Queueing Systems with Appointment-Driven Arrivals, Non-Punctual Customers, and No-Shows

Oualid Jouini\textsuperscript{1} \qquad Saif Benjaafar\textsuperscript{2}

\textsuperscript{1}Laboratoire Génie Industriel, Ecole Centrale Paris, Grande Voie des Vignes, 92290 Châtenay-Malabry, France

\textsuperscript{2}Industrial and System Engineering Program, University of Minnesota, 111 Church Street SE, Minneapolis, MN 55455, USA

oualid.jouini@ecp.fr \quad saif@umn.edu

Working paper, May 2010

Abstract

We consider queueing systems where a finite number of customers arrive over time to a service system, consisting of either a single or multiple servers. The arrival of customers is driven by appointments, with a scheduled appointment time associated with each customer. However, customers are not necessarily punctual and may arrive either earlier or later than their scheduled appointment times. Customers may also not show up altogether. The arrival times of customers (relative to their scheduled appointments) and their service times are both stochastic. Customers are not homogeneous in their punctuality, show-up probabilities, and time between previous and subsequent appointments, which may vary from customer to customer. We develop an exact analytical approach to obtain various performance measures of interest and illustrate the usefulness of the approach by describing numerical results that examine the impact of not accounting for non-punctuality and no-shows.

Keywords: appointment-driven arrivals; finite arrivals; queueing systems; customer punctuality; customer no-shows.
1 Introduction

There are numerous service systems where the arrivals of customers are driven by scheduled appointments. Examples include arrivals to healthcare facilities, government agencies (e.g., immigration, social services, and internal revenue), the offices of tax and financial service providers, academic advising offices at universities, restaurants and spa treatment facilities, just to name a few. Despite this prevalence, analytical tools for the performance evaluation of these systems are relatively limited. Existing approaches from queueing theory cannot be readily applied because of several important differences between standard queueing systems and systems with appointment-driven arrivals (ADA). Systems with ADA are characterized by (1) a finite number of customers (e.g., the set of patients that have been scheduled at a clinic in a given day), so that steady state analysis cannot be applied, (2) arrivals that are in part determined by known scheduled appointment times, (3) appointment times that may not be equally spaced, and (4) the possibility of customer non-punctuality and no-shows. The difficulty of the analysis can be further compounded in settings in which customers are heterogeneous in their service time requirements, punctuality, and no-show probabilities.

In this paper, we consider such a setting. In particular, we consider a system with a finite number of customers, where each customer has a scheduled appointment, but customers are not necessarily punctual and may arrive earlier or later than their appointment times. Customers may also not show up. We allow for the degree of punctuality, a random variable with a general distribution, and the probability of show-up to be customer-specific. We consider service systems with both single and multiple servers, with exponentially distributed service times. In the single server setting, service times can be non-identically distributed. We allow for appointments to be arbitrarily spaced, so that they are not necessarily equally spaced. Under a relatively mild condition on customer arrivals, namely that customers arrive in the order of their appointments times, we develop an exact analytical approach and obtain various performance measures related to customer waiting time. We illustrate the usefulness of the approach by describing numerical results that examine the impact of not accounting for non-punctuality and no-shows and show that doing so can lead to significant errors.

To our knowledge there are no papers that consider simultaneously appointment driven arrivals, non-punctuality, and no-shows, and to do so for a setting as general as ours. There is of course an extensive literature on systems where arrivals are determined by appointments times (the typical application is appointment scheduling in healthcare); see for example the reviews in Mondschein and Weintraub (2003), Gupta and Denton (2008), Cayirli and Veral (2003), and Preater (2001). However, in nearly all of that literature, customers are assumed to arrive on
time. In most of these papers, performance evaluation is carried out using simulation or traditional queueing analysis where steady state behavior, with an infinite number of arrivals and independent and identically distributed inter-arrival and service times, is assumed. There are few papers that consider no-shows. Examples include Kaandorp and Koole (2007), Hutzschenreuter (2005), and Hassin and Mendel (2008), and the references therein. However, in all of this literature customers that do show up are assumed to arrive on time and their service times are assumed to be identically distributed. The treatment in this literature is also limited to single server systems. Green and Savin (2008) consider a single server model with finite buffer, Poisson arrivals, and deterministic processing times. They model no-shows as customers that join the queue again immediately upon completing service. Their model does not capture punctuality (arrivals are not driven by specified appointments) and assumes an infinite number of arrivals. Parlar and Sharafali (2008) do consider a model with a finite number of arrivals motivated by the arrival process of customers to check-in counters for a particular flight. However, in their case, customers arrive independently of each other, with arrivals modeled as a “death process” from the population of N travelers booked on the flight. Mercer (1973) considers a system with appointment-driven arrivals but in his case the number of arrivals is infinite, the appointment times are equally spaced and customers have identical show up probabilities, service times and lateness.

In Section 2 we treat the case of a single server model. In Section 3, we consider various special cases. In Section 4, we extend the analysis to the multi-server case. In Section 5, we present numerical results. Finally in Section 6, we provide some concluding remarks and directions for future research.

2 Problem Description and Analysis

In this section we describe the single server model under consideration. We then provide a method to characterize the customer waiting time in the queue by deriving its probability distribution function. For ease of exposition, we first consider the case where service times are homogeneous (Section 2.1). We then extend the analysis to the case where service times are heterogeneous (Section 2.2).

We consider a queueing system with a single server and a finite number of customers that arrive over time. Successive service times are independent and exponentially distributed with mean service $\frac{1}{\mu_n}$ for the $n$-th customer\(^1\). There are $M$ customers that are scheduled to arrive.

\(^1\)This assumption, which is common in modeling in service operations, is reasonable for systems with high service time variability where service times are typically small but there are occasionally long service times.
We denote by $d_n$, for $n = 1, ..., M$, the appointment time of the $n$-th customer. We index the customers by their appointment times and we assume that $d_n \leq d_m$ if $n < m$. Customer $n$ has a probability $\alpha_n$ of showing up, independently of all other events. If a customer shows up, she may do so earlier or later than her appointment time. More specifically, the customer may show up at a random time between $d_n - l_n$ and $d_n + u_n$. In other words, the arrival time of customer $n$ can be described by a random variable with finite support $d_n - l_n$ and $d_n + u_n$. We refer to this random variable as $D_n$ and allow it to have a general distribution with probability distribution function (pdf) $f_n$. Note that we allow for this distribution function to be customer specific\(^2\). For mathematical tractability, we assume that customers arrive in the same order as their appointment times, so that $D_n < D_{n+1}$ or equivalently $d_n - d_{n-1} \geq \tau_{n-1}^u + \tau_n^l$. In other words, customer arrival times are non-overlapping. This assumption is reasonable in settings where the times between appointments are large relative to customer lateness (for example, appointments are 45 minutes apart but customers are at most 30 minutes early or late). A useful extension of our analysis would of course do away with this assumption and we leave this as a potential area for future research. We place no other assumptions on the distribution of customer arrival times.

Upon arrival, a customer goes immediately into service if the server is available. If not, the customer joins the queue where she waits for service. We assume that the server (e.g., the physician in a healthcare clinic) is available to start work exactly at $d_1$ (the scheduled time of the first customer). The server remains available until the last customer has completed service. The server has no prior knowledge of whether or not a particular customer will show up. Therefore if customer $M$ shows up, the server shuts down as soon as customer $M$ completes service. If not, the server shuts down at time $d_M + u_M$, i.e., after the latest possible arrival time of customer $M$. We assume that customers are processed in the order of their appointment times. We also assume that the system is work-conserving with the server never idle when there are customers in the queue.

### 2.1 The Case of Homogeneous Service Times

In this section we consider the case where service times are homogenous with $\mu_n = \mu$ for $n = 1, ..., M$. Our approach consists of first computing the stationary probabilities of the system state seen by a new arrival. We then compute the conditional waiting time, given the system state. Finally, we characterize the unconditional waiting time by averaging over all

---

\(^2\)This is an important feature in applications, such as healthcare, where data may be available, or can be collected, on the punctuality of different customers and where punctuality of different customers can vary significantly.
possibilities.

Let us consider a given schedule denoted by the vector $\delta = (d_1, \ldots, d_M)$. Without loss of generality we always choose $d_1 = \tau_1^l$ so that the origin of time is $d_1 - \tau_1^l$. We denote by $R_n$ the random variable that describes the number of customers found (would have been found) in the system by customer $n$ if she shows up (does not show up). This means that the total number of customers in the system at time $D_n$ is $R_n + 1$ ($R_n$) if customer $n$ shows up (does not show up). We let $p_{n,i} = \Pr\{R_n = i\}$ refer to the probability that the $n$-th customer finds (would have found), upon arrival if she shows up (she does not), $i$ customers already in the system (i.e., in the queue or in service), for $0 \leq i \leq n - 1$ and $1 \leq n \leq M$.

In what follows, we first characterize the probabilities $p_{n,i}$. Let $X_n$ be the random variable describing the inter-arrival time between customers $n$ and $n + 1$ (note that we associate an arrival time with a customer regardless of whether or not she actually shows; if the customer does not show; we refer to this as a virtual arrival), where $X_n = D_{n+1} - D_n$ for $1 \leq n \leq M - 1$. We denote by $h_n(.)$ the pdf of the random variable $X_n$. We have $d_{n+1} - d_n - \tau_{n+1}^l - \tau_n^u \leq X_n \leq d_{n+1} - d_n + \tau_n^u + \tau_l^l$. Note that the random variables $D_n$ and $D_{n+1}$ do not overlap; thus, $X_n \geq 0$.

For $n = 1$, $p_{1,0} = 1$ and $p_{1,i} = 0$ for $i \neq 0$, since the first customer always finds the system empty if she shows up. However she may have to wait to start service because the server starts work exactly at $d_1$. We shall discuss this matter later in this section. For $n = 2$, we have $p_{2,0} = 1 - p_{2,1}$. In what follows, we compute $p_{2,1}$. To do so, we separate the cases of whether customer 1 arrives early ($D_1 < d_1$), or she arrives late ($D_1 \geq d_1$). Recall that $d_1 = \tau_1^l$; then the probability $p_{2,1}$ may be written as

$$p_{2,1} = \alpha_1 \Pr\{D_1 < d_1\} p_{2,1|D_1<d_1} + \alpha_1 \Pr\{D_1 > d_1\} p_{2,1|D_1>d_1}, \tag{1}$$

$$= \alpha_1 \left( \int_0^{d_1} f_1(x) \, dx \right) p_{2,1|D_1<d_1} + \alpha_1 \left( \int_{d_1}^{d_1+\tau_l^l} f_1(x) \, dx \right) p_{2,1|D_1>d_1},$$

where $p_{2,1|D_1<d_1}$ ($p_{2,1|D_1>d_1}$) is the conditional probability that customer 2 sees customer 1 in the system upon arrival, given that customer 1 arrives early (late). Given that customer 1 arrives early, we have

$$p_{2,1|D_1<d_1} = \Pr\{D_2 - d_1 < \varepsilon_\mu\}, \tag{2}$$

where $\varepsilon_\mu$ is an exponential random variable with rate $\mu$. Then

$$p_{2,1|D_1<d_1} = \int_0^{d_2-d_1+\tau_l^l} \Pr\{x < \varepsilon_\mu\} f_2(x+d_1) \, dx = \int_{d_2-d_1-\tau_l^l}^{d_2-d_1+\tau_l^l} e^{-\mu x} f_2(x+d_1) \, dx. \tag{3}$$
Given that customer 1 arrives late, we may write

\[ p_{2,1|D_1 > d_1} = \Pr \{ D_2 - D_1 < \varepsilon_\mu \mid D_1 > d_1 \} = \int_{d_2 - d_1 - \tau_2^u}^{d_2 - d_1 + \tau_2^u} e^{-\mu x} h_{1|D_1 > d_1} (x) \, dx, \quad (4) \]

where \( h_{1|D_1 > d_1} (x) \) defined on \( d_2 - d_1 - \tau_2^u \leq x \leq d_2 - d_1 + \tau_2^u \) is the pdf of the random variable \( \hat{X}_1 = D_2 - D_1 \mid D_1 > d_1 \); the conditional inter-arrival time given that customer 1 arrives late. Using Equations (3) and (4) and substituting in Equation (2), we obtain \( p_{2,1} \) and then \( p_{2,0} \).

For \( 3 \leq n \leq M \), we separate the cases \( i = 0 \) and \( 1 \leq i \leq n - 1 \) to compute \( p_{n,i} \). Consider first \( p_{n,i} \) for \( 3 \leq n \leq M \) and \( 1 \leq i \leq n - 1 \). We condition on the number of customers found upon arrival by customer \( n - 1 \). One has

\[ p_{n,i} = \sum_{j=1}^{n-2} p_{n-1,j} \Pr \{ R_n = i \mid R_{n-1} = j \}. \quad (5) \]

We distinguish the cases that customer \( n - 1 \) shows up or not. For customer \( n \) to find \( i \) customers upon arrival given that customer \( n - 1 \) found (would have found) \( j \) customers, there must be \( j - i + 1 \) service completions (\( j - i \) service completions) between the arrival times of customer \( n - 1 \) and customer \( n \). This also means that, once customer \( n - 1 \) arrives but before customer \( n \) does (the duration is \( X_{n-1} \)), exactly \( j - i + 1 \) service completions (\( j - i \) service completions) occur. Since the server has an exponential service time, the number of customers served during \( X_{n-1} \) follows a Poisson process with rate \( \mu \). Thus, we have

\[ p_{n,i} = \alpha_{n-1} \sum_{j=1}^{n-2} p_{n-1,j} \int_{d_{n-1} - r_{n-1}^u - r_{n-1}^u}^{d_{n-1} - r_{n-1}^u + r_{n-1}^u} \frac{(\mu x)^{j-1-i}}{(j-1-i)!} e^{-\mu x} h_{n-1} (x) \, dx \]

\[ + (1 - \alpha_{n-1}) \sum_{j=1}^{n-2} p_{n-1,j} \int_{d_{n-1} - r_{n-1}^u - r_{n-1}^u}^{d_{n-1} - r_{n-1}^u + r_{n-1}^u} \frac{(\mu x)^{j-1-i}}{(j-1-i)!} e^{-\mu x} h_{n-1} (x) \, dx, \quad (6) \]

for \( 3 \leq n \leq M \) and \( 1 \leq i \leq n - 1 \). Note that \( \Pr \{ R_n = 0 \mid R_{n-1} = j \} \) may be viewed as the probability that at least \( j + 1 \) customers (\( j \) customers) are served during \( X_{n-1} \) if customer \( n - 1 \) shows up (if she does not). Using the expression of \( \Pr \{ R_n = i \mid R_{n-1} = j \} \) in Equation (6) leads to the following expression of \( p_{n,0} \)

\[ p_{n,0} = \alpha_{n-1} \sum_{j=0}^{n-2} p_{n-1,j} \left( 1 - \sum_{i=1}^{j+1} \int_{d_{n-1} - r_{n-1}^u - r_{n-1}^u}^{d_{n-1} - r_{n-1}^u + r_{n-1}^u} \frac{(\mu x)^{j+i-1}}{(j+i-1)!} e^{-\mu x} h_{n-1} (x) \, dx \right) \]

\[ + (1 - \alpha_{n-1}) \left( p_{n-1,0} + \sum_{j=1}^{n-2} p_{n-1,j} \left( 1 - \sum_{i=1}^{j} \int_{d_{n-1} - r_{n-1}^u - r_{n-1}^u}^{d_{n-1} - r_{n-1}^u + r_{n-1}^u} \frac{(\mu x)^{j+i-1}}{(j+i-1)!} e^{-\mu x} h_{n-1} (x) \, dx \right) \right), \quad (7) \]
for $3 \leq n \leq M$. An equivalent expression for $p_{n,0}$ is of course

$$p_{n,0} = 1 - \sum_{i=1}^{n-1} p_{n,i}, \quad (8)$$

for $1 \leq n \leq M$. Note that by convention an empty sum is equal to 0. Using the above expressions, the probabilities $p_{n,i}$ for $1 \leq n \leq M$ and $0 \leq i \leq n-1$ can now be computed recursively starting with $n = 1$.

Next we show how the above probabilities can be used to characterize various performance measures. Let $W_n$, a random variable, denote the waiting time in the queue of customer $n$, if she shows up, and let $E(W^k_n)$ be the corresponding $k$-th moment for $k \geq 1$. (For the rest of the paper, $E(Z)$ denotes the expected value of a given random variable $Z$ and $E(Z^k)$ the $k$-th moment). Then

$$E(W^k_n) = \sum_{i=1}^{n-1} p_{n,i} E(W^k_{n,i}), \quad 1 \leq n \leq M, \quad (9)$$

where $W_{n,i}$ is the random variable denoting the waiting time in queue of customer $n$, given that customer $n$ shows up and finds $i$ customers upon arrival, and $E(W^k_{n,i})$ is the corresponding $k$-th moment. Since service times are independent and exponentially distributed with parameter $\mu$, $W_{n,i}$ has an Erlang distribution with $i$ stages and parameter $\mu$. Using Equation (9) and knowing that $E(W_{n,i}) = \frac{i}{\mu}$ and $E(W^2_{n,i}) = \frac{i(i+1)}{\mu^2}$, we obtain

$$E(W_n) = \sum_{i=1}^{n-1} p_{n,i} \frac{i}{\mu} \quad \text{and} \quad E(W^2_n) = \sum_{i=1}^{n-1} p_{n,i} \frac{i(i+1)}{\mu^2}, \quad (10)$$

for $2 \leq n \leq M$.

The case $n = 1$ is treated separately. Assume Customer 1 shows up. If she arrives before $d_1$, then she has to wait for the server’s work starting time at $d_1$. If not, she immediately enters service with no waiting (recall that $d_1 - \tau_1^* = 0$). Then

$$E(W^k_1) = \int_0^{d_1} (d_1 - x)^k f_1(x) \, dx. \quad (11)$$

Let the random variable $W$ denote the waiting time in the queue of an arbitrary customer who shows up, then the $k$-th moment of $W$ is given by

$$E(W^k) = \frac{1}{M} \sum_{n=1}^{M} \alpha_n E(W^k_n). \quad (12)$$
From the probabilities \( p_{n,i} \), we can also characterize the distribution of \( W \). First, note that

\[
\Pr(W_{n,i} < t) = 1 - \sum_{j=0}^{i-1} \frac{(\mu t)^j}{j!} e^{-\mu t},
\]

for \( t \geq 0 \) and \( n \geq 2 \). Consequently, we have for \( n \geq 2 \)

\[
\Pr(W_n < t) = p_{n,0} + \sum_{i=1}^{n-1} p_{n,i} \Pr(W_{n,i} < t)
\]

\[= 1 - \sum_{i=1}^{n-1} \sum_{j=0}^{i-1} p_{n,i} \frac{(\mu t)^j}{j!} e^{-\mu t}.\]

As for \( n = 1 \), one has

\[
\Pr(W_1 < t) = \Pr(D_1 < d_1) \Pr(D_1 - d_1 < t \mid D_1 < d_1) + \Pr(D_1 > d_1)
\]

\[= (1 - \Pr(D_1 < d_1 - t \mid D_1 < d_1)) \Pr(D_1 < d_1) + \Pr(D_1 > d_1).\]

The pdf of the random variable \( \tilde{D}_1 = D_1 \mid D_1 < d_1 \) is defined on \([0,d_1]\) and is given by

\[
\frac{f_1(x)}{\Pr(D_1 < d_1)}.\]

After some algebra, Equation (15) leads to

\[
\Pr(W_1 < t) = 1 - \int_0^{\max(d_1 - t,0)} f_1(x) \, dx.
\]

Observing now that

\[
\Pr(W < t) = \frac{1}{M} \sum_{n=1}^{M} \alpha_n \Pr(W_n < t),
\]

we eventually obtain, after some manipulations,

\[
\Pr(W < t) = \sum_{n=1}^{M} \frac{\alpha_n}{M} - \frac{\alpha_1}{M} \int_0^{\max(d_1 - t,0)} f_1(x) \, dx - \frac{1}{M} \sum_{n=2}^{M} \sum_{i=1}^{n-1} \sum_{j=0}^{i-1} \alpha_n p_{n,i} \frac{(\mu t)^j}{j!} e^{-\mu t}.
\]

**Remark**

In some applications, system managers may prefer to exclude any waiting prior to the scheduled appointment. They indeed consider that waiting due to early arrival to be voluntary. In what follows we derive the expected waiting time (excluding any waiting prior to the appointment time) in the queue for customer \( n \), \( E(\hat{W}_n) \). For customer 1, we clearly have \( E(\hat{W}_1) = 0 \). If customer \( n \) shows up, we may write

\[
E(\hat{W}_n) = E(W_n) - \int_{d_n - \tau}^{d_n} (d_n - x) f_n(x) \, dx,
\]
for $2 \leq n \leq M$. Let $E(\hat{W})$ be the expected waiting time in the queue for an arbitrary customer who shows up. Then

$$E(\hat{W}) = \frac{1}{M} \sum_{n=2}^{M} \alpha_n E(\hat{W}_n).$$ \hfill (20)

Note also that

$$E(\hat{W}) = \frac{1}{M} \sum_{n=1}^{M} \alpha_n E(W_n) - \frac{1}{M} \sum_{n=1}^{M} \alpha_n \int_{d_n - \tau}^{d_n} (d_n - x) f_n(x) \, dx,$$

or equivalently

$$E(\hat{W}) = E(W) - \frac{1}{M} \sum_{n=1}^{M} \alpha_n \int_{d_n - \tau}^{d_n} (d_n - x) f_n(x) \, dx.$$ \hfill (21)

### 2.2 The Case of Heterogeneous Service Times

The analysis in Section 2.2 can be extended to the case of heterogeneous service times where the mean service times are now customer-dependent. In particular we now assume that the service time of customer $n$ follows an exponential distribution with rate $\mu_n$. In what follows, we describe the main differences in the analysis between the homogenous and heterogeneous cases.

We first characterize the probabilities $p_{n,i}$ for $0 \leq i \leq n-1$ and $1 \leq n \leq M$. For $n = 1$, $p_{1,0} = 1$ and $p_{1,i} = 0$ for $i \neq 0$, since the first customer always finds the system empty if she shows up. For $n = 2$, we have

$$p_{2,1} = \alpha_1 \left( \int_{0}^{d_1} f_1(x) \, dx \right) p_{2,1|D_1<d_1} + \alpha_1 \left( \int_{d_1}^{d_1+\tau_1} f_1(x) \, dx \right) p_{2,1|D_1>d_1},$$

where

$$p_{2,1|D_1<d_1} = \int_{d_2-d_1-\tau_1^2}^{d_2-d_1+\tau_1^2} \Pr\{x < \varepsilon_{\mu_1} \} f_2(x + d_1) \, dx = \int_{d_2-d_1-\tau_1^2}^{d_2-d_1+\tau_1^2} e^{-\mu_1 x} f_2(x + d_1) \, dx,$$ \hfill (24)

and

$$p_{2,1|D_1>d_1} = \Pr\{D_2 - D_1 < \varepsilon_{\mu_1} \mid D_1 > d_1\} = \int_{d_2-d_1-\tau_1^2}^{d_2-d_1+\tau_1^2} e^{-\mu_1 x} h_{1|D_1>d_1}(x) \, dx.$$ \hfill (25)

Next, we consider $p_{n,i}$ for $3 \leq n \leq M$ and $1 \leq i \leq n-1$. We condition on the number of customers found (would have been found) upon the arrival of customer $n-1$ if she shows up
Let us compute $\Pr\{R_n = i \mid R_{n-1} = j\}$. Assume first that customer $n-1$ shows up and finds $j$ customers in the system. In case of a system where all customers show up, these customers would be customer $n-j$, customer $n-j-1$, ..., customer $n-2$. Since the discipline of service is FCFS, customer $n-j-1$ would be the one being currently in service, and customers $n-j$, ..., $n-2$, $n-1$ are those waiting in the queue. They enter service according to the order $n-j$, $n-j+1$, ..., $n-2$, $n-1$. However in our system, a customer has the possibility of not showing up. Then if customer $n-1$ shows up and finds $j$ customers in the system, these $j$ customers are denoted by $S_{n-1,j,s} = \{S_{n-1,j,s}(1), \ldots, S_{n-1,j,s}(j)\} \subseteq \{1, 2, \ldots, n-2\}$, where $S_{n-1,j,s}(k) \neq S_{n-1,j,s}(l)$ if $k \neq l$. The equality $S_{n-1,j,s} \equiv \{1, 2, \ldots, n-2\}$ holds if $j = n-2$. Since each customer shows up according to a Bernoulli process, there are $\binom{n-2}{j}$ different and equiprobable (indexed by $s$) subsets $S_{n-1,j,s}$, each occurring with probability $\frac{1}{\binom{n-2}{j}}$, for $1 \leq s \leq \binom{n-2}{j}$.

We order the components of $S_{n-1,j,s}$ such that $S_{n-1,j,s}(k) < S_{n-1,j,s}(k+1)$ for $1 \leq k \leq j-1$. Thus, upon the arrival of customer $n-1$, the customer currently in service is customer $S_{n-1,j,s}(1)$. Customer $n-1$ joins customers $S_{n-1,j,s}(2)$, ..., $S_{n-1,j,s}(j)$ and wait with them in the queue. The order the latter enter service is $S_{n-1,j,s}(2)$, ..., $S_{n-1,j,s}(j)$, then $n-1$.

Therefore, $\Pr\{R_n = i \mid R_{n-1} = j\}$ may be viewed as the probability that customers $S_{n-1,j,s}(1)$, ..., $S_{n-1,j,s}(j-i+1)$ leave the system before customer $n$ would arrive, i.e., within the duration $X_{n-1}$. The remaining service time of customer $S_{n-1,j,s}(1)$ is still exponential with rate $\mu_{S_{n-1,j,s}(1)}$. Thus, we consider the following pure birth process: The states are $1$, $\ldots$, $j-i+2$. The rate at which the process moves from state $k$ to state $k+1$ is $\mu_{S_{n-1,j,s}(k)}$, for $1 \leq k \leq j-i+2$.

We denote by $Q_{k,l,s}(t)$ the transition probability function that the process is in state $l$ at time $t \geq 0$ given that it starts at state $k$ at $t = 0$. Then the pdf of the time within which customers $S_{n-1,j,s}(1)$, ..., $S_{n-1,j,s}(j-i+1)$ would leave the system is given by $Q_{1,j-i+2,s}(t)$. From the pure birth process, we write the following set of differential equations

$$
\begin{align*}
\frac{dQ_{k,l,s}(t)}{dt} &= -\mu_{S_{n-1,j,s}(l)}Q_{k,l,s}(t) + \mu_{l-1}Q_{k,l-1,s}(t), \quad k + 1 \leq l \\
\frac{dQ_{k,k,s}(t)}{dt} &= -\mu_{S_{n-1,j,s}(k)}Q_{k,k,s}(t).
\end{align*}
$$

(26)
From Kleinrock (1975) page 72, the solutions of Equations (27) are given by

\[
\begin{align*}
Q_{k,l,s}(t) &= \mu_{S_{n-1,j,s}(l-1)} e^{-\mu_{S_{n-1,j,s}(l)} t} \int_0^t Q_{k,l-1,s}(x) e^{\mu_{S_{n-1,j,s}(l)} x} \, dx, \quad k + 1 \leq l \\
Q_{k,k,s}(t) &= e^{-\mu_{S_{n-1,j,s}(k)} t}.
\end{align*}
\]  

(28)

Then

\[
\Pr\{R_n = i \mid R_{n-1} = j\} = \int_{d_n - d_{n-1} - \tau_n^u}^{d_n - d_{n-1} + \tau_n^u + \tau_n^l} \sum_{s=1}^{n-2} \frac{Q_{1,j-i+2,s}(x)}{{n-2\choose j}} h_{n-1}(x) \, dx,
\]

(29)

for \(i - 1 \leq j \leq n - 2\). Second, assume that customer \(n - 1\) does not show up. Customer \(n\) therefore finds \(i\) customers in the system if customers \(S_{n-1,j,s}(1), \ldots, S_{n-1,j,s}(j-i)\) leave the system during \(X_{n-1}\). As a consequence

\[
\Pr\{R_n = i \mid R_{n-1} = j\} = \int_{d_n - d_{n-1} - \tau_n^u}^{d_n - d_{n-1} + \tau_n^u + \tau_n^l} \sum_{s=1}^{n-2} \frac{Q_{1,j-i+1,s}(x)}{{n-2\choose j}} h_{n-1}(x) \, dx,
\]

(30)

for \(i \leq j \leq n - 2\). Thus

\[
p_{n,i} = \alpha_{n-1} \sum_{j=i-1}^{n-2} p_{n-1,j} \int_{d_n - d_{n-1} - \tau_n^u}^{d_n - d_{n-1} + \tau_n^u + \tau_n^l} \sum_{s=1}^{n-2} \frac{Q_{1,j-i+2,s}(x)}{{n-2\choose j}} h_{n-1}(x) \, dx
\]

\[
+ (1 - \alpha_{n-1}) \sum_{j=i}^{n-2} p_{n-1,j} \int_{d_n - d_{n-1} - \tau_n^u}^{d_n - d_{n-1} + \tau_n^u + \tau_n^l} \sum_{s=1}^{n-2} \frac{Q_{1,j-i+1,s}(x)}{{n-2\choose j}} h_{n-1}(x) \, dx.
\]

(31)

Finally for \(i = 0\), we have \(p_{n,0} = 1 - \sum_{i=1}^{n-1} p_{n,i}\). This completes the characterization of \(p_{n,i}\) for \(0 \leq i \leq n - 1\) and \(1 \leq n \leq M\).

The expected waiting time in the queue of customer \(n\), for \(2 \leq n \leq M\), is given by Equation (32). (For customer 1, this quantity can still be given by Equation (11)).

\[
E(W_n) = \sum_{i=1}^{n-1} p_{n,i} \frac{1}{{n-1\choose i}} \left( \sum_{s=1}^{n-1} \sum_{k \in S_{n,i,s}} \frac{1}{\mu_k} \right),
\]

(32)

where \(S_{n,i,s} = \{S_{n,i,s}(1), \ldots, S_{n,i,s}(i)\} \subseteq \{1, 2, \ldots, n - 1\}\).

Similarly to the case of homogeneous service times, one may characterize the \(k\)-th moment, for \(k \geq 2\), as well as the probability distribution function of the customer waiting time in the queue. This involves the characterization of the time it takes to empty the system of \(j\) customers having exponential service times with different rates \(\mu_1, \ldots, \mu_j\). Note that this random variable no longer has an Erlang distribution, instead its distribution is hypoexponential. We denote
by \( g_{\mu_1, \ldots, \mu_j}(\cdot) \) and \( G_{\mu_1, \ldots, \mu_j}(\cdot) \) the pdf and the cumulative distribution function (cdf) of this hypoeponential random variable, respectively. From Ross (1997), they are given, for \( x \geq 0 \), by

\[
g_{\mu_1, \ldots, \mu_j}(x) = \sum_{i=1}^{j} \left( \prod_{l=1, l \neq i}^{j} \frac{\mu_l}{\mu_l - \mu_i} \right) \mu_i e^{-\mu_i x}, \tag{33}\]

and

\[
G_{\mu_1, \ldots, \mu_j}(x) = \sum_{i=1}^{j} \left( \prod_{l=1, l \neq i}^{j} \frac{\mu_l}{\mu_l - \mu_i} \right) (1 - e^{-\mu_i x}). \tag{34}\]

### 3 Special Cases

Computing \( h_n \) involves computing convolution of random variables. For some special cases, it is possible to use the above results to obtain explicit expressions for the various performance measures. In this section, we do so for the case when arrival times follow uniform or triangle distributions. To simplify the presentation, we provide the analysis for the case of \( n = l = u = \tau \) for all \( n \).

#### 3.1 The case of Uniform Distribution

In this section, we assume that \( D_n \) follows a uniform distribution on \([d_n - \tau, d_n + \tau]\). That is, for \( 1 \leq n \leq M \), \( f_n(x) = \frac{1}{2\tau} \) if \( d_n - \tau \leq x \leq d_n + \tau \) and 0 otherwise. If customer \( n \) shows up, her expected arrival time is therefore \( d_n \). Consider the random variable \( X_{n-1} \) (inter-arrival time between customers \( n-1 \) and \( n \)) and let us compute its pdf, namely \( h_{n-1}(x) \). One has

\[
h_{n-1}(x) = \int_{-\infty}^{+\infty} f_{n-1}(u-x) f_n(u) \, du = \frac{1}{2\tau} \int_{d_n-\tau}^{d_n+\tau} f_{n-1}(u-x) \, du, \tag{35}\]

for \( 2 \leq n \leq M \). Knowing that \( f_{n-1}(u-x) \) is \( \frac{1}{2\tau} \) if \( d_{n-1} - \tau \leq u-x \leq d_{n-1} + \tau \) and is 0 otherwise, we obtain

\[
h_{n-1}(x) = \left\{ \begin{array}{ll}
\frac{1}{4\tau^2} \int_{d_n-d_{n-1}-2\tau}^{d_n-d_{n-1}-2\tau} dx + \frac{x+d_{n-1}-d_n}{4\tau^2}, & \text{for } d_n - d_{n-1} - 2\tau \leq x \leq d_n - d_{n-1} \\
\frac{1}{4\tau^2} \int_{d_n-d_{n-1}+2\tau}^{d_n-d_{n-1}+2\tau} dx - \frac{x+d_{n-1}-d_n}{4\tau^2}, & \text{for } d_n - d_{n-1} \leq x \leq d_n - d_{n-1} + 2\tau,
\end{array} \right. \tag{36}\]

which is a triangle distribution.

Let us now compute \( h_{1|D_1>d_1}(x) \), the pdf of the random variable \( X_1 = D_2 - D_1 | D_1 > d_1 \). One may see that \( h_{1|D_1>d_1}(x) \) is the convolution of \( f_2(x) \) defined on \([d_2-\tau, d_2+\tau]\) and \( f_{1|D_1>d_1}(x) = \frac{1}{\tau} \)
defined on \([-d_1 - \tau, -d_1]\), and is given by

\[
h_{1|D_1>d_1}(x) = \begin{cases} 
\frac{x+d_1 - d_2 + 2\tau}{2\tau^2}, & \text{for } d_2 - d_1 - 2\tau \leq x \leq d_2 - d_1 - \tau \\
\frac{1}{2\tau}, & \text{for } d_2 - d_1 - \tau \leq x \leq d_2 - d_1 \\
\frac{d_2 - d_1 + \tau - x}{2\tau^2}, & \text{for } d_2 - d_1 \leq x \leq d_2 - d_1 + \tau
\end{cases}.
\] (37)

Observing that

\[
\Pr\{D_1 < d_1\} = \Pr\{D_1 > d_1\} = \frac{1}{2},
\] (38)

and combining Equations (1)-(3) leads to

\[
p_{2,1} = \alpha_1 \frac{e^{-\mu(d_2-d_1)}}{4\mu^2\tau^2} \left( e^{\mu\tau} - e^{-\mu\tau} \right) \left( e^{\mu\tau} + \tau\mu - 1 \right).
\] (39)

The random variable measuring the duration of the \(j\) service completions has an Erlang distribution with \(j\) stages and parameter \(\mu\), say \(Er(j, \mu)\). Let \(g_j(.)\) and \(G_j(.)\) be the pdf and the cdf of \(Er(j, \mu)\), respectively. From Ross (1997), we have

\[
g_j(x) = \frac{\mu^j x^{j-1} e^{-\mu x}}{(j-1)!}, \text{ and } G_j(x) = 1 - \sum_{i=0}^{j-1} \frac{(\mu x)^i}{i!} e^{-\mu x},
\] (40)

for \(x \geq 0\) and \(j \geq 0\). Observe from Equation (40) that the Integral of \(\frac{\mu^j x^{j-1} e^{-\mu x}}{(j-1)!}\) is

\[
1 - \sum_{i=0}^{j-1} \frac{(\mu x)^i}{i!} e^{-\mu x}.
\]

Using Equation (7), we can then obtain

\[
p_{n,0} = 1 - \frac{\alpha_{n-1}}{4\tau^2} \sum_{j=0}^{n-2} \sum_{i=1}^{j+1} p_{n-1,j} \left( \frac{j + 2 - i}{\mu^2} (G_{j+3-i}(d_n - d_{n-1}) - G_{j+3-i}(d_n - d_{n-1} - 2\tau)) \right)
+ \frac{d_n - d_{n-1} + 2\tau}{\mu} (G_{j+2-i}(d_n - d_{n-1}) - G_{j+2-i}(d_n - d_{n-1} - 2\tau))
+ \frac{1 - \alpha_{n-1}}{4\tau^2} \sum_{j=1}^{n-2} \sum_{i=1}^{j} p_{n-1,j} \left( \frac{j + 1 - i}{\mu^2} (G_{j+2-i}(d_n - d_{n-1} + 2\tau) - G_{j+2-i}(d_n - d_{n-1})) \right)
+ \frac{d_n - d_{n-1} - 2\tau}{\mu} (G_{j+1-i}(d_n - d_{n-1} + 2\tau) - G_{j+1-i}(d_n - d_{n-1})),
\] (41)
for $3 \leq n \leq M$. Similarly using Equation (6), leads to
\[ p_{n,i} = \alpha_{n-1} \sum_{j=1}^{n-2} p_{n-1,j} \left( \frac{j + 2 - i}{\mu^2} (G_{j+3-i}(d_n - d_{n-1}) - G_{j+3-i}(d_n - d_{n-1} - 2\tau)) + \frac{d_n - d_{n-1} + 2\tau}{\mu} (G_{j+2-i}(d_n - d_{n-1} - 2\tau)) \right) \]
\[ + (1 - \alpha_{n-1}) \sum_{j=i}^{n-2} p_{n-1,j} \left( \frac{j + 2 - i}{\mu^2} (G_{j+2-i}(d_n - d_{n-1} + 2\tau) - G_{j+2-i}(d_n - d_{n-1})) + \frac{d_n - d_{n-1} - 2\tau}{\mu} (G_{j+1-i}(d_n - d_{n-1} + 2\tau) - G_{j+1-i}(d_n - d_{n-1})), \right) \]
for $3 \leq n \leq M$ and $1 \leq i \leq n - 1$.

One may now use Equations (9)-(22) to characterize the customer waiting time in the queue. Note that Equation (11) leads to $E(W_1) = \frac{\tau}{4}$, and ignoring waiting due to early arrivals, Equation (22) leads to $E(\hat{W}) = E(W) - \frac{\alpha_1 \tau}{4}$.

### 3.2 The case of Triangle Distribution

Here we assume that $D_n$ follows a triangle distribution on $[d_n - \tau, d_n + \tau]$, and its expected value is $d_n$. We have
\[ f_n(x) = \begin{cases} \frac{x - d_n + \tau}{\tau^2} & \text{for } d_n - \tau \leq x < d_n, \\ \frac{d_n + \tau - x}{\tau^2} & \text{for } d_n \leq x \leq d_n + \tau, \end{cases} \]
for $1 \leq n \leq M$. Consider now the random variable $X_{n-1} = D_n - D_{n-1}$, for $2 \leq n \leq M$, and let us compute its pdf $h_{n-1}(x)$. To do so, we determine in a first step the pdf, $\hat{h}_{n-1}(x)$, of $\hat{X}_{n-1} = \hat{D}_n + \hat{D}_{n-1}$, where $\hat{D}_{n-1} = D_{n-1} + d_{n-1} + \tau$ and $\hat{D}_n = D_n - d_n + \tau$. Having in hand $\hat{h}_{n-1}(x)$ and observing that $X_{n-1} = \hat{X}_{n-1} + d_n - d_{n-1} - 2\tau$, we obtain $h_{n-1}(x)$ as follows
\[ h_{n-1}(x) = \hat{h}_{n-1}(x - d_n + d_{n-1} + 2\tau). \]

Let us denote by $\hat{f}_{n-1}(x)$ and $f_n(x)$ the pdfs of the random variables $\hat{D}_{n-1}$ and $D_n$, respectively. Note that $\hat{D}_{n-1}$ and $\hat{D}_n$ are independent and identically distributed. Furthermore
\[ \hat{f}_{n-1}(x) = \hat{f}_n(x) = \begin{cases} \hat{f}_{1,n-1}(x) = \hat{f}_{1,n}(x) = \frac{x}{\tau} & \text{for } 0 \leq x < \tau \\ \hat{f}_{2,n-1}(x) = \hat{f}_{2,n}(x) = \frac{2\tau - x}{\tau^2} & \text{for } \tau \leq x \leq 2\tau. \end{cases} \]
Let us now compute the following integral
\[
\hat{h}_{n-1}(x) = \int_{-\infty}^{+\infty} \hat{f}_n(u)\hat{f}_{n-1}(x-u)\ du.
\] (46)

The random variable \(X_{n-1}\) takes values in the interval \([0, 2\tau]\). We distinguish the following four sub-intervals for the combination of the possible outcomes of the summation of the random variables \(\hat{D}_{n-1}\) and \(\hat{D}_n\): \([0, \tau], [\tau, 2\tau], [2\tau, 3\tau], [3\tau, 4\tau]\). Considering all possible combinations, the convolution may be computed as follows

\[
\hat{h}_{n-1}(x) = \int_{0}^{x} \hat{f}_{1,n}(u)\hat{f}_{1,n-1}(x-u)\ du, \text{ for } x \in [0, \tau],
\] (47)

\[
\hat{h}_{n-1}(x) = \int_{x-\tau}^{\tau} \hat{f}_{1,n}(u)\hat{f}_{1,n-1}(x-u)\ du + \int_{0}^{x-\tau} \hat{f}_{1,n}(u)\hat{f}_{2,n-1}(x-u)\ du + \int_{x-2\tau}^{2\tau} \hat{f}_{2,n}(u)\hat{f}_{1,n-1}(x-u)\ du, \text{ for } x \in [\tau, 2\tau],
\] (48)

\[
\hat{h}_{n-1}(x) = \int_{x-2\tau}^{\tau} \hat{f}_{1,n}(u)\hat{f}_{2,n-1}(x-u)\ du + \int_{x-\tau}^{2\tau} \hat{f}_{2,n}(u)\hat{f}_{1,n-1}(x-u)\ du + \int_{\tau}^{x-\tau} \hat{f}_{2,n}(u)\hat{f}_{2,n-1}(x-u)\ du, \text{ for } x \in [2\tau, 3\tau],
\] (49)

and

\[
\hat{h}_{n-1}(x) = \int_{x-2\tau}^{2\tau} \hat{f}_{2,n}(u)\hat{f}_{2,n-1}(x-u)\ du, \text{ for } x \in [3\tau, 4\tau].
\] (50)

Computing the previous integrals leads to

\[
\hat{h}_{n-1}(x) = \begin{cases} 
\frac{x^3}{6\tau^4}, & \text{for } x \in [0, \tau] \\
\frac{3}{4\tau} - \frac{2x}{\tau^2} + \frac{2x^2}{\tau^3} - \frac{x^3}{3\tau^4}, & \text{for } x \in [\tau, 2\tau] \\
-\frac{22}{3\tau} + \frac{10x}{\tau^2} - \frac{4x^2}{\tau^3} + \frac{x^3}{2\tau^4}, & \text{for } x \in [2\tau, 3\tau] \\
\frac{32}{3\tau} - \frac{8x}{\tau^2} + \frac{2x^2}{\tau^3} - \frac{x^3}{2\tau^4}, & \text{for } x \in [3\tau, 4\tau],
\end{cases}
\] (51)

and applying now Equation (44), we obtain \(h_{n-1}(x)\).

Let us now compute \(h_{1|D_1 > d_1}(x)\), the pdf of the random variable \(D_2 - D_1|D_1 > d_1\). One may apply the same method as above by noting that

\[
h_{1|D_1 > d_1}(x) = \hat{h}_{1|D_1 > d_1}(x - d_2 + d_1 + 2\tau),
\] (52)

where \(\hat{h}_{1|D_1 > d_1}(x)\) is the convolution of \(\hat{f}_2(x)\) defined on \([0, 2\tau]\) as in Equation (45) and \(\hat{f}_{1|D_1 > d_1}(x)\)
defined on $[0, \tau]$ by

$$\frac{1}{\int_{d_1}^{d_1+\tau} f_1(u) \, du} = \frac{x}{\tau^2}. $$

Here we only distinguish three sub-intervals $[0, \tau)$, $[\tau, 2\tau)$, $[2\tau, 3\tau]$. In the fourth interval $[3\tau, 4\tau)$, $\hat{h}_{1|D_1>d_1}(x) = 0$. As in Equations (47)-(50), we obtain

$$\hat{h}_{1|D_1>d_1}(x) = \begin{cases} \frac{x^3}{3\tau^3}, & \text{for } x \in [0, \tau] \\ \frac{-x}{\tau^2} + \frac{2x^3}{3\tau^3}, & \text{for } x \in [\tau, 2\tau] \\ \frac{3x}{\tau^2} - \frac{2x^3}{3\tau^3}, & \text{for } x \in [2\tau, 3\tau], \end{cases} \quad (53)$$

and applying now Equation (52) therefore leads to $h_{1|D_1>d_1}(x)$.

Observing that $\Pr\{D_1 < d_1\} = \Pr\{D_1 > d_1\} = \frac{1}{2}$ and that

$$p_{2,1|D_1<d_1} = \int_{d_1}^{d_1+\tau} e^{-\mu x} f_2(x + d_1) \, dx = \frac{e^{-\mu(d_2-d_1+\tau)}}{\mu^2\tau^2} \left( e^{2\mu\tau} - 2\mu\tau - 1 \right), \quad (54)$$

we can now apply Equation (1) to obtain $p_{2,1}$.

As in the uniform distribution analysis, one may also derive the expressions of $p_{n,i}$ for $3 \leq n \leq M$ and $0 \leq i \leq n - 1$. The main element that leads to the explicit solutions is to again see that the Integral of $\frac{\mu^j x^j - 1}{(j - 1)!} e^{-\mu x}$ is $1 - \sum_{i=0}^{j-1} \left( \frac{\mu x}{i!} \right)^i e^{-\mu x}$. Note that Equation (11) leads to $E(W_1) = \frac{\tau}{6}$, and ignoring waiting due to early arrivals, Equation (22) leads to $E(\hat{W}) = E(W) - \frac{\alpha_1 \tau}{6}$.

### 4 The Multiserver Case

In this section, we extend the analysis to the case of service systems with multiple servers. In particular, there are $s$ identical servers. The service times of each server are independent and exponentially distributed with mean $\frac{1}{\mu}$. If there is an available server, an arriving customer immediately starts service. Otherwise, the customer joins the queue and waits for its turn for service. Customers are processed on a first-come, first-served basis. We retain all other assumptions stated for the case of a single server.

We continue to use the same notations as in the single server case. It is easy to see that we again have $p_{1,0} = 1$, $p_{1,i} = 0$ for $i \neq 0$. For $n = 2$, we have $p_{2,0} = 1 - p_{2,1}$. In Equation (55), we
give the expression of $p_{2,1}$ which is the same as the one for the single server case.

$$
p_{2,1} = \alpha_1 \left( \int_0^{d_1} f_1(u) \, du \right) \left( \int_{d_2 - d_1 - \tau_1^u}^{d_1 + \tau_1^u} e^{-\mu(x)} f_2(x + d_1) \, dx \right) + \alpha_1 \left( \int_{d_1}^{d_1 + \tau_1^u} f_1(u) \, du \right) \left( \int_{d_2 - d_1 - \tau_1^u}^{d_2 - d_1 + \tau_1^u} e^{-\mu x} h_1|D_1 > d_1 \right) \, dx \right).$$  \hspace{1cm} (55)

For $3 \leq n \leq M$, we first compute the probability $p_{n,i}$ that customer $n$ finds (would have found) $i$ customers in the system upon arrival, for $1 \leq i \leq n - 1$. We again condition on the number of customers found (would have been found) by customer $n - 1$. This leads, for $3 \leq n \leq M$, to

$$p_{n,i} = \sum_{j=i+1}^{n-2} p_{n-1,j} \text{Pr}\{R_n = i \mid R_{n-1} = j\}. \hspace{1cm} (56)$$

For $1 \leq i \leq n - 1$ and $j \geq i - 1$, let us characterize the probabilities $\text{Pr}\{R_n = i \mid R_{n-1} = j\}$ in the two cases: Customer $n - 1$ shows up and customer $n - 1$ does not. In order for customer $n$ to find $i$ customers given that customer $n - 1$ found (would have found) $j$ customers, there must be exactly $j - i + 1$ service completions ($j - i$ service completions) between the arrival times of customers $n - 1$ and $n$, i.e., $X_{n-1}$. We then distinguish the following three cases.

**Case 1**, $1 \leq i \leq j + 1 \leq s$: Assume customer $n - 1$ shows up. In this case, customer $n - 1$ is in service and customer $n$ immediately enters service. Once customer $n - 1$ arrives but before customer $n$ does, we need to compute the probability that exactly $j + 1 - i$ among $j + 1$ finish their service. Since service times are independent and exponentially distributed, the remaining service time of each customer is exponentially distributed with rate $\mu$. Using the binomial distribution, noting that $\binom{j + 1}{i} = \binom{j + 1}{j + 1 - i}$, and using the same reasoning for the case when customer $n - 1$ does not show up, we obtain

$$p_{n,i} = \alpha_{n-1} \sum_{j=i+1}^{n-2} p_{n-1,j} \int_{d_n - d_{n-1} - \tau_n + \tau_{n-1}}^{d_n - d_{n-1} + \tau_n + \tau_{n-1}} \frac{j + 1}{i} \left( 1 - e^{-\mu x} \right)^{j + 1 - i} e^{-\mu x} h_{n-1}(x) \, dx \hspace{1cm} (57)$$

$$+ (1 - \alpha_{n-1}) \sum_{j=i}^{n-2} p_{n-1,j} \int_{d_n - d_{n-1} - \tau_n + \tau_{n-1}}^{d_n - d_{n-1} + \tau_n + \tau_{n-1}} \frac{j}{i} \left( 1 - e^{-\mu x} \right)^{j - i} e^{-\mu x} h_{n-1}(x) \, dx.$$

**Case 2**, $s \leq i \leq j + 1$: This is the simplest of the the three cases. Customer $n$ is queued. If she shows up, customer $n - 1$ is also queued. When all servers are busy, the departure process
is Poisson with rate $s\mu$. Thus

$$p_{n,i} = \alpha_{n-1} \sum_{j=-1}^{n-2} \sum_{j=1}^{n-2} p_{n-1,j} \int_{d_n-d_{n-1}+\tau_{n-1}^k + \tau_{n-1}^l}^{d_n-d_{n-1}+\tau_{n-1}^k + \tau_{n-1}^l} \frac{(s\mu x)^{j+i}}{(j+i)!} e^{-s\mu x} h_{n-1}(x) dx$$

$$+ (1 - \alpha_{n-1}) \sum_{j=1}^{n-2} p_{n-1,j} \int_{d_n-d_{n-1}+\tau_{n-1}^k + \tau_{n-1}^l}^{d_n-d_{n-1}+\tau_{n-1}^k + \tau_{n-1}^l} \frac{(s\mu x)^{j-i}}{(j-i)!} e^{-s\mu x} h_{n-1}(x) dx.$$  \hspace{1cm} (58)

**Case 3, 1 \leq i < s \leq j + 1:** Assume customer $n - 1$ shows up. In this case, the system starts busy with $j+1-s$ queued customers just after the epoch of the arrival of customer $n-1$. Before customer $n$ arrives, we need to compute the probability that the $j+1-s$ queued customers leave the queue (enter service) and $s-i$ customers finish their service afterward, i.e., $j + 1 - i$ service completions in total. When all servers are busy, the number of service completions follows a Poisson process with parameter $s\mu$. One may see that computing $Pr\{R_n = i \mid R_{n-1} = j\}$ reduces to compute the probability that after $j + 1 - s$ service completions but before the arrival of customer $n$, i.e., before $X_{n-1}$, exactly $s-i$ service completions occur. Using the binomial distribution, noting that \( \binom{s}{s-i} = \binom{s}{j} \), and using the same reasoning as when customer $n-1$ does not show up, we obtain

$$p_{n,i} = \alpha_{n-1} \sum_{j=-1}^{n-2} \sum_{j=1}^{n-2} p_{n-1,j} \int_{d_n-d_{n-1}+\tau_{n-1}^k + \tau_{n-1}^l}^{d_n-d_{n-1}+\tau_{n-1}^k + \tau_{n-1}^l} \int_0^x \binom{s}{i} (1 - e^{-\mu(x-t)})^{s-i} e^{-i\mu(x-t)} \frac{(s\mu)^{j+s} (s\mu)^{s-i}}{(j+s)!} e^{-s\mu t} h_{n-1}(x) dt dx$$

$$+ (1 - \alpha_{n-1}) \sum_{j=1}^{n-2} p_{n-1,j} \int_{d_n-d_{n-1}+\tau_{n-1}^k + \tau_{n-1}^l}^{d_n-d_{n-1}+\tau_{n-1}^k + \tau_{n-1}^l} \int_0^x \binom{s}{i} (1 - e^{-\mu(x-t)})^{s-i} e^{-i\mu(x-t)} \frac{(s\mu)^{j+s} (s\mu)^{s-i}}{(j+s)!} e^{-s\mu t} h_{n-1}(x) dt dx.$$ \hspace{1cm} (59)

What remains for us now to compute is the probability $p_{n,0}$ that customer $n$ finds the system empty upon arrival. This is simply given by $p_{n,0} = 1 - \sum_{i=1}^{n-1} p_{n,i}$.

Having in hand the probabilities $p_{n,i}$ for $0 \leq i \leq n-1$, we next describe how waiting times can be characterized. Recall that $W_n$ is the random variable that denotes waiting time in the queue of customer $n$ if she shows up and $E(W_n^k)$ is the corresponding $k^{th}$ moment for $k \geq 1$.

For $n = 1$, we follow the same reasoning as the one in Section 2.1. This leads to

$$E(W_n^k) = \int_0^{d_1} (d_1 - x)^k f_1(x) dx.$$ \hspace{1cm} (60)

For $2 \leq n \leq M$, we have

$$E(W_n^k) = \sum_{i=s}^{n-1} p_{n,i} E(W_n^{k,i}),$$ \hspace{1cm} (61)

where $W_n,i$ is the random variable denoting the waiting time in queue for customer $n$ if she shows up, given that customer $n$ finds $i$ customers in the system and $E(W_n^{k,i})$ is the $k^{th}$ moment.
for \( k \geq 1 \). We assume that \( s \geq 2 \) (if \( s = 1 \), the analysis is given in Section 2.1). For \( i \leq s - 1 \), \( W_{n,i} = 0 \). For \( i \geq s \), \( W_{n,i} \) has the Erlang distribution with \( i - s + 1 \) stages and parameter \( s \mu \). Consequently, we have

\[
E(W_n) = \sum_{i=s}^{n-1} \frac{i - s + 1}{s \mu}, \quad (62)
\]

and

\[
E(W_n^2) = \sum_{i=s}^{n-1} \frac{(i - s + 1)(i - s + 2)}{s^2 \mu^2}, \quad (63)
\]

for \( 2 \leq n \leq M \). (Recall that by convention an empty sum is equal to 0). Let \( W \), a random variable, denote the waiting time in the queue of an arbitrary customer among all \( M \) customers and let \( E(W^k) \) be the \( k \)th moment for \( k \geq 1 \). Then

\[
E(W^k) = \frac{1}{M} \sum_{n=s+1}^{M} \alpha_n E(W_n^k). \quad (64)
\]

As in Section 2.1, we can use the probabilities \( p_{n,i} \) to characterize the cdf of \( W \). For \( n = 1 \), we may again write

\[
\Pr(W_1 < t) = 1 - \int_{0}^{\max(d_1 - t, 0)} f_1(x) \, dx. \quad (65)
\]

For \( 2 \leq n \leq M \) and \( t \geq 0 \), we have

\[
\Pr(W_{n,i} < t) = 1 - \sum_{j=0}^{i-s} \frac{(s \mu t)^j}{j!} e^{-s \mu t}. \quad (66)
\]

Consequently,

\[
\Pr(W_n < t) = \sum_{i=0}^{s-1} p_{n,i} + \sum_{i=s}^{n-1} p_{n,i} \Pr(W_{n,i} < t), \quad (67)
\]

or equivalently

\[
\Pr(W_n < t) = 1 - \sum_{i=s}^{n-1} \sum_{j=0}^{i-s} \frac{(s \mu t)^j}{j!} e^{-s \mu t}. \quad (68)
\]

Observing that

\[
\Pr(W < t) = \frac{1}{M} \left( \alpha_1 \Pr(W_1 < t) + \sum_{n=s+1}^{M} \alpha_n \Pr(W_n < t) \right), \quad (69)
\]
we finally obtain
\[
\Pr(W < t) = \frac{\alpha_1}{M} - \frac{\alpha_1}{M} \int_0^{\max(d_1-t,0)} f_1(x) \, dx \\
+ \frac{1}{M} \sum_{n=s+1}^{M} \sum_{i=0}^{n-1} \alpha_n p_{n,i} - \frac{1}{M} \sum_{n=s+1}^{M} \sum_{i=s}^{n-1} \sum_{j=0}^{i-s} \alpha_n p_{n,i} \left( \frac{\tau}{j!} \right) e^{-\frac{\tau}{j!}},
\]
which completes the analysis of waiting times.

**Remark:** The analysis provided in this section can be used to determine the number of servers that ensure that a specified percentage of customers wait less than a specified amount of time (the so-called staffing problem that arises in some applications).

5 Numerical Results

In this section we briefly present numerical results to illustrate the usefulness of the performance evaluation models described in the previous sections. In particular, we examine the impact of customer non-punctuality on expected waiting time. This also allows us to assess the error in evaluating waiting times that would be introduced if customers were assumed to be always punctual. For the sake of consistency, all the results we present are for systems where the vector of appointments are constructed as follows. Appointment time for customer \( n \) is \( d_n = d_{n-1} + \frac{1}{\mu_{n-1}}, \) for \( 2 \leq n \leq M \). For customer 1, appointment time is \( d_1 = \tau \). This appointment scheme is perhaps consistent with some that are observed in practice (particularly in healthcare) that seek to limit the idleness of the servers (introducing additional slack time between appointments would reduce customer waiting time but increase server idleness)\(^3\). It is of course easy to consider other appointment schemes. Moreover, for the sake of brevity, we present results for only single server systems and for systems where arrival times are uniformly distributed. Results obtained (but not shown here) for multiple servers and for other distributions of arrival times yield similar insights.

To measure the impact of non-punctuality, we evaluate the percentage difference in expected waiting between a system where customers are non-punctual, \( E(W_{\text{non-punctual}}) \), and one where customers are always punctual (that is, customers show up on exactly their appointment times), \( E(W_{\text{punctual}}) \). We denote this percentage difference by \( \gamma \), where
\[
\gamma = \frac{E(W_{\text{non-punctual}}) - E(W_{\text{punctual}})}{E(W_{\text{punctual}})} \times 100.
\]
In both the punctual and non-punctual systems, we allow for no-shows, where the proba-

\(^3\)It is not our objective in this paper to develop methodology for optimal appointment scheduling and we leave this as an area for future research. However, our performance evaluation models would clearly be needed to carry out such an optimization.
The possibility of no-shows is the same in both systems (in the punctual system, the customers that do show up, they do it on time). The results, representative of results from a much larger set, are shown in Figures 1-4.

**Observation 1:** Non-punctuality can significantly increase customer waiting time. Thus, ignoring non-punctuality can lead to significant errors in estimating waiting time.

This observation is supported by Figures 1(a) and 1(b). As we can see not accounting for non-punctuality can significantly under-estimate the actual expected waiting time. This could have various negative impacts if the incorrect estimates of waiting times are used to make decisions about how many customers to schedule, how to determine appointment times, or how to optimize capacity. Figure 1(a) shows that the effect of non-punctuality is more significant when the number of customers is small. Figure 1(b) shows that the effect of non-punctuality is more significant when the no-show probability is high. These two effects are perhaps surprising, as one might expect the impact of non-punctuality to be greater when there are more customers in the system (either because of a larger $M$ or a higher $\alpha$). These effects appear due to the fact that, when either $M$ or $\alpha$ are large, expected waiting time is relatively large even when the customers are punctual. Introducing non-punctuality does increase waiting time, but the effect is relatively small.

**Observation 2:** Assigning appointment times based on customer non-punctuality can significantly reduce waiting times. Thus, ignoring customer punctuality can lead to poor appointment scheduling.

This observation is supported by Figures 2(a), 2(b) and 2(c). The figures correspond to different appointment assignment scenarios. Table 1 describes the appointment scenarios for each Figure. For Figure 2(a), scenario 1 corresponds to all customers being punctual (customers $n$ shows up
at time \( d_n \)); scenario 2 corresponds to all customers being punctual, except for customer 1 who has been assigned appointment time \( d_1 \); scenario 3 corresponds to all customers being punctual, except for customers 1 and 2 who have been assigned appointment times \( d_1 \) and \( d_2 \) and so on and so forth. For Figure 2(b), scenario 1 also corresponds to all customers being punctual. However, scenario 2 now corresponds to all customers being punctual, except for customer 10 who has been assigned appointment time \( d_{10} \); scenario 3 corresponds to all customers being punctual, except for customers 9 and 10, and so on and so forth. For Figure 2(c), scenario 1 is similarly defined but scenario 2 corresponds to customer 5 being the only non-punctual customer, scenario 3 corresponds to customers 4 and 5 being the only non-punctual customers, and so on and so forth. Hence, Figure 2(a) corresponds to a setting where the non-punctual customers are scheduled first, Figure 2(b) corresponds to a setting where the non-punctual customers are scheduled last, and Figure 2(c) to a setting where the non-punctual customers are scheduled in the middle. As we can see from the figures, how appointments are assigned can have a significant impact on expected waiting time. In particular, scheduling non-punctual customers early can significantly increase waiting times as the impact of their non-punctuality affects all customers with subsequent appointments. The figures also reveal that how appointments are assigned to the non-punctual customers can be more important than limiting the number of non-punctual customers. Consider for example the case where \( \alpha = 1 \). Then, we can see that having up to 8 non-punctual customers increases expected waiting by less than 5% relative to the all punctual case as long as the non-punctual customers are scheduled last. However, having a single non-punctual customer increases expected waiting time by over 20% if that customer is scheduled first.

**Observation 3:** Accounting for non-punctuality is more important for systems with heterogeneous service times than for those with homogenous service times.

This observation is supported by Figure 3. The figure shows the impact of different scenarios of heterogeneity in service time on the percentage difference in expected waiting time between a punctual and a non-punctual system. The four scenarios are described in Table 2. Scenario 1 corresponds to a system with homogenous service times where the mean service time is the same for all customers and equal to 5 time units. Scenario 2 corresponds to the first five customers having a mean service time of 4 and the last 5 having a mean of 6. Similarly, scenarios 3 (4) corresponds to the first five customers having a mean service time of 3 (2) and the last five having a mean service time of 7 (8). Note that customers with shortest mean service times are scheduled first. Such a scheduling rule minimizes overall expected waiting time and is consistent with scheduling practices in many applications. Note also that, to allow
Table 1: Description of the scenarios for Figure 2

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Figure 2(a)</th>
<th>Figure 2(b)</th>
<th>Figure 2(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \tau_n = 0 ) for all ( n )</td>
<td>( \tau_n = 0 ) for all ( n )</td>
<td>( \tau_n = 0 ) for all ( n )</td>
</tr>
<tr>
<td>2</td>
<td>( \tau_1 = 2 )</td>
<td>( \tau_n = 0 ) for ( n \neq 10 )</td>
<td>( \tau_5 = 2 )</td>
</tr>
<tr>
<td></td>
<td>( \tau_n = 0 ) for ( n \neq 1 )</td>
<td>( \tau_{10} = 2 )</td>
<td>( \tau_n = 0 ) for ( n \neq 5 )</td>
</tr>
<tr>
<td>3</td>
<td>( \tau_n = 2 ) for ( n = 1, 2 )</td>
<td>( \tau_n = 0 ) otherwise</td>
<td>( \tau_n = 2 ) for ( n = 5, 6 )</td>
</tr>
<tr>
<td></td>
<td>( \tau_n = 0 ) otherwise</td>
<td>( \tau_n = 2 ) otherwise</td>
<td>( \tau_n = 0 ) otherwise</td>
</tr>
<tr>
<td>4</td>
<td>( \tau_n = 2 ) for ( n = 1, \ldots, 3 )</td>
<td>( \tau_n = 0 ) for ( n = 1, \ldots, 7 )</td>
<td>( \tau_n = 2 ) for ( n = 4, \ldots, 6 )</td>
</tr>
<tr>
<td></td>
<td>( \tau_n = 0 ) otherwise</td>
<td>( \tau_n = 2 ) otherwise</td>
<td>( \tau_n = 0 ) otherwise</td>
</tr>
<tr>
<td>5</td>
<td>( \tau_n = 2 ) for ( n = 1, \ldots, 4 )</td>
<td>( \tau_n = 0 ) for ( n = 1, \ldots, 6 )</td>
<td>( \tau_n = 2 ) for ( n = 4, \ldots, 7 )</td>
</tr>
<tr>
<td></td>
<td>( \tau_n = 0 ) otherwise</td>
<td>( \tau_n = 2 ) otherwise</td>
<td>( \tau_n = 0 ) otherwise</td>
</tr>
<tr>
<td>6</td>
<td>( \tau_n = 2 ) for ( n = 1, \ldots, 5 )</td>
<td>( \tau_n = 0 ) for ( n = 1, \ldots, 5 )</td>
<td>( \tau_n = 2 ) for ( n = 3, \ldots, 7 )</td>
</tr>
<tr>
<td></td>
<td>( \tau_n = 0 ) otherwise</td>
<td>( \tau_n = 2 ) otherwise</td>
<td>( \tau_n = 0 ) otherwise</td>
</tr>
<tr>
<td>7</td>
<td>( \tau_n = 2 ) for ( n = 1, \ldots, 6 )</td>
<td>( \tau_n = 0 ) for ( n = 1, \ldots, 4 )</td>
<td>( \tau_n = 2 ) for ( n = 3, \ldots, 8 )</td>
</tr>
<tr>
<td></td>
<td>( \tau_n = 0 ) otherwise</td>
<td>( \tau_n = 2 ) otherwise</td>
<td>( \tau_n = 0 ) otherwise</td>
</tr>
<tr>
<td>8</td>
<td>( \tau_n = 2 ) for ( n = 1, \ldots, 7 )</td>
<td>( \tau_n = 0 ) for ( n = 1, \ldots, 3 )</td>
<td>( \tau_n = 2 ) for ( n = 2, \ldots, 8 )</td>
</tr>
<tr>
<td></td>
<td>( \tau_n = 0 ) otherwise</td>
<td>( \tau_n = 2 ) otherwise</td>
<td>( \tau_n = 0 ) otherwise</td>
</tr>
<tr>
<td>9</td>
<td>( \tau_n = 2 ) for ( n = 1, \ldots, 8 )</td>
<td>( \tau_n = 0 ) for ( n = 1, 2 )</td>
<td>( \tau_n = 2 ) for ( n = 2, \ldots, 9 )</td>
</tr>
<tr>
<td></td>
<td>( \tau_n = 0 ) otherwise</td>
<td>( \tau_n = 2 ) otherwise</td>
<td>( \tau_n = 0 ) otherwise</td>
</tr>
<tr>
<td>10</td>
<td>( \tau_n = 2 ) for ( n \neq 10 )</td>
<td>( \tau_{10} = 0 )</td>
<td>( \tau_n = 2 ) for ( n \neq 10 )</td>
</tr>
<tr>
<td></td>
<td>( \tau_{10} = 0 )</td>
<td>( \tau_n = 2 ) for ( n \neq 1 )</td>
<td>( \tau_{10} = 0 )</td>
</tr>
<tr>
<td>11</td>
<td>( \tau_n = 2 ) for all ( n )</td>
<td>( \tau_n = 2 ) for all ( n )</td>
<td>( \tau_n = 2 ) for all ( n )</td>
</tr>
</tbody>
</table>

Figure 2: The impact of appointment assignments (\( \alpha_n = \alpha, \mu_n = \mu = 0.2, M = 10 \), and \( \tau_n \) as in Table 1)
for a fair comparison between the scenarios, the total sum of mean service times \( \sum_{n=1}^{M} \frac{1}{\mu_n} \) is kept constant. As we can see, the impact of non-punctuality is more significant for scenarios with higher service time heterogeneity (that is, scenarios with larger differences in mean service times between the first and last five customers). This effect appears in part due to the fact that with increased service time heterogeneity the ratio \( \tau_n/\mu_n \) increases for the customers with the shorter mean service times, leading to relatively higher waiting times. This effect is compounded by the fact that customers with shorter service times are scheduled first and that customers regardless of their service times exhibit the same punctuality. It is of course possible to obtain different results if punctuality depends on service times and appointment times are selected based on both service times and punctuality. Nevertheless, the results shown here point to the importance of accounting for both non-punctuality and customer heterogeneity and the usefulness of performance evaluations models such ours that allow the explicit modeling of both features.

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Service rates (( \mu_n ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \mu_n = 1/5 ) for all ( n )</td>
</tr>
<tr>
<td>2</td>
<td>( \mu_n = 1/4 ) for ( n = 1, \ldots, 5 ); and ( \mu_n = 1/6 ) otherwise</td>
</tr>
<tr>
<td>3</td>
<td>( \mu_n = 1/3 ) for ( n = 1, \ldots, 5 ); and ( \mu_n = 1/7 ) otherwise</td>
</tr>
<tr>
<td>4</td>
<td>( \mu_n = 1/2 ) for ( n = 1, \ldots, 5 ); and ( \mu_n = 1/8 ) otherwise</td>
</tr>
</tbody>
</table>

Figure 3: The impact of Service time heterogeneity (\( \alpha_n = \alpha, M = 10, \tau_n = \tau = 1, \) and \( \mu_n \) as in Table 2)

**Observation 4:** Assigning appointment times based on customer show-up probabilities can significantly reduce waiting times. Moreover, the impact of non-punctuality can be sensitive to appointment assignments.
This observation is supported by Figures 4(a) and 4(b) where we consider a system with two sets of customers, those who are likely to show up ($\alpha_n = 0.9$) and those who are not ($\alpha_n = 0.1$). The figures depict the impact of four different scenarios, where as shown in Table 3, each scenario corresponds to a different assignment of appointments. In scenario 1, we alternate one by one the appointment of customers with low show-up probability with those with high show-up probability. In scenario 2, we alternate two by two the appointment of customers with low show-up probability with those with high show-up probability; while in scenarios 3 and 4, we alternate these appointments 3 by 3 and 6 by 6, respectively. Thus, scenario 1 corresponds to a perfect mixing of customers who are likely and unlikely to show up while scenario 2 corresponds to a complete segregation of these customers.

As we can see from Figure 4(a), the greater the mixing of customers who are likely and not likely to show up, the lower are the waiting times. This is perhaps consistent with intuition since such mixing allows more time between the arrivals of customers who are likely to show up (note that this effect can be quite significant with expected waiting times in scenario 4 more than double than those in scenario 1). However, perhaps not as intuitive is the impact of non-punctuality. As shown in Figure 4(b), this impact is greater for systems with a greater mixing of the two sets of customers. This appears due to the fact that even in the absence of punctuality waiting times can be large in systems where customers are more segregated and the additional waiting time introduced by non-punctuality is relatively small.

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Show up probabilities ($\alpha_n$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\alpha_1 = 0.1; \alpha_2 = 0.9; \alpha_3 = 0.1; \ldots; \alpha_{12} = 0.9$</td>
</tr>
<tr>
<td>2</td>
<td>$\alpha_1 = \alpha_2 = 0.1; \alpha_3 = \alpha_4 = 0.9; \alpha_5 = \alpha_6 = 0.1; \ldots; \alpha_{11} = \alpha_{12} = 0.9$</td>
</tr>
<tr>
<td>3</td>
<td>$\alpha_1 = \alpha_2 = \alpha_3 = 0.1; \alpha_4 = \alpha_5 = \alpha_6 = 0.9; \alpha_7 = \alpha_8 = \alpha_9 = 0.1; \alpha_{10} = \alpha_{11} = \alpha_{12} = 0.9$</td>
</tr>
<tr>
<td>4</td>
<td>$\alpha_1 = \alpha_2 = \ldots = \alpha_6 = 0.1; \alpha_6 = \alpha_7 = \ldots = \alpha_{12} = 0.9$</td>
</tr>
</tbody>
</table>

Figure 4: The impact of heterogeneity in show up probabilities ($M = 12, \mu_n = \mu = 0.2$, and $\alpha_n$ as in Table 3)
The numerical results described in this section highlight the importance of using models such as ours that accounts for non-punctuality and no-shows, particularly when both features are present. They also highlight the importance of accounting for differences in service times, punctuality, and no-shows among customers. This perhaps points to the importance for service systems to collect individualized information about their customers and to use this information to improve the quality of service it provides.

6 Concluding Remarks

In this paper, we described models for the performance evaluation and analysis for queueing systems where a finite number of customers arrive over time and their arrivals are driven by appointment times. We allowed for customers to vary in terms of time between appointments, punctuality, show-up probabilities, and service times. We illustrated the usefulness of the models by describing numerical results that examine the impact of not accounting for non-punctuality and no-shows.

There are several avenues for future research. It would be useful to extend our analysis to the case where customer arrival time distributions may overlap, although this may require developing an alternative approach to the analysis, particularly if the overlap involves multiple customers. It would also be useful to extend the analysis to the case of multiple servers with heterogeneous customer service times. It would also be of interest to embed the performance evaluation models described in this paper within an optimization model to determine optimal customer scheduling times. Because of the highly non-linear features of the resulting problem, it is unlikely that an exact approach would be possible. However, it may be possible to identify effective heuristics.

References


