Transient laws of non-stationary queueing systems and their applications

D. Bertsimas and G. Mourtizinou

WP# 3836-95-MSA June, 1995

Transient laws of non-stationary queueing systems and their applications

Dimitris Bertsimas [*] Georgia Mourtzinou [†‡]

June 1995

Abstract

We derive a set of transient distributional laws that relate the number of customers in the system (queue) at time t, denoted by $L(t)$ $(Q(t))$ and the system (waiting) time, $S(t)$ $(W(t))$ of a customer that arrived to the system (queue) at time t for non-stationary queueing systems that do not allow overtaking. These transient distributional laws provide a complete set of equations that describe the dynamics of the system in the transient domain, provide insight on the influence of the initial conditions and generalize the classical Little's law in the transient domain. Based on these transient distributional laws, we develop an efficient algorithm to analyze the performance of a single server queueing system with non-homogeneous Poisson arrivals and general time-dependent service requirements under arbitrary initial conditions. We further propose an asymptotic approach for single server multiclass systems and, for the single class case, report numerical results which are close to simulation and the traditional heavy traffic analysis via Brownian processes.

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

[†] Georgia Mourtzinou, Operations Research Center, MIT, Cambridge, MA 02139.

[‡] Research partially supported by a Presidential Young Investigator Award DDM-9158118 with matching funds from Draper Laboratory

1 Introduction

Non-stationary queueing systems have been used extensively to model complex production and service systems as well as communications and air-transportation systems.

In this paper we consider the general class of non-stationary queueing systems and we address the following questions: Are there "laws" of non-stationary queueing systems? In other words, is there a set of relationships between the fundamental quantities of non-stationary queueing systems that completely characterize the performance of such systems? If so, how can they be used in particular applications?

For systems in steady-state that do not allow overtaking, Keilson and Servi [14, 15] and Bertsimas and Mourtzinou [4, 5] demonstrated that the steady-state distributional laws constitute the "right" set of laws. For example, consider the $M/GI/1$ queue under FIFO and denote by L (Q) the steady-state number of customers in the system (queue) and by S (W) the steady-state system (waiting) time. Let also X be the service time. Then, we have from steady-state distributional laws (see Keilson and Servi [14, 15]) that

$$E[z^L] = E[e^{-\lambda(1-z)S}] \qquad \text{and} \qquad E[z^Q] = E[e^{-\lambda(1-z)W}].$$

Moreover, $S = W + X$ and the relation between Q and L is

$$E[z^L] = zE[z^Q] + (1-z)(1-\rho).$$

Combining the previous equations we obtain that:

$$E[e^{-sW}] = \frac{s\,(1-\rho)}{\lambda E[e^{-sX}] - \lambda + s}, \qquad E[e^{-sS}] = \frac{s\,(1-\rho)}{\lambda E[e^{-sX}] - \lambda + s} E[e^{-sX}],$$

$$E[z^Q] = \frac{(1-z)(1-\rho)}{E[e^{-\lambda(1-z)X}] - z}, \qquad E[z^L] = \frac{(1-z)(1-\rho)}{E[e^{-\lambda(1-z)X}] - z} E[e^{-\lambda(1-z)X}].$$

Hence, we obtain a complete description of the performance measures in the case of $M/GI/1$ queueing system under FIFO.

Motivated by the success of steady-state distributional laws, we develop in this paper a set of transient distributional laws that relate the transient performance measures of non-stationary queueing systems, i.e.,

- the number of customers in the system at time t, denoted by $L(t)$ and
- the system time, $S(t)$, of a customer that arrived to the system at time t.

It is important to notice that the form of the transient distributional laws depends on the initial conditions of the system and therefore it demonstrates the influence of the initial state on the evolution of the system. Furthermore, the transient distributional laws provide a complete set of equations that describe the dynamics of overtake-free non-stationary queueing systems.

Moreover, for general non-stationary queueing systems that may allow overtaking, we generalize the classical Little's law in the transient domain.

Finally, to demonstrate the power of transient distributional laws and the transient form of Little's law we apply them to a variety of particular queueing systems from single server systems with general non-stationary arrival and service distributions to infinite server systems with non-stationary Poisson arrivals and general non-stationary service distributions. For all specific systems we use the same approach: (a) Start the analysis by defining the performance measures of interest and (b) Relate the performance measures using the established set of laws. In this way we have a complete description of the system in the sense that we have a sufficient number of integral equations and unknowns. Our approach has parallels in physics, where there exist fundamental laws (laws of motion, Maxwell equations) that fully describe a physical system, and lead, using mathematical tools, to a complete solution for the quantities of interest. Once we formulate the stochastic system the next step is to actually solve it. For non-stationary Poisson arrival we can solve the system exactly. For general stationary arrivals, on the other hand, we use asymptotic expansions. Using the approach described above we are able to recover, in a unified way, existing results in the literature in the case of infinite server systems and also obtain new results in the case of single server systems.

The rest of this paper is structured as follows: In Section 2, we first review the steady-state distributional laws and then derive transient distributional laws for both single class and multiclass non-stationary queueing systems under arbitrary initial conditions. In Section 3 we develop a transient extension of the well-known Little's law, which holds under very general assumptions. In Section 4, we present the asymptotic expansions of the two kernels that are involved in the integral form of the transient distributional laws. In Section 5, we apply the transient laws to derive the transient performance analysis of several systems: infinite server systems with a single non-homogeneous Poisson arrival process and general time-dependent services, and multiclass single server systems with general time-dependent arrivals and services. Finally, in Section 6 we present some concluding remarks.

2 Transient distributional laws

In this section we present laws that relate the *distributions* of the number of customers in a queueing system and the system time for both single class queueing systems, where all the customers have the same characteristics, as well as multiclass systems, where each class of customers has some special characteristics and is treated differently by the system. These laws are called *distributional laws* and hold in both the steady state and the transient domain, for systems that satisfy the following assumptions:

Definition 1 ***(Distributional Laws Assumptions)***
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 of previous customers.
A.4 Arrival streams from different classes are mutually independent.

Assumption A.1 can be relaxed (see Mourtzinou [19]). Assumption A.2 is the crucial assumption that restricts the class of systems that admit distributional laws to the class of *overtake-free systems*, namely systems where customers exit in the order of their arrival. Assumption A.3 is exactly the lack of anticipation assumption needed for PASTA (see Wolff [26]) and is not particularly restrictive. Finally, Assumption A.4 is used only in the case of multiclass systems.

We define as **overtake free queueing systems** those systems that satisfy the Distributional Laws Assumptions and therefore, satisfy distributional laws. We use the notation $GI(t)/GI(t)/s$ to denote s-server systems with non-stationary arrival and service distributions, where successive interarrival and service times are mutually independent. The following systems are examples of overtake free systems:

(a) Multiclass $GI(t)/GI(t)/1$ queueing system under FIFO (where we can define "the system" to be either just the queue or the queue together with the server).

(b) Multiclass $GI(t)/D/s$ under FIFO (where we can define "the system" to be either just the queue or the queue together with the s servers).

(c) Multiclass $GI(t)/GI(t)/s$ under FIFO (where we 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) Non-stationary single-server systems where the server is unavailable for occasional intervals of time and customers are served under FIFO (see Bertsimas and Mourtzinou [4], Keilson and Servi [15]) (where, once again, we can define we can define "the system" to be either just the queue or the queue together with the server).

2.1 A review of steady-state distributional laws

In this section we first review steady-state distributional laws for single class systems, where all the customers have the same characteristics, and then we briefly review distributional laws for multiclass systems, where the system has N different customers classes.

The single class steady-state distributional law

Consider a general queueing system that satisfies Assumptions A.1-A.3. Customers arrive to the system according to a *single ordinary renewal* arrival process described by $N_a^o(t)$, the number of arrivals up to time t, where we use the term ordinary in the sense that all interarrival times, including the first one, are independently and identically distributed. We denote, also, by $N_a^e(t)$ the number of arrivals up to time t for the corresponding equilibrium renewal process, where the time for the first arrival is distributed as the forward recurrence time of the interarrival time of the ordinary renewal process (see Cox [8], p. 54).

We assume that the system is in steady-state and denote by L the steady-state number of customers in the system and by S the steady-state time a customers spend in the system, called the system time. Finally, we denote by $F_S(t) \triangleq P\{S \leq t\}$ the distribution function of S and by $G_L(z) \triangleq E[z^L]$ the generating function of L.

The single class steady-state distributional law can be stated as follows:

Theorem 1 *(Haji and Newell [10], Bertsimas and Nakazato [6]) For a system that satisfies Assumptions A.1-A.3 and has a single renewal arrival process, the steady-state number of customers, L, and the steady-state system time, S, are related in distribution by:*

$$L \overset{d}{=} N_a^e(S) \qquad \textit{equivalently} \qquad G_L(z) = \int_0^\infty K_e(z,t)\, dF_S(t), \tag{1}$$

where $K_e(z,t) \triangleq E[z^{N_a^e(t)}] = \sum_{n=0}^\infty z^n P\{N_a^e(t) = n\}$ *is the generating function of* $N_a^e(t)$.

Intuitively, (1) says that the number of customers in an overtake-free system in steady-state has the same distribution as the number of arrivals from the equilibrium renewal process during an interval of time distributed as the system time.

The multiclass steady-state distributional law

We consider now a *multiclass* queueing system, with N classes of customers. Customers of class i, $i = 1, \dots, N$ arrive at the system according to a renewal process with rate λ_i and have their own service requirements distributed according to a random variable X_i, $i = 1, \dots, N$. We assume that the system satisfies Assumptions A.1-A.4.

Let $N_{a_i}^o(t)$, $N_{a_i}^e(t)$ 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 S_i be the time spent in the system for class i customers in steady-state and let L_i be the number of class i customers in the system in steady-state. Finally let $L \triangleq \sum_{i=1}^N L_i$, $F_{S_i}(t) \triangleq P\{S_i \leq t\}$ and $G_{L_1,\dots,L_N}(z_1,\dots,z_N) \triangleq E[z_1^{L_1} \dots z_N^{L_N}]$.

The multiclass steady-state distributional law can be stated as follows:

Theorem 2 *(Bertsimas and Mourtzinou [5]) For a multiclass queueing system that satisfies Assumptions A.1-A.4, the joint generating function of the number of customers in the system from all classes and the individual system times are related as follows:*

$$G_{L_1,\dots,L_N}(z_1,\dots,z_N) = 1 + \sum_{i=1}^N \int_0^\infty \int_0^t \prod_{\substack{j=1 \\ j\neq i}}^N K_{e,j}(z_j,x)\ dK_{e,i}(z_i,x)\ dF_{S_i}(t), \tag{2}$$

where $K_{e,i}(z_i,t) \triangleq E[z_i^{N_{a_i}^e(t)}] = \sum_{n=0}^\infty z_i^n P\{N_{a_i}^e(t) = n\}$ *is the generating function of* $N_{a_i}^e(t)$.

Note that for each individual class Theorem 2 yields

$$G_{L_i}(z) = \int_0^\infty K_{e,i}(z,t)\ dF_{S_i}(t), \tag{3}$$

the single class distributional law of Theorem 1. Moreover, the generating function of the total number $L \triangleq \sum_{i=1}^N L_i$ in the system can be found if we set $z_1 = z_2 = \cdots = z_N = z$ in (2):

$$G_L(z) = 1 + \sum_{i=1}^N \int_0^\infty \int_0^t \prod_{\substack{j=1 \\ j\neq i}}^N K_{e,j}(z,x)\ dK_{e,i}(z,x)\ dF_{S_i}(t). \tag{4}$$

2.2 Transient single class distributional laws

In this section we generalize the single class distributional law to the transient domain for queueing systems satisfying Assumptions A.1-A.3.

We first introduce the necessary notation. We let T_j be the arrival time of the jth customer, with $T_0 = 0$ and $T_0 < T_1 < \cdots$, and S^j be his system time. We, also, let $N_a(t)$ be the number of arrivals in $(0, t]$ for all $0 < t \leq \infty$. Note that the counting process $N_a(t)$ is completely defined when we know the joint probability distribution of the T_j's via the relationship

$$N_a(t) \geq n \qquad \text{if and only if} \qquad T_n \leq t\,. \tag{5}$$

In the special case where $T_j - T_{j-1}$ for $j = 1, 2, \ldots$ are independent and identically distributed random variables, the arrival process is an ordinary renewal process and we use the notation $N_a^o(t)$ for the number of arrivals in $(0, t]$ for all $t > 0$ (see Section 2.1). Similarly, if $T_j - T_{j-1}$ for $j = 2, \ldots$ are independent and identically distributed random variables and T_1 is distributed as the forward recurrence time of $T_2 - T_1$, the arrival process is an equilibrium renewal process and we use the notation $N_a^e(t)$ for the number of arrivals in $(0, t]$ for all $t > 0$ (again see Section 2.1).

However, in the general case we assume, in accordance with Assumption A.3, that the time interval between two successive arrival epochs $A_i(T_i) \triangleq T_{i+1} - T_i$ is independent of the serial order i, for all $i = 1, 2, \ldots$, but it might depend on the value of T_i. Examples of such arrival processes, apart from the renewal processes we mentioned above, are (a) a non-homogeneous Poisson process of rate $\lambda(t)$ and although somewhat contrived (b) a counting process with $A_i(T_i)$ being uniformly distributed in $(2T_i, 2T_i + a)$ for fixed a.

We denote by $\sum_{i=1}^{n} A_i(x)$ the random variable that represents the sum of n sequential interarrival times with the first one starting at time x. We further define $N_a^o(t_o, t)$ to be the number of customers that arrived in the time interval $(t_o, t]$ given that t_o is an arrival epoch. The distribution of $N_a^o(t_0, t)$ can be calculated from the equivalence:

$$N_a^o(t_o, t) \geq n \qquad \text{if and only if} \qquad \sum_{i=1}^{n} A_i(t_o) \leq t - t_o\,. \tag{6}$$

Note, that for the case of renewal arrival processes $N_a^o(t_o, t)$ is the same as $N_a^o(t - t_o)$.

Finally, we let $h(t)\ \Delta t$ (as $\Delta t \rightarrow 0$) be the probability of an arrival in $(t, t + \Delta t]$. As we do not allow multiple arrivals (Assumption A.1), we have that

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t}\ P\{N_a(t + \Delta t) - N_a(t) > 0\}.$$

Note that for the special case of a nonhomogeneous Poisson arrival process with rate $\lambda(t)$, $h(t) = \lambda(t)$. On the other hand, for a (time homogeneous) renewal arrival process with arrival rate λ, the calculation of $h(t)$ depends on the distribution of the remaining time for the first customer to arrive to the system. If we assume that this is distributed as the forward recurrence time of the arrival process, then

$$h(t) = \lambda, \; t \geq 0.$$

This assumption physically means that we start counting arriving customers to the system at a random time relative to the arrival process. Moreover, naturally as $t \to \infty$, $h(t) \to \lambda$, as the influence of the initial distribution disappears.

The natural transient performance measures in such a system are

- $L(t)$ the number of customers in the system at time t characterized by its generating function

$$G_L(z,t) \triangleq E[z^{L(t)}] = \sum_{n=0}^{\infty} z^n P\{L(t) = n\},$$

- $S(t)$ the time that a customer who arrived at the system at time t spends in the system.

It is important to notice that $L(t)$ and $S(t)$ depend on the initial state of the system, i.e., the initial number of customers, $L(0)$, as well as on the initial work, $V(0) \triangleq \hat{V}(0) + \sum_{i=1}^{L(0)} X_i$, where $\hat{V}(0)$ is the set-up work in the system, which is independent of the number of initial customers, and X_i is the service requirement of the ith initial customer.

For ease of the presentation, we initially assume that the system is empty, i.e., $L(0) = 0$ with probability 1 (w.p.1) and $V(0) = 0$ w.p.1; we will relax this assumption later. In particular, the rest of this section is structured as follows. We first present the distributional law that relates $L(t)$ and $S(t)$ for a general system that satisfies Assumptions A.1-A.3 and starts empty. Next, we relax the initial condition assumption that the system starts empty and we extend the distributional laws to account for an arbitrary distribution of $L(0)$ and $V(0)$.

A transient law between L(t) and S(t) when the system starts empty

The transient distributional laws that relate the distributions of $L(t)$ and $S(t)$ when the system starts with no initial customers, i.e., $L(0) = 0$ w.p.1, and no initial work, i.e., $V(0) = 0$ w.p.1, is as follows.

Theorem 3 *For a queueing system that satisfies Assumptions A.1-A.3 and starts empty, the transient number in the system $L(t)$ and the transient system time $S(t)$ are related as follows:*

$$G_L(z,t) = 1 + (z-1) \int_0^t h(u) \; P\{S(u) > t-u\} \; K_o(z,u,t) \; du \; , \tag{7}$$

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

Proof: The proof of the relationship between $L(t)$ and $S(t)$ is based on the following observation. In an overtake-free system that starts empty, in order to have at least n ($n \geq 1$) customers in the system at time t, the nth most recently arrived customer with respect to t, i.e., the nth customer counting backwards in time, should still be in the system at time t.

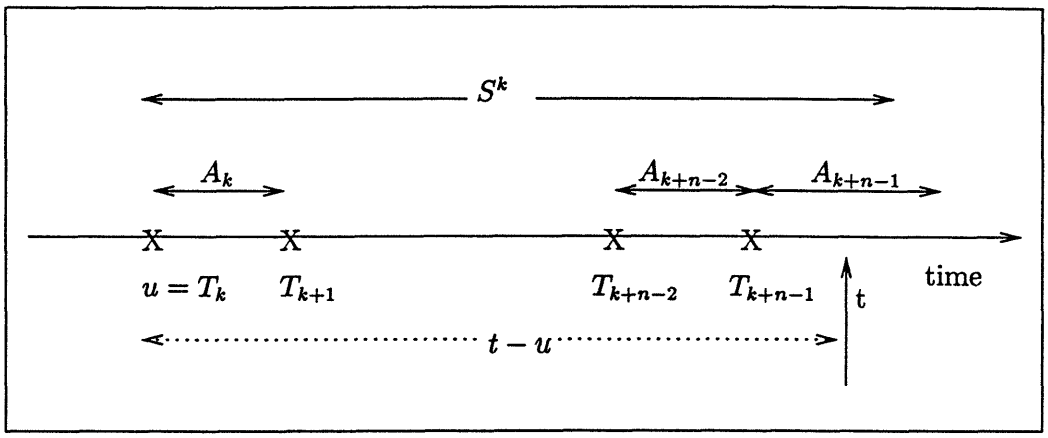


Figure 1: A scenario for a single class system in the transient regime.

This observation is based on Assumptions A.1 and A.2, since each customer arrives individually and stays in the system until served and also customers leave the system in the order of their arrival. Therefore, if the nth most recently arrived customer is in the system at time t, all the customers that came after him (and there are $n-1$ of those) are also still in the system at time t. Therefore, the event $\{L(t) \geq n\}$ is equivalent to the intersection of the following events:
$\mathbf{E}_1$: the nth most recently arrived customer with respect to t arrives at time u,
$\mathbf{E}_2$: his system time, given that he arrived at time u, is greater than $t-u$, for all $u \in (0,t]$.

We can further decompose event $\mathbf{E}_1$ into the event of an arrival at time u (that occurs with probability $h(u)\ du$) and the event of $n-1$ arrivals in $(u,t]$ given an arrival at u (that occurs with probability $P\{N_a^o(u,t) = n-1\}$). Furthermore, the probability of event $\mathbf{E}_2$ is $P\{S(u) > t-u\}$. Finally, according to Assumption A.3, $S(u)$ is independent of the path of the arrival process after time u and therefore events $\mathbf{E}_1$ and $\mathbf{E}_2$ are independent. The previous discussion leads to the relationship for $n \geq 1$:

$$P\{L(t) \geq n\} = \int_0^t h(u)\ P\{S(u) > t-u\}\ P\{N_a^o(u,t) = n-1\}\ du. \tag{8}$$

Given that $P\{L(t) \geq 0\} = 1$ and that $P\{L(t) = n\} = P\{L(t) \geq n\} - P\{L(t) \geq n+1\}$ we can easily calculate the generating function $G_L(z,t)$ to obtain (7). ■

Notice that from (6) we obtain the following alternative formula for $K_o(z,u,t) \triangleq E[z^{N_a^o(u,t)}]$,

$$K_o(z,u,t) = P\{T(u) > t-u\} + \sum_{n=1}^{\infty} z^n \left[P\{\sum_{k=1}^{n} A_k(u) \leq t-u\} - P\{\sum_{k=1}^{n+1} A_k(u) \leq t-u\} \right].$$

A transient law between L(t) and S(t) with arbitrary initial conditions

We, now, generalize the distributional law of Theorem 3 to account for the effect of initial customers.

We assume, that the system starts with k initial customers, i.e., $L(0) = k$ w.p.1 and initial work $V(0) = \hat{V}(0) + X_1 + \cdots + X_k$, where $\hat{V}(0)$ is the *set-up* work and X_i is the service requirement of the ith initial customer. We assume that the server finishes first the set-up work, then services the initial customers and then starts working on the customers that arrived after time 0.

Theorem 4 *For a queueing system that satisfies Assumptions A.1-A.3 and starts with $L(0) = k$ w.p.1 and $V(0) = \hat{V}(0) + X_1 + \cdots + X_k$, the transient number of customers in the system, $L(t)$, and the transient system time $S(t)$ are related as follows:*

$$G_L(z,t) = I^{(k)}(z,t) + P\{V(0) \leq t\} \left[1 + (z-1) \int_0^t h(u) P\{S(u) > t - u\} K_o(z,u,t) du\right], \tag{9}$$

where $K_o(z,u,t) \triangleq E[z^{N_a^o(u,t)}] = \sum_{n=0}^{\infty} z^n P\{N_a^o(u,t) = n\}$, $K(z,t) \triangleq E[z^{N_a(t)}] = \sum_{n=0}^{\infty} z^n P\{N_a(t) = n\}$, and $V_i \triangleq X_1 + \cdots + X_{k-i}$, $I^{(k)}(z,t) \triangleq K(z,t) \sum_{i=1}^{k} z^i \left[P\{\hat{V}(0) + V_i \leq t\} - P\{\hat{V}(0) + V_{i-1} \leq t\}\right]$.

Proof: Let $M(t)$ be the number of initial customers present in the system at time t, $M(t) \in \{1, \cdots, k\}$. Let also $V_i \triangleq X_1 + \cdots + X_{k-i}$. Then, since the server finishes first the set-up and starts servicing the initial customers we have that

$$\begin{aligned} &P\{M(t) = 0\} = P\{V(0) \leq t\} \\ &P\{M(t) = i\} = P\{\hat{V}(0) + V_i \leq t\} - P\{\hat{V}(0) + V_{i-1} \leq t\} \qquad \text{for all } i = 1, 2, \ldots, k. \end{aligned}$$

Let us define $G_{L^i}(z,t) \triangleq E[z^{L(t)} | M(t) = i]$, then

$$G_L(z,t) = P\{V(0) \leq t\} G_{L^0}(z,t) + \sum_{i=1}^{k} \left[P\{\hat{V}(0) + V_i \leq t\} - P\{\hat{V}(0) + V_{i-1} \leq t\}\right] G_{L^i}(z,t). \tag{10}$$

In the special case where $i = 0$ the analysis of Theorem 3 holds, i.e., in order to have at least n ($n \geq 1$) customers in the system at time t, given that no initial customer is present, the nth most recently arrived customer with respect to t should still be in the system at time t. Hence,

$$G_{L^0}(t) = 1 + (z-1) \int_0^t h(u)\ P\{S(u) > t - u\}\ K_o(z,u,t)\ du. \tag{11}$$

On the other hand, if $i = 1, 2, \ldots, k$ of the initial customers are present at time t we have that

$$\begin{aligned} &P\{L(t) = n \mid M(t) = i\} = P\{N_a(t) = n - i\} \qquad \text{for } n \geq i, \\ &P\{L(t) = n \mid M(t) = i\} = 0 \qquad \text{for } n < i. \end{aligned}$$

Therefore, if we define $K(z,t) \triangleq E[z^{N_a(t)}] = \sum_{n=0}^{\infty} z^n P\{N_a(t) = n\}$ we have that,

$$G_{L^i}(z,t) = \sum_{n=i}^{\infty} z^n P\{N_a(t) = n - i\} = z^i K(z,t). \tag{12}$$

Combining, (10), (11) and (12) we conclude the proof. ■

2.3 Transient multiclass distributional law

We, now, consider a general queueing system, with N classes of customers having independent arbitrary arrival streams and different service requirements. We assume that the system satisfies Assumptions A.1-A.4.

Let $N^o_{a_i}(u,t)$, be the number of customers from class i that arrived in the time interval $(u,t]$, given a class i arrival at time u, and $h_i(t)\Delta t$ (as $\Delta t \to 0$) the probability of a class i arrival in $(t, t+\Delta t]$. Furthermore, let $S_i(t)$ be the time spent in the system for class i customers that arrived at time t and let $L_i(t)$ be the number of class i customers in the system as observed at time t. Finally let $L(t) \triangleq \sum_{i=1}^N L_i(t)$, $\vec{z} \triangleq (z_1, \ldots, z_N)$ and $G_{L_1,\ldots,L_N}(\vec{z},t) \triangleq E[z_1^{L_1(t)} \ldots z_N^{L_N(t)}]$.

Assuming that the system starts empty the multiclass distributional law can be stated as follows:

Theorem 5 *For a queueing system that satisfies Assumptions A.1-A-4 and starts empty, we have that*

$$G_{L_1,\ldots,L_N}(\vec{z},t) = 1 - \sum_{j=1}^{N} \int_0^t \frac{\partial}{\partial a} K_{e,j}(z_j, a, t) \prod_{\substack{i=1 \\ i \neq j}}^{N} K_{e,i}(z_i, a, t) P\{S_j(a) > t - a\} da, \tag{13}$$

where $K_{o,i}(z_i, u, t) \triangleq E[z^{N^o_{a_i}(u,t)}] = \sum_{n=0}^{\infty} z_i^n P\{N^o_{a_i}(u,t) = n\}$ *and*

$$K_{e,i}(z_i, a, t) \triangleq 1 + (z_i - 1) \int_a^t h_i(u) K_{o,i}(z_i, u, t) du.$$

Proof: The essential observation of the proof is that, for all $i = 1, \ldots, N$, in order to have at time t at least n_i customers of the ith class in the system, where $n_i \geq 1$, we must have that the n_ith most recently arrived customer of the ith class is still in the system at t. Hence, the event $\{\bigcap_{i=1}^N (L_i \geq n_i)\}$ is equivalent to the intersection of the following events (for all $t_i \in (0,t]$ and for all $i = 1 \ldots N$):

$\mathbf{E_{1,i}}$: a customer of the ith class arrives at time t_i,
$\mathbf{E_{2,i}}$: the system time of the customer who arrived at t_i is greater than $t - t_i$,
$\mathbf{E_{3,i}}$: there are *exactly* $n_i - 1$ arrivals at $(t_i,\ t]$ given an arrival at t_i for the ith class.

Therefore taking probabilities we can write that:

$$P\{\bigcap_{i=1}^{N} (L_i \geq n_i)\} = \int_{t_1=0}^{t} \cdots \int_{t_N=0}^{t} P\{\bigcap_{i=1}^{N} E_{1,i} \bigcap_{i=1}^{N} E_{2,i} \bigcap_{i=1}^{N} E_{3,i}\}\ dt_1 \cdots dt_N.$$

From Assumption A.3 events $\mathbf{E_{1,i}}$, $\mathbf{E_{2,i}}$ and $\mathbf{E_{3,i}}$ are independent for any fixed t_i. Moreover, from Assumption A.4, the events $\mathbf{E_{1,i}}$ and $\mathbf{E_{3,i}}$ for all $i = 1, \ldots, N$, are also mutually independent. Hence, we can write that

$$P\{\bigcap_{i=1}^{N} (L_i \geq n_i)\} = \int_0^t \cdots \int_0^t P\{\bigcap_{i=1}^{N} E_{2,i}\} \prod_{i=1}^{N} P\{E_{1,i}\} P\{E_{3,i}\}\ dt_1 \cdots dt_N.$$

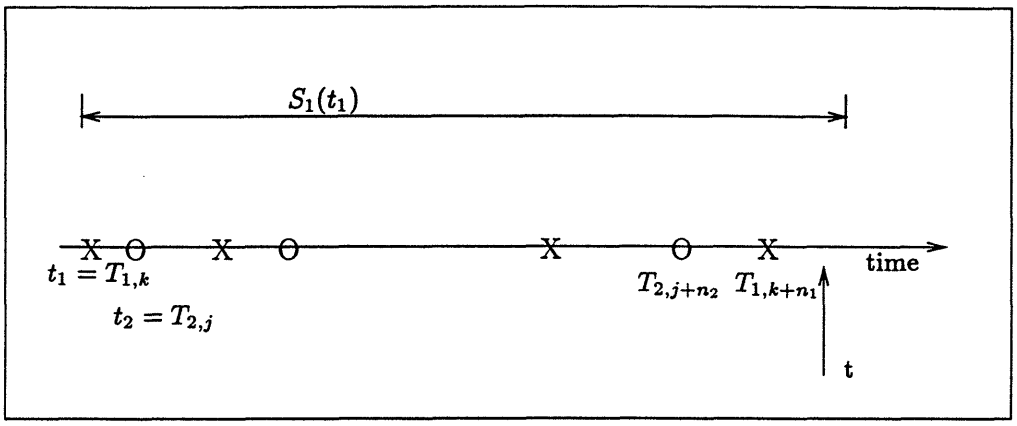


Figure 2: A scenario for a 2-class system in the transient regime.

Conditioning on the type of customer that arrived first to the system we have

$$P\{\bigcap_{i=1}^{N}(L_i \geq n_i)\bigcap(\tau_{j,n_j} = \min_i \tau_{i,n_i})\} =$$
$$\int_{t_j=0}^{t}\int_{t_1=t_j}^{t}\cdots\int_{t_{j-1}=t_j}^{t}\int_{t_{j+1}=t_j}^{t}\cdots\int_{t_N=t_j}^{t} P\{\bigcap_{i=1}^{N} E_{2,i}\} \prod_{i=1}^{N} P\{E_{1,i}\}P\{E_{3,i}\}\ dt_1\cdots dt_N.$$

Conditioning on the event $\mathbf{E_{2j}}$ we have that

$$P\{\bigcap_{i=1}^{N}(L_i \geq n_i)\bigcap(\tau_{j,n_j} = \min_i \tau_{i,n_i})\} =$$
$$\int_{t_j=0}^{t}\int_{t_1=t_j}^{t}\cdots\int_{t_N=t_j}^{t} P\{\bigcap_{i=1}^{N} E_{2,i} \mid E_{2,j}\}P\{E_{2,j}\} \prod_{i=1}^{N} P\{E_{1,i}\}P\{E_{3,i}\}\ dt_1\cdots dt_N.$$

Since the discipline is FIFO (Assumption A.2), for any arbitrary choice of time epochs t_i $i = 1, \dots, n$ such that $t_j = \min_i t_i$ we have that

$$P\{\bigcap_{i=1}^{N} E_{2,i} \mid E_{2,j}\} = 1,$$

i.e., if the customer that arrives first is still in the system at an observation epoch τ, all the customers that arrived after him are, also, still in the system at τ. Therefore,

$$P\{\bigcap_{i=1}^{N}(L_i \geq n_i)\bigcap(\tau_{j,n_j} = \min_i \tau_{i,n_i})\} = \int_{t_j=0}^{t} P\{E_{2,j}\}P\{E_{1,j}\}P\{E_{3,j}\} \prod_{\substack{i=1\\ i\neq j}}^{N} \int_{t_i=t_j}^{t} P\{E_{1,i}\}P\{E_{3,i}\}\ dt_i\ dt_j.$$

From the definitions of the events $\mathbf{E_{1,i}}$, $\mathbf{E_{2,i}}$ and $\mathbf{E_{3,i}}$ we have that

$$\int_{t_i=t_j}^{t} P\{E_{1,i}\}P\{E_{3,i}\}\ dt_i = \int_{t_j}^{t} h_i(t_i)\ P\{N^o_{a_i}(t_i,t) = n_i - 1\}\ dt_i\ , \qquad i \neq j,$$

$$P\{E_{2,j}\}P\{E_{1,j}\}P\{E_{3,j}\} = h_j(t_j)P\{S_j(t_j) > t - t_j\}\ P\{N^o_{a_j}(t_j, t) = n_j - 1\},$$

where in the second formula we use the fact that S_{n_j} conditioned on the arrival time of the n_jth customer does not depend on n_j. Hence,

$$P\{\bigcap_{i=1}^{N}(L_i \geq n_i)\} = \sum_{j=1}^{N} \int_0^t h_j(t_j)\ P\{S_j(t_j) > t - t_j\}P\{N_{a_j}(t_j, t) = n_j - 1\} \prod_{\substack{i=1\\ i\neq j}}^{N} H_i(t_j, t, n_i)dt_j, \quad (14)$$

where we define $H_i(t_j, t, n_i) \triangleq \int_{t_j}^{t} h_i(t_i)\ P\{N^o_{a_i}(t_i, t) = n_i - 1\}\ dt_i$.

In the general case where at time t there are *no* customers from class $k \in A \subset \{1, \cdots, N\}$ in the system, and there are $n_i \geq 1$ customers from class $i \notin A$ we can prove in a similar way

$$P\{\bigcap_{i\notin A}(L_i(t) \geq n_i)\} = \sum_{j\notin A} \int_0^t h_j(t_j)P\{S_j(t_j) > t - t_j\}P\{N^o_{a_j}(t_j, t) = n_j - 1\} \prod_{\substack{i\neq j\\ i\notin A}} H_i(t_j, t, n_i)dt_j. \quad (15)$$

We now compute $P\{\bigcap_{i=1}^{N}(L_i(t) = n_i)\}$ iteratively, using (14), (15) and the fact that for $n_i \geq 0$

$$P\{\bigcap_{k=1}^{i}(L_k(t) = n_k) \bigcap_{j=i+1}^{N} (L_j(t) \geq n_j)\} = P\{\bigcap_{k=1}^{i-1}(L_k(t) = n_k) \bigcap_{j=i}^{N}(L_j(t) \geq n_j)\}$$
$$-P\{\bigcap_{k=1}^{i-1}(L_k(t) = n_k) \bigcap(L_i(t) \geq n_i + 1) \bigcap_{j=i+1}^{N}(L_j(t) \geq n_j)\}.$$

Having calculated $P\{L_1(t) = n_1, \ldots, L_N(t) = n_N\}$, some tedious but straightforward manipulation yields (13). ■

In the case of a single class (13) yields:

$$G_L(z, t) = 1 - \int_0^t \frac{\partial}{\partial a} K_e(z, a, t)\ P\{S(a) > t - a\}\ da,$$

where $K_e(z, a, t) \triangleq 1 + (z - 1)\ \int_a^t h(u)\ K_o(z, u, t)\ du$.

Hence, substituting $\frac{\partial}{\partial a}K_e(z, a, t) = -h(a)\ (z - 1)K_o(z, a, t)$, we obtain (7). Moreover, the generating function of the total number of customers $L(t)$ in the system can be obtained if we set $z_1 = z_2 = \ldots = z_N$ in (13):

$$G_L(z, t) = 1 - \sum_{j=1}^{N} \int_0^t \frac{\partial}{\partial u} K_{e,j}(z, u, t) \prod_{i\neq j}^{N} K_{e,i}(z, u, t)\ P\{S_j(t - u) > u\}\ du. \quad (16)$$

Finally, notice that for renewal arrival processes $K_{o,i}(z_i, u, t) = K_{o,i}(z_i, t - u)$ and $K_{e,i}(z_i, u, t) = K_{e,i}(z_i, t - u)$, where $K_{o,i}(z_i, t - u) \triangleq E[z^{N^o_a(t-u)}]$ and $K_{e,i}(z_i, t - u) \triangleq E[z^{N^e_a(t-u)}]$ are the generating functions of the number of arrivals from an ordinary and an equilibrium renewal process, respectively.

3 Transient Little's law

One of the most celebrated results in queueing theory is that under natural and rather mild assumptions (see Heyman and Sobel [11]), the expected number of customers in the system $E[L]$ and the expected system time $E[S]$ in steady-state are linearly related via

$$E[L] = \lambda E[S],$$

where λ is the arrival rate.

Using the transient single class distributional laws of Theorem 1 we can obtain the following generalization of Little's law in the transient domain.

Corollary 1 *For a single class system that satisfies Assumptions A.1-A.3 and starts empty, we have that*

$$E[L(t)] = \int_0^t h(u)P\{S(u) > t-u\}du. \tag{17}$$

Proof: Since we consider a system that satisfies Assumptions A.1-A.3 the single class transient distributional law, (7), holds, i.e.,

$$G_L(z,t) = 1 + (z-1)\int_0^t h(u)\ P\{S(u) > t-u\}\ K_o(z,u,t)\ du,$$

where $K_o(z,u,t) \triangleq E[z^{N_a^o(u,t)}] = \sum_{n=0}^{\infty} z^n P\{N_a^o(u,t) = n\}$. We take derivatives $\frac{\partial}{\partial z}$ in the above equation, let $z = 1$, and prove the corollary. ■

The previous corollary raises the question whether (17) holds not only for overtake-free systems (Assumptions A.1-A.3) but more generally. The next theorem shows that this is indeed the case.

Theorem 6 *For a single class system that starts empty with k initial customers, $L(0) = k$, and initial work $V(0) = X_1 + \cdots X_k$, if we denote by $N_a(t)$ the number of arrivals in $(0,t]$, by $h(u)$ the probability of an arrival at time u, by $L(t)$ the number of customers in the system at time t and by $S(u)$ the time spent in the system for a customer that arrived at time u, we have that*

$$E[L(t)] = P\{V(0) \le t\}\int_0^t h(u)P\{S(u) > t-u\}du + P\{V(0) > t\}E[N_a(t)] \\ + \sum_{i=1}^{k} i\,[P\{V_i \le t\} - P\{V_{i-1} \le t\}], \tag{18}$$

where $V_i \triangleq X_1 + \cdots + X_{k-i}$.

Proof: To enhance our intuition let us first assume that the system starts empty and consider a particular realization of the system ω. We define $l(t;\omega)$ to be the number of customers in the system at time t for this particular realization and introduce the indicator function

$$f_t(u;\omega) = \begin{cases} 1 & \text{if we have an arrival at } u \text{ who is still in the system at } t \\ 0 & \text{otherwise.} \end{cases}$$

Then it is clear that

$$l(t;\omega) = \int_0^t f_t(u;\omega)\ du.$$

If we denote by $F_t(u)$ the stochastic process that corresponds to $f_t(u;\omega)$ we have that

$$E[L(t)] = E\left[\int_0^t F_t(u)\ du\right] = \int_0^t E[F_t(u)]\ du, \tag{19}$$

where the second equality follows from the bounded convergence theorem.

Moreover,

$$\begin{aligned}\int_0^t E[F_t(u)]\ du &= \int_0^t P\{\text{an arrival at } u \text{ who is still in the system at } t\}\ du \\ &= \int_0^t h(u)P\{S(u) > t-u\}\ du. \end{aligned}\tag{20}$$

Hence we proved (18) in the case where the system starts empty. In the general case where $L(0) = k$ w.p.1 and $V(0) = X_1 + \cdots + X_k$, let us again consider a particular realization of the system ω. We denote by $v(0;\omega)$ the initial work for this particular realization and with $m(t;\omega)$ the number of initial customers that are still present in the system at time t. We define $l(t;\omega)$ and $f_t(u;\omega)$ as before and we further define the indicator function

$$g(u;\omega) = \begin{cases} 1 & \text{if we have an arrival at } u \\ 0 & \text{otherwise.}\end{cases}$$

Then, for $t < v(0;\omega)$ the system is still working on the initial customers, so $l(t;\omega)$ is equal to $m(t;\omega)$ plus the number of customers that arrived to the system before t. On the other hand, if $t \geq v(0;\omega)$, no initial customer is present in the system and $l(t;\omega)$ is equal to the number of customers that arrived before t and are still in the system at t. In other words,

$$l(t;\omega) = \begin{cases} m(t;\omega) + \int_0^t g(u;\omega)\ du & \text{if } t < v(0;\omega) \\ \int_0^t f_t(u;\omega) du & \text{otherwise.}\end{cases}$$

If we denote by $G(u)$ and $M(t)$ the stochastic processes corresponding to $g(u;\omega)$ and $m(t;\omega)$ we have that

$$E[L(t)] = P\{V(0) \leq t\} E\left[\int_0^t F_t(u)\ du\right] + P\{V(0) < t\}\left(E[M(t)] + E\left[\int_0^t G(u)\ du\right]\right).$$

From the discussion in Theorem 4 we have that

$$E[M(t)] = \sum_{i=1}^{k} i\left[P\{V_i \leq t |\ V(0) > t\} - P\{V_{i-1} \leq t |\ V(0) > t\}\right].$$

Moreover we have from (19) and (20),

$$E\left[\int_0^t F_t(u)\ du\right] = \int_0^t E[F_t(u)]\ du = \int_0^t h(u)P\{S(u) > t-u\}\ du,$$

and similarly

$$E\left[\int_0^t G(u)\ du\right] = \int_0^t E[G(u)]\ du = \int_0^t h(u)\ du = E[N_a(t)].$$

Combining the last four relationships we prove (18). ■

Notice that unlike Little's law, $E[L(t)]$ depends on the entire distribution of $S(t)$, not just its expectation, and on the initial conditions.

If we further assume that the arrival process is renewal and that the initial interarrival time is distributed as the forward recurrence time of the interarrival distribution, i.e., $h(t) = \lambda$ we obtain in the case where the system starts empty

$$E[L(t)] = \lambda \int_0^t P\{S(u) > t-u\}\ du.$$

We will use this transient version of Little's law to obtain the mean number of customers in a $GI(t)/G(t)/\infty$ system in Section 5.1.

4 Asymptotic forms of $\mathbf{K_o(z, t)}$ and $\mathbf{K_e(z, t)}$

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 in both the transient and steady-state regime 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 of the number of customers in the system and the distribution of the system time. For example, for the single class system in steady state we have that:

$$G_L(z) = \int_0^\infty K_e(z,t)\ dF_S(t),$$

and for multiclass systems in the transient regime

$$G_{L_1,\dots,L_N}(\vec{z},t) = 1 - \sum_{j=1}^N \int_0^t \frac{\partial}{\partial a} K_{e,j}(z_j,a,t) \prod_{i\neq j} K_{e,i}(z_i,a,t)\ P\{S_j(a) > t-a\}\ da,$$

where the kernels $K_e(z,t)$ and $K_{e,i}(z_i,a,t)$ were defined in Theorem 1 and Theorem 8, respectively.

Since we only consider renewal processes where $K_{o,i}(z,a,t)$ depends only on z and the difference $t-a$, so in accordance with Section 2 will use the notation $K_{o,i}(z,t-a)$. Similarly, for $K_{e,i}(z,a,t)$ we will use the notation $K_{e,i}(z,t-a)$

It is important to notice that in the special case of a Poisson arrival process the kernel $K_e(z,t) =$

$e^{-\lambda t(1-z)}$ and the distributional laws are linear relationships between transforms, for example,

$$G_L(z) = \phi_S(\lambda(1-z)),$$

where $\phi_S(s)$ is the Laplace transform of the system time distribution.

For mixed generalized Erlang arrivals $K_e(z,t)$ is given explicitly in Bertsimas and Nakazato [6]. For arbitrary renewal arrivals, however, $K_e(z,t)$ is not known in closed form. In order to exploit the distributional laws we try to understand in this section the asymptotic behavior of $K_e(z,t)$ and $K_o(z,t)$ as $t \to \infty$ and $z \to 1$, for a renewal process with arrival rate λ and squared coefficient of variation c_a^2.

We use the notation that $h(x) \sim r(x)$ as $x \to a$ means that $\lim_{x\to a} \frac{h(x)}{r(x)} = 1$ and following the asymptotic approach introduced in Smith [24] (see also Cox [8], ch. 4-6) we obtain (see Mourtzinou [19]):

Proposition 1 *Asymptotically, as $t \to \infty$ and $z \to 1$ the kernels $K_e(z,t)$ and $K_o(z,t)$ behave as follows:*

$$K_e(z,t) \sim e^{-tf(z)}, \tag{21}$$

$$K_o(z,t) \sim [1 - \frac{1}{2}(1-z)(c_a^2 - 1) + O((1-z)^2)]e^{-tf(z)}, \tag{22}$$

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

Given that we will extensively use the asymptotic forms in later chapters we will evaluate numerically the accuracy of our asymptotic expansion as a function of time for different values of z and different arrivals processes. In the following figures the solid line corresponds to the exact value 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 [12] and in Abate and Whitt [1] which we programmed in Matlab and we got exactly the same results. The results are shown in Figures 4-7.

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 $t > 20$. It is also interesting to notice that our asymptotic expansion is more accurate for values of c_a^2 close to 1 and indeed is exact for Poisson arrivals $c_a^2 = 1$. In other words it performs better for Erlang 2 than Erlang 16 arrivals and it also performs better for hyperexponential arrivals with $c_a^2 = 1.5$ than for arrivals with $c_a^2 = 2$.

It is important to notice that according to the line of arguments in Mourtzinou [19], $-f(z)$ is the root of $1 - z\alpha(s) = 0$ for small values of s and for values of z close to 1, in other words

$$1 - z\alpha(-f(z)) = 0. \tag{23}$$

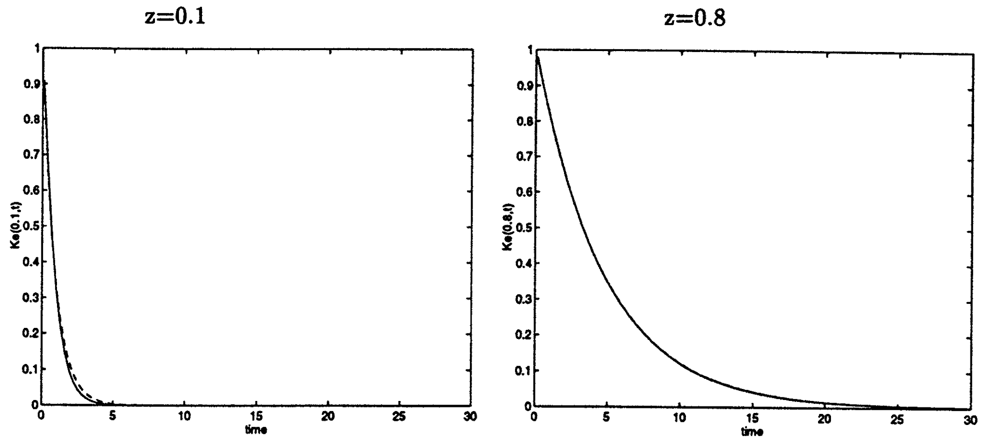


Figure 3: The function $K_e(z,t)$ for Erlang 2 arrivals.

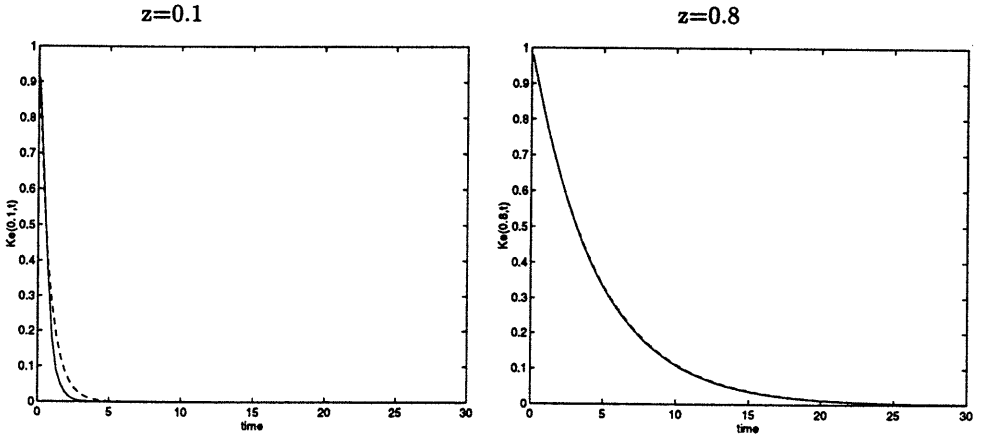


Figure 4: The function $K_e(z,t)$ for Erlang 16 arrivals.

5 Transient performance analysis

In this section we apply the transient distributional laws to derive the transient performance analysis of several systems: infinite server systems with a single non-homogeneous Poisson arrival process and general time-dependent services, and multiclass single server systems with general time-dependent arrivals and services.

5.1 The M(t)/G(t)/∞ queueing system

In this section we investigate the transient behavior of the $M(t)/G(t)/\infty$ queueing system; that often arises in air-traffic control and wireless communications systems where we use the nonhomogeneity to capture the important time-of-day effect and we ignore the resource constraints (limited number of lines) by assuming an infinite number of servers (see Massey and Whitt [18]). Since

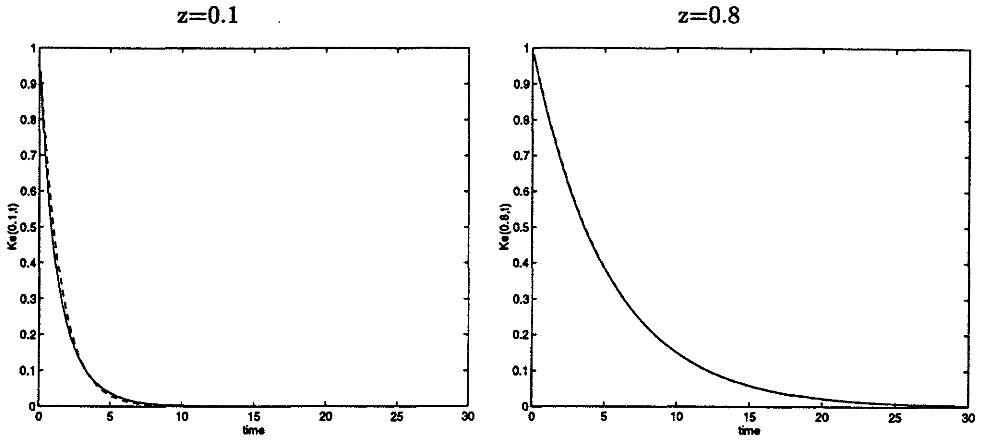


Figure 5: The function $K_e(z,t)$ for Hyperexponential arrivals with $c_a^2 = 1.5$.

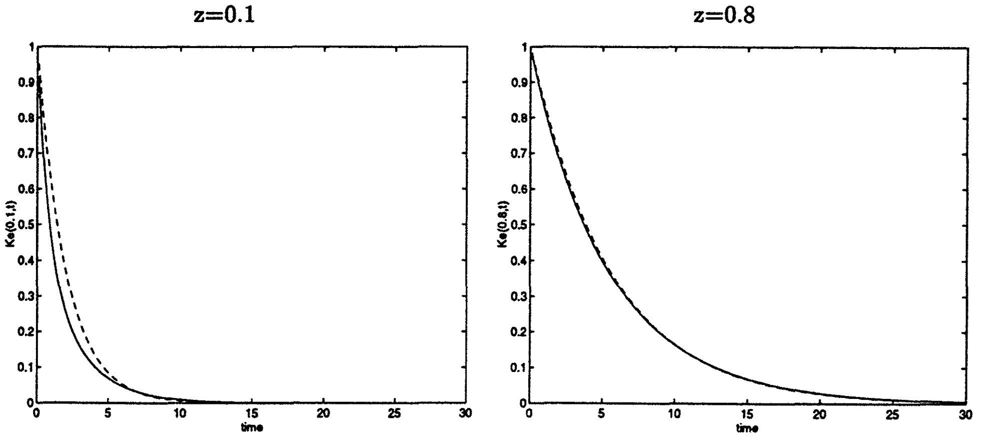


Figure 6: The function $K_e(z,t)$ for Hyperexponential arrivals with $c_a^2 = 2$.

this system is *not* overtake-free, the distributional laws presented in the previous sections do not directly apply. However, we can still use them as the building blocks of our analysis since they do apply in the special case of the $M(t)/D/\infty$ system, when all customers have the same deterministic service requirement and hence they leave the system in the order of their arrival.

Hence we start by proving the following proposition.

Proposition 2 *For a $M(t)/D/\infty$ queueing system with arrival rate $\lambda(t)$ and service time x that starts empty with no initial work, if we define $\Lambda(t) \triangleq \int_0^t \lambda(\tau)\, d\tau$, we have that*

$$G_L(z,t) = \begin{cases} e^{-(\Lambda(t)-\Lambda(t-x))(1-z)}, & \textit{if} \quad t \geq x \\ e^{-\Lambda(t)(1-z)} & \textit{otherwise.} \end{cases} \tag{24}$$

Proof: From Theorem 3 we have that

$$G_L(z,t) = 1 + (z-1)\int_0^t \lambda(u)\ P\{S(u) > t-u\}\ K_o(z,u,t)\ du, \tag{25}$$

where $S(u)$ denotes the system time of a customer that arrived at time u. Since there are infinitely many servers and no initial work there is *no* waiting time, so that $S(u) = x$. Moreover,

$$\begin{array}{llllll} \bullet\text{if} & t < x & \text{then} & P\{x > t-u\} = 1 & \text{for} & u \in [0,t) \\ \bullet\text{if} & t \geq x & \text{then} & P\{x > t-u\} = 1 & \text{for} & u \in [t-x, x) \\ & & & P\{x > t-u\} = 0 & \text{for} & u \in [0, t-x) \end{array} \tag{26}$$

On the other hand, since the arrival process is a non-homogeneous Poisson of rate $\lambda(t)$ we have that $K_o(z,u,t) = e^{-(\Lambda(t)-\Lambda(u))(1-z)}$. Substituting $K_o(z,u,t)$ and (25), (26) we obtain (24). ■

We next consider the $M(t)/G(t)/\infty$ queueing system and denote by $X(t)$ the service time of a customer entering service at time t and we prove the following theorem.

Theorem 7 *For a $M(t)/G(t)/\infty$ queueing system that starts empty, we have that the number of customers in the system at time t, $L(t)$, is a non-homogeneous Poisson process with rate $\int_0^t \lambda(\tau)\ P\{X(\tau) > t-\tau\}\ d\tau$, i.e.,*

$$G_L(z,t) = e^{-(1-z)\int_0^t \lambda(\tau)\ P\{X(\tau)>t-\tau\}\ d\tau}. \tag{27}$$

Proof: We can *decompose* this system into a number of $M(t)/D/\infty$ systems. Suppose that instead of having a general time-dependent service distribution the service time has $P\{X(t) = x_j\} = p_j(t)$ for $j = 1, 2, \ldots, k$. The customers with service times x_j can be treated as a separate class C_j of customers with arrival process being a non-homogeneous Poisson process of rate $\lambda(t)\ p_j(t)$. Therefore, if we denote by $\Lambda_j(t) \triangleq \int_0^t \lambda(\tau)\ p_j(\tau)\ d\tau$, we have that

$$G_{L_j}(z,t) = \begin{cases} e^{-(\Lambda_j(t)-\Lambda_j(t-x_j))(1-z)}, & \text{if} \quad t \geq x_j \\ e^{-\Lambda_j(t)(1-z)} & \text{otherwise.} \end{cases}$$

Moreover as discussed in Ross [23], p. 224, these processes are mutually independent and thus

$$G_L(z,t) = \prod_{j=1}^k G_{L_j}(z,t) = e^{-(1-z)\sum_{j=1}^k \Lambda_j(t)}\ e^{(1-z)\sum_{j:x_j\leq t}\Lambda_j(t-x_j)}. \tag{28}$$

We will evaluate now the exponents

$$\sum_{j=1}^k \Lambda_j(t) = \sum_{j=1}^k \int_0^t \lambda(\tau)p_j(\tau)\ d\tau = \int_0^t \lambda(\tau)\sum_{j=1}^k p_j(\tau)\ d\tau = \int_0^t \lambda(\tau)\ d\tau, \tag{29}$$

$$\begin{aligned}\sum_{j:x_j\le t}\Lambda_j(t-x_j) &= \sum_{j:x_j\le t}\int_0^{t-x_j}\lambda(\tau)p_j(\tau)\,d\tau = \sum_{j=1}^{k}\int_0^t \lambda(\tau)p_j(\tau)\;P\{\tau\le t-x_j\}\,d\tau \\ &= \int_0^t \lambda(\tau)\sum_{j:x_j\le t} p_j(\tau)\;P\{\tau\le t-x_j\}\,d\tau = \int_0^t \lambda(\tau)\;P\{X(\tau)\le t-\tau\}\,d\tau. \quad (30)\end{aligned}$$

Combining (28)-(30) we get (27) for this case. Since any general distribution is the limit of a sequence of mixtures of deterministic distributions, (27) holds in general. Moreover, the generating function $G_L(z,t)$ in (27) corresponds to a non-homogeneous Poisson process of rate $\int_0^t \lambda(\tau)\; P\{X(\tau) > t-\tau\}\; d\tau$. ∎

One can actually obtain the expected number of customers in the $M(t)/G(t)/\infty$ system,

$$E[L(t)] = \int_0^t \lambda(u)P\{X(u) > t-u\}du, \quad (31)$$

directly from the transient form of Little's law, (18), by substituting $h(u) = \lambda(u)$ and $S(u) = X(u)$ since there is no wait. Furthermore, (31) is independent of the Poisson assumption and gives the expected number of customers in any $GI(t)/G(t)/\infty$ system.

In the special case of the $M(t)/GI/\infty$ system, Theorem 7 can be traced to Palm [21], Bartlett [2], Doob [9], Khintchine [16] and Prékopa [22], all before 1958. For a recent reference on Theorem 7 and its extension to networks of infinite server queues with non-stationary Poisson input see Massey and Whitt [18].

5.2 The GI(t)/GI(t)/1 queueing system under FIFO

In this section we consider a single server system with general mutually independent non-stationary arrival and service time distributions, namely, the $GI(t)/GI(t)/1$ queue under FIFO. We denote by $Q(t)$ the number of customers waiting *in the queue* at time t and by $L(t)$ the number of customers *in the system*, i.e., the queue plus the server, at time t. Similarly, we denote by $W(t)$ the time that a customer who arrived at time t spends *in the queue* and by $S(t)$ the time that a customer who arrived at time t spends *in the system*, i.e., the queue and the server. We assume, without loss of generality, that the server of the system is working with unit speed as long as there is work in the system and we denote by $X(t)$ the service requirement of a customer that *enters the server at time t*. Finally, we denote by $G_L(z,t)$ (resp. $G_Q(z,t)$) the generating function of $L(t)$ (resp. $Q(t)$).

For this system we first prove another distributional law that relates $L(t)$ and $Q(t)$ and which in contrast with the laws presented in Section 2, does not hold for all overtake-free systems but it requires the existence of a single server, and it is, therefore, specialized to the case of a $GI(t)/GI(t)/1$ system under FIFO.

Proposition 3 *For a $GI(t)/GI(t)/1$ queueing system that starts with $L(0) = k$ w.p.1, set-up work $\hat{V}(0)$, and total initial work $V(0) = \hat{V}(0) + X_1 + \cdots + X_k$ and satisfies Assumptions A.1-A.3, the transient quantities $L(t)$ and $Q(t)$ are related as follows:*

$$G_L(z,t) = (1-z)idle(t) + (1-z)z^k P\{\hat{V}(0) > t\}K(z,t) + zG_Q(z,t), \quad (32)$$

where idle(t) is the emptiness function, i.e., $idle(t) \triangleq P\{\textit{the system is empty at time } t\}$ *and* $K(z,t) \triangleq E[z^{N_a(t)}] = \sum_{n=0}^{\infty} z^n P\{N_a(t) = n\}$.

Proof: Notice that at time t the system can be in either of the following states:

1. It is empty (with probability $idle(t)$).
2. It is still working on the set-up work, $\hat{V}(0)$ (with probability $P\{\hat{V}(0) > t\}$, as the server has unit speed).
3. It is busy servicing customers (with probability $1 - P\{\hat{V}(0) > t\} - idle(t)$).

In the first case, the number of customers in the queue, $Q(t)$, and in the system, $L(t)$, satisfy $Q(t) = L(t) = 0$. Similarly, in the second case, $Q(t) = L(t) = k + N_a(t)$, as in this case, all the initial customers plus the customers that arrived to the system up to time t are still waiting for the server to finish the set-up work, $\hat{V}(0)$. However, in the third case, $L(t) = Q(t) + 1$, as one of the customers is receiving service at time t. We can, therefore, decompose the generating functions of $Q(t)$ and $L(t)$, $G_Q(z,t) \triangleq E[z^{Q(t)}]$ and $G_L(z,t) \triangleq E[z^{L(t)}]$ as follows:

$$\begin{aligned} G_Q(z,t) &= idle(t) + z^k K(z,t) P\{\hat{V}(0) > t\} + (1 - idle(t) - P\{\hat{V}(0) > t\})\ G_{Q_B}(z,t), \\ G_L(z,t) &= idle(t) + z^k K(z,t) P\{\hat{V}(0) > t\} + z(1 - idle(t) - P\{\hat{V}(0) > t\})\ G_{Q_B}(z,t), \end{aligned}$$

where $G_{Q_B}(z,t) \triangleq E[z^{Q(t)} |$ the server is servicing customers]. Combining the last two relations we obtain (32). ■

The above proposition together with the transient distributional laws of Section 2 leads to a complete description of the $GI(t)/GI(t)/1$ system as a function of the emptiness function as the following theorem demonstrates.

Theorem 8 *For a GI(t)/GI(t)/1 system under FIFO with initial work* $V(0)$ *the probability distribution function of the waiting time of a customer who arrived to the system at time* t_o, $F_{W(t_o)}(x) \triangleq P\{W(t_o) \le x\}$, *satisfies the following integral equation*

$$\int_0^t h(u)\ K_o(z,u,t) \left[\int_0^{\infty} dF_{W(u)}(a)\ P\{X(u+a) > t-u-a\} - zP\{W(u) > t-u\} \right] du \\ = 1 - idle(t) - P\{V(0) > t\} K(z,t), \quad (33)$$

where $idle(t) \triangleq P\{\textit{the server is idle at time } t\}$, $dF_{W(u)}(\cdot)$ *is the pdf of* $W(u)$, $K_o(z,u,t) \triangleq E[z^{N_a^o(u,t)}] = \sum_{n=0}^{\infty} z^n P\{N_a^o(u,t) = n\}$ *and* $K(z,t) \triangleq E[z^{N_a(t)}] = \sum_{n=0}^{\infty} z^n P\{N_a(t) = n\}$.

Proof: We start by noticing that $W(t)$ given the initial work in the system , $V(0)$, is independent of the actual number of the initial customers; hence we assume wlog that the system starts with $L(0) = 0$ and $\hat{V}(0) = V(0)$. In this case, Theorem 4 holds for the pair ($L(t)$, $S(t)$), if we regard "the system" as the queue and the server, as well as the pair $(Q(t),\, W(t))$, if we regard "the system" as just the queue. Therefore,

$$G_Q(z,t) = 1 + (z-1) \int_0^t h(u)\ P\{W(u) > t-u\}\ K_o(z,u,t)\ du\ , \tag{34}$$

$$G_L(z,t) = 1 + (z-1) \int_0^t h(u)\ P\{S(u) > t-u\}\ K_o(z,u,t)\ du\ . \tag{35}$$

Moreover, from the definitions of $S(t)$, $W(t)$ and $X(t)$ we have that

$$P\{S(t) > x\} = \int_{a=0}^{\infty} P\{a \leq W(t) \leq a + da\}\ P\{X(t+a) > x - a\}, \tag{36}$$

so that from (35) we get

$$G_L(z,t) = 1 + (z-1) \int_0^t h(u) \int_0^{\infty} dF_{W(u)}(a)\ P\{X(u+a) > t-u-a\}\ K_o(z,u,t)\ du,$$

where $dF_{W(u)}$ is the pdf of $W(u)$ and $K_o(z,u,t) \triangleq \sum_{n=0}^{\infty} z^n P\{N_a^o(u,t) = n\}$.

Combining the last equation with (32) for $k = 0$ and $V(0) = \hat{V}(0)$, since we assumed that no customer is present, we complete the proof. ∎

By solving Equation (33) and then using (36) we also obtain the pdf of $S(t)$ as a function of $idle(t)$. Moreover, using the distributional laws of Theorem 3 we obtain the description of the $GI(t)/GI(t)/1$ system with no initial customers, again as a function of $idle(t)$. In the case where $L(0) = k$ we can use the distributional laws of Theorem 4. However, solving the equation of Theorem 8 for the general $GI(t)/GI(t)/1$ case is quite complicated and therefore in the sequel we consider two special cases: the $GI/GI/1$ queue and the $M(t)/GI(t)/1$ queue. For both cases we solve for the fundamental quantities of the system as function of $idle(t)$ and then we calculate $idle(t)$ from analytic properties of Laplace transforms.

5.2.1 Transient analysis of GI/GI/1 queueing system under FIFO

In this section we focus our attention to an important class of systems where customers arrive according to a single *equilibrium* renewal arrival process and have general (though time independent) service requirements.

We use the notation of the Section 5.2. Moreover, since the arrival process is renewal, the number of arrivals, $N_a^o(u,t)$ only depends on the difference $t-u$. Therefore, in this section we write $K_o(z,u,t)$ as $K_o(z,t-u)$. Moreover, as the arrival process is an equilibrium process $N_a(t) = N_a^e(t)$, $h(u) = \lambda$ for all $u \geq 0$, where λ is the arrival rate, and also $K(z,t) \triangleq E[z^{N_a(t)}] = K_e(z,t)$.

Since the $GI/GI/1$ queueing system is just a special case of the $GI(t)/G(t)/1$ system, Theorem

8 still holds and the integral equation takes the following form

$$\lambda \int_0^t K_o(z, t-u)(P\{W(u) + X > t-u\} - zP\{W(u) > t-u\})du$$
$$= 1 - idle(t) - P\{V(0) > t\}K_e(z,t). \quad (37)$$

The integral equation (37) is still difficult to solve analytically for general arrival processes. One possibility would be to solve it numerically, and then use the distributional laws to find the complete description of the $GI/GI/1$ system numerically. In the next section we follow another approach and we examine the behavior of the $GI/GI/1$ system for large times $t >> t_o$ and under the assumption that the traffic intensity $\rho \to 1$.

In the rest of this section we focus our attention to another pair of performance measures, namely, the expected number of customers in the system at time t, $E[L(t)]$, and the expected number of customers in the queue at time t, $E[Q(t)]$. We define $\mathcal{L}\{E[L(t)]\}$ and $\mathcal{L}\{E[Q(t)]\}$ to be the Laplace transform of $E[L(t)]$ and $E[Q(t)]$, respectively, i.e.,

$$\mathcal{L}\{E[L(t)]\} \triangleq \int_0^\infty e^{-st}E[L(t)]\ dt \qquad \text{and} \qquad \mathcal{L}\{E[Q(t)]\} \triangleq \int_0^\infty e^{-st}E[Q(t)]\ dt,$$

and we also define by $\phi_W(w,t)$ the Laplace transform of $W(t)$ and by $\Phi_W(w,s)$ the double Laplace transform of $W(\cdot)$, i.e.,

$$\phi_W(w,t) \triangleq \int_0^\infty e^{-wx}\ dF_{W(t)}(x) \qquad \text{and} \qquad \Phi_W(w,s) \triangleq \int_0^\infty e^{-st}\phi_W(w,t)\ dt.$$

Similarly, $S(t)$ is the system time of a customer that arrived at t and has Laplace transform $\phi_S(w,t)$ and double Laplace transform $\Phi_S(w,s)$. As an illustration of the importance of the transient version of Little's law we obtain the following theorem.

Theorem 9 *For a GI/GI/1 system under a work conserving policy, that starts empty with initial work $V(0)$, the Laplace transform of the expected number of customers in the system and in the queue are given as follows:*

$$\mathcal{L}\{E[Q(t)]\} = \frac{\lambda}{s^2} - \frac{s\ idle(s) - \phi_{V(0)}(s)}{s(\phi_X(s) - 1)}, \quad (38)$$

$$\mathcal{L}\{E[L(t)]\} = \frac{\lambda}{s^2} - \frac{s\ idle(s) - \phi_{V(0)}(s)}{s(\phi_X(s) - 1)}\ \phi_X(s), \quad (39)$$

where $idle(s) \triangleq \int_0^\infty e^{-st} idle(t)dt$ is the Laplace transform of the emptiness function and $\phi_{V(0)}(s) \triangleq \int_0^\infty e^{-st}dP\{V(0) \leq t\}$ is the Laplace transform of $V(0)$.

Proof: From the transient form of Little's law we have that

$$E[L(t)] = \lambda \int_0^t P\{S(u) > t-u\}\ du \quad \text{and} \quad E[Q(t)] = \lambda \int_0^t P\{W(u) > t-u\}\ du.$$

Taking Laplace transforms in the first of the previous two equations we obtain

$$\mathcal{L}\{E[L(t)]\} = \lambda \int_0^\infty e^{-st} \int_0^t P\{S(u) > t-u\}\, du\, dt = \lambda \int_0^\infty e^{-sa} \int_0^\infty e^{-su} P\{S(u) > a\}\, du\, da,$$

where we set $a \triangleq t-u$ and we changed integration variables. Equivalently, from the definition of the double Laplace transforms:

$$\mathcal{L}\{E[L(t)]\} = \frac{\lambda}{s^2} - \frac{\lambda}{s}\Phi_S(s,s), \tag{40}$$

$$\mathcal{L}\{E[Q(t)]\} = \frac{\lambda}{s^2} - \frac{\lambda}{s}\Phi_W(s,s). \tag{41}$$

Moreover, we know that $S(u) = W(u) + X$, so taking Laplace transforms

$$\Phi_S(s,s) = \phi_X(s)\ \Phi_W(s,s). \tag{42}$$

On the other hand we have from Proposition 3 that for a system that starts empty with initial work $V(0)$,

$$G_L(z,t) = (1-z)\ idle(t) + (1-z)P\{V(0) > t\}K_e(z,t) + zG_Q(z,t),$$

where $K_e(z,t) \triangleq E[z^{N_a^e(t)}] = \sum_{n=0}^\infty z^n P\{N_a^e(t) = n\}$. By differentiation we get that

$$E[L(t)] = -idle(t) - P\{V(0) > t\} + 1 + E[Q(t)],$$

or equivalently in the Laplace domain

$$\mathcal{L}\{E[L(t)]\} = -idle(s) + \frac{1}{s}\phi_{V(0)}(s) + \mathcal{L}\{E[Q(t)]\}. \tag{43}$$

Solving the linear system of equations (40)-(43), we conclude the proof. ■

It is important to notice that since the transient form of Little's law holds for any work conserving policy, (38) and (39) hold for any work conserving policy. However, the form of the emptiness function, which is not in general known and can *not* be obtained from the analytic properties of $\mathcal{L}\{E[L(t)]\}$ and $\mathcal{L}\{E[Q(t)]\}$, changes with the policy and so do $E[Q(t)]$ and $E[L(t)]$. It is, however, interesting to observe that the transient form of Little's law leads to a solution for the expected performance measures up to a function. Moreover, for FIFO policy we will actually use the asymptotic method to obtain a closed form expression for $idle(s)$, from the analytic properties of $\Phi_W(w,s)$, in the next section.

Finally notice that we can obtain the steady-state queue length, $E[Q]$, from the properties of the Laplace transforms as follows

$$E[Q] = \lim_{s\to 0} s\ \mathcal{L}\{E[Q(t)]\} = \lim_{s\to 0}\left[\frac{\lambda}{s} - \frac{s\ idle(s) - \phi_{V(0)}(s)}{\phi_X(s) - \lambda}\right].$$

In the Section 5.3.2 we will show using the asymptotic form of *idle*(s) that under FIFO

$$E[Q] \sim \frac{\rho(c_a^2 - 1) + \rho^2(c_x^2 + 1)}{2(1-\rho)},$$

where c_a^2 and c_x^2 are the squared coefficients of variation for the arrival and service process respectively. This is exactly the formula we obtained in Bertsimas and Mourtzinou [5].

The asymptotic heavy traffic analysis of the GI/GI/1 queue under FIFO

We, now, analyze the asymptotic heavy traffic transient behavior of the $GI/GI/1$ queueing system, where we define *asymptotic heavy traffic behavior* to mean the behavior as the traffic intensity $\rho \to 1$ and the observation time t is large, i.e., as $t \to \infty$. As we will see in the proof of the next theorem, we can equivalently define in the transform domain, where we are dealing with

$$G_Q(z,s) \triangleq \int_0^\infty e^{-st} E[z^{Q(t)}]dt \quad \text{and} \quad \Phi_W(w,s) \triangleq \int_0^\infty e^{-st} \int_0^\infty e^{-wx} dF_{W(t)}(x)dt,$$

the *asymptotic heavy traffic behavior* to mean the behavior for z relatively large, i.e., $z \to 1$, and s, w relatively small, i.e., $s, w \to 0$.

In particular, in the rest of this section we first obtain asymptotic expressions of the distributional laws in the transform domain under heavy traffic conditions. Using these expressions we obtain an asymptotic closed form expression of the double Laplace transform of the waiting time under heavy traffic conditions as a function of the Laplace transform of the emptiness function, *idle*(s). Then, we also obtain an asymptotic closed form expression for *idle*(s) under heavy traffic conditions, and hence we complete our asymptotic heavy traffic analysis of the $GI/GI/1$ queueing system.

To start, let us denote by λ the arrival rate and by c_a^2 the squared coefficient of variation of the arrival process (recall that the arrival process is an equilibrium renewal process). The main theorem of this section is the following.

Theorem 10 *In a $GI/GI/1$ queueing system under FIFO that starts empty with initial work $V(0)$, the distributional laws take the following form, asymptotically in heavy traffic:*

$$G_Q(z,s) \sim \frac{1}{s+f(z)} \; [1 + f(z) \; \Phi_W(s + f(z), s)], \tag{44}$$

$$G_L(z,s) \sim \frac{1}{s+f(z)} \; [1 + f(z) \; \Phi_S(s + f(z), s)], \tag{45}$$

with $f(z) = \lambda(1-z) - \frac{1}{2}\lambda(1-z)^2(c_a^2 - 1)$. Moreover, asymptotically under heavy traffic conditions

$$\Phi_W(w,s) \sim \frac{w \; idle(s) - \phi_{V(0)}(w)}{1 - \alpha(s-w) \; \phi_X(w)} \; \frac{1 - \alpha(s-w)}{(w-s)}, \tag{46}$$

where $\alpha(s)$ is the Laplace transform of the interarrival times, $\phi_{V(0)}(s)$ is the Laplace transform of the initial work and idle(s) is calculated from the analytic properties of $\Phi_W(w,s)$ in Proposition 4.

Proof: To justify (44) we can argue as follows: By taking the Laplace transform of the transient distributional law applied to the pair $(Q(t), W(t))$ we obtain

$$G_Q(z,s) = \frac{1}{s} + \lambda(z-1)\int_{t=0}^{\infty} e^{-st}\left[\int_{u=0}^{t} P\{W(u) > t-u\}K_o(z,t-u)\,du\right] dt. \tag{47}$$

We initially defined the *asymptotic heavy traffic behavior* of the system to mean the behavior as the traffic intensity $\rho \to 1$ and the observation time t is large, i.e., as $t \to \infty$. We know from the theory of Laplace transforms (Tauberian theorems, see Cox [8]) that the behavior of $G_Q(z,t)$ as $t \to \infty$ is associated with the behavior of $G_Q(z,s)$ as $s \to 0$. Moreover, as $\rho \to 1$ and $t \to \infty$ we have that $Q(t) \to \infty$. From the definition of $G_Q(z,t) \triangleq E[e^{-Q(t)\log(z)}]$ we observe that the behavior of $Q(t)$ when $Q(t) \to \infty$ is associated with the behavior of $G_Q(z,t)$ as $z \to 1$. Hence, the *asymptotic heavy traffic behavior* of $Q(t)$ is associated with the behavior of $G_Q(z,s)$ for small values of s and z close to 1.

Hence, we have to prove that for small values of s and large values of z the RHS of (47) yields the RHS of (44). Notice, now, that in (47) the second term is the Laplace transform of the function $\beta(t) \triangleq \int_{u=0}^{t} P\{W(u) > t-u\}K_o(z,t-u)\,du$. Therefore, its behavior for s relatively small is related to the behavior of $\beta(t)$ for t relatively large. Since we are also interested in z close to 1, we can substitute the asymptotic form of the kernel $K_o(z,t)$ from (22)

$$K_o(z,a) \sim \frac{f(z)}{\lambda(1-z)}\, e^{-f(z)a} \qquad \text{as} \quad a \to \infty \quad \text{and} \quad z \to 1.$$

Setting $a \triangleq t-u$ and changing the integration variables, we obtain

$$G_Q(z,s) = \frac{1}{s} - f(z)\int_{u=0}^{\infty} e^{-su}\left[\int_{a=0}^{\infty} e^{-sa}\, P\{W(u) > a\}\, e^{-f(z)a}\, da\right] du.$$

In other words,

$$G_Q(z,s) \sim \frac{1}{s} - f(z)\int_{u=0}^{\infty} e^{-su}\left[\frac{1-\phi_W(s+f(z),u)}{s+f(z)}\right] du, \quad \text{for small } s \quad \text{and } z \text{ close to 1.}$$

Equivalently,

$$G_Q(z,s) \sim \frac{1}{s+f(z)}\;[1 + f(z)\;\Phi_W(s+f(z),s)]\,, \quad \text{for small } s \quad \text{and } z \text{ close to 1.}$$

Similarly, we can prove (45).

Next, we take double Laplace transform of the relation $S(t) = W(t) + X$ to obtain

$$\Phi_S(s+f(z),s) = \Phi_S(s+f(z),s)\;\phi_X(s+f(z)). \tag{48}$$

Finally, we use (32) in the case where $L(0) = 0$ and the initial work is $V(0)$, i.e,

$$G_L(z,t) = (1-z)\; idle(t) + (1-z)P\{V(0) > t\}K_e(z,t) + zG_Q(z,t)$$

and we take Laplace transform with respect to t to obtain

$$G_L(z,s) = (1-z)\ idle(s) + \frac{(1-z)}{s+f(z)}\left[1-\phi_{V(0)}(s+f(z))\right] + zG_Q(z,s). \tag{49}$$

Combining (44),(45) with (48) and (49) we obtain

$$\Phi_W(s+f(z),s) = \frac{(s+f(z))\ idle(s) - \phi_{V(0)}(s+f(z))}{\phi_X(s+f(z))-z}\ \frac{1-z}{f(z)}. \tag{50}$$

Recall that $-f(z)$ is one of the roots of $1-z\alpha(s)=0$ where $\alpha(s)$ is the Laplace transform of the interarrival times (in particular it is a root of $1-z\alpha(s)=0$ for s small and z close to 1, see Section 2.4.2, equation (23)). Multiplying and dividing with $\alpha(-f(z))$ in (50) we obtain (46) where both s and w are small. The unknown function $idle(s)$ may be determined by insisting that the transform $\Phi_W(w,s)$ is analytic in the region $\Re(s)>0$ and $\Re(w)>0$. This implies that the zeroes of the numerator and denominator must coincide in this region (see Proposition 6.3). The same approach is used by Kleinrock [17] p. 229 to obtain $idle(s)$ for the $M/G/1$ queue. ■

Using (44) we can also find the heavy traffic form of $G_Q(z,s)$ if the system starts empty, once again for s relatively small and z close to 1.

It is important to note, that if the renewal process is Poisson, the asymptotic expressions of this theorem are *exact* with $f(z)=\lambda(1-z)$. Therefore, if we consider a system with Poisson arrivals, the asymptotic relations of Theorem 10 are exact under any traffic conditions and for any s. In particular, (46) is exact and yields:

$$\Phi_{W_{M/G/1}}(w,s) = \frac{\phi_{V(0)}(w) - w\ idle(s)}{\lambda+s-w-\lambda\phi_X(w)},$$

which is the exact transient solution for a $M/G/1$ queue (see Kleinrock [17]).

We can obtain the Laplace transform of the steady state waiting time, denoted by $\Phi_W(w)$, if we observe that $\Phi_W(w) = \lim_{s\to 0} s\ \Phi_W(w,s)$. Indeed,

$$\Phi_W(w) = \lim_{s\to 0} s\ \Phi_W(w,s) \sim \frac{1-\alpha(-w)}{1-\alpha(-w)\ \phi_X(w)}\ \lim_{s\to 0} s\ idle(s).$$

Moreover, we know (either from the properties of $\Phi_W(w)$ or from the physical meaning of $idle(s)$; see Section 2) that $\lim_{s\to 0} s\ idle(s) = 1-\rho$, where ρ is the traffic intensity of the system, and therefore,

$$\Phi_W(w) \sim \frac{(1-\rho)(1-\alpha(-w))}{1-\alpha(-w)\ \phi_X(w)}.$$

This is exactly the result obtained in Bertsimas and Mourtzinou [4].

We can also obtain an asymptotic closed form expression for $idle(s)$ as follows.

Proposition 4 *In a $GI/GI/1$ queue with FIFO service policy and initial work $V(0)$, asymptotically* *in heavy traffic the Laplace transform of the emptiness function for $\rho < 1$ is given as*

$$idle(s) \sim \frac{\phi_{V(0)}(w_2)}{w_2} \quad with \quad w_2 = \frac{-p_1(s) - \sqrt{(p_1(s))^2 - 4p_0(s)\, p_2(s)}}{2p_2(s)}, \tag{51}$$

where

$$p_0(s) \triangleq \frac{s}{\lambda} - \frac{(c_a^2+1)s^2}{2\lambda^2},$$

$$p_1(s) \triangleq \left(1 - \frac{s}{\lambda} + \frac{(c_a^2+1)s^2}{2\lambda^2}\right) E[X] - \frac{1}{\lambda} + \frac{(c_a^2+1)s}{\lambda^2},$$

$$p_2(s) \triangleq -\frac{1}{2}\left(1 - \frac{s}{\lambda} + \frac{(c_a^2+1)s^2}{2\lambda^2}\right) (c_x^2+1)\ (E[X])^2 + \left(\frac{1}{\lambda} - \frac{(c_a^2+1)s}{\lambda^2}\right) E[X] - \frac{c_a^2+1}{2\lambda^2}.$$

Proof: See Appendix.

Using the asymptotic expression for $idle(s)$ we can rewrite $\Phi_W(w,s)$ as follows

$$\Phi_W(w,s) \sim \frac{\frac{\phi_{V(0)}(w_2)}{w_2} - \phi_{V(0)}(w)}{p_2(s)(w-w_1)(w-w_2)} \frac{1-\alpha(s-w)}{(w-s)}. \tag{52}$$

On the other hand, using the Brownian approximation method for general arrival we get (see for example Kleinrock [17])

$$\Phi_W(w,s) \sim \frac{\frac{\phi_{V(0)}(\hat{w}_2)}{\hat{w}_2} - \phi_{V(0)}(w)}{\frac{1}{2\lambda}\rho^2(c_x^2+c_a^2)(w-\hat{w}_1)(w-\hat{w}_2)}, \tag{53}$$

with

$$\hat{w}_{1,2} = \frac{-\lambda(1-\rho)}{\rho^2(c_x^2+c_a^2)}\left[1 \mp \sqrt{1 + 2s\rho^2\frac{(c_x^2+c_a^2)}{\lambda(1-\rho)^2}}\,\right],$$

which is different from (52).

Using the asymptotic form of $idle(s)$ we can also obtain an asymptotic form of the Laplace transform of the expected queue length, $\mathcal{L}\{E[Q(t)]\}$ via Theorem 9. Indeed,

$$\mathcal{L}\{E[Q(t)]\} = \frac{\lambda}{s^2} - \frac{s\ idle(s) - \phi_{V(0)}(s)}{s(\phi_X(s)-1)} \sim \frac{\lambda}{s^2} - \frac{s\phi_{V(0)}(w_2) - w_2\phi_{V(0)}(s)}{w_2 s(\phi_X(s)-1)}. \tag{54}$$

It is interesting to note that if we calculate the asymptotic steady-state queue length, using the asymptotic value of w_2 we obtain

$$E[Q] = \lim_{s\to 0} s\ \mathcal{L}\{E[Q(t)]\} \sim \frac{\rho(c_a^2-1) + \rho^2(c_x^2+1)}{2(1-\rho)},$$

the same result we obtained in Bertsimas and Mourtzinou [4].

Another important performance measure is the expected waiting time of a customer that arrives at time t denoted by $E[W(t)]$. If we denote by $\mathcal{L}\{E[W(t)]\}$ its Laplace transform, i.e.,

$$\mathcal{L}\{E[W(t)]\} \triangleq \int_0^\infty e^{-st} E[W(t)]dt,$$

we have from the properties of Laplace transform that $\mathcal{L}\{E[W(t)]\} = \lim_{w\to 0} \frac{\partial}{\partial w}\Phi_W(w, s)$. Hence, we can prove the following corollary of Theorem 10.

Corollary 2 *In a $GI/G/1$ queue with FIFO service policy, and initial work $V(0)$, asymptotically in heavy traffic:*

$$\mathcal{L}\{E[W(t)]\} \sim \frac{idle(s)}{s} + \frac{E[V(0)]}{s} + \frac{E[X]\alpha(s)}{s(1-\alpha(s))} - \frac{1}{s^2}, \tag{55}$$

where $E[V(0)]$ is the expected initial work in the system.

Proof: Taking derivatives in (46) with respect to w and then calculating the limit as $w \to 0$ we we get (55). ■

If we calculate the steady-state expected waiting time $E[W] = \lim_{s\to 0} s\ \mathcal{L}\{E[W(t)]\}$ we get that

$$E[W] = \lim_{s\to 0} s\ \mathcal{L}\{E[W(t)]\} \sim \frac{\rho(c_a^2 - 1) + \rho^2(c_x^2 + 1)}{2\lambda(1-\rho)},$$

the same result we obtained in Bertsimas and Mourtzinou [4].

It is important to notice that although Theorem 9 holds for any traffic intensity and for any s, Corollary 2 only holds asymptotically in heavy traffic.

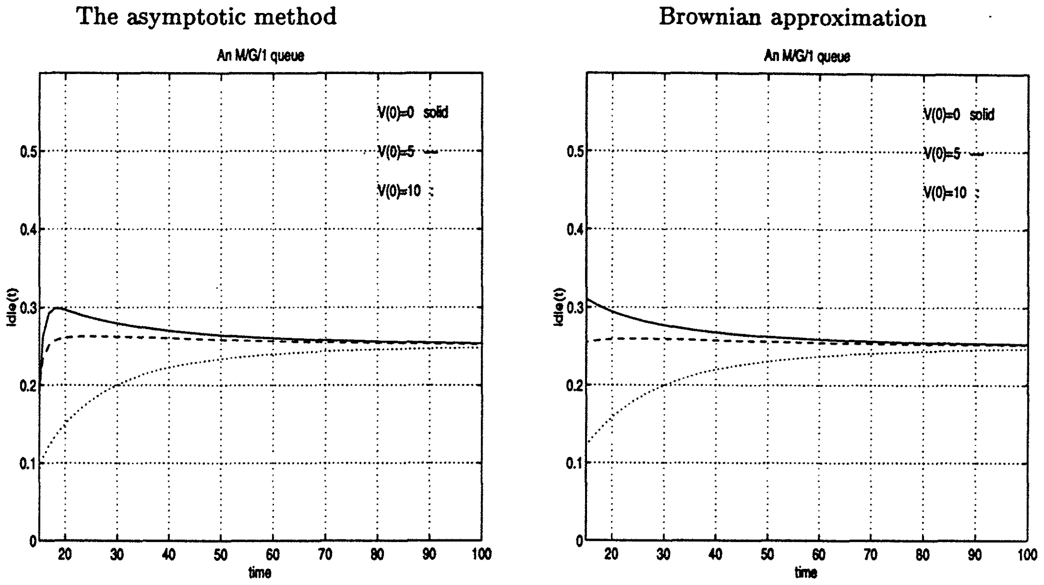
Numerical Results

In order to obtain a better understanding of the asymptotic method we now present some numerical results. We start by evaluating the function $idle(t)$ for an $M/G/1$ queue with $\lambda = 0.75$, $E[X] = 1$ and $c_x^2 = 2$ that starts empty with no initial work, the same queue if $V(0) = 5$ units and if $V(0) = 10$ units. Recall that $\lim_{t\to\infty} idle(t) = (1-\rho) = 0.25$. To invert the Laplace transform we used two algorithms, one proposed by Hosono in [12] and one proposed by Abate and Whitt in [1], and we got exactly the same results. Notice that we have asymptotically evaluated $idle(t)$ for $t >> t_o$, so our results for $t < 15$ are not very accurate and therefore we do not report them. The results for $idle(t)$ via our asymptotic method as well as the Brownian approximation are depicted in Figure 3. Notice that for times $t > 20$ the two methods produce identical results.

We next evaluate $idle(t)$ for an $E_2/E_2/1$ queue and an $H_2/H_2/1$ with $c_a^2 = 3$ and $c_x^2 = 1.5$, in Figure 4. In both cases we assume that $V(0) = 0$ units and we plot the results of both the asymptotic method and the Brownian approximation. For the $E_2/E_2/1$ queue the two methods give rise to identical results for $t > 12$; however in the case of the $H_2/H_2/1$ queue the two methods give rise to almost identical results only for $t > 30$.

Next, we calculate the difference $E[Q(t)] - E[Q]$ for an $E_2/H_2/1$ queue with $\lambda = 0.75$, $E[X] = 1$ and $c_x^2 = 3$, when $V(0) = 0$ units using our asymptotic method. Notice that for this system $E[Q] = 2.625$, according to our asymptotic method. For this particular system our results are relevant for $t > 80$ as the Figure 5 indicates.

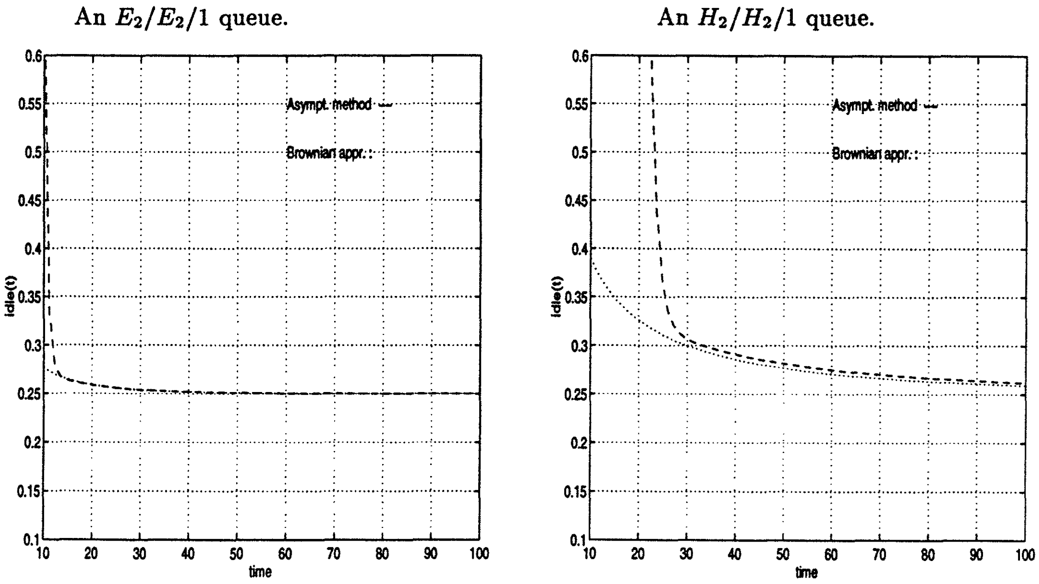
Furthermore, we calculate the difference $E[Q(t)] - E[Q]$ for an $H_2/E_{10}/1$ queue with $\lambda = 0.75$, $E[X] = 1$ and $c_a^2 = 1.5$, when $V(0) = 0$ units. Now, E[Q]=1.9875 and our results are relevant for $t > 20$.



The asymptotic method

Brownian approximation

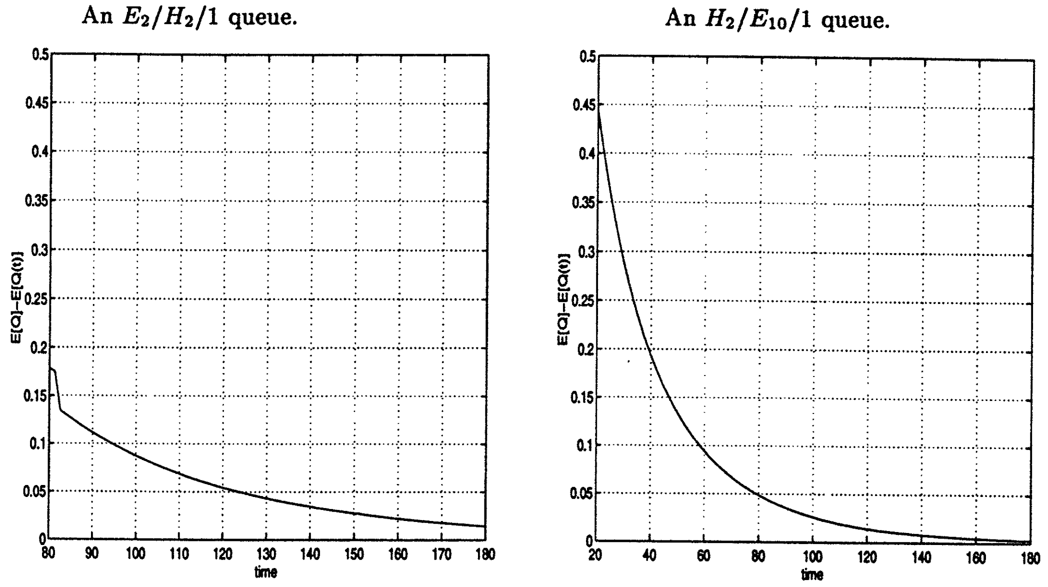
Figure 7: The function *idle*(t) for an M/G/1 queue.



An $E_2/E_2/1$ queue.

An $H_2/H_2/1$ queue.

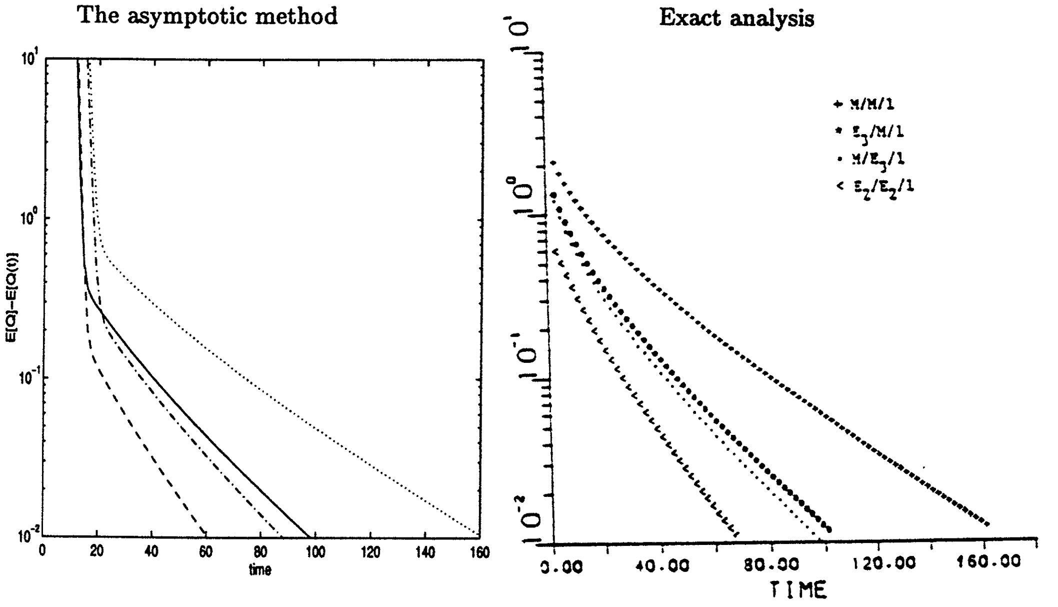
Figure 8: The function *idle*(t) for an $GI/GI/1$ queue with $V(0) = 0$.



An $E_2/H_2/1$ queue.

An $H_2/E_{10}/1$ queue.

Figure 9: The function $E[Q] - E[Q(t)]$ for an $GI/GI/1$ queue.



The asymptotic method

Exact analysis

Figure 10: A semilogarithmic plot of $E[Q] - E[Q(t)]$ -$\rho = 0.75$, $E[X] = 1$.

From the above figures we see that the performance of our asymptotic method in sensitive to the variance of the arrival and service time distributions. In particular, if we denote by t_o the earliest time for which our asymptotic method correctly predicts the behavior of the system, we observe that for systems where both the arrival and the service distributions are close to Poisson (i.e., c_a^2 and c_x^2 close to 1), $t_o \approx 20$. Moreover, $t_o \approx 20$ even if c_a^2 is big, provided that c_x^2 is small (see Figure 5, the case of the $H_2/E_{10}/1$ queue). On the other hand, for systems where c_a^2 is small and c_x^2 is big, t_o is bigger, for example $t_o \approx 80$ in Figure 5, the case of the $E_2/H_2/1$ queue.

It is also interesting to compare the predictions of the asymptotic method for $E[Q] - E[Q(t)]$ versus the exact values of $E[Q]-E[Q(t)]$; we do so in Figure 6, where use the exact results presented in Odoni and Roth [20] for various systems that start empty. Notice that the asymptotic method is performing very well and for all systems for $t > t_o \approx 30$.

The previous results for the GI/G/1 system can also be used in a GI/D/s queue. Since the service times are deterministic, every s customers are served by the same server. Therefore, as it is well known (see Iversen [13]), each customer sees a $GI^{(s)}/D/1$ queue, where $GI^{(s)}$ is the s fold convolution of the interarrival distribution. As a result, the waiting time in queue in the $GI/D/s$ queue is the same as in the $GI^{(s)}/D/1$ queue.

5.2.2 The M(t)/GI(t)/1 queueing system under FIFO

In this section we analyze single server systems with non-homogeneous Poisson arrivals and general time-dependent service time distributions that satisfy the following set of assumptions
Assumptions B :
B.1 There exists a set of ordered time epochs, $ta_i,\ i = 0, 1, 2, \ldots$ with $ta_0 \triangleq 0$, such that the arrival rate $\lambda(t)$ is piecewise constant with value $\lambda(t) = \lambda_i$ for $t \in [ta_i, ta_{i+1})$.
B.2 There exists a set of ordered time epochs, $ts_i,\ i = 0, 1, 2, \ldots$ with $ts_0 \triangleq 0$, such that the service time distribution $X(t) \overset{d}{=} X_i$ for $t \in [ts_i, ts_{i+1})$.

We define the set of all times epochs $T \triangleq \{ta_i,\ i \in \mathbb{Z}_+\} \bigcup \{ts_i,\ i \in \mathbb{Z}_+\}$ and let the set $O \triangleq \{0, t_1, t_2, \cdots\}$ be the ordering of the elements of T such that $t_i \leq t_j$ for $i \leq j$.

Since the arrival process is memoryless, we can decompose the system in the time intervals $[t_i, t_{i+1})$, for $i = 1, 2, \ldots$, in order to calculate the distribution of the waiting time. In other words, for $t \in [t_i, t_{i+1})$, if also $t \in [ta_k, ta_{k+1})$ and $t \in [ts_m, ts_{m+1})$, the original system behaves as an $M/GI/1$ queueing system with arrival rate λ_k, service time distribution represented by the random variable X_m and the appropriate initial work conditions.

Based on the above observation we define,

$$\Phi_{W_0}(w, s) = \frac{\frac{w}{\eta_0}\, \phi_{V(0)}(\eta_0) - \phi_{V(0)}(w)}{\lambda_0 \phi_{X_0}(w) - \lambda_0 - s + w}, \tag{56}$$

where $\phi_{V(0)}(w)$ is the Laplace transform of the initial work at $t = 0$, $V(0)$ and $\eta_0 \triangleq \eta_0(s)$ is the unique root of $\lambda_0 \phi_{X_0}(w) - \lambda_0 - s + w = 0$ in the region $\Re(s) > 0, \Re(w) > 0$. We also define $\phi_{W_0}(w, t)$ to be the inverse Laplace transform of $\Phi_{W_0}(w, s)$, i.e.,

$$\Phi_{W_0}(w, s) \triangleq \int_0^\infty e^{-st} \phi_{W_0}(w, t) \quad \text{equivalently} \quad \phi_{W_0}(w, t) = \mathcal{L}^{-1}\{\Phi_{W_0}(w, s)\}.$$

Finally, we define for all $i = 1, 2, \ldots$

$$\Phi_{W_i}(w,s) = \frac{\frac{w}{\eta_i}\,\phi_{W_{i-1}}(\eta_i, t_i) - \phi_{W_{i-1}}(w, t_i)}{\lambda_k \phi_{X_m}(w) - \lambda_k - s + w}, \tag{57}$$

where $\eta_i \triangleq \eta_i(s)$ is the unique root of $\lambda_k \phi_{X_m}(w) - \lambda_k - s + w = 0$ in the region $\Re(s) > 0, \Re(w) > 0$ (recall that Beneš [3] has shown that in this region this equation has a unique solution).

We next state the main theorem of this section (see Mourtzinou [19]).

Theorem 11 *For an $M(t)/GI(t)/1$ queueing system under FIFO that satisfies Assumptions B and starts with an arbitrary initial work $V(0)$ we can evaluate the Laplace transform of the distribution of $W(t)$ as follows.*

$$\phi_W(w,t) = \phi_{W_i}(w, t - t_i) \qquad for \quad t \in [t_i, t_{i+1}), \tag{58}$$

where $t_0 \triangleq 0$ and $\phi_{W_i}(w,t)$ is calculated recursively as follows

$$\phi_{W_0}(w,t) = \mathcal{L}^{-1}\{\frac{\frac{w}{\eta_0}\,\phi_{V(0)}(\eta_0) - \phi_{V(0)}(w)}{\lambda_0 \phi_{X_0}(w) - \lambda_0 - s + w}\},$$

$$\phi_{W_i}(w,t) = \mathcal{L}^{-1}\{\frac{\frac{w}{\eta_i}\,\phi_{W_{i-1}}(\eta_i, t_i) - \phi_{W_{i-1}}(w, t_i)}{\lambda_k \phi_{X_m}(w) - \lambda_k - s + w}\} \quad for\ t \in [t_i, t_{i+1}),$$

where $[t_i, t_{i+1}) \triangleq [ta_k, ta_{k+1}) \bigcap [ts_m, ts_{m+1})$.

The above theorem provides a recursive algorithm for obtaining the Laplace transform of the waiting time in a $M(t)/G(t)/1$ queue that satisfies Assumptions B.

Independently, Choudhury et. al. in [7] used a very similar approach to obtain the performance of the $M(t)/GI(t)/1$ queue under Assumptions B. The only difference is that we obtained the performance of the $M/GI/1$ queue using distributional laws and they obtained it using the Takács integrodifferential equation (see Takács [25]). In the same paper the authors also proposed an algorithm to numerically invert the Laplace transforms. We do not report numerical results since they coincide with those reported in Choudhury et. al. [7].

5.3 The ΣGI(t)/GI(t)/1 queue under FIFO

In this section we consider the multiclass $\Sigma GI(t)/G(t)/1$ queue under FIFO. We denote by $L_i(t)$ ($Q_i(t)$) the number of class i customers in the system (queue) at a random observation time t. We, also, denote by $G_{L_i}(z,t) \triangleq E[z^{L_i(t)}]$ the generating function of $L_i(t)$ and with $G_{L_i}(z,s)$ its Laplace transform (similar definitions hold for $G_{Q_i}(z,t), G_{Q_i}(z,s)$). Furthermore, $W_i(t)$ represents the waiting time of a customer that arrived at time t and $dF_{W_i(t)}(\cdot)$ is the pdf of $W_i(t)$. We assume, without loss of generality, that the server of the system is working with unit speed as long as there is work in the system and we denote by $X_i(t)$ the service requirement of a class i customer that *enters the server at time t.* Finally, we denote by $\vec{z} \triangleq (z_1, \dots, z_N)$ and by $G_{L_1,\dots,L_N}(\vec{z}, t)$ (resp. $G_{Q_1,\dots,Q_N}(\vec{z}, t)$) the joint generating function of $L_1(t), \cdots, L_N(t)$ (resp. $Q_1(t), \cdots, Q_N(t)$).

As in the single class case we first prove another distributional law that relates $G_{L_1,\dots,L_N}(\vec{z}, t)$ and $G_{Q_1,\dots,Q_N}(\vec{z}, t)$ and requires the existence of a single server.

Proposition 5 *In a $\Sigma GI(t)/GI(t)/1$ system with N-classes of customers that satisfies Assumptions A.1-A.4:*

$$G_{L_1,\dots,L_N}(\vec{z}, t) = idle(t) + \sum_{j=1}^{N} z_j \int_0^t h_j(a)\ K_{o,j}(z_j, a, t)\ M_j(a, t) \prod_{\substack{i=1 \\ i \neq j}}^{N} K_{e,i}(z_i, a, t)\ da, \tag{59}$$

$$G_{Q_1,\dots,Q_N}(\vec{z}, t) = idle(t) + \sum_{j=1}^{N} \int_0^t h_j(a)\ K_{o,j}(z_j, a, t)\ M_j(a, t) \prod_{\substack{i=1 \\ i \neq j}}^{N} K_{e,i}(z_i, a, t)\ da, \tag{60}$$

$$G_{L_i}(z, t) = zG_{Q_i}(z, t) + (1 - z) \left[idle(t) + \sum_{\substack{j=1 \\ j \neq i}}^{N} \int_0^t h_j(a) K_{e,i}(z, a, t) M_j(a, t) da \right], \tag{61}$$

where idle(t) is the emptiness process, $M_i(a, t) \triangleq P\{S_i(a) > t - a \geq W_i(a)\}$, and

$$K_{o,i}(z_i, a, t) \triangleq \sum_{n=0}^{\infty} z_i^n\ P\{N_i(a; t) = n\},$$

$$K_{e,i}(z_i, a, t) \triangleq 1 + (z_i - 1) \int_a^t h_i(u)\ K_{o,i}(z_i, u, t)\ du.$$

Proof: The proof of this theorem is based on the main ideas used to prove Theorem 5. Hence, we define τ_{i,n_i} to be the arrival time of the n_ith customer of the ith class and S_{i,n_i} to be his system time. Recall that according to the notational convention we keep using customers are indexed backwards in time.

The key idea is that in order to have at time t, *exactly* n_i customers of the ith class in the system, where $n_i \geq 1$, we must have that for $i = 1, \dots, N$ the n_ith customer of the ith class is still in the system at t and that the customer who arrived the first to the system (independent of class) is in the service facility.

In other words, if we denote by $\mathcal{A} \subset \{1, \dots, N\}$ the set of classes such that $k \in \mathcal{A}$ iff $n_k \geq 1$, the event

$$\{L_1(t) \geq n_1, \cdots, L_j(t) = n_j, \cdots, L_N(t) \geq n_N \text{ and } \tau_{j,n_j} = \min_i\ \tau_{i,n_i}\} \text{ for} j \in \mathcal{A}$$

is equivalent to the intersection of the following events, for all $i \in \mathcal{A}\ i \neq j$:

$\Gamma_{1,j}$: a customer of the jth class arrives at time $a_j \in (0, t]$,
$\Gamma_{2,j}$: the system time of the customer who arrived at a_j is greater than $t - a_j$,
$\Gamma_{3,j}$: the waiting time of the customer who arrived at a_j is less than $t - a_j$,

$\Gamma_{4,j}$: there are *exactly* $n_j - 1$ arrivals at $(a_j, t]$ for the j^{th} class, given an arrival at a_j,
$\Gamma_{1,i}$: a customer of the ith class arrives at time $a_i \in (a_j, t]$,
$\Gamma_{4,i}$: there are *at least* $n_i - 1$ arrivals at $(a_i, t]$ for the ith class, given an arrival at a_i.

From the definition of events $\Gamma_{1,i}$ and $\Gamma_{4,i}$ are independent for any fixed a_i, for $i = 1, \ldots, N$. Similarly, the events $\Gamma_{1,j}$ and $\Gamma_{4,j}$ are independent form each other and also independent of the events $\Gamma_{2,j}$ and $\Gamma_{3,j}$ for any fixed a_j. Moreover, from Assumption A.4, the events $\Gamma_{1,i}$ and $\Gamma_{4,i}$ for all $i = 1, \ldots, N$, are also mutually independent. Hence,

$$P\{L_1(t) \geq n_1, \cdots, L_j(t) = n_j, \cdots, L_N(t) \geq n_N \text{ and } \tau_{j,n_j} = \min_i \ \tau_{i,n_i}\}$$
$$= \int_{a_j=0}^{t} \int_{a_1=a_j}^{t} \cdots \int_{a_N=a_j}^{t} P\{\Gamma_{2,j} \bigcap \Gamma_{3,j}\} \prod_{i=1}^{N} P\{\Gamma_{4,i}\} \prod_{i=1}^{N} P\{\Gamma_{1,i}\} \ da_1 \cdots da_N.$$

Conditioning on the event $\Gamma_{2,i} \bigcap \Gamma_{3,j}$, and following the proof of Theorem 5, after some tedious but straightforward manipulations, we obtain relation (59).

Next, to prove (60) we observe that for $j, i \in \mathcal{A}$,

$$P\{Q_1(t) \geq n_1, \cdots, Q_j(t) = n_j, \cdots, Q_N(t) \geq n_N \text{ and } \tau_{j,n_j} = \min_i \ \tau_{i,n_i}\}$$
$$= P\{L_1(t) \geq n_1, \cdots, L_j(t) = n_j + 1, \cdots, L_N(t) \geq n_N \text{ and } \tau_{j,n_j} = \min_i \ \tau_{i,n_i}\}.$$

Finally, by combining (59) and (60) and setting $z_i = z$ and $z_j = 1, \ j \neq i$, we also obtain (61). ■

Using Proposition 5 together with the multiclass transient distributional and the fact that for all $i = 1, \ldots, N$

$$P\{S_i(t) > x\} = P\{a \leq W_i(t) \leq a + da\} \ P\{X_i(t + a) > x - a\},$$

we obtain a system on N integral equations on N unknowns, the cdf of $W_i(t)$ for $i = 1, \ldots, N$. This system constitutes a complete description of the fundamental quantities of a $\Sigma GI(t)/G(t)/1$ queue as functions of $idle(t)$ and can be solved numerically. We, next, focus our attention to the $\Sigma GI/G/1$ queueing system, where under heavy traffic conditions we can obtain closed form expressions.

Transient analysis of the ΣGI/GI/1 queue under FIFO

Consider the multiclass $\Sigma GI/GI/1$ queue under FIFO and denote by $\phi_{W_i}(w, t)$ the Laplace transform of $W_i(t)$ and by $\Phi_{W_i}(w, s)$ the double Laplace transform of $W_i(\cdot)$. Similarly, $S_i(t)$ has Laplace transform $\phi_{S_i}(w, t)$ and double Laplace transform $\Phi_{S_i}(w, s)$. Since all arrival processes are renewal we write $K_{oi}(z, u, t)$ as $K_{oi}(z, t - u)$. Moreover, as the arrival processes are equilibrium processes $h_i(u) = \lambda_i$ for all $u \geq 0$, and $N_{a_i}()$ where λ_i is the arrival rate for class i.

Theorem 12 *In a $\Sigma GI/GI/1$ system under FIFO that starts empty the Laplace transforms of the* *individual waiting times asymptotically under heavy traffic conditions are given by*

$$\Phi_{W_i}(w,s) \sim \frac{C(w,s)\ (1-\alpha_i(s-w))}{1-\alpha_i(s-w)\phi_{X_i}(w)-(1-\alpha_i(s-w))\rho_i\phi_{X_i^*}(w)}, \tag{62}$$

where $\alpha_i(s)$ is the Laplace transform of the interarrival distribution of the ith class and $C(w,s) \triangleq \frac{idle(s)}{1-D(w,s)}$ and

$$D(w,s) \triangleq \sum_{i=1}^{N} \frac{\rho_i\phi_{X_i^*}(w)(1-\alpha_i(s-w))}{1-\alpha_i(s-w)\phi_{X_i}(w)-(1-\alpha_i(s-w))\rho_i\phi_{X_i^*}(w)}.$$

Proof: For any class i we apply the single class transient distributional law to the pair $(Q_i(t), W_i(t))$ and also $(L_i(t), S_i(t))$ to obtain

$$G_{Q_i}(z,t) = 1 + \lambda_i(z-1)\int_0^t P\{W_i(u) > t-u\}\ K_{o,i}(z,t-u)\ du$$
$$G_{L_i}(z,t) = 1 + \lambda_i(z-1)\int_0^t P\{S_i(u) > t-u\}\ K_{o,i}(z,t-u)\ du$$

Taking Laplace transforms and using the same line of arguments used to prove (44) and (45) we obtain

$$G_{Q_i}(z,s) \sim \frac{1}{s+f_i(z)}\ [1 + f_i(z)\ \Phi_{W_i}(s+f_i(z),s)], \tag{63}$$
$$G_{L_i}(z,s) \sim \frac{1}{s+f_i(z)}\ [1 + f_i(z)\ \Phi_{S_i}(s+f_i(z),s)], \tag{64}$$

where $f_i(z) = \lambda_i(1-z) - \frac{1}{2}\lambda(1-z)^2(c_{a_i}^2 - 1)$.

We take double the Laplace transform of the relation $S_i(t) = W_i(t) + X_i$ to obtain

$$\Phi_{S_i}(s+f_i(z),s) = \Phi_{S_i}(s+f_i(z),s)\ \phi_X(s+f_i(z)). \tag{65}$$

Finally, we take Laplace transforms in (61) and we obtain

$$G_{L_i}(z,s) \sim zG_{Q_i}(z,s) + (1-z)\left[idle(s) + \sum_{\substack{j=1\\ j\neq i}}^{N} \lambda_j \int_0^\infty e^{-st}\int_0^t e^{-(t-a)(s+f_i(z))}M_j(a,t)dadt\right].$$

If we go through the algebra of Laplace transforms we obtain that

$$\int_0^\infty e^{-st}\int_0^t e^{-(t-a)(s+f_i(z))}M_j(a,t)dadt = \Phi_{W_j}(s+f_i(z),s)\ \phi_{X_j^*}(s+f_i(z))\ E[X_j],$$

where X_j^* is the age of the service time distribution and $E[X_j]$ is the expected service time. So we can write that

$$G_{L_i}(z,s) \sim zG_{Q_i}(z,s) + (1-z)\left[idle(s) + \sum_{\substack{j=1 \\ j\neq i}}^{N} \rho_j \Phi_{W_j}(s+f_i(z),s)\ \phi_{X_j^*}(s+f_i(z)) \right] \tag{66}$$

Combining (63)-(66), and multiplying and dividing by $\alpha_i(-f_i(z))$ (as in the single class case) and finally setting $w \triangleq s + f_i(z)$ we have for all $i = 1, \ldots, N$

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

The previous equations form a $N \times N$ linear system which can be solved by adding and subtracting $\rho_i(1-\alpha_i(s-w))\phi_{X_i^*}(w)\ \Phi_{W_i}(w,s)$. We can then solve for each $\Phi_{W_i}(w,s)$ as a function of $\sum_{j=1}^{N} \rho_j \phi_{X_j^*}(w)\ \Phi_{W_j}(w,s)$, from which (62) follows.

Notice that the function $idle(s)$ can be determined from the analytic properties of $\Phi_{W_i}(w,s)$ for all $i = 1, \cdots, N$. ■

6 Concluding Remarks

In this paper we established a set of "laws" that completely characterize the performance of a broad class of multiclass queueing systems that are operating in a time-varying environment. An important characteristic of the laws we derived, is that they provide insight on the influence of the initial conditions for systems that are operating under a time-varying environment. Moreover, they give rise to structural results such as a transient extension of Little's law. Finally, we applied this set of laws as well as the transient extension of Little's law to specific queueing systems and presented several insights and new results.

Although we demonstrated in this paper the power of the proposed approach in several applications there exist many systems widely used in real world applications that our method does not address, such as multiserver queueing systems and queueing networks. The major open problem is to identify queueing laws for these systems. A solution to this rather challenging but important problem will lead to a more complete theory of queues and is likely to provide very valuable new insights.

References

[1] J. Abate and W. Whitt, *Numerical inversion of Laplace transforms of probability distributions*, Journal on Computing **7** (1995), 36–43.

[2] M. S. Bartlett, *Some evolutionary stochastic processes*, J. Roy. Statist. Soc. B **11** (1949), 211–229.

[3] V. E. Beneš, *On queues with Poisson arrivals*, Annals of Mathematical Statistics **28** (1956), 670–677.

[4] D. Bertsimas and G. Mourtzinou, *A unified method to analyze overtake-free queueing systems*, Working paper, Operations Research Center, MIT, 1992, To appear in *Advances in Applied Probability.*

[5] ______, *Multiclass queueing systems in heavy traffic: An asymptotic approach based on distributional and conservation laws*, Working paper, Operations Research Center, MIT, 1993, To appear in *Operations Research.*

[6] D. Bertsimas and D. Nakazato, *The distributional Little's law and its applications*, Operations Research **43** (1995), 298–310.

[7] D. M. Choudhury, G. L. Lucantoni and W. Whitt, *Numerical solution of $M(t)/G(t)/1$ queues*, Working paper, AT&T Bell Labs, 1993.

[8] D. R. Cox, *Renewal theory*, Chapman and Hall, New York, 1962.

[9] J. L. Doob, *Stochastic processes*, John Wiley, New York, 1953.

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

[11] D. Heyman and M. Sobel, *Stochastic models in operations research: Vol. 1*, McGraw-Hill Book Company, New York, 1982.

[12] T. Hosono, *Numerical inversion of Laplace transform and some applications to wave optics*, Radio Science **16** (1981), 1015–1019.

[13] V. B. Iversen, *Decomposition of an $M/D/r \cdot k$ queue with FIFO into k $E_k/D/r$ queues with FIFO*, Operations Research Letters **2** (1983), 20–21.

[14] J. Keilson and L. Servi, *A distributional form of little's law*, Operations Research Letters **7** (1988), 223–227.

[15] ______, *The distributional form of Little's law and the Fuhrmann-Cooper decomposition*, Operations Research Letters **9** (1990), 239–247.

[16] A. Y. Khintchine, *Mathematical methods in the theory of queues*, (in Russian) Trudy Mat. Inst. Steklov, 49 (English translation by Charles Griffin & Co, London), 1955.

[17] L. Kleinrock, *Queueing systems; vol. 1: Theory*, Wiley, New York, 1975.

[18] W. A. Massey and W. Whitt, *Networks of infinite-server queues with non-stationary Poisson input*, Queueing Systems **13** (1993), 183–250.

[19] G. Mourtzinou, *An axiomatic approach to queueing systems*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Mass, 1995.

[20] A. Odoni and E. Roth, *An empirical investigation of the transient behavior of stationary queueing systems*, Operations Research **31** (1983), 432–455.

[21] C. Palm, *Intensity variations in telephone traffic*, Ericson Technics **44** (1943), 1–189 (in German)(English translation by North–Holland, Amsterdam, 1988).

[22] A. Prékopa, *On secondary processes generated by a random point distribution of Poisson type*, Annales Univ. Sci. Budapest de Eötvös Nom. Sectio. Math. **1** (1958), 153–170.

[23] S. Ross, *Introduction to probability models: 5th edition*, Academic Press, London, 1993.

[24] W. L. Smith, *Asymptotic renewal theorems*, Proc. Roy. Soc. Edinb. A **64** (1954), 9–48.

[25] L. Takács, *Investigation of waiting time problems by reduction to Markov processes*, Acta. Math. Acad. Sci. Hung. **6** (1955), 101–129.

[26] R. Wolff, *Stochastic modeling and the theory of queues*, Prentice Hall, 1982.

Appendix

In this Appendix we give a proof of Proposition 4: Recall that $idle(s)$ may be determined by insisting that the transform $\Phi_W(w,s)$ is analytic in the region $\Re(s) > 0$ and $\Re(w) > 0$, where

$$\Phi_W(w,s) \sim \frac{w\ idle(s) - \phi_{V(0)}(w)}{\alpha(s-w)\ \phi_X(w) - 1}\ \frac{1-\alpha(s-w)}{(w-s)}.$$

Since our asymptotic formula holds for both s, w small, we can expand $\alpha(s-w)$ as a Taylor series around $s-w$ and obtain

$$\alpha(s-w) = 1 - \frac{1}{\lambda}(s-w) + \frac{1}{2}\frac{c_a^2+1}{\lambda^2}(s-w)^2 + O((s-w)^3).$$

Hence, we have that:

$$\frac{1-\alpha(s-w)}{(w-s)} = -\frac{1}{\lambda} - \frac{1}{2}\frac{c_a^2+1}{\lambda^2}(w-s) + O((s-w)^2)$$

so that $\frac{1-\alpha(s-w)}{(w-s)}$ is analytic in the region $\Re(s) > 0$ and $\Re(w) > 0$. Therefore, $\Phi_W(w,s)$ is analytic in the region $\Re(s) > 0$ and $\Re(w) > 0$ if and only if $\frac{w\ idle(s) - \phi_{V(0)}(w)}{1-\alpha(s-w)\ \phi_X(w)}$ is analytic in the same region. Expanding the denominator around $w = 0$ and $s = 0$, we get that

$$1 - \alpha(s-w)\ \phi_X(w) \sim p_0(s) + p_1(s)w + p_2(s)w,$$

where if we denote by $E[X]$ the mean service time and by c_x^2 the squared coefficient of variation of X we have:

$$p_0(s) \triangleq \frac{s}{\lambda} - \frac{(c_a^2+1)s^2}{2\lambda^2},$$

$$p_1(s) \triangleq \left(1 - \frac{s}{\lambda} + \frac{(c_a^2+1)s^2}{2\lambda^2}\right) E[X] - \frac{1}{\lambda} + \frac{(c_a^2+1)s}{\lambda^2},$$

$$p_2(s) \triangleq -\frac{1}{2}\left(1 - \frac{s}{\lambda} + \frac{(c_a^2+1)s^2}{2\lambda^2}\right)(c_x^2+1)\ (E[X])^2 + \left(\frac{1}{\lambda} - \frac{(c_a^2+1)s}{\lambda^2}\right) E[X] - \frac{c_a^2+1}{2\lambda^2}.$$

Equivalently, we have that

$$1 - \alpha(s-w)\ \phi_X(w) \sim p_2(s)\ (w-w_1)(w-w_2) \quad \text{with} \quad w_{1,2} = \frac{-p_1(s) \mp \sqrt{(p_1(s))^2 - 4p_0(s)\ p_2(s)}}{2p_2(s)},$$

with w_1 corresponding to the + sign and w_2 to the - sign. Notice that for $s \approx 0$ we have that

$$p_1(s) \approx E[X] - \frac{1}{\lambda} = \frac{1}{\lambda}(\rho - 1) < 0 \quad \text{and} \quad p_2(s) \approx \frac{E[X]}{\lambda} - \frac{1}{2\lambda^2}\left(\rho^2(c_x^2+1) + c_a^2 + 1\right) < 0.$$

Therefore we have that $\Re(w_2) > 0$ and $\Re(w_1) < 0$ for s small. Then, from the analytic properties of $\Phi_W(w,s)$ we obtain (51). ■