Introduction

Smartphone technology is changing the paradigm for communication between customers and service systems. One example of this communication is delay announcements, which have become important tools for managers to inform customers of their estimated waiting time. As a result, there is tremendous value in understanding the impact of providing waiting time or queue length information to customers. These announcements can affect the decisions of customers as well as queue length dynamics of the system. Thus, the development of methods to support such announcements and interaction with customers, has attracted the attention of the operations research community and is growing steadily.

Recently, many healthcare providers have started to post their waiting times and queue lengths online, highway billboards, and even through apps. One example of this type of posting is given in Figure 1, which is an online snapshot of the waiting time at JFK Medical Center in Boynton Beach, Florida. In Figure 1, the average wait time is reported to be 12 minutes. However, in the top right of the figure we see that the time of the snapshot was 4:04pm while...
the time of a 12 minute wait is as of 3:44pm. 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.

**Constant Delay Queueing Model**

We present a new stochastic queueing model with customer choice based on the queue length with a constant delay. 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. However, we add the twist that the queue length information that is given to the customer is delayed by a constant $\Delta$ for all of the queues. Therefore, the queue length that the customer receives is actually the queue length $\Delta$ time units in the past. Motivated by the large number of customers, we introduce the following scaled queue length process by a parameter $\eta$

$$Q^n_\eta(t) = Q^n_\eta(0) + \frac{1}{\eta} \Pi_\eta \left( \eta \int_0^t \frac{\lambda \cdot \exp(-Q^n_j(s-\Delta))}{\sum_{j=1}^N \exp(-Q^n_j(s-\Delta))} ds \right) - \frac{1}{\eta} \Pi_\eta \left( \eta \int_0^t \mu Q^n_i(s) ds \right).$$  

(0.1)

We prove a limit theorem regarding the scaled queue length process and show that it is related to a system of delay differential equations (ddes).

**Theorem 0.1.** The sequence of stochastic processes \( \{Q^n(t) = (Q^n_1(t), Q^n_2(t), ..., Q^n_N(t))\}_{\eta \in \mathbb{N}} \) converges almost surely and uniformly on compact sets of time to \( \{q(t) = (q_1(t), q_2(t), ..., q_N(t))\} \) where

$$q_i(t) = \lambda \cdot \frac{\exp(-q_i(t-\Delta))}{\sum_{j=1}^N \exp(-q_j(t-\Delta))} - \mu q_i(t)$$  

(0.2)

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

This result states that as we let $\eta$ go towards infinity, the sequence of queueing processes converges to a system of ddes. Unlike ordinary differential equations, ddes are infinite dimensional and the smallest of delays can cause surprising dynamics. Recent work by Pender et al. [1] shows that Hopf bifurcations can occur in two dimensional queueing systems, however, this work is limited to the two dimensional case and does not immediately generalize to the multi-dimensional setting. Thus, here we generalize the critical delay analysis of Pender et al. [1] and derive the exact bifurcation threshold for $N$ queues.

**Theorem 0.2.** For the constant delay choice queueing model given in Equation (0.2) with arbitrary $N \geq 2$, the critical delay, $\Delta_{cr}(\lambda, \mu, N)$, is given by the following expression

$$\Delta_{cr}(\lambda, \mu, N) = \frac{N \cdot \arccos \left( \frac{-\mu \cdot N}{\lambda} \right)}{\sqrt{\lambda^2 - N^2 \cdot \mu^2}}.$$  

(0.3)

Theorem 0.2 provides a local characterization of the oscillation behavior of a queueing system with $N$ queues. If the delay $\Delta$ is larger than the critical delay $\Delta_{cr}(\lambda, \mu, N)$, then we should expect that the $N$ queues should oscillate in equilibrium and the dynamics are asynchronous. However, if the delay $\Delta$ is smaller than the critical delay $\Delta_{cr}(\lambda, \mu, N)$, then we should expect that the $N$ queues should converge to the limit $\frac{\lambda}{\mu N}$ and not oscillate around the equilibrium.
Numerical Results for Fluid Limits

In Figure 2 we simulate the queue length process and compare it to our fluid approximation. One observes in Figure 2 that our fluid limit does quite well at approximating the mean dynamics given by our simulation.

\[ \lambda = 10, \mu = 1, \Delta_{cr} = 0.3614, \Delta = 0.45, \eta = 100. \]

Figure 2: \( \lambda = 10, \mu = 1, \Delta_{cr} = 0.3614, \Delta = 0.45, \eta = 100. \)

References