Large Deviations of the Sojourn Time for Queues in Series

A. J. Ganesh[*]

Abstract

We consider an open queueing network consisting of an arbitrary number of queues in series. We assume that the arrival process into the first queue and the service processes at the individual queues are jointly stationary and ergodic, and that the mean inter-arrival time exceeds the mean service time at each of the queues. Starting from Lindley's recursion for the waiting time, we obtain a simple expression for the total delay (sojourn time) in the system. Under some mild additional assumptions, which are satisfied by many commonly used models, we show that the delay distribution has an exponentially decaying tail and compute the exact decay rate. We also find the most likely path leading to the build-up of large delays. Such a result is of relevance to communication networks, where it is often necessary to guarantee bounds on the probability of large delays. Such bounds are part of the specification of the quality of service desired by the network user.

1 Introduction

The problem considered here is motivated by applications to the design and operation of asynchronous transfer mode (ATM) networks. These are intended to integrate different traffic types like voice, video and data, and aim to exploit the efficiency gains of statistical multiplexing while at the same time providing a guaranteed quality of service (QoS) to the user. In an ATM network, the incoming traffic from each call is split into cells of fixed size (53 bytes). The cells from various calls are multiplexed onto a link for transmission, either on a first come first served basis, or using some priority rules. There are finite buffers at each node, and if these are full when a cell arrives, then the cell is lost.

Statistical multiplexing implies that guaranteed bandwidth is not available to the user, and therefore deterministic service criteria cannot be met. The QoS criteria take the form of bounds on the probability of cell loss, and of end-to-end transmission delays exceeding a threshold. While data traffic is sensitive to cell loss but usually not to transmission delays, the opposite is the case for voice and video. The problem facing the network operator is that of finding a call

[*] Department of Statistics, Birkbeck College, University of London, Malet Street, London WC1E 7HX, U.K. Research supported by a fellowship from BP and the Royal Society of Edinburgh.

admission policy which maximizes network utilization while ensuring that accepted calls enjoy a specified quality of service. Therefore, it is of interest to estimate cell loss probabilities, and the probability of large delays, given a description of the traffic characteristics. These are typically obtained in the context of a queueing model of the communication network. Exact expressions for the probabilities of interest are available only for a few special network models, namely Jackson networks and networks of quasi-reversible queues(see Kelly [7]). The assumptions on the traffic processes embodied in these models are unrealistic for the traffic types encountered in ATM networks. Furthermore, their use results in unduly optimistic performance predictions.

There has recently been considerable interest in the use of large deviations techniques to estimate cell loss probabilities. The case of a single deterministic queue multiplexing several independent traffic streams with fairly general arrival processes is considered by de Veciana and Walrand [4]. They show that the queue size distribution has an exponentially decaying tail and compute the decay rate. The result is extended to intree networks of such queues by Chang [3]. Ganesh and Anantharam [6] obtain the decay rate of the tail distribution for two exponential server queues in series fed by renewal arrivals. In a recent remarkable paper, Bertsimas *et al.* [2] consider acyclic queueing networks with fairly general arrival processes, and independent, identically distributed (*i.i.d.*) service times at each queue. They compute the decay rate of the stationary waiting time and queue length distributions at each node in the network.

While the tail of the queue size distribution can be easily related to the delay at a single queue, there are two problems when it comes to networks. One is that knowledge of the marginal queue length distributions is not adequate; at the very least, we would need to know the joint distribution of the queue lengths at different nodes. Even this is not enough; we need to be able to relate the joint distribution seen by arrivals (or in stationarity) to that seen by a customer moving from one queue to another, conditional on its history. This is a difficult problem and a solution is not available in general. The distribution of the sojourn time is known only in the case of overtake-free networks of quasi-reversible queues; see Walrand [12] for details.

In this paper, we consider the problem of estimating the total delay for a network consisting of an arbitrary number of queues in series, with quite general arrival and service processes. Our assumptions regarding these processes are stated in Section 3. We show that the distribution of the total sojourn time in the tandem has an exponentially decaying tail, and obtain the rate of decay. We introduce some notation and state the problem formally in the next section. We then use Lindley's recursion to obtain an expression for the total sojourn time in terms of the service times at the individual nodes and the inter-arrival times. We estimate the tail of the sojourn time distribution in Section 3. We also describe the most likely path of the process leading to large delays. A similar description in the context of a single queue has been obtained by Anantharam [1].

2 The Sojourn Time in Tandem Queues

Consider a system of M queues in series. Customers arrive into the system from outside requiring service at each of the M nodes. The arrival process and the required service times may be modeled jointly as a stochastic process. Customers enter the system at the first queue, traverse the queues in sequence, and leave the system after completing service at the last queue. There is a single server at each queue. The service discipline is first come first served (FCFS) and work-conserving (a server is never idle when its queue is non-empty). We assume that the system is in operation for all time, i.e., for $t \in (-\infty, \infty)$. We pick an arbitrary customer that we designate customer zero. Let S_n^m denote the service time required by the n^{th} customer at the m^{th} queue, $m = 1, \ldots, M$, $n = \ldots, -1, 0, 1, \ldots$. Let T_n^m denote the inter-arrival time of the n^{th} customer at the m^{th} queue, i.e., the time between the arrival of the n^{th} and $(n-1)^{\text{th}}$ customers to this queue. Define

$$\tau_{i,j}^m = \sum_{k=i}^{j} T_k^m, \qquad \sigma_{i,j}^m = \sum_{k=i}^{j} S_k^m$$

As usual, if $i > j$ then the sum is empty and is taken to be zero. We assume that T_n^1 and S_n^m, $m = 1, \ldots, M$ are jointly stationary and ergodic. We also assume that the system is stable, namely, that the mean inter-arrival time of customers exceeds their mean service time at each of the queues. In other words, $ET^1 > ES^m$ for all $m = 1, \ldots, M$.

Let W_n^m and D_n^m denote the waiting time and sojourn time respectively of the n^{th} customer in the m^{th} queue. The waiting time is the time from arrival until the start of service, and the sojourn time the time from arrival until the end of service. The waiting times satisfy Lindley's recursion (see [9])

$$W_n^m \;=\; (W_{n-1}^m + S_{n-1}^m - T_n^m)^+, \quad m = 1, \ldots, M,$$

where X^+ denotes $\max\{X, 0\}$. It was shown by Loynes [9] that, if the arrival and service time distributions satisfy the stability criterion, then Lindley's recursion has the solution

$$W_n^m \;=\; \max_{j_m \le n} \left(\sigma_{j_m, n-1}^m - \tau_{j_m+1, n}^m\right) \quad m = 1, \ldots, M, \tag{1}$$

and the maximum is achieved, almost surely, for $j_m > -\infty$, $m = 1, \ldots, M$. Also, the above is the unique (up to sets of measure zero) solution of Lindley's recursion for which W_n^m is finite almost surely. The solution in (1) has the following interpretation. Let $W_n^{m,k}$ denote the waiting time of the n^{th} customer in the m^{th} queue, in a system which is assumed to start empty at the arrival time of the k^{th} customer. Then, as k decreases to $-\infty$, $W_n^{m,k}$ increases monotonically to a limit, which is the solution W_n^m above. Since the sojourn time of a customer is the sum of its waiting time and its own service time, it follows from (1) that

$$D_n^m \;=\; \max_{j_m \le n} \left(\sigma_{j_m, n}^m - \tau_{j_m+1, n}^m\right) \tag{2}$$

Since the inter-arrival and service times were assumed to be stationary, so are the sojourn times given by the above expression. In addition, they are almost surely finite.

The inter-arrival time in the m^{th} queue is the inter-departure time from the $(m-1)^{\text{th}}$ queue, for $m \geq 2$. But the departure epoch of a customer is the sum of its arrival epoch and its sojourn time. Thus, for $m \geq 2$,

$$T_n^m = T_n^{m-1} + D_n^{m-1} - D_{n-1}^{m-1}$$

and so

$$\tau_{i,j}^m = \begin{cases} \tau_{i,j}^{m-1} + D_j^{m-1} - D_{i-1}^{m-1}, & \text{if } i \leq j+1 \\ 0, & \text{else} \end{cases} \tag{3}$$

Substituting in (2), we get

$$D_n^m = \max_{j_m \leq n} (\sigma_{j_m,n}^m - \tau_{j_m+1,n}^{m-1} - D_n^{m-1} + D_{j_m}^{m-1})$$

Hence

$$D_n^m + D_n^{m-1} = \max_{j_m \leq n} (\sigma_{j_m,n}^m - \tau_{j_m+1,n}^{m-1} + D_{j_m}^{m-1})$$

But, by (2), $D_{j_m}^{m-1} = \max_{j_{m-1} \leq j_m} (\sigma_{j_{m-1},j_m}^{m-1} - \tau_{j_{m-1}+1,j_m}^{m-1})$. Therefore,

$$\begin{aligned} D_n^m + D_n^{m-1} &= \max_{j_{m-1} \leq j_m \leq n} (\sigma_{j_m,n}^m - \tau_{j_m+1,n}^{m-1} + \sigma_{j_{m-1},j_m}^{m-1} - \tau_{j_{m-1}+1,j_m}^{m-1}) \\ &= \max_{j_{m-1} \leq j_m \leq n} (\sigma_{j_m,n}^m + \sigma_{j_{m-1},j_m}^{m-1} - \tau_{j_{m-1}+1,n}^{m-1}) \end{aligned} \tag{4}$$

Inductively, we obtain

$$\sum_{m=1}^{M} D_0^m = \max_{j_1 \leq \ldots \leq j_{M+1}=0} \left(\sum_{m=1}^{M} \sigma_{j_m,j_{m+1}}^m - \tau_{j_1+1,0}^1 \right) \tag{5}$$

3 The Tail of the Sojourn Time Distribution

The significance of the above result is that it provides a non-recursive relationship between the total sojourn time in a tandem, the external arrival process and the service processes at the individual nodes. We use it to estimate the probability of large total delay in the tandem. We are interested in particular in obtaining bounds on this probability that decay exponentially in the delay, i.e., in estimates of the form $\mathbf{P}(\sum_{m=1}^{M} D_0^m \geq x) \approx \exp(-\eta x)$. These are obtained as follows. We derive from (5) necessary and sufficient conditions on the arrival and service processes for the event $\{\sum_{m=1}^{M} D_0^m \geq x\}$. We use Chernoff's inequality to get an upper bound on the probability of the necessary conditions being met. We use the Gärtner-Ellis theorem from large deviations theory (see [5]) to obtain a lower bound on the probability that the sufficient conditions are satisfied.

While (5) holds for quite general arrival and service distributions, some additional assumptions are required in order to obtain exponential bounds. In particular, we require that the service times distributions have exponential tails. We shall also need restrictions on the correlations between successive inter-arrival times or service times, and between the arrival process and the service processes at the various queues. We consider two distinct settings. In one, the service processes at different queues are mutually independent, although successive service times at any one queue may be correlated. In the other, the service times of different customers are *i.i.d.*, though the service times required by any one customer at different queues may be correlated. In both cases, we assume that the arrival process is independent of the service requirements. We shall see that the evolution of large total delay can be very different in these two settings. We introduce below the assumptions that are used in the rest of the paper.

Assumptions

1. (a) The arrival process into the first queue and the service processes at the individual queues are mutually independent, stationary and ergodic, or

   (b) The arrival process is stationary and ergodic. Each arrival is handed a vector of service times required at the individual nodes. These vectors are identically distributed (with arbitrary joint distribution), independent from customer to customer, and independent of the arrival times.

2. The stability condition holds, i.e., $ET^1 > ES^m$, $m = 1, \ldots, M$. In other words, the mean inter-arrival time exceeds the mean service time at any of the queues.

3. For each $m = 1, \ldots, M$, and for all real θ, the limits

$$\Lambda_{T^1}(\theta) = \lim_{n\to\infty} \frac{1}{n} \log E[\exp(\theta \tau^1_{1,n})]$$

   and

$$\Lambda_{S^m}(\theta) = \lim_{n\to\infty} \frac{1}{n} \log E[\exp(\theta \sigma^m_{1,n})]$$

   exist as extended real numbers.

4. If 1(a) holds, then the functions $\Lambda_{T^1}(\cdot)$ and $\Lambda_{S^m}(\cdot)$, $m = 1, \ldots, M$ satisfy the assumptions of the Gärtner-Ellis theorem (see [5]).

The independence assumptions in 1 are not unreasonable for most models of practical interest. The stability criterion in 2 is essential to ensure that delays do not become infinite almost surely. Assumptions 3 and 4 are not very restrictive, and are satisfied by most commonly used traffic models. For example, if the inter-arrival and service times are *i.i.d.* with exponential tails, or alternatively, if they are (random) co-ordinate functions of Markov chains satisfying strong uniformity conditions on the transition kernel and the tails, then these assumptions are satisfied. In particular, Poisson, phase-type (see, *e.g.*, [12]) and deterministic processes satisfy these conditions. So do Markov modulated versions of these processes, where the modulating Markov chain has a finite number of states. Finally, while the above assumptions require that the inter-arrival times have an exponential tail, this requirement can be relaxed along the lines of Assumption B in Bertsimas *et al.*, [2].

We derive below some properties of the logarithmic moment-generating functions, $\Lambda(\cdot)$, defined in assumptions 3 and 4 above. These are needed to prove our main result regarding the tail of the delay distribution.

Lemma 1 *Suppose assumptions 2-4 above hold. Define*

$$\theta^m = \sup\{\theta > 0 : \Lambda_{T^1}(-\theta) + \Lambda_{S^m}(\theta) < 0\}, \quad m = 1, \ldots, M,$$

where the supremum of the empty set is $-\infty$*. Then* $\theta^m > 0$ *and*

$$\Lambda_{T^1}(-\theta) + \Lambda_{S^m}(\theta) < 0 \quad \textit{if} \quad \theta \in (0, \theta^m) \tag{6}$$

$$\Lambda_{T^1}(-\theta) + \Lambda_{S^m}(\theta) > 0 \quad \textit{if} \quad \theta \notin [0, \theta^m] \tag{7}$$

Proof : By the definition of Λ_{T^1} and Λ_{S^m} above, we have $\Lambda_{T^1}(0) = \Lambda_{S^m}(0) = 0$ for all $m = 1, \ldots, M$. Hence, by assumption 3,

$$\begin{aligned}\Lambda_{T^1}(-\delta) + \Lambda_{S^m}(\delta) &= -\delta\Lambda'_{T^1}(0) + \delta\Lambda'_{S^m}(0) + o(\delta) \\ &= \delta(ES^m - ET^1) + o(\delta)\end{aligned}$$

is less than zero for sufficiently small $\delta > 0$ by the stability assumption. Hence, the set over which the supremum in the definition of θ^m is taken is non-empty. Therefore, $\theta^m > 0$, $m = 1, \ldots, M$. By Lemma 2.3.9 in [5], Λ_{T^1} and Λ_{S^m} are convex and greater than $-\infty$ everywhere. Hence, so is $\Lambda_{T^1}(-\theta) + \Lambda_{S^m}(\theta)$. Together with the definition of θ^m and the fact that $\Lambda_{T^1}(0) + \Lambda_{S^m}(0) = 0$, this implies the claim of the lemma.

Lemma 2 *Let* $\theta^1, \ldots, \theta^M$ *be as above. Define*

$$\theta^0 = \sup\left\{\theta : E\left[\exp\Big(\theta(S_1^1 + \ldots + S_1^M)\Big)\right] < +\infty\right\}$$

Clearly $\theta^0 \geq 0$*. Define* $\theta^* = \min_{m=0,\ldots,M} \theta^m$*. If assumption 1(a) is satisfied, then* $\theta^* = \min_{m=1,\ldots,M} \theta^m$*, i.e.,* θ^0 *can be excluded in taking the minimum.*

Proof : Suppose assumption 1(a) holds. We shall show that if $\theta > \theta^0 \geq 0$, then $\theta \geq \theta^m$ for some $1 \leq m \leq M$, thereby proving the lemma. Now, if $\theta > \theta^0$, then

$$E\left[\exp\Big(\theta(S_1^1 + \ldots + S_1^M)\Big)\right] = \prod_{m=1}^{M} E\left[\exp(\theta S_1^m)\right] = +\infty$$

The first equality above holds because the service processes at the individual queues are independent by assumption 1(a), while the second equality follows from the definition of θ^0. Choose $m \in \{1, \ldots, M\}$ such that $E[\exp(\theta S_1^m)] = \infty$. Then, by the non-negativity of the service times, we have for every n,

$$\frac{1}{n} \log E\left[\exp\Big(\theta(S_1^m + \ldots + S_n^m)\Big)\right] \geq \frac{1}{n} \log E\left[\exp(\theta S_1^m)\right] = \infty$$

Letting n go to infinity, we get $\Lambda_{S^m}(\theta) = \infty$. Since $\Lambda_{T^1} > -\infty$ everywhere,

$$\Lambda_{S^m}(\theta) + \Lambda_{T^1}(-\theta) > 0$$

Therefore, $\theta \geq \theta_m$ by Lemma 1. Since $\theta > \theta^0$ was arbitrary, $\theta^0 \geq \theta^m$ for some $1 \leq m \leq M$. This establishes the claim of the lemma.

Below, we present our main result regarding the probabilities of large sojourn times in a tandem. The intuition underlying this result is as follows. Let $x > 0$ be given. By (5), a necessary condition for the total delay in the tandem to exceed x is that there exist $j_1 \leq \ldots \leq j_{M+1} = 0$ such that

$$\sum_{m=1}^{M} (\sigma_{j_m, j_{m+1}}^m - \tau_{j_1+1,0}^1) \geq x \tag{8}$$

Therefore, the probability that the total delay exceeds x is bounded above by the sum over $j_1 \leq \ldots \leq j_{M+1} = 0$ of the probabilities of the above events. This argument is the basis of the upper bound on (9), although the actual proof below uses a slightly simpler approach analogous to that in [4]. The lower bounds on (9) and (10) are obtained as a combination of the following results:

$$\begin{aligned}
\limsup_{x\to\infty} \frac{1}{x} \log \mathbf{P}\left(\sum_{m=1}^{M} D_0^m \geq x\right) &\geq -\theta^0, \\
\liminf_{x\to\infty} \frac{1}{x} \log \mathbf{P}\left(\sum_{m=1}^{M} D_0^m \geq x\right) &\geq -\min_{m=1,\ldots,M} \theta^m,
\end{aligned}$$

where $\theta^0, \ldots, \theta^M$ are as defined in the lemmas above. Bounding the total delay of a customer from below by the total service time required by that customer, we obtain the first claim above.

The second claim comes from choosing $j_1 \leq \ldots \leq j_{M+1} = 0$ to maximize the probability of the event in (8), where this probability is estimated using the Gärtner-Ellis theorem. The details are in Appendix A.

Theorem 1 *Suppose the inter-arrival and service processes satisfy assumptions 1-4 above. Let θ^* be defined as in the lemma above. Then,*

$$\limsup_{x\to\infty} \frac{1}{x} \log \mathbf{P}\left(\sum_{m=1}^{M} D_0^m \geq x\right) = -\theta^* \tag{9}$$

while

$$\liminf_{x\to\infty} \frac{1}{x} \log \mathbf{P}\left(\sum_{m=1}^{M} D_0^m \geq x\right) \geq -\min_{m=1,\ldots,M} \theta^m \tag{10}$$

If $\theta^* = \min\limits_{m=1,\ldots,M} \theta^m$, as is true in particular if assumption 1(a) is satisfied, then

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbf{P}\left(\sum_{m=1}^{M} D_0^m \geq x\right) = -\theta^* \tag{11}$$

The proof is given in Appendix A. We now consider the qualitative behaviour of the system that results in large total delay for a customer. A perusal of the proof shows that there are two distinct scenarios.

Suppose $\theta^* = \min_{1\leq m\leq M} \theta^M$, as is the case if assumption 1(a) holds, i.e., the service times at the different queues are mutually independent. Then we can identify a set of one or more *bottleneck* queues which are responsible for large delays in the following sense. The most likely cause of a given customer suffering a large delay is that a large number of its immediate predecessors require, at one of the bottleneck queues, service times in excess of their inter-arrival times. The number of such predecessors, and their mean inter-arrival and service times, may be found by maximizing $\mathbf{P}(\sigma_{1,n}^m - \tau_{1,n}^1 \geq x)$ over n, $\sigma_{1,n}^m$ and $\tau_{1,n}^1$. A formal characterisation of the most likely path is provided in Appendix B. We note that in this case the tail of the total delay distribution decays at the same exponential rate as the tail of the delay distributions at any of the bottleneck queues. In other words, solving for the delays at the individual queues and considering the worst case is adequate to describe the total delay in the tandem.

Suppose next that $\theta^* = \theta^0$. In that case, the most likely reason that a given customer suffers a large delay is that its own total service requirement is large. More precisely, there are arbitrarily large values of x for which the probability that the delay of a given customer exceeds x is roughly the probability that its own total service requirement exceeds x. In this case, the tail of the delay distribution at any single queue does not capture the tail behaviour of the total sojourn time distribution. The difference between the two cases is demonstrated by the example below.

The last point is an important respect in which the behaviour of tandems differs from that of a single queue. Anantharam [1] shows for GI/G/1 queues that the build-up of large delays can happen in one of two ways. If the service times have exponential tails, then it involves a large number of customers whose inter-arrival and service times differ from their true values according to an exponential twisting whose parameters are explicitly computed. This behaviour is analogous to that of tandems where service times at different queues are independent. If the service times don't have exponential tails, then large delays are caused by the arrival of a *single* customer with large service requirement. In contrast, we see that a single customer can create large delays in tandems even under the assumption of exponential service times, if service requirements at different queues are not independent. We don't consider service times with non-exponential tails in this paper.

Example 1 : Consider two queues in series, fed by Poisson external arrivals of rate λ. Let the service time requirements be *i.i.d.* for different customers, and independent of the arrival process. Suppose the service time required at each queue is exponentially distributed with mean $1/\mu$. Then,

$$\begin{aligned}\Lambda_{T^1}(\theta) &= \log\frac{\lambda}{\lambda-\theta}\cdot 1\{\theta<\lambda\}+\infty\cdot 1\{\theta\geq\lambda\},\\ \Lambda_{S^i}(\theta) &= \log\frac{\mu}{\mu-\theta}\cdot 1\{\theta<\mu\}+\infty\cdot 1\{\theta\geq\mu\},\ i=1,2.\end{aligned}$$

Solving $\Lambda_{S^i}(\theta)+\Lambda_{T^1}(-\theta)=0$, we get 0 and $\mu-\lambda$ as the solutions. Therefore, $\theta^1=\theta^2=\mu-\lambda$.

We now consider two cases, one in which the service times of a customer at the two queues are independent of each other, and another in which they are equal. In the former, we have $\theta^0=\mu$, and so, by Theorem 1,

$$\lim_{x\to\infty}\frac{1}{x}\log\mathbf{P}\left(D_0^1+D_0^2\geq x\right) = \mu-\lambda. \tag{12}$$

In the latter, we see from the definition of the service times that $\theta^0=\mu/2$, and that

$$\lim_{x\to\infty}\frac{1}{x}\log\mathbf{P}\left(S_0^1+S_0^2\geq x\right) = \lim_{x\to\infty}\frac{1}{x}\log\mathbf{P}\left(S_0^1\geq\frac{x}{2}\right) = \frac{\mu}{2}. \tag{13}$$

Since $S_0^1+S_0^2$ is a lower bound on $D_0^1+D_0^2$, it follows from (13) and the upper bound in (9) that

$$\lim_{x\to\infty}\frac{1}{x}\log\mathbf{P}\left(D_0^1+D_0^2\geq x\right) = \frac{\mu}{2}\quad\text{if } \lambda>\frac{\mu}{2}$$

Therefore, if $\lambda>\mu/2$, the tail of the total sojourn time distribution is determined by the total service requirement of a single customer.

Example 2 : The next example we consider is of a model that has gained wide popularity among queueing theorists following the pioneering work of Neuts [10]. We shall assume that the inter-arrival times and the service times at the individual queues are mutually independent and have phase-type distributions. A phase-type distribution is defined as the distribution of the time to absorption in an absorbing Markov chain. More formally, let $\{\phi_t, t \geq 0\}$ be a finite-state Markov chain with transition rate matrix

$$Q = \begin{pmatrix} S & s_0 \\ s_1 & s_2 \end{pmatrix}.$$

The last state of this chain is its unique absorbing state; s_0 is a column vector, s_1 is a row vector of zeros, and $s_2 = 0$. The time to absorption when this Markov chain is started with the initial distribution α on its non-absorbing states is defined to be the phase-type distribution with parameters α and S, denoted $PH(\alpha, S)$. Phase-type distributions constitute a very wide class; they include Erlang and hypergeometric distributions as special cases. Furthermore, by choosing α and S suitably, they can be made to approximate an arbitrary distribution to any desired level of accuracy. For more details, see [12].

Let the inter-arrival times be *i.i.d.*, phase-type, $PH(\alpha^0, S^0)$. Let the service times at the m^{th} queue be *i.i.d.*, $PH(\alpha^m, S^m)$, $m = 1, \ldots, M$. Suppose that the arrivals process and the service processes at the various queues are mutually independent. We first compute the moment-generating function for the inter-arrival time and thereby obtain $\Lambda_{T^1}(\theta)$. Let T^1 denote the first inter-arrival time. Then (see [12] for example) the distribution of T^1 is given by

$$P(T^1 > t) = \alpha^0 e^{S^0 t}\mathbf{1}, \quad t \geq 0, \tag{14}$$

where $\mathbf{1}$ denotes the column vector of ones. Hence,

$$\begin{aligned} E[e^{\theta T^1}] &= 1 + \theta \int_0^\infty e^{\theta t} P(T^1 > t) dt \\ &= \begin{cases} 1 - \theta\alpha^0(\theta I + S^0)^{-1}\mathbf{1}, & \text{if } \theta < \vartheta_0, \\ +\infty, & \text{else.} \end{cases} \end{aligned} \tag{15}$$

The first equality above can be verified by integrating by parts. The second is obtained by substituting for $P(T^1 > t)$ from (14) and integrating. Here, ϑ_0 is defined as

$$\vartheta_0 = \sup\Big\{\theta : \mathrm{sp}\Big(\exp[\theta I + S^0]\Big) < 1\Big\}, \tag{16}$$

where, for a matrix A, $\mathrm{sp}(A)$ denotes its spectral radius. In the case of a positive matrix like $\exp[\theta I + S^0]$, this is the same as its largest eigenvalue, which is real and positive. Similarly, we have for the moment-generating functions of the service times that, for each $m = 1, \ldots, M$,

$$E[e^{\theta S^m}] = \begin{cases} 1 - \theta\alpha^m(\theta I + S^m)^{-1}\mathbf{1}, & \text{if } \theta < \vartheta_m, \\ +\infty, & \text{else,} \end{cases} \tag{17}$$

where

$$\vartheta_m = \sup\Big\{\theta : \mathrm{sp}\Big(\exp[\theta I + S^m]\Big) < 1\Big\}. \tag{18}$$

Assumption 1(a) is satisfied by our definition of the arrival and service processes. We shall also suppose that the queues are stable, i.e., that Assumption 2 holds. Assumption 3 holds trivially for *i.i.d.* inter-arrival and service times; the quantities on the right hand side of the equalities don't depend on n, and we simply have

$$\Lambda_{T^1}(\theta) = \log E[e^{\theta T^1}], \quad \Lambda_{S^m}(\theta) = \log E[e^{\theta S^m}],\ m = 1, \ldots, M. \tag{19}$$

Furthermore, ϑ_0 and ϑ_m, $m = 1, \ldots, M$, defined in (16) and (18), are strictly positive and Assumption 4 is satisfied. Therefore, Theorem 1 holds and can be used to compute the tail of the sojourn time distribution in this system. The calculation of the quantities θ^m defined in Lemma 1 is straightforward from (15)-(19) using standard numerical techniques and the fact that Λ_{T^1} and Λ_{S^m}, $m = 1, \ldots, M$ are convex functions.

We conclude this section by analyzing a traffic model that is very popular for ATM traffic and working out the tail of the sojourn time distribution.

Example 3 : Suppose that the inter-arrival times and the service times at each queue are Markov modulated. The inter-arrival times are modulated by a Markov chain with finite state space $\mathcal{X}^0$ and transition probability matrix P^0. The jump times of this matrix are the arrival epochs, and we allow jumps from a state to itself. If the Markov chain is in state $i \in \mathcal{X}^0$ immediately after an arrival, then the time to the next arrival is deterministic and equal to t_i, where the t_i, $i \in \mathcal{X}^0$ are given constants. Service times at the k^{th} queue are governed by a finite state Markov chain, with state space $\mathcal{X}^k$ and transition probability matrix P^k. The jump times of this Markov chain are the service completion times; we allow jumps from a state to itself. Service times are described by a set of constants $\{s_j^k\}, k = 1, \ldots, M, j \in \mathcal{X}^k$. The service time of an arbitrary customer at the k^{th} queue is s_j^k if the modulating Markov chain is in state $j \in \mathcal{X}^k$ when this customer begins service at queue k. We assume that the Markov chains modulating arrival times and service times at the various queues are mutually independent.

We now compute the functions Λ_{T^1} and $\Lambda_{S^m}, m = 1, \ldots, M$, for this model. These calculations are not new (see [8] for example), but are presented here for completeness. Their use in this paper to compute sojourn times in tandems is new. We denote the state of the modulating Markov chain at queue k when customer j begins service at this queue by x_j^k. Consider the service times at the first queue, conditional on the modulating Markov chain being in state i when the first customer begins service. Define

$$f_1(i, n) = E\left[e^{\theta(S_1^1 + \ldots + S_n^1)} | x_1^1 = i\right], \tag{20}$$

and let $\mathbf{f}_1(n)$ denote the vector $\{f_1(i, n), i \in \mathcal{X}^1\}$. Then, by conditioning on the state when the first customer finishes service, we get

$$f_1(i, n) = e^{\theta s_i^1} \sum_{j \in \mathcal{X}^1} P_{ij}^1 f_1(j, n-1). \tag{21}$$

Define $R^1(\theta) = \text{diag}\Big(\exp(\theta s^1)\Big)P^1$, i.e., $R^1(\theta)$ is the matrix with ij^{th} entry $R^1_{ij}(\theta) = \exp(\theta s^1_i)P^1_{ij}$. Then it follows from (21) that

$$\mathbf{f}_1(n) = \Big[R^1(\theta)\Big]^n \mathbf{f}_1(0) = \Big[R^1(\theta)\Big]^n \mathbf{1},$$

where $\mathbf{1}$ denotes the vector of ones. Therefore, if π^1 denotes the stationary distribution of the modulating Markov chain at queue 1, we have

$$E_{\pi^1}\Big[\exp\{\theta(S^1_1 + \ldots + S^1_n)\}\Big] = \pi^1 \Big[R^1(\theta)\Big]^n \mathbf{1}.$$

Consequently,

$$\Lambda_{S^1}(\theta) \triangleq \lim_{n\to\infty} \frac{1}{n} \log E_{\pi^1}\Big[\exp\{\theta(S^1_1 + \ldots + S^1_n)\}\Big] = \log\Big[\text{sp}(R^1(\theta))\Big],$$

where, for a matrix A, $\text{sp}(A)$ denotes its spectral radius. Since $R^1(\theta)$ is a non-negative matrix, its spectral radius is equal to its largest eigenvalue, which is real and positive. Likewise, for each of the other queues, $m = 2, \ldots, M$, we have

$$\Lambda_{S^m}(\theta) = \log\left[\text{sp}(R^m(\theta))\right], \quad \text{where} \quad R^m(\theta) = \text{diag}\Big(\exp(\theta s^m)\Big)P^m. \tag{22}$$

Here, P^m denotes the transition probability matrix of the modulating Markov chain at queue m, and s^m_i denotes the (deterministic) service requirement of a customer who starts service at this queue when its modulating chain is in state $i \in \mathcal{X}^m$. Similarly, we have for the limiting cumulant generating function of the arrival times, the formula

$$\Lambda_{T^1}(\theta) = \log\Big[\text{sp}(R^0(\theta))\Big], \quad \text{where} \quad R^0(\theta) = \text{diag}\Big(\exp(\theta t)\Big)P^0. \tag{23}$$

Here P^0 is the transition probability matrix of the Markov chain modulating arrivals, and t_i, $i \in \mathcal{X}^0$ is the inter-arrival time when the inter-arrival period starts with the Markov chain modulating arrivals in state i.

The arrival and service processes defined above satisfy Assumption 1(a). We also saw that the limits $\Lambda_{T^1}(\theta)$ and $\Lambda_{S^m}(\theta)$ exist for all $\theta \in \mathbb{R}$, and are as defined in (22) and (23). In fact, these limits are finite for all $\theta \in \mathbb{R}$ and so Assumptions 3 and 4 are satisfied. If, in addition, the stability condition of Assumption 2 holds, then Theorem 1 is applicable and describes the tail of the sojourn time distribution. Calculating the quantities θ^m, $m = 1, \ldots, M$, defined in Lemma 1 is straightforward from (22) and (23) using standard numerical methods and the fact that Λ_{T^1} and Λ_{S^m}, $m = 1, \ldots, M$ are convex functions.

4 Conclusion

The problem of estimating packet loss probabilities in queueing networks has recently received considerable attention, motivated in large part by applications to broadband communication networks. A related problem, that of estimating the probability of large end-to-end packet delays, has been relatively neglected. In this paper, we obtain a description of this probability for a tandem queueing model, under mild conditions on the arrival and service processes. We also derive a simple expression for the total delay in (5), which could be useful in studying arrival and service process models other than the one we have considered here. From the viewpoint of applications, it would be of interest to study models with multiple classes of customers and priority service schemes at the individual queues. Extending the results of this paper to such a model remains an open problem.

A Proof of Theorem 1

Let $\theta \in (0, \theta^*)$. Then, for all $m \in \{1, \ldots, M\}$, $\Lambda_{T^1}(-\theta) + \Lambda_{S^m}(\theta) < 0$ by (6) and the definition of θ^*. In particular, $\Lambda_{T^1}(-\theta)$ and $\Lambda_{S^m}(\theta)$ are finite. Let $\epsilon > 0$ be given. Then, by assumption 2, there are finite positive constants c_m, c^m, k_1 and k^1 such that

$$c_m e^{n(\Lambda_{S^m}(\theta)-\epsilon)} \leq E[\exp(\theta\sigma^m_{1,n})] \leq c^m e^{n(\Lambda_{S^m}(\theta)+\epsilon)} \quad \forall\, n \geq 0 \tag{24}$$

$$k_1 e^{n(\Lambda_{T^1}(\theta)-\epsilon)} \leq E[\exp(\theta\tau^1_{1,n})] \leq k^1 e^{n(\Lambda_{T^1}(\theta)+\epsilon)} \quad \forall\, n \geq 0 \tag{25}$$

The constants c_m, c^m, k_1, k^1 depend on θ, ϵ but this is suppressed in the notation.

Suppose first that assumption 1(a) holds. Since the arrival and service processes are assumed to be mutually independent and stationary, we have from (5) that, for all $\theta \geq 0$,

$$\begin{aligned} E\left[\exp(\theta \sum_{m=1}^{M} D^m_0)\right] &= E\left[\max_{a_1,\ldots,a_M \geq 0} \exp(-\theta\tau^1_{1,a_1+\ldots+a_M}) \cdot \prod_{m=1}^{M} \exp(\theta\sigma^m_{0,a_m})\right] \\ &\leq \sum_{a_1,\ldots,a_M \geq 0} E\left[\exp(-\theta\tau^1_{1,a_1+\ldots+a_M})\right] \cdot \prod_{m=1}^{M} E\left[\exp(\theta\sigma^m_{0,a_m})\right] \end{aligned} \tag{26}$$

Thus, by (24) and (25), for all $\theta \in (0, \theta^*)$,

$$\begin{aligned} & E\left[\exp(\theta \sum_{m=1}^{M} D^m_0)\right] \\ & \leq \sum_{a_1,\ldots,a_M \geq 0} k^1 e^{(a_1+\ldots+a_M)(\Lambda_{T^1}(-\theta)+\epsilon)} \cdot \prod_{m=1}^{M} c^m e^{(a_m+1)(\Lambda_{S^m}(\theta)+\epsilon)} \end{aligned}$$

$$= \quad \hat{c} \prod_{m=1}^{M} \sum_{a_m=0}^{\infty} \exp\left[a_m\Big(\Lambda_{S^m}(\theta) + \Lambda_{T^1}(-\theta) + 2\epsilon\Big)\right] \tag{27}$$

where $0 < \hat{c} < +\infty$. Since $\theta \in (0, \theta^*)$, observe from Lemma 1 that we can choose $\epsilon > 0$ such that

$$\Lambda_{T^1}(-\theta) + \Lambda_{S^m}(\theta) + 2\epsilon < 0 \quad \text{for all} \quad m \in \{1, \ldots, M\}.$$

Therefore, by (27), $E[\exp(\theta \sum_{m=1}^{M} D_0^m)] \le c$ for some finite constant c. Hence, by Chernoff's inequality,

$$\mathbf{P}\left(\sum_{m=1}^{M} D_0^m \ge x\right) \;\le\; e^{-\theta x} E\left[\exp(\theta \sum_{m=1}^{M} D_0^m)\right] \;\le\; c e^{-\theta x}$$

Since the above holds for all $0 < \theta < \theta^*$, we get

$$\limsup_{x\to\infty} \frac{1}{x} \log \mathbf{P}\left(\sum_{m=1}^{M} D_0^m \ge x\right) \;\le\; -\theta^* \tag{28}$$

Suppose next that assumption 1(b) holds. Let $\theta \ge 0$. Observe from (5) and the stationarity of the arrival and service processes that

$$\begin{aligned}
& E\left[\exp(\theta \sum_{m=1}^{M} D_0^m)\right] \\
& = E\left[\max_{0=a_0\le\ldots\le a_M} \exp(-\theta \tau^1_{1,a_M}) \cdot \exp\Big(\theta \sum_{m=1}^{M} \sigma^m_{a_{m-1},a_m}\Big)\right] \\
& \le \sum_{0=a_0\le\ldots\le a_M} E\left[\exp(-\theta \tau^1_{1,a_1+\ldots+a_M})\right] \cdot E\left[\exp\Big(\theta \sum_{m=1}^{M} \sigma^m_{a_{m-1},a_m}\Big)\right]
\end{aligned} \tag{29}$$

In the last inequality above, we have used the fact that the service process is independent of the arrival process by assumption 1(b). Now, by the non-negativity of the service times and their independence from customer to customer, we have

$$\begin{aligned}
& E\left[\exp\Big(\theta \sum_{m=1}^{M} \sigma^m_{a_{m-1},a_m}\Big)\right] \\
\le\; & \prod_{m=1}^{M} E\left[\exp(\theta \sigma^m_{a_{m-1}+1,a_m-1})\right] \cdot \prod_{m=0}^{M} E\left[\exp\Big(\theta(S^1_{a_m} + \ldots + S^M_{a_m}\Big)\right] \\
=\; & \exp\left[\sum_{m=1}^{M} (a_m - a_{m-1} - 1)\Lambda_m(\theta)\right] \prod_{m=0}^{M} E\left[\exp\Big(\theta(S^1_{a_m} + \ldots + S^M_{a_m}\Big)\right]
\end{aligned} \tag{30}$$

If $\theta \in (0, \theta^*)$, then it follows from the definition of θ^* that

$$E\left[\exp\Big(\theta(S^1_{a_m} + \ldots + S^M_{a_m}\Big)\right] = c$$

for a finite constant, c. Combining this fact with (25), (29) and (30), we get, for $\theta \in (0, \theta^*)$ and any $\epsilon > 0$,

$$\begin{aligned} & E\left[\exp(\theta \sum_{m=1}^{M} D_0^m)\right] \\ \leq \quad & \sum_{0=a_0 \leq \ldots \leq a_M} \hat{c} \prod_{m=1}^{M} \exp\left[(a_m - a_{m-1})\Big(\Lambda_{T^1}(-\theta) + \Lambda_{S^m}(\theta) + \epsilon\Big)\right] \\ = \quad & \hat{c} \prod_{m=1}^{M} \sum_{b_m=0}^{\infty} \exp\left[b_m\Big(\Lambda_{T^1}(-\theta) + \Lambda_{S^m}(\theta) + \epsilon\Big)\right] \end{aligned} \tag{31}$$

where $\hat{c}$ is a finite constant. Since $\theta \in (0, \theta^*)$, observe from Lemma 1 that we can choose $\epsilon > 0$ such that

$$\Lambda_{T^1}(-\theta) + \Lambda_{S^m}(\theta) + \epsilon < 0 \quad \text{for all} \quad m \in \{1, \ldots, M\}.$$

Therefore, by (31), $E[\exp(\theta \sum_{m=1}^{M} D_0^m)] \leq c$, for some finite constant c. Hence, by Chernoff's inequality,

$$\mathbf{P}\left(\sum_{m=1}^{M} D_0^m \geq x\right) \;\leq\; e^{-\theta x} E\left[\exp(\theta \sum_{m=1}^{M} D_0^m)\right] \;\leq\; c e^{-\theta x}$$

Since the above holds for all $0 < \theta < \theta^*$, we get

$$\limsup_{x \to \infty} \frac{1}{x} \log \mathbf{P}\left(\sum_{m=1}^{M} D_0^m \geq x\right) \;\leq\; -\theta^* \tag{32}$$

We have, in (28) and (32), upper bounds on $\mathbf{P}\left(\sum_{m=1}^{M} D_0^m \geq x\right)$ under assumptions 1(a) and 1(b) respectively. We now turn to estimating lower bounds on the probability of large queue sizes.

The convex conjugate, $\Lambda^*(\cdot)$, of $\Lambda : I\!R \to [-\infty, +\infty]$ is defined as

$$\Lambda^*(x) = \sup_{\theta \in I\!R} [\theta x - \Lambda(\theta)]$$

Suppose $\theta > \min_{1 \leq m \leq M} \theta^m$. Let m be such that $\theta > \theta^m$. Define $\Lambda(\theta) = \Lambda_{S^m}(\theta) + \Lambda_{T^1}(-\theta)$. Since $\Lambda_{S^m}(\theta) > -\infty$ and $\Lambda_{T^1}(\theta) > -\infty$ for all θ, Λ is well-defined. Now, by assumptions 1 and 3,

$$\Lambda(\theta) = \lim_{n \to \infty} \frac{1}{n} \log E\left[\exp(\theta(\sigma^m_{1,n} - \tau^1_{1,n}))\right]$$

Also, by assumption 3, Λ is essentially smooth and lower semicontinuous. Hence, by the Gärtner-Ellis theorem (Theorem 2.3.6 in [5]), the process $\{(\sigma_{1n}^m - \tau_{1n}^1)/n\}$ satisfies a large deviations principle with rate function Λ^* which is the convex conjugate of Λ. In other words, for every closed set F and every open set G,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left(\frac{\sigma_{1n}^m - \tau_{1n}^1}{n} \in F\right) \quad \leq \quad -\inf_{x\in F} \Lambda^*(x) \tag{33}$$

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left(\frac{\sigma_{1n}^m - \tau_{1n}^1}{n} \in G\right) \quad \geq \quad -\inf_{x\in G} \Lambda^*(x) \tag{34}$$

Fix $\alpha > 0$. Given $x > 0$, define $n = x/\alpha$. Taking $j_1 = \ldots = j_m = -n$ and $j_{m+1} = \ldots = j_{M+1} = 0$ in (5), and using the stationarity of $\{S_i^m\}$ and $\{T_i^1\}$, we get

$$\mathbf{P}\left(\sum_{m=1}^{M} D_0^m > x\right) \;\geq\; \mathbf{P}\left(\sigma_{1n}^m - \tau_{1n}^1 > n\alpha\right)$$

Hence, taking $G = (\alpha, \infty)$ in (34), we see that

$$\begin{aligned}
\liminf_{x\to\infty} \frac{1}{x} \log \mathbf{P}\left(\sum_{m=1}^{M} D_0^m > x\right) \quad &\geq \quad \frac{1}{\alpha} \liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left(\frac{\sigma_{1n}^m - \tau_{1n}^1}{n} > \alpha\right) \\
&= \quad -\frac{1}{\alpha} \inf_{z>\alpha} \Lambda^*(z) \\
&\geq \quad -(1+\epsilon)\frac{\Lambda^*((1+\epsilon)\alpha)}{(1+\epsilon)\alpha} \qquad \forall\, \epsilon > 0
\end{aligned} \tag{35}$$

Since the above inequality holds for all $\alpha > 0$, we have

$$\liminf_{x\to\infty} \frac{1}{x} \log \mathbf{P}\left(\sum_{m=1}^{M} D_0^m > x\right) \geq -\inf_{\alpha>0} \frac{\Lambda^*(\alpha)}{\alpha} \tag{36}$$

By assumption 3, namely, that Λ is essentially smooth and lower semicontinuous, $\Lambda(\cdot)$ and $\Lambda^*(\cdot)$ are convex duals (see, for example, [11]). Therefore,

$$\Lambda(\theta) = \sup_{\alpha\in I\!\!R} [\theta\alpha - \Lambda^*(\alpha)]$$

Since $\theta > \theta^m$ by the choice of θ and m, we see from (7) that $\Lambda(\theta) > 0$. Hence, there exists $\alpha^* \in I\!\!R$ such that $\theta\alpha^* - \Lambda^*(\alpha^*) > 0$. Note that Λ^* is non-negative because $\Lambda(0) = 0$. Also, $\theta > 0$ because $\theta^m > 0$ as noted earlier. It follows that $\alpha^* > 0$. Consequently,

$$\inf_{\alpha>0} \frac{\Lambda^*(\alpha)}{\alpha} \leq \frac{\Lambda^*(\alpha^*)}{\alpha^*} < \theta$$

Since $\theta > \min_{1\leq m\leq M} \theta^m$ is arbitrary, we conclude from (36) that

$$\liminf_{x\to\infty} \frac{1}{x} \log \mathbf{P}\left(\sum_{m=1}^{M} D_0^m > x\right) \geq - \min_{1\leq m\leq M} \theta^m \tag{37}$$

Next, by taking $j_1 = \ldots = j_{M+1} = 0$ in (5), we see that

$$\sum_{m=1}^{M} D_0^m \;\geq\; \sum_{m=1}^{M} S_0^m$$

Therefore, if $\theta > \theta^0 \geq 0$, we have

$$E\left[\exp(\theta \sum_{m=1}^{M} D_0^m)\right] \geq E\left[\exp(\theta \sum_{m=1}^{M} S_0^m)\right] = +\infty$$

where the equality holds by definition of θ^0. It is an immediate consequence of the above that, for all $\epsilon > 0$,

$$\limsup_{x\to\infty} e^{(\theta+\epsilon)x}\mathbf{P}\left(\sum_{m=1}^{M} D_0^m > x\right) \;=\; +\infty,$$

as can be shown by contradiction. Therefore,

$$\limsup_{x\to\infty} \frac{1}{x} \log \mathbf{P}\left(\sum_{m=1}^{M} D_0^m > x\right) \;\geq\; -\theta - \epsilon$$

Since $\theta > \theta^0$ and $\epsilon > 0$ are arbitrary, we conclude that

$$\limsup_{x\to\infty} \frac{1}{x} \log \mathbf{P}\left(\sum_{m=1}^{M} D_0^m > x\right) \;\geq\; -\theta^0 \tag{38}$$

The inequality in (37) holds *a fortiori* if lim inf is replaced by lim sup. Together with (38), this implies that

$$\limsup_{x\to\infty} \frac{1}{x} \log \mathbf{P}\left(\sum_{m=1}^{M} D_0^m > x\right) \;\geq\; -\theta^* \tag{39}$$

Combining (28) and (32), which hold under assumption 1(a) and 1(b) respectively, with (39), we obtain the first claim of the theorem. The second claim is given by (37). The last claim of the theorem follows from the first two, and the definition of $\theta^0, \ldots, \theta^M$ and θ^*.

B Conditional limit theorem

In this section, we describe the most likely path leading to the build-up of large delays, under somewhat stronger assumptions than in Section 3. Let T and S be defined as in Section 2. For $n > 0$ and $t > 0$, define

$$\hat{T}_n^1(t) = \frac{1}{n} \sum_{k=0}^{\lfloor nt \rfloor} T_k^1, \qquad \hat{S}_n^m(t) = \frac{1}{n} \sum_{k=0}^{\lfloor nt \rfloor} S_k^m, \; m = 1, \dots, M, \tag{40}$$

where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. Clearly, $\hat{T}^1$ and $\hat{S}^m$ are piecewise constant, with jumps when t is an integer multiple of $1/n$. By interpolating linearly between the jump times, we obtain polygonal approximations that we denote $\tilde{T}_n^1$, $\tilde{S}_n^m$ respectively. Thus

$$\tilde{T}_n^1(t) = \left(\frac{\lfloor nt \rfloor + 1}{n} - t \right) \hat{T}_n^1(\lfloor nt \rfloor) + \left(t - \frac{\lfloor nt \rfloor}{n} \right) \hat{T}_n^1(\lfloor nt \rfloor + 1), \tag{41}$$

and $\tilde{S}_n^m$, $m = 1, \dots, M$, are defined analogously. Observe from (5) and the stationarity of the processes involved that the probability of the event $\{\sum_{m=1}^M D_0^m \geq n\}$ is the same as that of the event that there exists $t > 0$ such that

$$\sup_{0=t_1 \leq \dots \leq t_{M+1}=t} \sum_{m=1}^M [\tilde{S}_n^m(t_{m+1}) - \tilde{S}_n^m(t_m)] - [\tilde{T}_n^1(t) - \tilde{T}_n^1(0)] \;\geq\; 1$$

Indeed, the latter event is merely the former shifted in time by nt. We shall assume that the process $\{\tilde{T}_n^1(t), \tilde{S}_n^m(t), m = 1, \dots, M, t \in [0,1]\}$ satisfies a sample path large deviations principle, and use this to derive the conditional probability of paths leading to large delays.

Let $\mathcal{L}^{M+1}$ denote the space of L^∞ functions $\phi : [0,1] \to I\!R^{M+1}$, with non-decreasing components $\phi_0, \dots, \phi_M$, and having $\phi(0) = \mathbf{0}$. Let this space be endowed with the topology induced by the supremum norm,

$$\|\phi\|_\infty = \max_{i=0,\dots,M} \sup_{t \in [0,1]} |\phi_i(t)|,$$

where ϕ_i denotes the i^{th} component of ϕ. Let $\mathcal{A}^{M+1}$ denote the subspace $\mathcal{L}^{M+1}$ consisting of absolutely continuous functions.

We suppose in the following that assumptions 1(a), 2, 3 and 4 of Section 3 are satisfied, and that there is a *unique* bottleneck queue, m. In other words,

$$\exists \delta > 0 \;\; : \;\; \Lambda_{T^1}(-\theta) + \Lambda_{S^k}(\theta) < 0 \quad \forall \theta \in (0, \theta^m + \delta), \text{ if } k \neq m, \tag{42}$$

$$\forall \delta > 0 \;\; : \;\; \Lambda_{T^1}(-\theta^m - \delta) + \Lambda_{S^m}(\theta^m + \delta) > 0, \tag{43}$$

where θ^m is as defined in Lemma 1. We assume in addition that Λ_{S^m} is finite in a neighbourhood of θ^m. It follows from this and the analyticity of cumulant generating functions in the interior of the region where they are finite, that

$$\Lambda_{T^1}(-\theta^m) + \Lambda_{S^m}(\theta^m) = 0, \quad s \triangleq \Lambda'_{S^m}(\theta^m) - \Lambda'_{T^1}(-\theta^m) \text{ exists and is finite.} \tag{44}$$

Let the function $\Phi \in \mathcal{A}^{M+1}$ be defined as follows:

$$\Phi(0) = \mathbf{0}, \quad \forall t \in (0,1),\ \dot{\Phi}_k(t) = \begin{cases} \Lambda'_{T^1}(-\theta^m), & \text{if } k = 0, \\ \Lambda'_{S^m}(\theta^m), & \text{if } k = m, \\ \Lambda'_{S^k}(0), & \text{if } k \neq 0, m, \end{cases} \quad . \tag{45}$$

Given $\epsilon > 0$, let $B^\epsilon(\Phi)$ denote the open ϵ-ball around Φ, i.e.,

$$B^\epsilon(\Phi) = \{\psi \in \mathcal{L}^{M+1} : \|\Phi - \psi\|_\infty < \epsilon\}. \tag{46}$$

Define the function $D : \mathcal{L}^{M+1} \to I\!R$ by

$$D(\phi) = \sup_{0=t_1 \leq \ldots \leq t_{M+1}=1} \sum_{m=1}^{M} [\phi_m(t_{m+1}) - \phi_m(t_m)] - \phi_0(1) \tag{47}$$

Finally, for notational convenience, let $\tilde{X}_n = (\tilde{X}_n^0, \ldots, \tilde{X}_n^M)$ denote the vector $(\tilde{T}_n^1, \tilde{S}_n^1, \ldots, \tilde{S}_n^M)$. We are now ready to state our main result.

Theorem 2 *Suppose the assumptions above hold and also that $\{\tilde{X}_n\}$ satisfies the large deviation principle (LDP) (see [5]) in $\mathcal{L}^{M+1}$ with rate function I given by:*

$$I(\phi) = \begin{cases} \int_0^1 \left[\Lambda^*_{T^1}(\dot{\phi}_0(s)) + \sum_{k=1}^M \Lambda^*_{S^k}(\dot{\phi}_k(s))\right] ds, & \text{if } \phi \in \mathcal{A}^{M+1}, \\ +\infty, & \text{else,} \end{cases} \tag{48}$$

where $\Lambda^*_{T^1}$, $\Lambda^*_{S^k}$ are the convex conjugates of Λ_{T^1}, Λ_{S^k} respectively, as defined in Appendix A. Then, for all $\epsilon > 0$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left(\tilde{X}_n \notin B^\epsilon(\Phi) | D(\tilde{X}_n) \geq s\right) < 0,$$

for s, Φ, D defined as in (44), (45), (47) respectively.

Remarks : The above theorem says that, if the tandem is started empty and the n^{th} customer suffers a delay exceeding ns, then the most likely way this could have happened is as follows. The scaled process of arrival times and service requirements described by $\tilde{X}_n$ lies close to Φ. The probability that it deviates from Φ by more than ϵ, for any $\epsilon > 0$, decays exponentially in n. On the path Φ, the inter-arrival time of each customer is $\Lambda'_{T^1}(-\theta^m)$; its service time at the m^{th} queue is $\Lambda'_{S^m}(\theta^m)$, and its service time at any other queue $k \neq m$ is $ES^k = \Lambda'_{S^k}(0)$, the mean service time at that queue. In order to complete the description of the most likely path leading to large delays, it remains to show that, in order for an arbitrary customer to suffer a delay in excess of ns, the last time before its arrival that the tandem was empty was approximately n arrivals earlier. This is shown in the next lemma below.

Proof : Define $H^\epsilon = \{\phi \in \mathcal{L}^{M+1}\backslash B^\epsilon(\Phi) : D(\phi) \geq s\}$ to be the set of paths outside an ϵ-neighbourhood of Φ which satisfy the constraint $D(\phi) \geq s$. We want to show that

$$\inf_{\phi \in H^\epsilon} I(\phi) > I(\Phi) \tag{49}$$

If the infimum on the left is $+\infty$, then we are done. If not, we proceed as follows. Since Λ_{S^k} is finite in $[-\epsilon, \epsilon]$ for some $\epsilon > 0$ and all $k \in \{1, \dots, M\}$, we get

$$\begin{aligned} \lim_{|x|\to\infty} \Lambda^*_{S^k}(x) &= \lim_{|x|\to\infty} \sup_\theta \, [\theta x - \Lambda_{S^k}(\theta)] \\ &\geq \lim_{|x|\to\infty} |\epsilon x| - \max\{\Lambda_{S^k}(\epsilon), \Lambda_{S^k}(-\epsilon)\} \; = \; \infty. \end{aligned} \tag{50}$$

Likewise, $\lim_{|x|\to\infty} \Lambda^*_{T^1}(x) = +\infty$. Now, given $\delta > 0$, we claim that

$$\exists B < \infty : I(\psi) < \inf_{\phi \in H^\epsilon} I(\phi) + \delta \quad \Rightarrow \quad \psi_k(1) < B \;\; \forall \, k \in \{0, \dots, M\}. \tag{51}$$

Indeed, since all the Λ^* are non-negative and convex, we have from (48) that

$$I(\psi) \geq \int_0^1 \Lambda^*(\dot{\psi}_k(s))ds \geq \Lambda^*(\psi_k(1)) \quad \forall \, k \in \{0, \dots, M\}.$$

The last inequality is due to Jensen's inequality and the fact that $\psi(0) = \mathbf{0}$. Now, given $\inf_{\phi \in H^\epsilon} I(\phi)$, which is finite, and $\delta > 0$, it follows from (50) that we can choose B sufficiently large that

$$\psi_k(1) > B \quad \Rightarrow \quad \Lambda^*(\psi_k(1)) > \inf_{\phi \in H^\epsilon} I(\phi) + \delta,$$

completing the proof of the assertion in (51).

Let $\tilde{\psi}^1, \tilde{\psi}^2, \ldots \in H^\epsilon$ be such that

$$\lim_{n\to\infty} I(\tilde{\psi}^n) = \inf_{\phi \in H^\epsilon} I(\phi).$$

Since the right hand side above is assumed to be finite, we may, without loss of generality, restrict attention to a subsequence on which $I(\tilde{\psi}^n)$ is finite. Then, $\tilde{\psi}^n \in \mathcal{A}^{M+1}$, and so the supremum in (47) is achieved, i.e., there exist $0 = t_1^n \leq \ldots \leq t_{M+1}^n = 1$ such that

$$\sum_{m=1}^{M} [\tilde{\psi}^n_m(t^n_{m+1}) - \tilde{\psi}^n_m(t^n_m)] - \tilde{\psi}^n_0(1) \geq s. \tag{52}$$

Now, given $\delta > 0$, we have $I(\tilde{\psi}^n) < \inf_{\phi \in H^\epsilon} I(\phi) + \delta$ for all n sufficiently large. Therefore, by (51), $\tilde{\psi}^n_k(1) \leq B$ for all such n and all $k \in \{0, \ldots, M\}$. Since $\tilde{\psi}^n(0) = \mathbf{0}$ and $\tilde{\psi}^n$ is non-decreasing, we get that $\tilde{\psi}^n \in [0, B]^{M+1}$ for all $t \in [0, 1]$, if n is sufficiently large. Since $[0, B]^{M+1}$ is compact, as is the set of $(M+1)$-tuples $\{(x_1, \ldots, x_{M+1}) : 0 = x_1 \leq \ldots \leq x_{M+1} = 1\}$, there exists a subsequence γ of $I\!N$ such that the limits

$$t_k \triangleq \lim_{n \to \infty} t_k^{\gamma(n)}, \quad \tilde{\psi}(t_k) \triangleq \lim_{n \to \infty} \tilde{\psi}^{\gamma(n)}(t_k^{\gamma(n)}) \tag{53}$$

exist for each k in $\{0, \ldots, M\}$. Define $\psi \in \mathcal{A}^{M+1}$ as follows:

$$\psi(0) = \mathbf{0}, \qquad \dot{\psi}_0(t) = \tilde{\psi}^0(1), \; t \in (0,1),$$

and, for $k \in \{1, \ldots, M\}$,

$$\dot{\psi}_k(t) = \begin{cases} [\tilde{\psi}_k(t_{k+1}) - \tilde{\psi}_k(t_k)]/[t_{k+1} - t_k], & \text{if } t \in (t_k, t_{k+1}), \\ ES_k, & \text{if } t \notin (t_k, t_{k+1}). \end{cases}$$

Then,

$$\begin{aligned} D(\psi) &\geq \sum_{k=1}^{M} [\psi_k(t_{k+1}) - \psi_k(t_k)] - \psi_0(1) \\ &= \lim_{n \to \infty} \sum_{k=1}^{M} [\tilde{\psi}_k^{\gamma(n)}(t_{k+1}^{\gamma(n)}) - \tilde{\psi}_k^{\gamma(n)}(t_k^{\gamma(n)})] - \tilde{\psi}_0^{\gamma(n)}(1) \;\geq\; s, \end{aligned} \tag{54}$$

where the equality follows from the definition of ψ, and the last inequality from (52). Observe from the non-negativity and convexity of Λ^* that, by (48),

$$I(\tilde{\psi}^n) \geq \sum_{k=1}^{M} (t_{k+1}^n - t_k^n) \Lambda^*_{S^k} \left(\frac{\tilde{\psi}_k^n(t_{k+1}^n) - \tilde{\psi}_k^n(t_k^n)}{t_{k+1}^n - t_k^n} \right) + \Lambda^*_{T^1}(\tilde{\psi}_0^n(1)), \tag{55}$$

where we take $0 \cdot \Lambda^*(0/0) = 0$. But,

$$I(\psi) = \sum_{k=1}^{M} (t_{k+1} - t_k) \Lambda^*_{S^k} \left(\frac{\psi_k(t_{k+1}) - \psi_k(t_k)}{t_{k+1} - t_k} \right) + \Lambda^*_{T^1}(\psi_0(1)), \tag{56}$$

since $\Lambda^*_{S^k}(ES_k) = 0$, see [5]. Now, $x\Lambda^*_{S^k}(y/x) = \sup_\theta[\theta y - x\Lambda_{S^k}(\theta)]$, is the supremum of continuous functions of (x, y), and is therefore a lower semicontinuous function of (x, y). Hence, it follows from (53), (55) and (56) that

$$I(\psi) \leq \liminf_{n \to \infty} I(\tilde{\psi}^{\gamma(n)}) = \inf_{\phi \in H^\epsilon} I(\phi). \tag{57}$$

Finally, by the strict convexity of Λ^*, ([5]), the inequality above is strict unless $\liminf_{n\to\infty} \|\tilde{\psi}^{\gamma(n)} - \psi\|_\infty = 0$. But, if the last equality holds, then $\psi \notin B^\epsilon(\Phi)$, because $\tilde{\psi}^n \notin B^\epsilon(\Phi)$ for any n, and $B^\epsilon(\Phi)$ is open. Thus, we have,

$$I(\psi) = \inf_{\phi \in H^\epsilon} I(\phi) \quad \Rightarrow \quad \psi \notin B^\epsilon(\Phi) \tag{58}$$

We now show that $I(\psi) \geq I(\Phi)$, where Φ was defined in (45), with equality only if $\psi \equiv \Phi$. By (56),

$$\begin{aligned} I(\psi) &= \sup_{\lambda_0,\ldots,\lambda_M} \sum_{k=1}^{M} \Big[\lambda_k\Big(\psi_k(t_{k+1}) - \psi_k(t_k)\Big) - (t_{k+1}-t_k)\Lambda_{S^k}(\lambda_k)\Big] + \lambda_0\psi_0(1) - \Lambda_{T^1}(\lambda_0) \\ &\geq \sum_{k=1}^{M} [\theta^m(\psi_k(t_{k+1}) - \psi_k(t_k)) - (t_{k+1}-t_k)\Lambda_{S^k}(\theta^m)] - \theta^m\psi_0(1) - \Lambda_{T^1}(-\theta^m) \\ &\geq \theta^m s - \sum_{k=1}^{M}(t_{k+1}-t_k)\Big[\Lambda_{S^k}(\theta^m) - \Lambda_{T^1}(-\theta^m)\Big]. \end{aligned} \tag{59}$$

Here, θ^m is as defined in Lemma 1, and the last inequality follows from (54) and the fact that $\sum_{k=1}^{M}(t_{k+1}-t_k) = 1$. Now, $\Lambda^*(\Lambda'(\theta)) = \theta\Lambda'(\theta) - \Lambda(\theta)$, while $\Lambda^*_{S^k}(ES_k) = 0$, as noted above. Hence, by (45),

$$I(\Phi) = \theta^m(\Lambda'_{S^m}(\theta^m) - \Lambda'_{T^1}(-\theta^m)) - \Lambda_{S^m}(\theta^m) - \Lambda_{T^1}(-\theta^m). \tag{60}$$

Now, by (59) and (42), $I(\psi) \geq \theta^m s$, with equality only if $t_{k+1} - t_k = 0$ for all $k \neq m$, whereas, by (60) and (44), $I(\Phi) = \theta^m s$. Therefore, we see from the definition of ψ and Φ that $I(\psi) \geq I(\Phi)$, with equality only if $\psi = \Phi$. It now follows from (57) and (58) that $I(\Phi)$ is strictly less than $\inf_{\phi\in H^\epsilon} I(\phi)$, completing the proof of (49).

Now, $\tilde{X}_n$ satisfies the LDP in $\mathcal{L}^{M+1}$ with rate function I, and H^ϵ is closed. Therefore,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}(\tilde{X}_n \in H^\epsilon) \;\leq\; -\inf_{\phi\in H^\epsilon} I(\phi) \;<\; I(\Phi). \tag{61}$$

Since $D(\Phi) \geq s$, the LDP implies that for arbitrarily small $\delta > 0$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}(\tilde{X}_n \in \mathcal{L}^{M+1} : D(\tilde{X}_n) > s - \delta) \;\geq\; -I(\Phi). \tag{62}$$

The claim of the theorem follows from (61), (62) and the definition of H^ϵ.

In order for an arbitrary customer to suffer a delay of n, there must be a last time before its arrival, say nc^{-1} arrivals earlier, when the tandem was empty. Let us define the scaled process of arrival times and service requirements $\tilde{X}_n$ as before, but with time and space now scaled by nc^{-1} instead of n. Then, by (5) and (47), the condition for the delay of the last customer to be n is that $D(\tilde{X}_n) = c$. The statement of the LDP modified for the new scaling is:

$$-\inf_{\phi \in \Gamma^o} \frac{I(\phi)}{c} \le \liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}(\tilde{X}_n \in \Gamma) \le \limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}(\tilde{X}_n \in \Gamma) \le -\inf_{\phi\in\overline{\Gamma}} \frac{I(\phi)}{c},$$

for any $\Gamma \subseteq \mathcal{L}^{M+1}$, where Γ^o and $\overline{\Gamma}$ denote, respectively, the interior and closure of Γ. We need the following result.

Lemma 3 *For any positive $c \neq s$, we have*

$$\inf\{c^{-1}I(\phi) : D(\phi) = c\} > s^{-1}I(\Phi).$$

Proof : Let $c \neq s$ be fixed, positive. As in the proof of the above theorem, we can find a piecewise linear function ψ and $0 = t_1 \le \ldots \le t_{M+1} = 1$ such that

$$D(\psi) = \sum_{k=1}^{M} \psi_k(t_{k+1}) - \psi_k(t_k) - \psi_0(1) = c, \tag{63}$$

and

$$\inf_{\phi: D(\phi)=c} I(\phi) = I(\psi) = \sum_{k=1}^{M} (t_{k+1} - t_k)\Lambda^*_{S^k}\left(\frac{\psi_k(t_{k+1}) - \psi_k(t_k)}{t_{k+1} - t_k}\right) + \Lambda^*_{T^1}(\psi_0(1)).$$

Therefore,

$$\begin{aligned} & \inf\{c^{-1}I(\phi) : D(\phi) = c\} \\ = \; & \sup_{\underline{\lambda}} \sum_{k=1}^{M} \left[\frac{\lambda_k(\psi_k(t_{k+1}) - \psi_k(t_k)) - (t_{k+1} - t_k)\Lambda_{S^k}(\lambda_k)}{c}\right] + \frac{\lambda_0\psi_0(1) - \Lambda_{T^1}(\lambda_0)}{c} \\ \ge \; & \sum_{k=1}^{M} \left[\frac{\theta^m(\psi_k(t_{k+1}) - \psi_k(t_k)) - (t_{k+1} - t_k)\Lambda_{S^k}(\theta^m)}{c}\right] - \frac{\theta^m\psi_0(1) - \Lambda_{T^1}(-\theta^m)}{c} \\ \ge \; & \theta^m - \sum_{k=1}^{M} c^{-1}(t_{k+1} - t_k)\Big[\Lambda_{S^k}(\theta^m) - \Lambda_{T^1}(-\theta^m)\Big], \end{aligned} \tag{64}$$

where the last inequality is by (63) and the fact that $t_{M+1} - t_1 = 1$. Finally, observe from (42) and (44) that the last quantity is no smaller than θ^m, and is equal to θ^m only if $t_{k+1} - t_k = 0$ for all $k \neq m$ and in addition, $\psi_m(1) = \Lambda'_{S^m}(\theta^m)$, $\psi_0(1) = \Lambda'_{T^1}(-\theta^m)$. But, if these conditions hold, then $\psi = \Phi$ and $D(\psi) = s$. The proof of the lemma is completed by recalling that $D(\Phi) = s$ and that $s^{-1}I(\Phi) = \theta^m$, as noted following (60).

The above lemma implies that, conditional on customer number zero suffering a delay n, the probability that customer $-nc$ was the last to enter an empty tandem is exponentially smaller than the probability that customer $-ns$ was the last to enter an empty tandem, for any $c \neq s$. Together with Theorem 2, this completes our description of the most likely path leading to large queueing delay.

References

[1] V. Anantharam, "How large delays build up in a GI/G/1 queue", *Queueing Systems*, 5(4): 345-367, 1989.

[2] D. Bertsimas, I. Paschalidis and J. Tsitsiklis, "On the Large Deviations Behaviour of Acyclic Networks of $G/G/1$ Queues". To appear in *Annals of Applied Probability*.

[3] C.S. Chang, "Sample Path Large Deviations and Intree Networks", *Queueing Systems*, 20(1-2): 7-36, 1995.

[4] G. de Veciana and J. Walrand, "Effective Bandwidths : Call admission, Traffic policing and Filtering for ATM networks", *Queueing Systems*, 20(1-2): 37-59, 1995.

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

[6] A. Ganesh and V. Anantharam, "Stationary tail probabilities in exponential server tandems with renewal arrivals", *Queueing Systems*, 22: 203-247, 1996.

[7] F. P. Kelly, *Reversibility and stochastic networks*, Wiley, 1979.

[8] G. Kesidis, J. Walrand and C. S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources", *IEEE/ACM Trans. on Networking*, 1(4): 424-428, 1993.

[9] R.M. Loynes, "The stability of queues with non-independent inter-arrival and service times", *Proceedings of the Cambridge Philosophical Society* , 58: 497-520, 1962.

[10] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models*, Johns Hopkins, 1981.

[11] R. T. Rockafellar, *Conjugate Duality and Optimization*, Society of Industrial and Applied Mathematics, 1974.

[12] J. Walrand, *An Introduction to Queueing Networks*, Prentice Hall, 1988.