Join the Shortest Queue: Stability and Exact Asymptotics

R. D. Foley
D. R. McDonald*

July 1998

Key words and phrases: Join the shortest queue, rare events, change of measure, h transform
AMS 1991 subject classifications: Primary 60K25; Secondary 60K20.

Abstract

We consider the stability of a network serving a patchwork of overlapping regions where customers from a local region are assigned to a collection of local servers. These customers join the queue of the local server with the shortest queue of waiting customers. We then describe how the backlog in the network overloads. We do this in the simple case of two servers each of which receives a *dedicated* stream of customers in addition to customers from a stream of *smart* customers who join the shorter queue. There are three distinct ways the backlog can overload. If one server is very fast then that server takes all the smart customers along with its dedicated customers and keeps its queue small while the dedicated customers at the other server cause the overload. We call this the unpooled case. If the proportion of smart customers is large, then the two servers overload in tandem. We call this the strongly pooled case. Finally there is the weakly pooled case where both queues overload but in different proportions. The fact that strong pooling can be attained based on a local protocol for overlapping regions may have engineering significance. In addition, this paper extends the methodology developed in [14] to cover periodicities.

The emphasis here is on *sharp* asymptotics not rough asymptotics as in large deviation theory. Moreover the limiting distributions are for the unscaled process, not for the fluid limit as in large deviation theory. In the strongly pooled case for instance we give the limiting distribution of the difference between the two queues as the backlog grows. We also give the exact asymptotics of the mean time until overload.

*Research supported in part by NSERC grant A4551

1 Introduction

We will be analysing the following generalization of the classical problem of joining the shortest queue. Consider $m < \infty$ exponential servers numbered $1, \ldots, m$ each having an infinite capacity waiting area. The service times at the ith server form a sequence of independent, exponentially distributed random variables with rate $\mu_i > 0$. Each non-empty set of servers $A \subset M \equiv \{1, \ldots, m\}$ has an associated Poisson arrival process of customers with rate $\lambda_A \geq 0$ that join the shortest queue in A with ties broken randomly. We assume that the m sequences of exponential service times and the $2^m - 1$ Poisson arrival processes are mutually independent. To eliminate degeneracies, we assume that it is possible for customers to arrive to each of the servers; i.e., for every server $i \in M$, there exists $A \subset M$ such that $i \in A$ and $\lambda_A > 0$.

In particular, we will be analysing the stability conditions and the asymptotic behavior of this system. We determine the stability conditions for the general system, but our asymptotic results are for the case of $m = 2$ queues. In the case m = 2, we will use the notation $\lambda_i \equiv \lambda_{\{i\}}$ for $i \in \{1, 2\}$ for the arrival rates of customers *dedicated* to a single queue and $\gamma \equiv \lambda_{\{1,2\}}$ for the arrival rate of *smart* customers who join the shorter queue. It is possible to extrapolate from the analysis of the two queue system and develop some general principles for the design of more complicated systems.

As a by-product of analysing the asymptotics of joining the shortest queue, we hope that the methodology developed in [14] and this paper will be considered a viable alternative for analysing the exact asymptotics of other systems. This approach is limited in the sense that, unlike in the more general theory of large deviations [9], it does not handle systems in which the most likely (fluid limit) path to the rare event of interest is non-linear. However, in systems where the most likely path is linear, this approach may be easier and yield stronger results.

2 Results

Our results for the join the shortest queue system are divided into two groups: stability and exact asymptotics. In this paper determining the exact asymptotics of 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.,

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

In some cases, we must deal with periodicities. We extend the above notation to cover this situation and write $f(\ell, k) \sim c(\ell \mod p)\alpha^\ell \chi\{k \equiv \ell \mod p\}$ where χ is the indicator function by making the convention that $0/0$ is 1, and that the asymptotics must hold in ℓ for every k. We are primarily interested in asymptotic behavior of the join the shortest queue, but in order to analyse the asymptotics, we need to know the stability conditions which we now state.

For each non-empty subset $A \subset M$, define the traffic intensity on A as

$$\rho_A = \frac{\sum_{B \subset A} \lambda_B}{\mu_A} \tag{1}$$

where $\mu_A = \sum_{i \in A} \mu_i$. The numerator of (1) represents the total arrival rate of customers that must be serviced by the servers in A, and the denominator represents the total service rate of the servers in A. Note that the total rate at which customers are accepted to the servers in A may be greater than the numerator of (1) since other customers may be allowed to be served by some or all of the servers in A. Also, note that this total load represented by the numerator of (1) is not necessarily spread equally over the servers in A since some or all of the customers may be restricted to a subset of A. Let $\rho_{\max}$ represent the most heavily loaded subset; i.e., $\rho_{\max} \equiv \max_{A \subset M}\{\rho_A\}$.

Let $Q(t) = (Q_1(t), \ldots, Q_m(t))$ be the queue lengths at time t. Since we have assumed that for every server $i \in M$, there exists $A \subset M$ such that $i \in A$ and $\lambda_A > 0$, it is easy to see that $Q(\cdot)$ is an irreducible Markov process on the state space $\mathbb{Z}_+^m$ where $\mathbb{Z}_+ \equiv \{0, 1, 2, \ldots\}$.

Theorem 1. *Consider the generalized join the shortest queue system.*

1. *If $\rho_{\max} > 1$, then the Markov process $Q(\cdot)$ is transient.*
2. *If $\rho_{\max} = 1$, then the Markov process $Q(\cdot)$ is either transient or null recurrent.*

3. *If $\rho_{\max} < 1$, then the Markov process is positive recurrent and has a stationary probability distribution π. Furthermore, the expected number in the system in equilibrium is bounded above by $-m + \mu_M/c$ where c is given in (7).*

Remark 1. *When $\rho_{\max} = 1$, there are examples of both transience and null recurrence. Consider a system with only dedicated customers such that $\lambda_i = \mu_i = 1/(2m)$ for $i = 1, 2, \ldots, m$. The queue length process behaves like the absolute value of an m-dimensional simple symmetric random walk which is null recurrent if m is 1 or 2 and transient for $m \geq 3$.*

Let $\Pr_\pi\{\cdot\}$ denote the probability measure corresponding to the stationary distribution of $Q(t)$. Assume that there are $m = 2$ queues and a unique set $A \subset M$ such that $\rho_{\max} = \rho_A < 1$. (For technical reasons, we can only analyse the asymptotics when there is a unique set A with $\rho_A = \rho_{\max}$.) Note that with $m = 2$, $\rho_1 = \lambda_1/\mu_1$, $\rho_2 = \lambda_2/\mu_2$, and $\rho \equiv \rho_M = (\lambda_1 + \lambda_2 + \gamma)/(\mu_1 + \mu_2)$. Let T_ℓ denote first time that there are ℓ or more customers in the system. The following 3 theorems summarize the results in Theorems 9–14

Theorem 2 (strongly pooled servers). *If $\rho > \max\{\rho_1, \rho_2\}$ and $\gamma > |\rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2)|$ then*

$$\mathrm{E}[T_\ell \mid Q(0) = (0,0)] \sim g^{-1}\rho^{-\ell}/(\lambda_1 + \lambda_2 + \gamma + \mu_1 + \mu_2)$$

where the constant g is defined in (36). *Moreover, for non-negative integers k and l*

$$\Pr_\pi\{Q_1(t) + Q_2(t) = \ell, Q_1(t) - Q_2(t) = k\} \sim 2\frac{f(0)}{\tilde{d}_1}\rho^\ell \varphi(k)\chi\{k = \ell \mod 2\}$$

where $\tilde{d}_1 = \mu_1 + \mu_2 - (\lambda_1 + \lambda_2 + \gamma)$, $f(0)$ is defined in (36), *and*

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

Finally,

$$\Pr\{Q_1(T_\ell) - Q_2(T_\ell) = k \mid Q(0) = (0,0)\} \sim \quad 2\varphi(k)\chi\{k = \ell \mod 2\}.$$

Remark 2. *In (2) and (3), the 2 and the indicator function χ are a consequence of periodicity $p = 2$. In particular, note that $Q_1(t) + Q_2(t) = \ell$ and $Q_1(t) - Q_2(t)$ are either both odd or both even.*

Theorem 3 (weakly pooled servers). *If $\rho > \max\{\rho_1, \rho_2\}$ and $\gamma \leq |\rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2)|$, then*

$$\mathrm{E}[T_\ell \mid Q(0) = (0,0)] \sim g^{-1}\rho^{-\ell}/(\lambda_1 + \lambda_2 + \gamma + \mu_1 + \mu_2)$$

where g is given in (36). *Moreover*

$$\Pr_\pi\{Q_1(t) + Q_2(t) = \ell\} \sim \frac{f(0)}{\tilde{d}_1}\rho^\ell.$$

where $\tilde{d}_1$ is the same as in Theorem 2 and $f(0)$ is given in (36).

If $\rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2) > \gamma$ then

$$\lim_{\ell\to\infty}\left(\frac{Q_1(T_\ell)}{\ell}, \frac{Q_2(T_\ell)}{\ell}\right) = \left(\frac{\lambda_1\rho^{-1} - \mu_1\rho}{\tilde{d}_1}, \frac{(\lambda_2 + \gamma)\rho^{-1} - \rho\mu_2}{\tilde{d}_1}\right) \tag{3}$$

If $\rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2) < -\gamma$ then

$$\lim_{\ell\to\infty}\left(\frac{Q_1(T_\ell)}{\ell}, \frac{Q_2(T_\ell)}{\ell}\right) = \left(\frac{(\lambda_1 + \gamma)\rho^{-1} - \mu_1\rho}{\tilde{d}_1}, \frac{\lambda_2\rho^{-1} - \rho\mu_2}{\tilde{d}_1}\right) \tag{4}$$

If $|\rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2)| = \gamma$, then

$$\lim_{\ell\to\infty}\left(\frac{Q_1(T_\ell)}{\ell}, \frac{Q_2(T_\ell)}{\ell}\right) = \left(\frac{1}{2}, \frac{1}{2}\right) \tag{5}$$

Theorem 4 (unpooled servers). *If $\rho \equiv \rho_M < \max\{\rho_1, \rho_2\}$ and, without loss of generality, we assume $\rho_1 > \rho_2$) then*

$$\mathrm{E}[T_\ell \mid Q(0) = (0,0)] \sim g^{-1}\rho_1^{-\ell}/(\lambda_1 + \lambda_2 + \gamma + \mu_1 + \mu_2)$$

where g is defined in (41).

$$\Pr_\pi\{Q_1(t) + Q_2(t) = \ell, Q_2(t) = k\} \sim \frac{f}{\mu_1 - \lambda_1}\rho_1^{\ell-k}(1 - \frac{\lambda_2 + \gamma}{\mu_2})\left(\frac{\lambda_2 + \gamma}{\mu_2}\right)^k$$

and f is given in 40. Finally,

$$\Pr\{Q_2(T_\ell) = k \mid Q(0) = (0,0)\}\} \sim \quad c^{-1}\rho_1^k\mu(0,k). \tag{6}$$

where $c = \sum_{k=0}^{\infty}\rho_1^k\mu(0,k)$ and μ is a probability measure defined near (21) *and specifically for the unpooled case in Theorem 13.*

In addition, our analysis shows that there are basically four different ways in which the total number of customers in the system increases to some large level ℓ, depending on system parameters. Initially we guessed that there were only three ways depending on which was largest ρ_1, ρ_2, or ρ. (Thus far, we have not analysed the case of ties.) If ρ_1 were the largest, then we guessed that the system would be unpooled, and the most likely approach would bounce along the horizontal axis. Similarly, if ρ_2 were the largest, then we guessed that the most likely approach was along the y-axis. If ρ were the largest, we first guessed that the servers would pool and the most likely approach would be up the diagonal. However, this conjectured approach when ρ is largest was false. This pooled case splits into two subcases, weak and strong, depending upon system parameters. If ρ is the largest and $\gamma > |\rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2)|$, then the most likely approach does hug the diagonal, and we call this the strongly pooled case. However, if ρ is the largest and $\gamma < |\rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2)|$, then the boundary is approached by drifting in a direction with slope given by (3) if Q_2 lags behind Q_1 and by (4) if Q_1 lags behind Q_2. In these cases, we say the servers are weakly pooled. Note that this fourth way actually contains a whole spectrum of possible drift directions. Furthermore, in the strongly and unpooled cases, there is a restoring force which keeps the process close to the (fluid limit) approach path; i.e., one of the axes or the diagonal. However, when the servers are weakly pooled, there will be no restoring force towards the drift direction. Interestingly, the approach can be along the diagonal and *not* be strongly pooled. If $\gamma = |\rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2)|$ in the pooled case, the boundary is approached with slope 1, but the difference between the two queues behaves like a null recurrent Markov chain.

The weakly pooled case is in fact predicted by large deviation theory. A large deviation path where Q_2 lags behind Q_1 corresponds to the simultaneous large deviation of two independent $M|M|1$ queues where Q_1 has arrival rate λ_1 and service rate μ_1 while Q_2 has arrival rate $\lambda_2 + \gamma$ and service rate μ_2. During this large deviation there will be some time t when $Q_1(t)$ exceeds $a\ell$ while $Q_2(t)$ exceeds $(1-a)\ell$ where $0 < a < 1$. By Section 11.2 in [18], the cost or action associated with a trajectory of the first queue from 0 to $a\ell$ in time t plus the action associated with a trajectory of the second queue from 0 to $(1-a)\ell$ in time t is $t[L_1(a\ell/t) + L_2((1-a)\ell/t)]$ where

$$\begin{aligned} L_1(c) &= \left(c\log(\frac{c + \sqrt{c^2 + 4\lambda_1\mu_1}}{2\lambda_1}) + \lambda_1 + \mu_1 - \sqrt{c^2 + 4\lambda_1\mu_1}\right) \\ L_2(c) &= \left(c\log(\frac{c + \sqrt{c^2 + 4(\lambda_2+\gamma)\mu_2}}{2(\lambda_2+\gamma)}) + (\lambda_2 + \gamma) + \mu_2 - \sqrt{c^2 + 4(\lambda_2+\gamma)\mu_2}\right) \end{aligned}$$

If we first minimize this action on t and then on a we find $(a, 1-a)$ is precisely the pair given by expression (3) (see [19] for more details). However, large deviation theory seems to be unable to distinguish between weak pooling when (5) holds and strong pooling.

We believe this trichotomy among the weakly, strongly, and unpooled cases has engineering implications for properly balancing the load in a network. A priori, protocols like joining the shortest queue are implemented in order to keep all the servers busy rather than let some idle while others overload. Consider a system with only dedicated customers; e.g. two call centers one serving the western half of a region and the other the eastern. Typically, the system would be unbalanced and one of the operators would overload more frequently which corresponds to the unpooled case with a rate parameter $\max\{\rho_1, \rho_2\} > \rho$. To equalize the load, the operators could be moved to one location creating an $M/M/2$ queue with the best possible rate parameter ρ. An alternative solution that might be less costly and still obtain this best possible rate parameter ρ would be to leave the operators at their current locations and route all calls to the shorter queue. A third solution that might be the least costly is to leave the operators in their current locations and most of the customers remain dedicated to their current operator, but only route the customers in a small portion of the region to the shorter queue. The size of the region allowing rerouting should be large enough to achieve pooling and the best rate ρ. If it is also desired that the queues be roughly equal when overloading occurs, the region allowing rerouting should be large enough to achieve strong pooling. Note that even though several alternatives achieve this best rate ρ and have the same rough asymptotics, the exact asymptotics may differ since the coefficient may differ.

There is a huge literature. Flatto and McKean [10] investigated this system when $\lambda_1 = \lambda_2 = 0$ and $\mu_1 = \mu_2$. Using analyticity arguments they obtained an exact solution for the generating function of the stationary distribution $\pi(x,y)$. Adan, Wessels and Zijm [1] used a compensation procedure to represent the stationary distribution as an infinite sum of product measures in the asymmetric case when $\mu_1 \neq \mu_2$ but $\lambda_1 = \lambda_2 = 0$. $\gamma < \mu_1 + \mu_2$ was required for stability. Knessl, Matkowsky, Schuss and Tier [12] used a heuristic technique to give the stationary distribution for the model in this paper. The paper by van Houtum, Adan, Wessels, Zijm [20] studies the problem of assigning component types to the production/assembly of printed circuit boards on parallel insertion machines, which is naturally modelled as the join the shortest queue model with dedicated customers.

Shwartz and Weiss [18] developed a large deviation theory for Markov jump processes with a boundary and applied it to deviations of the number of customers in a join the shortest queue network like the one studied here. Shwartz and Weiss [17] also gave the first results on the exact asymptotics of backlogs in a network (describing the bathroom problem) using time reversal methods. The existence of a large deviation principle for this system is given in Dupuis and Ellis [9] and in Atar and Dupuis [2]. Turner [19] has used this large deviation principle to analyse overloads of the backlog in the system (without repacking) studied here and the system where the waiting room for each queue is of size C. A similar analysis of a system of $M|M|\infty$ servers can be found in Alanyali and Hajek [3], [4]. In [13] large deviations of a chosen queue in the join the shortest queue network are analysed, but the key steps (C.10 and C.12) are not given as they are in here.

Turner's work is motivated in part by his circle problem associated with alternate routing in a service network. In such a network a customer might be routed to the least busy node available to him. In order to balance the load on the network the best solution is to make every node available to every customer but this may not be practical. The question is whether resource pooling can be obtained by allowing each customer a small number of available servers. Louise Brown [6] has studied Turner's circle model for three servers and has found conditions for such strong pooling. Brown also gives ranges of parameter values where one obtains weak pooling or no pooling.

3 Analysis of Stability

We break up the analysis of stability into a series of lemmas. In proving stability we use an approach motivated by results in Markov decision processes. The fluid limit approach taken in [8] is more general, but does not yield bounds. First, we determine conditions for transience and null recurrence.

Lemma 1. *Consider the generalized join the shortest queue system. If $\rho_{\max} > 1$, then the Markov process $Q(\cdot)$ is transient. If $\rho_{\max} \geq 1$, then the Markov process $Q(\cdot)$ is either transient or null recurrent.*

Proof. For the proof of transience, note that the number of customers in A is stochastically larger than the number of customers in an $M/M/1$ queue with arrival rate $\sum_{B \subset A} \lambda_B$ and service rate μ_A. If $\rho_{\max} > 1$, then there exists $A \subset M$ with $\rho_A > 1$. Thus, the number of customers in A is stochastically larger than the number in an M/M/1 queue with traffic intensity greater than one; i.e., the number of customers in A diverges, and $Q(\cdot)$ is transient. Note that this argument with a slight modification also shows that if $\rho_{\max} = 1$, then $Q(\cdot)$ is either transient or null recurrent. □

In determining recurrence, we construct a related system with the same m servers, but a different arrival process. In the new system which only has dedicated arrivals, the arrival process to server i will be a Poisson process with rate α_i. To differentiate between the two systems, call the original system the λ-system, and the new system the α-system. The α-system will have the same service rates μ_i, but each server has a dedicated stream of

Poisson arrivals with rate $\alpha_i = \mu_i - c$ where c is given by

$$c \equiv \min_{A \subset M} \{(1-\rho_A)\mu_A/|A|\} \tag{7}$$

$$= \min_{A \subset M} \{(\mu_A - \sum_{B \subset A} \lambda_B)/|A|\} \tag{8}$$

and $|A|$ denotes the number of servers in A. Thus, c measures the average unused capacity per server of the most heavily loaded subset. The reason for selecting the α_i's so that $\mu_i - \alpha_i = c$ is that it will be needed in Lemma 3. Note that, $c > 0$ and

$$\mu_A - \sum_{B \subset A} \lambda_B - c|A| \geq 0. \tag{9}$$

We construct this related system in two steps. First, the arrival streams of customers that choose the shortest queue among sets of queues are decomposed into independent Poisson processes and assigned to particular servers. More precisely, the Poisson arrival process with rate λ_A is decomposed into $|A|$ independent Poisson processes with rates $\lambda_A(i) \geq 0$ and assigned to server i for $i \in A$. Clearly, we need $\lambda_A = \sum_{i \in A} \lambda_A(i)$. The arrival process to server i is the superposition of independent Poisson processes and has rate $\sum_{A \subset M} \lambda_A(i)$. The first part of the following lemma shows that for every $i \in M$, we have $\sum_{A \subset M} \lambda_A(i) \leq \mu_i - c$. The second step in constructing the related system is to note that the arrival rate to each of the queues can be increased as necessary so that $\alpha_i = \mu_i - c$ for $i \in M$. Note that we can think of the α-system as removing the customer's ability to choose the shortest queue in the λ-system and possibly adding additional customers.

Lemma 2. *If $\rho_{\max} < 1$, then there exists $\lambda_A(i) \geq 0$ for $i \in A \subset M$ satisfying the following:*

$$\lambda_A = \sum_{i \in A} \lambda_A(i) \text{ for } A \subset M, \text{ and} \tag{10}$$

$$\sum_{A \subset M} \lambda_A(i) \leq \mu_i - c > 0 \text{ for } i \in M \tag{11}$$

where c as defined in (7) is positive. Furthermore, the α_i satisfy the following:

$$\alpha_i \geq \sum_{A \subset M} \lambda_A(i) \text{ for } i \in M, \text{ where} \tag{12}$$

$$\alpha_i = \mu_i - c, i \in M. \tag{13}$$

Proof. To prove that the $\lambda_A(i)$'s exist, we reformulate the problem as the network flow problem depicted in Fig. 1. There is a source node s, a first column consisting of $2^m - 1$ nodes corresponding to and labelled by each non-empty subset of M, a second column of m nodes corresponding to each of the servers and labelled $1, \ldots, m$, and a sink node t. Note that $\{i\}$ and i are different nodes. There is an arc from the source s to each Node A with capacity λ_A for every non-empty $A \subset M$ which corresponds to the constraint (10). For each non-empty $A \subset M$, there is an infinite capacity arc to Node i for every $i \in A$. For each $i \in M$, there is an arc from Node i to the sink t with capacity $\mu_i - c$ which corresponds to the constraint (11). We claim the maximum flow through this network is $\sum_{A \subset M} \lambda_A$, and that this clearly implies the existence of the $\lambda_A(i)$'s satisfying (10) and (11).

The Max-Flow Min-Cut Theorem states that the maximum flow from s to t is equal to the minimum capacity of all cuts separating the source and destination. A cut is a set of arcs which when removed partitions the nodes into two sets L and R where L is the set of nodes accessible from the source and R is the set of nodes accessible to the sink. Since we are looking for the minimum capacity on a cut set, we do not need to include any of the infinite capacity arcs in a cut set. For such a cut set C, let $A \equiv \{(i,t) \mid (i,t) \in C\}$. For C to be a cut set, B must be a subset of C where $B \equiv \{(s,B) \mid B \not\subset A\}$. Hence, the capacity of this cut C is at least

$$\begin{aligned}\sum_{i \in A}(\mu_i - c) + \sum_{B \not\subset A} \lambda_B &= \sum_{i \in A}(\mu_i - c) + \sum_{B \subset M} \lambda_B - \sum_{B \subset A} \lambda_B \\ &= \mu_A - c|A| - \sum_{B \subset A} \lambda_B + \sum_{B \subset M} \lambda_B \\ &\geq \sum_{B \subset M} \lambda_B\end{aligned}$$

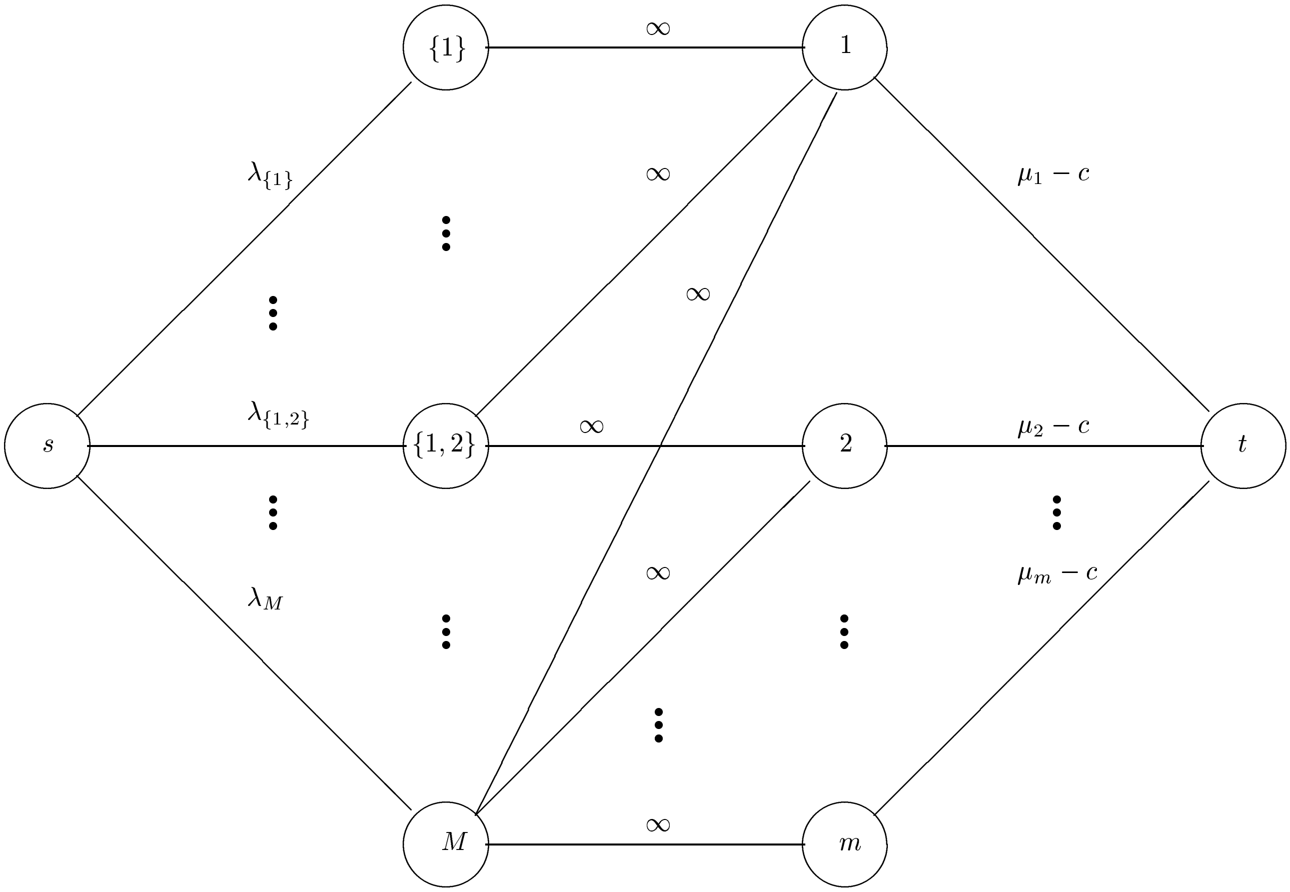

Figure 1: Network Flow Problem.

where the inequality follows from (9). Since the minimum capacity cut is necessarily of this form, we conclude the maximum flow is $\sum_{B \subset M} \lambda_B$.

Let $\lambda_A(i) \geq 0$ denote the flow from A in the first column of Fig. 1 to i in the second column for some solution attaining the maximum flow. By construction the $\lambda_A(i)$'s satisfy (10) and (11) and (12) follows from the definition of α_i's. $\square$

Note that the α-system is positive recurrent if $\rho_{\max} < 1$. The remainder of the stability argument basically shows that the λ-system behaves no worse than the α-system. Define

$$V(x) = \sum_{i \in M} \frac{x_i(x_i+1)}{2(\mu_i - \alpha_i)} \text{ for } x \in \mathbb{Z}_+^m.$$

Let $\{Q[n], n \geq 0\}$ denote the discrete time Markov chain obtained by uniformizing $\{Q(t), t \geq 0\}$ [7, Theorem 4.31]. Thus, Consider the Markov decision problem of minimizing the average number of customers in the system. Two actions α and λ are allowed in each state and correspond to the transition probabilities of the α-system or the λ-system. The equation in the next lemma states that V is the *bias* or *relative value function* for the α-policy which uses action α in all states.

Lemma 3. *If $\rho_{\max} < 1$ then*

$$V(x) = \sum_{i \in M} x_i - \sum_{i \in M} \alpha_i/(\mu_i - \alpha_i) + \mathrm{E}^{\alpha}[V(Q[n+1]) \mid Q[n] = x] \tag{14}$$

$$V(x) \geq \sum_{i \in M} x_i - \sum_{i \in M} \alpha_i/(\mu_i - \alpha_i) + \mathrm{E}^{\lambda}[V(Q[n+1]) \mid Q[n] = x] \tag{15}$$

where the superscript λ or α denotes whether the conditional expectations use the transition probabilities from the λ-system or the α-system.

Proof. It is straightforward to show that V satisfies the given discrete time Poisson equation (or even to find V by recursively solving for V in the single server situation). To obtain the inequality, we perform one step of the policy iteration algorithm as described in [16] or most texts on Markov decision processes. That is, for each state we choose the action which minimizes the r.h.s. of (14). In every state, action λ is at least as good as action α since action λ corresponds to allowing some of the customers to choose the shortest among several queues and possibly rejecting some customers. This follows from V being a symmetric, increasing function of x. Note that it is essential that $\mu_i - \alpha_i = c$ in V; otherwise, it would be better in some states for a customer to join the longer queue. □

Proof. Proof of Theorem 1 The first two parts of Theorem 1 follow directly from Lemma 1. Positive recurrence for $\rho_{\max} < 1$ follows by rewriting (15) as

$$\mathrm{E}^{\lambda}[V(Q[n+1]) \mid Q[n] = x] - V(x) \leq \sum_{i \in M} \alpha_i/(\mu_i - \alpha_i) - \sum_{i \in M} x_i$$

and noticing that V is a Foster-Lyapounov drift function for $Q[\cdot]$ since the right hand side is less than or equal to -1 for all but a finite number of states. The average queue length in the α-system is $\sum_{i \in M} \alpha_i/(\mu_i - \alpha_i) = -m + \mu_M/c$. Policy iteration decreases the average cost, i.e. average number of customers in the system; cf. Theorem 4.3(iii) of [16]. □

4 Exact Asymptotics for Rare Events

In this section we develop asymptotic expressions for a Markov chain $\{W[n], n \in \mathbb{Z}_+\}$ satisfying certain assumptions. In Section 5 we will show that $\{Q(t), t \geq 0\}$ can be transformed into such a Markov chain. The results in this section are similar to, and whenever possible we will use results and notation from, [14]. However, there are differences in the assumptions between the two papers, and we describe the differences in this section.

Let $W \equiv \{W[n], n \in \mathbb{Z}_+\}$ be an irreducible Markov chain on a countable state space S with transition kernel K and stationary distribution π. The process W needs to satisfy certain additional assumptions, which are most easily described by positing the existence of two other Markov chains, the *free* chain $W^\infty \equiv \{W^\infty[n], n \in \mathbb{Z}_+\}$ and the *twisted free* chain $\mathcal{W}^\infty \equiv \{\mathcal{W}^\infty[n], n \in \mathbb{Z}_+\}$ satisfying certain conditions. First, we give notation related to each of the three chains, and then we describe the conditions that the chains must satisfy. Note that the superscript ∞ is not being, and will not be, used in this paper to denote an infinite cross product.

Assume $S \subseteq \mathbb{Z}_+ \times \mathbb{Z}^{r-1} \times \hat{S}$ where $\hat{S}$ is some countable set. If $x \in S$, then $x = (\tilde{x}, \hat{x})$ with $\tilde{x} \in \mathbb{Z}^r$, $\tilde{x}_1 \geq 0$ and $\hat{x} \in \hat{S}$. Let $F_\ell \equiv \{x \in S \mid \tilde{x}_1 \geq \ell\} \neq \emptyset$ for $\ell \in \mathbb{Z}_+$. For each n, $W[n] = (\tilde{W}[n], \hat{W}[n])$ where $\tilde{W}[n] \in \mathbb{Z}^r$, $\tilde{W}_1[n] \geq 0$ and $\hat{W}[n] \in \hat{S}$. The state space S will be partitioned into two regions Δ and Θ, which will be referred to as the boundary and interior, respectively. Define $T_\Delta \equiv \inf\{n > 0 \mid W[n] \in \Delta\}$ and $T_\ell \equiv \inf\{n \geq 0 \mid W[n] \in F_\ell\}$.

The free chain $W^\infty[n] = (\tilde{W}^\infty[n], \hat{W}^\infty[n]), n \in \mathbb{Z}_+$ has state space $S^\infty \equiv \mathbb{Z}^r \times \hat{S}$ and transition kernel K^∞. For $x \in S^\infty$, we have $x = (\tilde{x}, \hat{x})$ with $\tilde{x} \in \mathbb{Z}^r$ and $\hat{x} \in \hat{S}$, and $F_\ell^\infty \equiv \{x \in S \mid \tilde{x}_1 \geq \ell\} \neq \emptyset$ for $\ell \in \mathbb{Z}_+$. The state space S^∞ is partitioned into Θ and $\blacktriangle$. Define $T_\blacktriangle^\infty \equiv \inf\{n > 0 \mid W^\infty[n] \in \blacktriangle\}$ and $T_\ell^\infty \equiv \inf\{n \geq 0 \mid W^\infty[n] \in F_\ell^\infty\}$. The chain $\mathcal{W}^\infty[n] = (\tilde{\mathcal{W}}^\infty[n], \hat{\mathcal{W}}^\infty[n]), n \in \mathbb{Z}_+$ has the same state space $S^\infty \equiv \mathbb{Z}^r \times \hat{S}$, but the transition kernel is $\mathcal{K}^\infty$, the twisted kernel defined below. Similarly, $\mathcal{T}_\blacktriangle^\infty \equiv \inf\{n > 0 \mid \mathcal{W}^\infty[n] \in \blacktriangle\}$ and $\mathcal{T}_\ell^\infty \equiv \inf\{n \geq 0 \mid \mathcal{W}^\infty[n] \in F_\ell^\infty\}$. Let $H(x) = \Pr_x\{\mathcal{T}_\blacktriangle^\infty = \infty\}$. Note that $S \subset S^\infty$, $F_\ell \subset F_\ell^\infty$, $\Delta \subset \blacktriangle$, and that F_ℓ and Δ may overlap.

The conditions that we need on the three processes are the following:

C.1: The marginal process $(\tilde{W}_1^\infty[n], \hat{W}^\infty[n])$ is a Markov additive process with

$$\Pr\left\{\tilde{W}_1^\infty[n+1] - \tilde{W}_1^\infty[n] = \tilde{x}_1, \hat{W}^\infty[n+1] = \hat{y} \mid W^\infty[n]\right\}$$
$$= \Pr\left\{\tilde{W}_1^\infty[n+1] - \tilde{W}_1^\infty[n] = \tilde{x}_1, \hat{W}^\infty[n+1] = \hat{y} \mid \hat{W}^\infty[n]\right\} \quad (16)$$

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 that W^∞ is a Markov additive process with

$$K^\infty\big((\tilde{x},\hat{x});(\tilde{y},\hat{y})\big) = P^\infty\big(\hat{x};\big(\tilde{y}-\tilde{x},\hat{y}\big)\big). \quad (17)$$

where for each $\hat{x}$, $P^\infty(\hat{x},(\cdot,\cdot))$ is a probability measure, holds. However, the weaker condition will prove useful in this paper when analyzing weak pooling, and the stronger condition will only prove useful in the joint bottleneck result.

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

$$K^\infty(x,y) = K(x,y) \text{ for } x \text{ and } y \text{ in } \Theta. \quad (18)$$

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

$$K^\infty(x,y) = K(x,y) \text{ for } x \in\Delta \text{ and } y \text{ in } \Theta. \quad (19)$$

This simplifying assumption could be omitted as in [14], but then one needs additional assumptions on the tail of the distribution of the first step from the boundary.

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 or h-transform of the free process W^∞; i.e.,

$$\mathcal{K}^\infty(x,y) = K^\infty(x,y)h(y)/h(x). \quad (20)$$

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

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

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

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

C.9: $\sum_{\hat{x}\in\hat{S}} \varphi(\hat{x})\hat{h}(\hat{x}) < \infty$.

C.10: Define the measure $\lambda(x) \equiv \pi(x)h(x)\chi\{x \in\Delta\}$ and the marginal measure $\hat{\lambda}(\hat{y}) \equiv \sum_{\tilde{y}} \lambda(\tilde{y},\hat{y})$. We need $\sum_x \lambda(x)\chi\{K(x,\Theta) > 0\} < \infty$.

C.11: Let $Y_\ell(\hat{y}) \equiv \{x \in S^\infty \mid \tilde{x}_1 = \ell, \hat{x} = \hat{y}\}$. 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: 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

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

which is Condition (V3) of [15]. Moreover we assume that $\sum_{\hat{y}} \hat{V}(\hat{y})\hat{\lambda}(\hat{y}) < \infty$.

Note that C.12 implies $\lambda(\Delta) < \infty$ and this implies C.10; however, C.12 will not be needed in Theorem 5. If $\hat{h}$ is bounded and $\sum_{x\in\Delta} \lambda(x) < \infty$, then C.10 and C.12 hold.

Now we need to determine the periodicity of the twisted free process. Let T be the first time when $\hat{\mathcal{W}}^\infty$ returns to a fixed state. Define p to be the largest integer such that the support of $\tilde{\mathcal{W}}_1^\infty[T] - \tilde{\mathcal{W}}_1^\infty[0]$ is a multiple of p. By the argument in [7, Corollary 10.2.24], this period p is independent of the state chosen, and we say that $\mathcal{W}^\infty$ *has period* p.

Fix some reference state $\delta = (\tilde{\delta}, \hat{\delta}) \in \Delta$. Define

$$A_j = \{\hat{z} : \mathcal{J}^n((\tilde{\delta}_1, \hat{\delta}), (\tilde{\delta}_1 + s, \hat{z})) > 0 \text{ for some } n, \text{ where } s \overset{p}{=} j\}$$

and $a \overset{p}{=} b$ means $a = b \mod p$. In Section 4.1, we show that $A_0, \ldots, A_{p-1}$ partitions $\hat{S}$. It will be convenient to know the location of any state; hence, for $\hat{z} \in \hat{S}$, define $A(\hat{z}) = j$ iff $\hat{z} \in A_j$. Notice that $A(\hat{\delta}) = 0$.

Let μ denotes the stationary distribution of $(\tilde{\mathcal{W}}_1^\infty[\mathcal{T}_\ell^\infty] - \ell, \hat{\mathcal{W}}^\infty[\mathcal{T}_\ell^\infty]), \ell = 0, 1, \ldots$. The stationary overshoot distribution μ is obtainable by fast simulation since $\tilde{\mathcal{W}}_1^\infty$ drifts to infinity. Note that even when the overshoot is always zero, $\mu(0, \hat{y})$ and $\varphi(\hat{y})$ are quite different. In the two server join the shortest queue system without dedicated customers, $\varphi(\cdot)$ is given in (2) while $\mu(0,0) = 1/2$, and $\mu(0,1) = \mu(0,-1) = 1/4$ while . This can be seen by noting that the first time ℓ is reached with ℓ even, the two queue lengths will be equal. However, for ℓ odd, there is an equal chance of either queue having the extra customer. The distribution μ may also be expressed as in Proposition 2.4 in [14]:

$$\mu(s, \hat{y}) = \tilde{d}_1^{-1} f^\infty(s, \hat{y}) \varphi(\hat{y}) \tag{21}$$

where $f^\infty(s, \hat{y})$ is the probability the time reversal of $\mathcal{W}^\infty$ with respect to $m^r \times \varphi$ (m is counting measure) jumps from $(s, \cdot, \hat{y})$ and the first additive component drifts away to minus infinity without ever becoming nonnegative again.

The following constants will be used in the statement of the theorems. Let

$$f(m) \equiv \sum_{\substack{x\in\Delta \\ A(\hat{x})-\tilde{x}_1 \overset{p}{=} m}} \pi(x)\, h(x)\, H(x) \quad \text{for } m = 0, \ldots, p-1 \tag{22}$$

$$g(m) \equiv f(m) p \sum_{\substack{(s,\hat{y})\in\mathbb{Z}_+\times\hat{S} \\ A(\hat{y})-s \overset{p}{=} m}} \alpha^{-s} \hat{h}(\hat{y}) \mu(s, \hat{y})$$

$$g \equiv g(0) + \cdots + g(p-1). \tag{23}$$

The constants $f(m)$ for $m = 0, \ldots, p-1$ and g are generally unknown, but can be obtained by fast simulation. Note that in the special case when the overshoot is always zero and $\hat{h}(\hat{y}) = 1$ for $\hat{y} \in \hat{S}$, then $f(m) = g(m)$ for $m = 0, \ldots, p-1$.

The following theorem provides an exact asymptotic description of the steady state distribution of the system. The proof in this paper eliminates the need for C.12 (Condition 7 in [14]), the uniform integrability condition.

Theorem 5 (Steady State). *Under C.1–C.11 and as* $\ell \to \infty$,

$$\pi(Y_\ell(\hat{y})) \sim \alpha^{-\ell} \hat{h}(\hat{y})\, p \varphi(\hat{y}) \frac{f((A(\hat{y}) - \ell) \mod p)}{\tilde{d}_1}$$

where $\tilde{d}_1$ *is given in C.7 and* f *is given at (22).*

Theorem 6 (Mean Hitting Time). *Let* $\sigma \in \Delta$ *with* $H(\sigma) > 0$. *Under C.1–C.12 and as* $\ell \to \infty$,

$$\mathrm{E}_\sigma T_\ell \sim \alpha^\ell / g \text{ as } \ell \to \infty,$$

where the constant g *is defined in* (23).

Theorem 7 (Hitting Distribution). *Let $\sigma \in \Delta$ with $H(\sigma) > 0$. Under C.1–C.12 and as $\ell \to \infty$,*

$$\Pr_\sigma \left\{ \tilde{W}_1[T_\ell] - \ell = s, \hat{W}[T_\ell] = \hat{y} \mid T_\ell < T_\sigma \right\} \sim \chi\{A(\hat{y}) - s \stackrel{p}{=} \ell\} c^{-1} \alpha^{-s} \hat{h}(\hat{y}) p \mu(s, \hat{y})$$

with the normalization constant

$$c = p \sum_{\substack{(t,\hat{z}) \in \mathbb{Z}_+ \times \hat{S} \\ A(\hat{z}) - t \stackrel{p}{=} \ell}} \alpha^{-t} \hat{h}(\hat{z}) \mu(t, \hat{z})$$

where T_σ is the return time to state σ.

It is easy to describe a sequence of "tubes" C_ℓ^c that contain the most likely approach to F_ℓ with high probability.

Theorem 8. *Let $C_\ell \in \sigma(W_z, 0 \le s \le T_\ell)$ be a sequence of sets of trajectories of W and let $\mathcal{C}_\ell$ be sets of trajectories for $\mathcal{W}^\infty$ such that $C_\ell \subseteq \mathcal{C}_\ell$ and $\Pr_x \{\mathcal{C}_\ell\} \to 0$ as $\ell \to \infty$. Let $\sigma \in \Delta$ with $H(\sigma) > 0$. Under C.1–C.12 and as $\ell \to \infty$,*

$$\lim_{\ell \to \infty} \Pr_\sigma \{C_\ell | T_\ell < T_\sigma\} = 0.$$

Corollary 1 (Large Deviation Tube). *Let $\sigma \in \Delta$ with $H(\sigma) > 0$. Assume C.1–C.12 and that the stronger Markov additive property* (17) *holds (or at least that* $\lim_{n\to\infty} \tilde{\mathcal{W}}^\infty[n]/n = \tilde{d}$*), then, as $\ell \to \infty$,*

$$\lim_{\ell \to \infty} \Pr_\sigma \left\{ \sup_{0 \le s \le T_\ell} |\tilde{W}[s] - \tilde{d}s| > \epsilon \ell | T_\ell < T_\sigma \right\} = 0.$$

This result means the nodes driven into overload by the first grow linearly with the length of the queue at the first node. The Joint Bottlenecks result in [14] follows immediately.

Corollary 2 (Joint Bottlenecks). *Let $\sigma \in \Delta$ with $H(\sigma) > 0$. Assume C.1–C.12 and the stronger Markov additive structure* (17) *(or at least that* $\lim_{n\to\infty} \tilde{\mathcal{W}}^\infty[n]/n = \tilde{d}$ *a.s.) hold. Then the conditional distribution of $\tilde{W}[T_\ell]/T_\ell$, given $T_\ell < T_\sigma$, converges to a unit point measure at $\tilde{d}/\tilde{d}_1$.*

Corollary 3. *Let f be a function on $\hat{S}$ such that $\sum_{\hat{y}} f(\hat{y})\varphi(\hat{y}) < \infty$. Let $\sigma \in \Delta$ with $H(\sigma) > 0$. then, as $\ell \to \infty$,*

$$\lim_{\ell \to \infty} \Pr_\sigma \left\{ \left|\frac{1}{T_\ell} \sum_{k=0}^{T_\ell} f(W[k]) - \sum_{\hat{y}} f(\hat{y})\varphi(\hat{y})\right| > \epsilon | T_\ell < T_\sigma \right\} = 0.$$

If we are interested in a starting point $\sigma \notin \Delta$, it is straightforward in most applications to show that C.1–C.12 still hold if Δ is enlarged to include *sigma*.

4.1 Proofs of Exact Asymptotic Results

The purpose of this section is to extend the results of [14] for countable state Markov chains to cover the join the shortest queue system. The results in [14] explicitly assume aperiodicity and tacitly assume $\Delta \cap F_\ell = \emptyset$ for ℓ sufficiently large. Lemma 7 shows that the uniform integrability C.12 (Condition 7 of [14]) implies that $\pi(\Delta \cap F_\ell)$ is asymptotically negligible. But first, we discuss some basic aspects of periodicity.

We now show that $\stackrel{p}{\sim}$ is an equivalence relation where $\hat{y} \stackrel{p}{\sim} \hat{x}$ if there exists n and s with $\mathcal{J}^n((0,\hat{x}),(s,\hat{y})) > 0$ and $s \stackrel{p}{=} 0$. ($\mathcal{J}$ was defined in C.1.) The reflexive property is obvious. To show symmetry, assume that $\hat{y} \stackrel{p}{\sim} \hat{x}$ and note that C.6 implies the existance of m and t such that $\mathcal{J}^m((0,\hat{y}),(t,\hat{x})) > 0$. Hence, $\mathcal{J}^{n+m}((0,\hat{x}),(s+t,\hat{x})) > 0$. Since $s + t \stackrel{p}{=} 0$, it follows that $t \stackrel{p}{=} 0$ implying $\hat{x} \stackrel{p}{\sim} \hat{y}$. Transitivity follows similarly. We now show that the equivalence classes generated by $\stackrel{p}{\sim}$ are the non-empty sets among $A_0, \ldots, A_{p-1}$. Clearly every point $\hat{y} \in \hat{S}$ is in some A_j for $j = 0, \ldots, p-1$, and $\hat{\delta} \in A_0$ by construction. Assume that $\hat{x} \in A_j$ and $\hat{y} \in A_i$. Then there exists n, m, s and t such that

$$\mathcal{J}^n((0,\hat{\delta}),(s,\hat{x})) > 0 \text{ where } s \stackrel{p}{=} j \text{ and } \mathcal{J}^m((0,\hat{\delta}),(t,\hat{y})) > 0 \text{ where } t \stackrel{p}{=} i.$$

Again, by C.6, there must exist a k such that $\mathcal{J}^k((0,\hat{y}),(u,\hat{\delta})) > 0$ for some u and by periodicity $u \stackrel{p}{=} -i$. Similarly, there exists an l such that $\mathcal{J}^l((0,\hat{x}),(v,\hat{y})) > 0$ for some v. Hence $\mathcal{J}^{n+l+k}((0,\hat{\delta}),(s+v+u,\hat{\delta})) > 0$ and by periodicity $s + v + u \stackrel{p}{=} 0$. Thus, $\hat{y} \stackrel{p}{\sim} \hat{x}$ iff $v \stackrel{p}{=} 0$ iff $i = j$.

Let the kernel of the chain $(\mathcal{R}^\infty[\ell], \hat{\mathcal{W}}^\infty[\mathcal{T}_\ell^\infty])$ indexed by ℓ where $\mathcal{R}^\infty[\ell] \equiv \tilde{\mathcal{W}}_1^\infty[\mathcal{T}_\ell^\infty] - \ell$ be denoted by R. Note that if $s > 0$, then $R((s,\hat{y}),(s-1,\hat{y})) = 1$. This along with C.6 implies that there is a single closed, irreducible set of recurrent states. Hence, R possesses an invariant measure which is unique up to rescaling. By the same argument as [11] Section 3, μ as given by (21) is the unique invariant probability measure for R. This Markov chain has period p since the first return to state $(0,\hat{\delta})$ will occur at a time ℓ where $\tilde{\mathcal{W}}_1^\infty[n] = \ell$ and $\hat{\mathcal{W}}^\infty[n] = \hat{\delta}$ and n is the first return time to $\hat{\delta}$ such that $\tilde{W}_1^\infty[n]$ is positive. Since $\mathcal{W}^\infty$ has period p it follows that ℓ must be a multiple of p. We partition the state space $\mathbb{Z}_+ \times \hat{S}$ into $B_0, \ldots, B_{p-1}$ by assigning $(s,\hat{y})$ to B_i iff $A(\hat{y}) - s \stackrel{p}{=} i$. Note that $(0,\hat{\delta}) \in B_0$. We leave it to the reader to prove the following lemma.

Lemma 4. *For $k = 0, \ldots, p-1$:*

1. *If $(\mathcal{R}^\infty[\ell], \hat{\mathcal{W}}^\infty[\mathcal{T}_\ell^\infty])$ is initially in B_k, then after ℓ steps, it is in $B_{k+\ell \mod p}$,*
2. *B_k is a closed set for the marginal twisted free Markov chain $(\tilde{\mathcal{W}}_1^\infty[n], \hat{\mathcal{W}}^\infty[n])$, and*
3. *$\mu(B_k) = 1/p$.*
4. *If $\hat{w} \in A_i$ and $\tilde{w}_1 = u$ then $(u,\hat{w}) \in B_k$ where $k \stackrel{p}{=} i - u$. After $\ell = np + m$ transitions, R^{np+m} takes B_k into B_{k+m}. By Theorem 6.3.10 in [7], we have*

$$\lim_{n\to\infty} R^{np+m}((u,\hat{w})(s,\hat{y})) = p\mu(s,\hat{y})\chi\{A(\hat{y}) - s \stackrel{p}{=} k+m\} \text{ and}$$
$$\Pr_x\left\{\mathcal{R}^\infty[\ell] = s, \hat{\mathcal{W}}^\infty[\mathcal{T}_\ell^\infty] = \hat{y}\right\} \sim p\mu(s,\hat{y})\chi\{A(\hat{x}) - \tilde{x}_1 \stackrel{p}{=} A(\hat{y}) - \ell - s\}. \tag{24}$$

Lemma 5. *Fix $\hat{y} \in A_j$. Define $\tilde{m}_Y^\infty(u,\hat{w})$ as the expected number of visits to $Y_0(\hat{y})$ by $\mathcal{W}^\infty$ starting from a state $w = (\tilde{w},\hat{w})$ with $\tilde{w}_1 = u$. Under C.5 and C.7,*

$$\sum_{u,\hat{w}} \tilde{m}_Y^\infty(u,\hat{w})\mu(u,\hat{w}) \;=\; \varphi(\hat{y})/\tilde{d}_1. \tag{25}$$

Proof. In [14] a stationary process $\mathcal{W}^\#$ is defined and μ is the hitting distribution of $\mathcal{W}^\#[\mathcal{T}_0]$; that is when $\tilde{\mathcal{W}}_1^\#$ first reaches or exceeds 0. Consequently, $\sum_{u,\hat{w}} \tilde{m}_Y^\infty(u,\hat{w})\mu(u,\hat{w})$ is just the expected number of visits to $Y_0(\hat{y})$ by $\mathcal{W}^\#$. By stationarity, this is just $\varphi(\hat{y})/\tilde{d}_1$. □

The following useful change of measure property follows from C.4 and C.5 and will be used in many of the remaining proofs. For $x_0, x_1, \ldots, x_n \in S^\infty$

$$\Pr_{x_0}\{W^\infty[1] = x_1, \ldots, W^\infty[n] = x_n\} = \frac{h(x_0)}{h(x_n)}\Pr_{x_0}\{\mathcal{W}^\infty[1] = x_1, \ldots, \mathcal{W}^\infty[n] = x_n\}. \tag{26}$$

Proof of Theorem 5. For $\ell > L(\hat{y})$, the steady state probability of $Y_\ell(\hat{y})$ is given by

$$\begin{aligned}\pi(Y_\ell(\hat{y})) = \pi(\Delta)\mathrm{E}_\Delta\left(\sum_{n=0}^{T_\Delta - 1}\chi\{W[n] \in Y_\ell(\hat{y})\}\right) &= \sum_{z\in\Delta}\pi(z)\mathrm{E}_z\left(\sum_{n=0}^{T_\blacktriangle^\infty - 1}\chi\{W^\infty[n] \in Y_\ell(\hat{y})\}\right)\\ = \sum_{z\in\Delta}\pi(z)\sum_{(u,\hat{w})} m_Y^\blacktriangle(\ell,u,\hat{w})&\Pr_z\left\{\hat{W}^\infty[T_\ell^\infty] = \hat{w}, R^\infty[T_\ell^\infty] = u, T_\ell^\infty < T_\blacktriangle^\infty\right\}\end{aligned} \tag{27}$$

where we have conditioned on the point where W^∞ overshoots ℓ, that is at $w = (\tilde{w},\hat{w})$, where $R^\infty[\ell] \equiv \tilde{W}_1^\infty[T_\ell] - \ell = u = \tilde{w}_1 - \ell$, and $m_Y^\blacktriangle(\ell,u,\hat{w}) = \mathrm{E}_w\left(\sum_{n=0}^{T_\blacktriangle - 1}\chi\{W^\infty[n] \in Y_\ell(\hat{y})\}\right)$ is the expected number of visits to $Y_\ell(\hat{y})$ obtained by W^∞ (or W) after hitting $\{\ell, \ell+1, \ldots\} \times \mathbb{Z}^{r-1} \times \hat{S}$, but before returning to $\blacktriangle$.

Let $N_Y^{\blacktriangle}(\ell)$ denote the number of visits and $T_Y^{\infty}[k]$ the time of the kth visit to $Y_\ell(\hat{y})$ before $T_{\blacktriangle}$. Hence,

$$\begin{aligned}
m_Y^{\blacktriangle}(\ell,u,\hat{w}) &= \sum_{k=1}^{\infty}\sum_{z\in Y_\ell(\hat{y})}\Pr_w\left\{N_Y^{\blacktriangle}(\ell)\geq k, W^{\infty}[T_Y^{\infty}[k]]=z\right\}\\
&= \sum_{k=1}^{\infty}\sum_{z\in Y_\ell(\hat{y})}\frac{h(w)}{h(z)}\Pr_w\left\{\mathcal{N}_Y^{\blacktriangle}(\ell)\geq k, \mathcal{W}^{\infty}[\mathcal{T}_{\ell+Y}^{\infty}[k]]=z\right\} \qquad (28)\\
&= \alpha^{-\ell}h(w)\hat{h}(\hat{y})\mathrm{E}_w\left(\sum_{n=0}^{\mathcal{T}_{\blacktriangle}^{\infty}-1}\chi\{\mathcal{W}^{\infty}[n]\in Y_\ell(\hat{y})\}\right)\\
&= \alpha^{-\ell}h(w)\hat{h}(\hat{y})\tilde{m}_Y^{\blacktriangle}(\ell,u,\hat{w}) \qquad (29)
\end{aligned}$$

where $\tilde{m}_Y^{\blacktriangle}(\ell,u,\hat{w})$ is the expected value of $\mathcal{N}_Y^{\blacktriangle}(\ell)$ conditioned on the first overshoot of ℓ occurring at some state $w=(\tilde{w},\hat{w})$ with $\tilde{w}_1=\ell+u$ and $u\geq 0$. The line (28) holds because the event $N_Y^{\blacktriangle}\geq k$ only depends on the trajectory up to the k^{th} visit to $Y_\ell(\hat{y})$ which occurs at some state $z\in Y_\ell(\hat{y})$. The likelihood ratio for this event for W^{∞} versus $\mathcal{N}_Y^{\blacktriangle}(\ell)\geq k$ for $\mathcal{W}^{\infty}$ is precisely $h(w)/h(z)$. We see this by using (26) on transitions from w until the k^{th} visit to $Y_\ell(\hat{y})$. Fortunately, $h(z)=\alpha^{\ell}/\hat{h}(\hat{y})$ for all states in $Y_\ell(\hat{y})$.

Substituting (29) into (27), doing a change of measure on the trajectory reaching F, and multiplying by α^{ℓ}, yields

$$\alpha^{\ell}\pi(Y_\ell(\hat{y}))=\hat{h}(\hat{y})\sum_{x\in\Delta}\lambda(x)\sum_{u,\hat{w}}\tilde{m}_Y^{\blacktriangle}(\ell,u,\hat{w})\Pr_x\left\{\mathcal{R}^{\infty}[\ell]=u,\hat{\mathcal{W}}^{\infty}[\mathcal{T}_\ell^{\infty}]=\hat{w},\mathcal{T}_\ell^{\infty}<\mathcal{T}_{\blacktriangle}^{\infty}\right\}.$$

Let $\ell=np+m$, and let $n\to\infty$. From (24) and since the event $\{\mathcal{T}_\ell^{\infty}<\mathcal{T}_{\blacktriangle}^{\infty}\}$ is asymptotically independent of the tail σ-field generated by $(\hat{\mathcal{W}}^{\infty}[\mathcal{T}_\ell^{\infty}],\mathcal{R}^{\infty}[\ell])$, it follows that,

$$\Pr_x\left\{\mathcal{R}^{\infty}[\ell]=u,\hat{\mathcal{W}}^{\infty}[\mathcal{T}_\ell^{\infty}]=\hat{w},\mathcal{T}_\ell^{\infty}<\mathcal{T}_{\blacktriangle}^{\infty}\right\}\to H(x)p\mu(u,\hat{w})\chi\{A(\hat{x})-\tilde{x}_1\overset{p}{=}A(\hat{w})-m-u\}. \qquad (30)$$

Also $\tilde{m}_Y^{\blacktriangle}(\ell,u,\hat{w})\to\tilde{m}_Y^{\infty}(u,\hat{w})\leq\tilde{m}_Y^{\infty}(0,\hat{y})$ where $\tilde{m}_Y^{\infty}(u,\hat{w})$ denotes the expected number of visits to $Y_0(\hat{y})$ starting from a point w with $\tilde{w}_1=u$. From the last inequality, C.10, and dominated convergence, it follows that as $n\to\infty$

$$\alpha^{\ell}\pi(Y_\ell(\hat{y})) \quad\to\quad \hat{h}(\hat{y})\sum_{\substack{x\in\Delta\\ A(\hat{x})-\tilde{x}_1\overset{p}{=}A(\hat{y})-\ell}}\lambda(x)H(x)\sum_{u,\hat{w}}\tilde{m}_Y^{\infty}(u,\hat{w})p\mu(u,\hat{w})$$

Note that we can only include states $(\tilde{x}_1,\hat{x})$, $(\ell+u,\hat{w})$, and $(\ell,\hat{y})$ which lie in the same closed set; i.e., with $A(\hat{x})-\tilde{x}_1\overset{p}{=}A(\hat{w})-\ell-u\overset{p}{=}A(\hat{y})-\ell$. This follows from the restricted region over which x is summed and since $\tilde{m}_Y^{\infty}(u,\hat{w})$ can only be non-zero if $A(\hat{w})-u\overset{p}{=}A(\hat{y})-0$ which occurs iff $A(\hat{w})-\ell-u\overset{p}{=}A(\hat{y})-\ell$. To complete the argument, use Lemma 5 and (22) to obtain

$$\pi(Y_\ell(\hat{y})) \quad\sim\quad \alpha^{-\ell}\hat{h}(\hat{y})f((A(\hat{y})-\ell)\mod p)p\frac{\varphi(\hat{y})}{\tilde{d}_1}.$$

□

We will use the following uniform integrability repeatedly.

Lemma 6. *Under C.1–C.12, the sequence of functions* $\mathrm{E}[\hat{h}(\hat{\mathcal{W}}^{\infty}[\mathcal{T}_\ell^{\infty}])\mid\mathcal{W}^{\infty}[0]=x]$ *indexed by* ℓ *are uniformly integrable with respect to the initial measure* λ.

Proof. We review the somewhat concise arguments in [14] and make adjustments for periodicity. As in Lemma 2.8 [14], to prove uniform integrability it suffices to show $\hat{h}(\hat{\mathcal{W}}^{\infty}[\mathcal{T}_\ell])$ converges in $L^1(\Pr_\lambda)$. Let $\hat{\alpha}\in\hat{S}$ and let $\mathcal{T}_{\hat{\alpha}}$ be the first time $\hat{\mathcal{W}}^{\infty}$ hits $\hat{\alpha}$. We can break $\hat{h}(\hat{\mathcal{W}}^{\infty}[\mathcal{T}_\ell])$ into

$$\hat{h}(\hat{\mathcal{W}}^{\infty}[\mathcal{T}_\ell])\chi\{\mathcal{T}_{\hat{\alpha}}>\mathcal{T}_\ell\}\text{ and }\hat{h}(\hat{\mathcal{W}}^{\infty}[\mathcal{T}_\ell])\chi\{\mathcal{T}_{\hat{\alpha}}\leq\mathcal{T}_\ell\}.$$

To bound the first term, remark that

$$\hat{h}(\hat{\mathcal{W}}^{\infty}[\mathcal{T}_\ell])\chi\{\mathcal{T}_{\hat{\alpha}} > \mathcal{T}_\ell\} \le \sum_{k=0}^{\mathcal{T}_{\hat{\alpha}}} \hat{h}(\hat{\mathcal{W}}^{\infty}[k])\chi\{\mathcal{T}_{\hat{\alpha}} > \mathcal{T}_\ell\}.$$

However the event $\{\mathcal{T}_{\hat{\alpha}} > \mathcal{T}_\ell\}$ tends to the empty set as $\ell \to \infty$. Moreover,

$$\mathrm{E}_\lambda \sum_{k=0}^{\mathcal{T}_{\hat{\alpha}}} \hat{h}(\hat{\mathcal{W}}^{\infty}[k]) = \mathrm{E}_{\hat{\lambda}} \sum_{k=0}^{\mathcal{T}_{\hat{\alpha}}} \hat{h}(\hat{\mathcal{W}}^{\infty}[k])$$

by the Markov additive structure. This is bounded by C.12 using Theorem 14.2.2 in [15]. Consequently, $\mathrm{E}_\lambda[\hat{h}(\hat{\mathcal{W}}^{\infty}[\mathcal{T}_\ell])\chi\{\mathcal{T}_{\hat{\alpha}} > \mathcal{T}_\ell\}] \to 0$ by dominated convergence.

The expected value of the second term with respect to λ may be expressed as

$$\sum_{x_1<\ell} \mathrm{Pr}_\lambda \{\mathcal{W}_1^{\infty}[\mathcal{T}_{\hat{\alpha}}] = x_1\} \mathrm{E}_{(\tilde{0},\hat{\alpha})} \hat{h}(\hat{\mathcal{W}}^{\infty}[\mathcal{T}_{(\ell-x_1)^+}]). \tag{31}$$

Moreover, $\mathrm{E}_{(\tilde{0},\hat{\alpha})}\hat{h}(\hat{\mathcal{W}}^{\infty}[\mathcal{T}_u])$ is uniformly bounded. This follows by time reversal, since

$$\begin{aligned}
&\varphi(\hat{\alpha})\mathrm{E}_{\tilde{0},\hat{\alpha}}\hat{h}(\hat{\mathcal{W}}^{\infty}[\mathcal{T}_u]) \\
= &\sum_{\hat{y}} \varphi(\hat{y})\hat{h}(\hat{y}) Pr(\mathcal{W}^*[t] < u, t > 0; \hat{\mathcal{W}}^*[t] = \hat{\alpha} \text{ for some } t > 0 | \hat{\mathcal{W}}^*[0] = \hat{y}) \\
\le &\sum_{\hat{y}} \varphi(\hat{y})\hat{h}(\hat{y}) < \infty
\end{aligned}$$

by C.9 where $\mathcal{W}^*$ is the time reversal of $\mathcal{W}^{\infty}$.

In the aperiodic case, Theorem 2.5 in [14] shows that, for any initial point x, $\mathrm{E}_x\hat{h}(\hat{\mathcal{W}}^{\infty}[\mathcal{T}_\ell])$ converges to $\sum \hat{h}(\hat{y})\mu(\hat{y})$ where μ is the limiting distribution of $\hat{\mathcal{W}}^{\infty}[\mathcal{T}_\ell]$ as $\ell \to \infty$. In the periodic case, using Lemma 4 part 4, and assuming without loss of generality that $\hat{\alpha} \in A_0$, we know that

$$\mathrm{E}_{(\tilde{0},\hat{\alpha})}\hat{h}(\hat{\mathcal{W}}^{\infty}[\mathcal{T}_{np+m}]) \to p \sum_{\hat{y}} \sum_{s \ge 0} \hat{h}(\hat{y})\mu(s,\hat{y})\chi\{A(\hat{y}) - s \overset{p}{=} m\}$$

Hence, by dominated convergence, $\mathrm{E}_{(\tilde{0},\hat{\alpha})}\hat{h}(\hat{\mathcal{W}}^{\infty}[\mathcal{T}_{\ell-x_1}])$ converges in L^1 relative to the measure $\mathrm{Pr}_\lambda \{\mathcal{W}_1^{\infty}[\mathcal{T}_{\hat{\alpha}}] = \cdot\}$. This gives our result. □

Lemma 7. *Under C.4, C.12, $\pi(\Delta \cap F_\ell)$ is asymptotically negligible since $\alpha^\ell \pi(\Delta \cap F_\ell) \to 0$.*

Proof. Note that,

$$\begin{aligned}
\alpha^\ell \pi(\Delta \cap F_\ell) &\le \sum_{x \in \Delta \cap F_\ell} \pi(x)\alpha^{\tilde{x}_1} \text{ since } \tilde{x}_1 > \ell \\
&= \sum_{x \in \Delta \cap F_\ell} \pi(x)h(x)\hat{h}(\hat{x}) \text{ by C.4} \\
&= \sum_{x \in \Delta \cap F_\ell} \lambda(x)\hat{h}(\hat{x}) \\
&= \sum_{x \in \Delta \cap F_\ell} \lambda(x)\mathrm{E}\left[\hat{h}(\hat{\mathcal{W}}^{\infty}[\mathcal{T}_\ell^{\infty}])\chi\{\mathcal{T}_\ell^{\infty} = 0\} \mid \mathcal{W}^{\infty}[0] = x\right] \\
&\le \sum_{x \in \Delta} \lambda(x)\mathrm{E}\left[\hat{h}(\hat{\mathcal{W}}^{\infty}[\mathcal{T}_\ell^{\infty}]) \mid \mathcal{W}^{\infty}[0] = x\right] \\
&< \infty \text{ by C.12.}
\end{aligned} \tag{32}$$

Since the r.h.s. of (32) is finite for fixed ℓ, the r.h.s. of (32) must go to zero as $\ell \to \infty$. □

In the proofs of Theorems 5–7, we will need the Comparison Lemma of [14] which changes only slightly to account for $\Delta \cap F_\ell \neq \emptyset$. The difficulties are resolved using Lemma 7 and the following Corollary.

Corollary 4. *Under C.4, C.12,*

$$\lim_{\ell\to\infty} \alpha^\ell \sum_{z\in S} \pi(z)\mathrm{Pr}_z \{T_\ell = T_\Delta\} = 0.$$

Proof. Use the time reversal W^* of W, so

$$\sum_{z\in F_\ell^c} \pi(z)\mathrm{Pr}_z \{T_\ell = T_\Delta\} = \sum_{x\in \Delta\cap F_\ell} \pi(x)\mathrm{Pr}_x \{W^*[1] \in F_\ell^c\} \leq \pi(\Delta \cap F_\ell).$$

The result now follows from Lemma 7. □

As in [14], define $f(x) = \mathrm{Pr}_x \{T_\sigma < T_\ell\}$ if $x \neq \sigma$ and $f(\sigma) = 1$. Similarly, define $\rho(x) = \mathrm{Pr}_x \{T_\Delta < T_\ell\}$ if $x \notin \Delta$ and $\rho(x) = 1$ if $x \in \Delta \cap F_\ell^c$. Note that $f(x) = \rho(y) = 0$ for $y \in F_\ell$. As in [14], $p_\sigma = \mathrm{Pr}_\sigma \{T_\ell < T_\sigma\}$ and

$$\pi(\sigma)p_\sigma = \pi(\sigma) \sum_{y\neq\sigma} (1 - f(y)) = \Lambda \equiv \sum_{y\in F_\ell} \pi(y) \sum_{x\in F_\ell^c} K(y,x)f(x).$$

The Comparison Lemma in [14] shows $\lim_{\ell\to\infty}(b - \Lambda)/b = 0$ where

$$\begin{aligned} b &\equiv \sum_{y\in F_\ell} \pi(y) \sum_{x\in F_\ell^c} K(y,x)\rho(x) && (33)\\ &= \sum_{z\in\Delta\cap F_\ell^c} \pi(z) \sum_{x\in S} K(z,x)(1-\rho(x)) \\ &= \sum_{z\in\Delta\cap F_\ell^c} \pi(z)\mathrm{Pr}_z \{T_\ell \leq T_\Delta\} \\ &\sim \sum_{z\in\Delta\cap F_\ell^c} \pi(z)\mathrm{Pr}_z \{T_\ell < T_\Delta\} \text{ by Corollary 4} \\ &= \alpha^{-\ell} \sum_{x\in\Delta\setminus F_\ell} \lambda(x)\mathrm{E}_x \left[\chi\{\mathcal{T}_\ell^\infty < \mathcal{T}_\blacktriangle^\infty\}\hat{h}(\hat{\mathcal{W}}^\infty[\mathcal{T}_\ell^\infty])\alpha^{-\mathcal{R}^\infty[\ell]}\right] && (34)\end{aligned}$$

Note that (34) times α^ℓ tends to a positive limit if $\ell = np + k$ as $n \to \infty$ by (30).

The key idea is to represent

$$\begin{aligned} b - \Lambda &= \sum_{y\in F_\ell} \pi(y) \sum_{x\in F_\ell^c} K(y,x)(\rho(x) - f(x)) \\ &= \sum_{z\in\Delta\cap F_\ell^c\setminus\sigma} \pi(z)U^*(z)V(z) \end{aligned}$$

where $U^*(z)$ is the probability the time reversal of W hits F_ℓ before σ (this probability tends to 0 as $\ell \to \infty$) and where $V(z) = \mathrm{Pr}_z \{T_\ell \leq T_\Delta\}$. Consequently,

$$\begin{aligned} b - \Lambda &= \sum_{z\in\Delta\cap F_\ell^c\setminus\sigma} \pi(z)U^*(z)\mathrm{Pr}_z \{T_\ell \leq T_\Delta\} \\ &\sim \sum_{z\in\Delta\cap F_\ell^c\setminus\sigma} \pi(z)U^*(z)\mathrm{Pr}_z \{T_\ell < T_\Delta\} \text{ using Corollary 4} \\ &= \alpha^{-\ell} \sum_{z\in\Delta\cap F_\ell^c\setminus\sigma} \lambda(z)U^*(z)\mathrm{E}_z \left[\chi\{\mathcal{T}_\ell^\infty < \mathcal{T}_\blacktriangle^\infty\}\hat{h}(\hat{\mathcal{W}}^\infty[\mathcal{T}_\ell^\infty])\alpha^{-\mathcal{R}^\infty[\ell]}\right]. && (35)\end{aligned}$$

Using (34) and the fact that $U^*(z) \to 0$ as $\ell \to \infty$ we can use Condition C.12 to show $\lim_{\ell\to\infty}(b - \Lambda)/b = 0$.

Proof of Theorem 6. We just follow the steps in [14]. There are no other changes other than to modify the asymptotics of (1.19) in [14] as above to take periodicity into account. Hence,

$$\alpha^{-\ell}\left(\mathrm{E}_\sigma T_\ell\right) \to g$$

where $g = g(0) + \cdots + g(p-1)$. $\square$

Let $p_\Delta = \Pr_\Delta\{T_\ell < T_\Delta\}$. Note that

$$b = \pi(\Delta)p_\Delta - \sum_{z\in\Delta\cap F_\ell}\pi(z)\sum_{x\in S}K(z,x)(1-\rho(x))$$

by (33). Hence, by Lemma 7, $\left(\mathrm{E}_\sigma T_\ell\right)^{-1} \sim \pi(\Delta)p_\Delta$.

Proof of Theorem 7. We modify the proof in [14]. Let $\Pr_\Delta\{\cdot\}$ denote the probability conditioned on the process starting in Δ with respect to π. For $\ell \geq L(\hat{y})$,

$$\begin{aligned}
\Pr_\Delta\{T_\ell < T_\Delta, W[T_\ell] \in Y_{\ell+s}(\hat{y})\} &= \sum_{x\in\Delta}\frac{\pi(x)}{\pi(\Delta)}\Pr_x\{T_\ell < T_\Delta, W[T_\ell] \in Y_{\ell+s}(\hat{y})\}\\
&= \sum_{x\in\Delta}\frac{\pi(x)}{\pi(\Delta)}\Pr_x\{T_\ell^\infty < T_\blacktriangle^\infty, W^\infty[T_\ell^\infty] \in Y_{\ell+s}(\hat{y})\} \text{ by C.2, C.3, and C.11}\\
&= \alpha^{-(\ell+s)}\sum_{x\in\Delta}\frac{\lambda(x)}{\pi(\Delta)}\hat{h}(\hat{y})\Pr_x\{\mathcal{T}_\ell^\infty < \mathcal{T}_\blacktriangle^\infty, \mathcal{W}^\infty[\mathcal{T}_\ell^\infty] \in Y_{\ell+s}(\hat{y})\} \text{ by (26).}
\end{aligned}$$

By the argument in [14] and using Lemma 4 part 4. we have,

$$\Pr_\sigma\left\{\tilde{W}_1[T_\ell] - \ell = s, \hat{W}[T_\ell] = \hat{y} \mid T_\ell < T_\sigma\right\}$$

is asymptotic to the limit of (1.22) in [14]. Using the above asymptotic result for $\ell = np + m$ as $n \to \infty$, the numerator of (1.22) in [14] tends to

$$\sum_{\substack{z\in\Delta\\ A(\hat{z})-\tilde{z}_1 \overset{p}{=} A(\hat{y})-s-\ell}}\pi(z)\,h(z)\,H(z)\,\alpha^{-s}\check{a}(\hat{y})p\mu(s,\hat{y}).$$

To see that the parity is correct, note that $(\tilde{W}_1[T_\ell], \hat{W}[T_\ell])$ must reach $(s+\ell, \hat{y}) \in B_k$ where $k \overset{p}{=} A(\hat{y}) - s - \ell$. Since B_k is closed for the marginal twisted free process, z must have been in B_k also. The denominator of (1.22) in [14] tends to $g(A(\hat{y}) - s - \ell \mod p)$ defined at (23). The ratio of these limits gives Theorem 7. $\square$

Proof of Theorem 8. The proof is an extension of Theorem 2.10 in [14] which in turn uses the ideas in [5]. Note that the proof does not need to require that all the components of $\vec{d}$ be positive as was assumed in [14]. Let H_ℓ^σ be the event $\{T_\ell < T_\sigma\}$. We first remark that

$$\Pr_\Delta\{H_\ell \cap C_\ell\} \sim \Pr_\sigma\{H_\ell^\sigma \cap C_\ell\}$$

where $H_\ell = \{T_\ell < T_\Delta\}$ as $\ell \to \infty$. This follows from the proof of the Comparison Lemma 1.8 in [14]. Similarly $\Pr_\Delta\{H_\ell\} \sim \Pr_\sigma\{H_\ell^\sigma\}$. Consequently it suffices to show

$$\lim_{\ell\to\infty}\Pr_\Delta\{C_\ell | T_\ell^\infty < T_\Delta^\infty\} = 0$$

since W never hits Δ during trajectories in H^ℓ.

Next,

$$\frac{\Pr_\Delta\{C_\ell, T_\ell^\infty < T_\Delta^\infty\}}{\Pr_\Delta\{T_\ell^\infty < T_\Delta^\infty\}} = \frac{\sum_{x\in\Delta}\lambda(x)\mathrm{E}_x(\chi\{\mathcal{C}_\ell \cap \mathcal{H}_\ell\}\hat{a}^{-1}(\hat{\mathcal{W}}^\infty[\mathcal{T}_\ell^\infty]))}{\sum_{x\in S_\Delta}\lambda(x)\mathrm{E}_x(\chi\{\mathcal{H}_\ell\}\hat{a}^{-1}(\hat{\mathcal{W}}^\infty[\mathcal{T}_\ell^\infty]))}$$

where $\mathcal{H}_\ell = \{\mathcal{T}_\ell^\infty < \mathcal{T}_\Delta^\infty\}$.

By hypothesis, $\Pr_x\{\mathcal{C}_\ell\} \to 0$ as $\ell \to \infty$. The result follows since $\mathrm{E}_x(\chi\{\mathcal{H}_\ell\}\hat{a}^{-1}(\hat{\mathcal{W}}^\infty[\mathcal{T}_\ell^\infty]))$ converges to a nonzero limit as $\ell \to \infty$ by Lemma 1.4 in [14]. $\square$

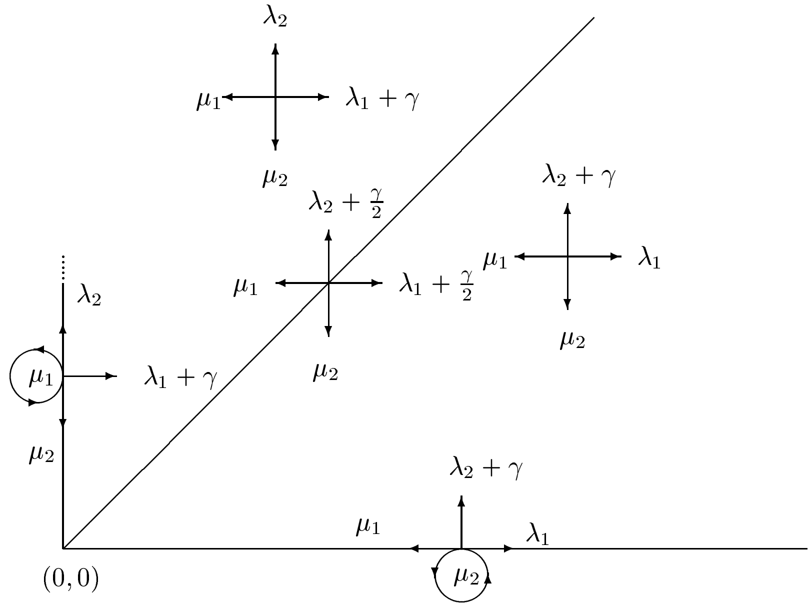

Figure 2: Transition probabilities of Q.

Proof of Corollary 1. Let $C_\ell = \chi\{\sup_{0\le s\le T_\ell} |\tilde{W}[s] - \tilde{d}s| > \epsilon\ell\}$. If the stronger Markov additive property (17) holds then by the law of large numbers, $\tilde{\mathcal{W}}^\infty[\ell]/\ell \to \tilde{d}$. This limit could also hold even if the stronger Markov additive property doesn't. Either way $\Pr_x\{\mathcal{C}_\ell\} \to 0$ so we can apply Theorem 8. ☐

Proof of Corollary 2. Since $\mathcal{T}_\ell^\infty/\ell \to 1/\tilde{d}_1$, we have $\tilde{\mathcal{W}}^\infty[\mathcal{T}_\ell^\infty]/\mathcal{T}_\ell^\infty \to \tilde{d}/\tilde{d}_1$. Let

$$C_\ell = \chi\{|(\tilde{W}[T_\ell]/T_\ell - \tilde{d}/\tilde{d}_1| > \epsilon\}.$$

Again, $\Pr_x\{\mathcal{C}_\ell\} \to 0$ so we can apply Theorem 8. ☐

Proof of Corollary 3. Since φ is the stationary distribution of $\hat{\mathcal{W}}^\infty$ it follows that

$$\sum_{k=0}^{n} f(\hat{\mathcal{W}}^\infty[k])/n \to \sum_{\hat{y}} f(\hat{y})\varphi(\hat{y}) \text{ almost surely.}$$

Let

$$C_\ell = \chi\{|\frac{1}{T_\ell}\sum_{k=0}^{T_\ell} f(W[k]) - \sum_{\hat{y}} f(\hat{y})\varphi(\hat{y})| > \epsilon\}.$$

Again, $\Pr_x\{\mathcal{C}_\ell\} \to 0$ so we can apply Theorem 8. ☐

5 Analysis of Asymptotics of Join the Shortest Queue

Now we apply the results of Section 4 to the two server join the shorter queue system. To ensure irreducibility, assume $\lambda_1 + \gamma > 0, \lambda_2 + \gamma > 0, \mu_1 > 0$, and $\mu_2 > 0$.

Our first step is to uniformize $Q(t)$, and let $Q[n] = (Q_1[n], Q_2[n])$ denote the state at the nth step of the discrete time Markov chain. For convenience, measure time in units such that $\lambda_1 + \lambda_2 + \gamma + \mu_1 + \mu_2 = 1$. Thus, these parameters can be interpreted both as rates and as transition probabilities, as shown in Fig. 2.

The second step to make a guess as to the direction taken when the total number of customers reaches some large level ℓ. We originally guessed that if ρ were strictly larger than ρ_1 and ρ_2, sample paths reaching the rare event of interest would hug the diagonal. We refer to this case as a pooled network and analyze it in Section 5.1. The actual behavior turned out to be more subtle.

If ρ_1 were strictly larger than ρ and ρ_2, then we guessed the most likely approach to a large number of customers would bounce along the horizontal axis, and similarly, for ρ_2 along the vertical axis. These cases are referred to as the unpooled cases and analyze them in Subsection 5.5.

The remaining steps for each of the cases is as follows:

Step 3: Define $W[n]$, which will be a relabelling of $Q[n]$. Of course, $\tilde{W}_1[n] = Q_1[n] + Q_2[n]$. The remaining components depend upon the conjectured behavior.

Step 4: Specify $\vartriangle$ and $\blacktriangle$ such that the resulting free chain W^∞, defined by C.2, satisfies C.3, C.1 and that there exists $L(\hat{y})$ satisfying C.11. We will not mention $L(\cdot)$ since it is easy to define $L(\cdot)$ in all of the cases we analyze.

Step 5: Find a harmonic function satisfying C.4 and use it to construct the twisted free process in C.5. Choose a reference state δ and determine the periodicity. Verify C.6 and compute $\varphi(\cdot)$.

Step 6: Verify C.7–C.10 and C.12.

Step 7: Write down the results.

5.1 Pooled Servers

If ρ were the largest, we initially guessed that the servers would pool, and the most likely way to reach a large number ℓ of customers in the system would be along the diagonal so that both queues become large, but their difference should remain small. Even though this guess turned out to be naive, we briefly describe the evolution of the steps.

Step 3: One alternative for defining is $\tilde{W}_2[n] = Q_1[n]$ and $\hat{W}[n]$ to be degenerate. However, a better alternative would be to define $\hat{W}[n]$ as $(Q_1[n] - Q_2[n])$. The alternative is better in the sense that the more information in $\hat{W}[n]$, the stronger the results provided all of the conditions can be verified. Hence, let $W[n] = (\tilde{W}[n], \hat{W}[n]) = (Q_1[n] + Q_2[n], Q_1[n] - Q_2[n])$. Note that $S = \{(\tilde{w}_1, \hat{w}) = (x + y, x - y); x, y \geq 0\}$ which implies S is a subset of $\mathbb{Z}_+ \times \mathbb{Z}$.

Step 4: The free chain is easily constructed by using the transition probabilities of W corresponding to when both queues are busy. Thus, $\vartriangle$ corresponds to at least one of the queues being idle, and Θ corresponds to the states in which both servers are busy. Let $\blacktriangle = \mathbb{Z}^2 \backslash \Theta$. The state space of the free (and twisted processes) is $\mathbb{Z}^2$. Some of the states around the origin are shown in Fig. 3. The states in $\vartriangle$ are labelled in the obvious way. The leftmost point labelled $\vartriangle$ is the origin. The states in $\blacktriangle \backslash \vartriangle$ are labelled with $\blacktriangle$. The points labelled with $\circ$ belong to Θ (and also to B_0). The W-chain lives on S; i.e., the points labelled with either $\vartriangle$ or $\circ$. Column (a) shows the transition probabilities for the free process used from any point in $\mathbb{Z}^2$ depending on whether the state is above, on, or below the x-axis. At first, the points labelled $\blacktriangle$ interspersed among $S \backslash \vartriangle$, e.g. (1,0), would appear to be useless since neither the W-chain nor the free chain starting on $\vartriangle$ would ever visit these points. However, they play a useful role in Lemma 5 and Theorem 7.

Step 5: A harmonic function is $h(x) = \rho^{-x_1}$ where x_1 is the total number of customers in the system. Note that $\hat{h}(\hat{x}) \equiv 1$ in this case. Column (b) of Fig. 3 shows the transition probabilities for the twisted free process. Choose $\delta = (0, 0)$. To be in state $(s, 0)$, s must be even. Clearly, $\hat{\mathcal{W}}^\infty$ is 2-periodic! To verify C.6, we would need to show that $\hat{\mathcal{W}}^\infty$, the difference in the queue lengths for the twisted process, is positive recurrent. If the difference between the lengths of Queue 1 and Queue 2 is positive, it is clear from Fig.3 that the expected drift of $\hat{\mathcal{W}}^\infty$ is $\rho\mu_2 + \lambda_1/\rho - \rho\mu_1 - (\lambda_2 + \gamma)/\rho$. If the difference between the lengths of Queue 1 and Queue 2 is negative, the expected drift of $\hat{\mathcal{W}}^\infty$ is $\rho\mu_2 + (\lambda_1 + \gamma)/\rho - \rho\mu_1 - \lambda_2/\rho$. For $\hat{\mathcal{W}}^\infty$ to be recurrent, we need both drifts to be towards zero; i.e., we need $\gamma > |\rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2)|$. However, it is possible to construct examples in which $\rho > \rho_1 \vee \rho_2$, but $\gamma \leq |\rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2)|$. Thus, C.6 did not hold; and, we realized that our guess about the way that the rare event would be approached when ρ was the largest needed to be refined.

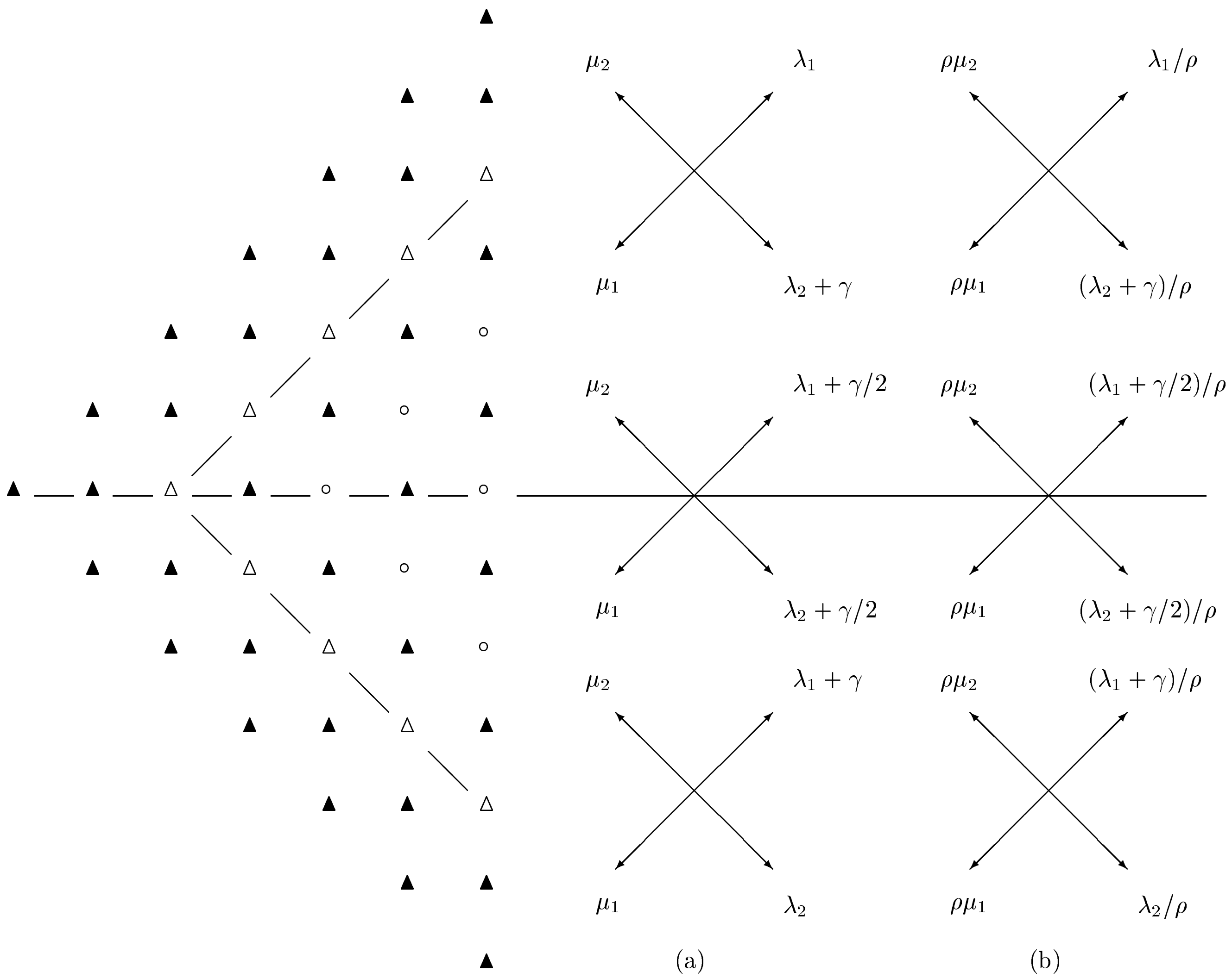

Figure 3: The transitions of (a) K^∞ and (b) $\mathcal{K}^\infty$ under pooling.

Thus, we need to split this case into two cases. If the difference in the queue lengths of the twisted free process is positive recurrent, we will refer to it as strongly pooled and analyze it in Section 5.2. If the difference is not positive recurrent, we refer to it as weakly pooled and analyze it in Section 5.3.

5.2 Strongly Pooled Servers

For the strongly pooled network, we are assuming in addition that $\gamma > |\rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2).|$ Steps 3, 4, 5, and verifying C.6 are identical to the arguments of the pooled section except that C.6 now holds due to the added condition. Not only is $\hat{\mathcal{W}}^\infty$ positive recurrent, it is simple to show that the stationary distribution φ of $\hat{\mathcal{W}}^\infty$ is given by (2).

Step 6: The remaining properties that need to be verified are:

C.7: $\tilde{\mathcal{W}}^\infty$ increases by 1 with probability $\mu_1 + \mu_2$ and decreases by 1 with probability $\lambda_1 + \lambda_2 + \gamma$ so

$$\tilde{d}_1 = \mu_1 + \mu_2 - (\lambda_1 + \lambda_2 + \gamma)$$

and $0 < \tilde{d}_1 < \infty$.

C.8: This condition is automatic since $\tilde{d}_1$ is positive and $\hat{\mathcal{W}}^\infty$ drifts towards 0. There is some probability $\mathcal{W}^\infty$ can reach a point $(L,0)$ for L arbitrarily large. It follows that if $\mathcal{W}^\infty$ were to hit $\blacktriangle$ from $(L,0)$ then the law of large numbers would apply. Given that $\hat{\mathcal{W}}^\infty$ drifts towards 0 it is therefore impossible that $\mathcal{W}^\infty$ hit $\blacktriangle$ with probability one.

C.9: This condition is automatic since $\hat{h} \equiv 1$.

C.10: The boundary Δ that was removed to transform the W-chain into the free chain corresponds to the states of the form $(x,0)$ and $(0,y)$ in the Q-chain. Thus, the analog to C.10 for the Q-chain is that

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

Thus, we need to find a Lyapunov function V such that $K_Q V - V < -f$ for "most" states $(x,y) \in \mathbb{Z}_+^2$ and $f(x,y) = \rho^{-(x+y)}\chi\{x = 0 \text{ or } y = 0\}$. By "most" states, we mean for all but a finite number of states. We leave this to Subsection 5.4.

C.12: This follows from C.10 since $\sum_{x\in\Delta} < \infty$.

Step 7: We may now draw our conclusions.

Theorem 9. *If $\rho > \max\{\rho_1, \rho_2\}$ and $\gamma > |\rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2)|$ we have:*

$$\pi_W((\ell, \hat{y})) \sim 2f(0)\rho^\ell \frac{1}{\tilde{d}_1}\varphi(\hat{y})\chi\{\hat{y} = \ell \mod 2\}$$

where φ is given at (2), where

$$f(0) = g = g(0) \equiv \sum_{x=1}^{\infty}\left[\pi_W(x,x)\rho^{-x}H(x,x) + \pi_W(x,-x)\rho^{-x}H(x,-x)\right] \tag{36}$$

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

Remark 3. *We are in the special case mentioned after* (23) *so $f(m) = g(m)$ for $m = 0, 1$. Also $f(1) = g(1) = 0$ since B_1 is not accessible from Δ.*

This means the stationary measure is a product for large ℓ. The constant $f(0)$ can only be obtained by simulation. This is not too onerous because we only need π_W near the origin.

Since we are taking $\delta = (0,0)$, we note that the sets A_0 and A_1 are the even and odd integers respectively. The chain $\{\hat{\mathcal{W}}^\infty[\mathcal{T}_\ell^\infty], \ell = 0, 1, \ldots\}$ is a periodic Markov chain with steady state μ which jumps back and forth between A_0 and A_1. If follows that $\mu(0, A_0) = \mu(0, A_1) = 1/2$. Now applying Theorem 7 we get:

Theorem 10. *Under the conditions of Theorem 9, as $\ell \to \infty$,*

$$\Pr_\sigma\left\{\hat{W}[T_\ell] = \hat{y} \mid T_\ell < T_\delta\right\} \sim \quad 2\mu(0,\hat{y})\chi\{\hat{y} = \ell \mod 2\}$$

for any initial point σ.

Theorem 11. *Under the conditions of Theorem 9,*

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

where σ is any initial point in S and $g = f(0)$ was given in Theorem 9.

5.3 Weakly Pooled Servers

In addition to stability, for the weakly pooled case we also assume that ρ is the largest and that $\gamma \leq |\rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2).|$

Step 3: Since we are guessing that $Q_1[T_\ell]$, $Q_2[T_\ell]$, and their difference all become large with ℓ, we pick one of them, and let $\tilde{W}[n] = (Q_1[n] + Q_2[n], Q_1[n] - Q_2[n])$ and $\hat{W}$ be trivial.

Step 4: Since $\hat{W}$ is trivial, to obtain the stronger Markov additive structure (17), we need to choose the transition structure either above, on, or below the x-axis for the free process. Initially we labelled the queues so that queue 1 grew faster and chose the transition structure above the axis, i.e., the topmost transition structure in column (a) of Fig. 3. Consequently all the points labelled "$\circ$" on or below the axis would become part of $\vartriangle$. When we did this, not only were we unable to verify C.10 in Step 6, we suspect that it does not hold. Instead we resorted to leaving the transition structure in column (a) intact which satisfies the weaker Markov additive structure (16). The set $\vartriangle$ remains the same as in the strongly pooled case, and we were able to verify C.10 for this process. However, to obtain a joint bottlenecks result, we need an additional argument.

Step 5: The harmonic function is the same as for the strongly pooled case. Since $\hat{S}$ has only one state, $\hat{W}$ is positive recurrent and aperiodic.

Step 6: To verify the remaining conditions:

C.7: This is the same as in the strongly pooled case.

C.8: If $\rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2) \geq \gamma$ then necessarily $\tilde{\mathcal{W}}_2^\infty$ has a non-negative drift. On the other hand $\tilde{\mathcal{W}}_2^\infty$ hits $\blacktriangle$ with probability strictly less than one. This follows from the law of large numbers and the fact that the slope of the drift is less than one. Above the x-axis this slope is

$$(\rho\mu_2 + \rho^{-1}\lambda_1 - (\rho\mu_1 + \rho^{-1}(\lambda_2 + \gamma))/(\mu_1 + \mu_2 - (\lambda_1 + \lambda_2 + \gamma)).$$

and this is less than one if $\rho^2 < (\lambda_2 + \gamma)/\mu_2$. This is true because $\rho < (\lambda_2 + \gamma)/\mu_2$ since

$$\rho = \frac{\mu_1}{\mu_1 + \mu_2}\frac{\lambda_1}{\mu_1} + \frac{\mu_2}{\mu_1 + \mu_2}\frac{\lambda_2 + \gamma}{\mu_2}$$

and $\rho > \rho_1$ in the pooling case.

If $\rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2) \leq -\gamma$ then necessarily $\tilde{\mathcal{W}}_2^\infty$ has a non-positive drift. On the other hand $\tilde{\mathcal{W}}_2^\infty$ hits $\blacktriangle$ with probability strictly less than one. This follows from the law of large numbers and the fact that the slope of the drift is greater than minus one. The slope below the x-axis is

$$(\rho\mu_2 + \rho^{-1}(\lambda_1 + \gamma) - (\rho\mu_1 + \rho^{-1}\lambda_2)/(\mu_1 + \mu_2 - (\lambda_1 + \lambda_2 + \gamma)).$$

and this is greater than minus one if $\rho^2 < (\lambda_1 + \gamma)/\mu_1$. Again, this is true because $\rho < (\lambda_1 + \gamma)/\mu_1$ since

$$\rho = \frac{\mu_1}{\mu_1 + \mu_2}\frac{\lambda_1 + \gamma}{\mu_1} + \frac{\mu_2}{\mu_1 + \mu_2}\frac{\lambda_2}{\mu_2}$$

and $\rho > \rho_2$.

C.9, C.10, and C.12: These are identical to those in the strong pooling case.

Step 7: We can now draw our conclusions. The asymptotics of the stationary distribution given in Theorem 3 follow because φ is a trivial measure. $\mathrm{E}_\delta T_\ell$ is the same as in Theorem 11.

If $\rho^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2) \geq \gamma$ then necessarily $\tilde{\mathcal{W}}_2^\infty$ has a non-negative drift so using the argument from C.8, we get

$$\lim_{\ell \to \infty}\left(\frac{Q_1[T_\ell] + Q_2[T_\ell]}{\ell}, \frac{Q_1[T_\ell] - Q_2[T_\ell]}{\ell}\right) = \left(1, \frac{\lambda_1/\rho - (\lambda_2 + \gamma)/\rho - \rho\mu_1 + \rho\mu_2}{\lambda_1/\rho - \rho\mu_1 + (\lambda_2 + \gamma)/\rho - \rho\mu_2}\right).$$

When $\tilde{\mathcal{W}}_2^\infty$ has a non-positive drift we get a similar result. Rewriting these gives the rest of Theorem 3.

5.4 Checking C.10

As a special case of the following proposition, we have

$$\sum_{x\geq 0} \rho^{-x}\pi_Q(x,0) < \infty \text{ and } \sum_{y\geq 0} \rho^{-y}\pi_Q(0,y) < \infty$$

which establishes C.10 for the strongly and weakly pooled cases.

Proposition 1. *If $\rho > \max\{\rho_1,\rho_2\}$ and $\rho < 1$, then*

$$\sum_{x\geq 0, y\geq 0} \rho^{-\sqrt{x^2+y^2}}\pi_Q(x,y) < \infty. \tag{37}$$

Proof. To establish (37), consider the Lyapunov function $V(x,y) = \rho^{-\sqrt{x^2+y^2}}$. We now calculate $K_QV(x,y) - V(x,y)$ for $x, y \geq 0$ and we will show

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

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

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

Consequently by Theorem 14.3.7 in Meyn and Tweedie (1993)

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

Except at the origin, $K_QV(x,y) - V(x,y)$ can be expressed as

$$[K_QV - V](x,y) = \begin{cases} \lambda_1\Delta_xV(x,0) - \mu_1\Delta_xV(x-1,0) + (\lambda_2+\gamma)\Delta_yV(x,0) & \text{if } 0 = y < x, \\ (\lambda_1+\gamma)\Delta_xV(0,y) + \lambda_2\Delta_yV(0,y) - \mu_2\Delta_yV(0,y-1) & \text{if } 0 = x < y, \\ (\lambda_1+\gamma/2)\Delta_xV(x,y) - \mu_1\Delta_xV(x-1,y) + (\lambda_2+\gamma/2)\Delta_yV(x,y) - \mu_2\Delta_yV(x,y-1) & \text{if } 0 < x = y, \\ \lambda_1\Delta_xV(x,y) - \mu_1\Delta_xV(x-1,y) + (\lambda_2+\gamma)\Delta_yV(x,y) - \mu_2\Delta_yV(x,y-1) & \text{if } 0 < y < x, \\ (\lambda_1+\gamma)\Delta_xV(x,y) - \mu_1\Delta_xV(x-1,y) + \lambda_2\Delta_yV(x,y) - \mu_2\Delta_yV(x,y-1) & \text{if } 0 < x < y, \end{cases}$$

where $\Delta_xV(x,y) \equiv V(x+1,y) - V(x,y)$ and $\Delta_yV(x,y) \equiv V(x,y+1) - V(x,y)$. Now we have to show that in each of the 5 cases, $K_QV(x,y) - V(x,y) < 0$ for x or y sufficiently large.

In the first case when $0 = y < x$,

$$K_QV(x,0) - V(x,0) = \rho^{-x}\left(\lambda_1(\rho^{-1}-1) + \mu_1(\rho-1) + (\lambda_2+\gamma)(\rho^{-(\sqrt{x^2+1}-x)}-1)\right)$$

Note that

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

Hence, for x large enough, this term is arbitarily small. It therefore suffices to show $\lambda_1(\rho^{-1}-1) + \mu_1(\rho-1) < 0$. This follows by noting that for $s > 0$, the function $\lambda_1 s + \mu_1 s^{-1} - (\lambda_1+\mu_1)$ is strictly convex and has zeros at 1 and ρ_1^{-1}. Since $1 < \rho^{-1} < \rho_1^{-1}$ in the pooled case, we have completed Case 1.

The second case when $0 = x < y$ follows similarly since the problem is completely symmetric in x and y; the analogous condition needed is $1 < \rho^{-1} < \rho_2^{-1}$ which also holds by hypothesis. In the last 3 cases, we find it more convenient to use polar coordinates. If $x = r\cos(\theta)$, $y = r\sin(\theta)$, then

$$\begin{aligned}
\Delta_x V(x,y) &= \rho^{-\sqrt{(x+1)^2+y^2}} - \rho^{-\sqrt{x^2+y^2}} \\
&= (\rho^{(-\sqrt{1+r^2+2r\cos\theta})} - \rho^{-r} \\
&\le \rho^{-r}(\rho^{-(\cos\theta+1/2r)} - 1) \\
-\Delta_x V(x-1,y) &= \rho^{-\sqrt{(x-1)^2+y^2}} - \rho^{-\sqrt{x^2+y^2}} \\
&= (\rho^{(-\sqrt{1+r^2-2r\cos\theta})} - \rho^{-r} \\
&\le \rho^{-r}(\rho^{\cos\theta-1/2r} - 1)
\end{aligned}
\qquad
\begin{aligned}
\Delta_y V(x,y) &= \rho^{-\sqrt{x^2+(y+1)^2}} - \rho^{-\sqrt{x^2+y^2}} \\
&= (\rho^{(-\sqrt{1+r^2+2r\sin\theta})} - \rho^{-r} \\
&\le \rho^{-r}(\rho^{-(\sin\theta+1/2r)} - 1) \\
-\Delta_y V(x,y-1) &= \rho^{-\sqrt{x^2+(y-1)^2}} - \rho^{-\sqrt{x^2+y^2}} \\
&= (\rho^{(-\sqrt{1+r^2-2r\sin\theta})} - \rho^{-r} \\
&\le \rho^{-r}(\rho^{(\sin\theta-1/2r)} - 1)
\end{aligned}$$

where we have repeatedly factored out r and used $\sqrt{1+x} \le 1 + x/2$ for $x > -1$.

Using these inequalities in the third case when $0 < x = y$ yields

$$K_Q V(x,y) - V(x,y) \le \rho^{-r}\left[\rho^{-1/(2r)}\left((\lambda_1+\gamma/2)\rho^{-\cos\theta} + \mu_1\rho^{\cos\theta} + (\lambda_2+\gamma/2)\rho^{-\sin\theta} + \mu_2\rho^{-\sin\theta}\right) - 1\right]$$

where $\theta = \pi/4$. Since $\rho^{-1/(2r)} \searrow 1$, it suffices to check that

$$(\lambda_1+\lambda_2+\gamma)\rho^{-1/\sqrt{2}} + (\mu_1+\mu_2)\rho^{1/\sqrt{2}} - 1 < 0.$$

But this follows since $(\lambda_1+\lambda_2+\gamma)s + (\mu_1+\mu_2)s^{-1} - 1$ for $s > 0$ is strictly convex with zeros at 1 and ρ^{-1} and $\rho^{-1/\sqrt{2}}$ lies between the two zeros.

Now consider the fourth case when $0 < y < x$. Again after converting to polar coordinates and bounding, we have

$$K_Q V(x,y) - V(x,y) \le \rho^{-r}\left[\rho^{-1/(2r)}\left[\lambda_1\rho^{-\cos\theta} + \mu_1\rho^{\cos\theta} + (\lambda_2+\gamma)\rho^{-\sin\theta} + \mu_2\rho^{\sin\theta}\right] - 1\right].$$

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

$$f(\theta) \equiv (\lambda_1\rho^{-cos\theta} + \mu_1\rho^{\cos\theta} + (\lambda_2+\gamma)\rho^{-\sin\theta} + \mu_2\rho^{\sin\theta}) - 1$$

is always negative for $0 \le \theta \le \pi/4$ radians.

First, rewrite f as

$$f(\theta) = f_1(\rho^{-\cos(\theta)}) + f_2(\rho^{-\sin(\theta)}) \tag{38}$$

where

$$\begin{aligned}
f_1(x) &\equiv \lambda_1 x + \mu_1 x^{-1} - (\lambda_1+\mu_1) \\
f_2(x) &\equiv (\lambda_2+\gamma)x + \mu_2 x^{-1} - (\lambda_2+\gamma+\mu_2).
\end{aligned}$$

We will find functions $\ell_1(x)$ and $\ell_2(x)$ which are upper bounds for $f_1(x)$ and $f_2(x)$ over the regions of interest. Then the last step will be to show that the last inequality in the following holds:

$$f(\theta) \le \ell_1(\rho^{-\cos(\theta)}) + \ell_2(\rho^{-\sin(\theta)}) \equiv \ell(\theta) < 0 \text{ for } 0 \le \theta \le \pi/4.$$

To find the upper bounds, note that $f_1(x)$ and $f_2(x)$ are strictly convex on $x > 0$ since the second derivatives are strictly positive. The zeros of $f_1(x)$ occur at $x = 1$ and $x = \rho_1^{-1}$; the zeros of $f_2(x)$ occur at $x = \mu_2/(\lambda_2+\gamma)$ and $x = 1$. We are interested in $f_1(x)$ for $x \in [\rho^{-1/\sqrt{2}}, \rho^{-1}]$, and $f_2(x)$ for $x \in [1, \rho^{-1/\sqrt{2}}]$. Let $\ell_1(x)$ be the line that agrees with $f_1(x)$ at the endpoints of its region of interest $[\rho^{-1/\sqrt{2}}, \rho^{-1}]$; similarly, define $\ell_2(x)$ to be the line that agrees with $f_2(x)$ at the endpoints of $[1, \rho^{-1/\sqrt{2}}]$. By convexity, each line $\ell_i(x)$ is an upper bound for $f_i(x)$ over its region of interest. Note that $\ell_1(x) < 0$ for $x \in [\rho^{-1/\sqrt{2}}, \rho^{-1}]$ since the endpoints of the region fall between the zeros of $f_1(x)$; i.e., $1 < \rho^{-1/\sqrt{2}} < \rho^{-1} < \rho_1^{-1}$. Now we consider two subcases. The easier case occurs when $\rho^{-1/\sqrt{2}} \le \mu_2/(\lambda_2+\gamma)$, since then $\ell_2(x) \le 0$ for $x \in [1, \rho^{-1/\sqrt{2}}]$. To see that $\ell_2(x) \le 0$ over this region, note that the left endpoint of the region is a zero of $f_2(x)$; i.e., $\ell_2(1) = f_2(1) = 0$. Hence, the line $\ell_2(x) \le 0$ for $x \in [1, \rho^{-1/\sqrt{2}}]$ if its less than or equal to zero at its right endpoint; i.e., if the right endpoint lies between the zeros of $f_2(x)$; i.e, if

$\rho^{-1/\sqrt{2}} \leq \mu_2/(\lambda_2+\gamma)$. Since $\ell_1 < 0$ and $\ell_2 \leq 0$ over their regions of interest, it follows that $\ell(\theta) < 0$ for $0 \leq \theta \leq \pi/4$ which completes the easier case. The more difficult case occurs when $\ell_2(x) > 0$ for $x \in (1, \rho^{-1/\sqrt{2}}]$; i.e., $\ell_2(x)$ has a positive slope. First note that $\ell(0) = f(0) = f_1(\rho^{-1}) < 0$, and $\ell(\pi/4) = f(\pi/4) < 0$ 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 $\pi/4$ with $\ell(\theta_0) \geq 0$. Since θ_0 is a local maximum,

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

where s_i is the slope of $\ell_i(x)$ and in this subcase $s_2 > 0$. It follows that

$$0 < \frac{\sin(\theta_0)}{\cos(\theta_0)}\rho^{\sin(\theta_0)-\cos(\theta_0)} = \frac{s_2}{s_1}$$

and since $s_2 > 0$ it follows that $s_1 > 0$.

This leads to a contradiction since

$$\begin{aligned}
\ell''(\theta_0) &= \frac{d^2}{d\theta^2}\ell_1(\rho^{-\cos(\theta)})|_{\theta=\theta_0} + \frac{d^2}{d\theta^2}\ell_2(\rho^{-\sin(\theta)})|_{\theta=\theta_0} \\
&= s_1(\ln(\rho)\sin(\theta_0))^2\rho^{-\cos(\theta_0)} + s_2(-\ln(\rho)\cos(\theta_0))^2\rho^{-\sin(\theta_0)} \\
&+ s_1\ln(\rho)\cos(\theta_0)\rho^{-\cos(\theta_0)} + s_2\ln(\rho)\sin(\theta_0)\rho^{-\sin(\theta_0)} \\
&> 0
\end{aligned}$$

which implies that θ_0 is a local minimum. Consequently, both $\ell(\theta)$ and $f(\theta)$ are strictly negative on the interval $0 \leq \theta \leq \pi/4$.

Case 5 when $0 < y < x$ follows from Case 4 by symmetry since the problem is symmetric in x and y. In Case 4, we used $\rho > \rho_1$ to show that $\ell_1(x) < 0$ over its region of interest. In Case 5, the analogous condition is $\rho > \rho_2$ and both of theses conditions hold by hypothesis. □

5.5 Unpooled Servers

In this section we suppose $\rho < \max\{\rho_1, \rho_2\}$ and we assume $\rho_2 < \rho_1$. (We avoid the case of exactly equal loads since this leads to additional subtleties.) In such a model the first queue overloads but the service rate of the second queue is so fast that even though the smart customers all join the second queue it still remains in steady state.

5.6 Steps for the Unpooled Case

Step 3: Define $W[n] = (\tilde{W}[n], \hat{W}[n]) = (Q_1[n] + Q_2[n], Q_2[n]) \in \mathbb{Z}_+ \times \hat{S}$ since we expect $Q_2[n] \in \hat{S} \equiv \mathbb{Z}_+$ to be stable when the backlog overloads.

Step 4: Define the set $\blacktriangle = \{(x_1, \hat{y}) \in \mathbb{Z} \times \hat{S} : x_1 \leq 2\hat{y} + 1\}$. Note that $\blacktriangle$ includes all states of the W-chain corresponding to states in which queue 1 has at most one more customer than queue 2. The set $\Delta = \{(x_1, \hat{y}) \in \blacktriangle : x_1 \geq 0, x_1 \leq 2\hat{y}+1, \hat{y} \leq x_1\}$. Consider a Markov kernel K^∞ defined in Figure 4. This transition structure can be thought of as changing the smart customers into customers dedicated to queue 2 and allowing negative customers at queue 1. Denote the chain with kernel K^∞ by W^∞ and remark that away from $\blacktriangle$, W agrees with W^∞. Moreover W^∞ can be viewed as a Markov additive process where the total number of customers $\tilde{W}^\infty$ will be the additive process and $\hat{W}^\infty$, the queue length at two, is the Markovian component.

Step 5: It is easy to check that $h(x_1, \hat{y}) = \rho_1^{-x_1}\rho_1^{\hat{y}}$ is a harmonic function for the kernel of the free process K^∞. The kernel by $\mathcal{K}^\infty$ of the twisted random walk $\mathcal{W}^\infty$ given in Figure 4. In fact this twist reverses the service rate μ_1 and the arrival rate λ_1 of the first queue. The additive increments between the times when $\hat{\mathcal{W}}^\infty$ returns to 0 is obviously aperiodic since it is possible to return in one transition.

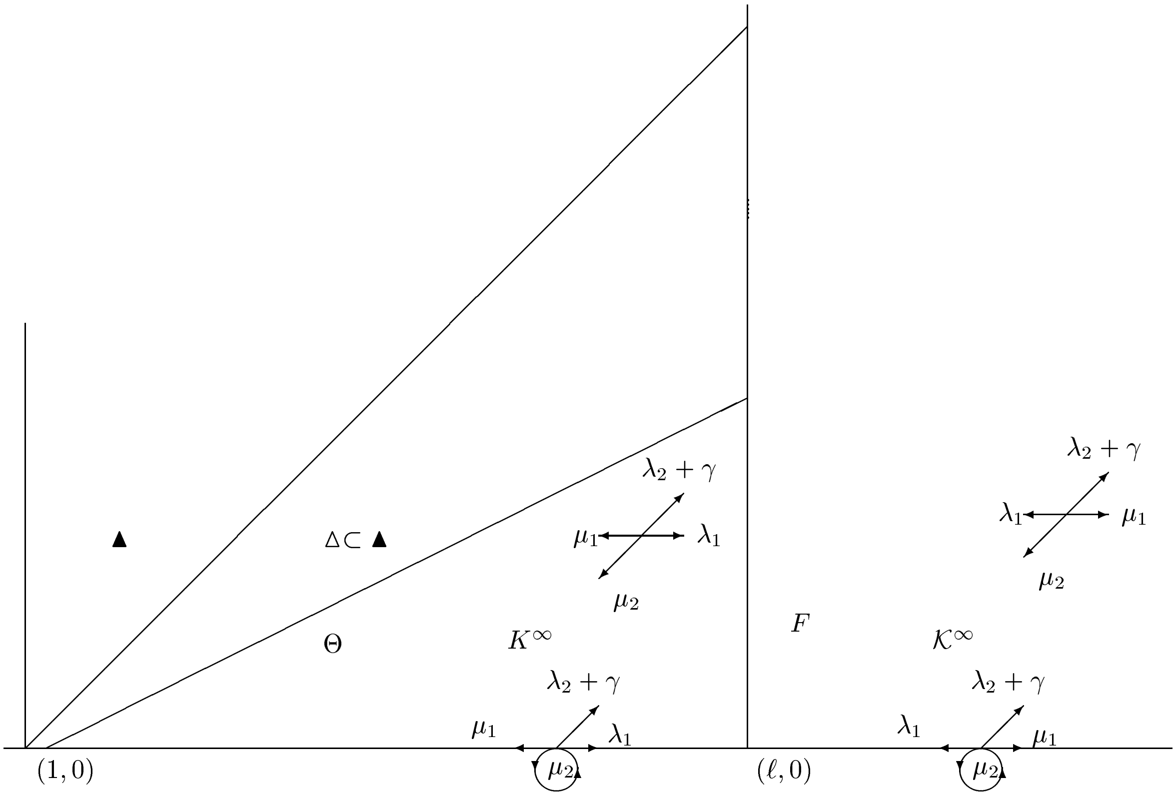

Figure 4: Transition probabilities of W^∞ and $\mathcal{W}^\infty$.

First, note that

$$\rho = \frac{\lambda_2+\gamma}{\mu_2}\frac{\mu_2}{\mu_1+\mu_2} + \frac{\lambda_1}{\mu_1}\frac{\mu_1}{\mu_1+\mu_2}.$$

Since $\rho_1 > \rho$ it follows that $\rho_{2\oplus\gamma} \equiv (\lambda_2+\gamma)/\mu_2 < \rho < \rho_1$. Consequently $\hat{\mathcal{W}}^\infty$ is stable and

$$\varphi(\hat{y}) = \left(1 - \frac{\lambda_2+\gamma}{\mu_2}\right)\left(\frac{\lambda_2+\gamma}{\mu_2}\right)^{\hat{y}}. \tag{39}$$

Step 6: We verify the remaining conditions:

C.7: It is easy to check that $\tilde{d}_1 = [1-\varphi(0)][(\lambda_2+\gamma+\mu_1)-(\lambda_1+\mu_2)]+\varphi(0)[(\lambda_2+\gamma+\mu_1)-\lambda_1] = \mu_1-\lambda_1 > 0$.

C.8: This condition follows from C.7 and the law of large numbers.

C.9: Notice that $\hat{h}(\hat{y}) = \rho_1^{-\hat{y}}$. Moreover, we showed above that $(\lambda_2+\gamma)/\mu_2 < \rho_1$ so $\rho_1^{-1}(\lambda_2+\gamma)/\mu_2 < 1$. This means $\sum \varphi(\hat{y})\hat{h}(\hat{y}) < \infty$.

C.10 The support of the measure λ is $\{(x_1,\hat{y}) \in \mathbb{Z}_+^2 \mid x_1 \le 2\hat{y}+1\}$. So it suffices to show that

$$\sum_{\hat{y}\ge 0} \pi_W(2\hat{y}+1,\hat{y})\rho_1^{-(2\hat{y}+1)}\rho_1^{\hat{y}} = \sum_{y\ge 0}\pi_Q(y+1,y)\rho_1^{-(y+1)} < \infty.$$

In fact to check C.12, we will have to prove more.

C.12: First note that $\hat{\lambda}(\hat{y}) = \sum_{x_1=0}^{2\hat{y}+1}\lambda(x_1,\hat{y})$. According to the discussion earlier and in [14] following Lemma 1.1, we need to find a $\hat{V}(\hat{y})$ such that

$$\hat{\mathcal{K}}^\infty\hat{V}(\hat{y}) - \hat{V}(\hat{y}) \le -\hat{h}(\hat{y}) + b\chi_C(\hat{y})$$

and $\sum_{\hat{y}\geq 0}\hat{V}(\hat{y})\hat{\lambda}(\hat{y})<\infty$ where C is a finite set and $b\geq 0$. Set $C=\{0\}$ and $\hat{V}(\hat{y})=\rho_1^{-\hat{y}}/|\psi_{2\oplus\gamma}(\rho_1^{-1})|$ where $\psi_{2\oplus\gamma}(s)=(\lambda_2+\gamma)(s-1)+\mu_2(s^{-1}-1)$. This works since

$$\begin{aligned}
\hat{\mathcal{K}}^{\infty}\hat{V}(\hat{y})-\hat{V}(\hat{y}) &= \left((\lambda_2+\gamma)(\rho_1^{-(\hat{y}+1)}-\rho_1^{-\hat{y}})+\mu_2\chi\{\hat{y}>0\}(\rho_1^{-(\hat{y}-1)}-\rho_1^{-\hat{y}})\right)/|\psi_{2\oplus\gamma}(\rho_1^{-1})| \\
&= \left(\rho_1^{-\hat{y}}((\lambda_2+\gamma)(\rho_1^{-1}-1)+\mu_2(\rho_1-1)+\mu_2(1-\rho_1)\chi\{\hat{y}=0\}\right)/|\psi_{2\oplus\gamma}(\rho_1^{-1})| \\
&= -\hat{h}(\hat{y})+\mu_2(1-\rho_1)/|\psi_{2\oplus\gamma}(\rho_1^{-1})|\chi\{\hat{y}=0\}
\end{aligned}$$

since $\psi_{2\oplus\gamma}(\rho_1^{-1})<0$.
It now suffices to check that

$$\sum_{\hat{y}\geq 0}\hat{V}(\hat{y})\hat{\lambda}(\hat{y})<\infty \text{ or equivalently } \sum_{\hat{y}\geq 0}\rho_1^{-\hat{y}}\sum_{x_1\leq 2\hat{y}+1}h(x_1,\hat{y})\pi_W(x_1,\hat{y})<\infty,$$

or equivalently

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

We leave this to Section 5.7.

Step 7: We may now draw our conclusions.

Theorem 12. *If $\rho<\max\{\rho_1,\rho_2\}=\rho_1$ we have:*

$$\pi_W((\ell,\hat{y}))\sim f\rho_1^{-\ell}\frac{1}{\tilde{d}_1}\rho_1^{-\hat{y}}\varphi(\hat{y})$$

where

$$f\equiv\sum_{\hat{y}\geq 0}\pi_W(2\hat{y}+1,\hat{y})\rho_1^{-(\hat{y}+1)}H(2\hat{y}+1,\hat{y}) \tag{40}$$

φ is given in (39), and $H(2\hat{y}+1,\hat{y})$ is the probability $\mathcal{W}^{\infty}$ starting at $(2\hat{y}+1,\hat{y})$ never hits $\blacktriangle$.

This means the stationary measure is a product for large ℓ. The constant f can only be obtained by simulation.

Theorem 13. *If $\rho<\max\{\rho_1,\rho_2\}=\rho_1$ then, as $\ell\to\infty$,*

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

where σ is some initial state and T_σ is the return time σ, where $\to^T$ denotes convergence in total variation and where $\mu(0,\cdot)$ denotes the stationary distribution of $\hat{\mathcal{W}}^{\infty}[\mathcal{T}_\ell^{\infty}],\ell=1,2,\ldots$ where $\mathcal{T}_\ell^{\infty}$ is the time first component of the twisted process first reaches ℓ.

Finally

Theorem 14. *If $\rho<\max\{\rho_1,\rho_2\}=\rho_1$ then*

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

where

$$g\equiv f\sum_{z\geq 0}\rho_1^{-z}\mu(0,z). \tag{41}$$

5.7 Checking C.12 for the unpooled network

In this section we establish the following proposition which establishes C.12 for the unpooled case.

Proposition 2. *If* $\rho < \rho_1 < 1$ *then*

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

Before proving the proposition, we introduce some notation and a lemma. Recall that $\rho_{2\oplus\gamma} \equiv (\lambda_2+\gamma)/\mu_2$. Since $\rho = \rho_1\mu_1/(\mu_1+\mu_2) + \rho_{2\oplus\gamma}\mu_2/(\mu_1+\mu_2)$, we know that if $\rho < \rho_1 < 1$ then $\rho_{2\oplus\gamma} < \rho$. Hence, if $\rho < \rho_1 < 1$, we have

$$1 < \rho_1^{-1} < \rho^{-1} < \rho_{2\oplus\gamma}^{-1} < \rho_2^{-1}. \tag{42}$$

Now define the following functions

$$\begin{aligned}\psi_M(s) &= (\lambda_1+\lambda_2+\gamma)(s-1) + (\mu_1+\mu_2)(s^{-1}-1),\\ \psi_i(s) &= \lambda_i(s-1) + \mu_i(s^{-1}-1) \text{ for } i = 1,2,\\ \psi_{2\oplus\gamma}(s) &= \lambda_{2\oplus\gamma}(s-1) + \mu_2(s^{-1}-1).\end{aligned}$$

which are strictly convex for $s > 0$. Furthermore,

$$\psi_M(s) < 0 \text{ for } s \in (1, \rho^{-1}), \tag{43}$$

$$\psi_i(s) < 0 \text{ for } s \in (1, \rho_i^{-1}), \text{ and} \tag{44}$$

$$\psi_{2\oplus\gamma}(s) < 0 \text{ for } s \in (1, \rho_{2\oplus\gamma}^{-1}). \tag{45}$$

In particular, if $\rho < \rho_1 < 1$ then $\psi_M(\rho_1^{-1})$, $\psi_2(\rho_1^{-1})$, and $\psi_{2\oplus\gamma}(\rho_1^{-1})$ are all strictly negative and $\psi_1(\rho_1^{-1}) = 0$.

Lemma 8. *If* $\rho < \rho_1 < 1$ *then*

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

Proof. Let $f(x,y) = \rho_1^{-y} > 0$, $c = -\psi_{2\oplus\gamma}(\rho_1^{-1}) > 0$, $V(x,y) = \rho_1^{-y}/c > 0$, and $s(x,y) = 1+(\lambda_2+\gamma)(\rho_1^{-1}-1)/c > 0$. To complete the proof, we need only show that $K_QV(x,y) - V(x,y) \le -f(x,y) + s(x,y)$ since by Theorem 14.3.7 of [15] we have $\pi_Q f < \pi_Q s$, and the latter is finite since s is bounded. Now,

$$K_QV(x,y) - V(x,y) = \begin{cases} \psi_2(\rho_1^{-1})\rho_1^{-y}/c & \text{for } y > x \ge 0,\\ ((\lambda_2+\gamma/2)(\rho_1^{-1}-1) + \mu_2(\rho_1-1))\rho_1^{-y}/c & \text{for } y = x > 0,\\ \psi_{2\oplus\gamma}(\rho_1^{-1})\rho_1^{-y}/c & \text{for } y > x > 0,\\ (\lambda_2+\gamma)(\rho_1^{-1}-1)/c & \text{for } x > y = 0,\\ (\lambda_2+\gamma/2)(\rho_1^{-1}-1)/c & \text{for } x = y = 0.\end{cases} \tag{46}$$

In the first case when $y > x \ge 0$, the r.h.s. of (46) simplifies to $-f(x,y)\psi_2(\rho_1^{-1})/\psi_{2\oplus\gamma}(\rho_1^{-1})$ which is less than $-f(x,y)+s(x,y)$ since $\psi_2(\rho_1^{-1})/\psi_{2\oplus\gamma}(\rho_1^{-1}) \ge 1$ and $s > 0$. In the second case when $y = x > 0$, use the inequality $((\lambda_2+\gamma/2)(\rho_1^{-1}-1) + \mu_2(\rho_1-1)) < \psi_{2\oplus\gamma}(\rho_1^{-1}) < 0$ to reduce the problem to considering the third case on the slightly larger region $y \ge x > 0$. However, in this third case, the r.h.s. of (46) simplifies to $-f(x,y)$. In the fourth case when $x > y = 0$, the r.h.s. is $-f(x,0) + s(x,0)$. The r.h.s. in the last case at the origin is smaller than the r.h.s. in the fourth case. Hence, $K_QV(x,y) - V(x,y) \le -f(x,y) + s(x,y)$. □

Proof of Proposition 2. Fix ϵ such that $1+\epsilon = \ln(\rho_{2\oplus\gamma}^{-1})/\ln(\rho_1^{-1})$ and note $\epsilon > 0$. Let $c = -\psi_M(\rho_1^{-1}) > 0$. Define

$$\begin{aligned} f(x,y) &= \rho_1^{-(x+y)}\chi\{x \le y+1\}/c\\ V(x,y) &= \begin{cases}\rho_1^{-(x+y)}/c & \text{for } 0 \le x, x-1 \le y\\ \rho_1^{-x(1+u(y/x))}/c & \text{for } 0 \le y < x-1\end{cases}\\ s(x,y) &= \left|(\lambda_1+\lambda_2+\gamma)(\rho_1^{-1}-1) + \mu_2(\rho_1-1)\right|\rho_1^{-y}/c + b\end{aligned}$$

where b is a positive constant whose value will be specified later, and u is a continuous, differentiable function with the following properties: There exists a $\delta > 0$ such that

$$\begin{aligned} u(s) &\geq 0 \\ u(s) &= 0 \text{ for } 0 \leq s \leq \delta/2 \\ u(s) &= 1 \text{ for } 1-\delta/2 \leq s \leq 1 \\ 0 \leq u'(s) &\leq 1+\epsilon \\ u(s) - su'(s) &\leq 0 \end{aligned}$$

At the end of the proof, we give a function u with these properties. Other than showing that such a u exists, we need only show that $K_Q V(x,y) - V(x,y) \leq -f(x,y) + s(x,y)$ since by Theorem 14.3.7 of [15] we have $\pi_Q f < \pi_Q s$, and the latter is finite. To see that $\pi_Q s < \infty$, use Lemma 8. Now we show that $K_Q V(x,y) - V(x,y) \leq -f(x,y) + s(x,y)$ in all cases except for the difficult region when $0 < y < x-1$ which we analyze later. Outside of $0 < y < x-1$,

$$K_Q V(x,y) - V(x,y) = \begin{cases} \psi_M(\rho_1^{-1})V(x,y)/c & \text{for } y \geq x-1 > 0, \\ \psi_1(\rho_1^{-1})V(x,y)/c & \text{for } x \geq y = 0, \\ (\lambda_1+\lambda_2+\gamma)\rho_1^{-1}/c & \text{for } y = x = 0, \\ \big((\lambda_1+\lambda_2+\gamma)(\rho_1^{-1}-1) + \mu_2(\rho_1 - 1)\big)\rho_1^{-y}/c & \text{for } y > x = 0 \end{cases} \tag{47}$$

In the first region $y \geq x-1 > 0$, the r.h.s. of (47) simplifies to $-V(x,y) \leq -f(x,y) \leq -f(x,y) + s(x,y)$ since $s(x,y) \geq 0$. In the second region, the r.h.s. is zero since $\psi_1(\rho_1^{-1}) = 0$. Fortunately, $f(x,y) = 0$ in this region also. In the third region, i.e. at the origin, the desired inequality will hold if we select $b \geq f(0,0) + (\lambda_1+\lambda_2+\gamma)\rho_1^{-1}/c$. In the fourth region when $y > x = 0$, we again have $f(x,y) = 0$ so we only need to show that the r.h.s. of (47) is smaller than $s(x,y)$, but this is straightforward.

Now we need to handle the difficult region $0 < y < x-1$. Let $h(x,y) = x(1+u(y/x))$. Since $u(s) = s$ for s in a neighbourhood of 1, it follows that $v(x,x-1) = \rho_1^{h(x,x-1)}$ except on the finite set $C = \{x : (x-1)/x < 1-\delta/2\}$. Therefore, for (x,y) such that $0 < y < x-1$ and $x \in C^c$,

$$\begin{aligned} &K_Q v(x,y) - v(x,y) \\ &= \lambda_1(\rho_1^{-h(x+1,y)} - \rho_1^{-h(x,y)}) + (\lambda_2+\gamma)(\rho_1^{-h(x,y+1)} - \rho_1^{-h(x,y)}) + \mu_1(\rho_1^{-h(x-1,y)} - \rho_1^{-h(x,y)}) + \mu_2(\rho_1^{-h(x,y-1)} - \rho_1^{-h(x,y)}) \\ &= \rho_1^{-h(x,y)}\left(\lambda_1\rho_1^{-(h(x+1,y)-h(x,y))} + (\lambda_2+\gamma)\rho_1^{-(h(x,y+1)-h(x,y))} + \mu_1\rho_1^{-(h(x-1,y)-h(x,y))} + \mu_2\rho_1^{-(h(x,y-1)-h(x,y))} - 1\right) \\ &= \rho_1^{-h(x,y)}\left(\lambda_1(\rho_1^{-\frac{\partial}{\partial x}h(x_+,y)} - 1) + (\lambda_2+\gamma)(\rho_1^{-\frac{\partial}{\partial y}h(x,y_+)} - 1) + \mu_1(\rho_1^{\frac{\partial}{\partial x}h(x_-,y)} - 1) + \mu_2(\rho_1^{\frac{\partial}{\partial y}h(x,y_-)} - 1)\right) \end{aligned} \tag{48}$$

where $x \leq x_+ \leq x+1$ and $x-1 \leq x_- \leq x$ and where y_+ and y_- are defined similarly.

Now we show that we can replace x_+ and x_- by x and y_+ and y_- by y in (48) and the change is negligible for x sufficiently large. Note that, $\frac{\partial}{\partial x}h(x,y) = 1 + u(y/x) - \frac{y}{x}u'(y/x)$. Hence,

$$\rho_1^{-\frac{\partial}{\partial x}h(x_+,y)} - \rho_1^{-\frac{\partial}{\partial x}h(x,y)} \text{ and } \rho_1^{\frac{\partial}{\partial x}h(x_-,y)} - \rho_1^{\frac{\partial}{\partial y}h(x,y)}$$

are arbitarily small for x large enough uniformly in y with $0 \leq x/y \leq 1$ using the the continuity of u'. Similarly, $\frac{\partial}{\partial y}h(x,y) = u'(y/x)$. Hence

$$\rho_1^{-\frac{\partial}{\partial y}h(x,y_+)} - \rho_1^{-\frac{\partial}{\partial y}h(x,y)} \text{ and } \rho_1^{\frac{\partial}{\partial y}h(x,y_-)} - \rho_1^{\frac{\partial}{\partial y}h(x,y)}$$

are arbitarily small for x large enough uniformly in y with $0 \leq x/y \leq 1$ using the continuity of u'. Hence, it suffices to show that the following is negative

$$\begin{aligned} &\left(\lambda_1(\rho_1^{-\frac{\partial}{\partial x}h(x,y)} - 1) + \mu_1(\rho_1^{\frac{\partial}{\partial x}h(x,y)} - 1)\right) + \left((\lambda_2+\gamma)(\rho_1^{-\frac{\partial}{\partial y}h(x,y)} - 1) + \mu_2(\rho_1^{\frac{\partial}{\partial y}h(x,y)} - 1)\right) \\ &\qquad = \psi_1(\rho_1^{-\frac{\partial}{\partial x}h(x,y)}) + \psi_{2\oplus\gamma}(\rho_1^{-\frac{\partial}{\partial y}h(x,y)}) \leq 0 \end{aligned}$$

since f is zero in this region and $b > 0$. First,

$$\begin{aligned}
\psi_{2\oplus\gamma}(\rho_1^{-\frac{\partial}{\partial y}h(x,y)}) &= \psi_{2\oplus\gamma}(\rho_1^{-u'(y/x)}) \\
&= \psi_{2\oplus\gamma}(\rho_{2\oplus\gamma}^{-u'(y/x)/(1+\epsilon)}) && \text{by the definition of } 1+\epsilon \\
&\leq \psi_{2\oplus\gamma}(\rho_{2\oplus\gamma}^{-1}) && \text{since } 0 \leq u' \leq 1+\epsilon \\
&= 0.
\end{aligned}$$

Next,

$$\begin{aligned}
\psi_1(\rho_1^{-\frac{\partial}{\partial x}h(x,y)}) &= \psi_1(\rho_1^{-(1+u(x/y)-(y/x)u'(x/y))}) \\
&\leq \psi_1(\rho_1^{-1}) && \text{since } u(s) - su'(s) \leq 0 \\
&= 0
\end{aligned}$$

Thus, b can be chosen to be the larger of $f(0,0)+(\lambda_1+\lambda_2+\gamma)\rho_1^{-1}/c$ and $\max_{0<y<x-1}\{f(x,y)+K_Q V(x,y)-V(x,y)\}$, both of which are finite. Hence, $K_Q V(x,y) - V(x,y) \leq -f(x,y) + s(x,y)$.

The only remaining step is to show that there exists a function u with the desired properties. Define $2\delta = \min\{1/2, \epsilon/(1+\epsilon)\}$. Now define

$$u(s) = \begin{cases} 0 & \text{for } s \leq \delta/2 \\ \frac{1-\delta}{2\delta(1-2\delta)}(s-\delta/2)^2 & \text{for } \delta/2 < s \leq 3\delta/2 \\ \frac{(1-\delta)(s-\delta)}{1-2\delta} & \text{for } 3\delta/2 < s \leq 1-3\delta/2 \\ s - \frac{1}{2(1-2\delta)}(s-(1-\delta/2))^2 & \text{for } 1-3\delta/2 < s \leq 1-\delta/2 \\ s & \text{for } s > 1-\delta/2 \end{cases}$$

We have chosen $\delta \leq 1/4$ to ensure that $3\delta/2 < 1/2$. The other constraint on δ ensures that $1/(1-2\delta) \leq 1+\epsilon$. We leave it to the reader to verify that u has the properties claimed. □

6 Acknowledgements

The authors wish to thank Stephen Turner and John Vande Vate for their comments. We also thank the referee's for their insightful reports.

References

[1] Adan, I. J., Wessels, J. and Zijm, W. H. M. (1991). Analysis of the asymmetric shortest queue problem, *Queueing Systems Theory Appl.,* **8** 1–58.

[2] Atar, R. and Dupuis, P. (1999) Large deviations and queueing networks: methods for rate function identification. *Stochastic Process. Appl.,* **84** 255–296.

[3] Alanyali, M., Hajek, B. (1998). On large deviations in load sharing networks, *Annals of Applied Probability,* **8**, 67–97.

[4] Alanyali, M., Hajek, B. (1998). On large deviations of Markov processes with discontinuous statistics, *Annals of Applied Probability,* **8**, 45–66.

[5] Asmussen, S. (1982). Conditioned limit theorems relating a random walk to its associate, with applications to risk reserve processes and the $GI/G/1$ queue. *Adv. in Appl. Probab.* **14**, 143–170.

[6] Brown, L. (1998). Asymptotic behaviour of an overloading queueing network with resource pooling. *Ph.D. Dissertation, Georgia Institute of Technology.*

[7] Cinlar, E. (1975). Introduction to Stochastic Processes. *Prentice Hall, N.J.*.

[8] CHERNOVA, N., FOSS, S. (1997). On The Stability of a Partially Accessible Multi-Station Queue with State-Dependent Routing, *submitted to Questa.*

[9] DUPUIS, P., ELLIS, R. S. (1995). The large deviation principle for a general class of queueing systems I. *TAMS,* **8**, 2689-2751.

[10] FLATTO, L. AND MCLEAN, H. P. (1977). Two Queues in Parallel, *Communications in Pure and Applied Math.,* **30**, 255–263.

[11] KESTEN, H. (1974). Renewals theory for functionals of a Markov chain with general state space, *Ann. Probab.* **2**, 355–386.

[12] KNESSL, C., MATKOWSKY,B. J., SCHUSS Z. AND TIER, C. (1986). Two parallel queues with dynamic routing, *IEEE Trans. Commun.,* **COM-34**, 1170-1175.

[13] MCDONALD, D. (1996). Overloading parallel servers when arrivals join the shortest queue. *Stochastic Networks: Stability and Rare Events,* P. Glasserman, K. Sigman and D. Yao Editors, Lecture Notes in Statistics 117, Springer Verlag, New York, 169-196.

[14] MCDONALD, D. (1998). Asymptotics of first passage times for random walk in a quadrant, *to appear Ann. Applied Probab.,*

[15] MEYN, S.P. AND TWEEDIE, R.L. (1993). *Markov Chains and Stochastic Stability,* Springer Verlag, New York.

[16] MEYN, S.P. (1997). The Policy Iteration Algorithm for Average Reward Markov Decision Processes with General State Spaces *manuscript.*

[17] SCHWARTZ, A., WEISS, A. (1993). Induced rare events: analysis via large deviations and time reversal. *Adv. Appl. Prob.* **25**, 667-689.

[18] SHWARTZ, A. AND WEISS, A. (1994). *Large deviations for performance analysis,* Chapman and Hall.

[19] TURNER, S. (1996). Large deviations for join the shorter queue, *manuscript.*

[20] VAN HOUTUM, G. J., ADAN, I. J. B. F., WESSELS, J., AND ZIJM, W. H. M. (1999). Performance Analysis of Parallel Identical Machines with a Generalized Shortest Queue Arrival Mechanism, *manuscript.*

R. D. Foley
School of Industrial and Systems Engineering
Georgia Institute of Technology
765 Ferst Drive
Atlanta, Georgia 30332-0205
Email: rfoley@isye.gatech.edu

D.R. McDonald
Department of Mathematics and Statistics
University of Ottawa
Ottawa, Ontario
Canada K1N 6N5
Email: dmdsg@omid.mathstats.uottawa.ca