Service Systems with Finite and Heterogeneous Customer Arrivals

Rowan Wang, Oualid Jouini, Saif Benjaafar

Service Systems with Finite and Heterogeneous Customer Arrivals

Rowan Wang[1] • **Oualid Jouini**[2] • **Saif Benjaafar**[1]

[1] Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, Minnesota 55455

[2] Laboratoire Génie Industriel, Ecole Centrale Paris, Châtenay-Malabry, France 92290

wang1075@umn.edu • oualid.jouini@ecp.fr • saif@umn.edu

Manufacturing & Service Operations Management*. To appear, 2014.*

Abstract

We consider service systems with a finite number of customer arrivals, where customer inter-arrival times and service times are both stochastic and heterogeneous. Applications of such systems are numerous and include systems where arrivals are driven by events or service completions in serial processes, and systems where servers are subject to learning or fatigue. Using an embedded Markov chain approach, we characterize the waiting time distribution for each customer, from which we obtain various performance measures of interest, including the expected waiting time of a specific customer, the expected waiting time of an arbitrary customer, and the expected completion time of all customers. We carry out extensive numerical experiments to examine the effect of heterogeneity in inter-arrival and service times. In particular, we examine cases where inter-arrival and service times increase with each subsequent arrival or service completion, decrease, increase and then decrease, or decrease and then increase. We derive several managerial insights and discuss implications for settings where such features can be induced. We validate the numerical results using a fluid approximation that yields closed form expressions.

Keywords: Queueing systems; finite arrivals; heterogeneous inter-arrival and service times; transient analysis; fluid approximation

1 Introduction

This paper is motivated by systems where a finite number of customer arrivals occur over a period of time followed by few or no arrivals for an extended period thereafter. During the period over which arrivals take place, inter-arrival times between consecutive customers can be different and so can be their service times. Examples of such systems are numerous.

Consider, for example, settings where arrivals are triggered by the start of an event or a service (e.g., the arrival of passengers to check-in for or to board a flight), the total number of arrivals is finite (and determined by the number of tickets sold). Passengers may belong to different classes (e.g., early, on-time, and late) or are assigned to different groups (e.g., priority boarding zones), so that arrivals occur in waves with each wave drawing from the population of the corresponding class or group.

Another example is one where a finite number of jobs go through a sequence of production stages. The arrival process to each stage (other than the first one) corresponds to the departure process from the preceding one. Because production times at a particular stage are stochastic and can vary in distribution from job to job, the inter-arrival times to the subsequent stage are also stochastic and vary from job to job.

A third example is one where arrivals are driven by appointments (e.g., patient appointments at a health clinic). Assuming customers are punctual (or nearly punctual), inter-arrival times coincide with time between appointments. Depending on how appointments are scheduled, the inter-arrival times between customers can vary. For example, spacing appointments equally leads to uniform inter-arrival times, while other rules, such as those that schedule more appointments at the beginning and at the end, and fewer in between, lead to increasing and then decreasing inter-arrival times.

All the above examples share four common characteristics: (1) a finite number of customers; (2) heterogeneous (and possibly stochastic) inter-arrival times; (3) heterogeneous (and possibly stochastic) service times; and (4) inter-arrival and service times that depend on the position of the customers in the arrival process.

Accounting for heterogeneity in arrival and service times is important in settings where inter-arrival and service times exhibit distinctive features that make it difficult to justify the common assumption of identically distributed inter-arrival and service times. Such features include (1) arrivals that decrease in intensity with each subsequent arrival; (2) arrivals that increase in intensity with each subsequent arrival; and (3) arrivals that exhibit the combinations of both the increasing and decreasing features. They also include (1) service times that increase with

each subsequent service completion, typical of settings where servers are subject to fatigue; (2) service times that decrease with each subsequent service completion, typical of systems where learning takes place; and (3) service times that exhibit the combinations of both the increasing and decreasing features (e.g., initial learning by the servers that is followed by eventual fatigue).

The modeling and analysis of systems with finite arrivals and varying inter-arrival and service times raise several important questions: (1) What is the impact of different inter-arrival and service time features on system performance (for example, does system performance deteriorate with increased heterogeneity in inter-arrival or service times)? (2) For a fixed number of arrivals, are there features which lead to better performance than others (for example, given a target time window for arrivals, is it best to have more arrivals early on, in the middle, or at the end of the arrival time window)? (3) How are the answers to the above questions affected by other problem parameters such as the overall arrival intensity and the total number of arrivals (for example, do higher levels of the parameters favor certain arrival features over others)? (4) Does the heterogeneity in service times affect performance the same way that the heterogeneity in inter-arrival times does, or are there fundamental differences between these two?

In this paper, we address these and other related questions. In particular, we consider a system with a finite number of arrivals, where the inter-arrival time between the m^{th} and $(m+1)^{th}$ customer is described by a random variable that has a general distribution which can be different from the distributions that describe the inter-arrival times between other consecutive customers. Customer service times are described by exponential distributions; however, the mean service times (or service rates) of different customers can be different. We consider systems with both single and multiple servers. Using an embedded Markov chain approach, in each case, we are able to characterize analytically the probability distribution of the number of customers seen by each arrival. This allows us to characterize the waiting time distribution for each customer, from which we obtain various performance measures of interest, including the expected waiting time of a specific customer, the expected waiting time of an arbitrary customer, and the expected completion time of all customers (makespan). These characterizations further simplify for several special cases of interest, including systems with exponential and deterministic inter-arrival times.

We carry out extensive numerical experiments to examine the effects of heterogeneity in inter-arrival and service times. In particular, we examine cases where, with each subsequent arrival or service completion, inter-arrival and service times (1) increase, (2) decrease, (3) increase and then decrease, or (4) decrease and then increase. We derive several managerial insights and

discuss implications for settings where such features can be induced. We validate the numerical results using a fluid approximation that yields closed form expressions. Some of our key findings are highlighted below:

- Arrival processes with different features can lead to significantly different expected waiting times. There is a considerable difference in performance between systems with homogeneous inter-arrival times and those with heterogeneous inter-arrival times. Therefore, ignoring the heterogeneity in arrival process can lead to significant errors in performance evaluation.

- Arrival processes with homogeneous inter-arrival times may not lead to the lowest waiting time. In fact, for a wide range of parameter values, systems with homogenous inter-arrival times perform poorly.

- Although there is no strict ordering in terms of performance among the arrival processes considered, for systems with homogeneous service times, arrival processes where inter-arrival times decrease, or increase and then decrease, lead to lower waiting time than those where inter-arrival times increase, or decrease and then increase, suggesting that it is generally better to postpone the busy (or peak) period.

- When inter-arrival times are homogeneous, systems in which customers with short service times arrive early (at the beginning of the arrival period) have lower waiting time than those in which such customers arrive later. This is perhaps consistent with results about the optimality of processing customers with shorter processing times first. However, this is not true when inter-arrival times are heterogeneous.

- Inter-arrival and service time features that lead to lower waiting time may not lead to lower makespan.

These insights show that there might be opportunities for system managers to improve system performance by inducing certain arrival features and by differentiating between customers or jobs with different service requirements. We illustrate how arrivals could be affected using two examples. The first one involves the sequencing of a finite number of jobs through two production stages in series. The second one involves the grouping of passengers into multiple boarding zones. For systems where arrivals cannot be controlled, we examine how arrival processes with different features affect the capacity needed to guarantee a specified level of performance (e.g., a maximum expected waiting time or makespan).

2 Related Literature

Although systems with a finite number of arrivals and distinct features in inter-arrival or service times are prevalent and perhaps even pervasive in practice, they have received relatively little attention in the service operations management literature (and more generally in the broader queueing literature). This appears to be, in part, due to the difficulty of analyzing these systems using standard queueing methodology which relies on steady state analysis (and therefore assumes an infinite number of arrivals) or requires homogenous inter-arrival and service times (see, Kleinrock 1975; Hall 1991).

There is an extensive literature that deals with finite population systems (see, for example, Takagi 1993; Sztrik 2005; Haque and Armstrong 2007). However, in that case, the finite population of customers cycles indefinitely through two phases of not needing service and needing service (e.g., machines that require repairs). The analysis typically assumes homogeneity in both arrival and service processes. Hence, this literature does not capture the essential features of the problem we consider here.

There is also an extensive literature on systems with time-dependent/state-dependent arrival or service processes (see, for example, Courtois and Georges 1971; Ross 1978; Green et al. 1991) where the arrival or service rates may depend on either time, the number of customers in the system, or the evolution of certain exogenous stochastic processes. This literature does not capture the settings we describe here where inter-arrival and service times depend on the order in which a particular customer arrives to the system and where the number of customers is finite.

The literature which is most related to ours is the one on transient analysis of queueing systems (see, for example, Kelton and Law 1985; Parthasarathy and Moosa 1989; Griffiths et al. 2006). However, this literature typically assumes homogenous inter-arrival and service time distributions and the existing results are for systems with Markovian arrivals. Other related papers include Hu and Benjaafar (2009), which treats a special case of our problem where all customers arrive at once (they refer to this as the rush hour regime). Parlar and Moosa (2008) also consider a special case of our problem where the arrivals are Markovian and determined by a pure death process so that the arrival rates are linearly decreasing. In our case, we allow for non-Markovian arrivals and arbitrary arrival rates. Hassin and Mendel (2008) consider a system with a single server and finite arrivals, but customer arrivals are determined by appointment times. Customers are assumed to be punctual and therefore there is no uncertainty regarding arrival times. The service times are exponentially and identically distributed.

There is an extensive body of literature in the area of scheduling which shares features of the problem we consider in this paper; namely a finite number of customers (or jobs) that are processed through one or more machines. The jobs are available for processing at specified release times. Jobs may vary in their processing times, delay costs, and due dates. In some cases the release and service times are stochastic. The focus of much of this literature is, on developing efficient algorithms for generating optimal job sequences, or on identifying structural properties of optimal sequences; see (Pinedo 2012; Emmons and Vairaktarakis 2013) for a discussion of important results and a review of relevant literature. Some of the literature treats the *online* version of the problem where jobs arrive over time and a decision on which job to process next is made with each job arrival and job completion (in the case where preemption is allowed); see for example (Chou et al. 2006; Chen and Shen 2007; Ouelhadj and Petrovic 2009). This literature is generally not concerned with developing performance evaluation models as we are in this paper.

Finally, there is a growing body of literature which deals with the scheduling of appointments, particularly in healthcare settings. A review of this literature can be found in (Preater 2001; Cayirli and Veral 2003). We also refer the reader to (Mondschein and Weintraub 2003; Gupta and Denton 2008; Jouini et al. 2014). Most of this literature assumes that customers are punctual and the objective is to identify the optimal spacing between appointments where the optimality is determined by a weighted measure of patient's delay, physician's idleness, and tardiness. Note that when customers are punctual and service times are exponential, the performance of a specified schedule can be evaluated using the approach described in this paper.

Some of the literature considers no-shows which introduces a particular form of stochasticity in patients' inter-arrival times. For example, Kaandorp and Koole (2007) develop a local search algorithm to identify optimal schedules in the presence of no shows and show that a so-called *dome-shaped* form where more appointments are scheduled at the beginning and at the end of the schedule, is particularly effective (see related discussion in Section 7). Zeng et al. (2010) extend Kaandorp and Koole (2007) to include heterogeneous no-show rates. Koeleman and Koole (2012) also generalize the model by considering both scheduled and emergency arrivals. Some recent papers consider patient scheduling based on an open access model with same day appointments; see Robinson (2010) and the references therein.

The rest of the paper is organized as follows. In Section 3, we describe the model and provide analysis for the single server system. In Section 4, we extend the analysis to the multi-server case. In Section 5, we present numerical results and discuss insights. In Sections 6, we describe

the fluid approximation. In Section 7, we discuss example applications. In Section 8, we provide a summary and concluding comments.

3 Problem Description and Analysis

We consider a queueing system with a single server and a finite number of customers arriving randomly over time. The total number of customers is M. We index customers by the order of their arrivals, so that customer m for $m = 1, ..., M$, is the m^{th} customer to arrive. The inter-arrival time between customer $m-1$ and customer m has a general distribution with a finite mean $\frac{1}{\lambda_m}$ for $m = 2, ..., M$. No other specific assumptions are made concerning inter-arrival times except that they are independent. Customer service times are independent and exponentially distributed with a strictly positive and finite mean $\frac{1}{\mu_m}$ for customer m. We make the exponential assumption regarding the distribution of service times for mathematical tractability, as it allows us to formulate the problem as an embedded Markov chain. This assumption is also useful in approximating the behavior of systems where service time variability is high. Doing away with this assumption without losing tractability is difficult, given the generality of the model otherwise (i.e., the heterogeneity in inter-arrival and service times). Upon arrival, a customer goes immediately into service if the server is available. If not, the customer joins the queue and waits. Customers waiting in the queue are served on a first-come, first-served (FCFS) basis.

Note that the inter-arrival and service times are indexed by the position of the customer in the arrival sequence ($m = 1, ..., M$) and not by time, as in a time-dependent process. This is because we are interested in settings, such as the ones we describe in Section 1, where the characteristics of the arrival and service processes are affected by the number of customers that have already arrived and not by the amount of time that has already elapsed. This is apparent for example when customers, who are drawn from a finite population, arrive independently from each other, when arrivals correspond to service completions from a preceding process, or when service times are affected by the number of customers previously processed, as in situations in which learning and fatigue can take place.

We are interested in characterizing customer waiting time. Our approach consists of first computing the probabilities of the system states 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 possibilities. We denote A_m as the random variable that describes the arrival time of customer m, and R_m as the random variable that describes the number of customers found in the system by customer m, upon her arrival at A_m. This means

that the total number of customers in the system immediately after A_m is R_m+1. We let $p_{m,i} = \Pr\{R_m = i\}$ refer to the probability that the m^{th} customer finds, upon arrival, i customers already in the system (in queue or in service) for $i = 0, ..., m-1$ and $m = 1, ..., M$.

In what follows, we first characterize the probabilities $p_{m,i}$. Let T_m be the random variable describing the inter-arrival time between customers $m-1$ and m, and let $f_m(.)$ be its probability density function. We have $T_m = A_m - A_{m-1}$ for $m = 2, ..., M$. Without loss of generality, we assume the first customer arrives at time 0 ($T_1 = 0$). For $m = 1$, we have $p_{1,0} = 1$ and $p_{1,i} = 0$ for $i \neq 0$, because the first customer always finds the system empty. For $2 \leq m \leq M$, we separate the two cases, $1 \leq i \leq m-1$ and $i = 0$. Let us first consider the case $1 \leq i \leq m-1$. Conditioning on the number of customers found, upon arrival, by customer $m-1$, we obtain

$$p_{m,i} = \sum_{j=i-1}^{m-2} p_{m-1,j} \Pr\{R_m = i \mid R_{m-1} = j\} \tag{1}$$

for $2 \leq m \leq M$. Note that we must have $i-1 \leq j \leq m-2$. Let us now characterize the probability $\Pr\{R_m = i \mid R_{m-1} = j\}$ for $1 \leq i \leq m-1$ and $i-1 \leq j \leq m-2$. We again separate the analysis into two cases, $i \leq j \leq m-2$ and $j = i-1$. Firstly, when $i \leq j \leq m-2$, in order for customer m to find i customers given that customer $m-1$ finds j, there must be exactly $j-i+1$ service completions during the time period $(A_{m-1}, A_m]$. It is easy to see that the $j-i+1$ customers who have finished their service are customers $m-j-1, m-j, ..., m-i-1$, and the one under service at time A_m is customer $m-i$. Let us define $B_{m,i,j}$ as the random variable describing the total duration of those $j-i+1$ service completions, and let $f_{B_{m,i,j}}(.)$ and $F_{B_{m,i,j}}(.)$ be its probability density function and cumulative distribution function, respectively. Noting that the underlying process is a pure death process, we can see that $B_{m,i,j}$ equals to the summation of exponential random variables, and thus, it is hypoexponentially distributed with parameters $\mu_{m-j-1}, \mu_{m-j}, ..., \mu_{m-i-1}$. From Ross (2009), we have (in the case where all the rates are distinct) $f_{B_{m,i,j}}(t) = \sum_{l=m-j-1}^{m-i-1} \mu_l o_{m,i,j,l} e^{-\mu_l t}$ and $F_{B_{m,i,j}}(t) = 1 - \sum_{l=m-j-1}^{m-i-1} o_{m,i,j,l} e^{-\mu_l t}$ for $t \geq 0$, where $o_{m,i,j,l} = \prod_{n=m-j-1,\, n\neq l}^{m-i-1} \frac{\mu_n}{\mu_n - \mu_l}$. (By convention, an empty product equals to 1.) We denote by ε_{m-i} the exponential random variable that describes the service time of the $(m-i)^{th}$ (yet to complete service) customer, and let $f_{\varepsilon_{m-i}}(.)$ be its probability density function, then we have $f_{\varepsilon_{m-i}}(t) = \mu_{m-i} e^{-\mu_{m-i} t}$ for $t \geq 0$. Let us now define the random variable $C_{m,i,j}$ by $C_{m,i,j} = B_{m,i,j} + \varepsilon_{m-i}$. One may easily see that $\Pr\{R_m = i \mid R_{m-1} = j\} = \Pr\{B_{m,i,j} < T_m < C_{m,i,j}\}$. Due to the independence between T_m, $B_{m,i,j}$ and ε_{m-i}, we have

$$\Pr\{R_m = i \mid R_{m-1} = j\} = \mu_{m-i} \sum_{l=m-j-1}^{m-i-1} \mu_l o_{m,i,j,l} \int_0^\infty \int_0^\infty \int_y^{y+z} f_m(x) e^{-\mu_l y - \mu_{m-i} z} dx dy dz$$

for $i \leq j \leq m-2$. Similarly, for $j = i-1$, we have

$$\Pr\{R_m = i \mid R_{m-1} = i-1\} = \mu_{m-i} \int_0^\infty \int_0^z f_m(x) e^{-\mu_{m-i} z} dx dz,$$

which leads to

$$p_{m,i} = \mu_{m-i} \sum_{j=i}^{m-2} \sum_{l=m-j-1}^{m-i-1} \mu_l p_{m-1,j} o_{m,i,j,l} \int_0^\infty \int_0^\infty \int_y^{y+z} f_m(x) e^{-\mu_l y - \mu_{m-i} z} dx dy dz$$
$$+ p_{m-1,i-1} \mu_{m-i} \int_0^\infty \int_0^z f_m(x) e^{-\mu_{m-i} z} dx dz \tag{2}$$

for $1 \leq i \leq m-1$. As for the quantity $p_{m,0}$, it is simply given by

$$p_{m,0} = 1 - \sum_{i=1}^{m-1} p_{m,i} \tag{3}$$

for $2 \leq m \leq M$. Using Equations (2) and (3), the probabilities $p_{m,i}$ for $1 \leq m \leq M$ and $0 \leq i \leq m-1$ can be recursively computed starting with $m = 1$.

Next we show how the above probabilities can be used to characterize various performance measures. Let X_m, a random variable, denote the waiting time in queue of customer m, and let $E(X_m^k)$ be the corresponding k^{th} moment for $k \geq 1$. (For the rest of the paper, we use $E(Z^k)$ to denote the k^{th} moment of a random variable Z for $k \geq 1$.) Note that $X_1 = 0$ with probability 1, since it corresponds to the waiting time of the first customer. For $2 \leq m \leq M$, we have

$$E(X_m^k) = \sum_{i=1}^{m-1} p_{m,i} E(X_{m,i}^k),$$

where $X_{m,i}$ is the conditional random variable denoting the waiting time in queue for customer m, given that customer m finds, upon arrival, i customers in the system. Obviously, $X_{m,0} = 0$ with probability 1. For $1 \leq i \leq m-1$, the i customers seen by the m^{th} arrival are customers $m-1, m-2, ..., m-i$. For the $(m-i)^{th}$ customer who is currently in service, the remaining service time is still exponentially distributed with rate μ_{m-i}. Since their service times are independent and exponentially distributed, $X_{m,i}$ has a hypoexponential distribution with parameters $\mu_{m-1}, \mu_{m-2}, ..., \mu_{m-i}$. Hence, the quantities $E(X_{m,i}^k)$ for $k \geq 1$ can be easily computed. For example, we have $E(X_{m,i}) = \sum_{l=m-i}^{m-1} \frac{1}{\mu_l}$ and $E(X_{m,i}^2) = \sum_{l=m-i}^{m-1} \frac{1}{\mu_l^2} + (\sum_{l=m-i}^{m-1} \frac{1}{\mu_l})^2$.

Let the random variable X denote the waiting time in queue of an arbitrary customer among the M ones. Then, we obtain $E(X^k) = \frac{1}{M} \sum_{m=2}^{M} E(X_m^k) = \frac{1}{M} \sum_{m=2}^{M} \sum_{i=1}^{m-1} p_{m,i} E(X_{m,i}^k)$ for $k \geq 1$. In particular, we have

$$E(X) = \frac{1}{M} \sum_{m=2}^{M} \sum_{i=1}^{m-1} \sum_{l=m-i}^{m-1} \frac{p_{m,i}}{\mu_l}$$

and

$$Var(X) = \frac{1}{M} \sum_{m=2}^{M} \sum_{i=1}^{m-1} p_{m,i} \left[\sum_{l=m-i}^{m-1} \frac{1}{\mu_l^2} + \left(\sum_{l=m-i}^{m-1} \frac{1}{\mu_l} \right)^2 \right] - \frac{1}{M^2} \left(\sum_{m=2}^{M} \sum_{i=1}^{m-1} \sum_{l=m-i}^{m-1} \frac{p_{m,i}}{\mu_l} \right)^2.$$

From the probabilities $p_{m,i}$, we can also characterize the distribution of X. Specifically,

$\Pr\{X \le t\} = \frac{1}{M}(1 + \sum_{m=2}^{M} \Pr\{X_m \le t\}) = \frac{1}{M} + \frac{1}{M}\sum_{m=2}^{M}(p_{m,0} + \sum_{i=1}^{m-1} p_{m,i}\Pr\{X_{m,i} \le t\})$ for $t \ge 0$. In case all the rates are distinct, we have

$$\Pr\{X \le t\} = 1 - \frac{1}{M}\sum_{m=2}^{M}\sum_{i=1}^{m-1}\sum_{l=m-i}^{m-1} p_{m,i}o_{m,0,i-1,l}e^{-\mu_l t}.$$

In addition to waiting time, an important performance measure for systems with finite arrivals is *makespan*, namely, the time it takes the system to complete serving all customers. Since the server starts working at time zero, makespan can be computed as the departure time of the last customer (customer M). We define D_m as the random variable describing the departure time of customer m. Then $D_M = A_M + X_M + \varepsilon_M$, which leads to

$$E(D_M) = E(A_M) + E(X_M) + E(\varepsilon_M) = \sum_{m=2}^{M}\frac{1}{\lambda_m} + \sum_{i=1}^{M-1}\sum_{l=M-i}^{M-1}\frac{p_{M,i}}{\mu_l} + \frac{1}{\mu_M}.$$

Other measures of interest, such as those discussed in Cayirli and Veral (2003), can also be easily obtained. For example, the expected total time in system (waiting time + service time) for an arbitrary customer is given by $\frac{1}{M}(\sum_{m=1}^{M} E(X_m) + \frac{1}{\mu_m})$, or equivalently $\frac{1}{M}\sum_{m=1}^{M}\frac{1}{\mu_m} + \frac{1}{M}\sum_{m=2}^{M}\sum_{i=1}^{m-1}\sum_{l=m-i}^{m-1}\frac{p_{m,i}}{\mu_l}$; while the expected server idle time is given by $E(D_M) - \sum_{m=1}^{M}\frac{1}{\mu_m}$, or equivalently $\sum_{m=2}^{M}\frac{1}{\lambda_m} + \sum_{i=1}^{M-1}\sum_{l=M-i}^{M-1}\frac{p_{M,i}}{\mu_l} - \sum_{m=1}^{M-1}\frac{1}{\mu_m}$; and the expected server utilization is given by $\frac{\sum_{m=1}^{M}\frac{1}{\mu_m}}{E(D_M)}$, which can also be rewritten as $(\sum_{m=1}^{M}\frac{1}{\mu_m})(\sum_{m=2}^{M}\frac{1}{\lambda_m} + \sum_{i=1}^{M-1}\sum_{l=M-i}^{M-1}\frac{p_{M,i}}{\mu_l} + \frac{1}{\mu_M})^{-1}$. Various service level measures can also be obtained, including the probability that a customer waits more than a specified threshold or that makespan exceeds a certain threshold.

In some applications where the arrival process can be controlled, another useful performance measure is the amount of time, starting from time zero, until a customer arrives. This can be viewed as the indirect or offline waiting time. The expected arrival time of an arbitrary customer is given by $\frac{\sum_{m=2}^{M}\sum_{i=2}^{m} E(T_i)}{M}$.

Next, we consider three special cases for which the analysis simplifies further.

The Case of Exponential Inter-arrival Times: In this case, computing the probability $p_{m,i}$ simplifies by noting that, the probability $\Pr\{R_m = i \mid R_{m-1} = j\}$ for $1 \le i \le m-1$ and $i-1 \le j \le m-2$, can now be expressed as

$$\Pr\{R_m = i \mid R_{m-1} = j\} = \left(\prod_{l=i+1}^{j+1}\frac{\mu_{m-l}}{\mu_{m-l} + \lambda_m}\right)\frac{\lambda_m}{\mu_{m-i} + \lambda_m}. \tag{4}$$

The Case of Deterministic Inter-arrival Times: In this case, T_m is constant and equals to $\frac{1}{\lambda_m}$ for $2 \le m \le M$. The probability density function $f_m(t)$ is now a Dirac delta function at $\frac{1}{\lambda_m}$, which leads to $\Pr\{R_m = i \mid R_{m-1} = j\} = e^{-\frac{\mu_{m-i}}{\lambda_m}}\sum_{l=m-j-1}^{m-i-1} o_{m,i,j,l}\frac{\mu_l}{\mu_{m-i} - \mu_l}(e^{\frac{\mu_{m-i}-\mu_l}{\lambda_m}} - 1)$ for $i \le j \le m-2$ and $\Pr\{R_m = i \mid R_{m-1} = i-1\} = e^{-\frac{\mu_{m-i}}{\lambda_m}}$.

The case of deterministic inter-arrival times is of interest in applications where arrivals are determined by appointments and customers are punctual. In this case, arrival times correspond

to appointment times. Note that the above allows for heterogeneous service time distributions and generalizes earlier treatments that consider service times with homogenous rates (see, for example, Kaandorp and Koole 2007; Hassin and Mendel 2008).

The Case of Instantaneous Arrivals: An extreme case of the arrival process is one where customers arrive all at once. In this case, the expected waiting time of the m^{th} customer corresponds to the sum of the expected service times of customers $1, 2, ..., m-1$, i.e. $E(X_m) = \sum_{i=1}^{m-1} \frac{1}{\mu_i}$. This leads to $E(X) = \frac{1}{M} \sum_{m=2}^{M} \sum_{l=1}^{m-1} \frac{1}{\mu_l}$ and $E(D_M) = \sum_{m=1}^{M} \frac{1}{\mu_m}$.

4 The Multi-Server Case

In this section, we consider the case of a queueing system with multiple servers. We assume that there are s parallel and identical servers. For tractability, we focus on the case where service times are independent and exponentially distributed with rate μ. An arriving customer immediately begins service if there is an available server. Otherwise, she waits in queue and will be served by the first available server. All other assumptions are the same as those for the single server case in Section 3, and we continue to use similar notations.

As in the single server case, let us first characterize the probability $\Pr\{R_m = i \mid R_{m-1} = j\}$ for $2 \leq m \leq M$, $1 \leq i \leq m-1$ and $i \leq j \leq m-2$. In order for customer m to find i customers given that customer $m-1$ finds j customers, there must exactly be $j-i+1$ service completions during the time period $(A_{m-1}, A_m]$. We distinguish the following three cases.

Case 1, $s \leq i \leq j+1$**:** Once customer $m-1$ arrives, she joins the queue (if $j+1 > s$) or occupies the last available server (if $j+1 = s$). In both cases, customer m joins the queue once she arrives, and all the servers are busy during the time period $(A_{m-1}, A_m]$. When all servers are busy, the departure process is Poisson with rate $s\mu$. The probability $\Pr\{R_m = i \mid R_{m-1} = j\}$ corresponds to the probability that $j-i+1$ customers finish their service during T_m. So we may write

$$\Pr\{R_m = i \mid R_{m-1} = j\} = \int_0^\infty \frac{(s\mu x)^{j-i+1}}{(j-i+1)!} e^{-s\mu x} f_m(x) dx.$$

Case 2, $1 \leq i \leq j+1 < s$**:** In this case, there is no queue. Both customer $m-1$ and m immediately enter service once they arrive, and $\Pr\{R_m = i \mid R_{m-1} = j\}$ corresponds to the probability that exactly $j-i+1$ among $j+1$ customers finish their service during T_m. Noticing that $\binom{j+1}{j-i+1} = \binom{j+1}{i}$, this leads to

$$\Pr\{R_m = i \mid R_{m-1} = j\} = \int_0^\infty \binom{j+1}{i} (1 - e^{-\mu x})^{j-i+1} e^{-\mu x i} f_m(x) dx.$$

Case 3, $1 \leq i < s \leq j+1$**:** In this case, the system starts busy with $j-s+1$ queued

customers immediately after A_{m-1}. The probability $\Pr\{R_m = i \mid R_{m-1} = j\}$ corresponds to the probability that, within T_m, the first $j - s + 1$ queued customers leave the queue and enter service (which implies that $j - s + 1$ customers finish their service) and then $s - i$ customers finish their service afterwards, i.e., $j - i + 1$ service completions in total. We denote by I the random variable that describes the time needed to complete those $j - s + 1$ services, then I has an Erlang distribution with $j - s + 1$ stages and parameter $s\mu$. Thus, the probability density function of I, say $f_I(t)$, is given by $f_I(t) = \frac{(s\mu)^{j-s+1}t^{j-s}e^{-s\mu t}}{(j-s)!}$ for $t \geq 0$. This leads to

$$\Pr\{R_m = i \mid R_{m-1} = j\} = \int_0^\infty \int_0^x \binom{s}{i} (1 - e^{-\mu(x-t)})^{s-i} e^{-\mu(x-t)i} \frac{(s\mu)^{j-s+1}t^{j-s}e^{-s\mu t}}{(j-s)!} f_m(x)dtdx.$$

As for the single server case, using Equations (1) and (3), we can obtain $p_{m,i}$ for $2 \leq m \leq M$ and $1 \leq i \leq m - 1$ recursively.

Having the probabilities $p_{m,i}$ on-hand, we can now compute various performance measures. In particular, we have

$$E(X_m^k) = \sum_{i=s}^{m-1} p_{m,i} E(X_{m,i}^k)$$

for $1 \leq m \leq M$. Obviously $X_{m,i} = 0$ with probability 1 for $i \leq s - 1$. For $i \geq s$, $X_{m,i}$ is Erlang distributed with $i - s + 1$ stages and parameter $s\mu$. Consequently, we have

$$E(X_m) = \sum_{i=s}^{m-1} p_{m,i} \frac{i - s + 1}{s\mu} \tag{5}$$

and $E(X_m^2) = \sum_{i=s}^{m-1} p_{m,i} \frac{(i-s+1)(i-s+2)}{s^2\mu^2}$. Higher moments can be similarly computed. Since $E(X_m) = 0$ for $m \leq s$, we have

$$E(X^k) = \frac{1}{M} \sum_{m=s+1}^{M} E(X_m^k).$$

From the cumulative distribution function of Erlang distribution, we obtain $\Pr\{X_{m,i} \leq t\} = 1 - \sum_{l=0}^{i-s} \frac{(s\mu t)^l}{l!} e^{-s\mu t}$ and then $\Pr\{X_m \leq t\} = 1 - \sum_{i=s}^{m-1} \sum_{l=0}^{i-s} p_{m,i} \frac{(s\mu t)^l}{l!} e^{-s\mu t}$. This leads to

$$\Pr\{X \leq t\} = 1 - \frac{1}{M} \sum_{m=s+1}^{M} \sum_{i=s}^{m-1} \sum_{l=0}^{i-s} p_{m,i} \frac{(s\mu t)^l}{l!} e^{-s\mu t}.$$

As in Section 3, we can also characterize the makespan. However, in contrast to the single server case, makespan in the multi-server system no longer necessarily coincides with the departure time of customer M. The reason is that, if there are other customers under service at the time when customer M enters service, since service times are random, customer M may finish service and leave the system earlier than someone else. But, note that, although customer M may not be the last one to leave the system, she is still the last one to enter service by assumption (FCFS). Therefore, makespan equals to, the sum of, the time it takes customer M to enter service, and the time it takes to empty the system after she enters service. When customer M

arrives, seeing i customers in system, there are two possibilities. The first possibility is $i \leq s-1$, which implies that there is at least one idle server, and customer M immediately enters service without waiting. In this case, the time to empty the system corresponds to the longest completion time among the $i+1$ services. This time has the hypoexponential distribution with parameters $(i+1)\mu, i\mu, ..., \mu$. Thus, if customer M finds i customers in the system upon her arrival and $i \leq s-1$, then the expected makespan is given by $\sum_{m=2}^{M} \frac{1}{\lambda_m} + \sum_{l=1}^{i+1} \frac{1}{l\mu}$.

The second possibility is $i \geq s$, which implies that customer M has to wait in queue before being served. In this case, the waiting time of customer M is Erlang distributed with $i-s+1$ stages and parameter $s\mu$, and the time to empty the system has the hypoexponential distribution with rates $s\mu, (s-1)\mu, ..., \mu$. Thus, if customer M finds i customers in system upon her arrival and $i \geq s$, then the expected makespan is given by $\sum_{m=2}^{M} \frac{1}{\lambda_m} + \frac{i-s+1}{s\mu} + \sum_{l=1}^{s} \frac{1}{l\mu}$.

Putting it all together, the unconditional expected makespan can be obtained as

$$E[Makespan] = \sum_{m=2}^{M} \frac{1}{\lambda_m} + \sum_{i=0}^{s-1} \left(p_{M,i} \sum_{l=1}^{i+1} \frac{1}{l\mu} \right) + \sum_{i=s}^{M-1} p_{M,i} \left(\frac{i-s+1}{s\mu} + \sum_{l=1}^{s} \frac{1}{l\mu} \right).$$

Other performance measures can be similarly obtained, and we omit the details for the sake of brevity.

The Case of Exponential Inter-Arrival Times: Using similar arguments as in the single server case and noting that, when there are l customers in the system, the service rate is $\mu\ min(l, s)$, we obtain

$$\Pr\{R_m = i \mid R_{m-1} = j\} = \left(\prod_{l=i+1}^{j+1} \frac{\mu \min(l,s)}{\mu \min(l,s) + \lambda_m} \right) \frac{\lambda_m}{\mu \min(i,s) + \lambda_m}.$$

The Case of Deterministic Inter-Arrival Times: This also follows the approach used for the single server case by setting $f_m(t)$ as a Dirac delta function at $\frac{1}{\lambda_m}$ and then computing $\Pr\{R_m = i \mid R_{m-1} = j\}$ using the corresponding equations.

The Case of Instantaneous Arrivals: In this case, the first s customers have zero waiting time, and customer $s+i$ $(1 \leq i \leq M-s)$ waits for i service completions to start service. This leads to $E(X) = \frac{(M-s)^2+(M-s)}{2s\mu M}$ and $E(D_M) = \frac{M-s}{s\mu} + \sum_{l=1}^{s} \frac{1}{l\mu}$.

5 Numerical Experiments

In this section, we describe results from the numerical experiments we carried out to examine the impact of features that are unique to the systems we consider, namely the finite number of arrivals, the heterogeneity in inter-arrival times, and the heterogeneity in service times. Our objective is three-fold: (1) to draw insights into how these specific features affect system performance, (2) to show that models which do not explicitly account for these features can lead

to significant errors in performance evaluation, and (3) to illustrate how the models we present in this paper can be used to support operational decision making, particularly as it pertains to capacity planning (see Section 7 for discussions on additional applications). In Sections 5.1 and 5.2, we consider respectively the impact of heterogeneity in inter-arrival times and service times, on various performance measures. In Section 5.3, we discuss the impact of heterogeneity on capacity levels. Throughout this section, we focus on the single server setting. We also studied the multi-server setting and obtained similar results; we omit the details for the sake of brevity.

5.1 The Impact of Heterogeneity in Inter-Arrival Times

To examine the impact of heterogeneity in inter-arrival times, we investigate five arrival processes with different inter-arrival time features that may arise naturally in practice (see our earlier discussion in the introduction section). These five processes are described in Table 1. To allow for a fair comparison between different processes, we maintain the same number of customers and the same average expected inter-arrival time (equal to $\frac{1}{\lambda}$) across processes. The first process corresponds to a setting where the expected inter-arrival times decrease with each subsequent arrival. Specifically, we let $E(T_m) = \frac{M-m+1}{M}\frac{2}{\lambda}$ for $m = 2, ..., M$. The other processes correspond similarly to settings where expected inter-arrival times (1) increase with each subsequent arrival, (2) decrease and then increase, (3) increase and then decrease, and (4) are constant. Note that $\{E(T_m)|m = 2, ..., M\}$ in the four heterogeneous processes are indeed four specific permutations of the sequence $\{\frac{1}{M}\frac{2}{\lambda}, ..., \frac{M-1}{M}\frac{2}{\lambda}\}$.

Inter-arrival Time Features	Expected Inter-arrival Times	
Decreasing	$E(T_m) = \frac{M-m+1}{M}\frac{2}{\lambda}$	for $m = 2, ..., M$
Increasing	$E(T_m) = \frac{m-1}{M}\frac{2}{\lambda}$	for $m = 2, ..., M$
Decreasing/Increasing	$E(T_m) = \frac{M-2m+3}{M}\frac{2}{\lambda}$	for $m = 2, ..., \frac{M+2}{2}$
	$E(T_m) = \frac{2m-M-2}{M}\frac{2}{\lambda}$	for $m = \frac{M+4}{2}, ..., M$
Increasing/Decreasing	$E(T_m) = \frac{2m-2}{M}\frac{2}{\lambda}$	for $m = 2, ..., \frac{M}{2}$
	$E(T_m) = \frac{2M-2m+1}{M}\frac{2}{\lambda}$	for $m = \frac{M+2}{2}, ..., M$
Constant	$E(T_m) = \frac{1}{\lambda}$	for $m = 2, ..., M$

Table 1: Inter-Arrival Time Features

A representative sample from an extensive set of numerical results on expected waiting time is shown in Figure 1 (additional results are available from the authors upon request). The results are shown for systems where inter-arrival times are exponentially distributed and service times are i.i.d. and exponentially distributed (the results are qualitatively the same for other common inter-arrival time distributions we tested). Note that by varying λ for fixed M and μ, the workload in system (i.e. the traffic intensity or the utilization of server) over the arrival

period, as measured by $\rho = \frac{\lambda}{\mu}$, is varied. On the other hand, by varying M for fixed λ and μ, the workload remains constant, but the period of arrivals, as measured by the expected time until the last customer arrives, is varied.

The following observations can be made regarding system performance in terms of the expected waiting time of an arbitrary customer.

- Arrival processes with different features can lead to significantly different expected waiting times. Moreover, there is a considerable difference between the performance of systems with constant expected inter-arrival times and those with heterogeneous expected inter-arrival times. Clearly, ignoring the heterogeneity in the arrival process can lead to significant errors in performance evaluation.

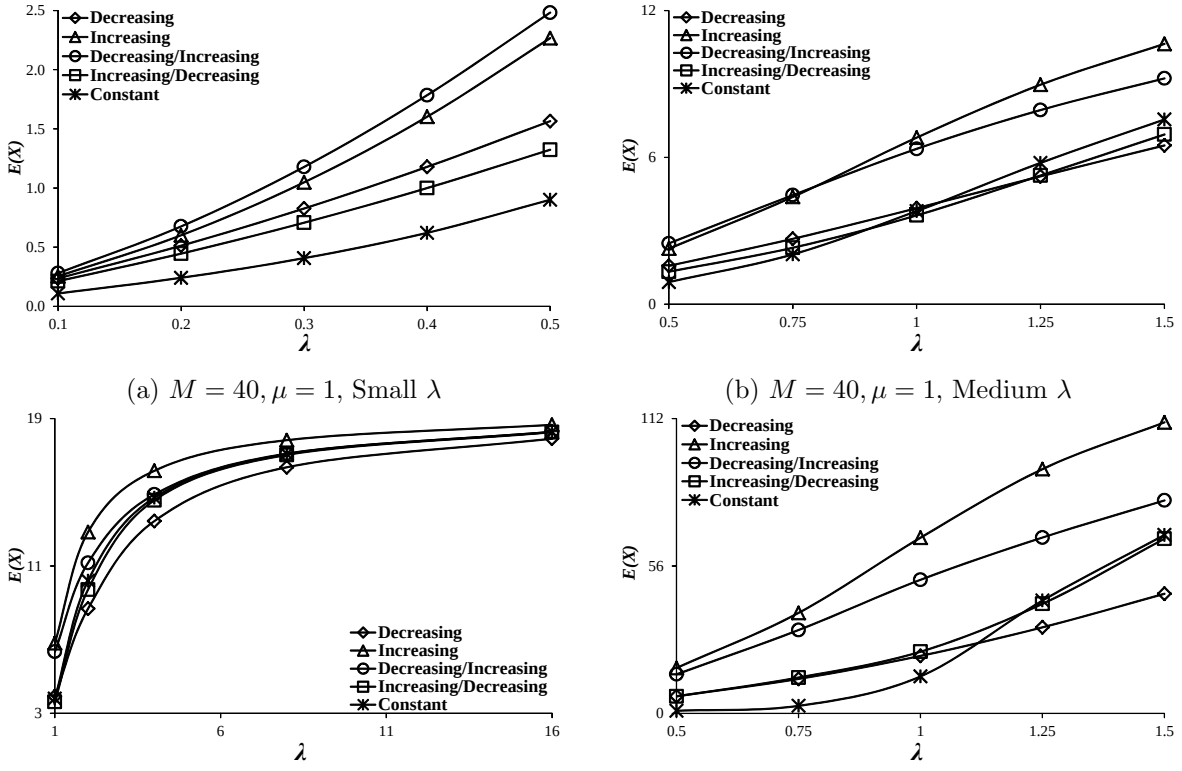
- Arrival processes with constant expected inter-arrival times does not guarantee better performance. In other words, arrivals with a fixed *intensity* may not necessarily be preferable to arrivals with variable intensity.

- Arrival processes with "Decreasing" inter-arrival times always perform better than processes with "Increasing" and "Decreasing/Increasing" inter-arrival times. In other words, processes where arrivals peak later leads to better performance than those where arrivals peak earlier. This is due to the fact that a peak in arrivals that occurs early in the process can delay all customers that arrive subsequently.

- The relative performance of different arrival processes depends on problem parameter values. For example, when ρ is small ($\rho << 1$), "Constant" is the best as it maximizes the spreading-out of arrivals, reducing the possibility of congestion. On the other hand, when ρ is large ($\rho >> 1$), congestion is inevitable. In that case, arrival processes, with features that can limit the number of customers affected by congestion, become more preferable, explaining, for example why "Decreasing" is the best.

- The difference in performance between different arrival processes decreases as λ increases. The performances become indistinguishable as λ gets very large, in which case, all customers arrive nearly instantaneously.

- The threshold on ρ that determines the relative performance of different arrival processes is affected by M. For example, the larger M is, the larger is the value of ρ under which "Constant" performs the best. In Section 6, we provide an approximation that allows us to specify these thresholds in closed form.



(a) $M = 40, \mu = 1$, Small λ

(b) $M = 40, \mu = 1$, Medium λ

(c) $M = 40, \mu = 1$, Large λ

(d) $M = 400, \mu = 1$, Medium λ

Figure 1: Impact of Inter-Arrival Time Features on Expected Waiting Time

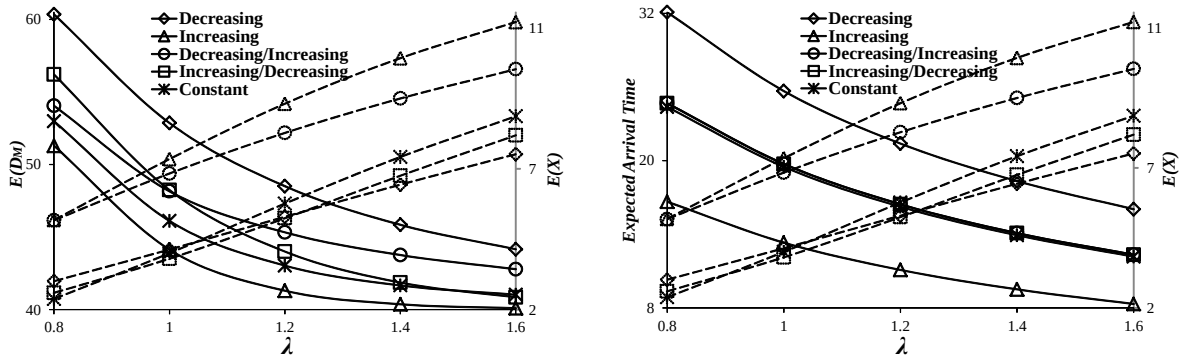
In addition to the expected waiting time, we also obtained results for the impact of different arrival processes on the variance of waiting time. For brevity, we omit these results (available from the authors upon request) and note the following.

- Most of the observations on expected waiting time continue to hold. For example, arrival processes with different features lead to significantly different variances, with the "Constant" inter-arrival time feature not always leading to the lowest variance. The difference in variances induced by different arrival processes decreases as λ increases, with the threshold on ρ that determines the relative performance of different processes affected by M.

- Systems with "Constant" and "Increasing/Decreasing" inter-arrival times always perform better than the others. In particular, for small ρ, "Constant" performs the best as it smoothes the arrival process and reduces the possibility of congestion. However, for large ρ, congestion is inevitable, and "Increasing/Decreasing" performs the best since it separates the arrival process into two sub-processes with each one having a lower peak value of congestion.

In Figures 2a and 2b, we present results that illustrate the impact of different arrival pro-

cesses on the expected makespan and the expected arrival time, with solid lines representing expected makespan and expected arrival time, respectively, and dashed lines representing expected waiting time. Here too, arrival processes with different inter-arrival time features can lead to significantly different expected makespans, with "Constant" not necessarily being the best. While the average expected inter-arrival time stays the same for all processes, makespan is minimized by minimizing the expected waiting time in queue of the last customer (or equivalently minimizing idleness of the server). This is achieved by maximizing the number of customers that arrive early, explaining why "Increasing" performs the best and "Decreasing" performs the worst. The relative performance of other processes depends on system utilization. For example, when utilization is low, "Decreasing/Increasing" performs better than "Increasing/Decreasing". Although the peak of arrivals occurs later under "Decreasing/Increasing", there is enough capacity in the system to ensure that most customers would clear before the last customer arrives. This is not the case when utilization is high. There, it is preferable to have the peak of arrivals occur as early as possible to minimize the idleness of the server, explaining why "Increasing/Decreasing" is more preferable. Same as for the expected waiting time, the difference in the expected makespan induced by different arrival processes decreases as λ increases. This difference approaches zero as λ becomes very large. Similar to the expected makespan, the expected arrival time is lower when more customers arrive earlier. Therefore, the relative performance of different arrival processes on the expected arrival time coincides with the one observed for the expected makespan.



(a) Makespan and Waiting Time

(b) Arrival Time and Waiting Time

Figure 2: Impact of Inter-Arrival Time Features on Makespan and Arrival Time ($M = 40$, $\mu = 1$)

5.2 Impact of Heterogeneity in Service Times

In results (the details of which are not shown here for the sake of brevity), we examine the impact of heterogeneity in service times. Here again, we investigate five service processes with different service time features, as shown in Table 2. These include settings where expected service times (1) decrease with each subsequent service completion, (2) increase, (3) decrease and then increase, (4) increase and then decrease, and (5) are constant. To allow for a fair comparison between different processes, we maintain the same number of customers and the same average expected service time (equal to $\frac{1}{\mu}$) across processes.

Service Time Features	Expected Service Times	
Decreasing	$E(\varepsilon_m) = \frac{M-m+1}{M+1}\frac{2}{\mu}$	for $m = 1, ..., M$
Increasing	$E(\varepsilon_m) = \frac{m}{M+1}\frac{2}{\mu}$	for $m = 1, ..., M$
Decreasing/Increasing	$E(\varepsilon_m) = \frac{M-2m+1}{M+1}\frac{2}{\mu}$	for $m = 1, ..., \frac{M}{2}$
	$E(\varepsilon_m) = \frac{2m-M}{M+1}\frac{2}{\mu}$	for $m = \frac{M+2}{2}, ..., M$
Increasing/Decreasing	$E(\varepsilon_m) = \frac{2m}{M+1}\frac{2}{\mu}$	for $m = 1, ..., \frac{M}{2}$
	$E(\varepsilon_m) = \frac{2M-2m+1}{M+1}\frac{2}{\mu}$	for $m = \frac{M+2}{2}, ..., M$
Constant	$E(\varepsilon_m) = \frac{1}{\mu}$	for $m = 1, ..., M$

Table 2: Service Time Features

Similar to what we have observed for the arrival process, service processes with different service time features can lead to significantly different expected waiting times, with the "Constant" service time feature again not necessarily being the best. Service processes with features that postpone congestion are preferable when utilization is high ($\rho >> 1$) (e.g., "Increasing" tends to perform the best). This is perhaps also consistent with known results from the scheduling literature regarding the optimality of the "shortest processing time first" scheduling rule. However, when utilization is low ($\rho << 1$), this is not the case, and "Constant" performs the best for reasons similar to those explained for the arrival process.

With regard to the variance of waiting time, again for the same reasons as explained in the previous section, when utilization is high, "Decreasing/Increasing" performs the best, and when utilization is low, "Constant" performs the best. For expected makespan, the order of preference tends to be reversed, with features that reduce congestion later in the arrival process being preferable (in other words, for the expected makespan, it is preferable that arrivals with shorter service times occur later in the arrival process).

5.3 On the Impact on Capacity Levels

In this section, we examine how arrival processes with different features affect the capacity needed to guarantee a specified level of performance (e.g., a maximum expected waiting time

or makespan). For single server systems, determining this capacity requires determining the minimum processing rate. For systems with multiple servers, this requires determining the minimum number of servers.

In Figure 3, we show the minimum service rate μ needed under each of the four heterogeneous arrival processes described in Table 1 to meet a specified minimum expected waiting time target. In this case, the specified target is the expected waiting time obtained under the arrival process with "Constant" inter-arrival times at $\mu = 1$. As we can see, the difference in the capacity levels needed under different arrival processes can be dramatically different. Ignoring the heterogeneity in inter-arrival times (and similarly in service requirements) can therefore lead to significant under or over investments in capacity, resulting in either poor service quality or unjustified additional capacity cost.

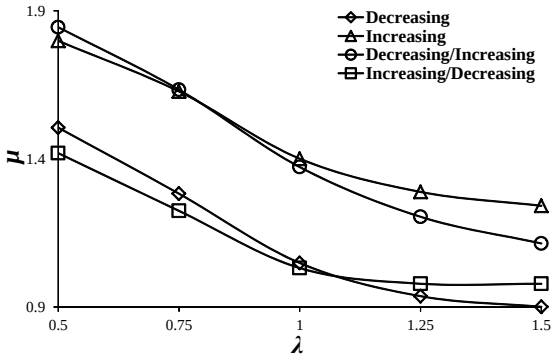


Figure 3: Impact of Inter-Arrival Time Features on Capacity Level ($M = 100$)

6 A Fluid Approximation

Although the performance analysis given in Sections 3 and 4 is exact, we resorted to numerical analysis in order to draw the conclusions in Section 5. This is because the exact results are not in closed form and therefore difficult to use to characterize structural results. To provide further support for the numerical results, we discuss in this section a deterministic *fluid* approximation that does yield closed form expressions and that allows us to capture key features of our setting. The objective from this approximation is of course not to substitute for the exact analysis which is easy to implement, but to analytically confirm the numerical findings of Section 5 and provide evidence of their robustness. The approximation may also be useful in investigating additional structural results and as a first step in examining first order effects. The approximation does not require the assumption of exponential service times and, therefore, is useful for the study of more general systems. For the sake of brevity, we describe the approximation in the context of

the single server model. However, extending the treatment to the multi-server case is relatively straightforward.

We treat all customer inter-arrival and service times as being deterministic and replace all corresponding random variables by their expected values. (For every quantity Z defined in Section 3 for the original model, we define a corresponding quantity Z^F for the fluid approximation). We treat the arrival of customers as fluid, one unit per customer, that is "pumped-in" to the system at a constant rate λ_m over the time period $(A^F_{m-1}, A^F_m]$ for $m = 2, ..., M$. Since $T_1 = 0$ in the original model, we assume all the fluid associated with the first customer is present in the system at time 0. Similarly, we treat the service process as fluid, also one unit per customer, that is "pumped-out" at a constant rate μ_m over the time period $(D^F_{m-1}, D^F_m]$ for $m = 2, ..., M$, and at the rate μ_1 over the time period $(0, D^F_1]$, where $D^F_m = \max(D^F_{m-1}, A^F_m) + \frac{1}{\mu_m}$ with $D^F_1 = \frac{1}{\mu_1}$. By induction, it is straightforward to show that $D^F_m = \max_{1\le i\le m}\{\sum_{j=2}^{i}\frac{1}{\lambda_j} + \sum_{j=i}^{m}\frac{1}{\mu_j}\}$ for $m = 1, ..., M$ (by convention, an empty sum equals to 0).

We define $A^F(t)$ and $D^F(t)$ as the cumulative arrivals to the system and the cumulative departures from the system by time t, respectively (with $A^F(0) = 1$). It is not difficult to see that, $A^F(t)$ and $D^F(t)$ are piecewise linear functions (see Figure 4 for an illustration). The area between $A^F(t)$ and $D^F(t)$ over the interval $[0, D^F_M]$ corresponds to the total time spent in the system for all customers, which, when divided by the total number of customers, yields the expected time in system of an arbitrary customer. Let us denote the expected time in system of an arbitrary customer by $E^F(Y)$. Then, we have

$$E^F(Y) = \frac{\int_0^{D^F_M}[A^F(t) - D^F(t)]dt}{M}.$$

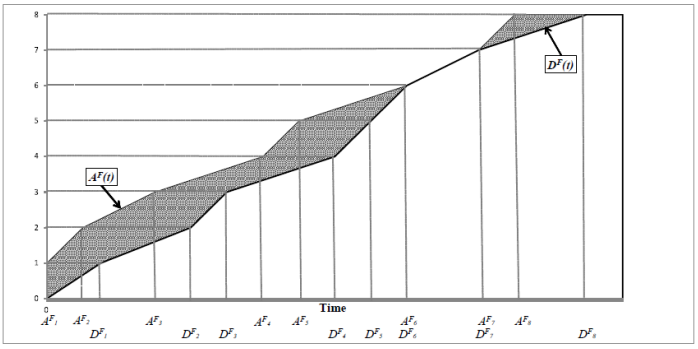


Figure 4: An Illustration of the Fluid Approximation

The area under $A^F(t)$ over the interval $[0, D^F_M]$ is the sum of the areas of $M - 1$ trapezoids and one rectangle. If we define $S^F_m(A)$ as the area of the m^{th} trapezoid from left, then $S^F_m(A) = (m + \frac{1}{2})\frac{1}{\lambda_{m+1}}$ for $m = 1, ...M - 1$, and the area of the rectangle, which we denote by $S^F_r(A)$,

equals to $M(D_M^F - A_M^F)$.

We now let $S^F(A)$ denote the total area under $A^F(t)$ for $t \in [0, D_M^F]$. Then, we can show that $S^F(A) = \sum_{m=1}^{M-1} S_m^F(A) + S_r^F(A) = \sum_{m=2}^{M} m\frac{1}{\lambda_m} - (M + \frac{1}{2})A_M^F + MD_M^F$.

Similarly, we denote $S^F(D)$ as the area under $D^F(t)$ over the interval $[0, D_M^F]$. This is the sum of the areas of one triangle and $M-1$ trapezoids. The area of the triangle, which we denote by $S_t^F(D)$, is $\frac{1}{2}D_1^F$. The area of the m^{th} trapezoid from left, which we denoted by $S_m^F(D)$, is given by $S_m^F(D) = (m + \frac{1}{2})(D_{m+1}^F - D_m^F)$ for $m = 1, ..., M-1$. This implies that $S^F(D) = \sum_{m=1}^{M-1} S_m^F(D) + S_t^F(D) = (M - \frac{1}{2})D_M^F - \sum_{m=1}^{M-1} D_m^F$.

Putting it together, the expected time in system can be written as

$$E^F(Y) = \frac{S^F(A) - S^F(D)}{M} = \frac{\sum_{m=2}^{M} m\frac{1}{\lambda_m} - (M + \frac{1}{2})A_M^F + \sum_{m=1}^{M} D_m^F - \frac{1}{2}D_M^F}{M}.$$

Using the above explicit expressions, we can evaluate each of the arrival and service time processes considered in the numerical study of the previous sections. For the sake of brevity, we focus on the relative performance of different arrival processes. Without loss of generality, we scale time such that $\mu_m = 1$ for $m = 1, ..., M$, and the sequences $\{\frac{1}{\lambda_m} | m = 2, ..., M\}$ are as those sequences in Table 1. For the four arrival processes with heterogeneous inter-arrival times, $\frac{1}{\lambda_m} \in \{\frac{1}{M}\frac{2}{\lambda}, ..., \frac{M-1}{M}\frac{2}{\lambda}\}$, and for the process with constant inter-arrival times, we have $\frac{1}{M}\frac{2}{\lambda} \leq \frac{1}{\lambda} \leq \frac{M-1}{M}\frac{2}{\lambda}$. In what follows, we consider the average time in system instead of the average waiting time in queue. Since the total service times of all customers are the same among all the arrival processes, the ordering of processes will not be affected by using time in systems instead of waiting time in queue. Let $E^F(Y)_{(C)}$, $E^F(Y)_{(D)}$, $E^F(Y)_{(I)}$, $E^F(Y)_{(DI)}$, and $E^F(Y)_{(ID)}$ refer respectively to the expected time in system for the arrival processes with "Constant", "Decreasing", "Increasing", "Decreasing/Increasing", and "Increasing/Decreasing" inter-arrival times.

We distinguish three different cases: Case 1 ($\frac{1}{M}\frac{2}{\lambda} \geq 1$); Case 2 ($\frac{M-1}{M}\frac{2}{\lambda} \leq 1$); and Case 3 ($\frac{1}{M}\frac{2}{\lambda} < 1 < \frac{M-1}{M}\frac{2}{\lambda}$).

Case 1: This is an obvious case. We have $D_M^F = \frac{M+(\lambda-1)}{\lambda}$ for all the processes. Therefore, it is easy to show that $E^F(Y)$ is the same for all the processes.

Case 2: In this case, $D_M^F = M$ for all the processes. After some algebra, we obtain $E^F(Y)_{(C)} = \frac{(\lambda-1)M^2+2M-1}{2\lambda M}$, $E^F(Y)_{(D)} = \frac{(3\lambda-4)M^2+9M-5}{6\lambda M}$, $E^F(Y)_{(I)} = \frac{(3\lambda-2)M^2+3M-1}{6\lambda M}$, $E^F(Y)_{(DI)} = \frac{(2\lambda-2)M^2+3M}{4\lambda M}$, and $E^F(Y)_{(ID)} = \frac{(2\lambda-2)M^2+3M}{4\lambda M}$. Then, we can easily show that

$$E^F(Y)_{(D)} < E^F(Y)_{(ID)} = E^F(Y)_{(DI)} < E^F(Y)_{(C)} < E^F(Y)_{(I)},$$

which is consistent with the results in Section 5.1.

Case 3: Denote $D_{M(C)}^F$, $D_{M(D)}^F$, and $D_{M(I)}^F$ as the makespan for the arrival processes with

"Constant", "Decreasing", and "Increasing" inter-arrival times, respectively. We can show that (see the detail derivations in the online supplement)

$D^F_{M(C)} = \begin{cases} \frac{M+(\lambda-1)}{\lambda} & \text{for } \lambda \in (2\frac{1}{M}, 1) \\ M & \text{for } \lambda \in [1, 2\frac{M-1}{M}) \end{cases}$, $D^F_{M(D)} = \frac{(\lambda^2+4)M+(2\lambda-4)}{4\lambda}$, and

$D^F_{M(I)} = \begin{cases} \frac{M+(\lambda-1)}{\lambda} & \text{for } \lambda \in (2\frac{1}{M}, 1) \\ M & \text{for } \lambda \in [1, 2\frac{M-1}{M}) \end{cases}$. Then, $E^F(Y)_{(C)} = \begin{cases} \frac{2\lambda M-\lambda}{2\lambda M} & \text{for } \lambda \in (2\frac{1}{M}, 1) \\ \frac{(\lambda-1)M^2+2M-1}{2\lambda M} & \text{for } \lambda \in [1, 2\frac{M-1}{M}) \end{cases}$,

$E^F(Y)_{(D)} = \frac{\lambda^3 M^2-(3\lambda^2-24\lambda)M-10\lambda}{24\lambda M}$ and $E^F(Y)_{(I)} = \begin{cases} \frac{\lambda^3 M^2-(3\lambda^2-6\lambda)M-\lambda}{6\lambda M} & \text{for } \lambda \in (2\frac{1}{M}, 1) \\ \frac{(3\lambda-2)M^2+3M-1}{6\lambda M} & \text{for } \lambda \in [1, 2\frac{M-1}{M}) \end{cases}$.

Applying the implicit function theorem, it is easy to show that there exits an $\alpha^F(M) \in (1, 2\frac{M-1}{M})$ increasing in M such that

$$E^F(Y)_{(C)} < E^F(Y)_{(D)} < E^F(Y)_{(I)} \text{ for } \lambda \in \left(2\frac{1}{M}, \alpha^F(M)\right), \text{ and}$$

$$E^F(Y)_{(D)} < E^F(Y)_{(C)} < E^F(Y)_{(I)} \text{ for } \lambda \in \left(\alpha^F(M), 2\frac{M-1}{M}\right),$$

which is again consistent with the results in Section 5.1. (We can obtain similar expressions for the expected time in system for the arrival processes with "Decreasing/Increasing" and "Increasing/Decreasing" inter-arrival times. For the sake of brevity, we omit the details. The relative ordering also coincides with the one observed in the previous section.)

Other results from Section 5.1 can also be confirmed using the fluid approximation. For example, the difference in performance between different arrival processes decreases as λ increases and approaches 0 as $\lambda \to \infty$. The limit case of $\lambda \to \infty$ corresponds to the case of instantaneous arrivals. In that case, the expression for the expected time in system reduces to $E^F(Y) = \frac{1}{M}\sum_{m=2}^{M}\sum_{j=1}^{m-1}\frac{1}{\mu_j} + \frac{1}{2M}\sum_{m=1}^{M}\frac{1}{\mu_j}$. It is straightforward to show that this expression converges asymptotically to the expression from the exact analysis in Section 3 as $M \to \infty$, with $\lim_{M\to\infty}\frac{E^F(Y)}{E(Y)} = 1$.

7 Example Applications

In this section, we describe example applications where the results from our analysis can be used to support operational decision making.

7.1 A Job Sequencing Problem

Consider the job sequencing problem described in the introduction section. In particular, consider a system with M jobs to be sequenced on two production stages (e.g., a manufacturing stage and an inspection stage) in series, with a single server at each stage (the extension to

multiple servers is straightforward). All M jobs are available at time 0. The processing time of job h for $h = 1, ..., M$, at stage r for $r = 1, 2$, is exponentially distributed with rate $\mu_{(h),r}$. Once a sequence is selected, the jobs are processed in that sequence on both stages without idling (i.e., a server never idles if there is a job available to be processed). For a given sequence, the expected waiting time of an arbitrary job at the first stage equals $\frac{1}{M}\sum_{m=2}^{M}\sum_{l=1}^{m-1}\frac{1}{\mu_{l,1}}$, where $\mu_{l,1}$ is the processing rate of the job assigned to position l (the l^{th} to process), and the corresponding total time spent in that stage equals to $\frac{1}{M}\sum_{m=1}^{M}\sum_{l=1}^{m}\frac{1}{\mu_{l,1}}$. To characterize the performance at the second stage, we must first characterize the inter-arrival time distributions to that stage. This can be done by recognizing that, given a job sequence, the distributions of inter-arrival times to the second stage are simply the distributions of processing times at the first stage. In particular, if job h is assigned position m ($m \geq 2$) in the sequence, then the time between the $(m-1)^{th}$ and m^{th} arrivals to the second stage is exponentially distributed with rate $\mu_{(h),1}$. Consequently, the expected waiting time for an arbitrary job at the second stage is given by $\frac{1}{M}\sum_{m=2}^{M}\sum_{i=1}^{m-1}\sum_{l=m-i}^{m-1}\frac{p_{m,i}}{\mu_{l,2}}$, where $p_{m,i}$ can be computed via the analysis we developed in Section 3, with λ_l and μ_l in Equation (4) replaced by $\mu_{l,1}$ and $\mu_{l,2}$ for all l, respectively. This leads to the expected total waiting time in the system of an arbitrary job as $\frac{1}{M}\sum_{m=2}^{M}\sum_{i=1}^{m-1}(\frac{1}{\mu_{i,1}} + \sum_{l=m-i}^{m-1}\frac{p_{m,i}}{\mu_{l,2}})$. Other performance measures can be similarly obtained. In particular, the expected makespan is given by $\sum_{m=1}^{M}\frac{1}{\mu_{m,1}} + \sum_{i=1}^{M-1}\sum_{l=M-i}^{M-1}\frac{p_{M,i}}{\mu_{l,2}} + \frac{1}{\mu_{M,2}}$.

From the above analysis, we can see that by controlling the job sequence, the system manager can control the distributions of inter-arrival times at the second stage, and therefore the corresponding system performance. Next, we present numerical results for an example system where $\mu_{(h),1} = \frac{M+1}{h}\frac{\varepsilon}{2}$ and $\mu_{(h),2} = \mu$, for $h = 1, ..., M$ and constants ε and μ. We evaluate four different sequences (four permutations of the sequence $\{\frac{M+1}{1}\frac{\varepsilon}{2}, ..., \frac{M+1}{M}\frac{\varepsilon}{2}\}$) as described in Table 3 (to be consistent with the other sections, we name the sequences according to the expected service times instead of the service rates). The first sequence corresponds to an ordering of the jobs in decreasing expected service times at stage 1, which implies an ordering of the jobs in decreasing expected inter-arrival times at stage 2. The second sequence corresponds to an ordering in increasing expected inter-arrival times at stage 2, while the third and fourth correspond respectively to, decreasing and then increasing, and, increasing and then decreasing, orderings of the expected inter-arrival times at stage 2.

Figure 5 provides comparisons of the four job sequences under different values of delay costs (consistent with the job scheduling literature, we assign a delay cost, w_r per job per unit time at stage r for $r = 1, 2$; without loss of generality, we let $w_1 = 1$ and vary w_2; the case of

Job Sequences	Expected Service Times at Stage 1
Decreasing	$E(\varepsilon_m) = \frac{M-m+1}{M+1}\frac{2}{\varepsilon}$ for $m = 1, ..., M$
Increasing	$E(\varepsilon_m) = \frac{m}{M+1}\frac{2}{\varepsilon}$ for $m = 1, ..., M$
Decreasing/Increasing	$E(\varepsilon_m) = \frac{M-2m+1}{M+1}\frac{2}{\varepsilon}$ for $m = 1, ..., \frac{M}{2}$ $E(\varepsilon_m) = \frac{2m-M}{M+1}\frac{2}{\varepsilon}$ for $m = \frac{M+2}{2}, ..., M$
Increasing/Decreasing	$E(\varepsilon_m) = \frac{2m}{M+1}\frac{2}{\varepsilon}$ for $m = 1, ..., \frac{M}{2}$ $E(\varepsilon_m) = \frac{2M-2m+1}{M+1}\frac{2}{\varepsilon}$ for $m = \frac{M+2}{2}, ..., M$

Table 3: Job Sequences

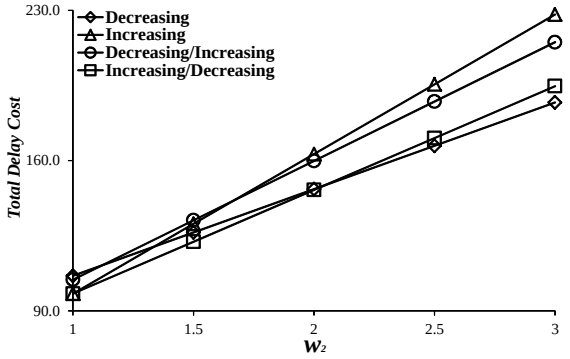


Figure 5: Impact of Job Sequence on Delay Cost ($M = 100$, $\varepsilon = 1$, $\mu = 0.5$)

$w_1 = w_2 = 1$ allows us to compare the expected total delay in the system for the four different job sequences). As we can see, the four job sequences lead to very different total delay costs. Perhaps surprisingly, the "Increasing" sequence which minimizes the delay cost at stage 1 does not necessarily minimize the expected total delay cost. In fact, for sufficiently large w_2, such a sequence performs the worst. This can be explained as follows. The "Increasing" sequence generates the "Increasing" inter-arrival times at stage 2, which, as discussed in Section 5.1, results in long waiting times. On the other hand, the "Decreasing" sequence, although leading to long waiting times at stage 1, generates the "Decreasing" inter-arrival times at stage 2 and therefore results in short waiting times at that stage. The net effect, when w_2 is large, is lower total delay cost.

Additional results (the details of which are not shown here for the sake of brevity) indicate that the four job sequences also lead to significant differences in makespan, with the "Increasing" sequence always performing the best. Note that characterizing the optimal sequence is difficult in general (even for the deterministic setting, the problem is strongly NP-hard; see discussions from Pinedo 2012), and is outside the scope of this paper.

7.2 A Flight Boarding Problem

Consider the flight boarding problem described in the introduction section. There are M passengers waiting to board a flight, and they are grouped into K equal size zones, each consisting of

$\frac{M}{K}$ passengers (assuming M is divisible by K). Passengers from a zone are called to embark only after all the passengers from a higher ranked zone have finished embarking. The announcement of each zone results in arrivals to the gate drawn from a population of $\frac{M}{K}$ passengers. Assuming each passenger takes an exponentially distributed amount of time to arrive, independent of other customers, then the arrival process for each zone corresponds to a pure death process, with the inter-arrival time between customer $m-1$ and customer m being exponentially distributed with rate $(\frac{M}{K}+1-m)\lambda$ for $m = 2, ..., \frac{M}{K}$ (the arrival time of the first customer is exponentially distributed with rate $\frac{M}{K}\lambda$). This also implies that the expected inter-arrival times within a zone is strictly increasing. Assuming that service times are exponentially distributed with rate μ, the results of Section 3 can be readily applied to obtain various measures of performance. In particular, the expected waiting time of an arbitrary passenger can be obtained by setting $\lambda_m = (\frac{M}{K}+1-m)\lambda$ for $m = 2, ..., \frac{M}{K}$ and $\mu_m = \mu$ for $m = 1, ..., \frac{M}{K}$ in Equation (4), and the expected makespan (the expected boarding completion time of all zones) is given by $K[\frac{1}{\lambda}\sum_{m=1}^{\frac{M}{K}}\frac{1}{m} + \frac{1}{\mu}(\sum_{i=1}^{\frac{M}{K}-1} i\ p_{\frac{M}{K},i} + 1)]$.

As we can see, by controlling the number of zones, the system manager can control the distributions of inter-arrival times and therefore the corresponding system performance. Two extreme cases are worth highlighting. The first is when $K = M$; in this case, the expected inter-arrival times are constant. The second is when $K = 1$; in that case, the expected inter-arrival times are strictly increasing. In between, the expected inter-arrival times exhibit a cyclical pattern of being strictly increasing within a cycle (a zone) and having a step decrease between cycles (the start of boarding of each zone). Fewer zones reduce makespan while more zones reduce waiting time. The system manager would typically want to balance the costs associated with these two measures; customers prefer to wait less while boarding (and there is an implied delay cost) while the airline would like to reduce the total boarding time (and there is an implied resource usage cost). There is of course indirect waiting time related to customers waiting for their zones to be called, but the cost of that waiting is lower since customers are less inconvenienced in that case than when they are waiting to board.

In Figure 6, we present numerical results for an example system with 120 passengers. The solid line represents the expected waiting times of an arbitrary customer, and the dashed line represents the expected makespan of the boarding process. It is interesting to note the diminishing value of having more zones. An initial increase in the number of zones significantly reduces expected waiting time while further increases lead to only marginal further reduction. Given that the increase in makespan due to more zones does not exhibit a similar diminishing

effect, the optimal number of zones would generally be relatively small.

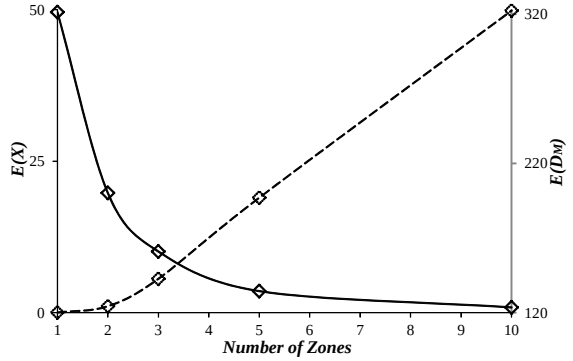


Figure 6: Impact of the Number of Zones on Expected Waiting Time and Makespan ($M = 120$, $\lambda = 0.1$, $\mu = 1$)

It is worth to note that results from the above examples, as well as those from the previous sections, show that in general, inter-arrival or service time features that reduce waiting time do not reduce makespan (in fact, the reverse is typically true). Thus, there is a need to trade off the benefit of lower waiting time against shorter makespan, in making decisions about which features to induce.

There are other related settings where arrivals exhibit features that are similar to the ones observed in the flight boarding problem. As mentioned in the introduction, this can be the case when the arrival of customers is triggered by the start of an event (e.g., the arrival of passengers to check-in for a flight or the arrivals of fans to a concert), and customers may belong to different classes that are differentiated by their risk attitudes toward being late for the event (with some classes preferring to arrive earlier than others). The arrival of customers within the same class can be modeled as a pure death process, which again leads to increasing mean inter-arrival times. Although controlling the number of customers within each class is more difficult in this case than in the flight boarding case, it may be possible, with sufficient incentives, to induce customers to arrive earlier or later. More importantly, recognizing the heterogeneity in inter-arrival times allows the system manager to plan for the necessary capacity (e.g., to meet target service levels as discussed in Section 5.3).

We conclude this section by noting that the insights provided so far also apply to settings where arrivals can be controlled in a more direct way, such as when arrivals to a particular process can be specified. This is the case, as we mentioned in the introduction, when arrivals are determined by appointment times. Assuming customers are punctual, inter-arrival times would be deterministic and would correspond to the time between appointment times. Depending on how appointments are scheduled, inter-arrival times may exhibit different features. For example,

scheduling more (fewer) appointments early on and then progressively fewer (more) leads to increasing (decreasing) inter-arrival times. Scheduling appointments differently could lead to inter-arrival times that exhibit combinations of both the increasing and decreasing features.

To evaluate the impact of different arrival and service time features, we carried out extensive experiments similar to those in Section 5 (for the sake of brevity, we omit the details). The results obtained are qualitatively consistent with those described there. Hence, our observations also provide insights into desirable features of appointment schedules for such settings. We note that some of these are consistent with the results from the appointment scheduling literature. For example, we observe that arrival processes with the "Increasing/Decreasing" inter-arrival time feature, although not always performing the best, do perform relatively well for all the performance measures considered. This "Increasing/Decreasing" feature is consistent with the "dome-shaped" appointment schedule shown in Kaandorp and Koole (2007) to perform well when the performance measure is a weighted cost of waiting time, idle time, and tardiness.

8 Concluding Comments

The results of this paper highlight the importance of accounting for the heterogeneity in customer inter-arrival and service times, when the number of customers is finite and customer inter-arrival or service times depend on their positions in the arrival sequence. This heterogeneity arises naturally in many service systems, but could also be engineered into how these systems are designed and managed. Accounting for this heterogeneity is important because different inter-arrival and service time features, even if resulting in the same total workload for the system, can lead to different levels of performance.

There are several possible avenues for future research. It would be useful to generalize our results to a broader class of systems (including queueing networks, systems with general service time distributions, and systems with customer priorities), and to investigate additional applications where systems with the type of features we studied arise naturally. It would also be interesting to study systems with other types of arrival processes such as those with time-dependent arrival rates. Moreover, it would be useful to explore other types of approximations (e.g., diffusion approximations). Finally, it would be meaningful to revisit principles that have been shown to be effective in the design and operation of service systems under steady state assumptions, and to determine whether or not they continue to be effective in systems with finite arrivals and heterogeneous inter-arrival and service times. One such principle is the benefit of pooling of servers and queues in systems with multiple servers.

Acknowledgments

The authors are grateful to Steve Graves, an anonymous associate editor, and three anonymous reviewers for their many constructive comments and suggestions.

References

Cayirli, T., Veral, E. (2003). Outpatient Scheduling in Health Care: A Review of Literature. *Production and Operations Management*. 12(4):519-549.

Chen, G., Shen, Z.M. (2007). Probabilistic Asymptotic Analysis of Stochastic Online Scheduling Problems. *IIE Transactions*. 39(5):525-538.

Chou, M.C., Liu, H., Queyranne, M., Simchi-Levi, D. (2006). On the Asymptotic Optimality of a Simple On-Line Algorithm for the Stochastic Single-Machine Weighted Completion Time Problem and Its Extensions. *Operations Research*. 54(3):464-474.

Courtois, P.J., Georges, J. (1971). On a Single-Server Finite Queuing Model with State-Dependent Arrival and Service Processes. *Operations Research*. 19(2):424-435.

Emmons, H., Vairaktarakis. G. (2013). *Flow Shop Scheduling: Theoretical Results, Algorithms, and Applications*. Springer.

Green, L., Kolesar P., Svoronos, A. (1991) Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems. *Operations Research*. 39(3):502-511.

Griffiths, J.D., Leonenko, G.M., Williams, J.E. (2006) The Transient Solution to $M/E_k/1$ Queue. *Operations Research Letters*. 34(3):349-354.

Gupta, D., Denton, B. (2008). Appointment Scheduling in Health Care: Challenges and Opportunities. *IIE Transactions*. 40(9):800-819.

Hall, R.W. (1991). *Queueing Methods: For Services and Manufacturing*. Prentice Hall.

Haque, L., Armstrong, M.J. (2007). A Survey of the Machine Interference Problem. *European Journal of Operational Research*. 179(2):469-482.

Hassin, R., Mendel, S. (2008). Scheduling Arrivals to Queues: A Single-Server Model with No-Shows. *Management Science*. 54(3):565-572.

Hu, B., Benjaafar, S. (2009). Partitioning of Servers in Queueing Systems During Rush Hour. *Manufacturing & Service Operations Management*. 11(3):416-428.

Jouini, O., Wang, R., Benjaafar, S. (2014). Queueing Systems with Appointment-Driven Arrivals, Non-Punctual Customers, and No-Shows. *Working Paper, University of Minnesota*.

Kaandorp, G.C., Koole, G. (2007). Optimal Outpatient Appointment Scheduling. *Health Care Management Science*. 10(3):217-229.

Kelton, W.D., Law, A.M. (1985). The Transient Behavior of the $M/M/s$ Queue, with Implications for the Steady-State Simulation. *Operations Research*. 33(2):378-396.

Kleinrock, L. (1975). *Queueing Systems, Volume 1: Theory*. Wiley-Interscience.

Koeleman, P.M., Koole, G.M. (2012). Optimal Outpatient Appointment Scheduling with Emergency Arrivals and General Service Times. *IIE Transactions on Healthcare Systems Engineering*. 2(1):14-30.

Mondschein, S.V., Weintraub, G.Y. (2003). Appointment Policies in Service Operations: A Critical Analysis of the Economic Framework. *Production and Operations Management*. 12(2):266-286.

Ouelhadj, D., Petrovic, S. (2009). A Survey of Dynamic Scheduling in Manufacturing Systems. *Journal of Scheduling*. 12(4):417-431.

Parlar, M., Moosa, S. (2008). Dynamic Allocation of Airline Check-In Counters: A Queueing Optimization Approach. *Management Science*. 54(8):1410-1424.

Parthasarathy, P.R., Moosa, S. (1989). Transient Solution to the Many-Server Poisson Queue: A Simple Approach. *Journal of Applied Probability*. 26(3):584-594.

Pinedo, M.L. (2012). *Scheduling: Theory, Algorithms, and Systems*. Springer.

Preater, J. (2001). A Bibliography of Queues in Health and Medicine. *Keele Mathematics Research Report, Keele University*.

Robinson, L.W., Chen, R.R. (2010). A Comparison of Traditional and Open-Access Policies for Appointment Scheduling. *Manufacturing & Service Operations Management*. 12(2):330-346.

Ross, S.M. (1978). Average Delay in Queues with Non-Stationary Poisson Arrivals. *Journal of Applied Probability*. 15(3):602-609.

Ross, S.M. (2009). *Introduction to Probability Models*. Academic Press.

Sztrik, J. (2005). Finite-Source Queueing Systems and Their Applications. Ferenczi, M., Pataricza, A., Ronyai, L. (Editors) *Formal Methods in Computing*. Akademia Kiado.

Takagi, H. (1993). *Queueing Analysis: A Foundation of Performance Evaluation, Volume 2: Finite Systems*. North-Holland.

Zeng, B., Turkcan, A., Lin, J., Lawley, M. (2010). Clinic Scheduling Models with Overbooking for Patients with Heterogeneous No-Show Probabilities. *Annals of Operations Research*. 178(1):121-144.