Sample Path Large Deviations and Intree Networks *

Cheng-Shang Chang
Dept. of Electrical Engineering
National Tsing Hua University
Hsinchu 30043
Taiwan, R.O.C.
Email: cschang@ee.nthu.edu.tw

(August, 1993; revision: June, 1994)

Abstract

Using the contraction principle, in this paper we derive a set of closure properties for sample path large deviations. These properties include sum, reduction, composition and reflection mapping. Using these properties, we show that the exponential decay rates of the steady state queue length distributions in an intree network with routing can be derived by a set of recursive equations. The solution of this set of equations is related to the recently developed theory of effective bandwidth for high speed digital networks, especially ATM networks. We also prove a conditional limit theorem that illustrates how a queue builds up in an intree network.

Keywords: ATM networks, effective bandwidth, large deviations, conditional limit theorem

*To appear in Queueing Systems

1 Introduction

In this paper, we consider a discrete-time intree network with J queues and multiple classes of customers. In such a network, there is a natural partial ordering among the J queues. If part of the departure process of queue i is an input process of queue j, we say queue j is a *successor* of queue i and queue i is a *predecessor* of queue j. Let $Suc(j)$ (resp. $Pre(j)$) be the set of queues that are *successors* (resp. predecessors) of queue j. We label the J queues in the *postorder* (left subtree, right subtree, root). Since the network is an intree, $Suc(i)$ only contains one element. Also, we have from the postorder labelling that $i < j$ for all $i \in Pre(j)$ and $Suc(j) > j$.

We assume that routing only depends on the queue that customers depart from. To model this, we let $\{p_j(n), n \geq 1\}, j = 1\ldots, J-1$, be $J-1$ sequences of routing random variables, i.e., $p_j(n) = 1$ if the n^{th} departure from queue j is routed to the successor of queue j, i.e., queue $Suc(j)$, and $p_j(n) = 0$ if the n^{th} departure from queue j leaves the intree network. Besides the arrivals from the queues in the predecessors, there are k_j external arrival processes coming to queue j with $\{a_{j,k}(t), t = 0, 1, 2, \ldots\}$, $k = 1, \ldots, k_j$, denoting the number of arrivals at time t.

Denote $q_j(t)$ and c_j the queue length of queue j at time t and its capacity, respectively. We assume that each queue is governed by Lindley's equation, i.e.,

$$q_j(t+1) = (q_j(t) + a_j(t+1) - c_j)^+, \;\; j = 1, \ldots, J, \tag{1}$$

where $a_j(t)$ is the total number of arrivals at queue j at time t, and $(x)^+ \equiv \max(0, x)$. The motivation for considering such a discrete-time model, as explained in our previous papers (Chang [6], Chang, Heidelberger, Juneja, and Shahabuddin [8], Chang [7]), is its suitability for modelling high speed digital networks, e.g., ATM networks. Various important results can be derived through such models for network managements such as admission control, traffic shaping, tariffs and congestion prevention. In Chang [6], Chang, Heidelberger, Juneja, and Shahabuddin [8] and Chang [7], we propose the following algorithm for analyzing and simulating such networks:

Step 1. Obtain the normalized logarithmic moment generating function (NLMGF) for each arrival process and routing process (provided that their limits exist).

$$\Lambda_{A,j,k}(\theta) \equiv \lim_{t\to\infty} \frac{1}{t} \log E(e^{\theta A_{j,k}(0,t)}) \tag{2}$$

$$\Lambda_{P,j}(\theta) \equiv \lim_{t\to\infty} \frac{1}{t} \log E(e^{\theta P_j(0,t)}) \tag{3}$$

where $A_{j,k}(t_1, t_2) = \sum_{s=t_1+1}^{t_2} a_{j,k}(s)$ and $P_j(t_1, t_2) = \sum_{s=t_1+1}^{t_2} p_j(s)$.

Step 2. For $j = 1, \ldots, J$, define the following functions recursively:

$$\Lambda_j(\theta) = \sum_k \Lambda_{A,j,k}(\theta) + \sum_{k \in Pre(j)} \tilde{\Omega}_k(\theta) \tag{4}$$

$$\Omega_j(\theta) = \begin{cases} \Lambda_j(\theta) & \text{if } \theta \le \tilde{\theta}_j \\ \theta c_j - \tilde{\theta}_j c_j + \Lambda_j(\tilde{\theta}_j) & \text{otherwise} \end{cases} \tag{5}$$

$$\tilde{\Omega}_j(\theta) = \Omega_j(\Lambda_{P,j}(\theta)) \tag{6}$$

where $\tilde{\theta}_j$ is a solution of $\Lambda'_j(\theta) = c_j$.

Step 3 . Approximate the steady state distribution $q_j(\infty)$ of queue j as follows:

$$\text{Prob}(q_j(\infty) \ge x) \approx e^{-\theta_j^* x} \tag{7}$$

where θ_j^* is the solution of $\Lambda_j(\theta) = c_j\theta$.

The intuition behind this algorithm is the concept of input-output relations of queueing networks. Basically, an intree network consists of the following network operations: (i) multiplexing (or sum), (ii) input-output relation through a queue (or reflection mapping), and (iii) routing (or composition). Suppose that the NLMGF is the desired way to characterize traffic, in contrast to using the average rate in Jackson's networks and using the average rate and the asymptotic variance in Brownian networks. The first rule in (4) suggests that the NLMGF of multiplexing (independent) sources of traffic is the sum of the NLMGF of each source and one can view $\Lambda_j(\theta)$ as the NLMGF of the total arrival process at queue j. Similarly, the second rule in (5) suggests the input-output relation through a queue, and one can view $\Omega_j(\theta)$ as the NLMGF of the departure process from queue j. The third rule in (6) suggests that the NLMGF of random thinning of a source, which is the model for routing, can be derived by the composition of the NLMGF of the original process and the NLMGF of the thinning process. Thus, $\tilde{\Omega}_j(\theta)$ can be viewed as the NLMGF of the process formed by the customers that departs from queue j and are routed to its successor. In Step 3, the tail of the steady state distribution $q_j(\infty)$ can be well approximated using the NLMGF of the arrival process coming to queue j. Since c_j is the capacity of queue j, $\Lambda_j(\theta_j^*)/\theta_j^*$ is then called the effective bandwidth (or equivalent bandwidth) of the arrival process to queue j. In view of (4), $\Lambda_{A,j,k}(\theta_j^*)/\theta_j^*$ is the equivalent amount of bandwidth consumed by the arrival process $a_{j,k}(t)$, given the tail distribution of queue j is exponentially decayed with rate θ_j^*. This concept of effective bandwidth, evidently suitable for bandwidth allocation, has recently generated a tremendous amount of interest in the field, e.g. Guérin, Ahmadi and Naghshineh [18], Kelly [22], Gibbens and Hunt [15], Chang [6], Kesidis, Walrand and Chang [24], Chang, Heidelberger, Juneja and Shahabuddin [8], Elwalid and Mitra [12], Whitt [35], Glynn and Whitt [16], Chang [7], Kesidis and Walrand [23], Nagarajan, Kurose and Towsley [28], Veciana and Walrand [36], and Veciana, Courcoubetis and Walrand [37].

It is of interest to compare these three rules in (4)-(6) with the key properties used in Jackson's networks: (i) the sum of independent Poisson processes is a Poisson process, (ii) the departure process is a Poisson process (when input is a Poisson processes), and (iii) geometrical splitting of a Poisson process forms independent Poisson processes. However, there is a major difference between the setting in Jackson's network and ours, namely *independence*. In our setting, the departure process is in general not independent of the future arrivals. Moreover, random splitting does not form independent subprocesses. As a consequence of lack of independence, we are not able to use the quasi-reversibility type of argument (e.g. see Kelly [21], Walrand [32]) for the treatment of nonfeedforward networks. Thus, our result in Step 3 is only for the marginal distribution, instead of having a product form joint distribution.

To see the connection of our approximation to heavy traffic two-moment approximation (see e.g. Whitt [35], Chang [7] under certain technical conditions), expanding $\Lambda_j(\theta)$ at $\theta = 0$ yields [cf. the definition in (2)]

$$\Lambda_j(\theta) = \bar{a}_j\theta + \frac{\bar{\sigma}_j^2}{2}\theta^2 + o(\theta^2), \tag{8}$$

where

$$\bar{a}_j = \lim_{t\to\infty} \frac{EA_j(0,t)}{t} \tag{9}$$

$$\bar{\sigma}_j^2 = \lim_{t\to\infty} \frac{EA_j(0,t)^2 - (EA_j(0,t))^2}{t}. \tag{10}$$

Thus, $\theta_j^* \approx 2(c_j - \bar{a}_j)/\bar{\sigma}_j^2$ and our approximation in Step 3 recovers the heavy traffic two-moment approximation.

These rules in (4)-(6) are derived based on the upper bounds in Chang [6], and Chang, Heidelberger, Juneja and Shahabuddin [8]. The lower bounds for sum and composition are also available in Chang [6], Chang, Heidelberger, Juneja and Shahabuddin [8], and Glynn and Whitt [16], and they coincide with the upper bounds. However, there is a gap for the lower bound when considering the input-output relation through a queue. One of our objectives in this paper is to fill this gap and provide a firm theoretical proof for these rules. Our approach for the lower bound is to use the sample path large deviation, a stronger concept of large deviation than that in Chang [6], Chang, Heidelberger, Juneja and Shahabuddin [8], Glynn and Whitt [16].

Using the sample path large deviation not only provides a proof for the lower bound, but also yields a physical insight on how congestion occurs. This permits us to extend the conditional limit theorem for a single queue (see e.g., Anantharam [1] and Dembo and Zajic [10]) to intree networks. Our main sources for sample path large deviation are Dembo and Zeitouni [9] and Dembo and Zajic [10]. Also, we note that the concept of sample path large deviation is also recently used in Glynn and Whitt [17] for inverting counting processes.

To state our results precisely, we make the following assumptions on the external arrival processes and routing processes:

(A1) All the external arrival processes and routing processes are *independent* and independent of all else.

(A2) For all j, k, $\{a_{j,k}(t), t = 0, 1, 2, \ldots,\}$ is bounded, nonnegative integer-valued, and adapted to a filtration $\mathcal{F}^A_{j,k,t}$ such that for all $\theta \in I\!R$, $s, t \geq 0$, there is a function $0 \leq \Gamma_{A,j,k}(\theta) < \infty$ (independent of s and t),

$$\Lambda_{A,j,k}(\theta)t - \Gamma_{A,j,k}(\theta) \leq \log E(e^{\theta A_{j,k}(s,t+s)}|\mathcal{F}^A_{j,k,s}) \leq \Lambda_{A,j,k}(\theta)t + \Gamma_{A,j,k}(\theta), \quad a.s. \tag{11}$$

where $\Lambda_{A,j,k}(\theta)$ is defined in (2). Moreover, $\Lambda_{A,j,k}(\theta)$ is differentiable for all θ.

(A3) For all j $\{p_j(t), t = 0, 1, 2, \ldots,\}$ is adapted to a filtration $\mathcal{F}^P_{j,t}$ such that for all $\theta \in I\!R$, $s, t \geq 0$, there is a function $0 \leq \Gamma_{P,j}(\theta) < \infty$ (independent of s and t),

$$\Lambda_{P,j}(\theta)t - \Gamma_{P,j}(\theta) \leq \log E(e^{\theta P_j(s,t+s)}|\mathcal{F}^P_{j,s}) \leq \Lambda_{P,j}(\theta)t + \Gamma_{P,j}(\theta), \quad a.s. \tag{12}$$

where $\Lambda_{P,j}(\theta)$ is defined in (3). Moreover, $\Lambda_{P,j}(\theta)$ is differentiable for all θ. (Since $p_j(t)$ only takes values 0 and 1, it is nonnegative and bounded.)

(A4) $\{a_{j,k}(t)t = 0, 1, 2, \ldots\}$, $k = 1, \ldots, k_j, j = 1, \ldots J$ and $\{p_j(t), t = 0, 1, 2, \ldots,\}$, $j = 1, \ldots, J$, are *stationary* and *ergodic* processes. Moreover, $\Lambda'_j(0) < c_j$ for all $j = 1, \ldots, J$.

The conditions in (A2-3) are related to the θ-EP defined in Chang [6] and Chang, Heidelberger, Juneja and Shahabuddin [8], where $(\hat{a}(\theta), \hat{\sigma}(\theta))$ is replaced by $(\Lambda(\theta)/\theta), \Gamma(\theta)/\theta)$. In Section 2, we will show that under (A1-3), the sample path large deviation principle holds for the joint process that consists of the external arrival processes and routing processes, and the rules in (4)-(6) can be derived from the contraction principle. Examples that satisfy (A2-3), as in Chang, Heidelberger, Juneja and Shahabuddin [8], are (bounded) finite state Markov arrival processes, renewal processes (with bounded supports), and (bounded) autoregressive processes.

Verification for these examples will be shown in Section 2. The condition $\Lambda_j'(0) < c_j$ in (A4) implies the usual traffic condition, i.e., the total arrival rate at queue j is smaller than its capacity (cf. (8) and Chang [6], Lemma 3.5). This condition, together with stationarity and ergodicity ensures the network has a steady state distribution.

Let $\mu_{A,J}^{(n)}$ be the distribution of the scaled process $A_J^{(n)}(t) = \frac{1}{n}\sum_{s=1}^{\lceil nt \rceil} a_J(s)$. (Note that $\{a_J(t), t = 0, 1, 2, \ldots\}$ is the total arrival process to queue J.) Also, let $q_J^{(n)}(t) = q_J(\lceil nt \rceil)/n$ and $\mu_{Q_J}^{(n)}$ be the distribution of $(q_J^{(n)}(\cdot) | q_J^{(n)}(1) > \alpha)$. Let $D[[0,1], (I\!R, \|\cdot\|_\infty)]$ be the space of right continuous and left limit R-valued functions on the interval $[0,1]$ with the supremum norm topology, i.e.,

$$\|\phi\|_\infty = \sup_{\{0 \le t \le 1\}} |\phi(t)|. \tag{13}$$

Our first theorem illustrates how the J^{th} queue builds up.

Theorem 1.1 *Under (A1-4), if there is a unique solution $\theta_J^* > 0$ of the equation $\Lambda_J(\theta) - c_J\theta = 0$ and there is a $\delta > 0$ such that $\alpha + \delta = \Lambda'_J(\theta_0) - c_J$ for some $\theta_0 < \infty$, then $\mu_{Q_J}^{(n)}$ converges in distribution on $D[[0,1], (I\!R, \|\cdot\|_\infty)]$ to δ_{q_α}, where δ_{q_α} denotes the probability distribution concentrated on*

$$q_\alpha(t) = \begin{cases} 0, & if \ \ 0 \le t \le 1-\beta; \\ \frac{\alpha}{\beta}(t - (1-\beta)), & if \ \ 1-\beta \le t \le 1, \end{cases} \tag{14}$$

and $\beta = \min[1, \alpha/(\Lambda'_J(\theta_J^*) - c_J)]$.

Intuitively, Theorem 1.1 implies that congestion occurs when the arrival process at queue J behaves as if it was another process with rate $\Lambda'_J(\theta_J^*)$. This information could be used for congestion prevention in the network. We note this observation is also used in Chang, Heidelberger, Juneja and Shahabuddin [8] to derive the conjugate process (the process that has the minimum Kullback-Leibler distance (relative entropy rate) with respect to the original arrival process) for a fast simulation algorithm, where only important sample paths are simulated.

Our second main theorem is regarding the tail distribution, which provides a theoretical foundation for the approximation in Step 3.

Theorem 1.2 *Under the conditions in Theorem 1.1, $q_J(t)$ converges in distribution to a random variable $q_J(\infty)$ that satisfies*

$$\lim_{x \to \infty} \frac{\log Prob(q_J(\infty) \ge x)}{x} = -\theta_J^*. \tag{15}$$

The rest of the paper is organized as follows: we introduce the notion of sample path large deviation in Section 2, and derive various closure properties via the contraction principle in Section 3. These results, along with the upper bounds in Chang [6], are used for strengthening the results for a single queue in Section 4 and proving Theorem 1.1 and Theorem 1.2 in Section 5.

2 Definition and examples of sample path large deviations

First, we briefly review the definition of large deviation principles on $I\!R^d$. For $y \in R^d$, let $|y|$ denote its Euclidean norm, i.e., $|y| = \sqrt{\sum_{i=1}^d y_j^2}$.

Definition 2.1 *A R^d-valued stochastic sequence $\{Y(n), n = 0, 1, 2 \ldots\}$ satisfies the large deviation principle (LDP) with the good rate function $I(y) : I\!R^d \mapsto I\!R$, if*

(i) *all the level sets $\{y \in R^d : I(y) \leq x\}$, $x \in I\!R$, are compact,*

(ii) *for any close set $F \in I\!R^d$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log Prob(Y(n) \in F) \leq - \inf_{y\in F} I(y), \quad \textit{and} \tag{16}$$

(iii) *for any open set $G \in I\!R^d$,*

$$\liminf_{n\to\infty} \frac{1}{n} \log Prob(Y(n) \in G) \geq - \inf_{y\in G} I(y). \tag{17}$$

Similarly, one can extend this definition to an abstract space. Let $D[[0,1], (I\!R^d, \|\cdot\|_\infty)]$ (resp. $AC[[0,1], (I\!R^d, \|\cdot\|_\infty)]$) be the space of right continuous and left limit (resp. absolutely continuous) functions on the interval $[0,1]$ with the supremum norm topology, i.e.,

$$\|\phi\|_\infty = \sup_{\{0\leq t\leq 1\}} |\phi(t)|. \tag{18}$$

Definition 2.2 *A sequence of probability measures $\mu^{(n)}$ on $D[[0,1], (I\!R^d, \|\cdot\|_\infty)]$ satisfies the sample path large deviation (LDP(sp)) with the good rate function $I(\phi) : D[[0,1], (I\!R^d, \|\cdot\|_\infty)] \mapsto I\!R$, if*

(i) *all the level sets $\{\phi : I(\phi) \leq x\}$, $x \in I\!R$, are compact,*

(ii) *for any close set $F \in D[[0,1], (I\!R^d, \|\cdot\|_\infty)]$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu^{(n)}(F) \leq - \inf_{\phi\in F} I(\phi), \quad \textit{and} \tag{19}$$

(iii) *for any open set* $G \in D[[0,1],(\mathbb{R}^d, \|\cdot\|_\infty)]$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu^{(n)}(G) \geq -\inf_{\phi\in G} I(\phi). \tag{20}$$

Suppose there is a function $\Lambda(\theta) : \mathbb{R}^d \mapsto \mathbb{R}$ *that is*

(E1) *convex and differentiable for all* $\theta \in \mathbb{R}^d$,

(E2) $\Lambda(\theta) < \infty$ *for all* $\theta \in \mathbb{R}^d$,

(E3) $\Lambda(\mathbf{0}) = 0$, *where* $\mathbf{0}$ *is a d-vector with all its elements being 0.*

Let $\Lambda^* : \mathbb{R}^d \mapsto \mathbb{R}$ *be its convex conjugate, i.e.,*

$$\Lambda^*(x) = \sup_{\{\theta\in\mathbb{R}^d\}} [\theta \cdot x - \Lambda(\theta)], \tag{21}$$

where $\theta \cdot x$ *is the inner product. If the the rate function* $I(\phi)$ *has the following form*

$$I(\phi) = \begin{cases} \int_0^1 \Lambda^*(\phi'(t)dt & \text{if } \phi \in AC[[0,1],(\mathbb{R}^d, \|\cdot\|_\infty)], \phi(0) = 0 \\ \infty & \text{otherwise} \end{cases} \tag{22}$$

Then we say the measures $\mu^{(n)}$ *satisfy LDP(sp) with respect to (w.r.t.) the function* $\Lambda(\cdot)$.

It is clear that under (E1-3), $\Lambda^*(x) \geq 0$ for all $x \in \mathbb{R}^d$. Since $\Lambda(\theta)$ is differentiable, $\Lambda^*(x)$ is strictly convex and $\Lambda^*(x) = \theta \cdot x - \Lambda(\theta)$ if $x = \Lambda'(\theta)$ (see e.g. Rockafellar [30], Theorem 26.3, Avriel [3], Theorem 5.3). This implies that $\Lambda^*(\Lambda'(\mathbf{0})) = 0$ and that

$$\Lambda(\theta) = \sup_{\{x\in\mathbb{R}^d\}} [\theta \cdot x - \Lambda^*(x)]. \tag{23}$$

Thus, $\Lambda(\theta)$ and $\Lambda^*(x)$ are convex conjugates of each other. Also, it is known that if $\mu_A^{(n)}$ satisfies LDP(sp) on $D[[0,1],(\mathbb{R}^d, \|\cdot\|_\infty)]$, then it can be extended to $D[[0,T],(\mathbb{R}^d, \|\cdot\|_\infty)]$ for any fixed $T \geq 0$ (see Dembo and Zajic [10]).

Consider a sequence of *nonnegative* random variables in $\mathbb{R}^d$ ($d \in \mathbb{N}$), $\{a(t), t = 0,1,2,\ldots\}$. Let $A(t_1,t_2) = \sum_{t=t_1+1}^{t_2} a(t)$. Define a sequence of stochastic processes in the fixed interval $[0,1]$

$$A^{(n)}(t) = \frac{1}{n} A(0, \lceil nt \rceil), \quad 0 \leq t \leq 1. \tag{24}$$

Let $\mu_A^{(n)}$ be the distribution of the process $\{A^{(n)}(t), 0 \leq t \leq 1\}$. The following proposition shows that LDP(sp) is stronger than LDP. The argument can be found in Glynn and Whitt [17], pp. 8, using the contraction principle and convexity of $\Lambda^*(x)$.

Proposition 2.3 *If $\mu_A^{(n)}$ satisfies LDP(sp) w.r.t. $\Lambda(\theta)$, then $\{A(0,t)/t, t = 0, 1, 2, \dots\}$ satisfies LDP with the convex good rate function $\Lambda^*(x)$ in (21).*

It is known in Mogulskii [27] if $\{a(t), t = 0, 1, 2, \dots,\}$ is a sequence of i.i.d. random vectors, then $\mu_A^{(n)}$ satisfies the LDP(sp) w.r.t. $\Lambda(\theta) = \log E \exp(\theta \cdot a(0))$. In the following, we present a uniform bounding condition, analogous to those in Chang [6] and Chang, Heidelberger, Juneja and Shahabuddin [8], that renders sample path large deviations.

Lemma 2.4 *Suppose that $\{a(t), t = 0, 1, 2, \dots\} \in \mathbb{R}^d$ is adapted to a filtration $\mathcal{F}_t$. If for all $\gamma \in \mathbb{R}$,*

$$\sup_{t,s} \frac{1}{t} \log E \exp(\gamma \sum_{l=s+1}^{s+t} |a(l)|) < \infty, \tag{25}$$

and for all $\theta \in \mathbb{R}^d$, $s, t \geq 0$, there is a differentiable function $\Lambda(\theta) < \infty$ and a function $0 \leq \Gamma(\theta) < \infty$ independent of s and t,

$$\Lambda(\theta)t - \Gamma(\theta) \leq \log E(e^{\theta \cdot A(s,t+s)} | \mathcal{F}_s) \leq \Lambda(\theta)t + \Gamma(\theta), \quad a.s. \tag{26}$$

then $\mu_A^{(n)}$ satisfies LDP(sp) w.r.t. $\Lambda(\theta)$.

Note that (26) implies the Gärtner and Ellis limit [14, 11], i.e.,

$$\Lambda(\theta) = \lim_{t \to \infty} \frac{1}{t} \log E(e^{\theta \cdot A(0,t)}). \tag{27}$$

Thus, $\Lambda(\theta)$ is convex and $\Lambda(\mathbf{0}) = 0$. In conjunction with (E1-3), the Gärtner-Ellis limit implies that $\{A(0,t)/t, t = 1, 2, \dots\}$ satisfies the LDP with the good rate function $\Lambda^*(x) = \sup_\theta[\theta \cdot x - \Lambda(\theta)]$. This is consistent with Proposition 2.3.

We also note if $\{a(t), t = 0, 1, 2, \dots\}$ is nonnegative, then (26) implies (25).

Proof. First, we show that for any $0 = t_0 < t_1 < \dots t_m = 1$,

$$\lim_{n \to \infty} \frac{1}{n} \log E e^{nz \cdot W^{(n)}} = \sum_{j=1}^m (t_j - t_{j-1}) \Lambda(z_j), \tag{28}$$

where $z = (z_1, \dots, z_m) \in (\mathbb{R}^d)^m$ and $W^{(n)} = (W_1^{(n)}, \dots, W_m^{(n)})$ with $W_j^{(n)} = A^{(n)}(t_j) - A^{(n)}(t_{j-1})$, $j = 1, \dots, m$. Since $a(t)$ is adapted to $\mathcal{F}_t$, conditioning on $\mathcal{F}_{\lceil nt_{m-1} \rceil}$ and using (26) yields

$$E(e^{nz \cdot W^{(n)}}) = E\Big(E(e^{nz \cdot W^{(n)}}) | \mathcal{F}_{\lceil nt_{m-1} \rceil}\Big) \tag{29}$$

$$= E\Big(e^{n \sum_{j=1}^{m-1} z_j \cdot W_j^{(n)}} E(e^{z_m \cdot A(\lceil nt_{m-1} \rceil, \lceil nt_m \rceil)} | \mathcal{F}_{\lceil nt_{m-1} \rceil})\Big) \tag{30}$$

$$\leq e^{\Lambda(z_m)(\lceil nt_m \rceil - \lceil nt_{m-1} \rceil) + \Gamma(z_m)} E\Big(e^{n \sum_{j=1}^{m-1} z_j \cdot W_j^{(n)}}\Big). \tag{31}$$

Repeating the argument for $t_j, j = m-2, \ldots, 1$, yields

$$\log E(e^{nz\cdot W^{(n)}}) \le \sum_{j=1}^{m} (\lceil nt_j \rceil - \lceil nt_{j-1} \rceil)\Lambda(z_j) + \Gamma(z_j). \tag{32}$$

Thus,

$$\limsup_{n\to\infty} \frac{1}{n} \log E(e^{nz\cdot W^{(n)}}) \le \sum_{j=1}^{m} (t_j - t_{j-1})\Lambda(z_j). \tag{33}$$

The argument for the lower bound of (28) is similar. Let $\Lambda^*(y_j) : I\!R^d \to I\!R$ be the convex conjugate of $\Lambda(z_j)$, i.e., $\Lambda^*(y_j) = \sup_{z_j}[y_j \cdot z_j - \Lambda(z_j)]$. In conjunction with (25), it follows from the Gärtner Ellis theorem ([14, 11], see e.g. Dembo and Zeitouni [9]) that $W^{(n)}$ satisfies the LDP principle with the convex good rate function $\Lambda_m^*(y)$, where

$$\Lambda_m^*(y) = \sup_z [y \cdot z - \sum_{j=1}^{m} (t_j - t_{j-1})\Lambda(z_j)] \tag{34}$$

$$= \sum_{j=1}^{m} (t_j - t_{j-1})\Lambda^*(\frac{y_j}{t_j - t_{j-1}}). \tag{35}$$

The last step is letting $m \to \infty$ to achieve the LDP(sp) in (22). This is established in Dembo and Zajic [10], Theorem 2 (along with the condition in (25)). See also Dembo and Zeitouni [9], the proof of Theorem 5.1.2. ■

In the following, we show some examples considered in Chang, Heidelberger, Juneja and Shahabuddin [8] indeed satisfies LDP(sp). Our first example is the Markov arrival process (MAP).

Example 2.5 (MAP) Consider an M-state irreducible and aperiodic Markov chain $\xi(t) \in I\!R$. A process $a(t)$ is called a *Markov Arrival Process* (MAP) if $(a(t), \xi(t))$ is a Markov chain with the transition kernel, $P_{i,j}(a) = \text{Prob}(a(t+1) \le a, \xi(t+1) = j | \xi(t) = i)$. Let $\Upsilon(\theta)$ be an $M \times M$ matrix with $\Upsilon_{i,j}(\theta) = \int e^{\theta a} dP_{i,j}(a)$. Also, let $sp(\Upsilon(\theta))$ and $f(\theta) = (f_1(\theta), \ldots, f_M(\theta))$ be the spectral radius of the matrix $\Upsilon(\theta)$ and the corresponding positive right eigenvector. It is shown in [Chang, Heidelberger, Juneja and Shahabuddin [8], Example 2.5 and (9)] that for any initial state $\xi(0) = \xi_0$

$$\Lambda(\theta)t - \Gamma(\theta) \le \log E(e^{\theta A(0,t)} | \xi(0) = \xi_0) \le \Lambda(\theta)t + \Gamma(\theta), \tag{36}$$

where $\Lambda(\theta) = \log sp(\Upsilon(\theta))$ and $\Gamma(\theta) = \max_{i,j} f_i(\theta)/f_j(\theta)$. If we let $\mathcal{F}_t$, $t = 0, 1, 2, \ldots$, be the σ-algebra generated by $a(t)$ and $x(t)$, then by the Markov property

$$E(e^{\theta A(s,s+t)} | \mathcal{F}_s) = E(e^{\theta A(s,s+t)} | \xi(s)). \tag{37}$$

Thus, the condition in (26) follows from (36). It is well known that $\Lambda(\theta) = \log(sp(\Upsilon(\theta)))$ is convex and differentiable for all θ (see Iscoe, Ney and Nummelin [19] and Miller [26]). Moreover, $\Lambda(\theta)$ is finite if $a(t)$ is bounded.

As a special case of MAP, consider a Markov modulated process (MMP) $a(t) = \eta_{\xi(t)}(t)$, where $\xi(t)$ is an irreducible and aperiodic Markov chain on the states $\{1, \ldots, M\}$ with the transition matrix $\mathbf{r}$, i.e., $r_{i,j}$ is the transition probability from state i to state j, and $\{\eta_i(t), t = 0, 1, \ldots\}$, $i = 1, \ldots, M$, are M sequences of i.i.d. random variables with the moment generating functions $\alpha_i(\theta) = E\exp(\theta\eta_i(0))$. Clearly, $\text{Prob}(a(t+1) \leq a, \xi(t+1) = j | \xi(t) = i) = \text{Prob}(\eta_j(t+1) \leq a) r_{i,j}$ and thus

$$\Lambda(\theta) = \log sp(\mathbf{r}\alpha(\theta)), \tag{38}$$

where $\alpha(\theta)$ is the diagonal matrix $\text{diag}\{\alpha_1(\theta), \ldots, \alpha_M(\theta)\}$ and $sp(\mathbf{r}\alpha(\theta))$ is the spectral radius of the matrix $\mathbf{r}\alpha(\theta)$.

For the usual voice model, one has for a two-state Markov chain that

$$\Lambda(\theta) = \log\Big(\frac{r_{11}\alpha_1(\theta) + r_{22}\alpha_2(\theta) + \sqrt{(r_{11}\alpha_1(\theta) - r_{22}\alpha_2(\theta))^2 + 4r_{12}r_{21}\alpha_1(\theta)\alpha_2(\theta)}}{2}\Big). \tag{39}$$

In the next example, we show the renewal process satisfies LDP(sp) when its interarrival times have bounded supports.

Example 2.6 (Renewal process) Consider a sequence of nonnegative i.i.d. $I\!R$-valued random variables $\{\xi(n), n \geq 1\}$. Let $\tau(n) = \tau(n-1) + \xi(n)$ and $\tau(0) = 0$. Then $a(t) = \sum_{n=1}^{\infty} \mathbf{1}_{\{\tau(n)=t\}}$ is a (discrete-time) renewal process. Assume that $1 \leq \xi(n) \leq \xi_{max} < \infty$ a.s. for some constant ξ_{max}. Then $a(t)$ can be modelled by a finite state Markov chain. Using the bounding result in Chang, Heidelberger, Juneja and Shahabuddin [8], pp. 12, it can be shown that (26) is satisfied with

$$\Lambda(\theta) = \hat{\xi}^{-1}(\theta), \quad and \tag{40}$$

$$\Gamma(\theta) = |\theta| + |\Lambda(\theta)|\xi_{max}, \tag{41}$$

where $\hat{\xi}^{-1}(\theta)$ is the inverse function of $\hat{\xi}(\gamma) = -\log E\exp(-\gamma\xi(1))$. That $\Lambda(\theta)$ is convex and differentiable follows from Example 2.5. We note that the LDP(sp) result (without the bounding assumption on $\xi(n)$) was recently obtained by Glynn and Whitt [17] using the inverse mapping and Mogulskii's result for i.i.d. random variables Mogulskii [27].

The third example is regarding the autoregressive process which is often used as a model for a video source.

Example 2.7 (Autoregressive process) Consider a standard autoregressive process $a(t)$ that is driven by a sequence of $M \times 1$ i.i.d. random vectors $u(t)$:

$$a(t) = \upsilon\xi(t), \quad \xi(t) = \mathbf{r}\xi(t-1) + u(t),$$

where $\mathbf{r}$ is an $M \times M$ constant matrix, v is a $1 \times M$ constant vector, and $\xi(t)$'s are $M \times 1$ state vectors. Assume that $sp(\mathbf{r}) < 1$ so that $\mathbf{I} - \mathbf{r}$ is invertible, where $\mathbf{I}$ is the identity matrix. Also, the matrices v and $\mathbf{r}$ are nonnegative and $u(t)$ is nonnegative and bounded, i.e., there exists a vector u_{max} such that $\mathbf{0} \le u(t) \le u_{max}$ a.s., where $\mathbf{0}$ is a zero vector with the appropriate dimension. Using the result in Chang, Heidelberger, Juneja and Shahabuddin [8], Example 2.6, one can show that the condition in (26) is satisfied with

$$\Lambda(\theta) = \log E e^{\theta v (\mathbf{I}-\mathbf{r})^{-1} u(t)} \tag{42}$$

$$\Gamma(\theta) = |\theta| v (\mathbf{I} - \mathbf{r})^{-1} u_{max} \tag{43}$$

It is clear that $\Lambda(\theta)$ is finite, differentiable and convex.

It is easy to see that if each R-valued process, as in Example 2.5-Example 2.7, satisfies (25) and (26) in Lemma 2.4, then the joint process also satisfies these two conditions if the individual processes are independent. This result is stated in the following corollary.

Corollary 2.8 *Suppose that $\{a(t) = (a_1(t), \dots, a_d(t)), t = 0, 1, 2, \dots, \}$ with $a_j(t) \in R^{d_j}$ being adapted to a filtration $\mathcal{F}_{j,t}$. Assume that $\mathcal{F}_{j,t}$, $j = 1, \dots, d$, are independent. If for all $\gamma \in I\!R$, $j = 1, \dots, d$,*

$$\sup_{t,s} \frac{1}{t} \log E \exp(\gamma \sum_{l=s+1}^{s+t} |a_j(l)|) < \infty, \tag{44}$$

and for all $j = 1, \dots, d$, $\theta_j \in I\!R^{d_j}$, $s, t \ge 0$, there is a differentiable function $\Lambda_j(\theta_j) < \infty$ and a function $0 \le \Gamma_j(\theta_j) < \infty$ independent of s and t,

$$\Lambda_j(\theta_j)t - \Gamma_j(\theta_j) \le \log E(e^{\theta_j \cdot A_j(s,t+s)} | \mathcal{F}_{j,s}) \le \Lambda_j(\theta_j)t + \Gamma_j(\theta_j), \quad a.s. \tag{45}$$

then $\mu_A^{(n)}$ satisfies LDP(sp) w.r.t. $\Lambda(\theta_1, \dots, \theta_d) = \sum_{j=1}^d \Lambda_j(\theta_j)$.

In view of Corollary 2.8 and (A1-3), the joint process of all external arrival processes and routing processes satisfies LDP(sp). In the next section, we will show this property is preserved under various network operations, including these three in (4)-(6).

3 Closure properties of sample path large deviations

In this section, we discuss closure properties of LDP(sp) via the contraction principle [Dembo and Zeitouni [9], Theorem 4.2.1].

Proposition 3.1 (Contraction Principle) *Let $\mu_A^{(n)}$ be the distribution of $\{A^{(n)}(t), 0 \le t \le 1\} \in D[[0,1], (I\!R^{d_1}, \|\cdot\|_\infty)]$, and $\mu_B^{(n)}$ be the distribution of $\{B^{(n)}(t), 0 \le t \le 1\} \in D[[0,1], (I\!R^{d_2}, \|\cdot\|_\infty)]$ such that $\{B^{(n)}(t), 0 \le t \le 1\} = f(\{A^{(n)}(t), 0 \le t \le 1\})$, where $f : D[[0,1], (I\!R^{d_1}, \|\cdot\|_\infty)] \mapsto D[[0,1], (I\!R^{d_2}, \|\cdot\|_\infty)]$ is a continuous function. If $\mu_A^{(n)}$ satisfies LDP(sp) with the good rate function $I_A(\phi)$, then $\mu_B^{(n)}$ satisfies LDP(sp) with the good rate function $I_B(\psi)$, where*

$$I_B(\psi) = \inf\{I_A(\phi) : \psi = f(\phi)\}, \tag{46}$$

with the convention the infimum over an empty set is ∞.

Note that Proposition 3.1 holds as long as the function f is continuous at every ϕ such that $I(\phi) < \infty$. In view of (22), the function f only needs to be continuous at every ϕ that is absolutely continuous. In the following, we will apply the contraction principle to various mappings, including sum, composition, reduction and reflection mapping. That these mappings are continuous for $\phi \in AC[[0,1],(\mathbb{R}^d,\|\cdot\|_\infty)]$ is well known (see e.g. Billingsley [5] and Whitt [33]). The proofs for the following results will be presented at the end of this section.

Lemma 3.2 (Sum) *Let $\mu_A^{(n)}$ be the distribution of $\{A^{(n)}(t), 0 \le t \le 1\} \in D[[0,1],(\mathbb{R}^d,\|\cdot\|_\infty)]$, and $\mu_B^{(n)}$ be the distribution of $\{B^{(n)}(t), 0 \le t \le 1\} \in D[[0,1],(\mathbb{R}^{d-1},\|\cdot\|_\infty)]$ such that $B_j^{(n)}(t) = A_j^{(n)}(t)$, $j = 1,\ldots,d-2$, and $B_{d-1}^{(n)}(t) = A_{d-1}^{(n)}(t) + A_d^{(n)}(t)$. If the sequence of measures $\mu_A^{(n)}$ satisfies LDP(sp) w.r.t. $\Lambda_A(\theta_1,\ldots,\theta_d)$, then $\mu_B^{(n)}$ satisfies LDP(sp) w.r.t. $\Lambda_B(\theta_1,\ldots,\theta_{d-1})$, where*

$$\Lambda_B(\theta_1,\ldots,\theta_{d-1}) = \Lambda_A(\theta_1,\ldots,\theta_{d-2},\theta_{d-1},\theta_{d-1}). \tag{47}$$

Note that if $\{A^{(n)}(t), 0 \le t \le 1\}$ is obtained from scaling a process $A(0,t)$ as in (24), then $\{B^{(n)}(t), 0 \le t \le 1\}$ is also obtained from scaling the process $B(0,t)$ with $B_j(0,t) = A_j(0,t)$, $j = 1,\ldots,d-2$, and $B_{d-1}(0,t) = A_{d-1}(0,t) + A_d(0,t)$. In view of Corollary 2.8, where all the subprocesses are independent and $\Lambda_A(\cdot)$ is a separable function, Lemma 3.2 implies the rule for multiplexing in (4) in our intree network setting.

Lemma 3.3 (Reduction) *Let $\mu_A^{(n)}$ be the distribution of $\{A^{(n)}(t), 0 \le t \le 1\} \in D[[0,1],(\mathbb{R}^d,\|\cdot\|_\infty)]$, and $\mu_B^{(n)}$ be the distribution of $\{B^{(n)}(t), 0 \le t \le 1\} \in D[[0,1],(\mathbb{R}^{d-1},\|\cdot\|_\infty)]$ such that $B_j^{(n)}(t) = A_j^{(n)}(t)$, $j = 1,\ldots,d-1$. If the sequence of measures $\mu_A^{(n)}$ satisfies LDP(sp) w.r.t. $\Lambda_A(\theta_1,\ldots,\theta_d)$, then $\mu_B^{(n)}$ satisfies LDP(sp) w.r.t. $\Lambda_B(\theta_1,\ldots,\theta_{d-1})$, where*

$$\Lambda_B(\theta_1,\ldots,\theta_{d-1}) = \Lambda_A(\theta_1,\ldots,\theta_{d-1},0). \tag{48}$$

Analogous to the argument after Lemma 3.2, Lemma 3.3 can be applied for the scaling in (24). In our network setting, Lemma 3.3 implies that those irrelevant processes (processes do not contribute arrivals to a particular queue) can be omitted (replaced by 0).

Lemma 3.4 (Composition) *Let $\mu_A^{(n)}$ be the distribution of $\{(A^{(n)}(t), P^{(n)}(t)), 0 \le t \le 1\} \in D[[0,1],(\mathbb{R}^d \times \mathbb{R}^{d_2},\|\cdot\|_\infty)]$, and $\mu_B^{(n)}$ be the distribution of $\{B^{(n)}(t), 0 \le t \le 1\} \in D[[0,1],(\mathbb{R}^{d+d_2-1},\|\cdot\|_\infty)]$ such that $B_j^{(n)}(t) = A_j^{(n)}(t)$, $j = 1,\ldots,d-1$, and $B_{d+j-1}^{(n)}(t) = P_j^{(n)}(A_d^{(n)}(t))$, $j = 1,\ldots,d_2$. Assume that $A_d^{(n)}(t)$ is increasing (nondecreasing) in t and $A_d^{(n)}(1) \le 1$. Suppose that the sequence of measures $\mu_A^{(n)}$ satisfies LDP(sp) w.r.t. $\Lambda_A(\theta_1,\ldots,\theta_d) + \Lambda_P(\theta_{d+1},\ldots,\theta_{d+d_2})$, where $\Lambda_A(\cdot)$ and $\Lambda_P(\cdot)$ satisfy (E1-3). Then $\mu_B^{(n)}$ satisfies LDP(sp) w.r.t. $\Lambda_B(\theta_1,\ldots,\theta_{d+d_2-1})$, where*

$$\Lambda_B(\theta_1,\ldots,\theta_{d+d_2-1}) = \Lambda_A(\theta_1,\ldots,\theta_{d-1},\Lambda_P(\theta_d,\ldots,\theta_{d+d_2-1})). \tag{49}$$

Since $\mu_A^{(n)}$ satisfies LDP(sp) on $D[[0,T],(I\!R^d,\|\cdot\|_\infty)]$ for any fixed $T \geq 0$, the condition $A_d^{(n)}(t) \leq 1$ can be relaxed to $A_d^{(n)}(t) \leq T < \infty$ for any fixed $T \geq 0$.

To apply Lemma 3.4 to our intree network setting, we use the scaling in (24) with $B_j(0,t) = A_j(0,t)$, $j = 1,\ldots,d-1$, and $B_{d+j-1}(0,t) = P_j(0,A_d(0,t))$, $j = 1,\ldots,d_2$. It is easy to verify that when $B^{(n)}(t) = B(0,\lceil nt\rceil)/n$, $P^{(n)}(t) = P(0,\lceil nt\rceil)/n$, and $A_d(0,t)$ is an integer, we have $B_j^{(n)}(t) = A_j^{(n)}(t)$, $j = 1,\ldots,d-1$, and $B_{d+j-1}^{(n)}(t) = P_j^{(n)}(A_d^{(n)}(t))$, $j = 1,\ldots,d_2$. Also, the conditions that $A_d^{(n)}(t)$ is increasing in t and $A_d^{(n)}(1) \leq 1$ correspond to that $a_d(t)$ ($A_d(0,t) = \sum_{s=1}^t a_d(s)$) is nonnegative and bounded above by a finite constant, as in our assumption (A2). Now view $A_d(0,t)$ as the number of arrivals within $(0,t]$ and $P_j(0,m)$ as the number of customers routed to a particular queue during the first m consecutive customers. Then $B_{d+j-1}(0,t)$ is the number of customers routed to a particular queue within $(0,t]$ (see also Chang [6] for marked point processes). Thus, in our network setting Lemma 3.4 implies the rule for routing in (6).

Lemma 3.5 (Reflection Mapping) *Let $\mu_A^{(n)}$ be the distribution of $\{A^{(n)}(t), 0 \leq t \leq 1\} \in D[[0,1],(I\!R^d,\|\cdot\|_\infty)]$, and $\mu_B^{(n)}$ be the distribution of $\{B^{(n)}(t), 0 \leq t \leq 1\} \in D[[0,1],(I\!R^d,\|\cdot\|_\infty)]$ such that $B_j^{(n)}(t) = A_j^{(n)}(t)$, $j = 1,\ldots,d-1$, and*

$$B_d^{(n)}(t) = \inf_{\{0\leq s\leq t\}} [A_d^{(n)}(s) + c(t-s)], \tag{50}$$

where c is a nonnegative constant. Assume that $A_d^{(n)}(t)$ is increasing in t. If the sequence of measures $\mu_A^{(n)}$ satisfies LDP(sp) w.r.t. $\tilde\Lambda_A(\theta_1,\ldots,\theta_{d-1}) + \Lambda_A(\theta_d)$, where $\Lambda_A(\cdot)$ and $\tilde\Lambda_A(\cdot)$ satisfy (E1-3), and $\tilde\theta \geq 0$ is a solution of $\Lambda'_A(\theta) = c$ (if $\Lambda'_A(\theta) < c$ for all $\theta > 0$, then $\tilde\theta = \infty$), then $\mu_B^{(n)}$ satisfies LDP(sp) w.r.t. $\Lambda_B(\theta_1,\ldots,\theta_d)$, where

$$\Lambda_B(\theta_1,\ldots,\theta_d) = \begin{cases} \tilde\Lambda_A(\theta_1,\ldots,\theta_{d-1}) + \Lambda_A(\theta_d) & \text{if } \theta_d \leq \tilde\theta \\ \tilde\Lambda_A(\theta_1,\ldots,\theta_{d-1}) + \theta_d c - \tilde\theta c + \Lambda_A(\tilde\theta) & \text{if } \theta_d > \tilde\theta \end{cases} \tag{51}$$

To apply this lemma to our intree network setting, we note that $B_d^{(n)}(t)$ is closely related to the departure process from a queue. To simplify the notations, consider the following one-dimensional Lindley's equation:

$$q(t+1) = (q(t) + a_d(t+1) - c)^+, \tag{52}$$

with $q(0) = 0$. It is easy to see that if $A_d(t)$ is increasing in t, then

$$\hat B_d(0,t) = \inf_{\{0\leq s\leq t\}} [A_d(s) + c(t-s)] \tag{53}$$

is the number of departures by time t (t is an integer) from the queue that satisfies (52). Now use the scaling $A_d^{(n)}(t) = A_d(0,\lceil nt\rceil)/n$ and $\hat B_d^{(n)}(t) = \hat B_d(0,\lceil nt\rceil)/n$. Though $\hat B_d^{(n)}(t) \neq B_d^{(n)}(t)$ for all t, they are close enough to guarantee that $\hat B_d^{(n)}(t)$ also satisfies LDP(sp) w.r.t. the same function in (51). To see this, note that

$$\sup_{\{0\leq t\leq 1\}} |B_d^{(n)}(t) - \hat B_d^{(n)}(t)| \leq \frac{c}{n}. \tag{54}$$

Thus, these two processes are exponentially equivalent and one can apply Dembo and Zeitouni [9], Theorem 4.2.13 to conclude that the distribution of $\hat{B}_d^{(n)}(t)$ also satisfies LDP(sp) w.r.t. the same function as $B_d^{(n)}(t)$. Thus, we can apply Lemma 3.5 to derive the rule for the input-output relation through a queue in (5). Note that the condition $\Lambda_j'(0) < c_j$ in (A4) ensures either there is a $\tilde{\theta} \geq 0$ such that $\Lambda_j'(\tilde{\theta}) = c_j$ (since $\Lambda_j'(\theta)$ is increasing and continuous from the convexity of $\Lambda_j(\theta)$ [Avriel [3], Theorem 4.28]), or $\tilde{\theta} = \infty$.

The rest of the section is devoted to the proofs of Lemma 3.2-Lemma 3.5.

Proof. [Lemma 3.2] Let ϕ and ψ be functions in $AC[[0,1], (\mathbb{R}^d, \|\cdot\|_\infty)]$ and $AC[[0,1], (\mathbb{R}^{d-1}, \|\cdot\|_\infty)]$ respectively. Assume that $\phi_j(0) = \psi_j(0) = 0$ for all j. Consider the sets

$$E_1 \equiv \{\phi | \phi_j(t) = \psi_j(t), j = 1, \ldots, d-2, 0 \leq t \leq 1\} \tag{55}$$

and

$$E_2 \equiv \{\phi | \phi_{d-1}(t) + \phi_d(t) = \psi_{d-1}(t), 0 \leq t \leq 1\}. \tag{56}$$

From Proposition 3.1, $\mu_B^{(n)}$ satisfy LDP(sp) with the rate function

$$I_B(\psi) = \inf_{E_1 \cap E_2} I_A(\phi) = \inf_{E_1 \cap E_2} \int_0^1 \Lambda_A^*(\phi'(t))dt. \tag{57}$$

Since $\phi_j(0) = 0, j = 1, \ldots, d$, and $\psi_j(0) = 0, j = 1, \ldots, d-1$,

$$E_1 = E_1' \equiv \{\phi | \phi_j'(t) = \psi_j'(t), j = 1, \ldots, d-2, 0 \leq t \leq 1\} \tag{58}$$

and

$$E_2 = E_2' \equiv \{\phi | \phi_{d-1}'(t) + \phi_d'(t) = \psi_{d-1}'(t), 0 \leq t \leq 1\}. \tag{59}$$

Thus,

$$I_B(\psi) = \inf_{E_1' \cap E_2'} \int_0^1 \Lambda_A^*(\phi'(t))dt \tag{60}$$

$$= \inf_{E_2'} \int_0^1 \Lambda_A^*(\psi_1'(t), \ldots, \psi_{d-2}'(t), \phi_{d-1}'(t), \phi_d'(t))dt \tag{61}$$

$$= \int_0^1 \Big[\inf_{\{\phi_{d-1}'(t) + \phi_d'(t) = \psi_d'(t)\}} \Lambda_A^*(\psi_1'(t), \ldots, \psi_{d-2}'(t), \phi_{d-1}'(t), \phi_d'(t)) \Big] dt. \tag{62}$$

From (62), we have

$$\Lambda_B^*(y_1, \ldots, y_{d-1}) = \inf_{\{(x_{d-1}, x_d) | y_{d-1} = x_{d-1} + x_d\}} \Lambda_A^*(y_1, \ldots, y_{d-2}, x_{d-1}, x_d). \tag{63}$$

Since $\Lambda_A^*(\cdot)$ is the Legendre transform of $\Lambda_A(\cdot)$, it follows that $\Lambda_A^*(\cdot)$ is a closed convex function. In conjunction with that fact that $\{(x_{d-1}, x_d) | y_{d-1} = x_{d-1} + x_d\}$ is a convex set, one can easily show that $\Lambda_B^*(\cdot)$ is also a closed convex function. Applying the Legendre transform yields

$$\Lambda_B(\theta_1, \ldots, \theta_{d-1}) = \sup_{\{y_1, \ldots, y_{d-1}\}} [\sum_{j=1}^{d-1} \theta_j y_j - \Lambda_B^*(y_1, \ldots, y_{d-1})] \tag{64}$$

$$= \sup_{\{y_1,\ldots,y_{d-1}\}} \sup_{\{(x_{d-1},x_d)|y_{d-1}=x_{d-1}+x_d\}} [\sum_{j=1}^{d-1} \theta_j y_j - \Lambda_A^*(y_1,\ldots,y_{d-2},x_{d-1},x_d)] \tag{65}$$

$$= \sup_{\{y_1,\ldots,y_{d-2},x_{d-1},x_d\}} [\sum_{j=1}^{d-2} \theta_j y_j + \theta_{d-1}(x_{d-1}+x_d) - \Lambda_A^*(y_1,\ldots,y_{d-2},x_{d-1},x_d)] \tag{66}$$

$$= \Lambda_A(\theta_1,\ldots,\theta_{d-2},\theta_{d-1},\theta_{d-1}). \tag{67}$$

Since $\Lambda_A(\cdot)$ satisfies (E1-3), it is clear that $\Lambda_B(\cdot)$ satisfies (E1-3). ■

Proof. [Lemma 3.3] Analogous to the proof of Lemma 3.2, let ϕ and ψ be functions in $AC[[0,1],(I\!R^d,\|\cdot\|_\infty)]$ and $AC[[0,1],(I\!R^{d-1},\|\cdot\|_\infty)]$ respectively. Assume that $\phi_j(0)=\psi_j(0)=0$ for all j. Consider the set

$$E \equiv \{\phi|\phi_j(t)=\psi_j(t), j=1,\ldots,d-1, 0\le t\le 1\}. \tag{68}$$

From Proposition 3.1, $\mu_B^{(n)}$ satisfy LDP(sp) with the rate function

$$I_B(\psi) = \inf_E I_A(\phi) = \inf_E \int_0^1 \Lambda_A^*(\phi'(t))dt. \tag{69}$$

Since $\phi_j(0)=0, j=1,\ldots,d$, and $\psi_j(0)=0, j=1,\ldots,d-1$,

$$E = E' \equiv \{\phi|\phi_j'(t)=\psi_j'(t), j=1,\ldots,d-1, 0\le t\le 1\}. \tag{70}$$

We note that the only free variable in E' is $\phi_d'(t)$. Thus,

$$I_B(\psi) = \inf_{E'} \int_0^1 \Lambda_A^*(\phi'(t))dt \tag{71}$$

$$= \int_0^1 \Big[\inf_{\phi_d'(t)} \Lambda_A^*(\psi_1'(t),\ldots,\psi_{d-1}'(t),\phi_d'(t))\Big]dt, \tag{72}$$

From (72), we have

$$\Lambda_B^*(y_1,\ldots,y_{d-1}) = \inf_{\{x_d\}} \Lambda_A^*(y_1,\ldots,y_{d-1},x_d). \tag{73}$$

Using a similar argument as in the proof of Lemma 3.2, one can easily show that $\Lambda_B^*(\cdot)$ is a closed convex function. Applying the Legendre transform yields

$$\Lambda_B(\theta_1,\ldots,\theta_{d-1}) = \sup_{\{y_1,\ldots,y_{d-1}\}} [\sum_{j=1}^{d-1} \theta_j y_j - \Lambda_B^*(y_1,\ldots,y_{d-1})] \tag{74}$$

$$= \sup_{\{y_1,\ldots,y_{d-1}\}} \sup_{\{x_d\}} [\sum_{j=1}^{d-1} \theta_j y_j - \Lambda_A^*(y_1,\ldots,y_{d-1},x_d)] \tag{75}$$

$$= \sup_{\{y_1,\ldots,y_{d-1},x_d\}} [\sum_{j=1}^{d-1} \theta_j y_j + 0\cdot x_d - \Lambda_A^*(y_1,\ldots,y_{d-1},x_d)] \tag{76}$$

$$= \Lambda_A(\theta_1,\ldots,\theta_{d-2},\theta_{d-1},0). \tag{77}$$

That $\Lambda_B(\cdot)$ satisfies (E1-3) follows from that $\Lambda_A(\cdot)$ satisfies (E1-3). ■

Proof. [Lemma 3.4] Let ϕ_A, ϕ_P and ψ be functions in $AC[[0,1],(I\!R^d,\|\cdot\|_\infty)]$, $AC[[0,1],(I\!R^{d_2},\|\cdot\|_\infty)]$ and $AC[[0,1],(I\!R^{d+d_2-1},\|\cdot\|_\infty)]$ respectively. Assume that $\phi_{A,j}(0)=\phi_{P,j}(0)=\psi_j(0)=0$ for all j. Consider the set

$$E=\{(\phi_A,\phi_P)|\phi_{A,j}(t)=\psi_j(t), j=1,\ldots,d-1, \phi_{P,j}(\phi_{A,d}(t))=\psi_{d+j-1}(t), j=1,\ldots,d_2, 0\le t\le 1\}. \tag{78}$$

Note that the convex conjugate of $\Lambda_A(\theta^1)+\Lambda_P(\theta^2)$ is $\Lambda_A^*(y^1)+\Lambda_P^*(y^2)$, where $\Lambda_A^*(y^1)$ and $\Lambda_P^*(y^2)$ are convex conjugates of $\Lambda_A(\theta^1)$ and $\Lambda_P(\theta^2)$ respectively. From Proposition 3.1, $\mu_B^{(n)}$ satisfies LDP(sp) with the good rate function

$$I_B(\psi)=\inf_E\Big[\int_0^1 \Lambda_A^*(\phi_A'(t))+\Lambda_P^*(\phi_P'(t))dt\Big]. \tag{79}$$

Since we assume that $A_d^{(n)}(1)\le 1$, we only need to consider the case $\phi_d(1)\le 1$. (The set $\{\phi_d(1)>1\}$ has a zero probability and the rate function is infinite.) Now decompose the integral $\int_0^1 \Lambda_P^*(\phi_P'(t))dt$ into two integrals over the time intervals $[0,\phi_d(1)]$ and $[\phi_d(1),1]$. In the second time interval, there is no constraint on $\phi_P(t)$ in E. Thus, we can choose $\phi_P'(t)=\Lambda_P{}'(\mathbf{0})$ for $t\in[\phi_d(1),1]$ and the second integral vanishes. Therefore,

$$I_B(\psi)=\inf_E\Big[\int_0^1 \Lambda_A^*(\phi_A'(t))dt+\int_0^{\phi_{A,d}(1)}\Lambda_P^*(\phi_P'(s))ds\Big]. \tag{80}$$

Since $\phi_{A,j}(0)=\phi_{P,j}(0)=\psi_j(0)=0$ for all j and they are differentiable,

$$E=\{(\phi_A,\phi_P)|\phi_{A,j}'(t)=\psi_j'(t), j=1,\ldots,d-1, \phi_{P,j}'(\phi_{A,d}(t))\phi_{A,d}'(t)=\psi_{d+j-1}'(t), j=1,\ldots,d_2, 0\le t\le 1\}. \tag{81}$$

Now we change the variable in the second integral of (80) using $s=\phi_{A,d}(t)$. Applying the identities in (81) yields

$$I_B(\psi)=\inf_E\Big[\int_0^1 \Lambda_A^*(\phi_A'(t))dt+\int_0^1 \Lambda_P^*(\psi_d'(t)/\phi_{A,d}'(t),\ldots,\psi_{d+d_2-1}'(t)/\phi_{A,d}'(t))\phi_{A,d}'(t)dt\Big]. \tag{82}$$

Observe that the only free variable is $\phi_{A,d}'(t)$ in E. Hence,

$$\begin{aligned}
I_B(\psi)=&\inf_{\{\phi_{A,d}'(t),0\le t\le 1\}}\Big[\int_0^1 \Lambda_A^*(\psi_1'(t),\ldots,\psi_{d-1}'(t),\phi_{A,d}'(t))dt\\
&+\int_0^1 \Lambda_P^*(\psi_d'(t)/\phi_{A,d}'(t),\ldots,\psi_{d+d_2-1}'(t)/\phi_{A,d}'(t))\phi_{A,d}'(t)dt\Big] \qquad (83)\\
=&\int_0^1 \inf_{\{\phi_{A,d}'(t)\}}\Big[\Lambda_A^*(\psi_1'(t),\ldots,\psi_{d-1}'(t),\phi_{A,d}'(t))\\
&+\Lambda_P^*(\psi_d'(t)/\phi_{A,d}'(t),\ldots,\psi_{d+d_2-1}'(t)/\phi_{A,d}'(t))\phi_{A,d}'(t)\Big]dt \qquad (84)
\end{aligned}$$

Since we assume that $A_d^{(n)}(t)$ is increasing in t, it follows that the set $\{\phi'_{A,d}(t) < 0\}$ has probability 0 and thus $\Lambda_A^*(x_1, \ldots, x_d) = \infty$ if $x_d < 0$. This implies

$$\Lambda_B^*(y_1, \ldots, y_{d+d_2-1}) = \inf_{\{x \geq 0\}} [\Lambda_A^*(y_1, \ldots, y_{d-1}, x) + \Lambda_P^*(y_d/x, \ldots, y_{d+d_2-1}/x)x]. \tag{85}$$

From the convexity of $\Lambda_P^*(\cdot)$, note that $\Lambda_P^*(y_d/x, \ldots, y_{d+d_2-1}/x)x$ is (jointly) convex in $(y_d, \ldots, y_{d+d_2-1}, x)$ for $x \geq 0$. Using a similar argument as in the proof of Lemma 3.2, one can easily show that $\Lambda_B^*(\cdot)$ is a closed convex function. Thus,

$$\Lambda_B(\theta_1, \ldots, \theta_{d+d_2-1}) = \sup_{\{y_1,\ldots,y_{d+d_2-1}\}} [\sum_{j=1}^{d+d_2-1} \theta_j y_j - \Lambda_B^*(y_1, \ldots, y_{d+d_2-1})] \tag{86}$$

$$= \sup_{\{y_1,\ldots,y_{d+d_2-1}\}} \sup_{x \geq 0} [\sum_{j=1}^{d+d_2-1} \theta_j y_j - \Lambda_A^*(y_1, \ldots, y_{d-1}, x) - \Lambda_P^*(y_d/x, \ldots, y_{d+d_2-1}/x)x]] \tag{87}$$

$$= \sup_{\{y_1,\ldots,y_{d-1},x \geq 0\}} \Big[\sum_{j=1}^{d-1} \theta_j y_j - \Lambda_A^*(y_1, \ldots, y_{d-1}, x)$$

$$+ \sup_{\{y_d,\ldots,y_{d+d_2-1}\}} [\sum_{j=d}^{d+d_2-1} \theta_j y_j - \Lambda_P^*(y_d/x, \ldots, y_{d+d_2-1}/x)x]\Big] \tag{88}$$

$$= \sup_{\{y_1,\ldots,y_{d-1},x \geq 0\}} \Big[\sum_{j=1}^{d-1} \theta_j y_j + x\Lambda_P(\theta_d, \ldots, \theta_{d+d_2-1}) - \Lambda_A^*(y_1, \ldots, y_{d-1}, x)\Big] \tag{89}$$

$$= \Lambda_A(\theta_1, \ldots, \theta_{d-1}, \Lambda_P(\theta_d, \ldots, \theta_{d+d_2-1})). \tag{90}$$

If $\Lambda_A(\theta_1, \ldots, \theta_d)$ is increasing in θ_d, then it follows from (E1-3) of $\Lambda_A(\cdot)$ and $\Lambda_P(\cdot)$ that $\Lambda_B(\cdot)$ also satisfies $(E1-3)$. Since $\Lambda_A^*(x_1, \ldots, x_d) = \infty$ if $x_d < 0$, we have from (23) that

$$\Lambda_A(\theta) = \sup_{\{x, x_d \geq 0\}} [\theta \cdot x - \Lambda_A^*(x)]. \tag{91}$$

Since $\theta \cdot x - \Lambda_A^*(x)$ is increasing in θ_d for all x with $x_d \geq 0$, $\Lambda_A(\theta)$ is increasing in θ_d. ∎

Proof. [Lemma 3.5] In view of the argument in the proof of Lemma 3.4, it is obvious that $\tilde{\Lambda}_A(\theta_1, \ldots, \theta_{d-1})$ does not play a role in the proof and it suffices to prove the one-dimensional case, namely $D[[0,1], (I\!\!R, \|\cdot\|_\infty)]$. To simplify the notation, we omit the subscript d. Let ϕ and ψ be functions in $AC[[0,1], (I\!\!R, \|\cdot\|_\infty)]$. Assume that $\phi(0) = \psi(0) = 0$. From Proposition 3.1, it follows that $\mu_B^{(n)}$ satisfy LDP(sp) with the rate function

$$I_B(\psi) = \inf\{I_A(\phi) : \psi(t) = \inf_{\{0 \leq s \leq t\}} [\phi(s) + c(t-s)], 0 \leq t \leq 1\}. \tag{92}$$

In physical world, ψ is analogous to the output process from a bucket with output rate c subject to the input process ϕ. The bucket level at time t is $\phi(t) - \psi(t)$. We will argue that the minimum in (92) is achieved when $\phi = \psi$. For a particular sample path ϕ, consider the following time epochs:

$$t_0 = 0, \tag{93}$$
$$t_{2n-1} = \min[1, \inf\{t > t_{2n-2} : \phi'(t) > c\}], n = 1, 2, \ldots, \tag{94}$$
$$t_{2n} = \min[1, \inf\{t > t_{2n-1} : \phi(t) - \phi(t_{2n-1}) \le c(t - t_{2n-1})\}], n = 1, 2, \ldots. \tag{95}$$

Since we assume that $A_d^{(n)}(t)$ is increasing in t, we only need to consider $\phi'(t) \ge 0$. From the reflection mapping in (50), the bucket is empty (resp. not empty) in the intervals $[t_{2n-2}, t_{2n-1})$ (resp. $[t_{2n-1}, t_{2n})$), $n = 1, 2, , \ldots$. Thus, $\psi(t) = \phi(t)$ in $[t_{2n-2}, t_{2n-1})$, $n = 1, 2, \ldots$, and $\psi'(t) = c$ in $[t_{2n-1}, t_{2n})$, $n = 1, 2, \ldots$. Since $\Lambda_A^*(\cdot)$ is a convex function, it follows from Jensen's inequality that

$$\frac{1}{t_{2n} - t_{2n-1}} \int_{t_{2n-1}}^{t_{2n}} \Lambda_A^*(\phi'(t))dt \ge \Lambda_A^*\Big(\frac{1}{t_{2n} - t_{2n-1}} \int_{t_{2n-1}}^{t_{2n}} \phi'(t)dt\Big) = \Lambda_A^*\Big(\frac{\phi(t_{2n}) - \phi(t_{2n-1})}{t_{2n} - t_{2n-1}}\Big). \tag{96}$$

If $t_{2n} < 1$, then we have from the continuity of ϕ and (95) that $\phi(t_{2n}) - \phi(t_{2n-1}) = c(t_{2n} - t_{2n-1})$. If $t_{2n} = 1$, then $\phi(t_{2n}) - \phi(t_{2n-1}) \ge c(t_{2n} - t_{2n-1})$ from (95). Thus, in both cases

$$\phi(t_{2n}) - \phi(t_{2n-1}) \ge c(t_{2n} - t_{2n-1}). \tag{97}$$

If $\Lambda_A^*(y)$ is increasing for $y \ge c$, then it follows from (96) and (97) that the minimum action path in $[t_{2n-1}, t_{2n})$ is achieved when $\phi'(t) = c = \psi'(t)$. To show $\Lambda_A^*(y) \ge \Lambda_A^*(c)$ for $y \ge c$, observe that

$$\Lambda_A^*(c) = \sup_\theta [\theta c - \Lambda_A(\theta)] = \tilde{\theta} c - \Lambda_A(\tilde{\theta}). \tag{98}$$

Since $\tilde{\theta} \ge 0$ and $y \ge c$,

$$\Lambda_A^*(y) = \sup_\theta [\theta y - \Lambda_A(\theta)] \ge \tilde{\theta} y - \Lambda_A(\tilde{\theta}) \ge \Lambda_A^*(c). \tag{99}$$

Thus, we conclude that the path $\phi = \psi$ achieves the minimum action and that

$$I_B(\psi) = \begin{cases} I_A(\psi) & \text{if } \psi'(t) \le c, 0 \le t \le 1 \\ \infty & \text{otherwise} \end{cases} \tag{100}$$

This is equivalent to defining $\Lambda_B^*(x) = \Lambda_A^*(x)$ if $x \le c$ and $\Lambda_B^*(x) = \infty$ if $x > c$. Since $\Lambda_A^*(x)$ is a closed convex function, $\Lambda_B^*(x)$ is also a closed convex function.

Now we show that $\Lambda_B(\theta)$ in (51) is the convex conjugate of $\Lambda_B^*(y)$, i.e., $\Lambda_B(\theta) = \sup_{y \le c}[\theta y - \Lambda_A^*(y)]$. Observe that for $\Lambda_A'(\theta) \le c$,

$$\theta \Lambda_A'(\theta) - \Lambda_A^*(\Lambda_A'(\theta)) \le \sup_{y \le c}[\theta y - \Lambda_A^*(y)] \le \sup_y [\theta y - \Lambda_A^*(y)] = \Lambda_A(\theta). \tag{101}$$

Since $\Lambda_A(\theta)$ is convex and differentiable, it follows that Avriel [3], Theorem 5.3 that

$$\Lambda_A(\theta) = \theta \Lambda_A'(\theta) - \Lambda_A^*(\Lambda_A'(\theta)). \tag{102}$$

This implies that $\Lambda_B(\theta) = \Lambda_A(\theta)$ for all $\Lambda'_A(\theta) \le c$. For the case $\Lambda'_A(\theta) > c$, one has from (102) and the fact $\Lambda_A(\theta)$ is the convex conjugate of $\Lambda^*_A(y)$ that $\sup_y[\theta y - \Lambda^*_A(y)]$ is achieved at $y = \Lambda'_A(\theta) > c$. Hence by concavity of $\theta y - \Lambda^*_A(y)$, the sup restricted to $y \le c$ is achieved at c. Thus, $\Lambda_B(\theta) = \theta c - \Lambda^*_A(c) = \theta c - \tilde{\theta} c + \Lambda_A(\tilde{\theta})$ for $\Lambda_A(\theta) > c$.

One can also verify that $\Lambda_B(\theta)$ satisfies (E1-3). ■

4 Preliminary results for a single queue

in this section we extend some known results for a single queue and strengthen the upper bound in Chang [6]. These results are then used to prove our main theorems in Section 5. Consider a discrete-time queue model with one class of customers. The service requirements of all customers are one unit of time. Also, the buffer size is infinite and the capacity of the server is c, i.e., at most c customers can be served per unit of time. Let $a(t)$, $k = 1, \ldots, K$, be the number of arrivals at time t. Denote $q(t)$ as the number of customers in the queue at time t. Then under a work-conserving policy the queue is governed by the following Lindley's equation:

$$q(t+1) = (q(t) + a(t+1) - c)^+. \tag{103}$$

Furthermore, we assume that the queue is empty at time 0, i.e., $q(0) = 0$.

Let $A^{(n)}(t) = A(0, \lceil nt \rceil)/n$ and $q^{(n)}(t) = q(\lceil nt \rceil)/n$. Let $\mu_A^{(n)}$ and $\mu_Q^{(n)}$ be the distributions of $A^{(n)}(t)$ and $(q^{(n)}(\cdot) | q^{(n)}(1) > \alpha)$, respectively. Now we state a conditional limit theorem in Dembo and Zajic [10], Proposition 4. The i.i.d. case is well known (see e.g. Anantharam [1]). For other related works, we refer to Parekh and Walrand [29], Anantharam, Heidelberger, and Tsoucas [2], Frater, Lennon, and Anderson [13], Heidelberger [20] and references therein.

Lemma 4.1 *Suppose that $\mu_A^{(n)}$ satisfies LDP(sp) w.r.t. $\Lambda_A(\theta)$. If there is a unique solution $\theta^* > 0$ of the equation $\Lambda_A(\theta) - c\theta = 0$ and there is a $\delta > 0$ such that $\alpha + \delta = \Lambda'_A(\theta_0) - c$ for some $\theta_0 < \infty$, then $\mu_Q^{(n)}$ converges in distribution on $D[[0,1], (I\!R, \|\cdot\|_\infty)]$ to δ_{q_α}, where δ_{q_α} denotes the probability distribution concentrated on*

$$q_\alpha(t) = \begin{cases} 0, & if \;\; 0 \le t \le 1-\beta; \\ \frac{\alpha}{\beta}(t - (1-\beta)), & if \;\; 1-\beta \le t \le 1, \end{cases} \tag{104}$$

and $\beta = \min[1, \alpha/(\Lambda'_A(\theta^*) - c)]$.

Remark 4.2 *We note that the conditions stated in Dembo and Zajic [10], Proposition 4, are for the rate function $\Lambda^*_A(x)$. Here we verify that the conditions for $\Lambda_A(\theta)$ in Lemma 4.1 indeed imply those in Dembo and Zajic [10]. Without loss of generality, assume $c = 0$ (the case $c > 0$ is a simple translation). Since we assume that $\Lambda_A(\theta^*) = 0$ and $\theta^* > 0$ is unique, in conjunction with $\Lambda_A(0) = 0$ and the convexity of $\Lambda_A(\theta)$ we have $\Lambda_A(\theta) < 0$, $0 < \theta < \theta^*$. This implies $\Lambda'_A(0) < 0$. By the assumption that $\Lambda_A(\theta)$ is differentiable everywhere and Rockafellar [30], Theorem 26.3, $\Lambda^*_A(y) > 0$ for all $y \ne \Lambda'_A(0)$ and $\Lambda^*_A(y)$ is strictly convex for all y in*

$\{y : \Lambda_A^*(y) < \infty\}$ *(cf. the remark after Dembo and Zajic [10], Proposition 4). Also, from the assumption that* $\Lambda_A'(\theta_0) = \alpha + \delta$, *it follows that*

$$\Lambda_A^*(\alpha + \delta) = \sup_\theta[\theta(\alpha + \delta) - \Lambda_A(\theta)] = \theta_0(\alpha + \delta) - \Lambda_A(\theta_0) < \infty. \tag{105}$$

Thus, it is left to show that

$$\Lambda_A'(\theta^*) = argmin_{y\geq 0}\frac{\Lambda_A^*(y)}{y}. \tag{106}$$

For $y \geq 0$,

$$\frac{\Lambda_A^*(y)}{y} = \frac{\sup_\theta[\theta y - \Lambda_A(\theta)]}{y} \geq \frac{\theta^* y - \Lambda_A(\theta^*)}{y} = \theta^*. \tag{107}$$

Since $\Lambda_A(\theta)$ *is convex and differentiable,* $\Lambda_A(\theta) = \theta\Lambda_A'(\theta) - \Lambda_A^*(\Lambda_A'(\theta))$ *[see Avriel [3], Theorem 5.3]. In conjunction with (107), one has for all* $y \geq 0$

$$\frac{\Lambda_A^*(y)}{y} - \frac{\Lambda_A^*(\Lambda_A'(\theta^*))}{\Lambda_A'(\theta^*)} = \frac{\Lambda_A^*(y)}{y} - \theta^* \geq 0. \tag{108}$$

In the following corollary, we show the property of LDP(sp) can be also used for deriving the lower bound for the tail distribution.

Corollary 4.3 *Suppose that* $\lim_{t\to\infty} q(t) \to q(\infty)$ *in distribution for some random variable* $q(\infty)$ *and* $Prob(q(t) \geq x) \leq Prob(q(\infty) \geq x)$ *for all* t *and* x. *Then under the conditions in Lemma 4.1,*

$$\liminf_{x\to\infty}\frac{\log Prob(q(\infty) > x)}{x} \geq -\theta^*. \tag{109}$$

Proof. We use an approach analogous to that in Chang [6], Theorem 3.9 (ii). Observe that

$$q(t) = \max_{1\leq s\leq t}[0, A(t-s,t) - sc] \geq A(0,t) - ct. \tag{110}$$

In conjunction with the assumption $\mathrm{Prob}(q(t) \geq x) \leq \mathrm{Prob}(q(\infty) \geq x)$, we have for all t

$$\liminf_{x\to\infty}\frac{\log \mathrm{Prob}(q(\infty) \geq x)}{x} \geq \liminf_{x\to\infty}\frac{\log \mathrm{Prob}(A(0,t) \geq ct + x)}{x}. \tag{111}$$

Since $\mu_A^{(n)}$ satisfies LDP(sp), it follows from Proposition 2.3 that $\{A(0,t), t = 1, 2, \ldots\}$ satisfies LDP. Letting $y = c + (x/t)$ for $y \geq c$ and applying the large deviation lower bound in (17) yields

$$\liminf_{x\to\infty}\frac{\log \mathrm{Prob}(q(\infty) \geq x)}{x} \geq \liminf_{t\to\infty}\frac{\log \mathrm{Prob}(A(0,t) \geq yt)}{(y-c)t} \geq -\frac{\Lambda_A^*(y)}{y-c}. \tag{112}$$

Optimizing y and using (107) yields the desired result. ■

To obtain the upper bound for the tail distribution, we need to introduce a stronger condition in Chang [6]. Define

$$\Lambda_A^u(\theta) \equiv \limsup_{t\to\infty}\frac{1}{t}\sup_{s\geq 0}\log Ee^{\theta A(s,s+t)}, \quad \theta \geq 0. \tag{113}$$

We note that $\Lambda_A^u(\theta)/\theta$ is the definition of the MER w.r.t. θ in Chang [6] and the domain $\theta < 0$ does not play a role in the derivation of the upper bound. If $a(t)$ is stationary and the Gärtner-Ellis limit exists, then

$$\Lambda_A^u(\theta) = \lim_{t\to\infty} \frac{1}{t} \log E e^{\theta A(0,t)} = \Lambda_A(\theta). \tag{114}$$

In view of the definition of $\Lambda_A^u(\theta)$, $\Lambda_A^u(\theta)$ is convex. In the following theorem, we state the asymptotic result for the tail distribution.

Theorem 4.4 *Suppose the conditions in Lemma 4.1 and Corollary 4.3 hold. If $\Lambda_A^u(\theta) = \Lambda_A(\theta)$ for all $\theta \geq 0$, then*

$$\lim_{x\to\infty} \frac{\log Prob(q(\infty) > x)}{x} = -\theta^*. \tag{115}$$

Unlike similar results in Chang [6] and Glynn and Whitt [16], we do not assume that the arrival process $\{a(t), t = 0, 1, 2, \ldots\}$ is *stationary*. This feature is important since the departure process from a queue may not be stationary (though there exists a stationary version). It is also clear that the conditions on $q(t)$ in Corollary 4.3 are satisfied if $a(t)$ is stationary and ergodic with $Ea(t) < c$ (see Chang [6], Theorem 3.9).

The proof of Theorem 4.4 follows directly from Corollary 4.3 and Chang [6], Theorem 3.8 stated in Proposition 4.5 below. Proposition 4.5 also contains an upper bound for the departure process. Let $b(t)$ be the number of departures at time t, i.e., $b(t) = \min[q(t-1) + a(t), c]$. Let $B(t_1, t_2) = \sum_{s=t_1+1}^{t_2} b(s)$.

Proposition 4.5 *If $\Lambda_A^u(\theta) < \theta c$ for some $\theta \geq 0$, then*

$$\sup_t E \exp(\theta q(t)) < \infty, \tag{116}$$

$$\Lambda_B^u(\theta) \equiv \limsup_{t\to\infty} \frac{1}{t} \sup_{s\geq 0} \log E e^{\theta B(s,s+t)} \leq \Lambda_A^u(\theta). \tag{117}$$

Thus, if $\lim_{t\to\infty} q(t) \to q(\infty)$ in distribution for some random variable $q(\infty)$, then

$$\limsup_{x\to\infty} \frac{\log Prob(q(\infty) > x)}{x} \leq -\theta^u, \tag{118}$$

where $\theta^u \equiv \sup\{\theta \geq 0 : \Lambda_A^u(\theta) < \theta c\}$.

In the following lemma, we strengthen the upper bound in (117) using the convexity of $\Lambda_B^u(\theta)$.

Lemma 4.6 *If $\Lambda_A^u(\theta)$ is differentiable, then*

$$\Lambda_B^u(\theta) \leq \Lambda_B(\theta) \tag{119}$$

where

$$\Lambda_B(\theta) = \begin{cases} \Lambda_A^u(\theta) & \text{if } 0 \leq \theta \leq \tilde{\theta} \\ \theta c - \tilde{\theta} c + \Lambda_A^u(\tilde{\theta}) & \text{if } \theta > \tilde{\theta} \end{cases} \tag{120}$$

and $\tilde{\theta}$ is a solution of ${\Lambda_A^u}'(\theta) = c$.

Proof. Since $b(t) \le c$, for all $\theta \ge 0$

$$\Lambda_B^u(\theta) \le c\theta. \tag{121}$$

In conjunction with Proposition 4.5, we have $\Lambda_B^u(\theta) \le \min[\theta c, \Lambda_A^u(\theta)]$ for all $\theta \ge 0$. Thus, we only need to show that $\Lambda_B^u(\theta) \le \Lambda_B(\theta)$ for $\theta > \tilde{\theta}$. We argue this by contradiction. Suppose that there is a $\theta_1 > \tilde{\theta}$ such that $\Lambda_B^u(\theta_1) > \theta_1 c - \tilde{\theta} c + \Lambda_A^u(\tilde{\theta})$. Since $\Lambda_B^u(\tilde{\theta}) \le \Lambda_A^u(\tilde{\theta})$,

$$\epsilon \equiv \frac{\Lambda_B^u(\theta_1) - \Lambda_B^u(\tilde{\theta})}{\theta_1 - \tilde{\theta}} - c > 0. \tag{122}$$

From convexity, it follows that for all $\theta_2 \ge \theta_1$

$$\frac{\Lambda_B^u(\theta_2) - \Lambda_B^u(\theta_1)}{\theta_2 - \theta_1} \ge \frac{\Lambda_B^u(\theta_1) - \Lambda_B^u(\tilde{\theta})}{\theta_1 - \tilde{\theta}} = (c + \epsilon). \tag{123}$$

Thus, for $\theta_2 > ((c+\epsilon)\theta_1 - \Lambda_B^u(\theta_1))/\epsilon$, $\Lambda_B^u(\theta_2) > \theta_2 c$ and we reach a contradiction. ■

In the following theorem, we show that LDP(sp) also ensures $\Lambda_B^u(\theta) = \Lambda_B(\theta)$. This allows us to extend Theorem 4.4 to intree networks.

Let $\mu_B^{(n)}$ be the distribution of the scaled departure process $B^{(n)}(t) = B(0, \lceil nt \rceil)/n$.

Theorem 4.7 *If $\mu_A^{(n)}$ satisfies LDP(sp) and $\Lambda_A^u(\theta) = \Lambda_A(\theta)$ for all $\theta \ge 0$, then $\mu_B^{(n)}$ satisfies LDP(sp) w.r.t to $\Lambda_B(\theta)$ and*

$$\Lambda_B^u(\theta) = \Lambda_B(\theta), \quad \textit{for all}\, \theta \ge 0, \tag{124}$$

where $\Lambda_B(\theta)$ is defined in (120).

Proof. In view of Lemma 4.6, we only need to show $\Lambda_B^u(\theta) \ge \Lambda_B(\theta)$. That $\mu_B^{(n)}$ satisfies LDP(sp) w.r.t to $\Lambda_B(\theta)$ follows from Lemma 3.5. As a result of Proposition 2.3, $\{B(0,t), t = 1, 2, \ldots,\}$ satisfies the LDP with the convex good rate function $\Lambda_B^*(y) = \sup_\theta[\theta y - \Lambda_B(\theta)]$. Since $\Lambda_B^u(\theta) \le c\theta < \infty$ for all $\theta \ge 0$, it then follows from Varadhan's integral lemma (Dembo and Zajic [10], Theorem 4.3.1) that

$$\lim_{t \to \infty} \frac{1}{t} \log E e^{\theta B(0,t)} = \Lambda_B(\theta). \tag{125}$$

In conjunction with the definition of $\Lambda_B^u(\theta)$ in (117), $\Lambda_B^u(\theta) \ge \Lambda_B(\theta)$. ■

5 Applications to discrete-time intree network: the proofs for the main theorems

In this section, we apply the closure properties and the single queue results in the previous sections to prove our main theorems (Theorem 1.1 and Theorem 1.2) for discrete-time intree networks.

First, we prove Theorem 1.1. Recall that $\{a_j(t), t = 0, 1, 2, \ldots\}$, $j = 1, \ldots, J$ is the overall arrival process to queue j. Let $\mu^{(n)}_{A,j}$ be the distribution of the process $A^{(n)}_j(t) = \frac{1}{n}\sum_{s=1}^{\lceil nt \rceil} a_j(s)$ and $\mu^{(n)}_{Q_j}$ be the distribution of $(q^{(n)}_j(\cdot)|q^{(n)}_j(1) > \alpha)$, where $q^{(n)}_j(t) = q_j(\lceil nt \rceil)/n$. In view of the conditional limit theorem in Lemma 4.1, we need to establish the LDP(sp) property for the overall arrival process at each queue.

Lemma 5.1 *Under (A1-4), $\mu^{(n)}_{A,j}$ satisfies LDP(sp) w.r.t. $\Lambda_j(\theta)$ in (4), $j = 1, \ldots, J$.*

Proof. From (A1-3) and Corollary 2.8, the joint process $\{(a_{j,k}(t), k = 1, \ldots, k_j, j = 1, \ldots, J, p_j(t), j = 1, \ldots, J), t = 0, 1, 2, \ldots\}$ satisfies LDP(sp) w.r.t.

$$\Lambda(\cdot) = \sum_{j=1}^{J}\sum_{k=1}^{k_j} \Lambda_{A,j,k}(\theta_{A,j,k}) + \sum_{j=1}^{J} \Lambda_{P,j}(\theta_{P,j}). \tag{126}$$

Consider the first queue in which there are only external arrival processes. Since $a_1(t) = \sum_{k=1}^{k_1} a_{1,k}(t)$, applying Lemma 3.2 for the sum and then Lemma 3.3 yields the desired LDP(sp) result for the arrival process at the first queue. For other queues, we simply apply recursively the closure properties in Lemma 3.2 for (4), Lemma 3.4 for (6), Lemma 3.5 for (5) and Lemma 3.3 for removing irrelevant processes [cf. the induction arguments in Chang, Heidelberger, Juneja and Shahabuddin [8] and Chang [7]]. Note that one needs the assumption $\Lambda'_j(0) < c$ in (A4) when applying Lemma 3.5. This condition ensures $\tilde{\theta}_j \geq 0$. ■

Proof. [Theorem 1.1] Theorem 1.1 is a direct consequence of Lemma 5.1 and Lemma 4.1. ■

Since we assume that the external arrival processes and routing processes are bounded in (A2-3), as in the proof of Theorem 4.7 one can apply Varadhan's integral lemma (Dembo and Zajic [10], Theorem 4.3.1) and Lemma 5.1 to derive the following corollary.

Corollary 5.2 *Under (A1-4), $\Lambda_j(\theta) = \lim_{t\to\infty} \frac{1}{t}\log Ee^{\theta\sum_{s=1}^{t} a_j(s)}$, $j = 1, \ldots, J$.*

In view of Theorem 4.4, we need the following lemma to prove Theorem 1.2.

Lemma 5.3 *Under (A1-4), $\Lambda^u_j(\theta) = \Lambda_j(\theta)$, $j = 1, \ldots, J$, where*

$$\Lambda^u_j(\theta) \equiv \limsup_{t\to\infty} \frac{1}{t} \sup_{s\geq 0} \log Ee^{\theta\sum_{l=s+1}^{s+t} a_j(l)}. \tag{127}$$

Proof. First note that the stationary assumption ensures $\Lambda^u_{A,j,k}(\theta) = \Lambda_{A,j,k}(\theta)$ and $\Lambda^u_{P,j}(\theta) = \Lambda_{P,j}(\theta)$ for all j, k, where

$$\Lambda^u_{A,j,k}(\theta) \equiv \limsup_{t\to\infty} \frac{1}{t} \sup_{s\geq 0} \log Ee^{\theta\sum_{l=s+1}^{s+t} a_{j,k}(l)} \quad \textit{and} \tag{128}$$

$$\Lambda^u_{P,j}(\theta) \equiv \limsup_{t\to\infty} \frac{1}{t} \sup_{s\geq 0} \log Ee^{\theta\sum_{l=s+1}^{s+t} p_j(l)}. \tag{129}$$

Analogous to the proof of Theorem 4.7, applying the LDP(sp) result in Lemma 5.1 and the Varadhan's integral lemma yields $\Lambda_j^u(\theta) \geq \Lambda_j(\theta)$ for all j. (Note $a_{j,k}(t)$ and $p_j(t)$ are bounded by (A2-3).) That $\Lambda_j^u(\theta) \leq \Lambda_j(\theta)$ can be shown inductively using Lemma 4.6 for input-output relation, Chang [6] Lemma 3.4 for sum, and Lemma 3.7 for composition. ■

Proof. [Theorem 1.2] From Corollary 5.2 and Jensen's inequality, it then follows that

$$\Lambda_j(\theta) \geq \theta \lim_{t\to\infty} \frac{1}{t} E \sum_{s=1}^{t} a_j(s). \tag{130}$$

Thus, $\Lambda_j'(0) \geq \lim_{t\to\infty} \frac{1}{t} E \sum_{s=1}^{t} a_j(s)$. In conjunction with the assumption of stationarity and ergodicity of all arrival processes and routing processes, it then follows from the well known stability result for feedforward networks (e.g. see Baccelli and Bremaud [4]) that the condition $\Lambda_j'(0) < c_j$ implies the mean arrival rate at queue j is not larger than the capacity c_j, and that $q_J(t)$ converges in distribution to a random variable $q_J(\infty)$. Since we assume the intree network starts from an empty network, using a standard sample argument (e.g. see Stoyan [31] and Whitt [34]) one can show that the departure process from each queue is stochastically dominated by its stationary version (in the sample path sense) and thus by the standard Loynes's construction Loynes [25] $\text{Prob}(q_J(t) \geq x) \leq \text{Prob}(q_J(\infty) \geq x)$ for all t and x. Thus, Theorem 1.2 follows from Lemma 5.3 and Theorem 4.4. ■

Acknowledgements:
This research was initiated and partially completed while the author was with IBM T.J. Watson Research Center. He thanks Philip Heidelberger and Marvin Nakayama for valuable discussions, and the referee for a detailed reading of the manuscript and valuable comments and corrections. This research is also supported in part by the National Science Council, Taiwan, R.O.C., under Grants NSC 83-0208-M007-091 and NSC 83-0404-E007-052.

References

[1] V. Anantharam, "How large delays build up in a GI/G/1 queue," *Queueing Systems*, Vol. 5, pp. 345-368, 1988.

[2] V. Anantharam, P. Heidelberger, and P. Tsoucas, "Analysis of rare events in continuous time Markov chains via time reversal and fluid approximation," *IBM RC 16280*, 1990.

[3] M. Avriel, "Nonlinear Programming: Analysis and Methods", *Prentice-Hall*, 1976.

[4] F. Baccelli and P. Bremaud. *Elements of Queueing Theory*. New York: Springer Verlag, 1990.

[5] P. Billingsley. *Convergence of Probability Measures*. New York: Wiley, 1968.

[6] C.S. Chang, "Stability, queue length and delay of deterministic and stochastic queueing networks," *IEEE Transactions on Automatic Control*, Vol. 39, pp. 913-931, 1994.

[7] C.S. Chang, "Approximations of ATM networks: effective bandwidths and traffic descriptors," *IBM RC 18954*, 1993.

[8] C.S. Chang, P. Heidelberger, S. Juneja and P. Shahabuddin, "Effective Bandwidth and Fast Simulation of ATM Intree Networks," *Performance Evaluation*, Vol. 20, pp. 45-66, 1994.

[9] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Boston: Jones and Barlett Publishers, 1992.

[10] A. Dembo and Tim Zajic, "Large deviations from empirical mean and measure to partial sums processes," *preprint*, 1993.

[11] R. Ellis, "Large deviations for a general class of random vectors, " *Ann. Probab.*, Vol. 12, pp. 1-12, 1984.

[12] A.I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *preprint.*

[13] M.R. Frater, T.M. Lennon, and B.D.O. Anderson, "Optimally efficient estimation of the statistics of rare events in queueing networks," *IEEE Trans. Aut. Contr.*, Vol. 36, pp. 1395-1405, 1991.

[14] J. Gärtner, "On large deviations from invariant measure," *Theory Probab. Appl.*, Vol. 22, pp. 24-39, 1977.

[15] R.J. Gibbens and P.J. Hunt, "Effective bandwidths for the multi-type UAS channel," *Queueing Systems*, Vol. 9, pp. 17-28, 1991.

[16] P.W. Glynn and W. Whitt, "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue," *J. Appl. Prob*, Vol. 31A, pp. 131-156, 1994.

[17] P.W. Glynn and W. Whitt, "Large deviations behavior of counting processes and their inverses," *Queueing Systems*, Vol. 17, pp. 107-128, 1994.

[18] R. Guérin, H. Ahmadi and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Select. Areas Commun.,* Vol. 9, pp. 968-981, 1991.

[19] I. Iscoe, P. Ney and E. Nummelin, "Large deviations of uniformly recurrent Markov Additive Processes," *Adv. Appl. Math.*, Vol. 6, pp. 373-412, 1985.

[20] P. Heidelberger, "Fast simulation of rare events in queueing and reliability models," *IBM RC 19028*, 1993.

[21] F.P. Kelly, *Reversibility and Stochastic Networks,* New York: John Wiley & Sons, 1979.

[22] F.P. Kelly, "Effective bandwidths at multi-class queues," *Queueing Systems*, Vol. 9, pp. 5-16, 1991.

[23] G. Kesidis and J. Walrand, "Traffic policing and enforcement of effective bandwidth constraints in ATM networks," *preprint.*

[24] G. Kesidis, J. Walrand and C.S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Tran. Networking*, Vol. 1, pp. 424-428, 1993.

[25] R.M. Loynes, "The stability of a queue with non-independent inter-arrival and service times," *Proc. Camb. Phil. Soc.*, Vol. 58, pp. 497-520, 1962.

[26] H.D. Miller, "A convexity property in the theory of random variables defined on a finite Markov chain," *Ann. Math. Stat.*, Vol. 32, pp. 1260-1270, 1961.

[27] A.A. Mogulskii, "Large deviations for trajectories of multidimensional random walk," *Th. Prob. Appl.*, Vol. 21, pp. 300-315, 1976.

[28] R. Nagarajan, J. Kurose and D. Towsley, "Local allocation of end-to-end quality-of-service in high speed networks," *preprint.*

[29] S. Parekh and J. Walrand, "A quick simulation method for excessive backlogs in networks of queues," *IEEE Trans. Aut. Contr.*, Vol. 34, pp. 54-66, 1989.

[30] R.T. Rochafellar. *Convex Analysis.* Princeton: Princeton University Press, 1970.

[31] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models*, Berlin: J. Wiley & Sons, 1983.

[32] J. Walrand, *An Introduction to Queueing Networks*, New Jersey: Prentice-Hall, 1988.

[33] W. Whitt, "Some useful functions for functional limit theorems," *Math. Operat. Res.*, Vol. 5, pp. 67-85, 1980.

[34] W. Whitt, "Comparing counting processes and queues," *Adv. Appl. Prob.*, Vol. 13, pp. 207-220, 1981.

[35] W. Whitt, "Tail probability with statistical multiplexing and effective bandwidths in multiclass queues," *Telecommunication Systems*, Vol. 2, pp. 71-107, 1993.

[36] G. De. Veciana and J. Walrand, "Effective Bandwidths: call admission, traffic policing & filtering for ATM networks," Submitted to *IEEE/ACM Trans. Networking*, 1993.

[37] G. De. Veciana, C. Courcoubetis and J. Walrand, "Decoupling bandwidths for networks: a decomposition approach to resource management," Submitted to *IEEE/ACM Trans. Networking*, 1993.