International Journal of Bifurcation and Chaos

k-Nearest Neighbor Queues with Delayed Information

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

Received (to be inserted by publisher)

In this paper, we analyze a model called the k-nearest neighbor queue with the possibility of having delayed queue length feedback. We prove fluid limits for the stochastic queueing model and show that the fluid limit converges to a system of delay differential equations. Using the properties of circulant matrices, we derive a closed form expression for the value of the critical delay, which governs whether the delayed information will induce oscillations or a Hopf bifurcation in our queueing system.

Keywords: Queueing theory, k-nearest neighbors, delayed information, delay differential equations, stochastic processes, fluid limits, Hopf bifurcation.

1. Introduction

Cloud computing services are pervasive in our society and are expanding across the world. These services are supported by very complex infrastructures in data centers. As demand for cloud computing services continues to increase it is important to understand how to manage these data centers so that they do not overload. It is not surprising that many researchers are now developing new load balancing algorithms for data center applications. Since data centers can be modeled well by queueing theory, often researchers use queueing theory to verify that their algorithms perform well. However, many data center managers must make the trade off between using storing information about current system state or component of the system load among several queues. Many algorithms have been developed to do this in the context of data centers, telecommunications networks, and even call centers.

In order to achieve load balancing many authors have analyzed stochastic queueing models that are modifications of join the shortest queue, see for example the work by [Mitzenmacher, 2001; Byers et al., 2004; Xie et al., 2015; Bramson et al., 2010; Lu et al., 2011; Bramson et al., 2013; Mukherjee et al., 2016, 2018, 2017; Banerjee et al., 2019; Budhiraja et al., 2017; Aghajani and Ramanan, 2017; Aghajani et al., 2015, 2017; Cybenko, 1989; Mitzenmacher, 2016]. Many of these variants of join the shortest queue allow for joining the shortest among d randomly selected queues. Selecting only $d \geq 2$ queues allows for less storage of all queue lengths and still achieves high quality performance.

Our paper is motivated by the load balancing literature, but is also motivated by the work of [Dong et al., 2018; Ding et al.] where customers have the option to choose which queue they join based on the queue length they observe in a delay announcement or smartphone app. The work of Dong et al. [2018]; Ding et al. explores the use of the multinomial logit choice model (MNL) as a probabilistic way for customers to choose which queue they will join. However, the current literature assumes two important,

but not realistic properties about the information that customers have when making their decision of which queue to join. The first assumption is that customers have information about all queues. In the context of telecommunication systems and data centers, this assumption is not realistic as it requires a substantial amount of data storage and knowledge. Thus, it is common in these settings that a queue might have information only about its neighbors. The second assumption is that customers receive information in a real-time fashion. This is also not realistic as there are many situations where either the information is delayed from a technological point of view or the decision about joining a queue must be made before actually joining the queue. In both situations, the information can be viewed as being delayed.

The queueing model that we present in this work tackles both of these gaps in the literature. First we consider a 2k-nearest neighbor set-up where a customer who will join the i^{th} queue also knows the queue length of the k neighbors to the left and k neighbors to right. One might view this nearest neighbor setup as each queue knowing some local information about the queues nearest to it. Second, we consider the fact that the information about the queues is delayed and is not given in real-time. This is important to consider in our model as the rate of information about the system is not infinitely fast. Thus, the queue length information about the neighbors must be delayed.

We should also mention that our work is closely related to the work by Mitzenmacher [2000]; Ding et al.; Pender et al. [2020]; Lipshutz and Williams [2015]; Novitzky et al. [2020]; Doldo et al. [2021] in that these papers either consider choice model dynamics in the construction of their queueing models or they consider delayed dynamics in the context of queueing theory. We should also mention that there is recent work by Atar and Lipshutz [2021] that considers heavy traffic limits for systems with delayed information. Thus, our work is similar in this spirit.

1.1. *Our Contributions*

This paper makes the following two major contributions:

- First, we develop a new model of k-nearest neighbor queues, where the information about each queue length is delayed by a constant Δ.
- We prove fluid limits for a scaled version of the queue length process and show that the fluid limit is a delay differential equation. Moreover, we prove the exact threshold for when oscillations in the queue length dynamics will occur in this fluid model.

1.2. *Organization of the Paper*

The rest of the paper is organized as follows:

In Section 2, we describe and construct the k-nearest neighbor queueing model. We prove the fluid limit of the scaled queue length process and derive an exact threshold where oscillations will occur if the delay in information is larger than the threshold. In Section 3, we conclude and provide new directions for future research. Finally, all proofs of our main results are given in the Appendix.

2. k-Nearest Neighbor Queueing Model

In this section, we present a new stochastic queueing model where customers that would arrive to the i^{th} queue are allowed to also join any of the k neighbors to the right or to the left of the i^{th} queue. Thus, any arriving customer will have the option of joining $2k+1$ queues. This choice reflects the fact that often in load balancing settings, one might not have the information of all of the possible queues one could join since it can be computationally expensive to store all of this information. In settings where queue length information is provided to customers via smartphone apps like in bike-sharing networks [Schuijbroek et al., 2017; Faghih-Imani et al., 2017] or waiting times at Disneyland [Nirenberg et al., 2018], a smartphone app will only indicate the nearest possible queues you can retrieve or return a bike. Thus, the information that is given to the customer is limited to the nearest stations to them. Thus, we begin with N infinite-server queues operating in parallel, where customers make a choice of which queue to join by taking the size of the queue length into account via a customer choice model. In addition to only being able to know the

queue length of one's k-nearest neighbors to the left and to the right, we also add the complication that the queue length information that is given to the customer is delayed by a constant Δ for all of the queues. Therefore, the queue length that the customer receives is not in real-time, which is commonly assumed. In fact, the queue length information that a customer actually receives is the queue length Δ time units in the past.

Thus, in a stochastic context with N queues, the probability of joining the i^{th} queue is given by the following expression

$$p_i(Q(t),\Delta) = \frac{g(Q_i(t-\Delta))}{\sum_{j=1}^{k} g(Q_{i-j}(t-\Delta)) + g(Q_{i+j}(t-\Delta)) + g(Q_i(t-\Delta))} \tag{1}$$

where $Q(t) = (Q_1(t), Q_2(t), ..., Q_N(t))$. We also make the following assumptions regarding the function $g(x)$.

- The function $g(x)$ maps from $\mathbb{R}_+ \to \mathbb{R}_+/\{0\}$ and is a continuously differentiable function with a uniformly bounded derivative. Moreover, we assume the function is strictly non-increasing.
- $g\left(\frac{\lambda}{\mu(2k+1)}\right)$ is bounded away from zero $\forall\ k \in \mathbb{N}$ and $k \leq \frac{N-1}{2}$.
- $g'(x) < 0,\ \forall x \geq 0$.

It is evident from the above expression that if the queue length in station i is larger than the other queue lengths, then the i^{th} station has a smaller likelihood of receiving the next arrival. This decrease in likelihood as the queue length increases represents the disdain customers have for waiting in longer lines. Using these probabilities for joining each queue allows us to construct the following stochastic model for the queue length process of our N dimensional system for $t \geq 0$

$$Q_i(t) = Q_i(0) + \Pi_i^a\left(\int_0^t \lambda \cdot p_i(Q(s),\Delta)ds\right) - \Pi_i^d\left(\int_0^t \mu Q_i(s)ds\right)$$

where each $\Pi(\cdot)$ is a unit rate Poisson process and $Q_i(s) = \varphi_i(s)$ for all $s \in [-\Delta, 0]$. In this model, for the i^{th} queue, we have that

$$\Pi_i^a\left(\int_0^t \lambda p_i(Q(s),\Delta)ds\right) \tag{2}$$

counts the number of customers that decide to join the i^{th} queue in the time interval $(0, t]$. Note that the rate depends on the queue length at time $t-\Delta$ and not time t, hence representing the lag in information. Similarly

$$\Pi_i^d\left(\int_0^t \mu Q_i(s)ds\right) \tag{3}$$

counts the number of customers that depart the i^{th} queue having received service from an agent or server in the time interval $(0, t]$. However, in contrast to the arrival process, the service process depends on the current queue length and not the past queue length.

2.1. *Special Cases of the Model*

In this section, we specify some forms for the function $g(x)$. In the case that $g(x) = e^{-\theta x}$, we have the choice model is given by a k-nearest neighbor version of the multinomial logit choice model (MNL). In the MNL choice model, there is an analagous utility perspective about the choices that customers make. In fact the utility for being served in the i^{th} queue with delayed queue length $Q_i(t-\Delta)$ is $u_i(Q_i(t-\Delta)) = -\theta Q_i(t-\Delta)$. Thus, there is an economic interpretation of the MNL model. We should also mention that the MNL model has two important asymptotic regimes, which are of independent interest. In the setting where we let the sensitivity parameter $\theta \to 0$, we find that our probabilities of choosing a queue to join converge to a uniform distribution over the k-nearest neighbors. Moreover, in the setting where we let the sensitivity parameter $\theta \to \infty$, we find that our probabilities of choosing a queue to join converge to an indicator function for

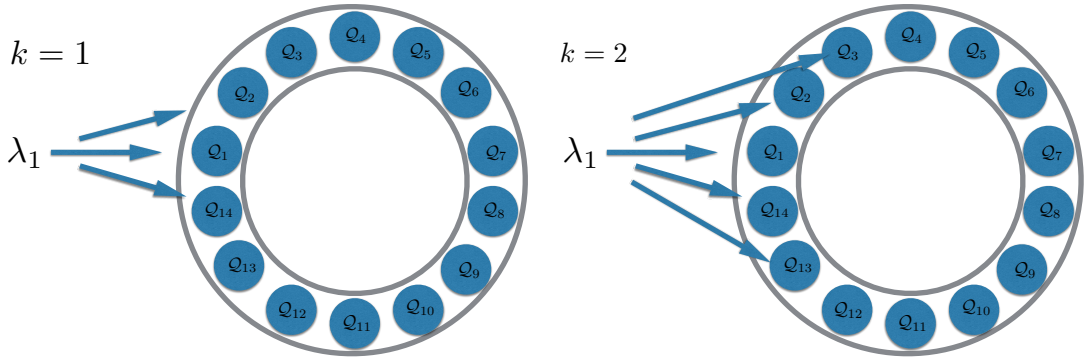

Fig. 1. k-nearest neighbor queue with $k = 1$ (left) and $k = 2$ (right).

shortest queue. If there is a tie, then it is uniform over those queues that are identical for the smallest queue length. As a result, we can view the MNL model outside of those two extremes as a smoothed and infinitely differentiable approximation of the join the shortest queue model.

Another function that could be used is the polynomial $g(x) = \frac{\gamma}{(x^\alpha + c)^\beta}$. This function is commonly used in the context of utility maximization and is a version of the Cobb-Douglas family of utlities used in economics. The interested reader for additional functions can see the many references in Hassin and Haviv [2003]; Hassin [2016].

Finally, we can also suggest that any tail cdf of a continuous non-negative distribution could also work as a potential function for $g(x)$. For example $g(x) = e^{-(x\theta)^k}$ is the tail cdf of a Weibull distribution. Another distribution potentially is the folded normal distribution where the tail cdf is given by the function $g(x) = \frac{1}{2}\left[\text{erf}\left(\frac{x+\mu}{\sqrt{2\sigma^2}}\right) + \text{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right]$ for $x \geq 0$ and where erf() is the error function.

2.2. *Fluid Scaling and Fluid Limits*

In many service systems, the arrival rate of customers is high. For example in Disneyland there are thousands of customers moving around the park and deciding on which ride they should join. Motivated by the large number of customers, we introduce the following scaled queue length process by a parameter η

$$Q_i^\eta(t) = Q_i^\eta(0) + \Pi_i^a\left(\int_0^t \lambda \cdot p_i(Q^\eta(s), \Delta)ds\right) - \frac{1}{\eta}\Pi_i^d\left(\eta\int_0^t \mu Q_i^\eta(s)ds\right). \tag{4}$$

Note that we write $Q_i^\eta(0)$ for the initial condition, however, it really should be interpreted as an initial function since it will contain all of the information needed to know the queue length in the past Δ time units. Now by letting the scaling parameter η go to infinity gives us our first result.

Theorem 1. *Let $\gamma_i(s)$ be a non-negative Lipschitz continuous function that knows all of the queue length values on the interval $[-\Delta, 0]$. Then, if $Q_i^\eta(s) \to \gamma_i(s)$ almost surely for all $s \in [-\Delta, 0]$ and for all $1 \leq i \leq N$, then the sequence of stochastic processes $\{Q^\eta(t) = (Q_1^\eta(t), Q_2^\eta(t), ..., Q_N^\eta(t)\}_{\eta\in\mathbb{N}}$ converges almost surely and uniformly on compact sets of time to $(q(t) = (q_1(t), q_2(t), ..., q_N(t))$ where*

$$\dot{q}_i(t) = \lambda \cdot p_i(q(t), \Delta) - \mu q_i(t) \tag{5}$$

and $q_i(s) = \gamma_i(s)$ for all $s \in [-\Delta, 0]$ and for all $1 \leq i \leq N$.

Proof. *See Appendix.* ■

This result states that as we let η go towards infinity, the sequence of queueing processes converges to a system of **delay differential equations**. Unlike ordinary differential equations, the existence and uniqueness results for delay differential equations is much less well known. However, we provide the result of existence and uniqueness for the delay differential system that we analyze in this paper below.

Theorem 2. *Given a non-negative Lipschitz continuous initial function $\varphi_i : [-\Delta, 0] \to \mathbb{R}_+$ for all $1 \leq i \leq N$ and a finite time horizon $T > 0$, there exists a unique Lipschitz continuous function $q(t) = \{q(t)\}_{-\Delta \leq t \leq T}$ that is the solution to the following delay differential equation*

$$\dot{q}_i(t) = \lambda \cdot p_i(q(t), \Delta) - \mu q_i(t) \tag{6}$$

and $q_i(s) = \varphi_i(s)$ for all $s \in [-\Delta, 0]$ and for all $1 \leq i \leq N$.

Proof. The proof of this result can be found in Hale [1971] as our model satisfies the Lipschitz continuity conditions of the right-hand side. ■

2.3. *Oscillations in the k-Nearest Neighbor Model*

Unlike ordinary differential equations, delay differential equations are truly infinite dimensional and the smallest of delays can cause surprising dynamics. Recent work by Pender et al. [2017] explores a two dimensional version of our fluid limit and uncovers that the two queues can oscillate in equilibrium when the delay is large enough. Pender et al. [2017] also characterizes the threshold in terms of the model parameters and provides an exact formula for the threshold in the two dimensional case. However, this analysis is limited and does not immediately generalize to the multi-dimensional setting. The main goal of this section is to generalize the critical delay analysis of Pender et al. [2017] and derive the exact threshold for an arbitrary number of queues. Before we state the formal theorem, we will need a lemma, which is stated below. This lemma provides an explicit expression for the sum of cosines in terms of the Dirichlet kernel. We will need the lemma for analyzing the stability of the delay differential equations that arise from our fluid model.

Lemma 1.

$$\sum_{n=1}^{N} \cos(n\theta) = -\frac{1}{2} + \frac{\sin\left(\left(N + \frac{1}{2}\right)\theta\right)}{2\sin\left(\frac{\theta}{2}\right)} \tag{7}$$

Proof. Start with the series

$$f(x) = \frac{1}{2} + \sum_{k=1}^{n} \cos(kx) \tag{8}$$

Then one should multiply both sides of the above by $2\sin(x/2)$ and use the trigonometric identity

$$\cos(a)\sin(b) = \frac{\sin(a+b) - \sin(a-b)}{2} \tag{9}$$

to reduce the right-hand side to

$$\sin((n + 1/2)x). \tag{10}$$

This completes the proof. ■

Lemma 2. *The equilibrium for the delay differential system of equation given in Equation 6 is unique.*

Proof. To mathematically verify that this is an equilibrium for the system of equations, one can substitute $\frac{\lambda}{N\mu}$ for $q_i(t)$ and $q_i(t - \Delta)$ and make the observation that the time derivatives for all of the equations are equal to zero. However, we may be unsure of whether the equilibrium is unique. We can show that the equilibrium in our setting is unique by noting that

$$\dot{q}_i(t) = 0 \tag{11}$$

and setting the equilibrium $q_i(\infty) = c_i$. Thus, for each i, we have that

$$\lambda \cdot p(c_i, \Delta) = \mu \cdot c_i. \tag{12}$$

This implies that

$$\frac{g(c_i)}{c_i} = \frac{\mu}{\lambda} \cdot \frac{g(c_i)}{\sum_{j=1}^{k} g(c_{i-j}) + g(c_{i+j}) + g(c_i)} = \text{constant}. \tag{13}$$

Now we observe that the function on the left $\frac{g(c_i)}{c_i}$ is a one-to-one function of $c_i \geq 0$. Therefore, all of the functions $\frac{g(c_i)}{c_i}$ are equal implies that all of the c_i terms are equal. This implies that our equilibrium is unique. ■

Theorem 3. *For the constant delay choice queueing model given in Equation 6 with arbitrary $N \geq 2$, the critical delay, $\Delta_{cr}(\lambda, \mu, N)$, is given by the following expression*

$$\Delta_{cr}(\lambda, \mu, N, k) = \frac{\arccos\left(\frac{\mu}{\min_{1 \leq j \leq N} \alpha_j}\right)}{\sqrt{(\min_{1 \leq j \leq N} \alpha_j)^2 - \mu^2}} \tag{14}$$

where α_j is given by

$$\alpha_j = \frac{\left(2k + 1 - \frac{\sin((k+1/2) \cdot 2\pi j/N)}{\sin(2\pi j/N}\right) \cdot \lambda \cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)^2 g\left(\frac{\lambda}{(2k+1)\mu}\right)}. \tag{15}$$

Proof.

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

$$\dot{q}_i(t) = 0 \tag{16}$$

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

$$\lambda \cdot p_i(q(t), \Delta) - \mu q_i(t) = 0 \tag{17}$$

Sometimes finding the equilibrium is non-trivial in many non-linear 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 Equation 17 are symmetric and this simplifies some of the analysis. In this case the N equations converge to the same point since in equilibrium each queue will receive exactly $1/(2k+1)$ of the arrivals and the service rates of all of the queues are the same. Thus, we have in equilibrium that for all $1 \leq i \leq N$

$$q_i(t - \Delta) = q_i(t) = \frac{\lambda}{(2k+1)\mu} \quad \text{as } t \to \infty. \tag{18}$$

To mathematically verify that this is an equilibrium for the system of equations, one can substitute $\frac{\lambda}{(2k+1)\mu}$ for $q_i(t)$ and $q_i(t - \Delta)$ and make the observation that the time derivative for all of the equations are equal to zero.

Now that we have established the equilibrium for Equation 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

$$q_i(t) = \frac{\lambda}{(2k+1)\mu} + u_i(t) \tag{19}$$

In this substitution, the $u_i(t)$ are perturbations about the equilibrium point $\frac{\lambda}{(2k+1)\mu}$. By substituting Equation 19 into Equation 6 we get the following equations

$$\dot{u}_i(t) = \lambda \cdot p_i\left(\frac{\lambda}{(2k+1)\mu} + u(t), \Delta\right) - \mu u_i(t) - \frac{\lambda}{(2k+1)\mu} \tag{20}$$

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

$$\begin{aligned}
\dot{w}_i(t) &= \frac{2k\lambda \cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)^2 g\left(\frac{\lambda}{(2k+1)\mu}\right)} \cdot w_i(t-\Delta) - \sum_{j=1}^{\lfloor (k+1)/2 \rfloor - 1} \frac{\lambda \cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)^2 g\left(\frac{\lambda}{(2k+1)\mu}\right)} \cdot w_{i-j}(t-\Delta) \\
&- \sum_{j=1}^{\lfloor (k+1)/2 \rfloor - 1} \frac{\lambda \cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)^2 g\left(\frac{\lambda}{(2k+1)\mu}\right)} \cdot w_{i+j}(t-\Delta) - \mu \cdot w_i(t) \\
&= \frac{\lambda \cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1) g\left(\frac{\lambda}{(2k+1)\mu}\right)} \cdot w_i(t-\Delta) - \sum_{j=1}^{\lfloor (k+1)/2 \rfloor - 1} \frac{\lambda \cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)^2 g\left(\frac{\lambda}{(2k+1)\mu}\right)} \cdot w_{i-j}(t-\Delta) \\
&- \sum_{j=0}^{\lfloor (k+1)/2 \rfloor - 1} \frac{\lambda \cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)^2 g\left(\frac{\lambda}{(2k+1)\mu}\right)} \cdot w_{i+j}(t-\Delta) - \mu \cdot w_i(t)
\end{aligned}$$

This can be written as a matrix system by

$$\begin{aligned}
\dot{w}(t) &= \frac{\lambda \cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1) g\left(\frac{\lambda}{(2k+1)\mu}\right)} \cdot \mathcal{I} \cdot w(t-\Delta) \\
&- \frac{\lambda \cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)^2 g\left(\frac{\lambda}{(2k+1)\mu}\right)} \cdot \mathcal{C} w(t-\Delta) - \mu \cdot \mathcal{I} w(t)
\end{aligned} \tag{21}$$

where $\mathcal{I}$ is an N dimensional identity matrix and $\mathcal{C}$ is a N dimensional symmetric circulant matrix. Circulant matrices are ideal since much is known about their eigenvalues. Thus, $\mathcal{C}$ has the following representation

$$\mathcal{C} = \begin{bmatrix} c_0 & c_1 & \dots & c_2 & c_1 \\ c_1 & c_0 & c_1 & & c_2 \\ \vdots & c_1 & c_0 & \ddots & \vdots \\ c_2 & & \ddots & \ddots & c_1 \\ c_1 & c_2 & \dots & c_1 & c_0 \end{bmatrix}.$$

With the representation of our linearized system in Equation 22, we can now exploit the fact that both $\mathcal{C}$ and $\mathcal{I}$ can be simultaneously diagonalized. Thus, we can write both $\mathcal{C}$ and $\mathcal{I}$ in terms of the eigenvectors of the matrix $\mathcal{C}$. If we denote S as the orthogonal matrix of the eigenvectors of $\mathcal{C}$ and denote Λ as diagonal matrix of the eigenvalues of $\mathcal{C}$, then we have that $\mathcal{C}$ and $\mathcal{I}$ can both be decomposed in terms of S, U^{-1}, and Λ as

$$\mathcal{C} = U\Lambda U^{-1} \tag{22}$$

$$\mathcal{I} = U\mathcal{I}U^{-1}. \tag{23}$$

The matrix $\mathcal{C}$ has rank $N - k$ and therefore has $N - k$ distinct eigenvalues. The eigenvalues of any real symmetric matrix are real. Thus, the corresponding eigenvalues of our circulant matrix have explicit

expressions in terms of the roots of unity i.e.

$$\lambda_j = c_0 + 2c_1\Re\omega_j + 2c_2\Re\omega_j^2 + \ldots + 2c_{n/2-1}\Re\omega_j^{n/2-1} + c_{n/2}\omega_j^{n/2} \tag{24}$$

for n even, and

$$\lambda_j = c_0 + 2c_1\Re\omega_j + 2c_2\Re\omega_j^2 + \ldots + 2c_{(n-1)/2}\Re\omega_j^{(n-1)/2} \tag{25}$$

for odd n. This can be further simplified by using the roots of unity identity that

$$\Re\omega_j^k = \cos(2\pi jk/N). \tag{26}$$

Thus, the j^{th} eigenvalue of the circulant matrix is given by the following expression

$$\lambda_j = 1 + \sum_{m=1}^{k} 2\cos(2\pi jm/N) = \frac{\sin\left(\left(k+\frac{1}{2}\right)\cdot(2\pi j/N)\right)}{\sin\left(\pi j/N\right)} \tag{27}$$

Using this knowledge of the eigenvalues of the matrix $\mathcal{C}$, we now we define, $v = U^{-1}w$ or $w = Uv$ and this leads us to the following delay differential system for v

$$\dot{w}(t) = U\dot{v}(t) \tag{28}$$

$$= \frac{\lambda\cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)g\left(\frac{\lambda}{(2k+1)\mu}\right)}\cdot\mathcal{I}w(t-\Delta) \tag{29}$$

$$- \frac{\lambda\cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)^2 g\left(\frac{\lambda}{(2k+1)\mu}\right)}\cdot\mathcal{C}w(t-\Delta) - \mu\cdot\mathcal{I}w(t)$$

$$= \frac{\lambda\cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)g\left(\frac{\lambda}{(2k+1)\mu}\right)}\cdot\mathcal{I}w(t-\Delta) \tag{30}$$

$$- \frac{\lambda\cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)^2 g\left(\frac{\lambda}{(2k+1)\mu}\right)}\cdot U\Lambda U^{-1}w(t-\Delta) - \mu\cdot\mathcal{I}w(t)$$

$$= \frac{\lambda\cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)g\left(\frac{\lambda}{(2k+1)\mu}\right)}\cdot\mathcal{I}Uv(t-\Delta) \tag{31}$$

$$- \frac{\lambda\cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)^2 g\left(\frac{\lambda}{(2k+1)\mu}\right)}\cdot U\Lambda U^{-1}Uv(t-\Delta) - \mu\cdot\mathcal{I}Uv(t)$$

$$= \frac{\lambda\cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)g\left(\frac{\lambda}{(2k+1)\mu}\right)}\cdot Uv(t-\Delta) \tag{32}$$

$$- \frac{\lambda\cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)^2 g\left(\frac{\lambda}{(2k+1)\mu}\right)}\cdot U\Lambda v(t-\Delta) - \mu\cdot Uv(t).$$

Now by multiplying both sides by U^{-1} we have the following delay differential system for v

$$\dot{v}(t) = \frac{\lambda \cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)g\left(\frac{\lambda}{(2k+1)\mu}\right)} \cdot U^{-1}Uv(t-\Delta) \tag{33}$$

$$- \frac{\lambda \cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)^2 g\left(\frac{\lambda}{(2k+1)\mu}\right)} \cdot U^{-1}U\Lambda v(t-\Delta) - \mu \cdot U^{-1}Uv(t)$$

$$= \frac{\lambda \cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)g\left(\frac{\lambda}{(2k+1)\mu}\right)} \cdot \mathcal{I}v(t-\Delta) \tag{34}$$

$$- \frac{\lambda \cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)^2 g\left(\frac{\lambda}{(2k+1)\mu}\right)} \cdot \Lambda v(t-\Delta) - \mu \cdot \mathcal{I}v(t).$$

Thus, for the i^{th} entry of the vector v, we have the following delay differential equation

$$\dot{v}_i(t) = \frac{\lambda \cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)g\left(\frac{\lambda}{(2k+1)\mu}\right)} \cdot v_i(t-\Delta) \tag{35}$$

$$- \frac{\lambda \cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)^2 g\left(\frac{\lambda}{(2k+1)\mu}\right)} \cdot \Lambda_{ii} \cdot v_i(t-\Delta) - \mu \cdot v_i(t)$$

$$= \frac{(2k+1-\Lambda_{ii}) \cdot \lambda \cdot g'\left(\frac{\lambda}{(2k+1)\mu}\right)}{(2k+1)^2 g\left(\frac{\lambda}{(2k+1)\mu}\right)} \cdot v_i(t-\Delta) - \mu \cdot v_i(t), \tag{36}$$

where Λ_{ii} is the i^{th} diagonal entry of the matrix Λ. One crucial observation is that this representation shows that system of delay equations given in Equation 36 are uncoupled and can be analyzed individually for stability purposes. To finish the proof, we observe that it only remains to analyze the stability of the each equation for $v_i(t)$. To do this we make the ansatz $v_i(t) = e^{rt}$ and derive an equation for the variable r. This yields the following transcendental equations for r

$$r = \alpha_i \cdot e^{-r\Delta} - \mu. \tag{37}$$

Note that this is the real difference between ordinary differential equations and delay differential equations. These types of transcendental equations do not appear in ordinary differential equations because Δ is typically equal to zero in the ordinary differential equation context. Now we complete the proof by analyzing our transcendental equation for r. If we substitute $r = i\omega$, we obtain two equations for the real and imaginary parts respectively using Euler's identity

$$\cos(\omega\Delta) = \frac{\mu}{\alpha_i} \tag{38}$$

$$\sin(\omega\Delta) = -\frac{\omega}{\alpha_i}. \tag{39}$$

Now by squaring both sides and adding the two equations together we arrive at the following equation

$$\cos^2(\omega\Delta) + \sin^2(\omega\Delta) = 1 = \frac{(\mu^2 + \omega^2)}{{\alpha_i}^2} \tag{40}$$

By moving all terms of Equation 40 that do not involve ω to the right, we can isolate an expression for ω. Thus, solving for ω, we arrive at the following expression

$$\omega = \sqrt{\alpha_i^2 - \mu^2}. \tag{41}$$

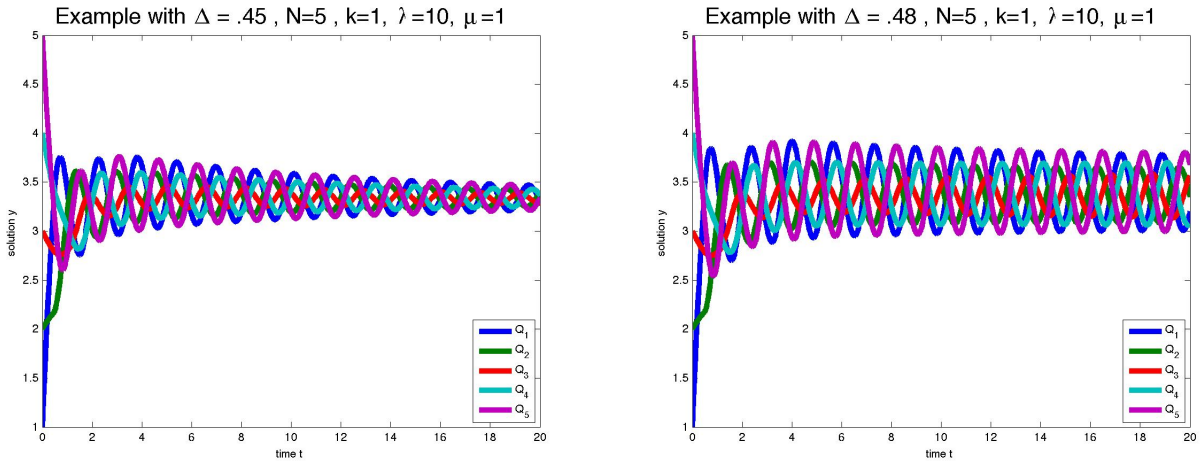

Fig. 2. $\Delta_{cr} = .468$.

Using this expression for ω, we can finally invert Equation 38 since it does not contain ω on the right hand side unlike Equation 39 to solve for the critical value of Δ. We find that our threshold Δ is equal to

$$\Delta^i_{cr}(\lambda, \mu, N, k) = \frac{\arccos\left(\frac{\mu}{\alpha_i}\right)}{\sqrt{\alpha_i^2 - \mu^2}}. \tag{42}$$

Thus our proof is complete. ■

Theorem 3 provides a local characterization of the oscillation behavior of an arbitrary queueing system with N queues. If the delay Δ is larger than the critical delay $\Delta_{cr}(\lambda, \mu, N)$, then we should expect that the N queues should oscillate in equilibrium. However, if the delay Δ is smaller than the critical delay $\Delta_{cr}(\lambda, \mu, N)$, then we should expect that the N queues should converge to the equilibrium point $\frac{\lambda}{\mu N}$ and not oscillate around the equilibrium.

2.4. *Numerical Results for Fluid Limits*

In this section, we describe some numerical results that compare the fluid limits before and after the critical delay values. On the left of Figure 2, we plot an example of N=5 queues and we let $k = 1$. Since $\Delta < \Delta_{cr} = .468$, we observe that the queues converge to the equilibrium and the initial oscillations vanish over time. However, on the right of Figure 2, we observe that because $\Delta > \Delta_{cr} = .468$, the queues oscillate around the equilibrium. Moreover, it is important to note that even though each of the queue lengths oscillate around the equilibrium, the size of the oscillations are not all equal. Thus, there is an asymmetry in the size of the oscillations.

In Figure 3, we plot an example of N=5 queues and we let $k = 1$, but we increase the arrival rate. Since $\Delta < \Delta_{cr} = .167$, we observe that the queues converge to the equilibrium and the initial oscillations vanish over time. However, inon the right of Figure 3, we observe that because $\Delta > \Delta_{cr} = .167$, the queues oscillate around the equilibrium. Moreover, it is important to note that even though each of the queue lengths oscillate around the equilibrium, the size of the oscillations are all actually equal this time. Thus, the size of the oscillations appears to be symmetric. This is a clear difference from the previous figures.

3. Conclusion and Future Research

In this paper, we analyze a new N-dimensional stochastic queueing model that incorporates customer choice and delayed queue length information in a k-nearest neighbor fashion. Our model considers the customer choice as a generalized multinomial logit choice model where the queue length information given to the customer is delayed by an amount of size Δ. We prove fluid limit theorems for our queueing process and

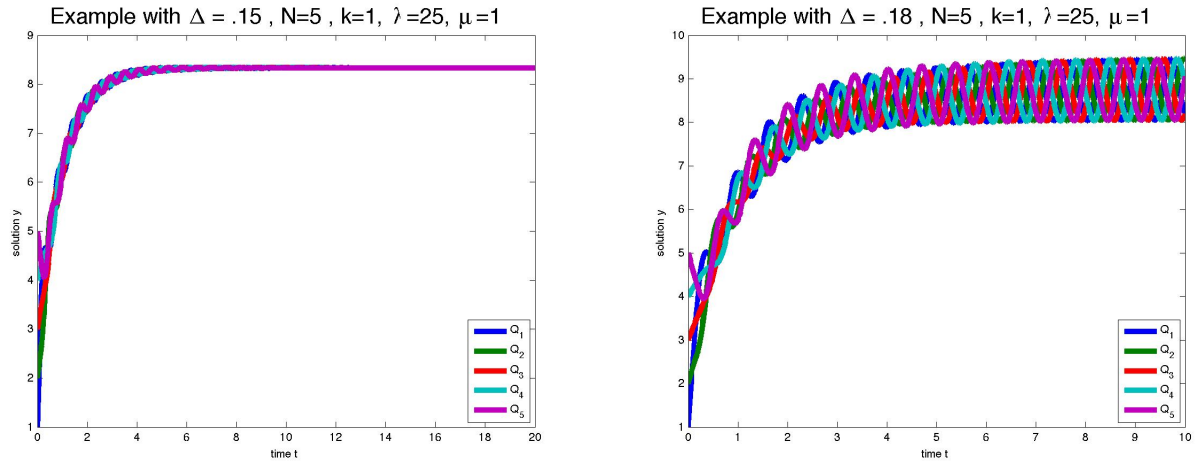

Fig. 3. $\Delta_{cr} = .167$.

show that the fluid limit is a system of delay differential equations. We also prove that the resulting fluid limit can experience a Hopf bifurcation. Using the properties of circulant matrices, we compute exactly when this Hopf bifurcation occurs in terms of our queueing model parameters and verify numerically our results. Although we consider an infinite server system, the analysis of multiserver queues like the Erlang-A queues can also be analyzed in an identical fashion yielding similar results. In fact, the results are identical to the infinite server case as long as the equilibrium does not linger around the number of servers, see for example Pender and Ko [2017]; Ko and Pender [2018].

There are many remaining questions for research. For one, what if the delay were non-stationary and not a constant? What if the topology of the network is more complicated? What if each queue has a different delay? We intend to answer these important questions in future work.

Acknowledgments

Jamol Pender acknowledges the generous support of NSF Career Grant CMMI # 1751975.

Appendix

Before we begin the proof, we present two lemmas that are vital to understanding and constructing the proof via strong approximation theory.

Lemma 3 [Kurtz 1978]. *A standard Poisson process* $\{\Pi(t)\}_{t\geq 0}$ *can be realized on the same probability space as a standard Brownian motion* $\{W(t)\}_{t\geq 0}$ *in such a way that the almost surely finite random variable*

$$Z \equiv \sup_{t\geq 0} \frac{|\Pi(t) - t - W(t)|}{\log(2 \vee t)}$$

has finite moment generating function in the neighborhood of the origin and in particular finite mean.

Lemma 4 [Kurtz 1978]. *For any standard Brownian motion* $\{W(t)\}_{t\geq 0}$ *and any* $\epsilon > 0$, $n \in \mathbb{N}$, *and* $T > 0$

$$\tilde{M} \equiv \sup_{u,v,\leq n\epsilon T} \frac{|W(u) - W(v)|}{\sqrt{|u-v|\left(1 + \log\left(n\epsilon T / |u-v|\right)\right)}} < \infty \quad a.s.$$

3.1. *Proof of Fluid Limit*

In this section we prove Theorem 1, which shows the convergence of the scaled queueing process to our system of delay differential equations.

Proof of Theorem 1

Proof.

$$Q_i^\eta(t) = Q_i^\eta(0) + \frac{1}{\eta}\Pi_i^a\left(\int_0^t \lambda \cdot p_i(Q^\eta(s),\Delta)ds\right) - \frac{1}{\eta}\Pi_i^d\left(\eta\int_0^t \mu Q_i^\eta(s)ds\right). \tag{43}$$

First we need to represent the difference of the scaled stochastic queue length minus the fluid limit. This is given by the following expressions

$$\begin{aligned}
Q_i^\eta(t) - q_i(t) &= Q_i^\eta(0) - q_i(0) + \frac{1}{\eta}\Pi_i^a\left(\eta\int_0^t \lambda \cdot p_i(Q^\eta(s),\Delta)ds\right) - \int_0^t \lambda \cdot p_i(q(s),\Delta)ds \\
&\quad -\frac{1}{\eta}\Pi_i^d\left(\eta\int_0^t \mu Q_i^\eta(s)ds\right) + \int_0^t \mu q_i(s)ds \\
&= Q^\eta(0) - q(0) + \frac{1}{\eta}\Pi_i^a\left(\eta\int_0^t \lambda \cdot p_i(Q^\eta(s),\Delta)ds\right) - \int_0^t \lambda \cdot p_i(Q^\eta(s),\Delta)ds \\
&\quad \int_0^t \lambda \cdot p_i(Q^\eta(s),\Delta)ds - \int_0^t \lambda \cdot p_i(q(s),\Delta)ds \\
&\quad -\frac{1}{\eta}\Pi_i^d\left(\eta\int_0^t \mu Q_i^\eta(s)ds\right) + \int_0^t \mu Q_i^\eta(s)ds - \int_0^t \mu Q_i^\eta(s)ds + \int_0^t \mu q_i(s)ds.
\end{aligned}$$

Now we have a representation of the queue length in terms of centered time changed Poisson processes and a deterministic part, we can now apply the strong approximations theory to the absolute value of the difference.

$$\begin{aligned}
|Q_i^\eta(t) - q_i(t)| &\le \left|Q_i^\eta(0) - q_i(0)\right| + \left|\frac{1}{\eta}\Pi_i^a\left(\eta\int_0^t \lambda \cdot p_i(Q^\eta(s),\Delta)ds\right) - \int_0^t \lambda \cdot p_i(Q^\eta(s),\Delta)ds\right| \\
&\quad + \left|\int_0^t \lambda \cdot p_i(Q^\eta(s),\Delta)ds - \int_0^t \lambda \cdot p_i(q(s),\Delta)ds\right| \\
&\quad + \left|\frac{1}{\eta}\Pi_i^d\left(\eta\int_0^t \mu Q_i^\eta(s)ds\right) - \int_0^t \mu Q_i^\eta(s)ds\right| + \left|\int_0^t \mu Q_i^\eta(s)ds - \int_0^t \mu q_i(s)ds\right|.
\end{aligned}$$

By the Lemma 3, we have the following strong approximation representation of the queue length as

$$\begin{aligned}
Q_i^\eta(t) &= \int_0^t \lambda \cdot p_i(Q^\eta(s),\Delta)ds + \frac{1}{\sqrt{\eta}}\mathcal{B}_i^a\left(\int_0^t \lambda \cdot p_i(Q^\eta(s),\Delta)ds\right) \\
&\quad - \int_0^t \mu Q_i^\eta(s)ds - \frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\left(\int_0^t \mu Q_i^\eta(s)ds\right) + \mathcal{O}\frac{\log\eta}{\eta}.
\end{aligned} \tag{44}$$

Using the strong approximation representation, we now have that the difference between the scaled queue length and the fluid limit is bounded by

$$\begin{aligned}
|Q_i^\eta(t) - q_i(t)| &\le \left|Q_i^\eta(0) - q_i(0)\right| + \left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^a\left(\int_0^t \lambda \cdot p_i(Q^\eta(s),\Delta)ds\right)\right| \\
&\quad + \left|\int_0^t \lambda \cdot p_i(Q^\eta(s),\Delta)ds - \int_0^t \lambda \cdot p_i(q(s),\Delta)ds\right| \\
&\quad + \left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\left(\int_0^t \mu Q_i^\eta(s)ds\right)\right| + \left|\int_0^t \mu Q_i^\eta(s)ds - \int_0^t \mu q_i(s)ds\right| + \mathcal{O}\frac{\log\eta}{\eta}.
\end{aligned}$$

Now it remains to show that

$$\lim_{\eta\to\infty}\sup_{t\le T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^a\left(\int_0^t \lambda \cdot p_i(Q^\eta(s),\Delta)ds\right)\right| = 0 \tag{45}$$

and

$$\lim_{\eta\to\infty}\sup_{t\leq T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\left(\int_0^t \mu Q_i^\eta(s)ds\right)\right| = 0 \tag{46}$$

For the first Brownian motion term we have that

$$\begin{aligned}
&\lim_{\eta\to\infty}\sup_{t\leq T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^a\left(\int_0^t \lambda\cdot p_i(Q^\eta(s),\Delta)ds\right)\right| \\
&\leq \lim_{\eta\to\infty}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^a\left(\lambda\cdot T\right)\right| \\
&= \lim_{\eta\to\infty}\left|\mathcal{B}_i^a\left(\frac{1}{\eta}\cdot\lambda\cdot T\right)\right| \\
&= 0.
\end{aligned}$$

For the second Brownian motion term we have that

$$\begin{aligned}
&\lim_{\eta\to\infty}\sup_{t\leq T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\left(\int_0^t \mu Q_i^\eta(s)ds\right)\right| \\
&\leq \lim_{\eta\to\infty}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\left((Q^\eta(0)+\lambda)\cdot\mu\cdot T\right)\right| \\
&= \lim_{\eta\to\infty}\left|\mathcal{B}_i^d\left(\frac{1}{\eta}\cdot(Q^\eta(0)+\lambda)\cdot\mu\cdot T\right)\right| \\
&= 0.
\end{aligned}$$

Thus, for every $\epsilon > 0$ there exists an η^* such that for all $\eta \geq \eta^*$

$$\left|Q_i^\eta(0) - q_i(0)\right| \leq \epsilon/4, \tag{47}$$

$$\sup_{t\leq T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^a\left(\int_0^t \lambda\cdot p_i(Q^\eta(s),\Delta)ds\right)\right| \leq \epsilon/4, \tag{48}$$

$$\sup_{t\leq T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\left(\int_0^t \mu Q_i^\eta(s)ds\right)\right| \leq \epsilon/4, \tag{49}$$

and

$$\mathcal{O}\frac{\log\eta}{\eta} \leq \epsilon/4 \tag{50}$$

so that we have

$$\begin{aligned}
|Q_i^\eta(t) - q_i(t)| &\leq \left|\int_0^t \lambda\cdot p_i(Q^\eta(s),\Delta)ds - \int_0^t \lambda\cdot p_i(q(s),\Delta)ds\right| \\
&\quad + \left|\int_0^t \mu Q_i^\eta(s)ds - \int_0^t \mu q_i(s)ds\right| + \epsilon \\
&\leq \int_0^t \left|\lambda\cdot p_i(Q^\eta(s),\Delta) - \lambda\cdot p_i(q(s),\Delta)\right|ds + \int_0^t \left|\mu Q_i^\eta(s) - \mu q_i(s)\right|ds + \epsilon
\end{aligned}$$

Since we assumed that the function $g(x)$ was continuously differentiable with uniformly bounded first derivatives, there exists a constant C such that

$$|Q_i^\eta(t) - q_i(t)| \leq C \int_0^t \sup_{-\Delta \leq r \leq s} \left| Q_i^\eta(r) - q_i(r) \right| ds + \epsilon \tag{51}$$

$$\leq C \cdot \left(\int_0^t \sup_{0 \leq r \leq s} \left| Q_i^\eta(r) - q_i(r) \right| ds \right) \tag{52}$$

$$+ C \cdot \left(t \cdot \sup_{-\Delta \leq r \leq 0} \left| Q_i^\eta(r) - q_i(r) \right| \right) + \epsilon.$$

Now we exploit the fact that we assumed that $Q_i^\eta(t) = q_i(t)$ for $t \in [-\Delta, 0]$ for our initial condition. This assumption yields the following new bound for the difference of the scaled queue length and the fluid limit by

$$|Q_i^\eta(t) - q_i(t)| \leq C \int_0^t \sup_{0 \leq r \leq s} \left| Q_i^\eta(r) - q_i(r) \right| ds + \epsilon. \tag{53}$$

Note that the difference between the two equations above is the interval of the supremum inside the integral. Now by invoking Gronwall's lemma in Hale [1969], we have that

$$\sup_{0 \leq t \leq T} |Q_i^\eta(t) - q_i(t)| \leq \epsilon \cdot e^{CT} \tag{54}$$

and since ϵ is arbitrary, we can let it go towards zero and this proves the fluid limit. ■

References

Reza Aghajani and Kavita Ramanan. The hydrodynamic limit of a randomized load balancing network. *arXiv preprint arXiv:1707.02005*, 2017.

Reza Aghajani, Xingjie Li, and Kavita Ramanan. Mean-field dynamics of load-balancing networks with general service distributions. *arXiv preprint arXiv:1512.05056*, 2015.

Reza Aghajani, Xingjie Li, and Kavita Ramanan. The pde method for the analysis of randomized load balancing networks. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1 (2):38, 2017.

Rami Atar and David Lipshutz. Heavy traffic limits for join-the-shortest-estimated-queue policy using delayed information. *Mathematics of Operations Research*, 46(1):268–300, 2021.

Sayan Banerjee, Debankur Mukherjee, et al. Join-the-shortest queue diffusion limit in halfin–whitt regime: Tail asymptotics and scaling of extrema. *The Annals of Applied Probability*, 29(2):1262–1309, 2019.

Maury Bramson, Yi Lu, and Balaji Prabhakar. Randomized load balancing with general service time distributions. In *ACM SIGMETRICS Performance Evaluation Review*, volume 38, pages 275–286. ACM, 2010.

Maury Bramson, Yi Lu, Balaji Prabhakar, et al. Decay of tails at equilibrium for fifo join the shortest queue networks. *The Annals of Applied Probability*, 23(5):1841–1878, 2013.

Amarjit Budhiraja, Debankur Mukherjee, and Ruoyu Wu. Supermarket model on graphs. *arXiv preprint arXiv:1712.07607*, 2017.

John W Byers, Jeffrey Considine, and Michael Mitzenmacher. Geometric generalizations of the power of two choices. In *Proceedings of the sixteenth annual ACM symposium on Parallelism in algorithms and architectures*, pages 54–63. ACM, 2004.

George Cybenko. Dynamic load balancing for distributed memory multiprocessors. *Journal of parallel and distributed computing*, 7(2):279–301, 1989.

Yichuan Ding, Mahesh Nagarajan, and Zhe George Zhang. Asymptotic analysis of multi-queue service systems with dynamic customer choice.

Philip Doldo, Jamol Pender, and Richard Rand. Breaking the symmetry in queues with delayed information. *International Journal of Bifurcation and Chaos*, 31(09):2130027, 2021.

Jing Dong, Elad Yom-Tov, and Galit B Yom-Tov. The impact of delay announcements on hospital network coordination and waiting times. *Management Science*, 2018.

Ahmadreza Faghih-Imani, Robert Hampshire, Lavanya Marla, and Naveen Eluru. An empirical analysis of bike sharing usage and rebalancing: Evidence from barcelona and seville. *Transportation Research Part A: Policy and Practice*, 97:177–191, 2017.

Jack K Hale. Ordinary differential equations. *Pure and Applied Mathematics*, 21, 1969.

Jack K Hale. Functional differential equations. In *Analytic theory of differential equations*, pages 9–22. Springer, 1971.

Refael Hassin. *Rational queueing*. Chapman and Hall/CRC, 2016.

Refael Hassin and Moshe Haviv. *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59. Springer Science & Business Media, 2003.

Young Myoung Ko and Jamol Pender. Strong approximations for time-varying infinite-server queues with non-renewal arrival and service processes. *Stochastic Models*, 34(2):186–206, 2018.

David Lipshutz and Ruth J Williams. Existence, uniqueness, and stability of slowly oscillating periodic solutions for delay differential equations with nonnegativity constraints. *SIAM Journal on Mathematical Analysis*, 47(6):4467–4535, 2015.

Yi Lu, Qiaomin Xie, Gabriel Kliot, Alan Geller, James R Larus, and Albert Greenberg. Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation*, 68 (11):1056–1071, 2011.

Michael Mitzenmacher. How useful is old information? *IEEE Transactions on Parallel and Distributed Systems*, 11(1):6–20, 2000.

Michael Mitzenmacher. The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems*, 12(10):1094–1104, 2001.

Michael Mitzenmacher. Analyzing distributed join-idle-queue: A fluid limit approach. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 312–318. IEEE, 2016.

Debankur Mukherjee, Sem Borst, Johan Van Leeuwaarden, and Phil Whiting. Universality of power-of-d load balancing schemes. *ACM SIGMETRICS Performance Evaluation Review*, 44(2):36–38, 2016.

Debankur Mukherjee, Souvik Dhara, Sem C Borst, and Johan SH van Leeuwaarden. Optimal service elasticity in large-scale distributed systems. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(1):25, 2017.

Debankur Mukherjee, Sem C Borst, Johan SH Van Leeuwaarden, and Philip A Whiting. Universality of power-of-d load balancing in many-server systems. *Stochastic Systems*, 8(4):265–292, 2018.

Samantha Nirenberg, Andrew Daw, and Jamol Pender. The impact of queue length rounding and delayed app information on disney world queues. In *2018 Winter Simulation Conference (WSC)*, pages 3849–3860. IEEE, 2018.

Sophia Novitzky, Jamol Pender, Richard H Rand, and Elizabeth Wesson. Limiting the oscillations in queues with delayed information through a novel type of delay announcement. *Queueing Systems*, 95 (3):281–330, 2020.

Jamol Pender and Young Myoung Ko. Approximations for the queue length distributions of time-varying many-server queues. *INFORMS Journal on Computing*, 29(4):688–704, 2017.

Jamol Pender, Richard H Rand, and Elizabeth Wesson. Queues with choice via delay differential equations. *International Journal of Bifurcation and Chaos*, 27(04):1730016, 2017.

Jamol Pender, Richard Rand, and Elizabeth Wesson. A stochastic analysis of queues with customer choice and delayed information. *Mathematics of Operations Research*, 45(3):1104–1126, 2020.

Jasper Schuijbroek, Robert C Hampshire, and W-J Van Hoeve. Inventory rebalancing and vehicle routing in bike sharing systems. *European Journal of Operational Research*, 257(3):992–1004, 2017.

Qiaomin Xie, Xiaobo Dong, Yi Lu, and Rayadurgam Srikant. Power of d choices for large-scale bin packing: A loss model. *ACM SIGMETRICS Performance Evaluation Review*, 43(1):321–334, 2015.