Queueing Syst (2017) 85:305–312
DOI 10.1007/s11134-016-9512-z

Large deviations for the total queue size in non-Markovian tandem queues

Anne Buijsrogge[1] · Pieter-Tjerk de Boer[1] · Karol Rosen[2] · Werner Scheinhardt[1]

Received: 28 July 2016 / Revised: 8 December 2016 / Published online: 9 January 2017

Abstract We consider a d-node tandem queue with arrival process and light-tailed service processes at all queues i.i.d. and independent of each other. We consider three variations of the probability that the number of customers in the system reaches some high level N, namely during a busy cycle, in steady state, and upon arrival of a new customer. We show that their decay rates for large N have the same value and give an expression for this value.

Keywords GG1 queue · Tandem queue · Decay rate · Large deviations

Mathematics Subject Classification 60K25 · 60F10

1 Introduction

Large deviations for the total queue size in (networks of) queues are of interest since they provide insight into how the probability of overflow decays as the overflow level increases. Such results are well-known for Markovian tandem queues (see, for

✉ Anne Buijsrogge
a.buijsrogge@utwente.nl

Pieter-Tjerk de Boer
p.t.deboer@utwente.nl

Karol Rosen
rosen.karol@gmail.com

Werner Scheinhardt
w.r.w.scheinhardt@utwente.nl

[1] Universiteit Twente, Enschede, The Netherlands

[2] Mexico City, Mexico

example, [4]), but not for non-Markovian tandem queues. Thus, in this short paper, our interest is in the probability that the number of customers in a non-Markovian tandem queue reaches some high level N *during a busy cycle*, and the related probabilities that this number exceeds N *in stationarity* and *upon arrival of a customer*. In Sadowsky [5] the probability in a busy cycle has been considered for a single $G|G|m$ queue. In Bertsimas et al. [1] the Palm probability of a single queue in a network reaching some high level N upon arrival of a customer is considered; the associated decay rate is characterized using the sojourn time of a specific customer. Very related to this work is Ganesh [3], in which the large deviations behavior of the sojourn time for queues in series is considered. The exact asymptotics of the sojourn time for tandem queues have been determined by Foss [2].

In this short paper we will consider a d-node $G|G|1$ tandem queue with renewal input and independent, i.i.d. service processes. We characterize the decay rate for the probability of reaching a total of N customers during a busy cycle of the system. Also we show that the stationary probability of having N customers in the system, as well as the probability of having N customers in the system upon arrival, have the same decay rate.

In Sect. 2 we provide the model and introduce our notation. Section 3 presents the main result of this paper, together with proofs.

2 Model and preliminaries

In this paper we consider d $G|G|1$ queues in tandem. Customers arrive at queue 1 according to a renewal process with inter-arrival times A_k (between customers k and $k+1$) distributed according to some positive random variable A. The service times at queue j, denoted as $B_k^{(j)}$ (for customer k), are independent and identically distributed according to some positive random variable $B^{(j)}$. Furthermore, we assume that all processes are independent and that customers are served based on a first come first served (FCFS) principle. After service completion at queue $j < d$, each customer enters queue $j+1$ immediately, and customers leave the system after service completion at queue d. For stability, we assume $\mathbb{E}\left[B^{(j)}\right] < \mathbb{E}[A]\ \forall j$. See Fig. 1 for a graphical illustration.

Starting with customer 1 entering queue 1 and all other queues empty, we are interested in the probability of overflow during the busy cycle of the total queue. This can be written as $\mathbb{P}(K_N < K_0)$, where K_N is the index of the first customer who reaches the overflow level N and K_0 is the index of the first customer to see an empty system upon arrival. The indices K_N and K_0 can be expressed in terms of the inter-arrival times A_k (at queue 1) and the inter-departure times D_k (from queue d), as follows.

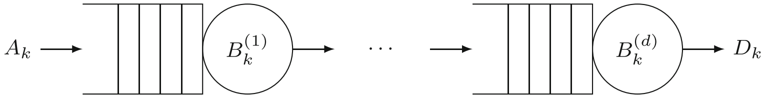

Fig. 1 The d-node tandem queue.

$$K_N = \min\left\{n \geq N : \sum_{k=1}^{n-1} A_k < \sum_{k=1}^{n-N+1} D_k\right\}, \tag{1}$$

$$K_0 = \min\left\{m : \sum_{k=1}^{m-1} A_k > \sum_{k=1}^{m-1} D_k\right\}. \tag{2}$$

For the inter-departure time D_k (between customers $k-1$ and k, for $k \geq 2$), we can write $D_k = B_k^{(d)} + I_k^{(d)}$, where $I_k^{(d)}$ is the, possibly zero, idle time of queue d after the departure of customer $k-1$, before customer k enters queue d. Consistently with this, D_1 is simply defined as the sojourn time of customer 1.

Other probabilities of interest that are related to $\mathbb{P}(K_N < K_0)$ are $\mathbb{P}(L \geq N)$ and $\mathbb{P}(L^{(a)} \geq N)$, where L denotes the total number of customers in the system in stationarity, and $L^{(a)}$ denotes the same number but immediately after an arbitrary arrival (including the customer that just arrived).

To characterize the decay rate, we need the following. For any random variable X, let $\Lambda_X(\theta) = \log \mathbb{E}\left[e^{\theta X}\right]$ denote its log moment generating function. For all $j = 1, \ldots, d$, we assume that $\Lambda_{B^{(j)}}(\theta)$ exists for some $\theta > 0$, and define θ_j as

$$\theta_j = \sup_{\theta}\left\{\Lambda_A(-\theta) + \Lambda_{B^{(j)}}(\theta) \leq 0\right\}.$$

Note that we only consider $\Lambda_A(-\theta)$ for $\theta \geq 0$ and so it always exists. Furthermore, we say $\theta_j = \infty$ when $\Lambda_A(-\theta) + \Lambda_{B^{(j)}}(\theta) < 0$ for all $\theta > 0$; note that this is equivalent to $\mathbb{P}(B^{(j)} > A) = 0$.

Finally, we define $\theta_{\min} = \min_j(\theta_j)$, and assume that $\theta_{\min} < \infty$, i.e., we do not have $\mathbb{P}(B^{(j)} > A) = 0$ for all queues, so that the number of customers can grow arbitrarily large and the decay rates of the probabilities of interest will be in $(0, \infty)$. The queue(s) j with $\theta_j = \theta_{\min}$ will be called the θ-bottleneck queue(s). Note that this notion can be different from the ρ-bottleneck queue, which is the queue with the smallest server utilization $\rho_j = \mathbb{E}[B^{(j)}]/\mathbb{E}[A]$.

3 Main result

In this section we present the main result of this paper, namely the characterization of the decay rates of $\mathbb{P}(K_N < K_0)$, $\mathbb{P}(L \geq N)$ and $\mathbb{P}(L^{(a)} \geq N)$. In order to achieve this result, we will prove both a lower bound and an upper bound for the decay of $\mathbb{P}(K_N < K_0)$, which will also turn out to hold for the other decay rates. We will start with the lower bound, with a proof based on a coupling argument.

Lemma 1 *(Lower bound) For the decay of* $\mathbb{P}(K_N < K_0)$ *it holds that*

$$\liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(K_N < K_0\right) \geq \Lambda_A(-\theta_{\min}).$$

Proof We compare the tandem queue to a single queue with the same arrival process A_k and the service process of the jth queue in the tandem, $B_k^{(j)}$. (This is equivalent to comparing our tandem queue to a tandem queue with the same arrival process and all service times set to 0, except the service times of queue j.) The idea of the proof is to show that overflow is more likely in the tandem queue than in the single queue.

Define $\widehat{D}_i$, $\widehat{K}_0$ and $\widehat{K}_N$ analogously to D_i, K_0 and K_N but for the single queue. Denote the inter-departure time of customer i at queue j in the tandem queue by $D_i^{(j)}$.

For $i < K_0$ it holds that $D_i^{(j)} = I_i^{(j)} + B_i^{(j)}$, and for $i < \widehat{K}_0$ it holds that $\widehat{D}_i = B_i^{(j)}$, as the single queue does not have idle times during its busy cycle. Since a customer cannot leave the last queue in the tandem before having left queue j, we find

$$\sum_{i=1}^{k} D_i \geq \sum_{i=1}^{k} D_i^{(j)} = \sum_{i=1}^{k} \widehat{D}_i + I_i^{(j)} \geq \sum_{i=1}^{k} \widehat{D}_i, \tag{3}$$

for all $k = 1, ..., \min(K_0 - 1, \widehat{K}_0 - 1)$, meaning that a customer leaves the tandem queue not earlier than that same customer leaves the coupled single queue.

Based on this we first show, by contradiction, that $\widehat{K}_0 \leq K_0$, i.e., the single queue empties not later than the tandem queue. Suppose that $\widehat{K}_0 > K_0$, then (3) still holds for k up to $K_0 - 1$. By using (2) and (3) we have

$$\sum_{k=1}^{K_0-1} A_k > \sum_{k=1}^{K_0-1} D_k \geq \sum_{k=1}^{K_0-1} \widehat{D}_k,$$

which implies by definition of $\widehat{K}_0$ that $\widehat{K}_0 \leq K_0$. Therefore, our assumption $\widehat{K}_0 > K_0$ is wrong and so we have shown $\widehat{K}_0 \leq K_0$.

Next, we show that the tandem queue reaches the overflow level not later than the single queue. Suppose we have reached overflow in a busy cycle of the single queue, that is, $\widehat{K}_N < \widehat{K}_0$. Then we have, by using (1) and (3),

$$\sum_{k=1}^{\widehat{K}_N-1} A_k < \sum_{k=1}^{\widehat{K}_N-N+1} \widehat{D}_k \leq \sum_{k=1}^{\widehat{K}_N-N+1} D_k,$$

and thus $K_N \leq \widehat{K}_N$.

Hence $\widehat{K}_N < \widehat{K}_0$ implies $K_N < K_0$, which means that overflow during a busy period in the single queue implies overflow during a busy period in the tandem queue. So we have for any j that

$$\liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(K_N < K_0\right) \geq \liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(\widehat{K}_N < \widehat{K}_0\right) = \Lambda_A(-\theta_j),$$

where the second step follows by Theorem 1 in [5]. In particular, the above holds for j such that $\theta_j = \theta_{\min}$, which completes the proof. □

The next step is to prove an upper bound. We will use a regenerative argument, for which we need that the expected total time spent at or above level N during a busy cycle in which level N is reached, is bounded from below, independently of N.

Even though this sounds very plausible, we could not find a reference. Hence the next lemma, the proof of which is based on first principles, together with the technical assumption $\mathbb{P}(B^{(d)} > A) > 0$ (which will not be a limitation for the main result).

Let $L(t)$ be the total number of customers in the system at time t, and let T be the length of the first busy cycle; then, we define the expected total time τ_N spent at or above level N during a busy cycle as $\tau_N = \int_0^T \mathbb{1}\{L(t) \geq N\}dt$.

Lemma 2 *Suppose that* $\mathbb{P}(B^{(d)} > A) > 0$. *Then some* $c > 0$ *exists such that for all* $N = 1, 2, \ldots,$

$$\mathbb{E}\left[\tau_N \mid K_N < K_0\right] \geq c.$$

Proof Consider a busy cycle in which the overflow level N is reached and denote the moment that N is reached for the first time by t. Then the first arrival after t occurs at time $t_1 = t + A_{K_N}$, while the *second* departure after t occurs at some time $t_2 \geq t + B^{(d)}_{K_N - N+2}$. (To see this, note that at time t, when customer K_N enters, customer $K_N - N + 1$ is the first to depart from the system, so the service of customer $K_N - N + 2$ at queue d cannot start earlier than at time t.) It is not difficult to check that if $t_1 < t_2$, there will be at least N customers in the system between t_1 and t_2. Thus, for any N we have $\mathbb{E}\left[\tau_N \mid K_N < K_0\right] \geq \mathbb{E}\left[\max(0, t_2 - t_1) \mid K_N < K_0\right] \geq \mathbb{E}\left[\max(0, B^{(d)} - A)\right]$, which is nonzero due to $\mathbb{P}(B^{(d)} > A) > 0$. □

We are now ready to prove the upper bound, based on a regenerative argument and a Chernoff bound.

Lemma 3 *(Upper bound) For the decay of* $\mathbb{P}(K_N < K_0)$, *under the condition that* $\mathbb{P}(B^{(d)} > A) > 0$, *it holds that*

$$\begin{aligned}\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(K_N < K_0\right) &\leq \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(L \geq N\right) \\ &\leq \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(L^{(a)} \geq N\right) \\ &\leq \Lambda_A(-\theta_{\min}),\end{aligned}$$

and a similar statement holds when we replace all limsups by liminfs.

Proof The proof for the liminfs and the limsups is similar; we only give it explicitly for the limsups. The same steps apply to prove the liminfs, in which the supremum has to be replaced by the infimum at the appropriate places.

The first inequality follows from a regenerative argument, as in [4], by which we have

$$\mathbb{P}(K_N < K_0) = \frac{\mathbb{E}\left[T\right]\ \mathbb{P}(L \geq N)}{\mathbb{E}\left[\tau_N \mid K_N < K_0\right]},$$

where T is the length of a busy cycle, which has a finite, constant expectation due to stability of the system, and τ_N is the total time spent above level N during a busy cycle, which is bounded from below independently of N; see Lemma 2.

The remainder of the proof considers the system in stationarity, so time 0 and customer 0 are not necessarily related to the start of a busy cycle. For the second inequality then, fix some arbitrary time t in stationarity, and consider the last customer to arrive before time t, call this customer k. If the number of customers at time t is $\geq N$, then the queue length $L_k^{(a)}$ observed by—and including—customer k is also $\geq N$, because there can only be departures between the arrival of customer k and time t. So $\mathbb{P}(L \geq N) \leq \mathbb{P}(L_k^{(a)} \geq N)$. Furthermore, $L_k^{(a)} \geq N$ if and only if the sojourn time of customer $k - N + 1$, denoted by S_{k-N+1}, exceeds the sum of $N - 1$ inter-arrival times. So we have

$$\mathbb{P}(L_k^{(a)} \geq N) = \mathbb{P}\left(S_{k-N+1} \geq \sum_{i=k-N+1}^{k-1} A_i\right).$$

Note that this probability is independent of the age of A_k at time t, as the inter-arrival times are independent, so in fact $L_k^{(a)}$ has the same distribution as $L^{(a)}$, i.e., customer k cannot be distinguished from an arbitrary customer in stationarity, which proves the second inequality.

For the last inequality, we analyze the right-hand side of the equation above (keeping customer index $k-N+1$ for convenience). We have for any $\theta > 0$, using the Chernoff bound, and the independence of S_{k-N+1} and $\sum_{i=k-N+1}^{k-1} A_i$,

$$\begin{aligned}\mathbb{P}\left(S_{k-N+1} \geq \sum_{i=k-N+1}^{k-1} A_i\right) &\leq \mathbb{E}\left[e^{\theta\left(S_{k-N+1}-\sum_{i=k-N+1}^{k-1} A_i\right)}\right] \\ &= \mathbb{E}\left[e^{\theta S_{k-N+1}}\right] \mathbb{E}\left[e^{-\theta \sum_{i=k-N+1}^{k-1} A_i}\right].\end{aligned}$$

In [3] it is shown that $\mathbb{E}[e^{\theta S_{k-N+1}}]$ is upper bounded by some constant C for all $\theta \in (0, \theta_{\min})$ (see just after equation (27) in the proof of Theorem 1). Note that the assumptions in [3] are more general than ours, so we can use this result. Hence, we have for any $\theta \in (0, \theta_{\min})$

$$\begin{aligned}&\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(S_{k-N+1} \geq \sum_{i=k-N+1}^{k-1} A_i\right) \\ &\quad\leq \limsup_{N\to\infty} \frac{1}{N}\left(\log C + \log \mathbb{E}[e^{-\theta \sum_{i=k-N+1}^{k-1} A_i}]\right) = \Lambda_A(-\theta),\end{aligned}$$

where the last step follows by independence of the inter-arrival times. Taking $\theta \to \theta_{\min}$ to achieve the best possible bound proves the statement. □

Theorem 1 *Consider a stable FCFS d-node $G|G|1$ tandem queue with arrival process and light-tailed service processes at all queues i.i.d. and independent of each other. If $\theta_{\min} < \infty$, it holds that*

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}(K_N < K_0) &= \lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}(L \geq N)\\ &= \lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}(L^{(a)} \geq N) = \Lambda_A(-\theta_{\min}). \qquad (4)\end{aligned}$$

Proof When $\mathbb{P}(B^{(d)} > A) > 0$, statement (4) follows immediately from Lemmas 1 and 3 since all liminfs and limsups (with respect to each of the three probabilities) are equal to $\Lambda_A(-\theta_{\min})$.

To show that (4) also holds in general, we consider a tandem queue where $\mathbb{P}(B^{(d)} > A) = 0$, and two corresponding systems, fed by the same arrival process. One is a queue in isolation as introduced in the proof of Lemma 1. More specifically, we consider a θ-bottleneck queue, i.e., some queue j for which $\theta_j = \theta_{\min}$. In this single queue we define $\widehat{K}_0$, $\widehat{K}_N$, $\widehat{L}$ and $\widehat{L}^{(a)}$ analogously to K_0, K_N, L and $L^{(a)}$ in the tandem queue. Note that $\mathbb{P}(B^{(j)} > A) > 0$ (otherwise we would have $\theta_{\min} = \theta_j = \infty$), and hence (4) holds for this single queue system.

The other system we consider is the original tandem queue augmented with a suitably chosen additional queue $d + 1$, for example, letting $B^{(d+1)} \sim B^{(j)}$ where queue j is a θ-bottleneck queue (another option is to choose $B^{(d+1)} \sim \exp(\mu)$ for some sufficiently large μ). In this system we analogously define $\widetilde{K}_0$, $\widetilde{K}_N$, $\widetilde{L}$ and $\widetilde{L}^{(a)}$. Clearly we then have $\mathbb{E}\left[B^{(d+1)}\right] < \mathbb{E}[A]$ and $\theta_{d+1} \geq \theta_{\min}$, while we also have $\mathbb{P}(B^{(d+1)} > A) > 0$. As a result, for this system (4) also holds.

All three probabilities for the original tandem queue can now be bounded by the corresponding probabilities in the two other systems, as follows:

$$\begin{array}{lllll}\mathbb{P}(\widehat{K}_N < \widehat{K}_0) & \leq & \mathbb{P}(K_N < K_0) & \leq & \mathbb{P}(\widetilde{K}_N < \widetilde{K}_0),\\ \mathbb{P}(\widehat{L} \geq N) & \leq & \mathbb{P}(L \geq N) & \leq & \mathbb{P}(\widetilde{L} \geq N),\\ \mathbb{P}(\widehat{L}^{(a)} \geq N) & \leq & \mathbb{P}(L^{(a)} \geq N) & \leq & \mathbb{P}(\widetilde{L}^{(a)} \geq N).\end{array}$$

Each of these inequalities follows similarly to the proof of Lemma 1 by coupling arguments; note that setting $B^{(d+1)} \equiv 0$ in the augmented tandem queue leads to the original tandem, and setting the service times of all but one queue in the original tandem queue leads to the single queue. Thus, the first inequality is straightforward from the proof of Lemma 1, and the second can be shown similarly. For the other two lines, we just need to consider the departure times in the three systems for the same customer to show that $\widehat{L}(t) \leq L(t) \leq \widetilde{L}(t)$ at any time t, and hence also in stationarity and upon arrivals.

Finally, we take logarithms above, then divide by N, and take limits. □

Note that when $\theta_{\min} = \infty$, the total number of customers cannot grow arbitrarily large (see Sect. 2), and hence the decay rates in (4) are not properly defined (or are equal to $-\infty$).

Remark 1 As mentioned in the introduction, Bertsimas et al. [1] and Ganesh [3] consider the decay of related overflow probabilities in a more general setting, where certain types of dependence for the arrival and service processes are allowed. We expect that the bounds in our current work can be extended to this case as well, but this will take different techniques and additional effort, in particular to relate $\mathbb{P}(K_N < K_0)$, $\mathbb{P}(L \geq N)$ and $\mathbb{P}(L^{(a)} \geq N)$ in the more general setting.

Acknowledgements This work is supported by the Netherlands Organisation for Scientific Research (NWO), project number 613.001.105.

References

1. Bertsimas, D., Paschalidis, I.C., Tsitsiklis, J.N.: Large deviations analysis of the generalized processor sharing policy. Queueing Syst. **32**(4), 319–349 (1999)
2. Foss, S.G.: On the exact asymptotics for the stationary sojourn time distribution in a tandem of queues with light-tailed service times. Probl. Inf. Transm **43**(4), 353–366 (2007)
3. Ganesh, A.J.: Large deviations of the sojourn time for queues in series. Ann. Oper. Res. **79**, 3–26 (1998)
4. Glasserman, P., Kou, S.G.: Analysis of an importance sampling estimator for tandem queues. ACM Trans. Model. Comput. Simul. **5**(1), 22–42 (1995)
5. Sadowsky, J.S.: Large deviations theory and efficient simulation of excessive backlogs in a $GI|GI|m$ queue. IEEE Trans. Autom. Control **36**(12), 1383–1394 (1991)