Adv. Appl. Prob. **28**, 588–625 (1996)
Printed in N. Ireland

A UNIFIED METHOD TO ANALYZE OVERTAKE FREE QUEUEING SYSTEMS

DIMITRIS BERTSIMAS* AND
GEORGIA MOURTZINOU,** *Massachusetts Institute of Technology*

Abstract

In this paper we demonstrate that the distributional laws that relate the number of customers in the system (queue), $L(Q)$ and the time a customer spends in the system (queue), $S(W)$ under the first-in-first-out (FIFO) discipline are special cases of the $H = \lambda G$ law and lead to a complete solution for the *distributions* of L, Q, S, W for queueing systems which satisfy distributional laws for both L and Q (*overtake free systems*). Moreover, in such systems the derivation of the distributions of L, Q, S, W can be done in *a unified way*. Consequences of the distributional laws include a generalization of PASTA to queueing systems with arbitrary renewal arrivals under heavy traffic conditions, a generalization of the Pollaczek–Khinchine formula to the $GI/G/1$ queue, an extension of the Fuhrmann and Cooper decomposition for queues with generalized vacations under mixed generalized Erlang renewal arrivals, approximate results for the distributions of L, S in a $GI/G/\infty$ queue, and exact results for the distributions of L, Q, S, W in priority queues with mixed generalized Erlang renewal arrivals.

PERFORMANCE ANALYSIS; DISTRIBUTIONAL LAWS; PRIORITY QUEUES; GENERALIZED VACATIONS; HEAVY TRAFFIC; MIXED GENERALIZED ERLANG ARRIVALS

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 60K25; 90B22

1. Introduction

What are the laws of electrodynamics? In order to address this question we should first define the fundamental quantities of electrodynamics, the electric field $\vec{E}$ and the magnetic field $\vec{B}$. The fundamental laws of electrodynamics are the Maxwell equations. The goal of electrodynamics is then to find $\vec{E}$ and $\vec{B}$ in various applications. The Maxwell equations form a *complete set* of laws in the sense that *just starting* from them and using the calculus of partial differential equations one is able to compute $\vec{E}$ and $\vec{B}$ either analytically or numerically in a variety of applications. What is important here is that the physics of a problem is summarized in the Maxwell equations, which then lead to a complete solution for $\vec{E}$ and $\vec{B}$ in a *unified way.*

Received 2 November 1992; revision received 5 December 1995.

* Postal address: Sloan School of Management and Operations Research Center, MIT, Cambridge, MA 02139, USA.

** Postal address: Operations Research Center, MIT, Cambridge, MA 02139, USA.

The research of D. Bertsimas was partially supported by a Presidential Young Investigator Award DDM-9158118 with matching funds from Draper Laboratory. The research of both authors was partially supported by the National Science Foundation under grant DDM-9014751.

Let us then ask the key question which motivated the present paper. What are the laws of queueing theory? The fundamental quantities in queueing theory are the stationary queue and system length (Q, L) and the waiting and system time (W, S) under the first-in-first-out (FIFO) discipline. Of course there are several other random variables of interest (often particular to the application studied), but these are the most widely used. The goal of queueing theory is then to find the distributions of Q, L, W, S in various applications. In its almost one hundred year history queueing theory has addressed a great variety of problems using a variety of techniques, which solve some problems but fail on others. What is interesting is the lack of a *unified way* to solve a particular application. Queueing theory research does not start from a set of well established laws and then proceed to the solution using some well established mathematical techniques. It rather uses the particular characteristics of the application to achieve its solution.

Coming to our original question regarding the laws of queueing theory, one would like to have a set of laws which, similar to Maxwell's equations in electrodynamics, lead to a *complete* solution of the queueing application. A first candidate might be Little's law [15]. However, as Little's law does not contain second moment information, it cannot lead to a complete solution. The next candidate might be the generalization of Little's law: $H = \lambda G$, as pointed out in Whitt [25] and Miyazawa [18], who also points out that the rate conservation law, $H = \lambda G$ and the Palm transformation for stationary marked point processes are essentially equivalent. Indeed, Heyman and Sobel [10] (see also Baccelli and Brémaud [1]) illustrate that the Pollaczek–Khinchine formula for the $M/G/1$ queue follows easily from $H = \lambda G$.

Our goal in this paper is to provide further evidence that $H = \lambda G$ is the right set of laws. In particular, we demonstrate that by applying $H = \lambda G$ to systems that do not allow overtaking, we derive the distributional laws first obtained by Haji and Newell [9]. We then demonstrate that the distributional laws lead to a complete solution for the stationary distributions of L, Q, S, W in overtake free systems. Moreover, in such systems the derivation of the distributions of L, Q, S, W can be done in a *unified way*. In this way not only do we obtain new simple derivations of known results providing new insights into old results, but we obtain several new results as well. We propose two methods of analysis: an asymptotic (in heavy traffic) method which applies to overtake free systems with arbitrary renewal arrivals and an exact method which applies to overtake free systems with mixed generalized Erlang arrivals.

For the case of Poisson arrivals Keilson and Servi [12], [13] found that the distributional laws have a very convenient form that can lead to complete solutions for some overtake free systems. For the case of mixed generalized Erlang renewal arrivals Bertsimas and Nakazato [2] gave another proof of the distributional laws that led to a very convenient form of the law as well. They also proposed a framework to find $\boldsymbol{E}[L]$, $\boldsymbol{E}[Q]$, $\boldsymbol{E}[S]$, $\boldsymbol{E}[W]$ in heavy traffic for overtake free queueing systems based on the distributional laws. In this paper we develop a

methodology to find the distributions of L, Q, S, W for overtake free systems with arbitrary renewal arrivals, thus generalizing all earlier work. Our apporach is to use asymptotic analysis (which is exact in heavy traffic) for the case of arbitrary renewal processes and exact analysis for the case of mixed generalized Erlang renewal arrivals.

The paper is structured as follows. In Section 2 we review the distributional laws and present a new derivation from $H = \lambda G$ providing further evidence of the power of the $H = \lambda G$ law. In Section 3 we present an asymptotic method of analysis for overtake free queueing systems based on the asymptotic properties of the distributional laws and a generalization of the well known result of Poisson arrivals see time averages (PASTA) to queueing systems with arbitrary renewal arrivals under heavy traffic conditions. Furthermore, we illustrate the efficiency of the method by deriving the distributions of L, Q, S, W in $GI/G/1$, $GI/D/s$ queues and obtaining approximate results for the distributions of L, S in a $GI/G/\infty$ queue. Our derivation unifies the heavy traffic results and leads to a generalization of the Pollaczek–Khinchine formula to the $GI/G/1$ queue. In Section 4 we present an exact method of analysis for overtake free systems with mixed generalized Erlang (MGE) renewal arrivals and we implement it in the case of the $MGE_M/G/1$ queue. This section demonstrates that there is a direct closed form expression for the number of customers in a $MGE_M/G/1$ system, while our approach reproduces the known results for the waiting time involving roots of a certain nonlinear equation in a direct way without the need for Hilbert factorization. In Section 5, as another application of the exact method of analysis for overtake free systems, we extend the decomposition results for queues with generalized vacations considered in Fuhrmann and Cooper [6] for the $M/G/1$ queue to MGE arrivals. In Section 6 we propose an algorithm to find the distributions of L, Q, S, W in priority queues with mixed generalized Erlang renewal arrivals, thus we generalize earlier results for Poisson arrivals. The derivations in this section are considerably more complicated compared with the results in previous sections. Finally, in Section 7 we include some concluding remarks and indicate directions for future research.

2. The distributional law

In this section we first review the distributional law for arbitrary arrivals and then consider the case in which the arrival process is a mixed generalized Erlang renewal process.

2.1. *A review of the distributional law.* Consider a general queueing system, with a renewal arrival process. As will become apparent later the 'system' may either correspond to a single queue, a queue plus a service facility as well as several queues in tandem. Let T_j be the arrival time of customer C_j, with $T_0 = 0$ and $T_0 < T_1 < \cdots < \infty$. Assume that the system satisfies the following set of assumptions.

Assumptions A (*Distributional law assumptions*).
A.1 All arriving customers enter the system one at a time, remain in the system until served (there is no blocking, balking or reneging) and leave also one at a time.
A.2 The customers leave the system in the order of arrival (FIFO).
A.3 New arriving customers do not affect the time in the system for previous customers.

Let us define S_j to be the total time customer C_j spends in the system and $L(t)$ to be the number of customers in the system at time t. Let $N_a(t)$ be the number of customers that arrived up to time t for the ordinary process (where the time of the first interarrival time has the same distribution as the stationary interarrival time). Let $N_a^*(t)$ be the number of arrivals up to time t for the equilibrium process (where the time of the first interarrival time is distributed as the forward recurrence time of the arrival process). Assuming that the system reaches, eventually, steady-state let L, L^-, L^+ be the steady-state number of customers in the system at a random observation time, just before an arrival or just after a departure, respectively. Similarly, let S be the steady-state system time. The distributional law can be stated as follows:

Theorem 1. (*Haji and Newell* [9]) *For a single class system that satisfies Assumptions A, the stationary number of customers, L, and the stationary system time, S, are related in distribution by*

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

while

$$G_L(z) = \int_0^\infty K(z, t)\, dF_S(t), \tag{2}$$

where

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

From $H = \lambda G$ to distributional laws. To derive the steady state distributional laws from the relation $H = \lambda G$, we first need to define a sequence of customer dependent random variables, $g_j(\cdot)$, a stochastic process $h(t; \cdot)$ and the corresponding admissible function, $f_j(t; \cdot)$ (see [10]), such that for an arbitrary realization of the queueing system ω:

$$h(t; \omega) \stackrel{\Delta}{=} \sum_{j=1}^{\infty} f_j(t; \omega) \qquad \text{and} \qquad g_j(\omega) \stackrel{\Delta}{=} \int_0^\infty f_j(t; \omega)\, dt. \tag{3}$$

More specifically, using lower case letters to define realizations of the corresponding upper case random variables, let us define the following indicator function:

$$f_j(t; \omega) = \begin{cases} 1 & \text{if } t_j < t \leqq t_j + s_j \quad \text{and} \quad n_a(t - t_j) = n - 1 \\ 0 & \text{otherwise.} \end{cases}$$

In words, $f_j(t; \omega) = 1$ if under realization ω customer C_j is the nth most recently arrived customer to the system at time t and he is still in the system at t. Using

indicator functions we rewrite $f_j(t;\omega)$ as follows:

$$f_j(t;\omega) = \mathbf{1}\{t_j < t \leqq t_j + s_j,\ n_a(t-t_j) = n-1\}.$$

From (3) we obtain

$$h(t;\omega) = \mathbf{1}\{l(t;\omega) \geqq n\} \quad \text{and} \quad g_j(\omega) = \int_0^\infty \mathbf{1}\{t_j < t \leqq t_j + s_j,\ n_a(t-t_j) = n-1\}\,dt,$$

hence, under realization ω, $h(t;\omega)$ is the indicator of the event that there are at least n customers in the system at time t and $g_j(\omega)$ represents the time period during which C_j as well as the $n-1$ customers that arrived after him are still in the system (notice the use of Assumptions A.1 and A.2).

Furthermore, using Assumption A.3 we can rewrite the above relation as follows:

$$h(t;\omega) = \mathbf{1}\{l(t;\omega) \geqq n\} \quad \text{and} \quad g_j(\omega) = \int_0^\infty \mathbf{1}\{\tau \leqq s_j\}\mathbf{1}\{n_a(\tau) = n-1\}\,d\tau,$$

where $\tau \triangleq t - t_j$. Using the sample path version of $H = \lambda G$ (see [25], Theorem 6.1) we obtain

$$\lim_{t\to\infty}(1/t)\int_0^t \mathbf{1}\{l(t;\omega) \geqq n\}\,dt = \lambda \lim_{k\to\infty}(1/k)\sum_{j=1}^{k}\int_0^\infty \mathbf{1}\{\tau \leqq s_j\}\mathbf{1}\{n_a(\tau) = n-1\}\,d\tau.$$

Moreover, if we define by $H(t)$, G_j the random variables corresponding to the realizations $h(t;\omega)$ and $g_j(\omega)$ we have that

$$H(t) \triangleq \mathbf{1}\{L(t) \geqq n\} \quad \text{and} \quad G_j \triangleq \int_0^\infty \mathbf{1}\{\tau \leqq S_j\}\mathbf{1}\{N_a(\tau) = n-1\}\,d\tau.$$

Hence, the steady state limits should be defined as follows:

$$H \triangleq \mathbf{1}\{L \geqq n\} \quad \text{and} \quad G \triangleq \int_0^\infty \mathbf{1}\{\tau \leqq S\}\mathbf{1}\{N_a(\tau) = n-1\}\,d\tau.$$

Using the steady state version of $H = \lambda G$ (see [25], p. 242 for the set of underlying assumptions) we conclude that

$$\boldsymbol{E}[\mathbf{1}\{L \geqq n\}] = \lambda \boldsymbol{E}\left[\int_0^\infty \mathbf{1}\{\tau \leqq S\}\mathbf{1}\{N_a(\tau) = n-1\}\,d\tau\right],$$

or equivalently,

$$\boldsymbol{P}\{L \geqq n\} = \lambda \int_0^\infty \boldsymbol{P}\{\tau \leqq S\}\boldsymbol{P}\{N_a(\tau) = n-1\}\,d\tau, \tag{4}$$

where we interchanged the expectation and the integral using Fubini's Theorem.

Combining (4) with the facts that $\boldsymbol{P}\{L = n\} = \boldsymbol{P}\{L \geqq n\} - \boldsymbol{P}\{L \geqq n+1\}$ and $\boldsymbol{P}\{L \geqq 0\} = 1$ we obtain a formula for $\boldsymbol{P}\{L = n\}$. Taking generating functions, we obtain, after some straightforward algebraic manipulations, that $G_L(z) = 1 + \lambda(z-1)\int_0^\infty K_0(z,t)\boldsymbol{P}\{S \geqq t\}\,dt$, where $K_0(z,t) \triangleq \sum_{n=0}^\infty z^n \boldsymbol{P}\{N_a(t) = n\}$. Integrating the above relation by parts and using the fact (see [4]) that $K(z,t) = 1 + \lambda(z-1)\int_0^t K_0(z,u)\,du$ we obtain (2).

Remarks.

1. Relation (1) holds even if we relax the assumption that the arrival process is renewal and we consider the broader family of stationary arrival processes (see [9]).
2. Similar relations hold for the number of customers in the system just before an arrival or just after a departure. Namely,

$$L^- \overset{\mathrm{d}}{=} L^+ \overset{\mathrm{d}}{=} N_a(S)$$

$$G_{L^-}(z) = G_{L^+}(z) = \int_0^\infty K_0(z, t)\, dF_S(t). \tag{5}$$

The Laplace transform of the renewal generating functions $K(z, t)$ and $K_0(z, t)$ are given by

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

$$K_a^*(z, s) \triangleq \int_0^\infty e^{-st} K_0(z, t)\, dt = \frac{1-\alpha(s)}{s(1-z\alpha(s))}. \tag{6}$$

3. For the case of Poisson arrivals $K(z, t) = K_0(z, t) = e^{-\lambda t(1-z)}$ and thus the distributional laws become a relation between transforms (Keilson and Servi [12]):

$$G_L(z) = \phi_S(\lambda(1-z)). \tag{7}$$

2.2. *A vector distributional law.* A vector generalization of (7) has been proposed in Bertsimas and Nakazato [2] under the assumption that the arrival process is a mixed generalized Erlang (MGE) process, which can approximate any renewal arrival process arbitrarily closely. While the class MGE is a special case of the phase type distribution PH (see [19]) it is the simplest class of distributions that is dense in the space of all distributions. The stage representation of the MGE distribution is presented in Figure 1, i.e. we conceive of the arrival process as an arrival timing channel (ATC) consisting of M consecutive exponential stages with rates $\lambda_1, \lambda_2, \cdots, \lambda_M$ and with probabilities $p_1, p_2, \cdots, p_M$ $(p_M = 1)$ of entering the system after the completion of the 1st, 2nd, $\cdots$, Mth stage.

Let $a_k(t)$ be the pdf of the remaining interarrival time if the customer in the ATC is in stage $k = 1, \cdots, M$. Therefore, $a(t) = a_1(t)$ is the pdf of the interarrival time. For notational convenience we will drop the subscript for $k = 1$. Also $1/\lambda$ denotes the mean interarrival time.

Let $\alpha_k(s)$ be the Laplace transform of $a_k(t)$. Let $a_i^j(t)$ be the probability to move from stage $i \leqq j$ of the ATC to stage j during the interval $[0, t)$ without having any new arrival. We will also use the notation: $\vec{a}_1(t) = (a_1^1(t), \cdots, a_1^M(t))$, $\vec{a}_k(t) = (0, \cdots, a_k^k(t), \cdots, a_k^M(t))$. $\vec{\alpha}_k(s)$ denotes the Laplace transforms of $\vec{a}_k(t)$. $\vec{e}_j = (0, \cdots, 1, \cdots, 0)$, $\vec{1} = (1, \cdots, 1, \cdots, 1)$.

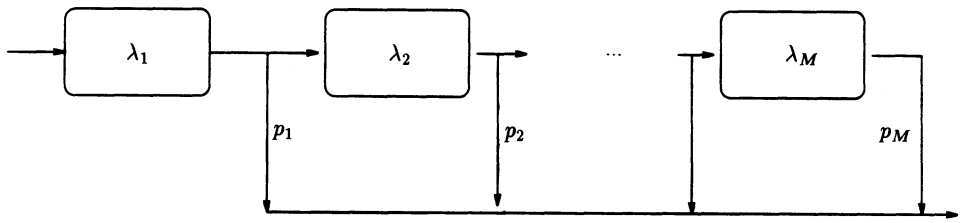

Figure 1. The Coxian class of distributions.

By introducing the following upper semidiagonal matrix A_0 and the dyadic matrix A_1

$$A_0 = \begin{bmatrix} \lambda_1 & -(1-p_1)\lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & -(1-p_2)\lambda_2 & \ddots & \vdots \\ \vdots & \ddots & & \ddots & \vdots \\ \vdots & & & \lambda_{M-1} & -(1-p_{M-1})\lambda_{M-1} \\ 0 & \cdots & \cdots & 0 & \lambda_M \end{bmatrix},$$

$$A_1 = \begin{bmatrix} -p_1\lambda_1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ -p_M\lambda_M & 0 & \cdots & 0 \end{bmatrix},$$

we can express compactly the transforms defined above as follows:

$$\vec{\alpha}_k(s) = \vec{e}_k(Is + A_0)^{-1},$$

$$\alpha_k(s) = -\vec{e}_k(Is + A_0)^{-1}A_1\vec{e}_1' = \sum_{r=k}^{M} p_r\lambda_r\alpha_k^r(s) = \sum_{r=k}^{M} p_r\lambda_r \frac{\prod_{i=k}^{r-1}(1-p_i)\lambda_i}{\prod_{i=k}^{r}(s+\lambda_i)},$$

$$\alpha(s) = -\text{trace}\,((Is + A_0)^{-1}A_1),$$

thus the interarrival pdf becomes $a(t) = -\text{trace}(e^{-A_0 t}A_1)$.

Note that a mixed generalized Erlang renewal process is fully characterized by the matrices A_0, A_1. In queueing systems with mixed generalized Erlang renewal arrival processes we introduce:

L^+ = the number of customers in the system immediately after a departure epoch.

L_t^- = the number of customers in the system just before a *transition* epoch of the arrival process. A transition includes both arrivals in the system and shifts to the next exponential stage of the ATC. We emphasize that L_t^- is *not* the number of customers before an *arrival* epoch. The motivation for considering L_t^- is that using uniformization the epochs of transition are Poisson distributed and thus we can apply PASTA.

R^+ = the ATC stage immediately after a departure epoch.
R_t^- = the ATC stage just before a transition epoch of the arrival process.

Finally we use the following row vectors: $\vec{P}_n^+ = [\boldsymbol{P}\{L^+ = n \cap R^+ = i\}]_{i=1}^{i=M}$, $\vec{P}_L^+(z) = \sum_{n=0}^{\infty} z^n \vec{P}_n^+$, $\vec{P}_n^- = [\boldsymbol{P}\{L_t^- = n \cap R_t^- = i\}]_{i=1}^{i=M}$, $\vec{P}_L^-(z) = \sum_{n=0}^{\infty} z^n \vec{P}_n^-$, and $\vec{P}_n = [\boldsymbol{P}\{L = n \cap R = i\}]_{i=1}^{i=M}$, $\vec{P}_L(z) = \sum_{n=0}^{\infty} z^n \vec{P}_n$. The vector distributional law is described in the following theorem.

Theorem 2. (*Bertsimas and Nakazato* [2]) *Under Assumptions A and for mixed generalized Erlang interarrival times characterized by the matrices* A_0, A_1,

$$\vec{P}_L(z) = \vec{P}_L^-(z), \tag{8}$$

$$\vec{P}_L(z)(A_0 + zA_1) = \lambda(1 - z)\vec{P}_L^+(z),$$

$$\vec{P}_L^+(z) = \vec{e}_1 \Phi_S(A_0 + zA_1), \tag{9}$$

$$\vec{P}_L(z) = \lambda(1 - z)\vec{e}_1 \Phi_S(A_0 + zA_1)(A_0 + zA_1)^{-1},$$

where for any matrix D *we symbolically define* $\Phi_S(D) \triangleq \int_0^{\infty} e^{-Dt}\, dF_S(t)$. *The kernel* $K(z, t)$ *in* (2) *is given by*

$$K(z, t) = \lambda(1 - z)\vec{e}_1 e^{-(A_0 + zA_1)t}(A_0 + zA_1)^{-1}\vec{1}',$$

which leads to

$$G_L(z) = \lambda(1 - z)\vec{e}_1 \Phi_S(A_0 + zA_1)(A_0 + zA_1)^{-1}\vec{1}'.$$

Once again in the case of Poisson arrivals the vector forms reduce to scalars and we obtain (7).

We define as *overtake free systems* those systems that satisfy Assumptions A if we define the 'system' to be either just the queue or the queue plus the service facility. Examples of such systems include (a) $GI/G/1$ under FIFO, (b) $GI/D/s$ under FIFO, (c) systems with vacations.

For such systems we can use Theorems 1 and 2 for both the queue and the queue plus the service facility. More specifically, if we keep the notation of the previous sections when the 'system' is the queue plus the service facility and we also denote by Q and Q^+ the steady-state number of customers in the queue at an observation epoch and just after a departure, respectively, and by W the steady-state waiting time (in the queue), then we have that the relations of both Theorem 1 and Theorem 2 also hold if we substitute Q for L and W for S. For example, similarly to (2), we can obtain

$$G_Q(z) \triangleq \boldsymbol{E}[z^Q] = \int_0^{\infty} K(z, t)\, dF_W(t). \tag{10}$$

3. An asymptotic method of analyis for overtake free systems

In this section we consider overtake free systems with general arrival processes that satisfy Assumptions A and have the property that whenever $\rho \to 1$,

$L, Q, S, W \to \infty$, and we propose a unified asymptotic method for the derivation of the distributions of L, Q, S, W, as well as L^+ and Q^+. This section is structured as follows. In Section 3.1 we derive the asymptotic form of the distributional law while in Section 3.2 we give an asymptotic generalization of the PASTA property. In Section 3.3 we present the asymptotic method of analysis for overtake free systems. Finally, in Section 3.4, we implement this method in specific examples, i.e. $GI/G/1$, $GI/D/s$ and $GI/G/\infty$ queues, to obtain new asymptotic results.

3.1. *The asymptotic distributional law.* The important advantage of the Poisson arrival process is that the kernel $K(z, t)$ in Theorem 2 has the very tractable form $K(z, t) = e^{-\lambda(1-z)t}$. As mentioned above, the distributional law then becomes a relation among transforms, i.e. $G_L(z) = \phi_S(\lambda(1-z))$. For mixed generalized Erlang arrivals $K(z, t)$ is given explicitly in Theorem 2. For arbitrary renewal arrivals, however, $K(z, t)$ is not known in closed form. In order to exploit the distributional laws we try to understand in this section the asymptotic behavior of $K(z, t)$. For systems in heavy traffic ($\rho \to 1$) both L, Q, S, W tend to infinity (we need to exclude systems with deterministic arrivals and deterministic service, i.e. $D/D/1$). We will use the notation that under heavy traffic conditions $h(x) \sim r(x)$ means that $\lim_{\rho \to 1} h(x)/r(x) = 1$.

As a result of the integral form of the distributional laws, for systems in heavy traffic we are interested in the behavior of $K(z, t)$, $K_0(z, t)$ as $t \to \infty$ and $z \to 1$. Following the asymptotic approach introduced in Smith [23] (see also Cox [4], ch. 4–6) we obtain:

Theorem 3. *Asymptotically, as* $t \to \infty$ *and* $z \to 1$ *the kernels in Theorem* 1 *behave as follows:*

$$K(z, t) \sim e^{-tf(z)}, \qquad \text{and} \qquad K_0(z, t) \sim [1 - \tfrac{1}{2}(1-z)(c_a^2 - 1) + O((1-z)^2)]e^{-tf(z)},$$

where

$$f(z) - \lambda(1-z) = \tfrac{1}{2}\lambda(1-z)^2(c_a^2 - 1),$$

and c_a^2 *is the square coefficient of variation of the interarrival process.*

Proof. From (6) by writing $K^*(z, s) = N(z, s)/s^2 D(s, z)$ and expanding $N(z, s)$, $D(s, z)$ as a Taylor series up to second-order terms in s (note that $t \to \infty$ in the time domain is equivalent to $s \to 0$ in the transform domain) we have

$$K^*(z, s) = \frac{(2z/\lambda) - \lambda(1-z)\boldsymbol{E}[A^2] + [z\boldsymbol{E}[A^2] - \frac{1}{3}\lambda(1-z)\boldsymbol{E}[A^3]]s + O(s^2)}{(s - s_1)(s - s_2)z\boldsymbol{E}[A^2]},$$

where we used the facts that $\dot{\alpha}(0) = 1/\lambda$, $\ddot{\alpha}(0) = \boldsymbol{E}[A^2], \cdots, \boldsymbol{E}[A^k]$ being the kth moment of the interarrival time, and s_1 and s_2 are the roots of

$$z\boldsymbol{E}[A^2]s^2 + 2(z/\lambda)s + 2z - 2 = 0.$$

If we denote by s_1 the smallest of the two roots, we can easily obtain that:

$$s_1 = -\frac{1}{zE[A^2]}\left[\sqrt{\frac{z^2}{\lambda^2} - 2z(z-1)E[A^2]} - \frac{z}{\lambda}\right] \quad \text{and} \quad s_2 = \frac{2}{\lambda E[A^2]} - s_1.$$

Using a partial fraction expansion we invert in the time domain to obtain that as $s \to 0$,

$$K^*(z, s) \sim g(z)/(s - s_1) + u(z)/(s - s_2),$$

with $g(z) = \lim_{s \to s_1}(s - s_1)K^*(z, s)$ and $u(z) = \lim_{s \to s_2}(s - s_2)K^*(z, s)$. Expanding, now around $z = 1$ we obtain after some tedious but straightforward manipulations that

$$s_1 \sim -\lambda(1-z) + \tfrac{1}{2}\lambda(1-z)^2(c_a^2 - 1) + O((1-z)^3)$$

and

$$g(z) \sim 1 + O((1-z)^2), \qquad u(z) \sim O((1-z)),$$

hence inverting back in the time domain

$$K(z, t) \sim (1 + O((1-z)^2))e^{s_1 t} = (1 + O((1-z)^2)) \times \exp\left[-t(\lambda(1-z) - \tfrac{1}{2}\lambda(1-z)^2(c_a^2 - 1))\right].$$

In a similar way, by expanding $K_a^*(z, s)$ as a Taylor series in terms of s and inverting in the time domain keeping only the most important term asymptotically, we obtain that

$$K_0(z, t) \sim [1 - \tfrac{1}{2}(1-z)(c_a^2 - 1) + O((1-z)^2)] \exp\left[-t(\lambda(1-z) - \tfrac{1}{2}\lambda(1-z)^2(c_a^2 - 1))\right].$$

Combining Theorems 1 and 2 we obtain the following asymptotic form of distributional laws for overtake-free systems.

Theorem 4. *In an overtake free queueing system that satisfies Assumptions A and assuming that as* $\rho \to 1$, $L, Q, S, W \to \infty$ *the following asymptotic relations hold as* $\rho \to 1$*:*

$$G_L(z) \sim \phi_S(f(z)), \tag{11}$$

$$G_Q(z) \sim \phi_W(f(z)), \tag{12}$$

$$G_{L^+}(z) \sim \frac{f(z)}{\lambda(1-z)}\phi_S(f(z)), \tag{13}$$

$$G_{Q^+}(z) \sim \frac{f(z)}{\lambda(1-z)}\phi_W(f(z)), \tag{14}$$

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

Proof. Substituting in (10), (2) and (5) the asymptotic form of $K(z, t)$ and $K_0(z, t)$ from the previous theorem we obtain (11), (12) and (13), (14), respectively.

Although only valid asymptotically, (11), (12) and (13), (14) are very useful since they are relations among transforms, which we will further exploit in this section. Also, the previous expressions are exact for the Poisson case ($c_a^2 = 1$). In order to develop some further insight into the asymptotic expressions of Theorem 4 we consider the case of E_2 arrivals, i.e. $\alpha(s) = (2\lambda/(2\lambda + s))^2$. Then,

$$K(z, t) = \frac{(1+\sqrt{z})^2}{4\sqrt{z}} \exp[-2\lambda(1-\sqrt{z})t] - \frac{(1-\sqrt{z})^2}{4\sqrt{z}} \exp[-2\lambda(1+\sqrt{z})t],$$

and

$$K_0(z, t) = \frac{(1+\sqrt{z})}{2\sqrt{z}} \exp[-2\lambda(1-\sqrt{z})t] - \frac{(1-\sqrt{z})}{2\sqrt{z}} \exp[-2\lambda(1+\sqrt{z})t].$$

As $z \to 1$ only the first of the two exponentials contributes to $K(z, t)$, $K_0(z, t)$. Expressions (11) and (13) are the Taylor series expansions of the first exponential in terms of $1 - z$.

3.2. *An asymptotic generalization of PASTA.* Theorem 4 leads to an interesting generalization of PASTA for systems in heavy traffic. Consider a queueing system that satisfies Assumptions A. Since in such systems the number of customers in the system always changes by one (for example a $GI/G/s$ queue), $L^+ = L^-$ in distribution. In the case of Poisson arrivals, PASTA implies that $L^- = L$ in distribution. For general arrival processes, though, the distributions of L^- and L are different. In heavy traffic ($\rho \to 1$); however, where Theorem 4 is applicable, we have

$$G_{L^-}(z) = G_{L^-}(z) \sim G_L(z)[1 - \tfrac{1}{2}(1-z)(c_a^2 - 1)]. \tag{15}$$

In particular the first moments are related by

$$\boldsymbol{E}[L^-] \sim \boldsymbol{E}[L] + \tfrac{1}{2}(c_a^2 - 1),$$

which means that in heavy traffic, where both $\boldsymbol{E}[L^-]$, $\boldsymbol{E}[L]$ are very large, their difference asymptotically depends only on the coefficient of variation of the arrival process. By similar arguments a relation similar to (15) holds for the number of customers in the queue. We remark that we need that L^-, L (or Q^-, Q) go to infinity as $\rho \to 1$. For example, in a $D/D/1$ queue, even if $\rho \to 1$, (15) does not hold, since L^-, L (and Q^-, Q) remain bounded and therefore the assumptions of Theorem 3 are not valid.

Relation (15) is an asymptotic formula, which provides insight mainly due to its simplicity. However, there exist in the literature other *exact* generalizations of PASTA (see for example [17]) that apply to a more general class of systems; they tend however to be complicated and, as a result, not always tractable.

3.3. *An asymptotic method of analysis.* Theorem 4 as well as (15) provide us with the necessary analytical tools to form a unified method that solves, asymptotically, overtake free systems.

Let L, Q be the number of customers in the system and queue respectively, and S and W be the time spent in the system and queue. Let the random variable X

denote the service time and let also $L^+(Q^+)$ be the number of customers in the system (or in the queue) just after a departure. We can describe the proposed method in an algorithmic way as follows.

Asymptotic method of analysis

1. Relate the transforms of L and S, using the asymptotic form of the distributional law (11).
2. Relate the transforms of Q and W, using the asymptotic form of the distributional law (12).
3. Relate the transforms of S and W using the fact that $S = W + X$.
4. Relate the transforms of L and Q using the characteristics of the system (see Section 3.4 for further details).
5. Solve the 4×4 system of equations from the previous 4 steps to find the transforms of L, Q, S and W.
6. Using the asymptotic generalization of PASTA, (15), find the transforms of L^+ and Q^+ from the transforms of L and Q.

3.4. *Application of the asymptotic method.*

The GI/G/1 and GI/D/s queues. As a first application we consider a $GI/G/1$ queue with a FIFO service discipline. Let $1/\lambda$, $\boldsymbol{E}[X]$, c_a^2, c_x^2 be the means and the square coefficients of variation for the interarrival and service time distributions. Let $\phi_X(s)$ be the Laplace transforms of the service time distribution.

Theorem 5. *In a* $GI/G/1$ *queue under FIFO as* $\rho \to 1$ *the Laplace transform of the waiting time distribution and the* z*-transform of the number of customers in the queue are given by*

$$\phi_W(s) \sim \frac{(1 - f^{-1}(s))(1 - \rho)}{\phi_X(s) - f^{-1}(s)}, \tag{16}$$

and

$$G_Q(z) \sim \frac{(1 - z)(1 - \rho)}{\phi_X(f(z)) - z}, \tag{17}$$

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

Proof. The distributional law holds for both L and Q. Performing the first two steps of the asymptotic method we obtain from (11) and (12), as $\rho \to 1$:

$$G_L(z) \sim \phi_S(f(z)), \qquad G_Q(z) \sim \phi_W(f(z)).$$

Performing the third step, since $S = W + X$ and W, X are independent we obtain

$$\phi_S(f(z)) = \phi_W(f(z))\phi_X(f(z)).$$

Finally, performing the fourth step, we obtain that the relation of the generating functions of L, Q is

$$G_L(z) = (1 - z)(1 - \rho) + zG_Q(z).$$

The previous equations form a system of four equations with four unknowns. By setting $s = f(z)$ and thus $z = f^{-1}(s)$ and solving the system of equations we obtain (16) and (17), as well as the transforms of the system time and the number of customers in the system.

Remarks.

1. Using (15) we can also find $G_{L^+}(z)$ or $G_{Q^+}(z)$ as $\rho \to 1$.
2. In the case of Poisson arrivals, it is important to notice that (16), (17) are exact and generalize the well known Pollaczek–Khinchine formulae for the $M/G/1$ queue.
3. By expanding $\phi_W(s)$ in powers of s we obtain

$$\phi_W(s) \sim 1 - s\frac{\rho^2(c_x^2+1) - \rho(1-c_a^2)}{2\lambda(1-\rho)} + As^2 + o(s^2),$$

with

$$A = \frac{1}{4}\left(\frac{(1-c_a^2)^2}{\lambda^2(1-\rho)^2} + \frac{\rho^4(1+c_x^2)^2}{\lambda^2(1-\rho)^2} - 2\frac{\rho^2(1-c_a^2)(1+c_x^2)}{\lambda^2(1-\rho)^2}\right).$$

Then, as $\rho \to 1$, $\boldsymbol{E}[W] \sim [\rho^2(c_x^2+1) - \rho(1-c_a^2)]/(2\lambda(1-\rho))$, and $\boldsymbol{E}[W^2] \sim 2A$. As a result, the coefficient of variation of W tends to one as $\rho \to 1$, which is consistent with the diffusion approximation for the waiting time in a $GI/G/1$ queue, i.e. W is exponentially distributed in heavy traffic.

4. The previous results for the $GI/G/1$ system can also be used in a $GI/D/s$ queue. Since the service times are deterministic, every s customers are served by the same server. Therefore, as is well known, 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.

The $GI/G/\infty$ queue. We now apply the asymptotic method to find approximate closed form expressions for the variance of the number in a $GI/G/\infty$ system.

Theorem 6. *In a $GI/G/\infty$ queue in heavy traffic conditions ($\boldsymbol{E}[X] \to \infty$)*

$$G_L(z) \approx \exp\left[-\lambda(1-z)\boldsymbol{E}[X] + \tfrac{1}{2}\lambda(1-z)^2(c_a^2-1)\int_0^\infty x f_X^2(x)\,dx\right],$$

$$\boldsymbol{E}[L] = \lambda \boldsymbol{E}[X],$$

and

$$\operatorname{Var}[L] \approx \lambda \boldsymbol{E}[X] + \lambda(c_a^2-1)\int_0^\infty x f_X^2(x)\,dx.$$

Proof. In a $GI/G/\infty$ system the distributional law does not hold because Assumption A.2 is violated (i.e. the system allows overtaking). In the special case of the $GI/D/\infty$ queue, however, the distributional law does hold because, due to the deterministic service distribution, the customers exit the system in the order they arrived. Thus we can write $L \overset{d}{=} N_a^*(s)$. Moreover, because of the presence of an

infinite number of servers there is no waiting and thus $S = X$, i.e. the time in the system is exactly the service time. But, the pdf of X is $f_X(t) = \delta(t - \boldsymbol{E}[X])$ and thus from (10)

$$G_L(x) = K(z, \boldsymbol{E}[X]). \tag{18}$$

We will now *decompose* the $GI/G/\infty$ system into a number of $GI/D/\infty$ systems. Suppose that instead of having a general service distribution the serivce time is $\boldsymbol{P}\{X = x_j\} = p_j$, $j = 1, \cdots, k$. The customers with service times x_j can be treated as a separate class C_j of customers with arrival process being a renewal process with Laplace transform $\alpha_j(s)$

$$\alpha_j(s) = \alpha(s)p_j \sum_{r=1}^{\infty} \alpha^{k-1}(s)(1-p_j)^{k-1} = \frac{\alpha(s)p_j}{1-(1-p_j)\alpha(s)},$$

i.e. the arrival rate and coefficient of variation for class C_j customers is

$$\lambda_j = \lambda p_j, \qquad c_{a_j}^2 = 1 + p_j(c_a^2 - 1).$$

If L_j, $j = 1, \cdots, k$ is the number of class C_j customers in the system, then $L = \sum_{j=1}^{k} L_j$. The random variables L_j are not independent since the arrival processes are not independent (in the special case of Poisson arrivals they are indeed independent). Using the *approximation* that they are indeed independent we obtain

$$G_L(z) \approx \prod_{j=1}^{k} G_{L_j}(z).$$

Each class C_j sees a $GI/D/\infty$ for which the distributional laws holds. Then applying (18), $G_{L_j}(z) = K(z, x_j)$. For large x_j the asymptotic form of the distributional law of Theorem 3 is valid and thus

$$K(z, x_j) \sim \exp\{-x_j[\lambda_j(1-z) - \tfrac{1}{2}\lambda_j(1-z)^2(c_{a_j}^2 - 1)]\}.$$

Therefore,

$$G_L(z) \approx \exp\left(-\lambda(1-z)\sum_{j=1}^{k} p_j x_j + \tfrac{1}{2}\lambda(1-z)^2(c_a^2 - 1)\sum_{j=1}^{k} p_j^2 x_j\right).$$

Since any general service distribution is the limit of a sequence of mixtures of deterministic distributions we obtain that

$$G_L(x) \approx \exp\left(-\lambda(1-z)\boldsymbol{E}[X] + \tfrac{1}{2}\lambda(1-z)^2(c_a^2 - 1)\int_0^{\infty} x f_X^2(x)\, dx\right),$$

which leads to $\boldsymbol{E}[L] = \lambda \boldsymbol{E}[X]$, and $\mathrm{Var}[L] \approx \lambda \boldsymbol{E}[X] + \lambda[c_a^2 - 1]\int_0^{\infty} x f_X^2(x)\, dx$.

Remarks.

1. For the case of Poisson arrivals ($c_a^2 = 1$) the expressions of the previous theorem are exact leading to the well known result $G_L(z) = \exp(-\lambda(1-z)\boldsymbol{E}[X])$, i.e. L has a Poisson distribution with rate $\lambda \boldsymbol{E}[X]$.
2. An alternative approach to obtain heavy traffic results for the $GI/G/\infty$ queue via Brownian approximations is presented in [8], where the authors use a sequence of systems with the arrival rate $\lambda \to \infty$ (instead of assuming that $\boldsymbol{E}[X] \to \infty$ as in

Theorem 6). It is interesting to notice that the approach in [8] yields the same results for the first and second moment as Theorem 6.

4. An exact method of analysis for overtake free systems

In this section we focus our attention on overtake free systems with mixed generalized Erlang (MGE) arrival processes that satisfy Assumptions A and we describe a unified exact method to obtain the distributions L, Q, S, W, L^+, and Q^+. We should point out that the same approach can be used for overtake free systems with phase-type (PH) renewal process, as introduced in [19], since Theorem 2 as well as the analysis of this section still hold if we use the appropriate matrices C and D instead of A_0 and A_1.

We first present the exact method in an algorithmic form in subsection 4.1 and then, in subsection 4.2, we illustrate its use in the case of $MGE_M/G/1$ and $MGE_M/D/s$ queues under FIFO.

4.1. *An exact method of analysis.* Theorem 2 enables us to present a unified exact method for solving overtake free systems with MGE arrivals under Assumptions A. We will use the notation of Section 2.2.

Exact method analysis.

1. Relate the transforms of L and S using the vector form of the distributional laws, (9).
2. Relate the transforms of Q and W applying (9) to the queue.
3. Relate the transforms S and W using the fact that $S = W + X$.
4. Relate the transforms $\vec{P}_L(z)$ and $\vec{P}_Q(z)$ using the characteristics of the system up to a vector term and use system properties to evaluate the vector (see Section 4.2 for further details).
5. Solve the 4×4 system of equations from the previous 4 steps to find $\vec{P}_L(z)$, $\vec{P}_Q(z)$, the transforms of S and W.
6. Find $\vec{P}_{L^+}(z)$, and $\vec{P}_{Q^+}(z)$ from (8).

We are going to illustrate how the method works through an application in the next section.

4.2. *The $MGE_M/G/1$ and $MGE_M/D/s$ queues under FIFO.* We consider in this subsection an $MGE_M/G/1$ queue, with a FIFO service discipline where the arrival process is a generalized Erlang process characterized by the matrices A_0 and A_1. Our goal is to evaluate the transforms of L, S, Q and W based on the exact method proposed in Section 4.1. We start with some definitions.

Let $\alpha(s) = \alpha_N(s)/\alpha_D(s)$ be the Laplace transform of the interarrival distribution where $\alpha_D(s)$, $\alpha_N(s)$ are polynomials of degree M and less than M respectively. We define the row vectors $\vec{H}$ and $\vec{H}^+$ with $H_i \triangleq \boldsymbol{P}\{L = 0, R = i\}$ and $H_i^+ \triangleq \boldsymbol{P}\{L^+ = 0, R^+ = i\}$, for $i = 1, \cdots, M$. As will become apparent in the sequel, knowledge of $\vec{H}$ is necessary to complete the third step of the exact method and

therefore we begin the analysis of the $MGE_M/G/1$ queue by evaluating $\vec{H}$ together with some other important quantities of the system.

In particular, let $\Gamma(s)$ be the busy period matrix with $[\Gamma(s)]_{i,j} = \gamma_{ij}(s)$, $i, j = 1, \cdots, M$ denoting the Laplace transform of a sub-busy period interval that ends with the $ATC = j$ given that it started with the $ATC = i$. Note that though a busy period interval is initialized by the first customer that arrives after an idle interval, a sub-busy period is initialized whenever a customer enters service (see, for example, [14] p. 210) and therefore at the beginning of a sub-busy period the ATC can be in any stage.

Let, also $C(s)$ be the busy cycle matrix with $[C(s)]_{i,j} = c_{ij}(s)$, $i, j = 1, \cdots, M$ denoting the Laplace transform of a busy cycle interval that ends with a customer leaving the system empty and the $ATC = j$ given that it started with a customer leaving the system empty and the $ATC = i$.

Proposition 1. *In a $MGE_M/G/1$ queueing system where the interarrival process is characterized by the matrices A_0 and A_1 and the service time is denoted by X we have that the busy period matrix $\Gamma(s)$ satisfies*

$$\Gamma(s) = \Phi_X(s + A_0 + A_1\Gamma(s)).$$

Furthermore, the busy cycle matrix $C(s)$ is given by

$$C(s) = (Is + A_0)^{-1}A_1\Gamma(s).$$

Finally, the vector $\vec{H}$ satisfies

$$\vec{H} = \lambda(1-\rho)\vec{H}^+A_0^{-1} \qquad \text{and} \qquad \vec{H}\vec{1}' = 1 - \rho,$$

with $\vec{H}^+$ the left eigenvector of $\lim_{s\to 0} C(s)$.

The proof of this proposition is presented in Appendix A.

Remarks.

1. The above results can also be obtained using an embedded Markov chain approach (see, for example, [16]). We presented the above proposition for the completeness of our analysis and because its proof is based on simple probabilistic arguments.
2. The transform $\gamma(s)$ of the busy period distribution is given by $\gamma(s) = \vec{e}_1\Gamma(s)\vec{1}'$.

Next, we prove the following theorem.

Theorem 7. *In a $MGE_M/G/1$ queue under FIFO*

$$\vec{P}_Q(z) = (1-z)\vec{H}(\Phi_X(A_0 + zA_1) - zI)^{-1}, \tag{19}$$

$$\vec{P}_L(z) = (1-z)\vec{H}(\Phi_X(A_0 + zA_1) - zI)^{-1}\Phi_X(A_0 + zA_1), \tag{20}$$

and

$$\phi_W(s) = \frac{\alpha_N(-s)}{\alpha_N(0)} \frac{(1-\rho)s}{\lambda(1-\alpha(-s)\phi_X(s))} \prod_{r=1}^{M-1} \frac{x_r - s}{x_r}, \tag{21}$$

where x_r, $r = 1, \cdots, M-1$ *are the* $M-1$ *roots of the equation* $\alpha(-s)\phi_X(s) = 1$, $\mathrm{Re}\,(s) > 0$, *and* $\vec{H}$ *is given in Proposition* 1.

Proof. Since this system is overtake free we will use the exact method of analysis described in the previous subsection. Thus, performing the first two steps of the exact method we use (9) for both the system and the queue and we obtain:

$$\vec{P}_L(z)(A_0 + zA_1) = \lambda(1-z)\vec{e}_1\Phi_S(A_0 + zA_1),$$

and

$$\vec{P}_Q(z)(A_0 + zA_1) = \lambda(1-z)\vec{e}_1\Phi_W(A_0 + zA_1).$$

Using the fact that $S = W + X$ we obtain that

$$\vec{P}_L(z) = \vec{P}_Q(z)\Phi_X(A_0 + zA_1). \tag{22}$$

Applying the fourth step, the number of customers in the queue and the number of customers in the system are also related as follows:

$$\vec{P}_L(z) = (1-z)\vec{H} + z\vec{P}_Q(z), \tag{23}$$

where $\vec{H}$ is an M-vector, with $H_i \triangleq \boldsymbol{P}\{L = 0, R = i\}$, that is given in Proposition 1.

We next follow the fifth step, where we combine (22) and (23) to obtain (19) and (20). Then, we use (9) in order to find the transform of the waiting time distribution and we obtain

$$\vec{e}_1\Phi_W(A_0 + zA_1)(\Phi_X(A_0 + zA_1) - zI) = (1/\lambda)\vec{H}(A_0 + zA_1). \tag{24}$$

We now choose a z such that $A_0 + zA_1$ has M linearly independent eigenvectors and thus it can be written as $A_0 + zA_1 = \Xi(z)\Theta(z)\Xi^{-1}(z)$, where $\Theta(z)$ is the diagonal matrix of the eigenvalues of $A_0 + zA_1$ which we denote by $\theta_i(z)$ for $i = 1, \cdots, M$. Bertsimas and Nakazato [3] have shown that the roots of the equation satisfy $z\alpha(-\theta_i(z)) = 1$, $i = 1, \cdots, M$. The columns of $\Xi(z)$ are the right eigenvectors of $A_0 + zA_1$, which we denote by $\vec{\xi}'(\theta_i(z))$. Moreover,

$$\Phi_W(A_0 + zA_1) = \Xi(z)\Phi_W(\Theta(z))\Xi^{-1}(z),$$

$$\Phi_X(A_0 + zA_1) - zI = \Xi(z)(\Phi_X(\Theta(z)) - zI)\Xi^{-1}(z),$$

and substituting in (24) we obtain

$$\vec{e}_1\Xi(z)\Phi_W(\Theta(z))(\Phi_X(\Theta(z)) - zI) = (1/\lambda)\vec{H}\Xi(z)\Theta(z)$$

or

$$\phi_W(\theta_1(z))\xi_1(\theta_1(z))(\phi_X(\theta_1(z)) - z) = (1/\lambda)\vec{H}\vec{\xi}'(\theta_1(z))\theta_1(z),$$

with $\xi_1(\theta_1(z))$ being the first component of $\vec{\xi}'(\theta_1(z))$ (the previous relation also holds for every eigenvalue $\theta_i(z)$, $i = 1 \cdots M$). Since $z\alpha(-\theta_i(z)) = 1$ we have

$$\phi_W(\theta_1(z)) = K\frac{\theta_1(z)\alpha(-\theta_1(z))}{\lambda(\alpha(-\theta_1(z))\phi_X(\theta_1(z)) - 1)}g(\theta_1(z)),$$

where the function $g(\theta_1(z))$ must have an appropriate form in order to maintain the analytical character of $\phi_W(\theta_1(z))$. Therefore,

$$\phi_W(s) = K\frac{s\alpha(-s)}{\lambda(\alpha(-s)\phi_X(s)-1)}g(s). \tag{25}$$

Since $\phi_W(s)$ is analytic, $g(s) = \alpha_D(-s)\prod_{r=1}^{M-1} x_r - s$, where x_r, $r = 1, \cdots, M-1$ are the $M-1$ roots of the equation $\alpha(-s)\phi_X(s) = 1$, $\mathrm{Re}\,(s) > 0$, and K is a constant such that $\lim_{s\to 0}\phi_W(s) = 1$, which leads to (21).

Remarks.

1. Equation (19) was also obtained in [16] for the more general case of a queue with vacations and non-renewal arrival processes, while (21) is a generalization of the Pollaczek–Khinchine formula for the $M/G/1$ queue. It is interesting to note that (21) could have been obtained using Hilbert factorization techniques [3]. Furthermore, it is the solution to the Volterra integral equations presented in [20] (see also [21]) in the case where the arrivals are MGE. It is remarkable that we were able to derive these formulae just from the distributional laws.
2. The previous results for the $MGE_M/G/1$ system can also be used in a $MGE_M/D/s$ queue (see Remark 4 after Theorem 6).

5. The $GI/G/1$ queue with generalized vacations

In this section we consider a class of $GI/G/1$ queueing models with a single server who is unavailable for occasional intervals of time. Whenever the server is either unavailable or idle we say that he is 'on vacation'. Formally the $GI/G/1$ queue with generalized vacations is defined as follows.

$GI/G/1$ with generalized vacations.

G1. The system satisfies Assumptions A. In particular, as long as the server is busy, customers are served in a non-preemptive FIFO order.

G2. The service mechanism need not be *exhaustive.* When the server begins his vacation he may leave customers behind depending on the service mechanism. We denote by Z_0 the number of customers present in the system in steady state when a vacation interval starts. Z_0 is determined by the service mechanism.

G3. Each vacation interval is distributed as a random variable V and has Laplace transform $\phi_V(s)$. We assume that the number of arrivals during V is *independent* of Z_0.

This system is a generalization of the $GI/G/1$ queue with *exhaustive vacations* considered in Doshi [5], and in Lucantoni *et al.* [16], in which $Z_0 = 0$ (although the analysis in [16] holds for a more general class of non-renewal arrivals). It also generalizes the $M/G/1$ system with generalized vacations considered in [6] (see also the discussion in [26], p. 457) in the sense that it allows more general arrival processes. In some of their results, however, Fuhrmann and Cooper [6] relax

Assumption G3 above, allowing the vacation time to depend on the arrival process. However, in order to prove sharper decomposition results they make exactly the same assumption (their Assumption 6). Our results also generalize the results of Keilson and Servi [13] in two respects: they consider Poisson arrivals and assume exhaustive service $Z_0 = 0$.

Our goal in this section is to illustrate a unified way based on the distributional laws to solve queues with generalized vacations based on the exact method of analysis from Section 4.1. Corollaries of our results include the decomposition results established in [5], [16], [6] and [13].

Examples of the class of $GI/G/1$ queues with generalized vacations that we consider in this section include:

1. The standard $GI/G/1$ queue, if all vacations correspond to idle periods (i.e. $V \to 0$).
2. The $GI/G/1$ queue with *exhaustive vacations,* in which, whenever the server is busy, he serves the system exhaustively, i.e. $Z_0 = 0$.
3. The $GI/G/1$ queue with *gated vacations,* in which the server accepts only those customers who were waiting when the server returned from vacation, i.e. Z_0 is distributed according to the number of customers who arrived after the server returned from vacation.
4. The $GI/G/1$ queue with *limited service,* in which the server serves up to k customers in each visit and then takes a vacation.
5. Queues served in cyclic order considered in [7]. The vacations associated with any particular queue correspond to times when the server is visiting the other queues.

5.1. *Analysis of $MGE_M/G/1$ queue with generalized vacations.* We consider the system in steady state and we let L_v, Q_v, and R_v be the number of customers in the system, the number of customers in the queue and the ATC stage of the arrival process respectively, when a random observer observes the system *with generalized vacations.* Let V^* be the elapsed time since the last vacation began (the backward recurrence time of V). Let B be the event that the server is busy at the time of observation. Obviously B' is the event that the server is on vacation at the time of observation.

Let R_0 and Z_0 be the ATC stage of the arrival process and the number of customers present in the system, when a vacation interval starts. We define $\vec{\zeta}_n \triangleq [\boldsymbol{P}\{Z_0 = n \cap R_0 = m \mid B'\}]_{m=1}^{M}$ and $\vec{\zeta}(z) \triangleq \sum_{n=0}^{\infty} z^n \vec{\zeta}_n$. We view the vector generating function $\vec{\zeta}(z)$ as defining the service mechanism. Our main theorem is as follows.

Theorem 8. *In an $MGE_M/G/1$ system with generalized vacations satisfying Assumptions G1–G3 that has mixed generalized Erlang interarrival times characterized by matrices A_0 and A_1, vacations distributed according to the random variable*

V and service mechanism characterized by the vector generating function $\vec{\zeta}(z)$ *the vector generating function of the number of customers in the queue and in the system is given by*

$$\vec{P}_{Q_v}(z) = (1-\rho)\vec{\zeta}(z)\Phi_{V^*}(A_0 + zA_1)(1-z)(\Phi_X(A_0 + zA_1) - zI)^{-1}, \tag{26}$$

$$\vec{P}_{L_v}(z) = (1-\rho)\vec{\zeta}(z)\Phi_{V^*}(A_0 + zA_1)(1-z)(\Phi_X(A_0 + zA_1) - zI)^{-1}\Phi_X(A_0 + zA_1). \tag{27}$$

Proof. Let S_v, W_v, X be the system, waiting and service time of a customer. Let ρ be the traffic intensity. Because of G1 using the exact method of analysis for overtake free systems and applying (22) for Q_v and L_v we obtain

$$\vec{P}_{L_v}(z) = \vec{P}_{Q_v}(z)\Phi_X(A_0 + zA_1). \tag{28}$$

Our goal is to establish another relation between $\vec{P}_{L_v}(z)$ and $\vec{P}_{Q_v}(z)$. Consider a random observer of the system. Recall that B is the event that the server is busy and B' is the event that the server is on vacation, at the time of observation. By applying Little's law to the server $\boldsymbol{P}\{B\} = \rho$ and $\boldsymbol{P}\{B'\} = 1 - \rho$. By conditioning on the event B we obtain

$$\boldsymbol{P}\{Q_v = n, R_v = i\} = \rho\boldsymbol{P}\{Q_v = n, R_v = i \mid B\} + (1-\rho)\boldsymbol{P}\{Q_v = n, R_v = i \mid B'\}. \tag{29}$$

Conditioning on Z_0, R_0, V^* we obtain

$$\begin{aligned}
&\boldsymbol{P}\{Q_v = n, R_v = i \mid B'\} \\
&\quad = \sum_{k=1}^{M}\sum_{m=0}^{n}\int_0^\infty \boldsymbol{P}\{Q_v = n, R_v = i \mid B', V^* = t, Z_0 = m, R_0 = k\} \\
&\qquad \times \boldsymbol{P}\{Z_0 = m, R_0 = k, V^* = t \mid B'\}\, dt \\
&\quad = \sum_{k=1}^{M}\sum_{m=0}^{n-1} \boldsymbol{P}\{Z_0 = m, R_0 = k \mid B'\}\int_0^\infty a_k(t) * a^{(n-m-1)}(t) * a_1^i(t)\, dF_{V^*}(t) \\
&\qquad + \sum_{k=1}^{M} \boldsymbol{P}\{Z_0 = n, R_0 = k \mid B'\}\int_0^\infty a_k^i(t)\, dF_{V^*}(t),
\end{aligned} \tag{30}$$

where we used the independence of V^* and (Z_0, R_0) (Assumption G3 in the definition of queues with generalized vacations). Let $\vec{B}(z) \triangleq [\sum_{n=0}^{\infty}\boldsymbol{P}\{Q_v = n, R_v = i \mid B\}z^n]_{i=1}^{M}$. Taking generating functions in (29) and using (45) to (30), we obtain

$$\vec{P}_{Q_v}(z) = \rho\vec{B}(z) + (1-\rho)\vec{\zeta}(z)\Phi_{V^*}(A_0 + zA_1).$$

Similarly,

$$\boldsymbol{P}\{L_v = n, R_v = i\} = \rho\boldsymbol{P}\{Q_v = n-1 \cap R_v = i \mid B\} + (1-\rho)\boldsymbol{P}\{Q_v = n \cap R_v = i \mid B\},$$

from where, by taking generating functions, we obtain

$$\vec{P}_{L_v}(z) = \rho z\vec{B}(z) + (1-\rho)\vec{\zeta}(z)\Phi_{V^*}(A_0 + zA_1).$$

Therefore,

$$\vec{P}_{L_v}(z) = z\vec{P}_{Q_v}(z) + (1-z)(1-\rho)\vec{\zeta}(z)\Phi_{V^*}(A_0 + zA_1), \tag{31}$$

which combined with (28) gives (27) and (26).

Remarks.

1. Equation (27), as well as (26), is not formally a decomposition result. It demonstrates, however the contributions of the various characteristics of the system to the system length distribution. The first term $\vec{\zeta}(z)$ represents the effect of the service mechanism used. The second term $\Phi_{V^*}(A_0 + zA_1)$ represents the effect of the vacation, while the third term $(1-\rho)(1-z)(\Phi_X(A_0 + zA_1) - zI)^{-1}\Phi_X(A_0 + zA_1)$ represents the contribution from the underlying $MGE_M/G/1$ queue without vacations.
2. In the case of Poisson arrivals we obtain

$$P_{L_v}(z) = \zeta(z)\phi_{V^*}(\lambda - \lambda z)\frac{(1-\rho)(1-z)\phi_X(\lambda - \lambda z)}{\phi_X(\lambda - \lambda z) - z},$$

which is a formal decomposition result obtained in [6]. The number of customers in the system is distributed as the sum of three independent random variables: (1) the number of customers that are left in the system when a vacation begins, (2) the number of customers that arrive in the system during a vacation period, and (3) the number of customers in a $M/G/1$ queue without vacations. A similar relation is, obviously, obtained for the queue length distribution.

3. Assumption G3 was only used in deriving (30). Without Assumption G3, instead of (31) we would obtain

$$\vec{P}_{L_v}(z) = z\vec{P}_{Q_v}(z) + (1-z)(1-\rho)\vec{P}_{L_v|B'}(z), \tag{32}$$

where $\vec{P}_{L_v|B'}(z)$ is the vector generating function of the number in the system given that the server is on vacation. Combining (32) with (28) we obtain

$$\vec{P}_{L_v}(z) = \vec{P}_{L_v|B'}(z)(1-\rho)(1-z)(\Phi_X(A_0 + zA_1) - zI)^{-1}\Phi_X(A_0 + zA_1),$$

which is the generalization of Proposition 5 in [6].

5.2. *Applications of the $MGE_M/G/1$ with generalized vacations.* In the previous subsection we have been able to derive a formula for the number of customers in the system and in the queue for an $MGE_M/G/1$ queue with generalized vacations as a function of $\vec{\zeta}(z)$. Thus, given that one is able to solve for $\vec{\zeta}(z)$, the queue and system length distributions are fully characterized, and from them the waiting and system time through the distributional laws. In this subsection we will consider some specific applications of the previous analysis that have interesting consequences.

The $MGE_M/G/1$ queue with exhaustive vacations. For the case of exhaustive vacations Theorem 8 implies the decomposition result of Doshi [5] that was also derived for a class of non-renewal arrival processes in [16].

Theorem 9. (Doshi [5]) For the $MGE_M/G/1$ with vacations V under FIFO, the waiting time is the sum of the waiting time of a $MGE_M/G/1$ and the backward recurrence time of the vacation V.

Proof. In this case $Z_0 = 0$ and therefore $\vec{\zeta}(z) = \boldsymbol{P}\{Z = 0, R_0 = i\}_{i=1}^{M} = \vec{R}$, i.e. a vector independent of z. Then (27) can be written, since all the matrices commute,

$$\vec{P}_{L_v}(z) = (1-\rho)\vec{R}(1-z)(\Phi_X(A_0 + zA_1) - zI)^{-1}\Phi_X(A_0 + zA_1)\Phi_{V^*}(A_0 + zA_1).$$

In a regular $MGE_M/G/1$ queue, however, (20) holds, i.e.

$$\vec{P}_L(z) = \vec{H}(1-z)(\Phi_X(A_0 + zA_1) - zI)^{-1}\Phi_X(A_0 + zA_1).$$

But $\vec{P}_{L_v}(1) = \vec{P}_L(1)$, since the ith component of each vector is the probability that the ATC is in stage i which is independent of the vacation. Taking limits as $z \to 1$ in the two previous equations we obtain

$$(1-\rho)\vec{R}\Phi_{V^*}(A_0 + A_1) = \vec{H}.$$

Therefore, in an $MGE_M/G/1$ with exhaustive vacations

$$\begin{aligned}\vec{P}_{L_v}(z) = {} & \vec{H}\Phi_{V^*}(A_0 + A_1)^{-1}(1-z)(\Phi_X(A_0 + zA_1) - zI)^{-1} \\ & \times \Phi_X(A_0 + zA_1)\Phi_{V^*}(A_0 + zA_1),\end{aligned} \tag{33}$$

where the vector $\vec{H}$ is computed in Proposition 1. Notice that (33) offers a complete solution of the $MGE_M/G/1$ queue with exhaustive vacations.

Following exactly the same approach leading to (25) in the proof of Theorem 7 we obtain that

$$\phi_{W_v}(s) = K\frac{s\phi_{V^*}(s)\alpha(-s)}{\lambda(\alpha(-s)\phi_X(s) - 1)}g(s) = \phi_W(s)\phi_{V^*}(s),$$

i.e. $W_v \overset{\mathrm{d}}{=} W \oplus V^*$.

The $MGE_M/G/1$ queue with gated vacations. In a gated vacation system our goal is to find $\vec{\zeta}(z)$. For this reason we define the following random variables.

Let J be the time the server spends in the system immediately after he returns from vacation until he starts a new one. Let $F_J(t) = \boldsymbol{P}\{J \leqq t\}$ and $\phi_J(s)$ be the Laplace transform of J. Let R_J be the ATC stage of the arrival process and N be the number of the customers that the server finds at the system just after the end of the vacation. We define $\vec{R}_J = \boldsymbol{P}\{R_J = m\}_{m=1}^{M}$ and $N(z) = \boldsymbol{E}[z^N]$. Finally, we define also the vectors $\vec{N}_n = \boldsymbol{P}\{N = n \cap R_J = m\}_{m=1}^{M}$ and $\vec{N}(z) = \sum_{n=0}^{\infty} z^n \vec{N}_n$. Note that $\vec{R}_J = \vec{N}(1)$.

From the definition of the service mechanism in a gated system, Z_0 is distributed according to the number of customers who arrived during J, thus:

$$\sum_{n=0}^{\infty} z^n \boldsymbol{P}\{Z_0 = n, R_0 = k \mid R_J = m\}$$
$$= \int_0^{\infty} a_m^k(t)\, dF_J(t) + \sum_{n=1}^{\infty} z^n \int_0^{\infty} a_m(t) * a^{(n-1)}(t) * a_1^k(t)\, dF_J(t),$$

which leads to

$$\sum_{n=0}^{\infty} z^n \boldsymbol{P}\{Z_0 = n, R_0 = k\}$$
$$= \sum_{m=1}^{M} \boldsymbol{P}\{R_J = m\}\left[\int_0^{\infty} a_m^k(t)\, dF_J(t) + \sum_{n=1}^{\infty} z^n \int_0^{\infty} a_m(t) * a^{(n-1)}(t) * a_1^k(t)\, dF_J(t)\right],$$

which in matrix notation becomes

$$\vec{\zeta}(z) = \vec{N}(1)\Phi_J(A_0 + zA_1). \tag{34}$$

Furthermore, the time interval J lasts as long as the server is servicing the N customers he finds upon his arrival. So

$$\phi_J(s) = N(\phi_X(s)). \tag{35}$$

Finally, we need to evaluate $N(z)$ from the characteristics of the system. Recalling the definition of the gated vacation system we see that N includes the customers that the server left behind in the system before starting his vacation as well as the customers that arrived during the vacation interval. Therefore, for $n \geqq 1$, $\boldsymbol{P}\{N = n, R_J = l\} = \sum_{k=0}^{n} \sum_{m=1}^{M} \boldsymbol{P}\{Z_0 = k, R_0 = m\} \int_0^{\infty} a_m(t) * a^{(n-k-1)}(t) * a_1^l(t)\, dF_V(t)$. Taking generating functions:

$$\vec{N}(z) = \vec{\zeta}(z)\Phi_V(A_0 + zA_1). \tag{36}$$

By combining (34), (35) and (36) we have

$$\vec{\zeta}(z) = \vec{\zeta}(1)\Phi_V(A_0 + A_1)\Phi_J(A_0 + zA_1), \tag{37}$$

where

$$\phi_J(s) = \vec{\zeta}(\phi_X(s))\Phi_V(A_0 + \phi_X(s)A_1)\vec{1}. \tag{38}$$

Equations (37) and (38) fully characterize $\vec{\zeta}(z)$ as we can solve for all moments. Moreover, if we use Theorem 8 and the distributional laws we can fully characterize the system.

Remark. Notice that in the Poisson case the recursion formula takes the form

$$\zeta(z) = \zeta(\phi_X(\lambda - \lambda z))\phi_V(\lambda - \lambda\phi_X(\lambda - \lambda z)).$$

6. Priority queues

Priority queues are important in communication and manufacturing systems where jobs of different significance need to be serviced. In addition, in several applications strict priority rules (for example, the so-called $c\mu$-rule) minimize a weighted combination of expected waiting times. It is therefore important to be able to analyze priority queues.

We consider single server priority queueing systems with mixed generalized Erlang arrivals, in which there are two distinct customer classes, numbered 1 and 2. Customers of class 1 have priority over those of class 2. Let $a(t)$, $b(t)$ be the pdf of the interarrival time for the high priority class 1 and the low priority class 2 respectively. We assume that they are mixed generalized Erlangs of order M_1, M_2 respectively. We denote with $1/\lambda_1$ and $1/\lambda_2$ the means of the arrival processes. The two classes have different (general) service time distributions with means $\boldsymbol{E}[X_1]$ and $\boldsymbol{E}[X_2]$, and they are served by a single server.

We assume that within the same class customers are served in a FIFO order. Although priority queues allow overtaking among classes, within the same class no overtaking can take place and therefore the distributional laws are applicable. In this section we use the distributional laws to derive the distributions of various performance measures. Our results generalize earlier work of Keilson and Servi [13] for Poisson arrivals.

We consider different types of priorities (preemptive repeat, preemptive resume, non-preemptive). The type of priority used does not affect the service time of class 1, but affects the service time of class 2. In order to develop a generic model to analyze priority queues in a unified way, we define the *effective service time*, G_i, $i = 1, 2$, as the time from the beginning of service until the customer of class i completes service ($G_1 = X_1$, regardless of the priority rule used). We can visualize the effective service time as the time spent in a *service box*. The service may be interrupted and resumed from where it was left or may start over, but the customer is assumed to stay in the service box until he is completely served. In this setting, the time in queue refers to the time from the arrival of the customer until the customer enters the service box.

The section is organized as follows. In Section 6.1 we find the effective service time distribution in various preemptive systems as a function of the busy period matrix. In Section 6.2 we analyze systems with preemptive priorities, while in Section 6.3 we analyze systems with non-preemptive priorities.

6.1. *Effective service time distribution in preemptive systems.* According to preemptive disciplines, whenever a high priority customer finds a lower priority customer in service, he interrupts the service in progress and starts his own immediately. Once there is no higher priority customer left in the system, the interrupted customer reenters service and depending upon the manner in which he is

serviced on his reentrance, the preemptive discipline can be further broken down into the following three categories.

(i) *Preemptive resume discipline.* Under this discipline the interrupted customer continues his service from the point of interruption.
(ii) *Preemptive repeat different discipline.* Under this discipline the interrupted customer continues his service by resampling.
(iii) *Preemptive repeat identical discipline.* Under this discipline the interrupted customer continues his service without resampling.

Each of these three preemptive disciplines is going to affect the effective service time of class 2 customers. In this section we calculate the effective service time in all the three preemptive categories as a function of the class 1 busy period matrix, $\Gamma_1(s)$.

We define random variables G_2^{ij}, $i, j = 1, \cdots, M_1$, to denote the effective service time of a class 2 customer such that the $ATC_1 = j$ when the class 2 customer finishes service given that the $ATC_1 = i$ when this class 2 customer started service. Let $\phi_{G_2^{ij}}(s)$ be the Laplace transform of G_2^{ij} and let $G_2(s)$ denote the matrix with elements $\phi_{G_2^{ij}}(s)$. Our goal in this section is to compute the matrix $G_2(s)$.

Preemptive resume discipline.

Proposition 2. *In a single server system with two priority classes each of which satisfies Assumptions A and has mixed generalized Erlang interarrival times characterized by matrices A_0, A_1 and B_0, B_1 respectively, the effective service time of the class* 2 *customers for the preemptive resume discipline is given as follows:*

$$G_2(s) = \Phi_{X_2}(A_0 + A_1\Gamma_1(s) + sI).$$

Proof. According to the preemptive resume discipline, whenever a low priority customer service is interrupted, the duration of the interruption is exactly the duration of a high priority customer busy period. Furthermore, due to the characteristics of the mixed generalized Erlang arrival process we condition on R_1^{bs}, the ATC_1 stage immediately before a low priority customer enters service. Let $\phi_{G_2^{ki}}(s)$ be the Laplace transform of the effective service time of a class 2 customer that ends leaving the $ATC_1 = i$ given that it started with the $ATC_1 = k$. Then

$$\begin{aligned}\boldsymbol{E}[e^{-sG_2^{ki}} \mid X_2 = x] = e^{-sx}\Big\{a_k^i(x) + \sum_{j_1=1}^{M_1} [\Gamma_1(s)]_{1,j_1} a_k(x) * a_{j_1}^i(x) \\ + \sum_{j_1=1}^{M_1}\sum_{j_2=1}^{M_1} [\Gamma_1(s)]_{1,j_1}[\Gamma_1(s)]_{1,j_2} a_k(x) * a_{j_1}(x) * a_{j_2}^i(x) + \cdots\Big\},\end{aligned}$$

where the first of the right-hand side terms represents the probability that there are no interruptions during the regular service time of the low priority customer, the

second the probability of having just one interruption, where we have to take into account the ATC stage of the high priority customer at the end of the type 1 busy period, and so on. By writing the previous formula in matrix notation we obtain

$$\boldsymbol{E}[e^{-sG_2^{ki}} \mid X_2 = x] = e^{-sx}\vec{e}_k \begin{bmatrix} a_1^1(x) & \cdots & a_1^{M_1}(x) \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{M_1}^{M_1}(x) \end{bmatrix} \vec{e}_i' + e^{-sx}\vec{e}_k \sum_{n=1}^{\infty} \begin{pmatrix} a_1(x) \\ \vdots \\ a_{M_1}(x) \end{pmatrix} * (a_1(x)[\Gamma_1(s)]_{1,1} + \cdots + a_{M_1}(x)[\Gamma_1(s)]_{1,M_1})^{(n-1)} * ([\Gamma_1(s)]_{1,1}\vec{a}_1(x) + \cdots + [\Gamma_1(s)]_{1,M_1}\vec{a}_{M_1}(x))\vec{e}_i'.$$

Using (45) we obtain

$$\boldsymbol{E}[e^{-sG_2^{ki}} \mid X_2 = x] = e^{-sx}\vec{e}_k e^{-(A_0 + A_1\Gamma_1(s))x}\vec{e}_i'.$$

Therefore,

$$\boldsymbol{E}[e^{-sG_2^{ki}}] = \vec{e}_k \Phi_{X_2}(A_0 + A_1\Gamma_1(s) + sI)\vec{e}_i',$$

and hence,

$$G_2(s) = \Phi_{X_2}(A_0 + A_1\Gamma_1(s) + sI).$$

Remark. For the Poisson case we obtain $\phi_{G_2}(s) = \phi_{X_2}(\lambda_1 - \lambda_1\gamma(s) + s)$, which is in agreement with Jaiswal [11].

Preemptive repeat disciplines. Let

$$\vec{a}(t) = (a_1(t), \cdots, a_k(t), \cdots, a_{M_1}(t)) \qquad \text{and} \qquad A(t) = \begin{bmatrix} \vec{a}_1(t) \\ \vdots \\ \vec{a}_{M_1}(t) \end{bmatrix}.$$

Proposition 3. *The effective service time G_2 for the preemptive repeat discipline under the assumptions of Proposition* 1 *is given as follows.*

- *In the case of the preemptive repeat different discipline*

$$G_2(s) = \int_0^\infty A(x)e^{-sx} f_{X_2}(x)\,dx \left[I - \int_0^\infty f_{X_2}(x) \int_0^x \vec{a}'(y)e^{-sy}\,dy\,dx\vec{e}_1\Gamma_1(s)\right]^{-1}.$$

- *In the case of the preemptive repeat identical discipline*

$$G_2(s) = \int_0^\infty A(x)\left[I - \int_0^x \vec{a}'(y)e^{-sy}\,dy\vec{e}_1\Gamma_1(s)\right]^{-1} e^{-sx} f_{X_2}(x)\,dx.$$

Proof. The underlying experiment is the following. Assume that a class 2 customer enters the service facility at τ_0 and his service time is given by a value of the r.v. X_2. At the moment he enters service there are no type 1 customers in the system and the $ATC_1 = k$. There are two possibilities for the remaining time until the next arrival of the high priority arrival process:

- either it is greater than the selected value of X_2 and in this case $G_2^{ki} = X_2$, where i is the stage of the ATC_1 when the low priority finishes service; or

- it is less than the selected value of X_2 and at the moment that the next type 1 customer arrives the service of the type 2 customer is interrupted and it starts over with a *new value* of the r.v. X_2 as soon as the busy period initialized by the type 1 customer is over for the preemptive repeat *different* discipline *or* with the *same value* of the r.v. X_2 for the preemptive repeat *identical* discipline.

So for the *repeat different* case, conditioning on X_2 we obtain

$$\boldsymbol{E}[e^{-sG_2^{ki}} \mid X_2 = x] = a_k^i(x)e^{-sx} + \int_0^x a_k(y)e^{-sy}\,dy\vec{e}_1\Gamma_1(s)G_2(s)\vec{e}_i'.$$

Thus,

$$\phi_{G_2^{ki}}(s) = \int_0^\infty a_k^i(x)e^{-sx}f_{X_2}(x)\,dx + \int_0^\infty f_{X_2}(x)\int_0^x a_k(y)e^{-sy}\,dy\,dx\vec{e}_1\Gamma_1(s)G_2(s)\vec{e}_i'.$$

And in matrix form:

$$G_2(s) = \int_0^\infty A(x)e^{-sx}f_{X_2}(x)\,dx\left[I - \int_0^\infty f_{X_2}(x)\int_0^x \vec{a}'(y)e^{-sy}\,dy\,dx\vec{e}_1\Gamma_1(s)\right]^{-1}.$$

Finally, for the *repeat identical* case:

$$G_2(s) = \int_0^\infty A(x)\left[I - \int_0^x \vec{a}'(y)e^{-sy}\,dy\vec{e}_1\Gamma_1(s)\right]^{-1} e^{-sx}f_{X_2}(x)\,dx.$$

In the case of Poisson arrivals we can obtain the results of Jaiswal [11], namely:

$$\phi_{G_2}(s) = \frac{\phi_{X_2}(s+\lambda_1)}{1 - \dfrac{\lambda_1}{s+\lambda_1}(1 - \phi_{X_2}(s+\lambda_1))\gamma_1(s)},$$

$$\phi_{G_2}(s) = \int_0^\infty \frac{e^{-(s+\lambda_1)x}}{1 - \dfrac{\lambda_1}{s+\lambda_1}(1 - e^{-(s+\lambda_1)x})\gamma_1(s)}\,dF_{X_2}(x),$$

for the preemptive repeat different and the preemptive repeat identical discipline, respectively.

6.2. *Preemptive priorities.* In this section we analyze a generic preemptive discipline in terms of the distribution of the effective service time. In this way we are able to analyze all preemptive disciplines we consider in a unified way.

Let L_i, Q_i, S_i, W_i, R_i, $i = 1, 2$ be the system and queue length, system and waiting time and ATC stage of the arrival process, respectively, of class $i = 1, 2$. Notice that the low priority customer that may be in the service box without being served *is not taken into account* in the number of low priority customers in the queue.

Let L_i^a, Q_i^a and R_i^a be the number of customers of class i in the system, in the queue and the ATC stage of class i, respectively, just before an *arrival* of a class 1 customer.

High priority customers. As long as the discipline is preemptive the high priority customers see a usual $MGE_{M_1}/G/1$ queue. Therefore, Theorem 8 can be used to find the distributions of L_1, Q_1, S_1, W_1.

Low priority customers. We will apply the exact method of analysis discussed in Section 4. We start by proving a relation between the vector z-transform (generating function) of the number of customers in the *system* from the low priority class, $\vec{P}_{L_2}(z)$, and the vector generating function of the number of customers in the *queue* from the low priority class, $\vec{P}_{Q_2}(z)$, i.e. implementing the fourth step of the method.

Let us first introduce the necessary notation. Let E_2 be the number of class 2 customers in queue given that no class 2 customer is in the service box but the system is not empty. Let Δ_2 be the number of class 2 customers in queue given that there is a class 2 customer in the service box. We introduce the vector generating functions

$$\vec{P}_{E_2}(z) \triangleq \sum_{n=0}^{\infty} z^n [\boldsymbol{P}\{E_2 = n \cap R_2 = i\}]_{i=1}^{i=M_2},$$

and

$$\vec{P}_{\Delta_2}(z) \triangleq \sum_{n=0}^{\infty} z^n [\boldsymbol{P}\{\Delta_2 = n \cap R_2 = i\}]_{i=1}^{i=M_2}.$$

Furthermore, let $\vec{h}$ and $\vec{k}$ be M_2 vectors such that $h_i \triangleq \boldsymbol{P}\{R_2 = i, L_1 = 0, L_2 = 0\}$ and $k_r \triangleq \boldsymbol{P}\{R_2^a = r, L_1^a = 0, L_2^a = 0\}$. Finally, let $\Gamma_{1,i}^*$ be the backward recurrence time (age) of a class 1 busy period that ended while $ATC_1 = i$. Then the Laplace transform of Γ_1^*, $\gamma_1^*(s)$ is given by

$$\gamma_1^*(s) = (1 - [\vec{e}_1 \Gamma_1(s) \vec{1}'])/s\boldsymbol{E}[\Gamma_1].$$

We also introduce the traffic intensities $\rho_i = \lambda_i \boldsymbol{E}[X_i]$, $\rho = \rho_1 + \rho_2$ and we define $p_{\Delta_2} \triangleq P\{\text{a class 2 customer is in the service box}\}$. We now prove the following proposition.

Proposition 4. *In a preemptive queueing system with two priority classes each of which satisfies Assumptions A and has mixed generalized Erlang interarrival times characterized by matrices A_0, A_1 and B_0, B_1, respectively, the following relation holds:*

$$\vec{P}_{L_2}(z) = (1 - z)\vec{h} + (1 - z)(\rho - p_{\Delta_2})\vec{k}\Phi_{\Gamma_1^*}(B_0 + zB_1) + z\vec{P}_{Q_2}(z), \tag{39}$$

where the vectors $\vec{h}$ and $\vec{k}$ are given by (48) *in Appendix B.*

Proof. We start by noticing that at a random observation time there are three possibilities for the state of the system: (a) the system is empty (with probability $1 - \rho$), (b) the system is not empty but there is no class 2 customer in the service box (with probability $\rho - p_{\Delta_2}$), and (c) the system is not empty and there is a class 2 customer in the service box (with probability p_{Δ_2}).

Conditioning on those three events and recalling that h_i is the probability that the system is empty and $R_2 = i$, we have that

$$\vec{P}_{Q_2}(z) = \vec{h} + p_{\Delta_2}\vec{P}_{\Delta_2}(z) + (\rho - p_{\Delta_2})\vec{P}_{E_2}(z),$$

and

$$\vec{P}_{L_2}(z) = \vec{h} + p_{\Delta_2} z\vec{P}_{\Delta_2}(z) + (\rho - p_{\Delta_2})\vec{P}_{E_2}(z).$$

Next we need to evaluate the vector generating function $\vec{P}_{E_2}(z)$. Notice that since we assumed a preemptive discipline, class 1 customers are not influenced by the fact that there is no low priority customer in the service box. In order for a random observer to see $n \geqq 1$ class 2 customers given that there is no class 2 customer in the service box, he has to arrive during a class 1 busy period. Therefore, if we denote by k_r the probability that the high priority customer who initialized the last class 1 busy period found, upon his arrival, the class 2 customer in stage r we have, for $n \geqq 1$,

$$\boldsymbol{P}\{E_2 = n, R_2 = i\} = \sum_{r=1}^{M_2} k_r \int_0^\infty b_r(t) * b^{(n-1)}(t) * b_1^j(t)\, dF_{\Gamma_1^*}(t).$$

Similarly,

$$\boldsymbol{P}\{E_2 = 0, R_2 = i\} = \sum_{r=1}^{M_2} k_r \int_0^\infty b_r^j(t)\, dF_{\Gamma_1^*}(t),$$

where $F_{\Gamma_1^*}(t)$ is the cdf of the forward recurrence time of a class 1 busy period. Taking generating functions (39) follows.

Notice that in the previous proposition the probability p_{Δ_2} is not known. We will however calculate it implicitly in the next theorem.

Theorem 10. *In a preemptive queueing system with two priority classes each of which satisfies Assumptions A and has mixed generalized Erlang interarrival times characterized by matrices A_0, A_1 and B_0, B_1 respectively, the Laplace transform of the waiting and system time of the low priority customers satisfy*

$$\Phi_{S_2}(B_0 + zB_1) - z\Phi_{W_2}(B_0 + zB_1) = \frac{1}{\lambda_2}[\vec{h} + (\rho - p_{\Delta_2})\vec{k}\Phi_{\Gamma_1^*}(B_0 + zB_1)](B_0 + zB_1)^{-1},$$

and

$$\Phi_{S_2}(sI) = \sum_{i=1}^{M_1} \vec{Y}\Phi_{W_2}(sI + A_0 + A_1)\vec{e}_i^{\,\prime}\Phi_{G_2^i}(sI).$$

Furthermore, the number of low priority customers in the queue is given as

$$\vec{P}_{Q_2}(z)(B_0 + zB_1) = \lambda_2(1 - z)\vec{e}_1\Phi_{W_2}(B_0 + zB_1),$$

where Γ_1^* denotes the age of the high priority busy period, $Y_k = (\lambda_1/\lambda_{1,k})\prod_{i=1}^{k-1}(1 - p_{1,i})$, the vectors $\vec{h}$ and $\vec{k}$ are given by (48) *in Appendix B and p_{Δ_2} is calculated by insisting that* $\lim_{z\to 1}\vec{P}_{Q_2}(z)\vec{1}' = 1$.

Proof. Since we assumed that among each class customers are served in a FIFO basis, the distributional laws are applicable. Therefore, performing the first two steps of the exact method we use (9) for both the queue plus the server and just the queue to obtain:

$$\vec{P}_{L_2}(z)(B_0 + zB_1) = \lambda_2(1-z)\vec{e}_1\Phi_{S_2}(B_0 + zB_1),$$

$$\vec{P}_{Q_2}(z)(B_0 + zB_1) = \lambda_2(1-z)\vec{e}_1\Phi_{W_2}(B_0 + zB_1).$$

Recall that in evaluating the number of customers in the queue we are not counting the customer in the service box and similarly the waiting time does not account for the time spent in the service box.

We proceed now to implement the third step, namely obtain a relation between S_2 and W_2. It is important to notice that, in this step, we need to use the notion of the *effective* service time which depends on the dynamics of the high priority customers. To be more specific, recall that we define in Section 6.1 the random variable G_2^{ij}, $i, j = 1, \cdots, M_1$, to be the effective service time of a class 2 customer such that the $ATC_1 = j$ when the class 2 customer finishes service given that the $ATC_1 = i$ when this class 2 customer started service. In Section 6.1 we actually calculated the Laplace transforms of G_2^{ij}, for all i, j. For the rest of our analysis let us also define G_2^i, $i = 1, \cdots, M_1$, to be the effective service time of a class 2 customer given that $ATC_1 = i$ when this class 2 customer started service. Clearly, the Laplace transform of G_2^i, $\phi_{G_2^i}(s) = \vec{e}_i G_2(s)\vec{1}'$, using the notation of Section 6.1.

Furthermore, we define R_1^{a2} to be the stage of the ATC_1 at an arrival epoch of class 2 customers and $\vec{Y}$, such that $Y_k \triangleq \boldsymbol{P}\{R_1^{a2} = k\}$ for $k = 1, \cdots, M_1$. Due to the independence of the arrival processes we can easily see that $Y_k = \boldsymbol{P}\{R_1 = k\} = (\lambda_1/\lambda_{1,k}) \prod_{i=1}^{k-1} (1 - p_{1,i})$, where the second equality follows from the properties of the MGE_M process.

Let us now proceed to relate S_2 and W_2 using G_2^i.

$$\boldsymbol{E}[e^{-sS_2} \mid R_1^{a2} = k, W_2 = w, ATC_1 = i \text{ at the end of the waiting}] = e^{-sw} \int_0^\infty e^{-sx}\, dF_{G_2^i}(x).$$

Unconditioning we get that

$$\Phi_{S_2}(sI) = \sum_{i=1}^{M_1} \vec{Y}\Phi_{W_2}(sI + A_0 + A_1)\vec{e}_i'\Phi_{G_2^i}(sI).$$

Applying the fourth step, the number of customers in the queue and the number of customers in the system are also related from Proposition 4 as follows:

$$\vec{P}_{L_2}(z) = (1-z)\vec{h} + (1-z)(\rho - p_{\Delta_2})\vec{k}\Phi_{\Gamma_1^*}(B_0 + zB_1) + z\vec{P}_{Q_2}(z),$$

where the vectors $\vec{h}$ and $\vec{k}$ are given in Appendix B.

We next follow the fifth step, where we combine the above equations to prove the theorem.

Remarks.

1. Having found W_2 we can use the rest of the relations to obtain S_2, L_2 and Q_2.
2. In the case of Poisson arrival processes we have that $\phi_{S_2}(s) = \phi_{W_2}(s)\phi_{G_2}(s)$, so that Step 5 yields

$$G_{L_2}(z) = \frac{1}{\lambda_2}(1 - p_{\Delta_2})(1 - \rho_1)\frac{\lambda_2(1-z) + \lambda_1(1 - \sigma_1(\lambda_2 - \lambda_2 z))}{\phi_{G_2}(\lambda_2 - \lambda_2 z) - z},$$

which is exactly the relation obtained in Keilson and Servi [13] using a different derivation. The probability p_{Δ_2} can be obtained either by requiring $\lim_{z\to 1} G_{Q_2}(z) = 1$, which in this case leads to $p_{\Delta_2} = \lambda_2 \boldsymbol{E}[G_2]$ or by applying Little's law in the service box.

6.3. *Non-preemptive priorities.* In this section we analyze the single server priority system under a non-preemptive discipline, where an arriving high priority customer that finds a low priority customer in service *does not* interrupt the service in progress. Therefore, the effective service time for class 2 customers under a non-preemptive priority discipline is $G_2 = X_2$. Furthermore, as no customer stays in the service box unless he is actually being served, the waiting time is in this case defined without ambiguity, exactly as in the case of a single $MGE_M/G/1$ queue. We will first calculate the distribution of the number of class 1 customers in the queue and in the system.

High priority customers. Due to the fact that we do not allow preemption, the number of class 1 customers in the queue as well as their waiting time are influenced by the possible existence of a class 2 customer in the service facility. Let R_1^{bs} be the state of the ATC_1, just before a class 2 customer enters service.

Let B_i be the event that the server is busy servicing a class i customer at a random time of observation.

Let Δ_1 be the number of class 1 customers in queue given that there is a class 1 customer in service. We introduce the vector generating function $\vec{P}_{\Delta_1}(z) \triangleq \sum_{z=0}^{\infty} z^n [\boldsymbol{P}\{\Delta_1 = n \cap R_1 = i\}]_{i=1}^{i=M_1}$, and the scalar generating function $G_{\Delta_1}(z) = \sum_{z=0}^{\infty} z^n \boldsymbol{P}\{\Delta_1 = n\}$. We also introduce the row vectors $\vec{g}$ and $\vec{f}$, such that $g_i \triangleq \boldsymbol{P}\{L_1 = 0,\ L_2 = 0,\ R_1 = i\}$ and $\pi_r \triangleq \boldsymbol{P}\{R_1^{bs} = r\}$.

Theorem 11. *In a non-preemptive queueing system with two priority classes each of which satisfies Assumptions A and has mixed generalized Erlang interarrival times characterized by matrices A_0, A_1 and B_0, B_1 respectively, the vector generating function of the number of class* 1 *customers in the queue and in the system is given as a function of the system characteristics as follows*:

$$\vec{P}_{Q_1}(z) = (1-z)[\rho_2 \vec{\pi} \Phi_{X_2^*}(A_0 + zA_1) + \vec{g}][\Phi_{X_1}(A_0 + zA_1) - zI]^{-1}, \tag{40}$$

$$\vec{P}_{L_1}(z) = (1-z)[\rho_2 \vec{\pi} \Phi_{X_2^*}(A_0 + zA_1) + \vec{g}][\Phi_{X_1}(A_0 + zA_1) - zI]^{-1} \Phi_{X_1}(A_0 + zA_1), \tag{41}$$

where $\vec{g}$ is given by (49) *in Appendix B and $\vec{\pi}$ satisfies:*

$$\rho_2 \vec{\pi} \Phi_{X_2^*}(A_0 + A_1) = \vec{H} - \vec{g}, \tag{42}$$

where $\vec{H}$ is given in Proposition 1.

Proof. From the vector distributional law (22) we have

$$\vec{P}_{L_1}(z) = \vec{P}_{Q_1}(z) \Phi_{X_1}(A_0 + zA_1). \tag{43}$$

We should establish a second relation between $\vec{P}_{L_1}(z)$ and $\vec{P}_{Q_1}(z)$. Consider a random observer of the system and let B_i be the event that the server is busy servicing a class i customer at the time of observation. By applying Little's law to the server $\boldsymbol{P}\{B_i\} = \rho_i$ and by conditioning on the events B_i we have, for $n \geqq 1$,

$$\boldsymbol{P}\{Q_1 = n, R_1 = i\} = \rho_1 \boldsymbol{P}\{Q_1 = n, R_1 = i \mid B_1\} + \rho_2 \boldsymbol{P}\{Q_1 = n, R_1 = i \mid B_2\},$$

or, by using the definition of Δ_1,

$$\boldsymbol{P}\{Q_1 = n, R_1 = i\} = \rho_1 \boldsymbol{P}\{\Delta_1 = n, R_1 = i\} + \rho_2 \boldsymbol{P}\{Q_1 = n, R_1 = i \mid B_2\},$$

and for $n = 0$ we also have

$$\boldsymbol{P}\{Q_1 = 0, R_1 = i\} = \rho_1 P\{\Delta_1 = 0, R_1 = i\} + \rho_2 \boldsymbol{P}\{Q_1 = 0, R_1 = i \mid B_2\} + \boldsymbol{P}\{L_1 = 0, L_2 = 0, R_1 = i\},$$

or equivalently

$$\boldsymbol{P}\{Q_1 = 0, R_1 = i\} = \rho_1 \boldsymbol{P}\{\Delta_1 = 0, R_1 = i\} + \rho_2 \boldsymbol{P}\{Q_1 = 0, R_1 = i \mid B_2\} + g_i.$$

Furthermore, if we denote by π_r the probability that the $ATC_1 = r$ just before a type 2 customer enters service we have that, for $n \geqq 1$,

$$\boldsymbol{P}\{Q_1 = n, R_1 = i \mid B_2\} = \sum_{r=1}^{M_1} \pi_r \int_0^\infty a_r(t) * a^{(n-1)}(t) * a_1^i(t)\, dF_{X_2^*}(t),$$

and

$$\boldsymbol{P}\{Q_1 = 0, R_1 = i \mid B_2\} = \sum_{r=1}^{M_1} \pi_r \int_0^\infty a_r^i(t)\, dF_{X_2^*}(t).$$

By taking generating vector functions we get

$$\vec{P}_{Q_1}(z) = \rho_1 \vec{P}_{\Delta_1}(z) + \rho_2 \vec{\pi} \Phi_{X_2^*}(A_0 + zA_1) + \vec{g}.$$

Using the same analysis for the number of customers in the system we also obtain

$$\vec{P}_{L_1}(z) = \rho_1 z \vec{P}_{\Delta_1}(z) + \rho_2 \vec{\pi} \Phi_{X_2^*}(A_0 + zA_1) + \vec{g}.$$

Combining the last two equations we have

$$\vec{P}_{L_1}(z) = z\vec{P}_{Q_1}(z) + \rho_2(1 - z)\vec{\pi} \Phi_{X_2^*}(A_0 + zA_1) + (1 - z)\vec{g}. \tag{44}$$

From (43) and (44) we obtain (40) and (41). Finally, we need to calculate the vectors $\vec{\pi}$ and $\vec{g}$. Vector $\vec{g}$, where, $g_i \triangleq \boldsymbol{P}\{L_1 = 0, L_2 = 0, R_1 = i\}$, is calculated in Appendix B. To calculate π we recall that in a regular $MGE_M/G/1$ queue (20) holds, namely $\vec{P}_L(z) = (1-z)\vec{H}(\Phi_{X_1}(A_0 + zA_1) - zI)^{-1}\Phi_{X_1}(A_0 + zA_1)$. But $\vec{P}_{L_1}(1) = \vec{P}_L(1)$, since the ith component of this vector represents the probability that the ATC of the arrival process of class 1 is in stage i. Thus by taking the limits as $z \to 1$, we get (42).

Remarks.

1. Using (40) and (41) as well as the vector distributional law one can easily calculate the waiting time distributions, as in the case of the single $MGE_M/G/1$ queue.
2. Note that once again for Poisson arrivals (40) takes the form

$$G_{Q_1}(z) = (1-z)[\rho_2 \phi_{X_2^*}(\lambda_1 - \lambda_1 z) + (1 - \rho_1 - \rho_2)](\phi_{X_1}(\lambda_1 - \lambda_1 z) - z)^{-1},$$

which is exactly the result obtained in [13].

Low priority customers. The waiting time of the low priority customer equals in distribution the total unfinished work in the system at the moment of his arrival subject to generalized Erlang interruptions, corresponding to class 1 arrivals. As the work in the system as well as the distribution and duration of the interruptions *do not* depend on whether we give non-preemptive or preemptive resume priority to the class 1 customers we can conclude that the waiting time distribution for the low priority customer under a non-preemptive policy is the same as the waiting time under a preemptive resume policy (see [13]). However, this is not true for the waiting time in the system because of the notion of the effective service time that we used in the preemptive priority analysis. Nevertheless, we can calculate all the distributions of interest by using the distributional laws as well as the relation $S_2 = W_2 + X_2$.

7. Concluding remarks

We have demonstrated that overtake-free systems can be analyzed in a unified way through the distributional laws, which we believe deserve a more prominent place in queueing theory. More than providing a method of analysis for a class of systems, the paper identified a subdivision of queueing theory into overtake free systems, which can be analyzed using distributional laws, but are unfortunately a small subset of the systems encountered in applications, which allow overtaking, and therefore are not analyzable directly through the techniques of this paper.

In the case of overtake free systems, we showed several insights and new results that can be obtained. One which we consider particularly satisfying is the derivation

of heavy traffic results (usually derived using diffusion methods) and exact results can be achieved in a unified way using the asymptotic and exact method of analysis based on the distributional laws.

The distributional laws being special cases of $H = \lambda G$ provide further evidence that the law $H = \lambda G$ provides the right set of laws at least for overtake free systems. The major open problem is to identify queueing laws for systems that allow overtaking, which lead to a complete solution. This is a challenging but important problem because it includes well known open problems as special cases ($GI/G/s$, queueing networks, etc.). A solution to this problem will lead, however, to a more complete theory of queues and is likely to provide very valuable new insights.

Appendix A

We present the proof of Proposition 1. We will use a generalization of the classical sub-busy period decomposition argument for the evaluation of the busy period for the $M/G/1$ queue (Takács [24]). The duration of a busy period is invariant under the service discipline provided that the server is always busy if there are customers present. We may, therefore, use the last-come-first-serve (LCFS) service discipline. Let $\Gamma_{i,m}$ be the random variable that represents the duration of the sub-busy period that ends with the $ATC = m$ given that it started with the $ATC = i$. This definition is useful for the decomposition of the busy period into sub-busy periods. Let R^{as} be the ATC stage occupied by the customer just after the first customer of the sub-busy period is served. Let $N_i(x)$ be the number of arrivals during x given that the $ATC = i$ at the beginning. Then, conditionally on the event $U = \{R^{\text{as}} = j, X = x, N_i(x) = n\}$, we obtain the following decomposition, for $n \geqq 1$:

$$\boldsymbol{E}[\mathrm{e}^{-s\Gamma_{i,m}} \mid R^{\text{as}} = j, X = x, N_i(x) = n] = \boldsymbol{E}\left[\exp\left\{-s\left(x + \sum_{j_2,\cdots,j_n} \Gamma_{j,j_2} + \Gamma_{j_2,j_3} + \cdots + \Gamma_{j_n,m}\right)\right\}\right]$$

$$= \mathrm{e}^{-sx}\vec{e}_j[\Gamma(s)]^n\vec{e}'_m.$$

Unconditioning, we write the previous relation in matrix form:

$$\Gamma(s) = \int_0^\infty \mathrm{e}^{-sx}\begin{bmatrix} a_1^1(x) & \cdots & a_1^M(x) \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_M^M(x) \end{bmatrix} dF_X(x)$$

$$+ \int_0^\infty \mathrm{e}^{-sx} \sum_{n=1}^\infty \begin{pmatrix} a_1(x) \\ \vdots \\ a_M(x) \end{pmatrix} * a_1^{(n-1)}(x) * (a_1^1(x) \cdots a_1^M(x))[\Gamma(s)]^n \, dF_X(x).$$

To obtain a more concise form of the above equation we use the fact that,

for every pair of matrices D_0 of full rank and D_1 of rank 1, $(D_0+D_1)^{-1} = D_0^{-1} - D_0^{-1}D_1D_0^{-1}/[1+\text{trace}\,(D_0^{-1}D_1)]$. Therefore,

$$(Is + A_0 + zA_1)^{-1} = (Is + A_0)^{-1} + \frac{z}{1 - z\alpha_1(s)} \begin{bmatrix} \alpha_1(s)\vec{\alpha}_1'(s) \\ \vdots \\ \alpha_M(s)\vec{\alpha}_1'(s) \end{bmatrix},$$

which expressed in real time gives

$$(45)\quad e^{-(A_0+zA_1)t} = \begin{bmatrix} a_1^1(t) & \cdots & a_1^M(t) \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_M^M(t) \end{bmatrix} + \sum_{n=1}^{\infty} z^n \begin{bmatrix} a_1(t) \\ \vdots \\ a_M(t) \end{bmatrix} * a_1^{(n-1)}(t) * (a_1^1(t) \cdots a_1^M(t)).$$

Using (45) we have that the busy period matrix $\Gamma(s)$ satisfies the equation $\Gamma(s) = \int_0^\infty e^{-sx} e^{-(sI+A_0+A_1\Gamma)} dF_X(x) = \Phi_X(s + A_0 + A_1\Gamma(s))$. The above implicit equation completely characterizes $\Gamma(s)$. Notice that for $M=1$, this reduces to $\gamma(s) = \phi_X(s + \lambda - \lambda\gamma(s))$, which is the equation that the transform of the busy period satisfies in an $M/G/1$ queue.

We now proceed in our analysis of the busy cycle $C(s)$. A careful comparison of the definitions of $C(s)$ and $\Gamma(s)$ reveals that $C_{i,j}(s) = \alpha_i(s)\vec{e}_1\Gamma(s)\vec{e}_j'$, or in words, the busy cycle that ends with the $ATC = j$ given that it starts with the $ATC = i$, is equal to the remaining interarrival given that the $ATC = i$ plus a busy period that ends with the $ATC = j$.

In matrix notation

$$C(s) = \begin{bmatrix} \alpha_1(t) \\ \vdots \\ \alpha_M(t) \end{bmatrix} \vec{e}_1\Gamma(s).$$

Using the fact that $\alpha_i(s) = -\vec{e}_i(Is + A_0)^{-1}A_1\vec{e}_1'$ and that $A_1\vec{e}_1'\vec{e}_1 = A_1$ we have that

$$(46)\qquad C(s) = (Is + A_0)^{-1}A_1\Gamma(s).$$

Now, we are in a position to obtain $\vec{H}$ and $\vec{H}^+$. Let C be a matrix with elements $c_{i,j} \triangleq \lim_{s\to 0} c_{i,j}(s)$, then we have that $c_{i,j}$ denotes the probability of a departing customer to leave the system empty and the $ATC = j$ given that the previous departing customer that left the system empty left the $ATC = i$. Furthermore, H_i^+ is the probability of a departing customer leaving the system empty and the $ATC = i$. Therefore, the following relation must be satisfied:

$$(47)\qquad \vec{H}^+C = \vec{H}^+,$$

i.e. $\vec{H}^+$ is a left eigenvector of C with eigenvalue 1. Notice that the above system of equations completely characterizes $\vec{H}^+$ up to a constant. On the other hand we know from (8) that $\vec{H} = \lambda\vec{H}^+A_0^{-1}$, and also that $\vec{H}\vec{1}' = (1-\rho)$. The last system of equations together with (47) provides us with a complete description of both $\vec{H}^+$ and $\vec{H}$.

Appendix B

We consider single server priority queueing systems with mixed generalized Erlang arrivals, in which there are two distinct customer classes, numbered 1 and 2. Customers of class 1 have priority over those of class 2 and within the same class customers are served in a FIFO order.

We first assume that customers of class 1 have preemptive priority over those of class 2. We will evaluate the vectors $\vec{h}$ and $\vec{k}$ such that $h_i \triangleq \boldsymbol{P}\{R_2 = i, L_1 = 0, L_2 = 0\}$ and $k_r \triangleq \boldsymbol{P}\{R_2^a = r, L_1^a = 0, L_2^a = 0\}$.

We first define a set of matrices $\Gamma_2^{r,k}(s)$ (for $r, k = 1, 2, \cdots, M_1$) with $[\Gamma_2^{r,k}(s)]_{i,j}$ to be the Laplace transform of a class 2 busy period that ended with $ATC_1 = k$ and $ATC_2 = j$ given that it started with $ATC_1 = r$ and $ATC_2 = i$.

Using the same ideas as in Proposition 1 we obtain the set of matrices by solving the following $M_1 \times M_1$ system of nonlinear equations:

$$\Gamma_2^{r,k}(s) = \int_0^\infty e^{-sx} \begin{bmatrix} b_1^1(x) & \cdots & b_1^{M_2}(x) \\ \vdots & \ddots & \vdots \\ 0 & \cdots & b_{M_2}^{M_2}(x) \end{bmatrix} dF_{G_2^{r,k}}(x)$$

$$+ \int_0^\infty e^{-sx} \sum_{n=1}^{\infty} \begin{pmatrix} b_1(x) \\ \vdots \\ b_M(x) \end{pmatrix} * b_1^{(n-1)}(x) * (b_1^1(x) \cdots b_1^M(x))\, dF_{G_2^{r,r_1}}(x) \sum_{r_1} U_n^{r_1}(s),$$

where $U_n^{r_1}(s) = \sum_{r_2, r_3, \cdots, r_n} \Gamma_2^{r_1,r_2}(s) \times \Gamma_2^{r_2,r_3}(s) \times \cdots \times \Gamma_2^{r_n,k}(s)$.

We next define the matrix $T^{r,k}(s)$ with $[T^{r,k}(s)]_{i,j}$ to be the Laplace transform of the duration of an interval that ends with a departure leaving the system empty, the $ATC_1 = k$ and $ATC_2 = j$ given that it started with a departure leaving the system empty $ATC_1 = r$ and $ATC_2 = i$.

To obtain $T^{r,k}(s)$ we follow the same line of arguments as in Proposition 1, and we obtain that

$$\boldsymbol{E}[e^{-sT_{i,j}^{r,k}} \text{ with 1st arrival from class 2}] = \int_0^\infty e^{-sx} \sum_{k_1=1}^{M_1} a_r^{k_1}(x) \vec{e}_1 \Gamma_2^{k_1,k}(s) \vec{e}_j' b_i(x)\, dx,$$

$\boldsymbol{E}[e^{-sT_{i,j}^{r,k}}$ with 1st arrival from class 1]

$$= \int_0^\infty a_r(x) e^{-sx} \vec{b}_i(x) \int_0^\infty e^{-sy} \begin{bmatrix} b_1^1(y) & \cdots & b_1^{M_2}(y) \\ \vdots & \ddots & \vdots \\ 0 & \cdots & b_{M_2}^{M_2}(y) \end{bmatrix} \vec{e}_j'\, dF_{\Gamma_{1,1,k}}(y)$$

$$+ \int_0^\infty a_r(x) e^{-sx} \vec{b}_i(x) \int_0^\infty \sum_{n=1}^{\infty} \begin{pmatrix} b_1(y) \\ \vdots \\ b_M(y) \end{pmatrix} * b_1^{(n-1)}(y) * (b_1^1(y) \cdots b_1^M(y))$$

$$\times \sum_{r_1} U_n^{r_1}(s) \vec{e}_j' e^{-sy}\, dF_{\Gamma_{1,1,k}}(y).$$

Next, we define the $M_1 \times M_2$ matrices $\boldsymbol{E}^+$ and $\boldsymbol{E}$ with elements

$$[\boldsymbol{E}^+]_{r,j} \triangleq \boldsymbol{P}\{\text{at a departure epoch } L^+ = 0, R_1^+ = r, R_2^+ = j\},$$

$$[\boldsymbol{E}]_{r,j} \triangleq \boldsymbol{P}\{L = 0, R_1 = r, R_2 = j\}.$$

As in the proof of Proposition 1, from the definition of $T^{r,k}(s)$ and $\boldsymbol{E}^+$ we obtain that

$$\sum_{r,j} E^+_{r,j} \lim_{s \to 0} [T^{r,k}(s)]_{j,i} = E^+_{k,i} \qquad \text{for} \quad k = 1, \cdots, M_1, i = 1, \cdots, M_2.$$

Furthermore, from the conservation of flow around the state $(L = 0, R_1 = r, R_2 = j)$ we can obtain that $\boldsymbol{E}^+ = A_0' \boldsymbol{E} + \boldsymbol{E} B_0$, so that knowing $\boldsymbol{E}^+$ we can evaluate the matrix $\boldsymbol{E}$.

Finally,

$$h_j = \sum_{r=1}^{M_1} E_{r,j}, \qquad k_j = \sum_{r=1}^{M_1} \lambda_{1,r} p_{1,r} E_{r,j}. \tag{48}$$

Now, in the case where class 1 customers have non-preemptive priority over the class 2 customers we need to evaluate the vector $\vec{g}$ with $g_r \triangleq \boldsymbol{P}\{L = 0, R_1 = r\}$. Notice that the total number of customers in a priority system as well as the ATC are independent of the priority policy (given that it is work conserving). Therefore, to calculate $\vec{g}$ we can use the matrix $\boldsymbol{E}$ to obtain that

$$g_r = \sum_{j=1}^{M_2} E_{r,j}. \tag{49}$$

Acknowledgements

We should like to thank the reviewer of the paper for helpful suggestions and Professor David Stanford whose critical comments improved the paper.

References

[1] BACCELLI, F. AND BREMAUD, P. (1994) *Elements of Queueing Theory: Palm–Martingale Calculus and Stochastic Recurrences.* Springer, New York.

[2] BERTSIMAS, D. AND NAKAZATO, D. (1995) The distributional Little's law and its applications. *Operat. Res.* **43,** 298.

[3] BERTSIMAS, D. AND NAKAZATO, D. (1992) Transient and busy period analysis of the $GI/G/1$ queue: The method of stages. *Queueing Systems* **10,** 153–184.

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

[5] DOSHI, B. (1985) A note on stochastic decomposition in a $GI/G/1$ queue with vacations or setup times. *J. Appl. Prob.* **22,** 419–428.

[6] FUHRMANN, S. W. AND COOPER, R. B. (1985) Stochastic decompositions in a $M/G/1$ queue with generalized vacation. *Operat. Res.* **33,** 1117–1129.

[7] FUHRMANN, S. W. (1985) Symmetric queues served in cyclic order. *Operat. Res. Lett.* **4,** 139–144.
[8] GLYNN, P. W. AND WHITT, W. (1991) A new view of the heavy traffic limit theorem for the infinite-server queues. *Adv. Appl. Prob.* **23,** 188–209.
[9] HAJI, R. AND NEWELL, G. (1971) A relation between stationary queue and waiting time distributions. *J. Appl. Prob.* **8,** 617–620.
[10] HEYMAN, D. AND SOBEL, M. (1982) *Stochastic Models in Operations Research. Vol 1.* McGraw-Hill, New York.
[11] JAISWAL, N. K. (1968) *Priority Queues.* Academic Press, New York.
[12] KEILSON, J. AND SERVI, L. (1988) A distributional form of Little's law. *Operat. Res. Lett.* **7,** 223–227.
[13] KEILSON, J. AND SERVI, L. (1990) The distributional form of Little's law and the Fuhrmann–Cooper decomposition. *Operat. Res. Lett.* **9,** 239–247.
[14] KLEINROCK, L. (1975) *Queueing Systems. Vol. 1: Theory.* Wiley, New York.
[15] LITTLE, J. (1961) A proof of the theorem $L = \lambda W$. *Operat. Res.* **9,** 383–387.
[16] LUCANTONI, D. M., MEIER-HELLSTERN, K. AND NEUTS, M. F. (1990) A single-server queue with server vacations and a class of non-renewal arrival processes. *Adv. Appl. Prob.* **22,** 676–705.
[17] MELAMED, B. AND WHITT, W. (1990) On arrivals that see time averages. *Operat. Res.* **38,** 156–172.
[18] MIYAZAWA, M. (1994) Rate conservation laws: a survey. *Queueing Systems* **15,** 1–58.
[19] NEUTS, M. F. (1975) Probability distributions of phase type. *Liber Amicorum Prof. Emeritus H. Florin* Department of Mathematics, University of Louvain, 173–206.
[20] NEUTS, M. F. (1986) Generalizations of the Pollaczek–Khinchine integral equation in the theory of queues. *Adv. Appl. Prob.* **18,** 952–990.
[21] NEUTS, M. F. (1989) *Structured Stochastic Matrices of M/G/1 Type and their Applications.* Marcel Dekker, New York.
[22] ROSS, S. (1985) *Introduction to Probability Models.* 3rd edn. Academic Press, New York.
[23] SMITH, W. L. (1954) Asymptotic Renewal Theorems. *Proc. R. Soc. Edinburgh.* A **64,** 9–48
[24] TAKÁCS, L. (1962) *Introduction to the Theory of Queues.* Oxford University Press, New York.
[25] WHITT, W. (1991) A review of $L = \lambda W$ and extensions. *Queueing Systems.*
[26] WOLFF, R. (1989) *Stochastic Modeling and the Theory of Queues.* Prentice Hall, New York.