A UNIFIED APPROACH TO FAST TELLER QUEUES AND ATM

B. Beck A. R. Dabrowski D. R. McDonald∗

October 19, 1998

Key words and phrases: Rare events, change of measure, \( h \)-transform, Asynchronous Transfer Mode, Hot spot simulation.
AMS 1991 subject classifications: Primary 60K25; Secondary 60F10, 60K20, 68M20, 90B22

Abstract

This paper examines a problem of importance to the telecommunications industry. In the design of modern ATM switches, it is necessary to use simulation to estimate the probability that a queue within the switch exceeds a given large value. Since these are extremely small probabilities, importance sampling methods must be used. Here we obtain a change of measure for a broad class of models with direct applicability to ATM switches.

We consider a model with \( A \) independent sources of cells where each source is modeled by a Markov renewal point process with batch arrivals. We do not assume the sources are necessarily identically distributed, nor that batch sizes are independent of the state of the Markov process. These arrivals join a queue served by multiple independent servers, each with service times also modeled as a Markov renewal process. We only discuss a time-slotted system. The queue is viewed as the additive component of a Markov additive chain subject to the constraint that the additive component remains nonnegative. We apply the theory in McDonald (1996) to obtain the asymptotics of the tail of the distribution of the queue size in steady state plus the asymptotics of the mean time between large deviations of the queue size.

Corresponding author;
André Dabrowski,
Mathematics and Statistics, University of Ottawa,
585 King Edward PO Box 450 STN A,
Ottawa Ontario, Canada, K1N 6N5.
ardsg@uottawa.ca

∗Research of the last two authors supported in part by NSERC grants.
1 Introduction

1.1 The engineering context: ATM Multiplexor.

ATM (Asynchronous Transfer Mode) is a prominent standard in modern telecommunications. Source signals (e.g., telephone or computer) are digitized, supplied with a destination address and sent to the ATM switch. This switch serves a number of sources. It reads the destination address and routes the packet of digitized signal through the appropriate output port of the switch. This routing must be accomplished with very little delay — e.g., a small delay between syllables in a telephone conversation would be unacceptable — even though incoming cells must be queued in a buffer before being served. A further loss of service quality may arise from cell losses due to overflowing buffers within the switch. In this paper we provide rapid simulation techniques to estimate the risk of a buffer overflowing under broad assumptions on the type of incoming traffic, and on the performance of the servers.

An ATM switch may be represented as a collection (usually a cascade) of discrete-time queues. The input to the switch is the superposition of independent, but not identical, arrival streams. Further, the arrival streams are not generally Poisson streams, but are "bursty". During a burst, packets of cells arrive at a fixed rate. The burst is followed by a silent period. Each queue is served deterministically: the servers will serve one cell in the queue per time unit. Although the entire system is stable (the global arrival rate is less than the global service rate), the arrival rate within bursts may exceed the available service rate.

A critical problem is to predict the performance of a switch before it is built. To decide whether or not the switch will meet industry standards, we must obtain the cell loss probability (CLP) of a given queue (the "hotspot") within the switch. This problem can be addressed by straightforward simulation, but the CLP is very small ($\sim 10^{-9}$), and so importance sampling methods must be used to reduce the simulation times. In these methods, the Markov chain describing the queues of the switch is "twisted" to provide a second system of transition probabilities for the switch. For this second transition kernel the switch is unstable and the queue overloads quickly. The CLP can now be obtained relatively quickly by a weighted average of trials for the twisted system.

In this paper we provide the appropriate twist for a specific queue within an ATM switch with Markov-modulated input streams. These will follow from more general results we establish on fast teller models, but we will return to ATM switches and we provide a complete solution to this practical problem.

1.2 ATM as a special case of fast teller models.

A fast teller system (generalizing the model in Sadowsky and Szpankowski (1995)) consists of a number of independent servers working in parallel to serve a single queue which collects arrivals from a number of independent sources. Our model is formulated for Markov-modulated arrivals and servers, where the inter-event waiting times are not necessarily geometric. The length of the hotspot queue within ATM switch can be embedded into a fast teller system by considering the fast teller queue to be the total workload directed at the hotspot buffer. The sources are then still Markov-modulated, but the services are deterministic.

A fast teller queue receives cells from $A$ independent sources and is served by $S$ independent servers. Denote the number of cells in the queue by $Q$. Source $a \in \{1 \ldots A\}$ supplies batches of cells (cus-
servers, for each server of an ATM multiplexor. The queue in all these models may be viewed as the additive component, and Szpankowski (1995). There is also an enormous literature of corresponding results for the queue. The most general results to date are for a fast teller system with heterogeneous servers by Sadowsky and Szpankowski (1995) or McDonald (1996) for recent results. Here, we will use that Markov additive asymptotics results when the queue size becomes large (see Bucklew (1990) for a survey and Sadowsky and Szpankowski (1995) or McDonald (1996) for recent results). Here, we will use that Markov additive structure to analyse a queue with multiple sources and servers. A more detailed comparison of results is found in Section 2.5.

A guide to the paper.

Many authors have studied the tail probabilities of the steady state of a generalized GI/G/1 queue. The most general results to date are for a fast teller system with heterogeneous servers by Sadowsky and Szpankowski (1995). There is also an enormous literature of corresponding results for the queue of an ATM multiplexor. The queue in all these models may be viewed as the additive component, \( Q \), of a Markov additive chain, \((Q, Z)\), subject to the constraint that the additive component remains nonnegative. This common Markov additive structure has been exploited by many authors to obtain asymptotics results when the queue size becomes large (see Bucklew (1990) for a survey and Sadowsky and Szpankowski (1995) or McDonald (1996) for recent results). Here, we will use that Markov additive structure to analyse a queue with multiple sources and servers. A more detailed comparison of results is found in Section 2.5.

The tail probabilities for \( Q \) will be obtained by studying those of the additive component, \( V^\infty \), of a related Markov additive process, \((V^\infty, Z^\infty)\). This process is defined from the first by removing the
boundary at 0 and allowing the “queue size” \( V^\infty \) to also take on negative values. The kernel \( K^\infty \) of the new process agrees with \( K \) when the queue size is positive, and extends that definition to negative queue sizes. The tail probabilities of \( Q \) and \( V^\infty \) are essentially the same. In the next section, we rephrase the general Markov additive framework of Ney and Nummelin (1987) for our \( V^\infty \).

We will use the tools developed in McDonald (1996) to give the asymptotic distribution of the queue length and mean time until the queue exceeds a given high level. We present a related method for accelerating the simulation of certain rare events associated with large deviations of the queue. In Section 3 we specialize to a model in which the arrival process is the sum of independent Markov modulated batches of arrivals and the queue is served by several independent Markov modulated batch servers.

We then specialize to the case when the batches of arrivals and the batched services occur according to independent but not necessarily identical renewal processes. This extends the fast teller system studied in Sadowsky and Szpankowski (1995) which only has one renewal stream of batched arrivals. Our model further permits inter-arrival times and block sizes to depend on the underlying Markov chain. This is the first result to allow for this dependence. In this case we verify all the conditions of Theorem 2.4 and 2.5 to obtain Theorem 4.2.

Finally, in Section 5, we apply our results to an ATM multiplexer. In Section 6 we discuss several theoretic and practical issues raised by the previous sections: effective bandwidths, and difficulties in simulation. We will use \( \mathbb{R} \) to represent the real numbers, \( \mathbb{Z} \) the integers, \( \mathbb{N}_0 \) the non-negative integers and, \( \mathbb{N} \) the strictly positive integers. Here \( I \{ A \} \) will represent the indicator function of the event \( A \). This paper draws on results from several papers, and the notation here is a compromise among the notations used in those source papers. We will use \([ \] \) to indicate dependence on time, and \( () \) for other functions. Variables and other elements belonging to the boundary-free processes will carry a \( \infty \) superscript.

## 2 General Framework

Let \( W^\infty := (V^\infty, Z^\infty) \) be a Markov additive (MA) chain defined on the probability space \((\Omega, \Sigma, P)\) with values in the measurable space \((\mathcal{E}^\infty, \mathcal{B}^\infty) := (\mathcal{E}_V^\infty \times \mathcal{E}_Z^\infty, \mathcal{B}_V^\infty \otimes \mathcal{B}_Z^\infty)\) where \( V^\infty := \{ V^\infty[n] : n \in \mathbb{N}_0 \} \), the additive part, is a family of random variables on \((\mathcal{E}_V^\infty, \mathcal{B}_V^\infty) := (\mathbb{Z}, \mathcal{B}(\mathbb{Z}))\), the integers and the associated maximal \( \sigma \)-algebra and where \( Z^\infty = \{ Z^\infty[n] : n \in \mathbb{N}_0 \} \) is the Markovian part. We will adopt the following notation for each point \( w \) of \( \mathcal{E}^\infty \): \( w = (q, z) \) where \( q \in \mathcal{E}_V^\infty \) and \( z \in \mathcal{E}_Z^\infty \). We define the increment associated with the jump of \( Z^\infty \) at time \( n \in \mathbb{N} \) to be \( \Delta V^\infty[n] := V^\infty[n] - V^\infty[n - 1] \). Then for all \( w, w' \in \mathcal{E}^\infty \), the kernel of \( W^\infty \) is defined by

\[
K^\infty(w, w') \equiv P \left[ \Delta V^\infty[1] = q' - q, Z^\infty[1] = z' \ \bigg| \ Z^\infty[0] = z \right] =: H^\infty \left( z, (q' - q, z') \right) \quad (2.1)
\]

where we assume that \( H^\infty \), the MA-transition kernel of \( W^\infty \), is measurable with respect to \( z \).

Note that the distribution of the additive increment from time \([n]\) to \([n+1]\) only depends on the value of the Markovian part at time \([n]\). We define the following conditions: A0 through A4.

A0 The Markovian part, \( Z^\infty = \{ Z^\infty[n] : n \in \mathbb{N}_0 \} \), is an irreducible Markov chain on a countable state space \((\mathcal{E}_Z^\infty, \mathcal{B}_Z^\infty)\) with associated kernel \( K_Z^\infty \) and associated stationary probability measure \( \pi_Z^\infty \).
Periodicities of \( Z \) do not matter because we can introduce a null transition with probability 1/2 without changing the steady state or the hitting distribution of \( W^\infty \). The mean hitting time would be doubled.

A1 The increments \( \Delta V^\infty [n] \) may of course be negative or zero, and we assume a negative average increment,

\[
\mu := \sum_{z \in \mathcal{E}_Z} \pi_Z^\infty (z) E_{(0, z)} \left[ \Delta V^\infty [1] \right] = \sum_{z \in \mathcal{E}_Z} \pi_Z^\infty (z) \sum_{\delta \in \mathbb{Z}} \delta \sum_{z' \in \mathcal{E}_Z} H^\infty \left( z, (\delta, z') \right) < 0. \tag{2.2}
\]

A2 We also assume \( W^\infty \) is aperiodic; that is \( P_{(0, z)} [S^\infty_1 \in \cdot] \) is an aperiodic distribution where \( z \) is some state in \( \mathcal{E}_Z^\infty \) and

\[
S^\infty_1 := V^\infty \left[ T^\infty_1 (z) \right] \tag{2.3}
\]

where \( T^\infty_n (z) \) is the \( n^{th} \) return time to \( z \).

A3 The additive increment between returns to some Markovian state \( z \) has some probability of being positive; that is \( P_{(0, z)} (S^\infty_1 > 0) > 0 \). It follows that the increment has this property for all initial states. This is so because any return to a state \( z' \) has positive probability of passing through \( z \) and therefore of passing through \( z \) an arbitrary number of times. The loops through \( z \) may produce positive increments which can make the total increment positive.

A4 The generating function associated to the Markov additive chain \( W^\infty \) exists for \( \gamma \) in some neighbourhood of 0 and is given by

\[
\hat{K}^\infty_{\gamma} (z, z') := E \left[ \exp (\gamma \Delta V^\infty [1]) I \{ Z^\infty [1] = z' \} \bigg| Z^\infty [0] = z \right] = \sum_{\delta \in \mathbb{Z}} e^{\gamma \delta} H^\infty \left( z, (\delta, z') \right). \tag{2.4}
\]

Since the increments \( \Delta V^\infty [n] \) are integer-valued variables, we can define a modified Markov modulated queue \( W := (Q, Z) \) on the measurable space \((\mathcal{E}, \mathcal{B}) := (\mathcal{E}_Q \times \mathcal{E}_Z, \mathcal{B}_Q \otimes \mathcal{B}_Z)\) where \( \mathcal{E}_Q \) and \( \mathcal{B}_Q \) denote respectively the positive integers (\( \mathbb{N}_0 \)) and the associated maximal \( \sigma \)-algebra (\( \mathcal{B}(\mathbb{N}_0) \)) and where in fact \( (\mathcal{E}_Z, \mathcal{B}_Z) = (\mathcal{E}_Z^\infty, \mathcal{B}_Z^\infty) \). For all \( q, q' \in \mathbb{N}_0 \), we define the transition kernel of the new chain \( W \) by

\[
K \left( (q, z), (q', z') \right) \equiv P \left[ Z^\infty [1] = z', (q + \Delta V^\infty [1])^+ = q' \bigg| Z^\infty [0] = z \right] = \begin{cases} H^\infty \left( z, (q' - q, z') \right) & \text{if } q' \neq 0 \\ \sum_{i \in \mathbb{N}_0} H^\infty \left( z, (-i - q, z') \right) & \text{if } q' = 0 \end{cases}. \]

Thus, \( Q[n] \) represents the queue size at time \( [n] \), and \( \Delta Q[n] = Q[n] - Q[n - 1] \) represents the batch size of customers added to or subtracted from the queue during the interval \( (n - 1, n] \).

It is important to realise that by construction the two chains \( W \) and \( W^\infty \) agree for transitions from \( \mathcal{E} \) to \( \mathcal{E}^\infty \setminus \Delta \), where \( \Delta = \{(q, z) \in \mathcal{E}^\infty : q \leq 0\} \) represents the boundary region. We will often have to work with the edge of that boundary region, \( \Delta = \Delta \cap \mathcal{E} = \{(0, v) \in \mathcal{E}^\infty : v \in \mathcal{E}_Z^\infty \} \).
It is also important to remark that \( W \) is not necessarily irreducible. Trajectories from some points \((0, z)\) may drift to states with a positive additive component and never return. Nevertheless there is one large irreducible component. Define \( T_n(z) \) to be the \( n^{th} \) return time to \( z \) and \( S_1 := Q[T_1(z)] \) to be the associated additive increment of \( W \). If we start from a point \( (q, z) \) where \( q \) is sufficiently large it is clear the increment \( S_1 \) shares the properties of \( S_1^\infty \), i.e. assumption A3. Hence the increment \( S_1 \) has a positive probability of being positive for some state \( z \). Moreover, for any \( z \in \mathcal{E}_W^\infty \), we know that \( S_1 \) is an aperiodic increment with negative mean. It follows that the support of the points \( S_1 := Q[T_n(z)] \) includes all integers greater than some \( L(z) \). Since the Markovian component \( Z \) is irreducible it follows that all points \( \{ (q, z) ; q > L(z) \} \) belong to an irreducible component \( \mathcal{I} \) of \( W \). By Condition A1 the mean drift of the additive component is negative so there must be at least one state \( (0, z) \in \mathcal{I} \).

Our first lemma states that the queueing system \( W \) is stable (i.e. positive recurrent). In such case, we will denote the stationary distribution of \( W \) by \( \pi \). By construction \( \hat{K}_Z(z, z') \equiv K_Z^\infty (z, z') \) and thus \( \pi_Z(z) \equiv \pi_Z^\infty (z) \) where \( \pi_Z \) is the marginal distribution of \( \pi \) on \( \mathcal{E}_Z \equiv \mathcal{E}_W^\infty \). Note that \( \pi \) gives no weight to points outside \( \mathcal{I} \). The proof of Lemma 2.1 is found in Section 7.

**Lemma 2.1 (Stability)** If the chain \( W^\infty \) with Markovian part kernel \( K_Z^\infty \) and associated stationary probability measure \( \pi_Z^\infty \) satisfies Conditions A0-A3, then the chain \( W \) has a stationary distribution \( \pi \).

### 2.1 Definition of the Twisted Chain

We are interested in the rare event corresponding to the queue, \( Q \), reaching a large size \( \ell \). The approach developed in McDonald (1996) requires a non-negative harmonic function \( h \) for the kernel \( K^\infty \),

\[
\text{i.e. } \sum_{w \in \mathcal{E}^\infty} K^\infty (w, w') h(w') = h(w), \quad \text{of the form } \forall w \in \mathcal{E}^\infty, \ h(w) = \exp(\gamma q) \hat{a}(z),
\]

(2.5)

where \( \hat{a}(z) \) is the right-eigenfunction of the generating function \( \hat{K}^\infty \) associated with eigenvalue 1. This \( h \) is used to “twist” \( K^\infty \) into a new kernel, \( K^\infty \). We define the \( h \)-transformed or twisted kernel of \( K^\infty \) by

\[
K^\infty(w, w') := K^\infty(w, w') \frac{h(w')}{h(w)} = H^\infty \left( z, (q' - q, z') \right) \exp \left( \gamma (q' - q) \right) \frac{\hat{a}(z')}{\hat{a}(z)}.
\]

(2.6)

We denote a realization of the Markov chain associated with this kernel by \( \mathcal{W}^\infty \). We can also split the chain \( \mathcal{W}^\infty \) into its additive and Markovian parts \( \mathcal{W}^\infty = (\mathcal{V}^\infty, \mathcal{Z}^\infty) \). The Markov chain \( Z^\infty \) has state space \( \mathcal{E}_Z^\infty \), and again \( \mathcal{V}^\infty \) can be viewed as an additive component taking integer values in \( \mathcal{E}_V^\infty \equiv \mathbb{Z} \). As with the untwisted chain, we define the increment associated with the jump of \( Z^\infty \) (i.e. from \( Z^\infty[n-1] \) to \( Z^\infty[n] \)) at time \( n \in \mathbb{N} \) to be \( \Delta V^\infty[n] := V^\infty[n] - V^\infty[n - 1] \). Then, the kernel of \( \mathcal{W}^\infty \) represents

\[
P \left[ \Delta V^\infty[1] = q' - q, Z^\infty[1] = z' \mid Z^\infty[0] = z \right] =: H^\infty \left( z, (q' - q, z') \right) \equiv K^\infty (w, w').
\]

(2.7)

The Markovian part \( Z^\infty \) is again a Markov chain with transition kernel given by

\[
K^\infty(z, z') = \sum_{\delta \in \mathbb{Z}} H^\infty \left( z, (\delta, z') \right).
\]
In the two next subsections we construct $h$ and develop versions of the key Steady State and Mean Hitting Time Theorems of McDonald (1996) which link the asymptotic distributional properties of $Q$ to $h$.

### 2.2 Construction of $h$

The above results require a particular value of $\gamma > 0$, noted here by $\theta$, for which the “Feynman-Kac” operator $\hat{K}_\gamma^\infty$ of (2.4) has Perron-Frobenius eigenvalue equal to 1, and has positive valued eigenfunctions $\ell(z \mid \gamma)$ and $r(z \mid \gamma)$ such that

$$\sum_{z \in E_Z^\infty} \ell(z \mid \gamma) r(z \mid \gamma) = 1. \tag{2.8}$$

Ney and Nummelin (1987) studied a general form of the question of existence, and we follow their approach below. Before we start, we remark that Condition (M1) of Ney and Nummelin (1987) is vacuous for a discrete time chain on a countable state space like $E^\infty$.

**Definition 2.2** If $T_1(z)$ represents the first time $Z^\infty$ returns to the state $z$ in $E_Z^\infty$, then $\forall \gamma \in \mathbb{R}$, $\Lambda \in \mathbb{R}$ define

$$\Psi(\gamma, \Lambda) := E_{(0,z)} \left[ \exp (\gamma V^\infty[T_1(z)] - \Lambda T_1(z)) \right],$$

$$D := \left\{ (\gamma, \Lambda) \in \mathbb{R}^2 \mid \Psi(\gamma, \Lambda) < \infty \right\},$$

$$D_1 := \left\{ \gamma \in \mathbb{R} \mid \exists \Lambda \in \mathbb{R} : (\gamma, \Lambda) \in D \right\}.$$

Under Assumption A3, and for $\Lambda > 0$ fixed, $\lim_{\gamma \to \infty} \Psi(\gamma, \Lambda) = \infty$. If $\Psi(\gamma, \Lambda_0) = 1$ for some $\Lambda_0 < \infty$ then define $\Lambda(\gamma) := \Lambda_0$. When it exists, the function $\Lambda(\gamma)$ is well-defined because $\Lambda_0$ is unique, i.e. $\Psi(\gamma, \Lambda)$ is monotone in $\Lambda$. Further, under assumption A3, $\lim_{\gamma \to \infty} \Lambda(\gamma) = \infty$.

If we choose $\gamma \in D_1$ and if we assume $D$, the domain of convergence of $\Psi$, is open (necessarily containing $(0,0)$) then by Theorem 4.1 in Ney and Nummelin (1987), $\exp(\Lambda(\gamma))$ is the Perron-Frobenius eigenvalue of $\hat{K}_\gamma^\infty$ with associated nonnegative eigenvector pair $\ell(z \mid \gamma)$ and $r(z \mid \gamma)$ (where $\sum_{z} \ell(z \mid \gamma) r(z \mid \gamma) = 1$ : Lemma 4.1 Ney-Nummelin (1987) ). Under the same assumptions, by Theorem 5.1 in Ney and Nummelin (1987), we have that for any state $z \in E_Z^\infty$,

$$\lim_{n \to \infty} \frac{1}{n} \log \left( E_{(0,z)} \left[ \exp (\gamma V^\infty[n]) \right] \right) = \Lambda(\gamma).$$

In Section 3 we will show that, under the assumptions A1, A3 plus $D$ open, the strict convexity of $\Lambda(\gamma)$ (Corollary 3.3 of Ney and Nummelin (1987)) implies there exists a unique $\theta > 0$ such that $\exp(\Lambda(\theta)) = 1$. For this $\theta$ there exist $r(z \mid \theta)$, harmonic $h(w) := \exp (\theta q) \hat{a}(z) = \exp (\theta q) r(z \mid \theta)$ and stationary distribution $\varphi(z) = \ell(z \mid \theta) r(z \mid \theta)$. For practical applications it is often difficult to find the eigenvectors and eigenvalues of $\hat{K}_\gamma^\infty$. In Section 3 we will consider fast teller systems and we will guess a suitable harmonic function.
2.3 Steady State and Mean Hitting Time Theorems

We now establish sufficient conditions for Conditions (1-7) in McDonald (1996). These will lead to the Steady State and Mean Hitting Time Theorems.

Lemma 2.3 Assume $h(w) = \exp(\theta q)\tilde{a}(z)$ is a positive harmonic function for $K^{\infty}$. Define $\lambda_{\Delta}(z) := \pi(0, z)\tilde{a}(z)$. Conditions (1-7) in McDonald (1996) hold for $(K^{\infty}, K^{\infty})$ under Conditions A0, A1, A2 and the following five additional conditions.

$Z^{\infty}$ has a stationary probability measure $\varphi$. \hfill (2.9)

$$d := E_{\varphi}\Delta V^{\infty}[1] > 0 \hfill (2.10)$$

$$\sum_{z \in E_{\Delta}^{\infty}} \left(\tilde{a}(z)\right)^{-1} \varphi(z) < \infty. \hfill (2.11)$$

$$\sum_{z \in E_{\Delta}^{\infty}} \lambda_{\Delta}(z) = \sum_{z \in E_{Z}^{\infty}} \pi(0, z)\tilde{a}(z) < \infty. \hfill (2.12)$$

$\lambda_{\Delta}(z)$ is $\left(\tilde{a}(z)\right)^{-1}$—regular for the chain $Z^{\infty}$ in the sense of Meyn and Tweedie (1993). \hfill (2.13)

Remark: Condition (2.12) clearly holds if $\sum_{z} \pi Z^{\infty}(z)\tilde{a}(z) < \infty$ since $\pi Z^{\infty} \equiv \pi Z$ is the marginal distribution of $\pi$ on $E_{Z}$. The above sufficient conditions are further examined for the fast teller queue in Sections 3.2 and 4.

Now we can restate two theorems from McDonald (1998). Let

$$f := \sum_{w \in \Delta} \pi(w)\tilde{a}(z) H(w), \hfill (2.14)$$

where for all $w = (0, z) \in \Delta = \blacktriangle \cap E$, $H(w)$ represents the probability that $W^{\infty}$ does not return to the boundary region $\blacktriangle$ starting from $w$. The constant $f$ is generally unknown, but can be obtained by simulating the queue (see section 2.4).

Theorem 2.4 (Steady State) Under conditions of Lemma 2.3

$$\pi((\ell, z)) \sim \exp(-\theta \ell) \left(\tilde{a}(z)\right)^{-1} \varphi(z) \frac{f}{d}$$

where $d$ is given in line (7.38), and $f$ by (2.14).
This theorem provides a complete description of the steady state distribution of the system as the queue size becomes large.

**Theorem 2.5 (Mean Hitting Time)** Assume the conditions of Lemma 2.3. Let $\tau_\ell$ denote the first time the queue $Q$ reaches or exceeds the level $\ell$ and let $\mu(k, z)$ denote the steady state of the ladder height process $(V^\infty[\tau^\infty_\ell] - \ell, Z^\infty[\tau^\infty_\ell])$ where $\tau^\infty_\ell$ is the first time $V^\infty$ reaches or exceeds $\ell$ starting from an empty queue. Then, for any initial state $z$ of the modulating Markov chain $Z$,

$$E_{(0,z)}[\tau_\ell] \sim \exp(\theta\ell)g^{-1}$$

where $g = f \cdot \sum_{z \in E_Z} \sum_{k=0}^{\infty} (\hat{a}(z))^{-1} \exp(-\theta k)\mu(k, z)$

and where $f$ is given in line (2.14).

Since the twisted chain $W^\infty$ is transient to infinity, $g$ may be approximated by a fast simulation. Modulo this constant we see the mean time for the queue to reach a high level $\ell$ increases exponentially in $\ell$.

**2.4 Simulating large deviations of $Q$.**

Usually it is impractical to simulate events such as $Q$ crossing a high level directly since these events are so rare that a direct Monte Carlo simulation would take an inordinate amount of time to complete. We now discuss a general method for accelerating these simulations and how this is related to our choice of $K^\infty$.

The estimation of $\pi(A)$, the steady state probability of a rare event $A$, is based on cycles of the original Markov additive chain $W = (Q, Z)$, where cycles are defined by successive returns to $\Delta$ by $W$. We can represent the steady state probability of a rare event $A$ as in Heidelberger (1995),

$$\pi(A) = (E_{\pi|\Delta}[\tau_\Delta])^{-1} E_{\pi|\Delta} \left[ \sum_{t=1}^{\tau_\Delta} I\{W[t] \in A\} \right] \quad (2.15)$$

where $\tau_\Delta$ is the return time to $\Delta$ and $\pi|\Delta$ denotes the steady state conditioned on starting in $\Delta$. If we denote the $i$ th cycle of the chain by $W^i$ and the duration of the $i$ th cycle by $\tau^i_\Delta$ then the above probability is estimated by

$$\hat{\pi}(A) = \left( \sum_{i=1}^{n} \tau^i_\Delta \right)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{\tau^i_\Delta} I\{W^i[t] \in A\}. \quad (2.16)$$

Since in our case the original process is recurrent to $\Delta$, standard Monte Carlo simulation is quite effective for the denominator. Successive returns of the process to $\Delta$ simulate the equilibrium measure restricted to $\Delta$ so, if $n$ is the number of cycles generated, $\sum_{i=1}^{n} \tau^i_\Delta / n \to E_{\pi|\Delta}[\tau_\Delta]$. Hence $\sum_{i=1}^{n} \tau^i_\Delta / n$ is a consistent estimator of the numerator of (2.15).
Next, dividing the numerator of (2.16) by the number of cycles gives a consistent estimator of the numerator of (2.15):

\[ \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{\tau_i} I \{ W^i[t] \in A \} \rightarrow E_{\pi|\Delta} \left[ \sum_{t=1}^{\tau_M} I \{ W[t] \in A \} \right]. \]

The naive simulation of the numerator of (2.15) consists of picking starting states according to \( \pi \| \Delta \) obtained by generating cycles \( W_i \), and then generating a trajectory according to the kernel, \( K \), of the given chain until it again hits \( \Delta \) (a cycle). If the trajectory hits \( A \), we count 1 for each \( t \) the chain remains in \( A \).

The alternative to direct simulation is to use a change of measure on the trajectories of the Markov chain in a manner analogous to importance sampling in Monte Carlo simulation (see Heidelberger (1995) for a review of the literature). A trajectory is simulated for a Markov chain for which the event in question is not rare but common. This alternative simulation consists of picking starting states according to \( \pi \| M \) as before but then generating a trajectory using a different twisted kernel, \( K' \), until the chain hits \( F := \{ (q, z) : q \geq \ell \} \supset A \), and then using the original \( K \) from that point until the end of the cycle. If the chain hits \( \Delta \) before hitting \( F \), the simulation ends without the switch back to \( K \). As before, we count 1 for each \( t \) spent by this composite chain in \( A \), but each cycle will receive a weighting equal to the relative likelihood of the path using \( K \) compared to \( K' \). If \( K' \) is chosen so that \( Q \) drifts to \(+\infty\), then this composite simulation should quickly move to \( F \), and then quickly return to \( \Delta \).

For each transition from \((q, z)\) to \((q', z')\), we can certainly compute the ratio

\[ \frac{K\left((q, z), (q', z')\right)}{K'\left((q, z), (q', z')\right)} \]

and then multiply these factors together over the portion of the trajectory simulated under the changed measure, and so obtain the weighting to be applied to that trajectory. Aside from the numerical problems associated with products of very small factors, there remains the probabilistic problem that the random variables being simulated (one per cycle) have highly skewed distributions with the vast majority of their mass placed at 0.

These difficulties are circumvented by choosing \( K' \) in such a way that the multiplicative factors collapse into a simple ratio of a function, \( h \), computed at \((q_0, z_0)\) at the start of the twisted portion (when the state belongs to \( \Delta \)) divided by that function computed at \((q, z)\) at the end of the twisted portion (when the trajectory first hits \( F \)). In between these end times the trajectory is inside \( E^\infty \setminus \Delta \) where \( K \) agrees with \( K^\infty \) so the selection of \( K' = K^\infty \) achieves these goals. Note that we only simulate the twisted chain outside \( \Delta \). This avoids the complications noted by Glasserman and Kou (1995).

In addition to the reduction in computational effort, there is an additional probabilistic benefit. Generally \( h(q, z) \) is of the form \( \exp(\theta q) \bar{a}(z) \). For \( A \subset F = \{ (q, z) \in E^\infty : q \geq \ell, z \in E_z^\infty \} \), \( \exp(\theta (q_0 - q)) \) will contain a fixed exponential \( \exp(-\theta \ell) \) and can be pulled out of the numerator. The ratio \( \bar{a}(z_0) / \bar{a}(z) \) now generally is not heavily skewed, and usually exhibits nice limiting distributional properties as \( \ell \to \infty \).

Finally, it is of practical importance that the twisted chain is like the original chain except that some of the rates are altered. This means that a simulation written for the untwisted system will serve for the twisted system simply by altering the rates!
2.5 Literature Review

Consider a random walk \( V^∞[n] := V^∞[n] + ΔV^∞[n] \) made from i.i.d. increments \( ΔV^∞[n] \) with negative mean. High level crossing probabilities for such a random walk have been studied extensively. The approach known as the twisting technique is described in Section XII.4(b) in Feller (1971), where the twisted random walk is called the associated random walk. In particular the twisting technique gives Cramer’s estimate for the probabilities of ruin as discussed in Section XII.5(d) in Feller (1971). Kesten (1974) Section 4 studied high level crossing probabilities of a Markov additive chain using the twist. Ney and Nummelin (1987) used the twist to study large deviations of the additive component of a Markov additive chain. This idea was also used by Siegmund (1976) and Asmussen (1985) to accelerate simulations of ruin probabilities using importance sampling. In all these applications the optimal twist is an \( h \)-transformation.

Now consider a queueing variable \( Q[n] := (Q[n − 1] + ΔV^∞[n])^+, \) \( Q[0] = 0 \) with an underlying random walk \( V^∞[n] \): \( V^∞ \) is called the free process since the boundary at 0 has been removed from \( Q \). The steady state \( π \) of \( Q \) has the same distribution as the maximum of an underlying random walk (see VI.9 Theorem 24 in Feller Volume II), i.e. \( P_π(Q[n] > q) = P_0(\max V^∞[k] > q) \). This representation has been extended by Loynes (1962) using time reversal to the case where the underlying random walk is a Markov modulated. Asmussen (1995) has another extension for \textit{stochastically monotone} Markov chains.

The asymptotics of the tail probability of the queueing variable may now be obtained from the above representation. Since the equilibrium probabilities are expressed as ruin probabilities we can obtain the asymptotics by twisting the underlying random walk. In Section XII.5(b), Feller used the associated random walk to obtain the asymptotics of the tail probability of the stationary distribution of the waiting time in a \( GI/G/1 \) queue (also see Asmussen (1987) Chapter III.7). The extensions by Loynes and Asmussen may be handled in the same way (see Asmussen (1995)).

Another representation of \( P_π(Q[n] > q) \) is given by the expected number of times \( Q \) or \( V^∞ \) exceeds \( q \) before returning to \( (−∞, 0] \) divided by the mean time for \( V \) to return to \( (−∞, 0] \). This cycle representation for general Markov chains is found in Orey (1971). The cycle representation was used in the seminal paper by Parekh and Walrand (1989) and subsequently in Nicola et al (1992) in a simulation context. Sadowsky and Szpankowski (1995) masterfully used this representation for calculating the exact asymptotics of a fast teller system with an i.i.d stream of batch arrivals and \( c \) independent servers which do not necessarily have identically distributed service times. This representation is also the basis of McDonald (1998) which gives the asymptotics of level crossing probabilities for Markov chains which can be modeled as a Markov additive process with a boundary. In all these studies the key remark is that \( Q = V^∞ \) during a cycle so the twisting technique above can be used to calculate level crossing probabilities of the free process.

The fast teller system of Sadowsky and Szpankowski (1995) is studied here as a special case since we allow Markov modulated interarrival times and service times. Sadowsky and Szpankowski only treat the case where all interarrival and service times have \textit{spread out} distributions while we only treat the case where these times are integer valued in order to include ATM queues.

Sadowsky and Szpankowski (1995) studied both the workload and the queue length of the fast teller system while we only studied the queue length. Both processes can be seen as Markov additive processes with a boundary (so we also could have studied the workload process). Both papers develop the same twist although we feel searching for a harmonic function is conceptually easier. The main result in
Sadowsky and Szpankowski (1995) is their Theorem 2.3. The tail of the steady state distribution of the queue just after a block arrival, $Q^+$, is shown to be die off exponentially, i.e. $P_π(Q^+ > q) \sim c_Q \omega^q$. The tail of the waiting time distribution is also given as are the tails of the distributions of the cycle maxima of the waiting times and the queues just after a block arrival.

The above result should be compared with Theorem 4.2 here which gives a result of the form $P_π(Q > q) \sim c_0 \exp(-\theta q)$. We study the tail of the steady state distribution of the queue at an arbitrary time so in fact our results are complementary. We could have modified our technique to arrive at a result like the above just as Sadowsky and Szpankowski could have studied the steady state at an arbitrary time (see the remark following Lemma 3.6 there). Either way the exponential rate is the same; i.e. $\omega = \exp(-\theta)$. The big difference is in the constant. Our constant is based on Theorem 1.6 in McDonald (1998) while the constant $c_Q$ is given by (39) in Sadowsky and Szpankowski (1995) divided by the mean cycle time. This is what one would get applying Theorem 1.5 in McDonald (1998). The expression in Theorem 4.2 is much more explicit. The constant is almost determined except for $f$ which has a probabilistic interpretation that invites a close approximation. One should also remark that we give the mean time for the queue to hit a high level (there is no such result in Sadowsky and Szpankowski (1995)).

In parallel with the work on generalizations of the $GI/G/1$ queue, much work has been done on the steady state of the queue of an ATM multiplexor. A bound on the probability of a large queue is essential for the design of ATM switches. The seminal paper by Anick, Mitra and Sondhi (1982) gave an eigenfunction expansion of the steady state of the queue. Unfortunately such an expansion is not practical when dealing with many Markov modulated sources. Instead one focuses on the term with the smallest eigenvalue; that is, the term of the form $c_0 \exp(-\theta \ell)$ where $\ell$ is the buffer size. Duffield and O’Connell (1995) specifically identified the harmonic function associated with the model of a buffer in an ATM multiplexor driven by Markov modulated sources. They used the harmonic function to obtain the Perron-Frobenius eigenvalue which gave bounds on the asymptotic behaviour of the tail probabilities of the queue in the buffer. The exact asymptotics in Theorem 2.4 show the harmonic function obtained by Duffield and O’Connell (1995) (a special case of the harmonic functions derived here) determines the asymptotics of the overflow probabilities to be $c_0 \exp(-\theta \ell)$.

Rare event simulation is also closely related to the twist (see Heidelberger (1995) for a review of the literature). The twisted transition kernel can be used to accelerate the simulation of large deviations of a Markov chain. For example, Bonneau (1996) used the $h$-transform explicitly to simulate a leaky-bucket controller. Lamothé (1998) used the $h$-transform to derive overflow probabilities for a queue buried deep in an ATM switch.

### 3 The Fast Teller Queue.

Here we obtain the Markov additive representation of a fast teller queue. The source and server processes will be used to build the boundary-free process $(V^\infty, Z^\infty)$. The introduction of the boundary will define $(Q, Z)$. In this setting, this section establishes the existence of $\ell$, $r$, and $\Lambda$. 
3.1 Definitions.

We regard \( \{ \Delta V^a[k] = (A^a_k, B^a_k) : k \in \mathbb{N} \} \) as the increments of the additive components of a Markov additive chain with underlying Markovian components \( \{ M^a_k : k \in \mathbb{N} \} \). The corresponding increments of the additive components for the servers are \( \Delta U^s[k] \).

Let
\[
\pi(s) \equiv \pi^s \equiv \pi^s_{\infty} \equiv \pi^s_{\infty}(s) = E[\pi(s) | IIN] \equiv \pi^s = \pi^s_{\infty} \equiv \pi^s_{\infty}(s)
\]
be the stationary probability measure of the kernel \( K^a_M(m_0, m_1) \).

Remark: To simplify the notation, we write \( k_a \) for \( k_a^a \) and \( k_s \) for \( k_s^a \). For simplicity we assume that free servers each take customers from the queue in an order specified by the index of the server; i.e. if servers \( j \) and \( j' \) are free and \( j < j' \) then the customers go to server \( j \).

We model the Markovian component of the fast teller system by the vector

\[
Z[t] = Z^\infty[t] := \left( (M[t], \mathbf{X}[t]) , (N[t], \mathbf{Y}[t]) : \mathbb{N} \rightarrow \prod_{a=1}^{A} (E^a_M \times \mathbb{N}_0) \times \prod_{s=1}^{S} (E^a_N \times \mathbb{N}_0) \right)
\]

where
\[
\mathbf{X}[t] \equiv \left( X^a[t] \right)_{a \in \{1...A\}} = \left( A^a_{1,k_a[t]} - t \right)_{a \in \{1...A\}} \quad \text{and} \quad \mathbf{Y}[t] \equiv \left( Y^s[t] \right)_{s \in \{1...S\}} = \left( S^s_{1,k_s[t]} - t \right)_{s \in \{1...S\}}
\]
represent, respectively, the residual number of time slots until the next arrival for each source and the residual service time for each server at the end of the \( t \)th time slot while
\[
M[t] \equiv \left( M^a[t] \right)_{a \in \{1...A\}} = \left( M^a_{k_a[t]} \right)_{a \in \{1...A\}} \quad \text{and} \quad \mathbf{N}[t] \equiv \left( N^s[t] \right)_{s \in \{1...S\}} = \left( N^s_{k_s[t]} \right)_{s \in \{1...S\}}
\]
indicate the state of the modulating chains at time \( [t] \). In the notation of Section 2, \( \mathcal{E}_Z \equiv \mathcal{E}_Z^\infty = \prod_a (E^a_M \times \mathbb{N}_0) \times \prod_s (E^a_N \times \mathbb{N}_0) \).

3.2 Conditions.

3.2.1 Condition A0.

Stationary measure. Let \( \pi^a_M \) denote the stationary probability measure of the kernel \( K^a_M(m_0, m_1) \). Let \( F^a_A(x|m) \) and \( \mu^a_M \) denote the cumulative distribution and the mean associated with the density \( f^a_A(x|m) \) of the sojourn time in state \( m \) by source \( a \). Let the associated mean block size added during that sojourn time in state \( m \) be \( \mu^a_B(m) \). Further assume \( \mu^a_A := \sum_m \pi^a_M(m) \mu^a_A(m) < \infty \) and \( \mu^a_B := \sum_m \pi^a_M(m) \mu^a_B(m) < \infty \). Define

\[
\pi^\infty_a \left( (m, x) \right) := \pi^a_M(m) \frac{1 - F^a_A(x|m)}{\mu^a_a}.
\]
The corresponding quantities for the servers are $\pi^*_s$, $K^*_s(n_0,n_1)$, $F^*_s(y|n)$, $\mu^*_s(n)$, $f_s^*(y|n)$, $\mu^*_R(n)$, $\mu^*_S$, and $\pi^*_s((n,y))$. Since the sources and servers are independent, $Z = Z^\infty$ is a Markov chain with stationary distribution (i.e (2.9)) given by

$$\pi^*_Z(z) := \prod_{a=1}^A \pi^*_a \left( (m_a, x_a) : \prod_{s=1}^S \pi^*_s \left( (n_s, y_s) \right) \right).$$

**Irreducibility.** For the theory of Section 2 to apply, the chain $Z^\infty = ((M, X), (N, Y))$ must be irreducible. If, for instance, two arrival streams $a$ and $a'$ have the same deterministic inter-arrival times (or one deterministic inter-arrival time is a multiple of the other) then $(X^a[t], X^{a'}[t])$ is not irreducible in the state space $\mathbb{N}_0 \times \mathbb{N}_0$. To apply the theory we would have to reduce the state space to the orbit of $(X^a[t], X^{a'}[t])$ determined by the starting point. We assume the state space has been reduced in this way to make $Z^\infty$ irreducible. The same phenomenon could occur for combinations of deterministic inter-service and inter-arrival times or/and for combinations of deterministic modulating Markovian paths. In this case $\pi^*_Z$ becomes

$$\pi^*_Z((\bar{m}, \bar{x}, \bar{n}, \bar{y})) := C \ I_O((\bar{m}, \bar{x}, \bar{n}, \bar{y})) \prod_{a=1}^A \pi^*_a \left( (m_a, x_a) : \prod_{s=1}^S \pi^*_s \left( (n_s, y_s) \right) \right)$$

\[ \text{(3.17)} \]

where $I_O((\bar{m}, \bar{x}, \bar{n}, \bar{y}))$ is an indicator of the orbit of $Z^\infty[0] \in \prod_{a=1}^A (\mathcal{E}_M^a \times \mathbb{N}_0) \times \prod_{s=1}^S (\mathcal{E}_N^s \times \mathbb{N}_0)$. The constant $C$ ensures $\pi^*_Z$ is still a probability measure.

### 3.2.2 Condition A1: Stability.

Here we look at Lemma 2.1 in the fast teller case. Let $\mathcal{I}_a(w) \equiv \{ a \in \{1 \ldots A \} : x_a = 0 \}$ denote the set of indices of those sources who will deliver a new batch by the end of the $(t - 1)^{th}$ time slot. Let $\mathcal{I}_s(w) \equiv \{ s \in \{1 \ldots S \} : y_s = 0 \}$ denote the set of indices of those servers which are free at the end of the $(t - 1)^{th}$ time slot. If the residual service time is zero at the end of a time slot then the server is free to start a new service at the start of the next time slot.

\[ \mu := E_{\pi^*_Z} \Delta V^\infty[1] = \sum_{z \in \mathcal{E}_Z^\infty} \pi^*_Z(z) E_{(0,z)} \left[ - \sum_{s=1}^S R^s_{[1]} I \{ y_s = 0 \} + \sum_{a=1}^A B^a_{[1]} I \{ x_a = 0 \} \right] \]

\[ = \sum_{z \in \mathcal{E}_Z^\infty} \pi^*_Z(z) E_{(0,z)} \left[ - \sum_{s=1}^S R^s_{[1]} I \{ y_s = 0 \} + \sum_{a=1}^A B^a_{[1]} I \{ x_a = 0 \} \right] \]

\[ = -\sum_{s=1}^S \sum_{n_s \in S^N} \pi^*_s(\{ n_s, 0 \}) E_{(n_s,0)} \left[ R^s_{[1]} \right] + \sum_{a=1}^A \sum_{m_a \in S^M_R} \pi^*_a(\{ m_a, 0 \}) E_{(m_a,0)} \left[ B^a_{[1]} \right] \]

\[ = -\sum_{s=1}^S \frac{1}{\mu^*_S} \sum_{n_s \in S^N} \pi^*_s(\{ n_s \}) E_{(n_s,0)} \left[ R^s_{[1]} \right] + \sum_{a=1}^A \frac{1}{\mu^*_A} \sum_{m_a \in S^M} \pi^*_a(\{ m_a \}) E_{(m_a,0)} \left[ B^a_{[1]} \right]. \]
For (2.2) we require that the average service rate exceeds the average arrival rate, i.e. A1:

$$\sum_{a=1}^{A} \left( E_{\pi_\infty} \left[ B_{[1]}^{a} \right] \left( E_{\pi_\infty} \left[ A_{[1]}^{a} \right] \right)^{-1} \right) - \sum_{s=1}^{S} \left( E_{\pi_\infty} \left[ R_{[1]}^{s} \right] \left( E_{\pi_\infty} \left[ S_{[1]}^{s} \right] \right)^{-1} \right) < 0. \quad (3.18)$$

We have assumed $E_{\pi_\infty} \left[ B_{[1]}^{a} \right]$ and $E_{\pi_\infty} \left[ R_{[1]}^{s} \right]$ are finite. Condition (3.18) is sufficient for stability of $W$ by Lemma 2.1. As before we denote the steady state of $K$ by $\pi$.

### 3.2.3 Condition A2: Sufficient conditions for aperiodicity.

It is sufficient that there exist a source $a$ having states $m_a$ and $m'_a$ and $\alpha, \beta \in \mathbb{N}_0$ such that

$$\min\{P_{AB}^{a} \left[ \alpha, \beta \left| m_a, m'_a \right. \right], P_{AB}^{a} \left[ \alpha, \beta + 1 \left| m_a, m'_a \right. \right] \} > 0$$

or that there exist a server $s$ having states $n_s$ and $n'_s$ and $\sigma, \rho \in \mathbb{N}_0$ such that

$$\min\{P_{SR}^{s} \left[ \sigma, \rho \left| n_s, n'_s \right. \right], P_{SR}^{s} \left[ \sigma, \rho + 1 \left| n_s, n'_s \right. \right] \} > 0.$$  

If either of these conditions hold the increment after a return to some state $z$ has a nonzero probability of taking consecutive integer values and this implies the increment is aperiodic.

If the blocks are of size one then it is sufficient that

$$\min\{P_{AB}^{a} \left[ \alpha, 1 \left| m_a, m_a \right. \right], P_{AB}^{a} \left[ \alpha', 1 \left| m_a, m_a \right. \right], P_{AB}^{a} \left[ \alpha + \alpha', 1 \left| m_a, m_a \right. \right] \} > 0.$$  

If this condition holds, the increment after a return to some state $z$ has a nonzero probability of taking consecutive integer values because the interarrival time of one customer $\alpha + \alpha'$ may be broken into the arrival times $\alpha$ and $\alpha'$ of two customers. A similar condition for one server would also be sufficient.

### 3.2.4 A3: a counter-example.

It is possible for a random system to fail A3. Consider one server with a service time of exactly 5 time slots. The server serves blocks of 5 customers, if available. Suppose the interarrival time for a single customer is geometric with a mean of 1 time slot. Clearly the queue size is bounded by 5. There is no asymptotic theory because A3 fails.

### 3.2.5 The kernel $K_\infty$ and Condition A4.

Recall the notation of Section 1.2, and define

$$\forall \; u \in \mathbb{R} \; \forall \; v \in \mathbb{R} \; \phi_{AB}^{a} (u, v \mid m_a, m'_a) = \sum_{\alpha=1}^{\infty} \sum_{\beta=0}^{\infty} e^{(u\alpha+v\beta)} P_{AB}^{a} \left[ \alpha, \beta \left| m_a, m'_a \right. \right] \quad (3.19)$$
and \( \phi_{SR}^s(u, v \mid n_s, n'_s) \) similarly. For \( a \in \{1 \ldots A\} \) and \( m_a \in S_M^a \) the generating function of the \( a \)th arrival process is

\[
\hat{K}_M^a(m_a, m'_a \mid \alpha^a, \gamma^a) = E \left[ \exp \left( \alpha^a A^a_I + \gamma^a B^a_{II} \right) I \{ M^a_I = m'_a \} \mid M^a_0 = m_a \right] = \phi_{AB}^a(\alpha^a, \gamma^a) m_a, m'_a \right) \right) \hat{K}_M^a(m_a, m'_a).
\]

For all \( \zeta_A^a = (\alpha^a, \gamma^a) \in \mathbb{R}^2 \) and \( \Lambda_A^a \in \mathbb{R} \), we may define

\[
\Psi_A^a(\zeta_A^a, \Lambda_A^a) := E_{m_a} \left[ \exp (\zeta_A^a \cdot V^a[T^a_I(m_a)] - \Lambda_A^a T^a_I(m_a)) \right]
\]

where \( T^a_I(m_a) \) represents the first time \( M^a \) returns to the state \( m_a \) in \( S_M^a \). If \( \Psi_A^a(\zeta_A^a, \Lambda_A^a) = 1 \) for some \( \Lambda_A^a < - \infty \) then define \( \Lambda_A^a(\zeta_A^a) := \Lambda_A^a \). Because \( \Psi_A^a(\zeta_A^a, \Lambda_A^a) \) is monotone in \( \Lambda_A^a \), \( \Lambda_A^a \) is unique, and so when it exists, the function \( \Lambda_A^a(\zeta_A^a) \) is well-defined. Let \( \exp \Lambda_A^a(\zeta_A^a) \) be the Perron-Frobenius eigenvalue of \( \hat{K}_M^a(m_a, m'_a \mid \zeta_A^a) \) with associated nonnegative eigenvector pair \( \ell(m_a \mid \zeta_A^a) \) and \( r(m_a \mid \zeta_A^a) \) (where \( \sum_{m_a} \ell(m_a \mid \zeta_A^a) r(m_a \mid \zeta_A^a) = 1 \)). Similarly for \( s \in \{1 \ldots S\} \), \( \Delta U^s[k] := (S^s_k, R^s_k) \) and \( n_s \in S_N^s \), the left and right Perron-Frobenius eigenvectors for

\[
\hat{K}_N^s(n_s, n'_s \mid \beta^s, \gamma^s) = E \left[ \exp \left( \beta^s S^s_I + \gamma^s R^s_{II} \right) I \{ N^s_I = n'_s \} \mid N^s_0 = n_s \right]
\]

are \( \ell_S^s(\beta^s, \gamma^s) \) and \( r_S^s(\beta^s, \gamma^s) \). Again these are associated with the eigenvalue given by \( \exp \Lambda_S^s(\beta^s, \gamma^s) \).

The boundary-free system: \( (V^\infty, Z^\infty) \) We now define the free fast teller system. Let \( \mathcal{E}^\infty = (\mathcal{E}_V^\infty, \mathcal{E}_Z^\infty) = (\mathcal{Z}, \mathcal{E}_Z^\infty) \). For this section, let

\[
w = \left( q, \left( (\bar{m}, \bar{x}), (\bar{n}, \bar{y}) \right) \right) = \left( q, \left( (m_1, \ldots, m_A), (x_1, \ldots, x_A) \right), \left( (n_1, \ldots, n_S), (y_1, \ldots, y_S) \right) \right) \in \mathcal{E}^\infty
\]

and \( w' \in \mathcal{E}^\infty \) be respectively the state of \( W^\infty \) at the end of the \( (t-1) \)th time slot and at the end of the \( t \)th time slot.

To define the transition kernel \( K^\infty \) consider a transition from \( w \) to \( w' \). Necessary conditions for this transition to occur \( (K^\infty(w, w') > 0) \) are

\[
\forall a \in \mathcal{I}_A^c(w) := \{1 \ldots A\} \setminus \mathcal{I}_A(w), \quad (m'_a, x'_a) = (m_a, x_a - 1)
\]

and

\[
\forall s \in \mathcal{I}_S^c(w) := \{1 \ldots S\} \setminus \mathcal{I}_S(w), \quad (n'_s, y'_s) = (n_s, y_s - 1).
\]

We will write \( \mathcal{I}_A := \mathcal{I}_A(w), \mathcal{I}_A^c := \mathcal{I}_A^c(w), \mathcal{I}_S := \mathcal{I}_S(w) \) and \( \mathcal{I}_S^c := \mathcal{I}_S^c(w) \). If \( \land \) represents the logical conjunction, let

\[
\Xi(w, w') := \bigwedge_{a \in \mathcal{I}_A^c} \{ x'_a + 1 = x_a \} \land \{ m'_a = m_a \} \bigwedge_{s \in \mathcal{I}_S^c} \{ y'_s + 1 = y_s \} \land \{ n'_s = n_s \}
\]
and if \( b_a \) represent the number of cells (customers) arriving from source \( a \), and \( r_s \) the number of cells (customers) served by the server \( s \), let

\[
\Upsilon\left(w, w'\right) := \left\{ \left(\vec{b}, \vec{r}\right) \in (\mathbb{N}_0)^{A+S} : \forall a \in I_{n_A}^c, b_a = 0 ; \forall s \in I_{n_S}^c, r_s = 0 ; q' = q - \sum_{s=1}^{S} r_s + \sum_{a=1}^{A} b_a \right\}
\]

The probability of the transition of \((V^\infty, Z^\infty)\) from \( w \) to \( w' \) is then given by:

\[
K^\infty\left(w, w'\right) = \mathbb{E}\left\{ \Xi\left(w, w'\right) \right\} \sum_{(\vec{b}, \vec{r}) \in \Upsilon(w, w')} \left( \prod_{a \in I_A} \left( K^a_M\left(m_a, m'_a\right) P_{AB}^{a} \left[ x'_a + 1, b_a \mid m_a, m'_a \right] \right) \right) \times \left( \prod_{s \in I_S} \left( K^s_S\left(n_s, n'_s\right) P_{SR}^{s} \left[ y'_s + 1, r_s \mid n_s, n'_s \right] \right) \right)
\]

Now define the kernel \( K \) associated with the chain \( W = (Q, Z) \) by introducing the boundary \( \Lambda = \{ w \in \mathcal{E}^\infty : q \leq 0 \} \). Then, \( Q[t] \) represents the number of customers queued (but not being served) at the end of the \( t \)th time slot.

### 3.3 Harmonic Function for \( K^\infty \).

We will build up a harmonic function for \( K^\infty \) (see Section 2.2) from the transforms of the individual batch arrival and batch service processes. The next theorem gives sufficient conditions for \( h(w) = \exp(\theta q) \hat{a}(z) \) to be an eigenvector for the kernel \( K^\infty \) where

\[
\hat{a}(z) = \prod_{a=1}^{A} r_a^a \left( m_a \mid \alpha^a, \theta \right) \exp\left( \alpha^a x_a \right) \prod_{s=1}^{S} r_s^s \left( n_s \mid \beta^s, -\theta \right) \exp\left( \beta^s y_s \right).
\]

**Theorem 3.1** Assume that \( \alpha^a, \beta^s \) are function of \( \gamma \) through the following constraints;

(A) \( \forall a \in \{1 \ldots A\}, \Lambda^a_A (\alpha^a, \gamma) = 0 \) \hspace{1em} (B) \( \forall s \in \{1 \ldots S\}, \Lambda^s_S (\beta^s, -\gamma) = 0 \)

Then \( h(w) = \exp(\theta q) \hat{a}(z) \) with \( \hat{a}(z) \) given in (3.20) is an eigenvector for the kernel \( K^\infty \) with eigenvalue

\[
\exp \Lambda(\gamma) = \exp\left( - \left( \sum_{a=1}^{A} \alpha^a + \sum_{s=1}^{S} \beta^s \right) \right).
\]

**Proof:** In order for \( h \) to be an eigenvector with eigenvalue \( \exp \Lambda(\gamma) \) we need to show

\[
E\left[ \frac{h(W^\infty[1])}{h(W^\infty[0])} \mid W^\infty[0] = w \right] = \frac{h(W^\infty[1])}{h(w)} = \exp \Lambda(\gamma).
\]
Recall that $\forall \ a \in \mathbb{II}_A, \ X^a[1] = A^a_{[1]} - 1$ and $\forall \ s \in \mathbb{II}_S, \ Y^s[1] = S^s_{[1]} - 1$, we obtain that for any transition with a strictly positive probability the ratio takes the following structure:

$$
\frac{h(W^\infty[1])}{h(w)} = \exp \left(-\sum_{a=1}^{A} \alpha^a - \sum_{s=1}^{S} \beta^s \right) \prod_{a \in \mathbb{II}_A} \frac{r_A^a(M_a^a \mid \alpha^a, \gamma)}{r_A^a(m_a \mid \alpha^a, \gamma)} \exp \left(\alpha^a A_{[1]}^a + \gamma B_{[1]}^a \right)
$$

$$
\prod_{s \in \mathbb{II}_S} \frac{r_S^s(N_s^s \mid \beta^s, -\gamma)}{r_S^s(n_s \mid \beta^s, -\gamma)} \exp \left(\beta^s S_{[1]}^s - \gamma R_{[1]}^s \right).
$$

Using

$$
E_w \left[ \exp \left(\alpha^a A_{[1]}^a + \gamma B_{[1]}^a \right) r_A^a(M_a^a \mid \alpha^a, \gamma) \right] = \exp \left(\Lambda^a_A(\alpha^a, \gamma) \right) r_A^a(m_a \mid \alpha^a, \gamma),
$$

$$
E_w \left[ \exp \left(\beta^s S_{[1]}^s - \gamma R_{[1]}^s \right) r_S^s(N_s^s \mid \beta^s, -\gamma) \right] = \exp \left(\Lambda^s_S(\beta^s, -\gamma) \right) r_S^s(n_s \mid \beta^s, -\gamma),
$$

and the assumption of independence, we obtain

$$
E_w \left[ \frac{h(W^\infty[1])}{h(W^\infty[0])} \right] = \exp \left(-\left(\sum_{a=1}^{A} \alpha^a + \sum_{s=1}^{S} \beta^s \right) \right) \prod_{a \in \mathbb{II}_A} \exp \left(\Lambda^a_A(\alpha^a, \gamma) \right) \prod_{s \in \mathbb{II}_S} \exp \left(\Lambda^s_S(\beta^s, -\gamma) \right).
$$

Under the conditions of the proposition, the above expression is equal to $\exp \Lambda(\gamma)$.

**Theorem 3.2** Assume that a solution exists for $\alpha^a$, $\beta^s$ and $\theta$ to the following constraints:

- (A) $\left(\sum_{a=1}^{A} \alpha^a + \sum_{s=1}^{S} \beta^s \right) = 0$ ;
- (B) $\forall \ a \in \{1 \ldots A\}, \ \Lambda^a_A(\alpha^a, \theta) = 0$
- (C) $\forall \ s \in \{1 \ldots S\}, \ \Lambda^s_S(\beta^s, -\theta) = 0$.

Then $h(w) = \exp(\theta q) \hat{a}(z)$ with $\hat{a}(z)$ given in (3.20) is a harmonic function for the kernel $K^\infty$.

**Proof:** In order for $h$ to be harmonic (i.e. $K^\infty h = h$) we need to show that

$$
E \left[ \frac{h(W^\infty[1])}{h(W^\infty[0])} \mid W^\infty[0] = w \right] = E_w \left[ \frac{h(W^\infty[1])}{h(w)} \right] = 1. \quad (3.24)
$$

But under the conditions of the theorem, this follows from Theorem 3.1.
3.4 Construction of the Twisted Kernel.

Once we have determined $\theta$ and hence the $\alpha^s$'s and the $\beta^s$'s by (3.23 B, C), we see that the twisted chain is again a version of a fast teller system. The kernel $K^\infty (w, w')$ of the free (i.e. without boundary) twisted chain is in fact given by

$$
\sum_{(\tilde{b}, \tilde{r}) \in \Upsilon (w, w')} \prod_{a \in I_A} K^a_M (m_a, m'_a) P^a_{AB} \left[ x'_a + 1, b_a \bigg| m_a, m'_a \right] \prod_{s \in I_S} K^s_N (n_s, n'_s) P^s_{SR} \left[ y'_s + 1, r_s \bigg| n_s, n'_s \right] \times
$$

$$
I \left\{ \Xi (w, w') \right\} \times \frac{\exp \left\{ \theta q \right\} \prod_{a=1}^A r^a_A (m'_a \bigg| \alpha^a, \theta) \exp \left\{ \alpha^a x'_a \right\} \prod_{s=1}^S r^s_S (n'_s \bigg| \beta^s, -\theta) \exp \left\{ \beta^s y'_s \right\}}{\exp \left\{ \theta q \right\} \prod_{a=1}^A r^a_A (m_a \bigg| \alpha^a, \theta) \exp \left\{ \alpha^a x_a \right\} \prod_{s=1}^S r^s_S (n_s \bigg| \beta^s, -\theta) \exp \left\{ \beta^s y_s \right\}}.
$$

Using $\Xi (w, w')$, we have that the above sum is equivalent to

$$
K^\infty (w, w') = I \left\{ \Xi (w, w') \right\} \exp \left\{ - \left( \sum_{a=1}^A \alpha^a + \sum_{s=1}^S \beta^s \right) \right\} \times
$$

$$
\sum_{(\tilde{b}, \tilde{r}) \in \Upsilon (w, w')} \prod_{a \in I_A} K^a_M (m_a, m'_a) \frac{r^a_A (m'_a \bigg| \alpha^a, \theta)}{r^a_A (m_a \bigg| \alpha^a, \theta)} P^a_{AB} \left[ x'_a + 1, b_a \bigg| m_a, m'_a \right] \exp \left\{ \alpha^a (x'_a + 1) + \theta b_a \right\} \times
$$

$$
\prod_{s \in I_S} K^s_N (n_s, n'_s) \frac{r^s_S (n'_s \bigg| \beta^s, -\theta)}{r^s_S (n_s \bigg| \beta^s, -\theta)} P^s_{SR} \left[ y'_s + 1, r_s \bigg| n_s, n'_s \right] \exp \left\{ \beta^s (y'_s + 1) - \theta r_s \right\} \right\}.
$$

We define the twisted transition kernels of the modulating Markov chains for the sources by

$$
K^s_M (m_a, m'_a) := K^a_M (m_a, m'_a) \phi^a_{AB} \left( \alpha^a, \theta \bigg| m_a, m'_a \right) \frac{r^a_A (m'_a \bigg| \alpha^a, \theta)}{r^a_A (m_a \bigg| \alpha^a, \theta)}
$$

and the twisted associated probabilities

$$
P^s_{AB} \left[ x_a, b_a \bigg| m_a, m'_a \right] := P^a_{AB} \left[ x_a, b_a \bigg| m_a, m'_a \right] \exp \left\{ \alpha^a x_a + \theta b_a \right\} \phi^a_{AB} \left( \alpha^a, \theta \bigg| m_a, m'_a \right). \tag{3.25}
$$

The servers are treated similarly. Using the last definitions and (3.23, A) we can rewrite the twisted kernel as

$$
K^\infty (w, w') = I \left\{ \Xi (w, w') \right\} \sum_{(\tilde{b}, \tilde{r}) \in \Upsilon (w, w')} \left\{ \prod_{a \in I_A} \left( K^a_M (m_a, m'_a) P^a_{AB} \left[ x'_a + 1, b_a \bigg| m_a, m'_a \right] \right) \right\}
$$

$$
\prod_{s \in I_S} \left( K^s_N (n_s, n'_s) P^s_{SR} \left[ y'_s + 1, r_s \bigg| n_s, n'_s \right] \right) \right\}.
$$
In this form we see that $K^\infty$ is again the kernel of a fast teller system where the kernels of the new modulating chains are the twisted versions, $K^a_M$ and $K^a_N$, of the original chains, and the joint distributions of the waiting time and batch sizes are reweighted versions of the original joint distributions. This means that each twisted source and server is represented by a conjugate Markov additive chain (see Section 5 in Ney and Nummelin (1987)).

### 3.5 Condition (2.9)

We now consider Condition (2.9). Let $\mathcal{M} = \{M^a[t] : a \in \{1, \ldots, A\}\}$ denote the state of the twisted sources and let $\mathcal{N} = \{N^s[t] : s \in \{1, \ldots, S\}\}$ denote the state of the twisted servers. The steady state of the chain $\mathcal{M}^a$ is

$$\varphi^a_M(m_a) := \varphi^a_M(m_a \mid \alpha^a, \theta) = e^a_A(m_a \mid \alpha^a, \theta) r^a_A(m_a \mid \alpha^a, \theta)$$

while the steady state of $\mathcal{N}^s$ is $\varphi^s_N(n_s)$. Note that $\alpha^a = \alpha^a(\theta)$ (by (3.23,B)) and $\beta^s = \beta^s(-\theta)$ (by (3.23,C)), so that in fact $\varphi^a_M(m_a)$ and $\varphi^s_N(n_s)$ depend only on $\theta$.

Denote the excess lifetimes (i.e. the time remaining until the next batch arrival) of the twisted sources by $\mathcal{X}[t] = \{X^a[t] : a \in \{1, \ldots, A\}\}$ and denote the excess lifetimes of the twisted servers by $\mathcal{Y}[t] = \{Y^s[t] : s \in \{1, \ldots, S\}\}$. In practice, we just calculate the moments of the sojourn times for the twisted chains. Denote these mean sojourn times from $m_a$ to $m'_a$ by $\hat{\mu}^a_M(m_a, m'_a)$ and from $n_s$ to $n'_s$ by $\hat{\mu}^s_S(n_s, n'_s)$. Denote the distribution of the sojourn of the twisted source in the state $m_a$ by $F^a_A(x \mid m_a)$ and denote the distribution of the sojourn of the twisted server in the state $n_s$ by $F^s_S(y \mid n_s)$.

Define

$$\hat{\mu}^a_A := \sum_{m_a \in E^a_M} \varphi^a_M(m_a) \sum_{m'_a \in E^a_M} K^a_M(m_a, m'_a) \hat{\mu}^a_M(m_a, m'_a).$$

(3.26)

We conclude that the Markovian component

$$Z^\infty[t] := \left( (\mathcal{M}[t], \mathcal{X}[t]), (\mathcal{N}[t], \mathcal{Y}[t]) \right)$$

of the twisted chain has a stationary probability measure given by

$$\varphi(\bar{m}, \bar{x}, \bar{n}, \bar{y}) := \hat{C} \cdot I_O(\bar{m}, \bar{x}, \bar{n}, \bar{y}) \prod_{a=1}^A \varphi^a_M(m_a) \frac{1 - F^a_A(x_a \mid m_a)}{\hat{\mu}^a_A} \prod_{s=1}^S \varphi^s_N(n_s) \frac{1 - F^s_S(y_s \mid n_s)}{\hat{\mu}^s_S},$$

where the function $I_O$ is the function introduced in (3.17) to assure the irreducibility of the chain $Z^\infty$, and $\hat{C}$ is a normalizing constant.

Also note that the structure of

$$\hat{a}(z) = r(\bar{m}, \bar{x}, \bar{n}, \bar{y} \mid \theta) = \prod_{a=1}^A r^a_A(m_a \mid \alpha^a(\theta), \theta) \exp(\alpha^a(\theta) x_a)$$

$$\times \prod_{s=1}^S r^s_S(n_s \mid \beta^s(-\theta), -\theta) \exp(\beta^s(-\theta) y_s)$$
is given by a product of right Perron-eigenfunctions associated with the Markovian parts and right eigenfunctions associated with the renewal parts. The link between the two kinds of eigenfunctions is in fact given by the parameter $\theta$.

## 3.6 Particular Cases

Here we consider two particular models. The first deals with deterministic inter-arrival and inter-departure times and the second with deterministic inter-departure times and service batches. The second will be useful when we look at an ATM multiplexer in Section 5.

### 3.6.1 Deterministic Inter-Arrival and Inter-Departure Times

We state explicitly the harmonic function for the model in which we restrict the inter-event waiting times to be one. Consequently the excess times remaining until an arrival or a departure are always zero, and so the joint distribution of the variables $A^a_i$ and $S^s_i$ is degenerate, i.e. $\forall \beta \in \mathbb{N}_0$ and $\forall \rho \in \mathbb{N}_0$

$$P_{AB}^a \left[ 1 \times \beta \left\lfloor m_a, m'_a \right\rfloor \right] = P_B^a \left[ \beta \left\lfloor m_a, m'_a \right\rfloor \right]$$

and

$$P_{SR}^s \left[ 1 \times \rho \left\lfloor n_s, n'_s \right\rfloor \right] = P_R^s \left[ \rho \left\lfloor n_s, n'_s \right\rfloor \right]$$

where $P_B^a$ and $P_R^s$ are two probability distributions on $\mathbb{N}_0$. We then obtain $\phi_{AB}^a (u, v \mid m_a, m'_a) = \exp (u) \phi_B^a (v \mid m_a, m'_a)$ and $\phi_{SR}^s (u, v \mid n_s, n'_s) = \exp (u) \phi_R^s (v \mid n_s, n'_s)$. Note $\mathbb{I}_A \equiv \{1 \ldots A\}$ and $\mathbb{I}_S \equiv \{1 \ldots S\}$. We denote by $\exp \Lambda^a_A (\gamma), r^a_A (m_a \mid \gamma),$ and $\ell^a_A (m_a \mid \gamma)$ respectively the Perron-Frobenius eigenvalue and positive eigenvectors of

$$\hat{K}_M^a \left( m_a, m'_a \mid \gamma \right) = \phi_B^a \left( \gamma \mid m_a, m'_a \right) K^a_M (m_a, m'_a)$$

and similarly for the server processes.

**Corollary 3.3** Assume $\theta > 0$ exists such that

$$(A) \quad \Lambda (\theta) = \sum_{a=1}^A \Lambda^a_A (\theta) + \sum_{s=1}^S \Lambda^s_S (-\theta) = 0$$

Then

$$h (w) = \exp (\theta q) \prod_{a=1}^A r^a_A \left( m_a \mid \theta \right) \prod_{s=1}^S r^s_S \left( n_s \mid -\theta \right)$$

is a harmonic function for the kernel $K^\infty$.

**Proof:** We have just to use Theorem 3.2 with

$$\Lambda^a_A (\alpha^a, \gamma) = \Lambda^a_A (\gamma) + \alpha^a \quad \Lambda^s_S (\beta^s, -\gamma) = \Lambda^s_S (-\gamma) + \beta^s,$$

$$r^a_A (m_a \mid \alpha^a, \gamma) = r^a_A (m_a \mid \gamma) \quad r^s_S (n_s \mid \beta^s, -\gamma) = r^s_S (n_s \mid -\gamma)$$

and the fact that $x_a, y_s$ are always equal to zero in this particular case. \[Q.E.D.\]

**Remark.** If $P_B^a \left[ 0 \left\lfloor m_a, m'_a \right\rfloor \right]$ and $P_R^s \left[ 0 \left\lfloor n_s, n'_s \right\rfloor \right]$ are both independent of their respective Markovian transitions then we have geometric inter-arrival times and geometric inter-departure times.
3.6.2 One Customer Served per Time Slot.

For an ATM multiplexer (See Section 5), we must deal with a reduced model in which exactly one customer is served in each time slot. In this special case, the joint distribution of the two variables \( S_i \) and \( R_{si} \) has a degenerate form, \( P_{S|R}^i[1 \times 1 | n_s, n'_s] = 1. \)

**Corollary 3.4** Assume \( \theta > 0 \) and \( \alpha^a \) exist such that

\[
\begin{align*}
(A) & \quad \Lambda(\theta) = - \left( \sum_{a=1}^{A} \alpha^a + S \theta \right) = 0 ; \\
(B) & \quad \forall a \in \{1 \ldots A\} \ , \ \Lambda_A^a(\alpha^a, \theta) = 0
\end{align*}
\]  

(3.27)

Then

\[
h(w) = \exp(\theta q) \prod_{a=1}^{A} r_{A}^{a} \left( m_a \mid \alpha^a, \theta \right) \exp(\alpha^a x_a)
\]

is a harmonic function for the kernel \( K^\infty. \)

**Proof:** We apply Theorem 3.2 with \( \Lambda_{s}^{s} (\beta^s, -\gamma) = -\gamma + \beta^s \) and \( r_{S}^{s} (n_s \mid \beta^s, -\gamma) = 1, \) and use the fact that \( g_s \) is always equal to zero in this particular case.

3.7 Existence questions.

At several points we have tacitly assumed the existence of certain functions and constants. The existence is not immediate, but has been thoroughly studied by Ney and Nummelin (1987). The key condition that the sets of Definition 2.2 be open. Let

\[
\mathcal{D}_A^a := \left\{ (\zeta^a_A, \Lambda^a_A) \in \mathbb{R}^3 \mid \Psi^a_A (\zeta^a_A, \Lambda^a_A) < \infty \right\}
\]

and

\[
(\mathcal{D}_A^a)_1 := \left\{ \zeta^a_A \in \mathbb{R}^2 \mid \exists \Lambda^a_A \in \mathbb{R} : (\zeta^a_A, \Lambda^a_A) \in \mathcal{D}_A^a \right\}.
\]

If \( \mathcal{D}_A^a \) and \( \mathcal{D}_S^s \) are open, then by Lemma 5.3 in Ney and Nummelin (1987) the mean inter-arrival and inter-service times, \( \mu^a_A \) and \( \mu^s_S, \) are finite, as are the mean batch sizes, \( E_{\pi^\infty} \left[ B_{[1]}^{0} \right] \) and \( E_{\pi^\infty} \left[ R_{[1]}^{0} \right]. \)

If we choose \( \zeta^a_A \in (\mathcal{D}_A^a)_1 \) and if we assume \( \mathcal{D}_A^a, \) the domain of convergence of \( \Psi^a_A, \) is open (necessarily containing \( ((0, 0), 0) \)) then by Theorem 4.1 in Ney and Nummelin (1987), the eigensystem \( \ell_A^{a} (m_a \mid \alpha^a, \gamma^A_A), r_{A}^{a} (m_a \mid \alpha^a, \gamma^A_A), \) and \( \exp \Lambda^a_A (\alpha^a, \gamma^A_A) \) exist. As for the sources, if we choose \( \zeta^s_S \in (\mathcal{D}_S^s)_1 \) and if we assume \( \mathcal{D}_S^s \) is open, then we have the existence of the eigenpair and eigenvalue for source \( s. \)

**Lemma 3.5** Suppose \( \mathcal{D}_A^a \) and \( \mathcal{D}_S^s \) are open for each source and server. If \( A0, A1 \) and \( A3 \) hold, then \( \Lambda(\cdot) \) is finite in a neighbourhood of \( 0, \) the eigenvalue-eigenfunction pairs exist, and \( \Lambda \) is a convex function with \( \Lambda(0) = 0, \Lambda'(0) < 0 \) and \( \lim_{\gamma \to \infty} \Lambda(\gamma) = +\infty. \) If further \( \Lambda(\gamma) \) increases continuously to \( +\infty, \) then there is a unique positive value \( \theta \) for which \( \Lambda(\theta) = 0. \)
Here we note that this first positive root can be found reliably by numeric means. The proof is in Section 7.

**Corollary 3.6** Under Conditions A0, A1, and A3, and if $\mathcal{D}$ is open, then $\Lambda(\cdot)$ increases continuously to $+\infty$. A solution to the constraints of Theorem 3.2 exists.

**Proof:** The additional hypothesis that $\mathcal{D}$ is open implies that $\Lambda$ will increase *continuously* to $\infty$. This follows from Ney and Nummelin (1987).

$$\Lambda(\cdot) = \text{open} \implies \Lambda(\cdot) \text{ increases continuously to } +\infty.$$  

If $(\alpha^a(\theta), \theta) \in (\mathcal{D}_A^a)_1$ and if we assume $\mathcal{D}_A^a$ is open, then $\tilde{\mu}_A^a$ is finite by Lemma 5.3, applied to the conjugate Markov additive chain for source $a$. This also implies that for all $m_a, m'_a$, $\tilde{\mu}_A^a(m_a, m'_a) < \infty$. Similar statements apply to the servers.

4 **Results for a fast teller model of Sadowsky and Szpankowski.**

We have established that Conditions A0 through A4 and Conditions (2.9, 2.10) hold under simple conditions for fast teller queues. The remaining conditions of Lemma 2.3, (2.11) through (2.13), are now examined in a special case. This special case is our fast teller queue where the blocks arriving or served are independent of the inter-arrival and inter-service times. This special case remains a generalization of the model of Sadowsky and Szpankowski. Theorem 4.2 below provides complete conditions for the application of Section 2.

The source/server streams have i.i.d. inter-arrival times (whose distributions may depend on the source/server) and i.i.d. batches of cells for each arrival/removal (where the distributions may vary with the source). The batch sizes are independent of the inter-arrival times. Each server serves blocks of cells at the head of the queue.

The inter-arrival density, the twisted inter-arrival density, cumulative distribution, and twisted cumulative distribution for source $a$ are $f_A^a(a)$, $\tilde{f}_A^a(a)$, $F_A^a(a)$, and $\bar{F}_A^a(a)$. The means are $\mu_A^a$ and $\tilde{\mu}_A^a$, and the generating function is $\phi_A^a(u)$. The block arrival cumulative distribution and the twisted block arrival cumulative function, for source $a$ are $F_B^a(b)$ and $\bar{F}_B^a(b)$. The means are $\mu_B^a$ and $\tilde{\mu}_B^a$, and the generating function is $\phi_B^a(u)$. The servers are characterized by $f_S^s(s)$, $\tilde{f}_S^s(s)$, $F_S^s(s)$, $\bar{F}_S^s(s)$, $\mu_S^a$, $\tilde{\mu}_S^a$, $\phi_S^a(u)$, $F_R^r(r)$, $\bar{F}_R^r(r)$, $\mu_R^a$, $\tilde{\mu}_R^a$, and $\phi_R^a$.

The state of the sources and servers is described by the Markov chain $Z^\infty$ whose states are $z = (\vec{x}, \vec{y})$, where $x_a$ is the time remaining until source $a$ delivers a batch and $y_s$, is the time remaining until server $s$ removes a batch. The steady state of $Z^\infty$ is given by (see (3.17))

$$\pi_{Z^\infty}(\vec{x}, \vec{y}) := CI_O(\vec{x}, \vec{y}) \prod_{a=1}^A \frac{1 - F_A^a(x)}{\mu_A^a} \prod_{s=1}^S \frac{1 - F_S^s(y)}{\mu_S^s},$$

where we assume the state space has been reduced to make $Z^\infty$ irreducible on $O$ and $\pi_{Z^\infty}$ was renormalized by $C$ to this irreducible component. The Markovian part of the twisted process has steady state

$$\varphi(\vec{x}, \vec{y}) := I_O(\vec{x}, \vec{y}) \prod_{a=1}^A \frac{1 - F_A^a(x_a)}{\tilde{\mu}_A^a} \prod_{s=1}^S \frac{1 - F_S^s(y_s)}{\tilde{\mu}_S^s}.$$
The expected one step increment of $\mathcal{V}^\infty$ is

$$d := \sum_{z \in E^Z_2} \varphi(z) E_{(0,z)} \left[ \Delta \mathcal{V}^\infty(1) \right]$$

$$= \sum_{a=1}^A \tilde{\mu}_A^a - \sum_{s=1}^S \tilde{\mu}_S^s$$

since the twisted positive increments $\mathcal{B}^{[a]}$ occur only when $x_a = 0$ and the twisted negative increments $\mathcal{R}^{[s]}$ occur only when $y_s = 0$. Define

$$\Lambda(\gamma) = -\left( \sum_{a=1}^A (\phi_A^a)^{-1} \left( 1/\phi_B^a (\gamma) \right) \right) + \sum_{s=1}^S (\phi_S^s)^{-1} \left( 1/\phi_R^s (-\gamma) \right).$$

**Lemma 4.1** Sufficient conditions for the existence of a strictly positive root of the function $\Lambda(\gamma)$ are (3.18) and

$$\lim_{\gamma \to \Gamma} (\Lambda(\gamma)) > 0 \quad (4.31)$$

where $\Gamma := \min(G_B, G_S)$ and where

$$g_S^a := \sup \{ x \in \mathbb{R} : \phi_S^a (x) < +\infty \}, \quad L_S^a := \lim_{x \to g_S^a} \phi_S^a (x), \quad G_S := \min_{s \in \{1, \ldots, S\}} (\phi_S^s)^{-1} (1/L_S^s),$$

$$g_B^a := \sup \{ x \in \mathbb{R} : \phi_B^a (x) < +\infty \}, \quad G_B := \min_{a \in \{1, \ldots, A\}} g_B^a.$$

Again, to prove the existence of such a solution we would have to check Condition A3, (3.18), and that $D$ is open.

**Proof:** If we choose $\gamma > G_S$ then one of the $\beta^s (-\gamma)$ does not exist and hence $\Lambda(\gamma)$ also fails to exist. On the other hand, if we choose $\gamma > G_B$, one of the $\alpha^a (\gamma)$ is equal to $-\infty$ so, if $\Lambda(\gamma)$ exists ($\gamma < G_S$) but the condition (4.31) fails, then the function $\Lambda$ jumps to $+\infty$ without going through zero for a strictly positive value of the parameter $\gamma$. \hfill \blacksquare

**Remark.** There does not exist a constraint on the tail of the distributions of “$A$” and “$R$”. This is intuitively obvious because the exponential nature of the tail probabilities of the stationary queue is conserved if we increase the removed block sizes or increase the inter-arrival times. For $\xi_A^a = (\alpha^a, \gamma_A^a) \in \mathbb{R}^2$ and $\Lambda_A^a \in \mathbb{R}$,

$$\Psi_A^a (\xi_A^a, \Lambda_A^a) := E_{na} \left[ \exp \left( \alpha^a A^a_{[k]} + \gamma_A^a B^a_{[k]} - \Lambda_A^a \right) \right] = \exp(-\Lambda_A^a) \phi_A^a (\alpha^a) \phi_B^a (\gamma_A^a).$$

If the conditions of Lemma 4.1 hold then clearly $\gamma_A (\theta) \in (D_A^a)_1$ and $\alpha^a (\theta) \in (D_A^a)_1$. It follows that for all $a$, the moments $\mu_A^a$ and $\tilde{\mu}_A^a$ are finite. Under the same conditions $\mu_S^s$ and $\tilde{\mu}_S^s$ are finite for all $s$. 

Theorem 4.2 If (4.31) holds, then there are positive constants \( \theta \), \( \{ \alpha^a \colon a = 1, \ldots, A \} \) and \( \{ \beta^s \colon s = 1, \ldots, S \} \) such that \( \exp (\theta q) \hat{a}(z) \) is a harmonic function where

\[
\hat{a}(z) = \exp \left( \sum_{a=1}^{A} \alpha^a x_a \right) \exp \left( \sum_{s=1}^{S} \beta^s y_s \right).
\]

If further A2 and A3 hold, the steady state distribution of \( W \) for \( \ell \to \infty \) and \( z = (\vec{x}, \vec{y}) \) is given by

\[
\pi \left( (\ell, z) \right) \sim \exp (-\theta \ell) \frac{f}{d} \varphi (\vec{x}, \vec{y}) \hat{a}(z)
\]

where

\[
f := \sum_{w=(0,z)} \pi ((w)) \hat{a}(z) H (w),
\]

and where \( H (w) \) represents the probability that \( W^\infty \) does not return to the boundary region \( \{(q, z) : q \leq 0\} \) starting from \( w \).

Finally, let \( (0, \delta) \) be an initial point such that \( \pi^\infty (\delta) > 0 \) and \( \pi (0, \delta) > 0 \). Let \( \tau_\ell \) denote the first time the queue \( Q \) reaches or exceeds the level \( \ell \) and let \( \mu (k, z) \) denote the steady state of the ladder height process \( (V^\infty | \tau^\infty_\ell - \ell, Z^\infty | \tau^\infty_\ell) \), where \( \tau^\infty_\ell \) is the first time \( V^\infty \) reaches or exceeds \( \ell \) starting from an empty queue. Then,

\[
E_{(0, \delta)} \left[ \tau_\ell \right] \sim \exp \left( \theta \ell \right) g^{-1} \quad \text{where} \quad g = f \cdot \sum_{z \in E^\infty} \sum_{k=0}^{\infty} \exp (-\theta k) \hat{a}(z) \mu (k, z).
\]

Proof: The goal is to apply Theorems 2.4 and 2.5. We first have to show that we can find a solution to

\[
\phi^A_\alpha (\alpha^a) \phi^B_\beta (\theta) = 1, \quad \phi^S_\beta (\beta^s) \phi^R_\gamma (-\theta) = 1, \quad \sum_{a=1}^{A} \alpha^a + \sum_{s=1}^{S} \beta^s = 0.
\]

We must find a positive root, \( \theta \), to \( \Lambda (\gamma) = 0 \). We note that \( \alpha^a (\theta) \leq 0 \) and \( \beta^s (-\theta) \geq 0 \) since \( \theta > 0 \). Now we are ready to check the conditions of Lemma 2.3. Conditions A0, A2 and A3 are assumed. Sufficient conditions were given in the last section. Condition A1 follows from (4.31) since the drift of \( \mu \) of \( V^\infty \) given by (3.18) is \( \Lambda' (\gamma) |_{\gamma=0} \), which is negative by convexity and the fact that \( \Lambda (0) = \Lambda (\theta) = 0 \).

The first condition of Lemma 2.3 (2.9), holds with \( \varphi = \varphi \left( (\lambda, \beta, \theta) \right) \) and \( r () \) given respectively by

\[
\varphi (\vec{x}, \vec{y}) := \tilde{C} I_O (\vec{x}, \vec{y}) \prod_{a=1}^{A} \frac{1 - \mathcal{F}^a_A (x_a)}{\tilde{\mu}^a_A} \prod_{s=1}^{S} \frac{1 - \mathcal{F}^s_S (y_s)}{\tilde{\mu}^s_S}
\]

and

\[
r (\vec{x}, \vec{y}) = \prod_{a=1}^{A} \exp (\alpha^a x_a) \prod_{s=1}^{S} \exp (\beta^s y_s).
\]
The second condition (2.10) holds because \( d \), given at (4.30), is \( \Lambda'(\gamma)|_{\gamma=\theta} \) which is positive by convexity. If we substitute \( r(\overline{x}, \overline{y}) \) and the formula for \( \pi_\theta \) into (2.11), Condition (2.11) will hold if for each source

\[
\sum_{x_a \in \mathbb{N}_0} \frac{1 - F^a_A(x_a)}{\mu^a_A} \exp(-\alpha^a x_a) < \infty \quad (4.34)
\]

and similarly for each server. The server sum is finite because \( \beta^s \) is positive. The sum of (4.34) is finite because

\[
\sum_{x_a \in \mathbb{N}_0} \left(1 - F^a_A(x_a)\right) \exp(-\alpha^a x_a) = \sum_{x_a \in \mathbb{N}_0} \sum_{z > x_a} \tilde{f}^a_A(z) \exp(-\alpha^a x_a) \\
= \sum_{z \geq 0} \tilde{f}^a_A(z) \sum_{x_a < z} \exp(-\alpha^a x_a) \\
\leq \frac{1}{\exp(-\alpha^a) - 1} \sum_{z \geq 0} \exp(-\alpha^a z) \tilde{f}^a_A(z) \\
= \frac{1}{\exp(-\alpha^a) - 1} \sum_{z \geq 0} f^a_A(z) < \infty
\]

Condition (2.12) will hold if for each server

\[
\sum_{y_s \in \mathbb{N}_0} \frac{1 - F^s_S(y_s)}{\mu^s_S} \exp(\beta^s y_s) < \infty. \quad (4.35)
\]

and for each source. The source sum is finite because \( \alpha^a \) is negative, while (4.35) holds because

\[
\sum_{y_s \in \mathbb{N}_0} \left(1 - F^s_S(y_s)\right) \exp(\beta^s y_s) = \sum_{y_s \in \mathbb{N}_0} \sum_{z > y_s} f^s_S(z) \exp(\beta^s y_s) \\
\leq \frac{1}{\exp(\beta^s) - 1} \sum_{z \geq 0} f^s_S(z) \exp(\beta^s z) \\
= \frac{1}{\exp(\beta^s) - 1} \phi^s_S(\beta^s) < \infty.
\]

To check Condition (2.13) it suffices to show \( \pi_Z^{\infty}(\overline{x}, \overline{y}) r(\overline{x}, \overline{y} \mid \theta) \) is \( r^{-1} \)--regular for the chain \( Z^{\infty} \). Since \( \alpha^a \leq 0 \) and \( \beta^s \geq 0 \) it suffices to show \( \exp(-\sum a \alpha^a x_a) \) is regular with respect to the measure

\[
\prod_{a=1}^A \frac{1 - F^a_A(x_a)}{\mu^a_A} \exp(\alpha^a x_a)
\]

for the chain \( X[t] \). Since the sources are independent it suffices that for any source \( a \) that \( \exp(-\alpha^a x_a) \) is regular with respect to the measure

\[
\lambda^a_A(x_a) := \frac{1 - F^a_A(x_a)}{\mu^a_A} \exp(\alpha^a x_a)
\]
for the chain $X^a[t]$. This is easy because, taking $\xi = 0$ to be a tagged point, the time $\tau_\xi$ for $X^a[t]$ to hit $\xi$ starting from $x_a$ is exactly $x_a$ since the excess is just reduced by 1 in each time slot. Along the way

$$E_{x_a} \left[ \sum_{t=0}^{\tau_\xi-1} \exp(-\alpha^a X^a[t]) \right] \leq \frac{\exp(-\alpha^a x_a)}{\exp(-\alpha^a) - 1}.$$ 

Therefore

$$E_{\lambda^\infty} \left[ \sum_{t=0}^{\tau_\xi-1} \exp(-\alpha^a X^a[t]) \right] \leq \sum_{x_a=0}^{\infty} \frac{1 - F^\infty_a(x_a)}{\bar{\mu}_a} \exp(\alpha^a x_a) \frac{\exp(-\alpha^a x_a)}{\exp(-\alpha^a) - 1} = \frac{1}{\exp(-\alpha^a) - 1} < \infty.$$ 

This gives Condition (2.13).

5 Application: The ATM Multiplexor

Suppose we wish to simulate the situation where a hot spot develops (i.e. the queue at that point builds up) in the switching fabric of an ATM switch. We shall suppose that each of the input ports receive cells from sources like those in Section 3. For source $a$, the underlying chain $M^a$ will indicates to which output port the last cell processed was directed. Now $A^a$ will be the period between cell arrivals. The batch size $B^a$ will be 0 or 1 depending on whether or not a cell must pass through the hot spot on its way to its output port. We will let $Q$ denote the workload in the switch, that is the number of cells queued within the switch which will pass through the hot spot. Now define the Markov chain $(Q, \vec{M}, \vec{X})$ representing the workload of the hot spot, the state of each source and the time until the next cell arrival for each source.

Suppose one of the buffers at the hot spot switching point has a finite buffer of size $\ell$ and the quantity of interest is the cell loss rate caused by overflowing this particular buffer. We will assume the protocol used to manage the hot spot buffers is work conserving, so that as long as one cell is in a hot spot queue one cell will be transmitted in the next time slot. We will assume the buffers upstream of the hot spot are infinite (or so big they won’t overflow). Lamothé (1998) removes this constraint keeping track of the extra factors introduced into the likelihood function of any trajectory which hits the maximum buffer size. Since transfers between buffers do not affect the workload directed at a hotspot, Lamothé also consider switches with different link rates for their buffers. For example, internal link rates are typically higher than input/output port link rates.

Now suppose the free chain $(V^\infty, \vec{M}^\infty, \vec{X}^\infty)$ starts at some point $w = (v, \vec{m}, \vec{x})$. By the above assumptions the free workload $V^\infty$ evolves from $v$ to $v - 1 + \sum_{a\in\mathcal{A}} \|l_a(w)\| b_a$ where $\|l_a(w)\|$ is the set of sources $a$ such that $x_a = 0$. If the state $a$ produces cells which pass through the hot spot then define $b_a = 1$ and 0 otherwise. This situation now fits into the single-server case of Corollary 3.4.

For $a \in \{1, \ldots, A\}$ let $r^\infty_a(m_a, \alpha^a, \theta)$ be the right Perron-Frobenius eigenvectors of $\tilde{K}^a_a(m, \alpha^a, \gamma)$ associated with the eigenvalue $\exp(\Lambda^a_a(\alpha^a, \gamma))$. We construct a harmonic function for the kernel $K^\infty$ of the form in (3.28) if we pick $\alpha^a$ and $\gamma$ to satisfy (3.27).
Now generate normal cycles of the untwisted queue which end if any of the hot spot buffers empties. Next starting from a generated end point, say \((v_0, (\vec{m}_0, \vec{x}_0))\), generate the twisted cycle and consider a twisted cycle which again ends when any one of the hot spot buffers empties. If the workload at the hot spot overloads, i.e. reaches or exceeds the workload level \(\ell\), we turn off the twist and calculate the factor \(f := h(v_0, (\vec{m}_0, \vec{x}_0)) / h(v, (\vec{m}, \vec{x}))\) where \((v, (\vec{m}, \vec{x}))\) is the state of \(W^\infty\) when the workload reaches or exceeds the level \(\ell\). Now let the chain evolve normally (untwisted) until this cycle ends. We count the number of cells lost, say \(N\). In some cases the cycle will end without the hotspot buffer ever reaching overload; then \(N := 0\) and the factor \(f\) is irrelevant. For each cycle we calculate the product of the above factor times the number of cells lost.

If we alternate between normal and twisted cycles a large number of times we can estimate the steady state length of a normal cycle by the average cycle length. Moreover we can estimate the expected value of \(f \cdot N\) by averaging over the twisted cycles. Finally the expected number of cells lost per unit time is given by the expected value of \(f \cdot N\) divided by the expected normal cycle length and we have estimates for both these quantities.

The advantage of the above method is that the factor \(f\) does not have a large standard deviation relative to its expected value. In fact for all cycles, \(f\) contains a common factor of \(\exp(-\gamma\ell)\). It may be that overloading the hot spot buffer is still a rare event even after the workload at the hot spot reaches \(\ell\). Nevertheless this event is \(\exp(\gamma\ell)\) times more likely than the probability the untwisted queue overloads before one of the hot spot buffers empties. This means the simulation time is reduced by the same factor.

## 6 Remarks

### 6.1 Effective Bandwidth – an Engineering Perspective

Stability is not a sufficient criterion for engineering applications; more precise specifications are needed. The effective bandwidth of a source or server is a more informative way to summarize the performance capability of that source or server. Using the effective bandwidth of the sources and servers one can specify the asymptotic rate of decay of the tail of the distribution of the queue size. By adding servers or removing sources, it is then possible to design the queue so that there is a specified (very small!) probability that it will overflow. The reader is referred to Kesidis, Walrand and Chang (1993) or Kelly (1991, 1996) for additional material on this topic.

Here we develop a notion of effective bandwidth corresponding to our development, and provide a rule of thumb for combining sources and servers of a single queue to obtain a specified distributional tail decay. We have seen that the asymptotic rate of \(P_\pi(Q \geq \ell)\) is given by \(\exp(-\gamma\ell)\) where \(\gamma\) is chosen correctly (i.e. \(\gamma = \theta\)).

In the setting of Section 2, and following Kelly (1996), define

\[
bw(\gamma, t) := \log E_{(0,2)} \left[ \frac{\exp(\gamma V^\infty[t])}{\gamma t} \right]
\]

to be the effective bandwidth function of \(W^\infty\). Hence

\[
\lim_{t \to \infty} bw(\gamma, t) = \Lambda(\gamma) / \gamma.
\]
Recall $\Lambda(\theta) = 0$, so when the queue is in equilibrium, the excess effective bandwidth of the system is 0 and the tail of the distribution of the queue decays at rate $\theta$.

If we have only one source and only one server with geometric arrival and service, we found by Corollary 3.3 that $\Lambda'(\theta) = \Lambda_A'(\theta) + \Lambda_S'(\theta)$. Dividing this last equation by $\theta$ we see $bw_A(\theta) - bw_S(\theta) = 0$, where the effective bandwidth of the source is $bw_A(\gamma) := \Lambda_A(\gamma)/\gamma$ and that of the server is $bw_S(\gamma) := \Lambda_S(-\gamma)/-\gamma$. For a stable queue, the bandwidth of the source and server must balance.

The fast teller application of Section 3 leads to a similar decomposition of $\Lambda(\cdot)$ as $\Lambda_A(\cdot) + \Lambda_S(\cdot)$ where the arrival and service terms are both sums over several components. Consequently the bandwidth of the combined sources and servers can be decomposed as sums of the bandwidths of the components. The effective bandwidth of source $a$ is given by $bw_A^a(\gamma) := \alpha^a(\gamma)/\gamma$, and the effective bandwidth of server $s$ is given by $bw_S^s(\gamma) := \beta^s(-\gamma)/-\gamma$. Hence the bandwidth of the whole system is

$$\Lambda(\gamma) = \sum_{a=1}^A bw_A^a(\gamma) - \sum_{s=1}^S bw_S^s(\gamma).$$

(6.36)

In equilibrium the bandwidth of the system is 0.

For a given arrangement of sources and servers attached to a single queue, (6.36) tells us whether or not a desired asymptotic tail behaviour of rate $\theta$ will be achieved. If the difference is negative, $P\{Q > \ell\} = O(\exp(-\theta\ell))$. If the difference exceeds 0, then $P[Q > \ell]$ will decay more slowly than the desired rate. We also see how to compensate for adding sources by adding additional servers.

### 6.2 Additional Problems.

Here, we assume conditions on the generating functions of arrival blocks ($B^a$) and on inter-service times ($S^s$). Of course there is no problem if all distributions are finite but heavy tailed distributions are excluded. See Asmussen and Klüppelberg (1997) for new results on asymptotics of an $(M/G/1)$-queue when $G$ has a heavy tail.

We have not discussed a continuous time fast teller queue. This would require an extension of the theory in McDonald (1996) beyond Markov jump processes. This is possible and the harmonic function should not change.

### 7 Proofs

**Lemma 7.1** Recall $S_1$ from the discussion preceding Lemma 2.1. Under Condition A1 and if $z_0 \in \mathcal{I}$, $E_{z_0}[S_1^\infty] < 0$.

**Proof:** Consider the sequence of return times $T_n$ to a state $z_0$ by $Z^\infty$. Since $Z^\infty$ has a steady state $\pi_{Z^\infty}$ and $E_{z_0}[T_1] < \infty$, the steady state $\pi_{Z^\infty}(y)$ for any state $y \in \mathcal{E}Z^\infty$, is given by $\pi_{Z^\infty}(x_0) E_{z_0}[N(y)]$ where

$$E_{x_0}[N(y)] = E_{x_0}[\text{Number of visits to } y \text{ before returning to } x_0]$$
Therefore, for any function $g$,

$$
\sum_{y \in E_\infty^Z} g(y) \pi_{\infty}^Z(y) = \pi_{\infty}^Z(x_0) \sum_{y \in E_\infty^Z} g(y) E_{x_0}[N(y)] = \pi_{\infty}^Z(x_0) \sum_{k=1}^{T_1} g(Z_{\infty}[k])
$$

Use the above expression by taking $g(y) = E_y[\Delta V_{\infty}[1]]$ and use the fact that

$$
E_{x_0} \left[ \sum_{k=1}^{T_1} E_{Z_{\infty}[k]} \Delta V_{\infty}[1] \right] = E_{x_0} \left[ \sum_{k=1}^{T_1} \Delta V_{\infty}[k] \right]
$$

(which holds because $E[\Delta V_{\infty}[k] | Z_{\infty}[k]] = E_{Z_{\infty}[k]} \Delta V_{\infty}[1]$). This gives

$$
E_{x_0} [S_{1}^\infty] = E_{x_0} \left[ \sum_{k=1}^{T_1} \Delta V_{\infty}[k] \right] = \sum_{y \in E_\infty^Z} \pi_{\infty}^Z(y) E_y[\Delta V_{\infty}[1]] = \frac{\mu}{\pi_{\infty}^Z(z_0)} =: v. \tag{7.37}
$$

By A1, $v < 0$. It follows that the average increment in $V_{\infty}$ between returns to $z_0$ is negative.

**Proof of Lemma 2.1:** Define $T_1$ as above, and define the chain $Y_n = Q[T_n]$ on $\mathbb{N}_0$. By Conditions A2 and A3 (see the remarks preceding Lemma 2.1) $Y_n$ is irreducible on a component including $(L, \infty)$, where $L = L(z_0)$ is some positive integer. The transition kernel of $Y_n$ agrees with that of $V_{\infty}$ for transitions in $\mathbb{N}$ so, starting from a point $y$ sufficiently large, the transitions will agree with high probability. Hence for all $\epsilon < |v|$ ($v$ of (7.37)) we can pick a $B$ such that $E_y[Y_1 - y] < v + \epsilon := -\delta$ for $y \geq B$ where $\delta > 0$. We can define a Liapounov function $U(y) = y/\delta$ such that

$$
E_y \left[ U(Y_1) - U(y) \right] \leq -1 + bI \left\{ y \in \{1 \ldots B\} \right\}
$$

where $b$ is some constant. The positive recurrence of $Y_n$ follows from Theorem 11.3.11 in Meyn and Tweedie (1993).

Pick some point $q_0$ such that $(q_0, z_0)$ is in the irreducible component of $W, I$. Let $N$ denote the first return time $n > 0$ such that $Y_n = q_0$. By the above $E_0[N] < \infty$. Next, starting from $(q_0, x_0)$, the total number of transitions for the chain $W$ to return to $(q_0, x_0)$ is $\sum_{k=1}^N T_k$. Since the random variables $\{T_k : k \in \mathbb{N}\}$ are i.i.d. and positive and since $N$ is a stopping time for this sequence it follows from Wald’s lemma and Lemma 7.1 that

$$
E_{(0, z_0)} \left[ \sum_{k=1}^N T_k \right] = E_0[N] E_{x_0}[T_1] < \infty.
$$

We conclude that the chain $W = (Q, Z)$ is positive recurrent on the irreducible component $I$ and therefore has a stationary distribution putting its mass on $I$.

**Proof of Lemma 2.3:** If (2.9) holds then the Markovian component $Z_{\infty}$ is positive recurrent, and hence Condition (1) of McDonald (1996) holds. The aperiodicity Condition (2) of McDonald (1996)
has been assumed in the condition (2.3). To complete Condition (2) we recall that Condition (M1) of Ney and Nummelin (1987) is vacuous for discrete time chain on a countable state space like $\mathcal{E}^\infty$.

Using Lemma 2.3 (the drift lemma) in McDonald (1996) it is easy to check that

$$d := \sum_{z \in \mathcal{E}_Z^\infty} \varphi(z) E_{(0,z)} \left[ \Delta \nu^\infty[1] \right] > 0$$

(7.38)

since by hypothesis $\mu < 0$ (2.2). Hence Condition (3) holds. This also means the queue of the twisted system has a non-zero probability of drifting toward plus infinity without ever emptying. Hence Condition (4) holds. Condition (5) will hold by line (2.11). Using the remark found after the Lemma 1.1 in McDonald (1996), it is easy to show that Condition (6) follows from line (2.12).

By definition $\lambda_\alpha(z)$ is $\left(\tilde{a}(z)\right)^{-1}$—regular for the chain $\mathcal{Z}^\infty$ if for each $B \in \mathcal{B}_Z$ such that $\sum_{z \in B} \pi_{Z0}^\infty(z) > 0$

$$E_{\lambda_\alpha} \left[ \sum_{k=0}^{\tau_B-1} \left( \tilde{a}(\mathcal{Z}^\infty[k]) \right)^{-1} \right] < \infty$$

where $\tau_B$ is the time until $\mathcal{Z}^\infty$ hits $B$. Rewriting this expression, we get

$$\sum_{z} \pi(0, z) + E_{\lambda_\alpha} \left[ \sum_{k=1}^{\tau_B-1} \left( \tilde{a}(\mathcal{Z}^\infty[k]) \right)^{-1} \right]$$

$$= \sum_{z} \pi(0, z) + \sum_{z} \pi(0, z) \tilde{a}(z) \sum_{q} \sum_{z'} K^\infty \left( (0, z), (q', z') \right) E_{(0,z')} \left[ \sum_{k=0}^{\tau_B-1} \left( \tilde{a}(\mathcal{Z}^\infty[k]) \right)^{-1} \right]$$

$$\geq \sum_{z'} \lambda_\alpha(z') E_{(0,z')} \left[ \sum_{k=0}^{\tau_B-1} \left( \tilde{a}(\mathcal{Z}^\infty[k]) \right)^{-1} \right] = E_{\lambda_\alpha} \left[ \sum_{k=0}^{\tau_B-1} \left( \tilde{a}(\mathcal{Z}^\infty[k]) \right)^{-1} \right]$$

so $\lambda_\alpha$ as defined in Condition (7) is $\left(\tilde{a}(z)\right)^{-1}$—regular. $\blacksquare$

**Proof of Lemma 3.5**: That the $\mathcal{D}_A^+$ sets are open implies the existence of the $\Lambda_A^a$ and the eigenfunctions $\ell_A^a$ and $r_A^a$ used below (see Section 3.1 of Ney and Nummelin (1987)). The server case is similar. The $\Lambda_A^a$ functions are convex as Laplace transforms. The finiteness of $\Lambda$ in a neighbourhood of 0, the existence of its eigenfunctions and its convexity follow from its definition in terms of the source and server functions.

Note that at $\gamma = 0$, $\alpha^a(0) = 0$, and $\beta^a(0) = 0$. Thus $\Lambda(0) = 0$. We now establish that $\Lambda'(0) < 0$. Note that $r_A^a(m_a \mid 0, 0) = 1$ and $\ell_A^a(m_a \mid 0, 0) =: \pi_M^a(m_a)$ is the stationary probability for $K_M^a$. Also note that for all $\gamma$ in a neighbourhood of 0

$$\sum_{m_a \in \mathcal{E}_M^a} \ell_A^a \left( m_a \mid \alpha^a(\gamma), \gamma \right) r_A^a \left( m_a \mid \alpha^a(\gamma), \gamma \right) = 1.$$

Now take the derivative with respect to $\gamma$ in the above expression and evaluate it at $\gamma = 0$. Hence

$$\sum_{m_a \in \mathcal{E}_M^a} \pi_M^a(m_a) \frac{d}{d\gamma} \left[ r_A^a \left( m_a \mid \alpha^a(\gamma), \gamma \right) \right]_{\gamma=0} + \sum_{m_a \in \mathcal{E}_M^a} \frac{d}{d\gamma} \left[ \ell_A^a \left( m_a \mid \alpha^a(\gamma), \gamma \right) \right]_{\gamma=0} \cdot 1 = 0.$$
Exchanging the order of integration and differentiation is permitted since the summands are non-negative and the series are summable. Using stationarity, the above gives

\[
\sum_{m_a \in E_M^a} \pi_M^a(m_a) \frac{d}{d\gamma} \left[ r_M^a \left( M^a_{[1]} \big| \alpha^a(\gamma), \gamma \right) \right]_{\gamma=0} = - \sum_{m_a \in E_M^a} \frac{d}{d\gamma} \left[ \ell_A^a \left( m_a \big| \alpha^a(\gamma), \gamma \right) \right]_{\gamma=0}. \tag{7.39}
\]

Next remark that

\[
1 = \exp \Lambda_A^a (\alpha^a(\gamma), \gamma) = \sum_{m_a \in E_M^a} \ell_A^a \left( m_a \big| \alpha^a(\gamma), \gamma \right) \exp \Lambda_A^a (\alpha^a(\gamma), \gamma) r_A^a \left( m_a \big| \alpha^a(\gamma), \gamma \right)
\]

\[
= \sum_{m_a \in E_M^a} \sum_{m_a' \in E_M^a} \ell_A^a \left( m_a \big| \alpha^a(\gamma), \gamma \right) K_M^a \left( m_a, m_a' \big| \alpha^a(\gamma), \gamma \right) r_A^a \left( m_a' \big| \alpha^a(\gamma), \gamma \right)
\]

\[
= \sum_{m_a \in E_M^a} \ell_A^a \left( m_a \big| \alpha^a(\gamma), \gamma \right) \overline{E} (\alpha^a(\gamma), \gamma) \tag{7.40}
\]

where

\[
\overline{E} (\alpha^a(\gamma), \gamma) = E \left[ \exp \left( \alpha^a(\gamma) A^a_{[1]} + \gamma B^a_{[1]} \right) r_A^a \left( M^a_{[1]} \big| \alpha^a(\gamma), \gamma \right) \bigg| M^a_{[0]} = m_a \right].
\]

Take the derivative of (7.40) with respect to \( \gamma \) and evaluate at \( \gamma = 0 \).

\[
\sum_{m_a \in E_M^a} \left( \frac{d}{d\gamma} \left[ \ell_A^a \left( m_a \big| \alpha^a(\gamma), \gamma \right) \right]_{\gamma=0} \cdot 1 + \pi_M^a(m_a) \cdot \frac{d}{d\gamma} \left[ \overline{E} (\alpha^a(\gamma), \gamma) \right]_{\gamma=0} \right) = 0,
\]

\[
\frac{d\overline{E}}{d\gamma} \bigg|_{\gamma=0} = E \left[ \left. 1 \cdot \frac{d}{d\gamma} \left[ r_M^a \left( M^a_{[1]} \big| \alpha^a(\gamma), \gamma \right) \right]_{\gamma=0} \right| + \left( A^a_{[1]} \frac{d}{d\gamma} \left[ \alpha^a(\gamma) \right]_{\gamma=0} + B^a_{[1]} \right) \cdot 1 \bigg| M^a_{[0]} = m_a \right].
\]

Using line (7.39) and again interchanging \( E \) and \( \frac{d}{d\gamma} \), we have

\[
0 = \sum_{m_a \in E_M^a} \pi_M^a(m_a) E \left[ A^a_{[1]} \frac{d}{d\gamma} \left[ \alpha^a(\gamma) \right]_{\gamma=0} + B^a_{[1]} \bigg| M^a_{[0]} = m_a \right] + \sum_{m_a \in E_M^a} \pi_M^a(m_a) E \left[ A^a_{[1]} \bigg| M^a_{[0]} = m_a \right] + \pi_M^a \left[ B^a_{[1]} \bigg| M^a_{[0]} = m_a \right].
\]

In other words

\[
\frac{d}{d\gamma} \left[ \alpha^a(\gamma) \right]_{\gamma=0} = - E_{\pi_M} \left[ B^a_{[1]} \left( E_{\pi_M} \left[ A^a_{[1]} \right] \right)^{-1} \right].
\]

A similar calculation shows

\[
\frac{d}{d\gamma} \left[ \beta^s(-\gamma) \right]_{\gamma=0} = E_{\pi_N} \left[ R^s_{[1]} \left( E_{\pi_N} \left[ S^s_{[1]} \right] \right)^{-1} \right].
\]
Since
\[
\sum_{a=1}^{A} E_{\pi_M}^a \left[ B_{[1]}^a \right] \left( E_{\pi_M}^a \left[ A_{[1]}^a \right] \right)^{-1} - \sum_{s=1}^{S} E_{\pi_N}^s \left[ R_{[1]}^s \right] \left( E_{\pi_N}^s \left[ S_{[1]}^s \right] \right)^{-1}
\]
is negative by hypothesis (see line (3.18)), the function \( \Lambda(\gamma) \) takes the value 0 and has a negative derivative when \( \gamma = 0 \). As stated in Definition 2.2, condition A3 assures that \( \Lambda(\gamma) \) will attain \( +\infty \) as \( \gamma \to \infty \).

Since \( \Lambda(0) = 0 \) and \( \Lambda'(0) < 0 \), and since \( \Lambda \) is convex, it follows that if \( \Lambda(\gamma) \) increases continuously to \( +\infty \) then there exists a unique point \( \gamma = \theta > 0 \) such that \( \Lambda(\theta) = 0 \).
Acknowledgments

The authors would like to acknowledge the assistance of Gilles Lamothe in the preparation of this work. The first author wants especially to acknowledge the financial assistance received from Professor McDonald and the Ottawa-Carleton Institute for Graduate Studies and Research in Mathematics and Statistics. The last author thanks the School of Industrial and Systems Engineering at Georgia Tech for their hospitality and technical support. The authors would also extend their thanks and appreciation to the referee for his valuable comments.

References


Benoit Beck
Institut de Statistique
Université Catholique de Louvain
Voie du Roman Pays, 20
B-1348 Louvain-La-Neuve, Belgium

A.R. Dabrowski and D.R. McDonald
Department of Mathematics and Statistics
University of Ottawa
Ottawa, Ontario
Canada K1N 6N5