Large Deviations of the Sojourn Time for Queues in Series

A. J. Ganesh*

July 3, 1995

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 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

*Dept. of Computer Science, University of Edinburgh, JCMB, Kings Buildings, Edinburgh EH9 3JZ. Research supported by a fellowship from BP and the Royal Society of Edinburgh

criteria take the form of bounds on the probability of cell loss, and of end-to-end transmission delays exceeding a threshold. Typically, we would like these probabilities to be of the order of 10^{-8} or lower. 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 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 [6]). 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 [3]. 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 [2]. In [5], Ganesh and Anantharam 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.* [1] 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.

The cell loss probability when the number of buffers at each queue is large is related directly to the tail distribution of the queue size. The relation between total delay in a network and waiting times at the individual queues is complicated by possible dependencies between these waiting times. 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.

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 $(n-1)^{th}$ 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 [7])

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

where X^+ denotes $\max\{X, 0\}$. It was shown by Loynes [7] that, if the arrival and service time distributions satisfy the stability criterion, then Lindley's recursion has the solution

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

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

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

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 $(m-1)^{th}$ 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

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

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. That is, we are interested in estimates of the form $P(\sum_{m=1}^M D_0^m \geq x) \approx \exp(-\eta x)$. The approach to obtaining these estimates is 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 [4]) 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 succesive 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. Starting with some definitions, we introduce below the assumptions that are used in the rest of the paper.

Define the effective domain of an extended real-valued function Λ as $\mathcal{D}_\Lambda = \{x : \Lambda(x) < +\infty\}$, and let $\mathcal{D}_\Lambda^0$ denote its interior. A convex function $\Lambda : I\!R^d \to (-\infty, +\infty]$ is called essentially smooth if

1. $\mathcal{D}_\Lambda^0$ is non-empty.
2. $\Lambda(\cdot)$ is differentiable throughout $\mathcal{D}_\Lambda^0$.
3. $\Lambda(\cdot)$ is steep, namely, $\lim_{n\to\infty} |\nabla\Lambda(\lambda_n)| = \infty$ whenever $\{\lambda_n\}$ is a sequence in $\mathcal{D}_\Lambda^0$ converging to a boundary point of $\mathcal{D}_\Lambda^0$.

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

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

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

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

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. The functions $\Lambda_{T^1}(\cdot)$ and $\Lambda_{S^m}(\cdot)$, $m = 1, \ldots, M$ are essentially smooth and lower semicontinuous, and the origin is in the interior of their effective domains.

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 required by the Gärtner-Ellis theorem which we use below. They 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.*, [9]) 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.*, [1].

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

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

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

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 [4], 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.

$\square$

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.

$\square$

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

$$\sum_{m=1}^{M} (\sigma^m_{j_m, j_{m+1}} \; - \; \tau^1_{j_1+1,0}) \geq x \tag{8}$$

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 [3]. The lower bounds on (9) and (10) are obtained as a combination of the following results:

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

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 given below.

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

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

while

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

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

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

Proof: 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

$$c_m e^{n(\Lambda_{S^m}(\theta) - \epsilon)} \leq E[\exp(\theta \sigma_{1,n}^m)] \leq c^m e^{n(\Lambda_{S^m}(\theta)+\epsilon)} \quad \forall\, n \geq 0 \tag{12}$$

$$k_1 e^{n(\Lambda_{T^1}(\theta) - \epsilon)} \leq E[\exp(\theta \tau_{1,n}^1)] \leq k^1 e^{n(\Lambda_{T^1}(\theta)+\epsilon)} \quad \forall\, n \geq 0 \tag{13}$$

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 $\theta \geq 0$,

$$\begin{aligned}
& E\left[\exp(\theta \sum_{m=1}^{M} D_0^m)\right] \\
&= E\left[\max_{a_1,\ldots,a_M \geq 0} \exp(-\theta \tau_{1,a_1+\ldots+a_M}^1) \cdot \prod_{m=1}^{M} \exp(\theta \sigma_{0,a_m}^m)\right] \\
&\leq \sum_{a_1,\ldots,a_M \geq 0} E\left[\exp(-\theta \tau_{1,a_1+\ldots+a_M}^1)\right] \cdot \prod_{m=1}^{M} E\left[\exp(\theta \sigma_{0,a_m}^m)\right]
\end{aligned} \tag{14}$$

Thus, by (12) and (13), for all $\theta \in (0, \theta^*)$,

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

$$\leq \sum_{a_1,\ldots,a_M \geq 0} k^1 e^{(a_1+\ldots+a_M)(\Lambda_{T^1}(-\theta)+\epsilon)} \cdot \prod_{m=1}^{M} c^m e^{(a_m+1)(\Lambda_{S^m}(\theta)+\epsilon)}$$

$$= \hat{c} \prod_{m=1}^{M} \sum_{a_m=0}^{\infty} \exp\left[a_m\Big(\Lambda_{S^m}(\theta) + \Lambda_{T^1}(-\theta) + 2\epsilon\Big)\right] \tag{15}$$

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

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

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

$$\limsup_{x\to\infty} \frac{1}{x} \log P\left(\sum_{m=1}^{M} D_0^m \geq x\right) \;\leq\; -\theta^* \tag{16}$$

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

$$\begin{aligned}
& E\left[\exp(\theta \sum_{m=1}^{M} D_0^m)\right] \\
&= E\left[\max_{0=a_0\leq\ldots\leq a_M} \exp(-\theta \tau^1_{1,a_M}) \cdot \exp\Big(\theta \sum_{m=1}^{M} \sigma^m_{a_{m-1},a_m}\Big)\right] \\
&\leq \sum_{0=a_0\leq\ldots\leq a_M} E\left[\exp(-\theta \tau^1_{1,a_1+\ldots+a_M})\right] \cdot E\left[\exp\Big(\theta \sum_{m=1}^{M} \sigma^m_{a_{m-1},a_m}\Big)\right]
\end{aligned} \tag{17}$$

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

$$\begin{aligned}
& E\left[\exp\Big(\theta \sum_{m=1}^{M} \sigma^m_{a_{m-1},a_m}\Big)\right] \\
&\leq \prod_{m=1}^{M} E\left[\exp(\theta \sigma^m_{a_{m-1}+1,a_m-1})\right] \cdot \prod_{m=0}^{M} E\left[\exp\Big(\theta(S^1_{a_m} + \ldots + S^M_{a_m}\Big)\right] \\
&= \exp\left[\sum_{m=1}^{M} (a_m - a_{m-1} - 1)\Lambda_m(\theta)\right] \prod_{m=0}^{M} E\left[\exp\Big(\theta(S^1_{a_m} + \ldots + S^M_{a_m}\Big)\right]
\end{aligned} \tag{18}$$

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 (13), (17) and (18), we get, for $\theta \in (0, \theta^*)$ and any $\epsilon > 0$,

$$\begin{aligned}
& E\left[\exp(\theta \sum_{m=1}^{M} D_0^m)\right] \\
\leq & \sum_{0=a_0 \leq \ldots \leq a_M} \hat{c} \prod_{m=1}^{M} \exp\left[(a_m - a_{m-1})\Big(\Lambda_{T^1}(-\theta) + \Lambda_{S^m}(\theta) + \epsilon\Big)\right] \\
= & \hat{c} \prod_{m=1}^{M} \sum_{b_m=0}^{\infty} \exp\left[b_m\Big(\Lambda_{T^1}(-\theta) + \Lambda_{S^m}(\theta) + \epsilon\Big)\right] \qquad (19)
\end{aligned}$$

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

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

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

$$\limsup_{x \to \infty} \frac{1}{x} \log P\left(\sum_{m=1}^{M} D_0^m \geq x\right) \;\leq\; -\theta^* \qquad (20)$$

We have, in (16) and (20), upper bounds on $P\left(\sum_{m=1}^{M} D_0^m \geq x\right)$ under assumptions 1(a) and 1(b) respectively. We now turn to estimating lower bounds on the probability of large queue sizes.

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,

$$\Lambda(\theta) = \lim_{n \to \infty} \frac{1}{n} \log E\left[\exp(\theta(\sigma_{1,n}^m - \tau_{1,n}^1))\right]$$

Also, by assumption 3, Λ is essentially smooth and lower semicontinuous. Hence, by the Gärtner-Ellis theorem (Theorem 2.3.6 in [4]), 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,

$$\limsup_{n \to \infty} \frac{1}{n} \log P\left(\frac{\sigma_{1n}^m - \tau_{1n}^1}{n} \in F\right) \;\leq\; -\inf_{x \in F} \Lambda^*(x) \qquad (21)$$

$$\liminf_{n \to \infty} \frac{1}{n} \log P\left(\frac{\sigma_{1n}^m - \tau_{1n}^1}{n} \in G\right) \;\geq\; -\inf_{x \in G} \Lambda^*(x) \qquad (22)$$

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

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

Hence, taking $G = (\alpha, \infty)$ in (22), we see that

$$\begin{aligned}
\liminf_{x\to\infty} \frac{1}{x} \log P\left(\sum_{m=1}^{M} D_0^m > x\right) \;&\geq\; \frac{1}{\alpha} \liminf_{n\to\infty} \frac{1}{n} \log P\left(\frac{\sigma_{1n}^m - \tau_{1n}^1}{n} > \alpha\right) \\
&= \; -\frac{1}{\alpha} \inf_{z>\alpha} \Lambda^*(z) \\
&\geq \; -(1+\epsilon)\frac{\Lambda^*((1+\epsilon)\alpha)}{(1+\epsilon)\alpha} \quad \forall\, \epsilon > 0 \qquad (23)
\end{aligned}$$

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

$$\liminf_{x\to\infty} \frac{1}{x} \log P\left(\sum_{m=1}^{M} D_0^m > x\right) \geq -\inf_{\alpha>0} \frac{\Lambda^*(\alpha)}{\alpha} \qquad (24)$$

By assumption 3, namely, that Λ is essentially smooth and lower semicontinuous, $\Lambda(\cdot)$ and $\Lambda^*(\cdot)$ are convex duals (see, for example, [8]). 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 (24) that

$$\liminf_{x\to\infty} \frac{1}{x} \log P\left(\sum_{m=1}^{M} D_0^m > x\right) \geq -\min_{1\leq m\leq M} \theta^m \qquad (25)$$

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 $\epsilon > 0$,

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

as can be shown by contradiction. Therefore,

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

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

$$\limsup_{x\to\infty} \frac{1}{x} \log P\left(\sum_{m=1}^{M} D_0^m > x\right) \geq -\theta^0 \tag{26}$$

The inequality in (25) holds *a fortiori* if lim inf is replaced by lim sup. Together with (26), this implies that

$$\limsup_{x\to\infty} \frac{1}{x} \log P\left(\sum_{m=1}^{M} D_0^m > x\right) \geq -\theta^* \tag{27}$$

Combining (16) and (20), which hold under assumption 1(a) and 1(b) respectively, with (27), we obtain the first claim of the theorem. The second claim is given by (25). The last claim of the theorem follows from the first two, and the definition of $\theta^0, \ldots, \theta^M$ and θ^*.

□

We now consider the qualitative behaviour of the system that results in large total delay for a customer. As is apparent from the proof above, 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 $P(\sigma_{1,n}^m - \tau_{1,n}^1 \geq x)$ over n, $\sigma_{1,n}^m$ and $\tau_{1,n}^1$. Using large deviations theory to approximate the above probability, we get the equivalent problem:

$$\min_{n,y,z} n\left[\Lambda_{S^m}^*(\frac{y}{n}) + \Lambda_{T^1}^*(\frac{z}{n})\right] \quad \text{subject to} \quad y - z \geq x$$

Here m indexes one of the bottleneck queues, which are those queues for which $\theta^m = \theta^*$. 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 following example.

Example : 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,

$$\lim_{x \to \infty} \frac{1}{x} \log P\left(D_0^1 + D_0^2 \geq x\right) = \mu - \lambda. \tag{28}$$

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

$$\lim_{x \to \infty} \frac{1}{x} \log P\left(S_0^1 + S_0^2 \geq x\right) = \lim_{x \to \infty} \frac{1}{x} \log P\left(S_0^1 \geq \frac{x}{2}\right) = \frac{\mu}{2}. \tag{29}$$

Since $S_0^1 + S_0^2$ is a lower bound on $D_0^1 + D_0^2$, it follows from (29) and the upper bound in (9) that

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

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

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.

References

[1] D. Bertsimas, I. Paschalides and J. Tsitsiklis, "On the Large Deviations Behaviour of Acyclic Networks of $G/G/1$ Queues", *Preprint*, 1994.

[2] C.S. Chang, "Sample Path Large Deviations and Intree Networks", *IBM RC 19118*, 1993.

[3] G. de Veciana and J. Walrand, "Effective Bandwidths : Call admission, Traffic policing and Filtering for ATM networks", *Preprint*, 1993.

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

[5] A. Ganesh and V. Anantharam, "Stationary tail probabilities in exponential server tandems with renewal arrivals". *To appear in Queueing Systems*, 1995.

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

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

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

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