The Annals of Applied Probability
1998, Vol. 8, No. 4, 1070–1129

LARGE DEVIATION PROPERTIES OF DATA STREAMS THAT SHARE A BUFFER[1]

BY KAVITA RAMANAN AND PAUL DUPUIS

Brown University

Using large deviation techniques, we analyze the tail behavior of the stationary distribution of the buffer content process for a two-station communication network. We also show how the associated rate function can be expressed as the solution to a finite-dimensional variational problem. Along the way, we develop a number of results and techniques that are of independent interest, including continuity results for the input–output mapping for certain multiclass fluid models and a new technique for obtaining large deviation principles for invariant distributions from sample path large deviation results.

1. Introduction. A problem that has attracted a great deal of interest in recent years is that of design and admission control for packet switched digital data networks. In such networks, data streams from different types of sources share the network's resources. Buffers are inserted into the network to reduce data loss due to large fluctuations in the traffic offered to a given switch. This turns out to complicate the situation from the point of view of analysis, since this sharing of a buffer couples data streams that may previously have been statistically independent.

One approach that has been discussed extensively involves the use of what is known as "effective bandwidth." The basic motivation for this concept is an attempt to characterize the properties of the various data sources that use the network in such a way that an equivalent circuit switched model of the network can be used for network management. There are many papers [14, 10, 15, 3] that show how this can be done in the context of a single switch and for a variety of data stream models. In the single buffer setting, a function of the form $H_i(\alpha)/\alpha$ is associated with the ith source, where $H_i(\alpha)$ can be defined as a certain limit of suitably normalized logarithmic moment generating functions of increments from the ith data stream. Suppose that the switch processes data at a (deterministic) rate c and that any work conserving service policy is used. Moreover, assume that the buffer content is an ergodic process with invariant distribution μ. When the data streams are independent, the following rough asymptotics can be established:

$$\sum_{i=1}^{I} \frac{H_i(M)}{M} \le c \quad \Rightarrow \quad \lim_{n\to\infty} \frac{1}{n} \log \mu([n, \infty)) \le -M. \tag{1.1}$$

Received October 1995; revised May 1997.

[1]Supported in part by NSF Grant DMS-94-03820, Army Research Office Grant DAAH04-93-G-0070 and AT&T Bell Laboratories.

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

Key words and phrases. Large deviations, data networks, effective bandwidths, fluid models, Skorokhod problem, rate function.

Such a result is quite appealing, since it implies that a relatively simple test can be used to determine whether or not (within the accuracy of the large deviations approximation) the stationary probability of exceeding any given buffer size is less than some specified value. In Section 9.1 we discuss the usefulness of this concept in the network setting. One can verify that $\mu([n, \infty))$ provides a conservative estimate of the stationary probability that data arriving at a system with a buffer of size n will find the buffer full and, as we discuss in Section 9.2, that the normalized logarithm of this quantity and $\log \mu([n, \infty))/n$ have the same asymptotic behavior as $n \to \infty$.

A considerable amount of work has been done on the single switch problem, but relatively little has been done to extend these ideas to the setting of a network of switches. A basic question one could ask is the following. Suppose a source shares a buffer at one switch with several other sources, and then proceeds downstream to a second switch that is shared with other (independent) sources. Can one establish some simple criteria which guarantee that the stationary distribution at each of the buffers satisfies a stipulated constraint? In particular, can the same bandwidth function be used in the same way for the estimation of the marginal invariant distribution μ_2 of the second switch? In other words, if I_2 represents the set of sources feeding into the second switch, and c_2 the capacity of the second switch, then is it true that (1.1) is satisfied and also for any $M_2 < \infty$,

$$\sum_{i \in I_2} \frac{H_i(M_2)}{M_2} \le c_2 \quad \Rightarrow \quad \lim_{n \to \infty} \frac{1}{n} \log \mu_2([n, \infty)) \le -M_2?$$

If such a condition holds, then the streams are said to "decouple" and the probability constraints throughout the network can be met just by ensuring that the streams entering the network satisfy simple "effective bandwidth" constraints. One would therefore like to determine if such a decoupling phenomena is to be typically expected.

In order to address these issues, one has to fully understand the nature of interactions between the various sources in the first buffer that lead to overflow of the buffer downstream. Since the important interactions can be nonstationary and depend on the timing of certain critical events at the different buffers, it becomes necessary to examine the large deviation properties of the streams at the level of sample paths. Using the same scaling as that in (1.1), in this paper we carry out a complete analysis of the asymptotic behavior of the invariant distribution at the two switches. In particular, we characterize the tails of the joint invariant distribution in terms of three finite-dimensional variational problems. As discussed in Section 9.1, a simple analysis of these finite-dimensional problems reveals that in general the decoupling property does not hold since the distribution of the data stream (and also its bandwidth function) is corrupted due to interactions with other sources in the first buffer. However, the explicit form that we obtain for the rate function should be useful in testing any other criteria proposed as an alternative to the effective bandwidth criterion. The constructions we use can be extended to larger

networks and we elaborate on this point in Section 9.3. However, the reduction of an infinite-dimensional variational problem to a finite-dimensional one that is carried out in Section 8 is quite detailed even for the two-switch model. In addition, the notation rapidly becomes unwieldy as the number of switches increases. For these reasons we focus our attention on the two-switch case. We note that the methods we use also allow the study of delays in the network, although we are not concerned with that here. Another extension that can be dealt with using the techniques developed here allows the number of users and the buffer sizes to tend to infinity together. The single switch model is considered in [24].

In the course of our analysis, we develop a number of results and techniques that are of independent interest. Chief among these are continuity properties of a "reflection mapping" for multiclass fluid models (Section 4), a method for the construction of Lyapunov functions for such models (Appendix A) and a new technique for connecting sample path large deviation properties to large deviation properties of associated invariant distributions (Section 6).

While preparing this paper we became aware of the work of O'Connell [19]. This paper uses a construction to prove continuity that is similar to the one we use in Section 4. Although the continuity result stated in Section 4 is not explicitly covered by his results, it can easily be obtained using his arguments. More recent work by Majewski [17] generalizes a construction due to Loynes [16] to obtain certain large deviation results for general feedforward networks and stationary sources.

The outline of the paper is as follows. In Section 2 we formulate the network model and define the scaling. We consider only Markov fluid models for the data streams, but the results may be extended to other types of sources. Due to the non-Markovian nature of our network model, we append a state variable to Markovianize the process in Section 3 and discuss the associated topology. The continuity of the mapping that takes the input processes into the buffer content processes is stated and proved in Section 4. In Section 5 we give the large deviation principle at the level of sample paths for the buffer content processes, and in Section 6 use these results to obtain the large deviation principle for the invariant distribution. This section assumes exponential tightness and stability results that are stated in Section 7 and proved in Appendix A. The rate function for the invariant distribution that is obtained in Section 6 is given in the form of an infinite-dimensional variational problem. In Section 8 we describe how this can be reduced to the problem of solving three finite-dimensional variational problems. Each of the simplified variational problems has an interesting interpretation, which is also discussed in Section 8. The proof of the reduction is deferred to Appendix B. We close the paper in Section 9 with remarks on extensions and a discussion of decoupling bandwidths.

We end this introduction with remarks on terminology and notation. In the current literature, the term "fluid model" seems to be applied in two quite different senses. On the one hand it is used to describe a dynamical system related to a given queueing model or similar process (e.g., a reflecting Brown-

ian motion). Roughly speaking, the dynamical system is obtained by defining at each point in the state space of the process a vector field which gives the "local mean behavior" at that point. However, the term is also employed while referring to certain data models in communication networks. In this case, a fluid model is a process with continuous sample paths that is used in lieu of a (presumably more accurate) discrete valued model. In this paper we use the term in both senses. We hope that the intended use is clear from the context. Finally, throughout the paper we use capital letters to denote stochastic processes and use lower case letters for their deterministic analogues.

2. Description of the network. In this section we describe the network model under consideration. The network is represented in Figure 1. We consider a two-queue, multiple class feedforward network. In particular, we consider a two-switch communication network with I different data sources entering the first buffer, a subset of which follows a predetermined route to the second buffer while the complement leaves the network. Generalizations are discussed in Section 9. Data is processed in a FIFO (first in–first out) manner, and a stochastic process $\xi_i(t)$ is used to model the data emitted from each source. We make the following assumption on $\xi_i(t)$.

ASSUMPTION 2.1. For each $i = 1, 2, \ldots, I$, $\xi_i(t)$ is an ergodic Markov process taking values in a finite state space $\mathscr{S}_i$.

The processes ξ_i will be right continuous with limits from the left. Let the invariant distribution of $\xi_i(t)$ be π_i and let r_i be a nonnegative function on $\mathscr{S}_i$. Suppose that $X_i(t)$ denotes the cumulative input to buffer A due to source i over the time interval $[0, t]$. We adopt the well-known fluid model [1], by which we mean that the rate of data output $\dot{X}_i(t)$ is modeled by

$$\dot{X}_i(t) = r_i(\xi_i(t)).$$

With the convention $X_i(0) = 0$, integration of the last equation yields

$$X_i(t) = \int_0^t r_i(\xi_i(s))\, ds.$$

REMARK. In this paper we will restrict our models for the data sources to be of this type: fluid models defined in terms of finite state Markov processes.

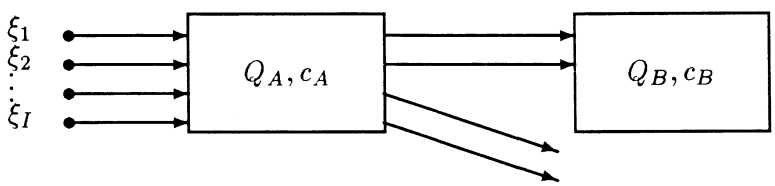

FIG. 1. *Network model.*

The analysis is applicable in far greater generality, and in fact does not require either the continuous nature of fluid model sample paths or the Markovian assumption. However, in order to simplify the exposition and keep the length of the paper within limits, we will restrict our attention to such fluid models. In particular, if one wishes to deal with sources where ξ_i is non-Markov or where the state space is not finite, many annoying additional uniformity conditions need to be checked for each particular model.

In this section we will describe the network only for the case of zero initial conditions. The case of arbitrary initial conditions is the subject of the next section.

Let $\Phi(t)$ denote the total cumulative input to buffer A over the interval $[0, t]$. Then

$$\Phi(t) = \sum_{i=1}^{I} X_i(t).$$

In order to describe the buffer content processes, we first introduce the one-dimensional Skorokhod map (see [7]). Fix $T < \infty$, and let $\mathscr{C}[0, T]$ denote the set of continuous functions from $[0, T]$ to $\mathbb{R}$. We use the standard sup norm metric on $\mathscr{C}[0, T]$: $d(f, g) = \sup_{t\in[0, T]} |f(t) - g(t)|$, and recall that with this metric $\mathscr{C}[0, T]$ is a Polish space.

The Skorokhod map that we consider is a standard tool in one dimensional queueing and reflecting diffusion models. Although it can be defined in greater generality, for our purposes we need only consider this mapping on $\mathscr{C}[0, T]$. In this case Γ: $\zeta \in \mathscr{C}[0, T] \to \psi \in \mathscr{C}[0, T]$ is given by

$$\psi(t) = \zeta(t) - \Big(\inf_{s\in[0, t]} \zeta(s) \Big) \wedge 0. \tag{2.1}$$

The mapping provides the natural "constrained version" of a path that is consistent with restricting the path to stay in the domain $[0, \infty)$ with the "least effort." Note that Γ is Lipschitz continuous with constant 2. In fact, our only interest in exhibiting the general form of the Skorokhod map is because it is in this form that this continuity is obvious. In the special case where ζ is absolutely continuous, which is the only case that we will have to consider, the Skorokhod map takes the form $\psi = \Gamma(\zeta)$ if and only if (see [8], Theorem 2)

$$\begin{aligned} &\psi(0) = \zeta(0) \vee 0, \\ &\dot{\psi} = \begin{cases} \dot{\zeta}(t), & \text{if } \psi(t) > 0, \\ \dot{\zeta}(t) \vee 0, & \text{if } \psi(t) = 0. \end{cases} \end{aligned} \tag{2.2}$$

We will assume that each buffer processes data continuously and with a given fixed rate. Let $Q_A(t)$ denote the content of buffer A at time t and let c_A be its processing rate. The equation for the buffer A content is then given by

$$\dot{Q}_A(t) = \begin{cases} \dot{\Phi}(t) - c_A, & \text{if } Q_A(t) > 0, \\ [\dot{\Phi}(t) - c_A] \vee 0, & \text{if } Q_A(t) = 0. \end{cases}$$

With the definition $g_A(t) = c_A t$, the buffer content process can be expressed succinctly in terms of the Skorokhod map as

$$Q_A = \Gamma(\Phi - g_A). \tag{2.3}$$

Suppose we define $D(t)$ to be the delay experienced by the data that exits buffer A at time t. Since we need to keep track of the mutual dependencies of the data streams that are induced through the sharing of the buffer, this quantity plays a key role in the subsequent development. For any t, the total cumulative input to buffer A at time t must equal the amount of data that has already been processed plus the amount of data currently in the buffer. The total amount of data that has exited the buffer at any time is equal to the amount of data that entered before the data currently leaving. Since this data had a delay $D(t)$, the amount of data already processed at time t is $\Phi(t-D(t))$. This implies the equation

$$\Phi(t) = \Phi(t - D(t)) + Q_A(t).$$

However, it turns out that this equation does not uniquely characterize $D(t)$ for certain t, and so we use the following definition:

$$D(t) = \inf\{\delta \geq 0\colon \Phi(t) = \Phi(t - \delta) + Q_A(t)\}. \tag{2.4}$$

Simple manipulations give the following explicit relation:

$$D(t) = t - \Phi^{-1}(\Phi(t) - Q_A(t)) \wedge t,$$

where for the continuous nondecreasing function ϕ, ϕ^{-1} is defined by

$$\phi^{-1}(s) = \sup\{t\colon \phi(t) = s\}.$$

Since such a function ϕ^{-1} is always right continuous, the continuity of Φ and Q_A imply that D is always right continuous as well. Moreover, if $Q_A(t) = 0$, then $D(t) = 0$, as one would expect. A last fact we will need is the following implicit relation:

$$Q_A(t - D(t)) = c_A D(t). \tag{2.5}$$

This equation asserts that if $D(t)$ is the delay associated with data that exits at t, then the buffer size at the time this data entered must be $c_A D(t)$. If $Q_A(t) = 0$, then $D(t) = 0$ and therefore $Q_A(t - D(t)) = Q_A(t) = 0$. To prove (2.5) when $Q_A(t) > 0$, we use the fact that $Q_A(s) > 0$ for all $s \in (t - D(t), t)$. Since this implies $\dot{\Phi}(s) - \dot{Q}_A(s) = c_A$ for $s \in (t - D(t), t)$, we obtain

$$\Phi(t) - \Phi(t - D(t)) - [Q_A(t) - Q_A(t - D(t))] = c_A D(t).$$

If we use the fact that the infimum in (2.4) is achieved at $D(t)$, we get

$$\Phi(t) - \Phi(t - D(t)) - Q_A(t) = 0.$$

Subtracting the second equation from the first produces (2.5).

Now let $Y_i(t)$ be the cumulative output from buffer A at time t that originated from source i. Note that the bound on the processing rate guarantees

that $Y_i(t)$ is almost surely differentiable in t. One can in fact give an explicit formula for $\dot{Y}_i(t)$, as shown in (B.2). From the definition of $D(t)$,

$$Y_i(t) = X_i(t - D(t)).$$

In the particular feedforward network structure considered here, only some of the sources that exit buffer A enter buffer B and the remaining sources leave the network. By grouping the sources in each of these categories, one may assume without loss of generality that there are only two sources ($I = 2$), where source 1 continues on to buffer B and source 2 leaves the network. This network is shown in Figure 2.

Let $Q_B(t)$ denote the contents of buffer B at time t. Then $Q_B(t)$ satisfies

$$\dot{Q}_B(t) = \begin{cases} \dot{Y}_1(t) - c_B, & \text{if } Q_B(t) > 0, \\ [\dot{Y}_1(t) - c_B] \vee 0, & \text{if } Q_B(t) = 0. \end{cases}$$

If we define $g_B(t) = c_B t$, then in terms of the Skorokhod map we can write this as

$$Q_B = \Gamma(Y_1 - g_B).$$

Recall that our goal is to calculate the invariant probability that the buffer contents exceed certain values. We shall use large deviations techniques to analyze the asymptotic behavior of the invariant distribution of a scaled version of the process (ξ_1, ξ_2, Q_A, Q_B). As in the one-dimensional case, the particular probability we will focus on in Section 8 provides a conservative estimate for the stationary probability that data in the corresponding finite buffer model is lost. It is well known in many analogous settings that the asymptotic behaviors of these two quantities are the same (in the large deviation sense) as the buffer sizes tend to infinity. This equivalence continues to hold here as well. If desired, one could in fact introduce a model with finite buffers, calculate the large deviation properties for such a model directly, and thereby verify this fact. The only difference between the finite buffer model and the model used here is the form of the Skorokhod map used at each switch. For example, if the buffer size at switch A is scaled as na_A, then the Skorokhod map used at that switch would constrain the buffer content to the domain $[0, na_A]$, and to $[0, a_A]$ after rescaling. Since this Skorokhod map has the same continuity properties as the one we use [7], the analysis can be completed in much the same way as in the case we consider. Since the notation and arguments

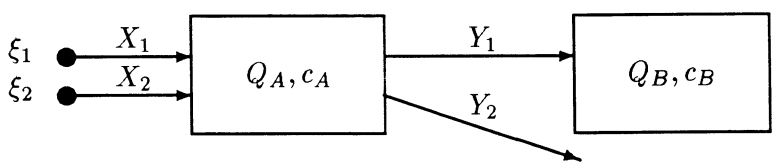

FIG. 2. *Simplified network model.*

even without these additional reflection maps are rather complicated, we will content ourselves with the treatment of the simpler case and provide a few additional remarks in Section 9.2.

The appropriate large deviation scaling for our problem is given by

$$\xi_i^n(t) = \xi_i(nt)$$

$$X_i^n(t) = \frac{1}{n} X_i(nt), \qquad Y_i^n(t) = \frac{1}{n} Y_i(nt),$$

$$Q_A^n(t) = \frac{1}{n} Q_A(nt), \qquad Q_B^n(t) = \frac{1}{n} Q_B(nt).$$

It can easily be seen that the scaled processes also satisfy the system of equations satisfied by the original processes. We remind the reader that up until now we have considered only the case of zero initial conditions: $X_1^n(0) = X_2^n(0) = Q_A^n(0) = Q_B^n(0) = Y_1^n(0) = Y_2^n(0) = 0$.

System of equations (*A*):

$$\dot{X}_i^n(t) = r_i(\xi_i^n(t)), \tag{2.6}$$

$$\Phi^n(t) = X_1^n(t) + X_2^n(t),$$

$$\dot{Q}_A^n(t) = \begin{cases} \dot{\Phi}^n(t) - c_A, & \text{if } Q_A^n(t) > 0, \\ [\dot{\Phi}^n(t) - c_A] \vee 0, & \text{if } Q_A^n(t) = 0, \end{cases} \tag{2.7}$$

$$D^n(t) = t - (\Phi^n)^{-1}(\Phi^n(t) - Q_A^n(t)) \wedge t,$$

$$Y_i^n(t) = X_i^n(t - D^n(t)),$$

$$\dot{Q}_B^n(t) = \begin{cases} \dot{Y}_1^n(t) - c_B, & \text{if } Q_B^n(t) > 0, \\ [\dot{Y}_1^n(t) - c_B] \vee 0, & \text{if } Q_B^n(t) = 0. \end{cases}$$

3. Markov model. From the system of equations (A) for the scaled process $(\xi_1^n, \xi_2^n, Q_A^n, Q_B^n)$ defined in Section 2 for zero initial conditions, we see that the evolution of the contents of buffer B at time t depends not only on the values of the process at time t, but also on the past values of $(\dot{X}_1^n, \dot{X}_2^n)$, and therefore on (ξ_1^n, ξ_2^n) during the time interval $[t - D^n(t), t]$. Thus $(\xi_1^n, \xi_2^n, Q_A^n, Q_B^n)$ is a non-Markovian process. This is an important feature of the network problem. In the case of a single queue, from (2.7) the process $(\xi_1^n, \xi_2^n, Q_A^n)$ is seen to be Markovian since the evolution of the buffer A content depends only on the current values of $(\dot{X}_1^n, \dot{X}_2^n)$, which are in turn determined by the current values of (ξ_1^n, ξ_2^n). The behavior of the single queue process is therefore much easier to analyze. In the two-queue problem the analysis is complicated by the delay experienced by the input processes (X_1^n, X_2^n) while in the buffer

A. The order in which the input streams enter buffer A becomes important since the stochastic properties or "burstiness" of an incoming stream may be significantly altered by interactions with other streams in the buffer.

In order to facilitate the analysis of the problem, we Markovianize the process $(\xi_1^n, \xi_2^n, Q_A^n, Q_B^n)$. We carry out the natural Markovianization by adding another state variable that captures all the past information required to fully determine the evolution of the distribution of the process. The new state variable will be defined in terms of a Borel measurable function f^n: $[0,\infty) \times [0,\infty) \to [0,1]$. The first variable of the function f^n will coincide with the time index t of the processes (ξ_1^n, ξ_2^n). For any given $s \in [0, Q_A^n(t)/c_A]$, we define $f^n(t,s)$ to be the fraction of the data due to source 1 *currently in the buffer* that will exit at the (scaled) time $t+s$. For $s > Q_A^n(t)/c_A$, we define $f^n(t,s) = 0$.

In Figure 3, $f^n(t,s)$ is shown as a function of s for a fixed t. An explicit representation for $f^n(t,s)$ can be given in terms of the restrictions of the functions ξ_1^n and ξ_2^n to $[0,t]$. However, we do not include the representation since it is never used. What is important for our purposes is that with the addition of this variable the state becomes Markovian and also that we can topologize the state space of this variable in a nice way.

For each $t \in [0,\infty)$, we define the functions $U_{i,t}^n(s)$ for $i = 1,2$ as the cumulative amount of data due to source i that is in the buffer at time t and exits the buffer by time $t+s$. Thus

$$U_{1,t}^n(s) = c_A \int_0^s f^n(t,u)\,du \quad \text{and} \quad U_{2,t}^n(s) = c_A \int_0^s [1 - f^n(t,u)]\,du$$

for $s \le Q_A^n(t)/c_A$, while $U_{1,t}^n(s) = U_{1,t}^n(Q_A^n(t)/c_A)$ and $U_{2,t}^n(s) = U_{2,t}^n(Q_A^n(t)/c_A)$ for $s \ge Q_A^n(t)/c_A$. $U_{i,t}^n(\cdot)$ represents all of the "history" that is needed to properly define the evolution of the two queues from time t on. Note that for $i = 1,2$, $U_{i,t}^n(s)$ has the explicit form

$$U_{i,t}^n(s) = \begin{cases} X_i^n(t+s-D^n(t+s)) - X_i^n(t-D^n(t)), & \text{if } s \le Q_A^n(t)/c_A, \\ U_{i,t}^n(Q_A^n(t)/c_A), & \text{if } s \ge Q_A^n(t)/c_A. \end{cases}$$

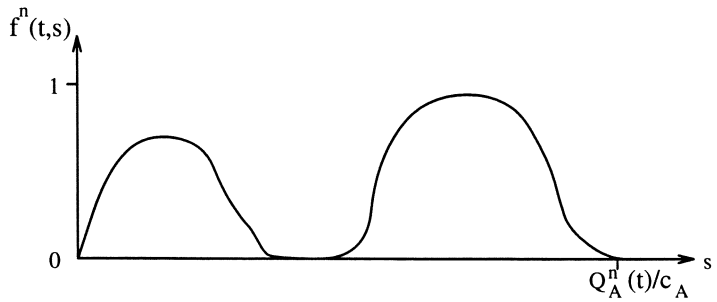

FIG. 3. *The Markovianizing function* $f^n(t,s)$.

We consider each $U^n_{i,t}$ as taking values in the space

$$\mathscr{U} = \{u \in \mathscr{I}[0,\infty)\colon u(0) = 0 \text{ and } \dot{u}(s) \in [0, c_A] \text{ a.s.}\},$$

where $\mathscr{I}[0,\infty) \subset \mathscr{C}[0,\infty)$ represents the space of absolutely continuous nondecreasing functions from $[0,\infty)$ to $[0,\infty)$. We use the standard metric

$$\gamma_C(u,v) = C\sum_{i=1}^{\infty} 2^{-i} \sup_{s\in[0,i]} |u(s) - v(s)| \tag{3.1}$$

on $\mathscr{C}[0,\infty)$, where the constant C will be chosen for our convenience in the next section. Since $\mathscr{U}$ is a closed subset of $\mathscr{C}[0,\infty)$, it is a Polish space with the inherited metric.

For $M < \infty$ we define the set $\mathscr{U}(M) = \{u \in \mathscr{U}\colon u(s) = u(M) \text{ for } s \geq M\}$. It is easy to check that $\mathscr{U}(M)$ is compact for each finite M, since it can be identified with a closed subset of the Lipschitz continuous functions on $[0, M]$ that have constant c_A. Note that the functions $U^n_{i,t}$ always take values in the space $\mathscr{U}(Q^n_A(t))$. This will prove to be quite convenient, since it implies that whenever the variables $Q^n_A(t)$ take values in a compact set K for some set of n, t or ω, the corresponding variables $U^n_{i,t}$ also take values in a compact set. This property will not always be specifically noted when used in the sequel. It will become evident later on that the variables $U^n_{i,t}$ are simply an unpleasant nuisance which must be included to Markovianize our process, but which play no significant role otherwise.

Let $\mathscr{S}_1 = \mathscr{T}_1 \times \mathscr{T}_2$, $\mathscr{S}_2 = \mathbb{R} \times \mathbb{R}$ and $\mathscr{S}_3 = \mathscr{U} \times \mathscr{U}$. Then we can define metrics such that $\mathscr{S}_1, \mathscr{S}_2, \mathscr{S}_3$ and therefore $\mathscr{S}_1 \times \mathscr{S}_2 \times \mathscr{S}_3$ are complete separable metric spaces. One can verify that the processes $Z^n = (\xi^n_1, \xi^n_2, Q^n_A, Q^n_B, U^n_1, U^n_2)$ are Markovian and take values in the set

$$\mathscr{S} = \mathscr{S}_1 \times \mathscr{S}_{2,3},$$

where

$$\mathscr{S}_{2,3} = \Big\{(q_A, q_B, u_1, u_2) \in \mathscr{S}_2 \times \mathscr{S}_3\colon (u_1, u_2) \in \mathscr{U}(q_A)^2 \text{ and } u_1(t) + u_2(t) = c_A t \text{ for } t \in [0, q_A/c_A]\Big\}.$$

If $z = (z_1, z_2, z_3) \in \mathscr{S}$, then z_1 represents the state of the modulating processes that determine the data input rates, z_2 gives the current buffer sizes, and z_3 represents a history that is compatible with these current buffer levels. The processes $\{Z^n\}$ are right continuous with limits from the left (in fact, the last four components are continuous). Since the space $\mathscr{S}$ is a closed subset of $\mathscr{S}_1 \times \mathscr{S}_2 \times \mathscr{S}_3$, it is obviously a Polish space.

4. Continuity of the mapping. In this section we will prove the continuity of the mapping that takes the scaled input processes and initial values to the scaled buffer content processes (Q^n_A, Q^n_B). The continuity of the mappings holds at the level of the sample paths of the processes. Thus, we consider the following deterministic system of equations (B).

System of equations (*B*). Let $(q_A(0), q_B(0), u_{1,0}, u_{2,0}) \in \mathscr{S}_{2,3}$ and inputs $x_i \in \mathscr{I}[0,\infty)$ with $x_i(0) = 0$ for $i = 1, 2$ be given. Define

$$\phi(t) = x_1(t) + x_2(t),$$

$$\dot{q}_A(t) = \begin{cases} \dot{\phi}(t) - c_A, & \text{if } q_A(t) > 0, \\ [\dot{\phi}(t) - c_A] \vee 0, & \text{if } q_A(t) = 0, \end{cases} \tag{4.1}$$

$$d(t) = t - \phi^{-1}(\phi(t) - q_A(t)) \wedge t, \tag{4.2}$$

$$y_i(t) = \begin{cases} u_{i,0}(t), & \text{if } t \in [0, q_A(0)/c_A], \\ x_i(t - d(t)) + u_{i,0}(q_A(0)/c_A), & \text{if } t \in [q_A(0)/c_A, \infty), \end{cases} \tag{4.3}$$

$$\dot{q}_B(t) = \begin{cases} \dot{y}_1(t) - c_B, & \text{if } q_B(t) > 0, \\ [\dot{y}_1(t) - c_B] \vee 0, & \text{if } q_B(t) = 0, \end{cases} \tag{4.4}$$

where $\phi^{-1}(s) = \sup\{t\colon \phi(t) = s\}$.

Fix $T \in [0,\infty)$, and let $\mathscr{I}[0,T]$ denote the set of all absolutely continuous increasing functions from $[0,T]$ to $[0,\infty)$. We consider both $\mathscr{I}[0,T]$ and $\mathscr{C}[0,T]$ with the usual supremum metric: $d(f,g) = \sup_{t\in[0,T]} |f(t) - g(t)|$. For the given value of T, we will select the metric on $\mathscr{U}$ to be $\gamma = \gamma_C$, with γ_C defined as in (3.1), and with C large enough that γ is bounded below by d. We consider product spaces with the sup metric, that is, $d[(f_1,f_2),(f_1',f_2')] = d(f_1,f_1') \vee d(f_2,f_2')$, and therefore $\mathbb{R}^2$ is equipped with the metric $\theta[(a_1,a_2),(b_1,b_2)] = |a_1 - b_1| \vee |a_2 - b_2|$. Finally, we consider the product space $\mathscr{Y} = \mathscr{I}^2 \times \mathscr{S}_{2,3}$, and let $z = (f_1, f_2, w_A, w_B, h_1, h_2)$ represent an element of $\mathscr{Y}$. We define the metric on $\mathscr{Y}$ by

$$\rho(z,\bar{z}) = d[(f_1,f_2),(\bar{f}_1,\bar{f}_2)] \vee \theta[(w_A,w_B),(\bar{w}_A,\bar{w}_B)] \vee \gamma[(h_1,h_2),(\bar{h}_1,\bar{h}_2)].$$

We define $\tilde{F}_T\colon \mathscr{Y} \to \mathscr{C}[0,T] \times \mathscr{C}[0,T]$ by

$$\tilde{F}_T(z) = (q_A, q_B),$$

and $F_T\colon \mathscr{Y} \to \mathbb{R}^2$ by

$$F_T(z) = (q_A(T), q_B(T)).$$

REMARK. A comparison of the system of equations (B) with the network dynamics described in Sections 2 and 3 shows that for every ω,

$$\tilde{F}_T(X_1^n, X_2^n, Q_A^n(0), Q_B^n(0), U_{1,0}^n, U_{2,0}^n)(\omega) = (Q_A^n, Q_B^n)(\omega).$$

THEOREM 4.1. *For every $T < \infty$, the mappings $\tilde{F}_T$ and F_T are Lipschitz continuous.*

PROOF. Consider $z=(x_1, x_2, q_A(0), q_B(0), u_{1,0}, u_{2,0})$ and $\bar{z}=(\bar{x}_1, \bar{x}_2, \bar{q}_A(0), \bar{q}_B(0), \bar{u}_{1,0}, \bar{u}_{2,0})$ in $\mathscr{Y}$, and let the related functions ϕ, $\bar{\phi}$, q_A, $\bar{q}_A$ and so on be defined through the system of equations (B). Let $g_A(t)$ and $g_B(t)$ denote the functions $c_A t - q_A(0)$ and $c_B t - q_B(0)$, respectively, and let $\bar{g}_A(t)$, $\bar{g}_B(t)$ be the analogous functions associated with $\bar{z}$. Suppose that $\rho(z, \bar{z}) < \delta$. Without loss of generality, we assume that $\bar{q}_A(0) \geq q_A(0)$ and denote the difference $\bar{q}_A(0) - q_A(0)$ by Δq_A. We noted after (2.1) that the Skorokhod map Γ is Lipschitz continuous with constant 2. Since $d[(x_1, x_2), (\bar{x}_1, \bar{x}_2)] < \delta$ and $\Delta q_A < \delta$, we have $d(\phi, \bar{\phi}) < 2\delta$ and $d(g_A, \bar{g}_A) < \delta$, from which it follows that $d(\phi - g_A, \bar{\phi} - \bar{g}_A) < 3\delta$. The relations $q_A = \Gamma(\phi - g_A)$ and $\bar{q}_A = \Gamma(\bar{\phi} - \bar{g}_A)$ introduced in (2.3) then imply that $d(q_A, \bar{q}_A) < 6\delta$. If we replace ζ and ψ in (2.1) by $\phi - g_A$ and q_A, respectively, and rearrange terms, then we also obtain the bound

$$d(\phi - q_A, \bar{\phi} - \bar{q}_A) < 4\delta. \tag{4.5}$$

We now establish the nearness of the output processes y_1 and $\bar{y}_1$. Recall (4.3), which describes the dynamics of y_1:

$$y_i(t) = \begin{cases} u_{i,0}(t), & \text{if } t \in [0, q_A(0)/c_A], \\ x_i(t - d(t)) + u_{i,0}(q_A(0)/c_A), & \text{if } t \in [q_A(0)/c_A, \infty). \end{cases}$$

In order to compare y_1 with $\bar{y}_1$, it is convenient to divide the time domain $[0, T]$ into three intervals: $I_1 = [0, q_A(0)/c_A]$, $I_2 = [q_A(0)/c_A, \bar{q}_A(0)/c_A]$ and $I_3 = [\bar{q}_A(0)/c_A, T]$. We can without loss of generality prove the assertion only for $T > \bar{q}_A(0)/c_A$, since the continuity for smaller T follows as a consequence. The fact that $d(u_{1,0}, \bar{u}_{1,0}) \leq \gamma(u_{1,0}, \bar{u}_{1,0}) < \delta$ immediately implies that

$$\sup_{t \in I_1} |\bar{y}_1(t) - y_1(t)| < \delta. \tag{4.6}$$

We recall that $u_{1,0}(t) = y_1(q_A(0)/c_A)$ for $t \in I_2$ and that the output rate at the first buffer is bounded above by c_A. These facts imply the second and third inequalities in

$$\begin{aligned} |\bar{y}_1(t) - y_1(t)| &\leq |\bar{y}_1(t) - y_1(q_A(0)/c_A)| + |y_1(t) - y_1(q_A(0)/c_A)| \\ &\leq |\bar{u}_{1,0}(t) - u_{1,0}(t)| + |y_1(\bar{q}_A(0)/c_A) - y_1(q_A(0)/c_A)| \\ &\leq |\bar{u}_{1,0}(t) - u_{1,0}(t)| + |\bar{q}_A(0) - q_A(0)| \\ &< 2\delta. \end{aligned} \tag{4.7}$$

Finally, when $t \in I_3$,

$$|\bar{y}_1(t) - y_1(t)| \leq 2\delta + |\bar{x}_1(t - \bar{d}(t)) - x_1(t - d(t))|. \tag{4.8}$$

Now since

$$\phi(t - d(t)) = \phi(t) - q_A(t) \quad \text{and} \quad \bar{\phi}(t - \bar{d}(t)) = \bar{\phi}(t) - \bar{q}_A(t), \tag{4.9}$$

from (4.5) we deduce that

$$d(\phi(t - d(t)), \bar{\phi}(t - \bar{d}(t))) < 4\delta. \tag{4.10}$$

For any given $t \in I_3$, let $s^* = t - d(t) \in [0, T]$. If $\phi(s^*) \geq 6\delta$, we choose $v_1 < s^*$ such that $\phi(v_1) = \phi(s^*) - 6\delta$, while if $\phi(s^*) < 6\delta$ we take $v_1 = 0$. Similarly, if $\phi(s^*) \leq \phi(T) - 6\delta$ we choose $v_2 > s^*$ such that $\phi(v_2) = \phi(s^*) + 6\delta$, while if $\phi(s^*) > \phi(T) - 6\delta$ we choose $v_2 = T$. We claim that the estimate in (4.10) together with the fact that $d(\phi, \bar{\phi}) < 2\delta$ imply $t - \bar{d}(t) \in [v_1, v_2]$. We first establish that $t - \bar{d}(t) \geq v_1$. Since $t - \bar{d}(t) \geq 0$ for $t \in I_3$, there is nothing to show if $v_1 = 0$. If $v_1 > 0$ and $t - \bar{d}(t) < v_1$, then (4.9) and the fact that $\bar{\phi}(\cdot)$ is nondecreasing imply

$$\bar{\phi}(t) - \bar{q}_A(t) = \bar{\phi}(t - \bar{d}(t)) \leq \bar{\phi}(v_1) < \phi(v_1) + 2\delta = \phi(s^*) - 4\delta = \phi(t) - q_A(t) - 4\delta,$$

which contradicts (4.10). A similar argument confirms that $t - \bar{d}(t) \leq v_2$.

We now use the fact that both x_1 and x_2 are nondecreasing, that they sum to ϕ and the definitions of v_1 and v_2 in terms of s^* to deduce that

$$x_1(v_1) \geq x_1(s^*) - 6\delta \quad \text{and} \quad x_1(v_2) \leq x_1(s^*) + 6\delta.$$

Since $t - \bar{d}(t) \in [v_1, v_2]$,

$$|x_1(t - \bar{d}(t)) - x_1(s^*)| \leq 6\delta.$$

Together with $s^* = t - d(t)$ and the bound $d(x_1, \bar{x}_1) < \delta$, this shows that $|\bar{x}_1(t - \bar{d}(t)) - x_1(t - d(t))| < 7\delta$, which when substituted into (4.8) yields

$$\sup_{t \in I_3} |\bar{y}_1(t) - y_1(t)| < 9\delta. \tag{4.11}$$

Combining (4.6), (4.7) and (4.11), we obtain $d(\bar{y}_1, y_1) < 9\delta$. Lastly, we use the representations $q_B = \Gamma(y_1 - g_B)$ and $\bar{q}_B = \Gamma(\bar{y}_1 - \bar{g}_B)$ of the processes in terms of the Skorokhod map. The Lipschitz property of Γ and the fact that $d(\bar{y}_1 - \bar{g}_B, y_1 - g_B) < 10\delta$ then imply that $d(q_B, \bar{q}_B) < 20\delta$.

We have shown that $\rho(z, \bar{z}) < \delta$ implies $d(q_A, \bar{q}_A) < 6\delta$ and $d(q_B, \bar{q}_B) < 20\delta$. Sending $\delta \downarrow \rho(z, \bar{z})$, we obtain

$$d((q_A, q_B), (\bar{q}_A, \bar{q}_B)) \leq 20\rho(z, \bar{z}),$$

and therefore in particular $d((q_A(T), q_B(T)), (\bar{q}_A(T), \bar{q}_B(T))) \leq 20\rho(z, \bar{z})$, which proves the theorem. □

REMARK. The proof incidentally establishes the Lipschitz property of the mapping $(x_1, x_2) \to y_1$, and by symmetry it also shows that $(x_1, x_2) \to y_2$ is Lipschitz continuous. The reader can verify that the assumption of two sources is unimportant, and in fact the result can be extended to include any finite number of sources $x_1, \ldots, x_k$ and resulting outputs $y_1, \ldots, y_k$. Thus the theorem easily extends to much more complicated networks, as long as there is no "feedback." For such a network it shows joint continuity for the mapping from the collection of inputs and initial conditions to the collection of buffer contents. Further comments along these lines are provided in Section 9.3. Finally, we note that we have proved more than we actually need, since all that is used in the sequel is continuity, rather than Lipschitz continuity.

5. LDP for the buffer content processes. In this section we introduce a few definitions to characterize the large deviation properties of the input processes and then prove the large deviation properties of the buffer content processes that are necessary for the analysis of the model.

DEFINITION 5.1 (Laplace principle). Let a function I from S into $[0,\infty]$ with compact level sets be given. A sequence of measures $\{\mu^n,\ n\in\mathbb{N}\}$ defined on a Polish space S is said to satisfy the Laplace principle with rate function I if for every bounded continuous function F on S,

$$\lim_{n\to\infty} -\frac{1}{n}\log\left[\int_S e^{-nF(x)}\mu^n(dx)\right] = \inf_{x\in S}[F(x)+I(x)].$$

DEFINITION 5.2 (Large deviation principle). Let a function I from S into $[0,\infty]$ with compact level sets be given. A sequence of measures $\{\mu^n,\ n\in\mathbb{N}\}$ defined on a Polish space S is said to satisfy the large deviation principle with rate function I if the following hold:

(i) For each closed set $C\subset S$,

$$\limsup_{n\to\infty}\frac{1}{n}\log\mu^n(C) \le -\inf_{\phi\in C} I(\phi).$$

(ii) For each open set $G\subset S$,

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

REMARK. A sequence of measures satisfies the Laplace principle with some rate function I if and only if it satisfies the large deviation principle with the same rate function (see [6], Theorem 1.2.3).

REMARK. A sequence of random variables $\{Z^n,\ n\in\mathbb{N}\}$ defined on a probability space $(\Omega,\mathscr{F},P)$ and taking values in a Polish space is said to satisfy the large deviation principle or Laplace principle with rate function I if the corresponding sequence of distributions $\{\mu^n,\ n\in\mathbb{N}\}$ defined by $\mu^n(dz)=P\{Z^n\in dz\}$ satisfies the large deviation principle or Laplace principle with that rate function.

Recall π_i, the invariant distribution of ξ_i introduced in Section 2, and the Polish space $\mathscr{S}$, defined at the end of Section 3. We continue to use the notation of Section 3, so that an element of $\mathscr{S}$ is written in the form (z_1,z_2,z_3) and we let P_{z_1} denote probability conditioned on $(\xi_1^n(0),\xi_2^n(0))=z_1$.

LEMMA 5.3. *Fix $T\in(0,\infty)$. Define the rate function I_T on $\mathscr{C}[0,T]^2$ by $I_T(x_1,x_2)=I_T^1(x_1)+I_T^2(x_2)$, where*

$$I_T^i(\phi)=\int_0^T L_i(\dot\phi)\,dt$$

if ϕ is absolutely continuous with $\phi(0) = 0$, and ∞ otherwise. The functions $L_i(u)$ are convex, nonnegative, have compact level sets, and $L_i(u) = 0$ if and only if $u = b_i$, where

$$b_i = \sum_{j \in \mathscr{F}_i} r_i(j)\pi_i(j).$$

The processes $\{(X_1^n, X_2^n),\ n \in \mathbb{N}\}$ satisfy a uniform Laplace principle on $\mathscr{C}[0,T]^2$, in the sense that for any bounded, continuous function H on $\mathscr{C}[0,T]^2$,

$$\begin{aligned} &-\lim_{n\to\infty} \frac{1}{n} \log \sup_{z_1 \in \mathscr{S}_1} E_{z_1}[\exp(-nH(X_1^n, X_2^n))] \\ &\quad = -\lim_{n\to\infty} \frac{1}{n} \log \inf_{z_1 \in \mathscr{S}_1} E_{z_1}[\exp(-nH(X_1^n, X_2^n))] \\ &\quad = \inf_{(x_1,x_2)\in\mathscr{C}[0,T]^2} [H(x_1,x_2) + I_T(x_1,x_2)]. \end{aligned}$$

This result is quite standard, and can be found in a number of places, including [12]. It can also be derived easily from the well-known result on the large deviation principle for the occupation measure of ξ_i due to Donsker and Varadhan [4, 6, 23].

We can now derive a type of uniform Laplace principle for the buffer content processes $(q_A(T), q_B(T))$. We recall the continuous map F_t defined at the beginning of Section 4. (Recall that θ is the metric on $\mathscr{S}_2 = \mathbb{R}^2$, and that $z_2 \in \mathbb{R}^2$ gives the initial level of the two buffers. If $(z_2, z_3) \in \mathscr{S}_{2,3}$, then z_3 represents a "history" that is consistent with these buffer levels.) For $\varepsilon > 0$, let $\bar{N}_\varepsilon = \{z_2 \in \mathscr{S}_2 \colon \theta(z_2, 0) \le \varepsilon\}$ and let the compact set $K'_\varepsilon = \mathscr{S}_1 \times \{(z_2, z_3) \in \mathscr{S}_{2,3} \colon z_2 \in \bar{N}_\varepsilon\}$. Furthermore, for any $t \in (0,\infty)$, we define G_t by

$$G_t(\eta) = \inf_{(x_1,x_2)\colon F_t(x_1,x_2,0,0,0,0)=\eta} \int_0^t [L_1(\dot{x}_1) + L_2(\dot{x}_2)]\, ds \tag{5.1}$$

if the set over which the infimum is taken is nonempty and $G_t(\eta) = \infty$ otherwise. (Thus G_t is defined only for the case of zero initial conditions in the z_2 and z_3 components.)

THEOREM 5.4. *Given any $0 < T_1 \le T_2 < \infty$, the sequence $\{(Q_A^n(t), Q_B^n(t)),\ n \in \mathbb{N}\}$ satisfies a uniform Laplace principle in the sense that given any bounded continuous function g on $\mathscr{S}_2$,*

$$\begin{aligned} &-\lim_{\varepsilon\downarrow 0} \limsup_{n\to\infty} \frac{1}{n} \log \sup_{z\in K'_\varepsilon} E_z[\exp(-ng(Q_A^n(t), Q_B^n(t)))] \\ &\quad = -\lim_{\varepsilon\downarrow 0} \liminf_{n\to\infty} \frac{1}{n} \log \inf_{z\in K'_\varepsilon} E_z[\exp(-ng(Q_A^n(t), Q_B^n(t)))] \\ &\quad = \inf_{\eta\in\mathscr{S}_2} [g(\eta) + G_t(\eta)] \end{aligned} \tag{5.2}$$

uniformly in $t \in [T_1, T_2]$, where $G_t(\eta)$ is defined as in (5.1).

PROOF. This result is an immediate consequence of Lemma 5.3. From the definition of G_t and the fact that I_t is a rate function, it follows directly that for each $t \in [0, \infty)$, G_t possesses compact level sets and is nonnegative. To prove the uniform limit asserted in the theorem, we let g be a bounded continuous function on $\mathscr{S}_2$, and note that $g \circ F_t$ defines a bounded continuous function on $C[0, t]^2 \times \mathscr{S}_{2,3}$. Owing to the continuity of F_t proved in Theorem 4.1, Theorem 5.4 holds if and only if (5.2) is satisfied when K'_ε is replaced by $(z_1, 0, 0)$ (i.e., zero initial conditions for the two buffers). However, by Lemma 5.3, for each $t \in [T_1, T_2]$ we have the following limit:

$$
\begin{aligned}
&- \lim_{n\to\infty} \frac{1}{n} \log \sup_{z_1 \in \mathscr{S}_1} E_{(z_1, 0, 0)}[\exp(-ng(Q_A^n(t), Q_B^n(t)))] \\
&\quad = - \lim_{n\to\infty} \frac{1}{n} \log \inf_{z_1 \in \mathscr{S}_1} E_{(z_1, 0, 0)}[\exp(-ng(Q_A^n(t), Q_B^n(t)))] \\
&\quad = \inf_{(x_1, x_2) \in C[0, t]^2} [g(F_t(x_1, x_2, 0, 0, 0, 0)) + I_t(x_1, x_2)].
\end{aligned}
$$

The Lipschitz continuity of the sample paths of $(Q_A(t), Q_B(t))$ and the continuity of g establish that the families of functions in the first two terms above are equicontinuous in t. As a consequence, the third term is seen to be continuous in t and the convergence holds uniformly in $t \in [T_1, T_2]$. Since

$$
\begin{aligned}
&\inf_{(x_1, x_2) \in C[0, t]^2} [g(F_t(x_1, x_2, 0, 0, 0, 0)) + I_t(x_1, x_2)] \\
&\quad = \inf_{\eta \in \mathscr{S}_2} \Big[g(\eta) + \inf_{(x_1, x_2):\, F_t(x_1, x_2, 0, 0, 0, 0) = \eta} I_t(x_1, x_2) \Big],
\end{aligned}
$$

the theorem is proved. □

6. LDP for the invariant distribution. In this section the (Q_A^n, Q_B^n)-marginal of the invariant distribution of the Markov process $Z^n = (\xi_1^n, \xi_2^n, Q_A^n, Q_B^n, U_1^n, U_2^n)$ is shown to satisfy a large deviation principle. For notational convenience, let $Z_1^n = (\xi_1^n, \xi_2^n)$, $Z_2^n = (Q_A^n, Q_B^n)$, $Z_3^n = (U_1^n, U_2^n)$. The three-component Markov process $Z^n = (Z_1^n, Z_2^n, Z_3^n)$ takes values in $\mathscr{S}$, where the description of the Polish space $\mathscr{S}$ was given at the end of Section 3. We will show in Section 7 that under stability conditions this process possesses an invariant distribution which we denote by μ^n. The Z_2^n-marginal of μ^n is denoted by μ_2^n. In this section we will formulate conditions under which $\{\mu_2^n\}$ satisfies a large deviation principle. Since only a certain marginal of the invariant distribution is shown to satisfy the large deviation principle, we need a number of definitions that explicitly identify the uniformity of various estimates with respect to variations in the other components.

DEFINITION 6.1 (Exponential tightness). A sequence of measures $\{\mu^n,\ n \in \mathbb{N}\}$ on a Polish space S is said to be exponentially tight if, given any $M < \infty$, there exists a compact set $K \subset S$ such that

$$
\limsup_{n\to\infty} \frac{1}{n} \log \mu^n(K^c) \le -M.
$$

A useful tool in deriving the Laplace principle (and hence the large deviation principle) for the invariant distributions is the following analogue of Prohorov's theorem. The theorem appears to be independently due to Puhalskii [20] and O'Brien and Vervaat [18]. For a proof, see [6], Theorem 1.3.7.

THEOREM 6.2. *If a sequence of measures* $\{\mu^n,\ n \in \mathbb{N}\}$ *on a Polish space* S *is exponentially tight, then there exists a subsequence which satisfies the Laplace principle with some rate function.*

We next define a type of stability that will be needed. As in the last section, $\bar{N}_\varepsilon$ denotes the closed neighborhood of radius ε about the origin: $\bar{N}_\varepsilon = \{z_2 \in \mathscr{S}_2\colon \theta(z_2, 0) \le \varepsilon\}$. Let $\tau_\varepsilon^n = \inf\{t \ge 0\colon Z_2^n(t) \in \bar{N}_\varepsilon\}$ and let P_z denote probability conditioned on $Z^n(0) = z$.

DEFINITION 6.3 (Uniform exponential attraction in probability). Consider the sequence of Markov processes $\{Z^n,\ n \in \mathbb{N}\}$: $\{Z_2^n,\ n \in \mathbb{N}\}$ is said to be uniformly exponentially attracted in probability to the origin if, given any $M < \infty$ and compact set $K \subset \mathscr{S}$, there exists $T < \infty$ such that for any $\varepsilon > 0$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \sup_{z\in K} P_z\{\tau_\varepsilon^n > T\} \le -M.$$

We now describe the conditions that the sequence of Markov processes $\{Z^n,\ n \in \mathbb{N}\}$ must satisfy in order for the main theorem of this section to hold. In Section 5, part (c) of this condition was shown to hold for our process. Parts (a) and (b) of the condition will be verified in Section 7, with details of the proof provided in Appendix A.

CONDITION 6.1. (a) $\{Z^n(t),\ n \in \mathbb{N}\}$ is exponentially tight uniformly with respect to initial conditions in compact sets and $t \in [0, \infty)$. In other words, given any $M < \infty$ and any compact set $C \subset \mathscr{S}$, there exists a compact set K such that

$$\limsup_{n\to\infty} \frac{1}{n} \sup_{t\in[0,\infty),\, z\in C} P_z\{Z^n(t) \in K^c\} \le -M.$$

(b) $\{Z_2^n,\ n \in \mathbb{N}\}$ is uniformly exponentially attracted in probability to the origin.

(c) $\{Z_2^n(T),\ n \in \mathbb{N}\}$ satisfies a uniform Laplace principle in the sense of Theorem 5.4 with rate function $G_T(\eta)$.

We now state the main result of this section. The theorem allows us to directly relate the rate functions for the sequence of random variables $\{Z_2^n(T)\}$ for $T < \infty$ to the rate function for the sequence of measures $\{\mu_2^n\}$. It requires existence and uniqueness of $\{\mu^n\}$ and exponential tightness of $\{\mu_2^n\}$, all of which will be proved in the next section.

THEOREM 6.4. *Assume that* $\{Z^n\}$ *satisfies Condition* 6.1. *Moreover, suppose that for all sufficiently large* n, Z^n *possesses a unique invariant distribution* μ^n, *and that the set of second marginals* $\{\mu_2^n\}$ *is exponentially tight. Then* $\{\mu_2^n\}$ *satisfies the large deviation principle with rate function* $J(\eta)$, *where*

$$J(\eta) = \inf_{T>0} G_T(\eta). \tag{6.1}$$

PROOF. Since the sequence $\{\mu_2^n\}$ is assumed to be exponentially tight, by Theorem 6.2 there exists a subsequence that satisfies the Laplace principle. We denote this subsequence also by $\{\mu_2^n\}$ and let $J(\eta)$ be its rate function. Thus for any bounded, continuous function g and any $T > 0$,

$$\lim_{n\to\infty} -\frac{1}{n} \log E_{\mu^n}[\exp(-ng(Z_2^n(T)))] = \inf_{\eta\in S_2} [J(\eta) + g(\eta)], \tag{6.2}$$

where E_{μ^n} denotes expectation conditioned on starting from the invariant distribution.

A rough outline of the intuition behind the proof is as follows. Suppose we start the process off at time 0 with initial distribution μ^n. Using the exponential tightness, we can effectively assume that μ^n is supported on some compact set. The uniform exponential attraction to the origin then implies that there exists $\hat{T} < \infty$ such that, save on a set of negligible probability, the process enters any arbitrarily small neighborhood of the origin by time $\hat{T}$. Using the strong Markov property and the fact that these neighborhoods can be made arbitrarily small permits a representation of the distribution at time T, where T is large compared to $\hat{T}$, in terms of the rate functions G_t for $t \in [T - \hat{T}, T]$. Using the fact that G_t is monotonically nonincreasing and that the distribution at T is also the stationary distribution, one can bound the tail behavior and hence the rate function for the stationary distribution in terms of $G_{T-\hat{T}}$ and G_T. It turns out that, for any given η, $G_t(\eta)$ is constant for all sufficiently large t, and thus we obtain (6.1).

Before proceeding with the proof, we select a number of parameters. To facilitate the discussion later on, we carefully note the dependencies of each parameter. Let $\delta > 0$, the function g and $M \in (\|g\|_\infty, \infty)$ be fixed. We recall the set $K'_\varepsilon = \mathscr{S}_1 \times \{(z_2, z_3) \in \mathscr{S}_{2,3}\colon z_2 \in \bar{N}_\varepsilon\}$, corresponding to small buffers at time zero.

1. According to part (c) of Condition 6.1, we can choose $\varepsilon \in (0, 1)$ such that if $0 < T_1 \le T_2 < \infty$, then uniformly for $t \in [T_1, T_2]$ and $z \in K'_\varepsilon$,

$$\begin{aligned} \inf_{\eta\in S_2} [G_t(\eta) + g(\eta)] - \delta &\le \liminf_{n\to\infty} -\frac{1}{n} \log E_z[\exp(-ng(Z_2^n(t)))] \\ &\le \limsup_{n\to\infty} -\frac{1}{n} \log E_z[\exp(-ng(Z_2^n(t)))] \\ &\le \inf_{\eta\in S_2} [G_t(\eta) + g(\eta)] + \delta. \end{aligned} \tag{6.3}$$

The number ε depends on δ and g.

2. Consider the compact set K_1', which is obtained from the definition above for K_ε' by setting $\varepsilon = 1$. Then part (a) of Condition 6.1 implies that there exist a compact set $K_2 \subset \mathcal{S}_2$ and $N < \infty$ such that for $n \geq N$ and for every $s \in [0, \infty)$,

$$\sup_{z \in K_1'} P_z\{Z_2^n(s) \in K_2^c\} \leq e^{-3Mn}. \tag{6.4}$$

Since $\{\mu_2^n\}$ is exponentially tight, by enlarging K_2 and taking N larger if necessary, we can also assume that for all $n \geq N$,

$$\mu_2^n(K_2^c) \leq e^{-3Mn}. \tag{6.5}$$

The set K_2 depends only on M. We define the compact set $K = \mathcal{S}_1 \times \{(z_2, z_3) \in \mathcal{S}_{2,3}\colon z_2 \in K_2\}$, so that K_2 is the projection of K onto $\mathcal{S}_2$.

3. Last, we use part (b) of Condition 6.1. According to this condition, for $M < \infty$ and compact K as given above, there exists $T_0 < \infty$ such that, given $\varepsilon > 0$,

$$\limsup_{n \to \infty} \frac{1}{n} \log \sup_{z \in K} P_z\{\tau_\varepsilon^n \geq T_0\} \leq -2M. \tag{6.6}$$

The number T_0 depends on K and M.

For sufficiently large n, (6.5) implies that for any $T > 0$,

$$E_{\mu^n}[1_{\{Z_2^n(T) \in K_2^c\}} \exp(-ng(Z_2^n(T)))] \leq \exp(-3Mn)\exp(Mn) = \exp(-2Mn).$$

The last inequality, and the fact that $\|g\|_\infty \leq M$, implies

$$\begin{aligned} &\lim_{n \to \infty} -\frac{1}{n} \log E_{\mu^n}[\exp(-ng(Z_2^n(T)))] \\ &\qquad = \lim_{n \to \infty} -\frac{1}{n} \log E_{\mu^n}\big[1_{\{Z_2^n(T) \in K_2\}} \exp(-ng(Z_2^n(T)))\big]. \end{aligned} \tag{6.7}$$

For $\varepsilon > 0$, we recall the random time

$$\tau_\varepsilon^n = \inf\{t > 0\colon Z_2^n(t) \in \bar{N}_\varepsilon\} \tag{6.8}$$

and the closed neighborhood $\bar{N}_\varepsilon = \{z_2 \in \mathcal{S}_2\colon \theta(z_2, 0) \leq \varepsilon\}$. Let $\{\mathcal{F}_t^n\}$ be the filtration generated by the process Z^n and let $\mathcal{F}_t^n \subset \mathcal{F}^n$ for every $t \geq 0$. According to Proposition 2.1.5 of [11], for every $n \in \mathbb{N}$ and $\varepsilon > 0$, τ_ε^n is a stopping time. Recall the standard definition of $\mathcal{F}_{\tau_\varepsilon^n}^n$ as the σ-algebra that contains all the information that an observer of the process Z^n knows at time τ_ε^n:

$$\mathcal{F}_{\tau_\varepsilon^n}^n = \{A \in \mathcal{F}^n\colon A \cap \{\tau_\varepsilon^n \leq t\} \in \mathcal{F}_t^n \text{ for all } t \geq 0\}.$$

It follows from the definition of the compact set K given below (6.5) that whenever Z_2^n is in the set K_2, Z^n must lie in the set K. Thus for all sufficiently large n,

$$\mu^n(K^c) \leq e^{-3Mn}. \tag{6.9}$$

Together, (6.9) and (6.6) imply that for any $\hat{T} \geq T_0$,

$$\begin{aligned} &\lim_{n\to\infty} -\frac{1}{n}\log E_{\mu^n}\bigl[1_{\{Z_2^n(T)\in K_2\}}\exp(-ng(Z_2^n(T)))\bigr] \\ &\qquad = \lim_{n\to\infty} -\frac{1}{n}\log E_{\mu^n}\bigl[1_{\{\tau_\varepsilon^n\leq\hat{T}\}}1_{\{Z_2^n(T)\in K_2\}}\exp(-ng(Z_2^n(T)))\bigr]. \end{aligned} \tag{6.10}$$

Suppose $T > \hat{T} \geq T_0$. From the properties of conditional expectation and the strong Markov property for Z^n,

$$\begin{aligned} &E_{\mu^n}\bigl[1_{\{\tau_\varepsilon^n\leq\hat{T}\}}1_{\{Z_2^n(T)\in K_2\}}\exp(-ng(Z_2^n(T)))\bigr] \\ &\qquad = E_{\mu^n}\bigl[1_{\{\tau_\varepsilon^n\leq\hat{T}\}}E_{\mu^n}\bigl[1_{\{Z_2^n(T)\in K_2\}}\exp(-ng(Z_2^n(T)))|\mathscr{F}_{\tau_\varepsilon^n}^n\bigr]\bigr] \\ &\qquad = E_{\mu^n}\bigl[1_{\{\tau_\varepsilon^n\leq\hat{T}\}}E_{Z^n(\tau_\varepsilon^n)}\bigl[1_{\{Z_2^n(T-\tau_\varepsilon^n)\in K_2\}}\exp(-ng(Z_2^n(T-\tau_\varepsilon^n)))\bigr]\bigr]. \end{aligned} \tag{6.11}$$

Recall the compact set $K'_\varepsilon = \mathscr{S}_1 \times \{(z_2, z_3) \in \mathscr{S}_{2,3}\colon z_2 \in \bar{N}_\varepsilon\}$. We define a measure on $[0,\infty)\times K'_\varepsilon$ by

$$\nu_\varepsilon^n(ds\times dz) = P_{\mu^n}\{\tau_\varepsilon^n \in ds \text{ and } Z^n(\tau_\varepsilon^n)\in dz\}.$$

With this definition (6.11) can be rewritten as

$$E_{\mu^n}\int_0^{\hat{T}}\int_{K'_\varepsilon} E_z[1_{\{Z_2^n(T-s)\in K_2\}}\exp(-ng(Z_2^n(T-s)))]\nu_\varepsilon^n(ds\times dz). \tag{6.12}$$

Following the chain of equalities in (6.7), (6.10), (6.11) and (6.12) yields

$$\begin{aligned} &\lim_{n\to\infty} -\frac{1}{n}\log E_{\mu^n}[\exp(-ng(Z_2^n(T)))] \\ &= \lim_{n\to\infty} -\frac{1}{n} \\ &\qquad \times \log E_{\mu^n}\int_0^{\hat{T}}\int_{K'_\varepsilon} E_z\bigl[1_{\{Z_2^n(T-s)\in K_2\}}\exp(-ng(Z_2^n(T-s)))\bigr]\nu_\varepsilon^n(ds\times dz). \end{aligned} \tag{6.13}$$

We next use the fact that (6.4), $K'_\varepsilon \subset K'_1$ and $\|g\|_\infty \leq M$ imply that for all $z \in K'_\varepsilon$ and $s \in [0,\hat{T}]$,

$$E_z\bigl[1_{\{Z_2^n(T-s)\in K_2^c\}}\exp(-ng(Z_2^n(T-s)))\bigr] \leq \exp(-2Mn).$$

This leads to the equality

$$\begin{aligned} &\lim_{n\to\infty} -\frac{1}{n}\log E_{\mu^n}\int_0^{\hat{T}}\int_{K'_\varepsilon} E_z\bigl[1_{\{Z_2^n(T-s)\in K_2\}}\exp(-ng(Z_2^n(T-s)))\bigr]\nu_\varepsilon^n(ds\times dz) \\ &\qquad = \lim_{n\to\infty} -\frac{1}{n}\log E_{\mu^n}\int_0^{\hat{T}}\int_{K'_\varepsilon} E_z[\exp(-ng(Z_2^n(T-s)))]\nu_\varepsilon^n(ds\times dz). \end{aligned}$$

Combining the last display with (6.13) we find that, for $T > \hat{T}$,

$$\begin{aligned} &\lim_{n\to\infty} -\frac{1}{n}\log E_{\mu^n}[\exp(-ng(Z_2^n(T)))] \\ &\qquad = \lim_{n\to\infty} -\frac{1}{n}\log E_{\mu^n}\int_0^{\hat{T}}\int_{K'_\varepsilon} E_z[\exp(-ng(Z_2^n(T-s)))]\nu_\varepsilon^n(ds\times dz). \end{aligned} \tag{6.14}$$

The fact that $G_t(\eta)$ is monotonically nonincreasing in t implies that for $T > \hat{T}$, $\eta \in \mathscr{S}_2$, and any $s \in [0, \hat{T}]$,

$$G_T(\eta) \le G_{T-s}(\eta) \le G_{T-\hat{T}}(\eta).$$

Now fix $T_1 > 0$ and $T \ge \hat{T} + T_1$. It follows from the last display and (6.3) that, given $\gamma > 0$, there exists $N < \infty$ such that for all $n \ge N$, $s \in [0, \hat{T}]$ and $z \in K'_\varepsilon$,

$$\begin{aligned} \inf_{\eta \in S_2} [G_T(\eta) + g(\eta)] - \delta - \gamma &\le -\frac{1}{n} \log E_z[\exp(-ng(Z_2^n(T-s)))] \\ &\le \inf_{\eta \in S_2} [G_{T-\hat{T}}(\eta) + g(\eta)] + \delta + \gamma. \end{aligned} \tag{6.15}$$

Then (6.15) and the fact that

$$1 - e^{-nM} \le \nu_\varepsilon^n([0, \hat{T}] \times K'_\varepsilon) \le 1$$

imply

$$\begin{aligned} &\inf_{\eta \in S_2} [G_T(\eta) + g(\eta)] - \delta - \gamma \\ &\quad \le \lim_{n \to \infty} -\frac{1}{n} \log E_{\mu^n} \int_0^{\hat{T}} \int_{K'_\varepsilon} E_z[\exp(-ng(Z_2^n(T-s)))] \nu_\varepsilon^n(ds \times dz) \\ &\quad \le \inf_{\eta \in S_2} [G_{T-\hat{T}}(\eta) + g(\eta)] + \delta + \gamma. \end{aligned}$$

Since $\gamma > 0$ is arbitrary, we may let $\gamma \downarrow 0$ and use (6.14) to obtain

$$\begin{aligned} \inf_{\eta \in S_2} [G_T(\eta) + g(\eta)] - \delta &\le \lim_{n \to \infty} -\frac{1}{n} \log E_{\mu_n}[\exp(-ng(Z_2^n(T)))] \\ &\le \inf_{\eta \in S_2} [G_{T-\hat{T}}(\eta) + g(\eta)] + \delta. \end{aligned}$$

This inequality can be rewritten using the fact that $\{\mu^n,\ n \in \mathbb{N}\}$ satisfies the Laplace principle with rate function $J(\eta)$, as in (6.2), thereby obtaining

$$\begin{aligned} \inf_{\eta \in S_2} [G_T(\eta) + g(\eta)] - \delta &\le \inf_{\eta \in S_2} [J(\eta) + g(\eta)] \\ &\le \inf_{\eta \in S_2} [G_{T-\hat{T}}(\eta) + g(\eta)] + \delta. \end{aligned} \tag{6.16}$$

Thus, given any $\delta > 0$, we have shown that there exists $T_0 < \infty$ such that the inequality above is satisfied for all bounded continuous functions g and for $T > \hat{T} \ge T_0$. Note that $\hat{T}$ depends on g through (6.6) only in the sense that we require $\|g\|_\infty \le M$. Fix $\bar{\eta} \in \mathscr{S}_2$, and consider a sequence $\{g_k,\ k \in \mathbb{N}\}$ of bounded, nonnegative, continuous functions with $\|g_k\|_\infty \le M$ and $g_k(\bar{\eta}) = 0$ that converge uniformly on each closed subset of $\mathscr{S}_2 \backslash \{\bar{\eta}\}$ to the function

$$g(\eta) = \begin{cases} 0, & \text{if } \eta = \bar{\eta}, \\ M, & \text{otherwise.} \end{cases}$$

Since each of the functions $G_{T-\hat{T}}$, J, and G_T is a rate function, each of them has compact level sets. As a consequence, upon taking limits in (6.16) with g replaced by g_k we obtain

$$[G_T(\bar{\eta}) \wedge M] - \delta \leq J(\bar{\eta}) \wedge M \leq [G_{T-\hat{T}}(\bar{\eta}) \wedge M] + \delta. \tag{6.17}$$

Since $\bar{\eta}$ is arbitrary, we can assume that (6.17) holds for all $\eta \in \mathcal{S}_2$ and $T > \hat{T} \geq T_0$. Because T_0 depends on M but not on δ, we can let $\delta \to 0$ in the last display. Let $i \in \{2, 3, \ldots\}$. Choosing $T = i\hat{T}$ then shows that

$$G_{i\hat{T}}(\eta) \wedge M \leq J(\eta) \wedge M \leq G_{(i-1)\hat{T}}(\eta) \wedge M. \tag{6.18}$$

Since $G_t(\eta)$ is nonincreasing in t, letting $i \to \infty$ in (6.18) gives

$$\inf_{T>0} G_T(\eta) \wedge M \leq J(\eta) \wedge M \leq \inf_{T>0} G_T(\eta) \wedge M.$$

Observing that the inequalities above are independent of T_0, we can let $M \uparrow \infty$ to obtain

$$J(\eta) = \inf_{T>0} G_T(\eta).$$

Thus we have proved the desired result (6.1) for the rate function of the subsequence $\{\mu^n, n \in \mathbb{N}\}$. In order to show that the result holds for the original sequence, one can use the usual argument by contradiction. □

7. Stability properties. In this section we verify parts (a) and (b) of Condition 6.1 for the sequence of Markov processes $\{Z^n = (\xi_1^n, \xi_2^n, Q_A^n, Q_B^n, U_1^n, U_2^n)\}$. Under natural assumptions, in Theorem 7.1 we first establish the stability of the associated "fluid model" by constructing an appropriate Lyapunov function. This Lyapunov function is then used in Theorem 7.2 and Corollary 7.4 to establish the required stability properties and exponential tightness of the processes. A number of proofs are deferred to Appendix A. In Lemma 7.5 we verify the other assumptions that are needed for Theorem 6.4, namely, the existence and uniqueness of the invariant distribution μ^n (for all large n) and the exponential tightness of the (Q_A^n, Q_B^n)-marginals $\{\mu_2^n\}$. The result one thereby obtains is summarized in Theorem 7.6.

THEOREM 7.1. *Consider the system of equations* (*B*) *in the special case when* $\dot{x}_1 = b_1$ *and* $\dot{x}_2 = b_2$. *Then* $(0, 0)$ *is a stable point for* (q_A, q_B) *if* $b_1 + b_2 < c_A$ *and* $b_1 < c_B$. *In other words, given any initial condition* $(q_A(0), q_B(0)) = (z_a, z_b)$, *there exists* $T < \infty$ *such that* $(q_A(t), q_B(t)) = (0, 0)$ *for every* $t \geq T$.

The proof is given in Appendix A.

The next theorem states that the stochastic processes (Q_A^n, Q_B^n) are uniformly exponentially attracted to the origin (τ_ε^n, $\bar{N}_\varepsilon$ and P_z are all defined as in Section 6).

THEOREM 7.2. *If $b_1 + b_2 < c_A$ and $b_1 < c_B$, then the sequence $\{(Q_A^n, Q_B^n)\}$ defined by the system of equations* (A) *is uniformly exponentially attracted to the origin. In other words, given any $M < \infty$ and compact set $K \subset \mathscr{S}$, there exists $T < \infty$ such that, for all $\varepsilon > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \sup_{z\in K} P_z\{\tau_\varepsilon^n > T\} \le -M.$$

The proof is given in Appendix A.

The following theorem establishes the uniform exponential tightness of the family of random variables $Z^n(t)$. The exponential tightness of the measures then follows from the uniform exponential tightness of these random variables.

THEOREM 7.3. *The family of random variables $\{(Q_A^n(t), Q_B^n(t)),\ n \in \mathbb{N}\}$ is exponentially tight uniformly with respect to $t \in [0, \infty)$ and initial conditions in compact sets. In other words, given any compact set $D \subset \mathscr{S}$ and $M < \infty$, there exists a compact set $C \subset \mathscr{S}_2$ and $N < \infty$ such that for every $n \ge N$,*

$$\sup_{z\in D,\, t\in[0,\infty)} P_z\{(Q_A^n(t), Q_B^n(t)) \notin C\} \le e^{-nM}.$$

The proof is given in Appendix A.

COROLLARY 7.4. *The family of random variables*

$$\{(\xi_1^n(t), \xi_2^n(t), Q_A^n(t), Q_B^n(t), U_{1,t}^n, U_{2,t}^n),\ n \in \mathbb{N}\}$$

is exponentially tight uniformly for $t \in [0, \infty)$ and initial conditions in compact sets.

PROOF. By Assumption 2.1, $\{(\xi_1^n(t), \xi_2^n(t)),\ t \in [0, \infty),\ n \in \mathbb{N}\}$ live on a finite state space $\mathscr{F}_1 \times \mathscr{F}_2$ and are thus trivially exponentially tight. The set $\{(Q_A^n(t), Q_B^n(t)),\ n \in \mathbb{N}\}$ for $t \in [0, \infty)$ was shown to be exponentially tight uniformly with respect to initial conditions in Theorem 7.3. Recall that if $Q_A^n(t)$ takes values in a compact set for some collection of n, ω and t, then $(U_{1,t}^n, U_{2,t}^n)$ also takes values in a compact set for the same collection. Thus the uniform exponential tightness of $\{(Q_A^n(t), Q_B^n(t)),\ n \in \mathbb{N}\}$ automatically implies the uniform exponential tightness of $\{U_{1,t}^n, U_{2,t}^n,\ n \in \mathbb{N}\}$ and the corollary is established. □

We now prove the remaining assumptions of Theorem 6.4 for the process $\{Z^n\}$.

LEMMA 7.5. *For all sufficiently large n, the process $(\xi_1^n, \xi_2^n, Q_A^n, Q_B^n, U_1^n, U_2^n)$ possesses a unique invariant measure. In addition, the sequence of (Q_A^n, Q_B^n)-marginals $\{\mu_2^n\}$ is exponentially tight.*

PROOF. A (much simpler) version of Theorem 7.3 which proves the uniform exponential tightness of the processes can be used to show that for all large n and any given initial condition the random variables $\{(\xi_1^n(t), \xi_2^n(t), Q_A^n(t), Q_B^n(t), U_{1,t}^n, U_{2,t}^n),\ t \in [0,\infty)\}$ are tight. Following a standard argument (e.g., [11], page 247), this implies the existence of an invariant distribution. For each fixed n, it is easy to show that given any compact set of initial conditions, one can find $T < \infty$ such that the probability of hitting the origin [i.e., $(Q_A^n(t), Q_B^n(t), U_{1,t}^n, U_{2,t}^n) = (0,0,0,0)$] before time T is greater than zero uniformly in the initial condition. Since the sets $\mathscr{F}_1$ and $\mathscr{F}_2$ are finite, this shows that given any compact set of initial conditions there is a point z and $T < \infty$ such that the probability that $(\xi_1^n(t), \xi_2^n(t), Q_A^n(t), Q_B^n(t), U_{1,t}^n, U_{2,t}^n) = z$ for some $t \le T$ is strictly positive, uniformly for initial conditions in the compact set. Thus one can use the standard argument (e.g., [2]) to establish the uniqueness of the invariant distribution.

Finally we consider the exponential tightness of the marginals $\{\mu_2^n\}$. However, this follows directly from the exponential tightness of the random variables proved in Theorem 7.3 and the fact that the invariant distribution is equal to the a.s. limit of the normalized occupation measures of the process. □

We can now state the main result of the section.

THEOREM 7.6. *Consider the model defined in Sections* 2 *and* 3. *The sequence* $\{\mu_2^n\}$ *of* (Q_A^n, Q_B^n)*-marginals of the invariant distribution* $\{\mu^n\}$ *satisfies the large deviation principle with rate function*

$$J(\eta) = \inf\{T > 0\colon G_T(\eta)\}.$$

PROOF. The process $\{Z^n,\ n \in \mathbb{N}\}$ was shown to satisfy parts (a), (b) and (c) of Condition 6.1 in Corollary 7.4, Theorem 7.2 and Theorem 5.4, respectively. Lemma 7.5 establishes existence and uniqueness of the measures $\{\mu^n\}$ and also the exponential tightness of $\{\mu_2^n\}$. Thus the process $\{Z^n,\ n \in \mathbb{N}\}$ satisfies all the assumptions of Theorem 6.4 and Theorem 7.6 follows. □

8. Simplification of the rate function for the invariant distribution. In Theorem 7.6, it was shown that the rate function J for $\{\mu_2^n\}$, the sequence of (Q_A^n, Q_B^n)-marginals of the invariant distributions μ^n, can be expressed in terms of the rate functions $G_T(\eta)$ for the sequence of random variables $\{(Q_A^n(T), Q_B^n(T))\}$:

$$J(\eta) = \inf_{T>0} G_T(\eta),$$

where

$$G_T(\eta) = \inf_{(x_1,x_2)\colon F_T(x_1,x_2)=\eta} \int_0^T [L_1(\dot{x}_1) + L_2(\dot{x}_2)]\,dt.$$

Since we need only consider the system of equations (B) for the special case of zero initial conditions, throughout this section we write $F_T(x_1, x_2)$ in place of $F_T(x_1, x_2, 0, 0, 0, 0)$.

In typical applications, one wants to estimate the stationary probability of data loss at either buffer. As discussed in Sections 2 and 9.2, an estimate is provided by

$$\mu_2^n(([0, a_A] \times [0, a_B])^c),$$

where a_A and a_B are (scaled) buffer sizes. A useful property of the rate function J is that it is affine in radial directions. That is, for any $\eta \in \mathbb{R}^2$ and $\alpha \in [0, \infty)$, $J(\alpha\eta) = \alpha J(\eta)$. To see this, let $T \in [0, \infty)$ and let x_1 and x_2 be any pair of inputs in the definition of $G_T(\eta)$ with finite cost. For $\alpha \in [0, \infty)$, we define inputs $\bar{x}_1$ and $\bar{x}_2$ for $G_{\alpha T}(\alpha\eta)$ by

$$\dot{\bar{x}}_1(t) = \dot{x}_1(t/\alpha), \qquad \dot{\bar{x}}_2(t) = \dot{x}_2(t/\alpha).$$

With such a definition, the system of equations (B) implies $F_{\alpha T}(\bar{x}_1, \bar{x}_2) = \alpha\eta$. Since the cost of $\bar{x}_1$ and $\bar{x}_2$ over the interval αT is precisely α times the cost of x_1 and x_2 over the interval T, this shows that $G_{\alpha T}(\alpha\eta) = \alpha G_T(\eta)$. The definition of J then immediately gives the scaling property $J(\alpha\eta) = \alpha J(\eta)$.

This facilitates the problem of estimating $\mu_2^n(([0, a_A] \times [0, a_B])^c)$, since it implies that the infimum in the corresponding variational problem

$$\inf_{\eta \notin [0, a_A] \times [0, a_B]} J(\eta)$$

is achieved on the boundary:

$$\inf_{\eta \notin [0, a_A] \times [0, a_B]} J(\eta) = \Big(\inf_{\eta_2 \in [0, a_B]} J((a_A, \eta_2)) \Big) \wedge \Big(\inf_{\eta_1 \in [0, a_A]} J((\eta_1, a_B)) \Big).$$

For our purposes it will be easier to compute this quantity using the representation

$$\Big(\inf_{\eta_2 \in [0, \infty)} J((a_A, \eta_2)) \Big) \wedge \Big(\inf_{\eta_1 \in [0, \infty)} J((\eta_1, a_B)) \Big), \tag{8.1}$$

the equivalence between the two being another consequence of the scaling property $J(\alpha\eta) = \alpha J(\eta)$. The two variational problems in the last display are simpler in that the constraints take a nicer form, for example, $\eta_2 \in [0, \infty)$ rather than $\eta_2 \in [0, a_B]$.

In this section, we will describe how one can simplify the infinite-dimensional variational problem

$$K(\theta) = \inf_{\eta_1 \in [0, \infty)} J((\eta_1, \theta)) \tag{8.2}$$

by using the convexity of the rate functions L_1 and L_2 and the nature of the mapping F_T. In Theorem 8.1, $K(\theta)$ is shown to be linear in θ and $K(1)$ is expressed as the smallest of the solutions to three finite-dimensional variational problems. Recall from the system of equations (B) that the second coordinate $(F_T(x_1, x_2))_2$ is equal to $q_B(T)$, the buffer B content at time T. It is then convenient to think of $K(1)$ as the least cost that must be incurred by the deterministic inputs (x_1, x_2) to raise the buffer B content to the value 1, where

$\int_0^T [L_1(\dot{x}_1)+L_2(\dot{x}_2)]\,dt$ is the cost function. Each of the finite-dimensional variational problems for $K(1)$ then corresponds to a particular interaction between the inputs within the buffers that raises the contents of buffer B to the value 1. The original variational problem reduces to the finite-dimensional problems when (x_1, x_2) are restricted to certain sets of piecewise linear functions. In the proof of Theorem 8.1, it is shown that without loss of generality, solutions to the original variational problem can be assumed to lie in these sets. The minimum of the solutions to the three finite-dimensional problems is thus shown to solve the original problem. The restrictions imposed on the functions (x_1, x_2) in each case, the resulting properties of the cost functions, and the associated finite-dimensional variational problems are first stated without proof in Section 8.1. The rigorous derivation of the simplification of the variational problem is then presented in Appendix B.

The other variational problem, namely $\inf_{\eta_2\in[0,\infty)} J((a_A, \eta_2))$, that appears in (8.1) is quite easily put into finite-dimensional form, and in fact equals

$$\inf_{\beta_1,\beta_2} \frac{[L_1(\beta_1) + L_2(\beta_2)]}{\beta_1 + \beta_2 - c_A}$$

subject to $\beta_1 + \beta_2 \geq c_A$.

As is well known, the location of the minimizer in $\inf_{\eta_2\in[0,a_B]} J((a_A, \eta_2)) \wedge \inf_{\eta_1\in[0,a_A]} J((\eta_1, a_B))$ indicates which of the buffers is more likely to be exceeded, and the corresponding trajectory indicates the most likely way in which the overflow occurs [12, 22].

8.1. *The finite dimensional variational problems.*

Problem 1. *Buffer B ignores buffer A.* We impose the following restrictions on the functions (x_1, x_2).

ASSUMPTION 8.1. The input functions (x_1, x_2) are linear throughout the time interval on which they are defined and their constant velocities are denoted by $(\dot{x}_1, \dot{x}_2) = (\beta_1, \beta_2)$.

ASSUMPTION 8.2. The input velocities satisfy $\beta_1 + \beta_2 \leq c_A$.

From (4.1) it can be seen that for constant velocity inputs, Assumption 8.2 is equivalent to the constraint that buffer A remain empty throughout the time interval $[0, T]$. The situation is illustrated in Figure 4. In this case, x_1 and x_2 do not interact with each other in buffer A and the exit velocity of each input equals its entrance velocity. Thus buffer B behaves as though buffer A were not present. Let $K_1(\theta)$ be the infimum of the cost over the restricted set of functions (x_1, x_2) that satisfy Assumptions 8.1 and 8.2 and have a final buffer B content of θ. For functions in this set, the minimum cost incurred by the inputs to raise the buffer B to a level θ is linear in θ. Thus $K_1(\theta) = \theta K_1(1)$

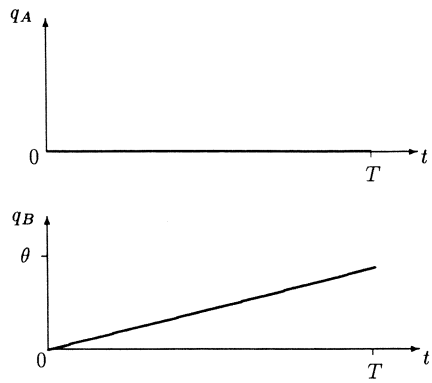

FIG. 4. *Buffer B ignores buffer A.*

and $K_1(1)$ solves the following variational problem:

$$K_1(1) = \inf_{\beta_1, \beta_2} \frac{[L_1(\beta_1) + L_2(\beta_2)]}{\beta_1 - c_B}, \tag{8.3}$$

subject to

$$\beta_1 + \beta_2 \le c_A, \qquad \beta_1 \ge c_B.$$

Problem 2. Buffer A builds up with buffer B. We now state the set of properties that must be satisfied by functions that lead to the second variational problem.

ASSUMPTION 8.3. The input functions (x_1, x_2) are piecewise linear. The time domain on which these functions are defined is divided into two nonempty intervals, on each of which the functions are linear. Thus $(\dot{x}_1, \dot{x}_2) = (\delta_1, \delta_2)$ in the first interval and $(\dot{x}_1, \dot{x}_2) = (b_1, b_2)$ in the second interval.

ASSUMPTION 8.4. The input velocities satisfy the constraints

$$\delta_1 + \delta_2 > c_A, \qquad \frac{\delta_1}{\delta_1 + \delta_2} c_A > c_B.$$

Note that the inputs (b_1, b_2) referred to in Assumption 8.3 represent the mean flows which have zero cost because $L_1(b_1) = L_2(b_2) = 0$. Since the velocities are piecewise constant, (4.1), (4.3) and (4.4) show that trajectories satisfying Assumption 8.4 are such that buffer A and buffer B fill up simultaneously. The situation is illustrated in Figure 5. Let $K_2(\theta)$ be the infimum of the cost functional over those trajectories in this set for which the final buffer B content is equal to θ. For functions in this set, the minimum total

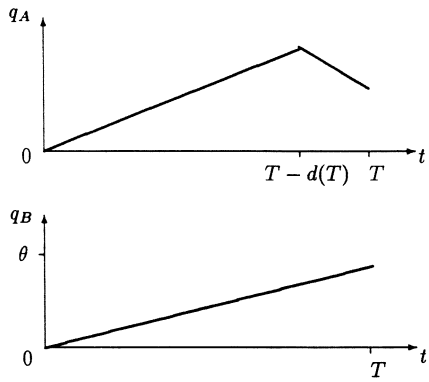

FIG. 5. *Buffer A builds up with buffer B.*

cost is found to be linear in θ and thus $K_2(\theta) = \theta K_2(1)$, where $K_2(1)$ solves the variational problem given by

$$K_2(1) = \inf_{\delta_1, \delta_2} \frac{[L_1(\delta_1) + L_2(\delta_2)]}{(\delta_1/(\delta_1 + \delta_2))c_A - c_B}, \tag{8.4}$$

subject to

$$\delta_1 + \delta_2 > c_A, \qquad \frac{\delta_1}{\delta_1 + \delta_2} c_A \geq c_B.$$

Problem 3. *Buffer B exploits buffer A.* Finally, we describe another set of restrictions on (x_1, x_2) which leads to the last finite-dimensional variational problem.

ASSUMPTION 8.5. The input functions (x_1, x_2) are piecewise linear. The time domain on which the functions are defined is divided into two nonempty intervals, on each of which the functions are linear. Thus $(\dot{x}_1, \dot{x}_2) = (\beta_1, \beta_2)$ in the first interval and $(\dot{x}_1, \dot{x}_2) = (\delta_1, \delta_2)$ in the second interval.

ASSUMPTION 8.6. The input velocities satisfy the constraints

$$\beta_1 + \beta_2 > c_A, \qquad \frac{\beta_1}{\beta_1 + \beta_2} c_A \leq c_B, \qquad \frac{\delta_1}{\delta_1 + \delta_2} c_A > c_B, \qquad \delta_1 + \delta_2 < c_A.$$

Since the velocities are piecewise constant, (4.1) and (4.4) show that trajectories satisfying Assumption 8.6 are such that buffer A is nonempty and fills up to a positive height at the end of the first interval, during which buffer B remains empty. In the second interval, the contents of buffer A decrease while buffer B fills up as long as buffer A remains nonempty. It will be seen that for the infimizer, buffer A empties precisely at the terminal time T. Now, let $K_3(\theta)$ be the infimum of the cost functional over the restricted set of functions

that satisfy Assumptions 8.5 and 8.6 and have a final buffer B content θ. The total cost incurred by these trajectories clearly depends on the values of the inputs in each subinterval and the relative lengths of the two subintervals. For certain inputs, it will turn out that buffer B can be built up at a faster rate and hence lower cost by exploiting the existing nonempty buffer content at A. Thus among the potential minimizing trajectories that one must consider when the system starts empty, are those which raise buffer A during an initial period and then exploit this situation to raise the content of buffer B. The height to which buffer A is raised dictates the amount of time that buffer A remains nonempty and consequently the duration for which buffer B can fill up at a faster rate. Thus we define $K_4(\chi)$ to be the minimum cost that must be incurred to raise the buffer A content to a level χ, when the input functions (x_1, x_2) satisfy the relevant parts of Assumptions 8.5 and 8.6:

$$K_4(\chi) = \inf_{T>0} \inf_{(x_1,x_2):\, q_A(T)=\chi} \int_0^T [L_1(\dot{x}_1) + L_2(\dot{x}_2)]\, dt.$$

The form of the dynamics implies that $K_4(\chi) = \chi K_4(1)$ and $K_4(1)$ solves

$$K_4(1) = \inf_{\beta_1, \beta_2} \frac{[L_1(\beta_1) + L_2(\beta_2)]}{\beta_1 + \beta_2 - c_A}, \tag{8.5}$$

subject to

$$\beta_1 + \beta_2 \geq c_A, \qquad \frac{\beta_1}{\beta_1 + \beta_2} c_A \leq c_B.$$

Figure 6 illustrates the behavior of the system after this initial buildup of buffer A. In the figure, inputs β_1 and β_2 are used to raise buffer A to the level χ at a time $s_1 - d(s_1)$. At this time, new inputs δ_1 and δ_2 are applied. If the condition $\delta_1 + \delta_2 \geq c_A$ holds, then buffer A will not drain after time $s_1 - d(s_1)$. However, in this case one can bound the cost in terms of the solution to Problem 2, and in fact show that the behavior over $[0, s_1 - d(s_1)]$ is suboptimal.

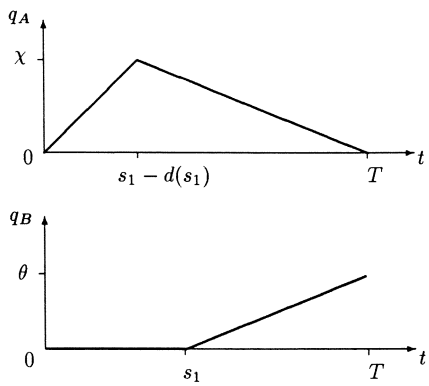

FIG. 6. *Buffer B exploits buffer A.*

Thus we can assume that the condition $\delta_1 + \delta_2 < c_A$ holds and that buffer A drains at some finite time T.

The effect of the change in input rates will not be seen at buffer B until all the input to buffer A due to β_1 and β_2 has been processed. This will take precisely $d(s_1) = \chi/c_A$ units of time. Thus buffer B begins to rise only at time s_1, which is the first time that the new input rates δ_1 and δ_2 affect buffer B. The new inputs will be able to exploit the positive buffer content of A only for the time it takes A to drain under the new rates: $\chi/(c_A - \delta_1 - \delta_2)$. Since the rate of increase of buffer B after time s_1 is $(\delta_1 c_A - (\delta_1 + \delta_2)c_B)/(\delta_1 + \delta_2)$, the inputs δ_1 and δ_2 must be applied for

$$\frac{\chi}{c_A} + \theta \frac{\delta_1 + \delta_2}{\delta_1 c_A - (\delta_1 + \delta_2)c_B}$$

units of time to raise buffer B to level θ. However, all of this must be accomplished before buffer A drains, which imposes the constraint

$$\frac{\chi}{c_A} + \theta \frac{\delta_1 + \delta_2}{\delta_1 c_A - (\delta_1 + \delta_2)c_B} \leq \frac{\chi}{(c_A - \delta_1 - \delta_2)}.$$

We can now define the minimum cost that must be incurred to raise the contents of buffer B to a level θ given that the buffer A content is initially at the value χ. This cost is seen to depend on χ only through the ratio $\chi/\theta = \chi'$ and can thus be expressed as $K_5(\chi', \theta)$, a function of χ' and θ. Then $K_5(\chi', \theta)$ is linear in θ and $K_5(\chi', 1)$ satisfies

$$K_5(\chi', 1) = \inf_{\delta_1, \delta_2} \left(\frac{\chi'}{c_A} + \frac{\delta_1 + \delta_2}{\delta_1 c_A - (\delta_1 + \delta_2)c_B} \right) [L_1(\delta_1) + L_2(\delta_2)],$$

subject to

$$\frac{\delta_1}{\delta_1 + \delta_2} c_A \geq c_B, \qquad \frac{c_A(c_A - \delta_1 - \delta_2)}{\delta_1 c_A - (\delta_1 + \delta_2)c_B} \leq \chi', \qquad \delta_1 + \delta_2 < c_A.$$

Proceeding, we note that the total cost is equal to the sum of the costs incurred during the first two time intervals and is a function of the height $\chi = \theta\chi'$ to which buffer A is raised during the first interval. Thus the total cost, as a function of χ' and θ, is linear in θ and is given by

$$K_4(\chi) + K_5(\chi', \theta) = \chi K_4(1) + \theta K_5(\chi', 1) = \theta[\chi' K_4(1) + K_5(\chi', 1)].$$

The minimum total cost, $K_3(\theta)$, is clearly achieved only by inputs that optimally balance the tradeoff in the cost incurred to raise buffer A to a certain height and the corresponding decrease in the cost of filling buffer B. Thus, minimizing the total cost over all possible positive values of χ', the scaled buffer A content at the end of the first interval yields $K_3(\theta)$. The total cost is affine in χ' with a positive coefficient and also depends on this quantity through the constraint $c_A(c_A - \delta_1 - \delta_2)/(\delta_1 c_A - (\delta_1 + \delta_2)c_B) \leq \chi'$. Thus for any value of χ' that minimizes the total cost, the equality must hold in the constraint. This translates in physical terms to the fact that for a minimizing

trajectory, buffer A empties precisely at the terminal time, and is not raised to a greater height than is necessary.

Thus we obtain

$$K_3(\theta) = \theta K_3(1),$$

where

$$K_3(1) = \inf_{\delta_1, \delta_2} c_A \left[\frac{c_A - \delta_1 - \delta_2}{\delta_1 c_A - (\delta_1 + \delta_2) c_B} K_4(1) + \frac{[L_1(\delta_1) + L_2(\delta_2)]}{\delta_1 c_A - (\delta_1 + \delta_2) c_B} \right], \tag{8.6}$$

subject to

$$\frac{\delta_1}{\delta_1 + \delta_2} c_A \geq c_B, \qquad \delta_1 + \delta_2 < c_A.$$

Thus the three finite-dimensional variational problems (8.3), (8.4) and (8.6) have been described and the corresponding minimizing trajectories have been illustrated in Figures 4, 5 and 6, respectively. We now state the simplification of the rate function in Theorem 8.1, the proof of which is given in Appendix B.

THEOREM 8.1. *Assume $b_1 + b_2 < c_A$ and $b_1 < c_B$. Let $K(\theta)$ be the rate function for $\{\nu^n\}$, the sequence of Q_B^n-marginals of the invariant distributions μ^n. Then $K(\theta)$ has the following finite-dimensional variational representation*:

$$K(\theta) = \theta K(1)$$

and

$$K(1) = \min\{K_1(1), K_2(1), K_3(1)\},$$

where $K_1(1)$, $K_2(1)$ and $K_3(1)$ are as defined in (8.3), (8.4) *and* (8.6), *respectively*.

We close this section with several remarks. In a number of large deviation problems for one-dimensional queueing models, one can easily simplify the rate function through the use of Jensen's inequality. The goal of this section was clearly to describe the corresponding simplification for our model. However, it turns out that while the description of the simplified form just given is reasonably intuitive, the proof is remarkably detailed. In the typical one-dimensional problem, the simplification is completely straightforward. One first shows that the minimization in the definition of the rate function can be restricted to inputs for which the corresponding output never touches the origin after time 0. From the form of the Skorokhod map (2.2), this implies that the output at time t minus the output at time 0 depends only on the input at time t minus the input at time 0. A consequence of the convexity of the integrand in the sample path rate function and Jensen's inequality is that the infimization can be taken over linear inputs, thus leading to the corresponding simplification of the rate function.

The situation here is not so simple. We have two switches to consider, and it is in general simply not true that the minimizer is linear. In fact, it takes one

of three possible forms, involving up to two intervals on which the inputs can be assumed to have constant derivatives. This complexity is obviously due to the fact that there are interesting interactions taking place. An unfortunate consequence is the level of detail required in the proof. The main difficulty is that, when replacing an input on a given interval by its average value over that interval, the highly nonlinear nature of the input–output map does not allow one to assume the remaining portion of the output is not perturbed in a significant way. For example, if we start with an input which has the property that $q_B(T) = \theta$, then in general it is not true that this condition will still be satisfied after replacement. Thus at each stage, it is necessary to understand the effect of the modification on the remaining portion of the output.

The reader will note that in a number of places in the statements of the finite-dimensional problems just given, the minimization involves one or more strict inequalities. It is natural to ask whether these can be relaxed, so that the minimization takes place over a closed set. It turns out that this can be done, and in fact the relaxation for any given case corresponds to a cost and a set of variables already included in one of the other cases. For example, if we include δ_1 and δ_2 in Problem 2 that satisfy $\delta_1 + \delta_2 = c_A$ and $(\delta_1/(\delta_1 + \delta_2))c_A \geq c_B$, we obtain potential values for the infimum that were already considered in Problem 1.

9. Decoupling bandwidths and network extensions.

9.1. *Effective bandwidth revisited.* As part of the motivation for the topic studied in this paper, in the introduction we described the use of the "effective bandwidth" concept as a means of satisfying constraints on the tail of the invariant distribution in the one buffer setting. Associated to each source i is a function H_i and, for a collection of models that include the Markov fluid models we use, one can show that

$$\sum_{i=1}^{I} \frac{H_i(\delta)}{\delta} \leq c \quad \Rightarrow \quad \lim_{n\to\infty} \frac{1}{n} \log \mu([n, \infty)) \leq -\delta, \tag{9.1}$$

where c is the processing rate at the buffer and μ is the Q-marginal of the associated invariant distribution. Using the characterization of H_i as a limit of normalized logarithmic moment generating functions, it is easy to show that for all δ, $H_i(\delta) = \sup_\beta[\delta\beta - L_i(\beta)]$, where the functions L_i were introduced in Section 5. In other words, H_i is the Legendre transform of L_i (and conversely). The effective bandwidth function is defined to be $H_i(\delta)/\delta$ for $\delta > 0$, and to be the mean flow rate b_i for $\delta = 0$.

It is natural to examine the degree to which this concept can be extended to the network setting. As the analysis in this paper reveals, a complete study of the effects due to the interactions of streams within each buffer in a network is a nontrivial task. One might therefore first attempt the seemingly less ambitious task of establishing conditions under which streams "decouple" so that they no longer significantly alter each other's stochastic properties. Consider a specific network for which a large deviation analysis along the lines

of that given in Sections 2–8 can be carried out. Suppose that the buffers are indexed by the parameter θ. For a collection of constraints defined in terms of the parameters δ_θ, we will say that the input streams decouple throughout the network (or simply that they decouple) if a condition analogous to (9.1) applies at each buffer in the network. That is, we require for all buffers θ in the network that

$$\sum_{i \in I_\theta} \frac{H_i(\delta_\theta)}{\delta_\theta} \leq \frac{c_\theta}{a_\theta} \quad \Rightarrow \quad \limsup_{n\to\infty} \frac{1}{n} \log \mu_\theta([na_\theta, \infty)) \leq -\delta_\theta. \tag{9.2}$$

Here I_θ is the subset of sources that share buffer θ, c_θ is the processing rate of buffer θ, na_θ is the size of buffer θ and μ_θ is the Q_θ-marginal of the invariant distribution for the Markov process that models the entire network.

For example, a simple but not very useful constraint which guarantees that the inputs decouple is that the sum of the peak rates of all sources entering a buffer is less than the processing rate of the buffer. For simplicity, we consider the network in Figure 2 and let $a_A = a_B = 1$. Recall that the peak rate of a source is denoted by R_i. If $R_1 + R_2 < c_A$, then the effective bandwidth of the exiting stream Y_1 is clearly equal to that of the entering stream X_1, in the sense that

$$\frac{H_1(\delta_B)}{\delta_B} \leq c_B \quad \Rightarrow \quad \lim_{n\to\infty} \frac{1}{n} \log \mu_2^n([0,\infty) \times [1,\infty)) \leq -\delta_B.$$

The idea of streams "decoupling" was introduced and a sufficient condition under which decoupling would occur was proposed in [5] and [13]. The notion of what it means to decouple used in these papers is stated in a form that differs from the definition given above, although the intended result is the same. Unfortunately, Theorem 8.1 implies that the condition given in [5] and [13] is not always sufficient. Consider the model of Sections 2–8. For the remainder of this section, all references will be to [5]. In that paper, the effective bandwidth $H_i(\delta)/\delta$ is denoted in Corollary 2.1 by $\alpha_i(\delta)$ and the definition of a decoupling bandwidth $\alpha_i^*(\delta)$ is introduced in Corollary 3.1. In our notation, the decoupling bandwidth is given by

$$\alpha_i^*(\delta) = \arg\max_\alpha [\alpha\delta - L_i(\alpha)].$$

In other words,

$$H_i(\delta) = \alpha_i^*(\delta)\delta - L_i(\alpha_i^*(\delta)).$$

An obvious consequence of the fact that $\dot{X}_i$ never exceeds R_i, where R_i is the peak rate of source i, is that $L_i(\beta) = +\infty$ if $\beta > R_i$. It follows directly from the definitions that both $\alpha_i(\delta)$ and $\alpha_i^*(\delta)$ are bounded by R_i. In Corollary 3.2, it is asserted that a sufficient condition for the input stream X_1 and the output stream Y_1 to have the same effective bandwidth is that

$$\alpha_1^*(\delta_A) + \alpha_2(0) = \alpha_1^*(\delta_A) + b_2 < c_A. \tag{9.3}$$

By this assertion, what is meant is that the bandwidth function $H_1(\delta)/\delta$ can be used at the next buffer in the same way it was used at the first; that is,

(9.2) should hold for the first two buffers. In particular, if true, the assertion would imply that if (9.3) holds and if the effective bandwidth constraint,

$$\alpha_1(\delta_B) < c_B, \tag{9.4}$$

is satisfied at buffer B, then

$$\lim_{n\to\infty} \frac{1}{n} \log \mu_2^n\big([0,\infty) \times [1,\infty)\big) \le -\delta_B. \tag{9.5}$$

We now show that this is not necessarily true. Assume that for the network in Figure 2,

$$R_1 + b_2 < c_A, \qquad R_1 < c_B. \tag{9.6}$$

Then (9.3) and (9.4) are satisfied for all values of δ_A and δ_B. Taking the limit $\delta_B \to \infty$ in (9.5) leads to the conclusion that

$$\lim_{n\to\infty} \frac{1}{n} \log \mu_2^n\big([0,\infty) \times [1,\infty)\big) = -\infty.$$

In our notation, this corresponds to the assertion that $K(1) = \infty$. However, the analysis of Section 8 shows that under certain conditions (described in Case 3), the first source can exploit the second source to increase its exit velocity above its peak rate R_1 and thus fill buffer B to a level greater than 1 with positive probability. In particular, suppose that in addition to the constraints (9.6), we also have

$$b_1 + R_2 > c_A, \qquad \frac{R_1}{R_1 + b_2} c_A > c_B. \tag{9.7}$$

Then one can check that $K_1(1) = K_2(1) = \infty$ since $R_1 < c_B$. Since under the invariant distribution for ξ_i, $r_i(\xi_i) = R_i$ with positive probability, we know that $L_i(R_i) < \infty$. It follows from the first constraint in (9.7), that $R_1 + R_2 > c_A$, and, from this inequality and the second inequality of (9.6) that $R_1 c_A/(R_1 + R_2) \le c_B$. Then $K_4(1)$ is clearly finite since the inputs (R_1, R_2) satisfy the constraints below (8.5) and yield a finite value for the functional that is being infimized. Similarly, $K_3(1)$ is also finite since $K_4(1)$ is finite, the inputs (R_1, b_2) satisfy the constraints below (8.6) and the functional evaluated at (R_1, b_2) is finite. Thus we have

$$\lim_{n\to\infty} \frac{1}{n} \log \mu_2^n\big([0,\infty) \times [1,\infty)\big) \ge -K_3(1) > -\infty.$$

This implies that there exist $\delta_A, \delta_B \in (0,\infty)$ for which the proposed decoupling condition (9.3) and the effective bandwidth condition at the second buffer (9.4) holds, but the probability constraint for the second buffer in (9.2) is violated.

Thus the description of useful conditions under which decoupling occurs remains an open problem.

9.2. *Finite buffer models.* In this subsection we elaborate on a remark made in Section 2. Suppose that in place of the model used in Sections 2–8 we use the finite buffer analogue with (scaled) buffer sizes a_A and a_B. This can be constructed by replacing the Skorokhod map Γ by the analogous map that constrains Q_A^n at both the endpoints 0 and a_A and Q_B^n at the endpoints 0 and a_B. To simplify the discussion, let us consider just the one buffer problem and use the analogous notation c, a, ξ^n, X^n, Q^n. We model X^n just as X_1^n and X_2^n were modeled in Section 2. For this set-up, the Skorokhod map relating the input X^n to the buffer content Q^n can be formulated as follows [7]. Let $Q^n(0) \geq 0$ be given. For every ω in the underlying probability space, there exist a unique pair of continuous, nondecreasing and real-valued processes (L_1^n, L_2^n) and a continuous $[0, a]-$valued process Q^n such that the following properties hold:

1. $Q^n(t) = Q^n(0) + X^n(t) - ct + L_1^n(t) - L_2^n(t)$;
2. L_1^n increases only when $Q^n = 0$ and L_2^n increases only when $Q^n = a$;
3. L_1^n and L_2^n are of bounded variation on every finite time interval.

The process Q^n is the desired model for the buffer content. We can write $Q^n = \tilde{\Gamma}(X^n - g)$, where $g(t) = ct - Q^n(0)$. It follows from [7] that $\tilde{\Gamma}$ can be defined as a Lipschitz continuous map on $\mathscr{C}[0, \infty)$ with constant 2 with respect to the supremum norm. The "local time" L_1^n provides the proper constraining action at the origin, while L_2^n realizes the constraint at a. Moreover, it is easily seen that $L_2^n(t)$ is precisely the total data lost due to overflow in the interval $[0, t]$. We are interested in the asymptotic behavior of

$$E \lim_{T \to \infty} \frac{L_2^n(T)}{T}$$

which gives the expected amount of data lost per unit time.

Let $\tilde{\mu}^n$ denote the invariant distribution of (ξ^n, Q^n). For any $T \in (0, \infty)$, we can represent the average amount of data lost per unit time as

$$E_{\tilde{\mu}^n} \frac{L_2^n(T)}{T}.$$

Using the techniques of Section 6 to relate the asymptotics of this quantity to the rate function for the buffer process, one can readily establish the following formula:

$$-\lim_{n \to \infty} \frac{1}{n} \log E_{\tilde{\mu}^n} \frac{L_2^n(T)}{T} = \tilde{J}(a),$$

where, for $g(t) = ct$, and with $\mathscr{I}[0, \infty)$ defined as in Section 4,

$$\tilde{J}(a) = \inf \left\{ \int_0^T L(\dot{\phi})\, dt \colon T \in [0, \infty),\ q = \tilde{\Gamma}(\phi - g),\ \phi \in \mathscr{I}[0, \infty), \right.$$
$$\left. \text{with } \phi(0) = 0 \text{ and } q(T) = a \right\}.$$

At this point one can easily see why the large deviation asymptotics for $\{E_{\tilde{\mu}^n}(L_2^n(T)/T)\}$ in the finite buffer case are the same as those for

$\{\mu_2^n([a,\infty))\}$, where μ_2^n is the second marginal of the joint invariant distribution of the Markovianized process in the infinite buffer case. The rate function in the infinite buffer case takes the form

$$J(a) = \inf\left\{\int_0^T L(\dot{\phi})\,dt\colon\ T \in [0,\infty),\ \ q = \Gamma(\phi - g),\ \ \phi \in \mathscr{I}[0,\infty),\right.$$
$$\left.\text{with } \phi(0) = 0 \text{ and } q(T) = a\right\}.$$

In other words, the only difference between the two is in the form of the mapping that takes the input $\phi - g$ to the output q. We note that for any given input, the outputs of the two mappings coincide on an interval $[0, T]$ if the outputs satisfy $q(t) < a$, $t < T$. This is because the constraint at a is never activated in the mapping for the finite buffer model until possibly $t = T$. Since the nonnegativity of L implies that the infimization can be restricted to such paths, we obtain $\tilde{J}(a) = J(a)$.

This situation is preserved in the multidimensional case, although the notation is more involved.

9.3. *Other network models.* The techniques developed in this paper can be extended to treat a number of more complicated and higher dimensional models. One can consider generalizations in a number of different directions. One type of extension that is easily accommodated allows more general data source models. More interesting generalizations involve modifying the network structure and increasing the number of buffers.

A simple and natural extension of the two buffer network we have considered so far allows a third independent source with scaled cumulative data process X_3^n to share the second buffer with source 1. We will suppose that this source is modeled via a finite state Markov chain as in Section 2. Let b_3 denote the mean output rate for this source. Then the stability conditions obviously become

$$b_1 + b_2 < c_A, \qquad b_1 + b_3 < c_B,$$

and the Lyapunov function can be constructed just as in Appendix A. The continuity of the mapping that takes the inputs and initial conditions into the buffer content processes follows from Theorem 4.1, and the obvious analogue of Theorem 5.4 gives the rate function for the buffer content at time $t \in [0,\infty)$. Theorem 6.4, which connects this rate function to that of the (scaled) invariant distribution, goes through without change. The only significant difference is in the form the finite-dimensional simplification takes in Section 8. The form must now reflect the fact that there is a third independent source that can be used to help raise the level of the second buffer. In particular, if L_3 is the integrand for the rate function for $\{X_3^n\}$, then we obtain the following replacements for $K_1(1)$, $K_2(1)$ and $K_3(1)$ (the interpretations of these problems are analogous to those given in Section 8):

$$K_1(1) = \inf_{\beta_1,\beta_2,\beta_3} \frac{[L_1(\beta_1) + L_2(\beta_2) + L_3(\beta_3)]}{\beta_1 + \beta_3 - c_B},$$

subject to

$$\beta_1 + \beta_2 \le c_A, \qquad \beta_1 + \beta_3 \ge c_B;$$

$$K_2(1) = \inf_{\delta_1, \delta_2, \delta_3} \frac{[L_1(\delta_1) + L_2(\delta_2) + L_3(\delta_3)]}{(\delta_1/(\delta_1 + \delta_2))c_A + \delta_3 - c_B},$$

subject to

$$\frac{\delta_1}{\delta_1 + \delta_2} c_A + \delta_3 \ge c_B, \qquad \delta_1 + \delta_2 > c_A;$$

$$K_3(1) = \inf_{\delta_1, \delta_2, \delta_3} \Bigg[\frac{c_A - \delta_1 - \delta_2}{\delta_1 c_A - (\delta_1 + \delta_2)(\delta_3 - c_B)} K_4(1) + \frac{[L_1(\delta_1) + L_2(\delta_2) + L_3(\delta_3)]}{\delta_1 c_A - (\delta_1 + \delta_2)(\delta_3 - c_B)} \Bigg],$$

subject to

$$\frac{\delta_1}{\delta_1 + \delta_2} c_A + \delta_3 \ge c_B, \qquad \delta_1 + \delta_2 < c_A,$$

where

$$K_4(1) = \inf_{\beta_1, \beta_2} \frac{[L_1(\beta_1) + L_2(\beta_2)]}{\beta_1 + \beta_2 - c_A},$$

subject to

$$\beta_1 + \beta_2 \ge c_A, \qquad \frac{\beta_1}{\beta_1 + \beta_2} c_A \le c_B.$$

One can also consider networks with more than two buffers. We will not provide a detailed discussion on this topic, but will simply describe the networks to which the methods of this paper can be applied in a more or less straightforward manner. The first restriction occurs if we want to apply the continuity result in Theorem 4.1 for the two-buffer model to obtain continuity for a more general network. The restriction is that we must consider only networks for which there is an ordering of the buffers, with the property that the output from a higher order buffer is never fed back as the input to a lower order buffer. For such a network, Theorem 4.1 (or more precisely, the elementary generalization of Theorem 4.1 that allows an arbitrary but finite number of input streams) can be applied sequentially. One starts with the lowest numbered buffer and proceeds up to the highest in order to prove the continuity of the map that takes initial conditions and all the inputs to the joint buffer content process for the network.

An example of a network that does not possess such an ordering property is given in Figure 7. Even if a continuity result were available for such a network and all the other results could be adapted to prove a large deviation principle for the invariant distribution, it would probably be significantly more difficult

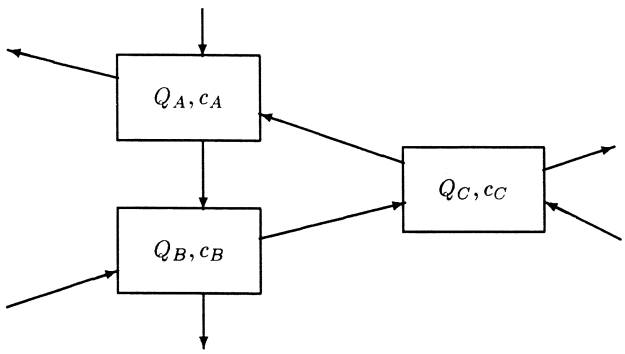

FIG. 7. *Three buffer cyclic network.*

to prove an analogue of Theorem 8.1. The problem is related to the difficulty in proving the continuity result: one cannot automatically rule out the possibility that the "most likely" way for large queue sizes to occur involves some type of cycling behavior, wherein a sequence of well-timed but relatively small inputs exploits a kind of system resonance to build a large queue size.

Thus we restrict the rest of our discussion to networks that satisfy the ordering property described above. It is useful to review the results (besides Theorem 4.1) proved earlier in the paper. The techniques used in two of the other main results are broadly applicable. These results are the theorem that produces a large deviation principle for the buffer contents at a given time from the sample path result for the inputs (Theorem 5.4) and the theorem that connects the rate function of the process to that of the associated invariant distribution (Theorem 6.4). The latter result uses in an essential way certain stability properties, all of which require the construction of a Lyapunov function as in Appendix A. The method used in Appendix A to construct Lyapunov functions can also be generalized to larger networks. For example, consider the networks shown in Figures 8 and 9. The dynamics are described by the obvious extension of the dynamics for the two-station network. A detailed calculation of the Lyapunov functions associated with these networks can be found in [21], Section 9.1.

Finally, we note that more elaborate scalings can also be dealt with using much the same techniques. For example, one could simultaneously scale the buffer size and the number of users as in [22].

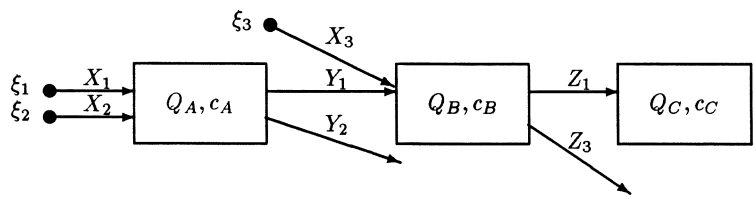

FIG. 8. *Series network.*

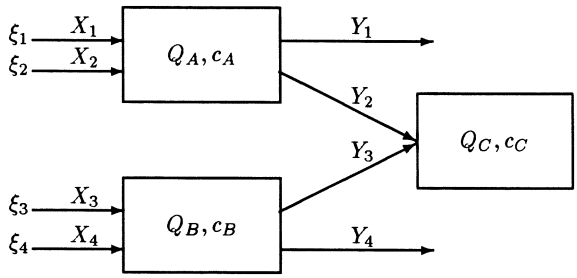

FIG. 9. *Three buffer network.*

APPENDIX A

Construction of Lyapunov functions to establish stability. In this Appendix, we prove the first three results stated in Section 7. We first prove the stability of the associated "fluid model," which is another way of saying the stability of the system under the mean flow. In the course of the proof we will construct a Lyapunov function that will also be used to establish stability properties of the process itself.

PROOF OF THEOREM 7.1. It has been assumed that $\dot{x}_1 = b_1$, $\dot{x}_2 = b_2$, $b_1 + b_2 < c_A$ and $b_1 < c_B$. In this case $(0, 0)$ is a critical point for (q_A, q_B) described by the system of equations (B). This follows because if $(q_A(t), q_B(t)) = (0, 0)$, then $\dot{q}_A(t) = [b_1 + b_2 - c_A] \vee 0 = 0$ from (4.1). This implies $d(t)$ is identically equal to zero, and thus $\dot{q}_B(t) = [b_1 - c_B] \vee 0 = 0$ from (4.4).

Let t_a be the time taken for buffer A to drain out under the mean flow: $t_a = \inf\{t \geq 0: q_A(t) = 0\}$. If $q_A(0) = z_a = 0$, then $t_a = 0$. If $q_A(0) = z_a > 0$, then $q_A(t) > 0$ for all $t \in [0, t_a)$ and from (4.1),

$$\dot{q}_A(t) = b_1 + b_2 - c_A \quad \text{for all } t \in [0, t_a),$$
$$\dot{q}_A(t) = 0 \quad \text{for all } t \geq t_a.$$

Integrating $\dot{q}_A(t)$ over the interval $[0, t_a)$ gives

$$q_A(t_a) = z_a + (b_1 + b_2 - c_A)t_a.$$

From the definition of t_a, $q_A(t_a) = 0$, and thus

$$t_a = -\frac{z_a}{b_1 + b_2 - c_A}. \tag{A.1}$$

Moreover, $q_A(t) = 0$ for all $t \geq t_a$. During $[0, t_a]$, the maximum rate of increase (or least rate of decrease) possible in the contents of buffer B is $(c_A - c_B)$. Hence the maximum possible buffer B content at t_a is

$$q_B(t_a) = [z_b + (c_A - c_B)t_a] \vee 0. \tag{A.2}$$

Let $t_b = \inf\{t \geq t_a\colon q_B(t) = 0\}$. Since $q_A(t) = 0$ for $t \geq t_a$, $d(t) = 0$ for $t \geq t_a$, and it follows that

$$\dot{q}_B(t) = \begin{cases} b_1 - c_B & \text{for all } t \in [t_a, t_b), \\ 0 & \text{for all } t \geq t_b. \end{cases}$$

Thus from the dynamics, $(q_A(t), q_B(t)) = (0, 0)$ for all $t \geq t_b$. □

We now explicitly identify t_b, which will be used below as a Lyapunov function. Integrating $\dot{q}_B(t)$ over the interval $[t_a, t_b)$ yields

$$0 = q_B(t_a) + (b_1 - c_B)(t_b - t_a).$$

Substituting for t_a from (A.1) and $q_B(t_a)$ from (A.2) in the last equation, we obtain

$$\left[z_b - \frac{(c_A - c_B)z_a}{b_1 + b_2 - c_A}\right] \vee 0 + (b_1 - c_B)\left(t_b + \frac{z_a}{b_1 + b_2 - c_A}\right) = 0,$$

which when rearranged gives

$$\begin{aligned} t_b &= \frac{1}{c_B - b_1}\left(\left[z_b + \frac{(c_A - c_B)z_a}{c_A - b_1 - b_2}\right] \vee 0\right) + \frac{z_a}{c_A - b_1 - b_2} \\ &= \left[\frac{z_a(c_A - b_1)}{(c_A - b_1 - b_2)(c_B - b_1)} + \frac{z_b}{c_B - b_1}\right] \vee \frac{z_a}{c_A - b_1 - b_2}. \end{aligned}$$

It is clear that t_b depends on the initial condition (z_a, z_b). If we define $V(z_a, z_b) = t_b$, then

$$V(z_a, z_b) = \langle \alpha_1, (z_a, z_b)\rangle \vee \langle \alpha_2, (z_a, z_b)\rangle, \tag{A.3}$$

where

$$\alpha_1 = \left(\frac{(c_A - b_1)}{(c_A - b_1 - b_2)(c_B - b_1)}, \frac{1}{c_B - b_1}\right) \quad \text{and} \quad \alpha_2 = \left(\frac{1}{c_A - b_1 - b_2}, 0\right).$$

Then $V(z_a, z_b)$ is obviously linear on each of the sets $\mathscr{D}_1$ and $\mathscr{D}_2$ defined by

$$\mathscr{D}_1 = \left\{z \in \mathbb{R}^2\colon \langle \alpha_1, z\rangle > \langle \alpha_2, z\rangle\right\} \quad \text{and} \quad \mathscr{D}_2 = \left\{z \in \mathbb{R}^2\colon \langle \alpha_2, z\rangle > \langle \alpha_1, z\rangle\right\}.$$

Let $\mathscr{D}_{1,2}$ denote the common boundary of $\mathscr{D}_1$ and $\mathscr{D}_2$, so that

$$\mathscr{D}_{1,2} = \{z \in \mathbb{R}^2\colon \langle \alpha_1, z\rangle = \langle \alpha_2, z\rangle\}.$$

We have constructed a function which provides an upper bound on the time for the fluid model to reach the origin, given initial buffer contents z_a and z_b. Although it is not in precise analogy with the method used in [9], it is nevertheless plausible that the function V could serve as a Lyapunov function for the associated stochastic model. As we will see, this is indeed the case. Generalizations of the method used here and a second construction that is closer in spirit to the method used in [9] are outlined in [21]. We now establish that the function V defined in (A.3) is indeed a Lyapunov function.

THEOREM A.1. *Consider deterministic functions x_1, x_2, q_A and q_B as described by the system of equations (B). Assume that $b_1 + b_2 < c_A$ and $b_1 < c_B$. The function V defined in* (A.3) *has the following properties*:

(a) $V(z_a, z_b)$ *is piecewise linear*;
(b) $V(z_a, z_b) \geq 0$;
(c) $V(z_a, z_b) = 0$ *iff* $(z_a, z_b) = (0, 0)$;
(d) *define* $L_t V(q_A(t), q_B(t))$ *to be the orbital derivative of* V, *that is, the derivative in* t *of the composed function. Then a.s. for* t *such that* $q_A(t) \vee q_B(t) > 0$, $L_t V$ *is given by*

$$L_t V(q_A(t), q_B(t)) = \langle \nabla V(q_A(t), q_B(t)), (\dot{q}_A(t), \dot{q}_B(t)) \rangle, \tag{A.4}$$

where $\nabla V(z_a, z_b)$ *can be defined as either* α_1 *or* α_2 *for* $(z_a, z_b) \in \mathscr{D}_{1,2}$.

PROOF. The first three properties follow immediately from the form of V defined in (A.3). Thus we only have to show property (d). Since V is Lipschitz continuous and q_A and q_B are absolutely continuous, $V(q_A(t), q_B(t))$ is absolutely continuous. However, the fact that V is not differentiable everywhere means that a little care must be taken when identifying the derivative of $V(q_A(t), q_B(t))$. Clearly, the only complication arises when calculating the derivative for t such that $(q_A(t), q_B(t)) \in \mathscr{D}_{1,2}$. We will use the fact that the set

$$\{t: (q_A(t), q_B(t)) \in \mathscr{D}_{1,2}, (\dot{q}_A(t), \dot{q}_B(t)) \notin \mathscr{D}_{1,2}\}$$

has Lebesgue measure zero [6], Theorem A.6.3. It is easy to check that the projection of α_1 onto $\mathscr{D}_{1,2}$ equals the projection of α_2 onto $\mathscr{D}_{1,2}$. We denote the common projection by $\alpha_{1,2}$. Then the definition of the derivative and the last three sentences imply

$$\begin{aligned}\frac{d}{dt} V(q_A(t), q_B(t)) &= \langle \alpha_{1,2}, (\dot{q}_A(t), \dot{q}_B(t)) \rangle \\ &= \langle \alpha_1, (\dot{q}_A(t), \dot{q}_B(t)) \rangle = \langle \alpha_2, (\dot{q}_A(t), \dot{q}_B(t)) \rangle\end{aligned}$$

a.s. for $t \in \{t: (q_A(t), q_B(t)) \in \mathscr{D}_{1,2}$ and $q_A(t) \vee q_B(t) > 0\}$. Hence without ambiguity, we can set $\nabla V(x)$ equal to either α_1 or α_2 when $x \in \mathscr{D}_{1,2}$, and have (A.4) true a.s. in t. □

We now use the Lyapunov function V to establish Theorem 7.2.

PROOF OF THEOREM 7.2. Consider the Lyapunov function defined in (A.3). We also consider the deterministic functions x_1, x_2, q_A and q_B described in the system of equations (B), and divide $\mathbb{R}^2 \backslash \{(0, 0)\}$ into two regions.

Region 1 $(q_A(t), q_B(t))$: $q_A(t) > 0$ and $q_B(t) \geq 0$. Here $q_A(t)$ and $q_B(t)$ satisfy

$$\begin{aligned}\dot{q}_A(t) &= \dot{x}_1(t) + \dot{x}_2(t) - c_A, \\ \dot{q}_B(t) &= \alpha(t) c_A - c_B\end{aligned}$$

a.s. in t, where $\alpha(t)$ is a measurable function that takes values in $[0, 1]$ when $q_B(t) > 0$ and almost surely takes the value c_B/c_A when $q_B(t) = 0$. [Here we have again used the fact that for an absolutely continuous function $f\colon [0, T] \to \mathbb{R}$, the set of t such that $f(t) = 0$ and $\dot{f}(t) \neq 0$ is a set of Lebesque measure zero.] Recall that $L_t V$, the orbital derivative of V, was shown in Theorem A.1 to equal

$$L_t V(q_A(t), q_B(t)) = \langle \nabla V(q_A(t), q_B(t)), (\dot{q}_A(t), \dot{q}_B(t)) \rangle$$

a.s. for t such that $q_A(t) \vee q_B(t) > 0$, where $\nabla V(z_a, z_b)$ can be defined as either α_1 or α_2 for $(z_a, z_b) \in \mathscr{D}_{1,2}$. If $(q_A(t), q_B(t)) \in \bar{\mathscr{D}}_1$, then either $q_B(t) > 0$, in which case $\alpha(t) \in [0, 1]$, or else $q_B(t) = 0$ and $q_A(t) > 0$, in which case $\alpha(t) = c_B/c_A$ a.s. Now in the latter case there exists a vector $(v_A, 0) \in \mathscr{D}_1$ with $v_A > 0$, which implies $(c_A - b_1)/(c_B - b_1) \geq 1$, or $c_A \geq c_B$. Thus $\alpha(t) \in [0, 1]$ a.s. for all t such that $(q_A(t), q_B(t)) \in \bar{\mathscr{D}}_1$. For such t, we have

$$\begin{aligned}
L_t V &= \frac{(c_A - b_1)(\dot{x}_1(t) + \dot{x}_2(t) - c_A)}{(c_A - b_1 - b_2)(c_B - b_1)} + \frac{\alpha(t)c_A - c_B}{c_B - b_1} \\
&= \frac{(c_A - b_1)(\dot{x}_1(t) + \dot{x}_2(t) - b_1 - b_2)}{(c_A - b_1 - b_2)(c_B - b_1)} \\
&\quad - \frac{(c_A - b_1)(c_A - b_1 - b_2)}{(c_A - b_1 - b_2)(c_B - b_1)} + \frac{\alpha(t)c_A - c_B}{c_B - b_1} \\
&= \frac{(c_A - b_1)(\dot{x}_1(t) + \dot{x}_2(t) - b_1 - b_2)}{(c_A - b_1 - b_2)(c_B - b_1)} + \frac{(\alpha(t) - 1)c_A}{c_B - b_1} - 1 \\
&\leq \frac{(c_A - b_1)(\dot{x}_1(t) + \dot{x}_2(t) - b_1 - b_2)}{(c_A - b_1 - b_2)(c_B - b_1)} - 1.
\end{aligned}$$

On the other hand, if $(q_A(t), q_B(t)) \in \bar{\mathscr{D}}_2$, $L_t V = \langle \alpha_2, (\dot{q}_A(t), \dot{q}_B(t)) \rangle$, which gives

$$L_t V = \frac{\dot{x}_1(t) + \dot{x}_2(t) - c_A}{c_A - b_1 - b_2} = \frac{\dot{x}_1(t) + \dot{x}_2(t) - b_1 - b_2}{c_A - b_1 - b_2} - 1$$

Region 2 $(q_A(t), q_B(t))$: $q_A(t) = 0$ and $q_B(t) > 0$. It can be easily seen that Region 2 is contained in $\mathscr{D}_1$ and therefore $\nabla V(q_A(t), q_B(t)) = \alpha_1$. The dynamics in this region are given by

$$\begin{aligned}
\dot{q}_A(t) &= 0, \\
\dot{q}_B(t) &= \dot{x}_1(t) - c_B
\end{aligned}$$

a.s. in t, once again using the property of absolutely continuous functions that was stated above. Therefore

$$L_t V = \frac{\dot{x}_1(t) - c_B}{c_B - b_1} = \frac{\dot{x}_1(t) - b_1}{c_B - b_1} - 1.$$

For notational convenience we define

$$k_1 = \frac{(c_A - b_1)}{(c_A - b_1 - b_2)(c_B - b_1)} \vee \frac{1}{c_A - b_1 - b_2} \quad \text{and} \quad k_2 = \frac{1}{c_B - b_1},$$

and note that $k_1 \geq k_2$. From the stability conditions, $k_1 \wedge k_2 > 0$. Integrating and using the definitions of k_1 and k_2, we obtain

$$\begin{aligned} &V(q_A(t), q_B(t)) - V(q_A(0), q_B(0)) + t \\ &\quad \leq \int_0^t \big[k_1 1_{\{q_A(s)>0\}}(\dot{x}_1(s) + \dot{x}_2(s) - b_1 - b_2) \\ &\qquad\qquad + k_2 1_{\{q_A(s)=0\}}(\dot{x}_1(s) - b_1)\big]\, ds \end{aligned} \tag{A.5}$$

for all t such that $(q_A(s), q_B(s)) \neq (0,0)$ for $s \in [0,t]$. This inequality provides the intuition behind the theorem. If $(Q_A^n(t), Q_B^n(t))$ remains nonzero, then $V(Q_A^n(t), Q_B^n(t)) - V(Q_A^n(0), Q_B^n(0))$ remains bounded from below since $V(z_a, z_b)$ is a Lyapunov function for the dynamical system. This implies that the integral on the right must increase approximately linearly with t. However, the large deviation property of the inputs (X_1^n, X_2^n) suggests that this event happens with exponentially small probability. In order to make this argument rigorous, we first define the set

$$Q_\lambda = \{(z_a, z_b)\colon V(z_a, z_b) \leq \lambda\}.$$

We show that if the process starts in a region of the form Q_λ, then with probability exponentially close to 1 it must enter the region $\tilde{N}_\varepsilon$ within some finite time that is independent of ε.

Given that the initial conditions lie in the compact set K, let K_2 be the projection of K onto $\mathscr{S}_2$. Then let $\lambda = \sup_{(z_a, z_b)\in K_2} V(z_a, z_b)$, so that $K_2 \subset Q_\lambda$. Let $N_\varepsilon = \{x \in \mathscr{S}_2\colon \theta(x, 0) < \varepsilon\}$. We define the closed set $\mathscr{A} \subset \mathscr{C}[0,T]^2 \times \mathscr{S}_{2,3}$ by

$$\begin{aligned} \mathscr{A} = \big\{&(x_1, x_2, q_A(0), q_B(0), u_{1,0}, u_{2,0})\colon (q_A(0), q_B(0)) \in Q_\lambda, \\ &(q_A(t), q_B(t)) \notin N_\varepsilon \text{ for all } t \in [0,T]\big\}. \end{aligned}$$

In the last display (q_A, q_B) are the trajectories associated to $(x_1, x_2, q_A(0), q_B(0), u_{1,0}, u_{2,0})$ through the system of equations (B). We will show that the probability of this event is exponentially small for large enough T. In order to do so, we first bound the event in terms of an event that involves only the input processes (X_1^n, X_2^n) so that we can then use large deviation estimates to estimate its probability.

If $(x_1, x_2, q_A(0), q_B(0), u_{1,0}, u_{2,0}) \in \mathscr{A}$, then $V(q_A(0), q_B(0)) \leq \lambda$ and for all $t \in [0,T]$, $V(q_A(t), q_B(t)) > 0$ [since $(q_A(t), q_B(t)) \notin N_\varepsilon$]. By (A.5), for such $(x_1, x_2, q_A(0), q_B(0), u_{1,0}, u_{2,0})$ and $t \in [0,T]$,

$$\begin{aligned} &\int_0^t k_1 1_{\{q_A(s)>0\}}(\dot{x}_1(s) - b_1)\, ds + \int_0^t k_1 1_{\{q_A(s)>0\}}(\dot{x}_2(s) - b_2)\, ds \\ &\qquad + \int_0^t k_2 1_{\{q_A(s)=0\}}(\dot{x}_1(t) - b_1)\, ds \geq t - \lambda. \end{aligned}$$

Henceforth, let $T \geq 3\lambda$. Then since $k_1 \geq k_2$, the last display implies

$$\begin{aligned}\left(\int_0^T 1_{\{q_A(s)>0\}}(\dot{x}_1(s)-b_1)\,ds\right) \vee \left(\int_0^T 1_{\{q_A(s)>0\}}(\dot{x}_2(s)-b_2)\,ds\right) \\ \vee \left(\int_0^T 1_{\{q_A(s)=0\}}(\dot{x}_1(t)-b_1)\,ds\right) \geq \frac{2T}{9k_1}.\end{aligned} \tag{A.6}$$

Recall that it was established in Lemma 5.3 that for each $i = 1, 2$ the sequence X_i^n satisfies the large deviation principle uniformly in the initial condition $\xi_i^n(0)$ with rate function $I_T^i(\phi) = \int_0^T L_i(\dot{\phi}(s))\,ds$, and that each L_i is nonnegative, convex, has compact level sets, with $L_i(u) = 0$ iff $u = b_i$. Let us consider the first term in the last display, and suppose that this term is bounded below by $2T/9k_1$. Let m denote the Lebesgue measure of the Borel set $\{s \in [0, T]\colon q_A(s) > 0\}$. Using the convexity of L_1, Jensen's inequality and the fact that $L_1(b_1) = 0$,

$$\begin{aligned}\int_0^T 1_{\{q_A(s)>0\}}L_1(\dot{x}_1(s))\,ds &\geq mL_1\left(\int_0^T 1_{\{q_A(s)>0\}}\dot{x}_1(s)\,ds/m\right) \\ &\geq mL_1\left(\frac{2T}{9k_1m}+b_1\right) \\ &\geq TL_1\left(\frac{2}{9k_1}+b_1\right).\end{aligned}$$

By estimating the second and third terms in (A.6) in a similar fashion, we obtain the bound

$$\begin{aligned}\left(\int_0^T 1_{\{q_A(s)>0\}}L_1(\dot{x}_1(s))\,ds\right) \vee \left(\int_0^T 1_{\{q_A(s)>0\}}L_2(\dot{x}_2(s))\,ds\right) \\ \vee \left(\int_0^T 1_{\{q_A(s)=0\}}L_1(\dot{x}_1(s))\,ds\right) \geq T\alpha,\end{aligned}$$

where $\alpha = L_1((2/9k_1)+b_1) \wedge L_2((2/9k_1)+b_2) > 0$. Thus

$$I_T^1(x_1) \vee I_T^2(x_2) \geq T\alpha. \tag{A.7}$$

For the remainder of the proof, we set $T = 3\lambda \vee M/\alpha$. Note that this definition is independent of $\varepsilon > 0$. Now define the set

$$\mathscr{B} = \{(x_1, x_2)\colon (x_1, x_2, q_A(0), q_B(0), u_{1,0}, u_{2,0}) \in \mathscr{A}\}.$$

Then $\mathscr{B}$ is closed because $\mathscr{A}$ is closed, $\{(q_A(0), q_B(0), u_{1,0}, u_{2,0}) \in \mathscr{S}_{2,3}\colon (q_A(0), q_B(0)) \in Q_\lambda\}$ is compact and the map $F_T\colon (x_1, x_2, q_A(0), q_B(0), u_{1,0}, u_{2,0}) \to (q_A, q_B)$ is continuous (cf. Theorem 4.1). In addition, (A.7) implies that for all $(x_1, x_2) \in \mathscr{B}$,

$$I_T^1(x_1) \vee I_T^2(x_2) \geq M. \tag{A.8}$$

If z is any initial condition such that $z_2 \in Q_\lambda$, then the definitions of $\mathscr{A}$ and $\mathscr{B}$ imply

$$\begin{aligned} P_z\{\tau_\varepsilon^n > T\} &\le P_z\{(X_1^n, X_2^n, Q_A^n(0), Q_B^n(0), U_{1,0}^n, U_{2,0}^n) \in \mathscr{A}\} \\ &\le P_{z_1}\{(X_1^n, X_2^n) \in \mathscr{B}\}. \end{aligned}$$

The lower bound (A.8) and the uniform large deviation principles for $\{X_1^n\}$ and $\{X_2^n\}$ stated in Lemma 5.3 then imply

$$\limsup_{n\to\infty} \frac{1}{n} \log \sup_{z\in K} P_z\{\tau_\varepsilon^n > T\} \le -M,$$

which establishes the theorem. □

We now prove the exponential tightness of the family of random variable $Z^n(t)$. As before, we use the notation $Z_1^n = (\xi_1^n, \xi_2^n)$, $Z_2^n = (Q_A, Q_B)$ and $Z_3^n = (U_1^n, U_2^n)$.

PROOF OF THEOREM 7.3. Given a compact set D, consider the compact set D_2 obtained from the projection of D onto $\mathscr{S}_2$. Recall the definition of the compact set $Q_\lambda \subset \mathscr{S}_2 = \mathbb{R}^2$ in terms of the Lyapunov function defined in (A.3):

$$Q_\lambda = \{(z_a, z_b)\colon V(z_a, z_b) \le \lambda\}.$$

Let $\lambda = \sup_{(z_a, z_b)\in D_2} V(z_a, z_b)$. We define the compact set K by $K = \mathscr{S}_1 \times \{(z_2, z_3) \in \mathscr{S}_{2,3}\colon z_2 \in Q_\lambda\}$. Observe that $D \subset K$. Choose $\varepsilon > 0$ such that $\sup_{(z_a, z_b)\in \bar{N}_\varepsilon} V(z_a, z_b) < \lambda$ and define the following series of random times:

$$\begin{aligned} \tau_0^n &= 0, \\ \sigma_k^n &= \inf\{t > \tau_{k-1}^n\colon (Q_A^n(t), Q_B^n(t)) \in \bar{N}_\varepsilon\}, \\ \tau_k^n &= \inf\{t > \sigma_k^n\colon (Q_A^n(t), Q_B^n(t)) \notin Q_\lambda^0\}, \end{aligned}$$

where Q_λ^0 denotes the interior of the set Q_λ. By Proposition 2.1.5 of [11], σ_k^n and τ_k^n are stopping times for every $n, k \in \mathbb{N}$. Using the strong Markov property of the process Z^n, for $k \in \mathbb{N}$,

$$P_z\{\sigma_{k+1}^n - \tau_k^n > T\} = P_{Z^n(\tau_k^n)}\{\sigma_1^n > T\} \le \sup_{z\in K} P_z\{\sigma_1^n > T\}.$$

Given any $M < \infty$, Theorem 7.2 implies the existence of $T < \infty$ and $N < \infty$ such that for $n \ge N$,

$$\sup_{z\in K} P_z\{\sigma_1^n > T\} \le e^{-nM}. \tag{A.9}$$

Combining the last two inequalities shows that for every $k \in \mathbb{N}$ and $n \ge N$,

$$P_z\{\sigma_{k+1}^n - \tau_k^n > T\} \le e^{-nM}.$$

We now characterize the probability that the buffer content processes leave some compact set C, where $Q_\lambda \subset C$. The indicator function $1_{\{(Q_A^n(t),\, Q_B^n(t))\notin C\}}$ of the event of interest can be expressed as

$$1_{\{(Q_A^n(t),\, Q_B^n(t))\notin C\}} = \sum_{k=0}^{\infty} 1_{\{(Q_A^n(t),\, Q_B^n(t))\notin C\}} 1_{[\tau_k^n,\, \sigma_{k+1}^n)}(t),$$

since $(Q_A^n(t), Q_B^n(t)) \in Q_\lambda \subset C$ for $t \in [\sigma_k^n, \tau_k^n)$. If we let $[a, b) = \varnothing$ whenever $b \le a$, then for any $T \in \mathbb{R}$, we can also write

$$\begin{aligned} 1_{\{(Q_A^n(t),\, Q_B^n(t))\notin C\}} = \sum_{k=0}^{\infty} \Big[& 1_{\{(Q_A^n(t),\, Q_B^n(t))\notin C\}} 1_{[\tau_k^n,\, \tau_k^n+T)}(t) \\ & + 1_{\{(Q_A^n(t),\, Q_B^n(t))\notin C\}} 1_{[\tau_k^n+T,\, \sigma_{k+1}^n)}(t) \Big]. \end{aligned} \tag{A.10}$$

Then $P_z\{t \in [\tau_k^n + T, \sigma_{k+1}^n)\}$ can be rewritten as

$$P_z\{t \in [\tau_k^n + T, \sigma_{k+1}^n) | t \in [\tau_k^n, \sigma_{k+1}^n)\} P_z\{t \in [\tau_k^n, \sigma_{k+1}^n)\},$$

which, by the strong Markov property, equals

$$P_{Z^n(\tau_k^n)}\{t \in [T, \sigma_1^n) | t \in [0, \sigma_1^n)\} P_z\{t \in [\tau_k^n, \sigma_{k+1}^n)\}.$$

By (A.9), for any $M < \infty$ one can choose $T, N < \infty$ such that for $n \ge N$,

$$P_{Z^n(\tau_k^n)}\{t \in [T, \sigma_1^n) | t \in [0, \sigma_1^n)\} \le \sup_{z \in k} P_z\{\sigma_1^n > T\} \le e^{-nM}.$$

Therefore, by the last three statements,

$$P_z\{t \in [\tau_k^n + T, \sigma_{k+1}^n)\} \le e^{-nM} P_z\{t \in [\tau_k^n, \sigma_{k+1}^n)\}.$$

Since $\sum_{k=0}^{\infty} P_z\{t \in [\tau_k^n, \sigma_{k+1}^n)\} \le 1$,

$$\sum_{k=0}^{\infty} P_z\{t \in [\tau_k^n + T, \sigma_{k+1}^n)\} \le e^{-nM} \sum_{k=0}^{\infty} P_z\{t \in [\tau_k^n, \sigma_{k+1}^n)\} \le e^{-nM}. \tag{A.11}$$

Now by the defining equation (2.6) for X_i^n and Assumption 2.1, the cumulative input velocities $\Phi^n = X_1^n + X_2^n$ are bounded above by $R = R_1 + R_2$, where

$$R_i = \max_{\xi_i \in \mathscr{F}_i} r_i(\xi_i).$$

Then R clearly also serves as a bound for the rate of increase of the contents of buffer A. We note that c_A provides the corresponding bound for buffer B, and define

$$C = Q_\lambda \cup \{(a + 2Rt, b + 2c_A t) \colon (a, b) \in Q_\lambda,\ t \in [0, T]\}.$$

If we start with $(Q_A^n(0), Q_B^n(0)) \in Q_\lambda$, then for $t \in [0, T]$ the buffer processes cannot leave the set C. Thus for $t \in [0, T]$,

$$\sup_{z \in K} P_z\{(Q_A^n(t), Q_B^n(t)) \notin C\} = 0.$$

By the strong Markov property and the fact that $Z_2^n(\tau_k^n) \in Q_\lambda$, for every k, $n \in \mathbb{N}$, we obtain for $t \in [\tau_k^n, \tau_k^n + T]$,

$$(Q_A^n(t), Q_B^n(t)) \in C \tag{A.12}$$

w.p.1. Thus taking expectations of all the terms in (A.10), and using (A.11) and (A.12) gives

$$\sup_{z \in K} P_z\{(Q_A^n(t), Q_B^n(t)) \notin C\} \le e^{-nM}.$$

Since $D \subset K$,

$$\sup_{z \in D} P_z\{(Q_A^n(t), Q_B^n(t)) \notin C\} \le e^{-nM},$$

and the theorem is proved. □

APPENDIX B

Proof of Theorem 8.1. Here we provide a rigorous proof of the simplification of the variational problem for the rate function $K(\theta)$ given in (8.2). A physical interpretation of each of the finite-dimensional variational problems obtained here was provided in Section 8.

PROOF OF THEOREM 8.1. We recall that the rate function $K(\theta)$ can be expressed in terms of the rate function $G_T(\eta)$ as

$$K(\theta) = \inf_{T>0,\, \eta_1 \in [0,\infty)} G_T((\eta_1, \theta)), \tag{B.1}$$

where

$$G_T(\eta) = \Big[\inf_{(x_1, x_2):\, F_T(x_1, x_2) = \eta} \int_0^T [L_1(\dot{x}_1) + L_2(\dot{x}_2)]\, dt \Big].$$

Recall also that $F_T(x_1, x_2) = (q_A(T), q_B(T))$, and that $K(\theta)$ can be interpreted as the minimum cost that must be incurred by the input trajectories (x_1, x_2) in order to raise the contents of buffer B to θ, where the cost function takes the form $\int_0^T [L_1(\dot{x}_1) + L_2(\dot{x}_2)]\, dt$ with T being the time at which buffer B reaches the height θ. We now show that the solution $K(\theta)$ to the infinite-dimensional variational problem in (B.1) is linear in θ and can be represented as the minimum of the solutions to the three finite-dimensional variational problems described in (8.3), (8.4) and (8.6). The case $\theta = 0$ follows trivially, and so for the remainder of the proof we assume $\theta > 0$.

From the system of equations (B) in Section 4, one can derive the following explicit expression for $\dot{y}_i(t)$ for the case of $q_A(0) = q_B(0) = 0$. For $i = 1, 2$,

$$\dot{y}_i(t) = \begin{cases} \dfrac{\dot{x}_i(t - d(t))}{\dot{x}_1(t - d(t)) + \dot{x}_2(t - d(t))} c_A, & \text{if } q_A(t) > 0, \\ \dot{x}_i(t), & \text{if } q_A(t) = 0. \end{cases} \tag{B.2}$$

The statement for $q_A(t) = 0$ is obvious. By differentiating (4.3) and using the definition of $d(t)$ given in (4.2) for $q_A(t) > 0$, one obtains

$$\begin{aligned}\dot{y}_i(t) &= \dot{x}_i(t - d(t))[1 - \dot{d}(t)] \\ &= \dot{x}_i(t - d(t))\left[1 - \left(1 - \frac{\dot{\phi}(t) - \dot{q}_A(t)}{\dot{\phi}(s)_{|s=\phi^{-1}(\phi(t)-q_A(t))}}\right)\right] \\ &= \dot{x}_i(t - d(t))\frac{c_A}{\dot{\phi}(t - d(t))},\end{aligned}$$

where the last equality follows by using the fact that $\dot{q}_A(t) = \dot{\phi}(t) - c_A$ when $q_A(t) > 0$ as given in (4.1).

We next define some important parameters of the system which help effect the simplification. Throughout this proof, x_1 and x_2 will be used to denote nondecreasing, absolutely continuous functions that start at 0 at time 0. For any pair (x_1, x_2), the corresponding trajectory (q_A, d, y_1, q_B) is obtained from the system of equations (B). For any pair of inputs (x_1, x_2), define

$$s_1 = \sup\{t \in [0, T]: q_B(t) = 0\},$$
$$s_2 = \sup\{t \in [0, s_1]: q_A(t) = 0\}.$$

Figure 10 shows s_1 and s_2 for a trajectory $(q_A(t), q_B(t))$ with $(q_A(0), q_B(0)) = (0, 0)$. We claim that without loss of generality one can assume that the inputs are such that buffer A is empty for $t \in [0, s_2]$ and buffer B is empty for $t \in [0, s_1]$. To verify this, we first note that the definitions of s_1 and s_2 and

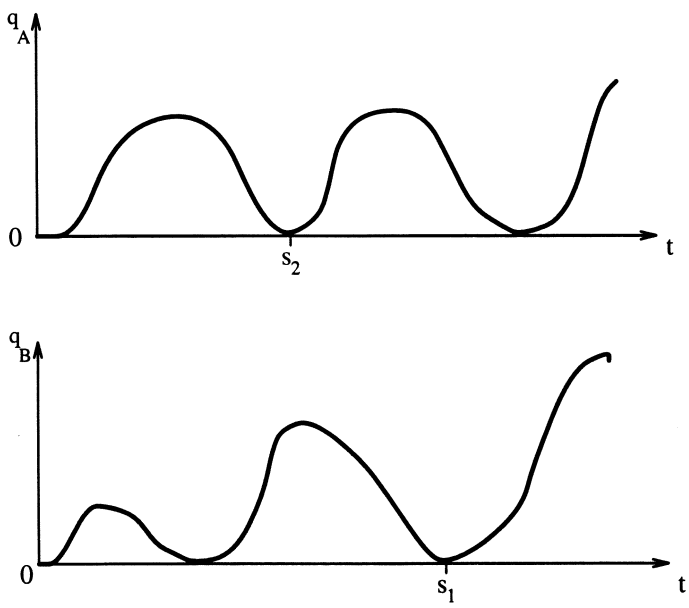

FIG. 10. *Characterization of s_1 and s_2.*

the continuity of q_A and q_B imply that $q_A(s_2) = q_B(s_1) = 0$. Since $q_B(s) > 0$ for $s \in [s_1, T]$, from (4.4) and (B.2), we see that

$$\begin{aligned} q_B(T) - q_B(s_1) &= \int_{s_1}^{T} (\dot{x}_1(t) - c_B) 1_{\{q_A(t)=0\}}\, dt \\ &\quad + \int_{s_1}^{T} \left(\frac{\dot{x}_1(t - d(t))}{\dot{x}_1(t - d(t)) + \dot{x}_2(t - d(t))} c_A - c_B \right) 1_{\{q_A(t)>0\}}\, dt. \end{aligned} \tag{B.3}$$

This equation shows that $q_B(T)$ is fully determined by the values of the inputs (x_1, x_2) and the set of times when q_A is zero during the interval $[s_1 - d(s_1), T]$. However, as can be seen from (4.1), the set of times in $[s_1 - d(s_1), T]$ when q_A is zero, in turn, depends only on $q_A(s_1 - d(s_1))$ and the value of the inputs (x_1, x_2) during the interval $[s_1 - d(s_1), T]$. For any $s \in [s_1 - d(s_1), T]$,

$$\begin{aligned} q_A(s) &= q_A(s_1 - d(s_1)) + \int_{s_1 - d(s_1)}^{s} (\dot{x}_1 + \dot{x}_2 - c_A) 1_{\{q_A(t)>0\}}\, dt \\ &\quad + \int_{s_1 - d(s_1)}^{s} [(\dot{x}_1 + \dot{x}_2 - c_A) \vee 0] 1_{\{q_A(t)=0\}}\, dt. \end{aligned}$$

Hence the value of $q_B(T)$ depends only on $q_A(s_1 - d(s_1))$ and the values of the functions (x_1, x_2) on the interval $[s_1 - d(s_1), T]$. Consider the trajectory (q_A, d, y_1, q_B) corresponding to the inputs (x_1, x_2) during $[0, T]$ and s_1, s_2 as defined above. For this trajectory, let $q_A(s_1 - d(s_1)) = \chi$ and $q_B(T) = \theta$. Then from (2.5), $d(s_1) = \chi / c_A$, while the cost incurred by this trajectory is by definition

$$\int_0^T [L_1(\dot{x}_1) + L_2(\dot{x}_2)]\, dt.$$

Now consider the trajectory (q_A', d', y_1', q_B') of another set of inputs (x_1', x_2') on $[0, T]$ defined by

$$\begin{aligned} &(\dot{x}_1', \dot{x}_2') = (b_1, b_2) \quad \text{for all } t \in [0, s_2), \\ &(\dot{x}_1', \dot{x}_2') = (\beta_1, \beta_2) \quad \text{for all } t \in [s_2, s_1 - d(s_1)), \\ &(\dot{x}_1', \dot{x}_2') = (\dot{x}_1, \dot{x}_2) \quad \text{for all } t \in [s_1 - d(s_1), T], \end{aligned} \tag{B.4}$$

where $\beta_1 = \bar{\dot{x}}_1$ and $\beta_2 = \bar{\dot{x}}_2$ equal the average input velocities over the interval $[s_2, s_1 - d(s_1)]$,

$$\bar{\dot{x}}_i = \frac{1}{s_1 - d(s_1) - s_2} \int_{s_2}^{s_1 - d(s_1)} \dot{x}_i(t)\, dt.$$

We have assumed the stability conditions $b_1 + b_2 < c_A$ and $b_1 < c_B$. So from (4.1) and (4.4), $q_A'(s_2) = q_B'(s_2) = 0$. By the definitions of s_1 and s_2, $q_A(t) > 0$ for $t \in [s_2, s_1)$ which implies that $\bar{\dot{x}}_1 + \bar{\dot{x}}_2 > c_A$. Then from (4.1),

$$\begin{aligned} q_A'(s_1 - d(s_1)) &= \int_{s_2}^{s_1 - d(s_1)} (\bar{\dot{x}}_1 + \bar{\dot{x}}_2 - c_A)\, dt \\ &= \int_{s_2}^{s_1 - d(s_1)} (\dot{x}_1 + \dot{x}_2 - c_A)\, dt = q_A(s_1 - d(s_1)). \end{aligned}$$

Equation (2.5) then implies that $d'(s_1) = d(s_1)$. Since (x_1', x_2') agrees with (x_1, x_2) on $[s_1 - d(s_1), T]$, $q_A(s_1 - d(s_1)) = q_A'(s_1 - d'(s_1))$, $q_B(s_1) = 0$ and $q_B'(s_1) \geq 0$, it follows from (B.3) that $q_B'(T') = q_B(T) = \theta$ for some $T' \leq T$. Now the convexity of L_1 and L_2 implies that for any interval $[t_1, t_2]$,

$$\int_{t_1}^{t_2} [L_1(\dot{x}_1) + L_2(\dot{x}_2)]\, dt \geq (t_2 - t_1)[L_1(\bar{\dot{x}}_1) + L_2(\bar{\dot{x}}_2)], \tag{B.5}$$

where $\bar{\dot{x}}_1$ and $\bar{\dot{x}}_2$ are the averages of the functions $\dot{x}_1$ and $\dot{x}_2$ respectively over the interval $[t_1, t_2]$. Consider the cost of the new trajectory (x_1', x_2') on $[0, T']$. Using (B.5) and the fact that $L_1(b_1) = L_2(b_2) = 0$,

$$\begin{aligned}
\int_0^{T'} [L_1(\dot{x}_1') + L_2(\dot{x}_2')]\, dt &\leq \int_0^{T} [L_1(\dot{x}_1') + L_2(\dot{x}_2')]\, dt \\
&= \int_{s_2}^{s_1 - d(s_1)} [L_1(\bar{\dot{x}}_1) + L_2(\bar{\dot{x}}_2)]\, dt \\
&\quad + \int_{s_1 - d(s_1)}^{T} [L_1(\dot{x}_1) + L_2(\dot{x}_2)]\, dt \\
&\leq \int_{s_2}^{T} [L_1(\dot{x}_1) + L_2(\dot{x}_2)]\, dt \\
&\leq \int_0^{T} [L_1(\dot{x}_1) + L_2(\dot{x}_2)]\, dt.
\end{aligned}$$

Thus given any input (x_1, x_2) with $(F_T(x_1, x_2))_2 = \theta$, there exists another input (x_1', x_2') with no greater cost, satisfying $(F_{T'}(x_1', x_2'))_2 = \theta$ for $T' \leq T$ such that its trajectory (q_A', d', y_1', q_B') satisfies

$$\begin{aligned}
&q_A'(s_2) = q_B'(s_2) = 0, \qquad q_A'(t) > 0 \quad \text{for all } t \in [s_2, s_1], \\
&q_B'(T') = \theta, \qquad q_B'(t) = 0 \quad \text{for all } t \in [s_2, s_1]
\end{aligned} \tag{B.6}$$

and

$$(\dot{x}_1', \dot{x}_2') = (\beta_1, \beta_2) \quad \text{for all } t \in [s_2, s_1 - d(s_1)].$$

This verifies the claim and shows that $K(\theta)$ is achieved on the set of functions that have the form described in (B.4). Now $K(\theta)$ involves an infimization over $T \in [0, \infty)$, where $[0, T]$ is the domain on which the functions are defined. Recall that $L_1(b_1) = L_2(b_2) = 0$, where b_1 and b_2 are the mean input rates of sources 1 and 2, respectively. This implies that the domain of the input functions can be augmented by appending the mean flow at the beginning of the interval without any change in the total cost. Thus minimizing trajectories are not unique. Since $(x_1'(t), x_2'(t)) = (b_1 t, b_2 t)$ during $[0, s_2]$, one can equivalently consider the functions (x_1', x_2') restricted to $[s_2, T']$. Since $q_A'(s_2) = q_B'(s_2) = 0$, the initial conditions are then satisfied at s_2. Thus one can without loss of generality set $s_2 = 0$ in (B.6) and only consider inputs (x_1, x_2) on $[0, T]$ for some

$T < \infty$ that satisfy

$$\begin{aligned} &q_A(t) > 0 \quad \text{and} \quad q_B(t) = 0 \quad \text{for all } t \in [0, s_1], \\ &(\dot{x}_1, \dot{x}_2) = (\beta_1, \beta_2) \quad \text{for all } t \in [0, s_1 - d(s_1)], \\ &q_B(t) > 0 \quad \text{for } t \in (s_1, T) \quad \text{and} \quad q_B(T) = \theta. \end{aligned} \tag{B.7}$$

We shall now further characterize the set on which the infimum $K(\theta)$ is achieved. We note that the conditions (B.7) satisfied by the input functions impose no restrictions on the values of q_A on the interval $(s_1, T]$. Thus we consider four cases corresponding to the different possible values that q_A can assume on the interval $(s_1, T]$ and ascertain the variational problems that the minimizing functions in each of those cases satisfy. We first consider the case when buffer A is empty throughout (s_1, T).

CASE 1: $q_A(t) = 0$ for all $t \in (s_1, T)$. From the definition of s_2, $s_2 = s_1$. Moreover, since it has been shown that without loss of generality s_2 can be taken to be zero, it follows that $s_1 = 0$. So in this case, buffer A is empty throughout the interval $[0, T]$. From (4.1), this happens if and only if for all $t \in [0, T]$,

$$\dot{x}_1(t) + \dot{x}_2(t) \le c_A. \tag{B.8}$$

Moreover, from (2.5) and (4.3), for all $t \in [0, T]$,

$$d(t) = 0 \quad \text{and} \quad y_1(t) = x_1(t).$$

We also note that the definition of s_1 and the conditions (B.7) imply that

$$q_B(t) > 0 \quad \text{for } t \in (0, T) \quad \text{and} \quad q_B(T) = \theta.$$

We now show that without loss of generality, the velocities of the inputs can be assumed to be constant throughout $[0, T]$. This follows from the fact that given any input velocities (x_1, x_2) during $[0, T]$, they can be replaced by the average velocities over the same interval and the resulting trajectory still achieves $q_B(T) = \theta$ at no greater cost. We define the new trajectory $(\dot{x}_1'(t), \dot{x}_2'(t)) = (\beta_1, \beta_2)$ for $t \in [0, T]$, where

$$\beta_1 = \frac{1}{T} \int_0^T \dot{x}_1(t)\, dt, \qquad \beta_2 = \frac{1}{T} \int_0^T \dot{x}_2(t)\, dt.$$

From (4.4) and (B.2),

$$\dot{q}_B'(t) = \dot{y}_1'(t) - c_B = \dot{x}_1'(t) - c_B.$$

Integrating both sides,

$$\begin{aligned} q_B'(T) &= \int_0^T (\beta_1 - c_B)\, dt = (\beta_1 - c_B)T \\ &= \left(\frac{1}{T} \int_0^T \dot{x}_1(t)\, dt - c_B \right) T = q_B(T) \\ &= \theta. \end{aligned} \tag{B.9}$$

The new trajectory incurs no greater cost since by the convexity of the L-functions and Jensen's inequality, as in (B.5),

$$\int_0^T [L_1(\dot{x}_1') + L_2(\dot{x}_2')]\,dt \le \int_0^T [L_1(\dot{x}_1) + L_2(\dot{x}_2)]\,dt.$$

Thus the input velocities may be assumed to be constant and the constraint (B.8) that ensures that $q_A(t) = 0$ for $t \in [0, T]$ becomes

$$\beta_1 + \beta_2 \le c_A.$$

Moreover, from equation (B.9),

$$q_B'(T) = \theta = (\beta_1 - c_B)T \quad \Rightarrow \quad T = \frac{\theta}{\beta_1 - c_B},$$

where $\theta > 0$. Note that here θ could as well represent the difference in the buffer B content at time T from its initial value and does not require that buffer B be initially empty. This property will be used in the simplification of Case 4. From the equation above, the fact that $T \in [0, \infty)$ leads to the constraint $\beta_1 > c_B$. The total cost incurred is

$$T[L_1(\beta_1) + L_2(\beta_2)] = \frac{\theta}{\beta_1 - c_B}[L_1(\beta_1) + L_2(\beta_2)].$$

Clearly the constraint $\beta_1 > c_B$ can be relaxed to $\beta_1 \ge c_B$. Thus in this case, $K_1(\theta)$, the minimum cost to raise buffer B to the value θ is given by the variational problem described in (8.3):

$$K_1(\theta) = \theta K_1(1),$$

where

$$K_1(1) = \inf_{\beta_1, \beta_2} \frac{[L_1(\beta_1) + L_2(\beta_2)]}{\beta_1 - c_B},$$

subject to

$$\beta_1 + \beta_2 \le c_A, \qquad \beta_1 \ge c_B.$$

CASE 2: $q_A(t) > 0$ for all $t \in (0, T)$ and $s_1 = 0$. In this case buffer A and buffer B are nonempty on $(0, T)$. From the system of equations (B), all functions in this case can be characterized as satisfying the conditions

$$\begin{aligned} &\int_0^t (\dot{x}_1(s) + \dot{x}_2(s) - c_A)\,ds > 0 \quad \text{for all } t \in (0, T), \\ &\int_0^t \left(\frac{\dot{x}_1(s - d(s))}{\dot{x}_1(s - d(s)) + \dot{x}_2(s - d(s))} c_A - c_B \right) ds > 0 \quad \text{for all } t \in (0, T). \end{aligned} \tag{B.10}$$

Using convexity arguments, we establish that the minimizing inputs can be assumed to have piecewise constant velocities. As in the previous case, this is a consequence of the fact that trajectories can be replaced by their averages on certain intervals with no change in the final buffer B content and at no

greater cost. Since we have zero initial conditions and $q_B(s) > 0$ for $s \in (0, T)$, from the system of equations (B) we obtain

$$\begin{aligned} q_B(T) &= y_1(T) - c_B T \\ &= x_1(T - d(T)) - c_B T. \end{aligned}$$

We replace the trajectories (x_1, x_2) by new trajectories (x_1', x_2') such that $(\dot{x}_1', \dot{x}_2') = (\delta_1, \delta_2)$ on $[0, T - d(T))$ and $(\dot{x}_1', \dot{x}_2') = (b_1, b_2)$ on $[T - d(T), T]$, where (δ_1, δ_2) are average velocities given by

$$\delta_1 = \frac{1}{T - d(T)} \int_0^{T-d(T)} \dot{x}_1(s)\, ds, \qquad \delta_2 = \frac{1}{T - d(T)} \int_0^{T-d(T)} \dot{x}_2(s)\, ds.$$

Let (q_A', d', y_1', q_B') be the trajectories of the inputs (x_1', x_2') as defined through the system of equations (B). Since $q_A(s) > 0$ on $s \in (0, T)$, we infer from (4.1) that

$$\begin{aligned} q_A'(T - d(T)) &= (\delta_1 + \delta_2 - c_A)[T - d(T)] \\ &= \int_0^{T-d(T)} (\dot{x}_1(s) + \dot{x}_2(s))\, ds \\ &= q_A(T - d(T)). \end{aligned}$$

Thus by (2.5), $d(T) = d'(T)$. If $q_B(T) = \theta$, then

$$\begin{aligned} q_B'(T) &= y_1'(T) - c_B T \\ &= x_1'(T - d'(T)) - c_B T \\ &= x_1(T - d(T)) - c_B T \\ &= y_1(T) - c_B T \\ &= \theta. \end{aligned}$$

Thus the new trajectories raise the contents of buffer B to the same value θ. Note that the proof does not require $q_B(0) = 0$ and hence θ can be regarded as the difference $q_B(T) - q_B(0)$. From the dynamics we have

$$q_B'(T) = \theta = [T - d(T)]\left(\frac{\delta_1}{\delta_1 + \delta_2} c_A - c_B \right). \tag{B.11}$$

Using the convexity of the L-functions as expressed in (B.5), we show that the new trajectory incurs lower cost since

$$\begin{aligned} \int_0^T [L_1(\dot{x}_1(s)) + L_2(\dot{x}_2(s))]\, ds &\geq \int_0^{T-d(T)} [L_1(\dot{x}_1(s)) + L_2(\dot{x}_2(s))]\, ds \\ &= [T - d(T)][L_1(\delta_1) + L_2(\delta_2)]. \end{aligned}$$

The last expression represents the cost for the new trajectories since $L_1(b_1) = L_2(b_2) = 0$ and hence no cost is incurred during the interval $[T - d(T), T]$. Thus we can assume without loss of generality that all trajectories have the

same form as (x_1', x_2'). We note that the conditions specified in (B.10) are satisfied if and only if (δ_1, δ_2) satisfy

$$\delta_1 + \delta_2 > c_A, \qquad \frac{\delta_1}{\delta_1 + \delta_2} c_A > c_B.$$

In order to determine the minimizing trajectories, it only remains to find the values of (δ_1, δ_2) that yield the lowest cost. Using (B.11), we see that the cost is linear in θ and can be expressed as

$$\frac{[L_1(\delta_1) + L_2(\delta_2)]}{(\delta_1/(\delta_1 + \delta_2))c_A - c_B}\theta.$$

If we define $K_2(\theta)$ to be the infimum of the cost over all functions satisfying conditions (B.10) and having a final buffer B content of θ, we obtain

$$K_2(\theta) = \theta K_2(1),$$

where

$$K_2(1) = \inf_{\delta_1, \delta_2} \frac{[L_1(\delta_1) + L_2(\delta_2)]}{(\delta_1/(\delta_1 + \delta_2))c_A - c_B},$$

subject to

$$\delta_1 + \delta_2 > c_A, \qquad \frac{\delta_1}{\delta_1 + \delta_2} c_A \geq c_B.$$

Note that the strict equality can be relaxed in the second constraint because the infimum is clearly not achieved when the equality holds. Thus we have derived the variational problem described in (8.4).

CASE 3: $q_A(t) > 0$ for all $t \in (s_1, T)$ and $s_1 > 0$. In this case, the definition of s_1 and the conditions (B.7) imply that $q_A(t) > 0$ for all $t \in (0, T)$, $q_B(t) = 0$ for all $t \in [0, s_1]$, $q_B(t) > 0$ for all $t \in (s_1, T)$ and $q_B(T) = \theta$. From the system of equations (B), it can be seen that this is satisfied if and only if

$$\begin{aligned}
&\int_0^t [\dot{x}_1(s) + \dot{x}_2(s) - c_A]\, ds > 0 \quad \text{for all } t \in (0, T), \\
&\frac{\dot{x}_1(t - d(t))}{\dot{x}_1(t - d(t)) + \dot{x}_2(t - d(t))} c_A - c_B \leq 0 \quad \text{for a.e. } t \in [0, s_1], \\
&\int_{s_1}^t \left(\frac{\dot{x}_1(s - d(s))}{\dot{x}_1(s - d(s)) + \dot{x}_2(s - d(s))} c_A - c_B \right) ds > 0 \quad \text{for all } t \in (s_1, T).
\end{aligned} \tag{B.12}$$

We assume for the rest of the discussion of this case that input functions satisfy these conditions. We will prove that minimizing input functions can be assumed to have piecewise constant velocities, using the fact that any function can be replaced by one with piecewise constant derivatives that incurs no greater cost to raise the level of buffer B to θ. Note from (B.3) that the value of $q_B(T)$ only depends on $q_A(s_1 - d(s_1))$ and the input (x_1, x_2) during the interval $[s_1 - d(s_1), T - d(T)]$. Let $q_A(s_1 - d(s_1)) = \chi$. Then it can be assumed that the velocities are constant in the interval $[0, s_1 - d(s_1)]$, since inputs with the

average rate during that time interval yield the same value for $q_A(s_1 - d(s_1))$ with no greater cost. This is similar to the argument made in Case 1. We define the new trajectories (x_1', x_2') with $(\dot{x}_1'(t), \dot{x}_2'(t)) = (\beta_1, \beta_2)$ for $t \in [0, s_1 - d(s_1)]$, where

$$\beta_1 = \frac{1}{s_1 - d(s_1)} \int_0^{s_1 - d(s_1)} \dot{x}_1(t)\, dt, \qquad \beta_2 = \frac{1}{s_1 - d(s_1)} \int_0^{s_1 - d(s_1)} \dot{x}_2(t)\, dt.$$

The constraints (B.12) for the interval $[0, s_1 - d(s_1)]$ then become

$$\beta_1 + \beta_2 > c_A, \qquad \frac{\beta_1}{\beta_1 + \beta_2} c_A \le c_B. \tag{B.13}$$

In the interval $[0, s_1 - d(s_1)]$, since $q_A(t) > 0$, from (4.1),

$$\begin{aligned} q_A'(s_1 - d(s_1)) &= (\beta_1 + \beta_2 - c_A)(s_1 - d(s_1)) \\ &= \int_0^{s_1 - d(s_1)} (\dot{x}_1(t) + \dot{x}_2(t) - c_A)\, dt \\ &= q_A(s_1 - d(s_1)). \end{aligned}$$

Thus from (2.5), $d(s_1) = d'(s_1)$ and the new trajectory incurs no greater cost during that interval because by convexity, as in (B.5), the new cost is less than or equal to the old cost.

Thus we have shown that the inputs on $[0, s_1 - d(s_1)]$ can be assumed to have constant velocity (β_1, β_2). We now try to determine the values of (β_1, β_2) that minimize the contribution to the cost in that interval, which is given by

$$(s_1 - d(s_1))[L_1(\beta_1) + L_2(\beta_2)].$$

Since from (4.1), $(s_1 - d(s_1)) = \chi/(\beta_1 + \beta_2 - c_A)$, where $\chi = q_A(s_1 - d(s_1))$, it follows that the cost to raise buffer A to a height χ is linear in χ. Thus $K_4(\chi) = \chi K_4(1)$, where $K_4(1)$ is the minimum cost that must be incurred to raise buffer A to a height 1. $K_4(1)$ is easily seen to be given by the variational problem stated in (8.5),

$$K_4(1) = \inf_{\beta_1, \beta_2} \frac{[L_1(\beta_1) + L_2(\beta_2)]}{\beta_1 + \beta_2 - c_A},$$

subject to

$$\beta_1 + \beta_2 \ge c_A, \qquad \frac{\beta_1}{\beta_1 + \beta_2} c_A \le c_B.$$

Note that the infimum above is certainly not achieved when $\beta_1 + \beta_2 = c_A$ since for those inputs, $K_4(\chi)$ is infinite. This justifies relaxing the strict inequality in (B.13) to obtain the constraint $\beta_1 + \beta_2 \ge c_A$.

We now consider the second interval. Since $q_A(t) > 0$ in the interval $(s_1 - d(s_1), T)$ and $q_B(t) > 0$ for $t \in (s_1, T]$, from (4.1) it follows that

$$q_A(T - d(T)) - \chi = \int_{s_1 - d(s_1)}^{T - d(T)} (\dot{x}_1(t) + \dot{x}_2(t) - c_A)\, dt,$$

and from (4.4) and (4.3) we conclude that

$$\begin{aligned}\theta = q_B(T) &= y_1(T) - y_1(s_1) - c_B(T - s_1)\\ &= x_1(T - d(T)) - x_1(s_1 - d(s_1)) - c_B(T - s_1).\end{aligned}$$

This last equation shows that replacing the input velocities by their respective averages in the interval $[s_1 - d(s_1), T - d(T)]$ leaves $q_A(T - d(T))$, $d(T)$, $x_1(T - d(T))$ and therefore $q_B(T)$ unaltered without any increase in the cost incurred during the interval. Thus without loss of generality, one can assume that the input velocities of the new trajectories $(\dot{x}_1', \dot{x}_2') = (\delta_1, \delta_2)$ are constant on $[s_1 - d(s_1), T - d(T)]$. Then the third condition in (B.12) takes the form

$$\frac{\delta_1}{\delta_1 + \delta_2} c_A > c_B.$$

At this point we must consider the two cases $\delta_1 + \delta_2 \geq c_A$ and $\delta_1 + \delta_2 < c_A$. However, the first case makes no use of the nonzero level of buffer A at time $s_1 - d(s_1)$. It is easy to check that if $\delta_1 + \delta_2 \geq c_A$, then the associated trajectory and cost are suboptimal when compared with the trajectory which uses these inputs during the first interval in Case 2. Thus we can assume $\delta_1 + \delta_2 < c_A$ for the remainder of this case.

As noted in Section 8, the new inputs δ_1 and δ_2 do not affect the output of buffer A until $d(s_1) = \chi/c_A$ units of time have passed. It will take an additional

$$\frac{\chi}{c_A} + \theta \frac{\delta_1 + \delta_2}{\delta_1 c_A - (\delta_1 + \delta_2) c_B}$$

units of time to raise buffer B to level θ. The constraint $q_A(t) > 0$ for $t \in (0, T)$ implies

$$\frac{\chi}{c_A} + \theta \frac{\delta_1 + \delta_2}{\delta_1 c_A - (\delta_1 + \delta_2) c_B} \leq \frac{\chi}{(c_A - \delta_1 - \delta_2)}. \tag{B.14}$$

However, since the combined cost over the two intervals,

$$\chi K_4(1) + \left(\frac{\chi}{c_A} + \frac{\theta(\delta_1 + \delta_2)}{\delta_1 c_A - (\delta_1 + \delta_2) c_B} \right) [L_1(\delta_1) + L_2(\delta_2)],$$

is increasing in χ, for any given pair δ_1, δ_2 the constraint (B.14) must be tight if the trajectory is to be optimal. This implies

$$\frac{\chi}{\theta} = \frac{c_A(c_A - \delta_1 - \delta_2)}{\delta_1 c_A - (\delta_1 + \delta_2) c_B}.$$

In particular, the inputs δ_1 and δ_2 must be applied for $\theta c_A / [\delta_1 c_A - (\delta_1 + \delta_2) c_B]$ units of time for the trajectory to be optimal. This implies $q_A(T) = 0$, and therefore $d(T) = 0$. Since it is evident that the total cost is linear in θ, we obtain the variational problem for $K_3(\theta)$ stated in (8.6):

$$K_3(\theta) = \theta K_3(1),$$

where

$$K_3(1) = \inf_{\delta_1,\,\delta_2} c_A\left[\frac{c_A - \delta_1 - \delta_2}{\delta_1 c_A - (\delta_1+\delta_2)c_B}K_4(1) + \frac{[L_1(\delta_1)+L_2(\delta_2)]}{\delta_1 c_A - (\delta_1+\delta_2)c_B}\right],$$

subject to

$$\frac{\delta_1}{\delta_1+\delta_2}c_A \geq c_B, \qquad \delta_1+\delta_2 < c_A.$$

CASE 4: $q_A(t) = 0$ for $\mathscr{K} \subset (s_1, T)$, $\mathscr{K} \neq \emptyset, (s_1, T)$. We finally consider the case when buffer A is not uniformly empty or nonempty on the interval (s_1, T). We define a *bump* of $f \in \mathscr{C}[0,T]$ to be a nonempty closed interval $[t_1, t_2] \in [0,T]$ such that $f(t_1) = f(t_2) = 0$, and $f(t) > 0$ for all $t \in (t_1, t_2)$. Since $q_A(t)$ is continuous, it can have only a countable number of bumps on $[0,T]$. From the definition of s_1, $q_A(t) > 0$ for all $t \in (0, s_1)$, and thus q_A cannot have bumps that are contained in $(0, s_1)$. Moreover, any bump $[t_1, t_2]$ with $t_1 < s_1$ satisfies $t_1 = 0$ and $t_2 > s_1$. Now consider an input function (x_1, x_2) for which $q_A(t)$ has a bump $[t_1, t_2]$ for some $t_1 \in [s_1, T]$. This implies that for all $t \in (t_1, t_2)$, $q_A(t) > 0$ and by the definition of s_1, $q_B(t) > 0$. So we define new trajectories (x_1', x_2') as follows:

$$(\dot{x}_1', \dot{x}_2') = (\bar{\dot{x}}_1, \bar{\dot{x}}_2) \quad \text{for all } t \in (t_1, t_2),$$
$$(\dot{x}_1', \dot{x}_2') = (\dot{x}_1, \dot{x}_2) \quad \text{otherwise,}$$

where $\bar{\dot{x}}_1$ and $\bar{\dot{x}}_2$ are the usual averages of ($\dot{x}_1$ and $\dot{x}_2$, respectively, over the interval (t_1, t_2),

$$\bar{\dot{x}}_1 = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \dot{x}_1(t)\,dt,$$
$$\bar{\dot{x}}_2 = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \dot{x}_2(t)\,dt.$$

Then from (4.4) and the fact that $d(t_1) = d(t_2) = 0$, it is clear that $q_B'(T) = q_B(T)$ and by (B.5), the cost of the new trajectory is less than or equal to the cost of the original one. Since $q_A(t_2) = 0$, (4.1) implies that $\bar{\dot{x}}_1 + \bar{\dot{x}}_2 \leq c_A$ and $q_A'(t) = 0$ for all $t \in (t_1, t_2)$. Thus the new trajectory has no bump contained in $[s_1, T]$. The argument generalizes in the obvious manner to input functions (x_1, x_2) with countably many bumps. Therefore, without loss of generalization, one can assume that the minimizing trajectory has at most one bump $[0, t_2]$ that satisfies $t_2 > s_1$. Then all possible configurations for the minimizing trajectories in this case are shown in Figure 11. The piecewise linear nature of the trajectories and the strictly monotonically increasing nature of q_B can be deduced from the dynamics using the same convexity arguments that were used in the previous cases. For all three configurations, the domain of the q_A trajectory can be broken up into a finite number of intervals, in each of which the assumptions of Case 1, Case 2 or Case 3 are satisfied. Thus one would expect the trajectories in each of those intervals to have the same velocities as the minimizing trajectory of the corresponding case. This argument

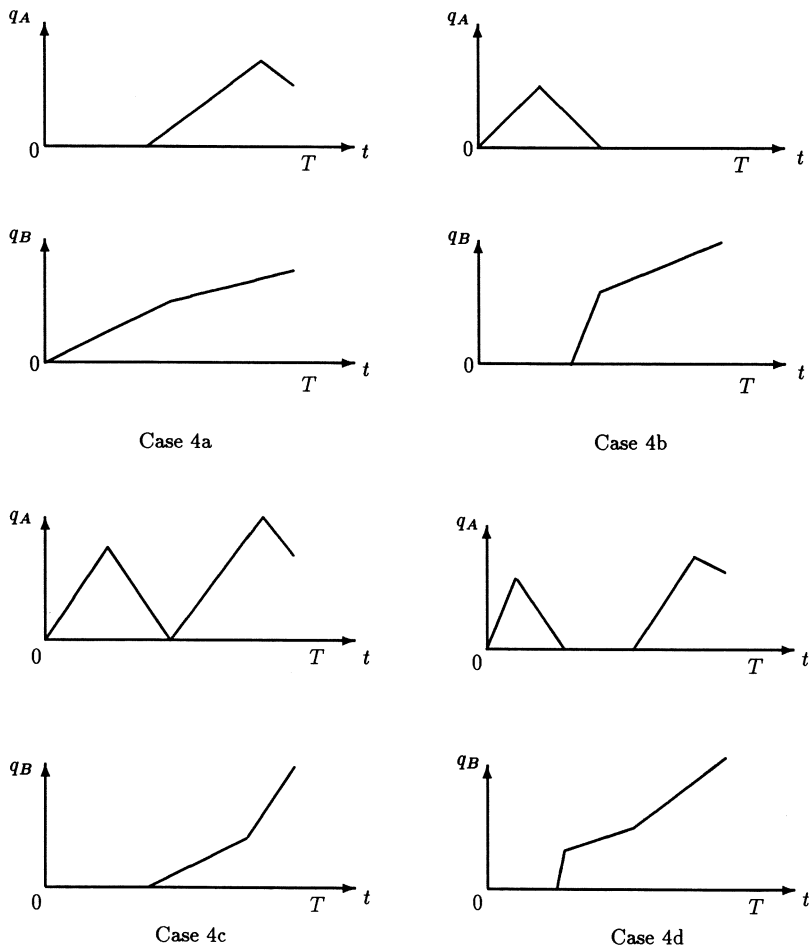

FIG. 11. *Possible configurations for Case* 4.

can be made rigorous by extending the techniques employed previously in a straightforward manner.

We have noted that the minimizing trajectories of Case 4 are concatenations of the minimizing trajectories of Cases 1, 2 and 3. It remains to show that any such concatenation incurs a cost no less than that incurred by the minimizing trajectory of the variational problem (B.1). We shall outline the argument only for the configuration Case 4b in Figure 11 since the arguments for the other two figures follow in the same manner. Let $\theta' = q_B(t_2)$ and let C_1 and C_2 be the costs incurred by the trajectory during $[0, t_2)$ and $[t_2, T]$ respectively. Then since $q_A(t) > 0$ for every $t \in (0, t_2)$ and $q_A(t_2) = 0$, the assumptions leading to the variational problem $K_3(1)$ in Case 3 are satisfied during that interval and so

$$C_1 \geq K_3(\theta') = \theta' K_3(1).$$

Similarly, since $q_A(t) = 0$ for every $t \in [t_2, T]$, the assumptions of Case 1 are satisfied during that interval and hence

$$C_2 \geq K_1(\theta - \theta') = (\theta - \theta')K_1(1).$$

Note that this holds even though $q_B(t_2) > 0$, since in the proofs of the earlier cases, K_1 and K_2 were established to be independent of the initial value of the buffer B and only a function of θ, the net change in the buffer B content. This does not hold for K_3. However, since in all the possible concatenations, the minimizing trajectories of the variational problem in $K_3(1)$ arise only during the first interval when the buffer B is initially empty, the same argument works for the other configurations as well.

Therefore the total cost of the trajectory in this case is

$$\begin{aligned} C_1 + C_2 &\geq K_3(\theta') + K_1(\theta - \theta') \\ &= \theta' K_3(1) + (\theta - \theta')K_1(1) \\ &\geq \theta(K_3(1) \wedge K_1(1)). \end{aligned}$$

Thus since the cost for functions considered in Case 4 is always higher than the minimum of the costs in Cases 1, 2 and 3, the theorem is proved. □

Acknowledgments. We thank Kurt Majewski for pointing out that the quantity $K_3(1)$ defined in Section 8 was off by a multiplicative constant in the original version of this paper.

REFERENCES

[1] ANICK, D., MITRA, D. and SONDHI, M. M. (1982). Stochastic theory of a data handling system with multiple sources. *Bell Syst. Tech. J.* **61** 1871–1894.

[2] BREIMAN, L. (1968). *Probability*. Addison-Wesley, Reading, MA.

[3] CHANG, C.-S. (1993). Approximations of ATM networks: effective bandwidths and traffic descriptors. Technical report IBM RC 18954, T.J. Watson Research Center.

[4] DEMBO, A. and ZEITOUNI, O. (1993). *Large Deviations Techniques and Applications*. Jones and Bartlett, Boston.

[5] DE VECIANA, G., COURCOUBETIS, C. and WALRAND, J. (1993). Decoupling bandwidths for networks: a decomposition approach to resource management. Memorandum UCB/ERL M93/50, Univ. California, Berkeley, CA.

[6] DUPUIS, P. and ELLIS, R. S. (1996). *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley, New York.

[7] DUPUIS, P. and ISHII, H. (1991). On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications. *Stochastics Stochastics Rep.* **35** 31–62.

[8] DUPUIS, P. and NAGURNEY, A. (1993). Dynamical systems and variational inequalities. *Ann. Oper. Res.* **44** 9–42.

[9] DUPUIS, P. and WILLIAMS, R. (1994). Lyapunov functions for semimartingale reflecting Brownian motions. *Ann. Probab.* **22** 680–702.

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

[11] ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes: Characterization and Convergence*. Wiley, New York.

[12] FREIDLIN, M. I. and WENTZELL, A. D. (1984). *Random Perturbations of Dynamical Systems*. Springer, New York.

[13] HSU, I. and WALRAND, J. (1995). Admission control for ATM networks. In *Stochastic Networks* (F. P. Kelly and R. J. Williams, eds.) 411–427. Springer, New York.
[14] KELLY, F. P. (1991). Effective bandwidths at multi-class queues. *Queueing Systems* **9** 5–16.
[15] KESIDIS, G., WALRAND, J. and CHANG, C. S. (1993). Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Trans. Networking* **1** 424–428.
[16] LOYNES, R. M. (1962). The stability of a queue with non-indepedent inter-arrivals and service times. *Math. Proc. Cambridge Philos. Soc.* **58** 497–520.
[17] MAJEWSKI, K. (1996). Large deviations of feedforward queueing networks. Ph.D. thesis, Ludwig-Maximilian-Univ., München.
[18] O'BRIEN, G. L. and VERVAAT, W. (1991). Capacities, large deviations and loglog laws. In *Stable Processes and Related Topics* (S. Cambanis, G. Samorodnitsky and M. Taqqu, eds.) 43–84. Birkhäuser, Boston.
[19] O'CONNELL, N. (1995). Large deviations in queueing networks. Preprint.
[20] PUHALSKII, A. A. (1991). On functional principles of large deviations. In *New Trends in Probability and Statistics* (V. Sazonov and T. Shervashidze, eds.) 198–218. VSP-Mokslas, Utrecht.
[21] RAMANAN, K. and DUPUIS, P. (1995). Large deviation properties of data streams that share a buffer. LCDS Technical Report 95-8, Brown Univ.
[22] SHWARTZ, A. and WEISS, A. (1995). *Large Deviations for Performance Analysis: Queues, Communication, and Computing*. Chapman and Hall, New York.
[23] STROOCK, D. W. (1984). *An Introduction to the Theory of Large Deviations*. Springer, New York.
[24] WEISS, A. (1986). A new technique for analyzing large traffic systems. *Adv. in Appl. Probab.* **21** 506–532.

LEFSCHETZ CENTER FOR DYNAMICAL SYSTEMS
DIVISION OF APPLIED MATHEMATICS
BROWN UNIVERSITY
PROVIDENCE, RHODE ISLAND 02912
E-MAIL: kavita_ramanan@brown.edu
dupuis@cfm.brown.edu