When to use Speedup: An Examination of Service Systems with Returns

Carri W. Chan
Decision, Risk, and Operations, Columbia Business School cwchan@columbia.edu

Galit Yom-Tov
Industrial Engineering and Management, Technion gality@tx.technion.ac.il

Gabriel Escobar
Division of Research, Kaiser Permanente gabriel.escobar@kp.org

In a number of service systems, there can be substantial latitude to vary service rates. However, while speeding up service rate during periods of congestion may address a present congestion issue, it may actually exacerbate the problem by increasing the need for rework. We introduce a state-dependent queueing network where service times and return probabilities depend on the 'overloaded' and 'underloaded' state of the system. We use a fluid model to examine how different definitions of 'overload' affect the long-term behavior of the system and provide insight into the impact of using speedup. We identify scenarios where speedup can be helpful to temporarily alleviate congestion and increase access to service. For such scenarios, we provide approximations for the likelihood of speedup to service. We also identify scenarios where speedup should never be used; moreover, in such a situation, an interesting bi-stability arises, such that the system shifts randomly between two equilibria states. Hence, our analysis sheds light onto the potential benefits and pitfalls of using speedup when the subsequent returns may be unavoidable.

Key words: Speedup, State-dependent queues, Erlang-R, Fluid models, Return to service, bi-stability

1. Introduction

We consider a queueing system where the service time of customers can be reduced at the expense of increased likelihood of the need to return to service. We refer to the mechanism of increasing the service rate of customers as *speedup*. The speedup phenomenon can arise in a number of settings such as the Intensive Care Unit (ICU) (Kc and Terwiesch 2012), production lines (Powell and Schultz 2004), email contact centers (Hasija et al. 2010), and general service systems (Ata and Shneorson 2006). The reduction in quality of service due to speedup manifests itself through the need for rework, which we refer to as customer *returns*. This work aims to understand the dynamics of a queueing system where speedup is used and the subsequent customer returns may be unavoidable.

We define the speedup dynamics by an operational control, in the form of a threshold, which specifies whether the system is considered to be overloaded. Hence, service rates and return probabilities are endogenous to the operational speedup control. We introduce a new *multi-server* queueing model where the parameters which define the system dynamics are *congestion dependent*; hence, they depend on the system

state. We examine these state-dependent dynamics using a fluid approximation. In doing so, we are able to characterize the system's stability conditions and long-term behavior.

A number of works have considered the impact of customer returns. For instance, Yom-Tov and Mandelbaum (2013), Jennings and de Véricourt (2008), Yankovic and Green (2011) consider staffing and resource provisioning in (healthcare) service systems with customer returns. de Vericourt and Zhou (2005) and Zhan and Ward (2013) consider routing in call-centers where customers may call-back. None of these works consider how the likelihood of return depends on the service rate, which may be altered depending on system congestion. In this work, we consider the impact of 'speedup' on these customer behaviors.

Also, there have been works which consider state-dependent dynamics (e.g. Armony and Maglaras (2004), Glazebrook and Whitaker (1992), Maglaras and Zeevi (2003), Powell and Schultz (2004)); however, to the best of our knowledge, our work is the first which considers both state-dependent service times *in addition* to state-dependent return probabilities. Combining these two effects reveals new phenomena which have not been previously observed when considering each dynamic separately. More specifically, we show an interesting bi-stability, i.e. the presence of two equilibria, can arise and identify conditions under which speedup is detrimental in the long run.

A number of works consider congestion-dependent service times. Whitt (1990) and Boxma and Vlasiou (2007) consider the steady-state behavior of state-dependent queues where the service times may increase or decrease with delay. From a control standpoint, Ata and Shneorson (2006), Anand et al. (2010) consider the quality-speed tradeoff of an M/M/1 queue and find that speedup can be beneficial. Bekker and Borst (2006), Bekker and Boxma (2007) consider the steady-state distribution and optimal control of single server queues with state-dependent service rates. These works do not consider returns to service. In contrast, our work examines a multi-server model which includes customer returns to service, as well as the increase in return probability due to speedup.

Mandelbaum and Pats (1998), Mandelbaum et al. (1998) consider state-dependent queueing networks with state-dependent routing. The focus of these works is to develop theoretical support for fluid and diffusion approximations of the network dynamics. These works assume state-dependent functions which are continuous and cannot be applied to our model which includes *discontinuities* in the state-dependent dynamics. These discontinuities require a different analytic approach: in this work we utilize fluid approximations and Filippov analysis (Filippov 1988).

We show that speedup–a mechanism which seems to alleviate congestion and increase access to service in a myopic manner–may create more congestion and exacerbate the situation in the long run under certain conditions. More precisely, we show that in some situations, speedup can be a useful operational tool to navigate periods of high congestion. In other instances, speedup will *increase* congestion due to the additional load of returning customers. A surprising bi-stability arises resulting in system dynamics which can

be misleading about whether speedup can help. Therefore, we seek to understand system dynamics under speedup and use this to develop insight into the benefits and pitfalls of using speedup. In analyzing our state-dependent model, we make the following key contributions:

- We introduce a new queueing model (Section 2) which, to the best of our knowledge, is the first such model which incorporates 1) congestion-dependent service times in addition to 2) congestion-dependent return probabilities. The interplay between speedup and customer returns is a phenomenon that has not yet been considered in the literature from an analytic view point.
- We specify conditions for when the queues of our state-dependent queueing system grow without bound (Theorem 4.2). We show that in some cases, speedup can make a stable system unstable; in other cases, speedup is necessary to maintain stability.
- We identify the long-term queueing dynamics and equilibria for our state-dependent queueing system (Section 4). We find that in some cases (Case 1), management can specify the desired system congestion and effective offered load by appropriately tuning the speedup threshold (N^*). Additionally, this implies that congestion is invariant to changes in the number of servers. This analysis provides a possible explanation for the observation of 'supply-sensitive demand' in healthcare, i.e. demand increases with supply.
- We also find that in some cases (Case 2) an interesting bi-stability arises: the long-term dynamics can converge to one of two states, depending on the initial condition. Using simulation, we demonstrate that the stochastic system will oscillate randomly between the two equilibria. This phenomenon demonstrates that while speedup may appear to reduce congestion in some instances (Case 1), its use may be extremely detrimental in other scenarios (Case 2). In such cases, other mechanisms may be necessary to navigate periods of congestion.

The rest of the paper is structured as follows: In Section 2, we present our queueing system which captures the main essence of a system with speedup and its influence on customer returns. We start by examining a system *without* speedup in Section 3. This provides a baseline for exploring the behavior of our system *with* speedup in Section 4. In Section 5, we extend our model to account for factors often seen in various service settings: multi-class customers and time-varying arrivals. We show that in both extensions, the main insights from our original model, such as the bi-stability effect, still hold. Finally, we conclude in Section 6.

2. Queueing model

We now formally introduce our state-dependent queueing model which captures new and returning customers as well as the effect occupancy levels and queue lengths may have on service times and returns.

We consider a queueing network with two stations as depicted in Figure 1. Following the terminology of Yom-Tov and Mandelbaum (2013), we distinguish between two customer states: *Needy* and *Content*. Needy customers require service at Station 1 and are either in service or waiting to begin service. When a Needy customer completes service at Station 1, he will either leave the system or transition into the Content

state. Content customers are customers who currently are being served at Station 2, but upon completion of service, they will transition back (return) to the Needy state and require additional service at Station 1. Station 1 represents a limited resource station with N servers. Station 2 represents an unlimited resource with an infinite number of servers. The service rate and return probability for Needy customers are state dependent and will be defined in the next subsection.

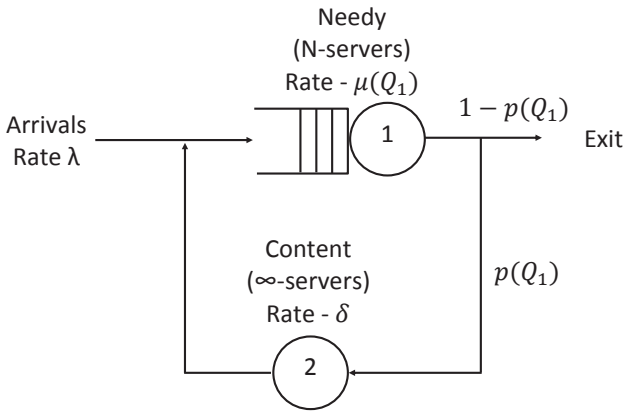


Figure 1 System model: Station 1 represents the N servers where Needy Customers are served. Station 2 represents the servers where Content Customers are served.

2.1. Stochastic Model

We now describe our stochastic model as a Continuous Time Markov Chain (CTMC), where all of the dynamics are Markovian. Let $\mathbb{Q} = (\mathbb{Q}(t), t \geq 0)$ be a two-dimensional stochastic queueing process, where $\mathbb{Q}(t) = (\mathbb{Q}_1(t), \mathbb{Q}_2(t))$: $\mathbb{Q}_1(t)$ is the number of Needy Customers at Station 1 at time t and $\mathbb{Q}_2(t)$ is the number of Content Customers at Station 2. We suppress the dependence on t when it is understood from the context of the discussion.

New **Needy customers** arrive to the system according to a Poisson random process with rate λ. Station 1 has N servers and an unbounded buffer. If a customer is being treated in one of the N servers at Station 1, his service rate, $\mu(\mathbb{Q}_1)$, depends on the number of Needy Customers, $\mathbb{Q}_1$. Customers discharged from Station 1 will return to Station 1 with probability $p(\mu(\mathbb{Q}_1))$.

We distinguish between two service rates: when the system is considered overloaded then Station 1 operates under congested dynamics, with *higher* service and return probabilities than the nominal operation. When Station 1 is not considered overloaded, then the servers operate normally, with nominal service rates and return probabilities. Let $N^* \geq 0$ be a control variable which determines the operation of Station 1. We refer to N^* as the speedup threshold. The system is considered to be overloaded when the number of Needy customers is greater than the speedup threshold, i.e. when $\mathbb{Q}_1 \geq N^*$. Note that if $N^* \leq N$, speedup will begin before a queue forms.

Let μ_L and μ_H ($\mu_H > \mu_L$) define the service rate during underloaded and overloaded periods, i.e. when the occupancy level is *low* and *high*, respectively. The state dependent service rates are given by:

$$\mu(\mathbb{Q}_1) = \begin{cases} \mu_L \ , \text{ if } \mathbb{Q}_1 < N^*, \\ \mu_H \ , \text{ if } \mathbb{Q}_1 \geq N^*. \end{cases} \tag{2.1}$$

Once a customer completes service at Station 1, he exits the system with probability $1 - p(\mu(\mathbb{Q}_1))$ and never requires additional service at Station 1. With probability $p(\mu(\mathbb{Q}_1))$ the customer enters the **Content** state. Content customers transition into the Needy state according to an exponential random variable with constant rate δ. Thus, this station actually models the delay between completion of initial service at Station 1 and the subsequent request for additional service. Note that the return probability, $p(\mu(\mathbb{Q}_1))$, depends on the service rate of the customer at discharge: when speedup is used, the return probability increases. Let p_L and p_H ($p_H > p_L$) denote the return probability during underloaded and overloaded periods:

$$p(\mu(\mathbb{Q}_1)) = \begin{cases} p_L \ , \text{ if } \mathbb{Q}_1 < N^*, \\ p_H \ , \text{ if } \mathbb{Q}_1 \geq N^*. \end{cases} \tag{2.2}$$

Thus, during their stay in the system, customers start in the Needy state and then alternate between Needy and Content states until they depart the system. When a customer becomes Needy and a server at Station 1 is available, he immediately begins service. However, if there are no available servers, customers must wait in a queue for an available one. The queueing policy is FCFS (First Come First Served).

Remark 2.1 *One could also consider return probabilities which depend on the mean congestion in Station 1 during service (e.g. to capture how much work was 'sped-up'). Doing so would require substantial notational overhead to keep track of the congestion each customer experienced and we leave such exploration for future work.*

Remark 2.2 *Note that in practice, $\mathbb{Q}_2$, which only tracks Content customers who eventually transition back to the Needy state, may be an unobservable quantity since we usually do not know which customers will return to service a priori. As such, we restrict our control to depend only on $\mathbb{Q}_1$.*

The queueing system we analyze is a CTMC, which, under conditions for stability (to be described later), has a long-term distribution. We can derive the equations for the equilibrium distribution (see Appendix B) and numerically evaluate or use simulation to compute desired performance measures. However, these methods fail to provide some insights into the behavior of the system, which a fluid analysis can.

2.2. The Fluid Model

In order to enable tractable analysis of the system dynamics of our state-dependent system, we introduce a deterministic fluid approximation to the stochastic model presented in Section 2.1. The fluid model is meant to provide insight into the use of speedup (some of which is unintuitive, as will be seen in Section 4.2).

We denote the fluid function of our queueing network by $Q = \{Q(t), t \geq 0\}$. Here $Q(t) = (Q_1(t), Q_2(t))$, where Q_1 and Q_2 are the fluid content of Needy and Content customers at time t. We derive the fluid formula directly. We assume that arrivals and departures occur deterministically at the specified rates and also regard the number of customers and servers as continuous quantities. Thus, the fluid arrives deterministically and continuously at constant rate λ. Fluid is served in station 1 deterministically at rate $\mu(Q_1)(Q_1 \wedge N)$, where $\wedge$ denotes the minimum function so that $(Q_1 \wedge N)$ is the number of occupied servers in station 1. A $p(Q_1)$ fraction of the fluid is transferred to station 2 after leaving station 1; the rest of the fluid exits the system. The fluid in station 2 is served deterministically and continuously at rate δQ_2. The service rate function, $\mu(\cdot)$, and the return probability function, $p(\cdot)$, are discontinuous in the amount of fluid content of the Needy customers, $Q_1(t)$. These functions are given by (2.1) and (2.2), respectively.

The dynamics of our model can be captured by the following Ordinary Differential Equations (ODE) with *discontinuous* right hand sides:

$$\begin{aligned} \dot{Q}_1(t) &= \lambda + \delta Q_2(t) - 1_{\{Q_1(t) < N^*\}} \mu_L (Q_1(t) \wedge N) - 1_{\{Q_1(t) \geq N^*\}} \mu_H (Q_1(t) \wedge N), \\ \dot{Q}_2(t) &= -\delta Q_2(t) + 1_{\{Q_1(t) < N^*\}} p_L \mu_L (Q_1(t) \wedge N) + 1_{\{Q_1(t) \geq N^*\}} p_H \mu_H (Q_1(t) \wedge N). \end{aligned} \tag{2.3}$$

This discontinuous ODE is discontinuous in Q, but continuous in t. From (2.3), it is easy to see that the derivative values, $\dot{Q}$, which specify the flow dynamics are discontinuous at $Q_1(t) = N^*$. We will analyze the long-term behavior of this fluid system, i.e. the behavior as $t \to \infty$. Let $\bar{q} = (\bar{q}_1, \bar{q}_2)$ be the long-term values such that:

$$\bar{q} \equiv \lim_{t \to \infty} [Q(t) | Q(0) = q_0]$$

Note that these limits may be infinite, may depend on the initial condition q_0, or may not exist. For notational simplicity, hence forth we will omit the dependence on the initial condition and specify explicitly if the limit depends on it.

2.3. Definitions

In our analysis of the long-term behavior of our fluid system, we will require a few definitions. Consider a dynamic system that is dictated by the ODE $\dot{q} = F(q),\ q \in \mathbb{R}^n_+$. In our system, $n = 2$ to capture the Needy and Content customers. We denote by $\Phi(q_0, t)$ the flow at time t, given initial condition q_0. Then, the flow dynamics over time are defined at time t by: $\frac{d}{dt}\Phi(q_0, t) = F(\Phi(q_0, t)),\ \Phi(q_0, 0) = q_0$. The system is considered to be unstable if the fluid content of customers in the system grows without bound over time. Formally,

Definition 2.1 ***Unstable System*** *We say that a system defined by the ODE $\dot{q} = F(q)$ is* unstable *if for any initial condition, q_0:*

$$\lim_{t \to \infty} [Q_1(t) + Q_2(t)] \to \infty$$

In cases where the system is not unstable, we wish to examine the behavior of the system and assess whether there is a limit to which the fluid system might converge to over time. The following definitions for equilibria can be found in di Bernardo et al. (2008).

Definition 2.2 ***Equilibrium (or fixed point)*** *A point $\bar{q}$ is an* equilibrium *of the ODE $\dot{q} = F(q)$ if*

$$\Phi(\bar{q}, t) = \Phi(\bar{q}, 0), \text{ for all } t.$$

The simplest form of equilibrium $\bar{q}$ is one that satisfies $F(\bar{q}) = 0$. Following di Bernardo et al. (2008), we call a *pseudo-equilibrium* an equilibrium that arises on the region of discontinuity in the ODE (e.g. on the switching boundary $\Sigma \equiv \{Q : Q_1 = N^*\}$ of (2.3)). This type of equilibrium is an equilibrium since a trajectory starting as that point will stay there, but it is different from standard equilibria because the derivatives may not be zero. This form of equilibrium happens when the forces that push the trajectory to this point are equal from all directions. Technical details of pseudo-equilibria are given in the Appendix.

Note that even if an equilibrium, $\bar{q}$, exists, it is not necessarily true that the system will converge to it as $t \to \infty$. Moreover, the limiting behavior may depend on the initial condition $q_0 \in \mathbb{R}^2_+$. Hence, we further differentiate between types of equilibria. An equilibrium is called *Lyapunov* stable if trajectories starting nearby to the equilibrium remain nearby for all time. This type of equilibrium is often referred to as a Locally Stable Equilibrium. Without loss of generality, we assume the equilibrium is at the origin; that is, $\Phi(0, t) = \Phi(0, 0)$ for all t.

Definition 2.3 ***(Lyapunov) Locally Stable Equilibrium*** *The origin is said to be (Lyapunov) locally stable if for any $\epsilon > 0$, there exists a $\delta > 0$ such that if*

$$\|q_0\| < \delta \Rightarrow \|\Phi(q_0, t)\| < \epsilon, \ \forall t > 0$$

We refer to an equilibrium as *Globally Stable* if for *every* starting point it will converge to the same stable equilibrium defined by Definition 2.3.

Definition 2.4 ***Globally Stable (in the sense of Lyapunov) Equilibrium*** *The origin is said to be globally stable (in the sense of Lyapunov) if the following two conditions hold:*

1. *It is locally stable;*
2. *For all initial conditions, q_0:* $\lim_{t\to\infty} \Phi(q_0, t) = 0$

Note that these definitions of stability do not mean that there exists a t_0 such that $Q(t) = \bar{q}$ for all $t > t_0$. They simply require that for every $\epsilon > 0$, there exists a t_0 such that for all $t > t_0$, $Q(t)$ is within ϵ of $\bar{q}$; in the case of local stability this is only true if the trajectory starts close enough to the equilibrium. We will actually see instances (for pseudo-equilibria) where the fluid oscillates with arbitrarily small fluctuations around the equilibrium point. Finally, we remark that there could exist long-term behavior which is not captured by Definitions 2.1-2.4, e.g. a trajectory could remain finite, but not converge to any single state.

3. A System Without Speedup

For comparison purposes, we first consider a system where speedup is never used. In this case, the fluid equations can be simplified to:

$$\begin{aligned}\dot{Q}_1(t) &= \lambda + \delta Q_2(t) - \mu_L(Q_1(t) \wedge N),\\ \dot{Q}_2(t) &= -\delta Q_2(t) + p_L\mu_L(Q_1(t) \wedge N).\end{aligned} \tag{3.1}$$

This is the fluid model of an Erlang-R queue (Yom-Tov and Mandelbaum (2013)). The queues grow without bound if $N < \frac{\lambda}{(1-p_L)\mu_L}$; otherwise, the system converges to a globally stable equilibrium, $\bar{q}$. More formally:

Theorem 3.1 *The dynamics of the fluid system described in* (3.1) *can be summarized as:*

1. *The system is unstable if* $N < \frac{\lambda}{(1-p_L)\mu_L}$*;*
2. *If* $N > \frac{\lambda}{(1-p_L)\mu_L}$*, the fluid converges to the following globally stable equilibrium:*

$$\lim_{t\to\infty} Q(t) = \bar{q} = \left(\frac{\lambda}{(1-p_L)\mu_L}, \frac{\lambda p_L}{(1-p_L)\delta}\right).$$

The proof of this result can be found in the Appendix. We note that if $N = \frac{\lambda}{(1-p_L)\mu_L}$, there are an uncountable number of equilibria. As an example, if the initial condition is such that $(q_0)_1 \geq N$ and $(q_0)_2 = \frac{\lambda p_L}{(1-p_L)\delta}$, then the fluid content stays at the initial condition, so $Q(t) = q_0$ for all t.

4. Analysis of System Dynamics

In this section, we analyze the long-term dynamics of the fluid model presented in Section 2.2. The main challenge is the discontinuity at $Q_1 = N^*$. The long-term dynamics are highly dependent on system parameters for arrival rate, service times, and return probabilities as well as the control variable for when to begin speedup, N^*.

To start, we leverage results from Filippov (1988) to establish the existence of a solution to our ODE.

Theorem 4.1 *There exists a solution of the problem defined by the ODE* (2.3) *for any initial condition* $q_0 = Q(0) \in [0, Q_{\max}] \times [0, Q_{\max}]$ *where* $Q_{\max} < \infty$ *is an arbitrary finite constant.*

This is a result of Theorem 1 on page 77, Chapter 2, Section 7 of Filippov (1988). The details of leveraging this result can be found in the Appendix.

We define the following parameters, which will be useful in describing the system dynamics:

$$q^L = \left(\frac{\lambda}{(1-p_L)\mu_L}, \frac{p_L\lambda}{(1-p_L)\delta}\right), \quad q^H = \left(\frac{\lambda}{(1-p_H)\mu_H}, \frac{p_H\lambda}{(1-p_H)\delta}\right), \quad \alpha = \frac{(N^* \wedge N) - q_1^H}{q_1^L - q_1^H}. \tag{4.1}$$

One can think of q^L and q^H as the offered load at Station 1 and 2 under low and high occupancy dynamics. This interpretation is clear when considering the system either i) *always* works under underloaded dynamics and so *never* speeds-up (i.e. the system analyzed in Section 3) or ii) *never* works under underloaded dynamics and so *always* speeds-up.

We begin our analysis with the question of when our system is unstable. The proof is given in the Appendix.

Theorem 4.2 *The instability conditions for the fluid system in* (2.3) *are broken into two cases.*

1. $q_1^H < q_1^L$.
   - *The system is unstable if* $N < q_1^H$.
   - *The system is unstable if* $N < q_1^L$ *and* $N^* = \infty$ *(i.e. speedup is never used)*
2. $q_1^H \geq q_1^L$.
   - *The system is unstable if* $N < q_1^L$.
   - *The system is unstable if* $N < q_1^H$ *and* $N^* < q_1^L$.

We will show in Theorem 4.3 that when the conditions of Theorem 4.2 are not satisfied, the system will converge to a finite equilibrium.

Note that the stability of the system depends on both system parameters (q_1^H, q_1^L, N) and the decision variable, i.e., the speedup threshold (N^*). Consequently, there are cases in which the system can be stabilized only if speedup is applied (e.g. under Case 1 if $q_1^H < N \leq q_1^L$); in such cases using speedup reduces the offered load so that it is not necessary to acquire additional servers to ensure that the queues do not grow without bound. On the other hand, there are cases where an otherwise stable system becomes unstable due to utilizing speedup (e.g. under Case 2 if $N^* < q_1^L < N < q_1^H$).

We now consider the long-term dynamics of our system. In the results which follow, we assume that N is large enough such that the queues in our system do not explode, i.e. the conditions of Theorem 4.2 are violated. Moreover, due to the potential for an uncountable number of equilibria of our fluid equations (as described in Section 3), we make the following assumption:

Assumption 4.1 *The number of servers,* N*, is such that the effective system load is strictly less than* 1*, i.e.*

$$N > \left(q_1^L \wedge q_1^H\right).$$

We then consider how the various system parameters impact the system. In particular, we identify scenarios where there is a unique, globally stable equilibrium as well as other scenarios where there may be multiple locally stable equilibria.

Theorem 4.3 *Given* N *and* N^* *such that Assumption 4.1 holds and the conditions of Theorem 4.2 are violated, the long-term dynamics of the fluid system in* (2.3) *can be broken in two cases with additional subcases:*

1. $q_1^H < q_1^L$

   1.1 If $N^* \leq q_1^H$*, then* q^H *is a globally stable equilibrium.*

1.2 If $q_1^H < N^* \wedge N < q_1^L$ then $(N^*, \alpha q_2^L + (1-\alpha) q_2^H)$ is a globally stable pseudo-equilibrium.

1.3 If $q_1^L \leq N^* \wedge N$, then q^L is a globally stable equilibrium.

2. $q_1^H \geq q_1^L$

2.1 $q_1^H < N$

2.1.1 If $N^* < q_1^L$, then q^H is a globally stable equilibrium.

2.1.2 If $q_1^L \leq N^* \leq q_1^H$, then q^L and q^H are locally stable equilibria. In addition, when $q_1^L \neq q_1^H$, then $(N^*, \alpha q_2^L + (1-\alpha) q_2^H)$ is a pseudo-equilibrium.

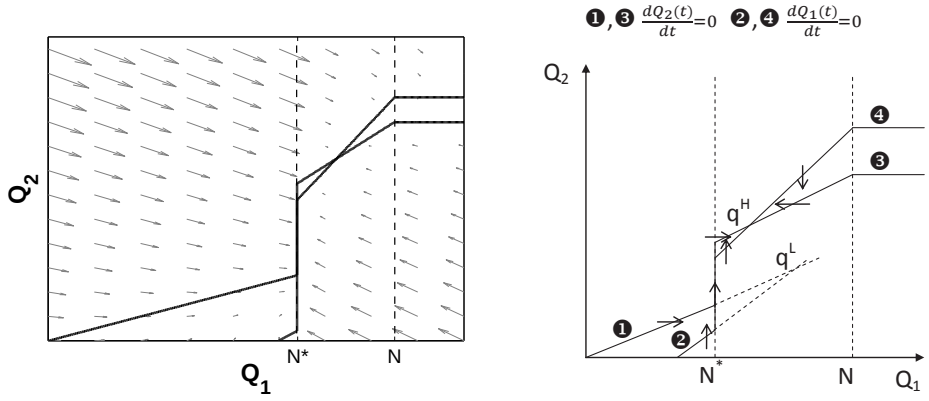
2.1.3 If $q_1^H < N^*$, then q^L is a globally stable equilibrium

2.2 $q_1^H \geq N > q_1^L$ and $N^* > q_1^L$. Then q^L is a locally stable equilibrium and $(N^*, \alpha q_2^L + (1-\alpha) q_2^H)$ is a pseudo-equilibrium.

The proof follows by Filippov and Lyapunov techniques and is given in the Appendix. We demonstrate the intuition behind the result for Case 1 via the phase portrait of each subcase with $N^* \leq N$ (The case for $N^* > N$ follows similarly). In Figure 2(a), we see the phase portrait when the equilibrium is at q^H–the arrows represent the magnitude and direction of the derivative at each state, while the solid lines represent points where the derivative is zero in one of the dimensions. Hence, the trajectory of the queueing system is pulled towards and along these lines. Figure 2(b) breaks down the phase portrait in Figure 2(a) to present a clearer view of the relationship between the different parameters. The dashed lines are a virtual continuation of the derivative lines. It is not necessarily the case that $\dot{q_1} = 0$ or $\dot{q_2} = 0$ along these lines due to the fact that the system dynamics change when crossing the N^* threshold. If the dynamics did not change, q^L, would be an equilibrium. However, because of the change in dynamics due to the speedup threshold, q^L is not an actual equilibrium in this case. Thus, we refer to q^L as an *inadmissable* equilibrium. Intuitively, when $Q_1 < N^*$, the system does not speedup and the trajectory is attracted to the point, q^L. Before reaching q^L, the number of Needy customers grows so that $Q_1 \geq N^*$ and speedup is used. At this point, the system dynamics switch to the overloaded dynamics and the trajectory is attracted to the point q^H. Because $N^* \leq q^H < N$ in Case 1.1, the derivatives at q_H, $\dot{q_1} = \dot{q_2} = 0$. We thus refer to q^H as an *admissible* point and can conclude it is the equilibrium point of the system.

This intuition can be extended to Case 1.2 and 1.3. The stripped down phase portraits for these cases are in Figure 3, which depict the pull of two points which attract trajectories: q^L and q^H. Each point represents the equilibrium when the system never or always speeds up. The relationship between q^L, q^H and the speedup threshold, N^*, dictates whether the equilibrium is at q^H (Case 1.1), q^L (Case 1.3), or N^* in which case the trajectories oscillate across the switching boundary between the speedup/no-speedup regions (Case 1.2). Similar phase portraits can be generated for Case 2.

To understand the impact of different parameters on the equilibrium values, we use bifurcation diagrams. Bifurcation diagrams are often used to show the possible long-term values (equilibria or periodic orbits) of



(a) (b)

Figure 2 Phase portrait for Case 1.1 ($N^* \leq N$): Dark lines represent points where the derivative is zero in one of the dimensions.

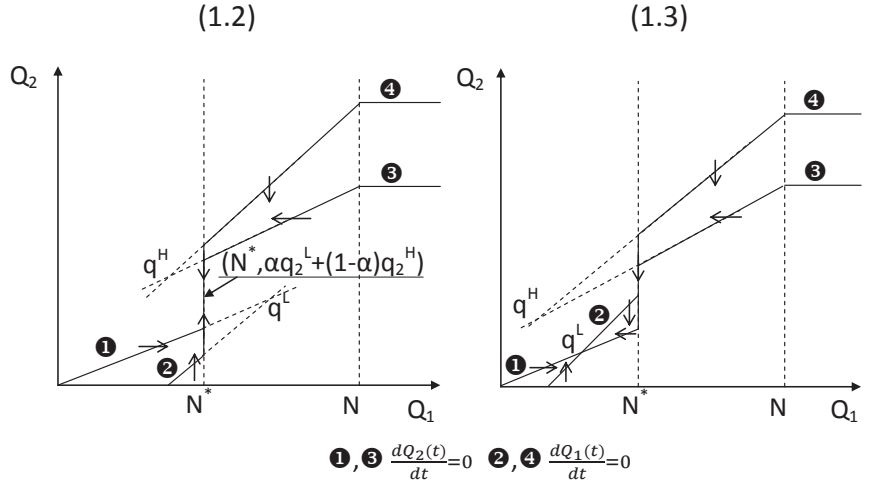
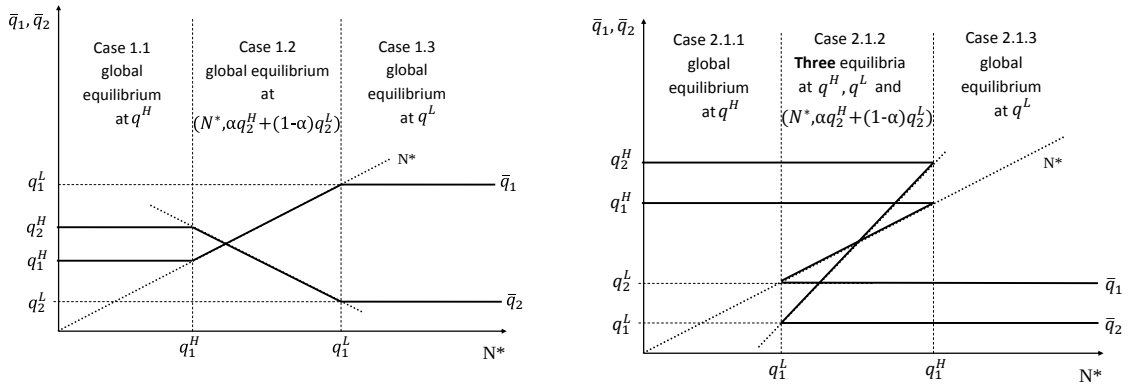


Figure 3 Phase portraits for Case 1.2 and 1.3 ($N^* \leq N$).

a dynamical system as a function of a parameter which may dictate the system's behavior. In our case, our main interest is in understanding how the speedup threshold, N^*, affects the equilibria (in Section 5.2 we will see cases where the long-term values are actually periodic orbits). To examine the influence of N^*, we assume that all other parameters, including the number of servers, are fixed. For consistency, we consider the case where $N^* \leq N$. The case of $N^* > N$ follows very similarly, assuming Assumption 4.1 holds. Figure 4(a) summarizes the equilibria for Case 1 as a function of N^*. The long-term number of Needy customers, $\bar{q}_1$, increases with N^*, while the number of Content customers, $\bar{q}_2$, decreases with N^*. When N^* is larger than q_1^L, no speedup is applied; when N^* is smaller than q_1^H, speedup is applied most of the time. Finally, in the middle range ($q_1^H \leq N^* \leq q_1^L$), speedup is applied a fraction of the time (therefore only some of the customers will be sped-up). This graph demonstrates that N^* is not only the threshold of speedup, but also the equilibrium of the system. From Theorems 3.1 and 4.2, we recognize that, in Case 1, $N > q_1^L$

guarantees the queue does not grow without bound irrespective of whether or not speedup is used. However, by utilizing speedup, we can achieve a long-term backlog of $\bar{q}_1 < q_1^L$ and maintain finite queues with *fewer* servers. Hence, in Case 1, utilizing speedup (i.e., reducing N^*) increases access to service by reducing the overall workload on Station 1, despite the increase in readmission likelihood.

Figure 4(b) summarizes the equilibria for Case 2.1 as a function of the parameter N^*. In this case, both the number of Needy and Content customers is higher when utilizing speedup compared to never using it. While speedup may seem like a reasonable action to take during periods of congestion, it is a myopic action which can exacerbate congestion issues in the long run. Hence, unlike Case 1, it is *undesirable* to utilize speedup as it can increase the overall load on Station 1, which is already congested.



(a) Case 1 (b) Case 2.1

Figure 4 Bifurcation diagram as the speedup threshold, N^*, varies.

At the extremes (high/low N^*) when speedup is *always* or *never* used, the basic insights from Case 1 and Case 2 are not surprising. However, because systems may elect to operate at intermediary values where speedup is used some of the time (Case 1.2 and Case 2.1.2), it is important to further understand the dynamics in these regions.

4.1. Case 1.2: $q_1^L > q_1^H$

We now discuss a number of interesting insights which can be extracted by our analysis of Case 1.2. Recall that in this case, speedup can increase access to service.

We first examine the impact of the number of servers, N, on the system dynamics. Fix an occupancy threshold, $0 \leq r$ at which speedup begins; hence, $N^* = rN$. Figure 5 demonstrates the long-term behavior as we vary the number of servers, but maintain the speedup threshold at $N^* = rN$. This introduces an interesting phenomenon where adding more servers does not seem to reduce congestion. More specifically, as the number of servers, N, increases, the occupancy level at Station 1, Q_1/N, remains at r. This is because

N^* is not only the threshold of speedup, but also the resulting equilibrium of the system. Hence, Station 1 still seems 'busy' even with the addition of servers. Though adding servers doesn't appear to reduce congestion, it does result in fewer customers who are sped-up. Our analysis suggests that large additions may be required before there will be any noticeable change in occupancy levels.

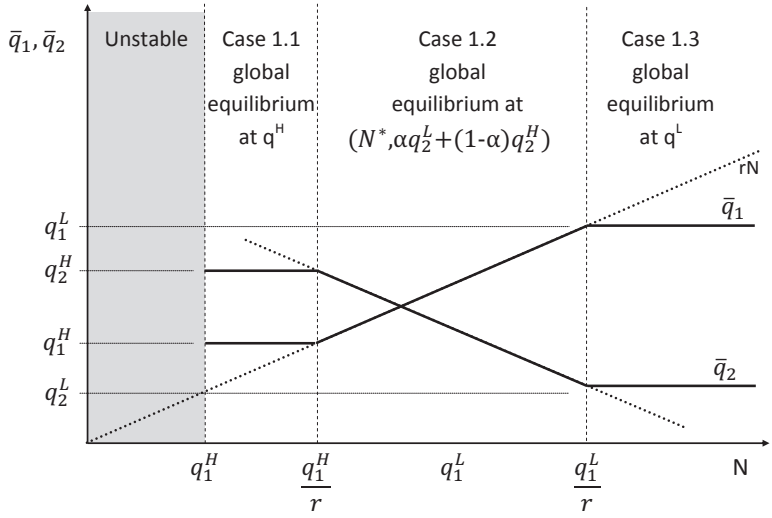


Figure 5 Bifurcation diagram of Case 1 as the number of servers, N, varies; Speedup begins at occupancy level $r \leq 1$. (The diagram is similar for $r > 1$.)

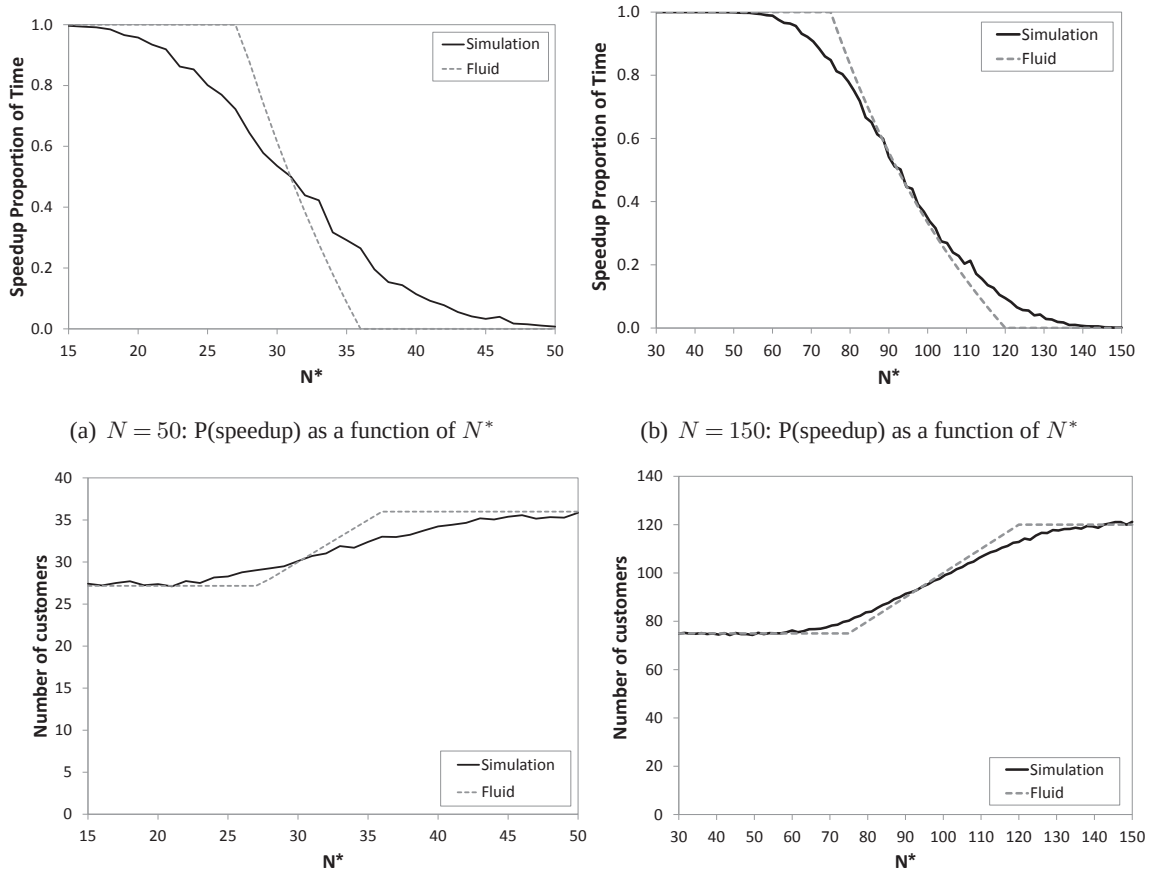
We now delve further into the behavior of the system in Case 1.2, where it oscillates frequently between overloaded and underloaded regions. Note that these fluctuations are arbitrarily small such that the fluid state remains close to the globally stable pseudo-equilibrium. Hence, while the derivatives are non-zero, the system is arbitrarily close to the equilibrium point. As a consequence of the proof of Theorem 4.3, we can establish the proportion of time spent in overload and underload when the system oscillates between these two regions.

Corollary 4.1 *If the fluid system is stable and $q_1^H \leq N^* \wedge N \leq q_1^L$, then the proportion of time the fluid process spends speeding up is given by:*

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T 1_{\{Q_1(t) \geq N^*\}} = \frac{\lambda + \delta \bar{q_2} - \mu_L(\bar{q_1} \wedge N)}{(\mu_H - \mu_L)(\bar{q_1} \wedge N)} = \frac{\lambda + \delta(\alpha q_2^L + (1-\alpha) q_2^H) - \mu_L(N^* \wedge N)}{(\mu_H - \mu_L)(N^* \wedge N)}. \quad (4.2)$$

This corollary is based on Filippov's convex method (Filippov 1988) which provides expressions for the proportion of time a trajectory spends above the switching boundary. This proportion–from the fluid model–can be used as an approximation for the probability of speedup in our original stochastic model, i.e.

$$P(Speedup) \equiv P(\mathbb{Q}_1(t) \geq N^*) \approx \lim_{T\to\infty} \frac{1}{T} \int_0^T 1_{\{Q_1(t) \geq N^*\}}$$



(a) $N = 50$: P(speedup) as a function of N^*

(b) $N = 150$: P(speedup) as a function of N^*

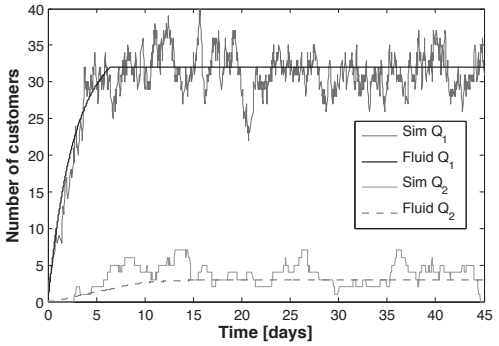
(c) $N = 50$: Q_1 as a function of N^*

(d) $N = 150$: Q_1 as a function of N^*

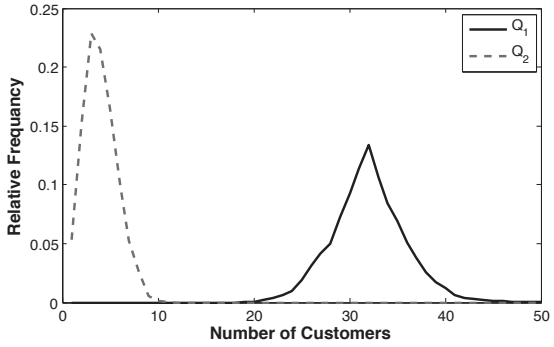
Figure 6 Case 1 ($q_1^H \leq N^* \leq q_1^L$): Simulation vs. Fluid.

We simulate the long-term behavior of our original stochastic system and compare it to our fluid approximation. Figure 6(a)-(b) shows the probability of speedup as we vary N^* for both the simulation and the fluid approximation. We use parameters which satisfy the criteria for Case 1: a) A small system with $N = 50$ servers and $\mu_L = .0164, \mu_H = .0224, p_L = .0667, p_H = .0973, \delta = .0161$[1] and b) A large system with $N = 150$ servers and $\mu_L = .01, \mu_H = .02, p_L = .5, p_H = .6, \delta = .01$. We observe that for large N the fluid is very accurate; this accuracy degrades as the size of the system decreases and when N^* is close to q_1^H or q_1^L. This is due to the non-smooth dynamics of the fluid approximation when N^* is relatively large or small. This phenomenon also arises when considering the expected number of Needy customers, $E[\mathbb{Q}_1]$, as seen in Figure 6(c)-(d). Upon further investigation, we noticed that the fluid model provides a more accurate estimate for the *mode* of $\mathbb{Q}_1$, i.e. the most frequently observed value of $\mathbb{Q}_1$. $\mathbb{Q}_1$ typically does not have a symmetric distribution, so $E[\mathbb{Q}_1]$ is not necessarily equal to the mode of $\mathbb{Q}_1$. As the system gets larger, the symmetry of the distribution increases, so the fluid approximation improves.

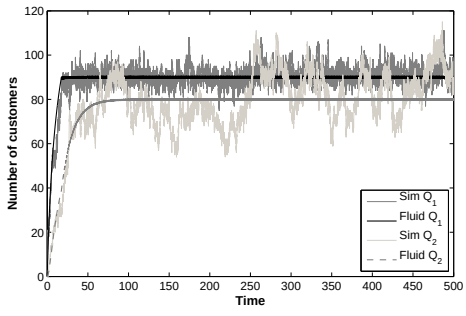
[1] These parameters were inspired by an ICU application.



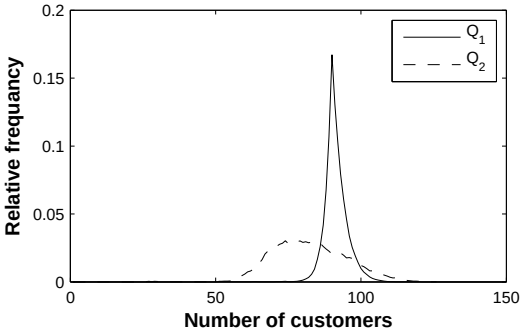
(a) A Single Sample Path for $N = 50$.



(b) Steady state distribution of $\mathbb{Q}_1(t)$ and $\mathbb{Q}_2(t)$ for $N = 50$.



(c) A Single Sample Path for $N = 150$.



(d) Steady state distribution of $\mathbb{Q}_1(t)$ and $\mathbb{Q}_2(t)$ for $N = 150$.

Figure 7 Case 1 Simulation: $q_1^H \leq N^* \leq q_1^L$.

We next examine the variation of our stochastic process with respect to the fluid approximation. Figures 7(a) and 7(c) show a sample path of the system in Case 1.2, and Figures 7(b) and 7(d) show the long-term distribution of $\mathbb{Q}$ (using the same parameters as before for the small and large systems). In this case, the equilibrium of the fluid model is exactly $\bar{q}_1 = N^*$. When considering the stochastic model, we observe the distribution for $\mathbb{Q}_1(t)$ has an unusual shape–similar to a bilateral exponential distribution–that is tight around the threshold N^* and can be observed as rapid changes in the sample path. On the other hand, $\mathbb{Q}_2(t)$ exhibits the typical Poisson distribution (this is more visible in the larger system). The rapid changes in $\mathbb{Q}_1(t)$ suggest a very strong *pull* towards the equilibrium from above and below the equilibrium N^* for Needy customers. This observation suggests that the methodology considered in Perry and Whitt (2011), which also observes tight drifts for a different queueing system, could be used to generate an approximation for the distribution of $\mathbb{Q}_1$.

4.1.1. Approximating $\mathbb{Q}$ under Case 1.2 Following ideas from Perry and Whitt (2011), we develop an approximation to our original stochastic process $\mathbb{Q}_1$ while operating under Case 1.2 conditions. Such an approximation provides insight into the behavior of the variation of the queue length process, which the

fluid system does not allow. We develop the approximation as a heuristic. We consider an approximation with a very simple structure: a two-sided birth-death process with constant rates on each side. Because of this simple structure, we are able to easily derive approximations for the steady-state distribution of $\mathbb{Q}_1$ as well as provide an approximation for the probability of speedup in our original stochastic model, $P(\mathbb{Q}_1 \geq N^*)$. While the approximation for P(speedup) from this approach is the same as the one developed in Corollary 4.1 using the Filippov method for the fluid model, we now also have more detailed insight into the distribution of the number of Needy customers in our original stochastic model than when considering the results of the fluid analysis alone.

Define a CTMC process $\tilde{Q} = (\tilde{Q}(t),\ t \geq 0) \in \mathbb{R}$. Let λ^+ and μ^+ be the birth and death rates of $\tilde{Q}(t)$ when $\tilde{Q}(t) \geq \bar{q}_1$ and λ^- and μ^- be the birth and death rates when $\tilde{Q}(t) < \bar{q}_1$. Our approximation defines these rates as:

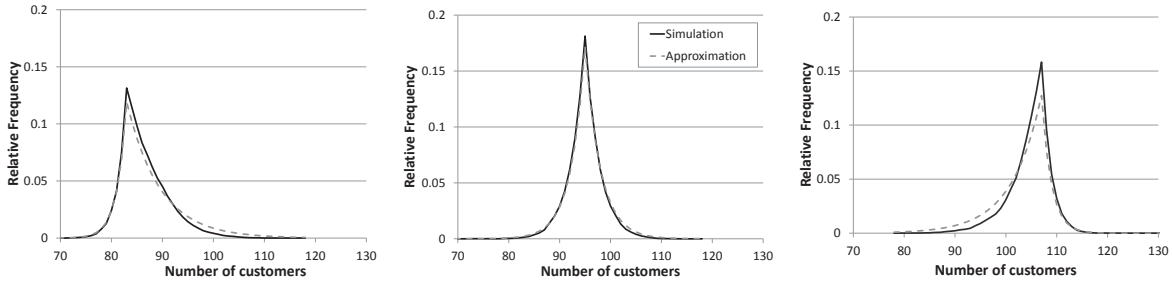
$$\begin{aligned} \lambda^+ &= \lambda + \delta\bar{q}_2 \\ \mu^+ &= \mu_H(\bar{q}_1 \wedge N) \\ \lambda^- &= \lambda + \delta\bar{q}_2 \\ \mu^- &= \mu_L(\bar{q}_1 \wedge N). \end{aligned} \tag{4.3}$$

Because of the constant birth and death rates, the process $\tilde{Q}(t)$ evolves as an M/M/1 queue in each of the regions $\tilde{Q}(t) \geq \bar{q}_1$ and $\tilde{Q}(t) < \bar{q}_1$. This allows us to easily determine the steady state probability of being in state i:

$$\begin{aligned} P(\tilde{Q} = \bar{q}_1 - 1) &= \frac{\left(1 - \frac{\lambda^+}{\mu^+}\right)\left(1 - \frac{\mu^-}{\lambda^-}\right)}{1 - \frac{\lambda^+}{\mu^+}\frac{\mu^-}{\lambda^-}} \\ P(\tilde{Q} = i) &= \begin{cases} \left(\frac{\mu^-}{\lambda^-}\right)^{-(i-\bar{q}_1+1)} \times P(\tilde{Q} = \bar{q}_1 - 1)\ , \text{ if } i < \bar{q}_1 - 1, \\ \left(\frac{\lambda^+}{\mu^+}\right)^{i-\bar{q}_1+1} \times P(\tilde{Q} = \bar{q}_1 - 1) \quad , \text{ if } i > \bar{q}_1 - 1, \end{cases} \end{aligned} \tag{4.4}$$

The intuition behind this process construction is as follows: The stochastic process $\mathbb{Q}$ we are trying to approximate, has state-dependent drifts depending on the number of customers in service; however, we observed in Figures 7(c) that that number of Needy customers is almost deterministic and equal to $\bar{q}_1 = N^*$. Hence, we remove the state-dependency and instead use constant drifts in the process $\tilde{Q}$, similar to a single server queue rather than the N-server queue we are approximating. The death rates differ on each side because of speedup; speedup is used when $\tilde{Q} \geq \bar{q}_1$, while speedup is not used when $\tilde{Q} < \bar{q}_1$. As a result, the rates of $\tilde{Q}$ are the same as the process $\mathbb{Q}$ if the number of customers were fixed $\mathbb{Q} = \bar{q} = (N^*, \alpha q_2^L + (1 - \alpha)q_2^H)$. This is irrespective of what the actual queue length $\mathbb{Q}$ is and allows us to derive simple expressions for the distribution of $\tilde{Q}$, as given in (4.4).

Previously, we used the fluid model to provide an approximation for the probability of speedup in our original stochastic system. We now consider a different approximation approach, which uses the process



(a) 75% (b) 50% (c) 25%

Figure 8 Case 1: Simulation vs. Approximation based on Two-sided M/M/1 queue.

$\tilde{Q}(t)$ to approximate $\mathbb{Q}_1(t)$; thus, we measure when $\tilde{Q}(t)$ is equal to or greater than N^*. Therefore, $P(\mathbb{Q}_1 \geq N^*) \approx P(\tilde{Q} \geq \bar{q}_1) = \sum_{i=\bar{q}_1}^{\infty} \left(\frac{\lambda^+}{\mu^+}\right)^{i-\bar{q}+1} \times P(\tilde{Q} = \bar{q}_1 - 1) = \frac{\frac{\lambda^+}{\mu^+}\left(1-\frac{\mu^-}{\lambda^-}\right)}{1-\frac{\lambda^+}{\mu^+}\frac{\mu^-}{\lambda^-}}$. Using (4.3), and noting that $\lambda^+ = \lambda^-$, gives:

$$P(speedup) = P(\mathbb{Q}_1 \geq N^*) \approx \frac{\lambda^+ - \mu^-}{\mu^+ - \mu^-} = \frac{\lambda + \delta\bar{q}_2 - \mu_L(\bar{q}_1 \wedge N)}{(\mu_H - \mu_L)(\bar{q}_1 \wedge N)}.$$

This is exactly the same approximation as from Corollary 4.1.

Figure 8 compares the steady state distribution of our approximation, $\tilde{Q}$, to the simulated distribution of the original process $\mathbb{Q}_1$ in various cases. As expected, the fit is very good when N^* is such that we expect the speedup probability should be close to 50%. As we deviate from that value of N^* (e.g., when the speedup probability is close to 25% or 75%), the fit degrades. Earlier, we observed in Figure 4.1 that the fluid model provides a very accurate approximation when P(speedup) is close to 50%, but its accuracy degrades as N^* approaches q_1^L or q_1^H (equivalently, as P(speedup) approaches 0 or 1). We expect this inaccuracy to also arise as we consider our approximation for the whole distribution for $\mathbb{Q}_1$. The fact that the shape of the distribution is still quite accurate in the latter cases suggests that with improved approximations for $\bar{q}$, the approximation for the distribution of $\mathbb{Q}_1$ could also improve.

4.2. Case 2.1.2: $q_1^L \leq q_1^H$

We now examine the analogous scenario in Case 2–Case 2.1.2–and consider the insights our fluid analysis provides for our original stochastic system. There are *three* equilibria in Case 2.1.2. However, the equilibrium $(N^*, \alpha q_2^L + (1-\alpha)q_2^H)$ is not stable. That is, if the fluid starts there, it stays there; however, even small deviations in the initial conditions from the equilibrium will drive the system away from it. Hence, it is unlikely to be observed in our original stochastic system. The other two equilibria, q^H and q^L, are locally stable. Hence, whether speedup can alleviate congestion at Station 1 or whether it will lead to worse congestion resulting in perpetual overload (even if the system could be operated in underload without using speedup) will depend on the initial condition. In the stochastic model, the behavior of the queues will depend on the distance between q^H and q^L. If they are very far from each other, the steady state of the stochastic

system will primarily depend on the initial condition. By starting near q^L, speedup will not need to be used; however, starting near q^H will require that speedup is always used. Even if q^L and q^H are far away from each other, there exists sample paths such that the number of Needy customers will increase (decrease), thereby effectively increasing (decreasing) the system load and transitioning to state q^H (q^L). For example, a transition from q^L to q^H may occur due to a 'burst' of arrivals. Because of stochastic fluctuations, it is possible that the stochastic queue will oscillate between q^H and q^L. If these two equilibria are very far apart, the transition times in the stochastic system could be very long–long enough that such transitions are never observed in practice. However, if the equilibria are close to one another, small bursts will be sufficient to cause the stochastic system to transition and so it may oscillate between the two equilibria frequently. As an example, we chose to demonstrate a scenario where both locally stable equilibrium coexist.

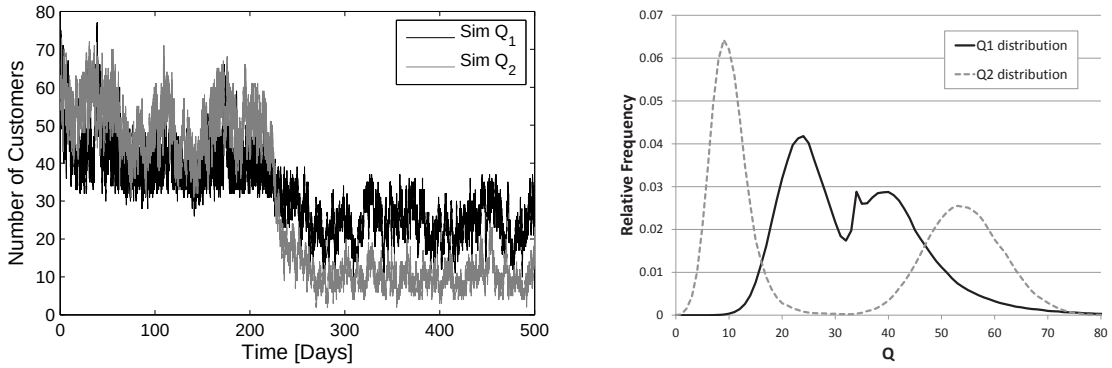
Figure 9(a) presents a sample path of the stochastic state $\mathbb{Q}(t) = (\mathbb{Q}_1(t), \mathbb{Q}_2(t))$, under Case 2.1.2. We observe shifting from one equilibrium to the second one in the middle of the run, after approximately 220 days[2] . The system begins around the q^L equilibrium and shifts to the q^H equilibrium. When examining the distribution of $\mathbb{Q}(t)$ in Figure 9(b), we observe the two equilibria at $q^L = (24, 9.6)$ and $q^H = (40, 54.4)$. Interestingly, there is another peak at $\mathbb{Q}_1 = N^* = 35$. This peak does not indicate the pseudo-equilibrium, but rather is a product of the system shifting from one region to the next. During the transition, when the fluid flow encounters the switching boundary Σ (where $Q_1 = N^*$), the flow slides along the switching boundary. Therefore, for a significant part of the time, Q_1 is constant and equal N^*, while Q_2 changes. This behavior is described as a *sliding mode* in the dynamical systems literature and occurs when $\frac{\mu_L(N^* \wedge N)-\lambda}{\delta} \leq Q_2 \leq \frac{\mu_H(N^* \wedge N)-\lambda}{\delta}$, which corresponds to $18.4 \leq Q_2 \leq 46.4$ in our example. More details can be found in the Appendix. While this sliding motion is a phenomenon of the fluid system, we can see that it still provides important insight into the behavior of the stochastic system.

The fluid analysis allowed us to identify these two operating modes. Gibbens et al. (1990) also used fixed point analysis of a deterministic system to demonstrate the existence of bi-stability, albeit in communication networks without feedback. Recognizing such behavior can exist will help avoid poor speedup decision making.

5. Model Extensions

Thus far, the focus of this work has been on the model presented in Section 2. We now consider a number of extensions to our stylized model which capture additional dynamics which can arise in various service settings. In particular, we look at the impact of including prioritization of customers and time-varying arrival rates. In both cases, we find that, although one can garner some additional insights from analyzing these extensions, the primary insights from our original analysis carry over to these extended models.

[2] The timing for such shift is unpredictable: we have simulation runs where the shift occurs within a few days and others which take much longer.



(a) Sample path of Case 2.1.2

(b) $\mathbb{Q}_1$ and $\mathbb{Q}_2$ distributions

Figure 9 Case 2.1.2 Simulation.

5.1. New Versus Return Customers

In this section, we consider differentiating between return and first time customers. Return customers may warrant higher priority in order to limit the total time customers spend in the system (e.g. Huang et al. (2012)). In addition, their service rates may differ, as seen in Durbin and Kopel (1993). We now examine the dynamics of our queueing model where the service rates and return probabilities depend, not only on congestion, but also on whether the customer is new versus returning. We assume that returning customers have preemptive priority over new customers. Again, we use fluid analysis to generate insights about our stochastic model. $\mu_{F,L}$ ($\mu_{R,L}$) denotes the service rate for first-time (return) Needy customers when the system is considered underloaded, while $\mu_{F,H}$ ($\mu_{R,H}$) represents the same when the system is considered overloaded. Similarly, $p_{F,L}$ ($p_{R,L}$) denotes the probability of return for first-time (return) Needy customers when the system is considered underloaded, while $p_{F,H}$ ($p_{R,H}$) represents the same when the system is considered overloaded. Denote by Q_1^F and Q_1^R the fluid content of first-time and return Needy customers, respectively. Thus, when $Q_1^F + Q_1^R \geq N^*$, the system is considered overloaded and speedup is used. Because we give preemptive priority to return customers, capacity will first be allocated to them ($Q_1^R \wedge N$); any remaining service capacity, $(N - Q_1^R)^+$, is allocated to the first-time Needy customers. The modified ODE under consideration is now:

$$\begin{aligned}
\dot{Q_1^F} &= \lambda - (Q_1^F \wedge (N - Q_1^R)^+)[\mu_{F,L} 1_{\{Q_1^F + Q_1^R < N^*\}} + \mu_{F,H} 1_{\{Q_1^F + Q_1^R \geq N^*\}}], \\
\dot{Q_1^R} &= \delta Q_2 - (Q_1^R \wedge N)[\mu_{R,L} 1_{\{Q_1^F + Q_1^R < N^*\}} + \mu_{R,H} 1_{\{Q_1^F + Q_1^R \geq N^*\}}], \\
\dot{Q_2} &= -\delta Q_2 + (Q_1^F \wedge (N - Q_1^R)^+)[p_{F,L}\mu_{F,L} 1_{\{Q_1^F + Q_1^R < N^*\}} + p_{F,H}\mu_{F,H} 1_{\{Q_1^F + Q_1^R \geq N^*\}}] \\
&\quad + (Q_1^R \wedge N)[p_{R,L}\mu_{R,L} 1_{\{Q_1^F + Q_1^R < N^*\}} + p_{R,H}\mu_{R,H} 1_{\{Q_1^F + Q_1^R \geq N^*\}}]
\end{aligned} \tag{5.1}$$

For this model, we utilize numerical approaches as the increased complexity of this model introduces additional challenges making it cumbersome to employ the generalized Lyapunov analysis used to prove

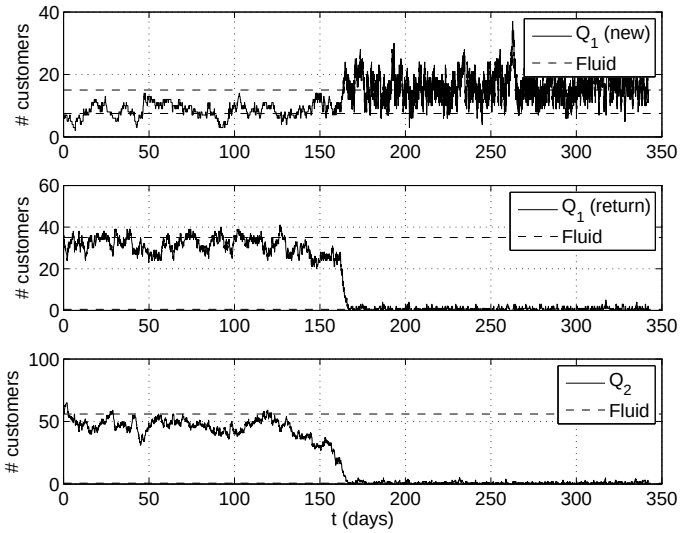


Figure 10 Simulation: New Versus Return Customers.

Theorem 4.3. Similar to our original model, we find that this extended fluid model also has two cases: one with a single globally stable equilibrium and another with bi-stability.

We translate the insight generated from the numerical analysis of the fluid model to a stochastic model via simulation of a system with $N = 45$ servers and speedup threshold $N^* = 35$. We use the following parameters in this example: $\mu_{F,L} = .01, \mu_{F,H} = .02, \mu_{R,L} = .015, \mu_{R,H} = .02, p_{F,L} = .05, p_{R,L} = .06, p_{F,H} = .7, p_{R,H} = .85; \lambda = .15, \delta = .0125$. Figure 10 shows the result of a single trace of this extended model. We see there exists a bi-stability effect in which the system transitions, after nearly 5 months, from a 'bad' equilibrium, where the system is always under speedup, to a 'good' equilibrium, where speedup is hardly used. Note that under the 'good' equilibrium, most of the customers are *new* customers and there are very few return customers; however, under the 'bad' equilibrium most of the customers are *returning* customers. Similar to our original model in Section 2, we see that in this case, utilizing speedup can result in even more congestion. We see again that when such a bi-stability exists, other mechanisms, such as admission control, may be more effective in navigating periods of high congestion.

5.2. Time Varying Arrivals

Another marked property of service systems is that customers arrivals are often time-varying (e.g. Gans et al. (2003), Green et al. (2006), Yom-Tov and Mandelbaum (2013)). We now explore the implications of having time-varying arrivals.

As discussed in Yom-Tov and Mandelbaum (2013) for a closely related queueing system (with returns but no speedup), the impact of time-varying arrivals depends on the relationship of the period and amplitude of the arrival rate versus the service duration. Time-variation can substantially impact the dynamics of our queueing system, especially when the scale of the service time is long but of the same order as the

time-variation. Here, we discover speedup control can sometimes smooth the time-variability. A complete analysis of the time-variability case is beyond the scope of this paper and there is currently little theory to support analysis of time-varying Filippov systems. Therefore, most of the observations we present here are based on numerical and simulation analysis.

We now consider a queueing system with the same stochastic dynamics as the system described in Section 2, except that the arrival process no longer has constant rate. We now model the arrival rate as a nonhomogeneous Poisson Process with time-varying arrival rate $\lambda(t)$. We again use fluid models to provide insight for the stochastic model. Accordingly, we can modify our original ODE in Equation (2.3), to derive an ODE to describe the fluid dynamics of this system with time-varying arrival rate as follows:

$$\begin{aligned}\dot{Q}_1(t) &= \lambda(t) + \delta Q_2(t) - 1_{\{Q_1(t)<N^*\}}\mu_L(Q_1(t)\wedge N) - 1_{\{Q_1(t)\geq N^*\}}\mu_H(Q_1(t)\wedge N),\\ \dot{Q}_2(t) &= -\delta Q_2(t) + 1_{\{Q_1(t)<N^*\}}p_L\mu_L(Q_1(t)\wedge N) + 1_{\{Q_1(t)\geq N^*\}}p_H\mu_H(Q_1(t)\wedge N).\end{aligned} \tag{5.2}$$

In our analysis of this modified system, we find the distinction between Case 1 and 2 still exists. In Case 1 we have a distinct solution to the ODE, while in Case 2, the system is quite chaotic (i.e. very dependent on the specific starting point and the phase of the arrival rate). Hence, we concentrate on Case 1. In this case, the solution may not be an equilibrium point as it was before, but could be an orbit, which is a periodic function which the trajectory follows over time. This orbit, which we denote by $\bar{q}(t)$, is closely related to the solution of a (time-varying) ODE which always or never uses speedup. We define $q^H(t)$ as the solution for the following ODE when speedup is always used.

$$\begin{aligned}\dot{q}_1^H(t) &= \lambda(t) + \delta q_2^H(t) - \mu_H q_1^H(t),\\ \dot{q}_2^H(t) &= -\delta q_2^H(t) + p_H\mu_H q_1^H(t).\end{aligned} \tag{5.3}$$

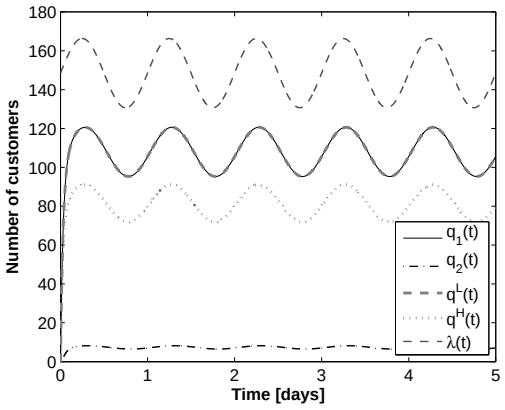
We similarly define $q^L(t)$ as the solution for the ODE when speedup is never used. A complete analysis of such an ODE is given in Yom-Tov and Mandelbaum (2013). If the arrival rate is periodic (as is the case in many service systems), $q^H(t)$ and $q^L(t)$ are cyclic functions that exhibit similar time-variation which lags after the arrival rate function $\lambda(t)$. This orbit's period is the same as the period of the arrival rates, though the phase is shifted. Since we are in Case 1, one can view never using speedup as a worst case scenario, i.e. the average number of customers is the highest possible. In a sense, $q^L(t)$ is an upper bound for the long term dynamics of our fluid system: consider two trajectories which start at the same initial point. One follows the dynamics described by (5.3) while the other follows the dynamics described by (5.2). The fluid content of Needy customers in the latter will always be larger. Hence, $q_1^L(t)$ is an upper bounding function for $\bar{q}_1$. Similarly, $q_1^H(t)$ is a lower bounding function for $\bar{q}_1$.

We start by considering a sinusoidal arrival process: $\lambda(t) = 148.5 \times (1 + 0.12\sin(2\pi t/f), t \geq 0$. The period f is 24 hours, $\mu_L = 1.474, \mu_H = 2.018, p_L = 0.667, p_H = 0.973, \delta = 1.445, N = 150$. Using numeric analysis, we find that in Case 1, the orbit function $\bar{q}$ is a function that during various points of its cycle

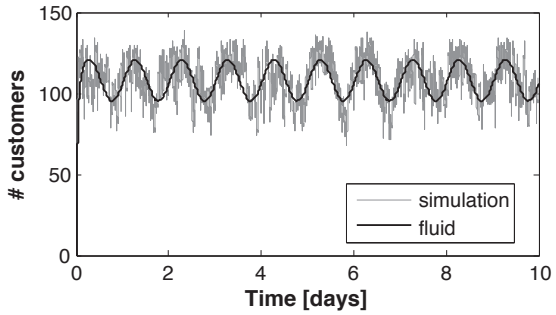
(determined by the cycle of time-variability in the arrival process) will follow either the upper bounding function, $q^L(t)$, the lower bounding function, $q_1^H(t)$, or stay along the speedup threshold, N^*.

Figure 11 presents some typical fluid approximations and simulated sample paths of our stochastic system under different threshold values. In Figure 11(a), the trajectory converges to the orbit $q^H(t)$ where speedup is always used. Because of the periodic nature of the arrival process, we see that the trajectory on the fluid model follows a cyclic orbit with the same period as the arrival process. In Figure 11(e), the trajectory converges to N^*. This is similar to the pseudo-equilibrium in Case 1.2 without periodic arrivals where q^H and q^L reside on opposite sides of the speedup threshold, N^*, so that the trajectory is pulled rapidly back and forth making N^* an equilibrium. What is interesting in the case of time-varying arrival rates is that this behavior creates a non-time-varying equilibrium, N^*. We see that using speedup improves access to service at Station 1, by reducing the offered load. It also has another benefit in that it also has the power to remove time-variation and smooth the occupancy level at Station 1. Thus, despite the fact that $\lambda(t)$ and, consequently, $q^H(t)$ and $q^L(t)$ are periodic functions with a period of 24 hours, the fluid content of Needy customers is *time-invariant* and fixed at N^*. Another possible trajectory of the fluid content is depicted in Figure 11(d). The orbit function, $\bar{q}(t)$, can follow 2 of the trajectories: it follows $q^H(t)$, but when it hits the speedup threshold, N^*, it stays there until the arrival rate falls again, at which point it returns to tracking $q^H(t)$. Thus, there is some smoothing of the occupancy level at Station 1 (when $Q_1(t) = N^*$); however, because the speedup threshold is higher than in Figure 11(e), it is not held constant for all time and the trajectory exhibits some (but not all) of the time variation of $q^H(t)$. Figures 11(b)-11(f) present simulated sample paths of the fluid systems depicted in Figures 11(a)-11(e). We see that the fluid approximation is quite accurate in describing the time-varying system dynamics.

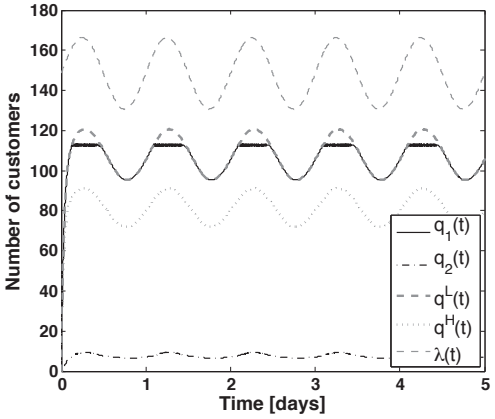
While we see some very interesting dynamics arise when incorporating time-variation into our model, we focused on a numeric setting which allows us to observe the nuances. We also wish examine the impact of time-varying arrivals in the ICU setting. In the ICU–unlike the ED setting in Yom-Tov and Mandelbaum (2013) and Green et al. (2006)–the LOS is quite long compared to the time variability. Specifically, the arrival rate varies at the time scale of hours, while ICU LOS is typically 3-4 days, spanning a few arrival rate cycles. Due to this discrepancy in the time scale of variation versus service time, Yom-Tov and Mandelbaum (2013) suggests that the impact of time-variation is likely to be small. We also find this to be true when considering our system with speedup. In Figures 13(a) and 13(b), we present the fluid approximation and simulated sample path of $\mathbb{Q}_1$ using identical parameters as in Figures 7(a) and 7(b), except the arrival rate is according to the empirical time-varying arrival rates depicted in Figure 12. We observe the system still varies around the chosen threshold and it is difficult to ascertain substantial differences from Figure 7(a). While we find that in this setting incorporating daily variability does not significantly alter the system dynamics, we can see in the previous analysis that the dynamics can change dramatically when incorporating time-varying arrivals. We leave further exploration of this type of time-varying, state-dependent queueing system for future research.



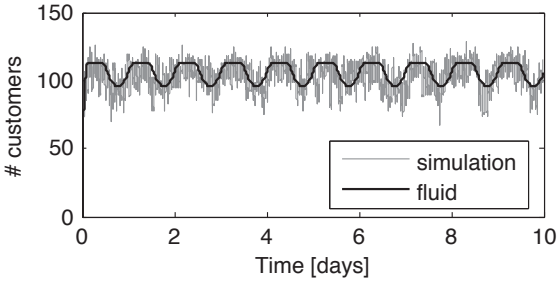
(a) Equilibrium is $q_1^H(t)$ $(r = 0.85)$: Fluid



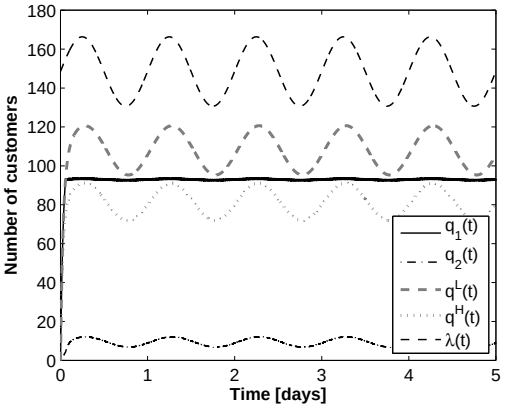
(b) Equilibrium is $q_1^H(t)$ $(r = 0.85)$: Sample path



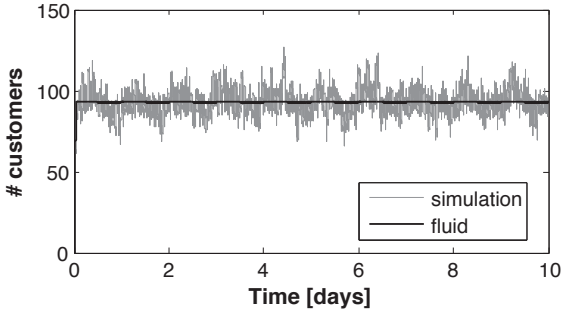
(c) Equilibrium is a mix between N^* and $q^H(t)$ $(r = 0.75)$: Fluid



(d) Equilibrium is a mix between N^* and $q^H(t)$ $(r = 0.75)$: Sample path



(e) Equilibrium is N^* $(r = 0.62)$: Fluid



(f) Equilibrium is N^* $(r = 0.62)$: Sample path

Figure 11 Fluid approximations and sample paths for time varying arrivals with different threshold under Case 1 conditions.

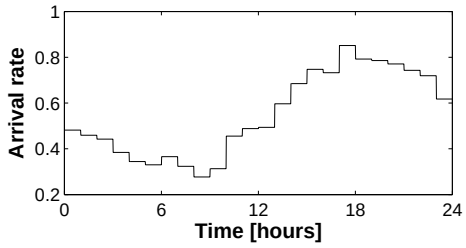
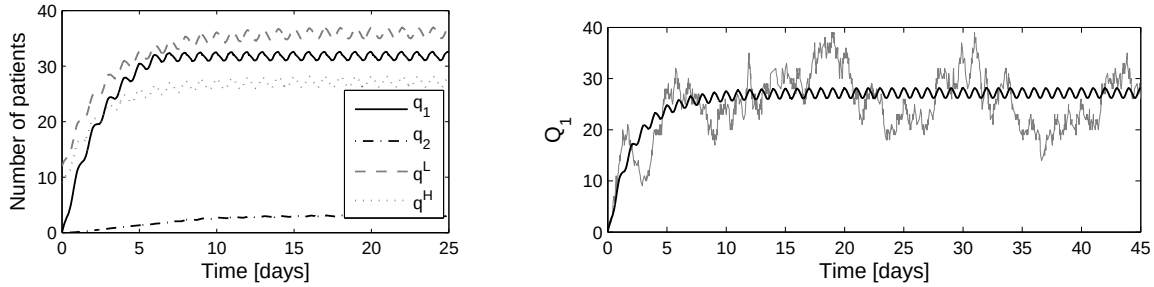


Figure 12 Time varying arrival rate to ICU (in number of patients/hour).



(a) Fluid approximation

(b) Sample Path

Figure 13 Time-varying ICU: Fluid approximation and sample path.

6. Conclusions

In this work, we consider a queueing model where service rates and return probabilities increase when the system is overloaded. We analyze the dynamics of this state-dependent queueing model to gain insight into the impact speedup and returns have on system dynamics. The model presented here provides insights into the pros and cons of using speedup in a service system where customers may return to service.

We find that there are two main parameter regimes which define whether speedup can be a beneficial or detrimental operational tool to help alleviate temporary congestion. Such analysis provides tools to enable practitioners to assess the potential benefits and pitfalls of different speedup policies. We find that in some cases speedup can be beneficial to help alleviate congestion. In such situations, the amount of congestion and frequency of speedup can be specified via the speedup threshold, N^*. In other cases, the use of speedup can exacerbate congestion. Moreover, an interesting bi-stability can arise, which demonstrates the potential problems associated with using speedup.

We demonstrate via simulation that the fluid approximation to our state-dependent queueing system can be very accurate. However, there are scenarios where the accuracy suffers–particularly in small systems and/or when speedup is used around 25% or 75% of the time. In this work, we derived the fluid directly. Establishing a proof of the limit in a Functional Weak Law of Large Numbers sense introduces several technical challenges due to the discontinuity of the ODE. However, it would be useful to be able to show such a result. Additionally, it would be interesting to consider refinements to the fluid approximation.

Finally, we consider two important extensions for our model: i) differing dynamics for new and returning customers and ii) time-varying arrivals. This analysis provides some additional insights, but also suggests that our original stylized model has value in shedding light on the much more complex reality. We observe, for example, that in the ICU application one need not explicitly consider time-varying dynamics. Instead, may draw important conclusions on the impact of using speedup from the time-stationary model. Nevertheless, we find the time-varying dynamics can be very interesting in it's own right and plan to investigate it further in future work.

Acknowledgments

The authors would like to thank the anonymous reviewers, an associate editor, and the department editor for their valuable suggestions. The authors are grateful to Mor Armony, Linda Green, and Ward Whitt for their support and feedback during the course of this research. A previous version of this paper placed second in the INFORMS 2012 JFIG best paper competition under the title "When to use Speedup: An Examination of Intensive Care Units with Readmissions". The authors would also like to thank Marla Gardner and John Greene for their help in preparing the ICU data. This research was conducted while the second author was a post-doctoral researcher at Columbia University – she would like to gratefully acknowledge the wonderful hospitality of this institution.

Appendix

A. Miscellaneous Proofs

PROOF OF THEOREM 3.1:

1. We begin with the instability result. Recall for instability, we must have the total fluid content of jobs in the system grow without bound. That is, we consider $Q_T = Q_1 + Q_2$. The dynamics of Q_T can be summarized as:

$$\dot{Q}_T = \dot{Q}_1 + \dot{Q}_2 = \lambda - (1-p_L)\mu_L(Q_1 \wedge N).$$

   If the system is unstable, then $\lim_{t\to\infty} \frac{Q_T(t)}{t} > 0$. We integrate and solve for $Q_T(t)$. We have:

$$\begin{aligned} \lim_{t\to\infty} \frac{Q_T(t)}{t} &= \lambda - (1-p_L)\mu_L \lim_{t\to\infty} \frac{1}{t}\int_0^t (Q_1(\tau) \wedge N) d\tau \\ &\geq \lambda - (1-p_L)\mu_L \lim_{t\to\infty} \frac{1}{t}\int_0^t N d\tau \\ &= \lambda - (1-p_L)\mu_L N > 0 \text{ if } N < \frac{\lambda}{(1-p_L)\mu_L} \end{aligned} \tag{A.1}$$

2. For the stability and equilibrium result, we first show that $\bar{q} = \left(\frac{\lambda}{(1-p_L)\mu_L}, \frac{\lambda p_L}{(1-p_L)\delta}\right)$ is a globally stable equilibrium. The stability result follows from the finiteness of $\bar{q}$. To show global stability, we use the following Lyapunov function:

$$V(Q) = |Q_1 - \bar{q}_1| + |Q_2 - \bar{q}_2|$$

   We must show that for all $Q \neq \bar{q}$, $\dot{V}(Q) < 0$. To do this, we must examine a few cases:

(a) $Q_1 > \bar{q}_1, Q_2 > \bar{q}_2$.

$$\dot{V}(Q) = \dot{Q}_1 + \dot{Q}_2 = \lambda - (1-p_L)\mu_L(Q_1 \wedge N) < \lambda - (1-p_L)\mu_L\bar{q}_1 = 0$$

(b) $Q_1 < \bar{q}_1, Q_2 < \bar{q}_2$.

$$\dot{V}(Q) = -\dot{Q}_1 - \dot{Q}_2 = -\lambda + (1-p_L)\mu_L(Q_1 \wedge N) < -\lambda + (1-p_L)\mu_L\bar{q}_1 = 0$$

(c) $Q_1 > \bar{q}_1, Q_2 < \bar{q}_2$.

$$\dot{V}(Q) = \dot{Q}_1 - \dot{Q}_2 = \lambda + 2\delta Q_2 - (1+p_L)\mu_L(Q_1 \wedge N) < \lambda + 2\delta\bar{q}_2 - (1+p_L)\mu_L\bar{q}_1 = 0$$

(d) $Q_1 < \bar{q}_1, Q_2 > \bar{q}_2$.

$$\dot{V}(Q) = -\dot{Q}_1 + \dot{Q}_2 = -\lambda - 2\delta Q_2 + (1+p_L)\mu_L(Q_1 \wedge N) < -\lambda - 2\delta\bar{q}_2 + (1+p_L)\mu_L\bar{q}_1 = 0$$

(e) $Q_1 = \bar{q}_1, Q_2 > \bar{q}_2$.

$$\dot{V}(Q) = \dot{Q}_2 = -\delta Q_2 + p_L\mu_L(Q_1 \wedge N) < -\delta\bar{q}_2 + p_L\mu_L\bar{q}_1 = 0$$

(f) $Q_1 = \bar{q}_1, Q_2 < \bar{q}_2$.

$$\dot{V}(Q) = -\dot{Q}_2 = \delta Q_2 - p_L\mu_L(Q_1 \wedge N) < \delta\bar{q}_2 - p_L\mu_L\bar{q}_1 = 0$$

(g) $Q_1 > \bar{q}_1, Q_2 = \bar{q}_2$.

$$\dot{V}(Q) = \dot{Q}_1 = \lambda + \delta Q_2 - \mu_L(Q_1 \wedge N) < \lambda + \delta\bar{q}_2 - \mu_L\bar{q}_1 = 0$$

(h) $Q_1 < \bar{q}_1, Q_2 = \bar{q}_2$.

$$\dot{V}(Q) = -\dot{Q}_1 = -\lambda - \delta Q_2 + \mu_L(Q_1 \wedge N) < -\lambda - \delta\bar{q}_2 + \mu_L\bar{q}_1 = 0$$

■

A.1. Proofs for Discontinuous Ordinary Differential Equations

Our system is a piecewise-smooth set of ordinary differential equations. As such, it fits in to the framework of Filippov (1988). In our analysis, we use Lyapunov techniques as well the methods outlined in di Bernardo et al. (2008).

Primitives

To begin, we represent our dynamic system by the following differential equation using the Filippov convex method. More details of this method can be found in di Bernardo et al. (2008) and Filippov (1988). The basic premise is to divide the state space into regions where the ODE is smooth and continuous in order to leverage existing results of smooth dynamical systems. A separate region, the switching boundary[3] is defined as the states of discontinuity in the ODE. The approach is to transform the differential equation into a *differential inclusion*, where the differential function is now a set-valued function. Additionally, Filippov (1988) proves that solutions to the original discontinuous differential equation coincide with solutions to the appropriately defined differential inclusion. In what follows, we will discuss first how to transform Equation (2.3) into the appropriate differential inclusion. Next, we will demonstrate the desired results for the differential inclusion, which will imply the result holds for the original differential equation. Note that in our case, the differential equation (and subsequently the differential inclusion) does not depend on t, but only on Q.

To start, we separate the state space, $\mathbb{R}^2_+$ into two regions, $\mathcal{D}_L$ and $\mathcal{D}_H$, and the switching boundary, Σ, between them as follows:

$$\mathcal{D}_L = \{Q : Q_1 < N^*\}$$
$$\mathcal{D}_H = \{Q : Q_1 > N^*\}$$
$$\Sigma = \{Q : Q_1 = N^*\}.$$

In the regions $\mathcal{D}_L$ and $\mathcal{D}_H$, the ODE is smooth. However, the ODE is discontinuous at the switching boundary Σ. The Filippov methodology overcomes this by transforming the differential equation into a differential inclusion by using a convex combination of the smooth flows defined in $\mathcal{D}_L$ and $\mathcal{D}_H$ on the switching boundary, Σ. We define the fluid function $F_i(Q)$, $Q \in \mathcal{D}_i$, as the smooth ODE in these regions:

$$F_L(Q) = \begin{pmatrix} \lambda + \delta Q_2 - \mu_L(Q_1 \wedge N) \\ -\delta Q_2 + p_L \mu_L(Q_1 \wedge N) \end{pmatrix}, F_H(Q) = \begin{pmatrix} \lambda + \delta Q_2 - \mu_H(Q_1 \wedge N) \\ -\delta Q_2 + p_H \mu_H(Q_1 \wedge N) \end{pmatrix}.$$

Note that even though the ODE is non-differentiable at $Q_1 = N$, as is customary, it is still considered smooth, and not discontinuous, at this point. The real challenge comes at the switching boundary , i.e. when $Q_1 = N^*$. Now, our ODE $\dot{Q} = F(Q)$ can be represented via a Filippov ODE (a.k.a. a differential inclusion):

$$\dot{Q} \in \mathcal{F}(Q) = \begin{cases} F_L(Q) & , \text{ if } Q \in \mathcal{D}_L, \\ F_H(Q) & , \text{ if } Q \in \mathcal{D}_H, \\ \{(1-\psi)F_L(Q) + \psi F_H(Q) | 0 \le \psi \le 1\} & , \text{ if } Q \in \Sigma. \end{cases} \tag{A.2}$$

PROOF OF THEOREM 4.1: We start by stating the existence result in Filippov (1988) in terms of our notation. The result is for a differential inclusion; however, the Filippov method utilizes the fact that solutions of the differential inclusion coincide with solutions of the original discontinuous differential equation.

[3] The switching boundary is also often referred to as a discontinuity set, discontinuity boundary or switching manifold.

Hence, if our differential inclusion satisfies the conditions of the following theorem, this will imply existence of a solution to our ODE (2.3).

Theorem A.1 *(Theorem 1, Chapter 2, Section 7 of Filippov (1988)) Let $\mathcal{F}(Q)$ be a differential inclusion that satisfies the following conditions in the domain G:*

1. *$\mathcal{F}(Q)$ is non-empty for all $Q \in G$.*
2. *$\mathcal{F}(Q)$ is bounded and closed for all $Q \in G$.*
3. *$\mathcal{F}(Q)$ is convex for all $Q \in G$.*
4. *The function $\mathcal{F}$ is upper semicontinuous in Q.*

Then for any point $q_0 \in G$, there exists a solution of the problem

$$\dot{Q} \in \mathcal{F}(Q), Q(0) = q_0.$$

We will consider the domain $G = [0, Q_{\max}] \times [0, Q_{\max}]$ for some arbitrary finite constant, $Q_{\max} < \infty$. Now, we just have to demonstrate that the conditions hold for all $Q \in G$. It is easy to see that conditions 1-4 hold for all $Q \in \mathcal{D}_L \cup \mathcal{D}_H$, as in these regions $\mathcal{F}$ is a continuous real-valued function (rather than a set-valued function). Thus, $\mathcal{F}(Q)$ is a single point, which is bounded above by $\max\{\lambda + \delta Q_{\max}, p_H \mu_H (Q_{\max} \wedge N)\}$ and bounded below by $\min\{\lambda - \mu_H(Q_{\max} \wedge N), -\delta Q_{\max}\}$. Any continuous function is also upper semicontinuous, so the fourth condition follows.

It remains to show the four conditions hold for any Q on the switching boundary, Σ. By the same argument as for $Q \in \mathcal{D}_L \cup \mathcal{D}_H$, $\mathcal{F}(Q)$ is bounded for any $Q \in \Sigma$. By definition of $\mathcal{F}$ in (A.2), $\mathcal{F}(Q)$ is closed and convex for $Q \in \Sigma$ as it is a convex combination of F_H and F_L. Since both of these functions are non-empty, so is $\mathcal{F}$. Finally, to show $\mathcal{F}$ is upper semicontinuous on Σ, we need to show that $\mathcal{F}$ is upper semicontinuous for every $Q \in \Sigma$. The set-valued function $\mathcal{F} : \Sigma \to Y \subset \mathbb{R}^2_+$ is upper semicontinuous at a point $Q \in \Sigma$ provided that for each open set V in Y containing $\mathcal{F}(Q)$, there is an open set U in Σ containing Q such that if $Q' \in U$, then $\mathcal{F}(Q') \subseteq V$. By the definition of the inclusion, for an $Q \in \Sigma, \mathcal{F}(Q) = \{(1-\psi)F_L(Q) + \psi F_H(Q) | 0 \leq \psi \leq 1\}$. Consider an open set V which contains $\mathcal{F}(Q)$: there exists an $\epsilon > 0$, such that for every $f \in \mathcal{F}(Q)$, $f + \epsilon \in V$. Now by the continuity of F_H and F_L, there exists $\delta > 0$, such that if $|Q' - Q| < \delta$, then $|F_H(Q) - F_H(Q')| < \epsilon/2$ and $|F_L(Q) - F_L(Q')| < \epsilon/2$. Thus, $|[(1-\psi)F_L(Q) + \psi F_H(Q)] - [(1-\psi)F_L(Q') + \psi F_H(Q')]| < \epsilon$ for all $0 \leq \psi \leq 1$. Hence, $\mathcal{F}(Q') \subset V$ and we have derived the necessary open set $U = \{Q' | |Q' - Q| < \delta\} \cap \Sigma$ (recall that the intersection of two open sets is open.). This demonstrates that $\mathcal{F}$ is upper semicontinuous in Σ. All conditions hold on the switching boundary. Therefore, there exists a solution to the differential inclusion, and subsequently our ODE.

■

PROOF OF THEOREM 4.2: We consider each of the two cases.

1. $[q_1^H < q_1^L]$

(a) We first consider the case where $N < q_1^H$. Similar to the instability proof of Theorem 3.1, we consider the total number of jobs in the system and show that $\lim_{t\to\infty} \frac{Q_T(t)}{t} > 0$. We have that:

$$\dot{Q}_T = \dot{Q}_1 + \dot{Q}_2 = \lambda - 1_{\{Q_1(t)<N^*\}}(1-p_L)\mu_L(Q_1 \wedge N) - 1_{\{Q_1(t)\geq N^*\}}(1-p_H)\mu_H(Q_1 \wedge N)$$

We integrate both sides, divide by t and take the limit as $t \to \infty$:

$$\begin{aligned}
\lim_{t\to\infty} \frac{Q_T(t)}{t} &= \lambda - \lim_{t\to\infty} \frac{1}{t} \int_0^t (Q_1(\tau) \wedge N) \left[(1-p_L)\mu_L 1_{\{Q_1(\tau)<N^*\}} + (1-p_H)\mu_H 1_{\{Q_1(\tau)\geq N^*\}}\right] d\tau \\
&> \lambda - (1-p_H)\mu_H \lim_{t\to\infty} \frac{1}{t} \int_0^t (Q_1(\tau) \wedge N) d\tau \\
&\geq \lambda - (1-p_H)\mu_H \lim_{t\to\infty} \frac{1}{t} \int_0^t N d\tau \\
&= \lambda - (1-p_H)\mu_H N > \lambda - (1-p_H)\mu_H q_1^H = 0
\end{aligned}$$

The first inequality comes from the fact that in this case $q_1^H < q_1^L$, which implies that $(1-p_L)\mu_L < (1-p_H)\mu_H$. The last inequality comes from the assumption that $N < \frac{\lambda}{(1-p_H)\mu_H} = q_1^H$. Hence, the system is unstable if $N < q_1^H$.

(b) We next consider the case where $N < q_1^L$ and $N^* = \infty$ (i.e. speedup is never used). This is simply the result of Theorem 3.1.

2. $[q_1^H \geq q_1^L]$

(a) We first consider the case where $N < q_1^L$. Using the same argument as before, we now have:

$$\begin{aligned}
\lim_{t\to\infty} \frac{Q_T(t)}{t} &= \lambda - \lim_{t\to\infty} \frac{1}{t} \int_0^t (Q_1(\tau) \wedge N) \left[(1-p_L)\mu_L 1_{\{Q_1(\tau)<N^*\}} + (1-p_H)\mu_H 1_{\{Q_1(\tau)\geq N^*\}}\right] d\tau \\
&\geq \lambda - (1-p_L)\mu_L \lim_{t\to\infty} \frac{1}{t} \int_0^t (Q_1(\tau) \wedge N) d\tau \\
&\geq \lambda - (1-p_L)\mu_L \lim_{t\to\infty} \frac{1}{t} \int_0^t N d\tau \\
&= \lambda - (1-p_L)\mu_L N > 0
\end{aligned}$$

The first inequality is because $q_1^H \geq q_1^L$. The last inequality comes from the assumption that $N < \frac{\lambda}{(1-p_L)\mu_L} = q_1^L$. Hence, the system is unstable if $N < q_1^L$.

(b) Now we consider the case where $N < q_1^H$ and $N^* < q_1^L$. Again, we focus on the total workload in the original system. We have that:

$$\begin{aligned}
\lim_{t\to\infty} \frac{Q_T(t)}{t} &= \lambda - \lim_{t\to\infty} \frac{1}{t} \int_0^t (Q_1(\tau) \wedge N) \left[(1-p_L)\mu_L 1_{\{Q_1(\tau)<N^*\}} + (1-p_H)\mu_H 1_{\{Q_1(\tau)\geq N^*\}}\right] d\tau \\
&> \lambda - \lim_{t\to\infty} \frac{1}{t} \int_0^t \left[N^*(1-p_L)\mu_L 1_{\{Q_1(\tau)<N^*\}} + q_1^H(1-p_H)\mu_H 1_{\{Q_1(\tau)\geq N^*\}}\right] d\tau \\
&> \lambda - \lim_{t\to\infty} \frac{1}{t} \int_0^t \left[q_1^L(1-p_L)\mu_L 1_{\{Q_1(\tau)<N^*\}} + q_1^H(1-p_H)\mu_H 1_{\{Q_1(\tau)\geq N^*\}}\right] d\tau \\
&= \lambda - \lim_{t\to\infty} \frac{1}{t} \int_0^t \lambda d\tau = 0.
\end{aligned}$$

The inequalities come from the assumptions that $N < q_1^H$ and $N^* < q_1^L$. Hence, the system is unstable.

■

PROOF OF THEOREM 4.3:

Defining Equilibria points and the ODE flow. Our theorem distinguishes between seven cases and three equilibria points. We start by identifying these points and the conditions under which they arise as equilibria; we then prove stability. To do so, we refer back to the primitives defined earlier for Filippov ODEs, which transform the original ODE into a differential inclusion.

We start by examining each region, $\mathcal{D}_H$ and $\mathcal{D}_L$. If the solution trajectory never leaves a region $\mathcal{D}_i$ once it enters the region, then the solution can be characterized with standard methods for regular continuous ODEs (Guckenheimer and Holmes (2002)). Therefore, we first define the equilibria of $F_L(Q)$ and $F_H(Q)$. That is, we consider the long-term behavior of a system with dynamics which are defined by the continuous ODE $F_L(Q)$ ($F_H(Q)$) across the entire state space in the case where speedup is never (always) used. We assume there is no switching boundary or speedup threshold. By the methods of di Bernardo et al. (2008), if a locally or globally stable equilibrium, $\bar{q}_L$ ($\bar{q}_H$), of the continuous ODE $F_L(Q)$ ($F_H(Q)$) is in $\mathcal{D}_L$ ($\mathcal{D}_H$), then this directly translates into local stability of $\bar{q}_L$ ($\bar{q}_H$) in the original discontinuous ODE.

Lemma A.1 *If $q_1^L < N$, then the continuous ODE F_L has a globally stable equilibrium at q^L. If $q^L \in \mathcal{D}_L$ and $q_1^L < N$, then q^L is a locally stable equilibrium in the original discontinuous ODE defined in* (2.3).

PROOF: The first part comes from Theorem 3.1. The second part is typical for piecewise-smooth ODEs and follows from Guckenheimer and Holmes (2002) and our stability assumption. Intuitively, by the stability of q^L in the continuous ODE defined by F_L, there exists a small ball around q^L such that trajectories which start within the ball will converge to q^L and stay there. If q^L is in $\mathcal{D}_L$, then one can also select a small ball around q^L such that 1) trajectories which start within the ball will converge to q^L and 2) the ball is entirely contained in $\mathcal{D}_L$. Then, starting a trajectory in this small ball, but letting the system dynamics be defined according to our original, discontinuous ODE in (2.3), the trajectory will stay in the ball, which means they will stay in $\mathcal{D}_L$ and follow the same dynamics as the continuous ODE F_L, since it won't hit the switching boundary, Σ, or enter the other region, $\mathcal{D}_H$. Thus, q^L is locally stable. ■

Lemma A.2 *If $q_1^H < N$, then the continuous ODE F_H has a globally stable equilibrium at q^H. If $q^H \in \mathcal{D}_H$ and $q_1^H < N$ then q^H is a locally stable equilibrium in the original discontinuous ODE defined in* (2.3).

PROOF: The results can be derived with the same techniques as in the proof for Theorem 3.1 and Lemma A.1.

Lemma A.1 proves the existence of a locally stable equilibrium at q_L in cases 1.3, 2.1.2, 2.1.3, and 2.2 because $q^L \in \mathcal{D}_L$ holds in these cases, as demonstrated for Case 1.3 in Figure 3. Similarly, Lemma A.2

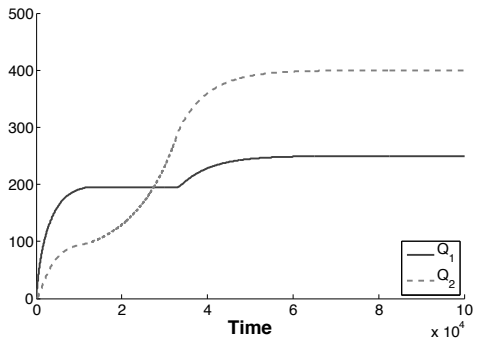


Figure 14 Example of a sliding motion.

proves the existence of a locally stable equilibrium at q_H in cases 1.1, 2.1.1 and 2.1.2. We will later use Lyaponov techniques to prove the global stability of these equilibria in the appropriate cases.

Next, we analyze the dynamics of the system on the switching boundary, Σ, and identify our third equilibrium which lies in this space. To do this, we look at the forces on either side of the switching boundary. Specifically, we look at the component of F_i normal to Σ, which we denote by $\mathcal{L}_i$. These measures are often referred to as the Lie derivatives.

$$\begin{aligned} \mathcal{L}_L &= \lambda + \delta Q_2 - \mu_L(N^* \wedge N), \\ \mathcal{L}_H &= \lambda + \delta Q_2 - \mu_H(N^* \wedge N) \end{aligned} \tag{A.3}$$

We have the following three cases to consider:

1. $\mathcal{L}_L < 0$ and $\mathcal{L}_H < 0$: a flow that hits Σ will transition from $\mathcal{D}_H$ to $\mathcal{D}_L$. If the flow is already in $\mathcal{D}_L$, it will stay in $\mathcal{D}_L$. Because $\mu_L < \mu_H$, for all $\left\{ Q \in \Sigma : Q_2 < \frac{\mu_L(N^* \wedge N) - \lambda}{\delta} \right\}$, both $\mathcal{L}_L$ and $\mathcal{L}_H$ are negative, and the transition will be from $\mathcal{D}_H$ to $\mathcal{D}_L$;
2. $\mathcal{L}_L > 0$ and $\mathcal{L}_H > 0$: a flow that hits Σ will transition from $\mathcal{D}_L$ to $\mathcal{D}_H$. If the flow is already in $\mathcal{D}_H$, it will stay in $\mathcal{D}_H$. For all $\left\{ Q \in \Sigma : Q_2 > \frac{\mu_H(N^* \wedge N) - \lambda}{\delta} \right\}$, both $\mathcal{L}_L$ and $\mathcal{L}_H$ are positive, and the transition will be from $\mathcal{D}_L$ to $\mathcal{D}_H$.
3. $\mathcal{L}_L \mathcal{L}_H < 0$: a flow that hits Σ will stay on Σ–potentially forever. The flow is said to be in a *sliding* mode, in which Q_1 is constant and equal to N^* while only Q_2 changes. Figure 14 illustrates an example of a sliding mode flow where $N^* = 200$. If $\frac{\mu_L(N^* \wedge N) - \lambda}{\delta} \leq Q_2 \leq \frac{\mu_H(N^* \wedge N) - \lambda}{\delta}$, we have that $\mathcal{L}_L < 0$ and $\mathcal{L}_H > 0$, so then the ODE is in a sliding mode on the switching boundary Σ. The importance of this sliding mode is there could be an additional equilibrium inside the sliding set. This point is typically not a fixed point in the standard sense since the ODE is non-zero at that point. However, it is still an equilibrium by definition 2.2 because if a flow starts at that point, it stays there. In some cases, this special equilibrium is a unique and stable equilibrium. We call this type of equilibrium a *pseudo-equilibrium* as in di Bernardo et al. (2008).

Definition A.1 *We call a point $\bar{q}$ a* pseudo-equilibrium *if it is an equilibrium of the sliding flow, i.e. for some scalar $0 < \xi < 1$,*

$$F_s = (1-\xi)F_L(\bar{q}) + \xi F_H(\bar{q}) = 0,\ \bar{q} \in \Sigma \tag{A.4}$$

Note the similarities to the Filippov ODE (A.2). When the convex combination is active in (A.2), i.e. $\psi \in (0,1)$, there exists an equilibrium on the switching boundary Σ and $\xi = \psi$.

Lemma A.3 *The point $(N^*, \alpha q_2^L + (1-\alpha)q_2^H)$ is a pseudo-equilibrium in Cases 1.2 and 2.1.2.*

PROOF: Solving (A.4) yields

$$\xi = \frac{\lambda - (1-p_L)\mu_L(N^* \wedge N)}{(1-p_H)\mu_H(N^* \wedge N) - (1-p_L)\mu_L(N^* \wedge N)}. \tag{A.5}$$

We have that $0 < \xi < 1$ if the following three conditions hold simultaneously:

(a) $\lambda - (1-p_L)\mu_L(N^* \wedge N) > 0 \ \Rightarrow\ N^* \wedge N < q_1^L$,

(b) $(1-p_H)\mu_H(N^* \wedge N) - (1-p_L)\mu_L(N^* \wedge N) > 0 \ \Rightarrow\ q_1^H < q_1^L$, and

(c) $\lambda - (1-p_L)\mu_L(N^* \wedge N) < (1-p_H)\mu_H(N^* \wedge N) - (1-p_L)\mu_L(N^* \wedge N) \ \Rightarrow\ q_1^H < N^* \wedge N$.

These are exactly the conditions of Case 1.2, which means that in this case there exists a pseudo-equilibrium on the switching boundary Σ. Substituting ξ from (A.5) into (A.4) gives the resulting equilibrium: $(N^*, \frac{(N^* \wedge N) - q_1^H}{q_1^L - q_1^H} q_2^L + \frac{q_1^L - (N^* \wedge N)}{q_1^L - q_1^H} q_2^H)$. We will show later that in this case this point is a globally stable equilibrium. By defining $\alpha = \frac{(N^* \wedge N) - q_1^H}{q_1^L - q_1^H}$, we can express the equilibrium as $(N^*, \alpha q_2^L - (1-\alpha)q_2^H)$. Note that under the above three conditions (Case 1.2) $0 < \alpha < 1$.

Another scenario where $0 < \xi < 1$ is if the following three conditions hold simultaneously:

(a) $\lambda - (1-p_L)\mu_L(N^* \wedge N) < 0 \ \Rightarrow\ N^* \wedge N > q_1^L$,

(b) $(1-p_H)\mu_H(N^* \wedge N) - (1-p_L)\mu_L(N^* \wedge N) < 0 \ \Rightarrow\ q_1^H > q_1^L$, and

(c) $\lambda - (1-p_L)\mu_L(N^* \wedge N) > (1-p_H)\mu_H(N^* \wedge N) - (1-p_L)\mu_L(N^* \wedge N) \ \Rightarrow\ q_1^H > N^* \wedge N$.

These are exactly the conditions of Cases 2.1.2 (for $q_1^L \neq q_1^H$) and 2.2, which means that in these cases there also exists a pseudo-equilibrium on the switching boundary Σ. Again this pseudo-equilibrium is $(N^*, \alpha q_2^L + (1-\alpha)q_2^H)$. Experiments show that this point is not stable. If we start the system at that point it will stay there, but very small changes from that point will bring us to one of the other two equilibria of the system. ■

Locally Stable Equilibria: Combining the results of Lemmas A.1, A.2, and A.3 yields our (Lyapunov/local) stability results for Case 2.1.2 and 2.2, and the existence of the third equilibrium in case 2.1.2.

Globally Stable Equilibria: We continue by proving the global stability results (Cases 1.1-1.3, 2.1.1, and 2.1.3). To show global stability in the sense of Lyapunov, we need to identify a Lyapunov function and prove

that for *all* $Q \in \mathbb{R}^2_+ \backslash \{\bar{q}\}$, the derivative of the Lyaponov function is strictly negative. We use the following Lyapunov function:

$$V(Q) = |Q_1 - \bar{q}_1| + |Q_2 - \bar{q}_2| \tag{A.6}$$

where $\bar{q}$ is the specified equilibrium. We use the Filippov methodology as describe earlier, which redefines the ODE as a differential inclusion so that on the switching boundary, Σ, is the convex combination of the surrounding smooth ODEs in (A.2) for $\mathcal{F}(Q)$. We continue to use this definition of our ODE and utilize the generalized Lyapunov theory for set valued functions. Using the approach in Shevitz and Paden (1994), we need to show that the set value map for our generalized Lyapunov derivative is negative for all states not equal to the equilibrium in order to establish global stability.

We have two cases to consider for our set value map, generalized Lyapunov derivative:

1. $[Q_1 \neq N^*]$.

$$\dot{\tilde{V}}(Q) = \begin{cases} \dot{V}(Q), & Q_1 \neq \bar{q}_1, Q_2 \neq \bar{q}_2; \\ \dot{Q}_1, & Q_1 > \bar{q}_1, Q_2 = \bar{q}_2; \\ -\dot{Q}_1, & Q_1 < \bar{q}_1, Q_2 = \bar{q}_2; \\ \dot{Q}_2, & Q_1 = \bar{q}_1, Q_2 > \bar{q}_2; \\ -\dot{Q}_2, & Q_1 = \bar{q}_1, Q_2 < \bar{q}_2. \end{cases} \tag{A.7}$$

2. $[Q_1 = N^*]$. In this case, the flow is on the switching boundary, Σ.

$$\dot{\tilde{V}}(Q) = \begin{cases} \psi \dot{V}^L(Q) + (1-\psi)\dot{V}^H(Q), \psi \in [0,1], & Q_1 \neq \bar{q}_1, Q_2 \neq \bar{q}_2; \\ \psi \dot{Q}_1^L + (1-\psi)\dot{Q}_1^H, \psi \in [0,1], & Q_1 > \bar{q}_1, Q_2 = \bar{q}_2; \\ -\psi \dot{Q}_1^L - (1-\psi)\dot{Q}_1^H, \psi \in [0,1], & Q_1 < \bar{q}_1, Q_2 = \bar{q}_2; \\ \psi \dot{Q}_2^L + (1-\psi)\dot{Q}_2^H, \psi \in [0,1], & Q_1 = \bar{q}_1, Q_2 > \bar{q}_2; \\ -\psi \dot{Q}_2^L - (1-\psi)\dot{Q}_2^H, \psi \in [0,1], & Q_1 = \bar{q}_1, Q_2 < \bar{q}_2. \end{cases} \tag{A.8}$$

where V^L and V^H correspond to the Lyapunov function when $Q_1 < N^*$ and $Q_1 > N^*$, respectively.

In what follows, we will use this generalized Lyapunov theory to prove the global stability in Theorem 4.3. Due to the immense amount of algebra involved in this proof, we only include here the proof for Case 1.2 while noting the rest of the cases (1.1, 1.3, 2.1.1, and 2.1.3) will follow similarly. We need to show that for all $Q \neq \bar{q}$, $\dot{\tilde{V}}(Q) < 0$.

Case 1.2 $q_1^H < (N^* \wedge N) < q_1^L$: Define the stability point as $\bar{q} = (\bar{q}_1(N^*), \bar{q}_2(N^*)) = (N^*, \alpha q_2^L + (1-\alpha)q_2^H)$, where $\alpha = \frac{(N^* \wedge N) - q_1^H}{q_1^L - q_1^H}$. We will suppress the dependence on the control factor, N^*, for notational compactness. Note that $\bar{q}_1 \wedge N \in (q_1^H, q_1^L)$ and $\bar{q}_2 \in (q_2^L, q_2^H)$. Moreover, by assumption 4.1, $N > q_1^H$. There are a number of subcases to consider within each of our two cases:

i. $[Q_1 \neq N^*]$

(a) $Q_1 > \bar{q}_1 = N^*, Q_2 > \bar{q}_2$.

$$\begin{aligned} \dot{\tilde{V}}(Q) &= \dot{V}(Q) = \dot{Q}_1 + \dot{Q}_2 = \lambda + \delta Q_2 - \mu_H(Q_1 \wedge N) - \delta Q_2 + p_H \mu_H(Q_1 \wedge N) \\ &= \lambda - (1-p_H)\mu_H(Q_1 \wedge N) < \lambda - (1-p_H)\mu_H(\bar{q}_1 \wedge q_1^H) = \lambda - (1-p_H)\mu_H q_1^H = 0 \end{aligned}$$

(b) $Q_1 < \bar{q}_1 = N^*, Q_2 < \bar{q}_2$.

$$\begin{aligned}\dot{\tilde{V}}(Q) &= \dot{V}(Q) = -\dot{Q}_1 - \dot{Q}_2 = -\lambda - \delta Q_2 + \mu_L(Q_1 \wedge N) + \delta Q_2 - p_L\mu_L(Q_1 \wedge N) \\ &= -\lambda + (1-p_L)\mu_L(Q_1 \wedge N) < -\lambda + (1-p_L)\mu_L q_1^L = 0\end{aligned}$$

(c) $Q_1 < \bar{q}_1 = N^*, Q_2 > \bar{q}_2$

$$\begin{aligned}\dot{\tilde{V}}(Q) &= \dot{V}(Q) = -\dot{Q}_1 + \dot{Q}_2 = -\lambda - \delta Q_2 + \mu_L(Q_1 \wedge N) - \delta Q_2 + p_L\mu_L(Q_1 \wedge N) \\ &= -\lambda - 2\delta Q_2 + (1+p_L)\mu_L(Q_1 \wedge N) < -\lambda - 2\delta\bar{q}_2 + (1+p_L)\mu_L(\bar{q}_1 \wedge N) \\ &< -\lambda - 2\delta q_2^L + (1+p_L)\mu_L q_1^L = 0\end{aligned}$$

(d) $Q_1 > \bar{q}_1 = N^*, Q_2 < \bar{q}_2$

$$\begin{aligned}\dot{\tilde{V}}(Q) &= \dot{V}(Q) = \dot{Q}_1 - \dot{Q}_2 = \lambda + \delta Q_2 - \mu_H(Q_1 \wedge N) + \delta Q_2 - p_H\mu_H(Q_1 \wedge N) \\ &= \lambda + 2\delta Q_2 - (1+p_H)\mu_H(Q_1 \wedge N) < \lambda + 2\delta\bar{q}_2 - (1+p_H)\mu_H(\bar{q}_1 \wedge q_1^H) \\ &\leq \lambda + 2\delta q_2^H - (1+p_H)\mu_H q_1^H = 0\end{aligned}$$

(e) $Q_1 > \bar{q}_1 = N^*, Q_2 = \bar{q}_2$.

$$\dot{\tilde{V}}(Q) = \dot{Q}_1 = \lambda + \delta Q_2 - \mu_H(Q_1 \wedge N) < \lambda + \delta\bar{q}_2 - \mu_H(\bar{q}_1 \wedge q_1^H) \leq \lambda + \delta q_2^H - \mu_H q_1^H = 0$$

(f) $Q_1 < \bar{q}_1 = N^*, Q_2 = \bar{q}_2$.

$$\dot{\tilde{V}}(Q) = -\dot{Q}_1 = -\lambda - \delta Q_2 + \mu_L(Q_1 \wedge N) < -\lambda - \delta q_2^L + \mu_L q_1^L = 0$$

ii. $[Q_1 = N^*]$ We want to show that for all $\psi \in [0,1]$,$\dot{\tilde{V}}(Q) < 0$:

(a) $Q_1 = \bar{q}_1 = N^*, Q_2 > \bar{q}_2$.

$$\begin{aligned}\dot{\tilde{V}}(Q) &= \psi\dot{Q}_2^L + (1-\psi)\dot{Q}_2^H = \psi[-\delta Q_2 + p_L\mu_L(Q_1 \wedge N)] + (1-\psi)[-\delta Q_2 + p_H\mu_H(Q_1 \wedge N)] \\ &< \psi[-\delta\bar{q}_2 + p_L\mu_L(\bar{q}_1 \wedge N)] + (1-\psi)[-\delta\bar{q}_2 + p_H\mu_H(\bar{q}_1 \wedge N)] \\ &< \psi[-\delta q_2^L + p_L\mu_L q_1^L] + (1-\psi)[-\delta\bar{q}_2 + p_H\mu_H(\bar{q}_1 \wedge N)] = (1-\psi)[-\delta\bar{q}_2 + p_H\mu_H(\bar{q}_1 \wedge N)] \\ &= (1-\psi)\left[p_H\mu_H(N^* \wedge N) - \delta\left(\frac{(N^* \wedge N) - q_1^H}{q_1^L - q_1^H}q_2^L + \frac{q_1^L - (N^* \wedge N)}{q_1^L - q_1^H}q_2^H\right)\right] \\ &= (1-\psi)\left[p_H\mu_H(N^* \wedge N) - \delta\left(\frac{(N^* \wedge N)(q_2^L - q_2^H) - q_1^H q_2^L + q_1^L q_2^H}{q_1^L - q_1^H}\right)\right] \\ &= (1-\psi)\left[p_H\mu_H(N^* \wedge N) - \frac{(N^* \wedge N)(p_L - p_H)\mu_L\mu_H - \lambda(p_L\mu_L - p_H\mu_H)}{(1-p_H)\mu_H - (1-p_L)\mu_L}\right] \\ &= (1-\psi)\left[p_H\mu_H(N^* \wedge N) - (N^* \wedge N)\frac{(p_L - p_H)\mu_L\mu_H - \frac{\lambda}{N^* \wedge N}(p_L\mu_L - p_H\mu_H)}{(1-p_H)\mu_H - (1-p_L)\mu_L}\right] \\ &< (1-\psi)\left[p_H\mu_H(N^* \wedge N) - (N^* \wedge N)\frac{(p_L - p_H)\mu_L\mu_H - \frac{\lambda}{q_1^H}(p_L\mu_L - p_H\mu_H)}{(1-p_H)\mu_H - (1-p_L)\mu_L}\right] \\ &= (1-\psi)\left[p_H\mu_H(N^* \wedge N) - (N^* \wedge N)\frac{(p_L - p_H)\mu_L\mu_H - (1-p_H)\mu_H(p_L\mu_L - p_H\mu_H)}{(1-p_H)\mu_H - (1-p_L)\mu_L}\right] = 0, \ \forall\psi\end{aligned}$$

(b) $Q_1 = \bar{q}_1 = N^*, Q_2 < \bar{q}_2$.

$$\begin{aligned}
\dot{\tilde{V}}(Q) &= -\psi\dot{Q}_2^L - (1-\psi)\dot{Q}_2^H = \psi[\delta Q_2 - p_L\mu_L(Q_1 \wedge N)] + (1-\psi)[\delta Q_2 - p_H\mu_H(Q_1 \wedge N)] \\
&< \psi[\delta\bar{q}_2 - p_L\mu_L(\bar{q}_1 \wedge N)] + (1-\psi)[\delta\bar{q}_2 - p_H\mu_H q_1^H] \\
&\leq \psi[\delta\bar{q}_2 - p_L\mu_L(\bar{q}_1 \wedge N)] + (1-\psi)[\delta q_2^H - p_H\mu_H q_1^H] = \psi[\delta\bar{q}_2 - p_L\mu_L(\bar{q}_1 \wedge N)] \\
&= (1-\psi)\left[-p_L\mu_L(N^* \wedge N) + \delta\left(\frac{(N^* \wedge N) - q_1^H}{q_1^L - q_1^H}q_2^L + \frac{q_1^L - (N^* \wedge N)}{q_1^L - q_1^H}q_2^H\right)\right] \\
&= (1-\psi)\left[p_H\mu_H(N^* \wedge N) - \delta\left(\frac{(N^* \wedge N)(q_2^L - q_2^H) - q_1^H q_2^L + q_1^L q_2^H}{q_1^L - q_1^H}\right)\right] \\
&= (1-\psi)\left[-p_L\mu_L(N^* \wedge N + \frac{(N^* \wedge N)(p_L - p_H)\mu_L\mu_H - \lambda(p_L\mu_L - p_H\mu_H)}{(1-p_H)\mu_H - (1-p_L)\mu_L}\right] \\
&= (1-\psi)\left[-p_L\mu_L(N^* \wedge N) + (N^* \wedge N)\frac{(p_L - p_H)\mu_L\mu_H - \frac{\lambda}{N^* \wedge N}(p_L\mu_L - p_H\mu_H)}{(1-p_H)\mu_H - (1-p_L)\mu_L}\right] \\
&< (1-\psi)\left[-p_L\mu_L(N^* \wedge N) + (N^* \wedge N)\frac{(p_L - p_H)\mu_L\mu_H - \frac{\lambda}{q_1^L}(p_L\mu_L - p_H\mu_H)}{(1-p_H)\mu_H - (1-p_L)\mu_L}\right] \\
&= (1-\psi)\left[-p_L\mu_L(N^* \wedge N) + (N^* \wedge N)\frac{(p_L - p_H)\mu_L\mu_H - (1-p_L)\mu_L(p_L\mu_L - p_H\mu_H)}{(1-p_H)\mu_H - (1-p_L)\mu_L}\right] = 0, \ \forall\psi
\end{aligned}$$

This concludes the proof for the global stability of Case 1.2. ■

B. Markov Chain Performance Measures

$$\begin{aligned}
P(Speedup) &= \pi(0,0)\left(\frac{(1-p_h)\mu_h}{(1-p_l)\mu_l}\right)^{T-1}\sum_j\frac{1}{j!}\left(\frac{\lambda p_l}{(1-p_l)\delta}\right)^j\left(\sum_{i=T}^{N}\frac{1}{i!}\left(\frac{\lambda}{(1-p_h)\mu_h}\right)^i\right. \\
&\quad \left. + \sum_{i=N+1}^{\infty}\frac{1}{N!N^{i-N}}\left(\frac{\lambda}{(1-p_h)\mu_h}\right)^i\right) \\
P(Wait>0) &= \pi(0,0)\left(\frac{(1-p_h)\mu_h}{(1-p_l)\mu_l}\right)^{T-1}\sum_j\frac{1}{j!}\left(\frac{\lambda p_l}{(1-p_l)\delta}\right)^j\sum_{i=N+1}^{\infty}\frac{1}{N!N^{i-N}}\left(\frac{\lambda}{(1-p_h)\mu_h}\right)^i \\
\pi(0,0) &= \left[\sum_j\frac{1}{j!}\left(\frac{\lambda p_l}{(1-p_l)\delta}\right)^j\left(\sum_{i=0}^{T-1}\frac{1}{i!}\left(\frac{\lambda}{(1-p_l)\mu_l}\right)^i + \left(\frac{(1-p_h)\mu_h}{(1-p_l)\mu_l}\right)^{T-1}\sum_{i=T}^{N}\frac{1}{i!}\left(\frac{\lambda}{(1-p_h)\mu_h}\right)^i\right.\right. \\
&\quad \left.\left. + \left(\frac{(1-p_h)\mu_h}{(1-p_l)\mu_l}\right)^{T-1}\sum_{i=N+1}^{\infty}\frac{1}{N!N^{i-N}}\left(\frac{\lambda}{(1-p_h)\mu_h}\right)^i\right)\right]^{-1}.
\end{aligned} \tag{B.1}$$

References

Anand, K., M. F. Pac, S. Veeraraghavan. 2010. Quality-speed conundrum: Tradeoffs in customer-intensive services. *Management Science* **57** 40–56.

Armony, M., C. Maglaras. 2004. Contact centers with a call-back option and real-time delay information. *Operations Research* **52** 527–545.

Ata, B., S. Shneorson. 2006. Dynamic control of an M/M/1 service system with adjustable arrival and service rates. *Management Science* **52** 1778–1791.

Bekker, R., S.C. Borst. 2006. Optimal admission control in queues with workload-dependent service rates. *Probability in the Engineering and Informational Sciences* **20** 543–570.

Bekker, R., O.J. Boxma. 2007. An M/G/1 queue with adaptable service speed. *Stochastic Models* **23** 373–396.

Boxma, O.J., M. Vlasiou. 2007. On queues with service and interarrival times depending on waiting times. *Queueing Systems* **56** 121–132.

de Vericourt, Francis, Yong-Pin Zhou. 2005. Managing response time in a call-routing problem with service failure. *Operations Research* **53**(6) 968–981.

di Bernardo, M., C.J. Budd, A.R. Champneys, P. Kowalczyk. 2008. *Piecewise-Smooth Dynamical Systems: Theory and Applications*. Springer.

Durbin, C.G., R.F. Kopel. 1993. A case-control study of patients readmitted to the intensive care unit. *Critical Care Medicine* **21** 1547–1553.

Filippov, A.F. 1988. *Differential Equations with Discontinuous Righthand Sides*. Kluwer Academic Publishers, Dortrecht.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* 79–141.

Gibbens, R. J., P. J. Hunt, F. P. Kelly. 1990. Bistability in communication networks. *Disorder in Physical Systems*. Oxford University Press, 113–128.

Glazebrook, K.D., L.R. Whitaker. 1992. Single-machine stochastic scheduling with dependent processing times. *Advances in Applied Probability* **24** 635–652.

Green, L. V., J. Soares, J. F. Giglio, R. A. Green. 2006. Using queuing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* **13** 61–68.

Guckenheimer, J., P. Holmes. 2002. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer.

Hasija, S., E. Pinker, R. A. Shumsky. 2010. Om practice–work expands to fill the time available: Capacity estimation and staffing under parkinson's law. *Manufacturing & Service Operations Management* **12** 1–18.

Huang, J., B. Carmeli, A. Mandelbaum. 2012. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. Working paper, Technion.

Jennings, O.B., F. de Véricourt. 2008. Dimensioning large-scale membership services. *Operations Research* **56**(1) 173–187.

Kc, D., C. Terwiesch. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* **14** 50–65.

Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Scaling relations and approximate solutions. *Management Science* **49** 1018–1038.

Mandelbaum, A., W. Massey, M. Reiman. 1998. Strong approximations for markovian service networks. *Queueing Systems* **30**(1-2) 149–201.

Mandelbaum, A., G. Pats. 1998. State-dependent stochastic networks. part I: Approximations and applications with continuous diffusion limits. *The Annals of Applied Probability* **8**(2) 569–646.

Perry, O., W. Whitt. 2011. An ODE for an overloaded X model involving a stochastic averaging principle. *Stochastic Systems* **1** 17–66.

Powell, S. G., K. L. Schultz. 2004. Throughput in serial lines with state-dependent behavior. *Management Science* **50** 1095–1105.

Shevitz, D., B. Paden. 1994. Lyapunov stability theory of nonsmooth systems. *IEEE Transactions on Automatic Control* **39**(9) 1910–1914.

Whitt, W. 1990. Queues with service times and interarrival times depending linearly and randomly upon waiting times. *Queueing Systems* **6** 335–352.

Yankovic, N., L. V. Green. 2011. Identifying Good Nursing Levels: A Queuing Approach. *Operations Research* 942–955.

Yom-Tov, G., A. Mandelbaum. 2013. Erlang-R: A time-varying queues with reentrant customers, in support of health-care staffing. *MSOM, to appear* .

Zhan, D., A. R. Ward. 2013. Routing to minimize waiting and callbacks in large call centers. *MSOM, to appear* .