International Journal of Bifurcation and Chaos, Vol. 27, No. 4 (2017) 1730016 (20 pages)

DOI: 10.1142/S0218127417300166

Queues with Choice via Delay Differential Equations

Jamol Pender
School of Operations Research and Information Engineering,
Cornell University, 228 Rhodes Hall, Ithaca, NY 14853, USA
jjp274@cornell.edu

Richard H. Rand
Sibley School of Mechanical and Aerospace Engineering,
Department of Mathematics, Cornell University,
535 Malott Hall, Ithaca, NY 14853, USA
rand@math.cornell.edu

Elizabeth Wesson
Department of Mathematics, Cornell University,
582 Malott Hall, Ithaca, NY 14853, USA
enw27@cornell.edu

Received December 12, 2016

Delay or queue length information has the potential to influence the decision of a customer to join a queue. Thus, it is imperative for managers of queueing systems to understand how the information that they provide will affect the performance of the system. To this end, we construct and analyze two two-dimensional deterministic fluid models that incorporate customer choice behavior based on delayed queue length information. In the first fluid model, customers join each queue according to a Multinomial Logit Model, however, the queue length information the customer receives is delayed by a constant Δ. We show that the delay can cause oscillations or asynchronous behavior in the model based on the value of Δ. In the second model, customers receive information about the queue length through a moving average of the queue length. Although it has been shown empirically that giving patients moving average information causes oscillations and asynchronous behavior to occur in U.S. hospitals, we analytically and mathematically show for the first time that the moving average fluid model can exhibit oscillations and determine their dependence on the moving average window. Thus, our analysis provides new insight on how operators of service systems should report queue length information to customers and how delayed information can produce unwanted system dynamics.

Keywords: Queueing theory; choice models; delay differential equations; oscillations; moving averages; healthcare; Disneyland.

1. Introduction

Understanding the impact of providing delay information to customers in service systems is a very important problem in the operations management literature. Smartphones and internet technology have changed the possibilities for communication between service systems and their potential customers. Currently, many companies and system managers choose to provide their customers with valuable information that has the potential to influence their choice of using the service. One example of this communication is delay announcements, which have become important tools for customers to know how long they will wait on average for someone to start serving them. These announcements are not only important because they give the customer information about the quality of the service, but also they have the possibility of influencing the possibility that a customer will return to use the service again. As a consequence, understanding the impact of providing delay or queue length information to customers on customer choices and system operations, as well as the development of methods to support such announcements, has attracted the attention of the Operations Research and Management communities in the past few years.

One example of this new communication between customers and services is in the healthcare industry. In Fig. 1, we show an example of a typical billboard sign that many hospitals use for marketing as well as a way of providing information to potential patients. Since emergency room waiting times can be very long, giving waiting time or queue length information to potential patients is a useful tool for hospitals to communicate to patients when their emergency rooms are relatively underloaded. Much of the current literature that explores the impact of giving customers information about queue lengths and waiting times has been applied in the context of telecommunication systems such as telephone call centers. However, understanding the impact in a healthcare context is much more complicated. For one, in healthcare, the service discipline is not necessarily first come first serve and can be quite arbitrary. Given the triage system that is prevalent in hospitals, one could be tempted to model the emergency room with a priority queue. However, understanding the impact of waiting times and queue lengths on the dynamics in the priority setting is also quite difficult, see for example [Pender, 2017]. Moreover, unlike the call center literature where callers are likely to only speak with one agent, the patient experience often involves multiple servers that each have a different purpose in the service process of the patient. Thus, the patient experience is more like movement through a queueing network. For example, in a typical emergency room, a patient might interact with a nurse, a doctor, various administrative staff, and even laboratory technicians when tests need to be performed.

Most of the current research on providing queue length or waiting time information to customer focuses on the impact of delay announcements with respect to call centers and telecommunications applications. There is a vast literature on this subject, which is mostly segmented into three different areas of research. The first part emphasizes making accurate real-time delay announcements to customers. In fact, in [Ibrahim & Whitt, 2008, 2009, 2011a, 2011b] the authors develop new estimators for estimating delays in various queueing systems. They primarily study two types of estimators. The first estimator is the head of the line (HOL) estimator, which provides the current amount of time that the next customer to get service has waited in line. The second type of estimator that they study is the delay of the last customer who entered the agents service (LES). These two estimators are different and the papers [Whitt, 1999b; Ibrahim & Whitt, 2008, 2009, 2011a, 2011b] provide a detailed

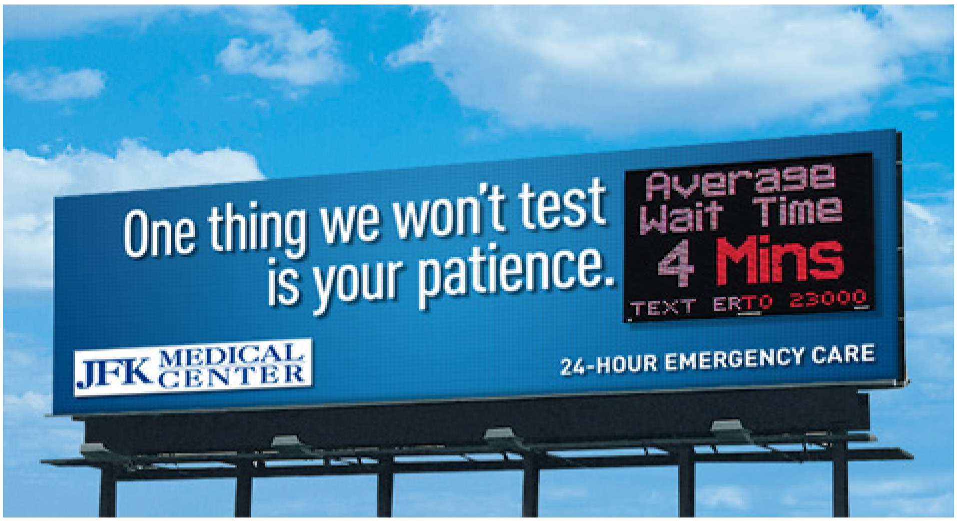


Fig. 1. Highway signs posting emergency room wait times.

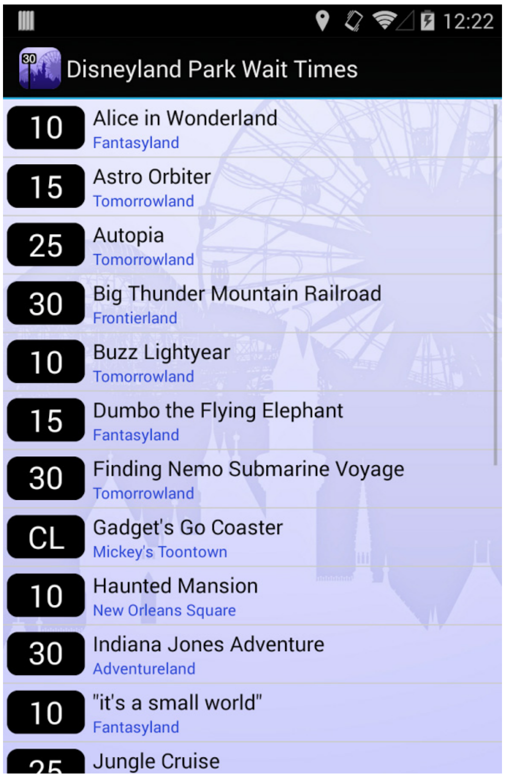
analysis of these estimators. The second part of the literature addresses how the delay announcements impact the dynamics of the queueing process and how customers respond to the announcements. The works of Armony and Maglaras [2004], Guo and Zipkin [2007], Hassin [2007], Armony *et al.* [2009], Guo and Zipkin [2009], Jouini *et al.* [2009], Jouini *et al.* [2011], Allon and Bassamboo [2011], Allon *et al.* [2011], Ibrahim *et al.* [2016], Whitt [1999a] and references therein analyze the impact of delay announcements on the queueing process and the abandonment process of the system. Finally, the third part of the literature analyzes the customer psychology of waiting. The works of Hui and Tse [1996], Hui *et al.* [1997], Pruyn and Smidts [1998], Munichor and Rafaeli [2007], Sarel and Marmorstein [1998], Taylor [1994] explore the behavioral aspect of customer waiting and how delays affect customer decisions. This paper is most related to the second area of research; however, it is unique in that it includes customer choice with delay differential equations.

More recently, there is work that also considers how information can impact queueing systems. Work by Jennings and Pender [2016] compares ticket queues with standard queues. In a ticket queue, the manager is unaware of when a customer abandons and is only notified of the abandonment when the customer would have entered service. This artificially inflates the queue length process and the work of Jennings and Pender [2016] compares the difference in queue length between the standard and ticket queues. Follow-up work by Pender [2015b, 2015c] also considers the case when there are dependencies between balking and reneging customers and when the server spends time clearing a customer who has abandoned the system respectively. However, this work does not consider the aspect of choice and delays in providing the information to customers, which is the case in many healthcare settings.

Since hospital networks are more complicated than telephone call centers, understanding the waiting and queueing dynamics is a much harder problem, see for example [Armony *et al.*, 2015]. Even designing a delay estimator in hospital systems is very difficult and recent work by Plambeck *et al.* [2014] develops a new emergency department delay estimator that combines methodology from statistical learning theory and queueing theory. Because of this complexity, it is common that hospitals publish historic average waiting times using a 4-hour moving average and this has been noticed in the work of Dong *et al.* [2015].

Another useful application of our work is for amusement parks like Disneyland or Six Flags. In Fig. 2, we show a snapshot of the Disneyland app. The Disneyland app lists waiting times of various rides in the themepark and customers get to choose which ride that they would want to go on given the waiting times. However, the wait times on the app are not posted in real-time and are calculated based on moving average of the waiting times. Thus, our queueing analysis is useful for Disney to synchronize their waiting times for rides across the themepark.

This paper introduces two new fluid models, which describe the dynamics of customer choice and delay information that customers use to make decisions. In the first fluid model, the customer receives information about the queue length which is delayed by a parameter Δ. In the second fluid model, we use a moving average of the queue length over the time interval Δ to represent the queue length information given to the customer. The models that we present are useful in two major contexts. The first context is where the software that communicates



12:22

Disneyland Park Wait Times

10	Alice in Wonderland	Fantasyland
15	Astro Orbiter	Tomorrowland
25	Autopia	Tomorrowland
30	Big Thunder Mountain Railroad	Frontierland
10	Buzz Lightyear	Tomorrowland
15	Dumbo the Flying Elephant	Fantasyland
30	Finding Nemo Submarine Voyage	Tomorrowland
CL	Gadget's Go Coaster	Mickey's Toontown
10	Haunted Mansion	New Orleans Square
30	Indiana Jones Adventure	Adventureland
10	"it's a small world"	Fantasyland
25	Jungle Cruise	

Fig. 2. Disneyland Park wait times app.

with customers is delayed in some fashion, which is common in many hospitals who outsource the computation of their waiting times and queue lengths. The second context is where the customer reaction to the information is delayed. This can happen in the Disney example where there is a delay between customers viewing the wait times and them joining the queue. Thus, the delay does not necessarily need to be a function of the software or a lag in information, it can be caused by the customer behavior and distance from the queue that they are joining. What these fluid models are able to show is that when the delay is small the two queues are balanced and synchronized; however, when the delay is large enough, the two queues are not balanced and asynchronous. We determine the exact threshold where the dynamics of the two queues are different for both the constant and moving average models. Our analysis combines theory from delay differential equations, customer choice models, and stability analysis of differential equations.

1.1. *Main contributions of paper*

The contributions of this work can be summarized as follows:

- We develop two new two-dimensional fluid models that incorporate customer choice based on delayed queue length information. One model uses a constant delay and another model uses a moving average.
- We show that the constant delay queueing model can experience oscillations where the two queues are not synchronized and derive the exact threshold where the oscillatory behavior is triggered in terms of the model parameters. Moreover, we show that the threshold is monotone in terms of the arrival rate.
- We show that the moving average queueing model can experience oscillations where the two queues are not synchronized. However, unlike the constant delay system, the threshold is not monotone as a function of the arrival rate.

1.2. *Organization of paper*

The remainder of this paper is organized as follows. Section 2 describes a constant delay fluid model. We derive the critical delay threshold under which the queues are balanced if the delay is below the threshold and the queues are asynchronized if the delay is above the threshold. We also show that the instability is preserved as long as the delay is increased. Section 3 describes a constant moving average delay fluid model. We derive the critical delay threshold under which the queues are balanced if the delay is below the threshold and the queues are asynchronized if the delay is above the threshold. We also show that the instability is preserved as long as the delay is increased in certain regions of the parameter space. Finally in Sec. 4, we conclude with directions for future research related to this work.

2. Constant Delay Fluid Model

In this section, we present a new fluid model with customer choice based on the queue length with a constant delay. Thus, we begin with two infinite-server queues operating in parallel, where customers choose which queue to join by taking the size of the queue length into account. However, we add the twist that the queue length information that is reported to the customer is delayed by a constant Δ. Therefore, the queue length that the customer receives is actually the queue length Δ time units in the past. An example of this delay is given in Fig. 3, which is JFK Medical Center in Boynton Beach, Florida. In Fig. 3, the average wait time is reported to be 12 minutes. However, on the top right of the figure we see that the time of the snapshot was 4:04 pm while the time of a 12 minute wait is as of 3:44 pm. Thus, there is a delay of 20 minutes in the reporting of the wait times in the emergency room and this can have an important impact on the system dynamics as we will show in the sequel.

The choice model that we use to model these dynamics is identical to that of a Multinomial Logit Model (MNL) where the utility for being served in the ith queue with delayed queue length $Q_i(t-\Delta)$ is $u_i(Q_i(t-\Delta)) = Q_i(t-\Delta)$. Thus, in a stochastic context with two queues, the probability of going to the first queue is given by the following expression

$$p_1(Q_1(t), Q_2(t), \Delta) = \frac{\exp(-Q_1(t-\Delta))}{\exp(-Q_1(t-\Delta)) + \exp(-Q_2(t-\Delta))} \tag{1}$$

and the probability of going to the second queue is

$$p_2(Q_1(t), Q_2(t), \Delta) = \frac{\exp(-Q_2(t-\Delta))}{\exp(-Q_1(t-\Delta)) + \exp(-Q_2(t-\Delta))}. \tag{2}$$

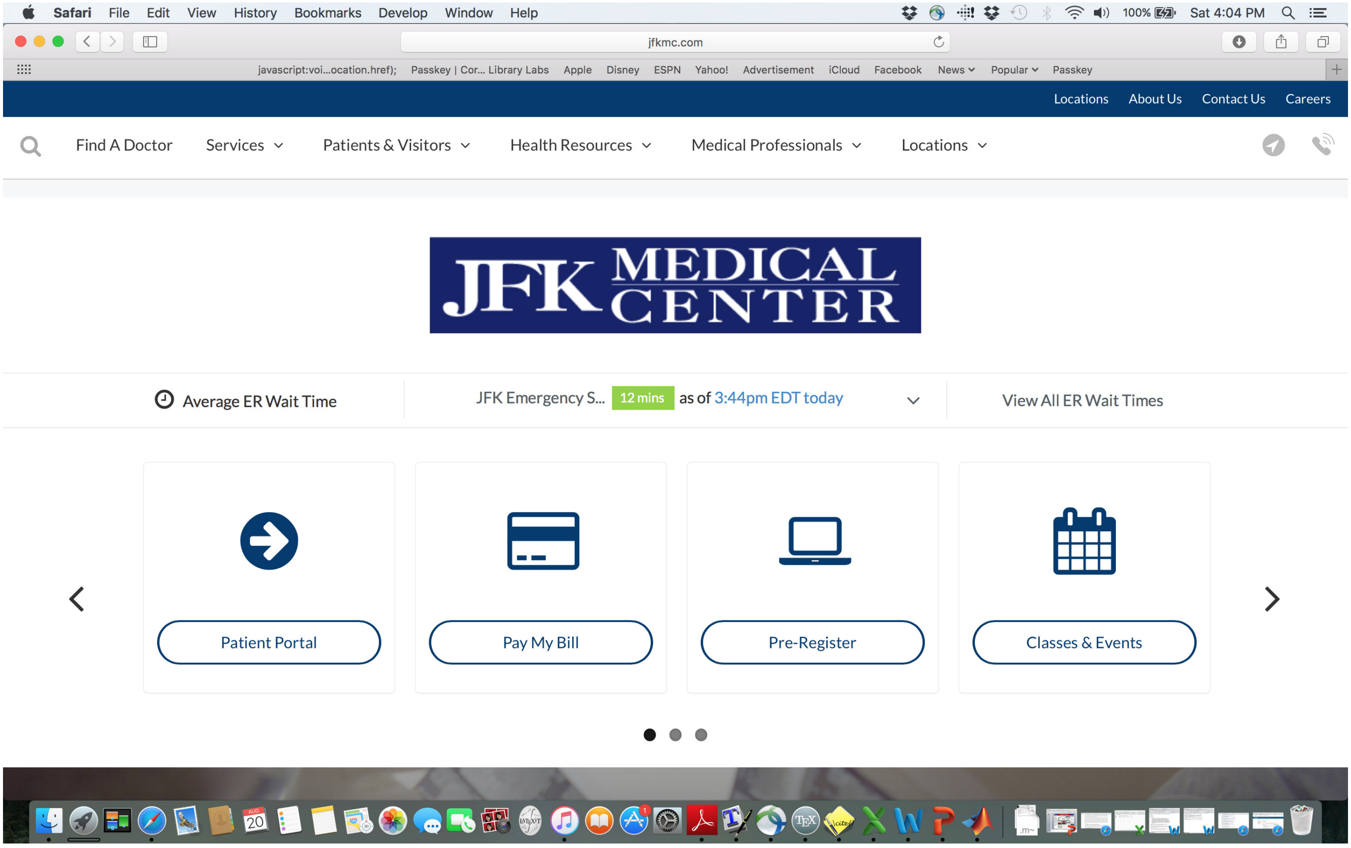

Safari File Edit View History Bookmarks Develop Window Help
100% Sat 4:04 PM
jfkmc.com
javascript:voi...ocation.href); Passkey | Cor... Library Labs Apple Disney ESPN Yahoo! Advertisement iCloud Facebook News Popular Passkey
Locations About Us Contact Us Careers
Find A Doctor
Services
Patients & Visitors
Health Resources
Medical Professionals
Locations
JFK MEDICAL CENTER
Average ER Wait Time
JFK Emergency S... 12 mins as of 3:44pm EDT today
View All ER Wait Times
Patient Portal
Pay My Bill
Pre-Register
Classes & Events

Fig. 3. JFK Medical Center online reporting.

Since the main goal of our analysis is to provide insight into the dynamics of the system when delayed information is given to customers, we analyze a fluid model of the system instead of the actual stochastic process, which is more difficult. Moreover, the fluid model enables us to understand the mean dynamics of the system when the number of arrivals in the system is large, which is the case in themeparks like Disneyland. However, since we analyze the fluid model instead of the real stochastic system, we no longer have probabilities in our choice model. Instead, we now have rates at which customers join each of the two queues. In our fluid model, we assume that the sum of the arrival rate of customers to both queues is equal to the constant rate λ, the service rate of both queues is equal to μ, and the information about the queue length is delayed by the constant Δ. Therefore, in this model, customers join the first queue, $q_1(t)$, at rate

$$\lambda \cdot \frac{\exp(-q_1(t-\Delta))}{\exp(-q_1(t-\Delta)) + \exp(-q_2(t-\Delta))} \tag{3}$$

and customers join the second queue, $q_2(t)$, at rate

$$\lambda \cdot \frac{\exp(-q_2(t-\Delta))}{\exp(-q_1(t-\Delta)) + \exp(-q_2(t-\Delta))}. \tag{4}$$

Thus, our model for customer choice infinite server queues with delayed information can be represented by the two-dimensional system of delay differential equations

$$\begin{aligned} \dot{q}_1(t) &= \lambda \cdot \frac{\exp(-q_1(t-\Delta))}{\exp(-q_1(t-\Delta)) + \exp(-q_2(t-\Delta))} \\ &\quad - \mu q_1(t), \qquad (5) \\ \dot{q}_2(t) &= \lambda \cdot \frac{\exp(-q_2(t-\Delta))}{\exp(-q_1(t-\Delta)) + \exp(-q_2(t-\Delta))} \\ &\quad - \mu q_2(t), \qquad (6) \end{aligned}$$

where we assume that $q_1(t)$ and $q_2(t)$ start with different initial functions $\varphi_1(t)$ and $\varphi_2(t)$ on the interval $[-\Delta, 0]$.

Remark 2.1. When the two delay differential equations are started with the same initial functions,

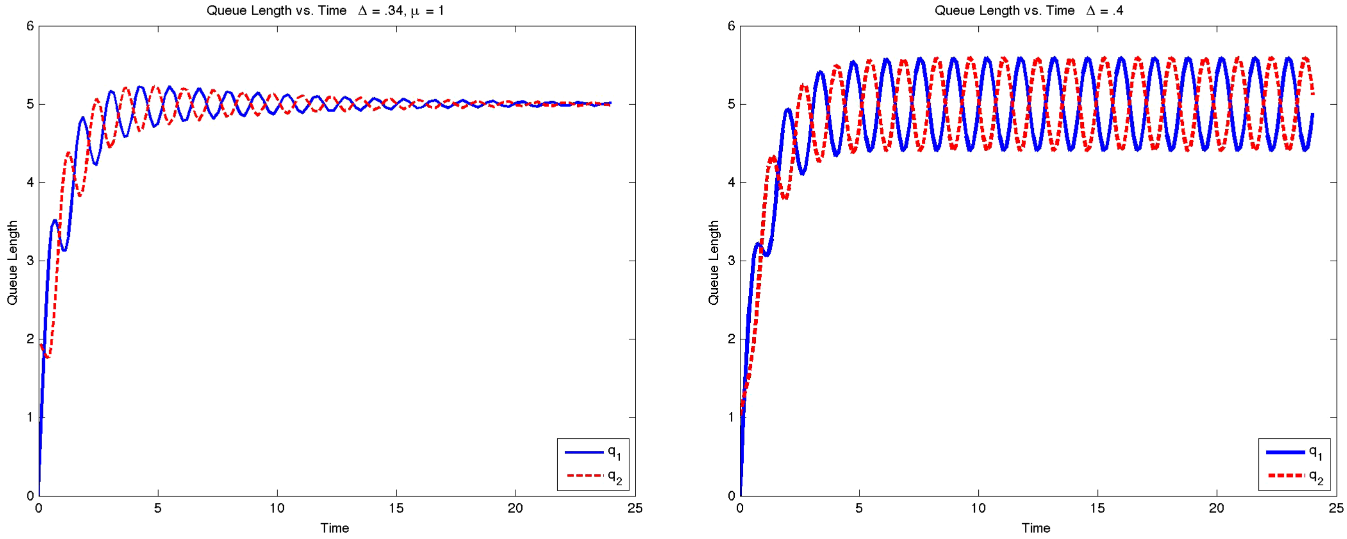


Fig. 4. Numerical integration of fluid model for $\lambda = 10$, $\mu = 1$, $\Delta_{\rm cr} = 0.3614$: (left) $\Delta = 0.34$ and (right) $\Delta = 0.4$.

they are identical for all time because of the symmetry of the problem. Therefore, we will start the system with nonidentical initial conditions so the problem is no longer trivial and the dynamics are not identical.

In Fig. 4, we see qualitatively different behavior of the fluid model equations when $\Delta = 0.34$ and $\Delta = 0.4$. It is clear from the left plot in Fig. 4 that the two queues are synchronized and converge to the same equilibrium solution when $\Delta = 0$. However, on the right plot of Fig. 4, we see that the two queues are not synchronized and exhibit oscillatory and asynchronous behavior. It turns out that for the model parameters presented in Fig. 4, asynchronous dynamics will not occur as in right plot of Fig. 4 if the delay $\Delta < 0.3614$. Otherwise, oscillations and asynchronous dynamics will exist for both queues. However, this change in behavior can be explained by the fact that the equilibrium points of the queue length delay differential equations transition from stable to unstable where a limit cycle is born. This situation is known as a Hopf bifurcation and will be explained in more detail later in the paper, however, the reader is referred to [Guckenheimer & Holmes, 2013] to learn more about Hopf bifurcations and the general analysis of ordinary differential equations. In the next theorem, we show how to derive the critical delay $\Delta_{\rm cr}(\lambda, \mu)$, which depends on the arrival rate λ and the service rate μ. The critical delay $\Delta_{\rm cr}(\lambda, \mu)$ separates the region of oscillatory and nonoscillatory dynamics of the queueing model and can be determined by a stability analysis of the delay differential equations given in Eqs. (5) and (6). Moreover, to the left of the critical delay, the real part of all eigenvalues of the linearized equations is negative and to the right of the critical delay, at least one of the eigenvalues is positive. Our first result, given below, determines the critical delay's dependence on the arrival and service rates of our queueing fluid model with choice.

Theorem 1. *For the constant delay choice queueing model given in Eqs.* (5) *and* (6), *the critical delay* $\Delta_{\rm cr}(\lambda, \mu)$ *is given by the following expression*

$$\Delta_{\rm cr}(\lambda, \mu) = \frac{2 \arccos\left(-\dfrac{2\mu}{\lambda}\right)}{\sqrt{\lambda^2 - 4\mu^2}}. \tag{7}$$

Moreover, if $\Delta < \Delta_{\rm cr}$, *then the two queues will be synchronized in equilibrium and when* $\Delta \geq \Delta_{\rm cr}$ *the two queues will be asynchronous and oscillate in equilibrium.*

Proof. See the Appendix for the proof. ■

In Fig. 5, we show the dependence of the critical delay $\Delta_{\rm cr}$ as a function of the model parameter λ while keeping μ constant. On the left of Fig. 5, we plot Hopf curves when $\mu = 0.5$ and $\mu = 1$. We see that the curves are very similar for both values of μ and that the critical delay value decreases as λ increases. Moreover, on the right of Fig. 5, we use the function in Matlab called **EZ-Plot**, to plot the various critical delay values. We see that there

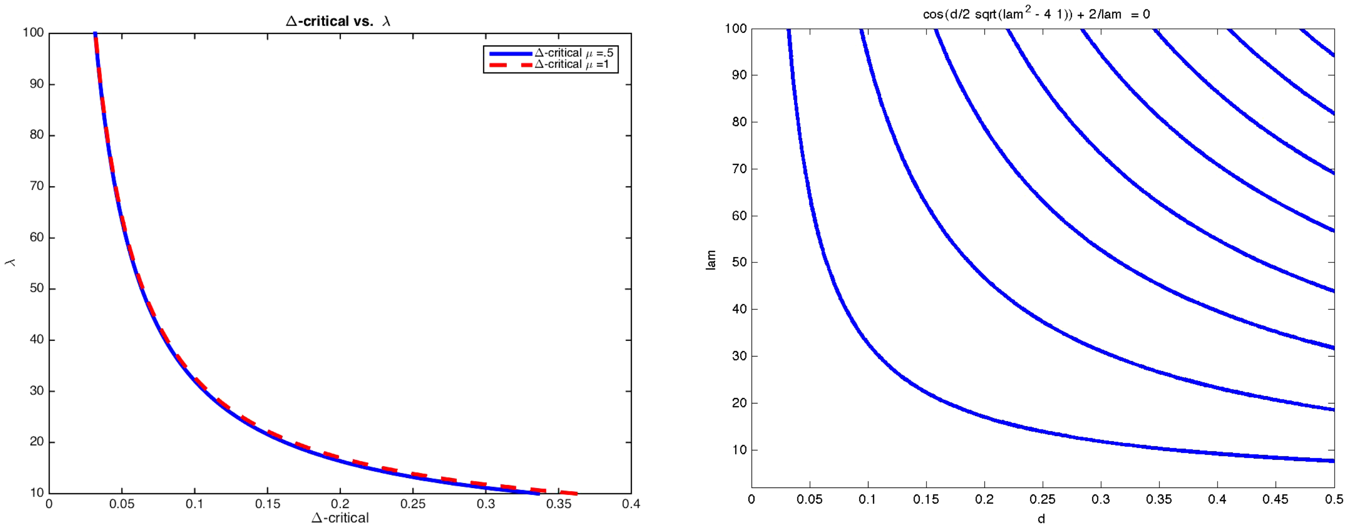


Fig. 5. (Left) Δ_{cr} as a function of λ when $\mu = 0.5$ and $\mu = 1$ and (right) EZ-plot of Δ_{cr} as a function of λ when $\mu = 1$.

are several Hopf bifurcation curves, which indicate that there could possibly be several regions where the stability of the delay differential equation system might change, i.e. the example of Fig. 4 could be reversed. However, we will show in the sequel that this reversal of stability is impossible since all roots pass from the left half-plane to right half-plane, which prevents the system from becoming stable again.

2.1. *Hopf bifurcation curves in the constant delay model*

On the right of Fig. 5, as the first Hopf curve is crossed, we see a stable limit cycle is born. However, after the next Hopf curve is crossed (as parameters are slowly changed), there are various possibilities:

(1) The pair of roots which crossed into the right half-plane in the first Hopf, may cross back into the left half-plane. Thus, all roots are in the left half-plane again and the equilibrium reverts to stability. The limit cycle which was born in the first Hopf bifurcation shrinks to nothing and disappears.
(2) Another pair of roots may cross into the right half-plane, so that two pairs of roots are now in the left half-plane. A new limit cycle may be born, but the stability of the equilibrium does not change. The new limit cycle is expected to be unstable.
(3) One of many possible degenerate cases: multiple pairs may cross at once, or the first pair of roots may recross simultaneously with the next pair crossing, or a pair of roots may touch the imaginary axis but not cross.

Numerical integration with the Matlab delay differential equation package **dde23** showed that there was only one stable limit cycle observed, no matter how many of the Hopf curves are crossed. However, this does not tell us what happens to the various pairs of imaginary roots which occur on the various Hopf curves. Do they all pass from left half-plane to right half-plane, or do some of them come back in the opposite direction? The purpose of the remainder of this subsection is to address this issue.

Now suppose that the delay Δ is close to a critical value for a Hopf bifurcation. We then make a slight perturbation from the critical value for the Hopf bifurcation i.e.

$$\Delta = \Delta_0 + \epsilon\Delta_1 \tag{8}$$

where $\epsilon \ll 1$. Then the root r will be slightly perturbed from the pure imaginary value it would take at $\Delta = \Delta_0$. That is, we can write

$$r = i\omega + \epsilon(ir_1 + r_2) \tag{9}$$

where r_1 and r_2, the imaginary and real parts of the perturbation, may be determined in terms of Δ_0 and Δ_1.

Proposition 1. *Suppose that we make a slight perturbation on the order of $\epsilon\Delta_1$ near a critical delay*

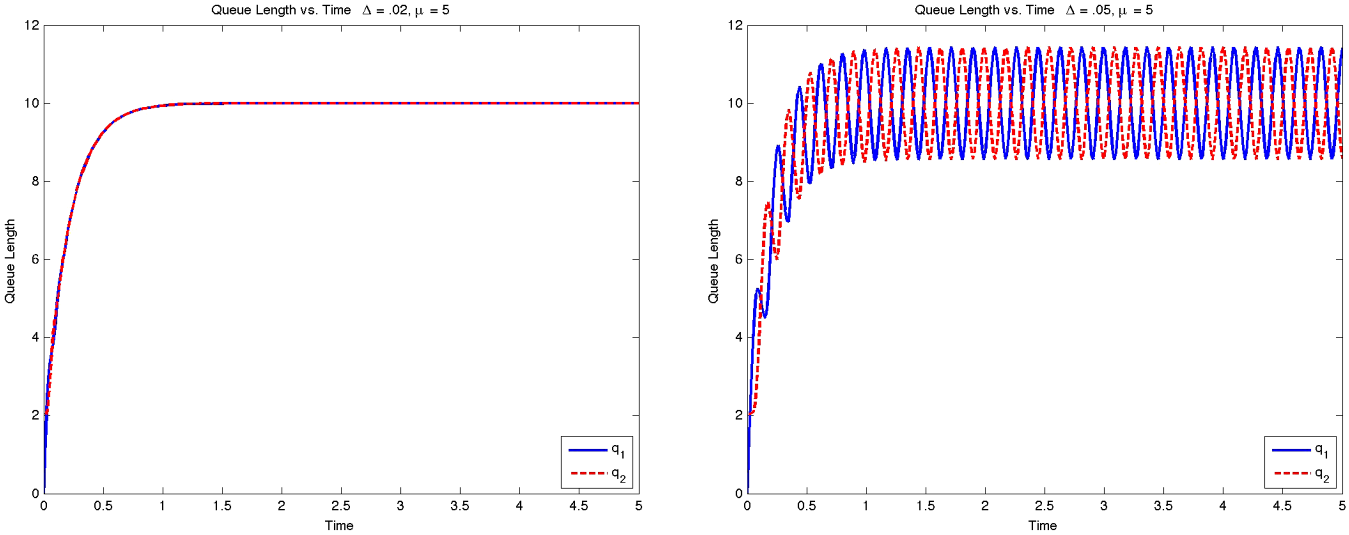


Fig. 6. Numerical integration of fluid model for $\lambda = 100$, $\mu = 5$, (left) $\Delta = 0.02$ and (right) $\Delta = 0.05$.

value for a Hopf bifurcation, then the real part of r is equal to

$$r_2 = \frac{4\omega^2\Delta_1}{8\Delta_0\mu + \Delta_0^2\lambda^2 + 4}. \tag{10}$$

In particular, r_2 has the same sign as Δ_1.

Proof. See Appendix. ■

Thus r_2 has the same sign as Δ_1. This means that as Δ increases past Δ_0, the critical Hopf value, r crosses the imaginary axis from left to right. This analysis holds for every Hopf curve, which implies that the initial instability that occurs after passing through the first Hopf curve is preserved and does not change as we pass through more Hopf curves. In Fig. 6, we provide an additional numerical example to illustrate the change in stability before and after our critical delay Δ_{cr}. Once again, we see that for all values of the delay before the critical delay threshold, the two queues synchronize, balance is achieved, and the system is stable near the equilibrium point. However, for all values after the critical delay threshold, the two queues exhibit asynchronous behavior and are not stable near the equilibrium point. However, not all real systems use a constant delay to report to their customers about the queue length or waiting time. It has been observed in [Dong *et al.*, 2015] that some service systems such as hospitals use a moving average. Thus, a moving average fluid model will be analyzed in the subsequent section.

3. Moving Average Delay Fluid Model

In this section, we present another fluid model with customer choice and where the delay information presented to the customer is a moving average. This model assumes that customers are informed about the queue length, but in the form of a moving average of the queue length between the current time and Δ time units in the past. Like in the previous model, customers also have the choice to join two parallel infinite server queues and they join according to the same multinomial logit model. We also assume that the total rate at which customers show up to the system is given by the parameter λ which is a constant. However, unlike the previous model, we assume that customers join the first queue at rate

$$\lambda \cdot \frac{\exp\left(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_1(s)ds\right)}{\exp\left(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_1(s)ds\right) + \exp\left(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_2(s)ds\right)} \tag{11}$$

and join the second queue at rate

$$\lambda \cdot \frac{\exp\left(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_2(s)ds\right)}{\exp\left(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_1(s)ds\right) + \exp\left(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_2(s)ds\right)}. \tag{12}$$

Thus, our model for customer choice with delayed information in the form of a moving average can be represented by a two-dimensional system of functional differential equations

$$\dot{q}_1(t) = \lambda \cdot \frac{\exp\left(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_1(s)ds\right)}{\exp\left(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_1(s)ds\right) + \exp\left(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_2(s)ds\right)} - \mu q_1(t), \tag{13}$$

$$\dot{q}_2(t) = \lambda \cdot \frac{\exp\left(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_2(s)ds\right)}{\exp\left(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_1(s)ds\right) + \exp\left(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_2(s)ds\right)} - \mu q_2(t), \tag{14}$$

where we assume that q_1 and q_2 start at different initial functions $\varphi_1(t)$ and $\varphi_2(t)$ on the interval $[-\Delta, 0]$.

Remark 3.1. We should also mention that if we initialize the differential equations with the same initial conditions, then the moving average delay differential equations are identical for all time because of the symmetry of the problem. Starting the two queues with identical initial conditions places both queues on an invariant manifold from which it cannot escape. Therefore, we start the system with nonidentical initial conditions so the problem is no longer trivial and the two queues start off the invariant manifold.

On the onset this problem is seemingly more difficult than the constant delay setting since the ratio now depends on a moving average of the queue length during a delay period Δ. To simplify the notation, we find it useful to define the moving average of the ith queue over the time interval $[t-\Delta, t]$ as

$$m_i(t, \Delta) = \frac{1}{\Delta}\int_{t-\Delta}^{t} q_i(s)ds. \tag{15}$$

A key observation to make is that the moving average itself solves a delay differential equation. In fact, by differentiating Eq. (15) with respect to time, it can be shown that the moving average of the ith queue is the solution to the following delay differential equation

$$\dot{m}_i(t, \Delta) = \frac{1}{\Delta} \cdot (q_i(t) - q_i(t-\Delta)), \quad i \in \{1, 2\}. \tag{16}$$

Leveraging the above delay equation for the moving average, we can describe our moving average fluid model with the following four-dimensional system of delay differential equations

$$\dot{q}_1 = \lambda \cdot \frac{\exp(-m_1(t))}{\exp(-m_1(t)) + \exp(-m_2(t))} - \mu q_1(t), \tag{17}$$

$$\dot{q}_2 = \lambda \cdot \frac{\exp(-m_2(t))}{\exp(-m_1(t)) + \exp(-m_2(t))} - \mu q_2(t), \tag{18}$$

$$\dot{m}_1 = \frac{1}{\Delta} \cdot (q_1(t) - q_1(t-\Delta)), \tag{19}$$

$$\dot{m}_2 = \frac{1}{\Delta} \cdot (q_2(t) - q_2(t-\Delta)). \tag{20}$$

In Fig. 7, we plot two examples of the moving average delay differential equations. In the example on the left of Fig. 7, the differential equations converge to the equilibrium of $\lambda/(2\mu)$ when $\Delta = 0.02$ and the dynamics are stable. It also seems like the plot on the right of Fig. 7 also converges to the equilibrium and is also stable. However, it is not stable and requires even closer observation.

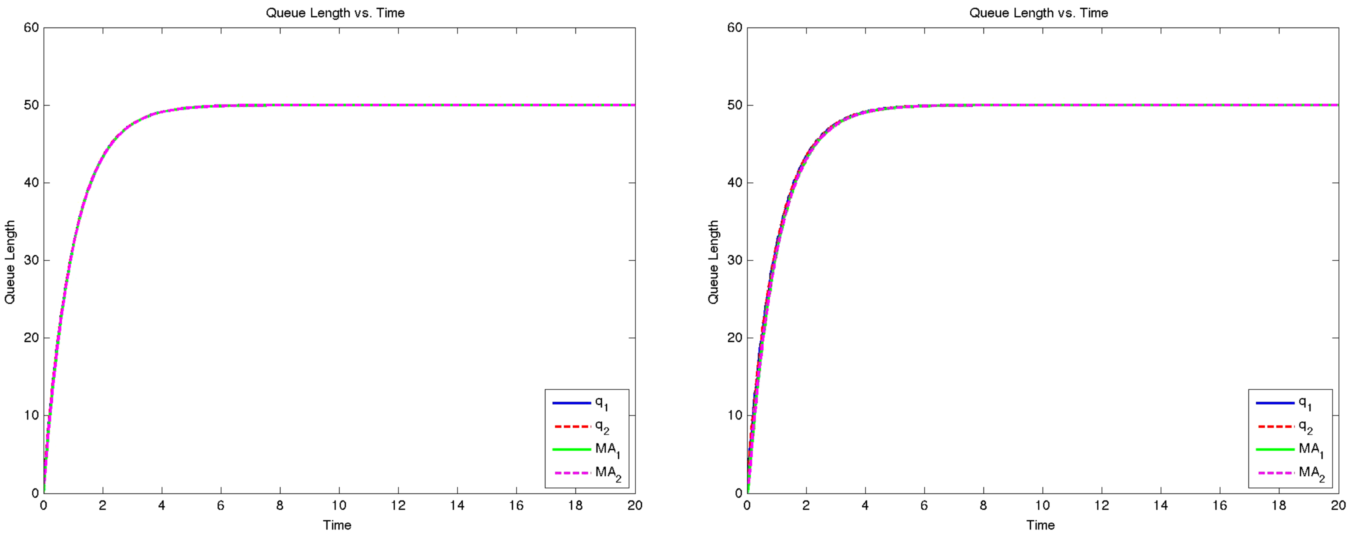


Fig. 7. $\lambda = 100$, $\mu = 1$, (left) $\Delta = 0.02$ and (right) $\Delta = 0.1$.

In Fig. 8, we zoom in and look at the dynamics of each example. On the left of Fig. 8, a closer look reveals that the differential equations are indeed stable. However, on the right when $\Delta = 0.1$, it is observed that the differential equations are not stable. Although the two queues appear to be stable in Fig. 7, the amplitude is too small to detect the asynchronous dynamics of the two queues. Thus, like in the previous fluid model, we need to understand the dynamics of the fluid model near the equilibrium to determine when the two queues will exhibit asynchronous behavior. The next theorem provides insight for understanding when the equilibrium behavior will be stable or unstable.

In Fig. 9, we plot the Hopf curves for the moving average model when $\mu = 1$. These curves are different from what would be plotted in the Matlab function **EZ-Plot**. One reason is that we square the cosine and sine functions, which introduces extraneous roots that do not exist. Thus, the plot given in Fig. 9 excludes these extraneous roots. Moreover, unlike the constant delay case, we also see a linear curve at the bottom of Fig. 9. This line represents where $\omega = 0$ and is another root of the equation.

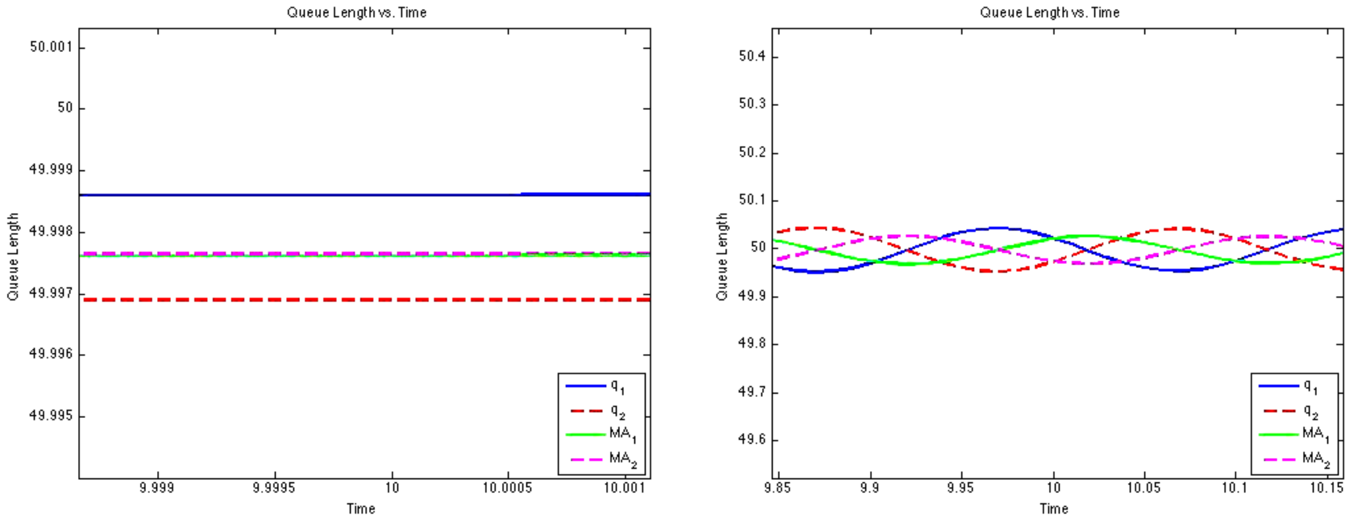


Fig. 8. Zoomed in. $\lambda = 100$, $\mu = 1$, (left) $\Delta = 0.02$ and (right) $\Delta = 0.1$.

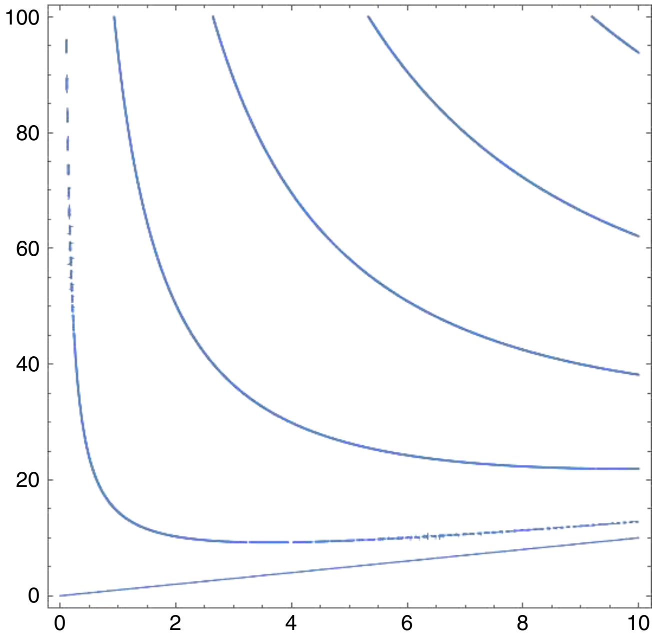


Fig. 9. Mathematica-plot of Hopf curves for moving average model ($\mu = 1$).

However, the stability is unchanged on this line, therefore it is not interesting to study.

Theorem 2. *For the moving average fluid model given by Eqs. (17)–(20), $\Delta_{\rm cr} = \Delta_{\rm cr}(\lambda, \mu)$ is given by the following transcendental equation*

$$\sin\left(\Delta_{\rm cr} \cdot \sqrt{\frac{\lambda}{\Delta_{\rm cr}} - \mu^2}\right) + \frac{2\mu\Delta_{\rm cr}}{\lambda} \cdot \sqrt{\frac{\lambda}{\Delta_{\rm cr}} - \mu^2} = 0. \tag{21}$$

Proof. See the Appendix for the proof. ■

In Fig. 10, we plot the dynamics for the moving average model when $\Delta = 2$ and when $\Delta = 4$. When $\Delta = 2$ we see that the dynamics are stable and the queues will converge to the equilibrium point. However, when $\Delta = 4$, we see that the dynamics are not stable and the queues are not synchronized. These dynamics are the same as in Fig. 11. Like in the constant delay example, we see that the stability of the queues is also given by the first Hopf curve. To the left and bottom of the Hopf curve, the two queues will eventually converge to their equilibrium values; however, to the right and above the Hopf curve, the two queues will be forever asynchronous.

3.1. *Hopf bifurcation curves in the moving average model*

Numerical integration with the Matlab delay differential equation package **dde23** showed that there was only one limit cycle observed in the moving average model, no matter how many of the Hopf curves are crossed. However, like in the constant delay setting, this does not tell us what happens to the various pairs of imaginary roots which occur on the various Hopf curves. We will show that as perturbation increases the size of the delay, the roots always pass from the left to the right, which implies that the equilibrium remains unstable as the delay increases forever.

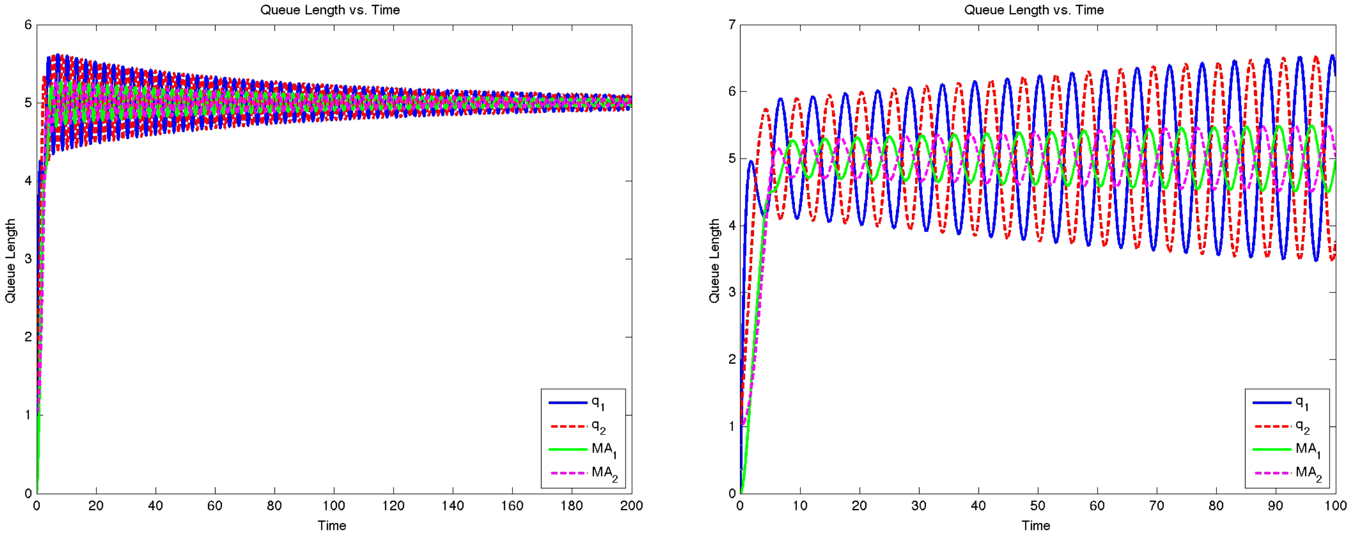


Fig. 10. $\lambda = 10$, $\mu = 1$, (left) $\Delta = 2$ and (right) $\Delta = 4$.

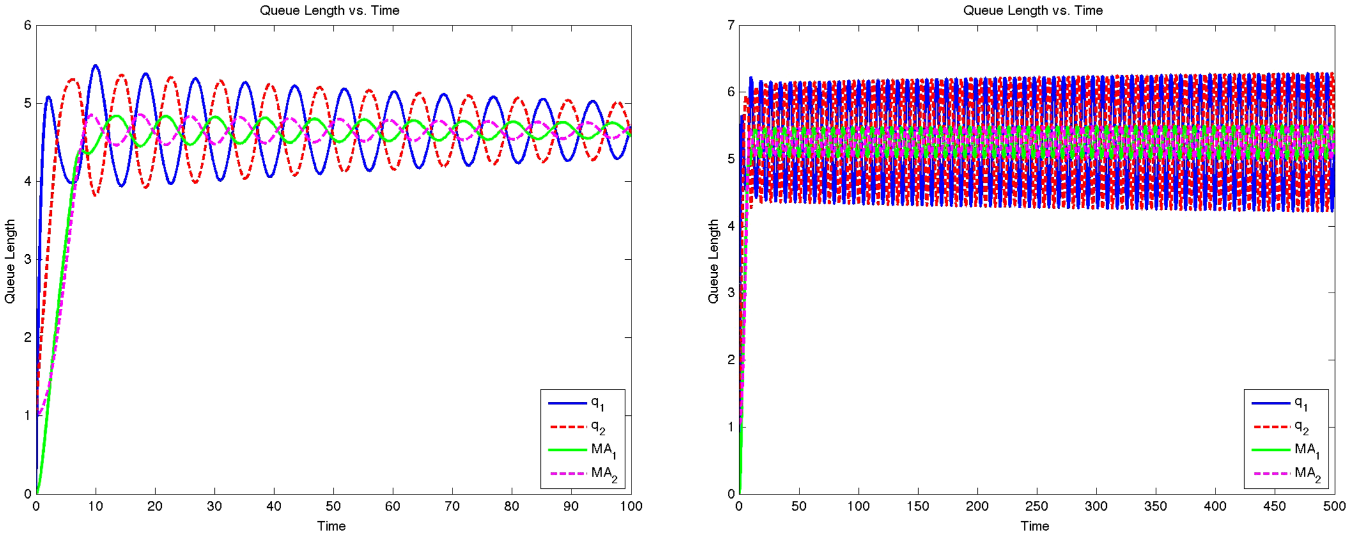


Fig. 11. (Left) $\lambda = 9.3$, $\Delta = 6.5$ and (right) $\lambda = 10.5$, $\Delta = 6.5$.

Now suppose that the delay Δ is close to a critical value for a Hopf bifurcation. We then make a slight perturbation from being near the critical value of the Hopf bifurcation i.e.

$$\Delta = \Delta_0 + \epsilon \Delta_1 \tag{22}$$

where $\epsilon \ll 1$. When the root r is close to the pure imaginary value, the critical delay is near the value $\Delta = \Delta_0$. Thus, when we make a slight perturbation the real and imaginary parts of the root will also be slightly perturbed i.e.

$$r = i\omega + \epsilon(ir_1 + r_2). \tag{23}$$

Proposition 2. *For the moving average model, suppose that we make a slight perturbation on the order of $\epsilon\Delta_1$ near a critical delay value for a Hopf bifurcation, then the real part of r is equal to*

$$r_2 = \frac{2\Delta_1\omega^2 \cdot (2\Delta_0\omega^2 - 2\mu\lambda)}{8\Delta_0^2\mu\omega^2 + 12\Delta_0\omega^2 + 4\Delta_0\lambda\mu + \Delta_0\lambda^2 + 4\lambda}. \tag{24}$$

Proof. In order to prove this, we can follow the same steps as in Proposition 1. When $\epsilon = 0$, we reduce back to the original critical threshold of Eq. (A.67). Now we substitute Eq. (A.38) into Eq. (A.61) and do a Taylor expansion for small values of ϵ. Then solve for r_1 and r_2. Solving for the real part of r, we find that r_2 is equal to the following value

$$r_2 = \frac{2\Delta_1\omega^2 \cdot (2\Delta_0\omega^2 - 2\mu\lambda)}{8\Delta_0^2\mu\omega^2 + 12\Delta_0\omega^2 + 4\Delta_0\lambda\mu + \Delta_0\lambda^2 + 4\lambda}. \tag{25}$$

■

In the moving average model, we see that the real part of the roots has the same sign as Δ_1 when $\Delta_0\omega^2 > \mu\lambda$ and has the opposite sign when $\Delta_0\omega^2 < \mu\lambda$. Therefore, when $\Delta_0\omega^2 > \mu\lambda$, the roots move from the left to the right and the stability is preserved. However, when $\Delta_0\omega^2 < \mu\lambda$ this may not be the case. This is caused by the fact that the Hopf curve is not monotone as a function of λ and can be seen in Fig. 9.

4. Conclusion and Future Research

In this paper, we analyze two new two-dimensional fluid models that incorporate customer choice and delayed queue length information. The first model considers the customer choice as a multinomial logit model where the queue length information given to the customer is delayed by a constant Δ. We derive an explicit threshold for the critical delay where below the threshold the two queues are balanced and converge to the equilibrium. However, when Δ is larger than the threshold, the two queues have asynchronous dynamics and the equilibrium point is unstable. In the second model, we consider customer choice as a multinomial logit model where the queue length information given to the customer is a moving average over an interval of Δ.

We also derive an explicit threshold where below the threshold the queues are balanced and above the threshold the queues are asynchronous. It is important for businesses and managers to determine and know these thresholds since using delayed information can have such a large impact on the dynamics of the business. Even small delays can cause oscillations and it is of great importance for managers of these service systems to understand when oscillations can arise based on the arrival and service parameters.

Since our analysis is the first of its kind in the queueing literature, there are many extensions that are worthy of future study. One extension that we would like to explore is the impact of nonstationary arrival rates in the spirit of Massey and Pender [2013], Engblom and Pender [2014], Pender [2014, 2016, 2015a, 2015d]. This is important not only because arrival rates of customers are not constant over time, but also because it is important to know how to distinguish and separate the impact of the time varying arrival rate from the impact of the delayed information given to the customer. Other extensions include the use of different customer choice functions and incorporating customer preferences in the model. With regard to customer preferences, this is a nontrivial problem because the equilibrium solution is no longer a simple expression, but the solution to a transcendental equation. This presents new challenges for deriving analytical formulas that determine synchronous or asynchronous dynamics. A detailed analysis of these extensions will provide a better understanding of what information and how the information that operations managers provide to their customers will affect the dynamics of the system. We plan to explore these extensions in subsequent work.

Acknowledgments

The authors would like to thank the ORIE Department and the Center for Applied Mathematics seminar participants for their feedback and suggestions.

References

Allon, G. & Bassamboo, A. [2011] "The impact of delaying the delay announcements," *Operat. Res.* **59**, 1198–1210.

Allon, G., Bassamboo, A. & Gurvich, I. [2011] "'We will be right with you': Managing customer expectations with vague promises and cheap talk," *Operat. Res.* **59**, 1382–1394.

Armony, M. & Maglaras, C. [2004] "On customer contact centers with a call-back option: Customer decisions, routing rules, and system design," *Operat. Res.* **52**, 271–292.

Armony, M., Shimkin, N. & Whitt, W. [2009] "The impact of delay announcements in many-server queues with abandonment," *Operat. Res.* **57**, 66–81.

Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y., Yom-Tov, G. B. *et al.* [2015] "On patient flow in hospitals: A data-based queueing-science perspective," *Stochast. Syst.* **5**, 146–194.

Asl, F. M. & Ulsoy, A. G. [2003] "Analysis of a system of linear delay differential equations," *J. Dyn. Syst. Measur. Contr.* **125**, 215–223.

Dong, J., Yom-Tov, E. & Yom-Tov, G. B. [2015] "The impact of delay announcements on hospital network coordination and waiting times," Technical Report, Working Paper.

Engblom, S. & Pender, J. [2014] "Approximations for the moments of nonstationary and state dependent birth-death queues."

Guckenheimer, J. & Holmes, P. J. [2013] *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Vol. 42 (Springer Science & Business Media).

Guo, P. & Zipkin, P. [2007] "Analysis and comparison of queues with different levels of delay information," *Manag. Sci.* **53**, 962–970.

Guo, P. & Zipkin, P. [2009] "The impacts of customers' delay-risk sensitivities on a queue with balking," *Probab. Engin. Inform. Sci.* **23**, 409–432.

Hale, J. K. [1971] "Functional differential equations," in *Analytic Theory of Differential Equations* (Springer), pp. 9–22.

Hassin, R. [2007] "Information and uncertainty in a queuing system," *Probab. Engin. Inform. Sci.* **21**, 361–380.

Hui, M. K. & Tse, D. K. [1996] "What to tell consumers in waits of different lengths: An integrative model of service evaluation," *J. Market.*, pp. 81–90.

Hui, M. K., Dube, L. & Chebat, J.-C. [1997] "The impact of music on consumers' reactions to waiting for services," *J. Retail.* **73**, 87–104.

Ibrahim, R. & Whitt, W. [2008] "Real-time delay estimation in call centers," *Proc. 40th Conf. Winter Simulation*, Winter Simulation Conference, pp. 2876–2883.

Ibrahim, R. & Whitt, W. [2009] "Real-time delay estimation in overloaded multiserver queues with abandonments," *Manag. Sci.* **55**, 1729–1742.

Ibrahim, R. & Whitt, W. [2011a] "Real-time delay estimation based on delay history in many-server service systems with time-varying arrivals," *Product. Operat. Manag.* **20**, 654–667.

Ibrahim, R. & Whitt, W. [2011b] "Wait-time predictors for customer service systems with time-varying demand and capacity," *Operat. Res.* **59**, 1106–1118.

Ibrahim, R., Armony, M. & Bassamboo, A. [2016] "Does the past predict the future? The case of delay announcements in service systems," *Manag. Sci.* **62**.

Jennings, O. B. & Pender, J. [2016] "Comparisons of standard and ticket queues in heavy traffic," *Queue. Syst.* **84**, 145–202.

Jouini, O., Dallery, Y. & Akşin, Z. [2009] "Queueing models for full-flexible multi-class call centers with real-time anticipated delays," *Int. J. Product. Econ.* **120**, 389–399.

Jouini, O., Aksin, Z. & Dallery, Y. [2011] "Call centers with delay information: Models and insights," *Manufact. Serv. Operat. Manag.* **13**, 534–548.

Massey, W. A. & Pender, J. [2013] "Gaussian skewness approximation for dynamic rate multi-server queues with abandonment," *Queue. Syst.* **75**, 243–277.

Munichor, N. & Rafaeli, A. [2007] "Numbers or apologies? Customer reactions to telephone waiting time fillers," *J. Appl. Psychol.* **92**, 511.

Neimark, J. I. [1973] "D-decomposition of the space of quasi-polynomials," *Amer. Math. Soc. Transl.* **102**, 95–132.

Pender, J. [2014] "Gram charlier expansion for time varying multiserver queues with abandonment," *SIAM J. Appl. Math.* **74**, 1238–1265.

Pender, J. [2015a] "Nonstationary loss queues via cumulant moment approximations," *Probab. Engin. Inform. Sci.* **29**, 27–49.

Pender, J. [2015b] "Heavy traffic limits for unobservable queues with clearing times," submitted for publication.

Pender, J. [2015c] "The impact of dependence on unobservable queues," submitted for publication.

Pender, J. [2015d] "The truncated normal distribution: Applications to queues with impatient customers," *Operat. Res. Lett.* **43**, 40–45.

Pender, J. [2016] "An analysis of nonstationary coupled queues," *Telecommun. Syst.* **61**, 823–838.

Pender, J. [2017] "Sampling the functional Kolmogorov forward equations for nonstationary queueing networks," *Informs J. Comput.* **29**, 1–17.

Plambeck, E., Bayati, M., Ang, E., Kwasnick, S., Aratow, M. *et al.* [2014] "Forecasting emergency department wait times," Technical report.

Pruyn, A. & Smidts, A. [1998] "Effects of waiting on the satisfaction with the service: Beyond objective time measures," *Int. J. Res. Market.* **15**, 321–334.

Rand, R. H. [2012] "Lecture notes on nonlinear vibrations".

Sarel, D. & Marmorstein, H. [1998] "Managing the delayed service encounter: The role of employee action and customer prior experience," *J. Serv. Market.* **12**, 195–208.

Smith, H. [2010] *An Introduction to Delay Differential Equations with Applications to the Life Sciences*, Vol. 57 (Springer Science & Business Media).

Strogatz, S. H. [2014] *Nonlinear Dynamics and Chaos*: *With Applications to Physics*, *Biology*, *Chemistry*, *and Engineering* (Westview Press).

Taylor, S. [1994] "Waiting for service: The relationship between delays and evaluations of service," *J. Market.*, pp. 56–69.

Whitt, W. [1999a] "Improving service by informing customers about anticipated delays," *Manag. Sci.* **45**, 192–207.

Whitt, W. [1999b] "Predicting queueing delays," *Manag. Sci.* **45**, 870–888.

Yi, S. & Ulsoy, A. G. [2006] "Solution of a system of linear delay differential equations using the matrix lambert function," *2006 American Control Conf.* (IEEE).

Yi, S., Nelson, P. W. & Ulsoy, A. G. [2010] *Time-Delay Systems*: *Analysis and Control Using the Lambert W Function* (World Scientific).

Appendix

We believe that it is important to outline the proof in words before we give the proof in mathematics to give the reader some intuition and to alert the reader to the variety of mathematical methods involved. Since we are analyzing a fluid model or deterministic model, we start with a nonlinear delay differential equation, however, the stability analysis of this nonlinear delay differential equation is nontrivial. Nonetheless, we can still analyze the stability of the nonlinear delay differential equation by using the principle of linear stability analysis from the ordinary differential equation literature. To analyze the stability of the nonlinear delay differential equation, we exploit a generalization of Lyapunov's indirect method for delay differential equations given in Theorem 4.8 of [Smith, 2010]. For readers who are unfamiliar with Theorem 4.8 of [Smith, 2010], the theorem tells us that in order to analyze the original nonlinear system of delay equations it now suffices to linearize the original delay equations about the chosen equilibrium points and analyze the stability of the linearized delay differential equations. Thus, the stability of the linearized system governs the stability of the nonlinear system near the equilibrium point. The fact that the linearized system allows us to determine the stability of the original system is an important and fundamental result in dynamical systems analysis. For more details about linear stability analysis and its applications see [Hale, 1971].

In order to analyze the stability of the linearized system, we can use two approaches. The first approach to determine stability is to exploit the knowledge and properties of the Lambert-W function, see for example, [Asl & Ulsoy, 2003; Yi *et al.*, 2010; Yi & Ulsoy, 2006]. The second approach is to use the D-decomposition method of Neimark [1973], which exploits complex analysis and the eigenvalues of linear operators. We only consider the D-decomposition approach because the Lambert-W function analysis only provides a numerical way for computing the threshold for stability, while the D-decomposition method provides analytical formulas that we provide in our theorems.

A.1. *Proof of Theorem 1*

Proof [Proof of Theorem 1]. Since our result is not common in the queueing literature, we split the proof into several parts to help readers understand the important ingredients that are necessary to prove the theorem. Our proof uses some theory developed in the analysis of nonlinear dynamical systems and the reader is referred to [Strogatz, 2014] for an elementary discussion of linear stability analysis and dynamical system analysis. Moreover, we outline the main parts of the proof in bold text.

A.1.1. *Computing the equilibrium*

The first part of the proof is to compute an equilibrium for the solution to the delay differential equations. In standard ordinary differential equations, one sets the time derivative of the differential equations to zero and solve for the value of the queue length that makes it zero. This implies that we set

$$\dot{q}_1(t) = 0, \tag{A.1}$$

$$\dot{q}_2(t) = 0. \tag{A.2}$$

This further implies that we need to solve the following two nonlinear equations

$$\lambda \cdot \frac{\exp(-q_1(t-\Delta))}{\exp(-q_1(t-\Delta)) + \exp(-q_2(t-\Delta))} - \mu q_1(t) = 0, \tag{A.3}$$

$$\lambda \cdot \frac{\exp(-q_2(t-\Delta))}{\exp(-q_1(t-\Delta)) + \exp(-q_2(t-\Delta))} - \mu q_2(t) = 0. \tag{A.4}$$

Sometimes finding the equilibrium is nontrivial in many nonlinear systems. In our system, we also have the complication that the differential equations are delay differential equations and have an extra complexity. However, in our case, the delay differential equations given in Eqs. (5) and (6) are symmetric and this simplifies some of the analysis. Moreover, in the case when $\Delta = 0$, the two equations converge to the same point since in equilibrium each queue will receive exactly one half of the arrivals and the two service rates are identical. This is also true in the case where the arrival process contains delays in the queue length since in equilibrium, the delayed queue length is equal to the nondelayed queue length. Thus, we have in equilibrium that

$$q_1(t-\Delta) = q_2(t-\Delta) = q_1(t) = q_2(t) = \frac{\lambda}{2\mu} \quad \text{as } t \to \infty. \tag{A.5}$$

To mathematically verify that this is the equilibrium, one can substitute $\frac{\lambda}{2\mu}$ for $q_1(t), q_2(t), q_1(t-\Delta), q_2(t-\Delta)$ and observe that the time derivative of Eqs. (5) and (6) are both equal to zero.

A.1.2. *Understanding the stability of the equilibrium*

Now that we know the equilibrium for Eqs. (5) and (6), we need to understand the stability of the delay differential equations near the equilibrium. The first step in doing this is to set each of the queue lengths to the equilibrium points plus a perturbation. With this in mind, we substitute the following values for each of the queue lengths

$$q_1(t) = \frac{\lambda}{2\mu} + u_1(t), \tag{A.6}$$

$$q_2(t) = \frac{\lambda}{2\mu} + u_2(t). \tag{A.7}$$

In this substitution, $u_1(t)$ and $u_2(t)$ are perturbations about the equilibrium point $\frac{\lambda}{2\mu}$. By substituting Eqs. (A.6) and (A.7) into Eqs. (5) and (6) we get the following equations

$$\dot{u}_1(t) = \lambda \cdot \frac{\exp(-u_1(t-\Delta))}{\exp(-u_1(t-\Delta)) + \exp(-u_2(t-\Delta))} - \mu u_1(t) - \frac{\lambda}{2}, \tag{A.8}$$

$$\dot{u}_2(t) = \lambda \cdot \frac{\exp(-u_2(t-\Delta))}{\exp(-u_1(t-\Delta)) + \exp(-u_2(t-\Delta))} - \mu u_2(t) - \frac{\lambda}{2}. \tag{A.9}$$

Now if we linearize around the point $u_1(t) = u_2(t) = 0$, which is equivalent to performing a Taylor expansion and keeping only the linear terms, we have that the linearized version of $u_1(t)$ and $u_2(t)$, which we now define as $w_1(t)$ and $w_2(t)$ solve the following linear delay differential equations

$$\dot{w}_1(t) = -\frac{\lambda}{4} \cdot (w_1(t-\Delta) - w_2(t-\Delta)) - \mu \cdot w_1(t), \tag{A.10}$$

$$\dot{w}_2(t) = -\frac{\lambda}{4} \cdot (w_2(t-\Delta) - w_1(t-\Delta)) - \mu \cdot w_2(t). \tag{A.11}$$

Now the reader might ask about the validity of the Taylor expansion or the linearization. This Taylor expansion is valid since we expect that the perturbations are near the value of zero. This is because, we expect the two queues to be near the equilibrium values $\frac{\lambda}{2\mu}$ and therefore, the perturbations are expected to be near zero and this is common in linear stability analysis. However, to reiterate again, we are only using the Taylor expansion or linearization since it suffices to analyze the linearized system in order to determine the stability of the original nonlinear system. An additional question a reader not familiar with dynamical systems theory might ask is how does the stability of the linear differential equations in Eqs. (A.10) and (A.11) relate to the shifted nonlinear system given in Eqs. (A.8) and (A.9)? By Lyapunov's linearization theorem as discussed on page 6 of [Rand, 2012], we see that if the real part of the eigenvalues of the linearized system are negative, then the equilibrium point in question is locally asymptotically stable. For more information and results on linear stability analysis or differential equations, the reader is also encouraged to review the first three chapters of [Strogatz, 2014]. Thus, to understand the local stability of our original queueing system, it suffices to study the linearized version using Lyapunov's linearization theorem or Theorem 4.8 of [Smith, 2010].

A.1.3. *Uncoupling the differential equations*

In their current form the delay differential equations for the perturbations or Eqs. (A.10) and (A.11) do not yield any immediate insight since they are a system of coupled equations. However, we can apply a simple transformation to Eqs. (A.10) and (A.11) and the resulting delay differential equations will become uncoupled. Thus, we apply the following transformation to uncouple the system of equations:

$$v_1(t) = w_1(t) + w_2(t), \tag{A.12}$$

$$v_2(t) = w_1(t) - w_2(t). \tag{A.13}$$

This transformation yields the following delay differential equations for the transformed perturbations $v_1(t)$ and $v_2(t)$

$$\dot{v}_1(t) = -\mu \cdot v_1(t), \tag{A.14}$$

$$\dot{v}_2(t) = -\frac{\lambda}{2} \cdot v_2(t-\Delta) - \mu \cdot v_2. \tag{A.15}$$

Remark A.1. Before we analyze the above delay equations, we would like to alert the reader to the fact that $v_1(t)$ and $v_2(t)$ are **linear** equations. This implies that **any** scalar multiple of the solution to these equations is also a solution since they form a vector space. Moreover, any solution can be added to another solution and the sum is also a solution. Thus, the superposition of two solutions is a solution as well.

Since Eq. (A.14) is linear, an ordinary differential equation, and does not depend on the delay parameter Δ, we can explicitly solve for its solution. The general solution of Eq. (A.14) is $v_1(t) = c_1 \exp(-\mu t)$ and is bounded and stable. Thus, it remains for us to analyze the stability of Eq. (A.15). The analysis of this delay differential equation is nontrivial since it depends on the delay parameter Δ. However, to start to analyze the delay differential equation, we substitute the following expression for $v_2(t)$

$$v_2(t) = \exp(r \cdot t). \tag{A.16}$$

By performing the substitution, we see that

$$r \cdot \exp(r \cdot t) = -\frac{\lambda}{2} \cdot \exp(r \cdot (t - \Delta)) - \mu \cdot \exp(r \cdot t) \qquad (A.17)$$

and multiplying both sides by $\exp(-r \cdot t)$, we have the following transcendental equation for the parameter r which is given by the equation

$$r = -\frac{\lambda}{2} \cdot \exp(-r\Delta) - \mu. \qquad (A.18)$$

One should note that this transcendental equation is not a polynomial and involves an exponential function, which implies that the solution is infinite dimensional and has an infinite number of *complex* roots. With the transcendental equation for the parameter r, it only remains for us to find the transition between stable and unstable solutions. Characteristic equations of the form (A.18) are often studied in order to understand changes in the *local stability* of equilibria of delay differential equations. It is therefore important to determine the values of the delay at which there are roots with zero real part. This value of the delay is given by $\Delta_{\text{cr}}(\lambda, \mu)$. When the parameter r crosses the imaginary axis, the stability of the equilibrium changes. In the fully nonlinear system given in Eqs. (5) and (6), this transition generally occurs in a Hopf bifurcation, in which a pair of roots crosses the imaginary axis and a limit cycle is born. Thus, to find the critical delay or $\Delta_{\text{cr}}(\lambda, \mu)$ for the change of stability, we set $r = i\omega$, which yields the following equation

$$i\omega = -\frac{\lambda}{2}(\cos \omega\Delta - i \sin \omega\Delta) - \mu. \qquad (A.19)$$

Writing the real and imaginary parts of Eq. (A.19), we have that:

$$0 = -\frac{\lambda}{2} \cos \omega\Delta - \mu \qquad (A.20)$$

for the real part and

$$\omega = \frac{\lambda}{2} \sin \omega\Delta \qquad (A.21)$$

for the imaginary part. However, in letting $r = i\omega$ one might be tempted to say that we are now making v_2 a complex valued function. This is true, however, we can easily make v_2 a real-valued function by adding the complex conjugate solution to v_2. For example, if we substitute

$$v_2(t) = c \cdot \exp(r \cdot t) + c^* \cdot \exp(r^* \cdot t) \qquad (A.22)$$

where r^* and c^* are the complex conjugates of r and c respectively, then we have a real expression and it is also a solution due to the vector space property of the *linear* equations we are analyzing. Furthermore, we should make it clear that we are not really interested in whether or not v_2 is complex or not. By Theorem 4.8 of [Smith, 2010], we are only concerned about the boundary that separates the eigenvalues of the linear system from being positive or negative. Moreover, one should note that if one can find $\Delta_{\text{cr}}(\lambda, \mu)$ satisfying Eqs. (A.20) and (A.21), then Eq. (A.18) will have pure imaginary roots. The reader should also keep in mind that just because we set $r = i\omega$ and the solution to v_2 is complex does not imply that the original nonlinear delay differential equations q_1 and q_2 are complex valued. Since we are concerned with the stability of the equilibrium point, we use the linearized equation v_2 to determine the stability of the equilibrium point of the original queueing equations.

Our goal in the remaining part of the proof is to find solutions of Eqs. (A.20) and (A.21), which will yield transitions between bounded behavior and unbounded behavior and give us the critical value of the delay Δ_{cr}. If we solve Eqs. (A.20) and (A.21) for the functions $\sin \omega\Delta_{\text{cr}}$ and $\cos \omega\Delta_{\text{cr}}$ we get that

$$\cos \omega\Delta_{\text{cr}} = -\frac{2 \cdot \mu}{\lambda} \qquad (A.23)$$

for the real part and

$$\sin \omega\Delta_{\text{cr}} = \frac{2 \cdot \omega}{\lambda} \qquad (A.24)$$

for the imaginary part. Now by squaring both equations and adding them together we get that

$$\lambda^2 = 4 \cdot (\omega^2 + \mu^2), \qquad (A.25)$$

which by some rearranging yields

$$\omega = \frac{1}{2}\sqrt{\lambda^2 - 4\mu^2}. \qquad (A.26)$$

Now if we go back to use Eq. (A.20) to find an expression for the critical delay $\Delta_{\text{cr}}(\lambda, \mu)$. From Eq. (A.20) we know that

$$\cos \omega\Delta_{\text{cr}} = -\frac{2 \cdot \mu}{\lambda} \qquad (A.27)$$

and therefore by taking the arcosine of both sides, we have that

$$\Delta_{\text{cr}} = \frac{\arccos\left(-\dfrac{2\mu}{\lambda}\right)}{\omega}. \tag{A.28}$$

Finally substituting our expression for ω in Eq. (A.26), we are able to obtain the final expression for the critical delay:

$$\boxed{\Delta_{\text{cr}} = \frac{2\arccos\left(-\dfrac{2\mu}{\lambda}\right)}{\sqrt{\lambda^2 - 4\mu^2}}.} \tag{A.29}$$

■

A.2. *Proof of Proposition 1*

Proof. When $\epsilon = 0$, we reduce back to the original critical threshold of Eq. (A.29):

$$\Delta_0 = \frac{2\arccos\left(-\dfrac{2\mu}{\lambda}\right)}{\sqrt{\lambda^2 - 4\mu^2}}. \tag{A.30}$$

Now we substitute Eq. (8)

$$\Delta = \Delta_0 + \epsilon\Delta_1 \tag{A.31}$$

and (9)

$$r = i\omega + \epsilon(ir_1 + r_2) \tag{A.32}$$

into Eq. (A.18)

$$r = -\frac{\lambda}{2} \cdot \exp(-r\Delta) - \mu \tag{A.33}$$

and do a Taylor expansion for small values of ϵ. Collecting real and imaginary parts, Eq. (A.18) becomes

$$\begin{aligned} 0 &= i\left(\omega - \frac{1}{2}\lambda\sin(\Delta_0\omega)\right) + \frac{1}{2}\lambda\cos(\Delta_0\omega) + \mu \\ &\quad + \epsilon\left[i\left(r_1 - \frac{1}{2}\lambda(\Delta_0 r_2\sin(\Delta_0\omega) + \cos(\Delta_0\omega)\right.\right. \\ &\quad \left.\times(\Delta_0 r_1 - \Delta_1\omega))\right) + r_2 - \frac{1}{2}\lambda(\sin(\Delta_0\omega) \\ &\quad \left.\times(\Delta_1\omega + \Delta_0 r_1) + \Delta_0 r_2\cos(\Delta_0\omega))\right] \\ &\quad + O(\epsilon^2). \end{aligned} \tag{A.34}$$

We set the real and imaginary parts of the $O(\epsilon)$ term separately equal to 0, and solve for r_1 and r_2 since there are two linear equations and two unknowns. Solving for the real part of r, we find that r_2 is equal to the following value

$$r_2 = \frac{-2\Delta_1\lambda\omega\sin\omega\Delta_0}{4\Delta_0\lambda\cos\omega\Delta_0 - \Delta_0^2\lambda^2 - 4}. \tag{A.35}$$

Now if we substitute the expressions for $\sin\omega\Delta_0$

$$\omega = \frac{\lambda}{2}\sin\omega\Delta \tag{A.36}$$

and $\cos\omega\Delta_0$

$$0 = -\frac{\lambda}{2}\cos\omega\Delta - \mu \tag{A.37}$$

from Eqs. (A.20) and (A.21), then we have that

$$r_2 = \frac{4\omega^2\Delta_1}{8\Delta_0\mu + \Delta_0^2\lambda^2 + 4}. \tag{A.38}$$

■

A.3. *Proof of Theorem 2*

Proof [Proof of Theorem 2]. Like in the constant delay setting, we will split the proof into several parts to help readers understand the important ingredients that are necessary to prove the theorem. We should take the time to emphasize that making the observation given in Eq. (16) was crucial to this analysis. Otherwise, we could not exploit the delay differential equation literature for the moving average equations.

A.3.1. *Computing the equilibrium*

The first part of the proof is to compute an equilibrium for the solution to the delay differential equations. In our case, the delay differential equations given in Eqs. (17)–(20) are symmetric. Moreover, in the case where there is no delay, the two equations converge to the same point since in equilibrium each queue will receive exactly one half of the arrivals and the two service rates are identical. This is also true in the case where the arrival process contains delays in the queue length since in equilibrium, the delayed queue length is equal to the non-delayed queue length. It can be shown that there is only one equilibrium where all of the states are equal to each other. One can prove this by substituting $q_2 = \lambda/\mu - q_1$ in the steady state version of Eq. (17) and solving for q_1. One eventually sees that $q_1 = q_2$ is the only solution since any other solution does not obey Eq. (17). Thus, we have in equilibrium that

$$\begin{aligned} q_1(t) &= q_2(t) = m_1(t) = m_2(t) \\ &= \frac{\lambda}{2\mu} \quad \text{as } t \to \infty. \end{aligned} \tag{A.39}$$

A.3.2. *Understanding the stability of the equilibrium*

Now that we know the equilibrium for Eqs. (17)–(20), we need to understand the stability of the delay differential equations around the equilibrium. The first step in doing this is to set each of the queue lengths to the equilibrium values plus a perturbation. Thus, we set each of the queue lengths to

$$q_1(t) = \frac{\lambda}{2\mu} + u_1(t), \tag{A.40}$$

$$q_2(t) = \frac{\lambda}{2\mu} + u_2(t), \tag{A.41}$$

$$m_1(t) = \frac{\lambda}{2\mu} + u_3(t), \tag{A.42}$$

$$m_2(t) = \frac{\lambda}{2\mu} + u_4(t). \tag{A.43}$$

Substitute Eqs. (A.40)–(A.43) into Eqs. (17)–(20) and perform a Taylor expansion or linearize about the point $u_1(t) = u_2(t) = u_3(t) = u_4(t) = 0$, giving

$$\dot{u}_1 = \frac{\lambda}{4} \cdot (u_4(t) - u_3(t)) - \mu \cdot u_1(t), \tag{A.44}$$

$$\dot{u}_2 = \frac{\lambda}{4} \cdot (u_3(t) - u_4(t)) - \mu \cdot u_2(t), \tag{A.45}$$

$$\dot{u}_3 = \frac{1}{\Delta} \cdot (u_1(t) - u_1(t - \Delta)), \tag{A.46}$$

$$\dot{u}_4 = \frac{1}{\Delta} \cdot (u_2(t) - u_2(t - \Delta)). \tag{A.47}$$

Once again, this Taylor expansion or linearization is valid because of the two Lyapunov theorems that are discussed on page 6 of [Rand, 2012] or Theorem 4.8 of [Smith, 2010], which rigorously describe why the linearization is valid and suffices for analyzing stability.

A.3.3. *Uncoupling the differential equations*

In their current form the delay differential equations for the perturbations do not yield any insight since they are coupled together. However, we can make a simple transformation and the resulting delay differential equations will become uncoupled. Thus, we apply the following transformation to uncouple the system of equations:

$$v_1(t) = u_1(t) + u_2(t), \tag{A.48}$$

$$v_2(t) = u_1(t) - u_2(t), \tag{A.49}$$

$$v_3(t) = u_3(t) + u_4(t), \tag{A.50}$$

$$v_4(t) = u_3(t) - u_4(t), \tag{A.51}$$

which gives

$$\dot{v}_1(t) = -\mu \cdot v_1(t), \tag{A.52}$$

$$\dot{v}_2(t) = -\frac{\lambda}{2} \cdot v_4(t) - \mu \cdot v_2(t) \tag{A.53}$$

$$\dot{v}_3(t) = \frac{1}{\Delta} \cdot (v_1(t) - v_1(t - \Delta)), \tag{A.54}$$

$$\dot{v}_4(t) = \frac{1}{\Delta} \cdot (v_2(t) - v_2(t - \Delta)). \tag{A.55}$$

The general solution of Eq. (A.52) is $v_1 = c_1 \exp(-\mu t)$ and is stable (bounded). This also implies that Eq. (A.54) is also stable and bounded since it only depends on the solution of Eq. (A.52). To study Eqs. (A.53) and (A.55), we let

$$v_2 = A \exp(rt), \tag{A.56}$$

$$v_4 = B \exp(rt). \tag{A.57}$$

These solutions imply the following relationships between the constants A, B, and r.

$$Ar = -\frac{\lambda}{2} B - \mu A, \tag{A.58}$$

$$Br = \frac{1}{\Delta}(A - A \exp(-r\Delta)) \tag{A.59}$$

solving for A yields

$$A = -\frac{\lambda}{2(\mu + r)} B \tag{A.60}$$

and rearranging yields the following equation for r

$$r = \frac{\lambda}{2\Delta \cdot r}(\exp(-r\Delta) - 1) - \mu. \tag{A.61}$$

Now it remains for us to understand the transition between the stable and unstable solutions once again.

A.3.4. *Understanding the transition between stable and unstable solutions*

To find the transition between stable and unstable solutions, set $r = i\omega$, giving us the following equation

$$i\omega = \frac{\lambda}{2\Delta i\omega}(\exp(-i\omega\Delta) - 1) - \mu. \tag{A.62}$$

Multiplying both sides by $i\omega$ and using Euler's identity, we have that

$$\frac{\lambda}{2\Delta}(\cos(\omega\Delta) - i\sin(\omega\Delta) - 1) - \mu i\omega + \omega^2 = 0. \tag{A.63}$$

Writing the real and imaginary parts of Eq. (A.63), we get:

$$\cos(\omega\Delta) = 1 - \frac{2\Delta\omega^2}{\lambda} \tag{A.64}$$

for the real part and

$$\sin(\omega\Delta) = -\frac{2\Delta\mu\omega}{\lambda}. \tag{A.65}$$

Once again by squaring and adding $\sin\omega\Delta$ and $\cos\omega\Delta$ together, we get:

$$\omega = \sqrt{\frac{\lambda}{\Delta} - \mu^2}. \tag{A.66}$$

Finally, substituting the expression for ω into Eq. (A.65) gives us the final expression for the critical delay, which is the solution to the following transcendental equation:

$$\boxed{\sin\left(\Delta \cdot \sqrt{\frac{\lambda}{\Delta} - \mu^2}\right) + \frac{2\mu\Delta}{\lambda} \cdot \sqrt{\frac{\lambda}{\Delta} - \mu^2} = 0.} \tag{A.67}$$

■

A.4. ***Proof of Proposition 2***

Proof. In order to prove this, we can follow the same steps as in Proposition 1. When $\epsilon = 0$, we reduce back to the original critical threshold of Eq. (A.67).

$$\sin\left(\Delta \cdot \sqrt{\frac{\lambda}{\Delta} - \mu^2}\right) + \frac{2\mu\Delta}{\lambda} \cdot \sqrt{\frac{\lambda}{\Delta} - \mu^2} = 0. \tag{A.68}$$

Now we substitute Eq. (22)

$$\Delta = \Delta_0 + \epsilon\Delta_1 \tag{A.69}$$

and Eq. (A.35)

$$r = i\omega + \epsilon(ir_1 + r_2) \tag{A.70}$$

into Eq. (A.61)

$$r = \frac{\lambda}{2\Delta \cdot r}(\exp(-r\Delta) - 1) - \mu \tag{A.71}$$

and do a Taylor expansion for small values of ϵ. Then solve for r_1 and r_2. Solving for the real part of r, we find that r_2 is equal to the following value

$$r_2 = \frac{2\Delta_1\omega^2 \cdot (2\Delta_0\omega^2 - 2\mu\lambda)}{8\Delta_0^2\mu\omega^2 + 12\Delta_0\omega^2 + 4\Delta_0\lambda\mu + \Delta_0\lambda^2 + 4\lambda}. \tag{A.72}$$

■