]}{2E[\Delta_i]}, \tag{32}$$

where $E[W_i^{GI/G/1}]$ is the mean waiting time in a regular $GI/G/1$ queue.

Proof. The distributional laws hold for both the queue and the queue plus the service facility, thus

$$G_{L_i}(z) = \int_0^\infty K_i(z, t)\, dF_{S_i}(t) \quad \text{and}$$

$$G_{Q_i}(z) = \int_0^\infty K_i(z, t)\, dF_{W_i}(t).$$

Differentiating the above relations twice with respect to z we obtain (see Bertsimas and Nakazato 1995):

$$E[L_i] = \lambda_i E[S_i], \tag{33}$$

$$E[Q_i] = \lambda_i E[W_i], \tag{34}$$

$$E[L_i^2] = \lambda_i E[S_i] + 2\lambda_i E\left[\int_0^{S_i} E[N_{a_i}(\tau)]\, d\tau\right], \tag{35}$$

$$E[Q_i^2] = \lambda_i E[W_i] + 2\lambda_i E\left[\int_0^{W_i} E[N_{a_i}(\tau)]\, d\tau\right]. \tag{36}$$

Now let B_i be the event that at the arrival epoch of a random observer the server is servicing class i and $(B_i)^c$ the

complement of B_i, i.e., the event that the server is either switching among classes or is servicing class $j \neq i$ (equivalently the server is in the intervisit period of class i). By applying Little's law to the server we have that $P\{B_i\} = \rho_i$ and hence $P\{(B_i)^c\} = 1 - \rho_i$. By conditioning on the state of the server we have that:

$$G_{Q_i}(z) = \rho_i E[z^{Q_i}|B_i] + (1 - \rho_i)E[z^{Q_i}|(B_i)^c],$$

$$G_{L_i}(z) = \rho_i E[z^{Q_i+1}|B_i] + (1 - \rho_i)E[z^{Q_i}|(B_i)^c].$$

Combining the above relations we obtain

$$G_{L_i}(z) = zG_{Q_i}(z) + (1 - \rho_i)(1 - z)E[Q_i|(B_i)^c].$$

Differentiating twice with respect to z and then taking limits as $z \to 1$ we obtain

$$E[L_i^2] = 2E[Q_i] + \rho_i + E[Q_i^2] - 2(1 - \rho_i)E[Q_i|(B_i)^c]. \quad (37)$$

We next calculate $E[Q_i|(B_i)^c]$. Given the event $(B_i)^c$, the arrival of the random observer occurs during intervisit time Δ_i. Moreover, as the service policy is gated, the customers that are waiting in queue upon the arrival of the random observer must have arrived during the elapsed time from the beginning of the cycle C_i until the random observation time which occurred during Δ_i; let us denote this elapsed time by A_i. Due to the heavy traffic assumptions we have therefore that

$$E[Q_i|(B_i)^c] \sim E[N^*_{a_i}(A_i)|(B_i)^c] = \lambda_i E[A_i|(B_i)^c]. \quad (38)$$

On the other hand, we have that

$$E[e^{-sA_i}|(B_i)^c] = \int_0^\infty E[e^{-s(T_i+\Delta_i^*)}|\Delta_i = x, (B_i)^c] \cdot P\{x \leq \Delta_i \leq x + dx|(B_i)^c\}$$

$$= \int_0^\infty E[e^{-sT_i}|\Delta_i = x, (B_i)^c] \cdot E[e^{-s\Delta_i^*}|\Delta_i = x, (B_i)^c] \cdot P\{x \leq \Delta_i \leq x + dx|(B_i)^c\}.$$

Now, from renewal theory given the fact that there was a random incidence in an interval Δ_i that has a p.d.f $dP_{\Delta_i}(x)$, the new p.d.f. is

$$P\{x \leq \Delta_i \leq x + dx|(B_i)^c\} = \frac{x}{E[\Delta_i]} dP_{\Delta_i}(x).$$

Moreover, given that $\{\Delta_i|(B_i)^c\} = x$, the age Δ_i^* is uniformly distributed in $[0, x]$. Hence,

$$E[e^{-sA_i}|(B_i)^c] = \int_{x=0}^\infty \int_{t=0}^\infty e^{-st} P\{t \leq T_i \leq t + dt|\Delta_i = x\} \cdot \left[\int_{u=0}^x \frac{1}{x} e^{-su} du\right] \frac{x}{E[\Delta_i]} dP_{\Delta_i}(x).$$

Differentiating the above relation we get that

$$E[A_i|(B_i)^c] = \frac{E[(\Delta_i)^2]}{2E[\Delta_i]} + \frac{E[T_i\Delta_i]}{E[\Delta_i]}. \quad (39)$$

Combining (33)–(39) and using $S_i = W_i + X_i$, we have that:

$$E\left[\int_0^{S_i} E[N_{a_i}(\tau)]\, d\tau\right] \sim E[W_i] + E\left[\int_0^{W_i} E[N_{a_i}(\tau)]\, d\tau\right] - (1 - \rho)\left[\frac{E[(\Delta_i)^2]}{2E[\Delta_i]} + \frac{E[T_i\Delta_i]}{E[\Delta_i]}\right].$$

Since $S_i = W_i + X_i$ and W_i, X_i are independent we obtain

$$E\left[\int_{W_i}^{W_i+X_i} E[N_{a_i}(\tau)]\, d\tau\right] \sim E[W_i] - (1 - \rho)\left[\frac{E[(\Delta_i)^2]}{2E[\Delta_i]} + \frac{E[T_i\Delta_i]}{E[\Delta_i]}\right].$$

Finally, since $W_i \to \infty$, we can use the fact (Cox 1962) that $E[N_{a_i}(\tau)] \sim \lambda\tau + (c_{a_i}^2 - 1)/2$ and combine it with (22):

$$E[W_i^{GI/G/1}] \sim \frac{2\rho_i E[X_i^*] + E[X_i](c_{a_i}^2 - 1)}{2(1 - \rho_i)},$$

to prove (32). □

Remarks.

1. For Poisson arrival processes the previous relation is exact (see also the analysis in Takagi 1975) and can be written as follows:

$$E[W_i] = E[W_i^{M/G/1}] + \frac{E[T_i\Delta_i]}{E[\Delta_i]} + E[\Delta_i^*].$$

It agrees with the decomposition result in Cooper et al. (1995). Moreover, it generalizes the result for systems with general renewal arrival processes, under heavy traffic conditions.

2. A similar decomposition result can be proved in the case where the server services the queues cyclically and exhaustively, i.e., the queue at a station must be empty before the server moves to the next queue. In this case we can prove, using the same line of arguments as in Proposition 2, that

$$E[W_i] \sim E[W_i^{GI/G/1}] + \frac{E[(\Delta_i)^2]}{2E[\Delta_i]}, \quad (40)$$

where $E[W_i^{GI/G/1}]$ is the mean waiting time in a regular $GI/G/1$ queue.

The above decomposition result generalizes the decomposition result in polling systems with Poisson arrivals, in which $W_i = W_i^{M/G/1} \oplus \Delta_i^*$ (see Fuhrmann and Cooper 1985). Our result shows that in heavy traffic the expected waiting time decomposes even if we have general renewal arrivals for both the gated and the exhaustive policy.

Based on the above proposition we need to calculate $E[\Delta_i]$, $E[T_i\Delta_i]$, and $E[(\Delta_i)^2]$. For this reason, we next present the equations that describe the system.

Fundamental Equations of the System

From the definitions that we introduced in the previous section we obtain:

$$\theta_i^k = d_{i+1} + T_i^k, \tag{41}$$

$$C_i^k = \sum_{j=1}^{i-1} \theta_j^k + \sum_{j=i}^{N} \theta_j^{k-1}, \tag{42}$$

$$\Delta_i^k = C_i^k - T_i^{k-1} = C_i^k - \theta_i^{k-1} + d_{i+1}, \tag{43}$$

$$C_{i+1}^k = C_i^k - \theta_i^{k-1} + \theta_i^k. \tag{44}$$

Before stating the rest of the equations of the system we should notice that under heavy traffic conditions the intervisit time $\Delta_i^k \to \infty$ for all queues $i = 1, \ldots, N$ and visits k. Hence, under heavy traffic conditions, the moment that the server enters queue i, constitutes a random incidence for the ith arrival process.

Now let N_i^k be the number of customers that the server finds upon his arrival in the ith queue at his kth visit. Due to the nature of the cyclic model, these customers must have arrived during the cyclic time C_i^k. According to the previous discussion, the arrival of the server to queue i constitutes a *random incidence* for the arrival process of the ith queue, under heavy traffic; hence we have that

$$N_i^k \sim N_{a_i}^*(C_i^k).$$

Moreover, we know that T_i^k, the time the server spends servicing the ith queue in the kth visit, is exactly the time it takes the server to service those N_i^k customers. Hence,

$$T_i^k \sim \sum_{l=1}^{N_{a_i}^*(C_i^k)} X_{i,l}, \tag{45}$$

where $X_{i,l}$ represents the service time distribution for the lth customer among N_i^k. Therefore,

$$\theta_i^k \sim d_i + \sum_{l=1}^{N_{a_i}^*(C_i^k)} X_{i,l}. \tag{46}$$

Relations (41)–(46) constitute the equations that characterize the polling system. Based on them we will prove the following theorem.

Theorem 9. *For a gated polling system in heavy traffic the mean waiting times $E[W_i]$ for all $i = 1, \ldots, N$ are given as*

$$E[W_i] \sim \frac{1+\rho_i}{2\bar{C}} \operatorname{var}[C_i] + \frac{(1+\rho_i)\bar{C}}{2} + \frac{(c_{a_i}^2 - 1)E[X_i]}{2}, \tag{47}$$

where the $\operatorname{var}[C_i]$ *satisfy an $N \times N$ linear system of Equations* (61).

Proof. Our strategy is to find $E[\Delta_i]$, $E[T_i\Delta_i]$ and $E[(\Delta_i)^2]$ as functions of $\operatorname{var}[C_i]$.

STEP 1. Evaluation of $E[\Delta_i]$.
Using (42) and (43) and letting $k \to \infty$ we have that in steady-state:

$$E[\theta_i] = d_{i+1} + E[T_i], \quad E[\Delta_i] = E[C_i] - E[T_i]$$

$$\text{and} \quad E[C_i] = \sum_{j=1}^{N} E[\theta_j].$$

Notice that $E[C_i]$ is independent of i and we denote it by $\bar{C}$. Therefore,

$$\bar{C} = \sum_{j=1}^{N} E[\theta_j], \tag{48}$$

$$E[\theta_i] = d_{i+1} + E[T_i], \tag{49}$$

$$E[\Delta_i] = \bar{C} - E[T_i]. \tag{50}$$

Furthermore, from (45) we have that $E[T_i] \sim \lambda_i \bar{C} E[X_i] = \rho_i \bar{C}$, where $\rho_i \triangleq \lambda_i E[X_i]$ is the traffic intensity of class i. Combining the last equation with (50) and (49) we obtain:

$$E[\Delta_i] \sim \bar{C}(1-\rho_i) \quad \text{and} \quad E[\theta_i] \sim d_i + \rho_i \bar{C}. \tag{51}$$

Substituting in (48) we, finally, obtain:

$$\bar{C} \sim \frac{\sum_{i=1}^{N} d_i}{1 - \sum_{i=1}^{N} \rho_i}. \tag{52}$$

STEP 2. Evaluation of $E[T_i\Delta_i]$ and $E[(\Delta_i)^2]$.
Notice first that from (43) and Step 1 we have:

$$\begin{aligned} E[T_i\Delta_i] &= \lim_k E[T_i^{k-1}C_i^k] - \lim_k E[(T_i^{k-1})^2] \\ &= \lim_k E[C_i^k\theta_i^{k-1}] + \lim_k E[T_i^{k-1}]\bar{C} \\ &\quad - \lim_k \operatorname{var}[T_i^{k-1}] - \lim_k E[T_i^{k-1}]^2 \\ &= \gamma_i + \rho_i(1-\rho_i)\bar{C}^2 - \operatorname{var}[\theta_i], \end{aligned} \tag{53}$$

where we defined $\gamma_i \triangleq \lim_{k\to\infty} \operatorname{Cov}[C_i^k, \theta_i^{k-1}]$ and we used the facts that $E[T_i] = \rho_i\bar{C}$, and that from (41) we have $\operatorname{var}[T_i] = \operatorname{var}[\theta_i]$.

On the other hand we have from (43) that $\operatorname{var}[\Delta_i^k] = \operatorname{var}[C_i^k] + \operatorname{var}[\theta_i^{k-1}] - 2\operatorname{Cov}[C_i^k, \theta_i^{k-1}]$. Taking limits as $k \to \infty$ and adding and subtracting $E[\Delta_i]^2$ we have that

$$\begin{aligned} E[(\Delta_i)^2] &= \operatorname{var}[C_i] + \operatorname{var}[\theta_i] - 2\gamma_i \\ &\quad + (1-\rho_i)^2\bar{C}^2. \end{aligned} \tag{54}$$

Next, we need to evaluate $\operatorname{var}[\theta_i]$. By differentiating (46) twice and using heavy traffic expansions, we obtain

$$\begin{aligned} E[(\theta_i)^2] &\sim d_{i+1}E[\theta_i] + d_{i+1}\rho_i\bar{C} + \rho_i^2 E[(C_i)^2] \\ &\quad + \lambda_i E[(X_i)^2]\bar{C} + \lambda_i(c_{a_i}^2 - 1)(E[X_i])^2\bar{C}. \end{aligned}$$

Now, combining the previous relation with (51) we obtain

$$\begin{aligned} \operatorname{var}[\theta_i] &\sim \rho_i^2 \operatorname{var}[C_i] + \lambda_i E[(X_i)^2]\bar{C} \\ &\quad + \lambda_i(c_{a_i}^2 - 1)(E[X_i])^2\bar{C}. \end{aligned} \tag{55}$$

We are in position now to combine Step 1 and Step 2 together with Proposition 2 to obtain (47). To conclude the proof of the theorem we need to formulate the $N \times N$ system that yields the $\operatorname{var}[C_i]$.

STEP 3. Formulation of an $N \times N$ linear system.
To achieve our goal we will evaluate γ_i, for all $i = 1, \ldots, N$, using two different approaches and then we will equate the results. In particular, we start by taking variances in both sides of (44),

$$\begin{aligned}\text{var}[C_{i+1}^k] = \text{var}[C_i^k] + \text{var}[\theta_i^{k-1}] + \text{var}[\theta_i^k] \\ + 2(\text{Cov}[C_i^k, \theta_i^k] - \text{Cov}[C_i^k, \theta_i^{k-1}] \\ - \text{Cov}[\theta_i^k, \theta_i^{k-1}]). \qquad (56)\end{aligned}$$

We then evaluate $E[C_i^k, \theta_i^k]$ as follows $E[C_i^k, \theta_i^k] = E[C_i^k E[\theta_i^k | C_i^k]]$. Substituting for $E[\theta_i^k | C_i^k]$ from (51) we get

$$\begin{aligned}E[C_i^k, \theta_i^k] &\sim E[C_i^k(d_{i+1} + \rho_i C_i^k)] \\ &= d_{i+1} E[C_i^k] + \rho_i E[(C_i^k)^2].\end{aligned}$$

Using $\text{Cov}[Z_1, Z_2] = E[Z_1 Z_2] - E[Z_1]E[Z_2]$, and taking limits in the previous relation, we obtain

$$\lim_{k\to\infty} \text{Cov}[C_i^k, \theta_i^k] \sim \rho_i \text{ var}[C_i]. \qquad (57)$$

Following similar arguments (see Sarkar and Zangwill 1989 for a more detailed derivation for the Poisson case) we obtain:

$$\lim_{k\to\infty} \text{Cov}[\theta_i^{k-1}, \theta_i^k] \sim \rho_i \gamma_i. \qquad (58)$$

Substituting (55), (57), and (58) to (56) we obtain

$$\begin{aligned}\gamma_i \triangleq \lim_{k\to\infty} \text{Cov}[C_i^k, \theta_i^{k-1}] \sim \frac{1 + 2\rho_i + 2\rho_i^2}{2(1+\rho_i)} \text{var}[C_i] \\ - \frac{1}{2(1+\rho_i)} \text{var}[C_{i+1}] + \frac{H_i}{1+\rho_i}, \qquad (59)\end{aligned}$$

and

$$H_i \triangleq \bar{C}\lambda_i(\text{var}[X_i] + c_{a_i}^2 (E[X_i])^2).$$

We, now, follow exactly the analysis of the polling system with Poisson arrivals presented in Sarkar and Zangwill (1989) to obtain a second relation for γ_i. Namely, we use (42) to obtain that:

$$\begin{aligned}\text{Cov}[\theta_i^{k-1}, C_i^k] = \sum_{j=1}^{i-1} \text{Cov}[\theta_i^{k-1}, \theta_j^k] \\ + \sum_{j=i}^{N} \text{Cov}[\theta_i^{k-1}, \theta_j^{k-1}],\end{aligned}$$

or equivalently,

$$\gamma_i = \text{var}[\theta_i] + \sum_{j=1}^{i-1} y_{ij} + \sum_{j=i+1}^{N} x_{ji}, \qquad (60)$$

where $x_{ij} \triangleq \lim_{k\to\infty} \text{Cov}[\theta_i^k, \theta_j^k]$ and $y_{ij} \triangleq \lim_{k\to\infty} \text{Cov}[\theta_i^{k-1}, \theta_j^k]$ are linear in $\text{var}[C_k]$ (see Sarkar and Zangwill 1989).

By combining the last equation with (59) we obtain the following $N \times N$ linear system in $\text{var}[C_k]$. because they are identical with the analysis in Sarkar and Zangwill (1989).

$$\begin{aligned}&\left[\frac{1 + 2\rho_i - 2\rho_i^3}{2(1+\rho_i)} - \sum_{j=i+1}^{N} E_{j,i}^{(i)} - \sum_{j=1}^{i-1} F_{i,j}^{(i)}\right] \text{var}[C_i] \\ &- \left[\frac{1}{2(1+\rho_i)} + \sum_{j=i+1}^{N} E_{j,i}^{(i+1)} + \sum_{j=1}^{i-1} F_{i,j}^{(i+1)}\right] \text{var}[C_{i+1}] \\ &- \sum_{k \neq i, i+1} \left[\sum_{j=i+1}^{N} E_{j,i}^{(k)} + \sum_{j=1}^{i-1} F_{i,j}^{(k)}\right] \text{var}[C_k] \\ &\sim \frac{\rho_i H_i}{1+\rho_i} + \sum_{j=i+1}^{N} E_{j,i}^{(0)} + \sum_{j=1}^{i-1} F_{i,j}^{(0)}, \qquad (61)\end{aligned}$$

where $E_{i,j}^{(k)}$ and $F_{i,j}^{(k)}$ are recursively given as

$$\begin{aligned}E_{i,j}^{(0)} &\sim (a_i - \rho_i e_j) E_{i-1,j}^{(0)} \\ &\quad - a_i f_j E_{i,j+1}^{(0)} + \frac{H_{i-1}\rho_i}{a_{i-1}\rho_{i-1}} \quad \text{for } i - j = 2, \\ E_{i,j}^{(k)} &\sim (a_i - e_j \rho_i) E_{i-1,j}^{(k)} \\ &\quad - a_i f_j E_{i-1,j+1}^{(k)} + f_j E_{i,j+1}^{(k)} \quad \text{for } i - j = 2, \\ E_{i,j}^{(k)} &= (a_i - e_j \rho_i) E_{i-1,j}^{(k)} \\ &\quad - a_i f_j E_{i-1,j+1}^{(k)} + f_j E_{i,j+1}^{(k)} \quad \text{for } i - j \geq 3, \\ F_{i,j}^{(k)} &\sim (a_i - e_j \rho_i) F_{i-1,j}^{(k)} - a_i f_j F_{i-1,j+1}^{(k)} + f_j F_{i,j+1}^{(k)},\end{aligned}$$

for $k = 0, 1, 2, \ldots, N$ and $i - j \geq 2$, where

$$a_i \sim \frac{\rho_i(1+\rho_{i-1})}{\rho_{i-1}}, \quad f_i \sim \frac{1}{a_{i+1}}, \quad e_i \sim \frac{\rho_i}{(1+\rho_i)},$$

$$E_{j,j}^{(0)} \sim A_j, \quad E_{j,j}^{(k)} \sim \begin{cases} \rho_j^2 & \text{if } k = j, \\ 0 & \text{else,} \end{cases}$$

$$E_{j+1,j}^{(0)} \sim \frac{A_j \rho_{j+1}}{(1+\rho_j)},$$

$$E_{j+1,j}^{(k)} \sim \begin{cases} \dfrac{\rho_j(1+2\rho_j)\rho_{j+1}}{2(1+\rho_j)} & \text{if } k = j, \\ \dfrac{\rho_j \rho_{j+1}}{(1+\rho_{j+1})} & \text{if } k = j+1, \\ 0 & \text{else, } (j = 1, 2, \ldots, n-1), \end{cases}$$

$$F_{j,j}^{(0)} \sim \frac{\rho_j}{1+\rho_j} A_j, \quad (j = 1, 2, \ldots, n),$$

$$F_{j,j}^{(k)} \sim \begin{cases} \dfrac{\rho_j(1+2\rho_j+2\rho_j^2)}{2(1+\rho_j)} & \text{if } k = j, \\ -\dfrac{\rho_j}{2(1+\rho_j)} & \text{if } k = j+1, \\ 0 & \text{else,} \end{cases}$$

$$F_{j+1,j}^{(0)} \sim \frac{e_j \rho_{j+1}}{1+\rho_j} A_j + \frac{f_j \rho_{j+1}}{1+\rho_{j+1}} A_{j+1},$$

$$(j = 1, 2, \ldots, n-1),$$

Table II
The Expected Waiting Time in a $E_4/M/1$ and an $E_2/M/1$ Queue

	The $E_4/M/1$ Queue					The $E_2/M/1$ Queue				
ρ	Act.	DL	HT	DL dev	HT dev	Act.	DL	HT	DL dev	HT dev
0.40	0.234	0.042	0.417	−82.05%	+78.06%	0.366	0.250	0.500	−31.69%	+36.61%
0.50	0.416	0.250	0.625	−39.90%	+50.24%	0.600	0.500	0.750	−16.66%	+25.00%
0.60	0.707	0.563	0.937	−20.37%	+32.60%	0.963	0.875	1.125	−9.14%	+28.57%
0.70	1.208	1.084	1.458	−10.27%	+34.50%	1.573	1.500	1.750	−1.96%	+11.25%
0.80	2.228	2.125	2.500	−3.50%	+12.21%	2.804	2.750	3.000	−1.93%	+6.99%
0.90	5.302	5.250	5.625	−0.98%	+6.09%	6.550	6.500	6.750	−0.77%	+3.05%

$$F^{(k)}_{j+1,j} \sim \begin{cases} \dfrac{e_j\rho_j(1+2\rho_j)\rho_{j+1}}{2(1+\rho_j)} & \text{if } k=j, \\ \dfrac{e_j\rho_j\rho_{j+1}}{2(1+\rho_j)} + \dfrac{f_j\rho_{j+1}(1+2\rho_{j+1}+2\rho_{j+1}^2)}{2(1+\rho_{j+1})} & \text{if } k=j+1, \\ -\dfrac{f_j\rho_{j+1}}{2(1+\rho_{j+1})} & \text{if } k=j+2, \\ 0 & \text{else.} \end{cases}$$

□

Remarks.

1. The above asymptotic method is exact for a system with Poisson arrivals under any traffic intensity $\rho < 1$, and we obtain the results presented in Sarkar and Zangwill (1989).
2. The previous approach can be easily generalized to allow general random delays d_i.

5. NUMERICAL RESULTS

Our goal in this section is to evaluate numerically our proposed asymptotic method for the following systems:

(1) a single class $GI/G/1$ queue under FIFO,
(2) a multiclass $\Sigma GI/G/1$ queue under FIFO,
(3) a multiclass $\Sigma GI/G/1$ queue under a strict priority discipline,
(4) a polling system with general renewal arrivals.

Our goal is to address the following questions:

(a) What is the accuracy of our methods compared with simulation?
(b) How large ρ has to be for the results to be accurate?
(c) How does our method compare to the traditional heavy traffic approach?

5.1. The Single Class GI/G/1 Queue

We first consider a single class queue with the arrival process being either an Erlang-2 (E_2) or Erlang-4 (E_4) and the service time process being exponential of rate 1. In Table II we give the expected waiting time as a function of the traffic intensity for the simulation (Act.), our method (DL) and the traditional heavy traffic approach (HT) as well as the percent deviation of the two methods (dev) from the simulation.

As expected, the efficiency of both methods increases with the traffic intensity, and it is of approximately the same order of magnitude, although our method is slightly closer. Also the results for the $E_2/M/1$ are better than the results for the $E_4/M/1$. This is expected since our method is exact for the Poisson case, the closer the arrival process is to a Poisson process, the better our method becomes.

Comparing our results with those of the traditional heavy traffic we obtain:

$$E[W_{DL}] - E[W_{HT}] = \frac{\rho^2(c_x^2+1) - \rho(c_a^2+1)}{2\lambda(1-\rho)}$$

$$- \frac{\rho^2(c_x^2+c_a^2)}{2\lambda(1-\rho)} = \frac{E[X]}{2}(c_a^2-1),$$

so that as $\lambda \to 1/E[X]$, and hence $\rho \to 1$, the difference between the two methods remains constant. Moreover, depending on the sign of $c_a^2 - 1$ our method either provides smaller or larger predictions than the traditional HT approach. Finally, as $\rho \to 1$ both $E[W_{DL}]$ and $E[W_{HT}] \to \infty$ and therefore their difference vanishes and both become exact.

In Table III we present results for a $H_2/M/1$ queue with unit service rate and interarrival distribution $f_a(x) = pr_1e^{-r_1x} + (1-p)r_2e^{-r_2x}$. Changing the parameters p, r_1, r_2 we obtain the following table, where we just indicate the resulting ρ, c_a^2, r_1 and the waiting times.

It is instructive to note that for $c_a^2 < 1$ or $c_a^2 \sim 1$ our method slightly outperforms the HT, although things are reversed when $c_a^2 \gg 1$.

5.2. Three-class GI/G/1 Queue Under FIFO

We consider a $GI/G/1$ queue under FIFO with three customer classes: classes 1 and 3 have E_2 arrivals while class 2 has E_4 arrivals. All services are identical and either exponential of rate 1 or hyperexponential with rate 1 and $c_x^2 = 2$ (the parameter $r_1 = 1.5$).

Table III
The Expected Waiting Time for $H_2/M/1$ Queues

ρ	c_a^2	Act.	DL	HT	DL dev	HT dev	r_1
0.54	1.65	1.48	1.867	1.528	+25.5%	+3.3%	0.25
0.62	1.34	2.38	2.63	2.48	+10.7%	+4.1%	0.25
0.76	1.44	3.66	4.05	3.83	+10.7%	+4.8%	0.25
0.83	1.57	6.15	6.72	6.42	+9.2%	+4.3%	0.25
0.90	2.00	13.26	14.00	13.50	+5.6%	+1.8%	1.2
0.75	1.04	2.92	2.96	2.99	+1.3%	+2.3%	0.25

Table IV
Numerical Results for the Waiting Time in a Three-class FIFO *GI/G/1* Queue

				Exp. Service			Hyp. Service		
ρ	ρ_1	ρ_2	ρ_3	Act.	DL dev	HT dev	Act.	DL dev	HT dev
0.5	0.1	0.1	0.3	0.674	−32.41%	+81.75%	1.155	−17.52%	+92.64%
0.6	0.1	0.2	0.3	1.000	−22.47%	+56.20%	1.726	−11.76%	+62.95%
0.7	0.2	0.2	0.3	1.605	−13.80%	+35.00%	2.758	−7.58%	+38.99%
0.8	0.2	0.3	0.3	2.737	−12.74%	+21.03%	4.698	−5.43%	+23.72%
0.9	0.3	0.3	0.3	6.297	−1.54%	+9.17%	10.770	−1.73%	+3.29%

The performance of our asymptotic method as well as the heavy traffic method as described in Iglehart and Whitt (1970) is depicted in Table IV as a function of the traffic intensity. Notice that, once again, our method is closer. Moreover, it performs better for hyperexponential services as they increase the waiting time.

Furthermore, it is interesting to notice that for the same total traffic intensity both methods perform slightly worse in the case of the multiclass queue than in the single-class case (see Table II).

5.3. Two-class *GI/G/1* Queue Under Absolute Priority Policy

We consider a *GI/G/1* system with two classes of customers, under an absolute priority rule that gives nonpreemptive priority to class 1. The data for the system are presented in Table V.

The performance of the asymptotic *approximation* method is summarized in Table VI as a function of the vector of traffic intensities $\{\rho_1, \rho_2\}$. Notice that as long as the low priority class is concerned, the method performs better than in the case of a single class *GI/G/1* queue (see also Table II). This is expected since our asymptotic method performs better as the waiting time increases. Furthermore, by taking a single class *GI/G/1* queue, with any arrival process as input, adding a second class and imposing a nonpreemptive priority rule, we cause an increase of the waiting time for the initial class and consequently we improve the performance of our method in evaluating the waiting time of that class. Consequently, the accuracy of the method in evaluating the mean waiting time of the low priority class is extremely good even when this class has a low traffic intensity as long as ρ_1 is greater or equal to 0.4, and hence the waiting time for the second priority class is high.

Table V
Data for a Two-class Priority Queue

Class	Interarrival Distr.	Arrival Rate	Service Distr.	Service Rate
1	Erlang 2	ρ_1	Exponential	1
2	Erlang 3	$0.5 * \rho_2$	Exponential	2

5.4. Four-class *GI/G/1* Queue Under Absolute Priority Policy

In order to further check the robustness of our method we consider in this section a *GI/G/1* system with four classes of customers under an absolute priority nonpreemptive rule. The service time distributions for all nodes are Exponential with unit rate (recall that in order for the strong conservation laws to hold for such a system we require that all classes have the same service time distribution) and the characteristics of the different arrival processes are being summarized in Table VII:

Table VIII verifies that our method is accurate even when the traffic intensity is small (for example, we have an −18.8% deviation for $\rho_1 = 0.2$). Moreover, it constitutes an accurate estimate of the actual waiting time of class i if the total traffic intensity for all classes that have priority greater or equal to class i is greater than 0.4.

5.5. Ten-node Polling System

We consider a polling system with ten nodes under a gated cyclic policy. The performance of our method (DL) is presented in Table IX for five different systems. For all the

Table VI
Numerical Results for the Waiting Time in a Two-class Priority *GI/G/1* Queue

	High Priority Class				Low Priority Class			
ρ	ρ_1	Actual	DL	Dev. of DL	ρ_2	Actual	DL	Dev. of DL
0.6	0.4	0.542	0.416	−23.25%	0.2	1.411	1.25	−11.41%
0.7	0.4	0.625	0.500	−20.00%	0.3	2.094	1.945	−7.12%
0.7	0.5	0.813	0.700	−13.90%	0.2	2.776	2.612	−5.91%
0.8	0.5	0.914	0.800	−12.46%	0.3	4.566	4.417	−3.26%
0.8	0.6	1.228	1.125	−8.84%	0.2	6.192	6.042	−2.42%
0.8	0.4	0.707	0.584	−17.40%	0.4	3.447	3.334	−3.28%
0.9	0.5	1.005	0.900	−10.54%	0.4	9.923	9.834	−0.90%
0.9	0.6	1.351	1.250	−7.48%	0.3	13.35	13.34	−0.07%

Table VII
Data for a Four-class Priority *GI/G/1* Queue

System	Class 1 Arrivals		Class 2 Arrivals		Class 3 Arrivals		Class 4 Arrivals	
	Distr.	Rate	Distr.	Rate	Distr.	Rate	Distr.	Rate
A	Erlang 2	0.4	Erlang 3	0.2	Erlang 2	0.1	Erlang 3	0.1
B	Erlang 2	0.2	Erlang 3	0.1	Erlang 2	0.1	Erlang 3	0.4

Table VIII
Numerical Results for a Four-class *GI/G/1* Under Absolute Priorities

	Class 1			Class 2			Class 3			Class 4		
	DL	Act.	Dev.	DL	Act.	Dev.	DL	Act.	Dev.	DL	Act.	Dev.
A	0.92	1.04	−11.6%	2.08	2.36	−11.5%	4.44	4.76	−6.6%	8.47	8.94	−5.2%
B	0.69	0.84	−18.8%	0.86	1.17	−26.0%	1.29	1.55	−16.7%	4.10	4.40	−6.9%

systems the service distribution is common for all nodes and it is Exponential with rate 1 and the delay $d_i = 2$ for all i. The rest of the data is contained in Tables X and XI.

It is interesting to note that the asymptotic method performs extremely well even when the total traffic intensity is relatively small (0.4). Furthermore, by comparing the results we presented for different queueing systems we see that the performance of our method as a function of the traffic intensity, in polling systems is better than for any other system.

Notice that systems A and E are symmetric, where systems B, C, and D are highly asymmetric. In all cases, however, the performance of the method is affected only slightly.

Table IX
Numerical Results for a Ten-nodes Polling System

System	Total Traffic Intensity	DL Mean Waiting Time	Actual Mean Waiting Time	Deviation
A	0.40	17.48	17.48	0.000%
B	0.70	38.89	38.61	0.725%
C	0.90	144.30	143.36	0.655%
D	0.94	240.11	237.90	0.928%
E	0.85	75.59	75.90	0.410%

Table X
Data for the First Five Nodes of the Ten-node Polling System

	Node 1		Node 2		Node 3		Node 4		Node 5	
Syst.	ρ_1	$c_{a_1}^2$	ρ_2	$c_{a_2}^2$	ρ_3	$c_{a_3}^2$	ρ_4	$c_{a_4}^2$	ρ_5	$c_{a_5}^2$
A	0.04	1/2	0.04	1/2	0.04	1/2	0.04	1/2	0.04	1/2
B	0.05	1/2	0.05	1/2	0.05	1/2	0.05	1/2	0.05	1/2
C	0.01	1/2	0.01	1/2	0.01	1/2	0.01	1/2	0.41	1/2
D	0.01	1/2	0.02	1/4	0.01	1/6	0.02	1/4	0.41	1/2
E	0.09	1/2	0.09	1/8	0.09	1/2	0.09	1/8	0.04	1/2

Table XI
Data for the Last Five Nodes of the Ten-node Polling System

	Node 6		Node 7		Node 8		Node 9		Node 10	
System	ρ_6	$c_{a_6}^2$	ρ_7	$c_{a_7}^2$	ρ_8	$c_{a_8}^2$	ρ_9	$c_{a_9}^2$	ρ_{10}	$c_{a_{10}}^2$
A	0.04	1/4	0.04	1/4	0.04	1/4	0.04	1/4	0.04	1/4
B	0.05	1/4	0.05	1/4	0.05	1/4	0.05	1/4	0.25	1/4
C	0.01	1/4	0.01	1/4	0.01	1/4	0.01	1/4	0.41	1/4
D	0.01	1/6	0.02	1/6	0.01	1/2	0.02	1/4	0.41	1/2
E	0.09	1/8	0.09	1/2	0.09	1/8	0.09	1/2	0.09	1/8

5.6. A Two-node Polling System

To further check the robustness of our method, we consider a 2-node polling system, whose corresponding data is presented in Table XII.

Table XIII presents the performance of our method as a function, only, of the traffic intensity of both queues.

Notice once again that the the proposed method performs very well, even under moderate traffic, i.e., even for $\rho = 0.5$.

5.7. Insights from the Numerical Results

The following conclusions can be drawn from the numerical results, as well as from the nature of our method.

1. Our asymptotic method performs better as the waiting time increases. Therefore, the method performs substantially better when it predicts that the answer is large. Under this light it should not be surprising that the method performs extremely well in polling systems, (the presence of delays further increases the waiting time), very well in priority systems, and satisfactorily for systems under FIFO, even for moderate traffic. Interestingly, the performance of our method is inversely proportional to the difficulty of the system.

Table XII
Data for the Two-node Polling System

Node	Interarrival Distr.	Arrival Rate	Service Distr.	Service Rate	d
1	Erlang 2	ρ_1	Exponential	1	2
2	Erlang 4	ρ_2	Exponential	1	2

Table XIII
Numerical Results for a Two-node Polling System with Exponential Service

Traffic Intensity			Asymptotic Mean Waiting Time	Actual Mean Waiting Time	Deviation
ρ	ρ_1	ρ_2			
0.5	0.4	0.1	5.870	5.807	+1.080%
0.6	0.4	0.2	7.487	7.440	+0.632%
0.6	0.2	0.4	7.363	7.333	+0.409%
0.6	0.3	0.3	7.220	7.220	+0.000%
0.7	0.4	0.3	10.384	10.367	+0.164%
0.7	0.6	0.1	11.800	11.637	+1.229%
0.7	0.3	0.4	10.318	10.307	+0.107%
0.8	0.4	0.4	16.438	16.439	+0.000%
0.8	0.2	0.6	17.500	17.170	+1.922%
0.8	0.6	0.2	17.847	17.559	+1.640%
0.9	0.3	0.6	35.969	35.298	+1.900%
0.9	0.6	0.3	36.420	35.632	+2.211%

2. As our method is exact for Poisson arrivals, the closer the arrival processes are to Poisson, the better the performance of the method.

ACKNOWLEDGMENT

Research supported in part by a Presidential Young Investigator Award DDM-9158118 with matching funds from Draper Laboratory and by the National Science Foundation under grant DDM-9014751.

REFERENCES

ABATE, J. AND W. WHITT. 1995. Numerical Inversion of Laplace Transforms of Probability Distributions. *J. Comput.* **7,** 36–43.

BERTSIMAS, D. AND G. MOURTZINOU. 1996. A Unified Method to Analyze Overtake Free Queueing Systems. *Advances in Applied Probability* **28,** 588–625.

BERTSIMAS, D. AND D. NAKAZATO. 1995. The General Distributional Little's Law and Its Applications. *Opns. Res.* **43,** 298–310.

COFFMAN, E. G., A. A. PUHALSKII, AND M. I. REIMAN. 1993. Polling Systems with Zero Switchover Times: A Heavy-traffic Averaging Principle. Working Paper.

COOPER, R., S. NIU, AND M. SRINIVASAN. 1995. A Decomposition Theorem for Polling Models: The Switchover Times are Effectively Additive. *Opns. Res.* **44,** 629–634.

COX, D. R. 1962. *Renewal Theory*. Chapman and Hall, New York.

FEDERGRUEN, A. AND H. GROENEVELT. 1988a. M/G/c Queueing Systems with Multiple Customer Classes: Characterization and Control of Achievable Performance under Nonpreemptive Priority Rules. *J. Appl. Prob.* **24,** 709–724.

FEDERGRUEN, A. AND H. GROENEVELT. 1988b. Characterization and Optimization of Achievable Performance in Queueing Systems. *Opns. Res.* **36,** 733–741.

FENDICK, K. W., V. R. SAKSENA, AND W. WHITT. 1989. Dependence in Packet Queues. *IEEE Trans. Commun.* **37,** 1173–1183.

FUHRMANN, S. W. AND R. B. COOPER. 1985. Stochastic Decompositions in a M/G/1 Queue with Generalized Vacation. *Opns. Res.* **33,** 1117–1129.

GELENBE, E. AND I. MITRANI. 1980. *Analysis and Synthesis of Computer Systems*. Academic Press, London.

HAJI, R. AND G. NEWELL. 1971. A Relation Between Stationary Queue and Waiting Time Distributions. *J. Appl. Prob.* **8,** 617–620.

HEYMAN, D. AND M. SOBEL. 1982. *Stochastic Models in Operations Research*. Vol. 1, McGraw-Hill Book Company, New York.

HOSONO, T. 1981. Numerical Inversion of Laplace Transform and Some Applications to Wave Optics. *Radio Sci.* **16,** 1015–1019.

IGLEHART, D. L. AND W. WHITT. 1970. Multiple Channel Queues in Heavy Traffic: I and II. *Advances in Applied Probability*, **2,** 150–177, 355–364.

KEILSON, J. AND L. SERVI. 1988. A Distributional Form of Little's Law. *O. R. Lett.* **7,** 5, 223–227.

KEILSON, J. AND L. SERVI. 1990. The Distributional Form of Little's Law and the Fuhrmann-Cooper Decomposition. *O. R. Lett.* **9,** 4, 239–247.

KLEINROCK, L. 1975. *Queueing Systems. Vol. 1: Theory*. Wiley, New York.

LEMOINE, A. 1974. On Two Stationary Distributions for the Stable GI/G/1 Queue. *J. Appl. Prob.* **11,** 849–852.

MOURTZINOU, G. 1995. An Axiomatic Approach to Queueing Systems. Ph.D. Thesis, Massachusetts Institute of Technology.

REIMAN, M. AND B. SIMON. 1990. A Network of Priority Queues in Heavy Traffic: One Bottleneck Station. *Queueing Sys.* **6,** 35–58.

SARKAR, D. AND W. I. ZANGWILL. 1989. Expected Waiting Times for Nonsymmetric Cyclic Queueing Systems-Exact results and Applications. *Mgmt. Sci.* **35,** 12, 1463–1474.

SHANTIKUMAR, J. G. AND D. D. YAO. 1992. Multiclass Queueing Systems: Polymatroidal Structure and Optimal Scheduling Control. *Opns. Res.* **40,** 293–299.

SMITH, W. L. 1954. Asymptotric Renewal Theorems. *Proc. Roy. Soc. Edinb.* A, **64,** 9–48.

TAKACS, L. 1962. *Introduction to the Theory of Queues*. Oxford University Press, New York.

TAKAGI, H. 1975. *Analysis of Polling Systems*. The MIT Press, Massachusetts.

WHITT, W. 1971. Weak Convergence Theorems for Priority Queues Preemptive Resume Discipline. *J. Appl. Prob.* **8,** 74–94.

WHITT, W. 1982. Approximating a Point Process by s Renewal Process I: Two Basic Methods. *Opns. Res.* **30,** 125–147.

WHITT, W. 1983. The Queueing Network Analyzer. *The Bell System Technical J.* **62,** 2779–2815.

WHITT, W. 1991. A Review of $L = \lambda W$ and Extensions. *Queueing Sys.* **9,** 235–268.