Fast Simulation of Discrete Time Queues with Markov Modulated Batch Arrivals and Batch Departures

Ad Ridder

Department of Econometrics, Vrije Universiteit Amsterdam

de Boelelaan 1105, 1081 HV Amsterdam, the Netherlands

e-mail: aridder@econ.vu.nl

Abstract

In this paper we consider a slotted queueing model with finite buffers and correlated input. Arrivals are generated by different sources and offered in batches at the end of each slot. The batch sizes depend on the states of the sources which change their states according to Markov schemes. In case that the buffer reaches its limits, the excess of the buffer is lost. At the end of a slot a random batch is removed from the buffer. We prove a large deviations result for the loss probability in this model, and apply importance sampling for estimation of this probability.

1 Introduction

In this paper we consider an application of importance sampling to perform rare event simulation in a queueing model. This topic has been studied and reported in many papers and applications over the last 10 years, see [2] and [7] for surveys. For long, simulation studies are used heavily in the performance analysis of telecommunication systems. With the increase of error sensitive data transfer and with the technology improvements, these systems become more and more error free. For instance, one guarantees errors—loss probabilities—in the order of 10^{-9}. This means that the simulations should be executed efficiently because otherwise they take too long to produce reliable estimates. The importance sampling technique is such a method.

In the importance sampling method the underlying probability measure to run the simulations has been changed in such a manner that rare events occur more frequently. A likelihood ratio takes care of getting the correct estimates of the parameters of the original model. So, in applications of this method, one searches for 'good', or 'optimal' new measures. When applied to queueing models, it is now known that under some conditions on the distributions of the arrival and departure processes, large deviations is a useful tool for this quest [2, 4, 7]. In queueing models typical rare events are high level crossing and custom loss in finite (but large) buffer models. If there is a large deviations result of the level crossing probabilities, it produces an exponential rate factor with which the probabilities decay to zero. This factor is used to exponentially twist (or tilt) the original probability measure.

Thus, rare event simulation in queueing models boils down typically to (i) prove a large deviations result, (ii) find the exponential decay factor, (iii) exponentially twist the distributions, (iv) prove the optimality of the new measure, (v) write the computer programs and execute the simulations (again, see the references cited above for more references of this approach). We have asked ourselves the question how well this approach works. The queueing model to which we apply the methodology has been studied in [5] in case of constant service. We allow random service which means that the large deviations results has to be proved in a different manner. Our studied queueing model will be described in the following section. Thereafter comes the large deviations analysis and the implementation.

2 The Model

In this section we describe the slotted queueing model to which we shall apply the importance sampling technique.

The input of the queue consists of customers (or *cells*) generated by external sources. Say there are L sources, then the ℓ-th source is characterised by a discrete time Markov chain $\{Z_\ell(t) : t = 0, 1, \ldots\}$ on a finite statespace $\mathcal{Z}_\ell$ and by batch sizes $\{A_\ell(t) : t = 0, 1, \ldots\}$ on the nonnegative integers. The Markov chains (Z_ℓ) are called phase chains. Suppose that $Z_\ell(t) = i_\ell,\ \ell = 1, 2, \ldots, L$ which says that at the beginning of the t-th time slot i_ℓ is the state of the ℓ-th phase chain. Then we let $a_\ell(k|i_\ell)$ be the probability that the ℓ-th source generates an input batch of size k:

$$a_\ell(k|i_\ell) = P\Big(A_\ell(t) = k|Z_\ell(t) = i_\ell\Big).$$

Hence, the total input during the t-th time slot of the queue is

$$A(t) = \sum_{\ell=1}^{L} A_\ell(t).$$

Writing $P_\ell = (p_\ell(j|i))_{i,j\in\mathcal{Z}_\ell}$ for the matrix of transition probabilities of the ℓ-th phase chain, an arbitrary source may be specified by the triplet (where we cancel the subscript ℓ) $(\mathcal{Z}, P, (a(\cdot|\cdot))$. For instance, a typical source is a so-called on-off source: it has a two-state phase chain $\mathcal{Z} = \{0,1\}$ with no cell generation in the off-state $a(0|0) = 1$ and a constant stream in the on-state $a(r|1) = 1$. Morover, we notice that our arrival model of an arbitrary source is a special case of the discrete version of the Batch Markovian Arrival Process (D-BMAP) [3]. It has been called a Markov Modulated Process (MMP) in [5]. The probability that the phase chain jumps to state j and that a batch of size k is generated during the t-th time slot, given that the phase chain is in state i at the beginning of the slot is in our model

$$P\Big(Z(t+1) = j, A(t) = k|Z(t) = i\Big) = p(j|i)a(k|i),$$

whereas in D-BMAP it may be $h(k)_{ij}$.

The total input $A(t)$ generated during time slot t arrives at a finite buffer of size B and joins the queue of waiting customers. When it does not fit in the buffer, the excess of B is lost. The accepted customers get their service starting in the next time slot. That is, during time slot t only customers who entered in preceeding slots can leave the buffer after their service completion. This number is generated by a random variable $D(t)$ on the positive integers with probability distribution $d = (d(k))_{k=1}^{\infty}$. We call $D(t)$ the potential number of services during the t-th time slot. The actual number of the departures is at most the number of customers present at the beginning of the slot. For instance in ATM models (with cells in stead of customers), the time slot duration is exactly the time to transmit c cells: $d(c) = 1$.

For modeling purposes it is convenient to assume that the arriving batches are scheduled at the end of the time slot, and the departuring batch at the beginning of the the next slot. Formally, let $X(t)$ be the buffer content at the beginning of slot t, then

$$X(t+1) = \min(X(t) + A(t), B) - \min(D(t), X(t)).$$

We call $\{X(t) : t = 0, 1, \ldots\}$ the *buffer process*. We make the following assumptions on the phase chains, the arrival and departure processes.

Assumptions

1. The phase chains (Z_ℓ) are independent.

2. Each phase chain (Z_ℓ) is irreducible and aperiodic admitting the stationary distribution π_ℓ on its state space.

3. The batch sizes generated by the ℓ-th source are conditionally independent, i.e.
$$P\Big(A_\ell(t_1) = k_1, \ldots, A_\ell(t_n) = k_n \Big| Z_\ell(t_1), \ldots, Z_\ell(t_n)\Big) = \prod_{s=1}^{n} P\Big(A_\ell(t_s) = k_s | Z_\ell(t_s)\Big).$$

4. The departure process $\{D(t) : t = 0, 1, \ldots\}$ consists of i.i.d. random variables.

5. The arrival processes and the departure process are independent.

6. The mean arrival rate is less than the mean departure rate,
$$\lambda := \sum_{\ell=1}^{L} \sum_{i \in \mathcal{Z}_\ell} \pi_\ell(i) \sum_{k=0}^{\infty} k a_\ell(k|i) < \sum_{k=1}^{\infty} k d(k) =: \mu < \infty. \tag{1}$$

7. $\{(X(t), Z_1(t), \ldots, Z_L(t)) : t = 0, 1, \ldots\}$ is an irreducible aperiodic Markov chain.

Notation

We denote vector states and processes by the ˜ symbol:

$$\begin{aligned}
\tilde{Z}(t) &:= (Z_1(t), \ldots, Z_L(t)), \\
\tilde{A}(t) &:= (A_1(t), \ldots, A_L(t)), \\
\tilde{\imath} &:= (i_1, \ldots, i_L).
\end{aligned}$$

Let $\mathcal{I}$ be all vector states $\tilde{\imath}$ such that at least one of the sources generates a positive input $A_\ell(\cdot)$ with positive probability. Then the buffer process is semi-regenerative (X.3 in [1]) with renegeration times the consecutive time epochs when $X(t) = 0$ and $\tilde{Z}(t) \in \mathcal{I}$. The elapsed time between two consecutive regeneration times is called a *regeneration cycle*. Notice that the buffer content $(X(t))$ may remain empty in a regeneration cycle.

3 Large Deviations

For performance evaluation one is interested in the *level crossing probability*, which is the probability that an overflow occurs in a regeneration cycle:

$$\alpha_B := P\Big(X(T_B) + A(T_B) > B | X(0) = 0, \tilde{Z}(0) \in \mathcal{I}\Big),$$

where T_B is the first epoch in the cycle where either the buffer content is zero again or an overflow occurs:

$$T_B := \inf\{t > 0 : X(t) = 0 \text{ or } X(t) + A(t) > B\}.$$

This probability represents the fraction of cycles in which loss of customers occurs. Let

$$\phi_B(\tilde{\imath}) := P\Big(\tilde{Z}(0) = \tilde{\imath} | X(0) = 0, \tilde{Z}(0) \in \mathcal{I}\Big)$$

be the initial state distribution at the beginnings of the regeneration cycles, and let $\alpha_B(\tilde{\imath})$ be the conditional probability of overflow in a cycle given that the cycle starts in $\tilde{\imath}$. Then

$$\alpha_B = \sum_{\tilde{\imath} \in \mathcal{I}} \phi_B(\tilde{\imath}) \alpha_B(\tilde{\imath}).$$

For buffer performance analysis, one is more curious to know the average fraction of customers that is lost in the long run, called the *loss fraction*:

$$\beta_B := \lim_{n \to \infty} \frac{\sum_{t=0}^{n} (X(t) + A(t) - B)^+}{\sum_{t=0}^{n} A(t)},$$

where $x^+ := \max(0, x)$, Because of the regenerative property, this limit exists and is expressable as ratio of two weighted means. Denote $v(\tilde{\imath})$ for the conditional expectation of the number of lost customers in a regeneration cycle given that the cycle starts in state $\tilde{\imath}$, and $w(\tilde{\imath})$ for the conditional expectation of the number of arriving customers in a regeneration cycle given that the cycle starts in state $\tilde{\imath}$. Then

$$\beta_B = \frac{\sum_{\tilde{\imath}} \phi_B(\tilde{\imath}) v(\tilde{\imath})}{\sum_{\tilde{\imath}} \phi_B(\tilde{\imath}) w(\tilde{\imath})}. \tag{2}$$

In case of a single source, the matrix-geometric method has been applied in [3] to obtain the steady-state distribution in the model. The loss probabilities can be extracted from this distribution. Multiple sources may be superimposed to form a meta-source, but then the number of states becomes too large to avoid numerical problems.

Now, let us concentrate on large deviations of α_B and β_B for large buffers ($B \to \infty$). In case of a deterministic service process [5] proves that for any $\tilde{\imath} \in \mathcal{I}$

$$\lim_{B\to\infty} \frac{1}{B} \log\ \alpha_B(\tilde{\imath}) = -\theta. \tag{3}$$

The proof of this result and the determination of $\theta > 0$ is based on the concept of *effective bandwidth.* The same asymptotic result for the conditional overflow probability $\alpha_B(\tilde{\imath})$ holds in case of the random batch services in our model. However, the proof and the determination of θ has to be done by another reasoning. This being technical, we prefer to go along and refer to the Appendix for the details.

For showing a large deviations of the unconditional level crossing probability α_B we need a lemma concerning the initial state distributions $\phi_B(\cdot)$ at the beginning of regeneration cycles.

Lemma 1 *The sequence $(\phi_B(\cdot))_B$ converges pointwise to a probability distribution $\phi(\cdot)$ on $\mathcal{I}$ as $B \to \infty$.*

The proof is in the Appendix.

Corollary 1

$$\lim_{B\to\infty} \frac{1}{B} \log \alpha_B = -\theta.$$

Proof. Let $N = |\mathcal{Z}_1 \times \cdots \times \mathcal{Z}_L|$ be the number of states of the phase chains.

$$\begin{aligned}
&\limsup_{B\to\infty} \frac{1}{B} \log \alpha_B \\
&= \limsup_{B\to\infty} \frac{1}{B} \log \sum_{\tilde{\imath}\in\mathcal{I}} \phi_B(\tilde{\imath})\alpha_B(\tilde{\imath}) \\
&\le \limsup_{B\to\infty} \frac{1}{B} \log \sum_{\tilde{\imath}\in\mathcal{I}} \alpha_B(\tilde{\imath}) \\
&\le \limsup_{B\to\infty} \frac{1}{B} \log N \max_{\tilde{\imath}\in\mathcal{I}} \alpha_B(\tilde{\imath}) \\
&\le \limsup_{B\to\infty} \frac{1}{B} \log N + \limsup_{B\to\infty} \max_{\tilde{\imath}\in\mathcal{I}} \frac{1}{B} \log \alpha_B(\tilde{\imath}) \\
&= -\theta.
\end{aligned}$$

Choose $\tilde{\imath} \in \mathcal{I}$ for which $\phi(\tilde{\imath}) > 0$ holds. Then

$$\liminf_{B\to\infty} \frac{1}{B} \log \alpha_B =$$

$$
\begin{aligned}
&\geq \liminf_{B\to\infty} \frac{1}{B} \log \phi_B(\tilde{\imath})\alpha_B(\tilde{\imath}) \\
&\geq \liminf_{B\to\infty} \frac{1}{B} \log \phi_B(\tilde{\imath}) + \liminf_{B\to\infty} \frac{1}{B} \log \alpha_B(\tilde{\imath}) \\
&= -\theta.
\end{aligned}
$$

Q.E.D.

Without much more effort we obtain a similar result for the loss fraction.

Corollary 2

$$\lim_{B\to\infty} \frac{1}{B} \log\ \beta_B = -\theta.$$

Proof. The conditional expectation $v(\tilde{\imath})$ in the numerator of eq. (2) is rewritten. Denote $v_{\text{loss}}(\tilde{\imath})$ for the conditional expectation of the number of lost customers in a regeneration cycle given that the cycle starts in state $\tilde{\imath}$ and given that the cycle is a loss cycle. Then

$$v(\tilde{\imath}) = \alpha_B(\tilde{\imath}) v_{\text{loss}}(\tilde{\imath}).$$

The conditional expectation $v_{\text{loss}}(\tilde{\imath})$ can be bounded uniformly in B (see [8] for a detailed proof in case of continuous time queues). Also the expectations in the denominator of eq. (2) are bounded uniformly.
Q.E.D.

4 Importance Sampling

The exponential decay factor θ of the previous section is used to define new distributions for the phase chains, the arrival and service processes:

$$
\left\{
\begin{aligned}
p^*_\ell(j|i) &= [p_\ell(j|i)M_{\ell j}(\theta)w_\ell(j)]/[\rho_\ell w_\ell(i)], \\
a^*_\ell(k|i) &= [a_\ell(k|i)e^{\theta k}]/M_{\ell i}(\theta), \\
d^*(k) &= [d(k)e^{-\theta k}]/M_D(-\theta).
\end{aligned}
\right. \tag{4}
$$

In these expressions, $M_{\ell i}(\cdot)$ stands for the conditional moment generating function of the input of source ℓ given the state i of its phase chain, and $M_D(\cdot)$ for the moment generating function of the departures:

$$M_{\ell i}(\theta) = E[\exp(\theta A_\ell(t))|Z_\ell(t) = i], \quad M_D(\theta) = E[\exp(D(t))].$$

Furthermore in eq. (4), $w_\ell(i)$ stands for the i-th entry of a $|\mathcal{Z}_\ell|$-dimensional positive (right) eigenvector w_ℓ belonging to the maximal eigenvalue ρ_ℓ of the $|\mathcal{Z}_\ell|\times|\mathcal{Z}_\ell|$ nonnegative matrix with ij-elements $p_\ell(j|i)M_{\ell j}(\theta)$ (Perron-Frobenius Theorem), see [4, 5, 6].

The simulation of a regeneration cycle with the importance sampling procedure works as follows. Suppose the cycle starts off in state $(X(0) = 0, \tilde{Z}(0) = \tilde{\imath})$. As long as the buffer content has not reached overflow the realisations of the next states of the phase chains, and of the arrival and departure batch sizes are drawn from distributions given in eq. (4). As soon as there is an overflow, i.e., $X(T_B) + A(T_B) > B$, the importance sampling is 'turned off' and the cycle is simulated until completion (buffer content back to 0) by drawing from the original distributions.

Consider a realisation of this happening. The part until overflow is a sequence of the realised random variables:

$$Y := [X(0) = 0, \tilde{Z}(0) = \tilde{\imath}, \tilde{A}(0), \tilde{Z}(1), \tilde{A}(1), D(1), \tilde{Z}(2), \ldots, \tilde{Z}(T_B), \tilde{A}(T_B)]. \qquad (5)$$

We represent by P^* the new probability measure under which the new distributions in eq. (4) are defined and by $L := \mathrm{d}P/\mathrm{d}P^*$ the likelihood ratio between the original and the new measure. Then the likelihood ratio of the regeneration cycle equals the likelihood of the realisation of eq. (5), viz. $L(Y) = \mathrm{d}P(Y)/\mathrm{d}P^*(Y)$ with

$$\begin{aligned} \mathrm{d}P(Y) &= \left[\prod_{t=0}^{T_B}\prod_{\ell=1}^{L} a_\ell(A_\ell(t)|Z_\ell(t))\right] \times \left[\prod_{t=0}^{T_B-1} d(D(t))\right] \\ &\quad \times \left[\prod_{t=1}^{T_B}\prod_{\ell=1}^{L} p_\ell(Z_\ell(t)|Z_\ell(t-1))\right], \qquad (6) \end{aligned}$$

and similarly for $\mathrm{d}P^*(Y)$ by replacing the original distributions by the new ones. We have to calculate this likelihood ratio of each regeneration cycle in which an overflow occurs.

Consider estimating the probability $\alpha_B(\tilde{\imath})$. The set of all realisations of regeneration cycles which start off in state $(X(0) = 0, \tilde{Z}(0) = \tilde{\imath})$ and which reach an overflow, is denoted by U_B. Then

$$\alpha_B(\tilde{\imath}) = P(Y \in U_B) = E[1\{Y \in U_B\}] = E^*[L(Y)1\{Y \in U_B\}].$$

This probability is estimated in importance sampling by the sample mean of n i.i.d. copies of $L(Y)1\{Y \in U_B\}$:

$$\frac{1}{n}\sum_{k=1}^{n} L(Y_k)1\{Y_k \in U_B\}.$$

with variance

$$\frac{1}{n}\Big(E^*[L^2(Y)1\{Y \in U_B\}] - \alpha_B^2(\tilde{\imath})\Big).$$

Comparing with the variance of the original estimator (obtained by drawing realisations from the original distributions only),

$$\frac{1}{n}\Big(E[1\{Y \in U_B\}] - \alpha_B^2(\tilde{\imath})\Big),$$

it is clear that the importance sampling simulations result in variance reduction if and only if

$$E^*[L^2(Y)1\{Y \in U_B\}] \leq E[1\{Y \in U_B\}] = \alpha_B(\tilde{\imath}).$$

The importance sampling simulations are called asymptotically optimal if the left hand side attaines its lower bound (apply Jensen's inequality) in the logarithmic asymptotic sense (see [7]):

$$\lim_{B\to\infty} \frac{\log E^*[L^2(Y)1\{Y \in U_B\}]}{\log \alpha_B(\tilde{\imath})} = 2. \tag{7}$$

We prove in the Appendix that this holds. Similar observations can be made for estimating the (unconditioned) probabilities α_B and β_B.

5 Example

We have implemented the importance sampling in the following (simple) model: two sources, each represented by a two-state Markov chain. State 1 is off-state, no input. State 2 is on-state. In the on-state, the first source generates a deterministic batch, the second source generates a geometric batch. The departure batch is Poisson distributed. The average input rate of source 1 is 0.375, of the 2-nd source 1.25. The output rate is 2.

We ran both the 'direct' simulations (under the original measure) and independently the 'quick' simulations (importance sampling, under the new measure). Each run consisted of the simulation of sufficient regeneration cycles. Since consecutive cycles are not independent, we applied the method of batch means with batch sizes of 50 cycles. The expressions of α_B and β_B contain the initial state probabilities $\phi_B(\tilde{\imath})$ (see section 3). However, we could avoid estimating these by continueing the simulation run. Thus, after completion of a cycle, the next cycle starts off in a realised (feasible) state. In this way the fraction of simulated cycles starting in state $\tilde{\imath}$ should estimate $\phi_B(\tilde{\imath})$. In the importance sampling simulations the initial state distribution $\phi_B^*(\cdot)$ was assumed to be the same as $\phi_B(\cdot)$. Thts implies that the first factor in the likelihood ratio of each cycle, $\phi_B(\tilde{\imath})/\phi_B^*(\tilde{\imath})$, was set to one. Since most importance sampling cycles finished under the original measure, this assumption did not influence much the simulation results.

The most dramatic differences are illustrated by tabeling the number of cycles n required in order to obtain the relative width of a 95%-confidence interval to be less that 2ϵ. Table 1 shows these numbers for ϵ=15%, where the confidence intervals are composed for the β estimation.

B	n	n_{IS}
10	1350	1550
15	3600	950
20	15350	1400
25	48450	950
30	204600	2000

Table 1: required number of cycles.

Also, we ran a fixed number of cycles in both the direct and quick simulations and compare the relative errors of the estimators. Table 2 shows these numbers for $n = 50000$ cycles.

B	RE($\hat{\alpha}$)	RE($\hat{\alpha}_{\text{IS}}$)	RE($\hat{\beta}$)	RE($\hat{\beta}_{\text{IS}}$)
10	9.448e-03	8.403e-03	1.016e-02	9.843e-03
15	1.812e-02	1.031e-02	2.230e-02	1.293e-02
20	3.224e-02	1.193e-02	4.235e-02	1.519e-02
25	6.171e-02	1.023e-02	8.086e-02	1.243e-02
30	1.038e-01	1.096e-02	1.307e-01	1.303e-02

Table 2: relative errors of estimators.

One clearly notices the increase of the RE of the direct estimators and the boundedness of the RE of the quick estimators. Finally, Table 3 give the estimates of the α's and β's obtained simultaneously with Table 1.

B	$\hat{\alpha}$	$\hat{\alpha}_{\mathrm{IS}}$	$\hat{\beta}$	$\hat{\beta}_{\mathrm{IS}}$
10	1.881481e-01	2.223653e-01	3.098606e-02	3.275982e-02
15	6.138889e-02	5.560288e-02	8.705741e-03	7.415723e-03
20	1.798046e-02	1.850408e-02	2.599191e-03	2.882089e-03
25	5.510836e-03	5.430967e-03	7.871924e-04	7.279734e-04
30	1.676442e-03	1.625380e-03	2.555095e-04	2.349516e-04

Table 3: estimates.

The numbers in these tables show that the quick estimates are as reliable as the direct estimates in case of small buffers ('large loss probabilities'). The other tables show the superiority of the importance sampling simulations in terms of relative error and required time. So, we conclude that in this (simple) example importance sampling works well. More studies should be made to find the limits of this method.

Acknowledgement

The author is grateful to the reviewers for there comments.

References

[1] S. Asmussen. *Applied probability and Queues.* Wiley, New York, 1985.

[2] S. Asmussen and R.Y. Rubinstein. Steady state rare events simulation in queueing models and its complexity properties. In: *Advances in Queueing, Theory, Methode, and Open Problems*, ed. J.H. Dshalalow. CRC press, Boca Raton, p. 429 – 461, 1995.

[3] C. Blondia. A Discrete-Time Batch Markovian Arrival Process as B-ISDN Traffic Model. *Belgian Journal of Operations Reserach, Statistics and Computer Science*, 32:3 – 23, 1993.

[4] J.A. Bucklew. *Large Deviations Techniques in Decision, Simulation, and Estimation.* Wiley, New York, 1990.

[5] C.-S. Chang, P. Heidelberger, S. Juneja and P. Shahabuddin. Effective bandwidth and fast simulation of ATM intree networks. *Performance Evaluation*, 20:45 – 65, 1994.

[6] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Jones and Bartlett, Boston, 1993.

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

[8] M. Mandjes. Importance sampling of buffer overflows in batch-arrival queues. Discussion paper TI 95-33. Tinbergen Institute, Amsterdam 1995.

Appendix

In this section we prove some large deviations results for our queueing model. First, we need the following concepts. The asymptotic cumulant generating function of the input of source ℓ is

$$H_\ell(\theta) := \lim_{n\to\infty} \frac{1}{n} \log E\left[\exp\left(\theta \sum_{t=0}^{n-1} A_\ell(t)\right)\right], \quad \theta \in R.$$

In chapter 3 of [6] is shown that this function can be evaluated as the eigenvalue $\rho_\ell(\theta)$ which we described in section 4. Similarly, the cumulant generating function of the potential output process is

$$\begin{aligned} H_D(\theta) &:= \lim_{n\to\infty} \frac{1}{n} \log E\left[\exp\left(\theta \sum_{t=0}^{n-1} D(t)\right)\right] \\ &= \log E[\exp(\theta D(\cdot))], \quad \theta \in R. \end{aligned}$$

Due to the properties of the cumulant generating functions and due to the stability assumption eq. (1) their is a unique $\theta > 0$ such that (cf. [2, 4, 7])

$$\sum_{\ell=1}^{L} \rho_\ell(\theta) + H_D(-\theta) = 0. \tag{8}$$

For the new distributions in eq. (4) of section 4 we used this solution θ and wrote ρ_ℓ for the eigenvalues $\rho_\ell(\cdot)$ evaluated in θ.

We propose that the asymptotics in eqs. (3) and (7) hold for our queueing model with $\theta > 0$ given by eq. (8). As usual in large deviations, the proofs fall in two parts. One proves that the asymptotic is a lower bound and an upper bound.

Proof of the upper bound of eq. (3)

Recall that $\alpha_B(\tilde{\imath}) = E^*[L(Y)1\{Y \in U_B\}]$. We substitute the new distributions of eq. (4)

in the expression of the denumerator $dP^*(Y)$ of the likelihood $L(Y)$ (see eq. (6)) and obtain after collecting factors

$$L(Y) = G \exp[-\theta(X(T_B) + A(T_B))] \times \left[M_D(-\theta) \prod_{\ell=1}^{L} \rho_\ell \right]^{T_B},$$

with

$$G = \prod_{\ell=1}^{L} \left[w_\ell(Z_\ell(T_B)) M_{\ell Z_\ell(T_B)}(\theta) \right] \; [w_\ell(Z_\ell(0))]^{-1}.$$

If $Y \in U_B$, an overflow occurs at time T_B, i.e., $X(T_B) + A(T_B) > B$. The last factor of $L(Y)$ equals 1 due to eq. (8). Because the factor G consists of a finite product of positive constants, we get

$$\limsup_{B \to \infty} \frac{1}{B} \log \alpha_B(\tilde{\imath}) \leq -\theta.$$

Proof of the lower bound of eq. (3)

We consider only the case that the arrival batch sizes are bounded by a constant K with probability 1. Then an overflow $X(T_B) + A(T_B) > B$ implies that $X(T_B) + A(T_B) < B + LK$. Again writing out the likelihood ratio yields

$$E^*[L(Y) 1\{Y \in U_B\}] \geq G \exp[-\theta(B + LK)] E^*[1\{Y \in U_B\}].$$

Under the new measure P^* the drift of the buffer process is positive. This means that with positive probability bounded away from zero for all B an overflow occurs in a regeneration cycle. Hence, taking logarithms, dividing by B and letting $B \to \infty$:

$$\liminf_{B \to \infty} \frac{1}{B} \log \alpha_B(\tilde{\imath}) \geq -\theta.$$

Proof of eq. (7)

This follows the same line of reasoning as the proof of eq. (3) by working out the likelihood ratio.

Proof of lemma 1.

We assume that the arrival batch sizes are bounded by a constant K with probability 1. Let Q_B be the matrix of transition probabilities of the Markov chain $(X(t), \tilde{Z}(t))$ representing the queueing model with a finite buffer B, and $Q = Q_\infty$ when there is no buffer constraint. The associated stationary distributions are ν_B and ν. It suffices to show that $\nu_B \to \nu$ pointwise if $B \to \infty$.

Due to the model assumptions, ν_B and ν are the limiting distributions:

$$\nu_B(x,\tilde{\imath}) = \lim_{n\to\infty} Q_B^n((y,\tilde{\jmath}),(x,\tilde{\imath})), \quad \nu(x,\tilde{\imath}) = \lim_{n\to\infty} Q^n((y,\tilde{\jmath}),(x,\tilde{\imath})),$$

for any $(y,\tilde{\jmath}) \in \{0,1,\ldots,B\} \times \tilde{\mathcal{Z}}$. Furthermore, for any power n, $Q_B^n \to Q^n$ pointwise if $B \to \infty$. Hence, it suffices to show that the matrix sequence $(Q_B^n)_B$ converges (pointwise) to Q^n if $B \to \infty$ uniformly in n:

$$\begin{aligned}
&\lim_{B\to\infty} \nu_B(x,\tilde{\imath}) \\
&= \lim_{B\to\infty} \lim_{n\to\infty} Q_B^n((y,\tilde{\jmath}),(x,\tilde{\imath})) \\
&= \lim_{n\to\infty} \lim_{B\to\infty} Q_B^n((y,\tilde{\jmath}),(x,\tilde{\imath})) \\
&= \lim_{n\to\infty} Q^n((y,\tilde{\jmath}),(x,\tilde{\imath})) \\
&= \nu(x,\tilde{\imath}).
\end{aligned}$$

The uniform convergence may be argued using the exponential decay rate of the probability of reaching high levels (Corollary 1). Let $B_1 < B_2$ be two large buffer sizes. The transition matrices Q_{B_1} and Q_{B_2} agree for all elements $((y,\tilde{\jmath}),(x,\tilde{\imath})) \in \Big(\{0,1,\ldots,B_1 - LK\} \times \tilde{\mathcal{Z}}\Big) \times \Big(\{0,1,\ldots,B_1 - LK\} \times \tilde{\mathcal{Z}}\Big)$. The difference in the probabilities $Q_{B_1}^n((y,\tilde{\jmath}),(x,\tilde{\imath}))$ and $Q_{B_2}^n((y,\tilde{\jmath}),(x,\tilde{\imath}))$ consists of the probabilities (assuming buffer size B_1 respectively B_2) of all paths of length n that start in $(y,\tilde{\jmath})$, end in $(x,\tilde{\imath})$ and have crossed level $B_1 - LK$ in between. Clearly, we can bound these probabilities uniformly (in n) by the probabilities of ever crossing level $B_1 - LK$ starting in $(y,\tilde{\jmath})$.