UNIVERSITY OF CINCINNATI

Date: February 4, 2004

I, Ramya Dhamodaran **,**

hereby submit this work as part of the requirements for the degree of:

Master of Science

in:

Industrial Engineering

It is entitled:

Efficient Analysis of Rare Events Associated with
Individual Buffers in a Tandem Jackson Network

This work and its defense approved by:

Chair: Dr. Bruce C. Shultes

Dr. Sam Anand

Dr. Emmanuel Fernandez

Efficient Analysis of Rare Events Associated with Individual Buffers in a Tandem Jackson Network

A thesis submitted to the

Division of Research and Advanced Studies

of the University of Cincinnati

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in the Department of

Mechanical,Industrial and Nuclear Engineering

of the College of Engineering, University of Cincinnati

March 2004

by

Ramya Dhamodaran

B.E. (Industrial Engineering)

Anna University, Chennai, TN, INDIA.

June 2001

Thesis Advisor and Committee Chair: Dr. Bruce Shultes

ABSTRACT

For more than a decade, importance sampling has been a popular technique for the efficient estimation of rare event probabilities. This thesis presents an approach for applying balanced likelihood ratio importance sampling to estimate rare event probabilities in tandem Jackson networks. The rare event of interest is the probability that the content of the second buffer in a two node tandem Jackson network reaches some high level before it empties. Heuristic importance sampling distributions are derived that can be used to estimate this overflow probability in cases where the first buffer capacity is finite and infinite. In the proposed methods, the transition probabilities of the embedded discrete-time Markov chain are modified dynamically to bound the overall likelihood ratio of each cycle. The proposed importance sampling distributions differ from previous balanced likelihood ratio methods in that they are specified as functions of the contents of the buffers. When the first buffer capacity is infinite, the proposed importance sampling estimator yields bounded relative error except when the first server is the bottleneck. In the latter case, numerical results suggest that the relative error is linearly bounded in the buffer size. When the first buffer capacity is finite, empirical results indicate that the relative errors of these importance sampling estimators are bounded independent of the buffer size when the second server is the bottleneck and bounded linearly in the buffer size otherwise.

ACKNOWLEDGEMENTS

I take this opportunity to extend my sincere thanks and appreciation to many people who made this thesis possible.

First and foremost, I would like to thank my parents Vasantha and Dhamodaran for all their prayers, love and support throughout my life.

I express my sincere gratitude towards my advisor, Dr.Bruce Shultes for his continuous guidance, motivation and financial support. I thank him in particular for spending time at length to discuss the issues concerning my research and helping me out with them. I also greatly appreciate his efforts in helping me with the documentation of this thesis.

I would like to thank Dr. Sam Anand and Dr. Emmanuel Fernandez for taking their valuable time to review my thesis and serve in my thesis committee.

I am thankful to my friends in the Operations Research lab, for their help during the course of this work.

I would like to express my sincere thanks to my brother Ravishankar, sister Sripriya and brother-in-law Mohan, who provided constant encouragement, support and took care of me in many aspects. I would also like to thank my cousins, Swaminathan and Srividya for all the affection, camaraderie and caring they provided. I am also grateful towards the help and support provided by my friends in Cincinnati.

Contents

List of Tables

Chapter 1

Introduction

Performance measures of highly reliable systems are hard to compute since they depend upon the occurrence of rare events. Tandem Jackson networks (for an introduction to Jackson networks see Chapter 1, Serfozo 1999) serve as a simplified model for analyzing rare events in many reliable systems such as switched telecommunication networks, manufacturing systems and computer networks. System performance measures such as the probability that the system size or a specific queue length exceeds a given level are needed to accurately assess system reliability, particularly the time until one of these events occurs.

Standard Monte Carlo simulation is inefficient in producing accurate estimates of rare event probabilities since it requires prohibitively long run lengths. In standard Monte Carlo simulation, the stochastic behavior of the system is not modified to force the rare event to occur and the rare event is not observed very

often. Consequently, the number of simulation trials required to get a precise estimate of the probability of the rare event is very large.

1.1 Literature Review

Importance sampling is gaining popularity as an efficient method for analyzing rare events in queueing and reliability systems (see Asmussen and Rubinstein 1995, Glynn and Iglehart 1989, Heidelberger 1995). The main idea of importance sampling is to force a simulation to observe a rare event frequently. The application of importance sampling involves simulating the model using an auxiliary distribution designed to make the system experience rare events of interest more often. The auxiliary distribution modifies the stochastic behavior of the system such that events that lead towards a rare event are more likely to happen and thus more samples hit the rare event. The sample values obtained by using the auxiliary distribution are then adjusted by using likelihood ratios in order to account for the modifications to the stochastic process leading to an unbiased estimator. The auxiliary distribution should be formed such that variance reduction is achieved when compared to standard Monte Carlo simulation.

An importance sampling distribution that yields a constant value for every sample (zero-variance importance sampling) is almost impossible because it requires perfect knowledge of the quantity being estimated. Kuruganti and Strickland (1997) identify properties that characterize zero-variance importance sam-

pling distributions and use them to develop a method to compute an optimal measure for a tandem queueing system. Juneja (1993, 2001) develops these properties as a basis for identifying asymptotically optimal importance sampling distributions.

Large deviations theory has been used for deriving and analyzing importance sampling estimators. Using large deviations theory, a heuristic change of measure was derived for estimating the probability that total system size exceeds a given level before returning to zero in tandem Jackson networks (see Parekh and Walrand 1989). This exponential twisting or tilting change of measure interchanges the arrival rate and the smallest service rate in the network. This heuristic was later analyzed by Glasserman and Kou (1995) who established necessary and sufficient conditions for the asymptotic efficiency of this heuristic importance sampling estimator. An adaptive importance sampling method for estimating overflow probabilities by minimizing the cross-entropy between a zero-variance distribution and the proposed importance sampling distribution has been developed (de Boer et al. 2000). Recently, de Boer, Kroese and Rubinstein (2002) proposed a modified approach which utilizes an optimal tilting parameter to estimate the overflow probability in three stages.

The balanced likelihood ratio approach to importance sampling (see Alexopoulos and Shultes 1998, 2001) was developed for analyzing system performance in fault-tolerant repairable systems. This approach has been used to derive importance sampling estimators for limiting system unavailability and mean time to

system failure that yield bounded relative error. Shultes (2002) applied this approach to estimate the system overflow probability in tandem Jackson networks. This method yields a zero variance importance sampling distribution for a single node system. For systems with more than one node, this method yields asymptotically efficient results with some restrictions on the model parameters.

The rare event studied in this thesis is the buffer overflow probability at the second node in a two node tandem Jackson network. An exponential tilting technique was developed by Kroese and Nicola to estimate this overflow probability (see Kroese and Nicola 2002). These authors exponentially tilt a Markov additive process representation of the system to derive an importance sampling estimator. Their distribution is state dependent in that it depends on the contents of the first buffer.

1.2 Contributions of this thesis

In this thesis work, an importance sampling distribution for estimating the overflow probability at the second node in a two node tandem Jackson network is derived using balanced likelihood ratio approach. The proposed distributions have guaranteed variance reduction over standard Monte Carlo methods. The proposed distributions depend on the contents of the buffers and can be applied to any set of arrival and service rates. When the first buffer is infinite, the proposed estimator is asymptotically optimal except when the first server is bottleneck. In the latter

case, numerical results indicate that the relative error is linearly bounded in the buffer size.

Chapter 2 presents the model studied and provides an overview of importance sampling and the balanced likelihood ratio approach. Chapter 3 provides details of the proposed method for the infinite and finite first buffer cases. Chapter 4 contains experimental results. Conclusions and future research directions are presented in Chapter 5.

Chapter 2

Background

Consider a tandem Jackson network with two nodes. Customers arrive at the first queue according to a Poisson process with rate λ. The service time of a customer at the first node is exponential with rate μ_1, independent of the input process and service time at the second node. The output process of the first queue forms the input process of the second queue. The service time at the second node is exponential with rate μ_2, which is also independent of the input process and service time at the first node. Without loss of generality, assume that $\lambda + \mu_1 + \mu_2 = 1$. The queueing system is assumed to be stable, i.e., $\lambda < \min(\mu_1, \mu_2)$.

Let $X(t)$ and $Y(t)$ denote the number of customers at the first and second node at time t, respectively (including customers in service). Let b denote the size of the first buffer, which may be finite or infinite. The quantity of interest is the probability (γ) that the number of customers in the second queue reaches some

high level $B \in \mathbb{N}$ before hitting 0. We wish to estimate this probability given that the system starts in state $(X(0) = 0, Y(0) = 0)$ or $(X(0) = 1, Y(0) = 1)$. These probabilities are denoted as γ_0 and γ_1 respectively. To estimate γ_0 (γ_1), the simulation cycle starts from (0,0) ((1,1)) and ends when either the second queue reaches B or zero.

The system can be modeled as a Markov process with system state $Z(t) = (X(t), Y(t))$. Let

$$r(t) = \lambda + 1\left(X(t) > 0\right)\mu_1 + 1\left(Y(t) > 0\right)\mu_2$$

denote the total rate of event transitions out of $Z(t)$. The buffer overflow probability depends upon the embedded discrete-time Markov chain whose one-step transition probabilities at time t are: $\lambda/r(t)$ the probability the next event is an arrival, $1\left(X(t) > 0\right)\mu_1/r(t)$ the probability that the next event is a service completion at node one, and $1\left(Y(t) > 0\right)\mu_2/r(t)$ the probability that the next event is a service completion at node two.

2.1 Importance Sampling

Let Ω denote the set of all cycles and for each $\omega \in \Omega$, let $\beta(\omega)$ denote the largest number of customers at the second node within the cycle. Consider an indicator

function, $\phi(\omega)$, which is defined as follows:

$$\phi(\omega) = \begin{cases} 1 \text{ if } \beta(\omega) = B \\ 0 \text{ if } \beta(\omega) < B \end{cases}$$

The buffer overflow probability of interest $(\gamma_i, i = 0, 1)$ can be written as

$$\gamma_i = E_P[\phi(\omega)]$$

where the subscript P denotes sampling from the probability measure P . In standard Monte Carlo simulation, γ_i can be estimated by drawing N independent samples under the probability measure P as follows:

$$\bar{\gamma}_i = \frac{1}{N} \sum_{j=1}^{N} \phi(\omega_j)$$

The estimator $\bar{\gamma}_i$ is an unbiased estimator of γ_i and $E_P[\bar{\gamma}_i] = \gamma_i$. The variance of $\bar{\gamma}_i$ is $\gamma_i(1-\gamma_i)/N$. By using the central limit theorem, a confidence interval for $\bar{\gamma}_i$ can be constructed as $\bar{\gamma}_i \pm z_{\alpha/2} \sqrt{\gamma_i(1-\gamma_i)/N}$, where $z_{\alpha/2}$ is $100(1-\alpha/2)\%$ quantile for a standard normal distribution. The number of samples required to accurately estimate γ_i is very large since the event of interest is very rare.

Under importance sampling, an alternative estimator is used to estimate γ_i such that the rare event is experienced more often. The probability $P(\omega)$ of observing the cycle ω is the product of one-step transition probabilities. A new

importance sampling distribution P' is defined such that $P(\omega) > 0 \implies P'(\omega) > 0$ and

$$\begin{aligned} \gamma_i &= \sum_{\omega \in \Omega} \phi(\omega) \frac{P(\omega)}{P'(\omega)} P'(\omega) \\ &= \sum_{\omega \in \Omega} \phi(\omega) L(\omega) P'(\omega) \end{aligned}$$

where the likelihood ratio $L(\omega)$ is the Radon-Nikodym derivative of P with respect to P'. The likelihood ratio $L(\omega)$ can be decomposed into a product of one-step transition *event likelihood ratios* associated with each individual event within the cycle. An unbiased estimator $\hat{\gamma}$ for γ can then be obtained by drawing N independent samples under the probability measure P' and multiplying the samples by the corresponding likelihood ratios. Thus,

$$\hat{\gamma}_i = \frac{1}{N} \sum_{j=1}^{N} \phi(\omega_j) L(\omega_j)$$

When $E_{P'}\left[\phi(\omega)^2 L(\omega)^2\right] < \infty$, a confidence interval for $\hat{\gamma}_i$ can be constructed as described earlier using the central limit theorem. The probability measure P' should be selected such that the variance of the estimator is minimized. In general, P' should be chosen such that $E_{P'}\left[\phi(\omega)^2 L(\omega)^2\right] < E_P\left[\phi(\omega)^2\right]$ to obtain variance reduction.

2.2 Comparing approaches

2.2.1 Asymptotic Properties

The asymptotic efficiency of an estimator can be quantified by considering the relative error of the estimator. Relative error is defined as the ratio of the standard deviation of the estimator over its expected value. Bounded relative error refers to the behavior of the estimator as the quantity to be estimated approaches zero which occurs by varying a rarity parameter for the system under study. In this model, the quantity of interest γ_i approaches zero as the buffer size is increased to infinity. An estimator yields bounded relative error if the relative error remains bounded as the quantity to be estimated approaches zero. This implies that, the sample size required to achieve a desired level of accuracy remains bounded in the limit, which is the best possible result.

An estimator is said to be asymptotically efficient if the relative error grows at a sub-exponential rate as the quantity to be estimated approaches zero. This means that the number of samples grows at a sub-exponential rate to achieve the desired accuracy. An estimator is said to have linearly bounded relative error when the number of samples required to achieve a fixed relative error increases linearly in buffer size B. For importance sampling estimators, bounded relative error implies asymptotic efficiency.

2.2.2 Variance Reduction Ratio

To compare the performance of two importance sampling estimators, we need to take into account variance reduction and the computational effort required to achieve that reduction. The variance reduction ratio (VRR) measures the trade-off between variance reduction and the associated computational cost. VRRs are computed by multiplying a ratio of the variances of two estimators by a ratio of the corresponding computational effort, i.e., simulation time or number of events sampled to generate that variance. Typically, VRRs are estimated empirically by simulation. If the VRR is less than one, then the approach in the numerator is more efficient and a VRR greater than one implies that the approach in the denominator is more efficient.

2.3 Balanced Likelihood Ratio Approaches

The proposed importance sampling method is based on the balanced likelihood ratio approach. This approach was originally proposed to estimate the reliability of fault-tolerant repairable systems (see Alexopoulos and Shultes 2001) and was later adapted to estimate system overflow probabilities in tandem-Jackson networks (see Shultes 2002). The importance sampling distribution for estimating the buffer overflow probability is based on the concept of controlling the event likelihood ratios within the cycles. A key feature of this approach is that likelihood ratios associated with cycles are forced to be bounded from above by one.

The application of the balanced likelihood ratio approach to estimate γ_0 and γ_1 proceeds as follows. Classify all system events into 2 classes: events that move the system towards buffer overflow and events that move the system away from buffer overflow. Arrival events and service completion events at the first node belong to the first category and service completion events at the second node fall into the second category. The balanced likelihood ratio method balances the event likelihood ratios associated with events from these two classes.

Every service completion event at the second node must be preceded by an arrival event and a service completion event at the first node. The product of these three event likelihood ratios can be forced to be one for all customers. This assignment causes likelihood ratios associated with cycles to be bounded below one. The proposed method has the following basic balanced likelihood ratio properties established by Shultes (2002).

- Every event that moves the system closer to the rare event (arrival and service completion at the first node) has one corresponding event (service completion at the second node) that effectively cancels out the events that moved the system closer to overflow.

- Events that would complete a cycle before the system experiences a rare event have zero probability in the importance sampling distribution.

- If the events that move the system closer to buffer overflow are forced to be more likely, then the corresponding future event which would move the

system away from overflow is forced to be less likely.

To summarize, each customer in the system experiences a series of events. Each event accumulates an event likelihood ratio. At any given time, the product of the event likelihood ratios accumulated for a customer is less than one. When the customer leaves the system, the product of the corresponding event likelihood ratios becomes one. The overall likelihood ratio of a cycle is obtained by multiplying the accumulated event likelihood ratios of all the customers in the system when the cycle ends. Since the accumulated likelihood ratio of all customers in the system is below one, the overall likelihood ratio of the cycle is bounded from above by one.

Chapter 3

Tandem Queues

Balanced likelihood ratio methods for estimating the probabilities γ_0 and γ_1 when the first buffer capacity is infinite and finite are described in Sections 3.1 and 3.2 respectively. The importance sampling distribution is the same for estimating both γ_0 and γ_1. However, the method for estimating γ_0 includes cases which do not occur while estimating γ_1, i.e., when the starting state is (1,1). Hence, without loss of generality, the importance sampling distributions are described for the starting state (0,0).

Customer arrival events and service completion events at the first node generate event likelihood ratios. These event likelihood ratios are used as multipliers for biasing the probability of service completion at the second node. Let $l_a(i)$ denote the i^{th} arrival event likelihood ratio and $l_s(i)$ denote the i^{th} first node service completion event likelihood ratio. The importance sampling distribution

is formed such that the content of the second buffer reaches the bound B in all cycles. The idea is to avoid paths which fail to experience the rare event within the cycle.

The proposed importance sampling distribution depends on the sample path for the process $\{Z(t),\, t \geq 0\}$. Importance sampling probabilities are time dependent, but at any time within the simulation only three importance sampling probabilities are relevant. Let λ' denote the importance sampling probability of an arrival event. Let μ_1' and μ_2' denote the importance sampling probabilities of service completion events at the first and second nodes respectively.

3.1 Infinite First Buffer

The importance sampling approach described in Section 2.2 is directly applied to the infinite first buffer case. There are four cases to consider: (1) The system is empty, (2) All customers are at the first node, (3) All customers are at the second node, and (4) Customers are at both nodes in the system.

Case 1: The system is empty. The next event is a customer arrival with probability one. The event likelihood ratio for this event is replaced by $l_a' = \lambda / (\lambda + \mu_2)$ in the implementation because an arrival event likelihood ratio of one does not allow the service completion probability associated with this arrival to be reduced. This initial likelihood ratio is used to bias the service completion probability of this first customer at node one. It is easy to show that this deviation

from the basic balanced likelihood ratio approach maintains established likelihood ratio properties.

Case 2: All customers in the system are at the first node, i.e., the system state is $(X(t), Y(t) = x, 0)$ for $t \geq 0$ and some $x \in \mathbb{N}$. In this case, the next event could be either a customer arrival or a service completion at the first node. Deviating from the original balanced likelihood ratio description, the importance sampling probability for a service completion event at the first node is reduced to increase the arrival probability. The importance sampling probabilities in this case are:

$$\begin{aligned} \mu_1' &= l_a(x) \left(\frac{\mu_1}{\lambda + \mu_1} \right), \text{ and} \\ \lambda' &= 1 - \mu_1'. \end{aligned}$$

Case 3: All customers in the system are at the second node, i.e., the system state is $(X(t), Y(t) = 0, y)$ for $t \geq 0$ and some $y \in \mathbb{N}$. In this case, the next event could be either a customer arrival or a service completion at the second node. The importance sampling probabilities when $y > 1$ are:

$$\begin{aligned} \mu_2' &= l_a(y)\ l_s(y) \left(\frac{\mu_2}{\lambda + \mu_2} \right), \text{ and} \\ \lambda' &= 1 - \mu_2'. \end{aligned}$$

The service completion event is not allowed when $y = 1$ if the rare event has not yet occurred within the cycle. In this latter case, the customer arrival probability is one.

Case 4: Customers in the system are at node one and node two, i.e., the system state is $(X(t), Y(t) = x, y)$ for $t \geq 0$ and some $x, y \in \mathbb{N}^2$. The importance sampling probabilities in this case when $y > 1$ derive from:

$$\mu_2' = l_a(x+y)\; l_s(y)\; \left(\frac{\mu_2}{\lambda + \mu_1 + \mu_2}\right).$$

The remaining probability $(1 - \mu_2')$ is split between the customer arrival event and service completion event at the first node based on the number of customers in the system.

Let ρ_s and ρ_a denote the fraction of the importance sampling probability $(1 - \mu_2')$ assigned to the service completion at the first node and the arrival event respectively. The importance sampling probabilities for the arrival event and the service completion at node one are:

$$\begin{aligned} \mu_1' &= \rho_s\,(1 - \mu_2')\,, \text{ and} \\ \lambda' &= \rho_a\,(1 - \mu_2')\,. \end{aligned}$$

When the system size is lesser than or equal to the bound B, the service completion probability at the first node is not biased except when the first server is

the bottleneck. When the first server is the bottleneck, the importance sampling distribution increases the service completion probability at the first node by allocating a fraction of $(1 - \mu_2')$ for this purpose depending on the state of the system. Thus, when $x + y \leq B$,

$$\rho_s = \begin{cases} \max\left\{0.5, \dfrac{\mu_1}{1-\mu_2'}\right\} & \text{if } \mu_1 < \mu_2 \\ \dfrac{\mu_1}{1-\mu_2'} & \text{if } \mu_1 \geq \mu_2 \end{cases}$$

When the system size is greater than the bound B, the importance sampling probabilities allocated to the arrival event and the service completion at node one are proportional to the respective rates λ and μ_1. Thus, when $x + y > B$, $\rho_s = \dfrac{\mu_1}{\lambda + \mu_1}$ and $\rho_a = 1 - \rho_s$.

3.1.1 Initial Event Likelihood Ratio

When the system is in state $(1, 0)$, the initial likelihood ratio $l_a' = \lambda / (\lambda + \mu_2)$ is obtained by looking ahead one stage. At this point, we need a value less than one, to bias the service completion probability of the customer at node one. The likelihood ratio of an arrival event, when the system moves from state $(0, 1)$ to $(1, 1)$ would be $l_a' = \lambda / (\lambda + \mu_2)$ (as explained in Case 3). This likelihood ratio is not required to bias the service probability when the system is in state $(1, 1)$ because μ_2' is set to zero in order to prevent the cycle from ending before the

rare event occurs. Hence, this event likelihood ratio can be used for l'_a when the system is in state $(1,0)$.

Remark 1 While estimating γ_1, the starting state of the system is (1,1). In this case, the arrival likelihood ratio for the first and second customer is one. So, when the system is in state $(0,2)$, the arrival likelihood ratio of one does not allow the second node service completion probability associated with this arrival to be reduced. Hence, in this case the value of $l_a(2)$ is replaced by $l'_a = \lambda/(\lambda + \mu_2)$ by looking ahead one stage when the system is in state $(0,2)$ as described earlier.

3.1.2 Implementation

Define two stacks: L_a for storing arrival event likelihood ratios and L_s for storing likelihood ratios for service completion events at the first node. Initially each stack contains one multiplier, $l'_a = \lambda/(\lambda + \mu_2)$ is on stack L_a and $l'_s = 0$ is on stack L_s where the 0 guarantees that the cycle does not end without observing a buffer overflow event. After each arrival event, the event likelihood ratio (λ/λ') is pushed onto stack L_a. After each service completion event at the second node, one likelihood ratio from each stack is removed. For each service completion event at the first node, the event likelihood ratio (μ_1/μ'_1) is pushed onto stack L_s if the system is in state (x,y) for some $x \in \mathbb{N}$, $y \in \mathbb{N}$ and a likelihood ratio is removed from stack L_a when the system state is $(x,0)$ for some $x \in \mathbb{N}$.

3.1.3 Asymptotic Behavior in Infinite First Buffer Case

The proposed balanced likelihood ratio method forces each cycle to visit the rare event. Hence, the likelihood ratio of a cycle $L(\omega)$, consists of event likelihood ratios computed up to the time there are B customers at the second node. The method also forces likelihood ratios for service completion events at second node to cancel the event likelihood ratios for the corresponding arrival and first node service completion events. Hence, the overall likelihood ratio for a cycle is the product of the likelihood ratios of arrival events and service completion events at the first node associated with the customers in the system when the rare event happens. $L(\omega)$ has the following form,

$$L(\omega) = \left(\prod_{i=1}^{x+B} l_a(i)\right) \left(\prod_{j=1}^{B} l_s(j)\right) \tag{3.1}$$

where $l_a(i)$ is the i^{th} arrival event likelihood ratio in L_a and $l_s(j)$ is the j^{th} first node service completion event likelihood ratio in L_s. For notational purposes, let $L_1(\omega)$ denote the first term $\left(\prod_{i=1}^{x+B} l_a(i)\right)$ and $L_2(\omega)$ denote the second term $\left(\prod_{j=1}^{B} l_s(j)\right)$. Note that the system state is (x, B) when the system hits the rare event. As described earlier in the implementation, the event likelihood ratios of arrival and service completion events at the first node are stored in two separate stacks L_a and L_s respectively. The number of likelihood ratios in the stack L_a is equal to the number of customers in the system $(x+B)$. The number of likelihood ratios in the stack L_s is equal to the number of customers at the second node (B).

Let $L_m \geq \max_{\omega} L(\omega)$ be a upper bound on the likelihood ratio of a cycle while estimating γ_1. From (3.1),

$$L_m \geq L_1(\omega)\, L_2(\omega) \quad \text{for all } \omega \in \Omega. \tag{3.2}$$

A value for L_m can be found by finding upper bounds on $L_1(\omega)$ and $L_2(\omega)$.

Lemma 1. *The product of first node service completion likelihood ratios* ($L_2(\omega)$) *is bounded from above by one.*

Proof. The maximum possible value for the term $L_2(\omega)$ can be obtained by determining the maximum possible likelihood ratios $l_s(j)$ for all $j = 1$ to B. By the construction of the proposed balanced likelihood ratio method, the maximum possible likelihood ratio for the service completion event at first node is bounded from above by one. Hence, $L_2(\omega) \leq 1$. □

Let $M(i) \geq l_a(i)$ denote the upper bound for the i^{th} arrival event likelihood ratio. Then, an upper bound for $L_1(\omega)$ can be obtained by $\left(\prod_{i=1}^{x+B} M(i)\right)$. Thus,

$$L_1(\omega) \leq \left(\prod_{i=1}^{x+B} M(i)\right) \tag{3.3}$$

Lemma 2. *When $\mu_1 \geq \mu_2$, an upper bound for the i^{th} arrival event likelihood*

ratio is,

$$M(i) = \begin{cases} \dfrac{\lambda}{\lambda + \mu_2 - \mu_2 M(i-1)} & \textit{when } i > 3 \\ \dfrac{\lambda\,(\lambda + \mu_2)}{\lambda^2 + \lambda\mu_2 + \mu_2^2} & \textit{when } i = 3 \end{cases}$$

Proof. The i^{th} arrival event likelihood ratio is generated in one of the following cases: (1) System state is $(0, i-1)$. (2) System state is $(i-2, 1)$. (3) System state is (x, y), where $y > 1$ and $x + y = i - 1$. Note that cases 1 and 3 are the same when $i = 3$.

Let a_1,a_2,a_3 denote the i^{th} arrival event likelihood ratio in cases 1, 2 and 3 respectively. The maximum possible i^{th} arrival event likelihood ratio is $M(i) = max(a_1, a_2, a_3)$.

Case 1: The original arrival event probability in this case is $\lambda/(\lambda + \mu_2)$. The importance sampling probability for this arrival is

$$1 \; - \; \frac{\mu_2}{\lambda + \mu_2}\, l_a(i-1)\, l_s(i-1)\,.$$

As described earlier in Lemma 1, the value of $l_s(j)$ for all j is bounded from above by one. $l_a(i-1)$ is the $(i-1)^{th}$ arrival event likelihood ratio in L_a. The arrival event likelihood ratio is the largest when the importance sampling probability for the arrival event is at its smallest value. So, in order to get the maximum value for a_1, $l_a(i-1)$ should be the maximum possible $(i-1)^{th}$ arrival event likelihood ratio

used in the simulation. The arrival event likelihood ratio used in the simulation for $l_a(2)$ is $l'_a = \lambda/(\lambda + \mu_2)$ (explained in Remark 1). This value remains the same throughout the simulation in all cycles for $l_a(2)$. Hence, when $i = 3$,

$$a_1 = \frac{\lambda/(\lambda + \mu_2)}{1 - \frac{\mu_2}{\lambda + \mu_2} l'_a}$$

and when $i > 3$

$$a_1 = \frac{\lambda/(\lambda + \mu_2)}{1 - \frac{\mu_2}{\lambda + \mu_2} M(i-1)}. \tag{3.4}$$

Case 2: The original arrival probability in this case is equal to $\lambda/(\lambda + \mu_1 + \mu_2)$. The importance sampling probability for arrival is equal to $1 - \mu_1 - \mu'_2$. Since there is only one customer at first node, $\mu'_2 = 0$. Thus, for all $i \geq 3$,

$$a_2 = \frac{\lambda/(\lambda + \mu_1 + \mu_2)}{1 - \mu_1}. \tag{3.5}$$

Case 3: The original arrival probability in this case is equal to $\lambda/(\lambda + \mu_1 + \mu_2)$. The importance sampling probability for an arrival event is equal to $1 - \mu_1 - \mu'_2$. The importance sampling probability for a service completion event at the second node, μ'_2 is

$$l_a(i-1)\ l_s(y-1)\ \left(\frac{\mu_2}{\lambda + \mu_1 + \mu_2}\right).$$

The value of $l_s(y-1)$ is substituted by its upper bound value of one. The value of $l_a(i-1)$ should be the maximum possible $(i-1)^{th}$ arrival event likelihood ratio used in the simulation in order to get the maximum value of a_3. Thus,

$$a_3 \;=\; \frac{\lambda/(\lambda + \mu_1 + \mu_2)}{1 \;-\; \mu_1 \;-\; \left(M(i-1)\,\dfrac{\mu_2}{\lambda+\mu_1+\mu_2}\right)}\,.$$

Rearranging the terms and using the fact that $\lambda+\mu_1+\mu_2 \;=\; 1$, we find that,

$$\text{when } i \;=\; 3,\; a_1 \;=\; \frac{\lambda}{\lambda+\mu_2-\mu_2(l'_a)} \quad \text{and when } i \;>\; 3,\; a_1 \;=\; a_3 \;=\; \frac{\lambda}{\lambda+\mu_2-\mu_2(M(i-1))}\,.$$

For all $i \geq 3$, $a_2 \;=\; \lambda/(\lambda+\mu_2)\,.$

Hence, when $\mu_1 \geq \mu_2$,

$$M(i) \;=\; max(a_1, a_2) \;=\; \frac{\lambda}{\lambda+\mu_2-\mu_2(l'_a)} \;=\; \frac{\lambda\,(\lambda+\mu_2)}{\lambda^2+\lambda\mu_2+\mu_2^2}, \quad \text{for } i = 3$$

and

$$M(i) \;=\; max(a_1, a_2, a_3) \;=\; \frac{\lambda}{\lambda+\mu_2-\mu_2 M(i-1)}, \quad \text{for } i \geq 3\;.$$

□

Remark 2 When $\mu_1 < \mu_2$, the maximum likelihood ratio M(i) can be found in a similar way. Specifically, the values of a_1 and a_2 in cases 1 and 2 are the same

as in (3.4) and (3.5) respectively. The form of a_3 changes when the first server is the bottleneck. This is because, the importance sampling probability of the first node service completion changes depending on the state of the system.

Lemma 3. *To get an upper bound on $L_1(\omega)$, the number of arrival event likelihood ratios should be B.*

Proof. The product of arrival event likelihood ratios is bounded by, $L_1(\omega) \leq \left(\prod_{i=1}^{x+B} M(i)\right)$.

As mentioned earlier, the number of arrival likelihood ratios in any cycle should be greater than or equal to the buffer size B. Consider the first case when the number of customers in the system is equal to the buffer size B when the cycle ends by hitting the rare event (system state is $X(t), Y(t) = 0, B$). In this case, $L_1(\omega)$ is bounded by $\prod_{i=1}^{B} M(i)$, where $M(i)$ is the maximum possible value for i^{th} arrival event likelihood ratio.

Consider the second case when the system contains more than B customers when the cycle ends. This means that $x > 0$ and the system is in state $(X(t), Y(t) = x, B)$. $L_1(\omega)$ in this second case is bounded by $\prod_{i=1}^{x+B} M(i)$. This is equal to

$$\left(\prod_{i=1}^{B} M(i)\right) \left(\prod_{k=B+1}^{x+B} M(k)\right).$$

By construction of the balanced likelihood ratio method, all arrival event likeli-

hood ratios are bounded from above by one. Thus, $\left(\prod_{k=B+1}^{x+B} M(k)\right)$ is bounded from above by one and implies that the bound for $L_1(\omega)$ in the second case becomes lesser by multiplying the term $\left(\prod_{k=B+1}^{x+B} M(k)\right)$. Hence, to get an upper bound on $L_1(\omega)$, the number of arrival event likelihood ratios should be equal to the buffer size B. □

Using Lemmas 2 and 3, when $\mu_1 \geq \mu_2$, the product of arrival event likelihood ratios $L_1(\omega)$ is bounded by

$$L_1(\omega) \leq \prod_{i=1}^{B} M(i)\,. \tag{3.6}$$

Lemma 4. *When $\mu_1 \geq \mu_2$, a upper bound for the maximum likelihood ratio of a cycle is*

$$L_m = \frac{(\lambda+\mu_2)\,(\lambda^{B-2})}{\sum_{k=0}^{B-1} \lambda^{B-1-k}\mu_2^k}\,.$$

Proof. Using Lemma 1, $L_2(\omega) \leq 1$.

From (3.6), $L_1(\omega) \leq \prod_{i=1}^{B} M(i)$. The arrival event likelihood ratio for the first and second customer is one. Hence, $M(i) = 1$ for $i = 1, 2$. Using Lemma 2, the maximum possible value for i^{th} arrival event likelihood ratio is,

$$M(i) = \frac{\lambda}{\lambda + \mu_2 - \mu_2(l_a(i-1))}, \quad i > 2.$$

Thus, $L_1(\omega)$ is,

$$(1)\,(1)\left(\frac{\lambda\,(\lambda+\mu_2)}{\lambda^2+\lambda\mu_2+\mu_2^2}\right).....\left(\frac{\lambda\sum\limits_{p=0}^{k}\lambda^{k-p}\,\mu_2^p}{\sum\limits_{p=0}^{k+1}\lambda^{k+1-p}\,\mu_2^p}\right).....\left(\frac{\lambda\sum\limits_{p=0}^{B-2}\lambda^{B-2-p}\,\mu_2^p}{\sum\limits_{p=0}^{B-1}\lambda^{B-1-p}\,\mu_2^p}\right).$$

Simplifying the terms,

$$L_1(\omega)\leq\prod_{i=1}^{B}M(i)=\frac{(\lambda+\mu_2)\,(\lambda^{B-2})}{\sum\limits_{k=0}^{B-1}\lambda^{B-1-k}\mu_2^k}\ .$$

Using the bounds for $L_1(\omega)$ and $L_2(\omega)$, we get

$$L_1(\omega)\,L_2(\omega)\ \leq\ (1)\left(\frac{(\lambda+\mu_2)\,(\lambda^{B-2})}{\sum\limits_{k=0}^{B-1}\lambda^{B-1-k}\mu_2^k}\right)\tag{3.7}$$

Using (3.2) and (3.7), yields the desired result. □

Theorem 1. *The proposed importance sampling distribution achieves bounded relative error when* $\mu_1\ \geq\ \mu_2$.

Proof. The relative error (RE) is defined as the ratio of the standard deviation of the estimator over its expected value. The maximum likelihood ratio of a cycle is an upper bound for the standard deviation of the estimator. Lemma 4 implies that, the standard deviation of the proposed estimator $\sigma\,(\hat{\gamma}_1)$ is bounded from

above by

$$L_m \;=\; \frac{(\lambda+\mu_2)\,(\lambda^{B-2})}{\sum_{k=0}^{B-1}\lambda^{B-1-k}\mu_2^k}\,.$$

The rare event of interest is an exponentially rare event, i.e, γ_1 has exponential decay rate. Specifically, Remark 3.5 in Kroese and Nicola (2002) states that γ_1 is proportional to e^{-sB} when the first node has infinite capacity. They have also proven that if $\mu_1 \geq \mu_2$, $\gamma_1 = d\,\eta^B$ where d is a positive constant and $\eta = \lambda/\mu_2$ (see *Lemma* A.5 and *Remark* 3.6 in Kroese and Nicola (2002)).

The relative error of the proposed BLR method satisfies,

$$RE \leq \frac{\left(\dfrac{(\lambda+\mu_2)\,\lambda^{B-2}}{\sum_{k=0}^{B-1}\lambda^{B-1-k}\mu_2^k}\right)}{d\,\left(\dfrac{\lambda}{\mu_2}\right)^B}\,.$$

Simplifying the right hand side leads to,

$$RE \leq \frac{(\lambda+\mu_2)}{d\;\lambda^2}\left(\frac{\mu_2^B}{\sum_{k=0}^{B-1}\lambda^{B-1-k}\mu_2^k}\right)\,.$$

Hence,

$$\lim_{B\to\infty} \frac{\sigma\left(\hat{\gamma}_1\right)}{\gamma_1} \leq \frac{(\lambda+\mu_2)}{d\ \lambda^2} \left(\frac{\mu_2^B}{\sum_{k=0}^{B-1} \lambda^{B-1-k}\mu_2^k} \right) \leq \infty.$$

Thus, when $\mu_1 \geq \mu_2$ the proposed importance sampling estimator has bounded relative error. □

3.2 Finite First Buffer

The balanced likelihood ratio method for estimating the probability of buffer overflow in the second node when the first buffer has finite capacity is described below. The approach is similar to the infinite first buffer case.

Assume the system starts from state $(0,0)$. The same four cases as in the infinite first buffer case are considered. For cases 1, 2 and 3, i.e., when the system is empty and when the system state is $(x,0)$ and $(0,y)$ for some $x, y \in \mathbb{N}^2$, the importance sampling distribution is the same as in the infinite first buffer case. When the system is in state (x,y) for some $x, y \in \mathbb{N}^2$, the importance sampling probabilities derive from the same starting point as before:

$$\mu_2' = l_a(x+y)\ l_s(y) \left(\frac{\mu_2}{\lambda+\mu_1+\mu_2} \right).$$

As before, the remaining probability $(1-\mu_2')$ is split between the customer arrival

event and the service completion event at the first node based on the number of customers in the system. Since the first node has a finite capacity b, the fraction ρ_s of the importance sampling probability $(1 - \mu_2')$ assigned to the service completion at node one is increased, relative to the infinite first buffer case, by a factor c which depends on the number of customers at the first node. However, if $\mu_1 > \mu_2$ then this modification is not necessary, so $c = 0$ in this special case. The importance sampling probabilities for customer arrival events and service completion at node one are:

$$\begin{aligned} \mu_1' &= (\rho_s + c)(1 - \mu_2'), \text{ and} \\ \lambda' &= 1 - \mu_1' - \mu_2', \end{aligned}$$

where ρ_s is defined as before and

$$c = \frac{x}{b}\left(\frac{\mu_1}{\lambda + \mu_1} - \rho_s\right).$$

The method can be implemented in the same way as that of the infinite first buffer case using two stacks: L_a for storing arrival event likelihood ratios and L_s for storing likelihood ratios of service completion events at first node.

Chapter 4

Numerical Results

Experimental results for four, two node tandem Jackson network examples are presented. In the first example, the second server is the bottleneck ($\mu_1 > \mu_2$), in the second and third examples the first server is the bottleneck ($\mu_1 < \mu_2$) and in the fourth example the service rates at the two nodes are equal. Results from experiments that estimate the probability that the contents of the second buffer reach the bound B before reaching zero starting from state $(1, 1)$ and $(0, 0)$ are presented for both finite and infinite first buffer cases. These cases come directly from Kroese and Nicola (2002). The rates in the tables can be normalized so that the normalized rates sum to one.

The result from each simulation experiment is based on 1,000,000 cycles. Cycles end when the second node experiences buffer overflow or when the second node empties. Each simulation run provides an estimate for the overflow prob-

ability (Mean), a 95% confidence interval halfwidth (Halfwidth) and the relative error (RE), i.e., standard deviation divided by mean. Computation times (CPU) are displayed in terms of average number of events per cycle. The tables include estimates of the overflow probabilities obtained by applying the exponential change of measure technique (K-N) presented by Kroese and Nicola (2002). The numerical values for these probabilities presented by Kroese and Nicola (2002) are also provided. The numerical values can be obtained by using the algorithm outlined in Garvels and Kroese (1999). The results from the two methods (BLR and exponential change of measure) are compared using Variance Reduction Ratios (VRRs). If the VRR is less than one, then the K-N method is more efficient and the BLR method is more efficient if the VRR is greater than one . All simulations were implemented in C and run on an HP C3600 workstation.

Tables 1-4 display the results for the estimates of the probability γ_1 for the infinite first buffer cases. Tables 5-8 display the results for the estimates of the probability γ_1 for cases where the first buffer is limited to nine customers. Tables 9 and 10 present the estimates of the probability γ_0 for all four examples for the infinite and finite first buffer cases respectively.

The relative error of the BLR method is bounded independent of the buffer size when the second server is the bottleneck in both finite and infinite buffer cases. For the infinite first buffer case, this is consistent with Theorem 1. In the other two cases, i.e., when the first server is the bottleneck and when the service rates at both nodes are equal, the relative error appears to be linearly bounded.

Based on the numerical results, the BLR method is more efficient than the K-N method when the buffer at the first node is infinite. In contrast, the K-N method is more efficient than the BLR method for B larger than 25 in the finite first buffer cases. This is not surprising given that the BLR relative errors are only linearly bounded in this case while the relative errors for the K-N method are bounded.

The BLR method yields similar results when used to estimate the overflow probabilities γ_0 and γ_1. The K-N method also yields similar results except when the first server is the bottleneck and its capacity is infinite in which case the relative error increases sharply with B. Kroese and Nicola (2002) have suggested that a different change of measure is needed in this case when the starting state is (0,0).

Table 4.1: Estimates of γ_1 in Example 1 $(\lambda, \mu_1, \mu_2 = 1, 4, 2)$ with $b = \infty$

Buffersize	Numerical	Method	Mean	Halfwidth	RE	CPU	VRR
20	1.43e-06	BLR	1.43e-06	1.26e-09	0.45e-03	85	3.8
		K-N	1.43e-06	3.20e-09	1.13e-03	51	—
25	4.47e-08	BLR	4.47e-08	3.95e-11	0.45e-03	110	3.8
		K-N	4.51e-08	1.00e-10	1.13e-03	65	—
50	1.33e-15	BLR	1.33e-15	1.18e-18	0.45e-03	235	3.8
		K-N	1.35-15	3.01e-18	1.13e-03	136	—
60	1.30e-18	BLR	1.30e-18	1.15e-21	0.45e-03	285	3.8
		K-N	1.33e-18	2.95e-21	1.13e-03	164	—
100	1.18e-30	BLR	1.18e-30	1.05e-33	0.45e-03	485	3.8
		K-N	1.22e-30	2.72e-33	1.13e-03	276	—

Table 4.2: Estimates of γ_1 in Example 2 $(\lambda, \mu_1, \mu_2 = 1, 2, 3)$ with $b = \infty$

Buffersize	Numerical	Method	Mean	Halfwidth	RE	CPU	VRR
20	2.05e-11	BLR	2.05e-11	5.24e-14	1.30e-03	70	9.2
		K-N	2.05e-11	1.97-13	4.89e-03	46	—
25	4.61e-14	BLR	4.61e-14	1.35e-16	1.49e-03	89	9.1
		K-N	4.63e-14	5.07e-16	5.59e-03	57	—
50	4.31e-27	BLR	4.30e-27	1.93e-29	2.29e-03	186	8.5
		K-N	4.28e-27	7.27e-29	8.66e-03	112	—
60	2.96e-32	BLR	2.96e-32	1.49e-34	2.57e-03	224	8.4
		K-N	2.94e-32	5.62e-34	9.76e-03	133	—
100	8.60e-53	BLR	8.58e-53	6.02e-55	3.58e-03	378	8.4
		K-N	8.49e-53	2.32e-54	13.8e-03	218	—

Table 4.3: Estimates of γ_1 in Example 3 $(\lambda, \mu_1, \mu_2 = 3, 4, 6)$ with $b = \infty$

Buffersize	Numerical	Method	Mean	Halfwidth	RE	CPU	VRR
20	1.35e-08	BLR	1.35e-08	3.54e-13	1.34e-03	97	6.6
		K-N	1.35e-08	1.38e-20	5.20e-03	42	—
25	1.97e-10	BLR	1.97e-10	5.95e-13	1.54e-03	125	6.4
		K-N	1.98e-10	2.33e-12	5.99e-03	52	—
50	2.20e-19	BLR	2.20e-19	1.03e-21	2.39e-03	264	6.1
		K-N	2.22e-19	4.13e-21	9.49e-03	101	—
60	6.54e-23	BLR	6.53e-23	3.46e-25	2.70e-03	320	6.0
		K-N	6.68e-23	1.39e-24	10.7e-03	120	—
100	6.79e-37	BLR	6.79e-37	5.08e-39	3.80e-03	541	5.8
		K-N	6.96e-37	2.05e-38	15.2e-03	194	—

Table 4.4: Estimates of γ_1 in Example 4 $(\lambda, \mu_1, \mu_2 = 1, 2, 2)$ with $b = \infty$

Buffersize	Numerical	Method	Mean	Halfwidth	RE	CPU	VRR
20	2.79e-07	BLR	2.79e-07	7.84e-10	1.43e-03	94	1.9
		K-N	2.78e-07	1.59e-09	2.90e-03	43	—
25	7.66e-09	BLR	7.68e-09	2.33e-11	1.55e-03	122	1.8
		K-N	7.67e-09	4.67e-11	3.10e-03	54	—
50	1.56e-16	BLR	1.56e-16	5.92e-19	1.93e-03	256	1.6
		K-N	1.56e-16	1.16e-18	3.79e-03	107	—
60	1.38e-19	BLR	1.38e-19	5.55e-22	2.04e-03	308	1.6
		K-N	1.39e-19	1.08e-21	3.99e-03	127	—
100	9.62e-32	BLR	9.60e-32	4.45e-34	2.39e-03	518	1.5
		K-N	9.58e-32	8.63e-34	4.59e-03	208	—

Table 4.5: Estimates of γ_1 in Example 1 $(\lambda, \mu_1, \mu_2 = 1, 4, 2)$ with $b = 9$

Buffersize	Numerical	Method	Mean	Halfwidth	RE	CPU	VRR
20	1.43e-06	BLR	1.43e-06	1.27e-09	0.45e-03	85	3.8
		K-N	1.43e-06	3.20e-09	1.13e-03	51	—
25	4.45e-08	BLR	4.45e-08	3.99e-11	0.45e-03	110	3.7
		K-N	4.48e-08	9.99e-11	1.13e-03	65	—
50	1.30e-15	BLR	1.30e-15	1.21e-18	0.47e-03	235	3.5
		K-N	1.32e-15	2.99e-18	1.13e-03	136	—
60	1.26e-18	BLR	1.26e-18	1.19e-21	0.48e-03	285	3.5
		K-N	1.29e-18	2.92e-21	1.13e-03	164	—
100	1.12e-30	BLR	1.11e-30	1.10e-33	0.49e-03	485	3.3
		K-N	1.12e-30	2.66e-33	1.17e-03	277	—

Table 4.6: Estimates of γ_1 in Example 2 $(\lambda, \mu_1, \mu_2 = 1, 2, 3)$ with $b = 9$

Buffersize	Numerical	Method	Mean	Halfwidth	RE	CPU	VRR
20	1.89e-11	BLR	1.88e-11	4.15e-14	1.12e-03	68	2.8
		K-N	1.87e-11	8.69e-14	2.37e-03	43	—
25	3.76e-14	BLR	3.76e-14	1.00e-16	1.37e-03	87	1.8
		K-N	3.76e-14	1.75e-16	2.37e-03	53	—
50	1.25e-27	BLR	1.25e-27	6.90e-30	2.83e-03	182	0.4
		K-N	1.25e-27	5.80e-30	2.37e-03	107	—
60	5.06e-33	BLR	5.00-33	7.48e-35	9.62e-03	221	0.1
		K-N	5.06e-33	2.35e-35	2.37e-03	128	—
100	1.38e-54	BLR	1.39-54	2.48e-56	9.12e-03	371	0.04
		K-N	1.37e-54	6.39e-57	2.37e-03	214	—

Table 4.7: Estimates of γ_1 in Example 3 $(\lambda, \mu_1, \mu_2 = 3, 4, 6)$ with $b = 9$

Buffersize	Numerical	Method	Mean	Halfwidth	RE	CPU	VRR
20	1.15e-08	BLR	1.15e-08	2.74e-11	1.21e-03	91	1.6
		K-N	1.15e-08	5.12e-11	2.27e-03	41	—
25	1.41e-10	BLR	1.40e-10	3.89e-13	1.41e-03	117	1.2
		K-N	1.41e-10	6.25e-13	2.27e-03	52	—
50	3.89e-20	BLR	3.88e-20	1.93e-22	2.54e-03	246	0.3
		K-N	3.88e-20	1.73e-22	2.27e-03	103	—
60	5.84e-24	BLR	5.85e-24	3.57e-26	3.11e-03	299	0.2
		K-N	5.89e-24	2.61e-26	2.27e-03	124	—
100	2.98e-39	BLR	2.99e-39	3.87e-41	6.61e-03	506	0.1
		K-N	2.98e-39	1.33e-41	2.27e-03	207	—

Table 4.8: Estimates of γ_1 in Example 4 $(\lambda, \mu_1, \mu_2 = 1, 2, 2)$ with $b = 9$

Buffersize	Numerical	Method	Mean	Halfwidth	RE	CPU	VRR
20	2.56e-07	BLR	2.56e-07	3.29e-10	0.65e-03	96	3.8
		K-N	2.55e-07	9.56e-10	1.91e-03	43	—
25	6.40e-09	BLR	6.40e-09	9.98e-12	0.79e-03	125	2.5
		K-N	6.42e-09	2.40e-11	1.91e-03	54	—
50	6.34e-17	BLR	6.34e-17	1.84e-19	1.48e-03	268	0.7
		K-N	6.33e-17	2.37e-19	1.91e-03	110	—
60	3.99e-20	BLR	3.99e-20	1.40e-22	1.79e-03	324	0.5
		K-N	3.99e-20	1.49e-22	1.91e-03	132	—
100	6.24e-33	BLR	6.25e-33	3.92e-35	3.20e-03	552	0.1
		K-N	6.21e-33	2.33e-35	1.91e-03	221	—

Table 4.9: Estimates of γ_0 with $b = \infty$

Buffersize	Example	Method	Mean	Halfwidth	RE	CPU	VRR
50	1	BLR	7.40e-16	1.40e-18	0.97e-03	230	1.0
		K-N	7.40e-16	2.05e-18	1.41e-03	112	—
	2	BLR	1.03e-26	9.80e-29	4.86e-03	188	18.4
		K-N	9.06e-27	5.54e-28	31.1e-03	108	—
	3	BLR	3.86e-19	3.10e-21	4.10e-03	267	10.4
		K-N	3.82e-19	1.65e-20	22.1e-03	98	—
	4	BLR	1.54e-16	5.97e-19	1.98e-03	256	4.0
		K-N	1.54e-16	1.93e-18	6.39e-03	98	—
100	1	BLR	6.56e-31	1.23e-33	0.96e-03	480	1.0
		K-N	6.56e-31	1.82e-33	1.41e-03	224	—
	2	BLR	2.39e-52	5.76e-54	12.3e-03	380	17.8
		K-N	2.22e-52	3.29e-53	75.7e-03	207	—
	3	BLR	1.23e-36	1.87e-38	7.72e-03	543	11.8
		K-N	1.23e-36	1.09e-37	45.4e-03	187	—
	4	BLR	9.51e-32	4.53e-34	2.43e-03	518	2.5
		K-N	9.59e-32	1.18e-33	6.28e-03	187	—

Table 4.10: Estimates of γ_0 with $b = 9$

Buffersize	Example	Method	Mean	Halfwidth	RE	CPU	VRR
50	1	BLR	7.23e-16	1.40e-18	0.99e-03	233	1.0
		K-N	7.35e-16	2.06e-18	1.43e-03	111	—
	2	BLR	1.96e-27	1.16e-29	3.02e-03	185	1.9
		K-N	1.96e-27	2.15e-29	5.61e-03	103	—
	3	BLR	5.59e-20	3.05e-22	2.79e-03	251	0.8
		K-N	5.64e-20	4.30e-22	3.89e-03	100	—
	4	BLR	5.89e-17	1.78e-19	1.54e-03	269	1.0
		K-N	5.86e-17	2.89e-19	2.52e-03	101	—
100	1	BLR	6.21e-31	1.23e-33	1.00e-03	483	1.0
		K-N	6.45e-31	1.83e-33	1.45e-03	225	—
	2	BLR	2.16e-54	3.93e-56	9.27e-03	374	0.2
		K-N	2.18e-54	2.41e-56	5.65e-03	204	—
	3	BLR	4.30e-39	6.10e-41	7.23e-03	512	0.1
		K-N	4.33e-39	3.29e-41	3.88e-03	199	—
	4	BLR	5.83e-33	3.86e-35	3.38e-03	554	0.3
		K-N	5.78e-33	2.85e-35	2.51e-03	200	—

Chapter 5

Conclusions

This paper presents a balanced likelihood ratio importance sampling approach for estimating the overflow probability of the second buffer in a two node tandem Jackson network. The proposed importance sampling distributions depend on the state of the system. The importance sampling estimator is asymptotically efficient with bounded relative error when the first buffer capacity is infinite except when the first server is the bottleneck. This has been proved formally and corroborated using numerical results. When the first server is the bottleneck, numerical results indicate that the relative error is linearly bounded in the buffer size. Empirical evidence indicates that the BLR method outperforms existing importance sampling distributions when the first node buffer is infinite. More work is needed to determine why the BLR method struggles when the first node buffer is finite. The proposed methods can be readily extended to estimate individual buffer over-

flow probabilities in tandem Jackson networks with more than two nodes. The proposed method can also be used to estimated buffer overflow probabilities in non-Markovian networks.

Alexopoulos, C. and B. C. Shultes. 2001. Estimating reliability measures for highly dependable systems, using balanced likelihood ratios. *IEEE Transactions on Reliability* 50 (3): 265-280.

Asmussen, S., and R.Y. Rubinstein. 1995. Steady state rare events simulation in queueing models and its complexity properties. In *Advances in Queueing: Theory, Methods and Open problems*, ed. J. H. Dhashalow, CRC Press, Boca Raton, Florida, 429-462.

De Boer, P. T., V. F. Nicola, and R. Y. Rubinstein. 2000. Adaptive Importance Sampling simulation of queueing networks. In *Proceedings of the 2000 Winter Simulation Conference*, ed. J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, 646-655.

De Boer, P. T., D. P. Kroese, and R. Y. Rubinstein. 2002. Estimating buffer overflows in three stages using cross-entropy. In *Proceedings of the 2002 Winter Simulation Conference*, ed. E. Yucesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, 301-309.

Glasserman, P. and S.-G. Kou. 1995. Analysis of an importance sampling estimator for tandem queues. *ACM Transactions on Modeling and Computer Simulation* 5 (1): 22-42.

Garvels, M. J. J., and D. P. Kroese. 1999. On the entrance distribution in RESTART simulation. In *Proceedings of the Second Workshop on Rare Event Simulation (RESIM 99)*, Enschede, The Netherlands, 65-88.

Glynn, P., and D. Iglehart. 1989. Importance sampling for stochastic simulations, *Management Science* 35 (4): 1367-1392

Heidelberger, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation* 5 (1): 43-85.

Juneja, S. 1993. Efficient rare event simulation of stochastic systems. Ph. D. Thesis, Department of Operations Research, Stanford University, Palo Alto, California.

Juneja, S. 2001. Importance Sampling and the cyclic approach *Operations Research* 49 (6): 900-912

Kroese, D. P. and V. F. Nicola. 2002. Efficient simulation of a tandem Jackson network. *ACM Transactions on Modeling and Computer Simulation* 12 (2): 119-141.

Kuruganti, I. and Strickland. 1997. Optimal importance sampling for Markovian systems with applications to tandem queues. *Mathematics and Computers in Simulation* 44: 61-79

Parekh, S. and J. Walrand. 1989. A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control* 34 (1): 54-66.

Serfozo, R. 1999. *Introduction to Stochastic Networks.* New York: Springer.

Shultes, B. C. 2002. A balanced likelihood ratio approach for analyzing rare events in a tandem Jackson network. In *Proceedings of the 2002 Winter Simulation Conference*, ed. E. Yucesan, C.-H. Chen, J. L. Snowdon and J. M. Charnes, 424-432.