Efficient Calculation of Rare Event Probabilities in Markovian Queueing Networks

Linar Mikeev
Saarland University,
Saarbrücken, Germany
mikeev@cs.uni-saarland.de

Werner Sandmann
Clausthal University of Technology,
Clausthal-Zellerfeld, Germany
werner.sandmann@tu-clausthal.de

Verena Wolf
Saarland University,
Saarbrücken, Germany
wolf@cs.uni-saarland.de

ABSTRACT

We address the computation of rare event probabilities in Markovian queueing networks with huge or possibly even infinite state spaces. For this purpose, we incorporate ideas from importance sampling simulations into a non-simulative numerical method that approximates transient probabilities based on a dynamical truncation of the state space. A change of measure technique is applied in order to accomplish a guided state space exploration. Numerical results for three different example networks demonstrate the efficiency and accuracy of our method.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: PROBABILITY AND STATISTICS—*Queueing theory,Markov Processes*

Keywords

Queueing networks, Markov chains, Rare events, Importance sampling, Transient analysis

1. INTRODUCTION

Many performance measures in modern computer and communication networks are associated with probabilities of unwanted events such as buffer overflows, packet losses, excessive backlogs, extreme delays, or blocking, all of which can significantly degrade the intended network service. To achieve an acceptably high level of performance they must be rare, that is, their probability must be very small. Typical probability thresholds according to common network standards are of the order 10^{-9} or below. Thus, network design and performance engineering requires efficient methods for calculating rare event probabilities.

Markovian queueing networks are well-established in performance evaluation of computer and communication networks [2, 18, 20, 31] but except for very special cases rare event analysis becomes cumbersome. Even in the absence of rare events and if product-form solutions are available, computing exact solutions is often impossible for complex networks. In such cases, approximate numerical analysis and stochastic simulation constitute two common complementary approaches to network performance evaluation.

For numerical analysis, Markovian queueing networks are usually mapped to an underlying continuous-time Markov chain (CTMC) described at the stochastic process level, uniquely defined by an initial probability distribution and a generator matrix. The state space of the underlying CTMC is multi-dimensional where the components describe the number of customers in the network nodes or (in case of phase-type distributed interarrival and service times) the exponential phases assigned to the nodes. But the size of the state space typically increases exponentially with the number of network nodes or model components, hence the model dimensionality. This effect is known as state space explosion and often causes models to be numerically intractable due to the prohibitively large state space.

Different strategies to tackle the state space explosion have been employed, mostly with regard to steady state analysis, that is computing unique stationary distributions of ergodic Markov chains by solving a system of linear equations. Transient analysis, which we consider in this paper, is far less often addressed and computing transient probabilities is typically more complicated since it requires the solution of the Kolmogorov differential equations (KDEs). For comprehensive treatments see, e.g., [2, 6, 28].

Stochastic simulation does not suffer from state space explosion because the state space need not be explicitly enumerated. But stochastic simulation constitutes an algorithmic statistical estimation procedure that tends to be computationally expensive and only provides an estimate whose reliability and accuracy in terms of relative error or confidence interval half width depend on the variance of the corresponding simulation estimator. Estimating rare event probabilities by straightforward direct simulation is not effective because rare events occur too infrequently to compute reliable statistical estimates in reasonable time. Therefore, reducing the variance of simulation estimators is a primary goal of rare event simulation [3, 24]. More precisely, efficient rare event simulation techniques, in particular importance sampling [13, 15, 17], aim at constructing alternative estimators with much smaller variances than standard estimators.

Importance sampling is a variance reduction technique that makes use of a change of measure. The probability distribution (measure) in the model is changed such that the rare event of interest becomes less rare. Then the simulation is conducted under the importance sampling measure and the systematically biased results are weighted by the likelihood ratio in order to yield unbiased estimates. However, importance sampling by no means guarantees variance reduction but may be even counterproductive and increase the variance. The efficiency of corresponding simulation schemes and the statistical accuracy of simulation results, usually expressed by asymptotic robustness properties of the underlying importance sampling estimators [1, 19, 27, 29], strongly depend on the choice of the change of measure. Despite considerable recent progress with regard to provably asymptotically efficient changes of measure for Jackson networks [10], their construction is still highly demanding and the development of efficient importance sampling simulations remains a difficult task. Moreover, even if asymptotically efficient schemes are available, other schemes may be better for specific probability ranges.

In this paper, we combine the basic idea of importance sampling with a numerical solution approach that overcomes the state space explosion by using a state space truncation. The underlying principle is a guided state space exploration where paths that contribute significantly to the rare event probability are not truncated. We use change of measure strategies suggested by importance sampling to identify these significant parts and “guide” the exploration of the state space in such a way that an accurate approximation of the rare event probability is obtained.

We present a numerical algorithm that approximates the solution of the KDEs by truncating large state spaces in an iterative fashion. At a particular time instant t, we consider an approximation of the transient distribution and assume that only a tractable number of states have a non-zero probability. States with a probability smaller than a threshold δ are neglected, that is, their probability is approximated as zero. The KDEs are then solved for a small time step h during which the truncated state space is adapted to the distribution at time $t+h$. More precisely, certain states that do not belong to the truncated space at time t are added at time $t+h$ because in between they receive a significant amount of probability which exceeds δ. Other states are neglected because their probability drops below δ. This idea has been previously used to approximate the transient distribution up to a total error of the order of 10^{-5} and less [8]. The smaller the significance threshold δ is chosen the more accurate the approximation becomes.

In the following, we first introduce a useful model description based on transition classes and review the formal framework of importance sampling in Section 2 and 3, respectively. Subsequently, we present our dynamical state space truncation procedure in Section 4.1 and then combine it in Section 4.2 with a change of measure strategy similarly as in the underlying concept of importance sampling, which yields the guided state space exploration. We substantiate the usefulness of our approach in Section 5 with experimental results of three queueing network examples and conclude our work in Section 6.

2. MODEL DESCRIPTION

For model description we employ a unified transition class formalism that applies to Markovian queueing networks as well as to other Markovian population models. It is motivated by the need to handle large state spaces and in particular by the transition structure of the underlying CTMC.

In almost all relevant cases, the transition structure is not arbitrary but state transitions correspond to certain events where similar events essentially have the same effect, e.g. arrivals, service completions, departures, or moves between nodes in queueing networks. Hence, they can be taken as specific discrete event systems [4], which provides a structured model description on an intermediate level of abstraction. Thus, a model description that reflects the event system character of the model is well suited, in particular for simulation purposes. For Markovian models the events need not be scheduled and the setting of Markovian event systems is also useful for numerical solution [14].

In order to describe a Markovian event system we have to define its state space and to specify all relevant events that may trigger state transitions. It is necessary to define under which conditions a certain event may occur, how it affects the system state and at which rate it occurs. Diverse formal specifications of Markovian event systems can be found in the literature. Here, we adopt the transition class formalism of [26]. Without loss of generality we assume that the state space is $\mathcal{S} \subseteq \mathbb{N}^d$. All events that trigger state transitions are classified according to their effects which yields transition classes. Formally, a transition class is a triplet $\mathcal{C} = (\mathcal{U}, u, \alpha)$ where $\mathcal{U} \subseteq \mathbb{N}^d$ is the source state space containing all states in which the event or the corresponding state transition, respectively, is possible, $u : \mathcal{U} \to \mathbb{N}^d$ is the update function giving the new state $u(x) \in \mathbb{N}^d$ according to the state transition when the event occurs in state $x \in \mathcal{U}$, and $\alpha : \mathcal{U} \to \mathbb{R}$ is the transition rate function giving the rate $\alpha(x) \in \mathbb{R}$ at which the event or transition occurs in state $x \in \mathcal{U}$. A particularly appealing feature is that very different systems can be cast in the same formalism. Any Markovian event system, in particular any Markovian queueing network, can be uniquely described by a set of such transition classes together with an initial distribution. The numbering of this set is arbitrary but often closely reflects the actual meaning of the transitions.

As a queueing network example consider a d-node tandem Jackson network with exponentially distributed service times where arrivals occur only at the first node according to a Poisson process with arrival rate λ. The service rates are denoted by $\mu_1, \dots, \mu_d$ and the buffer capacities (queue sizes) by $\nu_1, \dots, \nu_d$. Hence, the different types of transitions are arrivals at node 1, moves from node i to node $i+1$, $0 < i < d$ and departures from node d. Therefore, $d+1$ transition classes are sufficient:

$\mathcal{C}_1 = (\mathcal{U}_1, u_1, \alpha_1)$, where

- $\mathcal{U}_1 = \{(x_1, \dots, x_d) \in \mathbb{N}^d : x_1 < \nu_1\}$,
- $u_1(x) = (x_1 + 1, x_2, x_3, \dots, x_d)$,
- $\alpha_1(x) = \lambda$;

$\mathcal{C}_i = (\mathcal{U}_i, u_i, \alpha_i), i = 2, \dots, d$, where

- $\mathcal{U}_i = \{(x_1, \ldots, x_d) \in \mathbb{N}^d : x_{i-1} > 0, x_i < \nu_i\}$,
- $u_i(x) = (x_1, \ldots, x_{i-2}, x_{i-1} - 1, x_i + 1, x_{i+1}, \ldots, x_d)$,
- $\alpha_i(x) = \mu_{i-1}$;

$\mathcal{C}_{d+1} = (\mathcal{U}_{d+1}, u_{d+1}, \alpha_{d+1})$, where

- $\mathcal{U}_{d+1} = \{(x_1, \ldots, x_d) \in \mathbb{N}^d : x_d > 0\}$,
- $u_{d+1}(x) = (x_1, \ldots, x_{d-1}, x_d - 1)$,
- $\alpha_{d+1}(x) = \mu_d$.

State-dependent rates can be easily incorporated just by corresponding transition rate functions. The state space may be infinite due to one or more infinite buffers, which can be expressed explicitly by setting the buffer size to infinity or implicitly just by dropping the corresponding restrictions on the respective source state spaces. Phase-type distributed interarrival and service times can be modeled by properly defined transition classes for any change from one to the next phase.

Let $\{\mathcal{C}_1, \ldots, \mathcal{C}_m\}$ be a finite set of transition classes and let $\mathcal{C}_j = (\mathcal{U}_j, u_j, \alpha_j)$ for $j \in \{1, \ldots, m\}$. We assume that they define a CTMC $\{X^{(t)}, t \geq 0\}$. Note that certain regularity conditions are necessary to ensure that X is uniquely determined [16]. We denote the transient state probabilities for $x \in \mathcal{S}$ and $t \geq 0$ by $p^{(t)}(x) := P(X^{(t)} = x)$. In particular, $p^{(0)}$ is the initial distribution. Now, we have to express the distribution or density, respectively, of sample paths of X. Denote by $t_1 < t_2 < \ldots$ the successive time instants at which transitions occur and by $\mathcal{C}_{j_i}, j_i \in \{1, \ldots, m\}$ the transition class that applies at time t_i. Let $\tau_i := t_{i+1} - t_i$ be the time between the i-th and the $(i+1)$-th transition. Hence, state $x(t_i)$ is reached due to the i-th transition according to $\mathcal{C}_{j_i}$ at time t_i and remains unchanged for a sojourn time of τ_i after which the $(i+1)$-th transition according to $\mathcal{C}_{j_{i+1}}$ occurs at time t_{i+1} and changes the state to $x(t_{i+1})$. Thus, the time evolution of the system is completely described by the sequence of states and corresponding sojourn times. In compact form, $(x(t_0), \tau_0), (x(t_1), \tau_1), (x(t_2), \tau_2), \ldots$ describes a trajectory where $t_0 := 0$ and $\tau_0 = t_1$ is the sojourn time in the initial state $x(0)$.

For a trajectory up to the K-th transition, considering the Markovian property which in turn implies exponentially distributed sojourn times, the path density is given by

$$dP((x(t_0), \tau_0), \ldots, (x(t_K), \tau_K)) =$$
$$p^{(t_0)}(x(0)) \cdot \prod_{i=1}^{K} \alpha_{j_i}\left(x(t_{i-1})\right) \exp\left(-\alpha_0(x(t_{i-1}))\tau_{i-1}\right) \quad (1)$$

where $\alpha_0(x(t_{i-1})) := \alpha_1(x(t_{i-1})) + \cdots + \alpha_m(x(t_{i-1}))$ is the parameter (reciprocal mean) of the exponential sojourn time in state $x(t_{i-1}), i = 1, \ldots, K$. Note that for a given time horizon the number K of transitions is not known in advance and not deterministic. Formally, it is a random stopping time, which is in accordance with dP being a density of a probability measure P defined on the path space of the Markov process.

3. IMPORTANCE SAMPLING

Importance sampling aims at variance reduction for simulation estimators by a change of measure. The original system is simulated under a different probability measure and weighting by a correcting factor, the likelihood ratio, yields unbiased estimates. In a general measure theoretic setting, importance sampling is based on the Radon-Nikodym theorem, and all applications of importance sampling can be derived from this setting.

Consider two probability measures P and Q on a measurable space $(\Omega, \mathcal{A})$, where P is absolutely continuous with respect to Q, that is $\forall A \in \mathcal{A} : Q(A) = 0 \Rightarrow P(A) = 0$, or equivalently, $\forall A \in \mathcal{A} : P(A) > 0 \Rightarrow Q(A) > 0$. Then, the Radon-Nikodym theorem guarantees the existence of the Radon-Nikodym derivative $L = dP/dQ$, often also referred to as the likelihood ratio, and

$$\forall A \in \mathcal{A} : \; P(A) = \int_A L dQ. \quad (2)$$

Importance sampling basically exploits that expectations with respect to P are identical to expectations with respect to Q when weighting by the likelihood ratio. Hence, for random variables X on $(\Omega, \mathcal{A})$,

$$E_P[X] = \int X dP = \int X L dQ = E_Q[XL]. \quad (3)$$

The probability of an event $A \in \mathcal{A}$ can be expressed as a special case by $P(A) = E_P[I_A]$ where I_A denotes the indicator function of A.

For CTMCs, the relevant probability measures are path distributions and absolute continuity corresponds to the condition that all paths that are possible under the original measure must remain possible under the importance sampling measure. This can be obviously achieved by the condition that for all positive rates in the original model the corresponding rates under importance sampling are positive. Since we deal with CTMCs given in terms of transition classes as described in Section 2, we need an appropriate framework for the application of importance sampling to this model specification.

With importance sampling, the underlying probability measure determined by the transition rate functions is changed. Since the only requirement is absolute continuity of the probability measures involved, there is much freedom in how to change the measure. It is only necessary that all paths that are possible (have positive probability) under the original measure remain possible. This can be achieved by changing the original transition rate functions α_i to 'importance sampling transition rate function' β_i such that for all $i \in \{1, \ldots, m\}$ we have $\beta_i(x) = 0 \;\Rightarrow\; \alpha_i(x) = 0, \; x \in \mathcal{S}$, or equivalently, starting with the original propensity functions, $\alpha_i(x) > 0 \;\Rightarrow\; \beta_i(x) > 0, \; x \in \mathcal{S}$. One then generates trajectories according to the changed transition rate functions and multiplies the results with the likelihood ratio to get unbiased estimates for the original system. The trajectory generation now yields a sequence of states with associated sojourn times and reaction path density as in (1). If we set $\beta_0(x(t_{i-1})) := \beta_1(x(t_{i-1})) + \cdots + \beta_m(x(t_{i-1}))$ and keep the same initial distribution at time t_0 as for the original model, the likelihood ratio of a trajectory ω is

$$L(\omega) = \prod_{i=1}^{K} \frac{\alpha_{j_i}\left(x(t_{i-1})\right) \exp\left(-\alpha_0(x(t_{i-1}))\tau_{i-1}\right)}{\beta_{j_i}\left(x(t_{i-1})\right) \exp\left(-\beta_0(x(t_{i-1}))\tau_{i-1}\right)} \quad (4)$$

which can be efficiently computed during trajectory generation without much extra computational effort by successively updating its value after each simulated reaction according to the running product.

However, the results possess variances and are thus subject to statistical uncertainty since stochastic simulation yet with importance sampling still remains an estimation procedure. If the change of measure is chosen properly, it yields enormous variance reduction compared to direct simulation, but as a serious drawback of importance sampling a badly chosen change of measure can even lead to infinite variance increase. Moreover, in practice, the true probability as well as the unknown variance of the estimator must be estimated in course of the simulation and both are often significantly underestimated, which then leads to wrong conclusions and much too narrow confidence intervals that may even not contain the rare event probability of interest. Hence, also the reliability of importance sampling simulation results is extremely sensitive to the change of measure.

Nevertheless, the change of measure is an advantageous strategy for systematically increasing the probability of certain events and thus provides useful hints how to guide the system under study in order to provoke rare events of interest. We shall therefore exploit it in order to efficiently compute rare event probabilities without resorting to stochastic simulation, thereby avoiding both the statistical uncertainty inherent in simulation results and the danger of accidentally neglecting relevant parts of the state space as often the case with conventional state space truncation procedures.

4. NUMERICAL COMPUTATION OF RARE EVENT PROBABILITIES

In the sequel we stepwise develop our method for the computation of rare event probabilities. We first present a numerical algorithm that approximates the distribution based on a dynamical truncation of the state space. Then we combine it appropriately with the change of measure approach suggested by importance sampling. This combination, called guided state space exploration, constitutes a novel numerical method specifically designed for the computation of rare event probabilities.

4.1 Dynamical State Space Truncation

The dynamics of the CTMC $\{X^{(t)}, t \geq 0\}$ are given by the Kolmogorov differential equation (KDE)

$$\frac{d}{dt} p^{(t)}(x) = \mathcal{M}(p^{(t)})(x) \tag{5}$$

where the operator $\mathcal{M}$ is defined for any real-valued function $g : \mathbb{N}^d \to \mathbb{R}$ such that $\mathcal{M}(g)$ is the function that maps a state x to the value

$$\begin{aligned} \mathcal{M}(g)(x) \quad = \quad & \textstyle\sum_{j,y:x=u_j(y)\in\mathcal{U}_j} \alpha_j(y) \cdot p^{(t)}(y) \\ & -\textstyle\sum_{j:x\in\mathcal{U}_j} \alpha_j(x) \cdot p^{(t)}(x). \end{aligned}$$

The system of linear differential equations in (5) is typically large or even infinite such that its solution with standard numerical integration methods becomes computationally infeasible. If the variances of the state variables remain small, however, one can exploit that only a tractable number of states have "significant" probability, that is, only relatively few states have a probability that is greater than a small threshold. Here, we present a method based on our previous work [8] for efficiently approximating the solution of Eq. (5). Many transient solution approaches can be applied for this purpose (see, for instance, [8]). Here, we use an approximation based on numerical integration of Eq. (5) with an explicit fourth-order Runge-Kutta method.

The main idea of the approximation is to integrate only those differential equations in Eq. (5) that correspond to significant states. All other state probabilities are set to zero. This reduces the computational effort significantly since in each iteration step only a comparatively small subset of states is considered. Based on the fixed probability threshold $\delta > 0$, we dynamically decide which states to drop or add, respectively. Due to the regular structure of the CTMC the approximation error of the algorithm remains small since probability mass is usually concentrated at certain parts of the state space. The farther away a state is from a "significant set" the smaller is its probability. Thus, in most cases the total error of the approximation remains small. Since in each iteration step probability mass may be "lost" the approximation error at time t is the sum of all probability mass lost (provided that the numerical integration could be performed without any errors), that is,

$$1 - \sum_{x\in\mathcal{S}} \hat{p}^{(t)}(x)$$

where $\hat{p}^{(t)}$ is the approximation at time t.

The standard explicit fourth-order Runge-Kutta method applied to Eq. (5) yields the integration step

$$\begin{aligned} p^{(t+h)}(x) = p^{(t)}(x) + \tfrac{h}{6}\Big(& k^{(1)}(x) \\ & +2k^{(2)}(x)+2k^{(3)}(x) + k^{(4)}(x)\Big), \end{aligned} \tag{6}$$

where $h > 0$ is the time step of the method. For $i \in \{1, 2, 3, 4\}$ the values $k^{(i)}(x)$ are defined recursively as

$$\begin{aligned} k^{(1)}(x) &= \mathcal{M}(p^{(t)})(x), \\ k^{(2)}(x) &= k^{(1)}(x) + \tfrac{h}{2}\mathcal{M}(k^{(1)})(x), \\ k^{(3)}(x) &= k^{(1)}(x) + \tfrac{h}{2}\mathcal{M}(k^{(2)})(x), \\ k^{(4)}(x) &= k^{(1)}(x) + h\,\mathcal{M}(k^{(3)})(x). \end{aligned} \tag{7}$$

In order to avoid the explicit construction of a matrix and in order to work with a dynamic set *Sig* of significant states that changes in each step, we use for a state x a data structure with the following components:

- a field $x.p$ for the current probability of state x,
- fields $x.k_1, \ldots, x.k_4$ for the terms $k^{(1)}(x), \ldots, k^{(4)}(x)$,
- for all j with $x \in \mathcal{U}_j$ a pointer to the successor state $u_j(x)$ as well as the rate $\alpha_j(x)$.

We start at time $t = 0$ and initialize the set *Sig* as the set of all states that have initially a probability greater than δ, i.e. $Sig := \{x \mid p^{(0)}(x) > \delta\}$. We perform a step of the iteration in Eq. (6) by traversing the set *Sig* five times. In the first four rounds we compute $x.k_1, \ldots, x.k_4$ and in the final round we accumulate the summands. While processing state x in round i, $i < 5$, for each reaction j, we transfer probability

Table 1: A single iteration step of the fast RK4 algorithm, which approximates the solution of the KDE.

```
1  choose step size $h$;
2  for $i = 1, 2, 3, 4$ do //traverse Sig four times
3     //decide which fields from state data structure
4     //are needed for $k_i$
5     switch $i$
6        case $i = 1$: $coeff := 1$; $field := p$;
7        case $i \in \{2, 3\}$: $coeff := h/2$; $field := k_{i-1}$;
8        case $i = 4$: $coeff := h$; $field := k_{i-1}$;
9     for all $x \in Sig$ do
10       $x.k_i := x.k_i + x.k_1$;
11       for $j = 1, \dots, m$ with $x \in \mathcal{U}_j$ do
12          $x.k_i := x.k_i - coeff \cdot x.field \cdot \alpha_j(x)$;
13          if $u_j(x) \notin Sig$ then
14             $Sig := Sig \cup \{u_j(x)\}$;
15          $u_j(x).k_i := u_j(x).k_i + coeff \cdot x.field \cdot \alpha_j(x)$;
16 for all $x \in Sig$ do
17    $x.p := x.p + \frac{h}{6} \cdot + (x.k_1 + 2 \cdot x.k_2 + 2 \cdot x.k_3 + x.k_4)$;
18    $x.k_1 := 0; x.k_2 := 0; x.k_3 := 0; x.k_4 := 0$;
19    if $x.p < \delta$ then
20       $Sig := Sig \setminus \{x\}$;
```

mass from state x to its successor $u(x)$, by subtracting a term from $x.k_i$ and adding the same term to $u(x).k_i$. A single iteration step is given in pseudocode in Table 1. In line 20, we ensure that *Sig* does not contain states with a probability less than δ. As step size h in line 1 of the algorithm, we choose the smallest average sojourn time of all states in *Sig*, that is,

$$h = \min_{x \in Sig} 1 / \sum_{j=1}^{m} \alpha_j(x).$$

In lines 2-15 we compute the values $k^{(1)}(x), \dots, k^{(4)}(x)$ for all $x \in Sig$. The fifth round starts in line 16 and in line 17 the approximation of the probability $p^{(t+h)}(x)$ is calculated. Note that the fields $x.k_1, \dots, x.k_4$ are initialized with zero.

The performance of the algorithm can be further improved if we additionally check in line 13 whether it is worthwhile to add $u_j(x)$ to *Sig*, that is, we guarantee that $u_j(x)$ will receive enough probability mass and that $u_j(x)$ will not be removed in the same iteration due to the check in line 19. Thus, we add $u_j(x)$ only if the inflow $coeff \cdot x.field \cdot \alpha_j(x)$ to $u_j(x)$ is greater or equal than a certain threshold $\tilde{\delta} > 0$. Obviously, $u_j(x)$ may receive more probability mass from other states and the total inflow may be greater than $\tilde{\delta}$. Thus, if a state is not a member of *Sig* and if for each incoming transition the inflow probability is less than $\tilde{\delta}$, then this state will not be added to *Sig* even if the total inflow is greater or equal than $\tilde{\delta}$. This small modification yields a significant speed-up since otherwise all states that are reachable within at most four transitions will always be added to *Sig* because of line 13, but many of the newly added states will be removed in the same iteration because of line 19.

We list some experimental results for an 8-node tandem network in Table 2. We chose arrival rate $\lambda = 0.04$, service rates $\mu_1 = \ldots = \mu_8 = 0.12$ and computed the probability to reach a state where the second network node contains at least 29 customers (within a time horizon of $t = 100$). The accuracy of the approximation depends on the threshold δ (listed in the first column). Here, for ease of description we choose $\tilde{\delta} = \delta$. The total approximation error in the last column equals the probability mass that "got lost" during the computation because of the threshold δ. In the third column we list the average number of significant states, that is, the average size of the set *Sig*. The execution time is listed in the second column. For $\delta = 0$, the number of states that are considered is maximal since no states are neglected. The solution is most accurate at the cost of a long running time and high memory requirements. For small positive values of δ, we obtain accurate approximations based on a significantly smaller number of considered states. For larger systems, the analysis with very small values of δ (e.g. $\delta \leq 10^{-20}$) is often impossible because the size of *Sig* becomes intractable. For instance, in the 8-node tandem network above, if $\delta = 0$ it is impossible to compute the probability to reach a state where the second network node contains at least 30 customers since the memory requirements grow beyond 8 GB.

Table 2: Results for an 8-node tandem network.

δ	ex. time	$\lvert Sig \rvert$	error
0	2317s	36984860	-
1e-20	178s	1419901	1e-11
1e-15	37s	348056	4e-10
1e-10	3s	43343	6e-6

For many practical applications, the accuracy of the approximation is sufficient for a moderately small choice of the truncation thresholds δ and $\tilde{\delta}$, respectively. If, however, the probabilities of rare events have to be calculated, then the truncation approach above is no longer appropriate. As it stands now, the main drawback of the truncation approach is that rare events of interest may be neglected, that is, the truncated state space may not include those paths that lead to a certain rare event because their probability is smaller than the corresponding truncation threshold. If smaller truncation thresholds are chosen then the paths that

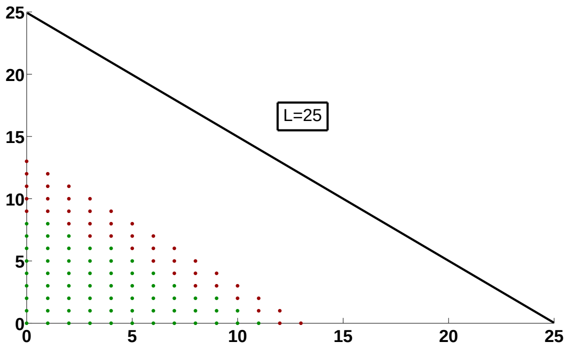

Figure 1: The event that the queues of a two-node tandem network contain L=25 customers is never reached.

significantly contribute to the rare event probability may not be truncated, but the number of states that have to be considered may become too large to be manageable.

We illustrate this problem for a two-node tandem network in Figure 1. The colored dots represent the states in the set Sig during different iteration steps. If we start with an empty system and approximate the probability distribution using threshold $\delta = 10^{-15}$, then the states where at least 25 customers are queuing are never reached because the paths that lead to these states have a probability smaller than δ. Note that after the system enters steady-state there are no significant changes in the set Sig. If we increase the time horizon, then eventually Sig will become smaller since in each step of the iteration probability mass is lost. If we choose smaller values for δ, then we reach $L = 25$ and we can accurately approximate the probability that at least 25 customers are present. While this is possible in the case of small examples, smaller values of δ render the solution intractable for more complex systems.

Note that other state space truncation approaches differ from ours in that they only generate the most probable states [7] or focus on calculating the steady-state distribution based on particular truncations [11, 30]. In contrast, our method relies on a truncation that is dynamically adapted in each step of the computation and we approximate the transient probability distribution. Moreover, for our truncation approach we never construct the infinitesimal generator matrix of the CTMC but use a structured transition class description to generate transition rates on-the-fly. The related approaches mentioned above are, however, similar in that they all avoid considering the entire state space and focus on the numerical solution of small subsets in order to overcome the state space explosion problem. But they do not appear promising in order to cope with rare event probabilities.

4.2 Guided State Space Exploration

In this section we propose an extension of the truncation approach presented in Section 4.1 that is based on ideas from importance sampling. Assume that we are interested in the probability $P(A)$ of a rare event A. Besides the CTMC X, we consider another CTMC $\{Y^{(t)}, t \geq 0\}$ that results from a change of measure in X, that is, if X has transition classes $\{\mathcal{C}_1, \ldots, \mathcal{C}_m\}$ with rate functions $\alpha_1, \ldots, \alpha_m$ then Y has the same transition classes except that each α_j is replaced by a rate function β_j. Here, absolute continuity is not necessary, i.e., we may choose β_j such that $\beta_j(x) = 0$ even if $\alpha_j(x) > 0$ for some x. We assume, however, that $\beta_1, \ldots, \beta_m$ are chosen such that the occurence of A is more likely.

The idea is to solve X and Y simultaneously using the dynamical state space truncation. Let $\hat{p}^{(t)}$ ($\hat{q}^{(t)}$) be the corresponding numerical approximation of the distribution of X (of Y) at time t, respectively. The algorithm for solving Y is exactly as in Section 4.1 whereas for X we slightly modify the dynamical state space truncation algorithm. The decision whether we remove a state x from the set Sig at time t depends only on $\hat{q}^{(t)}$ and not on $\hat{p}^{(t)}$. Thus, at all time instances t for both the solution of X and Y we use the same sets Sig. This ensures that we do not truncate the paths leading to the rare event A. Intuitively, Y shows the direction to the rare event. Therefore, we refer to this approach as guided state space exploration. If the change of measure is chosen appropriately, then the vectors $\hat{q}^{(t)}$ are computed using those paths that contribute most to $P(A)$. Hence, the vector $\hat{p}^{(t)}$ may loose lots of probability mass over time, i.e. $\sum_{x \in S} \hat{p}^{(t)}(x) \ll 1$. The probability mass that remains in $\hat{p}^{(t)}$ then contains those parts that contribute most to $P(A)$.

The guided state space exploration differs from the truncation algorithm in the following aspects:

- Instead of a single field $x.p$ for the current probability of state x we use two fields $x.p$ and $x.q$. The former refers to the current probability of state x in X and the latter refers to the probability of x in Y.
- In each iteration, we compute two different values for each field $x.k_i$, one for the probability flow in X and one for the flow in Y. Obviously, for Y we replace α_j by β_j (see line 12 of Table 1) and $x.p$ by $x.q$.
- We execute the two for-loops in lines 2-18 of Table 1 twice in order to compute $x.p$ and $x.q$, respectively. Lines 19 and 20, however, are only executed once in each iteration where we check whether $x.q < \delta$ (instead of $x.p < \delta$).

We remark that for the guided state space exploration the likelihood ratio is not needed to derive $P(A)$ from the probabilities $x.q$ of Y. Instead, $P(A)$ is directly approximated by the probabilities $x.p$ and the values $x.q$ are only used indirectly to determine the set of states that are considered in each step of the numerical integration. Actually, it would even be possible to solve Y and X not simultaneously but one after another. During the solution of Y, we would then record the elements of Sig for each time interval and use this information for the subsequent solution of X during which we truncate the state space in the same way as for Y. The simultaneous solution, however, has the advantage that it is faster than two subsequent solutions.

5. EXPERIMENTAL RESULTS

In this section, we present experimental results for specific rare event probabilities computed by the guided state space exploration described above. We consider certain buffer overflow probabilities corresponding to high numbers of customers in single nodes or in the overall system, respectively, for two variants of a two-node tandem Jackson network and an eight-node tandem Jackson network. The two-node networks are standard examples from the literature for which appropriate changes of measure have been studied extensively in the context of importance sampling. The eight-node network is a more complex example for which appropriate changes of measure for importance sampling have not been investigated yet.

In the sequel, we present results where we systematically vary the parameters determining the change of measure to study the sensitivity of our approach and consider different values for the truncation threshold δ.

5.1 Tandem Jackson Network

Our first example is a two-node tandem Jackson network with infinite buffers at both nodes, hence a special case of the

Table 3: Results for the two-node tandem network with parameters $\lambda = 0.04$, $\mu_1 = \mu_2 = 0.48$.

L	exact solution		guided state space exploration												
	Pr	$\|Sig\|$	δ	degenerated		0.48, 0.48, 0.04		0.6222,0.3333,0.0444		0.4,0.3,0.3		0.6,0.2,0.2		0.8,0.1,0.1	
				rel.err.	$\|Sig\|$	rel.err.	$\|Sig\|$	rel.err.	$\|Sig\|$	rel.err.	$\|Sig\|$	rel.err.	$\|Sig\|$	rel.err.	$\|Sig\|$
12	1.4693e-11	90	1e-20	3.461e-8	89	3.408e-14	58	2.012e-11	47	0	89	0	77	7.892e-8	39
			1e-15	4.881e-3	88	1.339e-9	46	4.589e-8	38	0	89	6.817e-15	60	3.171e-5	33
			1e-10	1	54	3.397e-5	34	8.166e-5	30	6.338e-9	89	3.632e-8	42	1.082e-2	25
25	2.8722e-25	350	1e-20	1	184	1.288e-9	235	3.777e-9	192	7.034e-15	343	7.034e-15	337	1.361e-4	155
			1e-15	1	103	1.257e-5	185	4.186e-6	155	1.403e-9	342	1.374e-9	269	3.519e-2	121
			1e-10	1	54	1.817e-3	132	2.488e-3	116	1.552e-4	337	4.987e-4	180	3.241e-1	75
50	6.0327e-52	1325	1e-20	1	184	5.349e-4	757	4.104e-6	650	2.746e-8	1259	1.245e-7	1145	2.225e-1	444
			1e-15	1	103	6.756e-2	591	1.149e-3	525	5.277e-4	1233	7.538e-3	877	3.960e-1	309
			1e-10	1	54	4.220e-1	398	9.555e-2	383	1.902e-1	1177	2.628e-1	524	6.209e-1	178

example described in Section 2 by transition classes where now $d = 2$ and $\nu_1 = \nu_2 = \infty$. The arrival and service rates are constant whereby we can skip the state dependence of the transition rate functions and concisely express them by a triplet $\alpha := (\alpha_1, \alpha_2, \alpha_3) := (\lambda, \mu_1, \mu_2)$. Similarly, we express the changed measure by $\beta = (\beta_1, \beta_2, \beta_3)$ where β_1 is the changed arrival rate and β_2, β_3 are the changed service rates at nodes 1 and 2, respectively. We are interested in the probability that the overall population in the system reaches a level L during a busy cycle, that is, the probability that starting with an arrival to the empty system the sum of the number of customers in both network nodes is at least L before the system empties again. Obviously, for low utilizations and/or high "target level" L, reaching L during a busy cycle is a rare event.

Though at a first glance seemingly simple, this example has received a lot of attention in the rare event simulation literature and has become a major reference example for judging change of measure strategies. This enormous interest was basically initiated by a change of measure proposed in [23] where the arrival rate and the smaller service rate (or, respectively, the service rate at the second node in case of equal service rates) are exchanged. Though initially supposed to be efficient, this change of measure has been subsequently proven in [12] to perform poorly in certain critical parameter (arrival and service rates) regions. A recent thorough analysis can be found in [5]. As the example has been so extensively studied in the context of importance sampling it enables us to demonstrate the advantages of our algorithm over importance sampling. For numerical analysis, we choose the parameter setting $\alpha = (0.04, 0.48, 0.48)$, hence arrival rate $\lambda = 0.04$ and service rates $\mu_1 = \mu_2 = 0.48$, which belongs to the critical parameter regions ascertained in [12]. We consider three different levels $L \in \{12, 25, 50\}$ and six different changes of measure $\beta^{(1)}, \ldots, \beta^{(6)}$ as follows.

First of all, we keep the original rates, that is, we do not at all apply a change of measure in this case, which we grasp as the "degenerated change of measure" $\beta^{(1)} = (0.04, 0.48, 0.48)$. The second change of measure is the interchange of the arrival rate and the service rate at the second node according to [23], hence $\beta^{(2)} = (0.48, 0.48, 0.04)$. The third one is $\beta^{(3)} = (0.6222, 0.3333, 0.0444)$ developed in [25] and shown to provide better results than $\beta^{(2)}$ when used for importance sampling. Furthermore, we consider $\beta^{(4)} = (0.4, 0.4, 0.3)$, $\beta^{(5)} = (0.6, 0.2, 0.2)$ and $\beta^{(6)} = (0.8, 0.1, 0.1)$, none of which developed for rare event simulation with importance sampling but rather ad hoc chosen by us. They are simply based on the intuitive reasoning that increasing the arrival rate and decreasing the service rate obviously guides the tandem network to a higher population level and thus increases the rare event probability of interest. Since these ad hoc changes of measure do not yield proper importance sampling simulation results, they are particularly well suited to highlight that our algorithm is far less sensitive to the change of measure than importance sampling and that in contrast to importance sampling our algorithm does not require intricate pre-analyses.

We list our experimental results in Table 3 where the first column contains the different values for L. For all parameters that we chose, the running time of our algorithm is less than one second. The second and the third column, respectively, contain numerical results obtained using the dynamical state space truncation outlined in Section 4.1 with truncation threshold $\delta = 0$. The rare event probabilities computed in this way (cf. column "Pr") are exact up to the numerical integration error. In the third column we list the average size of the set *Sig* during this exact computation, that is, the average number of states that were considered. For instance, there are 90 possible states if the number of customers is at most 12 and the algorithm quickly arrives at 90 states after starting with $|Sig| = 1$. The remaining part of the table shows the results of the guided state space exploration for different values of δ (listed in the fourth column). The six different changes of measures are indicated in the table by their respective parameter values except for the degenerated change of measure that is indicated by "degenerated". For each change of measure the column with heading "rel.err." lists the approximation error relative to the rare event probability.

For the degenerated change of measure, the guided state

Table 4: Results for the two-node tandem network with parameters $B = 100$ and $\lambda = 0.1$, $\mu_1 = 0.7$, $\mu_2 = 0.2$.

exact solution			guided state space exploration												
Pr	ex.time	$\|Sig\|$	δ	degenerated			0.2, 0.7,0.1			0.4,0.4,0.2			0.3,0.6,0.1		
				rel.err.	ex.time	$\|Sig\|$	rel.err.	ex.time	$\|Sig\|$	rel.err.	ex.time	$\|Sig\|$	rel.err.	ex.time	$\|Sig\|$
9.2034e-31	87s	36774	1e-20	1	1s	348	1.1419e-15	7s	1769	2.8028e-5	42s	8008	4.0291e-3	6s	1198
			1e-15	1	$< 1s$	198	1.2178e-10	4s	1022	2.8139e-3	22s	4561	5.3733e-2	3s	707
			1e-10	1	$< 1s$	93	8.9310e-6	2s	456	1.3226e-1	8s	1974	4.0188e-1	2s	325

space exploration is identical to the dynamical state space truncation presented in Section 4.1 (but in contrast to the exact solution with positive truncation threshold δ) and it yields accurate approximations only for $L = 12$ and $\delta \in \{10^{-15}, 10^{-20}\}$. Note that a relative error of one corresponds to approximating the rare event probability as zero. This is actually the same as what happens in direct simulation when due to the small probability the rare event is not observed and the probability of the non-observed event is estimated as zero. It is important that with our dynamical state space exploration the relative error is bounded by one since any error in the approximation is due to neglecting relevant states. In contrast, with importance sampling the relative error expressed in terms of the estimator's coefficient of variation or the relative half width of the corresponding confidence interval can become arbitrarily large in cases where the rare event has been observed but the estimator's variance is large (cf. [5, 12, 25]).

For $\beta^{(2)}$ according to [23] and $\beta^{(3)}$ according to [25] we can see that for all levels and all truncation thresholds the results provided by the guided state space exploration are very accurate. In [25] it is shown that with importance sampling $\beta^{(2)}$ performs very poor for high levels in that for $L = 25$ it yields results with a relative error of nearly 800%, whereas $\beta^{(3)}$ still yields statistically accurate results. For $L = 50$, neither $\beta^{(2)}$ nor $\beta^{(3)}$ yield useful results with importance sampling. Here, with the guided state space exploration both changes of measure perform extremely well even for $L = 50$ where $\beta^{(3)}$ is only slightly better $\beta^{(2)}$. For the latter case, the relative error is at most of the order of 10^{-4} if $\delta = 10^{-20}$ while it is at most of the order of 10^{-6} in the former case. Moreover, for $\beta^{(2)}$ the set Sig is slightly larger than for $\beta^{(3)}$. Hence, altogether both changes of measure perform quite similarly, which clearly indicates that the guided state space exploration is less sensitive to the change of measure, or, in other words, the impact of the specific change of measure on the reliability of the results is far lower.

For the ad hoc changes of measure $\beta^{(4)}$ and $\beta^{(5)}$ we observe that our results are again very accurate for small enough truncation threshold, though less accurate than the results for $\beta^{(2)}$ and $\beta^{(3)}$. However, only if δ is too large the relative error becomes high. Since neither of these changes of measure properly works with importance sampling, they are particularly well suited to further corroborate and highlight again that our algorithm is far less sensitive to the change of measure than importance sampling and that in contrast to importance sampling our algorithm does not require intricate pre-analyses.

Finally with the last case, $\beta^{(6)} = (0.8, 0.1, 0.1)$, we test a change of measure that has disastrous effects when used with importance sampling. It is beyond our scope here to provide a detailed analysis of this change of measure in the context of importance sampling but some explanation of the effects can be given by the notion of overbiasing. Overbiasing is a well known problem in importance sampling and basically means too much simulation acceleration. Although the goal is to provoke more of the rare events of interest in order to reduce the variance of the simulation estimator, it can be shown that provoking too many of them yields the contrary effect. More formally, overbiasing in importance sampling yields extremely small likelihood ratios for most of the corresponding simulation runs but very large likelihood ratios for a few simulation runs. This results in an enormous variance of the importance sampling estimators since this variance is mainly driven by the variance of the likelihood ratio. This effect is actually one of the main causes for the extreme sensitivity of importance sampling to the change of measure.

As we can see from our results in Table 3 the overbiasing effects seems to play a role also for the guided state space exploration but it is much less serious than for importance sampling. Our algorithm, though less efficient than for other changes of measure, still provides proper results for reasonably small truncation thresholds. Once again we get evidence that our algorithm is far less sensitive to the change of measure than importance sampling.

In Table 4, we list further results for the two-node tandem network but other than before we now do not consider the level of the overall population but the probability that the second queue reaches some high buffer content B. For the original network, we consider the parameter combinations studied in [22], namely $\lambda = 0.1$, $\mu_1 = 0.7$, $\mu_2 = 0.2$, and we compute the probability that the second queue contains at least $B = 100$ customers. In contrast to the previous type of rare event probabilities as presented in Table 3, we solve an infinite system without an indirectly given bound since now the number of customers in the first queue is not any longer implicitly bounded by a target level of the overall population. In addition to the relative error and the average size of Sig, we list the execution time of our method (cf. columns "ex.time"). The exact solution takes 87 seconds while a solution with the change of measure $(0.2, 0.7, 0.1)$ takes only a few seconds. The degenerated change of measure yields a relative error of one while the other heuristically chosen changes, $(0.4, 0.4, 0.2)$ and $(0.3, 0.6, 0.1)$, yield good results except if δ is chosen too large.

Table 5: Results for the two-node tandem network with slow-down and parameters $B = 100$, $\theta = 0.8$. Parameter set 1: $\lambda = 0.1$, $\mu_1 = 0.7$, $\mu_2 = 0.2$, $\nu_1 = 0.3$, parameter set 2: $\lambda = 0.1$, $\mu_1 = 0.7$, $\mu_2 = 0.2$, $\nu_1 = 0.15$.

pcom	exact solution			guided state space exploration												
	Pr	ex.time	$\|Sig\|$	δ	degenerated			as suggested in [22]			0.4,0.4,0.2,ν_1			0.3,0.6,0.1,ν_1		
					rel.err.	ex.time	$\|Sig\|$	rel.err.	ex.time	$\|Sig\|$	rel.err.	ex.time	$\|Sig\|$	rel.err.	ex.time	$\|Sig\|$
1	5.6009e-31	142s	54049	1e-20	1	$< 1s$	348	1.8764e-15	14s	2534	3.9418e-5	81s	13730	4.1024e-3	11s	1886
				1e-15	1	$< 1s$	198	1.0977e-10	8s	1444	3.8105e-3	40s	7009	5.4456e-2	6s	1089
				1e-10	1	$< 1s$	93	7.9622e-6	4s	631	1.6369e-1	14s	2699	4.0330e-1	3s	497
2	3.5471e-32	172s	71841	1e-20	1	$< 1s$	344	2.7777e-15	32s	5164	3.1021e-3	228s	30590	1.1562e-2	70s	10364
				1e-15	1	$< 1s$	197	2.6431e-10	21s	3464	1.2373e-1	150s	20777	1.2141e-1	44s	6590
				1e-10	1	$< 1s$	93	4.9194e-5	11s	1851	9.0224e-1	81s	11450	6.2285e-1	17s	2480

5.2 Tandem Jackson Network with Slow-down

Our next example network is a slight modification of the previous one, the two-node tandem Jackson network with server slow-down as considered in, e.g., [9, 22, 21]. When the length of the second queue reaches $B \cdot \theta$, the service rate of the first node changes from μ_1 to ν_1. Similar to the example above, the network with slow-down has become a reference example for judging change of measure strategies for systems with state-dependent rates in the original system. Here, we consider two different parameter combinations from [22]. The first combination is $(\lambda, \mu_1, \mu_2, \nu_1) = (0.1, 0.7, 0.2, 0.3)$ and the second one is $(\lambda, \mu_1, \mu_2, \nu_1) = (0.1, 0.7, 0.2, 0.15)$ where λ is the arrival rate, μ_2 is the service rate at the second node, and μ_1 and ν_1 are the service rates at the first node before and after slow-down. We compute the probability that starting with an arrival to the empty system the number of customers in the second network node is at least B before the system empties again.

We list results for both cases in Table 5 where the column "pcom" refers to the two parameter combinations. Also, in both cases we let $B = 100$ and $\theta = 0.8$. Similar as above, we consider an exact solution ($\delta = 0$) and the degenerated change of measure. The changes of measure suggested in [22] are such that, for pcom 1, arrival rate μ_2 is chosen, service rate at the second node is λ, and the service rate at the first node is chosen as μ_1 while the queue length is below $B \cdot \theta$ and ν_1 if the queue length is at least $B \cdot \theta$ (cf. column "as suggested in [22]"). For pcom 2, the parameters of the change of measure are also chosen as (μ_2, μ_1, λ) while the queue length at the first queue is below $B \cdot \theta$. If the queue length reaches $B \cdot \theta$, the parameters of the change of measure are calculated by solving an equation (see [22]) which yields $(0.177263, 0.177263, 0.095474)$. Again, we list the results of this change of measure in the column with heading "as suggested in [22]". Finally, we consider for both parameter combinations two heuristically chosen changes of measure, $(0.4, 0.4, 0.2, \nu_1)$ and $(0.3, 0.6, 0.1, \nu_1)$ where $\nu_1 = 0.3$ for pcom 1 and $\nu_1 = 0.15$ for pcom 2. Similar as for the network without slowdown, the exact solution takes significantly longer than the solutions based on the dynamical truncation of the state space. Moreover, the relative error is one for the degenerated case and high if δ is large and the change of measure is chosen heuristically. The change of measure suggested in [22] performs well even if δ is large.

5.3 Eight-Node Tandem Network

Our final example is a tandem network with eight nodes. We choose arrival rate $\lambda = 0.04$, and equal service rates $\mu_1 = \ldots = \mu_8 = 0.12$ and consider the probability that starting with an arrival to the empty system the sum of the number of customers in all network nodes is at least $L = 29$ before time T. Hence, as in our very first example we are concerned with the overall population in the system but we restrict our analysis to the time interval $[0, T]$ where $T = 50$ or $T = 100$ (cf. column "T"). To the best of our knowledge, no in-depth study of this system has been conducted in the context of importance sampling. As it is more complex than the two-node networks, of course, an appropriate change of measure is more difficult to obtain. Here, we consider heuristic changes of measure that seem reasonable with regard to the goal of provoking more rare events of interest. As before, we start by considering the degenerated change of measure where no rates are changed and the computation is done using the parameters of the original network. For the non-degenerated changes of measure we first apply a quite straightforward generalization of exchanging the arrival rate with the service rate at the last node, hence in our case now an exchange of λ and μ_8. Then we consider two ad hoc heuristics based on the reasoning that simply increasing the arrival rate without changing any service rate increases the probability of a high overall population. More precisely, we increase the arrival rate by a factor of two and by a factor of three, respectively. The results obtained by the guided state space exploration are given in Table 6. As we can see once more, for sufficiently small truncation threshold δ our algorithm provides very accurate results with quite low computational efforts.

6. CONCLUSIONS

In this paper, we have presented an efficient method for calculating rare event probabilities in Markovian queueing networks with huge or infinite state space. Our methods combines a dynamical state space truncation procedure with the change of measure idea of importance sampling. This combination yields a guided state space exploration where the change of measure is applied in order to guide the analysis algorithm to the relevant parts of the state space and to avoid truncation of important states. For the approximation of the transient distribution, we used an explicit fourth-order Runge-Kutta method. However, our approach can also be combined with other transient solution methods, e.g. with

Table 6: Results for the eight-node tandem network with parameters $\lambda = 0.04, \mu_1 = \cdots = \mu_8 = 0.12,\ L = 29$.

T	exact solution			guided state space exploration												
	Pr	ex.time	$\lvert Sig \rvert$	δ	degenerated			λ and μ_8 exchanged			λ two times faster			λ three times faster		
					rel.err.	ex.time	$\lvert Sig \rvert$	rel.err.	ex.time	$\lvert Sig \rvert$	rel.err.	ex.time	$\lvert Sig \rvert$	rel.err.	ex.time	$\lvert Sig \rvert$
50	1.6643e-23	796s	36984860	1e-20	1	13s	268689	5.9530e-5	106s	1514580	3.5570e-3	577s	3450157	3.3110e-4	99s	1257158
				1e-15	1	3s	72524	3.3907e-1	34s	479171	3.5541e-3	201s	1138686	6.7326e-1	29s	374744
				1e-10	1	$< 1s$	11161	1	6s	84349	3.4976e-3	30s	165803	1	5s	60968
100	1.2204e-16	2317s	36984860	1e-20	1	178s	1419901	3.8544e-9	1625s	9163634	1.3029e-4	1995s	4757788	1.6235e-5	1451s	7418540
				1e-15	1	37s	348056	1.8114e-4	729s	3943370	2.8527e-1	452s	1075486	5.3342e-4	581s	2833833
				1e-10	1	3s	43343	9.5890e-1	136s	684309	9.9755e-1	41s	110010	9.9189e-1	95s	458195

the uniformization method [28].

Our method has the general advantages of numerical methods over simulative approaches, namely, that it does not require the generation of trajectories and has only a numerical error but no statistical error. Moreover, our experimental results show that it is far less sensitive to the change of measure than importance sampling, that is, even if the change of measure is not well enough chosen for use with importance sampling, our numerical method performs well. The accuracy of our method is controlled by the truncation threshold δ. Obviously, as the truncation threshold $\delta \rightarrow 0$ the approximation becomes exact if we neglect the error introduced by the numerical integration method. For too large values of δ, say, $\delta \geq 10^{-15}$ the accuracy of the approximation can be degraded by a badly chosen change of measure, but even in this case the degradation is far less extreme than for importance sampling. A sufficiently small truncation threshold yields accurate results. Thus, determining an appropriate change of measure becomes easier as in the case of importance sampling or, in other words, the change of measure is less important than for importance sampling. Hence, our method provides an efficient means for numerical analysis of transient rare event probabilities.

At the current stage, exact formulas for the approximation error of the guided state space exploration are not yet available. As usual when inventing and presenting a novel computational method, we have demonstrated the accuracy by comparison with exact results. Clearly, it would be highly desirable to be able to compute a priori error bounds for given changes of measure and truncation thresholds, to determine the required truncation threshold for a prescribed maximum relative error, or to compute a posteriori the relative error for results obtained with a certain change of measure and a priori fixed truncation threshold. These issues are major topics of ongoing further research.

7. ACKNOWLEDGMENTS

L. Mikeev and V. Wolf have been partially funded by the German Research Council (DFG) as part of the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University and the Transregional Collaborative Research Center "Automatic Verification and Analysis of Complex Systems" (SFB/TR 14 AVACS).

8. REFERENCES

[1] J. H. Blanchet, P. W. Glynn, P. L'Ecuyer, W. Sandmann, and B. Tuffin. Asymptotic robustness of estimators in rare-event simulation. In *Proceedings of the 2007 INFORMS Simulation Society Research Workshop*, 2007.

[2] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi. *Queueing Networks and Markov Chains.* John Wiley & Sons, 2nd edition, 2006.

[3] J. A. Bucklew. *Introduction to Rare Event Simulation.* Springer-Verlag, 2004.

[4] C. G. Cassandras and S. Lafortune. *Introduction to Discrete Event Systems.* Springer-Verlag, 2nd edition, 2008.

[5] P. T. de Boer. Analysis of state-independent importance-sampling measures for the two-node tandem queue. *ACM Transactions on Modeling and Computer Simulation*, 16(3):225–250, 2006.

[6] E. de Souza e Silva and R. Gail. Transient solutions for Markov chains. In W. K. Grassmann, editor, *Computational Probability*, chapter 3, pages 43–79. Kluwer Academic Publishers, 2000.

[7] E. de Souza e Silva and P. M. Ochoa. State space exploration in Markov models. In *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pages 152–166. ACM, 1992.

[8] F. Didier, T. A. Henzinger, M. Mateescu, and V. Wolf. Fast adaptive uniformization of the chemical master equation. In *Proceedings of the High Performance Computational Systems Biology Workshop*, pages 118–127. IEEE Computer Society, 2009.

[9] P. Dupuis, K. Leder, and H. Wang. Large deviations and importance sampling for a tandem network with server slow-down. *Queueing Systems*, 57(2–3):71–83, 2007.

[10] P. Dupuis and H. Wang. Importance sampling for Jackson networks. *Queueing Systems*, 62(1–2):113–157, 2009.

[11] S. Gibson and E. Seneta. Augmented truncation of infinite stochastic matrices. *Journal of Applied Probability*, 24:600–608, 1987.

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

[13] P. W. Glynn and D. L. Iglehart. Importance sampling for stochastic simulations. *Management Science*,

35(11):1367–1392, 1989.

[14] W. K. Grassmann. Finding transient solutions in Markovian event systems through randomization. In W. J. Stewart, editor, *Numerical Solution of Markov Chains*, chapter 18, pages 357–372. Marcel Dekker, Inc., 1991.

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

[16] T. A. Henzinger, B. Jobstmann, and V. Wolf. Formalisms for specifying Markovian population models. In *Proceedings of the 3rd International Workshop on Reachability Problems*, volume 5797 of *LNCS*. Springer, 2009.

[17] S. Juneja and P. Shahabuddin. Rare event simulation techniques: An introduction and recent advances. In S. G. Henderson and B. L. Nelson, editors, *Simulation*, Handbooks in Operations Research and Management Science, chapter 11, pages 291–350. Elsevier, Amsterdam, The Netherlands, 2006.

[18] G. Kesidis. *An Introduction to Communication Network Analysis*. John Wiley & Sons, 2007.

[19] P. L'Ecuyer, J. H. Blanchet, B. Tuffin, and P. W. Glynn. Asymptotic robustness of estimators in rare-event simulation. *ACM Transactions on Modeling and Computer Simulation*, 20(1):6:1–4:41, 2010.

[20] D. A. Menasce, V. A. Almeida, and L. W. Dowdy. *Performance by Design*. Prentice Hall, 2004.

[21] D. I. Miretskiy. *Queueing Networks: Rare Events and Fast Simulations*. PhD thesis, University of Twente, Enschede, The Netherlands, 2009.

[22] D. I. Miretskiy, W. R. W. Scheinhardt, and M. R. M. Mandjes. Efficient simulation of a tandem queue with server slow-down. *Simulation*, 83(11):751–767, 2007.

[23] S. Parekh and J. Walrand. A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control*, 34:54–56, 1989.

[24] G. Rubino and B. Tuffin, editors. *Rare Event Simulation Using Monte Carlo Methods*. John Wiley & Sons, 2009.

[25] W. Sandmann. Fast simulation of excessive population size in tandem Jackson networks. In *Proceedings of 12th IEEE Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, MASCOTS*, pages 347–354. IEEE Computer Society Press, 2004.

[26] W. Sandmann. Structured description of Markovian network models and its potentials for efficient rare event simulation. In *Proceedings of the 2nd International Conference on Performance Modelling and Evaluation of Heterogeneous Networks, HetNets*, pages P39/1–10, 2004.

[27] W. Sandmann. Efficiency of importance sampling estimators. *Journal of Simulation*, 1(2):137–145, 2007.

[28] W. J. Stewart. *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, 1994.

[29] B. Tuffin, P. L'Ecuyer, and W. Sandmann. Robustness properties for simulations of highly reliable systems. In *Proceedings of the 6th International Workshop on Rare Event Simulation, RESIM*, pages 107–118, 2006.

[30] R. L. Tweedie. Truncation approximation of invariant measures for Markov chains. *Journal of Applied Probability*, 35(3):517–536, 1998.

[31] P. Van Mieghem. *Performance Analysis of Communications Networks and Systems*. Cambridge University Press, 2006.