WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

A Unified Method to Analyze Overtake Free Queueing Systems

Dimitris Bertsimas
and
Georgia Mourtzinou

WP# - 3486-92 MSA October, 1992

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139

A Unified Method to Analyze Overtake Free Queueing Systems

Dimitris Bertsimas
and
Georgia Mourtzinou

WP# - 3486-92 MSA October, 1992

A unified method to analyze overtake free queueing systems

Dimitris Bertsimas [*] Georgia Mourtzinou [†‡]

October 1992

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 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*. Our results include a generalization of PASTA to queueing systems with arbitrary renewal arrivals under heavy traffic conditions, a generalization of the Pollaczek-Khinchin 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, new approximate results for the distributions of L, S in a $GI/G/\infty$ queue, and new exact results for the distributions of L, Q, S, W in priority queues with mixed generalized Erlang renewal arrivals.

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

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

[‡] 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.

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.*

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 a 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 equations in electrodynamics, lead to a *complete* solution of the queueing application. One first candidate for a queueing law is Little's law [13] (see the recent review of Whitt [16] which traces the different forms of the law and its extensions). Let us examine whether Little's law leads to complete solution for the steady state $E[Q]$, $E[L]$, $E[W]$, $E[S]$ in a $GI/G/s$ queue. Let λ, μ, $\rho = \frac{\lambda}{s\mu} < 1$ be the mean arrival, service rate and traffic intensity. Then, from Little's law in the system and the queue

$$E[L] = \lambda E[S],\ E[Q] = \lambda E[W].$$

But, $E[S] = E[W] + \frac{1}{\mu}$, while the relation of $Q,\ L$ is

$$E[z^L] = z^s E[z^Q] + \sum_{n=0}^{s-1} P\{L = n\}[z^n - z^s],$$

from where

$$E[L] = s + E[Q] - \sum_{n=0}^{s-1} (s-n) P\{L = n\}.$$

Combining the previous equations we obtain that

$$\sum_{n=0}^{s-1} \frac{s-n}{s} P\{L = n\} = 1 - \rho,$$

which is exactly what Little's law would give if it were applied to a service box including the customers in service. For example, in a $GI/G/1$ queue one would be able to find that $P\{L = 0\} = 1 - \rho$, but it would not be possible to find $E[L]$. As a result, despite its importance, Little's law does not lead to a complete solution for expected performance measures.

Our goal in this paper is to demonstrate that *the distributional laws* first obtained by Haji and Newell [7] are the fundamental queueing laws for queueing systems which satisfy distributional laws for both the number in the system and the number in the queue (we will call them *overtake free systems*). We 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 we obtain new simple derivations of known results providing new insights to old results, but we obtain several new results as well. We propose two methods of analysis An asymptotic (as $\rho \rightarrow 1$) 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 [10], [11] 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 [1] gave another proof of the distributional laws that lead to a very convenient form of the law. They also proposed a framework to find $E[L]$, $E[Q]$, $E[S]$, $E[W]$ 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 approach 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. 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 new 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-Khinchin 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 $MGE_M/G/1$ 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 [5] 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 general distributional law

Consider a general queueing system, whose arrival process is a stationary process. Let $N_a(t)$ be the number of customers 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 customers up to time t for the equilibrium process (where the time of the first interarrival time is distributed as the forward recurrence time of the arrival process). Let also L^-, L^+ (Q^-, Q^+) be the number in the system (or in the queue) just before an arrival or just after a departure, respectively, for a system that satisfies the assumptions of Theorem 1 below. The distributional law can be stated as follows:

Theorem 1 *(Haji and Newell [7]) Let a given class C of customers have the following properties:*

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

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

3. *New arriving class C customers do not affect the time in the system (or the queue) for previous class C customers.*

Then, given that they exist in steady state, the stationary time spent in the system (queue) S (W) of the class C customers and the stationary number of the class C customers in the system (or queue) L (Q) are related in distribution by:

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

$$Q \overset{d}{=} N_a^*(W). \tag{2}$$

In addition,

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

$$Q^- \overset{d}{=} Q^+ \overset{d}{=} N_a(W).$$

We define as *overtake free systems* those systems that satisfy both (1) and (2). Note that for the general distributional law the arriving process need not be a renewal process. If we consider renewal arrivals, however, some interesting relations between the generating function of L and the Laplace transform of S have been proved in Bertsimas and Nakazato [1] and are reviewed in Theorem 2 below. For the rest of the paper let $\alpha(s)$ be the Laplace transform of the interarrival distribution, with arrival rate $\lambda = -1/\dot{\alpha}(0)$. Let $N_a(t)$ be the number of renewals up to time t for the ordinary renewal process and $N_a^*(t)$ be the number of renewals up to time t for the equilibrium renewal process.

Theorem 2 *(Bertsimas and Nakazato [1]) Arrivals of class C form a renewal process whose interarrival time has a transform $\alpha(s)$. Under the assumptions of Theorem 1, the distribution function $F_S(t) = P\{S \leq t\}$ of S and the generating functions $G_L(z), G_{L^-}(z)$, $G_{L^+}(z)$ satisfy the following relations:*

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

$$G_{L^-}(z) = G_{L^+}(z) = \int_0^\infty K_o(z,t)\, dF_S(t), \tag{4}$$

and the distribution function $F_W(t) = P\{W \leq t\}$ of W and the generating functions $G_Q(z), G_{Q^-}(z), G_{Q^+}(z)$ satisfy the relations:

$$G_Q(z) = \int_0^\infty K(z,t)\, dF_W(t), \tag{5}$$

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

with

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

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

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

$$K^*(z,s) = \int_0^\infty e^{-st}\, K(z,t)\, dt = \frac{1}{s} - \lambda \frac{(1-z)(1-\alpha(s))}{s^2(1-z\alpha(s))}, \tag{7}$$

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

For the case of Poisson arrivals $K(z,t) = K_o(z,t) = e^{-\lambda t(1-z)}$ and thus the distributional laws become a relation between transforms (Keilson and Servi [10]):

$$G_L(z) = \phi_S(\lambda(1-z)). \tag{8}$$

2.2 A vector distributional law

A vector generalization of (8) has been proposed in Bertsimas and Nakazato [1] under the assumption that the arrival process is a mixed generalized Erlang (MGE) process, which can approximate any renewal arrival process arbitrarily closely. The stage representation of the MGE distribution is presented in Figure 1, i.e., we conceive the arrival process as an arrival timing channel (ATC) consisting of M consecutive exponential stages with rates $\lambda_1, \lambda_2, ..., \lambda_M$ and with probabilities $p_1, p_2, ..., p_M$ ($p_M = 1$) of entering the system after the completion of the 1st, 2nd, $\ldots$, Mth stage.

Let $a_k(t)$ be the pdf of the remaining interarrival time if the customer in the ATC is in stage $k = 1, \ldots, M$. 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 $\frac{1}{\lambda}$ denotes the mean interarrival time.

Let $\alpha_k(s)$ be the Laplace transform of $a_k(t)$.

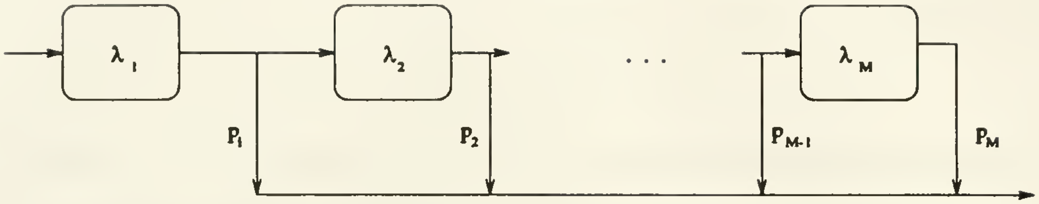

Figure 1: The Coxian class of distributions

Let $a_i^j(t)$ be the probability to move from stage $i \leq j$ of the ATC to stage j during the interval $[0,t)$ without having any new arrival.

We will also use the notation:

$\vec{a_1}(t) = (a_1^1(t),\dots,a_1^M(t))'$, $\vec{a_k}(t) = (0,\dots,a_k^k(t),\dots,a_k^M(t))'$.

$\vec{\alpha_k}(s)$ denotes the Laplace transforms of $\vec{a_k}(t)$.

$\vec{e_j} = (0,\dots,1,\dots,0)'$, $\vec{1} = (1,\dots,1,\dots,1)'$.

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:

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

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

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

thus the interarrival pdf becomes

$$a(t) = -\operatorname{trace}(e^{-A_0 t}A_1).$$

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^+, Q^+ = The number of customers in the system (or queue) immediately after a departure epoch.

L_t^-, Q_t^- = The number of customers in the system (or queue) 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 <u>not</u> 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.

$\vec{P}_n^+ = [P\{L^+ = n \cap R^+ = i\}]_{i=1}^{i=M}$, $\vec{P}_L^+(z) = \sum_{n=0}^{\infty} z^n \vec{P}_n^+$, $\vec{P}_n^- = \left[P\{L_t^- = n \cap R_t^- = i\}\right]_{i=1}^{i=M}$, $\vec{P}_L^-(z) = \sum_{n=0}^{\infty} z^n \vec{P}_n^-$, and $\vec{P}_n = [P\{L = n \cap R = i\}]_{i=1}^{i=M}$, $\vec{P}_L(z) = \sum_{n=0}^{\infty} z^n \vec{P}_n$.

We denote with $\vec{P}_Q^+(z)$, $\vec{P}_Q^-(z)$, and $\vec{P}_Q(z)$ the corresponding transforms for the number of customers in the queue. The vector distributional law is described in the following theorem.

Theorem 3 *(Bertsimas and Nakazato [1]) Under the assumptions of Theorem 1 and for mixed generalized Erlang interarrival times characterized by the matrices* A_0, A_1,

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

$$\vec{P}_Q(z) = \vec{P}_Q^-(z),$$

$$\vec{P}_L(z) = \lambda(1-z)\vec{P}_L^+(z)(A_0 + zA_1)^{-1}, \tag{9}$$

$$\vec{P}_Q(z) = \lambda(1-z)\vec{P}_Q^+(z)(A_0 + zA_1)^{-1}, \tag{10}$$

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

$$\vec{P}_Q^+(z) = \vec{e}_1{}'\Phi_W(A_0 + zA_1),$$

$$\vec{P}_L(z) = \lambda(1-z)\,\vec{e}_1{}'\Phi_S(A_0 + zA_1)(A_0 + zA_1)^{-1}, \tag{11}$$

$$\vec{P}_Q(z) = \lambda(1-z)\,\vec{e}_1{}'\Phi_W(A_0 + zA_1)(A_0 + zA_1)^{-1}, \tag{12}$$

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 (3) 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 arrival the vector forms reduce to scalars and we obtain (8).

3 An asymptotic method of analysis for overtake free queueing systems

In this section we consider overtake free systems with general arrival processes that satisfy the assumptions of Theorem 1 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 system. 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 3. 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 $D/D/1$). As a result, we are interested in the behavior of $K(z,t)$, $K_o(z,t)$ as $t \to \infty$ and $z \to 1$.

Theorem 4 *Asymptotically, as $t \to \infty$ and $z \to 1$ the kernels in Theorem 2 behave as follows:*

$$K(z,t) \sim e^{-tf(z)},$$

and

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

where

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

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

Proof

From (7) by writing $K^*(z,s) = \frac{N(z,s)}{s^2 D(s,z)}$ 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

$$K^*(z,s) = \frac{2\dot{\alpha}(0)z - \lambda(1-z)\ddot{\alpha}(0) + [z\ddot{\alpha}(0) - \frac{\lambda(1-z)E[A^3]}{3}]s + O(s^2)}{(s-s_1)(s-s_2)z\ddot{\alpha}(0)},$$

where the Taylor series expansion of the smaller root s_1 in terms of $(1-z)$ is

$$s_1 = -\lambda(1-z) + \frac{1}{2}\lambda(1-z)^2(c_a^2 - 1) + O((1-z)^3),$$

$$s_2 = -2\frac{\dot{\alpha}(0)}{\ddot{\alpha}(0)} - s_1.$$

Using a partial fraction expansion we invert in the time domain. Since we are interested in the behavior as $t \to \infty$ only the smaller root s_1 will be asymptotically important. As a result, after some tedious, but straightforward manipulations we obtain that

$$K(z,t) \sim (1 + O(1-z)^2)e^{s_1 t},$$

i.e.,

$$K(z,t) \sim (1 + O(1-z)^2)e^{-t(\lambda(1-z)-\frac{1}{2}\lambda(1-z)^2(c_a^2-1))}.$$

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

$$K_o(z,t) \sim [1 - \frac{1}{2}(1-z)(c_a^2-1) + O((1-z)^2)]e^{-t(\lambda(1-z)-\frac{1}{2}\lambda(1-z)^2(c_a^2-1))}.\square$$

Combining Theorems 3 and 4 the asymptotic form of the distributional Little's law becomes

Theorem 5 *In a queueing system that satisfies the assumptions of Theorem 1 and assuming that as $\rho \to 1$, $L, Q, S, W \to \infty$ the following asymptotic relations hold as $\rho \to 1$:*

$$G_L(z) \sim \phi_S(f(z)), \tag{13}$$

$$G_Q(z) \sim \phi_W(f(z)), \tag{14}$$

$$G_{L^+}(z) \sim [1 - \frac{1}{2}(1-z)(c_a^2-1)]\phi_S(f(z)), \tag{15}$$

$$G_{Q^+}(z) \sim [1 - \frac{1}{2}(1-z)(c_a^2-1)]\phi_W(f(z)), \tag{16}$$

with

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

Proof

Substituting in (3),(5) and (4), (6) the asymptotic form of $K(z,t)$ and $K_o(z,t)$ from the previous theorem we obtain (13), (14) and (15), (16), respectively. □

Although only valid asymptotically, (13), (14) and (15), (16) are very useful since they are relations among transforms, which we will further exploit in the section. Also, the previous expressions are exact for the Poisson case ($c_a^2 = 1$). In order to develop some further insight on the asymptotic expressions of Theorem 5 we consider the case of E_2 arrivals, i.e., $\alpha(s) = (\frac{2\lambda}{2\lambda+s})^2$. Then,

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

and

$$K_o(z,t) = \frac{(1+\sqrt{z})}{2\sqrt{z}} e^{-2\lambda(1-\sqrt{z})t} - \frac{(1-\sqrt{z})}{2\sqrt{z}} e^{-2\lambda(1+\sqrt{z})t}.$$

As $z \to 1$ only the first of the two exponentials contributes to $K(z,t)$, $K_o(z,t)$. Expressions (13) and (15) are the Taylor series expansions of the first exponential in terms of $1-z$.

3.2 An asymptotic generalization of PASTA

Theorem 5 leads to an interesting generalization of PASTA in systems in heavy traffic. Consider a queueing system that satisfies the assumptions of Theorem 1. 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 the distribution of L^- depends on the queueing discipline, while the distribution of L does not. In heavy traffic ($\rho \to 1$), however, where Theorem 5 is applicable we have that

$$G_{L^-}(z) = G_{L^+}(z) \sim G_L(z)[1 - \frac{1}{2}(1-z)(c_a^2 - 1)]. \tag{17}$$

In particular the first moments are related by

$$E[L^-] \sim E[L] + \frac{c_a^2 - 1}{2},$$

which means that in heavy traffic, where both $E[L^-]$, $E[L]$ are very large, their difference asymptotically depends only on the coefficient of variation of the arrival process. Apparently, a relation similar to (17) holds for the number of customers in the queue by a similar reasoning. 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 $\rho \to 1$, (17) does not hold, since L^-, L (and Q^-, Q) remain bounded and therefore the assumptions of Theorem 4 are not valid.

3.3 An asymptotic method

Theorem 5 as well as (17) 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 (13).*

2. *Relate the transforms of Q and W, using the asymptotic form of the distributional law (14).*

3. *Relate the transforms of S and W using the fact that $S = W \oplus X$.*

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, (17), find the transforms of L^+ and Q^+ from the transforms of L and Q .*

3.4 Applications 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 $1/\lambda$, $E[X]$, 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 transform of the service time distribution.

Theorem 6 *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:*

$$\phi_W(s) = \frac{(1 - f^{-1}(s))(1-\rho)}{\phi_X(s) - f^{-1}(s)}, \tag{18}$$

and

$$G_Q(z) = \frac{(1-z)(1-\rho)}{\phi_X(f(z)) - z}, \tag{19}$$

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 two first steps of the asymptotic method we obtain from (13) and (14), as $\rho \to 1$:

$$G_L(z) = \phi_S(f(z)),$$

$$G_Q(z) = \phi_W(f(z)).$$

Performing the third step, since $S = W \oplus X$ 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 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 (18) and

(19), as well as the transforms of the system time and the number of customers in the system. □

Remarks:

1. Using (17) 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 (18), (19) are exact and generalize the well known Pollaczek-Khinchin formulae for the $M/G/1$ queue.

3. By expanding $\phi_W(s)$ in powers of s we obtain

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

   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 $\rho \to 1$

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

   and

$$E[W^2] = 2A.$$

   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 it 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/ ∞ queue

We now apply the asymptotic method to find approximate closed form expressions for the variance of the number in a GI/G/ ∞ system.

Theorem 7 *In a $GI/G/\infty$ queue in heavy traffic conditions $(E[X] \to \infty)$*

$$G_L(z) \approx e^{-\lambda(1-z)E[X]+\frac{1}{2}\lambda(1-z)^2(c_a^2-1)\int_0^\infty x f_X^2(x)dx},$$

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

and

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

Proof

In a $GI/G/\infty$ system the distributional law doesn't hold because Assumption 2 in Theorem 1 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 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 $f_X(t) = \delta(t - E[X])$ and thus from (2)

$$G_L(z) = K(z, E[X]). \tag{20}$$

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 service time is $P\{X = x_j\} = p_j,\ j = 1, \ldots, k$. The customers with service times x_j can be treated as a separate class C_j of customers with arrival process being a 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

$$\lambda_j = \lambda p_j$$

$$c_{a_j}^2 = 1 + p_j(c_a^2 - 1).$$

If $L_j,\ j = 1, \ldots, k$ 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 an $GI/D/\infty$ for which the distributional law holds. Then applying (20)

$$G_{L_j}(z) = K(z, x_j).$$

For large x_j the asymptotic form of the distributional law of Theorem 4 is valid and thus

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

Therefore,

$$G_L(z) \approx e^{-\lambda(1-z)\sum_{j=1}^{k} p_j x_j + \frac{1}{2}\lambda(1-z)^2(c_a^2 - 1)\sum_{j=1}^{k} p_j^2 x_j}.$$

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

$$G_L(z) \approx e^{-\lambda(1-z)E[X] + \frac{1}{2}\lambda(1-z)^2(c_a^2 - 1)\int_0^\infty x f_X^2(x)dx},$$

which leads to

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

and

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

Remark: For the case of Poisson arrivals ($c_a^2 = 1$) the expressions of the previous theorem are exact leading to the well known result

$$G_L(z) = e^{-\lambda(1-z)E[X]},$$

i.e., L has a Poisson distribution with rate $\lambda E[X]$.

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 the assumptions of Theorem 1 and we describe a unified exact method to obtain the distributions L, Q, S, W, L^+, and Q^+. We will use the notation of Section 2.2. In order to accomplish our goal we first derive a relation between L^+ and Q^+, from first principles. Then, in subsection 4.1, we present the exact method in an algorithmic form and finally in subsection 4.2 we illustrate the method in the case of $MGE_M/G/1$ and $MGE_M/D/s$ queues under FIFO.

Proposition 1 *Under the assumptions of Theorem 1 and for mixed generalized Erlang interarrival times characterized by the matrices* A_0, A_1,

$$\vec{P}_L^+(z) = \vec{P}_Q^+(z)\Phi_X(A_0 + zA_1). \tag{21}$$

Proof

Conditioning on the length of the queue and the ATC stage just after a customer leaves the queue and enters service we obtain for $n \geq 1$

$$P\{L^+ = n, R^+ = i\} = \sum_{k=0}^{n}\sum_{m=1}^{M} P\{Q^+ = k, R^+ = m\}\int_0^\infty a_m(t) * a^{(n-k-1)}(t) * a_1^i(t)\, dF_X(t) \tag{22}$$

And for $n = 0$:

$$P\{L^+ = 0, R^+ = i\} = \sum_{m=1}^{M} P\{Q^+ = 0, R^+ = m\}\int_0^\infty a_m^i(t)\, dF_X(t)$$

For every pair of matrices C of full rank and D of rank 1,

$$(C + D)^{-1} = C^{-1} - \frac{C^{-1}DC^{-1}}{1 + \text{trace}(C^{-1}D)}.$$

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

$$e^{-(A_0+zA_1)t} = \begin{bmatrix} a_1^1(t) & \cdots & a_1^M(t) \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_M^M(t) \end{bmatrix} + \sum_{n=1}^{\infty} z^n \begin{pmatrix} a_1(t) \\ \vdots \\ a_M(t) \end{pmatrix} * a_1^{(n-1)}(t) * \begin{pmatrix} a_1^1(t) & \cdots & a_1^M(t) \end{pmatrix}. \tag{23}$$

Taking generating functions in (22) and using (23) we prove (21).□

Remark :

Equation (21) also follows from Theorem 3. The reason we have included a separate proof is that often in more general systems (like priority systems in Section 6) we need to generalize Proposition 1.

4.1 An exact method

Theorem 3 and Proposition 1 enable us to present an unified exact method for solving overtake free systems with MGE arrivals under the assumptions of Theorem 1. We will use the notation of Section 2.2.

Exact method of analysis

1. *Relate the transforms $\vec{P}_{L+}$ and $\vec{P}_L$ using (9).*

2. *Relate the transforms $\vec{P}_{Q+}$ and $\vec{P}_Q$ using (10).*

3. *Relate the transforms $\vec{P}_{L+}$ and $\vec{P}_{Q+}$ using (21).*

4. *Relate the transforms of $\vec{P}_L$ and $\vec{P}_Q$ using the characteristics of the system up to constant terms and use Little's law to evaluate the constants (see Section 4.2 for further details).*

5. *Solve the 4×4 system of equations from the previous 4 steps to find $\vec{P}_L$, $\vec{P}_Q$, $\vec{P}_{L+}$ and $\vec{P}_{Q+}$.*

6. *Find the transforms of S, and W, from (11) and (12).*

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

4.2 The $MGE_M/G/1$ and $MGE_M/D/s$ queues under FIFO

We consider in this subsection a $MGE_M/G/1$ queue,with a FIFO service discipline where the arrival process is a generalized Erlang process characterized by the matrices A_0 and A_1. Let $\alpha(s) = \frac{\alpha_N(s)}{\alpha_D(s)}$ 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.

Theorem 8 *In a $MGE_M/G/1$ queue under FIFO*

$$\vec{P}_Q(z) = (1-z)\vec{H}(\Phi_X(A_0+zA_1)-zI)^{-1}, \tag{24}$$

$$\vec{P}_L(z) = (1-z)\vec{H}(\Phi_X(A_0+zA_1)-zI)^{-1}\ \Phi_X(A_0+zA_1), \tag{25}$$

and

$$\phi_W(s) = \frac{\alpha_D(0)}{\alpha_D(-s)}\frac{(1-\rho)s}{\lambda(1-\alpha(-s)\phi_X(s))}\prod_{r=1}^{M-1}\frac{x_r - s}{x_r}, \tag{26}$$

where x_r, $r = 1, \ldots, M-1$ are the $M-1$ roots of the equation

$$\alpha(-s)\phi_X(s) = 1, \quad Re(s) > 0,$$

and $\vec{H}$ is an M vector whose ith component is

$$H_i = \frac{\lambda}{\lambda_i}(1 - \lambda_i p_i E[X])\prod_{k=1}^{i-1}(1-p_k). \tag{27}$$

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) and (10) and we obtain:

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

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

Combining the previous two equations with (21), third step, we obtain, since the matrices $\Phi_X(A_0 + zA_1)$, $(A_0 + zA_1)^{-1}$ commute,

$$\vec{P}_L(z) = \vec{P}_Q(z)\Phi_X(A_0 + zA_1). \tag{28}$$

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

$$\vec{P}_L(z) = (1-z)\vec{H} + z\vec{P}_Q(z), \tag{29}$$

where $\vec{H}$ is an M-vector with $H_i = P\{L = 0, R = i\}$.

Combining (28) and (29) we obtain (24) and (25).

To complete the fourth step we next compute $\vec{H}$.

$$H_i = P\{L = 0, R = i\} = P\{L = 0 | R = i\}P\{R = i\}.$$

Applying the usual Little's law to the server we find that:

$$1 - P\{L = 0 | R = i\} = (\lambda_i p_i)E[X].$$

In order to compute $P\{R = i\}$ we represent the ATC as a continuous time Markov chain with M states as shown in Figure 2. Solving for the steady-state distribution we obtain

$$P\{R = i\} = \frac{\lambda}{\lambda_i}\prod_{k=1}^{i-1}(1 - p_k), \tag{30}$$

and thus

$$H_i = \frac{\lambda}{\lambda_i}(1 - \lambda_i p_i E[X])\prod_{k=1}^{i-1}(1 - p_k).$$

At this point we have solved exactly for $\vec{P}_L(z)$ and $\vec{P}_Q(z)$ (fifth step). In order to find the transform of the waiting time distribution (sixth step) we combine (12) and (24) and obtain

$$\vec{e}_1{}'\Phi_W(A_0 + zA_1)(\Phi_X(A_0 + zA_1) - zI) = \frac{1}{\lambda}\vec{H}(A_0 + zA_1). \tag{31}$$

We now choose a z such that $A_0 + zA_1$ has M linear independent eigenvectors and thus it can be written as:

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

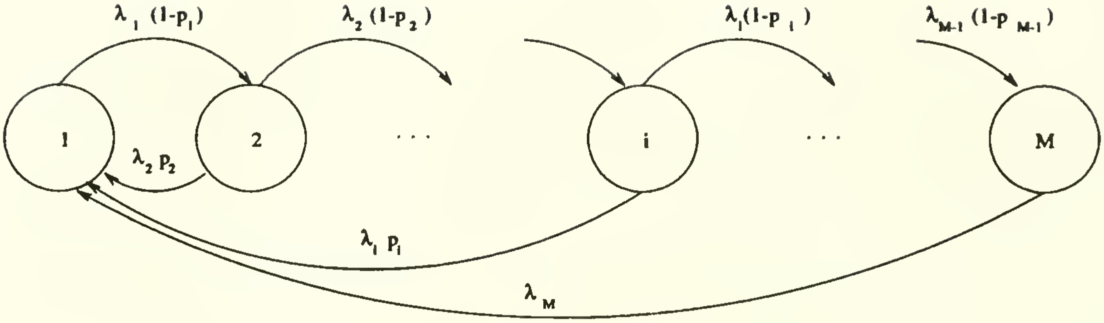

Figure 2: The Markov chain of the ATC

where $\Theta(z)$ is the diagonal matrix of the eigenvalues of $A_0 + zA_1$ which we denote by $\theta_i(z)$ for $i = 1, \ldots, M$. Bertsimas and Nakazato [2] have shown that the roots of the equation satisfy:

$$z\alpha(-\theta_i(z)) = 1, \ i = 1, \ldots, M.$$

The columns of $S(z)$ are the right eigenvectors of $A_0 + zA_1$ which we denote by $\vec{\xi}(\theta_i(z))$. Moreover,

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

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

and substituting to (31) we obtain

$$\vec{e}_1{}'S(z)\Phi_W(\Theta(z))(\Phi_X(\Theta(z)) - zI) = \frac{1}{\lambda}\vec{H}S(z)\Theta(z)$$

or

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

with $\xi_1(\theta_1(z))$ being the first component of $\vec{\xi}(\theta_1(z))$ (the previous relation also holds for every eigenvalue $\theta_i(z)$, $i = 1 \ldots M$). 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,

$$\phi_W(s) = K\frac{s\alpha(-s)}{\lambda(\alpha(-s)\phi_X(s)-1)}g(s). \tag{32}$$

Since $\phi_W(s)$ is analytic

$$g(s) = \frac{\prod_{r=1}^{M-1} x_r - s}{\alpha_D(s)},$$

where x_r, $r = 1, \ldots, M-1$ are the $M-1$ roots of the equation

$$\alpha(-s)\phi_X(s) = 1, \quad Re(s) > 0.$$

and K is a constant such that :

$$\lim_{s\to 0}\phi_W(s) = 1,$$

which leads to (26). □

Remarks:

1. Equation (24) is to the best of our knowledge new, while (26) is a generalization of the Pollaczek-Khinchin formula for the $M/G/1$ queue. It is interesting to notice that (26) could have been obtained using Hilbert factorization techniques. 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 the assumptions of Theorem 1. 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 [4], in which $Z_0 = 0$. It also generalizes the $M/G/1$ system with generalized vacations considered in Fuhrmann and Cooper [5] (see also the discussion in Wolff [17], p.457) in the sense that it allows more general arrival processes. In some of their results Fuhrmann and Cooper [5], however, relax Assumption G3 above, allowing the vacation time to depend on the arrival process. In order, however, to prove sharper decomposition results they make exactly the same assumption (their Assumption 6). Our results also generalize the results of Keilson and Servi [11] 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 [4], [5] and [11]. In this way we obtain insights on the extend to which the decomposition results depend on the Poisson assumption.

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 Fuhrmann [6]. 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 forward recurrence time of V). Let B 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 to be the ATC stage of the arrival process and the number of customers present in the system, when a vacation interval starts. We define

$$\vec{\zeta}_n = [P\{Z_0 = n \cap R_0 = m|B'\}]_{m=1}^{M} \text{ and } \vec{\zeta}(z) = \sum_{n=0}^{\infty} z^n \vec{\zeta}_n.$$

We view the vector generating function $\vec{\zeta}(z)$ as defining the service mechanism. Our main theorem is as follows:

Theorem 9 *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*

$$\vec{P}_{Q_v}(z) = (1-\rho)\ \vec{\zeta}(z)\ \Phi_{V^*}(A_0 + zA_1)(1-z)\ (\Phi_X(A_0 + zA_1) - zI)^{-1}, \tag{33}$$

$$\vec{P}_{L_v}(z) = (1-\rho)\ \vec{\zeta}(z)\ \Phi_{V^*}(A_0 + zA_1)(1-z)\ (\Phi_X(A_0 + zA_1) - zI)^{-1}\Phi_X(A_0 + zA_1). \tag{34}$$

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 (28) for Q_v and L_v we obtain

$$\vec{P}_{L_v}(z) = \vec{P}_{Q_v}(z)\Phi_X(A_0 + zA_1). \tag{35}$$

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 $P\{B\} = \rho$ and $P\{B'\} = 1 - \rho$. By conditioning on the event B we obtain

$$P\{Q_v = n, R_v = i\} = \rho P\{Q_v = n, R_v = i|B\} + (1-\rho)P\{Q_v = n, R_v = i|B'\}, \tag{36}$$

Conditioning on Z_0, R_0, V^* we obtain

$$\begin{aligned}
&P\{Q_v = n, R_v = i|B'\} = \\
&= \sum_{k=1}^{M}\sum_{m=0}^{n}\int_0^\infty P\{Q_v = n, R_v = i|B',\ V^* = t,\ Z_0 = m,\ R_0 = k\} \\
&\qquad\qquad\qquad\qquad P\{Z_0 = m,\ R_0 = k,\ V^* = t|B'\}dt \\
&= \sum_{k=1}^{M}\sum_{m=0}^{n-1} P\{Z_0 = m,\ R_0 = k|B'\}\int_0^\infty a_k(t) * a^{(n-m-1)}(t) * a_1^i(t)\ dF_{V^*}(t) \\
&\qquad + \sum_{k=1}^{M} P\{Z_0 = n,\ R_0 = k|B'\}\int_0^\infty a_k^i(t)dF_{V^*}(t),
\end{aligned} \tag{37}$$

where we used the independence of V^* and (Z_0, R_0) (Assumption G3 in the definition of queues with generalized vacations). Let $\vec{B}(z) = [\sum_{n=0}^{\infty} P\{Q_v = n, R_v = i|B\}z^n]_{i=1}^{M}$. Taking generating functions in (36) and using (23) to (37), we obtain

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

Similarly

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

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,

$$\vec{P}_{L_v}(z) = z\vec{P}_{Q_v}(z) + (1-z)(1-\rho)\ \vec{\zeta}(z)\ \Phi_{V^*}(A_0 + zA_1), \tag{38}$$

which combined with (35) gives (34) and (33). □

Remarks:

1. Equation (34), as well as (33), 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 Fuhrmann and Cooper [5]. 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 (37). Without Assumption G3, instead of (38) we would obtain

$$\vec{P}_{L_v}(z) = z\vec{P}_{Q_v}(z) + (1-z)(1-\rho)\vec{P}_{L_v|B'}(z), \tag{39}$$

where $\vec{P}_{L_v|B'}(z)$ is the vector generating function of the number in the system given that the server in on vacation. Combining (39) with (35) 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 Fuhrmann and Cooper [5].

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 a $MGE_M/G/1$ 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 9 implies the decomposition results of Doshi [4].

Theorem 10 *(Doshi [4]) 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 forward recurrence time of the vacation V.*

Proof

In this case $Z_0 = 0$ and therefore $\vec{\zeta}(z) = P\{Z = 0, R_0 = i\}_{i=1}^{M} = \vec{R}$, i.e., a vector independent of z. Then (34) 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 (25) 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 indepedent of the vacation. Taking limits as $z \rightarrow 1$ in the two previous equations we obtain

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

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

$$\vec{P}_{L_v}(z) = \vec{H}\ \Phi_{V^*}(A_0+A_1)^{-1}\ (1-z)\ (\Phi_X(A_0+zA_1)-zI)^{-1}\Phi_X(A_0+zA_1)\ \Phi_{V^*}(A_0+zA_1), \tag{40}$$

where the vector $\vec{H}$ is computed in (27). (40) offers a complete solution of the $MGE_M/G/1$ queue with exhaustive vacations.

Following exactly the same approach leading to (32) in the proof of Theorem 8 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.,

$$W_v \overset{d}{=} W \oplus V^*.\ \ \square$$

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 $F_J(t) = P\{J \leq t\}$ 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 $\vec{R}_J = \mathrm{P}\{R_J = m\}_{m=1}^{M}$ and $N(z) = E[z^N]$.

Finally, we define also the vectors

$\vec{N}_n = \mathrm{P}\{N = n \cap R_J = m\}_{m=1}^{M}$ 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 :

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

which leads to :

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

which in matrix notation becomes :

$$\vec{\zeta}(z) = \vec{N}(1) \, \Phi_J(A_0 + zA_1). \tag{41}$$

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

$$\phi_J(s) = N(\phi_X(s)). \tag{42}$$

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 ≥ 1 :

$$\mathrm{P}\{N = n, R_J = l\} = \sum_{k=0}^{n} \sum_{m=1}^{M} \mathrm{P}\{Z_0 = k, R_0 = m\} \int_0^{\infty} a_m(t) * a^{(n-k-1)}(t) * a_1^l(t) dF_V(t).$$

Taking generating functions :

$$\vec{N}(z) = \vec{\zeta}(z) \, \Phi_V(A_0 + zA_1). \tag{43}$$

By combining (41), (42) and (43) we have:

$$\vec{\zeta}(z) = \vec{\zeta}(1) \, \Phi_V(A_0 + A_1) \, \Phi_J(A_0 + zA_1), \tag{44}$$

where

$$\phi_J(s) = \vec{\zeta}(\phi_X(s))\,\Phi_V(A_0 + \phi_X(s)A_1)\,\vec{1}. \tag{45}$$

Equations (44) and (45) fully characterize $\vec{\zeta}(z)$ as we can solve for all moments. Moreover if we use Theorem 9 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. Let (A_0, A_1), (B_0, B_1) be the corresponding matrices for class 1 and 2 arrivals respectively. Then $A_0 + zA_1 = S_1(z)\Theta_1(z)S_1^{-1}(z)$, and $B_0 + zB_1 = S_2(z)\Theta_2(z)S_2^{-1}(z)$ where $\Theta_i(z)$ is the diagonal matrix of the eigenvalues and $S_i(z)$ is the matrix with columns the right eigenvectors $(i = 1, 2)$. 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 $\mathrm{E}[X_1]$ and $\mathrm{E}[X_2]$, 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 [11] 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 generalize the classical results of Takacs [15] for the M/G/1 queue for the busy period distribution to a matrix form. This generalization, which is also of independent interest, is essential since the service time of class 2 customers in a preemptive priority system depends on the busy period distribution of class 1 customers. In Section 6.2 we find the effective service time distribution in various preemptive systems as a function of the busy period matrix. In Section 6.3 we analyze systems with preemptive priorities, while in Section 6.4 we analyze systems with non-preemptive priorities.

6.1 The high priority customers busy period matrix

We denote with ATC_1 and ATC_2 the two arrival timing channels. In this section we will compute the busy period matrix $\Sigma_1(s)$ with $[\Sigma_1(s)]_{i,j} = \sigma_{ij}(s)$, $i,j = 1,\ldots,M_1$ denoting the Laplace transform of a sub-busy period interval for class 1 that ends with $ATC_1 = j$ given that it started with $ATC_1 = i$. 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 Kleinrock [12] p. 210) and therefore at the beginning of a sub-busy period ATC_1 can be in any stage.

Theorem 11 *In a $MGE_{M_1}/G/1$ queueing system where the interarrival process is characterized by the matrices A_0 and A_1 and the Laplace transform of the service time is $\phi_{X_1}(s)$ we have that:*

$$\Sigma_1(s) = \sum_{j=1}^{M_1} \phi_{X_1}(s - x_j(s))\vec{\xi}_j(s)\vec{\alpha_1}'(x_j(s)),$$

where $x_j(s)$ are the M_1 roots of the equation

$$\alpha(x)\phi_{X_1}(s-x) = 1, \quad Re(x) \leq 0 \textit{ for } Re(s) \geq 0,$$

and

$$\begin{bmatrix} \vec{\xi}_1(s) & \cdots & \vec{\xi}_{M_1}(s) \end{bmatrix} = \begin{bmatrix} \vec{\alpha_1}'(x_1(s)) \\ \vdots \\ \vec{\alpha_1}'(x_{M_1}(s)) \end{bmatrix}^{-1} .$$

Proof

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 (Takacs [15]). 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 then use the last-come-first-serve (LCFS) service discipline. Let $B_{i,m}$ be the duration of the sub-busy period for class 1 customers that ends with $ATC_1 = m$ given that it started with $ATC_1 = i$. This definition is useful for the decomposition of the busy period into sub-busy periods. Let R_1^{as} be the ATC_1 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 class 1 arrivals during x given that $ATC_1 = i$. Then, conditionally on the event $U = \{R_1^{as} = j, X_1 = x, N_i(x) = n\}$ we obtain the following decomposition, for $n \geq 1$

$$E[e^{-sB_{i,m}}|R_1^{as} = j, X_1 = x, N_i(x) = n] = E[e^{-s(x+\sum_{j_2,\ldots,j_n} B_{j,j_2}+B_{j_2,j_3}+\ldots+B_{j_n,m})}]$$

$$= e^{-sx}\vec{e}_j'[\Sigma_1(s)]^n\vec{e}_m.$$

Unconditioning, we write the previous relation in matrix form

$$\Sigma_1(s) = \int_0^\infty e^{-sx} \begin{bmatrix} a_1^1(x) & \cdots & a_1^{M_1}(x) \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{M_1}^{M_1}(x) \end{bmatrix} dF_{X_1}(x)+$$

$$+\int_0^\infty e^{-sx} \sum_{n=1}^{\infty} \begin{pmatrix} a_1(x) \\ \vdots \\ a_{M_1}(x) \end{pmatrix} * a_1^{(n-1)}(x) * \begin{pmatrix} a_1^1(x) & \cdots & a_1^{M_1}(x) \end{pmatrix} [\Sigma_1(s)]^n dF_{X_1}(x).$$

In order to compute $\Sigma_1(s)$ we will compute its eigenvalues and eigenvectors. Multiplying both parts of the previous equation with $\vec{\xi}(s)$, the right eigenvector of $\Sigma_1(s)$ corresponding to the eigenvalue $u(s)$, and using equation (23) we obtain:

$$\Sigma_1(s)\vec{\xi}(s) = u(s)\vec{\xi}(s) = \Phi_{X_1}(sI + A_0 + u(s)A_1)\vec{\xi}(s). \tag{46}$$

(Notice that for $M_1 = 1$, this reduces to $\sigma_1(s) = \phi_{X_1}(s + \lambda - \lambda\sigma_1(s))$, which is the equation that the transform of the busy period satisfies in a M/G/1 queue.)

Therefore $\vec{\xi}(s)$ must be a right eigenvector of $\Phi_{X_1}(sI + A_0 + u(s)A_1)$ and equivalently a right eigenvector of $A_0 + u(s)A_1$ with corresponding eigenvalue $-x(s)$. Bertsimas and Nakazato [2] have shown that $u(s)\alpha_1(x(s)) = 1$ and furthermore from (46) $u(s) = \phi_{X_1}(s - x(s))$. Therefore, the eigenvalues $u_j(s)$ $(j = 1, \ldots, M_1)$ of $\Sigma_1(s)$ are computed as follows: $u_j(s) = \phi_{X_1}(s - x_j(s))$, $j = 1, \ldots, M_1$ where $x_j(s)$ are the M_1 roots of the equation

$$\alpha(x)\phi_{X_1}(s - x) = 1, \quad Re(x) \leq 0 \text{ for } Re(s) \geq 0.$$

Moreover, $\vec{\xi_j}(s)$ is the right eigenvector of $A_0 + \phi_{X_1}(s - x_j(s))A_1$ corresponding to the eigenvalue $-x_j(s)$. The left eigenvectors are computed in [2] and are equal to $\vec{\alpha_1}'(x_j(s))$.

Having characterized the eigenvalues and eigenvectors of $\Sigma_1(s)$ we can spectrum decompose it as follows:

$$\Sigma_1(s) = \sum_{j=1}^{M_1} \phi_{X_1}(s - x_j(s))\vec{\xi_j}(s)\vec{\alpha_1}'(x_j(s)),$$

where

$$\begin{bmatrix} \vec{\xi}_1(s) & \cdots & \vec{\xi}_{M_1}(s) \end{bmatrix} = \begin{bmatrix} \vec{\alpha_1}'(x_1(s)) \\ \vdots \\ \vec{\alpha_1}'(x_{M_1}(s)) \end{bmatrix}^{-1} . \Box$$

Remark: The transform $\sigma_1(s)$ of the busy period distribution is given by

$$\sigma_1(s) = \vec{e}_1' \Sigma_1(s) \vec{1}.$$

6.2 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 reentry, the preemptive discipline can be further broken down into the following three categories:

- Preemptive resume discipline :

  Under this discipline the interrupted customer continues his service from the point of interruption.

- Preemptive repeat different discipline :

  Under this discipline the interrupted customer continues his service by resampling.

- 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.

We define random variables G_2^{ij}, $i, j = 1, \ldots, M_1$, which is the effective service time of a class 2 customer such that $ATC_1 = j$ when the class 2 customer finishes service given that $ATC_1 = i$ 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 the assumptions of Theorem 1 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:*

$$G_2(s) = \Phi_{X_2}(A_0 + A_1\Sigma_1(s) + sI).$$

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

$$E[e^{-sG_2^{ki}}|X_2 = x] = e^{-sx}\{a_k^i(x) + \textstyle\sum_{j_1=1}^{M_1}[\Sigma_1(s)]_{1,j_1}a_k(x) * a_{j_1}^i(x)$$

$$+ \sum_{j_1=1}^{M_1}\sum_{j_2=1}^{M_1}[\Sigma_1(s)]_{1,j_1}[\Sigma_1(s)]_{1,j_2}a_k(x) * a_{j_1}(x) * a_{j_2}^i(x) + \ldots\},$$

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:

$$E[e^{-sG_2^{ki}}|X_2 = x] = e^{-sx}\vec{e}_k' \begin{bmatrix} a_1^1(x) & \cdots & a_1^{M_1}(x) \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{M_1}^{M_1}(x) \end{bmatrix} \vec{e}_i + e^{-sx}\vec{e}_k' \textstyle\sum_{n=1}^{\infty} \begin{pmatrix} a_1(x) \\ \vdots \\ a_{M_1}(x) \end{pmatrix} *$$

$$(a_1(x)[\Sigma_1(s)]_{1,1}+\ldots+a_{M_1}(x)[\Sigma_1(s)]_{1,M_1})^{(n-1)}*([\Sigma_1(s)]_{1,1}\vec{a}_1'(x)+\ldots+[\Sigma_1(s)]_{1,M_1}\vec{a}_{M_1}'(x))\vec{e}_i.$$

Using (23) we obtain:

$$E[e^{-sG_2^{ki}}|X_2 = x] = e^{-sx}\vec{e}_k'\, e^{-(A_0+A_1\Sigma_1(s))x}\vec{e}_i.$$

Therefore,

$$E[e^{-sG_2^{ki}}] = \vec{e}_k'\, \Phi_{X_2}(A_0 + A_1\Sigma_1(s) + sI)\vec{e}_i,$$

and hence,

$$G_2(s) = \Phi_{X_2}(A_0 + A_1\Sigma_1(s) + sI). \quad \square$$

Remark: For the Poisson case we obtain

$$\phi_{G_2}(s) = \phi_{X_2}(\lambda_1 - \lambda_1\sigma(s) + s),$$

which is in agreement with Jaiswal [9].

Preemptive repeat disciplines

Let $\vec{a}(t) = (a_1(t), \ldots, a_k(t), \ldots, a_{M_1}(t))'$ and $A(t) = \begin{bmatrix} \vec{a_1}'(t) \\ \vdots \\ a_{\vec{M}_1}'(t) \end{bmatrix}$.

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*

$$G_2(s) = \int_0^\infty A(x)\ e^{-sx} f_{X_2}(x)dx \left[I - \int_0^\infty f_{X_2}(x)\int_0^x \vec{a}(y)e^{-sy}dydx\ \vec{e}_1'\ \Sigma_1(s)\right]^{-1}.$$

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

$$G_2(s) = \int_0^\infty A(x)\left[I - \int_0^x \vec{a}(y)e^{-sy}dy\ \vec{e}_1'\ \Sigma_1(s)\right]^{-1} e^{-sx} f_{X_2}(x)dx.$$

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 $ATC_1 = k$. 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

$$E[e^{-sG_2^{ki}}|X_2 = x] = a_k^i(x)e^{-sx} + \int_0^x a_k(y)e^{-sy}dy\ \vec{e}_1'\ \Sigma_1(s)G_2(s)\ \vec{e}_i.$$

Thus,

$$\phi_{G_2^{ki}}(s) = \int_0^\infty a_k^i(x)e^{-sx}f_{X_2}(x)dx + \int_0^\infty f_{X_2}(x)\int_0^x a_k(y)e^{-sy}dydx\ \vec{e}_1'\ \Sigma_1(s)G_2(s)\ \vec{e}_i,$$

And in matrix form :

$$G_2(s) = \int_0^\infty A(x)\ e^{-sx}f_{X_2}(x)dx\left[I - \int_0^\infty f_{X_2}(x)\int_0^x \vec{a}(y)e^{-sy}dydx\ \vec{e}_1'\ \Sigma_1(s)\right]^{-1}.$$

Finally for *the repeat identical* case :

$$G_2(s) = \int_0^\infty A(x)\left[I - \int_0^x \vec{a}(y)e^{-sy}dy\ \vec{e}_1'\ \Sigma_1(s)\right]^{-1} e^{-sx}f_{X_2}(x)dx.\Box$$

In the case of Poisson arrivals we can obtain the results of Jaiswal [9], namely :

$$\phi_{G_2}(s) = \frac{\phi_{X_2}(s+\lambda_1)}{1 - \frac{\lambda_1}{s+\lambda_1}(1 - \phi_{X_2}(s+\lambda_1))\sigma_1(s)} \quad \text{and}$$

$$\phi_{G_2}(s) = \int_0^\infty \frac{e^{-(s+\lambda_1)x}}{1 - \frac{\lambda_1}{s+\lambda_1}(1 - e^{-(s+\lambda_1)x})\sigma_1(s)}\, dF_{X_2}(x),$$

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

6.3 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 considered 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^+, Q_i^+ and R_i^+ be the number of customers of class i in the system, in the queue and the ATC stage of class i, respectively, immediately after a departure epoch of class 2. Let L_i^-, Q_i^- and R_i^- be the number of customers of class i in the system, in the queue and the ATC stage of class i, respectively, just before a *transition* epoch of the arrival process of class 2. A transition includes both arrivals in the system and shifts to the next exponential stage of the ATC according to the definitions of Section 2.2.

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.

We also define the matrices

$$\Pi_n^+ = \Big[P\{[L_2^+ = n \cap R_1^+ = m \cap R_2^+ = l\}\Big]_{m=1}^{m=M_1}{}_{l=1}^{l=M_2},$$

$$\Pi_n^- = \Big[P\{L_2^- = n \cap R_1^- = m \cap R_2^- = l\}\Big]_{m=1}^{m=M_1}{}_{l=1}^{l=M_2},$$

$$\Pi_n = [P\{L_2 = n \cap R_1 = m \cap R_2 = l\}]_{m=1}^{m=M_1}{}_{l=1}^{l=M_2},$$

and the matrix generating functions

$$\Pi_{L_2}^+(z) = \sum_{n=0}^\infty z^n \Pi_n^+, \quad \Pi_{L_2}^-(z) = \sum_{n=0}^\infty z^n \Pi_n^-, \text{ and } \Pi_{L_2}(z) = \sum_{n=0}^\infty z^n \Pi_n.$$

Exchanging L_2 with Q_2 we similarly define the generating functions $\Pi^+_{Q_2}(z)$, $\Pi^-_{Q_2}(z)$ and $\Pi_{Q_2}(z)$.

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 of Section 4. We will first establish relations between L_2^+, L_2^- and L_2 and L_2^+ and Q_2^+ that will be used in the analysis of preemptive systems.

Proposition 4 *Let* $\Pi^+_{L_2}(z)$, $\Pi^-_{L_2}(z)$ *and* $\Pi_{L_2}(z)$ *be the matrix pgf for the post-departures, the pre-transitions and the general time probabilities of a class 2 customer for a preemptive priority system satisfying the assumptions of Theorem 1. Then*

$$\Pi^-_{L_2}(z) = \Pi_{L_2}(z), \tag{47}$$

and

$$\lambda_2(1-z)\ \Pi^+_{L_2}(z) = (A_0 + A_1)'\ \Pi_{L_2}(z) + \ \Pi_{L_2}(z)\ (B_0 + zB_1). \tag{48}$$

Sketch of Proof

First we apply the uniformization technique to the two phase renewal processes and we choose the uniformization constant $\nu = \nu_1 + \nu_2$ such that $\nu_k \ \geq max\ \lambda_{k,i_k}$ for $k = 1, 2,\ i = 1, \ldots, M_k$. The epochs of transitions in both processes are therefore Poisson and thus (47) follows from PASTA.

In order to establish (48) we follow closely the approach of Bertsimas and Nakazato [1] to establish the relation between post-departures and the pre-transitions probabilities in stochastic processes with random upward and downward jumps. We first write down the flow balance equations for all states, where each state has four indexes corresponding to the two phase type arrival processes, and then we use the fact that our priority discipline is preemptive, i.e., class 2 departures can only happen if there are no class 1 customers in the system. Finally, by taking generating functions in the number of low priority customers in the system we obtain (48). The computations are algebraically involved but conceptually simple. □

Proposition 5 *In a preemptive single server priority system with two classes of customers each arriving according to a generalized Erlang distribution :*

$$\vec{e}_i' \, \Pi_{L_2}^{+}(z) \, \vec{e}_j = \sum_{k=1}^{M_1} \vec{e}_k' \, \Pi_{Q_2}^{+}(z) \, \Phi_{G_2^{k,i}}(B_0 + zB_1) \, \vec{e}_j. \tag{49}$$

Proof

Conditioning on the state of the queue that a class 2 customer left behind at the moment he started service and the duration of the effective service time we obtain

$$P\left\{L_2^+ = n, R_1^+ = i, R_2^+ = j\right\} =$$
$$\sum_{k=0}^{n} \sum_{m=1}^{M_1} \sum_{l=1}^{M_2} P\left\{Q_2^+ = k, R_1^+ = m, R_2^+ = l\right\} \int_0^\infty b_l(t) * b^{(n-k-1)}(t) * b_1^j(t) \, dF_{G_2^{m,i}}(t).$$

By writing the previous equation in matrix form we obtain (49). □

Let E_k, $k = 1, 2$ be the number of class k customers in queue given that no class k customer is in the service box. Let Δ_k be the number of class k customers in queue given that there is a class k customer in the service box. We introduce the matrix generating functions

$$\Pi_{E_2}(z) = \sum_{n=0}^{\infty} z^n \left[P\{E_2 = n \cap R_1 = i \cap R_2 = j\}\right]_{i=1}^{i=M_1} {}_{j=1}^{j=M_2} ,$$

$$\Pi_{\Delta_2}(z) = \sum_{n=0}^{\infty} z^n \left[P\{\Delta_2 = n \cap R_1 = i \cap R_2 = j\}\right]_{i=1}^{i=M_1} {}_{j=1}^{j=M_2} .$$

Furthermore, let Ξ be an $M_1 \times M_2$ matrix and $\vec{H}_2$ be an M_2 vector such that $\Xi_{i,j} = P\{R_1 = i, R_2 = j \mid L_1 = 0, L_2 = 0\}$ and $H_{2r} = P\{R_2^a = r \mid L_1^a = 0, L_2^a = 0\}$. Finally, let $B_{1,i}^*$ be the forward recurrence time of a class 1 busy period that ended while $ATC_1 = i$. Then the Laplace transform of $B_{1,i}^*$, $\sigma_{1,i}^*(s)$ is given by :

$$\sigma_{1,i}^*(s) = \frac{1 - [\Sigma_1(s)]_{1,i}}{sE[B_{1,i}]}.$$

We also introduce the traffic intensities $\rho_i = \lambda_i E[X_i]$, $\rho = \rho_1 + \rho_2$ and we define $p_{\Delta_i} = P\{\text{one class } i \text{ customer is in the service box}\}$. Our main result is

Theorem 12 *In a preemptive queueing system with two priority classes each of which satisfies the assumptions of Theorem 1 and has mixed generalized Erlang interarrival times characterized by matrices A_0, A_1 and B_0, B_1 respectively, the matrix generating function of the number of low priority customers in queue is calculated as a function of the system characteristics and the effective service time matrix from the following algorithm:*

1. *Calculate the matrix generating function $\Pi_{E_2(z)}$ such that:*

$$\Pi_{E_2(z)}\,\vec{e}_i = (1-\rho_1)\;\Xi\,\vec{e}_i + \rho_1\;\vec{H}_2{}'\;\Phi_{B^*_{1,i}}(B_0 + zB_1), \qquad i = 1,\ldots,M_2, \tag{50}$$

where

$$\Xi_{i,j} = \frac{1 - p_{1,i}\lambda_{1,i}E[X_1] - p_{2,j}\lambda_{2,j}E[X_2]}{1-\rho_1-\rho_2}\,\frac{\lambda_1}{\lambda_{1,i}}\prod_{r=1}^{i-1}(1-p_{1,r})\frac{\lambda_2}{\lambda_{2,j}}\prod_{r=1}^{j-1}(1-p_{2,r}),$$

$$H_{2r} = \frac{\sum_{l=1}^{M_1}\lambda_{1,l}p_{1,l}\Xi_{l,r}}{\sum_{l=1}^{M_1}\sum_{r=1}^{M_2}\lambda_{1,l}p_{1,l}\Xi_{l,r}},$$

and $B^*_{1,i}$ is the forward recurrence time of a class 1 busy period that ends while $ATC_1 = i$, and has Laplace transform $\sigma^*_{1,i}(s) = \frac{1-[\Sigma_1(s)]_{1,i}}{sE[B_{1,i}]}$

2. *For $i = 1,\ldots,M_1$ solve the system that would give the postdeparture probabilities*

$$\sum_{k=1}^{M_1}\vec{e}_k{}'\;\Pi^+_{Q_2}(z)\;\Phi_{G_2^{k,i}}(B_0 + zB_1)\;\; - z\;\vec{e}_i{}'\;\Pi^+_{Q_2}(z) =$$
$$\vec{e}_i{}'\frac{1}{\lambda_2}(1 - p_{\Delta_2})\left[(A_0 + A_1)'\;\Pi_{E_2}(z) + \;\Pi_{E_2}(z)(B_0 + zB_1),\right] \tag{51}$$

The constant p_{Δ_2} is calculated from the relation

$$\lim_{z\to 1}\vec{1}'\;\Pi^+_{Q_2}(z)\;\vec{1} = 1.$$

3. *The general time queue length distribution for class 2 customers is calculated by solving the system*

$$(A_0 + A_1)'\;\Pi_{Q_2}(z) + \;\Pi_{Q_2}(z)\;(B_0 + zB_1) = \lambda_2(1-z)\;\Pi^+_{Q_2}(z).$$

4. *The waiting time distribution for class 2 customers can be calculated by applying the distributional law*

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

Proof

Following the exact method of analysis of Section 4, our strategy for calculating the queue length distribution for class 2 customers is to find two relations between $\Pi_{L_2}(z)$ and $\Pi_{Q_2}(z)$ and then solve the underlying linear system. In (49) we have found the first relation between $\Pi^+_{L_2}(z)$ and $\Pi^+_{Q_2}(z)$. In order to find the second relation we condition on whether there is a class 2 customer in the service box and we obtain that

$$\Pi_{Q_2}(z) = p_{\Delta_2}\ \Pi_{\Delta_2(z)} + (1 - p_{\Delta_2})\ \Pi_{E_2(z)},$$

$$\Pi_{L_2}(z) = zp_{\Delta_2}\ \Pi_{\Delta_2}(z) + (1 - p_{\Delta_2})\ \Pi_{E_2}(z).$$

Hence,

$$\Pi_{L_2}(z) = z\Pi_{Q_2}(z) + (1-z)(1 - p_{\Delta_2})\ \Pi_{E_2}(z). \tag{52}$$

In order to find $\Pi_{E_2}(z)$ we use the following argument:

Because of the preemptive discipline, class 1 customers are not influenced by the fact that there is no low priority customer in the service box; so the server serves a class 1 customer with probability ρ_1 and does not serve class 1 customers with probability $1 - \rho_1$. In order for a random observer to see $n \geq 1$ 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 H_{2r} 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 $n \geq 1$

$$P\{E_2 = n,\ R_1 = i, R_2 = j\} = \ \rho_1 \sum_{r=1}^{M_2} H_{2r} \int_0^\infty b_r(t) * b^{(n-1)}(t) * b_1^j(t) dF_{B^\bullet_{1,i}}(t).$$

Similarly,

$$P\{E_2 = 0,\ R_1 = i, R_2 = j\} =$$

$$= (1 - \ \rho_1)P\{R_1 = i, R_2 = j \mid L_1 = 0, L_2 = 0\} + \ \rho_1 \sum_{r=1}^{M_2} H_{2r} \int_0^\infty b_r^j(t) dF_{B^\bullet_{1,i}}(t).$$

where $F_{B^{\bullet}_{1,i}}(t)$ is the cdf of the forward recurrence time of a class 1 busy period that ends with the class 1 customer being in stage i. Taking generating functions (50) follows.

We now proceed to calculate the constants appearing in (50).

$$\Xi_{i,j} = P\{R_1 = i, R_2 = j \mid L_1 = 0, L_2 = 0\} = \frac{P\{L_1 = 0, L_2 = 0, R_1 = i, R_2 = j\}}{P\{L_1 = 0, L_2 = 0\}}. \quad (53)$$

By applying Little's law to the server we obtain

$$P\{L_1 = 0,\ L_2 = 0 |\ R_1 = i,\ R_2 = j\} = 1 - p_{1,i}\lambda_{1,i}E[X_1] - p_{2,j}\lambda_{2,j}E[X_2],$$

and therefore, using (30) we have

$$P\{L_1 = 0,\ L_2 = 0,\ R_1 = i,\ R_2 = j\} =$$
$$\{1 - p_{1,i}\lambda_{1,i}E[X_1] - p_{2,j}\lambda_{2,j}E[X_2]\}\frac{\lambda_1}{\lambda_{1,i}}\prod_{r=1}^{i-1}(1 - p_{1,r})\frac{\lambda_2}{\lambda_{2,j}}\prod_{r=1}^{j-1}(1 - p_{2,r}).$$

We also know that $P\{L_1 = 0, L_2 = 0\} = 1 - \ \rho_1 - \ \rho_2$ and by substituting to (53) we obtain $\Xi_{i,j}$.

Finally

$$H_{2r} = P\{R^a_2 = r \mid L^a_1 = 0, L^a_2 = 0\} \ = \ \frac{\sum_{l=1}^{M_1} P\{L^a_1 = 0, L^a_2 = 0, R^a_1 = l, R^a_2 = r\}}{P\{L^a_1 = 0, L^a_2 = 0\}}.$$

But, because of the uniformization

$$P\{L^a_1 = 0, L^a_2 = 0, R^a_1 = l, R^a_2 = r\} = \lambda_{1,l}p_{1,l}P\{L_1 = 0, L_2 = 0, R_1 = l, R_2 = r\}$$

and thus

$$H_{2r} = \frac{\sum_{l=1}^{M_1} \lambda_{1,l}p_{1,l}\Xi_{l,r}}{\sum_{l=1}^{M_1}\sum_{r=1}^{M_2} \lambda_{1,l}p_{1,l}\Xi_{l,r}}.$$

Multiplying (52) with $(A_0 + A_1)'$ from the left and with $(B_0 + zB_1)$ from the right and using (49) we obtain (51). Notice that (51) determines $\Pi^+_{Q_2}(z)$ up to the constant p_{Δ_2} which is calculated from the relation

$$\lim_{z\to 1} \vec{1}'\ \Pi^+_{Q_2}(z)\ \vec{1} = 1.$$

Having found $\Pi^+_{Q_2}(z)$, we find $\Pi_{Q_2}(z)$ from (48), while the waiting time W_2 can be calculated by applying the distributional law (11):

$$\vec{1}'\Pi_{Q_2}(z) = \lambda_2(1-z)\vec{e_1}\,'\Phi_{W_2}(B_0 + zB_1)(B_0 + zB_1)^{-1}. \quad \square$$

Remarks:

1. In the case of Poisson arrival processes (51) gives :

$$\Pi_2(z) = \frac{1}{\lambda_2}(1 - p_{\Delta_2})(1-\rho_1)\frac{\lambda_2(1-z) + \lambda_1(1-\sigma_1(\lambda_2 - \lambda_2 z))}{\phi_{G_2}(\lambda_2 - \lambda_2 z) - z},$$

which is exactly the relation obtained in Keilson and Servi [11] using a different derivation. The probability p_{Δ_2} can be obtained either by requiring $\lim_{z\to 1} \Pi^+_{Q_2}(z) = 1$, which in this case leads to $p_{\Delta_2} = \lambda_2 E[G_2]$ or by applying Little's law in the service box.

2. The system time for class 2 customers is found from

$$S_2 = W_2 \oplus G_2,$$

while (52) offers a a way to calculate the distribution of the number of class 2 customers in the system once the distribution of the number of class 2 customers in the queue is determined.

6.4 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 stage of 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:

$$\vec{P}_{\Delta_1}(z) = \sum_{z=0}^{\infty} z^n [P\{\Delta_1 = n \cap R_1 = i\}]_{i=1}^{i=M_1},$$

and the scalar generating function $G_{\Delta_1}(z) = \sum_{z=0}^{\infty} z^n P\{\Delta_1 = n\}$. We also introduce the row vectors $\vec{E}$ and $\vec{H}_1$, such that:

$$E_i = P\{L_1 = 0, L_2 = 0, R_1 = i\} \quad \text{and} \quad H_{1r} = P\{R_1^{bs} = r\}.$$

Finally we will use $\vec{H}$ as defined in Section 4, i.e

$$H_i = \frac{\lambda_1}{\lambda_{1,i}}(1 - \lambda_{1,i} p_{1,i} E[X_1]) \prod_{k=1}^{i-1} (1 - p_{1,k}).$$

Theorem 13 *In a non-preemptive queueing system with two priority classes each of which satisfies the assumptions of Theorem 1 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 :*

$$\vec{P}_{Q_1}(z) = (1 - z)[\rho_2 \vec{H}_1 \Phi_{X_2^*}(A_0 + zA_1) + \vec{E}]\,[\Phi_{X_1}(A_0 + zA_1) - zI]^{-1}, \tag{54}$$

$$\vec{P}_{L_1}(z) = (1 - z)[\rho_2 \vec{H}_1\ \Phi_{X_2^*}(A_0 + zA_1) + \vec{E}]\,[\Phi_{X_1}(A_0 + zA_1) - zI]^{-1}\ \Phi_{X_1}(A_0 + zA_1), \tag{55}$$

where,

$$E_i = \{1 - p_{1,i}\lambda_{1,i} E[X_1] - \lambda_2 E[X_2]\} \frac{\lambda_1}{\lambda_{1,i}} \prod_{k=1}^{i-1} (1 - p_{1,k}), \tag{56}$$

and $\vec{H}_1$ satisfies :

$$\rho_2 \vec{H}_1 \; \Phi_{X_2^*}(A_0 + A_1)\vec{e}_i = \frac{\lambda_1}{\lambda_{1,i}}(1 - \lambda_{1,i} p_{1,i} E[X_1]) \prod_{k=1}^{i-1} (1 - p_{1,k}) - E_i$$

Proof

From the vector distributional law (28) we have:

$$\vec{P}_{L_1}(z) = \vec{P}_{Q_1}(z)\Phi_{X_1}(A_0 + zA_1). \tag{57}$$

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 $P\{B_i\} = \rho_i$ and by conditioning on the events B_i we have, for $n \geq 1$:

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

or by using the definition of Δ_1 :

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

and for $n = 0$ we also have :

$$\begin{aligned} &P\{Q_1 = 0, R_1 = i\} = \\ &\quad \rho_1 P\{\Delta_1 = 0, R_1 = i\} + \rho_2 P\{Q_1 = 0, R_1 = i | B_2\} + P\{L_1 = 0, L_2 = 0, R_1 = i\} \end{aligned}$$

or equivalently :

$$P\{Q_1 = 0, R_1 = i\} = \rho_1 P\{\Delta_1 = 0, R_1 = i\} + \rho_2 P\{Q_1 = 0, R_1 = i | B_2\} + E_i.$$

Furthermore, if we denote by H_{1r} the probability that $ATC_1 = r$ just before a type 2 customer enters service, we have that for $n \geq 1$:

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

and $\quad P\{Q_1 = 0, R_1 = i|B_2\} = \sum_{r=1}^{M_1} H_{1r} \int_0^\infty a_r^i(t)\, dF_{X_2^\bullet}(t)$.

By taking generating vector functions we get :

$$\vec{P}_{Q_1}(z) = \rho_1 \vec{P}_{\Delta_1}(z) + \rho_2 \vec{H}_1 \ \Phi_{X_2^\bullet}(A_0 + zA_1) + \vec{E}.$$

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

$$\vec{P}_{L_1}(z) = \rho_1 z \vec{P}_{\Delta_1}(z) + \rho_2 \vec{H}_1 \ \Phi_{X_2^\bullet}(A_0 + zA_1) + \vec{E}.$$

Combining the last two equations we have:

$$\vec{P}_{L_1}(z) = z\vec{P}_{Q_1}(z) + \rho_2(1-z)\vec{H}_2 \ \Phi_{X_2^\bullet}(A_0 + zA_1) + (1-z)\vec{E}. \tag{58}$$

From (57) and (58) we obtain (54) and (55). Finally we need to calculate the vectors $\vec{H}_1$ and $\vec{E}$. First note that $E_i = P\{L_1 = 0, L_2 = 0, R_1 = i\}$, so by applying Little's law to the server we get (56). In order to calculate $\vec{H}_1$ we recall that in a regular $MGE_M/G/1$ queue (25) 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

$$\rho_2 \vec{H}_1 \ \Phi_{X_2^\bullet}(A_0 + A_1) = \vec{H} - \vec{E},$$

where $H_i \ = \ \frac{\lambda_1}{\lambda_{1,i}}(1 - \lambda_{1,i} p_{1,i} E[X_1]) \prod_{k=1}^{i-1}(1 - p_{1,k})$. □

Remarks :

- Using (54) and (55) 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.

- Note that once again for Poisson arrivals (54) take the form:

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

  which is the exactly the result obtained in Keilson and Servi [11].

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 Keilson and Servi [11]). 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 $S_2 = W_2 \oplus X_2$.

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, and systems, which allow overtaking, which 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 only provide a partial answer (only for overtake free systems) to the question we raised in the first section of the paper regarding the laws of queueing theory. The major open problem is to identify queueing laws for systems that allow overtaking, which lead a complete solution. This is a rather challenging but important problem as 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.

References

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

[2] 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.

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

[4] Doshi B. (1985). "A note on Stochastic decomposition in a GI/G/1 queue with vacations or setup times", *Jour. Appl. Prob.*, 22, 419-428.

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

[6] Fuhrmann S.W. (1985). "Symmetric queues served in cyclic order", *Operation Research Letters*, Vol. 4, 3, 139-144.

[7] Haji R. and Newell G. (1971). "A relation between stationary queue and waiting time distributions", *Jour. Appl. Prob.*, 8, 617-620.

[8] Heyman D. and Sobel M. (1982). *Stochastic models in Operations Research*, Vol. 1, New York.

[9] Jaiswal N.K. (1968) *Priority Queues*, Academic Press, New York.

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

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

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

[13] Little J. (1961). "A proof of the theorem $L = \lambda W$", *Operations Research*, Vol.9, 383-387.

[14] Ross S. (1985). *Introduction to probability models*, 3rd edition, Academic Press, Florida.

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

[16] Whitt W. (1991). "A review of $L = \lambda W$ and extensions", *Queueing Systems.*

[17] Wollf R. (1989) *Stochastic modeling and the theory of queues*, Prentice Hall.