Multiclass queueing systems in heavy traffic: an asymptotic approach based on distributional and conservation laws

Dimitris Bertsimas *Georgia Mourtzinou*

OR 281-93 October 1993

Multiclass queueing systems in heavy traffic: an asymptotic approach based on distributional and conservation laws

Dimitris Bertsimas * Georgia Mourtzinou †‡

October 1993

Abstract

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 more 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.

*Dimitris Bertsimas, Sloan School of Management and Operations Research Center, MIT, Cambridge, Ma 02139.

†Georgia Mourtzinou, Operations Research Center, MIT, Cambridge, Ma 02139.

‡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.

1 Introduction

The goal of the present 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 non-preemptive) 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 points in the analysis of queueing systems:

1. Our analysis is based on the following principle: Define 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.

2. In contrast, traditional heavy traffic analysis in queueing systems focuses in approximating various processes involved by appropriate Brownian motions. We feel, however, that the proposed approach gives a clearer perspective to 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 as well as cost or profit functions. As there are several important applications of the systems we consider in telecommunication, computer, transportation and job-shop manufacturing systems, there is a huge literature in analyzing their performance.

Related to System 1 ($\Sigma GI/G/1$ under FIFO) Iglehart and Whitt [8] prove heavy traffic limit theorems. 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 in [8], but compared with simulation results are very accurate. Related to System 2 ($\Sigma GI/G/1$ with priorities) Gelenbe and Mitrani [6], Federgruen and Groenevelt [3], [4] and Shantikumar and Yao [15] derive conservation laws for ex-

pected 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 [17] for a survey). Perhaps the most efficient algorithm for the analysis of polling systems with Poisson arrivals is due to Sarkar and Zangwill [14], 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, Reiman [13] 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 [7] derive the distributional laws for an overtake free single class system, and for the case of Poisson arrivals Keilson and Servi [9], [10] 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 [2] and Bertsimas and Mourtzinou [1], 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 2, we develop the multiclass distributional laws. In Sections 3, 4 and 5 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 6 we report numerical results, comparing our results with the traditional heavy traffic approach and

simulation.

2 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.

2.1 A review of the single class distributional law

Consider a general queueing system, with a *single* stationary renewal arrival process of rate λ, in which the interarrival time has Laplace transform $\alpha(s)$. We assume that the system satisfies the following conditions:

Assumptions A:

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

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

A.3 New arriving customers do not affect the time in the system (or the queue) 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 S (W) be the stationary time a customer spends in the system (queue) and let L (Q) the stationary number of the customers in the system (or queue) for a system that satisfies Assumptions A. Let also L^-, L^+ (Q^-, Q^+) be the number in the system (or in the queue) just before an arrival or just after a departure, respectively. We denote with $F_S(t) = P\{S \leq t\}$

and $F_W(t) = P\{W \leq t\}$ the distribution functions of S and W respectively and with $G_L(z) = E[z^L]$ and $G_Q(z) = E[z^Q]$ the generating functions of L and Q.

The single class distributional law can be stated as follows:

Theorem 1 *(Haji and Newell [7], Bertsimas and Nakazato [2]) For a system that satisfies Assumptions A, L and S (Q and W) are related in distribution by:*

$$L \stackrel{d}{=} N_a^*(S), \tag{1}$$

$$Q \stackrel{d}{=} N_a^*(W), \tag{2}$$

while

$$G_L(z) = \int_0^\infty K(z,t)\, dF_S(t), \tag{3}$$

$$G_Q(z) = \int_0^\infty K(z,t)\, dF_W(t), \tag{4}$$

with

$$K(z,t) = \sum_{n=0}^{\infty} z^n P\{N_a^*(t) = n\},$$

where

$$K^*(z,s) = \int_0^\infty e^{-st}\, K(z,t)\, dt = \frac{1}{s} - \lambda \frac{(1-z)(1-\alpha(s))}{s^2(1-z\alpha(s))}.$$

Remarks :

1. Relations (1) and (2) hold even if we relax the assumption that the arrival process is renewal and we consider the broader family of stationary arrival processes (see Haji and Newell [7]).

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

$$L^- \stackrel{d}{=} L^+ \stackrel{d}{=} N_a(S), \quad Q^- \stackrel{d}{=} Q^+ \stackrel{d}{=} N_a(W),$$

$$G_{L^-}(z) = G_{L^+}(z) = \int_0^\infty K_o(z,t)\, dF_S(t), \tag{5}$$

$$G_{Q^-}(z) = G_{Q^+}(z) = \int_0^\infty K_o(z,t)\, dF_W(t), \tag{6}$$

with

$$K_o(z,t) = \sum_{n=0}^{\infty} z^n P\{N_a(t) = n\},$$

where

$$K_o^*(z,s) = \int_0^\infty e^{-st}\, K_o(z,t)\, dt = \frac{1-\alpha(s)}{s(1-z\alpha(s))}.$$

2.2 The multiclass distributional law

We, now, consider a general queueing system, with N classes of customers having independent arbitrary renewal arrival streams and different service requirements. We assume that the system satisfies Assumptions A. Let $\alpha_i(s)$ be the Laplace transform of the interarrival distribution for the i^{th} class, with arrival rate $\lambda_i = -1/\dot{\alpha}_i(0)$ and square coefficient of variation $c_{a_i}^2$.

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 i^{th} class respectively. Given that they exist in steady state, let S_i (W_i) be the stationary time spent in the system (queue) for class i customers and let L_i (Q_i) be the stationary number of class i customers in the system (or queue). Finally let $L = \sum_{i=1}^N L_i$ ($Q = \sum_{i=1}^N Q_i$), $F_{S_i}(t) = P\{S_i \le t\}$ ($F_{W_i}(t) = P\{W_i \le t\}$) and $G_{L_1,\ldots,L_N}(z_1,\ldots,z_N) = E[z_1^{L_1}\ldots z_N^{L_N}]$ ($G_{Q_1,\ldots,Q_N}(z_1,\ldots,z_N) = E[z_1^{Q_1}\ldots z_N^{Q_N}]$).

The multiclass distributional law can be stated as follows:

Theorem 2 *For a queueing system that satisfies Assumptions A,*

$$G_{L_1,\ldots,L_N}(z_1,\ldots,z_N) = 1 + \sum_{i=1}^{N} \int_0^\infty \int_0^t \prod_{j\neq i} K_j(z_j,x)\ \ dK_i(z_i,x)\ \ dF_{S_i}(t) \tag{7}$$

$$G_{Q_1,\ldots,Q_N}(z_1,\ldots,z_N) = 1 + \sum_{i=1}^{N} \int_0^\infty \int_0^t \prod_{j\neq i} K_j(z_j,x)\ \ dK_i(z_i,x)\ \ dF_{W_i}(t), \tag{8}$$

with

$$K_i(z_i,t) = \sum_{n=0}^{\infty} z_i^n P\{N_{a_i}^*(t) = n\}.$$

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_i^{th} customer of the i^{th} class and S_{i,n_i} be his system time. Note that within each class, the customer who is numbered 1 is the customer who arrived most recently. The customer currently served, if the server is actually busy, must have the highest ordinal number in his class. Therefore, τ_{i,n_i} and S_{i,n_i} are ordered in the reverse time direction.

Let $T_{i,1}^* = \tau - \tau_{i,1}$ for $i = 1, \ldots, N$, i.e., $T_{i,1}^*$ is distributed as the forward recurrence time of the i^{th} arrival process, and $T_{i,n_i} = \tau_{i,n_i-1} - \tau_{i,n_i}$, $n_i \geq 2$, i.e., T_{i,n_i} is the interarrival time of the i^{th} arrival process.

The key observation of the proof is that for an observer to see, at the random observation epoch τ, at least n_i customers of the i^{th} class in the system, where $n_i \geq 1$, we must have that for $i = 1, \ldots, N$ the n_i^{th} customer of the i^{th} class is still in the system at that moment τ. Then, for $n_i \geq 1 \quad i = 1, \ldots N$

$$L_1 \geq n_1, \ldots, L_N \geq n_N \text{ if and only if } S_{1,n_1} > \tau - \tau_{1,n_1}, \ldots, S_{N,n_N} > \tau - \tau_{N,n_N}. \quad (9)$$

Note, that we have used Assumptions A.1 and A.2 here. Thus,

$$P\{L_1 \geq n_1, \ldots, L_N \geq n_N\} = P\{S_{1,n_1} > \tau - \tau_{1,n_1}, \ldots, S_{N,n_N} > \tau - \tau_{N,n_N}\}.$$

We, then, condition on the type of the customer that arrived first to the system and obtain:

$$P\{L_1 \geq n_1, \ldots, L_N \geq n_N\} =$$
$$\sum_{i=1}^{N} P\{\tau - \tau_{i,n_i} = max_j \left(\tau - \tau_{j,n_j}\right), \; S_{1,n_1} > \tau - \tau_{1,n_1}, \ldots, \; S_{N,n_N} > \tau - \tau_{N,n_N}\}.$$

Since the discipline is FIFO (Assumption A.2), the event $(\tau - \tau_{i,n_i} \geq \tau - \tau_{j,n_j}) \bigcap (S_{i,n_i} > \tau - \tau_{i,n_i})$ implies that $S_{j,n_j} > \tau - \tau_{j,n_j}$, $j \neq i$. Therefore,

$$P\{L_1 \geq n_1, \ldots, L_N \geq n_N\} = \sum_{i=1}^{N} P\{\tau - \tau_{i,n_i} = max_j \left(\tau - \tau_{j,n_j}\right) \text{ and } S_{i,n_i} > \tau - \tau_{i,n_i}\}.$$

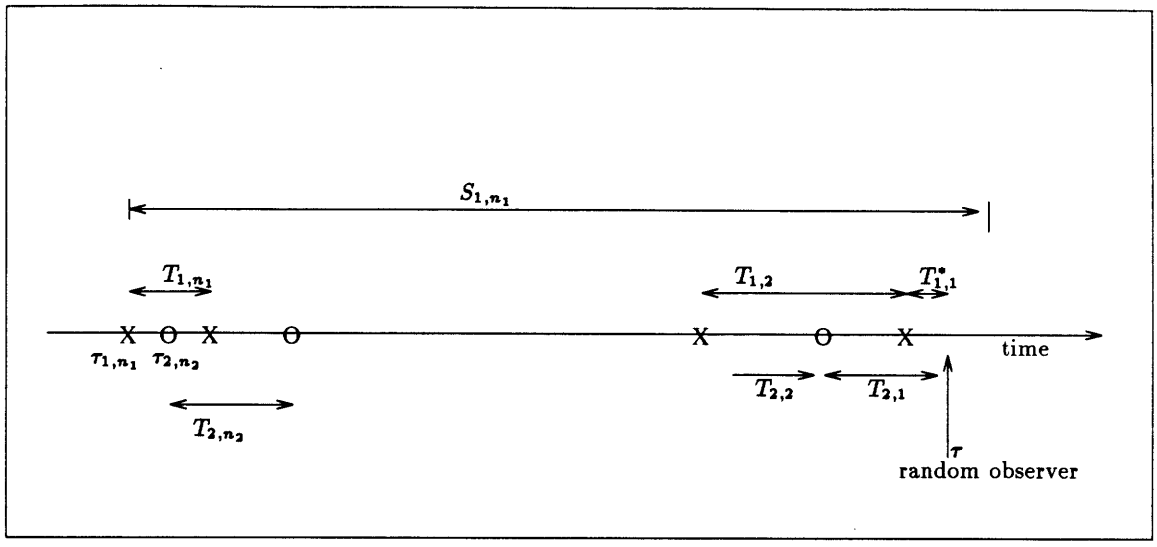

Figure 1: A possible observation scenario in the case of two customer classes.

Moreover, S_{i,n_i} is distributed as the stationary system time S_i, and because of Assumptions A.2 and A.3, S_{i,n_i}, $\tau - \tau_{i,n_i}$ are independent. We thus condition on S_i and obtain

$$P\{L_1 \geq n_1, \ldots, L_N \geq n_N\} = \sum_{i=1}^{N} \int_0^\infty P\{\bigcap_{j \neq i} (\tau - \tau_{i,n_i} \geq \tau - \tau_{j,n_j}),\ \tau - \tau_{i,n_i} < t\}\ dF_{S_i}(t).$$

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

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

and using the independence of $\tau - \tau_{j,n_j}$ for all $j = 1, \ldots, N$ (different arrival processes are independent) we obtain for $n_i \geq 1,\ \ i = 1, \ldots, N$

$$\begin{aligned} P\{L_1 \geq n_1, \ldots, L_N \geq n_N\} &= \sum_{i=1}^{N} \int_0^\infty \int_0^t \prod_{j \neq i} P\{\tau - \tau_{j,n_j} \leq x\}\ dA_{i,n_i}(x)\ dF_{S_i}(t) \\ &= \sum_{i=1}^{N} \int_0^\infty \int_0^t \prod_{j \neq i} A_{j,n_j}(x)\ dA_{i,n_i}(x)\ dF_{S_i}(t). \end{aligned} \tag{10}$$

We next consider the general case where the random observer, upon his arrival, does not see *any* customers from classes $k \in A \subset \{1, \ldots, N\}$ in the system, and sees $n_i \geq 1$

customers from class $i \notin A$. Similarly with relation (9), we obtain

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

Thus, following the derivation of (10), we obtain:

$$P\{\bigcap_{i \notin A}(L_i \geq n_i)\} = \sum_{i \notin A} \int_0^\infty \int_0^t \prod_{j \notin A, j \neq i} A_{j,n_j}(x)\, dA_{i,n_i}(x)\, dF_{S_i}(t), \tag{11}$$

for $n_i \geq 1, i \notin A$.

We next calculate $P\{L_1 = n_1, \ldots, L_N = n_N\}$ iteratively, based on (10) and (11) and using the fact that for $n_i \geq 0$

$P\{L_1 = n_1, \ldots, L_i = n_i, L_{i+1} \geq n_{i+1}, \ldots, L_N \geq n_N\} =$

$$P\{\bigcap_{k \leq i-1}(L_k = n_k), \bigcap_{j \geq i}(L_j \geq n_j)\} - P\{\bigcap_{k \leq i-1}(L_k = n_k),\ L_i \geq n_i + 1,\ \bigcap_{j \geq i+1}(L_j \geq n_j)\}.$$

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

$$G_{L_1,\ldots,L_N}(z_1, \ldots, z_N) = 1 + \sum_{i=1}^{N} \int_0^\infty \int_0^t \prod_{j \neq i} K_j(z_j, x)\ dK_i(z_i, x)\ dF_{S_i}(t),$$

where

$$\begin{aligned} K_i(z,t) = \quad & P\{T_{i,1}^* \geq t\} + \sum_{n=1}^{\infty} z^n \left\{ P\{T_{i,1}^* + \sum_{j=2}^{n} T_{i,n} > t\} - P\{T_{i,1}^* + \sum_{j=2}^{n+1} T_{i,n} > t\} \right\} \\ = \quad & \sum_{n=0}^{\infty} z^n P\{N_{a_i}^*(t) = n\}. \end{aligned}$$

Equation (8) is proved following exactly the same line of arguments if we restrict our attention to the number of customers in the queue. □

Remarks:

1. Note that for the case of a single class (7) reduces to (3).

2. The generating function of the total number L (Q) in the system (or in the queue) can be found if we set $z_1 = z_2 = \ldots = z_N = z$ in (7) and (8):

$$G_L(z) = 1 + \sum_{i=1}^{N} \int_0^\infty \int_0^t \prod_{j \neq i} K_j(z, x)\ dK_i(z, x)\ dF_{S_i}(t), \tag{12}$$

and

$$G_Q(z) = 1 + \sum_{i=1}^{N} \int_0^{\infty} \int_0^t \prod_{j \neq i} K_j(z,x)\ dK_i(z,x)\ dF_{W_i}(t). \tag{13}$$

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

(a) $\Sigma GI/G/1$ under FIFO for both L_i and Q_i,

(b) $\Sigma GI/D/s$ under FIFO for both L_i and Q_i,

(c) $\Sigma GI/G/s$ under FIFO for Q_i,

(d) multiclass systems with vacations (see [1], [10]).

2.3 Asymptotic forms of multiclass distributional laws

The distributional laws have a somewhat complicated form. Our goal in this section is to examine their implications as $L_i,\ Q_i,\ S_i,\ W_i \to \infty$. For the rest of this paper we only consider systems in which either the interarrival or the service times are *non-arithmetic*. It is well known that for these systems there is a natural parameter ρ, the traffic intensity, such that as $\rho \to 1$, $L_i,\ Q_i,\ S_i,\ W_i \to \infty$. The traffic intensity depends on the interarrival and service time characteristics of the particular system considered (for example in a $\Sigma GI/G/1$ queue, in which class i has arrival rate λ_i and mean service time $E[X_i]$, $\rho = \sum_{i=1}^{N} \lambda_i E[X_i]$). Therefore, whenever we say that a system is **under heavy traffic conditions**, we mean that $\rho \to 1$ and therefore, $L_i,\ Q_i,\ S_i,\ W_i \to \infty$. We will also use the notation that under heavy traffic conditions $g(x) \sim r(x)$ to mean that $\lim_{\rho \to 1} \frac{g(x)}{r(x)} = 1$.

As a preparation we need the following intermediate result:

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

$$K(z,t) = \sum_{n=0}^{\infty} z^n P\{N_a^*(t) = n\} \sim e^{-tf(z)},$$

and

$$K_o(z,t) = \sum_{n=0}^{\infty} z^n P\{N_a(t) = n\} \sim \frac{f(z)}{\lambda(1-z)} e^{-tf(z)},$$

where

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

Given a random variable Y, we will denote with $\phi_Y(s)$ the Laplace transform of Y. Then the asymptotic form of the distributional laws is as follows.

Theorem 4 *In a N-class queueing system that satisfies Assumptions A, the following asymptotic relations hold under heavy traffic conditions:*

$$G_{L_i}(z) \sim \phi_{S_i}(f_i(z)), \quad i = 1, \ldots, N \tag{14}$$

$$G_{Q_i(}(z) \sim \phi_{W_i}(f_i(z)), \quad i = 1, \ldots, N \tag{15}$$

$$G_{L_i^+}(z) \sim \frac{f_i(z)}{\lambda_i(1-z)}\phi_{S_i}(f_i(z)), \quad i = 1, \ldots, N \tag{16}$$

$$G_{Q_i^+}(z) \sim \frac{f_i(z)}{\lambda_i(1-z)}\phi_{W_i}(f_i(z)), \quad i = 1, \ldots, N \tag{17}$$

$$G_{L_1,\ldots,L_N}(z_1, \ldots, z_N) \sim \sum_{i=1}^{N} \frac{f_i(z_i)}{\sum_{j=1}^{N} f_j(z_j)} \phi_{S_i}(\sum_{k=1}^{N} f_k(z_k)), \tag{18}$$

$$G_{Q_1,\ldots,Q_N}(z_1, \ldots, z_N) \sim \sum_{i=1}^{N} \frac{f_i(z_i)}{\sum_{j=1}^{N} f_j(z_j)} \phi_{W_i}(\sum_{k=1}^{N} f_k(z_k)), \tag{19}$$

with

$$f_i(z) = \lambda_i(1-z) - \frac{1}{2}\lambda_i(1-z)^2(c_{a_i}^2 - 1) \quad i = 1, \ldots, N. \tag{20}$$

Proof

Substituting the asymptotic form of the individual kernels from Theorem 3 to (3), (4), (5) and (6), as well as (7) and (8) we obtain Theorem 4. □

The previous theorem is useful as it relates asymptotically the transform of the number of customers in the system (queue) to the transform of the time spent in the system (queue). Note that for Poisson arrivals the relations of the previous theorem are exact for all ρ, since $K(z,t) = K_o(z,t) = e^{-\lambda t(1-z)}$.

2.4 Another distributional law for the $\Sigma GI/G/1$ queue

In this section we consider a particular overtake free multiclass system, i.e., the $\Sigma GI/G/1$ queue. By generalizing the work of Lemoine [12] for the $GI/G/1$ queue we prove a new multiclass distributional law that involves the characteristics of the service time distribution.

There are N classes in system. Class i customers arrive at the system according to a renewal process of rate λ_i and square coefficient of variation $c^2_{a_i}$. Let X_i be the random variable corresponding to the service time of a class i customer. We denote with $E[X_i]$ and $c^2_{x_i}$ the mean and the square coefficient of variation of X_i. Let, also, X_i^* be the 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 $\rho_i = \lambda_i E[X_i]$ and $\rho = \sum_{i=1}^N \rho_i$. Let S_i (W_i) be the stationary time spent in the system (queue) for class i customers and by L_i (Q_i) the stationary number of the i class in the system (or queue), given that those quantities exist in steady state. Denote, also by L (Q) the stationary number of all the customers in the system (or queue).

Theorem 5 *In a $\Sigma GI/G/1$ queue that satisfies Assumptions A*

$$G_{Q_1,\ldots,Q_N}(z_1,\ldots,z_N) = (1-\rho) + \sum_{i=1}^{N} \rho_i \int_0^\infty K_{oi}(z_i,t) \prod_{j\neq i}^{N} K_j(z_j,t)\ \ dF_{W_i+X_i^*}(t), \quad (21)$$

and

$$G_{L_1,\ldots,L_N}(z_1,\ldots,z_N) = (1-\rho) + \sum_{i=1}^{N} z_i\ \rho_i \int_0^\infty K_{oi}(z_i,t) \prod_{j\neq i}^{N} K_j(z_j,t)\ \ dF_{W_i+X_i^*}(t), \quad (22)$$

where

$$K_i(z_i,t) = \sum_{n=0}^{\infty} z_i^n P\{N^*_{a_i}(t) = n\} \quad \textit{and} \quad K_{oi}(z_i,t) = \sum_{n=0}^{\infty} z_i^n P\{N_{a_i}(t) = n\}.$$

Furthermore, the following asymptotic relations hold under heavy traffic conditions:

$$G_{Q_1,\ldots,Q_N}(z_1,\ldots,z_N) \sim (1-\rho) + \sum_{i=1}^{N} \rho_i \frac{f_i(z_i)}{\lambda_i(1-z_i)} \phi_{W_i}\Big(\sum_{l=1}^{N} f_l(z_l)\Big)\ \phi_{X_i^*}\Big(\sum_{l=1}^{N} f_l(z_l)\Big), \quad (23)$$

and

$$G_{L_1,\ldots,L_N}(z_1,\ldots,z_N) \sim (1-\rho) + \sum_{i=1}^{N} z_i\ \rho_i \frac{f_i(z_i)}{\lambda_i(1-z_i)} \phi_{W_i}(\sum_{l=1}^{N} f_l(z_l))\ \phi_{X_i^*}(\sum_{l=1}^{N} f_l(z_l)), \quad (24)$$

where $f_i(z)$ is defined in (20).

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:

$$G_{Q_1,\ldots,Q_N}(z_1,\ldots,z_N) = (1-\rho) + \sum_{i=1}^{N} \rho_i\ E[z_1^{Q_1}\ldots z_N^{Q_N}|B_i], \quad (25)$$

and

$$G_{L_1,\ldots,L_N}(z_1,\ldots,z_N) = (1-\rho) + \sum_{i=1}^{N} z_i\ \rho_i\ E[z_1^{Q_1}\ldots z_N^{Q_N}|B_i]. \quad (26)$$

Moreover, due to 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 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 i^{th} process that constitutes a *random incidence* for the other arrival processes (see Figure 2).

Consequently, we must have n_i renewals of the i^{th} 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 j^{th} 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 j^{th} 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:

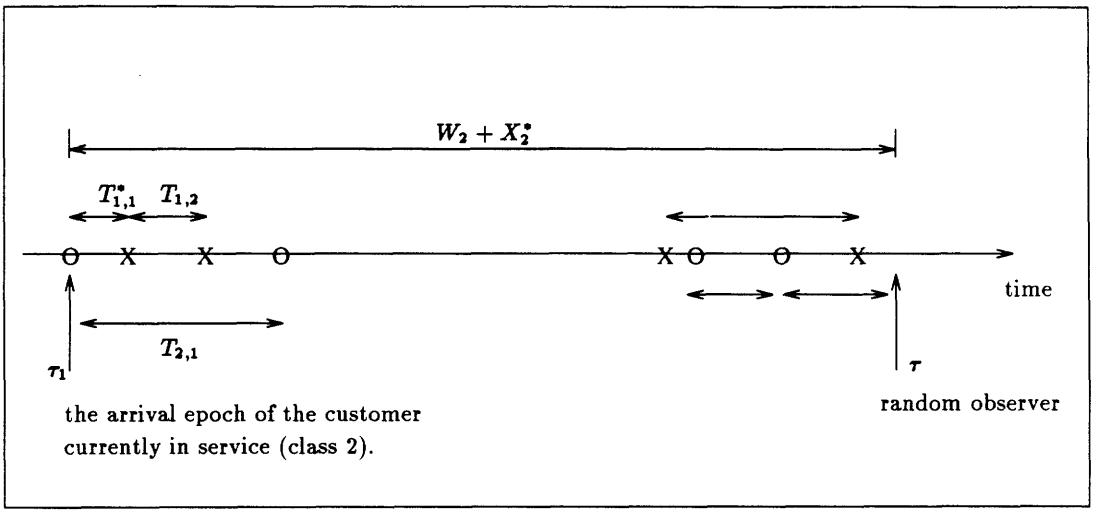

Figure 2: A possible observation scenario

$$P\{Q_1 = n_1, \ldots, Q_N = n_N | B_i\} =$$

$$P\{N^*_{a_1}(W_i+X_i^*) = n_1,\ N^*_{a_2}(W_i+X_i^*) = n_2, \ldots, N_{a_i}(W_i+X_i^*) = n_i, \ldots, N^*_{a_N}(W_i+X_i^*) = n_N\}. \tag{27}$$

By taking z-transforms we have:

$$E[z_1^{Q_1} \ldots z_N^{Q_N} | B_i] = \int_0^\infty K_{oi}(z_i, t) \prod_{j \neq i} K_j(z_j, t)\ \ dF_{W_i+X_i^*}(t), \tag{28}$$

where for $i = 1, \ldots, N$

$$K_i(z_i, t) = \sum_{n=0}^{\infty} z_i^n P\{N^*_{a_i}(t) = n\} \ \text{ and } \ K_{oi}(z_i, t) = \sum_{n=0}^{\infty} z_i^n P\{N_{a_i}(t) = n\}.$$

Substituting (28) into (25) and (26), we obtain (21) and (22). Moreover, using the asymptotic form of the kernels (Theorem 3) we establish (23) and (24). □

Remarks:

1. An interesting special case of (23) is a relation between $G_{Q_i}(z)$ $(G_{L_i}(z))$ for $i = 1, \ldots, N$ and the Laplace transforms of W_j, for $j = 1, \ldots, N$:

$$G_{Q_i}(z) \sim \quad (1-\rho) + \sum_{j=1, j\neq i}^{N} \rho_j \phi_{W_j}(f_i(z))\ \phi_{X_j^*}(f_i(z))$$

$$+\rho_i \frac{f_i(z)}{\lambda_i(1-z)} \phi_{W_i}(f_i(z))\ \phi_{X_i^*}(f_i(z)), \tag{29}$$

and

$$G_{L_i}(z) \sim \quad (1-\rho) + \sum_{j=1, j\neq i}^{N} \rho_j \phi_{W_j}(f_i(z))\ \phi_{X_j^*}(f_i(z))$$
$$+z\ \rho_i \frac{f_i(z)}{\lambda_i(1-z)} \phi_{W_i}(f_i(z))\ \phi_{X_i^*}(f_i(z)). \tag{30}$$

2. Another special case of (23) is a relation between $G_Q(z)$ ($G_L(z)$) and the Laplace transforms of W_j, for $j = 1, \ldots, N$, namely:

$$G_Q(z) \sim (1-\rho) + \sum_{i=1}^{N} \rho_i \frac{f_i(z)}{\lambda_i(1-z)} \phi_{W_i}(\sum_{l=1}^{N} f_l(z))\ \phi_{X_i^*}(\sum_{l=1}^{N} f_l(z)),$$

and

$$G_L(z) \sim (1-\rho) + z \sum_{i=1}^{N} \rho_i \frac{f_i(z)}{\lambda_i(1-z)} \phi_{W_i}(\sum_{l=1}^{N} f_l(z))\ \phi_{X_i^*}(\sum_{l=1}^{N} f_l(z)).$$

3. In the special case of a single class $GI/G/1$ queue (21) and (22) have been proved in Lemoine [12].

3 $\Sigma GI/G/1$ under FIFO

In this section we demonstrate that the distributional laws of the previous section lead to a complete solution of the $\Sigma GI/G/1$ under FIFO in heavy traffic. We use the notation of Section 2.4.

Theorem 6 *In a $\Sigma GI/G/1$ system under FIFO operating under heavy traffic conditions*

$$\phi_{W_i}(s) \sim (1-\rho)\ \frac{1+c(s)}{1-\rho_i \phi_{X_i^*}(s) \left[\frac{s}{\lambda_i(1-f_i^{-1}(s))} - 1\right]}, \tag{31}$$

and

$$G_{Q_i}(z) \sim (1-\rho)\ \frac{1+c(f_i(z))}{1-\rho_i \phi_{X_i^*}(f_i(z)) \left[\frac{f_i(z)}{\lambda_i(1-z)} - 1\right]}, \tag{32}$$

where $c(s) = D(s)/(1 - D(s))$ *and*

$$D(s) = \sum_{j=1}^{N} \frac{\rho_j \phi_{X_j^*}(s)}{1 - \rho_j \phi_{X_j^*}(s) \left[\frac{s}{\lambda_j (1 - f_j^{-1}(s))} - 1 \right]}.$$

The joint generating function of the number of customers in the queue is given by:

$$G_{Q_1,\ldots,Q_N}(z_1,\ldots,z_N) \sim \frac{(1-\rho)[1+c(g(\vec{z}))]}{g(\vec{z})} \sum_{i=1}^{N} \frac{f_i(z_i)}{1 - \rho_i \phi_{X_i^*}(g(\vec{z})) \left[\frac{g(\vec{z})}{\lambda_i (1 - f_i^{-1}(g(\vec{z})))} - 1 \right]}, \tag{33}$$

where $g(\vec{z}) = \sum_{k=1}^{N} f_k(z_k)$.

Proof

The distributional laws in Theorems 2 and 5 hold for both L_i and Q_i for all $i = 1, \ldots, N$. From (15) and (29), we obtain in heavy traffic for $i = 1, \ldots, N$

$$G_{Q_i}(z_i) \sim \phi_{W_i}(f_i(z_i)),$$

$$\begin{aligned} G_{Q_i}(z_i) \sim \quad & (1-\rho) + \sum_{j=1, j\neq i}^{N} \rho_j \phi_{W_j}(f_i(z_i))\ \phi_{X_j^*}(f_i(z_i)) \\ & + \rho_i \left(1 - \frac{1}{2}(1-z_i)(c_{a_i}^2 - 1)\right) \phi_{W_i}(f_i(z_i))\ \phi_{X_i^*}(f_i(z_i)) \quad i = 1, \ldots, N. \end{aligned}$$

Combining the previous equations pairwise, and setting for each i: $z_i = f_i^{-1}(s)$, we obtain for $i = 1, \ldots, N$:

$$\phi_{W_i}(s) \left(1 - \rho_i \phi_{X_i^*}(s) \frac{s}{\lambda_i (1 - f_i^{-1}(s))}\right) - \sum_{j \neq i} \rho_j \phi_{X_j^*}(s) \phi_{W_j}(s) \sim 1 - \rho.$$

The previous equations form a $N \times N$ linear system, which can be solved in closed form by adding and subtracting $\rho_i \phi_{X_i^*}(s) \phi_{W_i}(s)$. We can then solve each $\phi_{W_i}(s)$ as a function of $\sum_j \rho_j \phi_{X_j^*}(s) \phi_{W_j}(s)$, from where (31) follows. Moreover, because of (15), (32) follows.

Having found the transforms of $\phi_{W_i}(s)$, we obtain the joint transform of $(Q_1, \ldots, Q_N)$ from (19), which leads to (33). Note that we could use (23) instead. □

Remarks :

1. Since $S_i = W_i \oplus X_i$ and W_i, X_i are independent we obtain

$$\phi_{S_i}(s) = \phi_{W_i}(s)\,\phi_{X_i}(s) \quad i = 1, \ldots, N,$$

so that we can also find $\phi_{S_i}(s)$ and $G_{L_i}(z) \sim \phi_{S_i}(f_i(z))$.

2. The total number of customers in the queue can be found if we set $\vec{z} = (z, \ldots, z)$ in (33). If in addition all customer classes have the same service requirements, X, then we have from Theorem 4 as $\rho \to 1$:

$$G_L(z) \sim \phi_X(\sum_{i=1}^{N} f_i(z))\; G_Q(z),$$

and since $G_L(z) = zG_Q(z) + (1-\rho)(1-z)$, we obtain

$$G_Q(z) \sim \frac{1-\rho}{\phi_X(\sum_{i=1}^{N} f_i(z)) - z}. \tag{34}$$

3. In the case of a single class ($N = 1$) we obtain the results of [1] for the $GI/G/1$ queue.

4. For Poisson arrival processes $f_i(z) = \lambda_i(1-z)$, so that $\lambda_i(1 - f_i^{-1}(s)) = s$. Hence, we need to solve the following $N \times N$ system:

$$\phi_{W_i}(s)\left(1 - \rho_i \phi_{X_i^*}(s)\right) - \sum_{j \neq i} \rho_j \phi_{X_j^*}(s)\phi_{W_j}(s) = 1 - \rho \quad i = 1, \ldots, N,$$

from where we obtain, as it was expected,

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

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 $i = 1, \ldots N$*

$$E[W_i] \sim \frac{\sum_{j=1}^{N} \lambda_j E[X_j^2] + \rho_j E[X_j](c_{a_j}^2 - 1)}{2(1-\rho)} + \frac{1}{2} E[X_i](c_{a_i}^2 - 1). \tag{35}$$

Proof

From Little's law, :

$$E[Q_i] = \lambda_i E[W_i].$$

By differentiating (29) we obtain,

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

Combining the previous equations pairwise results in an $N \times N$ system of equations and solving the system yields (35). □

4 $\Sigma GI/G/1$ under general service disciplines

The techniques of the previous section lead to a complete solution only when the service discipline is FIFO. There are, however, many service disciplines (for example priority policies) that arise in practical situations and therefore it is interest to develop a methodology to analyze performance under arbitrary service disciplines. 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 arbitrary policies in heavy traffic.

4.1 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 non-anticipative* policies. For any policy $u \in \mathcal{U}$ and any class i, we let x_i^u to 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 $\mathbf{x}^u := (\mathrm{x}_i^u)_{i \in E}$ to be the performance vector under policy u. Finally, for any given permutation π of the N elements of E, we let $\mathbf{x}_i^\pi$ 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 formal definition of the strong conservation laws introduced in Shantikumar and Yao [15]:

Definition 1 *(Strong Conservation Laws) The performance vector* $\mathbf{x}$ *satisfies* strong conservation laws, *if there exists a set function* b*:* $2^E \to R_+$ *such that* $b(\emptyset) = 0$ *satisfying:*

$$\sum_{i \in A} x_i^\pi = b(A) \text{ for all } \pi : \{\pi(1), \ldots, \pi(|A|)\} = A \text{ and for all } A \subset E; \tag{36}$$

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

$$\sum_{i \in A} x_i^u \geq b(A) \text{ for all } A \subset E \text{ and } \sum_{i \in E} x_i^u = b(E). \tag{37}$$

In other words, a performance vector is said to satisfy strong conservation laws, if the total performance $\sum_{i \in E} x_i^u$ over all customer classes i is invariant under any admissible policy and the minimal performance $\sum_{i \in A} x_i^u$ over customer classes in a subset $A \subset E$ is achieved by an absolute priority policy giving priority to classes in the set A over all other classes in $E - A$.

The major result about systems that satisfy conservation laws is the following:

Theorem 7 *(Shantikumar and Yao [15]) Assume that the performance vector* $\mathbf{x}$ *satisfies strong conservation laws. Let* $P(b) = \{\mathbf{x} \in R^N | \sum_{i \in A} x_i^u \geq b(A), A \subset E \text{ and } \sum_{i \in E} x_i^u = b(E)\}$*. Then*

1. *$P(b)$ defines exactly those performance vectors that can be achieved under any policy u in $\mathcal{U}$.*

2. *The vertices of the polyhedron $P(b)$ are the performance vectors $\mathbf{x}^\pi$ of the absolute priority rules π. The performance vector of an absolute priority policy π, $\{\pi(1), \ldots, \pi(N)\} = E$, is given by:*

$$\begin{aligned} x^\pi_{\pi(1)} &= b(\{\pi(1)\}) \\ x^\pi_{\pi(2)} &= b(\{\pi(1), \pi(2)\}) - b(\{\pi(1)\}) \\ &\vdots \\ x^\pi_{\pi(N)} &= b(E) - b(\{\pi(1), \ldots, \pi(N-1)\}) \end{aligned}$$

3. *The polyhedron $P(b)$ is a* polymatroid, *i.e., the set function $b(\cdot)$ is supermodular, i.e., for any sets $A,\ B \subset E$, $b(A) + b(B) \leq b(A \cup B) + b(A \cap B)$.*

Therefore, an arbitrary policy in $\mathcal{U}$ gives rise to a performance vector $\mathbf{x}^u$ that is in $P(b)$. Moreover, if we know the set function $b(\cdot)$ we are able to calculate the performance of priority policies. Furthermore, as any policy $u \in \mathcal{U}$ can be obtained by an appropriate randomization among absolute priority policies, we can obtain the performance under any *work conserving and non-anticipative policy.* As a result, knowledge of the set function $b(\cdot)$ fully characterizes the achievable region.

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 [6]). 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 but we only deal with $\Sigma GI/G/1$ in this paper.

In Table 1 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 $b(\cdot)$ 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.

4.2 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 1 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 non-anticipative 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

System	Special characteristics		Performance measure	Evaluation of b
$\Sigma M/G/1$	N-classes	non-preemptive	$\rho_i E[W_i]$	[6]
$\Sigma GI/G/1$	N-classes	preemptive	$\rho_i E[W_i]$	Theorem 8a
$\Sigma GI/G/1$	N-classes same service	non-preemptive	$\rho_i E[W_i]$	Theorem 8b
$\Sigma GI/G/1$	2-classes	non-preemptive	$\rho_i E[W_i]$	Theorem 8c

Table 1: Systems satisfying strong conservation laws in steady state.

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,

$$b(A) \sim \frac{\rho_A \sum_{j \in A} \lambda_j E[X_j^2] + \sum_{j \in A} \rho_j E[X_j](c_{a_j}^2 - 1))}{2(1-\rho_A)}. \tag{38}$$

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

$$b(A) \sim \frac{\rho_A E[X^2] \sum_{i \in E} \lambda_i + E[X] \sum_{j \in A} \rho_j (c_{a_j}^2 - 1)}{2(1-\rho_A)}. \tag{39}$$

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

$$b(A) \sim \frac{\rho_A \sum_{i \in E} \lambda_i E[X_i^2] + \sum_{j \in A} \rho_j E[X_j](c_{a_j}^2 - 1)}{2(1-\rho_A)}. \tag{40}$$

Proof

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

$$b(A) = \sum_{i \in A} \rho_i E[W_i], \tag{41}$$

where $E[W_i]$ is the mean waiting time of the i^{th} 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 (35) we have evaluated $E[W_i]$ in heavy traffic. Substituting to (41) and rearranging (38) follows. (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:

$$E[Q_i] = \lambda_i E[W_i], \quad i \in E. \tag{42}$$

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

$$E[Q_i] = \sum_{j \in E} \rho_j E[Q_i |\ B_j], \quad i \in E. \tag{43}$$

In addition,

$$E[Q_i |\ B^j] = \lambda_i E[X^*] \quad i \in A,\ j \in E - A, \tag{44}$$

where $E[X^*]$ is the mean forward reccurence time of the service time distribution. This holds, because given the event B^j, the elapsed time since the initiation of the service of the class j customer is X^* and therefore, Q_i is exactly the number of customers of class $i \in A$ who arrived (according to the equilibrium renewal process) during X^*. Note that because we give priority to customers in A over those in $E - A$ we know that when the

service of the class j customer was initiated there were no customers present from class $i \in A$. From (27) we have that

$$E[Q_i|\ B^j] = E[N^*_{a_i}(W_j + X^*)] = \lambda_i\,(E[W_j] + E[X^*]) \quad i,j \in A,\ j \neq i, \qquad (45)$$

and

$$E[Q_i|\ B^i] = E[N_{a_i}(W_i + X^*)] \sim \lambda_i\,(E[W_i] + E[X^*]) + \frac{1}{2}(c^2_{a_i} - 1). \qquad (46)$$

Using equations (42)-(46) we obtain the following system of equations for $i \in A$:

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

Solving the above system yields (39).

(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 (44), (45) and (46) we use:

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

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

and

$$E[Q_i|\ B^i] \sim \lambda_i \left(E[W_i] + E[X^*_j]\right) + \frac{1}{2}(c^2_{a_i} - 1).$$

Using the above equations we form a $|A| \times |A|$ system, which, once solved, yields (40).

□

Remark:

For the case of Poisson arrivals and under non-preemption, (40) is exact. Moreover, under preemption, Poisson arrivals, and exponential service times ($\Sigma M/M/1$), (38) is also exact.

4.3 Applications of the achievable performance space

Having evaluated $b(A)$ in heavy traffic, our goal in this section is to illustrate how these closed form formulae can be used for various purposes.

Approximate 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 $\rho_1 E[W_1] = b(\{1\})$, $\rho_i E[W_i] = b(\{S_i\}) - b(\{S_{i-1}\})$, where $S_i = \{1, \ldots, i\}$.

We have evaluated $b(S_i)$ in heavy traffic, i.e., as long as $\rho_{S_i} \to 1$. But even if $\rho_{S_i} \not\to 1$ we can use the formulae for $b(S_i)$ as an approximation. In Section 6 we illustrate that this approximation is quite effective as long as $\rho_1 \geq 0.3$.

Optimization of a $\Sigma GI/G/1$ queue

The optimal solution for the problem $\min_{u \in \mathcal{U}} \sum_{i \in E} c_i E[W_i]$ is an absolute priority rule. In order to find which of the $n!$ priorities are optimal we do not need to know the set function $b(\cdot)$, as the optimal priority is the one that orders the classes according to the index $\frac{c_i}{\rho_i}$. As we argued before, we only need to know $b(\cdot)$ in order to understand the performance of the optimal policy. The situation is drastically different if we want to optimize a nonlinear objective function of the type $\min_{u \in \mathcal{U}} \sum_{i \in E} f(E[W_i])$. In this case we need to know $b(\cdot)$ in order to find the optimal policy, not only its performance. Again using the formulae we obtained for $b(\cdot)$ leads to an approximation of the optimal policy in this case.

5 Polling systems

In this section we consider the classical cyclic order polling system with general renewal arrival streams, independent service time distributions and an exhaustive service strategy. 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 an exhaustive 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 5.1 we introduce the model and our notation. In Section 5.2, we analyze

the system and express the expected performance measures in terms of the first two moments of a random variable related with the busy period in a $GI/G/1$, which are calculated in Section 5.3.

5.1 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 an exhaustive service discipline, i.e., if there are customers waiting to be serviced from the $i-1$ class when the server starts servicing this class, then the server processes all $i-1$ class customers until the system empties from them, and after encountering a random delay, d_i it starts servicing class i customers. One can visualize this process as if there were N queues in a circle and the server services them cyclically and exhaustively incurring a travel delay d_i when moving from the $i-1$ to the i queue. Traditionally these systems have been called *polling systems*. We use the notation of Section 2.4 for the arrival processes and service time distributions. Let $\rho = \sum_{i=1}^{N} \rho_i < 1$ be the traffic intensity. Notice that the stability condition is independent of the changeover times.

We also introduce the following additional notation:

T_i^k : the time that the server spends servicing the i^{th} class in the k^{th} visit;

θ_i^k : the station time, i.e., the time interval from the moment the server leaves class $i-1$ until he leaves class i, during the k^{th} visit;

C_i^k : the $(k-1)^{th}$ cycle with respect to class i, i.e., the time interval from the moment the server leaves class $i-1$ in the $(k-1)^{th}$ visit until he leaves class $i-1$ in the k^{th} visit $(C_{N+1}^k = C_1^{k+1})$;

Δ_i^k : the intervisit time with respect to class i, i.e., the time between the end of the $(k-1)^{th}$ visit and the beginning of the k^{th} visit to class i.

Furthermore, we let $\theta_i = \lim_{k\to\infty} \theta_i^k$, $C_i = \lim_{k\to\infty} C_i^k$, $\Delta_i = \lim_{k\to\infty} \Delta_i^k$.

5.2 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 and exhaustively, the expected waiting time of class i decomposes in heavy traffic as follows:*

$$E[W_i] \sim E[W_i^{GI/G/1}] + \frac{E[(\Delta_i)^2]}{2E[\Delta_i]}, \tag{47}$$

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

Proof

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

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

Furthermore, from Section 2.4 we have that:

$$E[Q_i|B_i] = E[N_{a_i}(W_i + X_i^*)] \sim \lambda_i(E[W_i] + E[X_i^*]) + \frac{1}{2}(c_{a_i}^2 - 1),$$

where X_i^* is the forward reccurence time of the service time distribution for class i. In addition,

$$E[Q_i|(B_i)^c] = E[N_{a_i}^*(\Delta_i^*)] = \lambda_i E[\Delta_i^*],$$

where Δ_i^* is the forward reccurence time of the intervisit time for class i. The reasoning for the above relation is that, given the event $(B_i)^c$, at the arrival of the random observer the elapsed time from the beginning of the intervisit time is Δ_i^* and therefore, as the service policy is exhaustive, the Q_i customers that are waiting in queue upon the arrival

of the random observer must have arrived during Δ_i^*. Combining the above relations we have that:

$$E[Q_i] \sim \rho_i \lambda_i (E[W_i] + E[X_i^*]) + \rho_i \frac{1}{2}(c_{a_i}^2 - 1) + (1 - \rho_i)\lambda_i E[\Delta_i^*]. \tag{48}$$

Using the fact that $E[Q_i] = \lambda_i E[W_i]$ and that as we proved in [2]

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

we prove (47). □

Remark:

The above decomposition result generalizes the decomposition result in polling systems with Poisson arrivals, in which $W_i = W_i^{GI/G/1} \oplus \Delta_i^*$ (see for example Fuhrmann and Cooper [5]). Our result shows that in heavy traffic the expected waiting time decomposes even if we have general renewal arrivals.

Based on the above proposition we need to calculate $E[\Delta_i]$ and $\text{var}[\Delta_i]$. 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:

$$\theta_i^k = d_i + T_i^k, \tag{49}$$

$$C_i^k = \sum_{j=1}^{i-1} \theta_j^k + \sum_{j=i}^{N} \theta_j^{k-1}, \tag{50}$$

$$\Delta_i^k = C_i^k - \theta_i^{k-1} + d_i, \tag{51}$$

$$C_{i+1}^k = C_i^k - \theta_i^{k-1} + \theta_i^k. \tag{52}$$

Before stating the rest of the fundamental equations of the system we should notice that under heavy traffic conditions the intervisit time $\Delta_j^k \to \infty$ for all queues $j = 1, \ldots, N$ and visits k. Hence the beginning of the busy period for queue j, denoted by B_j, constitutes a random incidence for the j^{th} arrival process. Subsequently, the beginning of the l^{th} sub-busy period for the j^{th} queue, denoted by $B_{l,j}$, is also a random incidence for the

j^{th} arrival process. Hence under heavy traffic conditions $B_{j,l} \stackrel{d}{=} B_j \quad \forall\, l, j$.

Let, now, N_j^k be the number of customers that the server finds upon his arrival in the j^{th} queue at his k^{th} visit. Due to the nature of the cyclic model these customers must have arrived during the intervisit time Δ_j^k. According to the previous discussion, the arrival of the server to queue j constitutes a *random incidence* for the arrival process of the j^{th} queue, hence by looking backwards in time as in the proof of the distributional laws we obtain

$$N_j^k \sim N_{a_j}^*(\Delta_j^k).$$

Moreover, we know that T_j^k, the time the server spends servicing the j^{th} queue in the k^{th} visit is independent of the service discipline. Hence, we can assume for the moment that we use non-preemptive LIFO (Last-In-First-Out) to conclude that under heavy traffic conditions:

$$T_j^k \sim \sum_{l=1}^{N_a^*(\Delta_j^k)} B_{j,l}, \tag{53}$$

where $B_{j,l}$ represents the l^{th} sub-busy period of the j^{th} queue, in which, due to the heavy traffic conditions , is identical distributed with the busy period B_j. Thus, for all i

$$\theta_i^k \sim d_i + \sum_{j=1}^{N_{a_i}^*(\Delta_i^k)} B_{i,j}. \tag{54}$$

Relations (49)-(54) constitute the equations that characterize the polling system. Our strategy to find $E[W_i]$ is to first find the first two moments of B_j, then proceed to find $E[\Delta_i]$ and $\mathrm{var}[\Delta_i]$.

Step 1: Evaluation of $E[B_i]$, $\mathrm{var}[B_i]$

These quantities in the expressions we have derived so far are calculated explicitly in Section 5.3 (Theorem 9) and are given as follows:

$$E[B_i] = \frac{E[X_i]}{1-\rho_i},$$

$$E[B_i^2] = (E[X_i])^2 \frac{c_{x_i}^2}{(1-\rho_i)^3} - \frac{(E[X_i])^2}{(1-\rho_i)^2} + (E[X_i])^2 \frac{E[\mathrm{var}[N_{a_i}^*(X_i)]]}{(1-\rho_i)^3}.$$

STEP 2: Evaluation of $E[\Delta_i]$.

Using (50), and (51) and letting $k \to \infty$ we have that in steady state:

$$E[\Delta_i] = E[C_i] - E[\theta_i] + d_i, \quad E[C_i] = \sum_{j=1}^{N} E[\theta_j].$$

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

$$\bar{C} = \sum_{j=1}^{N} E[\theta_j], \tag{55}$$

$$E[\Delta_i] = \bar{C} - E[\theta_i] + d_i. \tag{56}$$

Furthermore, from (54) we have that

$$E[\theta_i] \sim d_i + \lambda_i E[\Delta_i] E[B_i].$$

Combining the last equation with (56) we obtain :

$$E[\Delta_i] \sim \frac{\bar{C}}{1 + \lambda_i E[B_i]} \quad \text{and} \quad E[\theta_i] \sim d_i + \lambda_i E[B_i] \frac{\bar{C}}{1 + \lambda_i E[B_i]}. \tag{57}$$

Substituting in (55) we, finally, obtain:

$$\bar{C} \sim \frac{\sum_{i=1}^{N} d_i}{1 - \sum_{i=1}^{N} \frac{\lambda_i E[B_i]}{1 + \lambda_i E[B_i]}} \quad . \tag{58}$$

STEP 3: Evaluation of var$[\Delta_i]$.

The idea in this step is to express var$[\Delta_i]$ as a function of var$[C_i^k]$ for $k = 1, \ldots, N$ and then evaluate var$[C_i^k]$ by solving an $N \times N$ system. Notice, first, that from (51):

$$\text{var}[\Delta_i^k] = \text{var}[C_i^k] + \text{var}[\theta_i^{k-1}] - 2\text{Cov}[C_i^k, \theta_i^{k-1}].$$

Thus, in steady state

$$\text{var}[\Delta_i] = \text{var}[C_i] + \text{var}[\theta_i] - 2\gamma_i, \tag{59}$$

where $\gamma_i = \lim_{k \to \infty} \text{Cov}[C_i^k, \theta_i^{k-1}]$. In the next proposition we calculate var$[\theta_i]$ and γ_i as functions of var$[C_i]$.

Proposition 3 *Under heavy traffic:*

$$var[\theta_i] \sim \frac{var[C_i] - 2\gamma_i}{1-(\lambda_i E[B_i])^2} + \frac{\bar{C}\lambda_i(var[B_i] + c_{a_i}^2(E[B_i])^2)}{(1-(\lambda_i E[B_i])^2)(1+\lambda E[B_i])}, \tag{60}$$

$$\gamma_i \sim \frac{1}{2}var[C_i] - var[C_{i+1}]\frac{1-2\rho_i}{2} + \frac{A_i}{1-\rho_i.}, \tag{61}$$

where

$$A_i = E[C]\lambda_i(var[B_i] + c_{a_i}^2(E[B_i])^2)(1-\rho_i)^3.$$

Proof

From (54), we obtain

$$E[\theta_i] \sim d_i + \lambda_i E[\Delta_i]E[B_i],$$

$$\text{var}[\theta_i] \sim (\lambda_i E[B_i])^2 \text{var}[\Delta_i] + \lambda_i \text{var}[B_i]E[\Delta_i] + \lambda_i c_{a_i}^2 (E[B_i])^2 E[\Delta_i].$$

Now, combining the previous relation with (57) and (59) we obtain (60). By taking variances in both sides of (52) we obtain

$$\text{var}[C_{i+1}^k] = \text{var}[C_i^k] + \text{var}[\theta_i^{k-1}] + \text{var}[\theta_i^k] + 2\left(\text{Cov}[C_i^k, \theta_i^k] - \text{Cov}[C_i^k, \theta_i^{k-1}] - \text{Cov}[\theta_i^k, \theta_i^{k-1}]\right). \tag{62}$$

We first evaluate $E[\theta_i^{k-1}\theta_j^k]$ as follows:

$$E[\theta_i^{k-1}\theta_j^k] = E[\theta_i^{k-1}E[\theta_j^k \mid C_j^k, \theta_i^{k-1}]].$$

However, from (51) and (59) we have :

$$\theta_j^k \sim d_j + \sum_{l=1}^{N_{a_j}^*(C_j^k - \theta_j^{k-1} + d_j)} B_{j,l},$$

and therefore,

$$E[\theta_i^{k-1}\theta_j^k] \sim d_j E[\theta_i^{k-1}](1+\lambda_j E[B_j]) + \lambda_j E[B_j]\left(E[\theta_i^{k-1}C_j^k] - E[\theta_j^{k-1}\theta_i^{k-1}]\right).$$

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

$$\lim_{k\to\infty} \text{Cov}[\theta_i^{k-1}, \theta_i^k] \sim \lambda_i E[B_i](\gamma_i - \text{var}[\theta_i]), \tag{63}$$

Similarly,

$$\lim_{k\to\infty} \mathrm{Cov}[C_i^k, \theta_i^k] \sim \lambda_i E[B_i](\mathrm{var}[C_i] - \gamma_i), \tag{64}$$

Substituting (60), (63) and (64) to (62) we obtain (61). □

Until now we have expressed $\mathrm{var}[\Delta_i]$ as linear functions of the quantities $\mathrm{var}[C_j]$, $j = 1, \ldots, N$. We next form an $N \times N$ linear system to calculate $\mathrm{var}[C_j]$.

STEP 4: Formulation of an $N \times N$ linear system.

In this step we follow exactly the analysis of the polling system with Poisson arrivals presented in [14]. Namely, we use (50) to assert that:

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

or equivalently,

$$\gamma_i = \mathrm{var}[\theta_i] + \sum_{j=1}^{i-1} y_{ij} + \sum_{j=i+1}^{N} x_{ji}, \tag{65}$$

where $x_{ij} = \lim_{k\to\infty} \mathrm{Cov}[\theta_i^k, \theta_j^k]$ and $y_{ij} = \lim_{k\to\infty} \mathrm{Cov}[\theta_i^{k-1}, \theta_j^k]$. Then, we show that x_{ij} and y_{ij} are linear in $\mathrm{var}[C_k]$ and thus (65) can be written as:

$$\gamma_i = \mathrm{var}[\theta_i] + H_{i,j}^{(0)} + \sum_{j=1}^{i-1}\sum_{k=1}^{N} H_{i,j}^{(k)} \mathrm{var}[C_k] + G_{j,i}^{(0)} + \sum_{j=i+1}^{N}\sum_{k=1}^{N} G_{j,i}^{(k)} \mathrm{var}[C_k].$$

Finally we combine the last equation with (61) to obtain the following $N \times N$ linear system of equations, where we substitute for $E[B_i]$ from Step 1. We do not present the details because they are identical with the analysis in [14].

$$\begin{aligned}
&\left[\frac{1}{2} - \sum_{j=i+1}^{N} G_{j,i}^{(i)} - \sum_{j=1}^{i-1} H_{i,j}^{(i)}\right] \mathrm{var}[C_i] \\
&\quad - \left[\frac{1-2\rho_i+2\rho_i^2}{2} + \sum_{j=i+1}^{N} G_{j,i}^{(i+1)} - \sum_{j=1}^{i-1} H_{i,j}^{(i+1)}\right] \mathrm{var}[C_{i+1}] \\
&\quad\quad - \sum_{k\neq i,i+1} \left[\sum_{j=i+1}^{N} G_{j,i}^{(k)} - \sum_{j=1}^{i-1} H_{i,j}^{(k)}\right] \mathrm{var}[C_k] \\
&\quad\quad\quad \sim \frac{\rho_i A_i}{1-\rho_i} + \sum_{j=i+1}^{N} G_{j,i}^{(0)} - \sum_{j=1}^{i-1} H_{i,j}^{(0)},
\end{aligned}$$

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

$$G_{i,j}^{(k)} \sim (e_j - b_i\rho_i)G_{i,j+1}^{(k)} - a_i e_j G_{i-1,j+1}^{(k)} + a_i G_{i-1,j}^{(k)}, \tag{66}$$

$$H_{i,j}^{(k)} \sim (e_j - b_i\rho_i)H_{i,j+1}^{(k)} - a_i e_j H_{i-1,j+1}^{(k)} + a_i H_{i-1,j}^{(k)}. \tag{67}$$

for $k = 0, 1, 2, \ldots, N$ and $i - j \geq 3$, where

$$a_i \sim \frac{\rho_i}{\rho_{i-1}\,(1-\rho_i)}, \qquad b_i \sim \frac{\rho_i}{1-\rho_i}, \qquad e_j \sim \frac{\rho_j\,(1-\rho_j)}{\rho_{j+1}}, \tag{68}$$

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

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

$$G_{j+1,j}^{(0)} \sim \frac{\rho_j\,\rho_{j+1}}{(1-\rho_j)\,(1-\rho_{j+1})}\left[\frac{A_j}{\rho_j} - \frac{1-\rho_j}{1-\rho_{j+1}}A_{j+1}\right],$$

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

$$H_{j,j}^{(0)} \sim \left(\frac{\rho_j}{1-\rho_j}\right)^2 A_j,$$

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

$$H_{j+1,j}^{(0)} \sim \frac{\rho_j}{1-\rho_{j=1}}A_{j+1},$$

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

$$H_{j+2,j}^{(k)} \sim e_j H_{j+2,j+1}^{(k)} + \rho_j G_{j+2,j+1}^{(k)} \quad \text{for} \quad k \geq 0\,,$$

$$G_{j+2,j}^{(k)} \sim a_{j+2}G_{j+1,j}^{(k)} - e_j b_{j+2}H_{j+2,j+1}^{(k)} - \rho_j b_{j+2}G_{j+2,j+1}^{(k)} \quad \text{for} \quad k \geq 0\,.$$

After solving the system simple substitution into (59) yields the analytic formula for $\text{var}[\Delta_i]$ and thus we conclude the analysis.

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 [14].

2. The previous approach can be easily generalized to allow general random delays d_i.

5.3 Evaluation of the first two moments of B

In this section we evaluate the first two moments of B the busy period distribution of a queueing system, under the following condition:

Condition R:

The starting point of a busy period constitutes a random incidence for the arrival process. This condition naturally arises in analyzing polling systems in heavy traffic, since the server returns to a queue after a very long time and therefore, his arrival at the queue (and therefore, the initiation of a busy period) constitutes a random incidence for the arrival process. Notice, however, that B is not the actual busy period in a regular $GI/G/1$ queue (except if the arrival is Poisson). The technique we use is a generalization of the classical sub-busy period decomposition argument presented by Takacs in [16].

Consider a general queueing system with a *single* renewal arrival process with arrival rate λ and square coefficient of variation c_a^2. Denote by X the r.v. corresponding to the service time distribution and by $E[X]$ and c_x^2 its mean and square coefficient of variation, respectively. Denote, also, by ρ the traffic intensity. Furthermore, denote by B the r.v. that corresponds to the busy period distribution under condition R. Let $E[B]$, $E[B^2]$ be the first two moments of B.

Theorem 9 *In a $GI/G/1$ queue the following relations hold:*

$$E[B] = \frac{E[X]}{1-\rho}, \tag{69}$$

and

$$E[B^2] = (E[X])^2 \frac{c_x^2}{(1-\rho)^3} - \frac{(E[X])^2}{(1-\rho)^2} + (E[X])^2 \frac{E[var[N_a^*(X)]]}{(1-\rho)^3}, \tag{70}$$

Proof

We start by noticing that the duration of a busy period is invariant under any service discipline as long as it is work conserving. Hence, we can use the last-in-first-out (LIFO) service discipline. Assume that during the first customers waiting time K customers arrived. Each of these K customers initiates a sub-busy period, i.e., the time interval initialized by a customer entering service that lasts as long as all customers that arrived after him are being served (see also [11] p. 210).

Under Condition R, the number K of customers that arrive during the first service time that has duration X, is exactly $N_a^*(X)$. Moreover, the beginning of every sub-busy period constitutes a *random incidence* for the arrival process. If B_l is the duration of the lth sub-busy period

$$B = X + \sum_{l=1}^{N_a^*(X)} B_l,$$

where B_l has exactly the same distribution as B. Taking first and second moments we obtain:

$$E[B] = E[X] + E[N_a^*(X)]E[B], \tag{71}$$

$$E[B^2] = E[X^2] + 2E[X \sum_{i=1}^{N_a^*(X)} B_i] + E[(\sum_{i=1}^{N_a^*(X)} B_i)^2], \tag{72}$$

Since $E[N_a^*(X)] = \lambda E[X]$ we obtain:

$$E[B] = \frac{E[X]}{1-\rho}$$

Moreover,

$$E[X \sum_{i=1}^{N_a^*(X)} B_i] = \lambda E[B]E[X^2], \tag{73}$$

$$E[(\sum_{i=1}^{N_a^*(X)} B_i)^2] = \lambda E[X]E[B^2] - (E[B])^2 \left[\lambda E[X] - E[\text{var}[N_a^*(X)]] - \lambda^2 E[X^2]\right]. \quad (74)$$

Substituting, (73) and (74) in (72) we prove (70). □

6 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 multi-class GI/G/1 queue under FIFO,

(3) a multi-class 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) In the cases (1) and (2) above, in which there are alternative heavy traffic results, how the two methods differ?

6.1 The single class GI/G/1 queue

We 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. In Table 2 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 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. Furthermore, it is interesting to notice that our method provides a lower bound to the expected waiting time. We do not know if this happens accidentally. The fact that the heavy traffic method provides an upper bound is well known. Also the results for the $E_2/M/1$ are better than $E_4/M/1$. This is expected since our method is exact for the

	The $E_4/M/1$ Queue					The $E_2/M/1$ Queue				
ρ	Act.	DL	HT	Eff of DL	Eff of HT	Act.	DL	HT	Eff of DL	Eff of HT
0.40	0.234	0.042	0.417	17.95%	178.06%	0.366	0.250	0.500	68.31%	136.61%
0.50	0.416	0.250	0.625	60.1%	150.24%	0.600	0.500	0.750	83.34%	125.00%
0.60	0.707	0.563	0.937	79.63%	132.60%	0.963	0.875	1.125	90.86%	128.57%
0.70	1.208	1.084	1.458	89.73%	134.50%	1.573	1.500	1.750	98.04%	111.25%
0.75	1.610	1.500	1.875	93.17%	116.45%	2.060	2.000	2.250	97.08%	109.22%
0.80	2.228	2.125	2.500	96.50%	112.21%	2.804	2.750	3.000	98.07%	106.99%
0.85	3.256	3.167	3.542	97.27%	108.77%	4.041	4.000	4.250	98.98%	106.25%
0.90	5.302	5.250	5.625	99.02%	106.09%	6.550	6.500	6.750	99.23%	103.05%

Table 2: The expected waiting time in a $E_4/M/1$ and an $E_2/M/1$ Queue.

Poisson case, the closer the arrival process is to a Poisson process the better our method becomes.

6.2 3-Classes 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 exponential of rate 1. The performance of our asymptotic method as well as the heavy traffic method as

ρ	ρ_1	ρ_2	ρ_3	Act.	DL	HT	Eff. of DL	Eff. of HT
0.5	0.1	0.1	0.3	0.674	0.456	1.225	67.59%	181.75%
0.6	0.1	0.2	0.3	1.000	0.775	1.563	77.53%	156.2%
0.7	0.2	0.2	0.3	1.605	1.384	2.167	86.20%	135.00%
0.8	0.2	0.3	0.3	2.737	2.388	3.313	87.26%	121.03%
0.9	0.3	0.3	0.3	6.297	6.200	6.875	98.46%	109.17%

Table 3: Numerical results for the waiting time in a 3-classes FIFO GI/G/1 queue.

described in [8] is depicted in Table 3 as a function of the traffic intensity. Notice that, once again, our method is closer. Furthermore, it is interesting to notice that for the same total traffic intensity both methods perform worse in the case of the multi-class queue than in the single-class case (see Table 2).

6.3 2-Classes GI/G/1 queue under absolute priority policy

We consider a GI/G/1 system with 2 classes of customers,under an absolute priority rule that gives non-preemptive priority to class 1. The data for the system is presented in Table 4.

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

Table 4: Data for a 2-class priority queue.

The performance of the asymptotic *approximation* method is summarized in Table 5 as a function of the vector of traffic intensities $\{\rho_1, \rho_2\}$. Notice that as long as the high priority class is concerned, the method performs better than in the case of a single class GI/G/1 queue (see also Table 2). 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 non-preemptive 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.

ρ	High priority class				Low priority class			
	ρ_1	DL	Actual	Efficiency	ρ_2	DL	Actual	Efficiency
0.6	0.4	0.416	0.542	76.75%	0.2	1.25	1.411	88.59%
0.7	0.4	0.500	0.625	80.00%	0.3	1.945	2.094	92.88%
0.7	0.5	0.700	0.813	86.10%	0.2	2.612	2.776	94.09%
0.8	0.5	0.800	0.914	87.54%	0.3	4.417	4.566	96.74%
0.8	0.6	1.125	1.228	91.16%	0.2	6.042	6.192	97.58%
0.8	0.4	0.584	0.707	82.60%	0.4	3.334	3.447	96.72%
0.9	0.5	0.900	1.005	89.55%	0.4	9.834	9.923	99.10%
0.9	0.6	1.250	1.351	92.52%	0.3	13.34	13.35	99.93%

Table 5: Numerical results for the waiting time in a 2-classes priority GI/G/1 queue.

6.4 4-Classes 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 4 classes of customers under an absolute priority non-preemptive 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 6:

System	Class 1 arrivals		Class 2 arrivals		Class 3 arrivals		Class 4 arrivals	
	Distr.	Rate	Distr.	Rate	Distr.	Rate	Distr.	Rte
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 6: Data for a 4-classes priority GI/G/1 queue.

Table 7 verifies that our method is accurate even when the traffic intensity is small (for example we have an 81.2% efficiency 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.

	Class 1			Class 2			Class 3			Class 4		
	DL	Act.	Eff.	DL	Act.	Eff.	DL	Act.	Eff.	DL	Act.	Eff.
A	0.92	1.04	88.4%	2.08	2.36	88.5%	4.44	4.76	93.4%	8.47	8.94	94.8%
B	0.69	0.84	81.2%	0.86	1.17	74.0%	1.29	1.55	83.3%	4.10	4.40	93.1%

Table 7: Numerical results for a 4-classes GI/G/1 under absolute priorities.

6.5 10-Nodes polling system

We consider a polling system with 10 nodes under an exhaustive cyclic policy. The performance of our method (DL) is presented in Table 8 for 5 different systems. For all the 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 are contained in Tables 9 and 10.

System	Total traffic intensity	DL mean waiting time	Actual mean waiting time	Efficiency
A	0.40	15.96	16.43	97.1 %
B	0.75	30.54	30.50	100.1 %
C	0.90	69.60	68.67	101.4 %
D	0.94	123.65	119.75	96.8 %
E	0.85	64.67	63.59	101.6 %

Table 8: Numerical results for a 10-nodes polling system

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

Syst.	Node 1		Node 2		Node 3		Node 4		Node 5	
	ρ_1	$c^2_{a_1}$	ρ_2	$c^2_{a_2}$	ρ_3	$c^2_{a_3}$	ρ_4	$c^2_{a_4}$	ρ_5	$c^2_{a_5}$
A	.04	1/2	.04	1/2	.04	1/2	.04	1/2	.04	1/2
B	.05	1/2	.05	1/2	.05	1/2	.05	1/2	.05	1/2
C	.01	1/2	.01	1/2	.01	1/2	.01	1/2	.41	1/2
D	.01	1/2	.02	1/4	.01	1/6	.02	1/4	.41	1/2
E	.09	1/2	.09	1/8	.09	1/2	.09	1/8	.04	1/2

Table 9: Data for the first 5 nodes of the 10-node polling system.

System	Node 6		Node 7		Node 8		Node 9		Node 10	
	ρ_6	$c^2_{a_6}$	ρ_7	$c^2_{a_7}$	ρ_8	$c^2_{a_8}$	ρ_9	$c^2_{a_9}$	ρ_{10}	$c^2_{a_{10}}$
A	.04	1/4	.04	1/4	.04	1/4	.04	1/4	.04	1/4
B	.05	1/4	.05	1/4	.05	1/4	.05	1/4	.25	1/4
C	.01	1/4	.01	1/4	.01	1/4	.01	1/4	.41	1/4
D	.01	1/6	.02	1/6	.01	1/2	.02	1/4	.41	1/2
E	.09	1/8	.09	1/2	.09	1/8	.09	1/2	.09	1/8

Table 10: Data for the last 5 nodes of the 10-node polling system.

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,D are highly asymmetric. In all cases, however, the performance of the method is not affected.

6.6 A 2-Node polling system

In order to check the robustness of our method, we consider a 2-node polling system, whose corresponding data is presented in Table 11. Table 12 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 $\rho = 0.5$.

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 11: Data for the 2-node polling system.

6.7 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.

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

References

[1] Bertsimas, D. and Mourtzinou, G. (1992) "A unified method to analyze overtake free queueing systems", submitted for publication.

[2] Bertsimas, D. and Nakazato, D. (1991). "The general distributional Little's law and its applications", to appear in *Operations Research.*

Traffic intensity			Asymptotic	Actual	Efficiency
ρ	ρ_1	ρ_2	mean waiting time	mean waiting time	
0.5	0.4	0.1	2.968	3.253	91.2%
0.6	0.4	0.2	3.910	4.279	91.4%
0.6	0.2	0.4	3.851	4.188	92.0%
0.6	0.3	0.3	3.950	4.407	89.7%
0.7	0.4	0.3	5.270	5.784	91.1%
0.7	0.6	0.1	4.903	4.716	104.1%
0.7	0.3	0.4	5.234	5.730	91.4%
0.8	0.4	0.4	7.891	8.610	91.6%
0.8	0.2	0.6	7.789	7.450	104.5%
0.8	0.6	0.2	7.744	8.018	103.5%
0.9	0.3	0.6	15.945	15.813	100.8%
0.9	0.6	0.3	16.271	16.252	100.1 %

Table 12: Numerical results for a 2-nodes polling system with Exponential service.

[3] Federgruen, A. and Groenevelt, H., 1988a, M/G/c "Queueing Systems with Multiple customer Classes: Characterization and Control of Achievable Performance under Nonpreemptive Priority rules", *Journal of Applied Probability* 24, 709-724.

[4] Federgruen, A. and Groenevelt, H., 1988b, "Characterization and Optimization of Achievable Performance in Queueing systems", *Operations Research*, 36, 733-741.

[5] Fuhrmann S.W. and Cooper R.B. (1985). "Stochastic decompositions in a M/G/1 queue with generalized vacation", *Operations Research*, 33, 1117-1129.

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

[7] Haji, R. and Newell, G. (1971). "A relation between stationary queue and waiting time distributions", *Journal of Applied Probability*, 8, 617-620.

[8] Iglehart, D.L. and Whitt, W. (1970) "Multiple channel queues in heavy traffic, I and II", *Advances in Applied Probability,* 2, 150-177 and 355-364.

[9] Keilson, J and Servi, L. (1988) "A distributional form of Little's law", *Operations Research Letters*, Vol. 7, 5, 223-227.

[10] Keilson, J and Servi, L. (1990) "The distributional form of Little's law and the Fuhrmann-Cooper decomposition", *Operations Research Letters*, 9, 4, 239-247.

[11] Kleinrock, L. (1975). *Queueing systems; Vol. 1: Theory*, Wiley, New York.

[12] Lemoine, A. (1974) "On two stationary distributions for the stable GI/G/1 queue", *Journal of Applied Probability*, 11, 849-852.

[13] Reiman, M. (1993). Personal communication.

[14] Sarkar, D. and Zangwill, W.I., 1989, "Expected waiting times for nonsymmetric cyclic queueing systems-exact results and applications", *Management Science*, 35, 12, 1463-1474.

[15] Shantikumar, J.G. and Yao, D.D., 1992, "Multiclass Queueing Systems: Polymatroidal Structure and Optimal Scheduling Control", *Operations Research* 40, 293-299.

[16] Takacs, L. (1962). *Introduction to the theory of queues*, Oxford University Press, New York.

[17] Takagi, H. (1975). *Analysis of polling systems*, The MIT Press, Massachussets.

[18] Whitt, W. (1982). "Approximating a point process by s renewal process I: Two basic methods", *Operations Research* 30, 125-147.

[19] Whitt W. (1991). "A review of $L = \lambda W$ and extensions", *Queueing Systems* 9, 235-268.