Customer Allocation Policies in a Two Server Network: Stability and Exact Asymptotics

A Thesis
Presented to
The Academic Faculty

by

Jerome D. Coombs-Reyes

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

School of Industrial and Systems Engineering
Georgia Institute of Technology
July 2003

Customer Allocation Policies in a Two Server Network: Stability and Exact Asymptotics

Approved by:

Robert D. Foley, Committee Chair

Douglas G. Down, Co-Advisor

Richard Serfozo

Hayriye Ayhan

Amy R. Ward

Date Approved 7/31/03

ACKNOWLEDGEMENTS

First, I would like to express my appreciation for the efforts of my primary advisor Robert Foley, as well as those of David McDonald and co-advisor Douglas Down. The value of your technical advisement has been immeasurable. I would also like to thank the AT&T Labs Fellowship Program, the Office of Naval Research, and the ARCS and FACES fellowship programs for their financial support of my academic endeavors.

I must also express gratitude towards the entire OMED staff here at Georgia Tech, in particular Bob Hume and S. Gordon Moore, who have shown untiring support for my academic efforts as both a graduate and undergraduate student. Likewise, I would not be here without the help, support, and guidance (academic and otherwise) of Mrs. Dalinda Brown Clark and her staff at the Dual Degree Engineering Program office. I must also express appreciation for my colleague Junxia Chang. It seems as if you were always there to answer even the silliest of questions for me in times when I couldn't think straight. And finally, I send my best regards to Mrs. Jeannie Baggett who, as the ace employee of the Graduate Studies office, means more to the graduate students at Tech than she likely realizes. Jeannie, without your assistance, in all likelihood I would have left the program a long time ago.

This work is dedicated to all of those who have given me the moral support that I needed to survive the rigor of my Ph.D. program. Namely, I devote these pages

to my parents, the Ortega-Narvaez family, and my dear friends Terry, Kermit, and Aselka. Together you have given my life balance, and hence have made the past few years worthwhile despite my many trials and tribulations along the way. Without the love, support and stability that you have shown me, this entire experience would have been impossible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Consider a two node queueing network where each node has a single server with an infinite capacity waiting area. Each node has a dedicated stream of arrivals and, in addition, there is a stream of arrivals that may join either queue. For the customer that may join either queue, we investigate several different policies. For each policy, we first determine the stability conditions for the queueing network. We then find asymptotic expressions for the probability of a large deviation of the queue length. As in McDonald [10], the emphasis here is on *exact* asymptotics, not rough asymptotics as in the traditional large deviation theory.

CHAPTER I

PROBLEM DESCRIPTION

Consider a queueing network with two servers, each having its own queue. Each of these servers also has its own dedicated stream of customers; that is, a set of customers that can only join that particular server's queue. In addition, there is a stream of customers that may join either queue. These discretionary customers join the queue containing the smaller number of customers, with ties being broken arbitrarily. In using this join-the-shortest-queue policy (hereafter referred to as JSQ), a system with similarly paced servers will appear to be fairly efficient. However, once the service rates between the servers begin to differ significantly, the shortest queue may not be the best to join. In this case, a more natural policy may involve joining the queue with the shortest expected waiting time. Such a system will be referred to as a JSEW queueing system, and the corresponding discretionary customers will be called JSEW customers.

Closely related to the JSEW policy is the "shortest expected delay routing" policy, abbreviated JSED. The difference between the two policies is that the JSED policy takes into consideration the customers currently being served, while JSEW only considers the expected service times of the customers that are queued.

The primary objective of this research is to describe the overload characteristics

of the aforementioned JSEW queueing system. We will apply the methodology previously developed in [10] (and already applied to the JSQ system in [6]) in order to characterize the asymptotic behavior of the system described above. Precise definitions and explanations of the desired exact asymptotic expressions and other results will be described in the Summary of Main Results section.

1.1 The Model and Policies

Our network consists of two nodes. Each node has a single server and an infinite capacity waiting area. Customers arrive to the network according to a Poisson process with rate $\lambda > 0$. Each customer has an associated type, and the type defines how a customer decides whether to enter node 1 or node 2. The decision is allowed to depend upon the joint queue length at the time of the customer's arrival and to be randomized. After deciding which node to enter, the customer joins the end of the line at that node, waits for service, is serviced, and then departs from the system. The arrivals of a particular type form a Poisson process with a rate depending upon the type. Poisson processes for different customer types are mutually independent.

There are an unlimited number of types or decision rules, but we will be particularly interested in the following four rules. The two customer types with the simplest decision rules are those of the dedicated customers. Customers of type 1 always join node 1, and we let λ_1 denote the arrival rate of the customers dedicated to queue 1. Similarly, customers of type 2 always join node 2 at a rate λ_2. The third customer type is the customer that joins the shorter queue, breaking ties at random (JSQ), and we let λ_Q denote rate of JSQ customers. The fourth customer type joins the queue with the shorter expected waiting time, breaking ties at random (JSEW), and we let

λ_E denote the arrival rate of the JSEW customers.

If $\lambda = \lambda_1 + \lambda_2 + \lambda_E$, then we know that there are only three customer types in this system since λ represents the total arrival rate.

The total service rate in the network is denoted by μ. Just as the arrival rate is split between the two nodes, so is the service rate. We let μ_1 denote the service rate always dedicated to node 1. Similarly, μ_2 denotes the service rate always dedicated to node 2. Most of our results assume $\mu = \mu_1 + \mu_2$ implying that the service times at node i are exponentially distributed with parameter μ_i for $i = 1, 2$. However, in one section, we allow $\mu = \mu_1 + \mu_2 + \mu_L$ where μ_L is allocated to the node with the longer queue and split equally between the two nodes in the case of a tie.

Under the above assumptions, the joint queue length process will be a continuous-time Markov process. If we know the components of λ, their associated rules, and the components of μ, we can construct the generator of the joint queue length process. For example, if $\lambda = \lambda_1 + \lambda_2$ and $\mu = \mu_1 + \mu_2$, then the network consists of two independent M/M/1 queues. If we say that the system is a JSQ system, then we are assuming that $\lambda = \lambda_1 + \lambda_2 + \lambda_Q$ and $\mu = \mu_1 + \mu_2$. A JSEW system is quite similar except that $\lambda = \lambda_1 + \lambda_2 + \lambda_E$. A JSQ/SLQ (join shorter queue/serve longer queue) system allows $\lambda = \lambda_1 + \lambda_2 + \lambda_Q$ and $\mu = \mu_1 + \mu_2 + \mu_L$.

Let $\mathbf{Q}(t) = (Q_1(t), Q_2(t))$ denote the queue lengths at time t. For the policies being considered within this dissertation, $\mathbf{Q}(\cdot)$ being an irreducible Markov process on the state space $\mathbb{Z}_+^2$ follows from the assumption that $\lambda - \lambda_1 - \lambda_2$, μ_1 and μ_2 are all nonzero. Our rare event of interest will be the event that we reach any state in the set $F_\ell \equiv \{(x, y) \in \mathbb{Z}_+^2 \mid x + y \geq \ell\}$, where x and y are the queue lengths at nodes 1 and 2, respectively. In other words, we are analyzing large deviations of the total

number of customers in the system.

1.2 Summary of Main Results

Consider a two-node infinite capacity system with multiple types of customers and total arrival rate λ. The arrival rate of customers dedicated to node i will be denoted by λ_i, $i = 1, 2$, and all remaining customer types will arrive with combined rate $\lambda - \lambda_1 - \lambda_2$. Let $\rho_1 = \frac{\lambda_1}{\mu_1}$, $\rho_2 = \frac{\lambda_2}{\mu_2}$, $\rho = \frac{\lambda}{\mu}$, and $\rho_{\max} = \max\{\rho, \rho_1, \rho_2\}$. The service times at each node are independent and identically distributed (iid) with parameters μ_1 and μ_2, respectively. Customer arrivals are also iid, and occur from infinite sources with total rate λ.

Given the above framework, the following theorem provides stability conditions for a wide array of networks, including JSEW, JSQ, and combinations thereof. First note that whenever either queue is "close" to being empty, if the discretionary customers of a given policy obey JSQ routing then the policy will be called boundary-avoiding. This definition will be made precise in Section 2.2.

Theorem 1 *Given the queueing model described above, if the generator is work-conserving and boundary-avoiding with $\rho_{\max} < 1$, then the joint queue length process is positive recurrent.*

By work-conserving we mean that each server will remain busy when there is work to be done by that server. With the assumption that these stability conditions hold, we then move on to our exact asymptotic analysis. In using the methodology of [10] to obtain our results, several technical conditions have to be satisfied in order to utilize the implied results discussed in [10]. We will give a full account of these procedures in subsequent sections. Once these technical conditions are satisfied, we will be able

to make the results of [6] and [10] specific to our system and draw our conclusions. Note that in this paper, as in [6] and [10], finding an exact asymptotic expression for some function $f(\ell)$ will mean determining not only the rate α, but also the constant c such that $f(\ell) \sim c\alpha^{\ell}$; i.e.

$$\lim_{\ell \to \infty} \frac{f(\ell)}{c\alpha^{\ell}} = 1.$$

It turns out that there are two distinct ways that the JSEW system can overload. If one server is very fast, then that server will take all of the discretionary customers and will keep its queue small while the dedicated customers at the other queue cause the system to overload. This is called the unpooled case. On the other hand, if the proportion of discretionary customers is large enough, then the two queues will instead overload in tandem and this is called the pooled case. For each case, the results for the JSEW network with exponentially distributed service times and Poisson arrivals are given below. Theorem 2 gives the results for a "large portion" of the parameter space in the pooled case (we will make this idea rigorous in Section 5.4), while Theorem 3 addresses the unpooled scenario. Define $A_j \equiv \{k \mid \frac{y}{\mu_2} - \frac{x}{\mu_1} = k, x + y = j \text{ for some } x, y \in \mathbb{Z}\}$. When the ratio μ_2/μ_1 is rational, there will exist a $p \in \mathbb{Z}_+$ such that $A_j = A_i$ for $i = j \mod p$. Let $\mathrm{P}_\pi\{\cdot\}$ denote the probability measure conditioned on the process starting in Δ with respect to the stationary distribution of $\mathbf{Q}(t)$.

Theorem 2 *(Pooled Case) Let $b = \frac{\mu_2}{\mu_1}$. Assume b is rational and let T_ℓ denote the first time that there are ℓ or more customers in the system. If $\max\{\rho_1, \rho_2\} < \rho < 1$, and $\ln \rho_1 < \sqrt{b} \ln \rho$ then*

(i)

$$\mathrm{E}[T_\ell \mid Q[0] = (0,0)] \sim {f_1}^{-1} \rho^{-\ell} / (\lambda + \mu) \tag{1}$$

where f_1 is given in (52). Moreover, we also have

(ii)

$$\mathrm{P}_\pi\left\{Q_1(t)+Q_2(t)=\ell, \frac{Q_2(t)}{\mu_2}-\frac{Q_1(t)}{\mu_1}=k\right\} \sim f_1\rho^\ell \frac{1}{\tilde{d}}\varphi(k)p \tag{2}$$

(iii)

$$\mathrm{P}\left\{\frac{Q_2(T_\ell)}{\mu_2}-\frac{Q_1(T_\ell)}{\mu_1}=k \mid \boldsymbol{Q}[0]=(0,0)\right\} \sim p\ \varphi(k) \tag{3}$$

for $k \in A_\ell$, where $\tilde{d}=\mu-\lambda$, and the density $\varphi(k)$ and p will be discussed in Sections 4.1 *and* 5.2, *respectively.*

Theorem 3 *(Unpooled Case) Suppose $\rho < \max(\rho_1,\rho_2) < 1$. Without loss of generality, assume $\rho_1 > \rho_2$. Then,*

(i)

$$\mathrm{E}[T_\ell \mid \boldsymbol{Q}[0]=(0,0)] \sim g_1^{-1}{\rho_1}^{-\ell}/(\lambda+\mu) \tag{4}$$

where g_1 is given in (54). Moreover,

(ii)

$$\mathrm{P}_\pi\{Q_1(t)+Q_2(t)=\ell, Q_2(t)=k\} \sim \frac{f_1}{\mu_1-\lambda_1}{\rho_1}^{\ell-k}\left(1-\frac{\lambda_2+\lambda_E}{\mu_2}\right)\left(\frac{\lambda_2+\lambda_E}{\mu_2}\right)^k, \tag{5}$$

(iii)

$$\mathrm{P}\{Q_2(T_\ell)=k \mid \boldsymbol{Q}[0]=(0,0)\} \sim c^{-1}{\rho_1}^k\nu(0,k), \tag{6}$$

where c and the distribution ν are defined in Section 5.3.

When analyzing the JSQ system in [6], it was found that there were (not 2 but) 3 distinct ways that the system could approach F_ℓ depending on the traffic intensity parameters ρ_1, ρ_2 and ρ. Just as in our JSEW network, if ρ_1 was the largest, then queue 2 remained small while queue 1 grew large. This was again called the unpooled case, and the most likely approach to F_ℓ in this case bounced along the x-axis. Analogously, having ρ_2 as the largest in the JSQ system led to an approach along the

y-axis. However, when $\rho > \max(\rho_1, \rho_2)$ the classification of the network's behavior was more involved.

For the JSQ pooled case, an additional criterion was uncovered that divided the pooled case into two subcases, weak and strong pooling. In cases when this additional criterion held, the most likely approach to F_ℓ hugged the line $x = y$, and this was called the strongly pooled case. However, if ρ was the largest and the criterion did not hold true, then both queues still overloaded, but a large joint queue length could potentially be achieved via a slope other than 1. This scenario was called the weakly pooled case.

As we will shortly discover, it turns out that such a dichotomy does not exist in our JSEW network. Instead, in analyzing JSEW, the required condition developed analogous to the one that presented the strong/weak pooling dichotomy in JSQ turns out to be superfluous with respect to the pooling condition itself.

1.3 Literature Review

In this section, we first discuss the origin of the methodology used in this thesis to obtain our asymptotic results, as well as previous applications of this methodology. We then move on to consider other research contributions related to the JSEW policy considered within this thesis.

The methodology used in this thesis was originally developed in McDonald [10]. In his paper, McDonald modeled a general network of queues as a Markov additive chain, and proved three rare-event theorems. The first result in [10] provides an asymptotic expression for the mean time until the specified additive component hits a high level ℓ. For the Markovian component of the queueing network, the remaining two theorems provide both the limiting distribution at this hitting time, as well as

the steady state distribution when the specified additive component equals ℓ. As mentioned earlier, the emphasis in his work was on sharp asymptotics, not rough asymptotics as in the more traditional approach to large deviation theory [the reader is referred to the textbook of Shwartz and Weiss [15] for a treatment of this subject]. Moreover, the limiting distributions were for the unscaled process, not for the fluid limit as in the traditional large deviation theory. The proofs of the results in [10] were constructed such that, given a particular queueing network and policy, if the numerous conditions outlined in the paper could be established for the given network, then the main theorems of [10] could be made applicable to the desired network. In fact, showing that these conditions are satisfied for a JSEW policy within our two-server queueing network is what we seek to do in this thesis.

An application of this methodology is given in Brown [1]. In her work, Brown considers a system with N servers arranged in a circle, with N arrival streams, each having the discretion to choose between the shorter of two adjacent queues. In the case of $N = 3$ with exponential servers and Poisson arrivals, Brown considers the approach of McDonald [10] in order to characterize (both qualitatively and quantitatively) how this system overloads. In Foley and McDonald [6], the methodology was extended to cover periodicities, and also utilized to obtain the aforementioned asymptotic results for a two-server system under a JSQ policy.

The work most closely related to this thesis would be that of McDonald and Turner [11] in which they consider analogous results for a "join the shorter actual waiting time" policy (JSAW). In their paper, by using the results proven in this thesis as well as in [6] for the JSEW and JSQ policies, respectively, they compared both the rough and sharp asymptotics of the three policies across different overflow events. In short, for overflow events concerning the total amount of work in the system, McDonald and Turner found that the performance of all three polices is roughly the same in the

sense that the probability of an overflow only differed by a sub-exponential factor. However, when the overflow event concerned the waiting time in either queue, they found that the three policies differed by an exponential factor, and in fact that JSAW was substantially better.

With respect to the JSEW policy, there is an extensive literature detailing the optimality of this and similar policies. In [17], Turner used the traditional large deviation theory approach to find rough asymptotic expressions for the rare event discussed in this thesis. However, few others have considered the perspective of rare events and large deviations as it pertains to this policy. Such rare event considerations are particularly worthwhile when dealing with networks where queueing space is costly, or the dynamics of the queueing system change significantly after some predetermined queue length threshold. For the works listed below, the performance measure commonly used to consider optimality of a policy was the long run average delay per customer. Also note that, unlike our work, most of the following papers are limited to models with identically distributed servers and no dedicated customer arrival streams. Moreover, many of the considerations involve the JSED policy, which differs slightly from our JSEW policy. Nonetheless, we briefly detail some of the more well-known results achieved in this area.

Consider a two node network with a single server at each node, no dedicated customers, and service times for each server being identically distributed. For this system, Weber [18] has shown that JSED is optimal for arbitrary arrival streams with respect to average delay per customer. In [9], Houck has shown via simulation results that with Poisson arrivals and multiple servers at each node with exponentially distributed service times, the JSED policy results in a mean delay that is almost the same as the delay resulting from a system in which the servers are combined at one node. For single server nodes with general service times and multi-server nodes with

exponential service times, Whitt [19] has developed counterexamples showing that JSED is not optimal.

In the heavy traffic case with one server at each of two nodes, identical service times, and dedicated customers, Foschini & Salz [8] showed that the JSED system behaves as if it were a two-server system with a single queue. In addition, a closed form expression is obtained for the average delay advantage of the dynamic JSED rule over a static random policy where jobs are routed according to the appropriate fixed probability. In [7], Foschini extended this result to systems where the two servers have different service rates, and also to systems with more than two servers where separate arrival streams have to choose from different sets of servers. He showed that in heavy traffic, all of these systems behave as if the servers were pooled into one group of resources and able to serve any arriving customer. In [14], Reiman extended the work of Foschini and Salz to the iid setting.

1.4 Overview

The remainder of this dissertation will proceed as follows. In Section 2.1, we will obtain stability conditions for our JSEW system using a connection between our stochastic network and the corresponding fluid model. In Section 2.2, we will use the same approach to derive a much more general result that gives the stability conditions for a wide class of queueing networks.

As mentioned earlier, this paper will follow the methodology developed in [10] for analyzing the exact asymptotics of a queueing network. To facilitate easier reading, Chapter 3 contains an outline of this methodology suitable for discrete-state queueing systems. In Chapter 4, we will apply this methodology to our JSEW system and

provide the necessary proofs.

Upon completing the proofs for the JSEW system, the remainder of the dissertation can be outlined as follows. In Section 5.1, we will discuss our results for the JSEW policy as they relate to the qualitative behavior of the system as it approaches our rare event of interest. In particular, we will compare these findings with those obtained for the JSQ policy in [6]. In Section 5.2, we address the problem of periodicity within the JSEW policy, and then move on to give our main results in Section 5.3. In Section 5.4, we discuss an additional assumption that was necessary to complete a proof in Chapter 4. In Section 5.5, we use simulation results to briefly illustrate the usefulness of the exact asymptotic expressions derived in Chapter 4. In the final section of the thesis, we will consider a variation of the JSQ system that includes the addition of a third discretionary server that serves the longest of the two queues (hereafter referred to as a JSQ/SLQ policy).

CHAPTER II

STABILITY

Let $\{\mathbf{Q}(t), t \geq 0\}$ be a Markov process describing a queueing network; i.e., such that the joint queue length at time t can be determined from $\mathbf{Q}(t)$. We say that the queueing network is stable if $\mathbf{Q}(t)$ is positive Harris recurrent, which we will define below. In [3] and [4], Dai explores the connection between the stability of the fluid model and that of the network itself. He uses piecewise-linear Lyapunov functions to show stability of some fluid models, which in turn guarantees positive Harris recurrence for the corresponding stochastic networks. We will use such an approach in order to derive stability conditions for our JSEW network.

In the following paragraphs, we outline the steps required for our stability proofs. In Section 2.1, we follow these steps in order to obtain stability conditions for our JSEW system. Finally, in Section 2.2, we derive a more general result which gives the stability conditions for an entire class of queueing networks.

There are three major steps to consider in our proof of stability. Given the Markov Process $\mathbf{Q}(t)$ defined by our JSEW network, first we must define the corresponding fluid model. The equations that define this fluid model will be formulated to parallel the properties of our JSEW stochastic model. Once this is done, our next step will be to prove that each fluid limit of the stochastic model is a solution to the set of

equations that define our fluid model. This will ensure that we have correctly defined the fluid model corresponding to our stochastic network. Lastly, we must prove that this fluid model (under the associated queueing discipline) is stable. According to Theorem 4.2 of [3], the fluid model being stable will imply that the corresponding queueing network under the same queueing discipline is positive Harris recurrent; thus our proof for stability will be completed. By Theorem 10.2.1 in [13], the positive Harris recurrence of our Markov process implies the existence of a unique stationary distribution. This will be important in our characterization of the system's asymptotic behavior.

Let $X = \{X(t), t \geq 0\}$ be a Markov process that evolves on a state space $\mathbf{X}$ with Borel σ-field $\mathbb{B}_{\mathbf{X}}$. For any set $A \in \mathbb{B}_{\mathbf{X}}$, also define $\tau_A = \inf\{t \geq 0 : X_t \in A\}$. The following is the definition of positive Harris recurrence.

Definition 1 *The process X is Harris recurrent if there exists some σ-finite measure μ on $(X, \mathbb{B}_X)$, such that whenever $\mu(A) > 0$ and $A \in \mathbb{B}_X$,*

$$P_x\{\tau_A < \infty\} \equiv 1.$$

If μ is finite, then it may be normalized to a probability measure; in this case X is called positive Harris recurrent.

The next two definitions will also be helpful during our discussion of fluid models and the derivation of our stability conditions.

Definition 2 *A sequence of functions $f_n(\cdot) : \mathbb{R}_+ \to \mathbb{R}$ is said to be convergent to $f(\cdot)$ uniformly on compact sets (u.o.c.) if for every $t > 0$,*

$$\sup_{0 \leq s \leq t} |f_n(s) - f(s)| \to 0.$$

Definition 3 *A set $\{X_r : r \in \mathbb{R}\}$ is said to be relatively compact if for all sequences $\{r_n\} \subset \mathbb{R}$ there exists a subsequence $\{r_{n_k}\}$ such that X_{n_k} converges u.o.c.*

Consider a two-node infinite capacity system with J types of customers. For such a system, we define the following two cumulative processes. Let $S = \{S_k(n), n = 1, 2, \ldots\}$, where $S_k(n)$ is the total service requirement for the first n customers at node $k = 1, 2$. Let $E = \{E_j(t), t \geq 0\}$, where $E_j(t)$ is the number of arrivals of type j customers by time t. For the networks considered in this chapter, we assume that the strong law of large numbers holds for these cumulative processes. That is, with probability one, we must have

$$\lim_{t\to\infty} \frac{E_j(t)}{t} = \lambda_j \tag{7}$$

$$\lim_{n\to\infty} \frac{S_k(n)}{n} = \frac{1}{\mu_k} \tag{8}$$

for some λ_j, $\mu_k > 0$.

When considering the JSEW network, we will have $J = 3$. Cumulative processes E_1 and E_2 (with respective limits of λ_1 and λ_2) will correspond to dedicated customers at nodes 1 and 2, respectively. Type 3 customers will be the discretionary JSEW customers, and the corresponding limit will instead be denoted by λ_E. For the more general system considered in Section 2.2, the cumulative processes for any additional customer types can be assigned analogously. As done throughout the thesis, we will let λ and μ denote the total arrival and service rates of the system.

2.1 Stability Conditions for the JSEW Network

In this section, we will derive the stability conditions for a two-server queueing network which implements a JSEW queueing policy. To specify the dynamics of our

network, we first define the following descriptive processes. The number of arrivals to node k by time t will be denoted as $A_k(t)$, and likewise the number of departures as $D_k(t)$, $k = 1, 2$. Let $Q = \{Q(t), t > 0\}$ denote the two-dimensional queue length process. The other three processes, $W = \{W(t), t > 0\}$, $Y = \{Y(t), t > 0\}$ and $T = \{T(t), t > 0\}$ are also two-dimensional. As done for the arrival and departure processes, we will often use the subscript "k" to denote the k^{th} component of a multi-dimensional vector. W, the workload process, is used to denote the expected amount of work for each server (measured in units of remaining service time). We will let Y denote the total amount of time that the servers have been idle in the time interval $[0, t]$. And finally, T denotes the amount of time that the servers have spent serving their customers. For any function f, we will use $\dot{f}(t)$ to denote the derivative of f at t when it exists.

Let us now consider the formal deterministic analog of our JSEW queueing network. In giving our fluid model description of the JSEW queueing network, we will follow [3] with the intentional reuse of the symbols defined in the previous paragraph. This will be done in order to emphasize the parallel between the stochastic queueing network and its corresponding deterministic continuous fluid model. For example, as in the stochastic model, the queue length process in the fluid model will also be denoted as Q, and may instead be referred to as the fluid level. The components of our fluid model must satisfy the following criteria:

$$A_1(t) + A_2(t) = \lambda t, \tag{9}$$

$$D_k(t) = \mu_k T_k(t), \tag{10}$$

$$Q_k(t) = Q_k(0) + A_k(t) - D_k(t), \tag{11}$$

$$W_k(t) = \frac{1}{\mu_k}(Q_k(0) + A_k(t)) - T(t), \tag{12}$$

$$T(t) + Y(t) = et, \tag{13}$$

$$Y_k(t) \text{ can only increase when } W_k(t) = 0, \tag{14}$$

$$(\frac{1}{\mu_1})Q_1(t) > (\frac{1}{\mu_2})Q_2(t) \Longrightarrow \dot{A}_1(t) = \lambda_1, \tag{15}$$

$$(\frac{1}{\mu_2})Q_2(t) > (\frac{1}{\mu_1})Q_1(t) \Longrightarrow \dot{A}_2(t) = \lambda_2, \tag{16}$$

for all $t \geq 0$ and $k = 1, 2$, where e denotes a vector of ones. These criteria together define our fluid model; and their solutions, written as

$$\mathbb{X} = (A(t), D(t), T(t), W(t), Y(t), Q(t)), \ t \geq 0,$$

will be referred to as the *fluid model solutions.* Throughout our analysis, we will operate under the assumptions that $A(0) = D(0) = T(0) = Y(0) = 0$, and that all components are nonnegative. We also assume that the components A, T, and Y are nondecreasing.

Definition 4 *A function f is Lipschitz continuous if for some $M > 0$,*
$|f(t_2) - f(t_1)| \leq M(t_2 - t_1)$ for all $0 \leq t_1 \leq t_2$.

Proposition 1 *Each component of $\mathbb{X}$ is Lipschitz continuous.*

Proof Suppose $0 \leq s \leq t$. Consider the following inequalities:

$\mathrm{T}_k(t) - T_k(s) = (t - s) - (Y_k(t) - Y_k(s)) \leq t - s.$

$\mathrm{Y}_k(t) - Y_k(s) = (t - s) - (T_k(t) - T_k(s)) \leq t - s.$

$\mathrm{A}_k(t) - A_k(s) \leq (t - s)(\lambda_k + \lambda_E)$ for $k = 1, 2.$

$\mathrm{D}_k(t) - D_k(s) \leq (t - s)\mu_k$ for $k = 1, 2.$

Here the first two statements follow from the nondecreasing property of the respective components, and the fourth from equation (10). Lastly, note that $Q_k(t)$ and $W_k(t)$, for $k = 1, 2$, are linear combinations of the aforementioned components and hence

are also Lipschitz continuous. Hence, using (9) - (13), we have shown that each component of $\mathbb{X}$ is Lipschitz continuous.

□

Definition 5 *A function f is absolutely continuous if for each $s, t > 0$ and $\epsilon > 0$ there exists $\delta > 0$ such that for each collection $0 < s \leq s_1 < t_1 < s_2 < t_2 < ... < t_n \leq t$ with $\sum(t_i - s_i) < \delta$, we have $\sum |f(t_i) - f(s_i)| \leq \epsilon$.*

For a given Lipschitz continuous function f, choose $\delta = \epsilon/M$, where M is as defined in Definition 4. Thus for any $\epsilon > 0$, if some disjoint intervals (s_1, t_1),...,(s_n, t_n) are such that $\sum(t_i - s_i) < \delta$, then $\sum |f(t_i) - f(s_i)| \leq \sum M(t_i - s_i) \leq \epsilon$. Hence, as a result of Proposition 1, each component of $\mathbb{X}$ is also absolutely continuous, and therefore differentiable almost everywhere with respect to Lebesgue measure on $[0, \infty)$. Thus, whenever we are discussing the derivative of any component of $\mathbb{X}$ for a given time t, our assumption that the component is differentiable is justified.

Let $\bar{Q}^r(t) = \frac{1}{r}Q(rt)$ denote the fluid scaling of the queue length process [in this context, $Q(rt)$ is the stochastic network queue length]. The set of limit points obtained by letting r go to infinity is called the fluid limit, and will be denoted by $\bar{Q}(t)$. Similar notation will be used for the fluid limits of each component of $\mathbb{X}$. As alluded to earlier, when we say that a fluid model corresponds to a particular queueing network, we mean that each fluid limit of the queueing network is a solution to the fluid model. Hence, for this notion to make sense, we must be assured the existence of at least one limit point for the queueing network. This requirement is satisfied by Theorem 4.1 in [3] which tells us that, for any work-conserving queueing discipline, the scaled components of the queueing network are indeed relatively compact. By work-conserving, we simply mean that each server in the network works continuously

whenever there is work to be done by that server. In the following proposition, we establish the validity of our fluid model.

Proposition 2 *Each fluid limit of the original stochastic network is a solution to the fluid model defined by (9)-(16).*

Proof The first six equations, (9)-(14), are not specific to our particular fluid model, but instead are descriptive for any generic head-of-line queueing network (i.e., a network in which each queue is served in a FIFO manner, with each node having at most one job being served at a time). As such, by Theorem 4.1 in [3] we already know that each fluid limit of our stochastic model is a fluid model solution for these six equations. What is left to consider are the last two criteria which are what make the model specific to our particular queueing discipline. Thus to check whether this fluid model accurately represents our JSEW network, we need only show that the fluid limit of our stochastic model satisfies these two criteria, (15) and (16).

Assume that $\frac{1}{\mu_2}\bar{Q}_2(t) > \frac{1}{\mu_1}\bar{Q}_1(t)$.
Therefore, by continuity, there exists $\delta > 0$ such that $\frac{\mu_1}{\mu_2}\bar{Q}_2(s) - \bar{Q}_1(s) > 0$ for $s \in [t-\delta, t+\delta]$. Let $\epsilon = \min\left(\frac{\mu_1}{\mu_2}\bar{Q}_2(s) - \bar{Q}_1(s)\right)$ over $s \in [t-\delta, t+\delta]$. Clearly $\epsilon > 0$.

Suppose there exists $\{r_n\} \to \infty$ such that $(\bar{Q}^{r_n}(t), \bar{A}^{r_n}(t)) \to (\bar{Q}(t), \bar{A}(t))$ as $r_n \to \infty$. This would imply that $\frac{\mu_1}{\mu_2}\bar{Q}_2^{r_n}(s) - \bar{Q}_1^{r_n}(s) \to \frac{\mu_1}{\mu_2}\bar{Q}_2(s) - \bar{Q}_1(s)$ uniformly on compact sets. Hence, by the definition of convergence u.o.c., we would be ensured of the existence of an $N > 0$ such that

$$\sup_{s\in[t-\delta,t+\delta]} \left| \frac{\mu_1}{\mu_2}\bar{Q}_2^{r_n}(s) - \bar{Q}_1^{r_n}(s) - \left(\frac{\mu_1}{\mu_2}\bar{Q}_2(s) - \bar{Q}_1(s)\right)\right| \leq \frac{\epsilon}{2}$$

for $n > N$.

Therefore, $\frac{\mu_1}{\mu_2}\bar{Q}_2^{r_n}(s) - \bar{Q}_1^{r_n}(s) \geq \frac{\epsilon}{2}$ for $n \geq N, s \in [t-\delta, t+\delta]$;

i.e., $\frac{1}{\mu_2 r_n} Q_2(r_n s) \geq \frac{1}{\mu_1 r_n} Q_1(r_n s) + \frac{\epsilon}{2\mu_1}$;

i.e., $\frac{1}{\mu_2} Q_2(s) \geq \frac{1}{\mu_1} Q_1(s) + \frac{\epsilon}{2\mu_1} r_n$ for $s \in [r_n(t-\delta), r_n(t+\delta)]$.

Given the above inequality, our queueing discipline dictates that there are no discretionary customer arrivals to queue 2 in this scenario. That is, $\forall s \in [t-\delta, t+\delta]$, we obtain $A_2(r_n s) - A_2(r_n(t-\delta)) = E_2(r_n s) - E_2(r_n(t-\delta))$. By multiplying each side by $\frac{1}{r_n}$ and letting r_n go to infinity, we can rewrite this as $\bar{A}_2(s) - \bar{A}_2(t-\delta) = \lambda_2(s-t+\delta)$. Hence, $\dot{\bar{A}}_2(t) = \lambda_2$ as desired.

In the case when $(\frac{1}{\mu_1})\bar{Q}_1(t) > (\frac{1}{\mu_2})\bar{Q}_2(t)$, we analogously obtain that $\dot{\bar{A}}_1(t) = \lambda_1$.

□

As defined in [3], a fluid model is called *stable* if there exists a $\delta > 0$ such that for each fluid solution $\mathbb{X}$ with $|Q(0)| \leq 1, Q(t) = 0$ for $t \geq \delta$. Hence our goal is to show that the fluid limit of the model corresponding to our JSEW system must eventually reach zero and stay there regardless of the initial system configuration. This property is stated and proved in the following proposition.

Proposition 3 *The fluid model as defined by (9)-(16) is stable.*

Proof Let $\mathbb{X}$ be a fluid model solution of the model defined by equations (9)-(16). Define the piecewise-linear Lyapunov function $f(t) = \max(\frac{1}{\mu_1} Q_1(t), \frac{1}{\mu_2} Q_2(t))$. Let $\epsilon > 0$. Our goal is to show that $f(t) > 0 \implies \dot{f}(t) \leq -\epsilon$, which by Lemma 3 of [4] would yield our desired result that $f(t) = 0$ for all $t \geq f(0)/\epsilon$.

Let $t > 0$ be such that $f(t) > 0$, f is differentiable at t, and Q is differentiable at t. Recall that, by the absolute continuity of Q, we only need to consider t which satisfy

this criterion.

Case 1. Suppose $\frac{1}{\mu_1}Q_1(t) > \frac{1}{\mu_2}Q_2(t)$.

Then $\dot{f}(t) = \frac{1}{\mu_1}\dot{Q}_1(t) = \frac{\lambda_1}{\mu_1} - 1 < 0$ iff $\mu_1 > \lambda_1$.

Case 2. Suppose $\frac{1}{\mu_2}Q_2(t) > \frac{1}{\mu_1}Q_1(t)$.

Then $\dot{f}(t) = \frac{1}{\mu_2}\dot{Q}_2(t) = \frac{\lambda_2}{\mu_2} - 1 < 0$ iff $\mu_2 > \lambda_2$.

Case 3. Suppose $\frac{1}{\mu_1}Q_1(t) = \frac{1}{\mu_2}Q_2(t)$.

$$\begin{aligned}\text{Then } \dot{f}(t) &= \tfrac{1}{2}\tfrac{d}{dt}(\tfrac{1}{\mu_1}Q_1(t) + \tfrac{1}{\mu_2}Q_2(t)) \\ &< (2\min(\mu_1, \mu_2))^{-1}(\lambda - \mu) \\ &< 0 \text{ iff } \mu > \lambda.\end{aligned}$$

□

Define the traffic intensity parameters $\rho_1 = \frac{\lambda_1}{\mu_1}$, $\rho_2 = \frac{\lambda_2}{\mu_2}$, and $\rho = \frac{\lambda}{\mu}$. Let $\rho_{\max} \equiv \max\{\rho_1, \rho_2, \rho\}$. Through Propositions 2 and 3, we have proved the following theorem.

Theorem 4 *If $\rho_{\max} < 1$, then the JSEW Markov Process $\boldsymbol{Q}(t)$ is positive Harris recurrent.*

Observe that if we set $\mu_1 = \mu_2$, then the fluid model defined by equations (9) - (16) also corresponds to the deterministic analog of a general JSQ queueing network (by general, we mean that the service rates need not be equal at the two nodes). Moreover, the analog of Proposition 1 for the JSQ system also follows from the proof given for JSEW; and by setting $\mu_1 = \mu_2$ in the proof of Proposition 2 we also obtain the proof of the analogous proposition for the general JSQ queueing network. For the proof of the analog to Proposition 3, we instead use the Lyapunov function

$f(t) = \max(Q_1(t), Q_2(t))$, and the proof follows likewise. Hence, we have the following Corollary.

Corollary 1 *If $\rho_{\max} < 1$, then the JSQ Markov Process $\boldsymbol{Q}(t)$ is positive Harris recurrent.*

2.2 Stability for a General Two-Server Network

In this section, we will extend our stability results to a wide class of queueing systems. The description of this general system is given as follows.

Consider a two-node system with the following types of customers: a customer arrival stream dedicated to each node and any number of discretionary customer arrival streams that follow some predetermined well-defined queueing policies. The exact behavior of the discretionary customers (e.g., JSEW, JSQ, or some combination of the two) will purposely be left ambiguous.

Our claim is that the queueing system described above is stable for any work-conserving, boundary-avoiding queueing policy. By boundary-avoiding, we mean that there exists a region "close" to each axis such that within that region the discretionary customers are routed to the opposite queue. More formally, consider a graph of our system where the queue lengths Q_1 and Q_2 are represented respectively by the x and y coordinates. Furthermore, suppose that there exist wedges along the x and y axes defined by the lines $y = m_1x + b_1$, and $y = m_2x + b_2$, respectively, where $m_1 \leq m_2$. If the discretionary customers of a given policy obey JSQ routing whenever $y < m_1x + b_1$ and $y > m_2x + b_2$, then the policy will be considered boundary-avoiding.

In considering the stability conditions for this general system, we will use the

same fluid model framework defined for the JSEW scenario in Section 2.1, and make changes where necessary. As done before, to show stability for the fluid model, the first step is to define the fluid model corresponding to our stochastic network. This is done in a similar fashion to what was presented for JSEW. In fact, the first six equations used to define our fluid model, (9) - (14), remain unchanged. However, we must now replace (15) and (16) with new equations that correspond to the queueing discipline of our general system. Recall we have only specified that our queueing network obeys the JSQ policy within each of the wedges. For the fluid model, we state this constraint as follows.

$$Q_2(t) < m_1 Q_1(t) + b_1 \Longrightarrow \dot{A}_1(t) = \lambda_1 \tag{17}$$

$$Q_2(t) > m_2 Q_1(t) + b_2 \Longrightarrow \dot{A}_2(t) = \lambda_2 \tag{18}$$

Given these new equations, as with the JSEW model, we must next check whether this fluid model accurately represents the network in question. As before, the first six equations are already known to suffice, and we need only show that the fluid limit of our stochastic model is a fluid model solution to (17) and (18). The proof of this result is similar to that given in Proposition 2 for the JSEW network. However, for the sake of completeness, we present it here.

Proposition 4 *Each fluid limit of the original stochastic network described above is a solution to the fluid model equations (17) and (18).*

Proof Assume that $\bar{Q}_2(t) > m_2 \bar{Q}_1(t) + b_2$.

Therefore, by continuity, there exists $\delta > 0$ such that $\bar{Q}_2(s) - m_2 \bar{Q}_1(s) > b_2$ for $s \in [t-\delta, t+\delta]$. Let $\epsilon = \min\left(\bar{Q}_2(s) - m_2 \bar{Q}_1(s) - b_2\right)$ over $s \in [t-\delta, t+\delta]$. Clearly

$\epsilon > 0$.

Suppose there exists $\{r_n\} \to \infty$ such that $(\bar{Q}^{r_n}(t), \bar{A}^{r_n}(t)) \to (\bar{Q}(t), \bar{A}(t))$ as $r_n \to \infty$. This would imply that $\bar{Q}_2^{r_n}(s) - m_2\bar{Q}_1^{r_n}(s) \to \bar{Q}_2(s) - \bar{m}_2 Q_1(s)$ uniformly on compact sets. Hence, by the definition of convergence u.o.c., we would be ensured of the existence of an $N > 0$ such that

$$\sup_{s\in[t-\delta,t+\delta]} \left|\bar{Q}_2^{r_n}(s) - m_2\bar{Q}_1^{r_n}(s) - \left(\bar{Q}_2(s) - m_2\bar{Q}_1(s)\right)\right| \leq \frac{\epsilon}{2}$$

for $n > N$.

Therefore, $\bar{Q}_2^{r_n}(s) - m_2\bar{Q}_1^{r_n}(s) - b_2 \geq \frac{\epsilon}{2}$ for $n \geq N, s \in [t-\delta, t+\delta]$;
i.e., $\frac{1}{r_n}Q_2(r_n s) \geq \frac{m_2}{r_n}Q_1(r_n s) + b_2 + \frac{\epsilon}{2}$;
i.e., $Q_2(s) \geq m_2 Q_1(s) + b_2 r_n + \frac{\epsilon}{2}r_n$ for $s \in [r_n(t-\delta), r_n(t+\delta)]$.
Given the above inequality, our boundary avoiding policy dictates that there are no discretionary customer arrivals to queue 2 in this scenario. That is, $\forall s \in [t-\delta, t+\delta]$, we obtain $A_2(r_n s) - A_2(r_n(t-\delta)) = E_2(r_n s) - E_2(r_n(t-\delta))$. By multiplying each side by $\frac{1}{r_n}$ and letting r_n go to infinity, we can rewrite this as $\bar{A}_2(s) - \bar{A}_2(t-\delta) = \lambda_2(s-t+\delta)$. Hence, $\dot{\bar{A}}_2(t) = \lambda_2$ as desired.

In the case when $\bar{Q}_2(t) < m_1\bar{Q}_1(t) + b_1$, we analogously obtain that $\dot{\bar{A}}_1(t) = \lambda_1$.

□

To complete the proof of Theorem 1 (stated in Section 1.2), we will need the following lemma, which is presented as Lemma 2.2 in Dai and Weiss [5].

Lemma 1 *Let g be an absolutely continuous nonnegative function, and let $\dot{g}$ denote its derivative, whenever it exists. If $g(t) = 0$ and $\dot{g}(t)$ exists, then $\dot{g}(t) = 0$.*

Proof of Theorem 1

By Proposition 4 and the discussion that precedes it, we have shown that our fluid model is properly defined. Accordingly, we must now show that all fluid limits of the model eventually drain to zero. That is, we must consider the drainage in the different regions determined by the relationship between $Q_1(t)$ and $Q_2(t)$. Define $\rho = \frac{\lambda}{\mu}$, and let $\rho_{\max}$ and ρ_i for $i \in \{1, 2\}$ remain as previously defined. Define the piecewise-linear Lyapunov function $f(t) = Q_1(t) + Q_2(t)$. Once again, our goal is to show that having $f(t) > 0$ implies $\dot{f}(t) \leq -\epsilon$ for some $\epsilon > 0$, which by Lemma 3 of [4] would yield our desired result that $f(t) = 0$ for all $t \geq f(0)/\epsilon$. Again, for any such system, because of the non-idling constraint, we are guaranteed that both servers will be working anytime we are in the interior. Also, as usual, we assume that external arrivals are independent of the system state for all $t \geq 0$. As a result, for $t > 0$ such that both $Q_1(t)$ and $Q_2(t)$ are nonzero, we obtain the condition $\dot{f}(t) = \dot{Q}_1(t) + \dot{Q}_2(t) = \lambda - \mu < 0$. This of course holds iff $\rho < 1$.

Now suppose that $f(t) > 0$ and $Q_1(t) = 0$. Then by Lemma 1 we obtain $\dot{Q}_1(t) = 0$, and hence $\dot{f}(t) = \dot{Q}_2(t)$. Also, because of the wedge, we are ensured that the discretionary customers will be routed away from the boundary. As a result we obtain the condition $\dot{f}(t) = \dot{Q}_2(t) = \lambda_2 - \mu_2 < 0$. This of course holds iff $\rho_2 < 1$. Similarly, having $f(t) > 0$ and $Q_2(t) = 0$ would imply the condition that $\rho_1 < 1$. Hence we have shown stability is obtained in all regions of the state space if $\rho_{\max} < 1$. Again, recall that according to Theorem 4.2 of [3], stability of the fluid model implies that the corresponding queueing network is positive Harris recurrent. Thus we have proved Theorem 1 as stated in Section 1.2.

□

This generalized system clearly encompasses our particular systems of interest. For the JSEW system, given $\mu_1 < \mu_2$, the required wedges would be obtained by

setting $m_1 = 1$ and $m_2 = \mu_2/\mu_1$. For JSQ, we would simply have $m_1 = m_2 = 1$.

The need for the wedges becomes more evident if we consider unusual queueing disciplines such as JLQ (Join the Longest Queue). This policy is not boundary-avoiding, and as a result would require additional constraints to ensure proper drainage of the fluid model on either axis.

In Chapter 5, we will consider a variation of the aforementioned JSQ network which includes an additional (moveable) server. This policy will operate such that the third server will not be dedicated to either queue, but will instead serve the longer of the two queues. In the case of a tie, the service rate of the additional server will be split equally between the two nodes. Note that this policy may change the service rate at either node for a given period of time, but it does not change the service capacity. That is, each queue will still only be able to serve one customer at a time but may have added resources to do so. To extend our stability results to this system, denote the service rate of this discretionary server as μ_L, and define $\rho_{*1} = \frac{\lambda_1}{\mu_1+\mu_L}$ and $\rho_{*2} = \frac{\lambda_2}{\mu_2+\mu_L}$. We would like to show that stability is obtained in all regions of the state space if all three traffic intensity parameters (ρ, ρ_{*1}, and ρ_{*2}) are less than 1.

To show that all fluid limits of this model eventually drain to zero, we employ the same steps used in the proof of Theorem 1 to describe the system's behavior in the interior and on both axes. Again the policy is assumed to be non-idling, and thus we are guaranteed all three servers will be working anytime we are in the interior. As a result, for all $t > 0$ such that both $Q_1(t)$ and $Q_2(t)$ are nonzero, using the same piecewise-linear Lyapunov function as in the above proof, we easily obtain the condition that $\dot{f}(t) = \dot{Q}_1(t) + \dot{Q}_2(t) = \lambda - \mu < 0$. As expected, this will hold iff $\rho < 1$.

Now suppose that $f(t) > 0$ and $Q_1(t) = 0$. Using Lemma 1 again, we obtain $\dot{f}(t) = \dot{Q}_2(t)$. Given the JSQ routing policy, we are ensured that the discretionary

customers will be routed away from the boundary, while the SLQ service policy ensures that the movable server will be obligated to queue 2. As a result we obtain the condition that $\dot{f}(t) = \dot{Q}_2(t) = \lambda_2 - \mu_2 - \mu_L < 0$. This of course holds iff $\rho_{*2} < 1$. Similarly, having $f(t) > 0$ and $Q_2(t) = 0$ would require $\rho_{*1} < 1$, and we are done.

CHAPTER III

OUTLINE OF APPROACH TO ASYMPTOTICS

As mentioned earlier, this paper will follow the methodology developed in [10] and extended in [6] for analyzing the exact asymptotics of a queueing network. To facilitate reading, we first give an outline of the approach suitable for discrete-state queueing systems.

A general overview of the analysis can be given as follows. We start with the continuous-time Markov process that corresponds to our network of interest, denoted by $\mathbf{Q}(\cdot)$, and uniformize it to allow us to work in a discrete time setting. From this discrete time Markov chain, we will derive three other Markov chains that will be used during the analysis. The purpose of these transformations is to provide us with a setting that is more favorable in performing calculations and analysis. Using the final Markov chain, we will derive our desired results; and finally interpret these results for our original $\mathbf{Q}(\cdot)$ process.

In the following six steps, we will detail the construction of the four different Markov chains that will be used in our exact asymptotics analysis. In the subsequent chapter we will present the analysis as it pertains to the JSEW network described in Section 1.1. Note that throughout this dissertation, we have attempted to remain consistent with our usage of the terms "Markov process" (a continuous-time process)

and "Markov chain" (a discrete-time process). However, in adopting the terminology from [6] and [10], we will occasionally abuse this notation. For example, the final chain that we will construct is often called a "twisted free process", despite being a Markov chain by the above definition.

Step 1. We start with the continuous-time, discrete state Markov process $\mathbf{Q}(t)$ which describes the state of the queueing system. We assume that the elements on the diagonal of the generator of the Markov process are bounded. Since we prefer to work in discrete-time, we uniformize the process. Before uniformizing, we measure time in units such that the maximum diagonal element of the generator is equal to one. After the uniformization, the state of the discrete-time Markov chain at time n is denoted by $\mathbf{Q}[n] = (Q_1[n], Q_2[n]), n \in Z_+ \equiv \{0, 1, 2, \ldots\}$. Also, the Markov chain must be irreducible and positive recurrent on a countable state space $S_Q \subset \mathbb{Z}_+^2$.

Step 2. Make a guess as to the direction taken when the total number of customers reaches some large level ℓ. If this guess is wrong, it will be impossible to carry out all of the remaining steps. Similarly, if this guess is correct but does not fit into the framework of [10] (i.e., the approach to ℓ is not along a line), it will again be impossible to carry out the remaining steps. The remaining steps must be completed for each of the conjectured approaches.

Step 3. We next construct the first of three Markov chains related to $\mathbf{Q}[n]$. This chain, which we will call the W-chain, is simply a relabelling of $\mathbf{Q}[n]$ which incorporates some of our conjectured behavior. We transform $\mathbf{Q}[n]$ into an equivalent Markov chain $W \equiv \{W[n], n = 0, 1, \ldots\}$ with state space $S \equiv Z_+ \times Z^{r-1} \times \hat{S}$, where $\hat{S}$ is some countable set. For each n, $W[n] = (\tilde{W}[n], \hat{W}[n])$ where $\tilde{W}[n] \in Z^r$, $\tilde{W}_1[n] \geq 0$, and $\hat{W}[n] \in \hat{S}$. Hence, if $x \in S$, then $x = (\tilde{x}, \hat{x})$ with $\tilde{x}_1 \geq 0$. Let $F_\ell \equiv \{x \in S | \tilde{x}_1 \geq \ell\}$ for $\ell \in Z_+$. The W-chain shall be constructed such that each component of $\tilde{W}$ is

expected to diverge along the approach to F_ℓ while $\hat{W}$, the stable components, are expected to have a stationary distribution along the approach to F_ℓ. The transition matrix of the W-chain will be denoted by K, and its stationary distribution by π_W.

Step 4. The next step is to construct a Markov chain W^∞, which will be called the "free process" and is closely related to the W-chain. The construction of the free process also depends on our guess as to the most likely approach to F_ℓ. To construct the free process, we need the W-chain to have the following Markov additive transition structure along the large deviation path: There exists a set $\blacktriangle$ such that for $x, y \in S \backslash \blacktriangle$,

$$\begin{aligned} K^\infty((\tilde{x},\hat{x});(\tilde{y},\hat{y})) &\equiv \mathrm{P}(\tilde{W}^\infty[n+1]=\tilde{y}, \hat{W}^\infty[n+1]=\hat{y} \mid \tilde{W}^\infty[n]=\tilde{x}, \hat{W}^\infty[n]=\hat{x}) \\ &= \mathrm{P}^\infty(\hat{x};(\tilde{y}-\tilde{x},\hat{y})), \end{aligned}$$

where for each $\hat{x}, \mathrm{P}^\infty(\hat{x},(\cdot,\cdot))$ must be a probability measure. More will be said of this requirement within the description of Condition 1 below. The free process has state space $S^\infty \equiv Z^r \times \hat{S}$ and transition kernel K^∞. This chain is *free* in the sense that the state space is now enlarged, as the additive components are unconstrained in $\mathbb{R}^r$. The edge of the set $\blacktriangle$ will be denoted by $\Delta \equiv \blacktriangle \cap S$, and referred to as the boundary. Lastly we define $F_\ell^\infty \equiv \{x \in S^\infty | \tilde{x}_1 \geq \ell\}$ for $\ell \in Z_+$.

Step 5. We must now find a harmonic function $h(\cdot)$ of the form $h(x) = \alpha^{\tilde{x}_1}\hat{h}(\hat{x})$ for $x = (\tilde{x},\hat{x}) \in \mathbb{Z} \times \hat{S}$ for the free process, and then use this function to construct the "twisted free process" (see Condition 5 below). The discussion in Section 2.1 of [10] provides some solace by showing that such a function exists under quite general conditions; however, there is little guidance in finding such a function. Often in applications, it is possible to make an educated guess for the value of α.

Step 6. Now that we have successfully constructed the required backdrop for our analysis, the next step is to verify that our processes satisfy technical conditions 1 through 12, as enumerated in [6]. Once these conditions are met, we can then move

on to specialize the results of [6] and [10] to our particular queueing system. The conditions that we impose on our constructed Markov chains are as follows:

C.1: In step 4, we stated the condition that W^∞ must be a Markov additive process with

$$K^\infty\big((\tilde{x},\hat{x});(\tilde{y},\hat{y})\big) = P^\infty(\hat{x};\big(\tilde{y}-\tilde{x},\hat{y}\big)) \tag{19}$$

where for each $\hat{x}$, $P^\infty(\hat{x},(\cdot,\cdot))$ is a probability measure. However, it is worthy to note that the following weaker condition is typically equally sufficient: The marginal process $(\tilde{W}_1^\infty[n],\hat{W}^\infty[n])$ is a Markov additive process with

$$\begin{aligned}&\mathrm{P}(\tilde{W}_1^\infty[n+1]-\tilde{W}_1^\infty[n]=\tilde{x}_1,\hat{W}^\infty[n+1]=\hat{y}\mid W^\infty[n])\\ &\qquad=\mathrm{P}(\tilde{W}_1^\infty[n+1]-\tilde{W}_1^\infty[n]=\tilde{x}_1,\hat{W}^\infty[n+1]=\hat{y}\mid \hat{W}^\infty[n]).\end{aligned} \tag{20}$$

We let $\mathcal{J}$ denote the transition kernel of $(\tilde{W}_1^\infty[n],\hat{W}^\infty[n])$, and $\hat{K}^\infty$ the transition kernel of $\hat{W}^\infty[n]$. In most applications, the stronger condition holds.

C.2: The transition probabilities of W and W^∞ agree between states in the interior, i.e.,

$$K^\infty(x,y) = K(x,y) \text{ for } x \text{ and } y \text{ in } \Theta, \tag{21}$$

where $\Theta \equiv S\backslash\,\Delta$.

C.3: The transition probabilities of W and W^∞ agree from states on the boundary to states in the interior, i.e.,

$$K^\infty(x,y) = K(x,y) \text{ for } x \in \blacktriangle \text{ and } y \text{ in } \Theta. \tag{22}$$

C.4: The function h is a positive function on S^∞ of the form $h(x) = \alpha^{\tilde{x}_1}\hat{h}(\hat{x})$ with $\alpha > 1$, and h is harmonic for the free process; i.e., $K^\infty h = h$.

C.5: The process $\mathcal{W}^\infty$ is the twist (i.e., h-transform) of the free process W^∞. That is,

$$\mathcal{K}^\infty(x,y) = K^\infty(x,y)h(y)/h(x). \tag{23}$$

C.6: The marginal Markov chain $\{\hat{\mathcal{W}}^\infty[n], n \in \mathbb{Z}_+\}$, with transition kernel $\hat{\mathcal{K}}^\infty$ has a stationary probability distribution $\varphi(\cdot)$.

C.7: The first coordinate of the drift vector of the stationary version of $\mathcal{W}^\infty$ has a finite, strictly positive drift. That is, $0 < \tilde{d}_1 < \infty$ where

$$\tilde{d} = \sum_{\hat{x}\in\hat{S}} \varphi(\hat{x}) E[\tilde{\mathcal{W}}^\infty[1] \mid \mathcal{W}^\infty[0] = (0,\hat{x})].$$

C.8: The twisted free process starting from Δ has a positive probability of never hitting $\blacktriangle$; i.e.,

$$\sum_{x\in\Delta} \pi(x)H(x) > 0,$$

where $H(x) = P\{\mathcal{T}_{\blacktriangle}^\infty = \infty\}$ and $\mathcal{T}_{\blacktriangle}^\infty$ is the first return time to $\blacktriangle$ by $\mathcal{W}^\infty$.

C.9:

$$\sum_{\hat{x}\in\hat{S}} \varphi(\hat{x})\hat{h}(\hat{x}) < \infty.$$

C.10:

$$\sum_{x\in\Delta} \pi(x)h(x)\chi\{K(x,\Theta) > 0\} < \infty.$$

where χ is the indicator function, and $K(x,\Theta)$ is the probability that the free process jumps into the interior from x.

C.11: Let $Y_\ell(\hat{y}) \equiv \{x \in S^\infty | \tilde{x}_1 = \ell, \hat{x} = \hat{y}\}$. For each $\hat{y}$, there is an associated integer $L(\hat{y})$ such that $Y_\ell(\hat{y}) \cap \Delta = \emptyset$ if $\ell \geq L(\hat{y})$.

C.12: Either $\hat{h}$ is bounded or there must exist a function $\hat{V} : \hat{S} \to [1, \infty)$, a finite set $C \subset \hat{S}$ and a constant $b < \infty$ such that

$$\sum_{\hat{y} \in \hat{S}} \hat{\mathcal{K}}^{\infty}(\hat{x}, \hat{y})(\hat{V}(\hat{y}) - \hat{V}(\hat{x})) \leq -\hat{h}(\hat{x}) + b\chi\{\hat{x} \in C\}, \tag{24}$$

which is Condition (V3) of [13]. Moreover it must be the case that $\sum_{\hat{y}} \hat{V}(\hat{y})\hat{\lambda}(\hat{y}) < \infty$, where $\lambda(x) \equiv \pi(x)h(x)\chi\{x \in \Delta\}$ and $\hat{\lambda}(\hat{y})$ is the marginal measure $\sum_{\tilde{y}} \lambda(\tilde{y}, \hat{y})$.

Once these conditions have been satisfied, we can then apply Theorems 5-7 of [6] to obtain asymptotic expressions for the steady state distribution, mean hitting time, and hitting distribution of our particular network.

CHAPTER IV

ANALYSIS OF THE JSEW NETWORK

In this chapter, we focus on applying the methodology outlined in Chapter 3 to our JSEW queueing network. We delay the discussion of the implied asymptotic results to Chapter 5.

We expect that the approach to the rare event of interest will depend on the relationship between the traffic intensity parameters ρ_1, ρ_2, and ρ. Given the three types of customer arrivals, it is natural to conjecture that we will have at least two distinct ways that the system will behave as the number of customers in the system grows towards some large level ℓ. If $\rho_1 > \max\{\rho_2, \rho\}$, then it is logical to expect that the server at queue 2 will take all of the discretionary customers and keep its queue small while the dedicated customers at the other queue will cause the system to overload. Analogously, having $\rho_2 > \max\{\rho_1, \rho\}$ would cause queue 2 to be the primary cause of the system overload while the number of customers at queue 1 remained relatively small. These anticipated approaches will be called the unpooled case.

Suppose, however, that $\rho > \max\{\rho_1, \rho_2\}$. In this case, we would expect our system to exhibit resource pooling. That is, we anticipate that the workload will be spread over the entire network with both queues growing large as the system overloads. If

the proportion of discretionary customers is large enough, then we would anticipate that the expected waiting times of the two queues will remain relatively close to each other along the approach to our rare event. This case will be called the pooled case. In the section that follows, the conditions necessary for pooling to occur will be made rigorous, and the proofs for our asymptotic results in the pooled case will be completed. In Section 4.2, we will provide the proof for the unpooled case.

4.1 Pooled Case

For this section, we will complete the steps outlined in Chapter 3 for the case when

$$\rho > \max\{\rho_1, \rho_2\}. \tag{25}$$

Note that when $\mu_1 = \mu_2$ our JSEW network reduces to a system that simply employs the JSQ policy for the discretionary customers. Since McDonald's exact asymptotics methodology has already been applied to this system in [6], this case will not be considered within our work. To ensure that our problem is nontrivial, we will assume that $\lambda_E > 0$. Without loss of generality, we will also assume throughout this section that $\mu_2 > \mu_1$.

Step 1. We begin with the Markov process associated with the joint queue length process, $(Q_1[t], Q_2[t])$, as shown in Figure 1. The diagonal line in the figure, hereafter referred to as the JSEW line, corresponds to the set where the expected waiting times are equal for the two queues. To ensure irreducibility, we assume that $\mu_i > 0$, for $i = 1, 2$. Our assumption that $\lambda + \mu < \infty$ implies that the elements on the diagonal of the generator are bounded. This assumption permits us to rescale time such that $\lambda + \mu = 1$. With this rescaling, our system parameters can also be conveniently interpreted as transition probabilities, as shown in Figure 1. As given in Theorem 1, the requirement of positive recurrence is met by assuming $\rho_{\max} < 1$. After the uniformization, the

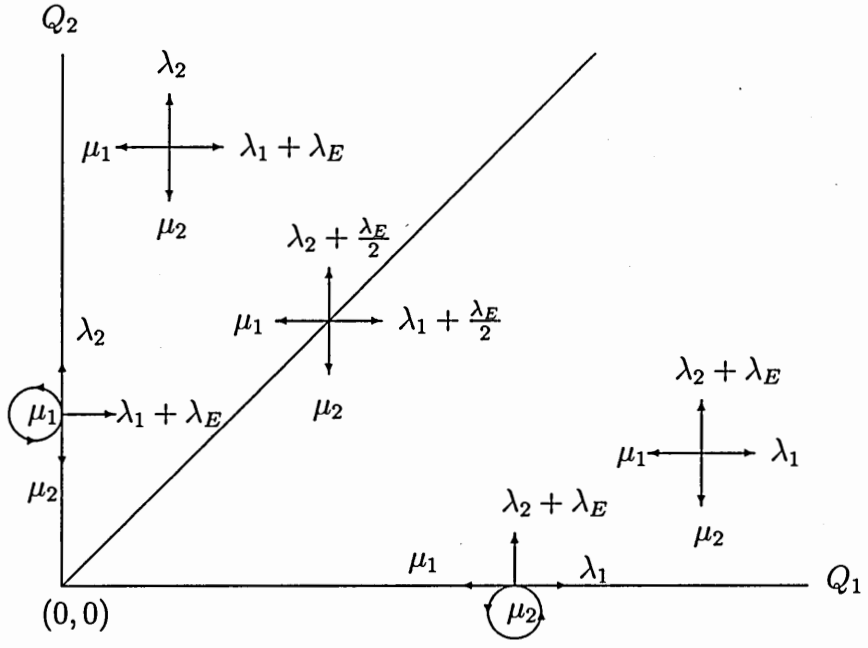

Figure 1: Transition probabilities for the JSEW system.

state of the Q-chain at time n is denoted by $\mathbf{Q}[n] = (Q_1[n], Q_2[n]), n \in Z_+$. Let K_Q and π_Q denote the corresponding transition kernel and stationary probability measure on S_Q. This Markov chain will be referred to as the Q-chain.

Step 2. Step 1 of the algorithm is identical for the pooled and unpooled cases. From there, the methodology and results differ depending on the anticipated system behavior as it grows towards having some large number of customers, ℓ. Again, if (25) holds, then it is natural to conjecture that the two servers would pool. In this section we consider the case when the difference between the expected waiting times of the two queues remains relatively small as the number of customers approaches ℓ. In other words, we will proceed under the presumption that the approach will be along the JSEW line.

Step 3. For our JSEW pooled case, the $\tilde{W}$ process has only one component. To have $\tilde{W}$ diverge along the approach to F_ℓ^∞ we define $\tilde{W}[n] \equiv Q_1[n] + Q_2[n]$. Hence, we

have $\tilde{W}[n] \geq \ell$ iff $(Q_1[n], Q_2[n]) \in F_\ell^\infty$. For the stable component, we define $\hat{W}$ as the difference between the expected waiting times of the two queues. Thus we have $\mathbf{W}[n] = (\tilde{W}[n], \hat{W}[n]) = (Q_1[n] + Q_2[n], \frac{1}{\mu_2}Q_2[n] - \frac{1}{\mu_1}Q_1[n])$. The state space for our W process is $S \subset \tilde{S} \times \hat{S} \equiv \{(x+y, \frac{y}{\mu_2} - \frac{x}{\mu_1}) \mid x, y \in \mathbb{Z}_+\} \subset \mathbb{Z}_+ \times \mathbb{R}$.

Note that, given any state in the W process, we can easily recover the corresponding state in the original Q process. Moreover, it is easy to see that we have achieved the desired one-to-one correspondence between the two. Note that there are numerous alternative ways that we could have defined the components of W. For example, we could have assigned $\tilde{W}$ to $Q_1[n] + Q_2[n]$ and left $\hat{W}$ to be degenerate. However, as discussed in [6], our initial choice for W is better in the sense that the higher the dimension we have in $\hat{W}$, the stronger the attained results once all of the conditions are verified.

Step 4. We now partition the state space S into two parts that we will call the interior and the boundary of S, denoted by Θ and Δ, respectively. In order to successfully construct the free process, our partition must satisfy the following Markov additive property. Given that we are at time n, the transition structure that determines the next stable component and the change in the additive component from time n to $n+1$ should depend only on $\hat{W}[n]$. To obtain our desired Markov additive structure, we use the transition probability structure of W corresponding to when both queues are busy. Hence, we define $\Delta \equiv \{(x,y) \in S_Q \mid x = 0 \text{ or } y = 0\}$. Figure 2 explicitly illustrates the transition structure of the W^∞ process. The ▲'s in the diagram represent the states added in the construction of the free process. More formally, we define $\blacktriangle \equiv (\mathbb{Z} \times \hat{S}) \setminus S$. As shown in Figure 2, our free process is constructed by using the transition probabilities of W corresponding to the states in $S \setminus \Delta$. By removing Δ as a drift barrier, we have enlarged the state space to $\mathbb{Z} \times \hat{S}$ and

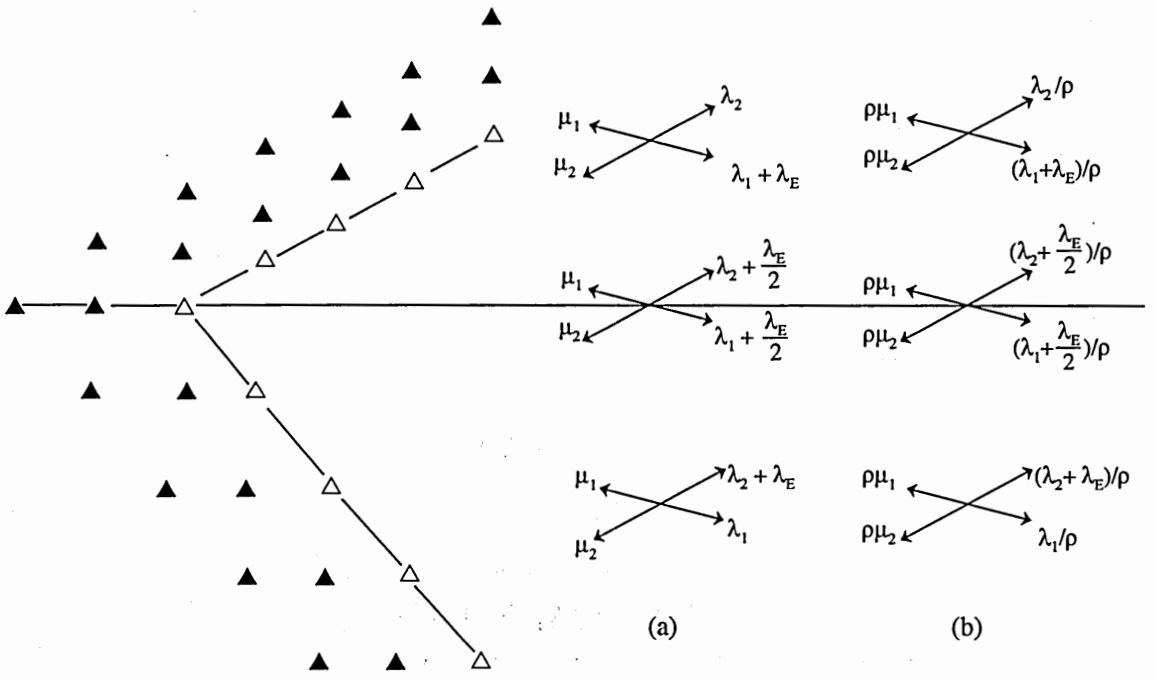

Figure 2: The transitions of (a) K^{∞} and (b) $\mathcal{K}^{\infty}$ under pooling.

for the free process we use the transition probabilities associated with P^{∞} everywhere.

Intuitively our choice of Δ seems justifiable since, in this case, we expect the approach to F_{ℓ}^{∞} to remain close to the JSEW line and spend little time in $\blacktriangle$. Of course, whether our choice of boundary is correct will be determined by whether we can satisfy all of the remaining technical conditions outlined in the previous chapter.

Step 5. A harmonic function that works is $h(x) = \rho^{-\tilde{x}}$ where $\tilde{x}$ is the total number of customers queued. Note that, in this case, $\alpha \equiv \rho^{-1}$, and $\hat{h}(\hat{x}) \equiv 1$. Using the transition probabilities of the free process that correspond to a positive difference in the expected waiting times (i.e., above the $\hat{x}$-axis), we obtain

$$\begin{aligned} E[h(W^{\infty}[1])|W^{\infty}[0] = (\tilde{x}, \hat{x})] &= (\lambda_1 + \lambda_2 + \lambda_E)\rho^{-(\tilde{x}+1)} + (\mu_1 + \mu_2)\rho^{-(\tilde{x}-1)} \\ &= \rho^{-\tilde{x}}[(\lambda_1 + \lambda_2 + \lambda_E)\rho^{-1} + (\mu_1 + \mu_2)\rho] \\ &= h(\tilde{x}, \hat{x}). \end{aligned}$$

Equality can similarly be shown to hold for the transition probabilities on and below the $\hat{x}$-axis of the free process. Hence, our choice of a harmonic function is valid since $K^{\infty}h = h$. Using this harmonic function, we then define our twisted free process $\mathcal{W}^{\infty} \equiv \{\mathcal{W}^{\infty}[n], n = 0, 1, 2, \ldots\}$, where $\mathcal{W}^{\infty}[n] = (\tilde{\mathcal{W}}^{\infty}[n], \hat{\mathcal{W}}^{\infty}[n])$. As in [6] and [10], we will be using caligraphic letters to correspond to the twisted free process. Recall that K^{∞} denotes the transition matrix for the free process. Given the harmonic function $h > 0$ for K^{∞}, the transition matrix $\mathcal{K}^{\infty}$ of the twisted free process has transition probabilities

$$\mathcal{K}^{\infty}(x, y) = K^{\infty}(x, y)h(y)/h(x) \text{ for } x, y \in \mathbb{Z} \times \hat{S}.$$

The resulting transition probabilities for the twisted free process are shown in Figure 2.

Step 6. While constructing our Markov chains, some of Conditions 1-12 were already taken into consideration and hence have been satisfied beforehand. For example, by construction, we have already ensured that our free process is a Markov additive process with $K_W((\tilde{x}, \hat{x}); (\tilde{y}, \hat{y})) = P^{\infty}(\hat{x}; (\tilde{y} - \tilde{x}, \hat{y}))$, and that the transition probabilities of W^{∞} and W agree for jumps between states in the interior. Hence, C1 and C2 have been met. Condition 3 is also satisfied since for any $x \in \blacktriangle$ and $y \in \Theta$, we have $K(x, y) = K^{\infty}(x, y)$.

In step 5 of the algorithm, our positive harmonic function has already been obtained.

Hence, C4 has been satisfied. Using this same harmonic function to construct the twisted free process $\mathcal{K}^{\infty}(x,y)$ from the free process leads to C5 also being satisfied.

Condition 6 requires that the marginal chain $\{\hat{\mathcal{W}}^{\infty}[n], n \in \mathbb{Z}_+\}$, with transition kernel $\hat{\mathcal{K}}^{\infty}$ has a stationary probability distribution $\varphi(\cdot)$. To verify this, we need to show that $\hat{\mathcal{W}}^{\infty}$ is positive recurrent. That is, we need the difference between the expected waiting times of the two queues to drift towards zero. This translates into the following conditions for the pooled case. For $\hat{x} > 0$, we obtain

$$\begin{aligned}
&E\left[\hat{\mathcal{W}}^{\infty}[n+1] - \hat{\mathcal{W}}^{\infty}[n] \mid \hat{\mathcal{W}}^{\infty}[n] = i\right] \\
&\qquad = E\left[\hat{\mathcal{W}}^{\infty}[n+1] \mid \hat{\mathcal{W}}^{\infty}[n] = i\right] - i \\
&\qquad = \textstyle\sum_{j \in S} j\mathcal{K}^{\infty}(i,j)^{\infty} - i \\
&\qquad = \rho\mu_2(i + \tfrac{1}{\mu_2}) + \tfrac{\lambda_1}{\rho}(i + \tfrac{1}{\mu_1}) + \rho\mu_1(i - \tfrac{1}{\mu_1}) + \tfrac{\lambda_2+\lambda_E}{\rho}(i - \tfrac{1}{\mu_2}) - i \\
&\qquad = i\left[\rho(\mu_1 + \mu_2) + \tfrac{1}{\rho}(\lambda_1 + \lambda_2 + \lambda_E)\right] + \rho + \tfrac{1}{\rho}(\tfrac{\lambda_1}{\mu_1}) - \rho - \tfrac{1}{\rho}(\tfrac{\lambda_2+\lambda_E}{\mu_2}) - i \\
&\qquad = \tfrac{1}{\rho}(\tfrac{\lambda_1}{\mu_1}) - \tfrac{1}{\rho}(\tfrac{\lambda_2+\lambda_E}{\mu_2}) \\
&\qquad = \tfrac{1}{\rho}(\tfrac{\lambda_1}{\mu_1} - \tfrac{\lambda_2+\lambda_E}{\mu_2}).
\end{aligned}$$

For a drift towards zero, this last equality implies that we must have $\frac{\lambda_1}{\mu_1} < \frac{\lambda_2+\lambda_E}{\mu_2}$ for the pooled case. Note that ρ can be rewritten as the following weighted sum:

$$\rho = \frac{\mu_1}{\mu_1+\mu_2}\frac{\lambda_1}{\mu_1} + \frac{\mu_2}{\mu_1+\mu_2}\frac{\lambda_2+\lambda_E}{\mu_2} \tag{26}$$

which is of the form $c\rho_1 + (1-c)\rho_{2\oplus\lambda}$, where $0 < c < 1$ and $\rho_{2\oplus\lambda} = \frac{\lambda_2+\lambda_E}{\mu_2}$. Therefore it must be the case that $\rho_{2\oplus\lambda} > \rho > \rho_1$, where the second inequality is provided by our pooled case assumption. Similarly, for $\hat{x} < 0$,

$$
\begin{aligned}
E\left[\hat{\mathcal{W}}^{\infty}[n+1]-\hat{\mathcal{W}}^{\infty}[n] \mid \hat{\mathcal{W}}^{\infty}[n]=i\right] & \\
&= E\left[\hat{\mathcal{W}}^{\infty}[n+1]) \mid \hat{\mathcal{W}}^{\infty}[n]=i\right]-i \\
&= \textstyle\sum_{j\in S} j\mathcal{K}^{\infty}(i,j)^{\infty}-i \\
&= \rho\mu_2(i+\tfrac{1}{\mu_2})+\tfrac{\lambda_1+\lambda_E}{\rho}(i+\tfrac{1}{\mu_1})+\rho\mu_1(i-\tfrac{1}{\mu_1})+\tfrac{\lambda_2}{\rho}(i-\tfrac{1}{\mu_2})-i \\
&= i\left[\rho(\mu_1+\mu_2)+\tfrac{1}{\rho}(\lambda_1+\lambda_2+\lambda_E)\right]+\rho+\tfrac{1}{\rho}(\tfrac{\lambda_1+\lambda_E}{\mu_1})-\rho-\tfrac{1}{\rho}(\tfrac{\lambda_2}{\mu_2})-i \\
&= \tfrac{1}{\rho}(\tfrac{\lambda_1+\lambda_E}{\mu_1})-\tfrac{1}{\rho}(\tfrac{\lambda_2}{\mu_2}) \\
&= \tfrac{1}{\rho}(\tfrac{\lambda_1+\lambda_E}{\mu_1}-\tfrac{\lambda_2}{\mu_2})
\end{aligned}
$$

which gives us the additional condition that $\frac{\lambda_1+\lambda_E}{\mu_1} > \frac{\lambda_2}{\mu_2}$. Using an argument analogous to the one above, we readily obtain $\rho_{1\oplus\lambda} > \rho > \rho_1$. Hence, given our pooling assumption, this condition is also satisfied. Thus it is worthy to note that while Condition 6 led to a dichotomy of the pooled case (into strong and weak pooling) when analyzing the JSQ system in Foley and McDonald [6], such a dichotomy does not exist in the JSEW case. In Chapter 5, we will say more about the pooling conditions of both systems as they relate to the qualitative behavior of the networks.

With Condition 6 satisfied we are assured of the existence of a stationary probability distribution $\varphi(\cdot)$ for $\hat{\mathcal{W}}^{\infty}[n]$. This distribution is defined on the additive subgroup $\{m/\mu_1 + n/\mu_2\}$ for $m, n \in \mathbb{Z}$, and can be obtained through fast simulation.

The next condition requires that the mean increment of $\tilde{\mathcal{W}}^{\infty}$ has a finite, strictly positive drift, $\tilde{d}$. Observe that for all $\hat{x} \in \hat{S}$,

$$
\begin{aligned}
\tilde{d} &= \rho^{-1}(\lambda_1+\lambda_2+\lambda_E)+\rho(\mu_1+\mu_2) \\
&= \mu_1+\mu_2-(\lambda_1+\lambda_2+\lambda_E)
\end{aligned}
$$

and recall our assumption that $\rho \equiv \frac{\lambda_1+\lambda_2+\lambda_E}{\mu_1+\mu_2} < 1$. Thus, we have ensured a positive

$\tilde{d}$ and Condition 7 has been satisfied.

Before we consider the remaining conditions, we first make the following observation. Recall that $Y_\ell(\hat{y}) \equiv \{x \in S^\infty | \tilde{x}_1 = \ell, \hat{x} = \hat{y}\}$. By Lemma 5 in [6], the expected number of visits to $Y_0(\hat{y})$ by $\mathcal{W}^\infty$ is $\varphi(\hat{y})/\tilde{d}$. Hence, by stationarity, the expected total number of hits to Δ is given by $\sum_{\hat{y}\in\hat{S}} \varphi(\hat{y})/\tilde{d} = 1/\tilde{d}$. Thus, we have the following theorem.

Theorem 5 *The twisted free process has a finite expected number of hits on Δ. That is,*

$$E_{z\in\Delta}\left[\sum_{m=0}^{\infty} \chi_{\{\mathcal{W}^\infty[m]\in\Delta\}}\right] < \infty.$$

Note that from the existence of φ and the fact that $\tilde{d}$ is strictly positive, the law of large numbers (LLN) implies that any visit to $\blacktriangle$ will eventually lead to a revisit to Δ. Moreover, suppose that for all initial states $z \in \Delta$, $\mathcal{W}^\infty$ drifts towards $\blacktriangle$ with probability one. The LLN would therefore imply that the twisted free process hits Δ infinitely many times, which contradicts Theorem 5. Hence with positive probability $\mathcal{W}^\infty$ wanders away from Δ without returning to $\blacktriangle$, and Condition 8 is therefore satisfied. Condition 9 is also automatic since $\hat{h} \equiv 1$ and $\hat{\mathcal{W}}^\infty$ is positive recurrent.

When written in terms of the Q-chain coordinate system, Condition 10 translates to showing that

$$\sum_{x>0} \pi_Q(x,0)\rho^{-x} + \sum_{y>0} \pi_Q(0,y)\rho^{-y} < \infty \tag{27}$$

Given our continued assumptions that $\rho < 1$ and $\mu_2 > \mu_1$, as a special case of the following proposition we will obtain (27), which establishes Condition 10 for a large portion of the parameter space.

Proposition 5 *Let* $b \equiv \frac{\mu_2}{\mu_1}$. *If* $\rho > \max\{\rho_1, \rho_2\}$ *and* $\frac{\ln \rho_1}{\ln \rho} > \sqrt{b}$, *then*

$$\sum_{x \geq 0, y \geq 0} \rho^{-\sqrt{bx^2+y^2}} \pi_Q(x,y) < \infty. \tag{28}$$

Note that in proving this proposition, we will have also added the following criterion for our results in the pooled case:

$$\frac{\ln \rho_1}{\ln \rho} > \sqrt{b}. \tag{29}$$

We will say more about the validity of this assumption in Section 5.4.

Proof

To establish (28), consider the Lyapunov function $V(x,y) = \rho^{-\sqrt{bx^2+y^2}}$. We will calculate $K_Q V(x,y) - V(x,y)$ for $x, y \geq 0$ and show

$$K_Q V(x,y) - V(x,y) \leq -cV(x,y) \text{ for } x, y \text{ large enough}$$

where c is some positive constant. By rescaling V it follows that in the region $x, y \geq 0$ we will have found a function V' such that

$$K_Q V'(x,y) - V'(x,y) \leq -\rho^{-\sqrt{bx^2+y^2}} + s(x,y) \text{ where } s \text{ has finite support.}$$

Consequently by Theorem 14.3.7 in Meyn and Tweedie [13],

$$\sum_{x \geq 0, y \geq 0} \rho^{-\sqrt{bx^2+y^2}} \pi_Q(x,y) \leq \sum_{x \geq 0, y \geq 0} s(x,y) \pi_Q(x,y) < \infty.$$

We will divide the proof into five cases. First we will consider the two cases when at least one server is idle, and then move on to analyze the system when the expected waiting times are equal. Lastly, in Cases 4 and 5 we will consider the busy regions

below and above the JSEW line, respectively.

Case 1. Suppose $0 = x < y$. Then

$$\begin{aligned} K_Q V(0,y) - V(0,y) &= (\lambda_1 + \lambda_E)\rho^{-\sqrt{b+y^2}} + \lambda_2 \rho^{-(y+1)} + \mu_1 \rho^{-y} + \mu_2 \rho^{-(y-1)} - \rho^{-y} \\ &= (\lambda_1 + \lambda_E)(\rho^{-\sqrt{b+y^2}} - \rho^{-y}) + \lambda_2(\rho^{-y-1} - \rho^{-y}) + \mu_2(\rho^{-y+1} - \rho^{-y}) \\ &= \rho^{-y}\left((\lambda_1 + \lambda_E)(\rho^{-\sqrt{b+y^2}+y} - 1) + \lambda_2(\rho^{-1} - 1) + \mu_2(\rho - 1)\right). \end{aligned}$$

Note that

$$\begin{aligned} \rho^{-(\sqrt{y^2+b}-y)} - 1 &= \rho^{-y(\sqrt{1+b/y^2}-1)} - 1 \\ &\leq \rho^{-y(1+b/2y^2-1)} - 1 = \rho^{-b/(2y)} - 1. \end{aligned} \tag{30}$$

For the above inequality, we have used $\sqrt{1+x} \leq 1 + x/2$ for $x > -1$. Hence, for y large enough, (30) is arbitrarily small. It therefore suffices to show that $\lambda_2(\rho^{-1} - 1) + \mu_2(\rho - 1) < 0$. This follows by noting that for $s > 0$, the function $\psi_2(s) \equiv \lambda_2 s + \mu_2 s^{-1} - (\lambda_2 + \mu_2)$ is strictly convex and has zeros at 1 and ρ_2^{-1}. Since $1 < \rho^{-1} < \rho_2^{-1}$ in the pooled case, we obtain $\psi_2(\rho^{-1}) < 0$ and thus have completed Case 1.

Case 2. Suppose $0 = y < x$.

In a similar fashion to Case 1, we obtain

$$K_Q V(x,0) - V(x,0) = \rho^{-\sqrt{b}x}\left(\lambda_1(\rho^{-\sqrt{b}} - 1) + (\lambda_2 + \lambda_E)(\rho^{-\sqrt{bx^2+1}+\sqrt{b}x} - 1) + \mu_1(\rho^{\sqrt{b}} - 1)\right).$$

In this case, note that

$$\begin{aligned} \rho^{-(\sqrt{bx^2+1}-\sqrt{b}x)} - 1 &= \rho^{-\sqrt{b}x(\sqrt{1+1/bx^2}-1)} - 1 \\ &\leq \rho^{-\sqrt{b}x(1+1/2bx^2-1)} - 1 \\ &= \rho^{-1/(2x\sqrt{b})} - 1. \end{aligned}$$

Hence, for x large enough, this term is arbitrarily small. Therefore, similar to Case 1, it suffices to show $\lambda_1(\rho^{-\sqrt{b}}-1)+\mu_1(\rho^{\sqrt{b}}-1)<0$. Note that for $s>0$, the function $\psi_1(s)\equiv\lambda_1 s+\mu_1 s^{-1}-(\lambda_1+\mu_1)$ is strictly convex and has zeros at 1 and ρ_1^{-1}. Recall our initial assumption that $\frac{\ln\rho_1}{\ln\rho}>\sqrt{b}$. Moreover, observe that this is equivalent to ensuring that $\rho^{-\sqrt{b}}<\rho_1^{-1}$. Thus, $\psi_1(\rho^{-\sqrt{b}})<0$ and we have completed Case 2.

Case 3. Suppose $0<y=\frac{\mu_2}{\mu_1}x$.

For this case and the two that follow, we find it more convenient to use polar coordinates. First, define $a(\theta)=\sqrt{b\cos^2\theta+\sin^2\theta}$. When θ is obvious, we will simply denote this function by a. Given $x=r\cos\theta$ and $y=r\sin\theta$, observe the following upper bounds.

$$\begin{aligned} V(x+1,y)=\rho^{-\sqrt{b(x+1)^2+y^2}} &= \rho^{\sqrt{a^2r^2+2br\cos\theta+b}} \\ &= \rho^{-ar\sqrt{1+\frac{2b\cos\theta}{a^2r}+\frac{b}{a^2r^2}}} \\ &\le \rho^{-ar\left(1+\frac{b\cos\theta}{a^2r}+\frac{b}{2a^2r^2}\right)} \\ &= \rho^{-ar}\rho^{-\left(\frac{b\cos\theta}{a}+\frac{b}{2ar}\right)} \end{aligned} \tag{31}$$

$$\begin{aligned} V(x,y+1)=\rho^{-\sqrt{bx^2+(y+1)^2}} &= \rho^{\sqrt{a^2r^2+2r\sin\theta+1}} \\ &= \rho^{-ar\sqrt{1+\frac{2\sin\theta}{a^2r}+\frac{1}{a^2r^2}}} \\ &\le \rho^{-ar\left(1+\frac{\sin\theta}{a^2r}+\frac{1}{2a^2r^2}\right)} \\ &= \rho^{-ar}\rho^{-\left(\frac{\sin\theta}{a}+\frac{1}{2ar}\right)} \end{aligned} \tag{32}$$

$$
\begin{aligned}
V(x-1,y) = \rho^{-\sqrt{b(x-1)^2+y^2}} &= \rho^{\sqrt{a^2r^2-2br\cos\theta+b}} \\
&= \rho^{-ar\sqrt{1-\frac{2b\cos\theta}{a^2r}+\frac{b}{a^2r^2}}} \\
&\leq \rho^{-ar\left(1-\frac{b\cos\theta}{a^2r}+\frac{b}{2a^2r^2}\right)} \\
&= \rho^{-ar}\rho^{\left(\frac{b\cos\theta}{a}-\frac{b}{2ar}\right)}
\end{aligned}
\tag{33}
$$

$$
\begin{aligned}
V(x,y-1) = \rho^{-\sqrt{bx^2+(y-1)^2}} &= \rho^{\sqrt{a^2r^2-2r\sin\theta+1}} \\
&= \rho^{-ar\sqrt{1-\frac{2\sin\theta}{a^2r}+\frac{1}{a^2r^2}}} \\
&\leq \rho^{-ar\left(1-\frac{\sin\theta}{a^2r}+\frac{1}{2a^2r^2}\right)} \\
&= \rho^{-ar}\rho^{\left(\frac{\sin\theta}{a}-\frac{1}{2ar}\right)}
\end{aligned}
\tag{34}
$$

where we have repeatedly factored out ar and used $\sqrt{1+x} \leq 1+x/2$ for $x > -1$. In this manner, to justify the inequality in (33), we needed

$$\frac{b}{a^2r^2} - \frac{2b\cos\theta}{a^2r} > -1.$$

By rewriting this condition as $a^2r^2 - 2br\cos\theta + b > 0$, it becomes clear that this inequality is satisfied for r large enough. The inequality in (34) was obtained in a similar manner.

Let $\vartheta \equiv \cos^{-1}(\frac{1}{\sqrt{1+b^2}})$ denote the angle between the x-axis and the JSEW line. Also, for notational purposes define $\alpha \equiv a(\vartheta)$. Using the above upper bounds in the third case when $0 < y = \frac{\mu_2}{\mu_1}x$ yields

$$K_Q V(x,y) - V(x,y) = (\lambda_1 + \frac{\lambda_E}{2})V(x+1,y) + (\lambda_2 + \frac{\lambda_E}{2})V(x,y+1)$$
$$+ \mu_1 V(x-1,y) + \mu_2 V(x,y-1) - V(x,y)$$

$$\leq \rho^{-\alpha r}\left[(\lambda_1 + \frac{\lambda_E}{2})\rho^{-\left(\frac{b\cos\vartheta}{\alpha}+\frac{b}{2\alpha r}\right)} + (\lambda_2 + \frac{\lambda_E}{2})\rho^{-\left(\frac{\sin\vartheta}{\alpha}+\frac{1}{2\alpha r}\right)}\right.$$
$$\left. + \mu_1\rho^{\left(\frac{b\cos\vartheta}{\alpha}-\frac{b}{2\alpha r}\right)} + \mu_2\rho^{\left(\frac{\sin\vartheta}{\alpha}-\frac{1}{2\alpha r}\right)}\right]$$

$$\leq \rho^{-\alpha r}\left[(\lambda_1 + \frac{\lambda_E}{2})\rho^{-\left(\frac{\sin\vartheta}{\alpha}+\frac{b}{2\alpha r}\right)} + (\lambda_2 + \frac{\lambda_E}{2})\rho^{-\left(\frac{\sin\vartheta}{\alpha}+\frac{b}{2\alpha r}\right)}\right.$$
$$\left. + \mu_1\rho^{\left(\frac{\sin\vartheta}{\alpha}-\frac{b}{2\alpha r}\right)} + \mu_2\rho^{\left(\frac{\sin\vartheta}{\alpha}-\frac{b}{2\alpha r}\right)}\right]$$

$$\leq \rho^{-\alpha r}\left[\rho^{-\frac{b}{2\alpha r}}\left[(\lambda_1+\lambda_2+\lambda_E)\rho^{-\frac{\sin\vartheta}{\alpha}} + (\mu_1+\mu_2)\rho^{\frac{\sin\vartheta}{\alpha}}\right] - 1\right].$$

For the second inequality, we have used our assumption that $b > 1$ and the relationship $b\cos\vartheta = \sin\vartheta$. Since $\rho^{-\frac{b}{2\alpha r}} \searrow 1$, it suffices to check that

$$(\lambda_1+\lambda_2+\lambda_E)\rho^{-\frac{\sin\vartheta}{\alpha}} + (\mu_1+\mu_2)\rho^{\frac{\sin\vartheta}{\alpha}} - 1 < 0.$$

Let $\psi_M(s) = (\lambda_1+\lambda_2+\lambda_E)s + (\mu_1+\mu_2)s^{-1} - 1$. Note that $\psi_M(s)$ is convex and has zeros of 1 and ρ^{-1}. Thus, to obtain $\psi_M(\rho^{-\frac{\sin\vartheta}{\alpha}}) < 0$, it suffices to show that $1 \leq \rho^{-\frac{\sin\vartheta}{\alpha}} \leq \rho^{-1}$. But this is trivially true since $a(\theta) > 1$, for all θ.

Case 4. Suppose $0 < y < \frac{\mu_2}{\mu_1}x$.

After converting to polar coordinates again and using the bounds from Case 3, we have

$$K_Q V(x,y) - V(x,y) \leq \rho^{-ar}\Bigg[\rho^{-\frac{b}{2ar}}\Big(\lambda_1\rho^{-\frac{b\cos\theta}{a}} + (\lambda_2+\lambda_E)\rho^{-\frac{\sin\theta}{a}} + \mu_1\rho^{\frac{b\cos\theta}{a}} + \mu_2\rho^{\frac{\sin\theta}{a}} - 1\Big]\Bigg].$$

Similar to Case 3, it is enough to check that the function

$$f(\theta) \equiv \Big(\lambda_1\rho^{-\frac{b\cos\theta}{a}} + (\lambda_2+\lambda_E)\rho^{-\frac{\sin\theta}{a}} + \mu_1\rho^{\frac{b\cos\theta}{a}} + \mu_2\rho^{\frac{\sin\theta}{a}}\Big) - 1$$

is always negative for $0 < \theta < \vartheta$. First, rewrite f as

$$f(\theta) = \psi_1(\rho^{-\frac{b\cos\theta}{a}}) + \psi_{2\oplus\lambda}(\rho^{-\frac{\sin\theta}{a}}) \tag{35}$$

where

$$\psi_1(x) \equiv \lambda_1 x + \mu_1 x^{-1} - (\lambda_1+\mu_1)$$
$$\psi_{2\oplus\lambda}(x) \equiv (\lambda_2+\lambda_E)x + \mu_2 x^{-1} - (\lambda_2+\lambda_E+\mu_2).$$

We will find functions $\ell_1(x)$ and $\ell_2(x)$ which are upper bounds for $\psi_1(x)$ and $\psi_{2\oplus\lambda}(x)$ over the regions of interest. Then the last step will be to show that the last inequality in the following holds:

$$f(\theta) \leq \ell_1(\rho^{-\frac{b\cos\theta}{a}}) + \ell_2(\rho^{-\frac{\sin\theta}{a}}) \equiv \ell(\theta) < 0$$

for $0 < \theta < \vartheta$. To find the upper bounds, note that $\psi_1(x)$ and $\psi_{2\oplus\lambda}(x)$ are strictly convex on $x > 0$ since the second derivatives are strictly positive. The zeros of $\psi_1(x)$ occur at $x = 1$ and $x = \rho_1^{-1}$, and the zeros of $\psi_{2\oplus\lambda}(x)$ occur at $x = \rho_{2\oplus\lambda}^{-1}$ and $x = 1$.

Observe that $a(0) = \sqrt{b}$, and recall that we are interested in $\psi_1(x)$ for $0 < \theta < \vartheta$. For the sake of the proof, we would like to be able to rewrite this region of interest as $\psi_1(x)$ for $x \in (\rho^{-\frac{b\cos\vartheta}{a}}, \rho^{-\sqrt{b}})$, where we have obtained the left and right endpoints by evaluating $\rho^{-\frac{b\cos\theta}{a}}$ at θ values of ϑ and 0, respectively. However, writing the region

of interest as such requires that the function $\rho^{-\frac{b\cos\theta}{a}}$ is monotone in θ. Note that for $0 \leq \theta \leq \pi/2$,

$$\begin{aligned} a^2 \equiv b\cos^2\theta \ + \ \sin^2\theta \ > b\cos^2\theta \ - \ \cos^2\theta & \\ \Rightarrow 1 > \frac{b\cos^2\theta \ - \cos^2\theta}{a^2} & \\ \Rightarrow \frac{\sin\theta}{a} > \frac{(b-1)\cos^2\theta \ \sin\theta}{a^3} & \\ \Rightarrow \frac{\sin\theta}{a} - a^{-3}(b-1)\cos^2\theta \ \sin\theta > 0 & \end{aligned}$$

Hence, we obtain

$$\frac{d}{d\theta}\rho^{-\frac{b\cos\theta}{a}} = \rho^{-\frac{b\cos\theta}{a}} b\ln\rho\left[\frac{\sin\theta}{a} - a^{-3}(b-1)\cos^2\theta \ \sin\theta\right] < 0, \tag{36}$$

for $0 \leq \theta \leq \pi/2$, and our monotonicity requirement is met. Likewise, observe that

$$\frac{d}{d\theta}\rho^{-\frac{\sin\theta}{a}} = \rho^{-\frac{\sin\theta}{a}} \ln\rho\left[-a^{-3}(b-1)\cos\theta \ \sin^2\theta \ - \frac{\cos\theta}{a}\right] > 0. \tag{37}$$

Therefore, we can also rewrite our region of interest for $\psi_{2\oplus\lambda}(\rho^{-\frac{b\cos\theta}{a}})$, $0 \leq \theta \leq \vartheta$, as $\psi_{2\oplus\lambda}(x)$ for $x \in [1, \rho^{-\frac{\sin\vartheta}{a}}]$.

Let $\ell_1(x)$ be the line that agrees with $\psi_1(x)$ at the endpoints of its region of interest $[\rho^{-\frac{b\cos\vartheta}{a}}, \rho^{-\sqrt{b}}]$. Similarly, define $\ell_2(x)$ to be the line that agrees with $\psi_{2\oplus\lambda}(x)$ at the endpoints of $[1, \rho^{-\frac{\sin\vartheta}{a}}]$. By convexity, each line is an upper bound for its respective function over its region of interest. Recall that $1 < \rho^{-\sqrt{b}} < \rho_1^{-1}$, where the second inequality holds by our pooled case assumption (29). Therefore, we are assured $\ell_1(x) < 0$ for $x \in [\rho^{-\frac{b\cos\vartheta}{a}}, \rho^{-\sqrt{b}}]$ since the endpoints of the region fall between

the zeros of $\psi_1(x)$. Now we consider two subcases.

Subcase 4A. Assume $\rho^{-\frac{\sin\vartheta}{a}} \leq \rho_{2\oplus\lambda}^{-1}$.

This is the easier of the two subcases. Note that the left endpoint of the region of interest $[1, \rho^{-\frac{\sin\vartheta}{a}}]$ is a zero of $\psi_{2\oplus\lambda}(x)$; i.e., $\ell_2(1) = \psi_{2\oplus\lambda}(1) = 0$. Hence, the line $\ell_2(x)$ would be negative for $x \in (1, \rho^{-\frac{\sin\vartheta}{a}})$ if it is less than or equal to zero at its right endpoint; i.e, if $\rho^{-\frac{\sin\vartheta}{a}} \leq \rho_{2\oplus\lambda}$. This is true by our subcase assumption. Hence $\ell_1 < 0$ and $\ell_2 \leq 0$ over their regions of interest, and it follows that $\ell(\theta) < 0$ for $0 \leq \theta \leq \vartheta$.

Subcase 4B. Assume $\rho^{-\frac{\sin\vartheta}{a}} > \rho_{2\oplus\lambda}^{-1}$.

In this subcase $\ell_2(x) > 0$ for $x \in (1, \rho^{-\sqrt{b}})$, which implies that $\ell_2(x)$ has a positive slope. Note that $\ell(0) = f(0) = \psi_1(\rho^{-\sqrt{b}}) < 0$, and $\ell(\vartheta) = f(\vartheta) < 0$ where the first inequality follows from Case 1 and the second inequality follows from Case 3. Hence, if $\ell(\theta)$ is ever non-negative, there must exist a local maximum at some point θ_0 between 0 and ϑ with $\ell(\theta_0) \geq 0$. Since θ_0 is a local maximum,

$$\begin{aligned}
\ell'(\theta_0) &= \frac{d}{d\theta}\ell_1(\rho^{-\frac{b\cos\theta}{a}})|_{\theta=\theta_0} + \frac{d}{d\theta}\ell_2(\rho^{-\frac{\sin\theta}{a}})|_{\theta=\theta_0} \\
&= s_1 \ln(\rho)\tau_1\rho^{-\frac{b\cos\theta_0}{a}} + s_2\ln(\rho)\tau_2\rho^{-\frac{\sin\theta_0}{a}} \\
&= 0
\end{aligned}$$

where s_i is the slope of the line $\ell_i(x)$, $i = 1, 2$, and τ_1 and τ_2 are defined as follows.

$$\tau_1 = \rho^{-\frac{b\cos\theta}{a}} b \ln\rho \left[\frac{\sin\theta_0}{a} - a^{-3}(b-1)\cos^2\theta_0\sin\theta_0\right] \tag{38}$$

$$\tau_2 = \rho^{-\frac{\sin\theta}{a}} \ln\rho \left[-a^{-3}(b-1)\cos\theta_0\sin^2\theta_0 - \frac{\cos\theta_0}{a}\right]. \tag{39}$$

In (36) and (37), it was shown that $\tau_1 < 0$ and $\tau_2 > 0$, respectively. Therefore, it

follows that

$$0 < -\frac{\tau_1}{\tau_2}\rho^{-\left(\frac{b\cos\theta_0}{a}+\frac{\sin\theta_0}{a}\right)} = \frac{s_2}{s_1}\,.$$

Since $s_2 > 0$ by our subcase assumption, we obtain $s_1 > 0$. Given that both slopes are positive, it can also be shown that the second derivative of $\ell(\theta)$ is positive for $\theta = \theta_0$. That is,

$$\begin{aligned}\ell''(\theta_0) &= \frac{d^2}{d\theta^2}\ell_1(\rho^{-\cos\theta})|_{\theta=\theta_0} + \frac{d^2}{d\theta^2}\ell_2(\rho^{-\sin\theta})|_{\theta=\theta_0} \\ &= s_1 \ln\rho\left[\tau_1^2 + \tau_3\rho^{-\frac{b\cos\theta_0}{a}}\right] + s_2\ln\rho\left[\tau_2^2 + \tau_4\rho^{-\frac{\sin\theta_0}{a}}\right] \\ &> 0\end{aligned}$$

where τ_1 and τ_2 are defined in (38) and (39), and $\tau_i = \frac{d}{d\theta}\tau_{i-2}|_{\theta=\theta_0}$, for $i = 3, 4$. But this leads to a contradiction since it implies that θ_0 is a local minimum. Thus, it must be the case that $\ell(\theta)$ and $f(\theta)$ are strictly negative on the interval $0 < \theta < \vartheta$.

Case 5. Suppose $0 < \frac{\mu_2}{\mu_1}x < y$.

For this fifth and final case, by again using the upper bounds (31) - (34), we obtain

$$K_Q V(x,y) - V(x,y) \le \rho^{-ar}\left[\rho^{-\frac{b}{2ar}}\left((\lambda_1+\lambda_E)\rho^{-\frac{b\cos\theta}{a}} + \lambda_2\rho^{-\frac{\sin\theta}{a}} + \mu_1\rho^{\frac{b\cos\theta}{a}} + \mu_2\rho^{\frac{\sin\theta}{a}}\right) - 1\right].$$

Thus, analogous to Case 4, it suffices to show that

$$f(\theta) \equiv \left((\lambda_1+\lambda_E)\rho^{-\frac{b\cos\theta}{a}} + \lambda_2\rho^{-\frac{\sin\theta}{a}} + \mu_1\rho^{\frac{b\cos\theta}{a}} + \mu_2\rho^{\frac{\sin\theta}{a}}\right) - 1 < 0$$

for $\theta_0 \le \theta \le \pi/2$. First, rewrite f as

$$f(\theta) = \psi_{1\oplus\lambda}(\rho^{-\frac{b\cos\theta}{a}}) + \psi_2(\rho^{-\frac{\sin\theta}{a}}) \tag{40}$$

where

$$\begin{aligned}\psi_{1\oplus\lambda}(x) &\equiv (\lambda_1+\lambda_E)x + \mu_1 x^{-1} - (\lambda_1+\lambda_E+\mu_1) \\ \psi_2(x) &\equiv \lambda_2 x + \mu_2 x^{-1} - (\lambda_2+\mu_2).\end{aligned}$$

As done in the previous case, we will find functions $\ell_1(x)$ and $\ell_2(x)$ which are upper bounds for $\psi_{1\oplus\lambda}(x)$ and $\psi_2(x)$ over their respective regions of interest. Then the last step will be to show that the last inequality in the following holds:

$$f(\theta) \leq \ell_1(\rho^{-\frac{b\cos\theta}{a}}) + \ell_2(\rho^{-\frac{\sin\theta}{a}}) \equiv \ell(\theta) < 0$$

for $\vartheta \leq \theta \leq \pi/2$. Note that $\psi_{1\oplus\lambda}(x)$ and $\psi_2(x)$ are both strictly convex on $x > 0$. The zeros of $\psi_{1\oplus\lambda}(x)$ occur at $x = 1$ and $x = \rho_{1\oplus\lambda}^{-1}$; while the zeros of $\psi_2(x)$ occur at $x = \rho_2^{-1}$ and $x = 1$.

Observe that $a(\pi/2) = 1$, and recall the notation $\alpha \equiv a(\vartheta)$. Hence, we are interested in $\psi_{1\oplus\lambda}(x)$ for $x \in (1, \rho^{-\frac{b\cos\vartheta}{\alpha}})$, and $\psi_2(x)$ for $x \in (\rho^{-\frac{\sin\vartheta}{\alpha}}, \rho^{-1})$. Let $\ell_1(x)$ be the line that agrees with $\psi_{1\oplus\lambda}(x)$ at the endpoints of its region of interest, and similarly, define $\ell_2(x)$ to be the line that agrees with $\psi_2(x)$ at the endpoints of its region of interest. By convexity, each line is an upper bound for its respective function over its region of interest. Note that $1 < \rho^{-\frac{\sin\vartheta}{\alpha}} < \rho^{-1} < \rho_2^{-1}$, where the last inequality holds by our pooled case assumption $\rho > \max\{\rho_1, \rho_2\}$. Therefore, we are assured $\ell_2(x) < 0$ for $x \in (\rho^{-\frac{\sin\vartheta}{\alpha}}, \rho^{-1})$ since the endpoints of the region fall between the zeros of $\psi_2(x)$. To complete the proof, we now consider the following two subcases.

Subcase 5A. Assume $\rho^{-\frac{b\cos\vartheta}{\alpha}} \leq \rho_{1\oplus\lambda}^{-1}$.

Analogous to what was done in Subcase 4A, this is the easier of the two subcases since the region of interest for $\psi_{1\oplus\lambda}$ will now fall between the zeros of the function. In this case we obtain both $\ell_1 < 0$ and $\ell_2 \leq 0$ over their regions of interest, and it

follows that $\ell(\theta) < 0$ for $\vartheta \leq \theta \leq \pi/2$. Hence, in this subcase we are done.

Subcase 5A. Assume $\rho^{-\frac{b\cos\vartheta}{a}} > \rho_{1\oplus\lambda}^{-1}$.

Our subcase assumption now implies that $\ell_1(x) > 0$ for $x \in (1, \rho^{-\frac{b\cos\vartheta}{a}})$; i.e., $\ell_1(x)$ has a positive slope. Note that $\ell(\frac{\pi}{2}) = f(\frac{\pi}{2}) = \psi_2(\rho^{-1}) < 0$, and $\ell(\vartheta) = f(\vartheta) < 0$ where the first inequality follows from Case 2 and the second inequality follows from Case 3. Hence, if $\ell(\theta)$ is ever non-negative, there must exist a local maximum at some point θ_0 between $\frac{\pi}{2}$ and ϑ with $\ell(\theta_0) \geq 0$. Since θ_0 is a local maximum,

$$\begin{aligned}
\ell'(\theta_0) &= \frac{d}{d\theta}\ell_1(\rho^{-\frac{b\cos\theta}{a}})|_{\theta=\theta_0} + \frac{d}{d\theta}\ell_2(\rho^{-\frac{\sin\theta}{a}})|_{\theta=\theta_0} \\
&= s_1\tau_1\rho^{-\frac{b\cos\theta_0}{a}}\ln\rho + s_2\tau_2\rho^{-\frac{\sin\theta_0}{a}}\ln\rho \\
&= 0
\end{aligned}$$

where s_1 is the slope of $\psi_1(x)$, s_2 the slope of $\psi_{2\oplus\lambda}(x)$, and τ_1 and τ_2 are defined in (38) and (39), respectively. Note that this expression for $\ell'(\theta_0)$ is identical to that of Case 4. Hence, the rest of the proof follows analogously, and we are done.

□

Condition 11 requires that the probability of $F_\ell \cap \Delta$ vanishes as $\ell \to \infty$. Recall $Y_\ell(\hat{y}) \equiv \{x \in S^\infty | \tilde{x}_1 = \ell, \hat{x} = \hat{y}\}$, and let $L(\hat{y}) = |\hat{y}| + 1$. Therefore, for each $\hat{y}$, $Y_\ell(\hat{y}) \cap \Delta = \emptyset$ if $\ell \geq L(\hat{y})$, and Condition 11 is satisfied. Since $\hat{h}$ is bounded, Condition 12 is trivial. Thus we have satisfied Conditions 1 - 12. We will discuss the implied results in Chapter 5.

4.2 Unpooled Case

We now consider the case when $\rho_i > \max\{\rho, \rho_j\}$, for $i, j \in \{1, 2\}, i \neq j$. We begin with the same continuous-time, discrete state Markov process $\mathbf{Q}(t)$ described in step 1 of the pooled case. For this section, we remove our assumption that $\mu_2 > \mu_1$.

Thus, without loss of generality, we can instead assume that $\rho_1 > \rho_2$. It is therefore expected that the second server will take all of the discretionary customers along with its dedicated customers and keep its own queue small while the dedicated customers at server 1 cause the system to overload. In other words, our guess is that the most likely approach to a large number of customers will bounce along the x-axis. This covers the first two steps of the procedure, and we now move to the third. Note that whenever it is reasonable we will again keep the same notation used for discussion of the pooled case.

Step 3. For the unpooled case, both components of $\mathbf{W}[n]$ will again be one dimensional. Since we expect Q_2 to be stable when the system overloads, we define $\mathbf{W}[n] = (\tilde{W}[n], \hat{W}[n]) = (Q_1[n] + Q_2[n], Q_2[n])$ with state space $S = \tilde{S} \times \hat{S} \equiv Z_+^2$.

Before proceeding, we will now divide the proof into two cases that depend on the relationship between μ_1 and μ_2. If $\mu_1 > \mu_2$, note that as the stable component increases from $\hat{W}[n] = 0$, the sample path of the free process will always encounter the JSQ line before reaching the JSEW line. This is identical to the behavior of the JSQ system considered in [6], and will lead to identical definitions of Δ and Θ. As a result, the remainder of the proof in this case will be identical to that which is presented in [6] for the JSQ system. On the other hand, if $\mu_1 < \mu_2$ then, as the stable component increases from $\hat{W}[n] = 0$, the sample path of the free process will encounter the JSEW line before reaching the JSQ line. Consequently, the proof for this case does not follow directly from the unpooled case in [6], but will be completed within the remainder of this section.

Step 4. To construct our free process, define the set $\blacktriangle = \{(\tilde{x}, \hat{x}) \in \mathbb{Z} \times \mathbb{Z}_+ \mid \tilde{x} \leq (\frac{\mu_1}{\mu_2} + 1)\hat{x} + 1\}$. In words, $\blacktriangle$ includes all states of the W-chain such that the total expected waiting time of queue 1 is either less than or equal to that of queue 2, or

greater than that of queue 2 by at most $\frac{1}{\mu_1}$. Again we define $\Delta \equiv \blacktriangle \cap S$, which in this case translates to $\Delta = \{(\tilde{x}, \hat{x}) \in \blacktriangle \mid \tilde{x} \geq 0, \hat{x} \leq \tilde{x}, \tilde{x} \leq (\frac{\mu_1}{\mu_2} + 1)\hat{x} + 1\}$. Once the boundary is removed, the resulting transition structure can be thought of as changing the discretionary customers into customers dedicated to queue 2 and allowing a negative amount of customers at queue 1. Note that since we do not expect the region $\blacktriangle$ to come into play during the approach to F_ℓ , removal of Δ as a drift barrier to form the free process seems reasonable.

Step 5. Again, we wish to obtain a function $h(x)$ such that $K^\infty h = h$. Consider the function $h(x) = {\rho_1}^{\hat{x}-\tilde{x}}$. Using either of the two transition probability structures present in the free process, we obtain

$$\begin{aligned} E[h(W^\infty[1])|W^\infty[0] = (\tilde{x}, \hat{x})] &= \lambda_1 \rho_1^{-(\tilde{x}+1)+\hat{x}} + \mu_1 \rho_1^{-(\tilde{x}-1)+\hat{x}} + (\lambda_2 + \lambda_E + \mu_2)\rho_1^{-\tilde{x}+\hat{x}} \\ &= {\rho_1}^{\hat{x}-\tilde{x}}(\lambda_1 \rho_1^{-1} + \mu_1 \rho_1 + \lambda_2 + \lambda_E + \mu_2) \\ &= h(\tilde{x}, \hat{x}) \end{aligned}$$

Hence, this choice for our harmonic function is valid. Using this function, we then define our twisted free process $\mathcal{W}^\infty$ as shown in (23). Figure 3 shows the resulting transition probabilities for the twisted free process.

Step 6. Observe that our free process is a Markov additive process where the total number of customers is the additive process and the queue length at node 2 is the Markovian component. Therefore Condition 1 is satisfied. Likewise, Conditions 2 and 3 were also already satisfied via our construction of the free process. Conditions 4 and 5 are also satisfied as we have already chosen our harmonic function for the free process and used it to construct the twisted free process. Note that the effect of the twist is to reverse the service rate μ_1 and the arrival rate λ_1 of the first queue. On the surface, this seems promising as it is our hope that a successful twist will lead to

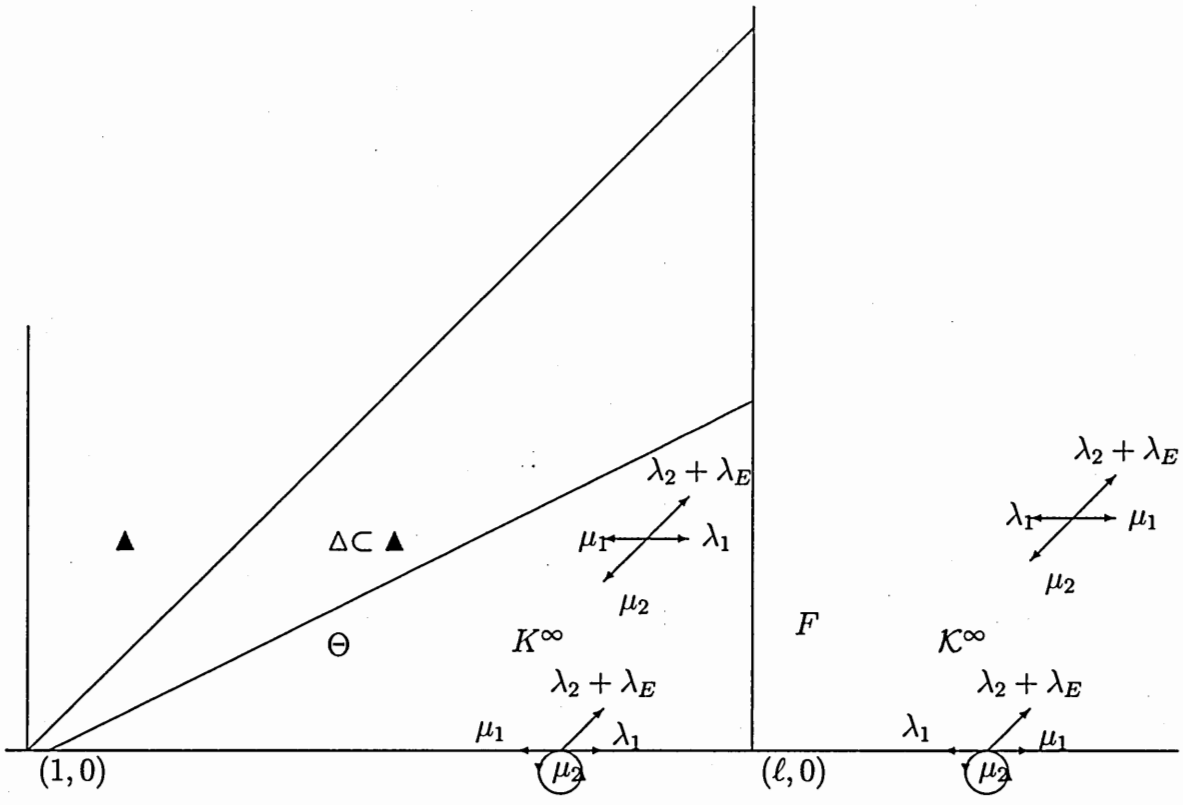

Figure 3: Transition probabilities of W^∞ and $\mathcal{W}^\infty$.

a reversal in the system's drift direction. This will be checked formally in Condition 7.

Condition 6 requires that the marginal chain $\{\hat{\mathcal{W}}^\infty[n], n \in \mathbb{Z}_+\}$, with transition kernel $\hat{\mathcal{K}}^\infty$ has a stationary probability distribution $\varphi(\cdot)$. Observe that $\hat{\mathcal{W}}^\infty[n]$ is simply an M/M/1 queue with arrival rate $\lambda_2 + \lambda_E$ and service rate μ_2. Hence for positive recurrence we need only show that $\rho_{2\oplus\lambda} \equiv \frac{\lambda_2+\lambda_E}{\mu_2} < 1$. Analogous to what was done in Section 4.1, we can express ρ as the following weighted sum:

$$\rho = \frac{\mu_1}{\mu_1+\mu_2}\frac{\lambda_1}{\mu_1} + \frac{\mu_2}{\mu_1+\mu_2}\frac{\lambda_2+\lambda_E}{\mu_2} \tag{41}$$

which is of the form $c\rho_1 + (1-c)\rho_{2\oplus\lambda}$, where $0 < c < 1$ and $\rho_{2\oplus\lambda} = \frac{\lambda_2+\lambda}{\mu_1}$. Using this expression and our case assumption that $\rho_1 > \rho$, we readily obtain $\rho_{2\oplus\lambda} < \rho < 1$,

where the second inequality is from our stability assumption. Thus we have

$$\varphi(\hat{y}) = \left(1 - \frac{\lambda_2 + \lambda_E}{\mu_2}\right)\left(\frac{\lambda_2 + \lambda_E}{\mu_2}\right)^{\hat{y}} \tag{42}$$

For Condition 7, observe that

$$\begin{aligned}\tilde{d} &= [1 - \varphi(0)]\,[(\lambda_2 + \lambda_E + \mu_1) - (\lambda_1 - \mu_2)] + \varphi(0)\,[(\lambda_2 + \lambda_E + \mu_1) - \mu_1] \\ &= \mu_1 - \lambda_1 > 0,\end{aligned}$$

where the last inequality again follows from our stability assumption. As in the pooled case, Condition 8 follows from the existence of φ, the law of large numbers, and the fact that $\tilde{d}$ is strictly positive.

Condition 9 equates to showing

$$\sum_{\hat{x}\in\hat{S}} \varphi(\hat{x})\rho_1^{-\hat{x}} < \infty \tag{43}$$

where $\varphi(\hat{x})$ is as defined in (42). Recall from our proof of Condition 6 that $\rho_{2\oplus\lambda} < \rho_1$. Hence, $\rho_1^{-1}\rho_{2\oplus\lambda} < 1$, and (43) can be expressed as a geometric series $\sum_{n=0}^{\infty} ax^n$, with $|x| < 1$. Thus Condition 9 is also satisfied.

We momentarily skip Condition 10 and move to Condition 11. Note that by defining $L(\hat{y}) = 2\hat{y} + 2$, for a given $\hat{y}$ and $\ell \geq L(\hat{y})$ we are ensured to be in the interior by definition of $\blacktriangle$. Hence Condition 11 is satisfied.

For Condition 10 recall that we wish to show

$$\sum_{x\in\Delta} \pi(x)h(x)\chi\{K(x,\Theta) > 0\} < \infty. \tag{44}$$

So, given the definition of Δ, it suffices to show that

$$\sum_{\hat{x}\geq 0} \pi_W((r+1)\hat{x}+1,\hat{x})\rho_1^{-((r+1)\hat{x}+1)}\rho_1^{\hat{x}} = \sum_{y\geq 0} \pi_Q(ry+1,y)\rho_1^{-(ry+1)} < \infty$$

where $r = \frac{\mu_1}{\mu_2}$. This inequality will be implied by our proof of Condition 12. For Condition 12, we take advantage of the similarities between the unpooled case of the JSEW policy, and that of the JSQ policy as analyzed in [6]. Recall that for this condition we need to find a function $\hat{V}(\hat{y})$ such that

$$\sum_{\hat{y}\in\hat{S}} \hat{\mathcal{K}}^{\infty}(\hat{x},\hat{y})(\hat{V}(\hat{y}) - \hat{V}(\hat{x})) \leq -\hat{h}(\hat{x}) + b\chi\{\hat{x}\in C\}.$$

for some $b < \infty$ and finite set C. For the JSQ policy, it is shown in [6] that this inequality is satisfied by defining $C = \{0\}$, and $\hat{V}(\hat{y}) = \rho_1^{-\hat{y}}/|\psi_{2\oplus\lambda}(\rho_1^{-1})|$, where $\psi_{2\oplus\lambda}(s) = (\lambda_2+\lambda_E)(s-1) + \mu_2(s^{-1}-1)$. Note that $\hat{\mathcal{K}}^{\infty}$ for the JSQ system is identical to our JSEW marginal twisted free kernel, while $\hat{h}$ and $\hat{S}$ are also identically defined. Hence (24) also holds for the JSEW policy, and we are left to check that

$$\sum_{\hat{y}\geq 0} \hat{V}(\hat{y})\hat{\lambda}(\hat{y}) < \infty,$$

or equivalently,

$$\sum_{\hat{y}\geq 0} \rho_1^{-\hat{y}} \sum_{\hat{y}\leq\tilde{x}\leq r\hat{y}+1} h(\tilde{x},\hat{y})\pi_W(\tilde{x},\hat{y}) < \infty,$$

or equivalently,

$$\sum_{y=0}^{\infty}\sum_{x=0}^{ry+1}\rho_1^{-(x+y)}\pi_Q(x,y)<\infty.$$

By Proposition 2 of [6], if $\rho<\rho_1<1$ then

$$\sum_{y=0}^{\infty}\sum_{x=0}^{y+1}\rho_1^{-(x+y)}\pi_Q(x,y)<\infty.$$

Since $0<r\leq 1$, this implies both Conditions 10 and 12 for our JSEW policy, and we are done.

CHAPTER V

EXACT ASYMPTOTIC RESULTS

In Chapter 4, we applied the methodology outlined in Chapter 3 to our JSEW system. Given the completed proofs, we now make the results of [6] and [10] specific to our network of interest. In Section 5.1, we begin with a summary of our findings with respect to the qualitative behavior of the system. We then discuss periodicity in Section 5.2, and provide our main theorems in the subsequent section. In Section 5.4, we discuss the validity of our added assumptions for our proof of the pooled case. In the final two sections, we continue to compare the JSEW and JSQ policies via simulation, and also present our results for a variation of the JSQ policy.

5.1 Strongly versus Weakly Pooled

It was revealed within our analysis of the JSEW queueing network that, based on the system parameters, the network will approach the rare event F_ℓ in one of three possible ways. If $\rho_1 > \max\{\rho_2, \rho\}$, then we concluded that the number of customers at queue 2 will form a positive recurrent Markov chain while the dedicated customers at the other queue will cause the system to overload. This will lead to an approach along the x-axis of Figure 4.1. Similarly, we proved that having $\rho_2 > \max\{\rho_1, \rho\}$ will cause queue 2 to be the primary cause of the system overload, causing the approach to F_ℓ to bounce along the y-axis. These two scenarios are called the unpooled case

and are identical for the JSEW and JSQ networks.

When $\rho > \max\{\rho_1, \rho_2\}$ we were said to be in the pooled case. When analyzing the JSQ system in [6], the pooled case divided into two subcases in which the network exhibited very different qualitative behaviors. The criterion for this dichotomy stemmed from satisfying Condition 6 of the methodology. To see how this compares to the JSEW system, we briefly describe this dichotomy and the derivation of the necessary conditions.

Recall that Condition 6 requires that the marginal chain $\hat{\mathcal{W}}^\infty[n], n \in \mathbb{Z}_+$, with transition kernel $\hat{\mathcal{K}}^\infty$ has a stationary probability distribution. To verify this for the JSQ system, it needed to be shown that $\hat{\mathcal{W}}^\infty \equiv Q_2 - Q_1$ was positive recurrent. The transition structure for $\hat{\mathcal{W}}^\infty$ of the JSQ system is similar to that of the JSEW system given in Figure 2. Using this figure, and noting that transitions of $\hat{\mathcal{W}}^\infty$ always have a jump size of 1, observe that for $\hat{x} > 0$ the vertical drift of $\hat{\mathcal{W}}^\infty$ is given as $\hat{d}_+ = \rho(\mu_1 - \mu_2) + \rho^{-1}(\lambda_2 - \lambda_1 - \lambda_Q)$. For a drift towards zero, this implies that we must satisfy the following condition for the JSQ pooled case.

$$\lambda_Q > \rho^2(\mu_1 - \mu_2) + (\lambda_2 - \lambda_1) \tag{45}$$

Similarly, for $\hat{x} < 0$, we obtain $\hat{d}_- = \rho(\mu_1 - \mu_2) + \rho^{-1}(\lambda_2 + \lambda_Q - \lambda_1)$. For a positive recurrent $\hat{\mathcal{W}}^\infty$, this implies the requirement that

$$\lambda_Q > \rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2) \tag{46}$$

By rearranging terms, it can be shown that (46) is equivalent to the following inequality.

$$\frac{(\lambda_2 + \lambda_Q)/\rho - \rho\mu_2}{\lambda_1/\rho - \rho\mu_1} > 1 \tag{47}$$

Moreover, by our assumption that $\mu_2 > \mu_1$, the above inequality will hold if

$$\frac{(\lambda_2 + \lambda_Q)/\rho - \rho\mu_2}{\lambda_1/\rho - \rho\mu_1} > \frac{\mu_2}{\mu_1} . \tag{48}$$

Multiplying both sides of (48) by $\lambda_1/\rho - \rho\mu_1$, and also adding $\rho\mu_2$ to each side of the inequality yields the following equivalent statement.

$$\frac{\lambda_2 + \lambda}{\mu_2} > \frac{\lambda_1}{\mu_1} . \tag{49}$$

Using an argument analogous to that which follows (41), it can be shown that (49) is implied by our pooling assumption (25). Hence, the JSQ strong pooling requirement (46) is implicitly satisfied by our underlying case assumption. On the other hand, the first requirement, (45), is not superfluous to our initial pooling assumption and hence is indeed an additional requirement for an approach to F_ℓ along the JSQ line. When this condition holds, the system is said to be strongly pooled. If this criterion does not hold true, then both queues still overload, but the approach to F_ℓ could be via a slope other than 1. This scenario was called the weakly pooled case. We summarize this JSQ pooling dichotomy in the following remark.

Remark 1 *Suppose* $\rho > \max\{\rho_1, \rho_2\}$, *and assume that* $\mu_2 > \mu_1$. *If*

$$\lambda_Q > \rho^2(\mu_1 - \mu_2) + (\lambda_2 - \lambda_1) \tag{50}$$

then the JSQ system will strongly pool along the JSQ line. If this inequality does not hold, then the system will be considered weakly pooled, and the approach to F_ℓ *will remain on or above the JSQ line.*

Recall that during our analysis of the JSEW network in Chapter 5, going through Condition 6 in a similar fashion led us to the following two conditions for positive recurrence of $\hat{\mathcal{W}}^{\infty}$.

$$\rho_1 < \rho_{2\oplus\lambda} \qquad\qquad \rho_2 < \rho_{1\oplus\lambda} \tag{51}$$

However, further inspection of these two inequalities implied the following conclusion.

Remark 2 *Given $\rho > \max\{\rho_1, \rho_2\}$, both of the inequalities in (51) will be superfluous. Hence, given our pooling assumption, the JSEW queueing network will necessarily strongly pool along the JSEW line.*

Thus our JSEW queueing network differs from that of JSQ in that there exists no dichotomy between strong and weak pooling. In this respect, the JSEW system is similar to the JSAW network (queueing by actual waiting time) as analyzed in [11] in that neither exhibits the weak pooling phenomenon. Consequently it can be said that these two queueing disciplines are more robust than JSQ since strong pooling is achieved without additional requirements on the system parameters. The drawback, of course, is that more information is required for implementing the JSEW policy since the JSQ policy does not require a priori knowledge of the service rates.

5.2 Periodicity

Before stating our main theorems, there are additional technicalities that must be addressed. We begin by introducing the notion of periodicity as it pertains to our $\mathcal{W}^{\infty}$ process defined in Chapter 4. Recall that for the JSEW strongly pooled case we defined $\mathcal{W}^{\infty}$ such that $\tilde{\mathcal{W}}^{\infty} = Q_1 + Q_2$ and $\hat{\mathcal{W}}^{\infty} = \frac{1}{\mu_2}Q_2 - \frac{1}{\mu_1}Q_1$. Here the state space of the marginal Markov chain $\{\hat{\mathcal{W}}^{\infty}[n], n \in \mathbb{Z}_+\}$ is the additive subgroup

$\hat{S} \equiv \{\frac{m}{\mu_1} + \frac{n}{\mu_2} \,|\, m, n \in \mathbb{Z}\}$. Define $A_j = \{k \,|\, \frac{y}{\mu_2} - \frac{x}{\mu_1} = k, x + y = j\}$. In what follows, we will show that for any fixed value of ℓ, having $Q_1 + Q_2 = \ell$ implies $A_\ell \cap \hat{S} \neq \emptyset$. Moreover, in order to ensure the existence of $\varphi(\cdot)$ we must also prove that the family of sets $\{A_j,\, j \in \mathbb{Z}\}$ exhibits periodicity. That is, there must exist a p such that $A_i = A_j$ for $i = j \mod p$.

Suppose $x + y = j$, for some $x, y \in \mathbb{Z}$. Hence we can express $k \in A_j$ as

$$\begin{aligned} \frac{y}{\mu_2} - \frac{x}{\mu_1} &= \frac{y}{\mu_2} - \frac{j-y}{\mu_1} \\ &= \frac{\mu_1}{\mu_1}\frac{y}{\mu_2} - \frac{\mu_2}{\mu_2}\frac{j}{\mu_1} + \frac{\mu_2}{\mu_2}\frac{y}{\mu_1} \\ &= \frac{(\mu_1+\mu_2)y - j\mu_2}{\mu_1\mu_2} \\ &= \frac{\mu_1+\mu_2}{\mu_1\mu_2}y - \frac{j}{\mu_1}. \end{aligned}$$

Thus for $j = 0$, $k \in A_0$ is of the form $\frac{\mu_1+\mu_2}{\mu_1\mu_2}y$, for $y \in \mathbb{Z}$. For $j = 1$ we obtain $A_1 = \{\frac{\mu_1+\mu_2}{\mu_1\mu_2}y - \frac{1}{\mu_1} \;\mid\; y \in \mathbb{Z}\}$. Since $\frac{1}{\mu_1} < \frac{\mu_1+\mu_2}{\mu_1\mu_2}$ we are ensured that $A_0 \cap A_1 = \emptyset$. Moreover, note that in general A_j can be thought of as a shift of A_0 by the amount $-\frac{j}{\mu_1}$. Accordingly we will use the notation $A_j = A_0 - \frac{j}{\mu_1}$. In this manner, it is obvious that in constructing the family of sets $\{A_0, A_1, A_2, \ldots, A_j, \ldots\}$, these sets will remain disjoint unless $\frac{j}{\mu_2}$ is a multiple of $\frac{\mu_1+\mu_2}{\mu_1\mu_2}$ for some $j \in \mathbb{Z}$. We now give a sufficient condition for this to occur.

Suppose $\frac{\mu_1}{\mu_2}$ is rational. Therefore, there exists $m, n \in \mathbb{Z}$ such that $\frac{m}{n} = \frac{\mu_1}{\mu_2} + 1$. Multiplying both sides of this equation by $\frac{n}{\mu_1}$, we equivalently obtain $\frac{m}{\mu_1} = \frac{\mu_1+\mu_2}{\mu_1\mu_2}n$. Hence,

$$\begin{aligned} A_m &= A_0 - \frac{m}{\mu_1} \\ &= \{\frac{\mu_1+\mu_2}{\mu_1\mu_2}y - \frac{\mu_1+\mu_2}{\mu_1\mu_2}n \,|\, y \in \mathbb{Z}\} \\ &= A_0. \end{aligned}$$

Similarly, we can obtain that $A_{um+v} = A_{v \bmod m}$, for $u, v \in \mathbb{Z}$. Thus if the ratio of service rates is rational, then we will be ensured the existence of a p such that $A_i = A_j$ for $i = j \mod p$.

5.3 Main Results

Recall that $H(z)$ is the probability that $\mathcal{W}^\infty$, starting at $z = (\tilde{z}, \hat{z})$, never visits $\blacktriangle$. Also, define

$$f_1 = \sum_{x \in \triangle} \left[\pi_W(x) \rho^{-\tilde{x}} H(x) + \pi_W(x) \rho^{-\tilde{x}} H(x) \right] \tag{52}$$

where

$$\triangle = \begin{cases} \{(\tilde{x}, \hat{x}) \in \mathbb{Z}^2 \mid \frac{1}{\mu_1}\tilde{x} + \hat{x} = 0 \text{ or } \frac{1}{\mu_2}\tilde{x} - \hat{x} = 0\} & \text{if } \rho > \max\{\rho_1, \rho_2\} \\ \{(\tilde{x}, \hat{x}) \in \mathbb{Z} \times \mathbb{Z}_+ \mid \tilde{x} \geq 0, \hat{x} \leq \tilde{x}, \tilde{x} \leq (\frac{\mu_1}{\mu_2} + 1)\hat{x} + 1\} & \text{otherwise.} \end{cases} \tag{53}$$

Note that f_1 is finite by Condition 10 of the methodology. The following theorems give our results for the pooled case. For Theorems 6 and 7, recall that in the pooled case $\hat{W}$ is defined as the difference between the expected waiting times of the two queues. Lastly, note that for Theorems 6 through 8 we impose the underlying assumption $\mu_2 > \mu_1$ as done in the corresponding proofs in Chapter 4.

Theorem 6 *Assume μ_1/μ_2 is rational and $\rho > \max\{\rho_1, \rho_2\}$. If (55) holds, then*

$$\pi_W((\ell, \hat{y})) \sim f_1 \rho^\ell \frac{1}{\tilde{d}} \varphi(\hat{y}) p \; \chi\{\hat{y} \in A_\ell\}$$

where φ, the stationary distribution for $\hat{\mathcal{W}}^\infty$, can be obtained via simulation, $\tilde{d} = \mu - \lambda$, f_1 is defined in (52), and p is the periodicity of $\{A_j, j \in \mathbb{Z}\}$.

Let ν denote the stationary overshoot distribution $(\tilde{\mathcal{W}}^{\infty}[\mathcal{T}_{\ell}^{\infty}] - \ell, \hat{\mathcal{W}}^{\infty}[\mathcal{T}_{\ell}^{\infty}]), \ell = 0, 1, \ldots$. For our systems of interest, note that the overshoot will always be zero; i.e., $\tilde{\mathcal{W}}^{\infty}[\mathcal{T}_{\ell}^{\infty}] - \ell = 0$, for all $\ell \in \mathbb{Z}_+$. Hence, we may instead use the notation $\nu(0, \cdot)$, and refer to this as the stationary distribution of $\hat{\mathcal{W}}^{\infty}[\mathcal{T}_{\ell}^{\infty}]$. Observe that ν is obtainable by fast simulation since $\tilde{\mathcal{W}}^{\infty}$ drifts to infinity.

Theorem 7 *Under the conditions of Theorem 6, as* $\ell \to \infty$,

$$\mathrm{P}_{\sigma}\{\hat{W}[T_{\ell}] = \hat{y} \mid T_{\ell} < T_{\triangle}\} \sim \; p\; \nu(0, \hat{y}) \chi\{\hat{y} \in A_{\ell}\}$$

for any initial point σ.

The following theorem gives an asymptotic expression for the expected hitting time to F_{ℓ} for the pooled case.

Theorem 8 *If* $\rho > \max\{\rho_1, \rho_2\}$, *then*

$$\mathrm{E}_{\sigma}[T_{\ell}] \sim f_1 \rho^{-\ell} \text{ as } \ell \to \infty,$$

where σ *is any initial point in* $\triangle$, *and* f_1 *is defined in (52).*

Finally, in the following theorems we consider the unpooled case. For these theorems, we remove our assumption that $\mu_2 > \mu_1$. Consequently, without loss of generality, we can present our results under the assumption that $\rho_1 > \rho_2$.

Theorem 9 *If* $\rho < \max\{\rho_1, \rho_2\} = \rho_1$ *we have:*

$$\pi_W((\ell, \hat{y})) \sim f_1 \rho_1^{\ell} \; \frac{1}{\tilde{d}} \; \rho_1^{-\hat{y}} \left(1 - \frac{\lambda_2 + \lambda_E}{\mu_2}\right) \left(\frac{\lambda_2 + \lambda_E}{\mu_2}\right)^{\hat{y}}$$

where f_1 *is defined in (52), and* $\tilde{d} = \mu_1 - \lambda_1$.

This means that the stationary measure is product-form for large ℓ. The constant f_1 can be obtained by fast simulation since we only need π_W near the origin.

Theorem 10 *If* $\rho < \max\{\rho_1, \rho_2\} = \rho_1$ *then, as* $\ell \to \infty$,

$$\mathrm{P}\{\hat{W}[T_\ell] = \hat{y} \mid T_\ell < T_\sigma\} \to^T \rho_1^{-\hat{y}}\nu(0,\hat{y}) / \left(\sum_{\hat{y}} \rho_1^{-\hat{y}}\nu(0,\hat{y})\right)$$

where σ *is some initial state in* Δ, T_σ *is the return time to* σ, $\to^T$ *denotes convergence in total variation,* $\nu(0,\cdot)$ *denotes the stationary distribution of* $\hat{\mathcal{W}}^\infty[\mathcal{T}_\ell^\infty]$, *and* $\mathcal{T}_\ell^\infty$ *is the first time the twisted free process first reaches* ℓ.

Theorem 11 *If* $\rho < \max\{\rho_1, \rho_2\} = \rho_1$ *then*

$$\mathrm{E}_\sigma T_\ell \sim \rho_1^{-\ell} g_1^{-1} \text{ as } \ell \to \infty$$

where σ *is any initial point in* Δ, f_1 *and* Δ *are defined in (52) and (53), and*

$$g_1 \equiv f_1 \sum_{z \geq 0} \rho_1^{-z}\nu(0,z). \tag{54}$$

5.4 Additional Pooled Case Assumption

In Chapter 4, in order to complete the proofs for the JSEW pooled case, we assumed that

$$\frac{\ln \rho_1}{\ln \rho} > \sqrt{\frac{\mu_2}{\mu_1}}. \tag{55}$$

In doing so, we have obviously reduced the size of the parameter space for which our strong pooling results hold true. A simulation of randomly generated parameter

sets suggests that we will be in our pooled case (i.e., have $\rho > \max\{\rho_1, \rho_2\}$) close to 65% of the time. In the remaining 35% of the parameter space, we will instead be in the unpooled scenario, for which the proofs do not require the additional condition (55). Moreover, for those parameters within the pooled case, (55) holds for approximately 92% of the parameters generated. As a result, we can say that overall the resulting theorems remain true for approximately 95% of the parameter space, and hence the required assumption does not seem too onerous.

Moreover, we suspect that these results indeed hold for the entire parameter space. Our conjecture is that the use of better upper bounds within the current proof or a different choice for the Lyapunov function could capture the remaining 5% of the parameter space. It is our intention to explore this conjecture within our future work, and to eventually make this additional assumption unnecessary.

5.5 Implementing Our Policies

Part of the novelty of this research is the ability to distinguish between two policies, namely JSQ and JSEW, in ways that we would otherwise be unable. In Section 5.1, we discussed the qualitative differences in the behavior of each system as the joint queue lengths grew large. In addition, in Section 5.3 we gave sub-exponential coefficients for our exact asymptotic expressions. In traditional large deviation theory, the analogous results for the JSQ and JSEW systems would lack the sub-exponential coefficients presented in Theorems 6 through 11. Consequently, we would be left with only the rough asymptotics for these two systems, which turn out to be identical. However, despite the rough asymptotics being the same for the two policies, it is intuitively clear that JSEW should perform at least as well as JSQ.

Note that if the service rate at each node is known beforehand, then implementing

Table 1: Simulation examples for exact asymptotics

Case	$\frac{\mu_2}{\mu_1}$	ρ	ρ_1	ρ_2	c_E	c_Q
(1)	1.026	0.792	0.724	0.023	0.562	0.610
(2)	1.391	0.787	0.162	0.695	0.343	0.373
(3)	5.352	0.353	0.2534	0.3176	0.759	2.139
(4)	33.345	0.941	0.404	0.028	0.081	0.921
(5)	34.148	0.821	0.072	0.207	0.136	20.537

JSEW over JSQ requires little or no additional work. Otherwise, the advantage of implementing JSQ instead of JSEW would be that it requires no prior knowledge of the service rates. Consequently, the cost/benefit of using JSEW over JSQ is not always obvious, and may differ depending on managerial objectives. As the difference between the two service rates increases, intuition would suggest that the JSEW policy should be the more efficient policy. But how much more efficient? And is it enough to warrant a change in policy? To quantify these differences, using the results in this thesis, we simulated the twisted free process to estimate the exact asymptotic coefficients.

In Table 1, we give examples which illustrate the usefulness of these asymptotic approximations. Recall that, for the pooled case under either policy, the expected time to reach ℓ customers in the system is of the form $c\rho^{-\ell}$, where c is our exact asymptotic coefficient. Using the same parameters for both the JSQ and JSEW systems, we varied the relationship between μ_1 and μ_2, and obtained the coefficient c for each system via fast simulation. In Table 1, C_E and C_Q denote the approximations of the exact asymptotic coefficients for the JSEW and JSQ networks, respectively. For the first two rows, the results are essentially what we would expect: for μ_2/μ_1 close to 1, the coefficients for the two policies are relatively close to each other, with that of the JSEW policy being slightly smaller as expected. Likewise, the third row also

illustrates what we'd expect; an increased difference in the service rates corresponding to a larger difference between the coefficients.

On the other hand, rows 4 and 5 show us that the relationship between the ratio of service rates and the two coefficients is not as straightforward as one may anticipate. In both cases, μ_2/μ_1 is of the same magnitude (μ_2 approximately 30 times larger than μ_1). However in Case 4, the JSQ coefficient is approximately 10 times larger than that of JSEW; whereas Case 5 illustrates the former being over 150 times larger than the latter. In words, this means that despite having similar service rate ratios, there is arguably a significant difference in the value of implementing JSEW over JSQ for these two systems. In Case 4, switching from JSQ to JSEW meant that it would take the system (on average) 10 times longer to reach ℓ customers in the queue. Whereas, switching policies in Case 5 lead to the system taking over 100 times longer to this rare event, making the switch more advantageous in this scenario. In general, such information may have engineering significance, and thus may be worthwhile when managing networks.

5.6 JSQ / SLQ Policy

In this section, we consider a variation of the JSQ system that includes the addition of a third discretionary server that serves the longest of the two queues. We will refer to a system implementing this policy as a JSQ/SLQ network. As mentioned in Section 2.2, the addition of this third server may change the service rate at either node, but not the service capacity. That is, at any given time, each queue will still only be able to serve one customer at a time but may have added resources to do so. In the case of a tie, the service rate of the additional server will be split equally between the two nodes. Our consideration of this policy is motivated by the work of Menich and Serfozo [12] in which they describe a Markovian model with state-dependent arrival

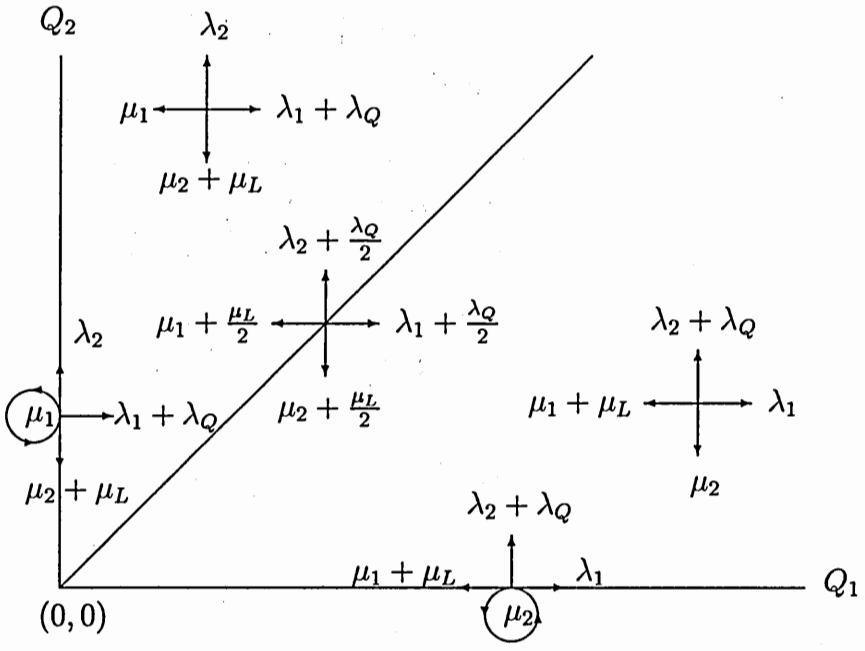

Figure 4: Transition probabilities for the JSQ/SLQ system.

process, service rates, and holding cost rate; and give conditions under which the JSQ/SLQ policy is optimal.

Let μ_L denote the service rate of this discretionary server, and define $\rho_{*1} = \frac{\lambda_1}{\mu_1+\mu_L}$, $\rho_{*2} = \frac{\lambda_2}{\mu_2+\mu_L}$, and $\rho_* = \frac{\lambda_1+\lambda_2+\lambda_Q}{\mu_1+\mu_2+\mu_L}$. It turns out that the proof is very similar to what was presented in Chapter 4 for the JSEW system, with only minor adjustments to account for the addition of an SLQ server. Most of the adjustments to the proof will only involve replacing ρ, ρ_1 or ρ_2 with the analogous "*" version of the respective traffic intensity parameter. Accordingly we will simply outline the proof for this system, highlighting the steps where the two proofs differ, and then give the implied results for the JSQ/SLQ system.

For the pooled case, we begin with the assumption that $\rho_* > \max\{\rho_{*1}, \rho_{*2}\}$. As done in previous sections, we will again assume without loss of generality that $\mu_2 > \mu_1$.

In Step 1 of the methodology, the Q-Chain is constructed in the same manner as done for JSEW in Chapter 4. For Step 2, given that we are in the pooled scenario, we anticipate an approach to F_ℓ along the JSQ line. That is, as in JSQ, we expect the queue lengths to remain close to each other, and we proceed accordingly. For Step 3, we define the free process as $\mathbf{W}[n] = (\tilde{W}[n], \hat{W}[n]) = (Q_1[n] + Q_2[n], Q_2[n] - Q_1[n])$. The state space for our W process is $S \equiv \tilde{S} \times \hat{S} \equiv \{(x+y,\ y-x) \mid x, y \geq 0\} \subset \mathbb{Z}_+ \times \mathbb{Z}$. For Step 4, we partition the state space exactly as done for the JSEW proof, defining $\triangle \equiv \{(x,y) \in S_Q \mid x = 0 \text{ or } y = 0\}$, and $\blacktriangle \equiv (\mathbb{Z} \times \hat{S}) \setminus S$. For Step 5 of the pooled case, observe that after replacing ρ with ρ^*, the same harmonic function used for JSEW works for the JSQ/SLQ system. We then use this harmonic function to define the transition structure of the twisted free process $\mathcal{W}^\infty \equiv \{\mathcal{W}^\infty[n], n = 0, 1, 2, \ldots\}$ in the usual fashion.

For Step 6, we must check that our three constructed Markov chains satisfy the 12 technical conditions outlined in Chapter 3. In doing so, it turns out that the first deviation from our JSEW proof does not occur until Condition 6. Recall that this condition requires $\hat{\mathcal{W}}^\infty$ to be positive recurrent. For the JSQ/SLQ policy, this translates to requiring that the difference in the two queue lengths has a drift towards zero. The vertical drift both above and below the $\hat{x}$-axis can be obtained exactly as done for the JSQ system in Section 5.1. In doing so, above the $\hat{x}$-axis we obtain the requirement

$$\lambda_Q > \rho_*^2(\mu_1 - \mu_2 - \mu_L) + (\lambda_2 - \lambda_1) \tag{56}$$

and likewise, when $\hat{x} < 0$, we obtain

$$\lambda_Q > \rho_*^2(\mu_2 - \mu_1 - \mu_L) + (\lambda_1 - \lambda_2). \tag{57}$$

As expected, by setting $\mu_L = 0$, these two inequalities reduce to the analogous conditions (45) and (46) described in Section 5.1 for the JSQ system. Also note that $\rho_* > \max\{\rho_{*1}, \rho_{*2}\}$ implies $\frac{\lambda}{\mu_1+\mu_2} > \max\{\rho_1, \rho_2\}$. In Section 5.1, we showed that the latter inequality, in addition to our assumption $\mu_2 > \mu_1$, implies that

$$\lambda_Q > \left(\frac{\lambda}{\mu_1 + \mu_2}\right)^2 (\mu_2 - \mu_1) + (\lambda_1 - \lambda_2). \tag{58}$$

This in turn implies (57) since $\rho_* \leq \frac{\lambda}{\mu_1+\mu_2}$. Thus, (57) is redundant to our pooling condition, and (56) becomes the condition which determines the dichotomy between strong and weak pooling in the JSQ/SLQ system. Given that this condition holds, we are assured that $\hat{\mathcal{W}}^\infty$ is positive recurrent with stationary distribution φ given as follows.

$$\varphi(k) = \begin{cases} \frac{\rho^{-1}(\lambda_2+\lambda_Q/2)+\rho(\mu_1+\mu/2)}{\rho^{-1}(\lambda_1+\lambda)+\rho(\mu_2+\mu_L)} \left(\frac{\rho^{-1}\lambda_2+\rho\mu_1}{\rho^{-1}(\lambda_1+\lambda)+\rho(\mu_2+\mu_L)}\right)^{k-1} \varphi(0) & \text{if } k > 0 \\ \frac{\rho^{-1}(\lambda_1+\lambda_Q/2)+\rho(\mu_2+\mu/2)}{\rho^{-1}(\lambda_2+\lambda)+\rho(\mu_1+\mu_L)} \left(\frac{\rho^{-1}\lambda_1+\rho\mu_2}{\rho^{-1}(\lambda_2+\lambda)+\rho(\mu_1+\mu_L)}\right)^{|k|-1} \varphi(0) & \text{if } k < 0 \\ \left(\frac{\rho^{-1}(\lambda_2+\lambda_Q/2)+\rho(\mu_1+\mu/2)}{\rho^{-1}(\lambda_1+\lambda_Q)+\rho(\mu_2+\mu_L)-(\rho^{-1}\lambda_2+\rho\mu_1)} + \frac{\rho^{-1}(\lambda_1+\lambda_Q/2)+\rho(\mu_2+\mu/2)}{\rho^{-1}(\lambda_2+\lambda_Q)+\rho(\mu_1+\mu_L)-(\rho^{-1}\lambda_1+\rho\mu_2)} + 1\right)^{-1} & \text{if } k = 0. \end{cases}$$

Recall from Chapter 1 that the JSQ/SLQ system is stable iff $\rho_{\max} < 1$. Thus $\tilde{d} = \mu - \lambda > 0$, and Condition 7 is satisfied. By substituting ρ with ρ_*, the proofs for Conditions 8 through 12 become identical to those done for the JSQ system in [6], and hence for the strongly pooled case we are done. The proof for the weakly pooled case also follows with only trivial modifications to the JSQ proofs in [6]. In the same fashion, once we replace ρ_i with ρ_{*i}, $i = 1, 2$, and substitute μ_i with $\mu_i + \mu_L$, $i = 1, 2$, where applicable, the unpooled case also follows exactly from that of JSQ in [6]. Accordingly, we now state our results for the JSQ/SLQ system. We begin by characterizing the qualitative behavior of the system.

Remark 3 *Suppose* $\rho_* > \max\{\rho_{*1}, \rho_{*2}\}$, *and assume that* $\mu_2 > \mu_1$. *If*

$$\lambda_Q > \rho_*^2(\mu_1 - \mu_2 - \mu_L) + (\lambda_2 - \lambda_1) \tag{59}$$

then the JSQ/SLQ system will strongly pool along the JSQ line. If this inequality does not hold, then the system will be considered weakly pooled, and the approach to F_ℓ *will remain on or above the JSQ line.*

So given the JSQ policy, we again observe a dichotomy between strong and weak pooling, similar to that of the JSQ network discussed in Section 5.1. However, the JSQ/SLQ strong pooling condition (59) is implied by that of the JSQ system, (58), but the converse is not true. Thus, implementing the SLQ policy within the JSQ network was not enough to eliminate the weakly pooled parameter region, but it did decrease the size of this region. In this manner, adding the discretionary server made the network more robust, as expected.

We conclude this section with the statements of our quantitative findings for the JSQ/SLQ network. Define $B_j = \{k \,|\, y - x = k, x + y = j \text{ for some } x, y \in \mathbb{Z}_+\}$, and note that $\{B_j, j \in \mathbb{Z}_+\}$ has a periodicity of 2 since the difference and sum of the queue lengths must be either both negative or both positive. Also define

$$f_2 = \sum_{x \in \Delta} \left[\pi_W(x)\rho^{-\tilde{x}}H(x) + \pi_W(x)\rho^{-\tilde{x}}H(x)\right]. \tag{60}$$

where

$$\Delta = \begin{cases} \{(\tilde{x}, \hat{x}) \in \mathbb{Z}^2 \mid \tilde{x} + \hat{x} = 0 \text{ or } \tilde{x} - \hat{x} = 0\} & \text{if } \rho_* > \max\{\rho_{*1}, \rho_{*2}\} \\ \{(\tilde{x}, \hat{x}) \in \mathbb{Z} \times \mathbb{Z}_+ \mid \tilde{x} \geq 0, \hat{x} \leq \tilde{x}, \tilde{x} \leq 2\hat{x} + 1\} & \text{otherwise} \end{cases}$$

Theorem 12 *(Strongly Pooled Case)*

If $\max\{\rho_{*1}, \rho_{*2}\} < \rho_* < 1$, *and (59) holds, then*

(i)

$$\mathrm{E}_\sigma[T_\ell] \sim f_2^{-1}\rho_*^{-\ell}/(\lambda+\mu) \tag{61}$$

where σ *is any initial point in* $\triangle$.

(ii)

$$\mathrm{P}_\pi\{Q_1(t)+Q_2(t)=\ell, Q_2(t)-Q_1(t)=k\} \sim 2f_2\frac{1}{\tilde{d}}\,\rho_*^\ell\varphi(k)\chi\{k=\ell \mod 2\} \tag{62}$$

(iii)

$$\mathrm{P}\{Q_2(T_\ell)-Q_1(T_\ell)=k \mid Q[0]=(0,0)\} \sim 2\varphi(k)\chi\{k=\ell \mod 2\}. \tag{63}$$

where $\tilde{d}=\mu-\lambda$, *and the density* $\varphi(k)$ *is defined above.*

For the weakly pooled case, we expect that $Q_1[T_\ell]$, $Q_2[T_\ell]$, and their difference all become large with ℓ. Hence we let $\tilde{W}[n] = (Q_1[n]+Q_2[n], Q_1[n]-Q_2[n])$ and $\hat{W}$ be trivial. As a result, φ is a trivial measure, and the asymptotics of the stationary distribution can be given as follows.

Theorem 13 *(Weakly Pooled Case) If* $\max\{\rho_{*1}, \rho_{*2}\} < \rho_* < 1$, *and (59) does not hold, then*

$$P_\pi\{Q_1(t)+Q_2(t)=\ell\} \sim \frac{f_2}{\tilde{d}}\rho_*^\ell$$

where $\tilde{d}$ *is the same as in Theorem 12 and* f_2 *is given in* (60).

For the weakly pooled case, $\mathrm{E}[T_\ell]$ is the same as in the strongly pooled case, given in Theorem 12.

For the results in the unpooled case, we again assume without loss of generality that queue 1 grows large, while queue 2 remains stable. Accordingly, for the following

theorem note that the Markov additive process $\mathcal{W}^\infty$ consists of the total number of customers $\tilde{\mathcal{W}}^\infty$ as the additive process, and the length of queue 2 as the Markovian component, $\hat{\mathcal{W}}^\infty$.

Theorem 14 *(Unpooled Case) Without loss of generality, suppose* $\rho_{*1} > \max(\rho_*, \rho_{*2})$. *Therefore,*

(i)

$$\mathrm{E}_\sigma[T_\ell] \sim {g_2}^{-1} {\rho_{*1}}^{-\ell}/(\lambda+\mu)$$

where

$$g_2 \equiv f_2 \sum_{z\geq 0} \rho_{*1}^{-z} \nu(0,z)$$

is obtainable by fast simulation.

(ii)

$$\mathrm{P}_\pi\{Q_1(t)+Q_2(t)=\ell, Q_2(t)=k\} \sim \frac{f_2}{\mu+\mu_1-\lambda_1} {\rho_{*1}}^{\ell-k} \left(1-\frac{\lambda_2+\lambda_Q}{\mu_2}\right)\left(\frac{\lambda_2+\lambda_Q}{\mu_2}\right)^k$$

(iii)

$$\mathrm{P}\{Q_2(T_\ell)=k \mid Q[0]=(0,0)\} \sim c^{-1} {\rho_{*1}}^k \nu(0,k),$$

where $c = \sum_{k=0}^\infty \rho_{*1}^k \nu(0,k)$.

APPENDIX

A THRESHOLD POLICY

In this section, we briefly consider the asymptotic behavior of a two server queuing system operating under a threshold policy. Note that the proofs for this system are incomplete. However, the purpose of this section is to illustrate an example of our current and future work in the area of exact asymptotics. The threshold system is also interesting from the standpoint that, thus far it is one of the few examples that we have come across where the form of the harmonic function (the rough asymptotics, α, in particular) is not easily guessed and in fact difficult to give explicitly. How the methodology works in this case is of interest since for more complicated models one may be in the same position. For this case, we do manage to give a set of equations that the harmonic functions must satisfy, but do not explicitly solve them.

The Model and Policy

We consider a network consisting of two nodes, each with a single server and an infinite capacity waiting area. Customers arrive to the network according to a Poisson process with total rate $\lambda > 0$. For the threshold model, λ can be decomposed into two customer types. Customers of type 1 always join node 1, and we let λ_1 denote the corresponding customer arrival rate. The second customer type operates via a threshold policy by which customers are routed to the second server's queue until the total number at this queue reaches some predetermined threshold value m. Once this threshold is reached, the overflow customers are then routed to queue 1 until queue 2 again falls below m customers. Let λ_T denote the arrival rate of this discretionary

customer.

The service times at nodes 1 and 2 are independent and exponentially distributed with parameters μ_1 and μ_2, respectively. The total service rate in the network is denoted by μ. We let $\mathbf{Q}(t) = (Q_1(t), Q_2(t))$ denote the queue lengths at time t, and note that the arrival processes remain mutually independent. If we assume that both service rates and the arrival rate of the discretionary customer are nonzero then it follows that $\mathbf{Q}(\cdot)$ is an irreducible Markov process on the state space $Z_+ \times \{0, ..., m\}$. Given some large level ℓ, our rare event of interest for this policy will be the event that we reach any state in the set $F_\ell \equiv \{(x, y) \in Z_+ \times \{0, ..., m\} \mid x \geq \ell\}$, where x and y represent the total number of customers in Q_1 and Q_2, respectively. In other words, we are analyzing large deviations of the total number of customers at queue 1.

There are several variations of this policy. The system is often modelled with only one arrival stream (what we refer to here as the discretionary customer arrivals). Our consideration of this policy is motivated by the work of Teh and Ward [16] in which they establish necessary and sufficient conditions for which the threshold policy with no dedicated customers asymptotically exhibits complete resource pooling in the heavy traffic limit. That is, under the conditions given in [16], the network behaves like a single server queue with total arrival λ_T and service rate $\mu_1 + \mu_2$. For more details on this policy and related works, the reader is asked to refer to [16].

For this policy and model, we will seek to prove the following conjectures.

Conjecture 1 *If $\mu_1 > \lambda_1 + \lambda_T b_m$, then*

$$\mathrm{P}_\pi\{Q_1(t) = \ell, Q_2(t) = k\} \sim \frac{f_3}{d} \frac{\varphi(j)}{h(\ell, j)} \tag{64}$$

where all of the components would be defined in an analogous fashion to the results

given in Chapter 5.

Conjecture 2

$$\mathrm{E}_\sigma[T_\ell] \sim f_3 \alpha^\ell \text{ as } \ell \to \infty,$$

where T_ℓ is the first time that there are ℓ customers in queue 1, and σ is any initial point of the form $(0, y)$.

Preliminary Analysis of the Threshold Model

Step 1. Again, our initial Markov process is the joint queue length process at the two queues, as shown in Figure 5. To ensure irreducibility we assume that λ_T, μ_1 and μ_2 are all nonzero. Furthermore, our assumption that $\lambda + \mu < \infty$ implies that the elements on the diagonal of the generator are bounded. As done before, rescaling time such that $\lambda + \mu = 1$ allows us to interpret our system parameters as transition probabilities. The requirement of positive recurrence is met by assuming $\lambda_1 + b_m \lambda_T < \mu_1$, where b_m is the steady state probability that queue 2 is full. For most of our analysis, we will keep the same notation as in Chapter 4. For example, after uniformization, the state of the discrete-time Markov chain at time n is again denoted by $\mathbf{Q}[n] = (Q_1[n], Q_2[n]), n \in Z_+$, while K_Q and π_Q denote the transition kernel and stationary probability measure of Q on S_Q.

Steps 2&3. To have $\tilde{W}$ diverge along the approach to F_ℓ, we define $\tilde{W}[n] = Q_1[n]$. Since the Markov chain $\{Q_2[n], n \geq 0\}$ will have some stationary distribution along the approach to F_ℓ, we define the Markovian component $\hat{W}$ as the queue length at queue 2. Thus we have $\mathbf{W}[n] = (\tilde{W}[n], \hat{W}[n]) = (Q_1[n], Q_2[n])$ where the state space is denoted by $S \equiv \tilde{S} \times \hat{S} \equiv Z_+^2$.

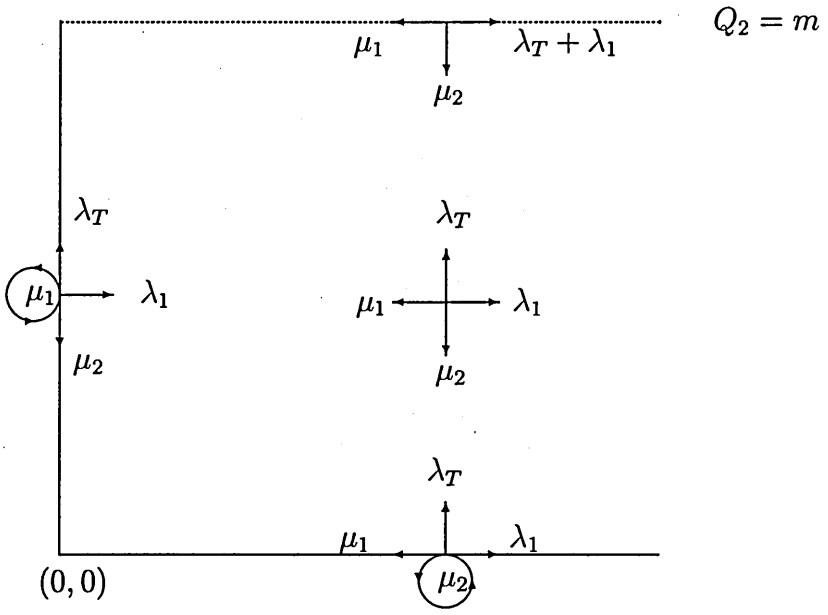

Figure 5: Transition probabilities for the Threshold policy

Step 4. We wish to again partition our state space S into the interior Θ and the boundary $\triangle$ of S such that the transition structure that determines the next stable component and the change in the additive component from time n to $n+1$ depends only on $\hat{W}[n]$. That is, given a plot of the W chain similar to Figure 5, the transition probability structure should not differ along any horizontal line excluding $\triangle$, which also must be defined. To obtain our desired Markov additive structure, we use the transition probabilities of W corresponding to when queue 1 is nonempty. More formally, we define $\triangle \equiv \{(x,y) \in S_Q \mid x = 0)$, and our free chain is constructed by using the transition probabilities of W corresponding to the states in $S \backslash \triangle$. An illustration of the resulting process would be the equivalent of removing the y-axis from Figure 5. In doing so, we enlarge the state space to $\mathbb{Z} \times \hat{S}$ for the free chain and our additive component now has no barriers to its drift in either direction. The states added in this process will be denoted by $\blacktriangle \equiv ((\mathbb{Z} \times \hat{S}) \setminus S)$.

Step 5. For a function h to be harmonic for the free process, it must be the case that $K^{\infty}h = h$, where h is of the form $h(x,y) = \alpha^x \hat{h}_y$, for $(x,y) \in S$. In this manner, using the transition structure of the free process, we obtain the following constraints on h:

$$\alpha^x \hat{h}_y = \begin{cases} \mu_2 \alpha^x \hat{h}_y + \mu_1 \alpha^{x-1} \hat{h}_y + \lambda_T \alpha^x \hat{h}_{y+1} + \lambda_1 \alpha^{x+1} \hat{h}_y & \text{for } y = 0, \\ \mu_2 \alpha^x \hat{h}_{y-1} + \mu_1 \alpha^{x-1} \hat{h}_y + \lambda_T \alpha^x \hat{h}_{y+1} + \lambda_1 \alpha^{x+1} \hat{h}_y & \text{for } 0 < y < m, \\ \mu_2 \alpha^x \hat{h}_{m-1} + \mu_1 \alpha^{x-1} \hat{h}_m + (\lambda_1 + \lambda_T) \alpha^{x+1} \hat{h}_m & \text{for } y = m, \end{cases} \tag{65}$$

where $\hat{h}_y = \hat{h}(y)$ for $(x,y) \in Z \times \hat{S}$, and $\alpha > 1$. Without loss of generality we can rescale the harmonic functions such that $h(0,0) = 1$. In doing so, we can numerically solve for our harmonic function in the following fashion. Observe that with $\hat{h}_0 = 1$, the first constraint in (65) leads to the following expression for $\hat{h}_1$ dependent only on α.

$$\lambda_T \hat{h}_1 = 1 - \mu_2 - \mu_1 \alpha^{-1} - \lambda_1 \alpha - 1. \tag{66}$$

We can then use the remaining constraints to iterate and obtain expressions for $\hat{h}_2, ..., \hat{h}_m$, also dependent on α. In this fashion, by using both the second and third equations in (65), we obtain two equations dependent on $\hat{h}_m$ and α which can be solved simultaneously to finally obtain α.

Alternatively, to solve for our harmonic function, we can also take advantage of the structure present in our constraints. Multiplying the second constraint of (65) by α^{-1} and rearranging terms yields the following difference equation,

$$\lambda_T h_{y+1} + (\alpha^{-1} + \lambda_1 - 1) h_y + \mu_2 h_{y-1} = 0 \tag{67}$$

which is known to have solution form $\hat{h}_y = c r_1{}^y + (1-c) r_2{}^y$, where r_1 and r_2 are roots

of the binomial equation

$$\lambda_T x^2 + (\lambda_1 \alpha + \mu_1 \alpha^{-1} - 1)x + \mu_2 = 0. \tag{68}$$

Step 6. As done in our JSEW system analysis, many of Conditions 1-12 were already taken into consideration while constructing our three processes and hence have already been satisfied. For example, by construction, we have already ensured that our free process is a Markov additive process with $K_W((\tilde{x}, \hat{x}); (\tilde{y}, \hat{y})) = P^{\infty}(\hat{x}; (\tilde{y} - \tilde{x}, \hat{y}))$, and that the transition probabilities of W^{∞} and W agree for jumps between states in the interior. Hence, Conditions 1 and 2 have been met.

Condition 3 requires that the transition probabilities of W^{∞} and W also agree for jumps from states on the boundary to states in the interior. This condition is also satisfied since for any $y \in \Theta$ we have

$$K(x, y) = K^{\infty}(x, y) = \begin{cases} \lambda_1 + \lambda_T & \text{for } x = (0, m), \\ \lambda_1 & \text{for } x = (0, i),\ i < m. \end{cases}$$

In step 5 of the algorithm, we have already chosen a positive function h which is harmonic for the free process. Condition 5 specifies that we use this harmonic function to construct the twisted free process $\mathcal{K}^{\infty}(x, y)$ from the free process. The resulting transition probabilities are shown in Figure 6.

Condition 6, which requires that the marginal chain $\hat{\mathcal{W}}^{\infty}[n], n \in \mathbb{Z}_+$, with transition kernel $\hat{\mathcal{K}}^{\infty}$ has a stationary probability distribution $\varphi(\cdot)$, is trivial since $Q_2[n]$ is finite. It can be shown that

$$\varphi(j) = \left(\frac{h_j}{h_0}\rho_2\right)^j \phi(0), \quad 0 < j \leq m$$

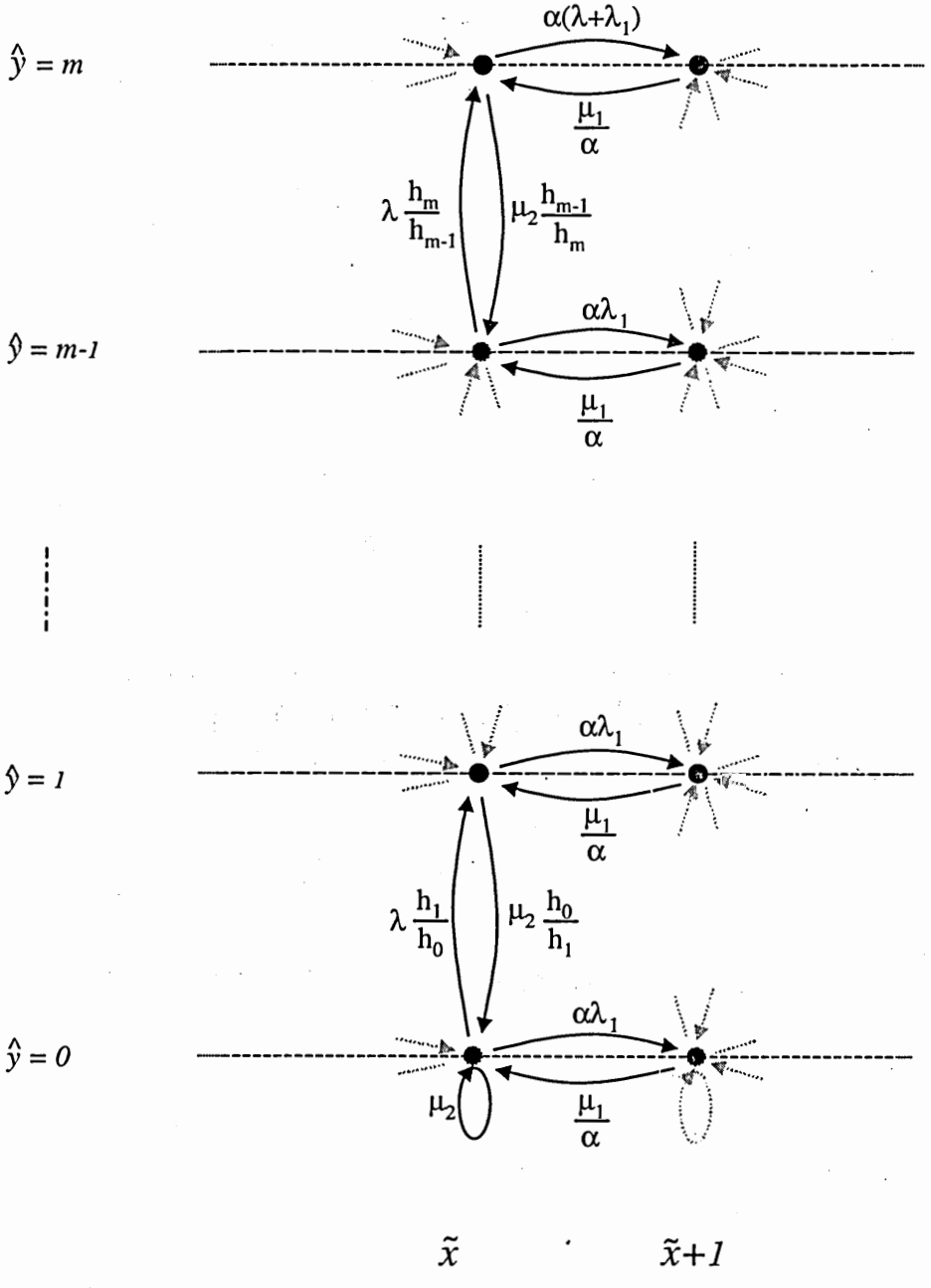

Figure 6: Twisted free process transition probabilities for the Threshold policy

$$\varphi(0) = \left[\sum_{0 \le j \le m} \left(\frac{h_j}{h_0} \rho_2 \right)^j \right]^{-1}.$$

We momentarily jump ahead and note that, Conditions 8,10 and 11 can be trivially checked for this system since Δ is finite. Hence, to complete the analysis of the threshold policy, Conditions 7, 9, and 12 must be satisfied. However, given that we have not explicitly solved for the harmonic function, showing that these conditions hold in general is not straightforward.

Future Work

As mentioned above, three conditions must still be satisfied to complete the analysis of the threshold policy. Moreover, there is also work to be completed with respect to choosing a proper harmonic function (Condition 4). More specifically, recall that the solution form of $\hat{h}_y$ was given to be $\hat{h}_y = cr_1{}^y + (1-c)r_2{}^y$, where r_1 and r_2 are roots of the binomial equation given in (68). To ensure that the roots are always real-valued, we must have $(\lambda_1\alpha + \mu_1\alpha^{-1} - 1)^2 > 4\lambda_2\mu_2$, which remains to be shown. It also remains to be shown that $\alpha > 1$.

It is also shown in [16] that the threshold policy with no dedicated customers is asymptotically optimal in that it exhibits complete resource pooling in the heavy traffic limit. For any given set of parameters, as the threshold m grows large, we would expect the system with dedicated customers to exhibit similar behavior. For a system with a single server queue with total arrival $\lambda_1 + \lambda_T$ and service rate $\mu_1 + \mu_2$, denote the traffic intensity parameter by α_0^{-1}. As m grows large, we expect that α, the rough asymptotics for the threshold network, would converge to α_0. From numerical examples performed thus far, we have observed that the value of α indeed seemed to converge to α_0, and did so very quickly. Hence as m grows large, not only will our

system mimic the $M/M/1$ system described above, but it also appears that the value of m does not have to be "very large" for such behavior to be observed. This concept will also be explored more formally in our future research.

REFERENCES

[1] BROWN, L. Asymptotic Behavior of an overloading queueing network with resource pooling.
Ph.D. dissertation, Georgia Institute of Technology, 1998.

[2] CHANG, W. Polling networks with limited service policies, and wavelet-based information fusion and dimension reduction.
Ph.D. dissertation, Georgia Institute of Technology, 2002.

[3] DAI, J.G. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Annals of Applied Probability* **5**, 49–77, 1995.

[4] DAI, J.G. Stability of open multiclass queueing networks via fluid models. *The IMA volumes in mathematics and its applications* **71**, *Stochastic networks*, 71–90, 1995.

[5] DAI, J.G. AND G. WEISS. Stability and instability of fluid models for re-entrant lines. *Mathematics of Operations Research* **21**, 115–134, 1996.

[6] FOLEY, R.D. and D. R. MCDONALD, Join the Shortest Queue: Stability and Exact Asymptotics. *Annals of Applied Probability* **26**, 105–119, 2001.

[7] FOSCHINI, G. J. Equilibria for diffusion models of pairs of communicating computers - symmetric case. *IEEE Transactions on Information Theory* **28**, 273–284, 1982.

[8] FOSCHINI, G. J. AND J. SALZ A basic dynamic routing problem and diffusion. *IEEE Transactions on Communications* **26**, 320–327, 1978.

[9] HOUCK, D.J. Algorithms for routing calls to parallel queueing systems. *Unpublished Work*, 1982.

[10] MCDONALD, D. Asymptotics of first passage times for random walk in a quadrant. *Annals of Applied Probability* **9**, 110–145, 1999.

[11] MCDONALD, D. AND S.R.E. TURNER. Comparing Load Balancing Algorithms for Distributed Queueing Networks. *Field Institute Communications*, **28**, 109–133, 2000.

[12] MENICH, R. P. AND RICHARD SERFOZO. Resource Allocation in Parallel Processing Systems.
Ph.D. Dissertation, Georgia Institute of Technology, 1991.

[13] MEYN, S.P. AND R. L. TWEEDIE. Markov Chains and Stochastic Stability. Springer, New York, 1993.

[14] REIMAN, M. I. Some Diffusion Approximations with State Space Collapse. *Modelling and Performance Evaluation Methodology, Springer-Verlag, 1984, pp. 209-240.*

[15] SHWARTZ, A. AND A. WEISS. Large Deviations for performance analysis: queues, communications and computing. *Chapman & Hall* 1995.

[16] TEH, YIH-CHOUNG AND AMY R. WARD. Critical Thresholds for Dynamic Routing in Queueing Networks. *Queueing Systems*, **42**, 297–316, 2002.

[17] TURNER, STEPHEN. Large Deviations for Join the Shorter Queue. *Field Institute Communications*, **28**, 95–105, 2000.

[18] WEBER, R.W. On the optimal assignment of customers to parallel servers. *Journal of Applied Probability* **15**, 406–413, 1978.

[19] WHITT, W., Deciding which queue to join: some counterexamples.
Operations Research **34** 55–62, 1986.

VITA

Jerome Daryl Coombs-Reyes was born and raised in Brooklyn, New York. Pegged as the class clown throughout elementary school, Jerome somehow managed to match his propensity for comedy with exceptional scholastic aptitude. Consequently, he was accepted into and chose to attend the Bronx High School of Science, one of the most prestigious schools in the Big Apple.

After four years of toil in the Bronx, Jerome went on to attend Morehouse College; a small historically Black college in Atlanta, Georgia. During his undergraduate studies, he managed to accumulate numerous academic awards and honors, culminating with the Engineering Student of the Year Award in 1996. In 1997, he graduated Phi Beta Kappa from Morehouse with a B.S. degree in Mathematics. As part of a dual degree program, he also received a B.S. degree in Industrial Engineering from the Georgia Institute of Technology.

Upon completing his undergraduate studies, Jerome decided to stay at Georgia Tech to pursue his Ph.D. It was both his teaching and learning experiences at Morehouse College that had confirmed his desire to become an educator, and hence seek the terminal degree. However, it was also the path of least resistance, and Mr. Coombs-Reyes was admittedly too indolent to move out of Atlanta and/or seek employment elsewhere.

During his tenure as a PhD student, Jerome was the recipient of numerous fellowship awards; including AT&T, NSF, Sloan Foundation, Georgia Tech Presidential fellowship, FACES fellowship program, and the Office of Naval Research. In addition to these accolades, as a true wildlife enthusiast, he is also the only Ph.D. student on record to have ever both (a) pulled over to save an alligator snapping turtle from its death as it attempted to cross an 8 lane interstate, and (b) accumulate a collection of photographs of the squirrels residing on his college campus.

In 1999, Jerome earned his M.S. degree in Operations Research, and in December, 2003, earned a Ph.D. in the same field. Jerome has a natural love for teaching, and a strong desire to become an educator. Accordingly, his post-graduation objective was to pursue a career in academia which emphasized the value of teaching. Besides his research and teaching interests, Jerome also enjoys travel, freelance photography, and writing vitae about himself in the third person.