Transient laws of non-stationary queueing systems and their applications

Dimitris Bertsimas [a,b,*] and Georgia Mourtzinou [b]

[a] *Sloan School of Management, MIT, Cambridge, MA 02139, USA*
[b] *Operations Research Center, MIT, Cambridge, MA 02139, USA*

Received 9 August 1995; revised 30 August 1996

In this paper we consider the general class of non-stationary queueing models and identify structural relationships between the number of customers in the system and the delay at time t, denoted by $L(t)$ and $S(t)$, respectively. In particular, we first establish a transient Little's law at the same level of generality as the classical stationary version of Little's law. We then obtain transient distributional laws for overtake free non-stationary systems. These laws relate the *distributions* of $L(t)$ and $S(t)$ and constitute a complete set of equations that describes the dynamics of overtake free non-stationary queueing systems. We further extend these laws to multiclass systems as well. Finally, to demonstrate the power of the transient laws we apply them to a variety queueing systems: Infinite and single server systems with non-stationary Poisson arrivals and general non-stationary services, multiclass single server systems with general non-stationary arrivals and services, and multiserver systems with renewal arrivals and deterministic services, operating in the transient domain. For all specific systems we relate the performance measures using the established set of laws and obtain a complete description of the system in the sense that we have a sufficient number of integral equations and unknowns. We then solve the set of integral equations using asymptotic expansions and exact numerical techniques. We also report computational results from our methods.

Keywords: non-stationary systems, transient analysis, Little's law, non-homogeneous Poisson process

1. Introduction

Transient analysis of queueing models has long been considered a very difficult problem that becomes even more complicated when we allow for the arrival and/or the service rates to change over time, i.e., when we consider the general class of non-stationary systems. On the other hand, it has long been recognized that non-stationary queueing models are needed to most appropriately model many complex production, service, communication and air transportation systems. Furthermore, even for the subclass of stationary systems, i.e., systems with constant arrival and service rates,

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

the convergence to steady-state is often so slow that the equilibrium behavior is not indicative of the system behavior. Therefore, the need arises for a better understanding of the transient behavior of queueing models.

In this paper we consider the general class of non-stationary queueing systems and we address the following questions: are there "transient laws" of non-stationary queueing systems? In other words, are there generic relationships between the fundamental quantities of interest for general queueing systems (such as the number of customers in the system and the delay at time t)? If so, how can they be used in particular applications?

Traditionally, transient analysis of queueing systems is addressed either via simulation or via approximate numerical methods. Computer simulation may provide insights on the system evolution, but many replication runs are required to obtain good estimates of time dependent probability distributions. In some cases, approximate numerical methods can be effective; see, for example, Odoni and Roth [25], Malone [22], Ong and Taaffee [26] and references cited therein.

In the last decades, work on the transient behavior of queueing systems has been mainly concentrated on exact numerical techniques; see, for example, Choudhury et al. [8], Green, Kolesar and Svoronos [12], and references cited therein. Moreover, there exist analytical results for infinite server systems (Massey and Whitt [23]), and for systems with phase type arrival and service distributions (Bertsimas and Nakazato [6]).

This paper takes a different standpoint, trying to identify structural relationships between the fundamental quantities of queueing systems that evolve over time. These relationships are next used in a variety of applications to obtain specific results. Our approach has its methodological foundation on Little's law and its extensions. In particular, Little [21] and Stidham [30] expressed via Little's law a fundamental principle of queueing theory: for stationary systems in steady-state, under very general conditions, the expected number of customers in the system, $E[L]$, is equal to the product of the arrival rate, λ, and the expected time a customer spends in the system, $E[S]$. An important generalization of Little's law, for overtake free systems, are the distributional laws first obtained by Haji and Newell [13]. The significance of these laws lies in the fact that, as demonstrated by Keilson and Servi [17,18] and Bertsimas and Mourtzinou [4,5], they can lead to complete solutions for a variety of systems in steady-state.

The major contributions of this work are as follows:

1. We establish a transient Little's law at the same level of generality as the classical stationary version of Little's law. This transient Little's law relates the expected number of customers in the system at time t, to the system time of customers joining the system in the interval $(0, t]$. It is important to notice that the form of the law depends on the initial conditions and therefore it demonstrates the influence of the initial state on the evolution of the system.

2. We obtain transient distributional laws for overtake free non-stationary systems. These laws relate the *distributions* of the most commonly used transient performance measures, i.e.,

   (a) the number of customers in the system at time t, denoted by $L(t)$ and

   (b) the system time, $S(t)$, of a customer that arrived to the system at $(t - \mathrm{d}t, t]$.

   They constitute a complete set of equations that describes the dynamics of overtake free non-stationary queueing systems. Moreover, the multiclass versions of these laws capture the interaction of customers from different classes, in the case of multiclass non-stationary systems.

3. To demonstrate the power of transient Little's law and the transient distributional laws we apply them to a variety of specific queueing systems. In particular, we consider

   (a) Infinite server systems with non-stationary Poisson arrivals and general non-stationary services. In this case, we establish the known exact formula for the number of customers in the system at time t.

   (b) Single server systems with non-stationary Poisson arrivals and general non-stationary services. In this case, we propose an algorithm to obtain exact numerical results for $L(t)$ and $S(t)$.

   (c) Multiclass single server systems with renewal arrivals and services, operating in the transient domain. In this case, we obtain a set of integral equations that completely characterize the distributions of the performance measures and then we obtain explicit asymptotic formulae. Furthermore, we numerically investigate the performance of our asymptotic expressions and assess their proximity to simulation results in a variety of settings. We also determine that they give rise to very similar numerical results with traditional Brownian approximations.

   (d) Multiserver systems with renewal arrivals and deterministic services, operating in the transient domain. We obtain integral equations that completely characterize the distributions of the performance measures and then we obtain explicit asymptotic formulae.

   For all specific systems we use the same approach: we 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. Once we formulate the stochastic system the next step is to actually solve it; in this step we use different math techniques depending on the application.

The rest of this paper is structured as follows: in section 2, we define some notation and present some mathematical preliminaries. In section 3, we establish the transient generalization of Little's law based on a sample path argument. Then, in section 4 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 5, we apply the transient laws to derive the transient performance analysis of several systems. Finally, in section 6 we present some concluding remarks.

2. Notation and preliminaries

In this section we consider a generic queueing system where customers arrive bringing random service requirements and upon completion of service they leave. To simplify the exposition, we assume a single class of customers; whenever we consider multiclass systems we will appropriately adapt our 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 $0 < t \leqslant \infty$. Note that the counting process $N_a(t)$ is completely defined when we know the probability distribution of the random variables $T_j,\ j = 1, 2, \ldots,$ via

$$N_a(t) \geqslant n \quad \text{if and only if} \quad T_n \leqslant t. \tag{1}$$

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 $t > 0$. 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 $t > 0$.

We also define $h(t)$ to be the "local rate" (see Cox and Isham [10]), i.e.,

$$h(t) \overset{\Delta}{=} \lim_{\Delta t \to 0} E\big[N_a(t) - N_a(t - \Delta t)\big], \tag{2}$$

assuming that the limit exist. In our analysis we do not allow multiple arrivals and therefore $h(t)\Delta t$ (as $\Delta t \to 0$) is the probability of an arrival in $(t - \Delta t, t]$ – given that the limit in (2) exists. In the special case of renewal arrival processes, for the limit to exist the interarrival distribution has to be absolutely continuous. Then, $h(t)$ is simply the derivative of the renewal function and 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 $h(t) = \lambda,\ t \geqslant 0$, where $1/\lambda$ is the mean interarrival. 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. On the other hand, for a nonhomogeneous Poisson arrival process with rate $\lambda(t)$, $h(t) = \lambda(t)$.

So far we have not imposed any assumptions on the counting process of the arrivals other than the ones necessary for the existence of the limit in (2). However for the analysis of section 4 and thereafter, we will assume that successive interarrivals satisfy a particular independence criterion:

Assumption A.1. The time interval between two successive arrival epochs, $A_i(T_i) \overset{\Delta}{=} T_{i+1} - T_i$, is independent of $\{A_j(T_j),\ j < i\}$, conditional on T_i, for all $i = 1, 2, \ldots$. Moreover, new arriving customers do not affect the time in the system of previous customers.

Examples of arrival processes satisfying Assumption A.1 include (a) all renewal processes, and (b) a nonhomogeneous Poisson process of rate $\lambda(t)$. An example of a non traditional, although somewhat contrived, process satisfying Assumption A.1 is a counting process with $A_i(T_i)$ being distributed as an Erlang 2 random variable with rate $\lambda(T_i)$.

Under Assumption A.1 we introduce some further notation. 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:

$$N_a^o(t_o, t) \geqslant n \quad \text{if and only if} \quad \sum_{i=1}^{n} A_i(t_o) \leqslant t - t_o. \tag{3}$$

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

Moreover, for the analysis of section 4 and thereafter, we will assume that customers bring to the system random stationary identically distributed service requirements. We will, however, allow for the service rate, i.e., the maximum number of units of work that the server can clear in one unit of time, to change over time. In this sense, the time a random customer spends in the service facility is a non-stationary random variable and it only depends on the time the customer enters the service facility. Furthermore, we assume that the interarrivals and service requirements are mutually independent.

We use the notation $GI(t)/G(t)/s$ to denote s-server systems with service process as described above and non-stationary arrivals satisfying Assumption A.1. In the case where successive service requirements are mutually independent, we use the notation $GI(t)/GI(t)/s$.

The natural transient performance measures in such a generic queueing system are

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

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

- S^j the time that the jth customer spends in the system.

For our analysis we introduce another performance measure, $S(t)$, by defining its distribution for each time $t > 0$ as follows:

$$h(t)\,\mathrm{d}tP\bigl\{S(t) > \tau\bigr\} \triangleq \sum_{n=1}^{\infty} P\{t - \mathrm{d}t < T_n \leqslant t\}P\bigl\{S^n > \tau \mid T_n = t\bigr\},$$

for $t, \tau > 0$. Intuitively,

- $S(t)$ is the time that a customer who arrived to the system at $(t - \mathrm{d}t, t]$ 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., on the initial number of customers, $L(0)$, as well as on the initial work,

$$V(0) \triangleq \widehat{V}(0) + \sum_{i=1}^{L(0)} X_i,$$

where $\widehat{V}(0)$ is the set-up work in the system, which is independent of the number of the initial customers, and X_i is the service requirement of the ith initial customer. Without loss of generality we will assume for the rest of paper that the initial customers – if any – are going to receive service in the following order $(X_1, X_2, \ldots, X_{L(0)})$.

3. Transient Little's law

One of the most celebrated results in queueing theory is that for systems in steady-state under natural and rather mild assumptions (see Heyman and Sobel [14]), 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. For general systems that are not functioning in steady-state, we prove the following transient generalizations of this result.

Theorem 1. For a single class system that starts empty, $L(0) = V(0) = 0$, if we denote 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 $(u - \mathrm{d}u, u]$, we have that

$$E\bigl[L(t)\bigr] = \int_0^t h(u)P\bigl\{S(u) > t - u\bigr\}\,\mathrm{d}u, \tag{4}$$

where

$$h(u) \triangleq \lim_{\Delta u \to 0} E\bigl[N_a(u) - N_a(u - \Delta u)\bigr].$$

Proof. 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

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

Then it is clear that

$$l(t;\omega) = \int_0^t f_t(u;\omega)\,\mathrm{d}u.$$

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

$$E\big[L(t)\big] = E\left[\int_0^t F_t(u)\,\mathrm{d}u\right] = \int_0^t E\big[F_t(u)\big]\,\mathrm{d}u, \tag{5}$$

where the second equality follows from Fubini's theorem. Moreover,

$$\begin{aligned} \int_0^t E\big[F_t(u)\big]\,\mathrm{d}u &= \int_0^t P\{\text{an arrival at } (u - \mathrm{d}u, u] \text{ who is still in the system at } t\} \\ &= \int_0^t h(u)P\big\{S(u) > t - u\big\}\,\mathrm{d}u. \end{aligned} \tag{6}$$

Hence we proved (4). □

Next, we extend Theorem 1 to the more general case where the systems starts with initial some customers $L(0) \neq 0$ and initial work $V(0)$. In particular we consider a system that starts with k initial customers, $L(0) = k$, and initial work $V(0) = \widehat{V}(0) + X_1 + \cdots + X_k$, where X_i is the service time requirement of the ith initial customer. To simplify the presentation, we assume that

Assumption B. The system clears first the set-up work $\widehat{V}(0)$, then the initial customers in the order $\{1, \ldots, k\}$, and then starts working on the customers who arrive after time 0. Moreover, until the system finishes the initial work $V(0)$, it does not idle.

The transient Little's law is as follows:

Theorem 2. For a single class system that starts with k initial customers, $L(0) = k$, and initial work $V(0) = \widehat{V}(0) + X_1 + \cdots + X_k$, and satisfies Assumption B, we have that

$$E\big[L(t)\big] = \int_0^t h(u)P\big\{S(u) > t - u\big\}\,\mathrm{d}u + \sum_{i=1}^{k} P\{V_i \geqslant t\}, \tag{7}$$

where $V_i \triangleq \widehat{V}(0) + X_k + \cdots + X_i$, for $i = k, \ldots, 1$.

Proof. 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. In this case,

$$l(t;\omega) = m(t;\omega) + \int_0^t f_t(u;\omega)\,\mathrm{d}u.$$

If we denote by $M(t)$ the number of initial customers still present in the system at time t, we get that

$$\begin{aligned} E\bigl[L(t)\bigr] &= E\left[\int_0^t F_t(u)\,\mathrm{d}u\right] + E\bigl[M(t)\bigr] \\ &= \int_0^t h(u)P\bigl\{S(u) > t-u\bigr\}\,\mathrm{d}u + E\bigl[M(t)\bigr]. \end{aligned} \tag{8}$$

Since $M(t)$ is the number of initial customers still present in the system at time t, $M(t) \in \{1,\dots,k\}$. In particular, since the server services the initial customers first, under Assumption B, we have that

$$P\bigl\{M(t) = k\bigr\} = P\bigl\{\widehat{V}(0) + X_k \geqslant t\bigr\}$$

and

$$P\bigl\{M(t) = i\bigr\} = P\{V_i \geqslant t\} - P\{V_{i+1} \geqslant t\},$$

with $V_i \overset{\Delta}{=} \widehat{V}(0) + X_k + \cdots + X_i$. Therefore,

$$\begin{aligned} E\bigl[M(t)\bigr] &= kP\{V_k \geqslant t\} + \sum_{i=1}^{k-1} i\bigl[P\{V_i \geqslant t\} - P\{V_{i+1} \geqslant t\}\bigr] \\ &= \sum_{i=1}^{k} P\{V_i \geqslant t\}. \end{aligned} \tag{9}$$

From (5), (6) and (9), (7) follows. □

It is important to notice that (8) holds independently of Assumption B; we introduced Assumption B to quantify $E[M(t)]$ and obtain (9). Hence, we can relax Assumption B and still get a transient Little's law.

Recall that for systems in steady-state the classical Little's law is

$$E[L] = \lambda \int_0^\infty P\{S > t\}\,\mathrm{d}t = \lambda E[S].$$

Notice that unlike the steady-state Little's law, $E[L(t)]$ in (7) and (4) 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

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

4. Transient distributional laws

In this section we present laws that relate the *distributions* of the number of customers in the 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 for general, possible non-stationary, systems in the transient domain as well as for stationary systems in steady-state, given that the systems satisfy the following assumptions:

Definition 3 (Distributional laws assumptions).

A.1. The time interval between two successive arrival epochs, $A_i(T_i)$, is independent of $\{A_j(T_j), j < i\}$ conditional on T_i, for all $i = 1, 2, \ldots$. Moreover, new arriving customers do not affect the time in the system of previous customers.

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

A.3. 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.4. Arrival streams from different classes are mutually independent.

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 can be relaxed (see Mourtzinou [24]). 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. Using the notation of section 2, the following systems are examples of overtake free systems:

(a) Multiclass $M(t)/G(t)/1$ 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 $GI/G/s$ 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 [5], Keilson and Servi [18]) (where, once again, we can define we can define "the system" to be either just the queue or the queue together with the server).

4.1. A review of steady-state distributional laws

In this section we first review steady-state distributional laws for single class systems, and then we briefly review distributional laws for multiclass systems with N different customers classes.

The single class steady-state distributional law

Consider a general stationary 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. We use the notation of section 2 and, therefore, denote by $N_a^e(t)$ the number of arrivals up to time t for the corresponding equilibrium renewal process.

We assume that the system is in steady-state and denote by L the number of customers in the system in steady-state and by S the time a customers spends in the system in steady-state, called the system time. Finally, we denote by

$$F_S(t) \overset{\Delta}{=} P\{S \leqslant t\}$$

the distribution function of S and by $G_L(z) \overset{\Delta}{=} E[z^L]$ the generating function of L. The single class steady-state distributional law can be stated as follows:

Theorem 4 (Haji and Newell [13], Bertsimas and Nakazato [7]). For a stationary 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

$$L \overset{\mathrm{d}}{=} N_a^e(S) \quad \text{and equivalently} \quad G_L(z) = \int_0^\infty K_e(z,t)\,\mathrm{d}F_S(t), \tag{10}$$

where

$$K_e(z,t) \overset{\Delta}{=} E\big[z^{N_a^e(t)}\big] = \sum_{n=0}^{\infty} z^n P\big\{N_a^e(t) = n\big\}$$

is the generating function of $N_a^e(t)$.

Intuitively, (10) 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* stationary 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 $N^o_{a_i}(t)$, $N^e_{a_i}(t)$ be the number of customers up to time t for the ordinary and equilibrium renewal process of the ith class, respectively. Given that they exist in steady-state, let 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

$$L \triangleq \sum_{i=1}^{N} L_i, \qquad F_{S_i}(t) \triangleq P\{S_i \leqslant t\}$$

and

$$G_{L_1,\dots,L_N}(z_1, \dots, z_N) \triangleq E\bigl[z_1^{L_1} \cdots z_N^{L_N}\bigr].$$

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

Theorem 5 (Bertsimas and Mourtzinou [4]). 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:

$$\begin{aligned} &G_{L_1,\dots,L_N}(z_1, \dots, z_N) \\ &\quad = 1 + \sum_{i=1}^{N} \int_0^\infty \int_0^t \prod_{\substack{j=1 \\ j \neq i}}^{N} K_{e,j}(z_j, x) \frac{\partial}{\partial x} K_{e,i}(z_i, x)\, \mathrm{d}F_{S_i}(t), \end{aligned} \tag{11}$$

where

$$K_{e,i}(z_i, t) \triangleq E\bigl[z_i^{N^e_{a_i}(t)}\bigr] = \sum_{n=0}^{\infty} z_i^n P\bigl\{N^e_{a_i}(t) = n\bigr\}$$

is the generating function of $N^e_{a_i}(t)$.

Note that for each individual class Theorem 5 yields the single class distributional law of Theorem 4. Moreover, the generating function of the total number of customers in the system, $L \triangleq \sum_{i=1}^N L_i$, can be found if we set $z_1 = \cdots = z_N = z$ in (11).

4.2. *Transient single class distributional laws*

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

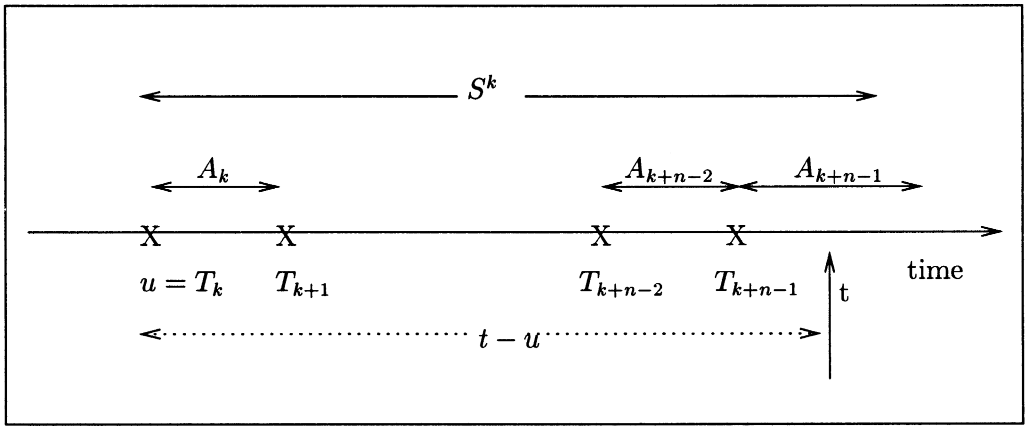


Figure 1. A scenario for a single class system in the transient regime.

We use the notation introduced in section 2. For ease of the presentation, we first prove the transient single class distributional law assuming that the system starts with $L(0) = 0$ with probability 1 (w.p.1) and $V(0) = \widehat{V}(0)$; we then state the more general theorem that accounts for an arbitrary distribution of $L(0)$ and $V(0)$.

A transient law between $L(t)$ and $S(t)$ when the system starts with no initial customers

Theorem 6. For a queueing system that satisfies Assumptions A.1–A.3 and starts with $L(0) = 0$ w.p.1, and $V(0) = \widehat{V}(0)$, the transient number in the system $L(t)$ and the transient system time $S(t)$ are related as follows:

$$G_L(z,t) = 1 + (z-1)\int_0^t h(u)P\{S(u) > t-u\}K_o(z,u,t)\,\mathrm{d}u, \tag{12}$$

where

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

is the generating function of $N_a^o(u,t)$.

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 ($n \geqslant 1$) 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.

This observation is based on Assumptions A.3 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 $\{L(t) \geqslant n\}$ is equivalent to the intersection of the following events:

E_1: the nth most recently arrived customer with respect to t arrives at time $(u-\mathrm{d}u, u]$,

E_2: his system time, is greater than $t-u$, for all $u \in (0,t]$.

We can further decompose event E_1 into the event of an arrival at time $(u-\mathrm{d}u, u]$ (that occurs with probability $h(u)\,\mathrm{d}u$) and the event of $n-1$ arrivals in $(u,t]$ given an arrival at $(u-\mathrm{d}u, u]$ (that occurs with probability $P\{N_a^o(u,t) = n-1\}$). Furthermore, the probability of event E_2 is $P\{S(u) > t-u\}$. Finally, according to Assumption A.1, $S(u)$ is independent of the path of the arrival process after time u and therefore events E_1 and E_2 are independent. The previous discussion leads to the relationship for $n \geqslant 1$:

$$P\bigl\{L(t) \geqslant n\bigr\} = \int_0^t h(u) P\bigl\{S(u) > t-u\bigr\} P\bigl\{N_a^o(u,t) = n-1\bigr\}\,\mathrm{d}u. \tag{13}$$

Given that $P\{L(t) \geqslant 0\} = 1$ and that $P\{L(t) = n\} = P\{L(t) \geqslant n\} - P\{L(t) \geqslant n+1\}$ we can easily calculate the generating function $G_L(z,t)$ to obtain (12). □

Note that although $\widehat{V}(0)$ is not explicitly present in (12), it does influence both $L(t)$ and $S(t)$, as it will become apparent in the sequel.

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

We, now, generalize the distributional law of Theorem 6 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 $V(0) = \widehat{V}(0) + X_1 + \cdots + X_k$, where $\widehat{V}(0)$ is the set-up work and X_i is the service time requirement of the ith initial customer.

Theorem 7. For a queueing system that satisfies Assumptions A.1–A.3, Assumption B and starts with $L(0) = k$ w.p.1 and $V(0) = \widehat{V}(0) + X_1 + \cdots + X_k$, the transient number of customers in the system, $L(t)$, and the transient system time $S(t)$ are related as follows:

$$\begin{aligned} G_L(z,t) = {} & I^{(k)}(z,t) + P\bigl\{V(0) < t\bigr\} \\ & \times \left[1 + (z-1)\int_0^t h(u) P\bigl\{S(u) > t-u \mid V(0) < t\bigr\} K_o(z,u,t)\,\mathrm{d}u\right], \end{aligned} \tag{14}$$

where

$$I^{(k)}(z,t) \triangleq K(z,t)\left[z^k P\{V_k \geqslant t\} + \sum_{i=1}^{k-1} z^i \bigl[P\{V_i \geqslant t\} - P\{V_{i+1} \geqslant t\}\bigr]\right],$$

with $V_i \overset{\Delta}{=} \widehat{V}(0) + X_k + \cdots + X_i,\ i = 1, \ldots, k$, and also

$$K_o(z,u,t) \overset{\Delta}{=} E\big[z^{N_a^o(u,t)}\big] = \sum_{n=0}^{\infty} z^n P\big\{N_a^o(u,t) = n\big\},$$

$$K(z,t) \overset{\Delta}{=} E\big[z^{N_a(t)}\big] = \sum_{n=0}^{\infty} z^n P\big\{N_a(t) = n\big\}.$$

Proof. Let $M(t)$ be the number of initial customers present in the system at time t, $M(t) \in \{1, \ldots, k\}$. Let also $V_i \overset{\Delta}{=} \widehat{V}(0) + X_k + \cdots + X_i,\ i = k, \ldots, 1$. Then,

$$P\big\{M(t) = 0\big\} = P\big\{V(0) < t\big\} \quad \text{and} \quad P\big\{M(t) = k\big\} = P\{V_k \geqslant t\}$$

and

$$P\big\{M(t) = i\big\} = P\{V_i \geqslant t\} - P\{V_{i+1} \geqslant t\}, \quad \text{for } i = 1, \ldots, k.$$

Let us define

$$G_{L^i}(z,t) \overset{\Delta}{=} E\big[z^{L(t)} \mid M(t) = i\big],$$

then

$$G_L(z,t) = P\big\{M(t) = 0\big\} G_{L^0}(z,t) + \sum_{i=1}^{k} P\big\{M(t) = i\big\} G_{L^i}(z,t). \tag{15}$$

In the special case where $M(t) = 0$ the analysis of Theorem 6 holds, i.e., in order to have at least n ($n \geqslant 1$) 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,

$$G_{L^0}(t) = 1 + (z-1) \int_0^t h(u) P\big\{S(u) > t - u \mid V(0) < t\big\} K_o(z,u,t)\, \mathrm{d}u. \tag{16}$$

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

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

Therefore, if we define

$$K(z,t) \overset{\Delta}{=} E\big[z^{N_a(t)}\big] = \sum_{n=0}^{\infty} z^n P\big\{N_a(t) = n\big\}$$

we have that

$$G_{L^i}(z,t) = \sum_{n=i}^{\infty} z^n P\big\{N_a(t) = n - i\big\} = z^i K(z,t). \tag{17}$$

Combining, (15), (16) and (17) we complete the proof. □

4.3. *Transient multiclass distributional law*

We, next, 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 $(u-\mathrm{d}u,u]$, and $h_i(t)\Delta t$ (as $\Delta t \to 0$) be the probability of a class i arrival in $(t-\Delta t,t]$. Furthermore, let $S_i(t)$ be the time spent in the system for class i customers that arrived at $(t-\mathrm{d}t,t]$ and let $L_i(t)$ be the number of class i customers in the system as observed at time t. Finally let

$$L(t) \overset{\Delta}{=} \sum_{i=1}^{N} L_i(t), \qquad \vec{z} \overset{\Delta}{=} (z_1,\ldots,z_N)$$

and

$$G_{L_1,\ldots,L_N}(\vec{z},t) \overset{\Delta}{=} E\bigl[z_1^{L_1(t)}\cdots z_N^{L_N(t)}\bigr].$$

Assuming that the system starts empty we have:

Theorem 8. For a queueing system that satisfies Assumptions A.1–A.4 and starts empty, we have that

$$\begin{aligned} &G_{L_1,\ldots,L_N}(\vec{z},t) \\ &\quad = 1-\sum_{j=1}^{N}\int_0^t \frac{\partial}{\partial a}K_{e,j}(z_j,a,t)\prod_{\substack{i=1\\ i\neq j}}^{N} K_{e,i}(z_i,a,t)P\bigl\{S_j(a)>t-a\bigr\}\,\mathrm{d}a, \end{aligned} \tag{18}$$

where

$$K_{o,i}(z_i,u,t) \overset{\Delta}{=} E\bigl[z^{N^o_{a_i}(u,t)}\bigr] = \sum_{n=0}^{\infty} z_i^n P\bigl\{N^o_{a_i}(u,t)=n\bigr\}$$

and

$$K_{e,i}(z_i,a,t) \overset{\Delta}{=} 1+(z_i-1)\int_a^t h_i(u)K_{o,i}(z_i,u,t)\,\mathrm{d}u.$$

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 $n_i \geqslant 1$, 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 $\{\bigcap_{i=1}^N (L_i(t) \geqslant n_i)\}$ is equivalent to the intersection of the following events (for all $t_i \in (0,t]$ and for all $i=1,\ldots,N$):

$E_{1,i}$: a customer of the ith class arrives at $(t_i-\mathrm{d}t_i,t_i]$,

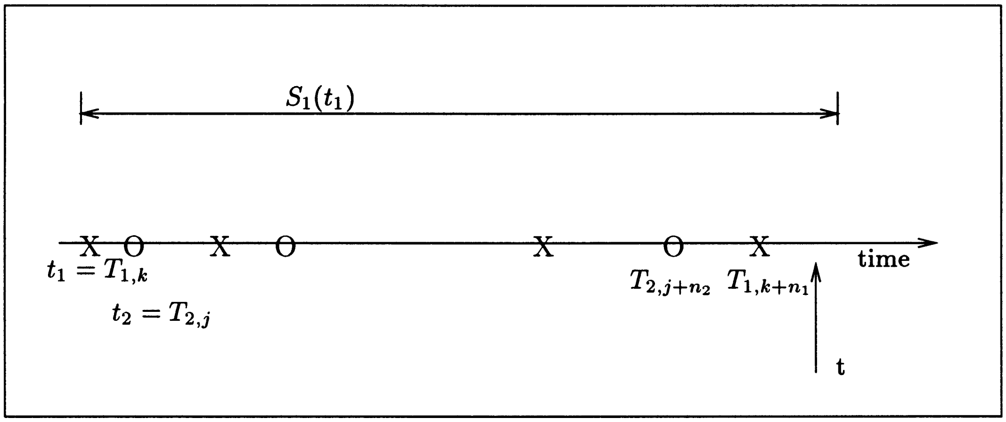


Figure 2. A scenario for a 2-class system in the transient regime.

$E_{2,i}$: the system time of the customer who arrived at $(t_i - \mathrm{d}t_i, t_i]$ is greater than $t - t_i$,

$E_{3,i}$: there are *exactly* $n_i - 1$ arrivals at $(t_i, t]$ given an arrival at $(t_i - \mathrm{d}t_i, t_i]$ for the ith class.

Therefore,

$$P\left\{\bigcap_{i=1}^{N}\big(L_i(t) \geqslant n_i\big)\right\} = \int_{t_1=0}^{t} \cdots \int_{t_N=0}^{t} P\left\{\bigcap_{i=1}^{N} E_{1,i} \bigcap_{i=1}^{N} E_{2,i} \bigcap_{i=1}^{N} E_{3,i}\right\}.$$

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

$$P\left\{\bigcap_{i=1}^{N}\big(L_i(t) \geqslant n_i\big)\right\} = \int_{t_1=0}^{t} \cdots \int_{t_n=0}^{t} P\left\{\bigcap_{i=1}^{N} E_{2,i}\right\} \prod_{i=1}^{N} P\{E_{1,i}\} P\{E_{3,i}\}.$$

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

$$\begin{aligned} &P\left\{\bigcap_{i=1}^{N}\big(L_i(t) \geqslant n_i\big) \cap \text{ (the customer who arrived the first is of class } j)\right\} \\ &= \int_{t_j=0}^{t} \int_{t_1=t_j}^{t} \cdots \int_{t_{j-1}=t_j}^{t} \int_{t_{j+1}=t_j}^{t} \cdots \int_{t_N=t_j}^{t} P\left\{\bigcap_{i=1}^{N} E_{2,i}\right\} \prod_{i=1}^{N} P\{E_{1,i}\} P\{E_{3,i}\}. \end{aligned}$$

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

$$P\left\{\bigcap_{i=1}^{N}\big(L_i(t) \geqslant n_i\big) \cap \text{ (the customer who arrived the first is of class } j)\right\}$$

$$= \int_{t_j=0}^{t} \int_{t_1=t_j}^{t} \cdots \int_{t_N=t_j}^{t} P\left\{ \bigcap_{i=1}^{N} E_{2,i} \mid E_{2,j} \right\} P\{E_{2,j}\} \prod_{i=1}^{N} P\{E_{1,i}\} P\{E_{3,i}\}.$$

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

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

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

$$P\left\{ \bigcap_{i=1}^{N} \bigl(L_i(t) \geqslant n_i\bigr) \cap \text{ (the customer who arrived the first is of class } j) \right\}$$
$$= \int_{t_j=0}^{t} P\{E_{2,j}\} P\{E_{1,j}\} P\{E_{3,j}\} \prod_{\substack{i=1\\ i\neq j}}^{N} \int_{t_i=t_j}^{t} P\{E_{1,i}\} P\{E_{3,i}\}.$$

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

$$\int_{t_i=t_j}^{t} P\{E_{1,i}\} P\{E_{3,i}\} = \int_{t_j}^{t} h_i(t_i) P\bigl\{N_{a_i}^{o}(t_i,t) = n_i - 1\bigr\}\,\mathrm{d}t_i, \quad i \neq j,$$

$$P\{E_{2,j}\} P\{E_{1,j}\} P\{E_{3,j}\} = h_j(t_j)\,\mathrm{d}t_j P\bigl\{S_j(t_j) > t - t_j\bigr\} P\bigl\{N_{a_j}^{o}(t_j,t) = n_j - 1\bigr\},$$

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, and therefore it is distributed as $S_j(t_j)$. Hence,

$$P\left\{ \bigcap_{i=1}^{N} \bigl(L_i(t) \geqslant n_i\bigr) \right\} = \sum_{j=1}^{N} \int_{0}^{t} h_j(t_j) P\bigl\{S_j(t_j) > t - t_j\bigr\}$$
$$\times P\bigl\{N_{a_j}(t_j,t) = n_j - 1\bigr\} \prod_{\substack{i=1\\ i\neq j}}^{N} H_i(t_j,t,n_i)\,\mathrm{d}t_j, \qquad (19)$$

where we define

$$H_i(t_j,t,n_i) \stackrel{\Delta}{=} \int_{t_j}^{t} h_i(t_i) P\bigl\{N_{a_i}^{o}(t_i,t) = n_i - 1\bigr\}\,\mathrm{d}t_i.$$

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

$$P\Big\{\bigcap_{i \notin A} \big(L_i(t) \geqslant n_i\big)\Big\} = \sum_{j \notin A} \int_0^t h_j(t_j) P\big\{S_j(t_j) > t - t_j\big\} \times P\big\{N^o_{a_j}(t_j, t) = n_j - 1\big\} \prod_{\substack{i \neq j \\ i \notin A}} H_i(t_j, t, n_i)\, \mathrm{d}t_j. \tag{20}$$

We now compute $P\{\bigcap_{i=1}^N (L_i(t) = n_i)\}$ iteratively, using (19), (20) and the fact that for $n_i \geqslant 0$

$$\begin{aligned} &P\Bigg\{\bigcap_{k=1}^{i} \big(L_k(t) = n_k\big) \bigcap_{j=i+1}^{N} \big(L_j(t) \geqslant n_j\big)\Bigg\} \\ &\quad = P\Bigg\{\bigcap_{k=1}^{i-1} \big(L_k(t) = n_k\big) \bigcap_{j=i}^{N} \big(L_j(t) \geqslant n_j\big)\Bigg\} \\ &\qquad - P\Bigg\{\bigcap_{k=1}^{i-1} \big(L_k(t) = n_k\big) \cap \big(L_i(t) \geqslant n_i + 1\big) \bigcap_{j=i+1}^{N} \big(L_j(t) \geqslant n_j\big)\Bigg\}. \end{aligned}$$

Having calculated $P\{L_1(t) = n_1, \dots, L_N(t) = n_N\}$, some tedious but straightforward manipulation yields (18). □

Remarks

1. In the case of a single class we define

$$K_e(z, a, t) \overset{\Delta}{=} 1 + (z-1) \int_a^t h(u) K_o(z, u, t)\, \mathrm{d}u,$$

and (18) yields (12). Moreover, the generating function of the total number of customers $L(t)$ in the system can be obtained if we set $z_1 = z_2 = \cdots = z_N$ in (18):

$$G_L(z, t) = 1 - \sum_{j=1}^{N} \int_0^t \frac{\partial}{\partial u} K_{e,j}(z, u, t) \times \prod_{i \neq j}^{N} K_{e,i}(z, u, t) P\big\{S_j(t-u) > u\big\}\, \mathrm{d}u. \tag{21}$$

2. Notice that (18) is the transient counterpart of (11), although in the latter we have performed an integration by parts. We can not perform the same integration in (18) since the distribution function of $S(t)$ depends on t.

3. Notice also that for renewal arrival processes

$$K_{o,i}(z_i, u, t) = K_{o,i}(z_i, t-u) \quad \text{and} \quad K_{e,i}(z_i, u, t) = K_{e,i}(z_i, t-u),$$

   where

$$K_{o,i}(z_i, t-u) \overset{\Delta}{=} E\left[z^{N_a^o(t-u)}\right] \quad \text{and} \quad K_{e,i}(z_i, t-u) \overset{\Delta}{=} E\left[z^{N_a^e(t-u)}\right]$$

   are the generating functions of the number of arrivals from an ordinary and an equilibrium renewal process, respectively.

4. Finally, one can prove multiclass transient distributional under arbitrary initial conditions by combining the proof techniques used in Theorems 6 and 7. The analysis is complicated and therefore it is omitted.

5. Transient performance analysis of specific queueing systems

In this section we apply transient Little's law and the transient distributional laws to derive the transient performance analysis of several systems. The following table summarizes our results:

System	Type of results
$M(t)/G(t)/\infty$	exact formulae
$GI(t)/GI(t)/1$	integral equations
transient $GI/GI/1$	integral equations & asymptotic formulae
transient $GI/D/s$	integral equations & asymptotic formulae
$M(t)/GI(t)/1$	algorithm for exact numerical solutions
$\Sigma GI(t)/GI(t)/1$	integral equations
transient $\Sigma GI/GI/1$	integral equations & asymptotic formulae

5.1. *The $M(t)/GI(t)/\infty$ queueing system*

We start by investigating the transient behavior of the $M(t)/GI(t)/\infty$ queueing system; which often arises in wireless communication 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 [23]). Since 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 9. For a $M(t)/D/\infty$ system with arrival rate $\lambda(t)$ and service time x that starts empty with no initial work, if we define $\Lambda(t) \stackrel{\Delta}{=} \int_0^t \lambda(\tau)\,\mathrm{d}\tau$, we have that

$$G_L(z,t) = \begin{cases} \mathrm{e}^{-(\Lambda(t)-\Lambda(t-x))(1-z)}, & \text{if } t \geqslant x, \\ \mathrm{e}^{-\Lambda(t)(1-z)}, & \text{otherwise.} \end{cases} \tag{22}$$

Proof. From Theorem 6 we have that

$$G_L(z,t) = 1 + (z-1)\int_0^t \lambda(u)P\bigl\{S(u) > t-u\bigr\}K_o(z,u,t)\,\mathrm{d}u, \tag{23}$$

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,

$$\begin{aligned} &\text{if } t < x \quad \text{then } P\bigl\{S(u) > t-u\bigr\} = 1 \quad \text{for } u \in [0,t), \\ &\text{if } t \geqslant x \quad \text{then } P\bigl\{S(u) > t-u\bigr\} = 1 \quad \text{for } u \in [t-x,x), \\ &\phantom{\text{if } t \geqslant x \quad \text{then }} P\bigl\{S(u) > t-u\bigr\} = 0 \quad \text{for } u \in [0,t-x). \end{aligned} \tag{24}$$

On the other hand, since the arrival process is a nonhomogeneous Poisson of rate $\lambda(t)$, we have that $K_o(z,u,t) = \mathrm{e}^{-(\Lambda(t)-\Lambda(u))(1-z)}$. Substituting $K_o(z,u,t)$ and (23), (24) we obtain (22). □

We next consider the $M(t)/GI(t)/\infty$ queueing system and denote by $X(t)$ the service time of a customer entering service at $(t-\mathrm{d}t, t]$, and we prove the following theorem.

Theorem 10. For a $M(t)/GI(t)/\infty$ system that starts empty, we have that the number of customers in the system at time t, $L(t)$, is distributed as a Poisson random variable with rate $\int_0^t \lambda(\tau)P\{X(\tau) > t-\tau\}\,\mathrm{d}\tau$, i.e.,

$$G_L(z,t) = \mathrm{e}^{-(1-z)\int_0^t \lambda(\tau)P\{X(\tau)>t-\tau\}\,\mathrm{d}\tau}. \tag{25}$$

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 nonhomogeneous Poisson process of rate $\lambda(t)p_j(t)$. Therefore, if we denote by $\Lambda_j(t) \stackrel{\Delta}{=} \int_0^t \lambda(\tau)p_j(\tau)\,\mathrm{d}\tau$, we have

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

Moreover as discussed in Ross [29, p. 24], these processes are mutually independent and thus

$$G_L(z,t) = \prod_{j=1}^{k} G_{L_j}(z,t) = \mathrm{e}^{-(1-z)\sum_{j=1}^{k} \Lambda_j(t)}\, \mathrm{e}^{(1-z)\sum_{j:\ x_j \leqslant t} \Lambda_j(t-x_j)}. \qquad (26)$$

Using simple algebraic manipulations on the exponents we obtain (25) for this case. Since any general distribution is the limit of a sequence of mixtures of deterministic distributions, (25) holds in general. Moreover, the generating function $G_L(z,t)$ in (25) for every time t corresponds to a Poisson random variable of rate $\int_0^t \lambda(\tau)P\{X(\tau) > t-\tau\}\,\mathrm{d}\tau$. □

Notice that one can actually obtain the expected number of customers in the $M(t)/GI(t)/\infty$ system,

$$E\big[L(t)\big] = \int_0^t \lambda(u)P\big\{X(u) > t-u\big\}\,\mathrm{d}u, \qquad (27)$$

directly from the transient form of Little's law, (7), by substituting $h(u) = \lambda(u)$ and $S(u) = X(u)$, since there is no waiting. Furthermore, (27) 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 10 can be traced back to Palm [27], Bartlett [2], Doob [11], Khintchine [19] and Prékopa [28], all before 1958. For a recent reference on Theorem 10 and its extension to networks of infinite server queues with non-stationary Poisson input see Massey and Whitt [23].

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

In this section we consider a single server system and use the notation of section 2. We assume that the system starts empty with initial work $V(0)$ equal the set-up work $\widehat{V}(0)$, and that customers are served according to a FIFO discipline. We furthermore impose a stronger version of Assumption B, namely:

Assumption C. The system clears first the initial work $V(0)$, and then starts working on the customers who arrive after time 0. Moreover, the server never idles as long as there is work in the system.

We denote by $X(t)$ the time a customer who enters service at $(t-\mathrm{d}t, t]$, spends in service. 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 $(t-\mathrm{d}t, t]$ spends waiting *in the queue* and by $S(t)$ the time that a customer who arrived at $(t-\mathrm{d}t, t]$ spends *in the system*. 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 derive another relationship between $L(t)$ and $Q(t)$, which in contrast with the laws presented in section 4, 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 11. For a $GI(t)/GI(t)/1$ queueing system with FIFO, that starts with $L(0) = 0$ w.p.1, initial work $V(0)$, and satisfies Assumptions A.1–A.3 and Assumption C, the transient quantities $L(t)$ and $Q(t)$ are related as follows:

$$G_L(z,t) = (1-z)\mathit{idle}(t) + (1-z)P\{V(0) \geqslant t\}K(z,t) + zG_Q(z,t), \tag{28}$$

where *idle*$(t) \overset{\Delta}{=} P\{$the system is empty at time $t\}$ and

$$K(z,t) \overset{\Delta}{=} E\bigl[z^{N_a(t)}\bigr] = \sum_{n=0}^{\infty} z^n P\bigl\{N_a(t) = n\bigr\}.$$

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. The server is working on the initial work, $V(0)$ (with probability $P\{V(0) \geqslant t\}$).
3. It is busy servicing customers (with probability $1 - P\{V(0) \geqslant t\} - \mathit{idle}(t)$).

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, $Q(t) = L(t) = N_a(t)$, as in this case, all the customers that arrived to the system up to time t are still waiting for the server to finish the initial set-up work, $V(0)$. 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 $L(t)$, $G_Q(z,t) \overset{\Delta}{=} E[z^{Q(t)}]$ and $G_L(z,t) \overset{\Delta}{=} E[z^{L(t)}]$ as follows:

$$\begin{aligned} G_Q(z,t) &= \mathit{idle}(t) + K(z,t)P\bigl\{V(0) \geqslant t\bigr\} + \bigl(1 - \mathit{idle}(t) - P\bigl\{V(0) \geqslant t\bigr\}\bigr)G_{Q_B}(z,t), \\ G_L(z,t) &= \mathit{idle}(t) + K(z,t)P\bigl\{V(0) \geqslant t\bigr\} + z\bigl(1 - \mathit{idle}(t) - P\bigl\{V(0) \geqslant t\bigr\}\bigr)G_{Q_B}(z,t), \end{aligned}$$

where $G_{Q_B}(z,t) \overset{\Delta}{=} E[z^{Q(t)} \mid$ the server is servicing customers]. Combining the last two relations we obtain (28). □

The above proposition together with the transient distributional laws of section 4 leads to a complete description of the $GI(t)/GI(t)/1$ system as a function of the emptiness function, *idle*(t), as the following theorem demonstrates.

Theorem 12. For a $GI(t)/GI(t)/1$ system under FIFO that starts with $L(0) = 0$ and initial work $V(0)$ and satisfies Assumptions A.1–A.3 and C, the probability distribution function of the waiting time of a customer who arrived to the system at $(t_o - \mathrm{d}t_o, t_o]$, $F_{W(t_o)}(x) \overset{\Delta}{=} P\{W(t_o) \leqslant x\}$, satisfies the following integral equation

$$\begin{aligned}\int_0^t h(u)K_o(z,u,t)\biggl[\int_0^\infty \mathrm{d}F_{W(u)}(a)P\bigl\{X(u+a) > t-u-a\bigr\} \\ - zP\bigl\{W(u) > t-u\bigr\}\biggr]\mathrm{d}u \\ = 1 - \mathit{idle}(t) - P\bigl\{V(0) \geqslant t\bigr\}K(z,t), \qquad (29)\end{aligned}$$

where

$$\mathit{idle}(t) \overset{\Delta}{=} P\{\text{the server is idle at time } t\},$$

$\mathrm{d}F_{W(u)}(\cdot)$ is the pdf of $W(u)$, $K_o(z,u,t) \overset{\Delta}{=} E[z^{N_a^o(u,t)}]$ and $K(z,t) \overset{\Delta}{=} E[z^{N_a(t)}]$.

Proof. Notice that Theorem 7 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,

$$G_Q(z,t) = 1 + (z-1)\int_0^t h(u)P\bigl\{W(u) > t-u\bigr\}K_o(z,u,t)\,\mathrm{d}u, \qquad (30)$$

$$G_L(z,t) = 1 + (z-1)\int_0^t h(u)P\bigl\{S(u) > t-u\bigr\}K_o(z,u,t)\,\mathrm{d}u. \qquad (31)$$

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

$$P\bigl\{S(t) > x\bigr\} = \int_{a=0}^\infty P\bigl\{a \leqslant W(t) \leqslant a + \mathrm{d}a\bigr\}P\bigl\{X(t+a) > x-a\bigr\}, \qquad (32)$$

so that from (31) we get

$$G_L(z,t) = 1 + (z-1)\int_0^t h(u)\int_0^\infty \mathrm{d}F_{W(u)}(a)P\bigl\{X(u+a) > t-u-a\bigr\}K_o(z,u,t)\,\mathrm{d}u.$$

Combining the last equation with (28) we complete the proof. □

By solving eq. (29) and then using (32) we also obtain the pdf of $S(t)$ as a function of *idle*(t). Moreover, using the distributional laws of Theorem 6 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 a similar analysis, see Mourtzinou [24]. However, solving the equation of Theorem 12 for the general $GI(t)/GI(t)/1$ case is quite complicated and therefore in the sequel we consider two special cases: the $M(t)/GI(t)/1$ and the $GI/GI/1$ queue. In 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.

5.3. The $M(t)/GI(t)/1$ queueing system under FIFO

In this section we analyze single server systems with nonhomogeneous Poisson arrivals and general time-dependent service time distributions as defined in section 2, that satisfy the following set of assumptions

Assumption D

D.1. There exists a set of ordered time epochs, $ta_i,\ i = 0, 1, 2, \ldots$, with $ta_0 \overset{\Delta}{=} 0$, such that the arrival rate $\lambda(t)$ is piecewise constant with value $\lambda(t) = \lambda_i$ for $t \in [ta_i, ta_{i+1})$.

D.2. There exists a set of ordered time epochs, $ts_i,\ i = 0, 1, 2, \ldots$, with $ts_0 \overset{\Delta}{=} 0$, 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 $T \overset{\Delta}{=} \{ta_i,\ i \in \mathbb{Z}_+\} \cup \{ts_i,\ i \in \mathbb{Z}_+\}$ and let the set $O \overset{\Delta}{=} \{0, t_1, t_2, \ldots\}$ be the ordering of the elements of T such that $t_i \leqslant t_j$ for $i \leqslant j$.

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

$$\Phi_{W_0}(w, s) = \frac{(w/\eta_0)\phi_{V(0)}(\eta_0) - \phi_{V(0)}(w)}{\lambda_0 \phi_{X_0}(w) - \lambda_0 - s + w}, \tag{33}$$

where $\phi_{V(0)}(w)$ is the Laplace transform of the initial work at $t = 0$, $V(0)$ and $\eta_0 \overset{\Delta}{=} \eta_0(s)$ 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.,

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

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

$$\Phi_{W_i}(w, s) = \frac{(w/\eta_i)\phi_{W_{i-1}}(\eta_i, t_i) - \phi_{W_{i-1}}(w, t_i)}{\lambda_k \phi_{X_m}(w) - \lambda_k - s + w}, \tag{34}$$

where $\eta_i \overset{\Delta}{=} \eta_i(s)$ 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 [24]).

Theorem 13. For an $M(t)/GI(t)/1$ queueing system under FIFO that satisfies Assumptions C, D and starts with an arbitrary initial work $V(0)$ we can evaluate the Laplace transform of the distribution of $W(t)$ as follows:

$$\phi_W(w,t) = \phi_{W_i}(w, t - t_i) \quad \text{for } t \in [t_i, t_{i+1}), \tag{35}$$

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

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

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

where $[t_i, t_{i+1}) \overset{\Delta}{=} [ta_k, ta_{k+1}) \cap [ts_m, ts_{m+1})$.

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 D.

Independently, Choudhury et al. in [8] used a very similar approach to obtain the performance of the $M(t)/GI(t)/1$ queue under Assumptions D. 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 [31]). 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. [8].

5.4. *Transient analysis of $GI/GI/1$ queueing system under FIFO*

In this section we focus on an important class of systems where customers arrive according to a single *equilibrium* renewal arrival process, have general service requirements and the single server has a constant rate.

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 $u \geqslant 0$, where λ is the arrival rate, and also $K(z,t) \overset{\Delta}{=} E[z^{N_a(t)}] = K_e(z,t)$. We also define by $\alpha(s)$ the Laplace transform of the interarrival times.

Since the $GI/GI/1$ queueing system is just a special case of the $GI(t)/GI(t)/1$ system, Theorem 12 still holds and the integral equation takes the following form:

$$\begin{aligned} &\lambda \int_0^t K_o(z, t-u)\big(P\{W(u) + X > t - u\} - zP\{W(u) > t - u\}\big)\,\mathrm{d}u \\ &\quad = 1 - \mathit{idle}(t) - P\{V(0) > t\}K_e(z,t). \end{aligned} \tag{36}$$

The integral equation (36) 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 $t \gg t_o$ 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 $\mathcal{L}_{E[L]}(s)$ and $\mathcal{L}_{E[Q]}(s)$ to be the Laplace transform of $E[L(t)]$ and $E[Q(t)]$, respectively, i.e.,

$$\mathcal{L}_{E[L]}(s) \stackrel{\Delta}{=} \int_0^\infty \mathrm{e}^{-st} E\big[L(t)\big]\,\mathrm{d}t \quad \text{and} \quad \mathcal{L}_{E[Q]}(s) \stackrel{\Delta}{=} \int_0^\infty \mathrm{e}^{-st} E\big[Q(t)\big]\,\mathrm{d}t,$$

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.,

$$\phi_W(w,t) \stackrel{\Delta}{=} \int_0^\infty \mathrm{e}^{-wx}\,\mathrm{d}F_{W(t)}(x) \quad \text{and} \quad \Phi_W(w,s) \stackrel{\Delta}{=} \int_0^\infty \mathrm{e}^{-st}\phi_W(w,t)\,\mathrm{d}t.$$

Similarly, $S(t)$ has Laplace transform $\phi_S(w,t)$ and double Laplace transform $\Phi_S(w,s)$.

Theorem 14. For a $GI/GI/1$ system that starts empty with initial work $V(0)$, and satisfies Assumption C, the Laplace transform of the expected number of customers in the system and the queue are given by

$$\begin{aligned} \mathcal{L}_{E[Q]}(s) &= \frac{\lambda}{s^2} - \frac{s\mathcal{L}_{\mathit{idle}}(s) - \phi_{V(0)}(s)}{s(\phi_X(s)-1)} \quad \text{and} \\ \mathcal{L}_{E[L]}(s) &= \frac{\lambda}{s^2} - \frac{s\mathcal{L}_{\mathit{idle}}(s) - \phi_{V(0)}(s)}{s(\phi_X(s)-1)}\phi_X(s), \end{aligned} \tag{37}$$

where

$$\mathcal{L}_{\mathit{idle}}(s) \stackrel{\Delta}{=} \int_0^\infty \mathrm{e}^{-st}\mathit{idle}(t)\,\mathrm{d}t$$

is the Laplace transform of *idle*(t) and

$$\phi_{V(0)}(s) \stackrel{\Delta}{=} \int_0^\infty \mathrm{e}^{-st}\,\mathrm{d}P\big\{V(0) \leqslant t\big\}$$

is the Laplace transform of $V(0)$.

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

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

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

$$\mathcal{L}_{E[L]}(s) = \lambda \int_0^\infty \mathrm{e}^{-st} \int_0^t P\bigl\{S(u) > t-u\bigr\}\,\mathrm{d}u\,\mathrm{d}t$$
$$= \lambda \int_0^\infty \mathrm{e}^{-sa} \int_0^\infty \mathrm{e}^{-su} P\bigl\{S(u) > a\bigr\}\,\mathrm{d}u\,\mathrm{d}a,$$

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

$$\mathcal{L}_{E[L]}(s) = \frac{\lambda}{s^2} - \frac{\lambda}{s}\Phi_S(s,s) \quad \text{and} \quad \mathcal{L}_{E[Q]}(s) = \frac{\lambda}{s^2} - \frac{\lambda}{s}\Phi_W(s,s). \tag{38}$$

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

$$\Phi_S(s,s) = \phi_X(s)\Phi_W(s,s). \tag{39}$$

On the other hand we have, from Proposition 11, that for a system that starts empty with initial work $V(0)$,

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

where

$$K_e(z,t) \overset{\Delta}{=} E\bigl[z^{N_a^e(t)}\bigr] = \sum_{n=0}^{\infty} z^n P\bigl\{N_a^e(t) = n\bigr\}.$$

By differentiation we get that

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

and by taking Laplace transforms

$$\mathcal{L}_{E[L]}(s) = -\mathcal{L}_{\mathit{idle}}(s) + \frac{1}{s}\phi_{V(0)}(s) + \mathcal{L}_{E[Q]}(s). \tag{40}$$

Solving the linear system of eqs. (38)–(40), we complete the proof. □

It is important to notice that since the transient form of Little's law holds independently of FIFO, (37) holds independently of the service discipline, as long as Assumption C is satisfied. However, the form of the emptiness function, which is not in general known and can *not* be obtained from the analytic properties of $\mathcal{L}_{E[L]}(s)$ and $\mathcal{L}_{E[Q]}(s)$, changes with the policy and so do $E[Q(t)]$ and $E[L(t)]$. Under the FIFO discipline we will use asymptotics to obtain a closed form expression for $\mathcal{L}_{\mathit{idle}}(s)$, 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:

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

In the sequel we will show, using the asymptotic form of $\mathcal{L}_{idle}(s)$, that under FIFO and as $\rho \to 1$, we obtain for $E[Q]$ exactly the formula we obtained in Bertsimas and Mourtzinou [4].

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

We, next, 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, in the transform domain, where we are dealing with

$$G_Q(z,s) \overset{\Delta}{=} \int_0^\infty \mathrm{e}^{-st} E\bigl[z^{Q(t)}\bigr]\,\mathrm{d}t$$

and

$$\Phi_W(w,s) \overset{\Delta}{=} \int_0^\infty \mathrm{e}^{-st} \int_0^\infty \mathrm{e}^{-wx}\,\mathrm{d}F_{W(t)}(x)\,\mathrm{d}t,$$

we can equivalently define 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, $\mathcal{L}_{idle}(s)$. Then, we also obtain an asymptotic closed form expression for $\mathcal{L}_{idle}(s)$ under heavy traffic conditions, and hence we complete our asymptotic heavy traffic analysis of the $GI/GI/1$ queueing system.

The starting point of our analysis is the following proposition (see Bertsimas and Mourtzinou [5]), where we use the notation that $h(x) \sim r(x)$ as $x \to a$ means that $\lim_{x\to a}(h(x)/r(x)) = 1$.

Proposition 15. Asymptotically, as $t \to \infty$ and $z \to 1$ the kernels $K_e(z,t)$ and $K_o(z,t)$ behave as follows:

$$K_e(z,t) \sim \mathrm{e}^{-tf(z)}, \tag{41}$$

$$K_o(z,t) \sim \bigl[1 - \tfrac{1}{2}(1-z)(c_a^2 - 1) + \mathrm{O}\bigl((1-z)^2\bigr)\bigr]\,\mathrm{e}^{-tf(z)}, \tag{42}$$

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

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

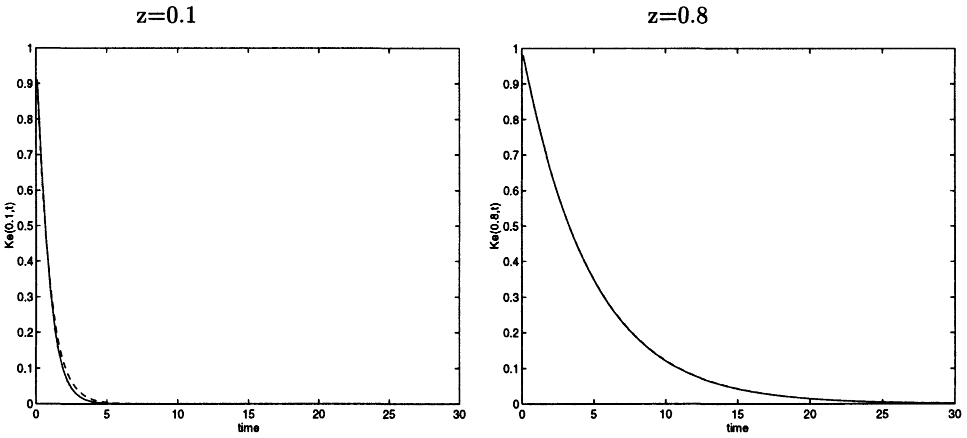


Figure 3. The function $K_e(z,t)$ for Erlang 2 arrivals.

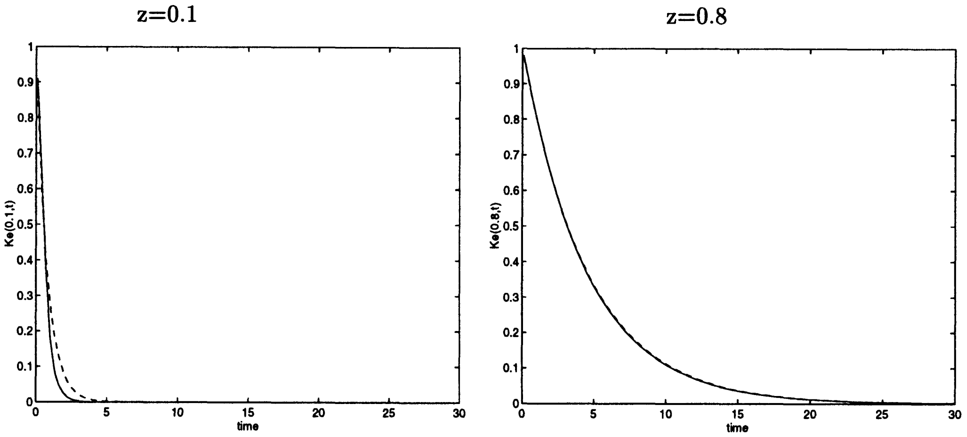
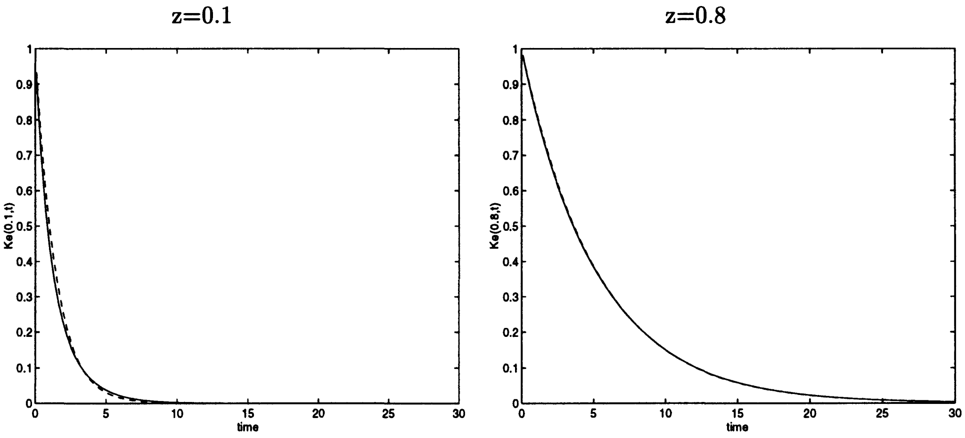


Figure 4. The function $K_e(z,t)$ for Erlang 16 arrivals.

Laplace transform of $K_e(z,t)$ we used the two algorithms in Hosono [15] and in Abate and Whitt [1] which we programmed in Matlab and we got exactly the same results. The results are shown in figures 3–6.

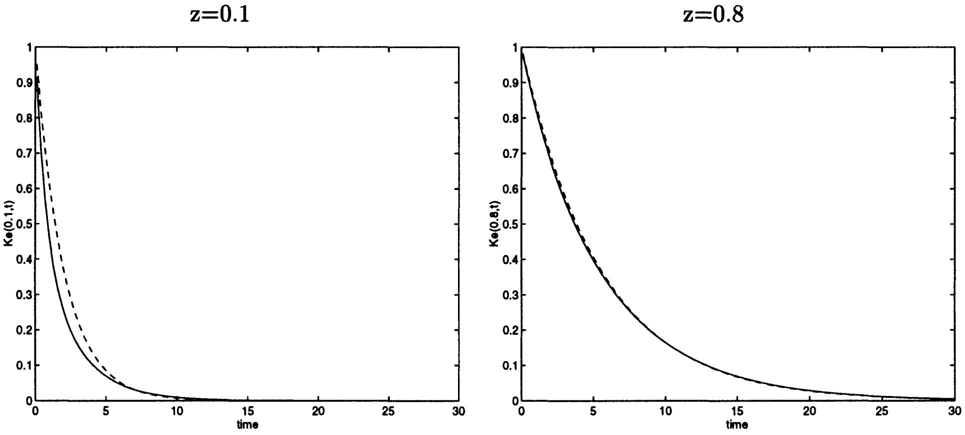
Notice that our expansion is indeed asymptotically exact as $z \to 1$ and $t \to \infty$. Moreover, in all the cases we consider, it is exact for $t > 20$. It is also interesting to notice that our asymptotic expansion is more accurate for values of c_a^2 close to 1 and indeed is exact for Poisson arrivals $c_a^2 = 1$. 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 [24], $-f(z)$ is the root of $1 - z\alpha(s) = 0$ for small values of s and for values of z close


z=0.1

z=0.8

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


z=0.1

z=0.8

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

to 1, in other words

$$1 - z\alpha\bigl(-f(z)\bigr) = 0. \tag{43}$$

Using the above asymptotic results we obtain the following theorem.

Theorem 16. In a $GI/GI/1$ queueing system under FIFO that starts empty with initial work $V(0)$, and satisfies Assumptions A.1–A.3 and C, the distributional laws take the following form, asymptotically in heavy traffic:

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

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

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

$$G_Q(z,s) \sim \frac{1}{s+f(z)}\left(1 + \frac{z-1}{z-\phi_X(s+f(z))}\big((s+f(z))\mathcal{L}_{idle}(s) - \phi_{V(0)}(w)\big)\right), \quad (46)$$

$$\Phi_W(w,s) \sim \frac{w\mathcal{L}_{idle}(s) - \phi_{V(0)}(w)}{1-\alpha(s-w)\phi_X(w)} \frac{1-\alpha(s-w)}{(w-s)}, \quad (47)$$

where $\phi_{V(0)}(s)$ is the Laplace transform of the initial work, $\mathcal{L}_{idle}(s)$ is calculated in Proposition 17 and $\alpha(s)$ is the Laplace transform of the interarrival times.

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

$$\begin{aligned} G_Q(z,s) = {} & \frac{1}{s} + \lambda(z-1) \\ & \times \int_{t=0}^{\infty} \mathrm{e}^{-st}\left[\int_{u=0}^{t} P\{W(u) > t-u\} K_o(z,t-u)\,\mathrm{d}u\right]\mathrm{d}t. \end{aligned} \quad (48)$$

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 [9]) 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

$$G_Q(z,t) \overset{\Delta}{=} E\left[\mathrm{e}^{-Q(t)\log(z)}\right]$$

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.

So, we have to prove that for small values of s and large values of z the RHS of (48) yields the RHS of (44). Notice, now, that in (48) the second term is the Laplace transform of the function

$$\beta(t) \overset{\Delta}{=} \int_{u=0}^{t} P\{W(u) > t-u\} K_o(z,t-u)\,\mathrm{d}u.$$

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 $K_o(z,t-u)$ from (42)

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

Using the above expression in (48) we obtain (44). Similarly, we prove (45).

Next, using a similar type of asymptotic expansions in (36) (see [24] for a more detailed proof) we obtain (47) and (46) and therefore we complete the proof. □

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 16 are exact under any traffic conditions and for any s. In particular, (47) is exact and yields

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

which is the exact transient solution for a $M/G/1$ queue (see Kleinrock [20]).

We can obtain the z-transform of the steady-state queue length, denoted by $G_Q(z)$, if we observe that $G_Q(z) = \lim_{s\to 0} sG_Q(z, s)$. Indeed,

$$G_Q(z) = \lim_{s\to 0} sG_Q(z, s) \sim \frac{z - 1}{z - \phi_X(f(z))} \lim_{s\to 0} s\mathcal{L}_{\mathit{idle}}(s) = \frac{(1 - \rho)(z - 1)}{z - \phi_X(f(z))}.$$

This is exactly the result obtained in Bertsimas and Mourtzinou [5].

To completely characterize the asymptotic behavior of the system we also need to obtain an asymptotic closed form expression for $\mathcal{L}_{\mathit{idle}}(s)$.

Proposition 17. In a $GI/GI/1$ queue with initial work $V(0)$, that is operating under FIFO and satisfies Assumptions A.1–A.3 and C, asymptotically in heavy traffic, the Laplace transform of the emptiness function for $\rho < 1$ is given as

$$\mathit{idle}(s) \sim \frac{\phi_{V(0)}(w_2)}{w_2} \quad \text{with } w_2 = \frac{-p_1(s) - \sqrt{(p_1(s))^2 - 4p_0(s)p_2(s)}}{2p_2(s)}, \tag{49}$$

where

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

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

$$\begin{aligned} p_2(s) \stackrel{\Delta}{=} {} & -\frac{1}{2}\left(1 - \frac{s}{\lambda} + \frac{(c_a^2 + 1)s^2}{2\lambda^2}\right)\left(c_x^2 + 1\right)\left(E[X]\right)^2 \\ & + \left(\frac{1}{\lambda} - \frac{(c_a^2 + 1)s}{\lambda^2}\right)E[X] - \frac{c_a^2 + 1}{2\lambda^2}. \end{aligned}$$

Proof. See Appendix. □

Using the asymptotic expression for $\mathcal{L}_{\mathit{idle}}(s)$ we rewrite $\Phi_W(w, s)$ as follows:

$$\begin{aligned} \Phi_W(w, s) &\sim \frac{(\phi_{V(0)}(w_2))/w_2 - \phi_{V(0)}(w)}{p_2(s)(w - w_1)(w - w_2)} \cdot \frac{1 - \alpha(s - w)}{(w - s)} \\ \text{with } w_2 &= \frac{-p_1(s) - \sqrt{(p_1(s))^2 - 4p_0(s)p_2(s)}}{2p_2(s)}, \end{aligned} \tag{50}$$

where $p_0(s)$, $p_1(s)$ and $p_2(s)$ are defined in Proposition 17.

On the other hand, using Brownian approximations for general arrival we get (see, for example, Kleinrock [20])

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

$$\text{with } \widehat{w}_{1,2} = \frac{-\lambda(1-\rho)}{\rho^2(c_x^2+c_a^2)}\left[1 \mp \sqrt{1+2s\rho^2\frac{(c_x^2+c_a^2)}{\lambda(1-\rho)^2}}\right], \tag{51}$$

which is different from (50).

Using the asymptotic form of $\mathcal{L}_{idle}(s)$ from [24] we also obtain an asymptotic form of the Laplace transform of the expected queue length, $\mathcal{L}_{E[Q]}(s)$ via Theorem 14. Indeed,

$$\mathcal{L}_{E[Q]}(s) = \frac{\lambda}{s^2} - \frac{s\mathcal{L}_{idle}(s) - \phi_{V(0)}(s)}{s(\phi_X(s)-1)} \sim \frac{\lambda}{s^2} - \frac{s\phi_{V(0)}(w_2) - w_2\phi_{V(0)}(s)}{w_2 s(\phi_X(s)-1)}. \tag{52}$$

It is interesting to note that if we calculate the asymptotic steady-state queue length, using the asymptotic value of w_2 we obtain the same result we obtained in Bertsimas and Mourtzinou [5].

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 $\mathcal{L}_{E[W]}(s)$ its Laplace transform, i.e.,

$$\mathcal{L}_{E[W]}(s) \stackrel{\Delta}{=} \int_0^\infty \mathrm{e}^{-st}E[W(t)]\,\mathrm{d}t,$$

we have from the properties of Laplace transform that

$$\mathcal{L}_{E[W]}(s) = \lim_{w\to 0}\frac{\partial}{\partial w}\Phi_W(w,s).$$

Hence, we can prove the following corollary of Theorem 16.

Corollary 18. In a $GI/G/1$ queue with FIFO service policy, and initial work $V(0)$, under Assumptions A.1–A.3 and C, asymptotically in heavy traffic:

$$\mathcal{L}_{E[W]}(s) \sim \frac{\mathcal{L}_{idle}(s)}{s} + \frac{E[V(0)]}{s} + \frac{E[X]\alpha(s)}{s(1-\alpha(s))} - \frac{1}{s^2}, \tag{53}$$

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

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

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

the same result we obtained in Bertsimas and Mourtzinou [5].

It is important to notice that although Theorem 14 holds for any traffic intensity and for any s, Corollary 18 only holds asymptotically in heavy traffic.

Numerical results

In order to obtain a better understanding of the asymptotic method and to quantify the range of its validity 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 [15] 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 $t \gg t_o$, 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 7.

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 8. 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 9 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 $t > 20$.

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 9, 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 10, where use the exact results presented in Odoni and Roth [25] 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

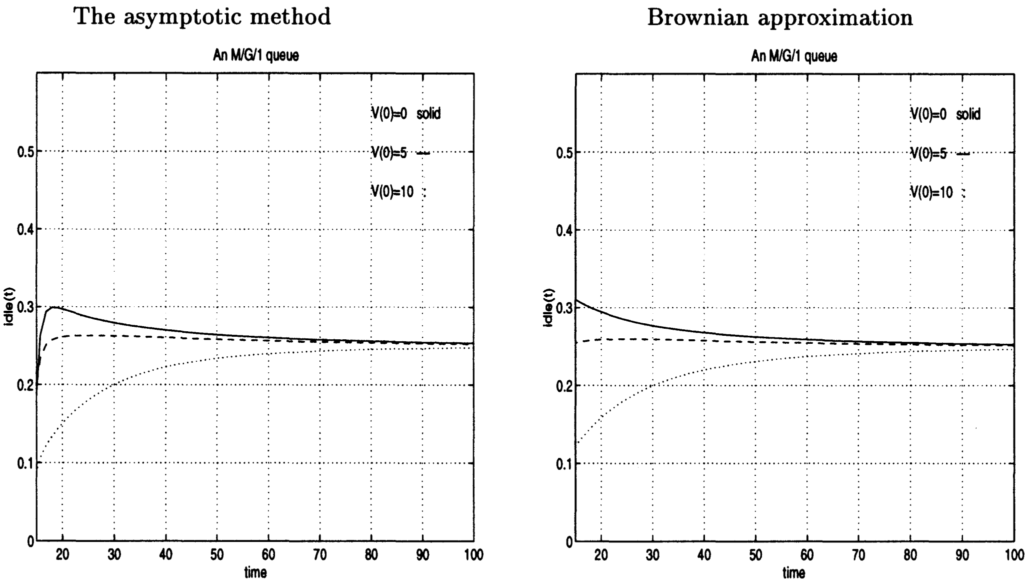


Figure 7. The function *idle*(t) for an $M/G/1$ queue.

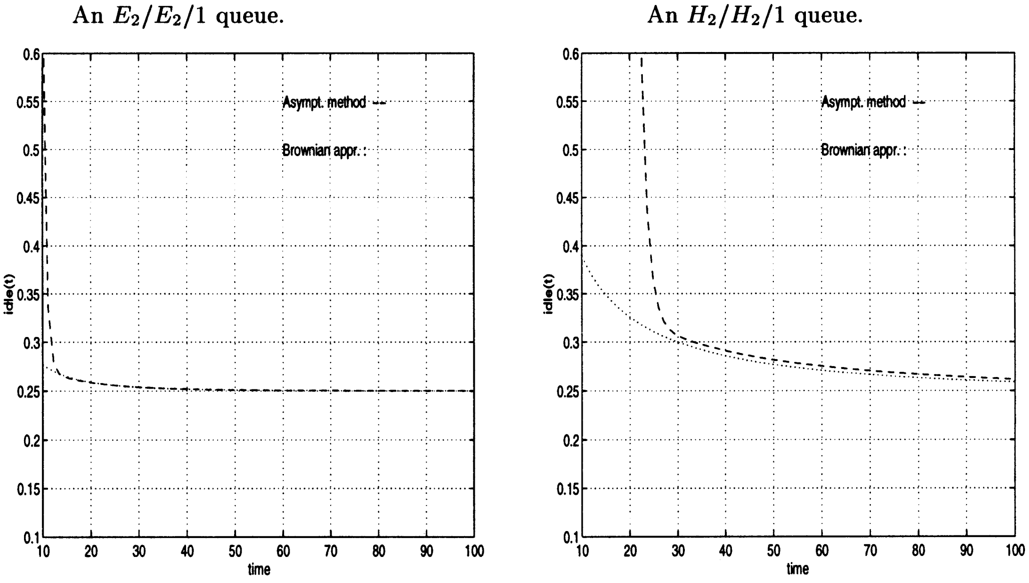


Figure 8. The function *idle*(t) for an $GI/GI/1$ queue with $V(0) = 0$.

same server. Therefore, as it is well known (see Iversen [16]), 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.

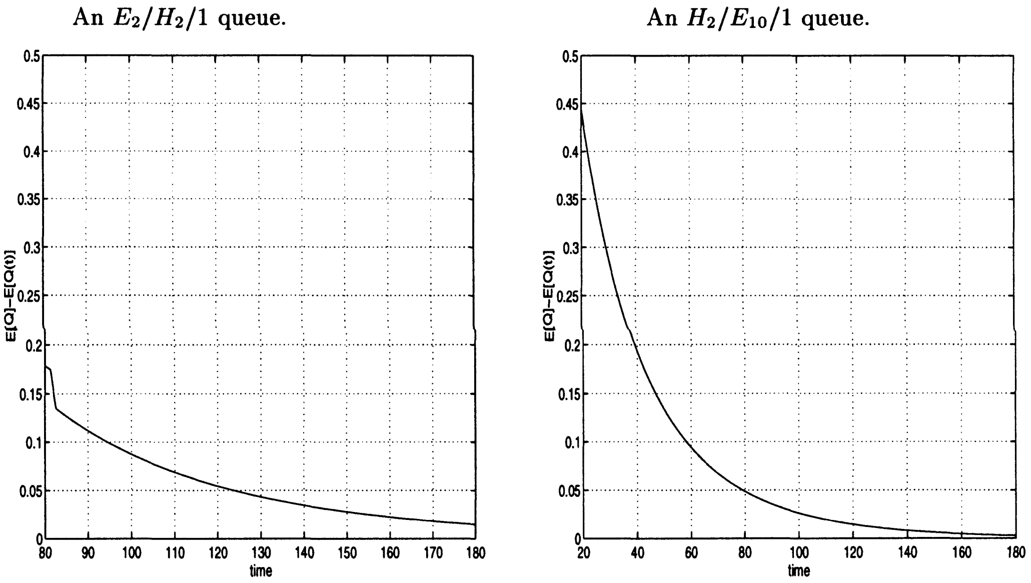


Figure 9. The function $E[Q] - E[Q(t)]$ for an $GI/GI/1$ queue.

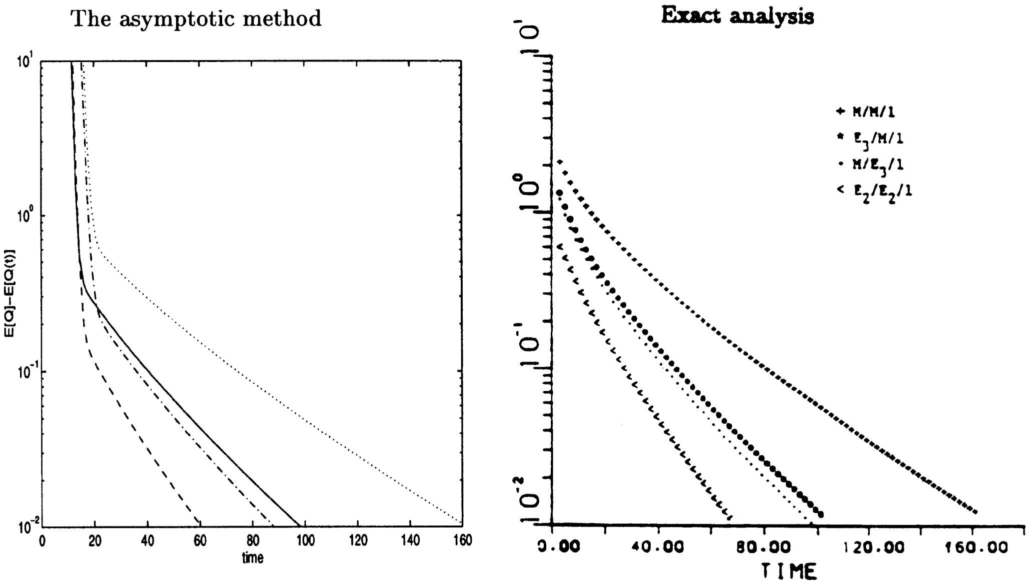


Figure 10. A semilogarithmic plot of $E[Q] - E[Q(t)] - \rho = 0.75,\ E[X] = 1$.

5.5. *The $\Sigma GI(t)/GI(t)/1$ 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 $(t-\mathrm{d}t,t]$ and $\mathrm{d}F_{W_i(t)}(\cdot)$ is the pdf of $W_i(t)$. Similarly to section 2, we denote by $X_i(t)$ the service time of a class i customer that *enters the server at* $(t-\mathrm{d}t,t]$. Finally, we denote by $\vec{z} \stackrel{\Delta}{=} (z_1,\dots,z_N)$ 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 $L_1(t),\dots,L_N(t)$ (resp. $Q_1(t),\dots,Q_N(t)$).

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 (see [24] for a proof).

Proposition 19. In a $\Sigma GI(t)/GI(t)/1$ system with N-classes of customers that satisfies Assumptions A.1–A.4:

$$G_{L_i}(z,t) = zG_{Q_i}(z,t) + (1-z)\left[\mathit{idle}(t) + \sum_{\substack{j=1\\ j\neq i}}^{N} \int_0^t h_j(a)K_{e,i}(z,a,t)M_j(a,t)\,\mathrm{d}a\right], \tag{54}$$

where *idle*(t) is the emptiness process,

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

and

$$K_{e,i}(z_i,a,t) \stackrel{\Delta}{=} 1 + (z_i-1)\int_a^t h_i(u)K_{o,i}(z_i,u,t)\,\mathrm{d}u,$$

and

$$M_i(a,t) \stackrel{\Delta}{=} P\{S_i(a) > t-a \geqslant W_i(a)\}.$$

Using Proposition 19 together with the multiclass transient distributional and the fact that for all $i=1,\dots,N$

$$P\{S_i(t) > x\} = P\{a \leqslant W_i(t) \leqslant a+\mathrm{d}a\}P\{X_i(t+a) > x-a\},$$

we obtain a system on N integral equations on N unknowns, the cdf of $W_i(t)$ for $i = 1,\dots,N$. 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. For the $\Sigma GI/G/1$ queueing system, under heavy traffic conditions we use asymptotic expansions to obtain the following theorem (see [24] for a proof).

Theorem 20. In a $\Sigma GI/GI/1$ system under FIFO that starts empty the Laplace transforms of the individual queue lengths asymptotically under heavy traffic conditions are given by

$$G_{Q_i}(z,s) \sim \frac{1}{s+f_i(z)}\left[1+\frac{f_i(z)C(z,s)(z-1)}{z-\phi_{X_i}(s+f_i(z))+(z-1)\rho_i\phi_{X_i^*}(s+f_i(z))}\right], \quad (55)$$

where

$$f_i(s) = \lambda_i(1-z) - \frac{(1-z)^2(c_{a_i}^2-1)}{2}, \qquad C(z,s) \triangleq \frac{\mathit{idle}(s)}{1-D(z,s)}$$

and

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

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 in 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 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.

Acknowledgements

We would like to thank the reviewer of the paper for his/her detailed and thoughtful report; the paper benefited substantially from the reviewer’s comments.

Appendix

In this Appendix we give a proof of Proposition 17: 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

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

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

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

Hence, we have that

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

so that $(1-\alpha(s-w))/(w-s)$ 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

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

is analytic in the same region. Expanding the denominator around $w = 0$ and $s = 0$, we get

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

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

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

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

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

Equivalently, we have that

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

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

$$p_1(s) \approx E[X] - \frac{1}{\lambda} = \frac{1}{\lambda}(\rho - 1) < 0$$

and

$$p_2(s) \approx \frac{E[X]}{\lambda} - \frac{1}{2\lambda^2}\big(\rho^2(c_x^2 + 1) + c_a^2 + 1\big) < 0.$$

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 (49).

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. Ser. B 11 (1949) 211–229.

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

[4] D. Bertsimas and G. Mourtzinou, 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 Oper. Res.

[5] D. Bertsimas and G. Mourtzinou, A unified method to analyze overtake-free queueing systems, Adv. Appl. Probab. 28 (1996) 588–625.

[6] D. Bertsimas and D. Nakazato, Transient and busy period analysis of the $GI/G/1$ queue: The method of stages, Queueing Systems 10 (1992) 153–184.

[7] D. Bertsimas and D. Nakazato, The distributional Little's law and its applications, Oper. Res. 43 (1995) 298–310.

[8] 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).

[9] D.R. Cox, *Renewal Theory* (Chapman and Hall, New York, 1962).

[10] D.R. Cox and V. Isham, *Point Processes* (Chapman and Hall, New York, 1980).

[11] J.L. Doob, *Stochastic Processes* (John Wiley, New York, 1953).

[12] P. Green, L. Kolesar and A. Svoronos, Some effects of nonstationarity on multiserver markovian systems, Oper. Res. 39 (1991) 502–511.

[13] R. Haji and G. Newell, A relation between stationary queue and waiting time distributions, J. Appl. Probab. 8 (1971) 617–620.

[14] D. Heyman and M. Sobel, *Stochastic Models in Operations Research: Vol. 1* (McGraw-Hill, New York, 1982).

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

[16] V.B. Iversen, Decomposition of an $M/D/r \cdot k$ queue with FIFO into k $E_k/D/r$ queues with FIFO, Oper. Res. Lett. 2 (1983) 20–21.

[17] J. Keilson and L. Servi, A distributional form of Little's law, Oper. Res. Lett. 7 (1988) 223–227.

[18] J. Keilson and L. Servi, The distributional form of Little's law and the Fuhrmann–Cooper decomposition, Oper. Res. Lett. 9 (1990) 239–247.

[19] 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).

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

[21] J. Little, A proof of the theorem $L = \lambda W$, Oper. Res. 9 (1961) 383–387.

[22] K.M Malone, Dynamic queueing systems: behavior and approximations for individual queues and for networks, Ph.D. thesis, Massachusetts Institute of Technology (Cambridge, MA, 1995).

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

[24] G. Mourtzinou, An axiomatic approach to queueing systems, Ph.D. thesis, Massachusetts Institute of Technology (Cambridge, MA, 1995).

[25] A. Odoni and E. Roth, An empirical investigation of the transient behavior of stationary queueing systems, Oper. Res. 31 (1983) 432–455.

[26] K.L. Ong and M.R. Taaffee, Non-stationay queues with interrupted Poisson arrivals and unreliable/repairable servers, Queueing Systems 4 (1989) 27–46.

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

[28] 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.

[29] S. Ross, *Introduction to Probability Models*, 5th edn. (Academic Press, London, 1993).

[30] S. Stidham Jr., A last word on $L = \lambda W$, Oper. Res. 22 (1974) 417–421.

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