Journal Pre-proof

Comparison of Fluid Approximations for Service Systems with State-Dependent Service Rates and Return Probabilities

Armann Ingolfsson, Eman Almehdawe, Ali Pedram, Monica Tran

PII: S0377-2217(19)30953-1
DOI: https://doi.org/10.1016/j.ejor.2019.11.041
Reference: EOR 16177

To appear in: *European Journal of Operational Research*

Received date: 4 October 2018
Accepted date: 20 November 2019

Please cite this article as: Armann Ingolfsson, Eman Almehdawe, Ali Pedram, Monica Tran, Comparison of Fluid Approximations for Service Systems with State-Dependent Service Rates and Return Probabilities, *European Journal of Operational Research* (2019), doi: https://doi.org/10.1016/j.ejor.2019.11.041

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2019 Published by Elsevier B.V.

Journal Pre-proof

Highlights

- We analyze the impact of inserting a delay between service completion and return
- Inserting a constant delay does not change fluid approximation equilibrium points
- A constant delay does changes the fluid approximation transient behavior
- Simulations indicate less impact on transient behavior if delay is exponential
- Bistable behavior can occur if return and service rates are state-dependent

Journal Pre-proof

Journal Pre-proof

Comparison of Fluid Approximations for Service Systems with State-Dependent Service Rates and Return Probabilities

Armann Ingolfsson*

Alberta School of Business, University of Alberta, Edmonton, AB T6G 2R6, Canada

Eman Almehdawe

University of Regina, Regina, SK, S4S 0A2, Canada

Ali Pedram, Monica Tran

Alberta School of Business, University of Alberta, Edmonton, AB T6G 2R6, Canada

Journal Pre-proof

Abstract

We compare two models of a multi-server queueing system with state-dependent service rates and return probabilities. In both models, upon completing service, customers are delayed prior to possibly returning to service. In one model, the determination of whether a customer will return occurs immediately upon service completion, at the beginning of the delay. In the other, that determination is made at the end of the delay, capturing the idea that it takes time for the customer's condition and needs to evolve or assess, before it becomes known whether a return to service is needed. Our comparison focuses on fluid approximations of the two models. The fluid approximation for the first model, which has been studied previously, consists of a system of two ordinary differential equations. The fluid approximation for the second model, which is new, consists of a delay differential equation. We find that the two fluid approximations have the same set of equilibrium points, but their transient behavior can differ markedly. Both fluid approximations can exhibit bistability for certain parameter values. We use discrete event simulation to illustrate the extent to which the findings from the fluid approximations carry over to the underlying stochastic models.

Keywords: Queueing; simulation; fluid approximation; delay differential equations.

*Corresponding author

Email addresses: armann.ingolfsson@ualberta.ca (Armann Ingolfsson), eman.almehdawe@uregina.ca (Eman Almehdawe), pedram@ualberta.ca (Ali Pedram), mt4@ualberta.ca (Monica Tran)

Journal Pre-proof

1. Introduction

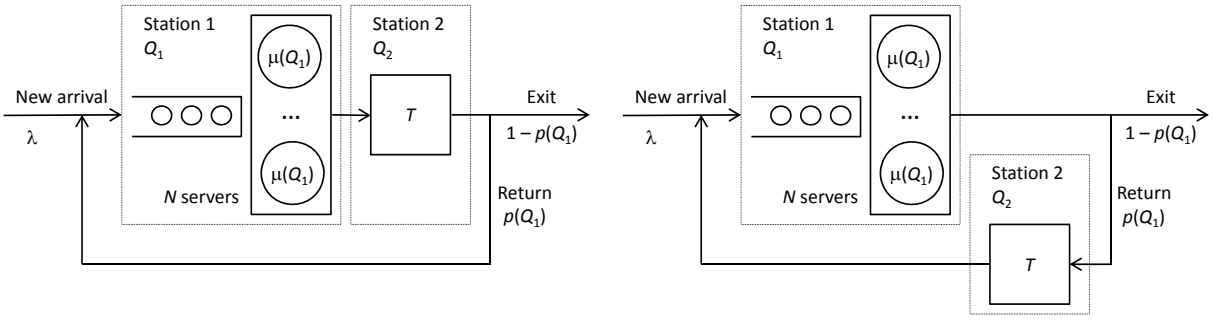
We study multi-server queueing systems with returns—systems in which, after completing a service, some customers return for another service, after a delay. Returns occur in a variety of contexts, including patient returns in intensive care units (ICUs) and emergency departments (EDs) in hospitals, part rework in manufacturing systems, and customer returns in contact centres. In addition to returns, our model features service rates and return probabilities that depend on the system congestion. These state-dependent rates allow us to investigate the impact on system performance of speedup accompanied by higher return probabilities—characteristics that are consistent with recent empirical evidence for ICUs. Our model differs from previous work in that we assume that (1) the determination of whether a customer will return occurs at the *end* of the delay between one service and the next, while (2) the return probability is determined by the system occupancy at the *beginning* of the delay.

We develop a fluid approximation of a system with returns and state-dependent rates based on a delay differential equation (DDE). We use the fluid approximation to study the characteristics of the transient and steady-state system behavior and we use a discrete event simulation (DES) model to assess the accuracy of the fluid approximation. We compare our model and its fluid approximation to a previous model (Chan et al., 2014) and its fluid approximation, in which the determination of whether a customer will return occurs at the beginning of the delay. In contrast to our DDE fluid approximation, the Chan et al. (2014) fluid approximation consists of a system of two ordinary differential equations (ODE).

Figure 1 provides diagrams of the two models. We view the system as a queueing network with two stations, and we refer to models corresponding to the two panels of Figure 1 as Model (a) (this is "our model") and Model (b) (this is the model from Chan et al. (2014)). In Model (a), we take the viewpoint that the probability that a customer returns for further service depends on the number of customers in Station 1 at the end of that customer's previous service and that it takes time for the customer's condition to either be measured or to evolve to the point where further service is needed. Therefore, it is not known whether a customer will return to service until after a delay.

A queueing system, in general, involves customers who arrive from a population, wait in line, receive service, return to the population, and potentially return to the queue at a later time.

Journal Pre-proof



(a) Delay before return routing.

(b) Return routing before delay.

Figure 1: Queueing network diagrams.

In this sense, all queueing systems involve returns and classical models of finite-source queueing systems model returns explicitly. In the settings that we focus on, however, a customer arrives from the population with a single *issue* that may require multiple *service episodes* before the issue is resolved. In these settings, issues that require a customer to join the queue occur infrequently for any given customer, but the service episodes for a particular issue are closely spaced in time.

The settings that we focus on include ICUs (KC & Terwiesch, 2012; Hu et al., 2018), with each ICU stay within a single hospital stay viewed as one service episode; hospital wards (Shi et al., 2019), with each hospital stay viewed as one service episode; manufacturing facilities, with each instance of rework (Owen & Blumenfeld, 2008) for a single unit of product viewed as one service episode; contact centers, where service episodes could take place via email, phone, or instant messaging (de Véricourt & Zhou, 2005; Tezcan & Zhang, 2014); and prisons, with each prison stay viewed as one service episode (Master et al., 2018). In all of these settings, one can envision customers flowing through a two-node queueing network akin to the one illustrated in Figure 1a, in which Station 1 is where customers receive service and Station 2 is where customers are delayed prior to returning to service.

2. Literature Review

Several researchers have recently formulated and analyzed models of service systems with returns. These models differ in many ways, including the following:

Admission of new customers: Some assume that new customers who arrive when the service system is at capacity are lost (Yom-Tov & Mandelbaum, 2014); others assume that new

Journal Pre-proof

customers wait for the first available server, either in a first-come-first-serve (FCFS) queue (Chan et al., 2014) or a priority queue (Barjesteh & Abouee-Mehrizi, 2018).

Routing of returning customers: Some assume that all service episodes for a given issue are with the same server (Campello et al., 2017); others assume pooling of servers (Yankovic & Green, 2011).

State-dependent rates: Most assume a constant service rate and a constant return probability but Chan et al. (2014) assume that the service rate and return probability increase when the number of busy servers is above a threshold and Barjesteh & Abouee-Mehrizi (2018) allow the service rate and return probability to depend in a more general fashion on Station 1 occupancy.

Table 1 compares several published models. The primary new feature in our Model (a) is that the determination of whether a customer will return is made after a delay.

Citation	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
de Véricourt & Jennings (2008)	multiple	closed	yes	N/A	random	no	no	single
de Véricourt & Jennings (2011)	multiple	closed	yes	N/A	random	no	no	single
Luo & Zhang (2013)	multiple	open	no	N/A	PS	yes	N/A	single
Tezcan & Zhang (2014)	multiple	open	no	N/A	PS	yes	N/A	single
Dong et al. (2015)	multiple	open	N/A	no return	N/A	yes	N/A	single
Apte et al. (1999)	multiple	open	no	N/A	zero	no	no	single
de Véricourt & Zhou (2005)	multiple	open	yes	N/A	zero	no	no	single
Zhan & Ward (2013)	multiple	open	yes	N/A	zero	no	no	single
Huang et al. (2015)	multiple	open	yes	N/A	zero	no	no	multiple
Furman et al. (2019)	multiple	open	yes	before	random	no	no	multiple
Owen & Blumenfeld (2008)	single	open	N/A	before	deterministic	no	no	single
Nakamura (1971)	single	open	N/A	before	random	no	no	single
Saghafian et al. (2014)	single	open	N/A	before	random	no	no	multiple
Mandelbaum et al. (1998), Section 5	multiple	open	yes	before	random	yes	yes	single
Yankovic & Green (2011)	multiple	open	yes	before	random	no	no	single
Yom-Tov & Mandelbaum (2014)	multiple	open	yes	before	random	no	no	single
Campello et al. (2017)	multiple	open	no	before	random	no	no	single
Chan et al. (2014)	multiple	open	yes	before	random	yes	yes	single
Barjesteh & Abouee-Mehrizi (2018)	multiple	open	yes	before	random	yes	yes	multiple
Model (a)	multiple	open	yes	after	random	yes	yes	single
Model (a) fluid approximation	multiple	open	yes	after	deterministic	yes	yes	single

Table 1: Summary of related models. Column headings: (1) number of servers, (2) open or closed network, (3) are returning customers pooled, (4) does the determination of whether a customer returns occur before or after the delay, (5) is the delay random, deterministic, zero, or modeled through processor sharing (PS), (6) are service rates state-dependent, (7) are return probabilities state-dependent, (8) single or multiple customer classes.

There is extensive empirical evidence, summarized in Delasay et al. (2018), indicating that service rates depend on system load through a variety of mechanisms. The evidence is less extensive regarding state-dependence of return probabilities, but several studies (Anderson et al., 2012;

Journal Pre-proof

Town et al., 2014; Chrusch et al., 2009) have shown that ICU readmission is associated with high ICU occupancy at the time of ICU discharge (consistent with the assumption in our model that the determination of whether a customer returns is influenced by the system occupancy at the beginning of the delay). Other studies have shown that ICU readmission is associated with earlier-than-predicted ICU discharge (KC & Terwiesch, 2012) and with after-hours ICU discharge (Utzolino et al., 2010). These studies provide indirect evidence that return probabilities depend on occupancy, assuming that earlier-than-predicted and after-hours discharges are more likely when ICU occupancy is high. In a manufacturing setting, Owen & Blumenfeld (2008) argue that the probability of rework will increase with machine speed.

Model (a), which is motivated by features of ICUs, differs from previous work primarily in the assumptions that we make about the delay that elapses before a customer returns to the queue to wait for another service. If a service failure is the *cause* and a return to service is the *effect*, then we take the viewpoint that the delay occurs because it takes time for it to become known whether a return to service is necessary. Common documented reasons for ICU readmissions, such as "complications arising from treatment" and "onset of new medical conditions" (Makris et al., 2010) are consistent with this viewpoint. Similar types of delayed feedback in a variety of physical, biological, and social systems have increasingly been modeled using DDEs. DDEs have been studied since the 1980s as models of systems with delayed feedback (Shampine & Thompson, 2009), including predator-prey systems, in which the predator birth rate depends on the predator and prey populations after a maturation delay (Faria, 2001) and the dynamics of epidemics, in which the infection rate depends on the population of infected people after an infection delay (Beretta et al., 1998).

DDEs have only rarely (e.g., Johari & Tan, 2001; Pender et al., 2017, 2018) been used to model queueing systems. Most fluid approximations of queueing models, including ones of service systems with returns (Chan et al., 2014; Barjesteh & Abouee-Mehrizi, 2018), can be represented as ODEs. Qualitative differences between DDEs and ODEs include the fact that DDE initial-value problems require one to specify the *history* of the state variables over a time interval of positive length, rather than simply values at a single point in time for ODEs, and that with DDEs, discontinuities in the state variables or their derivatives can be propagated forward in time, rather than being smoothed out as in ODEs.

Journal Pre-proof

It follows from the viewpoint that service failure causes return to service after a delay that it will not be known whether a return is necessary until at *the conclusion* of the delay and our Model (a) is consistent with this fact. In contrast, in the models in Chan et al. (2014) and Barjesteh & Abouee-Mehrizi (2018) the determination of whether a return to service occurs happens at *the beginning* of the delay.

Our primary findings are that the two fluid approximations that we study have an identical set of equilibrium points but that their transient behavior can differ markedly. Both models can exhibit bistability. Simulation experiments indicate that the accuracy of the fluid approximations increases with system size.

We define Models (a) and (b) in Section 3, discuss assumptions regarding the service rate and return probability functions in Section 4, and define fluid approximations in Section 5. We analyze fluid approximation equilibrium points in Section 6 and discuss their transient behavior in Section 7. We use simulation to demonstrate that Model (a) can exhibit bistable behavior in Section 8. Section 9 concludes.

3. Queueing Models

We formulate two stochastic models of an N-server queueing system with returns, in which both the service rate and the return probability depend on the number of customers that are receiving service or waiting. In this section, we elaborate on the formulation of the models, with reference to Figure 1. We denote the number of customers at Station i as $\mathbb{Q}_i(t), i = 1, 2$ and the number of busy servers at Station 1 as $\mathbb{B}(t) = \min(\mathbb{Q}_1(t), N)$. We use $\mathbb{X}(t)$ for a stochastic process and $X(t)$ for a fluid approximation to that stochastic process. We will sometimes use (a) and (b) as superscripts on the state variables $\mathbb{Q}_i(t), i = 1, 2$ and their fluid approximations $Q_i(t), i = 1, 2$.

The following assumptions are common to Models (a) and (b): New customers arrive to Station 1 according to a Poisson process with rate λ. Station 1 has N servers. Busy Station 1 servers serve customers at Markovian rate $\mu(\mathbb{Q}_1(t))$ per server. Some customers are delayed at Station 2 and the delay T is a random variable, with mean $\mathbb{E}[T] = \tau$. Chan et al. (2014) assume that T is exponentially distributed. We use deterministic, Erlang, and exponential distributions for T in our simulation experiments.

No new customers arrive to Station 2. Customers who return to Station 1 wait in a FCFS

Journal Pre-proof

infinite-capacity queue together with new customers and they receive service at the same rate as new customers.

Modeling delay before return:. In Model (a), after completing service at Station 1 all customers move to Station 2. Upon exit from Station 2 at time t, customers return for additional service from Station 1 with probability $p(\mathbb{Q}_1^{(a)}(t-T))$; otherwise, customers leave the system.

In Model (b), after completing service at Station 1 at time t, customers move to Station 2 with probability $p(\mathbb{Q}_1^{(b)}(t))$; otherwise customers leave the system. All customers who move to Station 2 return to Station 1 for additional service.

In Model (a), the delay occurs *before* it is determined whether the customer will return. In contrast, in previously published models (Chan et al., 2012; Barjesteh & Abouee-Mehrizi, 2018), the delay occurs *after* it is determined that the customer will return, as in Model (b). Modeling a delay before the determination of whether a customer will return is realistic in certain settings, which motivates our investigation, but it also makes the model more complicated, because the system evolution at time t becomes dependent on the system state at time $t-T$.

Model (a) is consistent with settings in which it does not become clear whether a customer needs to return until after a delay, during which the customer condition either changes or is measured. In an ICU, for example, patients are typically discharged to a "step-down unit." If a patient's condition deteriorates while in the step-down unit, then the patient may need to return to the ICU. In a manufacturing setting, inspection to determine whether a unit requires rework takes time, as another example. In these settings, modeling the delay as occurring after a customer is routed towards returning for service, as in Model (b) and in Chan et al. (2012) and Barjesteh & Abouee-Mehrizi (2018), underestimates the number of customers experiencing the delay.

Journal Pre-proof

4. Service Rate and Return Probability Functions:

We study situations in which the functions $\mu(x)$ and $p(x)$ are non-decreasing. Chan et al. (2014) used two-value step functions:

$$\mu(x) = \begin{cases} \mu_L, & x < N_\mu^* \\ \mu_H, & x \geq N_\mu^* \end{cases}, \quad p(x) = \begin{cases} p_L, & x < N_p^* \\ p_H, & x \geq N_p^* \end{cases}, \tag{1}$$

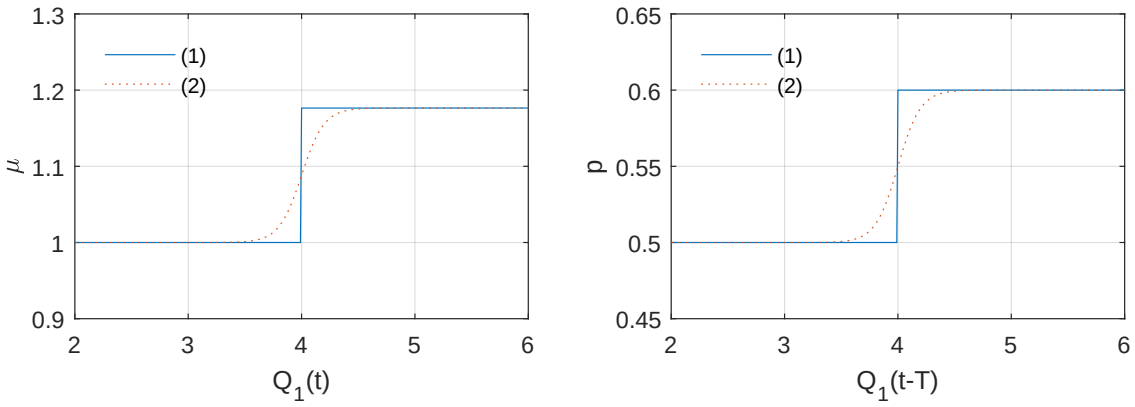


Figure 2: Two-value service rate and return probability functions (1) and logistic function approximation (2).

where $N^*_\mu = N^*_p = N^* \leq N$, $\mu_L < \mu_H$ and $p_L < p_H$. The interpretation is that if N^* or more servers are busy, then service speeds up, but the return probability increases.

Although the functions in (1) are simply stated, they are discontinuous with respect to their argument, x, which causes difficulties for the numerical solution and theoretical analysis of DDEs. Numerically, one has to enumerate time points at which the solution is discontinuous (Shampine & Thompson, 2009), which complicates programming and increases computation time.

An alternative is a logistic function approximation of (1) that is continuous in its argument:

$$\begin{aligned} \mu(x) &= \mu_L + \frac{\mu_H - \mu_L}{1 + \exp(-K_\mu(x - N^*_\mu))} \\ p(x) &= p_L + \frac{p_H - p_L}{1 + \exp(-K_p(x - N^*_p))}, \end{aligned} \tag{2}$$

where higher values of the additional parameters K_μ and K_p cause (2) to be closer to (1). Figure 2 compares (1) and (2) for a base case that we use in Section 7, with an arrival rate of $\lambda = 5$/day, $N = 11$ servers, a switching point of $N^* = N^*_\mu = N^*_p = 4$ servers, service rates of $\mu_L = 1$ and $\mu_H = 1/0.85 = 1.18$ per day, return probabilities of $p_L = 0.5$ and $p_H = 0.6$, and $K_\mu = K_p = 10$.

Conditions on μ and p:. In our analysis of equilibrium points and their stability for the fluid approximations, we investigate the consequences of four conditions on the functions μ and p and their derivatives, μ' and p'. Our first two conditions are that both functions are positive, bounded,

Journal Pre-proof

Journal Pre-proof

differentiable, and strictly increasing:

$$\mu(x) \in (0, \infty),\ \mu'(x) > 0 \text{ for } x \geq 0, \tag{3}$$

$$p(0) \in (0, 1),\ p'(x) > 0 \text{ for } x \geq 0 \tag{4}$$

Our second two conditions are expressed in terms of the product $\nu(x) \equiv \mu(x)(1 - p(x))$ (with derivative ν'), which is the rate at which customers leave the queueing network (after a delay at Station 2 in Model (a); after service completion at Station 1 in Model (b)).

The third condition is a stability condition:

$$\text{There exists } \tilde{x} > 0 \text{ such that if } x > \tilde{x}, \text{ then } N\nu(x) = N\mu(x)(1 - p(x)) > \lambda \tag{5}$$

The fourth condition is that the leaving rate is strictly increasing:

$$\nu'(x) = \mu'(x)(1 - p(x)) - \mu(x)p'(x) > 0 \text{ for } x \geq 0 \tag{6}$$

Condition (6) is stronger than Conditions (3)-(4). We will see that Condition (6) is a sufficient condition for the fluid approximations to have unique equilibrium points.

5. Fluid Approximations

We define $Q_1^{(a)}(t)$ and $B^{(a)}(t) = \min(Q_1^{(a)}(t), N)$ to be fluid approximations to $\mathbb{Q}_1^{(a)}(t)$ and $\mathbb{B}^{(a)}(t) = \min(\mathbb{Q}_1^{(a)}(t), N)$. The fluid arrives at a constant rate of λ to Station 1. The fluid is consumed at rate $B^{(a)}(t)\mu(Q_1^{(a)}(t))$. After service, customers are delayed by T, which is assumed constant and equal to τ in the Model (a) fluid approximation. After the delay, at time t, customers return to service with probability $p(Q_1^{(a)}(t - \tau))$.

The fluid amount $Q_1^{(a)}(t)$ changes as follows in an infinitesimal time interval $(t, t + \epsilon]$:

New arrivals: $\lambda\epsilon$ is added to $Q_1^{(a)}(t)$

Service completions: $B^{(a)}(t)\mu(Q_1^{(a)}(t))\epsilon$ is removed from $Q_1^{(a)}(t)$

Returns to service: $B^{(a)}(t - \tau)\mu(Q_1^{(a)}(t - \tau))p(Q_1^{(a)}(t - \tau))\epsilon$ is added to $Q_1^{(a)}(t)$

Journal Pre-proof

The resulting delay differential equation (DDE) that captures these system dynamics and corresponds to Model (a) is:

$$\frac{d}{dt}Q_1^{(a)}(t) = \lambda - B^{(a)}(t)\mu(Q_1^{(a)}(t)) + B^{(a)}(t-\tau)\mu(Q_1^{(a)}(t-\tau))p(Q_1^{(a)}(t-\tau)) \tag{7}$$

In general, in DDEs, the current value of a variable ($Q_1^{(a)}(t)$) influences not only the derivative of the variable at the current time ($\frac{d}{dt}Q_1^{(a)}(t)$), but also at one or more future times ($\frac{d}{dt}Q_1^{(a)}(t+\tau)$ in our setting), after a set of delays or time lags. In contrast to ODEs, which require a single value to specify an initial condition, for DDEs one needs to specify a *history*, consisting of an infinite set of initial values, corresponding to all past time points that can influence the first value of the derivative to be computed, at $t = 0$. For our DDE (7), it suffices to specify $Q_1^{(a)}(t)$ for $t \in [-\tau, 0]$. Typically, for brevity and simplicity, we specify $Q_1^{(a)}(t)$ to be equal to a constant value for all $t < 0$.

We reproduce the fluid approximation from Chan et al. (2014), generalized to arbitrary service rate and return probability functions, to facilitate comparison. This fluid approximation corresponds to Model (b):

$$\frac{d}{dt}Q_1^{(b)}(t) = \lambda - B^{(b)}(t)\mu(Q_1^{(b)}(t)) + Q_2^{(b)}(t)\delta, \tag{8}$$

$$\frac{d}{dt}Q_2^{(b)}(t) = B^{(b)}(t)\mu(Q_1^{(b)}(t))p(Q_1^{(b)}(t)) - Q_2^{(b)}(t)\delta, \tag{9}$$

where $\delta \equiv 1/\tau$.

Recall that $Q_2^{(b)}(t)$ is the number of customers in Station 2, assuming that it becomes known whether a customer who has completed service requires a return to service *before* the delay, which is consistent with Model (b) but not with Model (b). We are primarily interested in the Station 1 occupancy, $Q_1^{(b)}(t)$, however.

The following theorem summarizes conditions that guarantee the existence and uniqueness of solutions to the DDE (7) and the ODEs (8)-(9):

Theorem 5.1. *(a) If the history, $Q_1^{(a)}(t)$ for $t \in [-\tau, 0]$, is continuous and bounded, then (7) has a unique solution for $t \geq 0$.*

(b) If $Q_1^{(b)}(0), Q_2^{(b)}(0) \in [0, \infty)$, then the system (8)-(9) has a unique solution for $t \geq 0$.

Proof: See Appendix B.

Journal Pre-proof

6. Fluid Approximation Equilibrium Points and Stability

We present four theorems that characterize equilibrium points and their stability for the Model (a) and Model (b) fluid approximations. All proofs are in Appendix C. The first two theorems are for constant service rate and return probability.

Theorem 6.1. *Model (a) fluid approximation, constant μ and p: If $\mu > 0$ and $0 < p < 1$ are constant, then the DDE (7) has a unique equilibrium point $\bar{Q}_1^{(a)} = \lambda/\nu$ if and only if $\lambda \leq N\nu$. That equilibrium point is locally stable if $\lambda < N\nu$.*

Theorem 6.2. *Model (b) fluid approximation, constant μ and p: If $\mu > 0$ and $0 < p < 1$ are constant, then the ODE system (8)-(9) has a unique equilibrium point $\left(\bar{Q_1}^{(b)}, \bar{Q_2}^{(b)}\right) = \left(\frac{\lambda}{\nu}, \frac{\tau p\lambda}{1-p}\right)$ if and only if $\lambda \leq N\nu$. That equilibrium point is locally stable if $\lambda < N\nu$.*

The conditions in Theorems 6.1-6.2 are not a special case of Conditions (3)-(4), because the latter conditions require $\mu(x)$ and $p(x)$ to be strictly increasing in x.

Theorem 6.1 does not provide an equilibrium value for $Q_2^{(a)}$, but we can derive one using Little's Law. Assume that the Model (a) fluid system has reached equilibrium, at $\bar{Q}_1^{(a)} = \lambda/\nu$. The total number of visits by a customer to Station 1 is geometrically distributed with expected value $1/(1-p)$, and therefore the total arrival rate to Station 1 (new arrivals and returns combined) is $\lambda/(1-p)$ and this is also the arrival rate to Station 2. The time spent in Station 2 is τ. Therefore, Little's Law implies that $\bar{Q}_2^{(a)} = (\text{arrival rate})(\text{time in Station 1}) = \frac{\tau\lambda}{1-p}$. We see that $\bar{Q}_2^{(b)} = p\bar{Q}_2^{(a)}$, which implies $\bar{Q}_2^{(b)} < \bar{Q}_2^{(a)}$, as expected, because in the Model (a) fluid system, all customers are delayed in Station 2 before some of them exit the system, whereas in the Model (b) fluid system, the customers who exit the system do so before entering Station 2.

The next two theorems are for systems with state-dependent service rate and return probability functions that satisfy Conditions (3)-(5) and possibly also Condition (6).

Theorem 6.3. *Model (a) fluid approximation, state-dependent μ and p: If the functions $\mu(x)$ and $p(x)$ satisfy Conditions (3)-(5), then the DDE (7) has at least one equilibrium point x, which is a solution to the equation $\min(x, N)\nu(x) = \lambda$. If Condition (6) is added and if $x \neq N$, then (7) has a unique locally stable equilibrium point.*

Theorem 6.4. *Model (b) fluid approximation, state-dependent μ and p: If the functions $\mu(x)$ and $p(x)$ satisfy Conditions (3)-(5), then the ODE system (8)-(9) has at least one equilibrium*

Journal Pre-proof

point (x, y)*, where* x *is a solution to the equation* $\min(x, N)\nu(x) = \lambda$ *and* $y = \tau\mu(x)p(x)x$*. If Condition (6) is added and if* $x \neq N$*, then (8)-(9) has a unique locally stable equilibrium point.*

We see that under Conditions (3)-(5), the Model (a) and Model (b) fluid approximations have the same equilibrium values for Q_1, and if Condition (6) is added, that equilibrium value is unique and locally stable, at least if the equilibrium value does not equal N. (The condition $x \neq N$ is needed because we use standard proof techniques for stability of a differential equation, which require that the right side of the equation be continuously differentiable with respect to $Q_1^{(a)}$ or $Q_1^{(b)}$.)

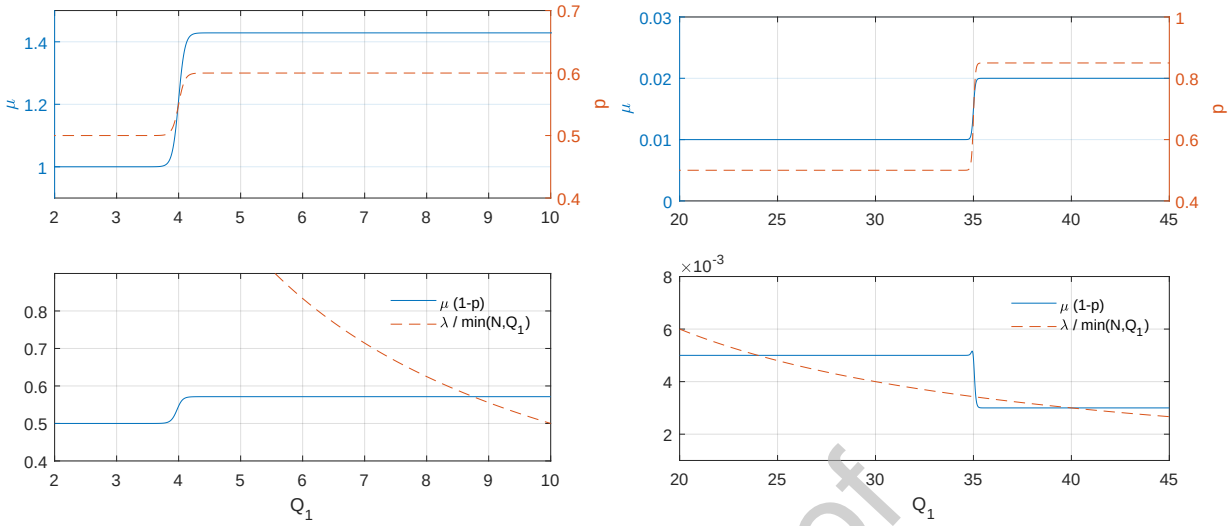
It is perhaps surprising that, under Conditions (3)-(6), the two fluid approximations have the same unique and locally stable equilibrium value for Q_1, whose value is independent of the delay, τ. The system state and, therefore, the return probability, could change drastically during a customer's delay at Station 2, especially if the delay is long. One might therefore expect that equilibrium values would depend on whether the return event occurs at the beginning or at the end of the delay. The fact that the equilibrium value is independent of the delay is consistent with the *snapshot principle* heavy-traffic approximation (Whitt, 2002, p. 187): That the system state remains constant during a customer's processing time (the delay τ in our setting).

We caution that the results that we have proven are for deterministic fluid approximations, in which the system state remains constant, indefinitely, once an equilibrium point is reached. It is a topic for future research to determine whether the conclusions of Theorems 6.3-6.4 continue to hold for the stochastic versions of Models (a) and (b).

Figure 3 shows two examples of service rate and return probability functions, and the resulting equilibrium points for Q_1. Figure 3a shows an example where Condition (6) holds, and a unique equilibrium point $x = 8.75$ is found by solving $\lambda/\min(N, x) = \nu(x)$. Figure 3b shows an example where Condition (6) does not hold, and the equation $\lambda/\min(N, x) = \nu(x)$ has 3 solutions, at $x = 24, 35.122$, and 40. We elaborate on the former example in Section 7 and we elaborate on the latter example in Section 8.

7. Transient Behavior

In this section we illustrate the transient behavior of the Model (a) fluid approximation and compare to the transient behavior of the Model (b) fluid approximation. We use Matlab's dde23



(a) $\lambda = 5, N = 11, N^* = 4, 1/\mu_L = 1, p_L = 0.5, 1/\mu_H = 0.7, p_H = 0.6, K_\mu = K_p = 20.$

(b) $\lambda = 0.12, N = 45, N^* = 35, 1/\mu_L = 100, p_L = 0.5, 1/\mu_H = 50, p_H = 0.85, K_\mu = K_p = 20.$

Figure 3: Examples of service rate and return probability functions.

solver (see Shampine & Thompson (2001)) to solve the DDE and the ode23 solver to solve the ODEs. We set the RelTol parameter to 10^{-6} and the AbsTol parameter to 10^{-7} throughout.

We simulate Models (a) and (b) using the Arena software and we present details in Appendix A. We use four separate random number streams in the models, for inter-arrival times of new arrivals, service times, returns, and duration of delays before return (if the delay is random). When comparing sample paths for Models (a) and (b) we use the same random number streams for corresponding elements in the two models.

We begin with a base case with constant parameters: An arrival rate of $\lambda = 5$/day, $N = 11$ servers, an average service time of $1/\mu = 1$ day, an average delay of $\tau = 1/\delta = 10$ days, and a return probability of $p = 0.5$. We assume that $Q_1^{(a)}(t) = 0$ for $t < 0$ and that $Q_1^{(b)}(0) = Q_2^{(b)}(0) = 0$.

In Figure 4a, we see the transient solutions. Both fluid approximations approach the same steady-state value of $\lambda/(\mu(1-p)) = \lambda/\nu = 10$. The Model (b) fluid approaches the steady-state value smoothly, whereas the Model (a) fluid goes through stages, in which the discontinuity in $dQ_1^{(a)}(t)/dt$ at $t = 0$ is propagated to discontinuities in $dQ_1^{(a)}(t)/dt$ that occur at $t = \tau, 2\tau, \cdots = 10, 20, \ldots$. In the first stage, for $t \in [0, \tau]$, no customers have returned, and $Q_1^{(a)}(t)$ approaches $\lambda/\mu = 5$, that is, the steady state average occupancy in the absence of returns. In the second stage, for $t \in [\tau, 2\tau]$, a proportion $p = 0.5$ of customers return once, and $Q_1^{(a)}(t)$ approaches

Journal Pre-proof

Journal Pre-proof

$\lambda/\mu(1+p) = 7.5$, that is, the steady state average occupancy if 50% of customers return once. In the third stage, $Q_1^{(a)}(t)$ approaches $\lambda/\mu(1+p+p^2) = 8.75$, and so on, until in the limit when $t \to \infty$, $Q_1^{(a)}(t)$ approaches $\lambda/\mu(1+p+p^2+\cdots) = \lambda/(\mu(1-p)) = 10$.

To gain further insight into the two transient solutions, we obtain closed-form solutions for $t \in [0, 20]$. For Model (b), standard analysis provides this solution:

$$Q_1^{(b)}(t) = 10 - 5.498e^{-0.0475t} - 4.502e^{-1.0525t},$$

which is valid for all $t \geq 0$. This expression confirms that $\lim_{t\to\infty} Q_1^{(b)}(t) = 10$.

For Model (a), we use the *method of steps* (Smith, 2011, Chapter 3). For any $t \in [0, \tau] = [0, 10]$, $B^{(a)}(t-\tau) = 0$, and therefore, on this interval, (7) reduces to:

$$\frac{d}{dt}Q_1^{(a)}(t) = \lambda - B^{(a)}(t)\mu(Q_1^{(a)}(t)) = \lambda - \min(Q_1^{(a)}(t), N)\mu = 5 - Q_1^{(a)}(t),$$

as long as $Q_1^{(a)}(t) < N = 11$. This is an ODE, whose solution is $Q_1^{(a)}(t) = 5(1-e^{-t})$. The expression approaches 5 as t increases, but is only valid for $t \in [0, 10]$. We have that $Q_1^{(a)}(10) = 5(1-e^{-10}) = 5.000$.

Stepping forward to the next interval, $[\tau, 2\tau] = [10, 20]$, (7) again reduces to an ODE, but one with a more complex forcing function:

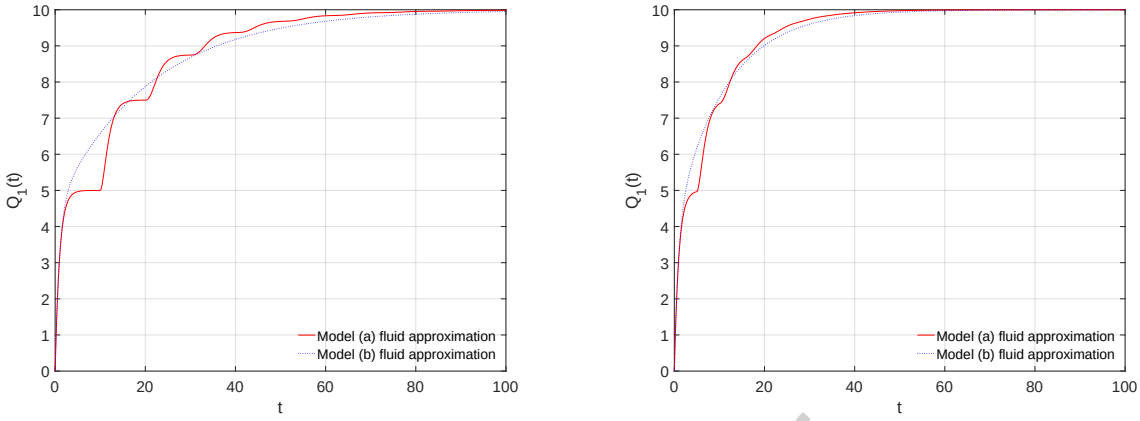
$$\begin{aligned}\frac{d}{dt}Q_1^{(a)}(t) &= \lambda - B^{(a)}(t)\mu(Q_1^{(a)}(t)) + B^{(a)}(t-\tau)\mu(Q_1^{(a)}(t-\tau))p(Q_1^{(a)}(t-\tau)) \\ &= \lambda - \min(Q_1^{(a)}(t), N)\mu + \min(Q_1^{(a)}(t-\tau), N)\mu p \\ &= 5 - Q_1^{(a)}(t) + 0.5 \times 5(1 - e^{-(t-10)}),\end{aligned}$$

as long as $Q_1^{(a)}(t) < N = 11$. The solution to this ODE is

$$Q_1^{(a)}(t) = 5 + 2.5\left(1 - e^{-(t-10)}(1 + t - 10)\right),$$

which approaches $5 + 2.5 = 7.5$ as t increases, is valid until $t = 20$, and we have $Q_1^{(a)}(20) = 7.499$. One can continue to obtain closed-form solutions for the DDE in this manner, step by step, but the forcing functions and the resulting solutions become increasingly cumbersome.

In Figure 4b, we show that when the length of the delay is reduced from $\tau = 10$ to 5, then the



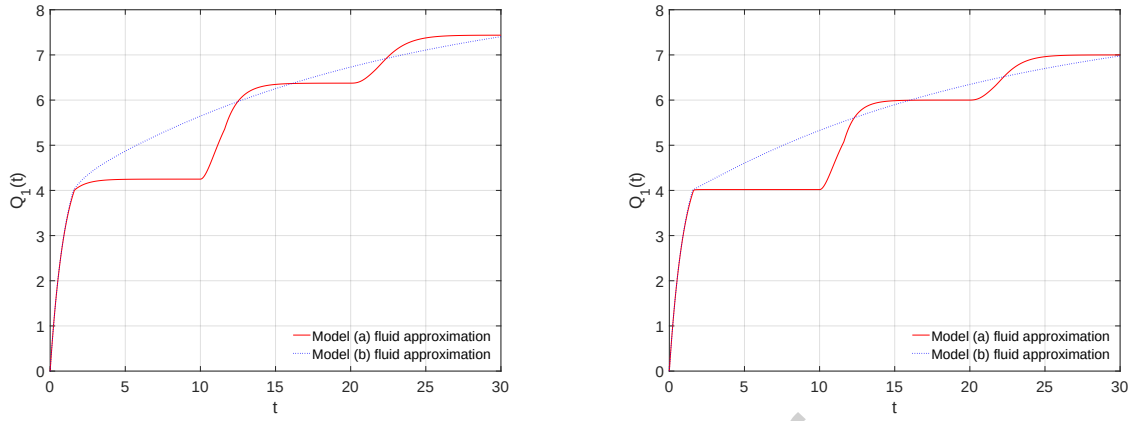
(a) $\tau = 10$ days (b) $\tau = 5$ days

Figure 4: Base case with no state dependence and different delays.
($\lambda = 5$, $N = 11$, $1/\mu = 1$ days, $p = 0.5$)

Model (a) fluid approximation becomes more similar to the Model (b) fluid approximation, and both approximations reach steady state sooner. The steady-state value is not impacted by the change in the length of the delay.

To see how state-dependence changes the DDE transient solution, we begin by changing from the constant-parameter base case to the two-value step functions in (1). We set $N^* = 4$, we keep the service rate and return probability for $Q_1 < N^*$ as before ($\mu_L = 1$ per day, $p_L = 0.5$), we keep the return probability for $Q_1 \geq N^*$ as before ($p_H = 0.5$), but we increase the service rate μ_H from 1 to $1/0.85 = 1.18$ per day. In Figure 5a, we see that $Q_1^{(a)}(t)$ reaches $N^* = 4$ at $t = 1.61$, at which point the slope changes discontinuously from $\lambda - \mu_L Q_1^{(a)}(1.61) = 5 - 1 \times 4 = 1$ to $\lambda - \mu_H Q_1^{(a)}(1.61) = 5 - (1/0.85) \times 4 = 0.29$ and the value that $Q_1^{(a)}(t)$ approaches in the interval $[0, 10]$ is reduced from $\lambda/\mu_L = 5$ to $\lambda/\mu_H = 4.25$. Furthermore, at $t = 1.61 + \tau = 11.61$, $Q_1^{(a)}(t - \tau)$ reaches $N^* = 4$, and the slope experiences another discontinuity—albeit one that is not clearly visible in the figure.

In Figure 5b, we have increased μ_H further, to $1/0.8$ per day, so that the slope after $Q_1^{(a)}(t)$ reaches $N^* = 4$ is $\lambda - \mu_H Q_1^{(a)}(1.61) = 5 - (1/0.8) \times 4 = 0$. If we increase μ_H past $1/0.8$ per day, then we see a situation where, in an interval after $t = 1.61$, the slope becomes positive if $Q_1^{(a)}(t)$ drops slightly below $N^* = 4$ and the slope becomes negative if $Q_1^{(a)}(t)$ increases slightly above $N^* = 4$. In other words, $Q_1^{(a)}(t)$ is attracted to the value $N^* = 4$, and remains there until the



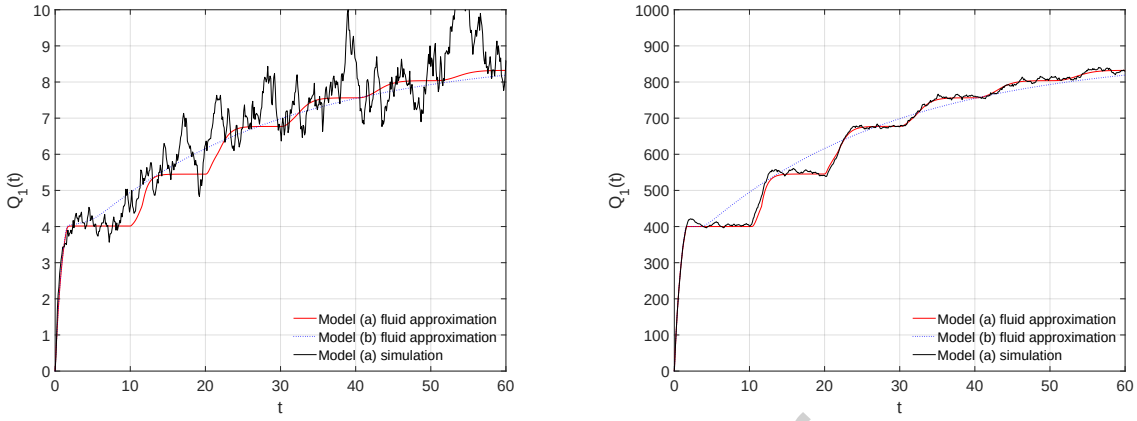
(a) $1/\mu_H = 0.85$ days (b) $1/\mu_H = 0.8$ days

Figure 5: Base case with state dependent service rate.
($\lambda = 5, N = 11, \tau = 10$ days, $1/\mu_L = 1$ days, $p_L = p_H = 0.5$, $N^* = 4$)

lagged value $Q_1^{(a)}(t-\tau)$ becomes large enough so that the $+\mu p Q_1^{(a)}(t-\tau)$ term in the expression for $dQ_1^{(a)}(t)/dt$ causes the slope to be positive regardless of whether $Q_1^{(a)}(t)$ is below or above N^*.

Next, we switch to the logistic approximation (2) and use the service rate and return probability functions shown in Figure 3a. In Figure 6a, we compare the Model (a) and (b) fluid approximations to the average of 30 simulated sample paths for Model (a). In order to investigate the impact of system size, we multiply λ, N, and N^* by a scaling factor η, keeping μ, p, τ, and K fixed. We observe that the average of the simulated Model (a) sample paths displays the same progression through stages of duration equal to τ as the Model (a) fluid approximation. We also observe that the Model (a) fluid approximation tends to underestimate the average of the Model (a) sample paths. This is consistent with results from Jiménez & Koole (2004), who prove that a fluid approximation for an $M/M/N$ system provides a lower bound on expected occupancy. If a similar result could be proved for Model (a), it would be of the form $E[\mathbb{Q}_1^{(a),\eta}(t)] \geq \eta Q_1^{(a)}(t)$.

Results in Mandelbaum et al. (2002) imply that if an $M/M/N$ system is scaled as we have described, then every sample path approaches the fluid approximation in the limit as $\eta \to \infty$. To investigate whether the same happens with Model (a) and its fluid approximation, we scale the system with $\eta = 100$. We see in Figure 6b that for this system, the Model (a) fluid approximation is much more accurate, mirroring the $M/M/N$ theoretical results from Mandelbaum et al. (2002). The Model (a) fluid approximation captures the progression through stages of the Model (a) sample



(a) $\eta = 1$ (b) $\eta = 100$

Figure 6: Base Case with different system sizes.
($\lambda = 5$, $N = 11$, $\tau = 10$ days, $1/\mu_L = 1$ days, $1/\mu_H = 0.7$ days, $p_L = 0.5$, $p_H = 0.6$, $N^* = 4$, $K_\mu = K_p = 20$)

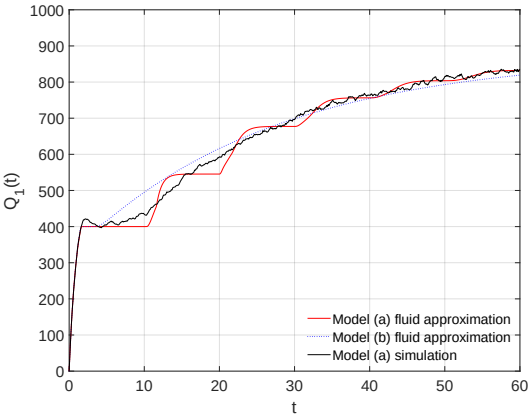
path average, for both $\eta = 1$ and $\eta = 100$, and for $\eta = 100$ the Model (a) fluid approximation is much more accurate than the Model (b) fluid approximation, if one assumes that Model (a) is correct.

The Model (a) fluid approximation makes the important assumption that the delay T is deterministic. Relaxing this assumption would add considerable complexity to the Model (a) fluid approximation, but we can easily relax this assumption in the Model (a) simulation. We see in Figure 7 that the Model (a) simulation sample path average is close to the Model (b) fluid approximation if T is exponentially distributed, and lies between the two fluid approximations when T is Erlang-8 distributed.

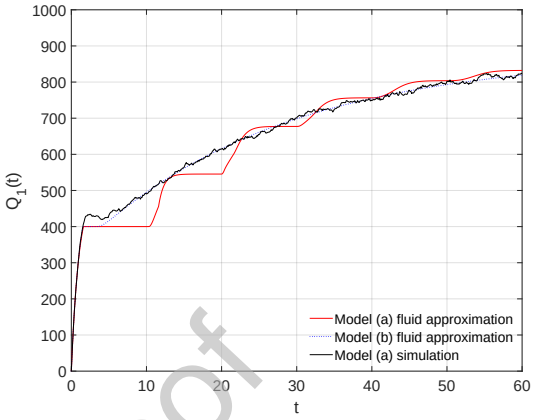
The stability limit for the Figure 4 base case system is $N\nu = 5.5$. In Figure 8, we illustrate the behavior of both fluid approximations for $\lambda = 6$, which is unstable, with $\tau = 10$ and 5 days. To understand the behavior of the fluid approximations for arrival rates above the stability limit, suppose that $Q_1^{(a)}(t) > N$ for $t > \bar{t} - \tau$ and that $\mu(x)$, $p(x)$, and $\nu(x)$ stabilize when x reaches N. This implies the following for the Model (a) fluid approximation:

$$\frac{d}{dt}Q_1^{(a)}(t) = \lambda - N\mu(N)(1 - p(N)) = \lambda - N\nu(N),$$

that is, the Station 1 occupancy increases at a rate of $\lambda - N\nu(N)$ per unit time, for $t > \hat{t}$. Similarly, suppose that $Q_1^{(b)}(t) > N$ for $t > \bar{t}$. Then we obtain the following for the Model (b) fluid

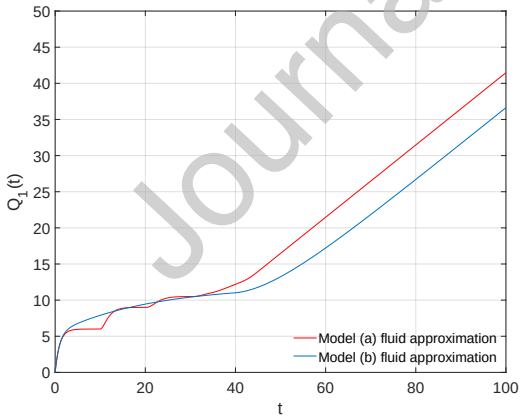


(a) T Erlang-8 distributed

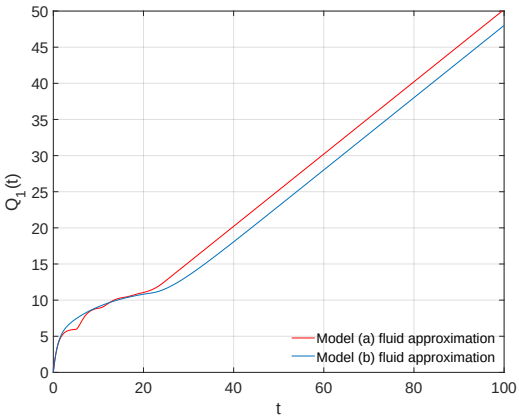


(b) T exponentially distributed

Figure 7: Base Case with large system size ($\eta = 100$), varying distribution for T.
($\lambda = 5$, $N = 11$, $\tau = 10$ days, $1/\mu_L = 1$ days, $1/\mu_H = 0.7$ days, $p_L = 0.5$, $p_H = 0.6$, $N^* = 4$, $K_\mu = K_p = 20$)



(a) $\tau = 10$ days



(b) $\tau = 5$ days

Figure 8: Base case with an unstable arrival rate.
($\lambda = 6, N = 11, \tau = 10$ days, $1/\mu = 1$ days, $p = 0.5$)

Journal Pre-proof

Journal Pre-proof

approximation:

$$\frac{d}{dt}(Q_1^{(b)}(t) + Q_2^{(b)}(t)) = \lambda - N\mu(N) + N\mu(N)p(N) = \lambda - N\nu(N),$$

that is, the combined occupancy of Stations 1 and 2 increases at the same rate as the Station 1 occupancy in the Model (a) fluid approximation. Furthermore, once $Q_1^{(b)}$ reaches N, the right side of the ODE (9) for $Q_2^{(b)}$ no longer involves $Q_1^{(b)}$ and can be solved explicitly. It follows from the explicit solution for (9) that

$$Q_2^{(b)}(t) \to \tau N\nu(N),$$

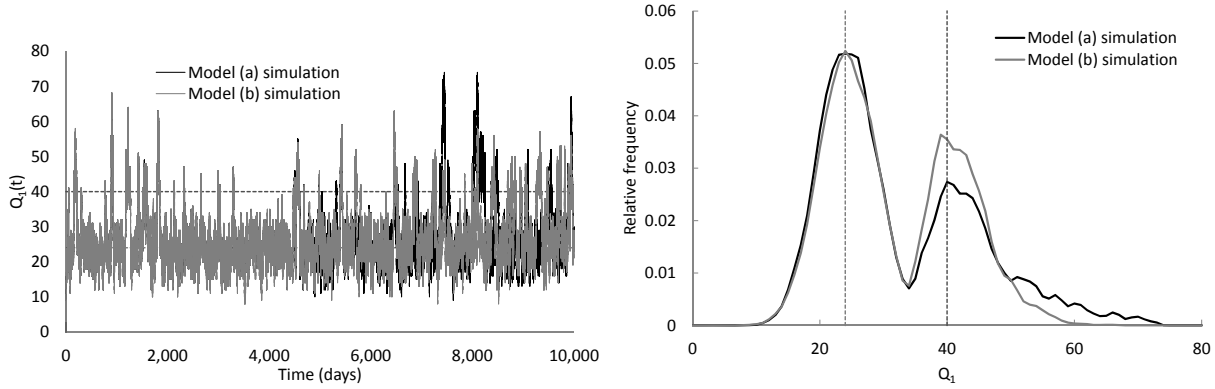
that is, the Station 2 occupancy approaches a constant. Taken together, these calculations imply that in the limit, the Station 1 occupancy grows at the same rate in both fluid approximations. The numerical solutions in Figure 8 confirm this, but also show that the Model (b) fluid approximation lags behind the Model (a) fluid approximation, by an amount that increases with τ.

8. Simulation of Bistability

In this section, we simulate a special case of the stochastic models in Figure 1. Chan et al. (2014) show that the fluid system, in certain settings, alternates between two equilibrium points—that is, the system exhibits bistability. We investigate a similar setting, obtained using the service rate and return probability functions shown in Figure 3b. As mentioned in Section 6, the corresponding equilibrium points are $x = 24, 35.122$, and 40.

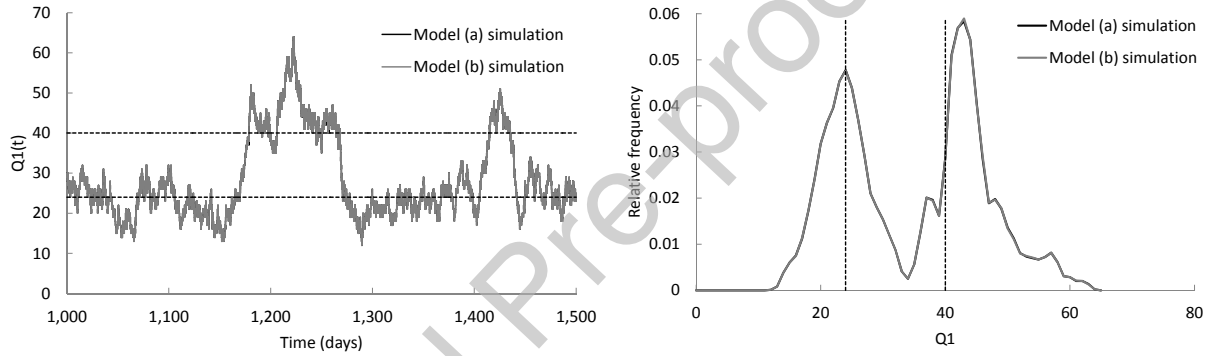
In Figure 9a, we compare simulated sample paths for Models (a) and (b) for the Station 1 occupancy. Both systems start empty, and as time passes, we observe that the sample paths alternate between the two locally stable equilibrium points $x = 24$ and 40 (the equilibrium point $x = 35.12$ is unstable). Figure 9b shows the Station 1 occupancy distribution corresponding to the two sample paths in Figure 9a. The distributions are similar but not identical—specifically, the Model (a) distribution has a lower mode and longer tail corresponding to the $x = 40$ equilibrium.

We used the same random number streams for corresponding elements in the two models, and we use an exponential distribution for the delay, T, in both models. As a result, the sample paths are identical for a while, but eventually (after about 4,000 days) they diverge. Figure 10 shows



(a) Sample paths (b) Frequency distributions

Figure 9: Simulated Station 1 occupancy for Models (a) and (b) for the bistable parameter settings listed in Figure 3b.



(a) Sample paths (b) Frequency distributions

Figure 10: Simulated Station 1 occupancy for Models (a) and (b) for the bistable parameter settings listed in Figure 3b, for the time period from 1,000 to 1,500 days.

more clearly that the sample paths coincide initially, by focusing on the time period from 1,000 to 1,500 days.

9. Conclusion

In this paper, we compared the transient behavior and the equilibrium points for fluid approximations of two systems that have state dependent service and return probabilities. Our proposed Model (a) is more realistic in certain settings than Model (b), which has been studied earlier in the literature. In Model (a), it takes time to decide whether a customer needs another stage of service. Different methodologies are used to analyze the two fluid approximations but the equilibrium results are similar. However, the transient behavior for Model (a) involves a progression

Journal Pre-proof

Journal Pre-proof

through stages. The nature of this transient behavior could be important in certain settings. For example, if an ICU behaves according to Model (a), then the rate at which patients return to the ICU from a step-down unit would increase in stages rather than continuously, with the duration of each stage corresponding to the length of stay in the step-down unit.

Our Model (a) fluid approximation assumes a deterministic delay after service. Simulation experiments with stochastic delay suggest that as the distribution of the delay after service becomes more similar to an exponential distribution, Station 1 occupancy in Model (a) becomes more similar to that in Model (b).

To summarize, our work suggests the following cautions with regard to using Model (b) in settings where Model (a) is closer to reality: (1) Model (b) underestimates Station 2 occupancy (customers experiencing delay) and should therefore not be used to choose capacity for Station 2, (2) equilibrium values of Station 1 occupancy are robust to the timing of return routing, (3) transient values of Station 1 occupancy are sensitive to the timing of return routing and the shape of the delay distribution.

Future work should investigate reformulation of the Model (a) fluid approximation to have a delay distribution that is either discrete (which is likely to be more tractable) or continuous. Another area that would benefit from further study is the effective system capacity in a system with state-dependent service rates and return probabilities. KC & Terwiesch (2012) found that "speeding up" might decrease an ICU's effective capacity. Future work could aim to determine a service speed that maximizes the effective system capacity. Such work could also be relevant in a manufacturing setting, in which one seeks the speed for a machine that maximizes capacity, as discussed in Owen & Blumenfeld (2008). Finally, the techniques used to prove Theorems 6.3 and 6.4 rely on the right sides of the differential equations being continuously differentiable. Methods from non-smooth analysis (Cortes, 2008) could perhaps be used to relax the condition $x \neq N$ for an equilibrium point x.

Acknowledgements

The authors thank three anonymous referees for their constructive comments, which helped to improve the paper. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [Discovery Grants 203534 and 06344; Undergraduate Student

Journal Pre-proof

Research Award 441121] and the Hill and Levene Research Stewardship Award.

References

Anderson, D., Golden, B., Jank, W., & Wasil, E. (2012). The impact of hospital utilization on patient readmission rate. *Health Care Management Science*, *15*, 29–36. doi:10.1007/s10729-011-9178-3.

Apte, U. M., Beath, C. M., & Goh, C.-H. (1999). An analysis of the production line versus the case manager approach to information intensive services. *Decision Sciences*, *30*, 1105–1129. doi:10.1111/j.1540-5915.1999.tb00920.x.

Barjesteh, N., & Abouee-Mehrizi, H. (2018). Multi-class multi-server state-dependent queueing systems with returns. Working paper.

Beretta, E., Kolmanovskii, V., & Shaikhet, L. (1998). Stability of epidemic model with time delays influenced by stochastic perturbations. *Mathematics and Computers in Simulation*, *45*, 269–277. doi:10.1016/S0378-4754(97)00106-7.

Breda, D. (2012). On characteristic roots and stability charts of delay differential equations. *International Journal of Robust and Nonlinear Control*, *22*, 892–917. doi:10.1002/rnc.1734.

Breda, D., Maset, S., & Vermiglio, R. (2014). *Stability of Linear Delay Differential Equations: A Numerical Approach with MATLAB*. Springer. doi:10.1007/978-1-4939-2107-2.

Campello, F., Ingolfsson, A., & Shumsky, R. A. (2017). Queueing models of case managers. *Management Science*, *63*, 882–900. doi:10.1287/mnsc.2015.2368.

Chan, C. W., Farias, V. F., Bambos, N., & Escobar, G. J. (2012). Optimizing intensive care unit discharge decisions with patient readmissions. *Operations Research*, *60*, 1323–1341. doi:10.1287/opre.1120.1105.

Chan, C. W., Yom-Tov, G., & Escobar, G. (2014). When to use speedup: An examination of service systems with returns. *Operations Research*, *62*, 462–482. doi:10.1287/opre.2014.1258.

Chrusch, C. A., Olafson, K. P., McMillan, P. M., Roberts, D. E., & Gray, P. R. (2009). High occupancy increases the risk of early death or readmission after transfer from intensive care. *Critical Care Medicine*, *37*, 2753–2758. doi:10.1097/CCM.0b013e3181a57b0c.

Cortes, J. (2008). Discontinuous dynamical systems. *IEEE Control Systems Magazine*, *28*, 36–73. doi:10.1109/MCS.2008.919306.

Cruz, F., Smith, J. M., & Medeiros, R. (2005). An $M/G/C/C$ state-dependent network simulation model. *Computers & Operations Research*, *32*, 919–941. doi:10.1016/j.cor.2003.09.006.

Delasay, M., Ingolfsson, A., Kolfal, B., & Schultz, K. (2018). Load effect on service times. *European Journal of Operational Research*, . doi:https://doi.org/10.1016/j.ejor.2018.12.028.

Dong, J., Feldman, P., & Yom-Tov, G. B. (2015). Service systems with slowdowns: Potential failures and proposed solutions. *Operations Research*, *63*, 305–324. doi:10.1287/opre.2015.1346.

Faria, T. (2001). Stability and bifurcation for a delayed predator–prey model and the effect of diffusion. *Journal of Mathematical Analysis and Applications*, *254*, 433–463. doi:10.1006/jmaa.2000.7182.

Furman, E., Diamant, E., & Kristal, M. (2019). Customer acquisition and retention: A fluid approach for staffing. Working paper.

Journal Pre-proof

Hayes, N. D. (1950). Roots of the transcendental equation associated with a certain difference-differential equation. *Journal of the London Mathematical Society*, *s1-25*, 226–232. doi:10.1112/jlms/s1-25.3.226.

Hu, W., Chan, C. W., Zubizarreta, J. R., & Escobar, G. J. (2018). An examination of early transfers to the ICU based on a physiologic risk score. *Manufacturing & Service Operations Management*, *20*, 531–549. doi:10.1287/msom.2017.0658.

Huang, J., Carmeli, B., & Mandelbaum, A. (2015). Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research*, *63*, 892–908. doi:10.1287/opre.2015.1389.

Jiménez, T., & Koole, G. (2004). Scaling and comparison of fluid limits of queues applied to call centers with time-varying parameters. *OR Spectrum*, *26*, 413–422. doi:10.1007/s00291-004-0162-x.

Johari, R., & Tan, D. K. H. (2001). End-to-end congestion control for the Internet: Delays and stability. *IEEE/ACM Transactions on Networking*, *9*, 818–832. doi:10.1109/90.974534.

KC, D. S., & Terwiesch, C. (2012). An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management*, *14*, 50–65. doi:10.1287/msom.1110.0341.

Khalid, R. M., Nawawi, M. K. M., Kawsar, L. A., Ghani, N. A., Kamil, A. A., & Mustafa, A. (2013). A discrete event simulation model for evaluating the performances of an $M/G/C/C$ state dependent queuing system. *PloS ONE*, *8*, e58402. doi:10.1371/journal.pone.0058402.

Luo, J., & Zhang, J. (2013). Staffing and control of instant messaging contact centers. *Operations Research*, *61*, 328–343. doi:10.1287/opre.1120.1151.

Makris, N., Dulhunty, J. M., Paratz, J. D., Bandeshe, H., & Gowardman, D. J. R. (2010). Unplanned early readmission to the intensive care unit: A case-control study of patient, intensive care and ward-related factors. *Anaesthesia and Intensive Care*, *38*, 723–731. doi:10.1177/0310057X1003800338.

Mandelbaum, A., Massey, W. A., & Reiman, M. I. (1998). Strong approximations for Markovian service networks. *Queueing Systems*, *30*, 149–201. doi:10.1023/A:1019112920622.

Mandelbaum, A., Massey, W. A., Reiman, M. I., Stolyar, A., & Rider, B. (2002). Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems*, *21*, 149–171. doi:10.1023/A:1020921829517.

Master, N., Reiman, M. I., Wang, C., & Wein, L. M. (2018). A continuous-class queueing model with proportional hazards-based routing. *Available at SSRN 3390476*, .

Nakamura, G. (1971). A feedback queueing model for an interactive computer system. In *Proceedings of the November 16-18, 1971, Fall Joint Computer Conference* (pp. 57–64). ACM. doi:10.1145/1479064.1479075.

Owen, J. H., & Blumenfeld, D. E. (2008). Effects of operating speed on production quality and throughput. *International Journal of Production Research*, *46*, 7039–7056. doi:10.1080/00207540701227833.

Pender, J., Rand, R. H., & Wesson, E. (2017). Queues with choice via delay differential equations. *International Journal of Bifurcation and Chaos*, *27*, 1730016. doi:10.1142/S0218127417300166.

Pender, J., Rand, R. H., & Wesson, E. (2018). An analysis of queues with delayed information and time-varying arrival rates. *Nonlinear Dynamics*, *91*, 2411–2427. doi:10.1007/s11071-017-4021-0.

Saghafian, S., Hopp, W. J., Van Oyen, M. P., Desmond, J. S., & Kronick, S. L. (2014). Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management*,

Journal Pre-proof

16, 329–345. doi:10.1287/msom.2014.0487.

Shampine, L. F., & Thompson, S. (2001). Solving DDEs in Matlab. *Applied Numerical Mathematics*, *37*, 441–458. doi:10.1016/S0168-9274(00)00055-6.

Shampine, L. F., & Thompson, S. (2009). Numerical solution of delay differential equations. In B. Balachandran, T. Kalmár-Nagy, & D. E. Gilsinn (Eds.), *Delay Differential Equations: Recent Advances and New Directions* chapter 9. (pp. 245–269). Boston, MA: Springer US. doi:10.1007/978-0-387-85595-0_9.

Shi, P., Helm, J., Deglise-Hawkinson, J., & Pan, J. (2019). Timing it right: Balancing inpatient congestion versus readmission risk at discharge. *Available at SSRN 3202975*, .

Sideris, T. C. (2013). *Ordinary differential equations and dynamical systems* volume 2 of *Atlantis Studies in Differential Equations*. Paris: Atlantis Press.

Smith, H. (2011). *An Introduction to Delay Differential Equations with Applications to the Life Sciences* volume 57 of *Texts in Applied Mathematics*. New York: Springer-Verlag.

Tezcan, T., & Zhang, J. (2014). Routing and staffing in customer service chat systems with impatient customers. *Operations Research*, *62*, 943–956. doi:10.1287/opre.2014.1284.

Town, J. A., Churpek, M. M., Yuen, T. C., Huber, M. T., Kress, J. P., & Edelson, D. P. (2014). Relationship between ICU bed availability, ICU readmission, and cardiac arrest on the general wards. *Critical Care Medicine*, *42*, 2037–2041. doi:10.1097/CCM.0000000000000401.

Utzolino, S., Kaffarnik, M., Keck, T., Berlet, M., & Hopt, U. T. (2010). Unplanned discharges from a surgical intensive care unit: Readmissions and mortality. *Journal of Critical Care*, *25*, 375–381. doi:https://doi.org/10.1016/j.jcrc.2009.09.009.

de Véricourt, F., & Jennings, O. B. (2008). Dimensioning large-scale membership services. *Operations Research*, *56*, 173–187. doi:10.1287/opre.1070.0464.

de Véricourt, F., & Jennings, O. B. (2011). Nurse staffing in medical units: A queueing perspective. *Operations Research*, *59*, 1320–1331. doi:10.1287/opre.1110.0968.

de Véricourt, F., & Zhou, Y.-P. (2005). Managing response time in a call-routing problem with service failure. *Operations Research*, *53*, 968–981. doi:10.1287/opre.1050.0230.

Whitt, W. (2002). *Stochastic-process limits: An introduction to stochastic-process limits and their application to queues*. Springer Science & Business Media.

Yankovic, N., & Green, L. V. (2011). Identifying good nursing levels: A queuing approach. *Operations Research*, *59*, 942–955. doi:10.1287/opre.1110.0943.

Yom-Tov, G. B., & Mandelbaum, A. (2014). Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, *16*, 283–299. doi:10.1287/msom.2013.0474.

Zhan, D., & Ward, A. R. (2013). Threshold routing to trade off waiting and call resolution in call centers. *Manufacturing & Service Operations Management*, *16*, 220–237. doi:10.1287/msom.2013.0463.

Journal Pre-proof

Journal Pre-proof

Appendix A. Arena Simulation Models

In this appendix, we document Arena discrete event simulation models of the queueing models defined in Section 3. We list the model components and outline the model logic. See Khalid et al. (2013) and Cruz et al. (2005) for other approaches to simulating state-dependent service times (these authors do not simulate returns). We focus on documenting the Model (a) simulation, but mention aspects where the Model (b) simulation differs.

Variables:

- ArrivalRate: Arrival rate, λ, of new (as opposed to returning) customers.
- ServiceRate: Array variable containing service rates $\mu(b)$ for $b = \mathbb{B}(t) = \min(N, \mathbb{Q}_1(t))$, where $b = 1, \ldots, N$.
- Probability: Array variable containing return probabilities $p(b)$ for $b = \mathbb{B}(t)$, where $b = 1, \ldots, N$, where t is the time at which a customer leaves service.
- BusyServers: The number of busy servers, $\mathbb{B}(t)$.
- TNOW: Current simulation time—a system variable

Attributes: (local variables, associated with Customer entities)

- ServiceTime: Remaining customer service time. Initalized when customer arrives and updated whenever BusyServers changes.
- ExitTime: The simulation time at which the current service is scheduled to end.
- CurrBusyServers: The value of the BusyServers variable at the time when the Customer entity entered the system, or the last time the BusyServers variable changed value, whichever occurred last.
- RetProb: The probability that the customer will return to service after the current service.

Model Logic: Figure A.11 shows the Model (a) simulation flowchart. The flowchart has two loops: An inner loop that models state changes during a single service, and an outer loop that models customer returns. Next, we zoom in on parts of the flowchart and explain the model logic.

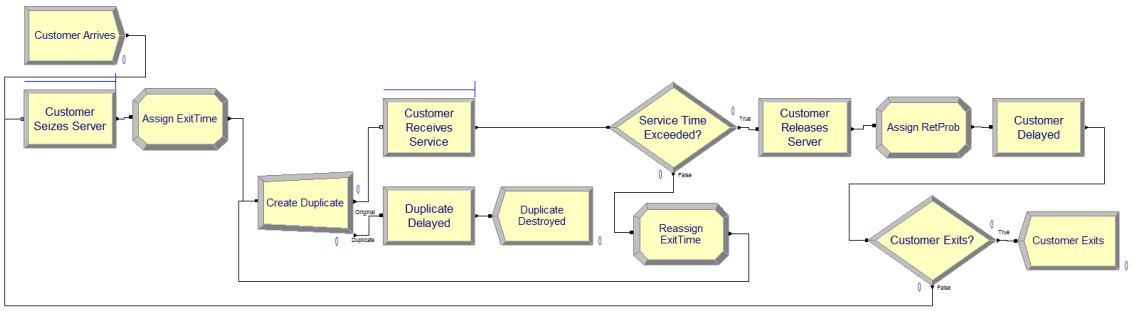


Figure A.11: Model flowchart

New customers arrive (Figure A.12) through a create module Customer Arrives with times between arrivals that are exponentially distributed with mean $1/\lambda$ (including the time from the start of the simulation until the arrival of the first customer). After a customer has arrived, he enters the Customer Seizes Server queue, a first-in-first-out queue, and seizes one out of a total of N servers when one becomes available. Then the customer enters an assign module Assign

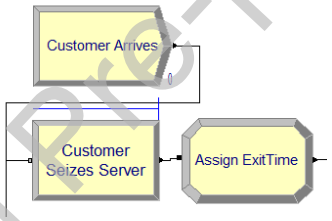


Figure A.12: Customer arrives and is assigned a server and initial attributes

ExitTime. This module updates the BusyServers system variable and sets the CurrBusyServers customer attribute to BusyServers. The ServiceTime customer attribute is set to a sample from an exponential distribution with mean $1/\mu(b) = 1/\text{ServiceRate}(\text{BusyServers})$. (The initial ServiceTime value could instead be drawn by a non-exponential distribution with state-dependent parameters.) The customer's ExitTime is computed as TNOW + ServiceTime.

After assignment, the customer is duplicated in Create Duplicate (Figure A.13). The original customer is held in a hold module Customer Receives Service until one of two conditions occur: (1) TNOW > ExitTime or (2) a customer's CurrBusyServers attribute no longer matches the system's BusyServers variable. Because the Customer Receives Service hold module only checks these conditions when an event occurs, a duplicate entity is created and delayed in Duplicate Delayed for the duration of ServiceTime so that an event occurs when the customer's ExitTime is reached. After

Journal Pre-proof

the duplicate is delayed, it is destroyed in Duplicate Destroyed. The customers in the Customer Receives Service hold module are sorted and processed in order of their ExitTime attribute.

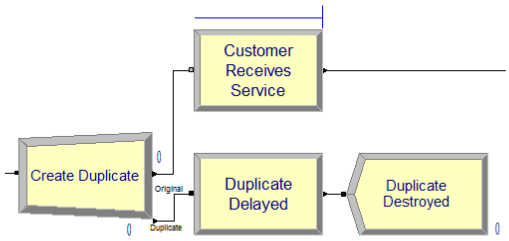


Figure A.13: A duplicate is created and customer receives service

Once one of the two conditions that are checked in the hold module is violated, the customer moves to a decide module Service Time Exceeded? (Figure A.14) that evaluates whether the customer's ExitTime has been reached or the customer's CurrBusyServers attribute no longer agrees with the BusyServers variable. If a customer's CurrBusyServers attribute no longer agrees

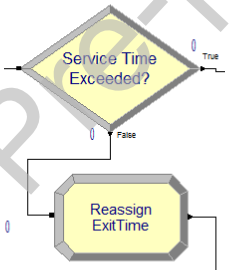


Figure A.14: Customer attributes are updated in response to a change in BusyServers

with the system's BusyServers variable, then the customer moves to an assign module Reassign ExitTime, where the customer's ServiceTime and ExitTime attributes are updated. There are two possible methods for adjusting the customer's remaining service time. One is to generate a new exponentially-distributed remaining service time with mean $1/\mu(b)$. The other is to scale the remaining service time by the ratio of two service rates—the old rate $\mu(b')$ (where b' is the previous value of BusyServers) and the new rate $\mu(b)$. We use this latter method, which can be interpreted to mean that a customer enters with a fixed amount of work that needs to be done, but the speed at which that work is completed changes whenever the number of busy servers changes. If R' and R are the customer's previous and new remaining service times, then $R = \frac{\mu(b')}{\mu(b)} \times R'$. Using the

Journal Pre-proof

Arena model variables and attributes, this can be written as

$$\mathsf{ServiceTime} = \frac{\mathsf{ServiceRate(CurrBusyServers)}}{\mathsf{ServiceRate(BusyServers)}} \times (\mathsf{ExitTime} - \mathsf{TNOW}). \tag{A.1}$$

After setting the customer's ServiceTime in this way, we update the customer's ExitTime to TNOW + ServiceTime and we update the CurrBusyServers attribute to match the current BusyServers variable. Once this is done, customers re-enter Create Duplicate and then wait in Customer Receives Service until one of the above conditions is again satisfied. Note that this looping does not represent customers completing and then returning for a new service. This inner loop is simply a device used in the simulation model to simulate state-dependent service times.

If instead a customer's ExitTime is reached, then the customer releases the server at Customer Releases Server and stops receiving service (Figure A.15). The customer does not exit the system—rather, the customer enters an assign block Assign RetProb which assigns the customer a reentry probability RetProb $= p(b) =$Probability(BusyServers). The BusyServers variable is updated. Next, the customer is delayed in Customer Delayed for T, whose distribution can be specified, with $\mathbb{E}[T]$ fixed to be τ. After the delay, the customer enter a decide module Customer Exits? that determines whether he returns for an additional service, which occurs with probability RetProb. Otherwise, the customer exits the system via Customer Exits. In the Model (b) simulation, the Customer Exits?

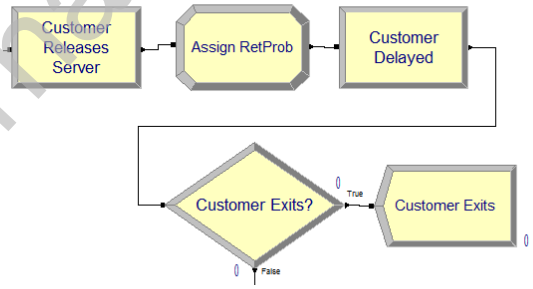


Figure A.15: Customer reenters or exits the system

decide module comes before the Customer Delayed module.

We used four separate random number streams (1) for new customer arrivals, (2) for service times, (3) for determining whether customers would return or not, and (4) for the delays, if modeled as random. In our models, only one random number is used to generate each service time, even if the remaining service time is re-scaled one or more times because of state changes.

Journal Pre-proof

Appendix B. Proof for Section 5

Proof of Theorem 5.1. Proof of (a): We use results from Chapter 3 of Smith (2011). Using the notation in that chapter, we set $x = Q_1^{(a)}$ and $r = \tau$ and write (7) as

$$\begin{aligned} x'(t) &= \lambda - \min(x(t), N)\mu(x(t)) + \min(x(t-r), N)\mu(x(t-r))p(x(t-r)) \\ &= f(t, x_t), t \geq 0 \\ x_0 &= \rho, \end{aligned} \tag{B.1}$$

where $x_t = \{x(t+\theta) : -r \leq \theta \leq 0\}$ and

$$f(t, x_t) = \lambda - \min(x(t), N)\mu(x(t)) + \min(x(t-r), N)\mu(x(t-r))p(x(t-r))$$

Let $||\phi|| = \sup\{|\phi(\theta) : -r \leq \theta \leq 0\}$ be a norm for a continuous function ϕ on an interval of length r. Let ρ, ϕ, and ψ be three such functions and assume that $||\rho||, ||\phi||, ||\psi|| \leq M$. Further, define $\mu_{\max}$ to be an upper bound for $\mu(x)$. Then we have:

$$\begin{aligned} &|f(t, \phi) - f(t, \psi)| \\ &\quad = |\lambda - \min(\phi(t), N)\mu(\phi(t)) + \min(\phi(t-r), N)\mu(\phi(t-r))p(\phi(t-r)) \\ &\quad - (\lambda - \min(\psi(t), N)\mu(\psi(t)) + \min(\psi(t-r), N)\mu(\psi(t-r))p(\psi(t-r)))| \\ &\quad \leq |\min(\psi(t), N)\mu(\psi(t)) - \min(\phi(t), N)\mu(\phi(t))| \\ &\quad + |\min(\phi(t-r), N)\mu(\phi(t-r))p(\phi(t-r)) - \min(\psi(t-r), N)\mu(\psi(t-r))p(\psi(t-r)))| \\ &\quad \leq |\psi(t) - \phi(t)|\mu_{\max} + |\phi(t-r) - \psi(t-r)|\mu_{\max} \\ &\quad \leq 2\mu_{\max}||\psi - \phi|| \\ &\quad = K||\psi - \phi||, \end{aligned}$$

where $K = 2\mu_{\max}$ is constant. This demonstrates that $f(t, x_t)$ is continuous and satisfies a Lipschitz condition with Lipschitz constant K. It follows, from Theorem 3.7 and Remark 3.8 in Smith (2011), that (7) has a unique solution for $t \geq 0$.

Proof of (b): Let $x = (Q_1^{(b)}, Q_2^{(b)})$. Then (8)-(9) can be written as $x'(t) = f(t, x(t))^T$, where

$$f(t, x(t)) = \begin{bmatrix} \lambda \\ 0 \end{bmatrix} + \begin{bmatrix} -\mu \min(N, x_1(t)) + \delta x_2(t) \\ \mu p \min(N, x_1(t)) - \delta x_2(t) \end{bmatrix}$$

We verify that $f(t, x(t))$ satisfies the assumption in Theorem 3.9 in Sideris (2013), that there exist continuous and nonnegative functions $c_1(t)$ and $c_2(t)$, such that $\|f(t, x(t))\| \leq c_1(t) \|x(t)\| + c_2(t)$ holds:

$$\begin{aligned} \|f(t, x(t))\| &\leq \left\| \begin{bmatrix} \lambda \\ 0 \end{bmatrix} \right\| + \left\| \begin{bmatrix} -\mu \min(N, x_1(t)) + \delta x_2(t) \\ \mu p \min(N, x_1(t)) - \delta x_2(t) \end{bmatrix} \right\| \\ &\leq \lambda + \left\| \begin{bmatrix} \mu x_1(t) + \delta x_2(t) \\ \mu p x_1(t) + \delta x_2(t) \end{bmatrix} \right\| \\ &= \lambda + \left\| \begin{bmatrix} \mu & \delta \\ \mu p & \delta \end{bmatrix} x(t) \right\| \\ &= \lambda + (\mu + \delta) \|x(t)\|, \end{aligned}$$

that is, we can set $c_1(t) = \mu + \delta$ and $c_2(t) = \lambda$. □

Appendix C. Proofs for Section 6

Appendix C.1. Model (a) Fluid Approximation Equilibria and Stability

Proof of Theorem 6.1. Assume that $Q_1^{(a)}(t)$ reaches a limit $\bar{Q}_1^{(a)}$ as $t \to \infty$. Then $\bar{Q}_1^{(a)}$ must satisfy $0 = \lambda - \min(N, \bar{Q}_1^{(a)})\nu$, which is obtained by setting the left hand side of (7) to zero and assuming that μ and p are constant. This equation implies:

$$\min(N, \bar{Q}_1^{(a)}) = \frac{\lambda}{\nu} \tag{C.1}$$

If $\lambda > N\nu$, then the right side of (C.1) is larger than N, while the left side must be less than or equal to N, which is a contradiction, and therefore (7) has no equilibrium point. If $\lambda \leq N\nu$, then the right side of (C.1) is less than or equal to N, which implies that the left side equals $\bar{Q}_1^{(a)}$, and

Journal Pre-proof

$\bar{Q}_1^{(a)} = \lambda/\nu$ is an equilibrium point. To analyse stability of $\bar{Q}_1^{(a)} = \lambda/\nu$, we express $Q_1^{(a)}(t)$ as

$$Q_1^{(a)}(t) = \bar{Q}_1^{(a)} + U(t), \tag{C.2}$$

where $U(t)$ is a perturbation. The assumption $\lambda < N\nu$ implies that $\bar{Q}_1^{(a)} < N$. Substituting (C.2) in (7), we obtain the following, which holds for $U(t)$ small enough that $U(t) \leq N - \bar{Q}_1^{(a)}$:

$$\frac{d}{dt}U(t) = -\mu U(t) + p\mu U(t-\tau) \tag{C.3}$$

This is a "Hayes equation," (Hayes, 1950) that is, a linear DDE with constant coefficients and one delay, τ. The stability of Hayes equations in the general form $dU(t)/dt = aU(t) + bU(t-\tau)$, where a and b are real, has been studied extensively (Breda et al., 2014). The stability is studied in terms of the characteristic equation, which is obtained by substituting $U(t) = e^{rt}$, where r is a characteristic root, resulting in the following for (C.3):

$$r = -\mu + p\mu e^{-r\tau} \tag{C.4}$$

We make use of Proposition 6 and Table I in Breda (2012), which imply that our Equation (C.3), with $a = -\mu < 0$, $b = p\mu > 0$, and $a + b = \mu(p-1) = -\nu < 0$, has a single real root that is negative and no complex roots that have positive real part. Therefore, $U(t) \to 0$ as $t \to \infty$ and the equilibrium point $\bar{Q}_1^{(a)}$ is locally stable. We only claim local (rather than global) stability, because our proof is based on the assumption that $U(t) \leq N - \bar{Q}_1$. □

Proof of Theorem 6.3. Assume that $Q_1^{(a)}(t) \to x$ as $t \to \infty$. Then x must satisfy:

$$f(x) \equiv \min(x, N)\nu(x) = \lambda, \tag{C.5}$$

which is obtained by setting $\frac{d}{dt}Q_1^{(a)}(t) = 0$. The values of $f(x)$ for $x = 0$ and $x = z = \max(N, \tilde{x})$ (where $\tilde{x}$ is from Condition (5)) are $f(0) = 0 < \lambda$ and $f(z) = N\nu(z) > \lambda$. The fact that $f(x) - \lambda$ switches signs from $x = 0$ to $x = z$ means that (C.5) has at least one root in $(0, z)$, which proves that (7) has at least one equilibrium point. If, furthermore, $\nu(x)$ is increasing (Condition (6)), then $f(x)$ is increasing, which implies that (C.5) has exactly one root in $(0, z)$, which proves that (7) has exactly one equilibrium point.

Journal Pre-proof

To analyse the stability of an equilibrium point x, we consider two cases: $x < N$ and $x > N$.

Case 1: $x < N$. We express $Q_1^{(a)}(t)$ as $x + U(t)$, where $U(t)$ is a perturbation. Assuming that $U(t) < N - x$, so that $Q_1^{(a)}(t) \leq N$, we can express (7) as

$$\frac{d}{dt}Q_1^{(a)}(t) = F(Q_1^{(a)}(t), Q_1^{(a)}(t-\tau)), \tag{C.6}$$

where the function F and its partial derivatives are:

$$F(u,v) = \lambda - u\mu(u) + v\mu(v)p(v), \tag{C.7}$$

$$f_u(u,v) = -\mu(u) - u\mu'(u), \tag{C.8}$$

$$f_v(u,v) = \mu(v)p(v) + v(\mu'(v)p(v) + \mu(v)p'(v)). \tag{C.9}$$

The linearized system for $U(t)$ (e.g, see Section 4.6 in Smith, 2011) is

$$\frac{d}{dt}U(t) = aU(t) + bU(t-\tau), \tag{C.10}$$

$$a = -\left(\mu(x) + x\mu'(x)\right), \tag{C.11}$$

$$b = \mu(x)p(x) + x(\mu'(x)p(x) + \mu(x)p'(x)). \tag{C.12}$$

For stability, the Hayes Equation sufficient condition that we used earlier is $a < 0$, $b > 0$, and $a + b < 0$. Our assumptions that μ and p and their derivatives are positive imply that $a < 0$ and $b > 0$. As for $a + b$, we have

$$a + b = -\left(\mu(x) + x\mu'(x)\right) + \mu(x)p(x) + x(\mu'(x)p(x) + \mu(x)p'(x)) \tag{C.13}$$

$$= -\mu(x)(1 - p(x)) - x\left(\mu(x)(1 - p(x))\right)' \tag{C.14}$$

$$= -\nu(x) - x\nu'(x) < 0. \tag{C.15}$$

Therefore, the linearized system (C.10) is stable, which implies that x is locally stable.

Case 2: $x > N$. Expressing $Q_1(t)$ as $Q_1(t) = x + U(t)$ and performing the same type of analysis as in Case 1 results in a Hayes equation with $a = -N\mu'(x) < 0$, $b = N(\mu'(x)p(x) + \mu(x)p'(x)) > 0$, and $a + b = -N\nu'(x) < 0$, which implies that x is locally stable. □

Journal Pre-proof

Appendix C.2. Model (b) Fluid Approximation Equilibria and Stability

Proof of Theorem 6.2. Assume $(Q_1^{(b)}(t), Q_2^{(b)}(t))$ reaches a limit $(\bar{Q}_1^{(b)}, \bar{Q}_2^{(b)})$ as $t \to \infty$. Then $(\bar{Q}_1^{(b)}, \bar{Q}_2^{(b)})$ must satisfy the following set of equations, obtained by setting the left sides of (8)-(9) to zero and assuming constant μ and p:

$$0 = \lambda - \mu \min(N, \bar{Q}_1^{(b)}) + \bar{Q}_2^{(b)}/\tau, \tag{C.16}$$

$$0 = \mu p \min(N, \bar{Q}_1^{(b)}) - \bar{Q}_2^{(b)}/\tau, \tag{C.17}$$

Adding (C.16) and (C.17), we obtain $0 = \lambda - \nu \min(N, \bar{Q}_1^{(b)})$, which implies:

$$\min(N, \bar{Q}_1^{(b)}) = \frac{\lambda}{\nu} \tag{C.18}$$

Substituting the solution from (C.18) into (C.17) and solving for $\bar{Q}_2^{(b)}$ results in $\bar{Q}_2^{(b)} = \tau\mu p\bar{Q}_1^{(b)}$. If $\lambda > N\nu$, then the right side of (C.18) is larger than N, while the left side must be less than or equal to N, which is contradiction, and therefore, no equilibrium point exists. On the other hand, if $\lambda \leq N\nu$, then the right side of (C.18) is less than or equal to N and therefore, the left side is $\bar{Q}_1^{(b)}$, and $(\bar{Q}_1^{(b)}, \tau\mu p\bar{Q}_1^{(b)})$ with $\bar{Q}_1^{(b)} = \lambda/\nu$ is an equilibrium point.

To analyze the stability of the equilibrium point, we express $(Q_1^{(b)}(t), Q_2^{(b)}(t))$ as:

$$Q_1^{(b)}(t) = \bar{Q}_1^{(b)} + U_1(t), \tag{C.19}$$

$$Q_2^{(b)}(t) = \tau\mu p\bar{Q}_1^{(b)} + U_2(t). \tag{C.20}$$

where $U_i(t), i = 1, 2$ are perturbations. If we assume $\lambda < N\nu$, then $\bar{Q}_1^{(b)} < N$ and $Q_1^{(b)}(t) \leq N$, if $U_i(t), i = 1, 2$ are sufficiently small. Under this assumption, substituting (C.19)-(C.20) in (8)-(9) results in the following, expressed in vector-matrix form:

$$\frac{d}{dt}U(t) = \begin{bmatrix} -\mu & 1/\tau \\ \mu p & -1/\tau \end{bmatrix} U(t) = AU(t) \tag{C.21}$$

The trace of A, $-\mu - 1/\tau < 0$, equals the sum of its eigenvalues. The determinant of A, $\mu/\tau - \mu p/\tau = \mu/\tau(1-p) > 0$, equals the product of its eigenvalues. The positive determinant implies that the two eigenvalues have the same sign, and given that their sum is negative, both eigenvalues must

Journal Pre-proof

be negative. Therefore, the equilibrium point $(\bar{Q}_1^{(b)}, \tau\mu p\bar{Q}_1^{(b)})$ is locally stable. □

Proof of Theorem 6.4. Assume $(Q_1^{(b)}(t), Q_2^{(b)}(t)) \to (x, y)$ as $t \to \infty$. Then (x, y) must satisfy:

$$0 = \lambda - \mu(x)\min(N, x) + y/\tau, \tag{C.22}$$

$$0 = \mu(x)p(x)\min(N, x) - y/\tau, \tag{C.23}$$

Adding (C.22) and (C.23) results in:

$$0 = \lambda - \nu(x)\min(N, x), \tag{C.24}$$

which is the same as (C.5). By the same argument as in the proof of Theorem 6.3, we conclude that (C.24) must have at least one solution, x, in the range $(0, \max(N, \tilde{x})]$. If we set $y = \tau\mu(x)p(x)\min(N, x)$, then (x, y) solves (C.22)-(C.23), which proves that the system (8)-(9) has at least one equilibrium point.

To analyse the stability of an equilibrium point $(x, \tau\mu(x)p(x)\min(N, x))$, we consider two cases: $x < N$ and $x > N$.

Case 1: $x < N$. We express $Q_1^{(b)}(t)$ and $Q_2^{(b)}(t)$ as:

$$Q_1^{(b)}(t) = x + U_1(t) \tag{C.25}$$

$$Q_2^{(b)}(t) = \tau\mu(x)p(x)x + U_2(t) \tag{C.26}$$

where $U_1(t)$, $U_2(t)$ are perturbations. Assuming that $U_1(t)$, $U_2(t)$ are small enough that $Q_1(t) \le N$, we can express (C.22)-(C.23) as:

$$Q_1^{(b)}(t) = F_1(Q_1(t), Q_2(t)), \tag{C.27}$$

$$Q_2^{(b)}(t) = F_2(Q_1(t), Q_2(t)), \tag{C.28}$$

where

$$F_1(u, v) = \lambda - u\mu(u) + v/\tau \tag{C.29}$$

$$F_2(u, v) = u\mu(u)p(u) - v/\tau \tag{C.30}$$

Journal Pre-proof

Journal Pre-proof

The Hessian matrix A is:

$$A = \begin{bmatrix} -(\mu(u) + u\mu'(u)) & 1/\tau \\ \mu(u)p(u) + u(\mu'(u)p(u) + \mu(u)p'(u)) & -1/\tau \end{bmatrix} \tag{C.31}$$

The linearized system is $\frac{d}{dt}U(t) = AU(t)$. The trace and determinant of A are:

$$\text{tr}(A) = -(\mu(u) + u\mu'(u)) - 1/\tau < 0 \tag{C.32}$$

$$\det(A) = ((\mu(u) + u\mu'(u)))(1/\tau) - (\mu(u)p(u) + u(\mu'(u)p(u) + \mu(u)p'(u)))(1/\tau) \tag{C.33}$$

$$= \frac{1}{\tau}\left(\nu(u) + u\nu'(u)\right) > 0 \tag{C.34}$$

The negative trace and positive determinant imply that both of A's eigenvalues are negative, and the equilibrium point is locally stable.

Case 2: $x > N$. The same type of linearization analysis as in Case 1 results in $\text{tr}(A) = -N\mu'(u) - 1/\tau < 0$ and $\det(A) = \frac{N}{\tau}\nu'(u) > 0$, and we conclude that the equilibrium point is locally stable. □

Journal Pre-proof