The Annals of Applied Probability
1998, Vol. 8, No. 4, 1027–1069

ON THE LARGE DEVIATIONS BEHAVIOR OF ACYCLIC NETWORKS OF $G/G/1$ QUEUES[1]

BY DIMITRIS BERTSIMAS, IOANNIS CH. PASCHALIDIS AND JOHN N. TSITSIKLIS

Massachusetts Institute of Technology, Boston University and Massachusetts Institute of Technology

We consider a single class, acyclic network of $G/G/1$ queues. We impose some mild assumptions on the service and external arrival processes and we characterize the large deviations behavior of all the processes resulting from various operations in the network. For the network model that we are considering, these operations are passing-through-a-single-server-queue (the process resulting from this operation being the departure process), superposition of independent processes and deterministic splitting of a process into a number of processes. We also characterize the large deviations behavior of the waiting time and the queue length observed by a typical customer in a single server queue. We prove that the assumptions imposed on the external arrival processes are preserved by these operations, and we show how to apply inductively these results to obtain the large deviations behavior of the waiting time and the queue length in all the queues of the network. Our results indicate how these large deviations occur, by concretely characterizing the most likely path that leads to them.

1. Introduction. Consider a single class, acyclic network of $G/G/1$ queues. Customers arrive at the network in a number of independent streams and are treated uniformly by the network. Different streams may share a queue and the first-come–first-serve (FCFS) policy is implemented. A constant fraction p_{ij} of customers departing a queue i is routed to queue j and a fraction p_{i0} leaves the network. The aim of this paper is to derive large deviations results for the waiting time and the queue length observed by an arbitrary customer at different queues of the network.

The main application area that motivates the study of such systems is the design and the operation of high speed, packet-switched communication networks. These networks will accommodate various types of traffic, namely, digitized voice, encoded video and data. The interesting problem arising is how to estimate and prevent congestion, which may cause long delays and

Received December 1994; revised November 1997.

[1] Supported in part by a Presidential Young Investigator Award DDM-91-58118 with matching funds from Draper Laboratory, by ARO Grant DAAL-03-92-G-0115 and by NSF Grant NCR-97-06148.

AMS 1991 *subject classifications.* 90B12, 60K30, 60F10, 60K25.

Key words and phrases. Communication networks, large deviations, queueing networks.

packet losses. It is desirable to operate the network in a regime where packet loss probabilities are very small, for example, on the order of 10^{-9}. Moreover, large delays should also have a correspondingly small probability. Thus, the need for understanding the large deviations behavior of such a network arises. In this paper, we consider single class networks, which from the application point of view means that we are dealing only with one type of traffic in the network. For this reason, the FCFS assumption can be made without loss of generality.

The problem of estimating tail probabilities of rare events in a single queue has received extensive attention in the literature and has been approached by two main methodologies. The first one is to use large deviations theory, as we do in this paper. This approach is used in [18] to estimate the tail probability of the queue length in a $G/G/1$ queue. In that paper, a discrete time model was used in contrast to the continuous time model that we use in this paper. Similar results are obtained in [11]. The second approach is to use spectral decomposition techniques. This second approach is used in [20] to estimate the tail probability of the queue length in a queue with a deterministic server and Markov modulated arrival process. Results for the single queue case were first obtained in [24], [26], [22] and later in [27], [32] and [23]. In all of these papers, the large deviations results obtained are used to derive appropriate admission control schemes for networks.

The extension of these ideas to networks appears to be a rather challenging problem. Researchers have been able to obtain some bounds on the tail probabilities for delays and queue lengths in various networks models (see [7, 12, 13, 33]), but it is not clear whether these bounds are tight. Recently, large deviations results for two queues in tandem, with renewal arrivals and exponential servers, were reported in [21]. In [16], a very interesting approach is used to obtain results for networks with deterministic servers. The departure process from a single $G/D/1$ queue is characterized in the large deviations regime, using a discrete time model, in an attempt to treat the whole network inductively. The main focus of [16] is to apply the large deviations results obtained to resource management for networks. A large deviations upper bound for the departure process appears also in [9]. It is important to point out that the departure process is a very difficult process for which to obtain exact results (see, e. g., [4]). However, we should note that it is not very clear to us how the large deviations result for the departure process in [16] can be applied inductively. The crux of the matter is that [16] uses a technical result from [14] in order to obtain the large deviations behavior of the departure process. The latter result holds under certain technical assumptions on the arrival process. Since the departure process from a queue is the arrival process in another downstream queue in the network, one would need at this point to verify that the same technical assumptions hold for the departure process. This is not done in [16] and appears to be rather difficult.

In the present paper, we consider a continuous time model and we extend the work in [16] to a network of $G/G/1$ queues. The objective is to obtain the large deviations behavior of waiting times and queue lengths in all the nodes of the network. To this end, we initially seek to characterize the large deviations behavior of the aggregate arrival process in each node. Our results are self-contained in the sense that we do not need the technical results of [14]. Instead, we impose certain assumptions on the external arrival processes and we characterize the large deviations behavior of all the processes resulting from various operations in the network. For the network model that we are considering, these operations are passing-through-a-queue (the process resulting from this operation being the departure process), superposition of independent processes and deterministic splitting of a process into a number of processes. We prove that the assumptions imposed on the external arrival processes are preserved by these operations, and thus we are able to apply these results inductively to obtain large deviations results for the aggregate arrival process in each node. As a by-product of our analysis we also obtain large deviations results for the internal traffic in the network. For a single queue, in isolation, we characterize the large deviations behavior of the waiting time incurred by a typical customer and, by using ideas from distributional laws (see [5, 3]), the large deviations behavior of the queue length observed by a typical customer. Finally, we compose the large deviations behavior of the aggregate arrival process in each node of the network with the results for a single queue to obtain the large deviations behavior of the waiting time and queue length in each node.

Our approach provides particular insight on how these large deviations occur, by concretely characterizing the most likely path that leads to them. Characterizations of most likely paths were obtained for the single queue case in [2], [1] and [14]. After the submission of the present paper the work in [8] and [10] was brought to our attention. In [8] the author independently obtained the large deviations behavior for a network model of $G/D/1$ queues similar to ours, when the external arrival processes are bounded. In [10] the authors obtain the large deviations behavior of the departure process of a $G/G/1$ queue, in isolation.

It is interesting to note that in order to obtain the large deviations behavior of the superposition operation we prove a general result that connects the *stationary distribution* (i.e., as it is seen at a random time) and the *Palm distribution* (i.e., as it is seen by a typical customer) of a point process in the large deviations regime. This result could be of independent interest.

Regarding the structure of the paper, we start in Section 2 by reviewing some results from the theory of large deviations that we use in the sequel. In Section 3 we present the network model that we are considering and establish our notation. In Section 4 we treat the single queue case. This section is comprised of two subsections. In Section 4.1 we review the existing result for the large deviations behavior of the waiting time and we completely characterize the most likely path along which the waiting time takes large values. In Section 4.2, using an idea from distributional laws, we obtain the tail probability of the queue length. In Section 5 we derive the large deviations behavior of the departure process from a $G/G/1$ queue. Particular attention is given to the way that such a deviation occurs. In Section 5.1, some special cases are studied. Namely, we apply the result for the departure process of a $G/G/1$ queue to a $G/D/1$ queue and an $M/M/1$ queue. For the latter case, Burke's theorem is verified in the large deviations regime. In Sections 6 and 7 we study the large deviations behavior of the processes resulting from the following operations: superposition of independent processes and deterministic splitting of a process into a number of processes, respectively. In Section 6.1 we prove a result that connects the Palm and the stationary distribution of a point process in the large deviations regime. This result is used in the rest of Section 6 to derive the large deviations behavior of the superposition process. In Section 8, we treat, as an example, a network consisting of two queues in tandem. We characterize the way that the waiting time in the second queue reaches large values and we include some numerical results. Finally, in Section 9 we provide some concluding remarks and discuss some open problems.

2. Preliminaries. In this section we review some basic results from large deviations theory that will be used in the sequel.

We first state that Gärtner–Ellis theorem (see [6] and [15]), which establishes a *large deviations principle* (LDP) for random variables. It is a generalization of Cramér's theorem, which applies to independent and identically distributed (iid) random variables.

Consider a sequence $\{S_1, S_2, \dots\}$ of random variables with values in $\mathbb{R}$ and define

$$\Lambda_n(\theta) \triangleq \frac{1}{n}\log \mathbf{E}[e^{\theta S_n}]. \tag{1}$$

For the applications that we have in mind, S_n is a partial sum process. Namely, $S_n = \sum_{i=1}^{n} X_i$, where X_i, $i \geq 1$, are identically distributed, possibly dependent random variables.

ASSUMPTION A.

(i) The limit

$$\Lambda(\theta) \triangleq \lim_{n\to\infty} \Lambda_n(\theta) = \lim_{n\to\infty} \frac{1}{n}\log \mathbf{E}[e^{\theta S_n}] \tag{2}$$

exists for all θ, where $\pm\infty$ are allowed both as elements of the sequence $\Lambda_n(\theta)$ and as limit points.

(ii) The origin is in the interior of the domain $D_\Lambda \triangleq \{\theta | \Lambda(\theta) < \infty\}$ of $\Lambda(\theta)$.

(iii) $\Lambda(\theta)$ is differentiable in the interior of D_Λ and the derivative tends to infinity as θ approaches the boundary of D_Λ.

(iv) $\Lambda(\theta)$ is lower semicontinuous, that is, $\liminf_{\theta_n \to \theta} \Lambda(\theta_n) \geq \Lambda(\theta)$, for all θ.

THEOREM 2.1 (Gärtner–Ellis). *Under Assumption* A, *the following inequalities hold.*

Upper bound: *For every closed set* F,

$$\limsup_{n\to\infty} \frac{1}{n}\log \mathbf{P}\left[\frac{S_n}{n} \in F\right] \leq -\inf_{a\in F} \Lambda^*(a). \tag{3}$$

Lower bound: *For every open set* G,

$$\liminf_{n\to\infty} \frac{1}{n}\log \mathbf{P}\left[\frac{S_n}{n} \in G\right] \geq -\inf_{a\in G} \Lambda^*(a), \tag{4}$$

where

$$\Lambda^*(a) \triangleq \sup_{\theta}(\theta a - \Lambda(\theta)). \tag{5}$$

We say that $\{S_n\}$ satisfies a LDP with *good rate function* $\Lambda^*(\cdot)$. The term "good" refers to the fact that the level sets $\{a|\Lambda^*(a) \leq k\}$ are compact for all $k < \infty$, which is a consequence of Assumption A (see [15] for a proof).

It is important to note that $\Lambda(\cdot)$ and $\Lambda^*(\cdot)$ are convex duals (Legendre transforms of each other). Namely, along with (5), it also holds that

$$\Lambda(\theta) = \sup_{a}(\theta a - \Lambda^*(a)). \tag{6}$$

The Gärtner–Ellis theorem intuitively asserts that for large enough n and for small $\varepsilon > 0$,

$$\mathbf{P}[S_n \in (na - n\varepsilon, na + n\varepsilon)] \sim \exp(-n\Lambda^*(a)).$$

However, in this paper, we are mostly estimating tail probabilities of the form $\mathbf{P}[S_n \leq na]$ or $\mathbf{P}[S_n \geq na]$. We therefore define large deviations rate functions associated with such tail probabilities.

Consider the case where $S_n = \sum_{i=1}^n X_i$, the random variables X_i, $i \geq 1$, being identically distributed, and let $m = \mathbf{E}[X_1]$. It is easily shown (see [15]) that $\Lambda^*(m) = 0$. Let us now define

$$\Lambda^{*+}(a) \triangleq \begin{cases} \Lambda^*(a), & \text{if } a > m, \\ 0, & \text{if } a \leq m, \end{cases} \tag{7}$$

and

$$\Lambda^{*-}(a) \triangleq \begin{cases} \Lambda^*(a), & \text{if } a < m, \\ 0, & \text{if } a \geq m. \end{cases} \tag{8}$$

Notice that $\Lambda^{*+}(a)$ is nondecreasing and $\Lambda^{*-}(a)$ is a nonincreasing function of a, respectively. The convex duals of these functions are

$$\Lambda^{+}(\theta) \triangleq \begin{cases} \Lambda(\theta), & \text{if } \theta \geq 0, \\ +\infty, & \text{if } \theta < 0, \end{cases} \tag{9}$$

and

$$\Lambda^{-}(\theta) \triangleq \begin{cases} \Lambda(\theta), & \text{if } \theta \leq 0, \\ +\infty, & \text{if } \theta > 0, \end{cases} \tag{10}$$

respectively. In particular, $\Lambda^{*-}(a) = \sup_\theta(\theta a - \Lambda^-(\theta))$ and $\Lambda^{*+}(a) = \sup_\theta(\theta a - \Lambda^+(\theta))$.

Using the Gärtner–Ellis theorem it can be shown that for all $\varepsilon_1, \varepsilon_2 > 0$ there exists n_0 such that for all $n \ge n_0$,

$$\exp(-n(\Lambda^{*-}(a) + \varepsilon_2)) \le \mathbf{P}[S_n \le na] \le \exp(-n(\Lambda^{*-}(a) - \varepsilon_1)), \tag{11}$$

and

$$\exp(-n(\Lambda^{*+}(a) + \varepsilon_2)) \le \mathbf{P}[S_n \ge na] \le \exp(-n(\Lambda^{*+}(a) - \varepsilon_1)). \tag{12}$$

More specifically, the lower bound in (11) can be obtained by noting that $\mathbf{P}[S_n \le na] \ge \mathbf{P}[S_n < na]$, and using the lower bound of the Gärtner–Ellis theorem for the open set $(-\infty, a)$. The upper bound can be obtained by an argument similar to the one we use in the proof of Lemma 4.2. A similar argument can be used for (12).

3. The network model. In this section, we formally define the network model for which we will derive the large deviations behavior. Moreover, we establish the notation that we will be using and state a set of assumptions on the arrival and service processes.

Consider a *directed acyclic graph* (dag) with J nodes. For reasons that will soon become apparent, we assume that any two directed paths do not meet in more than one node. Each node of the graph is equipped with an infinite buffer and a single server. Customers enter the network in a number of independent streams $A^1, A^2, \dots, A^J$. In particular, A^i is the stream of customers that enter the network at node i. Customers are treated uniformly by the network; that is, the network is assumed to be *single class*. Let $\mathbb{Z}$ denote the set of integers. By A_i^j, $i \in \mathbb{Z}$, we denote the interarrival time of the ith customer in the jth stream [the interval between the arrival epochs of the $(i-1)$st and the ith customer]. By B_i^j, $i \in \mathbb{Z}$, we denote the service time of the ith customer in the jth node. We assume that for each arriving stream j the process $\{A_i^j,\ i \in \mathbb{Z}\}$, is stationary, and A_i^j, $i \in \mathbb{Z}$, are possibly dependent random variables. Moreover, for each node j, the service times B_i^j, $i \in \mathbb{Z}$, are iid random variables. We also assume that interarrival and service times at a specific node are mutually independent and that service times at different nodes are independent.

Independent streams may share a queue and the FCFS policy is implemented. A fraction $p_{ij_1}, p_{ij_2}, \dots$ of customers departing node i, which is connected to nodes $j_1, j_2, \dots,$ are routed to these nodes, respectively, and a fraction p_{i0} leaves the network. The exact way that the routing is performed is not of importance in the large deviations regime. Roughly, out of every $1/p_{ij}$ customers leaving node i, the routing mechanism sends one to node j. Figure 1 depicts an example of the class of networks considered. Such a network is intended to model packet-switched communication networks.

We denote by $W^1, W^2, \dots, W^J$ and $L^1, L^2, \dots, L^J$ the steady-state waiting times and queue lengths incurred by a typical customer at nodes $1, 2, \dots, J$ of the network, respectively. For each node j, W_n^j (resp. L_n^j) denotes the waiting

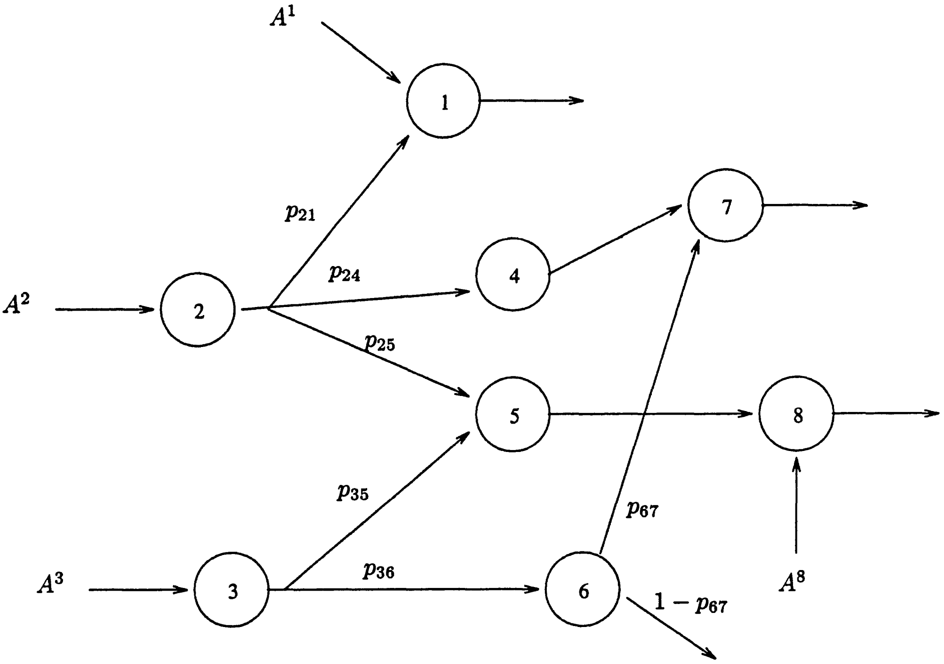

FIG. 1. *A network example.*

time incurred (resp. queue length observed) by the nth customer. We assume that the process $\{(W_n^j, L_n^j);\ n \in \mathbb{Z},\ j = 1, \dots, J\}$ is stationary. (The existence and stationarity of this process contains an implicit stability assumption.)

In this paper, we derive large deviations results for the steady-state waiting times $W^1, W^2, \dots, W^J$, and the corresponding queue lengths $L^1, L^2, \dots, L^J$, incurred at nodes $1, 2, \dots, J$ of the network, respectively (as these random variables are seen by a typical customer). Our strategy is first to obtain large deviations results for the steady-state waiting time and the corresponding queue length in a single $G/G/1$ queue. Then it suffices to derive a LDP for the partial sum of the aggregate arrival process in each queue of the network and apply the result for the single queue case. It is important to note that by the definition of the network all the streams sharing the same queue are independent. Therefore, from the model description, it is apparent that it suffices to obtain LDP's for the processes resulting from the following operations:

1. Passing-through-a-queue (the process resulting from this operation being the departure process).
2. Superposition of independent streams.
3. Deterministic splitting of a stream to a number of streams.

Let $\{A_i,\ i \in \mathbb{Z}\}$ be an arbitrary external arrival process and $\{B_i,\ i \in \mathbb{Z}\}$ be an arbitrary service process. Hereafter, we will be using the notation $S^X_{i,j} \triangleq \sum_{k=i}^{j} X_k;\ i \le j$ for the partial sums of the random sequence $\{X_i;\ i \in \mathbb{Z}\}$ along with the convention $S^X_{i,j} \triangleq 0;\ i > j$.

ASSUMPTION B.

(i) The sequence of partial sums $\{S^A_{1,n};\ n \ge 1\}$ satisfies

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left[S^A_{1,n} \le na\right] = -\Lambda^{*-}_A(a), \tag{13}$$

where

$$\Lambda^-_A(\theta) \triangleq \begin{cases} \lim_{n\to\infty}(1/n)\log \mathbf{E}\left[\exp\left(\theta S^A_{1,n}\right)\right], & \text{if } \theta \le 0, \\ +\infty, & \text{if } \theta > 0, \end{cases} \tag{14}$$

and

$$\Lambda^{*-}_A(a) \triangleq \sup_\theta \left(\theta a - \Lambda^-_A(\theta)\right). \tag{15}$$

The limit in the upper branch of (14) exists for all θ, where $\pm\infty$ are allowed both as elements of the sequence and as limit points. We will say that $\{S^A_{1,n};\ n \ge 1\}$ satisfies a one-sided LDP. Moreover, we assume that $\Lambda^{*-}_A(a)$ is a strictly convex function of a in the intersection of the interior of its domain and the interval $(-\infty, \mathbf{E}[A_1])$, and that A_1 has moments of all orders, that is, $\mathbf{E}[A_1^p] < \infty$ for all $p \ge 0$.

(ii) The sequence of partial sums $\{S^B_{1,n};\ n \ge 1\}$ satisfies the requirements of the Gärtner–Ellis theorem with limiting log-moment generating function

$$\Lambda_B(\theta) \triangleq \lim_{n\to\infty} \frac{1}{n} \log \mathbf{E}\left[\exp\left(\theta S^B_{1,n}\right)\right] \tag{16}$$

and large deviations rate function

$$\Lambda^*_B(a) \triangleq \sup_\theta \left(\theta a - \Lambda_B(\theta)\right). \tag{17}$$

Moreover, we assume that $\Lambda^*_B(a)$ is a strictly convex function of a in the interior of its domain and that B_1 has moments of all orders, that is, $\mathbf{E}[B_1^p] < \infty$ for all $p \ge 0$.

ASSUMPTION C.

(i) For every $\varepsilon_1, \varepsilon_2, a > 0$, there exists M_A such that, for all $n \ge M_A$,

$$\exp\left(-n\left(\Lambda^{*-}_A(a) + \varepsilon_2\right)\right) \le \mathbf{P}\left[S^A_{1,i} - ia \le \varepsilon_1 n,\ i = 1, \dots, n\right]. \tag{18}$$

(ii) For every $\varepsilon_1, \varepsilon_2, a > 0$, there exists M_B such that, for all $n \ge M_B$,

$$\begin{aligned} &\exp\left(-n\left(\Lambda^{*-}_B(a) + \varepsilon_2\right)\right) \\ &\qquad \le \mathbf{P}\left[S^B_{i,j} - (j - i + 1)a \le \varepsilon_1 n,\ 1 \le i \le j \le n\right] \end{aligned} \tag{19}$$

and

$$\exp(-n(\Lambda_B^{*+}(a)+\varepsilon_2)) \leq \mathbf{P}\big[S_{i,j}^B-(j-i+1)a \geq -\varepsilon_1 n,\ 1\leq i\leq j\leq n\big]. \tag{20}$$

We consider external arrival and service processes that satisfy Assumptions B and C. We will show that these assumptions are satisfied by the processes resulting from the three operations mentioned above. In this way, our approach provides a *calculus of acyclic networks* since we will be able to determine the large deviations behavior of each individual queue inductively.

Assumption B provides a LDP for the arrival and service processes. Based on these LDP's we will derive LDP's for all the processes of interest in the network. Note that only the tail probability of the external arrival processes corresponding to "many arrivals" is characterized by Assumption B. We will prove that in order to estimate probabilities of large waiting times and long delays, as we do in this paper, only such a tail probability of the aggregate arrival process in each queue of the network is needed. The strict convexity assumption on the large deviations rate functions of interest is needed to avoid some technical issues that have to do with the differentiability of the corresponding limiting log-moment generating functions.

Assumption C is needed in order to derive a LDP for the departure process of a $G/G/1$ queue. It intuitively asserts that besides the LDP for the partial sum random *variable* $S_{1,n}$, we also have a LDP for the partial sum *process* $\{S_{1,i},\ i=1,\dots,n\}$ for the arrivals and $\{S_{i,j},\ 1\leq i\leq j\leq n\}$ for the service times. In other words, (18) and (19) guarantee that, with high probability, the partial sum process follows a path that never overshoots the straight line of slope a, in order to reach an improbable level $S_{1,n}\leq na$. A similar interpretation can be given to (20). Mild mixing conditions on the arrival and service processes suffice to guarantee Assumption C. A thorough treatment is given in [14]. In the Appendix we provide some conditions under which Assumption C is satisfied, based on the results of [14]. In [8] a uniform bounding condition is given under which the above assumption is true. We should note here that we do not need the full power of the sample path large deviations results in [14] and [8] to establish our results. We only need Assumptions B and C, which, as we will show, are preserved by the internal traffic in the network.

Assumptions B and C are satisfied by processes that are used to model external arrival and services in communications networks, such as renewal processes, stationary processes with mild mixing conditions, as well as Markov-modulated processes with some uniformity assumptions on the stationary distribution (see [14], Section 4).

4. Large deviations of a $G/G/1$ queue. In this section, we establish a LDP for the Palm distributions of the steady-state waiting time and queue length (i.e., as these random variables are seen by a typical customer), in a $G/G/1$ queue with stationary arrivals and service times.

The setting is the same as in Section 3. We denote by $\{A_i,\ i \in \mathbb{Z}\}$ the stationary aggregate arrival process to the queue and we assume that it satisfies Assumption B(i). We also denote by $\{B_i,\ i \in \mathbb{Z}\}$ the stationary service process and we assume that it satisfies Assumption B(ii). For this section, the independence assumption for the service times can be relaxed. For stability purposes, we further assume $\mathbf{E}[A] > \mathbf{E}[B]$, where A (resp. B) denotes a typical interarrival (resp. service) time.

4.1. *Large deviations of the waiting time.* Let us first characterize the steady-state waiting time, W, incurred by a typical customer. By W_n we denote the waiting time of the nth customer. The condition $\mathbf{E}[A] > \mathbf{E}[B]$ is necessary for the existence and the uniqueness of a stationary process (see [31]). For sufficiency, ergodicity is also needed. From the Lindley equation, the waiting time of the 0th customer, at steady-state, is given by

$$\begin{aligned} W_0 &= [W_{-1} + B_{-1} - A_0]^+ \triangleq \max[W_{-1} + B_{-1} - A_0, 0] \\ &= \max_{i \geq 0} \left[S^B_{-i-1,-1} - S^A_{-i,0}, 0\right]. \end{aligned} \tag{21}$$

The intuitive meaning of this relation is the following: for a particular sample path, if i^* is the optimum i, then the customer with label $-i^* - 1$ is the one who initializes the busy period in which the 0th customer is served.

The next theorem establishes a LDP for W_0. This result is not new. The proof is almost identical with the proof in [19], Theorem 3.1, where a discrete time model is used and is therefore omitted. A similar argument is also given in [7]. An upper bound on the tail probability of the steady-state waiting time, for renewal arrival and service processes, was first obtained by Kingman [28].

THEOREM 4.1. *The tail of the Palm distribution of the steady-state waiting time, W, in a FCFS $G/G/1$ queue with arrivals and service times satisfying Assumption* B *is characterized by*

$$\lim_{U \to \infty} \frac{1}{U} \log \mathbf{P}[W \geq U] = \theta^*, \tag{22}$$

where $\theta^* < 0$ is the smallest root of the equation

$$\Lambda_A^-(\theta) + \Lambda_B(-\theta) = 0. \tag{23}$$

REMARKS. Intuitively, Theorem 4.1 asserts that for large enough U, we can state

$$\mathbf{P}[W \geq U] \sim e^{\theta^* U} \quad \text{where } \theta^* < 0 \text{ is such that} \quad \Lambda_A^-(\theta^*) + \Lambda_B(-\theta^*) = 0. \tag{24}$$

Note that θ^* exists as an extended real number since $\mathbf{E}[A] > \mathbf{E}[B]$ and the functions $\Lambda_A^-(\cdot), \Lambda_B(\cdot)$ are convex. This is proven under the conditions of Assumption B in [15], Lemma 2.3.9. Figure 2 depicts the function $\Lambda_A^-(\theta) + \Lambda_B(-\theta)$ and the root θ^*. If $\Lambda_A^-(\theta) + \Lambda_B(-\theta) < 0$ for all $\theta < 0$, we use the convention $\theta^* = \infty$. In Figure 2 we make use of the differentiability of

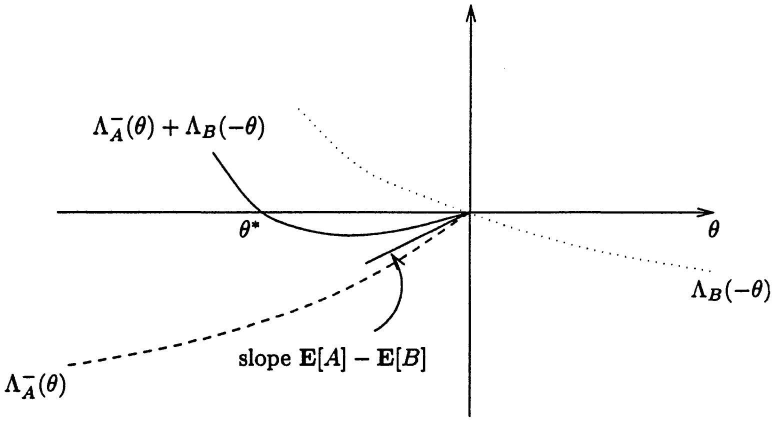

FIG. 2. *The root of* $\Lambda_A^-(\theta) + \Lambda_B(-\theta) = 0$.

$\Lambda_A^-(\cdot)$, $\Lambda_B(\cdot)$, which is guaranteed by the strict convexity assumption that we have imposed on $\Lambda_A^{*-}(\cdot)$ and $\Lambda_B^*(\cdot)$ (see Assumption B). More specifically, convex duality arguments (see [29], Theorem 26.3) guarantee that the convex dual of a strictly convex function is differentiable in the interior of its domain. Regarding $\Lambda_A^-(\theta)$, we are only interested in its differentiability for $\theta < 0$, which can be established using the strict convexity of $\Lambda_A^{*-}(a)$ for $a < \mathbf{E}[A_1]$. (As we have defined $\Lambda_A^-(\theta)$, it has only a left derivative at zero, which is positive and equal to $\mathbf{E}[A] - \mathbf{E}[B]$.)

It is instructive to characterize the most likely "path" along which the large deviation of the waiting time occurs. Such a characterization can also provide an alternative proof of Theorem 4.1. Let $a > 0$ and $x_1, x_2 \in \mathbb{R}^+$, such that $x_2 - x_1 = a$. Using (21), we have

$$\begin{aligned}\mathbf{P}[W_0 \geq (i+1)a] &\geq \mathbf{P}\big[S^B_{-i-1,-1} - S^A_{-i,0} \geq (i+1)a\big]\\ &\geq \mathbf{P}\big[S^A_{-i,0} \leq (i+1)x_1\big]\mathbf{P}\big[S^B_{-i-1,-1} \geq (i+1)x_2\big] \qquad (25)\\ &\geq \exp\big(-(i+1)\big[\Lambda_A^{*-}(x_1) + \Lambda_B^{*+}(x_2) + \varepsilon\big]\big),\end{aligned}$$

where the last inequality makes use of Assumption B and holds for any $\varepsilon > 0$ and for large i. Setting $U = (i+1)a$, we obtain

$$\mathbf{P}[W_0 \geq U] \geq \exp\left\{-U \inf_{a>0} \frac{1}{a} \inf_{x_2 - x_1 = a} [\Lambda_A^{*-}(x_1) + \Lambda_B^{*+}(x_2)] - U\varepsilon\right\}. \qquad (26)$$

Let $a^* > 0$ be a solution to the above optimization problem. Thus, for large U, and by taking $\varepsilon \to 0$ in (26), we obtain

$$\mathbf{P}[W_0 \geq U] \geq \exp\left\{-U \frac{\inf_{x_2 - x_1 = a^*}[\Lambda_A^{*-}(x_1) + \Lambda_B^{*+}(x_2)]}{a^*}\right\}. \qquad (27)$$

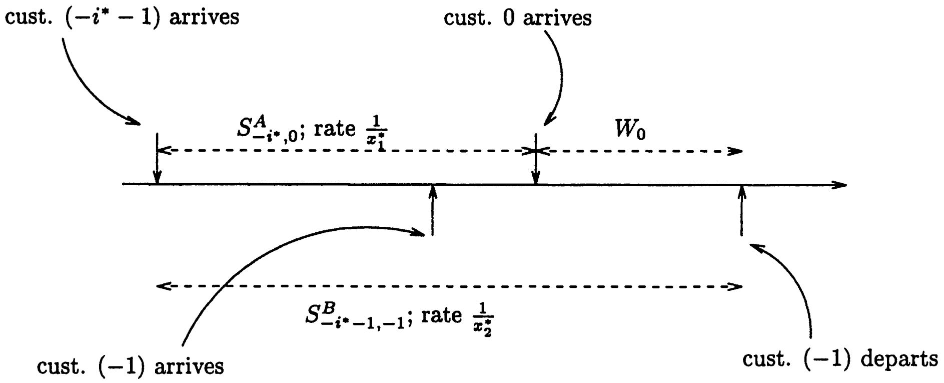
cust. $(-i^* - 1)$ arrives
cust. 0 arrives
$S^A_{-i^*,0}$; rate $\frac{1}{x_1^*}$
W_0
$S^B_{-i^*-1,-1}$; rate $\frac{1}{x_2^*}$
cust. (-1) arrives
cust. (-1) departs

FIG. 3. *The optimal path for large deviations in the waiting time.*

The tightness of this bound can be proven by obtaining a matching (i.e., with the same exponent) upper bound; the proof is omitted.

Let i^* be defined by the equation $i^* + 1 = U/a^*$. Let also x_1^* and x_2^* solve the optimization problem in (27). Consider a scenario where customers $-i^*, \dots, -1, 0$ arrive at an empirical arrival rate of $1/x_1^*$ and customers $(-i^* - 1), \dots, -1$ are served with an empirical service rate of $1/x_2^*$. Such a scenario, which is depicted in Figure 3, has probability comparable to the right-hand side of (27) and is therefore a most likely way for the large deviation of the waiting time to occur.

4.2. *Large deviations of the queue length.* In this section, we present a LDP for the steady-state queue length in a $G/G/1$ queue, as seen by a typical customer (Palm distribution). To accomplish this, we use the main argument used in deriving distributional laws; that is, a probabilistic relation between the waiting time and the queue length. A detailed discussion of distributional laws and their applications can be found in [5] and [3]. It is important to note that distributional laws have been proved there only for renewal arrival and service processes. However in the large deviations setting, we are able to relax the renewal assumption and state a result that holds even for correlated arrival and service processes.

Let us now characterize the steady-state queue length L seen by a typical customer (not including herself) upon arrival (this is sometimes denoted by L^- in the literature). The goal is to estimate $\mathbf{P}[L \geq n]$. Let us denote by L_n the queue length observed by the nth customer. As in Section 3, we assume that the process $\{(L_n, W_n);\ n \in \mathbb{Z}\}$ is stationary. The main idea, in order to establish a relation between the waiting time and the queue length, is to look backwards in time from the arrival epoch of the nth customer. Figure 4 depicts the situation. We denote with $T_0, T_1, \dots$ the arrival epochs of customers $0, 1, \dots,$ respectively. Recall that W_n and B_n denote the waiting and the service time of the nth customer, respectively.

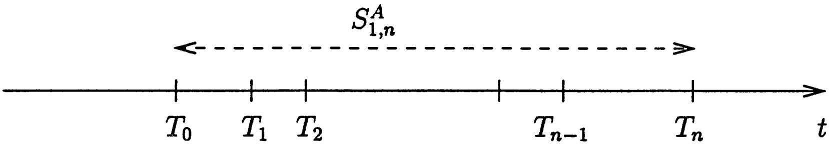

FIG. 4. *The system at time T_n.*

The main observation is the following: in order for the queue length right before T_n to be at least n, the 0th customer should be in the system at that time. Namely,

$$\mathbf{P}[L_n \geq n] = \mathbf{P}\big[W_0 + B_0 \geq S^A_{1,n}\big] \tag{28}$$

and by using (21) we obtain

$$\begin{aligned} \mathbf{P}[L_n \geq n] &= \mathbf{P}\Big[\max_{i \geq 0}\big[S^B_{-i-1,0} - S^A_{-i,n}, -S^A_{-i,n}\big] \geq 0\Big] \\ &= \mathbf{P}\Big[\max_{i \geq -1}\big[S^B_{-i-1,0} - S^A_{-i,n}\big] \geq 0\Big]. \end{aligned} \tag{29}$$

The next theorem establishes a LDP for L_n. We will need a technical lemma which we prove next. (This lemma is also used in the next section.)

LEMMA 4.2. *Under Assumption* B, *and for $\theta < 0$, satisfying $\Lambda_A^-(\theta) + \Lambda_B(-\theta) < 0$, it holds that*

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbf{E}\big[\exp\big(-\theta \max\nolimits_{i \geq -1}\big[S^B_{-i-1,0} - S^A_{-i,n}\big]\big)\big] \leq \Lambda_A^-(\theta). \tag{30}$$

PROOF. We have

$$\begin{aligned} &\mathbf{E}\big[\exp\big(-\theta \max\nolimits_{i \geq -1}\big[S^B_{-i-1,0} - S^A_{-i,n}\big]\big)\big] \\ &\qquad \leq \sum_{i \geq -1} \mathbf{E}\big[\exp\big(-\theta S^B_{-i-1,0}\big)\big]\mathbf{E}\big[\exp\big(\theta S^A_{-i,n}\big)\big]. \end{aligned}$$

From (16) it can be seen that for any $\varepsilon > 0$ there exists $j > 0$ such that for all $i > j$ it holds that

$$\mathbf{E}\big[\exp\big(-\theta S^B_{-i-1,0}\big)\big] \leq \exp((i+2)(\Lambda_B(-\theta) + \varepsilon)). \tag{31}$$

Also from (14), we have that for $\theta < 0$ and for any $\varepsilon > 0$ there exists N such that, for all $n > N$ and all $i \geq -1$,

$$\mathbf{E}\big[\exp\big(\theta S^A_{-i,n}\big)\big] \leq \exp((n+i+1)(\Lambda_A^-(\theta) + \varepsilon)). \tag{32}$$

Fix now some $\theta < 0$ satisfying $\Lambda_A^-(\theta) + \Lambda_B(-\theta) < 0$ and some $\varepsilon > 0$ such that $\Lambda_A^-(\theta) + \Lambda_B(-\theta) + 2\varepsilon < 0$. Note that the existence of such a θ is guaranteed by the condition $\mathbf{E}[A] > \mathbf{E}[B]$ (see Figure 2). We then have that, for all $n > N$,

$$\begin{aligned}
&\mathbf{E}\Big[\exp\Big(-\theta \max_{i\geq -1}\big[S^B_{-i-1,0} - S^A_{-i,n}\big]\Big)\Big] \\
&\quad\leq \sum_{i=-1}^{j} \mathbf{E}\big[\exp(-\theta S^B_{-i-1,0})\big]\mathbf{E}\big[\exp(\theta S^A_{-i,n})\big] \\
&\qquad + \sum_{i>j} \mathbf{E}\big[\exp(-\theta S^B_{-i-1,0})\big]\mathbf{E}\big[\exp(\theta S^A_{-i,n})\big] \\
&\quad\leq \exp\big(n(\Lambda^-_A(\theta) + \varepsilon)\big) \qquad (33)\\
&\qquad \times \sum_{i=-1}^{j} \mathbf{E}\big[\exp(-\theta S^B_{-i-1,0})\big]\exp((i+1)(\Lambda^-_A(\theta)+\varepsilon)) \\
&\qquad + \exp\big(n(\Lambda^-_A(\theta) + \varepsilon)\big)\sum_{i>j}\exp(2\Lambda_B(-\theta) + \Lambda^-_A(\theta) + 3\varepsilon) \\
&\qquad\qquad\qquad \times\exp\big(i(\Lambda_B(-\theta) + \Lambda^-_A(\theta) + 2\varepsilon)\big) \\
&\quad\leq K(\theta, j, \varepsilon)\exp\big(n(\Lambda^-_A(\theta) + \varepsilon)\big),
\end{aligned}$$

where $K(\theta, j, \varepsilon)$ is some constant depending on θ, j and ε but not on n. To see that, notice that in the last inequality above we use the fact that the first sum is finite and the infinite geometric series in the second sum converges to a constant independent of n. From (33) we obtain

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbf{E}\Big[\exp\Big(-\theta\max_{i\geq -1}\big[S^B_{-i-1,0} - S^A_{-i,n}\big]\Big)\Big] \leq \Lambda^-_A(\theta) + \varepsilon. \tag{34}$$

Since this is true for all small enough $\varepsilon > 0$, the result follows. □

THEOREM 4.3. *The tail of the Palm distribution of the steady-state queue length, L, in a FCFS $G/G/1$ queue with arrivals and service times satisfying Assumption* B *is characterized by*

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbf{P}[L \geq n] = \Lambda^-_A(\theta^*), \tag{35}$$

where $\theta^* < 0$ is the smallest root of the equation

$$\Lambda^-_A(\theta) + \Lambda_B(-\theta) = 0. \tag{36}$$

PROOF. Due to stationarity, it suffices to characterize the tail distribution of L_n. For an upper bound, define

$$G_n \triangleq \max_{i\geq -1}\big[S^B_{-i-1,0} - S^A_{-i,n}\big]. \tag{37}$$

Using the Markov inequality and (29), we obtain

$$\mathbf{P}[L_n \geq n] = \mathbf{P}[G_n \geq 0] \leq \mathbf{E}[e^{-\theta G_n}]$$

for $\theta < 0$. Taking the limit as $n \to \infty$, using Lemma 4.2, and optimizing over θ to get the best bound, we obtain

$$(38)\quad \limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}[L_n \geq n] \leq \inf_{\{\theta | \Lambda_A^-(\theta) + \Lambda_B(-\theta) < 0\}} [\Lambda_A^-(\theta)] = \Lambda_A^-(\theta^*),$$

where the last equality is justified by Figure 2.

For a lower bound, set $i = \delta n$ for $\delta \geq 0$ (δn is assumed integer), and notice that

$$\begin{aligned} \mathbf{P}[L_{n-1} \geq n] &= \mathbf{P}[G_n \geq 0] \\ &\geq \sup_{\delta \geq 0} \mathbf{P}\big[S^B_{-\delta n-1,0} - S^A_{-\delta n,n} \geq 0\big]. \end{aligned}$$

The limiting log-moment generating function of $S^B_{-\delta n-1,0} - S^A_{-\delta n,n}$ is

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{E}\big[\exp\big(-\theta\big(S^B_{-\delta n-1,0} - S^A_{-\delta n,n}\big)\big)\big] = \delta\Lambda_B(-\theta) + (1+\delta)\Lambda_A^-(\theta)$$

and by using Assumption B, we obtain

$$(39)\quad \begin{aligned} \liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}[L_n \geq n] &\geq \sup_{\delta \geq 0}\Big(-\sup_{\theta}\big[-\delta(\Lambda_A^-(\theta) + \Lambda_B(-\theta)) - \Lambda_A^-(\theta)\big]\Big) \\ &= \sup_{\delta \geq 0} \inf_{\theta} \big[\delta(\Lambda_A^-(\theta) + \Lambda_B(-\theta)) + \Lambda_A^-(\theta)\big] \\ &= \inf_{\{\theta | \Lambda_A^-(\theta) + \Lambda_B(-\theta) < 0\}} [\Lambda_A^-(\theta)] \\ &= \Lambda_A^-(\theta^*), \end{aligned}$$

where the second equality follows by dualizing the constraint $\Lambda_A^-(\theta) + \Lambda_B(-\theta) < 0$. The lower bound in (39) along with (38) proves (35). □

REMARK. Intuitively, Theorem 4.3 asserts that for large enough n, we can state

$$(40)\quad \begin{aligned} &\mathbf{P}[L \geq n] \sim \exp(n\Lambda_A^-(\theta^*)) \\ &\qquad \text{where } \theta^* < 0 \text{ such that } \Lambda_A^-(\theta^*) + \Lambda_B(-\theta^*) = 0. \end{aligned}$$

5. The departure process of a $G/GI/1$ queue.

In this section we obtain a LDP for the process resulting from the passing-through-a-queue operation of our network model. That is, we establish a LDP for the steady-state departure process of a $G/GI/1$ queue, as seen by a typical departing customer. We denote by D_i, $i \in \mathbb{Z}$, the interdeparture time of the ith customer [the interval between the departure epochs of the $(i-1)$st and the ith customer]. As in Section 3, we assume that the interarrival process $\{A_i, i \in \mathbb{Z}\}$ is stationary and A_i are possibly dependent random variables. The service times B_i are independent and identically distributed (iid) random variables. The arrival and service processes are also assumed to satisfy

Assumptions B and C. As explained in Section 3, we will prove that the departure process satisfies Assumptions B and C when the arrival and service processes do.

We denote by $S^D_{1,n} \triangleq \sum_{i=1}^n D_i$, the partial sum of the departure process. The objective of this section is to prove a LDP for $S^D_{1,n}$. The interdeparture times can be expressed as follows:

$$D_i = B_i + I_i, \tag{41}$$

where B_i denotes the service time of the ith customer and I_i the idling period of the system that ended with the arrival of the ith customer ($I_i = 0$ if the ith customer finds the system busy upon arrival). By using the Lindley equation, one can obtain an expression for I_i and after some algebra derive an expression for $S^D_{1,n}$ in terms of the partial sums for the arrival and the service process. Using such an expression, one can prove a LDP for $S^D_{1,n}$. In this paper we follow a more intuitive approach. We derive an upper bound and a matching lower bound on $\mathbf{P}[S^D_{1,n} \le na]$ based on sample path arguments. To that effect, we explicitly characterize the most likely path leading to the large deviation of the departure process. The next proposition establishes an upper bound for the tail probability of $S^D_{1,n}$.

PROPOSITION 5.1 (Upper bound). *Under Assumption* B, *the partial sum* $S^D_{1,n}$ *of the departure process of a* $G/GI/1$ *queue under FCFS satisfies*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}\big[S^D_{1,n} \le na\big] \le -\Lambda^{*-}_D(a), \tag{42}$$

where

$$\Lambda^{*-}_D(a) \triangleq \Lambda^{*-}_B(a) + \Lambda^{*-}_\Gamma(a) \tag{43}$$

and

$$\Lambda^{*-}_\Gamma(a) \triangleq \sup_{\{\theta \mid \Lambda^-_A(\theta) + \Lambda_B(-\theta) < 0\}} \big[\theta a - \Lambda^-_A(\theta)\big]. \tag{44}$$

PROOF. Since $D_i \ge B_i$ for all i we obtain

$$S^D_{1,n} \ge S^B_{1,n}. \tag{45}$$

Consider some $j \le 1$ and let $(j-1)$ be the customer who initializes the busy period in which the 0th customer is served. Let t be the time that the $(j-1)$st customer arrived, t' the time that the $(j-1)$st customer departed, and t'' the time that the nth customer departed. Figure 5 depicts the situation. Note that

$$B_{j-1} + S^D_{j,n} \ge S^A_{j,n}. \tag{46}$$

Since the system is busy from the arrival of the $(j-1)$st customer until the departure of customer 0, we have

$$S^D_{j,0} = S^B_{j,0}. \tag{47}$$

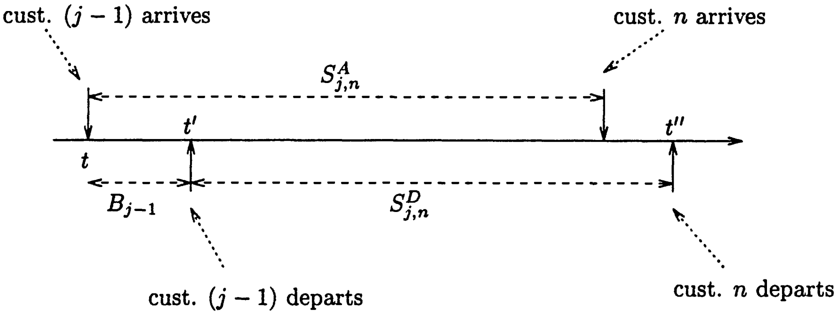

FIG. 5. *Deriving an upper bound on* $\mathbf{P}[S^D_{1,n} \le na]$. *Here, it is assumed that customer* $j-1$ *finds an empty queue.*

Therefore, from (47) and (46) we have

$$S^D_{1,n} = S^D_{j,n} - S^D_{j,0} \ge S^A_{j,n} - B_{j-1} - S^B_{j,0} = S^A_{j,n} - S^B_{j-1,0}. \tag{48}$$

Now, from (45) and (48) we obtain

$$\begin{aligned}\mathbf{P}\big[S^D_{1,n} \le na\big] &\le \mathbf{P}\big[S^B_{1,n} \le na,\ \exists\, j \le 1 \text{ s.t. } S^A_{j,n} - S^B_{j-1,0} \le na\big] \\ &= \mathbf{P}\big[S^B_{1,n} \le na\big]\mathbf{P}\Big[\min_{j\le 1}\big[S^A_{j,n} - S^B_{j-1,0}\big] \le na\Big],\end{aligned} \tag{49}$$

since the service times B_i are assumed to be independent and independent of the arrival process. Since $\min_{j\le 1}[S^A_{j,n} - S^B_{j-1,0}] = -\max_{j\le 1}[S^B_{j-1,0} - S^A_{j,n}]$, we use Lemma 4.2 to obtain

$$\limsup_{n\to\infty} \frac{1}{n}\log \mathbf{E}\Big[\exp\Big(\theta \min_{j\le 1}\big[S^A_{j,n} - S^B_{j-1,0}\big]\Big)\Big] \le \Lambda^-_A(\theta), \tag{50}$$

for $\theta < 0$, satisfying $\Lambda^-_A(\theta) + \Lambda_B(-\theta) < 0$.

Using Markov's inequality, we obtain

$$\limsup_{n\to\infty} \frac{1}{n}\log \mathbf{P}\Big[\min_{j\le 1}\big[S^A_{j,n} - S^B_{j-1,0}\big] \le na\Big] \le \Lambda^-_A(\theta) - \theta a.$$

Optimizing over θ to obtain the tightest bound, we finally find

$$\begin{aligned}&\limsup_{n\to\infty} \frac{1}{n}\log \mathbf{P}\Big[\min_{j\le 1}\big[S^A_{j,n} - S^B_{j-1,0}\big] \le na\Big] \\ &\quad \le - \sup_{\{\theta \mid \Lambda^-_A(\theta) + \Lambda_B(-\theta) < 0\}} \big[\theta a - \Lambda^-_A(\theta)\big].\end{aligned} \tag{51}$$

Moreover, from Assumption B we can assert that

$$\limsup_{n\to\infty} \frac{1}{n}\log \mathbf{P}\big[S^B_{1,n} \le na\big] \le -\Lambda^{*-}_B(a). \tag{52}$$

Combining (52) and (51) along with (49), we obtain (42). □

Obtaining a lower bound on the tail probability of $S^D_{1,n}$ is much more involved. Assumption B, which provides a LDP for the partial sums $S^A_{1,n}$, $S^B_{1,n}$ of the interarrival and service times, is not sufficient. Assumption C, which provides a LDP for the partial sum processes $\{S^A_{1,j},\ j = 1,\dots,n\}$ and $\{S^B_{i,j},\ 1 \le i \le j \le n\}$, is required. In the next proposition we derive a lower bound on the tail probability of $S^D_{1,n}$ and we prove that the departure process $\{S^D_{1,i},\ i = 1,\dots,n\}$ satisfies Assumption C(i).

PROPOSITION 5.2 (Lower bound). *Under Assumptions* B *and* C, *the partial sum $S^D_{1,n}$ of the departure process of a $G/GI/1$ queue under FCFS satisfies*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left[S^D_{1,n} \le na\right] \ge -\Lambda^{*-}_D(a). \tag{53}$$

Moreover, the departure process $\{S^D_{1,i},\ i = 1,\dots,n\}$ satisfies Assumption C(i).

PROOF. Fix $\varepsilon_1, \varepsilon_2 > 0$, $\zeta \ge 0$ and $y_1, y_2 \ge 0$ such that $y_1 - y_2 = a$ and $y_1/(1+\zeta) \ge a$. Consider the set of all sample paths that satisfy

$$S^B_{j,k} \le (k+1-j)a + \varepsilon_1 n, \qquad 1 \le j \le k \le n, \tag{54}$$

$$S^A_{-\zeta n,k} \le (\zeta n + k - 1)\frac{y_1}{1+\zeta} + \varepsilon_1 n, \qquad k = 1,\dots,n \tag{55}$$

and

$$S^B_{-\zeta n-1,0} \ge n y_2 - \varepsilon_1 n. \tag{56}$$

We state the following lemma, the proof of which is deferred until the end of the current proof.

LEMMA 5.3. *For any sample path that satisfies* (54), (55) *and* (56), *we have*

$$S^D_{1,k} \le ka + 4\varepsilon_1 n, \qquad k = 1,\dots,n. \tag{57}$$

Therefore,

$$\begin{aligned}
&\mathbf{P}\left[S^D_{1,k} \le ka + 4\varepsilon_1 n,\ k = 1,\dots,n\right] \\
&\quad\ge \mathbf{P}\left[S^B_{j,k} \le (k+1-j)a + \varepsilon_1 n,\ 1 \le j \le k \le n\right] \\
&\qquad\times \sup_{\{\zeta\ge 0 \mid y_1/(1+\zeta)\ge a\}} \ \sup_{y_1-y_2=a} \mathbf{P}\Big[S^A_{-\zeta n,k} \le (\zeta n + k - 1)(y_1/(1+\zeta)) \\
&\qquad\qquad\qquad\qquad + \varepsilon_1 n,\ k = 1,\dots,n\Big] \\
&\qquad\times \mathbf{P}\left[S^B_{-\zeta n-1,0} \ge n y_2 - \varepsilon_1 n\right] \qquad (58)\\
&\quad\ge \sup_{\{\zeta\ge 0 \mid y_1/(1+\zeta)\ge a\}} \ \sup_{y_1-y_2=a} \exp\Bigg\{-n\left(\Lambda^{*-}_B(a) + \varepsilon'\right) \\
&\qquad\qquad -n\left[\Lambda^{*-}_A\left(\frac{y_1}{1+\zeta}\right)(1+\zeta) + \varepsilon''\right] \\
&\qquad\qquad -n\left[\Lambda^{*+}_B\left(\frac{y_2-\varepsilon_1}{\zeta}\right)\zeta + \varepsilon'''\right]\Bigg\},
\end{aligned}$$

where the last inequality holds for large n and is obtained by applying Assumption C to the arrival and service processes. We can now choose appropriate ε', ε'' and ε''' such that, for sufficiently large n and given ε_2, we have

$$\begin{aligned}&\mathbf{P}\big[S^D_{1,k} \le ka + 4\varepsilon_1 n,\ k = 1,\dots,n\big]\\ (59)\qquad &\ge \sup_{\{\zeta\ge 0 \mid y_1/(1+\zeta)\ge a\}} \sup_{y_1-y_2=a} \exp\Bigg\{-n\Bigg[\Lambda_B^{*-}(a) + \Lambda_A^{*-}\bigg(\frac{y_1}{1+\zeta}\bigg)(1+\zeta)\\ &\qquad\qquad + \Lambda_B^{*+}\bigg(\frac{y_2}{\zeta}\bigg)\zeta + \varepsilon_2\Bigg]\Bigg\}.\end{aligned}$$

We now argue that the constraint $y_1/(1+\zeta) \ge a$ can be removed from the optimization in (59). Consider a choice of $y_1 = \tilde{y}_1$, $y_2 = \tilde{y}_2$ and $\zeta = \tilde{\zeta}$ such that $\tilde{y}_1 - \tilde{y}_2 = a$ and $\tilde{y}_1/(1+\tilde{\zeta}) < a$. Let us now consider a feasible solution of the above optimization problem with $\zeta = 0$, $y_1 = a$, $y_2 = 0$, and cost which is approximately $\exp(-n[\Lambda_B^{*-}(a) + \Lambda_A^{*-}(a)])$ (omitting the ε terms). Now note that since $\tilde{y}_1/(1+\tilde{\zeta}) < a$ and $\Lambda_A^{*-}(\cdot)$ nonincreasing, we have

$$\begin{aligned}&\exp\{-n[\Lambda_B^{*-}(a) + \Lambda_A^{*-}(a)]\}\\ &\qquad \ge \exp\Bigg\{-n\Bigg[\Lambda_B^{*-}(a) + \Lambda_A^{*-}\bigg(\frac{\tilde{y}_1}{1+\tilde{\zeta}}\bigg)(1+\tilde{\zeta}) + \Lambda_B^{*+}\bigg(\frac{\tilde{y}_2}{\tilde{\zeta}}\bigg)\tilde{\zeta}\Bigg]\Bigg\}.\end{aligned}$$

This shows that there exist choices of y_1, y_2 and ζ satisfying $y_1/(1+\zeta) \ge a$ that have a better exponent. Hence, the constraint $y_1/(1+\zeta) \ge a$ can indeed be removed.

We now use convex analysis to prove that $\Lambda_\Gamma^{*-}(a)$ as defined in (44) is equal to

$$-\sup_{\zeta\ge 0}\ \sup_{y_1-y_2=a} \bigg\{-(1+\zeta)\Lambda_A^{*-}\bigg(\frac{y_1}{1+\zeta}\bigg) - \zeta\Lambda_B^{*+}\bigg(\frac{y_2}{\zeta}\bigg)\bigg\},$$

thus proving that the lower bound in (59) (taking $\varepsilon_2 \to 0$) matches the upper bound obtained in Proposition 5.1. Dualizing the constraint $\Lambda_A^-(\theta) + \Lambda_B(-\theta) < 0$ we obtain [note that $\Lambda_A^-(\theta) + \Lambda_B(-\theta) < 0$ if and only if $\Lambda_A^-(\theta) + \Lambda_B^+(-\theta) < 0$]

$$\begin{aligned}-\Lambda_\Gamma^{*-}(a) &= -\sup_{\{\theta \mid \Lambda_A^-(\theta)+\Lambda_B(-\theta)<0\}} [\theta a - \Lambda_A^-(\theta)]\\ &= \inf_{\{\theta \mid \Lambda_A^-(\theta)+\Lambda_B(-\theta)<0\}} [-\theta a + \Lambda_A^-(\theta)]\\ (60)\qquad &= \sup_{\zeta\ge 0}\Big\{-\sup_{\theta}[\theta a - (1+\zeta)\Lambda_A^-(\theta) - \zeta\Lambda_B^+(-\theta)]\Big\}\\ &= \sup_{\zeta\ge 0}\Bigg\{-\inf_{y_1-y_2=a}\bigg[(1+\zeta)\Lambda_A^{*-}\bigg(\frac{y_1}{1+\zeta}\bigg) + \zeta\Lambda_B^{*+}\bigg(\frac{y_2}{\zeta}\bigg)\bigg]\Bigg\}\\ &= \sup_{\zeta\ge 0}\ \sup_{y_1-y_2=a}\bigg[-(1+\zeta)\Lambda_A^{*-}\bigg(\frac{y_1}{1+\zeta}\bigg) - \zeta\Lambda_B^{*+}\bigg(\frac{y_2}{\zeta}\bigg)\bigg].\end{aligned}$$

To see that, note that for convex functions f, f_1, f_2 and for a scalar $c \ge 0$, it holds that $(cf)^*(x^*) = cf^*(x^*/c)$, and $(f_1 + f_2)^*(x^*) = \inf_{x_1^* + x_2^* = x^*}[f_1^*(x_1^*) + f_2^*(x_2^*)]$ (see [29], Theorem 16.1, Theorem 16.4).

In summary, we have verified that Assumption C(i) holds for the departure process; that is,

$$\mathbf{P}\big[S^D_{1,i} \le ia + 4\varepsilon_1 n,\ i = 1, \dots, n\big] \ge \exp\big(-n(\Lambda_D^{*-}(a) + \varepsilon_2)\big). \tag{61}$$

By taking $\varepsilon_1, \varepsilon_2 \to 0$ and since $\mathbf{P}[S^D_{1,n} \le na]$ is clearly larger than the probability in (61), (53) is verified for the same region. □

Proof of Lemma 5.3. Note that for $k = 1, \dots, n$ from (55) and (56) we obtain

$$\begin{aligned} S^A_{-\zeta n,k} &\le (\zeta n + k - 1)\frac{y_1}{1 + \zeta} + \varepsilon_1 n \\ &\le (n y_2 - \varepsilon_1 n) + \big((k - 1)a + 2\varepsilon_1 n\big) \\ &\le (k - 1)a + 2\varepsilon_1 n + S^B_{-\zeta n - 1,0}, \end{aligned} \tag{62}$$

where the second inequality holds because the two sides are equal at $k = n + 1$ and because $y_1/(1 + \zeta) \ge a$. The third inequality is justified by (54) and (56).

Let t be the arrival time of customer $-\zeta n - 1$. Then customer k arrives at time $t + S^A_{-\zeta n,k}$. We distinguish two cases. In case 1, customer k finds an empty system upon arrival. Then it departs at time t' where

$$t' = t + S^A_{-\zeta n,k} + B_k \le ka + 3\varepsilon_1 n + t + S^B_{-\zeta n - 1,0}, \tag{63}$$

by using (54) and (62). Let t'' the departure time of the 0th customer. Clearly, $t'' \ge t + S^B_{-\zeta n - 1,0}$, which along with (63) implies that $t' - t'' \le ka + 3\varepsilon_1 n \le ka + 4\varepsilon_1 n$. However, according to their definition, $t' - t'' = S^D_{1,k}$.

In case 2, customer k finds a busy system upon arrival, in which case $D_k = B_k$. Then, if this is also true for all $i = 1, \dots, k - 1$, we have $S^D_{1,k} = S^B_{1,k} \le ka + \varepsilon_1 n \le ka + 4\varepsilon_1 n$. If not, let $i \in [1, \dots, k - 1]$ be the latest customer that finds the system empty (i.e., the one with maximum index). To bound $S^D_{1,i}$, we use the argument of case 1. Thus,

$$\begin{aligned} S^D_{1,k} &= S^D_{1,i} + S^D_{i+1,k} = S^D_{1,i} + S^B_{i+1,k} \le ia + 3\varepsilon_1 n + (k - i)a + \varepsilon_1 n \\ &= ka + 4\varepsilon_1 n, \end{aligned}$$

where we have used (54) in the last inequality. □

The proof of the above theorem indicates a most likely path along which the large deviation of $S^D_{1,n}$ occurs (in the sense that its probability is comparable to $\mathbf{P}[S^D_{1,n} \le na]$). Let ζ^*, y_1^* and y_2^* be a solution of the optimization problem in (59). The large deviation in $S^D_{1,n}$ occurs by the following:

1. Maintaining an empirical arrival rate of at least $(1 + \zeta^*)/y_1^*$ from the arrival of customer $-\zeta^* n - 1$, until the departure of the nth customer,

and an empirical service rate of at most ζ^*/y_2^* from the arrival of customer $-\zeta^* n - 1$, until the departure of the 0th customer.

2. Maintaining an empirical service rate of at least $1/a$ from the departure of the 0th customer until the departure of the nth customer.

Figure 6 illustrates the situation.

Combining Propositions 5.1 and 5.2, we obtain the following theorem.

THEOREM 5.4. *Under Assumptions* B *and* C, *the partial sum* $S_{1,n}^D$ *of the departure process of a* $G/GI/1$ *queue under FCFS satisfies*

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbf{P}\left[S_{1,n}^D \le na\right] = -\Lambda_D^{*-}(a), \tag{64}$$

where

$$\Lambda_D^{*-}(a) = \Lambda_B^{*-}(a) + \Lambda_\Gamma^{*-}(a)$$

and

$$\Lambda_\Gamma^{*-}(a) = \sup_{\{\theta \mid \Lambda_A^{-}(\theta) + \Lambda_B(-\theta) < 0\}} \left[\theta a - \Lambda_A^{-}(\theta)\right].$$

Throughout this section we have assumed that the service times B_i are iid. A close examination of the proofs of Propositions 5.1 and 5.2 suggests that a weaker condition is sufficient for our purposes. Namely, we only need the random variables $S_{j,0}^B$ and $S_{1,n}^B$ to be approximately independent for every $j \le 0$, as $n \to \infty$. A mixing condition of the type $\mathbf{E}[\exp(\theta S_{j,0}^B)\exp(\theta S_{1,n}^B)] = \mathbf{E}[\exp(\theta S_{j,0}^B)\mathbf{E}[\exp(\theta S_{1,n}^B)]\exp(n\varepsilon(n))$ for every $j \le 0$ and θ, where $\lim_{n \to \infty} \varepsilon(n) = 0$, is sufficient.

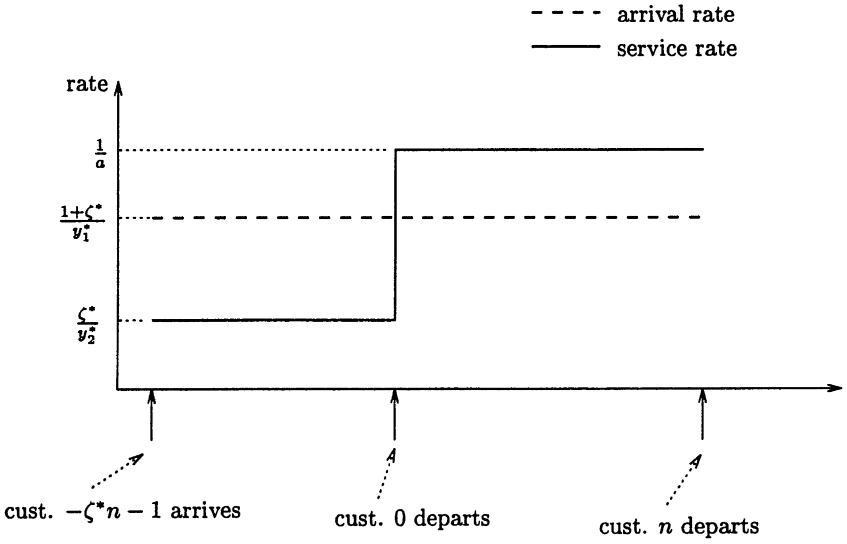

FIG. 6. *A most likely path for large deviations of* $S_{1,n}^D$.

An alternative expression for $\Lambda_D^{*-}(\cdot)$ which is a consequence of the defining equation (43) is

$$\begin{aligned}\Lambda_D^{*-}(a) &= \Lambda_B^{*-}(a) + \Lambda_\Gamma^{*-}(a)\\ &= \begin{cases}\Lambda_B^{*-}(a) + \Lambda_A^{*-}(a), & \text{if } a \geq \Lambda_A^{-\prime}(\theta^*),\\ \Lambda_B^{*-}(a) + \theta^* a - \Lambda_A^{-}(\theta^*), & \text{if } a < \Lambda_A^{-\prime}(\theta^*),\end{cases}\end{aligned} \tag{65}$$

where θ^* is defined in the statement of Theorem 4.1 and $\Lambda_A^{-\prime}(x)$ denotes the derivative of $\Lambda_A^{-}(\cdot)$ evaluated at x. To see that, consult Figure 2 and notice that the first branch of (65) corresponds to the region of a where the constraint $\Lambda_A^{-}(\theta) + \Lambda_B(-\theta) < 0$ is not tight and the second branch to the region of a where this constraint is tight.

We now argue that the passing-through-a-queue operation preserves Assumption B. Proposition 5.2 establishes that it preserves Assumption C(i). Notice first that we have proved a one-sided LDP for the departure process with large deviations rate function expressed as a function of the large deviations rate function of the arrival and service processes. Using Varadhan's integral lemma, the limiting log-moment generating function can be obtained as the convex dual of the large deviations rate function. (See [15], Section 4.3. That is, we let $\phi(x) = \theta x$ and use a one-sided version of Varadhan's lemma, where the left side of $\Lambda(\theta)$, that is, for $\theta < 0$, can be obtained as the convex dual of $\Lambda^*(a)$, for a less than the mean.)

Moreover, by (41) we have

$$D_i \leq B_i + A_i,$$

which implies that D_i has moments of all orders since B_i and A_i do. Finally, we establish that $\Lambda_D^{*-}(\cdot)$ is strictly convex in the intersection of the interior of its domain and the interval $(-\infty, \mathbf{E}[D_1])$. To this end we will use the expression in (65). By a stability argument $\mathbf{E}[D_i] = \mathbf{E}[A_i]$, thus the function in the upper branch of (65) is strictly convex in the interval $(\Lambda_A^{-\prime}(\theta^*), \mathbf{E}[D_1])$ due to the strict convexity of $\Lambda_A^{*-}(\cdot)$. Let now θ_1 be such that $\Lambda_A^{-\prime}(\theta_1) = \Lambda'_B(\theta_1)$. Notice that (see Figure 2)

$$\Lambda_A^{-\prime}(\theta^*) \leq \Lambda_A^{-\prime}(\theta_1) = \Lambda'_B(\theta_1) \leq \mathbf{E}[B_1],$$

which implies that in the lower branch of (65) we have $a < \mathbf{E}[B_1]$. Hence, the strict convexity of $\Lambda_B^{*-}(a)$ for $a < \mathbf{E}[B_1]$ suffices to guarantee the convexity of the function in the lower branch of (65).

To obtain the limiting log-moment generating function for the partial sum of the departure process, we take the convex dual of $\Lambda_D^{*-}(\cdot)$ in (65). Using the duality correspondences proved in [29], Section 16, we obtain the following corollary.

COROLLARY 5.5. *Under Assumptions* B *and* C *we have*

$$\Lambda_D^-(\theta) = \begin{cases} \inf_{\theta_1+\theta_2=\theta}\{\Lambda_B^-(\theta_1) + \Lambda_A^-(\theta_2)\}, & \text{if } \theta \ge \hat{\theta}, \\ \Lambda_B^-(\theta - \theta^*) + \Lambda_A^-(\theta^*), & \text{if } \theta < \hat{\theta}, \end{cases} \tag{66}$$

where

$$\hat{\theta} \triangleq \frac{d}{da}\left[\Lambda_B^{*-}(a) + \Lambda_A^{*-}(a)\right]_{a=\Lambda_A^{-\prime}(\theta^*)}. \tag{67}$$

It is instructive to determine the fluctuations of the queue length that lead to a large deviation in the departure process. Let ζ^* solve the optimization problem in (59). Let t be the arrival time of customer $-\zeta^* n - 1$. The 0th customer arrives at $t + S^A_{-\zeta^* n,0}$ and departs no earlier than $t + S^B_{-\zeta^* n-1,0}$. Thus, for the waiting time of customer 0 holds

$$W_0 \ge t + S^B_{-\zeta^* n-1,0} - t - S^A_{-\zeta^* n,0} = S^B_{-\zeta^* n-1,0} - S^A_{-\zeta^* n,0} \triangleq \tilde{W}_0. \tag{68}$$

A close examination of the proofs of Propositions 5.1 and 5.2 suggests that $\Lambda_\Gamma^{*-}(\cdot)$ is the large deviations rate function of the process

$$\left\{S^A_{-\zeta^* n,k} - S^B_{-\zeta^* n-1,0},\ k = 1,\dots,n\right\} \equiv \left\{S^A_{1,k} - \tilde{W}_0,\ k = 1,\dots,n\right\}. \tag{69}$$

From the above discussion and (65), we conclude that, depending on the value of a, we can distinguish two cases for the large deviation in the departure process to occur.

1. $a \ge \Lambda_A^{-\prime}(\theta^*)$: in this region, $\Lambda_\Gamma^{*-}(a) = \Lambda_A^{*-}(a)$ and from (69) it is clear that the most likely way for the large deviation in the departure process to occur is the 0th customer to incur $O(1)$ waiting time, which implies that it finds a queue length of $O(1)$ upon arrival.
2. $a < \Lambda_A^{-\prime}(\theta^*)$: in this region, $\Lambda_\Gamma^{*-}(a) = \theta^* a - \Lambda_A^-(\theta^*)$ and from (69) it is clear that the most likely way for the large deviation in the departure process to occur is the 0th customer to incur a large waiting time (recall from Theorem 4.1 that the large deviations rate function for the waiting time is linear with slope θ^*).

Hence, also taking into account Figure 6, we can infer for the queue length the cases depicted in Figure 7. In region 2 and in contrast with region 1, the queue builds up to lead to a large deviation in the departure process.

5.1. *Special cases.* In this section we apply Theorem 5.4 to two special cases. Namely, we study the departure process, in the large deviations regime, of an $M/M/1$ queue and a $G/D/1$ queue.

The departure process of a $G/D/1$ *queue.* We assume, as in Section 5, that the interarrival times process $\{A_i,\ i \in \mathbb{Z}\}$ is stationary and A_i are possibly dependent random variables. The service times B_i are iid random variables and equal to c w.p.1. Interarrival and service times are assumed independent.

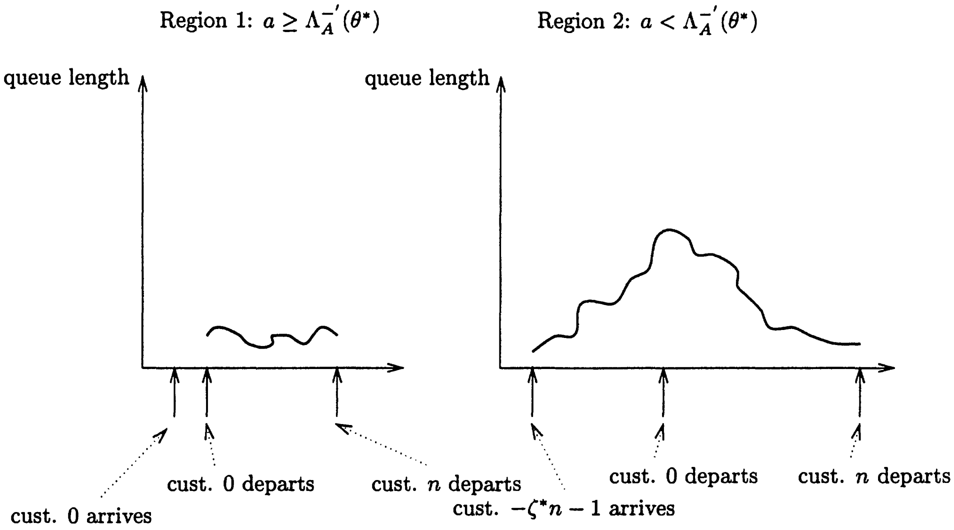
Region 1: $a \geq \Lambda_A^{-\prime}(\theta^*)$
Region 2: $a < \Lambda_A^{-\prime}(\theta^*)$
queue length
queue length
cust. 0 arrives
cust. 0 departs
cust. n departs
cust. $-\zeta^* n - 1$ arrives
cust. 0 departs
cust. n departs

FIG. 7. *Two cases for the queue length: In Region* 1, *the* 0*th customer finds an* $O(1)$ *queue upon arrival and until the nth customer departs the queue stays at an* $O(1)$ *level. In Region* 2, *the queue first builds up* (*see also the arrival and service rates in Figure* 6) *and then it is depleted, resulting in the large deviation in the departure process.*

It is straightforward that $\Lambda_B(\theta) = c\theta$. Therefore a simple calculation yields

$$\Lambda_B^{*-}(a) = \begin{cases} +\infty, & \text{if } a < c, \\ 0, & \text{if } a \geq c. \end{cases} \tag{70}$$

Moreover,

$$\Lambda_\Gamma^{*-}(a) = \sup_{\{\theta \mid \Lambda_A^-(\theta) - c\theta < 0\}} \left[\theta a - \Lambda_A^-(\theta)\right] = \hat{\theta} a - \Lambda_A^-(\hat{\theta}), \tag{71}$$

where $\hat{\theta}$ is the optimizing θ. Note that by taking $a \geq c$, we have $\Lambda_\Gamma^{*-}(a) = \Lambda_A^{*-}(a)$ (see Figure 8). Therefore, using (43),

$$\Lambda_D^{*-}(a) = \begin{cases} +\infty, & \text{if } a < c, \\ \Lambda_A^{*-}(a), & \text{if } a \geq c. \end{cases} \tag{72}$$

This is exactly the result obtained in [16] for a discrete time model. Taking the convex dual of the above we obtain

$$\Lambda_D^-(\theta) = \inf_{\theta = \theta_1 + \theta_2} \left[\Lambda_A^-(\theta_1) + \delta^*(\theta_2 \mid [c, \infty))\right], \tag{73}$$

where $\delta^*(\theta_2 \mid [c, \infty))$ is the *support* function of the set $[c, \infty)$ and is defined as

$$\delta^*(\theta \mid [c, \infty)) \triangleq \sup\{\theta x \mid x \in [c, \infty)\} = \begin{cases} \infty, & \text{if } \theta > 0, \\ c\theta, & \text{if } \theta \leq 0. \end{cases}$$

The departure process of an $M/M/1$ *queue.* We assume that the arrival process is Poisson with rate λ and the service times are iid, distributed according to an exponential distribution with parameter μ.

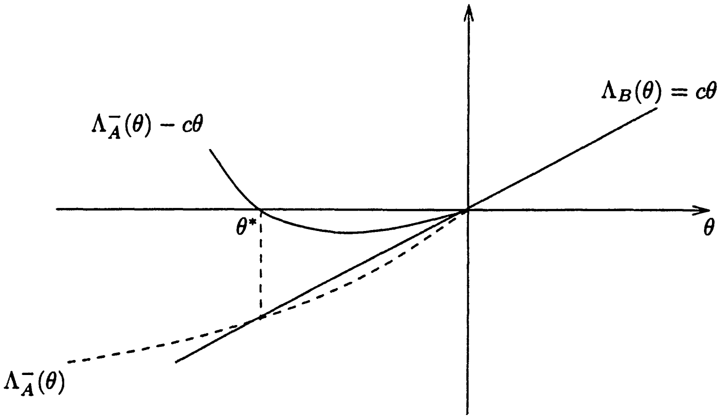

FIG. 8. *Notice that at θ^* we have $\Lambda_A^{-\prime}(\theta^*) \le c$. Thus, when $a \ge c$, the optimizer $\hat{\theta}$ satisfies $0 \ge \hat{\theta} \ge \theta^*$, which implies $\Lambda_\Gamma^{*-}(a) = \Lambda_A^{*-}(a)$.*

It is straightforward to calculate

$$\Lambda_A(\theta) = \log\left(\frac{\lambda}{\lambda - \theta}\right), \qquad \Lambda_B(\theta) = \log\left(\frac{\mu}{\mu - \theta}\right), \tag{74}$$

where $\Lambda_A(\theta)$ denotes the log-moment generating function of the arrival process. Now, notice that

$$\Lambda_A(\theta) + \Lambda_B(-\theta) = 0 \quad \Leftrightarrow \quad \frac{\lambda}{\lambda - \theta}\frac{\mu}{\mu + \theta} = 1 \quad \Leftrightarrow \quad \begin{matrix} \theta = 0, \\ \theta = \lambda - \mu, \end{matrix} \tag{75}$$

which implies that $\theta^* = \lambda - \mu$, where θ^* is defined in the statement of Theorem 4.1. Moreover, notice that

$$\Lambda'_A(\theta^*) = \frac{\lambda - \theta^*}{\lambda}\frac{\lambda}{(\lambda - \theta^*)^2} = \frac{1}{\mu}.$$

Thus, using (65), we obtain for $a \ge 1/\mu$,

$$\Lambda_D^{*-}(a) = \Lambda_B^{*-}(a) + \Lambda_A^{*-}(a) = \Lambda_A^{*-}(a), \tag{76}$$

since by definition $\Lambda_B^{*-}(a) = 0$ for $a \ge 1/\mu$. Using the second branch of (65), we obtain for $a < 1/\mu$,

$$\Lambda_D^{*-}(a) = \Lambda_B^{*-}(a) + a(\lambda - \mu) - \log(\lambda/\mu). \tag{77}$$

But

$$\Lambda_B^{*-}(a) = \sup_\theta\left[\theta a - \Lambda_B^-(\theta)\right] = a\mu - 1 - \log(a\mu),$$

since, by differentiating, the optimal θ is found equal to $(a\mu - 1)/a$. Thus, from (77), for $a < 1/\mu$,

$$\Lambda_D^{*-}(a) = a\lambda - 1 - \log(a\lambda) = \Lambda_A^{*-}(a). \tag{78}$$

Summarizing (76) and (78), we finally obtain

$$\Lambda_D^{*-}(a) = \Lambda_A^{*-}(a). \tag{79}$$

This result is in accordance with Burke's output theorem, which states that the departure process of an $M/M/1$ queue is Poisson with rate λ (see [25]).

6. Superposition of independent streams. In this section we treat the superposition operation of our network model. In particular, we derive a LDP for the process resulting from the superposition of independent arrival streams and we show that the superposition preserves Assumptions B and C(i). However, as will become clear in the sequel, in order to derive this LDP we need a result that connects, in the large deviations regime, the Palm distribution of the arrival process (i.e., as it is seen by a random customer) with its stationary distribution as seen at a random time. This result is presented in Section 6.1 and could be of independent interest.

Consider two independent arrival streams. By A_i^1 (resp. A_i^2), $i \in \mathbb{Z}$, we denote the interarrival time of the ith customer in stream 1 (resp. 2). We assume that the processes $\{A_i^1, A_i^2, i \in \mathbb{Z}\}$ are stationary and mutually independent. However, the interarrival times in each stream may be dependent. We impose Assumptions B and C(i) on the arrival process of each stream. We denote by $A_i^{1,2}$, $i \in \mathbb{Z}$, the interarrival times of the process resulting from the superposition. It should be noted that in order to derive the LDP for the superposition, Assumption C is not used.

The next theorem establishes a LDP for the partial sum $S_{1,n}^{A^{1,2}}$ of the aggregate process, resulting from the superposition of streams 1 and 2.

THEOREM 6.1. *Under Assumption* B, *the partial sum* $S_{1,n}^{A^{1,2}}$ *of the aggregate process, resulting from the superposition of the independent processes* A_i^1, A_i^2, $i \in \mathbb{Z}$, *satisfies*

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\log\mathbf{P}\left[S_{1,n}^{A^{1,2}} \le na\right] &= -\inf_{\substack{\delta_1+\delta_2=1\\ \delta_1,\delta_2\ge 0}}\left[\delta_1\Lambda_{A^1}^{*}(a/\delta_1) + \delta_2\Lambda_{A^2}^{*-}(a/\delta_2)\right]\\ &\triangleq -\Lambda_{A^{1,2}}^{*-}(a).\end{aligned} \tag{80}$$

PROOF. Consult Figure 9. Consider the partial sum $S_{1,n}^{A^{1,2}}$ and let H_1 (resp. H_2) denote the event that the first customer of the aggregate process originates from stream 1 (resp. 2). We first obtain an upper bound on

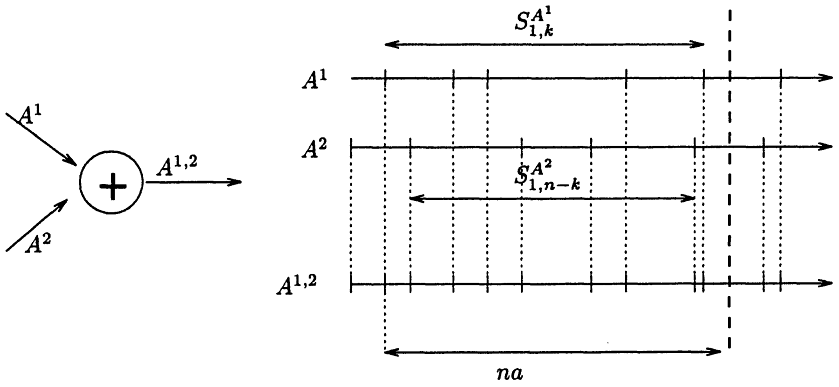

FIG. 9. *Superposition of two independent streams.*

$\mathbf{P}[S_{1,n}^{A^{1,2}} \le na|H_1]$. Notice that

$$(81)\qquad \mathbf{P}\Big[S_{1,n}^{A^{1,2}} \le na|H_1\Big] \le \sum_{k=1}^{n} \mathbf{P}\Big[S_{1,k}^{A^1} \le na\Big]\mathbf{P}_R\Big[S_{1,n-k}^{A^2} \le na\Big].$$

Here, $\mathbf{P}[\cdot]$ denotes the probability distribution seen by a random customer (Palm distribution) and $\mathbf{P}_R[\cdot]$ denotes the probability distribution seen at a random time. Due to the independence of the two arrival streams, an arrival originating from stream 1 constitutes a random incidence in the arrival process of stream 2, and therefore we are interested in the probability distribution seen at a random time for events concerning stream 2.

In Section 6.1 it is shown that

$$(82)\quad \lim_{n\to\infty} \frac{1}{n}\log \mathbf{P}_R\Big[S_{1,n}^{A^2} \le na\Big] = \lim_{n\to\infty} \frac{1}{n}\log \mathbf{P}\Big[S_{1,n}^{A^2} \le na\Big] = -\Lambda_{A^2}^{*-}(a).$$

Therefore, from (81), letting $k = n\delta$, $\delta \in [0,1]$ ($n\delta$ is assumed integer) and taking large n, we obtain

$$\begin{aligned} \mathbf{P}\Big[S_{1,n}^{A^{1,2}} \le na \mid H_1\Big] &\le \sum_{\delta\in[0,1]} \mathbf{P}\Big[S_{1,n\delta}^{A^1} \le na\Big]\mathbf{P}_R\Big[S_{1,n(1-\delta)}^{A^2} \le na\Big] \\ &\le n \sup_{\delta\in[0,1]} \mathbf{P}\Big[S_{1,n\delta}^{A^1} \le na\Big]\mathbf{P}_R\Big[S_{1,n(1-\delta)}^{A^2} \le na\Big], \end{aligned}$$

which implies

$$(83)\qquad \begin{aligned} &\limsup_{n\to\infty} \frac{1}{n}\log \mathbf{P}\Big[S_{1,n}^{A^{1,2}} \le na|H_1\Big] \\ &\quad\le -\inf_{\delta\in[0,1]} \big[\delta\Lambda_{A^1}^{*-}(a/\delta) + (1-\delta)\Lambda_{A^2}^{*-}(a/(1-\delta))\big] \\ &\quad= -\inf_{\substack{\delta_1+\delta_2=1\\ \delta_1,\delta_2\ge 0}} \big[\delta_1\Lambda_{A^1}^{*-}(a/\delta_1) + \delta_2\Lambda_{A^2}^{*-}(a/\delta)\big]. \end{aligned}$$

To obtain a lower bound notice that

$$\mathbf{P}\Big[S_{1,n}^{A^{1,2}} \le na | H_1\Big] \ge \sup_{\delta \in [0,1]} \mathbf{P}\Big[S_{1,n\delta}^{A^1} \le na\Big]\mathbf{P}_R\Big[S_{1,n(1-\delta)}^{A^2} \le na\Big]$$

which implies

$$\begin{aligned} &\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\Big[S_{1,n}^{A^{1,2}} \le na | H_1\Big] \\ &\quad \ge - \inf_{\delta \in [0,1]} \big[\delta \Lambda_{A^1}^{*-}(a/\delta) + (1-\delta)\Lambda_{A^2}^{*-}(a/(1-\delta))\big] \\ &\quad = - \inf_{\substack{\delta_1+\delta_2=1 \\ \delta_1,\,\delta_2 \ge 0}} \big[\delta_1 \Lambda_{A^1}^{*-}(a/\delta_1) + \delta_2 \Lambda_{A^2}^{*-}(a/\delta_2)\big]. \end{aligned} \tag{84}$$

Finally, observe that because of symmetry, (83) and (84) also hold for $\mathbf{P}[S_{1,n}^{A^{1,2}} \le na | H_2]$. This along with the fact that

$$\mathbf{P}\Big[S_{1,n}^{A^{1,2}} \le na\Big] = \mathbf{P}\Big[S_{1,n}^{A^{1,2}} \le na \,|\, H_1\Big]\mathbf{P}[H_1] + \mathbf{P}\Big[S_{1,n}^{A^{1,2}} \le na \,|\, H_2\Big]\mathbf{P}[H_2]$$

proves the theorem. □

REMARK. Let δ_1^*, δ_2^* be a solution to the optimization problem in (80). It can be seen that a most likely path to have a large deviation in the aggregate process is to maintain an empirical arrival rate of δ_1^*/a in stream 1 and a rate of δ_2^*/a in stream 2. Then, since $\delta_1^* + \delta_2^* = 1$, the empirical rate of the aggregate process is $1/a$.

Using induction on the number of streams superimposed, we generalize Theorem 6.1 to obtain the following corollary.

COROLLARY 6.2. *Under Assumption* B, *the partial sum* $S_{1,n}^{A^{1,\dots,m}}$ *of the aggregate process, resulting from the superposition of the m independent processes* $A_i^1, \dots, A_i^m$, $i \in \mathbb{Z}$, *satisfies*

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\Big[S_{1,n}^{A^{1,\dots,m}} \le na\Big] &= - \inf_{\substack{\delta_1 + \cdots + \delta_m = 1 \\ \delta_1, \dots, \delta_m \ge 0}} \sum_{k=1}^{m} \delta_k \Lambda_{A^k}^{*-}(a/\delta_k) \\ &\triangleq -\Lambda_{A^{1,\dots,m}}^{*}(a). \end{aligned} \tag{85}$$

Using convex duality, by Varadhan's integral lemma, one can obtain the limiting log-moment generating function $\Lambda_{A^{1,\dots,m}}^{-}(\cdot)$ of $S_{1,n}^{A^{1,\dots,m}}$ as the convex dual of its large deviations rate function $\Lambda_{A^{1,\dots,m}}^{*-}(\cdot)$. The latter function is convex by [29], Theorem 5.8.

We now proceed into proving that the aggregate process, resulting from the superposition of independent streams which satisfy Assumptions B and C(i) also satisfies the same assumptions. We have proved a one-sided LDP for the superposition with large deviations rate function expressed in terms of the large deviations rate functions of the superimposed processes. Using Varadhan's integral lemma, the limiting log-moment generating function can be obtained as the convex dual of the large deviations rate function. In addition, the superposition process has moments of all orders since

$$A_i^{1,2} \le A_i^1 \quad \text{and} \quad A_i^{1,2} \le A_i^2.$$

Finally, we show that $\Lambda^*_{\bar{A}^{1,2}}(a)$ is strictly convex in the intersection of the interior of its domain and the interval $(-\infty, \mathbf{E}[A_i^{1,2}])$. To this end, first note that for a strictly convex function $f(\cdot)$, and as long as $a/\delta \neq b/\delta'$, we have

$$\delta f(a/\delta) + \delta' f(b/\delta') > (\delta + \delta') f((a+b)/(\delta + \delta')). \tag{86}$$

Consider now the definition of $\Lambda^*_{\bar{A}^{1,2}}(a)$, which we rewrite as

$$\Lambda^*_{\bar{A}^{1,2}}(a) = \inf_{\delta \in [0,1]} \left[\delta \Lambda^*_{\bar{A}^1}(a/\delta) + (1-\delta)\Lambda^*_{\bar{A}^2}(a/(1-\delta)) \right].$$

Let δ (resp. δ') be the minimizer in the optimization problem corresponding to $\Lambda^*_{\bar{A}^{1,2}}(a)$ [resp. $\Lambda^*_{\bar{A}^{1,2}}(b)$]. We then have

$$\begin{aligned}
&\frac{1}{2}\Lambda^*_{\bar{A}^{1,2}}(a) + \frac{1}{2}\Lambda^*_{\bar{A}^{1,2}}(b) \\
&\quad = \frac{\delta}{2}\Lambda^*_{\bar{A}^1}\left(\frac{a}{\delta}\right) + \frac{1-\delta}{2}\Lambda^*_{\bar{A}^2}\left(\frac{a}{1-\delta}\right) + \frac{\delta'}{2}\Lambda^*_{\bar{A}^1}\left(\frac{b}{\delta'}\right) \\
&\qquad + \frac{1-\delta'}{2}\Lambda^*_{\bar{A}^2}\left(\frac{b}{1-\delta'}\right) \\
&\quad \ge \frac{\delta+\delta'}{2}\Lambda^*_{\bar{A}^1}\left(\frac{a+b}{\delta+\delta'}\right) + \frac{2-\delta-\delta'}{2}\Lambda^*_{\bar{A}^2}\left(\frac{a+b}{2-\delta-\delta'}\right) \\
&\quad = \zeta \Lambda^*_{\bar{A}^1}\left(\frac{(a+b)/2}{\zeta}\right) + (1-\zeta)\Lambda^*_{\bar{A}^2}\left(\frac{(a+b)/2}{1-\zeta}\right) \\
&\quad \ge \inf_{\zeta \in [0,1]} \left[\zeta \Lambda^*_{\bar{A}^1}\left(\frac{(a+b)/2}{\zeta}\right) + (1-\zeta)\Lambda^*_{\bar{A}^2}\left(\frac{(a+b)/2}{1-\zeta}\right) \right] \\
&\quad = \Lambda^*_{\bar{A}^{1,2}}\left(\frac{a+b}{2}\right).
\end{aligned}$$

The first inequality above is due to (86) and is strict unless $a/\delta = b/\delta'$ and $a/(1-\delta) = b/(1-\delta')$, which implies that $\delta = \delta'$ and $a = b$. Thus, as long as $a \neq b$, we have established the strict convexity of $\Lambda^*_{\bar{A}^{1,2}}(\cdot)$.

Theorem 6.3 establishes that the process resulting from the superposition satisfies Assumption C(i).

THEOREM 6.3. *Assume that the m independent processes $A_i^1, \dots, A_i^m$, $i \in \mathbb{Z}$, satisfy Assumption* C(i). *The aggregate process resulting from their superposition also satisfies Assumption* C(i).

PROOF. It suffices to prove the result for $m = 2$ since by using induction we can prove it for any m. We need to prove that for every ε_1, ε_2, $a > 0$, there exists M_S such that for all $n \geq M_S$,

$$\exp\left(-n\left(\Lambda^{*-}_{A^{1,2}}(a) + \varepsilon_2\right)\right) \leq \mathbf{P}\left[S^{A^{1,2}}_{1,j} - ja \leq \varepsilon_1 n, j = 1, \dots, n\right]. \tag{87}$$

Following the steps of the proof of Theorem 6.1, we consider the scenario that a fraction δ of customers of the aggregate process originates from the A^1 process. Again, H_1 denotes the event that customer 1 of the aggregate process originates from the A^1 process. We have

$$\begin{aligned} &\mathbf{P}\left[S^{A^{1,2}}_{1,j} - ja \leq \varepsilon_1 n, j = 1, \dots, n | H_1\right] \\ &\quad \geq \sup_{\delta \in [0,1]} \Big[\mathbf{P}\left[S^{A^1}_{1,j\delta} - ja \leq \varepsilon_1 n, j = 1, \dots, n\right] \\ &\qquad \times \mathbf{P}_R\left[S^{A^2}_{1,j(1-\delta)} - ja \leq \varepsilon_1 n, j = 1, \dots, n\right]\Big]. \end{aligned} \tag{88}$$

Using Assumption C(i) for the A^1 stream, we obtain for large enough n,

$$\mathbf{P}\left[S^{A^1}_{1,j\delta} - ja \leq \varepsilon_1 n, j = 1, \dots, n\right] \geq \exp\left(-n\delta\left(\Lambda^{*-}_{A^1}(a/\delta) + \varepsilon'\right)\right). \tag{89}$$

In Section 6.1 (Lemma 6.6), it is shown that for large enough n,

$$\begin{aligned} &\mathbf{P}_R\left[S^{A^2}_{1,j(1-\delta)} - ja \leq \varepsilon_1 n, j = 1, \dots, n\right] \\ &\quad \geq \exp\left(-n(1-\delta)\left(\Lambda^{*-}_{A^2}(a/(1-\delta)) + \varepsilon''\right)\right). \end{aligned} \tag{90}$$

To obtain (87) it suffices to choose appropriate ε' and ε'' such that for large enough n and given ε_2,

$$\begin{aligned} &\exp\left(-n \inf_{\delta \in [0,1]} \left[\delta\left(\Lambda^{*-}_{A^1}(a/\delta) + \varepsilon'\right) + (1-\delta)\left(\Lambda^{*-}_{A^2}(a/(1-\delta)) + \varepsilon''\right)\right]\right) \\ &\quad \geq \exp\left(-n\left(\Lambda^{*-}_{A^{1,2}}(a) + \varepsilon_2\right)\right). \qquad \square \end{aligned}$$

6.1. *Connection between Palm and stationary distributions in the large deviations regime.* In this subsection we show that the stationary and the Palm distribution of the same point process have the same large deviations behavior.

Consider a stationary arrival process satisfying Assumption B with the interarrivals A_i, $i \in \mathbb{Z}$. We have

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbf{P}\left[S^A_{1,n} \leq na\right] = -\Lambda^{*-}_A(a). \tag{91}$$

As explained in the proof of Theorem 6.1, $\mathbf{P}[\cdot]$ denotes the probability distribution seen by a random customer [customer 1 in the case of (91)]. Consider now a random time (say $t = 0$) and assume that customer 0 is the first customer to arrive after $t = 0$. Let U, V denote the duration and the age, respectively, of A_0. The situation is depicted in Figure 10. By $\mathbf{P}_R[\cdot]$ we denote the probability distribution seen at the random time $t = 0$ and we are interested in obtaining a LDP for $S^A_{1,n}$ under $\mathbf{P}_R[\cdot]$. The next theorem estab-

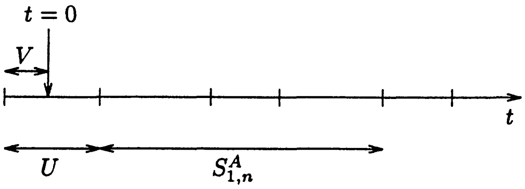

FIG. 10. *The arrival process seen at a random time.*

lishes the result. Moreover, we are also interested in obtaining a LDP result for the partial sum process $\{S^A_{1,j},\ j = 1,\dots,n\}$ under $\mathbf{P}_R[\cdot]$ when Assumption C(i) is satisfied. The latter result is obtained in Lemma 6.6.

THEOREM 6.4. *Under Assumption* B *we have*

$$\lim_{n\to\infty} \frac{1}{n}\log \mathbf{P}_R\big[S^A_{1,n} \le na\big] = -\Lambda^{*-}_A(a). \tag{92}$$

PROOF. Let $\mathbf{E}_R[\cdot]$ denote the expectation with respect to $\mathbf{P}_R[\cdot]$. We use a standard procedure to relate $\mathbf{E}_R[\cdot]$ to $\mathbf{E}[\cdot]$ (see [31]). Consider an arbitrary function $f(\cdot)$ of $S^A_{1,n}$. It can be shown ([31], Chapter 7) that

$$\mathbf{E}_R\big[f(S^A_{1,n}) \mid V = v, U = u\big] = \mathbf{E}\big[f(S^A_{1,n}) \mid A_0 = u\big].$$

Thus following the steps in [31], Chapter 7,

$$\begin{aligned}\mathbf{E}_R\big[f(S^A_{1,n})\big] &= \frac{1}{\mathbf{E}[A_1]}\int_{u=0}^{\infty}\int_{v=0}^{u} \mathbf{E}\big[f(S^A_{1,n}) \mid A_0 = u\big]\,dv\,dF_{A_0}(u)\\ &= \mathbf{E}\left[\int_{v=0}^{A_0} f(S^A_{1,n})\,dv\right]\\ &= \mathbf{E}\big[A_0 f(S^A_{1,n})\big],\end{aligned} \tag{93}$$

where we have assumed without loss of generality that $\mathbf{E}[A_1] = 1$, and we have used the notation $F_{A_0}(\cdot)$ for the distribution function of A_0.

To obtain an upper bound on $\mathbf{E}_R[\exp(\theta S^A_{1,n})]$, we set $f(\cdot) = e^{\theta\cdot}$ and use Hölder's inequality. Namely,

$$\begin{aligned}\mathbf{E}_R\big[\exp(\theta S^A_{1,n})\big] &= \mathbf{E}\big[A_0 \exp(\theta S^A_{1,n})\big]\\ &= \mathbf{E}\Big[\big(A_0^{1/p}\big)^p \exp\big((\theta/q) S^A_{1,n} q\big)\Big]\\ &\le \mathbf{E}\big[A_0^{1/p}\big]^p\,\mathbf{E}\big[\exp\big((\theta/q)S^A_{1,n}\big)\big]^q, \qquad p + q = 1,\end{aligned} \tag{94}$$

which for $\theta \le 0$ implies

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}\log\mathbf{E}_R\big[\exp(\theta S^A_{1,n})\big] &\le \limsup_{n\to\infty}\frac{p\log\mathbf{E}\big[A_0^{1/p}\big]}{n} + q\Lambda_A^-\left(\frac{\theta}{q}\right)\\ &= q\Lambda_A^-\left(\frac{\theta}{q}\right),\end{aligned}\tag{95}$$

since the first term of the right-hand side vanishes. Taking the limit now as $q \to 1$ in the above equation, we obtain for $\theta \le 0$,

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbf{E}_R\big[\exp(\theta S^A_{1,n})\big] \le \Lambda_A^-(\theta).\tag{96}$$

Therefore, using (96) and the Markov inequality, we obtain

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbf{P}_R\big[S^A_{1,n} \le na\big] \le -\Lambda_A^{*-}(a).\tag{97}$$

To obtain now a lower bound on $\mathbf{P}_R[S^A_{1,n} \le na]$, set $f(S^A_{1,n}) = \mathbf{1}\{S^A_{1,n} \le na\}$ in (93), where $\mathbf{1}\{\cdot\}$ denotes the indicator function. We have

$$\begin{aligned}\mathbf{P}_R\big[S^A_{1,n} \le na\big] &= \int_0^\infty u\mathbf{P}\big[S^A_{1,n} \le na \mid A_0 = u\big]\,dF_{A_0}(u)\\ &\ge \frac{1}{n^2}\int_{1/n^2}^\infty \mathbf{P}\big[S^A_{1,n} \le na \mid A_0 = u\big]\,dF_{A_0}(u)\\ &= \frac{1}{n^2}\mathbf{P}\left[S^A_{1,n} \le na,\ A_0 \ge \frac{1}{n^2}\right].\end{aligned}\tag{98}$$

We need the following lemma, the proof of which is deferred until the end of the current proof.

Lemma 6.5. *Under Assumption* B *and for every positive ε and a, there exists $N_{a,\varepsilon}$ such that for every $n \ge N_{a,\varepsilon}$ we have*

$$\mathbf{P}\left[S^A_{1,n} \le na,\ A_0 \ge \frac{1}{n^2}\right] \ge \exp(-n(\Lambda_A^{*-}(a) + \varepsilon)).\tag{99}$$

We now use Lemma 6.5 in (98) and take $\varepsilon \to 0$ to obtain

$$\liminf_{n\to\infty}\frac{1}{n}\log\mathbf{P}_R\big[S^A_{1,n} \le na\big] \ge -\Lambda_A^{*-}(a). \qquad \square$$

Proof of Lemma 6.5. Equation (91) implies that for every positive ε' and a there exists $N'_{a,\varepsilon'}$ such that for every $n \ge N'_{a,\varepsilon'}$,

$$\begin{aligned}\exp(-n(\Lambda_A^{*-}(a) + \varepsilon')) &\le \mathbf{P}\big[S^A_{1,n} \le na\big]\\ &\le \exp(-n(\Lambda_A^{*-}(a) - \varepsilon')).\end{aligned}\tag{100}$$

Fix now $a, \varepsilon' > 0$, and let $\delta = \varepsilon'$. We have

$$\begin{aligned}
&\mathbf{P}\left[S^A_{1,n} \le na,\ A_0 \ge \frac{1}{n^2}\right] \\
&\quad = \frac{1}{n\delta}\sum_{i=1}^{n\delta}\mathbf{P}\left[S^A_{i+1,i+n} \le na,\ A_i \ge \frac{1}{n^2}\right] \quad \text{(by stationarity)} \\
&\quad \ge \frac{1}{n\delta}\mathbf{P}\left[\exists\ i \in [1, n\delta] \text{ s.t. } S^A_{i+1,i+n} \le na,\ A_i \ge \frac{1}{n^2}\right] \quad \text{(union bound)} \\
&\quad \ge \frac{1}{n\delta}\mathbf{P}\left[S^A_{1,n(1+\delta)} \le na,\ \exists\ i \in [1, n\delta] \text{ s.t. } A_i \ge \frac{1}{n^2}\right] \qquad (101) \\
&\quad \ge \frac{1}{n\delta}\mathbf{P}\left[S^A_{1,n(1+\delta)} \le na,\ \sum_{i=1}^{n\delta} A_i \ge \frac{n\delta}{n^2}\right] \\
&\quad \ge \frac{1}{n\delta}\mathbf{P}\left[S^A_{1,n(1+\delta)} \le na\right] - \frac{1}{n\delta}\mathbf{P}\left[S^A_{1,n\delta} \le \frac{\delta}{n}\right] \\
&\qquad\qquad \text{(because } \mathbf{P}[A \cap B] \ge \mathbf{P}[A] - \mathbf{P}[B^C]\text{)} \\
&\quad \ge \frac{1}{n\delta}\exp\left(-n(1+\delta)\left(\Lambda^{*-}_A\left(\frac{a}{1+\delta}\right) + \varepsilon'\right)\right) - \frac{1}{n\delta}\mathbf{P}\left[S^A_{1,n\delta} \le \frac{\delta}{n}\right],
\end{aligned}$$

where the last inequality holds for all $n \ge N'_{a/(1+\delta),\varepsilon'}$. Note that we have used the notation B^C to denote the complement of B. We next show that for $n \to \infty$ (keeping a, δ, ε' fixed) we can neglect the second term in the right-hand side of (101). To see that, note that for all β positive there exists $N_{\beta,\varepsilon'}$ such that for all $n \ge N_{\beta,\varepsilon'}$, it holds

$$\mathbf{P}\left[S^A_{1,n\delta} \le \frac{\delta}{n}\right] \le \mathbf{P}\left[S^A_{1,n\delta} \le n\delta\beta\right] \le \exp(-n(\Lambda^{*-}_A(\beta) - \varepsilon')). \tag{102}$$

By taking β, δ and ε' small enough and $n \ge N_{\beta,\varepsilon'}$, we can achieve

$$\Lambda^{*-}_A(\beta) - \varepsilon' > (1+\delta)\left(\Lambda^{*-}_A\left(\frac{a}{1+\delta}\right) + \varepsilon'\right). \tag{103}$$

Here we are using the fact that for sufficiently small β,

$$\Lambda^{*-}_A(\beta) > \Lambda^{*-}_A(a/(1+\delta))$$

since $\Lambda^{*-}_A(\beta)$ is monotonically increasing as $\beta \downarrow 0$.

Observe now that the value of β which satisfies (103) is a function of a, δ and ε'. Therefore, using (102), there exists $N_{a,\delta,\varepsilon'}$ such that for all $n \ge N_{a,\delta,\varepsilon'}$ we have

$$-\frac{1}{n\delta}\mathbf{P}\left[S_{1,n\delta} \le \frac{\delta}{n}\right] \ge -\frac{1}{2n\delta}\exp\left(-n(1+\delta)\left(\Lambda^{*-}_A\left(\frac{a}{1+\delta}\right) + \varepsilon'\right)\right). \tag{104}$$

Combining (104) and (101), we conclude that there exists $\hat{N}_{a,\delta,\varepsilon'}$ such that for all $n \geq \hat{N}_{a,\delta,\varepsilon'}$, it holds

$$\mathbf{P}\left[S_{1,n}^A \leq na,\, A_0 \geq \frac{1}{n^2}\right] \geq \frac{1}{2n\delta}\exp\left(-n(1+\delta)\left(\Lambda_A^{*-}\left(\frac{a}{1+\delta}\right)+\varepsilon'\right)\right). \tag{105}$$

We now choose ε' such that (recall $\delta = \varepsilon'$)

$$\frac{1}{2n\varepsilon'}\exp\left(-n(1+\varepsilon')\left(\Lambda_A^{*-}\left(\frac{a}{1+\varepsilon'}\right)+\varepsilon'\right)\right) \geq \exp\left(-n\left(\Lambda_A^{*-}(a)+\varepsilon\right)\right),$$

for all $n \geq N_{a,\varepsilon}$. This can be done due to the lower semicontinuity of $\Lambda_A^{*-}(\cdot)$ (see the argument in proof of Lemma 2.2.5 in [15]). □

LEMMA 6.6. *Under Assumption* B *and* C(i), *we have that for every* $\varepsilon_1, \varepsilon_2, a > 0$ *there exists* $N_{a,\varepsilon_1,\varepsilon_2}$ *such that for all* $n \geq N_{a,\varepsilon_1,\varepsilon_2}$,

$$\mathbf{P}_R\left[S_{1,j}^A \leq ja + \varepsilon_1 n,\, j = 1,\dots,n\right] \geq \exp\left(-n\left(\Lambda_A^{*-}(a)+\varepsilon_2\right)\right). \tag{106}$$

PROOF. Following the proof of the lower bound in Theorem 6.4 [using the argument used to derive (98) but applied to the sample path $S_{1,j}^A \leq ja + \varepsilon_1 n$, $j = 1,\dots,n$], we have

$$\begin{aligned}&\mathbf{P}_R\left[S_{1,j}^A \leq ja + \varepsilon_1 n,\, j = 1,\dots,n\right]\\ &\quad\geq \frac{1}{n^2}\mathbf{P}\left[S_{1,j}^A \leq ja + \varepsilon_1 n,\, j = 1,\dots,n,\, A_0 \geq \frac{1}{n^2}\right].\end{aligned} \tag{107}$$

Now, as in the proof of Lemma 6.5, fixing $a, \varepsilon_1, \varepsilon_2 > 0$, we obtain

$$\begin{aligned}&\mathbf{P}\left[S_{1,j}^A \leq ja + \varepsilon_1 n,\, j = 1,\dots,n,\, A_0 \geq \frac{1}{n^2}\right]\\ &\quad= \frac{1}{n\delta}\sum_{k=1}^{n\delta}\mathbf{P}\left[S_{1+k,j+k}^A \leq ja + \varepsilon_1 n,\, j = 1,\dots,n,\, A_k \geq \frac{1}{n^2}\right]\\ &\quad\geq \frac{1}{n\delta}\mathbf{P}\Bigg[\exists\, k \in [1, n\delta] \text{ s.t. } S_{1+k,j+k}^A \leq ja + \varepsilon_1 n,\\ &\qquad\qquad\qquad j = 1,\dots,n,\, A_k \geq \frac{1}{n^2}\Bigg]\\ &\quad\geq \frac{1}{n\delta}\mathbf{P}\Bigg[\forall\, k \in [1, n\delta]\; S_{1+k,j+k}^A \leq ja + \varepsilon_1 n,\\ &\qquad\qquad j = 1,\dots,n, \exists\, k \in [1, n\delta] \text{ s.t. } A_k \geq \frac{1}{n^2}\Bigg]\\ &\quad\geq \frac{1}{n\delta}\mathbf{P}\left[\forall\, k \in [1, n\delta]\; S_{1+k,j+k}^A \leq ja + \varepsilon_1 n,\, j = 1,\dots,n\right]\\ &\qquad - \frac{1}{n\delta}\mathbf{P}\left[S_{1,n\delta}^A \leq \frac{\delta}{n}\right].\end{aligned} \tag{108}$$

Now notice that

$$\begin{aligned}
&\mathbf{P}\big[\forall\, k \in [1, n\delta]\ S^A_{1+k, j+k} \le ja + \varepsilon_1 n, j = 1, \dots, n\big]\\
&\quad = \mathbf{P}\big[\forall\, k \in [1, n\delta]\ S^A_{1, j+k} - S^A_{1,k} \le (j+k)a - ka + \varepsilon_1 n,\\
&\qquad\qquad\qquad\qquad\qquad\qquad j = 1, \dots, n\big]\\
&\quad \ge \mathbf{P}\bigg[S^A_{1, j+k} \le (j+k)a + \frac{\varepsilon_1 n}{2}, \forall\, k \in [1, n\delta], j = 1, \dots, n,\\
&\qquad\qquad\qquad S^A_{1,k} \ge ka - \frac{\varepsilon_1 n}{2}, \forall\, k \in [1, n\delta]\bigg] \qquad (109)\\
&\quad = \mathbf{P}\bigg[S^A_{1, j+k} \le (j+k)a + \frac{\varepsilon_1 n}{2}, \forall\, k \in [1, n\delta], j = 1, \dots, n\bigg]\\
&\quad \ge \exp\big(-n(1+\delta)(\Lambda^{*-}_A(a) + \varepsilon')\big),
\end{aligned}$$

where the last equality is obtained by choosing sufficiently small δ such that $n\delta a - \varepsilon_1 n/2 < 0$ which implies that $\mathbf{P}[S^A_{1,k} \ge ka - \varepsilon_1 n/2, \forall\, k \in [1, n\delta]] = 1$. The last inequality holds, due to Assumption C(i), for all $n \ge N'_{a, \varepsilon_1, \varepsilon'}$. Now, as in Lemma 6.5, it can be shown that there exists $N''_{a, \delta, \varepsilon'}$ such that for all $n \ge N''_{a, \delta, \varepsilon'}$, it holds

$$(110)\qquad -\frac{1}{n\delta}\mathbf{P}\bigg[S_{1, n\delta} \le \frac{\delta}{n}\bigg] \ge -\frac{1}{2n\delta}\exp\big(-n(1+\delta)(\Lambda^{*-}_A(a) + \varepsilon')\big).$$

Combining (107), (108), (109) and (110), we conclude that there exists $\hat{N}_{a, \varepsilon_1, \delta, \varepsilon'}$ such that for all $n \ge \hat{N}_{a, \varepsilon_1, \delta, \varepsilon'}$,

$$(111)\qquad \begin{aligned}
&\mathbf{P}_R\big[S^A_{1,j} \le ja + \varepsilon_1 n, j = 1, \dots, n\big]\\
&\quad \ge \frac{1}{2n^3\delta}\exp\big(-n(1+\delta)(\Lambda^{*-}_A(a) + \varepsilon')\big).
\end{aligned}$$

We now choose ε' and if necessary δ smaller than the one chosen above for the purposes of (109), such that

$$\frac{1}{2n^3\delta}\exp\big(-n(1+\delta)(\Lambda^{*-}_A(a) + \varepsilon')\big) \ge \exp\big(-n(\Lambda^{*-}_A(a) + \varepsilon_2)\big),$$

for $n \ge N_{a, \varepsilon_1, \varepsilon_2}$. □

7. Deterministic splitting of a stream. In this section we treat the splitting operation of our network model. In particular, we derive a LDP for the process resulting from the splitting of a stream to a number of streams and we show that splitting preserves Assumptions B and C(i).

Consider a stream with stationary interarrival times A_i, $i \in \mathbb{Z}$, which is split to two substreams. In particular, a fraction p of arrivals of the "*master*" stream is directed to substream 1 and a fraction $1 - p$ to substream 2. Theorem 7.1 provides a LDP for stream 1. Since stream 1 is chosen arbitrarily, by relabeling the streams one can obtain a LDP for stream 2. The more general case in which the master stream is split to more than two substreams can be handled by successive splitting to two substreams. Let us denote by A_i^1, A_i^2, $i \in \mathbb{Z}$, the interarrival times of substreams 1, 2, respectively. Here, $\Lambda_A^*(\cdot)$ and $\Lambda_A(\cdot)$ denote the large deviations rate function and the limiting log-moment generating function of the master stream.

THEOREM 7.1. *Under Assumption* B, *the partial sum* $S_{1,n}^{A^1}$ *of substream* 1 *satisfies*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left[S_{1,n}^{A^1} \le na\right] = -\frac{1}{p}\Lambda_A^{*-}(ap). \tag{112}$$

PROOF. To have n arrivals in substream 1 we need n/p arrivals of the master stream. Since we are interested in large values of n we will ignore integrality issues (i.e., we have $\lfloor n/p \rfloor / n \to 1/p$, as $n \to \infty$). Thus,

$$\mathbf{P}\left[S_{1,n}^{A^1} \le na\right] = \mathbf{P}\left[S_{1,n/p}^{A} \le na\right] \le \exp(-(n/p)(\Lambda_A^{*-}(ap) - \varepsilon)).$$

Similarly for the lower bound. □

We now argue that splitting preserves Assumptions B and C(i). It is clear that the process resulting from splitting satisfies Assumption B, since we have proved a one-sided LDP for this process with large deviations rate function expressed as a function of the large deviations rate function of the master process. Moreover, the process A^1 has moments of all orders since A has, and $(1/p)\Lambda_A^{*-}(ap)$ is strictly convex in $(-\infty, \mathbf{E}[A_i]/p)$ since $\Lambda_A^{*-}(a)$ is in the interval $(-\infty, \mathbf{E}[A_i])$. The next theorem establishes that the process resulting from splitting satisfies Assumption C(i).

THEOREM 7.2. *Assume that the process* $\{A_i,\ i \in \mathbb{Z}\}$, *satisfies Assumption* B *and* C(i). *Then the* A^1 *process satisfies Assumption* C(i).

PROOF. The proof is very similar to the proof of Theorem 7.1:

$$\begin{aligned}\mathbf{P}\left[S_{1,j}^{A^1} - ja \le \varepsilon_1 n, j = 1,\dots,n\right] &\ge \mathbf{P}\left[S_{1,j/p}^{A} - ja \le \varepsilon_1 n, j = 1,\dots,n\right]\\ &\ge \exp(-(n/p)(\Lambda_A^{*-}(ap) + \varepsilon)),\end{aligned}$$

for n large enough and all ε_1, $\varepsilon > 0$ by using Assumption C(i) for the master process. □

8. An example: queues in tandem.

In this section we apply the results derived so far to obtain LDP's for two $G/GI/1$ queues in tandem. Moreover, we work out a numerical example in order to get a qualitative understanding of the results. Large deviations results for tandem queues with renewal arrivals and exponential servers have been reported in [21].

Consider two $G/GI/1$ queues in tandem. Let A_i, $i \in \mathbb{Z}$, denote the interarrival times in the first queue and B_i^1, B_i^2, $i \in \mathbb{Z}$, the service times in the first and second queue, respectively. These processes are mutually independent, stationary and satisfy Assumptions B and C.

According to Corollary 5.5, the limiting log-moment generating function of the departure process from the first queue is given by

$$\Lambda_D^-(\theta) = \begin{cases} \inf_{x+y=\theta}\{\Lambda_{B^1}^-(x) + \Lambda_A^-(y)\}, & \text{if } \theta \geq \hat{\theta}, \\ \Lambda_{B^1}^-(\theta - \theta^*) + \Lambda_A(\theta^*), & \text{if } \theta < \hat{\theta}, \end{cases} \tag{113}$$

where

$$\hat{\theta} \triangleq \frac{d}{da}\left[\Lambda_{B^1}^{*-}(a) + \Lambda_A^{*-}(a)\right]_{a=\Lambda'_A(\theta_1^*)}.$$

Applying Theorem 4.1, we obtain that the tail probability of the stationary waiting time, W_2, seen by a customer in the second queue, is characterized by

$$\mathbf{P}[W_2 \geq U] \sim \exp(\theta_2^* U), \tag{114}$$

where U is large enough and $\theta_2^* < 0$ is the smallest root of the equation $\Lambda_D(\theta) + \Lambda_{B^2}(-\theta) = 0$. Since for $\theta \leq 0$ the equation $\Lambda_D(\theta) + \Lambda_{B^2}(-\theta) = 0$ has exactly the same roots as the equation $\Lambda_D^-(\theta) + \Lambda_{B^2}^+(-\theta) = 0$, it turns out that θ_2^* is the smallest root of the equation

$$\inf_{x+y=\theta}\{\Lambda_{B^1}^-(x) + \Lambda_A^-(y)\} + \Lambda_{B^2}^+(-\theta) = 0 \quad \text{if } \theta \geq \hat{\theta},$$

$$\Lambda_{B^1}^-(\theta - \theta^*) + \Lambda_A(\theta^*) + \Lambda_{B^2}^+(-\theta) = 0 \quad \text{if } \theta < \hat{\theta}.$$

It is instructive to characterize a most likely path along which the LDP for the waiting time occurs in the second queue. The remarks after the proof of Theorem 4.1, suggest that a most likely path for the waiting time in the second queue is characterized by

$$\begin{aligned} &\mathbf{P}\left[W_0^2 \geq (i+1)a\right] \\ &\quad \sim \sup_{x_2 - x_1 = a} \mathbf{P}\left[S_{-i,0}^D \leq (i+1)x_1\right]\mathbf{P}\left[S_{-i-1,-1}^{B^2} \geq (i+1)x_2\right], \end{aligned} \tag{115}$$

where W_0^2 denotes the waiting time of the 0th customer in the second queue and i is large enough. Setting $U = (i+1)a$, we obtain for large enough U,

$$\mathbf{P}\left[W_0^2 \geq U\right] \sim \exp\left\{-U \inf_{a>0} \frac{1}{a} \inf_{x_2 - x_1 = a}\left[\Lambda_D^{*-}(x_1) + \Lambda_{B^2}^{*+}(x_2)\right]\right\}. \tag{116}$$

Let (a^*, x_1^*, x_2^*) be an optimal solution of the optimization problem appearing in (116). Equation (116) suggests that the waiting time in the second queue builds up by maintaining an empirical rate of $1/x_1^*$ for the process D (departure from first queue) and an empirical service rate (process B^2) of $1/x_2^*$.

We use the remarks after Theorem 5.4 to characterize a most likely path for the process D to maintain an empirical rate of $1/x_1^*$. Let i^* be defined by the equation $i^* + 1 = U/a^*$. From (115), it can be seen that it suffices to characterize a most likely path along which the event $\{S^D_{-i^*,0} \leq (i^* + 1)x_1^*\}$ occurs. As shown in Theorem 5.4, this most likely path is characterized by

$$\begin{aligned}&\mathbf{P}\big[S^D_{-i^*,0} \leq (i^* + 1)x_1^*\big]\\ (117)\qquad &\sim \exp\left\{(i^*+1)\sup_{\zeta\geq 0}\ \sup_{y_1-y_2=a}\left[-(1+\zeta)\Lambda_A^{*-}\left(\frac{y_1}{1+\zeta}\right) - \zeta\Lambda_{B^1}^{*+}\left(\frac{y_2}{\zeta}\right)\right]\right.\\ &\qquad\qquad\left. -(i^*+1)\Lambda_{B^1}^{*-}(a)\right\}.\end{aligned}$$

Let (y_1^*, y_2^*, ζ^*) be the solution of the optimization problem appearing in (117). We depict a most likely path in Figure 11.

We now proceed with a numerical example. We choose the arrival process A to be a two-state *Markov modulated* deterministic process. More precisely, we consider a two-state Markov chain with transition probability matrix

$$(118)\qquad P = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix},$$

and we let the interarrival times be equal to $1/\lambda_1 = 1/5$ w.p.1 when the chain is at state 1, and equal to $1/\lambda_2 = 1/10$ w.p.1 when the chain is at state 2. The steady-state probability vector for this Markov chain is $[\pi_1\ \pi_2] =$

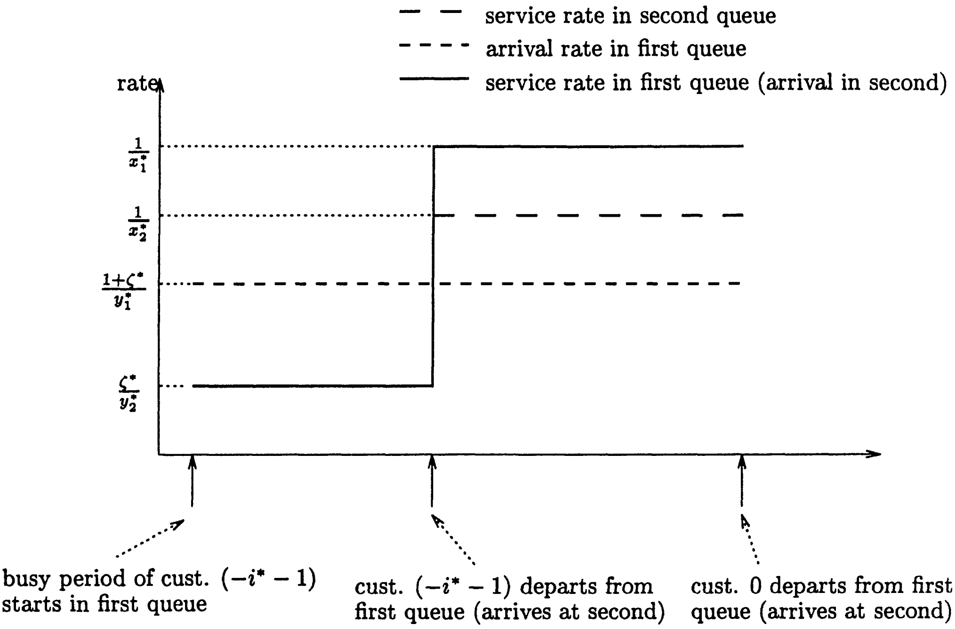

FIG. 11. *A most likely path for the waiting time in the second queue.*

[0.6 0.4] and thus the mean interarrival is $(1/\lambda_1)\pi_1 + (1/\lambda_2)\pi_2 = 0.16$. We chose a deterministic server for both queues 1 and 2 with service times $c = 0.13$.

Theorem 3.1.2 in [15] calculates the limiting log-moment generating function for the arrival process as the largest eigenvalue of the matrix $P_\theta \triangleq [p_{ij}e^{\theta/\lambda_j}]$, which in our case is

$$P = \begin{bmatrix} 0.8\,e^{\theta/5} & 0.2\,e^{\theta/10} \\ 0.3\,e^{\theta/5} & 0.7\,e^{\theta/10} \end{bmatrix}. \tag{119}$$

We performed several calculations using the software package *Matlab*. For the tail probability of the waiting time in the first queue we found that $\theta_1^* = -9.47$. We calculated the large deviations rate functions $\Lambda_A^{*-}(a)$ and $\Lambda_D^{*-}(a)$ for the arrival process and the departure process from the first queue, respectively. The results appear in Figure 12. To calculate $\Lambda_D^{*-}(a)$ we used (72). It can be seen that the first queue has a smoothing effect on the arrival process. In other words, the departure process deviates from its mean with smaller probability than the arrival process does. We also found that $\Lambda_D(\theta) + \Lambda_{B^2}(-\theta)$ is strictly negative for all $\theta < 0$, so that, as can be seen from the proof of Theorem 4.1, we have $\theta_2^* = -\infty$, which means that a large

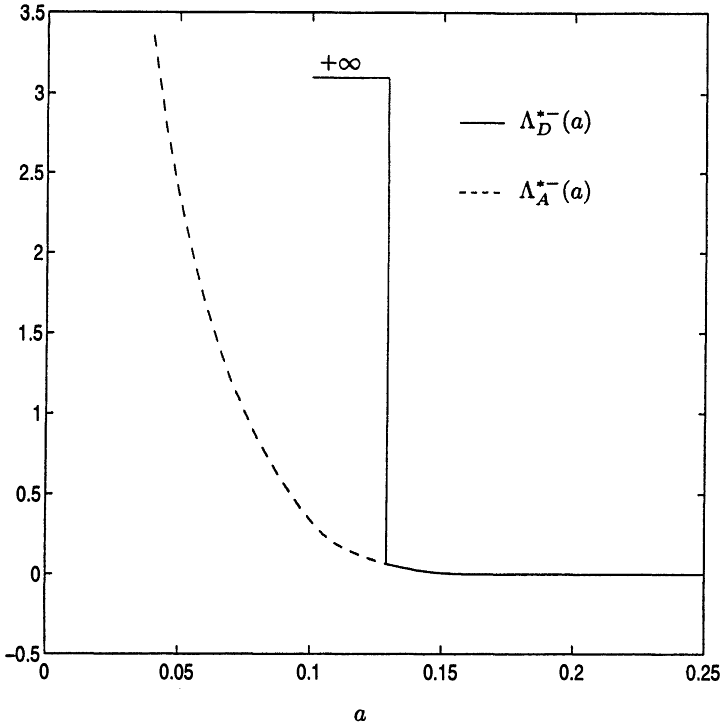

FIG. 12. $\Lambda_A^{*-}(a)$ *and* $\Lambda_D^{*-}(a)$ *for the numerical example.*

queue does not build up in the second queue. Finally, we found that the departure process D_2 from the second queue has large deviations rate function $\Lambda^{*}_{D_2}(a)$ equal to $\Lambda^{*-}_{D}(a)$. This can also be seen analytically. Namely, observe that in (71) we have $\Lambda^{*-}_{\Gamma}(a) = \Lambda^{*-}_{D}(a)$ which implies $\Lambda^{*-}_{D_2}(a) = \Lambda^{*-}_{D}(a)$.

9. Conclusions and open problems. We have considered a single class, acyclic network of $G/G/1$ queues and characterized the large deviations behavior of the waiting time and the queue length in all the queues of the network. We accomplished that by obtaining the large deviations behavior of all the processes resulting from various operations in the network, which for the network model that we considered were passage-from-a-queue, superposition of independent processes and deterministic splitting of a process to a number of processes. We concretely characterized the way that these large deviations occur.

These results are to the best of our knowledge among the few that study large deviations in a network. It is clear that more work is needed in this area, especially in view of the important applications in high speed communication networks. It is an interesting open problem to derive similar results for network models that have feedback and accommodate more than one type of traffic. It would also be interesting to study, in the large deviations regime, how different types of traffic interact and how to choose scheduling policies in order to satisfy certain performance criteria. Work relevant to the latter problem for the single queue case is reported in [30] and [17].

APPENDIX

Here we consider an arbitrary process $\{X_i,\ i \in \mathbb{Z}\}$ that satisfies Assumption B and the following: for every $\varepsilon_1, \varepsilon_2, \delta, a > 0$, there exists M_X such that for all $n \geq M_X$,

$$\begin{aligned}\exp\bigl(-n(\Lambda^{*-}_{X}(a) + \varepsilon_2)\bigr) \leq \mathbf{P}\bigl[S^{X}_{i,j} - (j-i+1)a \leq \varepsilon_1 n,\\ 1 \leq i \leq j \leq n \text{ s.t. } (j-i+1) > \delta n\bigr].\end{aligned} \tag{120}$$

Inequality (120) is implied by the results in [14], under some mild mixing assumptions on the process $\{X_i,\ i \in \mathbb{Z}\}$. We prove that the process $\{X_i,\ i \in \mathbb{Z}\}$ satisfies Assumption C for the service times [see (19)], that is, for every $\varepsilon_1, \varepsilon_2, a > 0$, there exists M'_X such that for all $n \geq M'_X$,

$$\begin{aligned}&\exp\bigl(-n(\Lambda^{*-}_{X}(a) + \varepsilon_2)\bigr)\\ &\quad\leq \mathbf{P}\bigl[S^{X}_{i,j} - (j-i+1)a \leq \varepsilon_1 n, 1 \leq i \leq j \leq n\bigr].\end{aligned} \tag{121}$$

Since Assumption C for the arrivals in (18) is a weaker version of the above, it is also satisfied by the process $\{X_i,\ i \in \mathbb{Z}\}$.

Fix positive ε_1, ε_2 and a. We have

$$\begin{aligned}
&\mathbf{P}\big[S^X_{i,j} - (j-i+1)a \le \varepsilon_1 n, 1 \le i \le j \le n\big]\\
&\quad = \mathbf{P}\big[S^X_{i,j} - (j-i+1)a \le \varepsilon_1 n, 1 \le i \le j \le n \text{ s.t. } (j-i+1) > \delta n,\\
&\qquad\quad S^X_{i,j} - (j-i+1)a \le \varepsilon_1 n, 1 \le i \le j \le n \text{ s.t. } (j-i+1) \le \delta n\big]\\
&\quad \ge \mathbf{P}\big[S^X_{i,j} - (j-i+1)a \le \varepsilon_1 n, 1 \le i \le j \le n \text{ s.t. } (j-i+1) > \delta n\big]\\
&\qquad - \mathbf{P}\big[\exists\, i \le j \in [1,n] \text{ s.t. } (j-i+1) \le \delta n\\
&\qquad\qquad \text{and } S^X_{i,j} - (j-i+1)a \ge \varepsilon_1 n\big],
\end{aligned} \tag{122}$$

where we have used the inequality $\mathbf{P}[A \cap B] \ge \mathbf{P}[A] - \mathbf{P}[B^C]$. Using the union bound and the Gärtner–Ellis theorem, we obtain that for all $\varepsilon_3 > 0$ there exists N_1 such that for all $n \ge N_1$,

$$\begin{aligned}
&\mathbf{P}\big[\exists\, i \le j \in [1,n] \text{ s.t. } (j-i+1) \le \delta n\\
&\qquad \text{and } S^X_{i,j} - (j-i+1)a \ge \varepsilon_1 n\big]\\
&\quad \le \sum_{\substack{i \le j \in [1,n]\\ (j-i+1) \le \delta n}} \mathbf{P}\big[S^X_{i,j} - (j-i+1)a \ge \varepsilon_1 n\big]\\
&\quad \le \sum_{\substack{i \le j \in [1,n]\\ (j-i+1) \le \delta n}} \mathbf{P}\big[S^X_{1,\delta n} \ge \varepsilon_1 n\big]\\
&\quad \le n^2 \exp\left(-n\delta\left(\Lambda^{*+}_X\left(\frac{\varepsilon_1}{\delta}\right) - \varepsilon_3\right)\right).
\end{aligned} \tag{123}$$

Now for given $\varepsilon_2' > 0$, choose ε_3 and δ small enough in order for large n to have

$$n^2 \exp\left(-n\delta\left(\Lambda^{*+}_X\left(\frac{\varepsilon_1}{\delta}\right) - \varepsilon_3\right)\right) \le \frac{1}{2}\exp(-n(\Lambda^{*-}_X(a) + \varepsilon_2')). \tag{124}$$

This can be done since $\Lambda^{*+}_X(\beta) \to \infty$ as $\beta \to \infty$.

Also, by using (120), we have that there exists N'' such that for all $n \ge N''$,

$$\begin{aligned}
&\mathbf{P}\big[S^X_{i,j} - (j-i+1)a \le \varepsilon_1 n, 1 \le i \le j \le n \text{ s.t. } (j-i+1) > \delta n\big]\\
&\qquad \ge \exp(-n(\Lambda^{*-}_X(a) + \varepsilon_2')).
\end{aligned} \tag{125}$$

Combining (125), (124) and (123) with (122), we obtain that there exists $\hat{N}$ such that for all $n \ge \hat{N}$,

$$\begin{aligned}
&\mathbf{P}\big[S^X_{i,j} - (j-i+1)a \le \varepsilon_1 n, 1 \le i \le j \le n\big]\\
&\qquad \ge \tfrac{1}{2}\exp(-n(\Lambda^{*-}_X(a) + \varepsilon_2')).
\end{aligned} \tag{126}$$

Finally, to obtain (121) it suffices to choose ε_2' such that for large enough n,

$$\tfrac{1}{2}\exp(-n(\Lambda_X^{*-}(a) + \varepsilon_2')) \geq \exp(-n(\Lambda_X^{*-}(a) + \varepsilon_2)). \qquad \square$$

Acknowledgments. The authors thank G. de Veciana, C. Courcoubetis, J. Walrand and T. Zajic for making available preprints of their work and D. Tse and O. Zeitouni for some helpful discussions. We also thank three anonymous reviewers for their careful reading of the paper and their very detailed and useful comments.

REFERENCES

[1] Anantharam, V. (1988). How large delays build up in a $GI/G/1$ queue. *Queueing Systems* **5** 345–368.

[2] Asmussen, S. (1982). Conditioned limit theorems relating a random walk to its associate, with applications to risk reserve processes and the $GI/G/1$ queue. *Adv. in Appl. Probab.* **14** 143–170.

[3] Bertsimas, D. and Mourtzinou, G. (1996). A unified method to analyze overtake free queueing systems. *Adv. in Appl. Probab.* **28** 586–625.

[4] Bertsimas, D. and Nakazato, D. (1990). The departure process from a $GI/G/1$ queue and its applications to the analysis of tandem queues. Working paper OR 245-91, Operations Research Center, MIT.

[5] Bertsimas, D. and Nakazato, D. (1995). The general distributional Little's law and its applications. *Oper. Res.* **43** 298–310.

[6] Bucklew, J. A. (1990). *Large Deviation Techniques in Decision, Simulation, and Estimation.* Wiley, New York.

[7] Chang, C. S. (1994). Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. Automat. Control* **39** 913–931.

[8] Chang, C. S. (1995). Sample path large deviations and intree networks. *Queueing Systems* **20** 7–36.

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

[10] Chang, C. S. and Zajic, T. (1995). Effective bandwidths of departure process from queues with time varying capacities. In *Proceedings IEEE Infocom '95* **3** 1001–1009. IEEE, New York.

[11] Courcoubetis, C. and Weber, R. (1995). Effective bandwidths for stationary sources. *Probab. Engrg. Inform. Sci.* **9** 285–296.

[12] Cruz, R. L. (1991). A calculus for network delay, I: network elements in isolation. *IEEE Trans. Inform. Theory* **37** 114–131.

[13] Cruz, R. L. (1991). A calculus for network delay, II: network analysis. *IEEE Trans. Inform. Theory* **37** 132–141.

[14] Dembo, A. and Zajic, T. (1995). Large deviations: from empirical mean and measure to partial sums processes. *Stochastic Process Appl.* **57** 191–224.

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

[16] de Veciana, G., Courcoubetis, C. and Walrand, J. (1993). Decoupling bandwidths for networks: a decomposition approach to resource management. Memorandum, Electronics Research Lab., Univ. California, Berkeley.

[17] de Veciana, G. and Kesidis, G. (1995). Bandwidth allocation for multiple qualities of service using generalized processor sharing. *IEEE Trans. Inform. Theory* **42**

[18] de Veciana, G. and Walrand, J. (1992). Traffic shaping for ATM networks: asymptotic analysis and simulations. Preprint.

[19] DE VECIANA, G. and WALRAND, J. (1993). Effective bandwidths: call admission, traffic policing and filtering for ATM networks. Memorandum, Electronics Research Lab., Univ. California, Berkeley.

[20] ELWALID, A. I. and MITRA, D. (1993). Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Transactions on Networking* **1** 329–343.

[21] GANES, H. A. and ANANTHARAM, V. (1996). Stationary tail probabilities in exponential server tandems with renewal arrivals. *Queueing Systems Theory Appl.* **22** 203–248.

[22] GIBBENS, R. J. and HUNT, P. J. (1991). Effective bandwidths for the multi-type UAS channel. *Queueing Systems* **9** 17–28.

[23] GLYNN, P. W. and WHITT, W. (1994). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Probab.* **31A** 131–156.

[24] HUI, J. Y. (1988). Resource allocation for broadband networks. *IEEE Journal on Selected Areas in Communications* **6** 1598–1608.

[25] KELLY, F. P. (1979). *Reversibility and Stochastic Networks.* Wiley, New York.

[26] KELLY, F. P. (1991). Effective bandwidths at multi-class queues. *Queueing Systems* **9** 5–16.

[27] KESIDIS, G., WALRAND, J. and CHANG, C. S. (1993). Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Transactions on Networking* **1** 424–428.

[28] KINGMAN, J. F. C. (1970). Inequalities in the theory of queues. *J. Roy. Statist. Soc.* **32** 102–110.

[29] ROCKAFELLAR, R. T. (1970). *Convex Analysis.* Princeton Univ. Press.

[30] TSE, D. (1994). Variable-rate lossy compression and its effects on communication networks. Ph.D. dissertation, MIT.

[31] WALRAND, J. (1988). *An Introduction to Queueing Networks.* Prentice-Hall, Englewood Cliffs, NJ.

[32] WHITT, W. (1993). Tail probability with statistical multiplexing and effective bandwidths in multi-class queues. *Telecommunication Systems* **2** 71–107.

[33] YARON, O. and SIDI, M. (1993). Performance and stability of communication networks via robust exponential bounds. *IEEE/ACM Transactions on Networking* **1** 372–385.

D. BERTSIMAS
SLOAN SCHOOL OF MANAGEMENT
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02139
E-MAIL: dbertsim@aris.mit.edu

I. C. PASCHALIDIS
DEPARTMENT OF MANUFACTURING
ENGINEERING
BOSTON UNIVERSITY
15 ST. MARY'S STREET
BOSTON, MASSACHUSETTS 02215
E-MAIL: yannisp@bu.edu

J. N. TSITSIKLIS
DEPARTMENT OF EECS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02139
E-MAIL: jnt@mit.edu